Trabajo Fin de Máster
Organización Industrial y Gestión de Empresas

Aplicación de algoritmos automáticos de aprendizaje supervisado para predecir el abandono de clientes en telefonía móvil

Autor: Juan José Cuevas Gómez
Tutor: Antonio Plácido Moreno Beltrán

Dpto. de Organización Industrial y Gestión de Empresas I
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla

Sevilla, 2023

Trabajo Fin de Máster
Organización Industrial y Gestión de Empresas

Aplicación de algoritmos automáticos de aprendizaje supervisado para predecir el abandono de clientes en telefonía móvil

Autor:
Juan José Cuevas Gómez

Tutor:
Antonio Plácido Moreno Beltrán

Dpto. de Organización Industrial y Gestión de Empresas I
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla
Sevilla, 2023

Trabajo fin de Máster: Aplicación de algoritmos automáticos de aprendizaje supervisado para predecir el abandono de clientes en telefonía móvil

Autor: Juan José Cuevas Gómez

Tutor: Antonio Plácido Moreno Beltrán

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Sevilla, 2023

El Secretario del Tribunal

A mi familia

A mis maestros

Resumen

Un aspecto clave para el éxito de una compañía telefónica es mantener la satisfacción de los usuarios que contratan sus servicios, lo que se conoce como fidelización de los clientes. Para ello, es necesario elaborar e implementar estrategias que respondan a las necesidades y preferencias de los clientes, teniendo en cuenta el mercado y la competencia. Para optimizar esas estrategias, es conveniente identificar a qué clientes enfocarse.

La empresa objeto de estudio es una compañía de telecomunicaciones que opera en todo el territorio estadounidense, compuesto por 51 estados. El contexto legal de este país se caracteriza por la apertura al mercado y el fomento de la competencia, garantizando al mismo tiempo el acceso universal al servicio. En este escenario, la rivalidad entre las empresas se ha intensificado en los últimos años, lo que ha impulsado la innovación en los departamentos de fidelización, que buscan estrategias para diferenciarse y mantenerse en el mercado.

Este trabajo busca disminuir el abandono de clientes de la empresa. Para lograrlo, se utilizan herramientas de aprendizaje automático supervisado que permiten crear un modelo predictivo para detectar a los clientes con mayor riesgo de cancelar su contrato con la empresa, y analizar las variables más relevantes en la decisión de estos de abandonar la compañía. Se utiliza una base de datos que contiene información de 3,333 clientes de una empresa telefónica.

El objetivo de este trabajo es identificar a los clientes que tienen más riesgo de abandonar la empresa. Para lograrlo, se han utilizado dos métodos de aprendizaje supervisado de manera consecutiva. El primero permite estimar la probabilidad de que un cliente se dé de baja y el segundo genera una serie de reglas que explican esa probabilidad en función de las características de los clientes.

Para predecir si un cliente cancelará o no los servicios que contrató con la compañía, se usa la regresión logística. Esta técnica permite estimar la relación entre una variable binaria (cancelación o no) y varias variables explicativas (categóricas y continuas) que influyen en ella. Se evaluaron diferentes criterios para seleccionar el mejor modelo, como el criterio de información bayesiano, el criterio de Akaike, el pseudo R^2 de McFadenn, la importancia de las variables y el factor de inflación de la varianza. Luego, se aplicó el árbol de decisión, otra técnica que clasifica a los clientes en dos grupos según la variable binaria. Esta técnica tiene la ventaja de ser más fácil de interpretar y comunicar que la regresión logística, además de obtener mejores métricas de rendimiento.

Para comparar las técnicas de árbol de decisión y regresión logística en la predicción de la cancelación de servicios de telefonía móvil, se construyen y se prueban varios modelos con diferentes parámetros. Se utiliza la matriz de confusión y otras métricas, como la precisión, la sensibilidad y la especificidad, para evaluar el rendimiento de los modelos. Los resultados muestran que el árbol de decisión alcanza una precisión del 95.65% en la predicción de la cancelación, mientras que la regresión logística logra una precisión del 85.30%.

Finalmente con el modelo predictivo obtenido, se diseñan políticas de retención de clientes basadas en las variables más influyentes.

Abstract

A key aspect for the success of a telephone company is to maintain the satisfaction of the users who contract its services, known as customer loyalty. To do this, it is necessary to develop and implement strategies that respond to customer needs and preferences, taking into account the market and the competition. To optimise these strategies, it is advisable to identify which customers to focus on.

The company under study is a telecommunications company that operates throughout the 51 states of the United States. The legal context of this country is characterised by openness to the market and the promotion of competition, while guaranteeing universal access to service. In this scenario, rivalry between companies has intensified in recent years, driving innovation in loyalty departments, which are looking for strategies to differentiate themselves and remain in the market.

This work aims to reduce customer churn. To achieve this, supervised Machine Learning tools are used to create a predictive model to detect the customers most at risk of cancelling their contract with the company, and to analyse the most relevant variables in their decision to leave the company. A database containing information on 3,333 customers of a telephone company is used.

The objective of this work is to identify the customers who are most at risk of leaving the company. To achieve this, two supervised learning methods have been used consecutively. The first method estimates the probability that a customer will churn and the second method generates a set of rules that explain this probability based on customer characteristics.

Logistic regression was used to predict whether or not a customer will cancel services contracted with the company. This technique allows estimating the relationship between a binary variable (cancellation or not) and several explanatory variables (categorical and continuous) that influence it. Different criteria were evaluated to select the best model, such as the Bayesian information criterion, the Akaike criterion, McFadenn's pseudo R^2, the importance of the variables and the variance inflation factor. Then, the decision tree, another technique that classifies customers into two groups according to the binary variable, was applied. This technique has the advantage of being easier to interpret and communicate than logistic regression, as well as obtaining better performance metrics.

To compare decision tree and logistic regression techniques in predicting mobile churn, several models with different parameters are built and tested. The confusion matrix and other metrics, such as accuracy, sensitivity and specificity, are used to evaluate the performance of the models. The results show that the decision tree achieves an accuracy of 95.65% in predicting cancellation, while the logistic regression achieves an accuracy of 85.30%.

Finally, with the predictive model obtained, customer retention policies are designed based on the most influential variables.

Índice

Resumen 8

Abstract 9

Índice 10

Índice de Tablas 11

Índice de Figuras 12

1 Introducción 14

2 Estado del arte 15

2.1 Aprendizaje no supervisado *16*

2.2 Aprendizaje supervisado *17*

3 Diseño y validación de la herramienta predictiva 20

3.1 Base de datos *20*

3.1.1 Estadística descriptiva 26

3.2 Modelo predictivo a partir de la Regresión Logística Binaria *32*

3.2.1 Construcción del modelo 32

3.2.2 Elección del mejor modelo 33

3.2.3 Validación del modelo 37

3.2.4 Contraste de hipótesis sobre los parámetros del modelo 46

3.2.5 Intervalos de confianza 47

3.2.6 Diagnóstico del modelo 48

3.2.7 Estudio de los datos influyentes 55

3.3 Modelo predictivo a partir de árboles de clasificación *57*

4 Conclusiones 63

Referencias 65

Anexo I 67

Anexo II 73

ÍNDICE DE TABLAS

Tabla 3.1 Matriz de correlación 27
Tabla 3.2 Medidas de bondad de ajuste de diferentes modelos 35
Tabla 3.3 Valores de los coeficientes de los parámetros (modelo 1) 35
Tabla 3.4 Odds ratio (modelo 1) 36
Tabla 3.5 Matriz de Confusión con P = 0.5 39
Tabla 3.6 Matriz de Confusión con P = 0.1 40
Tabla 3.7 Matriz de Confusión con P = 0.3 41
Tabla 3.8 Matriz de Confusión con P = 0.2 42
Tabla 3.9 Matriz de confusión para P*=0.3 (modelo 1) 43
Tabla 3.10 Matriz de confusión para P*=0.5 (modelo 1) 43
Tabla 3.11 Escala índice de concordancia Kappa de Cohen (k) 45
Tabla 3.12 Interpretación del valor de p 46
Tabla 3.13 Contraste de Wald para los parámetros de los coeficientes (modelo 1) 46
Tabla 3.14 Intervalos de confianza para los parámetros del modelo 47
Tabla 3.15 Intervalos de confianza de los cocientes de las ventajas 48
Tabla 3.16 Residuos significativos 48
Tabla 3.17 Distancias de Cook 49
Tabla 3.18 DFBETA 49
Tabla 3.19 DFFITS 50
Tabla 3.20 Parámetros ajustados (modelo 2) 55
Tabla 3.21 Modelo 1 frente a Modelo 2 56
Tabla 3.22 Interpretación de los coeficientes (modelo 2) 57
Tabla 3.23 Matriz de confusión (árbol 1) 58
Tabla 3.24 Matriz de confusión (árbol 2) 59
Tabla 3.25 Matriz de confusión (árbol 3) 59
Tabla 4.1 Métricas definitivas 63

ÍNDICE DE FIGURAS

Figura 2.1 Cuota de mercado de las principales empresas estadounidenses (diciembre de 2021) 15

Figura 2.2 Red neuronal Yasser Khan y otros (2019) 17

Figura 3.1 Gráfico de datos faltantes 21

Figura 3.2 Análisis básico de las variables. Rstudio 22

Figura 3.3 Estadísticos básicos de las variables. Rstudio 22

Figura 3.4 Histograma de la variable “duración del contrato” 23

Figura 3.5 Histograma de la variable “duración del contrato” en función de la variable respuesta “cancelación de los servicios” 23

Figura 3.6 Histograma de la variable “número de mensajes de texto” 23

Figura 3.7 Histograma de la variable “número de mensajes de texto” en función de la variable respuesta “cancelación de los servicios” 23

Figura 3.8 Histograma de la variable “total de minutos realizados a lo largo del tramo matinal” 24

Figura 3.9 Histograma de la variable “total de minutos realizados a lo largo del tramo matinal” en función de la variable respuesta “cancelación de los servicios” 24

Figura 3.10 Histograma de la variable "total de llamadas realizadas a lo largo del tramo de la tarde" 24

Figura 3.11 Histograma de la variable " total de llamadas realizadas a lo largo del tramo de la tarde " en función de la variable respuesta “cancelación de los servicios” 24

Figura 3.12 Histograma de la variable "Importe total generado en el tramo de la noche” 25

Figura 3.13 Histograma de la variable " Importe total generado en el tramo de la noche " en función de la variable respuesta “cancelación de los servicios” 25

Figura 3.14 Histograma de la variable "total de minutos internacionales" 25

Figura 3.15 Histograma de la variable " total de minutos internacionales " en función de la variable respuesta “cancelación de los servicios” 25

Figura 3.16 Histograma de la variable "total de llamadas realizadas al centro de atención al cliente de la compañía telefónica” 26

Figura 3.17 Histograma de lavariable " total de llamadas realizadas al centro de atención al cliente de la compañía telefónica " en función de la variable respuesta “cancelación de los servicios” 26

Figura 3.18 Gráfico de la variable “plan internacional de llamadas de voz” frente a “cancelación de los servicios” 29

Figura 3.19 Gráfico de la variable “llamadas realizadas al centro de atención al cliente” frente a “cancelación de los servicios” 29

Figura 3.20 Gráfico de la variable “estado” frente a “cancelación de los servicios” 31

Figura 3.21 Gráfico de dispersión de la variable “duración del contrato” frente a “total de

llamadas realizadas al centro de atención al cliente" frente a "cancelación de los servicios contratados" 31

Figura 3.22 Gráfico de dispersión de la variable "total de minutos realizados durante el tramo matinal" frente a "total de llamadas realizadas al centro de atención al cliente" frente a "cancelación de los servicios contratados" 31

Figura 3.23 Curva ROC. Datos de construcción. 38

Figura 3.24 Curva ROC. Datos de validación. 38

Figura 3.25 Diagrama de dispersión entre cada predictor y los valores logit 51

Figura 3.26 Distancias de Cook frente a Leverage 54

Figura 3.27 Árbol de clasificación (árbol 3) 61

1 INTRODUCCIÓN

Una de las principales preocupaciones de las empresas es no poder predecir cuándo un cliente dejará de comprar sus productos o servicios. Esta situación tiene un alto impacto económico, pues mantener a un cliente fiel suele ser más rentable que atraer a uno nuevo. Por eso, muchas empresas quieren crear modelos predictivos que les ayuden a detectar a los clientes con más riesgo de fuga y diseñar estrategias de fidelización personalizadas.

Una forma de mejorar la rentabilidad y la competitividad de una empresa es lograr que sus clientes se mantengan fieles a su marca y a sus servicios. Esto es especialmente importante en el sector de las telecomunicaciones, donde hay mucha competencia, los clientes pueden cambiar fácilmente de operador y los productos ofrecidos son muy similares. Ante esta situación, es conveniente detectar a los clientes que están en riesgo de irse a otra compañía y ofrecerles incentivos para que se queden, mediante acciones de fidelización personalizadas.

El objetivo de este trabajo es crear un modelo predictivo que clasifique a los clientes de una compañía telefónica estadounidense según cancelen o no sus servicios, usando algoritmos automáticos de aprendizaje supervisado. Para ello, se emplean dos técnicas estadísticas de clasificación binaria ampliamente desarrolladas en la minería de datos, la regresión logística y los árboles de decisión. La regresión logística estima la probabilidad de pertenencia a una clase (cancelación o no) a partir de una combinación lineal de variables predictoras. Los árboles de decisión generan reglas de clasificación simples e interpretables a partir de la división recursiva del espacio de las variables predictoras.

La base de datos utilizada para el desarrollo del modelo esta formada por información de clientes de una compañía telefónica de Estados Unidos, y contiene información de 21 variables, de las cuales una es la variable objetivo que indica si el cliente se da de baja o no de la empresa. Las otras 20 variables son predictoras y abarcan diferentes aspectos de los clientes, como por ejemplo:

- Datos demográficos: estado al que pertenece el cliente, código postal.
- Datos contractuales: duración del contrato, si el cliente contrato un plan de llamadas.
- Datos de uso: número de mensajes enviados, minutos totales de llamadas, llamadas totales, importe total.
- Datos de satisfacción: llamadas totales realizadas al centro de atención al cliente.

El objetivo del trabajo es comparar el rendimiento y la capacidad predictiva del modelo por regresión logística y árboles de decisión, así como analizar las variables más relevantes para explicar la cancelación de los servicios con la compañía.

Se desarrollan varios modelos con ambas técnicas y se comparan con métricas como la precisión, la sensibilidad, la especificidad y el área bajo la curva ROC, entre otras.

Finalmente se analizan los resultados y se determinan los factores que más afectan al abandono del cliente. Además, se ofrecen recomendaciones para aumentar la fidelización de los clientes.

2 ESTADO DEL ARTE

La ley de telecomunicaciones de 1996 fue un hito histórico en Estados Unidos, ya que abrió el mercado de las telecomunicaciones a la competencia. Esta ley cambió radicalmente el marco legal de las telecomunicaciones en ese país.

Desde hace más de 70 años, el mercado de las telecomunicaciones ha experimentado una gran transformación. Antes, solo existía una compañía que ofrecía el servicio de telefonía, pero en 1996 se abrió la competencia a otras empresas privadas. Con el avance tecnológico, el sector ha crecido mucho y se ha generado una rivalidad por captar más clientes. Sin embargo, esto también ha provocado un problema para las compañías: la elevada tasa de abandono de los usuarios.

En la Figura 2.1 se muestra la distribución del mercado de las telecomunicaciones en Estados Unidos para las principales compañías telefónicas del país.

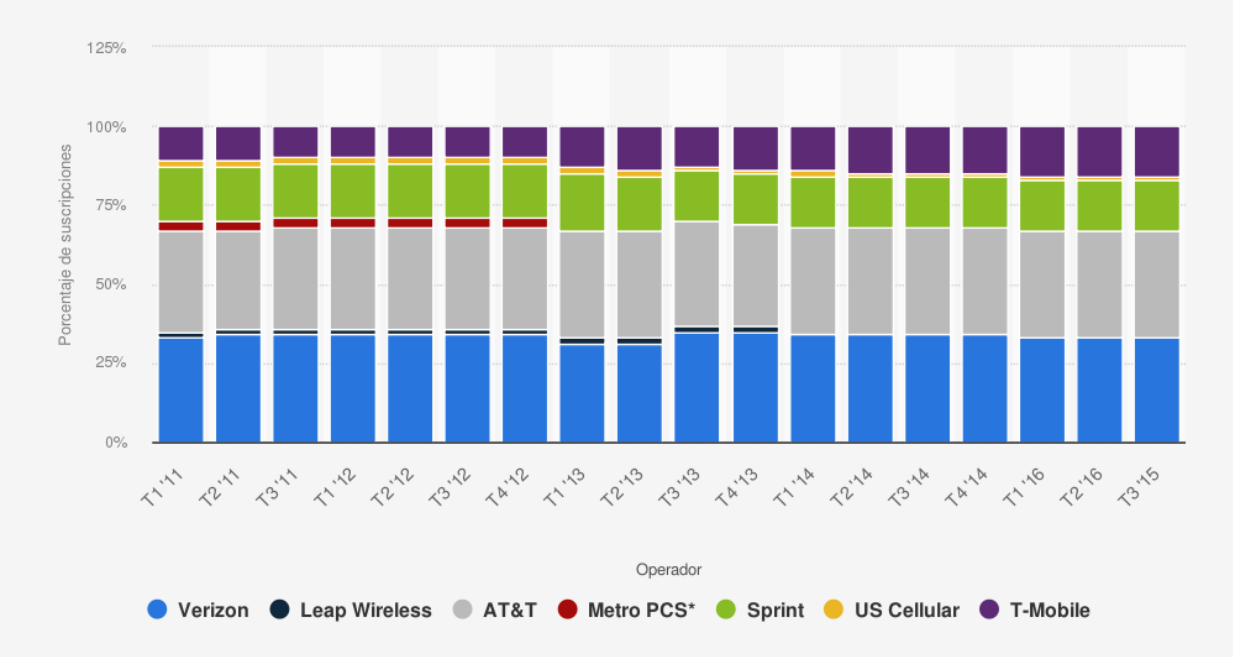


Figura 2.1 Cuota de mercado de las principales empresas estadounidenses (diciembre de 2021)

Las empresas de telecomunicaciones pueden ahorrar dinero si consiguen fidelizar a sus clientes, ya que captar nuevos clientes es seis veces más costoso. Para evitar que los clientes se den de baja, es importante analizar sus relaciones con la empresa mediante el CRM (Customer Relationship Management). Así, se pueden usar técnicas de aprendizaje automático y análisis de datos para detectar señales de que un cliente está pensando en cambiar de compañía. Esta estrategia se basa en la ley de telecomunicaciones que se aprobó en 1996 y que cambió el marco legal del sector en el país.

La empresa debe diseñar una estrategia de retención que le permita conservar a los clientes potenciales y evitar las cancelaciones. Para ello, puede ofrecer promociones especiales, un servicio al cliente de calidad, y una experiencia de compra personalizada. Estas acciones proactivas contribuyen a aumentar la retención y fidelización de los clientes a largo plazo.

Una de las aplicaciones del Machine Learning es la creación de modelos predictivos que permiten estimar la probabilidad de que un cliente contrate o cancele un servicio específico. Estos modelos se basan en

diferentes técnicas y metodologías que se adaptan a las necesidades y objetivos de cada área de negocio, como ventas, marketing, atención al cliente, etc. Algunas de las técnicas más utilizadas son:

- Una forma de medir cuánto tiempo permanecen los clientes con la compañía es el **análisis de supervivencia**. Esta técnica usa una función matemática llamada "función de supervivencia" para calcular la probabilidad de que un cliente no haya cancelado el servicio en un momento determinado. Para obtener la función de supervivencia, se aplican diferentes modelos a los datos disponibles y se examina cómo influyen las variables explicativas en el tiempo que transcurre hasta la cancelación. Los modelos más habituales para obtener la función de supervivencia son el modelo de Kaplan-Meier y el modelo de regresión de Cox.

- Análisis de redes sociales: esta técnica consiste en examinar las interacciones de los clientes en las redes sociales para identificar posibles indicios de abandono. Mediante algoritmos de extracción de texto y análisis de sentimiento, se evalúan los mensajes de los clientes y se detectan posibles inconvenientes.

- La minería de datos consiste en analizar grandes volúmenes de datos para encontrar patrones y relaciones entre las variables, con el fin de aprender y poder predecir el comportamiento de una variable de interés. Para ello, se aplican algoritmos y modelos estadísticos de aprendizaje automático.

Una forma de entender la minería de datos es como un proceso que utiliza el aprendizaje automático (Machine Learning) para extraer conocimiento de los datos. Dentro del aprendizaje automático, existen dos tipos principales de problemas: el aprendizaje supervisado, que consiste en aprender a predecir una salida a partir de unas entradas, y el aprendizaje no supervisado, que consiste en descubrir patrones o estructuras ocultas en los datos.

2.1 Aprendizaje no supervisado

Una forma de Machine Learning que no requiere etiquetas en las instancias es el aprendizaje no supervisado. Su objetivo es encontrar patrones de similitud, diferencia o asociación entre los datos, según el problema que se quiera resolver. Para ello, se utilizan diferentes modelos y técnicas que analizan datos históricos y descubren estructuras ocultas en ellos.

- Una forma de analizar datos es mediante el uso de **modelos de clustering**, que consisten en agrupar puntos de datos similares en distintos grupos. Estos modelos intentan que los elementos de un mismo grupo sean lo más parecidos posible y que los grupos sean lo más diferentes posible entre sí. Esta técnica tiene diversas aplicaciones, como la segmentación de clientes, el análisis de precios de mercado, etc.

  Kmeans es un tipo de modelo de clustering que no requiere etiquetas previas para los datos. Este método asigna cada observación al grupo cuya media es la más próxima.

- El propósito de la detección de anomalías es identificar observaciones que no se ajustan al patrón general. Esta técnica puede tener varias aplicaciones, como eliminar datos que contienen instancias anómalas, detectar fraudes en transacciones económicas, reconocer medidas inusuales en dispositivos, etc.

- Asociaciones, estos algoritmos analizan los datos para encontrar grupos de elementos que se presentan juntos con frecuencia, lo que indica una relación entre ellos. Esto permite estudiar los patrones de consumo de los clientes, asociar los productos para mejorar la distribución, conocer las preferencias de los clientes y los servicios contratados, entre otras aplicaciones.

2.2 Aprendizaje supervisado

El objetivo es obtener un conjunto de reglas que predigan el valor de la variable de estudio a partir de datos nuevos. Para ello, se usa un conjunto de datos ya clasificados, donde cada elemento tiene un conjunto de atributos y un valor para la variable de estudio. Las técnicas se aplican según el tipo de variable, que puede ser numérica (regresión) o categórica (clasificación).

Una forma de crear modelos de predicción es usar diferentes métodos que se engloban en el aprendizaje automático supervisado. Sin embargo, algunos de estos métodos también se pueden aplicar al aprendizaje automático no supervisado, que no requiere etiquetas previas para los datos.

- Regresión lineal, método basado en el cálculo de una recta que se ajusta a las observaciones, de forma que al introducir nuevas observaciones el modelo predice la posición de salida. Los atributos son de tipo cuantitativo.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1..N$$

- Algoritmo de Máquinas de Vectores de Soporte (SVM), el objetivo de este método es encontrar la mejor separación entre dos clases o la mejor aproximación de una función de regresión lineal en un espacio de alta dimensión. Este algoritmo tiene muchas áreas de aplicación: la clasificación de imágenes, la detección de spam, la predicción de precios, etc.

- Algoritmo K-Nearest Neighbours, trata de predecir la etiqueta de una observación desconocida en función de las etiquetas de las observaciones conocidas más cercanas a ella, de forma que encuentra los k vecinos más cercanos y asigna la etiqueta más común entre ellos. La elección del valor de k es importante, ya que, si k es demasiado pequeño, el algoritmo puede ser sensible a puntos outliers, mientras que si k es demasiado grande, el algoritmo puede perder la capacidad de detectar patrones locales en los datos. Se emplea en diferentes áreas, entre las que destacan: la clasificación de imágenes, la recomendación de productos o servicios, la detección de fraude, etc.

  Autores lo emplean junto con regresión logística para la predicción del abandono de clientes en el área de telecomunicaciones como es el caso del artículo Tianpei Xu y otros (2021).

- Redes neuronales, son técnicas para encontrar una combinación de parámetros con el objetivo de predecir un resultado. Primero se entrena la red neuronal para conseguir la combinación que mejor se ajusta para hacer predicciones o clasificaciones, una vez entrenada se aplica la combinación para predecir o clasificar. A medida que aumenta la complejidad de la aplicación para la que se entrena al modelo es necesario aumentar el número de capas. Puede verse a continuación una figura de Yasser Khan y otros (2019).

