UNIVERSIDAD JORGE TADEO LOZANO BOGOTÁ

FACULTAD DE CIENCIAS NATURALES E INGENIERÍA
MAESTRÍA EN INGENIERÍA Y ANALÍTICA DE DATOS

IDENTIFICACIÓN PREDICTIVA PARA INSTALACIONES FALLIDAS DE DATAFONOS EN COMERCIOS NUEVOS

PRESENTA:

OLIVER ANDRÉS RODRÍGUEZ BLANCO

DIRECTOR:

SEBASTIÁN ZAPATA RAMÍREZ

Bogotá D.C. Noviembre de 2020

Identificación predictiva para instalaciones fallidas de datafonos en comercios nuevos

Oliver Andrés Rodriguez Blanco, *Maestría en Ingeniería y Analítica de Datos, Universidad Jorge Tadeo Lozano*

***Resumen*—**La iniciativa del proyecto es presentar como el aplicar modelos de Aprendizaje Automático para identificar predictivamente las instalaciones fallidas de dispositivos Puntos de venta o Point of sales por sus siglas en inglés en comercios nuevos, los cuales se encontraron interesados en hacer parte de la red de Credibanco, pero por diferentes razones al momento de la visita de respectivo técnico, no aceptaron dicha instalación.

Este proyecto pretende aportar a uno de los principales objetivos estratégicos de CredibanCo, en lograr más participación en el mercado al instalar exitosamente una mayor cantidad de dispositivos POS a nivel nacional, adicionalmente se logrará disminuir el gasto operativo al identificar los comercios que son potencialmente propensos a no aceptar la instalación al momento de la visita, ya que cada una de estas tiene un costo asociado, se instale o no el dispositivo.

Se elaboraron modelos con el fin de predicción efectiva de aquellos comercios que no acepten la instalación del dispositivo, para tal fin se recurrió a diferentes fuentes de datos de la organización como radicaciones de solicitudes del área de servicio al cliente, información transaccional, datos del tipo de hardware de los dispositivos, datos demográficos de los comercios he información comercial.

En el proceso se alcanzó a un x porcentaje de Precisión junto con una Sensibilidad del x porcentaje, adicionalmente se generaron actividades de cara al proceso de visita para instalaciones mediante planes de retención para así garantizar que los comercios identificados por el modelo de ser posible instalación fallida se les ofrezca ofertas tentativas para garantizar dicha instalación y preferencia en los tiempos de agendamiento para la visita del técnico quien realizara la instalación.

Índice de Términos - Algoritmo,AUC, Aprendizaje Automático, Aprendizaje Supervisado, Comercio, Establecimiento, Precisión, Recall, ROC, Sensibilidad, Datafono

***Abstract*—**The project initiative is to present how the application of Machine Learning models allowed to predictively identify failed installations of POS devices (Points of Sale or Point) in new businesses, which were interested in being part of the CredibanCo network. but for different reasons at the time of the visit of the respective technician, they did not accept said installation.

This project objectives to contribute to one of the main strategic objectives of CredibanCo, to achieve a greater market share through the successful installation of a greater number of POS devices nationwide, additionally it will reduce operating expenses by identifying businesses potentially prone to not accept. installation at the time of the visit, since each of these has an associated cost, whether or not the device is installed.

Models were developed in order to automatically find those patterns that would be complex for a human to identify them and allow an effective prediction of those businesses that do not accept the installation of the device, for this purpose different data sources of the organization were used, such as requests customer service, transactional information, device hardware type data, business demographics and commercial information.

As a result of this project, an x percentage of precision was achieved together with a sensitivity of x percentage, additionally activities were generated for the process of visiting the facilities through retention plans in order to guarantee that the businesses identified by the model in what possible offers of failed tentative installation are offered to guarantee said installation and preference in scheduling times for the visit of the technician who will perform the installation.

***Index Terms*— Algorithm, AUC, Machine Learning, Supervised Learning, Commerce, Establishment, Accuracy, ROC, Sensitivity, Dataphone**

I. INTRODUCCIÓN

LAS organizaciones actualmente manejan y almacenan grandes volúmenes de información, provenientes de su operación diaria, dispositivos electrónicos, redes sociales, entre otros. Debido a la cuarta revolución industrial o la era de la transformación digital, han surgido herramientas capaces de tratar, transformar y analizar los datos que las organizaciones almacenan, con el fin de generar valor y tomar decisiones estratégicas en cuanto el cómo mejorar sus productos y/o servicios hacia sus clientes y mediante los análisis descriptivos y predictivos de aprendizaje automático podemos identificar patrones ocultos en el comportamiento de nuestros datos, llegando a predecir lo que puede hacer o no el consumidor [1].

Según el Reporte de Sistemas de Pago - 2020 del Banco de la Republica sobre percepción del uso de los instrumentos de pago de bajo valor (en los pagos habituales), la cual se aplicó a finales de 2019. Esta encuesta tiene como propósito fundamental identificar la disponibilidad y las preferencias del público en relación con estos instrumentos de pago, y su aceptación por parte de los establecimientos comerciales. Cabe resaltar que el efectivo continúa siendo el instrumento más utilizado por la población en sus pagos habituales mensuales (el 88,1% en número de pagos y 87,4% en su valor). Sin embargo, su uso en valor ha caído, ya que en la medición de 2017 registraba un 89,6%. A su vez, el nivel de aceptación de los comerciantes por instrumentos de pago diferentes al efectivo está en el 14,1% para tarjeta débito, 13,4% para tarjeta

crédito, 8,2% para transferencias y 1,8% para el cheque. La principal razón para el uso del efectivo es la ausencia de datáfonos en el negocio, esto nos indica el potencial que aún queda por explorar frente a la instalación de datafonos POS (Puntos de venta o Point of sales por sus siglas en inglés), frente a los comercios, cada vez más los compradores optan por preferir métodos de pago diferentes al efectivo [2].

Credibanco es una empresa colombiana vigilada por la Superintendencia Financiera, con más de 45 años de experiencia en la administración y desarrollo de sistema de pago de bajo valor. Actualmente, promueve los pagos electrónicos en el país a través de la estructuración de negocios que sustituyan el uso de dinero en efectivo, actuando como proveedor de datafonos para los establecimientos comerciales que desean el pago con medios como tarjetas bancarias de crédito y débito. [3]

Actualmente Credibanco se encuentra adentrando en los cambios que suponen convertirse en una empresa 4.0, abriendo las puertas a nuevas tecnologías como las inicialmente mencionadas, teniendo como una de sus metas emplear los datos con los que dispone en su operación para disminuir sus costos y generar oportunidades de negocio mediante el uso de técnicas de aprendizaje automático.

Este documento plantea usar técnicas de aprendizaje automático con el fin de mitigar una serie de sobrecostos que se generan como consecuencia de instalaciones fallidas de datafonos POS en comercios que inicialmente solicitaron el servicio, pero ¿qué se entiende por instalaciones fallidas? Es el resultado de una visita agendada de un técnico a un comercio, con el fin de instalar un datafono, sin embargo, por una serie de causales al momento de la instalación del dispositivo el comercio decide no seguir con el proceso, esto representa unas pérdidas cuantiosas anualmente para la organización, debido a que cada visita, se instale o no el dispositivo, tiene asociados unos costos operativos y es de vital importancia identificar cuáles podrían ser los comercios que caerían en instalaciones fallidas, esto como objetivo organizacional de crear soluciones preventivas, generadoras de ofertas diferentes para este tipo de comercios y dejar de ser reactivos frente a perdidas que se pueden mitigar con el uso de estas herramientas, sin embargo esto representa un reto para la organización debido a restricciones por políticas, el uso de estas herramientas se encuentra restringido por ser open source o de código abierto adicional a que las fuentes de información no se encuentran del todo depuradas obligando a realizar un fuerte proceso de minería de datos.

