



**DETECCION DE
PRESTAMOS FALLIDOS
ESTADOS UNIDOS (EE UU)**

INTRODUCCION

✓ EN EL PROYECTO ANTERIOR, A TRAVES DE BASES DE DATOS DEL REGISTRO DE LA PROPIEDAD, ANALIZABAMOS:

- EL CICLO INMOBILIARIO RESIDENCIAL Y SU RELACION CON LAS VARIABLES SOCIO-ECONOMICAS DEL PAIS
- CATEGORIZABAMOS LA DIFICULTAD DE ACCESO A VIVIENDA ACTUAL EN ESPAÑA, RELACIONANDO E INDIVIDUALIZANDO POR COMUNIDADES AUTONOMAS:
 - LAS VARIABLES DE OFERTA Y DEMANDA(PRECIO, COMPROVENTAS...)
 - CON LAS VARIABLES SOCIO-ECONOMICAS DE LAS MISMAS (SALARIO MEDIO, PLAZOS Y TIPOS DE INTERES MEDIOS ETC.)

Y LO HACIAMOS A TRAVES DE LA FORMULA DEL PRESTAMO HIPOTECARIO, CON EL QUE CALCULABAMOS Y EVIDENCIABAMOS LA DIFERENCIA ACTUAL EXISTENTE, ENTRE LOS INGRESOS TEORICOS REQUERIDOS POR UNA ENTIDAD BANCARIA PARA LA CONCESION DE UN PRESTAMO, Y LOS INGRESOS MEDIOS REALES POR COMUNIDADES.

✓ EN EL PROYECTO ACTUAL EN CAMBIO, NOS PONEMOS AL OTRO LADO DE LA MESA, COMO ENTIDAD PRESTAMISTA.

COMENZAREMOS ANALIZANDO EL NEGOCIO FINANCIERO Y LOS RIESGOS QUE COMPORTA, MEDIANTE:

- EL ANALISIS EXPLORATORIO DE DATOS (EDA) Y EL APRENDIZAJE AUTOMATICO,

CREAREMOS UNOS MODELOS DE MACHINE LEARNING (EVALUANDO Y COMPARANDO LOS MISMOS, **HASTA DETERMINAR** EL DE MEJORES RESULTADOS EN NUESTRA PROBLEMÁTICA,

Y CATEGORIZAREMOS EL RIESGO DE IMPAGO QUE POSEEN DICHOS CLIENTES, PARA QUE LA EMPRESA DETERMINE:

- CÓMO SE UTILIZA LA INFORMACIÓN PARA MINIMIZAR EL RIESGO DE PÉRDIDA DE DINERO EN LA CONCESIÓN DE PRESTAMOS A CLIENTES.

CONCESION DE PRESTAMOS



COMPRENSION DEL NEGOCIO

CONCESION DE PRESTAMOS

FUNCIONAMIENTO:

ANTE UNA SOLICITUD DE PRÉSTAMO, LA EMPRESA DEBE TOMAR UNA DECISIÓN SOBRE LA APROBACIÓN DEL PRÉSTAMO BASADA EN EL PERFIL DEL SOLICITANTE.

HAY DOS TIPOS DE RIESGOS ASOCIADOS CON LA DECISIÓN DEL BANCO:

- SI EL SOLICITANTE ES PROBABLE QUE DEVUELVA EL PRÉSTAMO, NO APROBARLO PUEDE RESULTAR EN UNA **PÉRDIDA DE NEGOCIO** PARA LA EMPRESA.
- SI EL SOLICITANTE ES PROBABLE QUE INCUMPLA, APROBARLO PUEDE RESULTAR EN UNA **PÉRDIDA FINANCIERA** PARA LA EMPRESA.

LOS DATOS PROPORCIONADOS CONTIENEN INFORMACIÓN SOBRE SOLICITANTES DE PRÉSTAMOS ANTERIORES Y SI HAN "INCUMPLIDO" O NO.

CUANDO SE SOLICITA UN PRÉSTAMO, EXISTEN DOS TIPOS DE DECISIONES QUE LA EMPRESA PODRÍA TOMAR:

PRÉSTAMO ACEPTADO: SI LA EMPRESA APRUEBA EL PRÉSTAMO, HAY 3 ESCENARIOS POSIBLES:

- PAGADO: EL SOLICITANTE HA PAGADO COMPLETAMENTE EL PRÉSTAMO (EL PRINCIPAL Y LA TASA DE INTERÉS)
- EN CURSO: EL SOLICITANTE ESTÁ EN PROCESO DE PAGAR LAS CUOTAS, ES DECIR, EL PLAZO DEL PRÉSTAMO AÚN NO HA FINALIZADO. ESTOS PRESTAMOS NO SE ETIQUETAN COMO "IMPAGADOS".
- A PÉRDIDA/IMPAGADO: EL SOLICITANTE HA INCUMPLIDO SUS OBLIGACIONES DEL PRÉSTAMO, Y TRAS LA INTERVENCIÓN DEL DEPARTAMENTO DE RECUPERACIONES, CON MAYOR O MENOR ÉXITO, SE CIERRA EXPEDIENTE, COMPUTANDO LA PARTE NO PAGADA COMO PERDIDA.

PRÉSTAMO RECHAZADO

LA EMPRESA HA RECHAZADO EL PRÉSTAMO, PORQUE EL CANDIDATO NO CUMPLE CON SUS REQUISITOS.

DADO QUE EL PRÉSTAMO FUE RECHAZADO, NO HAY HISTORIAL DE TRANSACCIONES DE ESOS SOLICITANTES CON LA ENTIDAD Y, POR LO TANTO, ESTOS DATOS NO ESTÁN DISPONIBLES EN LA EMPRESA, NI CONSECUENTEMENTE, EN SU CONJUNTO DE DATOS.

SI OTORGAR PRÉSTAMOS A SOLICITANTES "CON RIESGO" ES LA FUENTE MÁS GRANDE DE PÉRDIDA FINANCIERA, PRODUCIENDO "PÉRDIDA DE CRÉDITO."

OBJETIVO

IDENTIFICAR A ESTOS SOLICITANTES DE PRÉSTAMOS CON RIESGO ELEVADO, PARA ACTUAR EN CONSECUENCIA.



APLICACIÓN PRACTICA

CONCESSION DE PRESTAMOS

"QUÉ SIGNIFICA TODO ESTO PARA UN DATA-SCIENTIST?"

- ✓ **CASO DE CLASIFICACIÓN BINARIA**, EN EL QUE EL OBJETIVO ES QUE NUESTRO MODELO DETERMINE SI EXISTE RIESGO ANTICIPADO DE IMPAGO O NO.

LOS DATOS PROPORCIONADOS CONTIENEN INFORMACIÓN SOBRE SOLICITANTES DE PRÉSTAMOS ANTERIORES Y SI HAN "INCUMPLIDO" O NO, SIENDO EN NUESTRO CASO, EL VALOR 1, LA RESPUESTA POSITIVA DE IMPAGO= ("CHARGED OFF")

- ✓ **EL MODELO SELECCIONADO HA DE TENER UN BUEN DESEMPEÑO PARA EFECTUAR ESA CLASIFICACIÓN.**

Dicho desempeño se mide a través de la curva ROC (RECEIVER OPERATING CHARACTERISTIC) que es una herramienta gráfica que representa el rendimiento de un modelo de clasificación en todos los umbrales de decisión.

