Modelo Predictivo de Clientes en Mora

Carolina Bareño Amézquita

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesora

Daniela Serna Buitrago, Especialista (Esp) en Analítica y Ciencia de Datos

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2023

Dedicatoria

Este trabajo está dedicado:
A mi madre, quien con su esfuerzo, apoyo y mucha paciencia ha aportado a cumplir mis sueños y alcanzar las metas que me propongo.
A mi abuelita, quien fue parte fundamental para que hoy este acá.
A mis hermanos, Natalia y Lelian, por su apoyo incondicional.
A Bruno, por ser mi mayor apoyo emocional.
Finalmente, a mis amigos, por siempre estar ahí.

Agradecimientos

Expreso mi mas sincero agradecimiento a quienes hicieron posible la culminación de la especialización que concluye con este trabajo:
Muy especialmente a Astrid, por brindarme un gran soporte durante este año. Tu apoyo y ayuda es invaluable.
A mi familia y amigos por ser un gran circulo de apoyo y amor.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

IPC Índices de Precios al consumidor

KNN k-nearest neighbors algorithm (Algoritmo de los k vecinos más cercanos)

CRISP-DM Cross-Industry Standard Process for Data Mining (proceso estándar intersectorial para mineria de datos)

RESUMEN

En el documento se detalla la importancia de conocer el perfil de comportamiento de pago de los clientes de un Banco. Se busca lograr un mejor perfilamiento de clientes actuales y futuros a través del entrenamiento de un modelo Machine Learning.

Para la selección del mejor modelo de clasificación, se plantea el entrenamiento de modelos de machine learning con una data de históricos tomada de la nube. Los modelos usados son Regresión Logística, KNN y Random Forest, para cada modelo se determinan sus métricas dentro de las que se encuentra el accuracy y la precisión en la clasificación de cada uno de los valores de las variables de salida qué se dan como resultado del entrenamiento y prueba de cada uno de estos. Aquel que cuente en su conjunto con los mejores resultados en las métricas seleccionadas será el modelo que finalmente se presentará para determinar en qué grupo se clasifican los clientes.

Palabras clave — Banca, moroso, préstamo, perfil de cliente, Machine Learning, Regresión Logística, KNN, Random Forest.

ABSTRACT

In this document, it details the importance of know the payment profile of a client from a bank. One of the goals is to find a better way to profile the clients that already had a product and the clients that in the future, through the training of a machine learning model.

To choose the best classification model, detail it the training of Machine Learning models with data downloaded from the cloud. The models training are Logistic Regression, KNN and Random Forest, for each model its calculated their metrics, in those are the accuracy and the precision classifying the clients in one of the values that takes the output variable, these metrics are the result of the training and testing with the data. The one that gives the best values in the metrics selected previously, it will be the model selected, that finally it will classify in which one of the groups belong the clients.

Keywords — Bank, defaulter, loan, client profile, machine learning, logistic regression, KNN, Random Forest.

I. INTRODUCCIÓN

Este documento se realiza como trabajo final de la especialización de analítica y ciencia de datos. En este se presentará el modelamiento realizado sobre la data de estudio para lograr identificar a que grupo puede llegar a pertenecer un cliente de Banco, entre que en el futuro quede en mora o que no lo haga.

Puesto que el indicador de mora se tiene muy presente en los seguimientos que se realizan entres los productos que un banco maneja y en el general de las operaciones de un banco, su incremento desacelerado, puede ser una señal sobre los filtros realizados a los clientes a quienes se les realiza algún tipo de préstamo.

Así pues, se plantea generar un modelo, que con base en histórico de clientes y un indicador de comportamiento de pago, en donde se define si ha entrado en mora o no, se busca lograr identificar con datos comunes de un cliente si este tiene encaja más en el perfil de quedar en mora o no. Este pronóstico anticipado busca que aquellos con una probabilidad más alta tenga otro tipo de filtros o filtros adicionales, a los que tienen los demás, lo que los descarta totalmente para un préstamo o se les aceptaría con restricciones, implementar este tipo de medidas puede resultar en la reducción de los indicadores de mora del banco.

Teniendo en cuenta el problema planteado, se definió usar modelos de clasificación, ya que lo que se busca es ubicar a los clientes en una de las dos posibles salidas: moroso o no moroso. Se seleccionará el modelo que tenga mejores métricas.

Resumiendo, en este documento se presentarán las diferentes iteraciones realizadas sobre la data base para definir el mejor modelo para predecir la probabilidad de que un cliente entre en mora y cuál sería el seleccionado.

II. PLANTEAMIENTO DEL PROBLEMA

Dado que el indicador de mora es una referencia de la calidad de la cartera de un banco frente al resto del sector, es importante conservar este en un valor muy bajo. Dicho indicador se suele ver afectado de una manera casi directa por indicadores sociales como el desempleo.

En atención a qué como banco no se puede controlar los indicadores externos que afectan el comportamiento del cliente, se pretende optimizar los recursos en búsqueda de un mejoramiento en la calidad de la cartera y así brindar la posibilidad de generar mejores beneficios a los clientes que se proyectan dentro del mejor perfil, a su vez, fijar restricciones a aquellos que clasifiquen en un perfil inferior, todo esto teniendo siempre en cuenta los nichos preestablecidos

Se hace necesario plantear un modelo que ayude a clasificar a los clientes que ya se tienen y a los que vienen en un futuro para así conocer de antemano los perfiles de pago de los clientes y realizar ofertas de acuerdo a estos.

III. JUSTIFICACIÓN

El conocer de antemano cómo será el posible comportamiento de pago de un cliente, ayuda a que se le pueda hacer o no una oferta a este. Tener un modelo entrenado con una cantidad de datos históricos considerable, y con unas métricas aceptables para la clasificación de un cliente de acuerdo a su perfil, facilita el proceso de estudio que se les aplica.

Al crear un modelo que responda con el perfil de comportamiento de un cliente, facilita el proceso de filtros previos a un cliente y ayudaría a enfocar los métodos en las ofertas que se pueden llegar a realizar a estos, posterior a conocer a qué perfil pertenece.

Así mismo puede ser de gran ayuda al eliminar la intermediación comercial, quien actualmente es quien realiza el filtro de perfilamiento de un cliente. Ahora, con un modelo se podría evitar el error humano y tener ofertas más acertadas para los clientes.

IV. OBJETIVOS

A. Objetivo general

Clasificar a que grupo corresponde determinado cliente entre moroso o no, con base en una data histórica utilizando técnicas de aprendizaje automático.

B. Objetivos específicos

- Explorar y analizar las características de los datos que contiene la data.
- Evaluar la importancia de característica para el entrenamiento del modelo.
- Entrenar los modelos y evaluar sus métricas, principalmente el accurancy y la precisión en la clasificación especialmente en los morosos.
- Validar la capacidad de clasificación del modelo con data diferente a la de entrenamiento.

V. MARCO TEÓRICO

A. Antecedentes.

El indicador por mora o indicador de cartera castigada, es uno de los indicadores que controlan todas las instituciones financieras y en general el sector. Un indicador alto o bajo de este, indica un decrecimiento o crecimiento económico, respectivamente, tanto a nivel general como del sector o de una institución. Esto se describe en artículos como Mala cartera de créditos baja por los buenos hábitos de pago de la revista Portafolio, donde se señala el decrecimiento en los saldos por mora que reflejan, según mencionan, mejores hábitos de pago de los consumidores[1]. Esta idea también es respaldada por el artículo El indicador de cartera en mora de los bancos nacionales se ubicó en 3,9% en abril donde indica, y cito textualmente "A medida que avanza la recuperación económica, disminuye el índice de cartera vencida." [2]

Así mismo, este indicador puede demostrar la solidez de una economía frente a factores externos, tal como se infiere de la entrevista al superintendente financiero con Portafolio en el artículo Deterioro de la cartera de crédito obedece a factores donde menciona que a pesar de todo lo que está pasando en escenario internacional, la economía colombiana permanece estable y que esto se debe en parte al monitoreo de indicadores[3]. Esta estabilidad en el indicador del sector, se puede ver en el gráfico de Indicador de mora tomado del Anexo estadístico – Indicadores económicos con cifras a febrero 2023[4]:

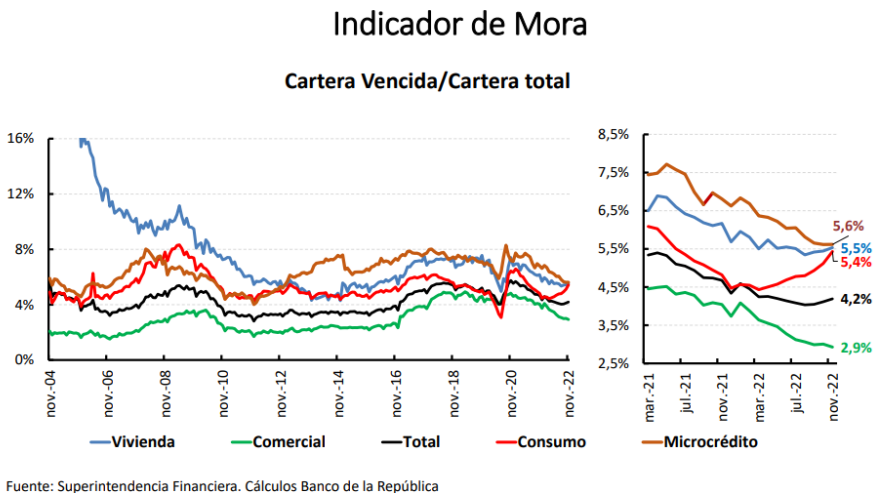


Ilustración 1 Grafico Indicador por mora - Indicadores Económicos Banco de la Republica

Nota. Fuente https://www.banrep.gov.co/sites/default/files/paginas/amjd_marzo_2023.pdf

Donde se evidencia un comportamiento estable entre los últimos 4 años, interrumpido por una caída que se da en pandemia debido a las condonaciones que hubo y que no se estaban generando créditos, y el posterior aumento cuando empezó la reactivación. Así mismo, se ve una caída en el último año y medio, de indicador general.

Factores como el incremento en el desempleo, el aumento de la inflación, y, en consecuencia, las altas tasas de interés para tratar de mitigar este, afectan directamente el comportamiento de pago de los clientes y por consiguiente la calidad de la cartera de las instituciones financieras y del sector. Un ejemplo de esto, se ve en el gráfico tomado del Anexo estadístico – Indicadores económicos con cifras a febrero 2023[4]:

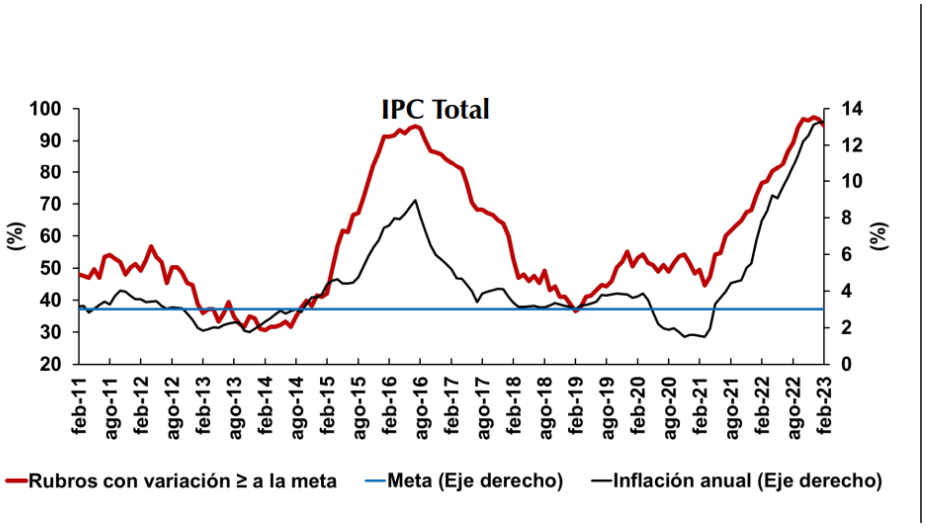


Ilustración 2 Grafico IPC vs Inflación Anual Banco de la Republica

Nota. Fuente https://www.banrep.gov.co/sites/default/files/paginas/amjd_marzo_2023.pdf

Aquí se ve cómo a pesar de tener una inflación en aumento desde agosto 2021, el indicador de mora se mantiene, excepto por picos para los productos de consumo, sin ser algo exponencial. Por esto, el monitoreo constante al total de saldo vencido sobre el total de la cartera, se hace necesario para tomar medidas en caso que se presente algún aumento.

El conocer de antemano como es el comportamiento de pago de un futuro cliente, es fundamental a la hora de decidir si es sujeto de un préstamo o no. Actualmente se manejan bases de datos en el sector, donde se le asigna una calificación a cada cliente y de acuerdo a esta calificación, y otros ítems como tipo de contrato, lugar de trabajo y edad, perfilan al cliente. La calificación, en muchos casos, es el punto de partida, para tener en cuenta a un cliente. Esta viene

dada por las letras A, B, C, D y E, principalmente, donde las letras A y B son los clientes con los mejores comportamientos de pago y las letras D, E y las que vienen de ahí en adelante, son los clientes con peor comportamiento. Esto ha funcionado hasta ahora como un buen filtro, para los nuevos clientes, para que las grandes instituciones mantengan la calidad de su cartera sin mayores fluctuaciones. En general, son las pequeñas instituciones de microcréditos las que se quedan con la población con peor calificación.

Pero, ¿qué pasa con los clientes actuales de las instituciones?, ¿cómo saber si su comportamiento va a continuar siendo tan bueno o van a entrar a la bolsa de la cartera vencida? La consultora Deloitte habla sobre una especie de cobranza preventiva, sobre los clientes actuales que no han caído en cartera vencida, pero tienen una alta probabilidad de incumplimiento de pago, a estos clientes se propone ofrecerles refinanciaciones de sus deudas de acuerdo a su situación, para evitar que caigan en malos hábitos de pago y afecten el indicador de cartera en mora de las instituciones.[5]

En conclusión, el conocer previamente cómo podría ser el comportamiento de un cliente o si los clientes actuales son propensos a desmejorar en sus hábitos de pagos, prepara a las instituciones y al sector para mantener una calidad de la cartera buena y a conciliar previamente con los clientes modalidades de pago que eviten incurrir en gastos como los de cobranza a las instituciones financieras.

B. Análisis Exploratorio de los datos.

A la base de datos que se usara para el entrenamiento de los datos, se inicia realizando un análisis exploratorio. Es de aclarar que esta es una base de datos tomada de Kaggel, sobre la que no se tiene ningún tipo de control.

Partiendo de este punto, lo primero que se realiza es una visualización de los datos. Estos se dividen en tres grupos para realizar su análisis: variable de salida, variable categórica y variable numérica.

1) Variable de salida: La variable de salida para este caso la nombran como Risk_Flag. Esta recibe dos valores, 0 para clasificar los clientes que no tienen una mora en su histórico y 1 para los clientes que si la tienen. Lo que se observa en su distribución es un desbalance entre las frecuencias de cada uno de los valores:

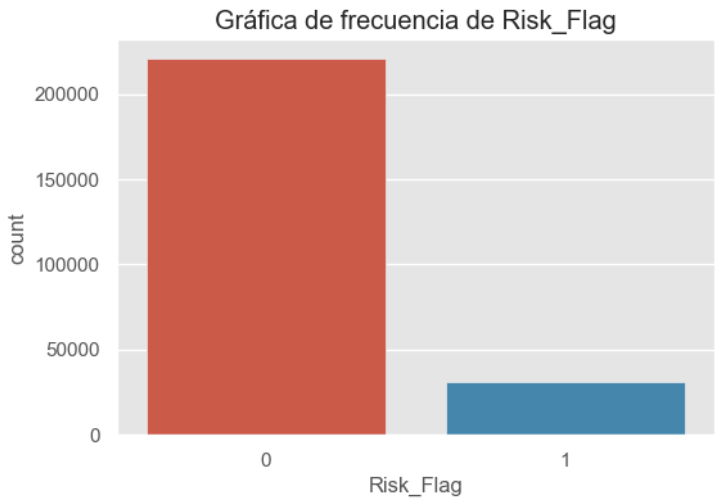


Ilustración 3 Grafico de Frecuencia Variable de Salida

2) Variables categóricas: Para la data las variables categóricas están dadas por: Married/Single, House_Ownership, Car_Ownership, Profession, City y State. Al tomar un análisis descriptivo de estas, resalta que para la variable City existe una gran cantidad de valores únicos en relación a las demás variables:

	Married/Single	House_Ownership	Car_Ownership	Profession	CITY	STATE
count	252000	252000	252000	252000	252000	252000
unique	2	3	2	51	317	29
top	single	rented	no	Physician	Vijayanagaram	Uttar_Pradesh
freq	226272	231898	176000	5957	1259	28400

Ilustración 4 Análisis descriptivo Variables Categóricas.