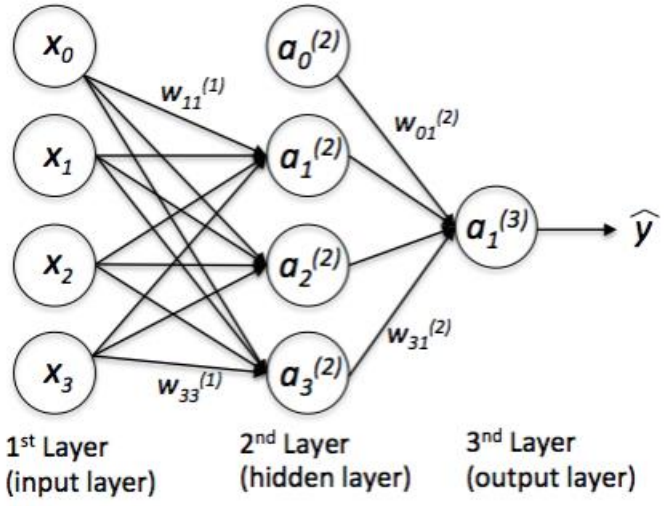


Figura 2.2 Red neuronal Yasser Khan y otros (2019)

El artículo de Yasser Khan y otros (2019), describe un estudio que utiliza redes neuronales artificiales para predecir la tasa de abandono de clientes en la industria de las telecomunicaciones. Los autores

explican que la predicción del abandono de clientes es un problema crítico en la industria de las telecomunicaciones, y que las técnicas de minería de datos, como las ANN, pueden ser útil para abordar este problema. Los autores reportan que su modelo de ANN tiene una precisión del 89,5% en la predicción de la tasa de abandono de clientes, lo que sugiere que las ANN pueden ser una herramienta útil para predecir el abandono de clientes en la industria de las telecomunicaciones.

- Random forest, es un algoritmo de aprendizaje automático que utiliza múltiples árboles de decisión para hacer predicciones precisas. Cada árbol de decisión en el bosque se construye de forma independiente utilizando una muestra aleatoria de los datos de entrenamiento y una selección aleatoria de características. Durante la predicción, el modelo combina las predicciones de todos los árboles para producir una predicción final. Debido a su alta precisión y capacidad de manejar datos de alta dimensión, Random Forest es uno de los algoritmos de aprendizaje automático más populares y ampliamente utilizados en la industria y en la investigación.

  Los autores de la investigación Adhikary y Gupta (2020), comparan 114 clasificadores de rendimiento para diferentes técnicas de predicción, obteniendo resultados con buen rendimiento para la familia de clasificadores Ramdon Forest, y arroja métricas de menor rendimiento para los métodos basados en SVM.

- Regresión logística binaria, este modelo requiere definir una función logística o sigmoide que transforme el resultado lineal en una probabilidad entre 0 y 1, y ajustar los coeficientes del modelo mediante un método de máxima verosimilitud. La ventaja de este modelo es que es simple, interpretable y fácil de implementar. Tambien presenta algunas limitaciones, como la asunción de linealidad entre las variables predictoras y el logit (el logaritmo de la razón de probabilidades), la sensibilidad a la multicolinealidad (alta correlación entre las variables predictoras) y la dificultad para captar interacciones complejas entre las variables.

  El trabajo de Tianpei Xu y otros (2021) emplea la regresión logística para la predicción de la cancelación de los servicios con la compañía y sus estudios concluyen un nivel de exactitud superior para la regresión logística que para los arboles de decisión o el algoritmo clasificador probabilístico basado en el teorema de Bayes.

  Otros autores proponen el uso de la regresión logística robusta en predicciones, ya que resulta menos sensible a datos atípicos o outlier, como se estudia en el trabajo de Henríquez (2008).

- Arboles de clasificación, son una técnica de aprendizaje automático supervisado no lineal, empleado para la clasificación de las instancias en función de una serie de atributos. Se representan gráficamente en forma de árbol, donde cada nodo representa una cuestión sobre un atributo y cada rama representa una respuesta en este caso la cancelación o no de los servicios con la compañía. Son útiles para la predicción y son de fácil interpretación. La desventaja de esta técnica es que puede ser propensa al sobreajuste o la inestabilidad si no se controla la profundidad o el tamaño mínimo del árbol. Su aplicación es muy amplia: en la predicción de enfermedades, de correo no deseado, en la clasificación de imágenes, etc.

  En la literatura se puede encontrar esta metodología aplicada en el área de las telecomunicaciones, como también en otro tipo de comercios electrónicos como es el caso del trabajo de Kima y Leeb (2022), obteniéndose en este estudio una exactitud en la predicción para los árboles de decisión del 82 %.

Dos métodos de aprendizaje automático supervisado que se pueden usar para predecir si un cliente va a cancelar su servicio son la regresión logística binaria y los árboles de clasificación. Estos métodos utilizan datos de entrada como el número total de llamadas, las llamadas internacionales, el plan de voz, el número de mensajes y otros datos de los clientes que están almacenados en los sistemas de información de las empresas.

Las predicciones del modelo se pueden contrastar con los valores reales observados usando diferentes medidas, como la precisión, la sensibilidad, la especificidad o el área bajo la curva ROC, entre otras. Estas medidas permiten evaluar el desempeño de ambas técnicas.

El análisis de datos ha experimentado avances significativos en los últimos tiempos, lo que ha permitido integrar diversas formas de sistemas de gestión de relaciones con los clientes (CRM) como métodos de análisis

de datos. Estos sistemas han despertado el interés de numerosos estudios y prácticas. Entre las dos técnicas que se utilizan con frecuencia en este ámbito se encuentran la regresión logística y los árboles de decisión, que tienen aplicaciones reales en diversos campos, como se muestra en la literatura. Algunos ejemplos son:

- Saha y otros (2023), este artículo emplea varios métodos de aprendizaje supervisado entre los que se encuentran los método citados anteriormente y otros como redes neuronales, k vecinos más cercanos. Se emplean dos base de datos públicas de Asia y de Estados Unidos para la construcción de los modelos. Las técnicas basadas en redes neuronales alcanzan una exactitud del 99%.

- Beeharry y Tsokizep (2022), estos autores emplean una base de datos pública con 7043 clientes y 21 atributos. Proponen varias técnicas de machine learnig para la predicción de la cancelación de los servicios, entre las que se encuentran la técnica de regresión logística y arboles de decisión, obteniéndose un 80.02 % y 81.17 % de exactitud respectivamente.

- Aleksandar y otros (2016), este artículo aborda la problemática del sector de las telecomunicaciones empleando la regresión logística para predecir la baja del cliente de una compañía de telefonía móvil en Macedonia. Obteniendo una exactitud del 94.35 %.

- Brandusoio y Toderean (2013), este artículo propone tres modelos predictivos para la rotación de suscriptores en empresas de telecomunicaciones móviles, utilizando algoritmos de árboles de decisión cart, chaid y quest. Se concluye un rendimiento similar para los tres modelos propuestos.

- Jadhav y otros (2011), en este estudio los autores predicen los clientes que están en riesgo de cancelar los servicios con una compañía telefónica de la India, obteniendo un resultado del 17,67 % de clientes que cancelarían los servicios, empleando un algoritmo de propagación.

- Xie y otros (2009), estudio realizado en China en el sector bancario para la predicción de la cancelación de los servicios con la compañía bancaria, a partir de una técnica de bosques aleatorios mejorados, comparándola con la técnica de los árboles de decisión. Se obtienen mejores resultados para el modelo mejorado.

- Dalvi y otros, este artículo propone el uso de regresión logística y arboles de decisión para estudiar el abandono de los clientes en el sector de las telecomunicaciones, estos métodos aportan explicaciones fácilmente deducibles sobre las razones causantes de la cancelación de los servicios con la compañía. La regresión logística proporciona conocimientos sobre que atributos y características influyen en la cancelación de los servicios con la compañía, por otro lado la técnica de árboles de decisión aportan una descripción gráfica de como se clasifican los datos y generar reglas para la posterior predicción de una nueva entrada.

La literatura muestra que hay muchas aplicaciones diferentes para analizar y predecir los clientes que tienen una alta probabilidad de cancelar los servicios que contrataron con una compañía y que luego se pasarán a la competencia. Esta problemática de la rotación de clientes afecta tanto a las compañías del sector de las telecomunicaciones como a otras muchas compañías de servicios.

3 DISEÑO Y VALIDACIÓN DE LA HERRAMIENTA PREDICTIVA

La regresión logística es un método que se usa cuando la variable que queremos estudiar tiene dos posibles valores, como por ejemplo, si un cliente cancela o no el servicio con la compañía. Este método nos permite estimar la probabilidad de que ocurra cada uno de estos valores. Cuando la variable tiene solo dos opciones, se llama regresión logística binaria.

En este capítulo se utiliza el software R versión 4.2.2 (R Core Team 2022) para realizar un análisis de una variable dicotómica en relación con diferentes variables explicativas. El proceso consiste en construir el modelo, validarlo y diagnosticarlo, usando paquetes complementarios de R para elaborar gráficos y estudios estadísticos.

3.1 Base de datos

La fuente de los datos utilizados es pública y pertenece a una empresa de telecomunicaciones que opera en los Estados Unidos de América.

El conjunto de datos que se usa para construir el modelo tiene veinte atributos para cada cliente, entre ellos la variable que se quiere pronosticar. A continuación se describen los atributos:

- Estado (variable cualitativa nominal), corresponde a los diferentes estados integrantes de los Estados Unidos de América. Se emplean las abreviaturas definidas por el propio gobierno estadounidense. Se encuentran 51 niveles.
- Duración del contrato (variable cuantitativa discreta), esta variable define el número de meses que permanece el cliente en la compañía, su permanencia o antigüedad. Queda definida por números enteros.
- Código postal (variable cualitativa nominal), corresponde con áreas de marcado telefónico. Y se identifican tres niveles "408","415","510".
- Plan de voz internacional (variable cualitativa nominal dicotómica), esta variable es de tipo binaria definida por “si”, “no”, y hace referencia a la contratación de un plan de llamadas internacionales por parte del cliente.
- Plan de llamadas (variable cualitativa nominal dicotómica), esta variable al igual que la anterior se define por “si”, “no”, y hace referencia a la contratación de un plan de llamadas.
- Número de mensajes (variable cuantitativa discreta), esta variable indica el número total de mensajes realizados por el cliente.
- Total de minutos realizados antes del medio día (variable cuantitativa continua), esta variable indica el número total de minutos realizados por el cliente en el tramo horario de la mañana.
- Total de llamadas realizadas antes del medio día (variable cuantitativa nominal), indica el número total de llamadas realizadas por el cliente en el tramo horario de la mañana.
- Importe total en el tramo de la mañana (variable cuantitativa continua), indica el cargo total por los minutos realizados durante el tramo horario de la mañana.
- Total de minutos realizados durante la el tramo horario de la tarde (variable cuantitativa continua),

indica el número de minutos realizados en el tramo horario de la tarde.

- Total de llamadas realizadas durante el tramo horario de la tarde (variable cuantitativa discreta), cuantifica el número de llamadas realizadas en el tramo horario de la tarde.
- Importe total en el tramo horario de la tarde (variable cuantitativa), indica el cargo total por los minutos realizados durante el tramo horario de la tarde.
- Total de minutos realizados en el tramo horario nocturno (variable cuantitativa continua), indica el número de minutos realizados en el tramo horario de la noche.
- Total de llamadas realizadas durante el tramo horario nocturno (variable cuantitativa discreta), define el número de llamadas realizadas en el tramo horario de la noche.
- Importe total nocturno (variable cuantitativa continua), indica el cargo total por los minutos realizados durante el tramo horario de la noche.
- Total de minutos internacionales (variable cuantitativa continua), indica los minutos realizados en llamadas internacionales.
- Total de llamadas internacionales (variable cuantitativa discreta), que indica el número total de llamadas internacionales realizadas.
- Importe total por minutos internacionales (variable cuantitativa continua), indica el importe por los minutos internacionales consumidos.
- Total de llamadas realizadas a atención al cliente (variable cuantitativa discreta), representa el número total de llamadas realizadas por el cliente al centro de llamadas a atención al cliente de la compañía.
- Cancelación de los servicios con la compañía (variable cualitativa nominal dicotómica), representa la cancelación de los servicios con la compañía por parte del cliente, y corresponde a la variable de interés.

Para construir el modelo, se utilizan 2,666 observaciones de un total de 3,333 registros disponibles. El resto de los registros se usan para validar el modelo.

Para verificar la integridad del conjunto de datos, se elabora la Figura 3.1, que muestra el número de datos faltantes por instancia en el eje vertical y el número de datos faltantes por atributo en el eje horizontal.

Figura 3.1 Gráfico de datos faltantes

Los registros que tienen un dato faltante se muestran en color azul, mientras que los que tienen todos los datos se muestran en color gris. Esto indica que la base de datos no tiene ningún vacío de información y está completa.

Se examinan los datos de forma preliminar para verificar que no haya datos faltantes y para observar los rangos de los valores. Con la función “str ( )”, el programa nos indica el tipo de variable (numérica o categórica, cuantitativa o cualitativa respectivamente) y los valores que asume, en el caso de variables categóricas, muestra los niveles. Puede verse la salida de la función en la Figura 3.2.

```
Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':        2666 obs. of  20 variables:
 $ State                 : Factor w/ 51 levels "AK","AL","AR",..: 17 36 32 36 37 2 20 25 50 40 ...
 $ Account length        : num  128 107 137 84 75 118 121 147 141 74 ...
 $ Area code             : Factor w/ 3 levels "408","415","510": 2 2 2 1 2 3 3 2 2 2 ...
 $ International plan    : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 2 1 2 2 1 ...
 $ Voice mail plan       : Factor w/ 2 levels "No","Yes": 2 2 1 1 1 1 2 1 2 1 ...
 $ Number vmail messages : num  25 26 0 0 0 0 24 0 37 0 ...
 $ Total day minutes     : num  265 162 243 299 167 ...
 $ Total day calls       : num  110 123 114 71 113 98 88 79 84 127 ...
 $ Total day charge      : num  45.1 27.5 41.4 50.9 28.3 ...
 $ Total eve minutes     : num  197.4 195.5 121.2 61.9 148.3 ...
 $ Total eve calls       : num  99 103 110 88 122 101 108 94 111 148 ...
 $ Total eve charge      : num  16.78 16.62 10.3 5.26 12.61 ...
 $ Total night minutes   : num  245 254 163 197 187 ...
 $ Total night calls     : num  91 103 104 89 121 118 118 96 97 94 ...
 $ Total night charge    : num  11.01 11.45 7.32 8.86 8.41 ...
 $ Total intl minutes    : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 11.2 9.1 ...
 $ Total intl calls      : num  3 3 5 7 3 6 7 6 5 5 ...
 $ Total intl charge     : num  2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 3.02 2.46 ...
 $ Customer service calls: num  1 1 0 2 3 0 3 0 0 0 ...
 $ Churn                 : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
```

Figura 3.2 Análisis básico de las variables. Rstudio

Una forma de obtener los estadísticos básicos de cada variable es mediante la función "summary ( )", que ofrece los siguientes datos para variables cuantitativas: el valor máximo y mínimo que alcanzan las variables, la mediana, la media, el primer y el tercer cuartil. Para las variables cualitativas, indica los distintos niveles que adoptan y el número de registros que corresponden a cada nivel. En la Figura 3.3 se puede observar la salida que arroja el programa para esta función.

```
     State          Account length  Area code  International plan Voice mail plan Number vmail messages Total day minutes
 WV     :  88   Min.   :  1.0   408: 669   No :2396           No :1933       Min.   : 0.000        Min.   :  0.0
 MN     :  70   1st Qu.: 73.0   415:1318   Yes: 270           Yes: 733       1st Qu.: 0.000        1st Qu.:143.4
 NY     :  68   Median :100.0   510: 679                                     Median : 0.000        Median :179.9
 VA     :  67   Mean   :100.6                                                Mean   : 8.022        Mean   :179.5
 AL     :  66   3rd Qu.:127.0                                                3rd Qu.:19.000        3rd Qu.:215.9
 OH     :  66   Max.   :243.0                                                Max.   :50.000        Max.   :350.8
 (Other):2241
Total day calls Total day charge Total eve minutes Total eve calls Total eve charge Total night minutes Total night calls
Min.   :  0.0   Min.   : 0.00    Min.   :  0.0     Min.   :  0     Min.   : 0.00    Min.   : 43.7       Min.   : 33.0
1st Qu.: 87.0   1st Qu.:24.38    1st Qu.:165.3     1st Qu.: 87     1st Qu.:14.05    1st Qu.:166.9       1st Qu.: 87.0
Median :101.0   Median :30.59    Median :200.9     Median :100     Median :17.08    Median :201.2       Median :100.0
Mean   :100.3   Mean   :30.51    Mean   :200.4     Mean   :100     Mean   :17.03    Mean   :201.2       Mean   :100.1
3rd Qu.:114.0   3rd Qu.:36.70    3rd Qu.:235.1     3rd Qu.:114     3rd Qu.:19.98    3rd Qu.:236.5       3rd Qu.:113.0
Max.   :160.0   Max.   :59.64    Max.   :363.7     Max.   :170     Max.   :30.91    Max.   :395.0       Max.   :166.0

Total night charge Total intl minutes Total intl calls Total intl charge Customer service calls   Churn
Min.   : 1.970     Min.   : 0.00      Min.   : 0.000   Min.   :0.000     Min.   :0.000          FALSE:2278
1st Qu.: 7.513     1st Qu.: 8.50      1st Qu.: 3.000   1st Qu.:2.300     1st Qu.:1.000          TRUE : 388
Median : 9.050     Median :10.20      Median : 4.000   Median :2.750     Median :1.000
Mean   : 9.053     Mean   :10.24      Mean   : 4.467   Mean   :2.764     Mean   :1.563
3rd Qu.:10.640     3rd Qu.:12.10      3rd Qu.: 6.000   3rd Qu.:3.270     3rd Qu.:2.000
Max.   :17.770     Max.   :20.00      Max.   :20.000   Max.   :5.400     Max.   :9.000
```

Figura 3.3 Estadísticos básicos de las variables. Rstudio

Para las variables cuantitativas:

- El número máximo de meses que un cliente mantiene servicios contratados con la compañía es de 243 meses. Siendo la media muy cercana a la medina, 100 meses. De esta variable podemos observar que el 75 % de los clientes mantienen un contrato inferior a 127 meses.
- El 90% de los clientes tiene contratado un plan de llamadas internacionales, mientras que el 72,5% de los clientes tiene contratado un plan de llamadas de voz.
- Los mensajes de texto no son una herramienta muy usada por los clientes, se puede observar que el 50% de los clientes no han usado el servicio.
- El tramo de día más usado por los clientes para comunicarse es el tramo horario de la tarde, seguido de la noche y por último el tramo matinal. La media se sitúa en 200.4, 201.2 y 179.5 de minutos respectivamente en cada tramo horario. Para todos los tramos la media se sitúa muy cercana a la mediana.
- Para el caso de llamadas internacionales, la media de minutos consumidos por los clientes se sitúa en 10.24 minutos, siendo la mediana de 10.20 minutos. El tercer cuartil indica que el 75% de los clientes han consumido menos de 12.10 minutos en llamadas internacionales.

- Las llamadas realizadas al centro de atención al cliente de la compañía es una variable importante en el estudio, de ella se concluye que el cliente que mas llamadas a realizado al centro de atención al cliente ha sido de 9 llamadas, el 25% ha realizado más de 2 llamadas y la media se sitúa en 1,5 llamadas por cliente.
- Por último del total de 2,666 clientes, un 14,5% de los clientes cancelo los servicios contratados con la compañía de telefonía móvil.

Aquí se presentan los gráficos de frecuencia de algunas variables numéricas continuas. Estos gráficos permiten visualizar cómo se distribuyen los datos y si presentan alguna tendencia, agrupación o dispersión. En el eje vertical se indica la cantidad de veces que el dato se encuentra en ese rango. Las gráficas de la derecha muestran también la frecuencia de cada variable según la variable de estudio que indica si el cliente se ha dado de baja (color más oscuro) o no.

A partir de estos gráficos se puede determinar la moda para cada variable.

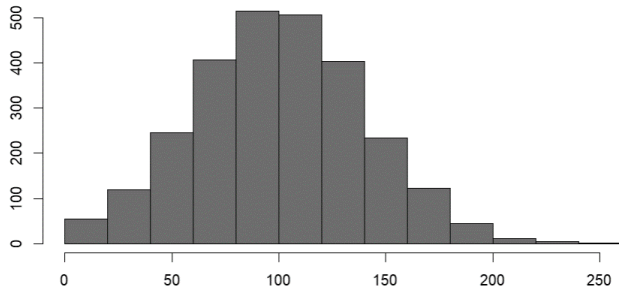


Figura 3.4 Histograma de la variable "duración del contrato"

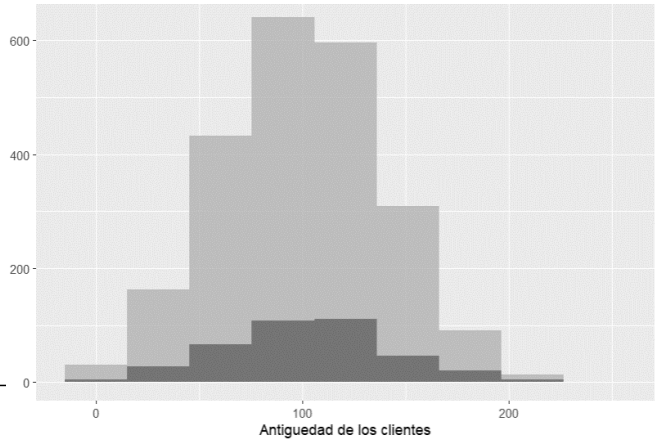


Figura 3.5 Histograma de la variable "duración del contrato" en función de la variable respuesta "cancelación de los servicios"

De la Figura 3.4 y Figura 3.5 se aprecia un histograma para esta variable en forma de campana, la moda se sitúa en 100 días, y la media y la mediana 100.6 y 100 respectivamente, lo que hace indicar que la distribución se asemeja a una normal.

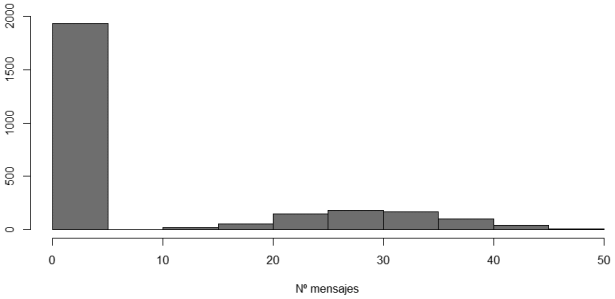


Figura 3.6 Histograma de la variable "número de mensajes de texto"

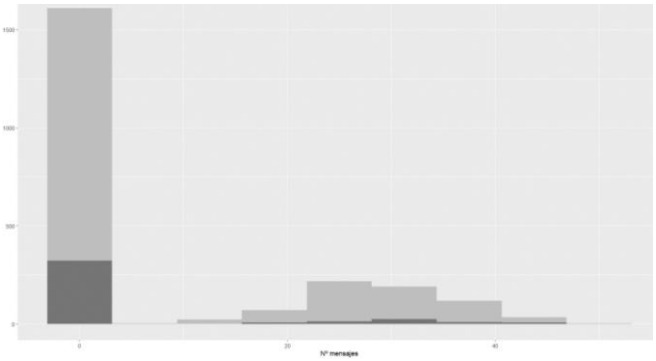


Figura 3.7 Histograma de la variable "número de mensajes de texto" en función de la variable respuesta "cancelación de los servicios"

De la Figura 3,6 y Figura 3.7 se puede concluir quela moda se situa en intervalo (0,5), con un número de repeticiones cercano a los 2,000, por lo que se entiende que de un total de 2,666 clientes, el 75% de ellos no ha usado el servico de mensajería de texto. La variable esta sesgada a la izquierda.

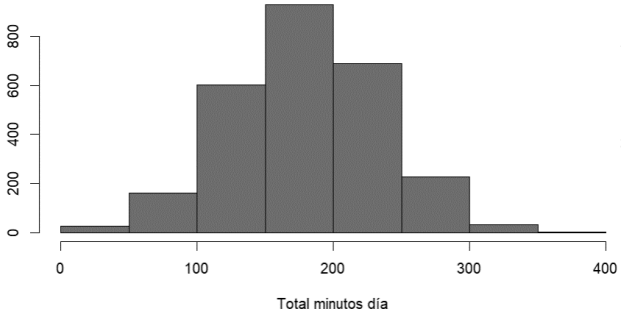


Figura 3.8 Histograma de la variable “total de minutos realizados a lo largo del tramo matinal”

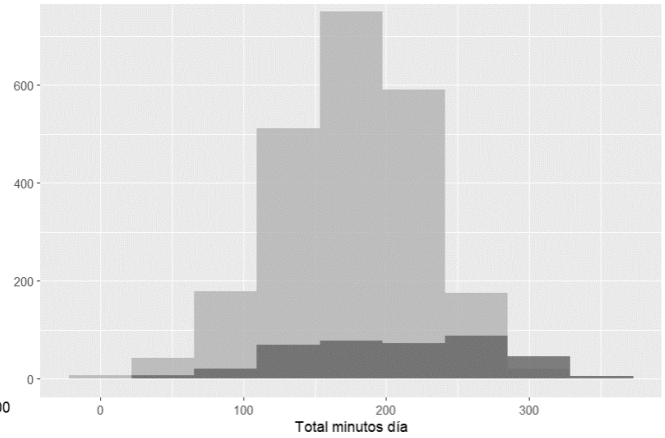


Figura 3.9 Histograma de la variable “total de minutos realizados a lo largo del tramo matinal” en función de la variable respuesta “cancelación de los servicios”

La moda para la Figura 3.8 y Figura 3.9 se sitúa en un tramo entre 150 y 200 minutos, tramo que incluye a la media y la mediana, por lo que puede asemejarse con una distribución normal.

