Universitat Oberta de Catalunya

uoc.edu

Predicción de potenciales compradores en un ecommerce

Autor: José A. Rodríguez Guzmán
Máster Universitario en Ciencia de Datos
Data Analysis y Big Data

Tutor: Santiago Rojo Muñoz
Profesor: Albert Solé Ribalta

2 de junio del 2021

Créditos/Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Predicción de potenciales compradores en un ecommerce*
Nombre del autor:	*José A. Rodríguez Guzmán*
Nombre del consultor/a:	*Santiago Rojo Muñoz*
Nombre del PRA:	*Albert Soler Ribalta*
Fecha de entrega (mm/aaaa):	06/2021
Titulación:	*Plan de estudios del estudiante*
Área del Trabajo Final:	*Máster Universitario en Ciencias de Datos*
Idioma del trabajo:	*Español*
Palabras clave	*Ecommerce, potenciales compradores, machine learning*

Resumen del Trabajo.

Este proyecto se enfocó en la predicción de potenciales compradores en un ecommerce con sede en Colombia. Es decir, con base en los datos que se han almacenado de los visitantes de una web de venta online, establecer su nivel potencial de compra. Inicialmente se contó con dos bases de datos, una con características de los usuarios y otra con los consumos realizados por ellos. Este segundo archivo, fue transformado en otro que resumiera para cada usuario sus consumos y así, tener en un repositorio final toda la información junta. Uno de los retos que se presentaron fue trabajar con datos desbalanceados, es decir, que menos del 1% de los usuarios eran clientes versus un 99% de no clientes, es decir, se contaba con poca información sobre los usuarios convertidos, lo cual era un problema para poder entrenar los modelos. El segundo desafío enfrentado fue la gran cantidad de nulos que tenían muchas de las variables, por lo que se tomó la decisión de crear dos grupos, uno donde se tuviera la información completa y el otro con los que no se tuviera todo. En ambos subconjuntos de datos se utilizaron técnicas estadísticas como el análisis de multicolinearidad y pruebas de chi-cuadrado, así como el algoritmo Boruta para la elección de las variables que finalmente se utilizaron en los modelos. Al igual que las investigaciones realizadas sobre proyectos similares, se emplearon los algoritmos de aprendizaje automático: árboles de decisiones, bosques aleatorios y regresión logística. El modelo de árboles de decisiones fue el que mostró los mejores resultados en ambos escenarios, logrando valores de AUC de 0.999, una precisión global de 0.997 y una especificidad del 100%. La única diferencia en ambos fue la proporción de usuarios que fueron predichos como clientes correctamente, donde el grupo 1 obtuvo un 81,35%, mientras que el del grupo 2 fue del 65%. Finalmente, se indica que la variable del indicador de que un usuario fue al botón de pago es la característica de mayor peso para la predicción de los potenciales compradores el ecommerce.

Abstract

This project was focused on the prediction of potential buyers in an ecommerce based in Colombia. So based on the data that has been stored from the visitors of an online sales website, establish their potential level of purchase. Initially, there were two databases, one with user data and the other with the consumption made by them. This second file was transformed into another that would summarize their consumption for each user and thus have all the information together in a final repository. One of the challenges was working with unbalanced data, it means that less than 1% of users were customers versus 99% of non-customers. The information on converted users was little, which made it difficult to train the models. The second challenge faced was the large number of nulls that many of the variables had. That is why the decision was made to create two groups, one with complete information and the other one with less features. In both subsets of data, statistical techniques such as multicollinearity analysis and chi-square tests were used, as well as the Boruta algorithm for the choice of variables that were finally used in the models. Like the research carried out on similar projects, machine learning algorithms were used: decision trees, random forests, and logistic regression. The decision tree model was the one that showed the best results in both scenarios, achieving AUC values of 0.999, an overall precision of 0.997 and a specificity of 100%. The only difference in both was the proportion of users who were predicted as clients correctly, where group 1 obtained 81.35%, while that of group 2 was 65%. Finally, the most important characteristic for the prediction of potential e-commerce buyers was the indicator of users who went to the payment button.

Dedicatoria

Este trabajo final del máster ha puesto ha prueba no sólo mis conocimientos técnicos y todo lo aprendido durante este máster, sino también, la lucha, entrega, paciencia y dedicación.

Han sido arduas horas las invertidas en el proyecto para buscar realizar un trabajo de gran calidad, no obstante, eso ha hecho que deba de sacrificar muchas otras cosas de la vida cotidiana y que, al no poder realizarlas, llegan a afectar de alguna manera.

Por esta razón, agradezco enormemente todo el apoyo que me brindó Priscilla Blanco, quien siempre tuvo palabras de apoyo, quien me vio en los momentos de mayor desesperación y estuvo ahí para abrazarme y animarme.

Gracias Pri, sin duda alguna vos has sido un pilar importante en el éxito de este TFM.

Así mismo, agradezco muchísimo todo el apoyo continuo e incondicional de mi madre y hermana, quienes siempre tuvieron las palabras adecuadas para motivarme y empujarme a dar un esfuerzo adicional.

Finalmente, pero no menos importante, gracias a Dios porque sin Él, nada de esto hubiese sido posible.

De corazón, gracias a todos.

Abstract

This project was focused on the prediction of potential buyers in an ecommerce based in Colombia. So based on the data that has been stored from the visitors of an online sales website, establish their potential level of purchase.

Initially, there were two databases, one with user data and the other with the consumption made by them. This second file was transformed into another that would summarize their consumption for each user and thus have all the information together in a final repository.

One of the challenges was working with unbalanced data, it means that less than 1% of users were customers versus 99% of non-customers. The information on converted users was little, which made it difficult to train the models.

The second challenge faced was the large number of nulls that many of the variables had. That is why the decision was made to create two groups, one with complete information and the other one with less features. In both subsets of data, statistical techniques such as multicollinearity analysis and chi-square tests were used, as well as the Boruta algorithm for the choice of variables that were finally used in the models.

Like the research carried out on similar projects, machine learning algorithms were used: decision trees, random forests, and logistic regression.

The decision tree model was the one that showed the best results in both scenarios, achieving AUC values of 0.999, an overall precision of 0.997 and a specificity of 100%. The only difference in both was the proportion of users who were predicted as clients correctly, where group 1 obtained 81.35%, while that of group 2 was 65%.

Finally, the most important characteristic for the prediction of potential e-commerce buyers was the indicator of users who went to the payment button.

Keywords

Ecommerce, Machine Learning, Data Science, Decision Trees, Unbalanced Data, Prediction, Potential Buyers, R

Resumen

Este proyecto se enfocó en la predicción de potenciales compradores en un ecommerce con sede en Colombia. Es decir, con base en los datos que se han almacenado de los visitantes de una web de venta online, establecer su nivel potencial de compra.

Inicialmente se contó con dos bases de datos, una con características de los usuarios y otra con los consumos realizados por ellos. Este segundo archivo, fue transformado en otro que resumiera para cada usuario sus consumos y así, tener en un repositorio final toda la información junta.

Uno de los retos que se presentaron fue trabajar con datos desbalanceados, es decir, que menos del 1% de los usuarios eran clientes versus un 99% de no clientes, es decir, se contaba con poca información sobre los usuarios convertidos, lo cual era un problema para poder entrenar los modelos.

El segundo desafío enfrentado fue la gran cantidad de nulos que tenían muchas de las variables, por lo que se tomó la decisión de crear dos grupos, uno donde se tuviera la información completa y el otro con los que no se tuviera todo. En ambos subconjuntos de datos se utilizaron técnicas estadísticas como el análisis de multicolinearidad y pruebas de chi-cuadrado, así como el algoritmo Boruta para la elección de las variables que finalmente se utilizaron en los modelos.

Al igual que las investigaciones realizadas sobre proyectos similares, se emplearon los algoritmos de aprendizaje automático: árboles de decisiones, bosques aleatorios y regresión logística.

El modelo de árboles de decisiones fue el que mostró los mejores resultados en ambos escenarios, logrando valores de AUC de 0.999, una precisión global de 0.997 y una especificidad del 100%. La única diferencia en ambos fue la proporción de usuarios que fueron predichos como clientes correctamente, donde el grupo 1 obtuvo un 81,35%, mientras que el del grupo 2 fue del 65%.

Finalmente, se indica que la variable del indicador de que un usuario fue al botón de pago es la característica de mayor peso para la predicción de los potenciales compradores el ecommerce.

Palabras clave

Comercio Electrónico, Aprendizaje Automático, Ciencia de Datos, Árboles de Decisiones, Datos Desbalanceados, Predicción, Compradores Poteciales, R.

Índice

Figuras y Tablas

Índice de figuras

Índice de tablas

1. Introducción

A continuación, se presentan las principales ideas que han incitado dicha investigación.

En primer lugar, el apartado 1.1 mostrará la contextualización, haciendo referencia a información relevante del comercio, como, por ejemplo: a qué se dedica, qué tipo de productos tiene u ofrece, cómo obtiene los datos, y una síntesis del flujo del negocio para así tener una mejor comprensión de este. Posteriormente la motivación de la realización de este irá a lo largo del apartado 1.2.

En segundo lugar, en el apartado 1.3 se presentarán los objetivos principales y parciales. Luego, el apartado 1.4 mostrará la metodología que se utilizará para desarrollar el proyecto, y finalmente, en el apartado 1.5 se describirá la planificación del proyecto.

1.1 Contexto y justificación

Los datos provienen de un ecommerce con sede en Colombia que se dedica a la venta de productos relacionados con la información de empresas del país: Informes Comerciales y módulos de información detallada sobre Datos Financieros, Prensa, Administradores, Incidencias, etc., Informes Sectoriales, Base de datos a medida, Productos de Marketing (mercadeo), Información de accionistas, Información de proveedores y clientes, etc.

Los usuarios (personas en su nombre o representando a una empresa u otra entidad), los cuales llamaremos potenciales compradores, llegan a la web del ecommerce por diferentes canales: directorios propios, webs de terceros, desde buscadores por labores de SEO y de SEM, etc.

Dichos usuarios tienen acceso a diferentes productos de menor contenido a cambio de registrarse en un formulario del sitio, en el que indican datos personales, email, teléfono, profesión, entre otros. De esta forma, el usuario pasa a ser “registrado”, y es lo que en el mundo del ecommerce se le denomina “lead”.

Tras el registro, se permite el acceso a información muy básica sobre las empresas buscadas (Ficha de empresa), y se otorga la posibilidad de consumir gratis 5 productos de información a los que llamamos perfiles de empresa. Dicha oferta caduca a los 30 días.

El Perfil de empresa es un producto con contenido básico y su objetivo es mostrar a los potenciales compradores el nivel del contenido de los productos que suministra el ecommerce.
Resumiendo, la dinámica es la siguiente:

- Un usuario busca una empresa en internet o en la web.
- Al seleccionar una de las listadas en la búsqueda, se presenta una Ficha (Identificación) y se crea un registro en el log. Se muestra los productos disponibles para consumir.

    - Productos Promocionales (Perfil de Empresa). Para usuarios Registrados, si el Usuario se registra.
    - Resto de Productos son de Pago. Es necesario estar registrado y pagar.
- El usuario consume un producto (registro en el log) o vuelve a buscar otra empresa.

Si el usuario Registrado está interesado en conocer más a fondo una empresa o un determinado producto de pago, se le ofrece la posibilidad de contratación:

- PPV: Compra puntual del producto
- Bonos: compra de un conjunto de unidades o una cantidad de productos a cambio de un pago anticipado.
- Suscripción: pagando una cantidad periódica permite el acceso y consumo de productos, limitado por el volumen de compra y por la fecha de caducidad de la suscripción

Cuando se produce una de estas contrataciones el usuario “Registrado”, pasa a ser “Cliente”.

El propietario del ecommerce tiene información del usuario de su plataforma de Google Analytics que utiliza para la captación en internet, uso de *cookies*, estrategias de SEO y SEM, acuerdos con portales.

Ahora desea dar un paso más y quiere conocer más sobre los usuarios registrados para determinar la probabilidad de conversión a cliente. Saber cuáles son potenciales compradores a partir de los datos de ese registro en sus sistemas, del hábito de consumo de productos, y del tipo de empresa buscada/consultada.

El conjunto de datos está muy desbalanceado por ello es muy importante determinar a qué usuarios registrados debe dirigir el ecommerce sus campañas de captación.

Por otra parte, pero relacionado con el mismo tema, desean saber que perfil tienen sus clientes y que variables son las mas influyentes/relevantes en la predicción de compra para modificar la estrategia de marketing y su posicionamiento en Internet.

1.2 Motivación

Debido a que siempre me he considerado una persona de retos, quise elegir un tema que me sacara de la zona de confort y me exigiera un poco más de lo “normal”.
Lo dicho anteriormente se debe a que, 1) los datos no se encuentran disponibles en la web o son de un concurso de Kaggle donde se puede tomar ventaja de las soluciones de otras personas o consejos y métodos que se encuentre en internet de cómo solucionar el problema, y 2) a que la base de datos es muy desbalanceada.

Ahora bien, lo interesante de este tema es que estamos en la era del internet, donde día a día millones de personas visitan páginas web, ya sea porque buscan algo en específico, o simplemente, porque fueron redirigidos a esa página por alguna razón. Pero, lo complicado de esto es identificar y predecir, cuales de todos esos usuarios

que visitaron una página web en específico, realmente es porque van a comprar algo. Es por ello, que el reto de predecir a un potencial comprador es alto.

Adicionalmente, el ecommerce ha explotado tanto en crecimiento como en popularidad en los últimos 10 años, incluso se dice que un tercio de las personas ahora realizan una compra en línea al menos una vez por semana.

Es por ello por lo que, con el crecimiento de esta industria, este proyecto me dará mucha experiencia y pericia de cómo desenvolverme en un futuro en un caso similar a este, ya sea para aplicar en un proyecto como consultor, o bien, dentro de una empresa.

1.3 Objetivos

En este apartado, se expondrá el objetivo principal y los objetivos secundarios de la investigación a realizar, recordando, que el comercio ha venido recolectando datos y desea conocer más sobre los usuarios.

1.3.1 Objetivo principal

El objetivo principal de este trabajo es:

- Creación de uno o varios modelos de clasificación de leads, donde se identifique una posible selección de potenciales compradores.

1.3.2 Objetivos Secundarios

Como objetivos derivados del anterior, se pretende también realizar y alcanzar:

- Realizar un análisis descriptivo de los datos existentes.
- Crear el repositorio final de datos con el que se trabajará.
- Identificar las variables que ayuden a predecir cuales usuarios serían potenciales compradores
- Comparar y elegir el o los mejores modelos que ayuden a alcanzar el objetivo principal

1.4 Metodología

Para la realización de este proyecto, se utilizará la metodología CRISP-DM, que significa por sus siglas en inglés, *cross-industry process for data mining*.

Esta metodología proporciona una descripción general del ciclo de vida de un proyecto de minería de datos. Contiene las fases de un proyecto, sus respectivas tareas y las relaciones entre estas tareas.
Es conocida por ser robusta y bien probada, así mismo, por su poderosa practicidad, su flexibilidad y su utilidad cuando se utiliza la analítica para resolver problemas comerciales.

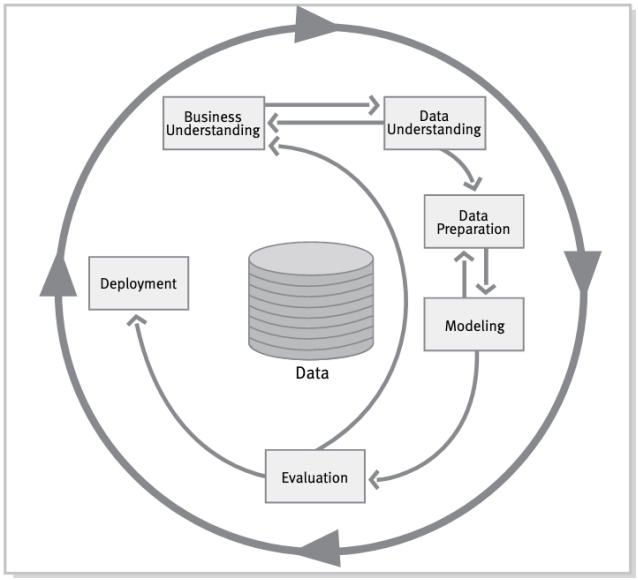


Figura 1: Fases del modelo de referencia CRISP-DM

El ciclo de vida de un proyecto de minería de datos consta de seis fases, que se muestran en la Figura 1. La secuencia de las fases no es rígida. Siempre es necesario moverse hacia adelante y hacia atrás entre las diferentes fases. El resultado de cada fase determina qué fase, o tarea particular de una fase, debe realizarse a continuación. Las flechas indican las dependencias más importantes y frecuentes entre fases.

El círculo exterior de la figura simboliza la naturaleza cíclica de la minería de datos en sí. La minería de datos no termina una vez que se implementa una solución. Las lecciones aprendidas durante el proceso y de la solución implementada pueden desencadenar nuevas preguntas comerciales, a menudo más centradas. Los procesos posteriores de minería de datos se beneficiarán de las experiencias de los anteriores. A continuación, se describirá brevemente cada fase:

- Entender el negocio: esta fase inicial se centra en comprender los objetivos y requisitos del proyecto desde una perspectiva comercial, luego convertir este conocimiento en una definición de problema de minería de datos y un plan preliminar diseñado para lograr los objetivos.
- Comprensión de los datos: esta comienza con la recopilación inicial de datos y continúa con actividades que le permiten familiarizarse con los datos, identificar problemas de calidad de los datos, descubrir los primeros conocimientos sobre los datos y / o detectar subconjuntos de interés para formar hipótesis sobre información oculta.
- Preparación de los datos: esta cubre todas las actividades necesarias para construir el conjunto de datos final (datos que se introducirán en las herramientas de modelado) a partir de los datos en bruto iniciales. Es probable que las tareas de preparación de datos se realicen varias veces y no en un orden prescrito. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de datos para las herramientas de modelado.
- Modelado: en esta fase, se seleccionan y aplican varias técnicas de modelado, y sus parámetros se calibran a valores óptimos. Normalmente, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas

tienen requisitos específicos sobre la forma de los datos. Por lo tanto, a menudo es necesario volver a la fase de preparación de datos.

