Modelo predictivo de rentabilidad para la atracción de clientes a un problema de negocio de una aseguradora

María Alejandra Ríos Herrera

Trabajo de grado presentado para optar al título de:

Especialista en Analítica y Ciencia de Datos

Asesora

Daniela Serna Buitrago

Especialista (Esp) en Analítica

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2022

Cita	Ríos Herrera [1]
Referencia Estilo IEEE (2020)	[1] Ríos Herrera, "Modelo Predictivo de Rentabilidad Para La Atracción De Clientes A Un Problema De Negocio De una Aseguradora", Trabajo de grado especialización, Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2022.

Especialización en Analítica y Ciencia de Datos, Cohorte II.

Dedicatoria

A mis hermanos: por llenar mis días de diversión. Por ser la mayor razón para perseverar.
Al amor de mi vida: por su acompañamiento y amor incondicional: Contigo volví a creer en mí, las mujeres podemos ser grandes científicas, ingenierías y matemáticas.
A mis papás: por el inmenso sacrificio de separarse de mi desde hace años para que pueda estudiar y porque solo gracias a ellos mis problemas de autoestima desaparecen.

Y a Efraín Francisco Ríos: Todos mis triunfos llevarán tu nombre.

Agradecimientos

A Daniela Serna por el acompañamiento y confianza para el desempeño de esta investigación.
A la Universidad de Antioquia por siempre abrirme las puertas.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

Esp.	Especialista
IA	Inteligencia Artificial
BI	Business Intelligence
E.T. L	Extraer, transformar, cargar
R.E.	Recuento estadístico
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
DANE	Departamento Administrativo Nacional de estadística
ML	“machine learning”
Fig	Figura

I. RESUMEN EJECUTIVO

Hoy en día, las compañías de seguros en Colombia están en búsqueda constante de conocer a sus clientes para determinar la contribución efectiva o la rentabilidad que ellos les generan. El "marketing" actual adopta el aprendizaje profundo de los clientes como una técnica para mejorar la rentabilidad de sus recaudos cuando ellos adquieren determinados productos en la compañía, es decir, a partir de estos se pueden modificar los resultados, mejorar las estrategias de atraer los clientes a los servicios y aumentar el valor agregado que los mismo le generan. Teniendo en cuenta que el desempeño de la rentabilidad técnica está inmersa a las posibilidades de distintos rendimientos financieros y a impactos que pueden generar quiebras o fallas en su estrategia de continuidad, se debe abordar en primer lugar, los aspectos coyunturales en la cotidianidad, por ejemplo: El descenso de la siniestralidad en algunos ramos[1] siendo ahora más afectada como consecuencia por la pandemia. Mencionado lo anterior, las aseguradoras en Colombia actualmente deben hacer frente al pago de una amplia gama de pólizas, que incluyen todo tipo de consecuencias, como la cancelación de estas, o eventos como pagos de las indemnizaciones de los clientes. El objetivo del "marketing" en potencia se enfoca en suplir a los clientes distintas herramientas que le brinden seguridad y confort con base a la buena identificación de sus necesidades. La presente investigación e implementación busca englobar la idea del conocimiento de los clientes de una empresa de seguros, y con base a esta información se realizará la implementación de un modelo analítico predictivo con el uso de herramientas de "machine learning", el cual le permite a la aseguradora estar preparada para analizar todas las tareas incluidas en los procesos de suscripción de un cliente a una póliza, y que a su vez presentan datos numéricos para evaluar el impacto en la cartera de negocios, es decir, a futuro es otra herramienta directamente relacionada con la labor que hoy en día consumen los asesores para inscribir clientes a la compañía y realizarles los estudios correspondientes como definición de su estrategia de mercadeo, y de como resultado que la aseguradora realice el despliegue y considere los cambios más eficientes de tarifas o de servicios.

Palabras clave — Ramo, Aprendizaje profundo, aseguradora, Análisis predictivo, "machine learning", Inteligencia Artificial.

[1] Ramo: Productos de una compañía de seguros, son ejemplos: Pólizas de vida, vivienda, daños a terceros, contraincendios, automóviles.

II. ABSTRACT

Today, insurance companies in Colombia are in constant search of knowing their clients to determine the effective contribution or profitability that they generate. Current "marketing" adopts deep learning and "machine learning" from customers as a technique to improve the profitability of their collections when they acquire certain products in the company, that is, from these the results can be modified, improve the strategies of attracting customers to services and increase the value-added that they generate. Considering that the performance of technical profitability is immersed in the possibilities of different financial returns and impacts that can generate bankruptcies or failures in its continuity strategy, the conjunctural aspects in daily life must be addressed first, for example: The decrease in the accident rate in some lines of business now being more affected because of the pandemic. Insurers in Colombia currently must face the payment of a wide range of policies, which include all kinds of consequences, such as the cancellation of these, or events such as payment of customer compensation. The objective of potential "marketing" focuses on supplying customers with different tools that provide security and comfort based on the good identification of their needs. This research and implementation seeks to encompass the idea of knowledge of the clients of an insurance company, and based on this information, the implementation of a predictive analytical model will be carried out with the use of "machine learning" tools, which allows the insurer be prepared to analyze all the tasks included in the processes of subscribing a client to a policy, and which in turn present numerical data to evaluate the impact on the business portfolio, that is, in the future it is another tool directly related to the work that consultants consume nowadays to register clients with the company and carry out the corresponding studies as a definition of their "marketing" strategy, and as a result that the insurer carries out the deployment and considers the most efficient changes in rates or services.

Keywords — insurance policy, Deep Learning, insurance company, predictive analytics, "machine learning", I.A.

III. INTRODUCCIÓN

Los modelos analíticos predictivos son modelos que están basados en algoritmos con complejidad estadística, los cuales se han venido incorporando en las aplicaciones de realidades empresariales. Actualmente, en el mercado financiero son distintos los modelos que se están ejecutando para la predicción de sucesos basados en probabilidades y más específicamente en el tema de la retención de clientes o en la cuantificación de satisfacción de estos. Para el desarrollo de un modelo predictivo se lleva a cabo un proceso de decisiones y familiarización del problema antes de entrar a desarrollar el modelo para la aseguradora, a esto se le conoce como un diagnóstico inicial, el cual involucra la identificación de aspectos críticos e identificación de patrones para el tratamiento de datos. El desarrollo y la utilización de modelos predictivos de seguros depende en gran medida de plataformas analíticas exclusivas que deben satisfacer una amplia gama de requisitos, incluidas las funcionalidades ETL (extraer, transformar, cargar); manejo, preprocesamiento, entrenamiento, comprobación, la producción y el monitoreo. Las aseguradoras recurren a las herramientas comerciales y de código abierto para justificar el coste de la implementación. Sin embargo, es necesario ser cuidadosos con las consideraciones pues a menudo esto podría conducir al consumo de tiempo y recursos. Es válido aclarar que las empresas aseguradoras usan mecanismos como el de la presente investigación para la optimización y mejora de ganancias, puesto que actualmente en el mundo, las compañías de seguros también hablan de ideas de uso de peritaje digital para el reconocimiento de imágenes y aplicación de modelos de aprendizaje automático en los cuales se analizan los daños siniestrales y a su vez se apoyan con gestiones de prevención del riesgo. La IA y el aprendizaje automático complementan, por ejemplo, en los casos de ramos asociados a automóviles, la identificación de carros, comportamientos o estilo de vida que tal vez hoy incrementan el nivel de riesgo económico de la compañía y de esta manera, apoyando y ejecutando soluciones se podrían implementar reducciones de costos administrativos. El rol del aprendizaje automático presenta correlación con frecuencia de datos y eventos, de tal forma que confine en la elaboración de predicciones mediante el uso de estadísticas computaciones. Para la industria y las compañías en las que se enfocan en B.I. como las compañías de pólizas de seguros, pueden incrementar el potencial de los clientes incorporando definiciones tales como la minería de datos o aprendizaje no supervisado, pues con el tiempo, las situaciones del mercado y los resultados están en un continuo cambio, es decir, no es correcto modelar a la humanidad y a su comportamiento en 1990 y esperar que los resultados hoy sean similares.

IV. PLANTEAMIENTO DEL PROBLEMA

La aseguradora en concreto manifiesta el deseo de minimizar los costos que se le atribuyen a la compañía en la adquisición de sus productos. Se sabe que la aseguradora obtiene los ingresos con los contratos de los productos adquiridos por los asegurados y mediante estos recursos se debe distribuir gran porcentaje entre los clientes que sufren de pérdidas aseguradas. Pero la idea de negocio y el desempeño de crecimiento para la compañía establece que las compañías de seguros puedan obtener ganancias menguando el riesgo de sus asegurados para suplir las necesidades que se presentan con mayor calidad. Esto permite que las aseguradoras tengan una ayuda financiera para enfrentar caídas financieras como ocurrió con la contingencia de covid-19 y pérdidas inesperadas, aumentando el valor agregado que los asegurados obtienen de los seguros. Es prudente afirmar que esta necesidad es fuerte ya que a diferencia de muchos otros productos cuyo coste se conoce antes de que se venda el producto, las pólizas de seguro se desconocen en el momento de la compra. Por lo tanto, vender un producto de seguro conlleva un gran riesgo financiero.

Antecedentes

El problema de negocio para la compañía surge concibiendo la idea de que la población colombiana busca en el sector de seguros una medida de contingencia para evitar consecuencias económicas negativas a futuro en caso de que se generen eventualidades críticas, además, buscan que el importe de los daños o pérdidas que sufre una parte se puedan mercantilizar entre una comunidad de personas que la soporta de forma conjunta, con un impacto mucho menor que si el daño se presentará de forma individual.

V. OBJETIVOS

A. Objetivo general

Implementar un modelo predictivo usando distintas herramientas de inteligencia artificial y más específicamente fundamentar toda la implementación del modelo en algoritmos de regresión que ofrecen las herramientas de "machine learning" para predecir la rentabilidad o el resultado técnico que puede generar un cliente cuando se suscriba a la compañía de seguros.

B. Objetivos específicos

- Probar el diseño de un modelo predictivo calibrado para la predicción del resultado técnico
- Establecer métricas de evaluación que de un indicio del desempeño para el objetivo del modelo
- Minimizar los rangos del error que arroja la predicción del resultado técnico mediante métodos estadísticos descritos en la algoritmia del aprendizaje automático.

VI. PREGUNTA DE NEGOCIO

¿Es posible predecir el valor de la rentabilidad que puede generar un cliente a partir de sus características sociodemográficas, cuando se le ofrecen los tres ramos más importantes de la compañía, usando herramientas de “machine learning” e inteligencia Artificial?

VII. HIPÓTESIS

A. Primera Hipótesis: Datos con ausencia de información

Los datos que conforman el problema presentan una gran cantidad de datos sin información en sus características, esto es algo realmente común en distintos problemas de naturaleza real. Los datos también consideran que un cliente tiene la opción de comprar cualquiera de los tres ramos ofrecidos: Seguros de vida, automóviles y salud, o incluso los 3 ramos. Por tal razón, existe la opción de que los datos tengan registros repetidos del mismo cliente con sus datos personales, esto es, que se podría presentar más de una vez registrado con sus datos personales, pero con ramos distintos. La primera hipótesis surge de corroborar que existen una cantidad considerable de datos **sin información** y que pueden estar documentados en otras filas del dataset asociado al mismo usuario. De acuerdo con Fig. 2, aproximadamente 312.031 clientes han comprado más de una póliza en la compañía, sin embargo, todos los registros asociados con los clientes presentan la misma información o falta de ella, como se ejemplifica en la Fig. 3, por lo tanto, la hipótesis de complementar los datos personales con los faltantes se descartó.

