Maestría en Ciencias Actuariales

Análisis Comparativo de Modelos de Machine Learning Y GLM de Procesos de Tarificación de Seguros de Automóviles

Wilhem Eisbey Castro Acosta

Bogotá D.C. 28 de noviembre de 2022

Análisis Comparativo de Modelos de Machine Learning Y GLM de Procesos de Tarificación de Seguros de Automóviles

Trabajo de Grado para Optar por el Título de
Magister en Ciencias Actuariales

Roberto Pérez
Director
Nicholás Metaxas
Codirector
Catalina Lozano
Jurado

Bogotá D.C 28 de noviembre de 2022

El trabajo de grado de maestría titulado:
"*Análisis Comparativo de modelos de Machine Learning y GLMs en procesos de tarificación de seguros de automóviles"*

Presentado por:
Wilhem Eisbey Castro Acosta

Cumple con los requisitos establecidos cumple con los requisitos establecidos para optar por el título de:
Magister en Ciencias Actuariales.

Catalina Lozano
Jurado

Roberto Pérez
Director

Nicholás Metaxas
Codirector

Bogotá D.C __ de diciembre de 2022

Agradecimientos

Mis más sinceros agradecimientos a mi familia que me apoya siempre en mis decisiones de crecimiento profesional, a mis asesores de trabajo de grado por sus consejos, a la universidad por permitirme hacer parte de una de sus cohortes de estudiantes que propenden aportar al progreso de nuestro hermoso país y finalmente, a todos los investigadores y académicos que nos preceden y nos permiten citando a Stephen Hawking, subirnos en “hombros de gigantes” para poder entender el mundo y la sociedad en la que vivimos y de esta manera soñar con que nuestro esfuerzo pueda redundar en su mejoramiento.

Resumen

En el presente trabajo se compara cuantitativa y cualitativamente el desempeño y los procesos metodológicos asociados a la predicción de la frecuencia y severidad para un caso particular en el ramo de automóviles, aplicando tanto técnicas clásicas como GLMs y algunas de sus variaciones como de Machine Learning, con el fin de evaluar sus potencialidades, ofreciendo mayor variedad en las herramientas usuales para este tipo de procesos, buscando detectar relaciones e interacciones no lineales en los predictores para disminuir escenarios de selección adversa en los modelos de tarificación. Se utilizó el marco metodológico CRISP-DM como guía para las etapas y procesos en el contexto analítico. Así mismo, se utilizó el software R y el paquete "caret" para desarrollar los diferentes modelos, buscando garantizar criterios adecuados de comparabilidad en la selección de los subconjuntos de validación cruzada. Finalmente, al comunicar los resultados de los modelos, las métricas de desempeño y los lift-charts asociados a cada uno de ellos, se concluye que al comparar las métricas de desempeño para el caso particular del dataset utilizado ("dataCar" de la librería "insuranceData" del software R) no existe ventaja cuantitativa considerable entre los dos enfoques (GLMs vs ML). Sin embargo, a través del análisis gráfico (Lift-charts) se aprecian diferencias en la capacidad de los modelos para detectar selección adversa, por lo que se concluye que las dos metodologías son complementarias pues ofrecen ventajas diferentes desde el punto de vista procedimental.

Contenido

Índice de Tablas

Índice de Figuras

Introducción

La industria aseguradora es una de las pocas en las cuales las tarifas de sus productos dependen de un componente fuertemente aleatorio, por lo que la rentabilidad de dichas compañías se encuentra altamente correlacionada con la calidad de sus procesos de tarificación. Hacia mediados de los años 60, los procedimientos de tarificación de seguros eran principalmente univariados y se movían alrededor del mismo principio, usando relatividades de precio de acuerdo con los diferentes tipos de riesgo de acuerdo con la desviación de la siniestralidad alrededor del promedio (Kuoa, 2020).

Posteriormente, hacia mediados de los años setenta, John Nelder y Robert Wedderburn (estadísticos inglés y escoces respectivamente), publican un artículo titulado "Generalized Linear Models" (GLMs) en donde se propone una generalización de los modelos lineales multivariados tradicionales, los cuales requerían que los datos cumplieran supuestos bastante rigurosos sobre todo en sus residuales, supuestos que no son cercanos a la naturaleza real de los fenómenos en donde conceptos como frecuencia y severidad inherentes al comportamiento de siniestralidad de la industria exigen técnicas más robustas. Es así como los GLMs, fueron adoptados progresivamente por compañías aseguradoras para desarrollar sus estrategias de tarificación desde finales del siglo XX hasta la fecha, obteniendo el aval de una de las organizaciones internacionales más importantes en el contexto actuarial, como lo es la Casualty Actuarial Society (CAS) que en 2016 publica una guía respecto a las metodologías de aplicación de los GLMs para procesos de P&C (Kuoa, 2020).

Sin embargo, a pesar de que los GLMs ayudaron a flexibilizar muchos de los supuestos asumidos por los modelos de regresión multivariados tradicionales, los GLMs aún requieren que las relaciones entre los predictores y la variable respuesta sea lineal en sus parámetros, lo que de cierta manera continua restringiendo las complejidades reales de los múltiples factores que explican los riesgos que se deben asumir por parte de las compañía de seguros, por lo que se identifica una gran oportunidad de mejora. Por otro lado, los GLMs permiten una mejor interpretabilidad de la tarifa final, generando que una mayor confianza en los altos inversionistas y los entes reguladores, quienes se encuentran a favor de su aplicación por la seguridad que ofrece el aval y la interpretación del actuario responsable en dichos procesos (Blier-Wong, 2001).

El concepto de Machine Learning se remonta a los años 50, pero no fue sino hasta comienzos del siglo XXI, que como resultado de los avances en la capacidad de procesamiento de información de las computadoras, se logró entender claramente su aplicación. Actualmente, el concepto de Machine Learning se puede clasificar en tres segmentos principales: Aprendizaje no supervisado, aprendizaje supervisado y aprendizaje por refuerzo (Riley, 2020).

El aprendizaje no supervisado se caracteriza por buscar patrones en los datos que no se conocen previamente sin tener como objetivo hacer ningún tipo de predicción particular. En contraste, el aprendizaje supervisado se proponen modelos de carácter predictivo, en donde los algoritmos aprenden patrones de un conjunto de datos de entrenamiento, conociendo el

valor de la variable (o variables) a predecir, para de esta manera, se logre estimar el valor de una variable objetivo en un conjunto de validación. En este ejercicio, se construyen métricas de desempeño que son utilizadas para cuantificar el desempeño y la capacidad predictiva de los diferentes algoritmos.

Por último, en el segmento de aprendizaje basado en refuerzo, se crean algoritmos que usan una estrategia de recompensa dependiendo de la efectividad de la predicción. Dichos modelos son usados principalmente en casos en los que se tengan que crear estrategias de recomendación personalizada (Wuthrich, 2020).

En la actualidad, se han documentado intentos (principalmente en Europa y EE. UU) de implementar modelos de Machine Learning en el contexto asegurador para predecir la frecuencia y la severidad de siniestros en algunas líneas de negocio, principalmente en líneas de Property and Casualty (P&C). En muchas de esas implementaciones, se ha encontrado que los modelos de Machine Learning han obtenido mejores métricas de desempeño que las obtenidas a través de la aplicación de GLMs clásicos, al lograr identificar patrones no lineales que difícilmente son capturados por modelos tradicionales de aprendizaje estadístico (Blier-Wong, 2001). Esta ventaja se ve no obstante restringida en la actualidad por la dificultad en la interpretabilidad de dichos modelos de cara a poder sustentar los resultados ante los entes reguladores de cada país (Kuoa, 2020).

Sin embargo, el presente y el futuro de la industria aseguradora muy seguramente dependerá de su capacidad de adaptación a las condiciones digitales cambiantes y sus complejidades. En ese sentido y más temprano que tarde, los entes reguladores tendrán que modificar sus criterios de aceptación de modelos que se ajustan mucho mejor a las características particulares de los asegurados, no para reemplazar los modelos tradicionales existentes, si no, para complementarlos y de esta manera usar por ejemplo técnicas de aprendizaje automático que ya se aplican activamente en otras industrias y que hacen parte de los procesos de valor para la toma de decisiones (Blier-Wong, 2001). En este sentido, queda el interrogante del rol del actuario contemporáneo de cara al proceso de adaptación de dichas técnicas que prometen mejorar los modelos de tarificación y de esta manera la capacidad de las compañías aseguradoras de medir los riesgos a los cuales se encuentran expuestas, riesgos que van a ir evolucionando de la mano con las características cambiantes de la población y del mercado (Riley, 2020).

1. Objetivo General

Analizar cuantitativa y cualitativamente el desempeño y los procesos metodológicos asociados a la tarificación de seguros en el ramo de automóviles, al aplicar técnicas de Machine Learning.

1.1. Objetivos Específicos

- Evaluar el impacto de la aplicación de modelos de Machine Learning en el desempeño de procesos de tarificación en el ramo de automóviles.
- Analizar comparativamente el desempeño de modelos tradicionales de tarificación de seguros (GLMs) con sus variaciones (Ej: GAMs, MARS, GLM Elastic Net)
- Analizar comparativamente el desempeño de modelos tradicionales de tarificación de seguros (GLMs) con modelos de Machine Learning (Ej:XGBoost, Random Forest, Redes Neuronales (MLP))
- Identificar cualitativamente las ventajas y desventajas en la aplicación de modelos de Machine Learning para estimar la tarifa en el ramo de automóviles.
- Proponer estrategias que permitan complementar los métodos tradicionales de tarificación de seguros generales (GLMs) con las potenciales ventajas en la implementación de modelos de Machine Learning.

2. Entendimiento del Contexto

2.1. Tarificación de Seguros

Los procesos de tarificación en seguros permiten a las aseguradoras fijar el precio de los riesgos individuales analizando la experiencia de las pérdidas de grupos de riesgos similares. Lo anterior con el objetivo de proteger al asegurador contra la selección adversa que significa el no ser capaz de diferenciar entre los altos y bajos riesgos en las compañías, lo que puede conducir a reducción de ganancias los inversionistas y pérdidas en la participación de mercado. Una construcción de tarifas eficiente puede proporcionar a las aseguradoras ventaja competitiva y mejora la posibilidad de mejorar la identificación del perfil de riesgos que el asegurador está dispuesto y puede asumir para suscribir de manera rentable (Geoff Werner, 2016).

2.1.1 Métodos Univariados

Dentro de los posibles enfoques de tarificación más sencillos se encuentran los enfoques univariados, cuyas principales debilidades son el no tener en cuenta el efecto de variables adicionales, lo cual puede mitigarse con un análisis bidireccional o algún ajuste manual. Sin embargo, los actuarios que trabajan en muchas líneas de negocios están analizando decenas o cientos de variables que hacen que el ajuste manual sea ineficiente, si no imposible. Otro enfoque utilizado para la elaboración de tarifas que fue popular durante la segunda mitad del siglo XX es la familia de procedimientos de mínimo sesgo. Dichos modelos son esencialmente univariantes estandarizados iterativamente. En estos enfoques, cada procedimiento implica la selección de una estructura de tarificación (por ejemplo, aditiva, multiplicativa o una combinación) y la selección de una función de sesgo (por ejemplo, principio de equilibrio, mínimos cuadrados, chi-2 y máximo funciones de sesgo de probabilidad). La función de sesgo es un medio para comparar las estadísticas de la pérdida observada del procedimiento (por ejemplo, costos de pérdida) con las estadísticas de pérdida indicadas y medir el desajuste (Geoff Werner, 2016), lo que debe ponderarse por las exposiciones en cada celda para ajustarse a una combinación desigual de negocios.

El término "sesgo mínimo" se refiere al principio de equilibrio comúnmente utilizado que requiere que la suma de las primas puras ponderadas indicadas es igual a la suma de los costos de pérdida observados ponderados para cada nivel de cada variable de tarificación, lo cual se conoce como "minimizar el sesgo" a lo largo de las dimensiones del sistema de clasificación. Los procedimientos de mínimo sesgo no son técnicamente métodos multivariados, y no fueron necesariamente basados directamente en la teoría estadística. Sin embargo, muchos de los procedimientos de sesgo mínimo son en realidad un subconjunto del método estadístico conocido como modelos lineales generalizados (GLMs). De hecho, al realizar un gran número de iteraciones en el procedimiento de sesgo mínimo se puede llegar a una convergencia con los resultados de GLM. Sin embargo, muchos argumentan que los procedimientos de mínimo sesgo implican una menor eficiencia computacional. El artículo de Stephen Mildenhall, "Una relación sistemática entre el sesgo mínimo y los modelos lineales generalizados" (Mildenhall 1999) demuestra que muchos de los procedimientos de sesgo mínimo corresponden directamente a modelos lineales generalizados (Geoff Werner, 2016).

2.1.2 Métodos Multivariados

Varios sucesos ocurrieron a fines del siglo XX y principios del siglo XXI que llevaron a la adopción de técnicas estadísticas, en particular modelos lineales generalizados, para la elaboración de procesos de tarificación:

- El poder de cómputo aumentó considerablemente.
- Los datos ya no tenían que ser agregados para ser analizados
- Lo qué anteriormente solo se podía lograr con grandes máquinas ahora se lograba con equipos de escritorio en una fracción del tiempo.

Por otro lado, las aseguradoras estaban instituyendo iniciativas de almacenamiento de datos que en gran medida mejoró la granularidad y la accesibilidad de los datos que podrían analizarse con fines de elaboración de tarifas. Así que, a pesar del hecho de que técnicas estadísticas sofisticadas existían mucho antes que esto, fueron las circunstancias de mayor poder de cómputo y mejores datos que permitieron su uso en procesos de tarificación. Un último y quizás el desencadenante más importante en la adopción generalizada de métodos multivariantes era la presión competitiva. Cuando una o más empresas comenzaron a implementar tarifas de clasificación mejoradas, obtuvieron una mayor ventaja competitiva, poniendo al resto de la industria en una posición de selección adversa y disminución de la rentabilidad. Esto ocurrió por ejemplo en el Reino Unido en la década de 1990 y en los mercados de automóviles de EE. UU. a principios de la década de 2000 (Geoff Werner, 2016).

Uno de los principales beneficios de los métodos multivariados es que consideran todas las variables de tarificación simultáneamente y ajustan automáticamente las correlaciones de exposición entre ellas la cual representa la principal deficiencia de los enfoques univariados. Por otro lado, los métodos multivariantes también intentan eliminar los efectos no sistemáticos en los datos (también conocidos como ruido) pues capturan los efectos sistemáticos (también conocidos como señal) tanto como sea posible a diferencia de los métodos univariados, que incluyen tanto la señal como ruido en los resultados. Adicionalmente, los métodos multivariados producen métricas para hacer diagnósticos a los modelos y de esta manera medir la certeza de los resultados. Finalmente, uno de los beneficios adicionales de los métodos multivariados es que permiten incluir consideraciones de interacción o interdependencia entre dos o más regresores del modelo (Geoff Werner, 2016).

3. Contexto Técnico

Los beneficios potenciales de los modelos multivariados varían considerablemente entre los diferentes métodos. Por ejemplo, una de las principales ventajas de los GLM es que son interpretables; la salida del modelo incluye estimaciones de parámetros para cada nivel de cada variable explicativa cualitativa en el modelo, así como un rango de diagnóstico estadístico y niveles de confianza. Por otro lado, otras técnicas multivariadas, como las redes neuronales son a menudo criticadas por su falta de interpretabilidad (Catalina Lozano, 2021), pues no importa cuán sofisticadas sean las matemáticas que soporten a un método, en la industria aseguradora es importante que los profesionales sean capaces de interpretar, traducir y comunicar los resultados para que estos puedan ser implementados en las operaciones de la compañía de seguros.

3.1.Modelos Lineales Generalizados

Los GLM (Generalized linear Models) se han convertido en la técnica estadística multivariante que representa el estándar para la tarificación de seguros para muchas líneas de negocio. Los GLM son una versión generalizada de modelos lineales que eliminan las restricciones de los supuestos de normalidad y de varianza constante de los modelos de regresión lineal tradicionales. También permiten que una función, llamada función de enlace o vínculo defina la relación entre la variable objetivo (por ejemplo, la gravedad de la reclamación) y los regresores o combinación lineal del predictor.

La elección de las funciones de enlace significa que las variables predictoras no tienen que relacionarse estrictamente de manera aditiva (como lo hacen con los modelos lineales tradicionales). Los GLM se ajustan a la experiencia de reclamaciones de seguros para fines de fijación especificando un enlace que asume que los predictores se relacionan multiplicativamente entre sí. Hay otros componentes de la formulación de GLM como pesos específicos que también pueden utilizarse en caso en que los registros se hayan agregado para ingresar al modelo (Ej: Exposición).

Para resolver un GLM, el actuario debe proporcionar un conjunto de datos con un número adecuado y suficiente de observaciones para la variable respuesta y las variables predictoras asociadas. También, se debe seleccionar una función de enlace para definir la relación entre los componentes sistemáticos y aleatorios, así como especificar la distribución del proceso aleatorio subyacente, que usualmente resulta ser un miembro de la familia de distribuciones exponenciales (ejemplo, normal, Poisson, gamma, binomial, inversa gaussiana). Esto, determinando la media y varianza de la distribución, siendo esta última una función de la primera (Geoff Werner, 2016).

Usualmente un GLM tiene la siguiente estructura:

$$g(u_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} \quad (3.1)$$

Siendo $g(u_i)$ la función de vínculo $\beta_i$los parámetros del modelo para la i-ésima observación. Cuando la función de vínculo es el logaritmo natural de u_i:

$$ln(u_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} \quad (3.2)$$

El modelo tiene la propiedad de generar una estructura de tarificación multiplicativa:

$$u_i = exp\,(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip)}$$
$$= exp\,(\beta_0)\; exp\; exp\left(\beta_1 x_{i1}\right)\; exp\; exp\left(\beta_2 x_{i2}\right) exp\,(\beta_3 x_{i3}) \ldots exp\,(\beta_p x_{ip}) \quad (3.3)$$

Esta estructura multiplicativa es la más común en el contexto de modelos de tarificación a partir de GLM por su facilidad en implementación e interpretabilidad.

3.1.1. Regularización (Redes Elásticas)

Cuando se construyen modelos de tarificación, generalmente hay un gran número de predictores por lo que el efecto de sobre ajuste es un peligro potencial. Cuando nos referimos a sobre ajustar un modelo queremos decir que, si aumentamos la complejidad de este para un conjunto específico de datos, el modelo puede capturar en exceso las asociaciones complejas entre los predictores y la variable objetivo, por lo que al cambiar el set de datos dicho modelo no podrá generalizarse de una manera significativa y confiable. Para evitar lo anterior, existen técnicas diseñadas para penalizar a los modelos en términos de su complejidad (número de variables) aumentando el valor de su desviación (error) con un término adicional que se encuentra asociado con los parámetros del modelo y con otros de penalización.

Dentro de las técnicas de penalización a los modelos para evitar el sobre ajuste, conocidas también como técnicas de regularización, se encuentran las llamadas redes elásticas cuya función de minimización está descrita por:

$$Desviación + \lambda \left(\alpha \sum \quad |\beta| + (1-\alpha) \frac{1}{2} \sum \quad \beta^2 \right) \quad (3.4)$$

El primer término aditivo de la expresión anterior es solo la desviación de GLM tradicional. La red elástica agrega un término de penalización compuesto por un promedio ponderado de la suma de los valores absolutos de los coeficientes más la suma reducida a la mitad de los coeficientes al cuadrado, con las ponderaciones determinadas por α, un parámetro entre 0 y 1 que se controla y se calibra a voluntad del modelador para agregar más peso al primer término con los valores absolutos (Lasso) o al segundo con los “betas” al cuadrado (Ridge). Lo importante a reconocer en el modelo de regularización anterior, es que los términos dentro de los paréntesis producen una función creciente proporcional a la magnitud de los coeficientes, así como el grado en que los coeficientes se desvían de cero. De esta manera, se aplica penalización para coeficientes más grandes.

Por otro lado, el parámetro de ajuste más importante en la ecuación anterior es λ, pues nos permite controlar la severidad de la sanción que se aplica. El efecto práctico de elevar λ es que obliga a los coeficientes a reducirse más cerca de cero, para compensar el aumento de la penalización, minimizando el error compuesto por la desviación más el término de penalización error.

3.1.2. Generalized Additive Models (GAMs)

El modelo aditivo generalizado o GAM por su nombre en inglés (Generalized Additive Model) maneja las no linealidades que no son capturadas por modelos GLM tradicionales.

Al igual que estos últimos, la variable a predecir o variable respuesta debe cumplir con la condición de una distribución que pertenezca a la familia de distribuciones exponenciales:

$$y_i = Familia\ Exponencial(\mu_i, \phi) \quad (3.5)$$

Con la diferencia de que los términos utilizados para hacer la estimación no son funciones lineales:

$$g(u_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + \cdots + f_p(x_{ip}) \quad (3.6)$$

Sino que pueden estar representados por cualquier función arbitraria de los predictores. Las formas de las curvas suaves representadas por las funciones son generalmente estimadas por las librerías de los softwares especializados.

3.1.3. Multivariate Adaptive Regression Splines (MARS)

Dentro de los modelos aditivos en el contexto de los GLM se encuentran los "Multivariate Adaptive regression Splines" (MARS), los cuales son un procedimiento adaptativo para la regresión, y es muy adecuado para problemas de alta dimensionalidad (es decir, un gran número de predictores). Se puede ver como una generalización de la regresión lineal escalonada (stepwise) o una modificación del CART (Classification and Regression Trees), empleado principalmente para mejorar el rendimiento de este último en el entorno de regresión.

En lugar de ajustar funciones suaves para los predictores, como lo hace el GAM, los modelos MARS operan incorporando funciones lineales por partes dentro de un GLM regular. Los modelos MARS crean las funciones y optimizan los puntos de corte automáticamente. MARS utiliza expansiones en funciones de base lineal por partes de la forma $(x - t)_+$ y $(t - x)_+$. El "+" significa parte positiva.

La estrategia de construcción de modelos es como una regresión lineal stepwise, pero en lugar de usar las entradas originales(predictores), se procede a usar las funciones del conjunto colección de las funciones básicas y sus productos. Así el modelo tiene la forma:

$$f(X) = \beta_0 + \sum \quad \beta_m h_m(X) \quad (3.7)$$

En donde $h_m(X)$ Es una función en el conjunto colección de funciones básicas o el producto de dos o más de ellas. Dada una opción para el h_m, los coeficientes β_m se estiman minimizando la suma del cuadrado de los residuos, es decir, mediante regresión lineal estándar. Sin embargo, la clave del proceso está en la construcción de las funciones $h_m(X)$.

