MACHINE LEARNING: PREDICCIÓN DE IMPAGOS EN PRÉSTAMOS BANCARIOS

Sebastián Segura

CODER HOUSE

ÍNDICE

PROBLEMÁTICA

La presente investigación tiene por objetivo analizar los datos de los clientes de un banco "x", que han contraído una deuda con el mismo. Para ello, se utilizará un dataset que contiene distintas variables propias de cada uno y se manejaran los datos utilizando el lenguaje de programación Python, a fin de poder gráficar y obtener vínculos entre los clientes del banco.

La problemática que va a tratar el proyecto es la mora en el pago de la deuda. Una de las fuentes de ingresos de los bancos comerciales es el interés cobrado por el financiamiento a terceros, siempre y cuando estos paguen. Para intentar reducir ese riesgo al máximo, se intentará detectar relaciones y en cierto punto, se podrá llegar a predecir que clientes pueden caer en default.

De este modo, se desarrollará el contenido que tenga que ver con el dataset utilizado. Luego, se realizará el Análisis Exploratorio de los Datos (EDA) y se expondrán gráficos que permitan visualizar los datos de forma más sencilla, pudiendo obtener diferentes insights. Finalmente, se realizará el proceso de *data wrangling* y *encoding* de los datos, para poder prepararlos y utilizarlos para la realización de un modelo de Machine Learning.

OBJETIVOS DEL MODELO

Con base a todas las variables que presenta el dataset relevado, se propone estudiar: ¿Cuáles son las variables comunes entre los clientes en mora? De esta forma, poder encontrar relaciones entre los mismo y en un futuro tener un trato especifico con los mismos.

En base a esta pregunta objetivo y con el fin de mitigar los clientes que no cumplen con el pago de la deuda, se buscara entrenar un modelo que encuentre las variables comunes entre los clientes en default, para que el sector encargado de otorgar créditos tenga en cuenta los riesgos que supone.

DATASET

Lo primero que debemos tener al hacer un proyecto de data science como este, es un dataset. En este caso, utilizaré uno encontrado en *Kaggle*, que fue descargado y posteriormente subido al notebook para poder ser utilizado.

El tablero se ve así:

id	grade	annual_inc	short_emp	emp_length_num	home_ownership	dti	purpose	term	last_delinq_none	last_major_derog_none	revol_util	total_rec_late_fee	od_ratio	bad_loan
11454641	A	100000.0	1	1	RENT	26.27	credit_card	36 months	1	NaN	43.2	0.0	0.160624	0
9604874	A	83000.0	0	4	OWN	5.39	credit_card	36 months	0	NaN	21.5	0.0	0.810777	0
9684700	D	78000.0	0	11	MORTGAGE	18.45	debt_consolidation	60 months	1	NaN	46.3	0.0	0.035147	1
9695736	D	37536.0	0	6	MORTGAGE	12.28	medical	60 months	0	NaN	10.7	0.0	0.534887	1
9795013	D	65000.0	0	11	MORTGAGE	11.26	debt_consolidation	36 months	0	NaN	15.2	0.0	0.166500	0

A través de algunas funciones, se expondrá información del dataset que nos permitirá conocer los tipos de datos y distintas métricas estadísticas. Para comenzar veamos una breve explicación de a que refiere cada una de las quince (teniendo en cuenta el ID, que más tarda será eliminado) variables:

Variable: GRADE

Esta columna es una clasificación que otorga el banco a los clientes, yendo desde la A (como la mejor calificación) hasta la F (la peor). Es una variable de tipo categórica y no tiene valores nulos.

Variable: ANNUAL_INC

Esta columna expresa el ingreso anual declarado por los clientes. Es de tipo numérica y puede tomar valores con coma.

Variable: SHORT_EMP

Esta variable hace referencia el tiempo en que el cliente es empleado. Es una variable binaria, sin valore nulos, que si toma el valor 1, el, cliente es empleado hace menos de un año y si toma el valor 0, hace más de un año.

Variable: EMP_LENGHT_NUM

Esta columna nos dice hace cuanto el cliente es empleado. Es una variable numérica, sin valores nulos. Los valores van desde 0 a 10. Siendo 0 menor a un año de empleado y 10, diez o más años.

Variable: HOME_OWNERSHIP

Esta variable categórica nos expone que tipo de método usa el cliente como tenencia del inmueble donde viven, que puede ser a través de hipoteca, alquiler o que sean dueños del inmueble. En este caso posee valores nulos, que serán arreglados más adelante.