El presente proyecto fue organizado con base en lo expuesto por Aurélien Géron en su libro Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, este autor propone una lista de chequeo bastante estructurada para proyectos de Aprendizaje Automático, adoptando definir el problema y mirar el panorama general, entendido lo anterior se procedió con la obtención de los datos, ya con esto se realizó un análisis exploratorio de estos datos, ya conociendo la información con la que se dispone se realizó la preparación de estos datos con el fin de exponer mejor los patrones en la exploración y selección de los modelos que para este caso se emplearon la Regresión logística, Arboles de decisión y Bosques Aleatorios por ser los más empleados en este tipo de proyectos y su facilidad en manejar variables categóricas, posteriormente de afinan estos modelos para identificar el que mejores métricas proporcionan llegando a mostrar una solución óptima a la organización. [4]

Seguido a lo anterior, el documento está organizado de la siguiente manera; en la siguiente sección el Marco teórico, en la sección 3. Estado del Arte, en la sección 4 la pregunta de Investigación, sección 5. Resultados, sección 6. Objetivos, sección 7. una explicación del modelo realizado, y finalmente se presentan las conclusiones.

II. MARCO TEÓRICO

Segun Tom Michell. "*El aprendizaje automático es un programa de computador que aprende de la experiencia E, respecto a una tarea T y con medida de rendimiento P, si el desempeño de la tarea T, medido por P, mejora la experiencia E*". [6]

El Aprendizaje Automatizado es una rama de la inteligencia artificial, en gran parte inspirada en el razonamiento humano, que comprende el aprendizaje a partir de experiencia, este aborda, a su vez, una serie de problemáticas que tributan a problemas específicos, entre ellos: los problemas de clasificación, asociación, agrupamiento, y selección de rasgos. [5]

El objetivo principal es hacer aprender un modelo, a partir de datos de entrenamiento etiquetados, que nos permite hacer predicciones sobre datos futuros o no vistos. Aquí el termino supervisado se refiere a un conjunto de muestras donde las señales de salida deseadas (etiquetas) que ya se conocen [8].

El aprendizaje supervisado proporciona una ruta directa para convertir datos en información real y procesable. Al utilizar los datos como un recurso, les permite a las organizaciones comprender y prevenir los resultados no deseados o impulsar los resultados deseados para lo que sea que estén tratando de predecir.

El aprendizaje supervisado se puede dividir en 2 categorías, Clasificación y Regresión, para el primer caso predice una categoría a la que pertenecen los datos por ejemplo: Detección de correo no deseado, predicción de abandono, análisis de sentimiento, detección de raza de perro y para el caso de Regresión predice un valor numérico basado en datos observados previamente por ejemplo: predicción del precio de la vivienda, score de fraude transaccional con tarjetas de crédito, Predicción de altura-peso, estos son algunos ejemplos de modelos de aprendizaje automático supervisado que permiten hacer tanto clasificación como regresión.

Árboles de clasificación

Su principio básico es generar particiones recursivas por reglas de clasificación hasta llegar a una clasificación final, tal que es posible identificar perfiles (nodos terminales) en los que

la proporción de clientes malos es muy alta (o baja) y de esta forma asignar su probabilidad, la idea surgió de la estructura de un árbol que se compone de una raíz, nodos (las posiciones donde las ramas se dividen), ramas y hojas; de manera similar, un árbol de clasificación se construye a partir de nodos que representan los círculos y las ramas son representadas por los segmentos que conectan los nodos. Un árbol de clasificación se inicia desde la raíz, se extiende hacia abajo y generalmente se dibuja de izquierda a derecha. El nodo inicial se llama nodo raíz, mientras los nodos en los extremos de la cadena se les conocen como nodos hoja. Dos o más ramas pueden extenderse desde cada nodo interno, es decir, desde un nodo que no es el nodo hoja [9].

Para los árboles de clasificación la bondad de una división se cuantifica por una medida de impureza; se dice que una división es pura si, después de la división, todas las instancias de la elección de una rama pertenecen a la misma clase [10].

Para el nodo $m_1 N_m$ es el número de instancias de entrenamiento que alcanza el nodo m. Para el nodo raíz, esto es N. N_m^i pertenecen a las clases C_1, donde $\sum_i N_m^i = N_m$. Dado que una instancia alcanza el nodo m, la estimación de la probabilidad de la clase Ci es:

$$\hat{P}(C_i nx, m) \equiv p_m^i = \frac{N_m^i}{N_m}$$

El nodo m es puro si p_m^i para todo i son 0 o bien 1, Es 0 cuando ninguna de las instancias del nodo m son de Clase Ci, y es 1 si todos estos casos son de Ci. Si la división es pura, no es necesario dividir más y se puede añadir un nodo hoja etiquetado con la clase para la cual p_m^i es 1.

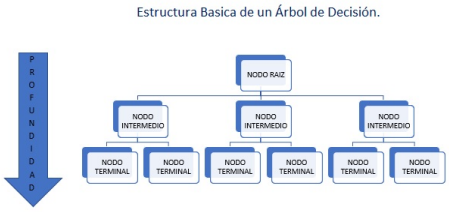

Figura 1. *Representación gráfica de la estructura de nodos en un árbol de decisión. Elaboración propia.*

Máquinas de Soporte Vectorial

La teoría de las Máquinas de Soporte Vectorial (SVM por su nombre en inglés Support Vector Machines) es una nueva técnica de clasificación y ha tomado mucha atención en años recientes [11], Una SVM primero mapea los puntos de entrada a un espacio de características de una dimensión mayor (si los puntos de entrada están en $\Re^2$ entonces son mapeados por la SVM a $\Re^3$) y encuentra un hiperplano que los separe y maximice el margen m entre las clases en este espacio como se aprecia en la figura 2 y la frontera de decisión debe estar tan lejos de los datos de ambas clases como sea posible. [12]

Si la distribución de las observaciones es tal, que se pueden separar linealmente de forma perfecta en las dos clases (+1 y -1), como se indica en la figura 3 y si el conjunto de puntos no es linealmente separable, como se indica en la figura 4.

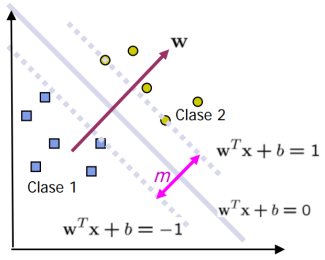

Figura 2. *Concepto de SVM [8]*

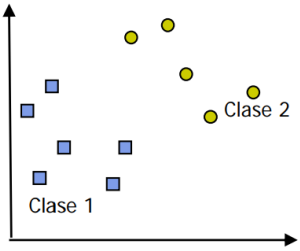

Figura 3. *Caso linealmente separable [8]*

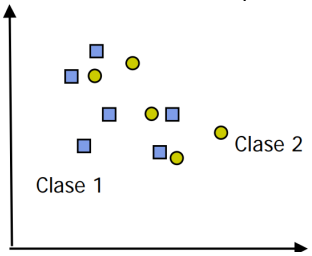

Figura 4. *Caso linealmente no separable [8]*

Las SMV se empleaban con mayor frecuencia antes de la llegada del Deep Learning, el uso de las SVM eranm incluso más confiables que los modelos de redes neuronales, ya que las matemáticas de los SVM se entiende muy bien y la propiedad de obtener el margen de separación máximo era muy atractivo, en este escenario las redes neuronales llegaban a presentar predicciones erradas, sin embargo actualmente las redes neurales de aprendizaje profundo generan una mayor capacidad de entrenarse y su capacidad de generalizar es mayor a los SVM.

