**Modelo Machine Learning para la clasificación de clientes**

**propensos a caer en Default**

**Trabajo Fin de Grado**

**Máster Universitario en Big Data Science**

**Curso académico 2023-2024**

AUTORES :

Jaime Vila Sagaseta de Ilúrdoz,

Mario Pérez Vicente

Unai Zabaleta Gañán

Luis Eduardo Lozada Vargas

Gonzalo Martínez Berzal

José López López-Jurado

Marcos Castro Cacho

TUTOR ACADÉMICO:

Alberto Iriarte Lanas

Madrid , 03/07/2024

**ÍNDICE**

[**1. RESUMEN 3**](#_heading=h.i2h588tq3vlw)

[**2. INTRODUCCIÓN 4**](#_heading=h.4s179cqod6dx)

[**3. LIMPIEZA Y PREPROCESAMIENTO DE DATOS 5**](#_heading=h.u56apmqzui7z)

[**4. ANÁLISIS ESTADÍSTICO DE LOS DATOS 6**](#_heading=h.aht5t5nsyji5)

[**5. REDUCCIÓN DE DIMENSIONALIDAD 23**](#_heading=h.omy5min07bo4)

[**6. ANÁLISIS DE LA IMPORTANCIA DE LAS VARIABLES 24**](#_heading=h.2p3zc6wpasyx)

[**7. CONSTRUCCIÓN DEL MODELO 25**](#_heading=h.9vq7oq72k0x9)

[**8. RETURN ON INVESTMENT (ROI) 26**](#_heading=h.cl09eonizwpk)

[**9. CASO DE USO 28**](#_heading=h.4vbkkggu8gu4)

[**10. POSIBLES NUEVAS VARIABLES 29**](#_heading=h.5mzlkhfq76tt)

[**11. REFERENCIAS 31**](#_heading=h.x0yc7pqyzdg5)

[**12. BIBLIOGRAFÍA 32**](#_heading=h.wdim0a10pzdf)

# RESUMEN

Este Proyecto se centra en la implementación de un sistema de detección de riesgo crediticio utilizando Machine Learning para clasificar a nuevos clientes como "Buenos Clientes" o "Posible Default". El flujo de trabajo consiste en varias etapas, empezando por el Análisis Exploratorio de Datos hasta la Evaluación del Rendimiento del Modelo, asegurando la calidad y la efectividad del modelo predictivo.

Se prioriza la limpieza de datos, el preprocesamiento y la selección del modelo. La optimización de hiper parámetros, la validación cruzada y la evaluación del modelo son fundamentales para garantizar la precisión y confiabilidad de las predicciones.

El objetivo final es mejorar la toma de decisiones crediticias de la empresa al distinguir entre clientes confiables y aquellos con mayor riesgo de incumplimiento, permitiendo mantener un equilibrio entre la expansión de la cartera de clientes y la mitigación del riesgo crediticio.

# INTRODUCCIÓN

En el contexto del Big Data, la capacidad de analizar

Diagrama

Descripción generada automáticamente

Esquema 1.- Diagrama de flujo empleado en el proyecto

El objetivo del proyecto es optimizar las estrategias de crédito y marketing de una financiera en expansión, frente al desafío de un aumento en la tasa de default. El flujo de trabajo incluye el análisis exploratorio de datos para entender patrones, la limpieza y preprocesamiento de datos, incluyendo la normalización y codificación de variables, y la división del conjunto de datos en entrenamiento y prueba. Se seleccionará y entrenará el modelo, con una posterior optimización y validación cruzada para estimar el rendimiento y evitar el sobreajuste. La evaluación del rendimiento del modelo permitirá ajustes necesarios, asegurando la precisión, confiabilidad y justicia del modelo para una toma de decisiones crediticias más informada y equilibrada.

# LIMPIEZA Y PREPROCESAMIENTO DE DATOS

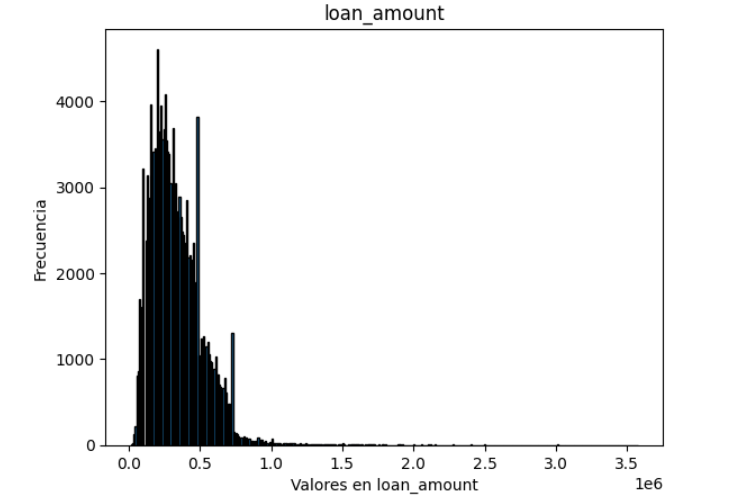
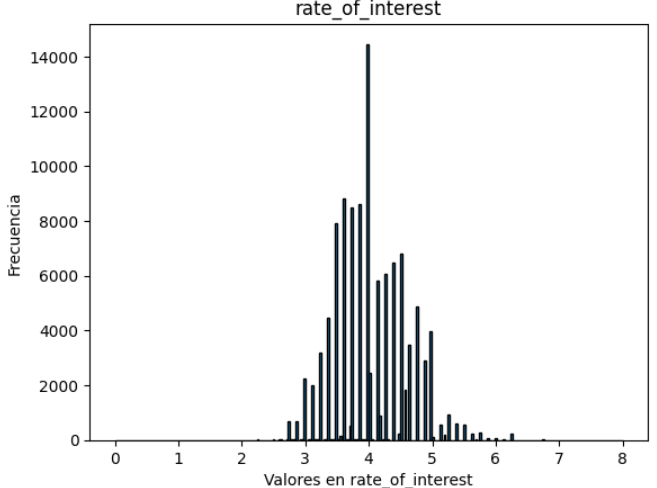
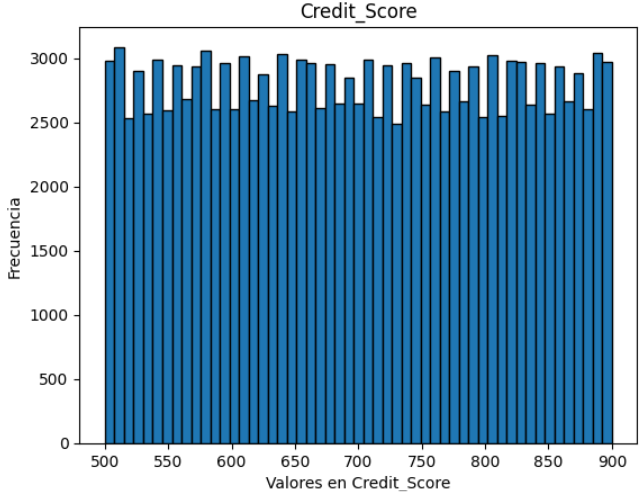
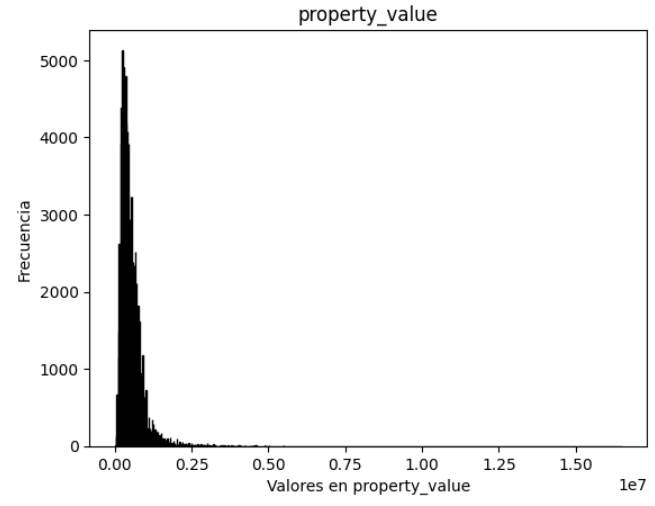
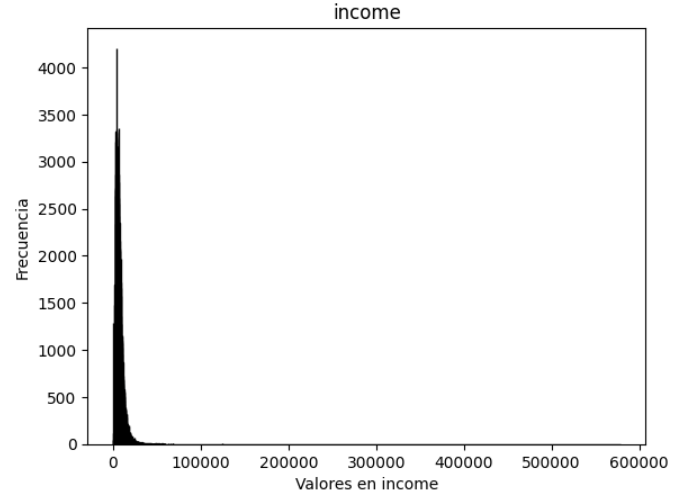
La limpieza de datos ha consistido en los siguientes pasos, adaptados a las necesidades específicas de la financiera y basados en las columnas presentes en el dataset:

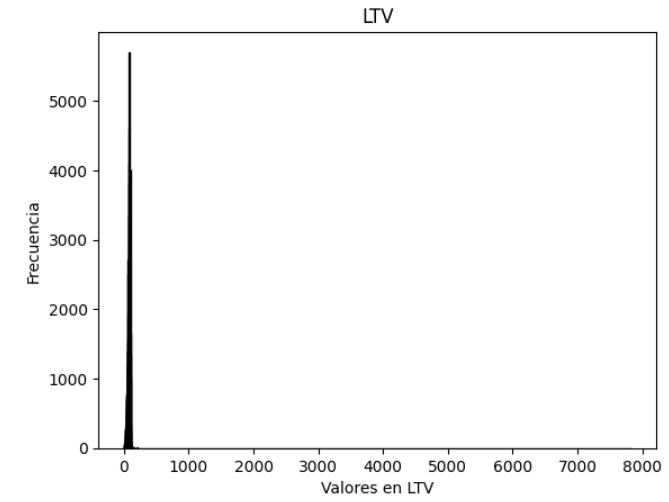
1. **Carga de los datos**: Este paso implica el uso de bibliotecas como Pandas para leer y cargar el conjunto de datos en estructuras de datos manipulables, como DataFrames, facilitando el acceso y análisis inicial de los datos.
2. **Identificación de Variables Categóricas y Continuas:** Se realiza un examen del conjunto de datos para clasificar las variables en categóricas (que representan categorías, como género) y continuas (que representan valores numéricos, como ingresos), preparándose para el preprocesamiento adecuado.
3. **Análisis y limpieza de Valores Nulos:** Consiste en identificar los valores faltantes dentro del conjunto de datos y decidir cómo tratarlos, ya sea eliminándolos o imputándolos con un valor representativo, para mantener la integridad del análisis.
4. **Eliminación de Outliers:** Este paso se enfoca en detectar y eliminar los valores atípicos, es decir, aquellos valores extremadamente altos o bajos que pueden sesgar el análisis y los modelos predictivos, utilizando técnicas estadísticas para identificarlos.
5. **Selección de Variables**: Implica determinar cuáles variables son más relevantes para el modelo predictivo, utilizando métodos estadísticos o de machine learning para evaluar su importancia y efecto en el modelo, optimizando así el rendimiento y la precisión del mismo.

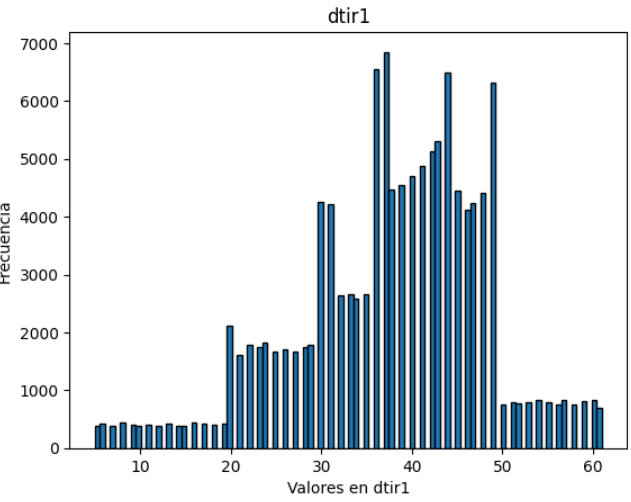
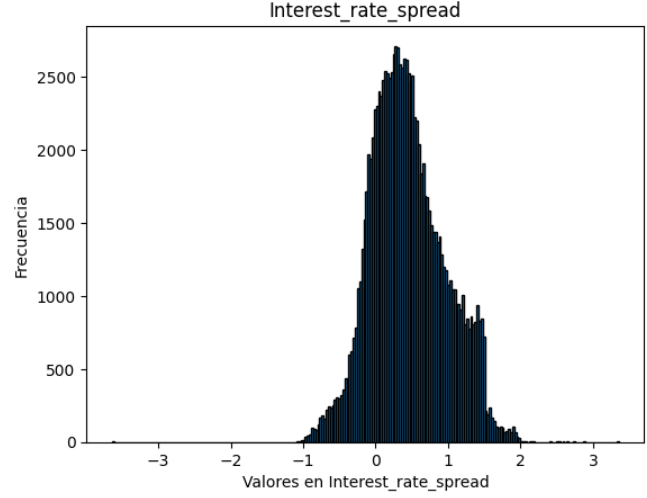
# ANÁLISIS ESTADÍSTICO DE LOS DATOS

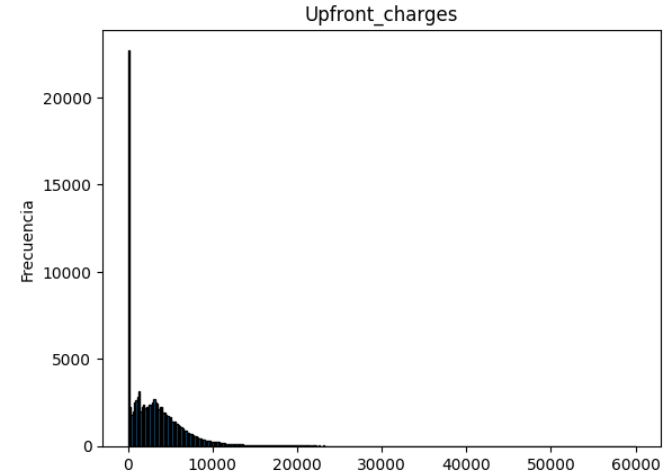
Tras realizar una análisis de las distintas variables, tanto numéricas como categóricas, que vamos a utilizar, hemos sacado ciertos patrones de cada una de ellas.

