Proyecto Final

Aprobación de Tarjetas de Crédito Utilizando Machine Learning

José Roberto
Torres Bello

Proyecto I

Profesor: María Fernanda Sánchez Puig
Ayudante: José Fernando Méndez Torres

agosto 2021

Facultad de Ciencias

Universidad Nacional Autónoma de México

Índice general

Capítulo 1

Introducción

Los bancos comerciales reciben muchas solicitudes de tarjetas de crédito, muchas de ellas son rechazadas por diversas razones, como saldos elevados de préstamos, bajos niveles de ingresos o demasiadas consultas sobre el informe crediticio de una persona, por poner algunos ejemplos.

Figura 1.1: Trajeta de Crédito.

El acreedor debe evaluar el riesgo de que el deudor incumpla con sus obligaciones en algún momento de la vida del préstamo, un método tradicional que utilizan los bancos es el de las 5Cs, las cuales son evaluar las siguientes cinco características:

- Capacidad de pago.
- Comportamiento de pago.
- Carácter.

- Colateral.
- Capital.

Las que se definen de la siguiente manera:

- **¿Qué es la capacidad de pago?** Es la evaluación de si tus ingresos son suficientes para cubrir los compromisos actuales más el nuevo préstamo.
- **¿Qué es el comportamiento de pago?** El historial crediticio es un registro del comportamiento que has tenido con tus créditos previos. Es como un récord de notas que dice que tan bien o mal has estado pagando tus créditos. Dicho comportamiento es el que toma en cuenta el banco a la hora de decidir el otorgarmiento de nuevo un crédito.
- **¿Qué es el carácter?** Es la probabilidad de que siempre cumplas con el compromiso de pago, aun cuando esto signifique que te quedes sin liquidez mensual o aunque tu situación financiera empeore en algún momento después de desembolsado el préstamo.
- **¿Qué es el colateral?** El colateral es también conocido como garantía. La garantía es un aval que respalda tu compromiso de pago y en caso de no poder cumplir, la misma es ejecutada como pago de lo adeudado.
- **¿Qué es el capital?** El capital o patrimonio es una resta del valor de venta de tus activos menos todo lo que debes o pasivos.

El análisis manual de dichos métodos se vuelve cada vez más obsoleto, propenso a errores y requiere mucho tiempo y claramente el tiempo es dinero. Afortunadamente, esta tarea se puede automatizar con el poder del aprendizaje automático y casi todos los bancos comerciales lo hacen hoy en día. En este proyecto, crearemos un predictor automático de aprobación de tarjetas de crédito utilizando técnicas de aprendizaje automático.

Se hará uso de una dataset llamado **Credit Card Approval**, es un dataset público que se puede consultar en http://archive.ics.uci.edu/ml/datasets/credit+approval perteneciente al repositorio: **UCI Machine Learning Repository**.

1.1. Objetivo

El objetivo del presente proyecto es utilizar algoritmos de machine learning para aprobar o rechazar solicitudes de tarjetas de crédito, por lo cual nos enfrentamos a un problema de clasificación binaria.

- Primero, se comienza cargando y analizndo el conjunto de datos.
- Se verá que el conjunto de datos tiene una combinación de características numéricas y categóricas, que contiene valores de diferentes rangos, además de que contiene una cantidad de datos faltantes.
- Se preprocesa el conjunto de datos para asegurarse de que el modelo de aprendizaje automático que se eligió pueda hacer buenas predicciones.
- Una vez que el conjunto de datos esté en buena forma, se hará un análisis exploratorio de datos.
- Finalmente, se creará un modelo de aprendizaje automático que puede predecir si se aceptará la solicitud de un cliente para una tarjeta de crédito.

El dataset a analizar se encuentra alojado en formato CSV en un bucket de Google Cloud Storage perteneciente al servicio de Google Cloud Platform (GCP), esto con la finalidad de que el dataset sea portable, replicable y puesta en producción de una forma rápida. Se puede consultar en https://storage.googleapis.com/proyecto-ml-fciencias/Credit_Card.csv.

Figura 1.2: GCP.

La metodología utilizada será la CRISP-DM que se puede visualizar en la figura 1.3.

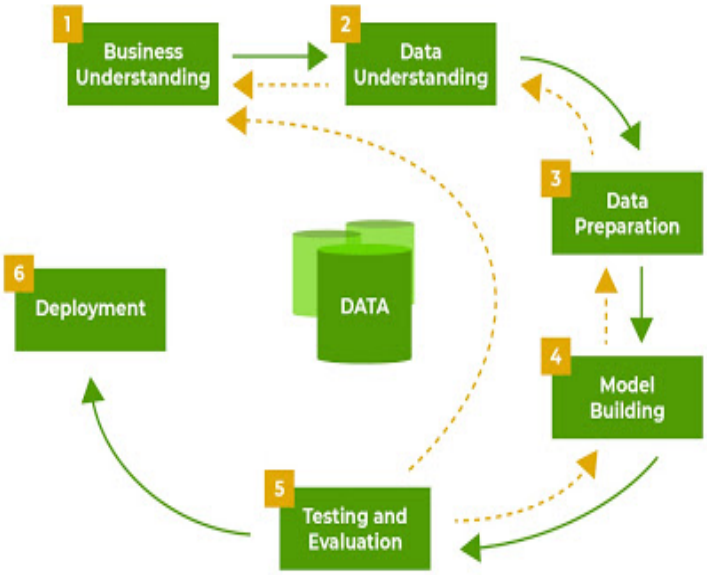

Figura 1.3: CRISP-DM.

El primer paso de la metodología CRISP-DM es entender el negocio, este caso corresponde al otorgamiento de tarjetas de credito por parte de algún banco, es decir, una línea de crédito que puede ser utilizada mediante una tarjeta.

Los demás pasos se detallarán en los capítulos siguientes.

Capítulo 2

Marco Teórico Parte I

Como se mencionó en la introducción, se hará uso del dataset llamado **Credit Card Approval**, es un dataset público que se puede consultar en http://archive.ics.uci.edu/ml/datasets/credit+approval perteneciente al repositorio: **UCI Machine Learning Repository**.

Este archivo se refiere a solicitudes de tarjetas de crédito. Todos los nombres y valores de los atributos se han cambiado a símbolos sin sentido para proteger la confidencialidad de los datos, en otas palabras el contribuyente del conjunto de datos ha anonimizado los nombres de las características.

El término dataset es una representación de datos conocido en español como conjunto de datos o serie de datos. Esta representación viene dada en una única tabla de base de datos o matriz de datos y se podría definir como una representación de datos residentes en memoria. El conjunto de datos se almacena por columnas y filas, siendo cada columna una variable (atributo como color, talla, edad, género...) y cada fila representa a un miembro determinado del conjunto de datos. La unión de todas las filas y columnas proporciona el conjunto de todos los valores que pueden tener las variables. El tipo de datos que pueden representar puede ser tanto texto, números o multimedia, por ejemplo.

Este conjunto de datos es interesante porque hay una buena combinación de atributos: continuo, entero, nominal con un número reducido de valores y nominal con un número mayor de valores. También como en cualquier ejemplo en la práctica el dataset utilizado cuenta con algunos valores faltantes.

La base se conforma de 690 registros donde cada registro es la solicitud de un cliente por una tarjeta de crédito.

