UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Estudios Superiores Acatlán

Toma de decisiones con Minería de datos para un portafolio de crédito: Simulación por componentes principales

Tesina para obtener el título de

Licenciado en Actuaría

Presenta

Juan Fabrizio Sánchez Jiménez

Asesor

Mtro. Gabriel Delgado Juárez

Abril 2019

Santa Cruz Acatlán, Naucalpan, Estado de México. 2019

Índice general

«Any one who considers arithmetical methods

of producing random digits is, of course,

in a state of sin.» — John Von Neumann

Agradecimientos

A mis padres que me apoyaron en cada momento de mi vida con cariño y comprensión. Mi amor y gratitud a mi madre quien me compartio el amor a las matematicas. A mi padre por enseñarme la importancia de la constancia.

A mi incasable esposa Julieta, quien nunca dejo que bajara los brazos y quien siempre tuvo una palabra de aliento y superación para mi. Para ti Todo mi amor.

A mis abuela Teresa y Zoila quienes me llenan de amor. A mis tios que vieron por mi cada dia. Al recuerdo de Felix y Alfonso.

A Felisa, Sergio y Laura por todo su asesoria, apoyo y cariño en todo momento.

A mi Asesor Gabriel, amigo y maestro.

A mis maestros Mahil, Victor, Gamaliel y Daniel a quienes les debo el amor a esta profesión. A mis sinodos Luz Maria, Pablo y Espartaco, quienes me retroalimentaron de manera formidable..

A mi querida Universidad.

Resumen

La simulación correlacionada por componentes principales, implementada través de minería de datos, es una herramienta análitica útil para la toma de decisiones estratégicas. La motivación del autor, es mostrar al lector el desarrollo funcional de la solución de una problemática en el sector crediticio. En el documento se muestran las comparativas de diversas técnicas de minería de datos y simulación correlacionada, así como, los softwares estadísticos utilizados. El análisis se realiza sobre una de las entidades bancarias P2P más relevantes en Estados Unidos para créditos revolventes activos. Para el desarrollo del proyecto se utiliza como marco de referencia a CRISP-DM, metodología de minería de datos que apoya de inicio a fin, la construcción de una solución con uso de información; teniendo como focos el conocimiento del negocio, las bases de datos, el modelado, la simulación hasta las conclusiones.

En los resultados se observa que es posible replicar el comportamiento del portafolio de crédito con una simulación correlacionada. Además, la función dependiente, que es el resultado deseado para el negocio, también cumple los requisitos deseados estadísticamente. Con esta herramienta de simulación, es posible la instrumentación de estrategias ante las modificaciones del portafolio de crédito que se podrían realizar.

Palabras clave: Minería de datos, Simulación Multivariada, Python, Portafolio de crédito, Análisis de Componentes Principales, Simulación Monte Carlo.

Abstract

The correlated simulation by principal components implemented through data mining, is a useful tool for making strategic decisions. The motivation of the author is to show the reader the functional development of the solution of a problem in retail banking. The document shows the comparatives of different methods of data mining and correlated simulation, as well as, the statistical software used. The analysis is performed on one of the most relevant P2P banks in the United States for active credit card. For the development of the project, CRISP-DM is used as reference framework, data mining methodology that supports, from end to end, the construction of a solution with the use of information, focusing on business knowledge, databases, modeling, simulation to conclusions.

The results show that it is possible to replicate the behavior of the credit portfolio with a correlated simulation. In addition, the dependent function, which is the desired result for the business, also meets the statistically desired requirements. With this simulation tool, it is possible to implement strategies before the modifications of the credit portfolio that will be made.

Keywords: Data Mining, Multivariate Simulation, Python, Credit Portfolio, Principal Component Analysis, Monte Carlo Simulation

INTRODUCCIÓN

Los avances científicos y tecnológicos son una parte importante de la toma de decisiones a partir de información. Hoy en día, los directivos y ejecutivos de forma profesional utilizan información estadística confiable, por ello el tratamiento de la información para obtener predicciones certeras requiere de procedimientos innovadores científicamente probadas que dan mayor certeza en términos probabilísticos al utilizar el modelaje para la simulación de situaciones reales en distintas áreas. En el presente trabajo se aborda una metodología y técnicas que representan un ejercicio de simulación robusta y minería de datos.

Dentro de las motivaciones de la construcción de este trabajo, el autor busca compartir al lector un camino para abordar el problema de toma de decisiones a partir de del uso de información, minería de datos y simulación que han sido probadas como eficientes en casos de negocios financieros.

El objetivo principal es mostrar una herramienta de predicción y simulación para la administración de un portafolio de crédito a través de un método de minería de datos que ayudará al equipo ejecutivo y directivo de una empresa en sus estrategias de negocio de forma sencilla y ágil. La minería de datos es uno de los marcos de referencia más aceptados en el sector bancario, pues además de proveer de resultados, está centrada en el conocimiento del negocio y las necesidades de los tomadores de decisiones, tales como la retroalimentación, el análisis descriptivo y la implementación al negocio. Para ello se realizó una comparación de distintas

metodologías de minería de datos y se seleccionó la metodología CRISP-DM para realizar el marco de referencia de la solución del problema. En el contexto de la minería de datos, las herramientas que apoyan el procesamiento masivo de los datos son software avanzados en métodos matemáticos, numéricos y estadísticos, donde el software Python ha tendio relevancia debido a su acelerado uso y crecimiento de sus librerías estadísticas.

El método propuesto en este documento, tiene como objetivo la generación de una herramienta de toma de decisiones por medio de técnicas predictivas y de simulación de variables correlacionadas, que, en sinergia con la minería de datos, aportará por medio de los pasos de la metodología un panorama previo, la representación y exploración de los datos y posteriormente la generación de escenarios en portafolios de crédito. La solución de una problemática de esta índole puede ser desarrollada por un profesionista como el actuario, el cual puede ser eje clave para la solución y optimización de procesos con minería de datos y ciencias afines..

En el primer capítulo se atenderán los antecedentes de minería de datos y análisis de información, la importancia de la correcta implementación de esta, así como las principales áreas de estudio. También se revisan temas de toma de decisiones estratégicas. Por último, se introducen estos temas al mundo del sector bancario y de portafolios de crédito.

En el segundo capítulo, incluimos las herramientas estadísticas que se requieren por la metodología implementada y son utilizadas en el desarrollo del caso de uso. Se realizó una revisión de técnicas descriptivas univariadas y multivariadas, predictivas, comparativas y de simulación.

En el tercer capítulo, se presentan las herramientas y marcos de referencias útiles para la solución de problemas. Se revisan diferentes marcos de referencia de minería de datos y contrastan bajo un enfoque de implementación. También se desarrollan los principales comandos de software seleccionado para este análisis, que es Python. Y al final se hace una revisión de

la metodología de implementación estadística que describe los pasos teóricos de la simulación de un portafolio de crédito.

En el cuarto capítulo, que se desarrolla bajo el enfoque de CRISP-DM y los resultados obtenidos con Python. Se realiza un análisis de la situación del portafolio, la metodología de predicción y la simulación de un portafolio de crédito bancario. Posteriormente, se realiza una primera simulación congruente con el portafolio actual y finalmente se hacen modificaciones en los parametros de la simulación para obtener modificaciones en el portafolio de crédito simulado.

La limitación principal del documento es la profundización en el contraste con otras técnicas que resuelven el mismo problema. En términos técnicos, la metodología de simulación utilizada solo es ocupada en un análisis numérico, dejando los análisis con métodos de variables categóricas para otras investigaciones. Además, se muestra una introducción al software Python, sin embargo, se sugiere al lector informarse más sobre el software y sus utilidades.

Capítulo 1

ANTECEDENTES

1.1. Importancia de la minería de datos para toma de decisiones

Las empresas, (grandes, medianas o pequeñas) requieren tomar decisiones de manera informada y con beneficios observables a corto o largo plazo, aquello que no tenga una mejora cuantitativa o cualitativa, apunta a ser dejado en segundo plano. De este modo, la minería de datos ha tomado terreno y hoy en día tiene un papel preponderante en toma de decisiones y diseño de estrategias de las empresas derivado a su implementación y resultados otorgados.

Algunos autores definen a la minería de datos como el proceso de obtención de conocimiento al examinar una gran cantidad de datos [Perez, 2007]. En otros casos, se toma como una aplicación conjunta de la estadística y las ciencias computacionales [Cruz Arrela, 2010]. Para otros más es la estrategia que a partir del modelado matemático intenta comprender el contenido de los datos [Ruiz Rangel, 2013].

En este documento se tomará la definición de minería de datos como el *"proceso de manejo de información a partir de técnicas computacionales, estadísticas, matemáticas, inferencia o*

simulación que nos apoyen a detectar patrones que nos indiquen un nuevo conocimiento de los datos".

La minería de datos tiene su origen principal en la integración de la estadística con el desarrollo de sistemas computacionales del siglo XX, que durante el proceso de automatización y evolución de las organizaciones, dieron como resultado de las mejoras tecnológicas en materia de procesamiento y almacenamiento, y abrieron de esta rama de aplicación. Para los defensores de las ramas puras, es difícil definir una línea clara entre las ciencias y utilidades que la minería de datos recoge, como lo son: la estadística, la simulación, la computación y la inteligencia artificial. Una de las líneas con mayor discusión es con la estadística, sin embargo, la principal diferencia con la minería de datos es que la estadística postula una hipótesis y el objetivo es aceptarla o rechazarla [UIAF, 2014], mientras que la minería de datos busca la practicidad y las multi-hipótesis.

La idea teórica de la minería de datos se desarrolló durante los 1960's; el manejo de base de datos [Marques, 2009] y el desarrollo de la idea de la extracción de información entonces conocida como *Data Dredging*, fue planteada por Stuart y Selvin en 1966. De manera teórica el Data Dredging o Fishing, postulaba que deberían utilizarse la mayor cantidad de información disponible, a diferencia de las pruebas comunes en su época, donde generalmente se utilizaban muestras para poder crear una probabilidad y el uso de éstas siempre conllevaba una desventaja, como que la muestra no siempre fuera representativa o que, a su vez, se perdieran relaciones importantes en el modelado. En ese momento, se sabía que estas ideas podían ser revolucionarias, en el momento que el almacenamiento de datos y el procesamiento de estos en grandes volúmenes fueran posibles y el análisis de datos no sería un problema para encontrar otras formas de encontrar conocimiento.

En esta época el almacenamiento de datos estaba acotado a dos megabytes. Con el cual

realizar pruebas de modelos que normalmente eran usados por los investigadores y descubiertos en siglos anteriores; Teorema de Bayes (1763) y el análisis de regresión lineal (1805), regresión logística (1958), Redes Neuronales (1943) y Clúster (1954), era una tarea larga y con gran pérdida de recursos y tiempo.

En los años 1970's, los principales avances se dieron en las ramas tecnológicas con la implementación de los sistemas de bases de datos de red, impulsados por Charles Bachman y bajo el sistema Integrated Data Store (IDS), el cual nos entregaría a lo que a la postre se conocería como sistema de red, que adicionaba velocidad y dinamismo a las consultas de datos.

En la misma década Edgar Frank Codd en San José, California, desarrollaría el postulado *"A Relational Model of Data for Large Shared Data Banks"* en el cual propuso el uso de un modelo relacional para el manejo de base de datos, pensado para las tareas del día a día de una institución como lo serían la consulta, carga y manipulación de datos. Representó la idea de un lenguaje universal de consulta y la normalización de las bases de datos. Con esta metodología se desarrolló la mayor parte de la industria de base de datos con DB2 (IBM) con la cual también se estandarizó una de las de ideas de Codd: SQL.

En 1979, se creó el IBM 3370 con 517MB de uso, pero su uso comercial llego a sus manos hasta 1985. Adicional los dispositivos portátiles de usuario se masificaron con el uso de los discos Floppy. Durante esta época ideas desarrolladas teóricamente toman fuerza como las redes neuronales con Teuvo Kohoen y Paul Werbos.

En los años 1980's se observa un desarrollo consistente en el manejo de base de datos incluyendo las consultas y desarrollos de empresas como Oracle, Sybase, DB2 y PostgreSQL. Se crea el concepto de data warehouse, el cual se formalizaría más tarde como un gran espacio lógico que permite el acceso, manipulación, almacenamiento, carga y descarga de información.

Los discos duros iniciaron con espacios desde cinco hasta veinte megabytes. Se masifico el uso de los Floppy y los Disketts para el uso de intercambio de datos. Es durante la época de la expansión de las PC (Personal Computer) en los 1990's, el manejo de datos y la innovación dio dirección hacia un mercado más global. Se generó un camino hacia los repositorios de datos como data warehouse de Inmon y los datamart de Kimball.

Inmon postuló en *"Building the Data Warehouse"*, que el manejo de las bases de datos debe estar basado en un gran almacén de datos, ya que la mayor parte de la información se encuentra en inventario de una entidad. Así, los modelos propuestos para las empresas ofrecen el guardado sistemático, progresivo y ordenado que felicite la extracción, transformación y carga de la información. En el planteamiento formal de Inmon, se define un Data Warehouse como:

- Orientado: Un conjunto de datos enfocado al desarrollo del negocio,
- Integrado: Con estándares en formatos, nombres, estructuras y atributos.
- Variable en el tiempo: diseñado para guardar y manejar un volumen importante de información.
- No Volátil: El significado de la volatilidad en Base de datos, resulta compleja y con un conocimiento incremental, no volátil no significa rígido o sin cambios, por el contrario, una BD busca tener cargas y actualizaciones periódicas. La volatilidad es como el cambio en términos de la estructura y de la consistencia de los datos.

Por su lado Kimball en 1998, presentó la definición de un datamart, los cuales son subconjuntos departamentales de una empresa, ya que las definiciones en un negocio no son iguales, por ejemplo; compras no tiene la misma definición en toda su información que el área de riesgos o recursos humanos. Así el Data Warehouse se integra de diversos Datamart donde cada uno es especificado según las necesidades o granulidad de la información y se basa en los principios:

- Orientación en el negocio
- Infraestructura integrada y de fácil acceso
- Incrementos significativos
- Entregas sistemáticas de valor al negocio
- Adaptativa a los cambios y consistente

Rivadera, en la metodología de Kimball para el diseño de almacenes de datos [Rivadera, 2010], nos comenta que el paradigma Kimball-Inmon es la preocupación o avance de una empresa para el desarrollo de la información.

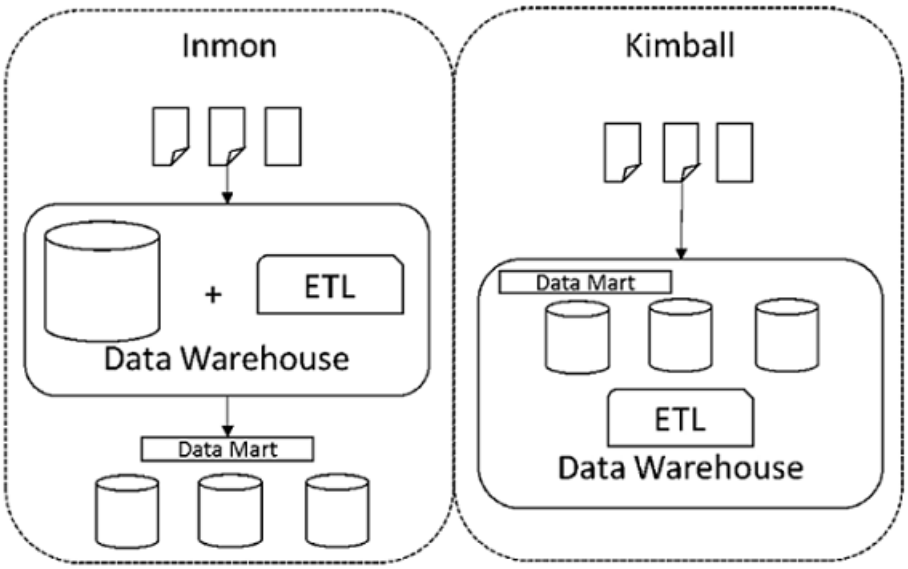


Figura 1.1.1: Diferencias entre los modelos Inmon-Kimball.Elaboración propia

En la figura 1.1.1 podemos observar las diferencias entre un data warehouse y un conjunto de datamart. Estos plantean ideas similares en temas de la integración de información, pero como propuestas distintas del manejo de datos.

En los 90's se popularizaron los CD-ROM, llegaron a los revolucionarios espacios en discos duros de 16 GB. Se inicia la investigación y comercialización de las tarjetas de memoria. Para la evolución de lo que hoy conocemos como Data Mining, fue en 1996 cuando en Gregory Piatetsky-Shapiro inició con descubrimiento de conocimiento a partir de las bases de datos y dió comienzo a la reestructura del Data Mining en la convención Knowledge Discovery in Databases de 1989. Para 1996, Usama Fayyad acuñaría la frase que definiría el camino de las aplicaciones de esta rama estadístico-computacional *"Es el proceso no trivial de identificación de patrones, útiles y novedosos implícitos en las bases de datos"*.

En los 2000, en terminos de almacenamiento se da una evolución desde los 36 gigabytes y termina con almacenamientos masivos de 1-4 terabytes. En el ámbito de los usuarios, se familiariza el uso de grandes repositorios de información de hasta 32 gigabytes, que en principios de la década estaban orientados a empresas y grandes consumidores.

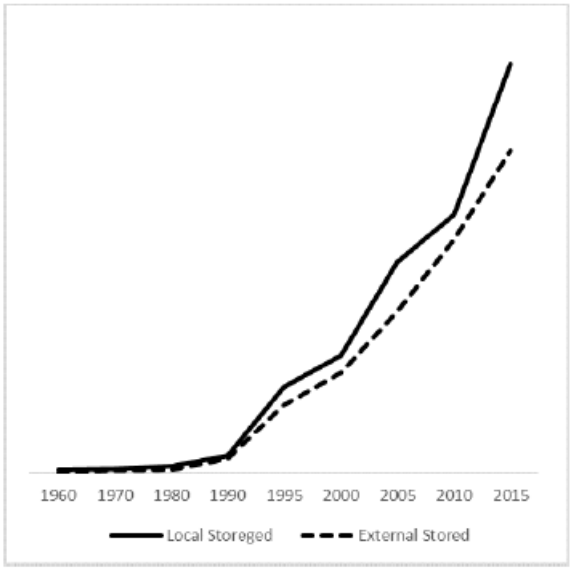


Figura 1.1.2: Crecimiento del almacenamiento de datos .Elaboración propia

En la figura 1.1.2, podemos observar que el avance exponencial en el crecimiento de almacenamiento de información ha sido vital para el manejo de grandes cantidades de información,

que a la postre sería la justificación y premisa de la minería de datos, ciencia de los datos y big data. Con lo cual a pesar de tener los algoritmos de análisis de información previamente, es adelante de los años 2000 que la explotación de estos repositorios de información son reales.

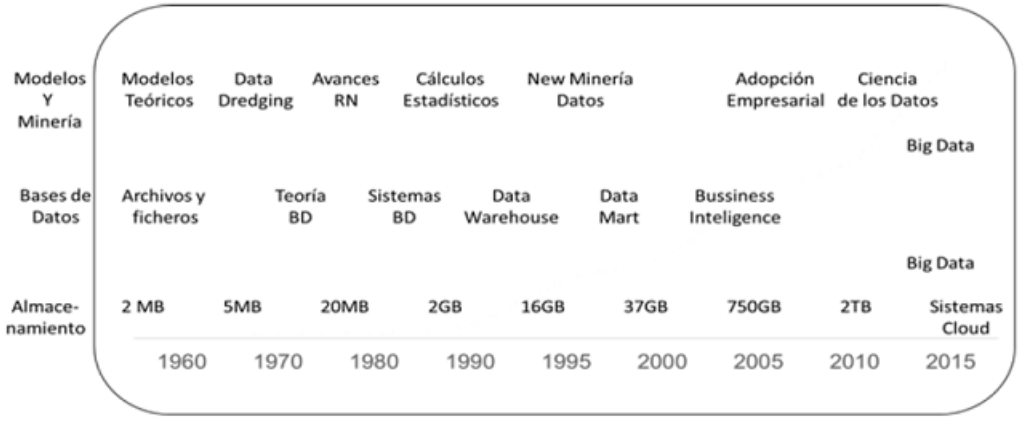


Figura 1.1.3: Evolución de la minería de datos: Modelos, Base de datos y Almacenamiento. Elaboración propia

En la figura 1.1.3 se resume la evolución de la minería de datos en tres componentes esenciales de la minería de datos, los cuales coadyuvaron a la evolución y las nuevas teorías y ramas de las aplicaciones en los últimos años. Bajo la explosión en la parte del almacenamiento de datos, el uso y explotación de la información y computadores que podrían hacer cálculos complejos la principal distancia entre el Data Dredging de Stuart y Selvin ya no existían, con lo cual se popularizan sistemas como SAS, Weka, SPSS y lenguajes estadísticos especializados como R y Python. Entre estos programas se generaron y adoptaron las metodologías de minería de datos.

1.2. Principales áreas de estudio

Los usos de la minería de datos más estudiados son el análisis de asociación, patrones de fuga, lavado de dinero, fraudes, comportamiento, banca, redes sociales, etc. [UIAF, 2014, Maimon y Rokach, 2010, Tuffery, 2008]

Análisis de asociación El análisis de asociación es útil para saber las preferencias de los consumidores y aplicar estrategias de ventas cruzadas útiles. Un ejemplo común del análisis de asociación es el análisis de la cesta de compra, en la cual, a partir de estudios de comportamiento de compra, los autoservicios pueden sugerir algunos productos asociados o codependientes de los primarios.

Patrones de fuga El análisis de patrones de fuga es usual en los negocios donde la competencia y las necesidades de un cliente, por lo que es importante saber, cuando un cliente está interesado de salirse del mercado y cuando pasara a manos de un competidor.

Patrones de lavado de dinero y fraudes En el lavado de dinero, la minería de datos es útil para distinguir las transacciones y movimientos que pueden ser parte de las herramientas delictivas para la suplantación de identidad, fraudes y lavado de dinero.

Patrones de comportamiento En los medios de telecomunicación con regreso de información tales como la telefonía celular e Internet; la minería de datos juega un papel importante a la hora de determinar perfiles de usuario, preferencias de usos, preferencias de transacciones, con ellos las empresas de telecomunicaciones pueden prepararse mejor para sus campañas publicitarias e infraestructura.

Banca Retail En la Banca retail y/o comercial, la minería de datos es fundamental en dos acciones tácticas del negocio; la aceptación de productos y la clasificación de riesgos de contraparte. Así como los patrones de fuga, es común que los clientes requieran hacer usos de los servicios y productos de la banca, a lo que hacer un proceso de aceptación de crédito reactivo, genera dudas acerca de los rechazos en que en ellos pueden ocurrir además de tiempos innecesarios, por lo que es común en el negocio bancario, que del cliente que ya se tiene información se hagan prospectos de crédito, previos a la solicitud del cliente. Y de forma análoga, a los clientes que ya tienen crédito evaluarlos para medir la probabilidad de impago.

Por ejemplo, en una cartera de un millón de clientes, detectar que el 10 % es susceptible de impago a corto plazo, te ahorra el 90 % de gestiones preventivas, que se traduce como una estrategia de ahorro, la cual es posible a partir de una segmentación y técnicas de minería de datos.

Redes sociales Con frecuencia se utiliza para la realización de análisis de redes sociales, el cual analiza las diferentes relaciones e interacciones entre los individuos, la fuerza de sus lazos y alto volumen de transaccionalidad e información, hacen de esta fuente de información un reto para los administradores de las plataformas.

1.3. Principales aspectos regulatorios de los datos

En términos del avance tecnológico, uno de los principales riesgos en México y el mundo, radica en uso inadecuado de la información de las personas para un lucro lícito o ilícito, por ello, diferentes países y específicamente en México se ha provisto de leyes que regulan y protegen el uso de los datos personales como un estado de seguridad para los ciudadanos.

En este aspecto en 2002 se crea la Ley Federal de Transparencia y Acceso a la Información Pública Gubernamental, con el cual se inaugura el Instituto Federal de Acceso a la Información (IFAI). Este organismo según su página de Internet tiene como misión *"Garantizar en el Estado mexicano los derechos de las personas a la información pública y a la protección de sus datos personales, así como promover una cultura de transparencia, rendición de cuentas y debido tratamiento de datos personales para el fortalecimiento de una sociedad incluyente y participativa."*, de este modo sus objetivos es cuidar los datos personales de los usuarios y aplicar la ley en caso de faltas. En 2007 se genera una reforma al artículo sexto constitucional con el cual se obliga a los estados a tener un mínimo piso de transparecía y acceso a la información [Guerrero, 2018]. Adicional se agregan los principios bajo los cuales se garantiza el derecho a la información. Con estos antecedentes en 2010 fue propicio enviar la propuesta de Ley General

de Transparencia y Acceso a la Información Pública.

La Ley Federal de Protección de Datos Personales (LFPDP) fue promulgada de acuerdo con el Diario oficial de la federación el día 5 de julio de 2010 con el siguiente calendario de actividades previas:

- Aprobado por diputados:15 de abril, 2010
- Aprobado por senadores: 27 de abril, 2010
- Publicado en Diario oficial de la Federación: 5 de julio, 2010
- Vigente a partir de: 6 de julio, 2010

1.4. Toma de decisiones con minería de datos

Dentro de los diferentes rubros donde la minería de datos se ha hecho de un lugar, es en la toma de decisiones ejecutivas y directivas, ya que permite encontrar soluciones o identificar problemas de manera más ágil y desglosada.

Figura 1.4.1: Cuadro de la evolución del conocimiento para la minería de datos Información:Garcia Reyes [2012]. Elaboración propia

En la figura 1.4.1 se muestra la relación, histórica y de valor del desarrollo del conocimiento. Con datos de fácil acceso y con información valiosa para el negocio, la forma de tomar decisiones

se da en función a una relación matemático-estadística, la cual no asegura que el análisis y resultados siempre sean los esperados, pero se tiene una mayor confianza que es incremental a parámetros como la profundidad del estudio, las pruebas realizadas y las herramientas que se utilizaron. El conocimiento de las evaluaciones y del negocio, así como sus relaciones entre los datos, crean mayor certeza a la hora de proponer cambios estratégicos. Sin embargo, terceras variables en el proceso serian el tiempo, el grado del cambio y la estructura. En el caso del tiempo dentro de los primeros paradigmas del conocimiento se encuentra la inversión del tiempo contra la agilidad y la oportunidad, ya que, en muchos casos, la inversión de un tiempo excesivo en alguna de las fases de la minería de datos puede generar que el ambiente de negocios o la regulación cambien drásticamente, esto genera tener ineficazmente un análisis o modelo con una certeza muy alta. En el caso del grado de cambio Roberto García [Garcia Reyes, 2012] propone en tres las formas de clasificación:

- Estratégicas: Modificaciones a procesos, áreas o políticas que tienen un impacto en el desarrollo y producto de una empresa.
- Tácticas: Modificaciones sobre áreas identificadas de la empresa, su impacto, aunque podría afectar a un porcentaje alto de la población, solo afecta a un área.
- Operativas: Cambios o elecciones de cambios o procesos que no involucran un impacto en el desarrollo global, aunque deben tomarse, lo cual genera un riesgo y una oportunidad.

En ellas, cada una requiere una complejidad distinta de habilidades y conocimientos.

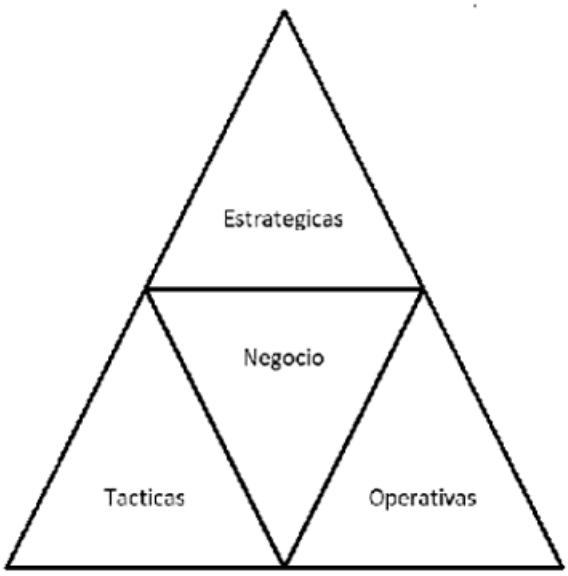


Figura 1.4.2: Conformación de las actividades del negocio. Fuente: Elaboración propia.

La figura 1.4.2, refleja las necesidades de acciones estratégicas, tácticas y operativas. Así un tomador de decisiones estratégicas como presidentes, directores generales, CEO y directores, con habilidades de negociación, innovación y análisis, requieren un panorama interno y externo de la situación de la empresa, un objetivo largoplacista claro y un plan para lograrlo. Su principal pregunta es el qué y cuándo hacer para lograr los objetivos con métricas financieras y no financieras que no permitan la ambigüedad al crear las condiciones del cómo es y cómo debe ser en un lenguaje no técnico, siempre pensando como cliente y para el cliente.

En épocas de crisis, generan cambios estructurales a la empresa, por ejemplo, el comité de dirección del banco A observa que, bajo estándares internacionales, el funcionamiento y delegación de sus cajeros se encuentra en administración de una empresa tercera, lo cual ha generado fricciones entre el personal interno, los clientes y algunos casos de fuga de información. Por lo tanto, el comité de dirección decide cambiar la infraestructura del banco para internalizar los cajeros, así como su gobierno corporativo. Los gastos inmediatos requieren inversión y nuevo talento humano, estos se mitigarán dentro de cinco años y la recuperación de la inversión en dos más. La decisión ha sido estratégica, por el periodo de retorno es probable que no se

encuentre el mismo comité de dirección. Las acciones para llegar al objetivo son variadas y afectan a varias áreas de la institución.

Las personas en las áreas tácticas están enfocadas en puestos como Gerentes, Lideres y Coordinadores, tienen avanzadas facultades de comunicación, proposición de procesos y distribución de funciones. Su conocimiento del negocio y funcionamiento es alto, así como su nivel de compromiso por las metas estratégicas y crean acciones para lograrlas, su principal cuestionamiento es el dónde y cuándo hacer las acciones. Las personas enlace entre la estrategia y la operación son quienes entienden las necesidades de ambas posiciones y tiene un conocimiento técnico avanzado. En situaciones de crisis, se buscan planes de mitigación para resolver problemas que las areas operativas no puede resolver.

En el mismo ejemplo del Banco A, la dirección de Riesgos observa que habrá una modificación en los insumos que se requieren para hacer reportes regulatorios, por lo que se tendrán que modificar procesos y realizar proyectos relevantes. Los gerentes realizan una reunión entre ellos para conversar los impactos técnicos de los cambios. Al finalizar la reunión agendan una cita con el comité de dirección para discutirlos, el cambio más relevante será un aumento en costos, pero el tiempo de retorno se observa más cercano. El comité de dirección a dado su visto bueno dentro del marco de actuación. Ahora es tarea de los gerentes en comunicárselo a las áreas de servicio y de infraestructura para su realización. Se observa que las decisiones de los gerentes están basadas en la comunicación y en los expertos técnicos, en ellos se sitúa la recomendación técnica. Es una decisión que afecta solo al área de riesgos y detalla el plan en acciones concretas.

La operación y las personas que la conforman son los personajes que ejecutan los planes y los logran como, por ejemplo: vendedores, ejecutivos, telefonistas, etc. independiente del marco de acciones de la empresa o giro. Tienen conocimientos avanzado del día a día y los procesos y herramientas del giro. Tienen en cada momento en que deben y en qué momento hacerlo,

tiene un amplio conocimiento en las oluciones rápidas y eficientes. Regularmente son el mayor porcentaje de los activos del banco. Por ejemplo, los ingenieros de los cajeros ahora requieren hacer modificaciones técnicas para el desarrollo de la data Warehouse interno del banco. Se requiere tomar decisiones técnicas de canales, bases de datos y componentes a instalar para el funcionamiento. Son decisiones que, en su rubro, son expertos y no requieren tomar en cuenta a funcionarios menos especializados. Para el cumplimiento de los planes estratégicos y tácticos, tiene un panel donde agregan el avance mensual y los impedimentos.

Cabe destacar que toda área requiere que los tres tipos de tomadores de decisión sean profesionales y expertos en sus temas. Sin un plan estratégico, los cambios estructurales no podrían ser posibles. Sin un decisor táctico, la idea estratégica hacia la operativización, tiene una brecha importante. Y sin un experto operativo, los planes tácticos no se cumplirán y los estratégicos se retrasarán. En este contexto, la minería de datos se centra en el conocimiento del negocio que no es tan sencillo de obtener y es parte del conocimiento de las decisiones tácticas y estratégicas.

1.5. Simulación en los portafolios de crédito

La simulación, como proceso de implantación de un modelo estadístico, busca realizar escenarios con el fin de mostrar de manera más fiable el aprendizaje y la evaluación de estrategias. Además, supone una mejora en costos, ya que se puede probar una hipótesis sin necesidad de realizar pruebas piloto. En diversos casos, la simulación es utilizada para resolver problemas de tiempos, costos, regulaciones o de solución determinista. Azofeifa [2004] propone la simulación como el desarrollo lógico-matemático que imita el sistema real.

