



CIENCIA DE DATOS

Modelo clasificación

Por: Javier Rodríguez Cifuentes

TABLA DE CONTENIDO



01

Acerca del Proyecto

El reto de la probabilidad de incumplimiento en crédito.

02

Requisitos principales

La consigna de los desafíos

03

Objetivo del proyecto

El modelo a desarrollar y su fin

04

Vistazo analítico

Algunas visualizaciones relevantes y hallazgos.

05

Etapas del proyecto

Hitos principales del proyecto

06

Modelado

Hallazgos y experimentación para selección modelo



HomeCredit

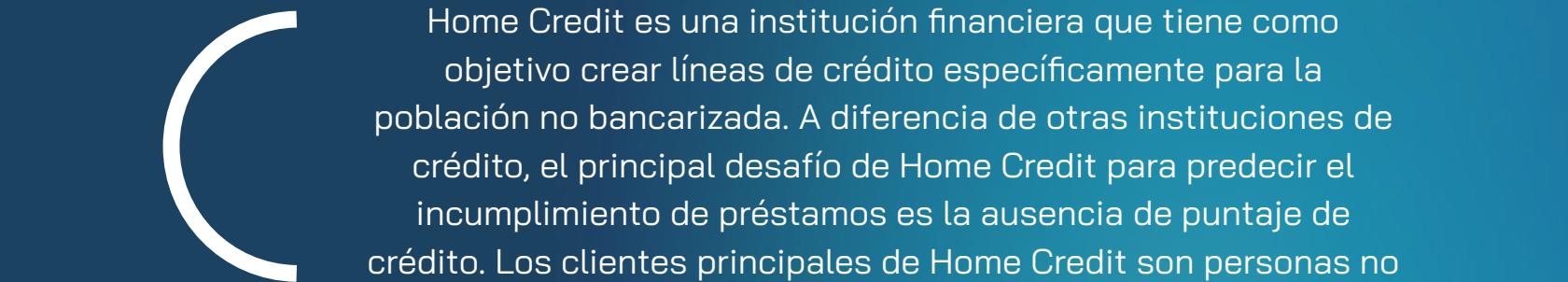
Home Credit es una empresa financiera internacional que se especializa en proporcionar servicios de financiamiento al consumidor, particularmente enfocados en la población no bancarizada o con acceso limitado a servicios bancarios tradicionales. La empresa opera en múltiples países de Europa, Asia, América Latina y otras regiones.





INTRODUCCIÓN

Home Credit es una institución financiera que tiene como objetivo crear líneas de crédito específicamente para la población no bancarizada. A diferencia de otras instituciones de crédito, el principal desafío de Home Credit para predecir el incumplimiento de préstamos es la ausencia de puntaje de crédito. Los clientes principales de Home Credit son personas no bancarizadas o con acceso limitado a servicios bancarios, que cuentan con un historial crediticio muy limitado. El objetivo de este proyecto es construir un **modelo de clasificación** eficaz y eficiente para predecir el incumplimiento del pago del crédito de los solicitantes de préstamos y mitigar el riesgo crediticio para Home Credit.



01

Acerca del proyecto

Contexto de los datos



Origen de los datos

Recopilamos nuestros datos de Kaggle, una comunidad en línea que permite a los profesionales de la ciencia de datos acceder a datos públicos y publicar sus soluciones. En total, se proporcionan siete tablas con el problema pero la más relevante y que se trabajará en este notebook es la que se denomina “application”.





Datos de entrenamiento

- ❖ “Aplication” es la tabla principal dividida en dos archivos uno para entrenamiento y otro para test. En el proyecto trabajamos netamente con el archivo de entrenamiento. El archivo de entrenamiento contiene una columna adicional llamada TARGET que nos indica si el cliente ha pagado el préstamo o no (1 es incumplimiento, 0 es pago).
- ❖ 'application_train' se usará para realizar análisis, construir un modelo y probar resultados. Esta tabla contiene la variable objetivo llamada “TARGET” y la información demográfica del cliente más otros datos importantes sobre el cliente, que incluyen si el solicitante es propietario de una casa/automovil, el número de miembros de la familia, etc



Elementos relevantes

Modelo de clasificación
para predecir
incumplimiento (riesgo)

Demasiadas variables
irrelevantes, valores nulos
y desbalance de clases

Alcance

Limitaciones



Motivación

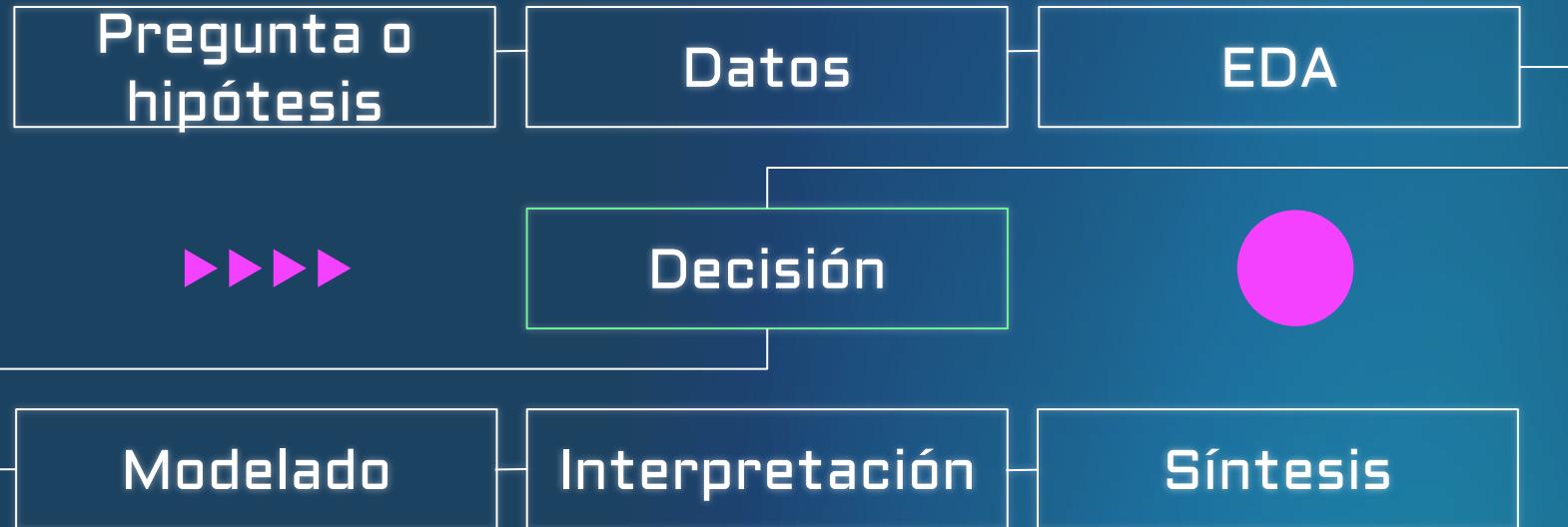
Evaluar riesgo crediticio
de solicitantes de
crédito para balancear
inclusión, exposición y
rentabilidad.