2.1. Atributos

2.1.1. Entendimiento y Preparación de los Datos

La base cuenta con 16 atributos los cuales se enlistan a coninuación con su tipo de dato:

- A1: b, a.
- A2: continuo.
- A3: continuo.
- A4: u, y, l, t.
- A5: g, p, gg.
- A6: c, re, cc, i, j, k, m, r, q, w, x, e, aa, ff.
- A7: v, h, bb, j, n, z, dd, ff, o.
- A8: continuo.
- A9: t, f.
- A10: t, f.
- A11: continuo.
- A12: t, f.
- A13: g, p, s.
- A14: continuo.
- A15: continuo.
- A16: +, - (atributo de clase)

Se puede observar que el conjunto de datos no cuenta con los nombres de las columnas. Como se mencionó anteriormente los datos son confidenciales ya que contienen información sensible de los clientes, en el mismo repositorio se identifica la fuente como confidencial, por lo tanto no existen nombres de las columnas pero según la página

https://nycdatascience.com/blog/student-works/credit-card-approval-analysis/
las columnas se pueden infererir quedano de la siguiente manera respectivamente:

- Male
- Age
- Debt
- Married
- BankCustomer
- EducationLevelw
- Ethnicity
- YearsEmployed
- PriorDefault
- Employed
- CreditScore
- DriversLicenset
- Citizen
- ZipCode
- Income
- Approved

Nos encontramos frente a un problema de aprendizaje supervisado, la variable objetivo, es decir, la variable a predecir es la columna A16 o Approved la cual se conforma de dos categorías: +, - (atributo de clase) donde:

- + Se le otorgó la tarjeta de crédito.
- - No se le otorgó la tarjeta de crédito.

El utilizar métodos de aprendizje automático o machine learning para predecir el otrogamiento de tarjetas de crédito conduce a un proceso más rápido que al hacerlo de manera manual, dicha tarea se puede automatizar y en cuestión de minutos el cliente puede saber si es candidato a recibir una tarjeta de crédito. Lo cual representa un ahorro para los bancos optimizando tiempo y costos.

Lo primero que se puede hacer es eliminar los atributos que no sean relevantes para el análisis, tal es el caso de ZipCode y DriversLicense que son el código postal y la licencia de conducrir respectivamente por lo cual ya no se considerarán desde ahora.

2.2. Valores Faltantes

Es común trabajar con bases de datos que no están complentas, es decir, se tiene la presencia de datos faltantes, esto suele ser muy común por diversas razones, es por ello que se tienen que aplicar técnicas de imputación e inclusive eliminar los registros que incluyen datos faltantes.

En la figura 2.1 se puede observar un mapa de calor de la base en donde se han píntado en color amarillo los datos faltantes.

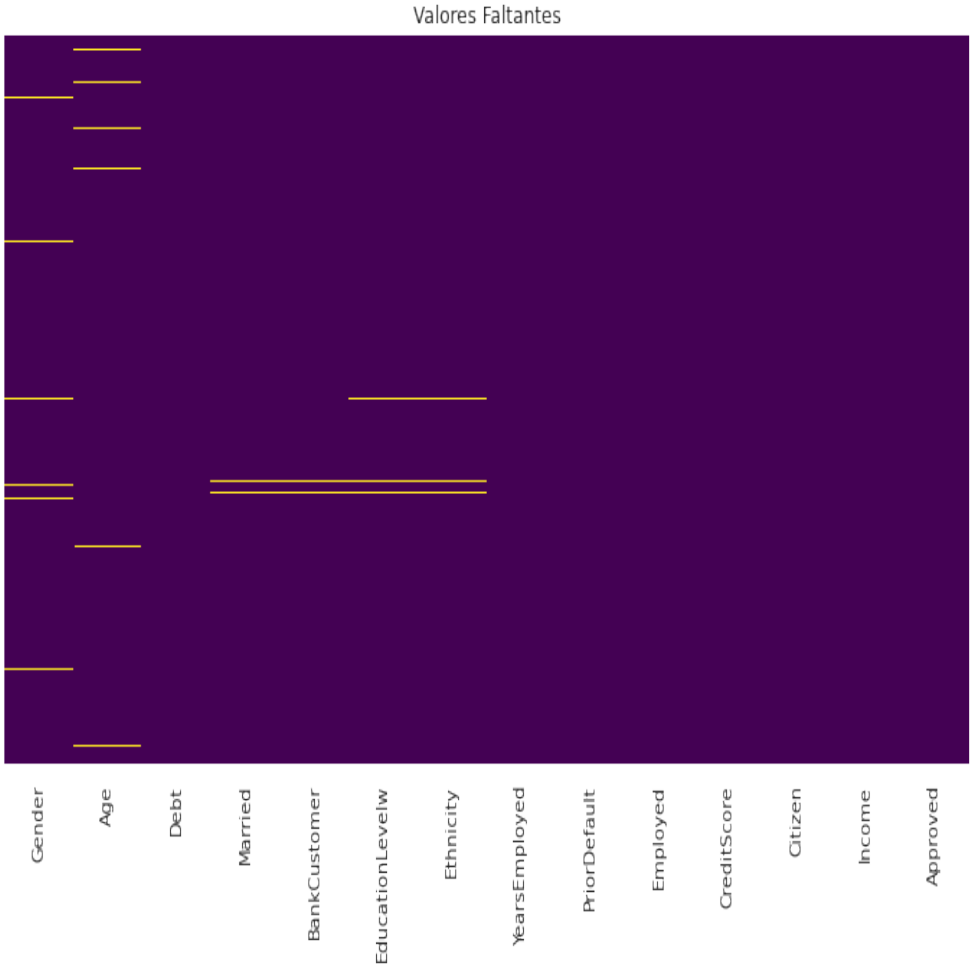

Figura 2.1: Mapa Datos Faltantes.

Ignorar los datos faltantes puede afectar en gran medida el rendimiento de un modelo de aprendizaje automático, el modelo puede perder información sobre el conjunto de datos que puede ser útil para su entrenamiento.

	Total	Porcentaje %
Age	12	1.739130
Gender	12	1.739130
Ethnicity	9	1.304348
EducationLevelw	9	1.304348
BankCustomer	6	0.869565
Married	6	0.869565
Approved	0	0.000000
Income	0	0.000000
Citizen	0	0.000000
CreditScore	0	0.000000
Employed	0	0.000000
PriorDefault	0	0.000000
YearsEmployed	0	0.000000
Debt	0	0.000000

Figura 2.2: Número Datos Faltantes.

El total y porcentaje de datos falatntes se pueden visualizar en la figura 2.2 y notamos que el porcentaje no es muy alto entonces, para evitar este problema, se van a imputar los valores faltantes con una estrategia llamada imputación de la media que aplica únicamente para variables numéricas. La técnica se trata de imputar los valores faltantes con la media del atributo.

Para variables categóricas se van a imputar los valores faltantes con los valores más frecuentes presentes en las columnas respectivas. Esta es una buena práctica cuando se trata de imputar valores faltantes para datos categóricos en general.

Haciendo uso de estas técnicas de imputación se tiene una base de datos completa, la cual se puede observar en el mapa de calor en la figura 2.3.