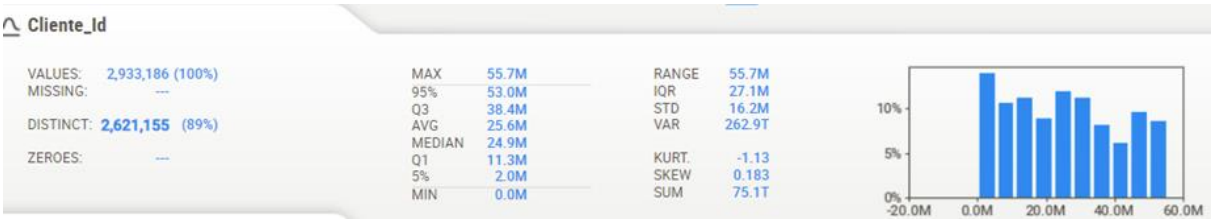


Fig. 1 Identificación de clientes

```
[ ]  df['Cliente_Id'].value_counts()

3984720     3
18182740    3
9498849     3
3603064     3
23203538    3
           ..
2641978     1
27805753    1
22193735    1
2631733     1
42376403    1
Name: Cliente_Id, Length: 2621155, dtype: int64
```

Fig. 2 Recuento del Cliente_Id

```
df[df['Cliente_Id']==3984720]
```

	Cliente_Id	Edad_Cliente	Sexo_Cd	Estado_Civil_cd	Nm_Estrato_Vivienda	Valor_Ingresos_Mensuales	Nombre_Ciudad	Nombre_Departamento	Numero_Hijos	Ocupacion_Desc	Ramo_Id	Ramo_Desc	Resultado_Tecnico	Produccion_Emitida
783109	3984720	69	F	-1	-1	?	RIONEGRO	ANTIOQUIA	2	ASESOR DE SEGUROS	26	SALUD FAMILIAR	8.815031e+06	12437802.0
1768158	3984720	69	F	-1	-1	?	RIONEGRO	ANTIOQUIA	2	ASESOR DE SEGUROS	168	AUTOMOVILES	-8.847959e+04	1855676.0
2634433	3984720	69	F	-1	-1	?	RIONEGRO	ANTIOQUIA	2	ASESOR DE SEGUROS	78	VIDA INDIVIDUAL	1.716039e+06	1247225.6

```
df[df['Cliente_Id']==18182740]
```

	Cliente_Id	Edad_Cliente	Sexo_Cd	Estado_Civil_cd	Nm_Estrato_Vivienda	Valor_Ingresos_Mensuales	Nombre_Ciudad	Nombre_Departamento	Numero_Hijos	Ocupacion_Desc	Ramo_Id	Ramo_Desc	Resultado_Tecnico	Produccion_Emitida
671947	18182740	35	F	-1	-1	?	PEREIRA	RISARALDA	0	QUIROPRÁCTICO	26	SALUD FAMILIAR	-2.511128e+06	3.523934e+06
835194	18182740	35	F	-1	-1	?	PEREIRA	RISARALDA	0	QUIROPRÁCTICO	168	AUTOMOVILES	1.035320e+06	1.464338e+06
1993613	18182740	35	F	-1	-1	?	PEREIRA	RISARALDA	0	QUIROPRÁCTICO	78	VIDA INDIVIDUAL	4.773886e+05	8.526923e+05

```
df[df['Cliente_Id']==9498849]
```

	Cliente_Id	Edad_Cliente	Sexo_Cd	Estado_Civil_cd	Nm_Estrato_Vivienda	Valor_Ingresos_Mensuales	Nombre_Ciudad	Nombre_Departamento	Numero_Hijos	Ocupacion_Desc	Ramo_Id	Ramo_Desc	Resultado_Tecnico	Produccion_Emitida
317863	9498849	41	F	D	4	6300000	MEDELLIN	ANTIOQUIA	0	CONSULTORES	26	SALUD FAMILIAR	-3.274097e+06	3.476511e+06
790296	9498849	41	F	D	4	6300000	MEDELLIN	ANTIOQUIA	0	CONSULTORES	168	AUTOMOVILES	1.199156e+06	1.503360e+06
2577555	9498849	41	F	D	4	6300000	MEDELLIN	ANTIOQUIA	0	CONSULTORES	78	VIDA INDIVIDUAL	1.081554e+06	2.450598e+06

Fig. 3 Información de clientes con las tres pólizas

B. Hipótesis estadística: Ingresos Mensuales y Ocupación

Los ingresos mensuales de un cliente es una variable que inicialmente no se descartó como entrada para el modelo predictivo, puesto que da un indicio del salario con el que cuenta cada cliente y puede fortalecer la seguridad del pago de su póliza y posible ganancia para la compañía. Cuando se realizó limpieza de datos se verificó que la variable tiene aproximadamente el 90% de los datos con información faltante registrados con el valor '?'. Para abordar este problema, se diseñó una función en la cual se relaciona la ocupación de los clientes de la compañía con los ingresos mensuales, de tal manera que se obtuvo los salarios que sí están registrados para cada ocupación y con la media del valor de ingresos mensuales se asignaron a los registros con información faltante que tenían la misma ocupación. Esto para todos los casos. Cabe resaltar que esta hipótesis también se sustentó en la relación de los datos y se corroboró una asignación estimada aproximada al registro de los ingresos mensuales de una fuente de datos de los ingresos mensuales según la ocupación, arrojada por las encuestas estadísticas realizadas por el DANE. [1]

VIII. MARCO TEÓRICO

A. *Aprendizaje supervisado y no supervisado:*

El aprendizaje supervisado es una técnica de "machine learning" donde dentro de los parámetros se encuentra la variable de salida y los demás parámetros son optimizados por minimizar la diferencia entre la salida y la entrada, el objetivo del aprendizaje supervisado es minimizar la salida entre la salida real y la salida computada. Este proceso se itera para minimizar el error o alcanzar la convergencia entre los parámetros. El concepto de aprendizaje no supervisado es otro paradigma del aprendizaje automático, en el cual los ejemplos de entrenamientos son datos no etiquetados, es decir, no presentan variables de salida. De este modo los parámetros no supervisados se agrupan en prototipos dependiendo de las similitudes encontrados en cada iteración. [2]

B. *Modelos de regresión:*

El presente modelo de negocio es modelo bajo el concepto de regresiones usando técnicas de "machine learning" y aprendizaje automático con el fin de determinar si existe, o no, relación causal entre una variable dependiente un conjunto de otras variables explicativas. De esta manera, el modelo buscará determinar cuál será el impacto sobre la variable salida ante un cambio en las variables explicativas. La definición de regresión corresponde al transcurso de eventualidades de acuerdo con una variable de interés, modelada mediante cuantificación de fórmulas matemáticas. Dentro de las ventajas de los modelos de regresión es la consideración de distintos escenarios bajo distintos panoramas para tratar los problemas de negocio teniendo en cuenta las variables de mayor influencia respecto a las demás, es ahí donde surge el concepto de la correlación.

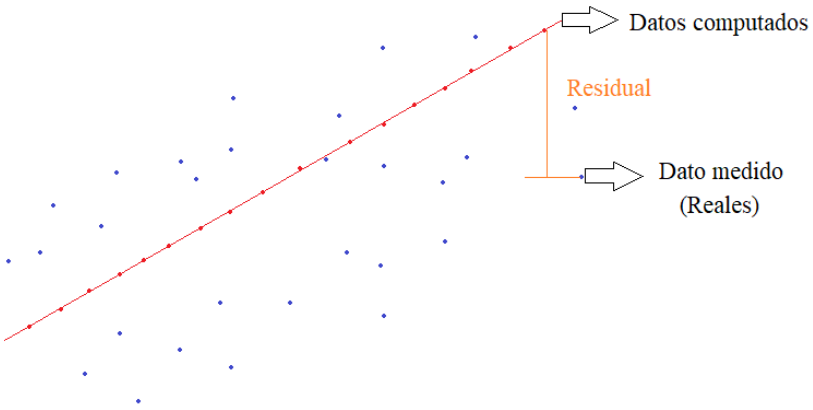


Fig. 4 Ejemplo de regresión lineal simple

C. Regresión Lineal Múltiple (MLR):

Este modelo está basado en técnicas estadísticas que corresponden a la extensión de una regresión lineal simple o la regresión de mínimos cuadrados (MCO), con la particularidad de que la MLR (Regresión lineal múltiple) se ajusta mejor a distintos problemas o fenómenos reales en los cuales se considera diferentes variables que puede ser representativas. El objetivo de la regresión lineal múltiple consiste en la relación de la variable a predecir o también llamada variable de interés y las variables de entrada, las cuales ocupan el papel de ser las variables predictoras. Las diferencias entre el comportamiento de la variable a predecir o la variable de salida puede tener un comportamiento similar o un patrón establecido por las variables regresores y su modelamiento matemático está dada por la ecuación:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{in} + \varepsilon \quad (1)$$

Donde:

- β_0 : es el termino independiente de la regresión cuando las variables predictoras son cero
- X_i : Son las variables predictoras
- ε : Es el error suponiendo que sigue una distribución normal de media nula
- Y : Es la variable para predecir

En el análisis de regresión lineal múltiple la construcción de su correspondiente ecuación se realiza seleccionando cada una de las variables y posteriormente, definir las variables más explicativas de la variable dependiente sin que ninguna de ellas sea combinación lineal de ellas y realizando la bondad de ajuste a cada parámetro. [3]

D. Decision tree regressor

Esta herramienta basada en arboles de decisión actúa como un método para aprender las regresiones lineales locales para aproximarse a la curva sinusoidal. Además, usa el mismo concepto de los árboles decisión para árboles de decisión de clasificación, es decir, dividir los datos usando un árbol binario de decisiones de los datos y separarlos de acuerdo con las condiciones en nodos de pruebas y en nodos de decisión para evaluar las variables de entrada y todos los puntos de división de datos para conseguir la predicción de la variable dependiente u objetivo que en este caso corresponde a “Resultado técnico”.

Para ver la profundidad máxima del árbol de decisión se tiene el parámetro “max_depth” con el objetivo de configurar el alto aprendizaje sobre el modelo. [4]

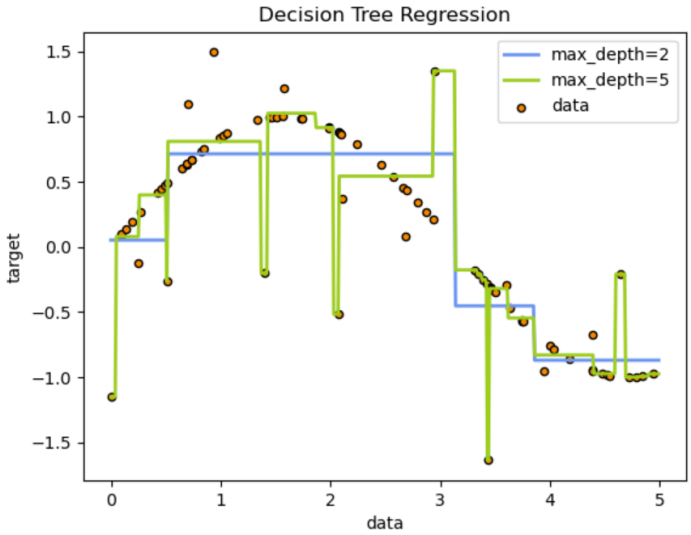


Fig. 5 Decision Tree Regression

Nota. fuente de Decision Tree Regression, scikit-learn.org

E. Random forest regressor

Este método de aprendizaje supervisado permite fijar un estimador el cuál divide la muestra de datos en distintos conjuntos, es una colección de árboles de decisión individuales convirtiéndolo para algunos casos un modelo un poco más robusto dependiendo de la naturaleza de los datos. El objetivo de *random forest regressor* consiste en entrenar las distintas muestras luego de la división de manera que cuando se entrene la primera iteración, las siguientes predicciones tendrán en cuenta las observaciones encontradas en las anteriores. Son distintas las ventajas del modelamiento basado en arboles para los casos de aprendizaje automático, este en particular, tiene una ventaja dado que su funcionamiento no se rige en una única ecuación matemática o estadística que definiría toda la muestra de los datos, esto debido a los estimadores que abarcan distintas técnicas supervisadas que mediante el agrupamiento lograría tratar el problema de manera más detallada o enfocada.