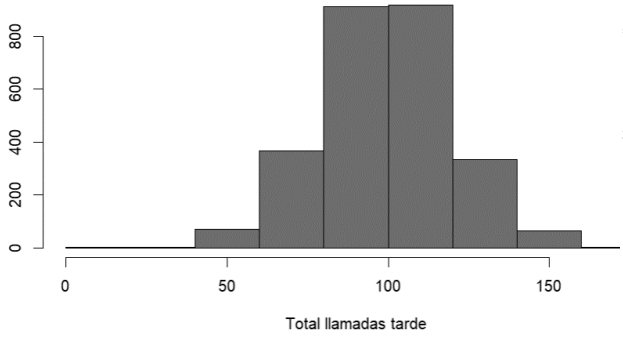


Figura 3.10 Histograma de la variable "total de llamadas realizadas a lo largo del tramo de la tarde"

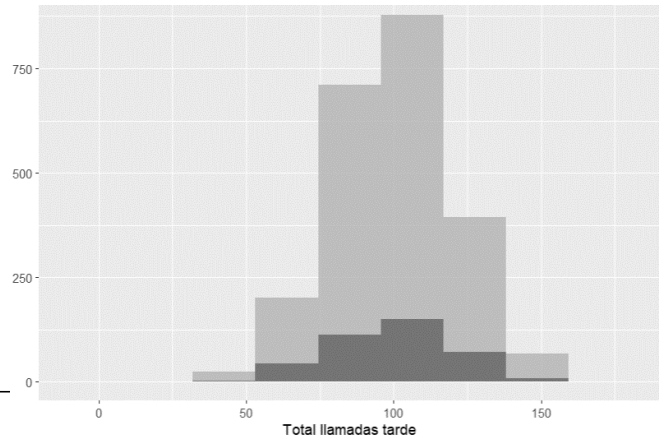


Figura 3.11 Histograma de la variable " total de llamadas realizadas a lo largo del tramo de la tarde " en función de la variable respuesta “cancelación de los servicios”

Para esta variable podría existir dos modas, ya que para ambos intervalos la frecuencia es prácticamente igual, se trata de los intervalos 80-100 y 100-120, se puede apreciar en la Figura 3.10

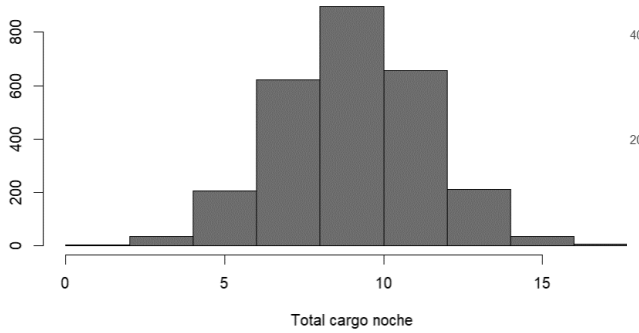


Figura 3.12 Histograma de la variable "Importe total generado en el tramo de la noche"

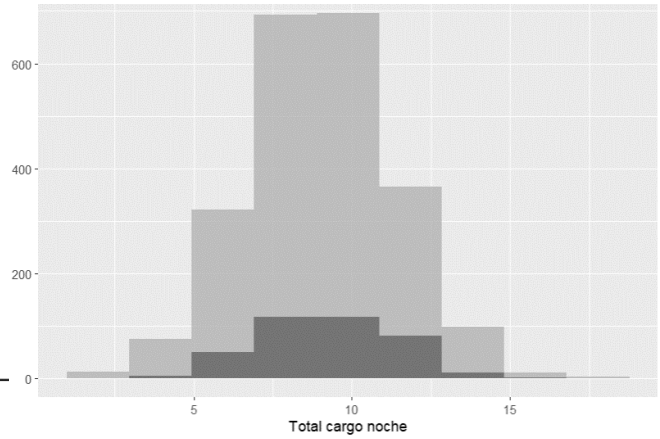


Figura 3.13 Histograma de la variable " Importe total generado en el tramo de la noche " en función de la variable respuesta "cancelación de los servicios"

La moda para la variable que representa la Figura 3.12 y Figura 3.13 se encuentra en el intervalo de 8 a 10, y su distribución se asemeja a una normal.

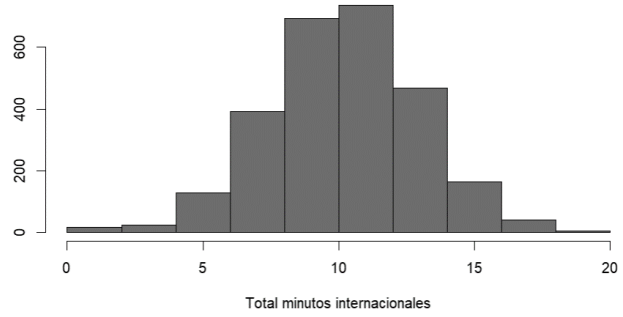


Figura 3.14 Histograma de la variable "total de minutos internacionales"

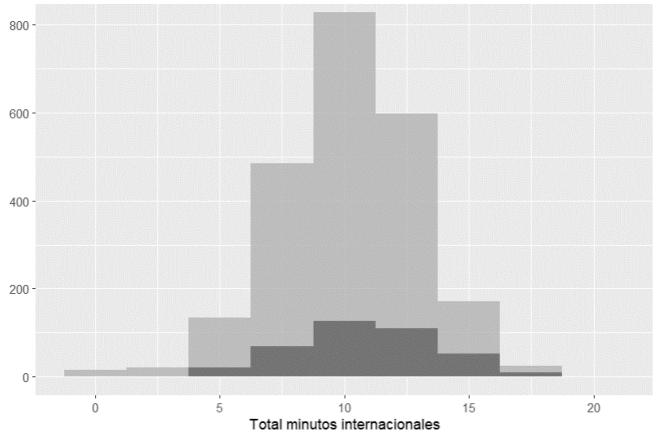


Figura 3.15 Histograma de la variable " total de minutos internacionales " en función de la variable respuesta "cancelación de los servicios"

De la Figura 3.14 y Figura 3.15 se intuye una moda para la variable representada entre 10 y 12 minutos, se asemeja a una distribución normal.

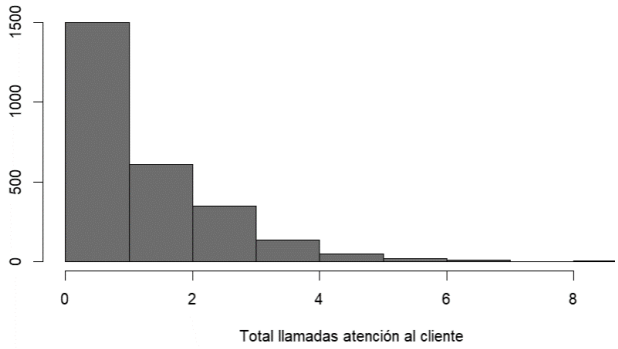


Figura 3.16 Histograma de la variable "total de llamadas realizadas al centro de atención al cliente de la compañía telefónica"

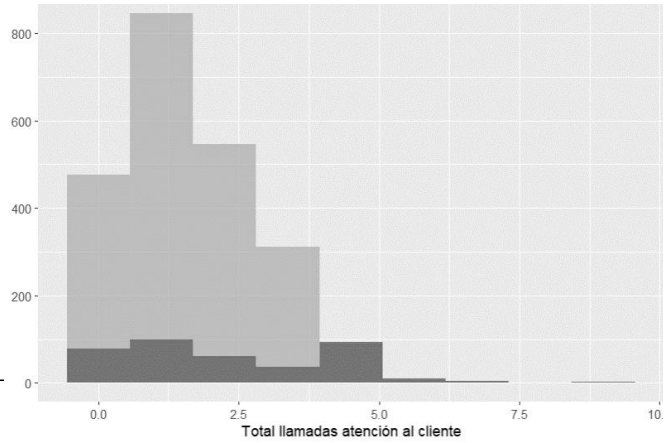


Figura 3.17 Histograma de la variable " total de llamadas realizadas al centro de atención al cliente de la compañía telefónica " en función de la variable respuesta "cancelación de los servicios"

La Figura 3.16 y Figura 3.17 representan al número total de llamadas realizadas al centro de atención al cliente, la moda se sitúa en el tramo 0 a 1, aunque puede verse que aquellos clientes que han realiza un mayor numero de llamadas al servicio de atención al cliente a abandonado la compañía. Esta variable tiene sesgo hacia la izquierda.

El comportamiento de las variables se puede analizar mediante sus distribuciones de frecuencia. La mayoría de ellas presentan una forma gaussiana o bimodal, lo que implica que los datos se agrupan alrededor de la media y tienen una variabilidad simétrica. Un ejemplo de este tipo de variables es la variable "duración del contrato, total de minutos en el tramo matinal". Otras variables, en cambio, no siguen una distribución normal, sino que tienen una forma asimétrica o sesgada. Esto ocurre con la variable "total de llamadas internacionales" y "total de llamadas realizadas al centro de atención al cliente", donde los datos se dispersan y no se concentran cerca de la moda. En el Anexo II se pueden observar los histogramas correspondientes a cada variable.

Una forma de visualizar la distribución y la variabilidad de los datos es mediante los gráficos de caja y bigotes. Estos gráficos representan la mediana, el rango intercuartil y los valores extremos de una muestra, lo que permite identificar posibles anomalías o valores atípicos. En el Anexo II se pueden observar algunos ejemplos de estos gráficos.

3.1.1 Estadística descriptiva

Se analizan las características de las variables cuantitativas mediante estadística descriptiva. Se evalúa la correlación entre las variables para descartar del modelo aquellas que presenten dependencia.

Una forma de medir la relación lineal entre dos variables numéricas es el coeficiente de correlación de Pearson. Este coeficiente varía entre -1 y +1, indicando el grado y la dirección de la asociación entre las variables. En este trabajo, se calcula el coeficiente de correlación de Pearson para las variables numéricas del modelo.

- Si la correlación toma valores negativos: se dice que la correlación lineal entre variables es negativa, es decir, cuando una crece la otra decrece.
- Si la correlación toma valores positivos, cuando una variable crece la otra crece.
- En cambio, si la correlación es 0, significa que no existe correlación lineal entre ellas.

Se emplea la función "cor ( )" para el estudio de la correlación. El resultado que muestra la función corresponde a una matriz que toma valores en su diagonal iguales a 1 y es simétrica.

Para ver las correlaciones de forma más visual pueden verse los gráficos que incluye la Tabla 3.1.

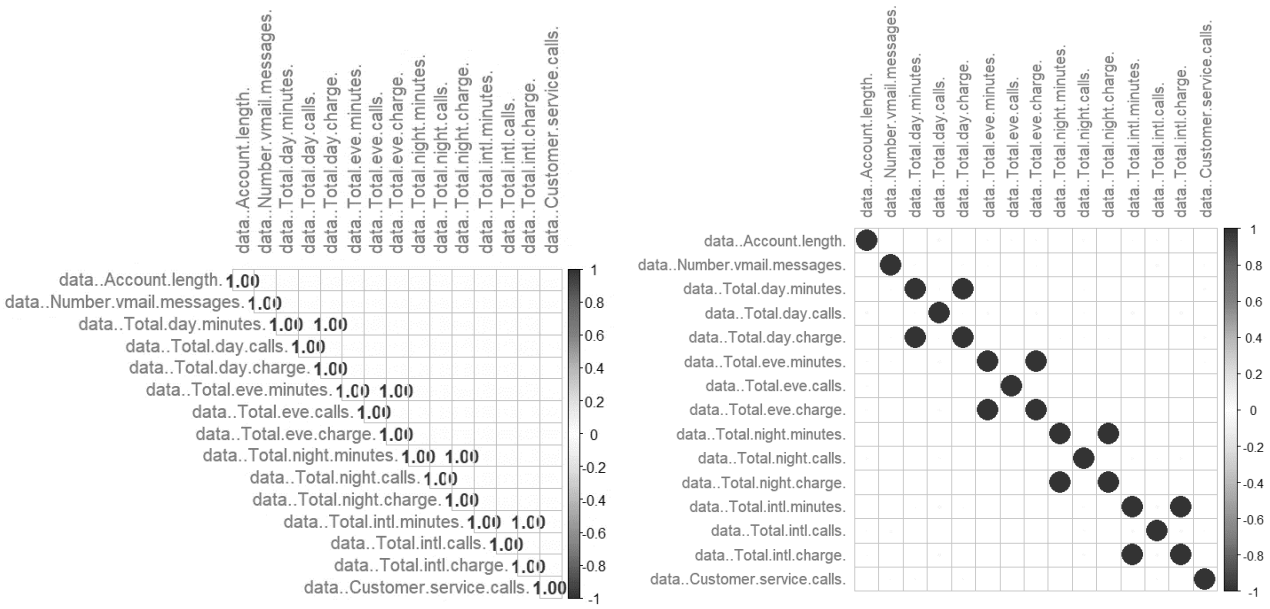


Tabla 3.1 Matriz de correlación

R studio permite representar la correlación de multitud de formas, en la Tabla 3.1 se pueden apreciar dos tipos de gráficos, el gráfico de la izquierda indica el valor numérico de la correlación entre las variables, mientras que la gráfica de la derecha nos ilustra la fuerza de la correlación mediante el tamaño de los círculos. Los datos más relevantes son los que se refieren a la correlación de:

- Total de minutos en el tramo matinal frente a Importe total por los minutos consumidos
- Total de minutos en el tramo de la tarde frente a Importe total por los minutos consumidos
- Total de minutos en el tramo de la noche frente a Importe total por los minutos consumidos
- Total de minutos internacionales frente a Importe total por los minutos consumidos

Estas variables presentan una correlación perfecta, lo que implica que existe multicolinealidad entre ellas. Por esta razón, se analizará más adelante cuál de las dos se puede descartar. Es lógico que al incrementar los minutos totales en cualquier intervalo del día, se incremente también el cargo total.

A continuación, se muestran varias gráficas que representan y estudian la influencia de las variables categóricas sobre la variable de estudio.

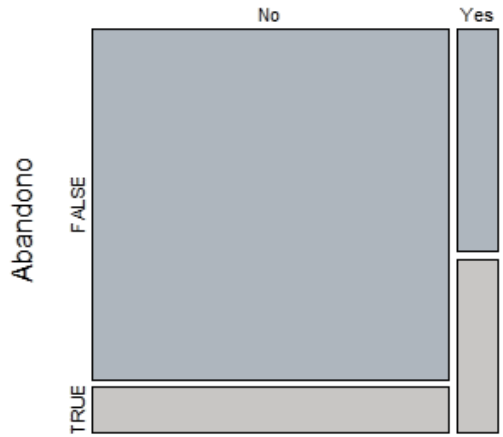


Figura 3.18 Gráfico de la variable "plan internacional de llamadas de voz" frente a "cancelación de los servicios"

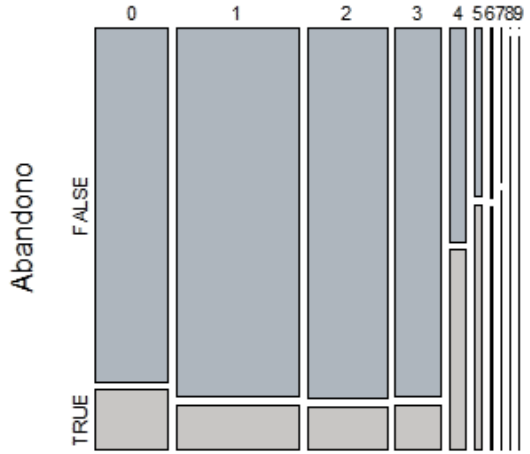


Figura 3.19 Gráfico de la variable "llamadas realizadas al centro de atención al cliente" frente a "cancelación de los servicios"

La Figura 3.18 muestra la relación entre la cancelación de los servicios con la compañía y la contratación de un plan de llamadas de voz internacional por parte de los clientes. Se observa que los clientes que tienen este plan tienden a abandonar la compañía con más frecuencia que los que no lo tienen.

La Figura 3.19 muestra la relación entre el número de llamadas que el cliente hace al centro de atención al cliente de la compañía y la cancelación de los servicios contratados. En el eje horizontal se representa el número de llamadas y en el eje vertical la cancelación de los servicios. La gráfica indica que hay una influencia de esta variable en la decisión del cliente, pues cuanto más llama al centro de atención al cliente, más probable es que cancele los servicios.

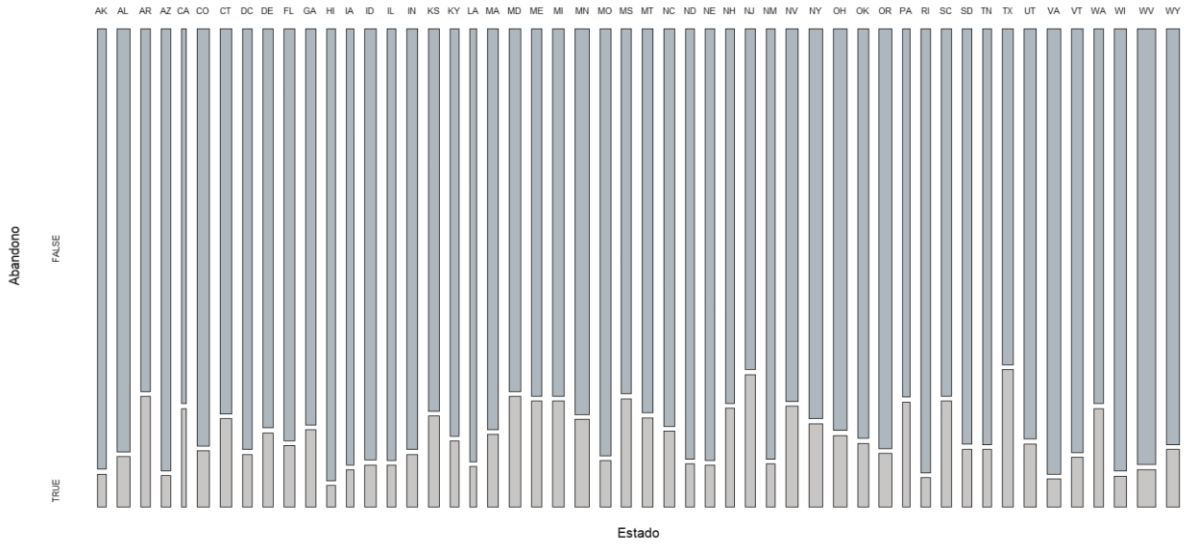


Figura 3.20 Gráfico de la variable "estado" frente a "cancelación de los servicios"

La relación entre el estado de origen del cliente y la decisión de cancelar los servicios se muestra en la Figura 3.20. No se observa una influencia clara de esta variable sobre el comportamiento de los clientes.

Una forma alternativa de visualizar la evolución de los datos es mediante los gráficos de dispersión, que muestran la relación entre dos variables y permiten identificar posibles patrones o tendencias.

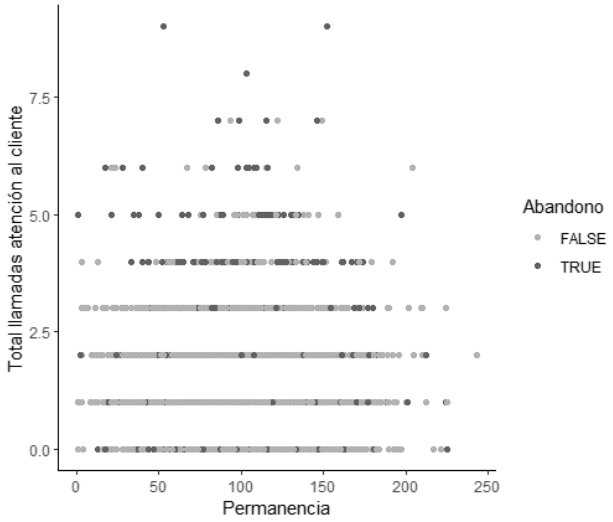


Figura 3.21 Gráfico de dispersión de la variable "duración del contrato" frente a "total de llamadas realizadas al centro de atención al cliente" frente a "cancelación de los servicios contratados"

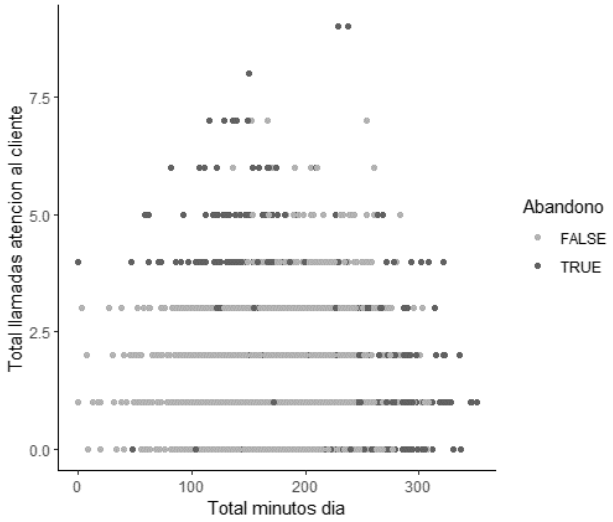


Figura 3.22 Gráfico de dispersión de la variable "total de minutos realizados durante el tramo matinal" frente a "total de llamadas realizadas al centro de atención al cliente" frente a "cancelación de los servicios contratados"

Como se aprecia del análisis de la Figura 3.21 y Figura 3.22, el factor que más influye en el abandono de los clientes es el número de llamadas al servicio de atención al cliente. La proporción de clientes que se dan de baja crece notablemente cuando superan las cinco llamadas, sin importar la otra variable que se analice. Aunque puede verse en la Figura 3.22 como el número de cancelaciones aumenta al sobrepasar los 200 minutos.

3.2 Modelo predictivo a partir de la Regresión Logística Binaria

Un modelo de regresión logística binaria es una herramienta estadística que permite estudiar cómo influyen una o más variables explicativas, que pueden ser numéricas o categóricas, sobre una variable respuesta de tipo binario. La finalidad de la regresión logística binaria es estimar la probabilidad de que la variable respuesta tome el valor 1 en función de los valores de las variables explicativas.

El objetivo de este estudio es analizar los factores que influyen en la decisión de los clientes de darse de baja de los servicios que ofrecen las empresas. Para ello, se utiliza una variable dependiente de tipo binario, que indica si el cliente canceló o no su contrato. Esta variable se codifica con los valores 0 y 1, donde 0 significa que el cliente no canceló y 1 significa que sí lo hizo.

Para estimar la probabilidad de que tome el valor 1, se usa una función logística que convierte la variable dependiente en una escala continua entre 0 y 1. Esta escala representa la probabilidad de que la variable dependiente sea 1. El método de máxima verosimilitud se aplica para calcular los coeficientes de regresión que mejor se ajustan a los datos.

Esta técnica se aplica en diversos campos de investigación, como en medicina, biología, economía y ciencias sociales, para modelar fenómenos que involucran dos opciones posibles.

3.2.1 Construcción del modelo

La generalización del modelo de regresión simple a múltiples variables explicativas (ya sean cualitativas o cuantitativas) se formula mediante la expresión:

$$p(x) = P(Y = 1|\mathrm{X} = \mathrm{x}) = E(Y) = \frac{1}{1 + e^{-x'\beta}}$$

Donde x', es el vector de las variables explicativas y β son los parámetros desconocidos. Desarrollando la ecuación anterior queda:

$$p(x_1, \dots, x_k) = P(Y = 1|x_1, \dots, x_k) = E(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

Las ventajas de respuesta y odds son las funciones exp (β_j), e indican la modificación en las probabilidades del suceso de interés, $Y = 1$, frente a $Y = 0$, por unidad de cambio en las x.

El modelo se puede expresar, mediante transformación logística, como un modelo lineal en función de las variables explicativas:

$$ln\frac{p(x_1, \dots, x_k)}{1 - p(x_1, \dots, x_k)} = x'\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

La función de regresión estimada es, despejando logaritmo:

$$\hat{p}(x_1, \dots, x_k) = \hat{P}(Y = 1|x_1, \dots, x_k) = \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)}} = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)}}$$

En términos de la transformación logística:

$$ln\frac{\hat{p}(x_1, \dots, x_k)}{1 - \hat{p}(x_1, \dots, x_k)} = x'\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

En general, los coeficientes $\hat{\beta}_i$ en el modelo logístico múltiple estima el cambio en los log-odds cuando x_i se incrementa en una unidad, manteniendo constantes el resto de variables explicativas. El antilogaritmo del coeficiente, $e^{\hat{\beta}_i}$, cuando se estima el odds-ratio es:

$$\frac{\frac{P(Y = 1|X_1, \dots, X_i = x_i + 1, \dots X_k)}{1 - (Y = 1|X_1, \dots, X_i = x_i + 1, \dots X_k)}}{\frac{P(Y = 1|X_1, \dots, X_i = x_i, \dots X_k)}{1 - (Y = 1|X_1, \dots, X_i = x_i, \dots X_k)}} = e^{\hat{\beta}_i}$$

Se interpreta como la estimación del porcentaje de incremento si $e^{\widehat{\beta}_i} > 1$ o decremento si $e^{\widehat{\beta}_i} < 1$ en los odds por cada incremento en una unidad en la variable explicativa X_i, manteniendo constantes el resto de variables explicativas. Teniendo en cuenta lo anterior un odds ratio mayor que 1 expresa un cambio positivo, mientras que si es menor que 1 (entre 0 y 1) representa un cambio negativo en las probabilidades estimadas.