Los principales ejemplos históricos del uso de la simulación podemos nombrar: La Perestroika, operativos militares en zonas desconocida, capacitación de vuelos y el Proyecto Monte

Carlo [Tarifa, 2018]. Este último fue desarrollado por John von Neumann, creador también del método de implosión que es precursor de la bomba atómica, quien con el uso primario de las computadoras utilizaba la simulación de números aleatorios en problemas que no podían ser resueltos de forma analítica. Este método de solución fue trabajado junto a Stanislao Ulam al cual nombraron método Monte Carlo [Rodriguez-Aragon, 2011]. Monte Carlo se basa en realizar registros aleatorios bajo una función de probabilidad [Grijalva, 2009]. Entre los problemas más usuales en la simulación en gestión de riesgo, podemos encontrar:

- Estimación de tasas de interés [Diez-Canedo et al., 2003]
- Valuación de derivados [Boyle, 1976]
- Pronóstico de acciones o bonos [Black et al., 1990, Black y Scholes, 1973, Olvera y Jimenez, 2013]
- Cálculo del Valor en riesgo [Tellez, 2010, Bolanos, 2010]
- Cálculo del Riesgo de crédito por CreditMetrics que utiliza matrices de transición [RiskMetrics, 2007] y Modelos Modernos [Altman, 1968, RiskMetrics, 2007]

El comportamiento de los clientes de un banco es una aplicación recurrente para los modelos de simulación, esto sucede dado que en la conducta de los clientes siempre existirá un elemento estocástico [Tarifa, 2018]. En el inicio de las regulaciones, se realizaron esfuerzos en Basilea I por tener un mínimo de requerimiento por crédito fijo para en proyección de las pérdidas esperadas por el incumplimiento de los clientes, sin embargo, esta medida no discriminaba las principales diferencias de las bancas privadas, patrimoniales, bancarias o de segmento de iniciación, por lo que el criterio lineal causaba muchas molestias a los bancos de segmentos altos e incertidumbres y bancarrotas para los segmentos bajos. Por ello las mediciones de riesgo usan la simulación estocástica para realizar estudios de CreditRisk y CreditMetrics realiza la evaluación del crédito en base Pérdidas Esperadas y Valor en Riesgo (VaR), sin embargo, no tiene en cuenta el riesgo sistemático de las probabilidades de incumplimiento y recuperación

[Luo y Shevchenko, 2013]. Con ello podemos observar que el VaR es una de las métricas más utilizadas en el mercado bancario.

Sin embargo, es recomendable que en la gestión de los portafolios se debe contar con herramientas para la toma de decisiones de otros tipos y no solo de carácter general, por ejemplo, en la evaluación de la gestión de riesgos es importante tener criterios de crecimiento o modificación de las reglas, estrategias y tácticas, en ese sentido las herramientas de evaluación según [Breeden y Ingram, 2003] se pueden separar en tres ramas.

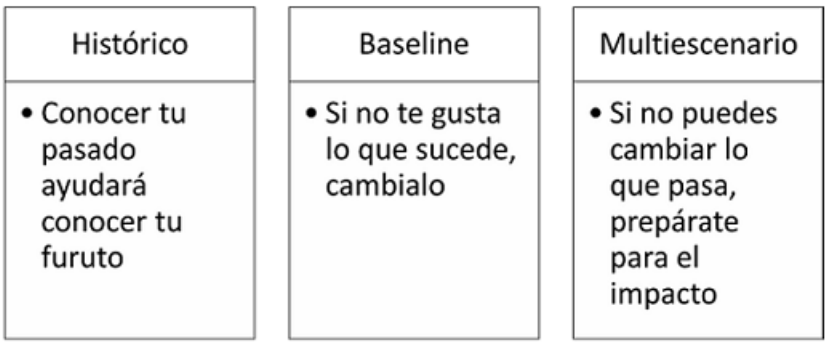


Figura 1.5.1: Evaluación y pronóstico de escenarios. Fuente: Elaboración propia.

Dentro de esta clasificación los multiescenarios requieren de técnicas avanzadas que ayuden al usuario a realizar cambios en un modelo complejo y tener múltiples vistas de las soluciones. En este aspecto el método Monte Carlo es ideal. Kreinin [2001] realiza una simulación en el mercado integrado y un portafolio de crédito, donde nos propone que los factores de riesgos de un portafolio se pueden describir bajo trasformaciones normalizantes, donde se requiere que se distribuyan de forma normal, con este resultado podemos generar muestras aleatorias por Método Monte Carlo que tenga una distribución normal. Y a partir de las muestras aleatorias aplicar las transformaciones inversas, llegar a simular el sistema.

Por su parte, Chadam et al. [2001] propone adicional que los componentes de riesgo se encuentran correlacionados, por lo que utiliza una matriz de correlaciones de los factores inicia-

les y con la descomposición de Cholesky, integra la correlación de los factores a la simulación Monte Carlo.

Para Altman [2004] la función de la pérdida esperada como la pérdida inesperada son subestimadas si se supone que los factores de riesgo de incumplimiento no están correlacionados. Por lo tanto, los modelos de crédito y las decisiones de negocio pueden dar lugar a reservas bancarias insuficientes y causar shocks innecesarios a los mercados financieros. En otros esfuerzos de la revisión de simulaciones Monte Carlo Breeden [2008] utiliza Modelos Arima para un portafolio de crédito.

La metodología de Kreinin ha tenido impactos en trabajos en Latinoamérica con Rodriguez y Trespalacios [2015] quienes con una metodología similar buscaron simular el valor en riesgo de un portafolio de crédito en Colombia. En el caso de fondos de pensiones, Chavez y Zanabria [2018] realizan un ejercicio de proyección de retornos a partir de la información histórica y las expectativas del mercado. Proponen utilizar el Análisis de Componentes Principales a las curvas de rendimientos en los mercados de renta fija.

La simulación Monte Carlo es utilizada en el sector bancario para realizar estimaciones del incumplimiento y sus posibles escenarios. [Ratings, 2018, Inbursa, 2017]. Además, Rosas y Benavides [2015] nos muestran en el artículo *"Stress-Testing para carteras de crédito del Sistema Bancario Mexicano"* como utilizá el método de Cholesky para el cálculo de pérdida esperada en escenarios adversos, con lo cual la correlación de las variables se integra a la prueba macroeconómica.

La búsqueda de escenarios más fiables y fáciles de resolver para predecir el futuro del sector bancario es una ardua tarea por descubrir e implementar. Este trabajo, buscará una solución a la problemática de la toma de decisiones en casos de modificación a la estrategia de portafolio,

con los antecedentes antes mostrados.

Capítulo 2

TÉCNICAS DE MINERÍA DE DATOS Y SIMULACIÓN

2.1. Estadística descriptiva

2.1.1. Medidas descriptivas

Medidas de tendencia central

Las medidas de tendencia central son útiles para describir el comportamiento de las variables y aproximarnos al posicionamiento de las variables dentro de su distribución. Sus principales funciones radican en sintetizar en medidas la mayor parte de la información para poder hacer conclusiones previas y saber si habrá algún análisis de mayor orden o complejidad que con los patrones o generalidades, y en pasos posteriores nos servirán para realizar modelos, aplicaciones o ajustes.

Media

La media se distingue por ser la medida de tendencia central más utilizada en términos prácticos. Indica en distribuciones normales, alrededor de qué valor se encuentran los datos. Tiene como definición poblacional:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \tag{2.1.1}$$

Rincón define a la media *"Se puede describir como el promedio ponderado de los diferentes valores que puede tomar la variable"* Rincon [2007].
Para la media muestral:

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2.1.2}$$

Que se interpreta como un estimador de la media poblacional. En ambos casos la sensibilidad de la media se ve afectada por los outliers de los datos.

Mediana

La mediana es definida como la medida central de los datos. Está definida por:

$$\bar{x} = \begin{cases} 1/2 \left[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right] & si\ es\ par \\ x_{(\frac{n}{2}+1)} & si\ es\ impar \end{cases} \tag{2.1.3}$$

La mediana es sumamente útil en el caso de outlier donde la media es afectada por los valores extremos. En la mediana existe el mismo número de registros en el lado superior e inferior.

Medidas de Posición Las medidas de posición se describen como los puntos en una variable donde describe algún comportamiento a partir de la distribución ordenada de los datos. El

ejemplo inicial ya definido es la mediana, pues es el número en la posición central intermedia de todos los datos.

Cuartiles

Los cuartiles describen la posición del 25 %, 50 % (mediana) y 75 %. El C1 (25 %) podría interpretarse de igual forma como la mediana de la distribución superior y el C3 (75 %) como la inferior.

Deciles y Percentiles

Los deciles y percentiles se describen de forma similar, en el caso de los deciles describen la posición de cada 10 % de la población y los percentiles el 1 %.

Medidas de Dispersión

Las medidas de dispersión miden la distancia entre las observaciones y que nos ayuda a saber que tan dispersa se encuentra una variable o base de datos. Definida como la diferencia de las variables respecto a la media de cada variable.

$$S_{jk} = \frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_i)(X_{ik} - \bar{X}_k)}{n} \tag{2.1.4}$$

Varianza poblacional

La varianza mide la posición central respecto a la media poblacional o muestral. Con ello tenemos una medida promedio de la distancia entre el punto medio y el punto de evaluación. En términos de la relación de dos variables es conocido como covarianzas poblacionales y se denota: La cual es el producto escalar de dos series de datos medidos contra la media de la serie. La varianza se da cuando se busca la relación de esta variable quedando.

$$S_{jk} = \frac{\sum_{i=1}^{n}(X_{ik} - \bar{X}_k)^2}{N} \tag{2.1.5}$$

Y tener como desviación estándar, la raíz de S_{ik} y como interpretación la distancia media de los datos en referencia con la media.

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_{ik} - \bar{X}_k)^2}{N}} \tag{2.1.6}$$

Varianza muestral

La varianza poblacional proporciona un estimador de la varianza poblacional.

$$s_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x_i})(x_{ik} - \bar{x_k})}{n-1} \tag{2.1.7}$$

La varianza se da cuando se busca la relación de esta variable quedando.

$$s_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x_k})(x_{ik} - \bar{x_k})}{n-1} \tag{2.1.8}$$

$$s_{jk} = \frac{\sum_{i=1}^{n}(x_{ik} - \bar{x_k})^2}{n-1} \tag{2.1.9}$$

Y tener como desviación estándar, la raíz de s_{ik} y como interpretación la distancia media de los datos en referencia con la media.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_{ik} - \bar{x_k})^2}{n-1}} \tag{2.1.10}$$

Amplitud Intercuartílica

La amplitud Intercuartílica, es la resta entre el cuartil tres y el cuartil uno. Esta medida no es sensible a outliers.

$$IQR = Q_3 - Q_1 \tag{2.1.11}$$

2.1.2. Análisis de correlación

El análisis de correlación fue acuñado y utilizado de manera rudimentaria para la astronomía por Augusto Bravais en 1846. Fue hasta 1889, que Francis Galton desarrolla la idea de correlación de dos variables. Y finalmente Karl Pearson obtiene el coeficiente de correlación [Pena, 2002].

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{2.1.12}$$

Donde

$$-1 \leq \rho_{ij} \leq 1 \tag{2.1.13}$$

- Para variables estandarizadas, de media cero y desviación típica unidad, la covarianza es el coeficiente de correlación.
- Si $x_{ij} = a + b_{ij}$entonces $|\rho_{ij}| = 1$
- El coeficiente de correlación mide el grado de asociación de dos o más variables.
- Si dos variables son ortogonales, es decir los vectores que las caracterizan forman un ángulo de 90 grados, llamando ρ al coeficiente de correlación como $\rho = \cos\theta = 0$, las variables están no correladas.

Es importante tomar en cuenta que en la interpretación que el coeficiente de correlación, solo mide asociación lineal y no causalidad [Nieto Barajas, 2018].

2.1.3. Análisis gráfico

El análisis gráfico es un recurso utilizado para describir las características de una población o reflejar un análisis de forma sencilla y básica. Con los gráficos y diagramas es común detectar comportamientos atípicos y realizar un análisis descriptivo.

Histograma

El histograma es una representación gráfica de una variable donde se relacionan los valores con la frecuencia de clases. Nos ayuda en tener una representación visual de la distribución de los datos.

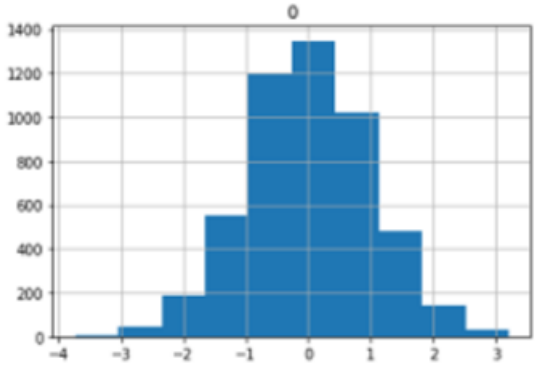


Figura 2.1.1: Gráfico ejemplo de Histograma. Fuente: Elaboración propia.

En el gráfico anterior, podemos observar un ejemplo clasico de un histograma con función de distribución Normal.

Diagrama de dispersión

El diagrama de dispersión es una representación bidimensional que nos muestra la dispersión de las variables. En una matriz de dispersión en la representación de la diagonal se muestra un histograma de las variables.

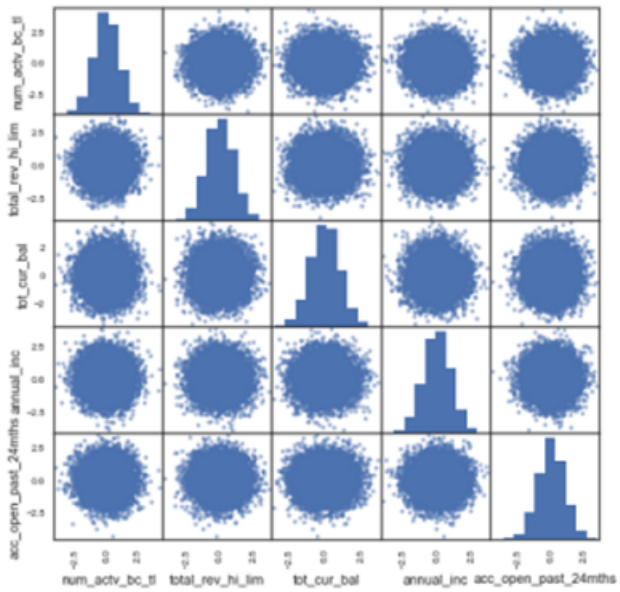


Figura 2.1.2: Gráfico ejemplo de gráfico de dispersión. Fuente: Elaboración propia.

El ultimo gráfico, podemos ver los diagramas de dispresión de varias variables, este gráfico nos apoya en saber el comportamiento bivariado de un conjuno de variables, ademas en la diagronal izquierda-derecha nos muestra en forma de hitograma las distribuciones de las variables.

Diagrama de caja-bigotes

El diagrama de caja-bigotes en una representación gráfica multivariada que nos muestra en una sola imagen la media, mediana, los cuartiles y las dispersiones, así como los principales outlier.

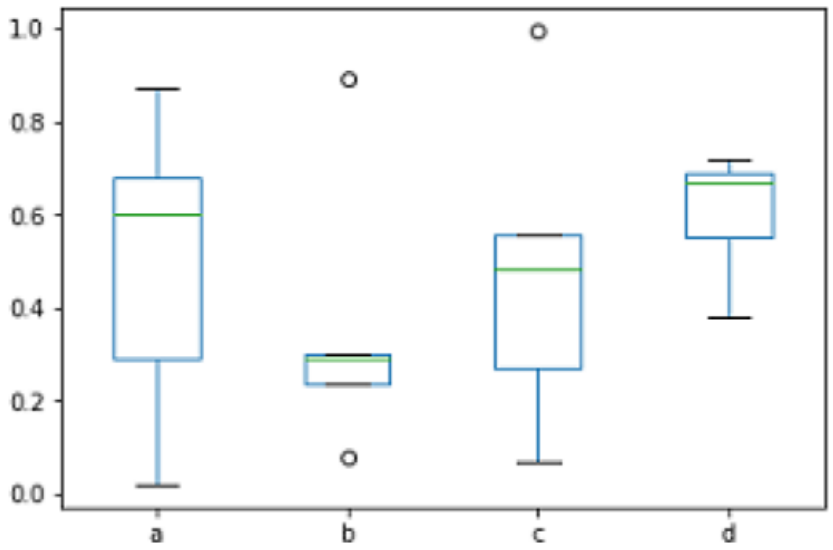


Figura 2.1.3: Gráfico ejemplo de Caja-bigotes. Fuente: Elaboración propia.

En la gráfica Caja-bijotes, podemos saber la distribuciónd de una variable por distintos segmentos, y saber las diferencias entre los grupos.

2.2. Pruebas de igualdad de función de distribución

Para casos donde queremos saber si las características de una variable aleatoria cumplen con alguna distribución o parámetro, es usual utilizar las pruebas de hipótesis, las cuales tienen como objetivo estudiar la información de las muestras y aceptar o rechazar una hipótesis. Las pruebas normalmente constan de cuatro elementos:

- Hipótesis nula, es la aseveración que se cumple hasta que exista evidencia suficiente de lo contrario.
- Hipótesis alternativa, resultado del rechazo de H_0
- Estadística de prueba, la decisión de rechazar se basa en la información en la prueba de la muestra.
- Región de rechazo, valor en el que se encuentra en un intervalo el cual rechaza H_0.

Las pruebas de igualdad en una función de distribución son herramientas para verificar si dos muestras tienen la misma distribución de probabilidad. Estas se clasifican en dos grupos:

- Pruebas paramétricas, son las derivadas de asumir que una muestra de datos tiene una distribución de probabilidad, esto basado en parámetros. Ejemplos:
  - t de Student, comparación de medias y equivalencias de muestras
  - F de Fisher, comparación de varianzas
  - X^2de Fisher, comparación proporciones, asociación entre variables cualitativas
- Pruebas no paramétricas, están basadas en una distribución libre o con parámetros no específicos.
  - Prueba Kolmogorov-Smirnoff
  - Prueba Anderson-Darling
  - Prueba τ Kendall

Prueba Kolmogorov-Smirnov

La prueba Kolmogorov-Sminov tiene como objetivo comparar si dos distribuciones son similares. Herrera la define como la comparación de la distribución acumulativa con la función de distribución empírica [Herrera Maldonado, 2008]. Para ello debemos garantizar que se cumple con:

$$\mathbb{P}\left[\lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = 0\right] = 1 \tag{2.2.1}$$

Con ello se garantiza que la función de distribución en prueba tiene a tener distribución teórica. Donde el estadístico de contraste es:

$$D = \sup_{-\infty<x<\infty} |F_n(x) - F(x)| \tag{2.2.2}$$

- x_i es el i-ésimo valor observado en la muestra
- $F_n(x)$ es un estimador de la probabilidad de observar valores menores o iguales que xi.
- $F(x)$ es la probabilidad de observar valores menores o iguales que xi cuando H_0 es cierta.

Se rechaza la hipotiposis nula para valores grande si,

$$D = \sqrt{n}D_n = K \tag{2.2.3}$$

Donde K se obtiene por medio de cuantiles de la distribución D_n. [UIAF, 2014] El estadístico busca encontrar las diferencias la distribución, media, dispersión y simetría [Nieto Murillo, 2010].

Prueba Anderson-Darling

Esta prueba asume que no se tienen parámetros a estimar, en cuyo caso la distribución de los datos es libre. Se define como

$$A^2 = -n - \sum_{i=1}^{n} \frac{2k-1}{N}[lnF(X_k) + ln(1 - F(X_{N+1-k})] \tag{2.2.4}$$

El estadístico de prueba se puede comparar con los valores críticos de la distribución teórica [Marques, 2018].

2.3. Estadística multivariada

2.3.1. Análisis de componentes principales

La dificultad y manejo de grandes cúmulos de información, en ocasiones son resueltos con software que manejan tamaños de procesamiento mayores, sin embargo, se tienen herramientas estadísticas que coadyuvan a encontrar formas de reducir el número de variables a utilizar. Con

ello se pueden realizar cálculos y representaciones con bajos niveles de pérdida de información.

Los componentes principales tienen una larga historia dentro de la estadística, ya que es para muchos autores se encuentra dentro de las técnicas multivariada más antiguas, pues se puede remontar el descubrimiento de la descomposición en valores singulares (SVD) en el siglo XVII que antecede al Análisis de Componentes Principales. Posteriormente Pearson en 1901 aportaría a la técnica con el ajuste de los vectores ortogonales por mínimos cuadrados, donde se establece que una de las principales limitaciones es el cálculo de la técnica, aunque demuestra la factibilidad del desarrollo. Con diferencia de 20 años, Fisher establece las bases del uso de SVD en análisis bidireccional. Fue en 1933 que Hotelling en el estudio *"Analysis of a Complex of Statistical Variables with Principal Components"* que se establecen las reglas básicas, así como el nombramiento de estas mismas. Para Hotelling, el punto central de investigación fue la reducción de variables independientes que determinan los valores de un conjunto más grande [Jolliffe, 2002].

Para Daniel Peña, el objetivo principal de la técnica es la reducción de un conjunto de datos en forma de columnas [Pena, 2002] que expliquen el comportamiento global de la base de estudio. Con lo cual se puede representar la varianza de un conjunto de variables, al tener una pérdida de información mínima. Los principales usos que establece del Análisis de Componentes Principales son la representación, por un lado, es decir, mostrar en un espacio de dimensión reducido la variabilidad de la dimensión general y por otro permite tener las variables originales correlacionadas en variables no correlacionadas con una matriz correlación, con la cual podemos identificar las posibles relaciones lineales entre las variables. La técnica tiene tres enfoques principales;

- Descriptivo, nos garantiza en caso de tener altas correlaciones, al menos n variables, poder representar la varianza estimada.
- Estadístico, se puede obtener campos genericas que describan el comportamiento general

de los datos, pues los primeros componentes principales tienen máxima correlación con todas las variables.

- Geométrico, en la representación gráfica de las variables, por cada variable tendremos un elipsoide, sin embargo, en el resultado de componentes principales, podemos tener un único elipsoide con dos variables, será la mejor aproximación a todas ellas.

En términos estadísticos, si tenemos el proceso estocástico que tiene i variables aleatorias, de media μ y varianza σ , lo podemos representar como un vector de puntos x_i y dirección a_i tal que,

$$Z_i = a_{11}x_{i1} + a_{12}x_{i2} \ + ... + a_{1n}x_{in} \ = \sum_{j=1}^{n} a_1 x_j \tag{2.3.1}$$

Y definiremos a r_i como la distancia x_i y su proyección a a_1.

$$\text{mín} \sum_{j=1}^{n} r^2 = \sum_{j=1}^{n} |x_j - z a_1|^2 \tag{2.3.2}$$

Y con ello,

$$x_2' x_i = z_i^2 + r_i^2 \tag{2.3.3}$$

Donde minimizar r_i^2 será equivalente a maximizar z_i^2, por lo que, al ser z_i^2 variables de media cero es posible decir

$$a_j \acute{} a_j = \sum_{j=1}^{n} a^2 = 1 \tag{2.3.4}$$

Con el primer componente,

$$y_1 = Z a_1 \tag{2.3.5}$$

con varianza

$$var(y_1) = var(Za_1) = a_1\sigma \tag{2.3.6}$$

para maximizar $Var(y_1)$ se tendrá sujeta la restricción 2.3.2, y la incluimos a partir de método de multiplicadores de Lagrange,

$$L(a_1) = a_1\sigma_1 - \lambda(a_1'a1 - 1) \tag{2.3.7}$$

con solución

$$\sigma_1 = \lambda a_1 \tag{2.3.8}$$

donde a_1 es vector propio σ y λ como el valor propio.

En general, el proceso estocástico Z_i se sabe que se tienen tantos componentes principales i, donde S es la matriz de varianza-covarianza y se obtienen los valores mediante

$$|S - \lambda a_1| = 0 \tag{2.3.9}$$

y vectores asociados

$$(S - \lambda I)a_1 = 0 \tag{2.3.10}$$

Se llama Y a la matriz relacionada con Z por

$$Y = ZA \tag{2.3.11}$$

donde $A'A = I$, con conclusión geométrica que al calcular los componentes principales es encontrar los vectores ortogonales del proceso estocástico.

Derivado a este procedimiento con la maximización de varianza en cada vector ortogonal, como consecuencia estadística tenemos que no se encuentren correlacionadas, los componentes tienen máxima varianza ascendente ordenado al procedimiento del cálculo, siendo que

$$Var(Y) = Var(y_1) + Var(y_2) + \ldots + Var(y_i) = Var(Z) \tag{2.3.12}$$

$$Var(y_1) > Var(y_2) > \ldots > Var(y_i) \tag{2.3.13}$$

Si la correlación entre las variables iniciales es alta, se tienen condiciones para concluir:

- El primer componente principal es un promedio ponderado de las variables
- El primer componente además en una representación gráfica aglomera la mayor parte del tamaño de la varianza.
- El resto de componente son conocidos como componentes de forma.

Con este proceso, podemos tener un resultado de variables ortogonales que ordenan la varianza entre ellas y que mantiene la estructura de covarianza [Rigollet, 2016].

2.3.2. Regresión logística

Para las técnicas estadísticas clásicas, los modelos lineales generalizados son una importante herramienta en el tratamiento de datos para la predicción. El termino regresión fue acuñado por Francis Galton en 1886, sin embargo, Adrien-Marie Legendre y Carl Friedrich Gauss publicaron trabajo con la idea general de la regresión. El termino de Modelos Lineales Generalizados (GML en inglés) fue desarrollado por John Nelder y Robert Wedderburn en la búsqueda de la unificación de los modelos: lineal, logístico y Poisson.

$$Y_i \sim N(\mu, \sigma^2) \tag{2.3.14}$$

$$\mathbb{E}(Y_i) = \mu(X) = X^T\beta \tag{2.3.15}$$

Donde la representación general es

$$g(\mathbb{E}(Y_i)) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \tag{2.3.16}$$

En el caso particular de la regresión logística, la cual es una función sigmoidal, al igual que el resto de los GML se basa de la ecuación 2.34. La definición de la regresión logística bajo una Y_i es:

$$\log\left(\frac{\mathbb{P}(y_i = 1|x)}{1 - \mathbb{P}(y_i = 1|x)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \tag{2.3.17}$$

Y en suposición de

$$Y_i \sim B(p_i) \tag{2.3.18}$$

Por lo que

$$\frac{p_i}{1 - p_i} = e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}} \tag{2.3.19}$$

para encontrar el valor de p

$$p_i = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}}} \tag{2.3.20}$$

Los usos ideales de la regresión logística son los casos donde lo más importante son la interpretación o pronóstico de los datos categóricos o binarios, ganancia/perdida, tomar o no decisiones. [Nylen y Wallisch, 2017, Maimon y Rokach, 2010, Rigollet, 2016]

2.4. Simulación

La simulación es una herramienta estadística que estudia y ocupa la generación de datos que duplican o se acercan a los comportamientos de un problema. Es recurrente el uso de la simulación en casos donde existe una parte de aleatoriedad o elemento no determinista, con lo que buscaremos aproximaciones a partir de supuestos. Dentro de numerosos sectores y áreas de investigación la simulación ha sido útil para el desarrollo y resolución de problemas donde la complejidad y costo es mitigado con estos resultados [Rodriguez-Aragon, 2011]. Para el uso

de simulación, es necesario tener conocimientos de probabilidad, estadística y programación matemática [Lopez Briega, 2017].

2.4.1. Simulación estocástica

La simulación estocástica comúnmente conocida como simulación por método Monte Carlo, está basado en la generación de números aleatorios por el método de transformación inversa, interpretando el volumen de simulación como una probabilidad. Es de mayor utilidad en casos donde la parte aleatoria, proporciona más información que el resto del problema. John Von Neumann desarrolló la teoría y las primeras prácticas de ello durante los estudios de difusión de neutrones.

Para Grijalva [2009], el proceso de construcción de la prueba más sencilla:

- Determinar las distribuciones acumuladas de interés

Sea X_i una variable aleatoria con una distribución $F(X)$ donde

$$F(x_1 \dots x_n) = \mathbb{P}(X_1 = x_1 \dots x_n = x_n) \tag{2.4.1}$$

Los valores de X_i siguen una distribución dada por

$$\gamma_i = \int f(x)\, dx \tag{2.4.2}$$

Siendo γ_i un número aleatorio de una distribución $U(0,1)$ y con ello,

$$x = F^{-1}(\gamma_i) \tag{2.4.3}$$

- Iterar a partir de algoritmos de números aleatorios en una $U(0,1)$.

$$y \sim x = F^{-1}(\gamma_i) \tag{2.4.4}$$

- Calcular los estadísticos descriptivos y realizar un histograma comparar

$$E\left(Y_k\right) - E\left(X_i\right) \sim 0 \tag{2.4.5}$$

entonces

$$X_i \sim Y_k \tag{2.4.6}$$

- Analizar los resultados para distintos tamaños de muestra.

El teorema de los grandes números y del límite central nos asegura

$$E(Yk) - E(Xi) \geq E(Y_{k+1}) - E(X_i) \sim 0 \tag{2.4.7}$$

Es de vital importancia tener generadores de números psudo aleatorios adecuados, ya que la generación y selección aleatoria nos asegura que no existirá correlación entre las variables y tener reglas de límites para los valores de los resultados. En los principales lenguajes de programación y paquetería estadística y simulación este tema se encuentra resuelto (R, Python, SAS, etc).

Si bien la computadora genera un espacio de volumen extenso de números, es la selección aleatoria la que realiza la tarea del ajuste con la probabilidad. A partir de la ley de los grandes números, se asegura que la simulación será más exacta con el aumento de volumen en las muestras [Cruz Arrela, 2010].

En caso el partícular, cuando una variable aleatoria tiene una distribución normal estandar $Normal(0,\ 1)$ puede esta descrito por la distribución $U(0,\ 1)$:

Sea $F(X)$ la función de distribución acumulada una X_i la cual es una variable aleatoria,

$$F_x(x) = \frac{1}{2}\left(1 + erf\left(\frac{y}{\sqrt{2}}\right)\right) \tag{2.4.8}$$

Donde

$$erf\,(x) = \frac{1}{\sqrt{\pi}} \int_0^x dz e^{-z^2} \tag{2.4.9}$$

Si tenemos X_1 y X_2 donde ambas se distribuyen como una $U(0,1)$

$$f(y_1 y_2) = f(y_1) f(y_2) = \frac{1}{2\pi} \exp\left(-\frac{y_1^2 + y_2^2}{2}\right) y21 + y2 \tag{2.4.10}$$

y se cambia y_1 y y_2 por un escalar y un valor en grados para dar dirección y fuerza a un vector

$$y_1 = R\cos\Phi \qquad y_2 = R\sin\Phi \tag{2.4.11}$$

entonces

$$f(y_1) f(y_2) dy_1 dy_2 = r \exp\left(-\frac{r^2}{2} dr\right)\left(\frac{1}{2} d\phi\right) \tag{2.4.12}$$

Por lo que

$$R = \sqrt{-2\ln\xi_1} \quad \Phi = 2\ln\xi_2 \tag{2.4.13}$$

Al realizar un muestreo a $X \sim N(0,1)$ y teniendo a ξ_1 y ξ_2

$$X = \sqrt{-2\ln\xi_1}\cos(2\pi\xi_2) \tag{2.4.14}$$

$$Y = \sqrt{-2\ln\xi_1}\sin(2\pi\xi_2) \tag{2.4.15}$$

Para realizar una demostración de la teoría, se revisará un ejemplo de simulación con Python, donde teniendo como referencia la ecuación 2.4.14 y 2.4.15 se realizarán las simulaciones de x y y que son variables aleatorias con distribución $U(0,1)$ con distintos ejemplos de número de repeticiones de 10, 100, 1,000 y 10,000 simulaciones. Se ha dejado en el anexo cuatro el

programa para generarlo.

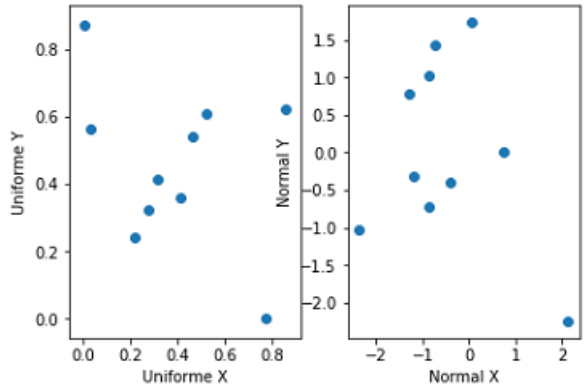


Figura 2.4.1: Simulación Monte Carlo de $U(0,\ 1)$ y $N(0,\ 1)$ con 10 repeticiones. Elaboración propia.

En la figura 2.4.1 se observa una primera simulación aleatoria bajo $U(0, 1)$ y $N(0, 1)$ con tan solo 10 repeticiones. Visualmente no se puede constatar alguna relación bajo sus distribuciones, derivado al número limitado de observaciones.

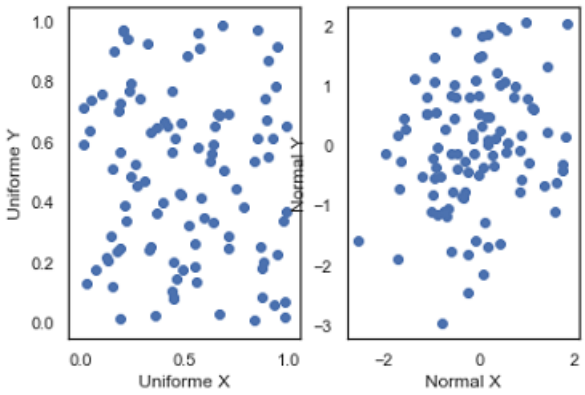


Figura 2.4.2: Simulación Monte Carlo de $U(0,\ 1)$ y $N(0,\ 1)$ con 100 repeticiones. Elaboración propia.

En la figura 2.4.2 se observa una segunda simulación aleatoria con 100 repeticiones. Se empieza a ver como el ejercicio de $N(0,\ 1)$ inicia a delimitarse en su figura.

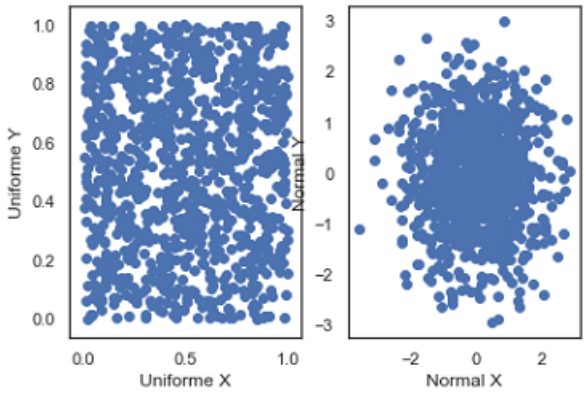


Figura 2.4.3: Simulación Monte Carlo de $U(0,1)$ y $N(0,1)$ con 1,000 repeticiones. Elaboración propia.

En la figura 2.4.3 se observa la tercera simulación con 1,000 repeticiones. Las figuras uniformes y redondas formadas por las distribuciones se conforman en su totalidad.