- ✓ **¿CÓMO CLASIFICARÍAMOS LOS DOS TIPOS DE RIESGOS ASOCIADOS CON LA DECISIÓN DEL BANCO?**

- SI EL SOLICITANTE ES PROBABLE QUE DEVUELVA EL PRÉSTAMO, NO APROBARLO PUEDE RESULTAR EN UNA **PÉRDIDA DE NEGOCIO** PARA LA EMPRESA.
 - FALSOS POSITIVOS (PREDECIMOS QUE NO VA A PAGAR Y ERRAMOS).
- SI EL SOLICITANTE ES PROBABLE QUE INCUMPLA, APROBARLO PUEDE RESULTAR EN UNA **PÉRDIDA FINANCIERA** PARA LA EMPRESA.
 - FALSOS NEGATIVOS (PREDECIMOS QUE EL CLIENTE VA A PAGAR, Y ERRAMOS)

SI BIEN AMBOS ERRORES, TIENEN CONSECUENCIAS, QUE LA ENTIDAD HABRÁ DE VALORAR, LO DETERMINANTE EN EL CASO QUE NOS OCUPA ES MINIMIZAR LOS FALSOS NEGATIVOS.

¿MOTIVO?

- PORQUE LOS FN, SON LOS QUE GENERAN UN QUEBRANTO ECONÓMICO DIRECTO EN LA CUENTA DE EXPLOTACIÓN DE LA EMPRESA.
- EL ERROR EN LOS FP, NO SÓLO ES MENOS GRAVOSO PARA LA ENTIDAD, SINO QUE ADEMÁS, SOBRE EL MISMO, SE PUEDE ACTUAR; NO SÓLO NEGANDO EL PRÉSTAMO, SINO REDUCIENDO LA CANTIDAD DEL PRÉSTAMO, PRESTANDO A UNA TASA DE INTERÉS MÁS ALTA, EXIGIENDO MAYORES GARANTÍAS ETC.

- ✓ EN CONSECUENCIA, LA MÉTRICA QUE RESULTA MÁS RELEVANTE EN NUESTRO CASO, ES EL RECALL, ÚTIL CUANDO EL COSTO DE LOS FALSOS NEGATIVOS ES ALTO Y QUEREMOS MINIMIZAR LA CANTIDAD DE INSTANCIAS POSITIVAS QUE SE CLASIFICAN INCORRECTAMENTE COMO NEGATIVAS.



OBJETIVO

PRIORIZAR RESULTADOS:

- ✓ CURVA ROC Y RECALL EN 1 ✓

PARA REDUCIR FALSOS NEGATIVOS

ESTRUCTURA PRINCIPAL DEL PROYECTO

INDICE DE CONTENIDO

 **DETECCION PRESTAMOS FALLIDOS**
1 cell hidden ...

 **Descripción de la DATA**
1 cell hidden ...

 **Carga del Dataset y Análisis inicial de los Datos**
87 cells hidden ...

 **Pre/procesamiento de datos**
82 cells hidden ...

 **✓ Salvado de data tras el análisis exploratorio de datos (EDA)**
3 cells hidden ...

 **El tarjet/objetivo y el estudio del Balanceo**
9 cells hidden ...

 **División de entrenamiento y prueba**
2 cells hidden ...

 **Normalizando la data - Estudio de Estandarización**
31 cells hidden ...

 **✓ Data para puebas de ML**
14 cells hidden ...

 **Construyendo el Modelo**
9 cells hidden ...

 **Machine Learning - Clasificación**
25 cells hidden ...

 **✓ Regresión Logística - Logistic Regression**
7 cells hidden ...

 **✓ Random Forest Classifier**
6 cells hidden ...

 **✓ XGBoost Classifier**
6 cells hidden ...

 **✓ KNN - KNearest Neighbors Classifier**
4 cells hidden ...

 **✓ ANN - Artificial Neural Networks / Redes Neuronales**

 **Comparando el Rendimiento de los Modelos de ML**
9 cells hidden ...

 **Conclusiones:**
2 cells hidden ...

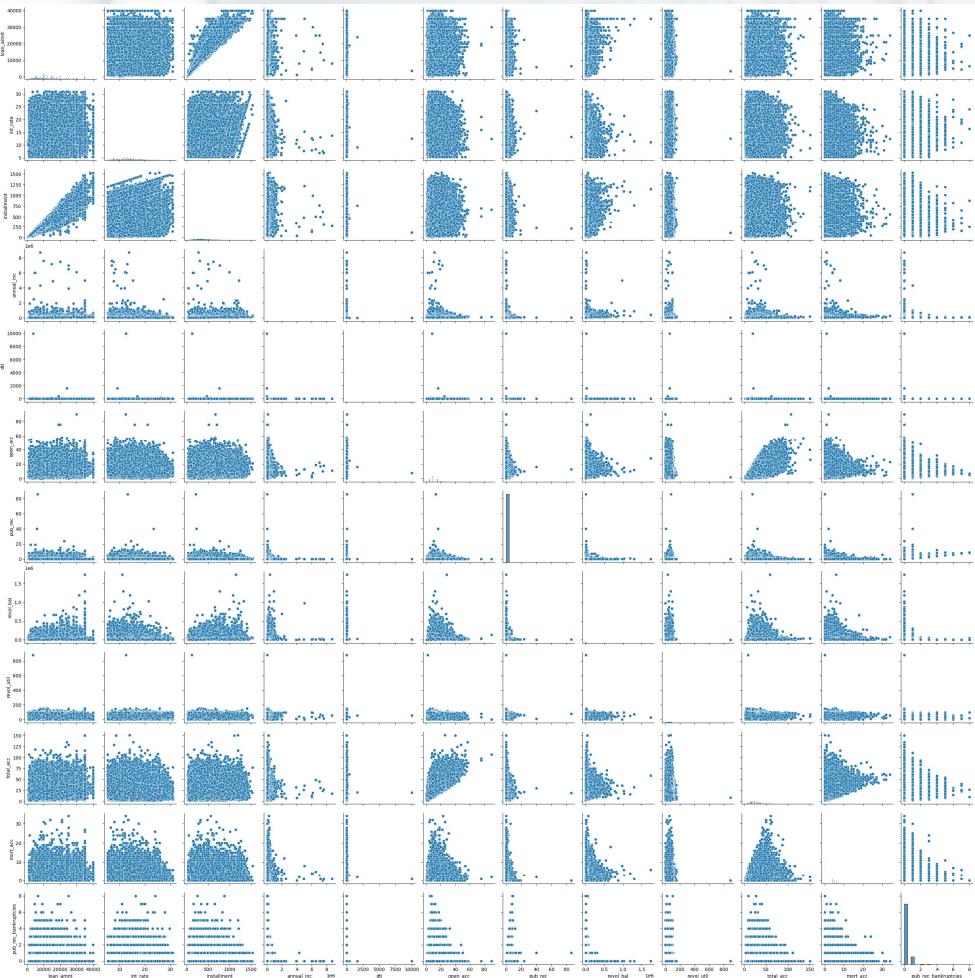
 **Salvado del mejor Modelo ML**
1 cell hidden ...