3.2.Machine Learning

El aprendizaje automático o Machine Learning es una rama en evolución de los algoritmos computacionales que están diseñados para emular la inteligencia humana aprendiendo del entorno circundante basado en datos. Son considerados fundamentales en la nueva era del llamado “Big data”. Las técnicas basadas en el aprendizaje automático se han aplicado con éxito en diversos campos que van desde el reconocimiento de patrones, la visión por computadora, la ingeniería de naves espaciales, las finanzas, el entretenimiento, la biología computacional hasta las aplicaciones biomédicas y médicas.

El concepto de Machine Learning en su sentido moderno es generalmente asociado con el nombre del psicólogo Frank Rosenblatt de la Universidad de Cornell, quien, basado en ideas sobre el funcionamiento del sistema nervioso humano creó un grupo de investigación que construyó una máquina capaz de reconocer las letras del alfabeto. La máquina llamada perceptrón usó señales analógicas y discretas incluyendo un elemento de umbral que convierte las señales analógicas en discretas convirtiéndose de esa manera en un prototipo de las redes neuronales modernas (Imran Chowdhury Dipto, 2020).

3.2.1. Aprendizaje Supervisado

La mayoría de los problemas estadísticos y de Machine Learning pertenecen al dominio de alguna de las siguientes 2 categorías, aprendizaje supervisado o no supervisado. Para los propósitos del siguiente estudio se hará énfasis en el aprendizaje supervisado, pues es a esta categoría a la que están asociados los métodos de tarificación, principalmente. Sin embargo, es importante aclarar que para algunos casos particulares puede que exista la necesidad de aplicar métodos no supervisados como pasos intermedios en procesos de tarificación.

El aprendizaje supervisado tiene que ver con metodologías para las cuales se cuenta con una o más variables respuesta en el set de datos, por lo que se puede evaluar el nivel de error con el cuál se pretende estimar dicha variable. De acuerdo con lo anterior, resulta natural el entender que los modelos predictivos como es el caso de los modelos de tarificación pertenecen a este tipo de aprendizaje.

Desde el punto de vista técnico y algorítmico, si se supone que los errores son aditivos y que el modelo $Y = f(X) + \varepsilon$, es en ese caso un supuesto razonable. El aprendizaje supervisado intenta aprender “f” a través de los valores de entrada observados para el sistema "x_i”, sistema conocido como algoritmo de aprendizaje (generalmente desarrollado por un programa de computadora), que también produce salidas $\hat{f}(xi)$ en respuesta a las entradas. El algoritmo de aprendizaje tiene la propiedad de que puede modificar su relación entrada/salida $\hat{f}$ en respuesta a las diferencias $y_i - \hat{f}(x_i)$ entre las salidas originales y generadas. Este proceso se conoce como aprendizaje a través del ejemplo (Gareth James, 2021).

Al finalizar el proceso de aprendizaje, el objetivo es que las estimaciones artificiales y los resultados reales estarán lo suficientemente cerca como para ser replicables en los conjuntos de entradas que probablemente encontrarán en la práctica. Sin embargo, teniendo en cuenta que los métodos de aprendizaje supervisado intentan aprender de los datos proporcionados para minimizar la función de error o costo que se defina y que se debe tener cuidados de no caer en sobreajuste, siempre se debe lidiar con el equilibrio entre sesgo y varianza.

3.2.1.1. Algoritmo tipo Bagging

La palabra "bagging" viene de la expresión "Bootstrap aggregating" y consiste en aplicar muestreo con reemplazo con el objetivo de minimizar el error de alta varianza de algunos modelos de Machine Learning, como los árboles de decisión. El objetivo es que, considerando un conjunto de variables independientes con igual varianza σ^2, se pueda obtener una la varianza de la media de las observaciones reducida en proporción al número de observaciones $\frac{\sigma^2}{n}$ (Jiménez, 2020).

La estrategia de "bagging" consiste en obtener varias muestras de una población y ajustar un modelo distinto para cada una de ellas. La predicción final se calcula como la media de las predicciones anteriores. La dificultad se encuentra en la consecución de múltiples muestras, para lo cual se recurre al bootstrapping (muestreo con reemplazo), que permite generar pseudo muestras.

Aunque el proceso de "bagging" generalmente consigue mejorar la capacidad predictiva de los modelos, la interpretación de éstos se torna más compleja. La representación gráfica pasa a ser menos intuitiva e identificar la importancia de los predictores ya no es un proceso tan directo, en tanto se emplea un elevado número de árboles de decisión de gran tamaño con poco sesgo pero alta varianza y de esta manera, mantener el mismo nivel de sesgo, reduciendo el nivel de varianza a través de agregaciones (Jiménez, 2020).

3.2.1.2. Algoritmo tipo Boosting

La estrategia de boosting consiste en entrenar, de manera iterativa (en serie), varios modelos sencillos con poco valor predictivo, de tal manera que cada nuevo modelo utiliza información del anterior para adaptarse y aprender de sus fallos y aciertos. En el caso de árboles de decisión, estos modelos sencillos serán árboles con una o pocas ramificaciones. A diferencia de la estrategia de bagging, a través de boosting se pretende conseguir árboles muy correlacionados entre sí. Otra diferencia con respecto a los algoritmos de bagging, es el gran número de hiperparámetros que utilizan. Los principales son los siguientes (Jiménez, 2020):

• **Número de modelos sencillos o número de iteraciones:** Si este número es demasiado alto, los algoritmos podrían sufrir de "overfitting" y perder capacidad predictiva.

• **Learning Rate:** Es un término de regularización que se utiliza para evitar el "overfitting" en los algoritmos de "boosting". Muestra el ritmo al que aprenden los modelos. Es recomendable que se sitúe entre 0,01 y 0,001.
• **Número de divisiones (d) de cada árbol:** Es conveniente utilizar valores pequeños (entre 1 y 10) (Jiménez, 2020).

3.2.1.3. Redes Neuronales y Deep Learning

El concepto de red neuronal hace parte de un subconjunto de los algoritmos de Machine Learning conocido como Deep Learning. Sus inicios se remontan a los años 80s en donde existía interés por esta aproximación en el contexto de sistemas de procesamiento de información neuronal (NeurIPS). Sin embargo, el advenimiento de algoritmos como support vector machines (SVMs), Boosting y Random Forest, modelos que requerían un menor esfuerzo interpretativo redujeron la aplicación de las redes neuronales a finales del siglo pasado. No fue sino hasta inicios del 2010 que con un nuevo nombre (Deep Learning) las redes neuronales tuvieron un nuevo aire en el contexto de la necesidad de mercados de nicho como clasificación de imágenes y video, así como procesamiento de audio y de texto (Gareth James, 2021).

En el contexto de aprendizaje supervisado las redes neuronales requieren como input un vector de variables de entrada para construir una función no lineal de interacciones entre ellos para predecir una variable respuesta. A pesar de que ya se habían mencionado modelos predictivos no lineales como los métodos de Bagging y Boosting, lo que distingue a las redes neuronales de estos métodos es la estructura particular del modelo:

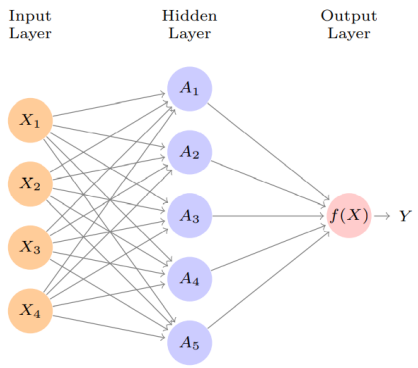


Figura 1 Arquitectura de Red Neuronal. Tomado de (Gareth, 2022)

La ilustración anterior muestra una red neuronal feed-forward para modelar una respuesta cuantitativa Y a través de 4 predictores $x_i \; para \; i = 1,2,3,4$ los cuales en la terminología de las redes neuronales se encuentran ubicados en la capa de entrada de la red. Las flechas

posteriores a ésta cada indican que cada una de dichas variables interactúan con las demás, en cada una de las neuronas de la siguiente capa conocida como capa oculta. Finalmente, la capa oculta se conecta con una capa de salida cuya representación analítica es:

$$f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k A_k \qquad (3.8)$$

Donde:

$$A_k = g(wk_0 + \sum_{j=1}^{p} w_{kj} X_j) \qquad (3.9)$$

Siendo $g(z)$ una función de activación no lineal. Como es común en las técnicas de machine learning el objetivo en este caso es encontrar los w_{kj} que minimizan el error de predicción del modelo. Sin embargo, por la estructura iterativa de la red, en realidad no se tiene visibilidad de los pesos w_{kj} ni de su interpretación, por lo que esto representa una desventaja en contextos como el mercado asegurador en donde los reguladores requieren tener un entendimiento profundo de los supuestos y parámetros de los modelos.

3.3 Evaluación de Desempeño

Existen múltiples opciones dentro de las métricas de desempeño disponibles en la literatura. La esencia de cada una de ellas en el contexto de regresión se enfoca generalmente en comparar los resultados de las predicciones con respecto a los valores reales de la variable objetivo. Es natural que conceptos como el de distancia en diferentes topologías sean candidatos para evaluar dichas comparaciones. A continuación, se muestran algunas de las métricas usuales en el contexto de regresión para GLMs y modelos de Machine Learning:

3.3.2 R-Squared:

$$R - Squared = \frac{SSRegression}{SSTotal} \qquad (3.10)$$

Donde $SSRegression$ es la suma de cuadrados de regresión y $SSTotal$ es la suma de cuadrados total.

3.3.2 Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{I=1}^{n} |y_i - \hat{y}_i| \qquad (3.11)$$

Siendo y_i el i-esimo valor de la variable objetivo y $\hat{y}_i$ la estimación para ese el mismo.

3.3.3 Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{I=1}^{n}(y_i - \hat{y}_i)^2 \qquad (3.12)$$

Esta métrica tiene la desventaja de no estar en las unidades de la variable objetivo.

3.3.4 Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{MSE} \qquad (3.13)$$

Existen métricas de desempeño que incorporan penalizaciones a la dimensionalidad de la vista miná ble que ingresa en los modelos, como es el caso del AIC (Akaike Information Criteria) que penaliza por el número de grados de libertad, y del BIC (Bayesian Information Criteria) que penaliza tanto por los grados de libertad como por el número de observaciones. En el presente trabajo nos centraremos en las métricas que pueden usarse en todos los contextos testeados en el estudio.

3.3.4 Lift-Charts:

Los lift-charts son un caso particular del segmento de gráficas conocido como “Quantile Plots” y están diseñadas para medir la habilidad de diferenciar entre los “buenos” de los “malos” riesgos desde el punto de vista del asegurador (Roberto Perez, 2021), es decir, evitar selección adversa. Se evalúan comparando la similitud entre las predicciones y los valores actuales de la variable objetivo, la monotonicidad y la distancia vertical entre el primer y último punto. Su eje horizontal está compuesto comúnmente por deciles de la exposición.

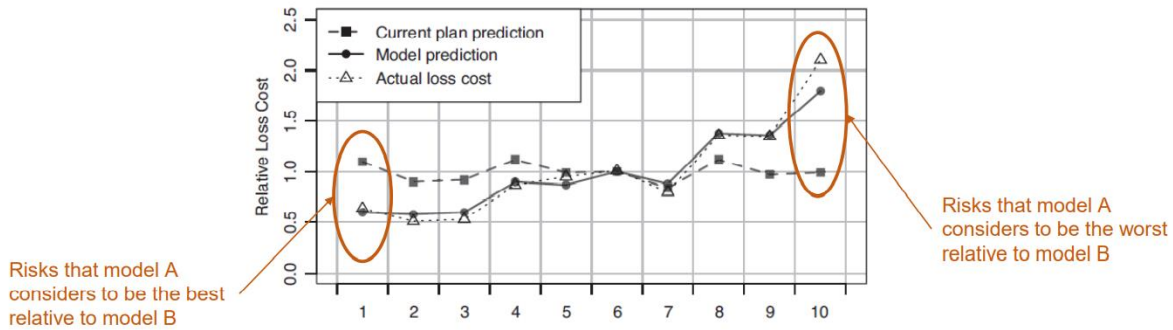


Figura 2. Ejemplo de Lift-Chart tomado de (Roberto Perez, 2021)

4. Metodología

El marco metodológico global del proyecto y por lo tanto del documento será CRSP-DM (ver figura 3).

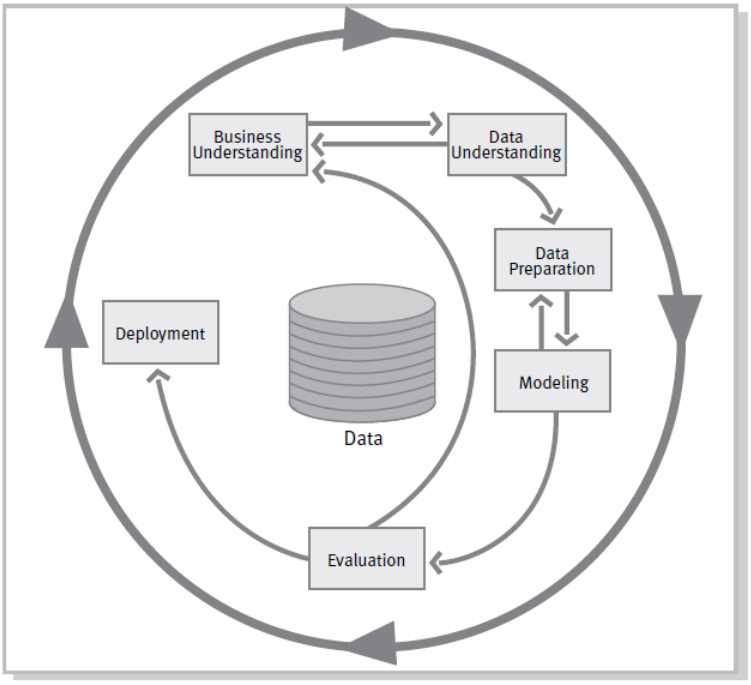


Figura 3. Tomado de 2000 SPSS Inc. CRISPMWP-1104

Éste es un marco metodológico ampliamente usado en proyectos de analítica de datos, concebido por IBM en 1996. En dicho marco se circunscriben los objetivos específicos del presente proyecto en sus etapas globales. El proceso inicia con la etapa de entendimiento del negocio en el mercado real, cuyo símil en el presente proyecto, representa la identificación del contexto de los datos, es decir, el ramo de automóviles. Posteriormente, en la etapa de entendimiento de los datos, se describen los metadatos junto con sus fortalezas y debilidades. En la etapa de preparación de datos, se adecúa la información para que cumpla los supuestos y tenga la estructura requerida para las vistas minables que ingresarán a los modelos. La etapa de modelamiento tiene el objetivo de identificar cuantitativamente las características de los modelos aplicados para estimar los elementos necesarios en la construcción de una tarifa en el ramo de automóviles.

Finalmente, en la etapa de "Evaluación" se analizarán comparativamente el desempeño de los modelos tipo GLMs y algunas de sus variaciones (Ej: GAMs, MARS, Elastic Net) versus modelos de Machine Learning (Ej: XGBoost, Random Forest, Redes Neuronales MLP). En esta misma etapa se describirán estrategias que permitan complementar los métodos tradicionales con las potenciales ventajas de modelos de Machine Learning. El set

de datos con el que se construirán los modelos provendrá de datos abiertos (Ej: Data pertinente incorporada en paquetes del software R).

4.1. Supuestos de Investigación

Dentro de los supuestos más importantes con los que se construye el presente trabajo se encuentran:

- El set de datos elegido representa fuente de información válida para cumplir el objetivo de construir un plan de tarificación en el ramo de automóviles.
- Los recursos computacionales locales y el software libre son suficientes para ejemplificar la situación problema.
- Las referencias consultadas son las más actuales y robustas dentro del cambiante contexto de la analítica de datos aplicada.
- Es posible encontrar una combinación adecuada entre técnicas de Machine Learning y técnicas tradicionales en el contexto de la tarificación de seguros de automóviles en la que exista un complemento adecuado entre ellas.

4.2. Restricciones

- Las fuentes de datos reales no se encuentran de manera disponible y abierta, lo que dificulta contrastar los resultados y técnicas del presente análisis.
- Los resultados de las diferentes técnicas de análisis dependen de la existencia de los paquetes de software libre que se encuentren disponibles para la realización del estudio.
- Es posible que el proceso de calibración de los hiperparámetros en la etapa de modelamiento impida llegar a las soluciones óptimas de los modelos que lo requieran.

4.3. Riesgos

- La posibilidad de que el set de datos elegido no proporcione suficiente variabilidad para representar una situación de contexto real.
- Teniendo en cuenta que el presente trabajo se realizará durante el año 2022, es posible que el versionamiento del software y de los paquetes utilizados impidan correr adecuadamente los modelos y los desarrollos en etapas finales del trabajo.
- Es posible que los resultados no se puedan generalizar o transferir a otras líneas de negocio en la industria de seguros.

5. Objetivos Minería de Datos

- Desarrollar procesos de exploración de datos para identificar las características de los campos del dataset, así como sus posibles asociaciones.
- Definir un proceso de preparación y evaluación de la calidad de los datos a utilizar.
- Construir una vista mináble que permita que los modelos a utilizar puedan ser aplicados a cabalidad.
- Construir modelos predictivos en el contexto de los GLMs y de Machine Learning para varios valores de hiper parámetros.
- Definir métricas de desempeño robustas para evaluar el desempeño de los modelos aplicados.

5.1. Criterios de Éxito de minería

- Lograr identificar una combinación adecuada entre técnicas de Machine Learning y técnicas tradicionales GLM en el contexto de la tarificación de seguros de automóviles, en la que coexistan y se complementen.
- Crecimiento profesional resultado del proceso de aprendizaje al integrar los diferentes conceptos y técnicas y aproximaciones a un caso particular.

6. Descripción de datos

6.1. Evaluación de Herramientas y Técnicas

Para el desarrollo del proyecto se utilizarán diversas herramientas y técnicas que nos ayuden a entender los datos, explorarlos y evaluar las soluciones planteadas, las principales herramientas que utilizaremos son:

Tabla 1 Herramientas a utilizar

Herramientas	Propósito
R	Proporcionará las herramientas necesarias para la fase de entendimiento y limpieza de los datos. También será la herramienta principal para la creación de los modelos finales de minería.
Google Drive	Los datos del cliente serán suministrados a partir de esta herramienta.
PowerBI	Se utilizará para la exploración inicial de los datos.
Python	Se utilizará para complementar los análisis desarrollados en R

7. Entendimiento de los Datos

En el presente trabajo se hará uso del conjunto de datos “dataCar”, el cual es uno de los 12 datasets del paquete “insuranceData” del software R creado en el año 2015 por Alicja Wolny Dominiak y Michal Trzesio con el objetivo de proporcionar un lienzo para probar modelos predictivos.

7.1. Obtención de los Datos

Para obtener los datos se instala el paquete “insuranceData” del software R y posteriormente se llaman el dataset “dataCar”

Tabla 2 Dataset

NOMBRE	TAMAÑO	N FILAS	N COLUMNAS	FORMATO
dataCar, librería insuranceData, software R	4.92 KB	67856	11	.txt

7.2. Descripción de los Datos

Tabla 3 Metadata

LABEL	DESCRIPCIÓN	TIPO DE VARIABLE	DOMINIO
veh_value	Valor del vehículo en escala de $10,000s	Continua	$[0, \infty)$
exposure	Exposición	Cuantitativa Continua	[0,1]
clm	Ocurrencia de reclamación	Dicotómica	{0,1}
numclaims	Número de reclamaciones	Cuantitativa Entera	$[0, \infty)$
claimcst0	Monto de la reclamación	Cuantitativa continua	$[0, \infty)$
veh_body	Tipo de vehículo	Cualitativa Nominal	{BUS, CONVT, COUPE, HBACK, HDTOP, MCARA, MIBUS, PANVN, RDSRT, SEDAN}
veh_age	Antigüedad del vehículo	Cualitativa Ordinal	{1,2,3,4}
gender	Género del tomador	Cualitativa Nominal	{F, M}

area	Área	Cualitativa Nominal	{A, B, C, D, E}
agecat	Categoría de Edad del tomador	Cualitativa Ordinal	{1,2,3,4,5,6}
X_OBSTAT	Factor de origen desconocido	Cualitativa	{01101,0,0,0}

7.3. Verificación de Calidad de Datos

Basic Statistics

Raw Counts

Name	Value
Rows	67,856
Columns	11
Discrete columns	4
Continuous columns	7
All missing columns	0
Missing observations	0
Complete Rows	67,856
Total observations	746,416