Regresión Logística

Es un procedimiento cuantitativo de gran utilidad para problemas donde la variable dependiente toma valores en un conjunto finito, esta es empleada cuando la variable de respuesta Y es polinómica, pero es especialmente útil en particular cuando

solo hay dos posibles respuestas (cuando la variable de respuesta es dicotómica), que es el caso más común, a pesar de su nombre, es un modelo lineal para clasificación en lugar de regresión. La regresión logística también se conoce en la literatura como regresión logit, clasificación de máxima entropía (MaxEnt) o clasificador log-lineal. En este modelo, las probabilidades que describen los posibles resultados de un solo ensayo se modelan utilizando una función logística . [13]

Sea Y una variable dependiente binaria (con dos posibles valores: 0 y 1). Sean un conjunto de k variables independientes,$(x_0, x_1, ..., x_k)$, observadas con el fin de predecir o explicar el valor de Y. El objetivo consiste en determinar:

$$P\left[Y = 1/X_1, X_2, ...X_k\right] \longmapsto P\left[Y = 0/X_1, X_2, ...X_k\right] = 1 - P\left[Y = 1/X_1, X_2, ...X_k\right]$$

Para ello, se construye el modelo $P\left[Y = 1/x_1, x_2, ...x_k\right] = p\left(X_1, X_2, ...X_k\right)$ donde: $p\left(X_1, X_2, ...X_k; \beta\right)$:$R^k \longrightarrow [0, 1]$ que depende de un vector de parámetros $\beta = (\beta_1, \beta_2, ..., \beta_k)$.

El modelo logístico establece la siguiente relación entre la probabilidad de que ocurra el suceso, dado que el individuo presenta los valores $(X = x_1, X = x_2, ...X = x_k)$:

$$P\left[Y = \frac{1}{x_1, x_2, ...x_k}\right] = \frac{1}{1+e^{(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - ... - \beta_k x_k)}}$$

El objetivo es hallar los coeficientes $(\beta_0, \beta_1, ..., \beta_k)$ que mejor se ajusten a la expresión funcional [14]. Siendo P la probabilidad positiva del evento de puede identificar la función de la regresión logística o Logit.

$$Logit\left(P_i\right) = ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i}$$

Teniendo en cuenta que se quiere predecir la probabilidad que una muestra sea de una clase determinada se calcula la inversa de la función Logit o función Sigmoide.

$$Y = \frac{1}{1 + e^{-f(x)}}$$

f(x) siendo la función analitica en X, pudiendoce expresar como una serie de potencias emergentes. [15]

Boques Aleatorios.

Este modelo es una combinación de árboles predictivos (clasificadores débiles), el cual trabaja con una colección de árboles no correlacionados y los promedia [16], pertenecen a modelos tipo Ensamblados, los cuales permiten alcanzar una mayor precisión y estabilidad del modelo. Estos proveen una mejora significativa a los modelos de árboles de decisión.

Los bosques aleatorios se consideran la panacea en todos los problemas de ciencia de datos, debido a que son útiles para regresión y clasificación, permiten combinar modelos débiles en uno robusto mediante la construcción de varios árboles como se representa en la figura 5, en donde cada uno da una clasificación votando por una clase, el resultado con mayor número de votos en todo el bosque es la salida seleccionada del modelo, en cuanto regresión se toma el promedio de las salidas de todos los árboles (Ver Fig 5) [17].

Los modelos de ensamble como los bosques aleatorios manejan un proceso denominado Bagging, siendo una técnica empleada para reducir la varianza de las predicciones a través de la combinación de los resultados de varios modelos, cada uno de estos modelos toman diferentes subconjuntos de la misma población (Ver Fig 6) [18].

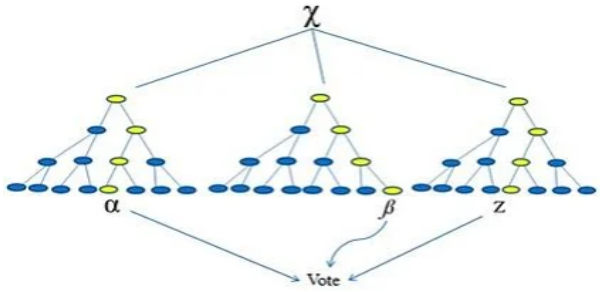

Figura 5. *Representación gráfica Bosque Aleatorio. [17]*

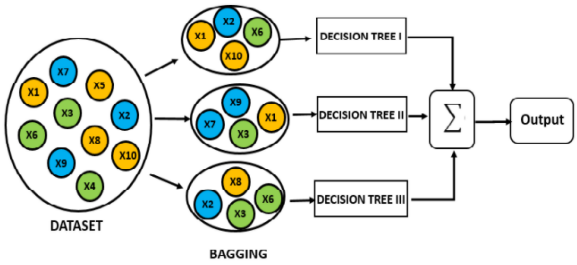

Figura 6. *Representación gráfica proceso Bagging. [18]*

En cuanto el aprendizaje no supervisado, no se abordará el tema en profundidad debido a que no es alcance de este proyecto, sin embargo, solo le damos las características al algoritmo, nunca las etiquetas. Queremos que nos agrupe los datos que le dimos según sus características. El algoritmo solo sabe que como los datos comparten ciertas características, de esa forma asume que pueda que pertenezcan al mismo grupo. [19]

Empleando la agrupación, permite encontrar grupos diferenciados en los datos suministrados, para ello, existen algoritmos de agrupamiento cuya función es encontrar la estructura en los datos de manera que los elementos del mismo clúster (o grupo) sean más similares entre sí en comparación con los de clústeres diferentes.

Los algoritmos de aprendizaje no supervisados son muy empleados y útiles para resolver problemas del mundo real como la detección de anomalías, la recomendación de sistemas, la agrupación de documentos o la búsqueda de clientes con intereses comunes basados en sus compras,

algunos de los algoritmos de agrupación más comunes empleados en aprendizaje no supervisado son [20]:

- K-Medias
- Clusterización Jerárquica
- Modelo de Agrupamiento Gaussiano

Rendimiento de los modelos:

Considerando lo anterior, se puede deducir la importancia que llega a tener un entrenamiento adecuado en los modelos empleados al ingresar los datos para tal fin, sin embargo, es igual de importante medir el rendimiento de estos modelos entrenados, lo bien que pueden llegar a generalizar las predicciones sobre datos nuevos, definiendo métricas de evaluación para valorar el rendimiento de un modelo de aprendizaje automático, siendo un componente integral de cualquier proyecto de ciencia de los datos. Su objetivo es estimar la precisión de la generalización de un modelo sobre los datos futuros (no vistos/fuera de muestra).

Una matriz de confusión es una representación matricial de los resultados de las predicciones de cualquier prueba binaria que se utiliza a menudo para describir el rendimiento del modelo de clasificación (o "clasificador") sobre un conjunto de datos de prueba cuyos valores reales se conocen. [21] Cada predicción que realizan los modelos pertenecerá a uno

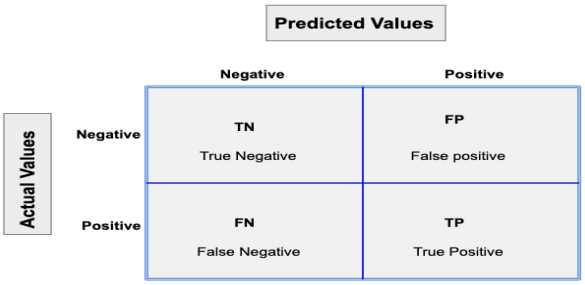

Figura 7. *Matriz de confusión con 2 etiquetas de clase. [21]*

de los cuatro posibles resultados expuestos en la figura 7, partiendo con el estado de su valor real.

