Universidad Nacional Autónoma de México

Facultad de Ciencias

Aplicación de un modelo clasificatorio para renovación de pólizas de seguro de vida individual

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

Actuario

PRESENTA:

Juan Carlos Badillo Martínez

DIRECTOR:

Fernando Herrera Contreras

Ciudad Universitaria, 2019
Cd. Mx.

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

A mis padres, hermanos y amigos por todo el apoyo recibido antes, ahora y siempre.
A los actuarios Fernando Herrera y Fernando Pérez por todo el apoyo recibido tanto en el terreno profesional como en el personal.
A Naomi por motivarme a retomar mis estudios.

Declaración de autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Juan Carlos Badillo Martínez. Ciudad Universitaria, 2019

Resumen

Desde mediados del siglo pasado los bancos han aplicado modelos matemáticos para discriminar a los clientes que pagarán un crédito de aquellos que no lo harán (default), este tipo de modelos pueden aplicarse por otro tipo de instituciones para hacer una selección y clasificación de sus clientes.

En esta tesis se utilizarán métodos estadísticos para analizar y seleccionar las variables que describen el comportamiento de los asegurados para posteriormente utilizando dichas variables costruir un modelo de regresión logística que permita clasificar a dichos clientes en dos grupos: Personas que cancelaran sus pólizas de seguro y personas que no cancelarán, ofreciendo también una introducción a las particularidades del seguro de vida que por su naturaleza a largo plazo es susceptible a cancelación por parte de los asegurados

Índice general

Índice de figuras

Índice de tablas

Introducción.

Un modelo clasificatorio es un modelo estadístico que permite separar a una población en grupos con respecto a una variable de interés, esta variable usualmente es una variable categórica e incluso puede ser una variable binaria. Para realizar esta clasificación se utilizan diversos métodos estadísticos que se apoyan en información conocida de los miembros de la población para asociar a estos con un grupo.

Desde la década de 1940 la banca ha utilizado modelos clasificatorios para analizar el comportamiento de sus clientes, en particular una de sus principales necesidades ha sido clasificar entre los solicitantes de crédito a aquellos que pagarán sus créditos de quienes no lo harán.

Asumiendo que las conductas del pasado se repetirán en el futuro los datos obtenidos de los clientes permiten a los otorgantes de crédito distinguir las características deseadas en sus contratantes de crédito, esta capacidad ayudó a la expansión masiva en los créditos revolventes para consumo permitiendo a los consumidores acceder al crédito como una herramienta que hace su vida más sencilla y productiva.

Breve historia de los modelos clasificatorios

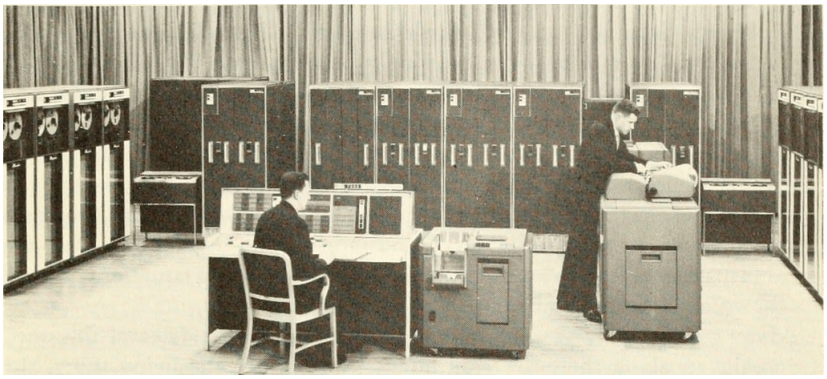
- En 1935 Sir Ronald Aylmer Fischer publicó un artículo sobre una técnica llamada "Análisis lineal discriminatorio" para clasificar especies correspondientes a plantas del género Iris.
- En 1941 David Durand mostró que la misma técnica puede ser utilizada para discriminar entre buenos y malos negocios, con datos sobre 7200 préstamos buenos y malos utilizando variables referentes a la edad y ocupación de los contratantes.
- En 1956 fue fundada Fair Issac Corporation la primer compañía en brindar servicios de puntaje crediticio.

- En 1963 Fair Issac obtuvo un contrato con Montgomery Ward para ofrecer créditos, con la llegada de las computadoras pudo tener una oficina de crédito en cada tienda, como describe Lewis (Lewis, 1994 [18]) una de las operaciones de crédito más eficientes y sin par hasta varios años después.

  Debido al éxito de las oficinas de crédito de MW Macy's, Gimbel's, Bloomingdale's y J.C. Penney siguieron sus pasos implementando sus propios modelos.

  Durante estos años la adopción de modelos estadísticos fue retrasada por dos factores, la resistencia de las organizaciones al uso de máquinas para tomar decisiones y lo complejo que eran los cálculos para el poder de cómputo de la época, una IBM 7090 de 1963 sólo tenía memoria suficiente para analizar los datos de 600 contratantes a la vez, en contraposición era grande, se calentaba y requería una habitación especial libre de polvo para su uso.

Figura 1: IBM 7090



Fuente: Biophysics, concepts and mechanisms (1962), Internet archive book images

  Conforme el costo y velocidad de las computadoras fue siendo más atractivo, se justificó la implementación de modelos para otro tipo de créditos, durante las décadas de 1970 y 1980 se aplicaron modelos de puntaje en créditos personales, créditos automotrices, incluso para pequeños negocios.

- En 1962 Cyert, Davidson y Thompson presentaron el primer trabajo académico para modelar la probabilidad de los modelos de puntaje crediticio, describiendo el problema como una cadena de Markov.
- A partir de la década de 1980 se fundan las primeras sociedades de información crediticia (Burós de crédito) para concentrar previa autorización de los clientes información sobre sus pagos a diversas instituciones permitiendo mejorar la precisión de los modelos añadiendo variables como el monto total de deuda y el impago a otras instituciones.
- Durante la década de 1990 los modelos de puntaje crediticio llegan a Latinoaméri-

ca, Miller y Rojas (M. Miller, 2004 [19]) apuntan que uno de los principales problemas de aplicación en la región, fue la falta de datos sobre los contratantes de crédito por lo que los primeros modelos tuvieron que utilizar la información conjunta de diversas instituciones para tener mayor confianza sobre los mismos.

- A inicios del milenio se comenzó a utilizar este tipo de modelos para analizar el riesgo de impago para compañías pequeñas y medianas

  La técnica estadística utilizada en los primeros modelos fue el Análisis discriminatorio y modelos lineales de probabilidad, actualmente y gracias al desarrollo de software estadístico el modelo más utilizado es la regresión logística aunque se han hecho avances en Redes neuronales buscando mejorar la precisión de los modelos.

Objetivo.

El objetivo de este trabajo es:

- Aplicar un modelo clasificatorio (regresión logística) a una parte de la experiencia mexicana en seguros de vida, con datos obtenidos del Reporte Regulatorio número 8 (RR-8) de la Comisión Nacional de Seguros y Fianzas (CNSF).

  En específico:

- Analizar los factores de mayor influencia en la decisión de los asegurados sobre la cancelación de pólizas en el seguro de vida individual.
- Ajustar un modelo de regresión logística utilizando las variables obtenidas en el punto anterior que permita distinguir entre los usuarios que cancelarán sus pólizas de los que no.
- Realizar un backtesting comparando la decisión que tomaron asegurados reales contra la predicha por el modelo utilizando una muestra que no fue considerada en el ajuste del modelo, para verificar la eficacia del mismo.

Motivación.

En los seguros el término Reserva se refiere a un fondo que la aseguradora debe mantener para cumplir con las obligaciones que tendrá con sus asegurados por la realización del riesgo. En el caso del seguro de vida estas reclamaciones siempre ocurren, excepto si el seguro es cancelado, a diferencia de otras operaciones del seguro en las que el riesgo cubierto puede o no realizarse. El conocimiento o estimación anticipada de las pólizas

canceladas le pueden permitir a una aseguradora ajustar el cálculo de sus reservas matemáticas para no sobre estimar o subestimar el valor de estas ofreciendo al público costos acordes al riesgo y tipo de seguro.

Por otro lado el artículo 237 de la Ley de Instituciones de Seguros y Fianzas (LISF) brinda a las aseguradoras la posibilidad de utilizar modelos internos para el cálculo de su Requerimiento de Capital de Solvencia por lo que el tener conocimiento sobre si la póliza será de corto o largo plazo le puede permitir a la aseguradora tener un estimado más preciso de sus obligaciones futuras con la posibilidad de reducir su cálculo de este capital respecto a la fórmula general propuesta por la ley.

Metodología.

Para cumplir con los objetivos de este trabajo se analizarán las pólizas de seguros de vida individuales reportado en el RR-8 de la CNSF durante el periodo de 2016. Utilizando el estadístico Kolmogórov-Smirnov, IV y WOE se elegirán las variables significativas para describir el comportamiento de los asegurados respecto a sus pólizas de seguro de vida y una vez elegidas las variables elegidas se ajustará un modelo de regresión logística que intentará explicar dicho comportamiento. Finalmente se compararán los resultados predichos por el modelo respecto a datos reales obtenidos de una muestra reservada que no será utilizada en la construcción del modelo y utilizará datos observados durante 2017.

Estructura de la tesis.

El presente trabajo desarrollará en cuatro grandes apartados:

- En el primer apartado se presenta una introducción y la motivación para desarrollar este proyecto; así como las características y particularidades del seguro de vida, que por su naturaleza de cobertura de largo plazo es susceptible a la de cancelación por parte de los asegurados. (Introducción y Capitulo 1).
- En el segundo apartado, se detalla el marco teórico referente al modelo clasificatorio logístico y la selección adecuada de las variables. (Capítulos 2 y 3)
- En el tercer apartado se presenta la aplicación de métodos estadísticos (IV, WOE y estadístico KS) para analizar las variables disponibles y seleccionar cuales son capaces de describir el comportamiento de los asegurados. Posteriormente utilizando las variables seleccionadas, se propone un modelo de regresión logística que permita clasificar a las pólizas en dos grupos: Pólizas que serán canceladas y pólizas que no lo serán. (Capítulo 4)

- Finalmente, en el cuarto apartado se presentan los resultados y se verifica la capacidad predictiva del modelo utilizando datos que no intervinieron en la construcción del mismo (es decir, una prueba backtesting); así como las conclusiones y hallazgos encontrados. (Capítulo 5 y Conclusiones)

Capítulo 1

Características generales de los seguros de vida.

Antes de ahondar en el desarrollo estadístico de esta tesis es conveniente desarrollar conceptos básicos sobre la operación de seguros más en particular sobre el seguro de vida para comprender los objetivos de la misma.

Los individuos y las organizaciones se enfrentan a pérdidas causadas por eventos aleatorios, eventos cuya, frecuencia, magnitud y ocurrencia no están bajo el control del afectado. Históricamente se ha buscado reducir el impacto financiero de estos eventos para que no causen grandes daños en la economía de los individuos y organizaciones, con este fin y de manera progresiva se han creado los seguros.

En el contexto de los seguros, a estos eventos se les llama riesgo, un riesgo no solo se refiere a un evento dañino, en general podemos definir al riesgo como el término jurídico que asume al mismo tiempo la configuración fortuita y del costo que se identifica con un evento aleatorio del cual puede originarse un daño (perjuicio económico) y en la promesa consiguiente de una prestación financiera con referencia al objeto de la garantía.

De la definición anterior se destaca que que término riesgo tradicionalmente asociado a un evento dañino se extiende para contemplar eventos que no necesariamente puedan considerarse dañinos, por ejemplo, pensiones y seguros educativos.

Un seguro es una operación mediante la cual una parte, el asegurado se hace prometer para si mismo o para un tercero en caso de la realización de un riesgo contemplado en el contrato (póliza) una prestación por la otra parte, el asegurador quien asumiendo un conjunto de riesgos los compensa conforme a las leyes de la estadística (Rodríguez, 1976 [23]).

Debido a que la muerte de una persona es un riesgo al que todos estamos expuestos y a que no es posible ponerle precio a la vida humana, los seguros de vida buscan cubrir

las pérdidas económicas producidas por la muerte de una persona o bien celebrar la supervivencia del individuo, en ambos casos el componente aleatorio está en función de la mortalidad del mismo.

1.1. Funciones de Utilidad.

Para analizar el pago de primas por parte de un asegurado, primero hay que analizar el comportamiento de los asegurados frente a las pérdidas, un asegurado tiene 3 posturas posibles frente al riesgo:

- **Reducción**: Prevenir o minimizar el riesgo.
- **Retención**: En esta situación se decide asumir o confrontar el riesgo, en general los que sean potencialmente pequeños y no afecten el patrimonio.
- **Transferencia**: Es traspaso del riesgo a un tercero por medio de un contrato de seguro, esto sucede cuando en caso de ocurrir el riesgo, el patrimonio del asegurado pudiera verse gravemente afectado.

Por lo anterior, podemos decir que el seguro se encargará de cubrir riesgos que puedan afectar de manera significativa el patrimonio de un individuo u organización. Este análisis puede hacerse desde el punto de vista de la utilidad que genera el patrimonio para el individuo.

Una función de utilidad $u(w)$ es una función que dado un patrimonio w nos indica que tanta satisfacción tiene su poseedor, en caso de que dicho patrimonio esté en términos monetarios cumple las siguientes propiedades:

- **Monótonicidad creciente**: Esta propiedad nos dice que un monto mayor genera una mayor utilidad para un individuo.

$$w_1 \leq w_2 \Rightarrow u(w_1) \leq u(w_2) \tag{1.1}$$

- **Insaciabilidad local**: Dado un monto w existe un monto ligeramente mayor que mejora la utilidad $u(w)$, es decir la satisfacción del individuo siempre se puede mejorar.

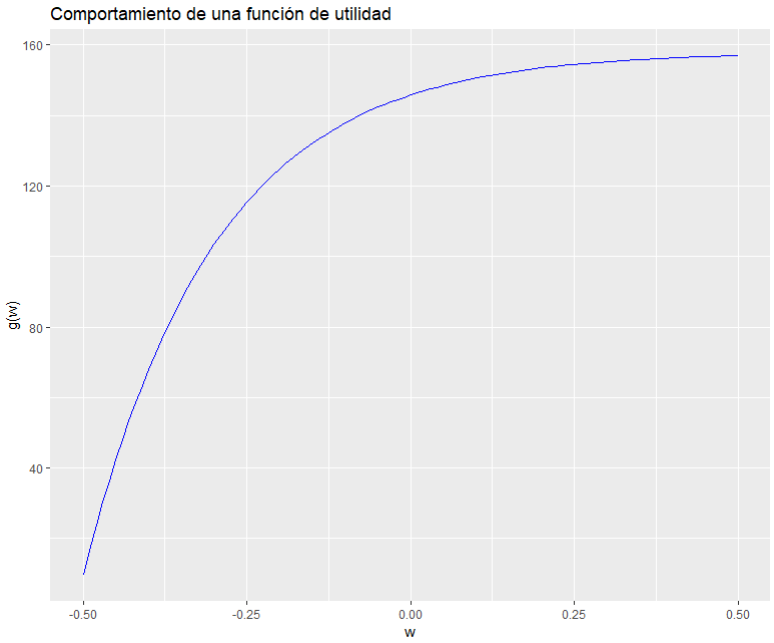
Dados w_1 y $\epsilon \geq 0$ existe w_2 tal que:

$$|w_1 - w_2| \leq \epsilon \rightarrow u(w_1) < u(w_2) \tag{1.2}$$

- **Concavidad**: Conforme más riqueza tenga un individuo su satisfacción crecerá menos al aumentar esta por un monto fijo.

$$u''(w) \leq 0 \tag{1.3}$$

Figura 1.1: Comportamiento gráfico para una función de utilidad



Fuente: Elaboración Propia con datos tomados de Rincon, 2012 [22]

Un tomador de decisiones racional debe de considerar la utilidad esperada para cada opción analizada, comparando sus resultados y seleccionando el que genere un mayor monto, si las opciones seleccionadas tienen la misma utilidad entonces la decisión dependerá de que tan a que tan averso al riesgo es el tomador de decisiones.

1.2. Primas de seguro.

En el ámbito de los seguros, la prima es el costo del seguro es decir, la aportación económica que ha de pagar un asegurado o contratante a una compañía aseguradora por la transferencia del riesgo bajo las coberturas que esta última ofrece a sus clientes durante un determinado período. El valor de esta debe ser tal que le permita a la aseguradora enfrentar sus obligaciones con los asegurados (pago de siniestros), el pago a los agentes por conseguir el negocio, los gastos administrativos de la compañía y además se esperaría que esta tuviera utilidades.

Una prima de seguro justa para el asegurado debería ser aquella que no le cause una perdida en caso de que ocurra el siniestro, es decir en caso de ocurrir el siniestro la riqueza del asegurado debería mantenerse, en términos de utilidad se puede expresar de la siguiente manera:

$$u(w - \mathcal{P}) = E[u(w - X)] \tag{1.4}$$

Donde:

$\mathcal{P}$ Es el costo de la prima.

w Es la riqueza del asegurado.

X Es una variable aleatoria que indica las pérdidas generadas por el siniestro en caso de ocurrir.

Intuitivamente $\mathcal{P} = E[u(X)]$ debería ser una prima justa. Sin embargo, este tipo de prima causaría pérdidas para la aseguradora debido a que que además de pagar a sus asegurados por los riesgos que se realicen debe pagar a sus trabajadores, la infraestructura que requiere para operar y el pago de comisiones correspondiente con la adquisición del seguro.

Para comprobar que la prima $\mathcal{P}$ de la ecuación 1.4 es justa nos apoyaremos con la desigualdad de Jensen [1], dicha desigualdad aplicada al contexto de una función de utilidad nos dice que:

$$u(E[X]) \geq E[u(X)] \tag{1.5}$$

Aplicando la desigualdad de Jensen a la expresión 1.4 tenemos que:

$$u(w - \mathcal{P}) = E[u(w - X)] \leq u(w - E[X]) \tag{1.6}$$

$$u(w - \mathcal{P}) \leq u(w - E[X]) \tag{1.7}$$

Como $u(x)$ es una función creciente por las propiedades expuestas en la sección 2.1 entonces:

$$\mathcal{P} \geq E[X] \tag{1.8}$$

El resultado anterior justifica que el costo de la prima de seguro sea superior a la pérdida esperada por el asegurado debido a la realización del riesgo permitiéndole a la aseguradora obtener los recursos extra necesarios para su operación.

1.3. El seguro de vida.

Hasta el momento se han mencionado características generales de los seguros, el seguro de vida tiene particularidades que lo diferencian de otras operaciones de seguros. En

[1] Se puede encontrar una demostración de esta desigualdad en el Apéndice A

particular la duración del mismo y como este requiere de que la compañía aseguradora acumule reservas para hacer frente a sus obligaciones con los asegurados.

El seguro de vida individual puede ser una transacción difícil de comprender en el sentido de que no se conoce el momento en el que una persona morirá. Sin embargo, haciendo un análisis de la población se puede prever cuantas personas morirán del grupo, tomando en cuenta la historia demográfica ya sea de la aseguradora o de un grupo poblacional más grande. Por lo anterior, en el seguro de vida podemos ver reflejado el principio de que un gran grupo de personas aceptan cooperar contra los trastornos económicos producidos por la muerte.

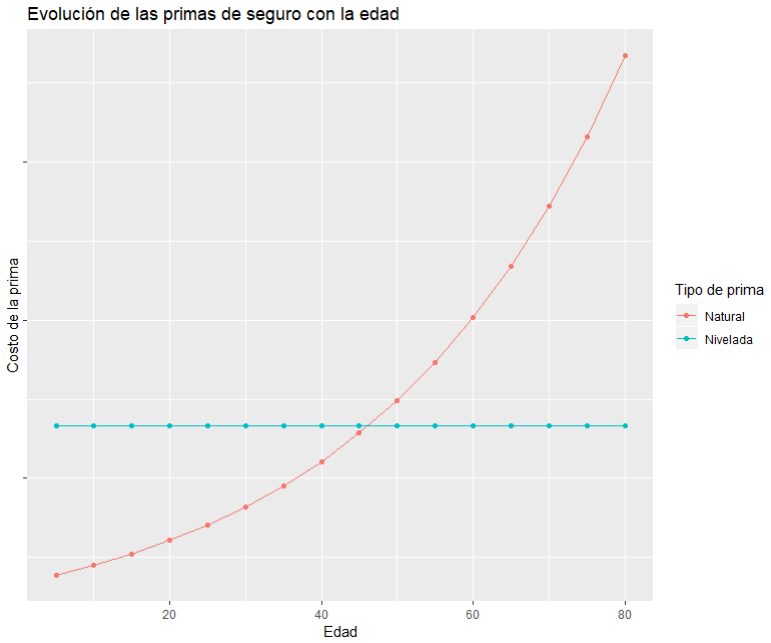
1.3.1. Cálculo de primas para el seguro de vida individual.

El seguro de vida en general es un seguro a largo plazo, la razón de ser de esto es la siguiente:

Tomemos el supuesto de que el seguro de vida fuera de renovación anual, si una persona quisiera asegurarse tendría que renovar cada año su póliza con distintos costos a los que llamaremos prima natural, ya que la probabilidad de morir aumenta con el tiempo, eso causaría que los seguros a edades tempranas fueran muy atractivos mientras que los seguros a edades avanzadas serían mucho más caros resultando incosteables o con sumas aseguradas muy pequeñas.

Por lo anterior, conviene hacer al seguro de vida de largo plazo para compensar los altos pagos de la edad avanzada con bajos pagos a edades menores, a una prima de este tipo se le llama prima nivelada.

Figura 1.2: Prima nivelada contra prima natural a través del tiempo



Fuente: Elaboración Propia. con datos de Bowers, 1997 [7]

El excedente que representa el ejemplo anterior es lo que en seguro de vida se conoce como reserva, desde el punto de vista de la aseguradora no tiene el significado tradicional de la palabra en el que reserva es una previsión para el futuro, debido a que en este caso, al ser un cobro adicional, la reserva mas bien es un fondo que permitirá a la aseguradora cubrir riesgos futuros es decir, un pasivo contable que será analizado más a fondo en la siguiente sección.

Para calcular la prima nivelada equivalente a la prima natural debemos hacer los siguientes supuestos:

- Tenemos una población estática (cohorte) de l_x asegurados con la misma edad.
- Contamos con un análisis sobre la mortalidad de la población y sabemos cuantas. personas pueden morir en n años después de iniciado el análisis, a esta cantidad de decesos se le representará con d_{x+n}.
- La aseguradora invierte sus recursos a una tasa fija de interés i.
- Las personas viven hasta una edad máxima w.

Para el cálculo de la prima pura[2] natural para el primer año, se tomará el pago esperado a los asegurados por parte de la aseguradora, supongamos que del grupo de asegurados

[2]Sin contemplar los gastos de operación y adquisición

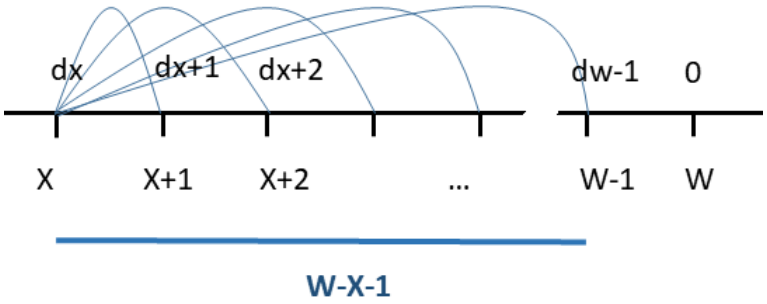
se espera que en el año x mueran d_x asegurados del total l_x a los que se les pagará una suma asegurada de \$1.00, las obligaciones de la aseguradora a final de año[3] serían de d_x pesos es decir a principio de año la aseguradora debería recibir:

$$\mathcal{P} = vd_x \tag{1.9}$$

Donde $v = (1+i)^{-1}$

Si extendemos este razonamiento a los años siguientes el total de dinero que la aseguradora recibirá por parte de los asegurados en valor presente será:

Figura 1.3: Valor Presente para los pagos de primas puras naturales de los asegurados



Fuente:Elaboración Propia.

$$\mathcal{P}_{total} = vd_x + v^2 d_{x+1} + v^3 d_{x+2} + \cdots + v^{w-x} d_{x+(w-x-1)} \tag{1.10}$$

Si todos los asegurados al inicio del primer año decidieran hacer un pago único A_x la aseguradora debería recibir en total:

$$\mathcal{P}_{total} = l_x A_x \tag{1.11}$$

Las expresiones 1.10 y 1.11 nos dan un cálculo del monto total que recibiría una aseguradora por los seguros contratados en el año x dada una población de tamaño l_x y estas deben ser equivalentes ya que lo que varía es el tipo de pago, así que se pueden igualar.

$$l_x A_x = vd_x + v^2 d_{x+1} + v^3 d_{x+2} + \cdots + v^{w-x} d_{x+(w-x-1)} \tag{1.12}$$

[3] Si bien para el cálculo se asume que el pago es a final del año en la práctica el pago se hace cuando la aseguradora confirma la muerte del asegurado

Por lo que el pago único que debería hacer un asegurado para estar cubierto de por vida sería:

$$A_x = \frac{vd_x + v^2 d_{x+1} + v^3 d_{x+2} + \cdots + v^{w-x} d_{x+(w-x-1)}}{l_x}$$
$$A_x = \frac{v^{x+1} d_x + v^{x+2} d_{x+1} + v^{x+3} d_{x+2} + \cdots + v^{w} d_{x+(w-x-1)}}{v^x l_x} \tag{1.13}$$

Por último, para obtener la prima neta nivelada hay que fraccionar la prima neta única en pagos periódicos vitalicios, dichos pagos en valor presente deben ser equivalentes al pago único que debería hacer el asegurado por la cobertura del seguro.

$$A_x = P_x \ddot{a}_x$$
$$P_x = \frac{A_x}{\ddot{a}_x} \tag{1.14}$$

Donde P_x es la prima neta nivelada que pagará un asegurado de edad x por el seguro de vida y $\ddot{a}_x$ es una anualidad vitalicia anticipada[4] para la persona de edad x

1.4. Reserva Matemática

Una compañía de seguros en su administración debe garantizar que tendrá solvencia, es decir que tendrá los recursos necesarios para hacer frente a las obligaciones (pagos) que tiene con sus asegurados.

Si regresamos al ejemplo de la sección anterior y comparamos un seguro de vida individual anual con un seguro de prima neta nivelada de las mismas características encontramos que existe una diferencia entre la prima pura natural y la prima neta nivelada, en los primeros años es a favor de la aseguradora mientras que en los últimos esta recibe menos de prima por un seguro anual equivalente.

Esta diferencia en el futuro debe ser compensada con la capitalización de los ingresos adicionales que recibió la aseguradora al inicio del contrato.

Al construir la prima neta nivelada llegamos a la siguiente igualdad:

$$A_x = P_x \ddot{a}_x \tag{1.15}$$

[4] Una descripción más profunda y la construcción de estas anualidades puede revisarse en el Apéndice A

Es decir, al inicio del contrato de seguro los compromisos adquiridos por la compañía y el asegurador son los mismos el equilibrio se rompe con el paso del tiempo debido a que los pagos hechos y las obligaciones adquiridas no son iguales, al cabo de t años se tendrá:

$$A_{x+t} \neq P_x \ddot{a}_{x+t} \tag{1.16}$$

La diferencia se debe a que la prima neta nivelada no corresponde con la prima neta única, una manera de analizar el monto de esta diferencia es la siguiente:

Debido al aumento en la probabilidad de muerte el monto de la prima neta nivelada correspondiente a un seguro aumenta con la edad.

$$P_x < P_{x+t} \tag{1.17}$$

Si multiplicamos ambos lados por $\ddot{a}_{x+t}$ tenemos:

$$P_x \ddot{a}_{x+t} < P_{x+t} \ddot{a}_{x+t} \tag{1.18}$$

Por 1.15 llegamos a la siguiente expresión

$$P_x \ddot{a}_{x+t} < A_{x+t} \tag{1.19}$$

El monto de las obligaciones de la aseguradora que es equivalente a la prima única de un seguro adquirido en el tiempo $x + t$ es mayor al valor presente de las primas que pagará el asegurado a ese momento, es decir, las obligaciones del asegurado. Esta obligación adicional es resultado del ingreso adicional que obtuvo la aseguradora al inicio del contrato.

Además la desigualdad 1.19 nos da el monto que debe tener reservado la aseguradora para mantener el equilibrio entre sus obligaciones y sus ingresos por prima, a este monto se le denominará reserva, representado al tiempo t por ${}_tV_x$.

$$\begin{aligned} P_x \ddot{a}_{x+t} +_t V_x &= A_{x+t} \\ {}_tV_x &= A_{x+t} - P_x \ddot{a}_{x+t} \end{aligned} \tag{1.20}$$

Capítulo 2

Modelo clasificatorio logístico.

Para clasificar a los asegurados, se utilizará un modelo clasificatorio logístico, este modelo es útil cuando la variable de interés es una variable dicotómica, 0 o 1, éxito o fracaso, renovar o no renovar... Entre otras.

La intención de los modelos de regresión lineal es describir el comportamiento de una variable respuesta o dependiente usualmente denominada y, por medio de una o varias variables explicativas llamadas $\{x_i\}$ a través de un conjunto de coeficientes $\{\beta_i\}$ que indican la influencia de cada variable explicativa sobre el comportamiento de la variable dependiente

El objetivo del análisis de comportamientos utilizando este tipo de modelos es encontrar el modelo que mejor se ajuste a los datos y a su vez explique la relación entre las variables.

2.1. Regresión logística simple.

Para comprender el modelo de regresión logística múltiple es conveniente familiarizarnos con la regresión logística simple, la idea principal de este modelo es la siguiente:

Imaginemos una colección de i individuos de los que conocemos la característica $\{X_i\}$ que pueden tomar una decisión $\{Y_i\}$, desde el punto de vista de quien analiza las decisiones tomadas por los individuos, estas decisiones pueden tomar sólo dos valores $\{0, 1\}$ a los que llamaremos éxito o fracaso.