Figura 4 Calidad de Datos

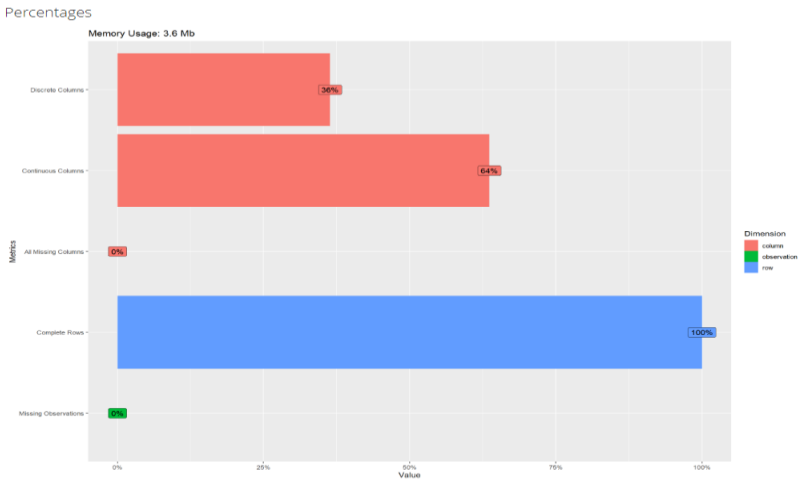


Figura 5 Calidad de datos II

7.4.Exploración inicial de Datos

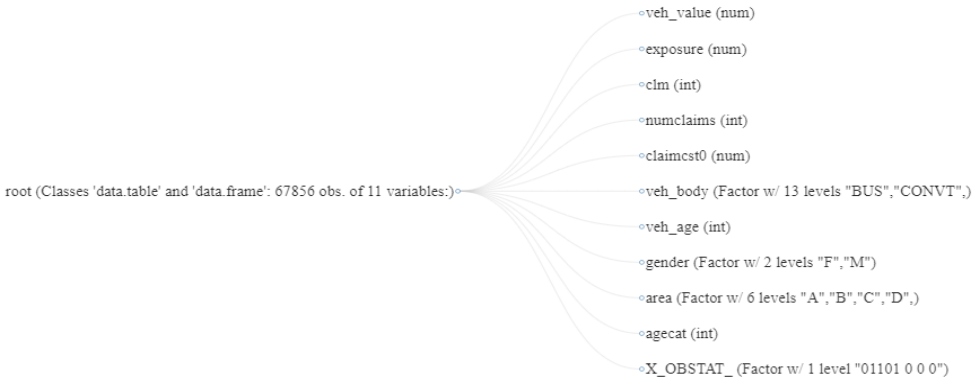


Figura 6 Campos

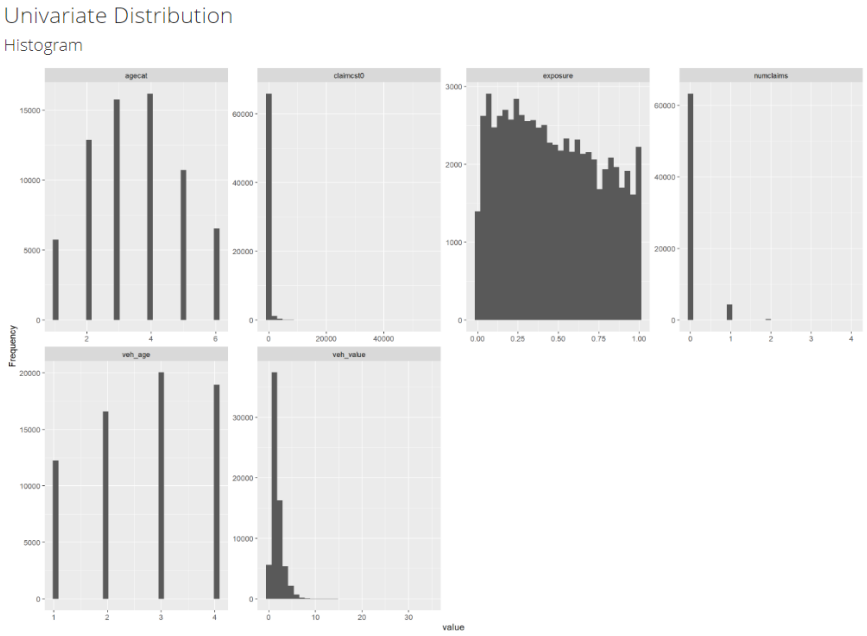


Figura 7 Histogramas Variables Continuas

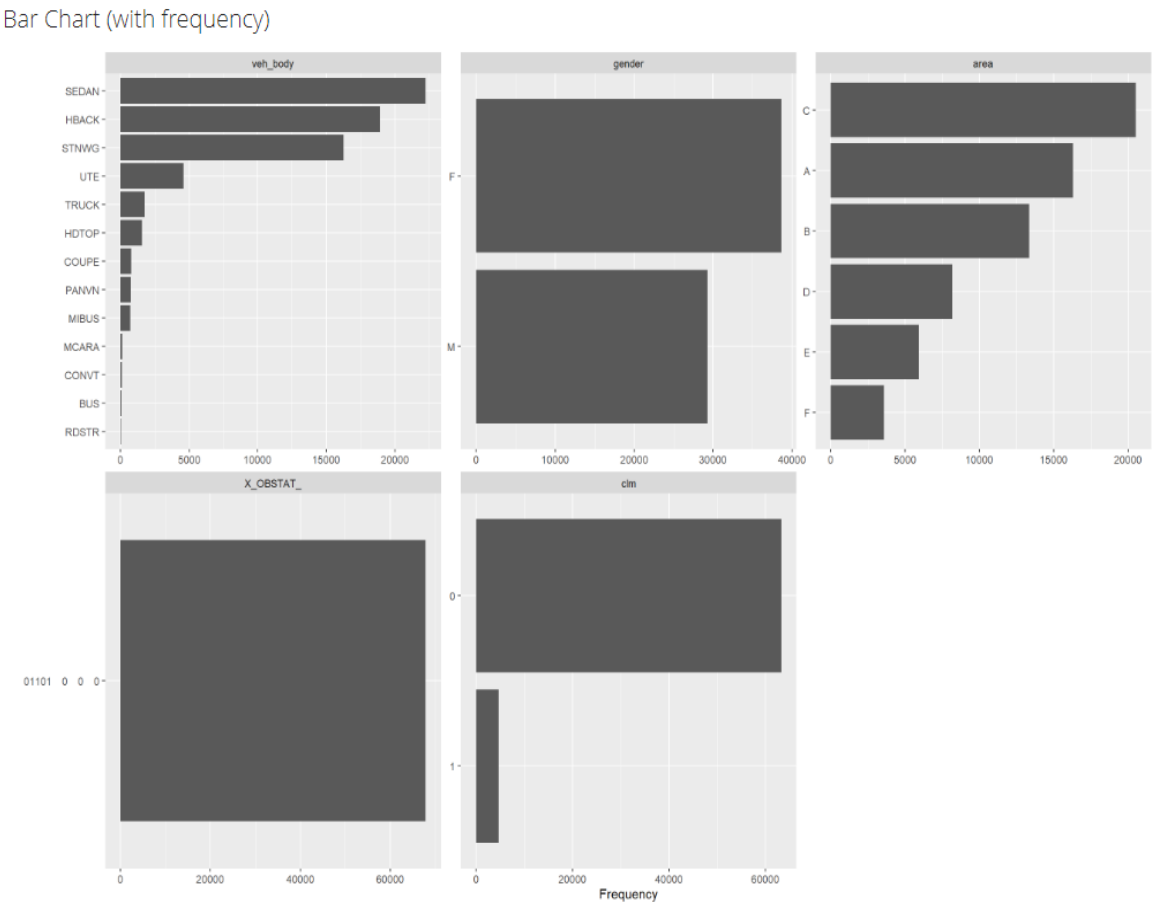


Figura 8 Distribución variables Categóricas

Tabla 4. Estructura del Set de Datos

```
'data.frame':   67856 obs. of  11 variables:
 $ veh_value: num  1.06 1.03 3.26 4.14 0.72 2.01 1.6 1.47 0.52 0.38 ...
 $ exposure : num  0.304 0.649 0.569 0.318 0.649 ...
 $ clm      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ numclaims: int  0 0 0 0 0 0 0 0 0 0 ...
 $ claimcst0: num  0 0 0 0 0 0 0 0 0 0 ...
 $ veh_body : Factor w/ 13 levels "BUS","CONVT",..: 4 4 13 11 4 5 8 4 4 4 ...
 $ veh_age  : int  3 2 2 2 4 3 3 2 4 4 ...
 $ gender   : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 2 1 1 ...
 $ area     : Factor w/ 6 levels "A","B","C","D",..: 3 1 5 4 3 3 1 2 1 2 ...
 $ agecat   : int  2 4 2 2 2 4 4 6 3 4 ...
 $ X_OBSTAT_: Factor w/ 1 level "01101    0    0    0": 1 1 1 1 1 1 1 1 1 1 ...
```

8. Preparación de Datos

8.1. Seleccionar datos

Teniendo en cuenta la buena calidad de los datos, la única variable que no brinda suficiente variabilidad y no tiene contexto interpretativo es la variable “X_OBSTAT” (Tabla 4), por lo que no será tenida en cuenta en los modelos.

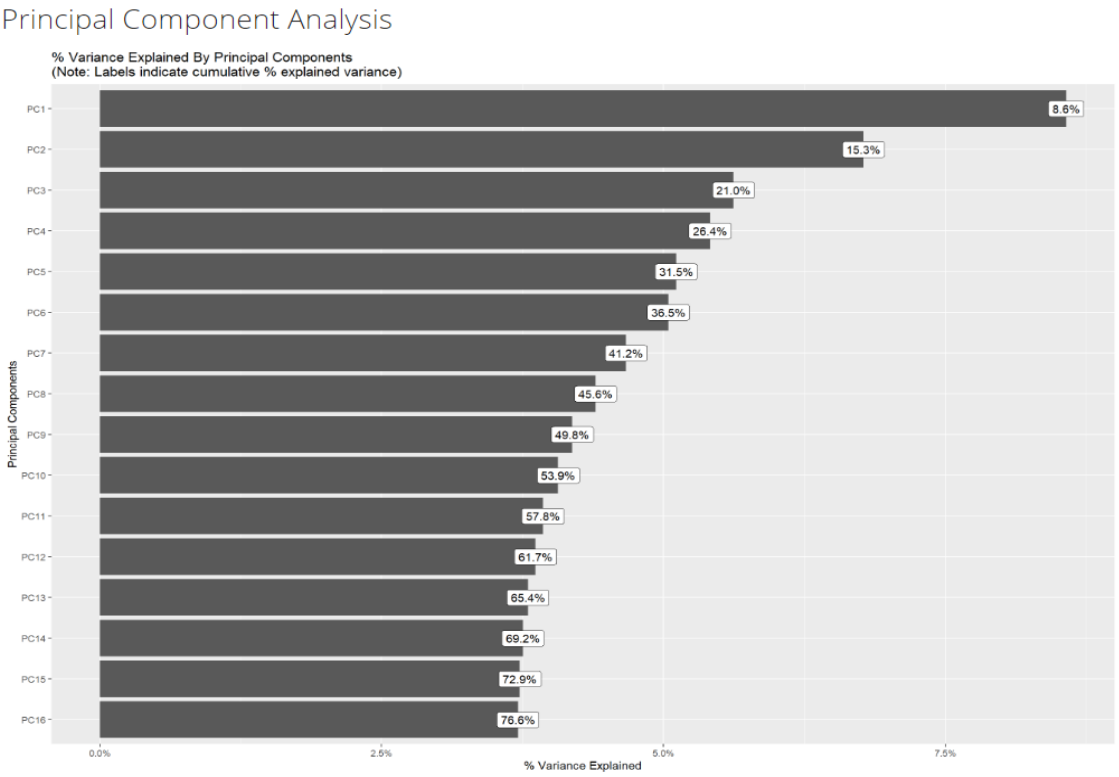


Figura 9 Componentes Principales

Por otro lado, mediante análisis de componentes principales se evidencia que no parece haber niveles de asociación fuerte entre las variables que ameriten la aplicación de técnicas de reducción de dimensionalidad, previas a las inherentes a los modelos que utilizan métodos de regularización.

8.2. Construcción de datos

Para incorporar los conceptos de frecuencia de reclamos y severidad conceptos base de una prima pura de riesgo, se procede a construir el "Claim Frequency" y el "Claim Severity" usando campos del dataset como el número de reclamos, la severidad y la exposición provenientes del set de datos, quedando definidas de la forma:

$$Claim\ Frequency = \frac{numclaims}{exposure}$$
$$Claim\ Severity = \frac{claimsct0}{numclaims} \quad (8.2)$$

Por otro lado, al incluir variables categóricas como predictores en los modelos de Machine Learning, se aplicará numerización n a n, también conocida como "one hot encoding". Una vez aplicada dicha técnica, se aumenta la dimensionalidad del set de datos de 11 variables iniciales a 28:

```
##  [1] "veh_value"       "exposure"        "claim_frequency" "veh_body.SEDAN"
##  [5] "veh_body.HBACK"  "veh_body.STNWG"  "veh_body.UTE"    "veh_body.TRUCK"
##  [9] "veh_body.HDTOP"  "veh_body.COUPE"  "veh_body.PANVN"  "veh_body.MIBUS"
## [13] "veh_body.MCARA"  "veh_body.CONVT"  "veh_body.BUS"    "veh_body.RDSTR"
## [17] "veh_age.3"       "veh_age.4"       "veh_age.2"       "veh_age.1"
## [21] "gender.F"        "gender.M"        "area.C"          "area.A"
## [25] "area.B"          "area.D"          "area.E"          "area.F"
```

Figura 10 Numerización

Es importante mencionar que en este dataset "dummyzado", los únicos predictores de carácter cuantitativo continuo serán el valor y la edad del vehículo "veh_value" y "veh_age" respectivamente, las demás variables serán dicotómicas.

Por otro lado, para el caso de los modelos GLM y sus variaciones, se decidió conservar el dataset original sin aumentar su dimensionalidad a través de "one hot encoding" debido a la facilidad de interpretación de sus parámetros al ser comparados con el nivel base de los campos categóricos.

8.3. Análisis Exploratorio "Claim Frequency"

La mayoría los predictores de la variable objetivo "Claim Frequency" son variables categóricas, por lo que se muestra a continuación una exploración visual de la potencial variabilidad de la variable objetivo dentro de los niveles de dichos campos para explorar la posible variabilidad que podrían aportar en el proceso predictivo:

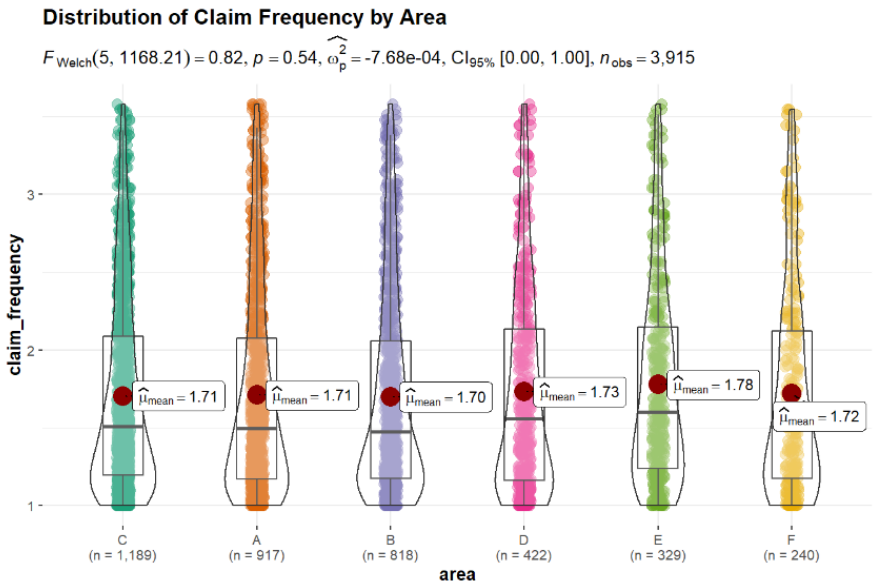


Figura 11. Claim frequency por Area

Se observa que las diferentes ubicaciones geográficas (“Area”) en la que se encuentra el vehículo no parecen explicar la frecuencia de reclamaciones si se analizan de manera aislada

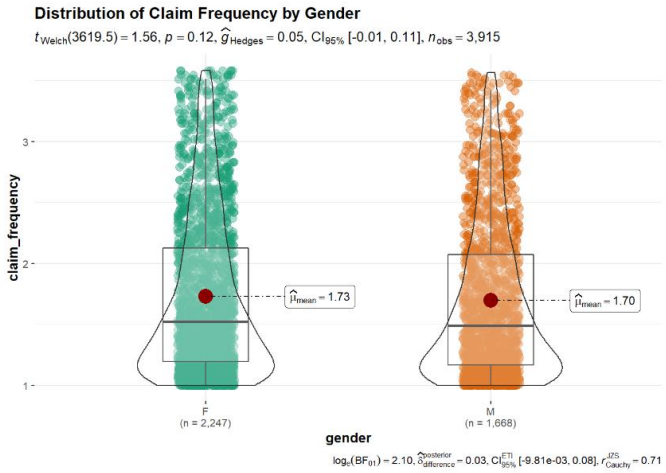


Figura 12. Claim frequency por Genero

De forma similar el género del poseedor del vehículo tampoco parece explicar la frecuencia de reclamación de forma aislada.

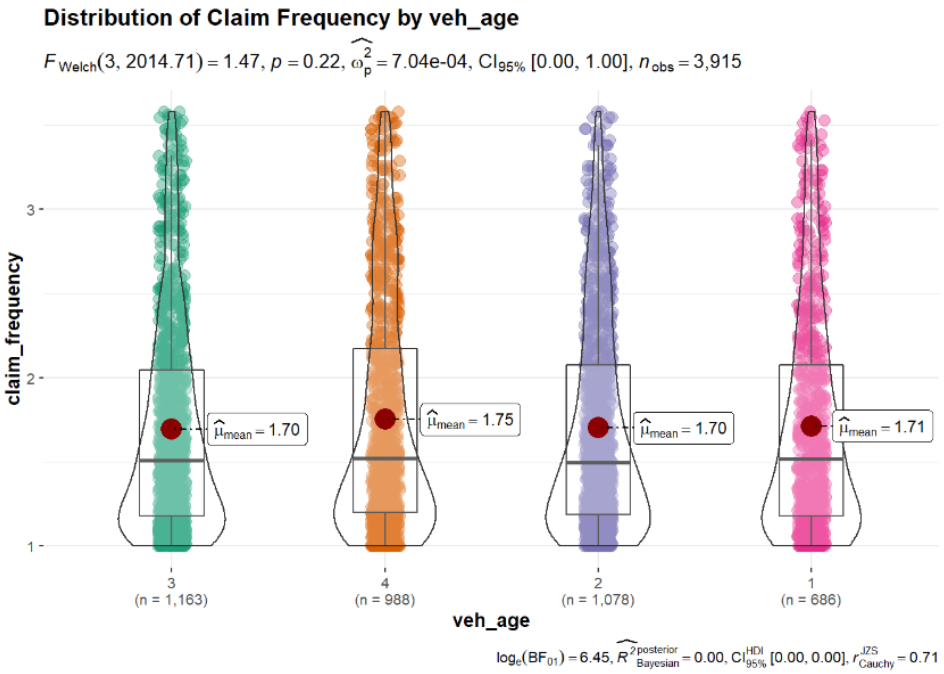


Figura 13. Claim frequency por Edad del vehículo

Aunque los vehículos de mayor edad "4" parecen tener una mayor variabilidad llegando a valores aparentemente más altos de reclamaciones, la variabilidad entre las edades de los vehículos parece ser similar.

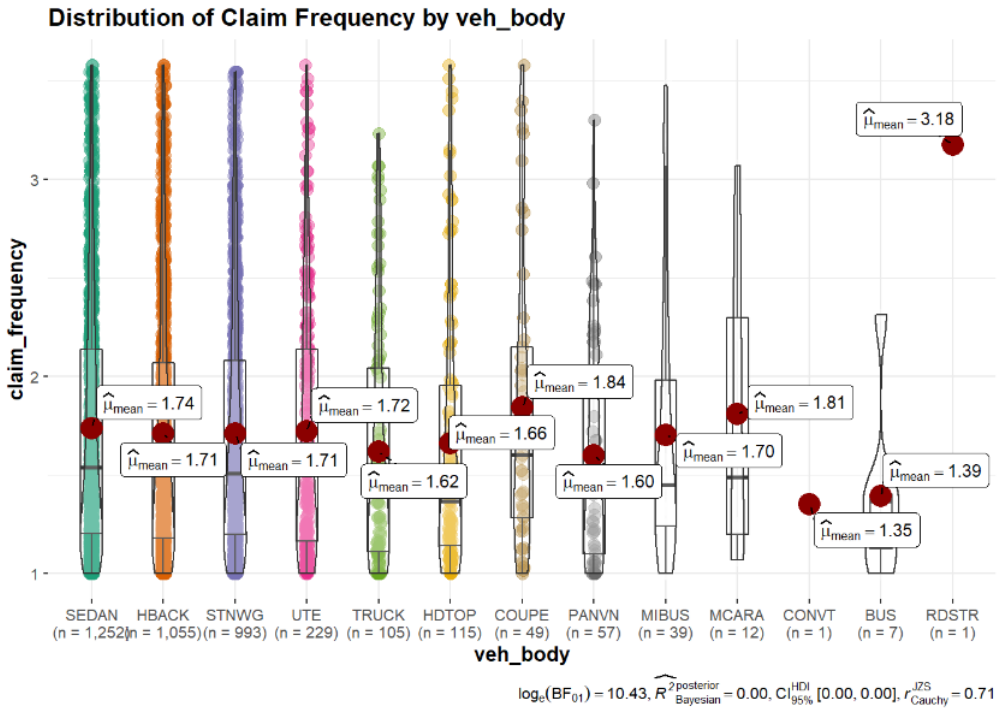


Figura 14. Claim frequency por Tipo de Carrocería

Finalmente, para el caso de la carrocería del vehículo (“veh_body”) se observa mayor heterogeneidad en la distribución de reclamaciones por tipo de carrocería, así como el desbalance en número de reclamaciones. También se observa que para algunos de los tipos de carrocería tales como “Bus”, “CONVT” y “RDSTR” hay escasez de casos registrados.