El método de estimación utilizado habitualmente para la estimación de los parámetros del modelo logístico es el método de máxima verosimilitud. Empleando la función "glm ( )", función base de R studio, se construye el modelo logístico binario a partir de la muestra con la variable respuesta dicotómica, k variables explicativas y $k + 1$ parámetros que han de ser estimados.

Principalmente se resuelve el modelo empleando el total de variables explicativas de las que se disponen, y resulta un modelo con un gran número de variables poco significativas, evaluadas por el p-valor. Para elegir el modelo que mejor explica a la variable de estudio se utiliza el criterio de Akaike, que mide la calidad del modelo, a menor criterio AIC más calidad tiene.

El modelo que incluye todas las variables resulta un valor **AIC** igual a **1,781.2**.

Otro criterio para seleccionar el modelo que mejor se ajusta, seria "Bayesian Information Criterion", conocido como BIC, el cual penaliza más la entrada de nuevas variables.

El **BIC** resultante para el modelo que contiene a todas las variables es de **2,193.385.**

3.2.2 Elección del mejor modelo

Para encontrar el modelo de mayor calidad, que explique la variable de estudio en función de ciertas variables explicativas, se emplea la función "step ( )", función base de R studio, la cual realiza iteraciones excluyendo del modelo una a una las variables explicativas y calcula el menor criterio de Akaike.

El modelo de mayor calidad obtiene un valor **AIC** de **1,752.3**.

Las variables que componen el modelo de mayor calidad son las siguientes:

- "International plan", que representa la contratación de un plan de llamadas internacionales.
- "Voice mail plan", representa la contratación de un plan de llamadas de voz.
- "Number vmail messages", representa al número de mensajes de texto enviados por el cliente.
- "Total day charge", representa al importe total por los minutos consumidos en el tramo horario matinal.
- "Total night charge", representa al importe total por los minutos consumidos en el tramo horario nocturno.
- "Total intl charge"
- "Customer service calls", representa al número de llamadas realizadas por el cliente al centro de atención al cliente.

Teniendo en cuenta el criterio BIC, para este modelo resulta un valor igual a 1,811.211 .

Para determinar el modelo que mejor se ajusta a los datos se emplean diferentes medidas usadas para comparar modelos alternativos que se ajustan a los mismos datos. Las medidas empleadas son las siguientes:

- **BIC (Bayesian Information Criterion)**, se calcula a partir de la función de verosimilitud del modelo y el número de parámetros libres en el modelo. El BIC penaliza los modelos con un mayor número de parámetros, lo que evita el sobreajuste y favorece los modelos más simples y parsimoniosos. Un valor más bajo indica un mejor ajuste.

$$BIC = -2 * LL + Log(N) * (k + 1)$$

Siendo:

k, es el número de variables explicativas.

LL, es la función de verosimilitud.

N, es el número de observaciones empleadas en la fase de construcción del modelo.

- **AIC (criterio de información de Akaike)**, es una medida que balancea la bondad del ajuste y la complejidad del modelo, se puede utilizar en modelos con diferentes distribuciones de errores y diferentes tamaños de muestra. Un valor más bajo indica un mejor ajuste.

$$AIC = -2\ log\ L\ +\ 2(k+1)$$

- **Pseudo R cuadrado de McFadden**, es una medida relativa que indica cuánta variabilidad en los datos puede explicar el modelo en comparación con un modelo nulo (un modelo sin variables predictoras). Oscila entre 0 y 1. Los valores cercanos a 0 indican que el modelo no tiene poder predictivo.

  Los valores suelen ser bajos, incluso para modelos que se ajustan bien a los datos. Esto se debe a que el modelo nulo es a menudo una predicción muy buena en sí misma. Por lo tanto, es importante no utilizar esta medida como la única medida de bondad de ajuste, sino complementarla con alguna de anteriormente propuestas.

$$R^2\_mf\ =\ 1\ -\ (logL\_modelo\ /\ logL\_nulo)$$

  Donde:

  logL_modelo es el logaritmo natural de la verosimilitud del modelo ajustado.

  logL_nulo es el logaritmo natural de la verosimilitud del modelo nulo.

- **Importancia de las variables**, usando la función "varImp ( )", del paquete "caret" versión 6.0.93 (Kuhn M, 2022), donde los valores más altos indican más importancia. Estos resultados suelen coincidir con los valores p-valor del modelo.
- **VIF (Factor de Inflación de la Varianza)**, es una medida estadística que se utiliza para evaluar la multicolinealidad, variables independientes que están altamente correlacionadas entre sí. Los valores superiores a 5 indican multicolinealidad severa.

Se construye la Tabla 3.2 para comparar varios modelos mediante las medidas mencionadas anteriormente. Para el caso de la importancia de las variables se muestra únicamente las variables con mayor valor.

Churn~.						
BIC	**AIC**	**R^2_mf**	**Importancia de Variables**		**VIF**	
2193.385	1781.2	0.2581	`International plan`	13.37	`Total day minutes` frente a `Total day charge`	3,164.09
					`Total eve minutes` frente a `Total eve charge`	1,492.82
			`Customer service calls`	11.71	`Total night minutes` frente a `Total night charge`	800.59
					`Total intl minutes` frente a `Total intl charge`	263.63

Churn ~ `International plan` + `Voice mail plan` + `Number vmail messages` + `Total day charge` + `Total eve minutes` + `Total night charge` + `Total intl calls` + `Total intl charge` + `Customer service calls`				
BIC	AIC	R^2_mf	Importancia de Variables	VIF
1811.211	1752.3	0.2169	`International plan` 13.18 `Total day charge` 10.39 `Customer service calls` 11.51	`Number vmail messages` 16.59 `Voice mail plan` 16.62

Churn ~ `International plan` + `Voice mail plan` + `Total day charge` + `Total eve minutes` + `Total intl calls` + `Total intl charge` + `Customer service calls`				
BIC	AIC	R^2_mf	Importancia de Variables	VIF
1803.636	1756.5	0.2132	`International plan` 13.18 `Total day charge` 10.40 `Customer service calls` 11.51	∄

Tabla 3.2 Medidas de bondad de ajuste de diferentes modelos

Los dos modelos finales identificados en la Tabla 3.2 son los que presentan un mejor ajuste, según las distintas métricas que se han utilizado. Entre ellos, se elige el último modelo, porque cumple con la hipótesis de multicolinealidad, que es un requisito importante para evitar problemas en las estimaciones del modelo. La multicolinealidad puede provocar que los coeficientes tengan una varianza alta, que no sean significativos o que varíen mucho, y que las estimaciones no sean confiables.

Una forma de obtener los parámetros estimados es usar la función "summary ( )", que es una función genérica de R versión 4.2.2 R Core Team (2022). Esta función resume los resultados del ajuste del modelo y muestra los valores de los coeficientes, sus errores estándar, los valores de prueba z y los p-valores.

Coeficientes	**Estimación** $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}})$	**Error estándar** $(\boldsymbol{S}_{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}})$	$\boldsymbol{z = \frac{(\widehat{\beta}_i)}{(S_{\widehat{\beta}_i})}}$	**p-valor**
Intercept	-6.911514	0.488185	-14.158	2e-16***
International plan Yes	2.091805	0.158659	13.184	2e-16***
Voice mail plan Yes	-0.894689	0.160706	-5.567	2.59e-08***
Total day charge	0.074302	0.007139	10.408	2e-16***
Total eve minutes	0.005472	0.001262	4.334	1.46e-05***
Total intl calls	-0.116978	0.028565	-4.095	4.22e-05***
Total instl charge	0.373817	0.084046	4.448	8.68e-06***
Customer service calls	0.504164	0.043791	11.513	2e-16***

Tabla 3.3 Valores de los coeficientes de los parámetros (modelo 1)

En la Tabla 3.3 se muestran los valores arrojados por la función "summary" para las variables que componen el modelo 1:

- La primera columna muestra el intercept y el nombre de las variables.
- La columna "estimación", indica el valor numérico de los coeficientes asociados a las variables explicativas tomadas en cuenta en el modelo. La interpretación de estos coeficientes depende de si son variables cuantitativas o cualitativas, en general, un coeficiente positivo indica que un aumento en el valor de la variable independiente se asocia con un aumento en la probabilidad de que ocurra el evento (variable dependiente), mientras que un coeficiente negativo indica lo contrario.
- La tercera columna representa el error estandar de las estimaciones, todas son muy cercanas a cero, que es lo ideal.
- La cuarta columna, "z value", corresponde al estadístico Z.
- La última columna corresponde al p-valor del estadístico Z, si el p-valor es menor al nivel de significacion escogido (habitualmente el 5% de significación), entonces el efecto que tiene la variable o categoría de una variable sobre la probabilidad de la respuesta igual a "true" será estadisticamente significativo. Como se aprecia en los resultados obtenidos, todas las variables son significativas a un nivel de significación del 1%.

La función de regresión estimada queda de la siguiente forma:

$$\widehat{\boldsymbol{p}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) = \frac{e^{(-6.9+2.09x_1-0.89x_2+0.07x_3+0.005x_4-0.116x_5+0.373x_6+0.504x_7)}}{1+e^{(-6.9+2.09x_1-0.89x_2+0.07x_3+0.005x_4-0.116x_5+0.373x_6+0.504x_7)}} =$$

$$\widehat{\boldsymbol{p}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) = \frac{1}{1+e^{-(-6.9+2.09x_1-0.89x_2+0.07x_3+0.005x_4-0.116x_5+0.373x_6+0.504x_7)}}$$

A continuación, en la Tabla 3.4 se estudian los coeficientes en forma de probabilidad mediante los odds ratio.

Coeficientes	$(e^{\widehat{\beta}_i})$
Intercept	0.000996248
International plan Yes	8.099521765
Voice mail plan Yes	0.408734709
Total day charge	1.077132081
Total eve minutes	1.005486548
Total intl calls	0.889604690
Total instl charge	1.453270820
Customer service calls	1.655600730

Tabla 3.4 Odds ratio (modelo 1)

Los resultados que se muestran en la Tabla 3.4 se interpretan de la siguiente forma:

- Para valores menores a 1, la asociación es negativa, la probabilidad de abandonar disminuye con el aumento de esa variable.
- Para valores iguales a 1, no existe asociación, es independiente el valor que tome con la respuesta.
- Para valores mayores a 1, la asociación es positiva, un aumento en la variable independiente se asocia con un aumento en la probabilidad del resultado.

En el caso de la variable "Customer service calls", se observa que la asociación es positiva, es decir por cada unidad que aumente la variable, el odds de abandonar se incrementa, en promedio, 1.655.

Cuando son variables categóricas, el resultado indica que la probabilidad es mayor con una categoría de referencia que con la otra. Por ejemplo para el caso de la variable "International plan", la probabilidad de abandonar si el cliente tiene contratado un plan de llamadas internacional es de 8,09 veces más probable que aquellos que no lo tienen contratado.

3.2.3 Validación del modelo

El modelo estimado debe someterse a una validación cruzada, que consiste en aplicar nuevas entradas y definir un valor de umbral o threshold. Este valor es un criterio que permite clasificar cada caso en una de las dos categorías posibles de la variable de estudio. A continuación, se explica cómo elegir el valor óptimo del umbral.

3.2.3.1 Validación cruzada

Mediante la función "predict ()", perteneciente al paquete base de R studio, se realizan las predicciones del modelo. Sin embargo, $\hat{p}(x)$ tiene que ser convertida en una variable discreta dicotómica, $\hat{Y} = 0$ e $\hat{Y} = 1$, de forma que:

- Si $\widehat{\hat{p}_\iota}(x) \geq P^*$ entonces $\hat{Y}_i = 1$
- Si $\widehat{\hat{p}_\iota}(x) < P^*$ entonces $\hat{Y}_i = 0$

Donde P^* es un valor predeterminado, habitualmente $P^* = 0{,}5$. Una vez se estiman las observaciones de $\hat{Y}_i$ se construye una tabla de doble entrada, comparando los valores observados con los estimados.

Para resumir la potencia de la estimación, se tienen en cuenta:

- **Sensibilidad**, es la probabilidad de obtener un **"verdadero positivo"**, probabilidad de que una observación se clasifique correctamente en la clase Y=1 cuando realmente pertenece a esa clase.
- **Especificidad**, es la probabilidad de obtener un **"verdadero negativo".**

Para estudiar el valor de P^*, se construye una gráfica de la sensibilidad en función de (1 – Especificidad) para los posibles umbrales de P^*. Esta gráfica se llama "Receiver Operating Characteristic", más habitualmente conocida por sus siglas como curva ROC.

A mayor área bajo la curva (AUC) mejores son las predicciones del modelo, esta curva resume el valor predictivo para todos los valores de P^*. Se considera que un modelo es bueno si el área bajo la curva está por encima de 0.7 y excelente si está por encima de 0.9.

A continuación se representan en la Ilustración 3.23 e Ilustración 3.24 la curva ROC que resulta de los datos de construcción y de validación respectivamente.

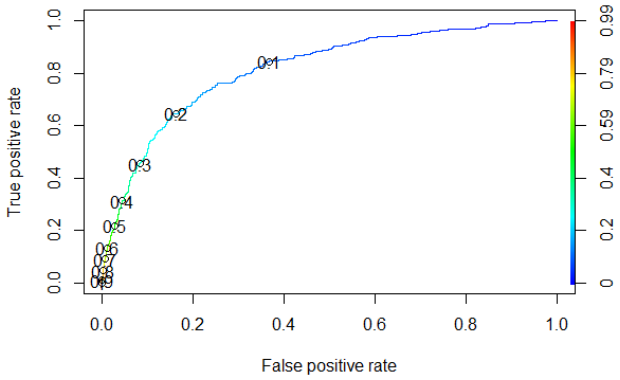


Figura 3.23 Curva ROC. Datos de construcción.

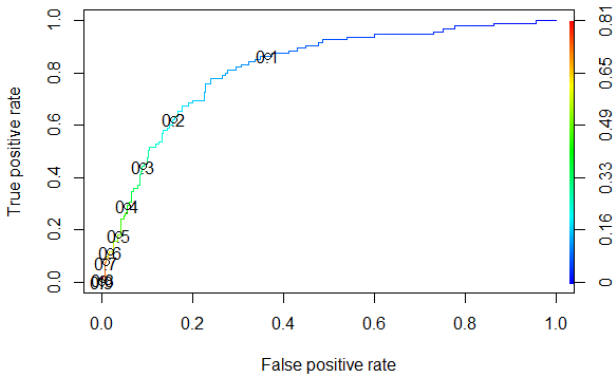


Figura 3.24 Curva ROC. Datos de validación.

Como se observa en la Ilustración 3.23 e Ilustración 3.24 se obtienen muy buenos valores para $P^* = 0.5$ ya que la probabilidad de falso positivo es muy cercana a 0, en ambas gráficas, aunque es conveniente ver los resultados con otros niveles de umbral.

Los resultados muestran que $P^* = 0.5$ es un nivel de umbral óptimo, pues la tasa de falsos positivos se aproxima a 0 en ambos casos. Sin embargo, para tener una visión más completa, se recomienda analizar los datos con otros umbrales distintos.

Una forma de evaluar la calidad del modelo es calcular el área bajo la curva ROC, que representa la relación entre la sensibilidad y la especificidad. En este caso, el valor obtenido es de 0.8186, lo que indica que el modelo tiene un buen desempeño para clasificar correctamente los casos positivos y negativos.

3.2.3.2 Matriz de confusion

En este apartado se presentan distintas tablas que corresponden a diferentes valores de P^*. El objetivo es elegir un valor que tenga una alta probabilidad de detectar correctamente los casos positivos y que al mismo tiempo minimice la probabilidad de clasificar erróneamente los casos negativos.

A continuación se muestra la tabla de clasificación, también conocida como matriz de confusión.

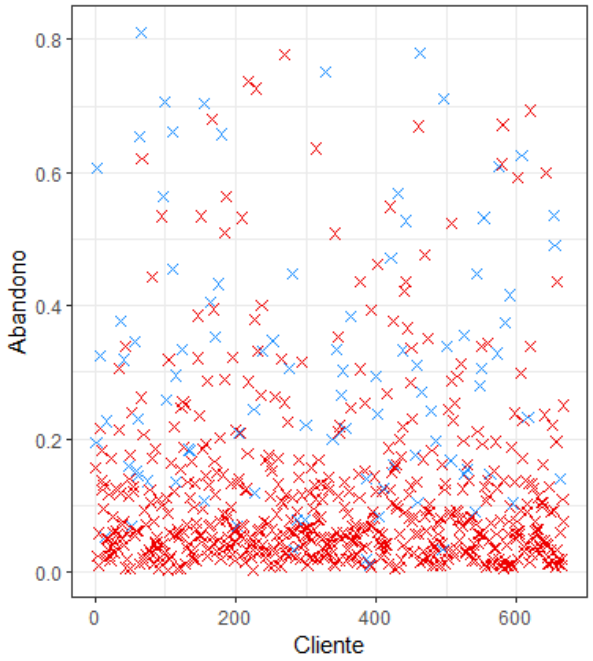
$P^* = 0,5$

Validación con datos de construcción

Predicción	Actual		
	0.8630908	FALSE	TRUE
	FALSE	2215	304
	TRUE	63	84

Validación con nuevos datos

Predicción	Actual		
	0.8530735	FALSE	TRUE
	FALSE	552	78
	TRUE	20	17



El modelo se validó con los datos de construcción y se logró predecir correctamente 2,517 casos, mientras que se cometieron 365 errores.

El objetivo es identificar a los clientes que dejan de usar el servicio. De los 388 casos de abandono registrados, el modelo estima 86 con un nivel $P^* = 0.5$. Esto representa el 22.16% del total de abandonos.

El modelo tiene una tasa de acierto del 86.30%.

El análisis de los datos para el caso de la validación con nuevos datos es el mismo que para el caso anterior. En el gráfico se puede observar que, cuando el valor en el eje vertical es menor a 0.25, todas las observaciones deben de ser de color rojo, lo que indica que los clientes siguen contratando los servicios de la compañía. Por el contrario, cuando el valor es mayor a 0.25, las observaciones deben de ser de color azul, lo que significa que los clientes han cancelado los servicios.

Tabla 3.5 Matriz de Confusión con P = 0.5

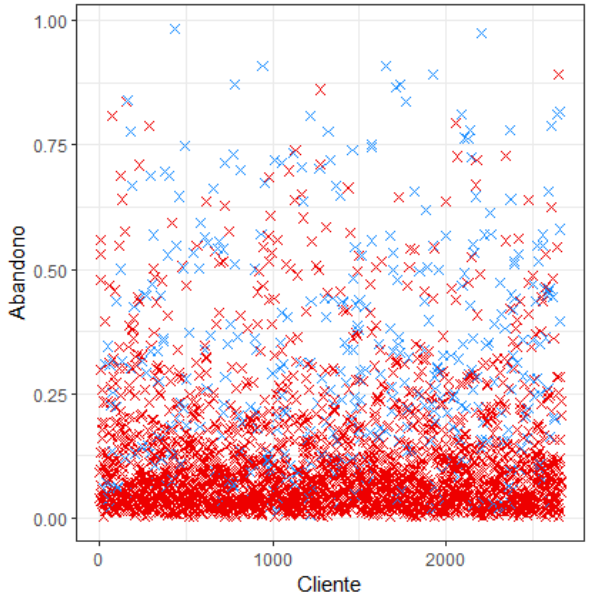
$P^* = 0,1$

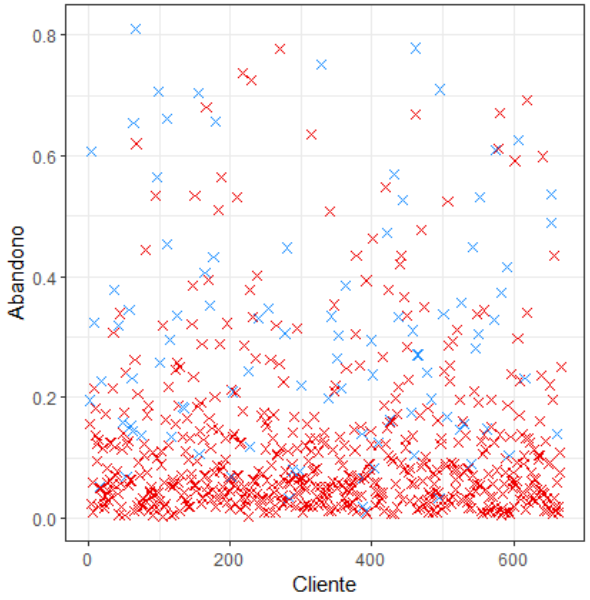
Validación con datos de construcción

	Actual		
Predicción	0.6650413	FALSE	TRUE
	FALSE	1447	61
	TRUE	837	327

Validación con nuevos datos

	Actual		
Predicción	0.6656672	FALSE	TRUE
	FALSE	364	13
	TRUE	208	82





La tabla presenta los resultados obtenidos al aplicar un nivel $P^* = 0.1$. Al validar el modelo con nuevos datos, se observa que el porcentaje de aciertos es de 66.56%. Este valor indica la capacidad predictiva del modelo para este nivel de significancia.

El modelo tiene una precisión de 444 sobre 667 observaciones, lo que significa que acierta en el 66,6% de los casos. Esto implica que el modelo es capaz de predecir correctamente más de la mitad de las veces, pero también tiene un margen de error considerable.

El modelo tiene una precisión del 87,3% al predecir los clientes que abandonan la compañía, pero solo acierta el 63,11% de los casos en los que los clientes se quedan. Esto implica que el modelo tiende a sobreestimar la tasa de cancelación, ya que hay muchos clientes que el modelo clasifica como posibles desertores pero que en realidad no lo son.

El valor de $P^* = 0.1$, que se destaca en la curva ROC, presenta una ventaja y una desventaja. Por un lado, aumenta la probabilidad de detectar correctamente los casos positivos. Por otro lado, también incrementa la probabilidad de clasificar erróneamente los casos negativos.

Tabla 3.6 Matriz de Confusión con P = 0.1

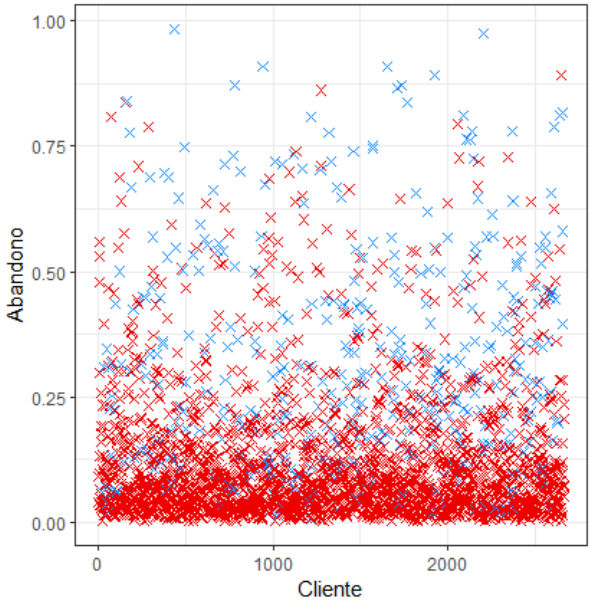
$P^* = 0{,}3$

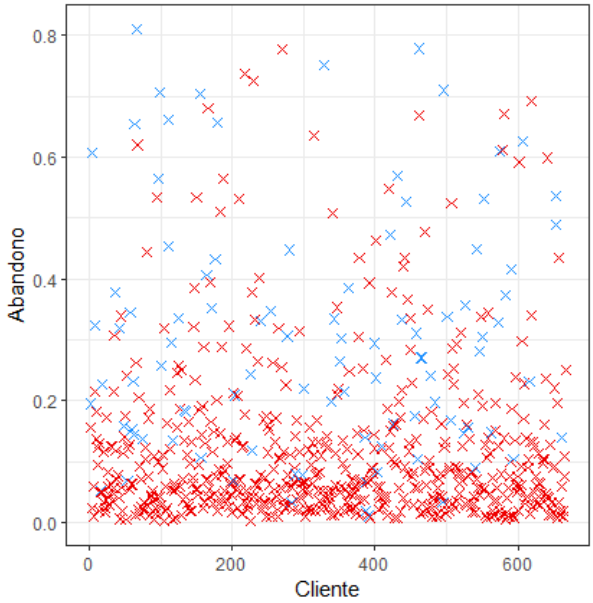
Validación con datos de construcción

	Actual		
Predicción	0.8537134	FALSE	TRUE
	FALSE	2091	212
	TRUE	187	176

Validación con nuevos datos

	Actual		
Predicción	0.8485757	FALSE	TRUE
	FALSE	522	53
	TRUE	50	42





Un valor óptimo para P^*es 0.3, ya que con este parámetro el modelo tiene una precisión de alrededor del 85%. Además, este valor de P^*maximiza la tasa de verdaderos positivos, que se acerca al 0.45 (ver Figura 3.23), y minimiza la tasa de falsos positivos, que se aproxima al 0.1.

