MODELO DE EVALUACIÓN Y VIABILIDAD PARA PRÉSTAMO DE PRODUCTOS A LOS USUARIOS DE UNA EMPRESA DE TELEFONÍA CELULAR PREPAGO EN COLOMBIA.

Edison Jair Gonzáles González, ejgonzalesg@libertadores.edu
Diego Felipe Jiménez Delgado, dfjimenezd@libertadores.edu
Jorge Armando Sánchez Jiménez, jasanchezj01@libertadores.edu

RESUMEN

La presente investigación plantea la comparación de dos modelos para evaluar el riesgo de crédito para el préstamo de productos prepago de una empresa de telecomunicaciones en Colombia utilizando técnicas de machine learning. Los modelos a presentar son árbol de decisión y regresión logística que posteriormente serán comparados por medio de la curva ROC y el método de validación cruzada K-Fold. La metodología que se usará es CRISP-DM la cual consta de 6 etapas que serán explicadas una a una en el documento y en el contexto de investigación. El mejor modelo para demostrar la viabilidad de préstamo de producto es el árbol de decisión, ya que comparándolo con el modelo de regresión logística y sus respectivas validaciones cruzadas el indicador de exactitud fue del 86%.

.

Palabras clave: Machine learning, Curva ROC, K-Fold.

ABSTRACT

This research proposes the comparison of two models to evaluate the credit risk for the loan of prepaid products of a telecommunications company in Colombia using machine learning techniques. The models to be presented are decision tree and logistic regression, which will be compared by means of the ROC curve and the K-Fold cross-validation method. The methodology to be used is CRISP-DM which consists of 6 stages that will be explained one by one in the document and in the research context. The best model to demonstrate the feasibility of product lending is the decision tree, since compared to the logistic regression model and its respective cross-validations, the accuracy indicator was 86%.

Keywords: Machine learning, ROC curve, K-Fold.

INTRODUCCIÓN

En Colombia existen una variedad de planes de telefonía celular brindados por los diferentes operadores presentes, para ello, suplen de minutos, datos y sms (plan pospago) por un valor dado el cual el usuario paga en una factura de forma mensual. Si desea cambiarlo, simplemente habla con el operador y automáticamente tiene otro plan.

La otra modalidad que manejan los operadores es el plan prepago, el cual los diferentes usuarios hacen una recarga de un valor X que contiene ciertas características un poco similares al plan pospago, donde los recursos de dicho plan se acaban, y es necesario hacer otra recarga.

Se empleará una metodología llamada CRISP-DM la cual consta de 6 fases, tales como:

Fase 1: Entendimiento del negocio

Fase 2: Entendimiento de los datos.

Fase 3: Preparación de los datos.

Fase 4: Modelado.

Fase 5: Evaluación.

Fase 6: Implementación.

En cada fase se describen el paso a paso de la aplicación del proyecto.

Por esto, brindaremos una solución para aquellos usuarios que aplican al uso de planes prepago y que, por su constante recarga de productos, puedan acceder a un préstamo de estos y que podría ser descontado de las futuras recargas. Usaremos técnicas de modelamiento con técnicas de predicción con machine learning y saber que usuario podría tener acceso a los préstamos de productos de la empresa, dependiendo del comportamiento de recargas en el tiempo reciente y otras diversas variables que se trabajarán.

En el capítulo de resultados se visualizarán la matriz de confusión, indicadores de rendimientos, curva ROC, validación cruzada correspondiente a los modelos árbol de decisión y regresión logística.

REFERENTES TEÓRICOS

A continuación, se presentan casos de estudio acordes a la identificación de riesgo de crédito y técnicas de machine learning específicamente con modelos logístico y árboles de decisión. Castaño y Pérez (2005), establecen el modelo logístico como la herramienta para poder evaluar el riego de crédito. Algunas decisiones se toman de manera automática, otras de manera lógica o de manera intuitiva pero que podrían llegar a ser pocos efectivas, por lo tanto, deben asumirse de una manera más profunda. Por eso, el poder disponer de un modelo es de gran utilidad ya que podrán saber si es posible otorgar o no un crédito. Concluyen que el modelo logístico es apropiado como herramienta para poder estimar si un evento ocurre, que, en el caso de estudio, el default del cliente.

