IMPAGO DE PRÉSTAMOS: ANÁLISIS DE DATOS Y ELABORACIÓN DE UN MODELO PREDICTIVO

Gonzalo Fernández de Córdoba García

Tutor del proyecto: Jaime Martín García-Cuerva

Proyecto de Investigación de Bachillerato de Excelencia

IES Arquitecto Ventura Rodríguez

Curso: 2022/2023

CONSEJERÍA DE EDUCACIÓN
E INVESTIGACIÓN
Comunidad de Madrid

ARQUITECTO VENTURA RODRÍGUEZ

IMPAGO DE PRÉSTAMOS: ANÁLISIS DE DATOS Y ELABORACIÓN DE UN MODELO PREDICTIVO

Realizado por Gonzalo Fernández de Córdoba García.

Coordinado por Jaime Martín García-Cuerva

Curso 2022/2023

Trabajo realizado en el programa de Bachillerato de Excelencia del IES Arquitecto Ventura Rodríguez.

Boadilla del Monte (Madrid) diciembre de 2022

1. AGRADECIMIENTOS

Este trabajo no habría sido posible sin el apoyo de diversas personas.

En primer lugar, querría agradecerle toda la ayuda brindada a mi primo, Gonzalo, el cual me dio la idea del proyecto, me inició en el mundo del *machine learning* y me ha recomendado diversos cursos que me han ayudado a la realización del trabajo.

También ha sido vital el apoyo de mi familia, que me ha apoyado en aquellos momentos más duros del proyecto.

Por último, quisiera darle las gracias a Jaime, mi tutor; el cual me ha guiado y me ha ayudado a adaptar y plasmar el trabajo en este documento, y a su vez me ha transmitido una positividad y una motivación que me han ayudado a seguir adelante.

2. ABSTRACT

We constantly rely on credits, as they allow us to make payments we can not afford by borrowing a certain amount of money, based on the promise to later pay back the same amount with interests. This is the main business of banks: to lend the capital deposited by their clients to borrowers. To be successful in the risky business of credit, however, it is important to minimize the losses derived from defaults or loans that are not payed back. Failure in credit policy is detrimental to banks and to society in general, as we already saw in the financial crisis of 2008. In order to improve their results, most financial entities have adopted the use of machine learning predictive models that are able to predict whether a loan is likely to be defaulted or not. In this project, using a data set containing information about 10,000 loans from 2008 to 2018, we built a predictive model capable of accurately assigning a probability of default to any give loan, and also draw conclusions from the data. The project in its entirety was carried out by coding in the Python programming language, also making use of several programming libraries. More specifically, the statistical analysis of the data was performed by employing machine learning methods and numerous hypothesis contrasts. This project accomplished the development of a predictive model, capable of assigning a probability of default to any given loan, as well as the identification of some of the factors that primarily contribute to credit failure and the general characteristics of defaulting loanees. Although the results are subjected to the data set used, the project provides a brief and accesible overview of credit default.

3. RESUMEN

Este proyecto consistirá en el análisis y obtención de conclusiones acerca de datos relacionados con el impago de crédito y la elaboración de un modelo predictivo, capaz de determinar el riesgo de impago de préstamos.

El procesamiento, transformación y análisis de los datos, asi como la elaboración y evaluación del modelo, se efectuarán mediante la programación en Python y el empleo de diversas librerías informáticas. El proyecto se realizará mediante una metodología objetiva y científica, contrastando y argumentando los resultados obtenidos. Del programa de Python se sacarán los resultados, gráficas e imágenes, a no ser que se indique lo contrario; y este se dejará a disposición del lector.

ÍNDICE

4. INTRODUCCIÓN

Un crédito o préstamo es un valor monetario que una empresa o persona entrega a otra, mediante un convenio formal o implícito, en donde el que lo recibe se compromete a devolverlo con un interés y en un plazo estipulado. Estos se suelen solicitar a la hora de efectuar pagos grandes como la compra de una casa o la adquisición de un negocio, o de pagos menores como la compra de un coche o de un teléfono móvil.

La sociedad consumista y materialista en la que vivimos se basa en el dinero y, por ende, en los préstamos, ya que estos nos dan la oportunidad de embarcarnos en oportunidades de negocio, estudiar, formar una familia o efectuar otro tipo de gastos ya sean necesarias o por capricho; sin disponer previamente del dinero.

La base del modelo de negocio de las entidades bancarias es la asignación de préstamos bancarios, por los que cobran una comisión o interés. Para ello emplean el capital depositado por sus clientes e inversores, que confían en que el dinero sea puesto a buen recaudo. Por lo que, al igual que para los bancos es vital maximizar los beneficios derivados de la concesión de préstamos, también lo es minimizar las perdidas provenientes del impago de préstamos; no solo por el interés de la entidad en sí, sino también el de los clientes y, con ello, el de la sociedad. El impago de créditos ya ha resultado ser fatal, como en la crisis financiera del 2008, la cual se originó, en gran medida, a raíz de la caída de bancos de inversión debido al impago en masa de préstamos hipotecarios que no habían sido concedidos de forma correcta.

En este proyecto se va a disponer de un set o conjunto de datos con 10.000 entradas. Cada una de las entradas tendrá información acerca de un préstamo, incluyendo si este fue impagado o no. Mediante la programación en el lenguaje informático de Python, en primer lugar se procesarán y se transformarán los datos iniciales, posteriormente se construirá y se evaluará un modelo informático predictivo, capaz de clasificar y asignar una probabilidad de impago a cada préstamo, y por último se tratará de obtener conclusiones acerca de los datos y el modelo.

5. CUERPO DEL TRABAJO

5.1 HIPÓTESIS Y OBJETIVOS

El objetivo de este proyecto es sacar conclusiones de un conjunto de datos bancarios y elaborar, a partir de ellos, un modelo capaz de hacer predicciones.

Las hipótesis del trabajo son las siguientes:

1. La variable numérica de entrada más relacionada con la variable de salida será el *ratio de cuota a ingresos*.
2. La variable numérica de salida con menor relación con la variable de salida será la *cantidad prestada*.
3. Los clientes con empleo serán aquellos cuya probabilidad de impago será menor, seguidos por los clientes que estén estudiando y los clientes sin empleo.
4. Los clientes casados serán aquellos cuya probabilidad de impago será menor, seguidos por los clientes que conviven, los solteros y, por último, los clientes divorciados.

5.2 MARCO TEÓRICO

Para entender el proyecto, es necesario repasar algunos de los conceptos o métodos utilizados:

5.2.1 TERMINOLOGIA BASICA

- Un **set** o **conjunto de datos** es, como su nombre indica, una colección de datos. En este caso se dispone de una tabla compuesta por filas y columnas.
- Las filas, también llamadas **entradas**, se corresponden a los préstamos. En este caso, se dispone inicialmente de 10.000.
- Las columnas o **variables** son determinadas características de las entradas que pueden tomar diferentes valores. Cuando la variable toma un valor por cada entrada, se obtienen unos **datos**. Un ejemplo serían los ingresos: esta variable se refiere a cuánto dinero ingresa el solicitante del préstamo al año, y los datos son los ingresos de cada solicitante.
- Una **variable de entrada** es una variable que aporta información acerca de las entradas o préstamos. En este caso, variables que aportan información acerca del

préstamo en sí (cantidad prestada, longitud del préstamo o fecha de solicitud del préstamo) o los solicitantes de este (ingresos, estado laboral o estado civil).