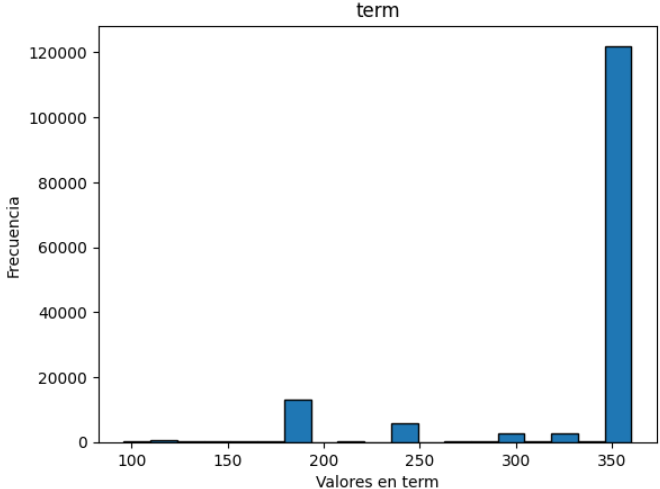
**NUMÉRICAS**

* **loan\_amount**: Variable sin nulos. Distribución bastante normal. Se ve que la distribución de los préstamos en mora es mayor que la de los préstamos no en mora.
* **rate\_of\_interest**: Muy poquitos nulos. Distribución bastante normal. Los intereses son menores para préstamos pagados.
* **Credit\_score**: Variable nada normal, muy uniforme. No se ve diferencia entre clases.
* **property\_value**: Hay outliers (propiedades de muchísimo valor). Los préstamos morosos tienen propiedades de menor valor. 
* **income**: Hay outliers. Los morosos tienen menores ingresos que los no morosos (como era esperado).
* **LTV**: Habría que quitar outliers. No se ve mucha diferencia entre clases.