Sea $\{Y_i^*\}$ una variable que desconocemos pero que nos dice la utilidad que le genera al individuo i tomar la decisión Y_i podemos asumir que si la característica observada de los individuos describe el comportamiento de estos existe el siguiente modelo de regresión:

$$Y_i^* = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2.1}$$

En el que β_j es un parámetro desconocido y ϵ_i es el error del modelo. Desconocemos los valores de Y_i^* y β definimos una variable asociada a esta utilidad $\{\tilde{Y}_i\}$ esta variable corresponde con la decisión tomada por el individuo i bajo el modelo mostrado en 2.1.

$$\begin{aligned} &\tilde{Y}_i = 1 \text{ si } Y_i^* > 0 \\ &\tilde{Y}_i = 0 \text{ e.o.c.} \end{aligned} \tag{2.2}$$

Si nuestros individuos toman decisiones de manera racional los valores de Y_i y $\tilde{Y}_i$ coinciden y podemos calcular las siguientes probabilidades[5]:

$$P(Y_i = 1|X_i) = P(Y_i^* > 0|X_i) = P(\epsilon_i > -(\beta_0 + \beta_1 X_i)) = F(\beta_0 + \beta_1 X_i) \tag{2.3}$$

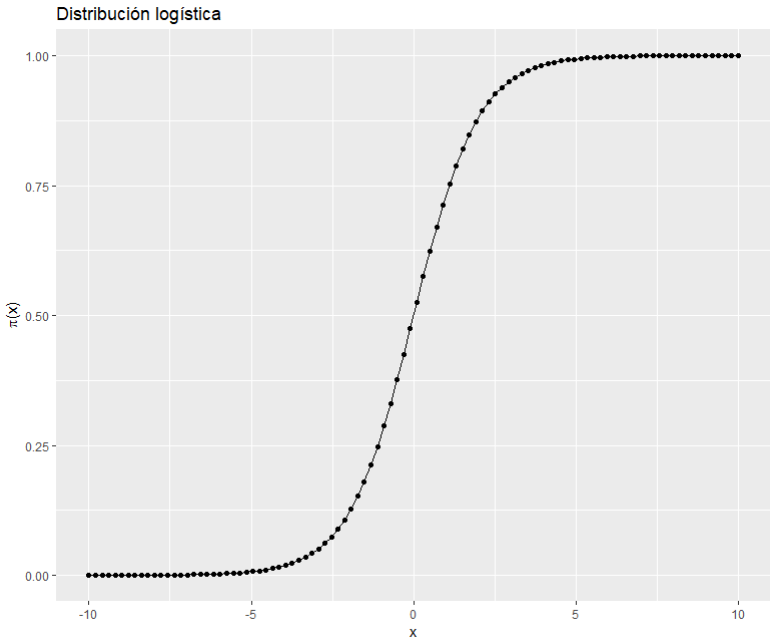
$$P(Y_i = 0|X_i) = P(Y_i^* \leq 0|X_i) = P(\epsilon_i < -(\beta_0 + \beta_1 X_i)) = 1 - F(\beta_0 + \beta_1 X_i) \tag{2.4}$$

Donde $F(\cdot)$ es la función de distribución del error ϵ_i, dependiendo que función de distribución asumamos para el error, será el modelo al que llegaremos. En caso de asumir una distribución Normal será el modelo conocido como *probit* y si asumimos la distribución logística obtendremos el modelo logístico, la función de distribución logística es la siguiente:

$$P[Y = 1|X = x] = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \tag{2.5}$$

[5]Para la última igualdad de cada ecuación estamos tomando el simétrico de la función de distribución

Figura 2.1: Gráfica de la distribución logística



Fuente: Elaboración Propia. con datos de D.W. Hosmer, 2013 [12]

La motivación para utilizar esta distribución contra otras distribuciones es su comportamiento ya que en los extremos de su dominio toma valores muy cercanos a 0 y 1 mientras que sólo cerca del cero tiene valores diferentes.

Para simplificar llamaremos $\pi(x)$ a $P(Y = 1|X = x)$, notemos además que podemos obtener una expresión para el modelo lineal en términos de $\pi(x)$.

$$g(x) = \beta_0 + \beta_1 x = log[\frac{\pi(x)}{1 - \pi(x)}] \tag{2.6}$$

La transformación anterior conocida como transformación *logit* es importante desde el punto de vista analítico ya que es continua en $(-\infty, \infty)$, es lineal en sus parámetros y permite heredar propiedades del análisis de regresión lineal para el modelo logístico debido a que en su forma coincide con el modelo de regresión lineal.

Debido a que Y_i es una variable dicotómica podemos verificar la siguiente igualdad que nos habla de la esperanza de la variable $Y|X$.

$$E[Y|X] = 0 \cdot P[Y = 0|X = x] + 1 \cdot P[Y = 1|X = x] = P[Y = 1|X = x] = \pi(x) \tag{2.7}$$

Por lo anterior tendremos que bajo un modelo logístico la aproximación de Y será $\pi(x)$.

$$\hat{y} = \pi(x) \tag{2.8}$$

Hay que tomar en cuenta que $\pi(x)$ es una variable que puede tomar cualquier valor entre cero y uno mientras que y puede tomar los valores $\{0, 1\}$ por lo anterior es necesario definir un criterio p que permita decidir a partir de que valor de $\pi(x)$ se considerará que $\hat{y} = 1$ siguiendo el siguiente criterio:

$$\begin{aligned} \pi(\underline{x}_i) \leq p \rightarrow \hat{y}_i = 0 \\ \pi(\underline{x}_i) > p \rightarrow \hat{y}_i = 1 \end{aligned} \tag{2.9}$$

En el capítulo 3 en el que se analizan los errores asociados al modelo se hará hincapié en la definición de este criterio.

2.1.1. Distribución del error.

En un modelo de regresión logística el error no sigue el supuesto del modelo de regresión lineal sobre el error, es decir el error no sigue una distribución de probabilidad Normal con media cero y una varianza constante, al ser nuestro modelo un modelo que busca una salida dicotómica (0 o 1) el error está restringido a ser cero cuando el modelo nos otorga el valor correcto o uno en caso de que no lo entregue.

Intuitivamente el error tiene una distribución Bernoulli, para confirmar lo anterior podemos ver el modelo lineal de la siguiente manera:

$$Y = E[Y/X] + \epsilon \tag{2.10}$$
$$\epsilon = Y - E[Y/X] \tag{2.11}$$
$$\epsilon = Y - \pi(x) \tag{2.12}$$

Por 2.3 si $Y = 0$ entonces $\epsilon = -\pi(x)$ con probabilidad $1 - \pi(x)$ mientras que si $Y = 1$ $\epsilon = 1 - \pi(x)$ con probabilidad $\pi(x)$ de lo anterior podemos concluir:

$$E[\epsilon] = 0 \tag{2.13}$$
$$var(\epsilon) = \pi(x)(1 - \pi(x)) \tag{2.14}$$

2.1.2. Ajuste del modelo.

Al analizar el modelo de regresión lineal el mejor método de ajuste es el de mínimos cuadrados ya que este método resulta ser de varianza mínima, este consiste en buscar los coeficientes β_0 y β_1 que minimicen el error promedio del modelo dados y_i y x_i es decir:

$$S = \sum_{i=1}^{n} \epsilon_i^2 \tag{2.15}$$

El procedimiento anterior se reduce a un problema de optimización de una función lineal continua de 2 variables mismo que se puede obtener por medio de las derivadas parciales de S respecto a los parámetros e igualando a cero.

Para el modelo logístico dicho procedimiento no es útil debido a que $\pi(x)$ no es lineal respecto a los parámetros, por esto el ajuste del modelo se apoya en la transformación logit la cual nos provee de un modelo lineal.

$$g(x) = \beta_0 + \beta_1 x \tag{2.16}$$

Debido a la distribución del error y a la debilidad del método de mínimos cuadrados respecto a la violación de supuestos la mejor manera para obtener estimadores para β_0 y β_1 es buscar los estimadores maximoverosímiles de estos parámetros.

Sea una muestra de tamaño n observaciones independientes de la pareja (x_i, y_i) donde y_i es el valor de una variable dependiente dicotómica y x_i es el valor de una variable independiente para la $i - esima$ observación nuestro modelo de regresión tiene la siguiente distribución:

$$P[Y = y_i] = \pi(x)^{y_i}(1 - \pi(x))^{1-y_i} \tag{2.17}$$

Su funciones de verosimilitud y Log-verosimilitud[6] están dadas por las siguientes expresiones:

$$\mathcal{I}(\beta_0, \beta_1, x_i, y_i) = \prod_{i=1}^{n} \pi(x)^{y_i}(1 - \pi(x))^{1-y_i} \tag{2.18}$$

$$\mathcal{L}(\beta_0, \beta_1, x_i, y_i) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} ln(1 + e^{\beta_0 + \beta_1 x_i}) \tag{2.19}$$

[6] La construcción de la Log-verosimilitud a partir de la verosimilitud se puede consultar en el Apéndice A

Al derivar la función de verosimilitud llegamos a la siguientes expresiones:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_0} &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \\ &= \sum_{i=1}^{n} y_i - \pi(x) \end{aligned} \tag{2.20}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_1} &= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \\ &= \sum_{i=1}^{n} x_i (y_i - \pi(x)) \end{aligned} \tag{2.21}$$

Para obtener β_0 y β_1 es necesario igualar a cero 2.20 y 2.21 de manera simultanea, en la práctica esto se hace por medio de algún paquete de software que utilizando métodos numéricos obtendrá el máximo buscado, los algoritmos numéricos más comunes para esto son Newton-Rabson o bien el algoritmo IWLS[7]

2.2. Regresión logística múltiple.

La sección anterior sirvió para introducir la idea general del modelo logístico misma que se replica en el modelo de regresión logística múltiple, la idea básica del modelo es la misma que en el anterior, explicar el comportamiento de una variable dicotómica de interés Y con un conjunto de variables explicativas, en este caso las variables explicativas serán un vector de dimensión k $X_i \in \mathbb{R}^{\mathrm{k}}$.

Este modelo nos da la ventaja de que no siempre es posible explicar una variable en términos de una característica y que en la realidad los factores que influyen en la toma de decisiones de una persona pueden ser múltiples.

2.2.1. Ajuste del modelo.

Para este modelo asumiremos que tenemos una muestra de variables independientes de tamaño n cada una de ellas es un vector de dimensión k con $i \in \{1, 2, \cdots n\}$ y

[7] Iterative Weighted Least Square

$j \in \{1, 2, \cdots k\}$ sea $\underline{\mathrm{x}}_i = x_{ij} = (x_{i1}, x_{i2}, \cdots, x_{ik})$ asumiremos que a lo menos son categóricas ordinales [8] y preferiblemente numéricas.

Si de manera análoga al modelo anterior definimos $P[Y = 1|\underline{\mathrm{X}}] = \pi(\underline{\mathrm{x}})$ podemos definir el modelo de regresión múltiple de la siguiente manera:

$$\pi(\underline{\mathrm{x}}_i) = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}} \tag{2.22}$$

La transformación logit está dada por las siguientes ecuaciones:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} \end{aligned} \tag{2.23}$$

Dichas ecuaciones son resultado de aplicar la transformación logit (2.7) al modelo de regresión múltiple (2.22), una por cada vector $\underline{\mathrm{x}}_i$

De manera análoga al modelo de regresión logística simple usaremos la semejanza entre (2.23) y el modelo de regresión lineal múltiple para describir cómo se obtienen los parámetros $\underline{\beta} = (\beta_0, \beta_1, \cdots, \beta_k)$

Las funciónes de verosimilitud y Log-verosimilitud son equivalentes a las del modelo univariado tomando en cuenta que $\underline{x}_i$ es un vector y son las siguientes:

$$\mathfrak{I}(\underline{\beta}, \underline{\mathrm{x}}_i, y_i) = \prod_{i=1}^{n} \pi(\underline{\mathrm{x}}_i)^{y_i} (1 - \pi(\underline{\mathrm{x}}_i))^{1-y_i} \tag{2.24}$$

$$\mathcal{L}(\underline{\beta}, \underline{\mathrm{x}}_i, y_i) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 \underline{\mathrm{x}}_i) - \sum_{i=1}^{n} ln(1 + e^{\beta_0 + \beta_1 \underline{\mathrm{x}}_i}) \tag{2.25}$$

Al derivar (2.24) respecto cada β_i llegamos al siguiente sistema de ecuaciones de tamaño $(k + 1)$ x $(k + 1)$

[8] Para las variables categóricas es necesario el orden para poder asignar un valor numérico a las mismas basados en dicho orden

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_0} &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_1 \underline{\mathrm{x}}_i}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_i \underline{\mathrm{x}}_i}} \\ &= \sum_{i=1}^{n} y_i - \pi(\underline{\mathrm{x}}_i) = 0 \end{aligned} \tag{2.26}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_m} &= \sum_{i=1}^{n} x_{im} y_i - \sum_{i=1}^{n} x_{im} \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_1 \underline{\mathrm{x}}_i}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_i \underline{\mathrm{x}}_i}} \\ &= \sum_{i=1}^{n} x_{im}(y_i - \pi(\underline{\mathrm{x}}_i)) = 0 \qquad \forall m \in \{1, 2, \cdots, k\} \end{aligned} \tag{2.27}$$

Al igual que en el cálculo de los estimadores del modelo univariado, estos estimadores se obtienen haciendo uso de software que hace la aproximación de estos por medio de métodos numéricos.

2.3. Varianza y covarianza de los estimadores

Que los estimadores de nuestros parámetros sean los máximos verosímiles es engorroso desde el punto de vista del cálculo de estos. Sin embargo, el análisis de varianza del modelo se simplifica debido a todo el trabajo teórico que hay detrás de estos estimadores.

Las segundas derivadas de la función de Log-verosimilitud son las siguientes:

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_m^2} = -\sum_{i=1}^{n} x_{im}^2 \pi(\underline{\mathrm{x}}_i)(1 - \pi(\underline{\mathrm{x}}_i)) \tag{2.28}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_l \partial \beta_m} = -\sum_{i=1}^{n} x_{im} x_{il} \pi(\underline{\mathrm{x}}_i)(1 - \pi(\underline{\mathrm{x}}_i)) \tag{2.29}$$

Constriur con los negativos de estas derivadas una matriz de tamaño $(k+1)$ x $(k+1)$ da lugar a la *Matriz de Información de Fischer* usualmente conocida como *Matriz de Información observada* $I(\beta)$.

Por ser los estimadores obtenidos máximo verosímiles estos son estimadores de varianza mínima, es decir, por teorema de *Rao-Blackwell* que menciona J. D. Gibbons, 2003 [17] la *Matriz de varianzas y covarianzas* coincidirá con el simétrico de la *Matriz de Información observada*.

$$\Sigma(\beta) = I^{-1}(\beta) \tag{2.30}$$

De esta matríz se puede extraer la siguente información:

- $Var(\beta_i)$ Será el $i - esimo$ elemento de la diagonal de $\Sigma(\beta)$ y será la varianza del estimador $\hat{\beta}_i$.
- $Cov(\beta_i, \beta_j)$ denota la covarianza entre $\hat{\beta}_i$ y $\hat{\beta}_j$.
- Los estimadores de los estadísticos anteriores $\widehat{Var(\hat{\beta}_i)}$ y $\widehat{Cov(\hat{\beta}_i, \hat{\beta}_j)}$ surgen de evaluar a Σ en los estimadores $\underline{\hat{\beta}}$.
- Se denotará por Error Standar del estimador β_i a $\widehat{SE(\hat{\beta}_i)} = \sqrt{\widehat{Var(\hat{\beta}_j)}}$.

2.4. Significancia del modelo

Una vez que se han obtenido los coeficientes del modelo surge la duda sobre que tanta repercusión tiene cada variable en el modelo y si puede haber alguna que no tenga relevancia para el mismo. Si bien en el siguiente capítulo se discutirán técnicas estadísticas sobre selección de variables, es conveniente que una vez ajustado el modelo se revise la significancia de los coeficientes obtenidos, es decir, si estos coeficientes son distintos o no de cero.

Una variable es significativa para un modelo de regresión si variaciones en dicha variable repercuten en el valor de la aproximación de la variable dependiente, en este punto no es importante la precisión del ajuste sólo será relevante si la variable o variables analizadas tienen injerencia sobre los resultados ofrecidos por el modelo.

Observando un modelo de regresión lineal múltiple (2.23) la significancia de las variables depende de si parámetros $\underline{\beta}$ son cero o distintos de este valor por lo que la significancia de las variables dependerá de pruebas estadísticas para las siguientes hipótesis:

$$\begin{aligned} \mathcal{H}_0 : &\quad \beta_i = 0 \\ \mathcal{H}_1 : &\quad \beta_i \neq 0 \end{aligned}$$

En el modelo de regresión lineal múltiple la prueba de significancia por excelencia es la prueba $\mathcal{F}$ que se deriva del análisis ANOVA[9], desgraciadamente esta prueba depende de la distribución del error por lo que no puede ni debe ser usada en un modelo de regresión logística.

[9] Analisis de Varianza

Debido a que los estimadores de $\hat{\beta}_i$ son resultado de la función de máxima verosimilitud las pruebas de significancia sobre estos parámetros se derivarán de esta función. Las pruebas más comunes de significancia para el modelo de regresión logística son las siguientes:

2.4.1. Razón de Verosimilitudes.

Si imaginamos un modelo que contiene a todas las variables observadas sin resumir, estaremos ante lo que se conoce como modelo saturado M_s, este modelo tiene la particularidad de que la variable dicotómica Y coincide con el valor $\hat{y}_i$ obtenido por el modelo. Si bien este modelo desde el punto de vista del error es deseable, desde el punto de vista de la predicción puede no serlo debido a que no aportará información sobre nuevas observaciones de las variables. Sin embargo, nos servirá para modelar la significancia del modelo de regresión logística.

Si definimos la función de verosimilitud del modelo saturado de manera análoga al modelo ajustado tenemos por (2.24):

$$\mathcal{J}(M_s) = \prod_{i=1}^{n} y_i^{y_i}(1-y_i)^{1-y_i} = 1 \tag{2.31}$$

Definimos la *Devianza* como el siguiente coeficiente, para simplificar llamaremos al modelo ajustado M_a

$$D(M_a) = -2log(\frac{\mathcal{J}(M_a)}{\mathcal{J}(M_s)})$$

Por (2.31)

$$D(M_a) = -2log(\mathcal{J}(M_a)) \tag{2.32}$$

Dentro del paréntesis de la primera expresión tenemos el coeficiente de verosimilitudes, que nos podría dar una idea que tan similares son los modelos, por lo que la devianza puede ser utilizada como una prueba de bondad de ajuste, en la siguiente sección se discutirá su uso para estos propósitos.

La devianza por si misma no es una prueba estadística sobre la significancia del modelo, pero mediante ella podemos construir una prueba sobre la significancia de cada variable seleccionada, para esta prueba hablaremos de dos modelos, el modelo ajustado, será el modelo que contiene nuestras variables y el modelo sin la variable X_i que será la variable sobre la que analizará la significancia, a este modelo lo llamaremos modelo reducido M_r.

Las hipótesis de la prueba de razón de verosimilitudes son:

$$\begin{aligned} \mathcal{H}_0 : &\quad \beta_i = 0 \\ \mathcal{H}_1 : &\quad \beta_i \neq 0 \end{aligned}$$

Definimos la razón de verosimilitudes de la siguiente manera:

$$G = -2log(\frac{\mathcal{J}(M_r)}{\mathcal{J}(M_a)}) = 2(D(M_a) - D(M_r)) \tag{2.33}$$

De manera análoga a la devianza la prueba de razón de verosimilitudes tiene por intención analizar que tan parecidos serán el modelo ajustado y el modelo reducido, en caso de que $\beta_i = 0$ esta razón de verosimilitudes debe de ser cercana a cero.

Por el teorema que enuncia Wilks en su trabajo (Wilks, 1938 [26]) y bajo la hipótesis nula $G \sim \chi^2_{(1)}$ entonces para un nivel de significancia α se rechaza $\mathcal{H}_0$ si $G > (\chi^2_{(1)})^{(1-\alpha)}$ donde $(\chi^2_{(1)})^{(1-\alpha)}$ es el quantil $1-\alpha$ de una distribución χ^2 con un grado de libertad.

Esta prueba se muestra de manera univariada pero se puede extender a un subconjunto de variables, las diferencias radican en que la hipótesis nula cambia a:

$$\mathcal{H}_0 : \beta_i = \beta_j = \cdots = \beta_k = 0$$

$$\mathcal{H}_1 : \beta_l \neq 0 \qquad \text{Para alguna } \beta_l \text{ en el conjunto de parámetros de } \mathcal{H}_0$$

Y además bajo $\mathcal{H}_0$ $G \sim \chi^2_{(k)}$ donde el número de grados de libertad es el número de variables retiradas del modelo ajustado para analizar, esta prueba múltiple tiene la desventaja de no garantizar que todos los parámetros no son cero y además en la práctica los paquetes de software entregan en los resultados una columna con la prueba para cada variable.

2.4.2. Prueba de Wald.

Se puede obtener otra prueba de significancia para parámetros individuales conocida como prueba de Wald, esta prueba se basa en que bajo la hipótesis de que $\beta_i = 0$ y el teorema central del límite, el estimador $\hat{\beta}_i$ se puede transformar en una variable aleatoria con distribución Normal estándar.

Las hipótesis de esta prueba son:

$$\begin{aligned} \mathcal{H}_0 : &\qquad \beta_i = 0 \\ \mathcal{H}_1 : &\qquad \beta_i \neq 0 \end{aligned}$$

Definimos el estadístico z_i de la siguiente manera:

$$z_i = \frac{\hat{\beta}_i}{\widehat{SE(\hat{\beta}_i)}} \tag{2.34}$$

Bajo la hipótesis nula z_i es la estandarización del estimador $\hat{\beta}_i$, por lo que con una confianza α se rechaza la hipótesis nula si $z_i > Z_{(1-\frac{\alpha}{2})}$ o bien si $z_i < Z_{(\frac{\alpha}{2})}$ donde $Z_{(1-\frac{\alpha}{2})}$ y $Z_{(\frac{\alpha}{2})}$ son los quantiles $Z_{(\frac{\alpha}{2})}$ y $1 - Z_{(\frac{\alpha}{2})}$ de una distribución Normal estándar.

Algunos paquetes muestran el estadístico z_i^2 que tiene una distribución χ^2 con un grado de libertad y cuya prueba de hipótesis es análoga a la razón de verosimilitudes

Existe una versión multivariada de esta prueba, esta tampoco distingue si todos los parámetros son diferentes de cero o sólo uno no es diferente de cero, las hipótesis de esta prueba son:

$$\mathcal{H}_0 : \beta_0 = \beta_1 = \cdots = \beta_k = 0$$

$$\mathcal{H}_i : \beta_l \neq 0 \qquad \text{Para alguna } \beta_i$$

La estadística de prueba en notación matricial se denota:

$$z = \underline{\hat{\beta}}^T \Sigma(\underline{\hat{\beta}}) \underline{\hat{\beta}} \tag{2.35}$$

Bajo la hipótesis nula el estadístico z obtenido tiene una distribución χ^2 con $(k+1)$ grados de libertad entonces para un nivel se significancia α se rechaza $\mathcal{H}_0$ si $z > (\chi^2_{(k+1)})^{(1-\alpha)}$ donde $(\chi^2_{(k+1)})^{(1-\alpha)}$ es el quantil $1-\alpha$ de una distribución χ^2 con $(k+1)$ grados de libertad.

2.5. Pruebas de bondad de ajuste.

Al tener un modelo ajustado con variables significativas es conveniente analizar la magnitud del error, es decir, que tanta diferencia hay entre las observaciones de la

variable dependiente Y_i y la aproximación obtenida por el modelo $\hat{y}_i$ permitiéndonos analizar que tanto es reflejada la realidad por el modelo.

En el caso del modelo de regresión lineal la bondad de ajuste se mide por medio del coeficiente de determinación R^2, este coeficiente también está sujeto al supuesto de normalidad en los residuos. Por lo que para el modelo de regresión logística múltiple se requiere utilizar de otros métodos estadísticos para analizar la bondad de ajuste.

Para el modelo de regresión lineal la medida de la bondad de ajuste es sencilla e intuitiva $(Y_i - \hat{y}_i)$ para el modelo de regresión logística surge el problema de que al ser la variable dependiente dicotómica puede haber combinaciones de variables idénticas con diferentes respuestas, por ejemplo, si en nuestras variables independientes tenemos sólo sexo y edad podríamos estar en el caso de que dos mujeres de 35 años elijan opciones diferentes.

Para distinguir estas situaciones, se requiere introducir un concepto adicional *Patrones de covariables*, estos contemplarán las combinaciones entre las variables que lo requieran para poder distinguir los casos posibles. Hay que notar que este concepto no tiene injerencia en la construcción, ajuste y significancia del modelo ya que en general, estos dependen del número de parámetros y no de los casos de ocurrencia, como es el caso de la bondad de ajuste.

Recordemos que en nuestro modelo tenemos una muestra de tamaño n de variables aleatorias x_i, cada variable es de dimensión k por lo que podemos describir a estas de la siguiente manera:

$$\underline{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ik}) \qquad i \in \{1, 2, \cdots, n\} \tag{2.36}$$

Sea J el número de valores únicos para la variable x_i observados en la muestra, denotaremos al número de personas que pertenecen al patrón de valor unico x_j con m_j $j \in \{1, 2, \cdots, J\}$ es evidente que $\sum_{i=1}^{J} m_j = n$, denotaremos al número de personas del patrón j para los que $y = 1$ como y_j por lo anterior $\sum_{j=1}^{J} y_j = \sum_{i=1}^{n} y_i = n_1$

Las distribuciones utilizadas para las pruebas de significancia son asintóticas cuando n crece (Wilks, 1938 [26]), para las pruebas de bondad de ajuste se utilizarán distribuciones asintóticas a m_j es intuitivo que conforme n crece los integrantes de cada patrón de covariables aumentarán, pero no necesariamente lo hará el número de patrones J por lo que suponer $J \approx n$ no siempre será lo correcto.

2.5.1. Prueba χ^2.

Debido a la naturaleza del método de máxima verosimilitud utilizado para obtener los estimadores para los parámetros $\underline{\hat{\beta}}_i$ es intuitivo hacer un análisis similar al hecho en la sección sobre la significancia del modelo asumiendo una distribución χ^2

Para cada patrón definiremos el valor ajustado de dicho patrón de la siguiente manera:

$$\hat{y}_j = m_j\hat{\pi}_j = m_j(\frac{e^{\hat{g}(x_j)}}{1+e^{\hat{g}(x_j)}}) \tag{2.37}$$

Donde $\hat{g}(x_j)$ es la suma de las transformaciones logit de los elementos del patrón j

Para cada patrón de covariables podemos definir el residual de Pearson de la misma manera que en el modelo de regresión lineal

$$r(y_j, \hat{\pi}_j) = \frac{y_j - \hat{y}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}} \tag{2.38}$$

Podemos definir el estadístico χ^2 como la suma de los residuales de pearson al cuadrado

$$\chi^2 = \sum_{j=1}^{J}(r(y_j, \hat{\pi}_j))^2 \tag{2.39}$$

En un escenario en el que $J \approx n$ el estadístico χ^2 tendría una distribución χ^2 con $(j-k-1)$ grados de libertad. Pudiéndose usar como prueba de bondad de ajuste bajo las hipótesis:

$$\mathcal{H}_0 : y_i \leq \hat{y}_i$$
$$\mathcal{H}_1 : y_i = \hat{y}_i$$

Bajo la hipótesis nula $\chi^2 \sim \chi^2_{(J-k-1)}$ entonces para un nivel se significancia α se rechaza $\mathcal{H}_0$ si $\chi^2 > (\chi^2_{(J-k-1)})^{(1-\alpha)}$ donde $(\chi^2_{(J-k-1)})^{(1-\alpha)}$ es el quantil $1-\alpha$ de una distribución χ^2 con $(J-k-1)$ grados de libertad.

2.5.2. Prueba de Hosmer-Lemeshow.

El supuesto de $J \approx n$ se debilita conforme el modelo tiene más variables categóricas debido a que estas limitan el número de patrones de covariables. Lo anterior hace que la prueba χ^2 no sea muy confiable para este tipo de datos.

Por lo anterior, en 1980 y 1982 Hosmer y Lemeshow publicaron de manera conjunta trabajos que proponen una prueba de bondad de ajuste basada en los valores de la probabilidad estimada $\hat{\pi}(\underline{x}_i)$.

El estadístico de Hosmer-Lemeshow parte de la comparación del modelo ajustado con los datos reales, de una manera similar a la metodología utilizada en pruebas de bondad de ajuste para distribuciones como la prueba de *Lillieforts*[10] para distribuciones Normal y exponencial, evitando suposiciones de distribuciones asintóticas cuando crece J que podrían no cumplirse.

La idea general es agrupar las probabilidades estimadas por el modelo de regresión logística ajustado según su valor para después formar G grupos conformados por una cantidad idéntica de integrantes $\frac{n}{G}$

Para cada grupo se calculan la suma de las frecuencias observadas y estimadas por el modelo de cada respuesta de la variable independiente.

