

**UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN**



**“PREDICCIÓN DE LA PRIMA DE LOS CLIENTES DE UNA
COMPAÑÍA ASEGURADORA USANDO EL MODELO LINEAL
GENERALIZADO TWEEDIE”**

**TRABAJO DE SUFICIENCIA PROFESIONAL
PARA OPTAR EL TÍTULO DE
INGENIERO ESTADÍSTICO E INFORMÁTICO**

ALBERTO REIMUNDO ARMAS ALVARADO

LIMA – PERÚ

2020

**UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN**

**“PREDICCIÓN DE LA PRIMA DE LOS CLIENTES DE UNA COMPAÑÍA
ASEGURADORA USANDO EL MODELO LINEAL GENERALIZADO
TWEEDIE”**

Presentado por:

ALBERTO REIMUNDO ARMAS ALVARADO

**TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL TÍTULO DE
INGENIERO ESTADÍSTICO E INFORMÁTICO**

Sustentado y aprobado ante el siguiente Jurado:

**Mg. Iván Dennys Soto Rodríguez
PRESIDENTE**

**Dr. César Higinio Menacho Chiok
ASESOR**

**Mg. Jesús Walter Salinas Flores
MIEMBRO**

**MS. Grimaldo José Febres Huamán
MIEMBRO**

Lima - Perú

2020

DEDICATORIA

A mi padre Raymundo Armas y a mi madre Plácida Alvarado por sus enseñanzas, guía y apoyarme en cada paso que doy.

A mi hermana Margoth por el apoyo invaluable y a mi hermana Anabel alegría del hogar.

A mis abuelos y abuelas por su cariño.

AGRADECIMIENTO

Agradezco a mi padre y a mi madre por el apoyo brindado durante el desarrollo de mi carrera.

Agradezco al Dr. César Higinio Menacho Chiok por las pautas brindadas durante el desarrollo del presente trabajo.

ÍNDICE GENERAL

I.	PRESENTACIÓN.....	1
II.	INTRODUCCIÓN	3
III.	OBJETIVOS	5
	3.1. Objetivo general.....	5
	3.2. Objetivos específicos	5
IV.	CUERPO DEL TRABAJO	6
	4.1. Funciones desempeñadas	6
	4.2. Puesta en práctica de lo aprendido en la carrera.....	6
	4.2.1. Descripción de las técnicas estadísticas e informáticas usadas	6
	4.2.2. Revisión de Literatura	10
	4.2.3. Propuesta de alternativa de solución	15
	4.3. Contribución en la solución del problema	48
	4.4. Análisis de la contribución en términos de competencias y habilidades	48
	4.5. Nivel de beneficio obtenido por el centro laboral.....	49
V.	CONCLUSIONES Y RECOMENDACIONES.....	50
	5.1. Conclusiones	50
	5.2. Recomendaciones	51
VI.	REFERENCIAS BIBLIOGRÁFICAS.....	52
VII.	ANEXOS	53

ÍNDICE DE TABLAS

Tabla 1: Tipos y funciones de enlace	8
Tabla 2: Principales variables objetivo de los Modelos Lineales Generalizados.....	12
Tabla 3: Resumen de los modelos Tweedie	14
Tabla 4: Funciones de enlace distribuciones de la familia exponencial.....	14
Tabla 5: Indicadores de riesgo muestra de Construcción	20
Tabla 6: Indicadores de riesgo muestra de Validación.....	20
Tabla 7: Análisis descriptivo variables cuantitativas Modelo 1	21
Tabla 8: Análisis descriptivo variables cuantitativas Modelo 2	29
Tabla 9: Matriz de correlaciones Modelo 1	37
Tabla 10: Matriz de correlaciones Modelo 2.....	38
Tabla 11: Coeficientes estimados Modelo 1	38
Tabla 12: Coeficientes estimados Modelo 2	39
Tabla 13: Bondad de ajuste Modelo 1	40
Tabla 14: Prueba Ómnibus Modelo 1	40
Tabla 15: Análisis de los estimadores Modelo 1	41
Tabla 16: Bondad de ajuste Modelo 2	42
Tabla 17: Prueba Ómnibus Modelo 2.....	42
Tabla 18: Análisis de los estimadores Modelo 2	43
Tabla 19: Nivel de calibración periodo de construcción	49
Tabla 20: Nivel de calibración periodo de validación.....	49

ÍNDICE DE FIGURAS

Figura 1: Diagrama Datamart de Autos.....	18
Figura 2: Histograma y diagrama de cajas Suma asegurada	22
Figura 3: Histograma y diagrama de cajas Edad cliente	23
Figura 4: Histograma y diagrama de cajas Edad cliente solo Personas.....	24
Figura 5: Gráfico de barras Cantidad de siniestros últimos 2 años	25
Figura 6: Histograma y diagrama de cajas Monto siniestro/Suma asegurada último año...	26
Figura 7: Gráfico de pie Tipo de cliente.....	27
Figura 8: Gráfico de pie Marca japonesa, Alta gama	28
Figura 9: Gráfico de barras Cantidad de siniestros por robo últimos 2 años	30
Figura 10: Gráfico de pie Marcas más robadas	30
Figura 11: Gráfico bivariado Suma asegurada Modelo 1	32
Figura 12: Gráfico bivariado Edad cliente Modelo 1	33
Figura 13: Gráfico bivariado Tipo cliente Modelo 1.....	33
Figura 14: Gráfico bivariado Marca japonesa Modelo 1.....	34
Figura 15: Gráfico bivariado Marca alta gama Modelo 1	35
Figura 16: Gráfico bivariado Suma asegurada Modelo 2.....	35
Figura 17: Gráfico bivariado Edad cliente Modelo 2.....	36
Figura 18: Gráfico bivariado Marcas más robadas Modelo 2	37
Figura 19: Gráfico de elevación modelo agregado, muestra construcción.....	44
Figura 20: Gráfico de elevación Modelo 1 y Modelo 2, muestra construcción	45
Figura 21: Gráfico de elevación modelo agregado, muestra validación	46
Figura 22: Gráfico de elevación Modelo 1 y Modelo 2, muestra validación	47

ÍNDICE DE ANEXOS

Anexo 1: Gráfico bivariado continuación Modelo 1	53
Anexo 2: Gráfico bivariado continuación Modelo 2	56
Anexo 3: Gráfico histórico mensual Modelo 1	59
Anexo 4: Gráfico histórico mensual Modelo 2	59

I. PRESENTACIÓN

La compañía de seguros, donde se desarrolló el presente trabajo, fue fundada en 1944, en su proceso de crecimiento tuvo diversas fusiones y adquisiciones, tal es el caso de 1992, ya en 1995 pasa a formar parte de un holding. En 1997 inicia operaciones para la parte de seguros vida, en 1999 completó su portafolio con el negocio de salud. Continuando su crecimiento, en el año 2011, en salud adquirió diversas clínicas y centros médicos. En el año 2015 firmó una alianza internacional. En agosto de 2017 se concretó la fusión de sus líneas de negocio vida, salud y seguros generales.

El objetivo de la compañía se centra en que sus clientes enfrenten la vida con absoluta confianza y tranquilidad de saber que están respaldados en todo momento. Esta filosofía es el resultado de un enfoque hacia el cliente, en el que se preocupan por ponerse en el lugar de éste y entender que está buscando de una compañía aseguradora. Se trabaja para ofrecer soluciones que se ajustan a sus necesidades y brindar el respaldo necesario ante situaciones inesperadas a través de sus líneas de negocios: Salud, Vida y Riesgos Generales.

La misión y visión de la compañía se enfoca en sus clientes:

MISIÓN.- Ayudar a los clientes a proteger su estabilidad económica, ofreciéndoles soluciones que protejan aquello que valoran y aseguren el cumplimiento de sus objetivos.

VISIÓN.- Ser una de las cinco mejores aseguradoras de Latinoamérica: simple, transparente, accesible, rentable y con colaboradores altamente competentes y motivados.

La División Técnico de Negocios tiene la labor de monitorear la suficiencia de primas y de reservas de siniestros, asegurando así los más altos estándares en la suscripción, tarificación y gestión de las reservas, para todos los ramos del negocio de seguros.

El Área de Modelo de Riesgo Técnico tiene como responsabilidades: Analizar reportes de seguimiento de los modelos y metodologías en uso e identificar posibles desviaciones, a través de indicadores de medición del riesgo. Se realiza el proceso de tarificación según las necesidades del Área de Negocios Vehiculares. Se desarrolla modelos estadísticos para estimación de precios de un producto de seguro vehicular ya sea existente o uno nuevo. Calibración de modelos según resultados de seguimientos y acuerdos con el área usuaria. Diseño, coordinación y análisis de reportes para seguros vehiculares. Desarrollo de modelos comerciales para estimación de precios de un producto vehicular existente o nuevo. Presentar informes de los modelos desarrollados. Elabora notas técnicas de los productos vehiculares. Es un área que brinda soporte y trabaja conjuntamente con el Área de Negocios Vehiculares. Cabe señalar que se cuenta con la asesoría de firmas independientes de servicios actuariales con reconocido prestigio, las cuales se encargan de hacer una revisión periódica respecto de la adecuada suficiencia de reservas para las empresas que conforman el grupo asegurador.

El Consultor de Modelo de Riesgo Técnico se relaciona directamente con el Área de Negocios Vehiculares de acuerdo con los requerimientos de ésta área desarrolla funciones en base a análisis o estudios que se soliciten, tales como calibración, desarrollo, participación activa durante la implementación de modelos, tarificación de canales y segmentos específicos. Para realizar las funciones del puesto se necesita información proveniente de las bases de datos de la compañía. Los resultados de los estudios, al finalizarlos, primero se presentan de manera interna con el jefe directo una vez aprobado y agregado los cambios pertinentes se presenta al Área usuaria de Negocios Vehiculares, a la cual se brinda soporte en temas analíticos.

II. INTRODUCCIÓN

El proceso de cálculo de precios (*pricing*) es fundamental en el campo de los seguros para garantizar la estabilidad y la solidez financiera, tiene como principal objetivo calcular primas de manera equitativa y suficiente, tomando en cuenta el nivel de riesgo del cliente.

Los Modelo Lineales Generalizados son de amplio uso en el campo Actuarial, permiten predecir la frecuencia, pérdida esperada, monto de siniestros, así en base a estos componentes calcular la prima pura de riesgo. El modelo Tweedie permite estimar el monto promedio de siniestros de manera directa, motivo por el cual se trabajó con este modelo.

En la División Técnico de Negocios de la Compañía de Seguros en estudio se tenía un modelo de tarificación construido entre los años 2013 y 2014, que a la fecha de estudio presentaba desajuste en las variables, debido a cambios en el parque automotor por políticas de gobierno, como la prohibición de importación de vehículos usados con más de 3 años de antigüedad, lo cual conlleva a renovar unidades e incentivar las ventas de vehículos nuevos en los distintos concesionarios, el riesgo de algunas marcas de vehículos ha variado respecto a lo que se tenía en años anteriores. La desviación de los modelos de precios no se podía corregir desplazando la curva, sino con la construcción de un nuevo modelo con población reciente, que recoja el comportamiento del mercado actual.

En el Perú el parque vehicular se incrementó en 63% en los últimos 10 años, al pasar de 1.3 millones de unidades a 2.2 millones según el documento estadístico Parque Vehicular en la Comunidad Andina 2004-2013 (Gestión, 2014).

Si se sigue estableciendo los precios de las primas con un modelo desviado, es posible que la compañía siga pagando siniestros elevados y cobrando primas por debajo del nivel de riesgo del cliente, lo cual llevará a pérdidas para la compañía, por hallar un déficit en la

suficiencia de primas y reservas de siniestros. Si la empresa no es sostenible no podrá satisfacer las necesidades de sus clientes en cuanto a los bienes asegurados con los que ya cuenta.

Por otro lado si no se establecen primas adecuadas, los siniestros ocurridos tales como accidentes de tránsito, choques, volcaduras, despistes, robo de autopartes y de vehículos seguirán en aumento perjudicando a la población y el parque automotor.

El presente trabajo se encuentra dividido en siete partes, luego de la presentación e introducción, se desarrollan los objetivos, en la cuarta parte se presenta el cuerpo del trabajo, en la quinta parte las conclusiones y recomendaciones, en el sexto apartado las referencias y por último los anexos.

III. OBJETIVOS

3.1. Objetivo general

Predecir la prima de los clientes de seguros vehiculares usando el modelo lineal generalizado Tweedie.

3.2. Objetivos específicos

- Estimar un modelo estadístico con indicadores de predicción dentro de los estándares requeridos por la compañía aseguradora.
- Identificar variables significativas en la predicción de la prima.
- Obtener estimaciones del precio de la prima para un cliente.

IV. CUERPO DEL TRABAJO

4.1. Funciones desempeñadas

A continuación, se describen las principales funciones desempeñadas en el Área de Modelos Riesgo Técnico: manejo de base de datos, sistematización de la información, automatización de procesos, modelamiento estadístico. Se detallan las funciones realizadas por el cargo:

- Desarrollo, calibración e implementación de modelos estadísticos de precios en el sistema de seguros.
- Seguimiento de los modelos vigentes para detectar posibles desviaciones y oportunidades.
- Búsqueda de datos y nuevas fuentes de información para mejorar la capacidad predictiva de los modelos.
- Liderar proyectos, definir uso de los modelos y apoyar en la integración para el uso en la gestión del negocio.
- Mejoramiento continuo de las metodologías para el desarrollo de los modelos así como la elaboración de propuestas.
- Generar señales de alerta temprana ante un posible deterioro del ramo de seguros vehiculares.

4.2. Puesta en práctica de lo aprendido en la carrera

4.2.1. Descripción de las técnicas estadísticas e informáticas usadas

Los modelos lineales generalizados son utilizados para determinar la relación existente entre la variable respuesta y variables explicativas, estos modelos fueron propuestos por John Nelder y Robert Wedderburn en 1972 citados por (Quishpe Tasiguano, 2015), los cuales se diferencian de los modelos de regresión lineal en tres aspectos de la variable respuesta:

- La distribución pertenece a la familia exponencial, la normal es un caso particular.
- El valor esperado se relaciona de manera lineal con las variables explicativas, a través de una función de enlace.
- La varianza no necesariamente es constante, es una función del valor esperado.

Debido a la naturaleza de los datos que difícilmente presentan distribución normal y tampoco presentan homocedasticidad, las características de los modelos lineales generalizados lo hacen muy útiles en aplicaciones de seguros.

Los Modelos Lineales Generalizados (MLG) son una extensión de los Modelos Lineales, los cuales permiten usar distribuciones no normales de los errores y varianzas no constantes. La relación no lineal se origina al restringir el rango de la variable respuesta, debido a esto es necesario aplicar transformaciones en ella o lo que es más común, emplear funciones de enlace pertinentes (Hardin & Hilbe, 2012). Estos modelos permiten calcular el valor esperado de una variable respuesta dado un conjunto de variables explicativas, pudiendo considerar inclusive sus interacciones, es decir el efecto de que una variable cambie de acuerdo con los niveles de otra variable.

Los MLG se usan cuando la variable aleatoria no sigue una distribución normal, pero su distribución pertenece a la familia exponencial (binomial, Poisson, Gamma, etc.) modeliza $E[y]$ como una función no lineal de $X\beta$. Es un conjunto de variables aleatorias independientes y_1, y, \dots, y_k con función de densidad, que puede escribirse como (De Jong & Heller, 2008):

$$p(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

Donde:

θ_i es el parámetro natural o canónico

ϕ es un parámetro adicional de escala o de dispersión

$a_i(\cdot)$, $b(\cdot)$ y $c(\cdot)$ son funciones específicas

Si ϕ es conocido este es un modelo de la familia exponencial lineal

Si ϕ es desconocido es un modelo de dispersión exponencial

Se quiere modelizar $u_i = E[y_i]$ en términos del predictor lineal $\beta'X_i$ formado con un conjunto de p covariables

$$\beta'X_i = \beta_0 + \beta_1x_i^1 + \dots + \beta_px_i^p$$

López de Castilla (2014), citado por (Quishpe Tasiguano, 2015), estos modelos tienen tres componentes:

1. **Componente Aleatorio.-** Identifica la variable respuesta y su distribución de probabilidad.
2. **Componente Sistemático.-** Identifica las variables explicativas usadas en una función predictor lineal. Conjunto de n variables respuesta independientes, de una distribución de la familia exponencial.
3. **Función de Enlace.-** Se encarga de linealizar la relación entre la variable respuesta y las variables independientes mediante la transformación de la variable respuesta. Es una función vínculo monótona y diferenciable que define la relación entre u_i y su predictor lineal.

$$g(\mu_i) = \beta'X_i$$

Permite modelizar distintas relaciones entre μ y el predictor lineal.

Vínculo Natural o Canónico:

Aquel que es igual a la función que define el parámetro natural o canónico de esa distribución. Por tanto $\theta = \beta'X$, se pueden apreciar los vínculos más usuales en la Tabla 1 (López, 2004):

Tabla 1: Tipos y funciones de enlace

Tipo de Enlace	Función
Logit	$\log \frac{\pi}{1 - \pi}$
Probit	$\Phi^{-1}(\pi)$
Complementario log-log	$\log[-\log(1 - \pi)]$
Identidad	μ
Inverso	$-\frac{1}{\mu}$
Logaritmo	$\log(\mu)$
Raíz cuadrada	$\sqrt{\mu}$

Fuente: Elaboración propia a partir de (López, 2004)

La elección de la función de enlace depende de la familia de distribuciones, del tipo de respuestas y de la aplicación (López, 2004).