Padilla, Tasgua (2016), señalan que el modelo de regresión logística ayuda al análisis de riesgo. El modelo considera otro tipo de variables que no se tienen en cuenta en análisis tradicional y ayudan a tomar mejores decisiones en la efectividad para medir el riesgo.

Narváes (2019), Implementa un modelo logístico para identificar la probabilidad de incumplimiento de crédito comercial. Aportando que teniendo en cuenta variables de tipo socioeconómicos y variables propias de la entidad permita evidenciar que la tasa de morosidad se mantenga en niveles adecuados y no perturbe la estabilidad financiera.

Támara, Aristizábal y Velásquez (2010), Evidencian la utilidad del uso de modelos econométricos para estimar el incumplimiento de deudores. Para ello ejecutan modelos como el árbol de decisión y logístico. Concluyen que, lo estimando con el modelo de árbol mantiene lo dicho en la teoría económica, donde variables como el endeudamiento, los ingresos y activos son significativos y fundamentales en la evaluación de los deudores. Por otra parte, el modelo logístico explica que las variables acordes en la teoría para medir la probabilidad de incumplimiento son fundamentales y completamente significativas para el análisis de crédito.

Cardona (2004) Utiliza modelos de árbol de decisión para calcular las probabilidades de incumplimiento de crédito. Logra identificar 6 tipos de perfiles de riesgo los cuales permitan asumir decisiones en los créditos. Cardona se basa de 3 premisas para la construcción de un modelo, la simplicidad, la potencia y estabilidad. Comentando que la simplicidad es que cualquier persona pueda comprender y entender por qué funciona y trata de predecir el modelo. La potencia como indicador para discriminar los clientes buenos de los malos y la estabilidad que hace referencia a la sostenibilidad del modelo en el tiempo para conservar la calidad en su discriminación. Concluye que el modelo de árboles de decisión es efectivo para

la predicción de incumplimiento y su uso ayuda a potencializar la predicción de estrategias en diferentes ámbitos.

Tello, Eslava y Tobías (2012) usan el modelo de árbol de decisión como técnica en la minería de datos para evaluar el riesgo del crédito financiero. Implementan dos tipos de algoritmos basados en técnicas de árboles de decisión para luego compararlos y determinar cual es mejor. Concluyen que el algoritmo ID3 tiene una mayor precisión para la clasificación de la información de los clientes.

Macias (2020) Aplica tres técnicas para elaborar un modelo de perfilamiento o puntuación de créditos. Dentro de las técnicas usa árboles de decisión, análisis discriminante y regresión logística. Las variables más influyentes en el modelo de árbol son el puntaje de crédito, estados civil y tiempo de trabajo, donde un segmento del puntaje mayor a 955 puntos obtiene un porcentaje del 6% de clientes malos. Concluye que el árbol de decisión es la técnica con mayor predictibilidad y ajuste determinando el porcentaje de clientes potenciales como malos.

Existen diferentes técnicas que nos permiten conocer qué tan viable es el préstamo de algún producto en particular que se ofrecen en las empresas de diversos sectores. Empresas financieras, automovilísticos, etc, brindan la posibilidad de adquirir una obligación a las personas que cumplan criterios impartidos y así saber si pueden o no aplicar a la necesidad del préstamo.

El riesgo de crédito incurre en que las empresas tienen la posibilidad de disminuir el valor de su activo por el incumplimiento de pago de los deudores a su obligación y que en dado caso de que ocurra, la empresa deberá contar con las medidas para poder obtener su respectivo cumplimiento de la obligación hacia el deudor.

Para la presente investigación no usaremos información de carácter financiero u otras obligaciones de los usuarios, pero si contamos con datos que permitan observar el comportamiento de cada uno de ellos en su tiempo con la entidad.

