Imagen que contiene Texto

Descripción generada automáticamente

Modelo de marketing para entidad TELEFÓNICa CON METODOLOGÍA crisp-dm

Por:

Daniel Esteban González Zuluaga

Gabriel Fernando Forero Ortiz

Profesor a cargo:

Jorge Andrés Alvarado Valencia

Pontificia Universidad Javeriana

Maestría de Analítica para la Inteligencia de Negocios

Métodos y aplicaciones de analítica1

2024

**1. Business Understanding**

**1.1 Background**

En el análisis estratégico del entorno y de la compañía Claro, utilizando los marcos de referencia PESTLE y las Fuerzas de Porter, se pueden identificar los factores clave que influyen en el mercado de las telecomunicaciones. El análisis PESTLE considera los factores Políticos, Económicos, Sociales, Tecnológicos, Legales y Ambientales. En el contexto de Claro, los factores políticos incluyen regulaciones gubernamentales y políticas de competencia que afectan la industria de telecomunicaciones. Económicamente, la saturación del mercado y la alta inversión inicial requerida para operar son desafíos críticos. Socialmente, el uso masivo de teléfonos celulares y la demanda de servicios de alta calidad influyen en las estrategias de retención de clientes, donde se ve un aumento el uso de la telefonía móvil, por ejemplo, entre el 2020 al 2021 se observó un aumento de un 1.9% en conexiones telefónicas móviles, es decir un aproximado de 1.1 millones de dispositivos móviles nuevos conectados en el territorio nacional. Desde el punto de vista tecnológico, la innovación constante y la actualización de infraestructuras son necesarias para mantenerse competitivos. Legalmente, el cumplimiento de normativas y la protección de datos son esenciales, mientras que los factores ambientales se relacionan con la sostenibilidad y la gestión de residuos electrónicos.

Por otro lado, el análisis de las Fuerzas de Porter revela la intensidad de la competencia en el mercado de telecomunicaciones. La amenaza de nuevos entrantes es baja debido a las altas barreras de entrada, como los costos significativos de infraestructura. La rivalidad entre competidores es alta, dado que hay un número limitado de operadores que compiten ferozmente por una base de consumidores ya saturada. El poder de negociación de los proveedores es moderado, mientras que el poder de los compradores es alto, ya que los clientes tienen múltiples opciones y pueden cambiar de operador fácilmente. La amenaza de productos o servicios sustitutos es moderada, considerando alternativas como la telefonía por internet.

Para combinar el análisis interno y externo, el marco SWOT permite evaluar las fortalezas, debilidades, oportunidades y amenazas de Claro. Las fortalezas incluyen una amplia infraestructura y una fuerte presencia de marca en el mercado. Las debilidades pueden ser las posibles altas tasas de deserción. Las oportunidades se presentan en la adopción de nuevas tecnologías y la expansión de servicios de valor añadido. Las amenazas incluyen la intensa competencia y las posibles regulaciones gubernamentales.

Frente a estos desafíos y oportunidades, la estrategia de Claro debe enfocarse en predecir y reducir la deserción de clientes. Utilizando técnicas analíticas avanzadas, como el modelado predictivo y el análisis del ciclo de vida del cliente (CLTV), Claro puede identificar patrones de comportamiento que indiquen la propensión a desertar. La implementación de estrategias de retención costo-efectivas, basadas en estos análisis, permitirá a Claro no solo mantener a sus clientes actuales, sino también optimizar el retorno a largo plazo, fortaleciendo su posición competitiva en el mercado, en las siguientes secciones se trabajará más a fondo el tema de las incidencias económicas que pueden traer las deserciones de los clientes.

**1.2 Business goal**

1. Optimizar la retención de los clientes según la información histórica identificando anticipadamente la población objetivo de retención y generando el plan de acción para minimizar su impacto en el negocio.

**1.3 Business success criteria**

1. **Tasa de retención de clientes:** Este KPI mide el porcentaje de clientes que continúan utilizando los productos o servicios de la empresa durante un período específico. Una tasa de retención más alta indica una mayor fidelidad de los clientes y una mejor retención a lo largo del tiempo.
2. **Valor de vida del cliente (CLV):** Este indicador representa el valor monetario total que un cliente genera para la empresa durante todo el tiempo que permanece como cliente. Se calcula teniendo en cuenta los ingresos generados por el cliente, descontando los costos asociados a su adquisición, servicio y retención.