Para el grupo g con $g \in \{1, 2, \cdots, G\}$ el número de respuestas positivas y negativas (0 y 1) esperadas son respectivamente:

$$\begin{aligned} E_{1g} &= \sum_{i=1}^{c_g} m_i \hat{\pi}(\underline{x}_{ig}) \\ E_{0g} &= \sum_{i=1}^{c_g} m_i (1 - \hat{\pi}(\underline{x}_{ig})) \end{aligned} \tag{2.40}$$

Donde c_g es el número de patrones en el grupo g

Mientras que el número de respuestas positivas y negativas por categoría será:

$$\begin{aligned} O_{1g} &= \sum_{i=1}^{c_g} Y_i \\ O_{0g} &= \sum_{i=1}^{c_g} (m_i - Y_i) \end{aligned} \tag{2.41}$$

La estadística $\hat{C}$ se calcula comparando los valores observados con los esperados para cada grupo de la siguiente manera:

$$\hat{C} = \sum_{g=1}^{G} [\frac{(O_{1g} - E_{1g})^2}{E_{1g}} + \frac{(O_{0g} - E_{0g})^2}{E_{0g}}] \tag{2.42}$$

La prueba se puede simplificar a la siguiente expresión:

[10] Esta prueba puede consultarse (J. D. Gibbons, 2003 [17])

$$\hat{C} = \sum_{g=1}^{G} \frac{(O_{1g} - n_g\pi_g)^2}{n_g\pi_g(1-\pi_g)} \tag{2.43}$$

Siendo π_g la probabilidad promedio para el grupo g

$$\pi_g = \frac{1}{n_g} \sum_{j=1}^{c_g} m_j \hat{\pi}(\underline{x}_{ig}) \tag{2.44}$$

Utilizando simulaciones publicadas D.W. Hosmer, 2013 [12], los autores de la prueba mostraron que cuando el modelo de regresión logística es el correcto, la distribución de $\hat{C}$ es χ^2 con $(g-2)$ grados de libertad y que dicha distribución es independiente de si se cumple o no el supuesto de $J \approx n$.

2.5.3. Pseudo-R^2.

Las pruebas anteriores nos dan una idea general sobre la bondad de ajuste del modelo, pero a diferencia del modelo de regresión lineal no tenemos un valor numérico que nos resuma la bondad de ajuste como es el caso del coeficiente de determinación R^2.

Para ello se han construido varios coeficientes que buscan cumplir con esta función, con la idea de resumir la similitud entre modelos, usualmente a estos coeficientes se les llama *pseudo-R^2*.

En 1973 McFadden propone una alternativa a la que el denominó "Índice de radio de verosimilitudes"[11], este índice y otros pseudo-R^2 fueron analizados por Bo, 2005 [6] .

La idea principal de este coeficiente es comparar el modelo ajustado M_a con un modelo en el que todos los coeficientes de las variables son cero excepto el coeficiente β_0, es decir $\beta_i = 0 \quad \forall i \neq 0$ al que llamaremos modelo nulo M_0.

El coeficiente pseudo-R^2 de McFadden se define con la siguiente expresión:

$$R^2_{McFadden} = 1 - \frac{(\mathcal{L}(M_a))}{(\mathcal{L}(M_0))} \tag{2.45}$$

Este estadístico se explica de la siguiente manera:

- En caso de que las variables no expliquen el comportamiento de la variable dependiente los coeficientes β_i de estas serán cero, por lo que la log-verosimilitud de ambos modelos será muy similar haciendo a su cociente cercano a uno resultando en un $R^2_{McFadden}$ cercano a cero.

[11] "likelihood-ratio index"

- En el caso opuesto de que todos los coeficientes sean distintos de cero y estos expliquen a la variable dependiente el modelo tendrá una verosimilitud similar a la del modelo saturado M_s cuyo valor es uno, siendo su logaritmo cercano a cero, por otro lado la verosimilitud del modelo M_0 será cercana a cero siendo su logaritmo cercano a uno resultando en un $R^2_{McFadden}$ cercano a uno como es de esperarse.

2.6. Intervalos de confianza.

Los intervalos de confianza son una herramienta que nos permiten con un grado de confianza α estimar en que rango se encontrarán los parámetros del modelo y por consiguiente la variable dependiente para una nueva observación de las variables independientes $\underline{\mathrm{x}}_i$

La normalidad del error en el modelo de regresión lineal, permite que los intervalos de confianza tengan una distribución $\mathcal{T}$ afectada por la varianza de los parámetros.

La estimación por máxima verosimilitud de los parámetros en el modelo de regresión logística nos permite utilizar una distribución normal $\beta_i \sim Normal(\hat{\beta}_i, \widehat{SE_{\hat{\beta}_i}})$ cuando el número de elementos de la muestra es suficientemente grande.

De la prueba de Wald (2.34) y del hecho de que no se cumple la hipótesis nula, es decir: $\beta_i \neq 0$ tenemos la siguiente propiedad:

$$\begin{aligned} C_i &= \frac{\beta_i - \hat{\beta}_i}{\widehat{SE}_{\hat{\beta}_i}} \\ C_i &\sim N(0,1) \end{aligned} \tag{2.46}$$

Por lo anterior los intervalos de confianza para los parámetros son:

$$\beta_i \in \{\hat{\beta}_i \pm Z_{(1-\frac{\alpha}{2})}\widehat{SE}_{\hat{\beta}_i}\} \tag{2.47}$$

Donde $Z_{(1-\frac{\alpha}{2})}$ es el cuantil $(1-\frac{\alpha}{2})$ correspondiente a una distribución Normal estándar.

2.7. Comparación entre los modelos de regresión lineal y logístico.

Como resumen es conveniente comparar las diferencias entre el modelo de regresión lineal y logística para analizar de una manera más sencilla las diferencias entre ambos.

Tabla 2.1: Diferencias entre los modelos de regresión linear y logístico

Propiedad	Modelo de Regresión Lineal	Modelo de Regresión logística
Tipo de variable dependiente	Real	Dicotómica $\{0, 1\}$
Distribución del error	$Normal(0, \sigma^2)$	$Binomial(1, \pi(x))$
Método de ajuste	Mínimos cuadrados	Máxima verosimilitud
Pruebas de significancia	Prueba $\mathcal{F}$	Razón de Verosimilitudes, Prueba de Wald
Pruebas de Bondad de Ajuste	Coeficiente de determinación (R^2)	Prueba χ^2, Prueba de Hosmer-Lemeshow
Intervalos de confianza	Distribución $\mathcal{T}$	Distribución $Normal$

Fuente: Elaboración Propia.

Capítulo 3

Validación y comprobación del modelo.

Un modelo de regresión logística nos da una aproximación matemática sobre la decisión que puede tomar una entidad al momento en caso de tener dos posibles opciones (variable dicotómica). Sin embargo, este modelo nos entrega el valor $\pi(\underline{x}_i)$ que puede tomar cualquier valor entre *0 y 1* para los valores en el extremo de este intervalo es intuitivo cual es el valor predicho. Sin embargo, conforme los valores de $\pi(\underline{x}_i)$ se alejan de dichos extremos puede no ser muy claro.

Una decisión intuitiva puede ser escoger como división el valor $\pi(\underline{x}_i) = 0.5$, pero puede ser posible que esta división no sea la correcta en términos de la intención del estudio particular que se esté haciendo, en este capítulo se analizará el efecto del criterio de clasificación en los errores del modelo.

3.1. Tablas de confusión y tipos de error.

Debido a la naturaleza de los tomadores de decisiones es posible distinguir a las entidades que bajo nuestro modelo y criterio de división tomen la decisión contraria a la predicha, teniendo 2 tipos posibles de error:

- Asignar 1 mediante el modelo cuando la decisión tomada será 0, llamado falso positivo.
- Asignar 0 mediante el modelo cuando la decisión tomada será 1, llamado falso negativo.

En el contexto de la decisión de renovar o no una póliza de seguros los errores anteriores se traducen a lo siguiente:

- Predecir que se va a cancelar una póliza que será renovada.
- Predecir que se renovará una póliza que será cancelada.

Una manera de representar estos errores para su fácil análisis es la construcción de tablas de contingencia, estas tablas contrastan las predicciones del modelo con los datos bajo los que fue construido.

Dado un criterio de división p es decir el valor de $\pi(\underline{x}_i)$ a partir del cual se considerará que un individuo tomará la decisión asociada al valor $y = 1$, es decir:

$$\pi(\underline{x}_i) \leq p \rightarrow \hat{y}_i = 0$$
$$\pi(\underline{x}_i) > p \rightarrow \hat{y}_i = 1$$

Una vez fijado el criterio puede construir la siguiente tabla de contingencia con las observaciones de la variable dependiente y las predicciones del modelo registrando la suma de los cuatro casos que pueden ocurrir al comparar las observaciones contra la predicción otorgada por el modelo.

Tabla 3.1: Matriz de confusión

		Predicción por el modelo	
		$\hat{y} = 0$	$\hat{y} = 1$
observaciones	$y = 0$	Verdadero negativo [**VN**]	Falso positivo [**FP**]
	$y = 1$	Falso negativo [**FN**]	Verdadero positivo [**VP**]

Fuente: Draper, 2002 [11]

De los resultados obtenidos en la matriz de confusión se obtienen los siguientes estadísticos:

- Sensibilidad: Este estadístico nos indica que tan probable es obtener un valor predicho de $\hat{y} = 1$ dado que el valor observado fue $y = 1$.

$$Sensibilidad = \frac{VP}{FN + VP} \tag{3.1}$$

- Especificidad: Este estadístico nos indica que tan probable es obtener un valor predicho de $\hat{y} = 0$ dado que el valor observado fue $y = 0$.

$$Especificidad = \frac{VN}{FP + VN} \tag{3.2}$$

- Valor predicho positivo (PV_{pos}): Este estadístico nos indica que tan probable es que una predicción $\hat{y} = 1$ sea correcta, es decir corresponda con una observación $y = 1$.

$$PV_{pos} = \frac{VP}{FP + VP} \tag{3.3}$$

- Valor predicho negativo (PV_{neg}): Este estadístico nos indica que tan probable es que una predicción $\hat{y} = 0$ sea correcta, es decir corresponda con una observación $y = 0$.

$$PV_{neg} = \frac{VN}{FN + VN} \tag{3.4}$$

Variando el valor de p, se puede encontrar el valor óptimo que nos permita minimizar el valor de algún error con cierta confianza estadística o el valor óptimo para alguno de los estadísticos previamente descritos. Lo anterior en función del contexto en el que se esté desarrollando el modelo.

Sería deseable minimizar ambos tipos de error para conseguir el mejor modelo posible. Sin embargo, ambos errores se conforman en sentido inverso cuando p varía ya que al aumentar el valor de p se reduce el número de falsos positivos pero al mismo tiempo aumenta el valor de falsos negativos por el aumento en los requisitos para decretar a una observación como $\hat{y} = 1$.

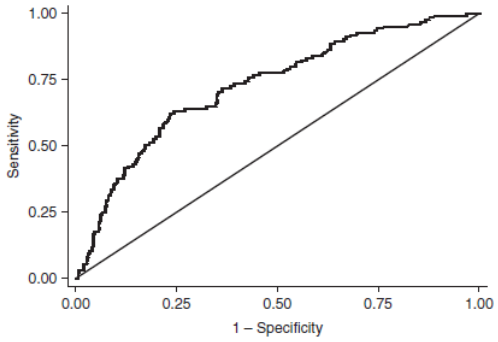
3.1.1. Curva ROC.

La información obtenida de las matrices de confusión para el modelo permite analizar el comportamiento de los errores para cada valor de p. Sin embargo, si buscamos analizar el comportamiento global del error tendremos múltipes tablas para analizar lo que puede ser tedioso, una manera facilitar este análisis es la curva ROC.

Esta curva se construye comparando la probabilidad de obtener una predicción correcta para un valor de $y = 1$ (Sensibilidad) contra la probabilidad de no obtener una medición correcta para un valor de $y = 0$ (1- Especificidad) para un amplio rango de valores de p. Permitiendo medir que tanto puede discriminar el modelo en los diversos valores de p.

En caso de que los valores de Sensibilidad y Especificidad fueran iguales para cada valor de p se tendría que la capacidad del modelo para discriminar sería incierta y equivalente a "lanzar una moneda"por lo que es deseable que sean distintas, una curva ROC deseable tiene el siguiente comportamiento.

Figura 3.1: Curva ROC



Fuente: D.W. Hosmer, 2013 [12]

Para medir la eficacia del modelo discriminando se utiliza el área bajo la curva ROC, es decir:

$$AUC = \int_0^1 ROC \tag{3.5}$$

Siendo el criterio para determinar la capacidad discriminatoria del modelo el siguiente:

Tabla 3.2: Reglas para evaluar el área bajo la curva ROC

Area bajo la curva ROC	Capacidad para discriminar
0.5 o menos	Sin capacidad discriminatoria
$0.5 < A \leq 0.7$	Poca capacidad discriminatoria
$0.7 < A \leq 0.8$	Capacidad discriminatoria aceptable
$0.8 < A \leq 0.9$	Capacidad discriminatoria excelente
0.9 o más	Capacidad discriminatoria excepcional

Fuente: Elaboración Propia. con información de D.W. Hosmer, 2013 [12]

3.2. Comprobación del modelo.

Una manera de evaluar la capacidad predictiva de un modelo antes de que este sea implementado es contrastar las predicciones hechas por el modelo, contra observaciones

que no formaron parte de la construcción del mismo, así estas observaciones pueden ser tomadas como nuevas observaciones y tienen la ventaja de que se dispone de los valores observados de la variable dependiente permitiendo medir la efectividad del modelo.

A la metodología anterior se le conoce como *backtesting* en Anderson, 2007 [4] se recomienda reservar para muestras grandes el 30 % de los datos para comprobar las capacidades del modelo. A esta muestra de los datos se le llama *Muestra de pruebas o Testing* mientras que al 70 % restante de los datos se le conoce como *Muestra de entrenamiento o Training*.

Para comprobar el modelo se sugiere el siguiente procedimiento:

- Obtener el valor predicho por el modelo para cada observación en la muestra de pruebas $\pi(\underline{x}_i)$.
- Con el valor de p seleccionado obtener los valores de la variable dependiente predichos por el modelo, es decir $\hat{y}(\underline{x}_i)$.
- Obtener la tabla de confusión y el estadístico AUC para la muestra de pruebas.

Con lo anterior se puede obtener una idea general de la capacidades del modelo para predecir el valor de la variable dependiente, comparando su eficacia contra la validación del modelo construido con la muestra de entrenamiento.

Capítulo 4

Métodos para selección de variables.

Si bien en las pruebas de significancia se tienen métodos para distinguir covariables que influyen en la variable independiente de las que no necesariamente tienen influencia, existen metodologías que pueden aplicarse antes de ajustar el modelo para seleccionar que variables de las que se poseen influyen en el modelo, este método de trabajo conocido como *Forward Stepwise* tiene la ventaja de que para grandes volúmenes de datos los algoritmos de ajuste de regresión pueden ser muy costos en términos de tiempo y poder de cómputo requerido debido a que son problemas de optimización, para el caso de la regresión logística múltiple, un estimador de máxima verosimilitud.

Por lo anterior, conviene hacer una selección previa de variables reduciendo el número de estas y reduciendo el tiempo que le tomaría al algoritmo de ajuste en ejecutarse.

4.1. Experiencia y relaciones de causalidad.

Un punto importante a notar en la selección de variables y en el trabajo estadístico en general, es el análisis de correlación y causalidad, que dos variables tengan correlación o que sean influyentes para un modelo no significa que dichas variables tengan una relación de causa y que puedan usarse para explicar un comportamiento, es por esto que el trabajo estadístico debe estar acompañado de experiencia en el tema que se analiza, ya sea que el conocimiento del tema lo tenga el encargado del análisis estadístico o por medio de un asesor experto en el tema de interés, de no ser así se puede llegar a conclusiones que no describan la realidad del fenómeno estudiado.

Como un ejemplo absurdo de lo mencionado tomemos la siguiente tabla que representa el número de muertes por caida de cama en Estados Unidos y las importaciones de petroleo canadiense por Estados Unidos[12]

[12]En millones de barriles

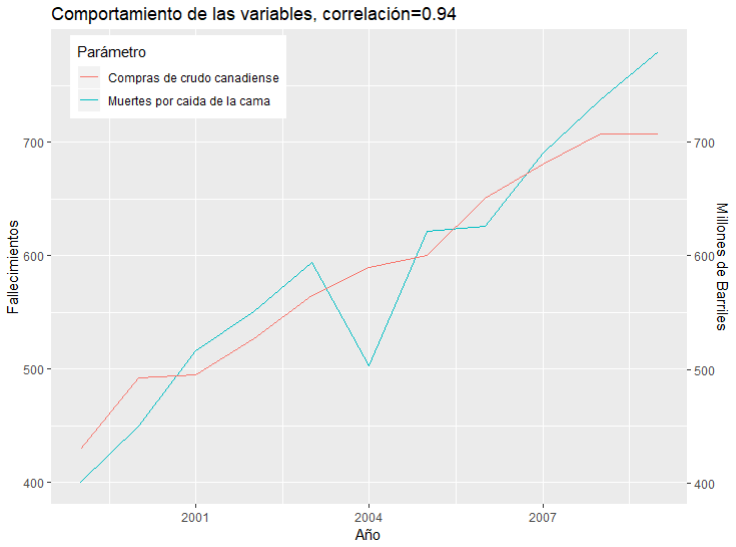
Tabla 4.1: Muertes por caidas de la cama e importaciones de crudo canadiense por USA

Año	Muertes por caida de la cama	Importaciones de crudo canadiense
1999	400	430
2000	450	492
2001	516	495
2002	551	527
2003	594	565
2004	503	590
2005	621	600
2006	626	651
2007	690	681
2008	737	707
2009	780	707

Fuente:Elaboración Propia. Con datos CDC, 2017 [8] y EIA, 2018 [13]

Si observamos la tendencia de ambas variables es similar, además al calcular el coeficiente de correlación para ambas variables se obtiene 0.94, es decir están muy relacionadas en un análisis que no toma en cuenta la naturaleza de las variables.

Figura 4.1: Comportamiento gráfico de las variables del ejemplo



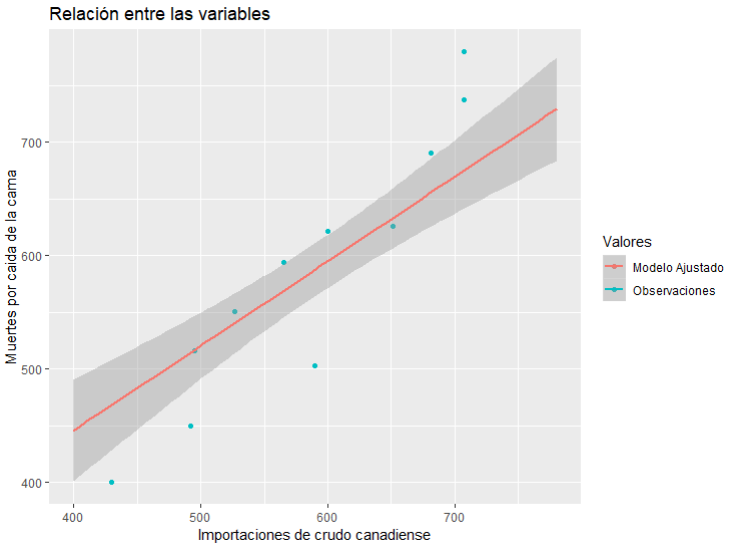
Fuente:Elaboración Propia. con los datos de la tabla 4.1

Si dejamos de lado el significado de las variables y la evidente ausencia de relación entre ellas podemos construir un modelo de regresión lineal que explique la variable, muertes por caida de cama y, en función de las compras de crudo canadiense x, el modelo sería el siguiente:

$$y = -106.51 + 1.18x$$

La relación gráfica entre las variables sería la siguiente:

Figura 4.2: Relación entre las variables del ejemplo



Fuente: Elaboración Propia. con los datos de la tabla 4.1

El coeficiente de determinación R^2 es de 0.88 las pruebas de bondad de ajuste y significancia se cumplen. El único problema es la evidente falta de relación entre las variables.

Este ejemplo es un absurdo de lo que puede ocurrir con una mala selección de variables, la herramienta estadística nos entrega un modelo aceptable. Sin embargo, la interpretación raya en lo ridículo debido a que las variables no tienen relación alguna, es por esto que es necesaria experiencia y conocimiento sobre el tema de interés como un complemento al trabajo estadístico.

4.2. Una variable predictiva ideal.

Antes de introducirnos en los métodos para seleccionar las variables predictoras es conveniente definir una variable predictora ideal. Debido a que en el proceso de selección de variables no se tiene un modelo multivariado, el análisis se hará de manera univariada comparando el comportamiento de la variable predictora contra el comportamiento de la variable independiente.

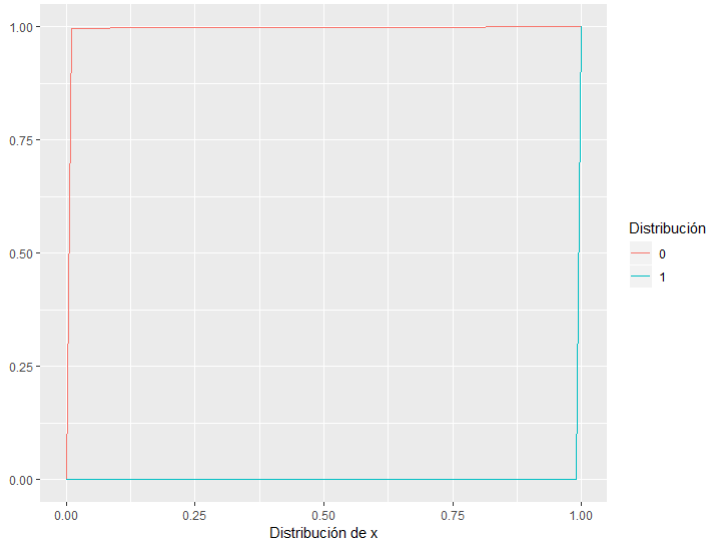
Como fue mencionado en el capítulo anterior, las variables de interés para construir un modelo clasificatorio cumplen con la característica de ser numéricas o categóricas ordinales por lo que si tenemos observaciones en forma de duplas (x_i, y_i) en donde x_i es

nuestra variable candidata a predictora y y_i es nuestra variable dependiente, podemos ordenar a estas duplas respecto a la variable x_i.

Con este orden lo deseable sería que y_i tuviera un comportamiento ordenado, al ser dicotómica entonces lo deseable es que primero se observarán todos los elementos con un valor y al final todos los elementos con el otro valor, siendo irrelevante el orden ya que sólo nos interesa que la variable x_i nos ayude a separar a la variable y_i.

Si graficamos la distribución de la variable x_i contra la distribución de cada valor que puede tomar y_i la gráfica se vería de la siguiente manera:

Figura 4.3: Una variable clasificadora óptima



Fuente:Elaboración Propia. con datos de Siddiqi, 2006 [24]

4.3. Estadístico Kolmogorov-Smirnov.

El estadístico Kolmogorov-Smirnov**KS**) es usualmente utilizado para medir la bondad de ajuste entre distribuciones de variables[13] permitiendo saber de una manera sencilla si dos variables aleatorias tienen la misma distribución.

Para el análisis de variables clasificatorias se usa con la intención opuesta, la idea es similar a la que se dió en la seccion anterior, es decir, comparar la distribución de cada valor posible de y_i contra la distribución de x_i, en este contexto el estadístico KS se define de la siguiente manera:

[13]Una descripción más profunda de este uso puede consultarse en J. D. Gibbons, 2003 [17]

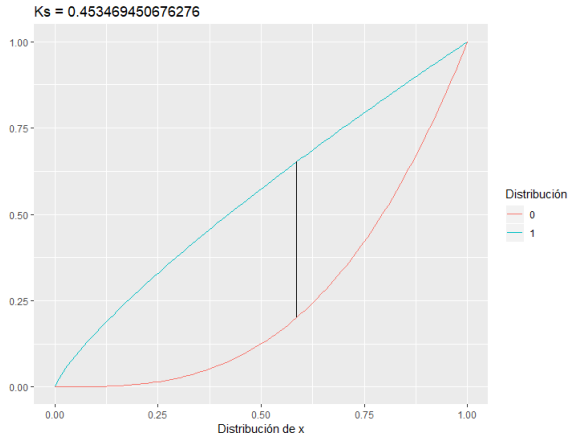
$$KS = max\{|F_P(x) - F_N(x)|\} \tag{4.1}$$

Donde $F_P(x)$ y $F_N(x)$ son respectivamente las funciones de distribución de los valores positivos y negativos de la variable dependiente.

En la literatura, May recomienda en su libro (Mays, 2011 [20]) rechazar variables en las que el estadístico KS sea menor a 0.15 o 15 % y dudar de aquellas con valor superior a 0.7 debido a que al ser demasiado buenas podrían indicar un error de selección.

Gráficamente una variable con buena capacidad predictora desde el punto de vista del estadístico KS sería la siguiente:

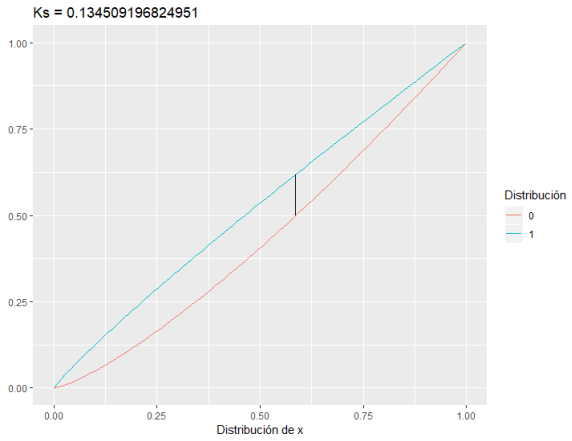
Figura 4.4: Una variable clasificadora buena bajo el estadístico KS



Fuente: Elaboración Propia.

Mientras que una mala variab le clasificatoria sería como la siguiente:

Figura 4.5: Una variable clasificadora mala bajo el estadístico KS



Fuente:Elaboración Propia.

4.4. IV y WOE.

El estadístico KS nos da una idea sobre que tan diferentes son dos distribuciones. Sin embargo, sólo nos da una idea general del comportamiento de la variable, no nos habla de la evolución de la variable, salvo al analizar la gráfica de las distribuciones.

Los estadísticos analizados en esta sección *IV* y *WOE* nos permiten analizar de manera simultánea el comportamiento predictivo de la variable por grupos o categorías *WOE* y de manera general *IV*, primero se analizará el *WOE* debido a que el *IV* es una medida resumen del mismo.

4.4.1. Weight Of Evidence

El estadístico *WOE* nos permite analizar la diferencia entre las distribuciones de los valores positivos y negativos de la variable independiente contra la variable que nos interesa añadir al modelo.

Para esto el estadístico usa la carácterística de que las variables son numéricas o categóricas ordenadas por lo que se pueden ordenar y agrupar bajo este orden, las razones para agrupar las variables son las siguientes:

- Ofrece una solución para lidiar con valores atípicos.
- Permite explicar la naturaleza de la relación entre la variable independiente y las covariables conforme estas avanzan.

- Permite analizar relaciones no lineales analizándolas de manera local.

Una vez categorizados los datos se obtienen las distribuciones de las observaciones de la covariable de interés que para cada categoría C_i son las siguientes:

$$F_P(C_i) = P[x_j \in C_i | y_j = 1] \qquad (4.2)$$
$$F_N(C_i) = P[x_j \in C_i | y_j = 0]$$

El *WOE* se basa en comparar la proporción entre los "buenos" y los "malos" de cada grupo, para la categoría C_i de la covariable de interés su *WOE* es el siguiente:

$$W_i = ln(\frac{F_P(C_i)}{F_N(C_i)})x100 = ln(F_P(C_i) - F_N(C_i))x100 \qquad (4.3)$$

Un valor negativo nos indica que hay una mayor proporción de observaciones negativas (*ceros*) de la variable dependiente, el comportamiento deseable de esta es que tenga una tendencia progresiva ya sea a crecer o decrecer con un único cambio de signo, Esto significa que uno de los comportamientos se da mas al inicio de la variable de interés y conforme esta crece dicho comportamiento va decreciendo en favor del otro. Haciéndolo consistente con un estadístico KS deseable.

Para ejemplificar un buen WOE observemos la tabla **4.2**, construida con datos ejemplo obtenidos de Siddiqi, 2006 [24]. Esta tabla se refiere a un modelo aplicado en pago de créditos donde los denominados "buenos" son las personas que pagan sus créditos, mientras los denominados "malos" son quienes dejaron de pagar un crédito.

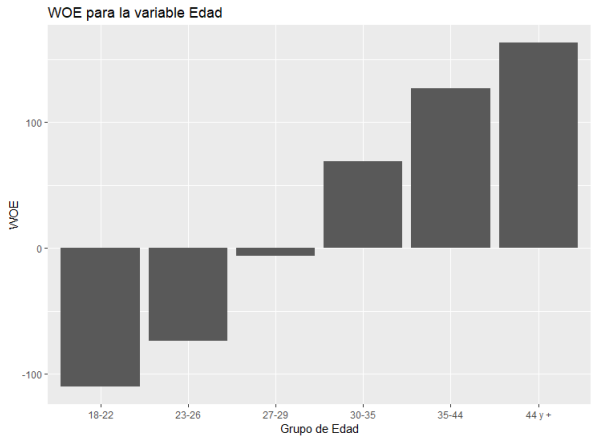
Tabla 4.2: Relación entre edad y la falta de pagos en créditos, ejemplo

Edad	Frecuencia Absoluta	Buenos	Malos	Distribución Buenos	Distribución malos Malos	WOE
18-22	4000	3040	960	0.09	0.26	-110.29
23-26	6000	4920	1080	0.14	0.29	-73.92
27-29	9000	8100	900	0.23	0.24	-5.83
30-35	10000	9500	500	0.27	0.14	68.89
35-44	7000	6800	200	0.19	0.05	127.08
44 y +	3000	2940	60	0.08	0.02	163.63

Fuente: Siddiqi, 2006 [24]

En la tabla **4.2** se puede observar como el WOE pasa progresivamente de valores negativos grandes a valores positivos, una manera común de analizar esto es mediante un histograma con los valores del *WOE*.