Variable: DTI

El DTI (deuda a ingreso) es un ratio que calcula el pago mensual de deuda que el cliente hace, dividido con el ingreso mensual que tiene. Si el cliente tiene un DTI mayor a uno quiere decir que esta endeudado por sobre sus posibilidades. En este caso está multiplicado por 100

para expresarlo como porcentaje. En este caso posee valores nulos, que serán arreglados más adelante.

Variable: PURPOSE

Esta variable es un motivo brindado por cliente para el cual se destinará el préstamo. Es categórica sin valores nulos.

Variable: TERM

La variable term es el plazo en el que cliente va a pagar la deuda. Puede tomar los valores 36 o 60 meses.

Variable: LAST_DELINQ_NONE

Esta columna contiene una variable binaria. Si el valor es uno significa que el cliente tiene antecedentes, en caso de que sea 0, no tiene.

Variable: LAST_MAJOR_DEROG_NONE

Variable binaria, que expresa si el cliente tiene más de noventa días de mala calificación. En este caso posee valores nulos, que serán arreglados más adelante.

Variable: REVOL_UTIL

Esta variable numérica, que puede tomar valores con coma, expresa el porcentaje de deuda que esa siendo utilizado por el cliente con respecto al total que tiene disponible.

Variable: TOTAL_REC_LATE_FEE

Esta variable numérica expresa las comisiones pagadas de forma tardía por el cliente. Es tomada como infracciones.

Variable: OD_RATIO

Variable numérica que representa el ratio de overdraft. El overdraft se produce cuando no hay dinero suficiente en una cuenta para cubrir una transacción o retirada, pero el banco permite la transacción de todos modos.

Variable: BAD_LOAN (TARGET)

Esta variable es binaria y es la variable target del proyecto, es decir, la que se va a intentar predecir a través del machine learning. Si el valor es 1 significa que el cliente está atrasado en el pago y es lo que estaría afectando el modelo de negocios del banco.

ANÁLISIS EXPLORATORIO DE LOS DATOS (EDA).

TIPOS DE DATOS

A continuación, los tipos de datos de cada una de las columnas. Los datos tipo "object" son datos categóricos, los datos tipo "int64" son números enteros y los de tipo "float64" permiten representar un número positivo o negativo con decimales, es decir, números reales.

```
id                      int64
short_emp               int64
emp_length_num          int64
last_delinq_none        int64
bad_loan                int64
annual_inc            float64
dti                   float64
last_major_derog_none float64
revol_util            float64
total_rec_late_fee    float64
od_ratio              float64
grade                  object
home_ownership         object
purpose                object
term                   object
```

```
float64    6
int64      5
object     4
```

Contamos con quince variables, de las cuales cuatro son categóricas y las diez faltantes son numéricas.

MÉTRICAS ESTADÍSTICAS

Lo primero que tenemos que conocer a la hora de explorar los datos, es que existen dos grandes grupos: los datos numéricos y los datos categóricos. Partiendo de esta base, comenzamos con el análisis.

Estadísticas de los datos numéricos:

	id	annual_inc	short_emp	emp_length_num	dti	last_delinq_none	last_major_derog_none	revol_util	total_rec_late_fee	od_ratio	bad_loan
count	20000.0	20000.0	20000.0	20000.0	19846.0	20000.0	574.0	20000.0	20000.0	20000.0	20000.0
mean	7590662.0	73350.0	0.0	7.0	17.0	1.0	1.0	56.0	0.0	1.0	0.0
std	1609593.0	45199.0	0.0	4.0	8.0	0.0	0.0	42.0	3.0	0.0	0.0
min	586040.0	8412.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	6206283.0	47000.0	0.0	3.0	11.0	0.0	1.0	39.0	0.0	0.0	0.0
50%	7378896.0	65000.0	0.0	7.0	16.0	1.0	1.0	57.0	0.0	1.0	0.0
75%	8766235.0	88000.0	0.0	11.0	22.0	1.0	1.0	74.0	0.0	1.0	0.0
max	11454641.0	1000000.0	1.0	11.0	35.0	1.0	1.0	5010.0	96.0	1.0	1.0

