PROPUESTA DE UN MODELO LOGÍSTICO PARA LA PROBABILIDAD DE INSTALACIÓN DE DATÁFONOS EN UNA EMPRESA UBICADA EN BOGOTÁ

Elkin Javier Cruz Hurtado

Manuel Francisco Romero

NORIA
INVESTIGACIÓN EDUCATIVA

NORIA
INVESTIGACIÓN EDUCATIVA
E-ISSN 2590-5791

PROPUESTA DE UN MODELO LOGÍSTICO PARA LA PROBABILIDAD DE INSTALACIÓN DE DATÁFONOS EN UNA EMPRESA UBICADA EN BOGOTÁ

Elkin Javier Cruz Hurtado
ejcruzh@libertadores.edu.co
Manuel Francisco Romero
mfromeroo@libertadores.edu.co
Fundación Universitaria Los Libertadores

RESUMEN

El objetivo de este articulo es dar a conocer una propuesta de un modelo de regresión logística múltiple comparado con un modelo de árbol de decisión que permita predecir la probabilidad de instalaciones exitosas en la empresa, con el fin de determinar el mejor modelo y así poder su aumentar su participación en el mercado realizando mejoras al interior de las áreas involucradas. Para el desarrollo del proyecto se tuvieron en cuenta nueve variables categóricas con 5534 individuos, con la información obtenida se identificaron las variables mas significativas para el modelo de regresión logística. En el desarrollo de este artículo se encuentra los datos analizados, las características de las variables, estimación de predicciones, matriz de confusión, diseño del área bajo la curva (curva ROC). En los resultados obtenidos la probabilidad de instalaciones de datáfonos exitosa es del 69% con las variables significativas arrojadas por el modelo de regresión logística, el análisis efectuado en el área bajo la curva de los dos modelos expuestos en el proyecto se determino que el mejor modelo para predecir los aciertos de las instalaciones tanto exitosas como fallidas frente a los datos observados es el modelo de la regresión logística múltiple con una efectividad del 76,7%.

Palabras claves*:* Datafono, Regresión Logística, Especificidad, Sensibilidad, Curva ROC.

NORIA
INVESTIGACIÓN EDUCATIVA
E-ISSN 2590-5791

ABSTRACT

The objective of this article is to present a proposal for a multiple logistic regression model compared to a decision tree model that allows predicting the probability of successful installations in the company, in order to determine the best model and thus be able to increase their participation in the market for improvements within the areas involved. for the development of the project, nine categorical variables with 5534 individuals were taken into account, with the information obtained, the most specific variables for the logistic regression model were identified. in the development of this article are the analyzed data, the characteristics of the variables, the predictions, the confusion matrix, the design of the area under the curve (roc curve). in the obtained results, the probability of successful data installations is 69% with the affected variables thrown by the logistic regression model, the analysis carried out in the area under the curve of the two affected models in the project will determine the best model to predict the successes of both successful and unsuccessful facilities against the observed data is the multiple logistic regression model with an efficiency of 76.03%.

Keywords: Point of Sale, Logistic regression, Specificity, Sensitivity, ROC curve,

INTRODUCCIÓN

En la actualidad el Datafono o P.O.S (Point of Sale) es un elemento fundamental como dispositivo que permite realizar transacciones de venta con tarjetas Visa, MasterCard, Diners, American Express, y tarjetas privadas tanto nacionales como internacionales. Este dispositivo genera beneficios a todas las partes interesadas en la transaccional de los mercados. Para el periodo del año 2019 en Colombia existía 46,12 millones de pasticos vigente de tarjetas débito y crédito (Monterrosa, 2019), lo cual aumenta la probabilidad de uso de los datafonos como un mecanismo de pago físico entre un establecimiento y un tarjetahabiente.

Los datafonos a través del tiempo han ido evolucionando debido a los cambios en las redes de comunicaciones actuales, para lo cual encontramos datafonos conectados por LAN, Wifi, GPRS, MPOS que han ayudado en la transformación de ventas de productos o servicios debido a que su respuesta para la aprobación de una transacción no supera los cinco segundos, adicional es de fácil manipulación, livianos y también es utilizado para realizar transacciones a domicilio ayudando a su vez a que el tarjetahabiente pueda adquirir sus necesidades sin tener que salir de casa. Uno de los grandes beneficios que tiene el comercio al adquirir un datafono es el aumento de ingresos monetarios debido a la facilidad del tarjetahabiente en obtener una tarjeta

débito o crédito por parte de las entidades financieras, la seguridad para el tarjetahabiente en cuanto a que no es necesario el manejo de dinero en efectivo, ya que para las ciudades de Bogotá el índice de denuncias por hurto han sido de 75.483, para Cali es de 14.546 y para Medellín con 17.527 en el periodo del año 2019 siendo este el segundo factor de robos después de los celulares (Nacional, 2019).