7.4 Análisis Exploratorio “Claim Severity”

Al explorar la variabilidad de la severidad (“Claim Severity”) en las reclamaciones dentro de los niveles de cada predictor categórico, se observa que dicha variabilidad es menor que la observada para el caso de la frecuencia de reclamaciones (“Claim Frequency”), comportamiento esperado dentro del contexto asegurador en donde la mayoría de las reclamaciones son de bajo monto. Los boxplots para este caso junto con sus métricas de tendencia central se muestran a continuación:

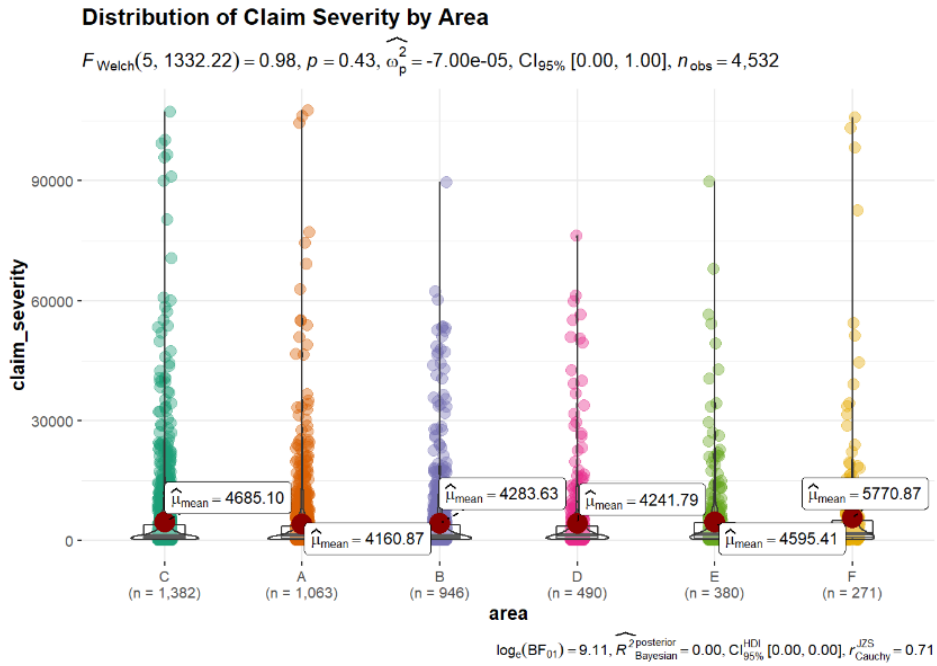


Figura 15. Claim Severity por Área

Se observa un leve incremento en el comportamiento de las medianas de la severidad para las áreas “F” y “C”

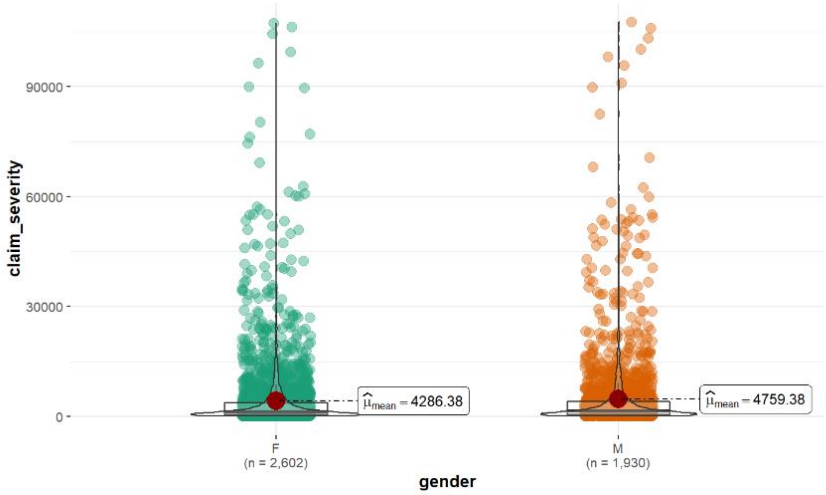


Figura 16. Claim Severity por Género

Respecto al género, se observa que los hombres(“M”) muestran una severidad ligeramente mayor que la de las mujeres(“F”)

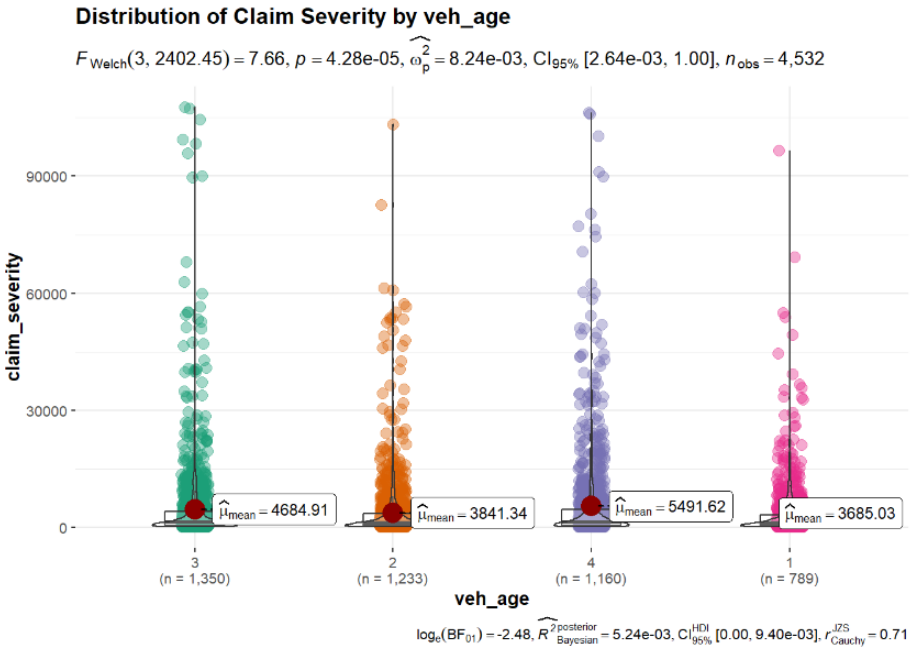


Figura 17. Claim Severity por Edad del Vehículo

Se observa que los vehículos de mayor edad (“4”) tienen ligeramente mayor nivel de severidad que los demás subgrupos.

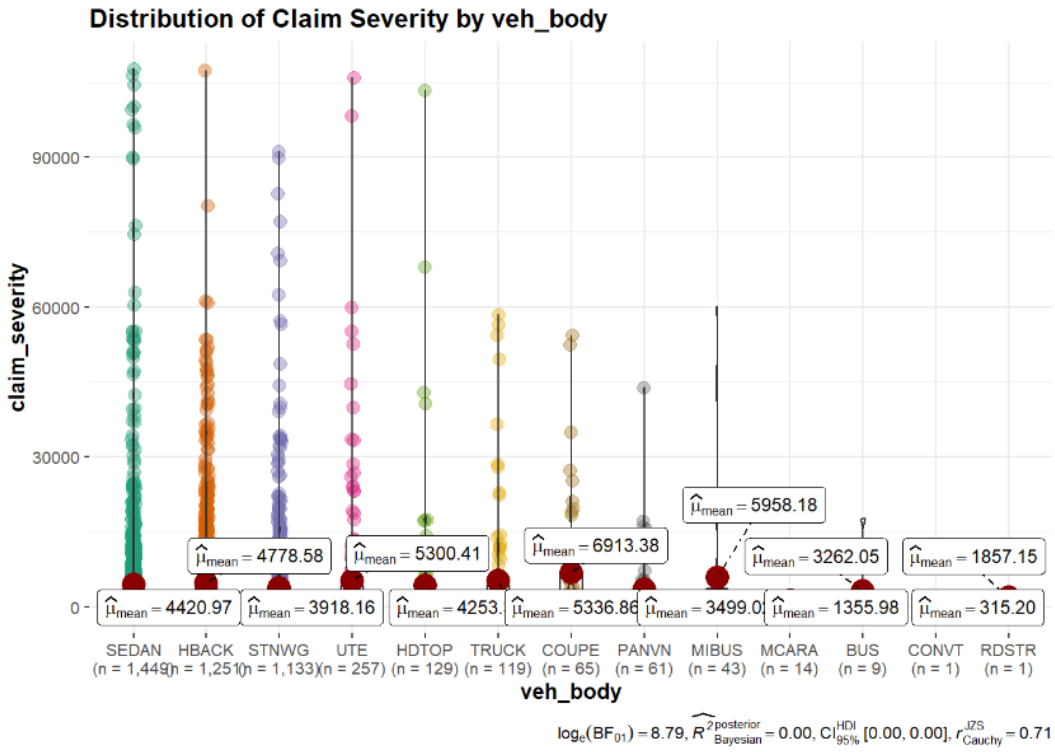


Figura 18. Claim Severity por Tipo de Carrocería

Al igual que para el caso de la frecuencia de reclamaciones (“Claim Frequency”) el caso de la carrocería del vehículo(“veh_body”) muestra mayor heterogeneidad en la distribución de severidad por tipo de carrocería, así como desbalance en número de reclamaciones. Dentro del proceso exploratorio, se aplicaron pruebas de independencia y contraste para explorar posibles interacciones entre las variables predictoras, encontrando interacciones entre algunas de ellas, que fueron capturadas por los modelos.

Finalmente, se aplicó un cambio de nivel base para garantizar que dicho nivel estuviera representado por el del mayor número de casos propendiendo mejorar la confianza en las estimaciones, principalmente para los modelos tipo GLM.

8.4.Tratamiento de datos atípicos

Como se observó en los boxplots anteriores, tanto para “Claim Severity” como para “Claim Frequency” se identificaron datos atípicos. Al analizar su baja proporción respecto al total de datos (6%) y con el propósito de evitar procesos de imputación que pudieran alterar la distribución natural del dataset “dataCar” para futuros procesos de reproducibilidad de resultados, se procedió a la construcción de una función para su remoción.

9. Modelamiento

8.1 Seleccionar técnicas de modelado

La metodología escogida para seleccionar los subconjuntos de entrenamiento y validación fue la de “hold out” del 20% para testeo. Con el 80% restante se aplicó “k-fold closs validation” para entrenar los modelos:

```
set.seed(1234)
#define the number of subsets (or "folds") to use
train_control <- trainControl(method = "cv", number = 5)
```

Figura 19. Estrategia de Muestreo Replicable

Dentro de cada tipo de modelo se corrieron en realidad múltiples submodelos adicionales ya que se usó “grid search” para hacer calibración de los hiperparámetros para los casos en donde era posible. En la siguiente sección se muestran los resultados de cada uno de los modelos, los cuales serán comentados en la sección posterior (Sección 9.3):

8.2 GLM

8.2.1 Claim Frequency GLM

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.56461    0.11715 -30.428  < 2e-16 ***
veh_value     0.25073    0.05151   4.867 1.13e-06 ***
veh_age      -0.04232    0.03476  -1.218   0.2234
veh_bodyHBACK  0.04891    0.08960   0.546   0.5851
veh_bodySTNWG -0.11114    0.10493  -1.059   0.2895
veh_bodyUTE  -0.14000    0.14911  -0.939   0.3478
veh_bodyTRUCK  0.13746    0.20245   0.679   0.4972
veh_bodyHDTOP  0.32591    0.18171   1.794   0.0729 .
veh_bodyCOUPE  0.12595    0.33881   0.372   0.7101
veh_bodyPANVN  0.47346    0.23899   1.981   0.0476 *
veh_bodyMIBUS -0.36608    0.38385  -0.954   0.3402
veh_bodyMCARA  0.52845    0.58452   0.904   0.3660
veh_bodyCONVT -11.08521  320.81004  -0.035   0.9724
veh_bodyBUS    0.45623    1.00339   0.455   0.6493
veh_bodyRDSTR -11.59923  312.35873  -0.037   0.9704
genderM       0.05233    0.07021   0.745   0.4561
areaA         0.04492    0.09179   0.489   0.6245
areaB         0.13468    0.09438   1.427   0.1536
areaD        -0.07481    0.11800  -0.634   0.5261
areaE        -0.26113    0.14406  -1.813   0.0699 .
areaF         0.06799    0.15121   0.450   0.6529
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5889.4  on 46551  degrees of freedom
Residual deviance: 5839.3  on 46531  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 12
```

```
Generalized Linear Model

46552 samples
    6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 37241, 37242, 37241, 37242, 37242
Resampling results:

  RMSE       Rsquared      MAE
  0.1535802  0.0004671627  0.06062978
```

Figura 20 Salida GLM para Claim Frequency

Se observa que muy pocas de las variables y de sus niveles resultan realmente significativos desde el punto de vista estadístico para el modelo.

8.2.2 Claim Severity GLM

```
Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.969177   0.088450  90.098  < 2e-16 ***
veh_value     -0.130694   0.033617  -3.888 0.000103 ***
veh_age        0.027621   0.022431   1.231 0.218266
veh_bodyHBACK  0.068727   0.063688   1.079 0.280622
veh_bodySTNWG  0.089521   0.070578   1.268 0.204752
veh_bodyUTE    0.058633   0.108136   0.542 0.587709
veh_bodyHDTOP  0.023740   0.135350   0.175 0.860782
veh_bodyTRUCK -0.102213   0.149621  -0.683 0.494567
veh_bodyCOUPE  0.155509   0.203605   0.764 0.445060
veh_bodyPANVN -0.075852   0.193108  -0.393 0.694499
veh_bodyMIBUS  0.027615   0.243633   0.113 0.909762
veh_bodyMCARA -0.395566   0.425072  -0.931 0.352144
veh_bodyBUS    0.557841   0.684652   0.815 0.415263
veh_bodyCONVT -2.011639   1.193332  -1.686 0.091952 .
veh_bodyRDSTR         NA         NA      NA       NA
genderM        0.043313   0.050079   0.865 0.387167
areaA         -0.005861   0.065374  -0.090 0.928568
areaB         -0.049400   0.068848  -0.718 0.473109
areaD         -0.066756   0.086129  -0.775 0.438361
areaE          0.114147   0.092965   1.228 0.219601
areaF          0.109049   0.104945   1.039 0.298843
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 1.04922)

    Null deviance: 2290.2  on 3000  degrees of freedom
Residual deviance: 2259.1  on 2981  degrees of freedom
AIC: 32964

Number of Fisher Scoring iterations: 6
```

```
Generalized Linear Model

3001 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 2401, 2401, 2401, 2401, 2400
Resampling results:

  RMSE      Rsquared     MAE
  3850.144  0.002176359  2522.912
```

Figura 21 Salida GLM para Claim Severity

Al igual que para claim frequency, se observa que muy pocas de las variables y de sus niveles resultan realmente significativos desde el punto de vista estadístico para el modelo.

8.3 GAMs

8.3.1 Claim Frequency GAM

```
Family: poisson
Link function: log

Formula:
.outcome ~ veh_bodyHBACK + veh_bodySTNWG + veh_bodyUTE + genderM +
    areaA + areaB + areaD + areaE + areaF + veh_age + s(veh_value)

Parametric coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.66331    0.10459 -35.024  < 2e-16 ***
veh_bodyHBACK  0.02935    0.08266   0.355  0.72252
veh_bodySTNWG -0.18874    0.09181  -2.056  0.03981 *
veh_bodyUTE   -0.23523    0.13665  -1.721  0.08518 .
genderM        0.08772    0.06561   1.337  0.18127
areaA          0.04783    0.08733   0.548  0.58392
areaB          0.15369    0.08971   1.713  0.08669 .
areaD         -0.04549    0.11161  -0.408  0.68361
areaE         -0.22388    0.13612  -1.645  0.10002
areaF          0.13584    0.14308   0.949  0.34242
veh_age       -0.10041    0.03281  -3.061  0.00221 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
              edf Ref.df Chi.sq p-value
s(veh_value) 3.927      9  41.01  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.000929   Deviance explained = 0.765%
UBRE = -0.83251  Scale est. = 1         n = 46552
```

Se observa que este modelo selecciona menos variables que las descritas por el GLM para claim frequency.

```
Generalized Additive Model using Splines

46552 samples
    6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 37242, 37241, 37242, 37242, 37241
Resampling results across tuning parameters:

  method   select  RMSE       Rsquared      MAE
  GCV.Cp   TRUE    0.1521446  0.0006281296  0.04167885
  GCV.Cp   FALSE   0.1521446  0.0006281296  0.04167885
  GACV.Cp  TRUE    0.1521446  0.0006281296  0.04167885
  GACV.Cp  FALSE   0.1521446  0.0006281296  0.04167885

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were select = TRUE and method = GCV.Cp.
```

Figura 22. Salida GAM para Claim Frequency

8.3.2 Claim Severity GAM

```
Family: Gamma
Link function: log

Formula:
.outcome ~ veh_bodyHBACK + veh_bodySTNWG + veh_bodyUTE + genderM +
    areaA + areaB + areaD + areaE + areaF + veh_age + s(veh_value)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.844035   0.071466 109.758   <2e-16 ***
veh_bodyHBACK  0.121724   0.058800   2.070   0.0385 *
veh_bodySTNWG  0.054390   0.064373   0.845   0.3982
veh_bodyUTE    0.045412   0.103543   0.439   0.6610
genderM        0.044516   0.048202   0.924   0.3558
areaA          0.007888   0.063626   0.124   0.9013
areaB         -0.015648   0.066679  -0.235   0.8145
areaD         -0.019556   0.082642  -0.237   0.8130
areaE          0.138326   0.088756   1.558   0.1192
areaF          0.099772   0.105338   0.947   0.3436
veh_age        0.029361   0.021750   1.350   0.1771
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
              edf Ref.df     F p-value
s(veh_value) 1.856      9 1.983 3.8e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.00679   Deviance explained = 1.42%
-REML =  27011  Scale est. = 1.6093    n = 3001
```

Se observa que este modelo selecciona menos variables que las descritas por el GLM para claim severity.

```
Generalized Additive Model using Splines

3001 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 2401, 2401, 2400, 2401, 2401
Resampling results across tuning parameters:

  method  select  RMSE      Rsquared     MAE
  GCV.Cp  TRUE    3822.099  0.004765577  2657.859
  GCV.Cp  FALSE   3822.099  0.004765577  2657.859
  REML    TRUE    3821.958  0.004642137  2658.079
  REML    FALSE   3821.958  0.004642137  2658.079
  ML      TRUE    3821.964  0.004635556  2658.082
  ML      FALSE   3821.964  0.004635556  2658.082

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were select = TRUE and method = REML.
```

Figura 23. Salida GAM para Claim Severity

8.4 MARS

8.4.1 Claim Frequency MARS

```
                              coefficients
(Intercept)                     0.02407549
h(3-veh_age)                    0.01460576
h(3-veh_age) * veh_bodySTNWG   -0.00501391
h(veh_value-3.97) * h(3-veh_age)  0.65436497
h(4.06-veh_value) * h(veh_age-3) -0.00302031
h(4.06-veh_value) * h(3-veh_age) -0.00589784

Selected 6 of 25 terms, and 3 of 20 predictors (nprune=6)
Termination condition: Reached nk 41
Importance: veh_value, veh_age, veh_bodySTNWG, veh_bodyHBACK-unused, veh_bodyUTE-unused, ...
Number of terms at each degree of interaction: 1 1 4
GCV 0.02314576    RSS 1076.856    GRSq 0.0007923563    RSq 0.001328904
```

Se observa la partición del modelo en funciones lineales para el valor y la edad del vehículo y las interacciones con los niveles de las variables cualitativas.

```
Multivariate Adaptive Regression Spline

46552 samples
    6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 37241, 37242, 37242, 37241, 37242
Resampling results across tuning parameters:

  degree  nprune  RMSE       Rsquared      MAE
  1       4       0.1521826  0.0004714868  0.04171276
  1       5       0.1521759  0.0005134780  0.04169701
  1       6       0.1521824  0.0005510276  0.04169445
  2       4       0.1521191  0.0014476004  0.04165153
  2       5       0.1521247  0.0014841868  0.04164822
  2       6       0.1520672  0.0023016506  0.04163373
  3       4       0.1522542  0.0003951499  0.04170818
  3       5       0.1522892  0.0003861614  0.04172696
  3       6       0.1523043  0.0003073750  0.04170137

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were nprune = 6 and degree = 2.
```

Figura 24. Salida MARS para Claim Frequency

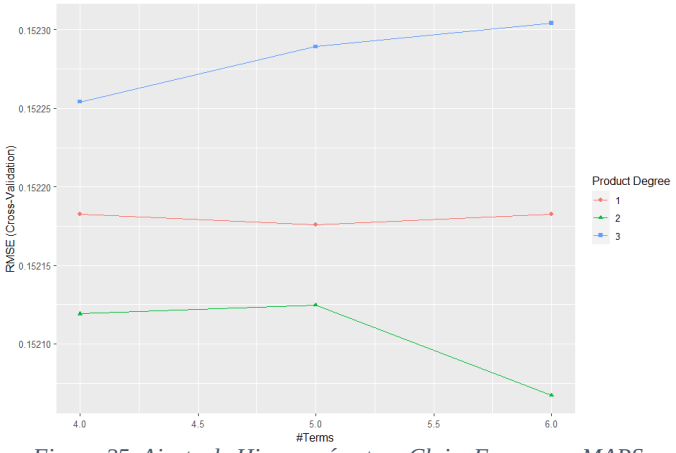


Figura 25. Ajuste de Hiperparámetros Claim Frequency MARS

8.4.2 Claim Severity MARS

```
Call: earth(x=matrix[3001,20], y=c(812,2769,6382...), keepxy=TRUE, degree=3, nprune=5,
            thresh=1e-08)

                                      coefficients
(Intercept)                             3164.0345
h(veh_value-0.7)                        -240.4066
veh_bodyBUS * areaD                    14590.5958
h(0.99-veh_value) * veh_age              785.5207
h(0.99-veh_value) * veh_bodyCOUPE * areaA  20254.8921

Selected 5 of 25 terms, and 6 of 20 predictors (nprune=5)
Termination condition: Reached nk 41
Importance: veh_value, veh_bodyCOUPE, areaA, veh_age, veh_bodyBUS, areaD, ...
Number of terms at each degree of interaction: 1 1 2 1
GCV 14494589    RSS 43179965294    GRSq 0.01359445    RSq 0.02015952
```

Se observa la partición del modelo en funciones lineales para el valor del vehículo y las interacciones con los niveles de las variables cualitativas.