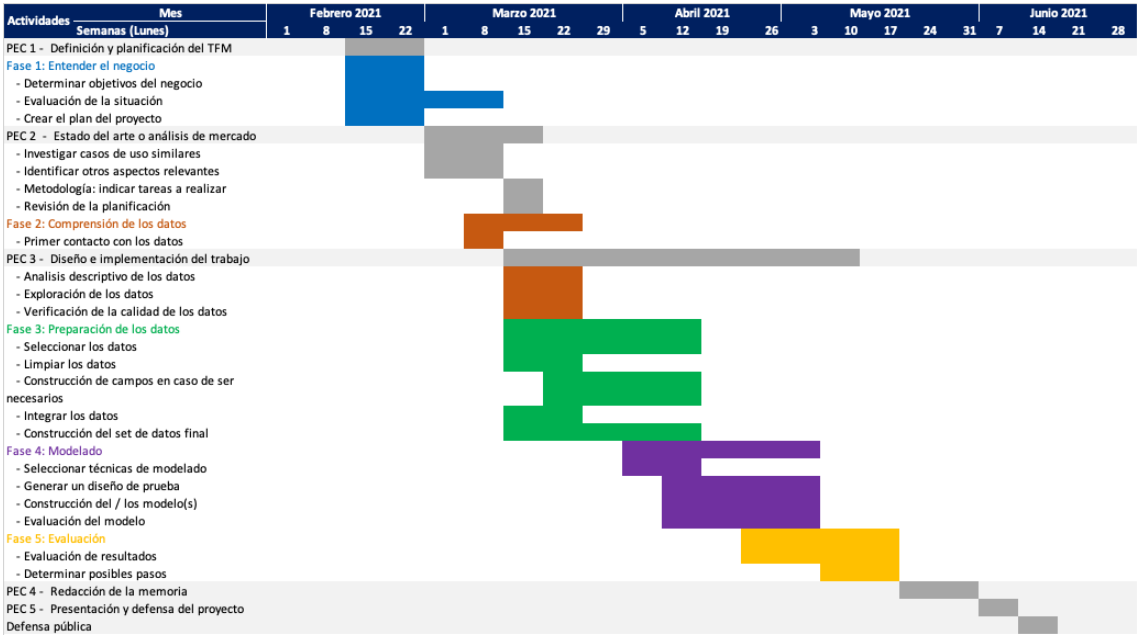
- Evaluación en esta etapa del proyecto, se ha construido un modelo (o modelos) que parece tener alta calidad desde la perspectiva del análisis de datos. Antes de proceder con la implementación final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, para asegurarse de que el modelo logre adecuadamente los objetivos comerciales. Un objetivo clave es determinar si existe algún tema comercial importante que no se haya considerado suficientemente. Al final de esta fase, se debe tomar una decisión sobre el uso de los resultados de la minería de datos.
- Implementación La creación del modelo generalmente no es el final del proyecto. Incluso si el propósito del modelo es aumentar el conocimiento de los datos, el conocimiento adquirido deberá organizarse y presentarse de manera que el cliente pueda utilizarlo. A menudo implica la aplicación de modelos "en vivo" dentro de los procesos de toma de decisiones de una organización, por ejemplo, la personalización en tiempo real de las páginas web o la puntuación repetida de bases de datos de marketing. Según los requisitos, la fase de implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minería de datos repetible en toda la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva a cabo los pasos de implementación. Sin embargo, incluso si el analista llevará a cabo el esfuerzo de implementación, es importante que el cliente comprenda de antemano qué acciones deben llevarse a cabo para hacer uso real de los modelos creados.

Debido a que este es un proyecto final de un máster y no se está trabajando directamente con la empresa de ecommerce, el paso de la implementación no se realizaría, no obstante, se mencionó arriba debido a que es parte de la metodología empleada.

Así mismo, al final del proyecto se podrían dar recomendaciones de como el o los modelos finales pueden ser implementados, sin embargo, esto se irá analizando durante la realización de este.

1.5 Plan de investigación del proyecto

Bajo la metodología indicada en el apartado anterior y siguiendo las fechas en las que se deben entregar las PEC durante la ejecución del proyecto, la planificación inicial sería la siguiente:



2. Estado del arte

En este capítulo, se abordan aspectos que contextualizan el proyecto. En primer lugar, a pesar de que no se conoce el nombre de la empresa que suministra los datos, se tiene conocimiento sobre el tipo de negocio de esta y algunos de sus competidores a nivel mundial, por lo que se expondrán algunos de ellos.

Seguidamente, se hará una revisión de literatura, para así ir comprendiendo cómo investigadores han abordado el tema, posteriormente se analizará cómo han buscado ellos hacer la predicción de potenciales compradores en un ecommerce por medio de técnicas de *machine learning*, y finalmente, se expondrán algunos de retos que se han enfrentado otras personas en el pasado y cómo los resolvieron.

2.1 Competencia

Anteriormente, en el capítulo del contexto y justificación, se mencionó que los datos son provenientes de un ecommerce con sede en Colombia que se dedica a la venta de productos relacionados con la información de empresas del país: Informes Comerciales y módulos de información detallada sobre Datos Financieros, Prensa, Administradores, Incidencias, etc., Informes Sectoriales, Base de datos a medida, Productos de Marketing (mercadeo), Información de accionistas, Información de proveedores y clientes, etc.

A pesar de que el tipo de producto que venden es muy específico y que la predicción de los potenciales compradores se realizará con datos que provienen de las herramientas que ellos utilizan, siempre es bueno hacer como un estudio de mercado y estudiar qué otras empresas ofrecen los mismos productos o productos similares, ya que un posible comprador de la empresa en estudio podría comprarles a ellos, o bien, a su competencia.

A continuación, se expondrán algunos de sus posibles competidores y los principales productos que venden.

- **Experian**
  Experian es una compañía tecnológica especializada en servicios de crédito, análisis, software y datos, con más de 20 años siendo líderes en servicios y soluciones de prevención del riesgo crediticio, prevención del fraude, estrategias de recobro y scoring. Opera en 45 países, dentro de los cuales está Colombia.

  Ellos dividen el negocio en tres: cliente, pequeña empresa y empresa; y ofrecen tanto soluciones (análisis avanzada y modelaje, decisiones de crédito, informes de perfil crediticio, soluciones de mercadeo, gestión de riesgos y fraude, soluciones de cobranza de deudas, etc.), como servicios especiales (servicios de información comercial, de asesoría, detección, de calidad de datos, etc.).

- **TransUnion**
  TransUnion es una empresa global de información y conocimientos. Ellos buscan tener una imagen precisa y completa de cada persona, y tratan de administrarla con cuidado para que cada consumidor esté representado de manera confiable y segura en el mercado. Sus oficinas centrales están ubicadas en Chicago, Illinois (Estados Unidos), y entre sus sedes regionales, tienen una ubicada en Colombia.

  Entre sus soluciones más destacadas, están los informes de crédito de los clientes, mercadeo digital, servicios de violación de datos, soluciones de fraude, etc. Y a nivel de productos para pequeñas empresas, se pueden destacar los siguientes productos: **CreditVision®,** que ofrece una visión más amplia y profunda del perfil crediticio de un consumidor; **Express Portfolio ReviewSM**, para medir las gestión de riesgos; **TLOxp®** donde tienen información sobre deudores y brindan el conocimiento, los datos y los análisis para tomar decisiones informadas y así tener a los clientes adecuados.

- **Equifax**
  Es una compañía global de datos, análisis y tecnología que crea soluciones y perspectivas innovadoras para ayudar a los clientes a impulsar el crecimiento y fomentar el progreso de las personas. Esta empresa tiene su sede central en Atlanta, y opera en 24 países. A pesar de que no tienen presencia en Colombia, si lo están en América del Sur en países como Argentina, Brasil, Chile, Ecuador, Paraguay, Perú y Uruguay.

  Equifax tiene su portafolio dividido en dos, empresas y personas. A nivel de empresas sus tres principales soluciones son: **Equifax Ignite,** que da acceso a múltiples fuentes de datos y conocimientos analíticos (*Marketing*, Riesgo, Inteligencia de Negocio y Cobranza). **Equifax InterConnect,** es un motor de decisión en la nube que permite evaluar de manera rápida y segura, las condiciones de crédito y riesgo de nuevas solicitudes y gestionar estrategias de riesgo de manera óptima. **CyberFinancial,** es un sistema de cobranza integral que busca automatizar todas las tareas de las personas involucradas en la cobranza de carteras de crédito, ayudando a las empresas a cobrar más en menos tiempo.

A nivel mundial existen otras compañías como Bisnode, Dun & Bradstreet o Crif, que a pesar de que venden bases de datos con información de todo tipo, incluida la crediticia y soluciones de mercadeo, de análisis datos, entre otros, no serán expuestos debido a que no tienen presencia en Colombia o países cercanos. Se mencionan debido a que pueden considerarse como competencia indirecta y capaz, que en un tiempo estas compañías extiendan sus horizontes y quieran penetrar otros negocios en Latinoamérica, o aún más específico, en Colombia.

2.2 Revisión de literatura

A pesar del rápido crecimiento de las ventas de comercio electrónico, las tasas de conversión en línea en todas las industrias suelen ser muy bajas, rara vez superan el 5% (eMarketer, 2014). La tasa de conversión se define como el porcentaje de visitas que generan compras. Las tasas de conversión bajas implican que la mayoría del tráfico de sitios web de comercio electrónico solo representa visitantes ocasionales en lugar de compradores serios.

Uno de los principales beneficios de hacer negocios en línea es que se pueden rastrear varios aspectos del comportamiento de los clientes con la ayuda de la tecnología moderna. Debido a la gran cantidad de datos de comportamiento disponibles en línea, una extensa literatura ha explorado la posibilidad de usar esos datos para comprender y predecir sus decisiones de conversión (da Silva, 2014; Fernandes, 2015).

En las últimas décadas, ha habido una serie de investigaciones que abordan la profundidad y la dinámica de búsqueda de los consumidores utilizando los datos del flujo de clicks. Por ejemplo, Johnson et al. (2004) utilizan los datos para caracterizar el comportamiento de búsqueda en tres niveles: profundidad, dinámica y actividad de búsqueda.

Hay diferentes opiniones sobre el impacto de los motores de búsqueda modernos en la profundidad de búsqueda de los consumidores. Peterson y Merino (2003) creen que la disponibilidad de herramientas de búsqueda aumentará la cantidad de información y, por lo tanto, aumentará la profundidad de búsqueda. Sin embargo, Holland y Mandry (2013) analizan una gran cantidad de datos del panel de Internet de varios sitios web de comercio electrónico y concluyen que la profundidad de búsqueda en todos los sectores es significativamente menor de lo esperado. En cuanto a la dinámica de búsqueda, la investigación empírica sobre las características temporales de la búsqueda es escasa, quizás debido a los datos de observación parcial de la búsqueda del consumidor.

Las excepciones notables son De Los Santos et al. (2015), quien explican que los consumidores vuelven a visitar los elementos que se buscaron anteriormente debido a los costos de aprendizaje o de búsqueda no estacionaria.

Este rico cuerpo de literatura profundiza en el comportamiento de búsqueda en comparación con aquellos que investigan los beneficios de la búsqueda en línea. Sin embargo, la profundidad y la dinámica son solo dos aspectos de alto nivel del comportamiento de búsqueda. En la comunidad de investigación se esperan más detalles de comportamiento a nivel de acción, como formulación de consultas,

escaneo y visualización de elementos, clics y tiempo de permanencia. Es posible que perder esas piezas no capte las señales potenciales para inferir el comportamiento de conversión.

En los últimos dos años, hay una serie de estudios en analítica empresarial que incorporaron modelos de aprendizaje automático para predecir las intenciones de compra de los clientes. Algunos ejemplos son la investigación de tesis de da Silva (2014) que construyó cuatro modelos de aprendizaje automático en un conjunto de datos de flujo de clics.

La investigación de tesis de Fernandes (2015) que construyó un modelo de secuencia para predecir la probabilidad de compra en tiempo real. Además, Vieira (2015) utilizó algoritmos de aprendizaje profundo para analizar el comportamiento de compra.

Finalmente, Varian (2014) dijo que “un gran conjunto de datos puede permitir relaciones más flexibles que los modelos lineales simples”. Por lo tanto, se recomiendan técnicas de aprendizaje automático para revelar e investigar relaciones complejas.

2.3 Prediciendo potenciales compradores en un ecommerce

Aplicar técnicas de *machine learning* no es solo tomar todos los modelos que existen y ver qué funciona mejor. Para este problema en particular, se necesitan utilizar modelos de clasificación y no de regresión, debido a que nuestra variable a predecir es de clasificación binaria, donde hay compradores y no compradores.

Estos modelos de predicción podrían ser divididos en dos categorías: modelos individuales y modelos de conjunto, según Huibing Zhang y Junchao Dong (2020).

Los modelos individuales serían, por ejemplo: la regresión logística (LR), la máquina de vectores de soporte (SVM), la red neuronal recurrente (RNN) y la percepción multicapa (MLP), mientras que los modelos de conjunto podrían ser: el bosque aleatorio (RF), el árbol de decisión de aumento de gradiente (GBDT) y XGBoost.

De acuerdo con D. Xu et al. (2018) y Guimei Liu et al. (2016), los resultados demuestran que los modelos de predicción de aprendizaje por conjuntos son superiores a los modelos de predicción individuales en términos de precisión y solidez; debido a que los modelos de conjuntos integran múltiples modelos de predicción individuales.

Dicho lo anterior, se expondrán algunos análisis que han hecho otros investigadores, así como sus resultados en términos generales. No obstante, es importante mencionar, que estos estudios han sido realizados en otros tipos de empresas que venden productos distintos a los estudiados, sin embargo, son una ayuda para ver qué métodos se han utilizado y cómo los han implementado.

- **Análisis 1**
  El primer análisis que se mostrará es el realizado por Martin Beck, quien publicó un artículo llamado "¿Se puede predecir si un cliente realizará una compra en un sitio web?", el cual, se basa en los resultados del proyecto final de Business Data Science.

  Sus compañeros y él decidieron utilizar tres modelos diferentes que fueron:

  1) SGDClassifier, donde el estimador utiliza un modelo lineal con aprendizaje de descenso de gradiente estocástico (SGD). En este, cada muestra estima el gradiente de pérdida y el modelo se actualiza en función de la tasa de aprendizaje. Este modelo simple con los parámetros predeterminados les dio buenos resultados, logrando una puntuación de precisión del 87,7% y una puntuación AUC de 0,76.

  2) Random Forest, este modelo requirió algunos ajustes de parámetros para optimizar su rendimiento. Inicialmente, ellos ejecutaron el modelo con todas las variables del conjunto de datos y lograr una precisión del 88,7%, sin embargo, la puntuación AUC disminuyó a 0.728.

  No obstante, para mejorar el rendimiento del modelo, utilizaron sklearn feature_selection's chi2 para seleccionar las variables con mayor importancia para el modelo. Esta técnica hizo que removieran 5 variables.

  Después de aplicar lo dicho anteriormente, volvieron a ejecutar el modelo de bosques aleatorios, el cual, dio como resultado un aumento en la precisión y el AUC, logrando un 89,9% y 0.771 respectivamente.

  3) XGBoost's Classifier, este modelo también utilizó el conjunto de datos donde fueron removidas 5 variables y dio como resultado la mejor puntuación AUC, siendo esta de 0.773 y con una precisión del 89,3%.

  Los investigadores indicaron que a pesar de que el XGB Classifier tuvo la puntuación AUC más alta, todos los modelos que ejecutaron puntuaron dentro de un rango relativamente pequeño entre sí, por lo que cualquiera de ellos sería Bueno para hacer las predicciones.

- **Análisis 2**
  El segundo estudio explora el comportamiento de búsqueda de los clientes en línea en un sitio web de comercio electrónico: Walmart.com. Xi Niu, Chuqin Li y Xing Yu utilizaron los métodos de aprendizaje automático bosque aleatorio y la regresión logística para desarrollar dos modelos computacionales.

  1) Random Forest. Para utilizar este método, ellos solo ajustaron el parámetro "mtry", en el cual, establecieron que mtry variara de 1 a 4 (≈15). Adicionalmente, para proporcionar una predicción más precisa, emplearon una validación cruzada de 10 con 15 repeticiones. Los mejores resultados los arrojó cuando mtry fue 3, logrando una precisión de 70,08%

2) Regresión Logística. Debido a que los 15 predictores que tenían estaban potencialmente correlacionados, utilizaron tanto el VIF (factor de inflación de la varianza) como el método LASSO para seleccionar variables para minimizar el problema de la multicolinealidad de las variables. Como resultado, se eliminaron cuatro variables, dejando el modelo final con 11 variables. Esto dio como resultado una precisión del 61%.

Al comparar ambos métodos en términos de la precisión, los bosques aleatorios dieron mejores resultados. Adicionalmente, los investigadores también hicieron una comparativa de los falsos negativos, ya que, en el contexto del comercio electrónico, el falso negativo, confundir a los compradores con no compradores, es más indeseado que el falso positivo, por lo que un mejor modelo también debe mantener una tasa baja de falsos negativos. Al hacer la compración de los modelos, las tasas de falsos negativos para el modelo de bosque aleatorio y el modelo de regresión logística fueron de 18,2% y 39,7% respectivamente. Desde este aspecto, el bosque aleatorio también es mejor en comparación con el modelo de regresión logística.