En el desarrollo de este trabajo se busca medir la efectividad de instalación del datafono para la empresa que proporcione una alternativa diferente de visualizar los resultados por medio de modelos estadísticos y sus predicciones, con el propósito de hacer planes de mejoras y capacitaciones al interior del equipo técnico y áreas involucradas, ayudando así en la posibilidad de aumentar su participación en el mercado actual.

REFERENTES TEÓRICOS

Navarro en 2007 define al datafono como" la transferencia electrónica de fondos mediante pago con tarjeta de banda magnética, desde el terminal en el punto de venta hasta un centro de validación de una entidad financiera" (p.27).

Por otra parte, el datáfono es una terminal dotada de un software especial que permite al tarjetahabiente el uso de un plástico (tarjeta débito o crédito), para acceder, mediante transacciones, a los recursos depositados en las cuentas de ahorro o corriente abiertas en los establecimientos de crédito o al cupo de crédito previamente asignado por éste, con el objeto de pagar bienes o servicios en establecimientos afiliados. (Colombia, DATÁFONOS, SERVICIOS Y COMISIONES, 2008)

Dispositivo empleado por los establecimientos comerciales a través de los cuales se efectúan pagos y se realizan otras operaciones. La abreviatura POS se deriva de su nombre en inglés "Point of Sale". (Colombia, INFORME DE TRANSACCIONES Y OPERACIONES, 2011)

El Terminal Punto de Venta (TPV) es un dispositivo usado en establecimientos comerciales para realizar gestiones de venta. Permite, entre otras cosas, realizar cobros con tarjeta de crédito o débito e imprimir tickets, gracias a los datáfonos, y controlar el inventario. (BBVA, 2017).

Tipo de datáfonos: En la actualidad existe variedad de datafonos que comprende con una configuración especial en su tecnología donde se pueden utilizar como integración a caja, domicilios entre otros.

LAN: el establecimiento de comercio debe contar con conexión a una red especial.

Gprs: datáfonos móviles con una tarjeta SIM incorporada.

- Cuentan con una batería que permite su uso sin línea telefónica ni cables de potencia.
- Ideales para comercios que deseen acercar el datáfono a su cliente

como: restaurantes, ferias, eventos y domicilios, y

- Realizan diversos tipos de transacciones como: compras, recargas, pago de servicios públicos, consulta de saldos (entre otras), según sean las necesidades de su negocio y configuración.

Mpos: datáfonos móviles conectados a un smartphone o tableta que permiten realizar compras a través de la aplicación MiPago (CREDIBANCO, 2020).

Regresión Logística: Kleinbaum.D y Klein. M (2002) afirman "la regresión Logística es un enfoque de modelado matemático que se puede usar para describir la relación de la variable X a una variable dependiente dicotómica, como D".

De igual manera, Fuentes. S (2011) la define como "sea Y una variable dependiente binaria (con dos posibles valores: 0 y 1). Sean un conjunto de k variables independientes, $(X_1, X_2, \cdots, X_k)$, observadas con el fin de predecir/explicar el valor de Y".

La regresión logística se usa principalmente para modelar una variable binaria (0,1) basada en una o más variables diferentes, llamadas predictores. La variable binaria que se está modelando generalmente se conoce como la variable de respuesta o la variable dependiente. (Hilbe J, 2016, p.6).

Formula de la regresión logística.

$$P(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \Sigma \beta_i X_i)}}$$

Sensibilidad: La sensibilidad puede definirse como la capacidad de la prueba para clasificar correctamente al enfermo como enfermo, o como la probabilidad de tener un resultado positivo si se tiene la enfermedad. (Medina, 2011)

La sensibilidad (Se) es la probabilidad de que la prueba dé positiva si la condición de estudio está presente (paciente enfermo o con patrón de referencia positivo). También se puede definir como la proporción de verdaderos positivos respecto al total de enfermos. (Ochoa, C y Orejas, G.,1999)

En cuanto al contexto del proyecto es necesario encontrar la sensibilidad ya que nos permite identificar la probabilidad de aciertos de las instalaciones de datafonos exitosas obtenidas de la modelación de la regresión logística múltiple frente a los datos observados (predicciones).