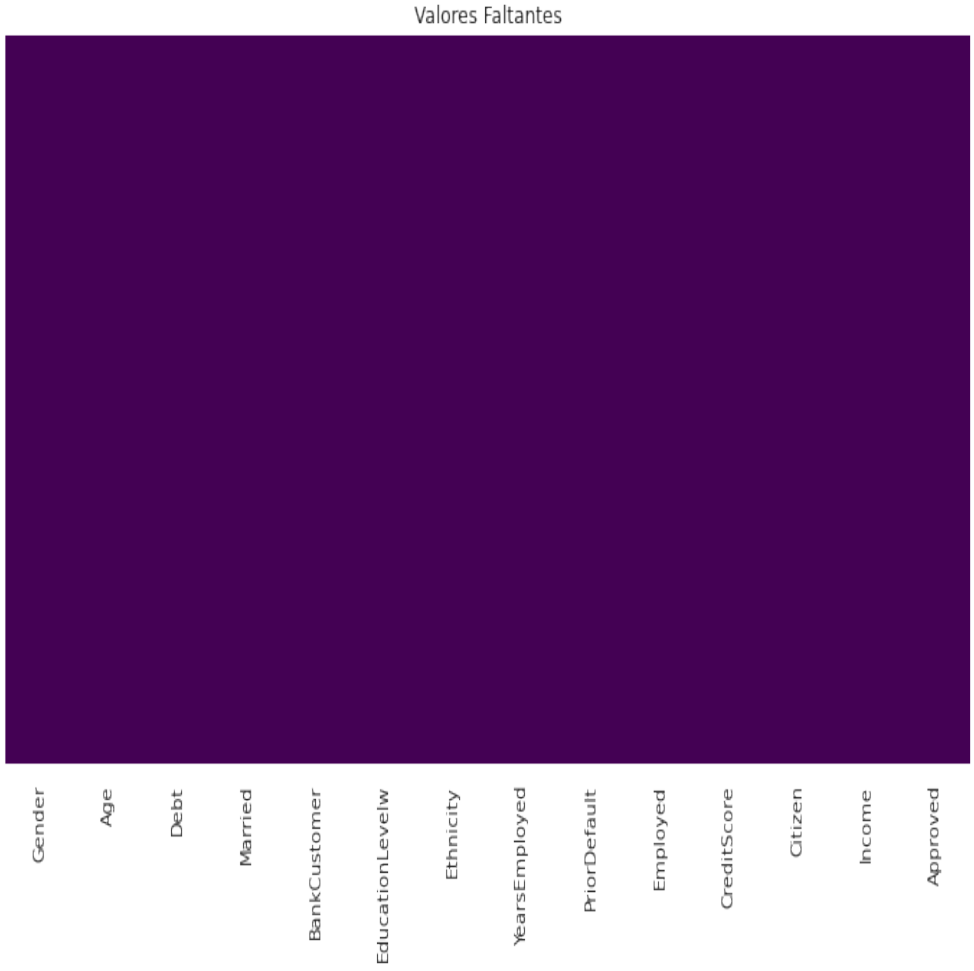

Figura 2.3: Mapa Datos Faltantes Limpio.

Capítulo 3

Marco Teórico Parte II

En este capítulo se detallan los métodos utilizados, una síntesis de cómo funciona el algoritmo.

El aprendizaje supervisado es cuando se tienen variables de entrada X y una variable de salida Y y se utiliza un algoritmo para aprender la función f de mapeo de la entrada a la salida, midiendo el ajuste mediante una función de pérdida.

$$Y = f(X)$$

El objetivo es aproximar la función de mapeo f tan bien que cuando se tengan nuevos datos de entrada X se puedan predecir las variables de salida Y para esos datos.

3.1. Regresión Logística

La Regresión Logística es un algoritmo de aprendizaje supervisado y se utiliza para clasificación. En términos de este problema particular, este método trata de explicar la probabilidad de que se otorgue o no una tarjeta de crédito a un cliente en función de una seride de variables explicativas. El valor principal de éste método reside en la buena interpretabilidad, otra ventaja reside en la modelización de probabilidades y en el hecho de que los modelos de regresión logística son menos sensibles a outliers.

La variable dependiente presenta dos categorías que representan la ocurrencia y la no ocurrencia del acontecimiento definido, en este caso la aprobación y la no aprobación de la tarjeta de crédito condificándose con los valores uno y cero respectivamente. En lo que se refiere a las variables explicativas o predictoras, pueden ser tanto numéricas como categóricas.

El modelo expresa la variable dependiente como la ocurrencia o no de un acotencimiento en términos de probabilidad, haciendo uso de la función logística para estimar la probabilidad de que ocurra el acontecimiento mediante la formulación:

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}} \tag{3.1}$$

siendo π_i la probabilidad de pertenecer a las clase buena y x_i las variables explicativas o características de un cliente.

Puesto que el modelo anterior no es lineal respecto a las variables independientes, se considera la inversa de la función logística, a lo que se llama **logit**, definiéndose como en cociente entre la probabilidad de que ocurra un acontecimiento y la probabilidad de que no ocurra, que es su complemento, como se puede observar:

$$g(x) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{3.2}$$

Dicha formulación hace más clara la interpretación del modelo y de sus coeficientes, que reflejan el cambio en el logit corresponiente a un cambio unitario en la variable independiente.

La probabilidad π_i obtenida por la ecuación 3.2 es el límite de la clasificación. El cliente es propenso a que no se le otorgue la tarjeta si es mayor a 0.5 o no propenso si es menor a 0.5. En este tipo de aplicación financiera, se puede asociar el término $\beta_i x_i$ a la calidad crediticia del cliente.

Se puede llegar a la regresión logística partiendo de una regresión lineal. Considere una regresión lineal con una variable independiente X y una variable dependiente y con las siguientes características:

$$y = \alpha + \beta \cdot X \tag{3.3}$$

- $y \in \{0, 1\}$
- $X \in [-\infty, \infty]$

Si se considera que P es la probabilidad condicionada de éxito o de fracaso condicionada a la presencia de la variable X.

- $P \in [0, 1]$

- $P = \alpha + \beta \cdot X$

$$\frac{P}{1-P} = \alpha + \beta \cdot X \in [0, +\infty]$$

$$\ln(\frac{P}{1-P}) = \alpha + \beta \cdot X \in [-\infty, +\infty]$$

$$\begin{cases} \frac{P}{1-P} \in [0,1] \Rightarrow \ln(\frac{P}{1-P}) \in [-\infty, 0] \\ \frac{P}{1-P} \in [1,\infty] \Rightarrow \ln(\frac{P}{1-P}) \in [0, \infty] \end{cases}$$

$$\ln(\frac{P}{1-P}) = \alpha + \beta \cdot X$$

$$\frac{P}{1-P} = e^{\alpha+\beta \cdot X}$$

$$P(y|X) = \frac{1}{1 + e^{-(\alpha+\beta \cdot X)}} \tag{3.4}$$

- Si $a + bX$ es muy pequeño (negativo), entonces P tiende a 0.
- Si $a + bX = 0$, $P = 0{,}5$
- Si $a + bX$ es muy grande (positivo), entonces P tiende a 1.

