

# Predicción de Riesgo de Crédito

Un modelo de ML para la detección anticipada de default crediticio

# Resumen

1. Contexto y Audiencia
2. Hipótesis / Preguntas de Interés
3. Análisis Exploratorio
4. Selección del Algoritmo
5. Feature Selection y Reevaluación del Modelo
6. Resultados Obtenidos

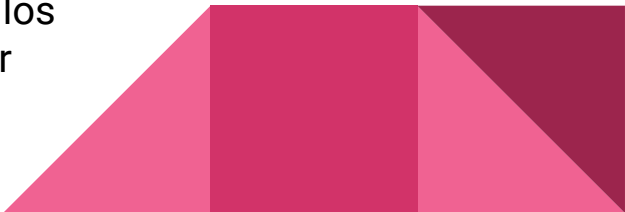


# 1. Contexto y Audiencia

Las **finanzas de consumo** están presentes en casi todas las decisiones económicas que las personas toman en su vida como comprar un automóvil o una casa, pagar la universidad o hacer un viaje en vacaciones. Todas estas experiencias por lo general requieren un gran gasto que se puede acomodar para realizar el pago en cuotas a lo largo del tiempo tomando préstamos.

En el otro lado de la mesa los prestamistas se arriesgan en cada operación, ya que incurren en pérdidas cuando los clientes no devuelven sus préstamos a tiempo. Desde su punto de vista, el riesgo es deseable porque es representativo de una oportunidad de ganancia mediante una tasa de interés, pero demasiado riesgo podría conducir a pérdidas masivas y potencialmente a la quiebra.

Es por eso que la predicción de incumplimiento crediticio es fundamental para **administrar el riesgo** en un negocio de **préstamos al consumidor**, ya que permite a los prestamistas optimizar las decisiones de préstamo y crear una mejor experiencia para el cliente.



## 2. Hipótesis / Preguntas de Interés

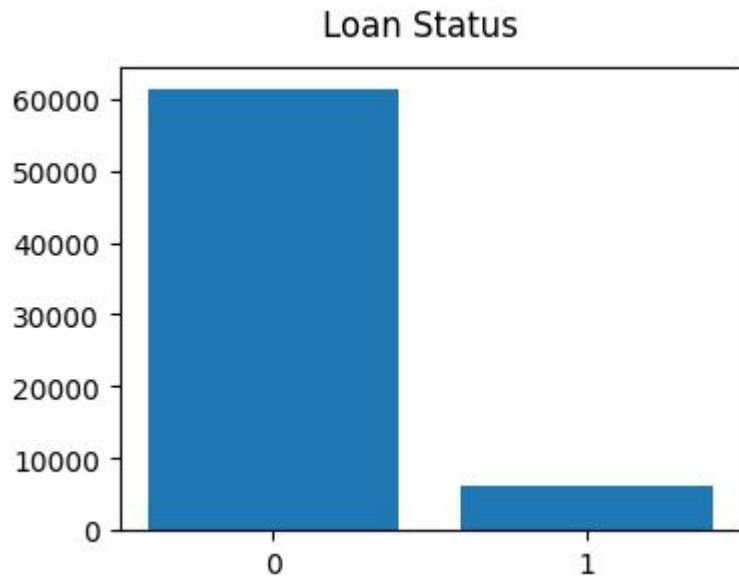
La **hipótesis** es que es posible predecir si un préstamo caerá en mora utilizando antecedentes e información financiera histórica sobre el tomador del préstamo y las condiciones del préstamo.

En el camino, podremos encontrar ideas para algunas **preguntas relevantes** del tema, tales como:

- ¿Cuánto influye el comportamiento financiero pasado en el futuro?
- ¿Cuál es el/los atributo/s del consumidor más relevante a considerar en la gestión de riesgos?
- ¿Qué tipo de préstamo es más riesgoso?
- ¿Existe alguna condición particular que tenga mayor relevancia en la morosidad de los préstamos?