Figura 4.6: WOE para el ejemplo de la variable edad



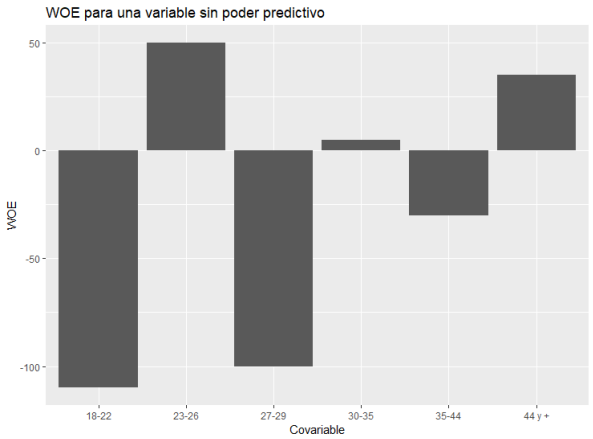
Fuente:Elaboración Propia. con los datos de la tabla 4.2

Un comportamiento deseable es el mostrado en el ejemplo anterior o el inverso (pasar de positivo a negativo) lo que no es aceptable en una variable de interés son múltiples cambios de signo debido a que esto reflejaría un comportamiento similar entre F_P y F_N.

Es posible que algunos cambio de signo no deseados puedan ser corregidos cambiando la manera en la que se agrupa la covariable de interés, el abuso de esto debe ser evitado para no caer en el sobreajuste del modelo (Siddiqi, 2006 [24]).

La siguiente gráfica ejemplifica un comportamiento no deseado para el *WOE* en las covariables de interés observando múltiples cambios de signo.

Figura 4.7: WOE para una covariable sin poder predictivo



Fuente: Elaboración Propia.

4.4.2. Information Value.

El comportamiento del *WOE* nos permite analizar que si una covariable puede ser considerada como predictiva o no, pero no nos dice que tan predictiva es, para esto hay un estadístico asociado al *WOE* que nos indica que tan predictiva es la variable.

El estadístico IV cumple con esta función y es la suma ponderada bajo las categorías del WOE.

$$IV = \sum_{i=1}^{n} (F_P(C_i) - F_N(C_i)) * ln(\frac{F_P(C_i)}{F_N(C_i)}) \tag{4.4}$$

Por como se define es destacable que el IV nunca toma valores negativos, es cero si $F_P(C_i) = F_N(C_i)$

El valor de este estadístico nos indicará que tan predictiva es la covariable de interés, la clasificación de esta predictibilidad es mencionada por Anderson, 2007 [4] de la siguiente manera:

Las variables con IV superiores a 0.5 pueden estar asociadas a un error de selección o con un factor dependiente de la variable, y su relación con la variable dependiente debe ser analizada de manera individual.

Tabla 4.3: Valores del estadístico IV y su relación con la predictibilidad de las covariables

Valor de IV	Capacidad predictiva
$IV < 0.02$	Sin valor predictivo
$0.02 \leq IV < 0.1$	Poca predictibilidad
$0.1 \leq IV < 0.3$	Predictibilidad media
$0.3 \leq IV < 0.5$	Alta predictibilidad
$0.5 \leq IV$	Valores atípicos

Fuente: Anderson, 2007 [4]

4.5. Uso de variables categóricas no ordinales.

Los supuestos del modelo de regresión logística requieren de variables numéricas o categóricas que se puedan ordenar. Sin embargo, hay variables que puedan ser de interés pero al no ser ordinales estas no podrían ser usadas.

Para poder utilizar este tipo de variables es conveniente separarlas en m o $m-1$ variables dicotómicas dependiendo de la situación, donde m es el número de categorías que tiene la variable de interés, convirtiendo a cada categoría en una variable con valor 0 o 1. (P. de Jong, 2008 [21])

Para ejemplificar esto tomemos el siguiente ejemplo para observaciones de una hipotética variable llamada color:

Tabla 4.4: Variable categórica ejemplo, color

	color
1	amarillo
2	verde
3	verde
4	blanco
5	azul
6	azul
7	verde
8	negro
9	negro

Fuente: Elaboración Propia.

La variable es categórica y no es útil de la forma en la que están presentados los datos, ya que no hay manera de ordenar algo como los colores. Pero, es posible utilizarlos cambiando la representacion de los datos con el procedimiento descrito anteriormente. Obteniendo la tabla **4.5**.

Tabla 4.5: Transformación de la variable color en múltiples variables

	amarillo	azul	blanco	negro	verde
1	1	0	0	0	0
2	0	0	0	0	1
3	0	0	0	0	1
4	0	0	1	0	0
5	0	1	0	0	0
6	0	1	0	0	0
7	0	0	0	0	1
8	0	0	0	1	0
9	0	0	0	1	0

Fuente:Elaboración Propia.

En algunos casos es posible que la variable tenga una categoría que englobe casos no contemplados o casos por eliminación como la categoría *otros*. P. de Jong, 2008 [21] propone nombrar a esta categoría como categoría base, si existe una categoría base para la variable, es posible convertirla utilizando una variable menos, siendo la categoría base el caso en el que todas las demás posibilidades son cero. Para ejemplificar este caso tomemos la siguiente tabla con observaciones de la variable mascota.

Tabla 4.6: Variable ejemplo, mascota

	mascota
1	perro
2	gato
3	otro
4	perro
5	otro
6	gato
7	gato
8	perro
9	otro
10	perro

Fuente: Elaboración Propia.

En este ejemplo la categoría base es *otro* por lo que en lugar de tener 3 variables dicotómicas nuevas se puede reducir a sólo 2. Siendo el tercer caso la ausencia de perro y gato.

Tabla 4.7: Transformación de la variable mascota en dos variables

	perro	gato
1	1	0
2	0	1
3	0	0
4	1	0
5	0	0
6	0	1
7	0	1
8	1	0
9	0	0
10	1	0

Fuente: Elaboración Propia.

Capítulo 5

Propuesta del Modelo.

Los datos utilizados en este trabajo provienen del **Reporte Regulatorio Número 8** que entregan las aseguradoras de manera anual a la **Comisión Nacional de Seguros y Fianzas**

Todos los cálculos fueron realizados con el apoyo del paquete estadístico **R** en su versión 3.5.1, el código generado para el ajuste y comprobación del modelo se puede revisar en el Anexo E

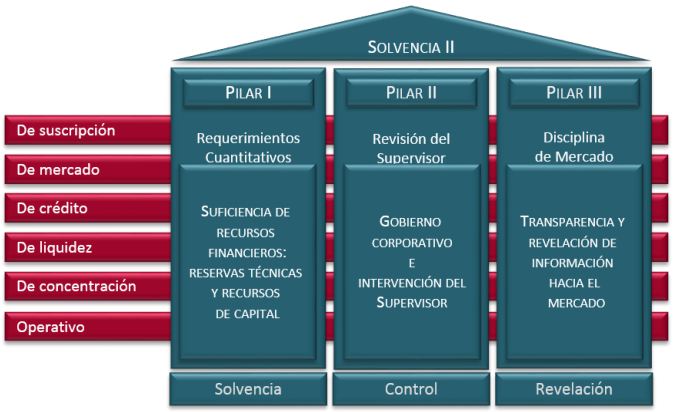
5.1. El Reporte Regulatorio Número 8 de Comisión Nacional de Seguros y Fianzas.

El primero de Abril de 2016 entró en vigor para México la *Ley de Instituciones de Seguros y Fianzas*, esta ley implementa el modelo mexicano de supervisión basada en riesgos que es compatible con el modelo europeo conocido como *Solvencia II*.

Este modelo se basa en 3 pilares fundamentales:

- Solvencia financiera.
- Control.
- Revelación de la información.

Figura 5.1: Pilares del modelo mexicano de supervisión basada en riesgos



Fuente: CNSF, 2017 [10]

El primer pilar se refiere a los requerimientos para garantizar la solvencia de las instituciones a través de cinco grandes conceptos: Reservas Técnicas, Requerimiento de Capital de Solvencia, Fondos Propios Admisibles, Inversiones y Reaseguro; asimismo, contempla la generación de un Balance Económico en el cual los activos y pasivos deben ser valuados considerando su valor de mercado.

El segundo pilar contempla aspectos relativos a la organización, al gobierno corporativo, así como a la vigilancia que las propias instituciones deben mantener sobre los riesgos inherentes a su operación.

El tercer pilar se refiere a la revelación de información por parte de las aseguradoras tanto al público en general como a las autoridades reguladoras, esta información es entregada a *CNSF* para su análisis por medio de 13 reportes regulatorios. El Reporte Regulatorio 8 (RR-8) denominado reporte de información estadística es el reporte en el que las instituciones de seguros revelan datos sobre sus pólizas en las operaciones de seguros que tienen autorizadas.

En dicho reporte se concentra información sobre todas las pólizas emitidas por las aseguradoras autorizadas en el país, siendo separada la información por operaciones en las distintas ramas del seguro.

Los datos utilizados provienen de aquellos reportados para primas emitidas del seguro de vida individual en los años 2016 y 2017.

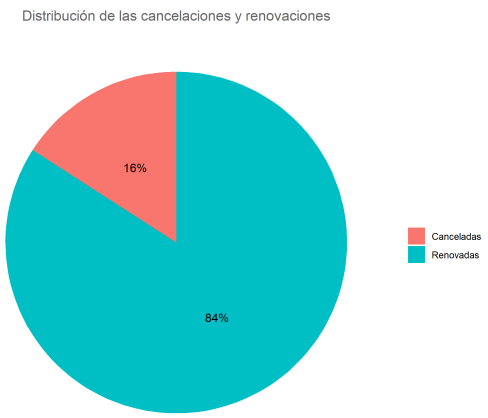
5.2. Descripción de los datos.

El criterio que se utilizó para definir si una póliza fue o no cancelada es: "Se considera cancelada una póliza si esta termina su vigencia antes de la fecha estipulada en la misma póliza por causas distintas al riesgo cubierto por el seguro"

Bajo el criterio anterior, no se consideran como canceladas aquellas pólizas que llegan al término de la misma por realizarse el riesgo cubierto o por cumplirse la vigencia de la misma.

Utilizando el criterio antes descrito, del total de $21,571,493$ pólizas analizadas, resultó que $3,428,932$ fueron canceladas, estas representan el $15.89\,\%$ del total

Figura 5.2: Distribución de las pólizas canceladas y renovada



Fuente: Elaboración Propia.

La tabla de primas emitidas del seguro de vida contiene 47 campos o variables con información sobre cada póliza de seguros de vida individual emitida en el periodo mencionado. Una descripción detallada de todas las variables que conforman el **RR-8** puede encontrarse en el Anexo B.

Fueron retiradas variables que no eran relevantes para el modelo, y las variables que son función de una o más de las demás variables, por ejemplo, los montos de prima ya que estas dependen de las sumas aseguradas. Las variables que no fueron retiradas, son las siguientes:

- MONEDA
- EDAD
- FUMADOR
- SEXO
- STATUS_POL
- STATUS_CERT
- PER_ESPERA
- SA_BEN1
- SA_BEN2
- SA_BEN3
- SA_BEN4
- SA_BEN6
- SA_BEN8
- SA_BEN9
- ENTIDAD
- FORM_VENTA
- MOD_POL
- PLAN_POL
- DIVIDENDO
- SUBT_SEG
- INI_COBER
- PZO_PGO_PMA
- TIP_RGO_ASOC
- EX_PMA_MBAS

Debido a que las variables MONEDA, FORM_VENTA, ENTIDAD, MOD_POL y PLAN_POL son categóricas no ordinales, fueron separadas en múltiples variables dicotómicas como se menciona en la sección 4.5

5.3. Selección de las variables.

De las variables no eliminadas se seleccionarán aquellas que sean consideradas como predictivas para el modelo utilizando los métodos descritos en el capítulo 4, se utilizarán *el estadístico KS, IV y WOE* para evaluar el grado de predictibilidad que tienen las variables de las que se dispone.

5.3.1. WOE e IV.

A continuación se muestra la tabla **5.1** que contiene las variables que por el valor de su IV se pueden considerar como predictivas, se puede consultar el valor de IV para cada variable en el Anexo C.1

Tabla 5.1: Variables predictivas bajo los criterios de *IV*

Variable	IV	Predictibilidad
STATUS_POL	19.696796	Valor atípico
SA_BEN1	1.244716	
SA_BEN3	0.756177	Alta predictibilidad
Entidad_ 9	0.510054	
SDO_FADMON	0.425667	
ANIO_POLIZA	0.253993	
PZO_PGO_PMA	0.121929	Predictibilidad media
PER_ESPERA	0.111008	
SA_BEN2	0.053575	Poca predictibilidad
Entidad_ 15	0.050294	
EDAD	0.041812	
Sexo_ F	0.034252	Poca predictibilidad
Sexo_ M	0.034252	
Entidad_ 7	0.029667	
EMISION	0.026219	
SA_BEN4	0.024499	

Fuente: Elaboración Propia.

La Variable: *STATUS_POL* tiene un *IV* atípico, estos valores indican que puede haber un sobre modelado o una relación directa con la variable dependiente haciendo necesaria una revisión más a fondo. En este caso si se observa este tipo de relación directa con la variable dependiente, ya que esta variable indica si la póliza está o no cancelada, por lo que no se utilizará en la elaboración del modelo.

Después de analizar el *WOE*[14] para cada una de las variables enunciadas en la tabla 5.1 los resultados fueron los siguientes:

Tabla 5.2: Variables predictivas bajo los criterios de *IV y WOE*

STATUS_POL	Valor atípico	Rechazada	Correlación con la variable dependiente
SA_BEN1	Valor atípico	Aceptada	
SA_BEN3	Alta predictibilidad	Aceptada	
Entidad_ 9	Alta predictibilidad	Aceptada	
SDO_FADMON	Alta predictibilidad	Aceptada	
ANIO_POLIZA	Alta predictibilidad	Rechazada	Múltiples cambios de signo
PZO_PGO_PMA	Predictibilidad media	Rechazada	Múltiples cambios de signo

Fuente: Elaboración Propia.

[14]Se pueden consultar las gráficas correspondientes al *WOE* para cada variable en el Anexo C.2

Tabla 5.2: Variables predictivas bajo los criterios de *IV y WOE*

PER_ESPERA	Predictibilidad media	Rechazada	Múltiples cambios de signo
SA_BEN2	Poca predictibilidad	Aceptada	
Entidad_ 15	Poca predictibilidad	Aceptada	
EDAD	Poca predictibilidad	Aceptada	
Sexo	Poca predictibilidad	Aceptada	
Entidad_ 7	Poca predictibilidad	Rechazada	Múltiples cambios de signo
EMISION	Poca predictibilidad	Aceptada	
SA_BEN4	Poca predictibilidad	Aceptada	

Fuente: Elaboración Propia.

5.3.2. Estadístico KS.

Si bien ya se disponen de algunas variables seleccionadas bajo los criterios de IV y WOE, es conveniente disponer de una segunda prueba para confirmar las decisiones o retirar algunas variables que no tengan el comportamiento deseado. En la sección 4.3 se presentó el uso de el estadístico *Kolmogorov-Smirnov* como un método para la selección de variables, teniendo como criterio rechazar variables en las que el estadístico KS sea menor a 0.15 o 15 % y dudar de aquellas con valor superior a 0.7 ya que al ser demasiado buenas podrían indicar un error de selección.

Después de analizar el estadístico *KS* para cada variable se obtienen las siguientes variables como predictoras[15] con sus respectivos valores para el estadístico *KS*.

[15]Se pueden consultar todos los valores y gráficas en el Anexo

Tabla 5.3: Variables predictoras bajo el criterio del estadístico *Kolmogorov-Smirnov*

Variable	KS
SA_BEN1	0.94
SA_BEN6	0.32
SA_BEN4	0.29
ANIO_POLIZA	0.29
Plan_ 1	0.29
SA_BEN8	0.27
SA_BEN9	0.27
SA_BEN3	0.24
Mod_ 1	0.23
Plan_ 3	0.21
EDAD	0.19
Mod_ 2	0.19
EMISION	0.18
SA_BEN2	0.16

Fuente: Elaboración Propia.

5.3.3. Variables seleccionadas.

Combinando ambos criterios y eligiendo como variables predictoras a las variables que cumplen con ambos, las variables que serán utilizadas en el ajuste del modelo son las siguientes:

Tabla 5.4: Variables predictivas bajo ambos criterios

Variable
SA_BEN1
SA_BEN3
SA_BEN2
EDAD
Sexo
EMISION
SA_BEN4

Fuente: Elaboración Propia.

5.4. Ajuste del modelo.

Previo al ajuste del modelo se separó una muestra aleatoria con el 30 % de los datos para comprobar el modelo contra nuevas observaciones de los datos como se menciona en el capítulo 4 a estos datos se les denominará como muestra de pruebas mientras que a los restantes muestra de entrenamiento.

Después de ajustar el modelo con la muestra de entrenamiento se obtuvieron los siguientes coeficientes:

Tabla 5.5: Coeficientes para el modelo de regresión logística

Variable	Coeficiente
(Intercept)	0.8009789082
SA_BEN1	-0.0000000347
SA_BEN2	-0.0000001586
SA_BEN3	-0.0000004102
SA_BEN4	0.0000004243
EDAD	0.0120757116
Sexo	0.4178572740
EMISION	0.2058676624

Fuente: Elaboración Propia.

5.5. Significancia del modelo.

El proceso de selección de variables debería garantizar que dichas variables sean significativas, sin embargo es posible que las variables que se encuentran en la frontera de las pruebas de predictibilidad realizadas no sean significativas para el modelo, es decir, $\beta_i = 0$, en la sección *3.4* mencionamos 2 pruebas de significancia marginal para las variables del modelo, dichas pruebas fueron aplicadas al mismo.

5.5.1. Razón de Verosimilitudes.

Las hipótesis de esta prueba son:

$$\begin{aligned} \mathcal{H}_0 : &\quad \beta_i = 0 \\ \mathcal{H}_1 : &\quad \beta_i \neq 0 \end{aligned}$$

Después de calcular la razón de verosimilitudes para cada variable escogida para el modelo, se obtuvieron los siguientes *p-value*:

Tabla 5.6: *p-value* obtenidos para la prueba Razón de verosimilitudes por variable

Variable	*p-value* razón de verosimilitudes
SA_BEN1	$p\text{-}value < 2^{-52}$
SA_BEN2	$p\text{-}value < 2^{-52}$
SA_BEN4	$p\text{-}value < 2^{-52}$
SA_BEN3	$p\text{-}value < 2^{-52}$
EDAD	$p\text{-}value < 2^{-52}$
Sexo	$p\text{-}value < 2^{-52}$
EMISION	$p\text{-}value < 2^{-52}$

Fuente: Elaboración Propia.

Para cada variable se obtuvo un valor inferior a la constante ϵ_{mach} que en un sistema de 64-bit corresponde con 2^{-52}, ese ínfimo valor permite que para la significancia $\alpha = 0.05$ podamos rechazar la hipótesis nula en favor de la hipótesis alternativa, es decir, que no se cumple que $\beta_i = 0 \qquad \forall i \in \{1...7\}$.

Cabe aclarar que para las variables correspondientes a las sumas aseguradas, el valor pequeño de los coeficientes correspondientes no se debe a que estadísticamente $\beta_i = 0$, se debe a la magnitud de las variables que están en miles y hasta millones de pesos.

5.5.2. Prueba de Wald.

Las hipótesis de esta prueba son las mismas que las de la anterior, es decir:

$$\begin{aligned} \mathcal{H}_0 : &\qquad \beta_i = 0 \\ \mathcal{H}_1 : &\qquad \beta_i \neq 0 \end{aligned}$$

A diferencia de la prueba anterior la prueba de Wald o estadístico z si toma en cuenta a β_0 usualmente conocido como parámetro interceptor, aunque desde el punto de vista de la significancia, el parámetro interceptor no es tan importante debido a que no permite el análisis sobre ninguna variable en particular.

El *p-value* obtenido de la prueba de Wald para cada variable son los siguientes:

Tabla 5.7: *p-value* obtenidos para la prueba de razón de verosimilitudes por variable

Variable	*p-value* razón de verosimilitudes
(Intercept)	$p\text{-}value < 2^{-52}$
SA_BEN1	$p\text{-}value < 2^{-52}$
SA_BEN2	$p\text{-}value < 2^{-52}$
SA_BEN4	$p\text{-}value < 2^{-52}$
SA_BEN3	$p\text{-}value < 2^{-52}$
EDAD	$p\text{-}value < 2^{-52}$
Sexo	$p\text{-}value < 2^{-52}$
EMISION	$p\text{-}value < 2^{-52}$

Fuente: Elaboración Propia.

Para cada variable se obtuvo un valor inferior a la constante ϵ_{mach}, ese ínfimo valor permite que para la significancia $\alpha = 0.05$ podamos rechazar la hipótesis nula en favor de la hipótesis alternativa, es decir, que no se cumple que $\beta_i = 0 \qquad \forall i \in \{0...7\}$.

5.6. Bondad de ajuste.

Una vez determinado que los coeficientes del modelo no son cero es conveniente revisar que tanto se ajustan los valores obtenidos del modelo a la realidad observada en los datos, en la sección 2.5 se mencionan tres pruebas estadísticas para analizar esta propiedad.

5.6.1. Prueba χ^2.

Esta prueba tiene las siguientes hipótesis:

$$\mathcal{H}_0 : y_i \leq \hat{y}_i$$
$$\mathcal{H}_1 : y_i = \hat{y}_i$$

En el escenario de que el número de patrones de covariables J es similar al número de variables n el *p-value* obtenido para esta prueba es menor a ϵ_{mach} por lo que bajo el

supuesto anterior se rechaza la hipótesis nula en favor de la hipótesis alternativa, es decir $y_i = \hat{y}_i$

5.6.2. Prueba de Hosmer-Lemeshow.

Esta prueba tiene las siguientes hipótesis:

$$\mathcal{H}_0 : y_i = \hat{y}_i$$
$$\mathcal{H}_1 : y_i \neq \hat{y}_i$$

Esta prueba no depende del supuesto de que $J \approx n$ pero a su vez tiene el problema de que se deben de conocer el número de patrones de covariables para los datos observados cosa que en la práctica puede volverse complejo más si el número de datos es muy grande, por lo que lo más correcto sería simular múltiples patrones buscando encontrar un resultado satisfactorio el dichos patrones.

Al hacer una simulación de 1,000 posibles patrones de covariables para los datos se obtuvo que en el 98.8 % de las observaciones el *p-value* era inferior a 0.05 por lo que en esta prueba con una confianza del 95 % podemos rechazar la hipótesis nula es decir, bajo esta prueba no tenemos ajuste aceptable en el modelo.

5.6.3. Pseudo-R^2.

Para el modelo el pseudo coeficiente de determinación de McFadden es de 0.67, esto nos indica que el modelo tiene un buen ajuste.

Se encontró, que de tres pruebas dos de ellas nos indican que el modelo tiene un ajuste aceptable por lo que aceptaremos que el modelo tiene buen ajuste.

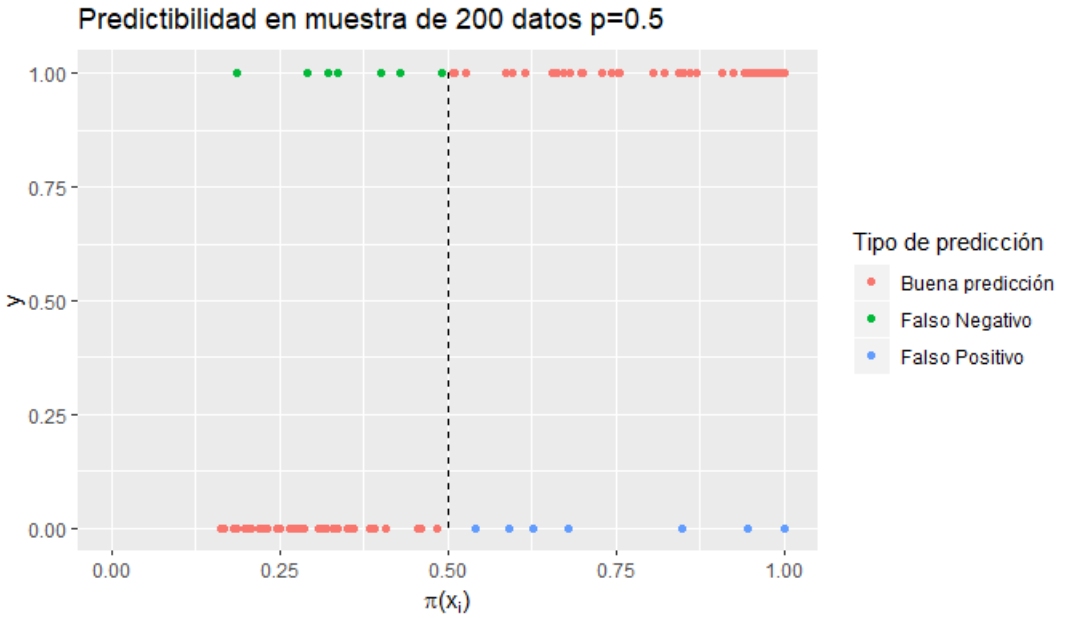
5.7. Validación del modelo.

Después de comprobar que el modelo de regresión ajustado es aceptable en términos de significancia y bondad de ajuste podemos definir el criterio de clasificación p y analizar la capacidad predictiva del modelo por medio de la curva ROC.

5.7.1. Matrices de Confusión y Estadísticos Asociados.

Antes de analizar las matrices de confusión, es conveniente presentar un ejemplo gráfico con los datos para visualizar el efecto que tiene la selección de p sobre los errores.

Figura 5.3: Errores de predicción en una muestra de 200 valores con $p = 0.5$



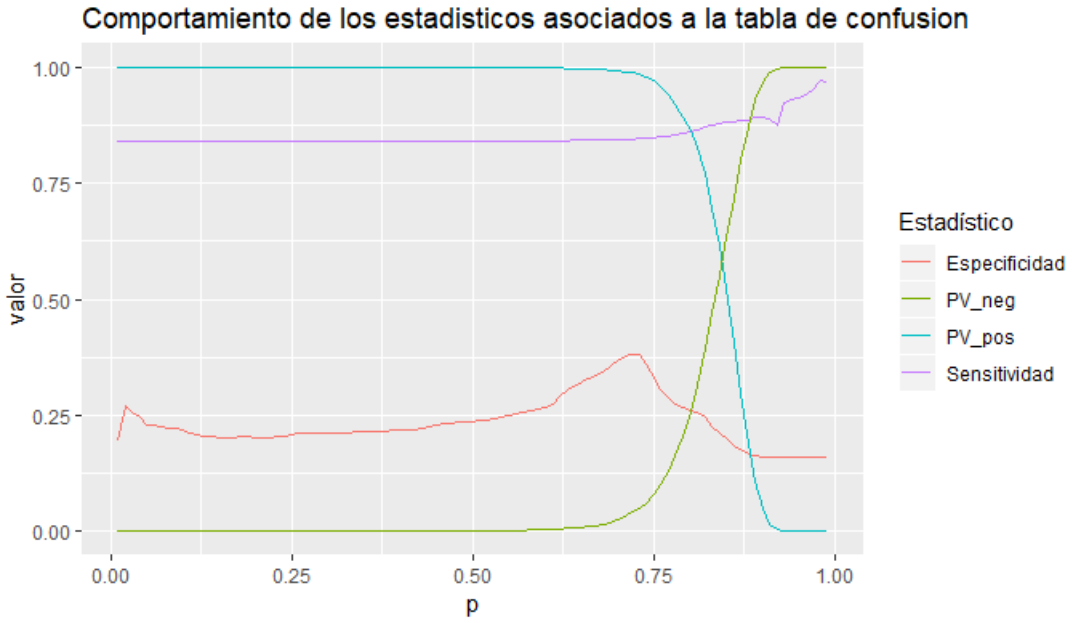
Fuente: Elaboración Propia.

Mover la linea punteada que representa el valor seleccionado de p puede servir para reducir los falsos positivos o falsos negativos pero como se observa en la gráfica esto aumentará el número observado para el otro error, por lo que se hace hincapié en la definición de criterios consistentes con las necesidades del modelos y el contexto en el que se utilizará.

El comportamiento de los estadísticos asociados a la tabla de confusión conforme varía el valor de p[16] es el siguiente:

[16]Se puede encontrar una tabla con los estadísticos y errores asociados a cada valor de p calculado en el Apéndice D

Figura 5.4: Comportamiento de los estadísticos asociados a la tabla de confusión

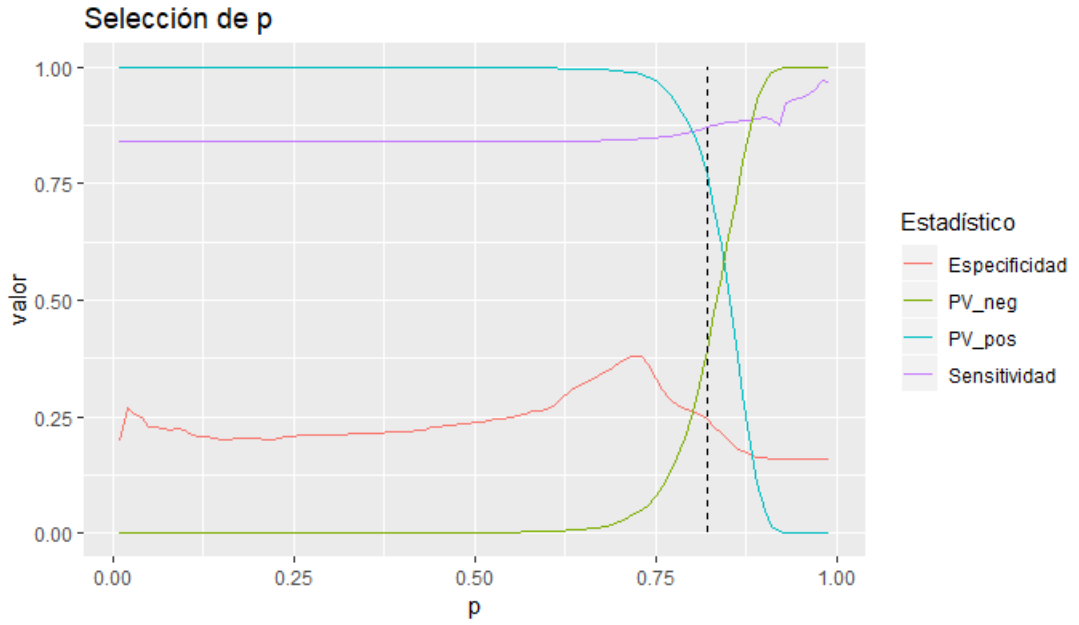


Fuente: Elaboración Propia.