F. AutoML: Automatic “machine learning”

Este concepto también interviene en el desarrollo del presente problema de negocio, debido a la investigación sobre la demanda que existe hoy en día de los expertos en sistemas de aprendizaje automático y que, para muchos casos, esta demanda ya ha superado la oferta. Lo mencionado

anteriormente también se debe a que para muchos de los problemas que se desea resolver no se requiere solo de conocimiento sino también de bastante experiencia para tomar decisiones relacionadas con qué tipos de modelo entrenar y cómo evaluarlos para obtener los mejores resultados. Para abordar este inconveniente hay avances en el desarrollo de software de aprendizaje automático que, en cierto modo, resulta ser más sencillo de usar y que perfectamente pueden utilizar las personas no expertas en el tema. Los primeros pasos hacia la simplificación del aprendizaje automático implican el desarrollo de interfaces simples y unificadas que permitan la implementación de una variedad de algoritmos de aprendizaje automático. [5]

El aprendizaje automático automatizado (AutoML) se puede considerar como la automatización de principio a fin de algunos de los pasos involucrados en el proceso de aprendizaje automático estándar (Automatizar ciertas partes de la preparación de datos, generación de modelos óptimos, escogencia del mejor modelo, entre otros). El AutoML tiene un gran potencial en permitir que personas de diversas áreas de conocimiento utilicen modelos de aprendizaje automático para abordar sus problemas en escenarios complejos del mundo real. Basados en lo anterior, se puede pensar que el AutoML puede ser una respuesta a todos los impedimentos mencionados anteriormente. [6]

Fig. 6 Proceso del aprendizaje automatizado

Nota. fuente de https://towardsdatascience.com/a-deep-dive-into-h2os-automl-4b1fe51d3f3e

IX. METODOLOGÍA

A. *Proceso de analítica*

En la Fig. 7 se presentan las etapas que hacen parte del proceso de modelamiento y entrenamiento de los modelos de regresión que hicieron parte del tratamiento del problema de negocio. La primera parte de esta fase se enfatizó en todo lo relacionado con la exploración e ingeniería de características, luego se hizo verificación y visualización de estas características para analizar conceptos estadísticos de cada una de estas variables. Posteriormente con el uso de las herramientas que proporciona la librería scikit-multilearn se realiza imputación de datos para la variable de ingresos mensuales debido a un problema de datos sin información lo que podría traer como consecuencia el desbalanceo de datos. Finalmente se realiza el entrenamiento del modelo y se evalúa de acuerdo con las métricas de desempeño que harán parte del proceso de investigación.

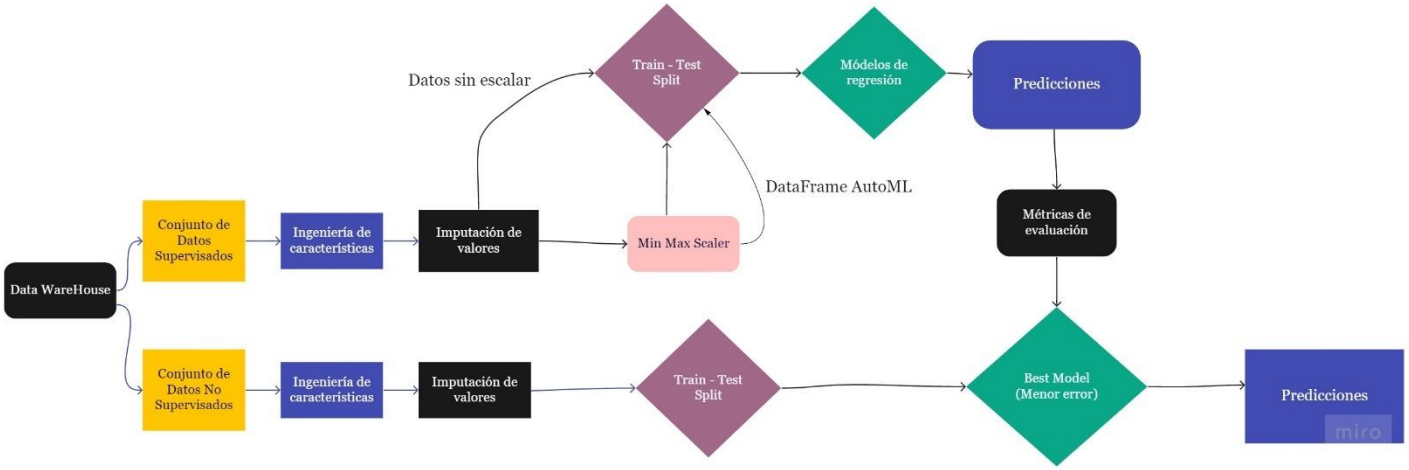


Fig. 7. Etapas de construcción del modelo predictivo

B. *Origen de los datos*

Los datos de los clientes inscritos en las pólizas de la aseguradora se encontraban inicialmente almacenados en una herramienta de Data WareHouse específico de la compañía, el cuál es un repositorio que funciona como servidor corporativo alojado en la nube y que colecciona los datos en comportamiento unificado y estructural para la futura ejecución analítica. De acuerdo con esta información, los especialistas de la aseguradora realizaron las distintas sentencias SQL y a partir de los cruces entre tablas estructuradas se obtuvieron dos dataset: DATA_SET_ENTRENAMIENTO y DATA_SET_TESTING.

C. Datos

Inicialmente, se cargaron los datos "Data_Set_Entrenamiento" a Google colaboratory, el cual es un sitio web que permite ejecutar y programar en Python desde el navegador. En cuanto a las características de los datos, constan de variables demográficas y también de variables de tipo moneda representadas en pesos colombianos (COP) correspondiente a los ingresos o salarios, rentabilidad y pago de compra por servicio. A continuación, se realizará la descripción de las características de los datos.

TABLA I
DESCRIPCIÓN DE LOS DATOS

Columnas o parámetros del dataset	**Descripción de las variables**
Cliente_id	ID del cliente afiliado a la aseguradora
Edad_cliente	Edad del cliente afiliado a la aseguradora
Sexo_cd	Sexo del cliente afiliado a la aseguradora
Estado_Civil_cd	Estado civil del cliente afiliado a la aseguradora
Nm_Estrato_Vivienda	Estrato del cliente afiliado a la aseguradora
Valor_Ingresos_Mensuales	Ingresos del cliente afiliado a la aseguradora
Nombre_Ciudad	Ciudad de residencia del cliente
Nombre_Departamento	Departamento de residencia del cliente
Número_Hijos	Número de hijos del cliente
Ocupación_Desc	Ocupación de las personas
Ramo_Id	Id del producto ofrecido por la aseguradora
Ramo_Desc	Detalle del producto ofrecido por la aseguradora
Resultado_Tecnico	La rentabilidad que le deja a la compañía desde el pago
Producción_Emitida	Dinero que le paga el cliente a la compañía por la compra de una póliza

D. Exploración de los datos

El presente dataset es un conjunto de datos de 2'933.186 registros (Filas), sin embargo, como se habló anteriormente muchas características tienen ausencia de información (*missing values*), puesto que tienen registrado valores de la siguiente manera: '?' y '-1'. En principio se realizó una copia de los datos en un nuevo DataFrame, y a partir de esto, se realiza el procedimiento

de exploración de datos haciendo un conteo de los datos que no tienen información o valor '?' en la variable de interés (Resultado técnico) y se encontraron 22 valores asignados con un símbolo "?", que fueron eliminados en la limpieza de datos.

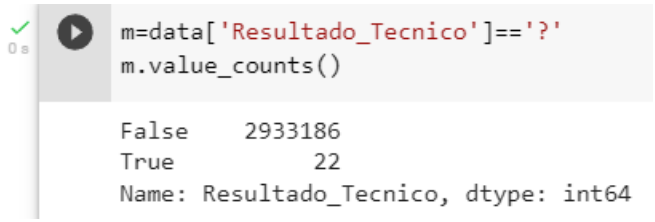


Fig. 8. Datos sin información para variable de interés

Al realizar un recuento estadístico sobre la variable objetivo (Resultado técnico) se visualiza datos atípicos con valores extremos en los mínimos y en los máximos, dado esto, para tratar el problema de negocio la compañía aseguradora testifica que estos datos son extremos y se decide eliminarlos al considerarlos un posible error en el almacenamiento de la información para estos registros; en correspondencia a la sugerencia de la compañía se elimina el percentil 99 y el percentil 1 para eliminar los outliers.

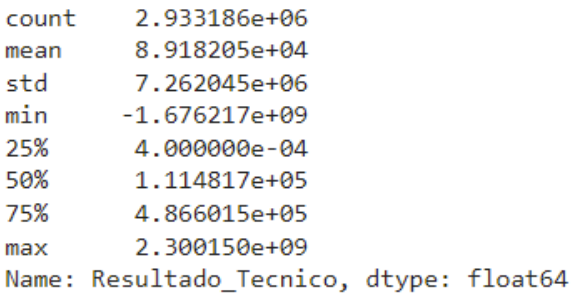


Fig. 9. R.E. de la variable objetivo

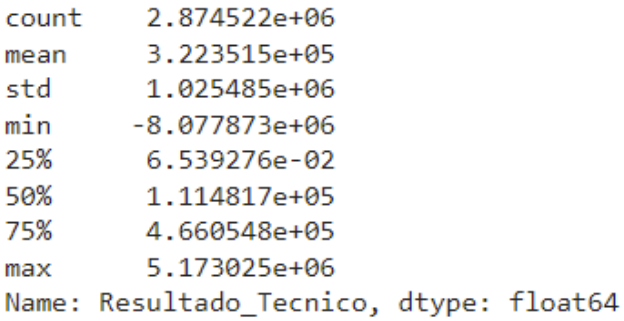


Fig. 10. R.E. de la variable objetivo sin percentil 1 y 99

Este procedimiento también se realizó para la variable “Producción emitida”, puesto que tiene un comportamiento similar que causaría desbalanceo en el entrenamiento del modelo.

```
count    2.874522e+06
mean     7.392987e+05
std      1.547905e+06
min     -1.446085e+08
25%      0.000000e+00
50%      1.586040e+05
75%      8.967840e+05
max      3.043288e+08
Name: Produccion_Emitida, dtype: float64
```

Fig. 11. R.E. de la variable Producción emitida

```
count    2.817071e+06
mean     6.824523e+05
std      1.150196e+06
min     -7.149960e+05
25%      0.000000e+00
50%      1.586220e+05
75%      8.596867e+05
max      6.239672e+06
Name: Produccion_Emitida, dtype: float64
```

Fig. 12. R.E. de la variable Producción emitida sin percentil 1 y 99

Además, se recuerda que la variable “Resultado técnico” y “producción emitida” son variables que tiene valores en pesos colombianos COP, por lo que se decide realizar un tipo de escalamiento manual de la variable, haciendo una división entre 1’000’000 de pesos para facilitar la exploración, visualización y limpieza.

E. Importancia de características

Como se mencionó anteriormente, el dataset exhibe distintas columnas con valores categóricas como son: La ciudad de residencia, departamentos, sexo, estrato o estado civil. De acuerdo con esto, se implementó un proceso que consiste en convertir todas las variables categóricas en datos numéricas. Para realizar este procedimiento se implementó un método conocido como Label Encoder, el cual le asigna un valor a cada valor de cada variable categórica. Detallando esto de manera específica, la función “sklearn.preprocessing.LabelEncoder” codifica etiquetas de una característica categórica en valores numéricos entre 0 y el número de clases menos 1, una vez instanciado, el método fit lo entrena (creando el mapeado entre las etiquetas y los

números) y el método transform transforma las etiquetas que se incluyan como argumento en los números correspondientes. El método fit_transform realiza ambas acciones simultáneamente. [7]