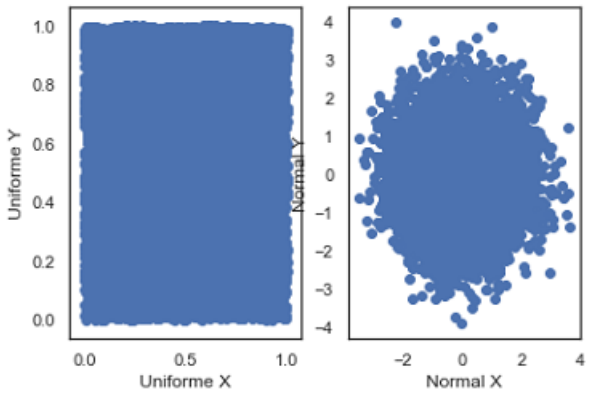


Figura 2.4.4: Simulación Monte Carlo de $U(0,1)$ y $N(0,1)$ con 10,000 repeticiones. Elaboración propia.

En la figura 2.4.4 se concluye que con la simulación con 10,000 repeticiones ha rellenado a la estructura uniforme perimetrada por el ejercicio anterior. Mismo caso ocurre con las $N(0,1)$.

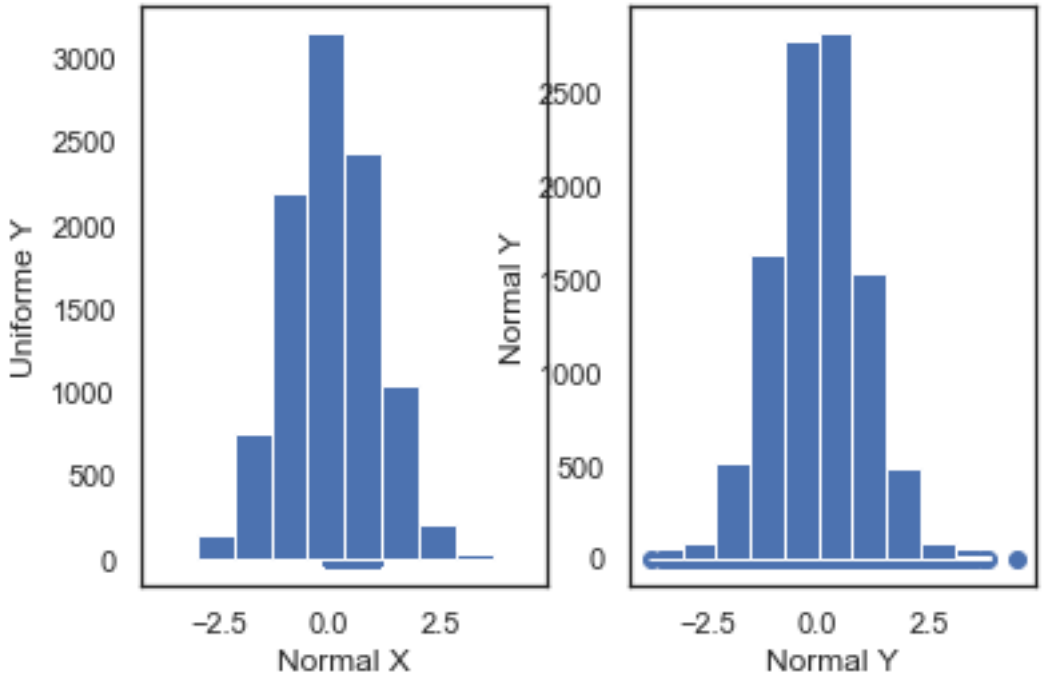


Figura 2.4.5: Histograma de la variable X y Y en 10,000 reperticiones. Elaboración propia.

La figura 2.4.5 nos muestra la generación de las variables normales generadas a partir de variables aleatorias uniformes. Podemos corroborar que en ambos casos se distribuyen de manera normal, bajo un análisis visual.

2.4.2. Simulación de variables correlacionadas

La simulación de variables aleatorias correlacionadas son un campo de la simulación Monte Carlo que se utiliza para simular modelos más complejos de estudio. Mike Giles nos menciona tres formas de obtener simulaciones correladas; por factorización de Cholesky, componentes principales y puente Brauniano [Giles, 2018], además en la tesis de simulación del cálculo de VaR de Emmanuel Malagon, integra la solución por copulas para encontrar la composición de factores [Bolanos, 2010].

Figura 2.4.6: Proceso de generación de escenarios de Loretan. Fuente: Elaboración propia.

Simulación por componentes principales

La quasi-Simulación por componentes principales tiene su primer acercamiento con el artículo *"Generating market risk scenarios using principal components analysis: methodological and practical considerations"* de Mico Loretan en 1997, el cual a partir de componentes principales busca la reducción de variables al encontrar los factores de riesgo, y con ellas buscar la prueba de estrés que se requiere [Loretan, 1997].

Loretan, en su artículo propone una solución para los riesgos de mercado, en el cual describe a ACP (Análisis de Componentes Principales) como una forma de determinar la estructura de los datos. Si la base de datos tiene una alta correlación, con uno o dos componentes es suficiente para determinar la mayor cantidad de varianza, así en consecuencia una baja correlación se necesitarán más variables para encontrar el número de dimensiones efectivas para el análisis. Estas dimensiones son caracterizadas por los vectores propios a los cuales llama *"meta-dimensiones"*. Con ello se planeta que al tener la metodología paso a paso hasta llegar a los CP (Componentes Principales), posteriormente al tener como supuesto que los componentes principales son variables aleatorias, se toma la función de distribución de cada componente, simula registros a partir de esta, se pueden realizar transformaciones inversas para encontrar los valores origen.

Alexander Kreinin, Leonid Merkoulovitch, Dan Rosen y Michael Zerbs en 1998, lanzaron un artículo nombrado *"Principal Component Analysis in Quasi Monte Carlo Simulation"* [Kreinin et al., 1998], en el muestran que la simulación Monte Carlo es una herramienta eficiente para realizar simulaciones de portafolios de riesgo con la reducción del problema a partir de componentes principales. El uso de componentes principales se usa por dos razones, en principio para

la reducción de espacios dimensionales, pero también las características de los componentes al cumplir la condición de ortonormalidad es inocuo para realizar las transformaciones inversas que nos lleve a la función de distribución acumulativa. Estas transformaciones se usarán para crear variables aleatorias y con la matriz *"meta-dimensiones"* realizar una ejecución de similaridad.
El fundamento central es, un grupo de variables ortonormales que describan el comportamiento del problema, por otro lado, realizar la simulación Monte Carlo independientes que con serán ajustadas con la matriz de covarianza de componentes principales y transformaciones lineales de lo obtenido en ACP, la cual a su vez ha generado variables ortogonales, y con realizar un ajuste a la dispersión y posición. Los pasos para generar escenarios son:

- Calcular la matriz de covarianza de los componentes principales
- Definir la dimensión k que se aceptará con un margen de error considerado
- Generar una muestra aleatoria k dimensiones basado en psudo números aleatorios.
- Transformar la muestra a una distribución uniforme de k dimensiones con una distribución de normal multivariada, con la se calcula la aproximación polinomial de la función inversa de componentes principales.
- Aplicar las transformaciones lineales previas a la generación de componentes principales.

Con este método, Kreinin demuestra que es un método eficiente para realizar una simulación. Esta metodología ha sido utilizada por otros investigadores como Bohdalova y Gregus [2010], Lai et al. [2005], Baez-Revueltas y Gamboa-Hirales [2013]. Van Son Lai, utiliza esta técnica para realizar un cálculo del VaR con la cartera de renta fija canadiense. En ella especifica los pasos a realizar, sea X_i un proceso estocástico

- Se obtienen la Q matriz de covarianza y los K factores de componentes principales

$$Q = AX_i \tag{2.4.16}$$

- Se realizan las K simulación aleatorias Monte Carlo

$$A_i = \sum_{i=1}^{n} \lambda_i \sim N(0,1) \sim Y_k = \sum_{i=1}^{n} y_i \tag{2.4.17}$$

- Realizar las transformaciones de los números aleatorios con el vector k dimensional de componentes principales

$$X_i' = Y_i Q^T \tag{2.4.18}$$

- Realizar los cálculos necesarios para la obtención del VaR

$$VaR = T(X_i') \tag{2.4.19}$$

Con ello, se encuentra una forma útil y rápida para obtener una simulación

Descomposición de Cholesky

La descomposición de Cholesky es un método usual en la solución de sistemas lineales. Se basa en los supuestos de tener una matriz Hermitiana definida positiva.

$$A = A^T \tag{2.4.20}$$

$$A = LL^T \tag{2.4.21}$$

En el caso de las simulaciones Monte Carlo se utiliza para descomponer la matriz de correlación, la cual puede expresarse como:

$$l_i = \sqrt{1 - \sum_{k=1}^{i-1} l_{ik}^2} \tag{2.4.22}$$

Con ellos se calcula el riesgo de los activos o factores de análisis, al simular la volatilidad de la cartera. Y con ello al integrarla a una simulación no correlada, lo convertirá en números aleatorio correlados.

La metodología para Bolanos [2010], en términos de simulación Monte Carlo por descomposición de Cholesky es:

- Generar n columnas de k números aleatorios uniformemente distribuidos.
- Aplicar la metodología Box-Muller a las variables generadas en el primer paso, para normalizarlas.
- Calcular las transformaciones lineales para normalizar las variables.
- Calcular la matriz de covarianzas y varianzas de las variables $N(0,\ 1)$.
- Emplear la factorización de Cholesky a la matriz de covarianzas y varianzas y multiplicar la matriz de variables normalizadas por la factorización de Cholesky.
- Transformaciones lineales para retornar a los valores iniciales

En el caso más sencillo donde tenemos

$$x_1, y_1 \sim N(0, 1) \tag{2.4.23}$$

donde existe el coeficiente de correlación

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} \tag{2.4.24}$$

y bajo Box-Muller tenemos

$$x_1', y_1' \sim N(0, 1) \tag{2.4.25}$$

Suponemos que $x_1 = x_2'$ entonces

$$y_2'' = \rho y_2 + \sqrt{1 - p^2 y_2} \tag{2.4.26}$$

Método Copulas

Las copulas, nombradas y postuladas por Abe Sklar, son una técnica de modelización de la distribución conjunta a través de una función de probabilidad multivariada conjunta, con lo

que se puede analizar la estructura de dependencia [Bolanos, 2010]. En el caso de tener los procesos estocásticos X_i y Y_i y tiene una función de distribución conjunta $H(x,\, y)$, entonces las funciones de distribución marginal de X, Y están dadas por $F(x)$ y $F(y)$ [Erdely, 2009].

Sklar establece

$$H(x,\, y) \,=\, C_{XY}(F(x),\, G(y)) \tag{2.4.27}$$

Para la simulación Monte Carlo por copulas [Bolance et al., 2015] propone la siguiente estrategia para una copula bidimensional:

- Calcular las pseudo-observaciones tal y como se describía en la expresión

$$U_{i1} \,=\, \frac{N}{N+1}\sum^{N} I(L_{j1} < L_{i1}) \tag{2.4.28}$$

y

$$U_{i2} \,=\, \frac{N}{N+1}\sum^{N} I(L_{j2} < L_{i2}) \tag{2.4.29}$$

- Estimar el parámetro de la cópula por el método de momentos
- Simular a partir de la cópula los valores, donde es el número de réplicas simuladas dependerá del proceso computacional.
- Obtener los resultados a partir de funciones inversas.

Se tienen ajustes que se pueden tomar como mejora para los resultados del comportamiento conjunto y marginal.

Elección de un modelo de simulación

La toma de decisiones sobre los modelos o técnicas a ocupar pueden resultar igual de complejas que la toma de decisiones sobre los resultados. En este documento nos guiaremos de las recomendaciones técnicas y metodológicas observadas durante el proceso de investigación.

Método	Recomendación Experta	Explicación	Implantación
PCA	Si	Si	Si
Cholesky	Inestable	No	Si
Cópulas	No Normal	Si	Si

Cuadro 2.1: Cuadro de decisión del modelo de simulación escogido. Fuente: Elaboración propia.

La primera opción eliminada ha sido el modelo por copulas, ya que para Bolance et al. [2015], que nos menciona que en el caso de simulación de distribuciones Normales, es más adecuado utilizar la descomposición de Cholesky, donde además de ser una técnica menos conocida en el medio financiero, existe mayor complejidad de los mineros de datos para poder ser explicada y aplicada en los análisis que se entregan a los tomadores de decisiones. En el caso de la descomposición de Cholesky, bajo el supuesto de matriz positiva definida, es menos efectiva ya que no pueden existir valores totalmente correlacionados, además de que existen casos de inestabilidad y problemas sin solución. Análisis de Componentes Principales, es una técnica con alta difusión y con una solución más global que la descomposición de Cholesky [Giles, 2018].

Capítulo 3

IMPLEMENTACIÓN DE LA MINERÍA DE DATOS EN UN PORTAFOLIO DE CRÉDITO

Dentro la implantación de las metodologías es importante tener en cuanta tres aspectos fundamentales, el problema, el camino y como lo caminaras, en referente con ello, a lo largo de este capítulo determinaremos los métodos y procesos existentes para el manejo de datos, que en este caso corresponde al camino, así como los programas que permiten el tránsito de la información y la técnica más idónea para la obtención del resultado deseado.

3.1. Metodologías de Minería de datos

En la conferencia KDD (Knowledge Discovery in Databases) que tuvo como principales expositores a Usama Fayyad,Gregory Piatetsky-Shapiro y Padhraic Smyth, fue dado conocer el concepto de Minería de Datos y se expusieron los pasos fundamentales para el descubrimiento en los datos [Nigro et al., 2018]. Fayyad como pionero de las metodologías de minería de datos postuló KDD, tomó en cuanta el contexto del momento y las necesidades que como investigador tenía. Posteriormente SAS Institute desarrolló SEMMA, metodología pensada y

generada del diseño de la interfaz de minería de datos que aplicaría en su sistema [Azevedo y Santos, 2008]. Posteriormente por parte de una iniciativa de las empresas Daimler-Benz, Integral Solutions, NCR, and OHRA es desarrollado CRISP-DM como metodología generalista y básica para el desarrollo no concesionado y no académico de minería de datos.

3.1.1. Knowledge discovery in databases

A mediados de los años 1990's, con el avance tecnológico de las bases de datos, potencializar los modelos estadísticos significaba un nuevo reto. [Piatetsky-Shapiro, 1990] postuló que para poder realizar una búsqueda del conocimiento se requiere de numerosas herramientas como sistemas expertos de bases de datos, machine learning, inteligencia de negocio y estadística. En la metodología KDD, la minería de datos es una parte más del proceso de descubrimiento de conocimiento. Con ello se postularon los siguientes pasos:

1. Entendimiento: La tarea fundamental de Minero de Datos es la de conocer el o las problemáticas reales de la situación.
    a) Entrevistas a expertos
    b) Análisis superficial
    c) Revisión teórica y practica
2. Limpieza de los datos: A partir de reglas de negocio, técnicas estadísticas y de imputación, se tratará de revisar, reparar y eliminar las aberraciones y datos inconsistentes.
    a) Transformaciones
        1) Discretización
        2) Normalización
    b) Matemáticas y Estadísticas
        1) Reducción de dimensionamiento

2) Transformación de Furier

3) Funciones matemáticas

c) Modificación

1) Imputación

2) Rotación

3) Filtrado

3. Modelación: Es la búsqueda de patrones en los datos que sean útiles para el problema a resolver. Es la parte central de la minería de datos, pues los pasos anteriores se requieren para tener un mayor contexto de la situación y la mitigación de conclusiones y los pasos siguientes serán para implementar la conclusión de esta fase.

a) Agrupamiento

1) K-Means

2) KNN

3) BDSCAN

b) Clasificación

1) Arboles de decisión

2) Redes neuronales

3) Modelos bayesianos

4) SVM

c) Model Lineal Generalizado

1) Lineal Múltiple

2) Logístico

3) Mínimos Cuadrados Parciales

4. 4. Implementación: en esta fase se validan los patrones, deberán expresarse de manera que los tomadores de decisiones y/o dueños tenada una interpretación sencilla y retro-alimentarle.

    a) Evaluación

        1) Pruebas de negocio

        2) Bootstrap

        3) Evaluación de Modelos ROC

    b) Visualización

    c) Simulación

Con estas herramientas Fayyad et al. [1996] proponen el proceso KDD, que otorga a los investigadores un proceso lógico para establecer las bases de la minería de datos. El principal aporte de esta metodología es importante por ser pionera en el campo de la implementación de la minería de datos. Sin embargo, algunas fases dentro de la metodología eran difusas y otras fuera de foco, derivado a las necesidades de los creadores de la metodología, por lo cual los especialistas de cada área argumentaban que se debería tener mayor explotación de alguna área u otra fase de KDD ante esto, se hicieron esfuerzos por parte de la industria de manejo de datos en proponer metodologías que conceptualizaran la minería de datos. Este las más reconocidas se encuentra SEMMA.

3.1.2. SEMMA

Conocido y desarrollado dentro de los estándares de la empresa tecnológica SAS, SEMMA es la propuesta estándar para los proyectos de minería de datos orientados a la oferta de software de SAS. Su desarrollo empezó en 1995 [SAS, 1998] en un esfuerzo de conglomerar las ideas revolucionarias de la minería de datos. Para 1998, se creaba el SAS Enterprise Miner el cual toma la minería de datos, como un proceso de integración. En este punto, SAS tiene

como propuesta el uso de la estadística al servicio de las empresas privadas, con un diseño de data Warehouse y minería, al explotar la inteligencia de negocios al máximo.

Figura 3.1.1: Diagrama fases del modelo SEMMA. Fuente: SAS [1998]Elaboración propia.

SEMMA es un acrónimo en inglés de Sample (Muestreo), Explore (exploración), Modify (modificación), Model (Modelación) y Assess (Evaluación).

- Sample: Funciona como la parte de creación y declaración de las tablas de desarrollo. Adicional tiene funcionalidades de creación de particiones para las pruebas en la evaluación (Entrenamiento, Validación y prueba), importación y muestreo estratégico (Proporcional o Equiprobable).

- Exploración: Es de fundamental importancia en el enfoque SEMMA, el entendimiento de los datos y el problema a resolver, por lo que la presentación de los valores es generada en esta fase con gráficos, agrupamiento de variables y agrupamiento de los registros.

- Modify: Al tener integrada y entendida la base, es necesario hacer las modificaciones necesarias que el modelo requiera; eliminación de registros aberrantes, imputación de valores perdidos, transformación de variables numéricas y cadena son parte de las etapas de la estrategia.

- Model: Es la fase principal de la metodología, en la cual, a partir de proponer y unir distintos modelos de minería de datos, se propone una solución válida para el problema en cuestión.

- Assess: Al tomarse en cuenta las soluciones propuestas en Model, se toman, evaluan y aplican.

La particularidad de esta metodología es la creación rápida y sencilla de modelos de minería de datos útiles, facilitado por el software de los desarrolladores. Dentro de la industria es difícil encontrar un software que desde la concepción tengan la metodología y las aplicaciones de la mano.

Las principales ventajas de esta metodología son, un software simbiótico entre el procedimiento y las herramientas, esto significa que existe literatura y crecimiento cognitivo gradual, en fáciles cursos técnicos puedes tener resultados fáciles y confiables. No requieres ser un experto técnico o estadístico para poder utilizar la herramienta, aunque si tener un contexto del problema de negocio y nociones básicas del modelado. Sin embargo, las dificultades de estar unido a un software propietario, hace que otros sistemas busquen ser diferentes con propuestas diferenciadas, lo que hace que la metodología no se masifique a quienes no utilizan el software. Defensores de KDD, mencionaban que las fases estaban sobre agrupadas y enfocadas a la parte estadística, la cual es la necesidad de desarrollador del producto.

3.1.3. CRISP-DM

Dentro de las metodologías más populares de las aplicaciones libres y privadas, consultorías y la investigación encontramos la metodología CRISP-DM, que es un acrónimo CRoss-Industry Standard Process for Data Mining. La metodología fue creada por SPSS, NCR y Daimer Chrysler [Chapman et al., 2000], en la cual se establecen conjuntos de tareas y marcos de referencia sin regular las actividades en ellos (Moine et al.).

Figura 3.1.2: Evolución CRISP-DM durante el desarrollo del proyecto. Fuente: [Goicochea, 2017] Elaboración propia.

El gran potencial en el negocio bancario, el cual está acostumbrado a tener metodologías lineales, es que, con esta metodología, el trabajo tiene una forma cíclica de elaboración de proyecto ya que la finalidad de esta no es la finalización del proyecto si no la búsqueda del conocimiento. En un punto de vista general, se tienen varios niveles para el proceso, se tienen seis fases las cuales se describen en tareas genéricas de acuerdo con situaciones específicas.

Figura 3.1.3: Fases del proceso CRISP-DM. Fuente: [Chapman et al., 2000]Elaboración propia.

En la figura 3.1.3 se puede observar las fases generales de la metodología. En las cuales se describe la relación de dependencia de cada una de las fases y subfases. Cabe destacar que la mención de las subfases, no son restrictivas o dependientes para el funcionamiento final del proceso [Goicochea, 2017]. En las tareas genéricas, busca hacer la mayor parte del proceso, en las cuales se tienen tareas especializadas.

El tercer nivel, está constituido de tareas particulares de casos de negocio y que constituyen pasos importantes para las tareas genéricas. En este trabajo se mencionan los ejemplos más importantes y utilizados por la metodología. Las instancias del proceso, último nivel de la metodología, son una serie de acciones en particulares que se realizan que especifican la función de cada uno de los procesos.

Entendimiento del negocio

Para iniciar la generación de conocimiento, en esta primera fase, se buscará conocer los antecedentes, perspectivas y reglas. Convertiremos estos conocimientos en entradas de las bases de datos o modelado. Además, de que definiremos el problema y las necesidades de la investigación.

- Determinar los objetivos de negocio: Esta es la fase de entendimiento del negocio, donde se buscará conocer las particularidades y factores importantes del problema a tratar.
    - Contexto: Recabar información de la situación del negocio para el inicio del proyecto.
    - Objetivos: Describe los objetivos primarios del negocio asociados a la necesidad del negocio y no a la necesidad de problema, comúnmente.
    - Criterios de Éxito: Determinar la medida de éxito asociada, la cual debe ser objetivamente medible.
- Evaluación de la situación : Esta tarea relaciona la importancia de descubrir y envolver al investigador sobre los recursos, supuestos que deben ser considerados. Si bien la tarea

anterior es para conocer el problema central de la investigación, esta tarea busca expandir el conocimiento de esta misma.

  - Inventario de recursos: Lista de recursos disponibles para el proyecto cuantificable en hardware, software, informacionales y humanos.
  - Requerimientos, supuestos: Lista de requerimientos y supuestos que nos apoyen y no limiten la investigación posteriormente.
  - Terminologías propias del negocio: Recopilar y entender las terminologías propias del sector, negocio y área con el fin de no obviar conocimiento.
  - Costos y beneficios: elaboración de análisis de costo-beneficio el cual compare los potenciales beneficios de la investigación.

- Establecer objetivos: Los objetivos de negocio se establecerán en función a la necesidad y terminología de la investigación o negocio.
  - Objetivos: Describe en términos técnicos los objetivos de la investigación.
  - Criterios de éxito: Define los criterios de éxito en términos técnicos de la investigación como; nivel de predicción esperada.
- Generación del plan del proyecto Esta tarea describe la dirección e intención del plan de los objetivos de negocio, reaccionados con los objetivos técnicos, además deberá contener los primeros pasos a detalle, así como las técnicas y herramientas a priori para resolver el problema.
  - Plan, herramientas, equipo y técnicas

Entendimiento de los datos

Se inicia con la recolección inicial de dato, busca que el investigador se familiarice con los datos para poder encontrar e identificar de forma temprana errores en la calidad de la informa-

ción, procesos y cumplimiento de los supuestos. Hay que destacar que errores, incoherencias o mala calidad de información da como resultado el fracaso de los proyectos.

- Recopilación inicial de datos : Es la identificación de la base de datos a utilizar, además se detalla cómo se obtuvo y problemas encontrados.
- Descripción de los datos : Se incluyen los análisis de formato de datos, calidad de los datos, identificación de los datos, así como reportar hallazgos importantes en el proceso.
- Exploración de los datos : Se da la búsqueda de las preguntas de los objetivos técnicos a partir de consultas, análisis visuales, descriptivos y exploratorios.
- Verificación de calidad de datos : se refuerza a una validación post exploratoria en la cual buscaremos que la calidad de los datos no se haya visto afectada por las transformaciones anteriores, al realizar preguntas de investigación se vean respondidas de manera inicial.

Preparación de los datos

En las instituciones y centros de investigación la información no siempre se encuentra con la estructura y formatos que nos permitan la explotación de esta, por lo cual se deben seguir tareas particulares para hacerlo.

- Selección de los datos : se decide que datos se usarán para el análisis y modelado de datos, tanto en volumen, calidad y objetivos.
- Limpieza de datos : se realizan modificaciones para la tener los datos de la mejor forma de ser tratados tanto para el tipo de modelado como para las conclusiones no redundantes. Se conoce como la auditoria de la calidad de los datos.
- Construcción de datos : Ya con los datos puros y limpios, se revisarán las posibles construcciones de datos que se podrán utilizar, se pueden transformar los datos, atributos, variables y registros en pro del resultado

- Integración de datos : se incluye en caso de que la información se encuentre en distintos repositorios de información o de que datos no incluidos naturalmente en el análisis se agreguen para dar valor y peso.

- Formateo de datos : En algunos casos, se buscará dar el formato correcto a las variables para dar un mayor peso o que simplemente puedan ser modelados. Esto en función de agregar valor y no limitarse a los datos iniciales y su construcción.

Modelado

- Es en esencia el valor agregado, la búsqueda de patrones y la identificación de los perfiles que el negocio requiere o necesita para realizar estrategias en mejora de los niveles. Es en este punto donde el investigador no se limita a utilizar uno o dos modelos, busca realizar una búsqueda exhaustiva de la mejor predicción con un sentido de negocio valido e implementable.

- Selección de la técnica de modelado : se refiere a la búsqueda de los mejores modelos que se ajusten a los tipos y volumen de datos. Diseño de la evaluación : Antes de realizar las pruebas de modelado, se debe tener claro, los criterios de calidad y predicción. Además, se crean las particiones de modelado, prueba y evaluación, ya sea de manera transversal aleatoria, longitudinal en el tiempo o totalmente aleatoria.

- Construcción del modelo : Ejecución y ajuste de los modelos seleccionados para la prueba.

- Evaluación del modelo : Se interpretan los resultados del modelado, teniendo en cuenta tres aristas: predicción estadística, valor al negocio e implementación. Se deben tener valores históricos de referencia, así como las particiones del modelo. Se debe conocer los lineamientos estadísticos y de negocio para encontrar conclusiones redundantes, poco valor estadístico y problemas en los pasos anteriores de construcción de la base de datos.

Evaluación

Es la revisión de los objetivos de negocio planteado en un inicio del análisis y se decidirá el uso de los resultados.

- Valuación de resultados : en este proceso, se realiza la valuación de los objetivos de negocio junto con el nivel de predicción del modelado. Se debe realizar un resumen de la evaluación y emitir los ganancia y perdidas respecto a los objetivos de negocio de cada modelado.

- Revisión del proceso : es un paso de retrospectiva de la metodológica, donde ya con los resultados, se evalúa si ha sido la mejor forma de construir los datos, modelar o evaluar el problema.

- Establecimiento de los siguientes pasos o acciones : es el punto, donde debe evaluarse si ha sido satisfactorio la solución y pasar a la fase de implementación o si es requerido pasar nuevamente por pasos anteriores de la metodología.

Aplicación

El conocimiento obtenido durante el ciclo del modelado y evaluación se debe aplicar al ámbito práctico, con lo cual la aplicación en vivo es fundamentalmente para el crecimiento de la investigación en caso de ser requerida posteriormente.

- Planificación de despliegue: Se crea un plan de implantación de modelo y documentar los hallazgos durante este proceso.

- Planificación de la motorización y del mantenimiento: es la evaluación constante del modelo, para localizar inconsciencias locales o mejoras no perceptibles durante el proceso de evaluación.

- Generación de informe final: Al término del proyecto, es importante documentar el proceso de generación de valor, experiencias y mejoras al proceso.

- Revisión del proyecto: Evaluación de constructiva de las cosas realizada correctamente y las cosas que requieren una mejora.

El proceso CRISP-DM fue diseñado para dar una guía a los practicantes de la minería de datos, con un enfoque realista y práctico, fuera de metodologías técnicas, teóricas o de carácter de la industria privada. Como tal de ser una guía, no contempla un manual para el conocimiento o los resultados efectivos de las investigaciones, las cuales quedan en responsabilidad de le investigador y su pericia para utilizar esta herramienta como un marco de referencia.
En el caso de CRISP-DM, en función de la implementación de un modelo de negocio, se observan las ventajas a la hora de la evaluación y el uso de los hallazgos de la minería de datos. Como punto a favor, es que es una metodología basada para los casos de usos aplicados diferentes industrias, además de que es un desarrollo no privado, lo cual resulta accesible y fácil de documentar. Como desventaja es que los softwares como Python y R no están desarrollados bajo esta metodología de trabajo. Se ha elegido CRISP-DM como metodología de trabajo como reflexión de mejora explicativa y de implementación.

3.2. Python

Python es un lenguaje de programación orientado a objetos, interpretado e interactivo creado por Guido van Rossum, en Países Bajos durante 1991, el cual a partir de una sintaxis sencilla busca facilitar la programación del usuario, generando así un despliegue de librerías, clases y módulos que coadyuvan a mejorar el desarrollo de programas y aplicaciones. Además, la interfaz es multiplataforma lo cual ha facilitado su uso entre los programadores [Python, 2018]. Tiene una licencia de código abierto, si bien el lenguaje está protegido por derechos de autor, es modificable y redistribuible. Se tiene una interfaz de entrada llamada Editor y una interfaz de salida llamada Terminal.

3.2.1. Anaconda

Anaconda es un paquete de distribución basado en Python y R, el cual está enfocado a los científicos de datos. Es utilizado por seis millones de usuarios, en código abierto. Contiene más de 1,400 paquetes relacionados con la ciencia de los datos [Anaconda, 2018]. Entre sus principales asociados IDEs se encuentran Jupyter, JupyterLab, Spyder y RStudio. Para análisis de datos Dask, numpy, pandas and Numba. Para visualización de datos Matplotlib, Bokeh y Datashader. Y para aprendizaje automático Scikit-learn, entre otros.

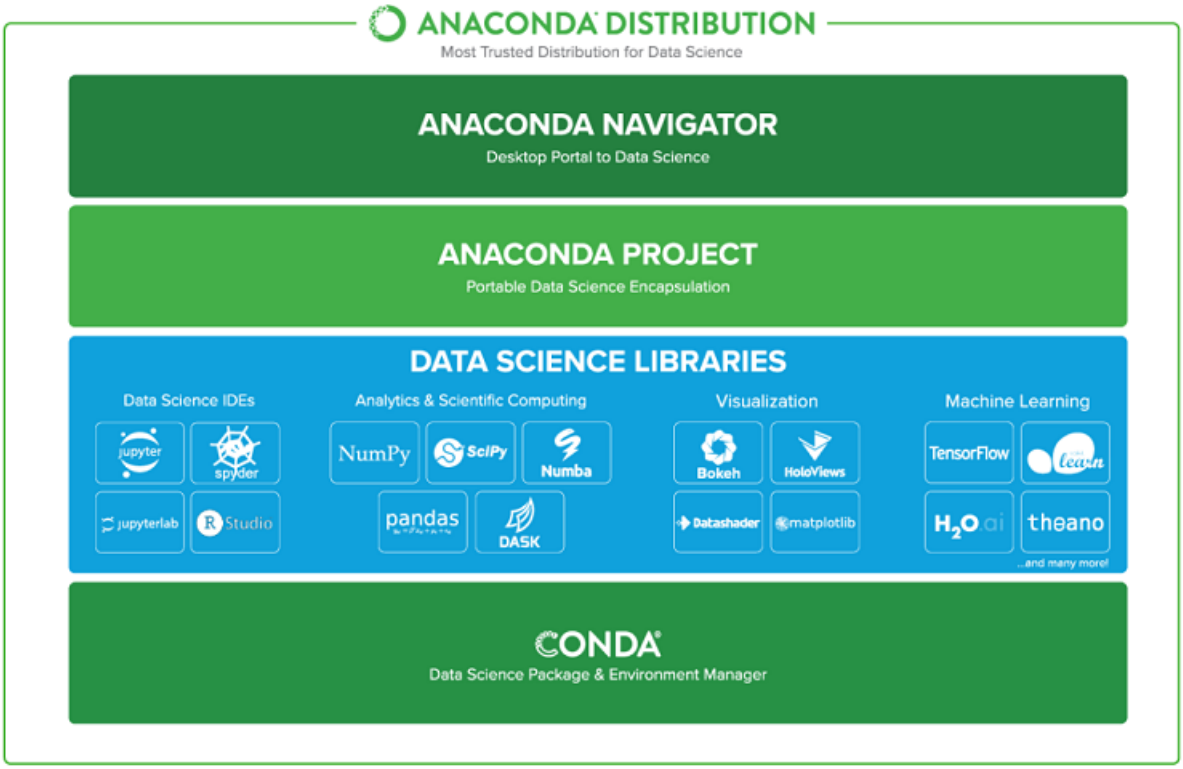


Figura 3.2.1: Mapa de interfaces incluidas en Anaconda. Fuente: Anaconda [2018]

3.2.2. Elección de software estadístico

En el desarrollo de herramientas estadísticas y de minería de datos, es frecuente el uso de herramientas licenciadas y libres, siendo para los investigadores las segundas de mayor aceptación por el uso sin pago [Rodriguez Sotelo, 2015]. Entre las aplicaciones con mayor uso

se encuentran Python, R, SAS, SPSS y aplicaciones basadas en Java (RapidMiner). Con ello parte de los objetivos del presente documento es mostrar el aprendizaje para el lector de una herramienta útil en el sector bancario.