Tabla 3.7 Matriz de Confusión con P = 0.3

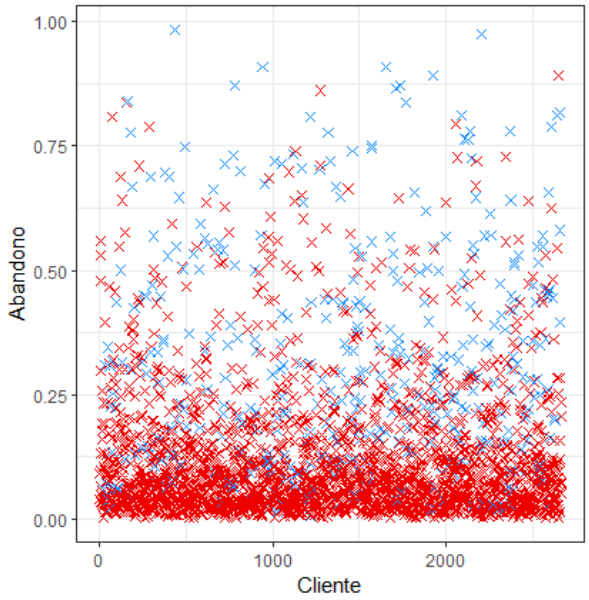
$P^* = 0,2$

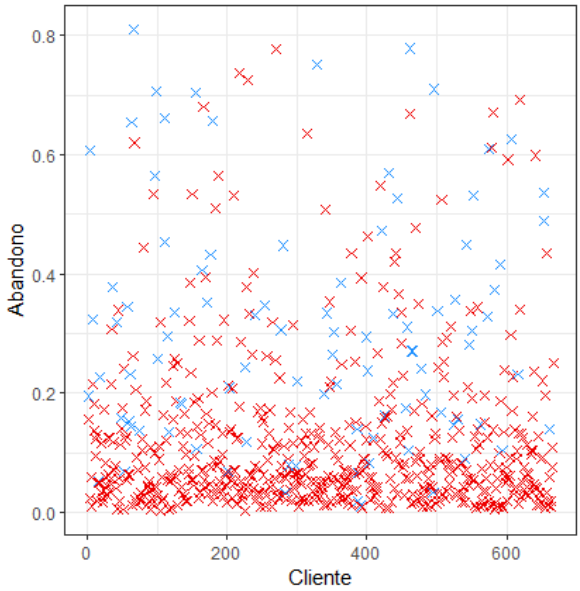
Validación con datos de construcción

Predicción	Actual 0.8098275	FALSE	TRUE
	FALSE	1912	137
	TRUE	366	251

Validación con nuevos datos

Predicción	Actual 0.8125937	FALSE	TRUE
	FALSE	483	36
	TRUE	89	59





Si establecemos un nivel de $P^* = 0.2$, obtenemos una probabilidad de falso positivo que es razonable y una probabilidad de verdadero positivo que aumenta significativamente, tal como se muestra en Ilustración 3.23. Así, la probabilidad de acertar en nuestra predicción es del 80%.

Tabla 3.8 Matriz de Confusión con P = 0.2

El modelo tiene una probabilidad de acierto de alrededor del 85% cuando se toma de referencia un nivel de $P^* = 0.5$ o $P^* = 0.3$, como es el caso de la Tabla 3.5 y Tabla 3.7 Estos son los valores para P^*que arrojan los mejores resultados, según el análisis realizado.

Una forma de evaluar el desempeño de un algoritmo de aprendizaje automático, especialmente cuando se trata de aprendizaje supervisado, es mediante la matriz de contingencia. Esta matriz, también conocida como tabla de confusión, muestra la relación entre los valores reales y los predichos por el algoritmo. En este estudio, se construye una matriz de confusión para los casos en que $P^* = 0.3\ y\ 0.5$.

<table>
<tr><th>Condiciones</th><th>Positiva predicha (y=1)</th><th>Negativa predicha (y=0)</th><td>Total</td><th colspan="2">Estadísticos</th></tr>
<tr><td rowspan="2">Positiva observada (y=1)</td><td rowspan="2">42
Verdadero positivo</td><td rowspan="2">53
Falso negativo</td><td rowspan="2">95</td><td>Sensitividad</td><td>Prevalencia</td></tr>
<tr><td>0.442105</td><td>0.142428</td></tr>
<tr><td rowspan="2">Negativa observada (y=0)</td><td rowspan="2">50
Falso positivo</td><td rowspan="2">522
Verdadero negativo</td><td rowspan="2">572</td><td>Especificidad</td><td>1-Especificidad</td></tr>
<tr><td>0.912587</td><td>0.08741</td></tr>
<tr><td rowspan="2">Total</td><td rowspan="2">92</td><td rowspan="2">575</td><td>Tamaño muestral (n)</td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td>667</td></tr>
<tr><td rowspan="2">Estadísticos</td><td>Valor predictivo positivo</td><td>Valor predictivo negativo</td><td colspan="3" rowspan="2">Exactitud = 0.8455</td></tr>
<tr><td>0.4565</td><td>0.9078</td></tr>
</table>

Tabla 3.9 Matriz de confusión para P*=0.3 (modelo 1)

El modelo ha realizado una clasificación correcta de 42 observaciones como verdaderos positivos y 522 como verdaderos negativos, según se observa en la matriz de confusión representada en la Tabla 3.9. No obstante, también ha cometido errores al clasificar 50 observaciones como falsos positivos y 53 como falsos negativos. Esto implica una exactitud del 84.55% en el modelo.

<table>
<tr><th>Condiciones</th><th>Positiva predicha (y=1)</th><th>Negativa predicha (y=0)</th><td>Total</td><th colspan="2">Estadísticos</th></tr>
<tr><td rowspan="2">Positiva observada (y=1)</td><td rowspan="2">17
Verdadero positivo</td><td rowspan="2">78
Falso negativo</td><td rowspan="2">95</td><td>Sensitividad</td><td>Prevalencia</td></tr>
<tr><td>0.178947</td><td>0.142428</td></tr>
<tr><td rowspan="2">Negativa observada (y=0)</td><td rowspan="2">20
Falso positivo</td><td rowspan="2">552
Verdadero negativo</td><td rowspan="2">572</td><td>Especificidad</td><td>1-Especificidad</td></tr>
<tr><td>0.965034</td><td>0.034965</td></tr>
<tr><td rowspan="2">Total</td><td rowspan="2">37</td><td rowspan="2">630</td><td>Tamaño muestral (n)</td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td>667</td></tr>
<tr><td rowspan="2">Estadísticos</td><td>Valor predictivo positivo</td><td>Valor predictivo negativo</td><td colspan="3" rowspan="2">Exactitud = 0.8080</td></tr>
<tr><td>0.4599</td><td>0.8761</td></tr>
</table>

Tabla 3.10 Matriz de confusión para P*=0.5 (modelo 1)

Para el caso de P*=0.5, el modelo ha realizado una buena clasificación de las observaciones, identificando correctamente a 17 de ellas como positivas y a 522 como negativas. No obstante, también ha cometido algunos errores, asignando la etiqueta positiva a 20 observaciones que eran negativas y la etiqueta negativa a 78 observaciones que eran positivas. Esto implica una exactitud del 80.80% en la clasificación. Pueden verse las métricas en la Tabla 3.10.

Las métricas empleadas en la Tabla 3.9 y 3.10 se desarrollan a continuación.

- La **exactitud** (accuracy) es la proporción de clasificaciones correctas. Se calcula como:

$$Exactitud = \frac{Verdadero\ positivo + Verdadero\ negativo}{Tamaño\ muestral}$$

- La s**ensitividad** (sensivity) es la tasa de verdaderos positivos, aquellos casos que se han predicho positivos y en realidad son positivos.

$$Sensitividad = \frac{Verdadero\ positivo}{Verdadero\ positivo + Falso\ negativo}$$

- La **especificidad** (specifity) es la tasa de verdaderos negativos, aquellas observaciones negativas que se clasifican correctamente.

$$Especificidad = \frac{Verdadero\ negativo}{Falso\ Positivo + Verdadero\ negativo}$$

- La **tasa de falsos positivos** (1-specifity) es la proporción de casos reales negativos que se han clasificado en otra categoría.

$$1 - Especificidad = \frac{Falso\ positivo}{Falso\ Positivo + Verdadero\ negativo}$$

- La **precisión** (precision) es el valor predictivo positivo, la proporción de casos positivos que se clasifican correctamente en la predicción.

$$Precisión = \frac{Verdadero\ Positivo}{Falso\ Positivo + Verdadero\ Positivo}$$

- El **valor predictivo negativo** (negative predictive value), es la proporción de casos negativos clasificados correctamente.

$$Valor\ Predictivo\ Negativo = \frac{Verdadero\ Negativo}{Falso\ Negativo + Verdadero\ Negativo}$$

- La **prevalencia** (prevalence) es la proporción de casos positivos observados.

$$Prevalencia = \frac{Verdadero\ Positivo + Falso\ Negativo}{Tamaño\ Muestral}$$

- La **exactitud balanceada** (balance accuracy), se define como:

$$Exactitud\ Balanceada = \frac{Sensitividad + Especificidad}{2}$$

- La **tasa de detección** (detection prevalence) es la proporción de casos positivos predichos.

$$Prevalencia\ de\ Detección = \frac{Verdadero\ Positivo + Falso\ Positivo}{Tamaño\ Muestral}$$

- **Índice Kappa de Cohen (k)**, es un índice de concordancia muy usado que toma valores entre 0 y 1, siendo uno el máximo acuerdo. Queda definido por la siguiente expresión:

$$k = \frac{Exactitud - P_E}{1 - P_E}$$

$$P_E = \frac{A + B}{Tamaño\ Muestral^2}$$

$$A = (Verdadero\ Positivo + Falso\ Negativo) * (Verdadero\ Positivo + Falso\ Positivo)$$

$$B = (Falso\ Positivo + Verdadero\ Negativo) * (Falso\ Negativo + Verdadero\ Negativo)$$

El índice Kappa de Cohen (k) para el modelo de estudio, con P*=0.3, resulta:

$$P_E = \frac{8930 + 327756}{667^2} = 0.756786$$

$$k = \frac{Exactitud - P_E}{1 - P_E} = 0.377403$$

Teniendo en cuenta la siguiente escala:

$\boldsymbol{k \leq 0.00 \rightarrow}$	**Sin acuerdo**
$\boldsymbol{0.00 \leq k \leq 0.20 \rightarrow}$	**Insignificante**
$\boldsymbol{0.20 \leq k \leq 0.40 \rightarrow}$	**Discreto**
$\boldsymbol{0.40 \leq k \leq 0.60 \rightarrow}$	**Moderado**
$\boldsymbol{0.60 \leq k \leq 0.80 \rightarrow}$	**Sustancial**
$\boldsymbol{0.80 \leq k \leq 1.00 \rightarrow}$	**Casi Perfecto**

Tabla 3.11 Escala índice de concordancia Kappa de Cohen (k)

Resulta un índice de concordancia del modelo discreto según la Tabla 3.11, esto significa que la concordancia observada entre los evaluadores o medidas es mayor que la concordancia esperada al azar, pero todavía existe una diferencia notable entre ellas.

3.2.4 Contraste de hipótesis sobre los parámetros del modelo

Una vez se ha consolidado el modelo en base a la bondad de ajuste del mismo, es importante hacer pruebas de hipótesis sobre los coeficientes de los parámetros que se han estimado en el modelo, para ver si son diferentes de cero de forma significativa.

- **Contraste de Wald.**

  Las hipótesis que se emplean en este contraste son:

  $$H_0: \beta_j = 0$$

  $$H_1: \beta_j \neq 0$$

  A continuación se calcula el estadístico para cada variable y su p-valor para comprobar si se rechaza la hipótesis nula o no se rechaza según los criterios establecidos en la Tabla 3.12. Para valores de p muy pequeños, se rechaza la hipótesis nula, ya que no existe un nivel de significación que quede por debajo del p-valor.

Un p-valor < 0.05 significa que se rechaza la hipótesis nula
Un p-valor >0.05 significa que no se rechaza la hipótesis nula

Tabla 3.12 Interpretación del valor de p

Coeficientes	**Estimación** $(\widehat{\boldsymbol{\beta}}_i)$	**Error estándar** $(\boldsymbol{S}_{\widehat{\beta}_i})$	$\boldsymbol{z = \frac{(\widehat{\beta}_i)}{(S_{\widehat{\beta}_i})}}$	$\boldsymbol{X^2 = \left(\frac{\widehat{\beta}_i}{S_{\widehat{\beta}_i}}\right)^2}$	**p-valor**
Intercept	-6.911514	0.488185	-14.158	200.44	0.0000***
International plan Yes	2.091805	0.158659	13.184	173.81	0.0000***
Voice mail plan Yes	-0.894689	0.160706	-5.567	30.99	2.591743e-08***
Total day charge	0.074302	0.007139	10.408	108.32	0.0000***
Total eve minutes	0.005472	0.001262	4.334	18.78	1.464285e-05***
Total intl calls	-0.116978	0.028565	-4.095	16.76	4.22174e-05***
Total instl charge	0.373817	0.084046	4.448	19.78	8.667371e-06***
Customer service calls	0.504164	0.043791	11.513	132.54	0.0000***

Tabla 3.13 Contraste de Wald para los parámetros de los coeficientes (modelo 1)

Todos los coeficientes del modelo son significativos al 1% de significación y su valor es distinto de 0.

- **Contraste condicional de razón de verosimilitud.**

  Se usa para contrastar si el modelo nulo es mejor que el modelo ajustado, permite evaluar la bondad de ajuste del modelo contrastando de forma global los parámetros de las variables, calculando directamente la log-verosimulitud de cada modelo.

  $$H_0: Modelo\ nulo\ mejor$$

  $$H_1: Modelo\ ajustado\ mejor$$

  El valor obtenido es 488.461, calculado el valor de p asociado, resulta 0. Por lo tanto se rechaza la hipótesis nula H_0 al 1% de significación bajo los criterios explicados en Tabla 3.12.

3.2.5 Intervalos de confianza

- **Intervalos de confianza basados en el estadístico de Wald**.

  Se calculan los intervalos de confianza para los parámetros $\hat{\beta}_k$, con un nivel de confianza del 95%.

Coeficientes	**Estimación** $(\widehat{\boldsymbol{\beta}}_i)$	**Intervalo de Confianza**
Intercept	-6.911514	$[-7.885325054 , -5.970551623]$
International plan Yes	2.091805	$[1.781691641 , 2.404202841]$
Voice mail plan Yes	-0.894689	$[-1.217280839 , -0.586454324]$
Total day charge	0.074302	$[0.060464498 , 0.088464567]$
Total eve minutes	0.005472	$[0.003007741 , 0.007959133]$
Total intl calls	-0.116978	$[-0.174059967 , -0.062057801]$
Total instl charge	0.373817	$[0.210248783 , 0.539902636]$
Customer service calls	0.504164	$[0.418878888 , 0.590694244]$

Tabla 3.14 Intervalos de confianza para los parámetros del modelo

Para los intervalos de confianza calculados en Tabla 3.14 se ha tenido en cuenta un nivel de confianza del 95%, esto quiere decir que existe una probabilidad del 95% de que el valor de los parámetros poblacionales se encuentre en ese intervalo.

La precisión de la muestra para estimar el parámetro poblacional se refleja en el ancho del intervalo de confianza. A menor ancho, mayor precisión. A mayor ancho, menor precisión. Los intervalos calculados son relativamente estrechos, lo que indica una buena estimación. Además, ninguno de los intervalos incluye al 0.

- **Intervalos de confianza para los cocientes de las ventajas**.

  Para estimar el efecto de cada variable explicativa sobre la respuesta, se calculan los cocientes de las ventajas y sus intervalos de confianza al 95%. Sin embargo, como el modelo incluye más de una variable, los parámetros no son independientes y esto afecta a los intervalos de confianza. Por lo tanto, los intervalos que se obtienen son:

Coeficientes	$(e^{\widehat{\beta}_i})$	**Intervalo de Confianza**
Intercept	0.000996248	$[0.0003762243 , 0.002552833]$
International plan Yes	8.099521765	$[5.9398960977 , 11.069602526]$
Voice mail plan Yes	0.408734709	$[0.2960340377 , 0.556296238]$
Total day charge	1.077132081	$[1.0623298816 , 1.092495542]$
Total eve minutes	1.005486548	$[1.0030122690 , 1.007990891]$

Total intl calls	0.889604690	[0.8402465088 , 0.939828562]
Total instl charge	1.453270820	[1.2339850168 , 1.715839793]
Customer service calls	1.655600730	[1.5202562222 , 1.805241258]

Tabla 3.15 Intervalos de confianza de los cocientes de las ventajas

Cuanto más pequeños sean los intervalos de confianza calculados en Tabla 3.15, mayor será la precisión de la estimación del parámetro y más significativo será el efecto de la variable independiente sobre la variable dependiente. Por el contrario, si el intervalo de confianza es grande, la precisión de la estimación del parámetro se reduce y el efecto de la variable independiente puede no ser significativo.

- **Intervalos de Wilson**, se estima la proporción poblacional de abandonos a partir de la muestra. Mediante un contraste de hipótesis:

$$H_0: Proporción\ de\ abandonos\ igual\ a\ 0.5$$
$$H_1: Proporción\ de\ abandonos\ distinto\ a\ 0.5$$

p-valor	2.2e-16***
Proporción	0.1455364
Intervalo de Confianza (95%)	[0.1324802 , 0.1596269]

Se rechaza la hipótesis nula a un nivel de significación del 1%. En la tabla anterior se muestra el valor de p, a partir del cual se rechaza la hipótesis nula del contraste, la proporción resultante y el intervalo de confianza a un nivel de significación del 95% para la proporción.

3.2.6 Diagnóstico del modelo

El diagnóstico del modelo estimado se centra en el análisis de los residuos. Suponiendo que el modelo es válido, la distribución de los residuos tendrá media cero y varianza unidad. Se definen distintos tipos de residuos que permiten evaluar la bondad de ajuste del modelo y ayudan a identificar puntos atípicos.

- **Residuos de la variable respuesta.**

  El residuo se define por la siguiente expresión:

$$e_i = y_i - \hat{p}_{1/i}$$

  Se muestra una tabla con la cantidad de residuos que en valor absoluto son mayor a dos, ya que es el umbral para concluir la significación del residuo.

FALSE	TRUE
2573	93

Tabla 3.16 Residuos significativos

- **Residuos de Pearson.**

  Corresponde al residuo tipificado, se calcula dividiendo el residuo de la variable respuesta calculado anteriormente entre la estimación de la desviación típica. Se obtiene una tabla igual a la Tabla 3.16.

  Los residuos de Pearson se consideran significativos si son grandes en comparación con su desviación estándar esperada. La desviación estándar esperada de los residuos de Pearson es de aproximadamente 2. Si un residuo de Pearson tiene una magnitud mucho mayor que 2, puede indicar que hay una discrepancia significativa entre los datos observados y los valores predichos por el modelo.

A continuación, se emplea diferentes medidas que cuantifican el cambio en algún estadístico crítico del modelo cuando se elimina un dato del conjunto de observaciones. Con estas medidas se estudian los datos influyentes en el modelo.

- **Distancia de Cook.**

  Mide el cambio en los coeficientes estimados tras prescindir de la i-ésima observación. Aquellos valores mayores a la unidad se consideran influyentes. Visualizando los resultados de forma gráfica en Tabla 3.17, se observa que no existen distancias mayores a la unidad, por lo tanto no influye ninguna observación al modelo.

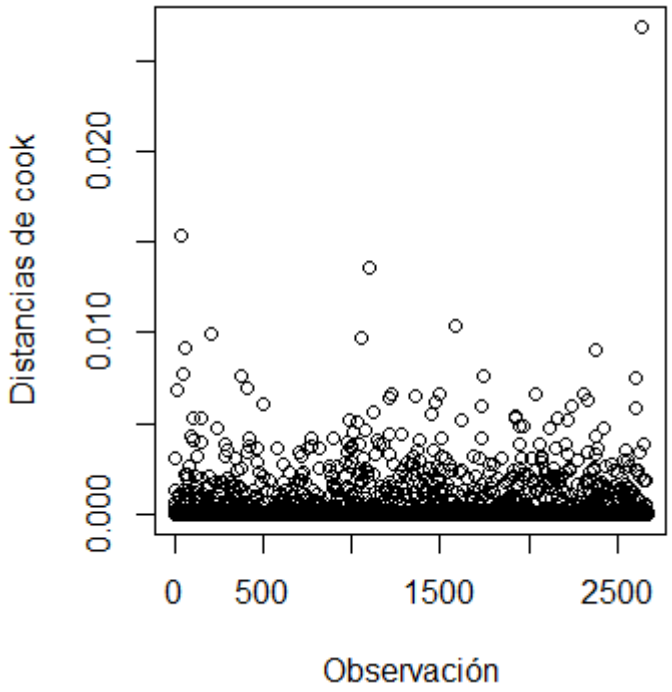


Tabla 3.17 Distancias de Cook

- **DFBETA.**

  Esta medida junto con otras se emplea para determinar si una observación es un valor atípico o un punto de datos influyente. Mide los cambios en los coeficientes estimados cuando se elimina una de las observaciones.

  En general, se considera que un valor DFBETA es significativo si su magnitud en valor absoluto es mayor que 2 dividido por la raíz cuadrada del número de observaciones en el modelo. Se considera que la observación correspondiente tiene una influencia significativa en el coeficiente de regresión correspondiente. La magnitud de referencia para el modelo toma un valor igual a 0.03873.

FALSE	TRUE
19458	1870

Tabla 3.18 DFBETA

- **DFFITS.**

  Mide los cambios en el valor ajustado de la observación i-ésima cuando se prescinde de dicha observación. Considerando una observación con alta influencia si el valor absoluto de la medida es mayor a dos veces la raíz cuadrada del total de parámetros entre el número de observaciones.

  El valor límite se establece para este modelo en 0.122489

FALSE	TRUE
2341	325

Tabla 3.19 DFFITS

A continuación en la Figura 3.25 se pueden observar gráficas para todas aquellas variables cuantitativas, se representa en el eje horizontal el logit. Se define como el logaritmo natural de la razón entre la probabilidad de que ocurra un evento y la probabilidad de que no ocurra el evento. En el eje vertical se muestra el valor predicho .

La línea diagonal en la gráfica representa la igualdad perfecta entre el valor predicho y el logit. Los puntos que están por encima de la línea diagonal indican que el valor predicho es mayor que el valor de logit correspondiente, mientras que los puntos que quedan por debajo de la línea diagonal indican lo contrario.

En general, los puntos en la gráfica están distribuidos en una línea diagonal, lo que sugiere que el modelo es consistente en sus predicciones. Sin embargo, hay algunos puntos que están bastante alejados, lo que sugiere que el modelo no es perfecto en sus predicciones.

En general, se puede decir que la gráfica muestra una buena correspondencia entre el logit y el valor predicho, lo que sugiere que el modelo es adecuado para predecir la variable de estudio.

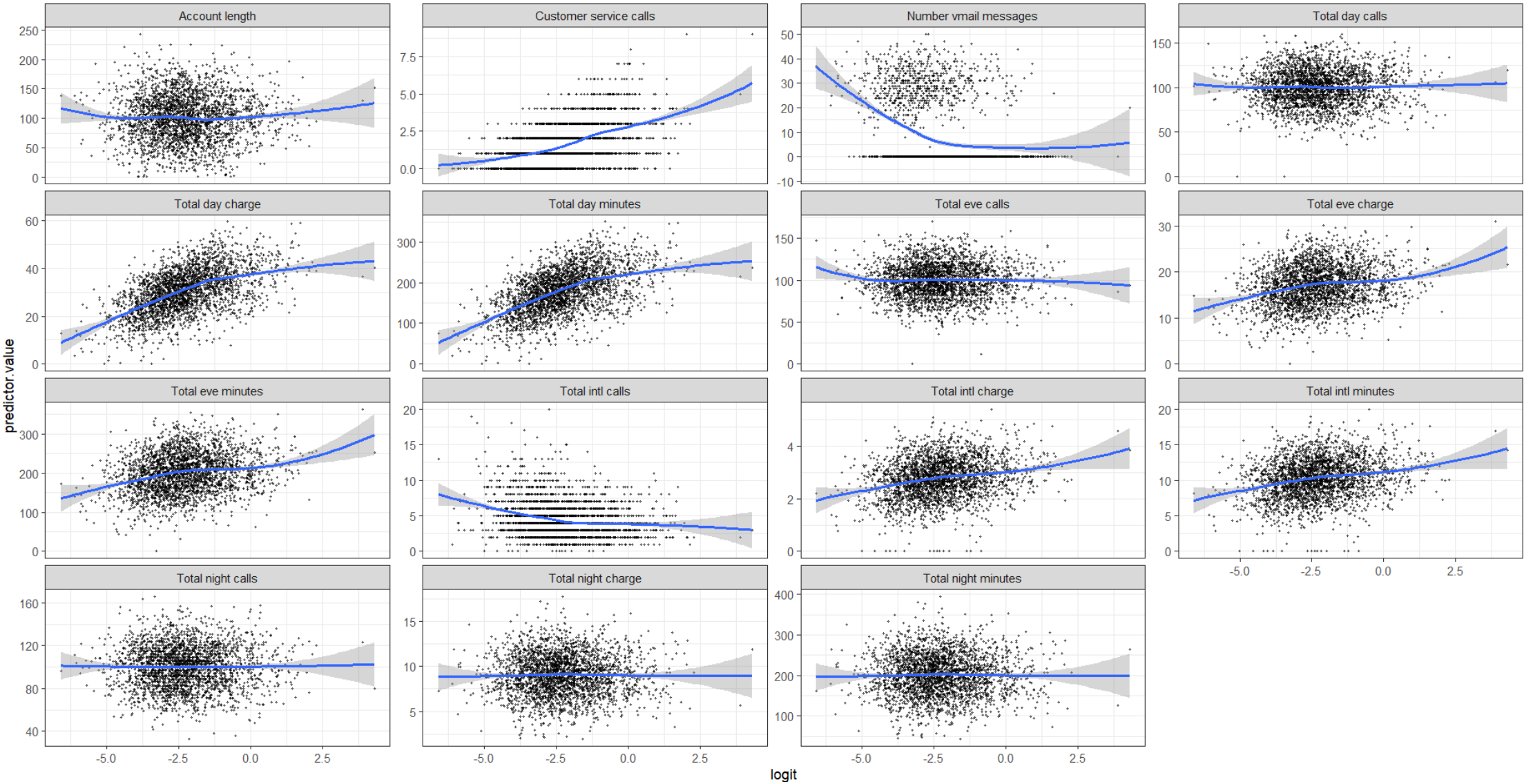
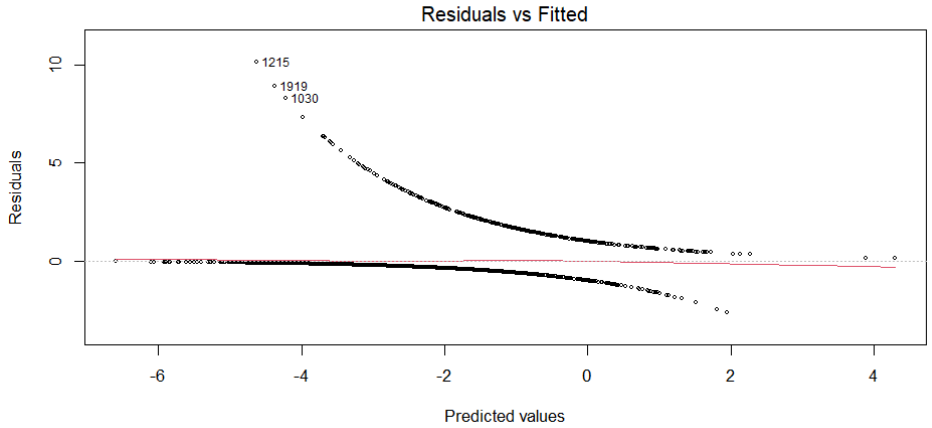


Figura 3.25 Diagrama de dispersión entre cada predictor y los valores logit

Las suposiciones del modelo se han validado a lo largo de los apartados, para estudiar la linealidad del modelo se estudian los diagramas de dispersión entre cada predictor y los valores logit, como puede verse en Figura 3.25. De estas gráficas es conveniente centrarse en aquellas variables predictoras que forman parte del modelo ajustado (Tabla 3.3).