**1.4 Data mining goal**

1. **Comparar modelos de predicción:** Realizar una comparativa entre una regresión y XGBoost con el objetivo de obtener un modelo de predicción preciso para prever la tasa de deserción.
2. **Análisis del comportamiento de deserción:** Identificar el porcentaje de usuarios potenciales para retención basándose en las predicciones. Esto permitirá una planificación financiera más sólida para la empresa, considerando la posibilidad de invertir en la adquisición de nuevos clientes.

**1.5 Data mining success criteria**  
  
Comparar modelos de predicción:

1. Obtener un AUC de mínimo un 70% para obtener un modelo de predicción sobresaliente, ya que este representa la capacidad del modelo para distinguir entre las clases positivas y negativas.

Análisis del comportamiento de deserción:

1. Encontrar relaciones entre variables para poder predecir las personas que se retirarán de la empresa. En este punto es importante tener en cuenta las posibles transformaciones e interacciones, que nos brinden la mayor cantidad de información para los modelos.

**2. Data Understanding**

**2.1 Describe data**

En la base de datos se identificaron 10,243 clientes, con variables categóricas y numéricas. De estos, 8,243 clientes se asignaron al conjunto de entrenamiento y 2,000 al conjunto de prueba.

**Identificadores de cliente**: Hay 1 variable que ayudan a identificar el id del cliente, esta variable no se tendrá en cuenta para el modelo.

**Variables tipo fecha**: Dentro de la base de datos encontramos 2 variables tipo fecha “d/m/y”, las cuales son “Fecha de nacimiento” y “Fecha inicio”, la variable fecha de inicio indica su primer plan pospago con la compañía.

**Variables categóricas**: Existe una variable categórica llamada 'tipo cliente' con tres categorías de clientes. La descripción detallada para las categorías son 1: Hombre, 2. Mujer, 3: Cliente empresarial.

**Variables binarias:** En la base de datos se encuentran 2 variables binarias llamadas “Factura online” y “Plan de datos”, la factura online indica si el cliente recibe la factura solamente por correo electrónico y el plan de datos indica si el cliente tiene un plan premium.

**Variables numéricas:** La base de datos contiene cuatro variables con valores numéricos: 'antigüedad' y 'mora' en días, 'facturación' en pesos, y 'minutos', la variable antigüedad indica los meses de antigüedad del equipo y los minutos indican la cantidad minutos consumidos en el mes.

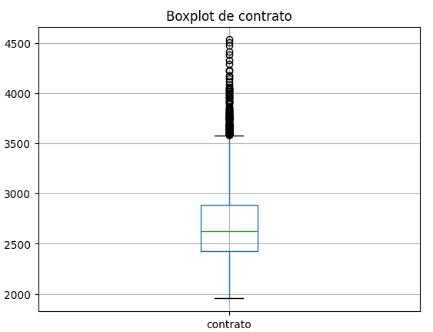
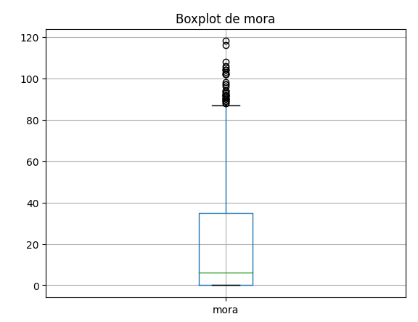
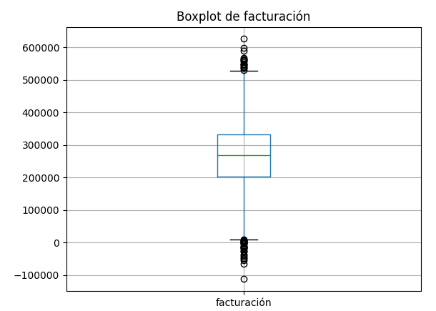
**Variable a predecir:** Por último, la variable 'resultado' es binaria y es la que vamos a predecir. Esta variable toma el valor de 1 si el cliente ha desertado de la compañía y 0 si aún sigue utilizando los servicios de la compañía.