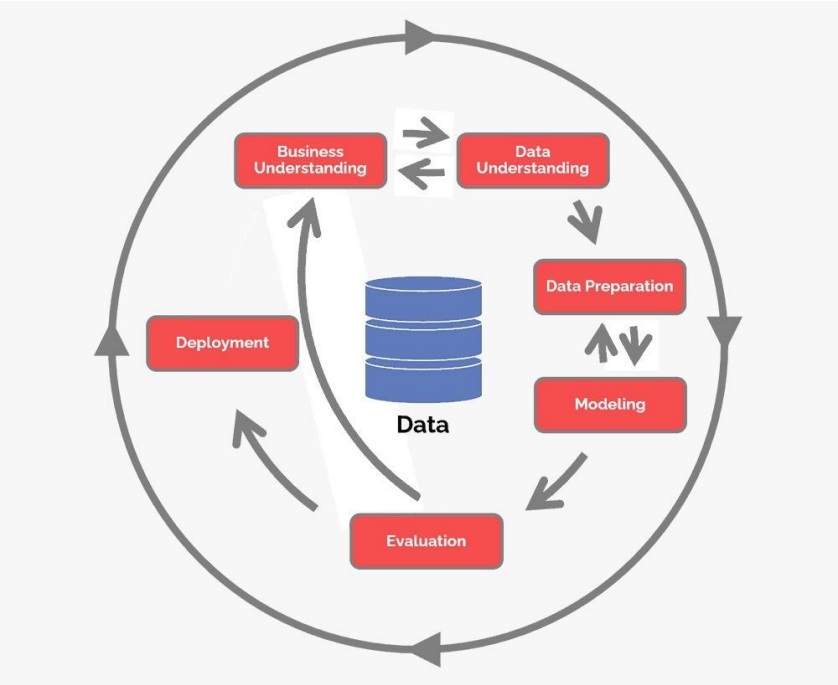
Para el escenario educativo quisimos implementar un modelo de regresión logística basado en la siguiente literatura según (ibm, s.f.), la regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de predictores. Es similar a un modelo de regresión lineal, pero está adaptado para modelos en los que la variable dependiente es dicotómica. Los coeficientes de regresión logística pueden utilizarse para estimar la razón de probabilidad de cada variable independiente del modelo. La regresión logística se puede aplicar a un rango más amplio de situaciones de investigación que el análisis discriminante.

METODOLOGÍA (CRIPS-DM)

Para la presente investigación utilizaremos la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), es un estándar que se maneja en proyectos de minería de datos y modelado, además permite su fácil uso en cualquier tipo de herramienta, nos basamos en la definición de (healthdataminer, s.f.).

Figura 1.

Flujo de cada etapa de la metodología CRISP-DM



Nota. Imagen tomada de (healthdataminer, s.f.).

Fase 1 Comprensión del negocio.

La presente investigación se basa es un operador virtual de telefonía prepago, que fue fundado hace más 9 años, llega a Colombia con un concepto disruptivo de llamadas de voz por segundo, ganando una posición del cuarto operador móvil en Colombia, debido a su concepción “ser prepago”, el comportamiento de los clientes es muy variable, en datos, voz y días de uso, por este motivo queremos encontrar patrones de uso mediante modelos de regresión logística y arboles de decisiones para contemplar la posibilidad de realizar un préstamo de producto a los clientes que cumplan con un mínimo de condiciones requeridas.

a. Definición de objetivos.

Evaluar el modelo de viabilidad de préstamo de producto que ayude a la estrategia de fidelización.

b. Definición de criterios de éxito.

El proyecto es exitoso si 30.000 clientes acceden a la estrategia y adquieren por lo menos un producto de valor promedio de $15.000, con esto la compañía mantendría un ingreso mensual aproximado de $450.000.000, ayudando a la compañía a no perder ingresos ni afectar el flujo de caja.

c. Valoración de la situación.

Actualmente la compañía no cuenta con una estrategia clara de fidelización y realiza envíos masivos de campañas en donde se satura al usuario de mensajes y se puede estar quemando el canal de envío. Es como si se quisiera matar una mosca con una bomba.

d. Inventario de recursos.

Generamos un ETL realizado en bash que se ejecuta de forma diaria en el demonio de una maquina Linux Ubuntu, allí traemos las compras de producto y recargas que han realizado los clientes, generando un histórico, la información se carga a una base de datos Postgres, generando el proceso de procesamiento, calidad y creación del dataset.

El dataset lo cargamos y procesamos con una conexión entre Python y Postgres.

e. Requisitos.