Se grafican las variables con valores únicos entre 2 y 3 versus la variable de salida para conocer su distribución, lo normal y lo que se presentó, es que estas se comportan de forma similar a la distribución de la variable de salida para los valores que toma:

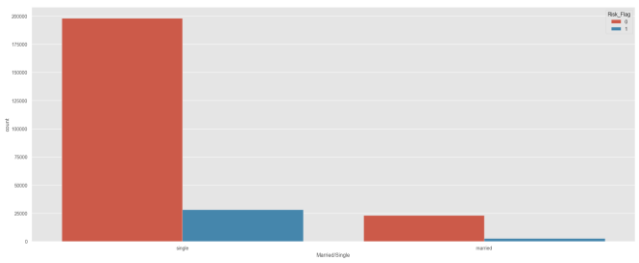


Ilustración 6 Relación Married/Single vs Risk_Flag

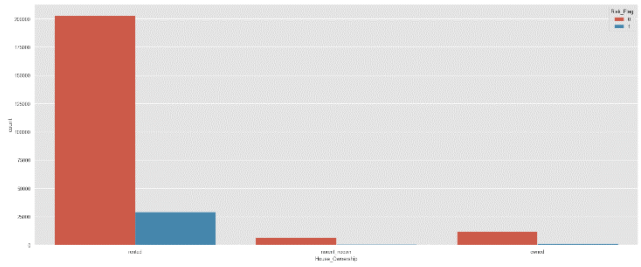


Ilustración 5 Relación House_Ownership vs Risk_Flag

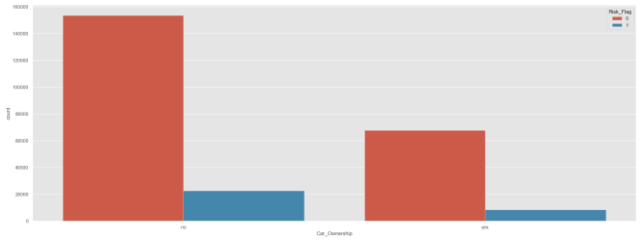


Ilustración 7 Relación Car_Ownership vs Risk_Flag

Para las variables Profession y State, se realizó un gráfico de frecuencias donde se pudiera observar la no presencia de valores extraños o atípicos. Al contener tantos valores únicos la variable City se decide no graficarla, ya que no era posible ver alguna información relevante:

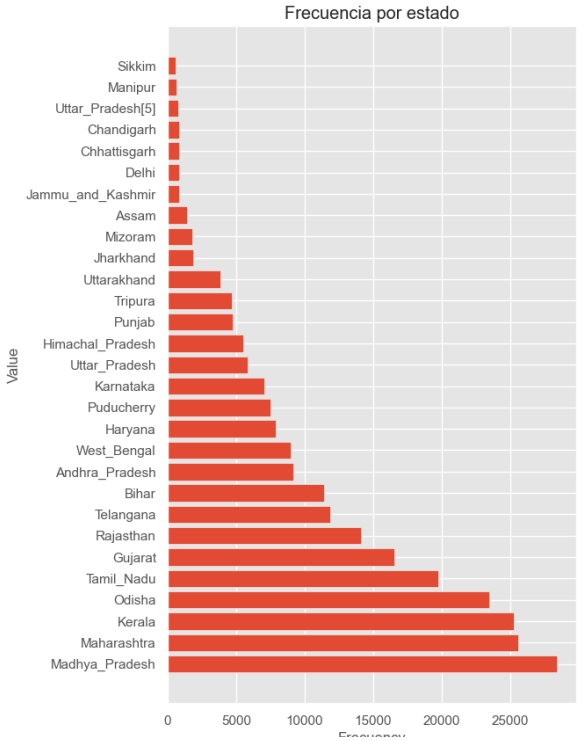


Ilustración 8 Grafico de Frecuencia por State

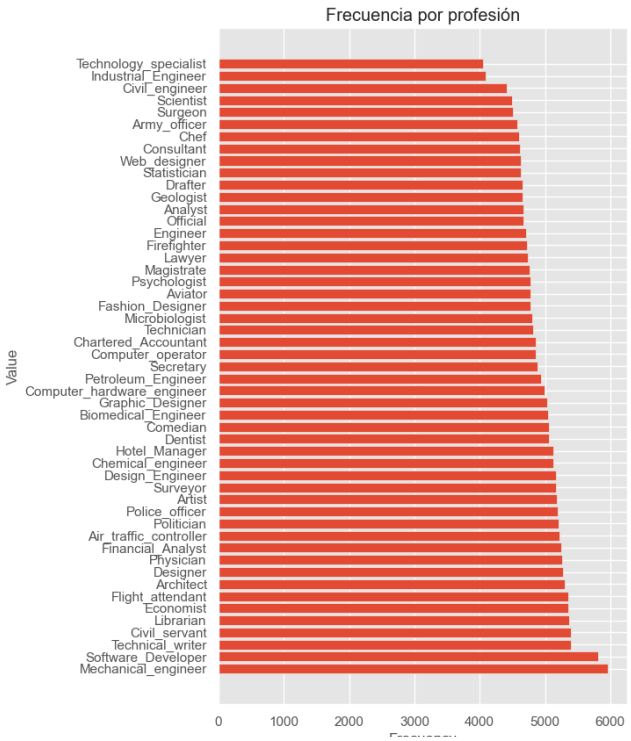


Ilustración 9 Grafico de Frecuencia por Profession

Por último, se analizó la relación entre las variables categóricas, donde se incluye la variable de salida, a través de una tabla de contingencia y sus coeficientes. Estos coeficientes se grafican en un mapa de calor para realizar una mejor visualización y así mismo tomar decisiones sobre los resultados:

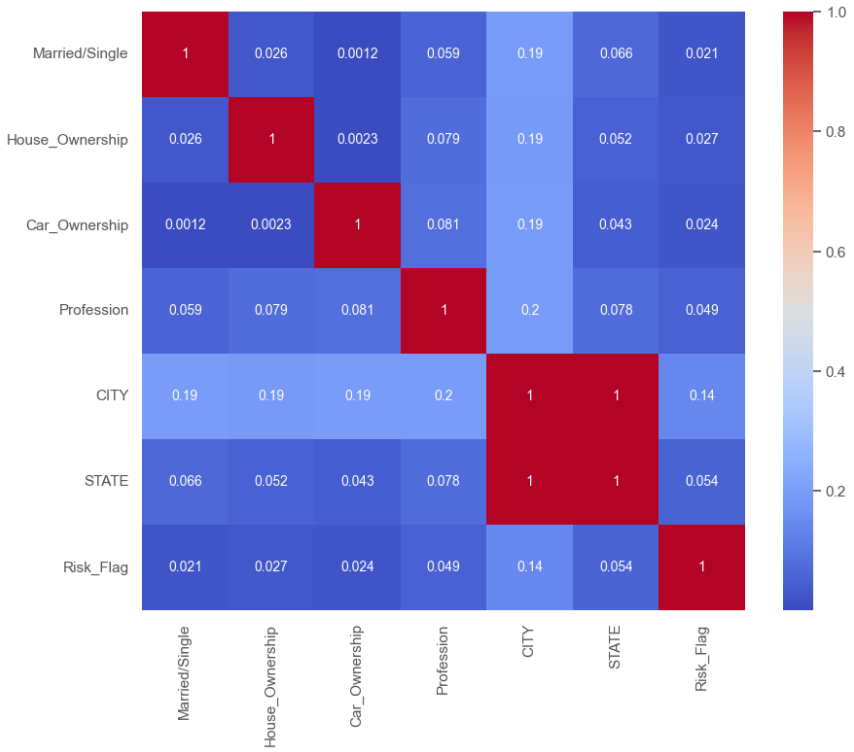


Ilustración 10 Mapa de Calor Coeficientes de Contingencia.

Como se observa en la diagramación de los coeficientes, hay una asociación de 1 entre las variables City y State. Esto se toma como que es posible eliminar una de las variables para entrenar los modelos y no se tendría mayor afectación a si se entrenara incluyéndolas todas, por lo que se decide eliminar la variable City que es la que tiene más valores únicos y al momento de realizar la creación de variables dummies va a representar la creación de más columnas que pueden acarrear un costo computacional no deseado.

