Pronóstico de abandono de un cliente de telefonía celular prepago mediante la aplicación de técnicas de Machine Learning

Presenta:

Yuri Tatiana Quevedo Rodríguez

Hernando Miguel Sang López

Director:

Jesús Antonio Villarraga Palomino

Bogotá, Colombia, junio de 2022

Tabla de contenido

RESUMEN 2

ABSTRACT 4

INTRODUCCIÓN 5

ESTADO DEL ARTE 7

METODOLOGÍA 11

Fase 1. Análisis Exploratorio de los datos 13

Fase 2. Selección de herramientas para los datos y el Modelo de pronóstico 14

Fase 3. Procesamiento de los datos 15

Fase 4 Análisis Descriptivo 18

Fase 5. Implementación de los modelos 21

CONCLUSIONES 37

REFERENCIAS BIBLIOGRÁFICAS 39

RESUMEN

En la actualidad, existen múltiples compañías que ofrecen servicios de telefonía celular con ofertas y paquetes similares en el mercado, lo cual hace que el movimiento de usuarios entre estas sea muy frecuente, generando abandonos masivos en función de la libre competencia. En este sentido, lo que se busca con el presente trabajo es determinar el comportamiento asociado a los clientes que abandonan la compañía anónima de telefonía prepagada y predecir a través del uso de técnicas de Machine Learning la similitud en el comportamiento de los usuarios activos e inactivos y clasificarlos dentro de un perfil de riesgo de abandono para años consecutivos. El análisis se hizo a través de diferentes técnicas de Machine Learning "Bosques Aleatorios, árboles de decisión, regresión logística, soporte máquinas vectoriales tipo polinómico y soporte máquinas vectoriales tipo rbf", que permite caracterizar los usuarios activos dentro de una correlación con los usuarios que han abandonado y a través de los modelos mencionados pronosticar el riesgo de abandono para diferentes usuarios.

Los resultados obtenidos muestran que los cinco modelos que han sido evaluados y probados bajo las diferentes métricas tienen valores dentro de un rango de variación en eficiencia de 0.77 a 0,89, demostrando que se tiene modelos de tipo discriminador que cumplen con el objetivo de la investigación. De acuerdo con la curva ROC comparativa de todos los modelos, si bien los bosques aleatorios y arboles decisión muestra mayor área bajo la curva indicando que estos tienen el mejor rendimiento de los cuales el mejor es el modelo de bosques aleatorios; los modelos de regresión nos permiten darle un manejo de incertidumbre mayor a las predicciones y de esta forma considerar una proporción adicional de resultados basado en la probabilidad. De acuerdo a lo anterior, el modelo más representativo de los regresores es el de máquina soporte

vectorial con un kernel de tipo rbf (Gausiano), ya que este tiene una curva ROC con mayor rendimiento.

Palabras clave: Machine Learning, bosques aleatorios, árbol de decisión, pronóstico de abandono, modelo logístico, soporte máquinas vectoriales, rbf, polinomial.

ABSTRACT

Currently, there are multiple companies that offer cell phone services with similar offers and packages in the market, which makes the movement of users between them very frequent, generating massive dropouts due to free competition. In this sense, what is sought with this work is to determine the behavior associated with customers who leave the anonymous prepaid telephone company and predict, through the use of Machine Learning techniques, the similarity in the behavior of active and inactive users. and classify them within a dropout risk profile for consecutive years. The analysis was done through different Machine Learning techniques "Random Forests, decision trees, logistic regression, polynomial type vector machine support and rbf type vector machine support", which allows characterizing the active users within a correlation with the users who have abandoned and through the mentioned models predict the risk of abandonment for different users.

The results obtained show that the five models that have been evaluated and tested under the different metrics have values within a range of variation in efficiency from 0.77 to 0.89, demonstrating that there are discriminator-type models that meet the objective of the research. According to the comparative ROC curve of all the models, although random forests and decision trees show a greater area under the curve, indicating that these have the best

performance, of which the best is the random forest model; regression models allow us to give greater uncertainty management to the predictions and thus consider an additional proportion of results based on probability. According to the above, the most representative model of the regressors is the support vector machine with an rbf (Gaussian) kernel, since it has a ROC curve with higher performance.

INTRODUCCIÓN

Los teléfonos celulares son hoy en día fuentes no solo de comunicación si no que dan alcance a diferentes opciones de desarrollo personal y empresarial, transformando la forma en que nos comunicamos. De esto han nacido diferentes compañías que se esfuerzan en brindar paquetes competitivos en el mercado con mejores y mayores opciones en voz y datos para consumo masivo.

La industria telefónica ha estudiado en diferentes ámbitos científicos un parámetro crucial para la gestión de sus clientes; la tasa de abandono y retención de clientes. Enfocado en este estudio y como alcance final, se desarrollará la tasa de abandono de clientes.

Diferentes estudios y artículos científicos se han realizado buscando clasificar clientes dentro de una probabilidad de que abandonen la empresa donde se encuentran los servicios de telefonía que están consumiendo. Artículos de investigación como la "Predicción de abandono de clientes en una empresa de telecomunicaciones" (Alegre, Repositorio Institucional de la UCM, 2020) y otros que podrán encontrar en este documento, utilizan técnicas de ciencia de datos para caracterizar y analizar los comportamientos de usuarios inactivos y activos para clasificar los usuarios activos dentro de una probabilidad de abandono. Aquí, es donde toma un papel fundamental las técnicas de Machine Learning y Deep Learning que se han venido madurando y aplicando para este tipo de análisis logrando resultados cada vez mejores que permiten caracterizar usuarios a través del riesgo y tomar acciones para su retención.

En el contexto empresarial, el "abandono" – churn rate en inglés- hace referencia tanto a la migración de un cliente como a la pérdida de valor de este; es decir, el "abandono" se refiere por un lado al porcentaje de clientes que ponen fin a su relación con la compañía, y por otro lado

a los clientes que siguen percibiendo sus servicios, pero en menor cantidad o en menor frecuencia.

Las compañías actuales enfrentan un desafío: buscar herramientas que les permita anticiparse al abandono de sus clientes y así evitar retiros masivos que representan grandes pérdidas económicas; aquí es donde se relacionan los objetivos del estudio, que son el detectar de forma temprana qué clientes pueden estar en riesgo de abandono, conocerlos en el ámbito del servicio y responder de acuerdo con quiénes son y cómo es su comportamiento. Por consiguiente, nace la idea de modelar un pronóstico de abandono que permita predecir de una cantidad de usuarios activos al cierre de un año, cuales podrían encontrarse en una condición de riesgo para poder tomar acciones que permitan su retención en un futuro.

Entendiendo esto, diferentes compañías en la industria de la telefonía celular han maximizado sus esfuerzos para adquirir nuevos usuarios y mantener los usuarios existentes con buenos comportamientos. Este estudio busca cómo en diferentes industrias reducir la pérdida de dinero a través de la clasificación de usuarios que pudieran tener un riesgo, ya que en términos de rendimiento de las compañías la tasa de abandono es un indicador que podría medir la insatisfacción general de clientes y las oportunidades que estas brindan.

Este Artículo busca proporcionar resultados de aplicación de técnicas de Machine Learning orientada a la clasificación de usuarios activos con probabilidad de convertirse en usuarios inactivos, teniendo en cuenta una base de datos con contenido específico tanto desde el contexto geográfico como de comportamiento que alimentarán los modelos y permitirán realizar este tipo de clasificación.

ESTADO DEL ARTE

La predicción de abandono de clientes constituye en la actualidad un aspecto preponderante y de gran importancia para las compañías de telecomunicaciones, quienes a través de la ciencia de datos buscan identificar los comportamientos de sus clientes.

Es por ello, que se ha acudido a la revisión documental de diversas fuentes con el ánimo de comprender los antecedentes que puedan dar soporte al presente estudio, partiendo de una búsqueda selectiva de artículos que permiten comprender cómo a partir de la aplicabilidad de diferentes técnicas se puede llegar a predecir el abandono de clientes en diferentes sectores.

Teniendo como punto de partida desde un contexto internacional, un estudio realizado por (Montero, 2018) en Madrid, desarrolla a través de modelos de Redes Neuronales Artificiales, la predicción del abandono de clientes en una empresa de gran consumo con datos reales de las ventas; aplicando las técnicas: como modelo MCE, LVQ y clasificación k-NN, donde los mejores resultados fueron aquellos obtenidos con el modelo MCE relacionados con sensibilidad y eficacia.

De otra parte, (Poyatos, 2021) frente a la predicción del abandono de clientes en una Empresa de Telecomunicaciones (TELCO), mediante el uso de algoritmos de Machine Learning utiliza las técnicas de regresión Logística, Árboles de Decisión, Random Forest y h2o, concluyendo que el modelo con mayor scoring a igualdad de AUC, es el de Regresión Logística.