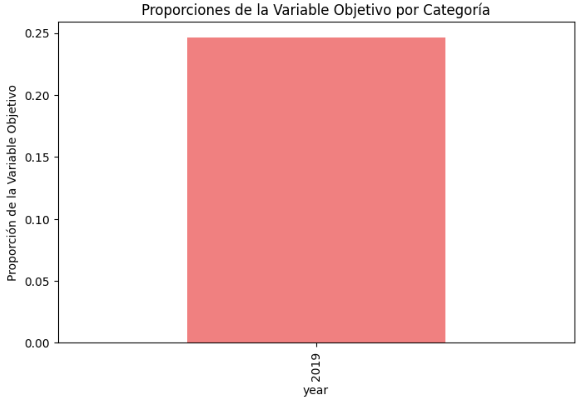
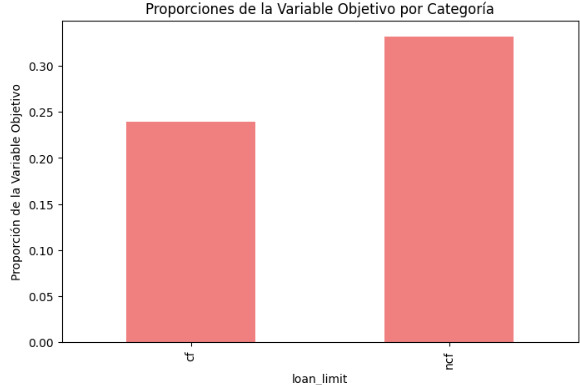
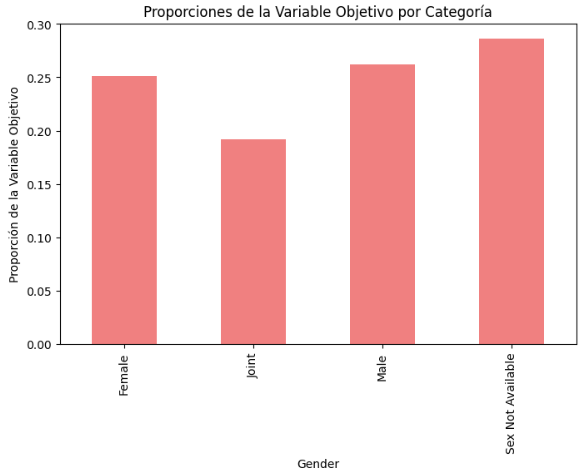
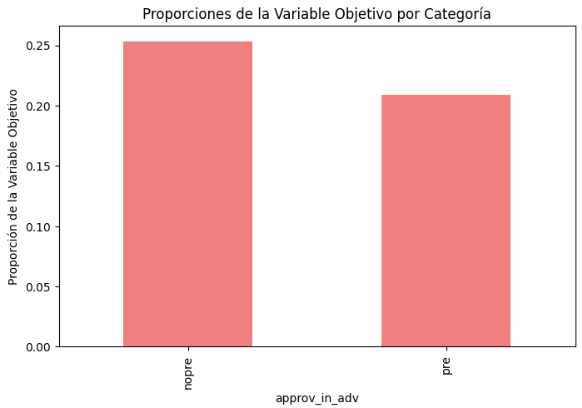
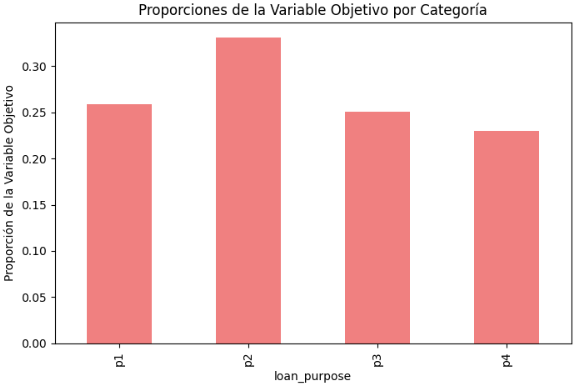
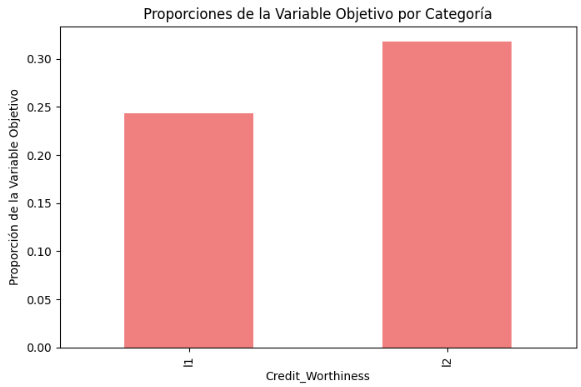
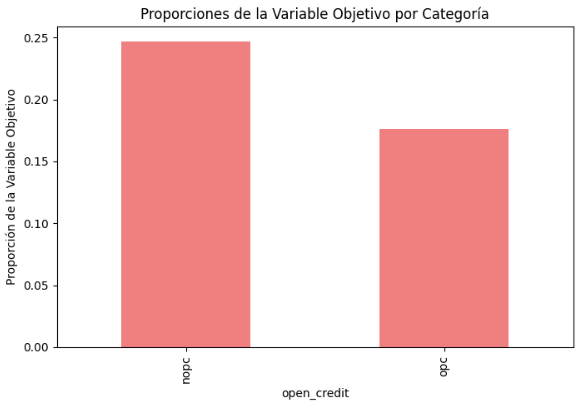
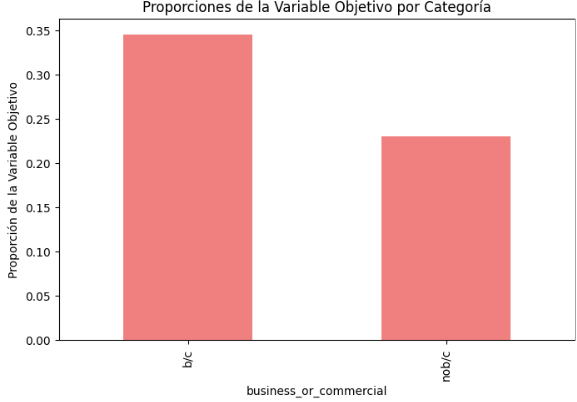
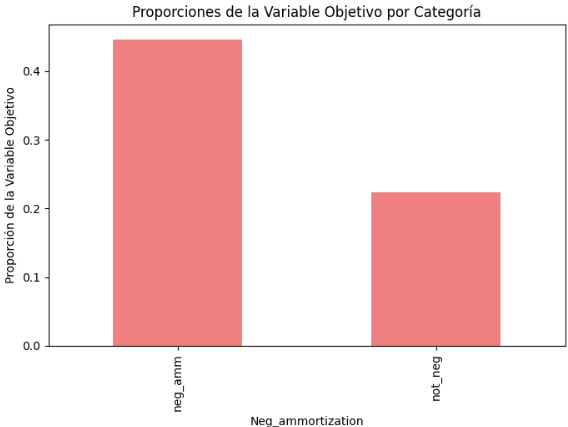
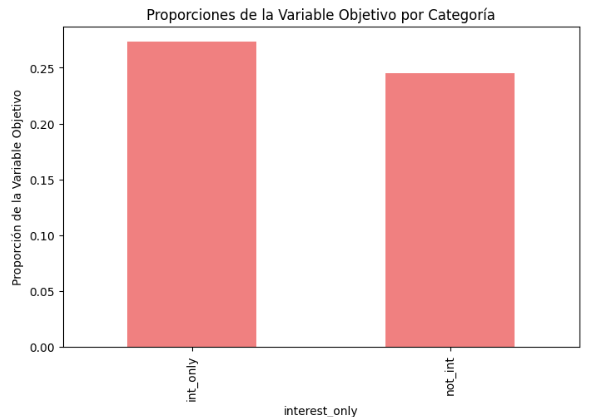
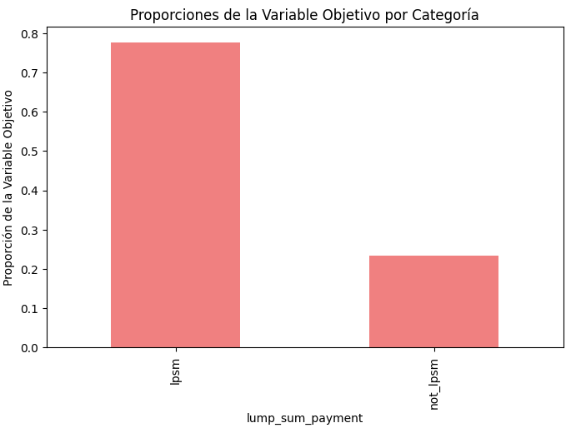
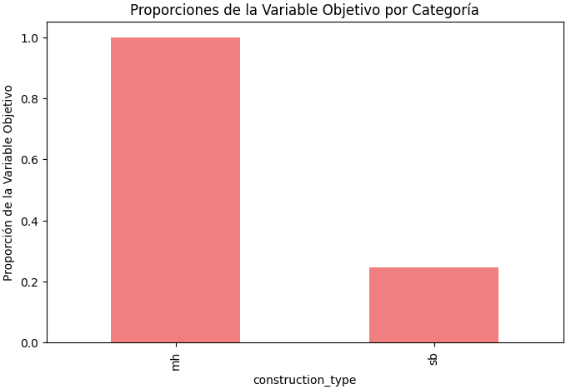
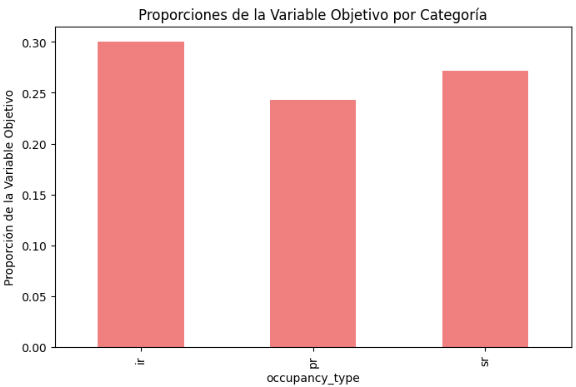
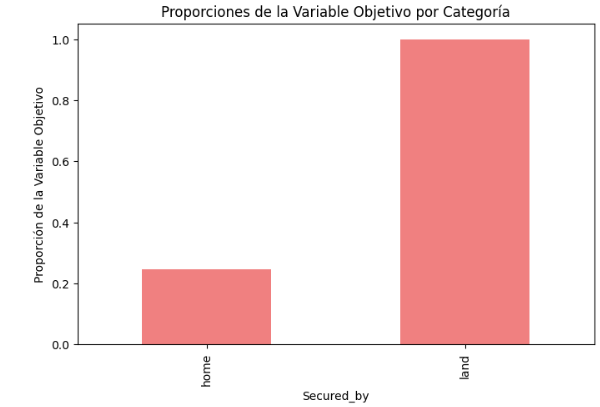
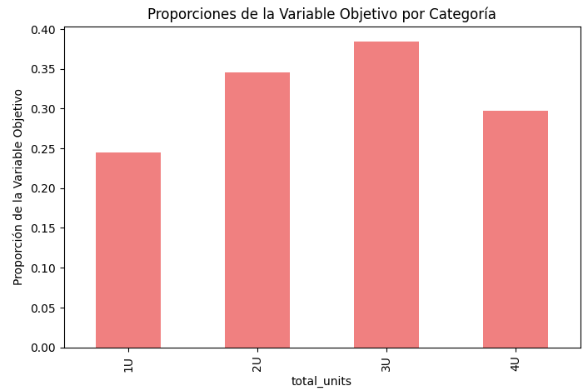
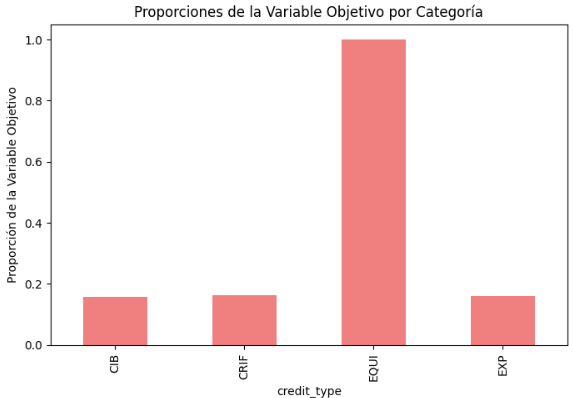


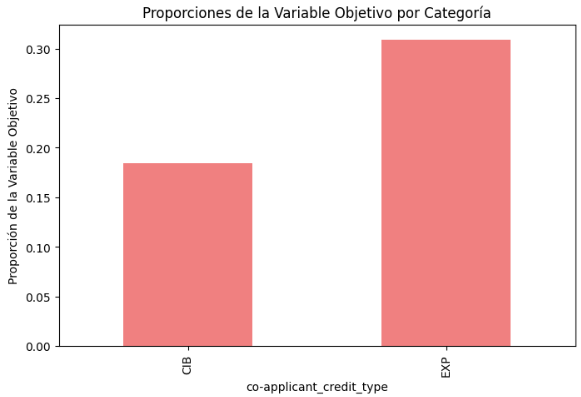
* **dtir1**: 16% de nulos. Variable poco normal. Distribuciones muy parecidas para las dos clases.
* **interest\_rate\_spread**: Distribución bastante normal (un poco de cola a la derecha). Misma información que el rate\_of\_interest.
* **Upfront\_charges**: Muchos valor 0 (no pagan en adelantado), se podría crear una variable binaria (pagan por adelantado o no). Cuanto más se paga de adelantado, menos probable que se vuelva en moroso. Está correlacionada con loan\_amount.