Audiencia

Consumidores no bancarizados
o con acceso limitado a
servicios financieros



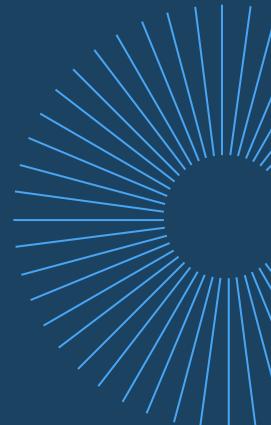
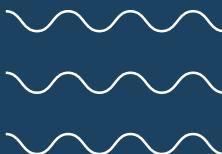
Arquitectura del proyecto de DS





“En ciencia de datos, no se trata solo
de encontrar respuestas, sino de
descubrir las preguntas correctas”

—Jeff Hammerbacher



Contexto Comercial



El contexto comercial del negocio de Home Credit se centra en la provisión de servicios financieros y créditos al consumo en mercados emergentes y en desarrollo. Home Credit opera en diversos países de Europa, Asia y América, y su modelo de negocio se caracteriza por varios aspectos clave:

- **Atención a población sub-atendida en el mercado financiero tradicional**
- **Amplia gama de productos de crédito con foco en consumo.**
- **Involucramiento en iniciativas de responsabilidad social.**
- **Base de clientes diversa en mercados emergentes y en desarrollo.**

Problema Comercial

El problema comercial clave que enfrenta Home Credit, al igual que otras instituciones financieras, es el riesgo crediticio.

Desafíos	<ul style="list-style-type: none">• Riesgo de incumplimiento• Equilibrio entre acceso a crédito y mitigación de riesgo• Recuperación de cartera y reducción de pérdidas por incumplimiento• Atracción y retención de clientes con una voluntad genuina de pago(buenos pagadores)
Metodología	Un conjunto de pasos y procesos organizados que se siguen al aplicar técnicas de aprendizaje automático.
Objetivo & Métricas	Lograr un modelo de clasificación con un desempeño en AUC superior al 70%



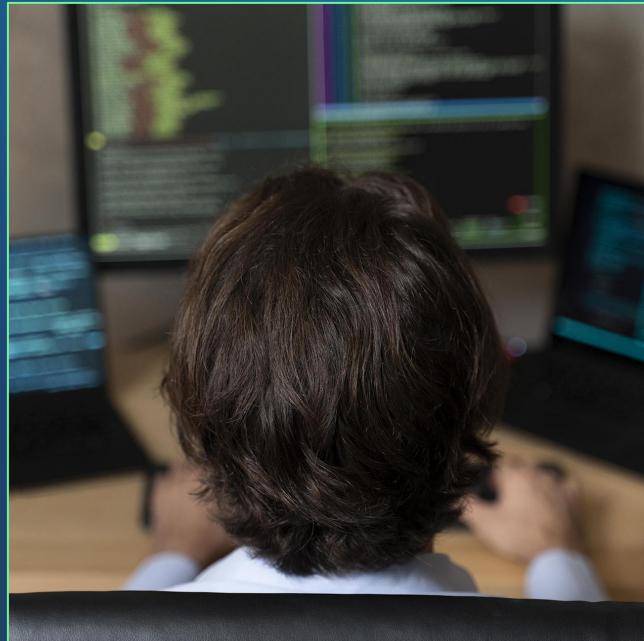


Hoy

HomeCredit atiende a población no bancarizada y tiene importantes retos para contener la cartera en mora desde el primer pago.

Futuro

Con modelos de clasificación cada vez mejores espera poder dar mayor acceso a crédito sin generar mayores pérdidas para la compañía en función de una probabilidad de incumplimiento controlada.



Borrador actividades proyecto

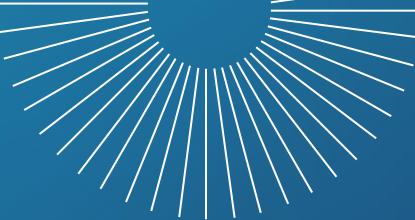
How a work plan
is organized:

- Elección de dataset potencial
- Visualización de variables
- Análisis exploratorio inicial
- Análisis exploratorio ampliado
- Presentación inicial del proyecto de DS
- Entrenamiento de modelos base ML
- Evaluación modelos ML
- Primera entrega proyecto final



Requisitos principales

Consigna de los desafíos



2



Requisitos principales



Datos

- Al menos 2000 registros
- Al menos 15 variables



Notebook

- Que contenga markdowns de documentación
- Preguntas, hipótesis e interpretación de resultados.

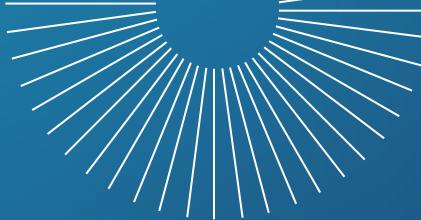


Presentación

Incorporar elementos como:

- Objetivo
- Problema
- Contexto
- Análisis





Objetivo del proyecto DS

3



El modelo a desarrollar y su fin



Objetivo del Proyecto

Construir un **modelo de clasificación** eficaz y eficiente para predecir el incumplimiento de pago de los solicitantes de préstamos y mitigar el riesgo crediticio para Home Credit, mejorando la toma de decisiones crediticias y gestionando de manera efectiva el riesgo crediticio al ofrecer líneas de crédito a la población no bancarizada.