En la regresión logística múltiple el caso es análogo al tener más de una variable independiente:

$$P(y|X) = \frac{1}{1 + e^{-(\alpha+\sum_{i=1}^{n} \beta \cdot X_i)}} \tag{3.5}$$

Donde P es la probabilidad condicional de éxito o de fracaso condicionada la presencia del vector de caracteísticas X.

La idea del algoritmo es encontrar los parámetros α y β_i que mejor se ajusten al conjunto de datos, para hacerlo se implementa el método de la máxima verosimilitud para la regresión logística, en donde se define la función de entorno $L(b)$ (también llamada función de pérdida).

$$L(\beta) = \sum_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i} \tag{3.6}$$

En donde se calculan las probabilidades para cada observación

$$Pi = P(x_i) = \frac{1}{1 + e^{-\sum_{j=0}^{k} \beta_j \cdot x_i}} \tag{3.7}$$

Se calcula la matriz diagonal W

$$W = diag(P_i \cdot (1 - P_i))_{i=1}^{n} \tag{3.8}$$

Se utilizan distinto algoritmos para la solución, uno de ellos es el método de Newton-Raphson.

$$\beta_{n+1} = \beta_n - \frac{f(\beta_n)}{f'\beta_n}$$

$$f(X) = X(Y - P)$$

$$f'(X) = XWX^T$$

Cabe mencionar que también se puede definir un umbral para la clasificación.

$$\varepsilon \in (0, 1), Y_p = \begin{cases} 0 & si\ p \leq \varepsilon \\ 1 & si\ p > \varepsilon \end{cases}$$

Como se puede observar, la regresión logística lleva en el núcleo de su método la funcón sigmoide; esta función es una curva en forma de S, que puede tomar cualquier número real y dar como resultado cualquier número entre cero y uno.

En el caso de machine learning, la función sigmoide relaciona la variable dependiente con las variables independientes; es una curva que puede tomar cualquier valor entre 0 y 1 y nunca valores por fuera de estos límites, así la ecuación que define la función sigmoidea es:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.9}$$

Una de las particularidades de esta función es de que cumple con la ecuación diferencial $f'(x) = f(x)(1 - f(x))$ por lo cual es fácil de evaluar su derivada para los problemas de optmización que involucran la primera derivada.

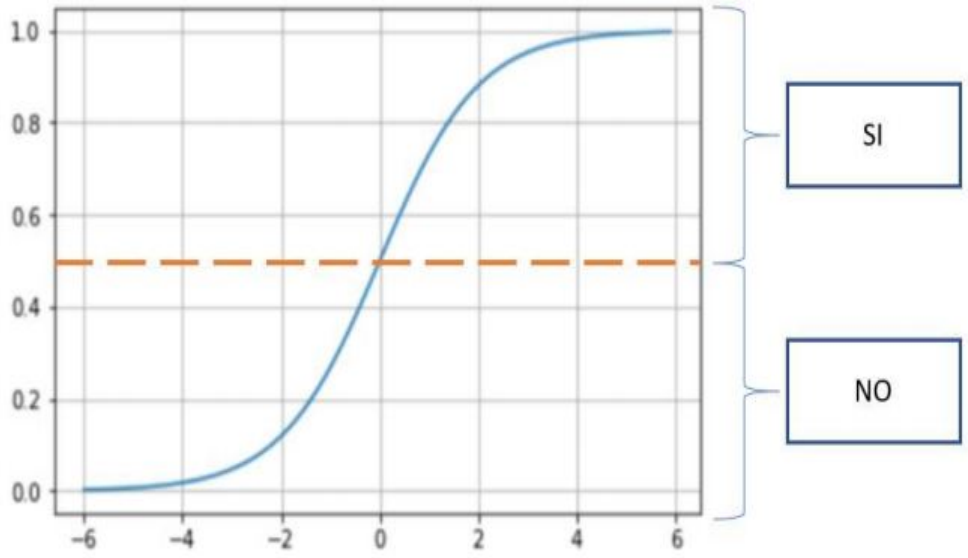

Figura 3.1: Función sigmoide.

3.1.1. Ventajas de la regresión logística

- Modelo de fácil interpretación.

- Costo computacional bajo.

La regresión logística es una técnica muy empleada por los científicos de datos debido a su eficacia y simplicidad, es los algortimos más sencillos de aplicar ya que los resultados son altamente interpretables. No se necesita contar con grandes recursos computacionales. Siendo esta una de sus principales ventajas respecto a otras técnicas más complejas. El peso de cada una de las características determina la importancia que tiene en la decisión final.

El funcionamiento de la regresión logística, al igual que la regresión lineal, es mejor cuando se utilizan atributos relacionados con la de salida. También es importante eliminar las características que muestran una gran multicolinealidad entre sí. Por lo que la selección de las características previa al entrenamiento del modelo es clave. Siendo aplicables las técnicas de ingeniería de características también utilizadas en la regresión lineal.

3.1.2. Desventajas de la regresión logística

En cuanto a sus desventajas se encuentra la imposibilidad de resolver directamente problemas no lineales. Esto es así porque la expresión que toma la decisión es lineal. Por ejemplo, en el caso de que la probabilidad de una clase se reduzca inicialmente con una característica y posteriormente suba no puede ser registrado con un modelo logístico directamente. Siendo necesario transforma esta característica previamente para que el modelo puede registrar este comportamiento no lineal. En estos casos es mejor utilizar otros modelos.

Una cuestión importante es que la variable objetivo esta ha de ser linealmente separable. En caso contrario el modelo de regresión logística no clasificará correctamente. Es decir, en los datos han de existir dos regiones con una frontera lineal.

Otra desventaja es la dependencia que muestra en las características. Ya que no es una herramienta útil para identificar las características más adecuadas. Siendo necesario identificar estas mediante otros métodos

Finalmente, la regresión logística tampoco es uno de los algoritmos más potentes que existen. Pudiendo ser superado fácilmente por otros más complejos.

Capítulo 4

Resultados

4.1. Análisis Exploratorio de Datos

En la presente sección se realiza un análisis exploratorio de datos.

El principal propósito del análisis exploratorio de datos es tener una idea de cómo son nuestros datos, antes de decidir qué técnica de Ciencia de Datos o de Machine Learning se usará.

Se debe entender su contenido, cuáles son las variables más relevantes y cómo se relacionan unas con otras, comenzar a ver algunos patrones y extraer conclusiones acerca de todo este análisis.

Y todo esto es precisamente el análisis exploratorio de datos, que es en resumen una forma de entender, visualizar y extraer información relevante del conjunto de datos para poder decidir cuál será la ruta o técnica más adecuada para su posterior procesamiento.

Y este es siempre el paso cero en cualquier proyecto de Machine Learning o Ciencia de Datos.

En la figura 4.1 se han graficado las distribuciones de frecuencia de las variables categóricas, con la finalidad de dectectar cuáles son las categorías con el mayor número de observaciones.

En la figura 4.2 se han graficado las tablas de contingencia normalizadas desgaregadas por la variable objetivo.