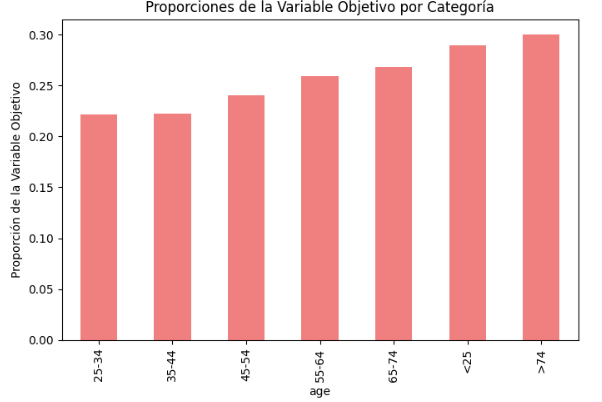
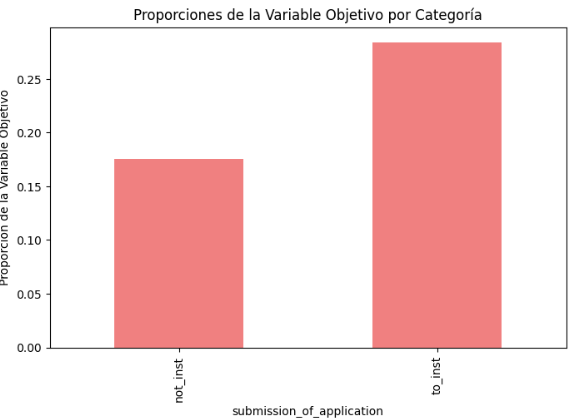
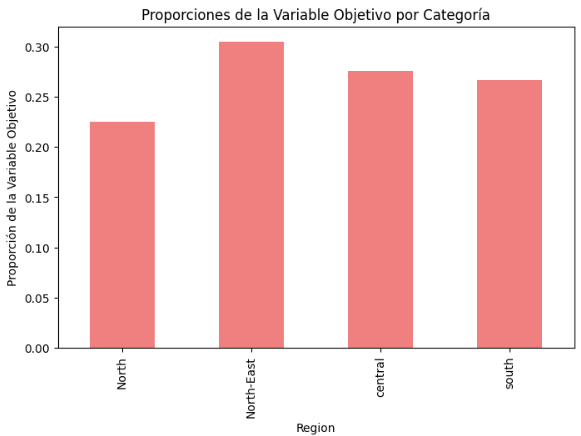


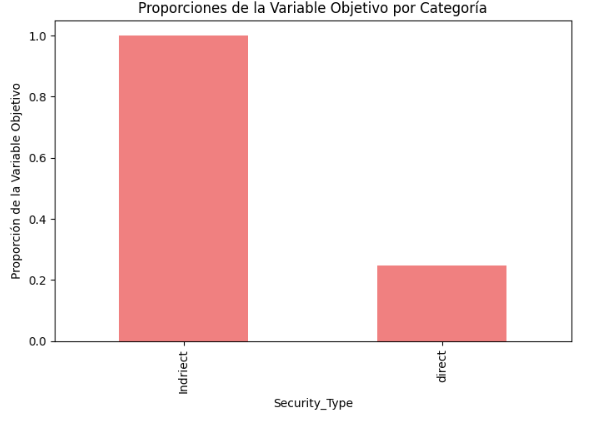
* **term**: Poca varianza respecto al objetivo.

**CATEGÓRICAS**

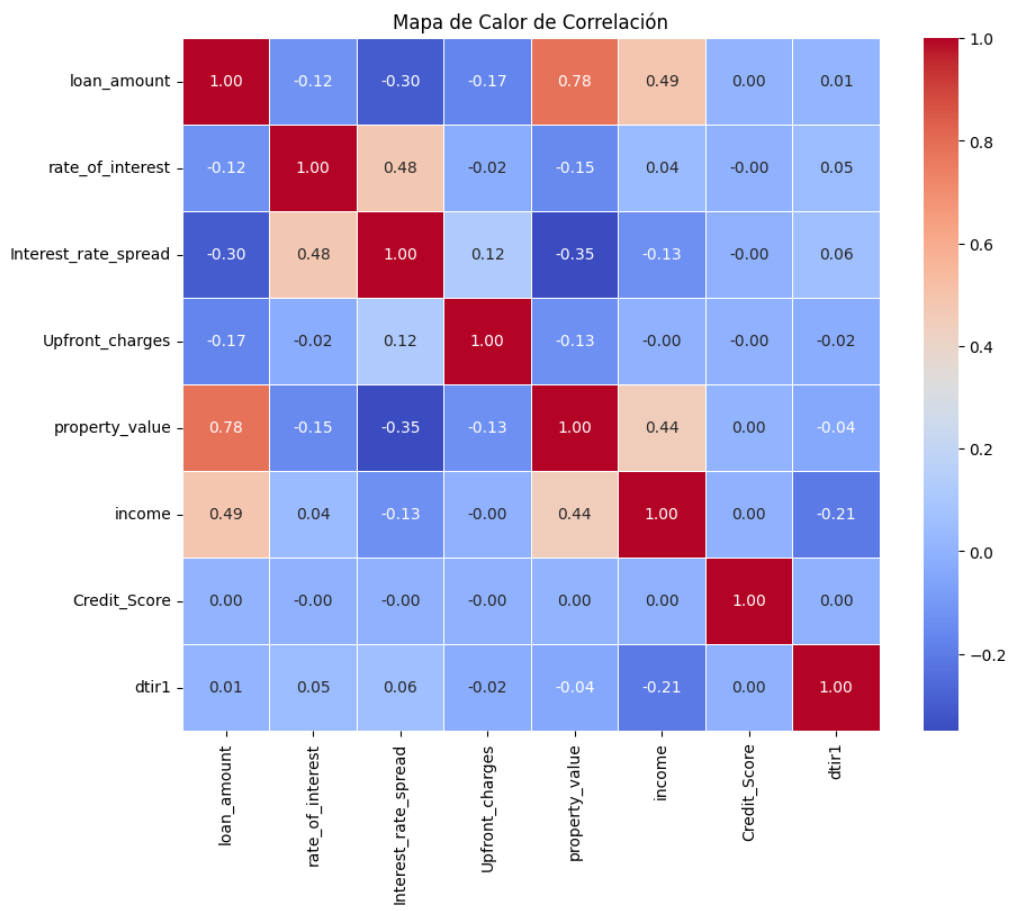
* **year**: Todos los datos son de 2019. Se puede eliminar.
* **loan\_limit**: Muy pocos nulos. Presenta desbalance y la clase minoritaria tiene una tasa de morosidad mayor. Mantén esta variable.
* **Gender**: Sin nulos. Menor morosidad cuando el préstamo es conjunto. Mantén esta variable.
* **approv\_in\_adv**: Muy pocos nulos. Presenta desbalance y los préstamos concedidos tienen una tasa de morosidad menor. Mantén esta variable.
* **loan\_purpose**: Muy pocos nulos. Tiene 4 categorías y la p2 es la que presenta la mayor tasa de morosidad. Mantén esta variable.
* **Credit\_Worthiness**: Sin nulos. Presenta desbalance y la clase minoritaria tiene una tasa de morosidad mayor. Mantén esta variable.
* **open\_credit**: Muy desbalanceada, donde la clase mayoritaria tiene un porcentaje de morosidad mayor. Mantén esta variable.
* **business\_or\_comercial**: Desbalanceada, donde la clase mayoritaria tiene un porcentaje de morosidad mayor. Mantén esta variable.
* **Neg\_ammortization**: Con pocos nulos. Variable de 2 categorías desbalanceada y la clase mayoritaria tiene el doble de tasa de morosidad que la minoritaria. Mantén esta variable.
* **interest\_only**: Sin nulos. Desbalanceada y aporta poca información. Puedes considerar eliminarla.
* **lump\_sum\_payment**: Sin nulos. Desbalanceada y separa un grupo pequeño con una tasa de morosidad alta. Mantén esta variable, pero ten en cuenta el desbalance.
* **construction\_type**: Separa un grupo de 33 préstamos, todos en morosidad. Mantén esta variable, pero ten en cuenta el grupo moroso.
* **occupancy\_type**: Con poca varianza y aporta poca información. Puedes considerar eliminarla.
* **Secured\_by**: Separa un grupo de 33 préstamos con 100% de morosidad. Puede que esté relacionado con "construction\_type". Mantén esta variable, pero ten en cuenta el grupo moroso.
* **total\_units**: 4 categorías desbalanceadas. Da poca información sobre la variable objetivo. Puedes considerar eliminarla.
* **credit\_type**: Separa un grupo de empresas con alta morosidad. Mantén esta variable, pero ten en cuenta el grupo moroso.
* **co-applicant\_credit\_type**: Dos categorías balanceadas, donde una tiene una tasa de morosidad mayor. Mantén esta variable.