Tipos de modelos

Se presentan las distribuciones a utilizar de acuerdo con el tipo de datos en la variable respuesta (López, 2004):

1. Modelos de datos discretos

a. Binomial:

Distribución Bi $\sim(n, \pi)$

$$E[y] = n\pi$$

Función de enlace $g(\mu) = \log \frac{\mu}{n-\mu} = \log \frac{\pi}{1-\pi}$ Logit

$$b(\theta) = n \log(1 + e^\theta), a(\emptyset) = 1$$

Otras funciones de enlace: Probit, Complementario Log-log

b. Poisson:

Distribución Po $\sim(\lambda)$

$$E[y] = \lambda$$

Función de enlace $g(\lambda) = \log \lambda$ Logaritmo

$$b(\theta) = e^\theta, a(\emptyset) = 1$$

Otras funciones de enlace: Identidad, Raíz cuadrada

2. Modelos de datos continuos

a. Normal:

Distribución N $\sim(\mu, \sigma^2)$

$$E[y] = \mu$$

Función de enlace $g(\mu) = \mu$ Identidad

$$b(\theta) = \frac{\theta^2}{2}, a(\emptyset) = \sigma^2$$

Otras funciones de enlace: Logaritmo, Raíz cuadrada

b. Gamma:

Distribución Gamma $\sim(\lambda, v)$

$$E[y] = \frac{\lambda}{v}$$

Función de enlace $g(\mu) = -\frac{1}{\mu} = -\frac{v}{\lambda}$ Inverso

$$b(\theta) = -\log(-\theta), a(\phi) = \frac{1}{\lambda}$$

Otras funciones de enlace: Logaritmo, Identidad

4.2.2. Revisión de Literatura

En (Quishpe Tasiguano, 2015) su estudio titulado “Factores de Riesgo de Siniestralidad y cálculo de primas de los vehículos asegurados en el Ecuador mediante Modelos Lineales Generalizados” tuvo como hallazgos: los factores que contribuyen a que exista el fenómeno de siniestralidad con mayor probabilidad, son los vehículos fabricados antes del 2012, de marcas diferentes a Chevrolet, suscribir una póliza nueva y asegurar vehículos por montos entre \$15,000 y \$50,000.

En (Villarino González, 2017) su estudio titulado: “Análisis de Riesgo para la Tarificación de Seguros de Automóvil mediante Modelos Lineales Generalizados” encontró que la distribución que mejor se ha comportado respecto a sus datos fue la log-normal, ya que se obtuvieron los mejores valores en máxima verosimilitud, AIC, BIC, Kolmogorov-Smirnov y Anderson-Darlin, posterior a ello se estimó el GLM Gamma con link logarítmico y se comparó con un GLM Log-normal mediante gráficos de diagnóstico de normalidad de los residuos llegando a la conclusión de que efectivamente la predicción es más válida si se considera que los errores siguen una distribución Log-normal. Tras ello mejoró el nuevo modelo mediante la eliminación de la variable del tipo de vehículo y de la interacción entre género y edad, y añadiendo la variable del valor del vehículo en forma lineal y cuadrática, el grupo de edad, el género, el área, la exposición al riesgo y la antigüedad del vehículo, las cuales son estadísticamente significativas y del mismo modo han mejorado el AIC y deviance, resultando en un modelo mejor.

(Martínez Gutiérrez, 2017) en su investigación realizada en la Ciudad de México llamada: “Análisis de las Primas de Riesgo en Seguros de Automóviles: Una aplicación de los

Modelos Lineales Generalizados” en la que tuvo como hallazgos: Tomó en cuenta factores de riesgo asociados a los asegurados que influyen en el comportamiento siniestral e intervienen en el proceso de tarificación. Debido a la diversificación de esos factores realizó segmentación con métodos aglomerativos de clusterización tuvo en cuenta su experiencia siniestral en grupos homogéneos logrando con esto que los asegurados que pertenecen a un mismo grupo paguen la misma tarifa, evitando la antiselección, tomó en cuenta diferente número de grupos para poder tener una mayor comparación y así pudo elegir la segmentación más idónea para el cálculo de la prima de riesgo. La frecuencia no presentó problemas debido a que modeló con distribución Poisson, en cambio para la severidad realizó el modeló con una distribución Gamma que es para valores positivos, los estados o carrocerías con siniestro cero se agruparon con los estados y carrocerías con siniestros ocurridos más bajos formando grupos con más estados y carrocerías y disminuyendo el número de grupos. Se obtuvo contribuciones considerables, debido a que la prima de riesgo en seguros de automóviles es muy importante para el sector asegurador, se estableció una metodología completa con métodos estadísticos multivariados que pueden ser utilizados como un parámetro de comparación a los resultados que presentan cada una de las compañías aseguradoras, con un buen cálculo de esta prima les permitirá cumplir con las demandas pactadas a las que sean sujetas sin tener que elevar sus reservas, por otro lado este cálculo impactaría directamente en la tarifa comercial ofertando tarifas competitivas dentro de todo sector asegurador, otra contribución fueron los intervalos de confianza para las estimaciones de las primas, este intervalo permite un margen de error con el cual se podría establecer las reservas necesarias y obtener una solvencia suficiente para cumplir con mayor eficiencia y sin problemas los posible siniestros que se presenten.

(Moura e Moura, 2017) en su estudio titulado: “Cálculo de la Prima Pura en un seguro de automóvil para la garantía de Daños Propios, mediante Modelos Lineales Generalizados y Segmentación de Clientes por Conglomerados” encontró que sus datos estudiados presentaron el problema de inflado en ceros, lo cual al estimar la frecuencia de siniestros, la diferenciación entre declarar cero siniestro o más es muy relevante. En su análisis, dicha estimación presenta diferencias significativas con los valores reales en los cuatro modelos utilizados. Se debe posiblemente a que se tendría que haber realizado un estudio más exhaustivo aplicando una distribución de probabilidad para frecuencia de ceros, otra para los valores intermedios, y por último una para los valores más elevados, obteniendo así un

modelo compuesto. En lo que respecta a la prima pura de los clústeres estudiados, los resultados se ajustaron correctamente a la realidad, debido en gran medida a que se utilizaron valores observados de las cuantías pagadas por los siniestros. Por último menciona que ese tipo de procedimiento para el cálculo de prima pura de la garantía de Daños Propios, puede ajustarse como un método de tarificación adecuado para una compañía aseguradora pequeña, pues una gran empresa de seguros necesitaría modelos mucho más elaborados.

La distribución Tweedie se define como una función de distribución Poisson compuesta con una Gamma, que por un lado asigna una probabilidad al cero igual a la probabilidad de Poisson de que no suceda ningún siniestro. De esta manera, acoge de forma natural la posibilidad de que una muestra tenga gran cantidad de ceros, sin necesidad de censurar una variable latente. Y por otro lado, establece una distribución para el resto de observaciones de la variable (Peña Sánchez, 2017).

La gran ventaja de elegir la distribución Tweedie es que estima la prima pura de riesgo tal como se visualiza en la Tabla 2, esto facilita temas de implementación. La desventaja es que se hace difícil controlar los componentes del siniestro como son, frecuencia y severidad.

Tabla 2: Principales variables objetivo de los Modelos Lineales Generalizados

Variable objetivo $Y=X/w$	Variable respuesta X	Peso w
Frecuencia	Número de siniestros	Exposición
Costo medio	Costo total	Número de siniestros
Prima pura de riesgo	Costo total	Exposición
Ratio de siniestralidad	Costo total	Prima imputada

Fuente: (Peña Sánchez, 2017) a partir de (Ohlsson & Johansson, 2010)

Este estudio se enfoca en la distribución Poisson compuesta, la cual es suma de una serie de variables aleatorias independientes e idénticamente distribuidas, donde el número se distribuye como función primaria Poisson, estas funciones se caracterizan por ser de tipo mixto y se pueden combinar con otras distribuciones secundarias para modelar la variabilidad de la variable respuesta, esta distribución es Poisson compuesta con una Gamma, siendo una distribución formada por una función primaria Poisson, discreta con probabilidad positiva en cero y otra secundaria Gamma de tipo continuo para el resto de

números reales positivos, esta distribución se enmarca dentro de los modelo Tweedie que tiene la siguiente forma (Peña Sánchez, 2017):

$$f(y) = \begin{cases} e^{-\lambda}, & y = 0 \\ e^{-\beta y} e^{-\lambda} \sum_{n=1}^{\infty} \frac{(\lambda \beta^\alpha)^n}{n! \Gamma(n\alpha)} y^{n\alpha-1}, & y > 0 \end{cases}$$

La distribución Tweedie pertenece a la familia exponencial de distribuciones, en ese sentido la función de densidad debe expresarse como:

$$f(y) = \exp\left\{\frac{y\theta - b\theta}{a(\phi)} + c(y, \phi)\right\}$$

Donde a , b y c son funciones arbitrarias, θ es el parámetro canónico de la distribución y ϕ el parámetro de escala. El modelo Tweedie se caracteriza por tener tres parámetros μ , ϕ y p , lo cuales tiene relación directa con las distribuciones Poisson y Gamma (Peña Sánchez, 2017):

$$\mu = \lambda \frac{\alpha}{\beta}, \quad p = \frac{\alpha + 2}{\alpha + 1}, \quad \phi = \frac{\lambda^{1-p}}{2-p} \left(\frac{\alpha}{\beta}\right)^{2-p}$$

Se pueden reparametrizar, considerando (λ, α, β) :

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \quad \alpha = \frac{2-p}{p-1}, \quad \frac{1}{\beta} = \phi(p-1)\mu^{p-1}$$

La suma depende de y , ϕ pero no de μ , entonces se puede definir la siguiente función $c(y, \phi)$ y $y > 0$

$$c(y, \phi) = \log \left[\sum_{n=1}^{\infty} \frac{(\lambda \beta^\alpha)^n}{n! \Gamma(n\alpha)} y^{n\alpha-1} \right], \quad c(0, \phi) = 1$$

$$f(y) = \exp\left\{-\beta y - \lambda + \log \left[\sum_{n=1}^{\infty} \frac{(\lambda \beta^\alpha)^n}{n! \Gamma(n\alpha)} y^{n\alpha-1} \right]\right\}$$

$$f(y) = \exp\{-\beta y - \lambda + c(y, \phi)\} = \exp\left\{\frac{-\beta \phi y - \lambda \phi}{\phi} + c(y, \phi)\right\}$$

El modelo Tweedie considera un conjunto de distribuciones de la familia exponencial, de tipo discreto, continuo o mixto, esta variedad se debe a que, adicionalmente a los parámetros usados en la familia exponencial, la distribución Tweedie incorpora un tercer parámetro p

denominado parámetro de potencia, en la Tabla 3 se presenta un resumen de los diferentes modelos Tweedie clasificados de acuerdo a parámetro p y la variable objetivo (Peña Sánchez, 2017):

Tabla 3: Resumen de los modelos Tweedie

p	Distribución	Tipo de variable	Variable objetivo
$p = 0$	Normal	Continua	-
$0 < p < 1$	-	Inexistente	-
$p = 1$	Poisson	Discreta	Frecuencia siniestro
$1 < p < 2$	Poisson compuesta	Mixta, no negativa	Prima pura
$p = 2$	Gamma	Continua, Positiva	Cuantía siniestro
$p = 3$	Normal inversa	Continua, positiva	Cuantía siniestro

Fuente: (Peña Sánchez, 2017) a partir de (Ohlsson & Johansson, 2010)

Muchas pólizas permiten múltiples siniestros. Si c es el número de siniestros de una póliza y z_1, \dots, z_c los montos individuales de cada siniestro. Entonces el monto total de siniestros es $y = z_1 + \dots + z_c$, con $y = 0$ si $c = 0$. Si c es Poisson y cada uno de los z_j son variables aleatorias independientes Gamma entonces y tiene distribución Tweedie. Esta distribución tiene una probabilidad distinta de cero cuando $y = 0$ igual a la probabilidad de Poisson cuando no se tiene siniestros. El resto de la distribución es similar a la Gamma. La distribución Tweedie pertenece a la familia exponencial en el cual $Var(y) = \phi\mu^p$, donde $1 < p < 2$ y permite estimar la prima pura de riesgo (De Jong & Heller, 2008).

En un modelo Lineal Generalizado la función de enlace $g(\mu_i)$ forma parte del modelo, su elección depende que sea una función diferenciable y estrictamente monótona. Cada distribución de la familia exponencial lleva asociada su función de enlace como se aprecia en la Tabla 4 (Peña Sánchez, 2017) :

Tabla 4: Funciones de enlace distribuciones de la familia exponencial

Distribución	Función de enlace	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Normal	Identidad	μ_i	η_i
Poisson/Tweedie	Logarítmica	$\ln(\mu_i)$	e^{η_i}
Binomial	Logit	$\ln(\mu_i/(1 - \mu_i))$	$e^{\eta_i}/(1 + e^{\eta_i})$
Gamma	Recíproca	$1/\mu_i$	$1/\eta_i$

Fuente: (Peña Sánchez, 2017) a partir de (Ohlsson & Johansson, 2010)

Jorgensen y de Souza (1994), Smyth y Jorgensen (2002) citados por (De Jong & Heller, 2008) han ajustado modelos Tweedie a datos de siniestros en seguros.

En la inferencia estadística se utilizó la prueba de Omnibus para determinar la significancia global del modelo, la prueba de Wald para las variables tomadas en cuenta.

4.2.3. Propuesta de alternativa de solución

A continuación se describe las etapas de desarrollo de este trabajo:

a. Fase de comprensión del negocio

En el Área de Riesgo Técnico para tarificar seguros se usaba un modelo denominado *InHouse* basado en un árbol de decisión para la venta de pólizas de clientes nuevos, este modelo presentaba, en la comparación de primas estimadas sobre primas observadas, descalibración de -11%, en el seguimiento realizado a noviembre 2017, lo cual mostraba primas insuficientes para soportar los siniestros del portafolio.

Hasta mayo 2016 las renovaciones de pólizas se emitían mediante la metodología *Bonus-Malus* que básicamente consistía en otorgar un descuento si el cliente no tuvo siniestros en el periodo anterior y adicionar un recargo si registró algún siniestro. El modelo Tweedie fue construido por una consultora usando data de los años 2012 y 2013, desde su implementación (mayo 2016) se usa para venta renovada, en el seguimiento se observó descalibración de +11% (seguimiento realizado noviembre 2017), las variables que usa este modelo como categorías de marca alta gama, marca japonesa, ya no se comportan de acuerdo a lo esperado por el modelo en el momento de su construcción, a causa de cambios en el parque automotor, en la actualidad existen automóviles más baratos y caros, el segmento de los vehículos utilitarios SUV, por sus siglas en inglés (*Standard Utility Vehicle*), ha ido ganado participación de mercado y el perfil del cliente que busca asegurar su bien ha cambiado. Debido a estos factores se solicitó calibrar el modelo Tweedie y actualizar la segmentación de variables para que puedan recoger el comportamiento del mercado actual, no se pidió incluir nuevas variables para evitar retrasos en la implementación del modelo. Se esperó con la calibración del modelo usarlo para la venta nueva y renovaciones. Se

encargó una nueva estimación de coeficientes con población más actualizada que aquella con la que se construyó el modelo, de manera que recoja el comportamiento del mercado actual y se obtenga mejor desempeño de las variables.

A continuación se definen conceptos clave que apoyarán la lectura del presente trabajo:

Seguro.- Acuerdo que realiza una persona o empresa con una compañía con el fin de que si la otra parte sufre algún daño y este se encuentra cubierto en el contrato de seguro, pueda ser indemnizado total o parcialmente.

Póliza.- Documento que detalla el contrato de seguro, en el cual se encuentran todas las coberturas pactadas entre la persona o empresa asegurada y la compañía de seguros.

Certificado.- Es un documento emitido por la compañía de seguros, verifica la existencia de una póliza de seguro, resume aspectos y condiciones clave de la póliza.

Vigencia.- Duración del periodo de tiempo previsto en la póliza del seguro durante el cual se activan sus coberturas.

Fecha Inicio de Vigencia.- Fecha en la cual inicia la cobertura de la póliza, periodo actual t.

Fecha Fin de Vigencia.- Fecha en la cual termina la cobertura de la póliza, periodo actual t.

Exposición.- Periodo de tiempo durante el cual un asegurado está expuesto a riesgo.

Contratante.- Persona o empresa que contrata el seguro, por el cual tiene que pagar una suma de dinero (prima) de manera periódica a la compañía aseguradora.

Asegurado.- Persona titular del contrato de seguro, se encuentra expuesto al riesgo.

Beneficiario.- Persona a indemnizar según las condiciones establecidas en el contrato de seguro, de ocurrir el evento.

Prima.- Monto de dinero, en dólares, abonado a la compañía por parte del asegurado para garantizar la cobertura del seguro.

Suma asegurada.- El valor atribuido a los bienes y el límite máximo de indemnización, está en dólares.

Bien asegurado.- Objeto sobre el que recae el contrato de seguro y en el que el asegurado tiene interés económico, para el caso desarrollado es un vehículo.

Cobertura.- Compromiso de la compañía aseguradora de pagar una indemnización al asegurado o beneficiario.

Siniestro.- Evento ocurrido de un riesgo contemplado en el contrato de seguro.

Monto de Reserva.- Gasto incurrido por la compañía para atender al asegurado, el monto está en dólares.

Causa de siniestro.- Motivo por el cual ocurrió el evento.

Fecha de cotización.- Fecha en la que se realiza al cliente la cotización de su seguro.

Cliente Nuevo.- Un cliente será clasificado como Nuevo, en caso haya una brecha de más de 4 meses entre la fecha fin de vigencia del periodo anterior y la fecha de cotización.

Cliente Renovado.- Aquello que no cumple con la definición de cliente nuevo.

Siniestro Pérdida Parcial.- Evento en el cual el daño ocurrido no involucra a todo el vehículo, siendo el monto de siniestro menor a la suma asegurada.

Siniestro Pérdida Total.- Evento en el cual el daño ocurrido involucra a todo el vehículo, en caso el monto de siniestro supere el 70% de la suma asegurada se considerará una pérdida total, se indemniza todo el valor de la suma asegurada.

Prima pura de riesgo.- Es el valor predicho por el modelo estadístico antes de aplicar estructura de costos, comisiones, impuesto general a las ventas (IGV) y derecho de emisión.

Frecuencia.- Suma de la cantidad de ocurrencia de siniestros sobre cantidad de asegurados.