Para IBM, el Accuracy es una medida que especifica qué parte del resultado del modelo de aprendizaje automático era precisa en comparación con el resultado del anotador humano [22]. En otras palabras, se define como la cantidad de veces que el modelo acerto una afirmación, sobre el total de datos de entrada, sin embargo, este valor puede en ocasiones parecer alto, cuando en verdad la parte relevante no lo es tanto, y es causado por un desbalance en la cantidad de muestras verdaderas y positivas.

$$Accuracy = \frac{Verdaderos\ Positivos + Verdaderos\ Negativos}{Total}$$

La precisión se define como la cantidad de casos verdaderos positivos sobre la cantidad total de todo lo que el modelo indica que era positivo. En otras palabras, de todo lo que el algoritmo predijo como positivo, se evalúa cuánto de eso era cierto.

$$Precisión = \frac{Verdaderos\ Positivos}{Total\ clasificados\ positivos}$$

Por otro lado, está esta otra métrica el Recall o Sensibilidad con un enfoque diferente. Se compara la cantidad de casos clasificados como verdaderos positivos sobre todo lo que realmente era positivo. Y a diferencia de la anterior (precisión), antes comparábamos lo que el algoritmo dice con lo que es cierto, en cambio acá, lo que él dice contra lo que no dijo que era cierto.

$$Sensibilidad = \frac{Verdaderos\ Positivos}{Total\ positivos}$$

El valor F1 se utiliza para combinar las medidas de precisión y Recall en un sólo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y el Recall entre varias soluciones, F1 se calcula haciendo la media armónica entre la precisión y la exhaustividad.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Adicionalmente la Especificidad la cual es la verdadera tasa negativa o la proporción de verdaderos negativos a todo lo que debería haber sido clasificado como negativo [23].

$$Especificidad = \frac{Verdaderos\ Negativos}{Verdaderos\ Negativos + Falsos\ Positivos}$$

Una representación gráfica que ilustra la relación entre la sensibilidad y la especificidad de un sistema clasificador para diferentes puntos de corte llamado curva ROC (Receiver Operating Characteristic), fue desarrollada por ingenieros eléctricos en la II Guerra Mundial, para medir la eficacia de la detección de objetos enemigos en el campo de batalla mediante señales de radar. Su uso está muy extendido en medicina, para validar técnicas diagnósticas. Más recientemente con el auge de las técnicas de aprendizaje automatizado, se han empleado las curvas ROC para evaluar diferentes algoritmos de clasificación (Ver Fig 8).

El mejor método posible de predicción se situaría en un punto en la esquina superior izquierda, o coordenada (0,1) del espacio ROC, representando un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún falso positivo). A este punto (0,1) también se le llama una clasificación perfecta. Por el contrario, una clasificación totalmente aleatoria (o adivinación aleatoria) daría un punto a lo largo de la línea diagonal, que se llama también línea de no-discriminación, desde el extremo inferior izquierdo hasta la esquina superior derecha (independientemente de los tipos de base positivas y negativas) [24].

Existen numerosos modelos de aprendizaje automático, cada uno con sus ventajas y desventajas para las empresas. Lo más importante en este caso es estar conectados con los objetivos del negocio, la idea es definir cuáles serán los atributos más sobresalientes del modelo a seleccionar y la evaluación del rendimiento del modelo siendo una de las fases principales en el proceso de ciencia de datos [26].

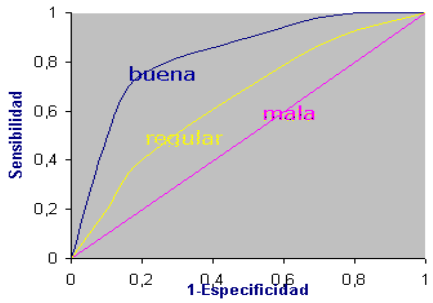

Figura 8. *Tipos de curva ROC. [25]*

III. ESTADO DEL ARTE

El Aprendizaje Automático ofrece un valor potencial a las empresas que tratan de aprovechar la información con la que disponen y les ayuda a comprender mejor los cambios sutiles en el comportamiento, las preferencias o la satisfacción del cliente. Los líderes de negocios están empezando a descubrir que muchas cosas que están sucediendo dentro de sus organizaciones e industrias no pueden ser entendidas a través de una consulta. No son las preguntas que normalmente se conocen; son los patrones ocultos en los datos que pueden aportar en diferentes actividades económicas [27].

Transporte y logística:

Sus aplicaciones en el sector logístico son muy amplias. Una de las más importantes es la de realizar predicciones relativas a la demanda, uno de los puntos clave en la cadena de suministro. Esta tecnología permite recoger datos, almacenarlos e interpretarlos en tiempo real, así como detectar patrones para entender los déficits o excesos de demanda y actuar en consecuencia.

Puede pronosticar condiciones de circulación de las mercancías y los vehículos, así como de las condiciones climatológicas. En la cadena logística, en la que el buen funcionamiento de cada eslabón de la cadena es determinante para que toda ella opere como debe, los modelos de aprendizaje automático son muy apreciados para anticiparse a los posibles errores técnicos que puedan darse en dispositivos tecnológicos.

El aprendizaje automático tiene aplicación en la planificación de rutas, comprueba por sí mismo cuál es la mejor ruta en cada momento para realizar el transporte de una mercancía gracias un sistema de aprendizaje basado en algoritmos que contribuye a evitar circulaciones lentas o restricciones de tráfico. Esto permite usar el recorrido más rentable en base a criterios de número de kilómetros o velocidad a la que puede circularse por una vía, lo que repercute en un menor consumo de combustible. También desempeña, por lo tanto, un papel básico en la optimización de procesos logísticos en la gestión de una flota vehicular [28].

Petróleo y gas:

Las compañías petroleras están pilotando el uso del aprendizaje automático en toda la cadena de valor. Desde la búsqueda de nuevos yacimientos petrolíferos meses más rápido de lo que sea posible hasta la predicción del fracaso de piezas de equipos multimillonarias, las compañías petroleras están muy entusiasmadas con sus perspectivas.

El mantenimiento predictivo de los equipos es un uso clave en el petróleo, donde técnicas como la agrupación en clústeres y el procesamiento del lenguaje natural han permitido a las empresas integrar conjuntos de datos de máquinas con registros de mantenimiento. El mantenimiento predictivo de algunos activos como bombas y compresores se ofrece como tecnología ”off- the-shelf” [29].

Salud:

En el ámbito de la salud, el objetivo del aprendizaje automático es dotar a las herramientas informáticas de un sistema que les permita procesar los datos de información sanitaria (procedentes sobre todos de la Historia Clínica) de tal forma que puedan realizar una serie de acciones relacionadas con la atención y gestión de la salud: desde emitir diagnósticos hasta arrojar datos predictivos sobre determinadas enfermedades o la efectividad que un tratamiento tendrá en un perfil de paciente, por ejemplo [30].

Esta tecnología desempeña un rol muy importante en la formación y actualización de todos los profesionales de la salud, ya que les permite disponer del conocimiento preciso en el momento concreto en que lo necesitan: *“Estamos usando el aprendizaje automático para comprender mejor en qué y cómo están trabajando los investigadores con el objetivo de proporcionarles la información más relevante para ellos. Manejamos una cantidad enorme de datos no estructurados como imágenes médicas, material audiovisual procedente de estudios académicos, artículos de revistas, libros, etc. El objetivo es extraer toda esa información, seleccionarla según el contexto y ofrecer los resultados”*, comenta Dan Olley [31].

Gobiernos:

Dependencias de gobierno como seguridad pública y los servicios públicos tienen una necesidad particular del aprendizaje automático porque tienen múltiples fuentes de datos de las que se pueden extraer insights. Por ejemplo, el análisis de datos de sensores identifica formas de incrementar la eficiencia y ahorrar dinero. Asimismo, el aprendizaje basado en máquina puede ayudar a detectar fraude y minimizar el robo de identidad [32].