* **age**: Las personas mayores de 65 y los menores de 25 son más propensos a caer en morosidad. Mantén esta variable.
* **submission\_of\_application**: Con 13% de nulos. Separa dos categorías; la grande tiene mayor tasa de morosidad que la otra. Mantén esta variable.
* **Region**: 4 categorías desbalanceadas que aportan poca información. Puedes considerar eliminarla.
* **Security\_type**: Separa 33 préstamos morosos. Puede que esté relacionado con "construction\_type". Mantén esta variable, pero ten en cuenta el grupo moroso.



# REDUCCIÓN DE DIMENSIONALIDAD



# 

# 

# 

# 

Viendo la matriz de correlación no podemos quitar ninguna variable numérica, ya que no hay ninguna correlación extremadamente alta en valor absoluto.

En un primer momento vamos a optar por eliminar las variables para las que todos o prácticamente todos los registros toman el mismo valor. Para ello, hemos tenido en cuenta el análisis descriptivo realizado anteriormente.

Las variables que hemos decido eliminar son: year, loan\_limit, Credit\_Worthiness, open\_credit, Neg\_ammortization, interest\_only, lump\_sum\_payment, construction\_type, Secured\_by, total\_units y Security\_Type

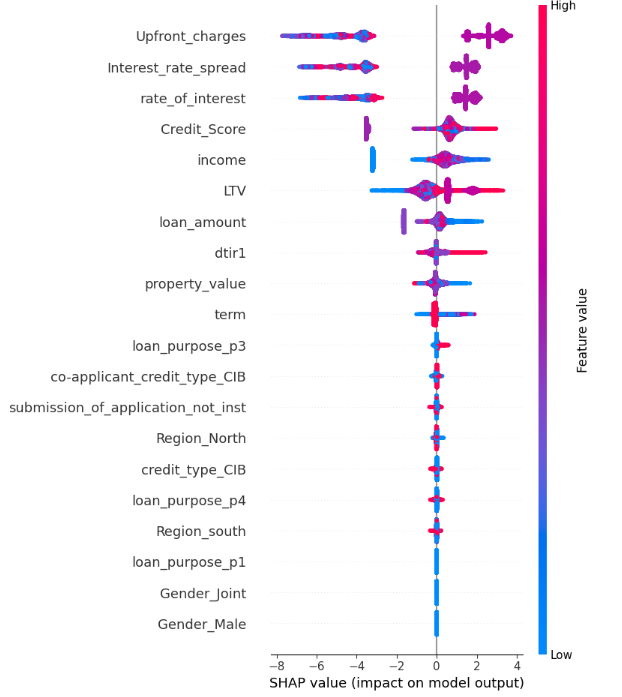
# 

# ANÁLISIS DE LA IMPORTANCIA DE LAS VARIABLES

En el análisis realizado, se ha examinado detalladamente el conjunto de datos,

Identificando aquellos con mayor influencia en la predicción de resultados.

Se ha distinguido entre datos la relevancia de cada tipo, tanto numéricos como categóricos, en la determinación de patrones de comportamiento.



# CONSTRUCCIÓN DEL MODELO

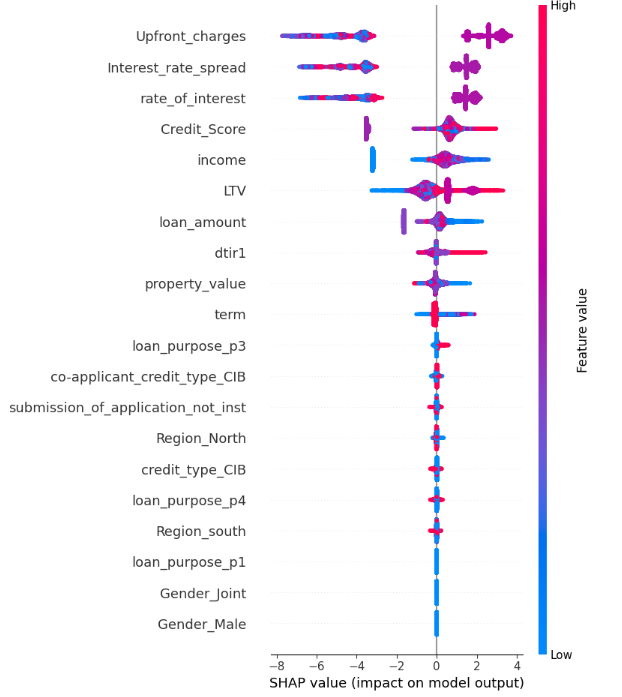
* + Selección y Entrenamiento del Modelo: Utilizaremos el algoritmo RandomForest para la clasificación. Se utilizaran las variables seleccionadas en el análisis anterior.
  + Optimización de los Hiper Parámetros: Ajustaremos los hiper parámetros de los modelos mediante XGBoost con validación cruzada para encontrar la mejor combinación que mejore el rendimiento del modelo.
  + Validación Cruzada y Modelo Entrenado: Aplicaremos la validación cruzada para obtener una estimación robusta del rendimiento del modelo y evitar el sobreajuste.
  + Evaluación del Modelo: Utilizaremos el conjunto de pruebas para evaluar métricas de clasificación como cross validation accuracy\_score y classification\_report de scikit-learn
  + Valores Predichos: Clasificaremos a los nuevos clientes en las categorías de 'Buenos Clientes' o 'Posible Default' basándonos en las predicciones del modelo.
  + Evaluación del Rendimiento del Modelo: Observaremos cómo se comporta el modelo en situaciones reales y haremos ajustes si es necesario para mantener un equilibrio entre aceptar nuevos clientes y minimizar el riesgo de crédito.