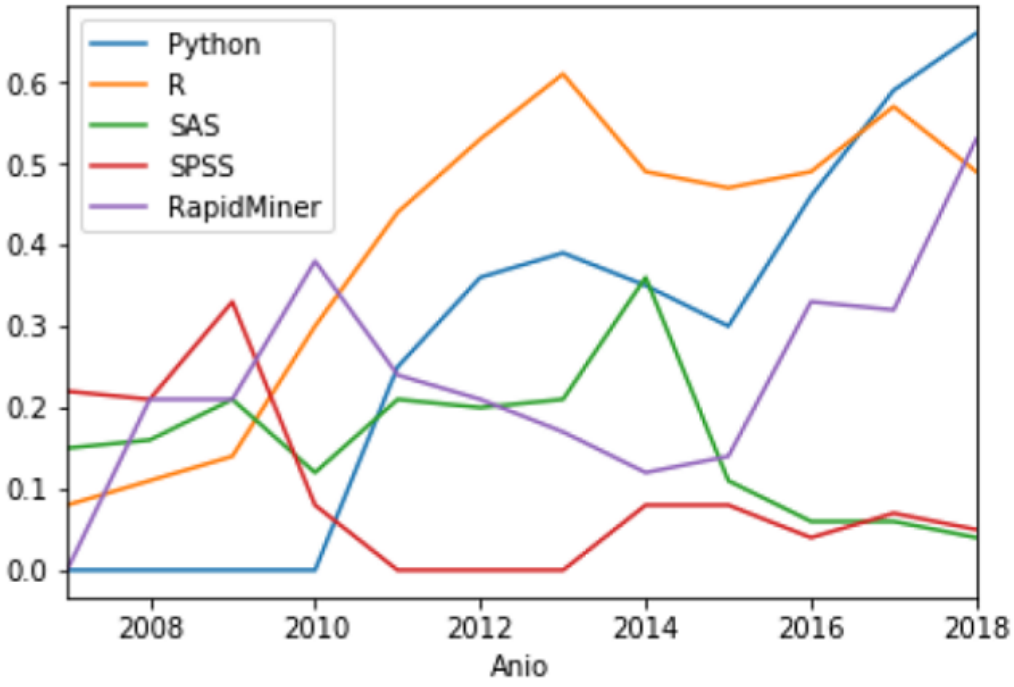


Figura 3.2.2: Estadísticas de uso de programas estadísticos. Fuente: Piatetsky-Shapiro [2008-2018] Elaboración propia.

En el caso de KdNuggets [Piatetsky-Shapiro, 2008-2018], pagina especializada en el uso de software analítico, minería de datos, big data y data scientist, se observa que en las encuestas anuales Python a lo largo de ocho años ha incrementado su tendencia de uso sobrepasando al mayor ocupante de los últimos años, R. Esto se puede derivar a que la curva de aprendizaje es muy corta [Willems, 2015]. También se tienen los datos de las tendencias de búsqueda para Google [Google, 2018].

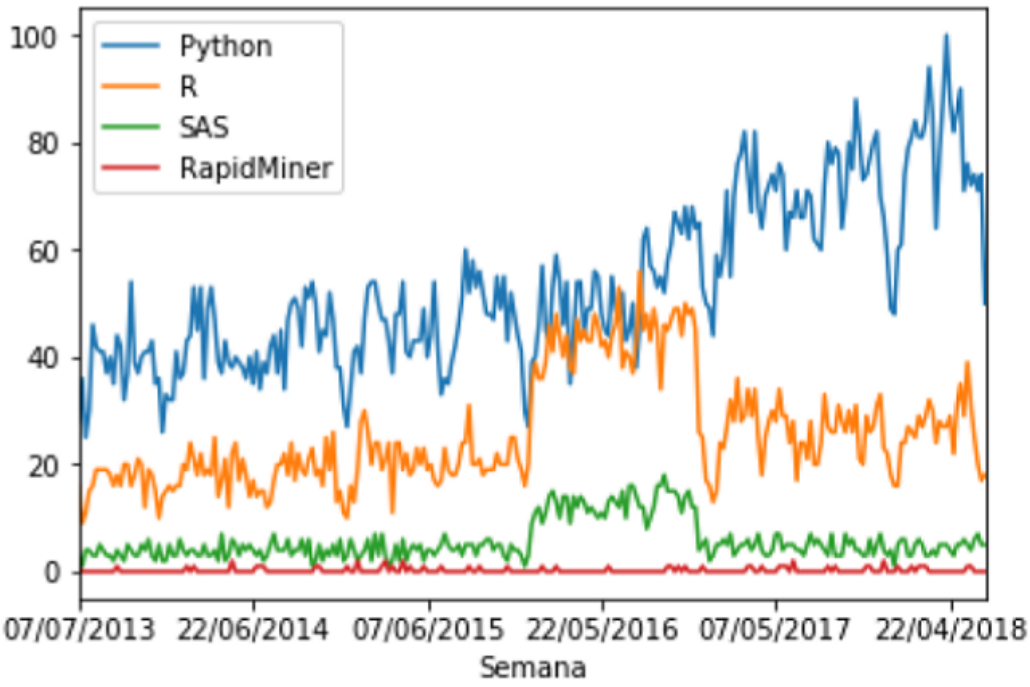


Figura 3.2.3: Estadísticas de búsqueda y tendencia en Google. Fuente: Google [2018]Elaboración propia.

A pesar de no ser una estadística de uso, si corresponde a la tendencia de interés, el cual podría ser no solo de programadores sino de otros perfiles que buscan conocer más de la herramienta [BBVA, 2016]. Siendo R uno de los principales representantes del uso de la minería de datos, es para mucho un lenguaje de programación lento [Rodriguez, 2015] y con re usos de sintaxis de Fortran y C++ [BBVAOpen4U, 2016]. Otro aliciente más para conocer Python es que el último año ha crecido su demanda y pago, de $72,000 dólares anuales a los $92,000. Mientras que R ha decrecido de $109,000 a los $78,000 [ziprecruiter, 2018, glassdoor, 2018]. El crecimiento de estas herramientas y el despunte contra las principales marcas licenciadas se han hecho a partir de la sana competencia entre los desarrolladores, el cual nos da un valor agregado a cada una de las herramientas[Jain, 2017].

3.2.3. Principales procesos numéricos y estadísticos

En el caso del presente trabajo, se utiliza con IDE Spyder. Y las siguientes librerías: Pandas. Panda es una librería de código abierto que incluye estructura y análisis de datos en Python, la cual ayuda al manejo de datos y estadística, sin ser un usuario experto de un programa destinado para este fin como R [Augspurger et al., 2018]. Las principales utilerías son el manejo de base de datos con herramientas de lectura y escritura sobre formatos de texto plano, archivos, bases de datos, entre otros. Además, se pueden realizar consultas programadas de las bases de datos como indexar, unir, cruzar y agrupar.

Llamado de la librería.

La librería de Pandas se llama con una sentencia de importación de librerías.

Editor	import pandas as pd
Terminal	import pandas as pd

Cuadro 3.1: Llamado de librerías. Fuente: Elaboración propia.

Integración de información

Entre las principales tareas del manejo de base de datos es la lectura de archivos. En Pandas, la lectura de archivos CSV. Se realiza con la sentencia read_csv.

Editor	data=pd.read_csv('C:/Users/Archivo.csv')
Terminal	data=pd.read_csv('C:/Users/Archivo.csv')

Cuadro 3.2: Inserción de información. Fuente: Elaboración propia.

Esta sentencia lee el archivo y lo transforma en una variable denominada data.

Visualizar datos precargados

Los datos previamente cargados son legibles al integrar el nombre de la variable.

Editor	data				
Terminal	s	loan_amnt	funded_amnt	term	int_rate
	0	10400	10400	36 mon	6.99 %
	1	15000	15000	60 mon	12.39 %
	2	9600	9600	36 mon	13.66 %
	3	21425	21425	60 mon	15.59 %
	4	12800	12800	60 mon	17.14 %
	5	7650	7650	36 mon	13.66 %
	6	17000	17000	36 mon	13.66 %
	7	21075	21075	60 mon	21.99 %
	8	16000	16000	60 mon	11.44 %
	9	23325	23325	36 mon	14.31 %

Cuadro 3.3: Despliegue de datos en pantalla. Fuente: Elaboración propia.

Información general

Para obtener la información general de los datos precargados. Se utiliza la sentencia .info.

Editor	data.info.()
Terminal	class pandas.core.frame.DataFrame> RangeIndex: 235633 entries, 0 to 235632 Columns: 145 entries, id to settlement_term dtypes: float64(108), object(37) memory usage: 260.7+ MB

Cuadro 3.4: Obtención de Información general. Fuente: Elaboración propia.

Exportación de información

Se tienen funciones de exportación sencillas con to_csv.

Editor	des_df.to_csv(C:/Users/salida.csv)
Terminal	des_df.to_csv(C:/Users/salida.csv)

Cuadro 3.5: Sentencia para exportar información en CSV. Fuente: Elaboración propia.

Filtrado de información

Se muestran el filtrado en la selección registros.

Editor	data_tdc = data	[(data.purpose	== çredit_card")]		
	loan_amnt	funded_amnt	funded_amnt_inv	term	purpose
0	10,400	10400	10400	36 mon	credit_card
Terminal 1	15,000	15000	15000	60 mon	credit_card
2	9,600	9600	9600	36 mon	credit_card
3	21,425	21425	21425	60 mon	credit_card

Cuadro 3.6: Filtrado de información. Fuente: Elaboración propia.

Se muestran el filtrado en la selección registros de forma negativa.

Editor	data_tdc = data	[(data.purpose	=! çredit_card")]		
	loan_amnt	funded_amnt	funded_amnt_inv	term	purpose
Terminal 1	10,000	10000	10000	60 mon	credit_card
2	15,600	20000	20000	36 mon	credit_card
3	20,000	20000	20000	42 mon	credit_card

Cuadro 3.7: Filtrado de información de forma negativa. Fuente: Elaboración propia.

Inhibición de variables

Para el retiro de variables se tiene la función

Editor	dg=	df.	drop	(['loan_amnt	'])
		funded_amnt	funded_amnt_inv	term	purpose
Terminal	1	10000	10000	60 mon	debt_consolidation
	2	20000	20000	36 mon	other
	3	20000	20000	42 mon	house

Cuadro 3.8: Quitar variables de una base de datos. Fuente: Elaboración propia.

Contención de variables

Para el mantener variables se tiene la siguiente sintaxis.

Editor	dg=df.	keep (['funded_amnt'])
		funded_amnt
Terminal	1	10000
	2	20000
	3	20000

Cuadro 3.9: Mantener variables de una base de datos. Fuente: Elaboración propia.

Cruce de información

En el caso de requerir que la información de dos bases de datos se una por una llave definida, lo podemos realizar con la tarea merge.

Editor	df=	dg.merge(pd_2,	left_on=New_ID,	right_on=New_ID,	how=inner)
		funded_amnt	funded_amnt_inv	term	purpose
Terminal	1	10000	10000	60 mon	debt_cons
	2	20000	.	36 mon	other
	3	20000	20000	42 mon	house

Cuadro 3.10: Cruce de información. Fuente: Elaboración propia.

Rellenado de nulos

Para forzar que una variable que contiene valores nulos se llene con algún valor predefinido, utilizaremos fillna.

Editor	df=	dg.merge(pd_2left_on,=	New_ID, right_on	=New_ID,	how=inner)
		funded_amnt	funded_amnt_inv	term	purpose
Terminal	1	10000	10000	60 mon	debt_cons
	2	20000	0.0	36 mon	other
	3	20000	20000	42 mon	house

Cuadro 3.11: Imputación de Nulos. Fuente: Elaboración propia.

Análisis descriptivo

En análisis de valores de tendencia central y dispersión es de vital importancia en la realización de estadísticas y análisis de la base de datos. Para tener un resumen general se puede utilizar describe.

Editor	dh=	df.	describe()			
Medidas	a	b	c	d	e	f
	count	4,949.00	4,949.00	4,949.00	4,949.00	4,949.00
	mean	0.0	0.0	0.0	0.0	0.0
	std	1.44	1.03	0.97	0.70	0.67
Terminal	min	6.14	3.83	3.65	2.38	2.80
	1Q	0.97	0.71	0.67	0.48	0.44
	2Q	0.01	0.01	0.03	0.05	0.01
	3Q	0.98	0.68	0.65	0.39	0.45
	max	9.75	3.92	5.16	5.50	8.61

Cuadro 3.12: Imputación de Nulos. Fuente: Elaboración propia.

Para el análisis de valores específicos como la media o conteo de valores. En el caso de la media utilizaremos mean().

Editor	dh=df.mean()
Terminal	a 0.519122 b 0.439866 c 0.551054 d 0.452423

Cuadro 3.13: Cálculo de la media. Fuente: Elaboración propia.

Para la desviación estándar

Editor	dh=df.mean()
Terminal	a 0.216855 b 0.239132 c 0.197108 d 0.280768

Cuadro 3.14: Cálculo de la desviación estándar. Fuente: Elaboración propia.

Análisis de Correlación

Para realizar el análisis de correlación se ocupa la sentencia

Editor	dg=df.	df.drop(['	loan_amnt ')		
		a	b	c	d
	a	1	0.362944	-0.288077	-0.465824
	b	0.362944	1	0.579711	-0.678864
Terminal	c	-0.288077	0.579711	1	-0.625552
	d	-0.465824	-0.678864	-0.625552	1

Cuadro 3.15: Cálculo del coeficiente de correlación. Fuente: Elaboración propia.

Determinación de cuantiles

Para encontrar el valor de cuantiles, se utiliza la función qcut.

Editor	df2.qcut(range(5), 4)
Terminal	[(-0.001, 1.0], (-0.001, 1.0], (1.0, 2.0], (2.0, 3.0], (3.0, 4.0]]

Cuadro 3.16: Cálculo de la desviación estándar. Fuente: Elaboración propia.

Visualización

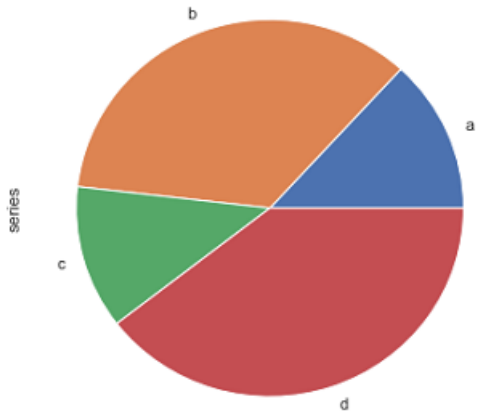
Pandas tiene uno de los anaqueles más amplios de gráficos en Python. Para ello se requiere importar la subclase Matplotlib.

Editor	import Matplotlib.pyplot as plt
Terminal	

Cuadro 3.17: Llamado de librería Marplotlib. Fuente: Elaboración propia.

La filosofía de esta utilería es fácil el manejo de gráficos sencillos y complejos, en una alta gama de gráficos como líneas, barras, histogramas, mapas de dispersión, espectrogramas, etc. El módulo pyplot da al usuario una interfaz similar a MATLAB que ayuda a la interacción de la librería con el usuario.

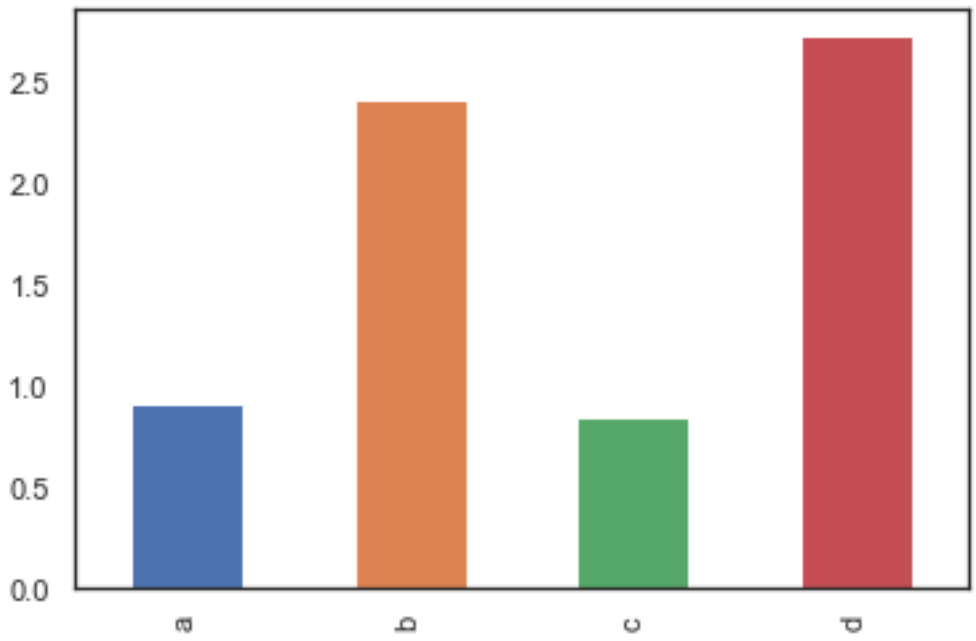
Editor	series = pd.Series(3 * np.random.rand(4), index=['a', 'b', 'c', 'd'], name='series') series.plot.pie(figsize=(6, 6))
Terminal	



Cuadro 3.18: Diagrama de pie Gráficos de barras. Fuente: Elaboración propia.

Gráficos de barras

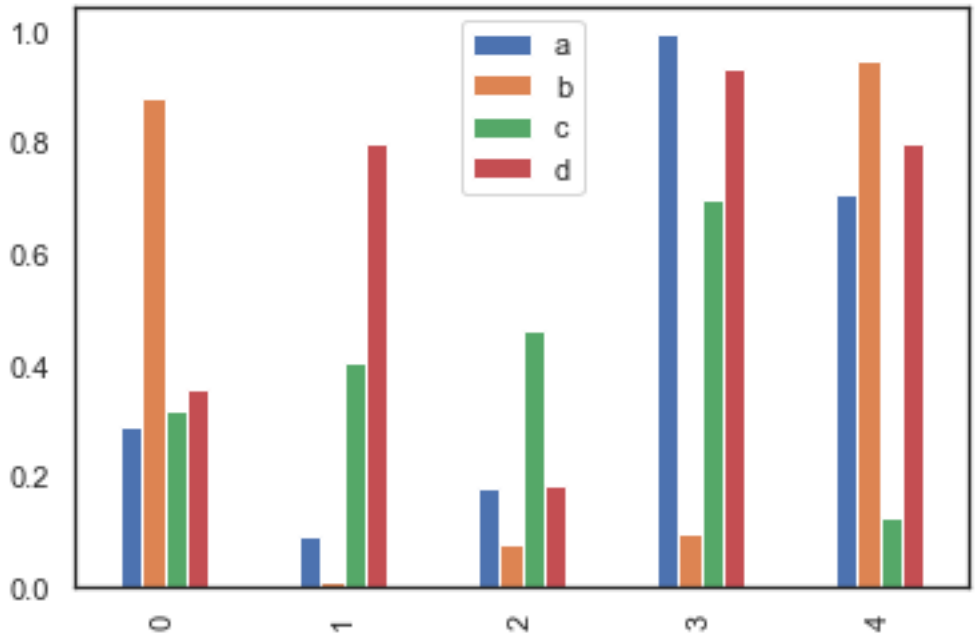
Editor	series.plot.bar()
Terminal	



Cuadro 3.19: Gráficos de barras. Fuente: Elaboración propia.

Diagrama de barra por clase

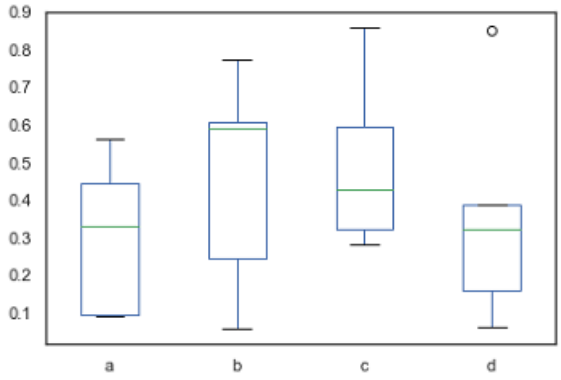
Editor	pd.DataFrame(np.random.rand(5, 4), columns=['a', 'b', 'c', 'd']).plot.bar()
Terminal	



Cuadro 3.20: Diagrama de barra por clase. Fuente: Elaboración propia.

Diagramas de medición de media y dispersión

Editor	pd.DataFrame(np.random.rand(5, 4), columns=['a', 'b', 'c', 'd']).plot.bar()
Terminal	



Cuadro 3.21: Diagrama de barra por clase. Fuente: Elaboración propia.

Numpy

Paquete fundamental de Python. Contiene arreglos dimensionales, funciones sofisticadas y herramientas de álgebra, lógica, transformaciones y números aleatorios.

Creación de arreglos

En Python, los arreglos numéricos se encuentran en estructuras matriciales y se pueden convertir a otras estructuras.

Editor	np.array([2,3,1,0])
Terminal	array([2, 3, 1, 0])

Cuadro 3.22: Arregle matricial en Python. Fuente: Elaboración propia.

Para la generación de variables aleatoria Numpy tiene una utilería fácil de usar para la generación de variables cuasi aleatorias. Con Rand, la generación es totalmente aleatoria.

Editor	np.random.rand(3,2)
Terminal	array([[0.71937001, 0.59680426], [0.12854059, 0.33543028], [0.39126029, 0.31921263]])

Cuadro 3.23: Generación de Variables Aleatorias. Fuente: Elaboración propia.

Existen en la documentación numerosas formas de generar variables a partir de esta librería. Es común utilizar las funciones Rand, Uniform y Normal.

Sklearn

Se define como una librería de aprendizaje automático. Contiene numerosas funcionalidades clasificación, regresión y agrupación [Pedregosa et al., 2018].

Componentes principales

La técnica de componentes principales es una función que se encuentra en la librería Sklearn. Para ello llamaremos a un arreglo de valores.

Llamado de librerías

Editor	from sklearn.decomposition import PCA
Terminal	

Cuadro 3.24: Carga de librería para PCA. Fuente: Elaboración propia.

Editor	pca = PCA(n_components=2)
Terminal	

Cuadro 3.25: Declaración de número de componentes principales. Fuente: Elaboración propia.

Ajuste de componentes principales

Editor	pca.t(df2)
Terminal	PCA(copy=True, iterated_power='auto', n_components=2, random_state=None svd_solver='auto', tol=0.0, whiten=False)

Cuadro 3.26: Ajuste del modelo por PCA. Fuente: Elaboración propia.

Revisión de varianza explicada por PCA

Editor	print(pca.explained_variance_ratio_)
Terminal	[0.35419602 0.28177304]

Cuadro 3.27: Ajuste del modelo por PCA. Fuente: Elaboración propia.

Revisión de valores propios

Editor	print(pca.singular_values_)
Terminal	[1.99490183 1.77930095]

Cuadro 3.28: Identificación de los valores propios. Fuente: Elaboración propia.

Matriz de covarianza

Editor	pca.get_covariance()
Terminal	[array([[0.10371745, -0.01643809, -0.0045944, -0.02479411]

Cuadro 3.29: Matriz de covarianza estimada. Fuente: Elaboración propia.

Matriz de inversa

Editor	pca.get_precision()
Terminal	array([[0.10371745, -0.01643809, -0.0045944, -0.02479411]

Cuadro 3.30: Matriz de inversa. Fuente: Elaboración propia.

Regresión Logística

Bases de desarrollo y prueba

Esta clase ayuda al analista a realizar cuatro subpoblaciones para realizar el ajuste y la prueba del modelo generado.

Editor	X_train, X_test, y_train, y_test= train_test_split(df3.iloc[:, 0:5], df3.a, test_size = .7, random_state=25)

Cuadro 3.31: Generación de bases de desarrollo y prueba del modelado. Fuente: Elaboración propia.

Ajuste del modelo

Se realizará el ajuste del modelo con la variable objetivo en y_train y X_train como variables explicativas

Editor	LogReg = LogisticRegression() LogReg.fit(X_train, y_train)
Terminal	LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100, multi_class=ovr, n_jobs=1, penalty=l2, random_state=None, solver=liblinear, tol=0.0001, verbose=0, warm_start=False)

Cuadro 3.32: Ajuste de la regresión logística. Fuente: Elaboración propia.

Predicción del modelo en forma del objetivo

Con esta herramienta se evaluó con el modelo generado una predicción del modelo.

Editor	LogReg.predict(X_train)
Terminal	array([1., 0., 1., 1., 1., 1., 1., 0., 0.])

Cuadro 3.33: Predicción del modelo en formato del objetivo. Fuente: Elaboración propia.

Predicción del modelo en probabilidad

Con esta herramienta se evalúa con el modelo generado una predicción del modelo.

Editor	LogReg.predict_ proba(X_train)
Terminal	array([[0.2110437, 0.7889563], 0.50533549, 0.49466451], 0.18363663, 0.81636337], 0.24274443, 0.75725557], 0.206961, 0.793039], 0.21152309, 0.78847691], 0.19206873, 0.80793127], 0.52023846, 0.47976154], 0.51273479, 0.48726521]])

Cuadro 3.34: Predicción del modelo en probabilidad. Fuente: Elaboración propia.

Statsmodels

Módulo enfocado en clases y funciones de estimación como modelos, pruebas y exploración estadística de datos. El cual en conjunto con Pandas y Nunpy se tienen una buena herramienta de análisis estadístico [Perktold et al., 2018].

Ajuste de parámetros a una regresión logística

Editor	regressorOLS = sm.OLS(y_train, X_train).fit()
Terminal	statsmodels.regression.linear_model.RegressionResultsWrapper

Cuadro 3.35: Ajuste de parámetros en regresión logística. Fuente: Elaboración propia.

Resumen de parámetros

Editor	regressorOLS.summary()
Terminal	

OLS Regression Results

Dep. Variable:	target	R-squared:	0.054
Model:	OLS	Adj. R-squared:	0.053
Method:	Least Squares	F-statistic:	50.93
Date:	Fri, 06 Jul 2018	Prob (F-statistic):	1.67e-51
Time:	23:13:50	Log-Likelihood:	390.45
No. Observations:	4482	AIC:	-770.9
Df Residuals:	4477	BIC:	-738.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
num_actv_bc_tl	0.0040	0.001	2.906	0.004	0.001	0.007
total_rev_hi_lim	-1.184e-07	1.2e-07	-0.988	0.113	-3.53e-07	1.17e-07
tot_cur_bal	-4.189e-08	2.68e-08	-1.563	0.118	-9.44e-08	1.07e-08
annual_inc	1.137e-07	7.84e-08	1.450	0.147	-4e-08	2.67e-07
acc_open_past_24mths	0.0078	0.001	7.122	0.000	0.006	0.010

Omnibus:	3551.112	Durbin-Watson:	2.024
Prob(Omnibus):	0.000	Jarque-Bera (JB):	48004.132
Skew:	3.960	Prob(JB):	0.00
Kurtosis:	16.940	Cond. No.	1.06e+05

Cuadro 3.36: Resumen y estadísticos de ajuste de parámetros en regresión logística. Fuente: Elaboración propia.

Scipy

Es una colección de algoritmos matemáticos y funciones que proveen de varias y eficientes rutinas numéricas para integración y optimizan en base de procesamiento de datos. Llamado de librería Sentencia para el llamado de librería

Editor	from scipy import stats

Cuadro 3.37: Llamado de librería Scipy. Fuente: Elaboración propia.

Prueba de Normalidad

Con esta prueba realizaremos un análisis de Ji cuadrada para revisar la normalidad de la curva en la distribución de los datos.

Editor	stats.mstats.normaltest(df3.b)
Terminal	NormaltestResult(statistic=3.802054, pvalue=0.14941)

Cuadro 3.38: Prueba de Normalidad. Fuente: Elaboración propia.

Seaborn

Seaborn es una librería de visualización basada en Matplotlib, el cual tiene como objetivo entregar a usuario de alto nivel una interfaz de gráfico estadísticos atractivos. Usualmente se usa para gráficos de nivel avanzado [Waskom, 2017].

3.3. Modelos de simulación de un portafolio

En el presente trabajo se tiene como objetivo la simulación de una variable dependiente con una relación logística de variables independientes con ello se busca conocer los criterios de crecimiento y los valores esperados. La metodología estadística que se ocupará y unirá con la metodología de minería de datos es la siguiente:

1. Conocer o definir la ecuación de probabilidad de incumplimiento.

   Sea Y_i una variable estocástica definido por

   $$Y = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \tag{3.3.1}$$

   Donde X_{ji} es una serie de variables aleatorias con $F(X)$ no tienden o no a comportase como $N(\mu, \sigma^2)$.

2. De las variables que intervienen en la probabilidad de incumplimiento, realizar transformaciones normalización

   Para toda X_{ji}

   $$Z_{ji} = T(X_{ij}) = \ln\left(\frac{x_{ij} - \mu}{\sigma^2}\right) \tag{3.3.2}$$

3. Calcular mediante Análisis de Componentes Principales los valores propios y la matriz de varianza-covarianza

   Por el Análisis de Componentes Principales

$$Q_{ji} = Z_{ji} A_{ji} \tag{3.3.3}$$

   Donde A_{ji} es una transformación ortogonal.

4. Analizar la función de distribución de los componentes principales

   Al tener que

$$Z_{ji} \sim \mathsf{N}(0,1) \tag{3.3.4}$$

   Entonces,

$$Q_{ji} \sim \mathsf{N}(0,1) \tag{3.3.5}$$

5. Con la función de distribución de los componentes principales, realizar simulación Monte Carlo. En caso de tener un análisis de casos, modificar la función de distribución de las variables deseadas

   Al utilizar el método Box-Muller

$$X = \sqrt{-2\ln\xi_1}\cos(2\pi\xi_2) \tag{3.3.6}$$

$$Y = \sqrt{-2\ln\xi_1}\sin(2\pi\xi_2) \tag{3.3.7}$$

   Se realizan las simulaciones

$$M_{ji} \sim \mathsf{N}(0,1) \tag{3.3.8}$$

donde $h \geq i$ y $\rho = 0$

6. Multiplicar la matriz de valores propios a la matriz de simulaciones Monte Carlo

   Al aplicar la matriz de valores propios

$$R_{ji} = A_{ji} M_{ji} \tag{3.3.9}$$

7. Aplicar las transformaciones inversas de des normalización

   Para R_{jh} existe,

$$S_{jh} = U(R_{jh}) = T^{-1}(R_{jh}) = e^{(x_{ij}+\mu)\sigma^2} \tag{3.3.10}$$

8. Comparar las funciones de distribución individuales de las variables independientes simuladas y real

   Con Anderson-Darling podemos comprobar

$$\mathbb{A}^2(S_{jh}, X_{ji}) > 0.05 \tag{3.3.11}$$

9. Comparar la función de distribución de la variable dependiente simulada y real

   Si tenemos S_{ji}, podemos aplicar en la fórmula de regresión

$$\mathbb{A}^2(P_{jh}, Y_{ji}) > 0.05 \tag{3.3.12}$$

   Bajo esta propuesta de simulación se da la búsqueda de tener una herramienta de análisis, descripción, simulación y comparación de un comportamiento de una variable dependiente.

Bajo los capítulos de recolección de metodologías y búsqueda de escenarios que requieran

un modelado multivariado, el resto del trabajo se centrará en la aplicación de los recursos propuestos.

Capítulo 4

PRUEBAS Y EVALUACIONES EN UN PORTAFOLIO DE CRÉDITO BANCARIO

En este capítulo realizaremos las pruebas y análisis de la unión de las herramientas; metodología de implantación: CRISP-DM, metodología estadística: simulación Monte Carlo por componentes principales y la herramienta computacional: Python.

4.1. Comprensión del negocio

Una entidad bancaria, de acuerdo con el banco de México, es un intermediario facultado para realizar la captación de fondos del público para la posterior colocación de créditos o Inversiones [BANXICO, 2005]. Una de sus principales funciones es mediar el riesgo de contra parte que se da ante los clientes de captación y de colocación. El riesgo de crédito existe cuando el contratante de crédito es incapaz de cumplir sus obligaciones financieras [Delgado, 2010].

Numerosos eventos en la historia del sector bancario son relevantes para la administración del riesgo de crédito, entre las más importantes a nivel internacional podemos nombrar la gran

depresión de EU (1929), Black Monday (1987) y Crisis subprime (2008). El antecedente regulatorio más relevante se encuentra en 1988 con el *"Basle committee on banking supervision"* conocido comúnmente como Basilea I [BIS, 1988]. En el postulan mejoras y recomendaciones hacia las entidades financieras y la supervisión acerca de tener el capital adecuado para hacer frente a las obligaciones en caso de perdidas. Estas recomendaciones fueron emitidas por el comité de supervisión bancaria en cual conocido como BIS por sus siglas en inglés (Bank for International Settlements) con sede en Basilea, Suiza. En el acuerdo, propone a las entidades regulatorias bancarias nacionales, la supervisión de los bancos en cuestión del Capital. Se define el capital regulatorio como el mínimo capital para hacer frente a los riesgos a través provisiones, ante la posibilidad de perdidas futuras en cinco niveles de riesgos al incumplimiento total bajo el incumplimiento de 0, 10, 20, 50 y 100 %. Sugiere el ratio de capital debe está en niveles de 8 % en sus activos ponderados por riesgo. Y propone tener 2 Tiers (Niveles) de supervisión. Tier 1 medirá el capital social y las reservas publicadas entre el valor de sus activos ponderados, por su parte, Tier 2 medirá las reservas no publicadas, revalorizaciones, etc.

En los hallazgos posteriores a Basilea I, se tienen:

- Las mediciones internas, son más sensibles al riesgo.
- En cinco años, todos los miembros del comité Basilea, han cumplido las reglas de capital
- Mas de 100 naciones no miembros, también han adoptado las recomendaciones.

Propuestas postuladas a partir de los hallazgos:

- Incentivar a los bancos que mejoren las mediciones de capital.
- Mejorar los beneficios de diversificación de portafolios de activos.
- Ante el alcance mundial del acuerdo, se propone realizar un nuevo marco de referencia con mejoras propuestas.