Para la construcción del dataset se requiere contar una historia de 6 meses de compra de productos y 6 meses de recargas, y realizar el procesamiento de los datos dejando un usuario único por las variables.

f. Restricciones.

Actualmente no tenemos restricciones, ya que contamos en la compañía con un ecosistema robusto que soporta el procesamiento de grandes cantidades de datos.

g. Riesgos y contingencias.

El proyecto se podría retrasar en la implementación debido a:

- Que el equipo financiero apruebe el préstamo del producto al usuario.
- Que el equipo de mercadeo nos ayude con la comunicación y la campaña
- Los planes que tenemos si estos acontecimientos ocurren son:
- Informar el valor y el caso de negocio que se tiene, demostrando que la compañía sin este modelo dejaría de percibir $450.000.000 de forma mensual.
- Ayudando a aterrizar los clientes objeto del prestamos con la realidad del cliente hoy en la compañía, esto con el fin de generar una comunicación acorde al segmento de a uno.

h. Terminología:

- ETL = Extracción, transformación y carga de información
- Bash = lenguaje nativo de linux
- Linux = Sistema operativo libre
- Python = Lenguaje de programación
- Postgres = Base de datos relacional SQL

Fase 2 Comprensión de los datos.

a. Recolección de datos iniciales

La información con la que se trabaja es tomada de las bases de datos transaccionales que tiene la empresa. No se tuvo ningún tipo de problema para acceder a ella. Se cuenta con información de clientes, cantidad de "recargas", tipo de canal, entre otras. Se tienen en cuenta 10 variables de negocio para poder usar en modelo las cuales son:

Tabla 1.

Variables utilizadas en la creación del dataset.

Cuantitativas	**tenure:** antigüedad del usuario en la compañía.	**inactivity days:** tiempo que un usuario no ha usado el servicio de telefonia.	**activity days:** tiempo de vida usanado la linea en la compañía.	**total recharge:** dinero que el cliente ha recargado en los ultimos 6 meses.	**count recharge:** numero de recargas que el cliente ha hecho en los ultimos 6 meses.	**purchase frecuency:** cada cuanto un cliente compra producto.	**age last purchase:** hace cuantos días el cliente no compra producto.	**clase:** variable dicotomica que nos dice si el cliente es apto para el prestamo de producto.
Cualitativas	**suscriberid:** identificador unicado de cada usuario.	**tipo:** indica si usuario compra digital o tradicional.						

Nota. Fuente: Elaboración propia.

Fase 3 Preparación de los datos.

a. Selección de los datos

Las siguientes variables: subscriberid, tenure, inactivity_days, activity_days, total_recharge, count_recharge, purchase_frequency, tipo, age_last_purchase, aportan con el entendimiento del comportamiento del cliente, los que ayuda de forma directa con la implementación de una estrategia de fidelización, hay un total de 359406 registros y 10 columnas en la base inicial.

b. Limpieza de los datos.

La base no tiene valores ausentes.

Se observa que no existen datos faltantes en la en la figura 2.

Figura 2.

Valores faltantes por columna

subscriberid	0
tenure	0
inactivity_days	0
activity_days	0
total_recharge	0
count_recharge	0
purchase_frequency	0
tipo	0
age_last_purchase	0
class	0

Nota. Fuente: Elaboración propia.

c. Estructuración de los datos.

Los registros de cada evento de producto y recarga llegan de forma online a la base de producción, y allí entra el ETL para traer la info y alimentar la bodega de datos

Fase 4 Modelado.

a. Para esta fase se evalúan dos técnicas de modelado, ya que son las más usadas para variables de respuesta dicotómicas y para modelo de riesgo de crédito:

- Regresión Logística, es una técnica estadística que estima la relación existente entre una variable dependiente y un conjunto de variables independientes
- Árbol de decisión, sirven para desarrollar sistemas de clasificación que predicen nuevas observaciones de acuerdo a un conjunto de reglas, por ejemplo, préstamos de alto riesgo vs. préstamos de bajo riesgo.

b. selección técnica de modelado

Se ejecutan ambas técnicas y el accuracy más alto es el de árbol de decisión con el 86%.

Fase 5 Evaluación.