Especificidad: La especificidad es la capacidad de la prueba para clasificar adecuadamente a los sanos como sanos; es el porcentaje de personas que no tienen la condición de estudio y dan resultados "negativos" o "normales". (Medina, 2011).

La especificidad (Es) es la probabilidad de que la prueba dé negativa si la enfermedad está ausente (paciente sano o con patrón de

referencia negativo). También se puede definir como la proporción de verdaderos negativos respecto al total de sujetos sanos. (Ochoa, C y Orejas, G, 1999).

En el marco del proyecto se necesita identificar la especificidad debido a que nos permite precisar la probabilidad de aciertos de instalaciones que fueron fallidas comparando los resultados entre el modelo de regresión logística versus los datos obtenidos con las predicciones. En la figura 1 se observa la tabla de contingencia 2 x 2 donde se determina los verdaderos positivos (a), falsos positivos (b), falsos negativos (c), verdaderos negativos (d), de esta forma se puede analizar qué tan efectivo es el modelo tanto los datos de la base con las predicciones.

Figura 1.Tabla de contingencia 2 x 2 para la evaluación de una prueba diagnóstica

		Patrón de referencia +	Patrón de referencia –	
Prueba diagnóstica	+	Verdaderos positivos (a)	Falsos positivos (b)	a+b
	–	Falsos negativos (c)	Verdaderos negativos (d)	c+d
		a+c	b+d	Total=a+b+c

Claves:
a Verdaderos positivos (VP): enfermos con la prueba positiva
b Falsos positivos (FP): no enfermos con la prueba positiva
c Falsos negativos (FN): enfermos con la prueba negativa
d Verdaderos negativos (VN): no enfermos con la prueba negat
a+c Casos con patrón de referencia positivo (enfermos)
b+d Casos con patrón de referencia negativo (no enfermos)
a+b Casos con la prueba diagnóstica positiva
c+d Casos con la prueba diagnóstica negativa

Fuente: Epidemiología y metodología científica aplicada a la pediatría (IV): Pruebas diagnósticas (1999)

- Árbol de decisión: Los árboles de decisión es una de las técnicas de aprendizaje inductivo supervisado no paramétrico, se utiliza para la predicción y se emplea en el campo de inteligencia artificial, donde a partir de una base de datos se construyen diagramas de construcción lógica, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar serie de condiciones que ocurren en forma repetitiva para la solución de un problema. Solarte G y Soto J. (2011).

- Matriz de Confusión: es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases. (Wikipedia, 2020)

Curva ROC: Hilbe. J (2016) afirma “La curva ROC se entiende como la relación óptima de la sensibilidad del modelo en uno menos la especificidad”.

La curva ROC o curva de "característica de funcionamiento del receptor" es un

método de evaluación que podemos utilizar para evaluar la eficacia de un algoritmo de clasificación binaria ("Características operativas del receptor", n.d.) así como elija el umbral óptimo en función de nuestra tolerancia a los falsos negativos y el deseo de verdaderos positivos. Logistic Regression and ROC Curve Primer (2020).

El área bajo la curva tiene como objetivo indicar el porcentaje de precisión de los aciertos tanto de instalaciones exitosas como fallidas. Hilbe. J (2016) afirma "representa un ROC estadística de 0.5. Los valores de 0.5 a 0.65 tienen poco poder predictivo. Valores de 0,65 a 0,80 tienen un valor predictivo moderado. Muchos modelos logísticos encajan en este rango. Los valores mayores que 0.8 y menores que 0.9 generalmente se consideran teniendo un fuerte poder predictivo. Los valores de 0.9 y mayores indican el más alto cantidad de poder predictivo, pero los modelos rara vez alcanzan valores en este rango".

METODOLOGÍA

Se utilizo el programa R libre en su versión 3.6.3 para examinar los datos con sus paquetes dplyr, pROC, randomForest, ROCR, vcd, rpart, rattle, rpart. plot, RColorBrewer. El lenguaje de programación fue diseñado por Robert Gentleman y Ross Ihaka, miembros del departamento de estadística de la universidad de Auckland, en Nueva Zelanda (R (lenguaje de programación), 2020).