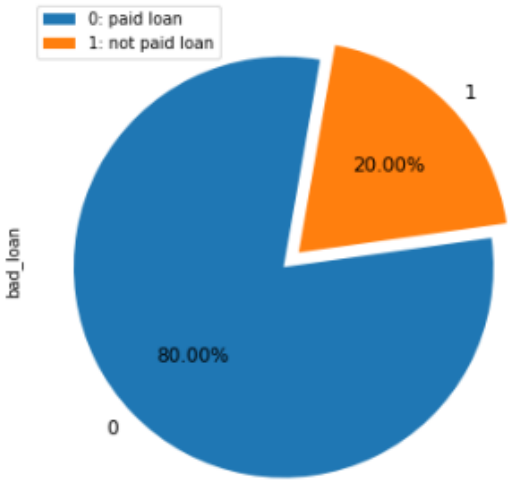
Estadísticas de los datos categóricos:

	grade	home_ownership	purpose	term
count	20000	18509	20000	20000
unique	7	3	12	3
top	B	MORTGAGE	debt_consolidation	36 months
freq	6042	9844	12337	14969

A través de estas dos tablas, lo primero que vemos es que hay 20.000 registros (filas), es decir que hay veinte mil clientes. Por otro lado, hay variables numéricas que son binarias, es decir que toman el valor 0 o 1; estas son las columnas: **`short_emp`, `last_delinq_none`, `total_rec_late_fee`, `bad_loan`.** A su vez, vemos distintas métricas estadísticas, como la media, el mínimo, máximo y los percentiles 25, 50 y 75.

INSIGHTS.

En este apartado se expondrán algunos insights, obtenidos a partir de funciones que se pueden ver en el notebook.



En primer lugar, y en base a la problemática central, vemos que un 20% de los clientes se encuentra en estado de default, esto equivale a 2000 personas.

No importa el tipo de calificación que el banco le haya otorgado al cliente, casi todos mantienen la misma relación en base a la tenencia de la vivienda, la mayoría a través de hipotecas, seguido por alquiler y la minoría son dueños de sus inmuebles. En la categoría G, la misma cantidad posee una hipotecta o alquila y en F son más los que alquilan.

Por otro lado, en base a la relación entre la calificación y el estado del pago de la deuda, tan solo los clientes de categoría A y B son mayoría con respecto a llevar la deuda pagada. Las demás categorías se encuentran, en proporción, mayormente en default.

Sorprendentemente, la tendencia dice que a medida que aumenta el ingreso, la probabilidad de impago es mayor.

Con respecto a la relación entre el ratio deuda/ingreso y el ingreso anual, se analizó que los ratios más altos se da en la gente que tiene menor ingreso, que es lo lógico, ya que los clientes más pudientes serán los que menos deuda necesiten. A la vez, observamos que a medida que aumenta el ingreso, los clientes en default son menos, salvo algunas excepciones. Y por otro lado, a mayor ratio, mas gente en default vemos.

Sigamos con la relación en base al estado de empleo. Los clientes que son empleados hace menos de un año, la mayoría esta en default, al contrario de los que son empleados hace más de un año. En promedio, los clientes trabajan hace 7 años. Los clientes que trabajan hace más de diez años son los que llevan la deuda al día. Hay algunas excepciones, como los cuatro años.

Por otro lado, se observó que, las vacaciones y los casamientos como propósito, suelen ser en promedio los préstamos que han terminado sin ser pagados.

Como era de esperarse, los clientes con mayor ratio deuda/ingreso son los que tienden a tener atrasos en el pago de sus deudas.

Por último, analicemos las correlaciones. Las variables más relacionadas con la variable target (bad_loan) son el dti, las comisiones atrasadas y el ingreso anual. Por otro lado, las variables de empleo reciente y duración tienen una correlación negativa bastante marcada, al igual que los antecedentes y la mala calificación por 90 días.

SELECCIÓN DE VARIABLES

Luego de haber realizado el análisis exploratorio de los datos, se seleccionaron las variables que serán utilizadas para entrenar el modelo.

En primer lugar se elimino la columna "ID", ya que es irrelevante. Otra cosa que se podría haber hecho, era colocarla como index.

Por otro lado, la variable "last_major_derog_none" fue borrada en el proceso de data wrangling por la alta cantidad de valores nulos que tenía.

A la hora de seleccionar las variables, se utilizó la correlación de Pearson, para así poder detectar que variables estarán mayormente relacionadas a nuestra variable target, que es el columna bad_loan. El "p_value" evalúa en qué medida sus datos rechazan la hipótesis nula, que afirma que no existe relación entre dos grupos comparados. A continuación, vemos la correlación de las variables numéricas:

	Pearson Corr.	p-value
annual_inc	-0.1227	0.0000
short_emp	0.0368	0.0000
emp_length_num	-0.0406	0.0000
dti	0.1394	0.0000
last_delinq_none	0.0216	0.0024
revol_util	0.1013	0.0000
total_rec_late_fee	0.0240	0.0007
od_ratio	0.0007	0.9214

De todas estas, se eliminará la variable "od_ratio", por tener un p_value mayor a 0.5.