La técnica de modelado seleccionada fue la de árbol de decisión, en la sección de resultados verteremos en detalle porque este modelo fue escogido.

Fase 6 Implementación del modelo.

Esta investigación es la primera fase del proyecto de viabilidad de préstamo de producto, el cual se presentará al área financiera de la compañía, con el fin de justificar la implementación y retorno de inversión del mismo, sostenido en datos estadísticos.

RESULTADOS

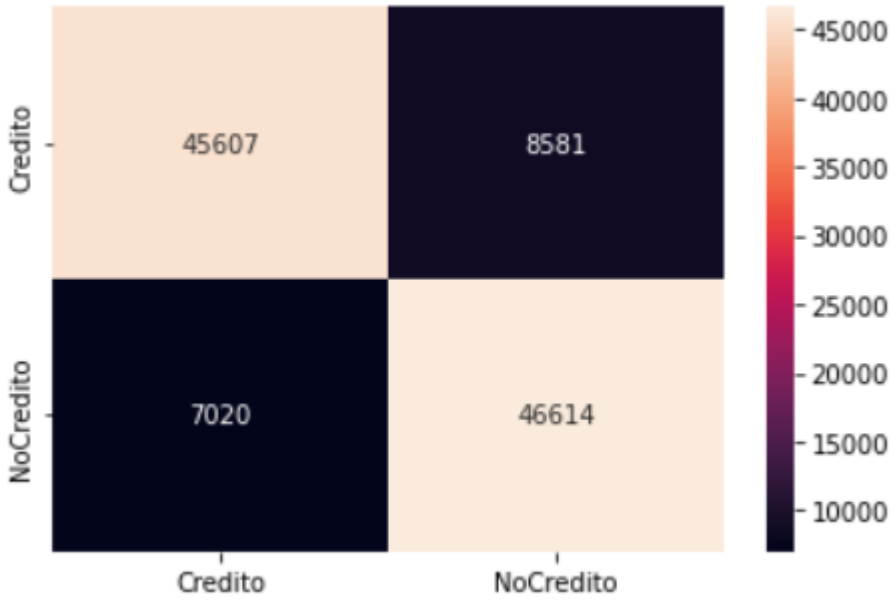
En el presente proyecto se aplicaron dos tipos de modelos supervisados, que se desarrollaron de manera satisfactoria, que para nos ayudaron entender de forma estadística la viabilidad de préstamo de producto a los clientes, a continuación, una breve explicación de los hallazgos encontrados en cada modelo.

Modelo de regresión logística

Para la aplicación del modelo de regresión logística se aplicó la proporción 70-30, donde la data de entrenamiento fue del 70% y una base de test fue del 30%, lo que arrojo como resultado la siguiente de matriz de confusión en la figura número 12.

Figura 3.

Matriz de confusión modelo logístico.



Nota. Fuente: Elaboración propia.

Para el cálculo de la predicción del modelo encontramos los siguientes indicadores:

Tabla 2.

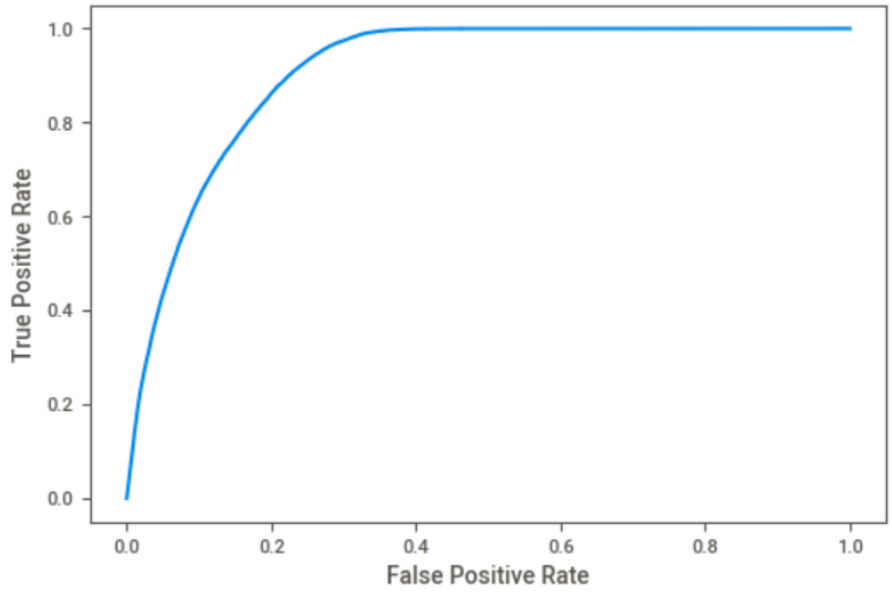
Indicadores de rendimiento del modelo logístico.