En este trabajo se buscará la mayor eficacia del modelo al momento de predecir, por lo que se desea obtener el valor de p que ofrezca en conjunto un mayor valor para los estadísticos asociados a la tabla de contingencia, para conseguir esto se decidió usar como criterio asignar el valor de p a aquel que produzca el máximo de la suma de los estadísticos asociados a la matriz de confusión.

Para este conjunto de datos el valor de p que nos garantiza lo anterior es 0.82

Figura 5.5: Selección de p



Fuente: Elaboración Propia.

Al observar en la gráfica, la selección cuenta con valores para *Sensitividad,* PV_{pos} *y* PV_{neg} cercanos al máximo siendo castigado solo el valor de Especificidad.

La matriz de confusión asociada a la elección de p es la siguiente:

Tabla 5.8: Matriz de confusión para p=0.82

		Predicción por el modelo	
		$\hat{y} = 0$	$\hat{y} = 1$
observaciones	$y = 0$	946,023	2,915,711
	$y = 1$	1,454,333	9,783,978

Fuente: Elaboración Propia.

Con este criterio el total de Falsos Negativos corresponde con el 19.3 % de las predicciones realizadas por el modelo mientras que el número de Falsos Positivos corresponde con el 9.6 % del total de las predicciones realizadas por el modelo, por lo que el total de las predicciones incorrectas asciende al 29.9 %.

Los estadísticos asociados a este valor de p son los siguientes:

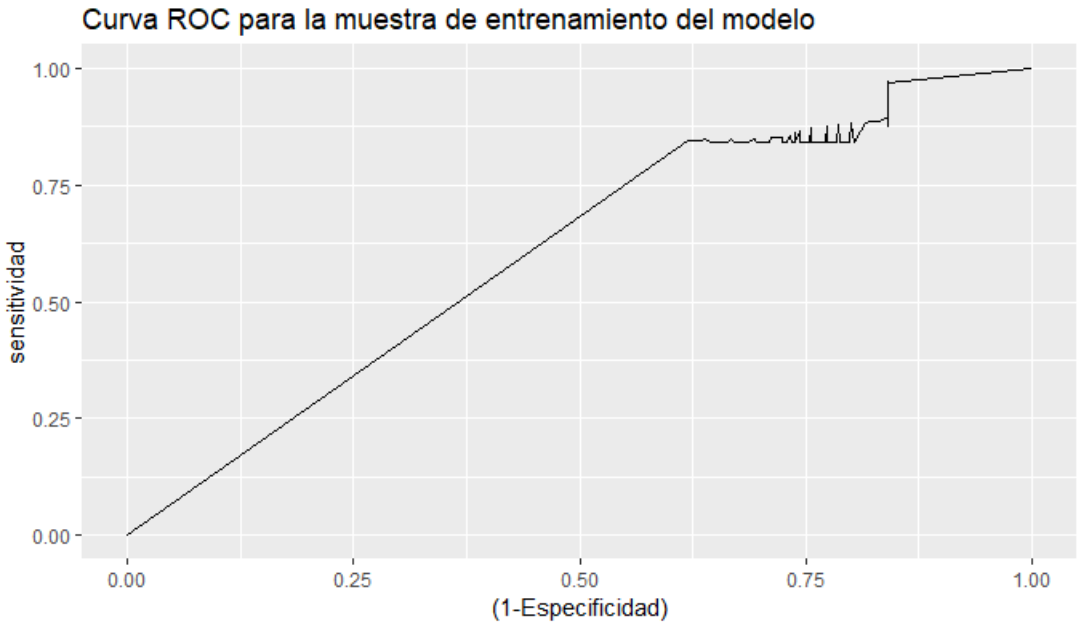
- Sensitividad: 0.8705, es decir dado que se tiene una observación en la que la póliza es renovada el modelo es capaz de determinar el 87.05 % de estas observaciones como renovadas.
- Especificidad: 0.2449, dada una observación en la que la póliza es cancelada el modelo es capaz de determinar al 24.49 % de estas observaciones como canceladas.
- PV_{pos} 0.7704, de las observaciones determinadas por el modelo como pólizas renovadas el 77.04 % corresponden con pólizas renovadas.
- PV_{neg} 0.6005 de las observaciones determinadas por el modelo como pólizas canceladas, el 60.05 % de estas corresponden con pólizas canceladas.

Los valores de $Especificidad$ y PV_{neg} nos indican que el criterio elegido favorece los Falsos Positivos reduciendo los casos de Falsos Negativos.

5.7.2. Curva ROC.

La curva ROC asociada al modelo es la siguiente:

Figura 5.6: Curva ROC para la muestra de entrenamiento del modelo



Fuente: Elaboración Propia.

El valor del área bajo esta curva ROC calculado utilizando el método del trapecio es el siguiente:

$$AUC = 0.6701 \tag{5.1}$$

Bajo el criterio expuesto en la sección 4.1.1 el modelo tiene una buena capacidad discriminatoria.

5.8. Comprobación del modelo.

Para comprobar la eficacia del modelo se aplicó a la muestra de pruebas, misma que no fue utilizada para el ajuste del mismo, una vez obtenidos los valores de $\pi(\underline{x}_i)$ se aplicará el criterio escogido de $p = 0.82$ para definir el valor predicho por el modelo con el siguiente criterio

$$\begin{aligned} \pi(\underline{x}_i) \leq p &\rightarrow \hat{y}_i = 0 \\ \pi(\underline{x}_i) > p &\rightarrow \hat{y}_i = 1 \end{aligned} \tag{5.2}$$

Se analizará la eficacia del modelo para estas nuevas observaciones utilizando los criterios de validación de la sección anterior y comparándolos con los obtenidos con la muestra de entrenamiento del modelo.

5.8.1. Errores y matriz de confusión

Aplicando el modelo y el criterio de clasificación $p = 0.55$ a los datos reservados en la muestra reservada para pruebas se obtiene la siguiente matriz de confusión.

Tabla 5.9: Matriz de confusión para el modelo aplicado a la muestra de pruebas

		Predicción por el modelo	
		$\hat{y} = 0$	$\hat{y} = 1$
observaciones	$y = 0$	404,804	1,248,634
	$y = 1$	623,722	4,194,238

Fuente: Elaboración Propia.

El total de Falsos Negativos corresponde con el 9.63 % de las predicciones realizadas por el modelo mientras que el número de Falsos Positivos corresponde con el 19.29 %

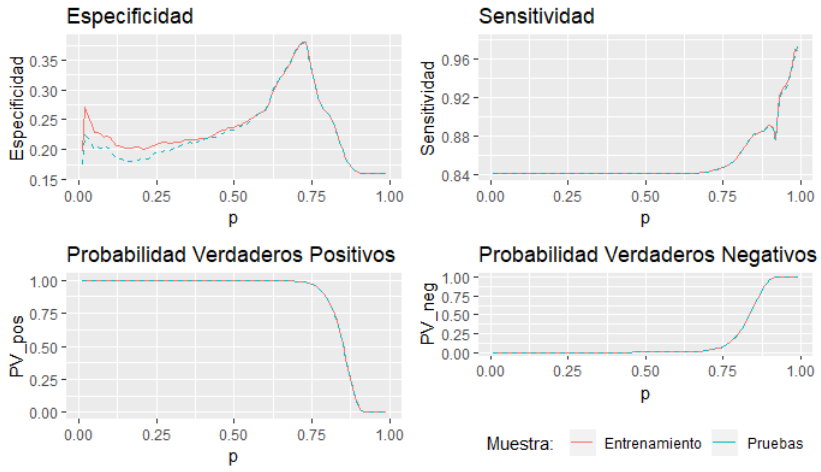
del total de las predicciones realizadas por el modelo por lo que el total de las predicciones incorrectas asciende al 29.92 % del total de observaciones en la muestra de entrenamiento.

Los estadísticos asociados a esta matriz de confusión son los siguientes:

- Sensitividad: 0.8705, es decir dado que se tiene una observación en la que la póliza es renovada el modelo es capaz de determinar el 87.05 % de estas observaciones como renovadas.
- Especificidad: 0.2448, dada una observación en la que la póliza es cancelada el modelo es capaz de determinar al 24.48 % de estas observaciones como canceladas.
- PV_{pos} 0.7705, de las observaciones determinadas por el modelo como pólizas renovadas el 77.05 % corresponden con pólizas renovadas.
- PV_{neg} 0.3935 de las observaciones determinadas por el modelo como pólizas canceladas, el 39.35 % de estas corresponden con pólizas canceladas.

Si comparamos las gráficas de los estadísticos asociados[17] a la matriz de confusión obtenemos lo siguiente:

Figura 5.7: Comparativo de los estadísticos asociados a la matriz de confusión para las muestras de entrenamiento y pruebas



Fuente: Elaboración Propia.

Podemos observar que las proporciones de los errores y por ende los estadísticos asociados a la matriz de confusión son muy similares a los obtenidos con la muestra de

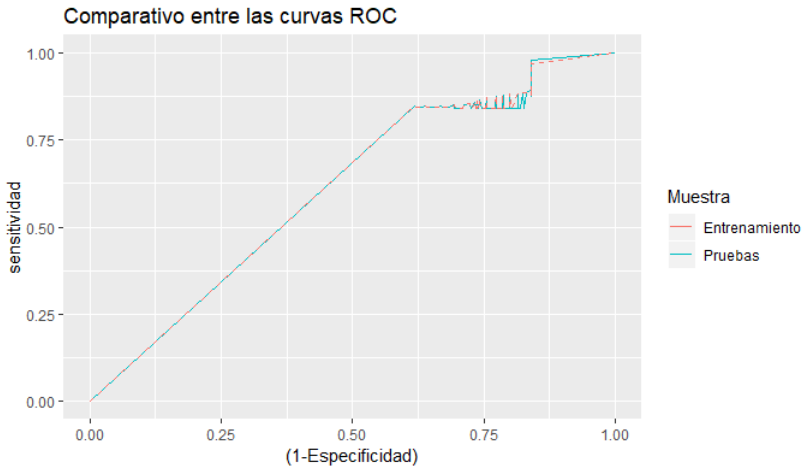
[17] Se puede encontrar una tabla con los estadísticos y errores asociados al modelo aplicado sobre la matriz de pruebas para cada valor de p calculado en el Anexo D.2

entrenamiento por lo que al variar los datos y tomar la muestra de pruebas como datos nuevos obtenemos una eficacia similar en el modelo, además en general el comportamiento de los mismos es muy similar para cada valor de p por lo que bajo estos criterios este modelo puede ser utilizado para predecir el comportamiento en pólizas nuevas.

5.8.2. Curva ROC.

La curva ROC asociada al modelo aplicado sobre la muestra de pruebas comparada con la curva ROC del modelo aplicado a la muestra de entrenamiento es la siguiente:

Figura 5.8: Comparativo entre las curvas ROC asociadas al modelo aplicado a las muestras de entrenamiento y pruebas



Fuente: Elaboración Propia.

En términos del área bajo la curva tenemos un comportamiento muy similar entre lo obtenido para cada muestra, confirmado por el valor del estadístico AUC.

$$AUC_{backtesting} = 0.6703 \tag{5.3}$$

Tenemos un valor muy similar al obtenido con la muestra de entrenamiento por lo que la predictibilidad del mismo no varía por cambiar de conjunto de datos.

Conclusiones

En este trabajo se aplicó un modelo clasificatorio utilizando una regresión logística sobre los datos provenientes del Reporte Regulatorio 8 (**RR-8**) de la Comisión Nacional de Seguros y Fianzas, con la intención de distinguir para el seguro de vida individual las pólizas que serán canceladas de aquellas que no lo serán.

Por medio de las pruebas estadísticas de IV, WOE y estadístico KS, se analizó que variables del **RR-8** tienen valor predictivo para nuestra variable binaria de interés.

Del total de 47 variables que conforman el **RR-8** sólo 7 cumplen con los criterios de predictibilidad para las tres pruebas realizadas, mismas que fueron utilizadas para el ajuste del modelo de regresión logística.

Una vez ajustado el modelo, con ayuda del paquete estadístico R, se realizaron pruebas estadísticas de razón de verosimilitud, Wald, χ^2, Hosmer-Lemeshow y Pseudo R^2 sobre el ajuste para determinar que se cumplen los supuestos del modelo.

Se observó que las pruebas estadísticas aplicadas al modelo, en general se cumplen, siendo una linea que puede desprenderse de este trabajo, investigar si esto se debe a que al selección de variables se hizo pensando en la aplicación de este modelo en particular o es un resultado asociado a la naturaleza de los datos utilizados.

Dentro del proceso de identificación de variables y ajuste, se encontraron problemas asociados a la cantidad de datos contenidos en la base de datos 21,571,493 registros con 47 campos cada uno, obteniendo un total de 1,013,860,171 valores en el conjunto de datos, requiriendo de un equipo con gran cantidad de memoria RAM disponible para poder hacer los cálculos necesarios, además de requerir de una gran cantidad de tiempo para ejecutar el modelo.

Al comprobar por medio de tablas de confusión la eficacia del modelo para los datos de entrenamiento y los datos de prueba, se observó que los resultados obtenidos entre ambos conjuntos son muy similares dándonos certeza sobre la consistencia del modelo consiguiendo una confianza de 70.10 % para la muestra de entrenamiento y 70.08 % para la muestra de pruebas.

De forma general, se puede mostrar que el modelo de regresión logística es adecuado para este tipo de análisis en seguros de vida individual, mediante un análisis similar podría expandirse a otros ramos de seguros.

Desde el punto de vista de comercialización puede utilizarse este o un modelo similar para identificar y evitar pérdidas de suscriptores de pólizas.

Considerando un modelo más preciso que podría combinarse con un sistema modificado de reservas que utilice la probabilidad de cancelación de las pólizas. O incluso utilizar la probabilidad obtenida $\pi(x_i)$ como insumo para el cálculo del requerimiento de capital de solvencia. Sin embargo es importante señalar que se requiere un análisis más profundo y una comparación de la eficacia de las probabilidades obtenidas por el modelo contra las metodologías existentes de cálculo.

Apéndice A

Demostraciones

En este capítulo se presentan alguna demostraciones y conceptos que serán útiles al lector para la lectura de este trabajo.

A.1. Desigualdad de Jensen

En 1906 Johan Jensen demostró la siguiente desigualdad:

Sea $\{\Omega, \mathcal{B}, \mu\}$ un espacio de medida donde $\mu(\Omega) = 1$ Si g es una función integrable en el espacio de medida y φ una función cóncava entonces:

$\varphi(\int_\Omega g d\mu) \leq \int_\Omega \varphi(g) d\mu$

Si limitamos a que $\{\Omega, \mathcal{B}, \mu\}$ sea un espacio de probabilidad y por consecuencia $\mu(\Omega) = 1$, tomamos $E[X]$ como la función integrable y $\varphi = u(x)$ tenemos

$$u(E[X]) \geq E[u(X)] \tag{A.1}$$

Demostración

$$u(x) = u(x_0) + u'(x_0)(x - x_0) + \frac{u''(x_0)(x - x_0)^2}{2} \tag{A.2}$$

como $u''(x) < 0$

$$u(x) \leq u(x_0) + u'(x_0)(x - x_0) \tag{A.3}$$

sea $x = X$ y $x_0 = E[X]$

$$u(X) \leq u(E[X])+u'(E[X])(X-E[X]) \Rightarrow E[u(X)] \leq E[u(E[X])]+E[u'(E[X])(X-E[X])] \tag{A.4}$$

Al ser $E[X]$ un valor conocido $u(E[X])$ y $u''(E[X])$ lo son

$$E[u(X)] \leq u(E[X]) + u'(E[X])E[(X - E[X])] \tag{A.5}$$

Como $E[(X - E[X])] = E[X] - E[X] = 0$

$$E[u(X)] \leq u(E[X]) \tag{A.6}$$

A.2. Anualidades Vitalicias

Para el cálculo de la prima neta única es necesario entender el concepto de anualidades vitalicias, una anualidad es un conjunto de pagos o depósitos generalmente iguales que se realizan en periodos regulares de tiempo, aunque su nombre lo indica no necesariamente son anuales mientras sean en periodos regulares de tiempo se les considera anualidades, en el caso de una anualidad vitalicia estos pagos se realizan durante toda la vida de un individuo.

Para el cálculo de anualidades vitalicias es necesario conocer la tasa de interés i con la que se acumularán los recursos financieros durante el transcurso de la anualidad y la probabilidad de que una persona de edad x sobreviva hasta la edad $x + n$ a esta probabilidad la denominaremos ${}_nP_x$.

Además supondremos que las personas viven hasta una edad w por lo que nuestro cálculo no será hasta el infinito sino hasta dicha edad.

A.2.1. Capital Diferido

Para facilitar la construcción de una anualidad diferida es conveniente introducir el concepto de *capital diferido*, este es es valor presente que tendrá un capital monetario pagadero dentro de n años considerando que quien realizará el pago tiene una edad x.

De la definición anterior se puede intuir que el capital diferido se conforma de el capital asegurado, un factor de descuento que depende de la tasa de interés y de la probabilidad de que una persona de edad x llegue con vida a la edad $x + n$, el capital diferido para el pago de \$1.00 en n años sería entonces:

$$ {}_nE_x = (1+i)^{-n}{}_nP_x \tag{A.7} $$

$$ {}_nE_x = v^n{}_nP_x \tag{A.8} $$

A.2.2. Cálculo de una anualidad vitalicia anticipada

Como su nombre lo indica una anualidad vitalicia anticipada es una anualidad vitalicia en la que el primer pago se realiza al inicio del periodo, si esta es para el pago de un servicio podemos decir que el primer pago es al contratar.

Se utiliza la notación $\ddot{a}_x$ para denotar a la anualidad vitalicia anticipada que debe pagar una persona de edad x debe pagar para que sea equivalente a sus pagos periódicos.

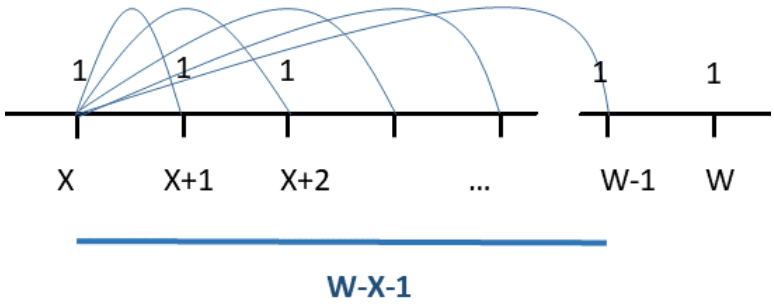


Figura A.1: Valor Presente para una anualidad Vitalicia Anticipada

Para un pago periódico de 1.00 el valor presente de cada pago sería la suma de los capitales diferidos correspondientes a cada pago el último pago se haría al inicio del año $w-x$ a partir del inicio es decir al completar $w-x-1$ años además al ser anticipada el primer pago se realiza al inicio es decir su valor presente es 1

$$ \ddot{a}_x = 1 + \sum_{t=1}^{t=w-x-1} {}_tE_x \tag{A.9} $$

Como ${}_0P_x = 1$ y $v^0 = 1$ podemos resumirlo en:

$$ \ddot{a}_x = \sum_{t=0}^{t=w-x-1} {}_tE_x \tag{A.10} $$

A.3. Log-Verosimilitud del modelo de regresión logística

Tenemos que la función de verosimilitud del modelo de regresión logística es:

$$\mathcal{I}(\beta_0, \beta_1, x_i, y_i) = \prod_{i=1}^{n} \pi(x)^{y_i}(1 - \pi(x))^{1-y_i} \tag{A.11}$$

El logaritmo de dicha función es su función de Log-verosimilitud:

$$\mathcal{L}(\beta_0, \beta_1, x_i, y_i) = log(\mathcal{I}(\beta_0, \beta_1, x_i, y_i)) \tag{A.12}$$

$$\begin{aligned}
&= \sum_{i=1}^{n} log(\pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}) \\
&= \sum_{i=1}^{n} log(\pi(x_i)^{y_i}) + \sum_{i=1}^{n} log((1 - \pi(x_i))^{1-y_i}) \\
&= \sum_{i=1}^{n} y_i log(\pi(x_i)) + \sum_{i=1}^{n} log((1 - \pi(x_i))) - \sum_{i=1}^{n} y_i log((1 - \pi(x_i))) \\
&= \sum_{i=1}^{n} y_i log(\frac{\pi(x_i)}{1 - \pi(x_i)}) + \sum_{i=1}^{n} log((1 - \pi(x_i)))
\end{aligned}$$

Por la ecuación 3.7

$$\mathcal{L}(\beta_0, \beta_1, x_i, y_i) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} log(1 + e^{\beta_0 + \beta_1 x_i}) \tag{A.13}$$

Apéndice B

Campos del RR-8

El reporte regulatorio 8 de Comisión Nacional de Seguros y Fianzas se compone de 47 campos estos campos, su descripción y requerimientos se encuentran detallados en el *ANEXO 38.1.9-b* de la Circular Única de Seguros y Fianzas, a continuación se presenta un extracto del mismo describiendo los campos.

1. Número de póliza: Se identificará a cada registro con el número de la póliza que la propia Institución le haya asignado, tanto el registro del titular del seguro como los correspondientes a cada uno de los dependientes o certificados. Dicho número deberá guardar consistencia con el archivo actual y futuro de emisión y siniestros.**Este campo al tener información del asegurado no es público**

2. Número de certificado: Se especificará el número de certificado en el caso de pólizas del seguro familiar; si el registro corresponde a una póliza no familiar, en este campo deberá capturarse el valor "metricconverterProductID1"1". El número de certificado no podrá repetirse dentro de una misma póliza, en caso de que la institución por cuestiones administrativas asigne el mismo número de certificado a dos o más integrantes de la póliza, deberá diferenciarlos de manera única considerando que los números de certificados asignados deberán ser consistentes con el archivo actual y futuro de emisión y siniestros.**Este campo al tener información del asegurado no es público**

3. Tipo de seguro: Se identificará a cada registro con alguna de las claves: **I** = Individual, **E** = Educativo y **F** = Familiar.

4. Modalidad de la póliza: Se identificará a cada registro la modalidad de cobertura contratada con alguno de los valores: **1** = Temporal, **2** = Vitalicio, **3** = VPL, **4** = Dotal, **5** = Rentas Diferidas o Pensiones Privadas, **6** = Saldado, **7** = Prorrogado y **8** = Otro.

5. Plan de la póliza: Se identificará a cada registro con las claves del catálogo 30.1, el plan que a la fecha de reporte tenga la póliza.

Tabla B.1: Catálogo 30.1 Planes para pólizas de vida individual

Clave	Póliza (plan)
1	Tradicional
2	Tradicional Inversión
3	Flexible, Vida Universal
4	Mancomunado
5	Otros

Fuente: CNSF, 2016 [9]

6. Moneda: Para cada registro se reportará la moneda en que fue emitida la póliza, **1**= Moneda Nacional **2**= Moneda Extranjera (Dólar) **3**= Moneda Indizada (UDI)

7. Entidad del contratante: Se especificará el estado de la República (o el extranjero) en donde radique el contratante especificado en la solicitud de la póliza, clasificado de acuerdo al catálogo 16.1.

8. Fecha de inicio de vigencia: Se reportará la fecha a partir de la cual la póliza del asegurado a quien corresponde el registro entró en vigor o la fecha en que entrará en vigor para el caso de pólizas de emisión adelantada. Para los casos de pólizas de seguro saldado o prorrogado se reportará la fecha correspondiente a la conversión.

9. Fecha de fin de vigencia: Se reportará la fecha en que finaliza la vigencia de la póliza. Para el caso de planes cuya fecha de fin de vigencia no sea conocida de antemano, este campo tendrá el valor "99991231". En los casos Saldado y Prorrogados, se deberá reportar la fecha de fin de vigencia que corresponda a la conversión. Para las pólizas emitidas del 1º de enero al 31 de diciembre cuya vigencia inicia en el ejercicio siguiente y la prima emitida se contabilizó en el ejercicio a reportar, también deberán formar parte del reporte.

10. Fecha de alta del certificado: Corresponde a la fecha en que se dio de alta, en la póliza el asegurado a quien corresponde el registro. Si el registro corresponde a una póliza individual, en este campo deberá reportarse la fecha de inicio de vigencia de la póliza, así como a todos aquellos certificados que se emitieron al mismo tiempo que la póliza.

11. Fecha de baja del certificado: Corresponde a la fecha en que se dio de baja de la póliza el certificado a quien corresponde el registro familiar o individual. Si el registro corresponde a un certificado que no se encuentra dado de baja, deberá reportarse vacío. Para el caso de cancelación, terminación o baja por muerte, la fecha que se reportará en este campo, será la que corresponda a la fecha de la cancelación, terminación u ocurrencia del fallecimiento que corresponda.

12. Fecha de nacimiento: Se especificará la fecha de nacimiento del asegurado a quien corresponda el registro, indicando el año, mes y día de su nacimiento.

13. Sexo: Se identificará el género del asegurado amparado por el certificado al que corresponda el registro, donde los valores permitidos son **F** = Femenino y **M** = Masculino.

14. Forma de venta: Se reportará mediante las claves del catálogo 1, el canal de distribución a través del cual se colocó o contrató el seguro de la póliza o certificado.

Tabla B.2: Catálogo 1 (extracto) Formas de venta de pólizas

Clave	Descripción
01	Agentes Persona Física
02	Agentes Persona Moral
05	Red de Sucursales Bancarias
06	Fuerza de Venta Interna o Casa Matriz
07	Módulos de Venta
08	Telemercadeo
10	Empresas Comerciales
11	Concesionarios Automotrices
12	Internet
13	Descuento por Nómina
14	Microcréditos
15	Otros Canales de Venta Masiva
99	Otra Forma de Venta

Fuente: CNSF, 2016 [9]

15. Estatus de la póliza: Se reportará mediante las claves del catálogo 22.1 y correspondiendo a las definiciones establecidas en la sección I, la situación en que se encuentra la póliza a la fecha de reporte y en el caso de pólizas "diferidas", tendrán el estatus de vigor. Entendiéndose como "diferidas", al hecho de que el inicio de la póliza sea posterior al periodo que se reporta, es decir, emisión anticipada.

En los casos en que la póliza sea rescatada y se convierta en Saldada o Prorrogada, el estatus de la póliza se reportará como Rescatada, y la Fecha de baja tendrá la fecha de conversión. Así mismo se procederá a reportar uno o más registros (uno por cada certi-

ficado que tenga la póliza que se convierta), y con estatus de Saldado o Prorrogado que corresponda, considerando los campos solicitados en la estadística y particularmente, en los campos Fecha de inicio de vigencia y Fecha de Alta del certificado se reportará la fecha de conversión.

Tabla B.3: Catálogo 22.1 (Extracto para vida) Estado de la póliza o certificado

Clave	Estatus de la póliza o certificado
1	Vigor
2	Expirada o terminada
3	Cancelada
4	Baja por muerte, invalidez o incapacidad
5	Rescatada
6	Saldada
7	Prorrogada

Fuente: CNSF, 2016 [9]

16. Estatus del certificado: Se reportará mediante las claves del catálogo 22.1 y correspondiendo a las definiciones establecidas en la sección I, la situación en que se encuentra el certificado a la fecha de reporte y en el caso de pólizas "diferidas", tendrán el estatus de vigor. Entendiéndose como "diferidas", al hecho de que el inicio de la póliza sea posterior al periodo que se reporta, es decir, emisión anticipada. Si el registro corresponde a una póliza individual, en este campo deberá reportarse el valor del campo "Estatus de Póliza".

17. Periodo de espera: Se registrará el número máximo de meses considerados como periodo de espera a partir de la fecha de la posible ocurrencia del siniestro, para los beneficios que utilicen dicho concepto. En caso de que no se haya emitido alguno de los beneficios que manejen periodo de espera, este campo se reportará en cero.

Cuando se presenten periodos de espera menor o igual a un mes, se reportarán como periodos de espera de un mes; en caso de que el número de días sea mayor a un mes deberán redondearse al mes que corresponda.

En el caso de que el registro considere más de un beneficio con periodos de espera diferentes, se procederá a registrar el número de meses que corresponda al mayor de los periodos de espera.

18. S.A. alcanzada beneficio 1: Se registrará el monto de suma asegurada alcanzada (sin decimales), de la póliza o certificado, para el beneficio 1 (**fallecimiento**). En caso de que no se haya contratado dicho beneficio, este campo se deberá reportar en cero.

19. S.A. alcanzada beneficio 2: Se registrará el monto de suma asegurada alcanzada (sin decimales), de la póliza o certificado, para el beneficio 2 (**pérdidas orgánicas**). En caso de que no se haya contratado dicho beneficio, este campo se deberá reportar en cero.

20. S.A. alcanzada beneficio 3: Se registrará el monto de suma asegurada alcanzada (sin decimales), de la póliza o certificado, para el beneficio 3 (**doble indemnización por muerte accidental**). En caso de que no se haya contratado dicho beneficio, este campo se deberá reportar en cero.

21. S.A. alcanzada beneficio 4: Se registrará el monto de suma asegurada alcanzada (sin decimales), de la póliza o certificado, para el beneficio 4 (**triple indemnización por muerte colectiva**). En caso de que no se haya contratado dicho beneficio, este campo se deberá reportar en cero.

22. S.A. alcanzada beneficio 6: Se registrará el monto de suma asegurada alcanzada (sin decimales), de la póliza o certificado, para el beneficio 6 (**pago adicional por invalidez o incapacidad, efectuada en una sola exhibición**). En caso de que no se haya contratado dicho beneficio, este campo se deberá reportar en cero.

23. S.A. alcanzada beneficio 8: Se registrará la suma de los montos de sumas aseguradas alcanzada (sin decimales), de la póliza o certificado, para el beneficio 8 (**otros beneficios**). En caso de que no se hayan contratado otros beneficios, este campo se deberá reportar en cero.