Costo medio.- Es la suma de montos de reserva por siniestros sobre la cantidad de siniestros.

PMPY.- Proviene de las palabras en inglés, *Per member per year*, el cual tiene como forma de cálculo suma de los montos de reserva por siniestros sobre la cantidad de asegurados.

b. Fase de comprensión de los datos

La fuente de información que se suministró para el desarrollo del modelo proviene de una base de datos interna (Datamart) almacenadas en Oracle, el cual contiene información del cliente, vehículo, pólizas emitidas, y los siniestros correspondientes a estas pólizas, de manera mensual. En la Figura 1 se puede observar de manera resumida el diseño del Datamart de autos que contiene información del ramo de

Seguros Vehiculares:



Figura 1: Diagrama Datamart de Autos

De acuerdo con los objetivos del presente trabajo se estimaron dos modelos de regresión, separados por la causa del siniestro:

Modelo 1: Estima el precio para causa de Choque. Se tomaron en cuenta las siguientes causas de siniestro:

- Choque
- Volcadura
- Despiste
- Rotura de vidrios

$$Y_1 = \log(\mu_1) = \text{offset} + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_7 X_7$$

A continuación se presenta la definición de variables para el Modelo 1:

Y_1 : Suma monto de reserva de siniestros por causa Choque

X_1 : Variable categórica que indica si un cliente es nuevo

X_2 : Variable categórica que indica vehículos japoneses, gama media y baja

X_3 : Variable categórica vehículos de alta gama

X_4 : Logaritmo natural de suma asegurada del vehículo

X_5 : Logaritmo natural de edad del cliente

X_6 : Logaritmo natural cantidad de siniestros últimos 2 años

X_7 : Logaritmo natural de monto de reserva sobre suma asegurada, en el último año

Modelo 2: Estima el precio para causa de Robo. Se tomaron en cuenta las siguientes causas de siniestros:

- Robo de autopartes
- Robo total
- Daño por intento de robo

$$Y_2 = \log(\mu_2) = offset + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_5 X_5$$

Se presenta la definición de variables para el Modelo 2:

Y_1 : Suma monto de reserva de siniestros por causa Robo

X_1 : Variable categórica que indica si un cliente es nuevo

X_2 : Variable categórica indica vehículos con mayor riesgo de ser robados

X_3 : Logaritmo natural de suma asegurada del vehículo

X_4 : Logaritmo natural de edad del cliente

X_5 : Cantidad de siniestros por robo en los últimos 2 años

En ambos modelos se excluyeron asistencias vehiculares, ya que estos no se ingresan como monto de reserva, sino que vienen incluidos dentro de la cobertura de un seguro Todo Riesgo.

c. Fase de preparación de datos

En esta fase se realizó el manejo de datos que incluye la selección limpieza y transformación de información relevante para el estudio:

Se tomó información de pólizas anuales emitidas desde enero 2015 hasta septiembre 2017, de manera mensual, para la muestra de construcción se tomaron emisiones de enero 2015 a diciembre 2016, haciendo un total de 44,056 observaciones, y la muestra de validación emisiones de octubre 2016 a septiembre 2017, se tuvo 25,169 observaciones, con el fin de completar un año en esta muestra y así capturar comportamientos estacionales. Se optó por esta partición debido a que los Costos

medios de reparación presentaron tendencia decreciente en el año 2017 debido a gestiones y convenios con talleres de reparación, este efecto pudo haber causado sesgo en las estimaciones, al presentar un nivel más bajo de lo que se tuvo los años anteriores.

En la siguiente tabla se observan los indicadores de riesgo, para la muestra de construcción en el Modelo 1 de Choque se tiene frecuencia de siniestro de 24.65%, costo medio de \$ 1,553, para el Modelo 2 de Robo frecuencia de siniestro 2.13%, costo medio de \$ 2,653, este periodo tuvo suma asegurada promedio de \$ 18,565.

Tabla 5: Indicadores de riesgo muestra de Construcción

	Frecuencia	Costo medio	Q Siniestros	SA promedio	Total n
Modelo 1	24.65%	1,553	10858	18,565	44056
Modelo 2	2.13%	2,635	939		

A continuación en la Tabla 6 se observan los indicadores de riesgo, para la muestra de validación en el Modelo 1 de Choque se tiene frecuencia de siniestro de 23.69%, costo medio de \$ 1,629, para el Modelo 2 de Robo frecuencia de siniestro 2.21%, costo medio de \$ 2,402, este periodo tuvo suma asegurada promedio de \$ 18,834.

Tabla 6: Indicadores de riesgo muestra de Validación

	Frecuencia	Costo medio	Q Siniestros	SA promedio	Total n
Modelo 1	23.69%	1,629	5963	18,834	25169
Modelo 2	2.21%	2,402	555		

Se realizaron filtros a la base de datos de manera que se quedó con pólizas de vehículos livianos, de uso particular, considerando Lima y provincias, del canal de ventas Corredores, producto todo riesgo, se consideran los tipos de cliente persona y empresa, pólizas anuales, venta nueva y renovada, se excluyen motos por ser unidades vehiculares con un riesgo y comportamiento distinto al de la población definida. Para el caso de venta nueva se considera en caso haya brecha de más de 4 meses entre la última fecha fin de vigencia y la fecha de cotización del seguro, la búsqueda se realizó con la llave número de documento del cliente más número de chasis del vehículo.

– **Análisis Univariado:**

En la Tabla 7 se aprecia las medidas de tendencia central, dispersión y posición de las variables cuantitativas que se utilizaron en el Modelo 1, así como de la variable respuesta. Las variables con mayor coeficiente de variación fueron aquellas referentes a historia de siniestros, cantidad y monto sobre suma asegurada con 211.91% y 1073.05%, la variable respuesta monto de siniestros por causa choque presentó variabilidad de 476.07%, esto debido a que el ratio, observaciones que tuvieron siniestros sobre el total de la muestra, es de 19.9%, similar comportamiento ocurre en las variables de historia de siniestros.

Tabla 7: Análisis descriptivo variables cuantitativas Modelo 1

	X4	X5	X6	X7	Y1
VÁLIDO	53630	42086	53630	53630	53630
PERDIDOS	0	0	0	0	0
MEDIA	17856.02	49.91	0.32	0.01	344.67
MEDIANA	15111.58	48.00	0.00	0.00	0.00
MODA	18000.00	42.00	0.00	0.00	0.00
DESVIACIÓN ESTÁNDAR	11824.57	14.72	0.67	0.12	1640.88
VARIANZA	139820378.90	216.76	0.45	0.02	2692472.16
RANGO	159999.99	117.00	8.00	9.51	101475.15
MÍNIMO	0.01	-1.00	0.00	0.00	0.00
MÁXIMO	160000.00	116.00	8.00	9.51	101475.15
COEFICIENTE DE VARIACIÓN	66.22%	29.50%	211.91%	1073.05%	476.07%
PERCENTIL 1	2400	22	0.0	0.000	0.00
PERCENTIL 2	3100	24	0.0	0.000	0.00
PERCENTIL 3	3719	26	0.0	0.000	0.00
PERCENTIL 4	4320	27	0.0	0.000	0.00
PERCENTIL 5	4868	28	0.0	0.000	0.00
PERCENTIL 25	10263	39	0.0	0.000	0.00
PERCENTIL 75	22229	60	0.0	0.000	0.00
PERCENTIL 95	39990	75	2.0	0.059	1805.05
PERCENTIL 96	42253	77	2.0	0.072	2188.91
PERCENTIL 97	45038	78	2.0	0.095	2725.48
PERCENTIL 98	50000	80	2.0	0.129	3709.85
PERCENTIL 99	62000	84	3.0	0.198	6061.69

Suma asegurada: En el histograma de esta variable, se observa que el 97% de los datos se encuentran por debajo de los \$ 50,000, del diagrama de caja se aprecia

presencia de valores *outliers*, los cuales serán excluidos posteriormente.

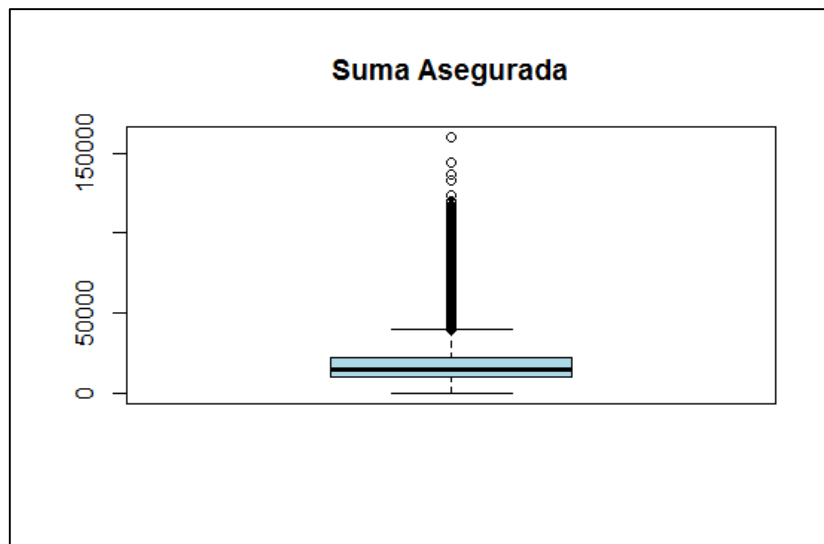
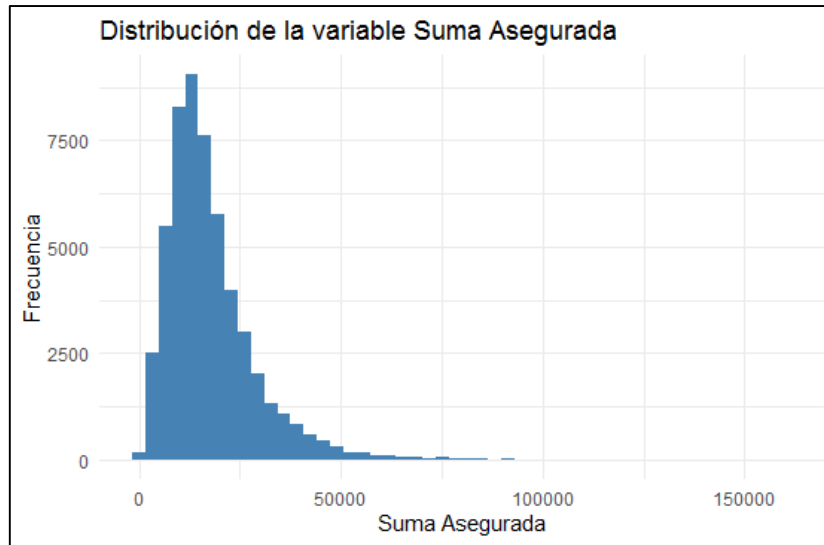


Figura 2: Histograma y diagrama de cajas Suma asegurada

Edad Cliente: En el histograma de esta variable se aprecia, concentración en un punto, estos clientes son los de tipo empresa, a los cuales se les asignó un valor constante de 32.6, posteriormente este valor será actualizado de acuerdo al comportamiento que presentó en la muestra de construcción.

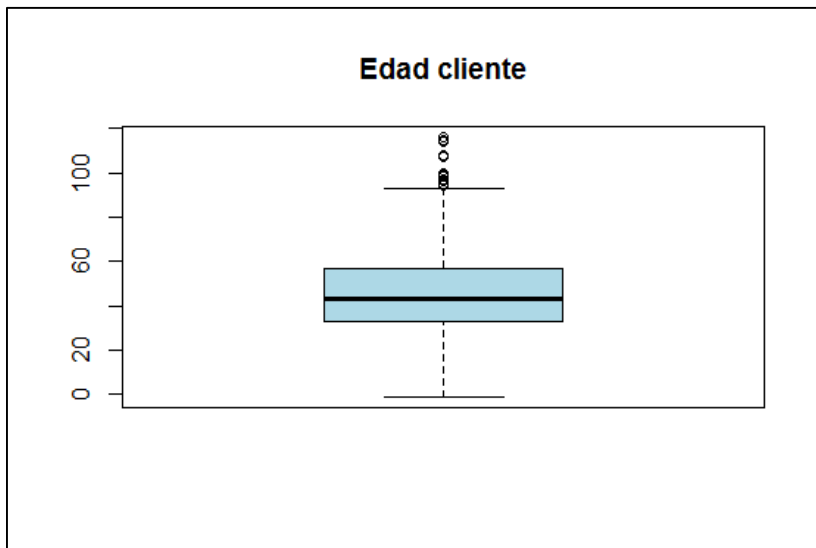
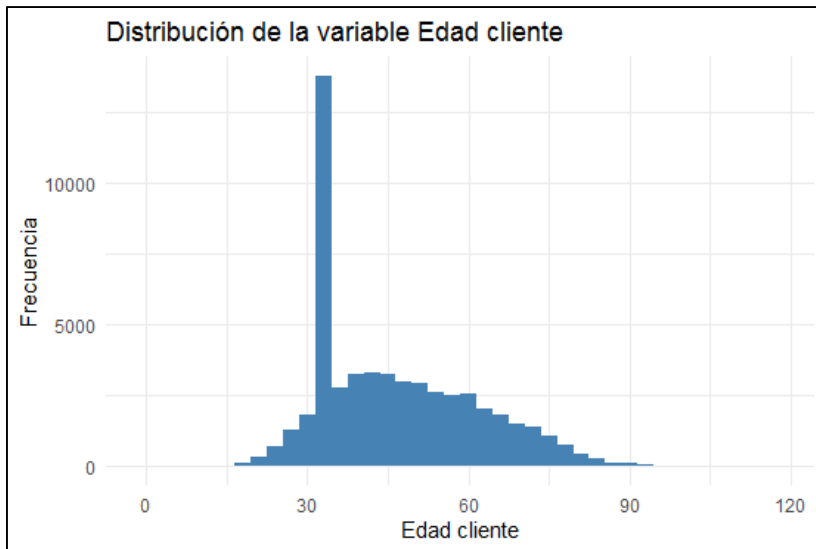


Figura 3: Histograma y diagrama de cajas Edad cliente

El histograma de la variable edad del cliente considerando solo personas, muestra composición de clientes maduros, el 75% tiene más de 39 años, en el diagrama de cajas se observa presencia de valores *outliers*.

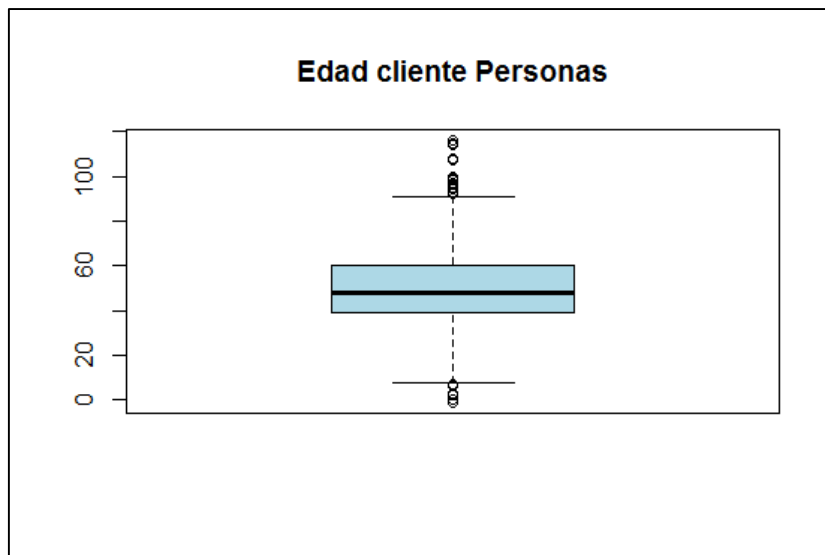
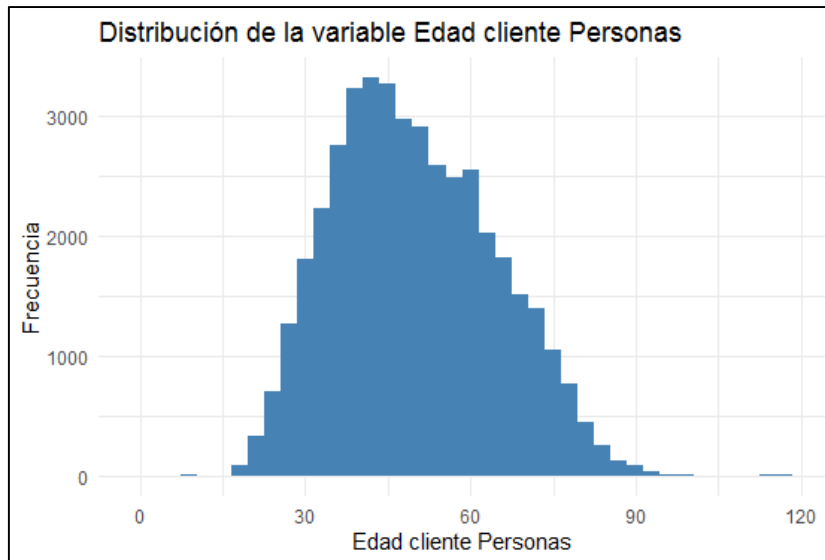


Figura 4: Histograma y diagrama de cajas Edad cliente solo Personas

Cantidad siniestros últimos dos años: En esta variable se aprecia mayoría con ceros, en este grupo se encuentran los clientes nuevos sin historia de siniestros y clientes que no se siniestraron en los últimos dos años.