Reto 1

Evaluación de riesgo crediticio precisa



Reto 2

Inclusión financiera



Reto 3

Reducción de pérdidas por el incumplimiento



Reto 4

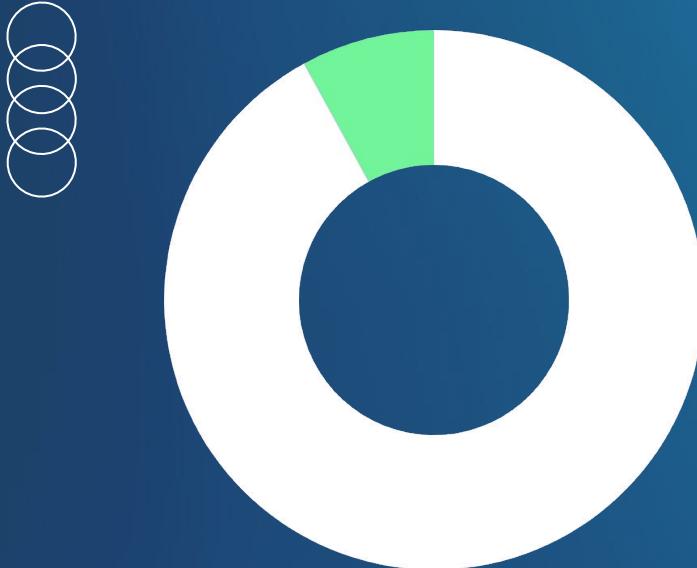
Personalización de ofertas de crédito

Vistazo analítico

Algunas visualizaciones relevantes y hallazgos



Variable TARGET



Buenos pagadores 92%

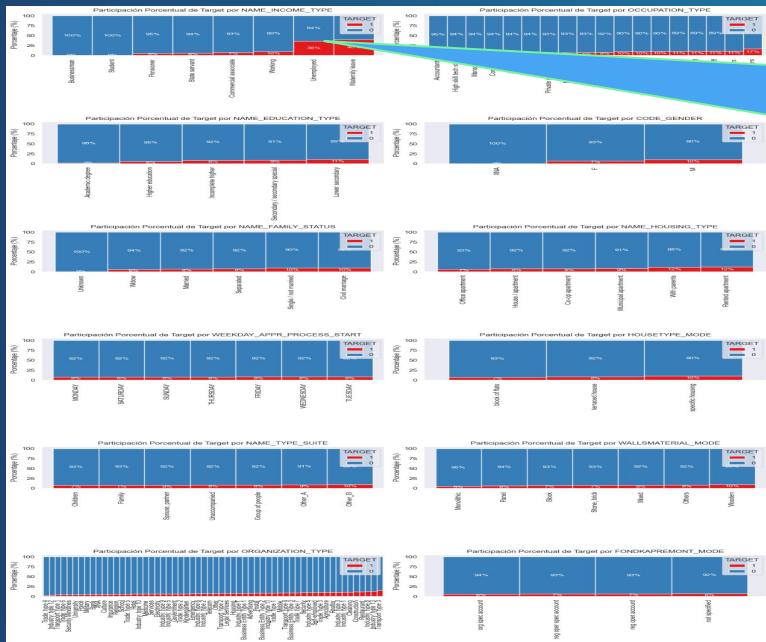
Pagan sin entrar en incumplimiento

Malos pagadores 8%

Entran en incumplimiento por no pago



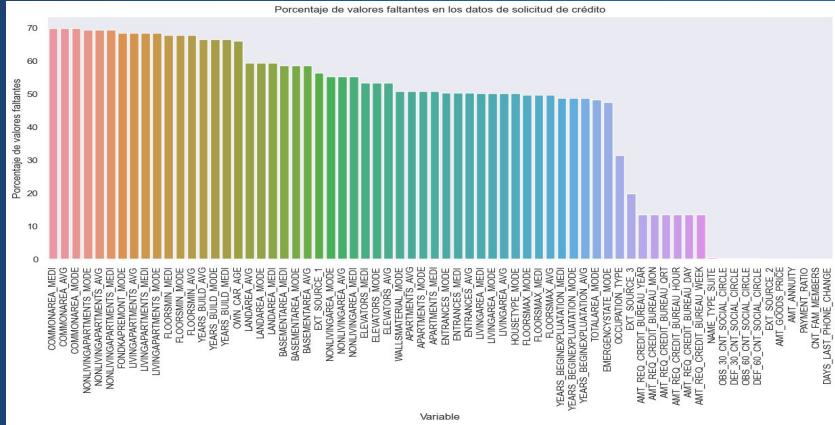
Variables categóricas vs TARGET



El tipo de ingreso (NAME_INCOME_TYPE) tiene un gran impacto en el riesgo de incumplimiento. Las personas desempleadas (Unemployed) o en licencia por maternidad (Maternity Leave) tienen un riesgo de incumplimiento significativo.