# 

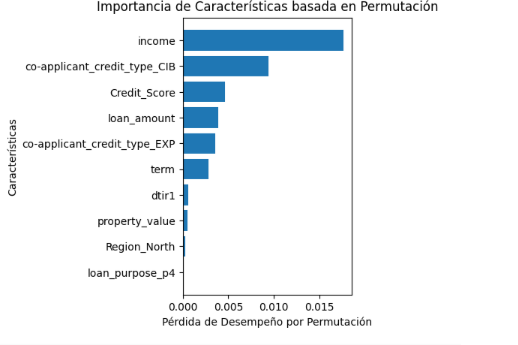
# IMPORTANCIA DE VARIABLES

Una vez hemos construido, optimizado y evaluado el modelo definitivo, en este apartado se tratará de analizar cuales son las variables que más influyen en las predicciones. Para ello, se hará uso de 3 técnicas diferentes:

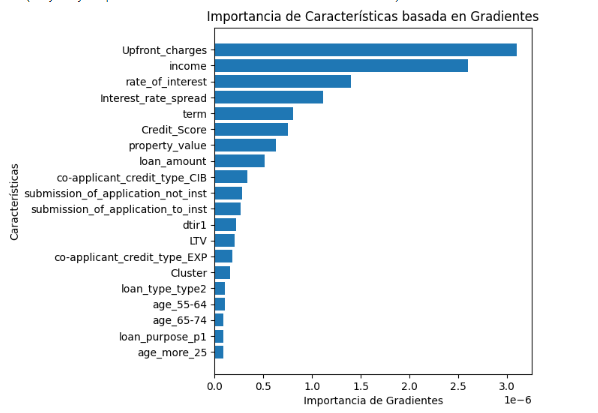
* **Shap values**:Los Shao values son una técnica basada en la teoría de juegos que se utiliza para interpretar la salida de modelos de aprendizaje automático. Específicamente, los valores SHAP asignan una medida de importancia a cada característica, indicando cuánto contribuye cada una a la predicción final del modelo. Esto se logra calculando la contribución marginal de cada característica a través de todas las posibles combinaciones de características, asegurando una explicación justa y consistente. Al proporcionar una descomposición aditiva de la predicción de cada instancia, los valores SHAP permiten entender cómo y por qué un modelo toma ciertas decisiones, facilitando la interpretación, la depuración y la mejora del modelo. Para nuestro modelo, se consigue lo siguiente:



* **Importancia de variables por permutación de variables:** Esta técnica consiste en ir recorriendo variable a variable y volver a entrenar el modelo permutando sus valores. Comparando el rendimiento del modelo normal con el rendimiento del modelo permutado, se puede afirmar que para aquellas variables donde más influye son las más importantes del modelo



* **Gradientes**: Esta técnica calcula la derivada de la salida del modelo con respecto a las entradas. Esto nos da una idea de cuánto cambia la salida del modelo en respuesta a pequeños cambios en cada entrada.



Analizando los resultados obtenidos por las 3 técnicas diferentes, que las 3 están bastante de acuerdo en que las variables más importantes son:

* Upfront\_charges: Cuánto se paga por adelantado
* income: Ingresos del solicitante
* rate\_of\_interest: Tasa de interés del crédito.
* Credit\_score: Rating del cliente evaluado por el departamento de riesgos
* Loan\_amount: Importe del prestamo.

Se puede ver que son las variables que en un primer momento podrían parecer las más importantes. En este caso, la intuición acierta.

# RETURN ON INVESTMENT (ROI)

El “Return on Investment” (ROI), que se traduce al español como “Retorno de la Inversión”, es una métrica financiera utilizada para evaluar la rentabilidad de una inversión. En términos simples, el ROI es una medida del rendimiento generado por una inversión en relación con su costo inicial.

La fórmula básica para calcular el ROI es:

Donde:

* Ganancia se refiere al beneficio generado por la inversión.
* Costo se refiere al monto total invertido.

El ROI se expresa típicamente como un porcentaje o un valor decimal. Un ROI positivo indica que la inversión generó ganancias, mientras que un ROI negativo indica pérdidas.

Es una herramienta importante para los inversores y las empresas, ya que les permite evaluar la eficacia de sus decisiones de inversión y comparar el rendimiento de diferentes oportunidades de inversión. Sin embargo, es importante tener en cuenta que el ROI no proporciona información sobre el plazo de recuperación de la inversión ni sobre el riesgo asociado a ella. Es útil considerar el ROI en conjunto con otras métricas y factores para tomar decisiones informadas sobre inversiones. A continuación, se va a calcular una estimación del ROI de este proyecto paso a paso.

Todos los préstamos de los que se tienen datos (148.670 registros) constan del año 2019. De ellos, 36.639 figuran como “default”, lo que representa el 25% del total de préstamos. Teniendo en cuenta la cantidad de dinero prestado (“loan amount”), la entidad bancaria ha perdido en promedio 319.275,18€ por cada préstamo en default. Esto supone unas pérdidas totales por impago de 11.697.923.500€ en 2019.

Tras este proyecto, consideramos muy factible que la tasa de morosidad baje del 25% al 15% gracias al sistema que desarrollaremos. De esta forma, las pérdidas totales anuales descenderían hasta los 7.119.996.150€. Estaríamos generando una ganancia para el banco de 4.577.927.350€ cada año (11.697.923.500€ - 7.119.996.150€).

Suponiendo que el modelo de predicción sea efectivo durante los próximos 10 años, las ganancias totales serían de aproximadamente 46 mil millones de euros.

En cuanto a los costos, hemos de tener en cuenta diversos aspectos:

* Falsos positivos. Ocurren cuando el modelo predice incorrectamente que un préstamo será incumplido (default), cuando en realidad el préstamo sería pagado correctamente por el prestatario. Suponiendo que el 10% de todos los préstamos solicitados acaben siendo falsos positivos, estimamos que la entidad dejaría de ganar unos 15 mil millones de euros para el periodo de referencia de los 10 próximos años.
* Gastos de personal. Estimamos que se necesitarán 20 empleados dedicados a jornada completa a este proyecto durante un año. La empresa ha de pagar de media 100.000€ por cada uno de ellos durante ese año en concepto de sueldo. Luego entonces, el costo por salarios asciende hasta los 2 millones de euros.
* Mantenimiento. Una vez el proyecto se haya dado por concluido y se implemente el modelo en producción, serán necesarios 3 empleados durante los próximos 10 años que se encarguen de monitorear el funcionamiento de la herramienta y de ofrecer soporte técnico cuando sea requerido. Suponiendo el mismo sueldo, se tiene, por tanto, que hay que dedicar 3 millones de euros al mantenimiento.

Ahora podríamos calcular fácilmente el ROI de este proyecto del siguiente modo:

Esto indica que, por cada euro invertido en el proyecto, se espera obtener un retorno de 2,06 euros adicionales.

# POSIBLES NUEVAS VARIABLES

Para este trabajo, se ha recibido un dataset con unas columnas fijas y se ha trabajado con ellas. En el caso hipotético en el que estuviéramos trabajando en un banco real, tendríamos acceso a una infinidad de otras variables que podrían enriquecer el modelo. A continuación, se recogen algunas de las propuestas que podrían valer para mejorar el rendimiento. Estas variables, son variables comunes en este tipo de modelos en la actualidad.