Desde otra perspectiva, en América Latina la mayoría de estudios encontrados, aplicaron la técnica de Regresión logística seguida por random forest; por ejemplo, en Chile (Jélvez Caamaño, 2014) utiliza la técnica de regresión logística multivariable y análisis de clúster con el método de las K-medias para crear un modelo de fuga de servicios en dos mercados de una

compañía de telecomunicaciones; donde el modelo de regresión logística multivariable clasifica en forma correcta el 71,2% y 72,1% en los dos mercados de los casos analizados.

También, (Beltran Cruz, 2019) para una compañía de telecomunicaciones y (Bohórquez, 2020) en el sector financiero, incluyen en sus técnicas de modelación la regresión Logística; sin embargo, los resultados no obtienen los mejores rendimientos, dado que frente al primer caso la metodología Deep Learning obtuvo mejores resultados en la práctica y respecto a la compañía financiera, de acuerdo a los resultados del área bajo la Curva ROC, el bosque aleatorio obtiene un 93% superando los otros modelos.

Se hace importante mencionar algunas investigaciones en Colombia, uno de ellos desarrollado por (Blanco, 2018) quien buscó desarrollar un modelo estadístico que aplica técnicas de aprendizaje de máquina y que permitiera identificar los clientes propensos a presentar fuga, a través de la recopilación de datos históricos; se implementaron tres tipos de modelos de clasificación : la regresión logística, redes neuronales y la técnica SVM (máquinas de soporte vectorial), concluyendo que el SVM es quien ofrece una tasa de error más baja siendo escogido como modelo clasificador adecuado.

Así mismo, de la Universidad ECCI en el trabajo de investigación titulado “Diseño de un Modelo Predictivo de Fuga de Clientes Utilizando Algoritmos de Machine Learning” desarrollado por (Pinto, 2020), utilizaron los algoritmos GBM (Gradient Boosting Machine) y Random Forest; para evaluar los rendimientos de los modelos utilizaron el criterio bajo la curva de AUC y la técnica de clasificación Random Forest fue quien arrojó los mejores resultados.

Finalmente, se evidencia la utilización de la técnica XG Boost por (Falla, 2021) en el trabajo denominado predicción de abandono de clientes en telecomunicaciones mediante el

aprendizaje automático y por (Echeverri, 2019) en su documento llamado Modelo predictivo de Churn de clientes para el negocio de Telecomunicaciones, donde como conclusión en ambos casos se aprecia que esta ofrece alta adaptabilidad, dada su robustez y capacidad para trabajar con datos desbalanceados.

Luego de haber desarrollado la revisión de diversas fuentes de información, las cuales permiten evidenciar el uso de diferentes técnicas de clasificación y regresión, se concluye que las más usadas son la Regresión Logística y Random Forest, dado que estas tienen las mejores respuestas a nivel de resultados obtenidos.

Dado lo anterior, para efectos del desarrollo de esta investigación se realizará una evaluación basada en un alcance de predicción dicotómica con 1 y 0 (activos e inactivos), a través de las siguientes técnicas de machine learning: Bosques Aleatorios, árboles de decisión, regresión logística, soporte máquinas vectoriales tipo polinómico y soporte máquinas vectoriales tipo Gausiano.

Un modelo *Random Forest* está formado por un conjunto de árboles de regresión individuales, cada uno ajustado empleando una muestra bootstrapping de los datos de entrenamiento. Una vez entrenado, la predicción de una nueva observación se obtiene promediando las predicciones de todos los árboles individuales que forman el modelo. (Rodrigo, 2020)

La técnica de regresión logística es un método estadístico, de machine learning que se usa para resolver problemas de clasificación binaria, donde el resultado solo puede ser de naturaleza dicotómica, o sea, solo puede tomar dos valores posibles; por ejemplo, se puede utilizar para detectar la probabilidad que ocurra un evento (Buitrago, 2020); Los árboles de

decisión son modelos predictivos formados por reglas binarias (si/no) con las que se consigue repartir las observaciones en función de sus atributos y predecir así el valor de la variable respuesta. (Rodrigo J. A., 2020).

SVM (máquina de vectores de soporte) es una técnica de clasificación y regresión que aprovecha al máximo la precisión de las predicciones de un modelo sin ajustar excesivamente los datos de entrenamiento. SVM es ideal para analizar datos con un gran número de campos predictores. (IBM, 2021)

Un árbol de decisión en Machine Learning es una estructura de árbol similar a un diagrama de flujo donde un nodo interno representa una característica (o atributo), la rama representa una regla de decisión y cada nodo hoja representa el resultado. Esta estructura tipo diagrama de flujo lo ayuda a tomar decisiones. Es una visualización como un diagrama de flujo que imita fácilmente el pensamiento a nivel humano. Es por eso que los árboles de decisión son fáciles de entender e interpretar.

METODOLOGÍA

El presente estudio se fundamenta en una investigación de tipo mixto, porque es de característica más estructurada y de tipo teórico, busca desarrollar todo en un conjunto de procesos que se manejen de manera consecuente y que no se puedan saltar, se basa en objetivos y preguntas de investigación que buscan construir un marco o perspectiva teórica. Lo anterior para probar para evaluar el abandono de clientes de acuerdo con sus patrones de consumo.

Para el desarrollo de la investigación se utilizará una metodología modificada basada en CRISP-DM, donde se tendrá en cuenta un análisis de tipo iterativo , soportado en la teoría del documento (Gil, 2021) que indica está compuesta por seis fases, las cuales dependen entre sí tanto en forma secuencial como cíclica, pudiendo encontrarse interacciones que permitan mejorar la aproximación obtenida en otras fases anteriores; sus fases son: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue, para este estudio solo se desarrollara hasta la fase 5, se culminará con la evaluación de los modelos.

La información para utilizar, son los datos disponibles de una compañía anónima de aproximadamente 9 años donde se obtienen datos de usuarios activos e inactivos a corte abril de 2022 doce meses hacia atrás, con las variables mencionadas en el dominio de los datos, que se consolidan a través de las plataformas digitales de la empresa y se estructuran mediante motores de bases de datos como SQL Server.

Fase 1. Análisis Exploratorio de los datos

Se consolidó la información a través de la herramienta SQL, donde se seleccionaron algunas variables asociadas al tráfico, el consumo de los usuarios y variables de caracterización como el departamento en donde presta el servicio la compañía anónima de telefonía prepago, identificando usuarios activos e inactivos.

Se consideraron variables mensuales que se procesaron para dejar una base de datos con comportamiento anual que pueda representar a cada usuario como un único registro. Y con esta información se trabajará el modelo de clasificación en función del riesgo.

Para el desarrollo del Modelo vamos a utilizar las siguientes variables, con sus respectivas características que se mencionan a continuación:

Subscriberid: Es el código asociado al cliente, variable cualitativa nominal.

Mno_name: Es el nombre de la compañía operadora donde está registrado el cliente, variable cualitativa nominal.

Tenure: Es el tiempo que lleva el usuario en la compañía, variable cuantitativa discreta.

Total_recharges: La recarga del mes hecha por el usuario, variable cuantitativa discreta.

Count_recharge: Cantidad de recargas en el mes, variable cuantitativa discreta.

Trafico_datos: Cantidad de datos consumidos en megas durante el mes, variable cuantitativa discreta.

Trafico_voz_saliente: Cantidad de llamadas salientes medidas en minutos durante el mes, variable cuantitativa discreta.

Trafico_voz_entrante: Cantidad de llamadas recibidas medidas en minutos durante el mes, variable cuantitativa discreta.

Trafico_sms: Cantidad de mensajes de texto realizados durante el mes, variable cuantitativa discreta.

Status_customer: Es el estado del usuario, activo o inactivo y es una variable cuantitativa discreta tipo dicotómica.

Depto_description: El departamento de donde se registró el usuario, variable cualitativa nominal.

Mun_descriptio: El municipio de donde se registró el usuario, variable cualitativa nominal.

Los datos fueron reestructurados, buscando la mejor estrategia para que la información que se pueda extraer de las variables tengan la mejor precisión; en este caso para poder dejar un único Subscriberid con su comportamiento correspondiente durante un año, se muestrearon usuarios activos e inactivos que cumplieran al menos doce meses de comportamiento desde el último mes reportado hacia atrás, buscando que los usuarios tengan el mismo peso y que se puedan usar promedios de sus variables sin perder información .

Fase 2. Selección de herramientas para los datos y el Modelo de pronóstico

Para el análisis descriptivo de los datos, se usa la herramienta R studio, donde se trabaja todo el análisis univariado, multivariado y de componentes principales.

Cabe mencionar, que según (R Studio, 2022), R studio es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicado a la computación estadística y gráficos. Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo.