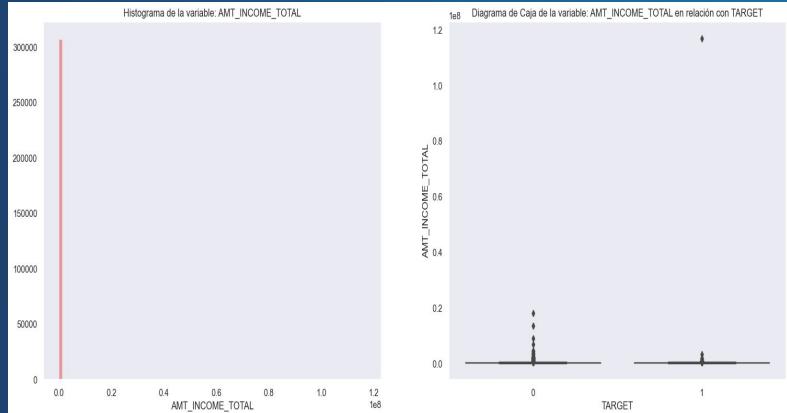


Valores faltantes



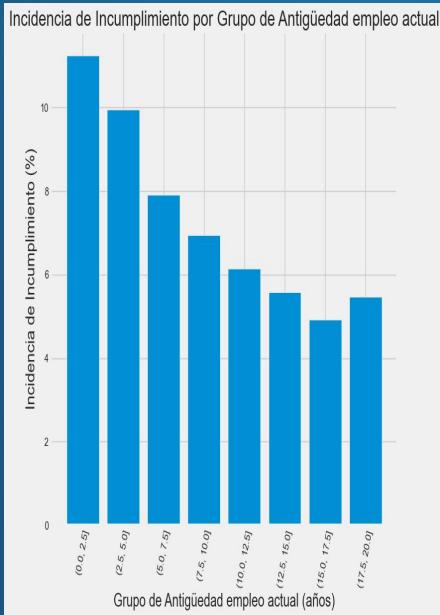
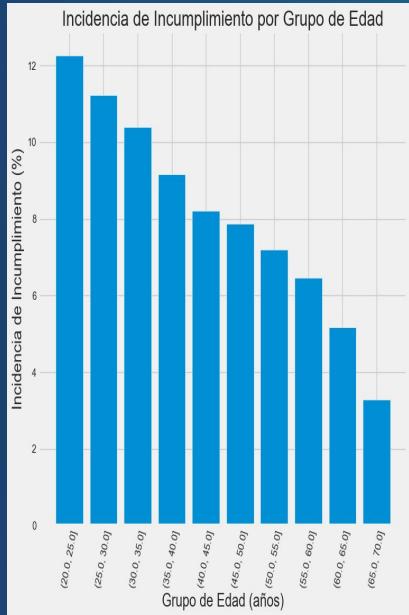
Algunas variables tienen una gran cantidad de valores perdidos (cerca del 70%). Sería interesante saber si el número de valores faltantes para cada cliente tiene un impacto en el rendimiento del modelo predictivo.

Valores atípicos

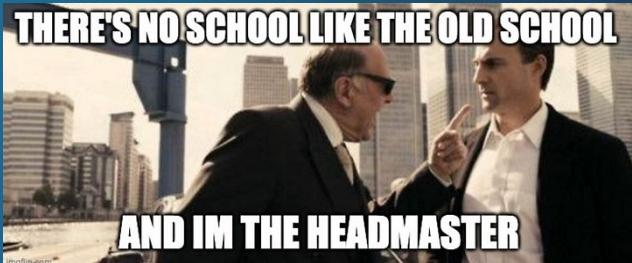


Hay valores atípicos en los datos: algunos clientes han estado empleados durante un número negativo de días. Algunos clientes pueden tener un ingreso superior a \$100 millones. Será importante identificar y tratar adecuadamente estos valores atípicos para evitar que distorsionen las predicciones del modelo de clasificación.

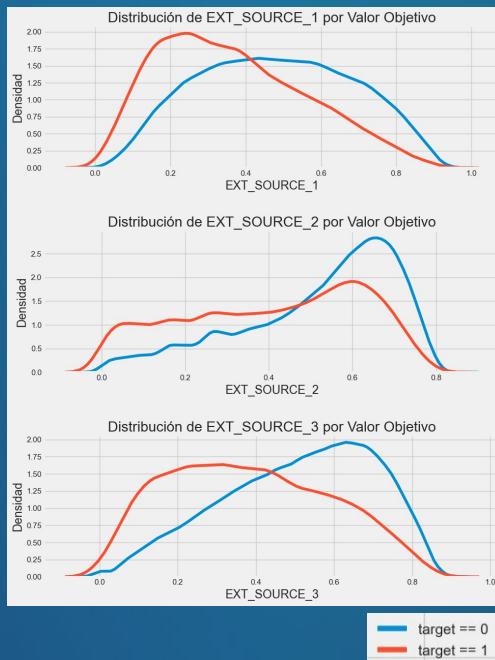
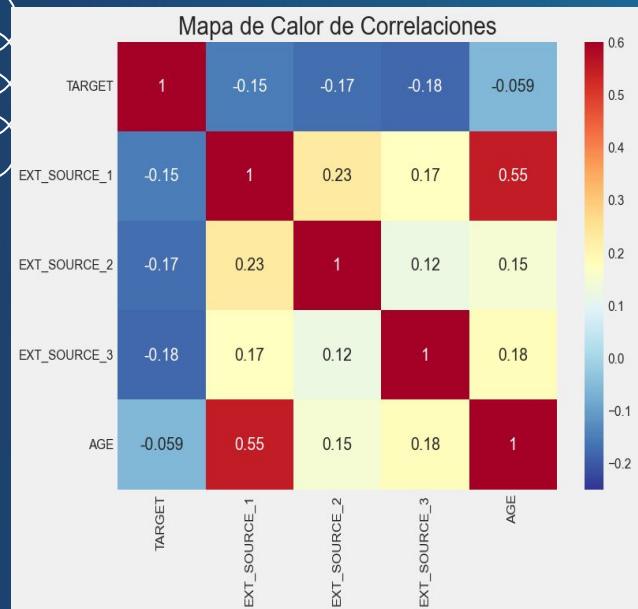
¿Los años tienen efecto en el impago?



Hay dos variables que contundentemente muestran que si hay efecto. Y en realidad entre más años mejor. Es decir, ¡que viva la vieja escuela!



¿Son relevantes los scores externos?



- Los scores externos son relevantes . Esto se evidencia en la correlación que tiene con una menor probabilidad de impago para puntajes altos (mayor en ext_source_3).
- El score ext_source_1 tiene una fuerte correlación positiva con la edad.
- El análisis visual se realiza sin imputar valores nulos dado que afectan tanto el resultado de correlaciones como la distribución de densidad de probabilidad.



Hallazgos análisis exploratorio

Insights impago en función del tipo de ingreso

- El tipo de ingreso impacta significativamente el riesgo de incumplimiento de pagos.
- Las personas no empleadas o en licencia de maternidad tienen la mayor probabilidad de incumplimiento de los pagos.