Label Encoder para las variables

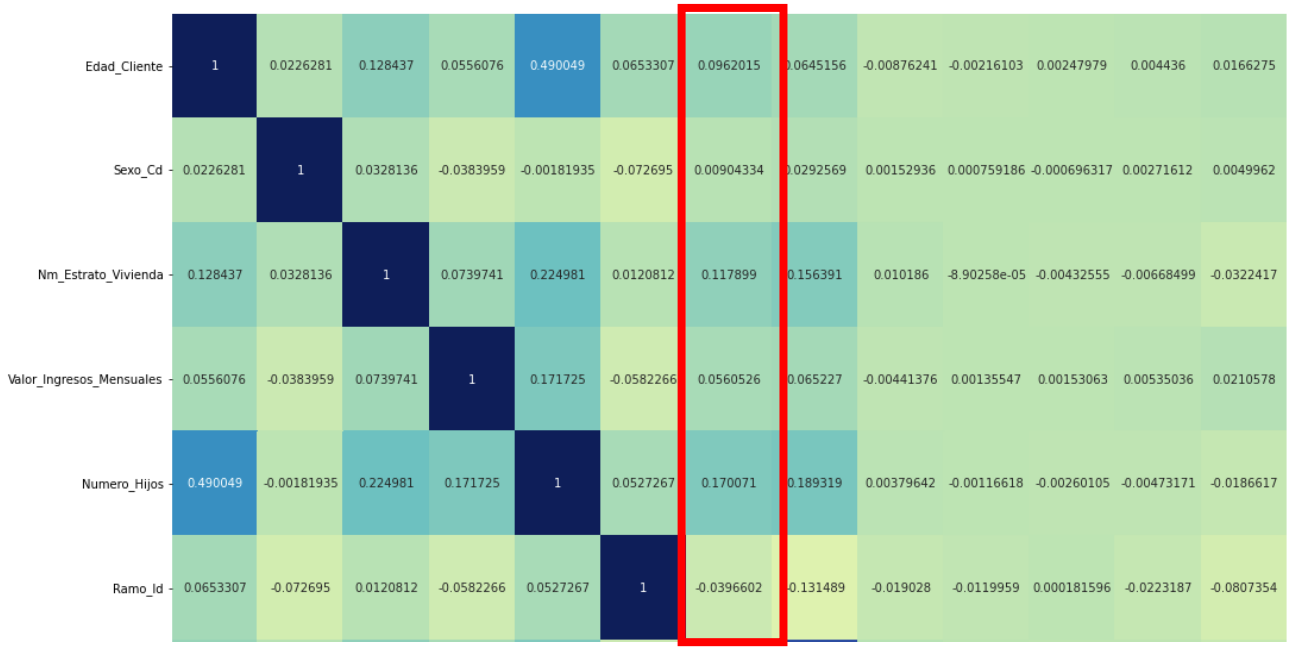
```
[ ]  data = data.drop(data[data['Sexo_Cd']=='-1'].index)
     data['Sexo_Cd']= pd.get_dummies(data["Sexo_Cd"])
     ########################################################################
     Encoder1= LabelEncoder()
     A=pd.DataFrame(Encoder1.fit_transform(data["Estado_Civil_cd"]))
     A.columns=['Estado_Civil']
     data = pd.concat([data, A], axis=1, join='inner')
     data = data.drop(columns = ['Estado_Civil_cd'])
     ########################################################################3#
     Encoder2= LabelEncoder()
     B=pd.DataFrame(Encoder2.fit_transform(data["Nombre_Ciudad"]))
     B.columns=['Ciudad']
     data = pd.concat([data, B], axis=1, join='inner')
     data = data.drop(columns = ['Nombre_Ciudad'])
     ########################################################################3#
     Encoder3= LabelEncoder()
     C=pd.DataFrame(Encoder3.fit_transform(data["Nombre_Departamento"]))
     C.columns=['Departamento']
     data = pd.concat([data, C], axis=1, join='inner')
     data = data.drop(columns = ['Nombre_Departamento'])
     ########################################################################3#
     Encoder4= LabelEncoder()
     D=pd.DataFrame(Encoder4.fit_transform(data["Ocupacion_Desc"]))
     D.columns=['Ocupación']
     data = pd.concat([data, D], axis=1, join='inner')
     data = data.drop(columns = ['Ocupacion_Desc'])
     ########################################################################
     Encoder= LabelEncoder()
     E=pd.DataFrame(Encoder.fit_transform(data["Ramo_Desc"]))
     E.columns=['Ramo']
     data = pd.concat([data,E], axis=1, join='inner')
     data = data.drop(columns = ['Ramo_Desc'])

     ########################################################################
     data['Valor_Ingresos_Mensuales'] = data['Valor_Ingresos_Mensuales'].replace(['?'], -1)
     data['Numero_Hijos'] = data['Numero_Hijos'].replace(['?'], -1)
     data['Valor_Ingresos_Mensuales']=data.Valor_Ingresos_Mensuales.astype(float)
     data['Numero_Hijos']=data.Numero_Hijos.astype(int)
```

Fig. 13. Label Encoder para datos originales

En la analítica de datos se vuelve fundamental el concepto de determinar la importancia de variables de todo el conjunto de datos, así como la información que le pueda entregar las demás variables a la variable de interés. Por consiguiente, en la exploración de datos se presenta *la matriz de correlación de spearman*. Este gráfico es un matriz de correlación basado en la correlación entre datos. Para mostrar los datos categóricos en la matriz representados de forma numérica se realizó el paso anteriormente mostrado (Label Encoder). El gráfico de matriz de correlación de spearman representa la fuerza y dirección entre dos variables el cuál es más robusto que el método de correlación de Pearson, puesto que es menos sensible a datos outliers (datos ruidosos). Para dar contexto a la anterior afirmación, las relaciones entre las variables se pueden explorar a través de distintos métodos estadísticos. El tipo de relación y la intensidad que existe entre distintas variables depende en buena medida, de la naturaleza de estas. [8]

La siguiente gráfica muestra que la variable de interés correspondiente al resultado técnico tiene mayor relación con la variable producción emitida, puesto que su valor de correlación es alto de acuerdo con el rango de valores positivos, así como existen otras variables que no tienen casi relación para la rentabilidad como la variable departamento, según infiere esta gráfica.



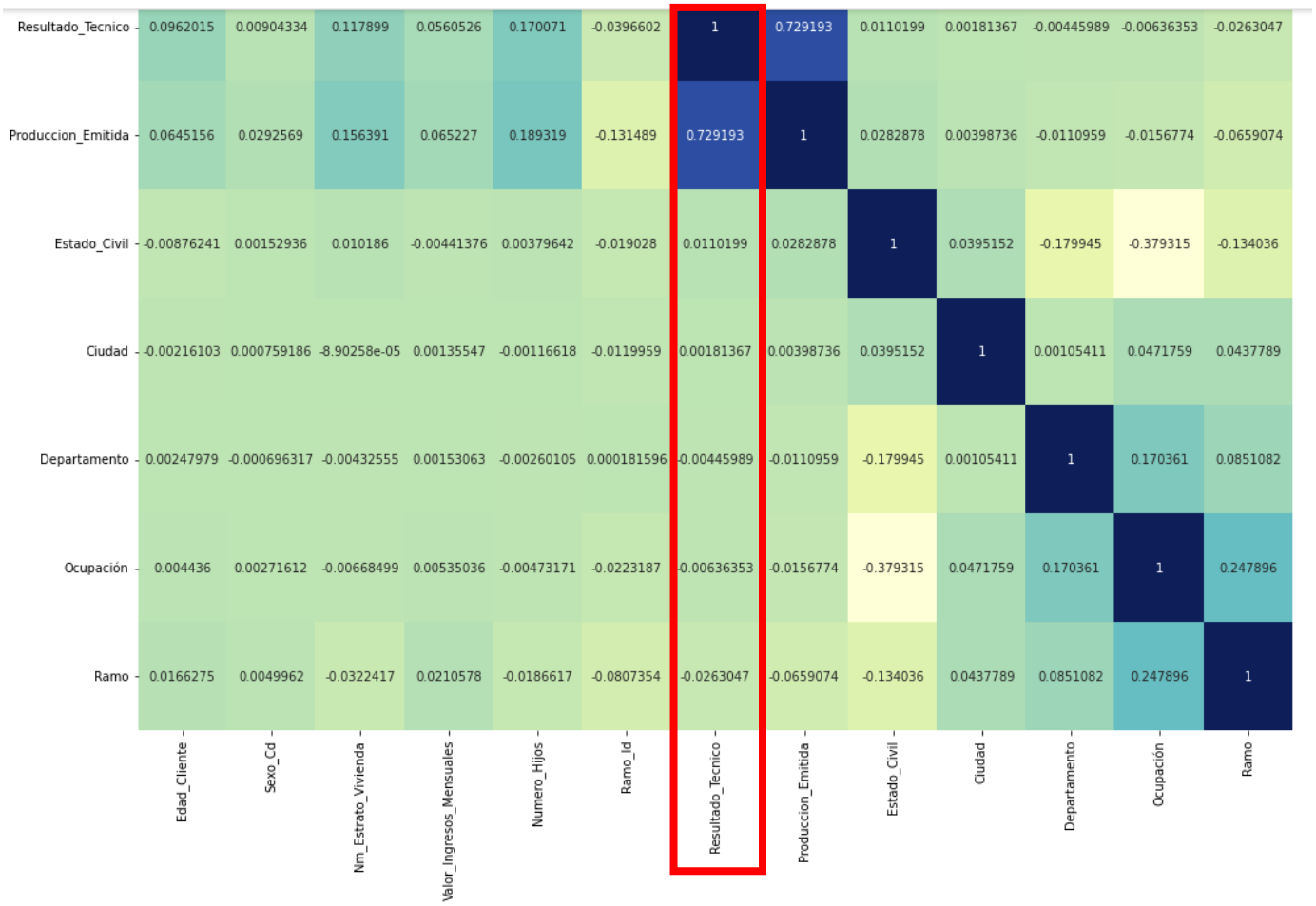


Fig. 14. Correlación de Spearman

Adicionalmente, como herramienta de exploración de los valores iniciales se graficó la entropía relativa. Este es un concepto que define el grado de desorden o poca información para el modelo que se va a implementar. Esta grafica indica que los valores de entropía entre variables cercanos a 1 tienen una naturaleza de aleatoriedad, lo que sugieren poca información para un modelo.

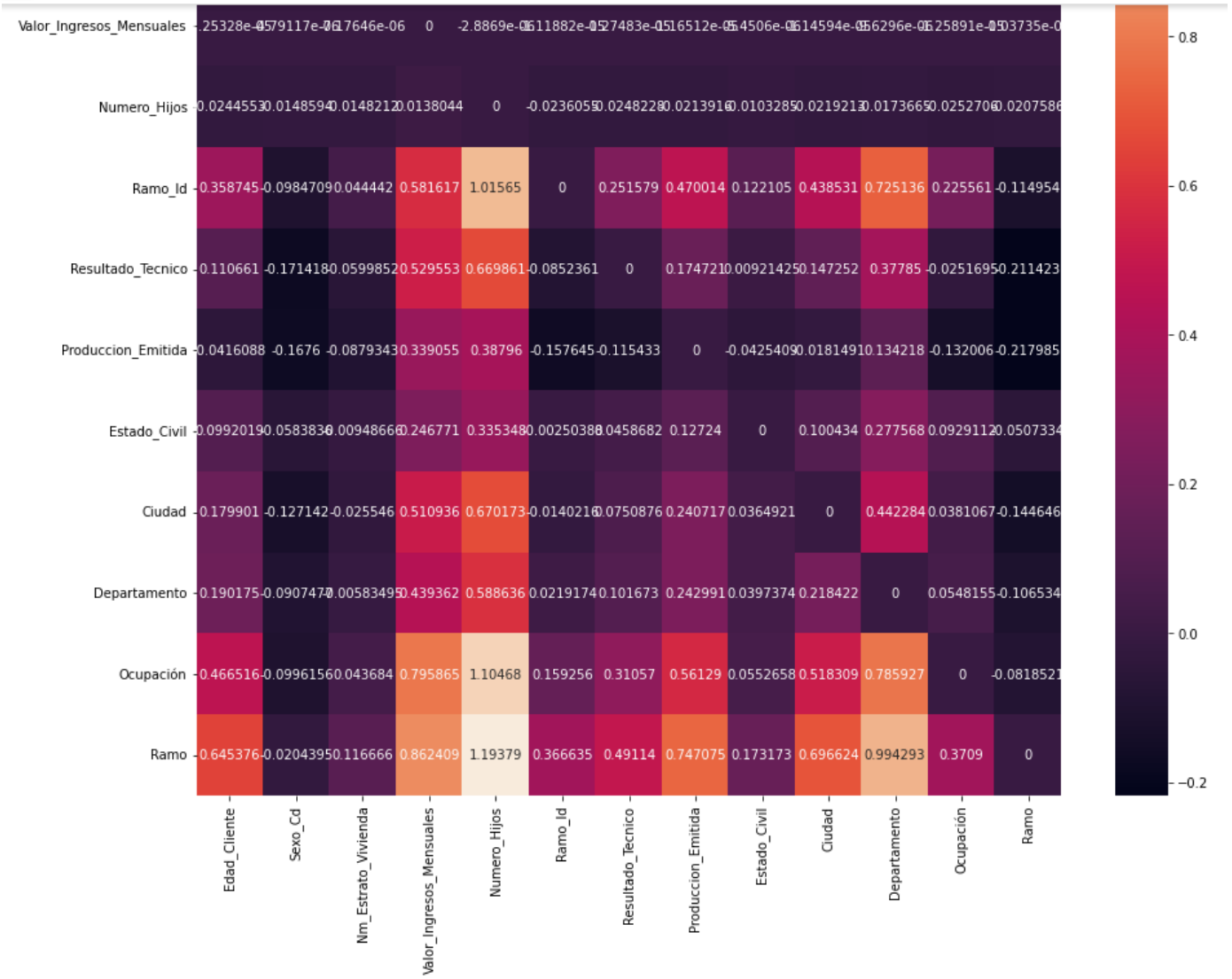


Fig. 15. Matriz de entropía relativa

F. Limpieza de datos: Cliente Id

La variable cliente es una variable que se considera eliminar debido a que solo es un número asociado al identificador de los distintos clientes que hay en las compañías. Es similar al número de cedula de cada persona, por lo tanto, no es un valor que resulte ser representativo que brinde información al modelo a la hora de querer saber la rentabilidad que deja un producto a la póliza de seguros.