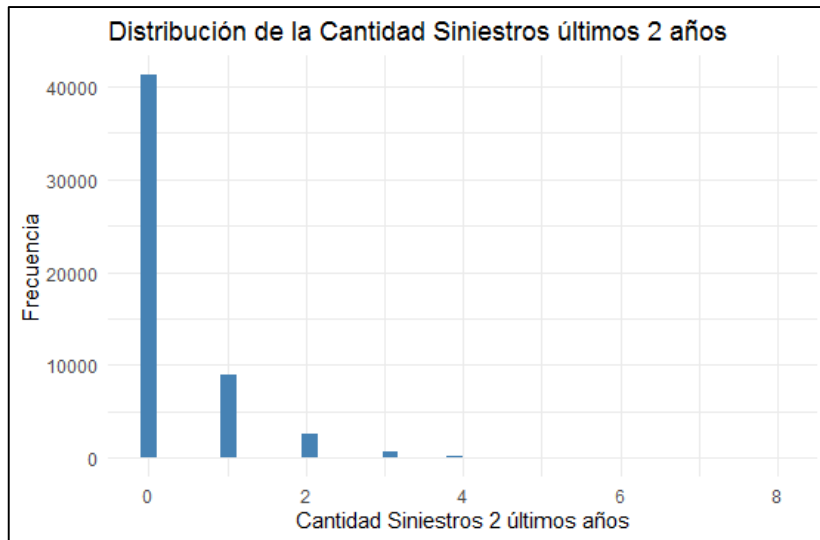


Figura 5: Gráfico de barras Cantidad de siniestros últimos 2 años

Logaritmo Monto de siniestro sobre suma asegurada: Esta variable presenta similar comportamiento a la anterior, ambas nos ayudarán a predecir siniestros en un futuro, gracias a su naturaleza histórica.

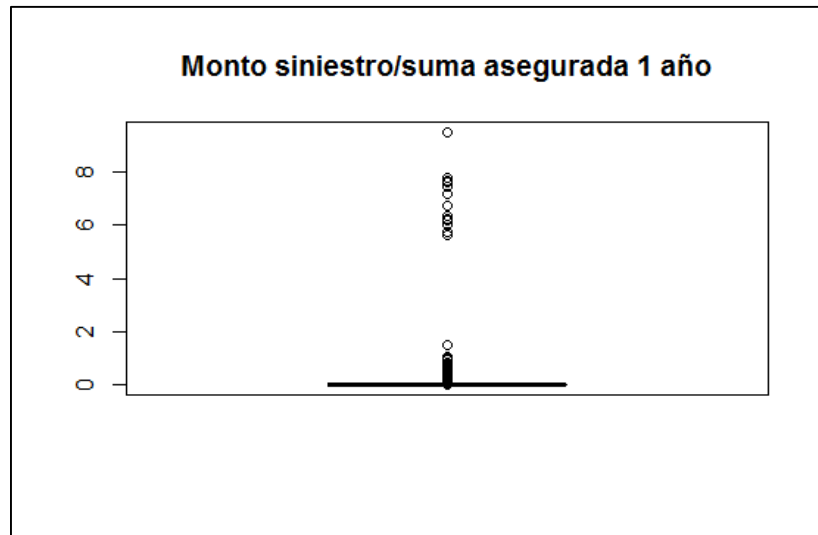
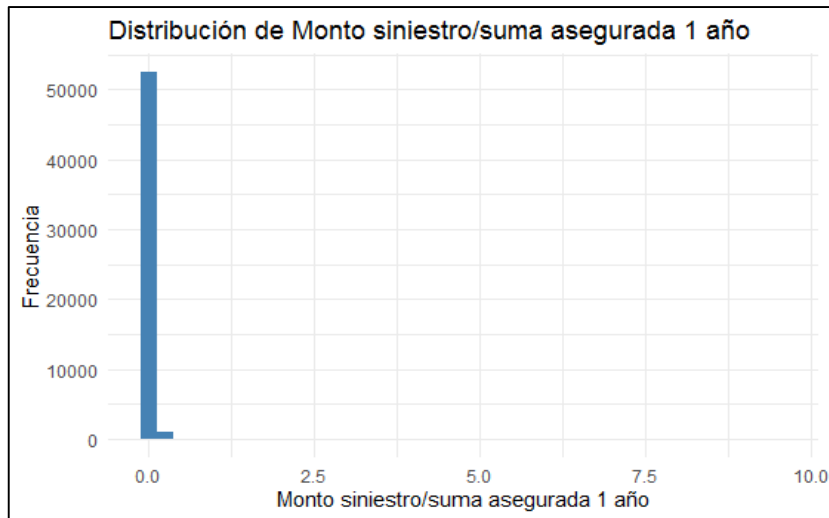


Figura 6: Histograma y diagrama de cajas Monto siniestro/Suma asegurada último año

Gráfico Pie variables cualitativas:

Respecto a las variables cualitativas en la Figura 7 se aprecia mayor cantidad de clientes Renovados con 70.4% de participación respecto de clientes Nuevos con 29.6% en la muestra de modelamiento.

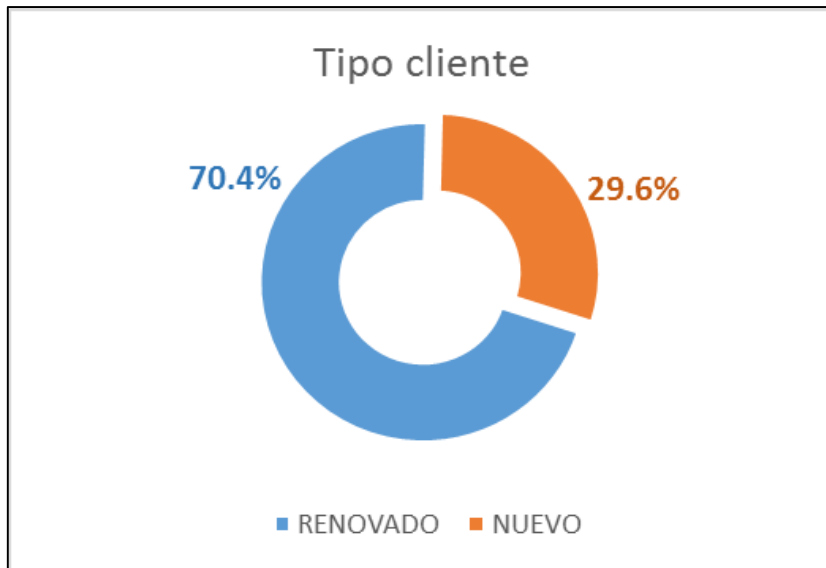


Figura 7: Gráfico de pie Tipo de cliente

En la Figura 8 las variable categórica Marca Japonesa posee 38.1% de participación, otras marcas 61.9% y la variable Marcas de Alta Gama representan 20.6% de la muestra de modelamiento, 70.4% otras marcas.

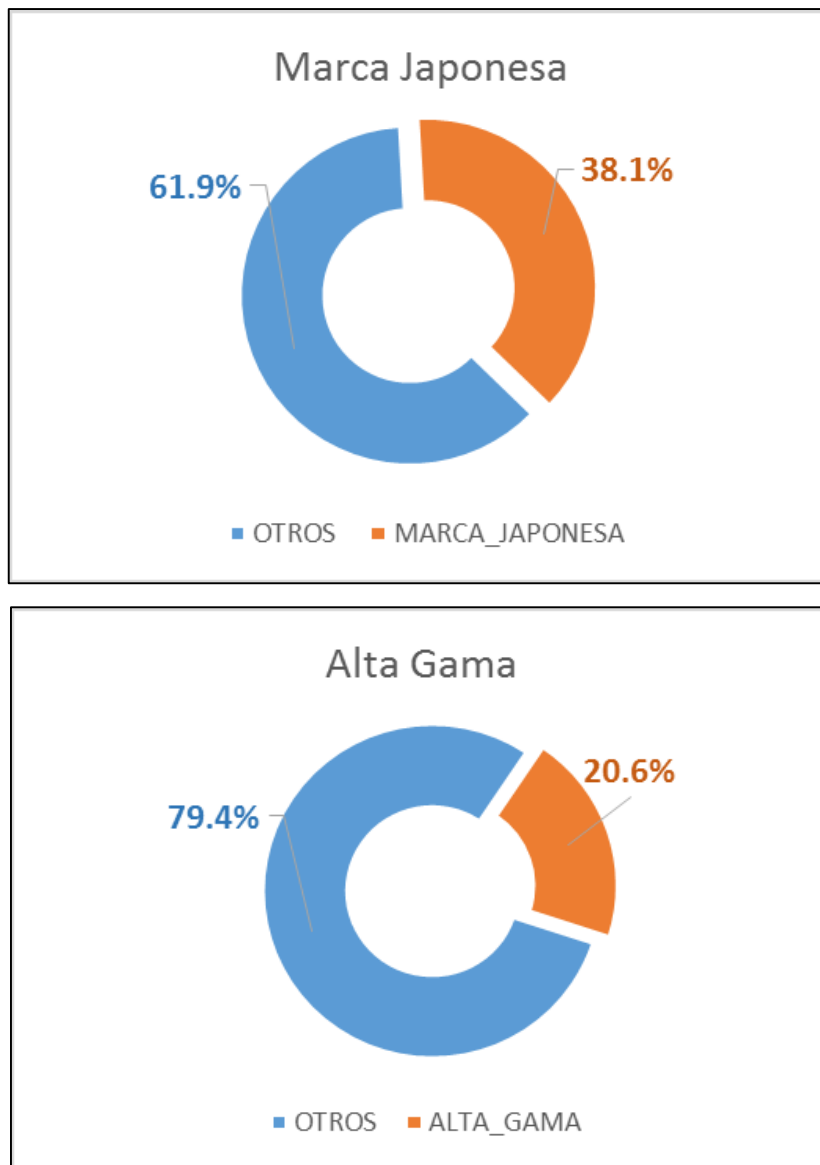


Figura 8: Gráfico de pie Marca japonesa, Alta gama

En la Tabla 8 se aprecia las medidas de tendencia central, dispersión y posición de las variables cuantitativas que se utilizaron en el Modelo 2, así como de la variable respuesta.

La variable independiente con mayor coeficiente de variación fue cantidad de siniestros por causa robo en los dos últimos años 700.30%, la variable respuesta monto de siniestros por causa robo presentó variabilidad de 1718.55%, esto debido a que el ratio, observaciones que tuvieron siniestros sobre el total de la muestra, es de 2.3%, por esta causa se encuentran más siniestros de pérdida total comparado con las pérdidas parciales lo cual origina una gran brecha entre una fracción de la suma

asegurada para algunas casos respecto del total de suma asegurada para otros casos que tuvieron siniestro.

Tabla 8: Análisis descriptivo variables cuantitativas Modelo 2

	X3	X4	X5	Y2
VÁLIDO	53630	42086	53630	53630
PERDIDOS	0	0	0	0
MEDIA	17856.02	49.91	0.02	53.75
MEDIANA	15111.58	48.00	0.00	0.00
MODA	18000.00	42.00	0.00	0.00
DESVIACIÓN ESTÁNDAR	11824.57	14.72	0.17	923.64
VARIANZA	139820378.90	216.76	0.03	853107.83
RANGO	159999.99	117.00	4.00	52549.77
MÍNIMO	0.01	-1.00	0.00	0.00
MÁXIMO	160000.00	116.00	4.00	52549.77
COEFICIENTE DE VARIACIÓN	66.22%	29.50%	700.30%	1718.55%
PERCENTIL 1	2400	22	0.0	0.00
PERCENTIL 2	3100	24	0.0	0.00
PERCENTIL 3	3719	26	0.0	0.00
PERCENTIL 4	4320	27	0.0	0.00
PERCENTIL 5	4868	28	0.0	0.00
PERCENTIL 25	10263	39	0.0	0.00
PERCENTIL 75	22229	60	0.0	0.00
PERCENTIL 95	39990	75	0.0	0.00
PERCENTIL 96	42253	77	0.0	0.00
PERCENTIL 97	45038	78	0.0	0.00
PERCENTIL 98	50000	80	1.0	273.76
PERCENTIL 99	62000	84	1.0	793.87

Cantidad siniestros últimos dos años causa robo: En esta variable se aprecia mayoría con ceros, debido a que es menos probable que te ocurra un robo parcial o robo total.

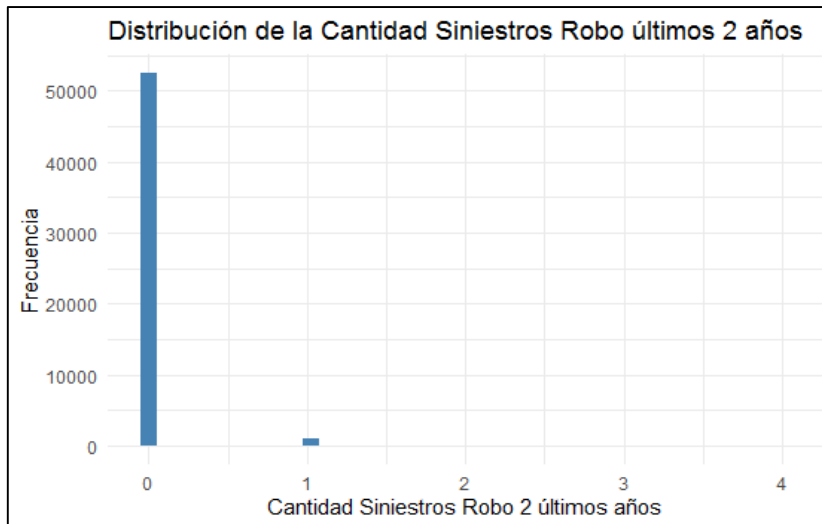


Figura 9: Gráfico de barras Cantidad de siniestros por robo últimos 2 años

Gráfico Pie variable cualitativa:

Respecto a la variables cualitativas se aprecia mayor cantidad de clientes con vehículo de marcas más robadas con 59.9% de participación, el resto de marcas representan el 40.1%.

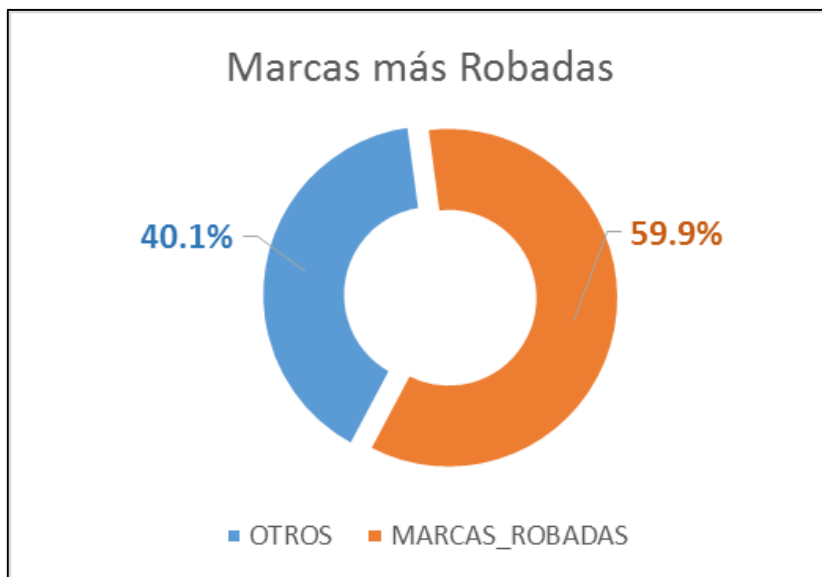


Figura 10: Gráfico de pie Marcas más robadas

– **Análisis de Valores Perdidos:**

Los datos obtenidos para el presente trabajo no presentan valores perdidos debido a que se trabajó en la extracción con bases de datos internas, las fuentes fueron cargadas desde los sistemas de la compañía hacia el Datamart que maneja el área.

– **Análisis de Valores Atípicos (*outliers*):**

Para estos valores se tomaron en cuenta percentiles 1, 5, 95 y 99, del análisis realizado para identificar a estos valores, se creó un campo el cual contiene los siguientes valores y criterios respectivos:

- 1:** Aquellas pólizas con sumas aseguradas menor a 500 USD
- 2:** Aquellas pólizas que tuvieron exposición menor a 6 meses y que no presentaron siniestros de pérdida total o parcial
- 3:** Aquellas pólizas que tuvieron monto de siniestro mayor a cero y menor a 50 USD
- 4:** Pólizas con clientes de edades menor a 18 años o mayor a 82 años
- 5:** Pólizas con sumas aseguradas mayor a 80,000 USD
- 6:** Vehículos con antigüedad menor a 0 años o mayor a 15 años
- 7:** Pólizas que presentaron en sus 3 vigencias anteriores a la vigencia en observación, como fecha fin de vigencia mayor a la fecha de inicio de vigencia del periodo en análisis
- 8:** Pólizas en las cuales el ratio suma del monto de siniestros sobre suma asegurada en un horizonte de tiempo de un año atrás, de la vigencia en análisis, es mayor a 1.5 veces la suma asegurada
- 9:** Pólizas con vehículos antiguos que no poseen coherencia con la suma asegurada registrada
- 10:** Pólizas óptimas que se utilizaron en el modelo de datos

Las observaciones marcadas del 1 al 9 fueron excluidas de la muestra de modelamiento según criterio por parte del Área Técnica y de Negocio.

– **Análisis Bivariado:**

A continuación se presenta el análisis bivariado del Modelo 1:

Suma asegurada: Esta variable presenta tendencia creciente en sus categorías, es decir a mayor suma asegurada mayor costo de prima, mostrando un salto cerca a los

\$ 800 de monto de reserva de siniestro por causa Choque para aquellos vehículos con suma asegurada por encima de los \$ 32,300.