El análisis de los datos se aplicó bajo un modelo de regresión logística múltiple resultante a partir de una base de datos suministrado por la empresa, en la cual cuenta con nueve variables categóricas y 5534 individuos. Las variables seleccionadas fueron previamente evaluadas por el personal encargado de la empresa con el fin de incluir información significativa que no genere alteraciones al modelo seleccionado. Para realizar el análisis de los datos se cargó la base en el lenguaje de Rstudio para su inspección e identificación de valores nulo que puedan interferir en los resultados de la modelación, además con la variable dependiente categoría (GESTION_CAMPO) se transformó en factor ya que debe tener contemplar valores binarios en el cual la representación del cero corresponde a las instalaciones exitosas y la representación de uno corresponde a las instalaciones fallidas, además permite establecer las variables significativas para el modelo y sus predicciones. Se utiliza la curva roc permitiendo seleccionar modelos óptimos y la representación gráfica de la sensibilidad frente a la especificidad. En la tabla 1 se observa las variables cualitativas de la base original de las cuales se podrán interpretar y seleccionar las que mejor se caracterizan para modelar y establecer el objetivo de medir la probabilidad de efectividad de instalación de datafonos.

Tabla 1: Variable cualitativa del modelo.

Nombre Variable	Descripción	Categoria	R	Cla
TIPO_COMERCIO	Clasificación de comercios			Inc
TIPO_TECNOLOGIA	Clasificación de dátafonos			Inc
UNIDAD_NEGOCIO	Clasificacion de areas integradas en la empresa			Inc
SECCIONAL_DMA	Nombrfe de la ciudad			Inc
REGIONAL_DMA	Nombre de la regional			Inc
GESTION_DE_CAMPO	Marcar si la instalación fue exitosa o fallida	Exitoso	0	De
		Fallida	1	
GEOGRAFIA	Ubicación de la instalación			Inc
NOMBRE_BANCO	Código de la entidad financiera			Inc
TIPOLOGIA	Segmento según ubicación			Inc

Fuente: Elaboración propia

RESULTADOS

Para la selección de las ciudades se toma como referencia las tres ciudades con mayor participación que son Bogotá, Medellín, Cali, Se realizó un filtro para obtener las ciudades mencionadas sin embargo entre las tres ciudades no se presenta un cambio significativo si se pone alguna como predictora, posterior a ello se realiza la simulación del modelo de regresión logística para determinar las variables significativas para el modelo, se ejecuta bajo el comando glm de la familia binomial y el modelo logit, con el comando summary se determina los coeficientes del modelo con su p- value.

En la Tabla 2 se presenta las variables significativas modelo de regresión logística múltiple, en donde como se aprecia unidad de negocio, seccional, tipo de comercio, nombre del banco son relevantes, el modelo de segmentación de la empresa es a partir de cantidad de transacciones autorizadas y abonadas se clasifica los establecimientos pymes, medianos y grandes. Con respecto a las ciudades Bogotá, Cali, Medellín son representativas debido a que son las ciudades con mayor solicitud de instalación a nivel nacional, también porque tiene gran volumen de habitantes y establecimientos creados. El tipo de tecnología es fundamental para el proceso ya que nos indica el tipo de datafono que se instala y cuál puede ser representativo en el mercado. El banco es esencial para la determinación del proceso de instalación debido a que realizan el proceso de afiliación con los comercios y asignan la red que va a ser integrada para destinar el datáfono requerido.

En la selección de las variables del modelo de regresión logística múltiple, se descartaron varios coeficientes ya que su p-value es menor a 0.05 indicando que no son significativos para el proyecto entre ellos se encuentra: el tipo de comercio mediano puesto que son limitados y están con poca demanda de datáfonos en sus establecimientos, el tipo de tecnología como Mpos de marcas ingenico y Wifi por causas de poca petición de los clientes. Por último, la validación de la variable con coeficientes de banco 02, banco 09, banco 12, banco 14, banco 64 al ser banco afiliador de los clientes no registra un alto número de asignación por lo cual dejan de ser relevante para desarrollar el propósito del proyecto.

Tabla 2. Modelo de regresión logística en R.