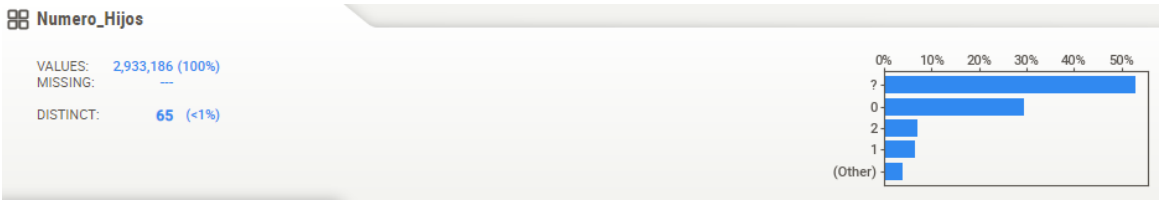
G. Limpieza de datos: Valores duplicados y valores nulos

Se corroboró de que el conjunto de datos almacenados en un dataframe el cual contienen los datos reales sin ningún tratamiento no tiene registros nulos ni registros duplicados, (registros de clientes exactamente iguales comprando la misma póliza). Luego, se verificó si después de la

eliminación del *cliente_id,* almacenado en otro dataframe podría contener datos duplicados, lo que arrojó como resultado que 52146 registros están duplicados, dando a concluir que clientes distintos tienen las mismas características. Aunque se pueda pensar que esto no sucede a menudo o resulte ser casi imposible, es decir, que clientes distintos tengan exactamente las mismas características demográficas y adicionalmente que al comprar la misma póliza éstos generen el mismo resultado técnico, se define que estos registros no aportan información al modelo, pues es exactamente la misma información repetida distintas veces, por tanto, se deciden eliminar.

H. Limpieza de datos: Número de hijos

Realizando la exploración sobre la característica actual, se identifica que el 53% de los datos no tienen información y adicionalmente los demás datos contienen muchos outliers. El tratamiento que se le realiza a la variable es con base al conocimiento de los clientes para la compañía y del promedio de hijos en Colombia de acuerdo con las encuentras estadísticas fundamentadas en el DANE. De acuerdo con la información descrita en la Fig. 16, se determina que el 1.6 millones de datos sin información se le imputó el valor de cero y tomando los valores arrojados en las encuestas y en la testificación de la compañía, los clientes que tienen más de 5 hijos se decidieron aproximar al valor máximo de 5 para evitar el desbalanceo de datos.



TOP CATEGORIES		
?	1.6M	53%
0	867,405	30%
2	206,025	7%
1	194,538	7%
3	78,817	3%
4	22,111	<1%
5	6,620	<1%
6	2,203	<1%
7	806	<1%
8	357	<1%
10	248	<1%
20	154	<1%
9	143	<1%
30	90	<1%
11	59	<1%
12	34	<1%
21	32	<1%
(Other)	282	<1%
ALL	2.9M	100%

Fig. 16. Distribución variable número de hijos

I. Limpieza de datos: Estado civil

A pesar de ser otra característica que no presenta suficiente información, para el desarrollo del modelo, se decide tener en cuenta a esta característica y no eliminar variables que puedan aportar información en las regresiones. A continuación, se presenta como era la distribución de los valores para la variable en mención:

```
-1    1624613
 3     363726
 2     222474
 4     201245
 5     150007
 6     109937
 1      76763
 0      16160
Name: Nm_Estrato_Vivienda, dtype: int64
```

Fig. 17. Valores únicos Estrato Vivienda

Luego, se realiza la conversión de datos categóricos en datos numéricos mediante la función *get_dummies* de PANDAS. Es importante explicar que en la función *get_dummies* se ajusta el parámetro de *drop_first=True*, haciendo mención que la idea de convertir esta variable categórica a numérica considera la definición de evitar la multicolinealidad en los modelos de regresión. La colinealidad significa que una de las características sea una combinación lineal de otra, convirtiéndose en un problema significativo para los modelos de aprendizaje automático.

J. Imputación de valores: Ingresos mensuales

En el capítulo de hipótesis, se describió que la columna de ingresos mensuales tiene el 90% de datos sin información como se presenta en la Fig. 18.

Fig. 18. Ingresos mensuales

El tratamiento que se le dio a esta variable tiene que ver con las imputaciones de valores. En primer lugar, se crea una función de tal forma que filtre los salarios por cada ocupación y se imputa un valor de ingreso mensual a los datos desconocidos con base a la mediana de los salarios conocidos para esa ocupación. De esta manera se llenaron algunos campos que estaban sin información. Posteriormente se notó que existían registros con ausencia de información de ingresos mensuales y ocupación, por tal razón el objetivo de la función no cubrió este problema, es entonces donde se recurrió al imputador iterativo IterativeImputer de la biblioteca de sikit learn. Este método funciona como un modelo predictivo y, además, usa técnicas de imputación e intenta adaptarse a los escenarios con características de modelos de regresión, resaltando que el funcionamiento depende mucho de la alteración y distribución de los datos. Se explica además que se modela cada característica con valores perdidos en función de otras características, y usa esa estimación para la imputación. Lo hace de una manera iterativa por turnos: en cada paso, una columna de características se designa como salida y, y las otras columnas de características se tratan como entradas X. Se ajusta un regresor en (X, y) para y conocida. Luego, el regresor se usa para predecir los valores perdidos de y. Esto se hace para cada característica de forma iterativa y luego se repite para las rondas de imputación max_iter. Se devuelven los resultados de la ronda de imputación final. [9]

K. Escalamiento de datos

Para iniciar con la implementación de los modelos de regresión, se agruparán los datos de las características predictoras en un DataFrame llamado “X”, es decir, a todas las variables que se decidieron conservar de acuerdo con la importancia de los datos, luego, en otro DataFrame

catalogado como “y”, se almacenará la variable dependiente o variable a predecir correspondiente al “resultado técnico”. Luego, se realiza un escalamiento a los datos contenidos en la variable X. Es decir, se utiliza el método de escalamiento de datos muy popular “Escala Mínima máxima”, de modo que este escalador toma cada valor y resta el mínimo y luego divide por el rango (máximo-mínimo), dando como resultado valores que oscilan entre cero (0) y uno (1).

L. Train Test Split

Es una operación común propio del aprendizaje supervisado que consiste en hacer dos divisiones en los dataframe “X” y “y”, después de la limpieza y exploración de datos: la primera división contiene los datos de entrenamiento que se usará para entrenar y calibrar el modelo de regresión y la otra división contiene los datos de prueba. Para realizar esta operación se usa el método *train_test_split* de la librería de scikit learn que recibe como parámetros los dataframes y mediante la opción “test_size” se ajusta la proporción o parte del dataset que se dejará como datos de prueba.

Para el caso del modelamiento de la regresión lineal, se decidió entrenar con el 20% de datos de entrenamiento y 80% para datos de prueba, la razón de proporcionar los datos de entrenamiento y testeo de esta manera responde al principio de invarianza estadística, el cual contiene los conceptos de exhaustividad y suficiencia, en este caso el dataset de entrada es una muestra con muchos registros, de manera que al particionar los datos de esta manera se puede modelar la regresión permitiendo que no haya perdida de información.

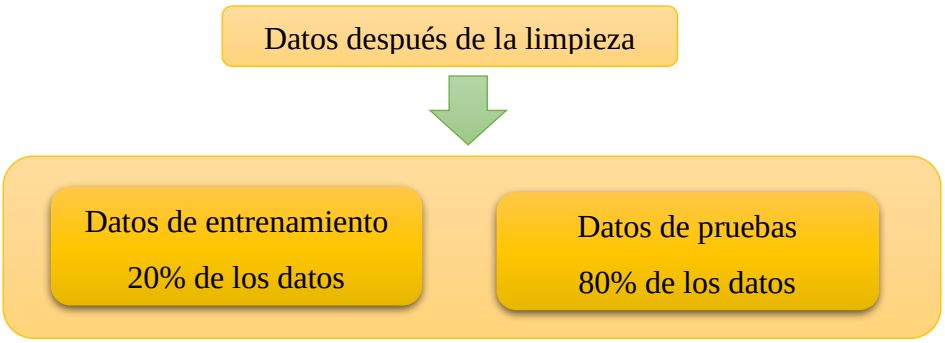


Fig. 19. Train Test Split Data

M. Métodos de modelamiento: Fit, Fit_Transform, predict

Dentro del proceso para la ejecución de un modelo predictivo desarrollado usando como herramienta el lenguaje de programación Python, usando bibliotecas las cuales consideran distintas herramientas que son soporte para la construcción un modelo de aprendizaje automático supervisado. El método FIT usado en los modelos de regresión se usa para hacer las iteraciones de aprendizaje, para el caso del escalamiento o imputación de datos se calcula cuál es la varianza y la media de todos los parámetros que participan en el entrenamiento y mediante el método transform se aplican los cálculos previos para aplicarlos a los demás conjuntos para escalarlos de manera uniforme. El método predict predice cuál es el valor de la variable objetivo (Resultado técnico) después de la ejecución del modelo de regresión.

X. HERRAMIENTAS DEL MODELAMIENTO

El costo computacional de cómo se abordó el problema de negocio se debió en gran parte a la cantidad del espacio muestral del tamaño de más de millón de registros de entradas para el modelamiento. Como se mencionó anteriormente, las iteraciones de las regresiones consideraron distintas herramientas que hicieron parte de la investigación y el desarrollo de la solución del problema de negocio; su ejecución se realizó bajo el entorno del lenguaje Python y las principales herramientas fueron:

TABLA II

HERRAMIENTAS

sklearn.model_selection	KFold, cross_val_score, train_test_split
sklearn.linear_model	LinearRegression, RidgeCV
sklearn.tree	DecisionTreeRegressor
sklearn.preprocessing	OrdinalEncoder, OneHotEncoder, LabelEncoder
sklearn.ensemble	RandomForestRegressor, StackingRegressor

AUTO ML – H2O

H2O es una plataforma de aprendizaje automático en memoria distribuida y de código abierto, compatible con los algoritmos estadísticos y de aprendizaje automático más utilizados. H2O tiene una funcionalidad de AutoML que permite automatizar el proceso de generación de una gran cantidad de modelos, con el fin de encontrar el mejor modelo de aprendizaje automático para el problema que se desea resolver. Aunque H2O puede facilitar el proceso de creación de modelos de aprendizaje automático, igualmente se requiere de conocimiento y experiencia en ciencia de datos para que los modelos de aprendizaje automático generados sean de alto rendimiento. Dentro del desarrollo del proyecto esta herramienta proporciona la ventaja de permitir realizar una gran cantidad de tareas relacionadas con el modelado que normalmente requerirían muchas líneas de código. Fundamentalmente, esto libera tiempo para centrarse en otros aspectos de relacionados con el problema de negocio, como en el análisis profundo del problema y un buen preprocesamiento de los datos, que particularmente en este problema fue bastante importante. El proceso llevado a cabo en el desarrollo del problema con esta librería consistió inicialmente en la importación de los módulos necesarios e inicialización del clúster de H2O. Como se mencionó anteriormente, la ejecución de los modelos se realizó en el repositorio de Google colaboratory, por lo tanto, para la

importación de esta librería se requirió de un procesamiento complejo debido a la ejecución de 8 modelos de regresión, este costo se lograba identificar en las iteraciones realizadas en máquinas suministradas dentro del repositorio en un tiempo aproximado de 1 hora de ejecución para toda la recopilación de los modelos de regresión con el objetivo de mejorar el aprendizaje. El procesamiento de la librería H2o tuvo las siguientes especificaciones:

```
h2o.init(nthreads=-1, max_mem_size=12)
```

```
Checking whether there is an H2O instance running at http://localhost:54321 ..... not found.
Attempting to start a local H2O server...
  Java Version: openjdk version "11.0.11" 2021-04-20; OpenJDK Runtime Environment (build 11.0.11+9-Ubuntu-0ubuntu2.18.04); OpenJDK 64-Bit Server VM (build 11.0.11+9-Ubuntu-0ubuntu2.18.04, mixed mode,
  Starting server from /usr/local/lib/python3.7/dist-packages/h2o/backend/bin/h2o.jar
  Ice root: /tmp/tmp1cpbbzw1
  JVM stdout: /tmp/tmp1cpbbzw1/h2o_unknownUser_started_from_python.out
  JVM stderr: /tmp/tmp1cpbbzw1/h2o_unknownUser_started_from_python.err
  Server is running at http://127.0.0.1:54321
Connecting to H2O server at http://127.0.0.1:54321 ... successful.
```

H2O_cluster_uptime:	03 secs
H2O_cluster_timezone:	Etc/UTC
H2O_data_parsing_timezone:	UTC
H2O_cluster_version:	3.34.0.3
H2O_cluster_version_age:	1 month and 22 days
H2O_cluster_name:	H2O_from_python_unknownUser_twis23
H2O_cluster_total_nodes:	1
H2O_cluster_free_memory:	12 Gb
H2O_cluster_total_cores:	2
H2O_cluster_allowed_cores:	2
H2O_cluster_status:	locked, healthy
H2O_connection_url:	http://127.0.0.1:54321
H2O_connection_proxy:	{"http": null, "https": null}
H2O_internal_security:	False
H2O_API_Extensions:	Amazon S3, XGBoost, Algos, AutoML, Core V3, TargetEncoder, Core V4
Python_version:	3.7.12 final

Fig. 20. Especificación de procesamiento H2O Auto ML

XI. RESULTADOS

Las siguientes figuras muestran el resultado después de implementar todas las modelaciones basadas en regresión.