```
Multivariate Adaptive Regression Spline

3001 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 2401, 2401, 2401, 2401, 2400
Resampling results across tuning parameters:

  degree  nprune  RMSE      Rsquared     MAE
  1       4       3816.385  0.008530914  2650.992
  1       5       3819.132  0.007917128  2652.653
  1       6       3819.132  0.007917128  2652.653
  2       4       3822.221  0.007086231  2652.435
  2       5       3824.199  0.006997724  2654.588
  2       6       3822.582  0.007635415  2656.479
  3       4       3814.393  0.009212819  2648.778
  3       5       3811.537  0.010930547  2640.996
  3       6       3824.213  0.007325737  2650.885

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were nprune = 5 and degree = 3.
```

Figura 26. Salida MARS para Claim Severity

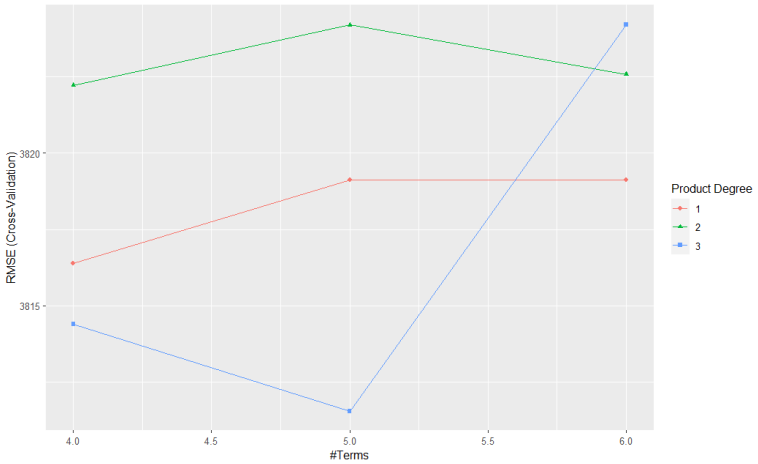


Figura 27. Ajuste de Hiperparámetros Claim Severity MARS

8.5 GLM Elastic Networks

8.5.1 Claim Frequency Elastic Network

```
glmnet

46552 samples
    6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 37241, 37242, 37241, 37242, 37242
Resampling results across tuning parameters:

  alpha  lambda  RMSE       Rsquared      MAE
  0e+00  0e+00   3.2567634  0.0004783407  3.24433706
  0e+00  1e-05   3.2567634  0.0004783407  3.24433706
  0e+00  5e-05   3.2567634  0.0004783407  3.24433706
  0e+00  7e-05   3.2567634  0.0004783407  3.24433706
  0e+00  8e-05   3.2567634  0.0004783407  3.24433706
  0e+00  9e-05   0.1535710  0.0004664768  0.06063774
  1e-04  0e+00   3.2567600  0.0004783791  3.24433484
  1e-04  1e-05   3.2567600  0.0004783791  3.24433484
  1e-04  5e-05   3.2567600  0.0004783791  3.24433484
  1e-04  7e-05   3.2567600  0.0004783791  3.24433484
  1e-04  8e-05   3.2567600  0.0004783791  3.24433484
  1e-04  9e-05   0.1535710  0.0004664819  0.06063775
  2e-01  0e+00   3.2610352  0.0004183323  3.24676563
  2e-01  1e-05   3.2610352  0.0004183323  3.24676563
  2e-01  5e-05   3.2596008  0.0004417016  3.24609368
  2e-01  7e-05   3.2589512  0.0004511245  3.24574608
  2e-01  8e-05   3.2586850  0.0004548284  3.24559651
  2e-01  9e-05   0.1535765  0.0004657808  0.06063339
  7e-01  0e+00   3.2611039  0.0004159642  3.24679769
  7e-01  1e-05   3.2606563  0.0004236283  3.24660716
  7e-01  5e-05   3.2574585  0.0004693357  3.24486588
  7e-01  7e-05   3.2566258  0.0004780289  3.24432002
  7e-01  8e-05   3.2562644  0.0004812618  3.24407341
  7e-01  9e-05   0.1535705  0.0004621629  0.06063982
  9e-01  0e+00   3.2611296  0.0004157964  3.24680848
  9e-01  1e-05   3.2603008  0.0004297855  3.24644236
  9e-01  5e-05   3.2569001  0.0004752838  3.24450504
  9e-01  7e-05   3.2559915  0.0004834695  3.24388559
  9e-01  8e-05   3.2555900  0.0004865328  3.24360193
  9e-01  9e-05   0.1535681  0.0004610527  0.06064237

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0.9 and lambda = 9e-05.
```

Figura 28. Salida Elastic Network para Claim frequency

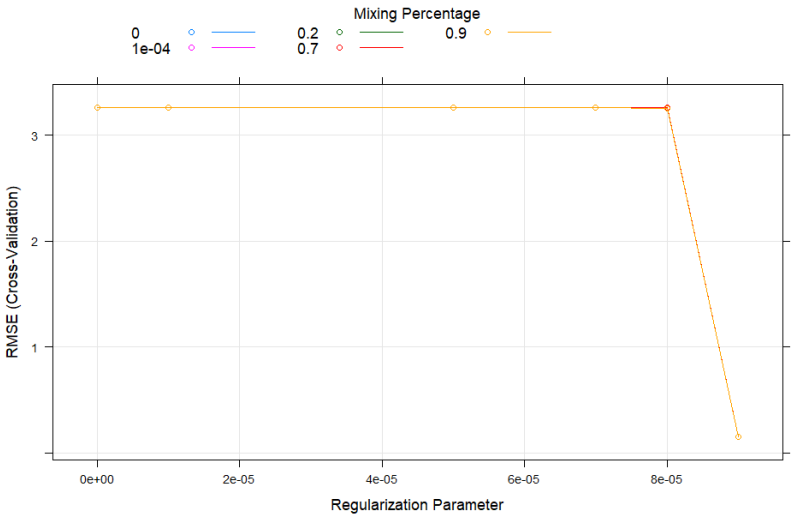


Figura 29. Ajuste de Hiperparámetros Claim Frequency Elastic Networks.

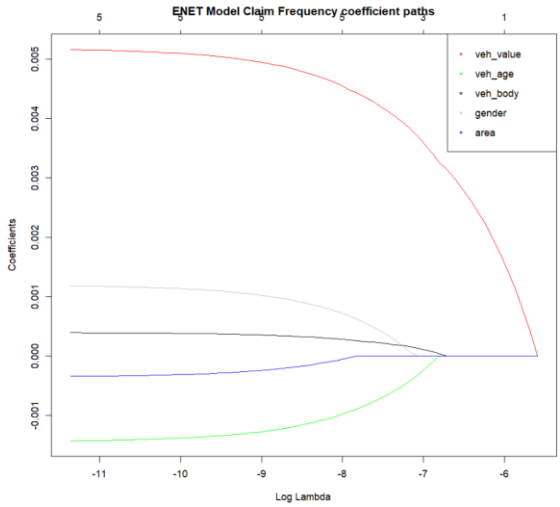


Figura 30. Coeficientes versus penalización para claim frequency

Se observa que el valor del vehículo es la variable que tarda más en llegar a cero cuando lambda crece.

8.5.2 Claim Severity Elastic Network

```
glmnet

3001 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 2401, 2401, 2400, 2401, 2401
Resampling results across tuning parameters:

  alpha  lambda  RMSE      Rsquared     MAE
  0e+00  0e+00   4846.478  0.004349030  2978.552
  0e+00  1e-06   4846.478  0.004349030  2978.552
  0e+00  5e-06   4846.478  0.004349030  2978.552
  0e+00  7e-06   4846.478  0.004349030  2978.552
  0e+00  8e-06   4846.478  0.004349030  2978.552
  0e+00  9e-06   3840.341  0.004075754  2518.689
  1e-04  0e+00   4846.478  0.004348104  2978.552
  1e-04  1e-06   4846.478  0.004348104  2978.552
  1e-04  5e-06   4846.478  0.004348104  2978.552
  1e-04  7e-06   4846.478  0.004348104  2978.552
  1e-04  8e-06   4846.478  0.004348104  2978.552
  1e-04  9e-06   3840.362  0.004074446  2518.711
  2e-01  0e+00   4846.478  0.004498556  2978.552
  2e-01  1e-06   4846.478  0.004498556  2978.552
  2e-01  5e-06   4846.478  0.004498556  2978.552
  2e-01  7e-06   4846.478  0.004498556  2978.552
  2e-01  8e-06   4846.478  0.004498556  2978.552
  2e-01  9e-06   3839.981  0.004264201  2518.464
  7e-01  0e+00   4846.478  0.004496888  2978.552
  7e-01  1e-06   4846.478  0.004496888  2978.552
  7e-01  5e-06   4846.478  0.004496888  2978.552
  7e-01  7e-06   4846.478  0.004496888  2978.552
  7e-01  8e-06   4846.478  0.004496888  2978.552
  7e-01  9e-06   3839.976  0.004264808  2518.469
  9e-01  0e+00   4846.478  0.004496504  2978.552
  9e-01  1e-06   4846.478  0.004496504  2978.552
  9e-01  5e-06   4846.478  0.004496504  2978.552
  9e-01  7e-06   4846.478  0.004496504  2978.552
  9e-01  8e-06   4846.478  0.004496504  2978.552
  9e-01  9e-06   3839.983  0.004264254  2518.488

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0.7 and lambda = 9e-06.
```

Figura 31. Salida Elastic Network Para Claim Severity

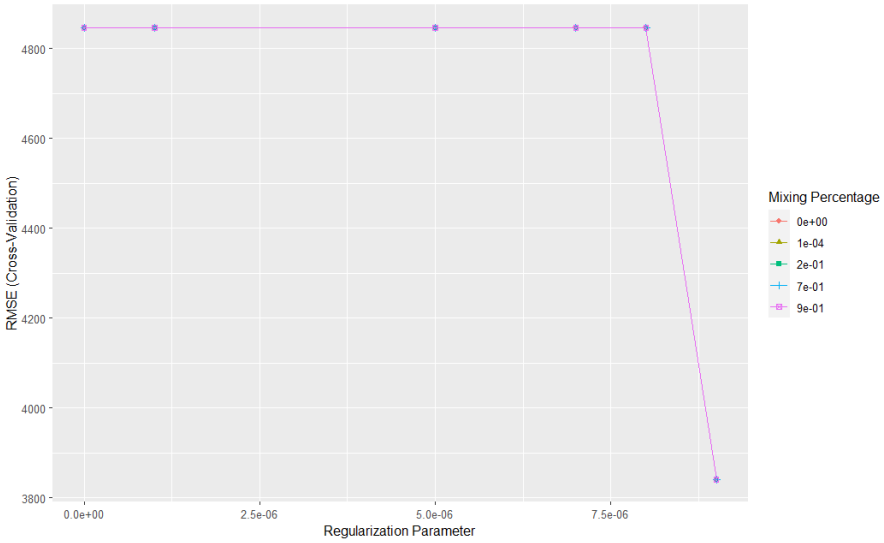


Figura 32. Ajuste de Hiperparámetros Claim Severity Elastic Networks

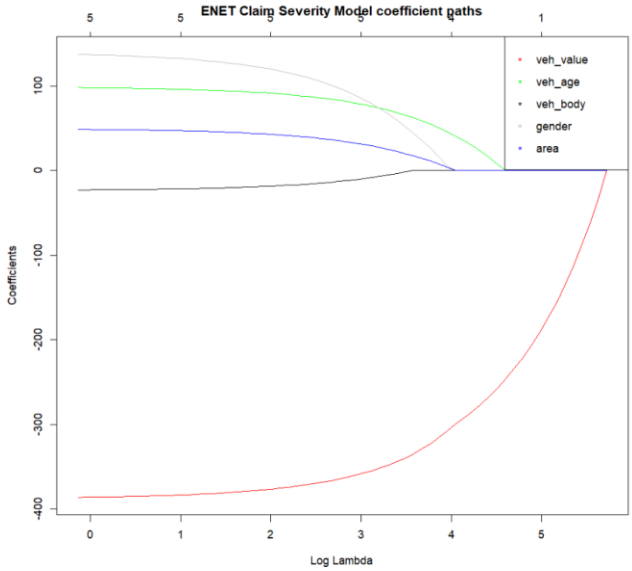


Figura 33. Coeficientes versus penalización para claim severity.

Al igual que para el caso de claim frequency, se observa que el valor del vehículo es la variable que tarda más en llegar a cero cuando lambda crece.

8.6 Bagging (Random Forest)

8.6.1 Claim Frequency Random Forest

```
Random Forest

46552 samples
   24 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 37241, 37242, 37241, 37242, 37242
Resampling results across tuning parameters:

  mtry  splitrule  min.node.size  RMSE       Rsquared      MAE
   5    variance     10           0.1543832  0.0006530299  0.05970334
   5    variance    100           0.1539944  0.0006208078  0.06009657
   5    variance   1000           0.1536082  0.0004794393  0.06043642
   5    variance   4600           0.1534643  0.0004922753  0.06062321
   5    maxstat      10           0.1533958  0.0005834956  0.06072661
   5    maxstat     100           0.1534082  0.0005050086  0.06076682
   5    maxstat    1000           0.1534070  0.0005240452  0.06075998
   5    maxstat    4600           0.1534171  0.0004076602  0.06079627
  15    variance     10           0.1682961  0.0009012019  0.05722637
  15    variance    100           0.1561292  0.0013007642  0.05902999
  15    variance   1000           0.1537264  0.0009476504  0.06000149
  15    variance   4600           0.1535089  0.0007276000  0.06038416
  15    maxstat      10           0.1535472  0.0010271542  0.05987555
  15    maxstat     100           0.1535075  0.0010221739  0.06024982
  15    maxstat    1000           0.1535098  0.0006070934  0.06057033
  15    maxstat    4600           0.1534719  0.0005025503  0.06063713
  23    variance     10           0.1739748  0.0009896848  0.05757010
  23    variance    100           0.1568044  0.0014135586  0.05901117
  23    variance   1000           0.1538433  0.0009847006  0.05988558
  23    variance   4600           0.1535454  0.0007603377  0.06035901
  23    maxstat      10           0.1540178  0.0010258272  0.05930426
  23    maxstat     100           0.1536959  0.0010449978  0.05992733
  23    maxstat    1000           0.1535625  0.0006561004  0.06042560
  23    maxstat    4600           0.1534936  0.0005553231  0.06058236

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were mtry = 5, splitrule = maxstat and min.node.size = 10.
```

Figura 34. Salida Random Forest para Claim Frequency

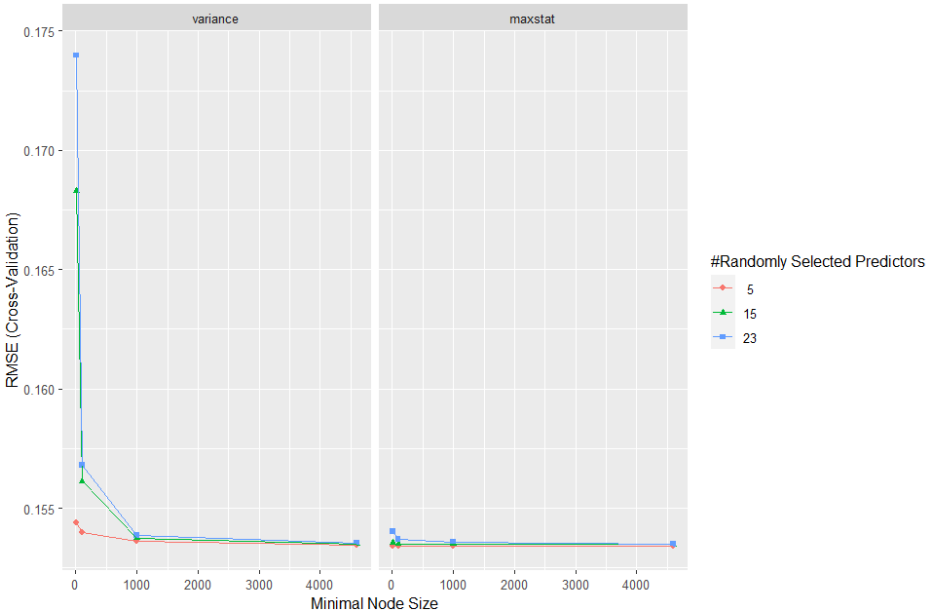


Figura 35.. Ajuste de Hiperparámetros Claim Frequency Random Forest.

8.6.2 Claim Severity Random Forest

```
Random Forest

3001 samples
  24 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 2401, 2401, 2400, 2401, 2401
Resampling results across tuning parameters:

  mtry  splitrule  min.node.size  RMSE      Rsquared     MAE
   5    variance     10           3884.622  0.003306296  2563.318
   5    variance    100           3849.669  0.003615040  2528.952
   5    variance   1000           3838.021  0.004154856  2515.931
   5    variance   2000           3839.063  0.003657192  2515.608
   5    maxstat      10           3840.022  0.004383437  2514.783
   5    maxstat     100           3839.993  0.004771279  2514.911
   5    maxstat    1000           3840.015  0.004938564  2515.217
   5    maxstat    2000           3840.682  0.005458247  2514.762
  15    variance     10           3981.245  0.002509732  2671.732
  15    variance    100           3865.826  0.002347793  2544.750
  15    variance   1000           3838.244  0.004799112  2515.190
  15    variance   2000           3835.635  0.005434929  2512.965
  15    maxstat      10           3839.588  0.005155534  2514.745
  15    maxstat     100           3839.655  0.004669783  2514.020
  15    maxstat    1000           3837.109  0.005126613  2512.425
  15    maxstat    2000           3837.665  0.005000647  2513.058
  23    variance     10           4019.043  0.002201644  2705.997
  23    variance    100           3874.875  0.001965198  2553.908
  23    variance   1000           3839.220  0.004680879  2516.764
  23    variance   2000           3836.593  0.005390526  2514.472
  23    maxstat      10           3842.615  0.004963179  2517.749
  23    maxstat     100           3840.743  0.004906682  2516.254
  23    maxstat    1000           3837.715  0.004857613  2512.708
  23    maxstat    2000           3836.596  0.005841541  2511.612

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were mtry = 15, splitrule = variance
 and min.node.size = 2000.
```

Figura 36. Salida Random Forest para Claim Severity

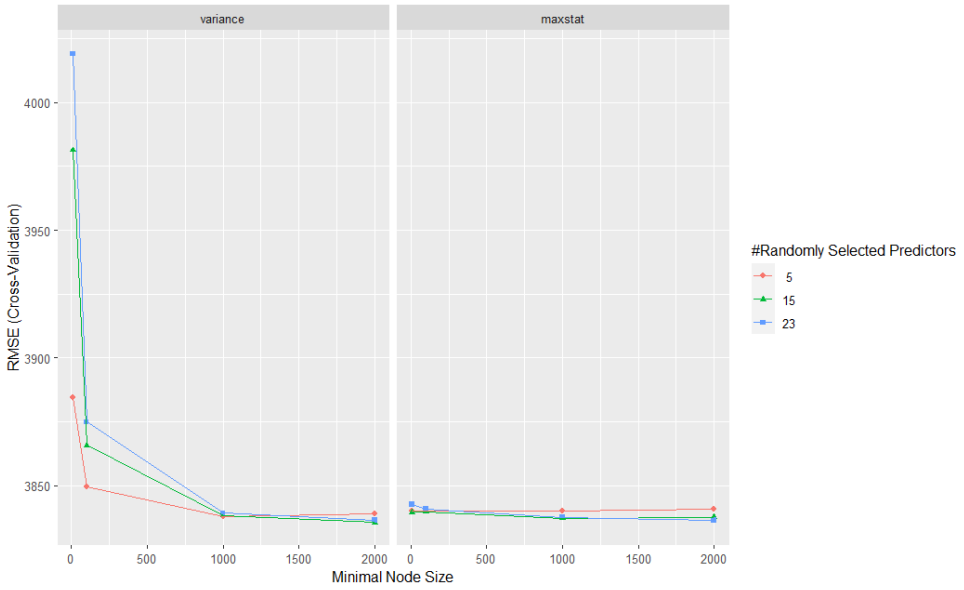


Figura 37. Ajuste de Hiperparámetros Claim Severity Elastic Networks

8.7 Boosting (XG Boost)

8.7.1 Claim Frequency XGBoost

```
eXtreme Gradient Boosting

46552 samples
   24 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 37241, 37242, 37241, 37242, 37242
Resampling results across tuning parameters:

  eta  max_depth  colsample_bytree  subsample  nrounds  RMSE       Rsquared      MAE
  0.3  1          0.6               0.50        50      0.1535651  0.0004868643  0.06035909
  0.3  1          0.6               0.50       100      0.1536679  0.0004744537  0.06048947
  0.3  1          0.6               0.50       150      0.1536681  0.0006115278  0.06058692
  0.3  1          0.6               0.75        50      0.1535798  0.0003731697  0.06068956
  0.3  1          0.6               0.75       100      0.1535677  0.0005493583  0.06041319
  0.3  1          0.6               0.75       150      0.1535999  0.0004539314  0.06031963
                                                ...
                                                ...
                                                ...
  0.4  3          0.8               0.50       150      0.1573253  0.0009005708  0.06167001
  0.4  3          0.8               0.75        50      0.1548478  0.0005872602  0.06017239
  0.4  3          0.8               0.75       100      0.1555005  0.0010250821  0.05974558
  0.4  3          0.8               0.75       150      0.1565196  0.0007699241  0.06059524
  0.4  3          0.8               1.00        50      0.1544087  0.0009490506  0.05996603
  0.4  3          0.8               1.00       100      0.1552138  0.0010778725  0.05991599
  0.4  3          0.8               1.00       150      0.1559754  0.0010358753  0.06009416

Tuning parameter 'gamma' was held constant at a value of 0
Tuning parameter
 'min_child_weight' was held constant at a value of 1
RMSE was used to select the optimal model using the smallest value.
```

Figura 38. Salida XGBoost para Claim Frequency

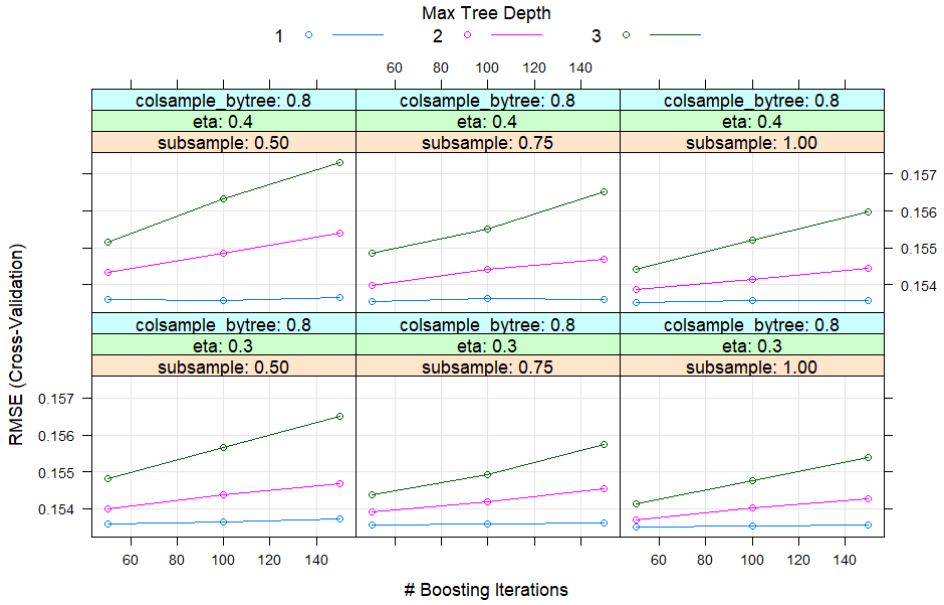


Figura 39. Ajuste de Hiperparámetros Claim Frequency XGBoost.

8.7.2 Claim Severity XGBoost