Ahora, para la manipulación de datos y modelo pronóstico se utiliza la herramienta Python, donde se implementa los modelos de tipo clasificación y regresión univariado y multivariada. Los modelos para desarrollar son: bosques aleatorios, árboles de decisión, regresión logística, soporte máquinas vectoriales tipo polinómico y soporte máquinas vectoriales tipo gaussiano.

Python es un lenguaje de programación multipropósito, que en la minería de datos o data mining es un proceso que permite analizar grandes bases de datos con el objetivo de predecir futuras tendencias. Se trata de un proceso complejo al que Python puede arrojar luz a través de la limpieza y organización de datos y del uso de algoritmos de aprendizaje automático que simplifica el análisis de datos. (Banco Santander S.A., 2021).

Fase 3. Procesamiento de los datos

Para el desarrollo de la investigación se cuenta inicialmente con 8’793.810 millones de registros y 15 variables, las cuales caracterizan el comportamiento de los clientes activos e inactivos con relación al consumo y la ubicación geográfica. Después se estructura base de datos en relación a los usuarios con movimiento igual a 12 meses, seguidamente se identificó valores faltantes o nulos y se agruparon de acuerdo a las características de la variable categórica “depto_description”, para mantener la mayor cantidad de información para un único usuario se

define utilizar el promedio de los valores en las variables de consumo proporcionadas, buscando la mejor caracterización y en las variables categóricas se utilizó la moda, dado lo anterior la base de datos queda conformada con un registro de 634.289 observaciones y 11 variables.

Una vez procesada la información, a través de la variable “status_customer” se identifica la cantidad de usuarios activos e inactivos, donde se observa un desbalance con una cifra de 633.162 usuarios activos y 1.127 inactivos (0.998% vs 0.002%), por lo que decidimos aplicar un balanceo estratificados interno de los algoritmos de machine learning , donde se observa que las predicciones tienen una tendencia preferencial hacia resultados positivo-positivo.

Teniendo en cuenta los resultados anteriores para lograr balancear los datos proporcionados para el modelado se decidió aplicar una técnica de muestreo estratificado proporcional, que busca considerar todos los grupos de interés en función de la variable categórica “depto_description”.

Para el desarrollo del muestreo estratificado se aplica la fórmula a continuación: se calcula la proporción de usuarios por cada departamento como users*depto = número de usuarios aleatorios en el departamento específico y usuarios inactivos es la cantidad de usuarios con la cual se hace balanceo de la variable respuesta de usuarios activos

$$n_{users} * depto = \left(\frac{\sum user_{id} * depto_{description}}{\sum_t depto_{description}}\right)\% * n_usuarios_inactivos(cantidad)$$

$\sum user_{id} * depto_{description}$: número de usuarios * departamento

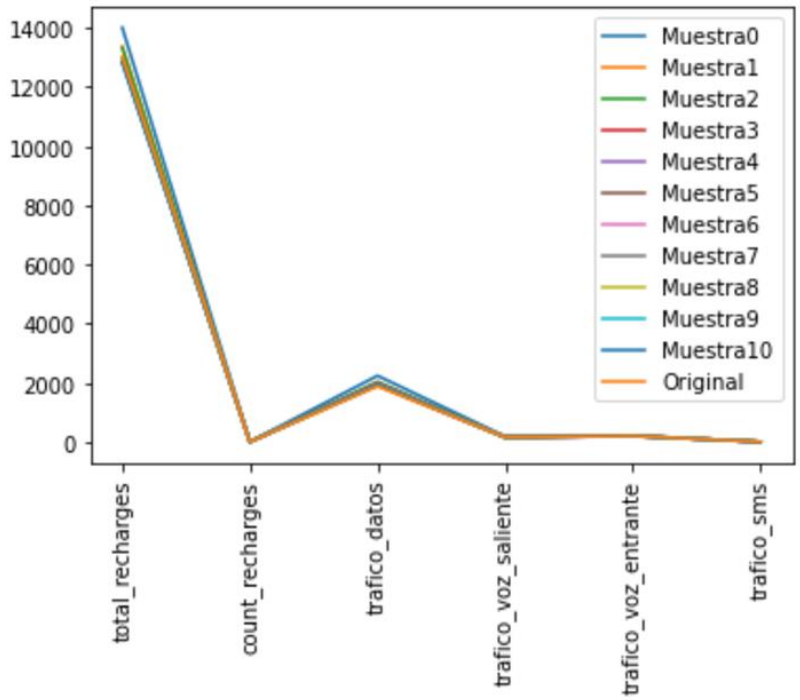
$\sum_t depto_{description}$: número total de usuarios de todos los departamentos

$n_usuarios_inactivos$: cantidad de usuarios inactivos.

Para efectos de este trabajo según (Roldán, 2019), se comprende que “el muestreo estratificado es una técnica o procedimiento en la que se divide la población bajo estudio en distintos subgrupos o estratos”. La muestra final después de aplicar la técnica constituye un dataset de 2247 registros y 10 variables. Para probar que aleatoriedad aplicada a diferentes muestras logran mantener medidas promedio en cada variable, se generaron 10 muestras aleatorias más la población original y a través de un gráfico de promedios se puede observar que la huella de todas las muestras incluyendo la original son similares, así como se ilustra en la Figura 1.

Figura 1.

Comportamiento de las 10 muestras y la población.



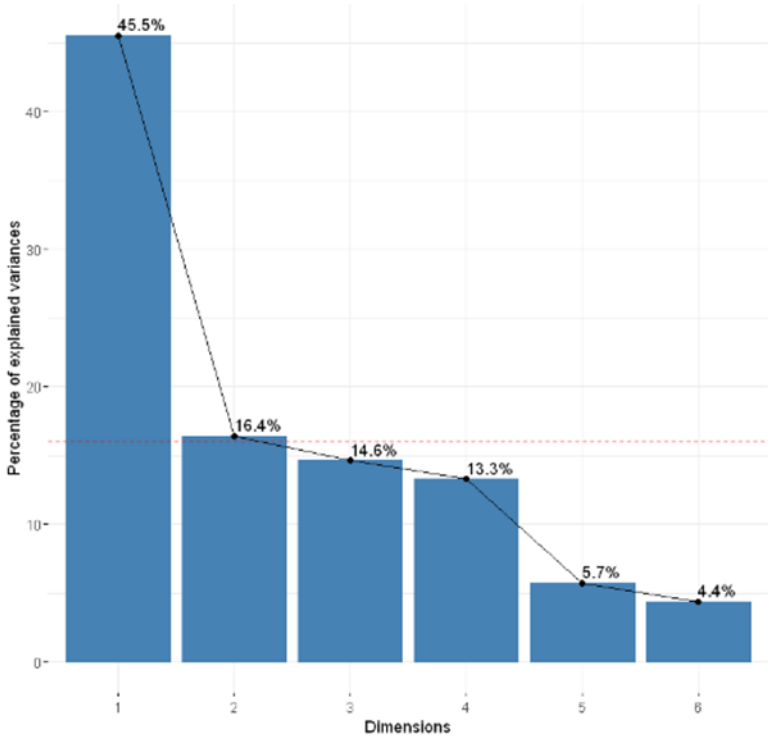
Fuente: elaboración propia

Fase 4 Análisis Descriptivo

Se realizó la matriz de varianza y covarianza para las variables a utilizar en el modelo, el determinante de la matriz es de 1,2717*e+25, valor positivo lo cual indica que se puede proceder con el análisis multivariado. Seguidamente se aplica la prueba de normalidad Jarque Bera, donde el p_value: nos arroja 0,029 > 0,001 lo cual indica que existe normalidad el comportamiento de los datos.

Figura 2.

Varianza explicada de las variables con 6 dimensiones.