Las curvas resultantes indican linealidad en el modelo, aunque para todas ellas existen algunos puntos que están bastante alejados de la línea diagonal.

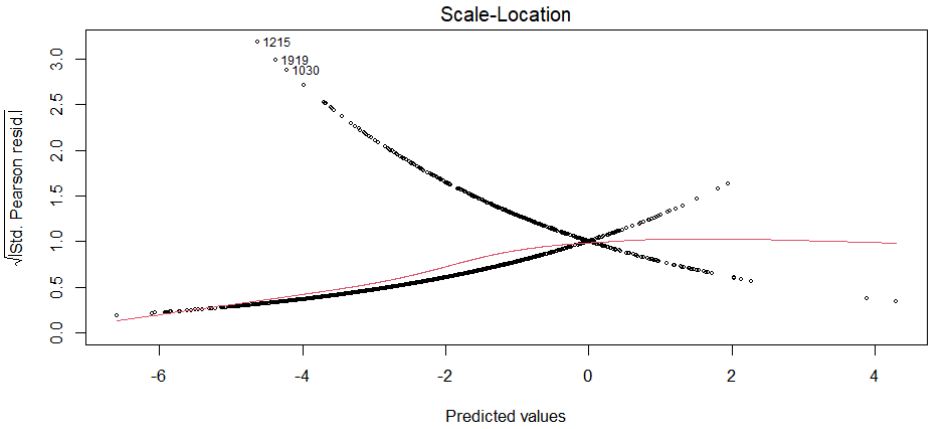


Esta gráfica representa los residuos (diferencia entre el valor observado y el ajustado) versus los valores ajustados por el modelo (probabilidad de que la variable explicada sea igual a uno), con el fin de evaluar la calidad de ajuste del mismo.

En este tipo de gráficos aplicados a modelos de regresión logística, se buscan patrones de valores atípicos o valores extremos en los residuos que puedan indicar que el modelo no está capturando adecuadamente alguna característica importante de los datos.

La gráfica señala la observación 1,215 , 1,030 y 2,307 , como valores extremos.

Este gráfico de dispersión ayuda a comprobar la hipótesis de linealidad del modelo, ya que los residuos de Pearson se distribuyen aleatoriamente alrededor de cero, sin patrones obvios, la curva no toma forma de U. El coeficiente de correlación entre las predicciones y los residuos de Pearson, resulta con un valor igual a -0.005101501, muy cercano a 0. Por lo tanto, no se puede decir que exista evidencia de no linealidad en el modelo ajustado.

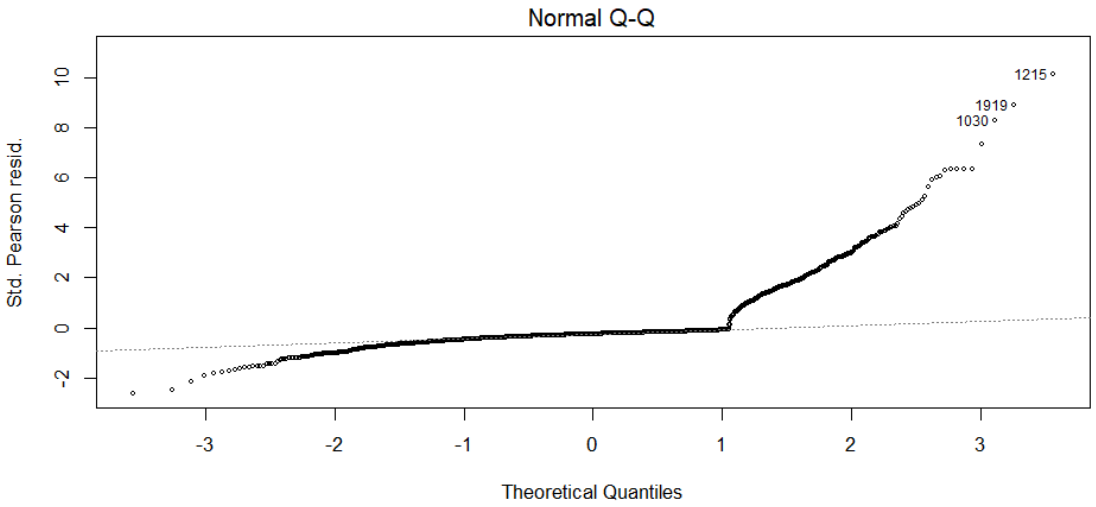


La segunda gráfica compara los valores predichos con la raíz de los residuos estandarizados de Pearson, se utiliza igual que la anterior para evaluar la bondad de ajuste del modelo.

La raíz de los residuos estandarizados de Pearson representa la magnitud de la discrepancia entre los valores observados y los valores predichos.

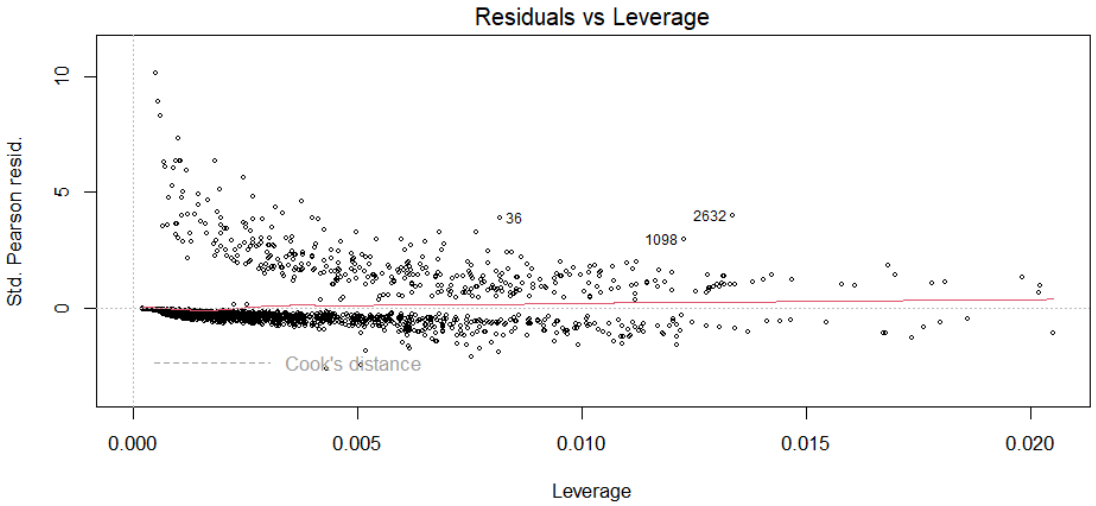
Si el modelo se ajusta bien a los datos, se espera que los puntos en la gráfica estén distribuidos aleatoriamente alrededor de una línea diagonal, lo que indica que la magnitud de los residuos no está correlacionada con los valores predichos.

Es importante tener en cuenta que esta gráfica no proporciona información sobre la precisión absoluta del modelo. En cambio, puede ayudar a evaluar la distribución de los errores del modelo.



Esta gráfica compara las cantidades teóricas y los residuos estandarizados de Pearson, se le conoce como gráfico Q-Q, este gráfico se utiliza para ver si los residuos del modelo siguen una distribución normal.

En el caso de un modelo de regresión logística binaria, los puntos se alinearán aproximadamente en una línea diagonal.



Este es un gráfico de influencia, puede ayudar a identificar observaciones influyentes en la estimación de los parámetros del modelo.

El apalancamiento (leverage) mide la influencia de una observación en la estimación de los coeficientes de regresión. Las observaciones con valores de apalancamiento más altos tienen más influencia en la estimación de los coeficientes.

Cada punto representa una observación en el conjunto de datos. Las observaciones con valores de apalancamiento más altos y residuos estandarizados de Pearson más altos que el resto de las observaciones pueden ser puntos atípicos o puntos influyentes en el modelo.

Las observaciones marcadas en la gráfica son la 1,475 , 2,632 y 36. Estos puntos deben ser examinados cuidadosamente para determinar si representan valores anómalos legítimos o errores en los datos. Deben ser examinadas con más detalle para determinar su impacto en los resultados del modelo.

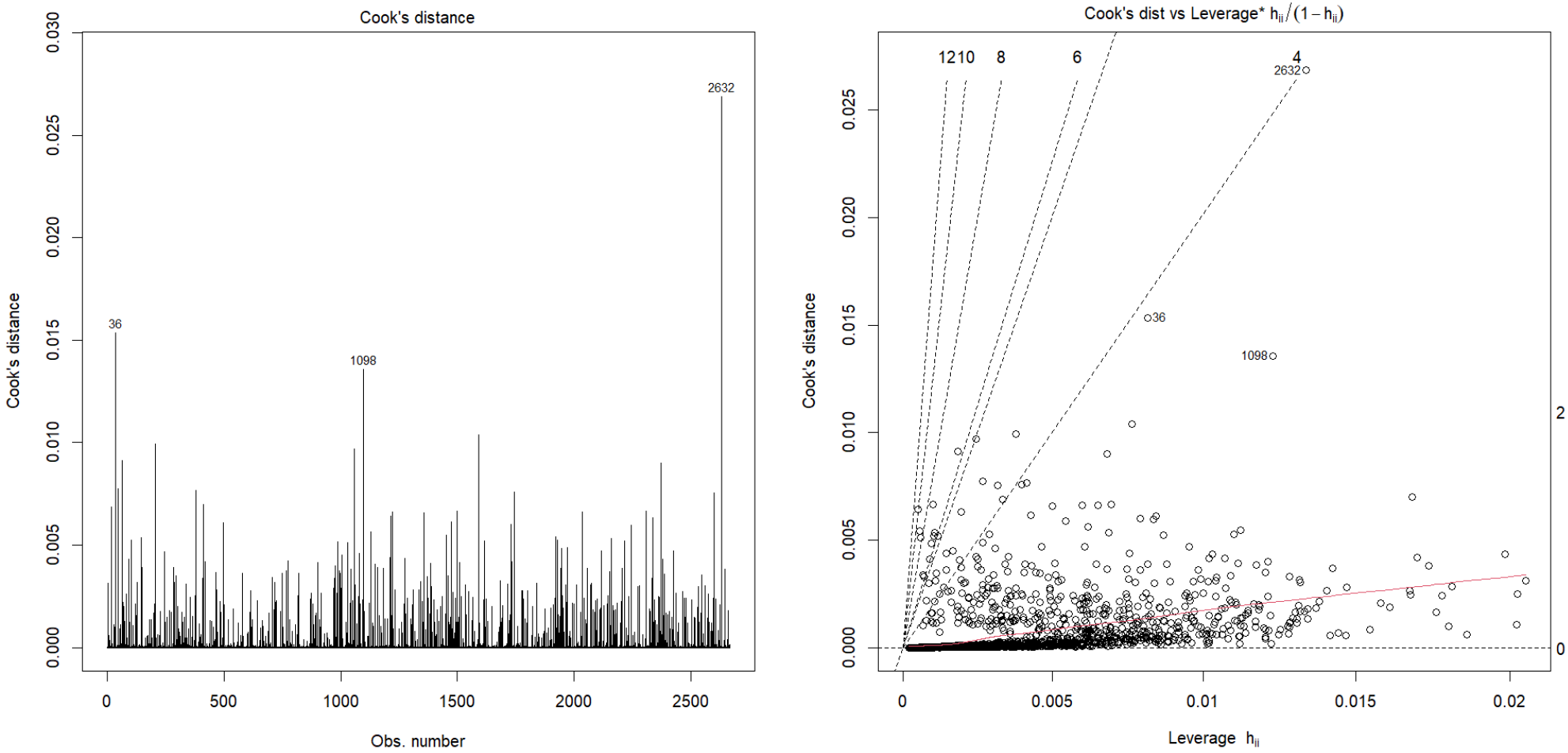


Figura 3.26 Distancias de Cook frente a Leverage

En la Figura 3.26 se presenta en primer lugar las distancias de cook calculados para cada observación, aquellas con mayor distancia de Cook son las observaciones 36, 1,475 y 2,632.

En segundo lugar se muestra una gráfica que compara el apalancamiento y la distancia de Cook, conocido como gráfico de la influencia de Cook. El apalancamiento mide la influencia de una observación en la estimación de los coeficientes de regresión, mientras que la distancia de Cook mide el cambio en los coeficientes de regresión cuando se elimina la observación.

Los coeficientes de regresión pueden verse afectados por las observaciones que se muestran en los gráficos anteriores, lo que puede alterar los resultados del modelo de manera significativa. Es importante determinar si una observación influyente corresponde a un valor atípico válido, y en ese caso, considerar si se debe incorporar al modelo.

3.2.7 Estudio de los datos influyentes

Es necesario analizar los puntos atípicos que se han identificado previamente, para decidir si conviene excluirlos del modelo o si conviene conservarlos para aumentar la precisión y eficacia de las predicciones. Dependiendo del caso, la exclusión de puntos influyentes puede beneficiar la precisión del modelo, o puede ser conveniente mantenerlos para obtener una predicción más fidedigna.

Las observaciones identificadas como atípicas son:

1,215
1,919
1,030

Estas observaciones han sido revisadas y no corresponden a fallos en el proceso de recolección del dato.

Los datos arrojados como influyentes por los gráficos anteriores son:

36
1,098
2,632

Para estos puntos se estudia un nuevo modelo que no contenga esas observaciones, para realizar una comparación con el modelo ajustado en Tabla 3.3.

Los nuevos parámetros del modelo resultante son:

Coeficientes	**Estimación** $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}})$	**Error estándar** $(\boldsymbol{S}_{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}})$	$\boldsymbol{z} = \frac{(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}})}{(\boldsymbol{S}_{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}})}$	**p-valor**	**Exp** $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}})$
Intercept	-6.831422	0.489468	-13.957	2e-16***	0.001079322
International plan Yes	2.076219	0.159643	13.005	2e-16***	7.974258121
Voice mail plan Yes	-0.929577	0.162693	-5.714	1.11e-08***	0.394720492
Total day charge	0.074661	0.007178	10.402	2e-16***	1.077518479
Total eve minutes	0.005503	0.001268	4.339	1.43e-05***	1.005518533
Total intl calls	-0.146844	0.029826	-4.923	8.51e-07***	0.863428961
Total instl charge	0.378170	0.084414	4.480	7.47e-06***	1.459611264
Customer service calls	0.508720	0.044017	11.557	2e-16***	1.663161051

Tabla 3.20 Parámetros ajustados (modelo 2)

$$P^* = 0,5$$

Modelo 1

Validación con datos de construcción

	Actual		
Predicción	0.8630908	FALSE	TRUE
	FALSE	2215	304
	TRUE	63	84

Validación con nuevos datos

	Actual		
Predicción	0.8530735	FALSE	TRUE
	FALSE	552	78
	TRUE	20	17

Modelo 2 (sin observaciones influyentes)

Validación con datos de construcción

	Actual		
Predicción	0.8640631	FALSE	TRUE
	FALSE	2217	301
	TRUE	61	84

Validación con nuevos datos

	Actual		
Predicción	0.8530735	FALSE	TRUE
	FALSE	552	78
	TRUE	20	17

Tabla 3.21 Modelo 1 frente a Modelo 2

La precisión de la clasificación de los casos no muestra diferencias significativas entre los distintos modelos utilizados como puede verse en la Tabla 3.21.

Una forma de comparar la calidad de dos modelos es mediante el contraste de razón de verosimilitud, que mide la diferencia entre la log-verosimilitud de cada modelo. Este contraste evalúa si los parámetros de las variables son significativos para explicar el modelo, y si se puede simplificar el modelo eliminando algunas variables.

$$H_0: Modelo\ 1\ explica\ mejor\ el\ abandono\ del\ cliente$$
$$H_1: Modelo\ 2\ explica\ mejor\ el\ abandono\ del\ cliente$$

El estadístico asociado al valor p, se calcula como se muestra:

$$Lv = -2 * logLink(modelo_2) + 2 * logLink\left(modelo_{Ajustado}\right) = 16.8094$$
$$\mathrm{P_{valor}} = 1 - pchisq\left(abs(Lv), df = df_{modelo_{wald}} - df_{modelo}\right) = 0.0007734799$$

Con un valor de p tan pequeño se rechaza la hipótesis nula, por lo tanto el modelo que mejor predice el abandono del cliente, es el modelo que no contiene a las observaciones influyentes y queda de la siguiente forma:

$$\widehat{\boldsymbol{p}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) = \frac{1}{1 + e^{-(-6.831 + 2.076x_1 - 0.929x_2 + 0.074x_3 + 0.005x_4 - 0.146x_5 + 0.378x_6 + 0.508x_7)}}$$

El significado de los coeficientes del modelo 2 (Tabla 3.20) se interpretan según la Tabla 3.22:

Coeficientes	**Estimación** $(\widehat{\boldsymbol{\beta}}_i)$	**Exp** $(\widehat{\boldsymbol{\beta}}_i)$	**Interpretación de los coeficientes del modelo**
Intercept	-6.831422	0.001079322	La probabilidad del abandono de un cliente sin considerar el resto de variables es 0.001
International plan Yes	2.076219	7.974258121	La probabilidad de abandonar perteneciendo a la categoría yes es 7.97 veces mayor que perteneciendo a la categoría no.
Voice mail plan Yes	-0.929577	0.394720492	La probabilidad de abandonar perteneciendo a la categoría yes es 0.394 veces mayor que perteneciendo a la categoría no.
Total day charge	0.074661	1.077518479	El incremento de probabilidad de abandonar de un cliente con un aumento en una unidad en esta variable manteniendo el resto de variables es de 1.077 veces mayor.
Total eve minutes	0.005503	1.005518533	El incremento de probabilidad de abandonar de un cliente con un aumento en una unidad en esta variable manteniendo el resto de variables es de 1.005 veces mayor.
Total intl calls	-0.146844	0.863428961	El incremento de probabilidad de abandonar de un cliente con una disminución en una unidad en esta variable manteniendo el resto de variables es de 0.863 veces mayor.
Total instl charge	0.378170	1.459611264	El incremento de probabilidad de abandonar de un cliente con un aumento en una unidad en esta variable manteniendo el resto de variables es de 1.459 veces mayor.
Customer service calls	0.508720	1.663161051	El incremento de probabilidad de abandonar de un cliente con un aumento en una unidad en esta variable manteniendo el resto de variables es de 1.663 veces mayor.

Tabla 3.22 Interpretación de los coeficientes (modelo 2)

3.3 Modelo predictivo a partir de árboles de clasificación

El árbol de decisión es una técnica de aprendizaje automático supervisado que se usa para clasificar variables categóricas. Consiste en crear un árbol donde cada nodo representa una pregunta o un criterio que permite separar el conjunto de datos en subconjuntos más homogéneos. De esta manera, se puede asignar una categoría a cada observación según el camino que siga en el árbol.

El algoritmo construye el árbol de clasificación eligiendo la variable predictora que mejor distingue las observaciones en categorías diferentes. Luego, repite este proceso de división de forma recursiva hasta alcanzar un criterio de parada.

El árbol de clasificación sirve para estimar la categoría de una nueva observación. Para hacerlo, se debe seguir el proceso de propagación, que consiste en responder a las preguntas de cada nodo del árbol y avanzar por el camino correspondiente hasta llegar a una hoja final. Esta hoja final indica la categoría que se predice para la observación.

En este apartado, se sigue el mismo procedimiento que en los anteriores: se ajustan y se comparan varios modelos, y se elige el que mejor predice la variable de estudio. Para los árboles de clasificación, se construyen tres modelos basados en árboles de decisión usando diferentes variables explicativas.

Usando todas las variables disponibles, se construye un árbol de decisión (árbol 1). Mediante la función "rpart ( )" del paquete con el mismo nombre versión 4.1.19 (Terry Therneau and Beth Atkinson 2022). La matriz de confusión que resulta de aplicar este modelo a los datos es la siguiente:

<table>
<tr><th>Condiciones</th><th>Positiva predicha (y=1)</th><th>Negativa predicha (y=0)</th><td>Total</td><th colspan="2">Estadísticos</th></tr>
<tr><td rowspan="2">Positiva observada (y=1)</td><td rowspan="2">68
Verdadero positivo</td><td rowspan="2">27
Falso negativo</td><td rowspan="2">95</td><td>Sensitividad</td><td>Prevalencia</td></tr>
<tr><td>0.7158</td><td>0.1424</td></tr>
<tr><td rowspan="2">Negativa observada (y=0)</td><td rowspan="2">16
Falso positivo</td><td rowspan="2">556
Verdadero negativo</td><td rowspan="2">572</td><td>Especificidad</td><td>1-Especificidad</td></tr>
<tr><td>0.9720</td><td>0.0279</td></tr>
<tr><td rowspan="2">Total</td><td rowspan="2">84</td><td rowspan="2">583</td><td>Tamaño muestral (n)</td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td>667</td></tr>
<tr><td rowspan="2">Estadísticos</td><td>Valor predictivo positivo</td><td>Valor predictivo negativo</td><td colspan="3" rowspan="2">Exactitud = 0.9355
Kappa = 0.7227</td></tr>
<tr><td>0.8095</td><td>0.9536</td></tr>
</table>

Tabla 3.23 Matriz de confusión (árbol 1)

Para identificar las variables predictoras más relevantes para el modelo, se utilizó la función "varImp ( )" del paquete "caret" versión 6.0-93 (Kuhn M, 2022), que permite evaluar la importancia de cada variable a partir del árbol de clasificación generado previamente. Con esta información, se construyó un nuevo árbol de clasificación que solo incluye las variables con mayor peso en el modelo.

"International plan"
"State"
"Total day minutes"
"Total day charge"
"Total eve minutes"
"Total eve charge"
"Total intl minutes"
"Total intl charge"
"Customer service calls"

Como se identificó anteriormente existe correlación entre varias de las variables resultantes, por lo tanto se suprimen aquellas variables correlacionadas del modelo y se emplean las siguientes variables para la construcción del árbol (árbol 2):

"International plan"
"State"
"Total day charge"
"Total eve charge"
"Total intl charge"
"Customer service calls"

Condiciones	**Positiva predicha (y=1)**	**Negativa predicha (y=0)**	Total	**Estadísticos**	
Positiva observada (y=1)	**60 Verdadero positivo**	**35 Falso negativo**	95	Sensitividad	Prevalencia
				0.6316	0.1424
Negativa observada (y=0)	**25 Falso positivo**	**547 Verdadero negativo**	572	Especificidad	1-Especificidad
				0.9563	0.0437
Total	85	582	Tamaño muestral (n)		
			667		
Estadísticos	Valor predictivo positivo	Valor predictivo negativo	**Exactitud = 0.91 Kappa = 0.6149**		
	0.7059	0.9399			

Tabla 3.24 Matriz de confusión (árbol 2)

Ahora se analiza el árbol de clasificación construido a partir de las variables que componen el modelo de regresión logística binaria, estas variables pueden verse en la Tabla 3.22.