Insights impago en función de los años

- Los aplicantes que acumulan más años tienen menor probabilidad de impago.
- Las variables de años que ordenan mejor la probabilidad de impago son la edad y la antigüedad en el empleo actual.

Insights impago en función de scores externos

- Los scores externos son variables relevantes para la predicción del impago.
- El score ext_souce_1 tiene una correlación positiva fuerte con la edad en años.

Recomendaciones modelado

- Debería evidenciarse la importancia de las variables identificadas en hallazgos.
- Revisar impacto del tipo de preprocesamiento en el desempeño del modelo.



Etapas del proyecto

Hitos principales del proyecto

Hitos del proyecto

ID	Tasks	Hito 1	Hito 2	Hito 3	Hito 4
1	Selección dataset				
2	Procesamiento inicial datos				
3	Análisis de los datos				
4	Modelado				
5	Presentación				

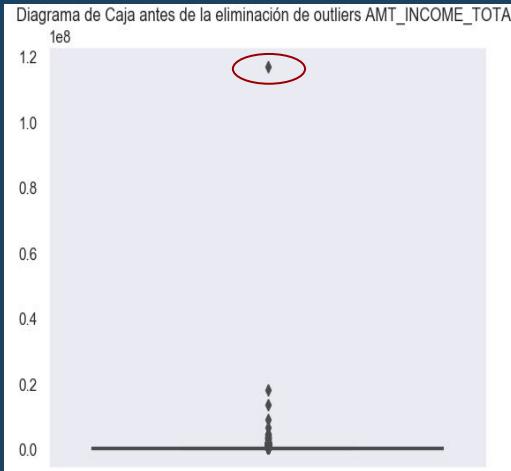
Modelado

Hallazgos y experimentación modelos

Datos atípicos, duplicados y nulos

Atípicos

Visualiza el valor extremo



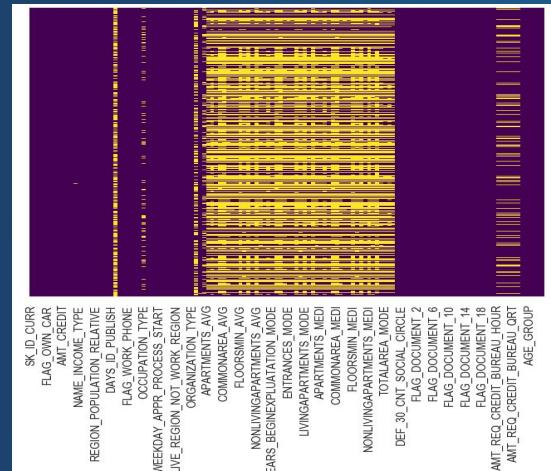
Duplicados

Indice de filas duplicadas

```
# Identificar duplicados  
duplicados = df[df.duplicated()]  
  
# Mostrar los duplicados encontrados  
print("Filas duplicadas:")  
print(duplicados.index)  
duplicados.shape  
  
Filas duplicadas:  
Int64Index([221256, 221257, 221258, 221259],
```

Nulos

Resalta nulos en amarillo



Tratamiento de atípicos, duplicados y nulos

Atípicos

Visualización con diagrama de cajas y de dispersión(ruido vs variable) para facilitar visualización de los atípicos y Eliminación en función del rango intercuartílico

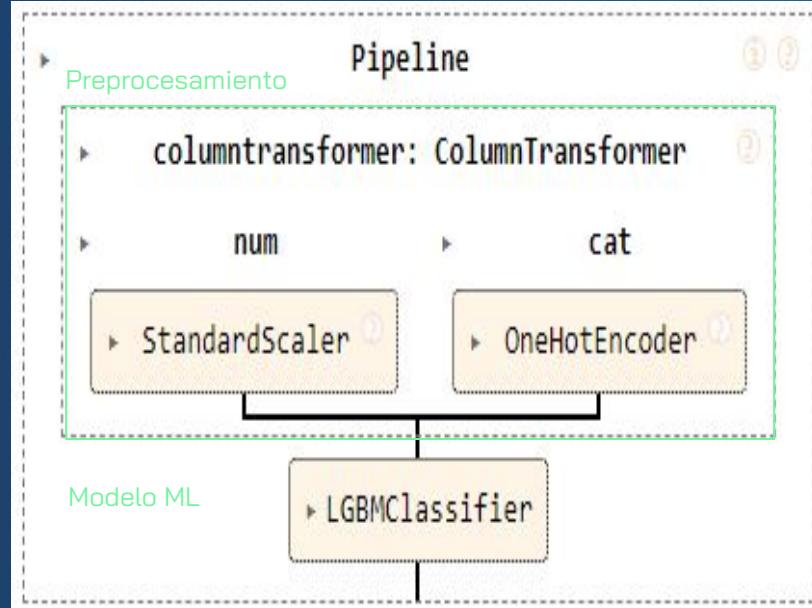
Duplicados

Identificación con el método `duplicated()` y eliminación con el método `drop_duplicates()`

Nulos

Imputación con un código en variables numéricas (-999) y en variables categóricas ('-999')

Preprocesamiento

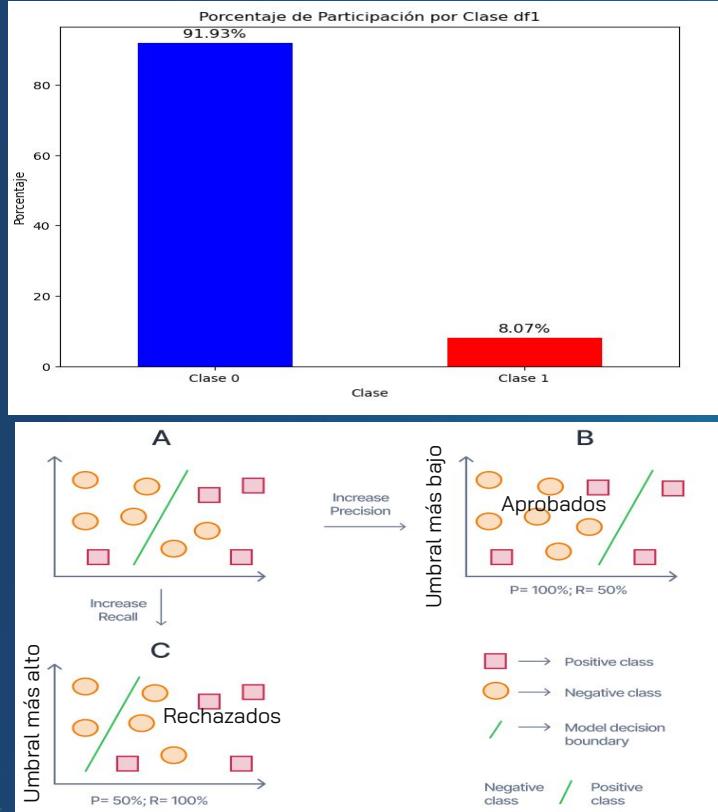


Pipeline: Secuencia de pasos de que se aplica a los datos.