	precision	recall	f1-score	support
0	0.87	0.84	0.85	54188
1	0.84	0.87	0.86	53634
accuracy			0.86	107822
macro avg	0.86	0.86	0.86	107822
weighted avg	0.86	0.86	0.86	107822

Nota. Fuente: Elaboración propia.

Figura 4.

Curva ROC modelo regresión logística.



Nota. Fuente: Elaboración propia.

- Podemos evidenciar que los promedios del stratified K-fold, son muy parecidos a la exactitud del modelo.

Figura 5.

Validación cruzada del modelo logístico.

```
Cross Validation Scores are [0.85589363 0.85261442 0.85319077 0.856152   0.85557278]
Average Cross Validation score :0.8546847211589723
```

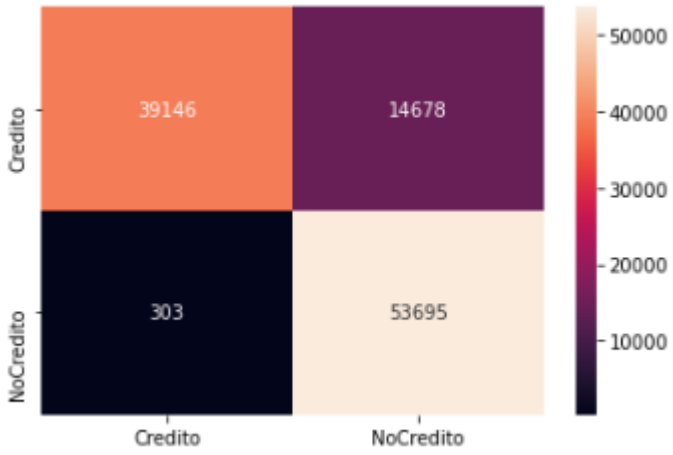
Nota. Fuente: Elaboración propia.

Modelo árbol de decisiones.

Para la aplicación del modelo de árboles de decisiones usamos la proporción 70-30, donde la data de entrenamiento fue del 70% y una base de test fue del 30%, lo que nos dio los siguientes resultados.

Figura 6.

Matriz de confusión árbol de decisiones.



Nota. Fuente: Elaboración propia.

Tabla 3.

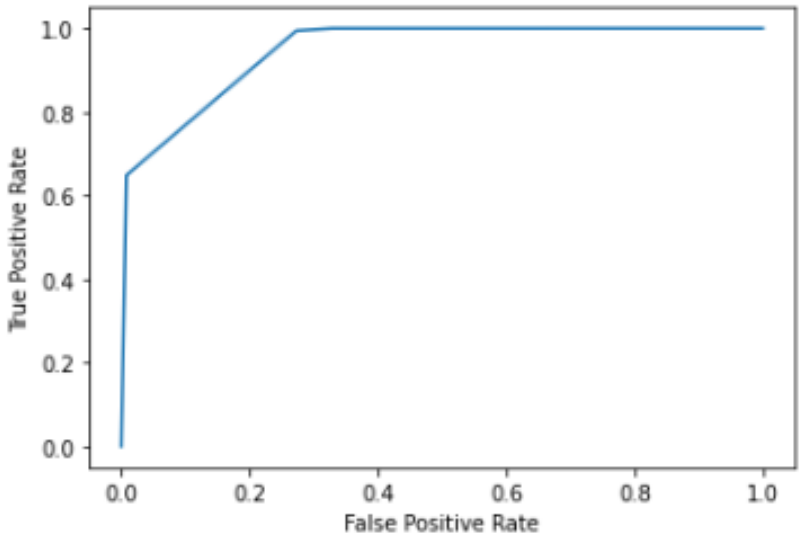
Indicadores de rendimiento del modelo árbol de decisión.

	precision	recall	f1-score	support
0	0.99	0.73	0.84	53824
1	0.79	0.99	0.88	53998
accuracy			0.86	107822
macro avg	0.89	0.86	0.86	107822
weighted avg	0.89	0.86	0.86	107822

Nota. Fuente: Elaboración propia.