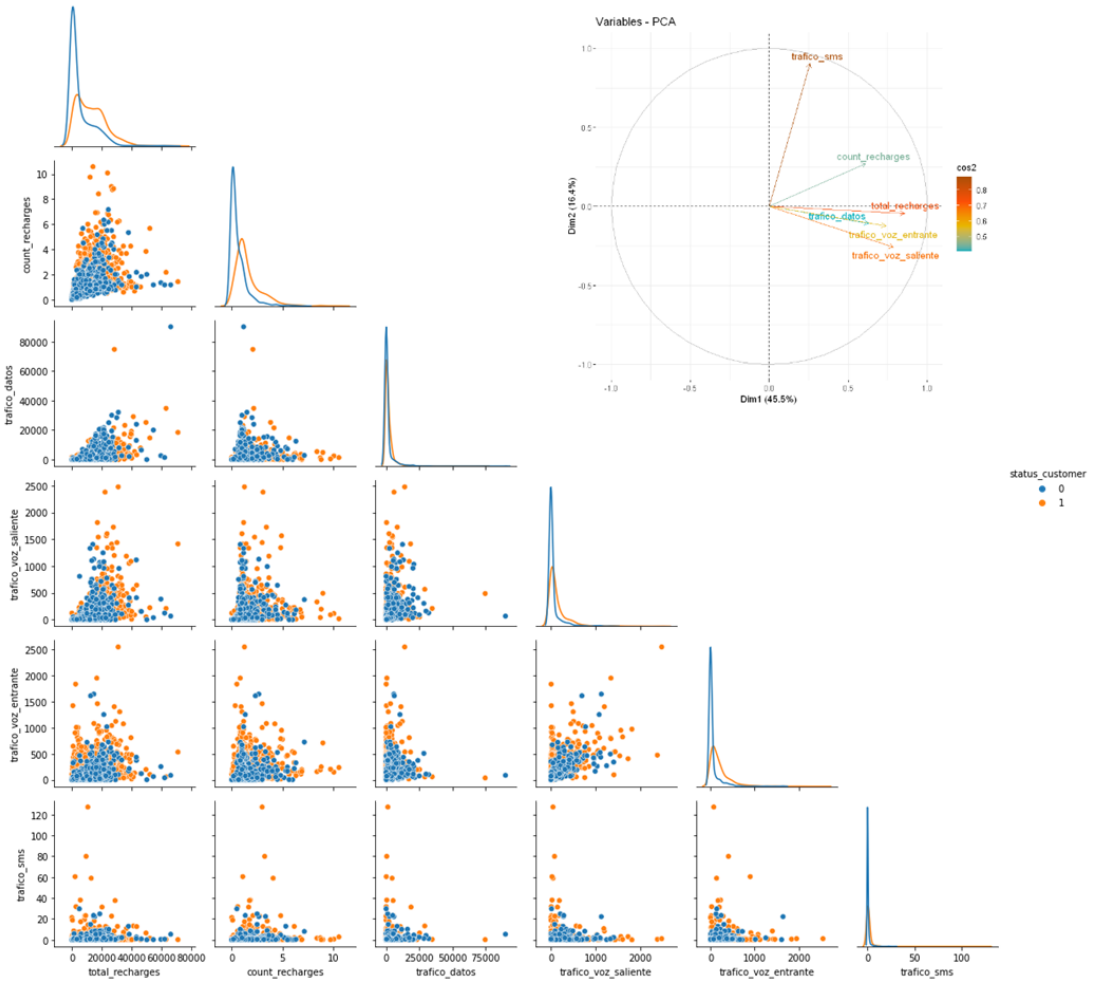
Fuente: elaboración propia

La figura 2 ilustra el análisis de ACP, donde se evidencia que entre las dimensiones 1 y 2 la varianza explicada es del 61%, seguidamente del 75% por la tercera, 89% la cuarta, 95% la cuarta y su totalidad con 6 dimensiones.

Al realizar el análisis bivariado de las variables cuantitativas en función a la variable “status_customer”, y en comparación con los biplots realizados en el análisis multivariado PCA, se pueden identificar las correlaciones entre las variables de la siguiente manera:

Figura 3.

Análisis bivariado y biplot de representación en las dimensiones 1 y 2.



Fuente: elaboración propia

La Figura.3 ilustra que las variables “trafico_voz_entrante” y “trafico_voz_saliente” son las que tienen mayor representación y contribución siendo altamente correlacionadas soportadas con el análisis multivariado en las dimensiones 1 y 2, igualmente las variables restantes tienen algo de correlación y en su interpretación “total_recharge” y “count_recharge” son seguidamente las que tienen mayor representación y contribución de este grupo.

Fase 5. Implementación de los modelos

Para la implementación de los modelos se utilizaron todas las variables del dataset final, listadas a continuación: “trafico_voz_entrante, trafico_voz, trafico_datos, count_recharge, status_customer, trafico_sms y total_recharge”.

De acuerdo con la exploración de diferentes artículos mencionados en este documento asociados a este tipo de análisis, donde se busca detectar clientes con riesgo de abandono, se identificaron algunos modelos que serán aplicados y otros que no están en los documentos. Los modelos que veremos a continuación son cinco:

- **Bosques Aleatorios**

Figura 4.

Bosques Aleatorios 1 – Predictor Abandono de Clientes.

Fuente: elaboración propia

Modelo a ejecutar utilizando la librería SKLEARN en Python; se trabajó con datos entrenamiento de 70% y datos prueba del 30%, y bajo el criterio de entropía, tiene por objetivo encontrar los nodos más puros y homogéneos posibles. (Amat, 2017). Su resultado en las métrica de rendimiento de 0.85 de eficiencia, se realizó una validación cruzada para 10 muestras arrojando los siguientes resultados: 0.95813953, 0.95348837, 0.9627907, 0.94883721, 0.95813953, 0.93488372, 0.9255814, 0.71028037, 0.56542056, 0.63084112, de acuerdo a estos, los primeros resultados son homogéneos alrededor del 0.90, sin embargo algunas anomalías en los resultados siguientes. Por lo anterior se buscó explorar un análisis de error cuadrado medio para validar la estabilización de los errores verificando problemas de Overfitting, ver figura 5.

Figura 5.

Test Overfitting MSE.

Fuente: elaboración propia

En la Figura.4 se ilustra el error cuadrado medio contra el número de bosques, en donde se puede observar a partir de 5 bosques aproximadamente se estabiliza el promedio del error cuadrado, indicando que no existe overfitting.

A continuación, se muestran las métricas del modelo de bosques aleatorios.

Tabla 1.

Métricas del modelo bosques aleatorios.

	precision	recall	f1-score	support
0	0.90	0.83	0.87	332
1	0.84	0.90	0.87	313
accuracy			0.87	645
macro avg	0.87	0.87	0.87	645
weighted avg	0.87	0.87	0.87	645

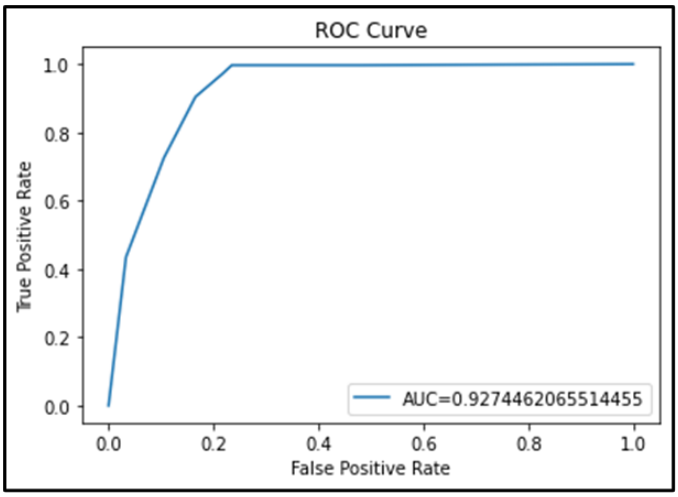
Fuente: elaboración propia

De acuerdo con la Tabla 1, se observa que las métricas proponen una buena precisión en los resultados del conjunto de datos al igual que la accuracy que nos muestra que los resultados del modelo tienen una cercanía del 0.87 respecto al valor real.

Se generó la curva ROC como métrica de puntaje del funcionamiento del modelo, para este caso el AUC (Area Under the Curve) es del 0,93 mostrando un funcionamiento cercano al discriminador ideal, ver figura 6.

Figura 6.

Curva ROC bosques aleatorios



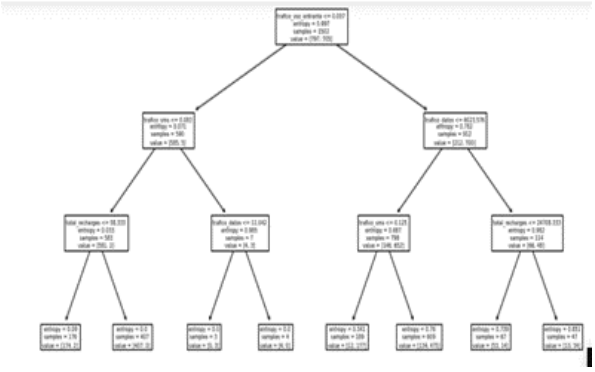
Fuente: elaboración propia

❖ Árbol de decisión

Modelo de clasificación que se ejecuta utilizando la librería SKLEARN en Python; se trabajó con datos entrenamiento de 70% y datos prueba del 30%, y bajo el criterio de entropía.

Figura 7.

Bosque aleatorio - predictor de abandono de clientes.