Servicios Financieros:

Las técnicas de aprendizaje automático tienen una enorme aplicación en el sector financiero, en diferentes ámbitos:

- **Negocio:** detectar patrones de comportamiento en los datos de los clientes permite un conocimiento mucho mayor de los mismos, para diseñar y ofrecer soluciones más específicamente personalizadas.
- **Riesgo:** el análisis avanzado de variables independientes en los estados financieros proporciona información valiosa acerca del comportamiento de los niveles de incumplimiento en el riesgo de crédito.
- **Fraude:** la detección temprana se basa en el descubrimiento y detección automáticos de asociaciones y reglas que pueden significar patrones interesantes; en la creación sistemas expertos para codificar la experiencia; en el reconocimiento de clases, clusters o patrones de comportamiento sospechosos; en técnicas de aprendizaje automático para identificar dichos patrones, etc.
- **Eficiencia:** la identificación automática de patrones de comportamiento puede contribuir al uso más eficiente de recursos, como por ejemplo en los servicios de call center, etc [33].
- **Operativo:** Identificación predictiva de fallos en datafonos o dispositivos POS. [34]

IV. PREGUNTA DE INVESTIGACIÓN

Una de las principales funciones de Credibanco, empresa vinculada al sector financiero de bajo valor, es ofrecer el servicio de arrendamiento de datafonos para compras con tarjetas bancarias en comercios tanto nuevos como antiguos a nivel nacional, sin embargo, no todos los comercios que solicitan la instalación, al momento de la visita del técnico aceptan que se instale la terminal o datafono en sus establecimientos.

Teniendo en cuenta lo anterior, algunos de los posibles motivos podrían ser que la competencia instalo primero o demora en el servicio entre otros, esto acarrea unos costos de $70.000 por visita de instalación fallida, desgaste operativo y administrativo para gestionar la instalación.

Se evidenció que para el año 2019 se presentaron en pérdidas operativas por este concepto, alrededor de doscientos millones de pesos, es de aclarar que los clientes son contactados vía telefónica antes de realizar la instalación en donde el problema son aquellos clientes que aceptan la instalación en la llamada, pero al momento de la llegada del técnico, no aceptan el Datafono **(Fuente: Datos de resultados operativos - CredibanCo)**.

Comprendiendo la necesidad de la organización, se desarrolló un modelo de aprendizaje automático supervisado, que permitiera llegar a la solución de la pregunta ¿Cómo emplear modelos de aprendizaje automático para identificar proactivamente, aquellas solicitudes de comercios nuevos, que al momento de la visita del técnico no aceptan la instalación del datafono para evitar perdida de participación en el mercado y pérdidas económicas operativas?

Como insumo para el modelo, se seleccionaron variables que describieran el tipo de tecnología de las terminales creadas para dicha instalación, la ubicación geográfica del comercio, horarios de agendamiento de visita, antigüedad del comercio en la industria y volumen transaccional por actividad económica dependiendo la ciudad, además, se modelaron técnicas de aprendizaje automático descritas en el marco teórico, con el fin de emplear el modelo con mejores métricas de predicción, mediante el uso de técnicas clásicas de ajustes de hiperparametros, entrenamiento y testing.

Teniendo que todos los comercios solicitantes deben visitarse por parte de los técnicos independientemente su resultado, el despliegue operativo de este proyecto no tendría mayor impacto frente a costos asociados a este, ya que se encontraría inmerso dentro de las actividades y recursos empleados actualmente, el cambio que generaría el modelo frente al proceso, es que al tener identificados los comercios que posiblemente puedan llegar a rechazar la instalación del datafono, a nivel de producto se ofrecerían planes de retención, siendo prioridad de visita en el agendamiento del técnico, esto con el fin de que el comercio identificado por el modelo como instalación fallida acepte dicha instalación con la oferta comercial para retención de clientes.

Con el fin de suplir lo requerido operativamente por la organización, se espera que el modelo seleccionado genere al menos un 70% de precisión, además se tendrá en cuenta la sensibilidad con el fin de impulsar mejores prácticas de servicio en visitas y desarrollar nuevas ofertas de valor que permitan aumentar participación en el mercado y aumento transaccional en los datafonos de la organización.

Los modelos empleados en este proyecto fueron entrenados con dos técnicas clásicas para tal fin, la partición 70% de los datos para entrenamiento y 30% para test, y la segunda técnica es K-fold Cross validation, esto se hace con el fin de garantizar que los modelos empleados generalizan adecuadamente y no se llega a presentar sobre entrenamiento.

Se procede a presentar los resultados de los modelos empleados:

Regresión Logística:

Actualmente siendo uno de los algoritmos más aceptados por la comunidad de ciencias de datos, por ser la clasificación binaria una de sus principales fortalezas, además de su simplicidad, se tomó en cuenta para este proyecto, alcanzando en entrenamiento una exactitud del 85%, con un total de datos para entrenamiento de 187.333 y unas métricas de test, con una exactitud igualmente al de entrenamiento del 85%,

una precisión del 39%, una sensibilidad del 44%, con un AUC de 71.6% y con un total de datos para test de 80.286.

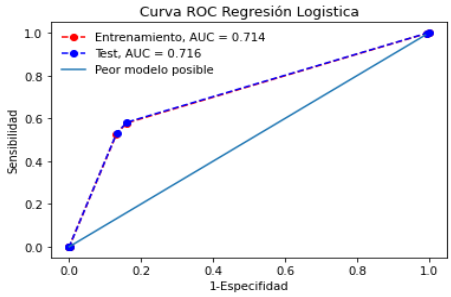

Figura 9. *Curva ROC y AUC Regresión Logística. Fuente: Resultados obtenidos a través de Python.*

Árbol de decisión:

Aunque estos modelos tienden a presentar sobre entrenamiento, se empleó este modelo como opción a tener en cuenta debido a su simplicidad para entender e interpretar, no requiere una preparación de los datos demasiado exigente (aunque la implementación de Scikit-Learn no soporta valores nulos) y permite trabajar tanto con variables cuantitativas como cualitativas, se compararon los resultados del árbol de decisión con y sin calibrar hiperparametros.

Para el modelo por defecto o sin calibrar hiperparametros, en entrenamiento se presentó una exactitud del 99.7%, identificando un sobre entrenamiento en los datos, sus resultados en test con una exactitud del 87.9%, una precisión de 58.3%, sensibilidad de 55.5% y un AUC de 76%. Para el modelo con hiperparametros como una profundidad de 19 niveles y un mínimo de muestras requeridas por nodo para realizar una división de 250 datos. Disminuyendo la exactitud a 91.1%, eliminando el sobre ajuste en entrenamiento, para test se incrementó la exactitud a 90.7%, una precisión de 49%, sensibilidad de 74.6% y un AUC de 86.8%, generando mejoría frente al resultado del modelo generado por defecto.

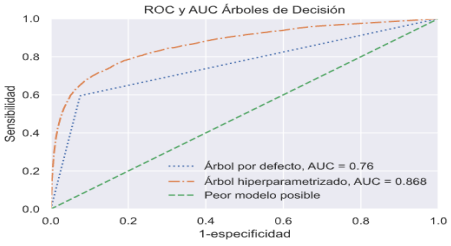

Figura 10. *Curva ROC y AUC Arboles de Decisión. Fuente: Resultados obtenidos a través de Python.*

Bosques Aleatorios:

Se empleo este modelo como candidato debido a que este algoritmo presenta la aleatoriedad como su principal fortaleza, pues le brinda flexibilidad suficiente como para poder obtener gran variedad de árboles, evitando el riesgo de caer en sobre entrenamiento, generando variedad de muestras que, producen una salida concreta, además funciona bien incluso sin llegar a emplear ajustes en sus hiperparámetros.

Al igual que el modelo de árboles de decisión, se apreció un sobre ajuste con los datos de entrenamiento, con una exactitud del 99.5%, en test presento una exactitud de 91%, una precisión de 77% y una sensibilidad de 49%. Con el fin de disminuir el sobre ajuste en entrenamiento se emplearon una serie de pruebas con diferentes hiperparámetros, disponiendo en entrenamiento una exactitud de 94%, el resultado en test se identifica una exactitud de 90%, precisión de 70%, una sensibilidad 57% y un AUC de 89%. Teniendo en cuenta los resultados de este modelo y las ventajas anteriormente indicadas, se optó por emplearlo como método para identificar predictivamente comercios con instalaciones fallidas.

V. OBJETIVOS

A. *OBJETIVO GENERAL*

Establecer un modelo predictivo con el fin de identificar los posibles comercios que no aceptan la instalación del datafono al momento de la visita del técnico, evitando pérdidas operativas para Credibanco.