En base a los resultados obtenidos, en 1999 BIS solicita propuestas para el nuevo marco de referencia que tendrá la finalidad de mejorar la suficiencia de capital. En 2001, se publica en 2001 con una revisión en 2004 del texto [BIS, 2001].
En él se proponen tres Pilares fundamentales:

- Requerimientos mínimos de Capital: para los requerimientos de capital por riesgo de crédito, operacional y de mercado, las definiciones son muy similares, en el caso del riesgo de crédito se agregan dos métodos de cálculo de reservas; método estándar e interno IRB (Básico y Avanzado). Se definen y proponen como métodos de medición la probabilidad de incumplimiento (PD, Probability of Default), pedida dado el incumplimiento (LGD, Loss Given Default) y exposición en el momento del incumplimiento (EAD, Exposure At Default). Se diferencia el método básico como el propuesto por el regulador nacional y el avanzado la metodología interna bajo restricciones [BIS, 2004].

- Proceso de examen supervisor: Propone que la supervisión del capital este bajo las reguladoras nacionales, y en base al riesgo permitido por ellas, se auditen, regulen y postulen las reglas mínimas de capital de las entidades bancarias.

- Disciplina del Mercado: el Comité introduce recomendaciones sobre la divulgación de información, buscando que el público tenga una participación en las decisiones de captación y colocación.

Con estos precedentes en la gestión y evaluación del riesgo, se abre la puerta a la administración del riesgo de crédito por medio de las probabilidades de incumplimiento. En Basilea III se desarrolla posterior a la crisis de hipotecas subprime, en 2011 [BIS, 2001]. Nace como una serie de iniciativas para la mejora de los precedentes de la crisis de 2008 como apalancamiento y liquidez. No sustituye en gran parte a los acuerdos de Basilea I y II. Ante estas evaluaciones los bancos mundiales o nacionales han adoptado por regulación (IFRS9) o recomendación (Basilea III) la forma de medición de los créditos en base al riesgo de contraparte por probabilidad de incumplimiento. La probabilidad de incumplimiento se ha vuelto para el sector financiero impor-

tante desde el cambio en las medidas para IFRS9 de contabilidad financiera, con probabilidad que al aprovisionar en el estado de resultados la reservas con la perdida incurrida a la perdida esperada, basada en la probabilidad de incumplimiento [IFRS, 2015].
Entre las necesidades básicas para una correcta gestión del portafolio de crédito [Santander, 2017, Banorte, 2015]:

- Gestionar la administración efectiva del portafolio.
- Cumplir en términos regulatorios obligaciones del banco (reservas).
- Mantener el marco regulatorio normativo.
- Proveer a la Dirección información para la toma de decisiones.

El Banco de México define la probabilidad de incumplimiento como una medida en que un cliente de colocación se encuentre en probabilidad de incumplir con sus obligaciones de pago. En México, las entidades bancarias utilizan métodos de para el cálculo de la probabilidad de incumplimiento:

- Escalas de calificadoras externas: Banorte [2015]
- Probabilidades de transición: HSBC [2013]
- Calificaciones Internas: Santander [2017], Bancomer [2017]

Por ejemplo, en el grupo BBVA [BBVA, 2014] tiene como metodología el uso de scores de riesgos que evalúan el comportamiento de los clientes retail por medio de evaluaciones en distintos puntos de la vida de solicitante:

- Comportamental: Realizan una evaluación posterior a la adquisición del crédito otorgado y a partir del comportamiento del crédito emitir una calificación.
- Reactivos: Son herramientas que apoyan a generar una aprobación o rechazo de una solicitud de crédito emitida por un cliente interno o externo.

- Proactivos: Realizan una evaluación similar a los scores comportamentales, con la diferencia de ser más usados para emisiones de calificación activa a los clientes con créditos previamente adquiridos.

- Buró: Con una similitud muy grande con los comportamentales, estas herramientas se apoyan de la información interna y externa para determinar el riesgo de crédito de los clientes.

Además, existen otros estudios de esta institución en la búsqueda de probabilidad de incumplimiento bajo una técnica predictivas de Bruce [2015] y Zhao [2015]. En este punto, es de interés para el sector bancario tener una medida de probabilidad de incumplimiento, los factores que la modifican y un estudio de hacia dónde puede crecer el portafolio son intereses naturales de la banca nacional e internacional. PO ello la simulación atiende en términos de decisiones de negocio una herramienta útil para conocer el riego adquirido por una institución bancaria en un plan de crecimiento, sin los gastos generados por pruebas pilotos o casos de negocio con fuertes supuestos que no se puedan cumplir. Ante esto una solución de medición del riesgo prospectivo es la unión de una medida de riesgo como la perdida esperada y la simulación estocástica de casos que se correlacionan con la medida de incumplimiento. En la presente prueba se realizará una de las posibles técnicas de obtención de la perdida esperada y la implementación de esta para un portafolio bancario real.

En la toma de decisiones, las políticas y consideraciones a tomar son multifactoriales, donde se integran posibles riesgos y oportunidades como lo pueden ser, la mejora de margen financiero, la reducción de costos, la mejora de reservas regulatorias. En este caso, para este documento tomaremos como métrica máxima de integración de crédito a los clientes que no superen al 95 % de la pérdida esperada. Para los casos que lo superen, se encuentran fuera de rango del apetito del riesgo del negocio y los inversionistas.

4.2. Comprensión de los datos

Para este ejercicio utilizaremos una base de datos del banco Online más grande en Estados Unidos el cual funciona con un sistema de préstamos P2P (Peer tú Peer) sistema conocido en México como préstamo entre particulares, el cual ofrece su información de manera pública [Lending Club, 2018]. Para el banco de estudio, como para cualquier banco; la medición, predicción, prevención y mitigación del riesgo de crédito es una de las tareas más importantes, donde se suelen tener áreas específicas y de apoyo que ejercen el día a día acciones de mitigación del riesgo, entre las más usuales son Riesgos, Finanzas, Normatividad, Jurídico y Sistemas, los cuales en caminos separados y en conjunto, buscan minimizar el incumplimiento para reportar número más sanos de la administración del capital.

Con el fin de ilustrar manejaremos la base que se reporta para el año 2014 en Lending Club. Esta base contiene los datos de los créditos y acreditados del año en revisión, incluido el estatus actual del préstamo. En esta base contiene 145 variables, entre los segmentos más relevantes de la base se encuentran:

- Datos financieros de originación del crédito (Monto otorgado, plazo, tasa de interés, estatus del crédito, propósito del crédito)
- Datos demográficos del prestatario (Rango de edad, Tipo de Vivienda, Dirección)
- Datos financieros del prestatario (Ingreso, Empleo, Verificación de Ingreso)
- Datos financieros de apalancamiento (Capacidad de pago, número de créditos finalizados, número de créditos totales)
- Datos de incumplimiento (Mora en los últimos dos años, estatus de pago actual, recuperaciones)

La base es elegida gracias al gran número de variables numéricas continuas que requiere la técnica de simulación seleccionada. Sin embargo, no todas las variables tienen información, ya que el banco en análisis actualiza su layout, pero no hace evaluación perspectivas retrospectiva

de los campos. Otra característica de elección es que la base se encuentra libre para descarga, con lo que para temas académicos es muy útil, pues tenemos las operaciones de un banco real. Podemos observar que tenemos una base con 235 mil cuentas y 145 variables, de las cuales 108 son numéricas y 37 alfanuméricas o cadena. El diccionario de datos y su interpretación de esta la incluimos en el 4.6.2 el cual tiene una traducción e interpretación al español y lenguaje académico. En el caso de las variables alfanuméricas, a pesar de no poderse integrar en el modelo de simulación, otorgan información importante para el ejercicio pues con ella haremos la limpieza de base de datos.

Iniciaremos con la variable Purpose, la cual describe el propósito para el uso del crédito. Como se tiene como objetivo, el análisis de las tarjetas de crédito, revisaremos como está conformada esta variable.

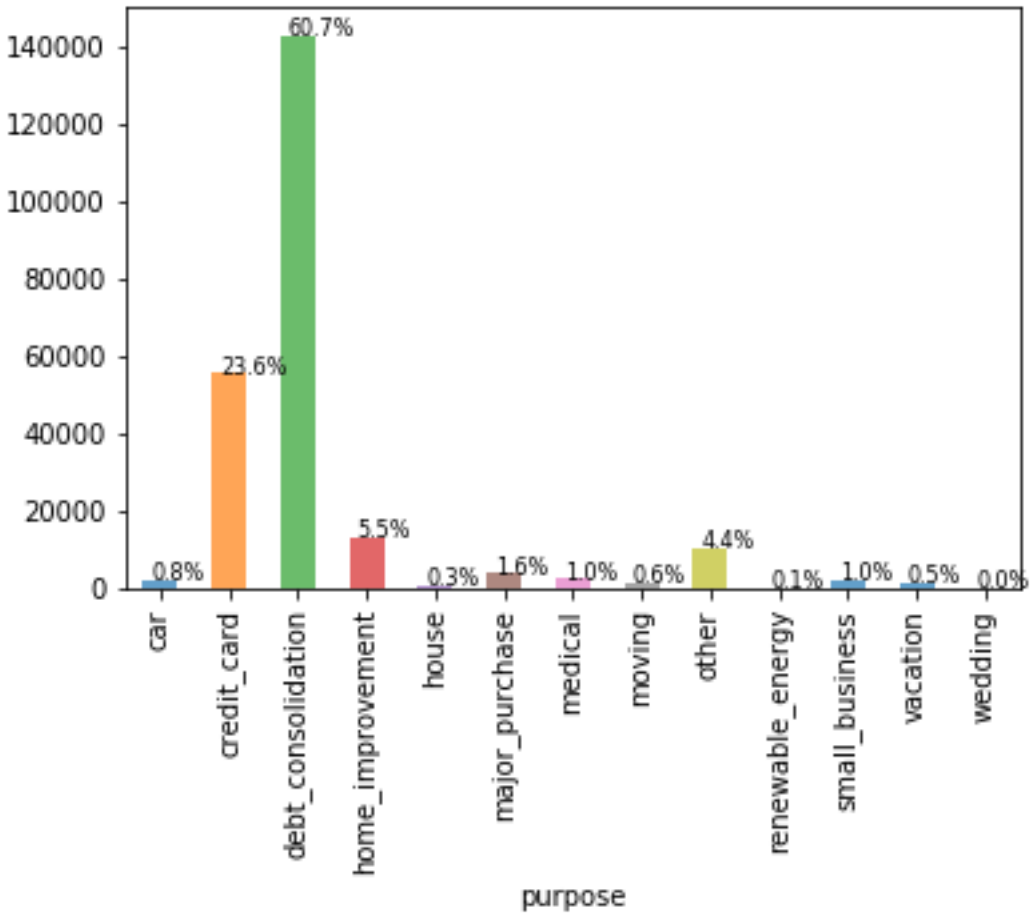


Figura 4.2.1: Gráfico de porcentaje de cartera por tipo de crédito. Fuente: Elaboración propia.

La figura 4.2.1 nos indica que, para la institución bancaria, tenemos el 60.7 % de los créditos otorgados son para consolidación de deudas, el 23.6 % de créditos con el propósito de tarjetas de crédito (credit_card), 5.5 % para mejoras en casa y el restante para otros propósitos como compra de casa, vacaciones, bodas, servicio médico entre otros. Derivado a que realizaremos el ejercicio para un portafolio de crédito nos quedaremos con 55 mil registros de tarjetas revolventes.

En el caso del estatus del crédito (loan_estatus), nos ayuda a discriminar los créditos activos en el portafolio.

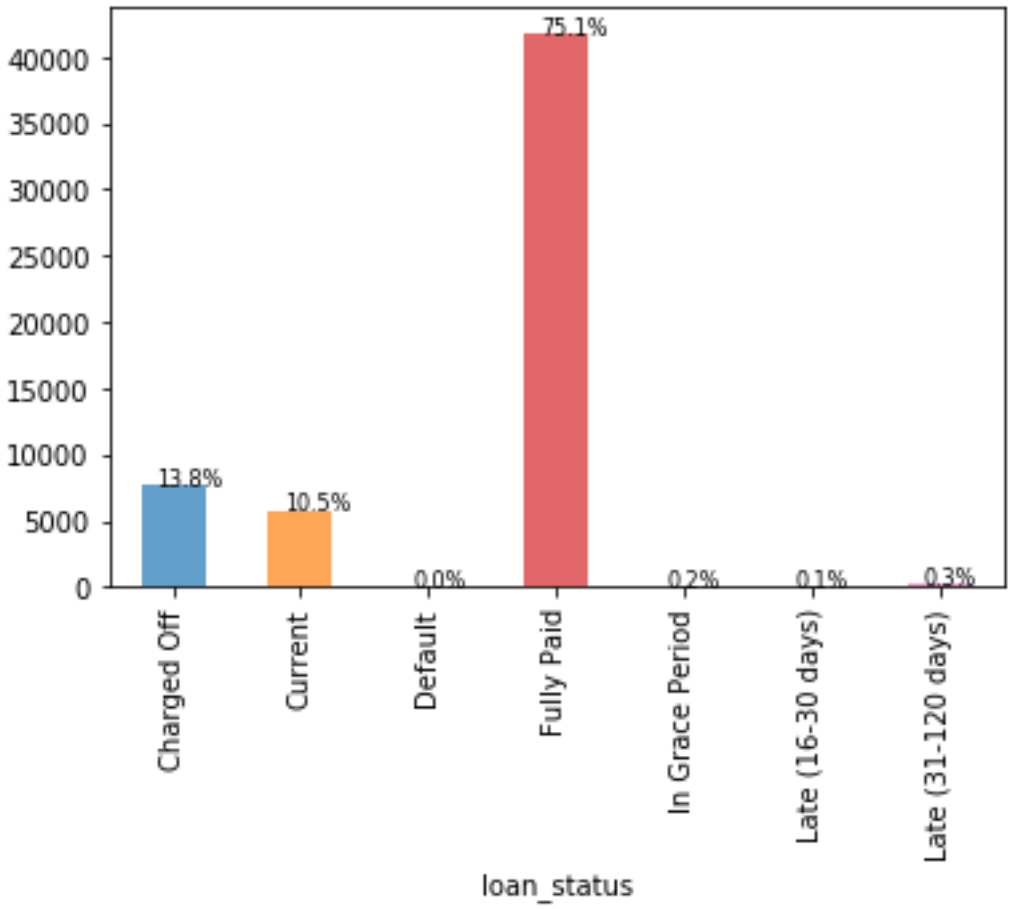


Figura 4.2.2: Gráfico de porcentaje de cartera credit card por situación crediticia. Fuente: Elaboración propia.

Para el ejercicio requeriremos créditos que a la fecha de la evaluación se encuentren transaccionando. en ese sentido se observa que más del 75 % de los contratos han pagado su crédito

(Fully Paid), por lo tanto, son créditos que nos será útiles dado que su probabilidad de que deje de paga es cero. También se nota que el 13 % se encuentran en situaciones irrecuperables (Charged Off, Default) de la base ya se encuentra en impago total, el cual su probabilidad seria uno por lo que no habrá nada que medir y descartaremos para el ejercicio. Nos quedaremos con los 6,046 créditos que no se encuentran en estos estatus.

La siguiente variable importante en la selección del universo es el incumplimiento de los dos últimos años (delinq_2yrs), dado que el ejercicio se encuentra en fechas posteriores a 2016, esta variable no nos ayudara a discriminar entre clientes con alta o baja probabilidad de incumplimiento.

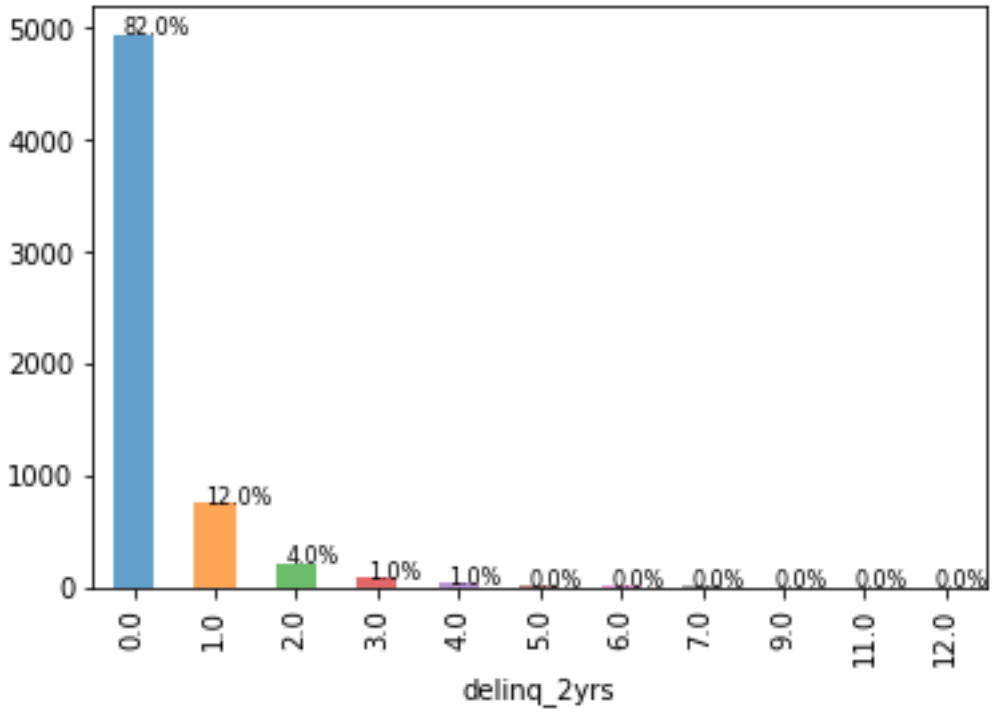


Figura 4.2.3: Gráfico de porcentaje de cartera credit card por incumplimiento en dos años. Fuente: Elaboración propia.

Donde nos quedaremos con el 81 % de la población que antes de la revisión no tenía incumplimientos, y el restante 19 % se encuentra con algún estatus posterior en mora. Por lo que podemos solo tener a los contratos que no han incumplido después del periodo de

evaluación.

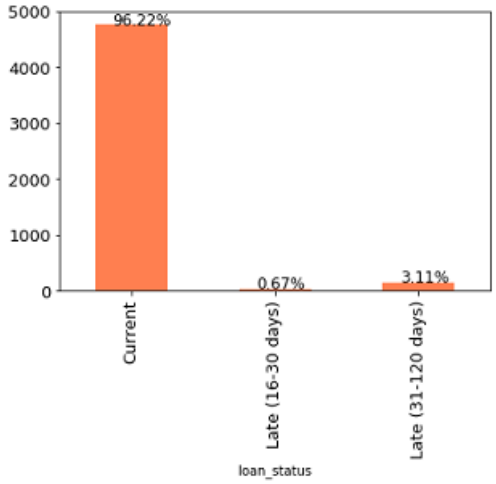


Figura 4.2.4: Gráfico de porcentaje de cartera credit card por estatus actual de impago. Fuente: Elaboración propia.

En la variable estatus de crédito (loan_status) se observa que 96 % de las cuentas se encuentran en un estatus al corriente, el resto ya presenta alguna probabilidad de incumplimiento mayor. Con ello podemos decir que la cartera tiene un 4 % de incumplimiento, con esta base de 4,949 créditos será la muestra para el ejercicio posterior.

4.3. Preparación de los datos

De manera inicial, solo nos quedaremos con los registros de Tarjeta de Crédito, con estatus activos y que no ha tenido incumplimiento total los últimos dos años. Podemos observar en términos técnicos que la base queda como la figura 4.2.4.

4.3.1. Limpieza de los datos

De las 145 variables de datos nos quedaremos con las variables numéricas, que nos detonen monto y nos sean consecuenciales de categorías, 108 variables.

- Se realizó un análisis para encontrar las variables que no contienen información, por lo

cual descartaremos todas las variables que no cumplan la condición de estar informadas, por ello nos quedaremos con 59 variables informadas.

- Posteriormente se realizó un análisis de correlaciones donde se encontró que existe un número considerable de variables con una correlación fuerte. En el gráfico 4.3.1 podemos observar que existen variables con alta correlación.

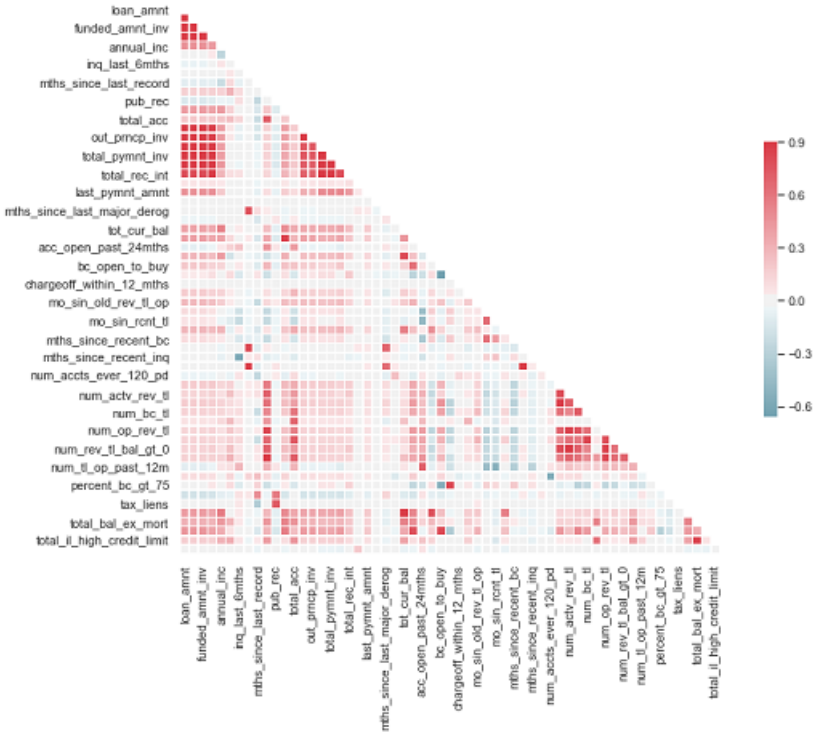


Figura 4.3.1: Gráfico de correlación de las variables. Fuente: Elaboración propia.

En el gráfico 4.3.1 podemos observar de una manera gráfica la correlación de las variables, donde si la correlación es positiva y alta, se tornará de color más fuerte y rojo, por el contrario, si la correlación es negativa, se tornará fuerte y azul. Entre las primeras variables con alta correlación y bajo un análisis experto del escritor del texto se decide que, de las siguientes nueve variables, se dejará solo la variable funded_amnt la cual el resto son partes o funciones de Monto del crédito otorgado:

- funded_amnt: Monto de crédito otorgado en algún momento del tiempo

- funded_amnt_inv : Monto de crédito comprometido con los inversores en algún momento del tiempo
- Installment : Pago del crédito
- out_prncp : Monto de capital remanente del monto del crédito otorgado
- out_prncp_inv : Monto de capital remanente del monto del crédito comprometido con los inversores
- total_pymnt : Pago total del cliente
- total_pymnt_inv : Pago total comprometido con los inversores
- total_rec_prncp : Monto de Capital Amortizado
- total_rec_int : Monto de los intereses Amortizados

Otra serie de variables con alta correlación es:

- open_acc: Número de cuentas abiertas
- num_actv_rev_tl: Número de activos revolventes
- num_bc_sats: Número de tarjetas bancarias satisfactorias
- num_bc_tl: Número de tarjetas bancarias
- num_il_tl: Número de préstamos
- num_op_rev_tl: Número de cuentas revolventes abiertas
- num_rev_accts: Número de cuentas revolventes
- num_rev_tl_bal_gt_0: Número de cuentas revolventes con saldo 0

En esta serie de variables nos quedaremos con open_acc la cual se puede interpretar como el subconjunto más grade del resto de las variables. Las siguientes variables también son eliminadas por una alta correlación:

Permanece	Descripción	Descartada	Descripción
Mths_since_ recent_revol_ delinq	Mesesdesdeel incumplimiento masreciente revolvente	Mths_since_ last_delinq Mths_since_ recent_bc Mths_since_ recent_bc_dlq Mths_since_ recent_inq	Mesesdesdeel incumplimiento reciente Mesesdesdeultima originación revelvente Meses desde el incumplimiento masrecientetarjetasbancarias Mesesdesdeultimaconsulta
Pub_rec	Número de registros publicos malos	Tax_liens Pub_rec_ bankruptcies	Número de embargos fiscales Número quiebra enregristros publicos
Total_bal_ ex_mort	Saldototal sinHipoteca	Total_il_high_ credit_limit	Linea de crédito totaldetarjetasyconsumos
Bc_util	Relación entre el saldo y limite de crédito	Percent_bc_gt_75 Bc_open_to_buy Total_bc_limit	Número de tarjetas con Lineadisponible de la tarjeta de crédito Lineadecréditototal
Tot_cur_bal	Saldoactual entodas sus cuentas	Tot_hi_cred_lim Avg_cur_bal	Linea de créditototal Saldo promedio de todas sus cuentas

Cuadro 4.1: Cuadro de descarte de variables por alta correlación. Fuente: Elaboración propia.

En el cuadro 4.1 podemos observar las variables con alta correlación mostradas en gráfico 4.3.1. De lado izquierdo se muestran las variables no descartadas o primarias, y en el lado derecho podremos observar las variables descartadas por la alta correlación. En la variable de meses desde el incumplimiento más reciente (mths_since_recent_revol_delinq) podemos resumir las cuatro variables subsecuentes. En el caso del número de registros malos, vemos que es una suma de las variables de la columna derecha. Por su parte, se observa que total_bal_ex_mort, bc_util y tot_cur_bal son variables sintéticas de las variables descartadas. Posterior a esta limpieza por correlación, buscaremos las variables que son parte del conocimiento de negocio como variables consecuenciales son las comisiones por pago tardío (total_rec_late_fee) y re-

cuperaciones (collections_12_mths_ex_med) por ejemplo.

Posterior a ello, usaremos una metodología para el cálculo de la pérdida esperada, teniendo como técnica predictora para el incumplimiento una regresión logística, con ello podemos determinar el valor de estas variables, el cual, además de crear perfiles de riesgo y calcular la probabilidad de default, apoyará en el descarte de variables no significativas para el ejercicio.

OLS Regression Results

Dep. Variable:	target	R-squared:	0.053
Model:	OLS	Adj. R-squared:	0.044
Method:	Least Squares	F-statistic:	5.884
Date:	Tue, 04 Jun 2019	Prob (F-statistic):	2.32e-11
Time:	09:55:50	Log-Likelihood:	421.38
No. Observations:	1484	AIC:	-814.8
Df Residuals:	1470	BIC:	-740.5
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
mths_since_last_record	0.0007	0.000	1.847	0.065	-4.53e-05	0.002
mo_sin_rcnt_tl	0.0007	0.000	1.466	0.143	-0.000	0.002
tax_liens	0.0500	0.032	1.565	0.118	-0.013	0.113
mths_since_recent_revol_delinq	0.0001	0.000	0.544	0.586	-0.000	0.001
pub_rec	-0.0609	0.033	-1.851	0.064	-0.125	0.004
num_actv_bc_tl	0.0039	0.002	1.735	0.083	-0.001	0.008
total_rev_hi_lim	-3.394e-07	4.51e-07	-0.753	0.452	-1.22e-06	5.45e-07
tot_cur_bal	-2.312e-08	4.35e-08	-0.532	0.595	-1.08e-07	6.22e-08
annual_inc	2.437e-07	1.36e-07	1.793	0.073	-2.29e-08	5.1e-07
acc_open_past_24mths	0.0061	0.002	2.879	0.004	0.002	0.010
funded_amnt	-1.062e-06	6.71e-07	-1.583	0.114	-2.38e-06	2.54e-07
revol_bal	9.496e-08	5.33e-07	0.178	0.859	-9.5e-07	1.14e-06
total_bc_limit	-8.721e-08	4.17e-07	-0.209	0.834	-9.05e-07	7.3e-07
num_il_tl	0.0008	0.001	1.207	0.228	-0.001	0.002

Omnibus:	1444.110	Durbin-Watson:	2.042
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37842.491
Skew:	4.915	Prob(JB):	0.00
Kurtosis:	25.702	Cond. No.	2.23e+06

Cuadro 4.2: Cuadro de variables con regresión. Fuente: Elaboración propia.

Utilizaremos una técnica de descarte de variables conocida como Stepwise, que combina la introducción de una variable independiente a la regresión y se selecciona o elimina alguna variable de acuerdo a parametro de la P. Con este procedimineto podemos descartar el mayor número

de variables, y solo nos quedaremos con las variables 'num_actv_bc_tl', 'total_rev_hi_lim', 'tot_cur_bal', 'annual_inc', 'acc_open_past_24m' las cuales tienen una significancia menor a .1. Al realizar la regresión con estos campos se comprueba su significancia.

OLS Regression Results

Dep. Variable:	target	R-squared:	0.046
Model:	OLS	Adj. R-squared:	0.045
Method:	Least Squares	F-statistic:	42.57
Date:	Tue, 04 Jun 2019	Prob (F-statistic):	5.61e-43
Time:	09:45:21	Log-Likelihood:	1096.6
No. Observations:	4454	AIC:	-2183.
Df Residuals:	4449	BIC:	-2151.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
num_actv_bc_tl	0.0032	0.001	2.807	0.005	0.001	0.005
total_rev_hi_lim	-1.514e-07	1.05e-07	-1.438	0.151	-3.58e-07	5.51e-08
tot_cur_bal	-8.847e-08	2.48e-08	-3.563	0.000	-1.37e-07	-3.98e-08
annual_inc	2.02e-07	6.91e-08	2.926	0.003	6.67e-08	3.37e-07
acc_open_past_24mths	0.0065	0.001	6.475	0.000	0.005	0.008

Omnibus:	4148.121	Durbin-Watson:	2.013
Prob(Omnibus):	0.000	Jarque-Bera (JB):	99554.303
Skew:	4.766	Prob(JB):	0.00
Kurtosis:	24.109	Cond. No.	1.06e+05

Cuadro 4.3: Cuadro de variables con regresión. Fuente: Elaboración propia.

A partir de estas variables y la significancia, podemos aceptar en base a las pruebas estadísticas y de negocio, que serán las variables que describen el comportamiento del incumplimiento. Utilizaremos el espacio bajo la curva ROC, como estadístico de eficiencia para el ejercicio.

		Estimado	
		0	1
Real	0	2308	2652
	1	50	109

Cuadro 4.4: Matriz de confusión de la regresión. Fuente: Elaboración propia.

En la matriz de confusión se observan que se pueden pronosticar el 70 % de las caídas a incumplimiento.

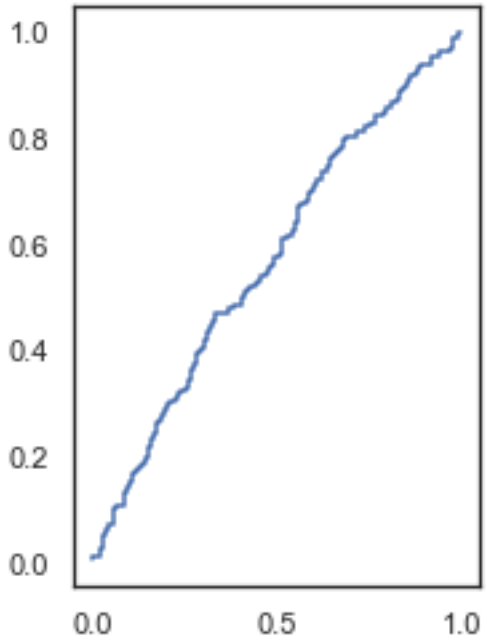


Figura 4.3.2: Curva ROC de la base de entrenamiento. Fuente: Elaboración propia.

A partir de esta Regresión tenemos que el área bajo la curva ROC es de 60 %, que bajo los criterios prudenciales en la banca de crédito normalmente muestra como satisfactorio.

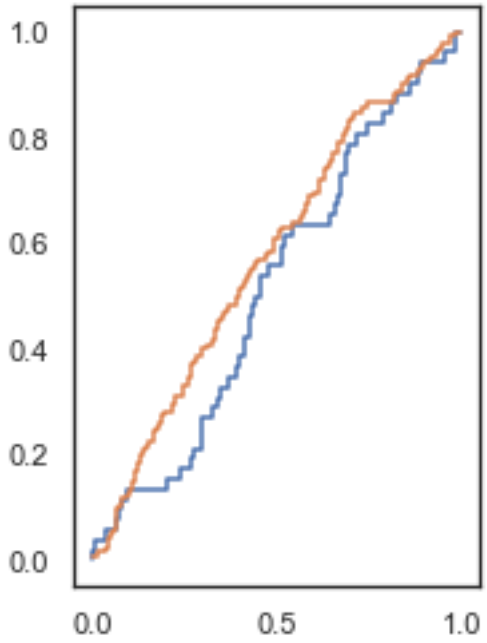


Figura 4.3.3: Curva ROC de la base de entrenamiento contra la de prueba. Fuente: Elaboración propia.

La comparación del entrenamiento contra la prueba, podemos observar que son muy similares, por lo que podemos concluir que la regresión es estable en muestras independientes. Con ello tenemos nuestras variables finales para el ejercicio de simulación.

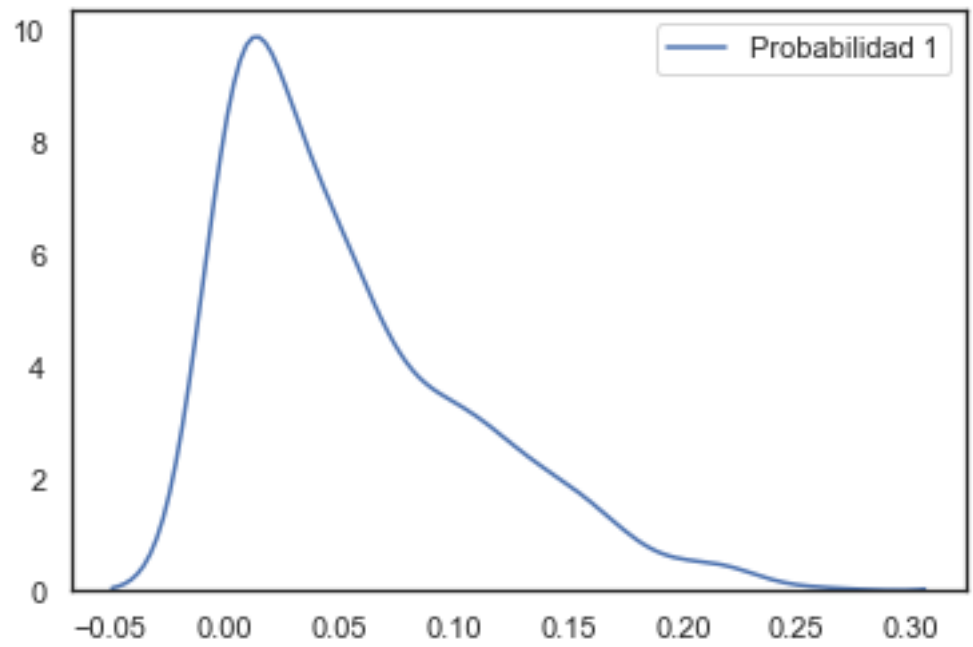


Figura 4.3.4: Gráfico de distribución de la probabilidad de no caer en impago. Fuente: Elaboración propia.