3) Variables numéricas: las variables numéricas de la data están nombradas como Income, Age, Experience, Current_job_yrs y Current_house_yrs. La variable de salida está representada como una variable de tipo numérica, pero para este análisis no se tendrá en cuenta. Se realiza un análisis descriptivo de las variables para empezar:

	Income	Age	Experience	CURRENT_JOB_YRS	CURRENT_HOUSE_YRS
count	2.520000e+05	252000.000000	252000.000000	252000.000000	252000.000000
mean	4.997117e+06	49.954071	10.084437	6.333877	11.997794
std	2.878311e+06	17.063855	6.002590	3.647053	1.399037
min	1.031000e+04	21.000000	0.000000	0.000000	10.000000
25%	2.503015e+06	35.000000	5.000000	3.000000	11.000000
50%	5.000694e+06	50.000000	10.000000	6.000000	12.000000
75%	7.477502e+06	65.000000	15.000000	9.000000	13.000000
max	9.999938e+06	79.000000	20.000000	14.000000	14.000000

Ilustración 11 Análisis descriptivo variables numéricas.

En este se evidencia que la variable Income recibe unos valores muy grandes con respecto a las demás variables.

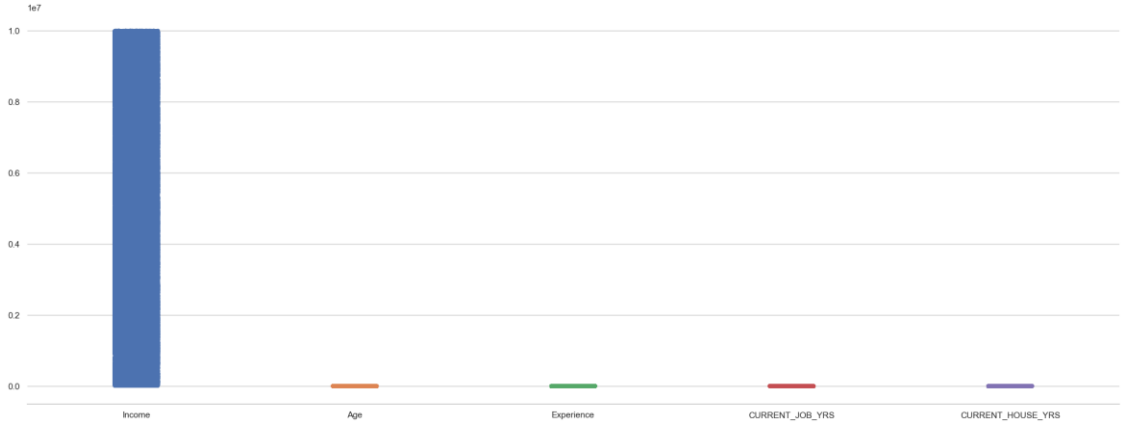


Ilustración 12 Grafico strip plot variables numéricas.

Se comprueba con un gráfico tipo strip plot, que se hace necesario normalizar las variables para realizar un buen análisis, en donde se pueda observar más detalladamente si hay presencia de valores atípicos.

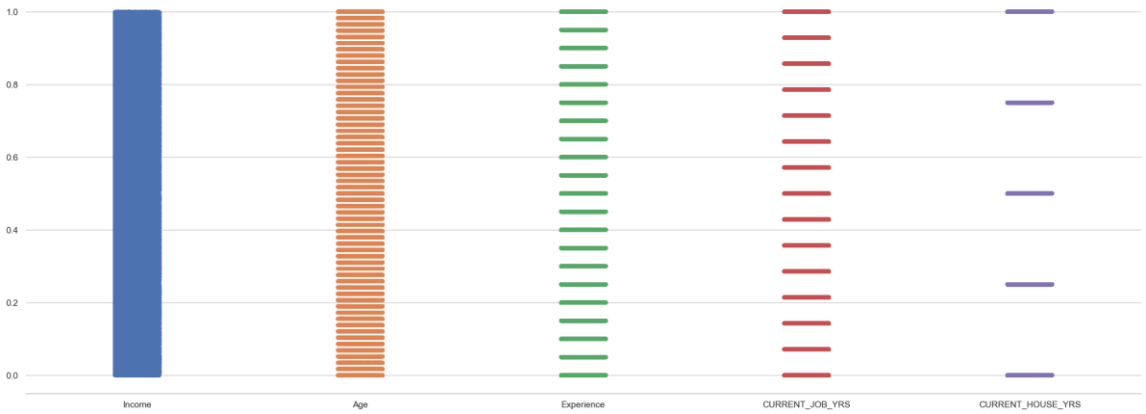


Ilustración 13 Grafico strip plot variables numéricas normalizadas

En el grafico se observa que no hay presencia de valores atípicos en ninguna de las variables, por lo que para este caso no se hace necesario aplicar alguna técnica para el tratamiento de estos. A continuación, se realiza una matriz de correlación, a través de un mapa de calor, aqui se espera ver la existencia o no de correlaciones altas. Se considera alta, aquella que se encuentre por encima de 0.5.

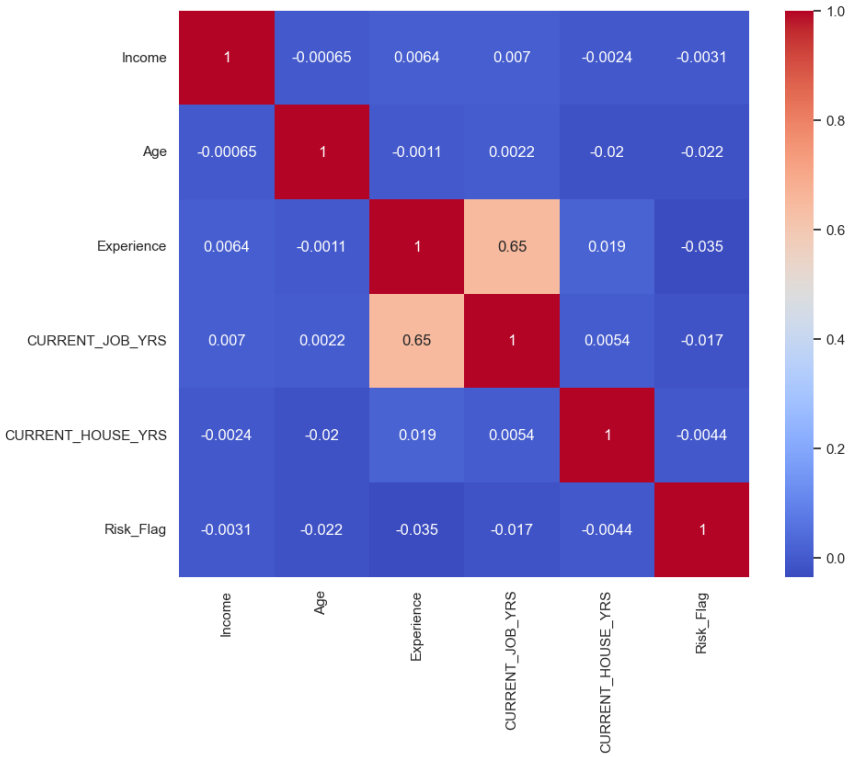


Ilustración 14 Matriz de correlación variables numéricas.

Se ve una correlación alta entre las variables Experience y Current_job_yrs, por lo que se decide eliminar y no tener en cuenta para el entrenamiento de los modelos, la variable Expérience. Se considera que aporta más información, para el caso, el conocer cuántos años lleva el cliente en su último empleo que el total de años de experiencia.

Para finalizar la exploración de los datos, se realiza una limpieza y preparación de estos. Aquí se analiza la presencia de datos nulos o vacíos, para el caso de la data se da que no existe este tipo de datos. Así mismo como parte de la preparación, se eliminan las variables City y Experience producto del análisis previamente hecho, esto hace que el dataset pase de tener 12 variables a 10, conservando los 252.000 registros.

VI. METODOLOGÍA

La metodología usada para este proyecto es CRISP-DM, a continuación, se explica cómo se aplicó cada una de las fases que la componen para el desarrollo del proyecto.

A. *Fase I. Comprensión del negocio. Definición de necesidades del cliente*

Inicialmente se define el objetivo del proyecto con la data que se elige. Se parte de conocer que la variable de salida es categórica y recibe dos tipos de valores, y que al buscar poder identificar a que grupo se ajusta mejor el perfil de un cliente entre uno de estos dos valores, lo más apropiado es entrenar modelos de clasificación.