B. *OBJETIVOS ESPECÍFICOS.*

- Definir las fuentes o bases de datos que se emplearan para este proyecto.
- Realizar preprocesamiento de datos de las fuentes.
- Diseñar los respectivos modelos predictivos propuestos (Regresión Logística, Árboles de Decisión y Bosques Aleatorios.) para clasificación de comercios con instalación de Datafono fallida.
- Estimar y evaluar los resultados de los modelos propuestos, con el fin de aplicar aquel modelo con las mejores métricas de desempeño (Exactitud, Precisión y Sensibilidad).

VI. MODELO

Este proyecto fue abordado bajo la metodología propuesta por Aurélien Géron en su libro Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition, adoptando 7 de los 8 pasos que el autor propone [35],

el primer paso se encuentra dado bajo la pregunta de investigación realizada anteriormente.

- Definir el problema y mirar el panorama general.
- Obtener los datos.
- Explorar los datos para obtener información.
- Preparación de los datos.
- Exploración y selección de modelos
- Afinar los modelos.
- Presentación de la solución.

Obtener los datos.

Para la extracción de la información que será insumo para el modelo, se empleó una consulta a una base de datos desarrollada con el fin de ser la fuente de datos para un proyecto de Aprendizaje Automático que permite la identificación predictiva de fallas en datafonos de CredibanCo [34].

El almacenamiento de esta consulta a la base de datos se encuentra alojada en un motor de base de datos analítico, licenciado por IBM llamado Netezza, Esta consulta consta de 19 variables, 1 etiqueta y 1 campos destinados a la identificación de los registros, para un total de 21 atributos y 267.635 registros, pertenecientes a solicitudes realizadas durante el año 2019.

Las variables independientes que alimentaran el modelo presentan las siguientes características:.

- Variables que describen la ubicación geográfica del comercio, incluidas 2 en el modelo (Categórica).
- Variables de tiempo, indican los tiempos de atención de la solicitud, incluidas 6 variables en el modelo (Numérica).
- Variables de caracterización económica y financiera, describe la actividad económica con la que se desempeña el comercio y su entidad financiera adquirente, incluidas 8 variables (Categórica).
- Variables legales, describen el ámbito contractual del comercio que se realiza la solicitud, incluida 1 variable (Categórica).
- Variables de hardware, describes atributos físicos del tipo de terminal o datafono, incluida 1 variable (Categórica).
- Variables transaccionales, describen el comportamiento transaccional de la actividad económica del comercio solicitante, incluida 2 variable (Numérica).
- Variable Objetivo, contiene la etiqueta a predecir, indica si se instalo el datafono o no (Categórica).

Explorar los datos para obtener información.

Ya contando con la información que alimentara el modelo, se realizaron los análisis invariados con el fin de identificar las medidas de tendencia central y la distribución para las variables numéricas, en cuanto las variables categóricas, siendo la mayoría se empleó un enfoque Bivariado, mediante el análisis con tablas de contingencia, con el fin de compararse frente a la variable objetivo, los errores de calidad en los datos fueron tratados y manejados durante la construcción de la base de datos de donde se realizó la consulta para este proyecto, por ende no fue necesario recurrir a una eliminación o imputación de valores por campos sin información.

Se identifico que los datos empleados para el entrenamiento del modelo, presentan desbalanceo con un 13.6% (Ver Fig 11) pertenecen a la etiqueta de instalaciones fallidas, siendo lo que se pretende llegar a predecir con el modelo, este es otro de los motivos por los cuales se adoptó al bosque aleatorio como opción al requerimiento de la organización, teniendo en cuenta que uno de sus hiperparámetros, permite mitigar este desbalanceo en la variable objetivo.

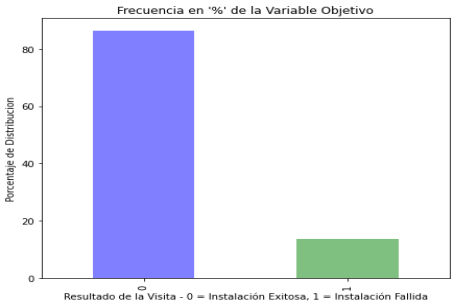

Figura 11. *Frecuencia Relativa de la Variable Objetivo. Fuente: Resultados obtenidos a través de Python.*

El análisis bivariado permitió identificar el comportamiento del resultado de las instalaciones de datafonos frente a las variables insumo para el modelo, esto con el fin de identificar dependencias e independencias asociadas a la variable objetivo, generando como resultado, un conocimiento de importancia que estas variables pueden tener para predecir el resultado de visita de instalación de datafono a un comercio.

	34C	EMERGENTE	POTENCIALES	VIP
Exitosa	44480	90468	53485	42600
Fallida	3794	24413	4777	3602
All	48274	114881	58262	46202
% Fallida	7,86%	21,25%	8,20%	7,80%

Figura 12. *Ejemplo análisis bivariado, etiqueta de resultado de instalación Vs el tipo de comercio según la organización, identificando que la participación mayoritaria en proporción con sus solicitudes se da para comercios tipo emergentes, seguido por los comercios tipo potenciales. Fuente: Elaboración propia*

Paso siguiente se realizó análisis de correlaciones, para este caso se empleó una técnica reciente llamada Phi K correlación, basado en varios refinamientos de la prueba de hipótesis de

Pearson de independencia de dos variables, Las características combinadas de Phi K forman una ventaja sobre los coeficientes existentes. En primer lugar, funciona de manera coherente entre variables categóricas, ordinales y de intervalo. En segundo lugar, captura la dependencia no lineal. En tercer lugar, vuelve al coeficiente de correlación de Pearson en el caso de una distribución de entrada normal bivariado. Estas son características útiles al estudiar la matriz de correlación de variables con tipos mixtos. [36]

Como resultado de este análisis, se identificó alta correlación que las variables transacciones por tipo de actividad económica y una baja correlación con las variables que describen tiempo, en la siguiente imagen se detalla el grado de correlación entre las variables (Ver Figura 13).

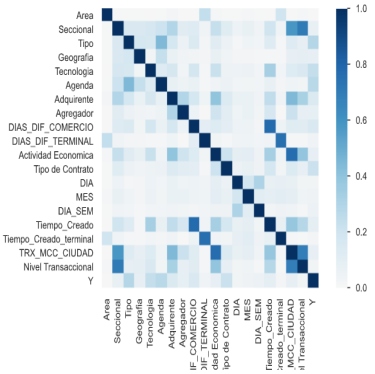

Figura 13. *Análisis de correlación. A mayor tonalidad indica mayor correlación. Fuente: Resultados obtenidos a través de Python.*

Durante el análisis se identificaron variables que presentaban gran cantidad de datos atípicos, a modo de ejemplo se identificó que para la variable que indica la cantidad de días entre la fecha de creación del comercio como negocio y la fecha solicitud de instalación del datafono un volumen importante de datos que fueron considerados atípicos frente a la etiqueta a predecir.

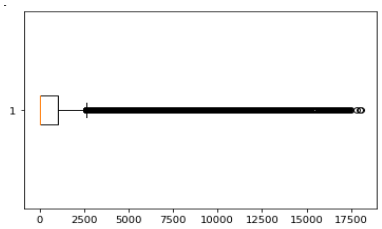

Figura 14. *Análisis de valores atípicos en la variable de tiempo. Fuente: Resultados obtenidos a través de Python.*

Preparación de los datos.

Durante el análisis de los datos se identificaron valores atípicos, debido a su gran cantidad de datos, se tomó la decisión de no eliminarlos y con base al conocimiento del negocio, se crearon categorías, agrupando valores por rangos logrando así no disminuir las muestras para el modelo y con el fin de evitar colinealidad entre la variable original y la variable con las agrupaciones además de evitar el aumento de la cardinalidad del insumo de datos, se optó por eliminar la variable original, dejado así la variable con las agrupaciones por rangos de valores. Teniendo en cuenta que la mayoría de variables son categorías, se optó por hacer transformación a dummy dichas variables.