 **Prueba de FUEGO**
8 cells hidden ...

ESTUDIOS PREVIOS DE VARIABLES

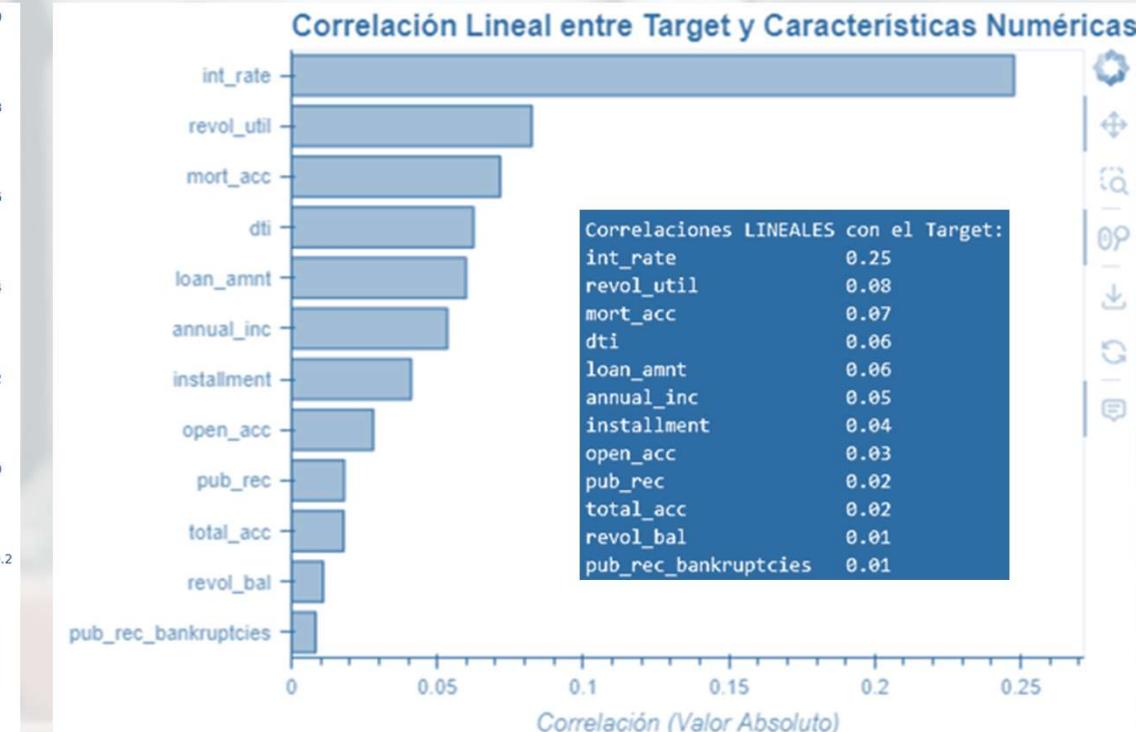
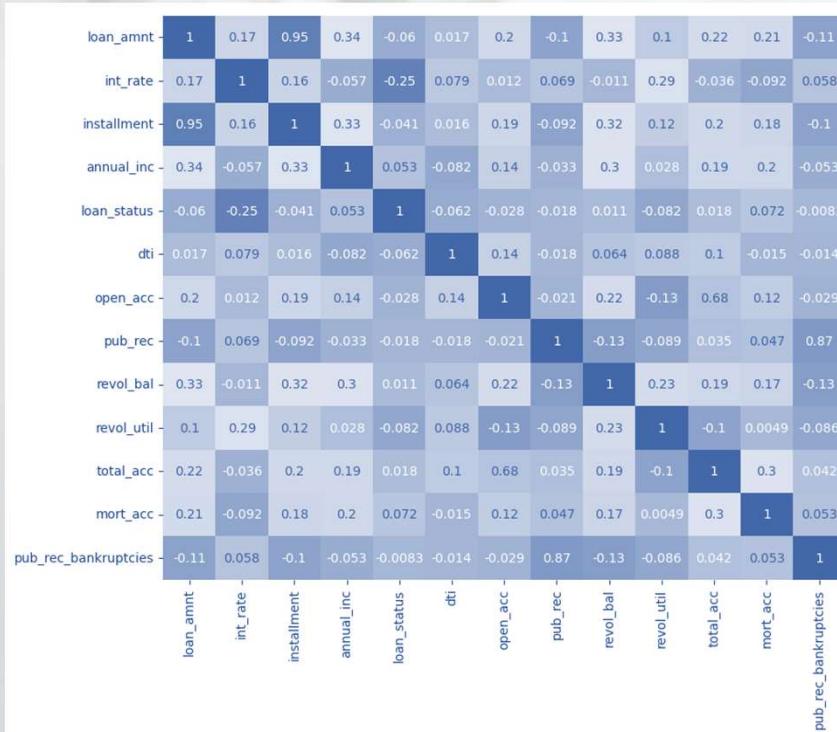
loan_amnt	La cantidad solicitada por el prestatario. Si en algún momento, se reduce la cantidad del préstamo, se reflejará en este valor.
term	Plazo del préstamo. Los valores están en meses y pueden ser 36 o 60.
int_rate	Tasa/Tipo de interés.
installment	Cuota mensual.
grade	Grado del préstamo asignado por LC.
sub_grade	Subgrado del préstamo asignado por LC.
emp_title	El título del trabajo proporcionado por el prestatario al solicitar el préstamo.
emp_length	Antigüedad laboral en años. Los valores posibles están entre 0 y 10, donde 0 significa menos de un año y 10 significa diez o más años.
home_ownership	El estado de propiedad de la vivienda proporcionado por el prestatario durante el registro o obtenido del informe de crédito. Nuestros valores son: Rent(alquiler), Own(propiedad), Mortgage(Hipoteca), Other(otro).
annual_inc	Los ingresos anuales informados por el prestatario durante el registro.
verification_status	Indica si los ingresos fueron verificados por LC, no verificados o si la fuente de ingresos fue verificada.
issue_d	El mes en que se financió el préstamo.
loan_status	Estado actual del préstamo.
purpose	Finalidad del préstamo.
title	El título del préstamo proporcionado por el prestatario.
zip_code	Los primeros 3 números del código postal proporcionado por el prestatario en la solicitud de préstamo.
addr_state	Estado de EEUU proporcionado por el prestatario en la solicitud de préstamo.
dti	Pagos mesuales por deudas/Ingresos Mesuales (%). Excluidos pagos de deuda por hipoteca.
earliest_cr_line	El mes en que se abrió la primera línea de crédito informada del prestatario.
open_acc	El número de líneas de crédito abiertas.
pub_rec	Número de registros públicos perjudiciales, por retrasos publicados.
revol_bal	Saldo total de sus créditos revolving/rotativos.
revol_util	Porcentaje de utilización de créditos revolving/rotativos.
total_acc	El número total de líneas de crédito actualmente en el archivo de crédito del prestatario.
initial_list_status	El estado de listado inicial del préstamo. Los valores posibles son: W, F.
application_type	Indica si el préstamo es una solicitud individual o una solicitud conjunta con dos co-prestatarios.
mort_acc	Número de cuentas hipotecarias.
pub_rec_bankruptcies	Número de veces que el prestatario ha solicitado la quiebra.