- **Análisis 3**

  Adil Mahmud Choudhury y Kamruddin Nur, propusieron un enfoque de aprendizaje automático para identificar clientes potenciales para una supertienda minorista. Ellos buscaron clasificar al cliente potencial basado en el comportamiento de compra previamente registrado.

  Después de haber corrido 4 modelos de aprendizaje automático, la precisión más alta que habían alcanzado fue de 56,78% con una regresión logística. Debido a ello, decidieron crear nuevas variables, para así poder mejorar los resultados.

  Ellos inicialmente habían analizado la cantidad de artículos comprados por cliente, sin embargo, luego decidieron calcular la media de compra de artículos para cada categoría. Posteriormente, estandarizaron las variables, debido a que tenían variables en diferentes escalas de medida. Finalmente, volvieron a aplicar los algoritmos.

  Debido a que no especifican si hubo ajustes en los parámetros, la figura 2 muestra los resultados de los algoritmos utilizados con las nuevas variables creadas.

Algorithm	Accuracy	Recall	Precision
Logistic Regression	98.49	99.56	97.00
Decision Tree Classifier	97.95	96.98	98.22
Support Vector Classifier	97.30	97.99	95.82
Random Forest Classifier	98.14	98.49	97.20
Multilayer Perceptron Classifier	99.41	98.93	99.68

Figura 2: resultados de los algoritmos de machine learning

El experimento de clasificación de clientes potenciales logró una precisión de predicción de hasta el 99,4% con una recuperación del 98,9% y una precisión del 99,7% utilizando el Multilayer Perceptron Classifier.

Ellos crearon nuevas variables para capturar la relación entre categorías, artículos, cantidad, unidad de medida y ventas; lo cual, provocó una mejora de hasta 42,6 puntos porcentuales con respecto a utilizar las variables originales.

Después de observar estos tres análisis, me llama poderosamente la atención que los bosques aleatorios resultaron ser el mejor algoritmo, o bien, el segundo mejor, dando como señales de que posiblemente sea un buen algoritmo para que yo aplique en el presente proyecto.

Adicionalmente, quiero resaltar que nunca había escuchado del algoritmo Multilayer Perceptron Classifier (MLP), por lo que intentaré indagar más en el tema para ver si este también pudiera ser utilizado para resolver mi problema.

2.4 Retos en las predicciones

Cuando trabajamos con datos, estos nunca son “color de rosa”, es decir, que presentan grandes retos el trabajar con ellos.

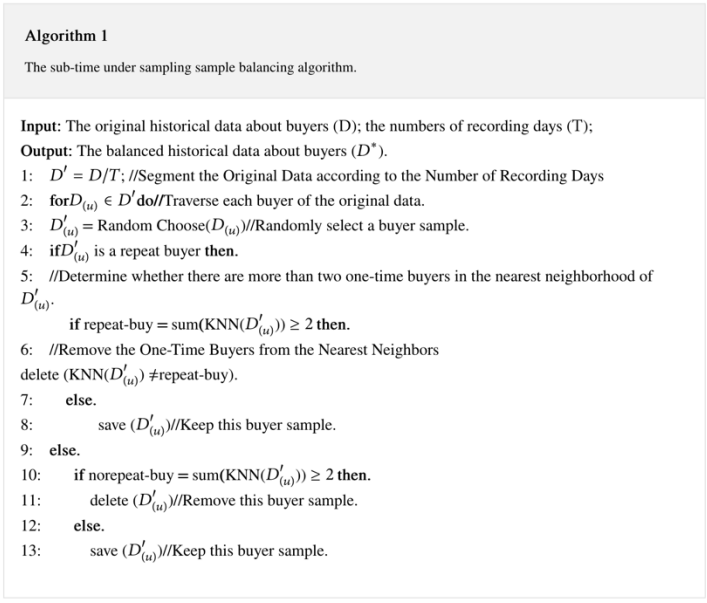
En este apartado, se quieren mostrar algunos de las dificultades que otras personas han encontrado al querer predecir potenciales compradores de un ecommerce y es que, por lo general, hay muy pocos compradores repetidos después de una promoción, lo que da como resultado el problema de que la muestra del comprador repetido y la muestra del comprador único no están equilibradas.

Esto quiere decir, que la base de datos no está balanceada y que es muy probable que los algoritmos fallen y no hagan buenas predicciones, pero que, a pesar de ello, muestren valores de precisión mayores al 90%. En otras palabras, cuando una base de datos tiene una variable de clasificación binaria (0 y 1) y el 98% son ceros y tan solo el 2% son unos, es muy probable que todas las predicciones sean 0, y por ende, la precisión en el resultado sea alta. El gran inconveniente es que, si todos son ceros, no estamos logrando nada, ya que nuestro objetivo es predecir los unos.

Una distribución desequilibrada de la variable independiente es un problema potencial para entrenar el modelo, ya que los algoritmos de clasificación más comunes minimizarían la tasa de error general en lugar de prestar especial atención a la clase minoritaria (Chawla, 2005).

A pesar de que este es un problema conocido y muchos lo mencionaron, no todos indicaron cómo lo resolvieron.

Para corregir este problema, Huibing Zhang y Junchao Dong (2020), utilizaron el método de sub-tiempo de muestreo que se muestra en la figura 3, para equilibrar las muestras.



Algorithm 1

The sub-time under sampling sample balancing algorithm.

Input: The original historical data about buyers (D); the numbers of recording days (T);
Output: The balanced historical data about buyers (D^*).
1: $D' = D/T$; //Segment the Original Data according to the Number of Recording Days
2: **for** $D_{(u)} \in D'$ **do//**Traverse each buyer of the original data.
3: $D'_{(u)}$ = Random Choose($D_{(u)}$)//Randomly select a buyer sample.
4: **if** $D'_{(u)}$ is a repeat buyer **then.**
5: //Determine whether there are more than two one-time buyers in the nearest neighborhood of $D'_{(u)}$.
if repeat-buy = sum(KNN($D'_{(u)}$)) ≥ 2 **then.**
6: //Remove the One-Time Buyers from the Nearest Neighbors
delete (KNN($D'_{(u)}$) ≠repeat-buy).
7: **else.**
8: save ($D'_{(u)}$)//Keep this buyer sample.
9: **else.**
10: **if** norepeat-buy = sum(KNN($D'_{(u)}$)) ≥ 2 **then.**
11: delete ($D'_{(u)}$)//Remove this buyer sample.
12: **else.**
13: save ($D'_{(u)}$)//Keep this buyer sample.

Figura 3: Algoritmo 1, El sub-tiempo bajo el algoritmo de balanceo de la muestra de muestreo.

De acuerdo con la característica “tiempo”, la muestra original de compradores repetidos y compradores únicos se segmenta según el día. Para cada comprador de la muestra original, los tres compradores vecinos más cercanos se determinan de acuerdo con su distancia euclidiana. Si el comprador es un comprador único y más de dos de sus tres compradores vecinos más cercanos son compradores habituales, se elimina. Si la muestra es un comprador habitual y más de dos de sus tres vecinos más cercanos son compradores únicos, se eliminan los compradores únicos de los vecinos más cercanos. El resto se mantiene en la muestra original del comprador.

Mientras que, Niu, Xi et al., decidieron adoptar el método de *under-sampling*, la cual es una técnica que realiza un submuestreo de la clase mayoritaria, lo que implica eliminar patrones para emparejar las clases utilizando alguna regla o criterio.

Por otro lado, algunos investigadores han mencionado que el utilizar únicamente las variables que vienen en la base de datos no han sido suficientes para hacer buenas predicciones, por lo que han tenido que hacer *feature engineering* para crear nuevas variables y así mejorar las predicciones.

Huibing Zhang y Junchao Dong (2020) dijeron que ellos tenían muy pocas características en los datos originales que se pudieran aplicar directamente para la predicción de compradores repetidos y que las predicciones no eran ideales, por lo que haciendo análisis estadísticos aplicaron la asignación de Dirichlet latente (LDA), el análisis de componentes principales (PCA) y los métodos de aprendizaje automático de la máquina de factorización (FM) para construir las características del comprador repetido.

Martin Beck (2019) lo que hizo fue cambiar las variables categóricas en variables One-Hot Encoded, es decir, en variables binarias.

Otro de los inconvenientes que se han presentado antes se indicó en el apartado anterior, donde Xi Niu, Chuqin Li y Xing Yu mencionaron haber tenido problemas con las correlacionados entre las variables, por lo que debieron de utilizar tanto el VIF

(factor de inflación de la varianza) como el método LASSO para seleccionar variables y así minimizar el problema de la multicolinealidad.

Otros de los métodos que utilizaron algunos de los analistas fue convertir algunas de las variables numéricas en variables más resumidas como promedios o medias por categorías u otras subdivisiones de las variables.

Finalmente, se podría decir que “el *feature engineering* es a menudo el factor más importante para el éxito de una tarea de predicción, pero no se puede encontrar mucho trabajo en la literatura sobre ingeniería de características para tareas de predicción en el comercio en línea.”, Liu, G. et al. (2016).

3. Diseño e implementación

Este proyecto siguió la metodología CRISP-DM, en la cual, muchas de las etapas se traslaparon conforme se iba avanzando y se iba obteniendo un mejor conocimiento de los datos y del negocio.

Como bien es sabido, lo normal en los proyectos de *machine learning* es tener procesos iterativos, donde se repiten múltiples veces las tareas, se reprocesa la información y constantemente se deben estar haciendo ajustes y modificaciones en los datos para poder alcanzar los objetivos planteados.

A continuación, se detallarán los pasos seguidos en este trabajo para poder hacer la implementación del modelo de predicción.

3.1 Recolección y descripción de los datos

El propietario del *ecommerce* tiene información que utiliza para la captación en internet, uso de cookies, estrategias de SEO y SEM, y acuerdos con portales.

Los datos sobre el comercio fueron compartidos por del profesor colaborador Santiago Rojo, quien es el tutor de este trabajo final.

Él hizo entrega de dos ficheros planos con campos delimitados, los cuales contienen datos completamente anonimizados para evitar tener acceso a información sensible y/o privada.

Adicionalmente se utilizó un archivo de Excel con los códigos de actividad CIIU (Clasificación Industrial Internacional Uniforme) adaptados para Colombia, los cuales fueron adquiridos por medio de la página web del Departamento Administrativo Nacional de Estadística (DANE).

Los conjuntos de datos con sus características son los siguientes:

1. **Usuarios**

   Este archivo tiene la información de 368.220 usuarios registrados en el comercio y sus 19 variables se muestran a continuación:

- ID_USUARIO. ID único del usuario
- TIPOUSUARIO. Tipo de Usuario: PJ=Persona Jurídica, PF= Persona física/natural, PX= Puede ser PJ pero no es seguro.
- FECHA_REGISTRO. Fecha en que el usuario se registró.
- CANAL_REGISTRO. Canal de registro del usuario.
- IND_CLIENTE. Indica si el usuario es cliente, es decir, se efectuó compra o no. (0= No Cliente, 1= Cliente)
- IND_ALTA. Indicador de alta. Antes de ir a la pasarela de pago para comprar, se marca este valor. (0= No fue a la pasarela de pago, 1= Sí fue a la pasarela de pago)
- FECHA_ALTA. Fecha del indicador de alta.
- FECHA_CLIENTE. Fecha en que el usuario se convierte a cliente.
- TIPOEMAIL. Tipo de email del usuario.
- BONDAD_EMAIL. Bondad/Ponderación del email. Resultado de campañas de emailing. (20= Verde [ok], 9= Naranja [ha dado un error temporal pero siguen enviando correos], 1= Spam, 0= Rojo [inválido], -10=dominio inválido [inválido], -20= no email).
- USU_TELF. Teléfono del usuario anonimizado.
- IPCASOS. Número de usuarios que utilizan la misma IP.
- IP_Country. País del usuario a partir de su IP.
- IP_Region. Región del usuario a partir de su IP.
- USU_TIPO. Tipología de la empresa si TIPOUSUARIO=PJ.
- USU_TAMANIO. Tamaño de la compañía. (GR= grande, MD=mediana, PQ=Pequeña, MC=Micro, SD=Sin Definir).
- USU_CIIU. Código de actividad si TIPOUSUARIO=PJ.
- USU_ESTADO. Estado/Situación de la compañía si TIPOUSUARIO=PJ.
- USU_DEPARTAMENTO. Departamento/Provincia del usuario si TIPOUSUARIO=PJ.

2. Consumos

Este set de datos tiene 928.816 filas (eliminando duplicados). Cada fila equivale a un consumo realizado por un usuario. Este set tiene 11 variables, las cuales se describirán a continuación:

- IDCONSUMO. Identificador único del consumo.
- IDUSUARIO. Identificador del usuario.
- IDPRODUCTO. Identificador de producto, del producto consumido.
- DESCPRODUCTO. Descripción del producto, del producto consumido.
- IDGRUPOPROD. Identificador del grupo de productos al que pertenece el producto consumido.
- DESCGRUPOPROD. Descripción del grupo de productos al que pertenece el producto consumido.
- FECHACONSUMO. Fecha del consumo.
- EMPCONSUL_ID. ID único de la empresa asociada al producto consumido.
- EMPCONSUL_CIIU. Código de actividad CIIU de la empresa asociada al producto consumido.
- EMPCONSUL_PROV. Departamento de la empresa asociada al producto consumido.
- EMPCONSUL_EST. Estado de la empresa asociada al producto consumido.

3. **Códigos CIIU**

   La tabla con códigos tiene un total de 710 filas y 4 columnas, las cuales se detallarán a continuación:

   - División.
   - Grupo.
   - Clase.
   - Descripción.

3.2 Análisis exploratorio de los datos

Sabiendo que la clave de un buen modelo de predicción está en la calidad de su información, en esta sección se realizará un análisis exploratorio de los datos, mejor conocido por sus siglas en inglés EDA (*Exploratory Data Analysis*).

El EDA consiste básicamente en analizar, investigar y resumir las características principales de los conjuntos de datos por medio de tablas y visualizaciones, buscando determinar las mejores formas de manipular la información para posteriormente obtener las respuestas que se necesitan. Parte del interés es descubrir si existen patrones, anomalías o información relevante que verifique alguna suposición.

A pesar de que las técnicas de EDA fueron desarrolladas originalmente por el matemático estadounidense John Tukey en la década de 1970, hoy en día se siguen utilizando principalmente para ver qué datos pueden revelar algo más allá del modelado formal o la tarea de prueba de hipótesis, proporcionando una mejor comprensión de las variables del conjunto de datos y las relaciones entre ellas.

Para realizar este proceso, se utilizaron dos herramientas: RStudio, el cuál, es un entorno de desarrollo integrado para R (un lenguaje de programación para gráficos y computación estadística), y Tableau, que es un software de visualización de datos.

A manera de resumen, EDA consiste en realizar análisis de una y dos variables y para ello, se siguen los siguientes pasos:

- Primer acercamiento a los datos
- Analizar variables categóricas
- Analizar variables numéricas
- Analizar categóricas y numéricas al mismo tiempo.

Parte de lo que se busca en estos análisis iniciales es ver si existen valores extremos, valores perdidos, tipo de datos mal asignados y conocer las distribuciones.

Como se cuentan con dos conjuntos de datos, primeramente, se trabajó con el de usuarios y posteriormente con el de consumo.

1. **Usuarios**

El primer paso fue cargar el archivo de texto plano, donde cada columna fue separada por “;” y todos los valores que estuviesen en blanco o nulos, fueron etiquetados como “NA”.

Al analizar la estructura de los datos, esta indica que hay 6 variables clasificadas como *integer* y 13 como factores, no obstante, no todas fueron asignadas correctamente, por lo que se deben de cambiar a la clase correspondiente.

Las variables FECHA_REGISTRO, FECHA_ALTA y FECHA_CLIENTE estaban clasificadas como factores y no como fechas, por lo que este fue el primer cambio que se hizo.

Posteriormente se tomaron las variables que estaban como *integer* ('ID_USUARIO', 'CANAL_REGISTRO', 'IND_CLIENTE', 'IND_ALTA', 'BONDAD_EMAIL') y fueron transformadas a factores, debido a que, a pesar de que tienen valores numéricos, son variables categóricas que simplemente fueron codificadas numéricamente. Finalmente, a la variable IPCASOS se le cambió la clase, pasando de *integer* a numérica.

Después de haber cambiado las clases de las variables, el siguiente paso fue analizar las variables de tipo fecha.

La fecha de registro muestra que hay usuarios que se registraron desde el primero de enero del 2018 al 31 de diciembre del 2019, abarcando así, 2 años de información. Esto es importante debido a que entre más histórico de datos se tenga, mejor pueden llegar a ser las predicciones.

La fecha de alta y de cliente muestran 364.500 y 365.600 nulos respectivamente, lo cual, es esperado debido a que no todos los usuarios fueron al botón de pago y no todos se convirtieron en cliente. Para validar que esos datos estuvieran correctos, se hizo una validación cruzada con las variables de indicador de alta y cliente, para buscar que todos los que tienen un indicador de alta, tengan fecha, y de manera similar con los clientes, si son clientes, se debe de tener la fecha en que se convirtieron.

En este análisis se encontraron distintos escenarios, donde el primero fue que tuviera el indicador de alta, es decir, un valor de 1, pero que la fecha fuera NA. En este escenario se encontraron 3 casos. El segundo fue cuando tienen fecha de alta, pero el indicador de alta es 0, donde se hallaron 4 observaciones. Finalmente, en el tercero hubo 6 filas donde se indicaba que el usuario era cliente (valor de 1) pero se desconocía la fecha es que se convirtió.