Variable	Estimación	P-value	
(Intercept)	-3.80598	< 2e-16	***
datos_mod$UNIDAD_NEGOCIOECOSISTEMAS DE PAGO	0.99297	< 2e-16	***
datos_mod$UNIDAD_NEGOCIOEMPRESAS Y PYMES	0.41261	6.70e-08	***
datos_mod$UNIDAD_NEGOCIOENTIDADES FINANCIERAS	1.70034	0.00186	**
datos_mod$UNIDAD_NEGOCIOMASIVO	1.13033	< 2e-16	***
datos_mod$UNIDAD_NEGOCIONA	1.39389	< 2e-16	***
datos_mod$UNIDAD_NEGOCIOREDES INTELIGENTES	0.68805	< 2e-16	***
datos_mod$UNIDAD_NEGOCIOSIN UEN ESTABLECIDA	2.59705	7.16e-05	***
datos_mod$SECCIONAL_DMACALI	-0.25674	1.15e-12	***
datos_mod$SECCIONAL_DMAMEDELLIN	-0.35960	< 2e-16	***
datos_mod$TIPO_COMERCIOComercio_Grande	0.69676	4.57e-09	***
datos_mod$TIPO_COMERCIOEMERGENTE	0.76355	< 2e-16	***
datos_mod$TIPO_COMERCIOSin_denominación	1.28122	7.40e-13	***
datos_mod$TIPO_TECNOLOGIADIAL-LAN	1.66691	1.77e-05	***
datos_mod$TIPO_TECNOLOGIAPar Aislado	3.13052	8.44e-07	***
datos_mod$TIPO_TECNOLOGIARompefilas	2.01229	2.20e-05	***
datos_mod$TIPO_TECNOLOGIATEF GPRS	1.64786	0.03900	*
datos_mod$NOMBRE_BANCOBANCO 03	0.75020	< 2e-16	***
datos_mod$NOMBRE_BANCOBANCO 13	0.20391	0.01659	*
datos_mod$NOMBRE_BANCOBANCO 19	0.23634	0.00093	***
datos_mod$NOMBRE_BANCOBANCO 23	0.49821	1.53e-08	***
datos_mod$NOMBRE_BANCOBANCO 43	1.00952	3.37e-05	***
datos_mod$NOMBRE_BANCOBANCO 51	0.92370	< 2e-16	***
datos_mod$NOMBRE_BANCOBANCO 52	0.74638	1.33e-07	***
AIC		**34469**	

Fuente: Elaboración propia

Con el fin de identificar si la población estadística es acorde a la muestra de dicha población se utilizó un método de validación del modelo de la regresión logística a partir de base de entrenamiento y una de test en el cual a la variable dependiente se destina un 70% de la información para la participación de la modelación, se efectúa de la misma manera con la que se realizó el modelo de regresión ajustado incluyendo las variables significativas, también se realiza una validación de los datos nulos para no tener alteraciones en los resultados. En la t

Tabla 3 se presenta las predicciones del testeo junto con la matriz de confusión donde se puede evidenciar la sensibilidad donde el modelo detecta el 87,04% de probabilidad que las instalaciones sean exitosas, por otra parte, la especificidad es del 46,31%, lo cual nos indica que el modelo de regresión Logística múltiple identifica mejor las instalaciones que son exitosas. En la figura 2 se presenta el valor porcentual del área bajo la curva (AUC) en el que indica que con un 76,03% el modelo es bueno debido a que tiene la capacidad de distinguir entre las instalaciones efectivas y las instalaciones fallidas entre

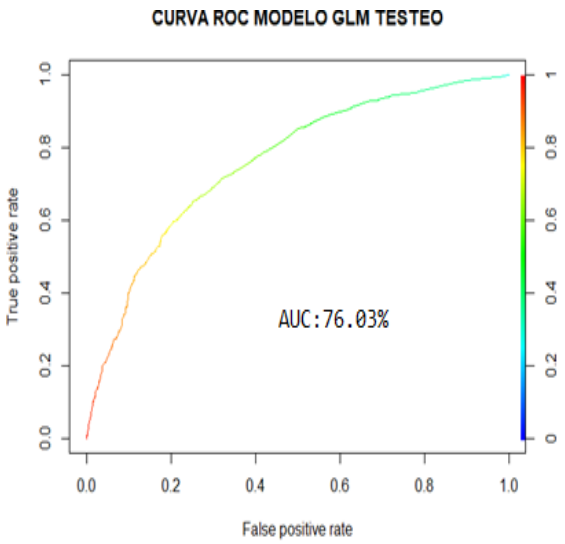
los datos de la base original con las predicciones.