|  |  |
| --- | --- |
| Nombre de columnas | Tipos de variables |
| id | Variable identificadora |
| Fecha de nacimiento y Fecha inicio | Variable tipo fecha |
| Tipo cliente | Variable categórica |
| Factura online y Plan de datos | Variable binaria |
| Antigüedad, Mora, Facturación y minutos | Variable numérica |
| Resultado | Variable binaria (A predecir) |

**2.2 Explore data**

En este análisis, se incluyen tanto análisis descriptivos univariados como multivariados relevantes para abordar las tres preguntas fundamentales y los dos objetivos del análisis descriptivo. Estos análisis se realizan atendiendo a las características clave del proceso: precisión y confianza. El enfoque principal es realizar una regresión logística para predecir la deserción de clientes. Es importante especificar que los análisis se han llevado a cabo sobre la base de datos de entrenamiento, para garantizar la validez y confiabilidad de los resultados. Aunque no se incluirán todos los gráficos o tablas producidos, se presentarán aquellos más significativos para ofrecer una visión clara y concisa de los hallazgos.

Para empezar, se realizó un análisis univariado de las variables utilizando gráficos Boxplot para mostrar su comportamiento. Se observó que algunas variables presentan valores atípicos, lo que podría afectar el resultado de nuestro modelo. Para identificar estos atípicos, se presenta en la imagen 1 el comportamiento de las variables más afectadas por ellos.



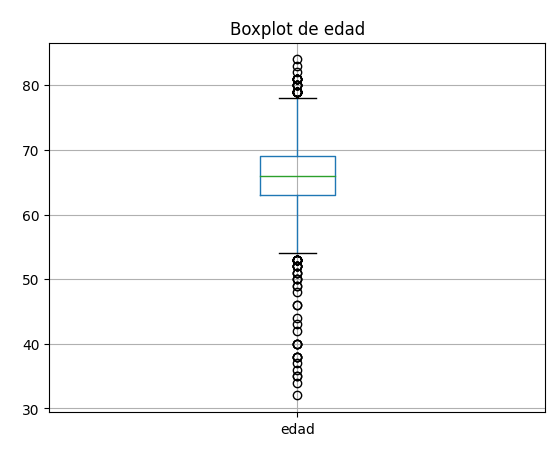
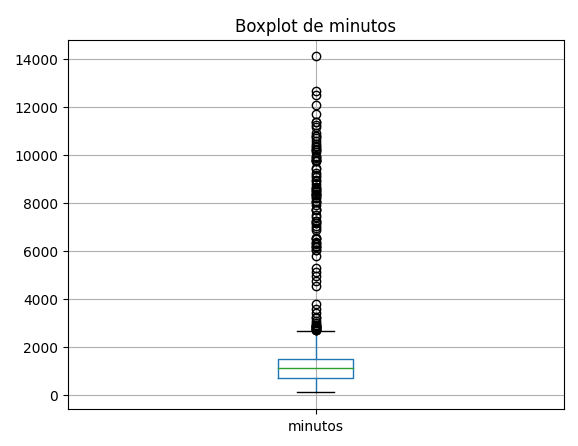
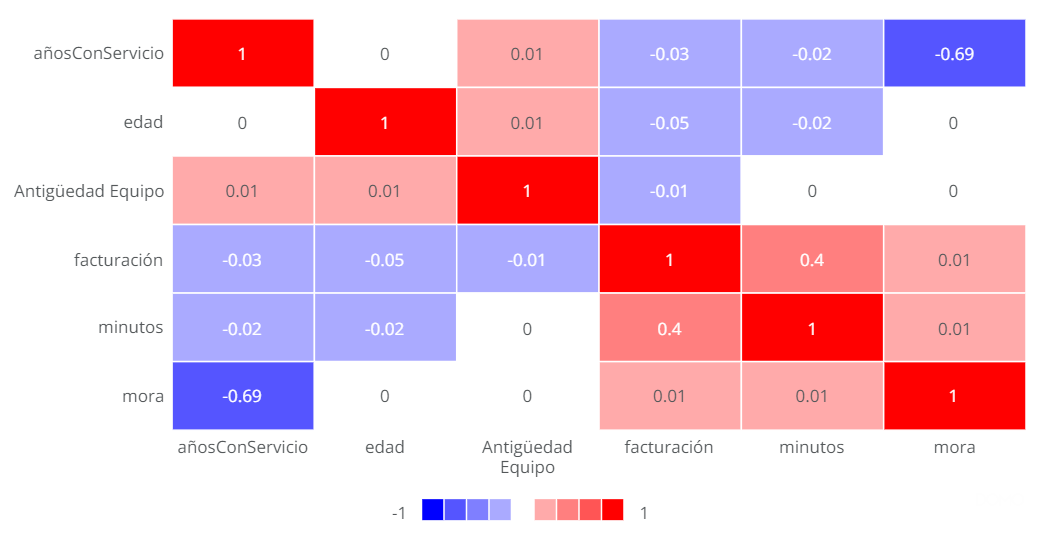