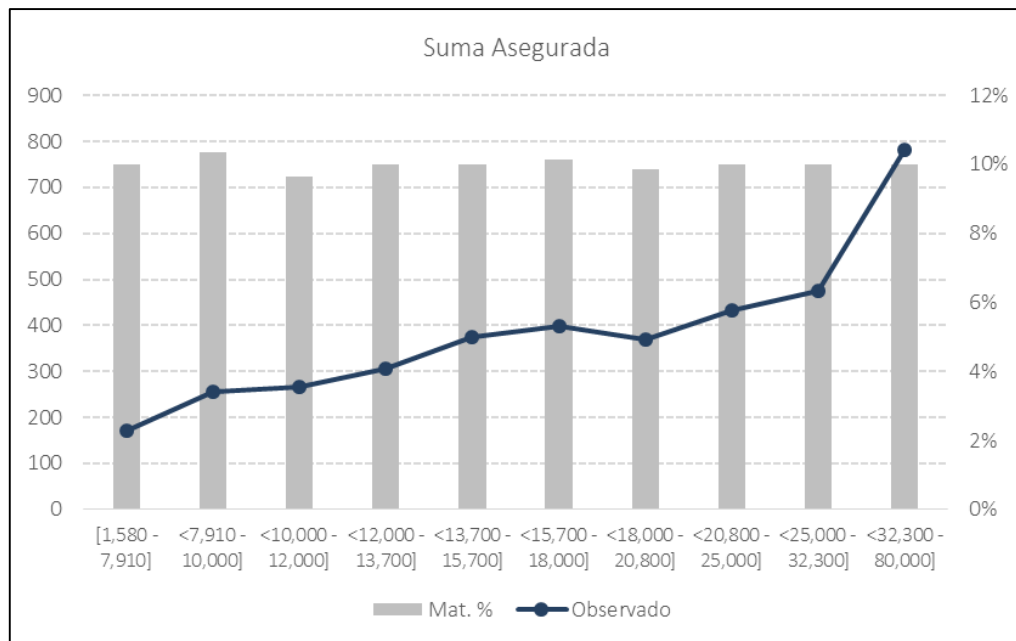


Figura 11: Gráfico bivariado Suma asegurada Modelo 1

Edad cliente: Esta variable presenta tendencia decreciente en sus categorías, es decir a mayor edad menor costo de prima, lo cual tiene relación con el conocimiento del negocio ya que se muestra que los jóvenes llegan a tener siniestros más severos. Se creó una categoría adicional para las empresas que muestra 32.6 años cuando el modelo fue construido, ahora ha cambiado su comportamiento y tiene similar comportamiento con aquellos clientes que se encuentra en el rango de <32.6 – 38] por lo que optaremos en actualizar el dato registrado de su edad tomando el límite superior de este rango, es decir 38 años.

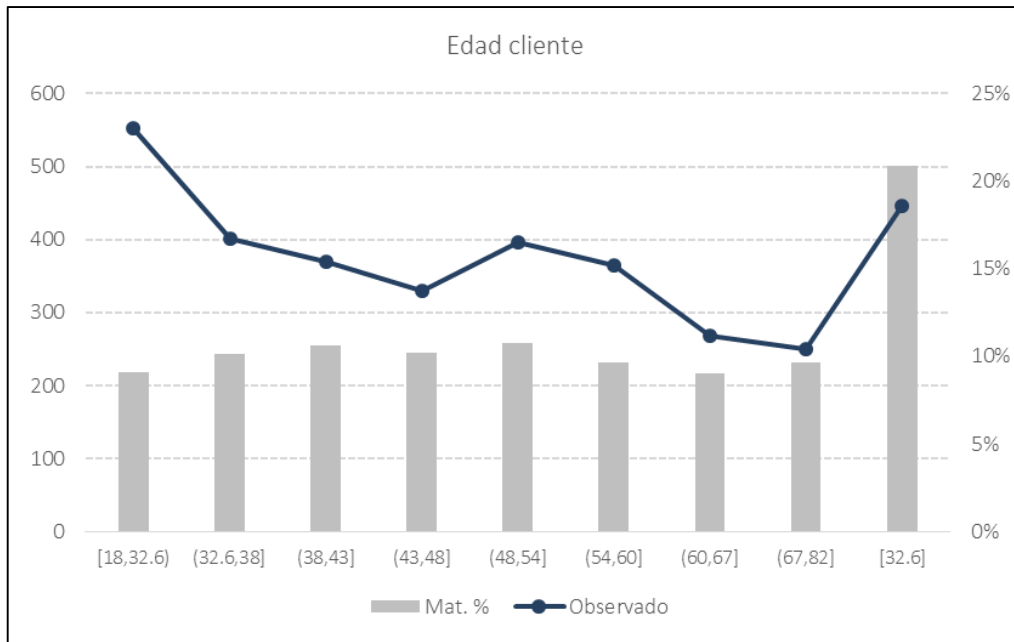


Figura 12: Gráfico bivariado Edad cliente Modelo 1

Tipo de cliente: Se observa que los clientes Nuevos tienen más riesgo que los clientes Renovados, esto debido a que un cliente Nuevo llega con un perfil poco conocido para la compañía, mientras que un cliente Renovado ya tiene un historial de comportamiento en la compañía y el programa de descuentos por renovación hace que tengan un mejor perfil de riesgo.

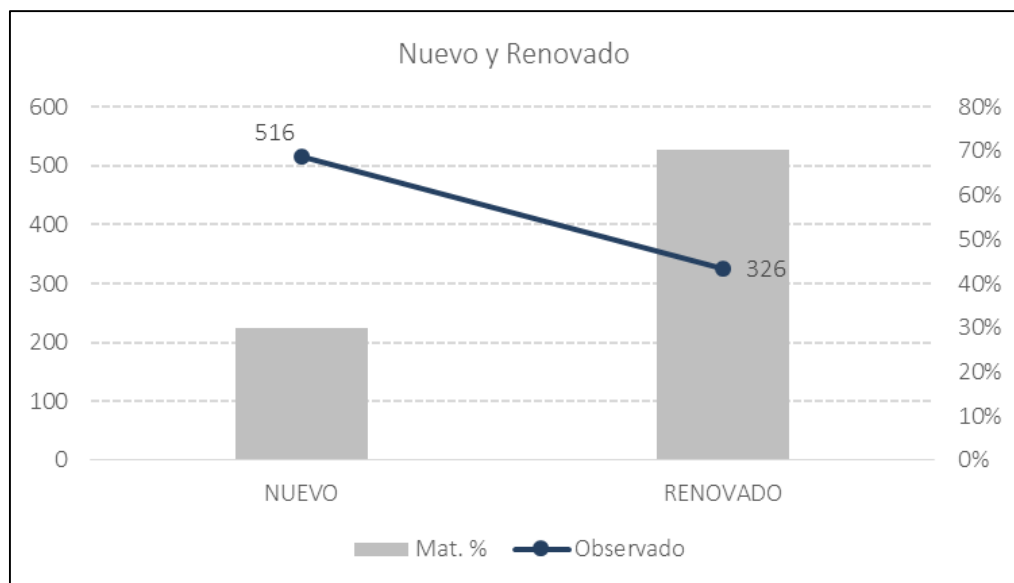


Figura 13: Gráfico bivariado Tipo cliente Modelo 1

Marca japonesa: De la Figura 13 se observa tendencia decreciente a las marcas presentes en este grupo respecto del resto de marcas, este comportamiento se relaciona con la cantidad de observaciones que hay por cada marca, así en Toyota se observa menor monto de reserva de siniestro por causa Choque y representa casi el 20% de participación en la muestra.

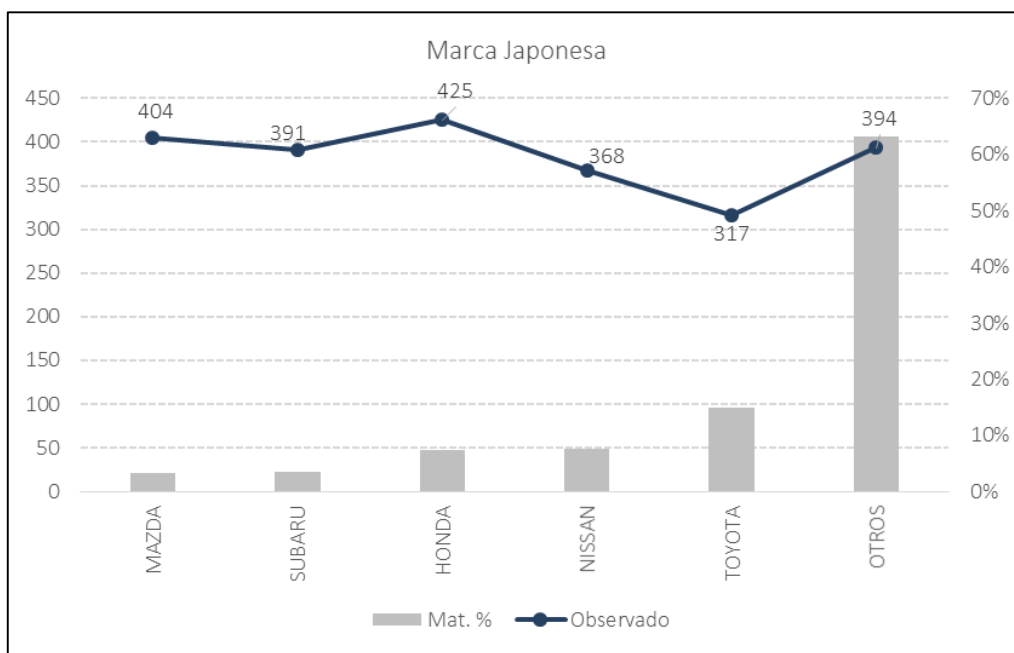


Figura 14: Gráfico bivariado Marca japonesa Modelo 1

Marca alta gama: Esta variable presenta mayor monto de reserva de siniestro por causa Choque debido a sus características, suma asegurada, precio de repuestos y atención en talleres. Por ejemplo salta a la vista la marca Jaguar debido a su baja participación en el mercado y a sus características ya que el chasis de los vehículos de esta marca está construido con piezas en fibra de carbono lo que hace ante un siniestro declararlo, generalmente como pérdida total.

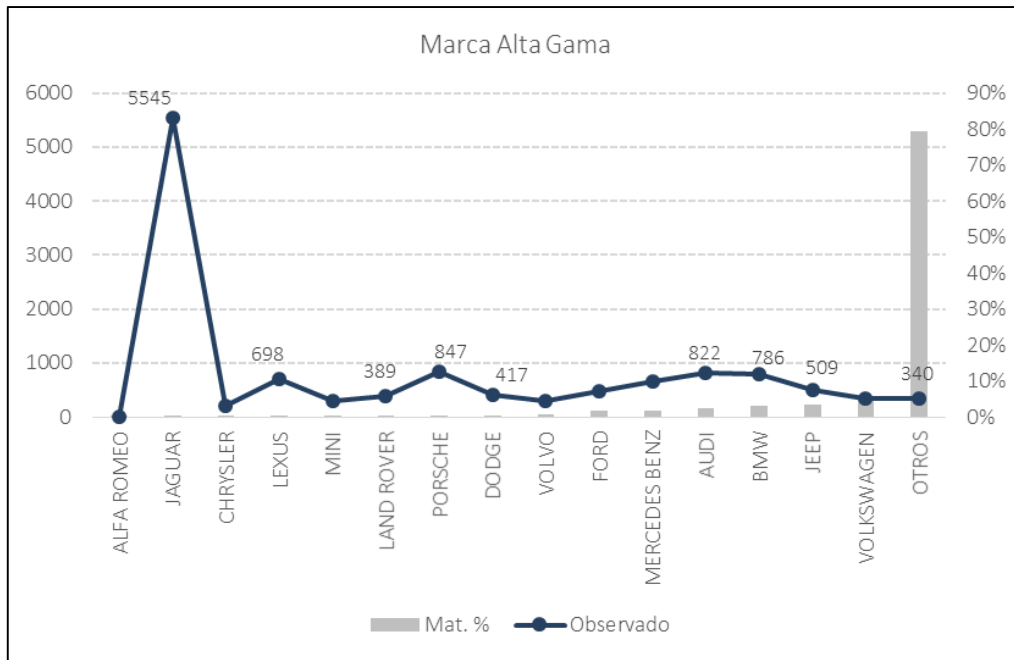


Figura 15: Gráfico bivariado Marca alta gama Modelo 1

A continuación se presenta el análisis bivariado del Modelo 2:

Suma asegurada: Esta variable presenta tendencia creciente en sus categorías, es decir a mayor suma asegurada mayor costo de prima, mostrando monto de reserva de siniestro por causa Robo cerca de \$ 100 para vehículos con suma asegurada por encima de los \$ 25,000.

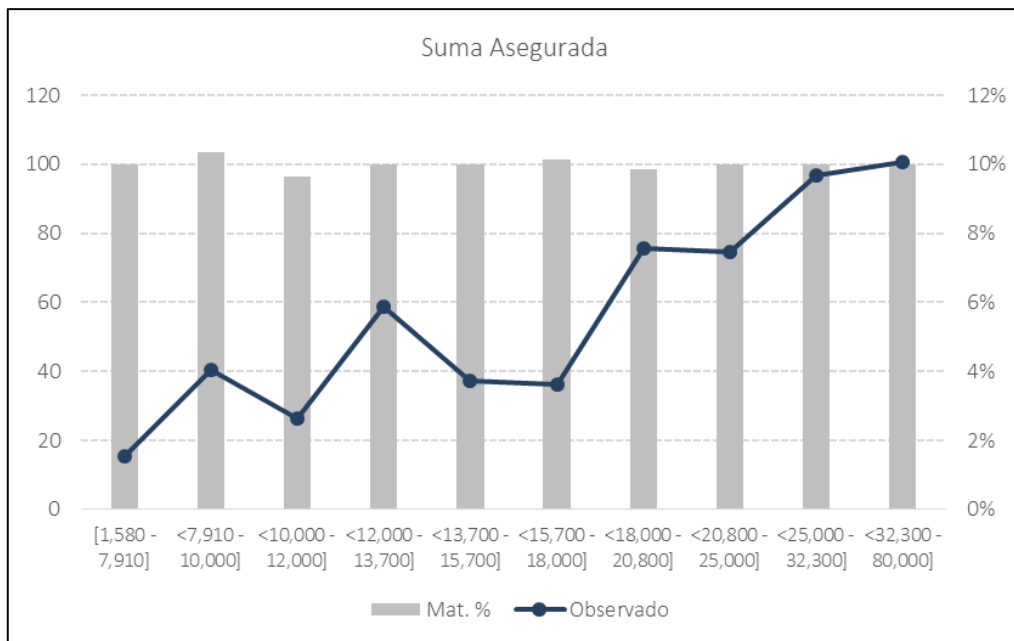


Figura 16: Gráfico bivariado Suma asegurada Modelo 2

Edad cliente: Esta variable presenta tendencia decreciente en sus categorías, es decir a mayor edad menor costo de prima. Se creó una categoría adicional para las empresas que muestra 32.6, ahora tiene comportamiento similar con aquellos clientes que se encuentra en el rango de <38 – 43] debido a que los montos de reserva de siniestro por causa Choque tienen más peso se optará por la edad ya definida de 38 años.

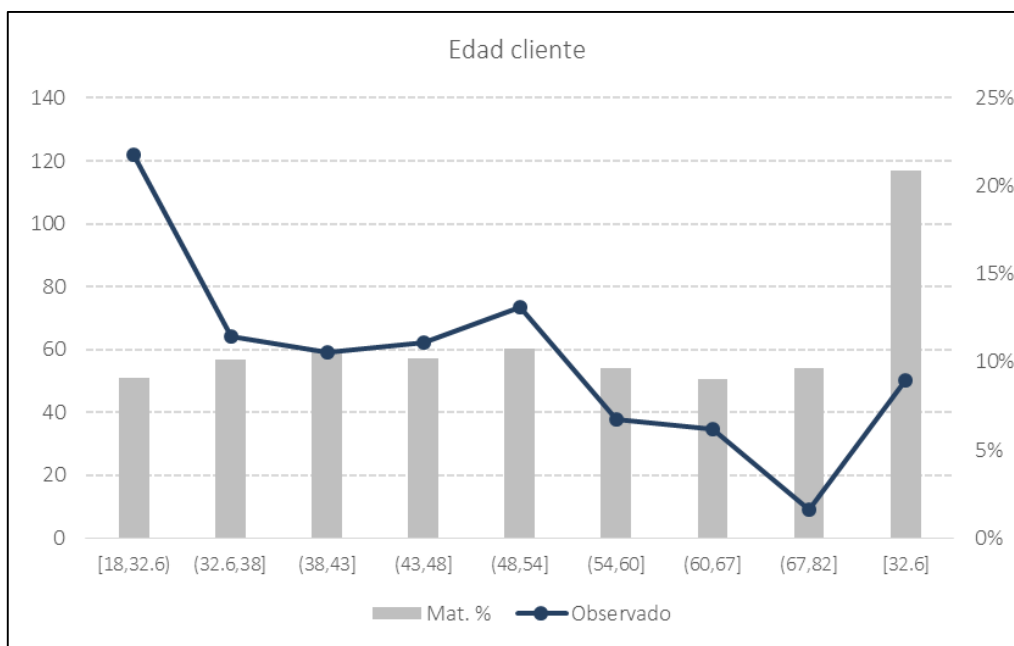


Figura 17: Gráfico bivariado Edad cliente Modelo 2

Marca más robada: Las marcas de vehículos presentes en este grupo presentan mayor monto de reserva de siniestro por causa Robo en comparación con el resto de marcas, a excepción de Nissan. Llama la atención el monto alto de BMW (\$ 176), esto se debe a que en los últimos años esta marca se ha estado robando con más frecuencia para usarla en actos delictivos.