Haciendo un análisis entre la fecha de alta y la fecha de cliente, se encontró que en el 92% de las veces ambas fechas son iguales. Bajo esta premisa y tomando en cuenta que cuando faltaba la fecha de alta tenían la de cliente y viceversa, se imputó la fecha faltante usando la información de la que sí se disponía.

En el caso del escenario dos donde se contaba con fecha, pero el indicador de cliente no estaba, se tomó la decisión de quitar las fechas y dejar el valor de NA. A pesar de que también existía la posibilidad de asignarle un valor de 1 al indicador, observando las fechas de alta parecía sospechoso que las fechas de los 4 usuarios fueran consecutivas (del 23 al 26 de enero del 2021) a pesar de que las fechas de

registro eran muy distintas entre ellas (algunas en 2018 y otras en 2019). Bajo esa premisa fue que se optó por el valor NA.

Siempre hablando de fechas, de los usuarios que hay en la base de datos, el primero que se hizo cliente fue el 2 de enero del 2018, mientras que el último lo hizo el 22 de enero del 2021. Si se analiza la cantidad de días que transcurrieron desde que un usuario se registró hasta que se hizo cliente, la moda y la media indican un valor de 0, es decir, que lo más común es que quienes pagaron para ser clientes, lo hicieron el mismo día en que se registraron, por otra parte, quienes duraron más en dar ese paso, lo hicieron en 1038 días, es decir, casi 3 años después de su registro.

Después de haber estudiado y analizado las variables tipo fecha, se prosiguió con las variables cualitativas que, en su defecto, R las toma como factores.

La primera variable en analizarse fue el ID del usuario, la cual, aunque pareciera ser insignificante, es importante porque al ser usuarios, estos deben de tener un registro único. Al crear una tabla con frecuencias absolutas, se logra ver que existen varios duplicados y estos deben de ser eliminados para contar con registros únicos. Al remover estos, el conjunto de datos pasó de 368.220 a 367.705.

La segunda variable en ser analizada fue el indicador de cliente debido a que esta es la variable que posteriormente vamos a querer predecir.

Observando la figura 4, encontramos que el 99,3% de los usuarios no son clientes, en contraste al 0,7% que sí lo son. Estas cifras lo que indican es que la base de datos está muy desbalanceada y que, para poder hacer una correcta predicción, hay que tomar esto en cuenta.

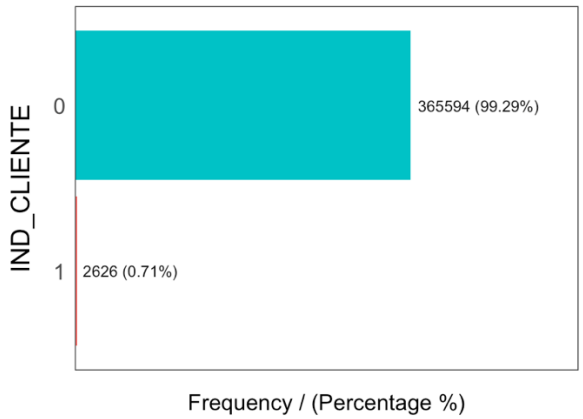


Figura 4: Frecuencia de usuarios según indicador de cliente

El principal problema de trabajar con bases tan desbalanceadas es que fácilmente se puede tener un modelo con un 99% de precisión, pero con el gran inconveniente de que a todos los usuarios registrados los etiqueta como que no serán futuros clientes, es decir, no sabe identificar futuros compradores.

Debido al resultado anterior, era de esperar que el indicador de alta tuviera valores similares al indicador de cliente, esto debido a que el indicador de alta se da cuando los usuarios son llevados al botón de pago del comercio. En términos generales, solo

el 1% de los usuarios registrados fueron a la página de pago, sin embargo, de ese grupo, solo el 71% de ellos se hizo cliente.

La siguiente variable es ser analizada fue el tipo de usuario, donde el 72,3% de los usuarios indicó ser una persona física, mientras que las personas jurídicas representan el 24,4%. Pero, al observar la figura 5, se puede apreciar que entre los usuarios que son clientes, el porcentaje de personas jurídicas y de las físicas son relativamente similares, teniendo estos un 44,9% y un 49,8% respectivamente.

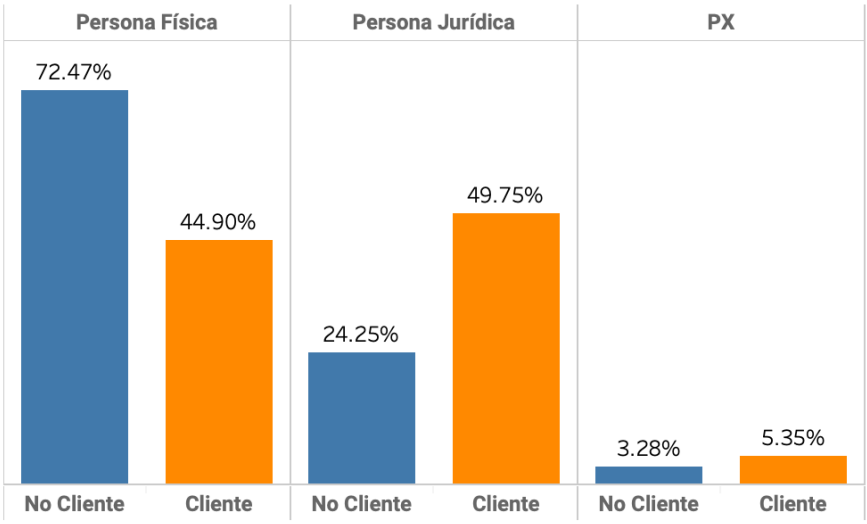


Figura 5: Tipo de usuario según indicador de cliente

En cuanto al canal de registro, no existe uno que sea tan predominante ante el resto, siendo el 3 el que tiene un mayor porcentaje (32,3%), seguido del 2 (23,0) y el 8 (12,95). Es importante mencionar que no se tiene el detalle de lo que representa cada valor en el canal de registro, por lo que se seguirá tratando con la etiqueta correspondiente. Nuevamente, si nos enfocamos en los canales de registro de los que se hicieron clientes, el orden cambia, siendo el canal 1 el de mayor porcentaje con un 22,3%, seguido del 7 y 3 con valores de 19,6% y 19,4%.

La quinta variable en ser estudiada fue el tipo de cuenta de email que utilizaron los usuarios para registrarse. De manera global, aproximadamente el 79% de las personas utilizaron un correo que fuese de Google, Yahoo, Hotmail u otra de estas empresas que brindan ese servicio. Si luego analizamos únicamente a los clientes, el porcentaje que utiliza esas plataformas es del 62%, mientras que los que utilizan un correo corporativo es del 23%, el cual es considerablemente mayor al 9% que representa entre la totalidad de los datos.

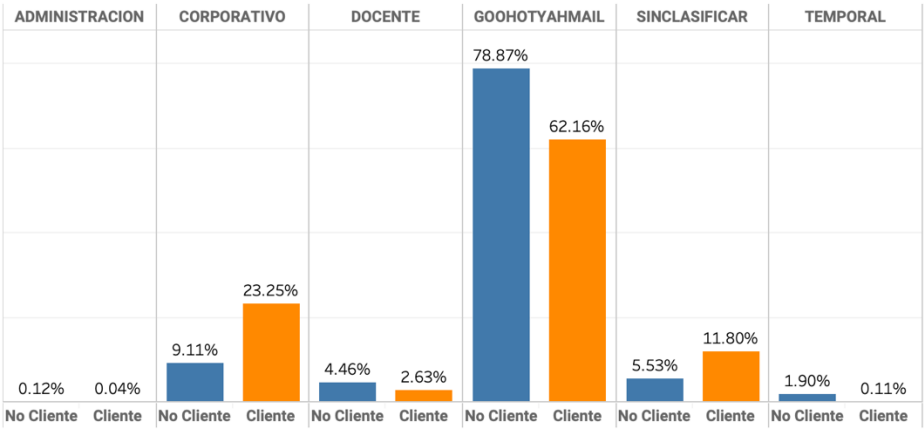


Figura 6: Tipo de email según indicador de cliente

Al estudiar el tipo de email, también es importante conocer si esas cuentas que fueron usadas para registrarse son válidas, es decir, que existen realmente. Para chequear eso, la base tiene una variable llamada bondad de ajuste, la cual busca capturar y medir esa información.

Los resultados indican que entre los usuarios que no son clientes, aproximadamente el 75% de ellos dieron un correo válido, mientras que los clientes lo hicieron en el 93% de los casos.

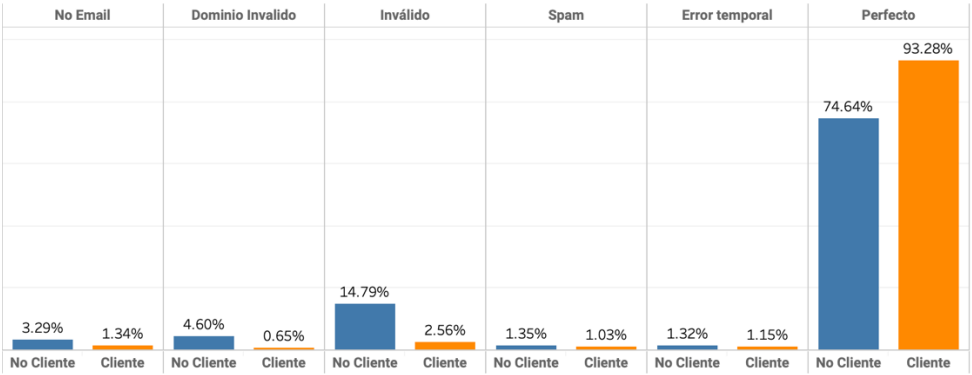


Figura 7: Bondad de email según indicador de cliente

Posteriormente, se analizaron las siguientes variables relacionadas a la empresa: tipología, tamaño, estado, departamento y código de actividad CIIU. En todas ellas, el porcentaje de valores nulos ronda el 77%. Este porcentaje es bastante alto, por lo que tratar de imputar tantos valores sería muy riesgoso y la probabilidad de que no se haga bien es elevada.

Una de las grandes dudas que surgían era si descartar esas variables por la cantidad de nulos o si tomarlas en cuenta. A pesar de que busqué literatura sobre eso, no había como un consenso, por lo que se tomó la decisión de dejarlas en la base de datos y posteriormente ver si estas generan algún valor.

Otro aspecto importante es que, entre los clientes, el 50% de ellos sí tenían esa información disponible, por lo que a futuro puede que alguna de esas variables logre aportar algo en la predicción.

Finalmente, las últimas variables cualitativas estudiadas fueron el país y la región obtenidas por medio de la IP. Con respecto al país, el 94% de los usuarios se conectaron desde Colombia, mientras que solo el 1,5% lo hizo desde otros países. El porcentaje faltante corresponde a valores nulos. Por otra parte, en el 41% de los casos no se logró determinar la región. En Bogotá (23%) y Antioquía (15%) fueron las dos regiones de donde se identificaron más usuarios.

El último paso del análisis exploratorio en los datos de usuarios es analizar las variables cuantitativas. La columna ipcasos es la única variable numérica en este *dataset*. Calculando las estadísticas de resumen tenemos que el mínimo de usuarios que utilizan la misma IP es 0, mientras que 16.393 es el valor máximo.

Al realizar un boxplot con esta variable, este no se observa con claridad ya que el percentil 75 es 6, pero como vimos anteriormente, el valor máximo es sumamente alto, lo que hace que alrededor del 19% de los datos se vean como valores extremos.

Lo que llama poderosamente la atención (ver tabla 1) es que las microempresas (MC) sean las que tienen más casos con más de 1000 usuarios conectados desde la misma IP, por supuesto, omitiendo los NA. En lo personal, hubiese esperado que ese comportamiento se diera en las empresas grandes.

	NA	SD	MC	PQ	MD	GR	Grand Total
Menos de 10	223,674	1,498	53,887	9,218	5,077	4,368	297,722
Entre 11 y 100	27,970	174	4,245	445	439	1,154	34,427
Entre 100 y 500	7,855	42	821	112	114	208	9,152
Entre 500 y 1000	5,482	33	441	71	64	52	6,143
Más de 1000	18,608	69	1,232	267	45	40	20,261
Grand Total	283,589	1,816	60,626	10,113	5,739	5,822	367,705

Tabla 1: Cantidad de Usuarios conectados desde la misma IP según tamaño de la empresa

Posteriormente, se tomó la variable USU_CIIU y se desagregó en 3 nuevas variables, siendo la primera el área, la cual corresponde a la letra inicial de cada código CIIU. La segunda fue la división, que son los primeros 2 dígitos del código, y la tercera fue la clase, la cual incluye los 4 números del código. Esta separación lo que busca es no solo poder analizar la variable en menos factores, sino también, ver si de forma separada puede llegar a tener mayor importancia al momento de crear un modelo. Estas 3 nuevas variables se les asignó la clase factor.

2. Consumos

Al igual que se hizo con el archivo de usuarios, la idea es realizar un análisis exploratorio del conjunto de datos de consumo.

El primer paso fue exactamente igual a lo realizado con el archivo anterior, se cargó el archivo de texto plano, se hizo la separación de columnas por medio de “;” y a los valores blancos o nulos se les asignó “NA”.

El segundo paso antes de iniciar el análisis de las variables fue eliminar los duplicados, pasando de 938.580 filas a 928.816.

Como se mencionó en el apartado de la descripción de los datos, el archivo de consumo tiene 11 variables, donde 1 es una fecha y las restantes son categóricas.

Lo tercero que se hizo fue un análisis de valores nulos, lo cual mostró que solo 3 variables tenían valores nulos y corresponden a 1848 usuarios a los que no se les identificó el estado, la provincia ni el código CIIU de la empresa. Estas 3 variables tienen más de 10 categorías cada una, por lo que no se comentaran en detalle.

En cuanto a las variables IDCONSUMO, IDUSUARIO, IDPRODUCTO, IDGRUPOPROD y EMPCONSUL_ID, estas, aunque son factores, simplemente son identificadores, por lo que existen muchos niveles diferentes y por ende se decidió no analizarlas, ya que no generan ningún valor.

La siguiente variable en ser estudiada fue la descripción del grupo de productos. Esta variable fue estudiada desde dos perspectivas diferentes, la primera, era conocer qué grupos eran los que más se consumían, dando como resultado que la ficha básica promocional fue el más popular con un 47,4%, seguido del perfil promocional (45,7%) y la ficha avanzada (5,8%). La segunda forma de analizar esta variable fue conocer el porcentaje de usuarios que consumió cada grupo, el cual indica que el 93% de los usuarios al menos una vez hicieron consumieron el perfil promocional, mientras que la ficha básica promocional fue consultada solo por el 31,9% de los usuarios. Los otros grupos de productos fueron los menos populares y es lógico debido a que son los grupos de productos de pago.

Al separar los grupos de productos entre los clientes y los que no, se logra ver que el perfil promocional a pesar de que es el grupo más consumido a nivel general, se debe a que es el grupo de producto más consumido por aquellos que solo usan los servicios gratuitos, mientras que aquellos que pagan, se inclinan en un 37% en la ficha avanzada,

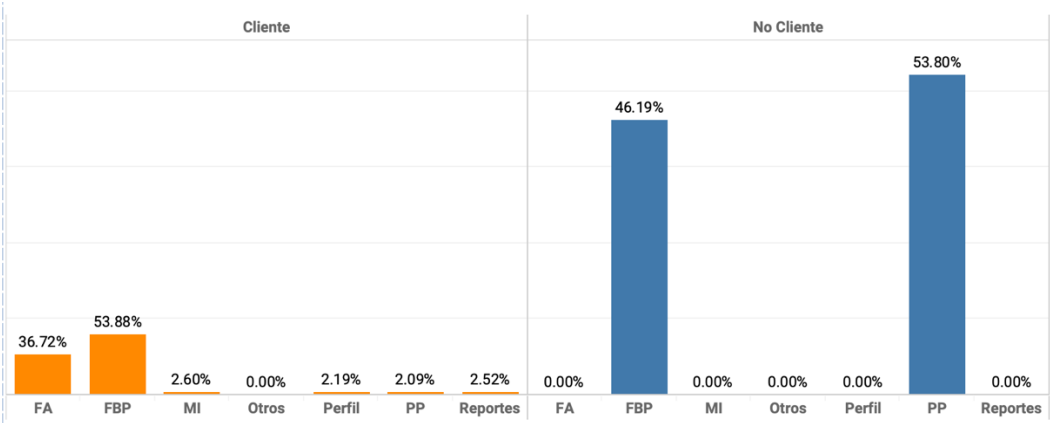


Figura 8: Consumo de usuarios por grupo de productos según indicador de cliente

Estudiar la descripción de los productos tampoco genera ningún valor, ya que lo único que hace es desagregar los grupos de productos de pago, es decir, los dos productos de acceso gratuito son los más consumidos y los de pago los menos.

Finalmente, se estudió el estado de las empresas consultadas por los usuarios, siendo activa el estado más común con un 81%, luego cancelación con un 12% y en liquidación, un 4% de ella.

3.3 Creación del repositorio final

Uno de los objetivos de este proyecto es crear un repositorio final, el cual, permita combinar la información de los usuarios como el de su consumo, para que, con base en la información de ambos archivos, se pueda crear el o los diferentes modelos de *machine learning* que faciliten la identificación de posibles compradores en el *ecommerce*.

Unir dos tablas se puede decir que es un trabajo relativamente sencillo, sin embargo, el reto que se presenta acá es que si combinamos los dos archivos a como están actualmente, lo que pasaría es que al *dataset* de consumo simplemente se le agregarían las columnas que hacen referencia a las características de los usuarios y, el principal inconveniente con esto, es que para poder hacer predicciones y perfiles de clientes, lo que se ocupa mas bien, es tener una sola fila por usuario y no múltiples.