EXPLORATORY DATA ANALYSIS:



VARIABLES: CORRELACIONES LINEALES

EXPLORATORY DATA ANALYSIS:



EN LOS DATOS, OBSERVAMOS QUE **LA INFORMACIÓN SE AGRUPA** EN TORNO A DOS TIPOS DE CARACTERÍSTICAS:

- LAS RELACIONADAS CON EL SOLICITANTE (VARIABLES DEMOGRÁFICAS COMO OCUPACIÓN, DETALLES DE EMPLEO, ETC.).
- LAS RELACIONADAS CON LA INFORMACIÓN DEL PRÉSTAMO (IMPORTE, PLAZO, TASA DE INTERÉS, PROPÓSITO DEL PRÉSTAMO, ETC.).

DEL ANÁLISIS DE DATOS EFECTUADO, COMPARANDO LAS CORRELACIONES LINEALES ENTRE LOS DATOS Y NUESTRA VARIABLE OBJETIVO, DEDUCIMOS NO OBSTANTE QUE HAY **INDICIOS CLAROS DE MULTICOLINEALIDAD**, (ALGUNAS VARIABLES SON EL RESULTADO DE LA FORMULACIÓN NUMÉRICA DE LOS PROPIOS PRESTAMOS), ASÍ COMO QUE **LA RELACIÓN ENTRE VARIABLES ES ESCASA**.

PROCEDEMOS POR TANTO A REALIZAR UN ANÁLISIS MAS EXHAUSTIVO.

MACHINE LEARNING DETECCION DE PRESTAMOS FALLIDOS - Mercado Estados Unidos

VARIABLES: ANALISIS PORMENORIZADO Y DEPENDENCIAS



EXPLORATORY DATA ANALYSIS:

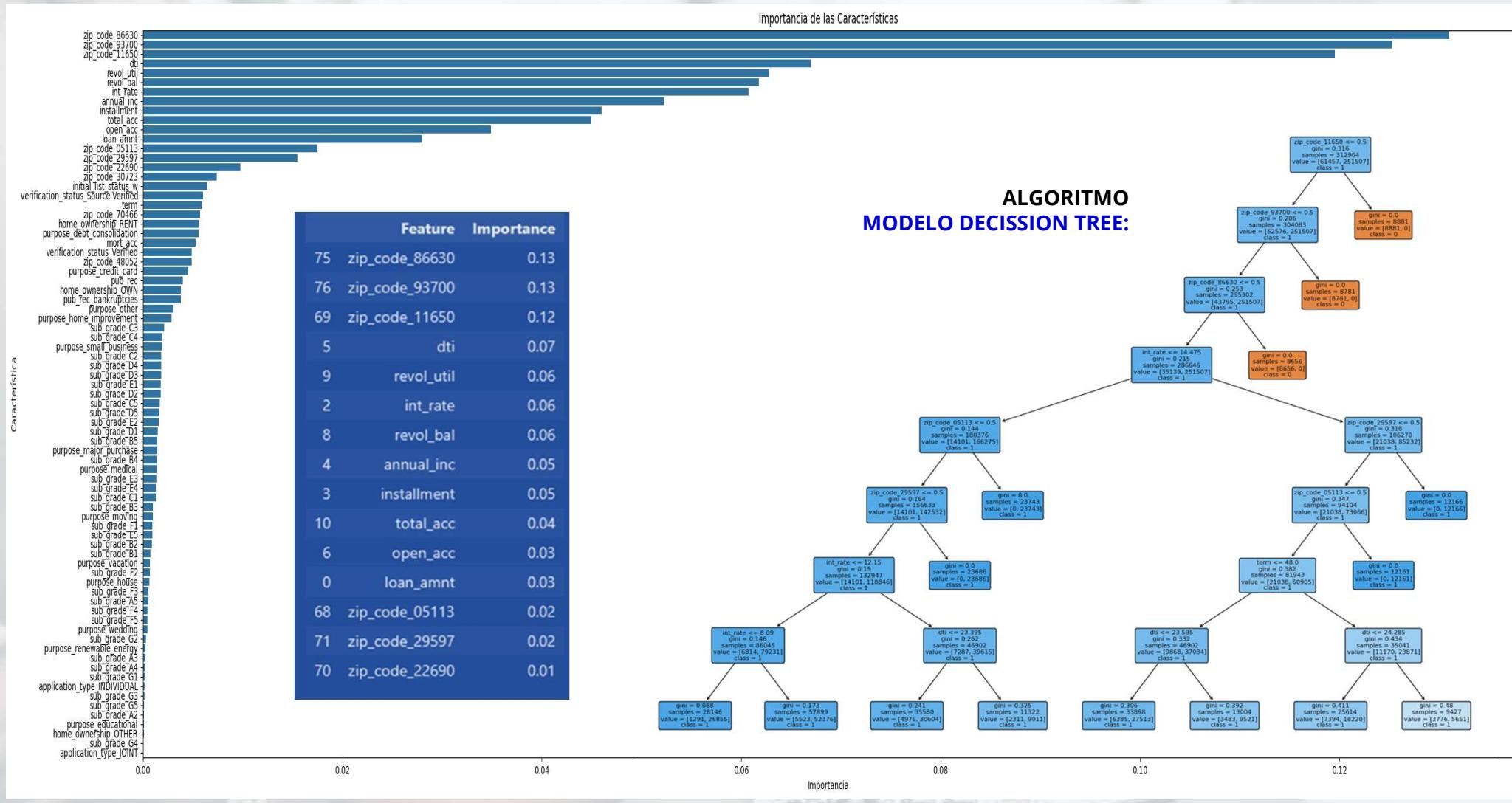


MACHINE LEARNING

DETECCION DE PRESTAMOS FALLIDOS - Mercado Estados Unidos

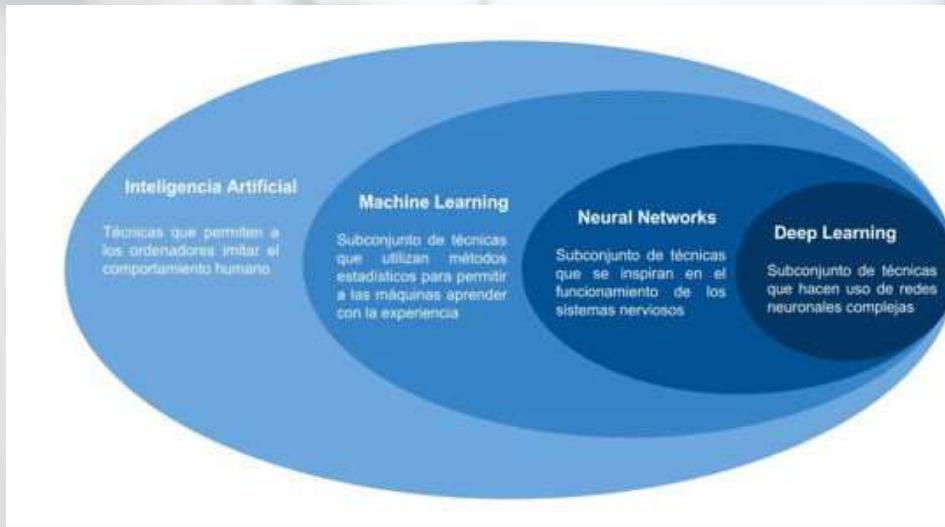
VARIABLES: CORRELACIONES NO LINEALES

EXPLORATORY DATA ANALYSIS:



MODELOS & ALGORITMOS EMPLEADOS

MACHINE LEARNING & REDES NEURONALES:



✓ **REGRESIÓN LOGÍSTICA - LOGISTIC REGRESSION**

LA REGRESIÓN LOGÍSTICA ES UN MODELO ESTADÍSTICO QUE SE UTILIZA EN PROBLEMAS DE CLASIFICACIÓN PARA PREDICIR LA PROBABILIDAD DE QUE UNA OBSERVACIÓN PERTENEZCA A UNA CATEGORÍA ESPECÍFICA.

✓ **RANDOM FOREST CLASSIFIER**

UTILIZA CONJUNTOS DE ÁRBOLES DE DECISIÓN ENTRENADOS DE MANERA INDEPENDIENTE, MANEJANDO ERRORES MEDIANTE EL PROMEDIO DE PREDICCIONES. OFRECE REGULARIZACIÓN MEDIANTE MUESTREO ALEATORIO Y ES EFICAZ PARA CONJUNTOS DE DATOS PEQUEÑOS A MEDIANOS.

✓ **XG BOOST CLASSIFIER**

EMPLEA ÁRBOLES DE DECISIÓN DÉBILES SECUENCIALES, CORRIGIENDO ERRORES DE FORMA ITERATIVA. OFRECE REGULARIZACIÓN AVANZADA, ES EFICIENTE EN CONJUNTOS DE DATOS GRANDES Y PUEDE MANEJAR DATOS FALTANTES DE MANERA INHERENTE, DESTACANDO POR SU RENDIMIENTO Y CAPACIDAD DE AJUSTE FINO.

✓ **KNN** - KNEAREST NEIGHBORS

EL ALGORITMO KNN SE BASA EN LA IDEA DE QUE LOS PUNTOS DE DATOS CON CARACTERÍSTICAS SIMILARES TIENDEN A AGRUPARSE JUNTOS EN EL ESPACIO

✓ **ANN** - ARTIFICIAL NEURAL NETWORKS / REDES NEURONALES

MODELOS DE APRENDIZAJE PROFUNDO INSPIRADOS EN LA ESTRUCTURA Y FUNCIONAMIENTO DEL CEREBRO. COMPUESTAS POR CAPAS DE NODOS (NEURONAS) INTERCONECTADAS, CON CAPAS DE ENTRADA, CAPAS OCULTAS Y CAPAS DE SALIDA. UTILIZAN ALGORITMOS DE RETROPROPAGACIÓN PARA APRENDER PATRONES COMPLEJOS Y REALIZAR TAREAS DE CLASIFICACIÓN, REGRESIÓN O RECONOCIMIENTO DE PATRONES EN DATOS. SON FUNDAMENTALES EN EL CAMPO DE LA INTELIGENCIA ARTIFICIAL Y EL APRENDIZAJE PROFUNDO.

MACHINE LEARNING DETECCION DE PRESTAMOS FALLIDOS - Mercado Estados Unidos

Machine Learning

MODELOS & ALGORITMOS EMPLEADOS

EVALUACION DE RESULTADOS POR MODELO: REGRESION LOGISTICA

Train Result:

```
Accuracy Score: 88.91%
```

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	0.95	0.88	0.89	0.92	0.90
recall	0.46	0.99	0.89	0.73	0.89
f1-score	0.62	0.94	0.89	0.78	0.87
support	61457.00	251507.00	0.89	312964.00	312964.00

Confusion Matrix:

```
[[ 28232 33225]
 [ 1491 250016]]
```

Test Result:

```
Accuracy Score: 88.86%
```

CLASSIFICATION REPORT:

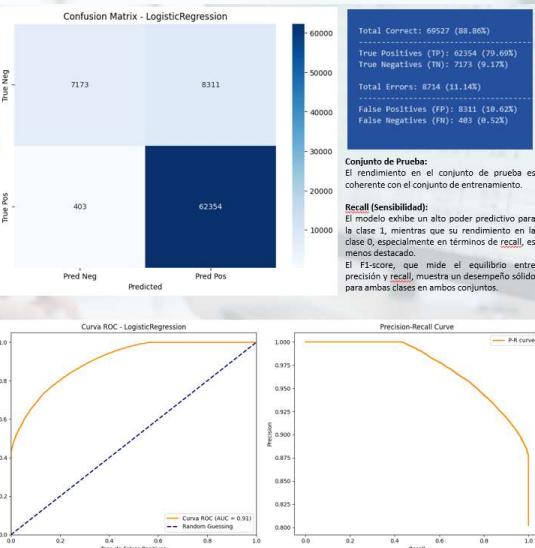
	0	1	accuracy	macro avg	weighted avg
precision	0.95	0.88	0.89	0.91	0.90
recall	0.46	0.99	0.89	0.73	0.89
f1-score	0.62	0.94	0.89	0.78	0.87
support	15484.00	62757.00	0.89	78241.00	78241.00

Confusion Matrix:

```
[[ 7173 8311]
 [ 403 62354]]
```

El rendimiento en el conjunto de prueba es coherente con el conjunto de entrenamiento.

Consideraciones sobre la Curva ROC:
El resultado de la Curva ROC de 0.91 indica que el modelo tiene una buena capacidad para discriminar entre las clases positiva y negativa, lo cual es positivo en la mayoría de los casos.



EVALUACION DE RESULTADOS POR MODELO: XGBoost Classifier

Train Result:

```
Accuracy Score: 88.77%
```

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	1.00	0.88	0.89	0.94	0.98
recall	0.43	1.00	0.89	0.71	0.89
f1-score	0.60	0.93	0.89	0.77	0.87
support	61457.00	251507.00	0.89	312964.00	312964.00

Confusion Matrix:

```
[[ 26318 35139]
 [ 0 251507]]
```

Test Result:

```
Accuracy Score: 88.78%
```

CLASSIFICATION REPORT:

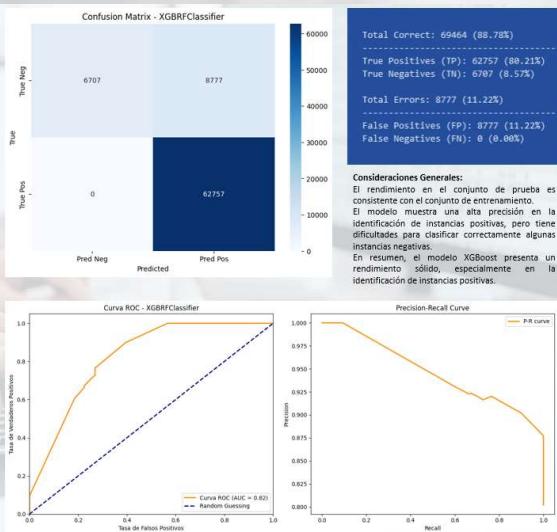
	0	1	accuracy	macro avg	weighted avg
precision	1.00	0.88	0.89	0.94	0.98
recall	0.43	1.00	0.89	0.72	0.89
f1-score	0.60	0.93	0.89	0.77	0.87
support	15484.00	62757.00	0.89	78241.00	78241.00

Confusion Matrix:

```
[[ 6707 8777]
 [ 0 62757]]
```

El rendimiento en el conjunto de prueba es consistente con el conjunto de entrenamiento.