- Una **variable de salida** es determinada por algunas de las variables de entrada. En este caso es el impago. Esta puede ser 1 si el préstamo ha sido impagado, o 0 si el préstamo no ha sido impagado. Dado que no todas las variables de entrada estarán necesariamente relacionadas con la variable de salida, se deberán identificar aquellas que sí lo estén.
- Una **variable numérica** guarda un valor continuo, es decir, puede tomar infinitos valores.
- Una **variable categórica** solo puede tomar un número finito de valores, llamados clases.

5.2.2 MACHINE LEARNING

- El ***machine learning*** es una rama de la inteligencia artificial que, grosso modo, dota a las máquinas de la capacidad de aprender. En este proyecto, se ha utilizado para elaborar el modelo predictivo y analizar y transformar los datos.
- Un **algoritmo informático** son una serie de procesos que se ejecutan de manera finita e independiente con un objetivo determinado.
- Un **algoritmo de regresión logística** es un algoritmo utilizado en problemas de clasificación, es decir, en los que la variable de salida es categórica. Cuando este se *entrena* o se prepara para poder predecir correctamente, trata de encontrar la función que describe el comportamiento de la variable de salida. Para ello, actualiza los coeficientes que varían el peso de las variables de entrada en la predicción con una función de coste, la cual trata de minimizar, y un algoritmo de gradiente descendente. El modelo predictivo de este proyecto consiste en un algoritmo de regresión logística que le asigna una probabilidad de impago a cada cliente representada por un valor entre 0 y 1, y que posteriormente es aproximada a 0 o a 1 (siendo 1 que va a impagar y 0 que no va a impagar).

**Ilustración 1.1 en anexo: función de coste y algoritmo de gradiente descendente.*

5.2.3 ESTADÍSTICA

- La **media** (μ) es el promedio de unos valores. Se calcula mediante la suma de todos los valores entre el número total de estos.
- Los **cuartiles** dividen los datos en secciones. Si los datos son agrupados de manera ascendente, en el primer cuartil se encuentran el 25% de los valores, en el segundo

cuartil o **mediana** el 50%, en el tercer cuartil el 75%, y en el cuarto cuartil el 100%.

- La **desviación típica** o **estándar (σ)** es una medida de dispersión. Es la raíz cuadrada de la varianza.
- Un **contraste de hipótesis** es una forma de comprobar si una hipótesis considerada cierta a priori (hipótesis nula o H_0) es correcta, o si en cambio, otra hipótesis que afirma lo contrario (hipótesis alternativa o H_1) lo es. Se han utilizado para puntuar las variables de entrada y evaluar la validez de los resultados obtenidos.
- El **valor p** es la probabilidad de que el valor obtenido sea cierto teniendo en cuenta que la hipótesis nula (H_0) es cierta. En este proyecto se ha utilizado en contrastes de hipótesis para comprobar si la diferencia entre dos poblaciones era estadísticamente significativa, siendo H_0 que no y H_1 que sí. Si este tomaba un valor inferior al nivel de significancia (del 0,05), se confirmaba la hipótesis alternativa.
- La **prueba chi cuadrado** se utiliza para estudiar la independencia entre dos variables categóricas. En este proyecto se ha utilizado para comprobar la independencia entre la variable de salida y las variables categóricas de entrada.
- El **test ANOVA** se utiliza para hallar la diferencia o varianza entre las medias de diferentes grupos. En este proyecto se ha utilizado para ver la correlación entre la variable de salida y las variables de entrada numéricas.
- Una **distribución de frecuencias** muestra cuántas veces se ha observado un valor o rango de valores determinado, y es descrita por los siguientes parámetros:
    - **Asimetría**: medida de asimetría de una distribución. Si esta es 0, la distribución es simétrica y la media y la mediana son la misma. Si es negativa, la media es menor que la mediana; mientras que si es positiva, la media es superior a la mediana.

      **Ilustración 1.2 en anexo: explicación visual de la asimetría de una distribución.*

    - **Curtosis o puntiagudez**: mide cómo de "puntiaguda" es la distribución comparada con la distribución normal, que presenta una curtosis de 0. Esta nos indica la concentración de datos respecto de la media (si esta es simétrica). Si la curtosis es positiva, la concentración será mayor (más puntiaguda); mientras que si es negativa, será menor (más achatada).

Para representar distribuciones de datos o frecuencias, se van a utilizar **histogramas** y **diagramas de cajas y bigotes**. Estos nos permiten analizarlas de una manera visual.

- Los **histogramas** tienen que presentar un número de divisiones que minimice el ruido de los datos pero que dé la suficiente información. Para ello se hallará el valor de Sturges, que nos da el número ideal de divisiones (k). Este viene dado por la siguiente expresión, en la que N es el número de observaciones:

$$k = 1 + \log_2(N)$$

- Los **diagramas de cajas**, en cambio, muestran una caja delimitada por el 1er y 3er cuartil, que a su vez presenta una recta roja que representa la mediana. Además, muestran unas rectas negras o *bigotes*, a partir de los cuales se representan aquellas entradas con valores 1,5 veces más altos o bajos que el rango intercuartílico (diferencia entre el primer y tercer cuartil de la distribución). Estos se denominan *outliers* o valores atípicos.

  **Ilustración 1.3 en anexo: ejemplos de histograma y diagrama de cajas.*

- La **estandarización de datos** es un proceso que consigue compactar los datos, convirtiendo la desviación estándar a 1 y la media a 0. En machine learning, es necesario estandarizar las distintas variables para tenerlas en una escala común sin cambiar la diferencia en el rango de valores. De esta manera, se mantiene la distribución, pero las variables se pueden comparar entre sí.

  **Ilustración 1.4 en anexo: expresión de la estandarización.*

  **Ilustraciones 1.5, 1.6 en anexo: ejemplo de la estandarización de unos datos.*

5.2.4 MÉTRICAS DEL MODELO

- Una entrada es **positiva** cuando su variable de salida es 1 (impago). Una entrada es **negativa** cuando su variable de salida es 0 (no impago)
- **Verdaderos positivos**: predicciones cuya variable de salida es 1 (impago) que son predichas como 1 (como impagos).
- **Verdaderos negativos**: predicciones cuya variable de salida es 0 (no impago) que son predichas como 0 (como no impagos).
- **Falsos positivos**: predicciones cuya variable de salida es 0 (no impago) que son predichas como 1 (como impagos).

- **Falsos negativos**: predicciones cuya variable de salida es 1 (impago) que son predichas como 0 (como no impagos).
- La **exactitud** es la proporción de predicciones correctas respecto del total.
- La **precisión** es la proporción de verdaderos positivos respecto del total de positivos predichos (verdaderos positivos y falsos positivos).
- La **sensibilidad** es la proporción de verdaderos positivos respecto del total de positivos (verdaderos positivos y falsos negativos). Es la probabilidad de predecir correctamente una entrada positiva o impagada
- La **especificidad** es la proporción de verdaderos negativos respecto del total de negativos (verdaderos negativos y falsos positivos). Es la probabilidad de predecir correctamente una entrada negativa o no impagada
- **Curva ROC**: curva que indica la proporción de verdaderos positivos y falsos positivos a medida que el modelo va prediciendo.
- **AUROC**: Es el área debajo de la curva ROC, y nos indica la capacidad discriminatoria del modelo.

  **Ilustración 1.7 en anexo*: *ejemplo de curva ROC.*

- **Coeficiente Gini**: se calcula a partir del AUROC ($2 \times AUROC - 1$) , y en este caso indica la efectividad del modelo a la hora de calcular la probabilidad de impago. Este varía de -1 a 1.
- **Puntuación F1**: puntuación que tiene en cuenta la sensibilidad y la precisión. Indica la exactitud teniendo en cuenta el desbalance de los datos.