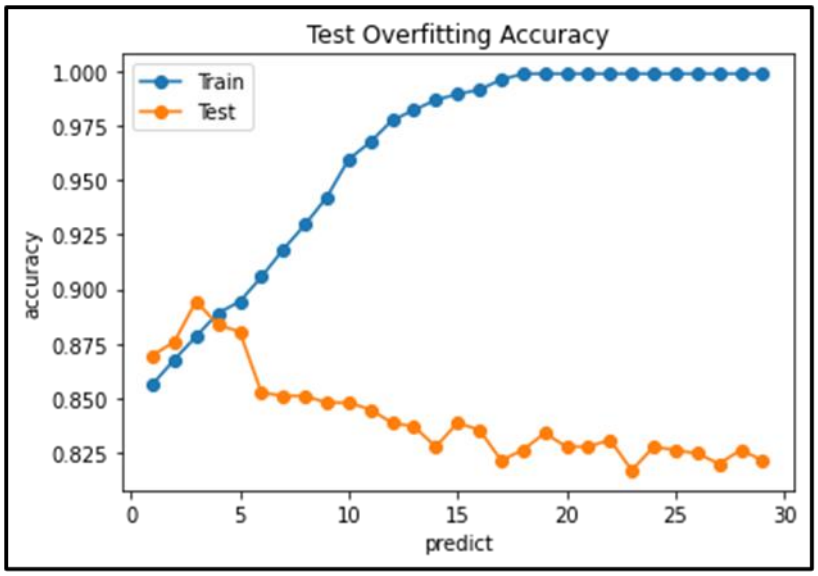


Fuente: elaboración propia

En la figura 8, se grafica un plot de los resultados de la exactitud del entrenamiento y prueba contra el accuracy score para los datos de entrenamiento y prueba a partir del modelo de árbol de decisión.

Figura 8.

Test Overfitting accuracy.



Fuente: elaboración propia

De acuerdo con estos resultados se observa un sobre ajuste únicamente en los tres primeros árboles y a partir del árbol número siete la separación y sobre el árbol 15 se estabilizan, lo anterior indica que no existe overfitting.

Al observar la Tabla 2 a continuación las métricas proponen una buena precisión en los resultados del conjunto de datos al igual que la accuracy que nos muestra que los resultados del modelo tienen una cercanía del 0.89 respecto al valor real.

Tabla 2.

Métricas de rendimiento modelo árbol de decisión.

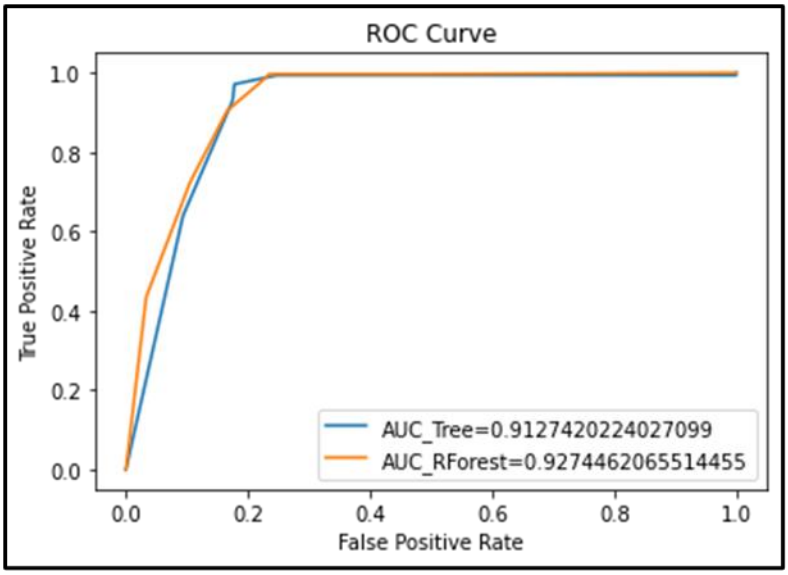
	precision	recall	f1-score	support
0	0.97	0.82	0.89	332
1	0.84	0.97	0.90	313
accuracy			0.89	645
macro avg	0.90	0.90	0.89	645
weighted avg	0.90	0.89	0.89	645

Fuente: elaboración propia

Se generó la curva ROC como métrica de puntaje del funcionamiento del modelo, para este caso el AUC (Area Under the Curve) es del 0,91 mostrando un funcionamiento cercano al discriminador ideal, ver figura 9.

Figura 9.

Curva ROC bosques aleatorios vs árbol de decisión.

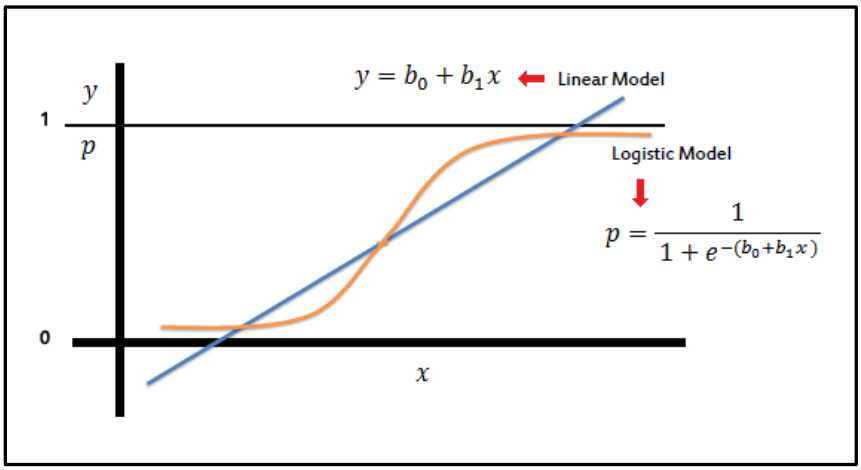


Fuente: elaboración propia

Regresión Logística

Figura 10.

Ejemplo matemático Regresión Logística.



Fuente: https://rubialesalberto.medium.com/regresi%C3%B3n-log%C3%ADstica-con-sklearn-4384c707075d

Modelo de regresión con resultados basado en la probabilidad, que se trabajó con la librería sklearn.linear en Python. se trabajó con datos entrenamiento de 70% y datos prueba del 30%.

De acuerdo con la Tabla 3., observamos que las métricas son menores en relación a los anteriores modelos, el resultado de precisión es de 0.76 para la predicción de los “0” y 0,79 para los “1”, el accuracy que nos muestra que los resultados del modelo tienen una cercanía del 0.77 respecto al valor real;

Tabla 3.

Métricas de rendimiento Modelo Regresión Logística.

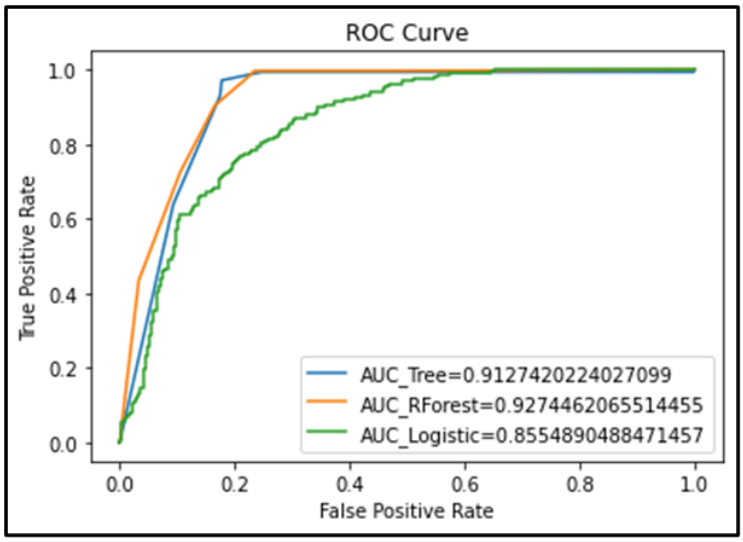
	precision	recall	f1-score	support
0	0.76	0.82	0.79	332
1	0.79	0.72	0.75	313
accuracy			0.77	645
macro avg	0.77	0.77	0.77	645
weighted avg	0.77	0.77	0.77	645

Fuente: elaboración propia

Se generó la curva ROC como métrica de puntaje del funcionamiento del modelo, para este caso el AUC (Area Under the Curve) es del 0,85 mostrando un funcionamiento cercano al discriminador ideal.

Figura 11.

Curva ROC de bosques aleatorios, árbol de decisión, regresión logística.