1. Ingresos del solicitante: En el modelo están considerados los ingresos totales del cliente. Sin embargo, se podrían incluir no solo el ingreso total mensual o anual, sino también la fuente de esos ingresos (trabajo, negocios propios, inversiones, etc.). La variabilidad de los ingresos (por ejemplo, si provienen de empleo por comisión o freelance) también es relevante.
2. Historial crediticio:

● Puntaje de crédito: Un indicador compuesto que refleja la salud crediticia del solicitante, basado en su historial de pagos, la cantidad de crédito utilizado, y la duración del historial crediticio. En la banca actual, se suele asignar una puntuación (a modo de letra AA-, B+, por ejemplo) a la fiabilidad del cliente.

● Préstamos anteriores y su estado: Detalles sobre cualquier préstamo previo, incluyendo montos, tipos, si fueron pagados a tiempo o si hubo retrasos.

● Saldo total de deuda actual: Cuánto debe el solicitante en total, incluyendo tarjetas de crédito, otros préstamos personales, hipotecas, etc

1. Duración del empleo: El tiempo en el empleo actual se puede utilizar como un indicador de estabilidad laboral. Las personas que han estado en su empleo actual durante un periodo más largo suelen tener ingresos más estables.

1. Educación: El nivel de educación alcanzado (por ejemplo, secundaria, universitario, posgrado) puede proporcionar información sobre el potencial de ingresos y la estabilidad financiera del solicitante.
2. Estado civil y dependientes:

● Estado civil: Casado, soltero, divorciado, viudo, etc. Puede influir en las responsabilidades financieras del solicitante.

● Número de dependientes: Cuántas personas dependen económicamente del solicitante, lo que puede afectar la cantidad de ingreso disponible para servir al préstamo.

1. Gastos mensuales: Incluye gastos fijos como alquiler o hipoteca, pagos de automóviles, seguros, gastos de manutención, etc. Esto ayuda a calcular el flujo de caja neto disponible para el pago del nuevo préstamo.
2. Ratio de deuda-ingreso: Este ratio es crítico y se calcula dividiendo la suma total de los pagos mensuales de deudas del solicitante entre su ingreso bruto mensual. Un ratio más alto indica un mayor riesgo de dificultades financieras.
3. Historial de morosidad: Información sobre morosidades previas, incluyendo la frecuencia y la gravedad de los eventos de mora. Esto puede indicar patrones de comportamiento financiero del solicitante.
4. Estabilidad residencial: Cuánto tiempo ha vivido el solicitante en su residencia actual. Cambios frecuentes pueden indicar inestabilidad, mientras que una residencia más prolongada podría implicar mayor estabilidad.
5. Condiciones económicas actuales: Indicadores macroeconómicos como la tasa de desempleo, tasas de interés y otros datos económicos regionales o nacionales que puedan afectar la capacidad de los individuos para pagar sus deudas.

Considerando alguna de estas variables, el modelo mejoraría considerablemente. Se puede dejar como trabajo futuro.

**11. IMPLEMENTACIÓN REAL**

Para valorar la utilidad de una posible implementación real del modelo, es importante analizar tanto los beneficios como los riesgos asociados. Desplegar este modelo puede ofrecer numerosos beneficios. Principalmente, permite la automatización y mejora de la eficiencia operativa al reducir el tiempo de procesamiento y minimizar las tareas repetitivas. Un modelo bien entrenado puede proporcionar insights valiosos basados en datos históricos, permitiendo una toma de decisiones más precisa y fundamentada. Además, la automatización y mejora en la precisión de las decisiones pueden llevar a una reducción significativa de los costos operativos. Por otro lado, al ofrecer un servicio más rápido y preciso, se puede mejorar considerablemente la satisfacción del cliente, lo que se traduce en una experiencia más positiva y una mayor fidelidad.

No obstante, implementar el modelo también implica ciertos riesgos, especialmente en términos de falsos positivos y errores de clasificación. Los falsos positivos, donde el modelo identifica incorrectamente a clientes de alto riesgo como de bajo riesgo, pueden resultar en la aprobación de créditos a personas que probablemente no puedan cumplir con sus obligaciones financieras, aumentando así las tasas de morosidad y las pérdidas para la empresa. Por otro lado, los falsos negativos pueden llevar a rechazar clientes que en realidad son de bajo riesgo, perdiendo así oportunidades de negocio y potenciales ingresos. Estos errores no solo impactan financieramente, sino que también pueden dañar la reputación de la empresa, disminuyendo la confianza del cliente. Además, errores en la clasificación pueden resultar en problemas de cumplimiento con las regulaciones financieras, lo que podría acarrear sanciones y multas.

Una posible segunda fase del proyecto debería centrarse en el refinamiento del modelo. Esto puede incluir ajustar los hiperparámetros, incorporar más datos y aplicar técnicas avanzadas de machine learning para mejorar la precisión. Además, es crucial realizar pruebas extensivas en entornos controlados para evaluar el rendimiento y la fiabilidad del modelo antes de su implementación completa. Una vez desplegado, se debe establecer un sistema de monitoreo continuo para evaluar el desempeño del modelo en tiempo real y realizar los ajustes necesarios para mantener su eficacia. También se puede considerar la expansión de las funcionalidades del modelo a otras áreas operativas, como la detección de fraude y la personalización de ofertas. Finalmente, es esencial capacitar al personal para que pueda interpretar correctamente los resultados del modelo y tomar decisiones informadas basadas en estos insights.

En resumen, aunque la implementación del modelo ofrece claros beneficios en términos de eficiencia, reducción de costos y mejora de la satisfacción del cliente, es crucial gestionar cuidadosamente los riesgos asociados con los falsos positivos y errores de clasificación. La segunda fase del proyecto debe centrarse en el refinamiento, validación, monitoreo y expansión del modelo, asegurando al mismo tiempo que el personal esté bien capacitado para maximizar el valor derivado del modelo.

# 12. REFERENCIAS

**Loan default predictability with explainable machine learning**

Este estudio investiga la capacidad predictiva de nueve modelos de aprendizaje automático para predecir el riesgo de incumplimiento de préstamos. Se encuentra que el modelo de random forest tiene una eficiencia y estabilidad destacadas en la predicción. Además, se emplea el método SHAP para comprender los factores importantes que afectan a los incumplimientos de préstamos, encontrando que la edad y los años de experiencia laboral del prestamista son factores significativos que reducen el riesgo de incumplimiento. Este análisis extenso proporciona información valiosa para las instituciones financieras en la gestión del riesgo crediticio.

Hemos detectado que en este proyecto han empleado variables predictivas que no hemos tenido en cuenta en nuestro trabajo, como por ejemplo, la profesión o los años en el puesto de trabajo actual.

(Li and Wu)

**Explainable prediction of loan default based on machine learning models**

El estudio examina la creciente tendencia de préstamos en línea y la importancia de modelos predictivos para predecir incumplimientos. Se comparan modelos como regresión logística, árboles de decisión, XGBoost y LightGBM, destacando el rendimiento superior de los dos últimos. Se enfatiza la necesidad de interpretabilidad en los modelos y se utiliza el enfoque LIME para analizar los hallazgos, identificando factores como el plazo del préstamo, la calificación crediticia y el monto del préstamo como determinantes clave de los resultados predichos.

(Zhu et al.)

# BIBLIOGRAFÍA

Li, Huan, and Weixing Wu. “Loan default predictability with explainable machine learning.” *Science Direct*, 5 March 2024, <https://www.sciencedirect.com/science/article/abs/pii/S1544612323012394>

Accessed 10 May 2024.

Zhu, Xu, et al. “Explainable prediction of loan default based on machine learning models.” *Science Direct*, 5 March 2024, <https://www.sciencedirect.com/science/article/pii/S2666764923000218>

Accessed 10 May 2024.