5.3 MARCO PRÁCTICO.

5.3.1 INFORMACIÓN BÁSICA DEL CONJUNTO DE DATOS.

Las variables numéricas de entrada son las siguientes:

- **Ingresos**, que representa los ingresos del cliente en un año.
- **Ratio cuota a ingresos**, que compara el pago mensual del préstamo con los ingresos mensuales del cliente.
- **Cantidad prestada**, que representa el capital prestado.
- **Longitud del préstamo**, que son los meses en los que se devuelve el crédito.
- **Puntuación externa**, que es una puntuación dada al cliente por una entidad denomidada *Schufa*.

Las variables categóricas de entrada son:

- **Ocupación**, que describe el estado laboral del cliente (estudiante, desempleado, empleado).
- **Estado civil**, que describe la situación personal del cliente (soltero, casado, divorciado, separado, conviviendo).
- **Número de solicitantes**, que hace alusión a si el número de personas que piden el préstamo es 1 o 2.

La variable de salida del conjunto es el **impago**, y se utiliza para clasificar las entradas en impagadas (1) o no impagadas (0).

También se dispone de la **fecha de solicitud del crédito**, que detalla cuándo se produjo el préstamo

	income	loan_amount	term_length	install_to_inc	occup	marital	schufa	num_applic	OBS_DATE	target_var
0	18785.517943694504	12300	67.0	0.009772532761805456	nan	Single	7106.014471346445	1	18JUL2018 - 00:00:00	nan
1	12861.495159877606	Not avail.	113.0	0.003509136597401741	Employee	Divorced	7694.806893720367	1	16JUL2018 - 00:00:00	0
2	14886.776341632107	10700	Not avail.	0.013310346283231944	Unemployed	Single	7142.496337537019	1	21DEC2010 - 00:00:00	1
3	Not avail.	33000	112.0	0.004760874885112928	Employee	Single	7446.1706118576485	1	05NOV2015 - 00:00:00	0
4	15897.753806874589	19900	59.0	0.02121608748572823	Unemployed	Separated	7241.656646188094	1	13JUL2015 - 00:00:00	1

Ilustración 2.1: muestra de parte del conjunto de datos

	variables	explicación / traducción
0	income	Ingresos del cliente
1	loan_amount	Cantidad prestada
2	term_length	Longitud del préstamo
3	install_to_inc	Ratio cuota a ingresos
4	occup	Situación laboral
5	marital	Estado civil
6	schufa	Puntuación externa
7	num_applic	Número de solicitantes
8	OBS_DATE	Fecha de solicitud del crédito
9	target_var	Variable de salida. Si se impagó o no

Ilustración 2.2: traducción de las variables del conjunto de datos.

5.3.2 DEPURACIÓN Y TRANSFORMACIÓN INICIAL DE LOS DATOS.

- DATOS AUSENTES

Algunas de las entradas están incompletas, es decir, hay variables con valores vacíos (nulls) o que no tienen un valor especificado. Si el conjunto de datos fuera más grande se podrían desechar, pero dado que solo se dispone de 10.000 entradas, que se verían reducidas a casi la mitad (5403), es indispensable tratar de minimizar las perdidas. Es vital disponer del máximo número de datos posibles, ya que así el modelo tendrá mejores resultados y el análisis será mejor.

	income	loan_amount	term_length	install_to_inc	occup	marital	schufa	num_applic	OBS_DATE	target_var
0	18785.517943694504	12300	67.0	0.009772532761805456	nan	Single	7106.014471346445	1	18JUL2018 - 00:00:00	nan
1	12861.495159877606	nan	113.0	0.003509136597401741	Employee	Divorced	7694.806893720367	1	16JUL2018 - 00:00:00	0
2	14886.776341632107	10700	nan	0.013310346283231944	Unemployed	Single	7142.496337537019	1	21DEC2010 - 00:00:00	1
3	nan	33000	112.0	0.004760874885112928	Employee	Single	7446.1706118576485	1	05NOV2015 - 00:00:00	0
4	15897.753806874589	19900	59.0	0.02121608748572823	Unemployed	Separated	7241.656646188094	1	13JUL2015 - 00:00:00	1

Ilustración 2.3: en rojo, valores vacíos en el conjunto de datos

Aquellas entradas que presenten las variables fecha, puntuación externa o impago vacías, serán inservibles. Estas variables no se pueden averiguar de ninguna manera y son necesarias para el proyecto, por lo que nos desharemos de estas (alrededor de 565 entradas). En cambio, hay ciertas variables que se pueden inferir a partir de otras: ratio de cuota a ingresos, ingresos, longitud del préstamo y cantidad del préstamo. Estas variables guardan la siguiente relación:

$$ratio\ de\ cuota\ a\ ingresos = \frac{\frac{cantidad\ prestada}{longitud\ del\ préstamo}}{ingresos}$$

Si a una entrada le falta una de estas variables, se podrá averiguar a partir de las otras. Si a alguna entrada le faltan más de una de estas variables, se utilizarán árboles de regresión que, basándose en los valores de las variables de otras entradas, las inferirá. Con esto se consigue que el número de entradas se vea reducido solamente en 1.618.

El conjunto de datos también está desbalanceado, es decir, presenta más entradas no impagadas que impagadas. Dado que igualar el número de entradas impagadas y no impagadas supondría eliminar alrededor del 85% de las entradas (4.947), se corregirá este sesgo en el momento de evaluar el modelo.

**Ilustración 2.4 en anexo: muestra del desbalance del conjunto de datos.*

- DIVISIÓN DE LOS DATOS EN SETS DE ENTRENAMIENTOS Y DE COMPROBACIÓN

Los algoritmos se dedican a buscar patrones para, en base a estos, dar una salida. A veces, estos tienden a adaptarse demasiado a los datos de entrenamiento y, eventualmente, pueden acabar estando sesgados hacia estos.

Nosotros buscamos un modelo que sea capaz de generalizar, por lo que para controlar esto y comprobar si realmente rinde como debería, se separarán los datos en dos sets: uno de entrenamiento, que se empleará para *entrenar* o preparar al modelo para predecir, y otro de comprobación, empleado para evaluar las predicciones de dicho modelo. El set de entrenamiento estará compuesto de alrededor del 70% de las entradas, mientras que el set de comprobación lo estará por alrededor del 30%. Además, se conservará la proporción de impagos y no impagos en ambos conjuntos de datos.

5.3.3 TRANSFORMACIÓN DE LAS VARIABLES CATEGÓRICAS

Como se ha mencionado anteriormente, las variables categóricas disponen de distintas clases o categorías, y suelen ser de gran utilidad para el modelo a la hora de predecir, ya que las distintas clases suelen guardar diferencias en lo que respecta a la variable de salida (en este caso, el impago). Sin embargo, aunque a veces los datos estén bien clasificados, puede que las distintas clases no "ayuden" necesariamente al modelo a la hora de clasificar; o que simplemente las distintas clases no guarden la suficiente "diferencia" entre ellas, en lo que respecta a la variable de salida, como para que la variable categórica sea determinante a la hora de predecir. Si disminuimos el número de clases pero hacemos que las diferencias entre la probabilidad de impago de estas sean mayores, el modelo le podrá dar más valor a la variable categórica, ya que ayudará a segmentar los datos.