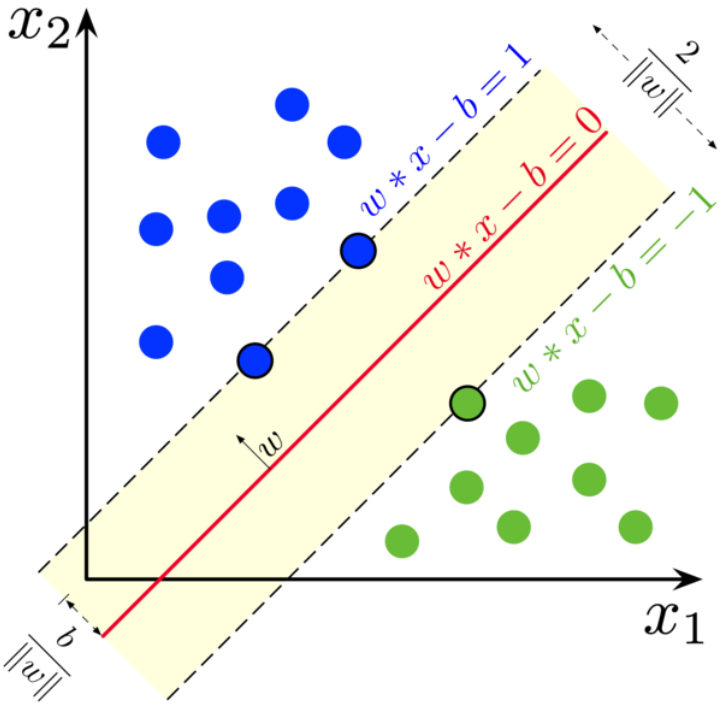


Fuente: elaboración propia

❖ **Soporte máquinas vectoriales núcleos polinómicos**

Figura 12.

Ejemplo Matemático de Support Vector Machine General.



Fuente: https://towardsdatascience.com/so-why-the-heck-are-they-called-support-vector-machines-52fc72c990a1

Modelo de regresión con resultados basado en la probabilidad, se trabajó con la librería sklearn.svm en Python mediante el kernel “Poly”, con datos entrenamiento de 70% y datos prueba del 30%.

De acuerdo con la tabla 4, observamos que las métricas no son muy buenas en relación con los modelos de clasificación, el resultado de precisión es de 0,83 para la predicción de los “0” y 0,81 para los “1”, el accuracy que nos muestra que los resultados del modelo tienen una cercanía del 0.82 respecto al valor real;

Tabla 4.

Métricas de rendimiento modelo svm – polinomial.

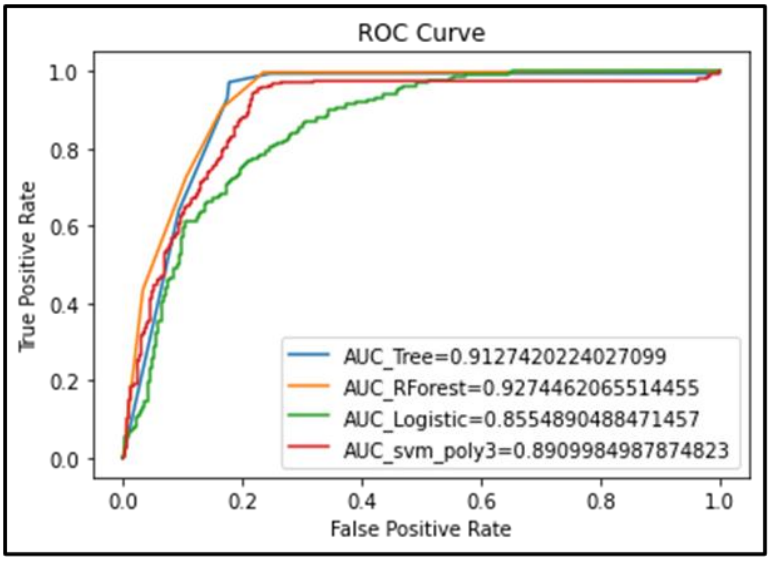
	precision	recall	f1-score	support
0	0.83	0.83	0.83	332
1	0.81	0.81	0.81	313
accuracy			0.82	645
macro avg	0.82	0.82	0.82	645
weighted avg	0.82	0.82	0.82	645

Fuente: elaboración propia

Se generó la curva ROC como métrica de puntaje del funcionamiento del modelo, para este caso el AUC (Area Under the Curve) es del 0,89 mostrando un funcionamiento cercano al discriminador ideal.

Figura 13.

Curva ROC SVM polinomial.



Fuente: elaboración propia

❖ **Soporte máquinas vectoriales a núcleos gaussianos.**

Modelo de regresión con resultados basado en la probabilidad, se trabajó con la librería sklearn.svm en Python mediante el kernel “rbf”, con datos entrenamiento de 70% y datos prueba del 30%.

De acuerdo a la tabla 5, observamos que las métricas son similares al anterior modelo, el resultado de precisión es de 0,83 para la predicción de los “0” y 0,81 para los “1”, el accuracy que nos muestra que los resultados del modelo tienen una cercanía del 0.82 respecto al valor real;

Tabla 5.

Métricas de rendimiento modelo svm- gaussiano.

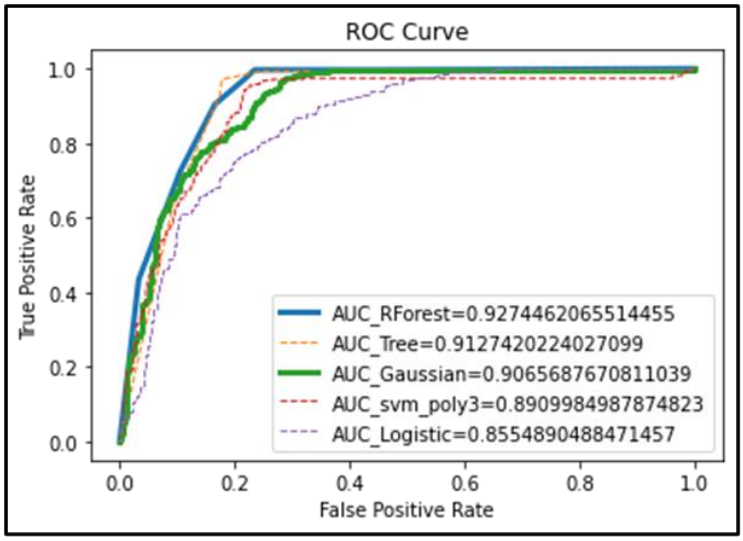
	precision	recall	f1-score	support
0	0.83	0.83	0.83	332
1	0.81	0.81	0.81	313
accuracy			0.82	645
macro avg	0.82	0.82	0.82	645
weighted avg	0.82	0.82	0.82	645

Fuente: elaboración propia

Se generó la curva ROC como métrica de puntaje del funcionamiento del modelo, para este caso el AUC (Area Under the Curve) es del 0,91 mostrando un funcionamiento cercano al discriminador ideal.

Figura 14.

Curva ROC SVM gaussiano.



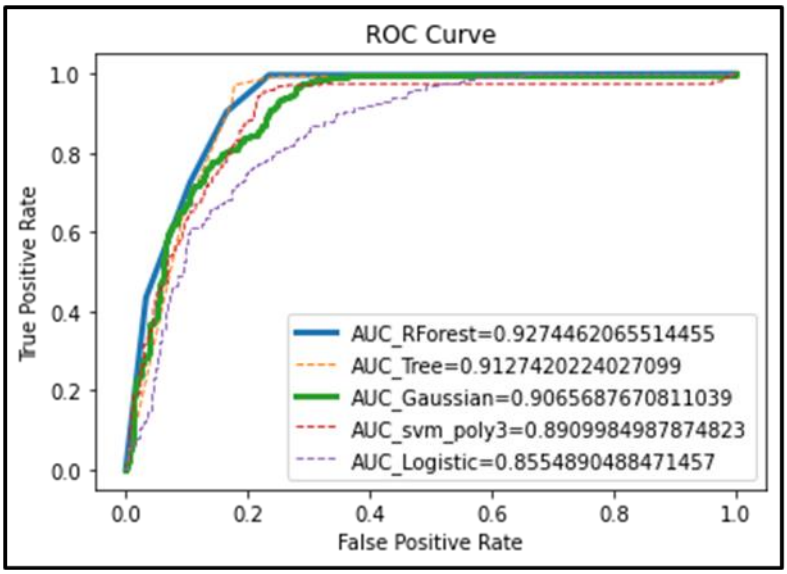
Fuente: elaboración propia

RESULTADOS

A continuación, se presenta la Curva ROC para todos los modelos clasificadores, es de acotar que se sombrean en verde y azul oscuros los modelos con mejores resultados, cabe resaltar que para los entrenamientos de los modelos se usa la muestra balanceada, y en los resultados a continuación se generan las matrices de confusión para el dataset desbalanceado.

Figura 15.

Curva ROC Comparativa.