B. *Fase II. Entendimiento de los datos. Estudio y comprensión de los datos*

En esta fase lo que se realiza es el proceso descrito en el análisis exploratorio de los datos. Se buscaba conocer más los datos, a través de gráficos y medidas estadísticas. Se calcularon medidas como coeficientes de correlación y contingencia, se concluyó que había características con altos valores en estos parámetros por lo que se evaluó la posibilidad de trabajar con solo algunas de ellas para el entrenamiento de los modelos.

C. *Fase III. Preparación de los datos. Análisis de los datos y selección de características*

En esta fase de limpieza de los datos para preparar ya la base que será usada para el modelamiento, se verifica la presencia o no de valores nulos o vacíos, así mismo se eliminaron las características que en el análisis exploratorio se encontró que no aportaban información relevante, y finalmente se creó una data con valores dummies para las variables categóricas. Como la variable de salida es de tipo categórica, se cambian los valores de 0 y 1 por Not Defaulter y Defaulter, respectivamente.

En esta fase se vio, que al ser una data tomada de una página de almacenamiento de datasets, está muy limpia. En esta no se encontraron ni valores atípicos o valores nulos, que si fuese una data propia sería más probable la existencia de estos.

Aquí también se decidió, que como la variable de salida estaba desbalanceada, se realizarían dos modelamientos, uno con la base completa y otro con la base a la que se le aplicaría undersampling. A esta base reducida se le aplico la misma limpieza y creación de variables dummies que se aplicó con el total de la base.

D. Fase IV. Modelado

Para el caso, se deciden aplicar 3 modelos, Regresión Logística, KNN y Random Forest. Para cada uno de estos modelos, se evaluó la precisión en la predicción de la variable de salida y el accurancy.

Los 3 modelos se entrenaron con las dos bases, tanto la base completa como la base undersamplig. Así mismo, los 3 modelos fueron entrenados con el 80% de la base y el restante se usó para realizar los test en cada entrenamiento.

Para el entrenamiento de cada uno de los modelos se tomaron los hiperparametros que aseguraban un accurancy más alto. Es decir, cada uno de los modelos, se entrenó de tal forma que se aseguró que las propiedades seleccionadas para cada uno, diera la mayor calidad del modelo.

E. Fase V. Evaluación (obtención de resultados)

Para la evaluación de los modelos entrenados, se tomó una base de prueba que también proporcionaba la página de donde se extrajo la base de entrenamiento. Lo que se buscó al hacer este tipo de evaluación, fue comparar las métricas de los modelos entrenados y las resultantes al alimentar los modelos con esta base de prueba, que, en cierto modo los modelos no conocen y verificar su exactitud.

A partir de las métricas que se dieron en este punto, y realizando una comparativa entre los diferentes modelos, se tomará como el modelo a implementar el que, en su conjunto, tenga un buen accurancy y unos valores considerables en la precisión de la predicción de la variable de salida.

F. Fase VI. Despliegue (puesta en producción)

En esta fase, que no se encuentra en el alcance del proyecto, se realiza la preparación del entorno de producción para el despliegue del modelo. Se debe establecer un plan donde se indique los puntos donde para el despliegue del modelo. Así mismo, en este punto, se deben realizar las pruebas pertinentes que aseguren el correcto funcionamiento del modelo en el entorno de producción.

Finalmente se debe establecer el cómo se monitorea el desempeño del modelo, en este se debe realizar los ajustes y mantenimiento del mismo para que este siga siendo preciso y útil.

VII. RESULTADOS

Los resultados se presentan por modelo. Como se mencionó previamente, cada uno es medido por la precisión en la predicción de la variable de salida y el accurancy del modelo, y cada uno se entrenó con dos bases y se evaluó con la misma base de prueba.

A. Modelo Regresión Logistica (RL).

Para definir los hiperparametros se crearon 3 modelos. Uno con los hiperparametros: multi_class = "ovr", solver='liblinear', otro con multi_class = "ovr", solver='lbfgs' y finalmente uno con multi_class = "multinomial", solver='lbfgs'.

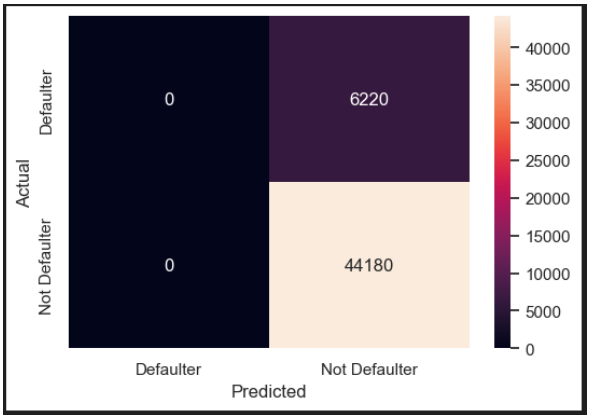
1) Métricas del modelo entrenado con la data completa. Para este caso se dio que para los tres modelos entrenados con la data completa el accurancy era el mismo así se cambiaran los hiperparametros:

```
==========Accuracy Logistic Regression ================
ovr - Linear       : 0.8765873015873016
ovr - lbfgs        : 0.8765873015873016
multinomial - lbfgs: 0.8765873015873016
```

Ilustración 15 Accurancy Modelos RL Data Completa

Para el cálculo de las métricas se tomo el modelo con los hiperparametros multi_class = "ovr", solver='lbfgs', que, según la teoría vista, son los que mas se ajustan al dataset manejado.

Al graficar la matriz de confusión y calcular las métricas para este, se ve como el modelo no logra identificar ningún valor que se clasifique como Defaulter.



	precision	recall	f1-score	support
Defaulter	0.00	0.00	0.00	6220
Not Defaulter	0.88	1.00	0.93	44180
accuracy			0.88	50400
macro avg	0.44	0.50	0.47	50400
weighted avg	0.77	0.88	0.82	50400

Ilustración 16 Matriz de Confusión y Métricas RL data Completa

Es un modelo que cuenta con un accurancy de 0.88 que se puede catalogar como aceptable, sin embargo, falla en la precisión de la predicción de la variable de salida cuando esta se clasifica como Defaulter.

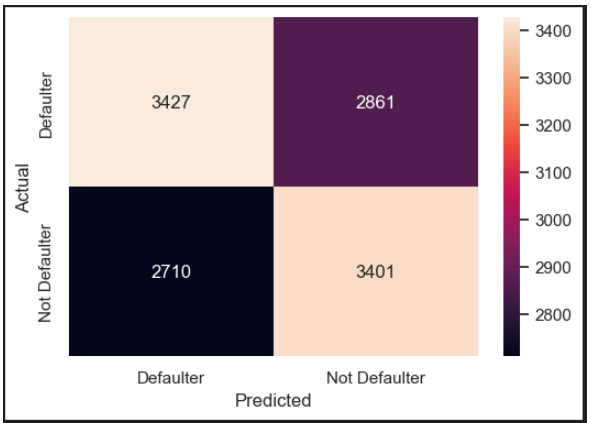
2) Métricas del modelo entrenado con la data undersamplig. Para la selección de hiperparametros y entrenamiento, se uso la base a la que se le aplico undersamplig. Se entrenan 3 modelos con la misma selección de hiperparametros que se hizo para entrenar el modelo con la base completa, es decir, uno con hiperparametros: multi_class = "ovr", solver='liblinear', otro con multi_class = "ovr", solver='lbfgs' y finalmente uno con multi_class = "multinomial", solver='lbfgs'.