Pero esto no solo se hará para mejorar la capacidad predictiva del modelo, ya que también nos servirá al analizar las variables. En algunas variables hay clases con un número de muestras muy pequeño y cuya clase únicamente viene descrita por un número, algo que no nos da mucha información, precisamente. Dado que estamos analizando los datos principalmente en función del impago y que estas clases apenas cambiarán la distribución de las otras categorías, esto favorecerá al análisis.

La hipótesis nula (H_0) será que las distintas clases no guardan una diferencia estadísticamente significativa entre sí en lo que respecta a la variable objetivo, es decir, que las probabilidades de impago de cada clase guardan diferencias estadísticamente significativas entre sí. La hipótesis alternativa (H_1), en cambio, afirmará que las variables sí guardan una diferencia estadísticamente significativa entre sí y que, por lo tanto, son de la misma clase en lo que respecta al impago.

Se compararán los datos pertenecientes a cada clase con aquellos de la clase con una probabilidad de impago inmediatamente superior e inferior, y se irán agrupando con aquellas clases con las que guarden una menor diferencia (mayor valor p).

- SITUACIÓN LABORAL

La variable situación laboral contiene las siguientes clases: empleado, estudiante, desempleado, 1, 2, 3 y no disponible.

occup	loan_amount	target_var	customer	PD
Employee	32923230.000	32	1614	0.020
2	1268100.000	7	63	0.111
1	1311600.000	8	62	0.129
Student	21858110.000	140	1067	0.131
Not avail.	6970740.000	66	358	0.184
Unemployed	54667990.000	647	2663	0.243
3	1055250.000	13	52	0.250

Ilustración 2.5: clases iniciales de la variable situación laboral.

occup_agg	loan_amount	target_var	customer	PD
Employee	32923230.000	32	1614	0.020
Student	31408550.000	221	1550	0.143
Unemployed	55723240.000	660	2715	0.243

Ilustración 2.8: clases resultantes.

**Ilustración 2.6, 2.7 en anexo: contrastes de hipótesis de la variable situación laboral.*

Al final, se han combinado todas las clases hasta quedar las clases empleado, estudiante y desempleado.

- ESTADO CIVIL

La segunda variable categórica es "estado civil", cuyas clases son las siguientes: casados, viviendo juntos, soltero, divorciado, separado y no disponible.

Se han ido combinando hasta quedar las clases de casados, viviendo juntos, soltero y divorciado o separado.

marital	loan_amount	target_var	customer	PD
Married	15760280.000	22	784	0.028
Living together	16447330.000	88	801	0.110
Not avail.	7271560.000	47	354	0.133
Single	46297080.000	407	2273	0.179
Divorced	17071820.000	167	825	0.202
Separated	17206950.000	182	842	0.216

Ilustración 2.9: clases iniciales de la variable estado

marital_agg	loan_amount	target_var	customer	PD
Married	15760280.000	22	784	0.028
Living together	16447330.000	88	801	0.110
Single	53568640.000	454	2627	0.173
Divorced_sep	34278770.000	349	1667	0.209

Ilustración 2.12: clases resultantes

**Ilustración 2.10, 2.11 en anexo: contrastes de hipótesis de la variable estado civil.*

➢ NÚMERO DE SOLICITANTES

La tercera variable categórica es "número de solicitantes", cuyas clases son 1, 2 y no disponible.

num_applic	loan_amount	target_var	customer	PD
2	32273760.000	190	1605	0.118
Not avail.	7042500.000	53	354	0.150
1	80738760.000	670	3920	0.171

Ilustración 2.13: clases de la variable número de solicitantes

num_applic_agg	loan_amount	target_var	customer	PD
2	39316260.000	243	1959	0.124
1	80738760.000	670	3920	0.171

Ilustración 2.16: clases resultantes

**Ilustración 2.14, 2.15 en anexo: contrastes de hipótesis de la variable número de solicitantes.*

Al final, han quedado las clases de 1 y 2 solicitantes.

Añadiremos las variables transformadas al conjunto de datos para comprobar si dan mejores resultados.

5.3.4 TRANSFORMACIÓN DE LAS VARIABLES NÚMERICAS

➢ ANÁLISIS DE CORRELACIÓN

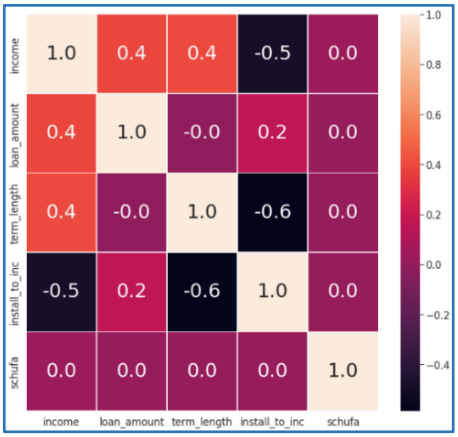

Ilustración 2.17: mapa de calor del análisis de correlación entre las variables

Primero analizaremos si las variables numéricas guardan una correlación alta entre sí. En el caso de que la guarden, o bien se combinarán, o se dejará aquella que dé una mejor puntuación posteriormente.

Dado que los valores de correlación son bajos entre todas las variables, no se hará cambio alguno.

➢ BREVE EXPLICACIÓN

Para representar distribuciones de datos o frecuencias, se van a utilizar histogramas y diagramas de cajas y bigotes.

Nuestro objetivo es hacer que las distribuciones de las distintas variables numéricas se asemejen lo máximo posible a la distribución normal, es decir, trataremos de normalizar

los datos. Para ello, modificaremos estas distribuciones para quitarle peso a aquellos valores altos que hagan que la distribución sea asimétrica y demasiado compacta o extensa, y trataremos de convertir la asimetría y la curtosis a 0. Las transformaciones que aplicaremos serán la transformación logarítmica (aplicar un logaritmo a los datos), la transformación raíz cuadrada (aplicar una raíz cuadrada) y las transformaciones box-cox (serie de transformaciones en función de un parámetro, λ, que varía según las características de la distribución).

- INGRESOS

La distribución de la variable ingresos presenta, inicialmente, una asimetría de 0,49 y una curtosis de -0,14.

(ver transformaciones en los anexos, *Ilustraciones 2.19, 2.20, 2.21*)

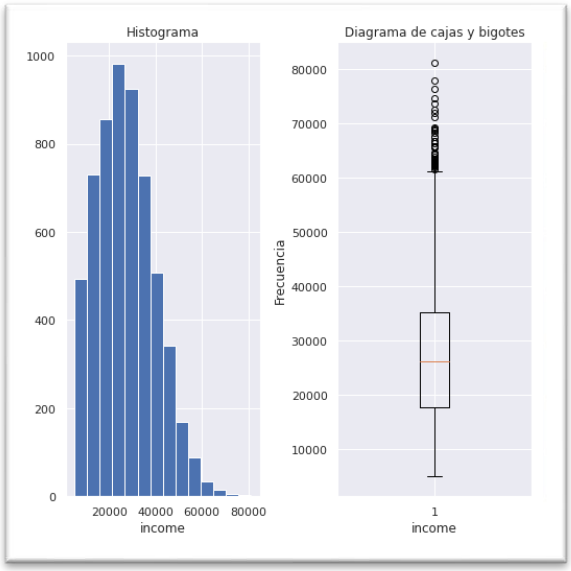

Ilustración 2.18: distribución de la variable ingresos sin transformar

- CANTIDAD PRESTADA

La distribución de la variable cantidad prestada presenta una asimetría de 0,27 y una curtosis de -0,17.