Ejemplos:

Concepto y suma asegurada: Llenado del campo:

Caso 1) Otros beneficios: 150,000 |150000|

Caso 2) Otros beneficios: 25,000 y 35,500 |60500|

24. S.A. alcanzada beneficio 9: Se registrará el monto de suma asegurada alcanzada (sin decimales), de la póliza o certificado, para el beneficio 9 (**sobrevivencia**) en una sola exhibición. En caso de que no se haya contratado dicho beneficio, este campo se deberá reportar en cero.

25. Prima emitida beneficio 1: Se registrará la prima emitida en el periodo de reporte (con 2 decimales) del beneficio 1 (**fallecimiento**). En caso de que no se haya contratado este beneficio, este campo se deberá reportar en cero.

26. Prima emitida beneficio 2: Se registrará la prima emitida en el periodo de reporte (con 2 decimales) del beneficio 2 (**pérdidas orgánicas**). En caso de que no se haya contratado este beneficio, este campo se deberá reportar en cero.

27. Prima emitida beneficio 3: Se registrará la prima emitida en el periodo de reporte (con 2 decimales) del beneficio 3 (**doble indemnización por muerte accidental**). En caso de que no se haya contratado este beneficio, este campo se deberá

reportar en cero.

28. Prima emitida beneficio 4: Se registrará la prima emitida en el periodo (con 2 decimales) del beneficio 4 (**triple indemnización por muerte colectiva**). En caso de que no se haya contratado este beneficio, este campo se deberá reportar en cero.

29. Prima emitida beneficio 5: Se registrará la prima emitida en el periodo (con 2 decimales) del beneficio 5 (**exención de pago de prima por invalidez, incapacidad o muerte**). En caso de que no se haya contratado este beneficio, este campo se deberá reportar en cero.

30. Prima emitida beneficio 6: Se registrará la prima emitida en el periodo (con 2 decimales) del beneficio 6 (**pago adicional por invalidez o incapacidad, efectuada en una sola exhibición**). En caso de que no se haya contratado este beneficio, este campo se deberá reportar en cero.

31. Prima emitida beneficio 7: Se registrará la prima emitida en el periodo (con 2 decimales) del beneficio 7 (**rentas diferidas**). En caso de que no se haya contratado este beneficio, este campo se deberá reportar en cero.

32. Prima emitida beneficio 8: Se registrará la suma de las primas emitidas en el periodo (con 2 decimales) del beneficio 8 (**otros beneficios**). En caso de que no se hayan contratado otros beneficios, este campo se deberá reportar en cero.

33. Prima emitida beneficio 9: Se registrará la prima emitida en el periodo (con 2 decimales) del beneficio 9 (**sobrevivencia**) en una sola exhibición. En caso de que no se haya contratado este beneficio, este campo se deberá reportar en cero.

34. Saldo del fondo en administración: Para aquellas pólizas que tengan asociado un fondo en administración, se reportará el saldo (con dos decimales) al cierre del ejercicio de que se trate. En caso de que la póliza a reportar no se encuentre asociada a un fondo en administración, este campo se reportará en cero. El monto correspondiente al saldo del fondo en administración deberá reportarse a prorrata en cada certificado de la póliza (en el caso de pólizas con certificados).

35. Monto de vencimiento: Se reportará el importe total (con dos decimales) del valor del vencimiento al cierre del ejercicio de que se trate. En el caso de las pólizas que no cuenten con este concepto, este campo se reportará en cero. El monto correspondiente al vencimiento deberá reportarse a prorrata en cada certificado de la póliza (en el caso de pólizas con certificados).

36. Monto de rescate: Se reportará el importe total (con dos decimales) del valor de rescate ocurrido al cierre del ejercicio de que se trate. En el caso de las pólizas que no cuenten con este concepto o no exista dicho monto, este campo se reportará en cero. El monto correspondiente al rescate deberá reportarse a prorrata en cada certificado de la póliza (en el caso de pólizas con certificados).

37. Monto de dividendo: Se reportará el monto correspondiente al incremento neto

(con dos decimales) que registre por concepto de participación en las utilidades obtenidas por la Institución, ya sea en la operación global de la cartera a la que pertenece dicha póliza, o bien con base en experiencia propia de cada contrato al cierre del ejercicio de que se trate. En caso de que no exista este concepto, el campo deberá reportarse en cero. El monto correspondiente al dividendo deberá reportarse a prorrata en cada certificado de la póliza (en el caso de pólizas con certificados).

38. Subtipo de seguro: Se reportará mediante la clave del catálogo 83.

Tabla B.4: Catálogo 83 (Extracto para Vida) Subtipo de Seguro

Clave	Subtipo de Seguro
1	Microseguros, pólizas que corresponden a productos registrados de acuerdo con el Capítulo 5.1 vigente.
2	Microseguros, pólizas que no corresponden a productos registrados como tal sino respecto de los límites de suma asegurada y que tengan el propósito de promover el acceso de la población de bajos ingresos a la protección del seguro mediante la utilización de medios de distribución y operación de bajo costo, es decir, que tengan un tratamiento como tal.
3	Producto Básico estandarizado, pólizas que corresponden a productos registrados de acuerdo con el Capítulo 5.4 vigente.
4	Otros

Fuente: CNSF, 2016 [9]

39. Emisión: Se reportará mediante los siguientes valores: **0** = Primer año, **1** = Renovación y **2** = Prima única, que corresponda a la póliza o certificado.

40. Año póliza: Se identificará a cada registro con el número de años de antigüedad de la póliza, considerándose como año póliza al periodo que comprende su aniversario. En caso de que la póliza a la fecha de reporte, se encuentre antes de cumplir su primer año póliza, deberá ser considerada con antigüedad igual a 1.

Ejemplos:

Aniversarios Llenado del campo

De cero hasta su 1er. Aniversario = |1|

1er. Aniversario cumplido + 1 día = |2|

2do. Aniversario cumplido + 1 día = |3|

41. Inicio de cobertura: Se especificará para cada registro de la póliza, la forma en que inicia su cobertura, donde los valores permitidos son **1** = Diferida y **2** = No Diferida. Entendiéndose como "diferido", al hecho de que el inicio de la cobertura de la póliza sea posterior al periodo que se reporta, es decir, emisión anticipada. Cuando una póliza diferida ya se encuentre en el periodo de vigencia, el valor para este campo deberá ser "2", es decir No diferida. Si el registro en este campo tiene el valor 1 y la póliza que le da origen se emite por primera vez (1er. año), entonces se reportará 1 en el campo Año póliza.

42. S.A. alcanzada de dotales a corto plazo: Se registrará el acumulado de los montos de sumas aseguradas alcanzadas (sin decimales), de la póliza o certificado, que presente dotales a corto plazo. En caso de que no se cuente con dotales a corto plazo, este campo se deberá reportar en cero.

43. Prima emitida de dotales a corto plazo: Se registrará la suma de las primas emitidas en el periodo (con 2 decimales). En caso de que no se cuente con dotales a corto plazo, este campo se deberá reportar en cero.

44. Plazo de pago de primas: Se reportará el número de años durante los cuales la Institución recibirá primas del contratante o asegurado. En caso de que la vigencia de la póliza sea menor o igual a un año, se reportará el valor 1.

45. Tipo de riesgo asociado: Se registrará de acuerdo con los valores **1** = Riesgo financiero a cargo de la Institución, **2** = Riesgo no financiero (mortalidad, morbilidad o sobrevivencia) y **3** = Otros (por ejemplo desempleo), que ampara la póliza o certificado.

46. Saldo del fondo de Inversión: Para aquellas pólizas que tengan asociado un fondo de inversión, se reportará el saldo (con dos decimales) al cierre del ejercicio de que se trate. En caso de que la póliza a reportar no se encuentre asociada a un fondo de inversión, este campo se reportará en cero. El monto correspondiente al saldo del fondo en inversión deberá reportarse a prorrata en cada certificado de la póliza (en el caso de pólizas con certificados). Este fondo se reportará en cero cuando no exista dicho concepto o bien cuando forme parte del fondo en administración, por lo que se reportará el o los saldos de los fondos de inversión (si existen) y que no formen parte del fondo en administración.

47. Extraprima médica básica: Se reportará para cada registro el grado de sub-normalidad que corresponda para la cobertura de fallecimiento por concepto médico, como múltiplo de la prima de dicha cobertura. Ejemplos: a) Para una sub-normalidad del 100 %, se reportará el valor 2.0; b) Sub-normalidad del 150 %, se deberá capturar el valor 2.5; c) Sub-normalidad del 75 %, se reportará el valor 1.75. En el caso de que no aplique dicho concepto, se reportará el valor 1.

Apéndice C

Tablas y gráficas para la selección de variables

C.1. IV para todas las variables

El valor del IV para cada variable y su capacidad predictiva bajo el criterio expuesto en la sección 5.4.2 es el siguiente:

Tabla C.1: IV y predictividad para cada variable analizada

Variable	IV	Predictibilidad
STATUS_CERT	11.660253	Valores atípicos
STATUS_POL	9.257941	
SA_BEN1	5.406985	
SDO_FADMON	1.443313	
SA_BEN6	1.374173	
SA_BEN4	1.237831	
SA_BEN8	1.013358	
SA_BEN3	0.630533	
ANIO_POLIZA	0.580554	
Plan_ 1	0.459795	Alta predictibilidad
Mod_ 1	0.323410	
Plan_ 3	0.304244	
Mod_ 2	0.279367	Predictibilidad media
PER_ESPERA	0.227832	
Plan_ 4	0.198939	
EDAD	0.197063	

Fuente: Elaboración Propia.

Tabla C.1: IV y predictividad para cada variable analizada

Variable	IV	Predictibilidad
EMISION	0.135063	
SA_BEN2	0.070417	Poca predictibilidad
Plan_ 2	0.047796	
Entidad_ 9	0.036997	
DIVIDENDO	0.032028	
INI_COBER	0.012513	Sin predictibilidad
Entidad_ 30	0.012025	
Entidad_ 15	0.011216	
MONEDA	0.010978	
SA_DCP	0.007308	
Entidad_ 14	0.005541	
Entidad_ 19	0.005327	
PMA_EMIDCP	0.002759	
Entidad_ 21	0.002566	
Entidad_ 34	0.002401	
Plan_ 5	0.001950	
VENCIMIENTO	0.001710	
Entidad_ 16	0.001632	
Sexo_ F	0.001544	
Sexo_ M	0.001544	
Entidad_ 2	0.001395	
Entidad_ 17	0.001085	
Entidad_ 27	0.001062	
RESCATE	0.000930	
Entidad_ 7	0.000624	
Entidad_ 8	0.000609	
Entidad_ 29	0.000429	
Entidad_ 12	0.000389	
Entidad_ 23	0.000378	
Entidad_ 28	0.000321	
Entidad_ 25	0.000289	
Entidad_ 20	0.000277	
Entidad_ 22	0.000236	
Entidad_ 32	0.000158	
Entidad_ 4	0.000128	
Entidad_ 11	0.000105	
Entidad_ 31	0.000092	
Entidad_ 24	0.000069	
Entidad_ 33	0.000059	
Entidad_ 1	0.000058	

Fuente: Elaboración Propia.

Tabla C.1: IV y predictividad para cada variable analizada

Variable	IV	Predictibilidad
Entidad_ 26	0.000049	
Entidad_ 18	0.000032	
Entidad_ 10	0.000030	
Entidad_ 6	0.000026	
Entidad_ 5	0.000008	
Entidad_ 13	0.000007	
Entidad_ 3	0.000002	

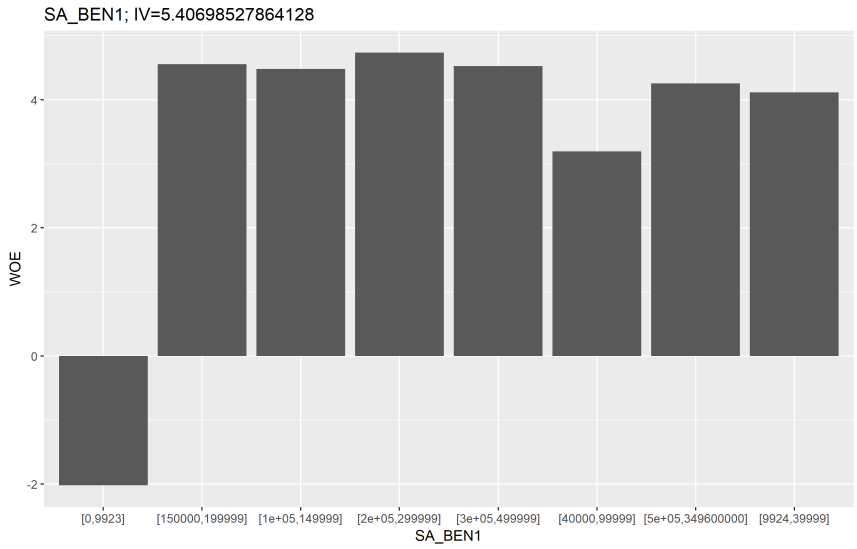
Fuente: Elaboración Propia.

C.2. Comportamiento del WOE

El comportamiento del *WOE* para cada variable analizada (Solo aquellas con predictibilidad) es el siguiente:

- **STATUS_CERT y STATUS_POL** Tienen relación directa con la variable dependiente ya que indican si el certificado o póliza están activos o no. Por lo anterior, no deben considerarse para el modelo
- **SA_BEN1 (muerte)** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.1: Comportamiento del WOE para la variable Suma asegurada Beneficio 1

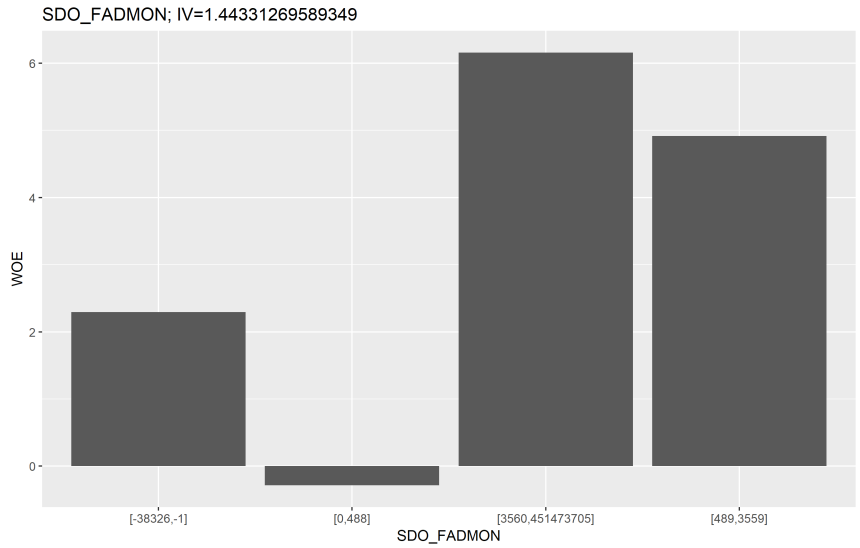


Fuente:Elaboración Propia.

Cumple con las características deseables de una variable, y además dentro del contexto del seguro de vida y la toma de decisiones del asegurado, tiene sentido ya que a mayor recompensa por mantener la póliza se reduce el incentivo para cancelar dicha póliza

- **SDO_FADMON** El comportamiento del WOE para la variable es el siguiente

Figura C.2: Comportamiento del WOE para la variable Saldo del fondo de administración

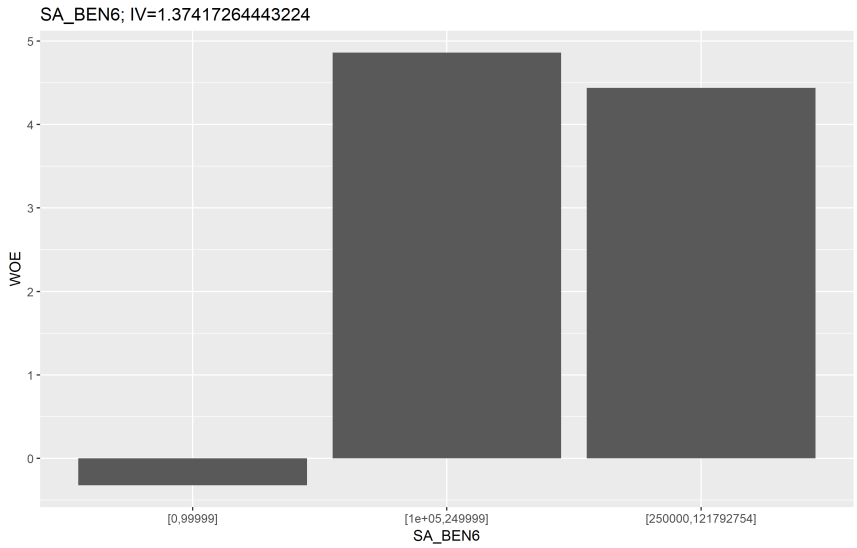


Fuente:Elaboración Propia.

Se observa que en la segunda columna se tiene un valor negativo, mientras que en las otras son valores positivos por lo que bajo este criterio no es una variable predictiva.

- **SA_BEN6 (Incapacidad)** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.3: Comportamiento del WOE para la variable Suma asegurada Beneficio 6

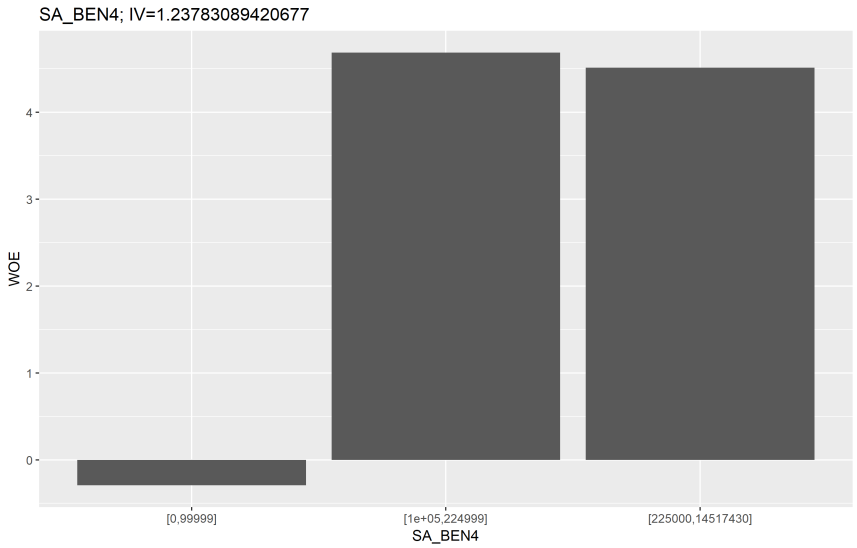


Fuente:Elaboración Propia.

Cumple con las características deseables de una variable, y además dentro del contexto del seguro de vida y la toma de decisiones del asegurado, tiene sentido debido a que a mayores beneficios por mantener la póliza se reduce el incentivo para cancelar dicha póliza

- **SA_BEN4 (Muerte colectiva)** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.4: Comportamiento del WOE para la variable Suma asegurada Beneficio 4

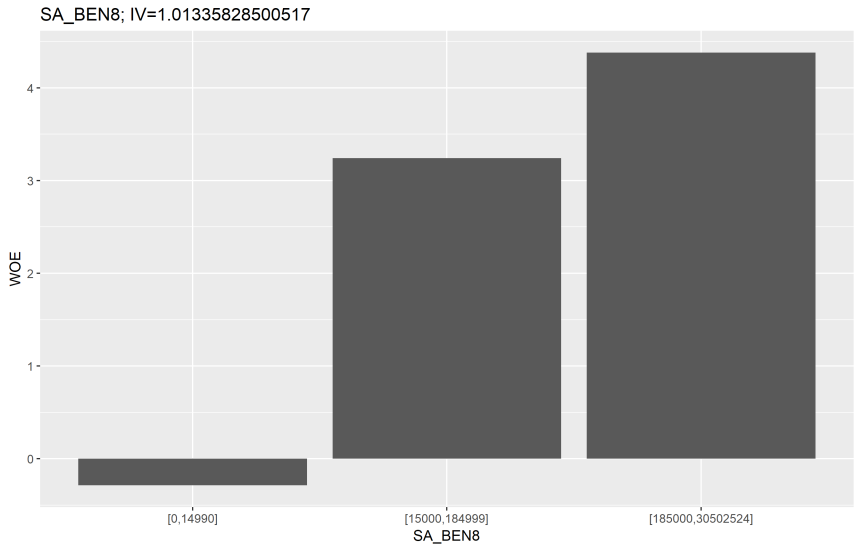


Fuente:Elaboración Propia.

Cumple con las características deseables de una variable, y además dentro del contexto del seguro de vida y la toma de decisiones del asegurado, tiene sentido ya que a mayores beneficios por mantener la póliza se reduce el incentivo para cancelar dicha póliza

- **SA_BEN8 (Otros beneficios)** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.5: Comportamiento del WOE para la variable Suma asegurada Beneficio 8

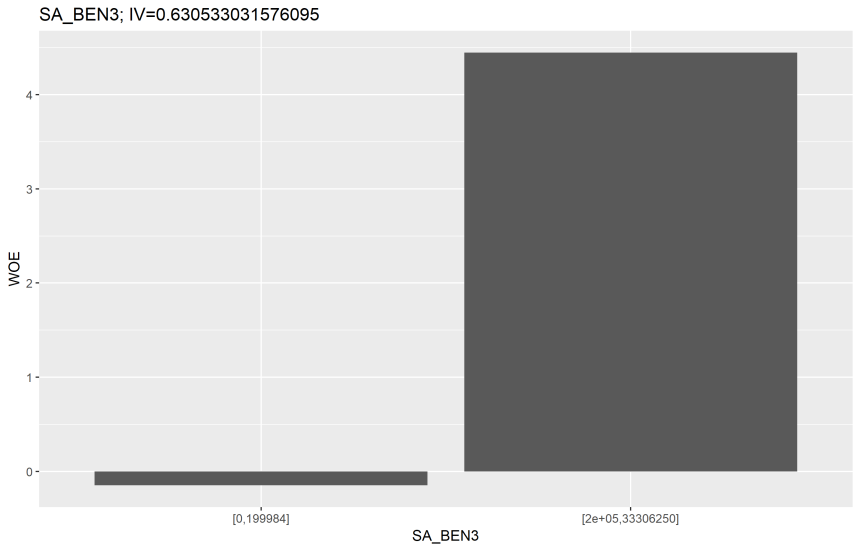


Fuente:Elaboración Propia.

Cumple con las características deseables de una variable, y además dentro del contexto del seguro de vida y la toma de decisiones del asegurado, tiene sentido ya que a mayores beneficios por mantener la póliza se reduce el incentivo para cancelar dicha póliza

- **SA_BEN3 (Muerte accidental)** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.6: Comportamiento del WOE para la variable Suma asegurada Beneficio 3

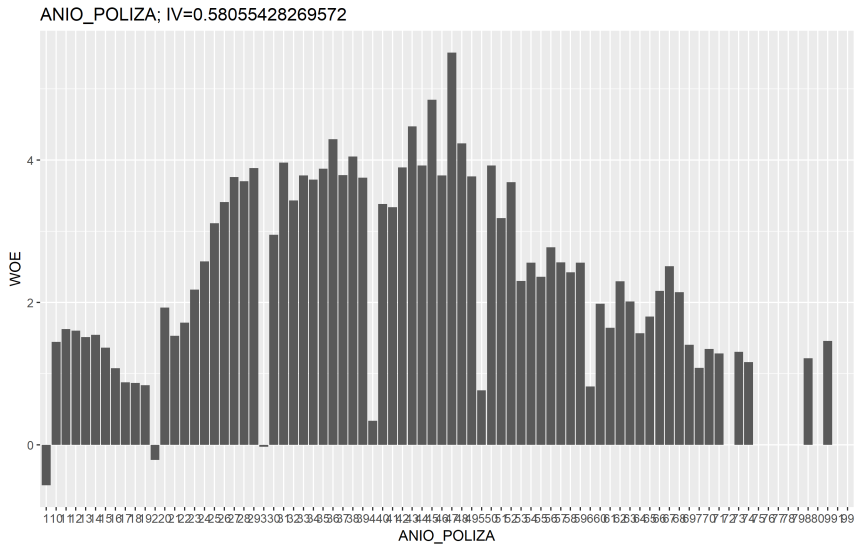


Fuente:Elaboración Propia.

Cumple con las características deseables de una variable, y además dentro del contexto del seguro de vida y la toma de decisiones del asegurado, tiene sentido ya que a mayores beneficios por mantener la póliza se reduce el incentivo para cancelar dicha póliza

- **ANIO_POLIZA** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.7: Comportamiento del WOE para la variable Año Pòliza

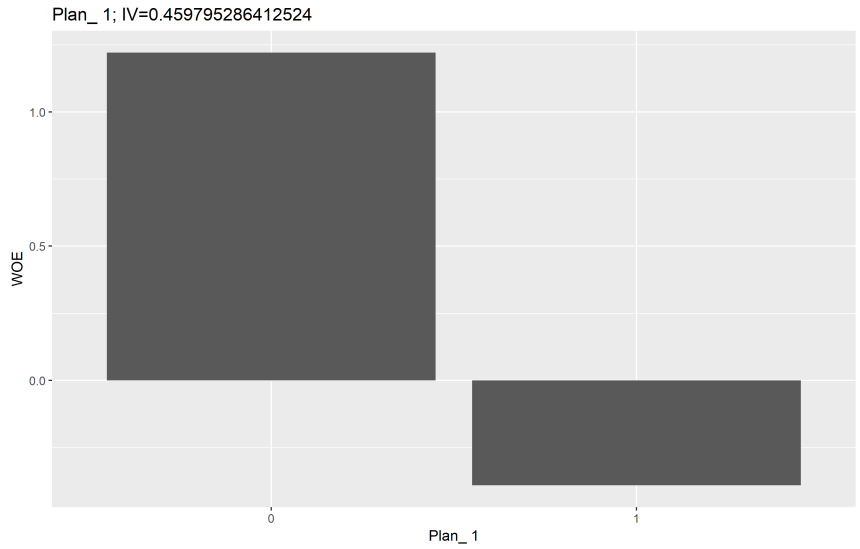


Fuente:Elaboración Propia.

El comportamiento del WOE correspondiente no permite aceptar a la variable como predictora debido a que tiene muchos cambios de signo.

- **PLAN_1 (Póliza tradicional)** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.8: Comportamiento del WOE para la variable póliza tradicional

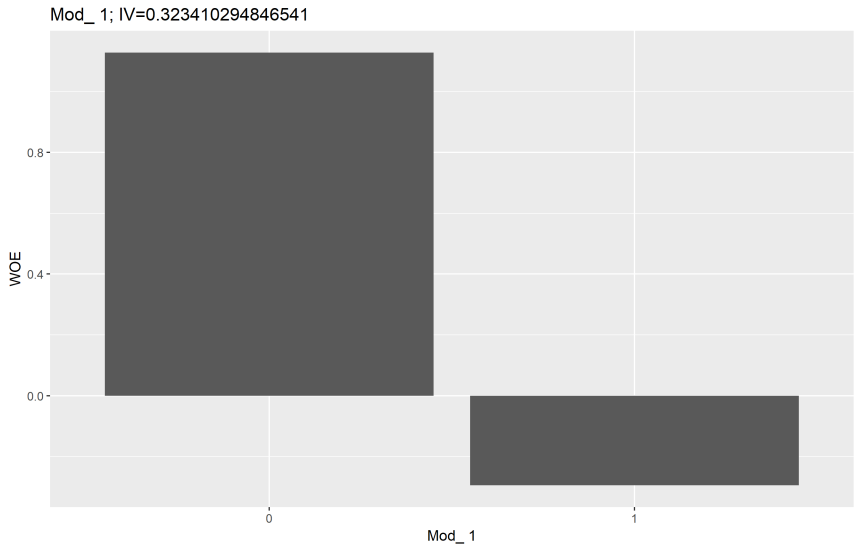


Fuente:Elaboración Propia.

El comportamiento del WOE permite aceptar a la variable como una variable predictora

- **Mod_1 (Póliza temporal)** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.9: Comportamiento del WOE para la variable póliza tradicional

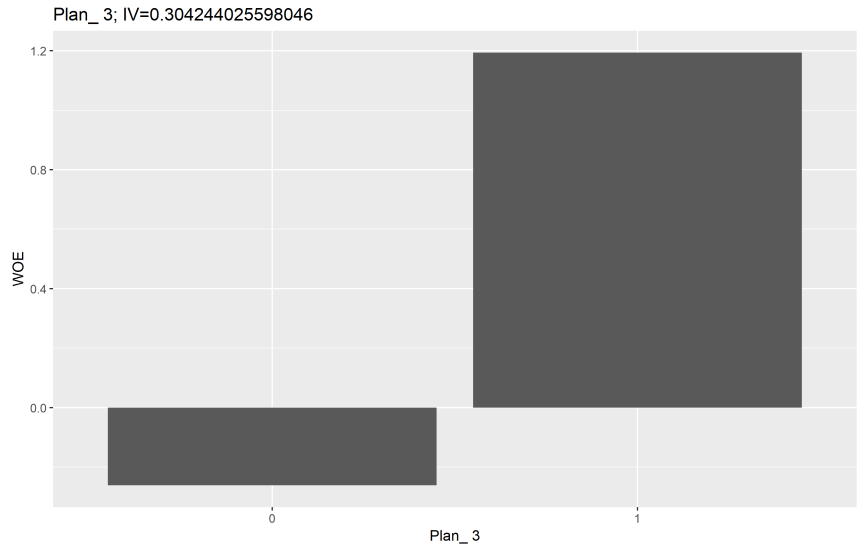


Fuente:Elaboración Propia.