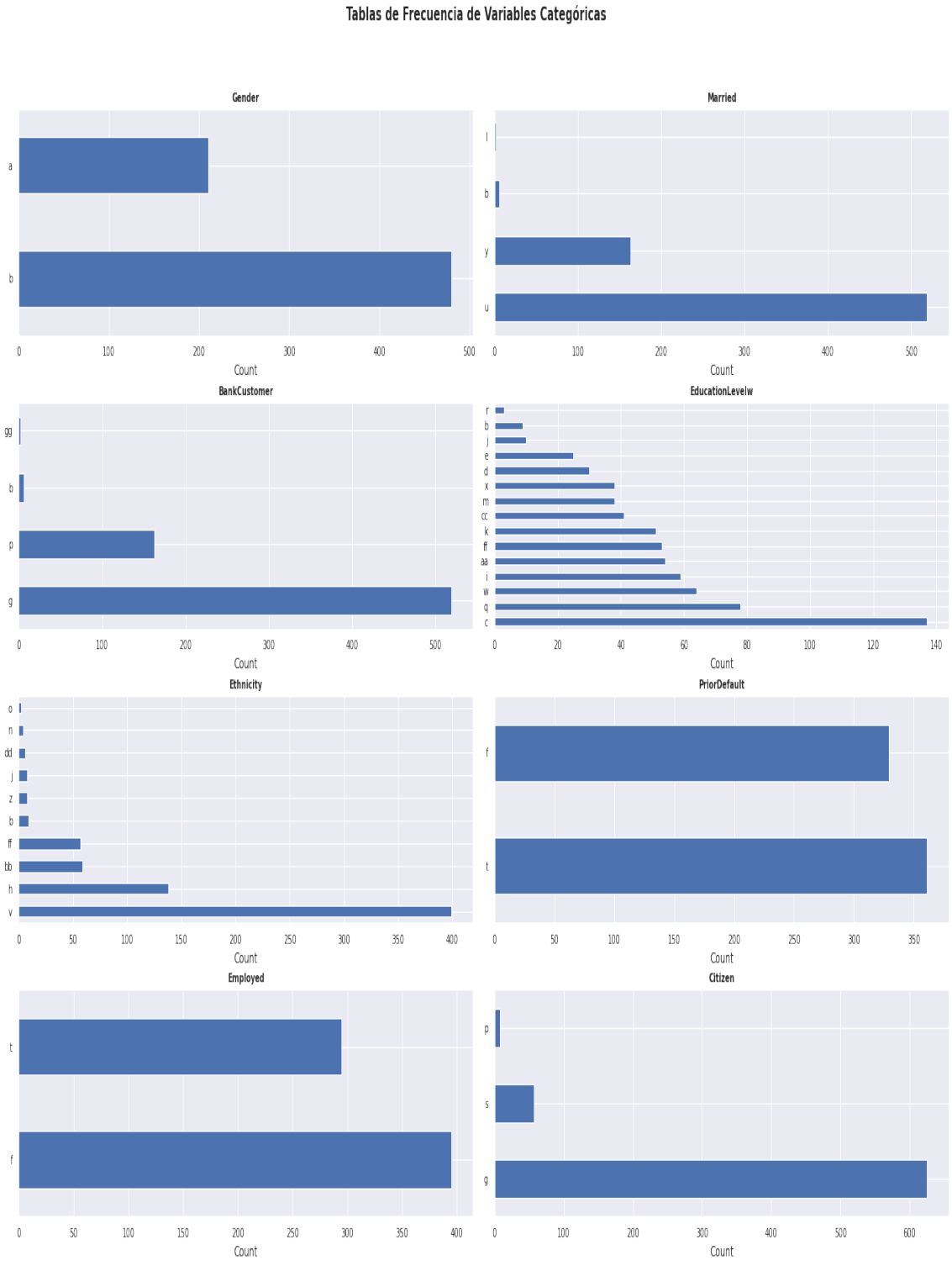

Figura 4.1: Tablas de Frecuencia de Variables Categóricas.

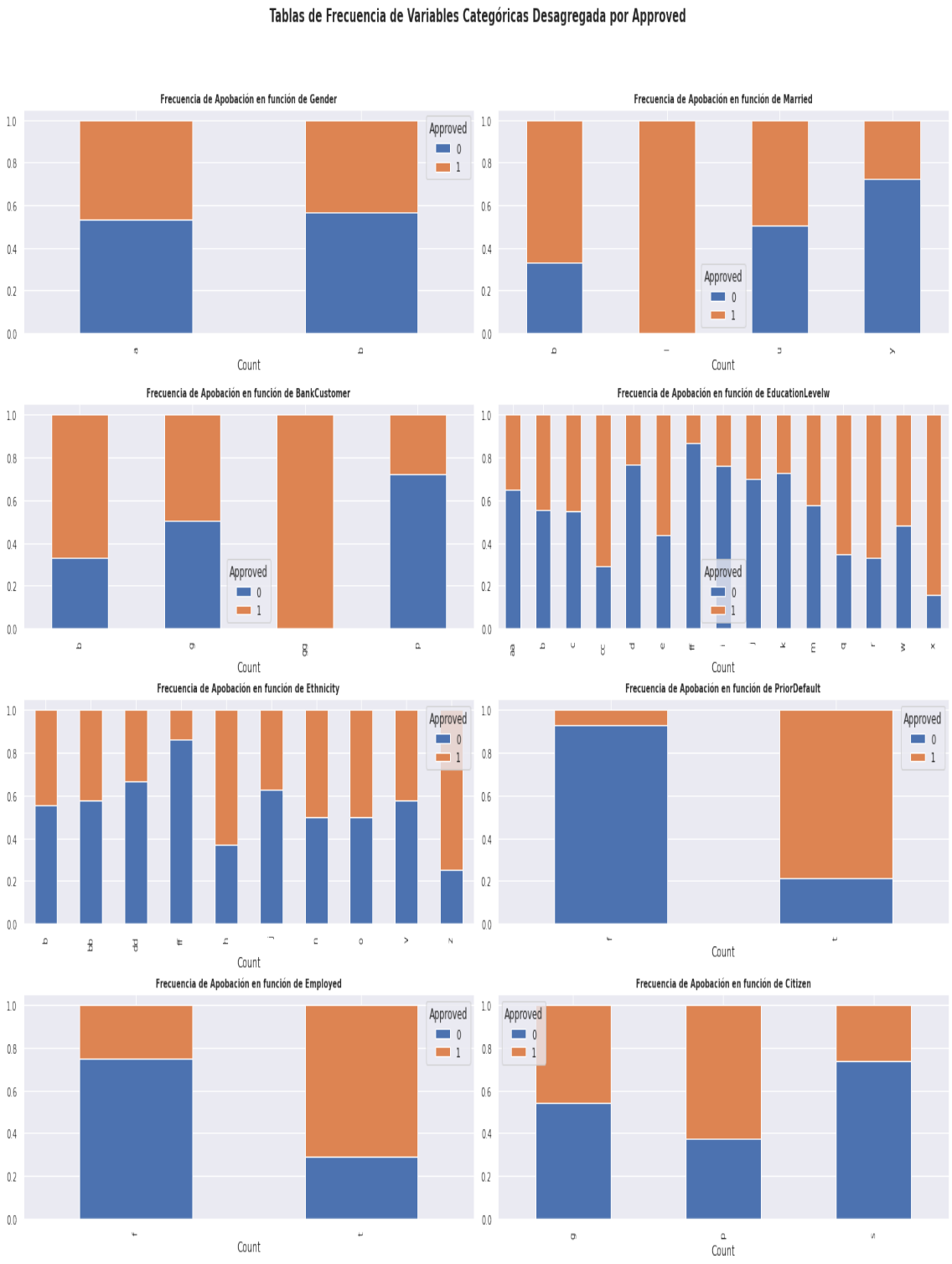

Figura 4.2: Tablas de Frecuencia de Variables Categóricas Desagregada por Approved.

	count	mean	std	min	25%	50%	75%	max
Age	690.0	31.118841	11.852887	13.0	22.000	28.00	37.0000	80.0
Debt	690.0	4.758725	4.978163	0.0	1.000	2.75	7.2075	28.0
YearsEmployed	690.0	2.223406	3.346513	0.0	0.165	1.00	2.6250	28.5
CreditScore	690.0	2.400000	4.862940	0.0	0.000	0.00	3.0000	67.0
Income	690.0	1017.385507	5210.102598	0.0	0.000	5.00	395.5000	100000.0
Approved	690.0	0.444928	0.497318	0.0	0.000	0.00	1.0000	1.0

Figura 4.3: Estadísticos Básicos.