Se da que 2 de los modelos entrenados tiene un mismo accurancy, más alto que el del tercero.

```
==========Accuracy Logistic Regression ================
ovr - Linear       : 0.5506895717396564
ovr - lbfgs        : 0.5506089200741995
multinomial - lbfgs: 0.5506895717396564
```

Ilustración 17 Accurancy Modelos RL Data Undersampling

Para el grafico de la matriz de confusión y las métricas, se toma el modelo con los hiperparametros multi_class = "ovr", solver='liblinear'.



	precision	recall	f1-score	support
Defaulter	0.56	0.55	0.55	6288
Not Defaulter	0.54	0.56	0.55	6111
accuracy			0.55	12399
macro avg	0.55	0.55	0.55	12399
weighted avg	0.55	0.55	0.55	12399

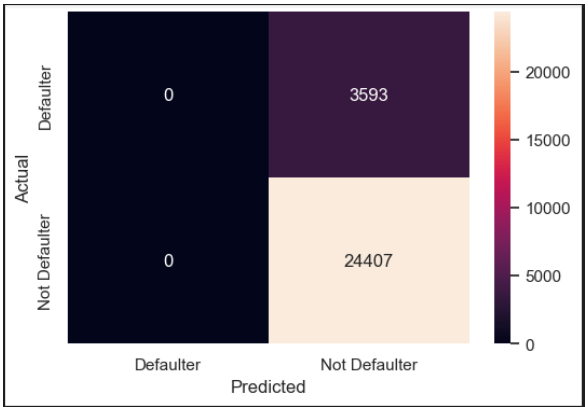
Ilustración 18 Matriz de Confusión y Métricas RL data Undersampling

Aquí se observó que el modelo tiene un accurancy de 0.55, que se considera aceptable, pero bajo en comparación con el modelo entrenado con toda la data, sin embargo, a diferencia del

modelo anterior, este tiene mejor precisión en la clasificación cuando la variable de salida es Defaulter.

3) Evaluación y comparación de los modelos entrenados previamente con la data de prueba. Aquí lo que se hizo, fue tomar los dos modelos entrenados, uno con la data completa y otro con la data undersampling, y se tomó los hiperparametros que tenían mejor accurancy para calcular las mismas métricas pasándole la data de prueba para poder comparar.

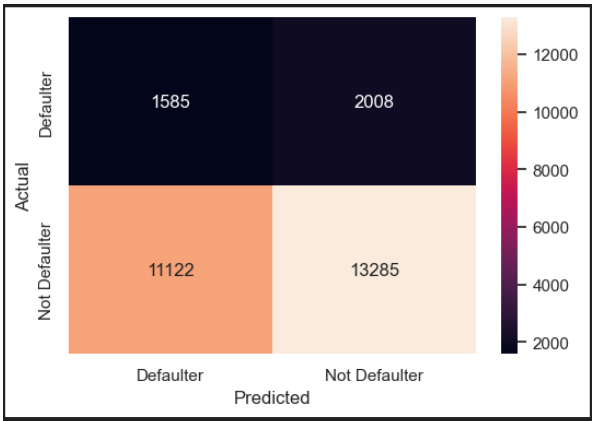
Para el modelo de regresión lineal entrenado con toda la data, como se esperaba, no logra tener una buena precisión al momento de clasificar en Defaulter. Sigue teniendo un accurancy bueno, pero pues para el caso no es suficiente solo que esta medida sea buena.



	precision	recall	f1-score	support
Defaulter	0.00	0.00	0.00	3593
Not Defaulter	0.87	1.00	0.93	24407
accuracy			0.87	28000
macro avg	0.44	0.50	0.47	28000
weighted avg	0.76	0.87	0.81	28000

Ilustración 19 Matriz de Confusión y Métricas Prueba RL Modelo Entrenado con la Data Completa

Para el modelo entrenado con la adata undersampling, al pasarle la base de prueba conserva un accurancy bajo, la precisión baja al momento de clasifica en Defaulter y aumenta para clasificar Not Defaulter, frente a las métricas obtenidas en el entrenamiento del modelo.



	precision	recall	f1-score	support
Defaulter	0.12	0.44	0.19	3593
Not Defaulter	0.87	0.54	0.67	24407
accuracy			0.53	28000
macro avg	0.50	0.49	0.43	28000
weighted avg	0.77	0.53	0.61	28000

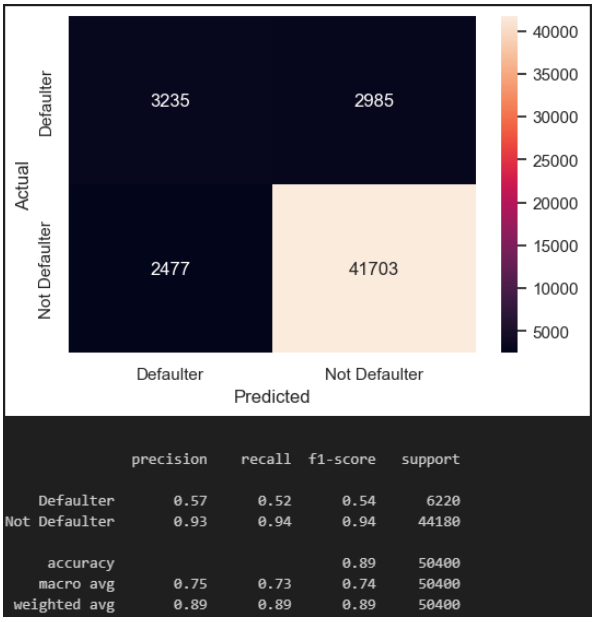
Ilustración 20 Matriz de Confusión y Métricas Prueba RL Modelo Entrenado con la Data Undersampling

Si se tuviera que determinar entre estos dos modelos el definitivo para cumplir el objetivo de este proyecto, se tomaría el modelo entrenado con la data Undersamplig puesto que, en las métricas que se están evaluando, tiene los mejores valores.

B. Modelo KNN

Para el modelo de KNN o vecinos más cercanos, se realizó una validación cruzada con la data completa, variando el número de vecinos entre 2,5,7 y 10. El resultado arrojo que se tendrían mejores métricas con 7, este valor también se aplicó para la data undersamplig, ya que al se una muestra de la data completa, y para que los modelos tuvieran las mimas condiciones, se consideró que este es el dato que generaría mejores resultados también para el caso.

1) Métricas del modelo entrenado con la data completa. Como se dijo previamente, este modelo se entrenó tomando como hiperparametro 7 vecinos, y la data completa. El accurancy para este modelo es de 0.8916. Al graficar la matriz de confusión y calcular las métricas de precisión, arrojo unos valores considerablemente buenos:

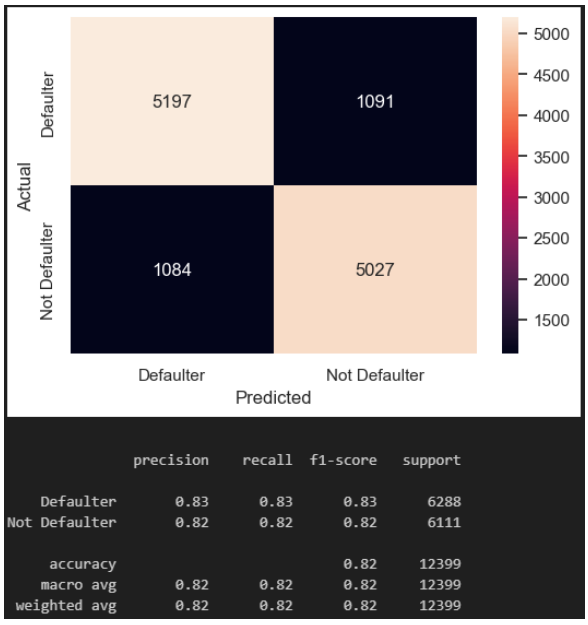


	precision	recall	f1-score	support
Defaulter	0.57	0.52	0.54	6220
Not Defaulter	0.93	0.94	0.94	44180
accuracy			0.89	50400
macro avg	0.75	0.73	0.74	50400
weighted avg	0.89	0.89	0.89	50400

Ilustración 21 Matriz de Confusión y Métricas KNN Data Completa

Se observo que tiene una precisión considerablemente buena para clasificar la data entre Defaulter y Not Defaulter.