Tabla 3. Matriz de confusión con su especificidad y sensibilidad del testeo en R.

		Referencias	
		Exitoso	Fallidas
Prediccion	Exitoso	5955	1646
	Fallidas	887	1420
Sensibilidad		87,04%	
Especificidad		46,31%	

Fuente: Elaboración propia

Figura 2. Curva Roc modelo glm testeo



Fuente: Elaboración propia

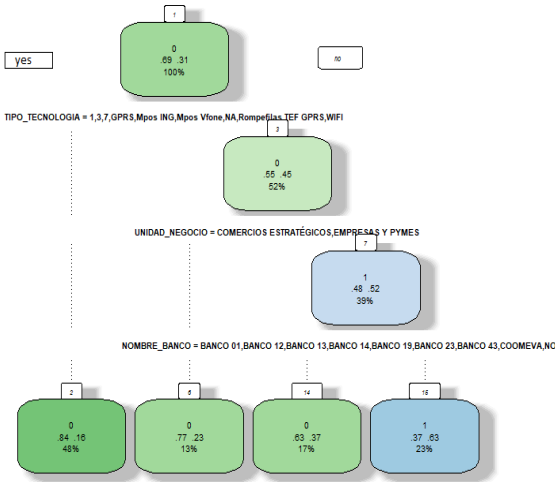
Adicionalmente, en la figura 3, se ilustra un árbol de decisión para efectuar la comparación entre modelos y analizar el mejor valor del área bajo la curva. Para comenzar a realizar el árbol de decisión, al igual que el test del modelo de regresión logística, se realiza un entrenamiento del 70% de la base con las variables

significativas del modelo, además se realiza las predicciones y se elabora la matriz de confusión con su especificidad y sensibilidad. La variables dependiente (GESTION_DE_CAMPO) al ser categóricas el árbol es de clasificación (Alvear, 2018), en donde se presenta en el primer nodo de la totalidad de los datos obtenidos la probabilidad de instalación de datafonos exitoso es del 69,05% y el 30,95 son instalaciones son fallidas, en el nodo dos se visualiza el tipo de datafono Gprs, Mpos ing, entre otros que aunque en la modelación de la regresión logística no son significativos, en el árbol de decisión nos demuestra que su instalación exitosa es de un 80% lo que realmente es bueno para la empresa. En el nodo tres se analiza los tipos de datafonos (Dial-lan) el cual su probabilidad de instalación es del 55%. Con respecto a los bancos encontrados en el nodo 15 se puede evidenciar la probabilidad de instalaciones fallidas para los bancos 02,03,51,52 y 64 con un porcentaje del 62,80%, esto es debido a que al tener mayor solicitud de instalación.

En la tabla 4 y figura 4 se presenta una presión del modelo (accuracy) del 73,16% lo que nos indica que es bueno dando la efectividad de detectar entre las instalaciones exitosas y las instalaciones fallidas, la especificidad del 46,28% y su

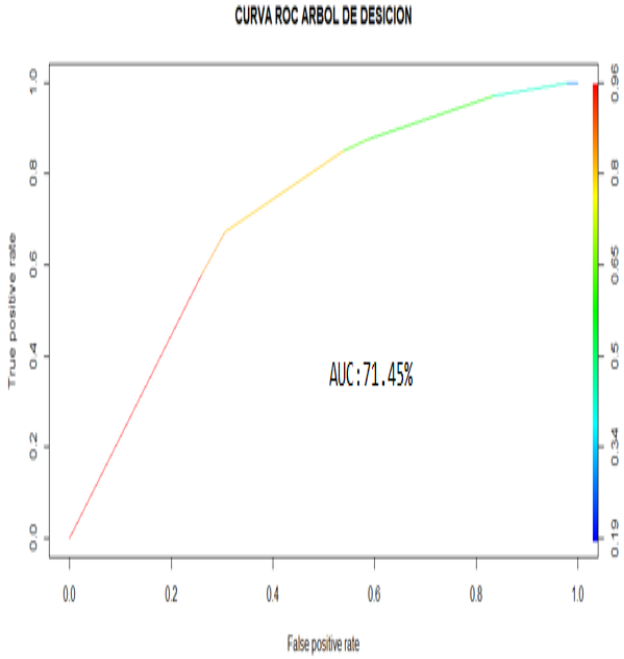
sensibilidad del 86,55%. La curva ROC para el árbol de decisión es del 71,45%.