### 3. Análisis Exploratorio - Variable Objetivo

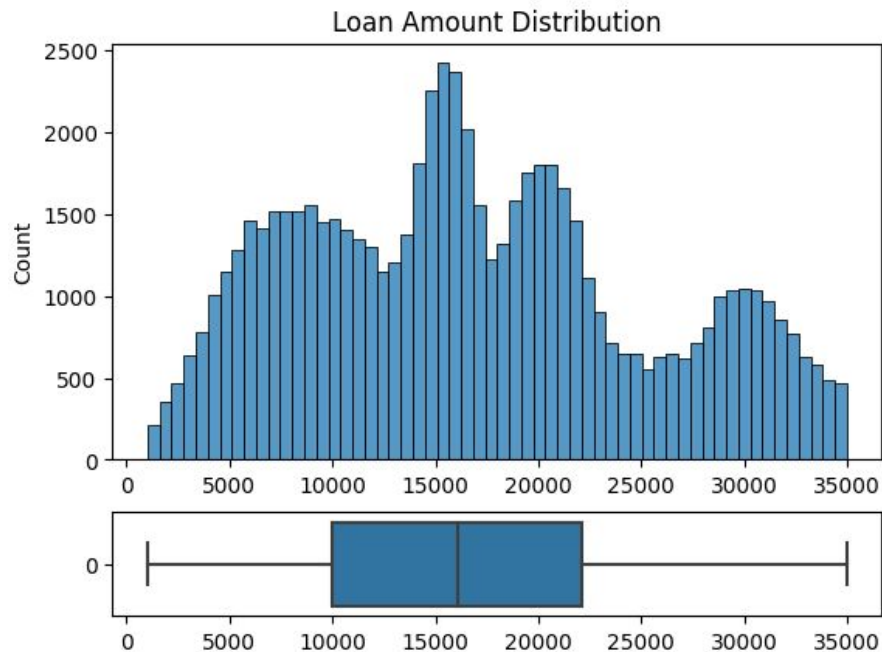


El primer dato importante al explorar el dataset, es que la variable objetivo se encuentra muy desbalanceada. Por lo tanto para poder utilizar los datos correctamente, en el desarrollo del trabajo se realizó un “resampling” para sortear esta dificultad.

En el gráfico:

- 0 indica los préstamos al día
- 1 indica los préstamos en mora.

### 3. Análisis Exploratorio - Variables importantes

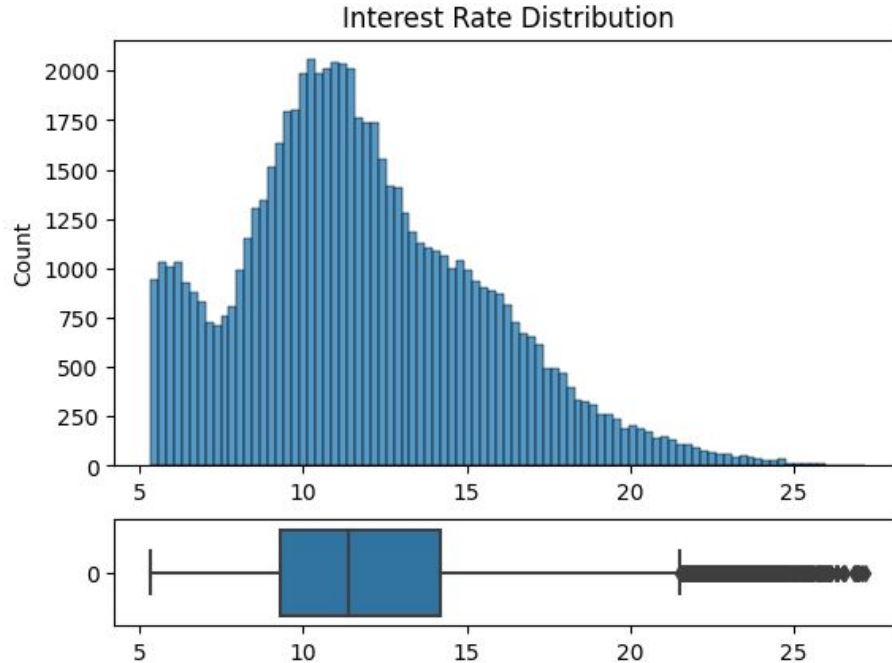


Para tener noción de los valores que se están analizando, veamos la distribución de la variable “Loan Amount”.

Los montos de los préstamos del dataset varían entre \$1.000 y \$35.000.

La mayor concentración se encuentra entre los \$10.000 y los \$22.000.

