UNIVERSIDAD DE ANTIOQUIA

1 8 0 3

Análisis estadístico de llamadas reincidentes a soporte técnico de Tigo Home (octubre 2020 -marzo 2021)

Jessica Torres Franco

Universidad de Antioquia
Facultad de Ciencias Exactas y Naturales
Instituto de Matemáticas
Medellín, Colombia
2021

Análisis estadístico de llamadas reincidentes a soporte técnico de Tigo Home (octubre 2020 -marzo 2021)

Jessica Torres Franco

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico

Jonatan A. González
Orientador Interno, Instituto de Matemáticas

Maria Angela Arrieta Argel
Orientador externo, Tigo

Universidad de Antioquia
Facultad de Ciencias Exactas y Naturales
Instituto de Matemáticas
Medellín, Colombia
2021

"No importa lo lento que vayas mientras no te detengas".

Resumen

En este trabajo, se presenta un análisis estadístico de tres bases de datos relacionadas con las llamadas de usuarios al soporte técnico de Tigo Home. Para el análisis, se utilizan técnicas de aprendizaje supervisado con el fin de modelar las llamadas y brindar a la empresa una base científica que sirva para la toma de decisiones relacionadas con la organización del personal que atiende a los usuarios telefónicamente.

Mediante técnicas de series de tiempo, se modela el número de llamadas diarias basándose únicamente en la estacionalidad, posible tendencia de la serie y su autocorrelación; además, se propone un sistema de predicción diaria basado en cinco modelos que han sido probados exitosos.

Con la ayuda de técnicas de regresión logística y árboles de decisión, se identifican los factores que influyen en que un usuario vuelva a llamar a soporte técnico por temas relacionados con el agendamiento de la cita, que ha sido programada de antemano a través de una primera llamada. A este respecto, se plantea y se prueba el rendimiento de un conjunto de cuatro modelos de acuerdo a la naturaleza de las bases de datos de aseguramiento y aprovisionamiento.

Palabras clave: Regresión logística, árbol de decisión, series de tiempo, aseguramiento, aprovisionamiento.

Abstract

This study presents a statistical analysis of three databases related to user calls to Tigo Home technical support. Supervised learning techniques are used to model the calls and provid the company with data that can be used to make decisions related to the organization of the person who serves the users through the telephone.

Using time series techniques, the number of daily calls is modeled based only on seasonality, possible trend of the series, and its autocorrelation; furthermore, a daily prediction system is designed based on five models that have been proven successfully.

With the help of logistic regression techniques and decision trees, it became easier to determine the factors that make a user to call back to technical support for issues related to the scheduling of the appointment determined in the first call. Regarding this, the performance of a set of four models is proposed and tested according to the nature of the assurance and provisioning databases.

Keywords: Logistic regression, decision tree, time series, assurance, provisioning.

Contenido

1. Introducción

Tigo es uno de los operadores de servicios de telecomunicaciones más relevantes en Colombia, cuyo principal objetivo es posicionarse como la empresa con mayor cobertura a nivel nacional, brindando variedad de servicios en tecnologías de información y comunicación y servicios de fijo y móvil a nivel internacional y nacional. Se encuentra presente en 32 departamentos del país con el propósito de llevar innovación y calidad en el estilo de vida digital de todos sus usuarios (Tigo, Tigo).

Unas de las finalidades de la compañía Tigo es garantizar una rápida y buena atención a los usuarios por medio de la implementación de metodologías ágiles, gestión que se encuentra a cargo de la Vicepresidencia de Experiencia a Clientes, cuyo propósito es comprender la experiencia del usuario a través de diferentes canales de servicio para asegurar que las dificultades que presente el usuario sean resueltas rápidamente.

El área de servicio al cliente, es la encargada en primera instancia de dar solución a los problemas que presente un usuario en cualesquiera de sus servicios proporcionados por Tigo home. Si el problema no pudo ser resuelto de manera remota por el asesor, él mismo procederá a agendar una visita técnica al cliente, en la que se buscará dar solución al problema.

Una vez agendada la cita, se notifica al usuario toda la información relacionada con la misma, a través de diferentes medios, buscando evitar que el cliente se vuelva a comunicar con soporte técnico. No obstante, se ha observado que, un 24 y un 30 % de clientes que reciben las notificaciones, se comunican de nuevo con soporte técnico, por lo que se hace necesario precisar las posibles variables que influyen en la reincidencia de las llamadas y predecir una aproximación de posibles llamadas futuras. Para ello, metodológicamente se hará uso de la técnica de aprendizaje supervisado.

2. Objetivos

2.1. Objetivo general

Implementar un conjunto de técnicas de análisis de datos relacionadas con el aprendizaje automático que revelen las razones estadísticas de la reincidencia de las llamadas. Estas razones permitirán la intervención en las decisiones de la compañía para la solución del problema.

2.2. Objetivos específicos

- Predecir las tasas de frecuencia de llamada de cada tipo de cliente durante el horizonte de planificación basándose en datos históricos.
- Identificar los factores que alteran el riesgo de reincidencia de llamada al soporte técnico por exposición (en el caso de que la aumenten) o ausencia (en el caso de que la disminuyan).

3. Marco Teórico

3.1. Regresión logística

Cuando se quiere clasificar un sujeto dentro de uno o más grupos previamente determinados a partir de un conjunto de características observadas, se puede hacer una regresión logística ya que esta es una técnica analítica que permite relacionar funcionalmente un suceso en función de un conjunto de variables independientes con el fin de modelar la influencia de estas sobre la probabilidad de ocurrencia de un evento en particular. Como resultado del análisis se obtienen unos coeficientes que miden la importancia de cada variable independiente, los cuales ayudan a generar una fórmula para estimar la probabilidad de pertenencia a uno de los grupos (Serna Pineda, 2009).

Es importante tener en cuenta que, aunque la regresión logística permite clasificar, se trata de un modelo de regresión que modela el logaritmo de la probabilidad de pertenecer a cada grupo. La asignación final se hace en función de las probabilidades predichas.

Regresión logística simple

Sea Y la variable dependiente que toma valores de $\{0,1\}$ en función de la variable independientes X la cual puede ser continua, discreta, dicotómica, ordinal o nominal y sea β_i son los parámetros del modelo. El modelo logístico simple esta dado por

$$p(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

Sea $P(Y=1|X) = P(\mathbf{X})$, donde la negrita $\mathbf{X}$ es una notación abreviada para la colección de variables independiente, en este caso solo sería X_1. La ecuación anterior puede estructurarse de la siguiente forma:

$$p(\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{1}$$

Regresión logística univariante múltiple

Esta regresión logística es una extensión de la regresión logística simple, en la que se predice la variable dependiente Y que toma valores de $\{0,1\}$ en función de las variables independientes $X_1, X_2, ..., X_p$.

$$p(Y=1|X_1, X_2, ..., X_p) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p)}}$$

Al igual que en el caso de regresión logística simple, se denota **X** como la colección de variables independientes. La ecuación anterior puede estructurarse de la siguiente forma:

$$p(\mathbf{X}) = \frac{e^{(\beta_0 + \Sigma \beta_p X_p)}}{1 + e^{(\beta_0 + \Sigma \beta_p X_p)}} \tag{2}$$

La gráfica de las ecuaciones 1 y 2 es una curva en forma de S ycomprende los valores de Y en el intervalo $[0, 1]$, como se observa en la figura 1.

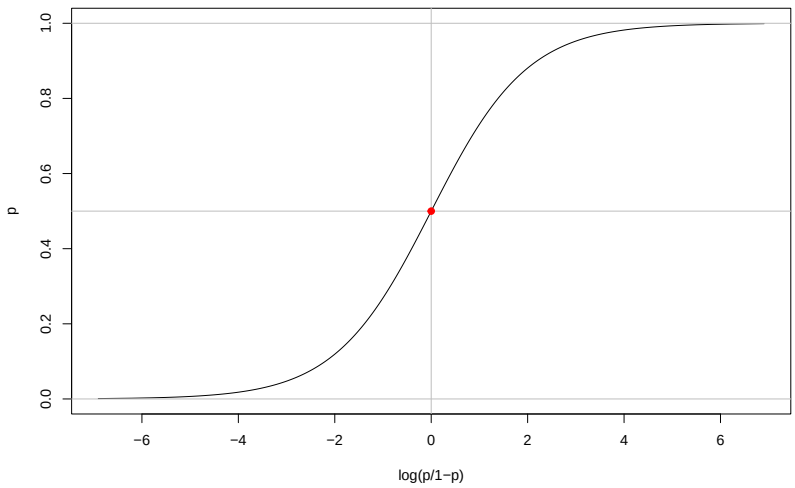

Figura 1: Función logística

3.2. Árbol de decisión

Los árboles de decisión son algoritmos versátiles de aprendizaje automatizado que pueden realizar tareas de clasificación y regresión, e incluso tareas de salida múltiple. Es un algoritmo muy poderoso y se pueden adaptar a conjuntos de datos complejos. Entre las ventajas de los árboles de decisión es su robustez a los outliers, que requieren muy poca preparación de los datos y que son fáciles a la hora de interpretar. Sin embargo, normalmente no son competitivos con los mejores enfoques de aprendizaje supervisado.

Los árboles de decisión se dividen en dos tipos:

- Árboles de regresión, cuya variable dependiente es continua y los valores de los nodos que indican el resultado definitivo (nodo terminal) es la media de las observaciones en esa región.
- Árboles de clasificación, cuya variable dependiente es cualitativa y el valor en el nodo terminal se reduce a la moda de las observaciones del conjunto de entrenamiento que han “caído”en esa región.

El objetivo de los árboles de decisión es encontrar múltiples esquemas de dicotomía o bifurcación anidados en forma de árbol, de modo que, al rastrear cada rama del árbol se obtenga, una predicción para la clase de pertenencia (clasificación) o para el valor que toman (regresión) los individuos que cumplen con las propiedades que se han ido exigiendo en las distintas bifurcaciones, todo esto se hace por medio de reglas de decisión inferidas en los datos de entrenamiento.

Los árboles de decisión se construyen mediante el algoritmo de segmentación recursiva y se pueden obtener de diferentes formas, unas de estas son: CHAID (Chi-Square Automatic Interaction Detector), QUEST (Quick Unbiased Efficient Statistical Tree) y la que se usa en este trabajo es CART (Classification And Regression Trees) y la implementación particular de CART que usa es Recursive Partitioning and Regression Trees o RPART (Vega, 2018).

CART (Serna Pineda, 2009) es un método no-paramétrico de segmentación binaria donde el árbol es construido dividiendo repetidamente los datos y en cada división los datos son partidos en dos grupos mutuamente excluyentes, cuyo objetivo es dividir la respuesta en grupos homogéneos mientras se mantiene un árbol bastante pequeño. Para dividir los datos, se necesita un estándar de clasificación para determinar la medida de impureza, que establecerá el grado de homogeneidad entre grupo.

Particionamiento recursivo

Sea Y una variable dependiente y sean $X_1, X_2, ..., X_p$, variables independientes donde las X son varibles fijas y Y es una variable aleatoria. El problema estadístico es establecer una relación entre la variable dependiente y las variables independiente de tal forma que sea posible predecir Y basado en los valores de las X. Matemáticamente, se quiere estimar la probabilidad condicional de la variable aleatoria Y, es decir, $P(Y|X_1, X_2, ..., X_p)$ o un funcional de su probabilidad tal como la esperanza condicional $E(Y|X_1, X_2, ..., X_p)$, según se trate de un árbol de regresión o de clasificación (Sepúlveda, 2012).

Construcción del árbol

La estructura de los árboles de clasificación se construyen mediante divisiones repetidas de subconjuntos de X en dos subconjuntos descendientes comenzando con X mismo, es decir, comienza con un nodo raíz (que sería X) y luego se ramifica en dos nodos internos X_2 y X_3, estos resultados crean nodos adicionales, que se ramifican en otras posibilidades, hasta llegar a los nodos terminales que en este caso serian X_6, X_8, X_{10}, X_{11}, X_{12}, X_{14}, X_{15}, X_{16} y X_{17} (Figura **??**). A cada nodo terminal se asigna a una clase que representa el valor objetivo (Breiman et al., 2017).

Los nodos internos se representan como círculos, mientras que los nodos terminales se denotan como triángulos o cuadrados. De cada nodo interno pueden salir dos o más ramas. Cada nodo se corresponde con una determinada característica y las ramas corresponden a un rango de valores. Estos rangos de valores deben ser mutuamente excluyentes y completos. Estas dos propiedades de disociación y completitud son importantes, ya que garantizan que cada instancia de datos se asigne a una instancia.

Las instancias se clasifican navegando desde la raíz del árbol hasta un nodo terminal según el resultado de las pruebas a lo largo del camino. Para esto se empieza con una raíz de árbol; luego se considera la característica que corresponde a la raíz y se procede a definir a qué rama corresponde el valor observado de la característica dada, este análisis se repite hasta llegar a un nodo terminal (Rokach and Maimon, 2014).

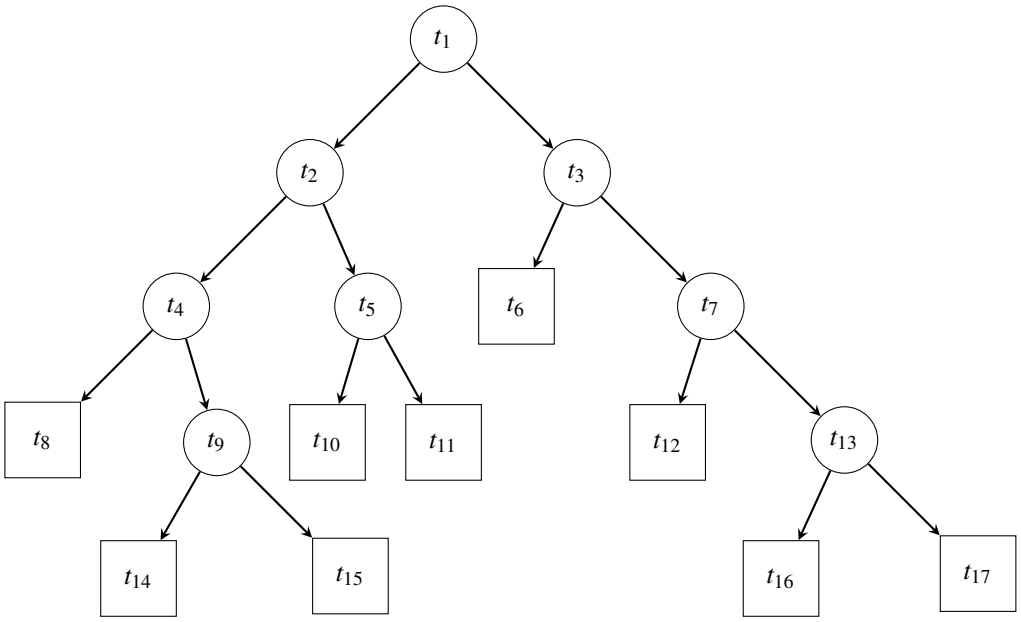

Figura 2: Ejemplo árbol de decisión

Impureza del nodo

La función de impureza es una medida que permite determinar la calidad de un nodo. Existen varias medidas de impureza (criterios de particionamiento) que permiten analizar varios tipos de respuesta, las tres medidas más comunes para árboles de clasificación presentadas por Breiman et al. (2017) son:

- Índice de información o entropía. Es una medida que indica el desorden de las caracteristicas con las etiquetas. La división óptima es elegida por la características con menos entropía y se obtiene su valor máximo cuando la probabilidad de las dos clases es la misma y un nodo es puro cuando la entropía es 0 (Rodríguez, 2020). El índice de entropía se define como:

$$i(t) = \sum_j p(j|t) \ln p(j|t) \tag{3}$$

  Cuyo objetivo es encontrar la partición que maximice $\Delta i(t)$ la ecuación anterior, es decir,

$$\Delta i(t) = -\sum_{j=1}^{p} p(j|t) \ln p(j|t) \tag{4}$$

  Donde $i(t)$ es la calidad del nodo, $j = 1...p$ es el número de clases de la variable categórica y $p(j|t)$ es la probabilidad de clasificación correcta para la clase j en el nodo t (Serna Pineda, 2009).