```
eXtreme Gradient Boosting

3001 samples
  24 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 2401, 2401, 2400, 2401, 2401
Resampling results across tuning parameters:

  eta  max_depth  colsample_bytree  subsample  nrounds  RMSE      Rsquared      MAE
  0.3  1          0.6               0.50        50      3866.638  0.0047659183  2539.118
  0.3  1          0.6               0.50       100      3861.597  0.0054698605  2532.690
  0.3  1          0.6               0.50       150      3861.144  0.0047353327  2531.503
  0.3  1          0.6               0.75        50      3862.494  0.0020022570  2533.003
  0.3  1          0.6               0.75       100      3867.691  0.0032377042  2532.947
  0.3  1          0.6               0.75       150      3875.509  0.0019422185  2533.114
                                                ...
                                                ...
                                                ...
  0.4  3          0.8               0.50       150      4053.184  0.0035786699  2691.985
  0.4  3          0.8               0.75        50      3936.770  0.0020213811  2584.326
  0.4  3          0.8               0.75       100      4007.016  0.0018798591  2646.606
  0.4  3          0.8               0.75       150      4046.034  0.0010274672  2689.826
  0.4  3          0.8               1.00        50      3926.262  0.0016407782  2568.359
  0.4  3          0.8               1.00       100      3980.563  0.0014434211  2612.340
  0.4  3          0.8               1.00       150      4024.752  0.0014185487  2639.927

Tuning parameter 'gamma' was held constant at a value of 0
Tuning
 parameter 'min_child_weight' was held constant at a value of 1
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were nrounds = 50, max_depth = 1, eta = 0.3, gamma
 = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample = 1.
>
```

Figura 40. Salida XGBoost para Claim Severity

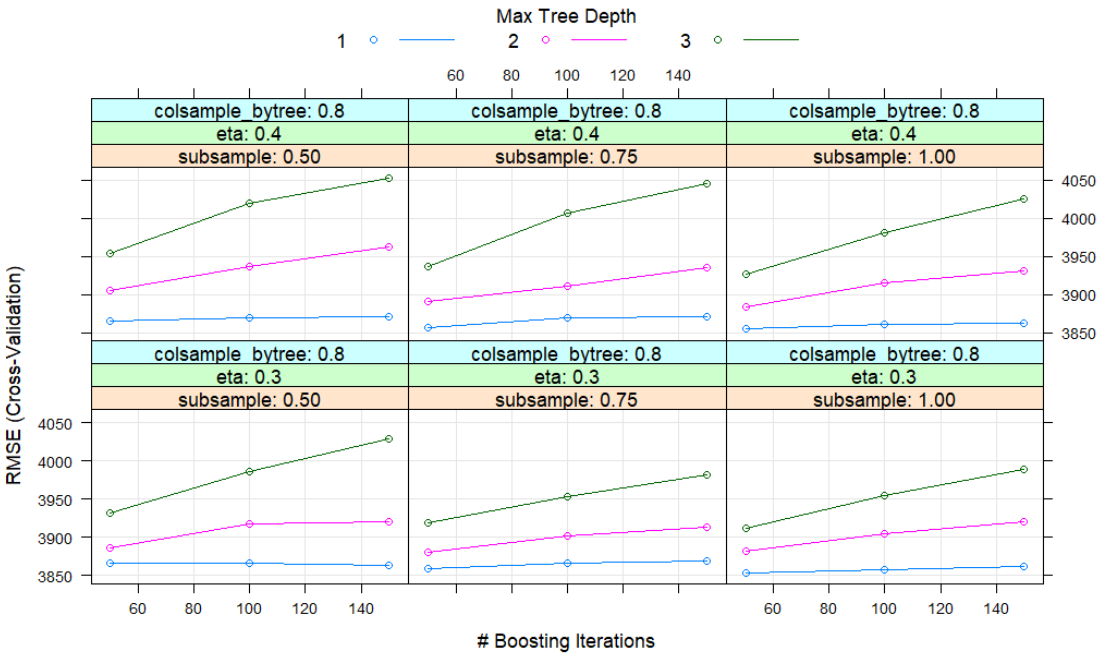


Figura 41. Ajuste de Hiperparámetros Claim Severity XGBoost.

8.8 MLP Deep Learning

8.8.1 Claim Frequency MLP

```
Neural Network

46552 samples
   24 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 37241, 37242, 37241, 37242, 37242
Resampling results across tuning parameters:

  size  decay  RMSE       Rsquared      MAE
  1     0e+00  0.1536097           NaN  0.02126747
  1     1e-04  0.1530182  0.0001674977  0.02947625
  1     1e-01  0.1521008  0.0006654805  0.04220154
  3     0e+00  0.1536097           NaN  0.02126747
  3     1e-04  0.1533492  0.0002178860  0.02544931
  3     1e-01  0.1520994  0.0006867781  0.04202168
  5     0e+00  0.1536097           NaN  0.02126747
  5     1e-04  0.1521580  0.0005772596  0.04159640
  5     1e-01  0.1521001  0.0006757665  0.04199843

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were size = 3 and decay = 0.1.
```

Figura 42. Salida MLP para Claim Frequency

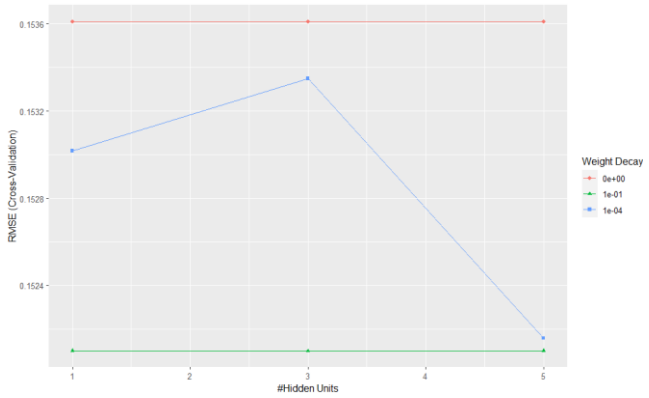


Figura 43. Ajuste de Hiperparámetros Claim Frequency MLP.

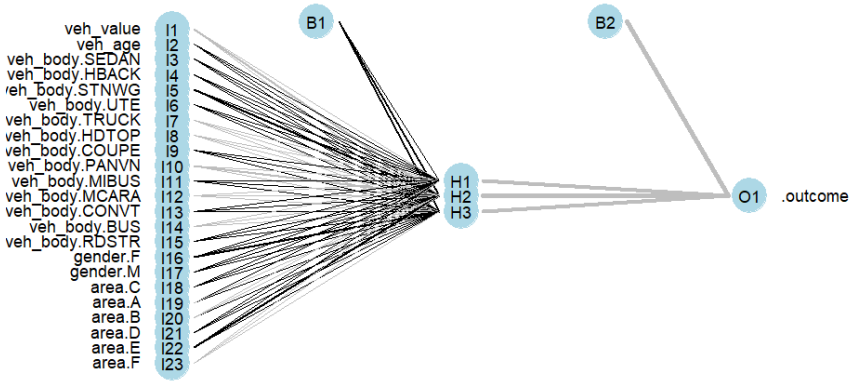


Figura 44. Arquitectura de la Red Neuronal para Claim Frequency.

8.8.2 Claim Severity MLP

```
Neural Network

3001 samples
  24 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 2401, 2401, 2400, 2401, 2401
Resampling results across tuning parameters:

  size  decay  RMSE      Rsquared     MAE
  1     0e+00  4850.704         NaN  2985.409
  1     1e-04  4850.704         NaN  2985.409
  1     1e-01  4850.704  0.003229649  2985.409
  3     0e+00  4850.704         NaN  2985.409
  3     1e-04  4850.704         NaN  2985.409
  3     1e-01  4850.704  0.000172371  2985.409
  5     0e+00  4850.704         NaN  2985.409
  5     1e-04  4850.704         NaN  2985.409
  5     1e-01  4850.704  0.004575475  2985.409