Curva AUC-ROC:
El resultado de la Curva AUC-ROC de 0.82 indica una buena capacidad de discriminación entre las clases, aunque no tan alto como en algunos modelos anteriores.



EVALUACION DE RESULTADOS POR MODELO: RANDOM FOREST

Train Result:

```
Accuracy Score: 91.66%
```

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	1.00	0.91	0.92	0.95	0.92
recall	0.58	1.00	0.92	0.79	0.92
f1-score	0.73	0.95	0.92	0.84	0.91
support	61457.00	251507.00	0.92	312964.00	312964.00

Confusion Matrix:

```
[[ 3558 26099]
 [ 0 251507]]
```

Test Result:

```
Accuracy Score: 88.83%
```

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	0.98	0.88	0.89	0.93	0.90
recall	0.44	1.00	0.89	0.72	0.89
f1-score	0.61	0.93	0.89	0.77	0.87
support	15484.00	62757.00	0.89	78241.00	78241.00

Confusion Matrix:

```
[[ 6871 8613]
 [ 125 62632]]
```

Indicador de ligero sobreajuste debido a la diferencia de rendimiento entre el conjunto de entrenamiento y prueba.

Curva ROC:
Resultado de la Curva AUC-ROC de 0.88 sugiere buena capacidad de discriminación entre clases. Indica tasas de verdaderos positivos altas y falsos positivos relativamente bajos.

EVALUACION DE RESULTADOS POR MODELO: KNN - Knearest Neighbors Classifier

Train Result:

```
Accuracy Score: 89.45%
```

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	0.90	0.89	0.89	0.90	0.89
recall	0.52	0.99	0.89	0.75	0.89
f1-score	0.66	0.94	0.89	0.88	0.88
support	61457.00	251507.00	0.89	312964.00	312964.00

Confusion Matrix:

```
[[ 31982 29475]
 [ 3552 247955]]
```

Test Result:

```
Accuracy Score: 87.97%
```

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	0.84	0.88	0.88	0.86	0.88
recall	0.48	0.98	0.88	0.73	0.88
f1-score	0.61	0.93	0.88	0.77	0.87
support	15484.00	62757.00	0.88	78241.00	78241.00

Confusion Matrix:

```
[[ 7468 8016]
 [ 1395 61362]]
```

El modelo KNN presenta un buen equilibrio entre precisión y recall en ambos conjuntos de datos.

Curva ROC:
La Curva ROC AUC de 0.86 sugiere una buena capacidad de discriminación entre clases. Como siempre, considera ajustar hiperparámetros y realizar una validación cruzada para optimizar el rendimiento del modelo.

MACHINE LEARNING & REDES NEURONALES:

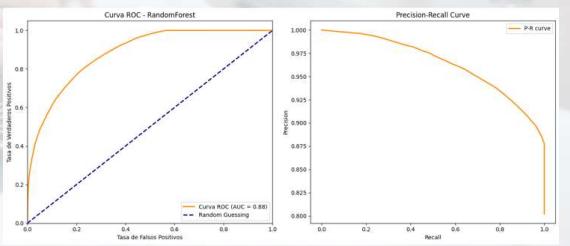
Confusion Matrix - Random Forest

Predicted	True Neg	True Pos
Pred Neg	6871	125
Pred Pos	8613	62632

Total Correct: 69593 (88.83%)
 True Positives (TP): 62632 (89.05%)
 True Negatives (TN): 6871 (8.78%)
 Total Errors: 8738 (11.17%)
 False Positives (FP): 8613 (11.01%)
 False Negatives (FN): 125 (0.16%)

Consideraciones Generales:
En resumen, el modelo Random Forest muestra un notable rendimiento en el conjunto de entrenamiento y prueba, si bien se detecta un ligero descenso en prueba, que podría ser indicativo de un grado menor de overfitting.

Buen resultado de recall en 1 (charged off loans), que denota una buena distinción entre clases.



Confusion Matrix - KNN

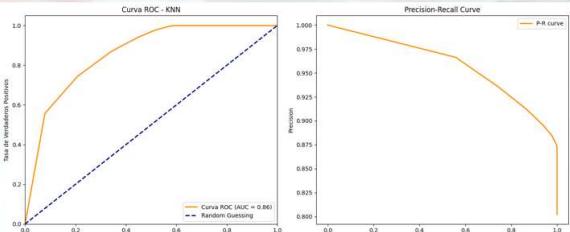
Predicted	True Neg	True Pos
Pred Neg	7468	1395
Pred Pos	8016	61362

Total Correct: 68838 (87.97%)
 True Positives (TP): 61362 (79.43%)
 True Negatives (TN): 7468 (9.54%)
 Total Errors: 8811 (12.03%)
 False Positives (FP): 8016 (10.25%)
 False Negatives (FN): 1395 (1.78%)

Consideraciones Generales:
El modelo KNN presenta un buen equilibrio entre precisión y recall en ambos conjuntos de datos.

La Curva ROC AUC de 0.86 indica una capacidad razonable para distinguir entre las clases.

En resumen, el modelo KNN demuestra un rendimiento generalmente sólido, con un equilibrio razonable entre precisión y recall.



EVALUACION DE RESULTADOS POR MODELO: ANN - Artificial Neural Networks

Accuracy Score: 88.77%

CLASSIFICATION REPORT:

	0.0	1.0	accuracy	macro avg	weighted avg
precision	1.00	0.88	0.89	0.94	0.90
recall	0.43	1.00	0.89	0.71	0.89
f1-score	0.60	0.93	0.89	0.77	0.87
support	61457.00	251507.00	0.89	312964.00	312964.00

Confusion Matrix:

```
[[ 26320 35137]
 [ 4 251503]]
```

2446/2446 [=====] - 3s 1ms/step

Test Result:

Accuracy Score: 88.78%

CLASSIFICATION REPORT:

	0.0	1.0	accuracy	macro avg	weighted avg
precision	1.00	0.88	0.89	0.94	0.90
recall	0.43	1.00	0.89	0.72	0.89
f1-score	0.60	0.93	0.89	0.77	0.87
support	15484.00	62757.00	0.89	78241.00	78241.00