En la figura 4.3 se tienen los estadísticos básicos de las variables numéricas como media, desviación estándar, mínimo, máximo y primer, segungo y tercer cuartil. Donde se puede observar que la edad promedio es de 31 años y un ingreso de 1017.385507, donde también se han graficado sus distribuciones en las figuras 4.4 y 4.5.

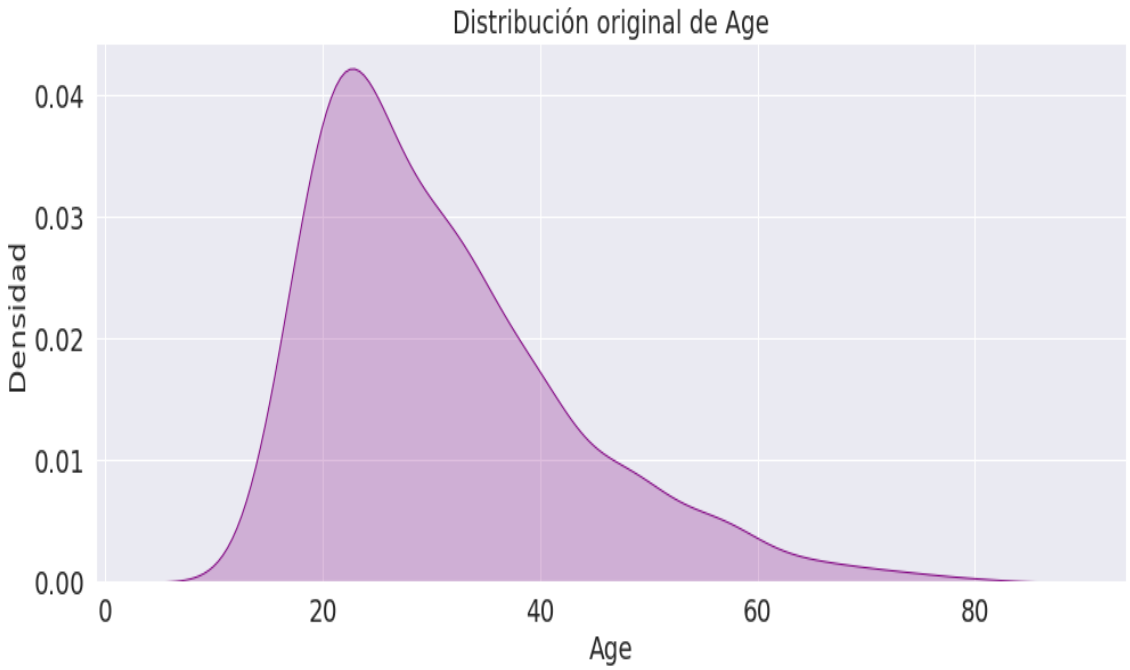

Figura 4.4: Distribución de la Edad.

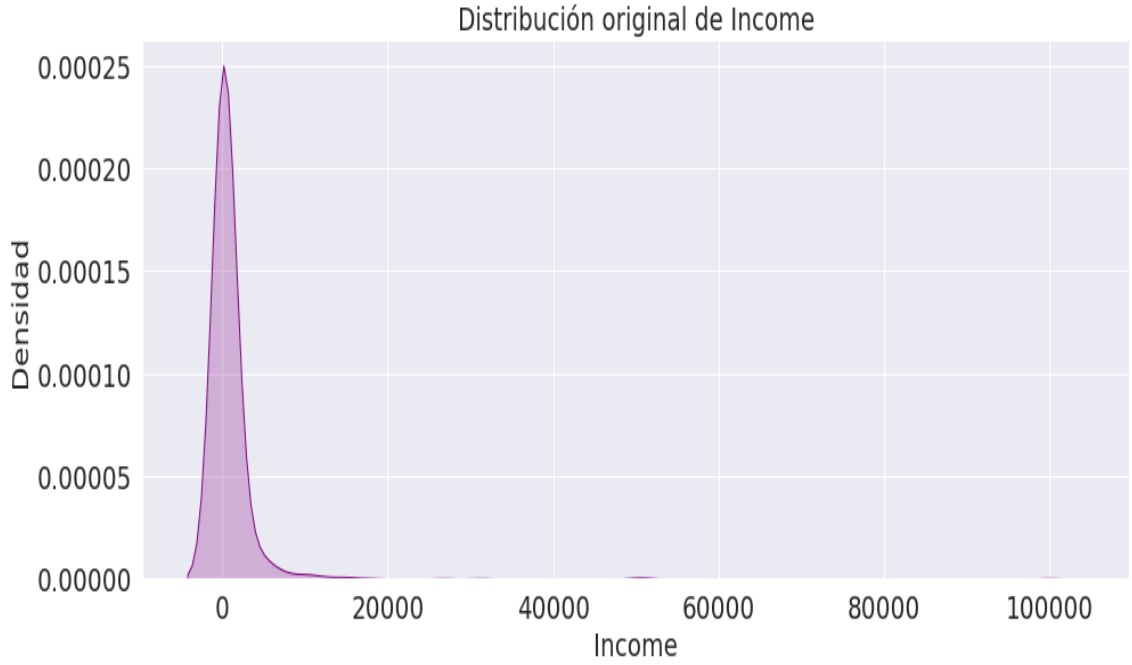

Figura 4.5: Distribución del Ingreso.