(ver transformaciones en los anexos, *Ilustraciones 2.23, 2.24, 2.25*)

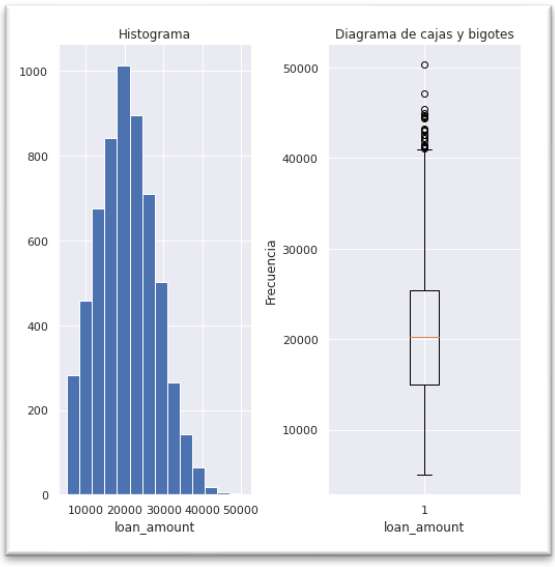

Ilustración 2.22: distribución de la variable cantidad prestada sin transformar

- LONGITUD DEL PRÉSTAMO

La segunda variable numérica cuya distribución vamos a analizar es la variable longitud del préstamo. Esta presenta una asimetría de 0,5, la cual es considerablemente alta, y una curtosis de -0,04.

(ver transformaciones en los anexos, *Ilustraciones 2.27, 2.28, 2.29*)

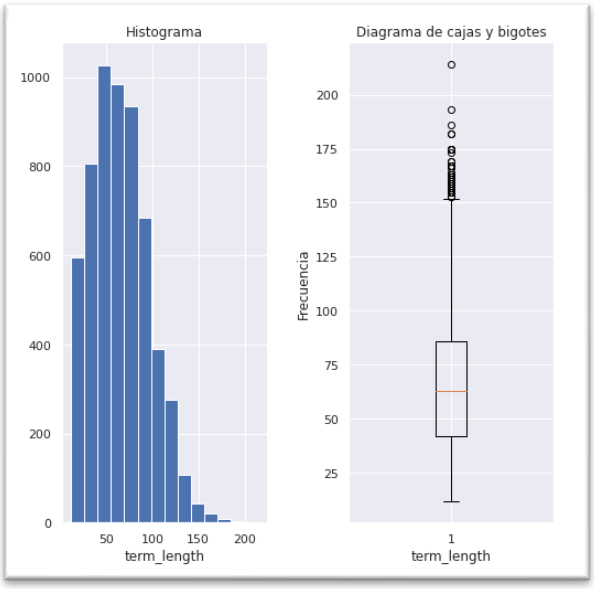

Ilustración 2.26: distribución de la variable longitud del préstamo sin transformar

- RATIO DE CUOTA A INGRESOS

La siguiente variable numérica cuya distribución vamos a analizar es la variable ratio de cuota a ingresos. Esta presenta una asimetría de 4,17 y una curtosis de 29, las cuales se salen completamente de lo normal.

(ver transformaciones en los anexos, *Ilustraciones 2.31, 2.32, 2.33*)

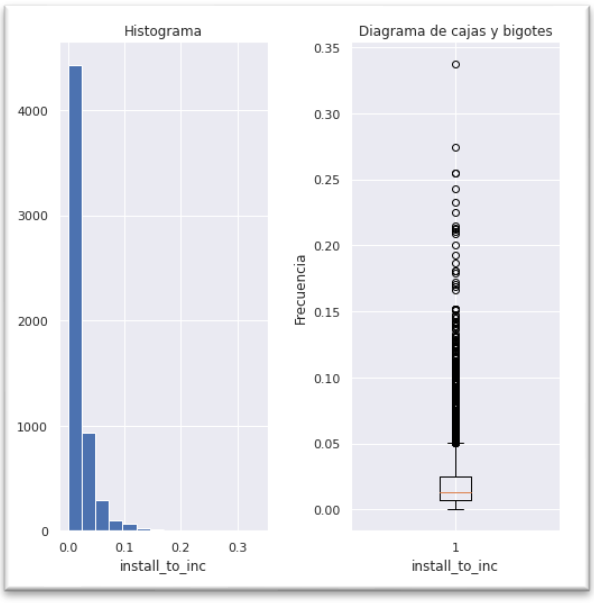

Ilustración 2.30: distribución de la variable ratio de cuota a ingresos sin transformar

- PUNTUACIÓN EXTERNA

La última variable numérica cuya distribución vamos a analizar es la variable "puntuación externa". Esta presenta una asimetría de 0,91 y una curtosis de 0,29.

(ver transformaciones en los anexos, *Ilustraciones 2.35, 2.36, 2.37*)

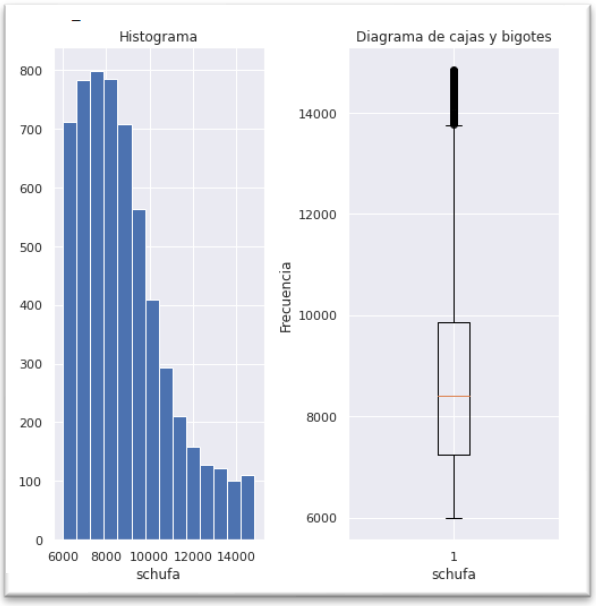

Ilustración 2.34: distribución de la variable puntuación externa sin transformar

A priori, se puede afirmar que, en general, la transformación que mejor ha funcionado ha sido la box-cox, ya que ha normalizado satisfactoriamente aquellas distribuciones tanto con mucha como con poca asimetría. La transformación logarítmica también ha funcionado bien con aquellas distribuciones con una gran asimetría, al igual que la de raíz cuadrada ha normalizado correctamente aquellas con una asimetría menor; pero la transformación box-cox ha sido más eficaz al poder adaptar la normalización en función de las características de cada distribución.

Añadiremos las variables transformadas para comprobar si estas dan mejores resultados.

**Ilustración 2.38 en anexo: parámetros de las transformaciones*

5.3.5 SELECCIÓN DE LAS VARIABLES

El *feature selection* o la selección de variables es, como su nombre indica, el proceso de escoger las variables más relevantes para el modelo predictivo.

En este caso vamos a utilizar el test ANOVA para seleccionar las mejores variables numéricas, y el chi cuadrado para hacer lo propio con las variables categóricas.

**Ilustración 2.39 en anexo: esquema de métodos de selección de variables*

➢ SELECCIÓN DE LAS VARIABLES NUMÉRICAS

Como se ha mencionado anteriormente, emplearemos el test ANOVA para la selección de las variables numéricas. En concreto, obtendremos el estadístico f, que indica la diferencia entre las distribuciones de las entradas positivas y de las entradas negativas de una variable (cuando mayor es el estadístico f, mayor es la diferencia). Con este seleccionaremos la mejor versión de cada variable. Posteriormente, seleccionaremos las variables finales que se pasarán al modelo.