Condiciones	**Positiva predicha (y=1)**	**Negativa predicha (y=0)**	Total	**Estadísticos**	
Positiva observada (y=1)	**72 Verdadero positivo**	**23 Falso negativo**	95	Sensitividad	Prevalencia
				0.7579	0.1424
Negativa observada (y=0)	**6 Falso positivo**	**566 Verdadero negativo**	572	Especificidad	1-Especificidad
				0.9895	0.0104
Total	78	589	Tamaño muestral (n)		
			667		
Estadísticos	Valor predictivo positivo	Valor predictivo negativo	**Exactitud = 0.9565 Kappa = 0.8077**		
	0.9231	0.9610			

Tabla 3.25 Matriz de confusión (árbol 3)

De la Tabla 3.25 se puede observar:

- El modelo de clasificación que se basa en las variables del modelo de regresión logística muestra un alto rendimiento, con una exactitud del 95.65% y un índice kappa de 0.8077. Estos valores indican que el modelo tiene una concordancia casi perfecta.

- Tanto el valor predictivo negativo como el positivo muestran una excelente capacidad para clasificar correctamente los casos, alcanzando más del 90% de acierto en ambos indicadores.

- El modelo tiene una sensitividad del 75.79%, lo que significa que de todos los clientes que se dan de baja de la compañía, solo el 24.21% son clasificados erróneamente como clientes que se quedan. Estos son los falsos negativos, que representan a los clientes que se van pero el modelo no los detecta.

- El objetivo de este estudio es identificar con mayor precisión a los clientes que se dan de baja del servicio, sin que aumente significativamente el número de falsos positivos. Estos son los casos en los que el modelo predice que el cliente se va a dar de baja, pero en realidad no lo hace. Para lograr este equilibrio, el modelo se ajusta a los requisitos establecidos. Así, se consigue un valor predictivo positivo del 92.31% y una tasa de falsos positivos del 1.04%.

- El índice de concordancia de kappa se ve mejorado con respecto a los otros modelos.

A continuación puede verse la representación del árbol de clasificación (árbol 3) en figura 3.28.

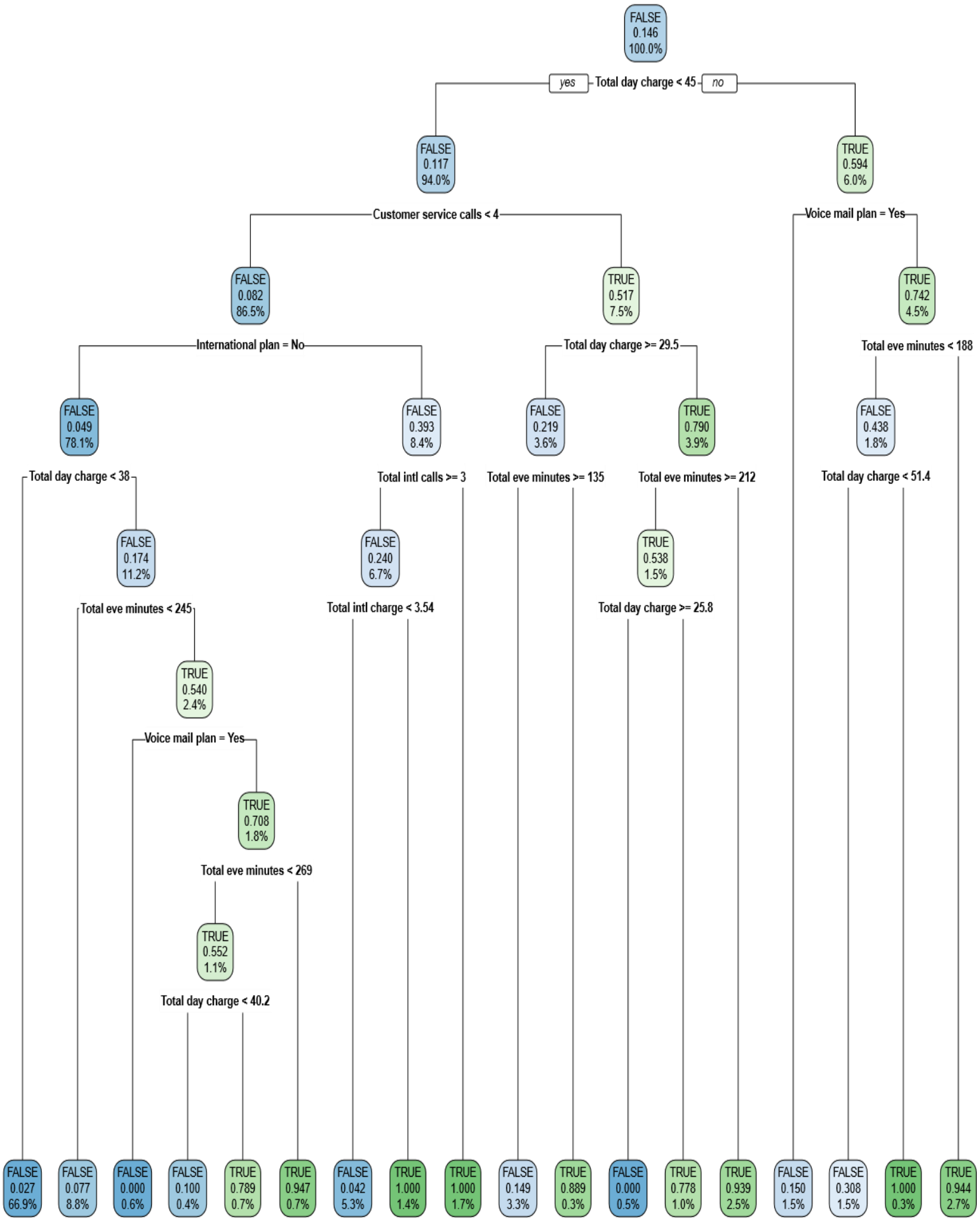


Figura 3.27 Árbol de clasificación (árbol 3)

La forma de interpretar la Figura 3.27 es la siguiente:

- El árbol consiste en nodos que plantean una cuestión sobre una variable del conjunto de datos, y en hojas que indican una predicción para una instancia.
- Para recorrer el árbol, se empieza por el nodo inicial y se plantea la pregunta que contiene. Dependiendo de si la respuesta es afirmativa o negativa, se sigue por la rama izquierda o derecha, respectivamente. Así se va avanzando de nodo en nodo hasta llegar a una hoja, que indica el final del recorrido.
- El modelo hace una predicción para cada instancia basándose en la hoja del árbol a la que llega. Luego, asigna la misma categoría que tienen la mayoría de las instancias en esa hoja a cualquier nueva instancia que llegue a la misma hoja.
- Cada nodo muestra información importante sobre la construcción del árbol y la toma de decisiones del modelo. Muestra la proporción de casos que pertenecen a la categoría y la proporción de datos que han sido agrupados en ese nodo. Para la hoja inferior izquierda, el 2.7% pertenece a la categoría "true", (100 – 2.7)% pertenecen a la categoría "false", y la proporción de datos de la muestra agrupados en ese nodo ha sido del 66.9%.
- Se ha obtenido un resultado perfecto de clasificación con dos hojas del árbol de decisión, que se corresponden con los nodos terminales de color verde más oscuro.

4 CONCLUSIONES

Para una compañía telefónica, es esencial retener a sus clientes, logrando que se sientan satisfechos y sean fieles a los servicios de la compañía. El proyecto que se ha desarrollado tiene dos metas principales. En primer lugar entender las razones por las que los clientes se dan de baja y en segundo lugar predecir con precisión cuáles son los clientes que se irán de la compañía. Para ello, se han utilizado dos técnicas muy populares en el campo del Machine Learning.

La regresión logística tiene como inconveniente que sus resultados son difíciles de interpretar, a diferencia de los árboles de decisión que se entienden fácilmente. Sin embargo, los árboles de decisión también tienen sus desventajas, como el alto coste en términos de tamaño de la muestra. Esto se debe a que al dividir los datos según cada predictor, el tamaño de la muestra se reduce.

Para determinar qué factores influyen en la cancelación de los servicios por parte de los clientes, se ha utilizado la regresión logística, que es un método adecuado para analizar variables de estudio con dos posibles respuestas. Sin embargo, para hacer predicciones sobre el abandono del cliente, se ha optado por los árboles de decisión, que han mostrado una mayor precisión. Así, se ha seleccionado las variables explicativas más relevantes mediante la regresión logística y se han empleado en la construcción del árbol de decisión, cuyas métricas de predicción se pueden observar en la Tabla 4.1.

MÉTRICA / MODELO	Exactitud	Kappa	Sensitividad	Prevalencia	Tasa de detección	Especificidad	Tasa de falsos positivos	Valor predictivo negativo	Valor predictivo positivo
Regresión logística (modelo 2)	0.8530	0.1927	0.1789	0.1424	0.055	0.9650	0.0349	0.8761	0.459
Árbol de decisión (árbol 3)	0.9565	0.8077	0.7579	0.1424	0.1169	0.9895	0.0104	0.9610	0.9231

Tabla 4.1 Métricas definitivas

La prevalencia indica el total de casos positivos en la muestra, y tiene un valor del 14.24%. Los modelos propuestos para detectar casos positivos tuvieron diferentes tasas de éxito. El árbol de decisión fue el más eficaz, con un 11.69%.

El modelo propuesto empleando arboles de decisión es el que mejor responde al objetivo de la problemática planteada, ya que predice con un 92.31% de precisión qué clientes van a cancelar los servicios. Estos clientes son los que requieren de políticas de retención personalizadas para asegurar su fidelidad a la compañía. El modelo a partir de regresión logística, en cambio, tiene una baja capacidad para identificar a estos clientes, con un valor predictivo positivo de solo el 45.90%, a pesar de tener una buena exactitud general. Por lo tanto, el árbol 3 supera al modelo 2 en las métricas más relevantes para el caso de estudio.

El modelo de regresión logística nos permite identificar las variables que influyen en el abandono del cliente y diseñar estrategias de retención de clientes adecuadas a cada caso. Algunas sugerencias de políticas para prevenir la pérdida de clientes son:

- Una forma de facilitar la comunicación internacional de nuestros clientes es ofrecerles planes de llamadas adaptados a sus necesidades y preferencias, que les permitan mantener el contacto con sus seres queridos en otros países sin tener que preocuparse por el precio o el tiempo de las llamadas. Estos planes cuentan con diversas opciones, como cuotas fijas, paquetes de minutos, ofertas por horarios o destinos, etc.

- Una forma de satisfacer las demandas de comunicación de los clientes que hacen muchas llamadas locales o nacionales es ofrecer planes de voz con minutos ilimitados o abundantes. Estos planes pueden tener beneficios como llamadas sin costo entre líneas de la misma empresa, números preferidos, servicios extra.

- Establecer un sistema de tarificación justo y transparente, que no genere cargos ocultos o sorpresas en la factura.

- Para que el cliente no se sienta frustrado e insatisfecho cuando llama a la compañía, hay que reducir el tiempo de espera para ser atendido. Si no se logra, el cliente puede decidir cambiar a otra compañía que le brinde un mejor servicio. Para lograr este objetivo, se debe optimizar los recursos humanos y técnicos del centro de atención al cliente, y también implementar sistemas que mejoren la gestión de colas, la atención personalizada o la devolución de llamada.

- Mejorar la calidad y la eficacia de las llamadas a atención al cliente, asegurando que el cliente reciba una respuesta rápida y satisfactoria a sus consultas, reclamaciones o sugerencias. Para ello, es necesario capacitar y motivar al personal encargado de esta función, así como dotarlo de las herramientas y los protocolos adecuados. Además, se debe medir y evaluar la satisfacción del cliente con este servicio y tomar medidas correctivas si fuera necesario.

- Fomentar el uso de las llamadas internacionales como una forma de diferenciación y valor añadido. Para ello, se puede crear una campaña de marketing que resalte los beneficios de comunicarse con otros países. Asimismo, se puede incentivar el consumo de estas llamadas mediante sorteos, premios o regalos.

- Ajustar el cargo internacional al costo real de las llamadas internacionales, evitando cobrar más de lo debido o aplicar recargos injustificados. Este aspecto puede influir en la percepción del cliente sobre la calidad y la honestidad de la compañía. Para lograrlo, es necesario negociar con los operadores internacionales las mejores tarifas posibles y trasladarlas al cliente final.

- Realizar un seguimiento de la experiencia del cliente, mediante encuestas, feedbacks o llamadas, para conocer su grado de satisfacción con el servicio, detectar posibles problemas o áreas de mejora, y ofrecer soluciones o compensaciones .

- Premiar la lealtad de los clientes con descuentos en la factura, ofertas y promociones en la adquisición de nuevos terminales, o programas de puntos o recompensas que se puedan canjear por productos o servicios .

- Generar un vínculo emocional con los clientes, mediante una comunicación cercana y personalizada, que les haga sentir parte de la comunidad de la marca, y que les motive a recomendarla a otros potenciales clientes.

Es importante actualizar estos análisis con frecuencia, por ejemplo anualmente, para adaptarse a los cambios en los datos y evitar que el modelo predictivo se vuelva obsoleto. Además, dependiendo de los datos, puede existir un modelo más adecuado que otro para la predicción, por lo que se recomienda hacer una evaluación periódica.

Realizar estos análisis es muy sencillo hoy en día y cualquier comercio puede emplear los datos recopilados en el CRM de la empresa para desarrollar análisis sobre sus clientes y elevar sus ingresos anticipándose a las necesidades de los mismos. `

Como trabajo futuro, se propone explorar otras técnicas de modelado predictivo, como redes neuronales o máquinas de vectores de soporte, que podrían mejorar el rendimiento y la precisión de la clasificación de los clientes. Asimismo, se sugiere incorporar más variables explicativas al análisis, que representen fielmente las características actuales de los servicios ofertados por las compañías de telefonía móvil. Como el uso de servicios de streaming, fibra óptica, datos móviles, entre otros, que podrían aportar más información sobre los factores que influyen en el abandono. Finalmente, se recomienda realizar un seguimiento periódico de los modelos desarrollados, para evaluar su robustez y estabilidad ante posibles cambios en el comportamiento de los clientes o en el mercado de la telefonía móvil.

Referencias

Adhikary, D. D., & Gupta, D. (2021). Applying over 100 classifiers for churn prediction in telecom companies. Multimedia Tools and Applications, 80(28–29), 35123–35144. https://doi.org/10.1007/s11042-020-09658-z

Ahn, J.-H., Han, S.-P., & Lee, Y.-S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications Policy, 30(10–11), 552–568. https://doi.org/10.1016/j.telpol.2006.09.006

Applying over 100 classifiers for churn prediction in telecom companies" (2020) por Debjyoti Das Adhikary1 & Deepak Gupta. (s/f).

Beeharry, Y., & Tsokizep Fokone, R. (2022). Hybrid approach using Machine Learning algorithms for customers' churn prediction in the telecommunications industry. Concurrency and Computation: Practice & Experience, 34(4). https://doi.org/10.1002/cpe.6627

Bock, T. (2018, octubre 24). Decision trees are usually better than logistic regression. Displayr. https://www.displayr.com/decision-trees-are-usually-better-than-logistic-regression/

Capítulo 8 Modelos logísticos. (s/f). Github.io. Recuperado el 8 de junio de 2023, de https://arcruz0.github.io/libroadp/logit.html

Churn prediction in telecommunication using data mining technology. International journal of advanced computer science and applications. (s/f).

Churn Prediction Modeling in Mobile Telecommunications Industry Using Decision Trees. (s/f).

Churn Prediction Modeling in Mobile Telecommunications Industry Using Decision Trees. (2013).

Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees. (s/f).

Customer Churn Prediction in Influencer Commerce: an application of decisión trees" (2022) por Sulim Kima. (s/f). Heeseok Leeb.

Cutanda Henríquez, F. (2008). Datos anómalos y regresión logística robusta en ciencias de la salud. Revista espanola de salud publica, 82(6), 617–625. https://doi.org/10.1590/s1135-57272008000600003

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. 2016 Symposium on Colossal Data Analysis and Networking (CDAN).

david_becks. (2017). Churn in Telecom's dataset [Data set].

Diaz-Aviles, E., Pinelli, F., Lynch, K., Nabi, Z., Gkoufas, Y., Bouillet, E., Calabrese, F., Coughlan, E., Holland, P., & Salzwedel, J. (2015). Towards real-time customer experience prediction for telecommunication operators. 2015 IEEE International Conference on Big Data (Big Data).

Hybrid approach using Machine Learning algorithms for customers churn prediction in the telecommunications industry" (2022) por Beeharry Y. (s/f). Tsokizep Fokone R.

J, R., & T., U. (2011). Churn prediction in telecommunication using data mining technology. International journal of advanced computer science and applications : IJACSA, 2(2). https://doi.org/10.14569/ijacsa.2011.020204

Khan, Y., Shafiq, S., Naeem, A., Ahmed, S., Safwan, N., & Hussain, S. (2019). Customers churn prediction using artificial neural networks (ANN) in telecom industry. International Journal of Advanced Computer Science and Applications : IJACSA, 10(9). https://doi.org/10.14569/ijacsa.2019.0100918

Lee, J., Lee, J., & Feick, L. (2001). The impact of switching costs on the customer satisfaction-loyalty link: mobile phone service in France. Journal of Services Marketing, 15(1), 35–48. https://doi.org/10.1108/08876040110381463

Malyar, M., Mykola Robotyshyn, M. V., & Sharkadi, M. (2020). Churn prediction estimation based on Machine Learning methods. 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC).

Muñoz Villalba, J. F., Lidon Lopez, I., & Rebollar Rubio, R. (2019). Diseño y validación de una herramienta predictiva de accidentes laborales en las obras de construcción. Dyna, 94(1), 88–93. https://doi.org/10.6036/8794

Operadores móviles: cuota de mercado de suscripciones EE. UU. T1 de 2011 - T2 de 2015. (s/f). Statista. Recuperado el 26 de junio de 2023, de https://es.statista.com/estadisticas/633997/operadores-moviles-cuota-de-mercado-de-suscripciones-ee-uu-t1-de-2011-t2-de/

Petkovski, A. J., Risteska Stojkoska, B. L., Trivodaliev, K. V., & Kalajdziski, S. A. (2016). Analysis of churn prediction: A case study on telecommunication services in Macedonia. 2016 24th Telecommunications Forum (TELFOR).

Por Aleksandar, J., Biljana, L. R., Stojkoska, K. V., & Trivodaliev, S. A. (2016). Analysis of Churn Prediction: A Case Study on Telecommunication Services in Macedonia.

Por, T., Xu, Y., & Ma, K. (2021). Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping.

RPubs - 27.6. Regresión logística binaria (matriz de confusión). (s/f). Rpubs.com. Recuperado el 8 de junio de 2023, de https://rpubs.com/hllinas/R_Logit_Binario_Confusion

Saha, L., Tripathy, H. K., Gaber, T., El-Gohary, H., & El-Kenawy, E.-S. M. (2023). Churn Prediction Method for Telecommunication Industry". Sustainability, 15.

Solano Dávila, O., Ramírez Torres, A., Bartolo Gotarate, F. M., Giraldo Laguna, O., & Salinas Moreno, A. (2014). Análisis de diagnóstico en el modelo de regresión logística: una aplicación. Pesquimat, 10(1). https://doi.org/10.15381/pes.v10i1.9431

Therneau T, Atkinson B (2022). _rpart: Recursive Partitioning and Regression Trees_. R package version 4.1.19, <https://CRAN.R-project.org/package=rpart>.

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research, 157(1), 196–217. https://doi.org/10.1016/s0377-2217(03)00069-9

Wei, C.-P., & Chiu, I.-T. (2002). Turning telecommunications call details to churn prediction: a data mining approach. Expert Systems with Applications, 23(2), 103–112. https://doi.org/10.1016/s0957-4174(02)00030-1

wilson.ci function - RDocumentation. (s/f). Rdocumentation.org. Recuperado el 8 de junio de 2023, de https://www.rdocumentation.org/packages/fastR2/versions/1.2.2/topics/wilson.ci

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), 5445–5449. https://doi.org/10.1016/j.eswa.2008.06.121

Xu, T., Ma, Y., & Kim, K. (2021). Telecom churn prediction system based on ensemble learning using feature grouping. Applied Sciences (Basel, Switzerland), 11(11), 4742. https://doi.org/10.3390/app11114742

ANEXO I

###Estudio de variables

```
#Importar datos de construcción del modelo
library(readr)
data <- read_csv("~/TFM/churn-bigml-80.csv")
View(data)

#Convertir variables categóricas
data$State <- as.factor(data$State)
data$`Area code` <- as.factor(data$`Area code`)
data$`International plan` <- as.factor(data$`International plan`)
data$`Voice mail plan` <- as.factor(data$`Voice mail plan`)
data$Churn <- as.factor(data$Churn)

#Importar datos de validación del modelo
library(readr)
datos_validacion <- read_csv("~/TFM/churn-bigml-20.csv")
View(datos_validacion)

#Convertir variables categóricas
datos_validacion$State <- as.factor(datos_validacion$State)
datos_validacion$`Area code` <- as.factor(datos_validacion$`Area code`)
datos_validacion$`International plan` <- as.factor(datos_validacion$`International plan`)
datos_validacion$`Voice mail plan` <- as.factor(datos_validacion$`Voice mail plan`)
datos_validacion$Churn <- as.factor(datos_validacion$Churn)

#Explorar datos
str(data)

#Analizar estadísticos básicos
summary(data)

#Histograma para variables numéricas
hist(data$`Customer service calls`,main="",cex.main=1,col="#934962",breaks = 10,
```

```
xlab="Total llamadas atención al cliente", ylab="Frecuencia")

#Histograma en función de la variable abandono
ggplot(data = data, mapping = aes(x = `Customer service calls`, fill = factor(Churn))) +
geom_histogram(bins = 9, position = 'identity', alpha = 0.8) + labs(title = , fill = 'Abandono', x = 'Total llamadas atención al cliente', y = 'Frecuencia', subtitle = , caption = )

#Estudiar correlación entre variable numéricas
cor(datos_numericos)

#Representar gráfico de correlación
library(corrplot)
correlacion<-round(cor(datos_numericos), 1)
corrplot(correlacion, method="number", type="upper")

library(corrplot)
correlations <- cor(datos_numericos)
corrplot(correlations, method="circle")

#Representar variables categóricas en función de variable abandono
Plan_internacional <- table(data$`International plan`, as.factor(data$Churn))
plot(Plan_internacional)
mosaicplot(Plan_internacional, xlab="Plan internacional", ylab="Abandono",
main="",color=c("#9FB6CD","#CDC5BF","#9FB6CD","#CDC5BF"))

Plan_llamadas <- table(data$`Customer service calls`, as.factor(data$Churn))
plot(Plan_llamadas)
mosaicplot(Plan_llamadas, xlab="Llamadas Atención al Cliente", ylab="Abandono",
main="",color=c("#9FB6CD","#CDC5BF","#9FB6CD","#CDC5BF"))

Estado <- table(data$State, as.factor(data$Churn))
plot(Plan_llamadas)
mosaicplot(Estado, xlab="Estado", ylab="Abandono",
main="",color=c("#9FB6CD","#CDC5BF","#9FB6CD","#CDC5BF"))

Codigo_area <- table(data$`Area code`, as.factor(data$Churn))
plot(Plan_llamadas)
mosaicplot(Codigo_area, xlab="Codigo de area", ylab="Abandono",
main="",color=c("#9FB6CD","#CDC5BF","#9FB6CD","#CDC5BF"))
```

#Gráfico de dispersión

```
library(ggplot2)
Abandono<- data$Churn
ggplot(data=datos_numericos)+ geom_point(mapping=aes(x=datos_numericos$data..Total.day.minutes., y=datos_numericos$data..Customer.service.calls., color= Abandono))+ xlab("Total minutos dia")+ ylab("Total llamadas atencion al cliente")+ theme_classic()
```

#Gráfico de caja y bigotes

```
par(mfrow=c(1,8))
for(i in 1:8) {   boxplot(datos_numericos[,i], main=names(datos_numericos)[i]) }
```

#Matriz de dispersión

```
pairs(data, col=data$Churn)
```

###Construcción del modelo de regresión logística

#Modelo con todas las variables explicativas

```
mod.est <- glm(Churn~., family=binomial(logit), data=data)
```

#Estadísticos del modelo

```
summary(mod.est)
BIC(mod.est)
pR2(mod.est)
varImp(mod.est)
vif.logit(mod.est)
vif.log
```

#Selección del mejor modelo

```
modelo_mejor <- step(mod.est, direction= "backward")
```

#Coeficientes en forma de exponente

```
exp(modelo_mejor$coefficients)
```

#Predicción del modelo con datos de validación

```
pred_validacion <- predict(modelo_mejor, newdata = datos_validacion, type= "response")
```