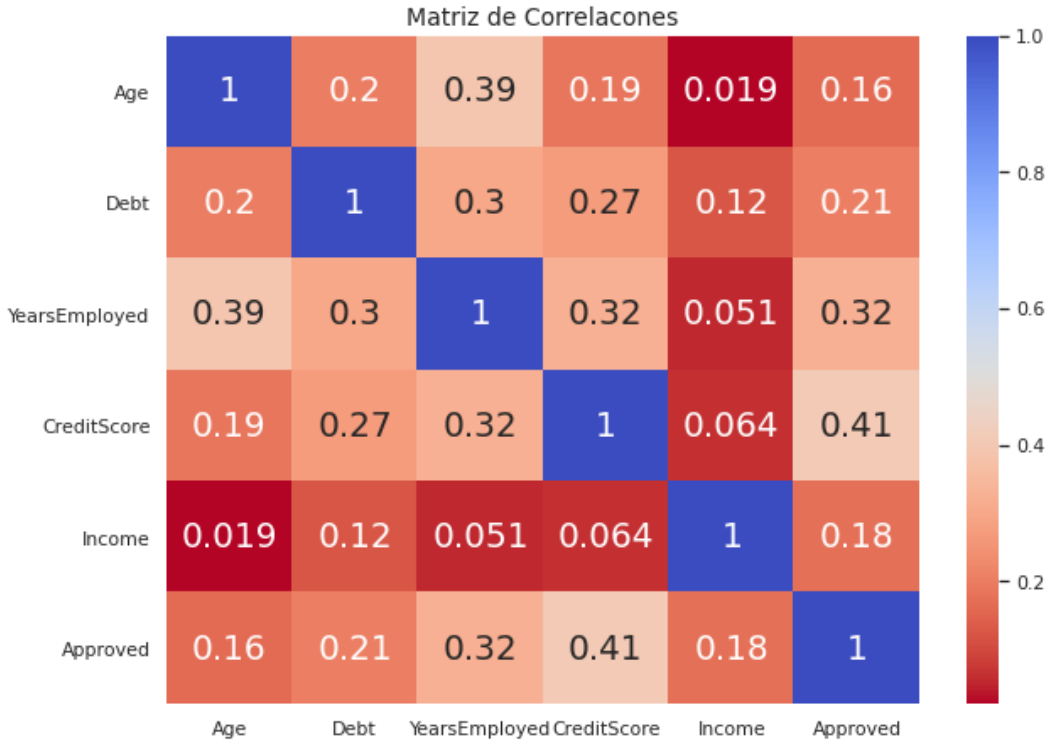

Figura 4.6: Matriz de Correlaciones.

En la figura 4.6 se muestra la matriz de correlaciones de las variables numéricas en la que se puede observar que no se presentan problemas de multicolinealidad.

En la tabla 4.1 se tiene la frecuencia obsuluta y relativa de la variable objetivo, donde se puede observar que no se presenta un problema de desbalanceo de clases.

Cuadro 4.1: Accuracy.

Valor	Total	Porcentaje
No Aprobada	3837	55.51 %
Aprobada	307	44.49 %

Figura 4.7: Distribución de la Variable Objetivo.

4.1.1. Construcción del Modelo

Una vez limpia la base y realizado un análisis exploratorio, se procede al tratamiento de las variables categóricas, es decir, codifcar características categóricas como una matriz numérica para poder aplicar el algoritmo ya que solamente recibe valores numéricos, esto se puede hacer una una función de la librería pandas llamada *get_dummies*.

Una vez hecho esto ahora sí se procede a separar los datos en datos de entrenamiento con los cuales se va a entrenar el modelo y datos de test con los que se va a evaluar el modelo. Una de las convenciones es separar los datos de la siguiente manera:

- Entrenamiento: 70 % Un subconjunto para entrenar el modelo, este subconjunto se utilizará en la práctica para ajustar los parámetros en los algoritmos.

- Test: 30 % Un subconjunto para probar el modelo entrenado, este subconjunto de datos se utilizará en la práctica para proporcionar una evaluación imparcial de un último ajuste del modelo y análisis de los resultados.

A los datos de entrenamiento se les aplicó el modelo de regresión logística con los hiperparámetros definidos por default salvo $max_iter = 1000$, dicho algoritmo ya se encuentra en la librería sklearn de python.

Al entrenar el modelo la primera métrica que se considera es el Accuracy, que es el total de predicciones correctas entre el total de predicciones.

A la hora del uso de los modelos es de gran importancia el uso de los conjuntos de datos que se están usando. Es aquí donde entran en práctica los términos overfitting o underfitting conocidos en castellano como sobreajuste y subajuste.

- Subajuste: se dice de un modelo estadístico o un algoritmo de aprendizaje automático en el que se obtiene un bajo rendimiento cuando al no proporcionar los suficientes datos no puede captar la tendencia subyacente de los datos. El modelo probablemente realizará un gran número de predicciones erróneas.

  Causas

  - Modelo demasiado simple, no es capaz de aprender relación entre datos de entrada y salida
  - Falta de datos, si no se cuenta con los datos necesarios.
  - Atributos relevantes insuficientes, es posible tener todos los datos pero que surja la necesidad de transformar estos para una mejor comprensión del modelo entre la relación de entrada salida

- Sobreajuste: se dice de un modelo estadístico o un algoritmo de aprendizaje automático en el que se obtiene un bajo rendimiento cuando se entrena con un número demasiado alto de datos. Cuando un modelo se entrena con datos de más, comienza a aprender del noise o ruido y de las entradas de datos inexactos y no es capaz de categorizar los datos correctamente debido a demasiados detalles. Causas:

  - Modelo demasiado complejo, podrá aprender muchos de los datos de memoria.

- Los datos tienen noise, es decir, como si hubiera valores atípicos y errores en los datos.
- El tamaño de los datos utilizados para los datos de entrenamiento puede que no sean suficientes.

4.1.2. Evaluación del Modelo

Al utilizar el algoritmo de regresión logística los resultados al evaluar el modelo con los siguientes:

En la tabla 4.4 se pueden observar el accuracy tanto para los datos de entrenamiento como los de test, donde no existe problema de subajuste y sobrajuste.

Cuadro 4.2: Accuracy.

Datos	Accuracy
Entrenamiento:	86.7 %
Test:	89.3 %

La matriz de confusión es una herramienta que nos muestra el desempeño de un algoritmo de clasificación, describiendo cómo se distribuyen los valores reales y las predicciones hechas por el modelo mediante 4 distintos casos basados en 4 variables que se comentan a continuación:

- Verdadero positivo (VP): Se predice que es positivo y es verdad, fue clasificado correctamente.
- Verdadero negativo (VN): Se predice que es negativo y es verdad, fue clasificado correctamente.
- Falso positivo (FP): Se predice que es positivo y es falso, fue clasificado incorrectamente.
- Falso negativo (FN): Se predice que es negativo y es falso, fue clasificado incorrectamente.

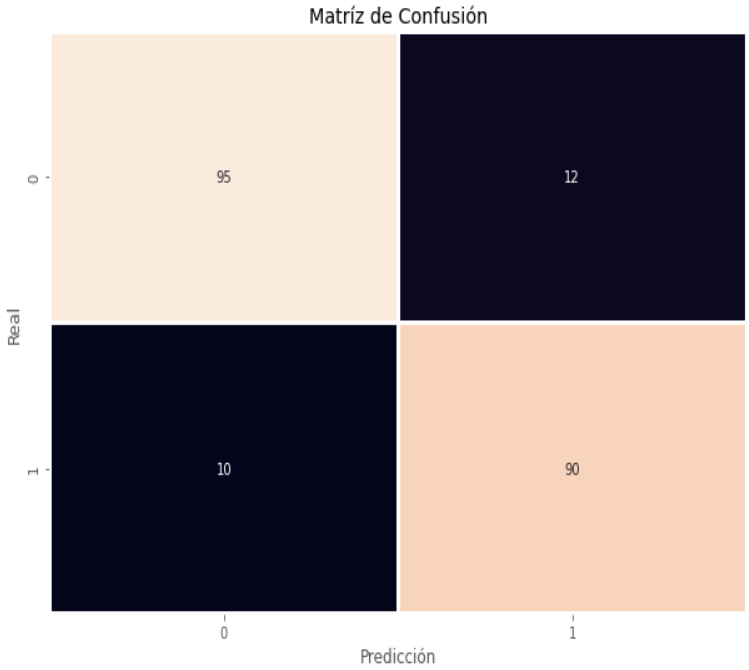

Figura 4.8: Matriz de Confusión.