- Índice de Gini. Tiende a separar la categoría mas grande en un grupo aparte a diferencia del indice de entropía que tiende a formar grupos con más de una categoría en las primeras decisiones. Este se define como:

$$i(t) = \sum_{i \neq j} p(j|t) p(i|t) \tag{5}$$

El objetivo es encontrar la partición que maximice $\Delta i(t)$ la ecuación anterior, es decir,

$$\Delta i(t) = \sum_{j}^{p} [p_j(t)]^2 \tag{6}$$

- Índice Towing. Busca dos clases que juntas formen más del 50 % de los datos, esto define dos super categorías en cada división para las cuales la impureza es definida por el índice Gini. El índice towing produce árboles mas balanceados. Para usar el índice towing selecciona la partición, que maximice (Breiman et al., 2017)

$$\frac{p_L p_R}{4} \left[\sum_{j} |p(j|t_L) - p(j|t_R)| \right]^2 \tag{7}$$

Donde t_L representa el nodo hijo izquierdo y t_R representa el nodo hijo derecho, p_L y p_R representan la proporción de observaciones en t que pasaron a t_L y a t_R en cada caso Serna Pineda (2009).

3.3. Series de tiempo

Una serie de tiempo es una serie de puntos de datos ordenados en el tiempo. Durante un evento en una serie de tiempo, las medidas son organizadas típicamente en tiempos sucesivos (Adhikari and Agrawal, 2013). Gracias a esto, existe la posibilidad de una correlación entre las observaciones. En gran medida, el análisis de las series de tiempo tiene como objetivo explicar esta correlación y las principales características de los datos, usando modelos estadísticos y métodos descriptivos apropiados (Paul S. P. Cowpertwait, 2009). En una serie de tiempo, el tiempo es a menudo la variable independiente y el objetivo suele ser hacer un pronóstico para el futuro. Se pueden extraer diversas características de las series de tiempo, como las tendencias y variaciones estacionales que pueden ser modeladas de forma determinista con funciones matemáticas del tiempo (Paul S. P. Cowpertwait, 2009).

Sea Y_t una serie temporal en la que t denota el momento en que se toma la observación, donde $t \in Z^+$. El objetivo es construir un modelo que describa la evolución de la serie a través del tiempo, para esto se asume que los datos se pueden expresar como una función de una componente de tendencia T_t, estacional S_t y un error E_t (Jonathan D. Cryer, 2009).

Ruido Blanco

Un proceso ε_t se denota ruido blanco de media 0 y varianza σ^2 si satisface

$$E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma^2 < \infty, Cov(\varepsilon_t, \varepsilon_{t-k}) = 0$$

Para todo $k \neq 0$. En particular, una sucesión de variables aleatorias independientes e idénticamente distribuidas, con media 0 y varianza σ_ε^2 representa un caso especial de un proceso de ruido blanco, y que se denota por $\varepsilon_t \sim N(0, \sigma 2_\varepsilon)$. Si además ε_t se distribuye normalmente, la serie se denomina ruido blanco gaussiano (Jonathan D. Cryer, 2009).

Hay tres formas de comprobar si la serie temporal se asemeja al ruido blanco:

- Trazando la serie temporal
- Comparando la media y la desviación estándar a lo largo del tiempo
- Examinando los gráficos de autocorrelación

Modelos AR(p)

Los modelos de autorregresivos de orden finito p, son una representación de un proceso aleatorio, en el que la variable de interés depende de sus observaciones pasadas. En general, para denotar el modelo autorregresivo AR se usa $AR(p)$ (Giraldo Gómez, 2006). Así, un modelo (AR) de orden p se puede escribir como

$$Y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t \tag{8}$$

para constantes $\phi_0, ..., \phi_p$ y $\varepsilon_t \sim RB(0, \sigma^2)$.

Modelos MA(q)

Los modelos de medias móviles de orden finito q, son una aproximación común para las series de tiempo univariadas. El modelo de medias móviles especifica que la variable de salida depende linealmente del valor actual y varios de los anteriores. En general para denotar un modelo de medias móviles MA se usa $MA(q)$ (Giraldo Gómez, 2006). Así, un modelo MA de orden q se puede escribir como

$$Y_t = \theta_0 + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} \tag{9}$$

para constantes $\theta_1, ..., \theta_q$ y $\varepsilon_t \sim RM(0, \sigma^2)$.

Modelo ARMA

un proceso $ARMA(p,q)$ es un modelo que combina las propiedades de memoria larga de los $AR(p)$ con las propiedades de ruido débilmente autocorrelacionado en los $MA(q)$, y que tiene suficiente flexibilidad y parsimonia para representar una variedad grande de procesos estacionarios en covarianza.

$$Y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j} \tag{10}$$

donde $\varepsilon_t \sim RB(0, \sigma^2)$ (Giraldo Gómez, 2006).

Modelos Sarima

Un proceso ARIMA estacional o SARIMA, utiliza la diferenciación con un retardo igual al número de estaciones (s) para eliminar los efectos estacionales aditivos. El modelo SARIMA tiene una componente ARIMA(P, D, Q) que modeliza la dependencia estacional, que está asociada a observaciones separadas por l periodos y contiene otra componente ARIMA(p, d, q) que modeliza la dependencia regular, que es la dependencia asociada a observaciones consecutivas y tambien tiene un proceso diferenciado $w_t = \nabla_l^D \nabla^d Y_t$ es un proceso estacionario que sigue el modelo ARMA estacional. Por tanto la ecuación general del modelo SARIMA es

$$\Phi_P(B^l)\phi_p(B)w_t = \Theta_Q(B^l)\theta_q(B)E_t \tag{11}$$

Con $E_{t\,t\in Z}$ un $RB \sim (0, \sigma^2)$ (Giraldo Gómez, 2006).

4. Metodología

4.1. Balanceo de Clases

A la hora de hacer los modelos de regresión logística y árboles de clasificación se evidencia que hay un desbalance en las clases, esto quiere decir, que las llamadas reincidentes por temas que no están relacionados al agendamiento es mayor que el número de llamadas reincidentes por agendamiento. Al momento de realizar estos modelos la proporción de una u otra clase en la variable dependiente debería ser aproximadamente la misma en los datos de entrenamiento, esto se debe a que el algoritmo tiende a favorecer la clase con la mayor proporción de observaciones, lo cual puede ocasionar que haya sesgo cuando se vaya a evaluar el modelo. Para evitar esto se realiza un balanceo de clases, es decir, se utilizan los valores de la variable clasificadora que ajusta automáticamente los pesos que son inversamente proporcionales a las frecuencias de clase en los datos (Rodríguez, 2020).

4.2. Métricas para medir la efectividad de los métodos utilizados

Para ver que tan efectivos son los modelos de regresión logística y de árboles de clasificación se utilizan métodos como error de clasificación, matriz de confusión y curvas de ROC.

Error de clasificación

El error de clasificación es el desajuste porcentual de los valores que se utilizan para el entrenamiento del modelo frente a los que se utilizan para verificar la eficacia del modelo. Cuanto menor es el error de clasificación, mejor es el modelo.

$$Error\ de\ clasificacion\ = \frac{Total\ de\ mal\ clasificados}{cantidad\ de\ datos} * 100\,\%$$

Matriz de confusión

Para medir la efectividad de los modelos de árboles de clasificación se analizan todos outputs que arroja una matriz de confusión. Esta permite visualizar y estimar el rendimiento de un algoritmo al informar las clasificaciones reales y previstas junto con sus exactitud. Cada columna de la matriz representa las ocurrencias en una clase real, mientras que cada fila representa las ocurrencias en una clase predicha.

	Clase real		
Clase predicha		Evento	No evento
	Evento	TP	FP
	No evento	FN	TN

Donde **TP**, verdadero positivo; **TN**, verdadero negativo; **FP**, falso positivo; **FN**, falso negativo.

- **Sensibilidad**: Es la tasa positiva verdadera o recuperación, muestra la proporción de la clase positiva predicha correctamente.

$$Sensibilidad = \frac{TP}{TP+FN}$$

- **Especificidad**: Es la tasa negativa verdadera, muestra la proporción de la clase negativa predicha correctamente.

$$Especificidad = \frac{TN}{TN+FP}$$

- **Precisión**: Es la proporción del número total de predicciones correctas.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

- **Valor predictivo positivo**: Muestra el número de la clase positiva predicha correctamente como una proporción del total de predicciones de la clase positiva realizadas.

$$VPP = \frac{TP}{TP+FP}$$

- **Valor predictivo negativo**: Muestra el número de clases negativas predichas correctamente como una proporción del total de predicciones de clases negativas realizadas.

$$VPN = \frac{TN}{TN+FN}$$

- **Prevalencia**: Muestra con qué frecuencia ocurre realmente la clase positiva en nuestra muestra.

$$Prevalencia = \frac{TP+FN}{TP+FP+FN+TN}$$

- **Tasa de detección**: Muestra el número de predicciones de clase positivas correctas realizadas como proporción de todas las predicciones realizadas.

$$Tasa\ de\ deteccion = \frac{TP}{TP+FP+FN+TN}$$

- **Prevalencia de detección** : Muestra el número de predicciones de clase positivas realizadas como proporción de todas las predicciones.

$$Prevalencia\ de\ deteccion = \frac{TP+FP}{TP+FP+FN+TN}$$

- **Exactitud equilibrada**: Esencialmente toma el promedio de las tasas positivas y negativas verdaderas.

$$Exactitud\ equilibrad = \frac{sensibilidad + especificidad}{2}$$

- **Tasa de ausencia de información**: Criterio que indica el valor más alto entre la prevalencia.
- **Kappa**: Dice qué tan bien coinciden las predicciones de clasificadores con las etiquetas de clase reales, mientras se controla la precisión de un clasificador aleatorio.
- **Valor-p del test de McNemar**: Verifica si hay una diferencia significativa entre los falsos positivos y los falsos negativos.

$$\chi^2 = \frac{(FP-FN)^2}{FP+FN}$$

Curva ROC

A la hora de ver que tan buenos son los modelos de regresión logística, se utiliza la curva de ROC (Receiver Operating Characteristics), esta muestra el porcentaje de verdaderos positivos pronosticados con precisión por el modelo logit, la curva debería subir bruscamente, indicando que el TPR (eje Y), es decir la tasa de verdaderos positivos aumenta más rápido que la tasa de falsos positivos FPR (eje X) a medida que disminuye la puntuación de corte. Cuanto mayor sea el área bajo la curva ROC, mejor será la capacidad de predicción del modelo.

$$Sensibilidad = \frac{TP}{TP+FN} \quad vs \quad 1 - Espacificidad = \frac{FP}{TN+FP}$$

4.3. Criterios de información

Criterio de información de Akaike (AIC) y criterio de información Bayesiano (BIC). Cuando se tiene una serie de modelos, una metodología para compararlos corresponde a la función de máxima verosimilitud. La máxima verosimilitud permite seleccionar el modelo que realiza el mejor ajuste de los datos, pero no penaliza su complejidad, lo que si sucede cuando se emplean medidas de contraste como el AIC y el BIC (P. and R., 2007).

Criterio de información de Akaike (AIC). El criterio combina la teoría de máxima verosimilitud e información teórica, este criterio tiene en cuenta los cambios en la bondad de ajuste y los mejores modelos son aquellos que presentaron el menor valor de AIC. Este criterio está definido por(P. and R., 2007):

$$AIC = -2\hat{ln}(\beta) + 2p$$

Criterio de información bayesiano (BIC). El BIC es calculado para los diferentes modelos como una función de la bondad de ajuste de ajuste del $\ln(\beta)$, el número de parámetros ajustados (p) y el número total de datos (n). El modelo con el más bajo valor de BIC es considerado el mejor en explicar los datos. Este criterio este definido por la ecuación (P. and R., 2007):

$$BIC = -2\hat{ln}(\beta) + 2log(n).$$

Donde

$$Ln(\beta) = \prod_{n}^{t=1} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(Y_t - f(t,\beta))^2}{2\sigma^2} \right\}.$$

4.4. Autocorrelación y Autocorrelación parcial

Las funciones de autocorrelación (ACF) y las funciones de autocorrelación parcial (PACF) se utilizan para describir la presencia o ausencia de correlación en los datos de las series temporales, indicando si influyen las observaciones del pasado en las observaciones del futuro. Es decir, que la autocorrelación hace referencia a los valores que toman una variable en el tiempo no son independientes entre sí, sino que un valor determinado depende de los valores del pasado.

El test de ACF, esta definido por

$$\rho(k) = \text{corr}(a_t, a_{t+k}),$$

para todo $k = 1, ..., 48$, construidos con los residuales de ajuste, y se tienen las siguiente hipótesis,

$$H_0 : \rho(k) = \text{corr}(a_t, a_{t+k}) = 0 \quad \forall k, \text{ vs. } \exists k, \rho(k) \neq 0.$$

y su estadístico de prueba es,

$$\hat{\rho}(k) = \frac{\sum_{i=1}^{n-k} \hat{a}_t \hat{a}_{t+k}}{\sum_{i=1}^{n} \hat{a}_t^2} \sim N\left(0, \tfrac{1}{n}\right),$$

El test PACF, esta definido por

$$\phi_{kk} = \text{corr}(a_t, a_{t+k}|a_{t+1}, ..., a_{t+k-1}),$$

para todo $k = 1, ..., 48$, construidos con los residuales de ajuste, y se tienen las siguientes hipótesis,

$$H_0 : \phi_{kk} = 0 \quad \forall k, \text{ vs. } \exists k, \phi_{kk} \neq 0.$$

y su estadístico de prueba es,

$$\hat{\phi}_{kk} = \widehat{\text{corr}}(a_t, a_{t+k}|a_{t+1}, ..., a_{t+k-1}) \sim N\left(0, \tfrac{1}{n}\right),$$

El criterio de rechazo de ambos test esta dado por el teorema de Bartlett.

4.5. Teorema de Bartlett

Suponga una serie de tiempo proveniente de un proceso estacionario en covarianza de media cero (Bartlett, 1946) y sea (ε_t, t e Z) un ruido blanco entonces las autocorrelaciones muestrales son $\hat{\rho} = \hat{Corr}(\varepsilon_t, \varepsilon_{t+k})$ con $k = 1, 2, ..., m$ (Giraldo Gómez, 2006). Si $\rho(k) = 0$ para $k > q$ entonces

$$VAR[\hat{\rho(k)}] \approx \frac{1}{n}\left(1 + 2\sum_{j=1}^{q}[\rho(j)]\right)$$

donde $1 < m < n$ es un entero arbitrario, con base en una muestra $\{\varepsilon_1, ..., \varepsilon_t\}$ cumple que $\hat{\rho(k)}$ son independientes y $\hat{\rho(k)} \overset{\text{aprox}}{\sim} N(0, \frac{1}{n})$. Si una serie de tiempo proviene de un proceso de ruido blanco, realizando $m = [n/4]$ pruebas de tipo

$$H_0 : \rho(k) = 0 \text{ Vs. } H_1 : \rho(k) \neq 0$$

para cada k, con $\alpha \approx 5\,\%$, se rechaza que el procesos es un ruido blanco si para algún k se observa $|\hat{\rho(k)}| > 2/\sqrt{n}$, es decir, cuando en al menos una de las m pruebas se rechaza la H_0 (Bartlett, 1946).