El gran desafío que se presenta en este punto es tomar la información de consumo y resumirla en nuevas variables que logren explicar y/o describir el consumo realizado.

No obstante, antes de buscar unir la información de ambos archivos, se describirán algunos pasos de limpieza, preprocesado y preparación de los datos que permitieron la creación del repositorio final y que durante el EDA no se profundizaron.

Cuando se trabaja con variables categóricas que no han sido etiquetadas, sino que están en formato texto, suele pasar que algunas de las palabras tengan algún acento y el problema es que como el programa R no las reconoce, las cambia por una combinación de caracteres. Para evitar trabajar con estas palabras con caracteres especiales, se hizo la recodificación manual buscando sustituir la letra con acento por una sin acento. Adicionalmente, también se encontró el problema de que en ocasiones los países o regiones estaban tanto en español como en inglés, por lo que se dejó únicamente la opción en español, modificando entonces las que estaban en otro idioma.

De momento, en este punto no se hará una codificación de etiquetas en las variables categóricas, ya que existen modelos que saben tratarlas, por lo que más adelante cuando se estén creando los modelos, se decidirá si se les dará otro tratamiento para mejorar algún algoritmo.

Dejando atrás el proceso de recodificación, se decidió tratar de imputar el país por medio del teléfono del usuario. Para ello, se seleccionaron los prefijos de los teléfonos que tuvieran asociado solamente un país y con base en ellos, se imputaron algunos valores faltantes. Esto redujo la cantidad de valores nulos, pero no los eliminó por completo.

Después de haberle dado tratamiento a las variables existentes, se pensó en nuevas variables que pudiesen aportar valor al modelo y así tener mayores opciones al momento de crear las predicciones.

3.3.1 Creación de nuevas variables

La base de usuarios y de consumo incluyen un código de actividad CIIU, la cual debería de constar de una letra y 4 dígitos. La letra indica el área de la actividad, los primeros dos dígitos son la división y finalmente, los 4 dígitos en conjunto forman la clase. Para reducir la cardinalidad y el exceso de clases en esta variable, se tomó la decisión de crear una variable en cada archivo indicando el área del usuario y el área de la empresa consultada. Utilizando el área se logra reducir la granularidad y se agrupan mejor las actividades, esperando que esta nueva variable aporte más al modelo.

En el momento en que se estaban creando las variables anteriores, me percaté de que algunos casos no tenían el código de área CIIU debido a que la letra estaba ausente y únicamente se contaba con el número de clase. Como cada área y número de clase son únicas y están asociadas entre ellas, se tomó el archivo de códigos CIIU y se importó la clase a las filas que tenían ese valor como nulo. No se logró identificar todas las áreas debido a que algunos códigos CIIU ya no existen o fueron descontinuados.

Otra de las variables que tiene muchos niveles (alrededor de 80 distintos países) es la de IP_Country. Esta variable fue transformada en una variable dicotómica basándose en lo observado en la figura 9 donde se ve que el 94,4% de los casos provienen de Colombia. Los valores que tomó esta variable son 1 cuando la IP corresponde a Colombia y 0 cuando corresponde a otro país.

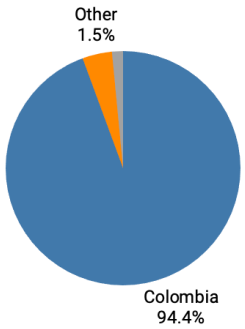


Figura 9: Distribución de los países según la IP

Como se indicó al principio de este apartado, el archivo de consumo no se puede unir con el de usuarios, sino que se debía buscar la manera de resumir la información de consumo basada en cada usuario para poder unir toda la información.

Para lograr lo anterior lo que se buscó fue crear un nuevo repositorio que incluyera únicamente las variables creadas a partir del resumen de consumo de los usuarios.

Como primer paso, en el archivo de consumo se renombró la variable IDUSARIO a ID_USUARIO para que, al momento de unir los datos, se pudiera hacer mediante la misma variable, la cual, se espera que tengan el mismo nombre.

Lo segundo fue agregarle a los datos de consumo la fecha en que los usuarios se hicieron clientes (en caso de que lo hayan hecho), para así, poder identificar cuántas compras y cuales productos se consumieron de forma gratuita y cuales pagando.

Utilizando la fecha de consumo y la de cliente, se creó una variable dicotómica llamada “SE_PAGO”, la cual indique que, si el consumo se hizo antes de hacerse cliente o si del todo nunca se hizo cliente, esta fue gratuita y se etiquetó con un 0; de lo contrario, si se el producto fue adquirido después de haberse hecho cliente, significa que pagó por él y se le asignó el valor de 1.

Debido a que solo nos interesa lo que haya pasado antes de que los clientes se hicieran clientes, se filtró la base para ver únicamente los consumos etiquetados con 0, es decir, por los que el usuario no pagó. La idea es buscar qué patrón o patrones hubo antes de convertirse en cliente. Esta nueva base representa el 85% de los consumos totales.

Con esta base filtrada, se crearon las primeras dos características que resumen el consumo, la primera fue el total de consumos y la segunda, el total de productos distintos consumidos.

Siempre bajo la misma línea, se sabe que cada consumo de los usuarios está asociado a un producto, el cual, pertenece a un grupo de productos. Como la mayoría de los productos solo pueden ser consumidos por los clientes (usuarios que pagaron), se decidió crear una variable por cada grupo de productos que indique cuantos consumos se hicieron de ese grupo. De este ejercicio se crearon 7 variables, sin embargo, se eliminó la del grupo “Otros” debido a que nadie la había consumido.

Adicionalmente, se crearon las variables “DIAS_CONSUMO” y “CON_PROM_DIA”, que hacen referencia a la cantidad de días en los que han hecho consumos y al promedio de consumos realizados por día.

Por otra parte, existe una variable asociada al estado de las empresas consumidas, es decir, indican si estas están activas, inactivas, extinguidas, etc. Como es información disponible en la base de consumo, se hizo un ejercicio similar al de los grupos de productos, es decir, se creo una variable por cada estado que indica cuantos consumos se hicieron sobre empresas con ese estado en específico, dando como resultado la creación de 13 nuevas columnas

Como último paso, ya que se desconoce qué variables o información pueden ser útil para llegar a predecir posibles clientes, se creó también la variable “MISMO_CIIU” que indica la cantidad de consumos donde el área del CIIU del usuario registrado es la misma de la empresa/usuario consumido.

La base de consumo que tenía 11 variables fue transformada en un nuevo *dataset* con 25 variables, en la cuál se intentó resumir de la mejor manera la información de consumo para cada cliente.

Este nuevo conjunto de datos fue unido al de usuarios por medio de la variable ID_USUARIO, dejando así un conjunto de datos con 44 variables.

3.3.2 Selección de variables inicial

Este segmento del trabajo lo que intenta es remover de manera natural las variables que de antemano se sabe que no van a aportarle al modelo y así, dejar listo el repositorio final de datos. Es importante mencionar que en este punto no se hará la selección de variables según su nivel de importancia ni se evaluará multicolinearidad.

A este momento el nuevo repositorio de datos contiene 44 columnas, unas que provenían del archivo de usuarios y otras que fueron creadas y agregadas posteriormente. No obstante, eso no significa que todas funcionan o aportan algún valor.

Las primeras columnas en ser removidas son las fechas y las variables con muchos niveles, debido a que estas no aportan en la creación de un modelo. En total, estas fueron 6 variables, fecha_registro, fecha_alta, fecha_cliente, usu_telf, usu_ciiu e ip_region.

Finalmente, el repositorio quedaría conformado por 38 variables.

3.4 Análisis predictivo

A pesar de que en la sección anterior se creo el repositorio final, no quiere decir que todas esas variables realmente aporten algo en la creación de un modelo predictivo.

En este apartado, no solo nos centraremos en la creación de el o los modelos, sino también, en saber elegir cuáles son las variables más relevantes o de mayor importancia que ayuden a crear modelos más precisos.

Antes de entrar en detalle con lo antes mencionado, se tomó la decisión de trabajar con dos grupos de poblaciones distintas.

El análisis exploratorio de los datos mostró que existe un gran conjunto de usuarios de los que no se tienen muchas de las características, por lo que el grupo 1 estará conformado por todos aquellos usuarios que indicaron el tamaño, el tipo, el estado en que se encuentran y su código CIIU; mientras que en el grupo 2 estarán todos aquellos donde estos valores sean nulos.

El objetivo de realizar esta separación es pensando que es mas difícil de encontrar un patrón con tantos valores nulos, por lo que esperaría que el modelo predictivo del grupo 1 al tener más información sea más preciso que el del segundo grupo.

Con esta separación, el grupo 1 quedaría con un total de 84.116 usuarios, de los cuales, el 98,5% corresponde a no clientes y el 1,5% a clientes. Por su parte, el grupo 2 que es el de mayor cantidad quedó con 283.589 usuarios donde el 0,5% de ellos consumió productos de pago mientras el 99,5% restante lo hizo únicamente con los de acceso gratuito.

En ambos grupos se presenta el problema de desbalanceo de los datos, ya que el porcentaje de usuarios que se convirtieron en clientes es sumamente bajo.

3.4.1 Elección de atributos a utilizar en el entrenamiento y testeo de los algoritmos

Al haberse creado 2 grupos, es posible que para cada set de datos las posibles variables predictoras sean diferentes, por lo que se hará un análisis en cada uno de los grupos de forma independiente.

En ambos casos, antes de iniciar con la selección de variables, el primer paso es separar los datos creando dos *datasets*, el de entrenamiento y el de prueba.

- **Grupo 1: Información completa**

Para separar los datos en el conjunto de *train* y *test*, se utilizó la relación de división 80:20, es decir, el 80% de los datos serán utilizados para entrenar el modelo, mientras que el 20% restante para comparar los resultados.

Luego separar los datos, se revisó que los datos quedaran bien balanceados, es decir, que el porcentaje de clientes y no clientes fuese similar al obtenido al analizar todos los datos. En ambos casos se obtuvo un 1,5% de clientes, siendo este igual al calculado previamente.

Definida ya la base de entrenamiento, el siguiente paso es hacer la selección de características para así reducir el número de variables de entrada a aquellas que se cree que son más útiles para un modelo a fin de predecir la variable objetivo.

Lo primero que se empleó es el método de filtros de selección, lo cual, implica buscar las relaciones entre las variables. Esto se dividió en dos pasos: el primero, analizar la correlación entre las variables numéricas para ver si existe multicolinealidad entre ellas, y el segundo, analizar las relaciones entre las variables categóricas con la variable a predecir.

En términos generales, la mulicolinealidad quiere decir que, al tener dos variables altamente relacionadas, podríamos estar sobre ajustando los datos si se utilizaran ambas variables, por lo que se debería de eliminar variables redundantes.

Observando el gráfico de correlaciones, se puede apreciar como hay variables altamente relacionadas entre sí, por ejemplo, el consumo total esta muy correlacionada con la ficha básica promocional, así como también con los usuarios en estado activo.

Lo otro que se observan son variables con signos de pregunta y esto significa que su desviación estándar es cero.

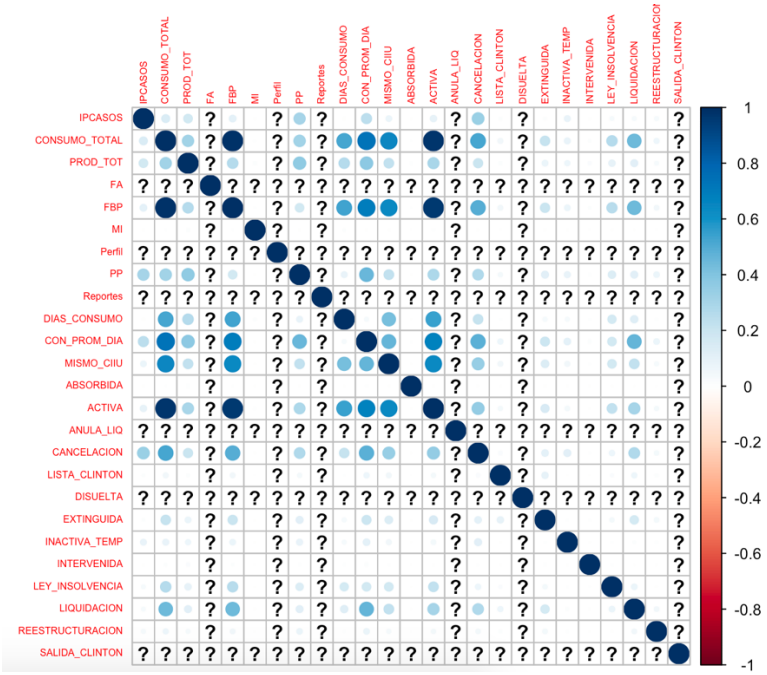


Figura 10: Correlación entre variables numéricas

Para eliminar la multicolinealidad hay dos opciones, eliminar las variables con una correlación mayor al 0.9, o bien, darles tratamiento. En este caso, se eligió la segunda opción.

Las variables de grupos de productos y los estados de los usuarios se hicieron calculando en términos absolutos el consumo realizado, pero una forma diferente de ver esto podría ser hacerlo en términos relativos, de esta forma, no solo se les estaría otorgando el peso relativo a cada una, sino que también normaliza las variables al hacer que todas estén en valores entre 0 y 1.

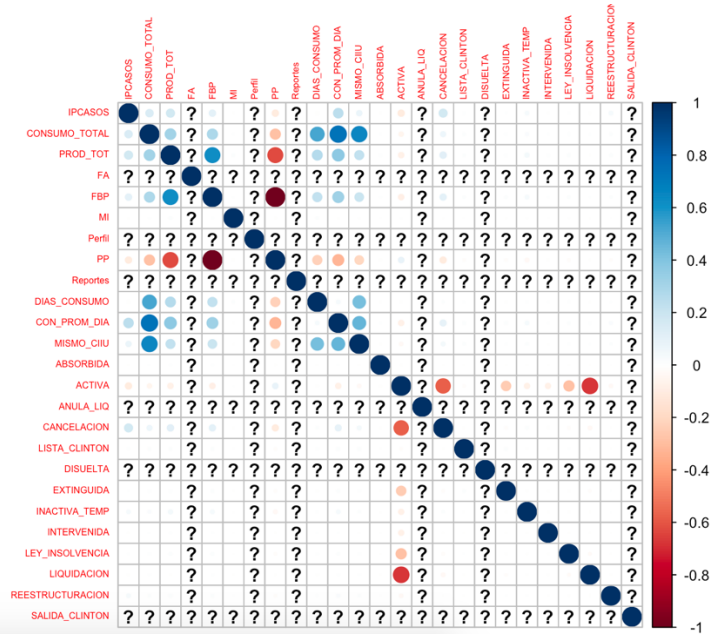


Figura 11: Correlación entre variables numéricas después de relativazarlas

Al aplicar esta técnica, la figura 11 muestra como se eliminó la fuerte correlación que existía en las variables antes mencionadas, dejando únicamente a las variables de ficha básica promocional y perfil promocional con una alta relación inversa.

Mediante este método, se decidió eliminar un total de 16 variables, 1 de ellas por su alta correlación inversa (Perfil Promocional), 6 debido a que su desviación estándar es 0, y finalmente, 9 porque no existe ninguna relación con las demás variables, sumado a que su porcentaje relativo en el consumo es casi nulo.

Para las variables categóricas se implementará la prueba de independencia de Chi-Cuadrado, la cual se utiliza para determinar si existe una relación significativa entre dos variables categóricas (nominales). Al ser esta una prueba de hipótesis, lo que se estaría contrastando sería:

H_0: no hay relación entre las variables
H_1: existe relación entre las variables

Como en cualquier prueba estadística, el resultado se contrasta con el valor p elegido (0.05). Si el valor p es significativo, podemos rechazar la hipótesis nula y afirmar que los hallazgos apoyan la hipótesis alternativa.

Al aplicar la prueba de chi cuadrado tanto con simulación Monte Carlo, como sin ella, nos dio el mismo resultado. A excepción de las variables de país y tipo de usuario, las demás sí tienen relación con el indicador de cliente.

Tabla 2: Resultados del test de chi cuadrado

Método	TIPOUSUARIO	CANAL_REGISTRO	IND_ALTA	TIPOEMAIL	BONDAD_EMAIL	USU_TIPO	USU_TAMANIO	USU_ESTADO	USU_CIIU_AREA	PAIS
Simulación MC	0.1794	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.2889
Sin Simulación	0.1976	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3444

Dicho esto, se procede a eliminar estas dos variables de la base de datos y así, dejar lista la base de datos para seguir con el siguiente paso.

La base de datos ha quedado con 20 variables, las cuales, se espera que sean las que ayuden a predecir qué usuarios se convertirán en clientes. Para lograr eso, primero hay que solucionar el problema de desbalanceo de los datos.

Esto puede ser solucionado con diferentes técnicas, las cuales se describirán 4 de las más usadas:

- Sobre muestreo: este método trabaja con la clase minoritaria. Lo que hace es replicar las observaciones de la clase minoritaria para equilibrar los datos.
- Sub muestreo: este método funciona con la clase mayoritaria. Reduce el número de observaciones de la clase mayoritaria para equilibrar el conjunto de datos.
- Ambos muestreos: este método es la combinación de los dos anteriores, reduciendo la clase mayoritaria y aumentando la minoritaria.
- ROSE: genera datos sintéticamente y proporciona una mejor estimación de los datos originales.

Como no se sabe bien cuál de los 4 métodos se ajusta mejor al set datos con el que se está trabajando, se decidió usarlos todos y posteriormente evaluar los resultados.