#Curva ROC

```
ROC_pred <- prediction(pred_validacion,datos_validacion$Churn)
ROC_pref <- performance(ROC_pred, "tpr", "fpr")
plot(ROC_pref, colorize=TRUE, print.cutoffs.at=seq(0.1, by=0.1))

#Matriz de confusión
clas_validacion <- ifelse(pred_validacion > 0.5, "TRUE", "FALSE")
mean(clas_validacion == datos_validacion$Churn)
table(pred_validacion=clas_validacion, actual=datos_validacion$Churn)

#Visualizar matriz de confusión
ind_validacion<- seq(1:667)
ggplot(datos_validacion, aes(ind_validacion, pred_validacion))+ geom_point(aes(color=as.numeric(Churn)), alpha=1,shape=4, stroke=1)+  scale_colour_gradient(low="#EE0000", high="#1E90FF")+ xlab("Cliente")+ ylab("Abandono")+ theme_bw()+theme(legend.position="none")

#Constraste de hipótesis sobre los coeficientes del modelo
x<-data.frame(177.6887,173.7817,9.5198,3.2441, 108.0735, 19.5267,5.1868,17.3577, 19.6200, 132.6804)
1-pchisq(177.6887,df=1)
1-pchisq(173.7817,df=1)
1-pchisq(9.5198,df=1)
1-pchisq(3.2441,df=1)
1-pchisq(108.0735,df=1)
1-pchisq(19.5267,df=1)
1-pchisq(5.1868,df=1)
1-pchisq(17.3577,df=1)
1-pchisq(19.6200,df=1)
1-pchisq(132.6804,df=1)

#Contraste condicional de razón de verosimilitud
modelo.nulo<- glm(data$Churn ~ 1, family=binomial, data=data)
RV=-2*logLik(modelo.nulo)+2*logLik(modelo_mejor)
1-pchisq(abs(RV),df=9)

#Intervalos de confianza para los parametros del modelo
confint.default(modelo_mejor, level=0.95)

#Intervalos de confianza para las ventajas
exp(confint.default(modelo_mejor, level=0.99))
```

#Cálculo de los residuos de la variable respuesta

```
res<-residuals (modelo_mejor)
res.significativos<-abs(res)>2
res.significativos
table(res.significativos)
```

#Residuos de Pearson

```
res.pearson<-residuals (modelo_mejor, type="pearson")
res.significativos<-abs(res)>2
table(res.significativos)
```

##Distancia de Cook

```
distancias.cook<-cooks.distance(modelo_mejor)
distancias.cook
distancia_sig<- abs(distancias.cook)>1
table(distancias_sig)
plot(distancias.cook, xlab = "Observación", ylab = "Distancias de cook")
```

#DFBETAS

```
DFBTAS <- dfbetas (modelo_mejor)
DFBTAS_significativos <- abs(DFBTAS)>0.0387346756
table(DFBTAS_significativos)
```

#Dfits

```
df <- dffits (modelo_mejor)
dffits_significativas <- abs(df)>0.122448
table(dffits_significativas)
```

#Gráficos

```
par(mfrow=c(2,2))
plot(modelo_mejor, cex=0.6)
par(mfrow=c(1,2))
plot(modelo_mejor, which=4)
plot(modelo_mejor,which=6)
```

###Construcción del Árbol de Clasificación

#Crear un set de datos con las variables seleccionadas para el modelo ajustado

```
data_arbol <- subset(data, select=c(`International plan` , `Voice mail plan` , `Number vmail messages` ,
`Total day calls` , `Total day charge` , `Total eve minutes` ,
`Total night charge` , `Total intl calls` , `Total intl charge` ,
`Customer service calls`,  `Churn`))
```

#Crear modelo

```
arbol_mejor <- rpart(Churn~ `International plan` + `Voice mail plan` + `Total day charge` + `Total eve minutes`
+ `Total intl charge` + `Total intl calls`+ `Customer service calls`, data=data, method = "class")
```

#Predicción

```
pred_arbol_mejor <-predict(arbol_mejor, newdata = datos_validacion, type = "class")
```

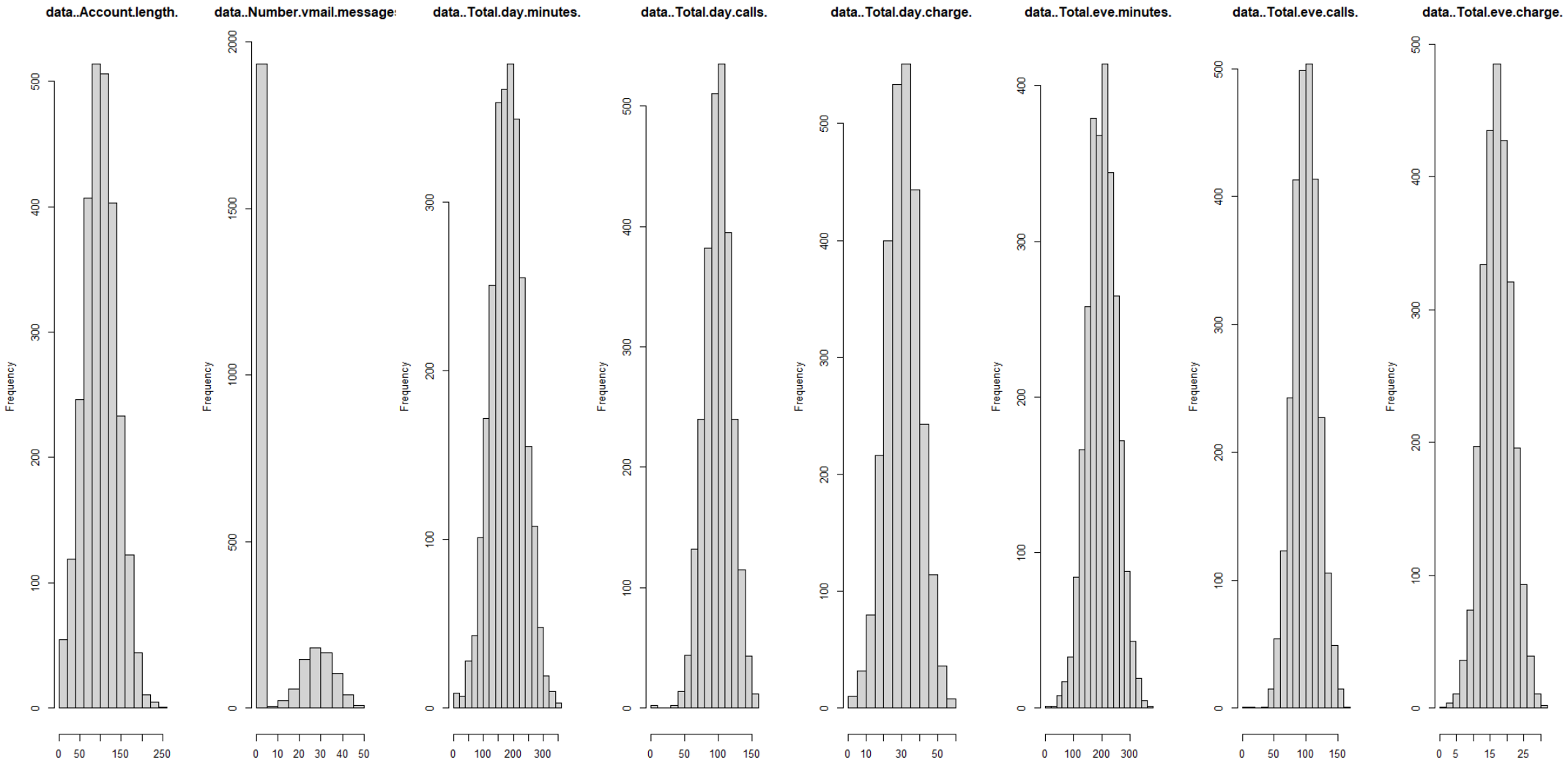
#Matriz de confusión

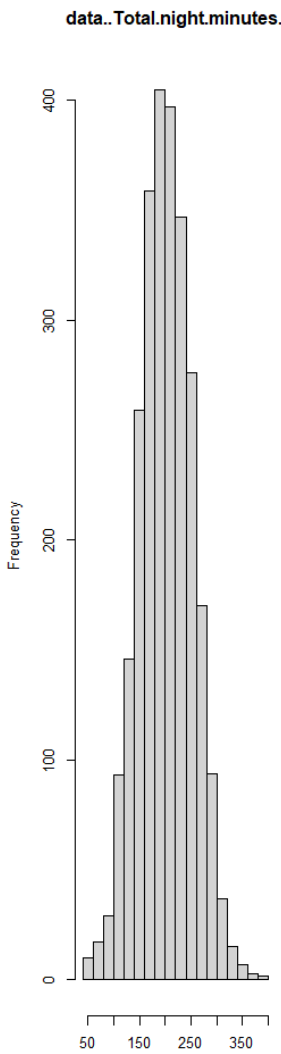
```
confusionMatrix(pred_arbol_mejor, datos_validacion$Churn)
```

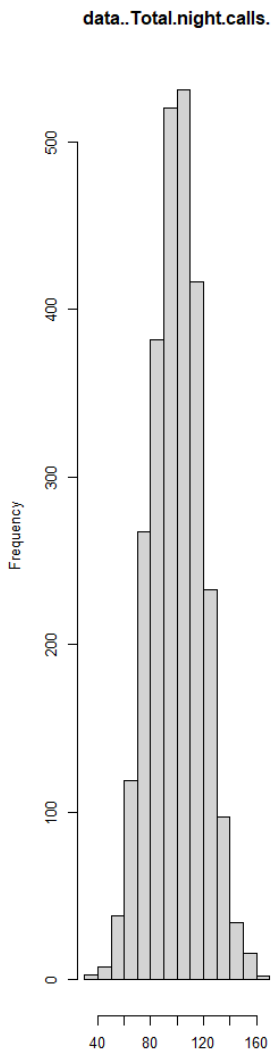
#Representar Árbol de Clasificación

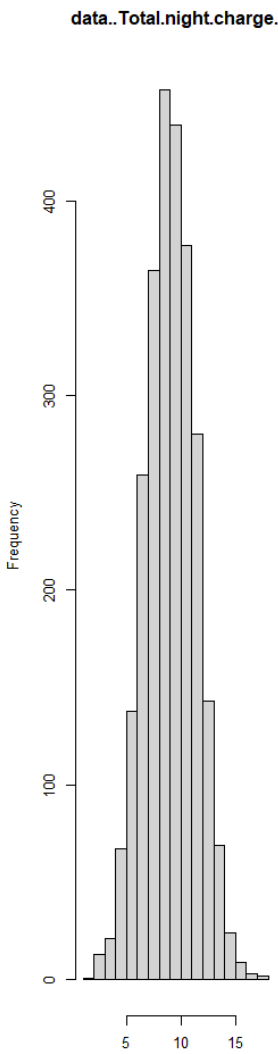
```
rpart.plot(arbol_mejor, digits = 3)
rpart.plot(arbol_mejor, digits = 4, fallen.leaves = TRUE, type = 3, extra = 101)
```

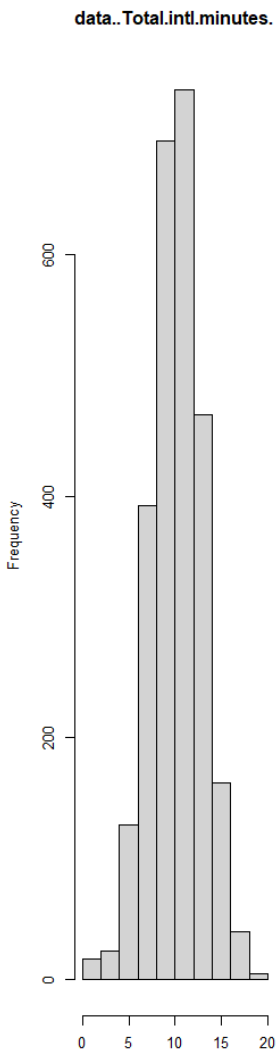
ANEXO II

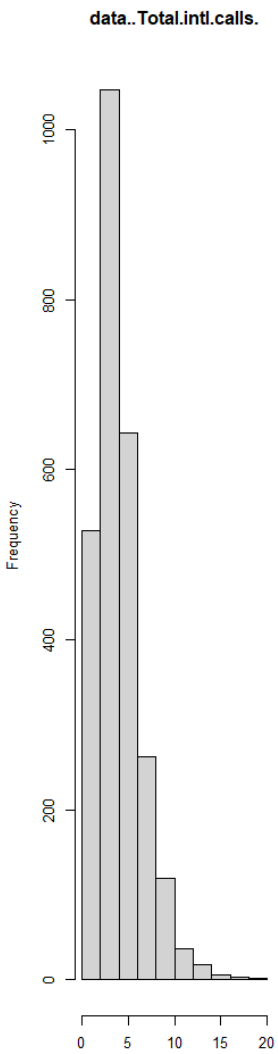

data..Account.length.
data..Number.vmail.messages
data..Total.day.minutes.
data..Total.day.calls.
data..Total.day.charge.
data..Total.eve.minutes.
data..Total.eve.calls.
data..Total.eve.charge.
Frequency

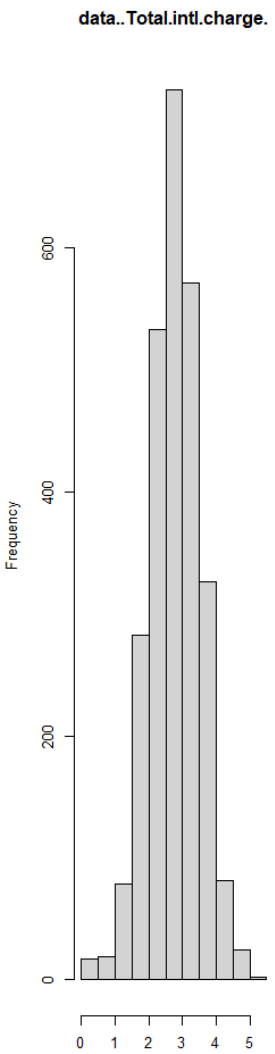

data..Total.night.minutes.
Frequency
0
100
200
300
400
50
150
250
350

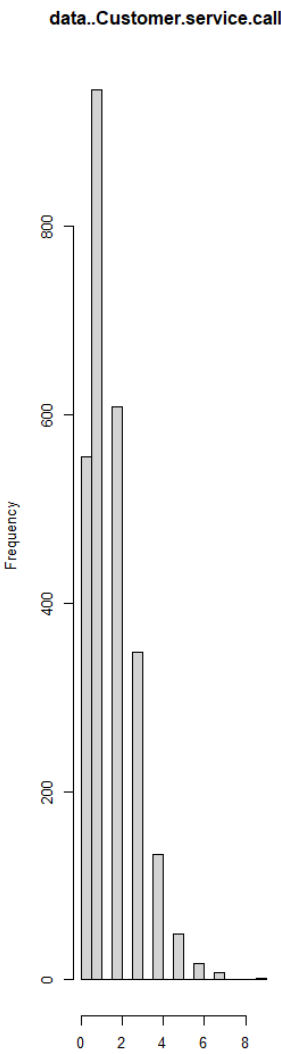

data..Total.night.calls.
Frequency
0
100
200
300
400
500
40
80
120
160

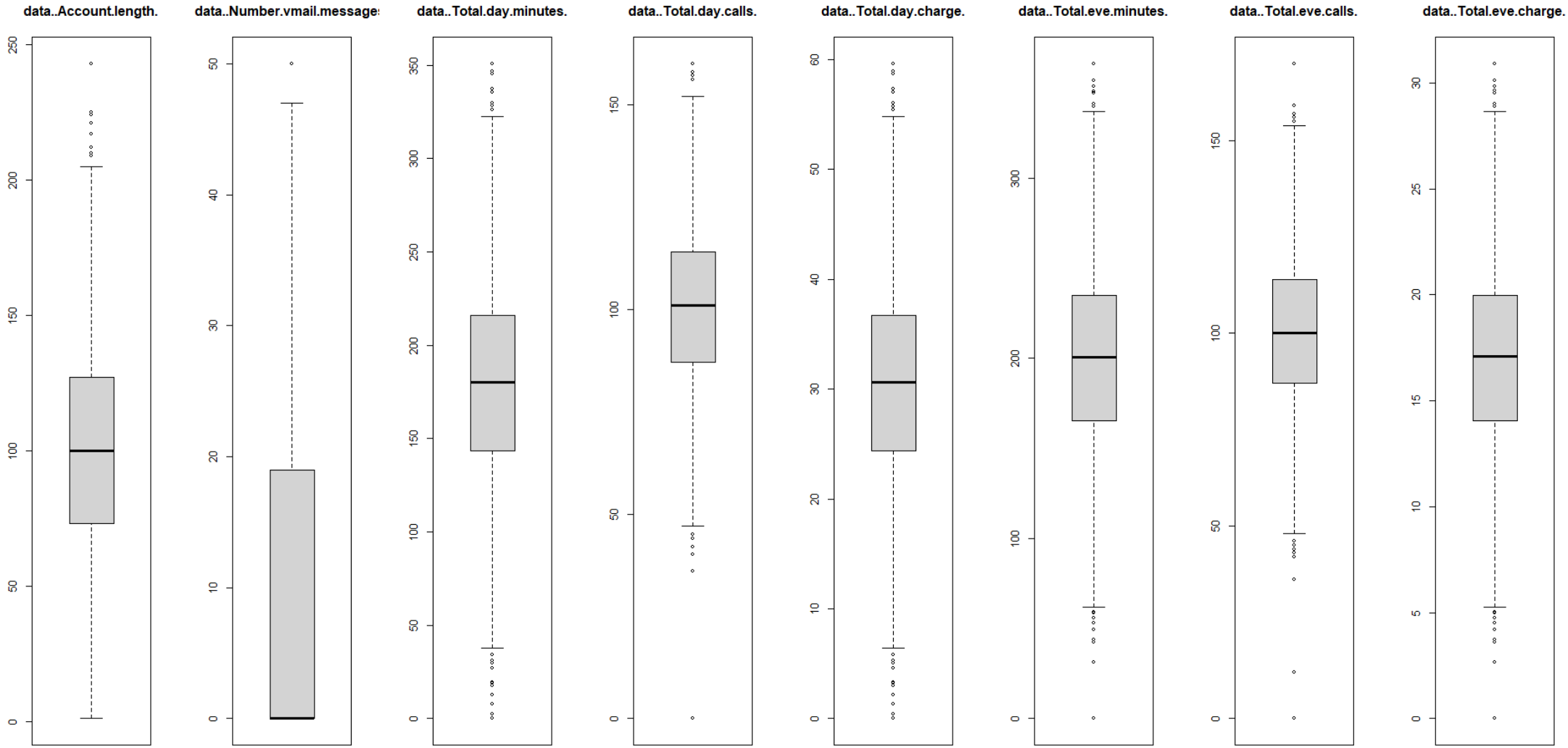

data..Total.night.charge.
Frequency
0
100
200
300
400
5
10
15

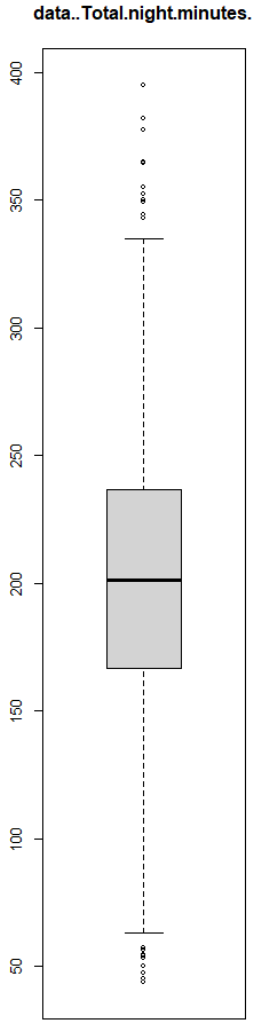

data..Total.intl.minutes.
Frequency
0
200
400
600
0
5
10
15
20

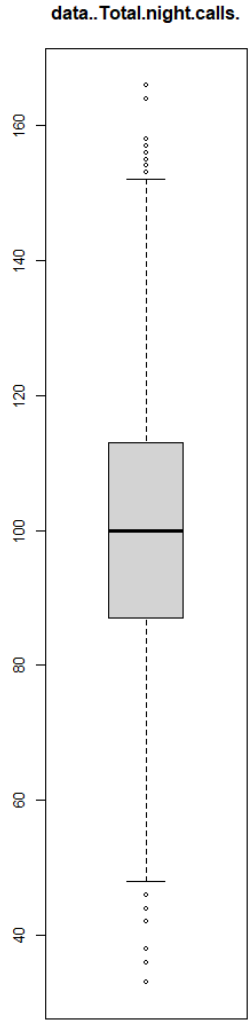

data..Total.intl.calls.
Frequency
0
200
400
600
800
1000
0
5
10
15
20

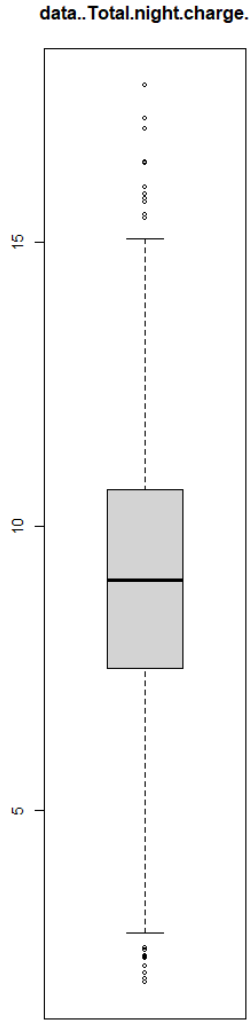

data..Total.intl.charge.
Frequency
0
200
400
600
0
1
2
3
4
5

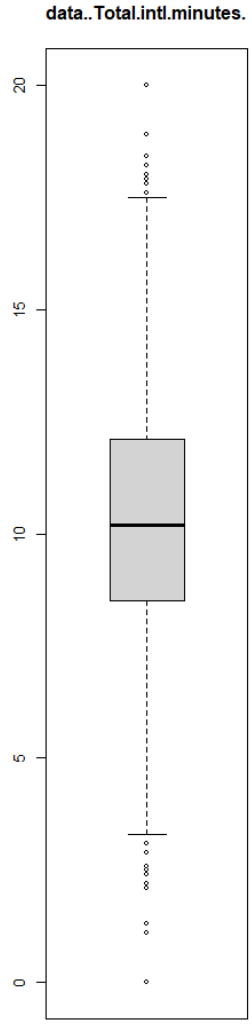

data..Customer.service.calls.
Frequency
0
200
400
600
800
0
2
4
6
8

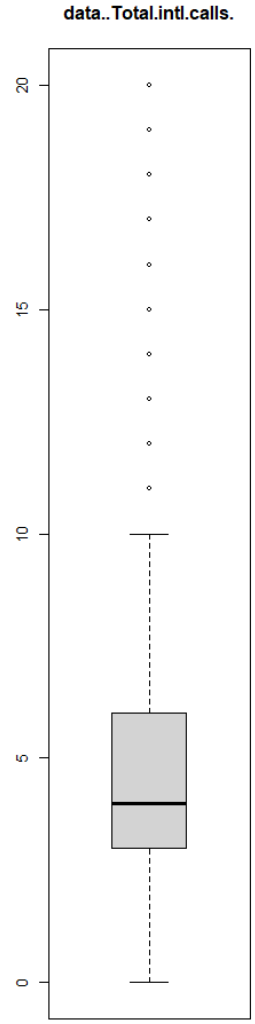

data..Account.length.
data..Number.vmail.message
data..Total.day.minutes.
data..Total.day.calls.
data..Total.day.charge.
data..Total.eve.minutes.
data..Total.eve.calls.
data..Total.eve.charge.

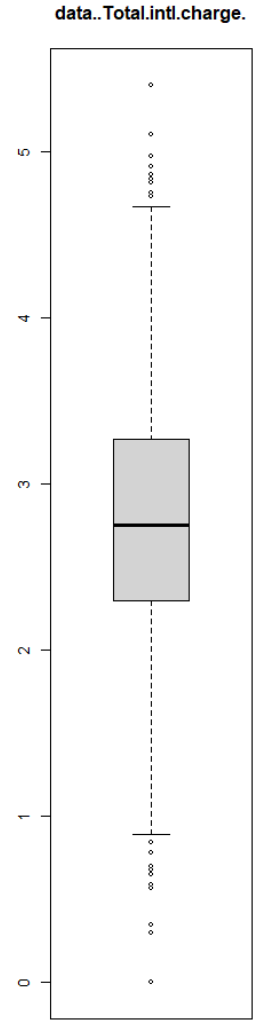

data..Total.night.minutes.
50
100
150
200
250
300
350
400

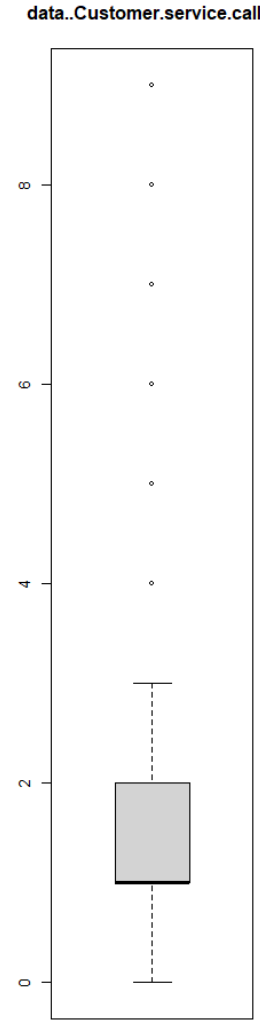

data..Total.night.calls.
40
60
80
100
120
140
160


data..Total.night.charge.
5
10
15


data..Total.intl.minutes.
0
5
10
15
20


data..Total.intl.calls.
0
5
10
15
20


data..Total.intl.charge.
0
1
2
3
4
5


data..Customer.service.calls.
0
2
4
6
8