Presentación de la solución.

Ya conociendo los resultados de los modelos y sus métricas (Ver Hoja 7 y 8), se procede a construir la solución al requerimiento de la organización, se procedió a diseñar el algoritmo seleccionado (Bosques Aleatorios) empleando como lenguaje de programación Python y diferentes librerías, como por ejemplo, Pandas para el manejo de Dataframes, Numpy para el uso de cálculos matriciales, Scikit learn para cálculos estadísticos y diseño de modelos predictivos, Seaborn y Matplotlib para el uso grafico de resultados y para el uso de métricas se emplearon Confusión matrix, roc curve, auc, recall score, precisión: score, f1 scores y para el modelo se empleó el módulo de Scikit learn RandomForestClassifier:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
from sklearn import model_selection
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import plot_confusion_matrix
from sklearn.metrics import mean_squared_error
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from sklearn.model_selection import ParameterGrid
from sklearn.inspection import permutation_importance
from sklearn import metrics
import seaborn as sb
```

Figura 15. *Paquetería empleada para la construcción del modelo. Fuente: Resultados obtenidos a través de Python.*

Ya con los datos listos para su modelamiento, se inició con la calibración del hiperparámetro *n_ estimators*, el cual permite ingresar al modelo el número de árboles de decisión que se requiera en el bosque aleatorio, se empleó el uso de un ciclo FOR para determinar en un rango de entre 5 a 150 árboles de decisión, el comportamiento del coeficiente de determinación o R2, esto significa que mientras más cerca esté del 1 estará más ajustada a la variable que se intenta probar, que para este caso, es el comportamiento incremental de la cantidad de árboles de decisión, mientras que en el caso contrario, es decir, cuanto más se acerca a 0, menos fiable será ya que estará menos ajustado el modelo, todo esto empleando el método Out-of-Bag Error o la estimación de la tasa de error OOB, siendo calculada a partir de observaciones fuera de la bolsa de entrenamiento. La estimación del error sugiere que cuando el modelo sea aplicado a nuevas observaciones, el modelo es exacto en un 90% empleando una cantidad de 146 árboles en este hiperparámetro.

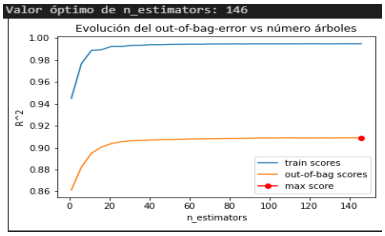

Figura 16. *Resultado de calibración del cálculo de la cantidad de árboles requerido para el modelo. Fuente: Resultados obtenidos a través de Python.*

Seguido, igualmente mediante un ciclo FOR en un rango de 3 a 50 niveles de profundidad, se calculó el valor óptimo para el hiperparámetro *max_ depth*, con el fin controlar la cantidad de ramas que pueda tener cada uno de los árboles generados, evitando así que el modelo genere vías de respuesta específicas para cada dato y genere sobre entrenamiento, llegando a una profundidad optima de 27 hojas.

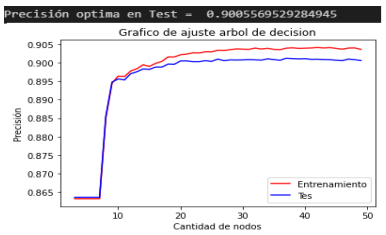

Figura 17. *Resultado de calibración del cálculo de la cantidad de ramas requeridas para el modelo. Fuente: Resultados obtenidos a través de Python.*

Teniendo en cuenta que los datos presentan un desbalance entre sus etiquetas (Ver Fig. 11), se calibro en hiperparámetro *class_ weight*, esto con el fin de dar mayor peso a la etiqueta con menos muestras, dando balance entre las etiquetas, aportando a generar resultados generalizables, en este caso la proporción de cada 7.5 solicitudes de instalación de datafono, 1 es fallida, sin embargo este hiperparámetro permite hacer de manera automática, la mejor combinación de parámetros, llegando a un resultado de igualdad de pesos entre las etiquetas usando la configuración *class_ weight = 'auto'*.

Se opto por emplear el hiperparámetro *criterion = 'entropy'* sobre *'gini'*, debido al aumento en 3 puntos la sensibilidad, siendo importante esta métrica al momento de clasificar las etiquetas verdaderas.

Presentación de la solución.

Como se ha indicado durante el desarrollo del modelo, la precisión y la sensibilidad son las métricas de mayor importancia, teniendo claramente en cuenta el resto de métricas para llegar a esta etapa, la base de datos para entrenamiento fue elaborada con el 70% del total de los datos y el restante 30% se destinó para test, siendo los resultados que se presentaran a continuación.

Con resultado de las pruebas de diferentes hiperparametros, se llegó a estructurar el *RandomForestClassifier* de la siguiente manera:

```
modelRFHP = RandomForestClassifier(n_estimators=146, max_depth=27,
                                   min_samples_split=10,class_weight='auto',
                                   random_state=1,criterion='entropy')
```

Figura 18. *Modelo con calibración final de hiperparámetros para Bosques Aleatorios. Fuente: Resultados obtenidos a través de Python.*

Se obtuvieron los siguientes resultados según matriz de confusión para test

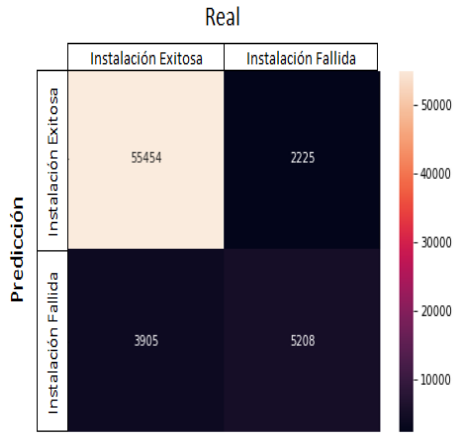

Figura 19. *Matriz de confusión en test. Fuente: Resultados obtenidos a través de Python.*

Empleando el método classification_ report incluido en la librería Sklearn se visualizan los siguientes resultados.

	precision	recall	f1-score	support
0	0.93	0.96	0.95	57679
1	0.70	0.57	0.63	9113
accuracy			0.91	66792
macro avg	0.82	0.77	0.79	66792
weighted avg	0.90	0.91	0.90	66792

Figura 20. *Métricas obtenidas en test. Fuente: Resultados obtenidos a través de Python.*

Según el resultado de la curva ROC, se cuenta con área para mejorar el poder predictivo del modelo, sin embargo, considerando la dificultad para conseguir el insumo y la capacidad de procesamiento limitada para este proyecto, se presentan resultados favorables.

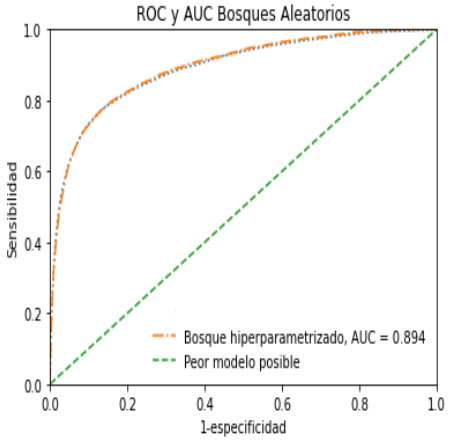

Figura 21. *Curva ROC y AUC del modelo de Bosques Aleatorios, con el 30 % de los datos para test. Fuente: Resultados obtenidos a través de Python.*

VII. PASOS SIGUIENTES.

- Se espera realizar antes del paso a producción, pruebas con intervalos de tiempo diarios durante 15 días, esto a fin de analizar el resultado de las predicciones mientras los técnicos cierran las solicitud en terreno, permitiendo identificar si para el paso a producción el modelo requiere o no una calibración adicional y garantizar que lo ofrecido en esta solución sea realmente lo que se evidencia al momento en que un comercio se comunica con la organización, solicitando la instalación de un datafono en sus instalaciones.
- Una vez validados los resultados con la anterior prueba, se espera realizar el paso a producción, teniendo seguimientos de resultados semanales y reentrenamientos cada tres meses, empleando las métricas descritas en este trabajo.

VIII. CONCLUSIONES.

El presente trabajo pretende dar solución a una necesidad especifica en la operación de CredibanCo, permitiendo ser más proactivo al momento de identificar un potencial comercio que puede llegar a retractarse del servicio solicitado, aprovechando las nuevas tecnologías y la capacidad de predicción del algoritmo de Bosque Aleatorios en función de variaciones en sus parámetros fundaménteles, además de su comparación frente a los resultados obtenidos empleado como competidores, la Regresión Logística y los Árboles de Decisión.

Así mismo se identificó que el algoritmo de Bosque aleatorios no requiere altos sacrificios de los datos de entrenamiento pues puede manejar variables de tipo numérico y categórico, además permite calibrar de forma dinámica dichos parámetros a fin de confirmar que tiene un potencial adecuado de generalización de predicciones ante datos nuevos, sin embargo, es complicada su interpretación y requiere alto nivel de procesamiento a mayor cantidad de datos.

La disponibilidad de recursos es un punto clave al momento de realizar las calibraciones en los hiperparámetros, este proyecto servirá como otro referente en la organización respecto a la importancia que tienen las ciencias de datos para la optimización de procesos mediante algoritmos de aprendizaje automático, beneficiando tanto actividades operativas como en reducción de costos asociados a la operación.

Actualmente en el mercado existen muchas técnicas, leguajes y programas que permiten el desarrollo del aprendizaje automático, sin embargo muchos de estos programas o software son licenciados, ocasionando que las empresas no están dispuestas a invertir en estas herramientas, ya que se debería sumar este costo al presupuesto anual de la operación llegando incluso a afectar los cierres de P&G del área, por tal motivo esta desde nuestras manos dar a conocer estas herramientas gratuitas y darles el uso enfocado a una necesidad en nuestros negocios o empresas, sin llegar a incurrir en costos adicionales.

Se espera que mediante la implementación de este modelo, se llegue a identificar al 70% de los comercios que se retracten del servicio al momento de la instalación del datafono, esto representaría un ahorro anual de alrededor de 140 millones de pesos, adicionalmente se incrementaría la participación en el mercado y seria un instrumento mas para alcanzar dos de los objetivos estratégicos de la compañía que son: llegar al 2022 con 700 mil comercios con terminales de credibanco e incrementar los niveles transaccionales.

REFERENCES

[1] Castro, Milton Felipe PROAÑO and Contreras, Shirley Yésica ORELLANA and Pazmiño, Italo Omar MARTILLO, (2018, May 07). *Los sistemas de información y su importancia en la transformación digital de la empresa actual.*

[2] Departamento de Seguimiento a la Infraestructura Financiera & Subgerencia Monetaria y de Inversiones Internacionales Banco de la Republica, (2020, Junio),ISSN - 2215 - 9363. Online Available: https://repositorio.banrep.gov.co/bitstream/handle/20.500.12134/9876/rept_sist_pag_2020?sequence=1&isAllowed=y

[3] CredibanCo Online Available: https://www.credibanco.com/sobre-credibanco

[4] Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.

[5] SAMMUT,C. y WEBB, G.I. Encyclopedia of machine learning. Springer Science and Business Media. 2011.

[6] ML Tom M. Mitchell. "Machine Learning"(1997). McGraw-Hill

[7] Branco, A. ¿Qué es la Inteligencia Artificial y cuáles son sus diferentes tipos? Omicrono El Español, (27 de Octubre de 2018).Online Available: https://www.elespanol.com/omicrono/tecnologia/20181027/inteligencia-artificial-diferentes-tipos/348715969_0.html

[8] Sebastian Raschka and Vahid Mirjalili *Python Machine Learning (2019),23-325*

[9] Ali, J., Khan, R., Ahmad, N., y Maqsood, I. (2012). Random forests and decision trees. IJCSI International Journal of Computer Science Issues, 9(5), 272-278.

[10] Breiman, L., Friedman, J., Stone, C., y Olshen, R. Classification and regression trees. California, Estados Unidos: Wadsworth, Inc (1984).

[11] C. Burges B. Schölkopf and A. Smola.*Advances in kernel methods: Support vector machines.Cambridge, MA: MIT Press, 1999.*

[12] C. Burges. A *tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, vol. 2, no. 2, 1998.*

[13] Logistic Regression - scikit-learn 0.23.2 Available: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
[14] Santiago F. *REGRESIÓN LOGÍSTICA - Facultad de Ciencias Económicas y Empresariales UAM* 2011
[15] W. H Greene. Econometric Analysis. Prentice Hall. Quinte Edición. 2010.
[16] Breiman, L., Friedman, J., Stone, C., y Olshen, R. (1984). Classification and regression trees. California, Estados Unidos: Wadsworth, Inc.
[17] Touw, W., Bayjanov, J., Overmars, L., Backus, L., Boekhorst, J., Wels, M. and van Hijum, S. (2012). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Briefings in Bioinformatics, 14(3), pp.315-326.
[18] Árboles de Decisión y Random Forest. bookdown.org, PBC 2016 - 2020. Online Available: https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html]
[19] ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA ANÁLISIS Y PREDICCIÓN DE DATOS. REVISTA TECNOLÓGICA N° 11. ENERO - DICIEMBRE 2018. Online Available: http://www.redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf
[20] Geoffrey Hinton, Terrence J. Sejnowski. 1999 Unsupervised Learning and Map Formation Foundations of Neural Computation, MIT Press, ISBN 026258168X
[21] Nagesh S. Métricas De Evaluación De Modelos En El Aprendizaje Automático. 2020
[22] IBM. Análisis del rendimiento del modelo de aprendizaje automático. 2020. Online Available: https://cloud.ibm.com/docs/watson-knowledge-studio?topic=watson-knowledge-studio-evaluate-ml&locale=es
[23] P. Flash y M.Kull. Precision-Recall-Gain Curves: PR Analysis Dine Right. 2014.
[24] Hand, D.J., and Till, R.J. (2001). A simple generalization of the area under the ROC curve to multiple class classification problems. Machine Learning, 45, 171-186
[25] Hanley J.A., McNeil B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 143: 29-36
[26] Microsoft. Evaluación del rendimiento de un modelo en Machine Learning. Online Available: https://docs.microsoft.com/es-es/azure/machine-learning/studio/evaluate-model-performance.2017.
[27] IBM, Aplicación de machine learning a las necesidades empresariales. Online Available: https://www.ibm.com/ar-es/analytics/machine-learning?
[28] Qué es el machine learning y cómo se aplica en la logística. Online Available: https://blog.pulpomatic.com/blog/aplicaciones-del-machine-learning-en-la-log
[29] WORLD ENERGY TRADE. 2020. Online Available: https://www.worldenergytrade.com/oil-gas/investigacion
[30] Elsevier. C. Qué es el machine learning. 2018. Online Available: https://www.elsevier.com/es-es/connect/ehealth/que-es-el-machine-learning-salud-digital
[31] Dan Olley, Global EVP and CTO of Elsevier, the science, technology and healthcare division of RELX, since 2013.
[32] Aprendizaje automático: Qué es y por qué es importante. Sas.com. Online Available: https://www.sas.com/es_co/insights/analytics/machine-learning.html
[33] Big Data y Banca: el valor del Machine Learning. Cleverdata. Online Available: https://cleverdata.io/big-data-banca
[34] Huérfano Lenis, C. (2020). Identificación predictiva de fallos POS A través del uso de algoritmos de aprendizaje automático.
[35] Géron, A. Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2017
[36] Phi K Correlation Version: 0.9.12. Released: May 2020. Online Available: https://phik.readthedocs.io/en/latest/