Mediante el algoritmo de árboles de decisiones se hizo la predicción de los clientes potenciales, el cuál, mostró buenos resultados. Se obtuvieron modelos con una precisión (accuracy) de 99,6% y con una predicción de posibles clientes (Neg Pred Value) en un 81% de los casos.

Tabla 3: Resultados del algoritmo de árboles de decisions en grupo 1

Sampling Method	Accuracy	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	AUC
Rose	0.9966	0.9965	1	1	0.8135	0.998
Under	0.7136	0.7133	0.73518	0.99436	0.03767	0.724
Over	0.9966	0.9965	1	1	0.8135	0.998
Both	0.9966	0.9965	1	1	0.8135	0.998

La tabla 3 muestra que elegir el método “rose”, “over” o “both” genera los mismos resultados, sin embargo, el método “under” es el que se ve más perjudicado.

Para buscar una mejora en alguno de los métodos, se aplicó adicionalmente el método de selección de variables llamado Boruta. A pesar de que previamente se habían elegido características por medio de las correlaciones y la prueba de chi-cuadrado, no se había realizado un método adicional para confirmar que esas era las variables con mayor importancia, por eso, se aplicó este otro método también.

Boruta es un algoritmo de selección de características. Precisamente, funciona como un algoritmo de envoltura (wrapper) alrededor de los arboles de decisiones. Este método tiene la ventaja de que particularmente busca comprender los mecanismos relacionados con la variable de interés, en lugar de simplemente construir un modelo predictivo de caja negra con buena precisión de predicción.

Tras ejecutar el algoritmo, este recomendó eliminar 5 variables: USU_TIPO, USU_ESTADO, USU_DEPARTAMENTO, USU_CIIU_AREA, dejando la base con un total de 16 variables, 14 si excluimos el ID_USUARIO e IND_CLIENTE.

Posteriormente, se volvieron a correr los modelos de arboles de decisiones, con la sorpresa de que, en lugar de encontrar diferencias entre los métodos de muestreo, ahora todos daban los mismos resultados vistos anteriormente. La única diferencia fue que el método “under” ahora lograba hacer bien las predicciones. Con esta disyuntiva, se tomó la decisión de seleccionar el método “rose”.

- **Grupo 2: Información incompleta**

Con las lecciones aprendidas de lo ejecutado en el grupo 1, el primer paso es dividir al grupo dos en la base de entrenamiento y de prueba.

En este caso, encontramos que en ambos conjuntos el porcentaje de clientes es de 0,05%. Un porcentaje sumamente bajo al cual, definitivamente habrá que aplicarles las técnicas de muestreo para buscar balancearlas.

En esta ocasión y sin necesidad del método de Boruta, se eliminan las variables USU_TIPO, USU_TAMANIO, USU_ESTADO, USU_DEPARTAMENTO, USU_CIIU_AREA y MISMO_CIIU debido a que el 100% de sus valores son nulos.

Como Boruta es un método bastante robusto, en el grupo 2 se ejecutó también, pero sin previamente eliminar variables por medio de la correlación de Pearson y la prueba Chi-Cuadrado. Esta vez, se eliminaron 27 variables y quedaron 11 variables de las cuales, solo 9 se utilizarán para la predicción.

Posterior a la reducción de variables, el otro paso realizado fue utilizar los 4 tipos de muestreo y compararlos nuevamente para evaluar y elegir el mejor. El resultado fue similar al del grupo 1, donde “under” fue el peor método.

Tabla 4: Resultados del algoritmo de árboles de decisiones en grupo 2

Sampling Method	Accuracy	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	AUC
Rose	0.9974	0.9974	1	1	0.6499	0.999
Under	0.675	0.6765	0.369	0.9955	0.0054	0.523
Over	0.9974	0.9974	1	1	0.6499	0.999
Both	0.9974	0.9974	1	1	0.6499	0.999

3.4.2 Algoritmos: generación de modelos

En esta sección, se utilizarán los sets de entrenamiento y pruebas elaborados anteriormente. Con ellos, se buscará encontrar qué algoritmo logra predecir de mejor manera a los potenciales compradores.

Entre los algoritmos que se pondrán a prueba están: bosques aleatorios, árboles de decisiones y regresión logística.

La elección de estos algoritmos se debe a que fueron los utilizados por Adil Mahmud y Kamruddin Nur en la investigación planteada al inicio del trabajo. De igual manera Xi Niu, Chugin Li y Xing Yu usaron también dos de esos algoritmos.

3.4.2.1 Arboles de decisiones en grupo 1

Un árbol de decisiones es un algoritmo de aprendizaje automático supervisado que parece un árbol invertido, en el que cada nodo representa una variable predictora (característica), el vínculo entre los nodos representa una decisión y cada nodo hoja representa un resultado (variable de respuesta).

El árbol de decisiones se considera uno de los algoritmos de aprendizaje automático más útiles, ya que se puede utilizar para resolver una variedad de problemas.

A continuación, se muestran algunas razones por las que debería utilizar el árbol de decisiones:

- Se considera el algoritmo de aprendizaje automático más comprensible y se puede interpretar fácilmente.
- Se puede utilizar para problemas de clasificación y regresión.
- A diferencia de la mayoría de los algoritmos de aprendizaje automático, funciona eficazmente con datos no lineales.

- La construcción de un árbol de decisión es un proceso muy rápido, ya que utiliza solo una función por nodo para dividir los datos.

Con esta introducción, entramos en lo que es el modelo donde se hicieron varias corridas utilizando la base de datos con la técnica de balanceo 'ROSE' y ajustando los siguientes parámetros:

- minsplit: establece el número mínimo de observaciones en el nodo antes de que el algoritmo realice una división
- maxdepth: establece la profundidad máxima de cualquier nodo del árbol final.

Probando diferentes combinaciones, se corrieron alrededor de unos 20 modelos con parámetros diferentes, de los cuales, a continuación, se muestran cinco de los modelos más relevantes.

Modelos	Min Split	Max Depth	AUC	Accuracy	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
Modelo 1	5	8	0.967	0.9957	0.9966	0.9368	0.999	0.8061
Modelo 2	9	10	0.955	0.9956	0.9969	0.913	0.9987	0.8163
Modelo 3	5	6	0.967	0.9957	0.9966	0.9368	0.999	0.8061
Modelo 4	5	5	0.999	0.9966	0.9965	1.0000	1.0000	0.8135
Modelo 5	5	4	0.998	0.9966	0.9965	1.0000	1.0000	0.8135

Figura 12: Comparación de modelos de árboles de decisiones

La figura 12 muestra que los mejores modelos fueron el 4to y 5to, donde la única diferencia que se hizo al ajustarlos fue en la profundidad máxima. A pesar de que cualquiera de los dos modelos podría escogerse, se eligió el modelo 4 debido a que tuvo un valor mayor en el área sobre la curva (0.999).

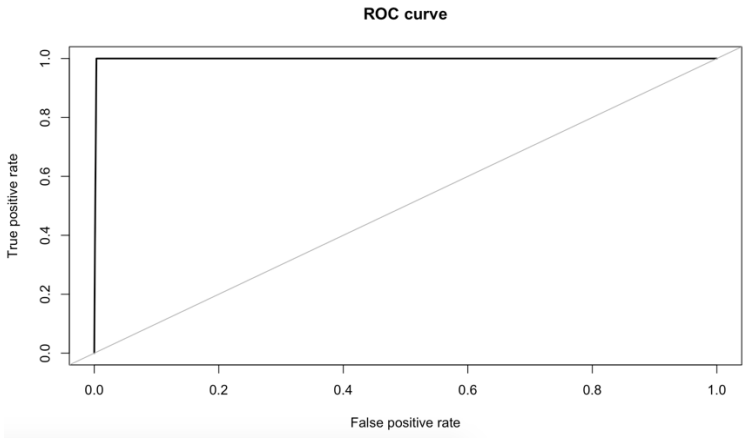


Figura 13: Gráfico de la curva ROC del modelo 4

En términos generales los resultados son buenos. Para este caso en particular nuestro interés no es solo conseguir una precisión alta, sino también, que los valores negativos predichos también sean altos ya que eso nos indica qué tan bien predice el modelo a los posibles clientes.

La figura 14, muestra que el modelo 4 tuvo una precisión global del 99,7%, mientras que la sensibilidad (predecir correctamente los no clientes) y especificidad (predecir correctamente a los clientes) fueron del 99,7% y 100% respectivamente.

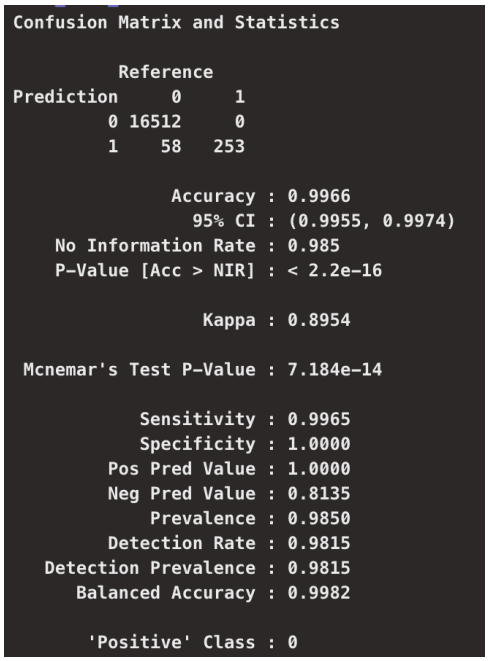


Figura 14: Resultados del modelo de árboles de decisiones en el grupo 1

3.4.2.2 Random Forest en grupo 1

Este algoritmo se basa en generar una gran cantidad de árboles de decisión, cada uno construido utilizando un subconjunto diferente de su conjunto de entrenamiento.

Estos subconjuntos generalmente se seleccionan mediante muestreo al azar y con reemplazo del conjunto de datos original. Luego, los árboles de decisión se utilizan para identificar un consenso de clasificación seleccionando la salida (modo) más común. Si bien los bosques aleatorios se pueden usar para otras aplicaciones (es decir, regresión), para este caso en particular, se utilizará para la clasificación.

Al igual que el árbol de decisiones, este modelo puede ser ajustado por diferentes parámetros, pero antes de iniciar a buscar la mejor combinación, el primer ejercicio fue correrlo con los valores preestablecidos y haciendo varios

Existen distintos parámetros con los que se puede jugar, pero en este trabajo me centraré en dos parámetros de ajuste: mtry y ntree. Esto debido a que son los que usualmente tienen un mayor efecto en la precisión del modelo.

- mtry: es el número de variables que se recopila aleatoriamente para muestrear en cada tiempo fraccionado.
- ntree: es el número de ramas crecerá después de cada división de tiempo.

Estos 2 parámetros fueron probados con unas 15 combinaciones diferentes, incluso, se ejecutó un algoritmo que buscara los parámetros ideales basados en una validación cruzada, con el objetivo de obtener una precisión mayor.

La figura 15 muestra el resumen de 4 de los modelos que se corrieron.

Modelos	mtry	ntree	AUC	Accuracy	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
Modelo 1	3	500	0.9986	0.9964	0.9964	1.0000	1.0000	0.6463
Modelo 2	14	500	0.9982	0.9961	0.9964	0.9762	0.9996	0.8092
Modelo 3	14	1000	0.9982	0.9961	0.9964	0.9762	0.9996	0.8092
Modelo 4	4	1000	0.9989	0.9965	0.9964	1.0000	1.0000	0.8129

Figura 15: Comparación de modelos de bosques aleatorios

Algo que me llamó poderosamente la atención es que a pesar de que el algoritmo que busca el mejor valor de mtry para lograr la mayor precisión indicó que este debía de ser 14, los mejores resultados se dieron con un mtry menor.

Debido a ello, se hizo una validación cruzada con un mtry=4 y ntree=1000, donde se rectificó que estos valores lograban las mejores métricas. Como resultado, se eligió el modelo 4 como el mejor entre los de bosques aleatorios.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 16270     0
         1    58   252

               Accuracy : 0.9965
                 95% CI : (0.9955, 0.9973)
    No Information Rate : 0.9848
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.895

 Mcnemar's Test P-Value : 7.184e-14

            Sensitivity : 0.9964
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8129
             Prevalence : 0.9848
         Detection Rate : 0.9813
   Detection Prevalence : 0.9813
      Balanced Accuracy : 0.9982

       'Positive' Class : 0
```

Figura 16: Resultados del modelo de bosques aleatorios en el grupo 1

Los resultados indican que se obtuvo una precisión global del 99,65%, una especificidad del 100% y una predicción de los valores negativos de 81,3%.

3.4.2.3 Regresión Logística grupo 1

La regresión logística se utiliza para predecir la clase (o categoría) de individuos en función de una o varias variables predictoras. Se utiliza para modelar un resultado binario, es decir, una variable, que solo puede tener dos valores posibles: 0 o 1, que en nuestro caso son los que no serán clientes y los que sí.

La regresión logística pertenece a una familia, denominada Modelo lineal generalizado (GLM), desarrollada para extender el modelo de regresión lineal a otras situaciones.

Un aspecto importante es que la regresión logística no devuelve directamente la clase de observaciones, si no mas bien la probabilidad (p) estimada de pertenencia a una clase. Como la probabilidad varia entre 0 y 1 se decidió utilizar como umbral que p = 0,5. Eso significa que todo lo que sea igual o mayor a ese valor se etiquetó como 1, mientras que lo menor fue etiquetado como 0.

Debido a que la regresión logística no puede lidiar con las variables categóricas, estas tuvieron que ser recodificadas para poder hacer uso de ellas en el modelo. Hacer eso, trae consigo pros y contras, pero de alguna forma hay que hacerlo para poder hacer uso de las variables

En este proyecto, se utilizó la técnica de '*label encoding*', el cual implica convertir cada valor de una columna a un número. Por ejemplo, en el tamaño de la empresa 1= GR, 2=MD, 3=PQ, 4=MC y 5=SD.

Al hacer esto, se tiene la ventaja de que la variable puede ser utilizada como un predictor, la desventaja, es que el algoritmo tome los valores y crea que los más altos tienen mayor peso.

Otra posible solución era la llamada 'One-Hot Encoding', que por cada etiqueta crea una nueva variable dicotómica, es decir, en el caso del tamaño de la empresa, se hubiesen creado 5 variables binarias donde el valor de 1 indicaría que es de ese tipo y 0 de que no lo es. La razón por la cual no se eligió este método se debe a que, al ser tantos niveles, es posible que haya multicolinealidad entre estas nuevas variables.

Una vez realizada la codificación de etiquetas tanto en el archivo de entrenamiento como en el de pruebas, se prosiguió a correr el modelo de regresión logística, el cual, logró un AUC =

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 16271     1
         1    57   251

               Accuracy : 0.9965
                 95% CI : (0.9955, 0.9973)
    No Information Rate : 0.9848
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8947

 Mcnemar's Test P-Value : 5.128e-13

            Sensitivity : 0.9965
            Specificity : 0.9960
         Pos Pred Value : 0.9999
         Neg Pred Value : 0.8149
             Prevalence : 0.9848
         Detection Rate : 0.9814
   Detection Prevalence : 0.9814
      Balanced Accuracy : 0.9963

       'Positive' Class : 0
```

Figura 17: Resultados del modelo de la regresión logística en el grupo 1

Los resultados parecieran ser buenos y que el modelo logra clasificar correctamente a los clientes y a los no clientes. La regresión logística consiguió una precisión global del 99,65%, una especificidad del 99,6% y una predicción de valores negativos del 81,49%.

3.4.2.4 Arboles de decisiones en grupo 2

Al igual que con el grupo 1, se buscó ajustar los parámetros ‘minsplit’ y ‘maxdepth’, donde se encontró que, con los valores de 3 y 2 respectivamente, se lograban los mejores resultados.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 56301     0
         1   146   271

               Accuracy : 0.9974
                 95% CI : (0.997, 0.9978)
    No Information Rate : 0.9952
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7866

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9974
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.6499
             Prevalence : 0.9952
         Detection Rate : 0.9926
   Detection Prevalence : 0.9926
      Balanced Accuracy : 0.9987

       'Positive' Class : 0
```

Figura 18: Resultados del modelo de árboles de decisiones en el grupo 2

Los árboles de decisiones alcanzaron un 99,74% de precisión global, una especificidad del 100% y una predicción correcta de clientes en un 65%. Así mismo, este modelo logró un AUC del 0.999

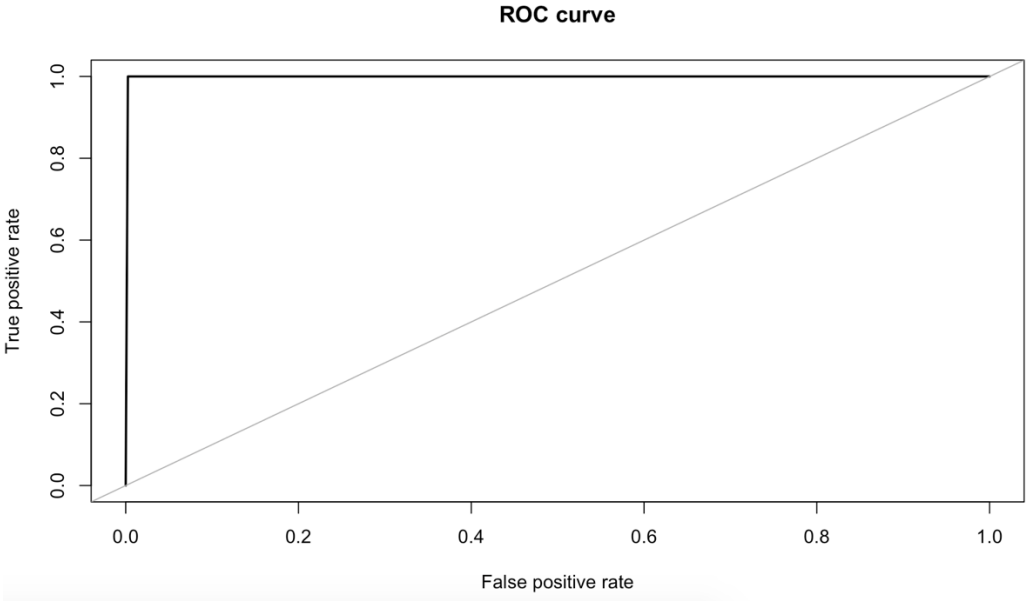


Figura 19: Gráfico de curva ROC del árbol de decisiones en grupo 2

3.4.2.5 Bosques aleatorios en grupo 2

Para ajustar este modelo, se hicieron varias combinaciones en los parámetros mtry y ntree, sin embargo, con todos se logró el mismo resultado, es decir, sin importar qué valores se utilizaron nada cambió. Usando una validación cruzada de 10 repeticiones, el que dio ligeramente más alto fue utilizando un mtry=3 y ntree =250, por lo que me quedé con ese modelo finalmente.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 55050     2
         1   141   128

               Accuracy : 0.9974
                 95% CI : (0.997, 0.9978)
    No Information Rate : 0.9977
    P-Value [Acc > NIR] : 0.881

                  Kappa : 0.6405

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9974
            Specificity : 0.9846
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.4758
             Prevalence : 0.9977
         Detection Rate : 0.9951
   Detection Prevalence : 0.9951
      Balanced Accuracy : 0.9910

       'Positive' Class : 0
```

Figura 20: Resultados del modelos de bosques aleatorios en el grupo 2

La figura 20 muestra que la precisión global fue de un 99,74% al igual que la sensibilidad, mientras que la especificidad fue del 98,46%. Como se ha mencionado durante el proyecto de investigación. El porcentaje de falsos negativos fue de un 47,58%, es decir, los clientes que fueron predichos como posibles clientes logró predecir a menos de la mitad, a pesar de que su AUC fuese del 0.9988

3.4.2.6 Regresión logística en grupo 2

Finalmente se realizó la regresión logística. En el grupo 2 no fue necesario recodificar ninguna variable, ya que, si recordamos, casi que todas las variables categóricas fueron descartadas debido a que únicamente contenían valores nulos, por lo que era imposible hacer uso de ellos.