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were size = 1 and decay = 1e-04.
```

Figura 45. Salida MLP para Claim Severity

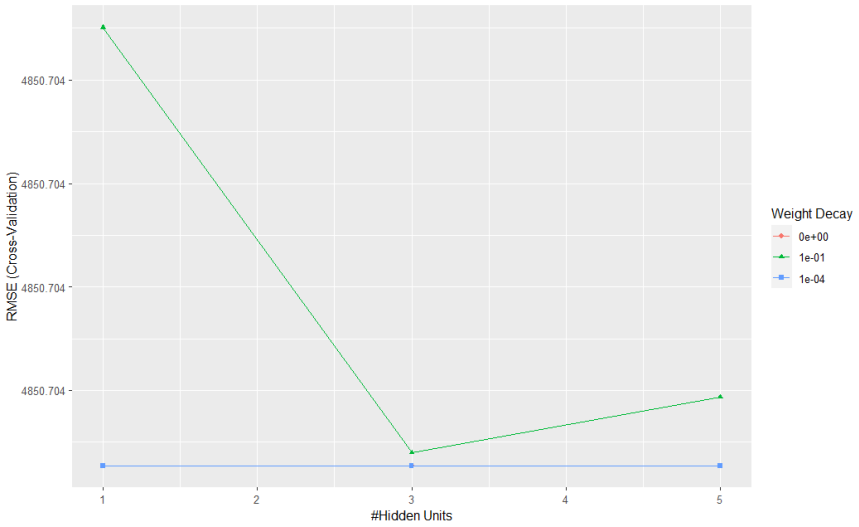


Figura 46. Ajuste de Hiperparámetros Claim Severity MLP.

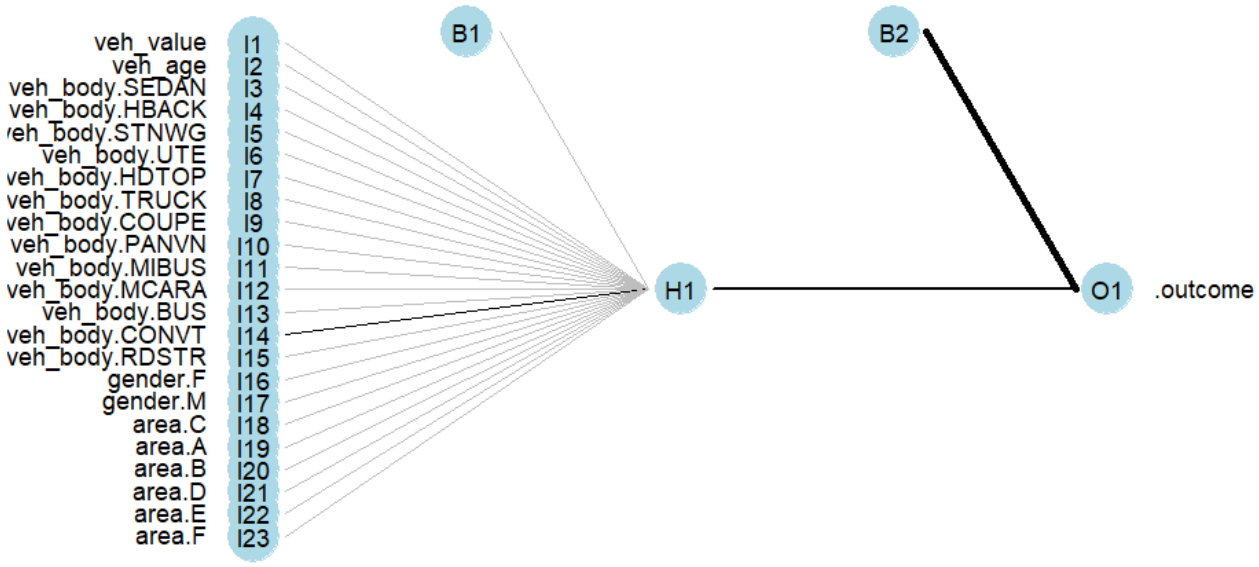


Figura 47. Arquitectura de la Red Neuronal para Claim Severity.

9. Evaluación

A continuación, se muestran las métricas de desempeño y los “lift charts” de cada uno de los modelos para los conjuntos de entrenamiento y validación con el objetivo de tener una descripción visual de la eficiencia con la que los modelos disminuyen la selección adversa. Dado que en cada modelo se corrieron realmente todas las combinaciones de hiperparámetros descritas a través del gridsearch, el mejor de cada tipo se seleccionó con base en el RMSE.

9.1 Lift-Charts Claim Frequency

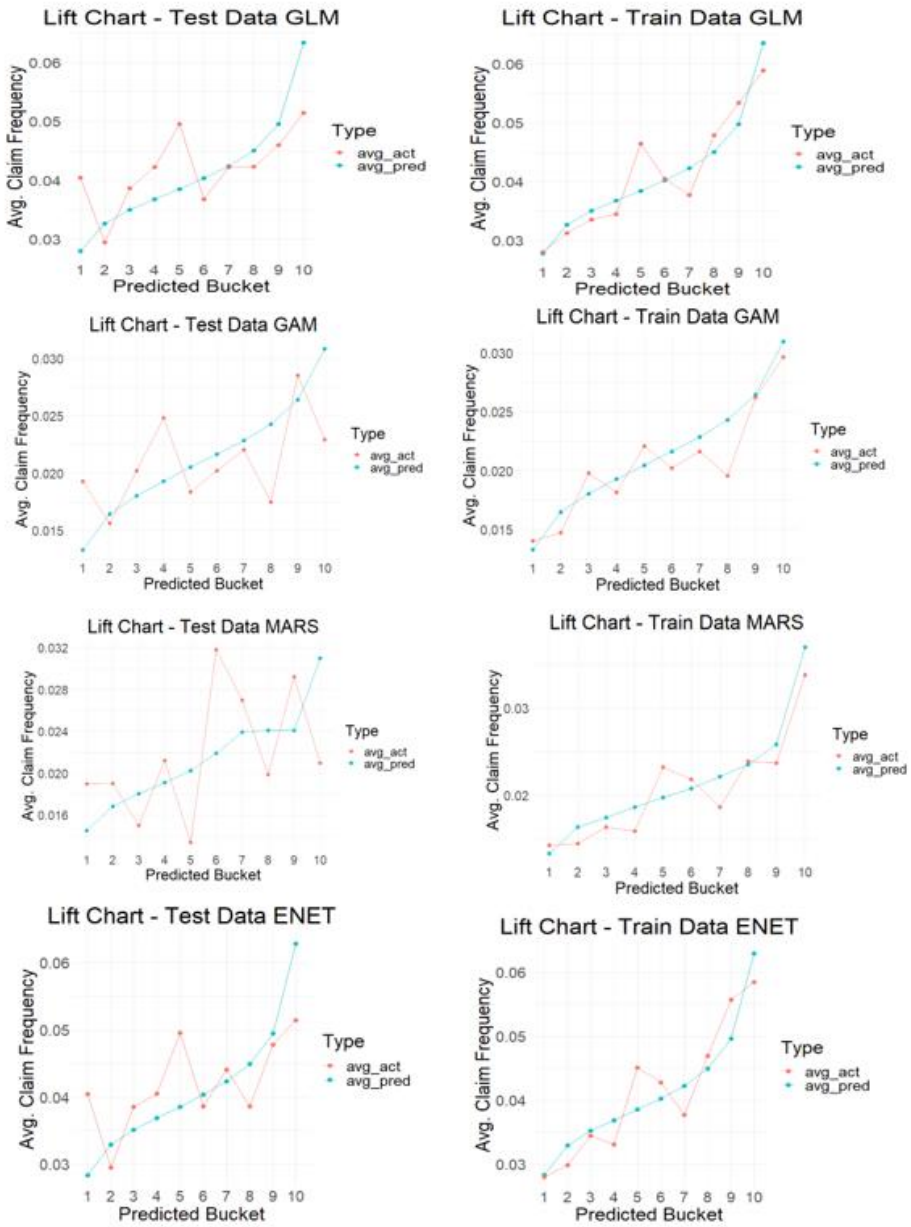
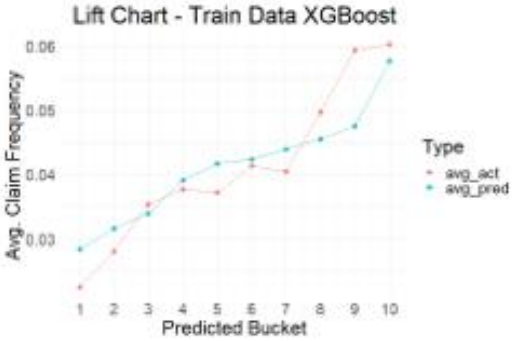
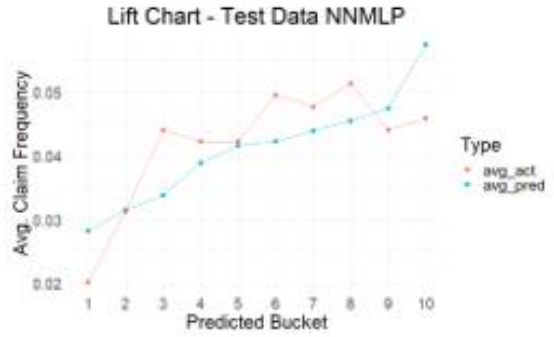
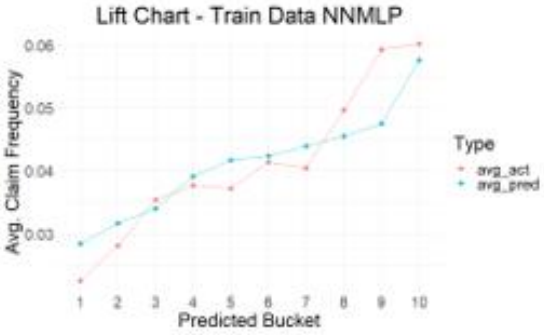


Figura 48. Lift-charts para Claim Frequency







Claim_Frequency_model_Performance_Test	R_squared	RMSE	MAE
GLM	0.0003551250	0.1538303	0.06084508
GAM	0.0003854509	0.1524348	0.04189059
MARS	0.0002005475	0.1524664	0.04189092
Elastic Net	0.0003491694	0.1538200	0.06085592
Random Forest	0.0007864018	0.1536334	0.06086662
XGBoost	0.0007726252	0.1536865	0.06078939
MLP	0.0007726252	0.1536865	0.06078939

Claim_Frequency_model_Performance_Train	R_squared	RMSE	MAE
GLM	0.001333001	0.1538303	0.06057706
GAM	0.001230651	0.1521009	0.04166673
MARS	0.003275428	0.1519449	0.04158630
Elastic Net	0.001325599	0.1534242	0.06059129
Random Forest	0.009083435	0.1531447	0.06064231
XGBoost	0.001913495	0.1533343	0.06058921
MLP	0.001913495	0.1533343	0.06058921

Figura 49. Métricas de Desempeño para Claim Frequency

9.2 Lift-Charts Claim Severity

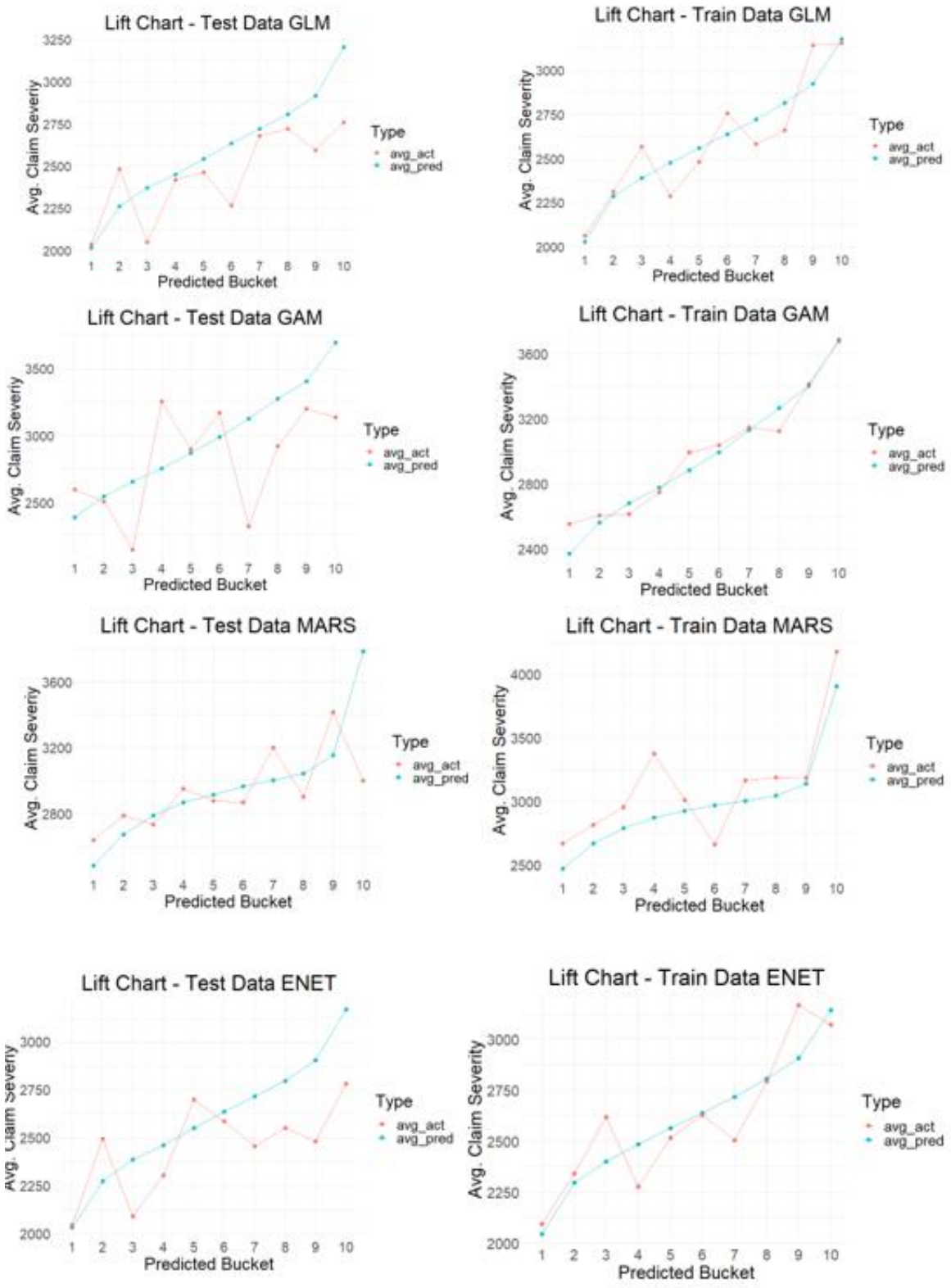
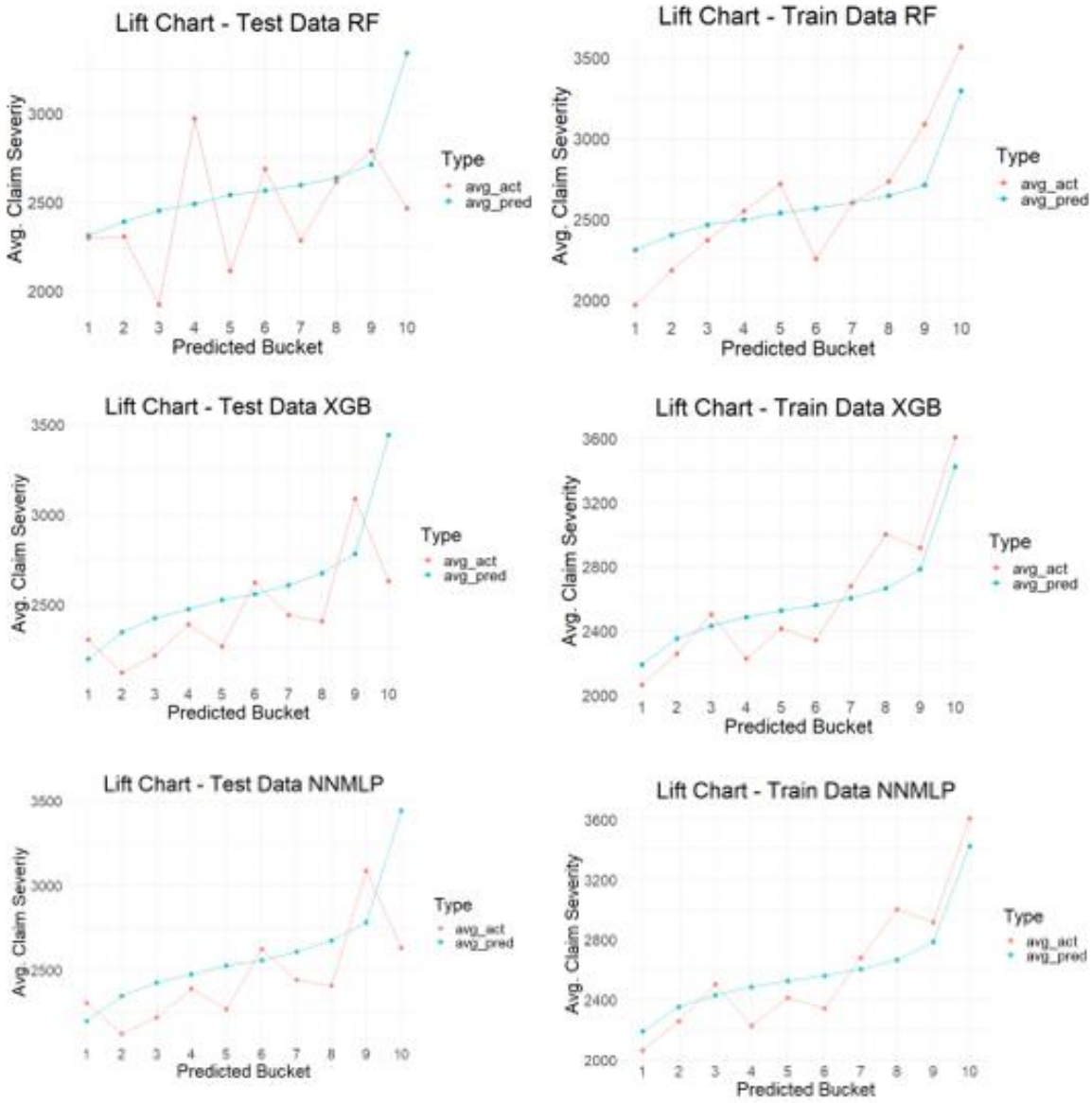


Figura 50. Lift-charts para Claim Severity



Claim_severity_mode_Performance_Test	R_squared	RMSE	MAE
GLM	0.0022101585	3771.234	2442.731
GAM	0.0021666592	3759.211	2586.996
MARS	0.0047421858	3749.074	2592.115
Elastic Net	0.0022815222	3770.240	2443.194
Random Forest	0.0013489214	3770.535	2459.945
XGBoost	0.0008226073	3776.172	2459.735
MLP	0.0008226073	3776.172	2459.735

Claim_severity_model_Performance_Train	R_squared	RMSE	MAE
GLM	0.01086952	3771.234	2506.454
GAM	0.01089603	3811.116	2644.609
MARS	0.02015952	3793.221	2633.806
Elastic Net	0.01099926	3830.849	2505.732
Random Forest	0.01594669	3826.229	2500.462
XGBoost	0.01702195	3821.385	2495.954
MLP	0.01702195	3821.385	2495.954

Figura 51. Métricas de Desempeño para Claim Severity

9.3 Comentarios sobre los modelos y su desempeño

Tabla 5 Comentarios a Nivel de Modelo

GLM	Se observan métricas de desempeño muy homogéneas tanto en el conjunto de entrenamiento como en el de validación por lo que hay muy poca posibilidad de sobre ajuste. Se controla fácilmente el nivel base de las variables categóricas empleadas y la interpretación de los niveles subsecuentes.
GAM	A pesar que se identifican métricas de desempeño ligeramente bajas en comparación con los demás modelos, se observa un comportamiento indeseado en el lift chart del conjunto de prueba. Se mantiene una la buena interpretabilidad de los parámetros heredada de los GLM tradicionales. No se observan mayores cambios en las métricas de desempeño al modificar los hipoerparámetros del modelo Exige mayor verificación de los supuestos versus los modelos de Machine Learning.
MARS	-Al igual que en los modelos GAM en donde se identifican métricas de desempeño ligeramente bajas en comparación con los demás modelos, se observa un comportamiento indeseado en el lift chart del conjunto de prueba principalmente para la estimación de Claim frequency. La interpretación de los parámetros se complejiza perdiendo los beneficios de los GLM tradicionales. Se identifica una fuerte reducción de variables usadas en el modelo. Exige mayor verificación de los supuestos versus los modelos de Machine Learning.
GLM Elastic Network	Este modelo hereda las dificultades de los dos modelos anteriores (GAMs y MARS) en el comportamiento del lift chart principalmente en el conjunto de prueba. Se observa una alta reducción de dimensionalidad al permitir que el tuning de los hiperparámetros sea realizado de manera automática por el modelo, en cuyo caso se seleccionan modelos demasiado reducidos generando beneficios mínimos en las métricas de desempeño, pero afectando el comportamiento de los lift charts y por consiguiente la posibilidad de aumentar el número de casos de selección adversa. Exige mayor verificación de los supuestos versus los modelos de Machine Learning.
Bagging (Random Forest)	Se observa un pobre comportamiento de este modelo tanto en las métricas de desempeño como en los lift charts. Se pierde la interpretabilidad de los GLM sin conseguir en contra parte un beneficio en las métricas de desempeño. Sus hiperparámetros son de fácil ajuste y no requiere verificación de supuestos. Su consumo computacional es bastante alto por lo que podría tener problemas si se encuentra en producción, incluso de manera local, si el número de datos es alto, lo que lo hace no recomendado para Big Data.

Boosting (XGBoost)	A pesar de que las métricas de desempeño son levemente inferiores a las de los demás modelos, muestra un comportamiento más robusto en los lift charts tanto para Claim Frequency como para Claim Severity, tanto en conjunto de entrenamiento como el de prueba. Su desempeño computacional es bastante bueno. No es un modelo "knowledge-based" sino enfocado en mejorar el poder predictivo por lo que se pierde la interpretabilidad respecto a los GLMs.
MLP	Este modelo muestra un comportamiento casi idéntico al del XGBoost tanto en lo concerniente a métricas de desempeño como a los lift charts.

10. Conclusiones

- El análisis cuantitativo de desempeño de los modelos construidos (GLMs y de Machine Learning) de cara al diseño de estrategias de tarificación en el ramo de autos al aplicar técnicas de Machine Learning, arroja que para el data set utilizado, no existe diferencia apreciable entre las métricas de desempeño de éstos, por lo que no se considera desde el punto vista puramente cuantitativo que se deba privilegiar uno de los dos enfoques en particular.