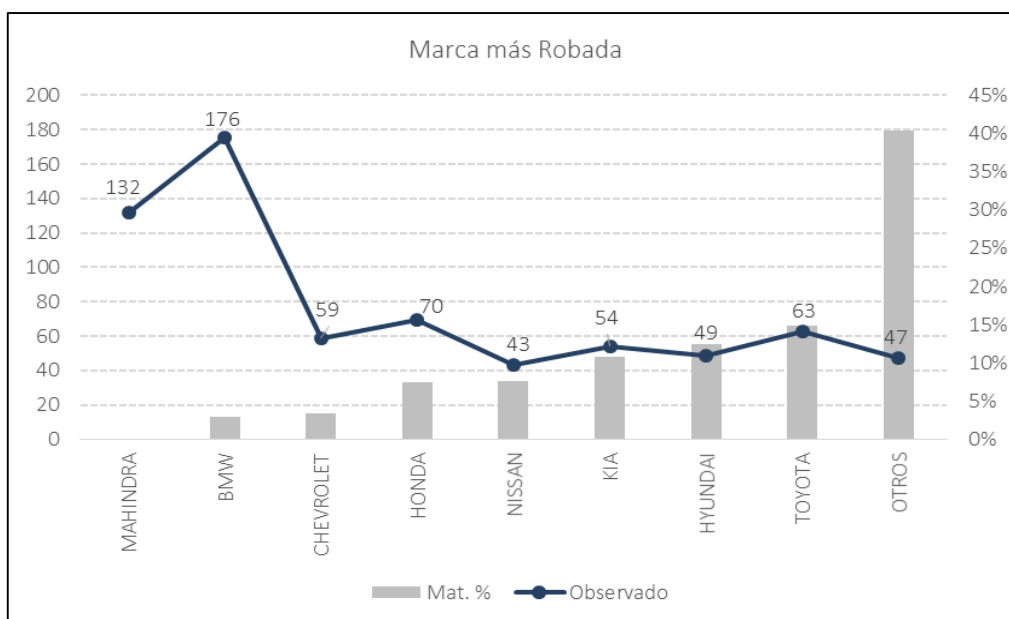


Figura 18: Gráfico bivariado Marcas más robadas Modelo 2

– **Transformación de variables:**

Para el Modelo 1 las variables independientes que se transformaron fueron la Suma asegurada, Edad cliente, Cantidad de siniestros en los últimos 2 años, se tomó el logaritmo de cada una de estas variables para estabilizar el sesgo y fue con esa transformación que ingresaron al modelo.

Para el Modelo 2 las variables independientes que se transformaron fueron la Suma asegurada y Edad cliente, se realizó la transformación logarítmica antes de que ingresen al modelo.

– **Análisis de Correlaciones:**

En la Tabla 9 se muestra la matriz de correlaciones para el Modelo 1, no se observa valores por encima de 0.6 en valor absoluto, por lo tanto no existe correlación entre las variables.

Tabla 9: Matriz de correlaciones Modelo 1

	X4	X5	X6	X7	Y1
X4	1.0	-0.140	0.027	-0.005	0.099
X5	-0.140	1.0	-0.058	-0.012	-0.044
X6	0.027	-0.058	1.0	0.168	0.047
X7	-0.005	-0.012	0.168	1.0	0.005
Y1	0.099	-0.044	0.047	0.005	1.0

En la Tabla 10 se muestra la matriz de correlaciones para el Modelo 2, no se observa valores por encima de 0.6 en valor absoluto, por lo cual no existe correlación entre las variables.

Tabla 10: Matriz de correlaciones Modelo 2

	X3	X4	X5	Y2
X3	1.0	-0.140	0.030	0.027
X4	-0.140	1.0	-0.042	-0.022
X5	0.030	-0.042	1.0	0.007
Y2	0.027	-0.022	0.007	1.0

d. Fase de modelado

Para la estimación de los modelos se usó el programa *IBM SPSS Statistics* versión 25. Para las variables categóricas se tomó como referencia la categoría de menor valor, para el parámetro de potencia p se consideró el valor de 1.5, este valor lo ofrece por defecto el programa *IBM SPSS Statistics*, con respecto a la función de enlace se tomó la función Logaritmo debido a que el modelo Tweedie es una función compuesta.

Se estimaron dos modelos estadísticos:

Modelo 1: Estima el monto de reserva de siniestros para causa Choque.

$$\hat{Y}_1 = \log(\hat{\mu}_1) = offset + \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \dots + \hat{\beta}_7 X_7$$

En la Tabla 11 se muestra las variables relevantes al Modelo 1:

Tabla 11: Coeficientes estimados Modelo 1

Variable	Estimador β	Peso
Intercepto	1.435	
X1: Variable categórica que indica si un cliente es nuevo	0.461	19.6%
X2: Variable categórica que indica vehículos japoneses, gama media y baja	0.128	1.5%
X3: Variable categórica vehículos de alta gama	0.185	2.5%
X4: Logaritmo natural de suma asegurada del vehículo	0.525	35.8%
X5: Logaritmo natural de edad del asegurado	-0.526	10.9%
X6: Logaritmo natural cantidad de siniestros últimos 2 años	0.579	28.3%
X7: Logaritmo natural de monto de reserva sobre suma asegurada, en el último año	0.446	1.5%

Modelo 2: Estima el monto de reserva de siniestros para causa Robo.

$$\hat{Y}_2 = \log(\hat{\mu}_2) = \text{offset} + \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \dots + \hat{\beta}_5 X_5$$

En la Tabla 12 se muestra las variables relevantes al Modelo 2:

Tabla 12: Coeficientes estimados Modelo 2

Variable	Estimador β	Peso
Intercepto	1.231	
X1: Variable categórica que indica si un cliente es nuevo	0.735	26.6%
X2: Variable categórica indica vehículos con mayor riesgo de ser robados	0.349	5.9%
X3: Logaritmo natural de suma asegurada del vehículo	0.648	22.8%
X4: Logaritmo natural de edad del asegurado	-1.353	34.5%
X5: Cantidad de siniestros por robo en los últimos 2 años	0.526	10.2%

e. Fase de evaluación

En esta parte se presenta procedimientos para validar la bondad de ajuste del modelo propuesto, mediante pruebas de hipótesis de los coeficientes, intervalos de confianza.

Modelo 1:

A continuación en la Tabla 13 podemos observar estadísticos de Bondad de ajuste del Modelo 1. El valor de la Devianza es un valor elevado debido a que la variable objetivo presenta gran cantidad de clientes que no presentaron siniestros (80.1%) esto hace que los valores observados (Monto reserva de siniestro causa Choque) y los valores estimados (prima) tengan grandes diferencias, lo cual no implica bajo ajuste del modelo, sino se puso mayor énfasis en la significancia del modelo estimado. El indicador AIC no se usó, pues solo se estimó un modelo para los montos de siniestro por causa Choque.

Tabla 13: Bondad de ajuste Modelo 1

Bondad de ajuste	Valor	gl	Valor/gl
Desviianza	4,029,146	44,048	91.47
Desviianza escalada	23,701	44,048	
Chi-cuadrado de Pearson	15,721,998	44,048	356.93
Logaritmo de la verosimilitud	-96,523		
Criterio de información Akaike (AIC)	193,064		
Criterio de información bayesiana (BIC)	193,142		

Prueba global del Modelo 1

$$H_0 : \beta_{1,\dots,7} = 0$$

$$H_1 : \text{Al menos un } \beta_{1,\dots,7} \neq 0$$

Tabla 14: Prueba Ómnibus Modelo 1

ChiSq de Verosimilitud	gl	Pr > ChiSq
1,245.86	7	0.00

Con un p-valor igual a 0% se puede afirmar que al menos uno de los coeficientes estimados es diferente a cero.

– **Significancia individual de los coeficientes estimados Modelo 1**

Para lo cual se han planteado 7 hipótesis, una por cada variable:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0 \text{ para } i = 1, \dots, 7$$

En la Tabla 15 se detalla el estadístico de Wald y el p-valor asociado a cada variable, debido a que todos los p-valores son menores a 5% y en algunos casos son iguales a cero se concluye estadísticamente que todos los coeficientes son diferentes a cero.

– **Interpretación de Odds ratio Modelo 1:**

El promedio del monto de reserva de siniestros por causa Choque aumenta 1.585 veces cuando el cliente es Nuevo, respecto a un cliente Renovado.

El promedio del monto de reserva de siniestros por causa Choque aumenta 1.136 veces cuando la marca del vehículo se encuentra en el grupo de Vehículos japoneses.

El promedio del monto de reserva de siniestros por causa Choque aumenta 1.203 veces cuando la marca del vehículo es de Alta gama.

El promedio del monto de reserva de siniestros por causa Choque aumenta 1.691 veces cuando el Logaritmo natural de Suma asegurada del vehículo aumenta en una unidad.

El promedio del monto de reserva de siniestros por causa Choque disminuye 0.591 veces cuando el Logaritmo natural de Edad cliente se adiciona una unidad.

El promedio del monto de reserva de siniestros por causa Choque aumenta 1.784 veces, por cada unidad adicional del Logaritmo natural de cantidad de siniestros últimos 2 años.

El promedio del monto de reserva de siniestros por causa Choque aumenta 1.562 veces, por cada unidad adicional del Logaritmo natural de monto de siniestros sobre suma asegurada último año.

Tabla 15: Análisis de los estimadores Modelo 1

Parámetros	Estimador β	Relatividad e^{β}	Límites de confianza 95%		ChiSq Wald	Pr > ChiSq	Peso
Intercepto	1.435	4.201	0.470	2.401	8.496	0.004	
X1	0.461	1.585	0.365	0.556	89.315	< 0.001	19.6%
X2	0.128	1.136	0.032	0.224	6.842	0.009	1.5%
X3	0.185	1.203	0.077	0.293	11.239	0.001	2.5%
X4	0.525	1.691	0.445	0.606	162.978	< 0.001	35.8%
X5	-0.526	0.591	-0.672	-0.379	49.585	< 0.001	10.9%
X6	0.579	1.784	0.479	0.679	128.733	< 0.001	28.3%
X7	0.446	1.562	0.112	0.781	6.836	0.009	1.5%

Modelo 2:

A continuación en la Tabla 16 se observa estadísticos de Bondad de ajuste del Modelo 2. El valor de la Devianza es un valor elevado debido a que la variable objetivo presenta gran cantidad de clientes que no presentaron siniestros (97.7%) esto hace que los valores observados (Monto reserva de siniestro causa Robo) y los valores estimados (prima) tengan grandes diferencias, lo cual no implica bajo ajuste del modelo, sino se puso mayor énfasis en la significancia del modelo estimado. El indicador AIC no se usó, pues solo se estimó un modelo para los montos de siniestro por causa Robo.

Tabla 16: Bondad de ajuste Modelo 2

Bondad de ajuste	Valor	Gl	Valor/gl
Desviación	2,163,113	44,050	49.11
Desviación escalada	3,607	44,050	
Chi-cuadrado de Pearson	72,771,188	44,050	1,652.01
Logaritmo de la verosimilitud	-13,648		
Criterio de información Akaike (AIC)	27,309		
Criterio de información bayesiana (BIC)	27,370		

Prueba global del Modelo 2

$$H_0 : \beta_{1,\dots,5} = 0$$

$$H_1 : \text{Al menos un } \beta_{1,\dots,5} \neq 0$$

Tabla 17: Prueba Ómnibus Modelo 2

ChiSq de Verosimilitud	gl	Pr > ChiSq
263.23	5	0.00

Con un p-valor igual a 0% se puede afirmar que al menos uno de los coeficientes estimados es diferente a cero.

– **Significancia individual de los coeficientes estimados Modelo 2**

Para lo cual se han planteado 5 hipótesis, una por cada variable:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0 \text{ para } i = 1, \dots, 5$$

En la Tabla 18 se detalla el estadístico de Wald y el p-valor asociado a cada variable, debido a que todos los p-valores son menores a 5% y en algunos casos son iguales a cero se concluye estadísticamente que todos los coeficientes son significativos.

– **Interpretación de Odds ratio Modelo 2:**

El promedio del monto de reserva de siniestros por causa Robo aumenta 2.085 veces cuando el cliente es Nuevo, respecto a un cliente Renovado.

El promedio del monto de reserva de siniestros por causa Robo aumenta 1.417 veces cuando la marca del vehículo se encuentra dentro del grupo Marcas más robadas.

El promedio del monto de reserva de siniestros por causa Robo aumenta 1.911 veces cuando el Logaritmo natural de Suma asegurada del vehículo aumenta en una unidad.

El promedio del monto de reserva de siniestros por causa Robo disminuye 0.259 veces cuando el Logaritmo natural de Edad cliente se adiciona una unidad.

El promedio del monto de reserva de siniestros por causa Robo aumenta 1.692 veces, por cada siniestro adicional de Cantidad de siniestros por robo en los últimos 2 años.

Tabla 18: Análisis de los estimadores Modelo 2

Parámetros	Estimador β	Relatividad e^{β}	Límites de confianza 95%		ChiSq Wald	Pr > ChiSq	Peso
Intercepto	1.231	3.425	-2.354	4.816	0.453	0.501	
X1	0.735	2.085	0.426	1.043	21.816	< 0.001	26.6%
X2	0.349	1.417	0.038	0.660	4.830	0.028	5.9%
X3	0.648	1.911	0.354	0.941	18.666	< 0.001	22.8%
X4	-1.353	0.259	-1.851	-0.854	28.324	< 0.001	34.5%
X5	0.526	1.692	0.170	0.882	8.405	0.004	10.2%

Los gráficos de elevación (*lift chart*), permiten comparar las estimaciones de prima pura de riesgo del modelo Tweedie, respecto a la variable observada monto de reserva de siniestros (Goldburd, Khare, & Tevet, 2016), primero se construyen grupos distribuidos de manera homogénea, luego se calcula el monto promedio de reserva de siniestros observado y el monto promedio de prima pura estimada.

En la siguiente figura, para la muestra de construcción (201501 - 201612), del modelo agregado, esto es Modelo 1 más Modelo 2, se observa buen ajuste del modelo, salvo una ligera sobreestimación en el noveno decil.

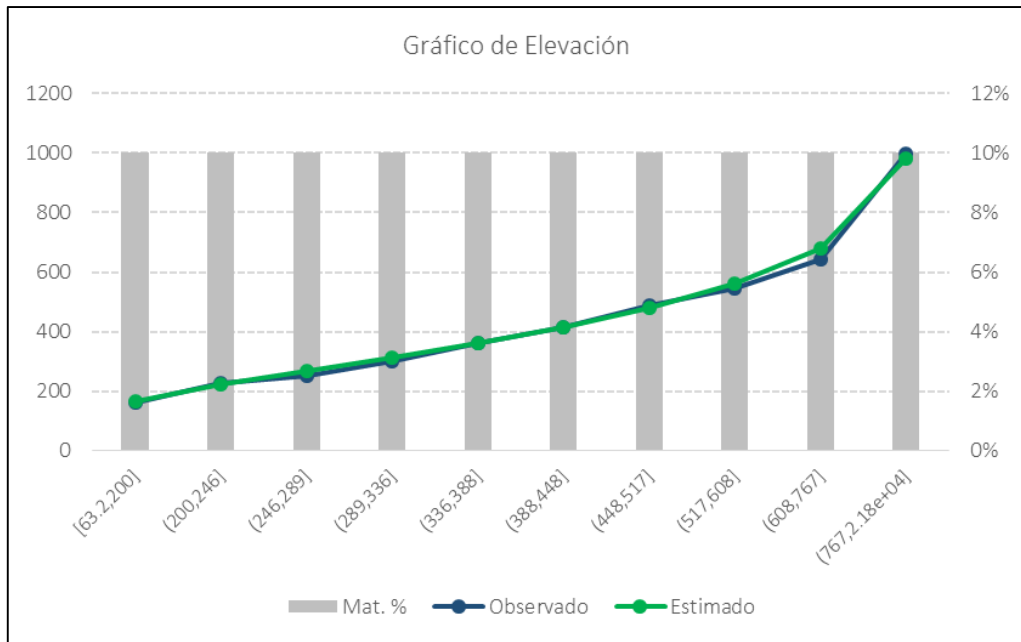


Figura 19: Gráfico de elevación modelo agregado, muestra construcción

En la Figura 20 se observan los gráficos de elevación de los Modelos 1 y 2 para la muestra de construcción, el Modelo 1 presentó buen ajuste con quiebre en el octavo decil, sin embargo el Modelo 2 presentó comportamiento más impredecible, esto debido a la baja cantidad de siniestros por robo, del cual el modelo no tuvo mucha historia para aprender.

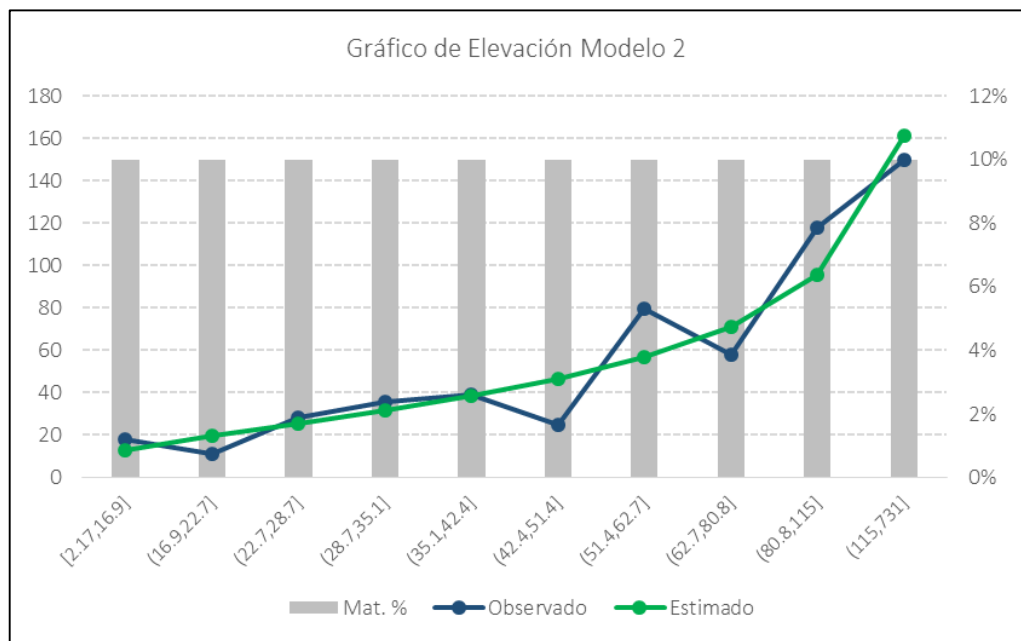
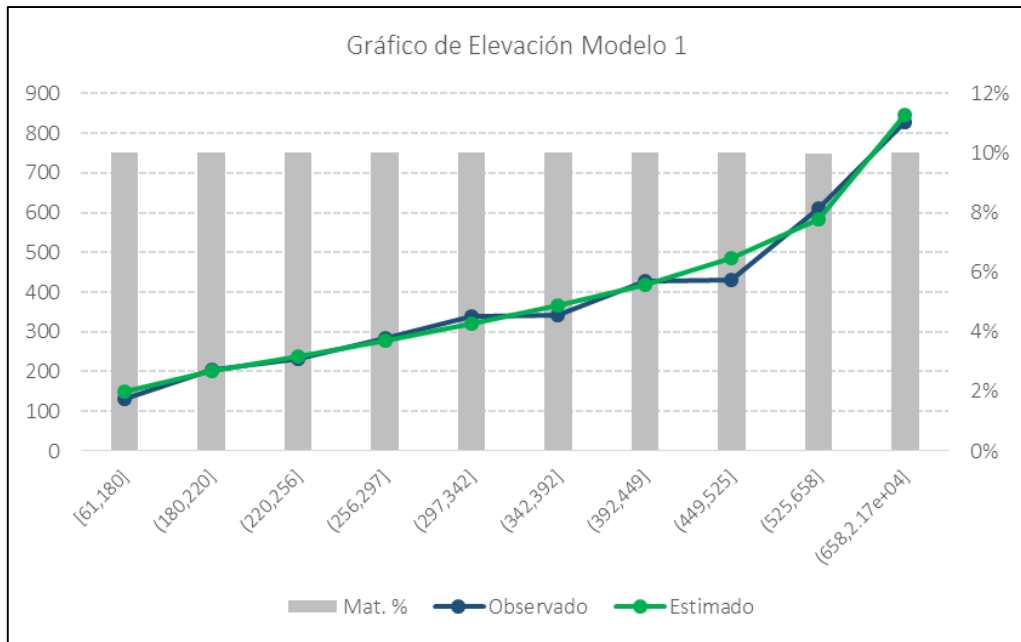


Figura 20: Gráfico de elevación Modelo 1 y Modelo 2, muestra construcción

En la Figura 21, para la muestra de validación (201610 - 201709), del modelo agregado, esto es Modelo 1 más Modelo 2, se observó sobreestimación en el primer, cuarto, quinto y sexto decil.