Teniendo el modelo del cuadro 4.3, tomamos una muestra independiente y aplicamos la ecuación del modelo en con los datos. Se observa que en la gráfica de la predicción nos dice que la mayor parte del portafolio no tienen una probabilidad de default alta. Dentro de la metodología propuesta, hemos encontrado en este punto la función de los modelos lineales generalizados a la que tenemos interés de simular, en base a la simulación correlacionada de los datos.

4.3.2. Análisis de variables finales

Annual_inc

La variable en el diccionario de datos se informa como el ingreso auto informado por el cliente de colocación al momento del otorgamiento del crédito. Esta variable, es de alta relevancia para saber si el cliente colocación tiene los ingresos suficientes para poder hacer

cargo de sus obligaciones.

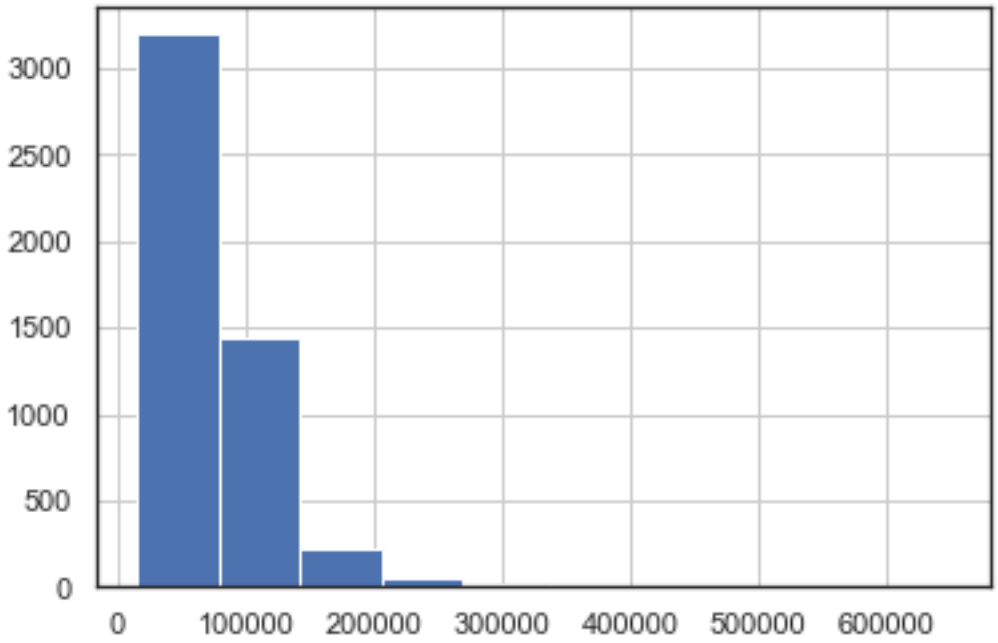


Figura 4.3.5: Distribución de la variable de ingreso. Fuente: Elaboración propia.

En análisis descriptivo de la variable annual_inc podemos observar en histograma, que la variable no tiene una distribución visual normal.

Medida	Valor
count	4,949
mean	77,010
std	46,457
min	16,000
Q1	50,000
Q2	66,000
Q3	90,000
max	650,000

Cuadro 4.5: Cuadro de estadísticas descriptivas del ingreso. Fuente: Elaboración propia.

Con 4949 datos tenemos una media de 77010 y una desviación estándar de 46457. A pesar de que los cuartiles se mantienen distribuidos con cierta distancia equitativa, el valor máximo es un outlier sobresaliente.

NormaltestResult(statistic=4348.5213652055327, pvalue=0.0)

Figura 4.3.6: Salida prueba de normalidad para el ingreso. Fuente: Elaboración propia.

En la prueba de normalidad, se observa que no hay evidencia de poder aceptar que esta variable se describe como una variable normal.

num_actv_bc_tl

La variable tiene como significado el número de tarjetas de crédito que tiene en este momento el cliente colocación. Su relevancia en negocio radica por el mix de deuda que tiene el cliente y los sobre costos que genera anualmente una tarjeta.

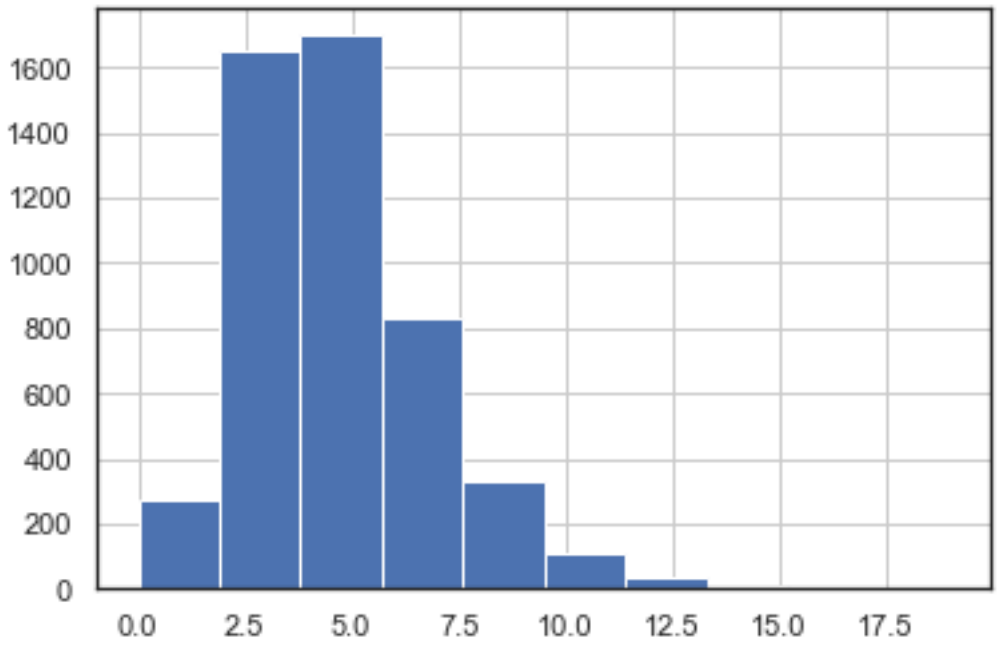


Figura 4.3.7: Distribución de la variable número de tarjetas que tiene el cliente. Fuente: Elaboración propia.

Dentro de la figura 4.3.7 podemos observar que la variable tiene una distribución normal aparente, además de que los datos se encuentran concentrados entre los números 2.5 y 7.5.

Medida	Valor
count	4,949
mean	4
std	2
min	0
Q1	3
Q2	4
Q3	6
max	19

Cuadro 4.6: Cuadro de estadísticas descriptivas de número de tarjetas que tiene el cliente. Fuente: Elaboración propia.

En el cuadro 4.6 se describe que, en promedio, los clientes de la entidad bancaria tienen en promedio 4.4 tarjetas de crédito, al ser la mediana 4.

NormaltestResult(statistic=948.50479792821068, pvalue=1.083428112889883e-206)

Figura 4.3.8: Salida prueba de normalidad para el número de tarjetas que tiene el cliente. Fuente: Elaboración propia.

En la prueba de normalidad no arroja que hay pruebas suficientes para decir que la variable num_actv_bc_tl se distribuye con una función normal.

total_rev_hi_lim

La variable describe la fórmula: $Limite\,de\,crédito\,total/Limite\,de\,crédito\,revolvente$. En límite de crédito total, corresponde al monto máximo que el cliente tiene disponible en la entidad de crédito. En el caso de los casos no revolventes (Hipotecario, consolidación de deuda, adquisición de bienes duraderos, créditos al consumo, autos, etc.), se traduce como el monto otorgado a ser amortizado. En el caso de los casos revolventes, se traduce como la línea de crédito total de las tarjetas. Esta variable mide la proporción inversa que tiene el límite de crédito revolventes contra el limite total. Y nos puede ofrecer qué tipo de mix de colocación tiene el cliente.

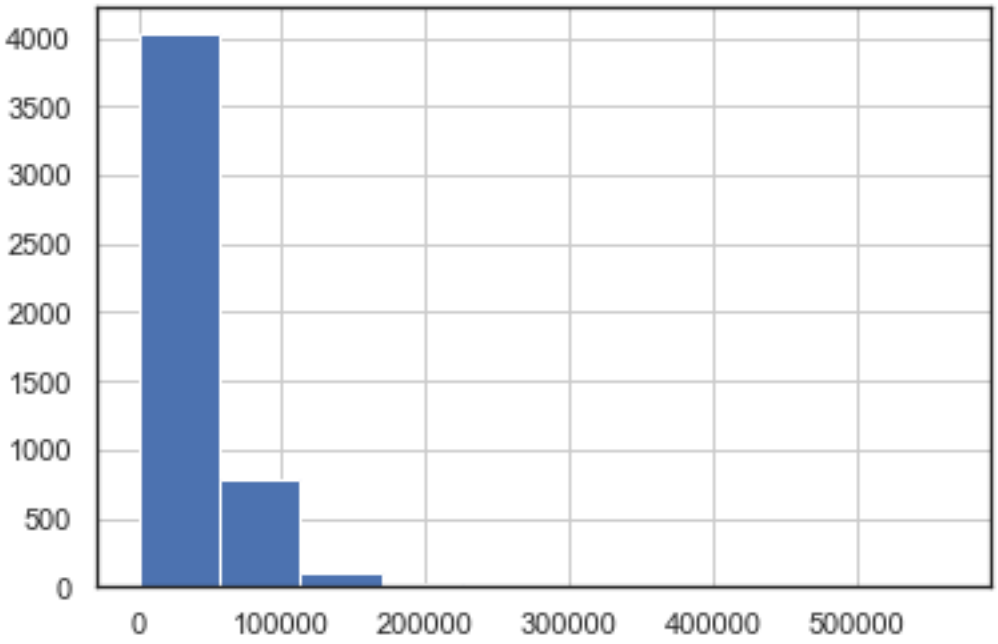


Figura 4.3.9: Distribución de la variable porcentaje de uso del crédito disponible. Fuente: Elaboración propia.

la figura 4.3.9 nos muestra que la mayor parte de los registros se tienden a encontrarse entre un monto de 0 a 100000.

Medida	Valor
count	4,949
mean	39,864
std	32,587
min	0
Q1	20,600
Q2	31,700
Q3	49,400
max	564,600

Cuadro 4.7: Cuadro de estadísticas descriptivas de porcentaje de uso. Fuente: Elaboración propia.

También se relacionan una dispersión de datos similar a la media y muy cercanas al primer cuartil. El número máximo se nota alejado de la distribución.

NormaltestResult(statistic=948.50479792821068, pvalue=1.083428112889883e-206)

Figura 4.3.10: Salida prueba de normalidad para el porcentaje de uso. Fuente: Elaboración propia.

En la prueba de normalidad observamos que no tenemos evidencia para comentar que la variable se distribuye bajo una normal.

tot_cur_bal

La descripción de la variable se basa en la fórmula: $totcurbal = \sum_{i=1}^{n} Saldo\ total$. Esta variable describirá el comportamiento del saldo de todas las cuentas que tiene el cliente. Es importarte esta variable para conocer el nivel de endeudamiento del cliente neto.

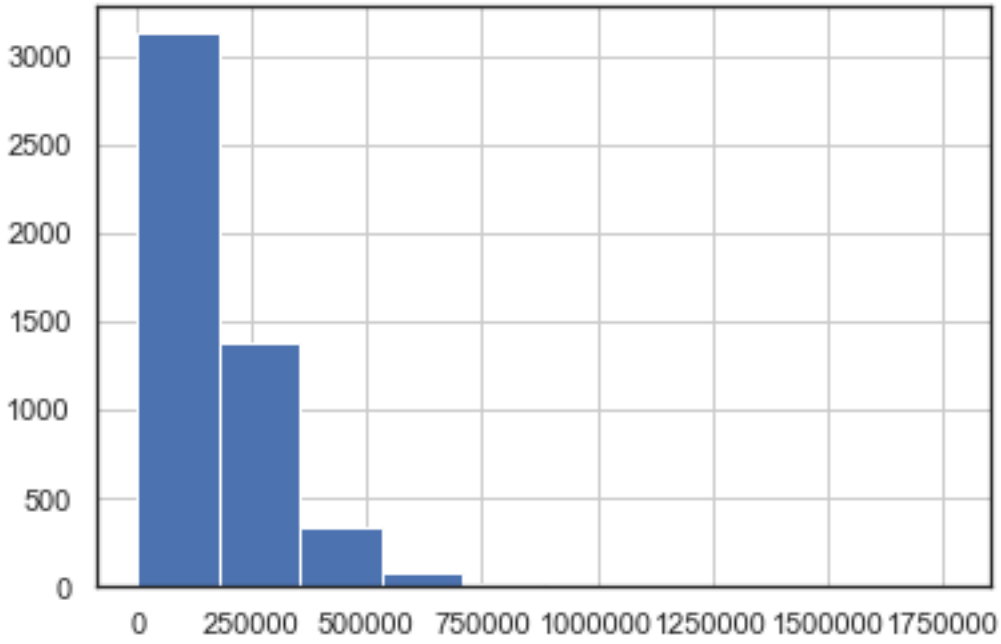


Figura 4.3.11: Distribución de la variable saldo total del cliente. Fuente: Elaboración propia.

La figura 4.3.11 nos muestra que la mayor parte de los registros se tienden a encontrarse

entre un monto de 0 a 100000.

Medida	Valor
count	4,949
mean	157,751
std	146,084
min	1,583
Q1	42,342
Q2	120,424
Q3	230,936
max	1,767,290

Cuadro 4.8: Cuadro de estadísticas descriptivas del saldo total del cliente. Fuente: Elaboración propia.

La distribución de la variable tiene a tener una distribución a los valores menores a la media. Si bien, la media y la mediana tienen comportamientos similares, a mayor parte de los créditos se encuentran en los cuartiles inferiores. Sin embargo, tenemos clientes con mayor valor que se reflejan a partir del 3er cuartil

NormaltestResult(statistic=2141.5500243211054, pvalue=0.0)

Figura 4.3.12: Salida prueba de normalidad para el saldo total del cliente. Fuente: Elaboración propia.

Esta variable no tiene evidencias que se distribuya de manera normal.

acc_open_past_24m

La variable describe el comportamiento de adquisidores de créditos nuevos en los últimos 24 meses. Esto describe la frecuencia con el que el cliente adquiere nuevos créditos. Esta frecuencia de créditos tiende a tener un comportamiento hacia tener pocos créditos adquiridos.

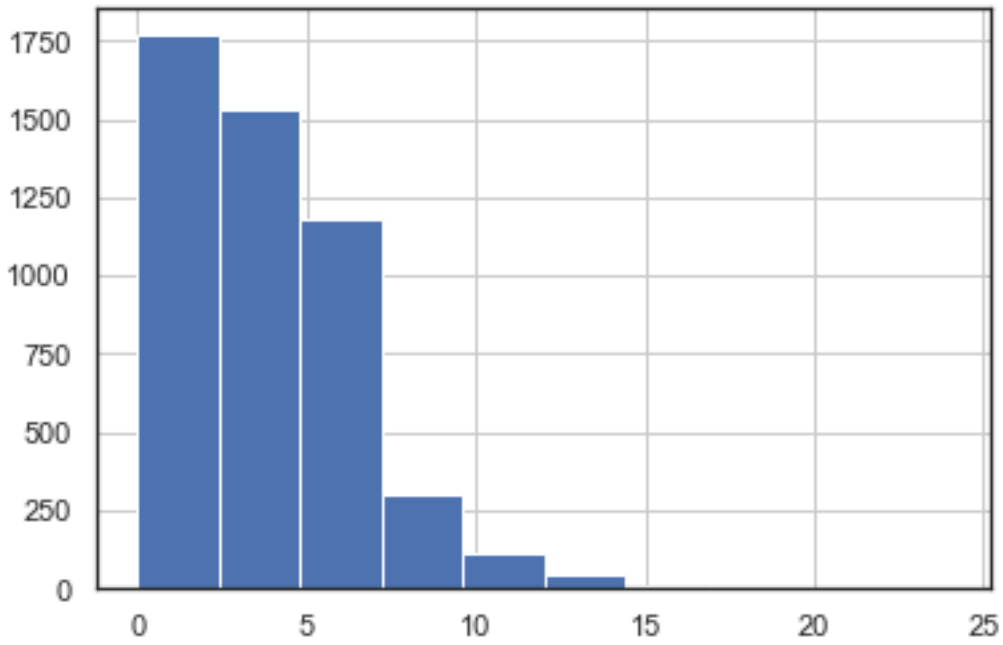


Figura 4.3.13: Distribución de la variable crédito en los últimos 24 meses. Fuente: Elaboración propia.

La figura 4.3.13 nos muestra que la mayor parte de los registros se tienden a encontrarse entre un monto de 0 a 100000.

Medida	Valor
count	4,949
mean	4
std	3
min	0
Q1	2
Q2	3
Q3	5
max	24

Cuadro 4.9: Cuadro de estadísticas descriptivas de crédito en los últimos 24 meses. Fuente: Elaboración propia.

La distribución de la variable tiende a tener una distribución a los valores menores a la media. Si bien, la media y la mediana tienen comportamientos similares, a mayor parte de los créditos se encuentran en los cuartiles inferiores. Sin embargo, tenemos clientes con mayor

valor que se reflejan a partir del 3er cuartil

NormaltestResult(statistic=1060.5414263316593, pvalue=5.0857540917162e-231)

Figura 4.3.14: Salida prueba de normalidad para crédito en los últimos 24 meses. Fuente: Elaboración propia.

No tenemos evidencia en la prueba que esta variable tenga una distribución normal. La distribución global de las variables vistas en una matriz de dispersión, podemos observar que las variables tienen distintas distribuciones.

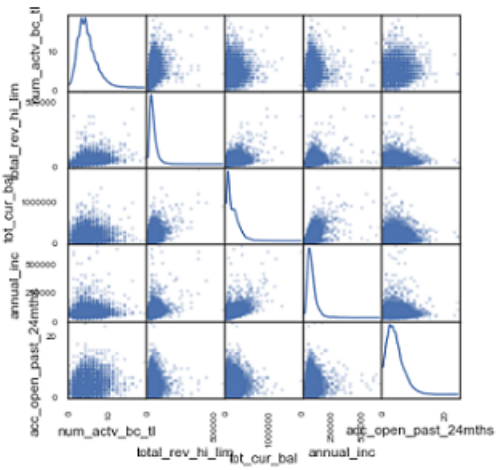


Figura 4.3.15: Distribución de las variables en un mapa de dispersión. Fuente: Elaboración propia.

En la distribución de variables en una vista conjunta se describen que la mayor parte de los valores se encuentran posicionados en el tercer cuadrante (inferior-izquierdo).

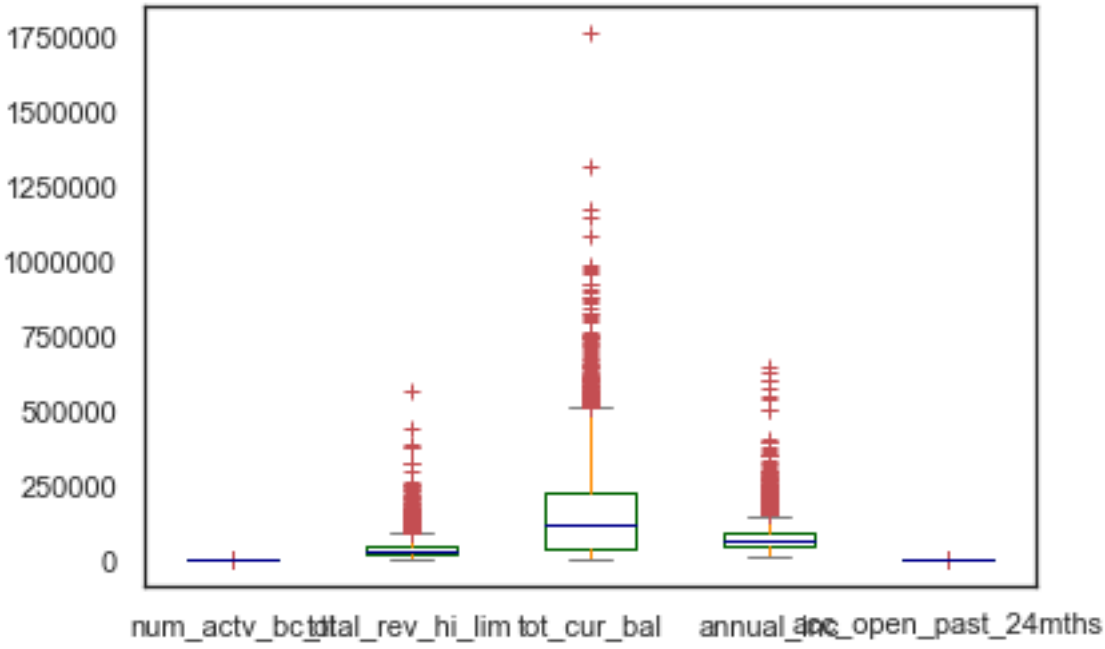


Figura 4.3.16: Diagrama de medias-dispersión. Fuente: Elaboración propia.

Por otro lado, se da a notar que la variable del saldo (tot_cur_bal) contiene el mayor número de valores aberrantes respecto a su media y dispersión.

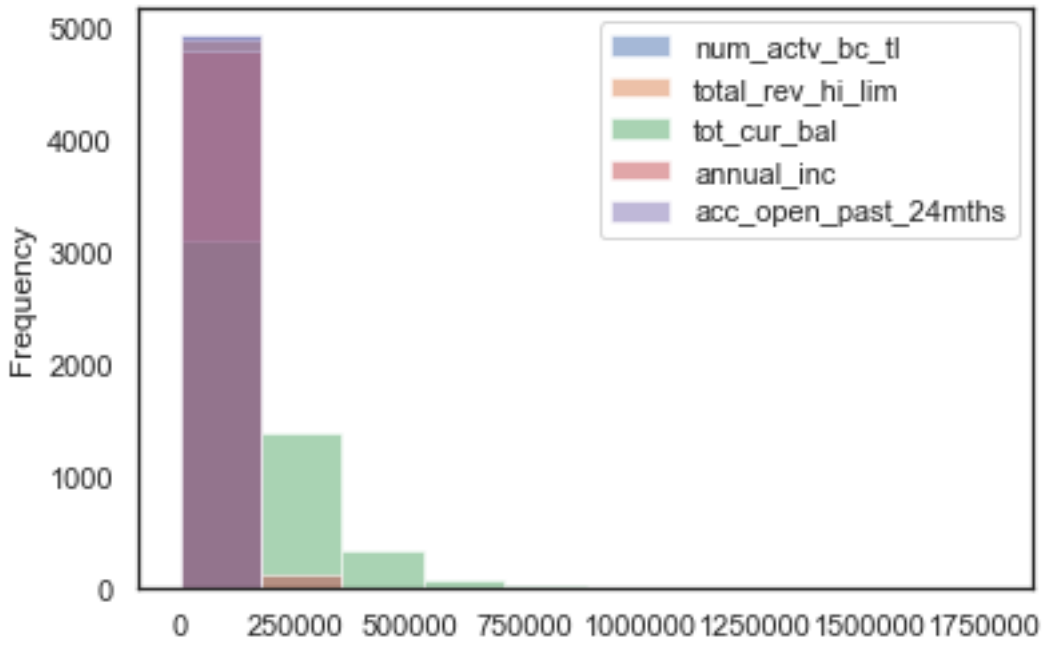


Figura 4.3.17: Distribución de las variables por su comportamiento. Fuente: Elaboración propia.

En la distribución de los datos conjuntos, se analizan que de igual forma tenemos distribuciones con tendencia logarítmica en las variables.

4.3.3. Transformación de los datos

Ya con la elección de las variables de la simulación, la cuales son las mismas planteadas por la regresión logística, realizaremos la transformación de las variables que nos apoyaran más adelante. El proceso de la figura 4.3.18 nos muestra que realizaremos una transformación logarítmica a la base para normalizar el comportamiento. A ello tendremos que imputar el valor 0.01 para no causar errores en el procesamiento de la base. Posterior, aplicaremos logaritmo a las cinco variables que se evaluaran. Y gráfico 4.3.18 muestra la diferencia en la transformación.

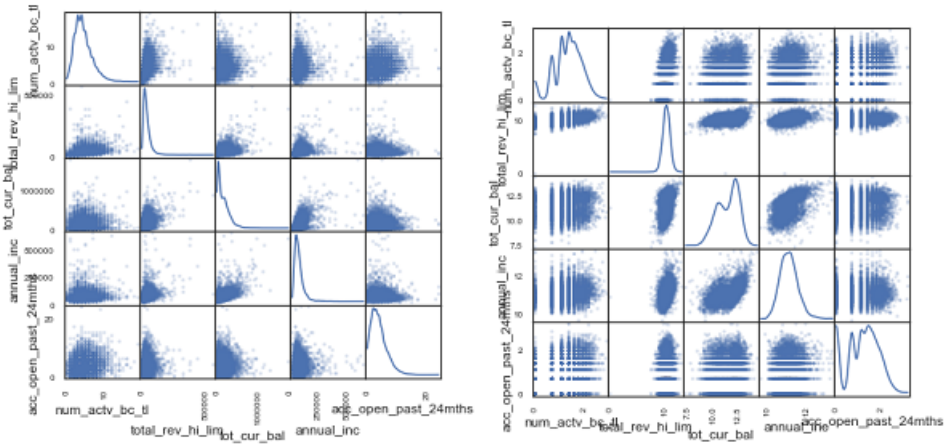


Figura 4.3.18: Contraste de la transformación logarítmica de las variables originales. Fuente: Elaboración propia.

Podemos observar que las variables se han trasladado, y modificado su comportamiento, cercano a una forma de distribución gaussiana. De forma adicional, haremos una normalización de los datos.

Esta nos ayuda a que las variables de entrada a los componentes principales. La prueba de normalidad de estas variables log-normalizadas, no muestran evidencia para aceptar que se comportan de forma normal.

pvalue=array([6.79290120e-039, 1.83892268e-267, 4.53632724e-065, 4.08850302e-072, 4.97081175e-128])

Figura 4.3.19: Salida prueba de normalidad para la transformación de variables originales. Fuente: Elaboración propia.

Con esta prueba se finalizan las transformaciones de las variables y se continua con el modelado de la prueba.

4.4. Modelado

En el modelado del ejercicio, realizaremos la generación del motor de variables. El principal objetivo será buscar los componentes principales y la matriz de vectores propios que nos apoyen en la simulación de nuevos escenarios. En el capítulo tres se postuló el marco teórico de componentes principales, a lo cual ocuparemos en Python.

	CP1	CP2	CP3	CP4	CP5
CP1	1.000067	0.441971	0.098952	0.143456	0.103519
CP2	0.441971	1.000067	0.335790	0.345260	0.046157
CP3	0.098952	0.335790	1.000067	0.488353	0.151132
CP4	0.143456	0.345260	0.488353	1.000067	0.084100
CP5	0.103519	0.046157	0.151132	0.084100	1.000067

Cuadro 4.10: Matriz de correlación de componentes principales. Fuente: Elaboración propia.

La matriz de correlaciones nos permite observar que existe correlaciones positivas entre las variables. Como la selección de estas mismas ha dependido de una regresión logística, ninguna tiene una correlación fuerte (arriba del 0.7).

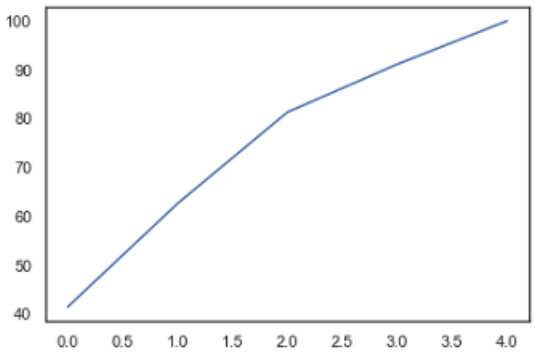
	CP1	CP2	CP3	CP4	CP5
num_actv_bc	0.388595	0.538488	0.508273	0.514906	0.188559
total_rev_lim	0.696860	0.353749	0.456401	0.365893	0.216954
tot_cur_bal	0.169939	0.132119	0.106943	0.229871	0.943074
annual_inc	0.273147	0.422787	0.487414	0.713418	0.010171
acc_open_24m	0.509796	0.623449	0.533242	0.197958	0.166988

Cuadro 4.11: Vectores Propios para Componentes Principales. Fuente: Elaboración propia.

Medias	CP1	CP2	CP3	CP4	CP5
count	4,949.00	4,949.00	4,949.00	4,949.00	4,949.00
mean	0.0	0.0	0.0	0.0	0.0
std	1.44	1.03	0.97	0.70	0.67
min	6.14	3.83	3.65	2.38	2.80
1Q	0.97	0.71	0.67	0.48	0.44
2Q	0.01	0.01	0.03	0.05	0.01
3Q	0.98	0.68	0.65	0.39	0.45
max	9.75	3.92	5.16	5.50	8.61

Cuadro 4.12: Estadísticos descriptivos de Componentes principales. Fuente: Elaboración propia.

En los estadísticos de componentes principales se puede observar que la media de todas las variables es cero y que la desviación estándar no es uno para todos los casos. Con esta media y varianza realizaremos las simulaciones Monte Carlo.



	CP1	CP2	CP3	CP4	CP5
VarIndividual	39.64 %	21.04 %	19.49 %	10.36 %	9.47 %
VarAcumulada	39.64 %	60.68 %	80.17 %	90.53 %	100.00 %

Figura 4.4.1: Varianza explicada por Componente Principal. Fuente: Elaboración propia.

En el conjunto de los componentes principales, como un criterio, el primer componente tiene la mayor parte de la varianza explicada, así hasta el último componente.

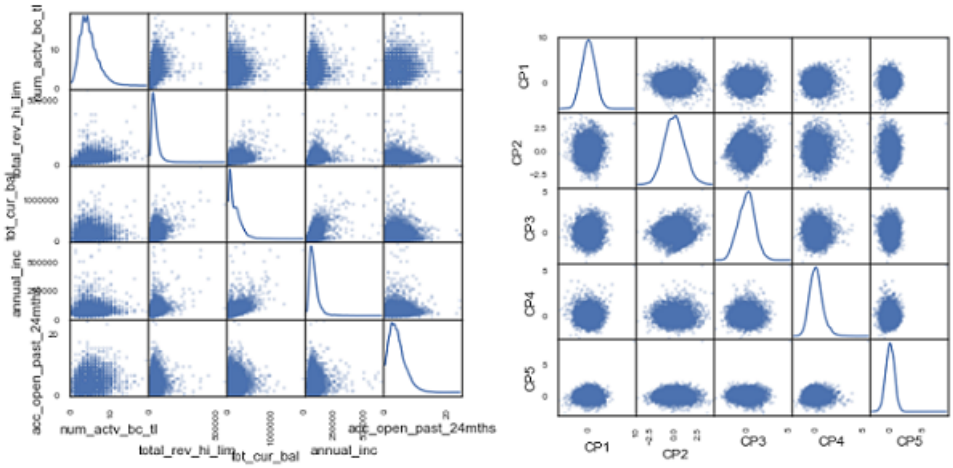
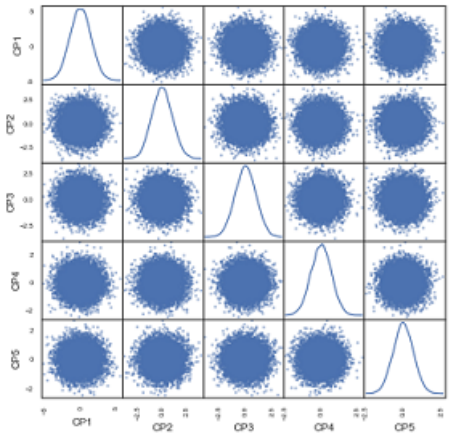


Figura 4.4.2: Contraste de la transformación por componentes principales lognormalizadas de las variables originales. Fuente: Elaboración propia.

Además, los datos mantienen una dispersión conforme a la técnica, ortonormal lo cual nos apoyara en la simulación Monte Carlo. Al tener estas distribuciones la generación de la simulación Monte Carlo.

4.4.1. Aplicación de componentes principales

Para la generación del universo utilizaremos el método Monte Carlo para realizar simulaciones aleatorias de múltiples, esta simulación será independientes. Para el ejercicio generaremos diez mil casos de simulación con la media y varianza de los componentes principales.



	Simulación1	Simulación2	Simulación3	Simulación4	Simulación5
pvalue	0.6828	0.9559	0.4700	0.7383	0.8407

Figura 4.4.3: Matriz de dispersión de simulación Monte Carlo que sigue una función normal. Fuente: Elaboración propia.

Con la prueba visual y de la prueba estadística, podemos observar que cinco variables son normales y no tienen una correlación observable dentro de ellas, con lo que de igual forma son normales.