A. Modelo 1: Modelo de regresión basado en árboles de decisión (DecisionTreeRegressor)

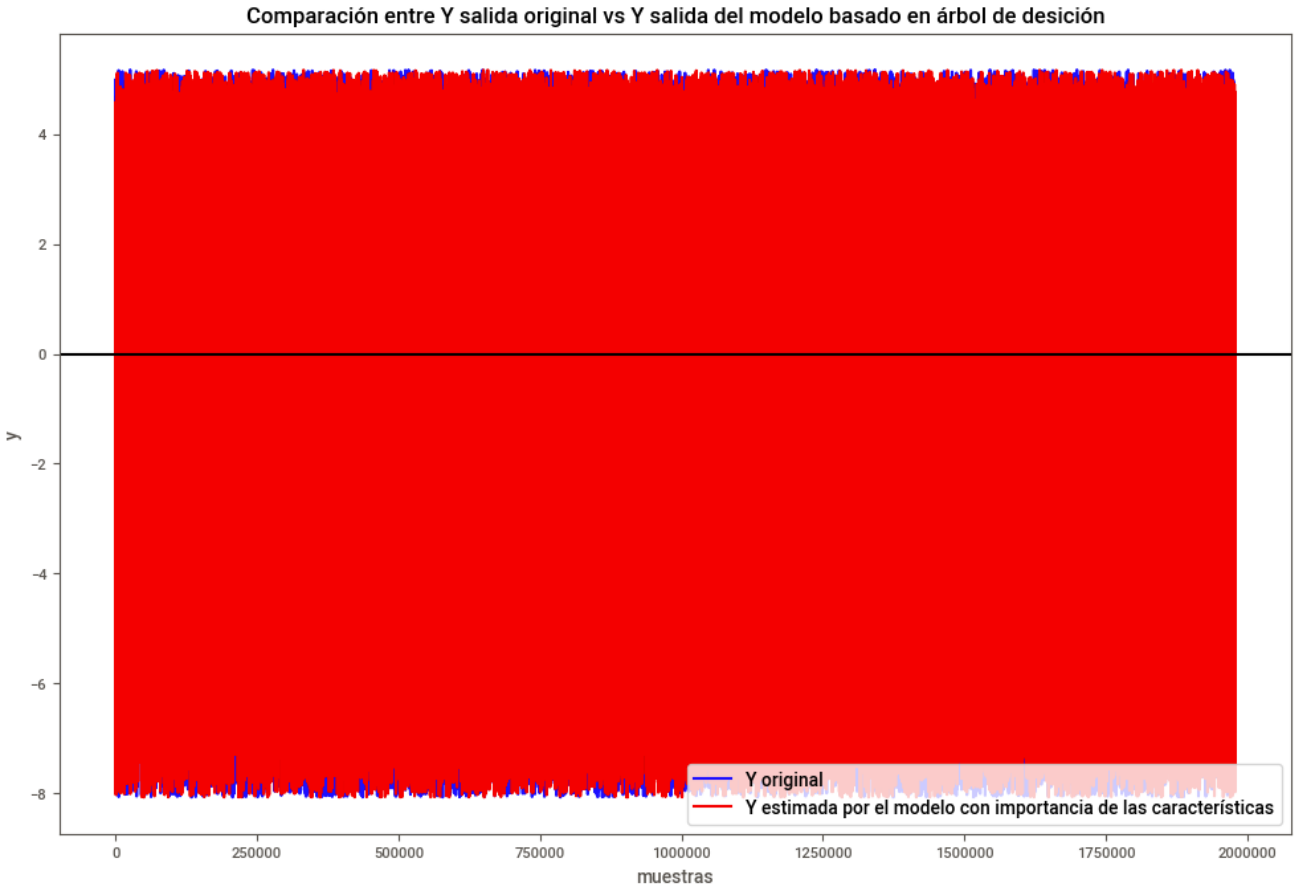


Fig. 21. Rentabilidad real VS Rentabilidad computada (Modelo Arboles de decisión)

B. Modelo 2: Modelo de regresión múltiple usando validación cruzada (MLR)

Teniendo en cuenta el concepto de la regresión lineal múltiple anteriormente explicado, en el cuál, el objetivo es predecir la variable respuesta o también llamada variable objetivo. El error presentado cuando se entrena el modelo suele ser el error que tuvo cuando se realizó el entrenamiento bajo el espacio muestral con el que fue entrenado. A partir de esto se decidió implementar el concepto de validación cruzada, el cual es una estrategia de validación que consiste en obtener un subconjunto de datos de test dentro de los datos de entrenamiento para hacer

evaluaciones internas, de modo que abarque todo el espacio muestral desde distintas perspectivas, resultando como consecuencia mejorar la estimación predictiva a nuevas observaciones.

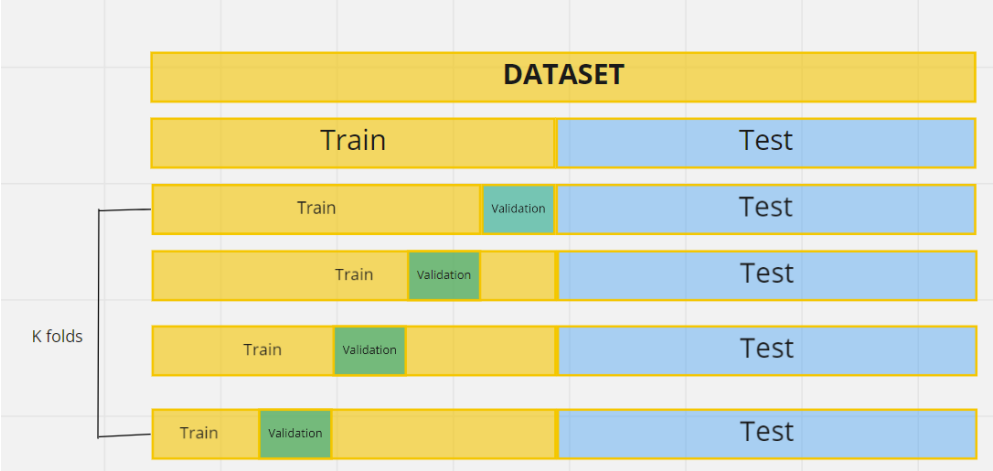


Fig. 22. Entrenamiento con estrategia de Cross Validation

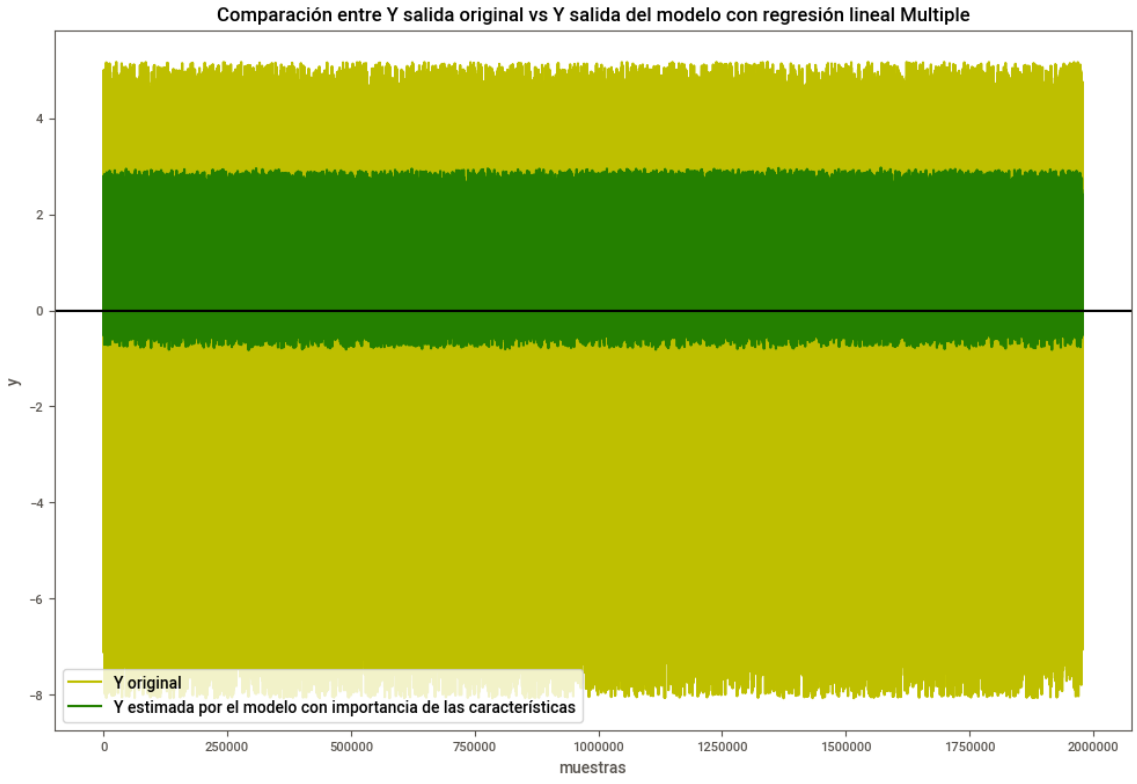


Fig. 23. Rentabilidad real VS Rentabilidad computada (Modelo de regresión lineal Múltiple)

C. Modelo 3: Modelo de regresión múltiple usando random forest

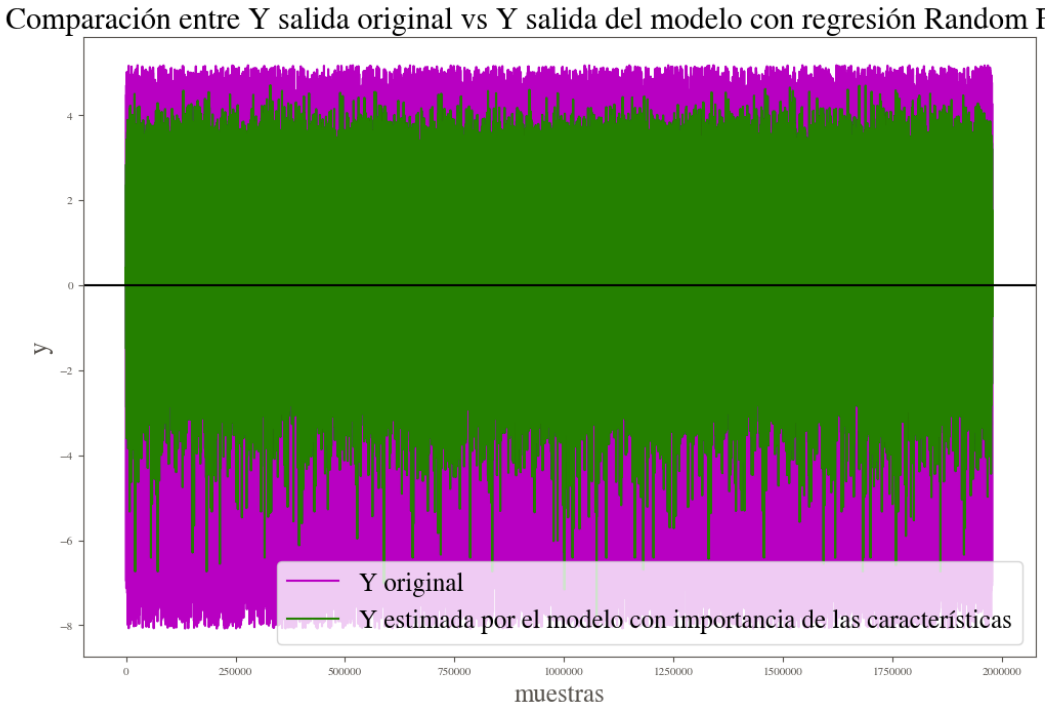


Fig. 24. Rentabilidad real VS Rentabilidad computada (Modelo de regresión Random Forest)