4.6. Test de Ljung-Box y Box-Pierce

Es una prueba estadística de si alguno de un grupo de autocorrelaciones de un series de tiempo son diferentes de cero. En lugar de probar aleatoriedad en cada desfase distinto, prueba la aleatoriedad general en función de varios desfases.

Los tests de Ljung-Box y Box-Pierce evalúan la siguiente hipótesis,

$$H_0 : \rho(1) = ... = \rho(m) = 0, \text{ vs. } \exists k, \rho(k) \neq 0.$$

Los estadísticos de prueba de Box-Pierce y Ljung-Box están dados, respectivamente, por

$$Q_{BP} = n \sum_{k=1}^{m} \hat{\rho}^2(k) \sim \chi^2_m, \qquad Q_{LB} = n(n+2) \sum_{k=1}^{m} \frac{\hat{\rho}^2(k)}{n-k} \sim \chi^2_m.$$

Si $H_0 : \varepsilon_t \sim RB(0, \sigma^2)$ se cumple que $Q_{LB} \overset{a}{\sim} \chi^2_m$.

Colocando Q_{obs} el estadístico observado y Valor-p = $P(\chi^2_m \geq Q_{obs}|H_0 cierto)$ si Valor-$p < 0.05$, se rechaza H_0 y si el Valor-$p > 0.05$, no se rechaza la H_0 (Giraldo Gómez, 2006).

4.7. Test de Lilliefors

La prueba de Lilliefors es una prueba de bondad de ajuste de dos caras que resulta adecuada cuando se desconocen los parámetros de la distribución nula y deben estimarse. Esto contrasta con la prueba de Kolmogorov-Smirnov de una muestra, que requiere que la distribución nula esté completamente especificada.

El estadístico de la prueba de Lilliefors es:

$$T = sup|F*(x) - S(x)|$$

Donde T es el supremo, sobre todo x, del valor absoluto de la diferencia $F*(x) - S(x)$; $F*(x)$ es la función de distribución acumulada de una distribución normal con media cero y desviación estándar uno y $S(x)$ es la función de distribución empírica de los valores de $Z_i = (X_i - \overline{X})/s$.

4.8. Precisión de los modelos de pronóstico

- **Error promedio de pronóstico (ME):** Da información sobre de la suma total de los valores observados que se utilizaron para el pronóstico divido los valores pronosticados en el modelo.

$$ME = \frac{1}{m} \sum_{L=1}^{m} e_n(L)$$

- **Error promedio absoluto de pronóstico (MAE):** Da información sobre de la suma total de los valores observados que se utilizaron para el pronóstico divido los valores pronosticados en el modelo.

$$MAE = \frac{1}{m}\sum_{L=1}^{m}|e_n(L)|$$

- **Raíz del error cuadrático medio (error estándar) de pronóstico (RMSE):** Dice en promedio cuanto se aleja el ajuste al valor real.

$$RMSE = \sqrt{\frac{1}{m}\sum_{m}^{L=1} e_n^2(L)}$$

- **Porcentaje medio de error (MPE):** Da el porcentaje de la estimación del error divido por la cantidad observa.

$$MPE = 100\,\%\frac{1}{m}\sum_{L=1}^{m} ER_n(L)$$

- **Porcentaje medio absoluto de error (MAPE):** Da el porcentaje de la estimación del error divido por la cantidad observa, pero en valor absoluto.

$$MAPE = 100\,\%\frac{1}{m}\sum_{L=1}^{m} |ER_n(L)|$$

Donde N es la longitud de Y_t, n observaciones $(n < N)$ usadas para ajuste, $m = N - n$ son las observaciones pronosticadas con el modelo ajustado. El error de pronóstico L es

$$e_n(L) = Y_{n+L} - \hat{Y}_n(L)$$

Y el error relativo de pronóstico.

$$ER_n(L) = \frac{e_n(L)}{Y_{n+L}}.$$

5. Experimento

La base de datos está conformada por 1.564.463 datos que fueron tomados diariamente entre los meses de octubre de 2020 a marzo de 2021, esta se encuentra dividida en dos partes que son procesos de aseguramiento (reparación) que consta de 689290 datos y procesos de aprovisionamiento (instalación) con 867286 datos, cada parte contiene 21 variables explicativas que hacen referencia al estado de la agenda y a la información del usuario. Dos de estas variables son clasificadoras, una indica si el usuario volvió a contactar al soporte técnico después de haber llamado por primera vez (llamada) y la otra dice si la llamada está relacionada por temas de agendamiento (llamada agenda).

5.1. Análisis descriptivo

Para hacer el análisis descriptivo se unen las dos bases de datos.

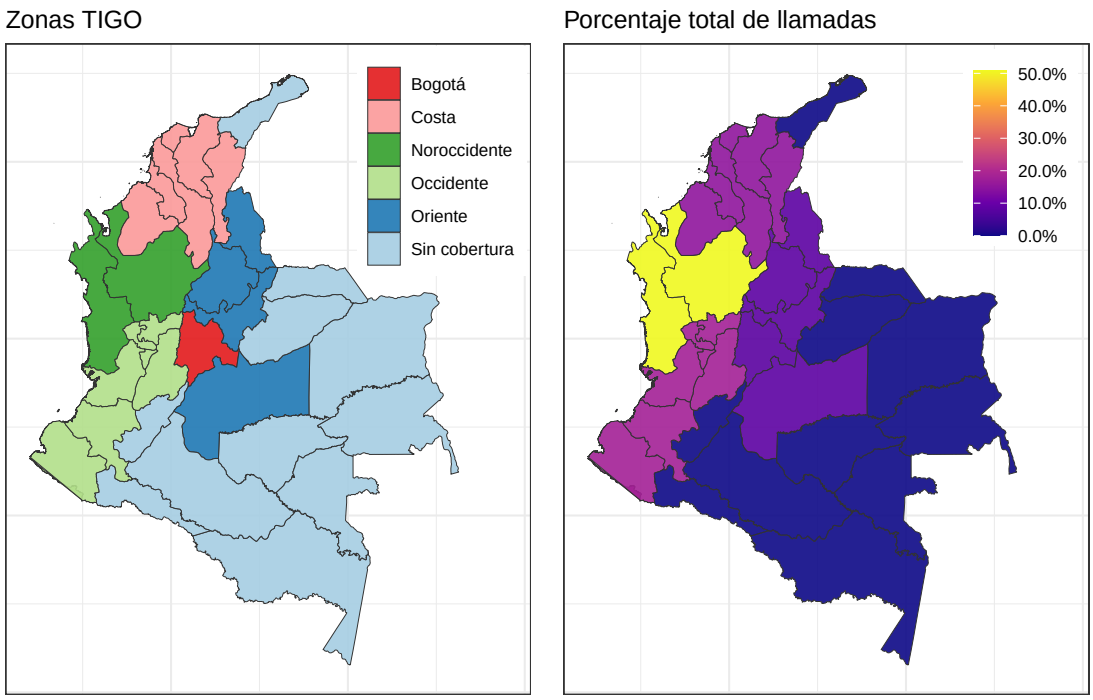

Figura 3: Gráfica izquierda: Distribución geográfica de la cobertura de Tigo. Gráfica derecha: Distribución geográfica de la proporción de llamadas totales de usuarios al servicio de soporte técnico.

En el primer gráfico de la figura 3 se observa que Tigo tiene dividido el país en 6 zonas, estas son: Bogotá, Costa, Noroccidente, Occidente, Oriente y las zonas donde no hay cobertura. En esta última zona pertenece gran parte de las regiones que están en el sur del país y la región de la Guajira, se dice que no hay cobertura porque desde octubre del 2020 a marzo del 2021 no se registraron llamadas a soporte técnico por temas relacionados a aseguramiento o aprovisionamiento de los servicios de Tigo Home. En el segundo gráfico de la figura 3 se observa el porcentaje total de llamadas que ingresaron a soporte técnico por temas relacionados a aseguramiento o aprovisionamiento de los servicios de Tigo Home, donde la zona del Noroccidente del país es donde se concentra el 50 % de todas las llamadas, las zonas de Costa y Occidente concentran entre un 25 a 15 % de las llamadas y las zonas con menor número de llamadas fueron Bogotá y Oriente.

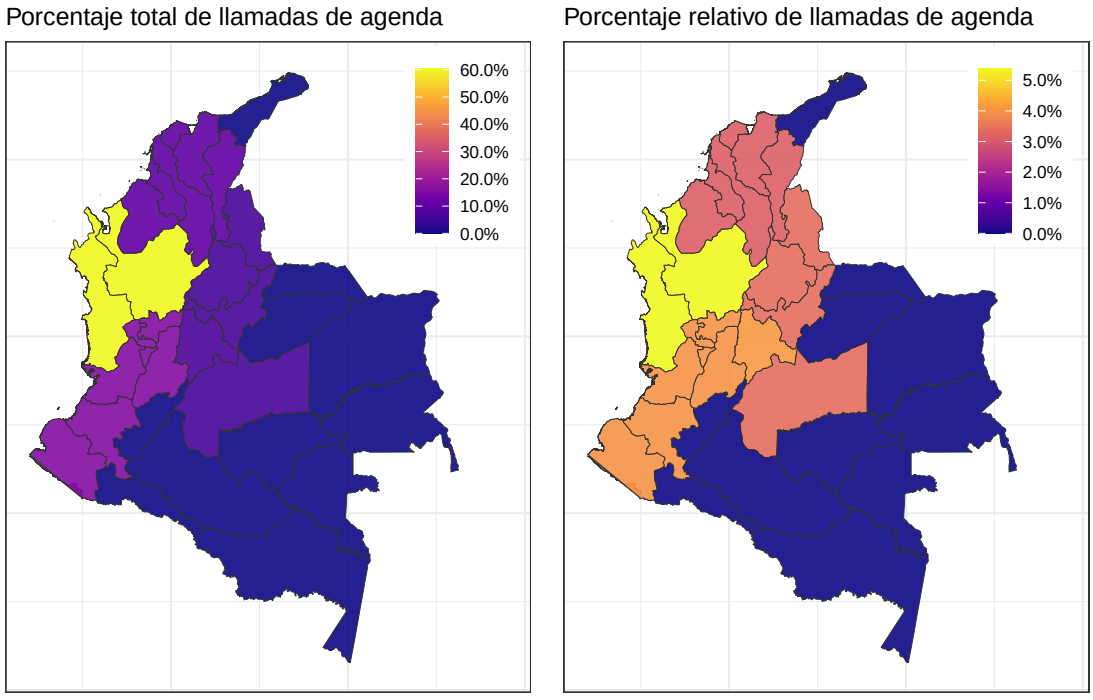

Figura 4: Gráfica izquierda: Distribución geográfica de la proporción de llamadas de agenda al servicio de soporte técnico de Tigo. Gráfica derecha: Distribución geográfica del porcentaje relativo de llamadas de agenda (con respecto al total de llamadas).

El primer gráfico de la figura 4 muestra el porcentaje total de llamadas de agenda a ingresaron a soporte técnico de Tigo Home, donde más del 50 % de las llamadas se concentran en la zona del Noroccidente del país, las zonas de Costa y Occidente concentran entre un 25 a 15 % de las llamadas y las zonas que concentran un 10 % de las llamadas fue Bogotá y Oriente. En el segundo gráfico de la figura 4 se observa el porcentaje relativo de llamadas de agenda, es decir, el porcentaje de llamadas en el total de llamadas que son de agenda, donde el 5 % del total de llamadas que ingresan a soporte técnico por temas de aprovisionamiento y aseguramiento son de agenda y provienen de la zona del Noroccidente del país y entre un 2 a 4 % del total de llamadas pertenecen a las otras cuatro zonas del país.

La primera gráfica de la figura 5 muestra el porcentaje de llamada por estrato socioeconómico, donde las personas que viven en el estrato 2 con un 37.36 % son las que más se comunican con soporte técnico, seguido del estrato 3 y 4 con un 28.36 % y un 15.97 % respectivamente y las personas que menos se comunican a soporte técnico de Tigo Home por temas de aseguramiento y aprovisionamiento son las que viven en el estrato 6 y 5.

La gráfica de porcentaje de conformación HH indica los servicios de Tigo Home que tiene los usuarios. Con un 62.9 % los usuarios con los servicios de televisión-internet banda ancha- telefonía son los que más llaman a soporte técnico de Tigo, seguido de los usuarios con televisión-internet banda ancha con un 15.41 % y los usuarios con internet banda ancha con un 10.23 %. Los usuarios con los servicios de televisión-telefonía son los que menos se comunican a soporte técnico.

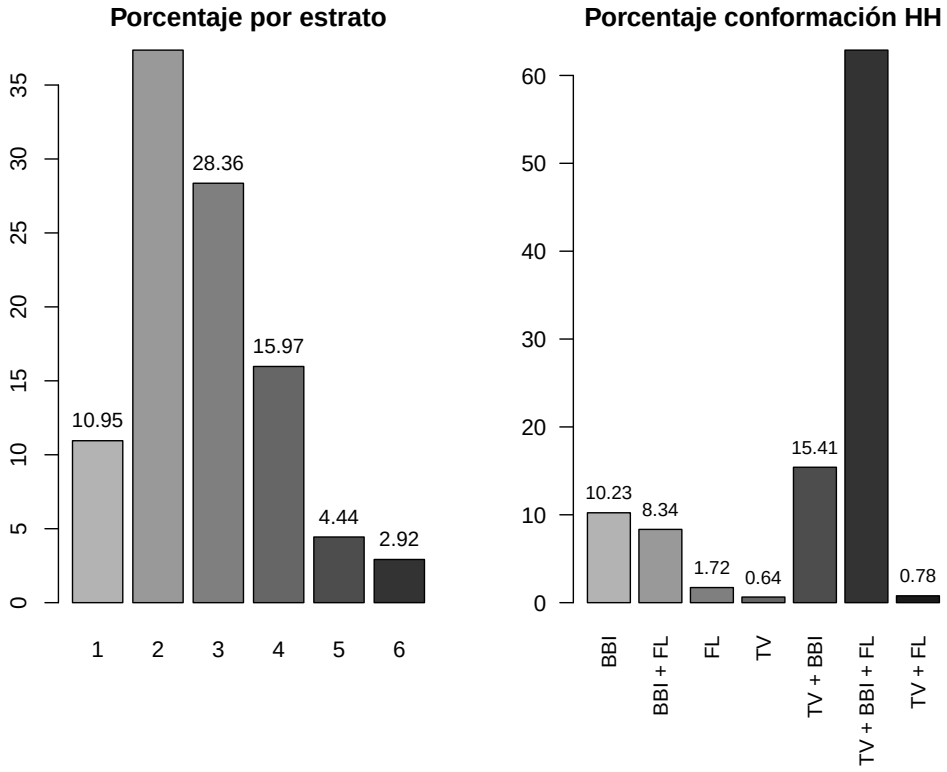

Figura 5: Porcentaje de las variables Región y la variable que indica los servicios que tiene el usuario (conformación HH).