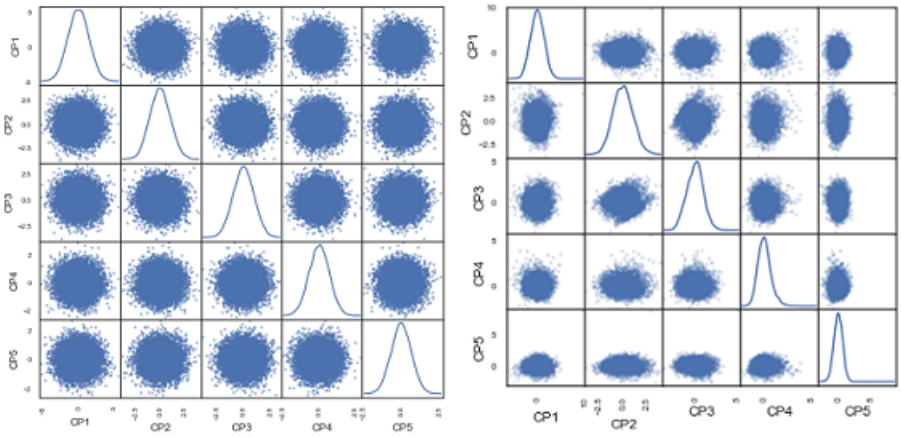


Figura 4.4.4: Contraste de la simulación Monte Carlo y las variables originales Componentes Principales. Fuente: Elaboración propia.

Al comparar ambos comportamientos, observamos existe un comportamiento distinto en el componente principal cinco, derivado a que es el componente que se lleva el rezago de la varianza.

4.4.2. Transformación de las variables

Con la generación de las variables normales, bajo la metodología de Kreinin et al. [1998] se realiza la multiplicación de la matriz de eigenvalores de componentes principales con 10,000 registros simulados.

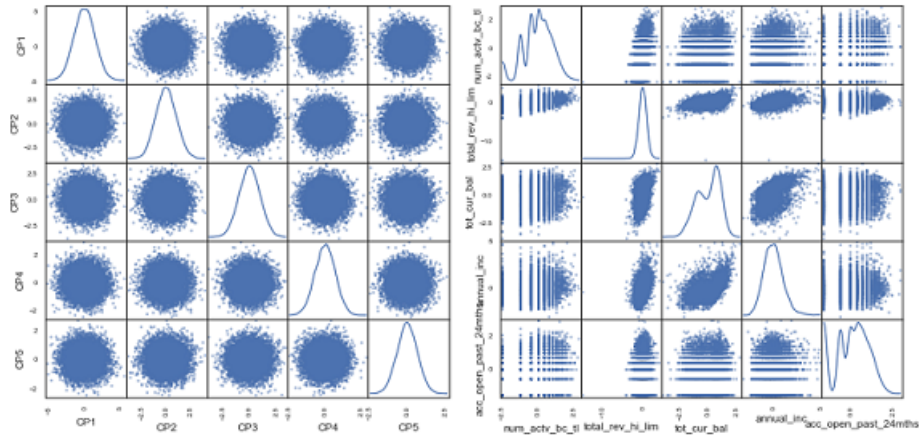


Figura 4.4.5: Contraste de la simulación Monte Carlo transformada de Componentes Principales y las variables originales Lognormalizadas. Fuente: Elaboración propia.

Simulación	num_actv_bc	total_rev_lim	tot_cur_bal	annual_inc	acc_open_24m
mean	0.005132	0.010774	0.008842	0.005597	0.001197
std	0.990667	1.00176	1.013529	0.9999	1.012398
min	4.399734	4.244362	3.826767	3.500525	3.974003
Q1	0.666646	0.690996	0.686421	0.685021	0.662808
Q2	0.00143	0.010863	0.006367	0.007417	0.00593
Q3	0.676011	0.661772	0.672186	0.668822	0.677657
max	3.646504	3.936357	3.967273	3.423323	3.889358
Test	num_actv_bc	total_rev_lim	tot_cur_bal	annual_inc	acc_open_24m
mean	0	0	0	0	0
std	1.000101	1.000101	1.000101	1.000101	1.000101
min	2.444951	15.01975	3.917529	3.140634	1.600169
Q1	0.457794	0.62056	0.803444	0.67719	0.616391
Q2	0.062562	0.004265	0.186973	0.076953	0.040918
Q3	0.795963	0.647367	0.803944	0.593599	0.684092
max	2.880916	4.178894	2.732268	4.868203	2.910416

Cuadro 4.13: Comparativos descriptivos Monte Carlo transformada de Componentes Principales y las variables originales Lognormalizadas. Fuente: Elaboración propia.

Para desnormalizar las distribuciones sumaremos la media y dividiremos el resultado entre la desviación estándar.

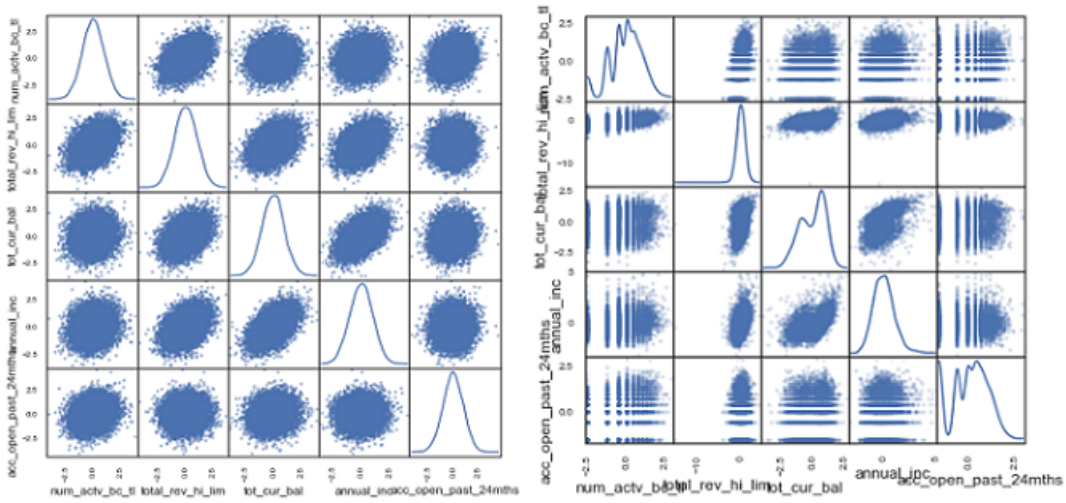


Figura 4.4.6: Contraste de la simulación Monte Carlo transformada de Componentes Principales desnormalizado y las variables originales normalizadas. Fuente: Elaboración propia.

Simulación	num_actv_bc	total_rev_lim	tot_cur_bal	annual_inc	acc_open_24m
mean	1.216363	10.167808	11.429922	11.150115	1.214722
std	0.586508	0.784151	1.150006	0.467075	0.689851
min	1.111951	7.013477	7.4677	9.329918	1.85188
Q1	0.828024	9.635089	10.649722	10.829664	0.757675
Q2	1.21873	10.161045	11.435936	11.147711	1.219662
Q3	1.605943	10.70656	12.217197	11.465835	1.677338
max	3.685069	12.994397	15.629901	12.863699	3.600215
Test	num_actv_bc	total_rev_lim	tot_cur_bal	annual_inc	acc_open_24m
count	4949	4949	4949	4949	4949
mean	1.213762	10.168041	11.441543	11.152582	1.221648
std	0.587683	0.782635	1.13569	0.463256	0.694591
min	0	0	0	9.798127	0
Q1	0.693147	9.705037	10.580632	10.819778	0.693147
Q2	1.386294	10.199882	11.619571	11.127263	1.386294
Q3	1.609438	10.666627	12.341709	11.429544	1.791759
max	3.044522	13.90233	14.644352	15.33285	3.218876

Cuadro 4.14: Comparativos descriptivos Monte Carlo transformada de Componentes Principales desnormalizados y las variables originales. Fuente: Elaboración propia.

Finalmente realizaremos las transformaciones inversas del logaritmo y tener simulación final.

4.5. Evaluación

4.5.1. Pruebas descriptivas de igualdad

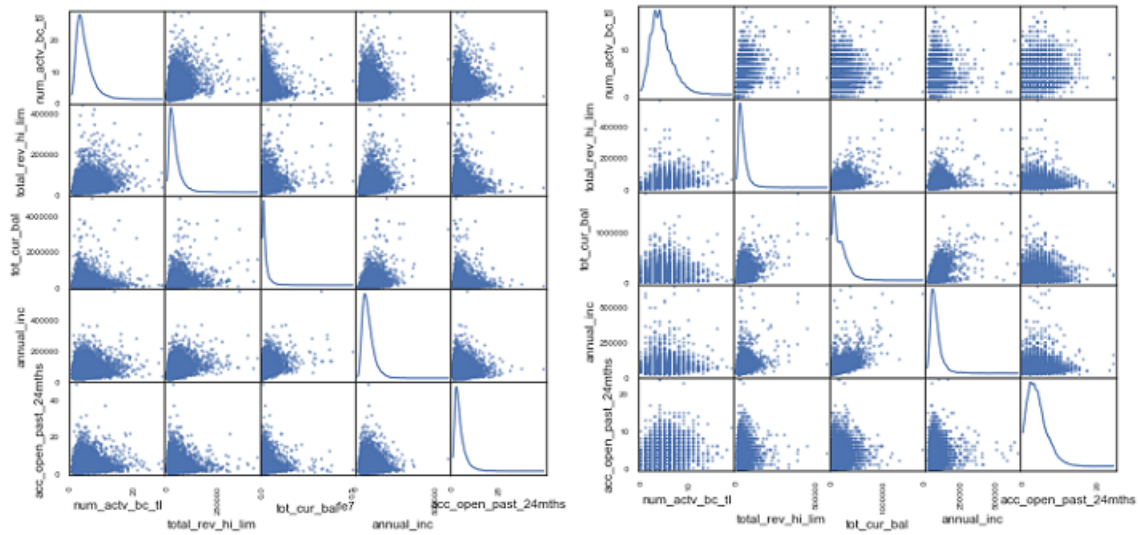


Figura 4.5.1: Contraste de la simulación Monte Carlo transformada de Componentes Principales desnormalizado y las variables originales. Fuente: Elaboración propia.

Simulación	num_actv_bc	total_rev_lim	tot_cur_bal	annual_inc	acc_open_24m
mean	4.5	39,586.7	167,395.1	75,556.3	4.0
std	2.7	31,235.7	224,668.3	36,936.7	3.1
min	0.4	2,771.9	2,129.4	10,772.5	0.2
Q1	2.6	19,742.4	49,122.9	49,292.6	1.9
Q2	3.8	31,174.1	98,371.5	67,667.0	3.1
Q3	5.6	49,314.0	198,684.4	92,947.8	5.0
max	28.5	632,658.4	4,945,232.0	39,4116.8	38.5
Test	num_actv_bc	total_rev_lim	tot_cur_bal	annual_inc	acc_open_24m
count	4949	4949	4949	4949	4949
mean	4.4	39864.5	157750.6	77,010.0	3.8
std	2.3	32,587.2	146,083.5	46,457.3	2.7
min	0.0	0.0	1,583.0	16,000.0	0.0
Q1	3.0	20,600.0	42,342.0	50,000.0	2.0
Q2	4.0	31,700.0	1,204,24.0	66,000.0	3.0
Q3	6.0	49,400.0	2,309,36.0	90,000.0	5.0
max	19.0	564,600.0	1,767,290.0	650,000.0	24.0

Cuadro 4.15: Comparativos descriptivos Monte Carlo transformada de Componentes Principales deslognormalizada y las variables originales. Fuente: Elaboración propia.

Con ayuda de la figura 4.5.1 y el cuadro 4.15 podemos contrastar el comportamiento de las variables originales y las variables simuladas.

Observamos que las desviaciones estadar y las medias son cercanas. Esto podría ser posible con una simulación Monte Carlo estándar, sin embargo, atrás de las variables, existen correlaciones entre las variables que nos ayudaran a explicar las variaciones del portafolio de crédito.

4.5.2. Pruebas igualdad de la función de distribución individual

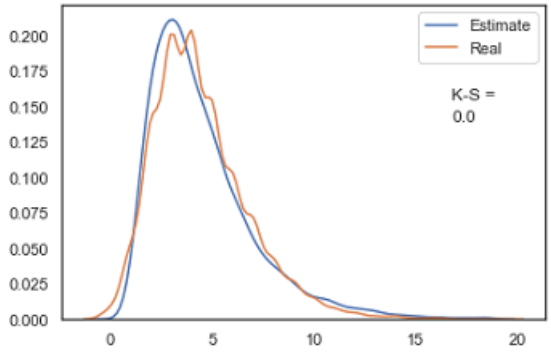


Figura 4.5.2: Comparación de distribución entre la simulación y real de número de tarjetas. Fuente: Elaboración propia.

En la variable de número de tarjetas, a pesar de que no se ha podido replicar el movimiento aleatorio del comportamiento, si podemos observar que ambas variables tienen el mismo comportamiento.

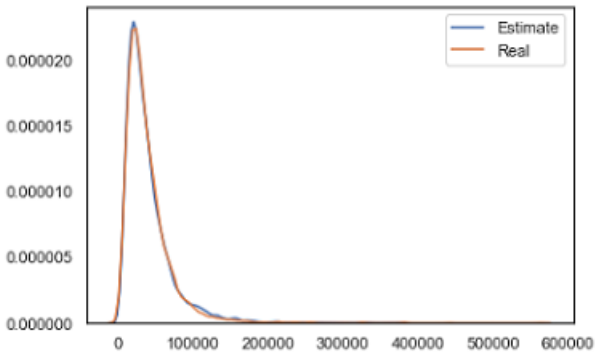


Figura 4.5.3: Comparación de distribución entre la simulación y real la línea amortizable. Fuente: Elaboración propia.

En la variable de limite amortizable del cliente, no existen evidencias de que las variables sigan una distribución distinta.

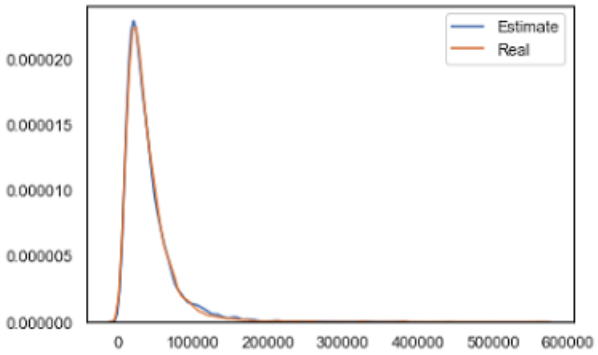


Figura 4.5.4: Comparación de distribución entre la simulación y real del saldo total de créditos. Fuente: Elaboración propia.

La variable saldo total del cliente, se observa que se sigue una misma distribución, sin embargo, algunos movimientos de la curva no son idénticos.

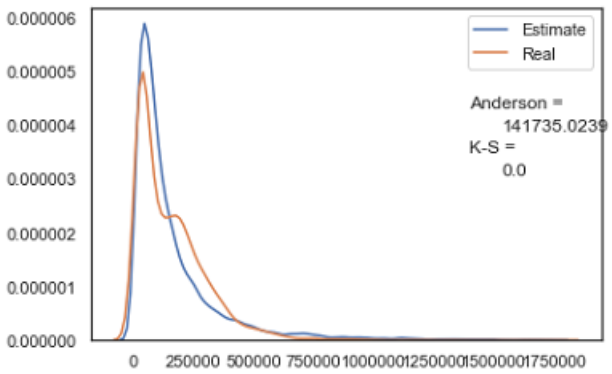


Figura 4.5.5: Comparación de distribución entre la simulación y real de ingresos anuales. Fuente: Elaboración propia.

En la variable ingreso, se observa que la simulación ha sido efectiva contra la variable real.

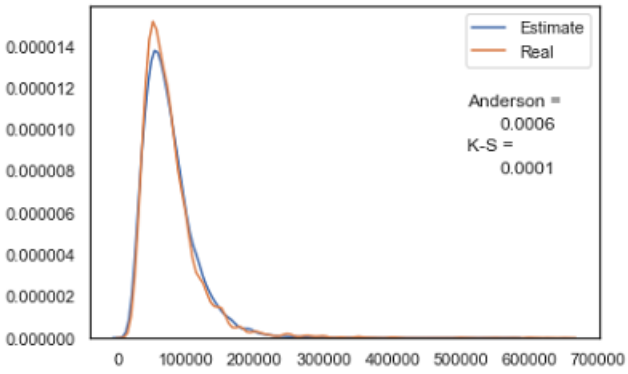


Figura 4.5.6: Comparación de distribución entre la simulación y real del número de créditos contratados en los últimos 24 meses. Fuente: Elaboración propia.

En la variable de contratación, es la simulación que nos muestra mayor diferencia respecto a las otras variables simuladas.

4.6. Implantación

4.6.1. Prueba de un método predictivo

Se tomó el modelo de probabilidad de default que se calculó y se aplica a la base simulada. Con lo que se tiene.

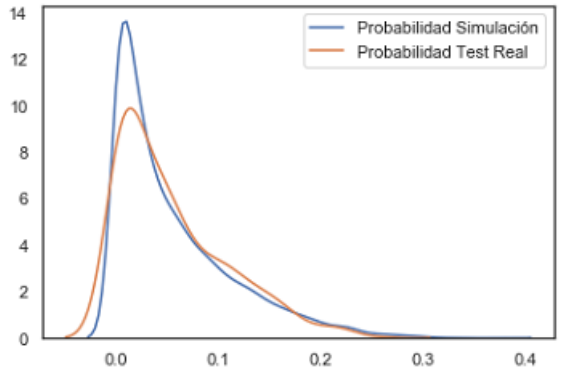


Figura 4.6.1: Comparación de distribución entre la simulación y real de pérdida esperada. Fuente: Elaboración propia.

Con ello podemos observar que la simulación de variables nos es efectiva dentro de sus componentes, con la matriz de covarianzas para poder responder a comportamientos predictivos. Si bien existe un número de valores dentro de cada una de las variables que impiden recrear un valor predicho por su sensibilidad. También es cierto que se generó un escenario al doble de las observaciones en la base de prueba.

4.6.2. Simulación para escenarios de toma de decisiones

Con las conclusiones de la sección anterior, realizaremos modificaciones en la generación de componentes principales, para observar las modificaciones en el portafolio. Se escoge el componente principal número cuatro. Y se realiza una modificación en el parámetro de la simulación de la varianza, la cual se ve afectada y mostrada en el siguiente cuadro.

	Varianza			
	0	1	2	3
Número de tarjetas	5	6	8	10
Revolvente del cliente	42,155	27,694	17,548	11,438
Saldo total	195,804	342,981	645,790	1,090,804
Ingreso declarado	76,031	70,578	66,478	61,632
Cuentas abietas 24m	4	3	3	3

Cuadro 4.16: Cambios descriptivos por modificaciones en la simulación. Elaboración propia.

En el cuadro 4.16 podemos observar que la modificación de la varianza en el componente cuatro, afecta positivamente el número de tarjetas, el saldo total del cliente. Por otro lado, afecta de forma negativa al saldo revolvente del cliente, el ingreso y número de cuentas abiertas en 24 meses. Este cambio en el comportamiento afecta a la probabilidad de incumplimiento como se muestra en la figura 4.6.2.

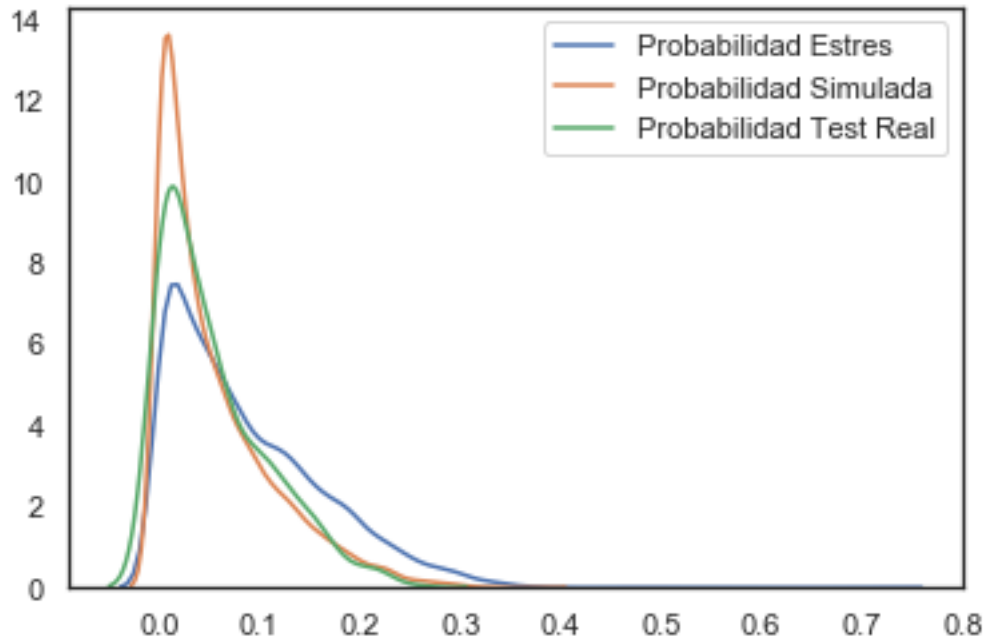


Figura 4.6.2: Comparación de distribución entre la simulación y real de pérdida esperada. Elaboración propia.

Podemos observar que las modificaciones del negocio han realizado que se tengan cambios significativos en la distribución global de la perdida espera. En términos de una estrategia de negocio, se puede determinar el camino a seguir en la gestión de los nuevos créditos admitidos bajo esta herramienta o proceso estadístico.

CONCLUSIONES

Entre los principales hallazgos estuvieron: la delimitación del universo de variables para estudiar, el cual, implicó diversos filtros y manipulación de base de datos. Donde se encontraron las variables que predicen la perdida esperada bajo los parámetros especificados con ayuda de una regresión logística, las cuales fueron cinco variables. Posteriormente se realizó un análisis descriptivo de las variables a nivel univariado, encontrado las distribuciones, medias y dispersiones, mismas que ayudarían a la simulación subsecuente.

Se profundiza sobre el comportamiento normal de las variables, donde se encontró que el comportamiento no seguía una distribución normal. Para poder seguir el modelado de la simulación por componentes principales se aplicó una transformación logarítmica y una normalización, con lo que se aseguró el comportamiento normal de las variables univariadas. Con ello se pudo aplicar la técnica de componentes principales, con el objetivo de crear variables ortonormales y la matriz de vectores propios que pide la teoria. Se tomó esta matriz de vectores propios como la correlación y características de las variables.

Por otro lado, se generaron cinco variables aleatorias por medio de método de simulación Monte Carlo con parámetros de media cero y varianza uno, con el fin de asemejar con el resultado de componentes principales. Este método nos asegura que cada una de las variables tienen un comportamiento normal e independiente. Con estas características similares entre las variables ortonormales de componentes principales y la simulación Monte Carlo, se optó

por transformar las variables de la simulación a una simulación correlacionada por medio de transformaciones inversas obtenidas del proceso de componentes principales. Con la simulación correlacionada, se realizó una comparación univariada en la distribución de cada una de las variables contra las variables originales y por medio del estadístico de comparación Anderson-Darling, se pudo comprobar que las variables por separado tienen la misma distribución de las originales.

También se realizó un análisis de la variable de perdida esperada, bajo el modelo anteriormente generado por la regresión logística. Las comparaciones de la perdida esperada simulada y la generada de variables originales siguen la misma distribución. Dado el comporntamiento de las distribuciones univariadas y de la función aplicada.

Se puede concluir que el ejercicio ha cumplido la expectativa de realizar una simulación que se asimile a la distribución original. Con la simulación y los parámetros de media y varianza de las variables, se realizó un ejercicio de simulación con modificaciones en los parametros, el cual, cambia el comportamiento en perdida esperada..

Llevado a cabo el ejercicio práctico con información real de una institución de crédito, se puede observar que el método de simulación Monte Carlo correlacionado puede ser utilizado como una herramienta de apoyo en las decisiones en un porfolio de crédito y la simulación de la pérdida esperada. Estas simulaciones pueden ser tomadas en función de parámetros al ser comprensibles para los tomadores de decisión de los negocios. Además, se muestra un camino de desarrollo e implementación con la metodología antes presentada.

El desarrollo de esta herramienta ha sido posible al descomponer el problema planteado en las fases que caracteriza al marco de referencia de minería de datos, CRISP-DM, ya que las fases que se incluyen nos muestran y guían en el camino de la investigación. Muestran al lector

un panorama previo del problema, las respectivas modificaciones, la generación del modelo, las simulaciones, las comparaciones con las distribuciones originales y un ejercicio de simulación modificada. Si bien los otros marcos de referencia nos hubieran ayudado a llegar al objetivo, la metodología CRISP-DM, tiene como ventaja incluir un apartado especial para la simulación, la bibliografía consultable es más extensa y explicativa de cada fase.

En mancuerna CRIP-DM y Python, se muestran como una alternativa para el lector al realizar análisis con minería de datos, tanto como por su flexibilidad como por lo robusto que pueden ser los ambientes de trabajo de ambos.

Además, el documento muestra un uso no convencional de los componentes principales el cual no es citado frecuentemente por los autores de la estadística clásica o de simulación, este uso corresponde a utilizar la matriz de valores propios como una fuente de generación de información. En términos de estadística, la prueba resulta de interés para quienes quieran conocer otra aplicación para los componentes principales, distintos a las aplicaciones clásicas de utilización como lo son la reducción de variables correlacionadas y la selección de unificación de varianza en menores componentes.

Dentro de los datos revisados, observamos que bajo los parametros especificados en los criterios de riesgos de la sección 4.1, existiran clientes con probabilidades menores al 20 % que podran tener incentivación de crédito, ya que el ingreso no es el unico componente de riesgo, existen otras variables que los denota como clientes susceptibles de crédito. Lo que puede aventajar al negoció bancario. Por otro lado existe una población mayor de clientes con Ingreso declarado menor, que son propensos al impago en caso de mover el portafolio a un segmento menor de ingresos. Estos clientes, dado el apetito de crédito del banco en cuestion, no son susceptibles de crédito.

ANEXO

Anexo I: Código Python

El codigo presentado se encuentra tambien en la siguiente ruta:

https://github.com/jfabrizios/Toma-de-decisiones-con-miner-a-de-datos-para-un-portafolio-de-cr-dito-Simulaci-n-por-componentes-pr/blob/master/Code

-*- coding: utf-8 -*- """ Created on Sat Jan 27 20:00:54 2018

@author: Juan Fabrizio Sanchez """

#Creación de un flujo para una simulación por componentes pricipales

#Pasos a seguir

análisis de la base

Transformacion

Componentes principales

Generación de espacio

Contaste de pruebas

######## LIBRERIAS ###########

import numpy as np from sklearn.decomposition

import PCA import pandas as pd

import matplotlib.pyplot as plt from sklearn.preprocessing import scale

import scipy from scipy import stats

import seaborn as sns

from pandas.tools.plotting import scatter_matrix

```
from scipy.stats import ks_2samp
from string import ascii_letters
from pandas import Series, DataFrame from pylab
import rcParams from sklearn import preprocessing
from sklearn.linear_model import LogisticRegression
from sklearn.cross_validation import train_test_split
from sklearn import metrics
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.metrics import classification_report
import statsmodels.api as sm
######## análisis DE LA MUESTRA ##########
### Introducimos la base
data=pd.read_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/LC_2014.csv',header=0)
### Información técnica de la base data.info()
### Información purpose
ax=data.groupby(["purpose"]).size().plot(kind='bar',alpha=.70,) totals = [] for i in ax.patches:
totals.append(i.get_height()) total = sum(totals) for i in ax.patches: ax.text(i.get_x()+.12,
i.get_height()-3, \ str(round((i.get_height()/total)*100, 1))+' %', fontsize=12, color='black')
data_tdc = data[(data.purpose == "credit_card") ]
### Información loan_status
ax=data_tdc.groupby(["loan_status"]).size().plot(kind='bar',alpha=.70,)
totals = [] for i in ax.patches: totals.append(i.get_height()) total = sum(totals)
for i in ax.patches: ax.text(i.get_x()+.12, i.get_height()-3, \
str(round((i.get_height()/total)*100, 1))+' %', fontsize=12, color='black')
### Filtrado de los P = 1 y 0
data_tdc1 = data_tdc[(data_tdc.loan_status != "Default")
& (data_tdc.loan_status != "Fully Paid") & (data_tdc.loan_status != "Charged Off")
```

```
& (data_tdc.loan_status != "In Grace Period")]
data_tdc1['target']=np.where(data_tdc1['loan_status']=="Current", 0, 1)
ax=data_tdc1.groupby(["delinq_2yrs"]).size().plot(kind='bar',alpha=.70,)
totals=[] for i in ax.patches: totals.append(i.get_height()) total = sum(totals)
for i in ax.patches: ax.text(i.get_x()+.12, i.get_height()-3, \
str(round((i.get_height()/total)*100, 0))+' %', fontsize=12, color='black')
data_tdc2 = data_tdc1[(data_tdc.delinq_2yrs == 0)]
### Información incumplimiento
ax=data_tdc2.groupby(["loan_status"]).size().plot(kind='bar',
color="coral", fontsize=13); totals = [] for i in ax.patches:
totals.append(i.get_height()) total = sum(totals) for i in ax.patches:
ax.text(i.get_x()+.12, i.get_height()-3, \
str(round((i.get_height()/total)*100, 2))+' %', fontsize=12, color='black')
### Diseño de datos
data_tdc2.info()
### Eliminación de cadenas
df = data_tdc2[data_tdc2.T[data_tdc2.dtypes!=np.object].index]
df.info()
### Quitamos variables sin valores
des_df=df.describe()
des_df.to_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/out.csv')
dg=df.drop(['member_id', 'url', 'annual_inc_joint', 'dti_joint', 'verification_status_joint',
'open_acc_6m', 'open_act_il', 'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il',
'total_bal_il', 'il_util', 'open_rv_12m', 'open_rv_24m', 'max_bal_bc', 'all_util',
'inq_fi', 'total_cu_tl', 'inq_last_12m', 'revol_bal_joint', 'sec_app_earliest_cr_line',
'sec_app_inq_last_6mths', 'sec_app_mort_acc', 'sec_app_open_acc',
'sec_app_revol_util', 'sec_app_open_act_il', 'sec_app_num_rev_accts',
```

```
'sec_app_chargeoff_within_12_mths', 'sec_app_collections_12_mths_ex_med',
'sec_app_mths_since_last_major_derog', 'deferral_term', 'hardship_amount',
'hardship_length', 'hardship_dpd', 'orig_projected_additional_accrued_interest',
'hardship_payoff_balance_amount', 'hardship_last_payment_amount',
'settlement_amount', 'settlement_percentage', 'settlement_term',
'delinq_2yrs', 'recoveries', 'collection_recovery_fee', 'policy_code',
'acc_now_delinq', 'delinq_amnt', 'num_tl_120dpd_2m', 'num_tl_30dpd',
'num_tl_90g_dpd_24m'], axis=1, inplace=True)
df.info()
#df.to_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/limpia.csv')
### Utilizaremos Corr para ver las variables correlacionadas
sns.set(style="white")
# Compute the correlation matrix corr = df.corr()
#corr.to_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/corr.csv')
# Generate a mask for the upper triangle mask =
np.zeros_like(corr, dtype=np.bool) mask[np.triu_indices_from(mask)] = True
# Set up the matplotlib figure f, ax = plt.subplots(figsize=(11, 9))
# Generate a custom diverging colormap cmap = sns.diverging_palette(220, 10, as_cmap=True)
# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.9, center=0, square=True,
linewidths=.5, cbar_kws={"shrink": .5})
#Variables con alta correlación
#df = data_tdc2[data_tdc2.T[data_tdc2.dtypes!=np.object].index] #dg=df.drop(['member_id',
'url', 'annual_inc_joint', 'dti_joint', 'verification_status_joint', 'open_acc_6m', 'open_act_il',
'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_12m',
'open_rv_24m', 'max_bal_bc', 'all_util', 'inq_fi', 'total_cu_tl', 'inq_last_12m', 'revol_bal_joint',
'sec_app_earliest_cr_line', 'sec_app_inq_last_6mths', 'sec_app_mort_acc', 'sec_app_open_acc',
```