D. Modelo 3: Modelo de regresión basado en regresión Robusta

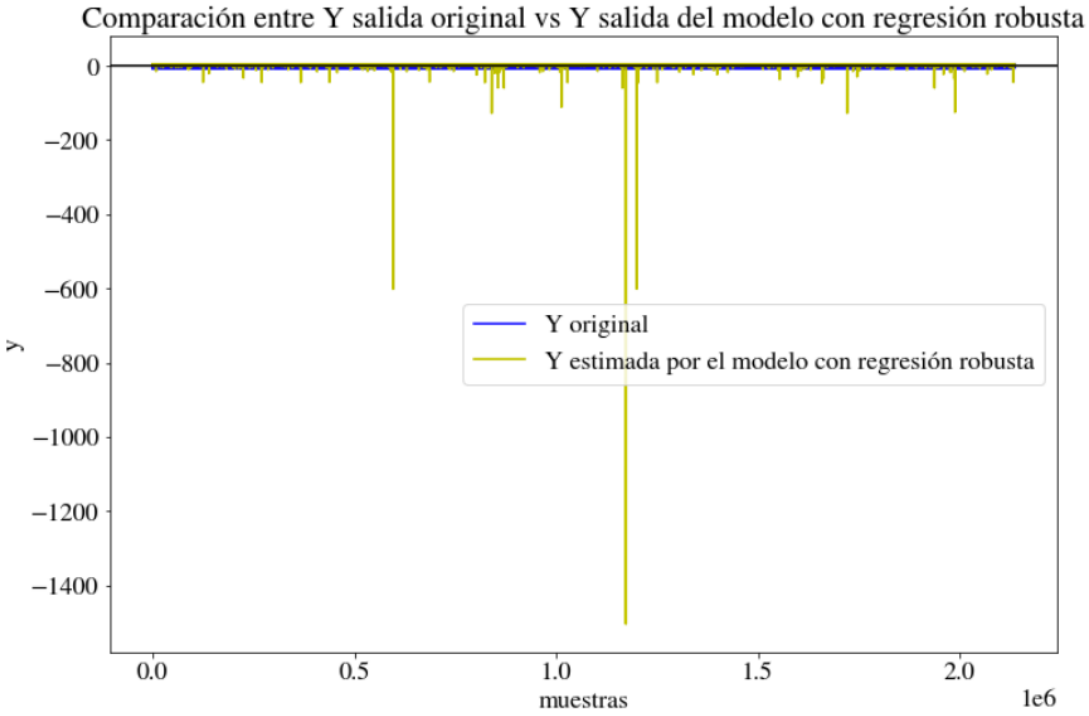


Fig. 25. Rentabilidad real VS Rentabilidad computada (Modelo de regresión Robusta)

E. Modelo 4: Modelo de regresión basado en métodos de ensamble

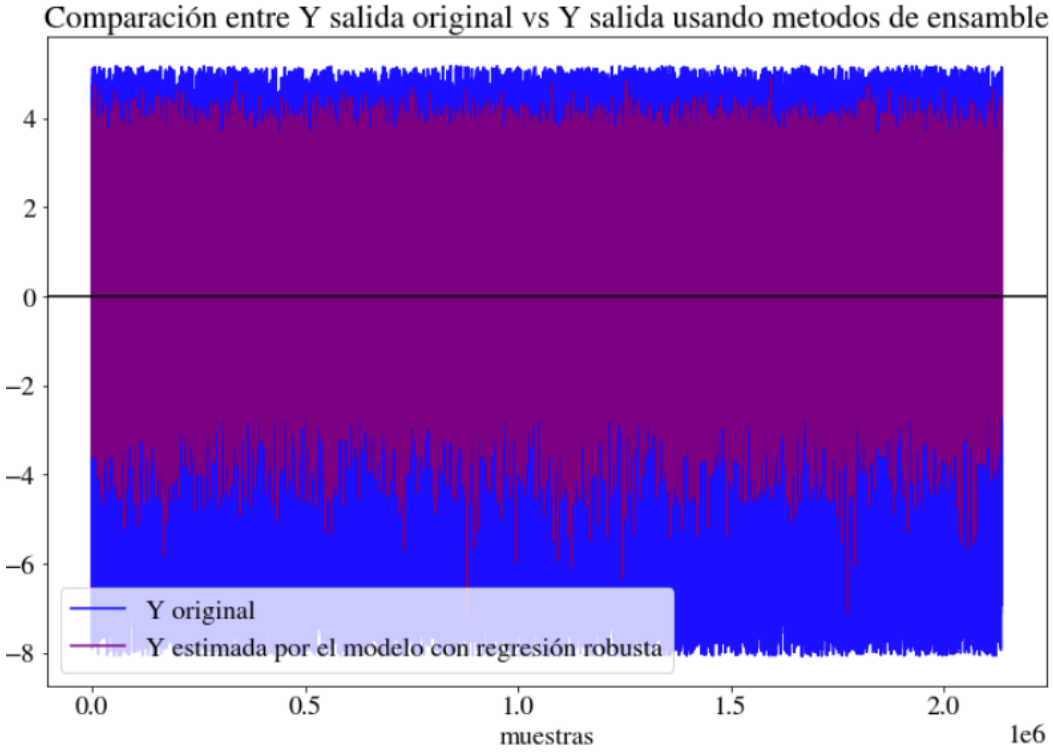


Fig. 26. Rentabilidad real VS Rentabilidad computada (Modelo de regresión con métodos de ensamble)

F. Modelos arrojados por H2O AutoML

Posteriormente, se procede con la generación y modificación de los datos previamente preprocesados para que sean compatibles y funcionen correctamente con H2O, los datos se deben almacenar completos en un objeto tipo H2OFrame. Es importante resaltar que se debe indicar en el H2OFrame creado que la variable objetivo a predecir es Resultado Técnico, y las demás corresponderán a las variables regresoras o explicativas del modelo. El siguiente paso es ejecutar la función de AutoML con determinados parámetros escogidos, debido a la gran cantidad de registros que se tienen en el data set se eligió entrenar un máximo de 10 modelos para evitar un colapso de memoria en el clúster. Durante este proceso se entrenan algoritmos de aprendizaje automático aleatorios con sus respectivos hiperparámetros determinados por la misma función, de igual manera, se entrenan dos tipos de modelos más robustos y optimizados empleando métodos de ensamble, uno de estos se genera a partir de todos los modelos obtenidos y el segundo se genera a partir de los mejores modelos obtenidos para cada tipo o familia.

En la siguiente figura se observan los modelos generados para este problema junto con las métricas de evaluación obtenidas para cada uno.

model_id	mean_residual_deviance	rmse	mse	mae
StackedEnsemble_AllModels_1_AutoML_1_20211129_180040	0.734371	0.856955	0.734371	0.299545
StackedEnsemble_BestOfFamily_2_AutoML_1_20211129_180040	0.734632	0.857107	0.734632	0.299707
StackedEnsemble_AllModels_4_AutoML_1_20211129_180040	0.734687	0.857139	0.734687	0.30107
GBM_2_AutoML_1_20211129_180040	0.734761	0.857182	0.734761	0.300784
StackedEnsemble_BestOfFamily_5_AutoML_1_20211129_180040	0.734949	0.857292	0.734949	0.300951
StackedEnsemble_AllModels_3_AutoML_1_20211129_180040	0.735382	0.857544	0.735382	0.300586
GBM_3_AutoML_1_20211129_180040	0.735996	0.857902	0.735996	0.302125
StackedEnsemble_BestOfFamily_4_AutoML_1_20211129_180040	0.736653	0.858285	0.736653	0.300874
StackedEnsemble_BestOfFamily_1_AutoML_1_20211129_180040	0.73815	0.859156	0.73815	0.30461
GBM_4_AutoML_1_20211129_180040	0.738836	0.859555	0.738836	0.304802
GBM_1_AutoML_1_20211129_180040	0.739239	0.85979	0.739239	0.305129
StackedEnsemble_BestOfFamily_3_AutoML_1_20211129_180040	0.743716	0.86239	0.743716	0.302177
StackedEnsemble_AllModels_2_AutoML_1_20211129_180040	0.743737	0.862402	0.743737	0.302026
GLM_1_AutoML_1_20211129_180040	0.764184	0.874176	0.764184	0.339642
XGBoost_2_AutoML_1_20211129_180040	0.764801	0.874529	0.764801	0.31361
DRF_1_AutoML_1_20211129_180040	0.771838	0.878543	0.771838	0.316805
XGBoost_1_AutoML_1_20211129_180040	0.791218	0.889504	0.791218	0.322225

Fig. 27. Modelos de regresión generados por librería auto ML

De acuerdo con los resultados se observa que el mejor modelo obtenido corresponde a uno de los dos generados por los métodos de ensamble (Stacked Ensemble All Models). En la Fig. 28 se puede observar cómo fue el rendimiento del modelo en sus etapas de entrenamiento y validación y en la Figura XX la importancia o contribución de cada uno del modelo en este modelo final generado.

```
ModelMetricsRegressionGLM: stackedensemble
** Reported on train data. **

MSE: 0.6704909513244691
RMSE: 0.8188351185217139
MAE: 0.29178369554315303
RMSLE: NaN
R^2: 0.28765494855256646
Mean Residual Deviance: 0.6704909513244691
Null degrees of freedom: 9880
Residual degrees of freedom: 9873
Null deviance: 9300.440424878088
Residual deviance: 6625.12109003708
AIC: 24109.182239069618
```

```
ModelMetricsRegressionGLM: stackedensemble
** Reported on cross-validation data. **

MSE: 0.7343714650262221
RMSE: 0.8569547625319682
MAE: 0.29954548089000443
RMSLE: NaN
R^2: 0.25889940307930825
Mean Residual Deviance: 0.7343714650262221
Null degrees of freedom: 494543
Residual degrees of freedom: 494536
Null deviance: 490055.4121025885
Residual deviance: 363179.00179992797
AIC: 1250787.4153599297
```