Figura 7.

Curva ROC modelo árbol de decisiones.



Nota. Fuente: Elaboración propia.

- Podemos evidenciar que los promedios del stratified K-fold, son muy parecidos a la exactitud del modelo.

Figura 8.

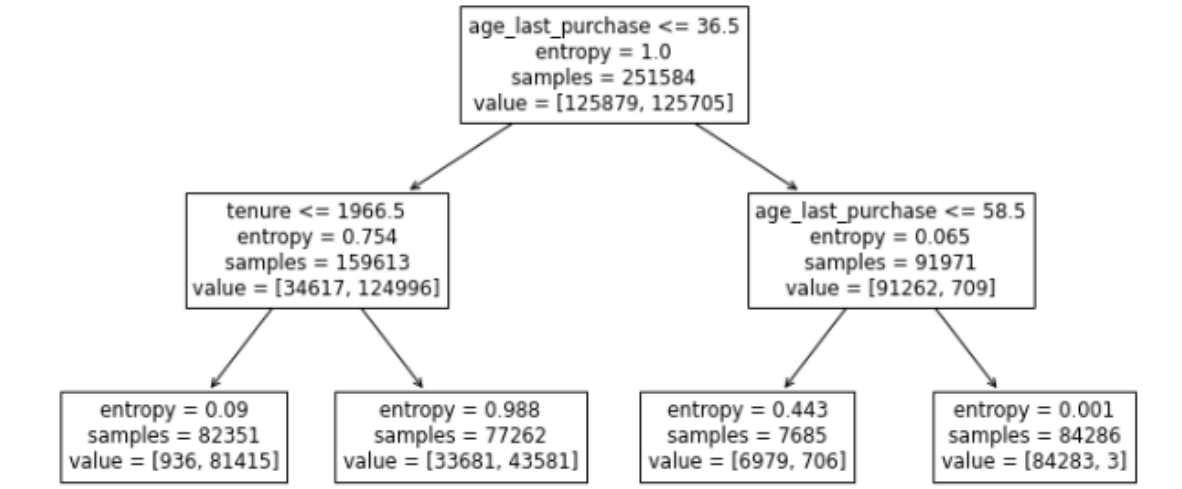
Validación cruzada del modelo árbol de decisiones.

```
Cross Validation Scores are [0.80195942 0.80084814 0.79999467 0.80163761 0.80209991]
Average Cross Validation score :0.8013079492496935
```

Nota. Fuente: Elaboración propia.

Figura 9.

Resultado árbol de decisiones.



Nota. Fuente: Elaboración propia.

DISCUSIONES

Dado el objetivo del presente trabajo, se puede concluir que la fiabilidad de los datos obtenidos con el modelo entrenado árbol de decisión es aprobado, al observar el error absoluto medio es de 0.16, nos indica que la diferencia media entre los valores reales y el valor obtenido es de 0.16. Este valor entre más cerca este de 0, el modelo tendrá una mayor capacidad para predecir bien. De acuerdo a la metodología establecida, se puede determinar que es viable realizar investigaciones en proyectos de minería de datos aplicada a diferentes sectores económicos. Aunque la metodología realizada demanda mayor tiempo al final, esta brinda confianza en el resultado obtenido.

Someter los datos a los modelos de árbol de decisión y regresión logística, con una proporción 70-30, donde la data de entrenamiento fue del 70% y una base de test fue del 30%, permite concluir que los resultados de predicción obtenidos son más robustos, dado que los falsos positivos y falsos negativos son valores iguales o menores al 15%.

CONCLUSIONES

El mejor modelo para demostrar la viabilidad de préstamo de producto es el árbol de decisión, ya que comparándolo con el modelo de regresión logística y sus respectivas validaciones cruzadas el indicador de exactitud fue del 86%.