	Variables	Puntuación
0	schufaboxcox	866.262
1	schufalog	789.661
2	incomesqrt	739.725
3	incomeboxcox	737.757
4	schufa	689.432
5	income	677.334
6	install_to_incboxcox	476.947
7	install_to_inclog	475.352
8	term_length	338.179
9	term_lengthsqrt	295.072
10	term_lengthboxcox	295.033
11	install_to_inc	242.319
12	loan_amount	59.233
13	loan_amountboxcox	59.073
14	loan_amountsqrt	58.723

Ilustración 2.40: puntuación de las variables numéricas en el test ANOVA.

Como podemos observar *(ilustración 2.40)*, las transformaciones han sido en general exitosas, especialmente la box-cox, dado que han obtenido mayores puntuaciones. Además, cabe destacar la baja puntuación que ha obtenido la variable cantidad prestada; algo esperable dado que esta, sobre el papel, no parecía influir demasiado en la variable objetivo.

Las variables numéricas que seleccionaremos serán las transformaciones box-cox de puntuación externa, ratio de cuota a ingreso y cantidad prestada; la transformación de raíz cuadrada de ingresos y la variable original de longitud del préstamo.

➢ SELECCIÓN DE LAS VARIABLES CATEGÓRICAS

Para seleccionar la mejor versión de cada variable categórica, en cambio, utilizaremos la prueba de chi cuadrado. Obtendremos una puntuación basada en el valor p obtenido.

Lo primero que haremos será asignar valores numéricos a las clases de las variables categóricas. Como podemos observar *(ilustración 2.41)*, las transformaciones no han dado buenos resultados. Aquellas variables no transformadas han obtenido una mayor puntuación, por lo que las mantendremos.

	features	score
0	occup	607.812
1	occup_agg	228.106
2	marital	130.840
3	marital_agg	73.286
4	num_applic	13.505
5	num_applic_agg	7.291

Ilustración 2.41: puntuación de las variables categóricas en la prueba chi cuadrado

- SELECCIÓN DE LAS MEJORES VARIABLES

Antes de elegir las mejores variables y entrenar el modelo, estandarizaremos los datos numéricos para que las variables numéricas puedan ser comparadas entre sí.

Ahora que tenemos la mejor versión de cada variable, escogeremos las variables óptimas para el modelo. Para ello utilizaremos la eliminación recursiva de las variables con validación cruzada: dividiremos el conjunto de datos de entrenamiento en 5 sets más pequeños, y se medirá el rendimiento de un modelo de regresión logística. Se irán eliminando las variables según su relación con la variable objetivo (de menor a mayor), y se verá cuál es el número óptimo de variables, así como cuáles son esas variables. La puntuación que se usará para valorar el desempeño del modelo será el AUROC o área debajo de la curva ROC.

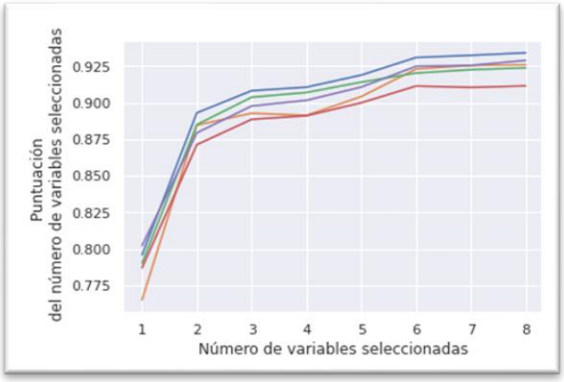

*Ilustración 2.42 en anexo: *datos estandarizados*

Ilustración 2.43: gráfica de la validación cruzada

En este caso, el modelo rinde mejor con todas las variables, por lo que se mantendrán.

5.3.6 ENTRENAMIENTO Y ELABORACIÓN DEL MODELO

Una vez ya preparados los datos, se procederá al ensamblaje final del modelo. Para ello utilizaremos la técnica de *grid search* con validación cruzada, con la que se construirán distintos modelos con diversas combinaciones de *hiperparámetros* (que son los parámetros del modelo) para así seleccionar aquellos que den un mayor AUROC.

Una vez ya seleccionados los hiperparámetros, procederemos al *entrenamiento* del modelo para que este "aprenda" a predecir. Para ello se usarán los datos de entrenamiento. Para verificar y puntuar su desempeño, en cambio, se emplearán los datos de comprobación.

En riesgo de crédito, los modelos de este tipo se utilizan para poder maximizar los beneficios teniendo en cuenta el riesgo que se asume. Lo que se busca con un modelo

predictivo en este ámbito es que sea capaz de evaluar el riesgo que supone cada préstamo, y poder así analizar si merece la pena o no otorgar dicho préstamo. Por lo que, si bien este es un modelo de regresión logística que a priori es empleado para la clasificación, se va a buscar principalmente que este sea capaz de asignar correctamente una probabilidad de impago a cada entrada.

Es por esto que se utilizará principalmente como referencia el coeficiente Gini. Para evaluar la capacidad de clasificar del modelo, teniendo en cuenta el desbalance de los datos, se utilizará la puntuación F1.

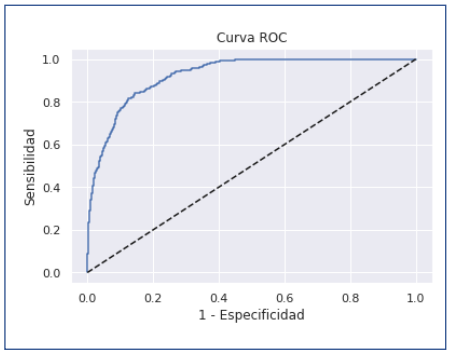

Ilustración 2.44: curva ROC

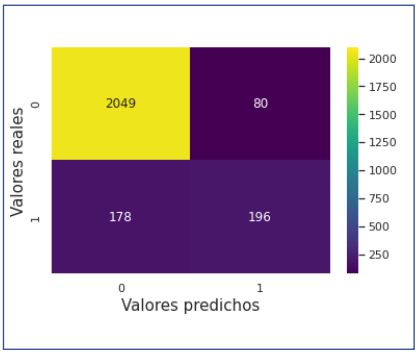

Ilustración 2.45: matriz de confusión

	Puntuaciones
Gini	0.851
AUROC	0.925
Exactitud	0.896
Precisión	0.705
Sensibilidad	0.524
Puntuación F1	0.601

Ilustración 2.46: puntuaciones del modelo obtenidas

Según se puede apreciar en la *ilustración 2.46*, los resultados son satisfactorios: el coeficiente Gini es alto, de alrededor de un 0,85 sobre 1; y la puntuación F1 es aceptable: de un 0,60 sobre 1.

La puntuación F1 muestra que el modelo no es especialmente efectivo en la clasificación, ya que el modelo no clasifica correctamente aquellas entradas dudosas, con una probabilidad de impago cercana al 0,5. Sin embargo, el alto Gini muestra que el modelo si asigna correctamente probabilidades de impago, que es lo que realmente se le pide.

5.3.7 ANÁLISIS Y OBTENCIÓN DE CONCLUSIONES DE LOS DATOS

- VARIABLES NUMÉRICAS: DIFERENCIAS DE LAS MEDIAS EN FUNCIÓN DEL IMPAGO.