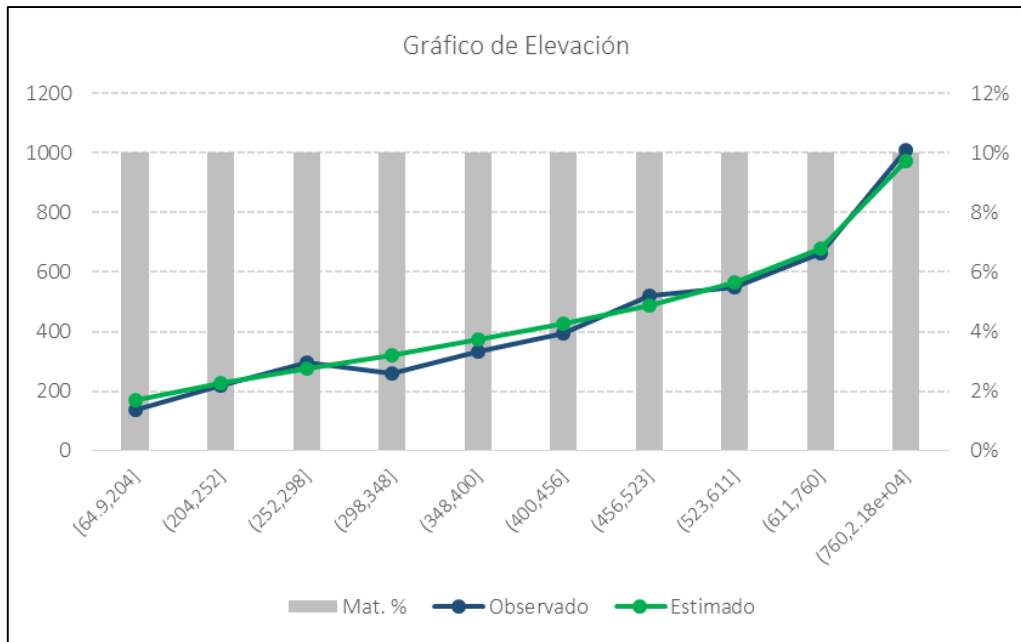


Figura 21: Gráfico de elevación modelo agregado, muestra validación

En la Figura 22 se observan los gráficos de elevación de los Modelos 1 y 2 para la muestra de validación, el Modelo 1 presentó subestimación en el último decil, el Modelo 2 presentó subestimación en el tercer y sexto decil, en el último decil este modelo presentó sobreestimación.

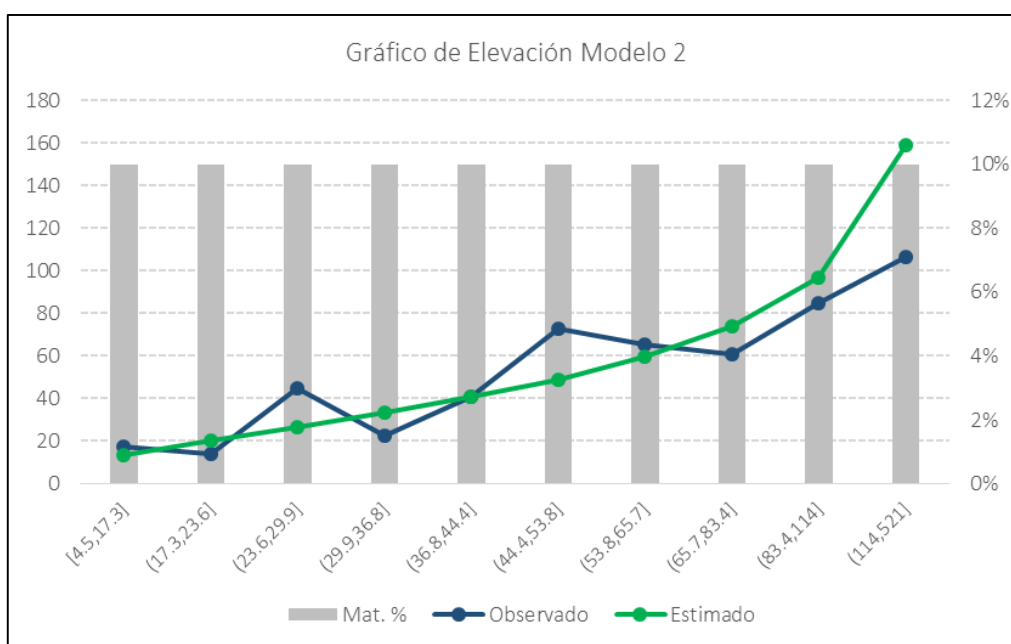
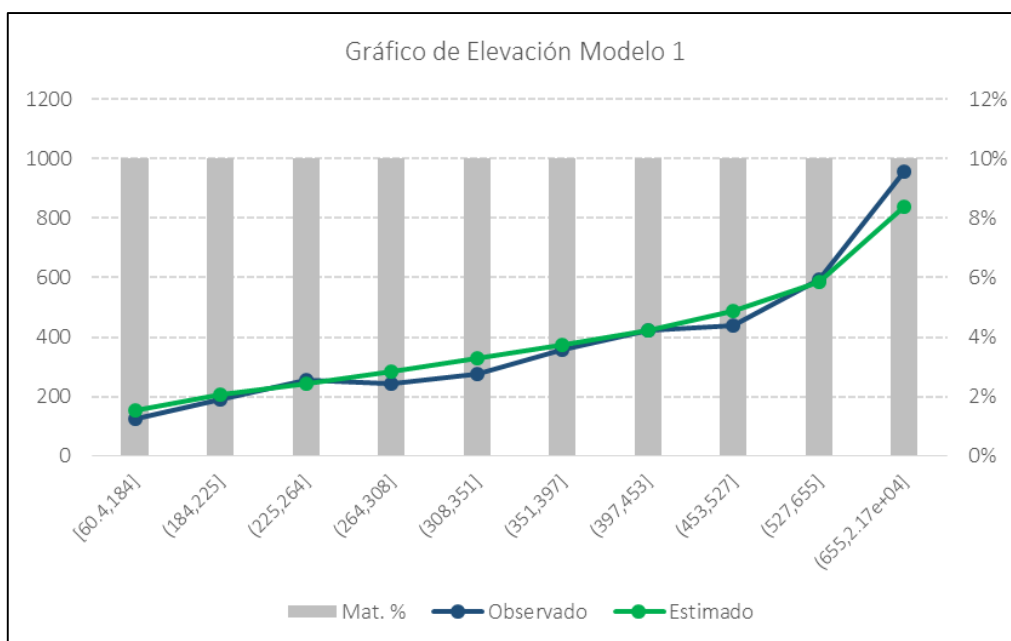


Figura 22: Gráfico de elevación Modelo 1 y Modelo 2, muestra validación

f. Fase de implementación

En esta fase se suministró al Área de Desarrollo de Tecnologías de Información los coeficientes estimados, las tablas de grupos marcas de vehículos, la nueva edad para Empresas, la lógica para contabilizar los siniestros por causa robo en los últimos 2 años, en libros de archivos hojas de cálculo (*Microsoft Excel*).

Esta información proporcionada fue cargada en los sistemas de la compañía por el Área de Desarrollo de Tecnologías de Información, los cuales se encargaron de realizar el plan, las pruebas y el proceso de implementación.

El Área Técnica participó como apoyo en la etapa de pruebas de desarrollo de implementación de los modelos de precios.

4.3. Contribución en la solución del problema

En el División Técnico de Negocios, se necesitaba actualizar un modelo de tarificación con el cual ya se estaba trabajando durante buen tiempo, sin haber realizado calibración de este pese a los indicadores desviados que presentaba.

En ese sentido se determinó necesario contar con un modelo que estime directamente el precio de la prima que un cliente paga por su seguro modelado con población más actual, es así como se eligió el Modelo Lineal Generalizado con función de distribución Tweedie, ya que este permite incluir en el modelamiento a toda la muestra en estudio, hayan tenido o no siniestros.

4.4. Análisis de la contribución en términos de competencias y habilidades

La sólida formación que ofrece el Departamento de Estadística e Informática a través de los cursos impartidos durante toda la carrera profesional ha facilitado la solución de problemas reales en el campo laboral donde me desempeño.

Los cursos Base de Datos, Sistemas Informáticos, correspondientes a la rama de Informática contribuyeron en todo lo que respecta a la gestión de base de datos para extraer información necesaria en el desarrollo de este modelo. La forma crítica de ver los datos, realizar análisis deductivo para examinar la información y el análisis exploratorio de datos se enmarca dentro de las competencias impartidas por el departamento académico. Para la etapa de modelamiento los cursos de Modelo Lineales, Modelos Lineales Generalizados, Técnicas Multivariadas, cimentaron las bases teóricas y prácticas necesarias para poder aplicarlos a un problema de negocios real; todo lo mencionado se complementa con el manejo de software especializado en el desarrollo de un proyecto herramientas informáticas como SQL,

Oracle para el manejo de base de datos, programación en Excel y uso de herramientas estadísticas que se fueron aprendiendo en los cursos de carrera llevados durante la etapa universitaria como SPSS, programación en R, fueron útiles para el modelamiento de datos. Adicionalmente los trabajos realizados en grupo que se hacían necesarios, fomentó el trabajo en equipo y desarrollo de habilidades blandas muy indispensables en el ámbito laboral.

4.5. Nivel de beneficio obtenido por el centro laboral

Con los resultados se permitió obtener un nivel de calibración, comparando los siniestros reales y lo estimado por el modelo, dentro de los límites permitidos por el negocio $\pm 5\%$. Se hizo la comparación utilizando el indicador PMPY (*Per member per year*), el cual tiene como forma de cálculo promedio de los montos de reserva por siniestros. Esto es para la muestra de construcción que comprende el periodo enero 2015 a diciembre 2016 un nivel de calibración que se puede observar en la Tabla 19 de 1.34%, y para la muestra de validación que comprende el periodo octubre 2016 a septiembre 2017, se muestra en la Tabla 20 un nivel de calibración de 2.47%, esto debido a que el costo medio para ese periodo presentó tendencia decreciente como efectos de las políticas establecidas por la compañía, tal es el caso de importación de repuestos por parte de la empresa y convenio con talleres Preferentes, los cuales tienen costos más bajos que los talleres de concesionarios.

Tabla 19: Nivel de calibración periodo de construcción

Modelo 1		Modelo 2		Total		
Pmpy observada	Pmpy estimada	Pmpy observada	Pmpy estimada	Pmpy observada	Pmpy estimada	% Var.
382.86	388.82	56.17	56.10	439.03	444.92	1.34%

Tabla 20: Nivel de calibración periodo de validación

Modelo 1		Modelo 2		Total		
Pmpy observada	Pmpy estimada	Pmpy observada	Pmpy estimada	Pmpy observada	Pmpy estimada	% Var.
385.93	392.52	52.97	57.23	438.89	449.75	2.47%

V. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

1. La prima pura de riesgo del modelo técnico se explica por el modelo lineal generalizado Tweedie propuesto.
2. Para la muestra de construcción del modelo, de la comparación Monto de reserva de siniestros total promedio observado, y el promedio estimado se obtiene un nivel de 1.34% dentro del límite de 5% de tolerancia por el negocio.
3. Para la muestra de validación del modelo, de la comparación Monto de reserva de siniestros total promedio observado, y el promedio estimado se obtiene un nivel de 2.47% dentro del límite de tolerancia al 5% por el negocio.
4. En el Modelo 1, que estima el precio por causa Choque, todas las variables resultaron significativas, siendo las tres variables más importantes el Logaritmo natural de la suma asegurada del vehículo con un nivel de importancia de 35.8%, seguido del Logaritmo natural de la cantidad de siniestros en los últimos 24 meses con 28.3% y la variable categórica que indica si un cliente es Nuevo con 19.6%.
5. En el Modelo 2, que estima el precio por causa Robo, todas las variables resultaron significativas, siendo las tres variables más importantes el Logaritmo natural de la edad del asegurado con nivel de importancia de 34.5%, seguido de la variable categórica que indica si un cliente es Nuevo con 26.6% y el Logaritmo natural de la suma asegurada del vehículo con 22.8%.
6. En el Modelo 1, se realizó la predicción para un cliente Renovado, que no tiene un vehículo en la lista de Marca japonesa, tampoco Alta gama, con Suma asegurada de \$ 18,000, con Edad de 42 años y que no tuvo siniestros en vigencias pasadas, obtuvo como prima pura de riesgo \$ 273.97.

7. En el Modelo 2, se realizó la predicción para el mismo cliente Renovado, que tiene un vehículo entre las Marcas más robadas, con Suma asegurada de \$ 18,000, Edad de 42 años y no tuvo siniestros por robo en vigencias pasadas, obtuvo como prima pura de riesgo \$ 48.04. En total este cliente obtuvo una prima pura de riesgo de \$ 322.01 que es resultado de la suma de Modelo 1 y Modelo 2.

5.2. Recomendaciones

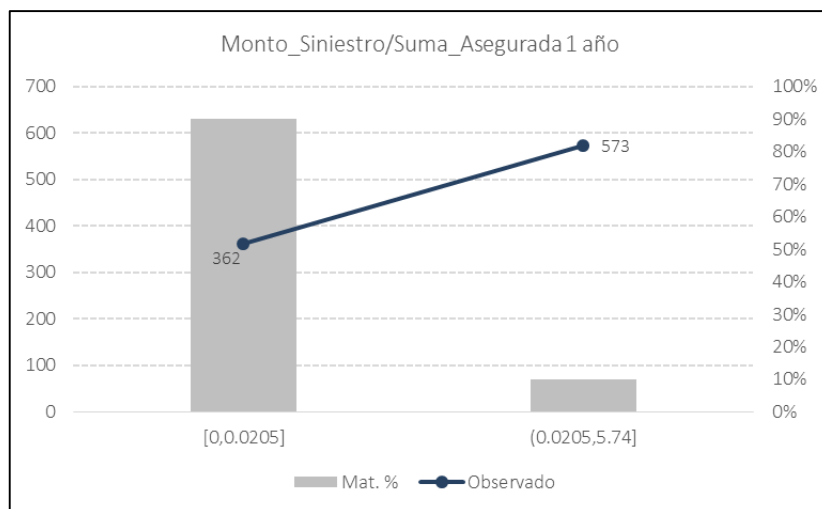
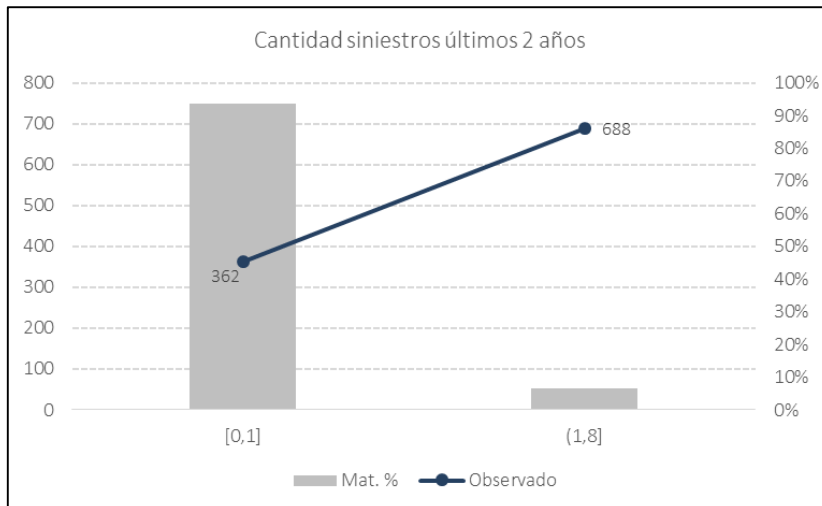
1. Para un control más sensible al riesgo de los clientes se recomienda estimar modelos separados por Frecuencia y Severidad.
2. Para recoger variables del comportamiento del conductor se recomienda obtener datos de los dispositivos de telemática que usan los vehículos.
3. Para obtener variables tales como el comportamiento financiero del cliente se recomienda usar bases de datos externas.
4. Para incluir nuevas fuentes de información, sin incurrir en costos adicionales, se recomienda implementar técnicas de *Web Scraping*, para extraer información de sitios web.