En la figura 6 la gráfica de porcentaje perfil digital, muestra que Tigo clasifica a sus usuarios en cinco perfiles, los que indican qué tan digitalizado es el usuario. El 63.26 % de los usuarios que llaman a soporte técnico de Tigo Home por temas de aseguramiento y aprovisionamiento tienen un perfil tradicional, es decir, que la mayoría de los usuarios no utilizan herramientas tecnológicas a la hora de hacer diligencias referentes a los servicios que tienen con Tigo. También se observa que gran parte de los usuarios que se

comunican con soporte técnico de Tigo Home tienen perfil Full- digital, digital e híbrido, mostrando que estos usuarios utilizan herramientas tecnológicas a la hora de hacer sus diligencias.

Por otra parte, la gráfica de porcentaje producto homologado, hace referencia a lo que se le esta instalando o reparando al usuario en sus servicios de Tigo Home, la gráfica muestra que un 37.45 % de los usuarios que se comunican a soporte técnico de Tigo Home se le está instalando o reparando el servicio de Internet, el segundo servicio por el que mas se comunican es internet-televisión- telefonía con un 20.87 %, seguido del servicio de televisión con un 18.57 % y el servicio de televisión satelital es por el que menos se comunican los usuarios.

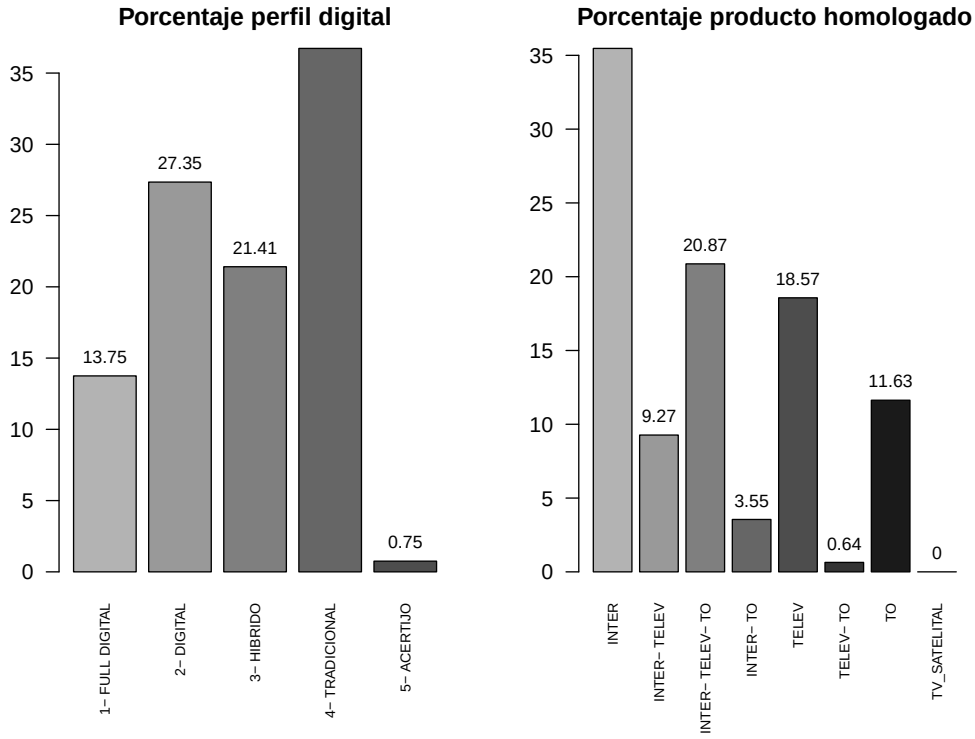

Figura 6: Porcentaje de las variables Perfil Digital y Producto Homologado.

El 29.42 % de los usuarios que se comunican a Tigo Home por temas relacionados a aseguramiento o aprovisionamiento tienen los servicios fijo y móvil y el 70.58 % de los usuarios tienen uno de los dos servicios. Al observar las variables clasificadoras se evidencia que un 38.5 % de las personas llaman por segunda vez a soporte técnico después de haberles agendado una cita, pero solo 6.22 % de estas llamadas son relacionadas por temas de agendamiento. El 60.12 % de los usuarios que se contactan a soporte técnico de Tigo Home por temas relacionados a aprovisionamiento o aseguramiento poseen tres servicios y un 23.12 % y un 12.62 % de los usuarios poseen dos y un servicio respectivamente.

5.2. Árboles de clasificación

Se pretende realizar dos árboles de clasificación, uno para la base de datos de Aprovisionamiento (instalación) que consta de 867286 datos y otro para la base de datos de Aseguramiento (reparación) con 689290 datos, ambas bases de datos tienen 9 variables.

Recuérdese que la variable dependiente es llamada agenda, la cual consta de dos clases, la primera clase (0) indica que el usuario volvió a llamar, pero no por temas relacionados a agendamiento y la segunda clase (1) hace referencia a que el usuario llamó por segunda vez por temas relacionados al agendamiento de la cita ya programada.

El objetivo es identificar qué variables pueden influir en que un usuario de Tigo Home se vuelva a comunicar con soporte técnico por segunda vez por temas relacionados al agendamiento de la cita que fue programada en la primera llamada.

5.2.1. Árboles de clasificación aprovisionamiento

Los datos de aprovisionamiento o instalación hacen referencia a los usuarios de Tigo Home que requieran visita de un técnico para la instalación de telefonía, internet o televisión en su lugar de residencia. Se procede a verificar la proporción de clases en la variable dependiente llamada agenda.

Sesgo de clase

0	1
828031	39255

Se evidencia que la proporción de llamadas reincidentes por temas que no están relacionados al agendamiento es aproximadamente 21 veces más que el número de llamadas reincidentes por agenda. Partiendo de lo dicho anteriormente, se procede a muestrear las observaciones en proporciones aproximadamente iguales para así poder obtener mejores modelos.

Árbol de clasificación

El primer nivel del árbol es sobre el servicio que se le está prestando al usuario (trabajo homologado). El lado derecho de la gráfica muestra que el 36 % de las llamadas reincidentes son por temas de agendamiento; y que los servicios que se le están prestando a los usuarios son: adición de extensión, cambio de algún plan que requiera visita técnica, productos de UNE tv, garantía y traslado. De estas llamadas, el 70 % no son por agendamiento y el 30 % si lo son.

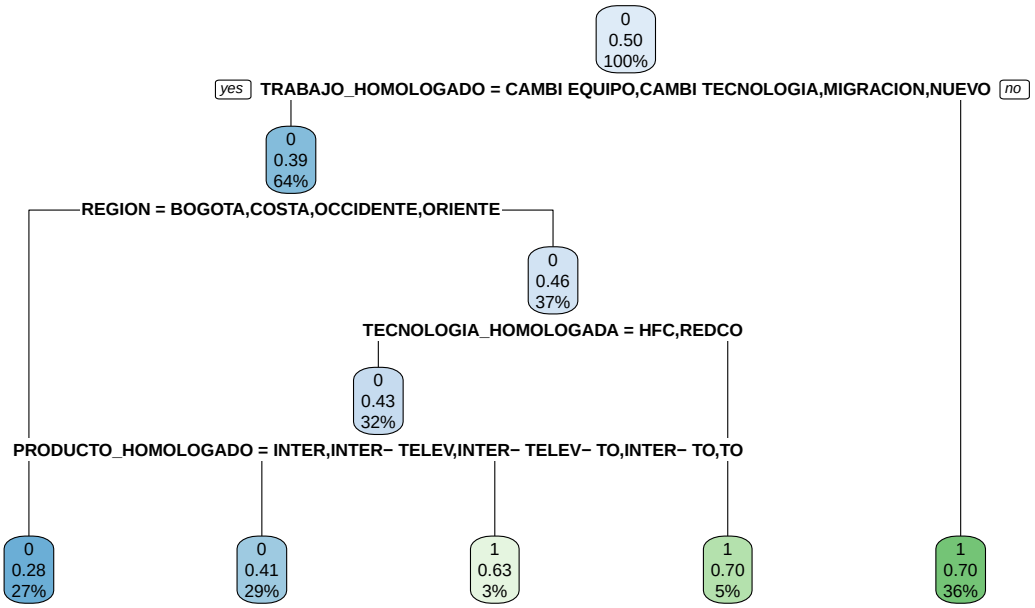

Figura 7: Arboles de decisión aprovisionamiento

Por otro lado, un 64 % de las llamadas reincidentes no son por temas de agendamiento. A este porcentaje de usuarios que están llamando, los servicios que se les están prestando son cambio de equipo, cambio de tecnología, migración a una nueva tecnología y servicios nuevos. De estas llamadas, el 39 % no son por agendamiento y el 61 % si lo son.

El segundo nivel hace referencia a la región del país a la que pertenecen los usuarios que se les está prestando los servicios de cambio de equipo, cambio de tecnología, migración y nuevo, abarcando un 64 % de los datos. El 27 % de estos usuarios pertenecen a las regiones de Bogotá, costa, occidente y oriente. En estas regiones, un 72 % de las llamadas si son por agendamiento y el 28 % no son. El 37 % restante pertenecen a la región del noroccidente del país. De estas llamadas, un 46 % no son por agendamiento, y un 53 % si lo son.

En el tercer nivel se relaciona información de los usuarios de la región noroccidental y el tipo de tecnología que tienen en los servicios de Tigo Home (tecnología homologada). El 5 % de los usuarios utilizan en sus hogares GPON, HFC, 2REDCO y REDCO. De estos, el 70 % de las llamadas no son por agendamiento y El 30 % restante si lo sonon. El 32 % restante de las personas, tienen como tecnología en sus hogares HFC y REDCO. De estos, el 43 % de las llamadas no son por agendamiento y El 57 % restante si lo son.

El último nivel hace referencia a los servicios que se le está instalando o reparando (producto homologado) al usuario, donde el 29 % de las personas tienen como tecnología HFC y REDCO y les instalan los servicios de internet, internet-televisión, internet-televisión-telefonía, internet-telefonía y telefonía. De estos, el 41 % de las llamadas no son por tema de agendamiento y el 59 % restante si lo son. El 3 % restante pertenece a las personas que tienen como tecnología en sus hogares HFC y REDCO y le están instalando los servicios de televisión y televisión-telefonía. De estos, el 37 % de las llamadas son por temas relacionados con agendamiento y el 63 % restante no lo son.

Evaluación del modelo

Para verificar que tan asertiva es la predicción de este modelo, utiliza la matriz de confusión, ya que esta es una de las métricas más intuitivas y sencillas para encontrar la precisión y exactitud de un modelo.

		Referencia	
		0	1
Predicción	0	582412	1564
	1	210290	2362

Precisión	0.7341	Valor-p del test de Mcnemar	<2e-16
95 % CI	(0.7331, 0.735)	Prevalencia de detección	0.73306
Valor-p	1	Valor predictivo positivo	0.99732
Kappa	0.0123	Valor predictivo negativo	0.01111
Sensibilidad	0.73472	Tasa de ausencia de información	0.9951
Especificidad	0.60163	Clase positiva	0
Prevalencia	0.99507		

En este modelo tiene 796628 casos, y se clasifican correctamente 582412 casos pertenecientes a las llamadas reincidentes que no fueron por temas de agendamiento y 2362 casos pertenecientes a las llamadas reincidentes que si fueron por temas de agendamiento. Por tanto, hay 584774 clasificaciones correctas, esto equivale a un 73.41 % a la hora de ver si una persona llama, o no, por temas relacionados al agendamiento de su cita. Este porcentaje tiene un intervalo de confianza del 95 % de 0.7331 y 0.735, lo que significa que hay una probabilidad del 95 % de que la verdadera precisión de este modelo se encuentra dentro de este rango.

La tasa de ausencia de información es del 0.995. Esta hace referencia a la precisión que se puede lograr al predecir siempre la etiqueta de la clase mayoritaria. En este caso, si se le pide que prediga si una persona llamará o no por temas de agenda, entonces se tiene que un 99.5 % de los usuarios no llaman por temas de agenda. Por tanto, la mejor suposición sin otra información es elegir la clase mayoritaria. Se evidencia que este clasificador identifica correctamente las llamadas reincidentes que no son por temas de agendamiento en un 73.47 % y en un 60.16 % que si lo son.

Se evidencia que la tasa de falsos positivos es de un 1 % (es decir, se predice que el usuario no llamó por temas relacionados con agendamiento, pero si lo hizo) y la tasa de verdaderos positivos es de un 99.7 % (es decir, se predice que el usuario no llama por temas relacionados con agendamiento y esto fue cierto). El porcentaje de casos positivos en la muestra es del 99.5 % y la prevalencia de detección

es de un 73.3 %, esta hace referencia a todos los positivos predichos en toda la población. El presente modelo toma como clase positiva a las llamadas reincidente que no son por temas relacionados con agendamiento.

El coeficiente Kappa dice qué tan bien coinciden nuestras predicciones de clasificadores con las etiquetas de clase reales, mientras se controla la precisión de un clasificador aleatorio. Para este trabajo el coeficiente es de 0.0123, lo que significa que la matriz es 1.23 % mejor que lo que podría resultar de aplicar un clasificador aleatorio.

La prueba de McNemar verifica si hay una diferencia significativa entre los falsos positivos y los falsos negativos. El *p*-valor del modelo es igual a 2e-16 este valor es menor a 0.05, lo que significa que hay una diferencia significativa entre los falsos positivos (es decir, se predijo que el usuario no llama por temas relacionados con agendamiento, pero si lo hace) y los falsos negativos (es decir, que se predice que el usuario llama por temas relacionados con agendamiento y esto fue cierto).

Variables importantes

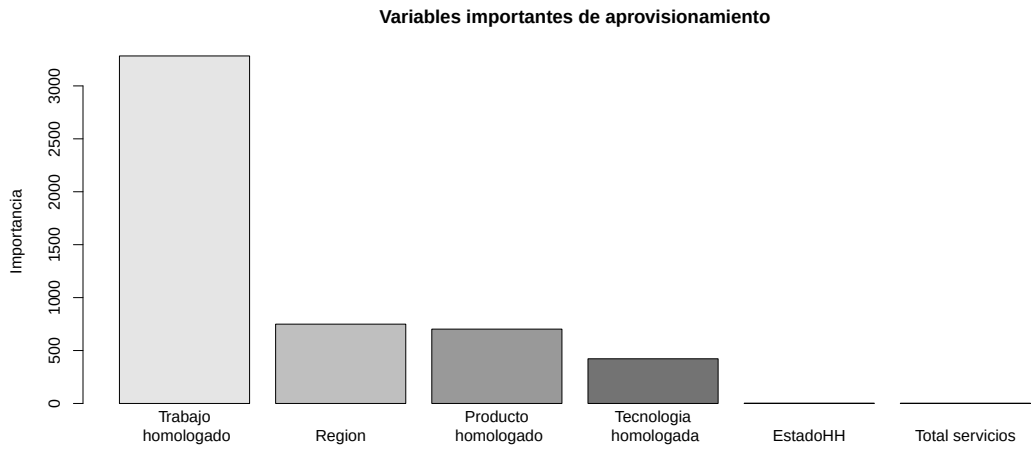

Figura 8: Variables importantes de aprovisionamiento

En la gráfica 8 muestra que trabajo homologado, región, producto homologado, tecnología homologada, estado HH y total de servicios son las variables más importantes a la hora de predecir si una persona vuelve a llamar a soporte técnico por temas relacionados al agendamiento. La variable con mayor importancia es trabajo homologado la que indica cuál es el servicio que se le está prestando al usuario, es decir, si es nuevo, cambio de algún plan, traslado, si es producto de UNE tv, entre otros. La segunda, tercera y cuarta variable con mayor importancia son: la región a la que pertenece la persona, el producto o los productos que se le están instalando y el tipo de tecnología que tiene el usuario en su vivienda. Las dos variables menos importantes son las que indican el estado del usuario y el total de servicios que este tiene con la empresa.