Gráfico 1. Comportamiento de las variables con atípicos.

Por último, realizamos un análisis multivariado para observar la correlación entre todas las variables numéricas. Esto es crucial para evitar que nuestro modelo se vea afectado por correlaciones muy altas, ya que estas pueden alterar de manera abrupta algunos coeficientes de las variables. En la gráfica 2 se presenta el gráfico de correlaciones entre las variables. Es importante mencionar que se realizaron modificaciones en las variables "edad" y "añosConServicio" para convertirlas en numéricas; estas transformaciones se explicarán posteriormente en la sección de transformación de variables.

Gráfico 2. Correlación entre las variables numéricas.

Se puede observar en esta matriz de correlación que no hay correlaciones tan fuertes. Aun así, se pueden percibir y reafirmar algunas ideas aplicables al modelo de negocio. Por ejemplo, la correlación entre mora y años con servicio es de –0.69, lo que indica que las personas con más años de servicio tienden a no estar mora o por lo menos que la variable mora sea alta. Esto sugiere que, en el negocio, las personas nuevas son más propensas a estar en mora. Otra correlación, aunque no tan fuerte pero aún positiva, es de 0.4 entre minutos y factura. Esto implica que, a mayor uso de minutos, la facturación tiende a ser más alta.

**2.3 Verify data quality**

En este análisis, es fundamental identificar y reportar la presencia de datos atípicos y datos perdidos en las variables del modelo de telefonía mencionado anteriormente. Los datos atípicos, tanto posibles como imposibles, pueden surgir debido a errores de entrada, mediciones incorrectas o variaciones extremas en el comportamiento de los clientes. Del mismo modo, los datos perdidos pueden deberse a problemas en la recopilación de información o a respuestas incompletas por parte de los usuarios. Proporcionar explicaciones potenciales sobre la aparición de estos datos es esencial para entender su impacto en el modelo y tomar decisiones informadas sobre cómo manejarlos para mejorar la precisión y confiabilidad de las predicciones de deserción de clientes, a continuación, nombraremos 2 inconsistencias que se encuentran en la base de datos.

**Facturacion negativa:** En las bases de datos de entrenamiento y prueba se encuentran valores negativos en la variable "facturación". En este caso, no está claro si estos valores representan una deuda hacia el cliente o si son el resultado de datos incorrectos.

**Fechas:** Las fechas más que ser un valor atípico o erróneo conduce a una perdida significativa de información, ya que no es cuantificable y no se pueden introducir en los modelos de regresión.

**Valores atípicos:** Como bien se mencionó en la sección pasada se encontraron en las variables datos atípicos por lo que es muy importante tenerlos en cuenta para que no afecte nuestro modelo.

**3. Data preparation**

Se utilizará toda la información de la base de datos de facturación de los últimos 6 meses de los clientes que Live Telecomunicaciones tiene en su registro, para esto se usaron diferentes criterios para seleccionar las variables.