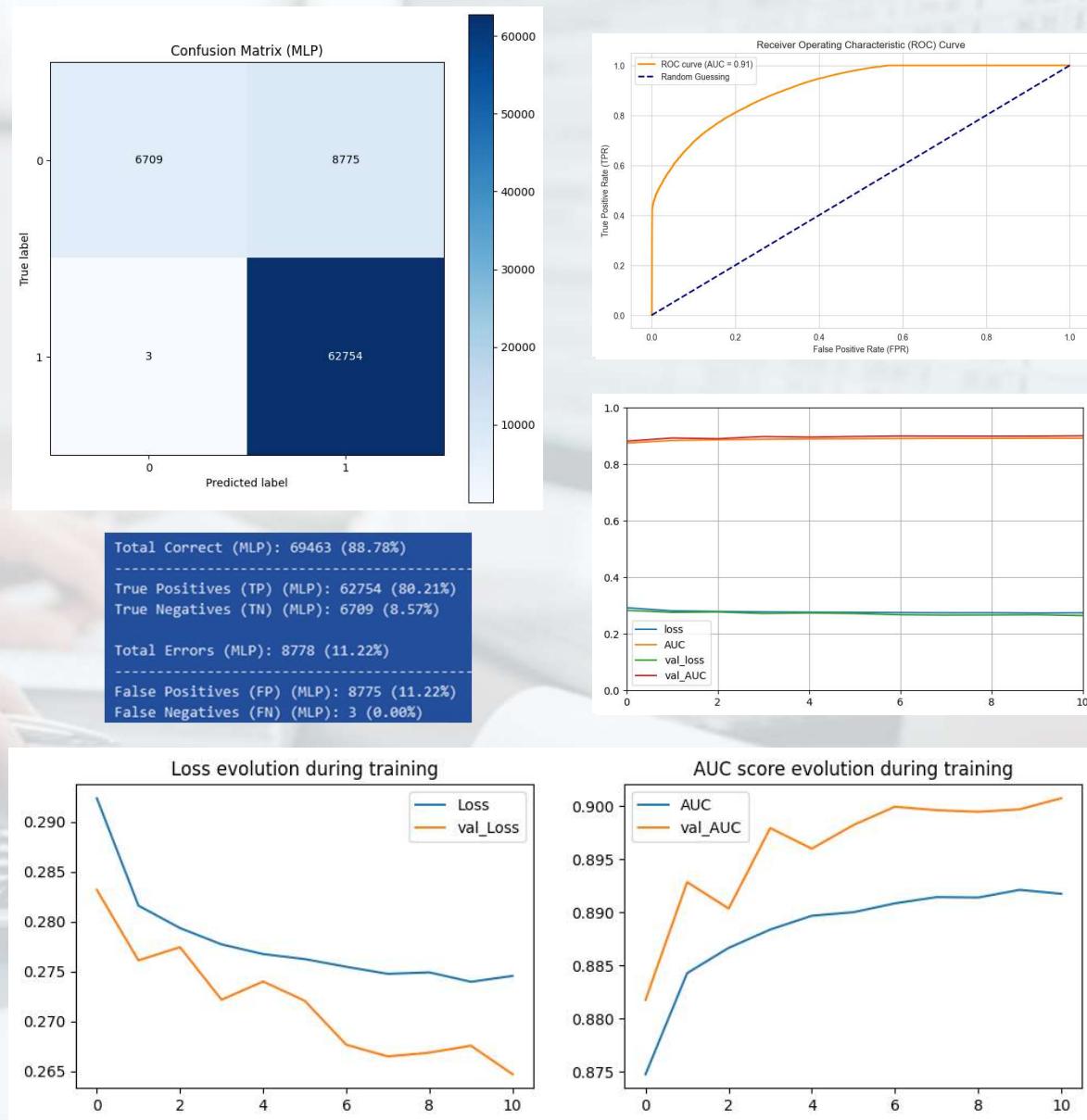
Confusion Matrix:

```
[[ 6709 8775]
 [ 3 62754]]
```

Consideraciones generales:

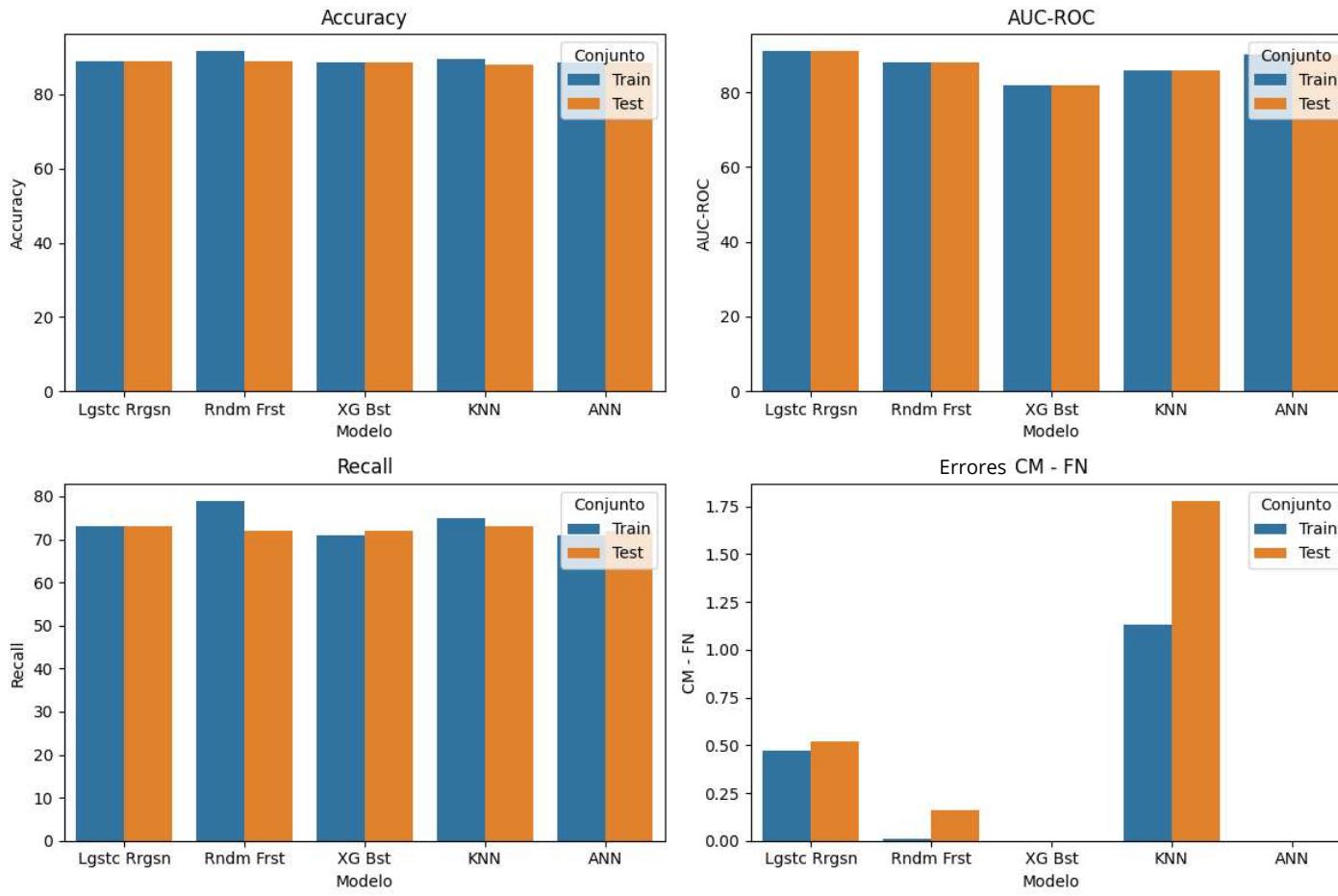
El modelo tiene una alta precisión, lo que sugiere un buen rendimiento en general.