Preprocesamiento: Procesos o pasos aplicados a los datos antes de entrenar el modelo de machine learning.

- StandardScaler: Estandariza las variables numéricas del conjunto de datos para que tengan una media de cero(0) y una desviación estándar de uno(1). Útil en especial para variables de valores monetarios con diferencias de millones.
- OneHotEncoder: Convierte variables categóricas en una representación numérica. Se representan como un vector binario con 1 en la categoría presente y 0 en las demás.

Tratamiento del desbalance de clases



Desbalance de clases: El número de casos de una clase es significativamente menor que el de la(s) otra(s).

Técnicas para tratar con el desbalance de clases:

Se utilizan las técnicas de ponderación de clases y punto de corte de la probabilidad de incumplimiento (probabilidad de clase 1). No se utilizan técnicas de generación de datos sintéticos porque el conjunto de datos es grande y aumenta el costo computacional.

- **Ponderación de clases:** Se asigna un mayor peso a la clases minoritaria. Inversamente proporcional a su frecuencia en el conjunto de datos.
- **Punto de Corte:** Se ajusta el umbral de decisión del modelo para lograr un balance entre falsos positivos (precisión) y falsos negativos (recall).

Ingeniería de variables

Con un enfoque de negocio podemos crear algunas variables calculadas que intenten capturar lo usualmente se cree en la industria financiera que puede ser importante para determinar si un cliente incumplirá con un préstamo. Aquí se van a utilizar algunas variables calculadas como las siguientes:

- **TIMES_INCOME**: El número de veces que el monto del préstamo supera los ingresos del cliente.
- **ANUITY_INCOME_PERCENT**: El porcentaje de la cuota del préstamo en relación con los ingresos del cliente
- **CREDIT_TERM**: La duración o plazo del pago en meses (ya que la anualidad es la cuota mensual de la deuda)
- **YEARS_EMPLOYED_PERCENT**: El porcentaje de los años de antigüedad como empleado en relación con la edad del cliente
- **INCOME_PER_PERSON**: El ingreso promedio por miembros del hogar o familia.
- **SCOREAVG**: Es el promedio de los tres puntajes del cliente en fuentes externas 'EXT_SOURCE_1', 'EXT_SOURCE_2' y 'EXT_SOURCE_3'.



Se están probando otras variables relacionadas con las que han dado más importantes que están relacionadas con razones financieras, ratios de tiempo y scores.

-



Métodos de selección de variables

- El método más eficiente en costo computacional de hallar las variables seleccionadas, costo computacional de entrenamiento con las variables seleccionadas y mejor desempeño es el de Boruta (basado en modelo RF para la selección).
- Usar conjuntamente Boruta y luego Forward requiere un alto procesamiento computacional de selección, pero menor que solo usar la búsqueda exhaustiva Forward.

Base original	Método Basado en modelos: Boruta	Método univariado: Correlación	Método de reducción dimensionalidad: PCA	Método de búsqueda exhaustiva: Forward
#Variables: 127 AUC: 0.788 Gini: 0.517 Tiempo CPU: 1 min 39s	#Variables: 44 AUC: 0.7561 Gini: 0.512 Tiempo CPU: 42.6s	#Variables: 86 AUC: 0.748 Gini: 0.5096 Tiempo CPU: 1 min 17s	#Variables: 44 componentes AUC: 0.7355 Gini: 0.4710 Tiempo CPU: 1 min 44s	#Variables: 29 AUC: 0.757 Gini: 0.5115 Tiempo CPU: 33.8s

Selección de variables de Boruta

- El método Boruta evalúa importancia de variables originales comparándola con sus correspondientes variables sombra(copias permutadas) mediante un algoritmo de aprendizaje automático (generalmente bosques aleatorios).
- Las variables sombra se usan en el método de Boruta para comparar la importancia de las variables originales con respecto a su importancia aleatoria.

Las variables seleccionadas o de importancia alta son las siguientes 44:

[**'AMT_CREDIT'**, '**'AMT_ANNUITY'**', '**'AMT_GOODS_PRICE'**', '**'REGION_POPULATION_RELATIVE'**', '**'DAYS_REGISTRATION'**, '**'EXT_SOURCE_1'**', '**'EXT_SOURCE_2'**', '**'EXT_SOURCE_3'**', '**'APARTMENTS_AVG'**', '**'FLOORSMAX_AVG'**', '**'LIVINGAREA_AVG'**', '**'APARTMENTS_MODE'**', '**'YEARS_BEGINEXPLUATATION_MODE'**', '**'FLOORSMAX_MODE'**', '**'LIVINGAREA_MODE'**', '**'APARTMENTS_MEDI'**', '**'YEARS_BEGINEXPLUATATION_MEDI'**', '**'FLOORSMAX_MEDI'**', '**'LIVINGAREA_MEDI'**', '**'TOTALAREA_MODE'**', '**'DAYS_LAST_PHONE_CHANGE'**', '**'PAYMENT_RATIO'**', '**'AGE'**', '**'YEARS_EMPLOYED'**', '**'YEARS_REGISTRATION'**', '**'YEARS_ID_PUBLISH'**', '**'TIMES_INCOME'**', '**'ANNUITY_INCOME_PERCENT'**', '**'CREDIT_TERM'**', '**'LOAN_TO_VALUE'**', '**'EMPLOYED_AGE_PERCENT'**', '**'SCOREMIN'**', '**'SCOREMAX'**', '**'SCOREAVG'**', '**'SCOREMEDIAN'**', '**'SCOREVAR'**', '**'SCORESUM'**', '**'DOCUMENT_MEAN'**', '**'DOCUMENT_VAR'**', '**'DOCUMENT_KUR'**', '**'NAME_CONTRACT_TYPE'**', '**'CODE_GENDER'**', '**'FLAG_OWN_CAR'**', '**'FLAG_OWN_REALTY'**]