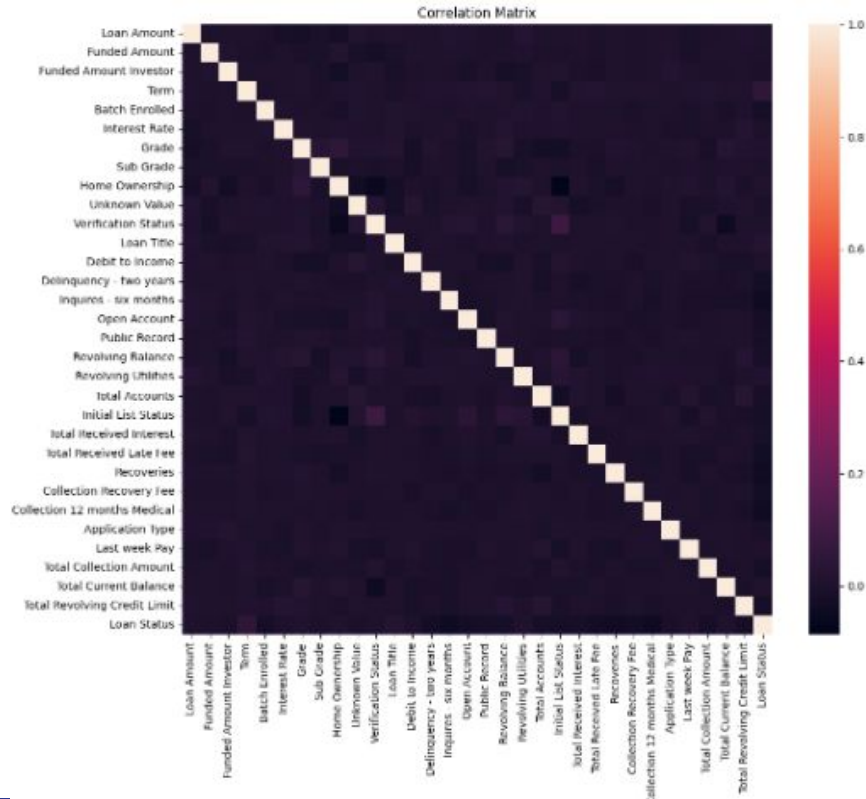
### 3. Análisis Exploratorio - Variables importantes



Tener idea de las tasas de interés aplicables a los préstamos, puede hacernos una idea de la rentabilidad del negocio de los prestadores.

Observando las tasas de interés, vemos que varían entre 5% y 30%, con la mayor proporción entre 8% y 14%

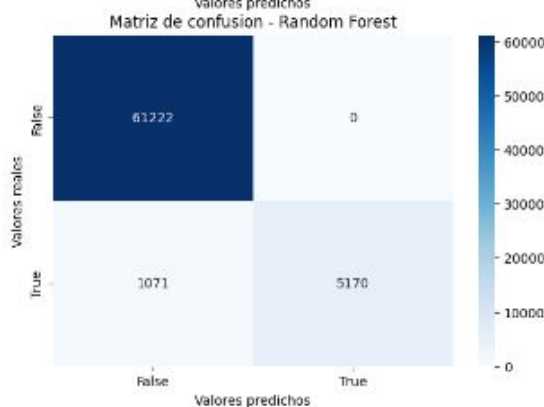
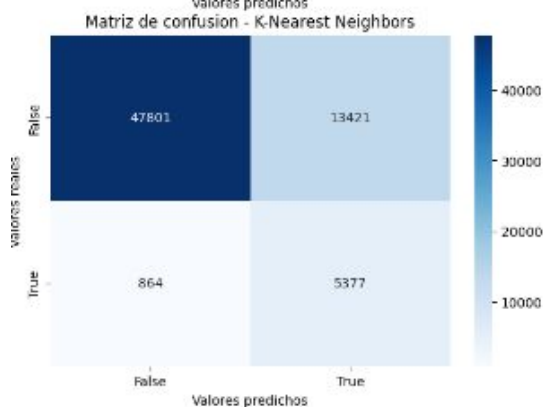
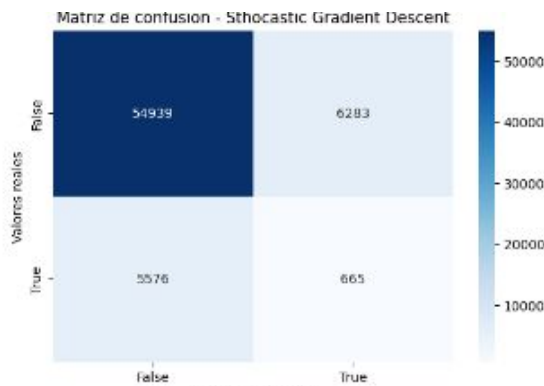
### 3. Análisis Exploratorio - Correlaciones



Claramente no existen correlaciones fuertes entre las variables originales del dataset. Esto puede suponer un obstáculo a la obtención de insights mediante la exploración visual, por lo que se optó por no seguir indagando en este sentido.