'sec_app_revol_util', 'sec_app_open_act_il', 'sec_app_num_rev_accts', 'sec_app_chargeoff_within_1
'sec_app_collections_12_mths_ex_med', 'sec_app_mths_since_last_major_derog', 'de-
ferral_term', 'hardship_amount', 'hardship_length', 'hardship_dpd', 'orig_projected_additional_accrued_i
'hardship_payoff_balance_amount', 'hardship_last_payment_amount', 'settlement_amount',
'settlement_percentage', 'settlement_term', 'delinq_2yrs', 'recoveries', 'collection_recovery_fee',
'policy_code', 'acc_now_delinq', 'delinq_amnt', 'num_tl_120dpd_2m', 'num_tl_30dpd',
'num_tl_90g_dpd_24m'], axis=1, inplace=True) #dg=df.drop(['collections_12_mths_ex_med',
'total_rec_late_fee','funded_amnt_inv', 'installment', 'out_prncp', 'out_prncp_inv', 'to-
tal_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'num_actv_rev_tl', 'num_bc_sats',
'num_bc_tl', 'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0',
'funded_amnt', 'mths_since_last_delinq', 'mths_since_recent_bc', 'mths_since_recent_bc_dlq',
'mths_since_recent_inq', 'num_sats', 'total_acc', 'tax_liens', 'pub_rec_bankruptcies', 'to-
tal_il_high_credit_limit', 'percent_bc_gt_75', 'bc_open_to_buy', 'total_bc_limit', 'tot_hi_cred_lim',
'avg_cur_bal'], axis=1, inplace=True)

```
df.info()
### Observamos la regresión logística
#Creamos la variable independiente
dh=df.fillna(int(0)) dh dh.info()
X_train, X_test, y_train, y_test = train_test_split( dh.iloc[:, 1:59], dh.target,
test_size = .7, random_state=25)
X_train.info()
regressorOLS = sm.OLS(y_train, X_train).fit() regressorOLS.summary()
#Variables no significativas
di=dh[['num_actv_bc_tl', 'total_rev_hi_lim', 'tot_cur_bal', 'annual_inc',
'acc_open_past_24mths' , 'target']]
di.info() #di.to_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/di.csv')
#dj=di[(di.target == 1)] #dj2=di[(di.target == 0)]
```

```
#dj3 = np.array(dj2)
#dj4=dj3[np.random.choice(dj3.shape[0], 187, replace=False)]
#dj5 = pd.DataFrame(dj4, columns=['num_actv_bc_tl', 'total_rev_hi_lim',
'tot_cur_bal', 'annual_inc', 'acc_open_past_24mths' , 'target'])
#dj5
#dj6=pd.concat([dj,dj5]) #dj6
#X_train, X_test, y_train, y_test = train_test_split( dj6.iloc[:, 0:5],
dj6.target, test_size = .7, random_state=25)
#regressorOLS = sm.OLS(y_train, X_train).fit() #regressorOLS.summary()
# Regresión para reestimación LogReg = LogisticRegression() LogReg.fit(X_train, y_train)
#x_pred = LogReg.predict(X_train)
#x_prob = LogReg.predict_proba(X_train)
#intercept=LogReg.intercept_ #intercept #coef=LogReg.coef_ #coef
predicted = LogReg.predict(X_test1)
probs = LogReg.predict_proba(X_test1)
#probs = pd.DataFrame(probs)
#probs.to_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/20182/X_test1.csv')
#x1.to_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/x1.csv')
metrics.accuracy_score(y_test1, predicted) metrics.roc_auc_score(y_test1, probs[:, 1])
metrics.confusion_matrix(y_test1, predicted)
metrics.classification_report(y_test1, predicted)
print(x_pred)
y_pred = LogReg.predict(X_test)
score_train=LogReg.score(X_train, y_train)
print ("Fit a model X_train, and calculate MSE with Y_train:",
np.mean((y_train - LogReg.predict(X_train)) ** 2)) print
("Fit a model X_train, and calculate MSE with X_test, Y_test:", np.mean((y_test -
```

```
LogReg.predict(X_test)) ** 2))
X_train, X_test, y_train, y_test = train_test_split( di.iloc[:, 0:5], di.target,
test_size = .1, random_state=25)
model_2=LogReg.fit(X_train,y_train)
#print (com1) ### ROC Train from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve logit_roc_auc = roc_auc_score(y_train, LogReg.predict(X_trair
fpr, tpr, thresholds = roc_curve(y_train, LogReg.predict_proba(X_train)[:,1])
plt.figure()
plt.subplot(121)
plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc)
plt.show()
### ROC Test from sklearn.metrics import roc_auc_score from
sklearn.metrics import roc_curve logit_roc_auc1 = roc_auc_score(y_test, LogReg.predict(X_test))
fpr1, tpr1, thresholds1 = roc_curve(y_test, LogReg.predict_proba(X_test)[:,1])
plt.figure() plt.subplot(121) plt.plot(fpr, tpr, label='Logistic Regression Real' )
plt.plot(fpr1, tpr1, label='Logistic Regression Test') plt.show()
intercept=model_2.intercept_ intercept coef=model_2.coef_ coef
probs = model_2.predict_proba(X_test)
regressorOLS = sm.OLS(y_train, X_train).fit() regressorOLS.summary()
pd.set_option('display.width', 150)
header = ['pd_0', 'pd_1' ]
header pd_1 = np.array(probs)
pd_1
pd_2 = pd.DataFrame(pd_1, columns=['pd_0', 'pd_1' ])
pd_2['New_ID']= pd_2.index pd_2
target = pd.DataFrame(y_train, columns=['target'])
target
```

```
target.insert(0, 'New_ID',
range(len(target)))
target['ID']= target.index
compara=target.merge(pd_2, left_on='New_ID', right_on='New_ID', how='inner')
sns.kdeplot(compara.pd_1,label="Probabilidad 1")
plt.show()
X_train.info()
### Nos quedamos con las variables de interes x1=di[['num_actv_bc_tl', 'total_rev_hi_lim',
'tot_cur_bal', 'annual_inc', 'acc_open_past_24mths']] x1
#x1.to_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/x1.csv')
####### análisis visual annual_inc
##histograma
x1.annual_inc.hist()
##Descriptivo
x1.annual_inc.describe()
## Pruebas de normalidad stats.mstats.normaltest(x1.annual_inc)
####### análisis visual num_actv_bc_tl
##histograma
x1.num_actv_bc_tl.hist()
##Descriptivo x1.num_actv_bc_tl.describe()
## Pruebas de normalidad stats.mstats.normaltest(x1.num_actv_bc_tl)
####### análisis visual
total_rev_hi_lim
##histograma
x1.total_rev_hi_lim.hist()
##Descriptivo
x1.total_rev_hi_lim.describe()
```

```
## Pruebas de normalidad
stats.mstats.normaltest(x1.total_rev_hi_lim)
####### análisis visual
tot_cur_bal
##histograma
x1.tot_cur_bal.hist()
##Descriptivo
x1.tot_cur_bal.describe()
## Pruebas de normalidad stats.mstats.normaltest(x1.tot_cur_bal)
####### análisis visual
acc_open_past_24mths
##histograma
x1.acc_open_past_24mths.hist()
##Descriptivo
x1.acc_open_past_24mths.describe()
## Pruebas de normalidad
stats.mstats.normaltest(x1.acc_open_past_24mths)
## Matriz de dispersión con gráfico de linea
##Con diagonal de lineas
scatter_matrix(x1, alpha=0.2, figsize=(6, 6), diagonal='kde')
scatter_matrix(x1, alpha=0.2, figsize=(6, 6), diagonal='hist')
##Con diagonal de histograma
#scatter_matrix(x1, alpha=0.2, figsize=(6, 6), diagonal='hist')
## Box Wiskas
color = dict(boxes='DarkGreen', whiskers='DarkOrange', medians='DarkBlue', caps='Gray')
x1.plot.box(color=color, sym='r+')
plt.figure()
```

```
x1.plot.hist(alpha=0.5)
######## TRANSFORMACION DE VARIABLES #######
####### Imputación de missing
snull=x1.fillna(int(0)) snull
#snull.to_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/snull.csv')
####### Ajuste de valor 0 para logaritmos
scero = snull.replace([0], .01)
scero
#scero.to_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/scero.csv')
####### Logaritmo Natural
logbase1=np.log(scero)
logbase= logbase1.clip(0)
#logbase.to_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/logbase.csv')
logbase.describe()
scatter_matrix(logbase, alpha=0.2, figsize=(6, 6), diagonal='kde')
medias=logbase.mean()
medias
medias1=np.array(medias).astype(np.float)
medias1
desv=logbase.std()
desv
desv1=np.array(desv).astype(np.float)
desv1
####### Estandarizaición
logbase_valor=logbase.values Normaliz = scale(logbase_valor)
Normaliz
pd.set_option('display.width', 150)
```

```
header = ['num_actv_bc_tl', 'total_rev_hi_lim',
'tot_cur_bal', 'annual_inc', 'acc_open_past_24mths'] header Normaliz_1 =
np.array(Normaliz) Normaliz_1
Normaliz_2 = pd.DataFrame(Normaliz_1, columns=['num_actv_bc_tl',
'total_rev_hi_lim', 'tot_cur_bal', 'annual_inc', 'acc_open_past_24mths'])
Normaliz_2
scatter_matrix(Normaliz_2, alpha=0.2, figsize=(6, 6), diagonal='kde')
#Normaliz_2.to_csv('C:/Users/bodyr/Dropbox/Bebitos/Thesis/Bases/Normaliz_2.csv')
stats.mstats.normaltest(Normaliz_2)
com_x=Normaliz_2.describe() com_x
######## COMPONENTES PRINCIPALES #######
####### cálculo de la técnica
pca = PCA(n_components=5)
####### Ajuste del modelo
pca1=pca.fit(Normaliz)
####### Ajuste del modelo
#The amount of variance that each PC explains
var= pca.explained_variance_ratio_ print(var)
####### Ajuste del modelo
#Cumulative Variance explains
var1=np.cumsum(np.round(pca.explained_variance_ratio_, decimals=4)*100)
print (var1)
plt.plot(var1)
####### ********************** Nombrar cada uno de los componenstes
####### cálculo de los estadísticos
value=pca.fit_transform(Normaliz)
value
```

```
matrix=pca.get_covariance()
matrix
parametros=pca.get_params(deep=True)
parametros
inversa=pca.get_precision()
inversa
Eigenvectors=pca.components_
Eigenvectors
eigenvalue=pca.explained_variance_
eigenvalue
singular_values=pca.singular_values_
singular_values
######## GENERACIÓN DE ESPACIO #######
####### cálculos para la generaicón del espacio a partir de la base real
pd.set_option('display.width', 150)
header = ['CP1','CP2','CP3','CP4','CP5' ]
header
datos1 = np.array(value) datos1
Normaliz_3 = pd.DataFrame(datos1, columns=['CP1','CP2','CP3','CP4','CP5' ])
Normaliz_3.describe()
scatter_matrix(Normaliz_3, alpha=0.2, figsize=(6, 6), diagonal='kde')
CP = {} for l in range(1,6): globals()['CP%s' % l] =Normaliz_3[['CP%s' % l]]
#CP=datos[['CPl']]
####### cálculo de posición
globals()['medias_cp%s' % l] =eval('CP%s' % l).mean()
####### cálculo de dispersión
globals()['desv_cp%s' % l] =eval('CP%s' % l).std()
```

```
#medias_sim = pd.DataFrame(medias_cp) #medias_sim
####### cálculo de las simulación
globals()['s%s' % l]=np.random.normal(eval('medias_cp%s' % l),eval('desv_cp%s' % l),10000)
#globals()['s%s' % l]=np.random.normal(0,1,10000) globals()['t%s' % l]=
np.array(globals()['s%s' % l]).astype(np.float) globals()['tt%s' % l] =
pd.DataFrame(eval('t%s' % l), columns=['CP%s' % l])
datosSCP=pd.concat([tt1, tt2, tt3, tt4, tt5],axis=1) datosSCP
scatter_matrix(datosSCP, alpha=0.5, figsize=(8, 8), diagonal='kde')
stats.mstats.normaltest(datosSCP)
datosSCP.describe()
####### Creación del espacio
reg_Normal=np.matmul(datosSCP,Eigenvectors)
####### Renombre de las variables
reg_Normal = pd.DataFrame(reg_Normal, columns=['num_actv_bc_tl',
'total_rev_hi_lim', 'tot_cur_bal', 'annual_inc', 'acc_open_past_24mths']) reg_Normal
stats.mstats.normaltest(reg_Normal)
scatter_matrix(reg_Normal, alpha=0.5, figsize=(8, 8), diagonal='kde')
scatter_matrix(Normaliz_2, alpha=0.5, figsize=(8, 8), diagonal='kde')
reg_Normal.describe() Normaliz_2.describe()
####### Posicionamiento y dispersión, desestandarización
reg_Normal_std=reg_Normal*desv1 reg_Normal_std reg_desNormal =
reg_Normal_std+medias1 reg_desNormal
reg_desNormal.describe()
scatter_matrix(reg_desNormal, alpha=0.5, figsize=(8, 8), diagonal='kde')
logbase.describe()
scatter_matrix(logbase, alpha=0.5, figsize=(8, 8), diagonal='hist')
####### Escalamiento
```

```
reg_desLogNormal= np.exp(reg_desNormal) reg_desLogNormal.describe()
scero.describe()
########## CONTRASTE DE PRUEBA #######
####### análisis descriptivo scatter_matrix(reg_desLogNormal,
alpha=0.5, figsize=(8, 8), diagonal='kde')
scatter_matrix(scero, alpha=0.5, figsize=(8, 8), diagonal='kde')
reg_desLogNormal
#Maximos reg_desLogNormal1 = reg_desLogNormal[(reg_desLogNormal.num_actv_bc_tl
<= 19)] reg_desLogNormal1 = reg_desLogNormal1[(reg_desLogNormal.tot_cur_bal
<= 1800000)] reg_desLogNormal1 =
reg_desLogNormal1[(reg_desLogNormal1.acc_open_past_24mths <= 24)]
#Minimos
reg_desLogNormal1.describe()
x1.describe()
reg_desLogNormal1.info()
####### Graficas bivariadas
sta, critucal, Sl =stats.anderson_ksamp([reg_desLogNormal1['num_actv_bc_tl'],x1['num_actv_bc_tl'
st, pvalue=ks_2samp(reg_desLogNormal1['num_actv_bc_tl'], x1['num_actv_bc_tl'])
sns.kdeplot(reg_desLogNormal1['num_actv_bc_tl'],label="Estimate") sns.kdeplot(x1['num_actv_bc_
#plt.figtext(.5, .65, "Anderson = ") #plt.figtext(.5, .6, format(round (Sl,4) )) plt.figtext(.75,
.65, "K-S =") plt.figtext(.75, .6, format(round (pvalue,4) )) plt.legend();
sta, critucal, Sl =stats.anderson_ksamp([reg_desLogNormal1['total_rev_hi_lim'],x1['total_rev_hi_lim
st, pvalue=ks_2samp(reg_desLogNormal1['total_rev_hi_lim'], x1['total_rev_hi_lim'])
sns.kdeplot(reg_desLogNormal1['total_rev_hi_lim'],label="Estimate") sns.kdeplot(x1['total_rev_hi_li
#plt.figtext(.7, .65, "Anderson = ") #plt.figtext(.75, .6, format(round (Sl,4) )) #plt.figtext(.7,
.55, "K-S =") #plt.figtext(.75, .5, format(round (pvalue,4) )) plt.legend();
sta, critucal, Sl =stats.anderson_ksamp([reg_desLogNormal1['tot_cur_bal'],x1['tot_cur_bal']])
```

```
    st, pvalue=ks_2samp(reg_desLogNormal1['tot_cur_bal'], x1['tot_cur_bal'])
    sns.kdeplot(reg_desLogNormal1['tot_cur_bal'],label="Estimate") sns.kdeplot(x1['tot_cur_bal'],label=
plt.figtext(.7, .65, "Anderson = ") plt.figtext(.75, .6, format(round (Sl,4) )) plt.figtext(.7, .55,
"K-S =") plt.figtext(.75, .5, format(round (pvalue,4) )) plt.legend();
    sta, critucal, Sl =stats.anderson_ksamp([reg_desLogNormal1['annual_inc'],x1['annual_inc']])
    st, pvalue=ks_2samp(reg_desLogNormal1['annual_inc'], x1['annual_inc'])
    sns.kdeplot(reg_desLogNormal1['annual_inc'],label="Estimate") sns.kdeplot(x1['annual_inc'],label="R
plt.figtext(.7, .65, "Anderson = ") plt.figtext(.75, .6, format(round (Sl,4) )) plt.figtext(.7, .55,
"K-S =") plt.figtext(.75, .5, format(round (pvalue,4) )) plt.legend();
    sta, critucal, Sl =stats.anderson_ksamp([reg_desLogNormal1['acc_open_past_24mths'],x1['acc_open_
    st, pvalue=ks_2samp(reg_desLogNormal1['acc_open_past_24mths'], x1['acc_open_past_24mths'])
    sns.kdeplot(reg_desLogNormal1['acc_open_past_24mths'],label="Estimate") sns.kdeplot(x1['acc_ope
plt.figtext(.7, .65, "Anderson = ") plt.figtext(.75, .6, format(round (Sl,4) )) plt.figtext(.7, .55,
"K-S =") plt.figtext(.75, .5, format(round (pvalue,4) )) plt.legend();
    ####### Evaluación en la prueba predictiva
    se_prob = model_2.predict_proba(reg_desLogNormal1)
    pd.set_option('display.width', 150) header = ['pd_0', 'pd_1' ] header pd_1_se = np.array(se_prob)
pd_1_se
    pd_2_se = pd.DataFrame(pd_1_se, columns=['pd_0', 'pd_1' ]) pd_2_se
    sns.kdeplot(pd_2_se.pd_1,label="Probabilidad Simulación") sns.kdeplot(pd_2.pd_1,label="Probabilid
Test Real") plt.show()
    ####### Simulación estresada CP = {} for l in range(1,6):
    globals()['s%s' % l]=np.random.normal(eval('medias_cp%s' % l),eval('desv_cp%s' % l),10000)
s5=np.random.normal(1,1,10000) globals()['t%s' % l]=np.array(globals()['s%s' % l]).astype(np.float)
globals()['tt%s' % l] = pd.DataFrame(eval('t%s' % l), columns=['CP%s' % l])
    datosSCP1=pd.concat([tt1, tt2, tt3, tt4, tt5],axis=1)
    ####### Creación del espacio
```

reg_Normal1=np.matmul(datosSCP1,Eigenvectors)

####### Renombre de las variables

reg_Normal1 = pd.DataFrame(reg_Normal1, columns=['num_actv_bc_tl', 'total_rev_hi_lim', 'tot_cur_bal', 'annual_inc', 'acc_open_past_24mths'])

reg_Normal_std1=reg_Normal1*desv1 reg_Normal_std1 reg_desNormal2 = reg_Normal_std1+media reg_desNormal2

reg_desLogNormal2= np.exp(reg_desNormal2)

reg_desLogNormal2.mean() reg_desLogNormal1.mean()

sns.kdeplot(reg_desLogNormal2['num_actv_bc_tl'],label="Estimate") sns.kdeplot(reg_desLogNormal1 Estres") plt.legend();

reg_desLogNormal2.mean()

reg_desLogNormal1.mean()

sf_prob = model_2.predict_proba(reg_desLogNormal2)

pd.set_option('display.width', 150) header = ['pd_0', 'pd_1'] header pd_1_sf = np.array(sf_prob) pd_1_sf

pd_2_sf = pd.DataFrame(pd_1_sf, columns=['pd_0', 'pd_1']) pd_2_sf

sns.kdeplot(pd_2_sf.pd_1,label="Probabilidad Estres") sns.kdeplot(pd_2_se.pd_1,label="Probabilida Simulada")

sns.kdeplot(pd_2.pd_1,label="Probabilidad Test Real") plt.show()

Anexo II: Diccionario de datos

Variable	**Descripción**
Purpose	Propósito para el uso del crédito
loan_estatus	Estatus del crédito a nivelo motatorio
delinq_2yrs	Incumplimiento de los dos últimos año (al 2018)
num_actv_bc_tl	número de tarjetas de crédito que tiene actualmente el cliente
total_rev_hi_lim	monto máximo que el cliente tiene en la entidad de crédito
tot_cur_bal	Saldo de todas las cuentas que tiene el cliente
annual_inc	Ingreso auto informado por el cliente al momento del otorgamiento
acc_open_past_ 24mths	Adquisidores de créditos nuevos en los últimos 24 meses
funded_amnt	Monto de crédito otrogado en algún momento del tiempo
Installment	Pago del crédito
out_prncp	Monto de capital remanente del monto del crédito otorgado
total_pymnt	Pago total del cliente
total_rec_prncp	Monto de Capital Amortizado
num_rev_accts	Número de cuentas revolventes
Pub_rec	Número de registros publicos malos
Bc_util	Relación entre el saldo y el limite de crédito
Tot_cur_bal	Saldo actual en todas sus cuentas
Total_bal_ex_mort	Saldo total sin hipoteca

Anexo III: Glosario de acrónimos y siglas

Palabra	**Significado**
SEMMA	Siglas en inglés. Sample, Emplore, Modify, Model and Assess.
CRISP-DM	Acronimo en inglés. Cross Industry Standard Process for Data Mining
UIAF	Sigla. Unidad de Información y Análisis Financiero
MB	Acronimo. Megabytes
CODASYL	Acronimo. Conference on Data Systems Languages
IDS	Siglas en inglés. Integrated Data Store
SQL	Siglas en inglés. Structured Query Language
BD	Sigla. Base de datos
SAS	Siglas en inglés. Statistical Analysis System
WEKA	Siglas en inglés. Waikato Environment for Knowledge Analysis
SPSS	Siglas en inglés. Statistical Package for the Social Sciences
IFAI	Acronimo. Instituto Federal de Acceso a la Información
LFPDP	Acronimo. La Ley Federal de Protección de Datos Personales
VaR	Sigla. Value At Risk

Palabra	**Significado**
C1	Acronimo. Cuartil 1
C3	Acronimo. Cuartil 3
GML	Sigla. Modelos Lineales Generalizados
ACP	Sigla. Análisis de Componentes Principales
PCA	Sigla. Principal component analysis
KDD	Sigla. Knowledge discovery in databases
ROC	Sigla en inglés. Receiver Operating Characteristic
BANXICO	Acronimo. Banco de México
BIS	Sigla en inglés. Bank for International Settlements
PD	Siglas en inglés. Probability of Default
LGD	Siglas en inglés. Loss Given Default
EAD	Siglas en inglés. Exposure At Default
IFRS	Siglas en inglés. International Financial Reporting Standard
HSBC	Sigla en inglés. The Hong Kong and Shanghai Banking Corporation
BBVA	Sigla. Banco Bilbao Vizcaya Argentaria
P2P	Siglas en inglés. Peer tú Peer

Anexo IV: Programa de generación de Normales aleatorias

```
"""Generación de variables aleatorias uniformes en generación inversa de una función normal
Autor: Juan Fabrizio Sanchez Jimenez """
import numpy as np
import matplotlib.pyplot as plt
def Circulo(N):
thetas = np.random.uniform(0,1,N)
alpha = np.random.uniform(0,1,N)
x = (-2*np.log(thetas))**(1/2)*np.cos(2*np.pi*alpha)
y = (-2*np.log(thetas))**(1/2)*np.sin(2*np.pi*alpha)
plt.subplot(1, 2, 1)
plt.scatter(thetas,alpha)
plt.xlabel("Uniforme X")
plt.ylabel("Uniforme Y")
plt.subplot(1, 2, 2)
plt.scatter(x,y)
plt.xlabel("Normal X")
plt.ylabel("Normal Y")
plt.subplot(2, 2, 1)
```

```
plt.hist(x)
plt.xlabel("Normal X")
plt.subplot(2, 2, 2)
plt.hist(y)
plt.xlabel("Normal Y")
x1=Circulo(10)
x1=Circulo(100)
x1=Circulo(1000)
x1=Circulo(10000)
```

Índice de cuadros

Índice de figuras

Bibliografía

Edward Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 1968.

Edward Altman. Default recovery rates and lgd in credit risk modeling and practice: an updated review of the literature and empirical evidence. *Stern School of Business*, 2004.

Anaconda. Anaconda distribution, 2018. URL `https://www.anaconda.com/distribution/`.

Tom Augspurger, Chris Bartak, Phillip Cloud, Andy Hayden, Stephan Hoyer, Wes McKinney, Jeff Reback, Chang She, Masaaki Horikoshi, y Joris Van den Bossche. Pandas, 2018. URL `https://pandas.pydata.org/`.

Ana Azevedo y Manuel Filipe Santos. Kdd, semma and crisp-dm: a parallel overview. *KDD*, 2008.

Carlos E. Azofeifa. Aplicacion de la simulacion monte carlo en el calculo del riesgo usando excel. *Tecnologia en marcha*, 2004.

Berenice Baez-Revueltas y Rafael G. Gamboa-Hirales. Generation of linearly representative samples. In *El seminario aleatorio*, 2013.

BBVA Bancomer. *Reporte sobre la solvencia y condicion financiera 2017*. Grupo Financiero BBVA Bancomer, 2017.

Banorte. *Politicas y metodologias para la administracion de riesgos*. Grupo Financiero Banorte, 2015.

BANXICO. *Definiciones basicas de riesgos*. BANXICO, 2005.

BBVA, 2014. URL `https://accionistaseinversores.bbva.com/microsites/bbvain2013/es/G/c2.html`.

BBVA, 2016. URL `https://www.bbva.com/es/ las-claves-entender-lenguaje-moda-python/`.

BBVAOpen4U, 2016. URL `https://bbvaopen4u.com/es/actualidad/ventajas-e-inconvenientes-de-python-y-r-para-la-ciencia-de-datos`.

BIS. *International convergence of capital measurement and capital standards*. Basle committee on banking supervision, 1988.

BIS. *Vision general del nuevo acuerdo de capital de basilea*. Comite de supervision bancaria de basilea, 2001.

BIS. *El Nuevo Acuerdo de Capital de Basilea*. Comite de supervision bancaria de basilea, 2004.

Fischer Black y Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 1973.

Fischer Black, Emanuel Derman, y Bill Toy. Impact of different interest rate models on bond value measures. *Financial Analysts Journal*, 1990.

Maria Bohdalova y Michal Gregus. Monte carlo simulation value at risk and PCA. *VEGA*, 2010.

Catalina Bolance, Montserrat Guillen, y Alemar Padilla. Estimacion del riesgo mediante el ajuste de copulas. *Research group on risk in insurance and finance*, 2015.

Emmanuel Malagon Bolanos. *Diseno de software para el calculo del var*. PhD thesis, UNAM, 2010.

Phelim Boyle. Options a montecarlo approach. *Journal of Financial Economics*, 1976.

Joseph L Breeden. Monte carlo scenario generation for retail loan portfolios. *Strategic Analytics*, 2008.

Joseph L Breeden y David Ingram. Portfolio forecasting tools what you need to know. *Retail Risk*, 2003.

Andrew Bruce. Using predictive models to improve loan portfolios. *turi*, 2015.

John Chadam, Joel Hanson, y Yuriy Kazmerchuk. Monte carlo simulation in the integrated market and credit portfolio model. *Mmaths in industry*, 2001.

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, y RÃŒediger Wirth. *Crisp dm 1 0*. SPSS, 2000.

Gonzalo Chavez y Paul Zanabria. Simulacion de curvas de rendimiento empleando componentes principales: una aplicacion para los fondos de pensiones. *PensionesBanco Central de Reserva del Peru*, 2018.

Liliana Cruz Arrela. *Mineria de datos con aplicaciones*. PhD thesis, UNAM, 2010.

Gabriel Juarez Delgado. *Evaluacion de la conversion a cafe organico usando la metodologia de opciones reales*. PhD thesis, UNAM, 2010.

Javier Marquez Diez-Canedo, Carlos E. Nogues Nivon, y Viviana Velez Grajales. Un metodo eficiente para la simulacion de curvas de tasas de interes. *BANCO DE MEXICO*, 2003.

Arturo Erdely. Copulas y dependencia de variables aleatorias: una introduccion. *Miscelanea Matematica*, 2009.

Usama Fayyad, Gregory Piatetsky Shapiro, y Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 1996.

Roberto Garcia Reyes. *Mineria de datos para la toma de decisiones e inteligencia de negocios: aplicaciones a la mercadotecnia*. PhD thesis, UNAM, 2012.

Mike Giles. Advanced monte carlo methods: quasi-monte carlo, 2018.

glassdoor, 2018. URL `https://www.glassdoor.com/Salaries/python-developer-salary`.

Anibal Goicochea. Crisp-dm, una metodologÃa para proyectos de mineria de datos, 2017. URL `https://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de-datos/`.

Google, 2018. URL `https://trends.google.com/trends/`.

Ybnias Eli Grijalva. *Metodos cuantitativos para los negocios*, chapter Introduccion al metodo de simulacion monte carlo. Facultad de ciencias administrativas y contables, 2009.

Guerrero, 2018. URL `http://transparencia.guerrero.gob.mx/historia-de-la-transparencia/`.

Mahil Herrera Maldonado. Pruebas de bondad de ajuste, 2008.

HSBC. *Politica general, metodologia e informacion cuantitativa*. Grupo Financiero HSBC, 2013.

IFRS. *Instrumentos Financieros*. NII9, 2015.

Banco Inbursa. Banco Inbursa. Technical report, Banco Inbursa, 2017.

Kunal Jain, 2017. URL `https://www.analyticsvidhya.com/blog/2017/09/sas-vs-vs-python-tool-learn/`.

I.T. Jolliffe. *Principal component analysis*. Springer, 2002.

Alexander Kreinin. Monte carlo simulation in the integrated market and credit portfolio model. *Algorithmics Inc*, 2001.

Alexander Kreinin, Leonid Merkoulovitch, Dan Rosen, y Michael Zerbs. Principal component analysis in quasi Monte Carlo simulation. *ALGO RESEARCH QUARTERLY*, 1998.

Van Son Lai, Youssef Sakni, y Issouf Soumare. A simple method for computing value at risk using PCA and QMC. *Journal of financial decision making*, 2005.

Bank Lending Club. Lending club statistics, 2018. URL `https://www.lendingclub.com/info/download-data.action`.

Raul E. Lopez Briega. Introduccion a los metodos de monte-carlo con python, 2017. URL `https://relopezbriega.github.io/ blog/2017/01/10/ introduccion-a-los-metodos-de-monte-carlo-con-python/`.

Mico Loretan. Generate market risk scenarios using principal components analysis metodogical and parctical considerations. *Federal Reserve Board*, 1997.

Xiaolin Luo y Pavel V. Shevchenko. Markov chain Monte Carlo estimation of default and recovery: dependent via the latent systematic factor. *CSIRO Mathematics*, 2013.

Oded Maimon y Lior Rokach. *Data mining and knowledge discovery handbook*. Springer, 2010.

Maria Jose Marques. *Estadistica basica un enfoque no parametrico*. UNAM, 2018.

Mercedes Marques. *Bases de datos*. PhD thesis, Universitat jeume i de castello, 2009.

Juan Miguel Moine, Ana Silvia Haedo, y Silvia Gordillo. Estudio comparativo de metodologias para mineria de datos. *Grupo de investigacion en Mineria de Datos*, 9999.

Luis E. Nieto Barajas. EstadÃstica y probabilidad, 2018.

Soraida Nieto Murillo. *Credito al consumo: la estadistica aplicada a un problema de riesgo crediticio*. PhD thesis, UAM, 2010.

Hector Oscar Nigro, Daniel Xodo, Gabriel Corti, y DamiÃ¡n Terren. Kdd knowledge discovery in databases: un proceso centrado en el usuario. *Departamento de computacin y sistemas facultad de ciencias exactas unicen*, 2018.

Erik Nylen y Pascal Wallisch. *Neural data science*. Elsevier, 2017.

Emilio David Olvera y Joshua Zenteno Jimenez. *Un comparativo entre las metodologias de optimizacion de portafolios de inversion entre el modelo de markowitz y el metodo de simulacion monte carlo con acciones pertenecientes al ipc 2007 2012*. PhD thesis, UAEM, 2013.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, y E. Duchesnay. Scikit-learn: machine learning in Python, 2018. URL `http://scikit-learn.org/stable/`.

Daniel Pena. *Analisis de datos multivariantes*. NA, 2002.

Cesar Perez. *Mineria de datos: tecnicas y herramientas*. Editorial Paraninfo, 2007.

Josef Perktold, Skipper Seabold, y Jonathan Taylor. Statsmodels, 2018. URL `https://www.statsmodels.org/stable/index.html`.

Gregory Piatetsky-Shapiro. Knowledge discovery in real databases. *AI Magazine*, 1990.

Gregory Piatetsky-Shapiro, 2008-2018. URL `https:/ /www.kdnuggets.com/ polls/ 2007/ data mining softwaretools.htm https:/ /www.kdnuggets.com/ polls/ 2008/ data mining software tools used.htm https:/ /www.kdnuggets.com/ polls/ 2009/ data mining tools used.htm https:/ /www.kdnuggets.com/ polls/ 2011/ languages for data mining analytics.html https:/ /www.kdnuggets.com/ polls/ 2012/ analytics data mining programming languages.html https:/ /www.kdnuggets.com/ polls/ 2013/ languages analytics data mining data science.html https:/ /www.kdnuggets.com/ polls/ 2014/ languages analytics data mining data science.html https:/ /www.kdnuggets.com/ 2016/ 06/ r python top analytics data mining data science software.html https:/ /www.kdnuggets.com/ 2017/ 08/ python overtakes r leader analytics data science.html https:/`

/www.kdnuggets.com/ 2018/ 05/ poll tools analytics data science machine learning results.html https:/ /www.kdnuggets.com/ 2015/ 05/ r vs python data science html.

Python. Beginners guide, 2018. URL https://wiki.python.org/moin/BeginnersGuide/Overview.

Fitch Ratings. Banco Azteca. Technical report, Fitch Ratings, 2018.

Philippe Rigollet. Statistics for applications, 2016.

Luis Rincon. *Curso elemental de probabilidad y estadistica*. Departamento de Matematicas, 2007.

RiskMetrics. CreditMetrics. Technical report, RiskMetrics, 2007.

Gustavo R. Rivadera. La metodologia de Kimball para el diseno de almacenes de datos (data warehouses). *Data werwhouse*, 2010.

Carlos Daniel Rodriguez. *Programacion de metodos numericos en python aplicados a problemas de ingenieria quimica*. PhD thesis, UNAM, 2015.

David Esteban Rodriguez y Alfredo Trespalacios. Medicion de valor en riesgo en cartera de clientes a traves de modelos logisticos y simulacion de montecarlo. *Universidad EAFIT*, 2015.

Licesio J. Rodriguez-Aragon. *Simulacion, metodo de montecarlo*. Area de Estadistica e Investigacion Operativa, 2011.

Carlos Daniel Rodriguez Sotelo. *Programacion de metodos en python aplicados a problemas de ingenieria quimica*. PhD thesis, UNAM, 2015.

Leslie Jimenez Rosas y Guillermo Benavides. Stress testing para carteras de credito del sistema bancario mexicano. *Revista Mexicana de Economia y Finanzas*, 2015.

Jonatha Ruiz Rangel. Mineria de datos como soporte a la toma de decisiones empresariales en una arquitectura soa. *Coruniamericana*, 2013.

Santander. *Informe anual administracion riesgos*. Grupo Financiero Santander, 2017.

SAS. *Data mining and the case for sampling*. SAS Institute, 1998.

Enrique Eduardo Tarifa. Teoria de modelos y simulacion. *Universidad Nacional de Jujuy*, 2018.

Marco Ricardo Tellez. *Medicion del riesgo en credito implementacion y calculo del VaR y el CVaR en tres modelos de incumplimiento.* PhD thesis, UAM, 2010.

Stephane Tuffery. *Data mining and statistics for decision making*. Wiley, 2008.

UIAF. Tecnicas de mineria de datos para la deteccion y prevencion del lavado de activos y la nanciacion del terrorismo. *Documentos UAIF*, 2014.

Michael Waskom. Seaborn, 2017.

Karlijn Willems, 2015. URL `https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis`.

Qing Zhao. Lending club data analysis and default rate prediction. *HARVARD UNIVERSITY CS109*, 2015.

ziprecruiter, 2018. URL `https://www.ziprecruiter.com/Salaries/R-Developer-Salary`.