- En el análisis cualitativo de desempeño y los procesos metodológicos asociados a los modelos aplicados se evidencian diferencias apreciables entre los lift charts construidos, aun cuando dichas diferencias sean poco perceptibles a través de las métricas de desempeño habituales.

- Aunque el impacto de la aplicación de modelos de Machine Learning no sobrepasa a los GLM desde la perspectiva de las métricas de desempeño, sí representan un complemento adicional pertinente y adecuado para contrastar los resultados de los modelos clásicos.

- Las variaciones de los GLMs aplicadas muestran una disminución en la interpretabilidad de los modelos tradicionales sin incrementar notablemente su desempeño. Sin embargo, al igual que ocurrió con los modelos de Machine Learning, representan una herramienta de contraste válida.

- Las métricas de desempeño no representan una medición absoluta y objetiva del desempeño de los modelos de tarificación por sí mismas, por lo que para evitar selección adversa es indispensable utilizar mecanismos adicionales como los métodos gráficos (lift-charts) que complementen dichas métricas.

- Se confirma la metodología CRISP DM como un marco metodológico adecuado y enriquecedor para enmarcar los procesos de tarificación de seguros de no vida, en tanto sus etapas garantizan de manera exhaustiva el ciclo de vida de los modelos.

- Una posible oportunidad de mejora y de continuidad para el presente trabajo radicaría en construir una metodología estándar con potencial de automatización y testeo de diferentes modelos de tarificación para diferentes data sets provenientes de ramos de no vida en un ambiente de producción.

- Se observó que el desempeño de los modelos fue menor para la predicción de la frecuencia de reclamaciones que para la severidad de estas, lo anterior debido a el desbalanceo natural por el bajo volumen de reclamaciones.

- Resulta enriquecedor para el contexto académico y real de la región, el desarrollar estrategias disruptivas que propendan por la mejora continua de las metodologías actuales de tarificación de seguros, incorporando mejoras que no provengan solamente de adaptaciones que han funcionado en mercados de mayor envergadura, sino de iniciativas locales en el marco de la investigación científica.

11. Bibliografía

Asimit, V. (2020). Special Issue “Machine Learning in Insurance”. *Risk*.

Blier-Wong, C. C. (2001). Machine Learning in P&C Insurance: AReview for Pricing and Reserving. Risks.

Catalina Lozano, F. P.-G. (2021). The current role of machine learning and explainability in actuarial science. *Short Papers of the 9th Conference on Cloud Computing Conference, Big Data & Emerging Topics*, 29,30,31.

Chapman, P. (2000). CRISP-DM 1.0. . *IBM*.

Denuit, M. (2021). Autocalibration and Tweedie-dominance for Insurance Pricing with Machine Learning. *arXiv:2103.03635v2 [stat.ML]*.

Gareth James, D. W. (2021). *An Introduction to Statistical learning.* 5th edition.

Geoff Werner, C. M. (2016). BASIC RATEMAKING. *CAS*.

Giorgio Alfredo Spedicato, C. D. (2018). Machine Learning Methods to Perform Pricing Optimization. A Comparison with Standard GLMs. Variance. *Casualty Actuarial Society*.

Guillen, M. (2018). MACHINE LEARNING AND PREDICTIVE MODELING FOR AUTOMOBILE INSURANCE PRICING. *Anales del Instituto de Actuarios Españoles, 4ª época*.

Imran Chowdhury Dipto, M. A. (2020). Prediction of Accident Severity Using Artificial Neural Network: A Comparison of Analytical Capabilities between Python and R. *Journal of Data Analysis and Information Processing, Vol.8 No.3*.

Jiménez, P. G. (2020). *Modelos de predicción del coste del siniestro en Seguros de Salud.* Madrid: Universidad Carlos II de Madrid.

Kshirsagar, R. (2019). Accurate and Interpretable Machine Learning for Transparent Pricing of Health. *Lumiata Data Science and Engineering*.

Kuoa, K. (2020). Towards Explainability of Machine Learning Models in Insurance Pricing. *arXiv:2003.10674v1 [q-fin.RM]*.

Riley, J. (2020). AI and Machine Learning Usage in Actuarial Science. *Williams Honors College, Honors Research Projects. 1081*.

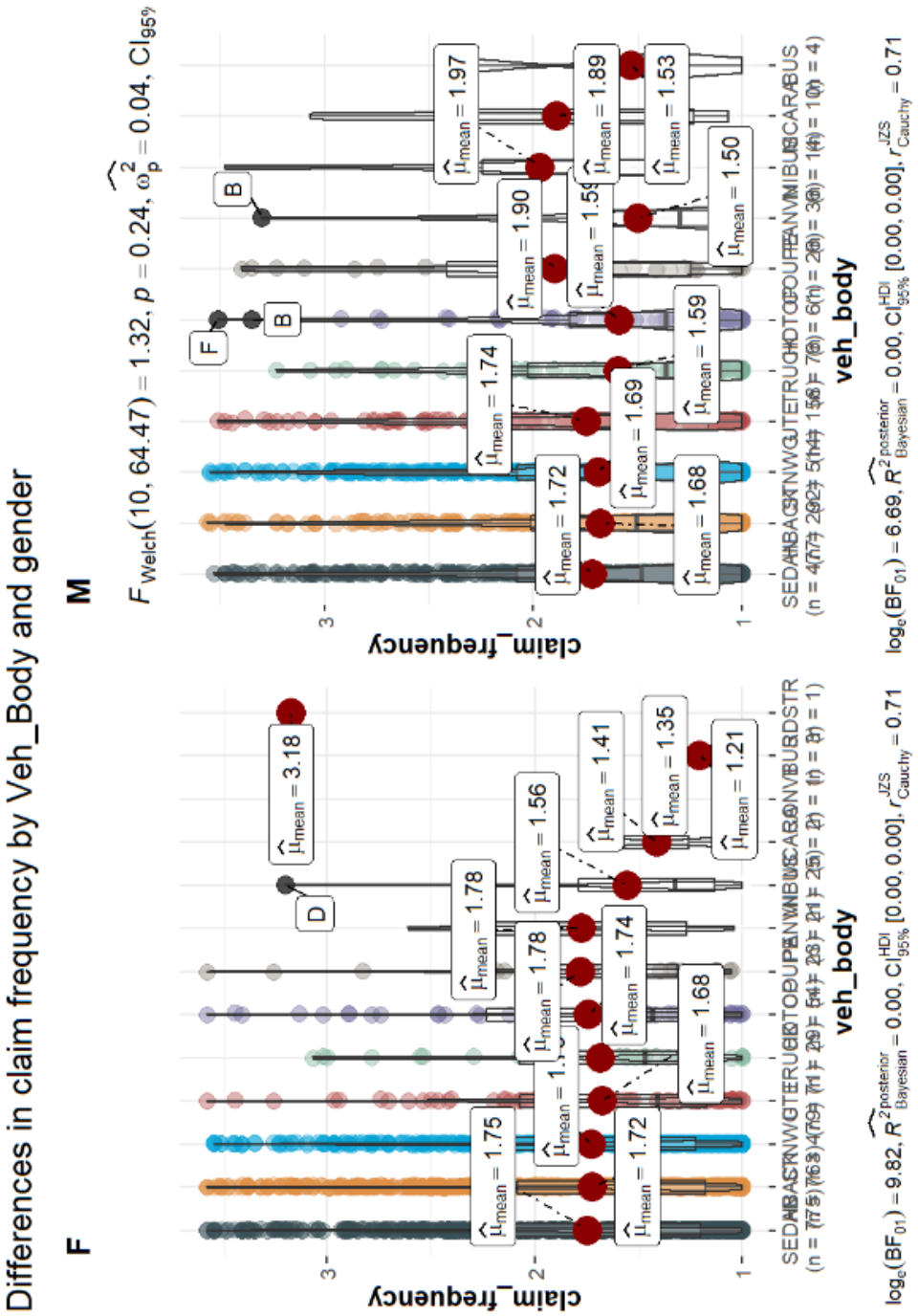
Roberto Perez, F. -N. (2021). Refinamiento y Validación. *CAS – Introducción al proceso de análisis predictivo en el mercado asegurador.* Bogotá: CAS – Escuela de Ingeniería Julio Garavito.

Wagner, Y. S. (2020). Comparison of Machine Learning and Traditional. *Swiss Finance Institute*.

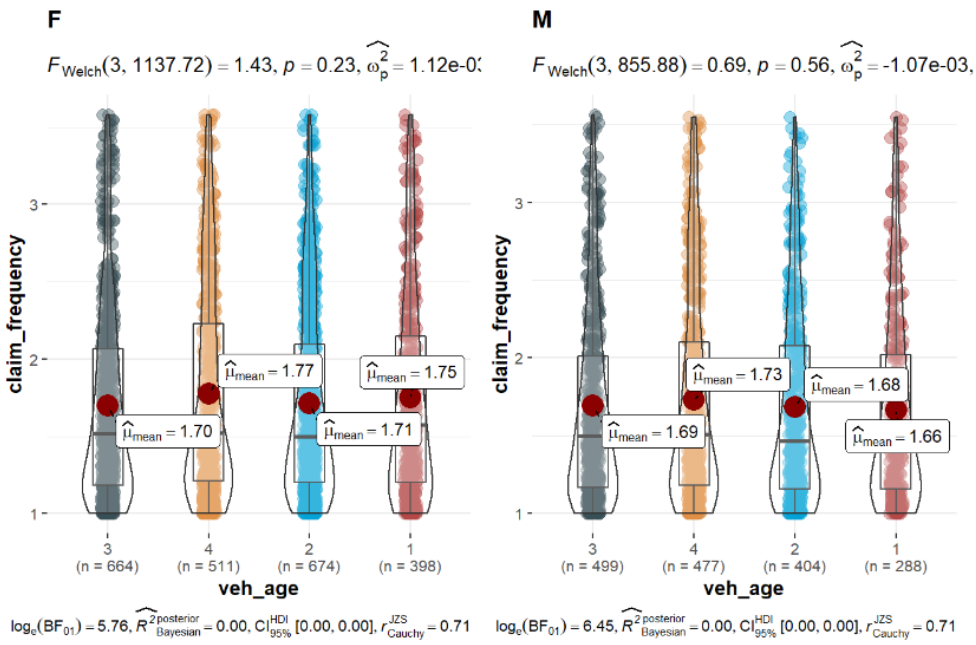
Wuthrich, M. V. (2020). Data Analytics for Non-Life Insurance Pricing. *Swiss Finance Institute Research Paper No. 16-68*.

12. Anexos

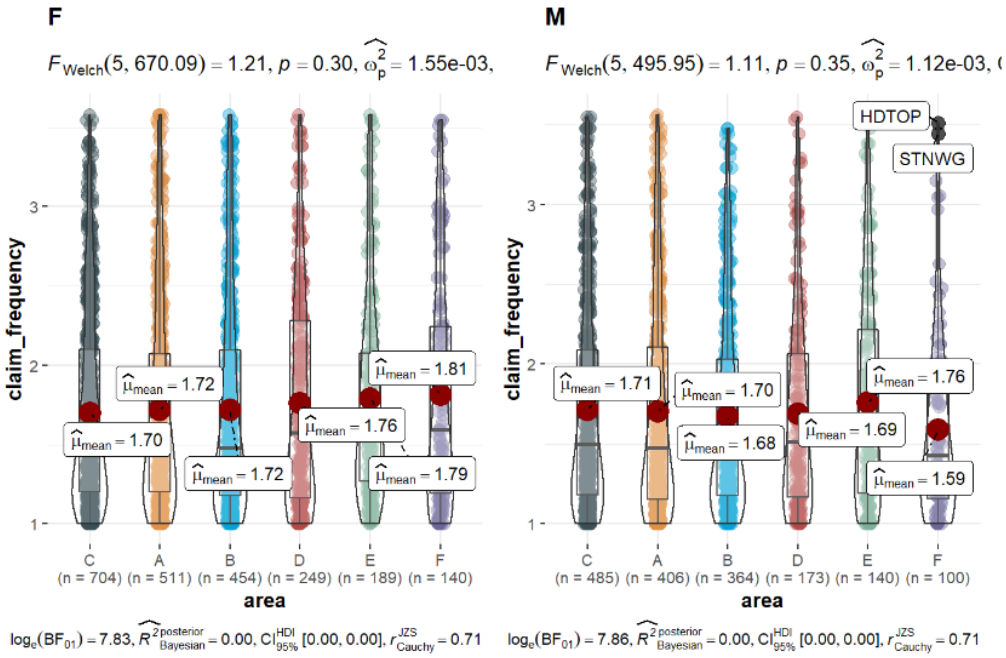
11.1 Exploración Grafica de interacciones de variables categóricas para "Claim Frequency"



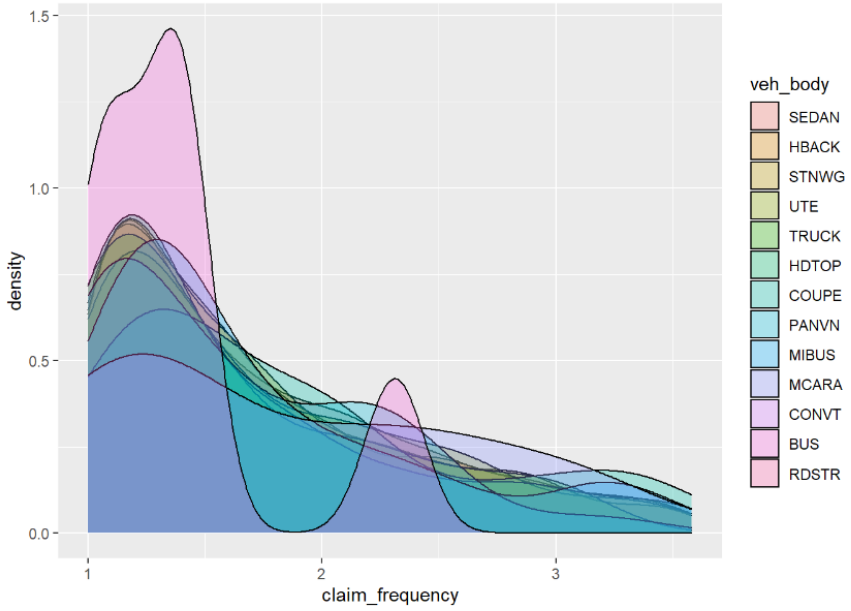
Differences in claim frequency by Veh_Age and gender

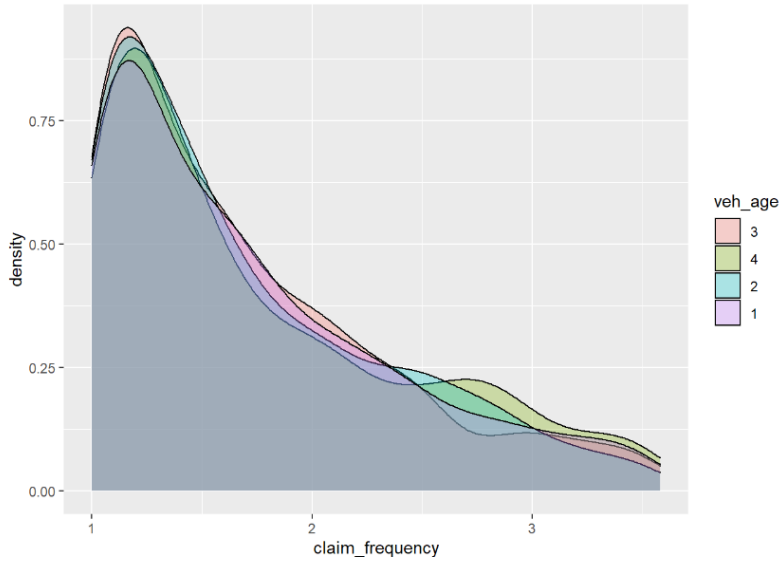


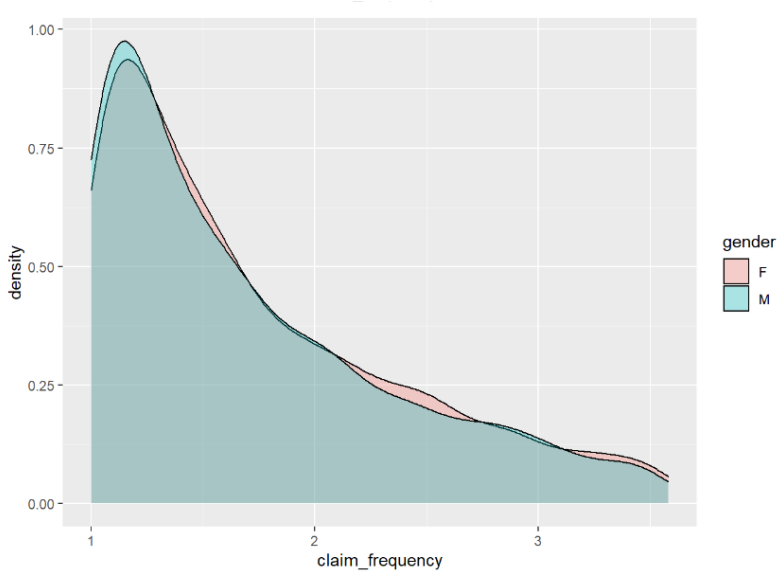
Differences in claim frequency by Area and gender

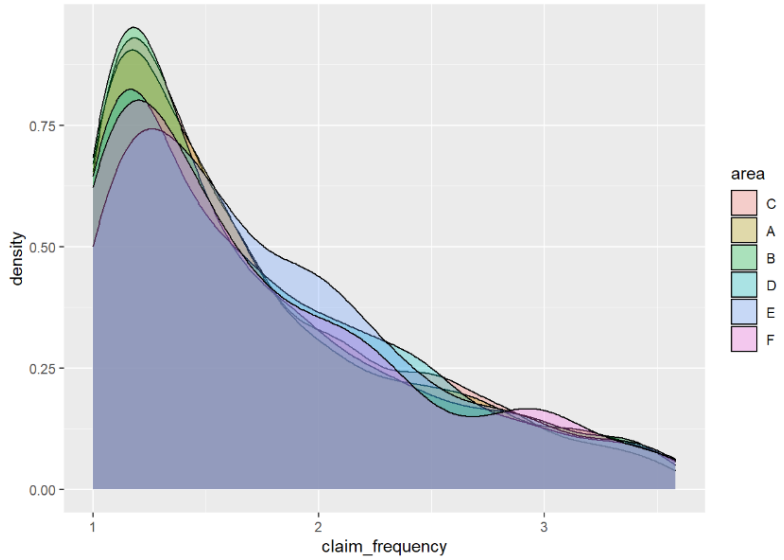


11.2 Análisis Exploratorio de Claim Frequency para el subconjunto de los registros que hicieron reclamaciones

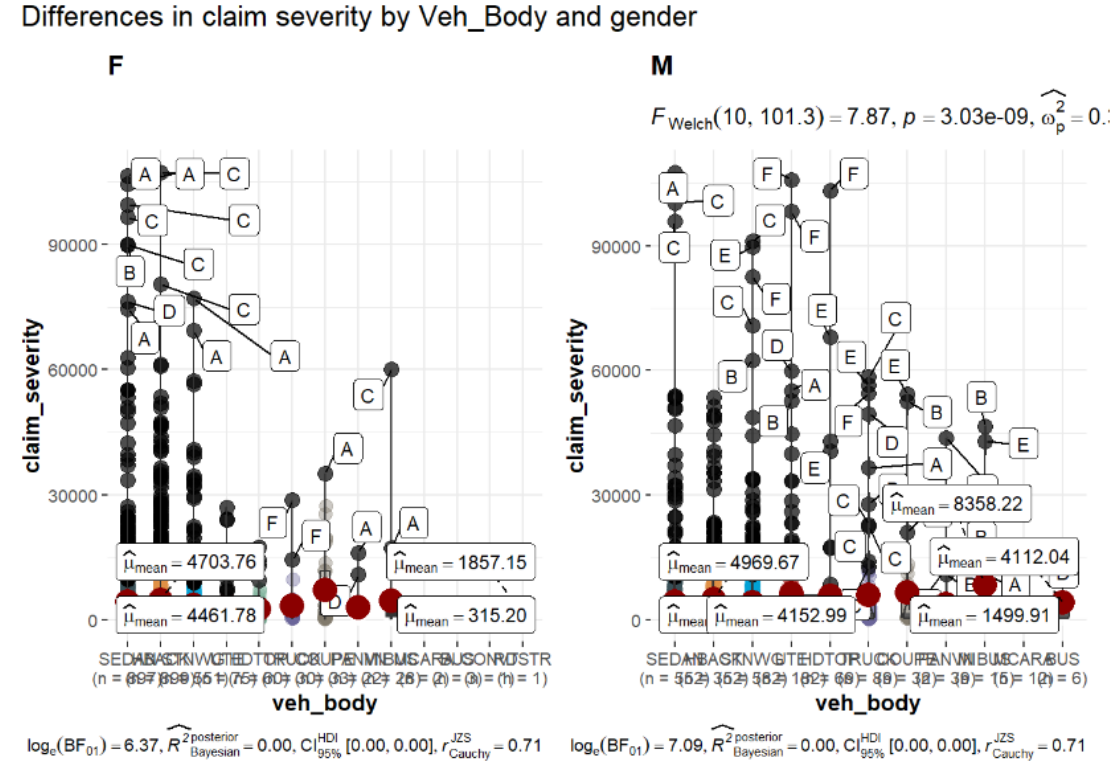


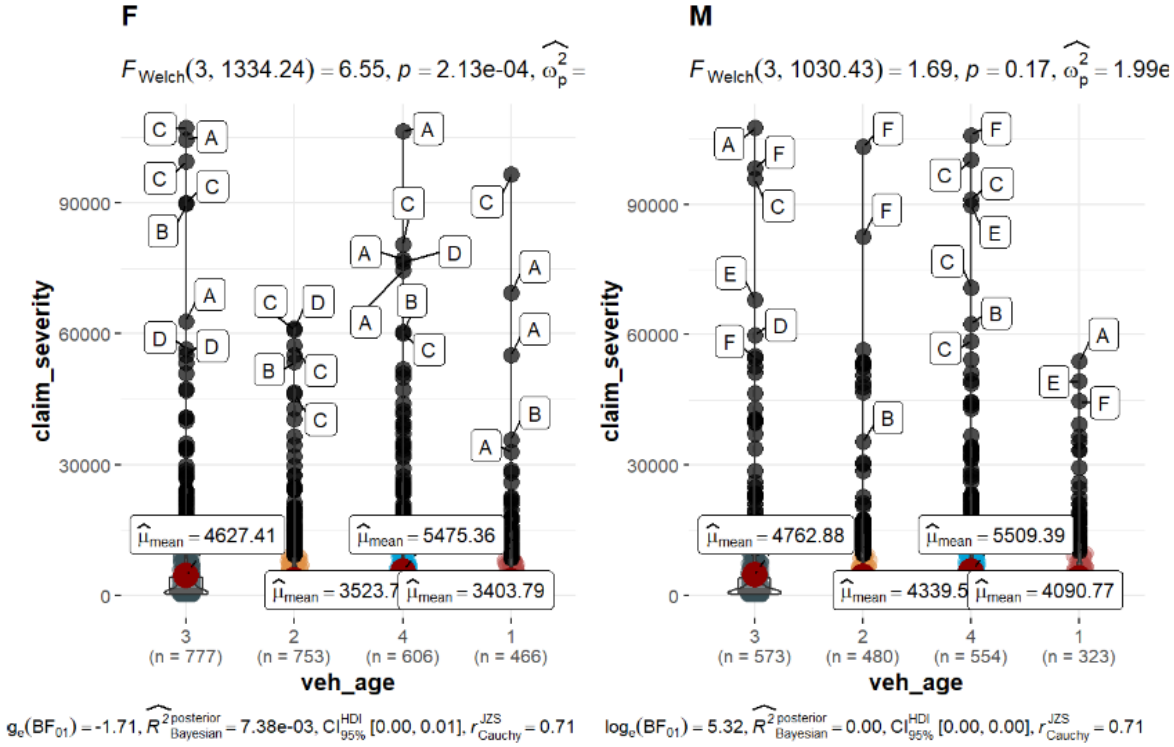


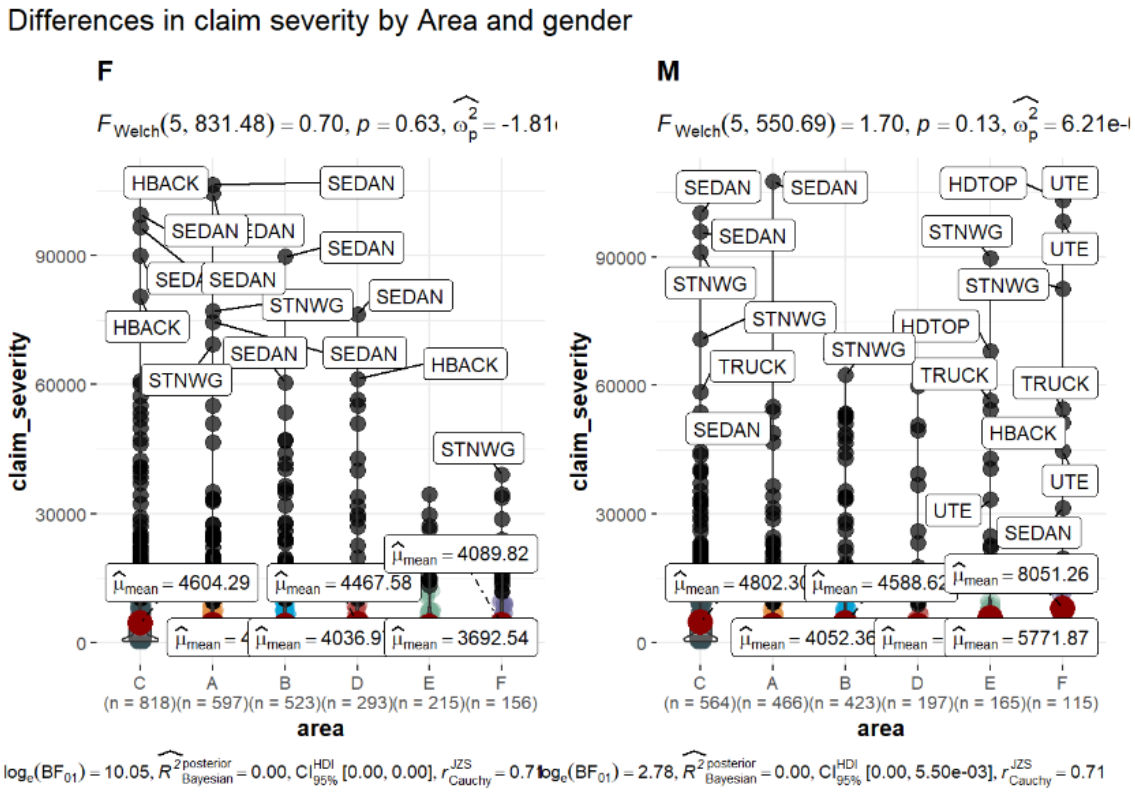

1.00
0.75
0.50
0.25
0.00
density
1
2
3
claim_frequency
gender
F
M


0.75
0.50
0.25
0.00
density
1
2
3
claim_frequency
area
C
A
B
D
E
F

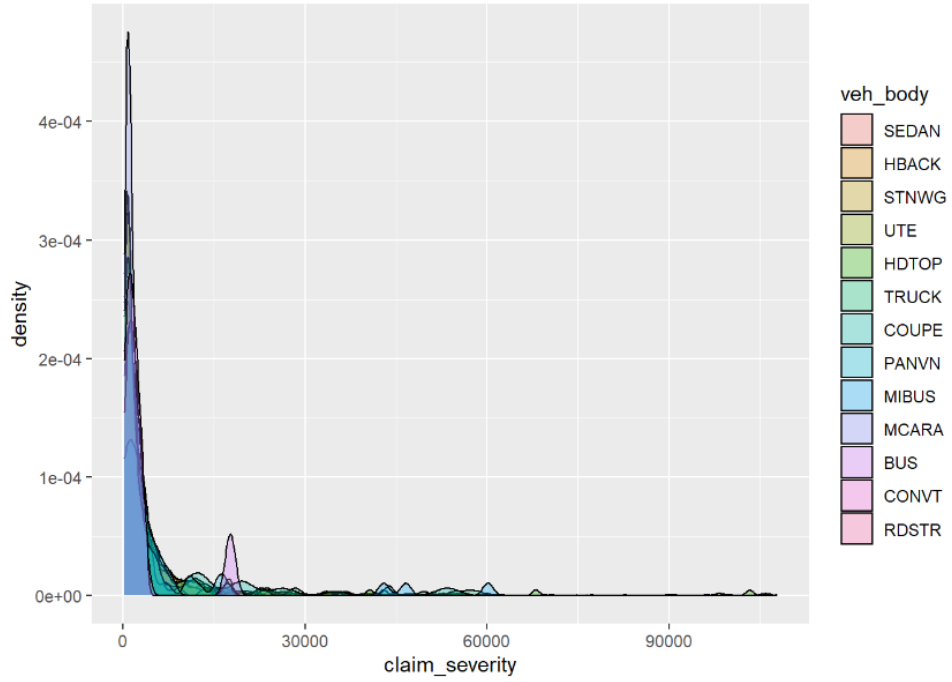
11.3 Exploración Grafica de interacciones de variables categóricas para "Claim Severity"

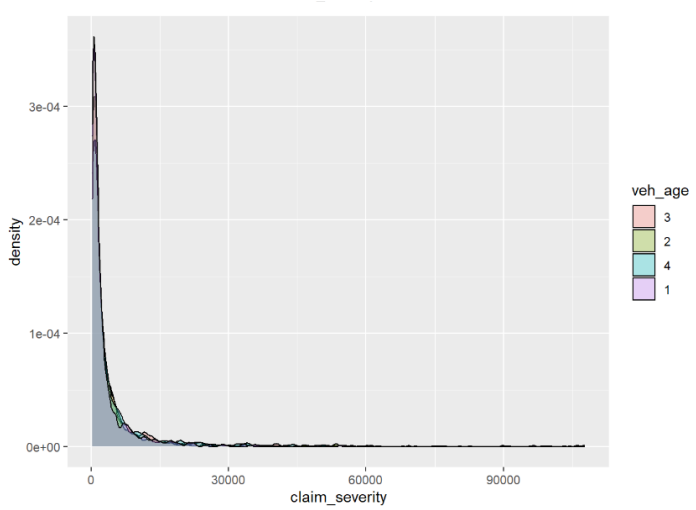


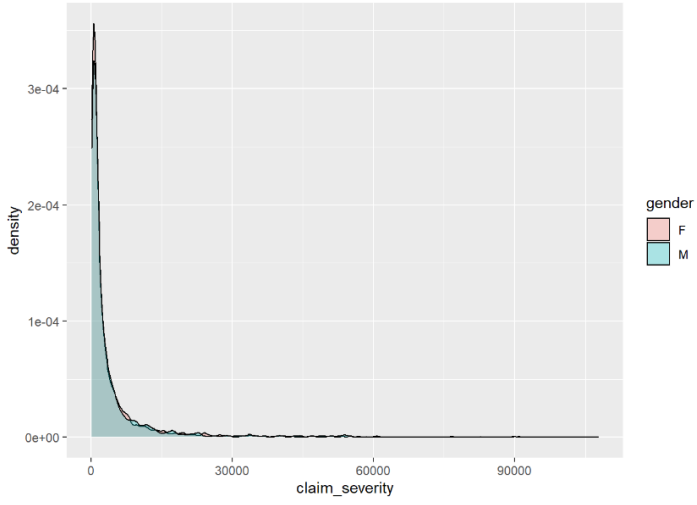


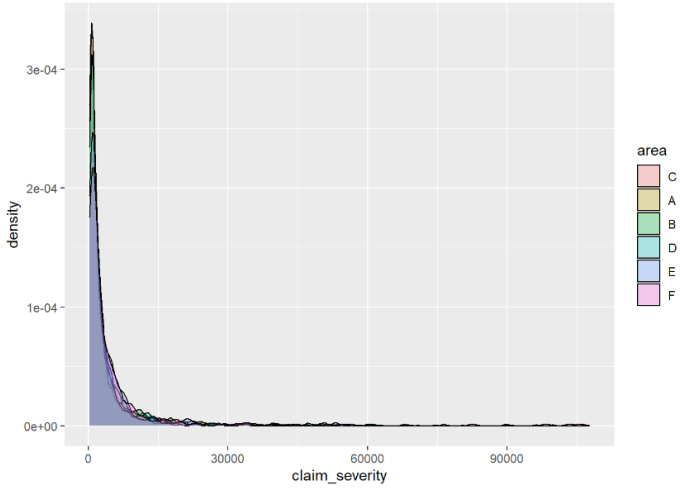


11.4 Análisis Exploratorio de densidad de Claim Severity para el subconjunto de los registros que hicieron reclamaciones

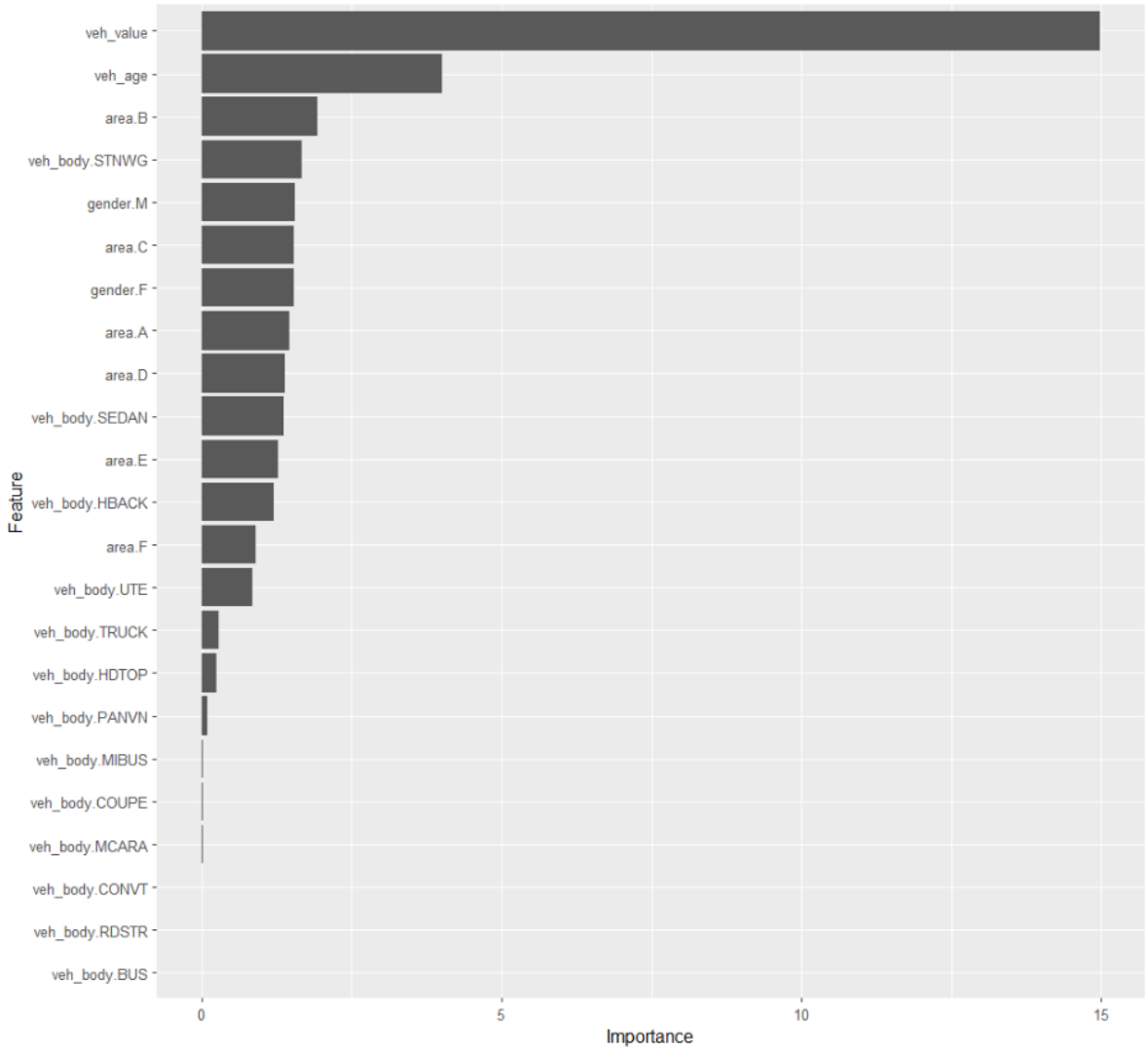



3e-04
2e-04
1e-04
0e+00
density
0
30000
60000
90000
claim_severity
veh_age
3
2
4
1


3e-04
2e-04
1e-04
0e+00
density
0
30000
60000
90000
claim_severity
gender
F
M


3e-04
2e-04
1e-04
0e+00
density
0
30000
60000
90000
claim_severity
area
C
A
B
D
E
F

11.5 Importancia de variables basada en impureza de nodos para el modelo de Claim Frequency usando Random Forest.



11.6 Importancia de variables basada en impureza de nodos para el modelo de Claim Severity usando Random Forest.