**3.1 Clean data**

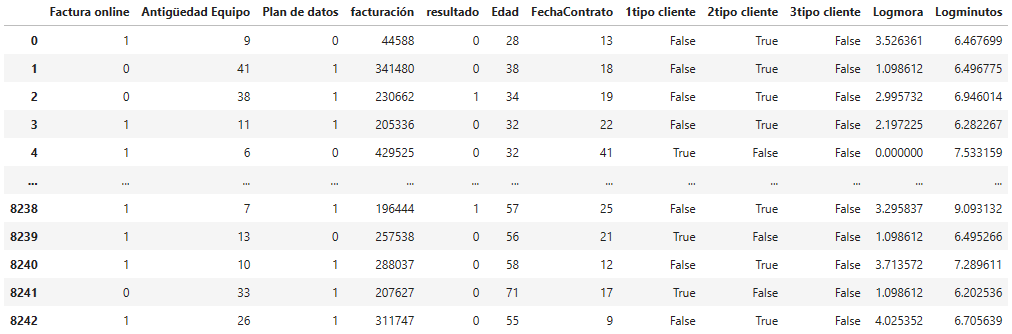
Los principales cambios de formato que se realizó a la base de datos fue cambiar a variable Dummy la columna tipo cliente y de las variables fecha de nacimiento y fecha de contrato para calcular la antigüedad en a la empresa (meses) y la edad (años). Una vez hecho estos cambios se eliminaron las columnas con fechas y el id

**3.2 Construct data**

Para la construcción del modelo el primer paso fue estandarizar las variables cuantitativas que no presentaban una concentración en los valores cercanos a 0. Las variables que no se vieron afectadas fueron minutos y mora, esto se debía a que su comportamiento era más parecido a la de una exponencial que a la de una normal. Por esta razón el tratamiento hecho en estas dos variables fue aplicarles un logaritmo natural (n+1). A pesar de que los minutos tienen valores demasiado altos que afectan la media no se hizo ningún tratamiento puesto que algunos planes postpago son de minutos ilimitado asi que no pueden considerarse como atípicos

**3.3 Dataset description**

A continuación, se describe el dataset de entrenamiento limpio con el que se entrenó el modelo de regresión logistica y XGBoost



|  |  |
| --- | --- |
| Nombre de columnas | Descripción |
| Factura Online | Binario que indica si el cliente recibe únicamente su factura online. |
| Antigüedad Equipo | Meses de antigüedad del equipo. |
| Plan de datos | Binario. 1 = el cliente tiene plan premium de datos. |
| Facturación | Total, facturación de los últimos seis meses (suma) |
| Resultado | 0 = No hubo deserción; 1 = Hubo deserción (churn) |
| Edad | Edad del cliente en a ñ os |
| Fecha contrato | Meses desde que inicio un contrato con la compañía |
| 1 tipo cliente | Si es cliente tipo 1 |
| 2 tipo cliente | Si es cliente tipo 2 |
| 3 tipo cliente | Si es cliente tipo 3 |
| Log mora | Logaritmo de los Días de mora acumulados |
| Log minutos | Logaritmo del total minutos consumidos en los últimos seis meses |

**4. Modeling**

**4.1 Select modeling techniques**

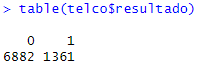
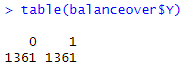
Al abordar el problema de la predicción de una variable binaria, es crucial seleccionar técnicas de modelado adecuadas que no solo proporcionen precisión en las predicciones, sino que también se ajusten a los supuestos y características de los datos. Dos técnicas que encontramos adecuadas para este propósito son la regresión logística y el clasificador XGBoost. A continuación, presentaremos una explicación de cada una de estas técnicas.

La regresión logística es un método estadístico utilizado para modelar una variable dependiente binaria. Utiliza la función logística para estimar las probabilidades de las clases y se basa en la relación lineal entre las variables independientes y la logit de la variable dependiente. Es especialmente útil porque su salida puede interpretarse como probabilidades, y sus coeficientes indican la dirección y magnitud del efecto de cada variable independiente sobre la variable dependiente. Por otro lado, XGBoost (Extreme Gradient Boosting) es una técnica de aprendizaje automático que utiliza árboles de decisión en un marco de boosting. A diferencia de la regresión logística, XGBoost no asume una relación lineal entre las variables independientes y la dependiente. En su lugar, construye múltiples árboles de decisión secuenciales, donde cada árbol intenta corregir los errores de predicción de los árboles anteriores. XGBoost es conocido por su capacidad para manejar datos con relaciones no lineales complejas y su eficacia en la reducción del overfitting a través de técnicas como la regularización.