5.2.2. Árboles de clasificación aseguramiento

Los datos de seguramiento o reparación hacen referencia a los usuarios de Tigo Home que requieran visita de un técnico para la reparación de telefonía, internet o televisión en su lugar de residencia. Se procede a verificar la proporción de clases en la variable dependiente llamada agenda.

Sesgo de clase

0	1
631689	57601

Se evidencia que la proporción de llamadas reincidentes por temas que no están relacionados al agendamiento es aproximadamente 11 veces más que el número de llamadas reincidentes por agenda. Se procede a muestrear las observaciones en proporciones aproximadamente iguales para así poder obtener mejores modelos.

Árbol de clasificación

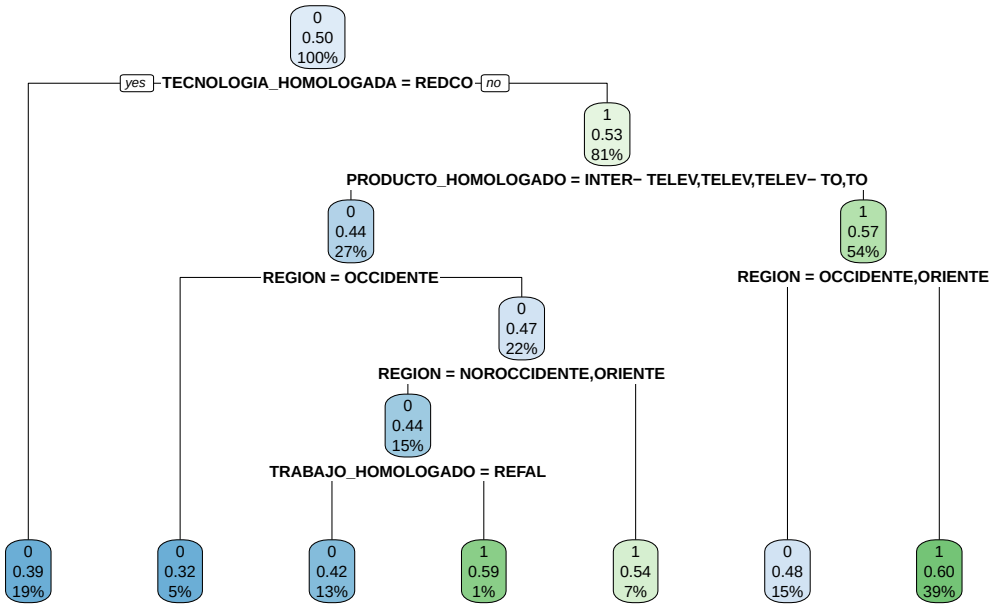

Figura 9: Arboles de decisión aseguramiento

Para interpretar la gráfica anterior hay que tener en cuenta que la variable dependiente se divide en dos clases, la **clase 0** significa que la llamada reincidente no es por temas de agendamiento y la **clase 1** significa que la llamada es por temas relacionados al agendamiento de la cita que ya había sido programada.

El primer nivel del árbol es sobre el tipo de tecnología que tiene el usuario en sus servicios de Tigo Home (tecnología homologada). El lado izquierdo de la gráfica indica que el 19 % de las llamadas reincidentes no son por temas de agendamiento. Este 19 % de los usuarios tienen como tecnología REDCO en sus hogares. De estas llamadas, el 39 % no son por agendamiento y el 61 % si lo son.

Se observa en el lado derecho que un 81 % de las llamadas reincidentes son por temas relacionados con agendamiento. A este porcentaje de usuarios que están llamando, la tecnología que tienen en sus hogares es GPON, HFC y REDCO. De estas llamadas, el 53 % no son por agendamiento y el 47 % si lo son.

El segundo nivel hace referencia a los usuarios que tienen como tecnología en sus hogares GPON, HFC y REDCO y se les están reparando o instalando diferentes productos (producto homologado). Estos abarcan un 81 % de los datos. Al 54 % de estos usuarios se le están reparando o instalando los servicios de internet, internet-telefonía. De estas llamadas el 57 % no son por agendamiento y el 43 % si lo son. Corresponde a este aparte, una de las divisiones del tercer nivel que se refiere a la categoría de Región, en la que un 39 % de las personas pertenecen a las regiones de Bogotá, costa y noroccidente. De estas llamadas el 60 % no son por agendamiento y el 40 % si lo son. Y el 15 % restante de los usuarios pertenecen a las regiones de occidente y oriente del país. De estas llamadas el 48 % no son por agendamiento y el 52 % si lo son.

Al analizar en el 27 % de los datos restantes del segundo nivel, que hace referencia a quienes se le están reparando o instalando los servicios de internet-televisión, televisión, televisión-telefonía y telefonía, se encuentra que: De estas llamadas el 44 % no son por agendamiento y el 56 % si lo son. La segunda división del tercer nivel se refiere a la categoría de Región, en la que un 5 % de las personas pertenecen a la región del occidente del país. De estas llamadas el 32 % no son por agendamiento y el 68 % si lo son. Y el 22 % restante de los usuarios pertenecen a las regiones de Bogotá, costa, noroccidente y oriente del país. De estas llamadas el 47 % no son por agendamiento y el 53 % si lo son.

El cuarto nivel hace referencia a la región a la que pertenecen los usuarios, donde un 7 % de las personas pertenecen a las regiones de la costa y Bogotá. De estas llamadas el 44 % no son por agendamiento y el 46 % si lo son. El 15 % restante de las personas pertenecen a las regiones del noroccidente y oriente del país. De estas llamadas el 44 % no son por agendamiento y el 56 % si lo son.

El último nivel se hizo sobre los servicios que se le está prestando al usuario (trabajo homologado), donde el 13 % de los usurarios se les están prestando los servicios de REFAL. De estas llamadas el 59 % no son por agendamiento y el 41 % si lo son. El 1 % restante pertenece a las personas que se le están prestando los servicios de producto UNE tv. De estas llamadas el 42 % no son por agendamiento y el 57 % si lo son.

Evaluación del modelo

		Referencia	
		0	1
Predicción	0	350048	2399
	1	229801	3362

Precisión	0.6035	Valor-p del test de Mcnemar	<2e-16
95 % CI	(0.6022, 0.6047)	Prevalencia de detección	0.60185
Valor-p	1	Valor predictivo positivo	0.99319
Kappa	0.0091	Valor predictivo negativo	0.01442
Sensibilidad	0.60369	Tasa de ausencia de información	0.9902
Especificidad	0.58358	Clase positiva	0
Prevalencia	0.99016		

En el presente modelo se tiene 585610 casos, y se clasificaron correctamente 350045 casos pertenecientes a las llamadas reincidentes que no son por temas de agendamiento y 3362 casos pertenecientes a las llamadas reincidentes que son por temas de agendamiento. Por tanto, hay 353407 clasificaciones correctas, esto equivale a un 60.35 % a la hora de ver si una persona llama, o no, por temas relacionados al agendamiento de su cita. Este porcentaje tiene un intervalo de confianza del 95 % de 0.6022 y 0.6047, lo que significa que hay una probabilidad del 95 % de que la verdadera precisión de este modelo se encuentra dentro de este rango.

La tasa de ausencia de información es del 0.990. Esta hace referencia a la precisión que se puede lograr al predecir siempre la etiqueta de la clase mayoritaria. En este caso, si se le pide que prediga si una persona llama, o, no por temas de agenda, entonces se tiene que un 99.0 % de los usuarios no llaman por temas de agenda. Por tanto, la mejor suposición sin otra información es elegir la clase mayoritaria. Se evidencia que el clasificador identifica correctamente las llamadas reincidentes que no son por temas de agendamiento en un 60.37 % y en un 58.36 % a las que si son.

Por su parte la tasa de falsos positivos es de un 1.4 % (es decir, se predijo que el usuario no llama por temas relacionados con agendamiento, pero si lo hace) y la tasa de verdaderos positivos es de un 99.3 % (es decir, se predijo que el usuario no llama por temas relacionados con agendamiento y esto fue cierto). El porcentaje de casos positivos en la muestra es del 99.3 % y la prevalencia de detección es de un 59.7 %, esta hace referencia a todos los positivos predichos en toda la población. Este modelo toma como clase positiva a las llamadas reincidente que no son por temas relacionados con agendamiento.

El coeficiente Kappa dice qué tan bien coinciden las predicciones de clasificadores con las etiquetas de clase reales, mientras se controla la precisión de un clasificador aleatorio. El coeficiente arrojado es de 0.091, lo que significa que la matriz es 9.1 % mejor que lo podría resultar de aplicar un clasificador aleatorio.

La prueba de McNemar verifica si hay una diferencia significativa entre los falsos positivos y los falsos negativos. El p-valor de este modelo es igual a 2e-16 este valor es menor a 0.05, lo que significa que hay una diferencia significativa entre los falsos positivos (es decir, se predijo que el usuario no llama por temas relacionados con agendamiento, pero si lo hace) y los falsos negativos (es decir, que se predijo que el usuario llama por temas relacionados con agendamiento y esto fue cierto).

Variables importantes

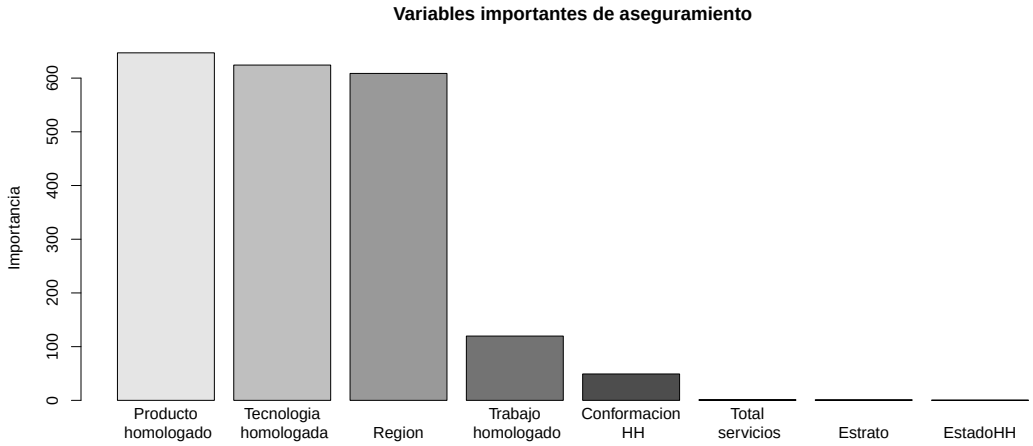

Figura 10: Variables importantes de aseguramiento.

La gráfica 10 muestra que producto homologado, tecnología homologada, región, trabajo homologado, conformación HH, total de servicios, estratos y estado HH son las variables más importantes a la hora de predecir si una persona vuelve a llamar a soporte técnico por temas relacionados al agendamiento. Las tres variables con mayor importancia son: producto homologado, está indica qué producto o productos se le están instalando al usuario, el tipo de tecnología que tiene el usuario en su vivienda y la región a la que pertenecen. La variable 4 (trabajo homologado) indican cuál es el servicio que se le está prestando al usuario, es decir, si es nuevo, cambio de algún plan, traslado, si es producto de UNE tv, entre otros. La variable 5 (conformación HH) indica los servicios que tiene el usuario. Las tres variables menos importantes son las que indican el estado del usuario, el total de servicios que este tiene con la empresa y el estrato socioeconómico al que pertenece.

5.3. Regresión logística

Se pretende realizar dos regresiones logísticas, una para la base de datos de Aprovisionamiento (instalación) que con 867.286 datos y otra para la base de datos de Aseguramiento (reparación) con 689.290 datos, ambas bases de datos tienen 9 variables.

La variable dependiente es llamada agenda, la cual consta de dos clases, la primera clase (0) indica que el usuario volvió a llamar, pero no por temas relacionados a agendamiento y la segunda clase (1) hace referencia a que el usuario llamó por segunda vez por temas relacionados al agendamiento de la cita ya programada.

El objetivo es identificar qué variables pueden influir en que un usuario de Tigo Home se vuelva a comunicar con soporte técnico por segunda vez por temas relacionados al agendamiento de la cita que fue programada en la primera llamada (clase 1).

5.3.1. Regresión logística aprovisionamiento

Los datos de aprovisionamiento o instalación hacen referencia a los usuarios de Tigo Home que requieran visita de un técnico para la instalación de telefonía, internet o televisión en su lugar de residencia. Al igual que en el árbol de clasificación se verifica la proporción de clases en la variable dependiente llamada agenda, como se muestra en la sección 5.2.1 en la parte de sesgo de clase.

Regresión logística

Para la interpretación de los coeficientes estimados para el modelo de regresión logística con los datos de aprovisionamiento (instalación), se tendrá en cuenta un nivel de significancia del 5 %, para observar qué variables influyen si un usuario volverá a contactar con soporte técnico por segunda vez por temas relacionados al agendamiento de la cita. Se procede a analizar cada variable y a comparar su valor-p con el nivel de significancia que se había establecido anteriormente. Esto se hace con el fin de saber si la variable es significativa o no.

La variable que indica qué servicio se le está reparando o instalando al usuario (producto homologado) toma como nivel de referencia **Internet** e indica que los servicios de Internet-Televisión-Telefonía, Internet-Telefonía y Telefonía son significativos y que están por debajo del nivel de referencia. Esto quiere decir, si un usuario posee alguno de estos servicios, entonces la probabilidad de que este se vuelva a contactar con soporte técnico es inferior al $1/(1+exp(-(\beta_0+\beta_1 x)))*100$ =43.21 %, conservando fijas las otras variables. Los servicios de Televisión y Televisión-Telefonía indican que son significativos y están por encima del nivel de referencia. Esto quiere decir, si un usuario posee alguno de estos servicios, entonces la probabilidad de que este se vuelva a contactar con soporte técnico es mayor al 60.5 %. Por otro lado, Televisión-Internet fue el único servicio que diferente al nivel de referencia, es decir, la probabilidad de que un usuario que posea este servicio y que se comunique con soporte técnico por temas de agendamiento es 45.5 %.

La variable que indica el servicio que se le está prestando al usuario (trabajo homologado) toma como nivel de referencia **adición de extensión** e indica que cambió de tecnología y los nuevos servicios que se le están instalando al usuario son significativos y que están por debajo del nivel de referencia. Esto quiere decir, que si un usuario posee alguno de estos servicios, entonces la probabilidad de que este se vuelva a contactar con soporte técnico es inferior al 40.1 %. Los servicios de migrar a una nueva tecnología, los productos de UNE tv, traslado, y algún cambio que requiera la visita de un técnico, indican que son significativos y están por encima del nivel de referencia, es decir, si un usuario posee alguno de estos servicios, entonces la probabilidad de que este se vuelva a contactar con soporte técnico es mayor al 63.21 %. Pero los servicios relacionados con la garantía y cambio de equipo son significativamente diferentes al nivel de referencia, es decir, que es muy poco probable que un usuario que posea estos servicios se vuelva a comunicar con soporte técnico por temas de agendamiento.