Se han hecho varios contrastes de hipótesis con las variables ingresos, ratio de cuota a ingresos y puntuación externa, con el fin de hallar diferencias estadísticamente significativas entre las medias de las entradas impagadas y no impagadas. Para ello se ha empleado el valor p del test ANOVA realizado previamente (*ilustración 2.40*).

	income	loan_amount	term_length	install_to_inc	schufa
valor p	0.000	0.000	0.000	0.000	0.000

Ilustración 2.47: valor p obtenido del test ANOVA

El valor p es menor al 0,05, por lo que se puede afirmar lo siguiente:

Los clientes que impagan tienen, de media, menos ingresos, un ratio de cuota a ingresos más alto y reciben una puntuación externa menor.

- VARIABLES CATEGÓRICAS: ESTUDIO DE LA PROBABILIDAD DE IMPAGO DE LAS CLASES DE CADA VARIABLE.

Procederemos a estudiar la probabilidad de impago de las clases de las variables categóricas para averiguar qué clases tienden a impagar más. Para ello utilizaremos las transformaciones de las variables, ya que, aunque no diesen buen resultado en la prueba de chi cuadrado, el contraste de hipótesis que se hizo demostró que no había diferencias estadísticamente significativas entre las probabilidades de impago de las clases.

En la variable situación laboral, los que más impagan son los clientes desempleados (probabilidad de impago = 0,23), luego los estudiantes (probabilidad de impago = 0,14) y por último aquellos con trabajo (probabilidad de impago = 0,02).

En cuanto al estado civil, los que más impagan son los clientes divorciados o separados (probabilidad de impago = 0,21), luego los solteros (probabilidad de impago = 0,17), después los clientes conviviendo (probabilidad de impago = 0,11) y por último los casados (probabilidad de impago = 0,03).

En la variable número de solicitantes, se impaga más cuando el préstamo es solicitado por una sola persona (probabilidad de impago = 0,17) frente a cuando es solicitado por dos (probabilidad de impago = 0,12)

occup_agg	PD
Employee	0.021561
Student	0.142339
Unemployed	0.239305

Ilustración 2.48: probabilidades de impago de las clases de la variable estado laboral.

marital_agg	PD
Married	0.025087
Living together	0.105727
Single	0.170653
Divorced_sep	0.212236

Ilustración 2.49: probabilidades de impago de las clases de la variable estado civil.

num_applic_agg	PD
2	0.122318
1	0.169173

Ilustración 2.50: probabilidades de impago de las clases de la variable número de solicitantes.

➢ CARACTERÍSTICAS GENERALES DEL CONJUNTO DE DATOS

Tomando en cuenta los anteriores análisis, se han analizado las características generales del conjunto de datos para así poder poner en contexto los resultados e identificar para qué tipo de préstamos está creado el modelo.

A. TIPOS DE PRÉSTAMOS: CANTIDAD PRESTADA

La mayoría de las entradas presentan cantidades entre los 10.000 y 30.000 €, lo que indica que este es un conjunto de datos de préstamos relativamente bajos. Los préstamos hipotecarios suelen concederse a partir de alrededor de los 60.000 €, por lo que, dado que el mayor préstamo en el set es de 50.000 €, se puede afirmar que el conjunto de datos presenta únicamente préstamos personales.

	cantidad prestada
count	8382.000000
mean	20523.962658
std	7451.741917
min	5100.000000
25%	15100.000000
50%	20300.000000
75%	25500.000000
max	50300.000000

Ilustración 2.51: características de la variable cantidad prestada.

B. TIPOS DE SOLICITANTES

La media de los salarios del set es de alrededor de 27.200 €, y la mayor parte de las entradas presentan unos salarios de entre 39.800 y 14.600 €.

Ilustración 2.52 en anexos: características de la variable ingresos.

El salario medio en España en el mismo periodo que el del conjunto de datos, de 2008 a 2018, es de 22.934 €, lo cual se acerca razonablemente a aquel del conjunto de datos; aunque este último es notablemente superior.

Si observamos los ingresos medios según el año, podemos apreciar que estos se mantienen constantes entre 2008 y 2018. Comparado con los ingresos medios de España desde 2008 a 2018; estos son mayores, y se comportan de una manera ligeramente diferente, ya que los ingresos medios de España en este periodo siguen una tónica levemente ascendente.

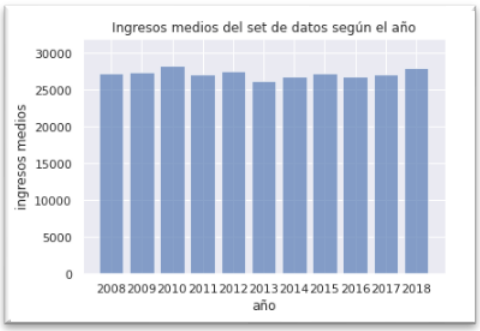

Ilustración 2.53: ingresos por año del conjunto de datos de 2008 a 2018

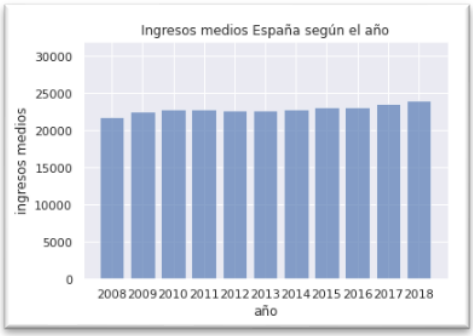

Ilustración 2.54 ingresos por año de España de 2008 a 2018

Sin embargo, parece que la muestra de clientes del conjunto de datos es razonablemente similar a la población española en cuanto a ingresos.

C. SITUACIÓN LABORAL

El set está compuesto en su mayoría por solicitantes desempleados (46%), solicitantes empleados (28%) y solicitantes estudiando (26%).

D. ESTADO CIVIL

El set está compuesto principalmente por solicitantes solteros (43,5%), seguidos por solicitantes divorciados o separados (22,6%), solicitantes casados (16%) y solicitantes conviviendo (14,3%).

5.3.8 ANÁLISIS DE RESULTADOS

El modelo calcula bien las probabilidades de impago de las entradas (Gini alto), aunque no es tan efectivo clasificándolas (Puntuación F1 media). El objetivo principal del modelo es el de asignar correctamente una probabilidad de impago a cada entrada, por lo que, si bien a la hora de clasificar puede no ser tan brillante, en este caso el modelo es muy apto para riesgo de crédito. En cambio, en otros ámbitos en los que se tiene en cuenta las clasificaciones del modelo y no las probabilidades que este da, sería necesario modificarlo para minimizar el número de falsos negativos. Además, el modelo será más eficaz prediciendo préstamos personales solicitados por clientes en una situación económica media.

La variable numérica que más ha influido en el impago ha sido la puntuación externa schufa, seguida de los ingresos, el ratio de cuota a ingresos, la longitud del préstamo y por último la cantidad prestada. Por lo tanto, la 1ª hipótesis propuesta ha resultado ser errónea (*La variable de numérica de entrada más relacionada con la variable de salida será el ratio de cuota a ingresos)*, mientras que la 2ª hipótesis propuesta ha terminado siendo correcta (*La variable numérica de salida con menor relación con la variable de salida será la cantidad prestada*).

La variable categórica que más ha influido en el impago ha sido la situación laboral, seguida por el estado civil y el número de solicitantes. Tras el análisis de estas variables, se han confirmado la 3ª hipótesis (*Los clientes con empleo serán aquellos cuya probabilidad de impago será menor, seguidos por los clientes que estén estudiando y los clientes sin empleo*) y la 4ª hipótesis (*Los clientes casados serán aquellos cuya probabilidad de impago será menor, seguidos por los clientes que conviven, los solteros y, por último, los clientes divorciados*) propuestas.