Fig. 28. Mejor modelo de regresión sin Cross validación VS con Cross Validation

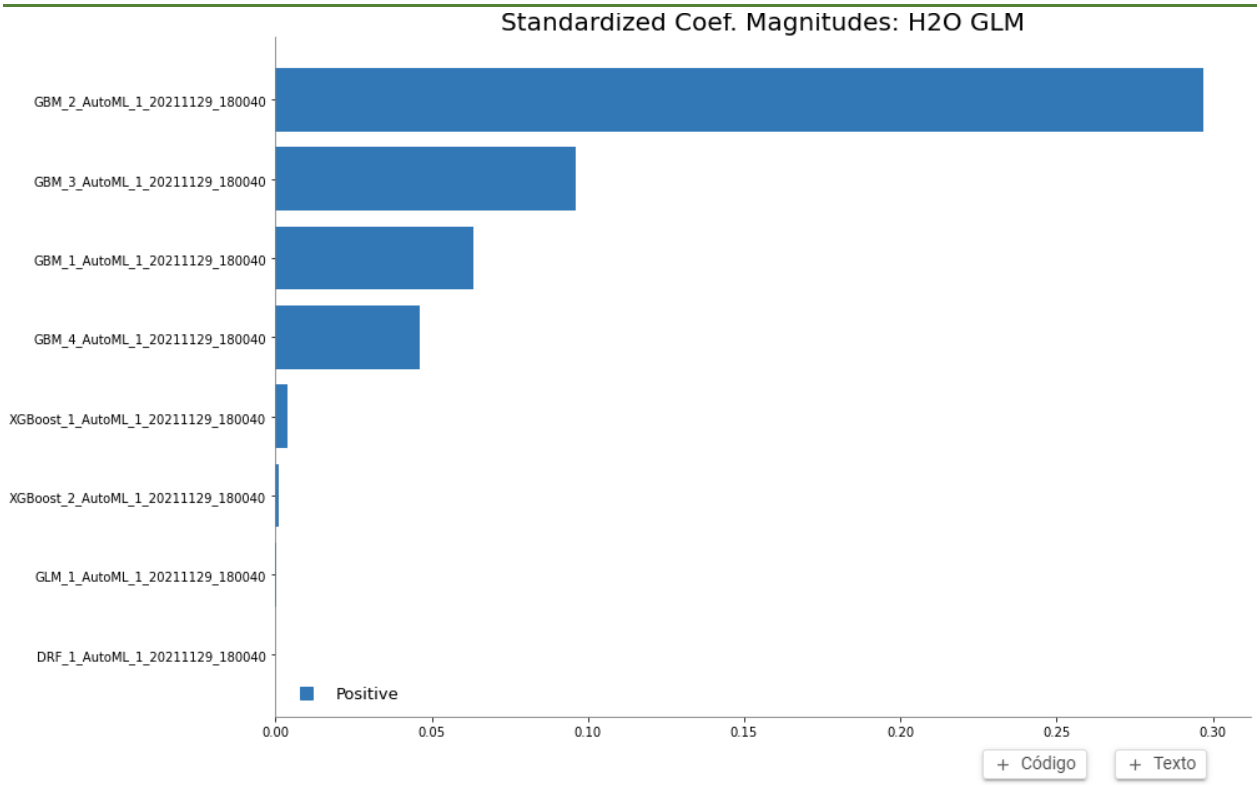


Fig. 29. Importancia de aporte de modelos de regresión para ensamble

Con los resultados anteriores se procede a evaluar el modelo con la partición de los datos generada para test dentro del conjunto de datos de entrenamiento y se obtuvieron los siguientes resultados.

```
best_model.model_performance(test)
```

```
ModelMetricsRegressionGLM: stackedensemble
** Reported on test data. **

MSE: 0.7317618540329128
RMSE: 0.8554308002596779
MAE: 0.2980739973526825
RMSLE: NaN
R^2: 0.2612758932177309
Mean Residual Deviance: 0.7317618540329128
Null degrees of freedom: 1978016
Residual degrees of freedom: 1978009
Null deviance: 1959375.531424639
Residual deviance: 1447437.3872286202
AIC: 4995652.067603278
```

Fig. 30. Mejor modelo de regresión para datos de prueba

Finalmente se guarda este modelo para poderlo importar cuando se requiera utilizar y no tener que realizar nuevamente todo el proceso de entrenamiento, ya que, la función de AutoML tarda un tiempo considerable en entrenar todos los modelos.

G. Métricas de evaluación

Mean Absolute Percentage Error

Esta métrica corresponde a ser uno de los principales evaluadores de los modelos de regresión pues se asocia como el pronóstico de demanda o el tamaño del error absoluto dado en porcentaje. Lo que significa que todas las diferencias entre las rentabilidades reales Vs las rentabilidades computadas se ponderan por igual en el promedio. Desde una perspectiva asociada al problema de negocio, esta métrica define la diferencia en pesos colombianos COP de las equivocaciones que tiene el modelo predictivo cuando intenta computar los resultados técnicos. Se debe recordar que el tratamiento de la variable resultado técnico tuvo una división en 1’000’000 para comodidad en el trabajo del modelo, por lo tanto, por cada valor de MAE se interpreta como la equivocación que se obtuvo de los distintos modelos de regresión medido en dinero de las rentabilidades de los clientes.

Mean Absolute Error

Este método de evaluación del desempeño del modelo considera el promedio de la diferencia absoluta entre los valores predichos y el valor observado.

TABLA III
MÉTRICAS DE EVALUACIÓN

Modelo de regresión	*Mean Absolute Percentage Error*	*Mean Absolute Error*
Decision Tree Regressor	19.37×10^{12}	0.4445...
Linear Regression	8.42×10^{12}	0.3374….
Random Forest Regressor	17.46×10^{12}	0.3582...
Robusta: RANSAC Regressor	3.29×10^{12}	0.3155…
Stacked Ensemble All Models	4.44×10^{12}	0.2980…

H. Métricas de evaluación de acuerdo con el problema de negocio

Para escoger el mejor modelo que se adapta con el espacio muestral del problema de negocio se implementó la relación entre la sumatoria de todas las rentabilidades reales para cada uno de los registros del conjunto de datos y las sumatoria de todas las rentabilidades que fueron calculadas por cada uno de los entrenamientos. Es válido aclarar que la relación está dada por una división entre ambos valores, considerando que los valores más próximos a 1 serían los que sugieren tener más similitud con los valores reales.

TABLA IV
PORCENTAJE DE IGUALDAD Y REAL VS Y COMPUTADA

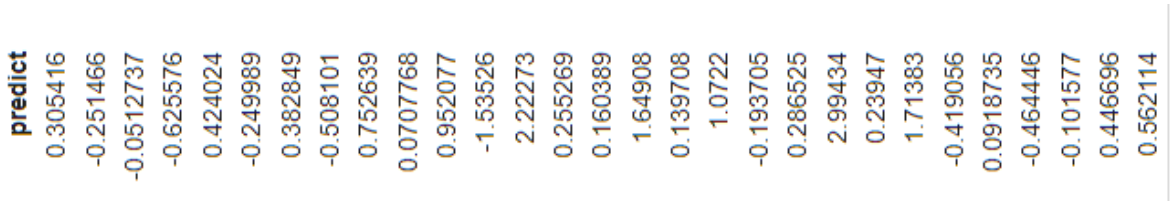
Modelo de regresión	Mean Absolute Percentage Error
DecisionTreeRegressor	0.929
LinearRegression	0.999
RandomForestRegressor	0.921
Robusta: RANSAC Regressor	1.69
Stacking Regressor	0.999

Resultado_Tecnico	predict
1.54281	1.1329
-0.00144442	-0.134635
0.406378	0.49427
0.00827439	0.00141367
0.150948	0.15584
-0.0152981	0.00074009
0	-0.0291897
-0.0380286	-0.0928657
0.0108844	0.0383844
0.218446	0.109136

Fig. 31. Predicciones entre valores reales VS valores computarizados

XII. CONSIDERACIÓN A PRODUCCIÓN

En correspondencia con las etapas de abordaje del problema de negocio, explicadas en la sección Proceso de analítica, se obtienen los datos que fueron abstraídos del DataWare House de la compañía, denominados: DATA_SET_TESTING. El presente dataset es un conjunto de datos de 900’000 registros (Filas), en el que contienen las cotizaciones que ha recibido la compañía de distintos clientes. Como bien se explicó, no representan conversiones de ser clientes afiliados a algunas de los servicios que ofrece la compañía sino el manifiesto o indicio de convertirse o no convertirse en clientes potenciales. Estos datos fueron cargados al mismo repositorio de Google colaboratory y recibieron el mismo tratamiento de datos, tal cual como se realizó con el data set denominado: DATA_SET_ENTRENAMIENTO. La diferencia consiste en que este conjunto de datos, al representar únicamente las cotizaciones de la compañía, no tienen la variable objetivo “Resultado técnico”. Es válido aclarar que este dataset recibió el mismo tratamiento debido a la similitud de contener las mismas características sociodemográficas o parámetros y cabe mencionar que también contenían ausencia de información (*missing values)*. De acuerdo con los resultados de todos los entrenamientos de modelos de regresión para predecir la rentabilidad o resultado de negocio del actual problema de negocio, se decidió escoger el mejor evaluado por las métricas de desempeño escogidas, en concreto, se decidió escoger el método de regresión basado en métodos de ensamble obtenido mediante la librería AUTO ML. Así pues, cuando se realizó la limpieza y tratamiento de datos del conjunto DATA_SET_TESTING, se envían como las nuevas entradas del modelo escogido, para la predicción de la rentabilidad para los clientes que ha realizado las cotizaciones, en el caso que se pueda vincular a la compañía. Las predicciones son visualizadas en la siguiente imagen.



predict
0.305416
-0.251466
-0.0512737
-0.625576
0.424024
-0.249989
0.382849
-0.508101
0.752639
0.0707768
0.952077
-1.53526
2.22273
0.255269
0.160389
1.64908
0.139708
1.0722
-0.193705
0.286525
2.99434
0.23947
1.71383
-0.419056
0.0918735
-0.464446
-0.101577
0.446696
0.562114

Fig. 32. Predicciones de las cotizaciones

XIII. CONCLUSIONES

- El problema de negocio se considera de alta complejidad en predicción para un entorno productivo pues la variable objetivo es un valor de ganancia o pérdida de dinero para la compañía de seguros. Debido al hecho de que la variable objetivo es una variable continua, el MAE o el error absoluto medio significa un valor en dinero, sin embargo, no es la única métrica adecuada para la evaluación del modelo de regresión, pues este concepto de ganancia o pérdida para la compañía puede ser subjetivo. Es decir, un cliente potencial puede ser una perdida pequeña de resultado técnico. Esta razón hizo que se considere como otra métrica de desempeño el MAPE o error absoluto porcentual medio, el cuál se establece en términos relativos, y da referencia del promedio de equivocación en la regresión.

- A pesar de que se esperaría que el MAPE sea bajo, no se debe descartar el escenario del problema negocio actual. Por un lado, el MAPE es un promedio o recuento estadístico de toda la muestra teniendo en cuenta los errores individuales, no obstante, se debe considerar otra métrica de desempeño que evalúe sobre la población total. Esta métrica consistió en la relación que existe entre la suma de todos los valores reales del resultado técnico y la suma de todos los valores predichos para determinar el grado de similitud entre ambas.

- De acuerdo con la métrica asociada a la evaluación del problema de negocio, se interpreta que el modelo basado en ensambles es capaz de reproducir la base de datos como un conjunto, sin embargo, resulta ser complicado predecir sobre cada cliente porque es posible que, para un solo cliente en la muestra total, haya sido afectado por alguna característica en particular.

- Se propone realizar como implementación de mejora al modelo de regresión, la transformación de variables continuas pertenecientes al resultado técnico en rangos. Para este procedimiento la compañía puede realizar el acompañamiento para definir los rangos entre ganancias o pérdidas de acuerdo con el conocimiento del panorama que tengan en el mercado.

XIV. BIBLIOGRAFÍA

[1] DANE, «DANE INFORMACIÓN PARA TODOS - Saber para decidir – Sistema nacional de información de demanda laboral,» [En línea]. Available: https://www.dane.gov.co/index.php/en/statistics-by-topic-1/education/informante#informacion-nacional. [Último acceso: 2018].

[2] T. Jo, "machine learning" Foundations: Supervised, Unsupervised, and Advanced Learning, Korea: Springer, 2021.

[3] M. J. RODRÍGUEZ JAUME y R. MORA CATALÁ, «"Analisis de regresión múltiple",» de *Estadística informática : casos y ejemplos con el SPSS*, Universidad de Alicante, ISBN 84-7908-638-6, pp. 3-17, sep-2001, pp. pp. 3-17.

[4] F. V. G. G. A. M. V. T. B. G. O. …. o. Pedregosa, «Scikit-learn: "machine learning" in Python,» 2011. [En línea]. Available: https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html.

[5] H2O AutoML, 17 Nov 2021. [En línea]. Available: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html.

[6] P. Pandey, «A Deep dive into H2O's AutoML,» *Towards Data Science*, Oct 14, 2019.

[7] D. Burrueco, «https://interactivechaos.com/es/python/function/labelencoder,» 14 04 2019. [En línea]. Available: https://interactivechaos.com/es/python/function/labelencoder.

[8] R. V.-B. M. T.-F. Mercedes Reguant-Álvarez, «La relación entre dos variables según la escala de,» *revista d'innovació i recerca en educació*, 2018.

[9] «6.4. Imputation of missing values — scikit-learn 1.0.1 documentation,» [En línea]. Available: https://scikit-learn.org/stable/modules/impute.html#iterative-imputer.