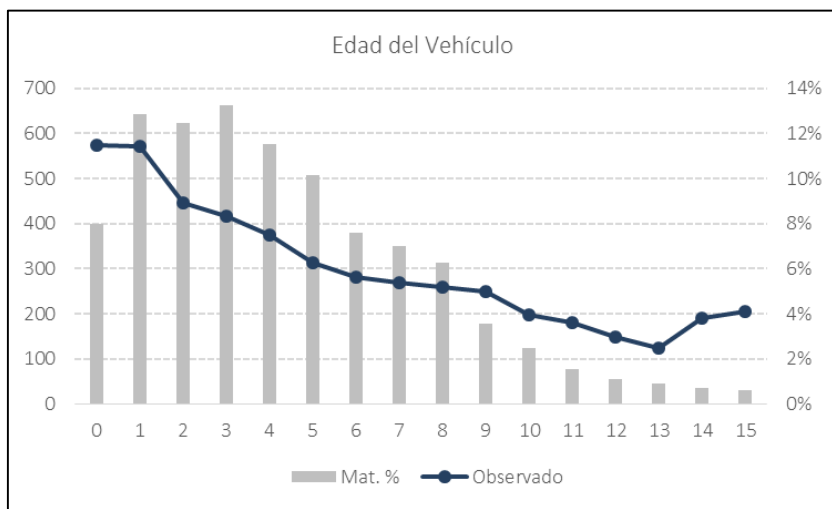
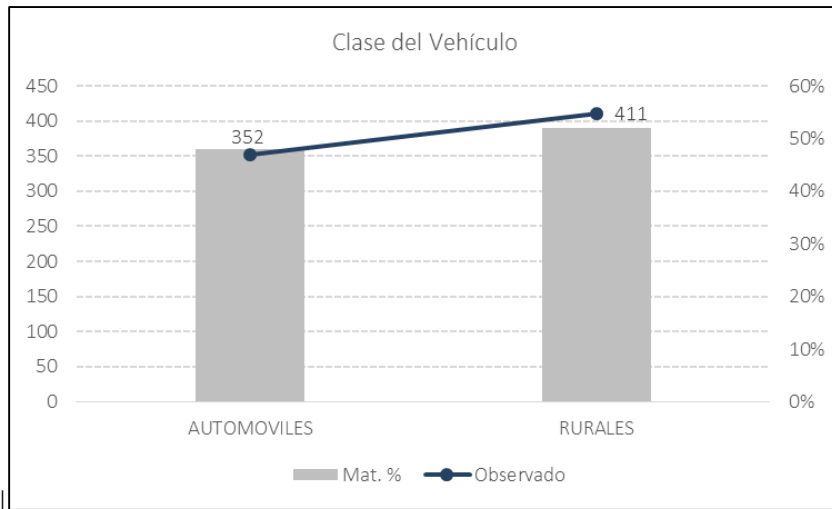
VI. REFERENCIAS BIBLIOGRÁFICAS

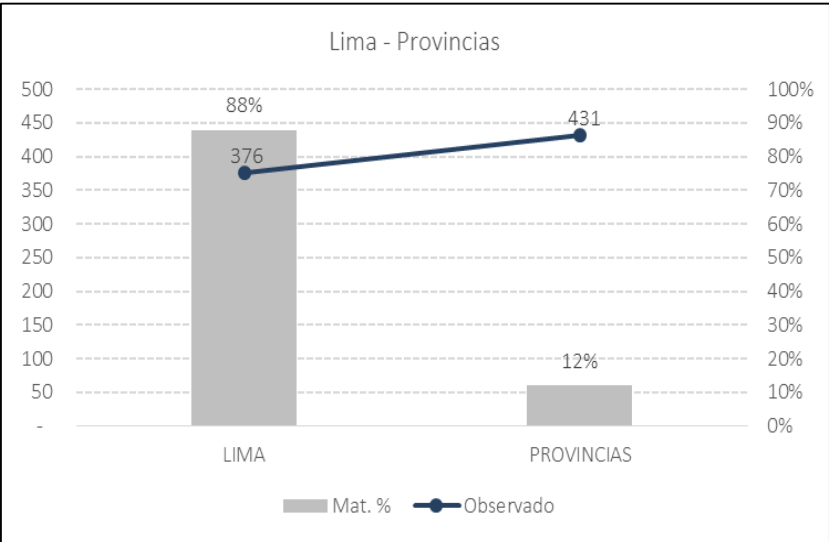
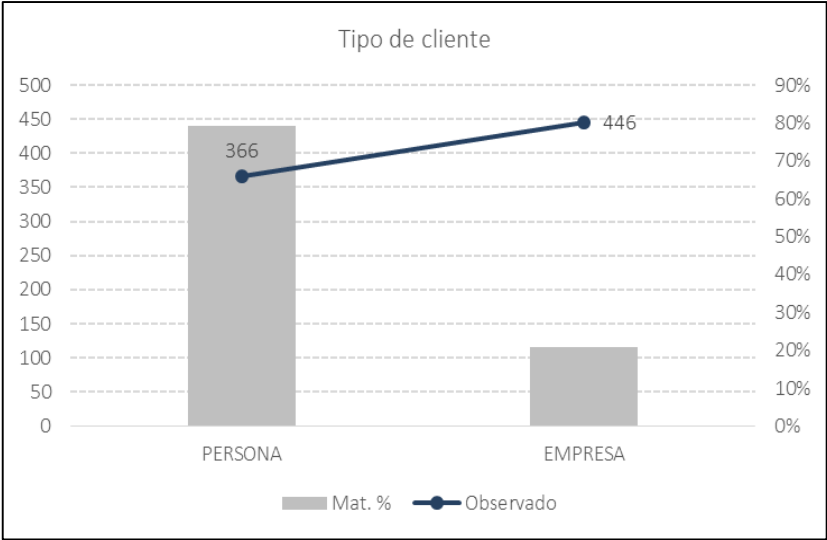
- De Jong, P. & Heller, G. Z. (2008). Generalized Linear Models for Insurance Data. En P. De Jong, & G. Z. Heller, *Generalized Linear Models for Insurance Data* (págs. 127-128). New York: Cambridge University Press.
- Gestión, R. (20 de 12 de 2014). *gestion.pe*. Obtenido de *gestion.pe*:
<https://gestion.pe/economia/parque-vehicular-peru-incremento-63-ultimos-10-anos-152332-noticia/>
- Goldburd, M.; Khare, A. & Tevet, D. (2016). *Generalized Linear Models for insurance rating*. Arlington: Casualty Actuarial Society.
- Hardin, J.M. & Hilbe, J.M. (2012). *Generalized Estimating Equations*. CRC Press.
- López, A. (2004). *Modelos Lineales Generalizados*. Valencia, España: Estadística Espacial en Epidemiología y Medio Ambiente.
- Martínez Gutiérrez, A. (2017). *Análisis de las Primas de Riesgo en Seguros de Automóviles: Una aplicación de los Modelos Lineales Generalizados*. Ciudad de México.
- Moura e Moura, M.M. (2017). *Cálculo de la Prima Pura en un Seguro de Automóvil para la Garantía de Daños Propios, mediante Modelos Lineales Generalizados y Segmentación de Clientes por Conglomerados*. Madrid.
- Ohlsson, E. & Johansson, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models*. Berlin: Springer.
- Peña Sánchez, I. (2017). *Tarificación de Microseguros: Una aplicación del Modelo Tweedie*. Madrid.
- Quishpe Tasiguano, I.D. (2015). *Factores de Riesgo de Siniestralidad y cálculo de primas de los vehículos asegurados en el Ecuador mediante Modelo Lineales Generalizados*. Quito.
- Villarino González, D. (2017). *Análisis de riesgos para la tarificación de seguros de automóvil mediante modelos lineales generalizados*. Cantabria.

VII. ANEXOS

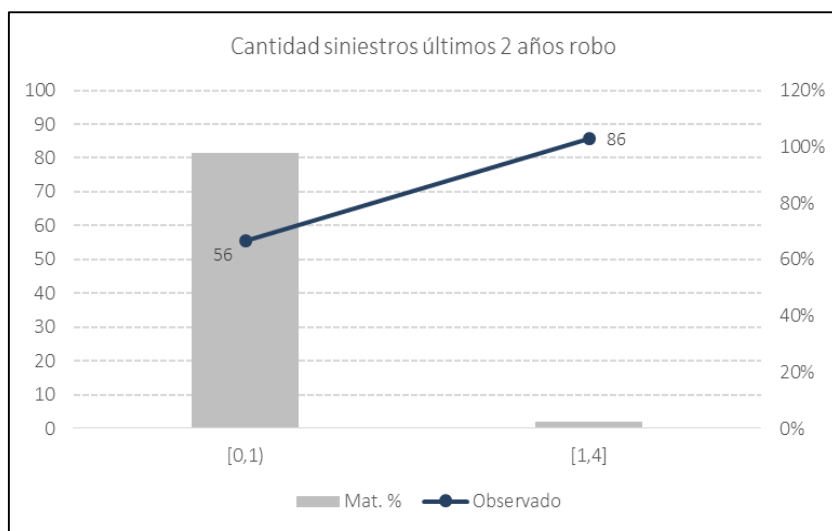
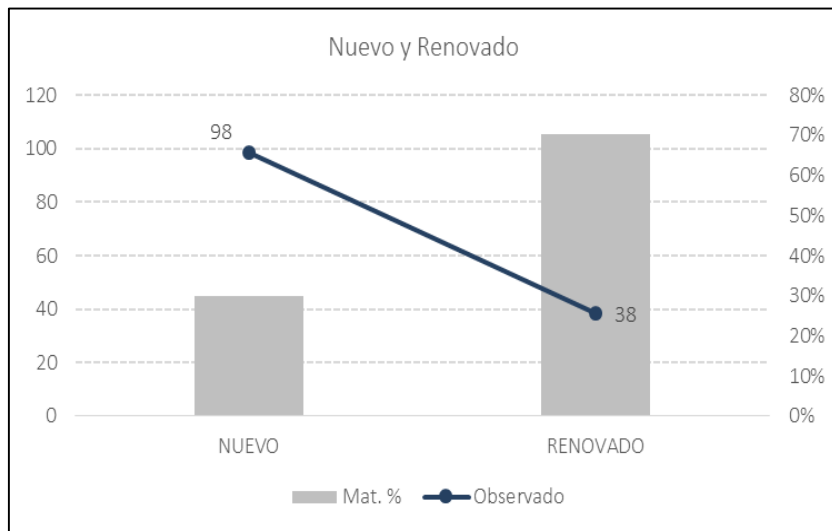
Anexo 1: Gráfico bivariado continuación Modelo 1

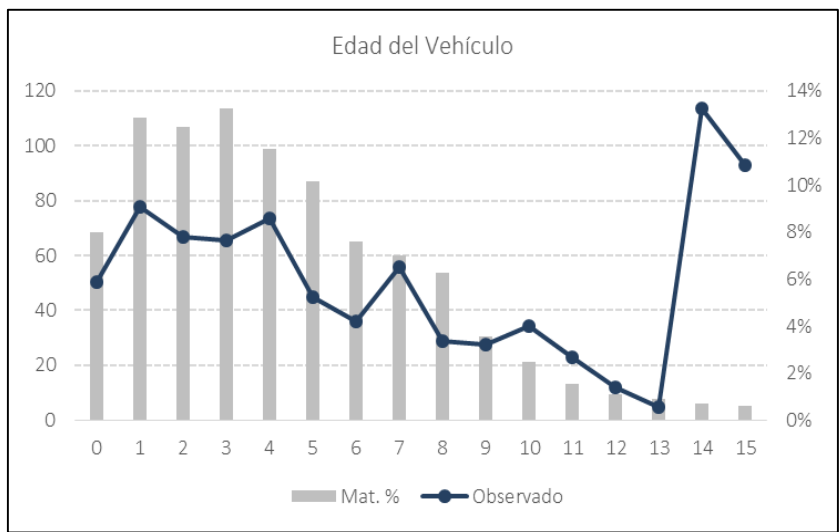
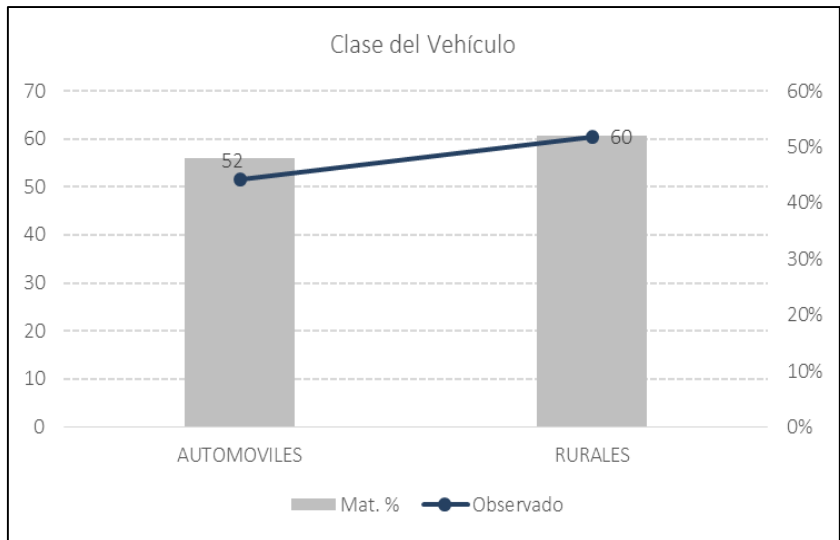


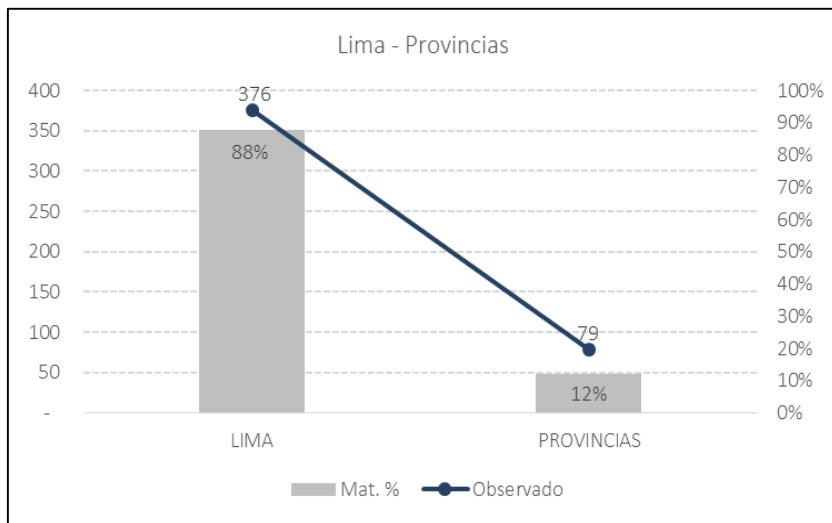
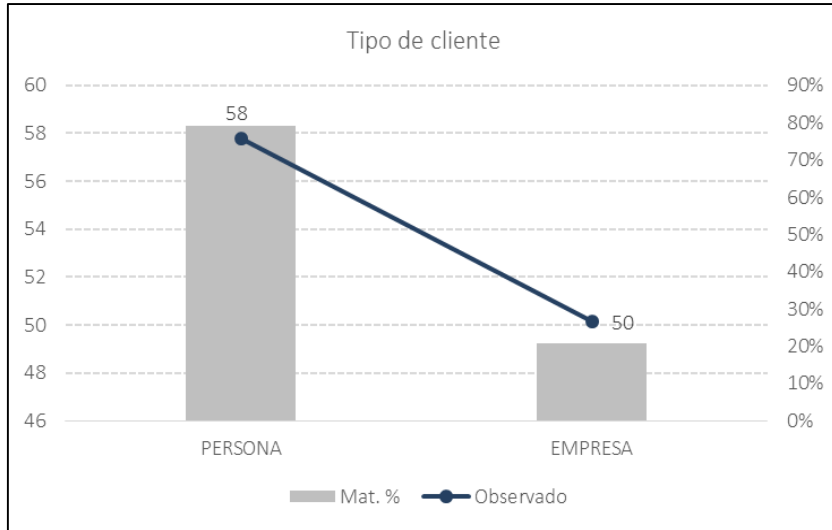




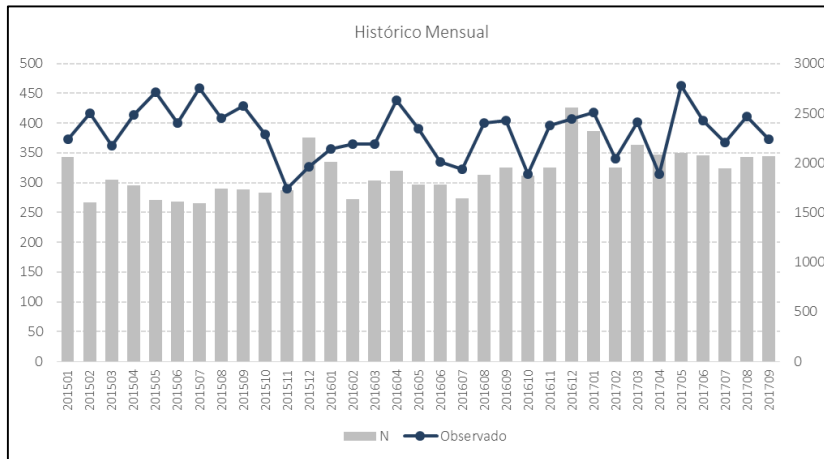
Anexo 2: Gráfico bivariado continuación Modelo 2







Anexo 3: Gráfico histórico mensual Modelo 1



Anexo 4: Gráfico histórico mensual Modelo 2