El comportamiento del WOE permite aceptar a la variable como una variable predictora

- **Plan_3 (Póliza flexible)** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.10: Comportamiento del WOE para la variable póliza tradicional

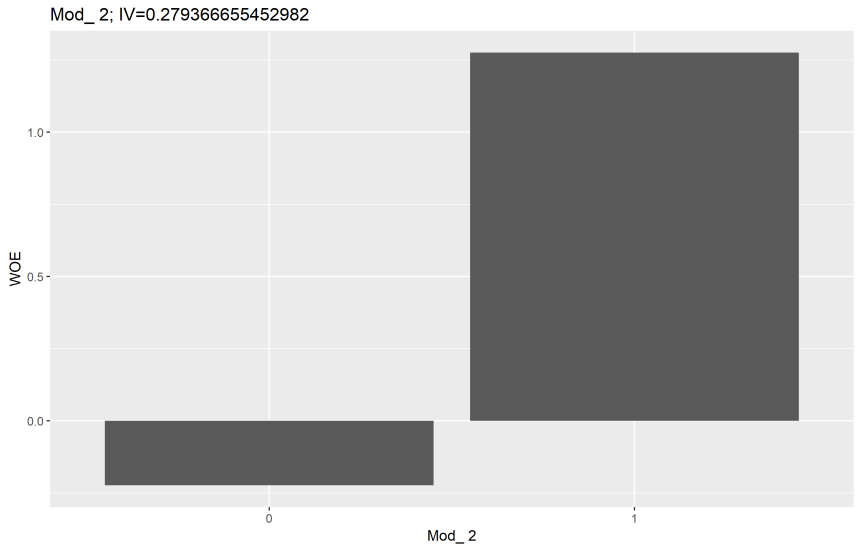


Fuente:Elaboración Propia.

El comportamiento del WOE permite aceptar a la variable como una variable predictora

- **Mod_2 (Póliza vitalicia)** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.11: Comportamiento del WOE para la variable póliza vitalicia

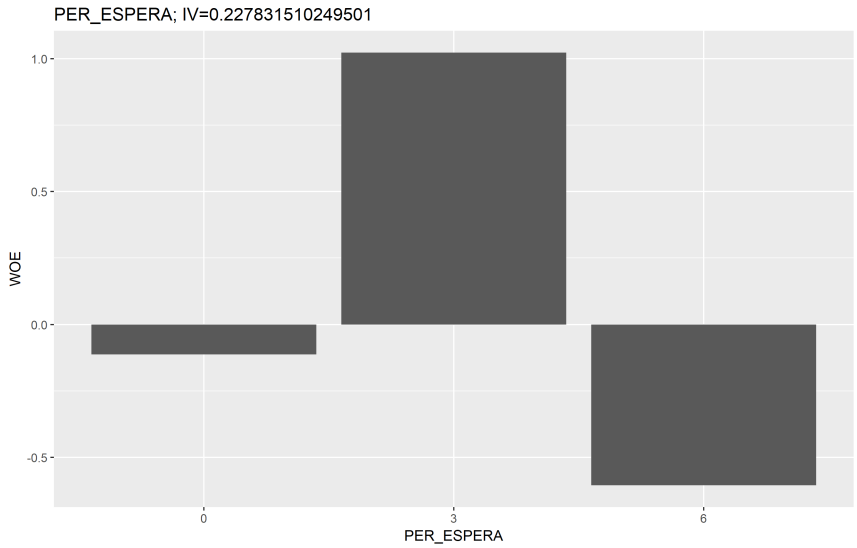


Fuente:Elaboración Propia.

El comportamiento del WOE permite aceptar a la variable como una variable predictora

- **PER_ESPERA** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.12: Comportamiento del WOE para la variable periodo de espera

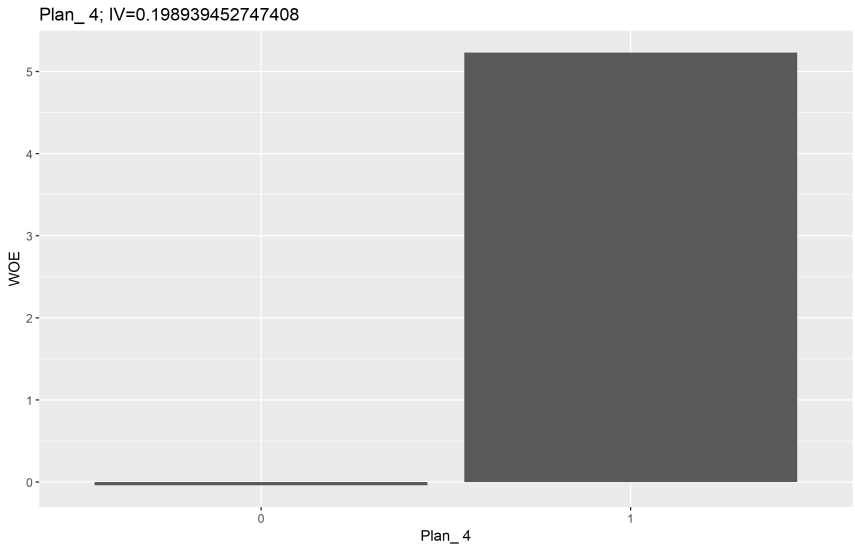


Fuente:Elaboración Propia.

El comportamiento del WOE permite rechaza a la variable como una variable predictora

- **Plan_4 (Seguro mancomunado)** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.13: Comportamiento del WOE para la variable periodo de espera

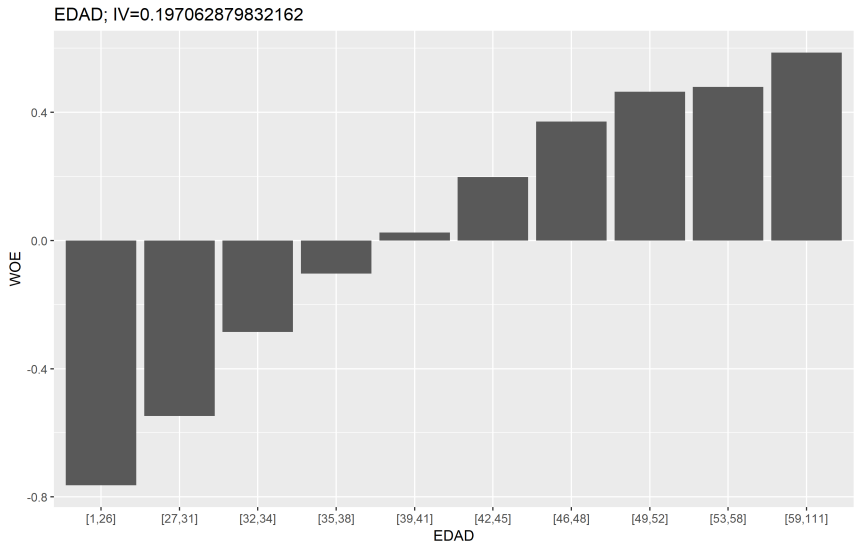


Fuente:Elaboración Propia.

El comportamiento del WOE permite rechaza a la variable como una variable predictora

- **EDAD** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.14: Comportamiento del WOE para la variable Edad

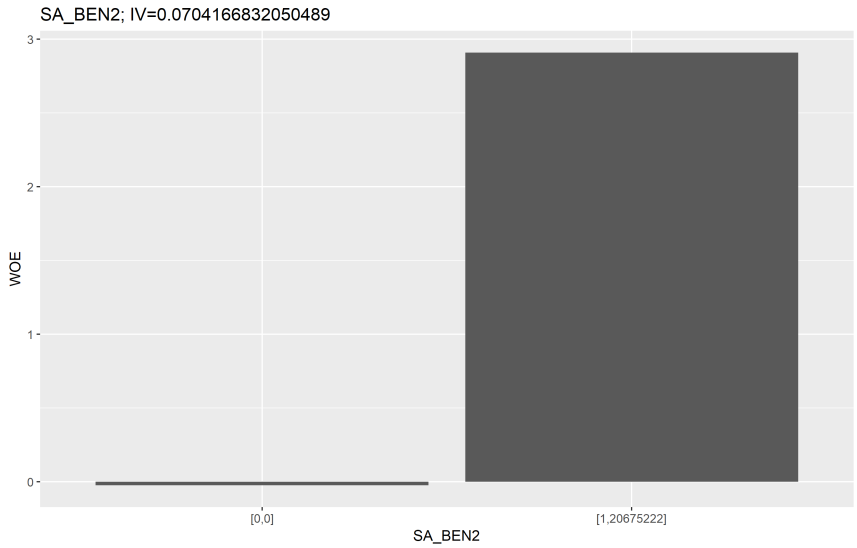


Fuente:Elaboración Propia.

El comportamiento del WOE permite aceptar a la variable como una variable predictora

- **SA_BEN2** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.15: Comportamiento del WOE para la variable periodo de espera

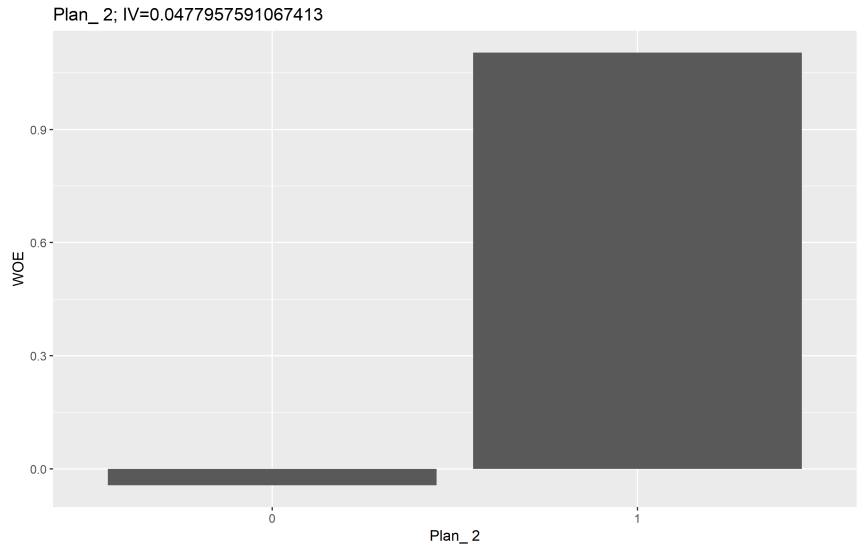


Fuente:Elaboración Propia.

El comportamiento del WOE permite rechazar a la variable como una variable predictora debido a que el valor para una categoría es muy cercano a cero.

- **Plan_2** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.16: Comportamiento del WOE para la variable periodo de espera

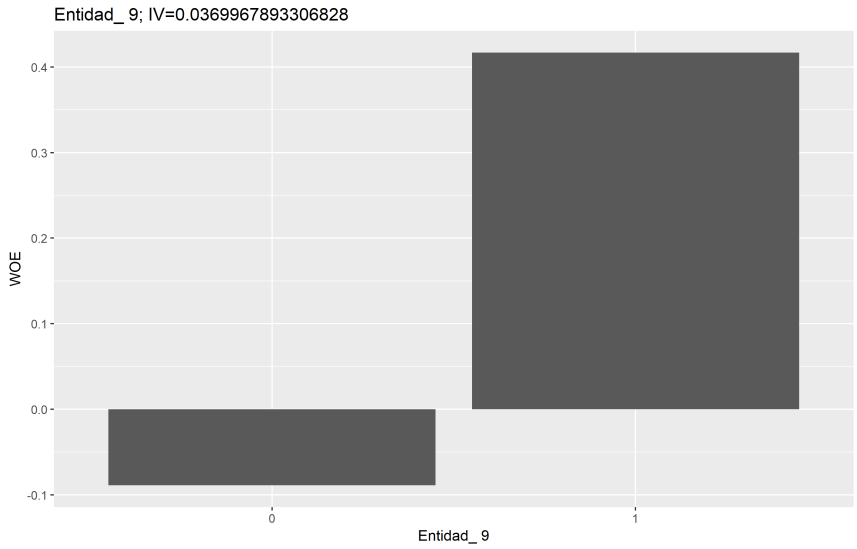


Fuente:Elaboración Propia.

El comportamiento del WOE permite aceptar a la variable como una variable predictora

- **Entidad_9 (CDMX)** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.17: Comportamiento del WOE para la variable entidad 9 (CDMX)

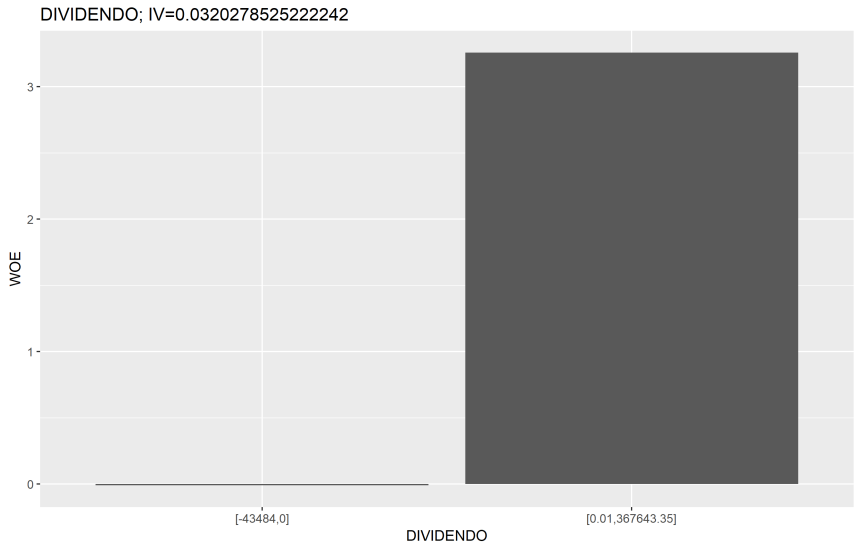


Fuente:Elaboración Propia.

El comportamiento del WOE permite aceptar a la variable como una variable predictora.

- **DIVIDENDO** Tiene el siguiente comportamiento respecto al valor del WOE

Figura C.18: Comportamiento del WOE para la variable dividendo



Fuente:Elaboración Propia.

El comportamiento del WOE permite rechaza a la variable como una variable predictora debido a que en una de sus categorías tiene un valor muy cercano a cero.

C.3. Estadístico KS

El valor del Estadístico KS para cada variable y su capacidad predictiva bajo el criterio expuesto en la sección 5.3 es el siguiente:

Tabla C.2: Estadístico KS y predictividad para cada variable analizada

Variable	KS	Predictibilidad
SA_BEN1	0.936504	Valor atípico
SA_BEN6	0.322636	Variable predictiva
SA_BEN4	0.290135	
ANIO_POLIZA	0.286784	
Plan_ 1	0.285143	
SA_BEN8	0.268722	
SA_BEN9	0.268722	
SA_BEN3	0.237062	
Mod_ 1	0.227360	
Plan_ 3	0.209063	
EDAD	0.191297	
Mod_ 2	0.186212	
EMISION	0.181732	
PER_ESPERA	0.086358	Variable no predictiva
Entidad_ 9	0.073119	
Plan_ 2	0.041654	
Plan_ 4	0.037736	
Entidad_ 15	0.033798	
Entidad_ 30	0.026217	
SA_BEN2	0.023995	
Sexo_ F	0.019587	
Sexo_ M	0.019587	
Entidad_ 14	0.018564	
Entidad_ 19	0.016753	
Mod_ 7	0.014986	
Mod_ 6	0.013722	
Mod_ 3	0.010212	
Entidad_ 21	0.009132	
Entidad_ 16	0.006805	

Fuente: Elaboración Propia.

Tabla C.2: Estadístico KS y predictividad para cada variable analizada

Variable	KS	Predictibilidad
Entidad_ 2	0.006347	
Mod_ 4	0.005349	
Entidad_ 8	0.004194	
Entidad_ 27	0.004077	
Entidad_ 7	0.003939	
Entidad_ 17	0.003847	
Plan_ 5	0.003310	
Entidad_ 28	0.003300	
Mod_ 8	0.003201	
Entidad_ 25	0.002775	
Entidad_ 12	0.002728	
INI_COBER	0.002631	
Entidad_ 20	0.002343	
Entidad_ 23	0.002207	
Entidad_ 11	0.001907	
Entidad_ 22	0.001821	
Entidad_ 29	0.001680	
Entidad_ 32	0.001282	
Entidad_ 34	0.001204	
Entidad_ 26	0.001156	
Entidad_ 24	0.001148	Variable no predictiva
Entidad_ 31	0.001118	
Entidad_ 4	0.001031	
SDO_FADMON	0.001017	
Entidad_ 1	0.000796	
Entidad_ 10	0.000625	
Entidad_ 18	0.000568	
SA_DCP	0.000560	
DIVIDENDO	0.000538	
PMA_EMIDCP	0.000490	
PZO_PGO_PMA	0.000490	
TIP_RGO_ASOC	0.000490	
SAL_FON_INV	0.000490	
EX_PMA_MBAS	0.000490	
Entidad_ 5	0.000448	
Entidad_ 6	0.000395	
Entidad_ 13	0.000332	
VENCIMIENTO	0.000138	
Entidad_ 3	0.000106	
Mod_ 5	0.000080	

Fuente: Elaboración Propia.

Tabla C.2: Estadístico KS y predictividad para cada variable analizada

Variable	KS	Predictibilidad
Entidad_ 33	0.000040	
RESCATE	0.000025	

Fuente: Elaboración Propia.

Apéndice D

Errores y estadísticos para distintos valores de p

En este apéndice se muestran los errores y estadísticos asociados para diferentes valores de p, estos valores se utilizaron para definir el criterio de clasificación y para construir la curva ROC asociada al modelo, para la construcción de esta tabla el valor de p fue incrementando en 1 % para construir los errores y estadísticos asociados

D.1. Muestra de entrenamiento

Tabla D.1: Errores y estadísticos asociados a la matriz de confusión para distintos valores de p en la muestra de entrenamiento

p	VN	FP	FN	VP	sens	esp	PV_pos	PV_neg	suma
0.55	1311565	429940	127413	5410546	0.98	0.75	0.93	0.91	3.57
0.54	1300641	415074	138337	5425412	0.98	0.76	0.93	0.90	3.57
0.56	1319209	448764	119769	5391722	0.98	0.75	0.92	0.92	3.56
0.57	1328360	466999	110618	5373487	0.98	0.74	0.92	0.92	3.56
0.53	1288408	400328	150570	5440158	0.97	0.76	0.93	0.90	3.56
0.58	1334692	483193	104286	5357293	0.98	0.73	0.92	0.93	3.56
0.52	1277421	388381	161557	5452105	0.97	0.77	0.93	0.89	3.56
0.59	1342037	502871	96941	5337615	0.98	0.73	0.91	0.93	3.56
0.51	1263636	376709	175342	5463777	0.97	0.77	0.94	0.88	3.55
0.60	1347785	523070	91193	5317416	0.98	0.72	0.91	0.94	3.55
0.50	1244192	362296	194786	5478190	0.97	0.77	0.94	0.86	3.54

Fuente: Elaboración Propia.

Tabla D.1: Errores y estadísticos asociados a la matriz de confusión para distintos valores de p en la muestra de entrenamiento

p	VN	FP	FN	VP	sens	esp	PV_pos	PV_neg	suma
0.61	1352910	550886	86068	5289600	0.98	0.71	0.91	0.94	3.54
0.49	1229249	349937	209729	5490549	0.96	0.78	0.94	0.85	3.54
0.62	1358001	573866	80977	5266620	0.98	0.70	0.90	0.94	3.53
0.48	1212489	338459	226489	5502027	0.96	0.78	0.94	0.84	3.53
0.63	1361750	595849	77228	5244637	0.99	0.70	0.90	0.95	3.53
0.64	1367313	619948	71665	5220538	0.99	0.69	0.89	0.95	3.52
0.47	1191914	324135	247064	5516351	0.96	0.79	0.94	0.83	3.52
0.46	1175296	311593	263682	5528893	0.95	0.79	0.95	0.82	3.51
0.65	1370803	650535	68175	5189951	0.99	0.68	0.89	0.95	3.51
0.66	1373856	672846	65122	5167640	0.99	0.67	0.88	0.95	3.50
0.45	1153170	297954	285808	5542532	0.95	0.79	0.95	0.80	3.50
0.67	1377261	695243	61717	5145243	0.99	0.66	0.88	0.96	3.49
0.44	1125132	277069	313846	5563417	0.95	0.80	0.95	0.78	3.48
0.68	1380198	720558	58780	5119928	0.99	0.66	0.88	0.96	3.48
0.43	1103860	262610	335118	5577876	0.94	0.81	0.96	0.77	3.47
0.69	1382361	748767	56617	5091719	0.99	0.65	0.87	0.96	3.47
0.70	1384773	772712	54205	5067774	0.99	0.64	0.87	0.96	3.46
0.42	1078711	249704	360267	5590782	0.94	0.81	0.96	0.75	3.46
0.71	1386788	802465	52190	5038021	0.99	0.63	0.86	0.96	3.45
0.41	1052887	236347	386091	5604139	0.94	0.82	0.96	0.73	3.44
0.72	1388365	824810	50613	5015676	0.99	0.63	0.86	0.96	3.44
0.73	1390445	848339	48533	4992147	0.99	0.62	0.85	0.97	3.43
0.40	1022039	224188	416939	5616298	0.93	0.82	0.96	0.71	3.42
0.74	1391711	874897	47267	4965589	0.99	0.61	0.85	0.97	3.42
0.75	1393403	897546	45575	4942940	0.99	0.61	0.85	0.97	3.41
0.76	1394590	922157	44388	4918329	0.99	0.60	0.84	0.97	3.40
0.39	983905	207861	455073	5632625	0.93	0.83	0.96	0.68	3.40
0.77	1396040	947442	42938	4893044	0.99	0.60	0.84	0.97	3.39
0.78	1397183	970186	41795	4870300	0.99	0.59	0.83	0.97	3.39
0.38	956682	196547	482296	5643939	0.92	0.83	0.97	0.66	3.38
0.79	1398400	992518	40578	4847968	0.99	0.58	0.83	0.97	3.38
0.80	1399705	1013870	39273	4826616	0.99	0.58	0.83	0.97	3.37
0.81	1400963	1032444	38015	4808042	0.99	0.58	0.82	0.97	3.36
0.37	924118	184406	514860	5656080	0.92	0.83	0.97	0.64	3.36
0.82	1402188	1052052	36790	4788434	0.99	0.57	0.82	0.97	3.36
0.83	1403436	1072496	35542	4767990	0.99	0.57	0.82	0.98	3.35
0.84	1404554	1094007	34424	4746479	0.99	0.56	0.81	0.98	3.34

Fuente: Elaboración Propia.

Tabla D.1: Errores y estadísticos asociados a la matriz de confusión para distintos valores de p en la muestra de entrenamiento

p	VN	FP	FN	VP	sens	esp	PV_pos	PV_neg	suma
0.36	894748	173112	544230	5667374	0.91	0.84	0.97	0.62	3.34
0.85	1405650	1116483	33328	4724003	0.99	0.56	0.81	0.98	3.34
0.86	1406593	1141936	32385	4698550	0.99	0.55	0.80	0.98	3.33
0.87	1407770	1168354	31208	4672132	0.99	0.55	0.80	0.98	3.32
0.88	1408890	1199710	30088	4640776	0.99	0.54	0.79	0.98	3.31
0.35	835189	155221	603789	5685265	0.90	0.84	0.97	0.58	3.30
0.89	1409852	1229418	29126	4611068	0.99	0.53	0.79	0.98	3.30
0.90	1411110	1260071	27868	4580415	0.99	0.53	0.78	0.98	3.29
0.34	800065	144544	638913	5695942	0.90	0.85	0.98	0.56	3.28
0.91	1412356	1293491	26622	4546995	0.99	0.52	0.78	0.98	3.28
0.92	1413687	1331424	25291	4509062	0.99	0.51	0.77	0.98	3.26
0.93	1415457	1370814	23521	4469672	0.99	0.51	0.77	0.98	3.25
0.33	762462	134076	676516	5706410	0.89	0.85	0.98	0.53	3.25
0.94	1417182	1419588	21796	4420898	1.00	0.50	0.76	0.98	3.24
0.95	1419159	1481667	19819	4358819	1.00	0.49	0.75	0.99	3.22
0.32	698305	119200	740673	5721286	0.89	0.85	0.98	0.49	3.20
0.96	1421634	1573421	17344	4267065	1.00	0.47	0.73	0.99	3.19
0.31	661103	108675	777875	5731811	0.88	0.86	0.98	0.46	3.18
0.97	1424012	1682970	14966	4157516	1.00	0.46	0.71	0.99	3.16
0.30	622468	99549	816510	5740937	0.88	0.86	0.98	0.43	3.15
0.98	1426083	1806108	12895	4034378	1.00	0.44	0.69	0.99	3.12
0.29	550761	85386	888217	5755100	0.87	0.87	0.99	0.38	3.10
0.99	1427244	1934174	11734	3906312	1.00	0.42	0.67	0.99	3.08
0.28	510148	76701	928830	5763785	0.86	0.87	0.99	0.35	3.07
0.27	437609	64108	1001369	5776378	0.85	0.87	0.99	0.30	3.02
0.26	399924	57230	1039054	5783256	0.85	0.87	0.99	0.28	2.99
0.25	333821	47814	1105157	5792672	0.84	0.87	0.99	0.23	2.94
0.24	302355	42813	1136623	5797673	0.84	0.88	0.99	0.21	2.91
0.23	244774	32179	1194204	5808307	0.83	0.88	0.99	0.17	2.88
0.22	195623	25426	1243355	5815060	0.82	0.88	1.00	0.14	2.84
0.21	167912	22228	1271066	5818258	0.82	0.88	1.00	0.12	2.82
0.20	120956	16516	1318022	5823970	0.82	0.88	1.00	0.08	2.78
0.19	78474	10903	1360504	5829583	0.81	0.88	1.00	0.05	2.74
0.18	43525	6237	1395453	5834249	0.81	0.87	1.00	0.03	2.71
0.16	6153	660	1432825	5839826	0.80	0.90	1.00	0.00	2.71
0.17	16965	2131	1422013	5838355	0.80	0.89	1.00	0.01	2.70
0.15	1750	586	1437228	5839900	0.80	0.75	1.00	0.00	2.55

Fuente: Elaboración Propia.

Tabla D.1: Errores y estadísticos asociados a la matriz de confusión para distintos valores de p en la muestra de entrenamiento

p	VN	FP	FN	VP	sens	esp	PV_pos	PV_neg	suma
0.14	566	558	1438412	5839928	0.80	0.50	1.00	0.00	2.31
0.13	156	545	1438822	5839941	0.80	0.22	1.00	0.00	2.02
0.12	96	542	1438882	5839944	0.80	0.15	1.00	0.00	1.95
0.11	57	541	1438921	5839945	0.80	0.10	1.00	0.00	1.90
0.10	19	540	1438959	5839946	0.80	0.03	1.00	0.00	1.84
0.09	17	535	1438961	5839951	0.80	0.03	1.00	0.00	1.83
0.07	16	527	1438962	5839959	0.80	0.03	1.00	0.00	1.83
0.08	16	531	1438962	5839955	0.80	0.03	1.00	0.00	1.83
0.04	15	517	1438963	5839969	0.80	0.03	1.00	0.00	1.83
0.01	11	381	1438967	5840105	0.80	0.03	1.00	0.00	1.83
0.05	15	520	1438963	5839966	0.80	0.03	1.00	0.00	1.83
0.06	15	524	1438963	5839962	0.80	0.03	1.00	0.00	1.83
0.03	14	504	1438964	5839982	0.80	0.03	1.00	0.00	1.83
0.02	13	487	1438965	5839999	0.80	0.03	1.00	0.00	1.83

Fuente: Elaboración Propia.

D.2. Muestra de pruebas

Tabla D.2: Errores y estadísticos asociados a la matriz de confusión para distintos valores de p en la muestra de pruebas

p	VN	FP	FN	VP	sens	esp	PV_pos	PV_neg	suma
0.55	561068	184202	55060	2319441	0.98	0.75	0.93	0.91	3.57
0.54	556358	177716	59770	2325927	0.97	0.76	0.93	0.90	3.56
0.56	564391	192313	51737	2311330	0.98	0.75	0.92	0.92	3.56
0.57	568454	200054	47674	2303589	0.98	0.74	0.92	0.92	3.56
0.53	551181	171325	64947	2332318	0.97	0.76	0.93	0.89	3.56
0.58	571156	207111	44972	2296532	0.98	0.73	0.92	0.93	3.56
0.52	546381	166230	69747	2337413	0.97	0.77	0.93	0.89	3.56
0.59	574254	215489	41874	2288154	0.98	0.73	0.91	0.93	3.56
0.51	540521	161331	75607	2342312	0.97	0.77	0.94	0.88	3.55
0.60	576824	224120	39304	2279523	0.98	0.72	0.91	0.94	3.55
0.50	532297	155139	83831	2348504	0.97	0.77	0.94	0.86	3.54
0.61	579018	236313	37110	2267330	0.98	0.71	0.91	0.94	3.54
0.49	525901	149769	90227	2353874	0.96	0.78	0.94	0.85	3.54

Fuente: Elaboración Propia.