A partir de la matriz de confusión se pueden obtener la precision, recall y f1-score que se pueden observar en la tabla 4.3. Una buena métrica que pondera la precision y el recall es la f1-score y para ambas clases es bastante alta por lo cual el modelo tiene un ajuste considerablemente bueno, es decir, es capaz de distiguir las dos clases.

Cuadro 4.3: Métricas.

Clase	Precision	Recall	f1-score
0 (No Aprobada)	0.90	0.89	0.90
1 (Aprobada)	0.88	0.90	0.89

Al aplicar una técnica de validación cruzada se obtiene un Accuracy promedio de 92.7 % con una desviación ± 3.8 %

Cuadro 4.4: My first table.

A	B	C
1	2	3
4	5	6

4.1.3. Curva ROC

Una curva ROC (curva de característica operativa del receptor) es una gráfica que muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasifi-

cación. Esta curva representa dos parámetros:

La curva ROC-AUC es una medida de rendimiento para el problema de clasificación en varios umbrales. ROC es una curva de probabilidad y AUC representa el grado o medida de separabilidad. Será a través de la métrica AUC (Area Under the ROC Curve) que representa un valor del área que queda por debajo de la curva ROC, la encargada de comparar unos modelos con otros indicando cuánto es capaz el modelo de distinguir entre clases. Cuanto más alto es el AUC, mejor es el modelo para predecir verdaderos como verdades y falsos como falsos. La curva ROC se traza con la Sensibilidad frente la Especificidad donde la Sensibilidad está en el eje y, y el eje x está compuesto por 1-Especificidad, la evaluación de esta medida se clasifica de la siguiente forma:

- Valor cerca de 1: El modelo es excelente, tiene una gran capacidad de separabilidad. En este caso es perfectamente capaz de distinguir entre la clase positiva y la clase negativa.
- Valor cerca de 0.5: El modelo no tiene ninguna capacidad de separación de clases. Es la peor situación y el modelo no tiene capacidad de discriminación para distinguir entre clase positiva y negativa.

La gráfica de la curva ROC del modelo se puede observar en la figura 4.9 donde se obtiene un AUC de 0.94 bastante aceptable.

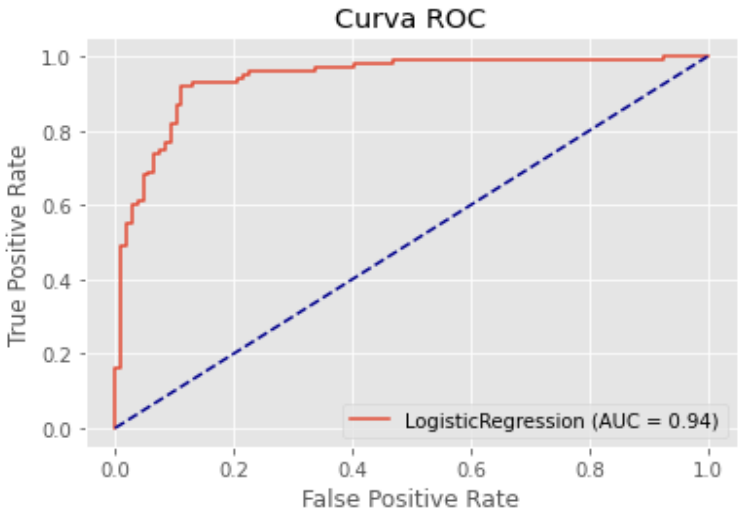

Figura 4.9: Curva ROC.

Capítulo 5

Conclusiones

El machine learning o aprendizaje automático abre para las organizaciones posibilidades sin precedentes que facilitan la automatización, la eficiencia y la innovación. Tal es esl caso en el sector financiero en donde la complejidad en la gestión de los riesgos financieros ha aumentado durante los últimos años, lo cual ha generado, en algunos casos, grandes pérdidas económicas por parte de las instituciones financieras derivadas de fallos en los modelos de medición de riesgos y en los esquemas de toma de decisiones soportados por estos. En especial, el riesgo de crédito supone una de las principales preocupaciones para las entidades. Para reducir este tipo de riesgo, muchas entidades utilizan sistemas automáticos de clasificación de clientes. Como resultado de la evolución de la tecnología, los modelos de machine learning están cada vez más presentes en el sector financiero.

Un modelo de referencia es el de regresión logística, que por diseño conduce a conclusiones que son fáciles de revelar y, por lo tanto,interpretable tanto por los gestores de riesgo de crédito como por los reguladores.

En este proyecto se consideró una dataset público conformado por clientes de un banco a los cuales se les ha aprobado o no una tarjeta de crédito. Se planteó un problema de aprendizaje supervisado en donde la variable objetivo se representa con + si se otorga la tarjeta o − en caso contrario. Una vez limpia la base y al aplicar un análisis exploratorio de datos se aplicó el algoritmo de regresión logísitca ya que se tiene una variable dicotómica. Con respecto al modelo el 89.3 % de los casos fueron correctamente clasificados utilizando un umbral de 0.5. También se calculó la curva ROC y se obtuvo un AUC de 0.94 que es bastante cercano a 1 por lo que se concluye que el modelo es capaz de hacer una buena clasificación.

Cabe mencionar que este algoritmo, al igual que cualquier otro, depende en gran

parte de los datos, por lo que sin una buena base de información de gran valor los sistemas no serán capaces de obtener un modelo confiable, de manera que sus limitaciones están condicionadas por otros factores.

Una de las líneas a seguir es tratar de cambiar los hiperparámetros del modelo para tratar de mejorar las métricas y obtener un mejor modelo para su posterior puesta en producción que es el último paso en la metodología CRISP-DM.

Bibliografía

[1] Wright, R. E. (1995). Logistic regression.

[2] Menard, S. (2002). Applied logistic regression analysis (Vol. 106). Sage.

[3] Carter, C., & Catlett, J. (1987). Assessing credit card applications using machine learning. IEEE Computer Architecture Letters, 2(03), 71-79.

[4] DAVIS, R. H., Edelman, D. B., & Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications. IMA Journal of Management Mathematics, 4(1), 43-51.

[5] Kibria, M. G., & Sevkli, M. (2021). Application of Deep Learning for Credit Card Approval: A Comparison with Two Machine Learning Techniques. International Journal of Machine Learning and Computing, 11(4).

[6] Dhankhad, S., Mohammed, E., & Far, B. (2018, July). Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In 2018 IEEE international conference on information reuse and integration (IRI) (pp. 122-125). IEEE.

[7] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 International Conference on Computing Networking and Informatics (ICCNI) (pp. 1-9). IEEE.

[8] Yee, O. S., Sagadevan, S., & Malim, N. H. A. H. (2018). Credit card fraud detection using machine learning as data mining technique. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(1-4), 23-27.