2) Métricas del modelo entrenado con la data undersamplig. El accurancy para este modelo entrenado con la data undersampling, da un valor de 0.8245. Como se menciono al inicio de esta sección, el modelo se entreno con 7 vecinos cercanos como hiperparametro, y este fue el modelo usado para graficar la matriz de confusión y las métricas:

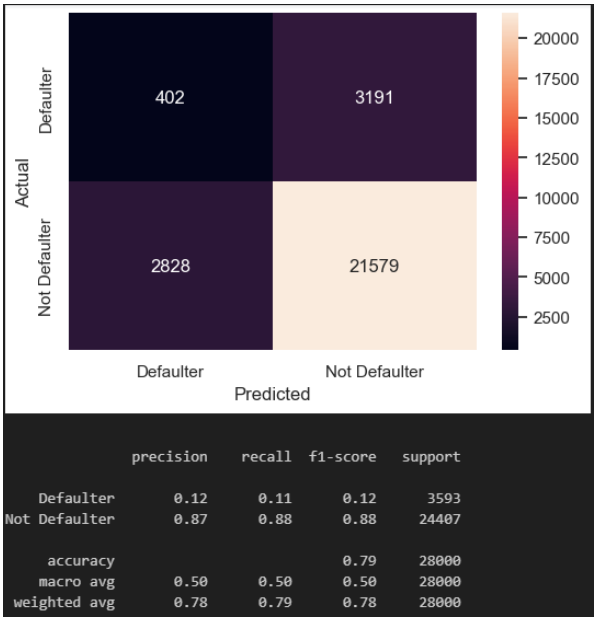


	precision	recall	f1-score	support
Defaulter	0.83	0.83	0.83	6288
Not Defaulter	0.82	0.82	0.82	6111
accuracy			0.82	12399
macro avg	0.82	0.82	0.82	12399
weighted avg	0.82	0.82	0.82	12399

Ilustración 22 Matriz de Confusión y Métricas KNN Data Undersampling

Se observa una mejora considerable en la precisión para clasificar en la variable de salida de Defaulter, en comparación con el modelo entrenado con toda la data, pasa de 0.57 a 0.83. Aunque se disminuye esta misma precisión cuando la variable de salida toma valores de Not Defaulter, esta no es alarmante y adicional, se observa como un equilibrio en la precisión para clasificar la variable de salida en los dos valores.

3) Evaluación y comparación de los modelos entrenados previamente con la data de prueba. Para realizar esta evaluación, se utilizo la misma data de prueba en los dos modelos. El primer modelo a evaluar, es el entrenado con el total de la data, arrojando la siguiente matriz de confusión y métricas:

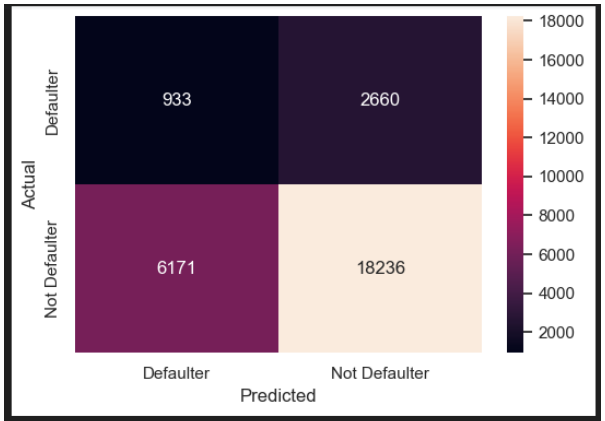


	precision	recall	f1-score	support
Defaulter	0.12	0.11	0.12	3593
Not Defaulter	0.87	0.88	0.88	24407
accuracy			0.79	28000
macro avg	0.50	0.50	0.50	28000
weighted avg	0.78	0.79	0.78	28000

Ilustración 23 Matriz de Confusión y Métricas Prueba KNN Modelo Entrenado con la Data Completa

Se observo una caída en la precisión y en el accurancy, en relación a las métricas del entrenamiento del modelo. Aunque para la precisión en clasificar Not Defaulter y el accurancy las caídas no son considerables, para clasificar Defaulter se ve una caída de 45 puntos y queda en una métrica muy baja.

A continuación, se evaluó con la misma base el modelo entrenado con la data undersampling, presentando la siguiente matriz de confusión y métricas:



	precision	recall	f1-score	support
Defaulter	0.13	0.26	0.17	3593
Not Defaulter	0.87	0.75	0.81	24407
accuracy			0.68	28000
macro avg	0.50	0.50	0.49	28000
weighted avg	0.78	0.68	0.72	28000

Ilustración 24 Matriz de Confusión y Métricas Prueba KNN Modelo Entrenado con la Data Undersampling

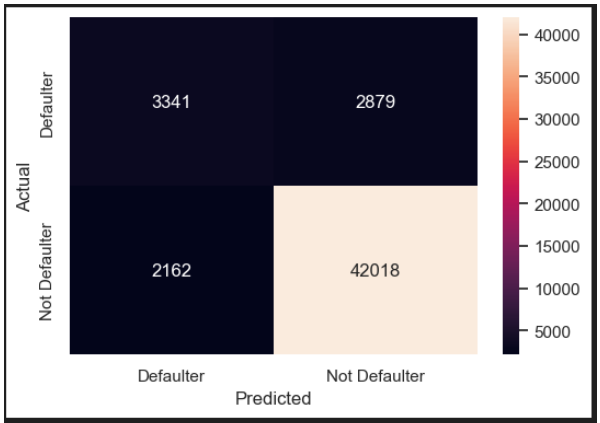
Así mismo, se ve una caída en las métricas que se están evaluando. Para este caso, se ve como también hay una caída considerable en el accurancy.

Si se tuviera que seleccionar el modelo definitivo para alcanzar el objetivo del trabajo, se tomaría el modelo entrenado con el total de la data. Aunque las métricas de precisión son similares en los dos modelos, tiene un mejor accurancy.

C. *Modelo Random Forest.*

Para la selección de hiperparametros del modelo se aplicaron dos técnicas, una fue aplicar un Grid Search basado en out-of-bag score y la otra aplicar un Grid Search basado en validación cruzada. Las dos técnicas se aplicaron tanto para la data completa como para la data undersamplig, al final para entrenar cada modelo, se seleccionó los hiperparametros que arrojaran mayor accurancy entre las dos técnicas.

1) Métricas del modelo entrenado con la data completa. Para entrenar este modelo se seleccionaron los hiperparametros criterion='entropy', max_depth=None, max_features=7, n_estimators=150, dado por la técnica de Grid Search basado en out-of-bag score que calculo un accurancy de 0.8994 en comparación con el calculado por los seleccionados por Grid Search basado en validación cruzada, que daba un accurancy de 0.8992. Para este modelo la matriz de confusión y métricas son las siguientes:

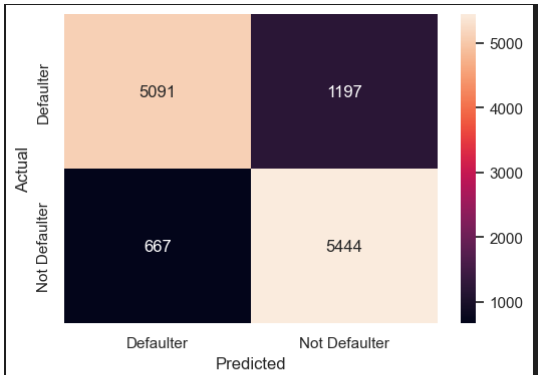


	precision	recall	f1-score	support
Defaulter	0.61	0.54	0.57	6220
Not Defaulter	0.94	0.95	0.94	44180
accuracy			0.90	50400
macro avg	0.77	0.74	0.76	50400
weighted avg	0.90	0.90	0.90	50400

Ilustración 25 Matriz de Confusión y Métricas Random Forest Data Completa

El modelo arrojo unas métricas considerablemente buenas. Cuenta con una precisión considerable para clasificar la variable de salida, así como un accurancy alto.

2) Métricas del modelo entrenado con la data undersamplig. Los hiperparametros seleccionados para entrenar el modelo con la data undersamplig, son los calculados por el grid search basado en out of bag, que dan un accurancy del 0,8513 en comparación con el que dan los hiperparametros del grid search basado en validación cruzada de 0,8464. Esto genera la siguiente grafica matriz de correlación y métricas:

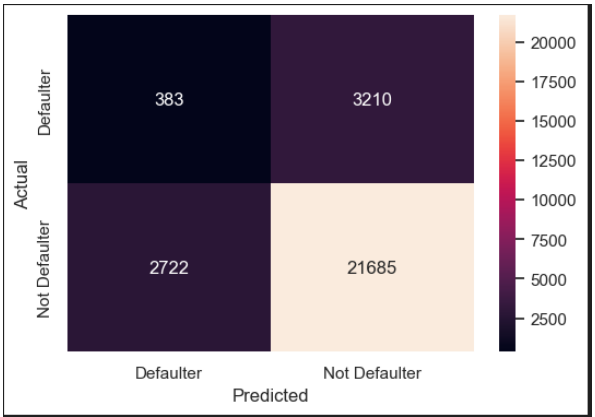


	precision	recall	f1-score	support
Defaulter	0.88	0.81	0.85	6288
Not Defaulter	0.82	0.89	0.85	6111
accuracy			0.85	12399
macro avg	0.85	0.85	0.85	12399
weighted avg	0.85	0.85	0.85	12399

Ilustración 26 Matriz de Confusión y Métricas Random Forest Data Undersampling

Se observo como se equilibra la clasificación en la predicción de la variable de salida con este modelo, el accurancy baja, pero no hasta un punto que denote que este sea malo. La precisión en clasificar los registros en Defaulter es bueno.