Comparando ambas técnicas, la regresión logística es más fácil de interpretar y adecuada cuando se asume una relación lineal entre las variables, mientras que XGBoost es más potente para capturar relaciones no lineales y complejas en los datos. Sin embargo, XGBoost puede ser computacionalmente intensivo y difícil de interpretar debido a su naturaleza de conjunto de modelos. Ambas técnicas son buenas opciones para predecir una variable binaria porque la regresión logística proporciona una interpretación clara de las probabilidades y los efectos de las variables independientes, mientras que XGBoost ofrece una alta precisión en la predicción mediante el manejo de interacciones no lineales y complejas. Al elegir entre estas técnicas, consideramos que podemos tener un mejor rendimiento con XGBoost por lo que haremos una comparación de estos dos modelos en la sección 5.

**4.2 Generate test design**

Para ambos modelos, se utilizó la técnica de undersampling para abordar el problema del desbalanceo de clases en los datos. El undersampling reduce el número de muestras de la clase mayoritaria para equilibrar la proporción entre clases, lo que ayuda a los modelos a aprender de manera más efectiva y evita que la clase mayoritaria domine las predicciones. Este enfoque es crucial para mejorar la capacidad predictiva de los modelos, a continuación, se pueden observar cómo cambió la base de datos a partir del undersampling.

Para implementar adecuadamente las técnicas de regresión logística y XGBoost, se implementaron diferentes estrategias. En el caso de la regresión logística, se utilizará el método stepwise para la selección de variables. Este enfoque iterativo agrega o elimina variables del modelo basado en el criterio AIC, asegurando que solo las variables más relevantes y significativas se incluyan en el modelo final. Esto ayuda a controlar el sobreajuste al evitar la inclusión de variables irrelevantes que podrían ajustar demasiado el modelo a los datos de entrenamiento.

Para el XGBoost, se utilizarán todas las variables disponibles junto con sus respectivas transformaciones. La preparación de la base de entrenamiento incluirá la división de los datos en conjuntos de entrenamiento y prueba. El diseño de pruebas incluirá la evaluación del rendimiento de ambos modelos mediante métricas como la precisión, el recall, la precisión. Estas métricas proporcionarán una visión integral de la capacidad predictiva de cada modelo. Además, se analizará la importancia de las variables en el modelo XGBoost para entender mejor qué características contribuyen más a las predicciones. Este enfoque integral asegura que los modelos sean robustos, interpretables y capaces de generalizar bien a nuevos datos.

**4.3 Build model**

El punto de corte es un aspecto crucial en la predicción de una variable binaria y se considera un hiperparámetro, ya que determina el valor a partir del cual se clasifica una observación en una de las dos categorías posibles. En la regresión logística y el modelo XGBoost, las predicciones inicialmente se generan como probabilidades. El punto de corte comúnmente utilizado es 0.5, lo que significa que cualquier observación con una probabilidad predicha mayor o igual a 0.5 se clasifica en la clase positiva, mientras que las observaciones con probabilidades menores se clasifican en la clase negativa. Sin embargo, este punto de corte puede ajustarse para obtener mejores resultados. En el caso de la regresión logística, se ajustó a 0.59, mientras que para XGBoost se dejó el valor predeterminado de 0.5.

**5. Evaluation**

**5.1 Assess model**

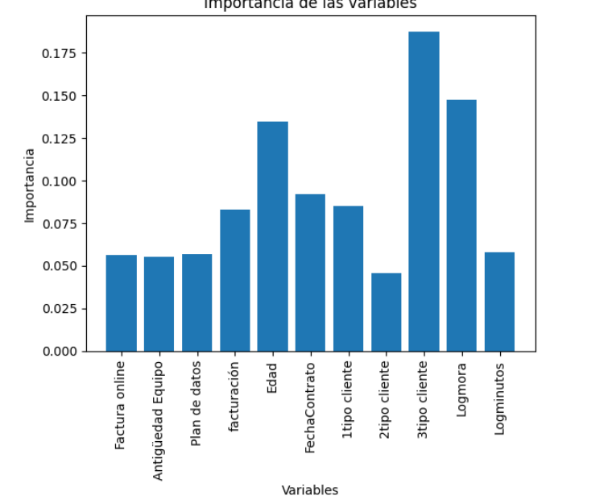
Se compararon los resultados del XGBoost y de la regresión logística para esto es importante tener en cuenta que los modelos se enteraron con los mismos datos para así poder comparar sus resultados