Por otro lado, todas las variables categóricas tienen un p_value casi nulo, por lo que podemos suponer que poseen un alto grado de importancia y relación con respecto a la variable target.

MACHINE LEARNING

Antes de poder entrenar un modelo de machine learning se debe pasar por el proceso de data wrangling, donde se manejan los datos con el fin de que no queden valores nulos ni outliers, y el proceso de encoding, preparar los datos para poder entrenar un modelo. Todo está realizado y explicado en el notebook.

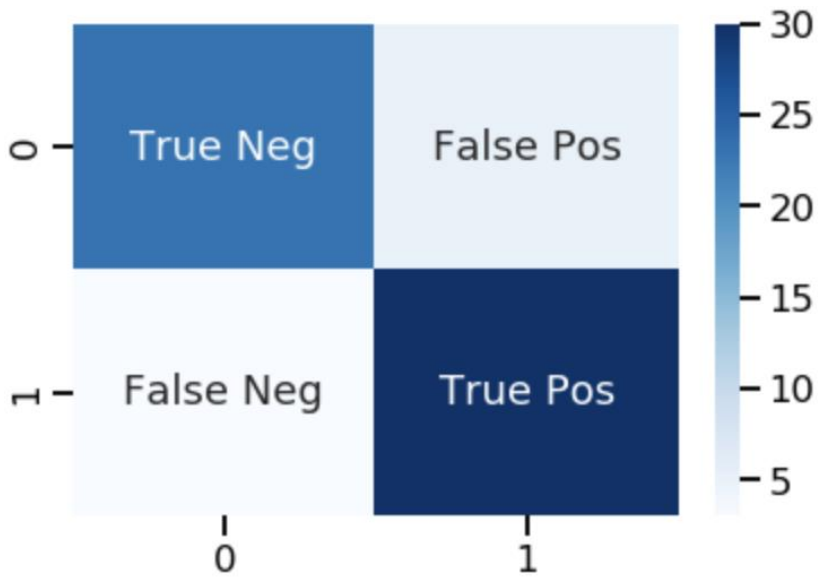
Por otro lado, para poder evaluar el modelo se usará la métrica **AUC ROC**. Esta métrica fue seleccionada porque se esta trabajando con datos desbalanceados con respecto al target.

Los modelos utilizados para predecir la variable bad_loan fueron los siguientes:

- Regresión logística.
- KNN
- SVC
- Árbol de decisión
- Random forest
- Red neuronal

En primer lugar, y para todos los modelos, se entrenaron con una división 80% entrenamiento y 20% testeo. Además, los modelos ya han sido optimizados.

A la vez, para evaluar se utilizará una matriz de confusión que nos dice que datos fueron falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos. La matriz es la siguiente:



Los resultados de los testeos fueron los siguientes:

MODELO	SCORE	AUC ROC	RECALL
REGRESIÓN LOGISTICA	0.65	0.71	0.67
KNN	0.81	0.51	0.014
SVC	0.65	0.68	0.59
ARBOL DE DECISIÓN	0.78	0.62	0.12
RANDOM FOREST	0.80	0.68	0.048
RED NEURONAL	0.81	0.71	0.04

CONCLUSIÓN

La variable objetivo, el default en la deuda (bad_loan), se encuentra desbalanceada, es decir existen muchos casos en una respuesta y muy pocos en otra, a saber el 80% de clientes lleva el pago de la deuda al día, mientras que el 20% restante se encuentra en mora.

El dataset recibido para el análisis propuesto contaba con 15 variables, 13 se utilizaron en los 5 modelos de Machine Learning (ML). Los modelos implementados fueron el Árbol de Decisión, el Random Forest, el Support Vector Machine, la Regresión Logística, el KNN y la red neuronal.

Se utilizó una subdivisión 80/20 del dataset, para los set de Train/Test. Se optimizaron los modelos, con técnicas como GridSearchCV y best_estimator

La optimización del modelo de Regresión logística fue el algoritmo que consiguió el mayor valor de área bajo la curva roc, métrica que se seleccionó como objetivo para seleccionar el modelo que mejor puede predecir si un cliente se encuentra en mora o no. El modelo obtuvo una precisión (precisión) o proporción entre el número de predicciones correctas respecto al total del 65.67%, una sensibilidad (Recall), casos positivos correctamente identificados del 67% y un área de la curva roc igual a 0,71.