La variable región y la variable que indica el tipo de tecnología que tiene el usuario en su residencia (tecnología homologada) toman como nivel de referencia **Bogotá** y **tecnología GPON** y dice que las regiones occidente, oriente y costa y las tecnologías HFC y REDCO son significativas y están por debajo del nivel de referencia. Eso quiere decir que, si un usuario vive en alguna de estas regiones, la probabilidad de que este se vuelva a contactar con soporte técnico es inferior al 43.37 % y si tiene alguna de estas tecnologías en su vivienda, entonces la probabilidad será inferior al 29.22 %. Por otro lado, la región del noroccidente y las tecnologías HFC-2REDCO y HFC-REDCO son significativas con respecto al nivel de referencia de cada variable, es decir, si un usuario vive en esta región, la probabilidad de que este se vuelva a contactar con soporte técnico es de 55.84 % y si tiene alguna de estas tecnologías en su

vivienda, entonces la probabilidad de que se vuelva a comunicar será mayor a 72.36 %.

La variable que indica la cantidad de servicios que tiene el usuario (total de servicios) y la variable que hace referencia a si el usuario está ?activo al día de hoy? (estado HH). Ambas variables no son significativas para este modelo, lo que indica que estas variables no arrojan información si un usuario se va a volver a contactar con soporte técnico respecto a temas de agendamiento.

La variable estrato socioeconómico y la variable que indica los servicios que tiene el usuario (conformación HH) toman como nivel de referencia **estrato 1** e **Internet banda ancha**. La variable que indica los estratos 2 y 6 son significativamente diferentes al nivel de referencia, es decir, si un usuario pertenece a alguno de estos estratos socioeconómicos la probabilidad de que este se vuelva a contactar con soporte técnico es inferior a 47.42 %. Los estratos 3, 4 y 5 y los servicios Internet banda ancha-Telefonía, Telefonía, Televisión, Televisión-Internet banda ancha, Televisión-Internet banda ancha-Telefonía y Televisión-Telefonía indican que son significativos y están por encima del nivel de referencia. Esto quiere decir, si un usuario vive en alguno de estos estratos y posee alguno de estos servicios, la probabilidad de que se vuelva a contactar con soporte técnico es alta en comparación con los otros servicios.

Error de clasificación

Error de clasificación	0.0145

El error de clasificación del modelo logístico para los datos de aprovisionamiento de los valores predefinidos frente a los reales es de 1.45 %.

Curva ROC

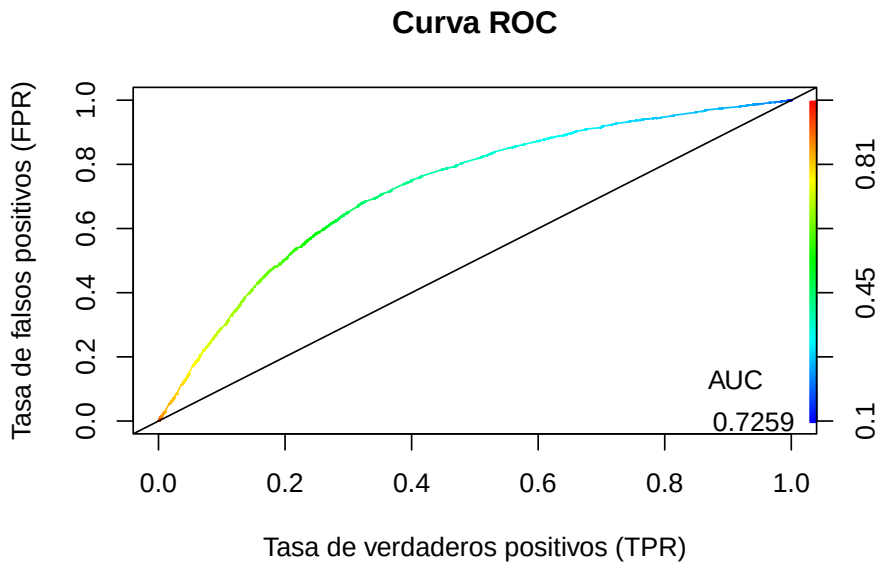

Figura 11: Curva de ROC para los datos de aprovisionamiento.

El área bajo la curva (AUC) es de 0.7258, esto quiere decir que el 73 % de los valores predichos son correctos, es decir, que es una buena probabilidad de clasificar correctamente si una persona se va a volver a contactar con soporte técnico por temas de agendamiento y se tiene que solo el 28 % restante se predice incorrectamente.

Precisión del modelo

Precisión	0.724000098482144

La precisión del modelo es de 0.7240 a la hora de ver si una persona llama o no por temas relacionados al agendamiento de su cita. Por tanto, se considera que es un buen resultado. Sin embargo, hay que tener en cuenta que este resultado depende en cierta medida de la división aleatoria de los datos que se hizo anteriormente.

Capacidad predictiva del modelo

Sensibilidad	0.0002547338
Especificidad	0.99995

El modelo identifica correctamente a las personas que llaman por temas que no están relacionados a agendamiento, esto quiere decir que si el modelo predice que la llamada no es por temas de agendamiento entonces se podría confiar en este resultado. En cambio si el modelo predice que las llamadas reincidentes fueron por temas relacionados al agendamiento, no es bueno confiar ya que la probabilidad de predicción en este caso es muy baja.

5.3.2. Regresión logística aseguramiento

Los datos de seguramiento o reparación hacen referencia a los usuarios de Tigo Home que requieran visita de un técnico para la reparación de telefonía, internet o televisión en su lugar de residencia. Al igual que en el árbol de clasificación se verifica la proporción de clases en la variable dependiente llamada agenda, como se muestra en la sección 5.2.2 en la parte de sesgo de clase.

Regresión logística

Para la interpretación de los coeficientes estimados del modelo de regresión logística con los datos de aseguramiento (reparación), se tendrá en cuenta un nivel de significancia del 5 %, para observar qué variables influyen si un usuario volverá a contactar con soporte técnico por segunda vez por temas relacionados al agendamiento de la cita. Se procede a analizar cada variable y a comparar su valor-p con el nivel de significancia que se había establecido anteriormente. Esto se hace con el fin de saber si la variable es significativa o no.

La variable que indica qué servicio se le está reparando o instalando al usuario (producto homologado) tomó como nivel de referencia **Internet** y muestra que Internet-Televisión, Televisión y Telefonía son significativos y que están por debajo del nivel de referencia. Esto quiere decir, si un usuario posee alguno de estos servicios, entonces la probabilidad de que este se vuelva a contactar con soporte técnico es inferior al 35.85 %. Por otro lado, Internet-Telefonía y Televisión-Telefonía fueron los únicos servicios

diferentes al nivel de referencia, es decir, es muy poco probable que un usuario que posea estos servicios se vuelva a comunicar con soporte técnico por temas de agendamiento.

La variable región y la variable que indica el tipo de tecnología que tiene el usuario en su residencia (tecnología homologada) toman como nivel de referencia **Bogotá** y **tecnología GPON** e indican que las regiones noroccidente, occidente, oriente y costa, y la tecnología REDCO son significativos y están por debajo del nivel de referencia. Esto quiere decir, si un usuario vive en alguna de estas regiones, la probabilidad de que este se vuelva a contactar con soporte técnico es inferior al 45.43 % y si tiene la tecnología REDCO en su vivienda, entonces la probabilidad de que se vuelva a contactar es de 35.23 %. Por otro lado, las tecnologías HFC y HFC-REDCO son diferentes al nivel de referencia, es decir, es muy poco probable que un usuario que posea estos servicios se vuelva a comunicar con soporte técnico por temas de agendamiento.

Las variables que indican el servicio que se le está prestando al usuario (trabajo homologado), la cantidad de servicios que tiene el usuario (total de servicios) y la variable que hace referencia a si el usuario está ?activo al día de hoy? (estado HH) . Las tres variables son significativas para este modelo, lo que indica que estas variables arrojan información de si un usuario se va a volver a contactar con soporte técnico respecto a temas de agendamiento.

La variable estrato socioeconómico y la variable que indica los servicios que tiene (conformación HH) toman como nivel de referencia **estrato 1** e **Internet banda ancha**. Esto indica que los estratos 2, 5 y 6 son significativamente diferentes al estrato 1, es decir, es muy poco probable que un usuario perteneciente a alguno de estos estratos se vuelva a comunicar con soporte técnico por temas de agendamiento. Los estratos 3 y 4 y los servicios de Internet banda ancha-Telefonía, Telefonía, Televisión, Televisión-Internet banda ancha, Televisión-Internet banda ancha-Telefonía y Televisión-Telefonía indican que son significativos y están por encima del nivel de referencia respecto al número de llamadas reincidentes, es decir, que hay mayor probabilidad de que un usuario que posea estos servicios y que viva en alguno de estos estratos socioeconómicos se vuelva a comunicar con soporte técnico por temas de agendamiento.

Error de clasificación

Error de clasificación	0.0284

El error de clasificación del modelo logístico para los datos de aseguramiento de los valores predefinidos frente a los reales es de 2.84 %.

Curva ROC

El área bajo la curva (AUC) es de 0.6106, esto quiere decir que el 61 % de los valores predichos son correctos, es decir, que es una buena probabilidad de clasificar correctamente si una persona se va a volver a contactar con soporte técnico por temas de agendamiento y solo el 39 % restante se predice incorrectamente.

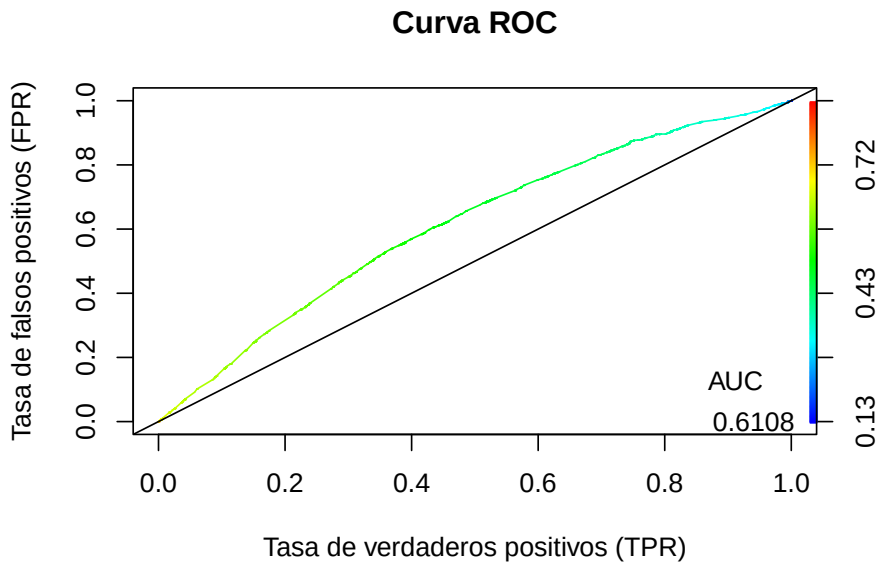

Figura 12: Curva de ROC para los datos de aseguramiento.

Precisión del modelo

Precisión	0.579485747145322

La precisión del modelo es de 0.579 a la hora de ver si una persona llama o no por temas relacionados al agendamiento de su cita. Por tanto, se considera que el resultado no es tan bueno. Sin embargo, hay que tener en cuenta que este resultado depende en cierta medida de la división aleatoria de los datos que se hizo anteriormente.

Capacidad predictiva del modelo

Sensibilidad	0.0001736011
Especificidad	0.9999476

El modelo identifica correctamente a las personas que llaman por temas que no están relacionados a agendamiento, esto quiere decir que si el modelo predice que la llamada no es por temas de agendamiento entonces se podría confiar en este resultado. En cambio si el modelo predice que las llamadas reincidentes fueron por temas relacionados al agendamiento, no es bueno confiar ya que la probabilidad de predicción en este caso es muy baja.

5.4. Serie de tiempo

Por medio de una serie de tiempo se desea analizar, modelar y predecir el número de llamadas reincidentes que entraron a soporte técnico de Tigo Home, cuyo objetivo de la llamada pudo haber sido por cuestiones de agendamiento o por temas diferentes.

La base de datos consta de $N = 182$ datos que son el número de llamadas reincidentes que fueron tomadas diariamente entre los meses de octubre de 2020 a marzo de 2021.

Mínimo	Mediana	Media	Máximo
2	3622	3307	5120

Desde el mes octubre de 2020 hasta marzo del 2021 se observa que en promedio ingresan 3307 llamadas diarías al área de soporte técnico desde todo el territorio Colombiano, siendo 5120 el número máximo de llamadas que se ha registrado en un día y 2 el número mínimo de llamadas que se registraron en un día dentro de este periodo de tiempo.

Serie en bruto

En la siguiente gráfica se hace un acercamiento de la serie en bruto, cuyo objetivo es ver su comportamiento en el tiempo.

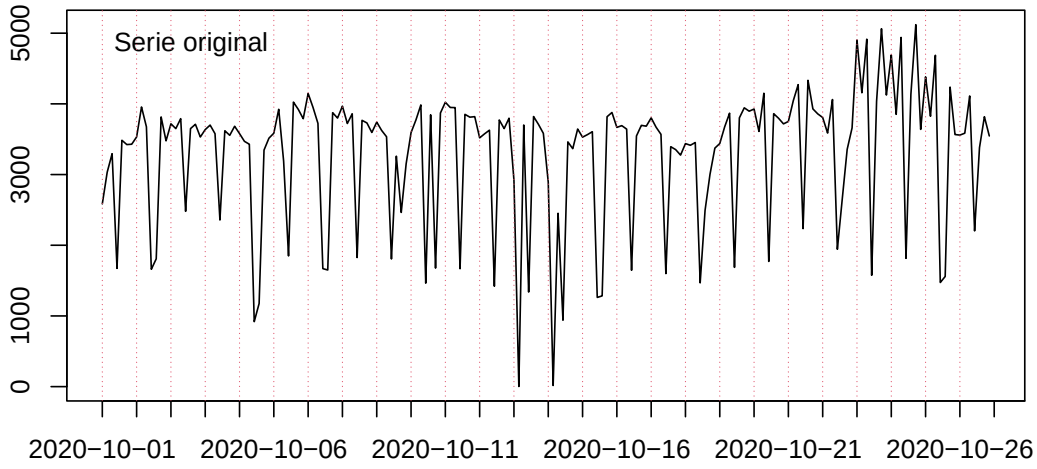

Figura 13: Serie de tiempo en bruto del número de llamadas reincidentes que ingresaron a soporte técnico de Tigo Home

En la serie de tiempo en bruto se observa que no hay tendencia creciente ni decreciente, esto quiere decir que el número de llamadas reincidentes se mantiene oscilando en un mismo rango; también se observa que hay estacionalidad ya que se presentan caídas y picos en ciertos días y la varianza aparentemente se observa constante a través de la serie.

Box-plot

En la siguiente gráfica se hace un acercamiento descriptivo de la serie de tiempo en bruto, cuyo objetivo es ver su comportamiento en el tiempo en los diferentes días de semana.