Al ejecutar el algoritmo, las únicas variables significativas, es decir, que aportan información para predecir si un usuario es un potencial comprador son: canal de registro con valores de 2, 3 y 8; bondad de email =20, la cantidad de usuarios utilizando la misma IP, la cantidad de productos distintos consumidos y el consumo promedio por día.

```
Call:
glm(formula = IND_CLIENTE ~ ., family = "binomial", data = rose_g2[,
    2:11], na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1090   0.0000   0.0187   0.0501   0.4977

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.320e+01  2.025e+02  -0.115   0.9088
CANAL_REGISTRO2 -8.138e-01  2.042e-01  -3.986 6.71e-05 ***
CANAL_REGISTRO3 -8.828e-01  1.963e-01  -4.497 6.90e-06 ***
CANAL_REGISTRO4 -2.959e-01  2.614e-01  -1.132   0.2576
CANAL_REGISTRO6 -1.619e-01  3.375e-01  -0.480   0.6314
CANAL_REGISTRO7 -2.212e-01  2.251e-01  -0.983   0.3257
CANAL_REGISTRO8 -6.909e-01  2.461e-01  -2.807   0.0050 **
CANAL_REGISTRO9 -4.961e-01  4.153e-01  -1.194   0.2323
IND_ALTA1        2.943e+01  2.025e+02   0.145   0.8844
BONDAD_EMAIL-10 -4.278e-01  6.152e-01  -0.695   0.4868
BONDAD_EMAIL0    3.006e-01  4.584e-01   0.656   0.5119
BONDAD_EMAIL1    9.966e-01  5.370e-01   1.856   0.0634 .
BONDAD_EMAIL9    1.499e+00  1.056e+00   1.419   0.1559
BONDAD_EMAIL20   1.514e+00  3.542e-01   4.275 1.91e-05 ***
IPCASOS         -2.035e-03  3.338e-04  -6.095 1.09e-09 ***
PROD_TOT        -9.503e-01  7.756e-02 -12.251  < 2e-16 ***
FBP              2.225e-02  1.862e-02   1.195   0.2321
PP               6.094e-02  6.372e-02   0.956   0.3389
CON_PROM_DIA    -1.445e-01  2.514e-02  -5.747 9.08e-09 ***
ACTIVA          -1.023e-02  2.261e-02  -0.452   0.6509
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 307639.0  on 221914  degrees of freedom
Residual deviance:   4037.4  on 221895  degrees of freedom
AIC: 4077.4

Number of Fisher Scoring iterations: 22
```

Figura 21: Resumen de la regresión logística en grupo 2

Al descartar las demás variables y correr el modelo únicamente con esas variables, se logró obtener un valor de AUC de 0.9987

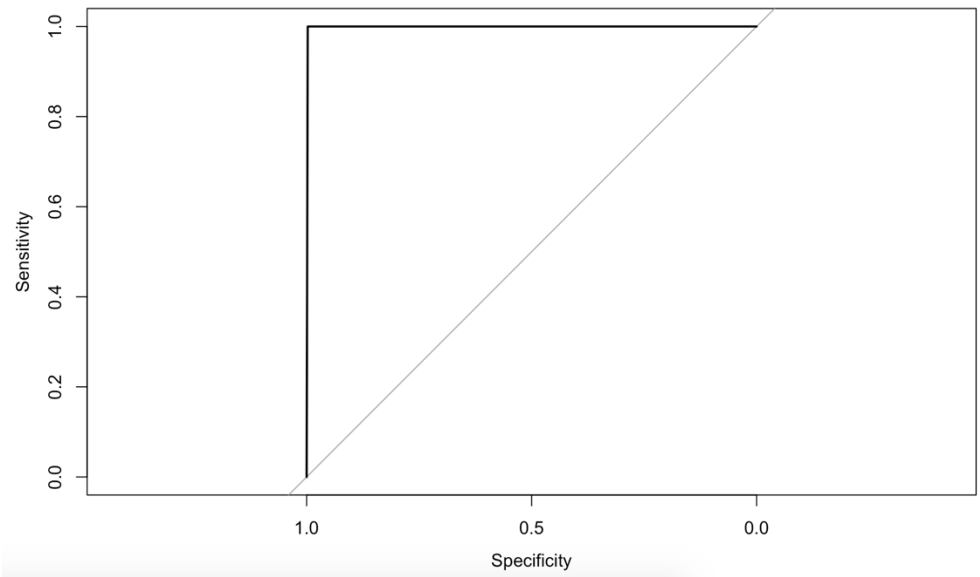


Figura 22: Gráfico de curva ROC de la regresión logística en grupo 2

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 55047     0
         1   144   130

               Accuracy : 0.9974
                 95% CI : (0.9969, 0.9978)
    No Information Rate : 0.9977
    P-Value [Acc > NIR] : 0.8971

                  Kappa : 0.6424

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9974
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.4745
             Prevalence : 0.9977
         Detection Rate : 0.9950
   Detection Prevalence : 0.9950
      Balanced Accuracy : 0.9987