Fuente: elaboración propia

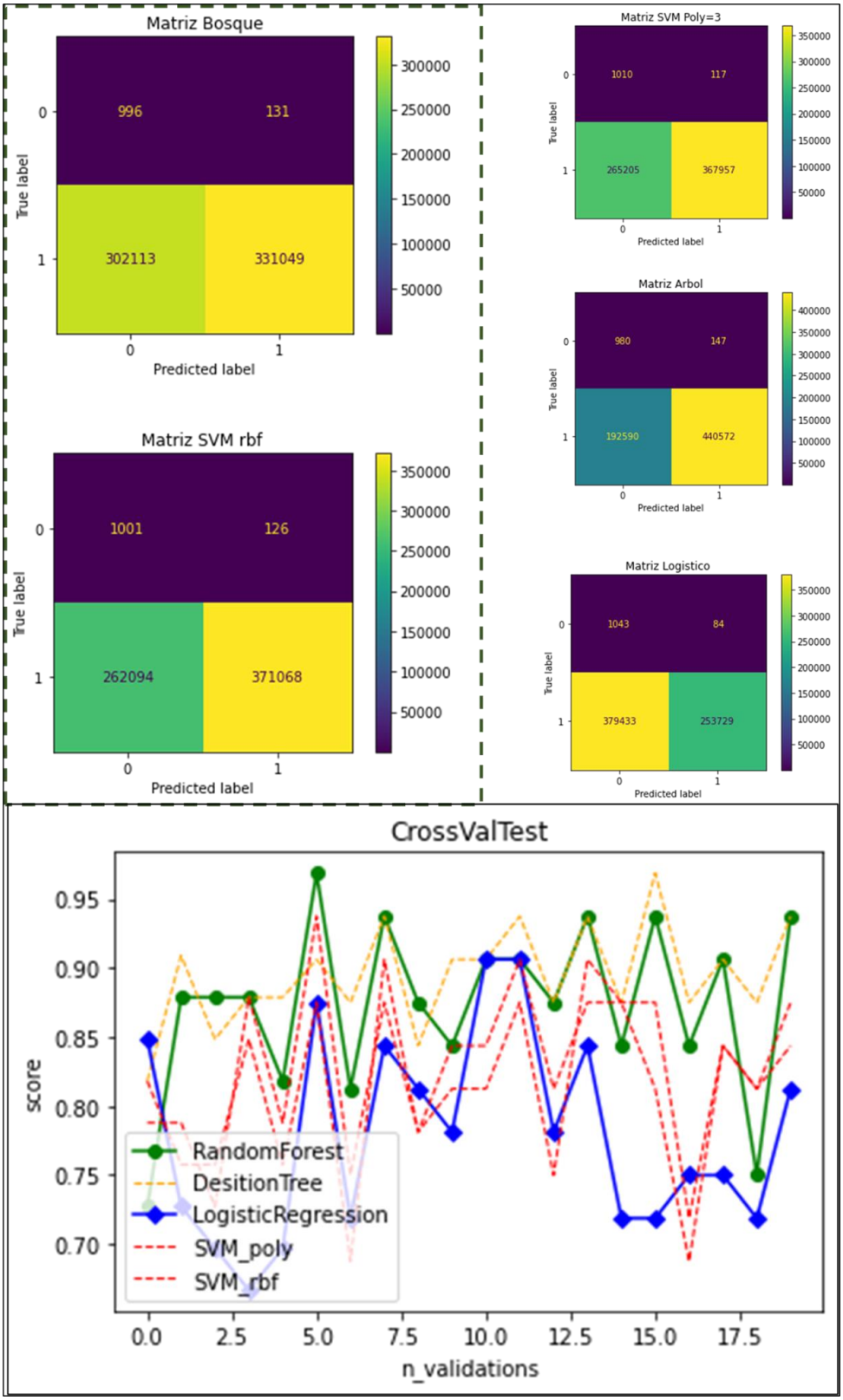
De acuerdo con la ilustración de la figura 11. la curva ROC, y las métricas analizadas el mejor modelo es bosques aleatorios, ya que tiene una mayor exactitud y precisión en la predicción de los usuarios inactivos y es el que presenta una mejor área bajo la curva de acuerdo con la gráfica complementaria.

En segunda posición y dándole prioridad a la misma métrica en exactitud y predicción de los usuarios inactivos; el modelo soporte máquina vectorial “rbf” presenta resultados razonables con una probabilidad más amplia de usuarios que pudieran pronosticarse como de riesgo.

La figura 16. ilustra la matriz de confusión donde se registra las predicciones para cada modelo en el dataset completo, como se puede observar se realizó validación cruzada para todos los modelos indicando que todos presentan un comportamiento estable y aprueban un supuesto de que no existe sobre ajuste, de la misma manera a pesar de que 3 modelos fueran descartados sus métricas son relativamente buenas y pudieran ser objeto de estudio en entrenamientos posteriores.

Figura 16.

Resumen de resultados Matrices cruzadas y Pruebas modelos dataset final.

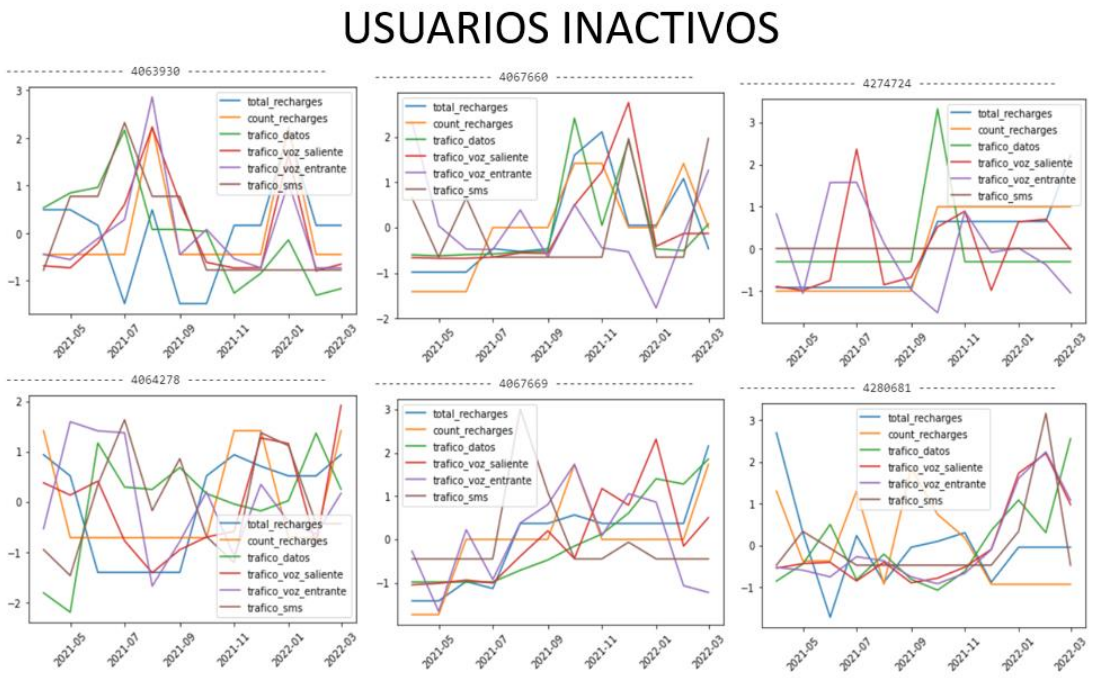


Fuente: elaboración propia

separando los resultados en función de la probabilidad se ven comportamientos buenos de usuarios activos pronosticados como inactivos comparando tendencias con usuarios inactivos. a continuación un set de gráficas con dicha comparación:

Figura 17.