Figura 3. Árbol de decisión



Fuente: Elaboración Propia

Figura 4. Curva Roc árbol de decisión



Fuente: Elaboración propia.

Tabla 4. Matriz de confusión con su especificidad y sensibilidad del testeo árbol de decisión.

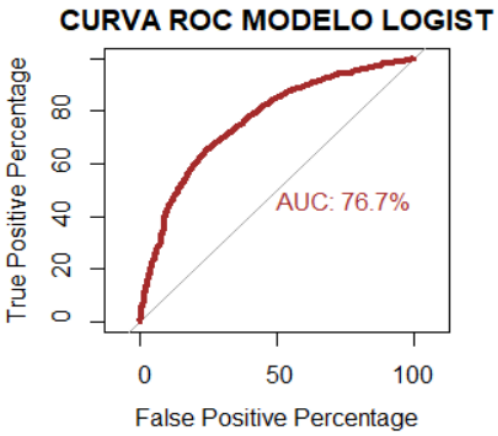
		Referencias	
		Exitoso	Fallidas
Prediccion	Exitoso	5922	1647
	Fallidas	920	1419
Sensibilidad		86,55%	
Especificidad		46,28%	

Fuente: Elaboración propia.

Finalmente, para concluir con el modelo propuesto con sus variables significativas, se realiza la curva ROC, en el cual se ilustra la relación entre la sensibilidad y la especificidad en los diferentes puntos de corte de la gráfica. En la figura 5 se presenta el área bajo la curva junto con las predicciones realizadas y de acuerdo a los resultados, su AUC es de 76,69% donde observamos que es un buen ajuste

detectando la efectividad de instalación exitosa y las instalaciones fallidas.

Figura 5. Curva Roc modelo logist ajustado



Fuente: Elaboración propia

CONCLUSIONES

En el desarrollo de este trabajo se busca medir la efectividad de instalación del datafono para la empresa que proporcione una alternativa diferente de visualizar los resultados por medio de modelos estadísticos y sus predicciones, con el propósito de hacer planes de mejoras y capacitaciones al interior del equipo técnico y áreas involucradas, ayudando así en la posibilidad de aumentar su participación en el mercado actual. Con los resultados obtenidos se observó que la probabilidad de éxito de instalación de datáfono es del 69% con las variables significativas determinadas por el modelo de regresión logística, además entre el modelo de regresión logística múltiple y la metodología de árbol de decisión se concluyo que el mejor modelo pronosticado es de la regresión logística múltiple ya que su valor bajo la curva es de 76,03% superior al del árbol de decisión, permitiendo identificar con mayor precisión los datos de la base original de

las instalaciones exitosas y las instalaciones fallidas frente a las predicciones estimadas presentando así un modelo con buen ajuste. Los resultados de los dos modelos expuestos en el proyecto (logístico y árbol de decisión) son predicciones medianas ya que se esperaba tener un 80% de efectividad de las instalaciones, sin embargo
las variables significativas para el proyecto (unidad de negocio, Banco, ciudad, tecnología) son necesarias para el proceso de instalación del datáfono, el banco al ser el aliado de sus clientes la empresa debe realizar estrategias comerciales y/o convenios para poder aumentar la probabilidad de la instalación exitosa de datáfonos y ser competitivos en el mercado participante.

REFERENCIAS BIBLIOGRÁFICAS

Alvear, J. O. (16 de 11 de 2018). *Arboles de decision y Random Forest.* Obtenido de bookdown: https://bookdown.org/content/2031/

BBVA. (2017 de 05 de 2017). *¿Qué es el TPV?* Obtenido de BBVA: https://www.bbva.com/es/que-es-el-tpv/

Blissett, R. (26 de 11 de 2017). *Logistic Regression in R.* Obtenido de RPubs by RStudio: https://rpubs.com/rslbliss/r_logistic_ws

Colombia, S. F. (2008). *Datáfonos, servicios y comisiones.*

Colombia, S. F. (2011). *Informe de transacciones y operaciones.*

Camargo, J. J., Joyanes, L. y Giraldo, L. M. (2016). La inteligencia de negocios como una herramienta en la gestión académica Revista Científica, 1(24). https://doi.org/10.14483/udistrital.jour.RC. 2016.24.a11

CREDIBANCO. (15 de 05 de 2020). *Soluciones en punto de Venta*. Obtenido de Credibanco: https://www.credibanco.com/soluciones-para-comercios/soluciones-en-punto-de-venta

Fernández, S. d. (2011). *REGRESIÓN LOGÍSTICA.* Madrid.