       'Positive' Class : 0
```

Figura 23: Resultados del modelos de la regresión logística en el grupo 2

A pesar de que a simple vista parece un modelo que ajusta y predice bien a los compradores potenciales, en realidad no lo es, ya que el porcentaje de clientes que fueron predichos correctamente fue de un 47,45%.

3.4.3 Comparación de modelos y elección del mejor

Como se ha visto en las secciones anteriores, cada algoritmo fue entrenado, ajustado y puesto a prueba para poder elegir los parámetros que mejor lograran predecir a los potenciales compradores.

La siguiente tabla, muestra un resumen de las principales métricas e información sobre cada algoritmo con el fin de poder elegir uno de ellos para una posible implementación.

Tabla 5: Comparación de modelos aplicados al grupo 1

Algoritmo	AUC	Accuracy	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
Árboles de decisiones	0.9990	0.9966	0.9965	1.0000	1.0000	0.8135
Bosques Aleatorios	0.9989	0.9965	0.9964	1.0000	1.0000	0.8129
Regresión Logística	0.9988	0.9965	0.9965	0.9960	0.9999	0.8149

Para evaluar los modelos, deben de analizarse distintos puntos. En primer lugar, se debe analizar el área bajo la curva (AUC: '*Area Under the Curve'*), que mide la capacidad de un clasificador para distinguir entre clases.

El AUC puede tener valores entre 0 y 1, siendo 1 el indicador de que logra distinguir correctamente las dos clases, mientras que 0, sería lo contrario. Entonces, cuanto mayor sea el AUC, mejor será el rendimiento del modelo para distinguir entre las clases positivas y negativas.

Observando la tabla 5, los valores del AUC son prácticamente iguales en los 3 modelos, por lo que es difícil tomar una decisión basándonos únicamente en esa medida.

Lo segundo en examinarse sería la precisión global del modelo (*accuracy*), donde nuevamente los valores son casi que los mismos y se podría decir que no hay diferencias entre ellos.

Ahora bien, debido a que el interés específico de esta investigación es predecir a los usuarios que podrían ser potenciales compradores, nos enfocaremos en dos parámetros adicionales. La especificidad que mide la proporción de clientes que se identificaron correctamente, es decir, que los que hayan pagado se hayan clasificado como clientes. En este apartado, los árboles de decisiones y los bosques aleatorios lograron un 100%.

El otro aspecto para considerar en este proyecto es la proporción de clientes que fue predicho como cliente y que realmente lo haya sido. En esta categoría, el mejor resultado lo obtuvo la regresión logística con un valor del 81,49%.

Los tres modelos son casi que igual de precisos, sus resultados son muy similares y decir que uno es mejor que otro sería irreal. Basado en los análisis presentados, cualquiera de los tres algoritmos podría ser implementado, no obstante, para este proyecto elegiré los árboles de decisiones debido a que es el más sencillo de explicar.

Cuando se explicó este algoritmo, se mencionó que este parece un árbol invertido y que cada uno de sus nodos indican una decisión, la cual, determina si va a clasificar a cada usuario en un potencial comprador o no. Las reglas de decisiones creadas en este modelo se muestran en la siguiente figura:

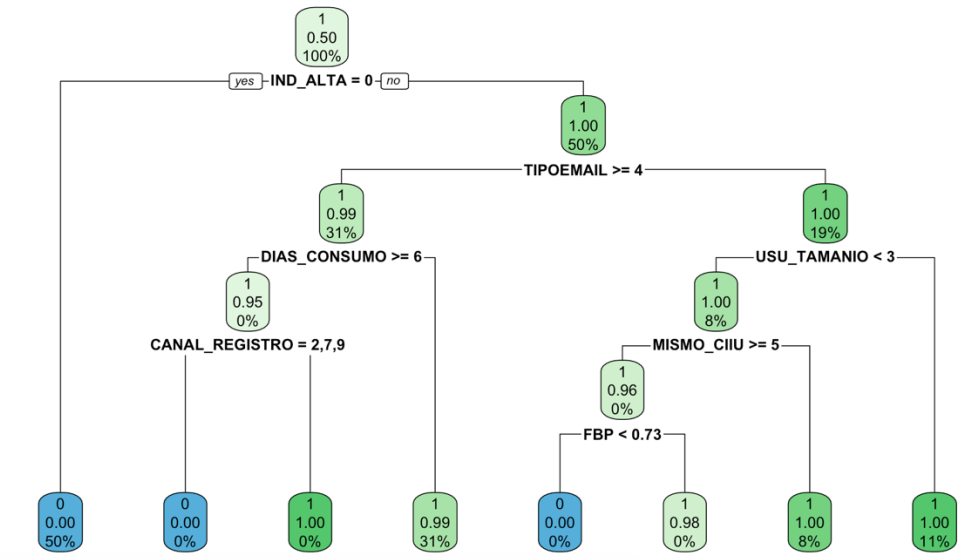


Figura 24: Árbol de decisiones del grupo 1

Este gráfico muestra que la primera decisión es si un cliente fue al botón de pago, si este lo hizo, la probabilidad es muy alta de que sí vaya a ser cliente, caso contrario, no pagará por los servicios.

Las únicas dos condiciones donde, aunque el usuario haya ido al botón de pago este no sea etiquetado como potencial comprador, es en el caso de: a) Que el valor normalizado del tipo de email sea menor a 4, haya realizado consumos en 5 o menos días distintos y que el canal de registro sea distinto al 2, 7 y 9; b) Cuando la empresa es media o grande, han realizado pocos consumos en empresas que tienen el mismo código CIIU de ellos y que el consumo de la "ficha básica promocional" no exceda el 73% del total de consumo.

Con base en esas condiciones, es que este algoritmo logra clasificar a un potencial cliente o no en el grupo 1.

En el caso del grupo 2, también se evaluaron tres modelos, los cuales se presenta un resumen de ellos en la tabla 6

Tabla 6: Comparación de modelos aplicados al grupo 2

Algoritmo	AUC	Accuracy	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
Árboles de decisiones	0.9990	0.9974	0.9974	1.0000	1.0000	0.6499
Bosques Aleatorios	0.9987	0.9974	0.9974	0.9846	1.0000	0.4758
Regresión Logística	0.9987	0.9974	0.9974	1.0000	1.0000	0.4745

Estos resultados nos llevan casi que a la misma conjetura enfrentada con el grupo 1, donde casi que todas sus métricas son iguales y las que no, son muy similares.

Empero, hay una de ellas que sí sobresale y es el '*Neg Pred Value*', el cual, indicamos anteriormente, que se refiere a la proporción de usuarios predichos como potenciales compradores que realmente eran clientes y, sin duda alguna, los arboles de decisiones obtuvieron el mejor puntaje con un 65% aproximadamente. A pesar de que este no es un valor alto y sus predicciones no son tan precisas, definitivamente es el mejor comparado con los otros dos modelos y por ende, fue elegido como el modelo a utilizarse.

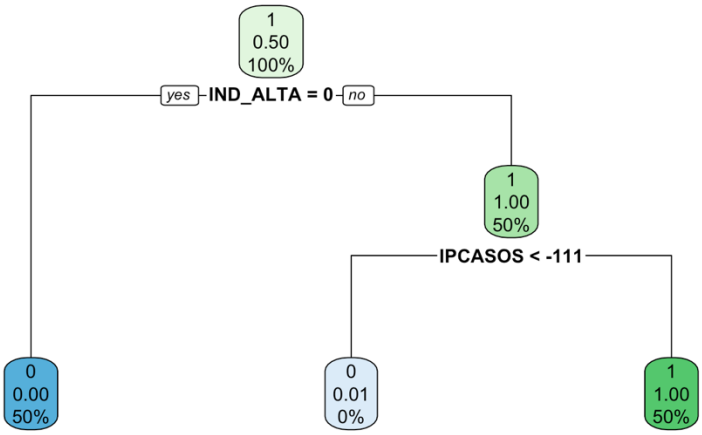


Figura 25: Árbol de decisiones del grupo 2

En el grupo 2 la toma de decisiones fue mucho más sencilla ya que el árbol tiene únicamente dos nodos. El primero separa a aquellos que fueron o no darse de alta, si el usuario va, será catalogado como un potencial comprador, caso contrario, no será un futuro cliente.

El segundo nodo indica que si el usuario tiene un valor “normalizado” mayor a -111, quiere decir que no será un cliente.

4. Conclusiones y líneas a futuro

Para concluir, en esta sección se presenta un resumen con los principales resultados, dando así respuesta a los objetivos planteados inicialmente. Consecutivamente, se expondrán algunas ideas sobre futuros análisis que se podrían realizar con el propósito de mejorar lo realizado en el presente proyecto.

4.1 Conclusiones

El objetivo principal de este proyecto es el de crear uno o varios modelos de clasificación de leads donde se logre identificar una posible selección de potenciales compradores.

Para llevar a cabo este, se utilizaron dos bases de datos, una de usuarios y otra de consumos, las cuales, debían de ser unidas para poder lograr el objetivo principal.

Antes de unir ambas bases, se realizó una limpieza de los datos en la cual se incluyó la reclasificación correcta de las variables, imputación de valores nulos, recodificación de variables, entre otros.

Ya con los datos listos, se procedió al EDA, el cual mostró que el 0,71% de los usuarios registrados eran clientes, es decir, que más del 99% de sus usuarios únicamente consumían productos de no pago durante el periodo de promoción.

Durante ese lapso de consumo gratuito que tienen, los productos más consumidos fueron los perfiles promocionales con un 54% y la ficha básica promocional con un 46%. No obstante, aquellos usuarios que se convirtieron en clientes, previo a su conversión casi que lo único que consumían era la ficha básica promocional.

El asunto con los consumos es que esa base de datos tiene múltiples registros (consumos) por cada usuario, pero para poder unirla con el archivo de clientes, se debía de tener resumida en una sola fila por usuario.

Inicialmente la base de consumo que tenía 11 variables y estas fueron transformadas en un nuevo *dataset* con 25 variables, intentando así resumir toda la información. De esta forma, el nuevo conjunto de datos fue unido al de usuarios por medio de la variable ID_USUARIO, dejando así un conjunto de datos con 44 variables.

El problema encontrado en este proyecto fue que muchas de las características de los usuarios eran nulas, es decir, no se conocían sus valores, por lo que se tomó la

decisión de separar la base en dos grupos, los que sí habían completado esa información y los que no.

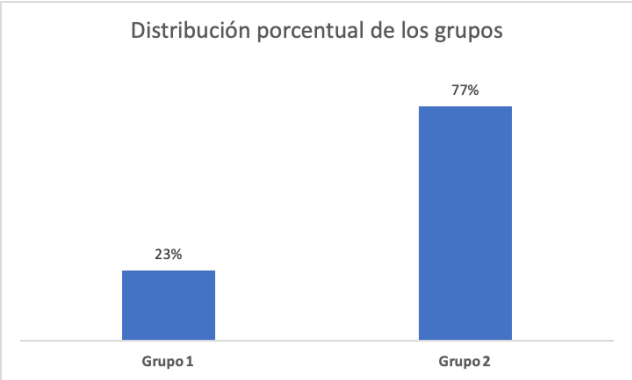


Figura 26: Distribución de los grupos de análisis

Con esta separación, el proyecto tomó dos caminos independientes, donde cada decisión fue basada en las características propias de cada grupo.

En el grupo 1, luego de realizar un análisis de multicolinearidad para eliminar las variables altamente correlacionadas; pruebas de chi cuadrado para borrar las características que no tuvieran relación con la variable a predecir y, el algoritmo de Boruta para elegir a las variables con mayor impacto en la predicción de potenciales compradores, se determinó que, de las 44 variables iniciales, solo 14 de ellas ayudarían realmente a construir modelos de *machine learning*.

Las 14 variables con las que se trabajó fueron:

- Índice de alta
- Canal de registro
- Tipo de email
- Bondad del email
- IP Casos
- Tamaño del usuario
- Consumo Total
- Productos distintos totales
- Porcentaje de consumo de la ficha básica promocional
- Cantidad de días en que hubo consumo
- Consumo promedio por día
- Consumos de empresas con el mismo CIIU
- Porcentaje de consumo de empresas en estado activa
- Porcentaje de consumo de empresas en estado de cancelación

Debido a que en el pasado otros investigadores habían utilizado los algoritmos de árboles de decisiones, bosques aleatorios y regresión logística, se tomó la decisión de utilizar los mismos.

Estos tres modelos resultaron ser igual de precisos, por lo que elegir cualquiera de ellos daría los mismos resultados, sin embargo, se eligieron los árboles de decisiones debido a que son los más sencillos de explicar. Este modelo, tuvo un valor de AUC del 0.999 y una precisión global del 99,7%. Las otras dos métricas de mayor interés eran la especificidad, ya que es la proporción de clientes clasificados correctamente y en la cual se obtuvo un 100%; y el 'Neg Pred Value', que identifica el porcentaje de clientes predichos como clientes y que realmente lo eran. En esta métrica se obtuvo un 81,35%.

Debido a lo aprendido durante lo trabajado en el grupo 1, se decidió no hacer el análisis de multicolinearidad ni las pruebas de chi cuadrado, sino que, únicamente se utilizó el algoritmo de Boruta para la selección de las variables.

Esta técnica, dejó una base con tan solo 9 variables, lo cual, era de esperarse que fuera menor a la del grupo 1 debido a la ausencia de información en muchas de sus características.

A continuación, se presentan las variables finales para el grupo 2:

- Índice de alta
- Canal registro
- Bondad del email
- IP Casos
- Productos distintos totales
- Consumo promedio por día
- Porcentaje de consumo de la ficha básica promocional
- Porcentaje de consumo del perfil promocional
- Porcentaje de consumo de empresas en estado activa.

Utilizando estas características, se corrieron los 3 modelos de aprendizaje automático, y aunque todos mostraron valores similares, definitivamente los árboles de decisiones esta vez sí fueron mejores que los otros.

Este modelo, alcanzó un AUC de 0.990, una precisión global del 99,74%, una especificidad del 100% y finalmente, un 65% de las predicciones sobre los potenciales clientes fue acertada. Aunque este no es un valor tan alto, definitivamente fue el mejor.

Es importante mencionar que ambos modelos lograron clasificar correctamente a los clientes, y donde fallaron, fue en indicar que algunos usuarios se iban a convertir y no lo hicieron, es decir, se comete el error tipo II.

Creo que cometer este tipo de error no genera ningún problema para la empresa, debido a que estaría apostando por posibles clientes que al fin y al cabo no se convirtieron, pero, sería peor dejar de lado a futuros clientes debido a que fueron predichos como que no serían posibles leads.

4.2 Líneas a futuro

Un sinsabor que me deja este proyecto es ver la importancia de las variables al momento de hacer la predicción y es que, si se observan las figuras 27 y 28, muestran que la característica de mayor peso en la predicción es si el cliente fue o no al botón de pago.

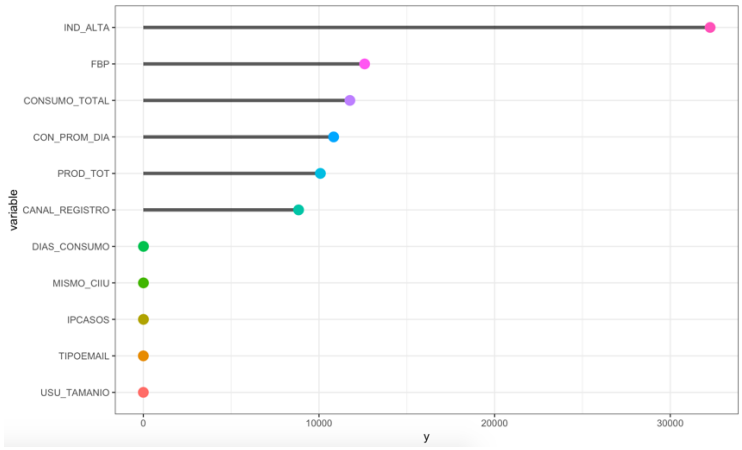


Figura 27: Importancia de las variables en el grupo 1

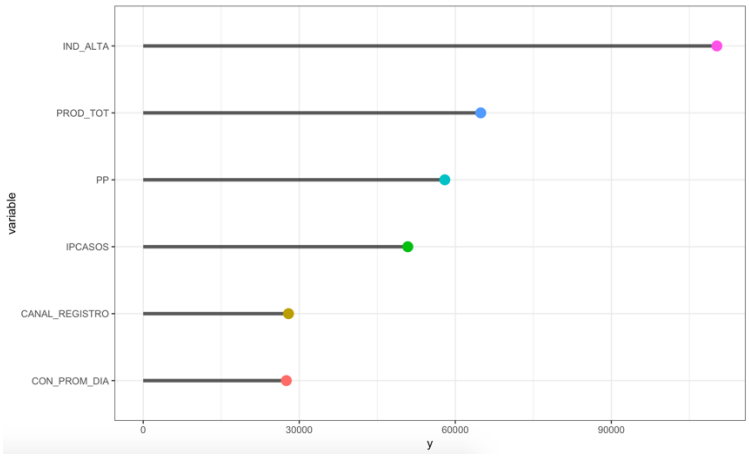


Figura 28: Importancia de las variables en el grupo 2

Esto tiene todo el sentido, y es porque el 98% de los usuarios que fueron al botón de pago se convirtieron en clientes.

Dicho esto, el dueño de la empresa podría decir que él no ocupa ningún algoritmo para predecir quién será o no un potencial comprador, con solo saber si es redirigido a hacer el pago, ya sabe si será o no un lead.

Por esa razón, creo que lo ideal sería crear un modelo conjunto, donde primero se identifique al usuario que potencialmente irá al botón de pago y posteriormente, cuales de ellos finalmente se convertirán en clientes.

5. Anexos

1- Variables seleccionadas para el grupo 1 mediante el algoritmo Boruta

	meanImp	medianImp	minImp	maxImp	normHits	decision
CANAL_REGISTRO	4.7177690	4.8206759	-0.06782531	9.263953	0.8571429	Confirmed
TIPOEMAIL	5.9599364	5.7817442	-2.01390237	13.151153	0.7619048	Confirmed
BONDAD_EMAIL	17.3976450	17.3499684	14.30204503	22.510436	1.0000000	Confirmed
IPCASOS	34.4811946	35.1920363	29.42157583	40.759930	1.0000000	Confirmed
USU_TIPO	-3.6880379	-3.5547878	-5.54630396	-1.744681	0.0000000	Rejected
USU_TAMANIO	7.2267568	7.3285100	2.14389799	13.129050	0.9523810	Confirmed
USU_ESTADO	-0.8935099	-0.7424443	-2.93146596	1.527334	0.0000000	Rejected
USU_DEPARTAMENTO	-1.3893847	-1.4337037	-3.22012865	1.750481	0.0000000	Rejected
USU_CIIU_AREA	1.3073600	1.3987806	-1.70600322	3.562341	0.0952381	Rejected
CONSUMO_TOTAL	12.3153975	12.3395722	7.78862293	15.524089	1.0000000	Confirmed
PROD_TOT	10.3990192	9.8765991	5.91593919	15.980543	1.0000000	Confirmed
FBP	37.8030117	37.6801702	33.10714890	42.210916	1.0000000	Confirmed
DIAS_CONSUMO	3.4701886	3.6814162	0.78963433	5.137774	0.7619048	Confirmed
CON_PROM_DIA	20.5473875	20.5638193	17.75737099	22.699572	1.0000000	Confirmed
MISMO_CIIU	6.3067125	6.3868262	4.71097807	8.012627	1.0000000	Confirmed
ACTIVA	6.2198288	6.1293437	3.01405157	10.948788	0.9761905	Confirmed
CANCELACION	4.5337229	4.6492582	1.97489895	8.719443	0.8333333	Confirmed

2- Variables seleccionadas para el grupo 2 mediante el algoritmo Boruta

	meanImp	medianImp	minImp	maxImp	normHits	decision
TIPOUSUARIO	1.7587895	1.9479657	-0.8971966	4.460655	0.04040404	Rejected
CANAL_REGISTRO	6.3125822	6.1454331	2.1051899	10.187057	0.84848485	Confirmed
IND_ALTA	104.9038284	109.2490770	69.2766665	130.318694	1.00000000	Confirmed
TIPOEMAIL	2.6509501	2.7106631	-1.9086878	6.464191	0.19191919	Rejected
BONDAD_EMAIL	7.0687391	6.8863342	3.4640355	11.636546	0.96969697	Confirmed
IPCASOS	8.3613059	8.4585559	6.2116817	10.440311	0.98989899	Confirmed
PAIS	2.8061897	2.8909215	-0.1164362	6.321944	0.25252525	Rejected
CONSUMO_TOTAL	3.1961613	2.9186199	1.9158922	6.435978	0.37373737	Rejected
PROD_TOT	6.1819156	6.1326536	3.7287063	8.267126	0.88888889	Confirmed
FA	0.0000000	0.0000000	0.0000000	0.000000	0.00000000	Rejected
FBP	5.5112552	5.5057766	3.4981244	7.809617	0.80808081	Confirmed
MI	0.0000000	0.0000000	0.0000000	0.000000	0.00000000	Rejected
Perfil	0.0000000	0.0000000	0.0000000	0.000000	0.00000000	Rejected
PP	4.6891559	4.7906406	1.6361352	6.341640	0.72727273	Confirmed
Reportes	0.0000000	0.0000000	0.0000000	0.000000	0.00000000	Rejected
DIAS_CONSUMO	3.2951564	3.3456259	-0.1080867	5.309317	0.44444444	Rejected
CON_PROM_DIA	4.8242664	4.5123909	2.6009010	8.797339	0.69696970	Confirmed
ABSORBIDA	0.0000000	0.0000000	0.0000000	0.000000	0.00000000	Rejected
ACTIVA	4.6670867	4.7590361	2.1320919	8.442506	0.68686869	Confirmed
ANULA_LIQ	0.0000000	0.0000000	0.0000000	0.000000	0.00000000	Rejected
CANCELACION	2.9964761	2.9860202	1.8369963	5.614620	0.26262626	Rejected
LISTA_CLINTON	0.0000000	0.0000000	0.0000000	0.000000	0.00000000	Rejected
DISUELTA	0.0000000	0.0000000	0.0000000	0.000000	0.00000000	Rejected
EXTINGUIDA	1.6420135	1.7688738	-0.9689651	3.358287	0.03030303	Rejected
INACTIVA_TEMP	0.7931105	0.9810983	-1.8961599	2.464319	0.01010101	Rejected
INTERVENIDA	-2.6473386	-2.8229366	-4.2322132	-1.023768	0.00000000	Rejected
LEY_INSOLVENCIA	1.4266254	1.3568869	0.1574338	3.474715	0.01010101	Rejected
LIQUIDACION	3.1919829	3.3012510	0.4739348	5.659766	0.39393939	Rejected
REESTRUCTURACION	0.8222549	1.0010015	-1.0010015	1.737270	0.00000000	Rejected
SALIDA_CLINTON	0.0000000	0.0000000	0.0000000	0.000000	0.00000000	Rejected

3- Enlace de GitHub donde se encuentra el código utilizado: https://github.com/jar2015/ecommerce_predict_buyers/blob/main/README.md

6. Bibliografía

Bhandari, Aniruddha. (2020). “AUC-ROC Curve in Machine Learning Clearly Explained”. Obtenido de: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Beck, Martin. (2019). “Can You Predict If a Customer Will Make a Purchase on a Website?”. Obtenido de: https://towardsdatascience.com/can-you-predict-if-a-customer-will-make-a-purchase-on-a-website-e6843ec264ae

Casas, Pablo (2018). “Exploratory Data Analysis in R (introduction”. Obtenido de: https://blog.datascienceheroes.com/exploratory-data-analysis-in-r-intro/

Chapman, Pete et. al. (2000). CRISP-DM 1.0 Step-by-step data mining guide. Obtenido de: https://the-modeling-agency.com/crisp-dm.pdf

Chawla, N. V. 2005. “Data mining for imbalanced datasets: An overview,” in Data mining and knowledge discovery handbook, Springer, pp. 853–867.

Choudhury, A. M. and Nur, K. (2019). “A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior”. *Conference: 2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST).* DOI: 10.1109/ICREST.2019.8644458. Obtenido de: https://www.researchgate.net/publication/331563946_A_Machine_Learning_Approach_to_Identify_Potential_Customer_Based_on_Purchase_Behavior

D. Xu, W. Yang, and L. Ma, (2018). “Repurchase Prediction Based on Ensemble Learning,” *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI),* pp. 1317–1322, Guangzhou, China, 2018. Obtenido en: https://ieeexplore.ieee.org/document/8560207/

da Silva, J. M. M. (2014). “The Road to Enlightenment: Generating Insight and Predicting Consumer Actions in Digital Markets,” Dissertation Research Submitted to University of Porto.

De los Santos, B., Hortaçsu, A., and Wildenbeest, M. R. (2015). “Search with learning for differentiated products: Evidence from e-commerce,” Obtenido de: https://host.kelley.iu.edu/mwildenb/learning.pdf

E-Marketer. (2014). “Global B2C Ecommerce Sales to Hit $1.5 Trillion This Year Driven by Growth in Emerging Markets - eMarketer.,”

Fernandes, R. F., & Teixeira, C. M. (2015). “Using clickstream data to analyze online purchase intentions”, Dissertation Research Submitted to University of Porto.

Gironés, Jordi, et al. (2017), “Minería de datos:modelos y algoritmos”, Editorial UOC. Machine learning, 29 (2-3), 131-163.

G. Liu, T. T. Nguyen, G. Zhao et al. (2016), “Repeat Buyer Prediction for E-Commerce,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–164, San Francisco, CA, USA. Obtenido de: https://dl.acm.org/doi/10.1145/2939672.2939674

Holland, C. P., and Mandry, G. D. (2013). “Online search and buying behaviour in consumer markets,” in 46th Hawaii International Conference on System Sciences (HICSS), 2013, pp. 2918–2927.

IBM Cloud Education (2020). “Exploratory Data Analysis”. Obtenido de: https://www.ibm.com/cloud/learn/exploratory-data-analysis

Johnson, E. J., Moe, W. W., Fader, P. S., Bellman, S., and Lohse, G. L. (2004). “On the depth and dynamics of online search behavior,” Management science, (50:3), INFORMS, pp. 299–308.

Kassambara, Alboukadel. (2018). “”Logistic Regression Essentials in R”. Obtenido de: http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

Lateef, Zulaikha. (2020). “A complete guide on decision tree algorithm”. Obtenido en: https://www.edureka.co/blog/decision-tree-algorithm/

Liu, G. et al. (2016). “Repeat Buyer Prediction for E-Commerce”.*In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 155–164. DOI: https://doi.org/10.1145/2939672.2939674 Obtenido de: https://www.kdd.org/kdd2016/papers/files/adf0160-liuA.pdf

Marfice, Christina (2020). The Evolution of ecommerce [10-50 years] [Timeline]. Obtenido de: https://www.plytix.com/blog/evolution-of-ecommerce-timeline

Niu, X., Li, C., Yu, X. (2017). “Predictive Analytics of E-Commerce Search Behavior for Conversion”. Twenty-third Americas Conference on Information Systems, Boston, 2017.

Pichardo, Francisco. (2017). “Apuntes de desbalance de clases en clasificación inteligente”. Obtenido en: https://franciscopichardoblog.files.wordpress.com/2017/06/apuntes-de-desbalance1.pdf

Peterson, R. A., and Merino, M. C. (2003). “Consumer information search behavior and the Internet,” Psychology & Marketing, (20:2), Wiley Online Library, pp. 99–121.

Varian, H. R. 2014. “Big data: New tricks for econometrics,” The Journal of Economic Perspectives, (28:2), American Economic Association, pp. 3–27.

Vieira, Armando. (2015). Predicting online user behaviour using deep learningalgorithms. The Computing Research Repository (CoRR) abs/1511.06247

Yadav, Dinesh. (2019). “Categorical encoding using Label-Encoding and One-Hot Encoder. Obtenido en: https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd

Zhang, H. and Dong, J. (2020). "Prediction of Repeat Customers on E-Commerce Platform Based on Blockchain", *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8841437, 15 pages. https://doi.org/10.1155/2020/8841437