La precisión, el recall y el F1-score para la clase 1 son buenos, lo que indica un rendimiento sólido en la predicción de esta clase.



COMPARANDO RESULTADOS: CONCLUSIONES FINALES

MACHINE LEARNING



Modelo	Conjunto	Accuracy	AUC-ROC	Recall	Precision	CM - FP	CM - FN
0 Lgsc Rrgsn	Train	88.91	91.00	73.00	92.00	10.61	0.47
1 Lgsc Rrgsn	Test	88.86	91.00	73.00	91.00	10.62	0.52
2 Rndm Frst	Train	91.66	88.00	79.00	95.00	8.33	0.01
3 Rndm Frst	Test	88.79	88.00	72.00	93.00	11.01	0.16
4 XG Bst	Train	88.77	82.00	71.00	94.00	11.22	0.00
5 XG Bst	Test	88.78	82.00	72.00	94.00	11.22	0.00
6 KNN	Train	89.45	86.00	75.00	90.00	9.41	1.13
7 KNN	Test	87.97	86.00	73.00	86.00	10.25	1.78
8 ANN	Train	88.77	90.00	71.00	94.00	11.22	0.00
9 ANN	Test	88.78	90.00	72.00	94.00	11.22	0.00

Consideraciones generales:

Todos los modelos muestran un buen desempeño en términos de Accuracy, AUC-ROC, y Recall, lo cual es positivo para un problema de clasificación binaria.

Logistic Regression, XGBoost, y ANN parecen ser los modelos más equilibrados en términos de rendimiento general.

Artificial Neural Network (ANN):

En general, si el AUC-ROC es la métrica principal, se busca un buen rendimiento en datos no vistos, priorizando la no aparición de falsos negativos, el modelo de Redes Neuronales Artificiales (ANNs) parece ser la mejor opción, justificado por:

- AUC-ROC próximo al (90.00%), lo que sugiere una buena capacidad para distinguir entre clases.
- Notable exactitud y Recall sólido que permite minimizar FN

Modelo	Accuracy	AUC-ROC	Recall	Precision	CM - FP	CM - FN
0 ANN	88.78	90.00	71.50	94.00	11.22	0.00
4 XG Bst	88.78	82.00	71.50	94.00	11.22	0.00
3 Rndm Frst	90.22	88.00	75.50	94.00	9.67	0.09
2 Lgsc Rrgsn	88.88	91.00	73.00	91.50	10.61	0.49
1 KNN	88.71	86.00	74.00	88.00	9.83	1.46

"PRUEBA DE FUEGO" - MEJOR MODELO:

ANN - Artificial Neural Networks MACHINE LEARNING

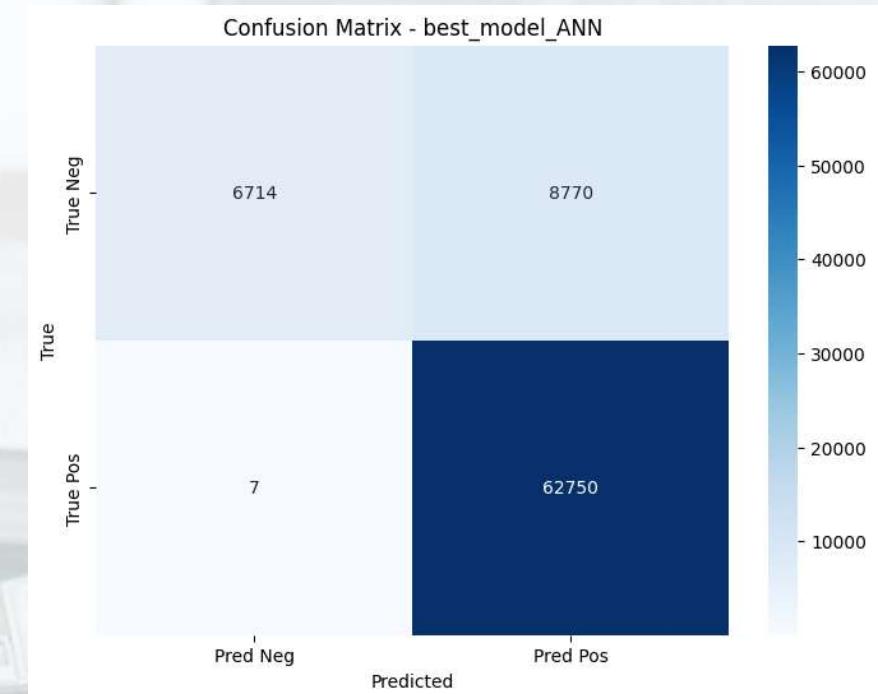
➡ SE DETRAJO UN 20% DE LOS DATA ORIGINAL, PARA TESTEAR EL RENDIMIENTO DE NUESTRO MEJOR MODELO ➡

⌚ OBJETIVO ⌚

OBTENER UN MODELO DE MACHINE LEARNING ROBUSTO, (PRIORIZANDO CURVA ROC & EVITAR FALSOS NEGATIVOS EN 1), QUE EN EL NEGOCIO BANCARIO ANALIZADO, SE TRADUCE EN PRESTAMOS FALLIDOS, QUE ORIGINEN PERDIDA ECONÓMICA A LA ENTIDAD.

PROCEDEMOS A REALIZAR LAS PREDICIONES Y EL RESULTADO DE NUESTRO MODELO ES:

Exactitud: 0.89				
Informe de Clasificación:				
	precision	recall	f1-score	support
0	1.00	0.43	0.60	15484
1	0.88	1.00	0.93	62757
accuracy			0.89	78241
macro avg	0.94	0.72	0.77	78241
weighted avg	0.90	0.89	0.87	78241
Total Correct: 69464 (88.78%)				
True Positives (TP): 62750 (80.20%)				
True Negatives (TN): 6714 (8.58%)				
Total Errors: 8777 (11.22%)				
False Positives (FP): 8770 (11.21%)				
False Negatives (FN): 7 (0.01%)				



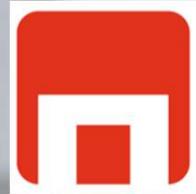
DE LOS 78.241 CASOS QUE HEMOS SUMINISTRADO ALEATORIAMENTE PARA ANALIZAR:

✓ ERROR EN FALSOS NEGATIVOS (FN) DE UN 0.01% - CAPACIDAD DE PREDICCIÓN DE FN DE 99.99% & ✓ CURVA ROC PRÓXIMA AL 90%
HABIENDO SUPERADO INCLUSO, EL HANDICAP INICIAL DE CONTAR CON UNA DATA POCO CORRELACIONADA

Por tanto, podemos afirmar que...

⭐⭐⭐ ¡¡OBJETIVO CUMPLIDO!! ⭐⭐⭐

¡GRACIAS POR SU ATENCIÓN!



DANIEL AZPITARTE SEOANE - az0110se@gmail.com