Las variables tentativas o de importancia débil son las siguientes 6:

[**'YEARS_BEGINEXPLUATATION_AVG'**, '**'ELEVATORS_AVG'**', '**'ENTRANCES_MODE'**', '**'ELEVATORS_MEDI'**', '**'ENTRANCES_MEDI'**', '**'DEF_30_CNT_SOCIAL_CIRCLE'**]

Selección de variables de Boruta

- Teniendo menos columnas por la selección de variables de boruta ahora se evalúa el efecto que tiene trabajar con menos registros del conjunto de datos mediante la curva de aprendizaje (learning_curve).
- Se observa que mínimo con 50.000 registros se podría entrenar un modelo relativamente competitivo, no obstante arriba de 100.000 registros es donde se tiene menor brecha de sesgo y varianza. El accuracy converge alrededor de 0.91 y el AUC alrededor de 0.76. Mayor brecha en AUC.



Algoritmos de clasificación



- Los algoritmos con mejor AUC son el Catboost y el LightGBM, superior en el CatBoostClassifier.
- En tiempo de procesamiento es más rápido el LGBMClassifier aunque vale la pena compararlo a detalle como score de probabilidad de incumplimiento.

TensorFlow

#Variables: 50
AUC: 0.74
Tiempo CPU: 25.2s

SVM

#Variables: 50
AUC: 0.64
Tiempo CPU: 29m 12s

CatBoost

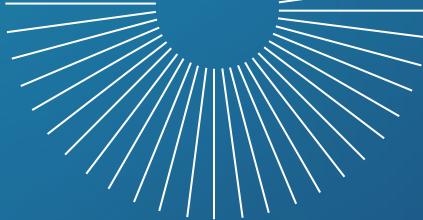
#Variables: 50
AUC: 0.76
Tiempo CPU: 43.78s

XGBoost

#Variables: 50
AUC: 0.75
Tiempo CPU: 3.80s

LightGBM

#Variables: 50
AUC: 0.76
Tiempo CPU: 3.80s



Comparativa Mejores Algoritmos

AUC>=76%

Revisión de los 2 mejores algoritmos



Comparativa Ctb y Lgb como score PI

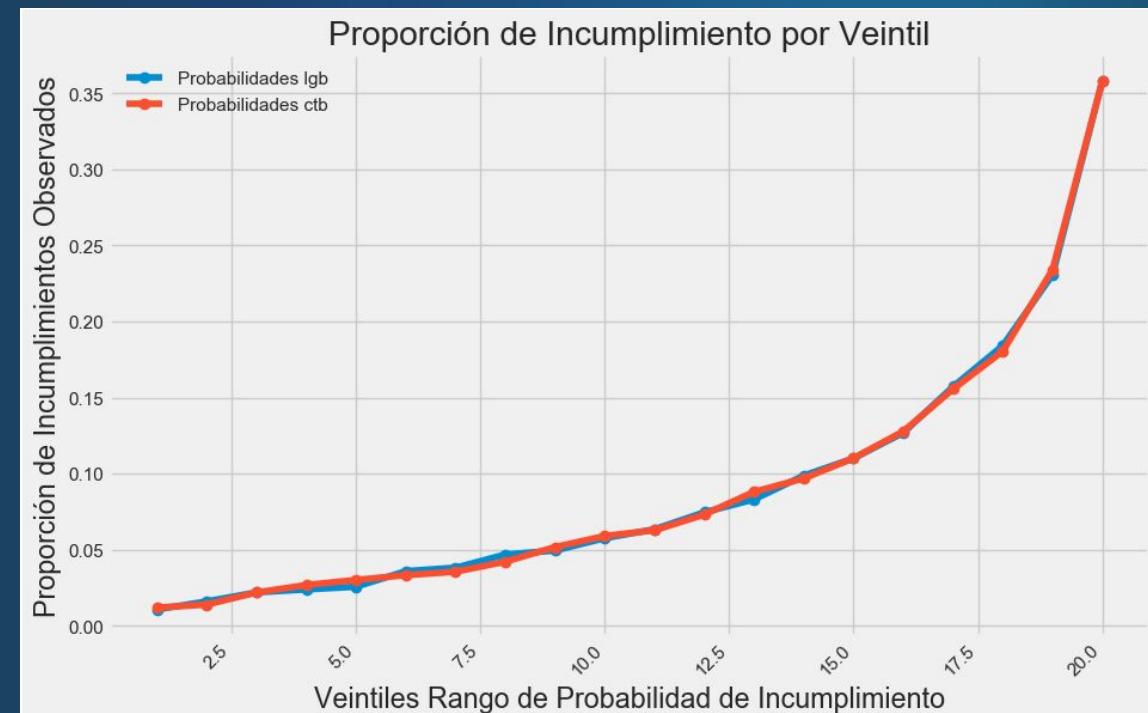
Se entrenamos los modelos CatBoost(lgb) y LightGBM(lgb) con sintonización de parámetros Halving, validación cruzada y 50 variables.

3.53%

Badrate CatBoostClassifier veintil 10

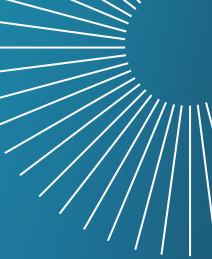
3.54%

Badrate LGBMClassifier veintil 10



Aprobando los primeros 11 veintiles (del 0 al 10) tendríamos una tasa de mora esperada menor en el CatBoost(ctb) que en el LightGMB(lgb). Se percibe como una menor pérdida esperada para el negocio y por ende mayor rentabilidad. Aceptando los primeros 11 veintiles ambos modelos aceptarían al 55% de la población pero sería más rentable el CatBoot que el LightGBM dada su menor tasa de mora en media y desviación estandar.