## 4. Selección del Algoritmo



Para seleccionar el modelo a utilizar se compararon los resultados de 4 algoritmos:

- Decision Tree
- Stochastic Gradient Descent
- K-Nearest Neighbors
- Random Forest

## 4. Selección del Algoritmo

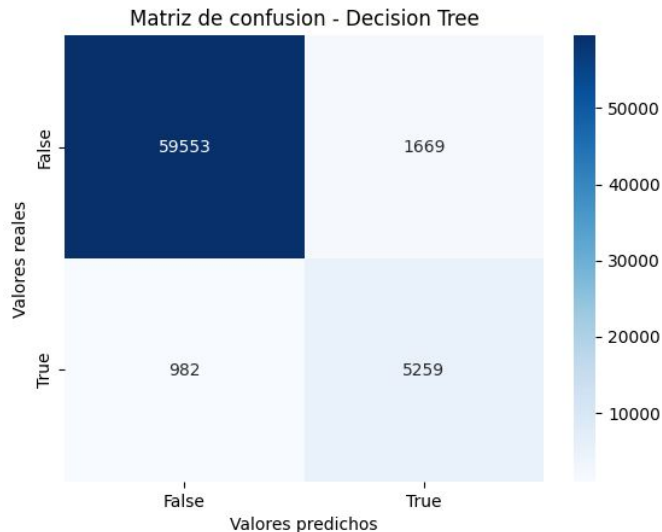
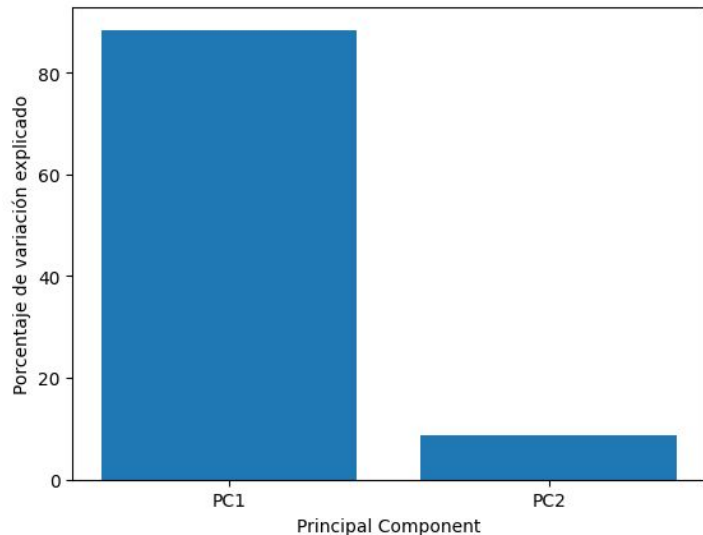
	metricas	Decision Tree	Sthocastic Gradient Descent	K-Nearest Neighbors	Random Forest
0	Accuracy	0.96	0.82	0.79	0.98
1	Precision	0.76	0.10	0.29	1.00
2	Recall	0.84	0.11	0.86	0.83
3	F1-Score	0.80	0.10	0.43	0.91

Las métricas indican que los algoritmos que mejor se desempeñaron son el Decision Tree y el Random Forest. El resto se descartan.

Si bien las métricas del Random Forest son más altas, podríamos estar frente a un caso de overfitting. Además el Recall del Decision Tree es mayor y esta medida es la más relevante para el contexto del negocio.

Esto es porque indica la cantidad de defaults detectados que se podrían evitar mediante la aplicación del algoritmo, esquivando así la pérdida total del capital.

## 5. Feature Selection y Reevaluación del Modelo



Un análisis de PCA, permitió encontrar que la cantidad de variables se pueden reducir drásticamente manteniendo la capacidad de predicción del modelo.

## 6. Resultados Obtenidos

Tal como inferimos a partir del resultado de la aplicación del PCA, utilizando sólo 2 variables de las provistas por el dataset inicial es suficiente para que el modelo mantenga una performance aceptable.

Respecto a las variables utilizadas, una de ellas ("Unknown Value") corresponde a una serie de datos que no estaba correctamente identificada en el dataset original.

Si se tratara de datos de un negocio real, resultaría muy importante solicitar al proveedor de los datos una aclaración sobre a qué se refiere la variable en cuestión.

Respecto al modelo, el haber reducido la cantidad de variables de las 32 iniciales a sólo 2 es una ganancia muy relevante ya que resultará ser un modelo de aplicación más simple a la hora de hacer el despliegue.