Hilbe, J. M. (2016). Practical Guide to Logistic Regression. En J. M. Hilbe, *Practical Guide to Logistic Regression* (pág. 6).

Kaggle. (2020). *Kaggle*. Obtenido de Logistic Regression and ROC Curve Primer: https://www.kaggle.com/captcalculator/logistic-regression-and-roc-curve-primer/code

Klein, D. G. (2002). *Logistic Regression: A Self-Learning Text, Second Edition.*

Lemeshow, D. W. (2000). *Applied Logistic Regression.*

Li, J. (07 de 11 de 2018). *Data Analysis on the Credit Card Fraud Detection dataset from Kaggle.* Obtenido de RPubs by RStudio: https://rpubs.com/Kertoky/438939

Logistic Regression. (2 de 10 de 2018). Obtenido de UC Business Analytics R: https://uc-r.github.io/logistic_regression

Lozano -Forero, S, Ballesteros-Ballesteros, V., & Nisperuza- Toledo, J. L. (2018). Gradient Statistic: An option for conducting hypothesis testing in small sample size scenarios. *International Journal of Applied Engineering Research*, *13*(23), 16368-16375.

Machine Learning con R. (28 de 02 de 2015). Obtenido de Apuntes-r.blogspot.com: http://apuntes-r.blogspot.com/2015/02/curva-roc-con-package-rocr.html

Medina, M. C. (2011). Generalidades de las pruebas diagnósticas,. *Rev. Colomb. Psiquiat*.

Méndez Hincapié, N. F. (2019). Introducción del concepto de probabilidad en Física desde la Mecánica Estadística. Revista Científica, 280-292. Recuperado a partir de https://revistas.udistrital.edu.co/index.php/revcie/article/view/14500

Monterrosa, H. (10 de 7 de 2019). *Cada día se emiten más de 12.000 tarjetas de crédito y más 15.000 de débito*. Obtenido de La republica.com: https://www.larepublica.co/finanzas/cada-dia-se-emiten-mas-de-12000-tarjetas-de-credito-y-15000-de-debito-2882840

Nacional, R. (23 de 07 de 2019). *En promedio, cerca de 1.136 personas son víctimas de hurto cada día en Colombia*. Obtenido de El Espectador: https://www.elespectador.com/noticias/nacional/en-promedio-cerca-de-1136-personas-son-victimas-de-hurto-cada-dia-en-colombia-articulo-872344

Navarro, Y. P. (2007). *Factibilidad para la creación de una empresa para pagos de servicios públicos con datáfono en barrancabermeja (tesis de pregrado).* Barrancabermeja.

Ochoa , C y Orejas, G. (1999). Epidemiología y metodología. *Epidemiología y*

metodología científica aplicada a la pediatría (IV), 303.

R (lenguaje de programación). (6 de 5 de 2020). Obtenido de Wikipedia: https://es.wikipedia.org/wiki/R_(lenguaje_de_programaci%C3%B3n)

R Notebook - Ejemplo ML para detección de fraude en TC. (s.f.). Obtenido de R Pubs: https://rstudio-pubs-static.s3.amazonaws.com/298528_845f68bfb9814f05bade678fc770bc99.html

Rickert, J. (01 de 03 de 2019). *Some R Packages for ROC Curves*. Obtenido de R Views: https://rviews.rstudio.com/2019/03/01/some-r-packages-for-roc-curves/

Rodrigo, J. A. (8 de 2016). *Regresión logística simple y múltiple* . Obtenido de RPubs by RStudio: https://rpubs.com/Joaquin_AR/229736

Solarte, G. y Soto, J. (2011). Arboles de decisiones en el diagnóstico de enfermedades cardiovasculares. *Scientia et Technica Año XVI*, 104

Wikipedia. (20 de 3 de 2020). *Matriz de confusión*. Obtenido de Wikipedia: https://es.wikipedia.org/wiki/Matriz_de_confusi%C3%B3n

Zapata Ceballos, H. A., Bustince, H., & Dimuro, G. (2020). Funciones t-migrativas t-overlap: una generalización de migratividad en funciones t-overlap. *Revista Científica*, *2*(38). https://doi.org/10.14483/23448350.15601