Tabla D.2: Errores y estadísticos asociados a la matriz de confusión para distintos valores de p en la muestra de pruebas

p	VN	FP	FN	VP	sens	esp	PV_pos	PV_neg	suma
0.62	581198	245993	34930	2257650	0.98	0.70	0.90	0.94	3.53
0.48	518883	144799	97245	2358844	0.96	0.78	0.94	0.84	3.53
0.63	582802	255272	33326	2248371	0.99	0.70	0.90	0.95	3.52
0.64	585163	265654	30965	2237989	0.99	0.69	0.89	0.95	3.52
0.47	510000	138714	106128	2364929	0.96	0.79	0.94	0.83	3.52
0.46	502717	133397	113411	2370246	0.95	0.79	0.95	0.82	3.51
0.65	586615	278730	29513	2224913	0.99	0.68	0.89	0.95	3.51
0.66	587927	288243	28201	2215400	0.99	0.67	0.88	0.95	3.50
0.45	493240	127650	122888	2375993	0.95	0.79	0.95	0.80	3.49
0.67	589373	297746	26755	2205897	0.99	0.66	0.88	0.96	3.49
0.44	481340	118707	134788	2384936	0.95	0.80	0.95	0.78	3.48
0.68	590670	308411	25458	2195232	0.99	0.66	0.88	0.96	3.48
0.43	472107	112547	144021	2391096	0.94	0.81	0.96	0.77	3.47
0.69	591627	320432	24501	2183211	0.99	0.65	0.87	0.96	3.47
0.70	592645	330697	23483	2172946	0.99	0.64	0.87	0.96	3.46
0.42	461417	107000	154711	2396643	0.94	0.81	0.96	0.75	3.46
0.71	593508	343396	22620	2160247	0.99	0.63	0.86	0.96	3.45
0.41	450453	101358	165675	2402285	0.94	0.82	0.96	0.73	3.44
0.72	594193	353195	21935	2150448	0.99	0.63	0.86	0.96	3.44
0.73	595085	363197	21043	2140446	0.99	0.62	0.85	0.97	3.43
0.40	437558	96160	178570	2407483	0.93	0.82	0.96	0.71	3.42
0.74	595676	374746	20452	2128897	0.99	0.61	0.85	0.97	3.42
0.75	596449	384344	19679	2119299	0.99	0.61	0.85	0.97	3.41
0.76	596915	394888	19213	2108755	0.99	0.60	0.84	0.97	3.40
0.39	421217	89068	194911	2414575	0.93	0.83	0.96	0.68	3.40
0.77	597534	405800	18594	2097843	0.99	0.60	0.84	0.97	3.39
0.78	598053	415717	18075	2087926	0.99	0.59	0.83	0.97	3.39
0.38	409453	84117	206675	2419526	0.92	0.83	0.97	0.66	3.38
0.79	598572	425135	17556	2078508	0.99	0.58	0.83	0.97	3.38
0.80	599175	434221	16953	2069422	0.99	0.58	0.83	0.97	3.37
0.81	599700	441879	16428	2061764	0.99	0.58	0.82	0.97	3.36
0.37	395464	78990	220664	2424653	0.92	0.83	0.97	0.64	3.36
0.82	600232	450304	15896	2053339	0.99	0.57	0.82	0.97	3.36
0.83	600803	458934	15325	2044709	0.99	0.57	0.82	0.98	3.35
0.84	601289	468296	14839	2035347	0.99	0.56	0.81	0.98	3.34
0.36	382884	74129	233244	2429514	0.91	0.84	0.97	0.62	3.34
0.85	601779	477938	14349	2025705	0.99	0.56	0.81	0.98	3.34

Fuente: Elaboración Propia.

Tabla D.2: Errores y estadísticos asociados a la matriz de confusión para distintos valores de p en la muestra de pruebas

p	VN	FP	FN	VP	sens	esp	PV_pos	PV_neg	suma
0.86	602202	488738	13926	2014905	0.99	0.55	0.80	0.98	3.33
0.87	602700	500027	13428	2003616	0.99	0.55	0.80	0.98	3.32
0.88	603201	513349	12927	1990294	0.99	0.54	0.79	0.98	3.31
0.35	357266	66439	258862	2437204	0.90	0.84	0.97	0.58	3.30
0.89	603636	526155	12492	1977488	0.99	0.53	0.79	0.98	3.30
0.90	604176	539309	11952	1964334	0.99	0.53	0.78	0.98	3.29
0.34	342214	61797	273914	2441846	0.90	0.85	0.98	0.56	3.28
0.91	604750	553778	11378	1949865	0.99	0.52	0.78	0.98	3.28
0.92	605359	570126	10769	1933517	0.99	0.51	0.77	0.98	3.26
0.93	606079	587142	10049	1916501	0.99	0.51	0.77	0.98	3.25
0.33	325858	57324	290270	2446319	0.89	0.85	0.98	0.53	3.25
0.94	606833	608042	9295	1895601	1.00	0.50	0.76	0.98	3.24
0.95	607619	634678	8509	1868965	1.00	0.49	0.75	0.99	3.22
0.32	298523	50916	317605	2452727	0.89	0.85	0.98	0.48	3.20
0.96	608626	673887	7502	1829756	1.00	0.47	0.73	0.99	3.19
0.31	282649	46612	333479	2457031	0.88	0.86	0.98	0.46	3.18
0.97	609636	720628	6492	1783015	1.00	0.46	0.71	0.99	3.16
0.30	266046	42598	350082	2461045	0.88	0.86	0.98	0.43	3.15
0.98	610503	773257	5625	1730386	1.00	0.44	0.69	0.99	3.12
0.29	235235	36727	380893	2466916	0.87	0.86	0.99	0.38	3.10
0.99	611068	827760	5060	1675883	1.00	0.42	0.67	0.99	3.08
0.28	218072	32988	398056	2470655	0.86	0.87	0.99	0.35	3.07
0.27	186815	27664	429313	2475979	0.85	0.87	0.99	0.30	3.02
0.26	170821	24654	445307	2478989	0.85	0.87	0.99	0.28	2.99
0.25	142354	20516	473774	2483127	0.84	0.87	0.99	0.23	2.94
0.24	128970	18342	487158	2485301	0.84	0.88	0.99	0.21	2.91
0.23	104088	13700	512040	2489943	0.83	0.88	0.99	0.17	2.88
0.22	83408	10822	532720	2492821	0.82	0.89	1.00	0.14	2.84
0.21	71575	9418	544553	2494225	0.82	0.88	1.00	0.12	2.82
0.20	51586	7021	564542	2496622	0.82	0.88	1.00	0.08	2.78
0.19	33782	4668	582346	2498975	0.81	0.88	1.00	0.05	2.74
0.18	18823	2663	597305	2500980	0.81	0.88	1.00	0.03	2.71
0.16	2652	294	613476	2503349	0.80	0.90	1.00	0.00	2.71
0.17	7331	894	608797	2502749	0.80	0.89	1.00	0.01	2.71
0.15	743	273	615385	2503370	0.80	0.73	1.00	0.00	2.54
0.14	232	262	615896	2503381	0.80	0.47	1.00	0.00	2.27
0.13	52	252	616076	2503391	0.80	0.17	1.00	0.00	1.97

Fuente: Elaboración Propia.

Tabla D.2: Errores y estadísticos asociados a la matriz de confusión para distintos valores de p en la muestra de pruebas

p	VN	FP	FN	VP	sens	esp	PV_pos	PV_neg	suma
0.12	30	251	616098	2503392	0.80	0.11	1.00	0.00	1.91
0.11	22	251	616106	2503392	0.80	0.08	1.00	0.00	1.88
0.03	2	230	616126	2503413	0.80	0.01	1.00	0.00	1.81
0.04	2	240	616126	2503403	0.80	0.01	1.00	0.00	1.81
0.05	2	240	616126	2503403	0.80	0.01	1.00	0.00	1.81
0.06	2	243	616126	2503400	0.80	0.01	1.00	0.00	1.81
0.07	2	245	616126	2503398	0.80	0.01	1.00	0.00	1.81
0.08	2	246	616126	2503397	0.80	0.01	1.00	0.00	1.81
0.09	2	246	616126	2503397	0.80	0.01	1.00	0.00	1.81
0.10	2	248	616126	2503395	0.80	0.01	1.00	0.00	1.81
0.01	1	175	616127	2503468	0.80	0.01	1.00	0.00	1.81
0.02	1	225	616127	2503418	0.80	0.00	1.00	0.00	1.81

Fuente: Elaboración Propia.

Apéndice E

Código en R

Para el manejo y procesamiento de la información, así como para el desarrollo de los cálculos, pruebas estadísticas, gráficas y tablas se utilizó el paquete estadístico R en su versión 3.5.1. y diversas bibliotecas con funciones útiles para el manejo de datos.

Para replicar el trabajo utilizando el código que aquí se incluye se recomienda utilizar un equipo de cómputo con al menos 16 gb de RAM debido a que por el tamaño de las tablas y ya que a que R almacena los datos en memoria RAM haciendo el consumo de este recurso es muy alto.

E.1. Bibliotecas utilizadas

Las bibliotecas utilizadas para trabajar con los datos utilizados son las siguientes:

- xtable: Esta biblioteca permite convertir matrices de R (data.table, data.frame, matrix...) en tablas listas para ser mostradas en LaTeX
- ggplot2: Presentación de gráficas.
- Information: Contiene funciones que ayudan al cálculo de IV y WOE para las variables.
- data.table: estructura para matrices ce datos especializada en grandes volúmenes de datos.
- lmtest: Contiene funciones para pruebas de bondad de ajuste y significancia de las variables para modelos lineales

```
library (xtable)
```

```
  library (ggplot2)
  library (Information)
  library (data.table)
  library (lmtest)
```

E.2. Importación y manejo de datos

```
######### FUNCIÓN PARA DETERMINAR POLIZAS VIGENTES Y CANCELADAS DURANTE
↪  EL AÑO#########

estado<-function (v){

  if  (v[3]=="NULL"){
    return (1)
  }
  if (v[1]==v[3]){
    return (1)
  }

  return (0)

}

#################################################################################################
```

```
#*****************************************************************************
#************* PREPARACIÓN DE LOS DATOS
↪  ******************************************#
#*****************************************************************************



########## LECTURA DEL ARCHIVO
↪  ##########################################################

datos<-fread (file.choose ())
gc ()

#############################################################################################

########## COMPROBACIÓN DE VARIABLES CATEGÓRICAS
↪  ######################################

datos$MOD_POL<-as.factor (datos$MOD_POL)
datos$PLAN_POL<-as.factor (datos$PLAN_POL)
datos$MONEDA<-as.factor (datos$MONEDA)
datos$ENTIDAD<-as.factor (datos$ENTIDAD)
datos$SEXO<-as.factor (datos$SEXO)
datos$STATUS_POL<-as.factor (datos$STATUS_POL)
datos$STATUS_CERT<-as.factor (datos$STATUS_CERT)
datos$PER_ESPERA<-as.factor (datos$PER_ESPERA)
```

```
datos$ANIO_POLIZA<-as.factor (datos$ANIO_POLIZA)
datos$SUBT_SEG<-as.factor (datos$SUBT_SEG)


##############################################################################################


##############################################################################################
########## SEPARAMOS LOS DATOS QUE NOS SIRVEN PARA DETERMINAR
↪  CANCELACIONES ##########
########## APLICAMOS LA FUNCIÓN PARA DETERMINAR LAS CANCELACIONES
↪  ####################
##############################################################################################


datos[,FEC_INIVIG:=NULL]
datos[,FEC_ALTACERT:=NULL]
datos[,FEC_BAJACERT:=NULL]

datos$y<-ifelse (datos$STATUS_CERT==3, 0, 1)


gc ()

##############################################################################################

##############################################################################################
########## CONVERTIMOS LAS VARIABLES CATEGÓRICAS NO ORDINALES
↪  ########################
```

```
########## EN MÚLTIPLES VARIABLES BINARIAS
↪  ##########################################
################################################################################

setDT (datos)[, c (paste ("forma_venta_",levels (datos$FORM_VENTA)),
↪  "FORM_VENTA") :=
               c (lapply (levels (datos$FORM_VENTA), function (x)
               ↪  as.factor (as.integer (x == datos$FORM_VENTA))), .
               ↪  (NULL))]

setDT (datos)[, c (paste ("Entidad_",levels (datos$ENTIDAD)),
↪  "ENTIDAD") :=
               c (lapply (levels (datos$ENTIDAD), function (x)
               ↪  as.factor (as.integer (x == datos$ENTIDAD))), .
               ↪  (NULL))]

setDT (datos)[, c (paste ("Sexo_",levels (datos$SEXO)), "SEXO") :=
               c (lapply (levels (datos$SEXO), function (x) as.factor
               ↪  (as.integer (x == datos$SEXO))), . (NULL))]

setDT (datos)[, c (paste ("Plan_",levels (datos$PLAN_POL)), "PLAN_POL")
↪  :=
               c (lapply (levels (datos$PLAN_POL), function (x)
               ↪  as.factor (as.integer (x == datos$PLAN_POL))), .
               ↪  (NULL))]

setDT (datos)[, c (paste ("Mod_",levels (datos$MOD_POL)), "MOD_POL") :=
```

```
              c (lapply (levels (datos$MOD_POL), function (x)
              ↪  as.factor (as.integer (x == datos$MOD_POL))), .
              ↪  (NULL))]
gc ()

################################################################################

########## GRÁFICA DE PROPORCIÓN CANCELACIONES
↪  ######################################


cancelaciones<-data.frame (total=summary (as.factor
↪  (datos$y)),variable=c ("Canceladas","Renovadas"))
cancelaciones$total<-cancelaciones$total/sum (cancelaciones$total)
cancelaciones$variable<-as.factor (cancelaciones$variable)
ggplot (cancelaciones, aes (x="", y=total, fill=variable) )+
  geom_bar (width = 1, stat = "identity")+
  coord_polar ("y", start=0)+geom_text (aes (label = paste0 (round
  ↪  (total*100), "%")), position = position_stack (vjust = 0.5))+
  labs (x = NULL, y = NULL, fill = NULL, title = "Distribución de las
  ↪  cancelaciones y renovaciones", colour="Tipo de póliza")+
  theme_classic () + theme (axis.line = element_blank (),
                            axis.text = element_blank (),
                            axis.ticks = element_blank (),
                            plot.title = element_text (hjust = 0.5, color
                            ↪  = "#666666"))

################################################################################
```

```
##############################################################################
########## FIJAMOS UNA SEMILLA ALEATORIA PARA
↪  ######################################
########## TENER LAS MISMAS MUESTRAS ALEATORIAS
↪  ######################################
########## GENERAMOS LAS MUESTRAS DE ENTRENAMIENTO Y PRUEBA
↪  ##########################
##############################################################################


set.seed (2017)

sub <- sample (nrow (datos), floor (nrow (datos) * 0.7))

train<-datos[sub,]
test<-datos[-sub,]

rm (datos) ##LIBERAMOS MEMORIA DE LOS DATOS ORIGINALES YA QUE TENEMOS
↪  LAS MUESTRAS

gc ()

##############################################################################
```

E.3. Selección de variables

```
↪ #*****************************************************************************#
#********** SELECCION DE VARIABLES
↪ ***********************************************#

↪ #*****************************************************************************#



↪ #/////////////////////////////////////////////////////////////////////////////#
#////////// IV Y WOE
↪ ////////////////////////////////////////////////////////#

↪ #/////////////////////////////////////////////////////////////////////////////#


tablas<-vector ()
IV<-NULL

y_pos<-which (names (train)=="y")

########## DEBIDO A COMO GESTIONA LA MEMORIA EL PAQUETE INFORMATION
↪ HAY QUE
########## FRACCIONAR LOS DATOS PARA LIBERAR MEMORIA AL TERMINAR CADA
↪ GRUPO DE VARIABLES
```

```
T1<-create_infotables (data=train[,c (1:5,y_pos), with=FALSE], y="y",
↪  parallel = TRUE)
gc ()
T2<-create_infotables (data=train[,c (6:10,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T3<-create_infotables (data=train[,c (11:15,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T4<-create_infotables (data=train[,c (16:20,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T5<-create_infotables (data=train[,c (21:25,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T6<-create_infotables (data=train[,c (26:30,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T7<-create_infotables (data=train[,c (31:35,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T8<-create_infotables (data=train[,c (36:40,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T9<-create_infotables (data=train[,c (41:45,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T10<-create_infotables (data=train[,c (46:50,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
```

```
gc ()
T11<-create_infotables (data=train[,c (51:55,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T12<-create_infotables (data=train[,c (56:60,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T13<-create_infotables (data=train[,c (61:65,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T14<-create_infotables (data=train[,c (66:70,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T15<-create_infotables (data=train[,c (71:75,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()
T16<-create_infotables (data=train[,c (76:78,y_pos), with=FALSE],
↪  y="y", parallel = TRUE)
gc ()



######### UNIMOS LOS RESULTADOS OBTENIDOS PARA ANALIZARLOS


IV<-rbind (T1$Summary, T2$Summary, T3$Summary, T4$Summary,
↪  T5$Summary, T6$Summary, T7$Summary, T8$Summary, T9$Summary,
↪  T10$Summary, T11$Summary, T12$Summary,T13$Summary, T14$Summary)
IV<-IV[order (-IV$IV),] ### TABLA IV ORDENADA
xtable (IV, digits = c (0,0,6)) ### TABLA FORMATO LATEX 6 DECIMALES
```

```
Tablas<-c (T1$Tables, T2$Tables, T3$Tables, T4$Tables, T5$Tables,
↪  T6$Tables, T7$Tables, T8$Tables, T9$Tables, T10$Tables,
↪  T11$Tables, T12$Tables,T13$Tables, T14$Tables,T15$Tables  )



######### GRAFICAS DE WOE PARA LAS VARIABLES CON PREDICTIBILIDAD
i=1
while (IV[i,2]>0.02){
nombre<-IV[i,1]
valor_IV<-IV[i,2]
T<-as.data.frame (Tablas[nombre])
p=ggplot (T,aes (x=T[,1], y=100*T[,4]))+geom_col ()+labs (x=nombre,
↪  y="WOE", title=paste (nombre, "; IV=", valor_IV, sep=""))
print (p)
ggsave (paste (nombre,"_WOE.png", sep=""), plot = p)
i=i+1
}
gc ()

########## RETIRAMOS LAS VARIABLES SIN PREDICTIBILIDAD

train[,MONEDA:=NULL]
train[,FUMADOR:=NULL]
train[,STATUS_POL:=NULL]
train[,STATUS_CERT:=NULL]
```

```
↪ #////////////////////////////////////////////////////////////////////////////#
#////////// ESTADISTICO KS
↪ ////////////////////////////////////////////////////#

↪ #////////////////////////////////////////////////////////////////////////////#


tabla_ks<-data.frame (variable=NULL, KS=NULL)
tablasks<-list ()

nombres<-names (train)
malos<-train[y==0]
buenos<-train[y==1]
for (i in 1:length (nombres)){

      nom_variable<-nombres[i]
  if (nom_variable!="y"){
    if (sapply (train,class)[which (names
    ↪  (train)==nom_variable)]=="integer"){
      #Print ("Entero")
      vec_buenos<-buenos[[nom_variable]]
      vec_malos<-malos[[nom_variable]]
      ran_min<-min (train[,..nom_variable])
      ran_max<-max (train[,..nom_variable])
      if ( (ran_max-ran_min+1)<500){
        dom<-seq (ran_min,ran_max, length.out =  (ran_max-ran_min+1))
      }else if ( (ran_max-ran_min+1)>=500){
```

```
    dom<-seq (ran_min,ran_max, length.out = 500)
  }
  }
if (sapply (train,class)[which (names
↪  (train)==nom_variable)]=="factor"){
  #print ("Factor")
  vec_buenos<-as.numeric (levels
  ↪  (buenos[[nom_variable]]))[buenos[[nom_variable]]]
  vec_malos<-as.numeric (levels
  ↪  (malos[[nom_variable]]))[malos[[nom_variable]]]
  ran_min<-min (min (vec_buenos), min (vec_malos))
  ran_max<-max (max (vec_buenos), max (vec_malos))
  dom<-as.numeric (levels (train[[nom_variable]]))
  }


if (sapply (train,class)[which (names
↪  (train)==nom_variable)]=="numeric"){
  #print ("Numerico")
  vec_buenos<-buenos[[nom_variable]]
  vec_malos<-malos[[nom_variable]]
  ran_min<-min (train[,..nom_variable])
  ran_max<-max (train[,..nom_variable])
  dom<-seq (ran_min,ran_max, length.out = 500)

}
```

```
    fbuenos<-ecdf (vec_buenos)
    fmalos<-ecdf (vec_malos)

    buenos_malos<-data.frame (x=dom,buenos=fbuenos (dom),
    ↪  malos=fmalos (dom))
    buenos_malos$diff<-abs (buenos_malos$buenos-buenos_malos$malos)
    KS<-max (buenos_malos$diff)
    posicion<-which (buenos_malos$diff==KS)

    tablasks[[nom_variable]]<-buenos_malos

    tabla_ks<-rbind (tabla_ks,data.frame (Variable=nom_variable, KS))

    g<-ggplot (buenos_malos)+
      geom_line (aes (x=x, y=buenos, colour="Renovaciones"), size=1)+
      geom_line (aes (x=x, y=malos, colour="Cancelaciones"), size=1)+
      geom_segment (aes (x=posicion, y=buenos_malos$buenos[posicion],
      ↪  xend=posicion, yend=buenos_malos$malos[posicion]),
      ↪  linetype="dashed", color="blue")+
      labs (x=nom_variable, y="Porcentaje de la población",
      ↪  title=paste ("Variable: ",nom_variable,"; KS: ",KS,sep=""),
      ↪  colour="Población")+
      ylim (0,1)
    print (g)
    ggsave (paste (nom_variable,"_KS.png", sep=""), plot = g)
    gc ()
  }
}
```

```
  tabla_ks<-tabla_ks[order (-tabla_ks$KS),] ##TABLA DE KS ORDENADA
  xtable (tabla_ks, digits = c (0,0,6)) ##IMPRESIÓN PARA LATEX CON 6
  ↪  DECIMALES


  ↪  ##############################################################################
```

E.4. Ajuste del modelo

```
#*****************************************************************************
#********** AJUSTE DEL MODELO
↪  ************************************************
#*****************************************************************************




#buenas<-c ("SA_BEN1","SA_BEN6","SA_BEN4","SA_BEN8","SA_BEN3", "Plan_
↪  1","Plan_ 3","Mod_ 1", "Mod_ 2", "EDAD", "EMISION",  "y") #
↪  VARIABLES SELECCIONADAS BAJO AMBOS CRITERIOS

buenas<-c ("SA_BEN1","SA_BEN2","SA_BEN3","SA_BEN4", "EDAD","Sexo_ M",
↪  "EMISION",  "y") # VARIABLES SELECCIONADAS BAJO AMBOS CRITERIOS
```

```
train<-train[,buenas,  with=FALSE]
test<-test[,buenas,  with=FALSE]
gc ()

modelo<-glm (y~., family = "binomial", data=train)



summary (modelo) ## LA PRUEBA DE WALD CORRESPONDE CON LA COLUMNA "z" DE
↪  LA SALIDA DEL MODELO

xtable (as.data.frame (modelo$coefficients), digits = c (0,10)) ###
↪  TABLA EN FORMATO LATEX DE LOS COEFICIENTES CON 6 DECIMALES
gc ()

################################################################################
```

E.5. Pruebas estadísticas

```
  #////////// PRUEBAS DE BONDAD DE AJUSTE
  ↪  ////////////////////////////////////

########## PRUEBA CHI-CUADRADA
↪  ##################################################
chisqtest<-list ()
for (i in 1:11){

temptest<-lrtest (modelo, i)
```

```
chisqtest[[buenas[i]]]<-temptest
gc ()
}

################################################################################


########## PRUEBA DE HOSMER-LEMESHOW
↪  ##########################################
pval<-vector ()

for (i in 1:1000){
subc<-sample (length (modelo$y), 1000)

pval[i]<-hoslem.test (modelo$y[subc], fitted (modelo)[subc],
↪  10)$p.value
gc ()
}

################################################################################


########## pSEUDO R2
↪  ############################################################

pR2 (modelo)

################################################################################
```

```
#////////////////////////////////////////////////////////////////////////////
```

E.6. Validación del modelo

```
  muestra<-sample (nrow (train), 200)
confusion<-data.frame (x=modelo$fitted.values[muestra],
↪  y=train$y[muestra])
confusion$pred<-ifelse (confusion$x>0.5,1,0)
confusion$grupo<-ifelse (confusion$y==confusion$pred, "Buena
↪  predicción", ifelse (confusion$y==1, "Falso Negativo", "Falso
↪  Positivo"))


ggplot (confusion)+geom_point (aes (x=x, y=y, colour=grupo))+labs
↪  (x=parse (text=TeX ('$\\pi (x_i)$')), colour="Tipo de predicción",
↪  title= parse (text=TeX ('Predictibilidad en muestra de 200 datos
↪  $p=0.5$')))+xlim (0,1)+ylim (0,1)+geom_segment (aes (x=0.5, y=0,
↪  xend=0.5,yend=1), linetype="dashed")

cortes<-seq (1,100, length.out = 100)/100
tabla_resumen<-data.frame ()

for (i in 1:100){
pred<-ifelse (modelo$fitted.values>cortes[i],1,0)
tabla_conf<-table (pred, train$y)
```

```
renglon_confusion<-data.frame
↪  (p=cortes[i],VN=tabla_conf[1,1],FP=tabla_conf[1,2],
↪  FN=tabla_conf[2,1], VP=tabla_conf[2,2])

tabla_resumen=rbind (tabla_resumen,renglon_confusion)
gc ()
}

tabla_resumen$sensitividad<-tabla_resumen$VP/
↪  (tabla_resumen$FN+tabla_resumen$VP)
tabla_resumen$especificidad<-tabla_resumen$VN/
↪  (tabla_resumen$VN+tabla_resumen$FP)
tabla_resumen$PV_pos<-tabla_resumen$VP/
↪  (tabla_resumen$VP+tabla_resumen$FP)
tabla_resumen$PV_neg<-tabla_resumen$VN/
↪  (tabla_resumen$VN+tabla_resumen$FN)
tabla_resumen$suma<-tabla_resumen$especificidad+tabla_resumen$sensitividad+tabla_res



tabla_resumen_ordenada<-tabla_resumen[order (tabla_resumen$suma),]


ggplot (tabla_resumen)+geom_line (aes (x=p, y=sensitividad,
↪  colour="Sensitividad"))+
  geom_line (aes (x=p, y=especificidad,colour="Especificidad"))+
  geom_line (aes (x=p, y=PV_pos,colour="PV_pos"))+
  geom_line (aes (x=p, y=PV_neg,colour="PV_neg"))+
```

```
  labs (y="valor", title="Comportamiento de los estadisticos asociados
  ↪  a la tabla de confusion", colour="Estadístico")+
  geom_segment (aes (x=0.82, y=0, xend=0.82,yend=1), linetype="dashed")



tabla_ROC<-tabla_resumen[,c ("p", "sensitividad")]
tabla_ROC$esp<- (1-tabla_resumen$especificidad)

tabla_ROC<-rbind (tabla_ROC, c (200,0,0), c (201,1,1))
tabla_ROC<-tabla_ROC[order (tabla_ROC$esp),]
ggplot (tabla_ROC)+geom_line (aes (x= (esp), y=sensitividad))+ylim
↪  (0,1)+xlim (0,1)+labs (x=" (1-Especificidad)", title="Curva ROC
↪  para la muestra de entrenamiento del modelo")

AUC<-0
for (i in 1:100){
  h=tabla_ROC[i+1,3]-tabla_ROC[i,3]
  a=tabla_ROC[i,2]
  b=tabla_ROC[i+1,2]
  ar<- (a+b)*h/2
  AUC<-AUC+ar
}
AUC
```

Bibliografía

[1] (2015). *Circular Unica de Seguros y Fianzas*, Ciudad de México. CONGRESO GENERAL DE LOS ESTADOS UNIDOS MEXICANOS.

[2] (2015). *Ley de Instituciones de Seguros y Fianzas*, Ciudad de México. CONGRESO GENERAL DE LOS ESTADOS UNIDOS MEXICANOS.

[3] (2016). *RR-8 Catálogos Seguros 1 a 231 excepto 4, 10, 16.2, 16.3, 41, 80, 81, 82, 150 y 151*, Ciudad de México. Comisión Nacional de Seguros y Fianzas.

[4] Anderson, R. (2007). *The Credit Scoring Toolkit*. Oxford University Press. 39, 50, 51

[5] Aristizabal, J. M. (2017). Regresión Logística multinomial.

[6] Bo, H. (2005). Pseudo-r2 in logistic regression model. 32

[7] Bowers, Gerber, e. A. (1997). *Actuarial Mathematics*. Society of Actuaries; 2nd edition, New York. 12

[8] CDC (2017). Fatal injury data https://www.cdc.gov/injury/wisqars/fatal.html. 42

[9] CNSF (2016). Manual del sistema estadístico de los seguros de vida individual. 82, 83, 84, 87

[10] CNSF (2017). Comportamiento regulatorio definitivo de los sectores aseguradory afianzador con cifras al 31 de marzo de 2017. 56

[11] Draper, P. (2002). Customer loyality in the insurance industry a logistic regression approach. Samos, Grecia. 36

[12] D.W. Hosmer, S. Lemeshow, R. S. (2013). *Applied Logistic Regression*. Wiley, New Jersey. 19, 32, 38

[13] EIA (2018). Petroleum and other liquids https://www.eia.gov/dnav/pet/hist/leafhandler.ashx. 42

[14] Franco, A. B. (2006). Teoría del seguro de vida.

[15] García, O. S. (1994). Sistemas de reserva modificada en el seguro de vida.

[16] Hal R. Varian, T. E. R. (1992). *Análisis Microeconómico*. Antoni Bosch, Barcelona.

[17] J. D. Gibbons, S. C. (2003). *Non parametric Statistical Inference*. Marcel Dekker, New York. 24, 31, 45

[18] Lewis, E. M. (1994). *Introduction to Credit Scoring*. Athena Press, New York. 2

[19] M. Miller, D. R. (2004). Improving access to credit for smes: An empirical analysis of the viability of pooled data sme credit scoring models in brazil, colombia and mexico. 3

[20] Mays, E. (2011). *Credit Scoring for Risk Managers: The Handbook for Lenders*. CreateSpace Independent Publishing Platform, New York. 46

[21] P. de Jong, G. H. (2008). *Generalized Linear Models for Insurance Data*. Cambridge, New York. 51, 53

[22] Rincon, L. (2012). *Introducción a la teoría del riesgo*. Facultad de Ciencias UNAM, Ciudad de México. 9

[23] Rodríguez, J. R. (1976). *El contrato de Seguro en el derecho mexicano*. AMIC Editor. 7

[24] Siddiqi, N. (2006). *Credit Risk Scorecards*. Wiley. 45, 48, 49

[25] Verzani, J. (2014). *Using R for introductory statistics*. Chapman and Hall/CRC.

[26] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62. 27, 29