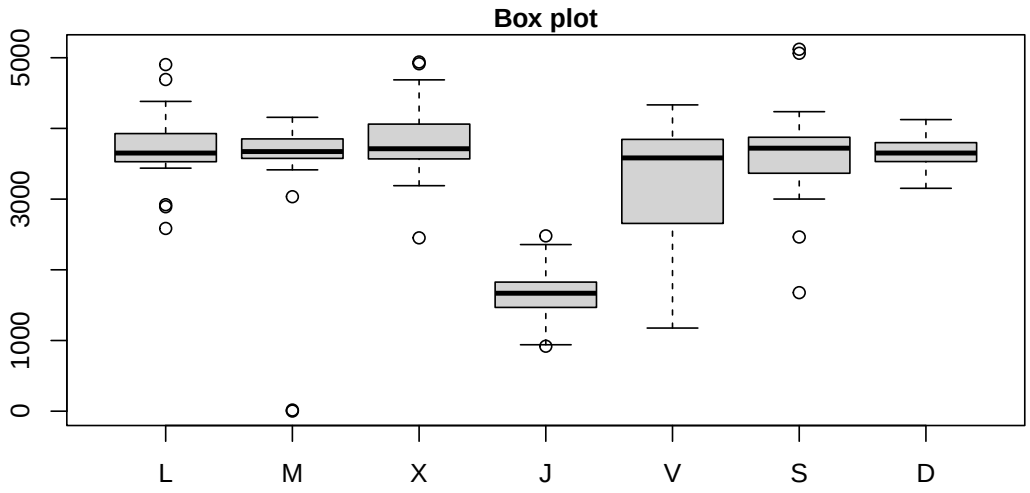

Figura 14: Box-plot.

En el gráfico de Box-plot se observa que en los días jueves se presenta el menor número de llamadas reincidentes, mientras que el resto de los días de la semana mantienen valores relativamente estables entre de 3000 a 5000 llamadas diarias, aunque, respecto al día viernes se observa una mayor dispersión del número de llamadas reincidentes respecto a los demás días, la disposición de las cajas puede indicar distribuciones leptocúrticas (Una distribución es leptocúrtica cuando es más apuntada y con colas menos anchas que la distribución normal) que no ofrecen evidencia para afirmar presencia de normalidad; por otra parte, la mediana coincide en casi todos los días excepto el jueves, tomando valores entre de 3700 y 3900 aproximadamente.

5.4.1. Modelos y ajuste

Para las predicciones se considera $m = 7$, es decir, se toma los últimos 7 días de nuestra base de datos, cuyo propósito es evaluar y hacer predicciones con este periodo de tiempo. Sea t un rango temporal sin considerar las últimas 7 observaciones de la serie.

Modelo 1: Polinomial con indicadoras

$$Y_t = \sum_{i=0}^{3} \beta_i t^i + \sum_{j=1}^{6} \delta_j I_{j,t} + E_t, \quad E_t \sim N(0, \sigma^2).$$

Este modelo es de grado tres en la parte estacional y posee variables indicadoras que toman como nivel de referencia el día domingo.

Modelo 2: AR(18)

$$Y_t = \sum_{i=0}^{3} \beta_i t^i + \sum_{j=1}^{6} \delta_j I_{j,t} + E_t.$$

$$E_t = \sum_{j=1}^{18} \phi_j E_{t-j} + a_t, a_t \sim RB(0, \sigma_a^2).$$

Polinomio autoregresivo:

$$\phi_{18}(B) = 1 - \sum_{j=1}^{18} \phi_j B^j.$$

Modelo 3: ARMA(18,6)

$$Y_t = \sum_{i=0}^{3} \beta_i t^i + \sum_{j=1}^{6} \delta_j I_{j,t} + E_t.$$

$$E_t = \phi_1 E_{t-1} + \phi_{17} E_{t-17} + \phi_{18} E_{t-18} + a_t + \theta_6 a_{t-6}.$$

Polinomios autoregresivo y de media móviles:

$$\phi_{18}(B) = 1 - \phi_1 B - \phi_{17} B^{17} - \phi_{18} B^{18}, \qquad \theta_6(B) = 1 + \theta_6 B^6.$$

Modelo 4: SARIMA I

$$Y_t \sim ARIMA(0,0,2)(0,1,2)_7.$$

$$(1 - B^7) Y_t = (1 + \theta_1 B + \theta_2 B^2)(1 + \Theta_1 B^7 + \Theta_2 B^{14}) E_t, \quad E_t \sim RBN(0, \sigma^2)$$

Modelo

$$\nabla_7 Y_t = E_t + \theta_1 E_{t-1} + \theta_2 E_{t-2} + \Theta_1 E_{t-7} + \Theta_2 E_{t-14}, \quad E_t \sim RBN(0, \sigma^2).$$

Modelo 5: SARIMA II

$$Y_t \sim ARIMA(0,0,1)(0,1,2)_7.$$

$$(1-B^7)Y_t = (1+\theta_1 B+)(1+\Theta_1 B^7 + \Theta_2 B^{14})E_t, \quad E_t \sim RBN(0, \sigma^2)$$

Modelo

$$\nabla_7 Y_t = E_t + \theta_1 E_{t-1} + \Theta_1 E_{t-7} + \Theta_2 E_{t-14}, \quad E_t \sim RBN(0, \sigma^2).$$

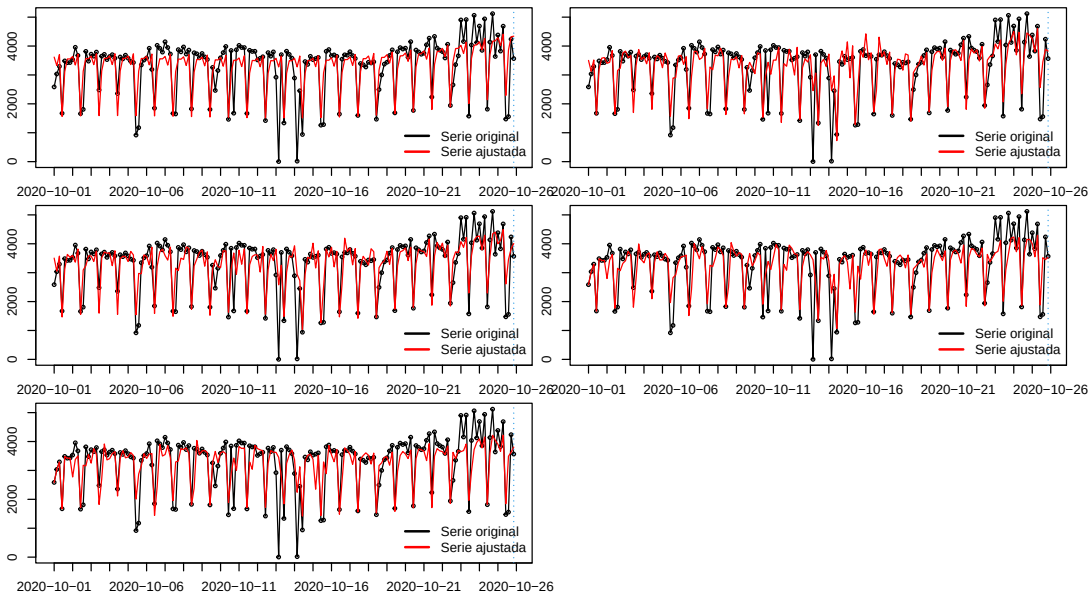

Figura 15: Gráfica de los valores ajustados de los cinco modelos.

El modelo con indicadoras, el modelo AR(18) y el modelo ARMA(18,6) se muestran que el intercepto es la única componente significativa de la tendencia y los únicos días que fueron significativos son el jueves y el viernes, ambos siendo inferiores al nivel de referencia, esto quiere decir, que hay menor número de llamadas reincidentes respecto al día domingo, las únicas componentes significativas de la parte autoregresiva de los modelo AR(18) y ARMA(18,6) son ϕ_1, ϕ_{17}, ϕ_{18}. Se observa que todas las componentes de los modelos 4 y 5 son significativas, es decir, que la parte de medias móviles regular es

significativo y así mismo la parte estacional de medias móviles de uno y de dos. Todo lo anterior respecto a un nivel de significancia del 0.05.

En las gráficas de los ajustes de los cinco modelos planteados, se observa que tienen un buen ajuste a la serie original, es decir, que los cinco modelos explican bien el número de llamadas reincidentes a soporte técnico de Tigo Home.

5.4.2. Criterios de información

	Base	AR(18)	ARMA(18,6)	SARIMA I	SARIMA II
AIC	13.01229	12.99760	12.91666	12.94481	12.95963
BIC	13.19314	13.50397	13.16984	13.01715	13.01388

Con los criterios de información se observa que los modelos SARIMAI y SARIMAII son los que tienen menor AIC y BIC, y también tienen el menor número de parámetros, es decir, que son los más parsimoniosos, por tanto, los cinco modelos planteados podrían explicar el número de llamadas reincidentes en Tigo Home, pero los dos modelos SARIMA son los que pueden hacer mejores predicciones con mayor detalle dentro de los datos y son los que mejor se explican en términos de desviación.

5.4.3. Análisis de residuales

Residuales de los ajustes vs el tiempo

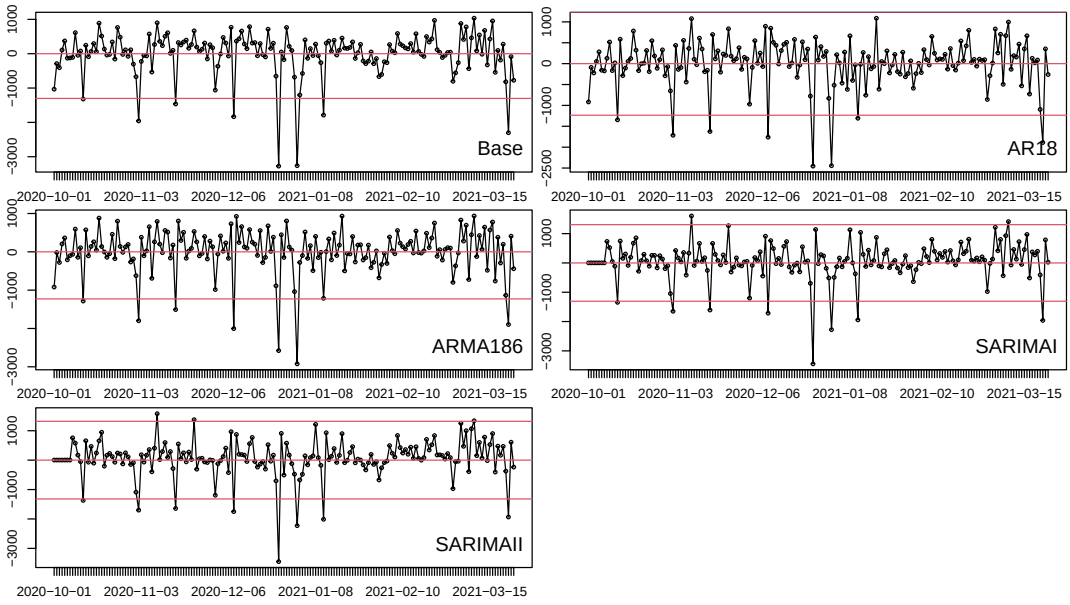

Figura 16: Gráfica de residuales de los valores ajustados vs el tiempo.

En las diferentes gráficas se aprecia un comportamiento de los errores muy parecido a lo largo del tiempo, ya que estos suben y bajan de forma errática alrededor del 0, lo cual indica que no hay presencia de ningún patrón, también se puede ver que hay algunos outliers en todos los modelos. Por tanto, se concluye que no hay evidencia gráfica en contra del supuesto de independencia de los errores respecto al número de llamadas reincidentes.

Residuales de los ajustes vs los valores ajustados

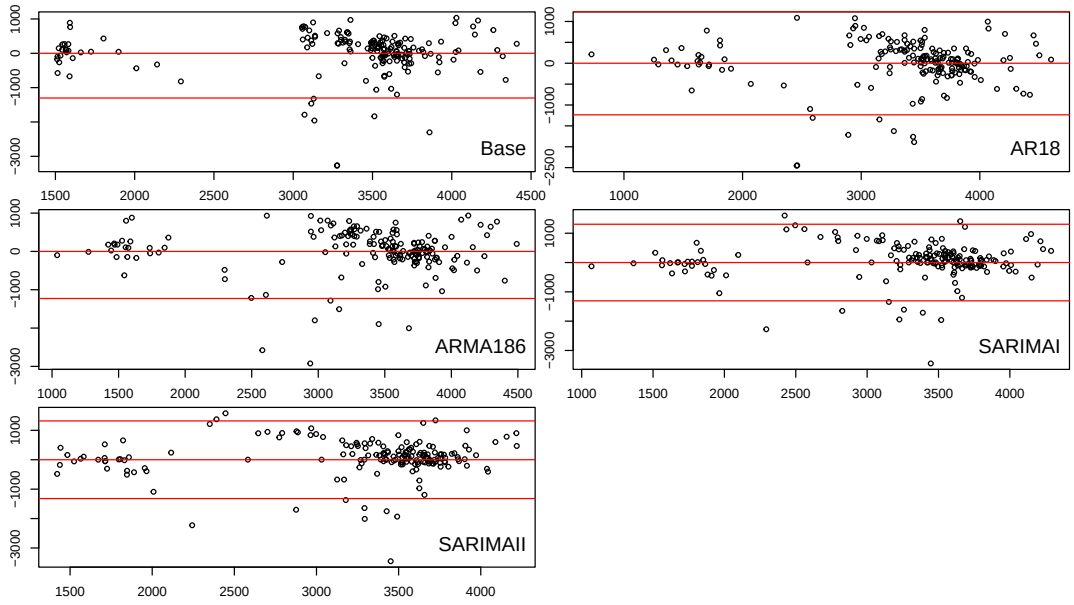

Figura 17: Gráfica de residuales de los valores ajustados vs los valores ajustados.

En las gráficas se observa que no hay evidencia en contra del supuesto de varianza constante, en ninguno de los modelos hay presencia de patrones ni formas de embudo, así que puede que el error no esté influenciado por otras variables. Para verificar lo dicho anteriormente se calcula el ACF de los residuales.

Funciones de autocorrelación (ACF)

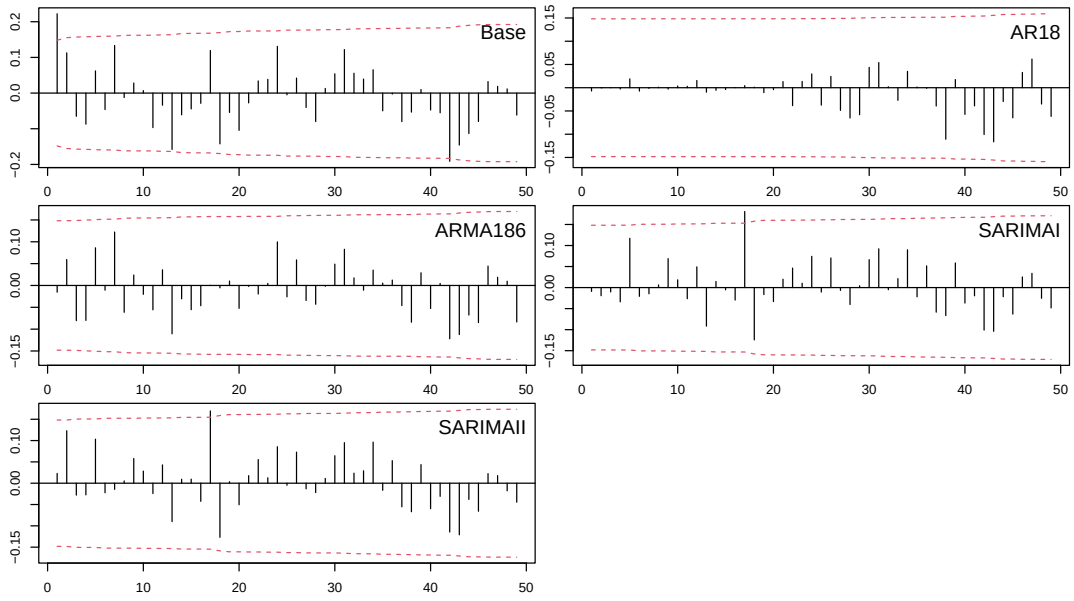

Figura 18: Gráfica función de autocorrelación.