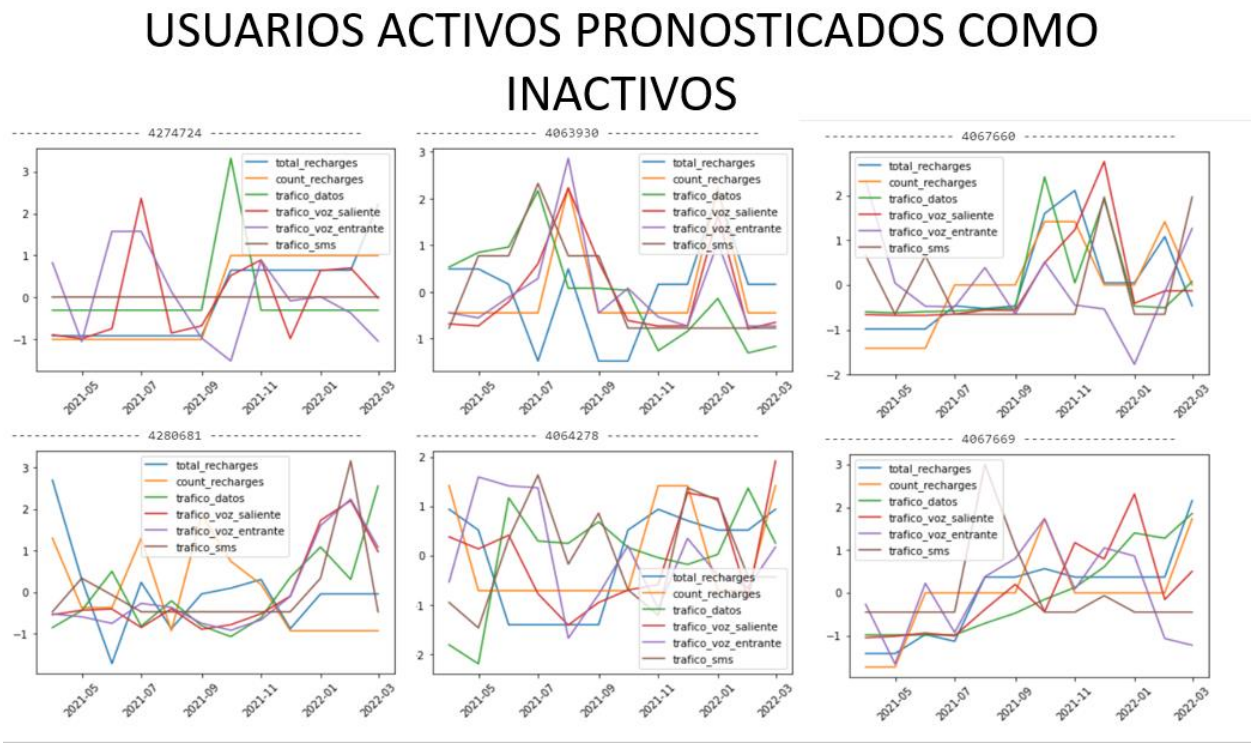
Tendencias set de usuarios aleatorios Inactivos .



Fuente: elaboración propia

Figura 18.

Tendencias set de usuarios aleatorios Activos pronosticados como inactivos con una probabilidad de riesgo > 90%.



Fuente: elaboración propia

CONCLUSIONES

1. Es fundamental realizar en análisis bivariado y multivariado ya que este nos da un contexto de cómo estar correlacionados nuestros datos, su contribución y representación del modelo, facilitando la limpieza y el procesamiento

2. En presencia de desbalance es válido aplicar técnicas de muestreo, estratificación o duplicación con la finalidad de lograr ajustes en los modelos predictivos, teniendo en cuenta que se debe garantizar la menor pérdida de información.

3. Si se realizan muestreos aleatorios, es primordial validar que los comportamientos estadísticos sean capaces de mantenerse para tener la tranquilidad de que la muestra representa el mundo de datos.

4. En general en los modelos de acuerdo a la validación cruzada y métodos utilizados sobre bosques aleatorios y árbol de decisión: “MSE” & “Accuracy vs Score”, se puede decir que las pruebas tienen estabilidad, haciendo evidente la No presencia de sobre ajustes en los Modelos.

5. Teniendo en cuenta la curva ROC y las métricas asociadas a: “AUC, Accuracy, Precisión y otros”. Los modelos más predictivos y con mejor Exactitud son:

- Modelo Bosques Aleatorios
- Modelo SVM – rbf

REFERENCIAS BIBLIOGRÁFICAS

Luis F Restrepo. (5 de Agosto de 2011). Obtenido de Scielo: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-06902012000200011#:~:text=El%20an%C3%A1lisis%20de%20componentes%20principales,variables%20(Pe%C3%B1a%2C%202002).

<retamal, M. E. (Diciembre de 2011). *Repositorio Universidad Bío Bío*. Obtenido de http://repobib.ubiobio.cl/jspui/bitstream/123456789/2308/1/Ovalle_Retamal_Victor_Francisco_Javier.pdf

Alegre, F. (7 de Septiembre de 2020). Obtenido de Repositorio Institucional de la UCM: https://eprints.ucm.es/id/eprint/61986/

Alegre, F. (2020). *Predicción de abandono de clientes en una empresa de telecomunicaciones*. Obtenido de https://eprints.ucm.es/id/eprint/61986/1/TFM-Miner%C3%ADa2020%20Federico-Alegre.pdf

Alegre, F. (7 de Septiembre de 2020). *Repositorio Institucional de la UCM*. Obtenido de https://eprints.ucm.es/id/eprint/61986/

Amat, J. R. (Febrero de 2017). *Ciencia de datos. net*. Obtenido de https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting

Banco Santander S.A. (9 de abril de 2021). *Santander, Becas*. Obtenido de https://www.becas-santander.com/es/blog/python-que-es.html

Beltran Cruz, V. (2019). *Repositorio Institucional UDD*. Obtenido de https://repositorio.udd.cl/handle/11447/2665

Blanco, C. B. (2018). *Repositorio unilibertadores*. Obtenido de https://repository.libertadores.edu.co/bitstream/handle/11371/2077/Baena_Camilo_2018.pdf?sequence=1&isAllowed=y

Bohórquez, M. (Abril de 2020). *Researchgate.net*. Obtenido de https://www.researchgate.net/publication/343186410_MODELOS_DE_PREDICCION_DE_DESERCION_DE_CLIENTES_PARA_UNA_ADMINISTRADORA_DE_FONDOS_ECUATORIANA

Buitrago, B. (2017 de Septiembre de 2020). *Medium.com*. Obtenido de https://medium.com/iwannabedatadriven/regresi%C3%B3n-log%C3%ADstica-i-machine-learning-84ffe9d6be15#:~:text=El%20m%C3%A9todo%20de%20regresi%C3%B3n%20log%C3%ADstica,probabilidad%20que%20ocurra%20un%20evento.

Echeverri, A. A. (2019). *Repositorio Institucional Universidad de Antioquia*. Obtenido de https://bibliotecadigital.udea.edu.co/handle/10495/15142

Falla, J. D. (2021). *Repositorio Universidad Tadeo*. Obtenido de https://expeditiorepositorio.utadeo.edu.co/handle/20.500.12010/22247

Galindo, D. A. (2020). *Diseño de un Modelo Predictivo de Fuga de Clientes Utilizando Algoritmos Machine Learning* . Obtenido de Repositorio ECCI: https://repositorio.ecci.edu.co/bitstream/handle/001/773/Dise%C3%B1o%20de%20un%2

0Modelo%20Predictivo%20de%20Fuga%20de%20Clientes.pdf?sequence=1&isAllowed=y

Gil, D. Á. (14 de Enero de 2021). *Adictos al trabajo.com*. Obtenido de https://www.adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/

IBM. (17 de Agosto de 2021). *IBM.com*. Obtenido de https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=models-about-svm

Jélvez Caamaño, A. M. (15 de Octubre de 2014). *Scielo,Universidad Ciencia y Tecnología*. Obtenido de http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1316-48212014000300004&lng=es&nrm=iso&tlng=es

Montero, A. D. (Julio de 2018). *Repositorio Universidad Pontificia Comillas*. Obtenido de https://repositorio.comillas.edu/xmlui/handle/11531/26705

Pinto, D. A. (2020). *Repositorio.ecci.edu.co*. Obtenido de https://repositorio.ecci.edu.co/bitstream/handle/001/773/Dise%C3%B1o%20de%20un%20Modelo%20Predictivo%20de%20Fuga%20de%20Clientes.pdf?sequence=1&isAllowed=y

Poyatos, R. A. (2021). *RPubsbyRstudio*. Obtenido de https://rpubs.com/Raul_A_P/Telco_h2o

R Studio. (junio de 2022). *R project*. Obtenido de https://www.r-project.org/

RETAMAL, M. E. (Diciembre de 2011). *Repositorio Universidad del Bío-Bío*. Obtenido de http://repobib.ubiobio.cl/jspui/bitstream/123456789/2308/1/Ovalle_Retamal_Victor_Francisco_Javier.pdf

Rodrigo, J. A. (junio de 2017). *cinciadedatos.net*. Obtenido de https://www.cienciadedatos.net/documentos/35_principal_component_analysis

Rodrigo, J. A. (Octubre de 2020). *Ciencia de datos .net*. Obtenido de https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting

Roldán, P. N. (9 de Febrero de 2019). *Economipedia.com*. Obtenido de https://economipedia.com/definiciones/muestreo-estratificado.html#:~:text=El%20muestreo%20estratificado%20es%20una,excluyentes%20(no%20se%20superponen).

Troncoso Espinosa, F. H. (22 de Octubre de 2020). *Researchgate.net*. Obtenido de ttps://www.researchgate.net/publication/346994600_PREDICCION_DE_FUGA_DE_CLIENTES_EN_UNA_EMPRESA_DE_DISTRIBUCION_DE_GAS_NATURAL_MEDIANTE_EL_USO_DE_MINERIA_DE_DATOS