3) Evaluación y comparación de los modelos entrenados previamente con la data de prueba. Para realizar esta prueba se usa la misma data de prueba en los dos modelos entrenados previamente, con los hiperparametros seleccionados a través de los grid search aplicados. EL primer modelo que se evalúo con esta data, fue el creado con la data completa y su matriz de confusión y métricas son:

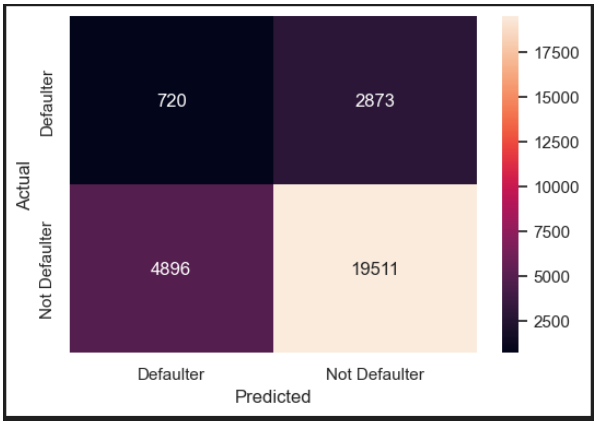


	precision	recall	f1-score	support
Defaulter	0.12	0.11	0.11	3593
Not Defaulter	0.87	0.89	0.88	24407
accuracy			0.79	28000
macro avg	0.50	0.50	0.50	28000
weighted avg	0.78	0.79	0.78	28000

Ilustración 27 Matriz de Confusión y Métricas Data Prueba Random Forest Data Completa

Se destaca una caída en los valores de precisión en la clasificación de la variable de salida, especialmente en al clasificar los registros como Defaulter versus las métricas resultado del entrenamiento del modelo. También hay una disminución en el valor del accurancy, sin embargo, sigue siendo un valor considerable como bueno.

A continuación, se presentan las métricas resultantes sobre probar el modelo entrenado con la data undersampling, pasándole la data de prueba.



	precision	recall	f1-score	support
Defaulter	0.13	0.20	0.16	3593
Not Defaulter	0.87	0.80	0.83	24407
accuracy			0.72	28000
macro avg	0.50	0.50	0.50	28000
weighted avg	0.78	0.72	0.75	28000

Ilustración 28 Matriz de Confusión y Métricas Data Prueba Random Forest Data Undersampling

Al igual que el modelo entrenado con el total de la data, hay una caída en las métricas de precisión en la clasificación de la variable de salida y el accurancy.

En este caso si se tuviera que seleccionar uno de los dos modelos como el definitivo, se tomaría el entrenado con la data undersamplig. Aunque su accurancy es bajo con respecto al modelo entrenado con toda la data la precisión es ligeramente más alta para identificar "Defaulter", que se considera como los críticos a identificar, igual que el recall o verdaderos positivos es más alta en este modelo y el costo computacional usando un modelo entrenado con data undersamplig tiende a ser más bajo.

C. *Selección del Modelo.*

Después de conocer las métricas de los modelos seleccionados tanto al momento de entrenarlos como con la data de prueba, el modelo que se selecciona es Random Forest entrenado con la base undersampling. Como se observa en la tabla, este modelo conserva los valores más altos en casi todas las métricas de evaluación con las diferentes bases con las que se entrenó y se probó el modelo.

A pesar de que el modelo de Random Forest entrenado con la base original, arroja un accurancy mas alto cuando se evalúa con la base de prueba, el modelo entrenado con la base undersampling tiene un valor ligeramente mas alto en la precisión para clasificar como Defaulter, que se considera un punto crítico, ya que, para el objetivo del trabajo es importante conocer quien

en un punto va a entrar en mora. Se presenta algo curioso y es que para todos los modelos el valor en la precisión para clasificar como Not Defaulter con la base de prueba es de 0.87.

	Accurancy				Presicion							
					Defaulter				Not Defaulter			
	Base Original	Base undersampling	Base prueba		Base Original	Base undersampling	Base prueba		Base Original	Base undersampling	Base prueba	
			Modelo Base original	Modelo Base undersampling			Modelo Base original	Modelo Base undersampling			Modelo Base original	Modelo Base undersampling
Regresion logistica	0,88	0,55	0,87	0,53	0,00	0,56	0,00	0,12	0,88	0,54	0,87	0,87
KNN	0,89	0,82	0,79	0,68	0,57	0,83	0,12	0,13	0,93	0,82	0,87	0,87
Random Forest	0,90	0,85	0,79	0,72	0,61	0,88	0,12	0,13	0,94	0,82	0,87	0,87

Tabla 1 Resumen de las métricas obtenidas por modelo y base usada.

X. CONCLUSIONES

- Al usar para el entrenamiento del modelo una data no propia, en la exploración de datos se observó ser una data que ya venía limpia, no se encontraron valores atípicos o nulos.
- Se observó que la información aportada por ciertas características no era relevante frente a las demás, o dentro de estas existía un símil, por lo que se decidió no contar con estas para el entrenamiento de los modelos.
- Contar con un modelo que asegure una precisión superior al 80% al momento de clasificar un cliente entre sí será moroso o no, representa un punto de partida valioso al momento de tomar decisiones en la estrategia comercial que se pueda llegar a manejar en un Banco.
- El conocer de antemano el posible comportamiento de un cliente con sus pagos, puede reducir costos como generados al realizar cobranzas, es un buen punto de para plantear a los entes reguladores reducir las penalidades por la posibilidad de que un cliente no pague su cartera o en dado caso, ser más rentable por conocer de antemano a qué clientes se les puede dar créditos a un plazo sin incurrir en el pago de multas por carteras vencidas.
- Que el modelo nos asegure una exactitud del 80% clasificando a los clientes correctamente en su perfil de pago, se considera valioso en el sector de la banca, ya que es un gran aporte para la gestión del riesgo al momento de realizar los créditos y la toma de decisiones a nivel gerencial.

  Sin embargo, considerando que quienes representan un mayor peso en los indicadores de cartera vencida, son los clientes que se perfilaron como morosos, para este caso, la precisión al momento de clasificar los morosos es igual de importante que la exactitud. Al momento de testear los modelos con la base de prueba, ninguno tuvo una precisión que se considere como buena, por lo que cualquiera de estos que se saque a producción debe ser monitoreado constantemente hasta lograr una mejor precisión en la clasificación de los morosos.

Dentro del análisis para la decisión del modelo, se debe tener en cuenta el costo de errores de clasificación. Este cálculo no está dentro del alcance de este documento, pero es importante que se considere en la implementación.

REFERENCIAS

[1] «Mala cartera de créditos baja por los buenos hábitos de pago», *PORTAFOLIO*, 22 de 2022. [En línea]. Disponible en: https://www.portafolio.co/economia/finanzas/los-creditos-en-colombia-crecen-y-continuan-mejorando-en-2022-575524

[2] «El indicador de cartera en mora de los bancos nacionales se ubicó en 3,9% en abril», *La República*, 28 de 2022. [En línea]. Disponible en: https://www.larepublica.co/finanzas/el-indicador-de-cartera-en-mora-de-los-bancos-nacionales-se-ubico-en-3-9-en-abril-3412079

[3] «Deterioro de la cartera de crédito obedece a factores macroeconómicos», *PORTAFOLIO*, 26 de 2023. [En línea]. Disponible en: https://www.portafolio.co/economia/finanzas/entrevista-a-superintendente-financiero-de-colombia-580454

[4] «Anexo estadístico Informe de política monetaria». 28 de febrero de 2023. [En línea]. Disponible en: https://www.banrep.gov.co/sites/default/files/paginas/amjd_marzo_2023.pdf

[5] Galaz, Yamazaki, Ruiz Urquiza, y S.C., «Tendencias de Cobranza y Recuperacion de Cartera en el Sector Financiero a Partir de la Crisis». 2012. [En línea]. Disponible en: https://www2.deloitte.com/content/dam/Deloitte/pa/Documents/financial-services/2015-01-Pa-FinancialServices-CobranzaCartera.pdf

ANEXOS

Anexo A. Autoarchivo en Repositorio y documentos de interés

- Repositorio: https://github.com/cabareno/Loan_Behavior