Ensamblaje vs Catboost

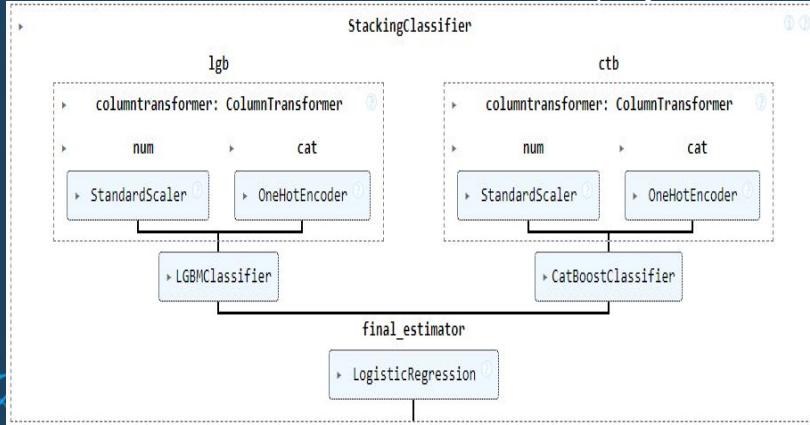


Con el ensamblaje se mejora el AUC de entrenamiento pero desmejora el AUC en base de test y el f1-score disminuye capacidad de identificar clase minoritaria (1).



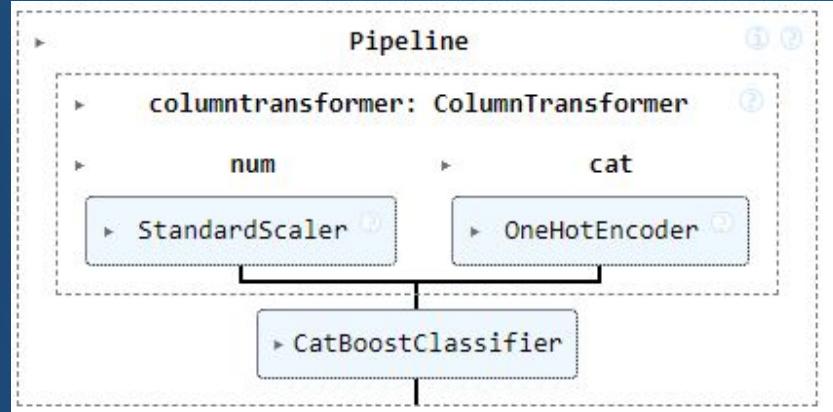
StackingClassifier

- AUC train: 0.812
- AUC test: 0.746
- F1-score 1: 0.0
- Estructura más compleja



CatBoostClassifier

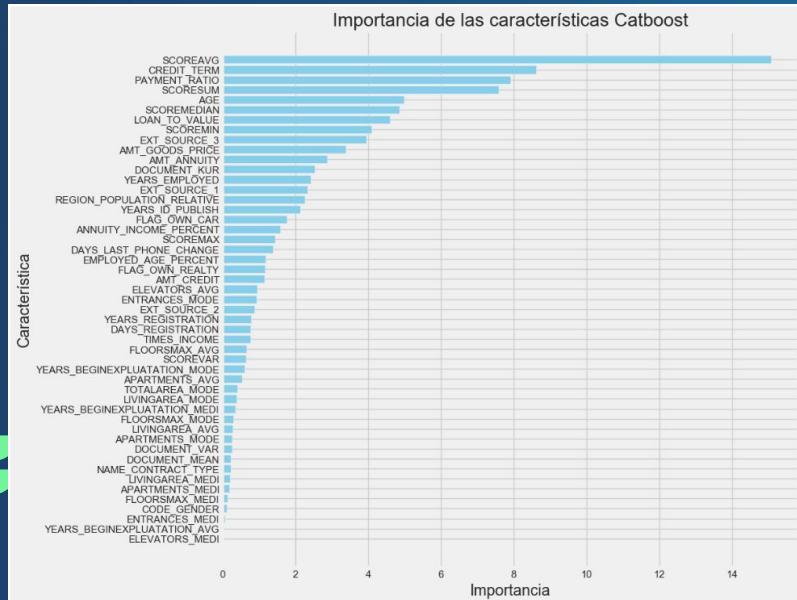
- AUC train: 0.7581
- AUC test: 0.7504
- F1-score 1: 0.29
- Estructura más sencilla



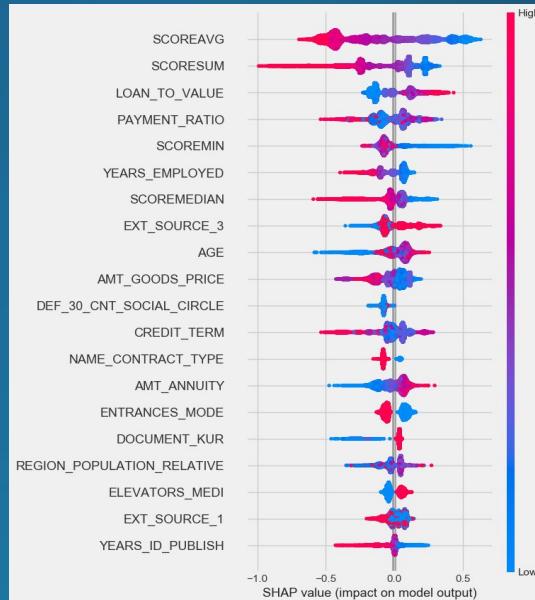
Variables importantes Catboost

El análisis de importancia de variables ayuda a identificar qué características son más relevantes para el modelo, mientras que el análisis de impacto de variables proporciona información sobre cómo cada característica contribuye a las predicciones del modelo. Ambas técnicas coinciden en que la variable de mayor importancia o impacto es la denominada "SCOREAVG".

Importancia de Variables



Impacto de variables SHAP





¡Gracias!

¿Preguntas?

javierrodriguezcifuentes316@gmail.com



<https://www.linkedin.com/in/javier-rodriguezcifuentes316>