XGBOOST

|  |  |  |  |
| --- | --- | --- | --- |
| XGBOOST | Predicciones | | |
| Observados |  | 0 | 1 |
| 0 | 1344 | 33 |
| 1 | 149 | 123 |

|  |  |
| --- | --- |
| Precision | 0,788462 |
| Exahustividad/recall | 0,452206 |
| Falso Positivo | 0,023965 |
| Falso Negativo | 0,547794 |

Es decir que el XGBOOST es capaz de identificar el 45% de los casos correctamente con una precisión del 78.84%, Por otro lado, el acuracy del modelo es de **88.96%** es decir que predijo correctamente en más de un 88% los casos de abandono y finalmente tenemos un AUC de 0.7782

A continuación, se muestra la importancia de las variables en el modelo al momento de predecir, siendo el cliente tipo 3 que es el empresarial el que mejor ayuda en la predicción seguido por logmora y edad.

Regresión Logística

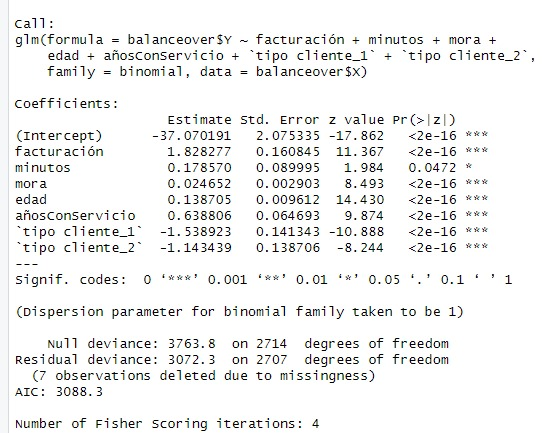
|  |  |  |  |
| --- | --- | --- | --- |
| Regresion Logistica | Predicciones | | |
| Observados |  | 0 | 1 |
| 0 | 1119 | 109 |
| 1 | 246 | 170 |

|  |  |
| --- | --- |
| Precision | 0,609319 |
| Exahustividad/recall | 0,408654 |
| Falso Positivo | 0,088762 |
| Falso Negativo | 0,591346 |

Es decir que la regresion es capaz de identificar el 40% de los casos correctamente con una precisión del 60.09%, Por otro lado, el acuracy del modelo es de 78**.10%** es decir que predijo correctamente en más de un 78% los casos de abandono y finalmente tenemos un AUC de 0.7948

Para el caso de la regresión se usó el método de feature seleccion step wise dejandonos los siguientes resultados en las variables

Como se puede observar a diferencia del XGBoost la regresion no considero la variable de plan de datos como significativa. Y su variable más importante es la edad, seguida por facturación y mora



Finalmente concluimos que la regresión tiene 10 puntos porcentuales menos de accuracy que el XGBOOST sin embargo tiene un mejor AUC!! Por temas de negocio valoramos más el accuracy ya que este es el que más impacto tienen en el negocio. De igual manera cumplimos con el objetivo de minería ya que se logró para cada modelo un AUC superior al 70%

**5.2 Produce final report**

Dado que el XGBOOST tiene un mejor acuracy el reporte final se hará con los resultados de ese modelo

En el caso de los verdaderos positivos, podemos determinar que la ganancia de haber acertado en que una persona se iba a irse corresponde al dinero invertido en el intento de retención. Es importante considerar que solo se toma una fracción de estas ganancias, ya que hay un costo de adquisición de clientes. En el caso de Claro la probabilidad de reemplazo es del 50%, y con esta probabilidad se calculó un valor esperado de las ganancias que dejará el nuevo cliente

Por otro lado, si nuestra predicción falla, las pérdidas serán el costo de retención invertido en esta persona. Estimamos que la compañía estaba dispuesta a perder el 25% de las ganancias producidas por este cliente para crear una estrategia atractiva de retención. Es importante destacar que las ganancias de la compañía por cada pago se estimaron en un 15%, pues este es el estimado de utilidad por cliente para Claro Colombia en el ano 2023.