La ACF de los residuales del modelo Base se sale de la región de aceptación en los rezagos temporales $k = 1, 13, 42$, los modelos SARIMA I y SARIMA II se salen en $k = 17$, esto quiere decir que hay evidencia para rechazar la hipotesis nula y en los modelos AR(18) y ARMA(18,6) no hay evidencia para el rechazar la hipotesis nula.

Función de autocorrelación parcial(PACF)

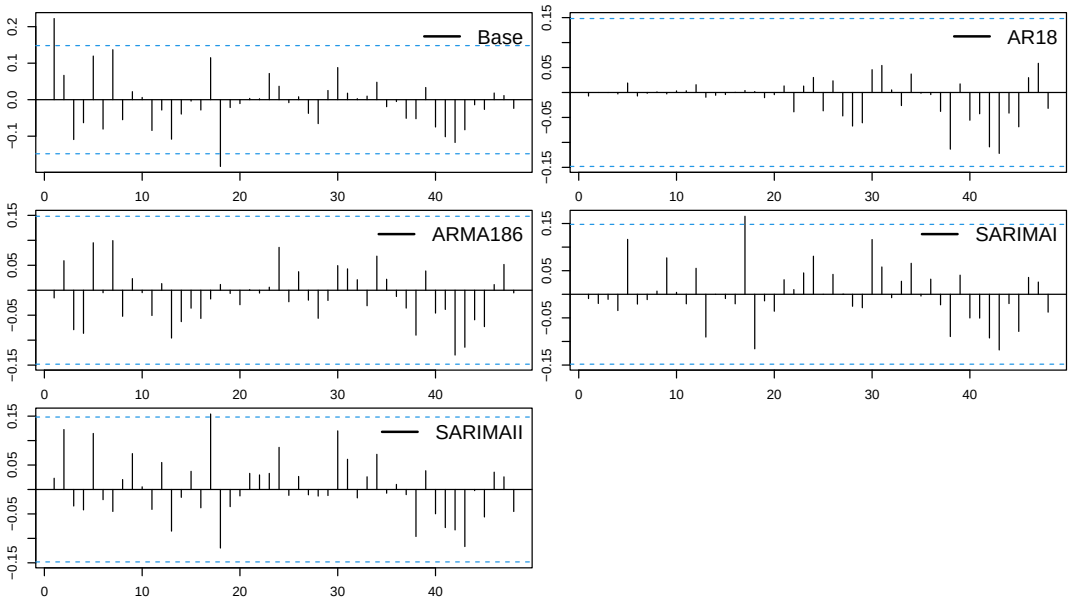

Figura 19: Gráfica función de autocorrelación parcial.

La PACF de residuales del modelo Base excede la región de aceptación en los rezagos $k = 1, 18$, los modelos SARIMA I y SARIMA II se salen de las bandas en $k = 17$ al igual que en el ACF, en ningún de los modelos AR(18) y ARMA(18,6) hay evidencia para el rechazo. Los rezagos temporales pueden salir de las bandas de aceptación debido a la influencia de observaciones extremas en la estimación de ACF y PACF. Es por eso que se hace necesario ejecutar un test estadístico que controle el error de tipo I.

Prueba de ruido blanco

Tests de Ljung-Box

	B Base	B AR(18)	BARMA(18,6)	B SARIMAI	B SARIMAII
m=6	0.02967393	0.9999886	0.6447016	0.8360562	0.5484798
m=12	0.08546164	1	0.7483643	0.9782263	0.9078571
m=18	0.03137578	1	0.8630152	0.7057162	0.607813
m=24	0.04687909	1	0.9467381	0.8870976	0.796396
m=30	0.1267878	1	0.9854149	0.9588408	0.9182805

Test de Box-Pierce

	B Base	B AR(18)	BARMA(18,6)	B SARIMAI	B SARIMAII
m=6	0.026349	0.9999872	0.6256086	0.8224368	0.5290088
m=12	0.07158122	1	0.7153038	0.9730828	0.8946657
m=18	0.01870539	1	0.8260522	0.6089674	0.5124041
m=24	0.02380482	1	0.9181271	0.8168383	0.7028971
m=30	0.06771798	1	0.9714859	0.9149388	0.8520991

Los resultados del test Ljung-Box y Box-Pierce muestran que los $p-$valores de todos los modelos excepto el modelo 1 son mayores a 0.5 esto quiere decir que para ninguno de los modelos se encuentra evidencia en contra del supuesto de ruido blanco para los errores.

Normalidad

Se procede a comparar la normalidad de los residuos para comprobar si los errores, además de ser un ruido blanco, también son un ruido blanco normal. Para ello se dibuja primero los qq-plots y luego, para confirmar los hallazgos, se hace un test de normalidad. En este caso no se va usar el test de Shapiro-Wilks, ya que es sumamente sensible a los outliers; por tanto, se utiliza **el test de Lilliefors** (libraría nortest), este es una modificación del test de Kolmogorov-Smirnov donde la media y la varianza son desconocidas.

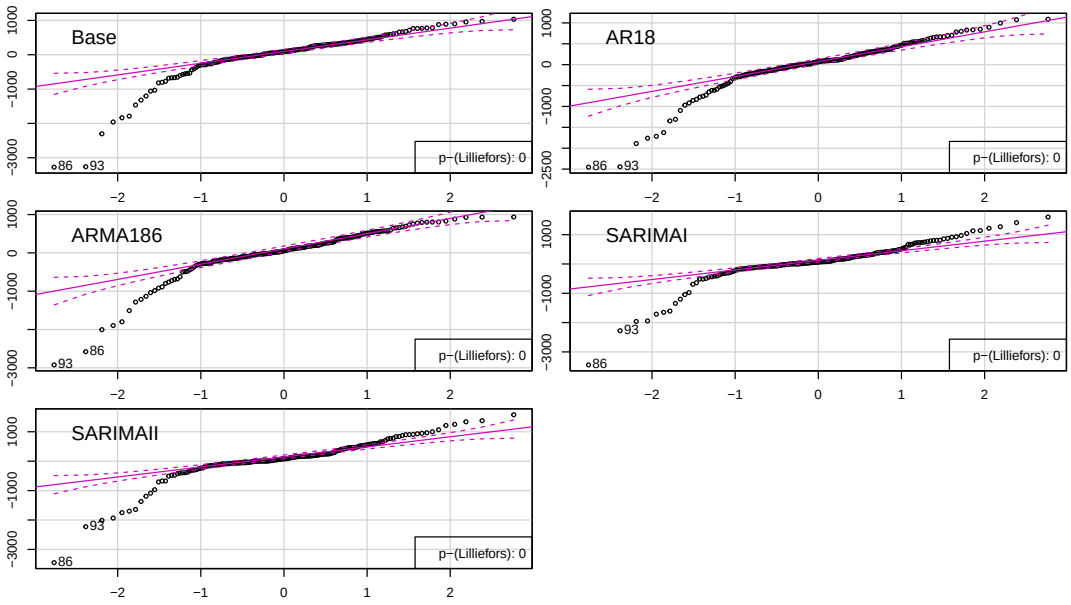

Figura 20: Gráfica normalidad de los residuos y test de Lilliefors.

En las gráficas anteriores se puede observar que los datos no se ajustan tan bien a la línea recta, es decir, que hay algo de desviación con respecto a la recta de probabilidad normal y los valores-p son casi cero, por tanto, se rechaza el supuesto de normalidad de los cinco modelos planteados, es decir, los datos no provienen de una distribución normal.

5.4.4. Forcasting

	ME	RMSE	MAE	MPE	MAPE	Amplitud	Cobertura
Base	-645.346	683.434	645.346	-18.474	18.474	2767.240	100.000
AR(18)	-742.398	831.458	742.398	-21.577	21.577	2517.344	85.714
ARMA(18,6)	-696.663	722.542	696.663	-20.525	20.525	2492.216	100.000
SARIMA I	49.749	379.448	281.264	3.026	9.439	2670.893	100.000
SARIMA II	76.509	362.883	272.363	3.828	9.201	2653.192	100.000

Se puede ver que el error promedio de pronóstico (ME) de los modelos Base, AR(18) y ARMA(18,6) muestra que se sobreestima la serie en 645.35, 742.40 y 696.66 llamadas que ingresaron a soporte técnico por temas de agendamiento o por temas diferentes. Por su parte, los modelos SARIMA I y SARIMA II subestimaron la serie en 49.75 y 76.51 llamadas menos de las llamadas que ingresan a soporte técnico. Al mismo tiempo, estos dos modelos son los que menos se alejan al número real de llamadas, puesto que ambos son los que tienen menor RMSE (raíz del error cuadrático medio). Cabe resaltar que los modelos Base, AR(18) y ARMA(18,6) son los que tienen mayor RMSE, es decir, que están más alejados al número real de llamadas en el intervalo de predicción escogido.

Ahora bien, en otro análisis de los modelos SARIMA I y SARIMA II se puede evidenciar que son los que tienen menor error promedio absoluto del pronóstico (MAE), pues cada modelo tiene un error de 281.3 y 272.4 llamadas para las llamadas que ingresan a soporte técnico por temas de agendamiento o por temas diferentes. Mientras que los modelos Base, AR(18) y ARMA(18,6) son los que tienen mayor error promedio. Esto es posible afirmarlo porque cada uno tiene un error de 645.3, 742.3 y 696.6 llamadas. Estos modelos a la vez sobreestimaron la serie original en -18.5 %, -21.6 % y -20.5 % de las llamadas que ingresan a soporte técnico por temas de agendamiento o por temas diferentes. Los modelos SARIMA I y SARIMA II subestimaron la serie original en 3.0 % y 3.8 %.

Además, se tiene que los modelos Base, AR(18) y ARMA(18,6) son los que tienen el porcentaje medio absoluto de error (MAPE) más alto de los cinco modelos, mientras que los modelos SARIMA I y SARIMA II con 9.4 % y 9.2 % son los que tienen menor error absoluto del pronóstico.

Teniendo en cuenta lo dicho anteriormente y lo que se aprecia en la gráfica, se puede concluir que los mejores modelos son los SARIMA, ya que estos poseen una cobertura del 100 %. Esto significa que cualquiera de los dos modelos es bueno para predecir el número de llamadas que ingresan a soporte técnico de Tigo Home por temas de instalación o reparación en sus productos. Esto no quiere decir que los modelos restantes no sean buenos a la hora de predecir, sin embargo, estos no son los más parsimoniosos y suelen sobrestimar las llamadas que ingresan a soporte técnico.

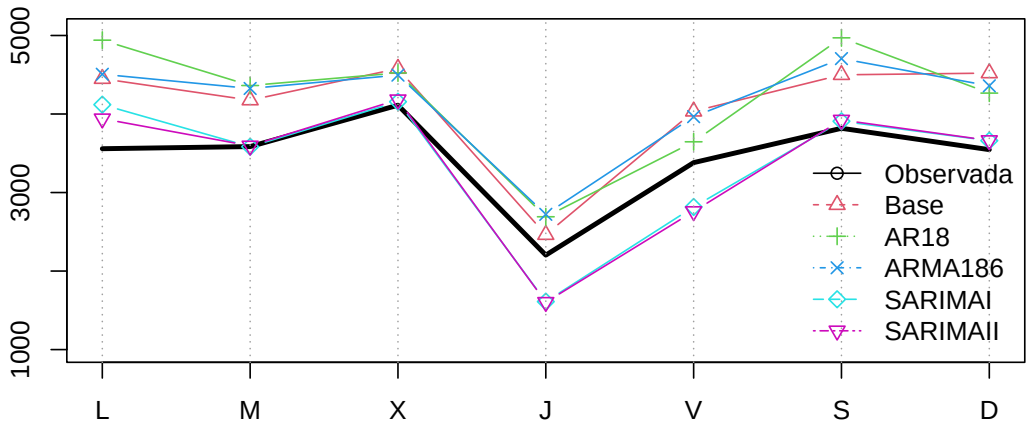

Figura 21: Gráfica de predicción.

En esta gráfica se observa que en el primer día de predicción (lunes) todos los modelos planteados sobrestiman la cantidad de llamadas que entran a soporte técnico, siendo todos muy similares. Ahora bien, los días martes y miércoles, algunos modelos siguen sobrestimando la serie, sin embargo, los modelos SARIMA I y SARIMA II aciertan con exactitud el valor real. El resto de días de la semana, los modelos base, AR18 y ARMA(18,6) sobrestiman la serie y parecen alejarse un poco. Mientras tanto, los modelos SARIMA I y SARIMA II subestiman los días jueves y viernes, pero los días del fin de semana, sábado y domingo, la serie sobrestima por muy poco el valor real de la serie. De este análisis gráfico, se puede pensar que los modelos SARIMA I y SARIMA II son relativamente buenos para la predicción del número de llamadas que ingresan a soporte técnico por temas de agendamiento o por diferentes causas.

6. Conclusiones y recomendaciones

- De las 6 zonas en que Tigo divide el territorio nacional, se puede decir que la zona Noroccidental presenta un mayor porcentaje total de llamadas que ingresaron a soporte técnico por aseguramiento o aprovisionamiento y llamadas de agenda, representando un 50 % de las llamadas.
- Se obtuvieron cuatro modelos buenos para mirar qué variables influían en que un usuario se volviera a contactar con soporte técnico por temas de aseguramiento o agendamiento, donde los modelos de árboles de clasificación son los que mejor predicción tienen.
- Los modelos SARIMA son los mejores para la predicción del número de llamadas reincidentes que entraron a soporte técnico de Tigo Home, cuyo objetivo de la llamada pudo haber sido por cuestiones de agendamiento o por temas diferentes.

Referencias

Adhikari, R. and R. Agrawal (2013). *An Introductory Study on Time Series Modeling and Forecasting*. Lap Lambert Academic Publishing GmbH KG.

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (2017). *Classification and Regression Trees*. CRC Press.

Giraldo Gómez, N. D. (2006). *Técnicas de pronósticos : aplicaciones con R*.

Jonathan D. Cryer, K.-S. C. (2009, October). *Time Series Analysis with Applications in R*. Springer-Verlag.

P., S. L. and R. R. (2007). Comparación de modelos matemáticos: una aplicación en la evaluación de alimentos para animales. *Revista Colombiana de Ciencias Pecuarias 20, 141–148.*

Paul S. P. Cowpertwait, A. M. (2009, June). Introductory Time Series with R. *Springer-Verlag GmbH.*

Rodríguez, C. M. C. (2020). Modelo para reducción de contactos para la gerencia de calidad y metodología ágiles en tigo.

Rokach, L. and O. Maimon (2014). Data Mining with Decision Trees: Theory and Applications. *Series in machine perception and artificial intelligence. World Scientific.*

Sepúlveda, J. F. D. (2012). Comparación entre árboles de regresión CART *y regresión lineal. Master's thesis, Universidad Nacional de Colombia.*

Serna Pineda, S. C. (2009). Comparación de árboles de regresión y clasificación y regresión logístcia. Master's thesis, Universidad Nacional de Colombia.

Tigo. ¿quiénes somos?

Vega, J. B. M. (2018, abril). árboles de decisión con r- clasificación.