A su vez, se ha demostrado que los préstamos tienden a ser impagados más frecuentemente por clientes con ingresos más bajos, ratios de cuota a ingresos más altos y puntuaciones externas más bajas. Además, estos tienden a pedir los préstamos individualmente, a estar en un estado laboral desfavorable (estudiando, sin empleo) y, en lo que respecta a su estado civil, estos suelen estar solteros, divorciados o separados.

6. CONCLUSIONES

6.1 VALORACIÓN DE LOS RESULTADOS

En mi opinión, se han cumplido los objetivos preestablecidos. Se buscaba crear un modelo que fuese apto para la predicción del impago de préstamos, lo cual se ha conseguido; y a su vez se pretendía sacar algunas conclusiones acerca de los datos, lo que también se ha logrado.

Al principio se propusieron una serie de hipótesis. El proyecto no estaba enteramente enfocado a la confirmación de estas, pero sí que reflejaban mis expectativas iniciales.

Si bien se trató de hacer afirmaciones aparentemente obvias, dado mi desconocimiento sobre el tema; me sorprendió desmentir la primera hipótesis propuesta. Inicialmente se creía que el ratio de cuota a ingresos sería la variable numérica que mejor describiría el esfuerzo de pago de los solicitantes, ya que compara los ingresos mensuales con los pagos mensuales del cliente; y que, por ende, sería la que más influiría en la variable de salida. Sin embargo, se ha probado que la variable numérica que mejor ha descrito el impago ha sido la puntuación externa Schufa, lo cual tiene sentido, ya que tiene en cuenta muchos factores que influyen en que un cliente termine impagando o no el préstamo.

La segunda hipótesis se ha cumplido, lo cual no me ha sorprendido, dado que la cantidad prestada al cliente no debería influir tanto en el impago, si no lo que le supone al cliente el devolver todo ese capital, lo cual es indicado por el ratio de cuota a ingresos.

La tercera hipótesis también se ha confirmado. Esta tampoco era muy arriesgada, ya que aquellos solicitantes con empleo se encuentran en una situación económica más favorable que las otras dos clases, mientras que los estudiantes reciben en muchos casos apoyo económico de sus padres, y los clientes desempleados se encuentran en una situación más comprometida económicamente.

La cuarta hipótesis se ha confirmado pero era, a priori, la menos clara de todas, ya que la variable estado civil está más ligada a la situación personal que a la situación económica del cliente, y por lo tanto no había indicios muy claros acerca de qué clases tenderían a impagar más o menos, o si realmente llegaría haber diferencias entre todas estas o no. Se suponía que los clientes casados y viviendo juntos impagarían menos al disponer de personas que serían capaces de asistir económicamente, lo cual se cumplió. Respecto a

las otras dos clases, se supuso que al no disponer de ese apoyo sería más improbable que acabasen impagando, lo cual también se confirmó.

6.2 LIMITACIONES

La realización del proyecto me ha supuesto un gran reto:

Por un lado, durante estos últimos 8 meses he tenido que iniciarme en el mundo de la ciencia de datos y de la programación. Para poder llevar a cabo el proyecto, he tenido que previamente formarme, y posteriormente ir averiguando cómo hacer mi objetivo una realidad. Si bien hay una gran disponibilidad de recursos en Internet en forma de cursos, blogs y páginas web, que posibilitan la iniciación y formación en este ámbito; frecuentemente (especialmente al principio) he sentido que el contenido estaba fuera de mi alcance y dirigido a personas con un nivel superior en la materia.

Por otro lado, plasmar el proyecto en este documento escrito me ha sido tremendamente complicado; en gran parte debido a la complejidad y extensión del proyecto, el cual he tenido que resumir, suprimiendo gran parte del trabajo realizado. Es un tema muy complejo y que requiere un cierto grado de especialización en la materia, y hacerlo entendible para virtualmente cualquier persona que quiera leer el trabajo ha sido especialmente difícil.

Estoy muy satisfecho con el proyecto. Sin embargo, creo que si la extensión del proyecto pudiese haber sido mayor se podría haber explicado todo más detenidamente, y se podría haber ahondado más en algunas partes del proyecto. A su vez, creo que si hubiese tenido un mayor grado de formación previo a la realización del trabajo, me habría sido más fácil desde un primer momento.

6.3 CONCLUSIÓN PERSONAL

Dicho todo esto creo que, en su totalidad, ha sido una experiencia positiva. El proyecto me ha servido para adentrarme en el mundo de la ciencia de datos , el cual tiene muchas salidas en el ámbito laboral, y para reafirmarme en mi voluntad de estudiar ingeniería informática. Estoy agradecido de poder haber tenido la oportunidad de realizar este proyecto.

La realización de este tipo de trabajos le permiten al alumno ampliar sus conocimientos lejos de las aulas y aprender a ser más autosuficiente. Frecuentemente, los estudiantes se quejan de la poca preparación para el "mundo real" que suponen la educación secundaria y el Bachillerato. En mi opinión, los proyectos de esta índole son un buen complemento al trabajo realizado en las aulas.

ACCESO AL ENTORNO EN EL QUE SE HA ELABORADO EL PROYECTO

Como se ha mencionado anteriormente, el trabajo se ha realizado mediante la programación en *Python*. Para que el lector pueda observar cómo se ha realizado el proyecto, el programa se ha dejado a disposición del lector.

Código QR al programa de Python.

https://colab.research.google.com/drive/1AwULtft7ex_57Tp2rjhrn4c1OCiouDhs#scrollTo=3TsFjWEzmU7I

7. BIBLIOGRAFÍA

[1] Abdatum. (2021) *Curvas ROC.*

https://abdatum.com/ciencia/curvas-roc

[Accedido el 17/11/22].

[2] Barrios Arce, J.I. (2019). *La matriz de confusión y sus métricas.*

https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

[Accedido el 19/11/22].

[3] BBVA. *Qué es un préstamo financiero: tipos y diferencias con un crédito.*

https://www.bbva.com/es/salud-financiera/que-es-un-prestamo-financiero-tipos-y-diferencias-con-un-credito/

[Accedido el 15/9/22].

[4] BBVA. *Tipos de préstamos, cuáles son sus características.*

https://www.bbva.es/finanzas-vistazo/ef/prestamos/tipos-de-prestamos.html

[Accedido el 15/9/22].

[5] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer.

https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

[Accedido el 30/8/22].

[6] G.E.P. Box y D.R. Cox (1964). *An Analysis of Transformations, Series B (Methodological)*. Journal of the Royal Statistical Society B, vol. 26, p. 211-252.

https://www.jstor.org/stable/2984418

[Accedido el 7/12/22].

[7] Medium. (2018) *Understanding AUC – ROC Curve.*

https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
[Accedido el 17/11/22].

[8] Medium. (2020) *Using the Gini coefficient to evaluate the performance of credit score models.*

https://www.bbva.es/finanzas-vistazo/ef/prestamos/tipos-de-prestamos.html

[Accedido el 17/11/22].

[9] Shalev-Shwartz, S. y Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press.

https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/

[Accedido el 28/8/22].

[10] Statista. (2022) *Salario bruto medio al año en España de 2008 a 2019.*

https://es.statista.com/estadisticas/478099/espana-salario-medio-al-ano/

[Accedido el 17/11/22].

[11] Statistics How To. *Box Cox Transformation: Definition, Examples.*

https://www.statisticshowto.com/probability-and-statistics/normal-distributions/box-cox-transformation/

[Accedido el 1/9/22].

[12] *Supervised Machine Learning: Regression and Classification* – Coursera, Stanford University.

https://coursera.org/share/1bd100f620e35029edc3601672059494

[Accedido por primera vez el 10/4/22].