El CLTV (Customer Lifetime Value) de los clientes de la empresa es de 24 meses según los datos de claro Colombia, adicional la facturación mensual es de $31.627 COP y el costo de adquisición promedio de la industria colombiana de telefonía es de $23.768 pesos según el último estudio de mercado, adicionalmente claro tiene una tasa de cancelación (churn) del 6% anual

La estrategia será enfocada a la retención, para esto se propone crear un modelo de clustering buscando crear perfiles de clientes y para cada perfil de diseñar un modelo de retención de la primera compra y los costos aproximados en el mercado colombiano es

Desarrollo (3 a 6 meses):

* Personal Técnico (mensual):
  + Científico de Datos: $6,000,000 COP
  + Ingeniero de Datos: $5,250,000 COP
  + Desarrollador Backend: $5,250,000 COP
  + DevOps: $4,500,000 COP
  + Total Mensual: $21,000,000 COP
  + Total 6 Meses: $126,000,000 COP

Equipos de Cómputo:

* Workstations (3 unidades): $24,000,000 COP
* Servidor y Almacenamiento: $40,000,000 COP
* Total Equipos: $64,000,000 COP

Software y Herramientas (anual prorrateado):

* Licencias: $3,000,000 COP

Implementación (1 a 2 meses adicionales):

* Salarios de Implementación (2 meses): $42,000,000 COP

Infraestructura en la Nube (opcional, 6 meses):

* Costo Nube: $3,000,000 COP mensuales
* Total Nube 6 meses: $18,000,000 COP

Total Aproximado del Proyecto

* Total Desarrollo (6 meses): $126,000,000 COP
* Equipos de Cómputo: $64,000,000 COP
* Software (6 meses): $18,000,000 COP
* Implementación (2 meses): $45,000,000 COP
* Infraestructura en la Nube (opcional, 6 meses): $15,000,000 COP

Total, Proyecto (con nube): $268,000,000 COP adicional los costos de mercadeo, distribución y logística es de $345.000.000 y el paquete de incentivos y promociones para que no abandonen la compañía es de $950.000.000

El total de clientes de Claro en el 2023 es de 39,2 Millones donde 18,3 Millones es de telefonía móvil, por lo que el clustering se hará sobre los 18 Millones sin embargo al ser tan complejo computacionalmente decidio por hacerlo con 1.000.000 de clientes escogidos de manera aleatoria para luego implementarlo sobre la base total. El cálculo de ganancias con estos números por lo que aproximadamente cancelan 60.000(Multiplicación de % de churn con todos los clientes) Ante esto el costo del proyecto por cliente es de $10.130 COP mensual por lo que el costo de fracaso es $10.130 y el costo de ganancia son de $21.514COP ($31.527 COP - $10.130 COP)

La ganancia neta seria de = 1Millon\* Exhaustividad\*(Ingreso de excito – Costo fracaso / Precisión)

=1.000.000\*0,4\*(21541-10130/0,61)

= 1.974.124.590COP en ganancias con esta estrategia lo cual es muy apetitoso para claro

Para complementar un poco más la estrategia, el costo de marketing incluye promociones por mensaje de texto adicional a llamadas a los clientes preferenciales y con mayor facturación para ofrecerle descuentos durante los siguientes 3 primeros meses por la mitad del valor de su factura e incluso llegar a acuerdos comerciales con aliados estratégicos como Apple y Samsung para que los clientes puedan adquirir los teléfonos de última tecnología, con esto se espera mejorar la tasa de retencion en 3 puntos porcentuales y a su vez aumentar el tiempo de vida de un cliente en Claro

**Bibliografía 1.** Kotler-Keller, (2015) Marketing Management.. Pearson.