Los dos modelos creados no presentan sobre entrenamiento, por lo tanto, cualquiera de los dos puede ser usados para predecir el comportamiento de préstamo de producto.

REFERENCIAS BIBLIOGRÁFICAS

agenciab12. (24 de 05 de 2021). agenciab12. Obtenido de agenciab12: https://agenciab12.com/noticia/que-son-arboles-de-decision-inteligencia-artificial

bctsconsulting. (11 de 05 de 2020). bctsconsulting. Obtenido de bctsconsulting: https://bctsconsulting.com/2020/05/11/los-modelos-analiticos-la-base-de-todo-sistema-de-inteligencia-artificial/

cleverdata. (s.f.). Obtenido de cleverdata: https://cleverdata.io/que-es-machine-learning-big-data/

Goicochea, A. (11 de 08 de 2009). Tecnologías de la Información y Estrategia. Obtenido de Tecnologías de la Información y Estrategia: https://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de-datos/

healthdataminer. (s.f.). healthdataminer. Obtenido de healthdataminer: https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

ibm. (s.f.). ibm. Obtenido de ibm: https://www.ibm.com/docs/es/spss-statistics/SaaS?topic=regression-logistic

Joaquín Amat Rodrigo. (06 de 2016). Obtenido de Joaquín Amat Rodrigo: https://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal

opain. (11 de 2016). Obtenido de opain: https://www.opain.co/files/2016-11-el_peligro_aviar_desde_la_perspectiva_de_una_aerolinea.pdf

Otalora Dueñas, L. M. (29 de 04 de 2013). repositoryunimilitar. Obtenido de repositoryunimilitar: https://repository.unimilitar.edu.co/handle/10654/3459

Portafolio. (27 de 12 de 2017). Portafolio. Obtenido de Portafolio: https://www.portafolio.co/negocios/empresas/empresas-en-colombia-que-usan-tecnologia-de-analisis-de-datos-512836

QuestionPro. (18 de 11 de 2021). Obtenido de QuestionPro: https://www.questionpro.com/blog/es/metodologia-de-la-investigacion/

Roman, V. (07 de 02 de 2019). Ciencia & Datos. Obtenido de Ciencia & Datos: https://medium.com/datos-y-ciencia/machine-learning-supervisado-fundamentos-de-la-regresi%C3%B3n-lineal-bbcb07fe7fd

Tactic. (11 de 18 de 2021). Tactic. Obtenido de Tactic: https://tatic.net/es/blog/inteligencia-artificial-en-el-sector-de-las-telecomunicaciones/

unioviedo. (s.f.). unioviedo. Obtenido de unioviedo: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html

ustadistancia. (11 de 2021). Obtenido de ustadistancia: http://soda.ustadistancia.edu.co/enlinea/sandracarvajal_METODOLOGIADELAINVESTIGACION2/poblacin_y_muestra.html

Fernández Castro, H., Pérez Ramírez, F. (2005). El modelo logístico: Una herramienta estadística para evaluar el riesgo de crédito. Revista Ingenierías Universidad de Medellín. Vol. 4 (Número. 6). pp. 55-75.

Narváez García, A. (2019). Variables Determinantes De La Probabilidad De Incumplimiento De Los Créditos Comerciales En Una Institución Financiera Del Ecuador, Aproximación Bajo El Modelo De Regresión Logística Binaria. Universidad Técnica de Ambato

Támara Ayús, A., Aristizábal Velásquez, R., Velásquez Ceballos, H. (2010). Estimación De Las Provisiones Esperadas En Una Institución Financiera Utilizando Modelos Logit Y Probit. Revista Ciencias Estratégicas. Vol. 18 (Número. 24). pp. 259-270.

Cardona Hernández, P. (2004). Aplicación de árboles de decisión en modelos de riesgo crediticio. Revista Colombiana de Estadística Vol. 27 (Número. 2). pp. 139-151.

Padilla Guamán, D., Tagua Guambo, N. (2016). Análisis E Impacto En El Otorgamiento De Créditos Mediante Un Modelo Logístico En La Cooperativa De Ahorro Y Crédito “Mushuc Runa” De La Ciudad De Riobamba.