



xepelin

---

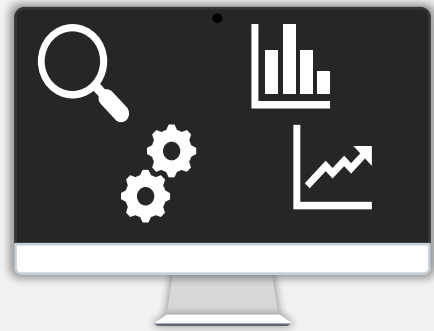
# XEPELIN

## Machine learning Model

Guillermo J. Bergues



# Introducción



**EDA:** Análisis exploratorio de los datos para comprender las características de lo requerido.

**FE:** Ingeniería de los datos previa al entrenamiento de varios modelos. Se trabajaron los datos de 3 maneras diferentes.

**Modelos:** Se realizaron 3 modelos. Primero, un regresor lineal para comprender los datos. Segundo, un Catboost regresor buscando un desempeño mejor y, finalmente, un estudio de varios modelos entrenados con un AutogluOn.



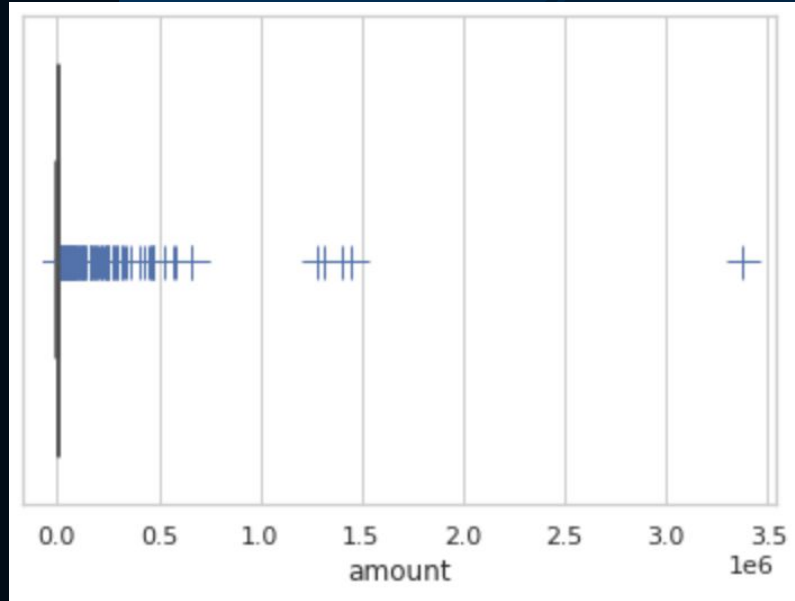
# Análisis exploratorio (EDA)

- ☐ Distribución de los datos.
- ☐ Nulos, duplicados y outliers.
- ☐ Correlaciones.

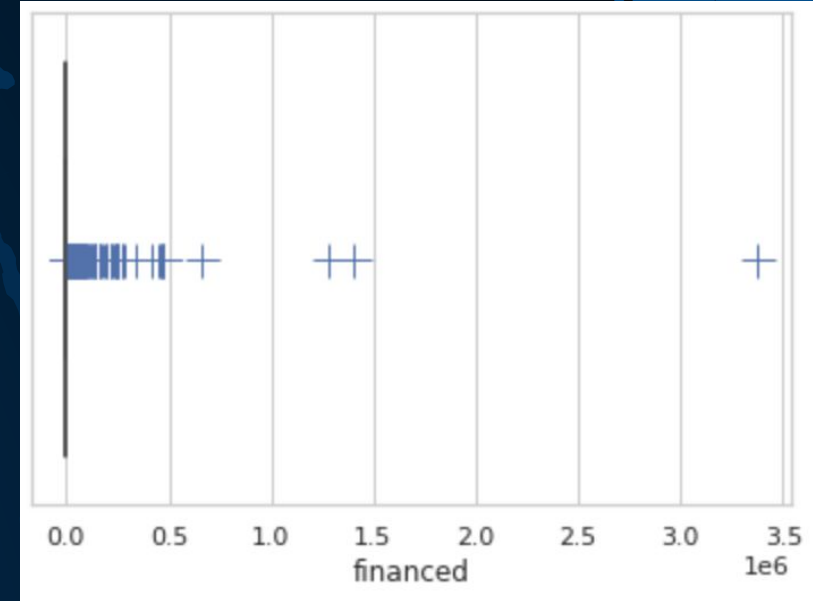


# Distribuciones

Amount



Financiado

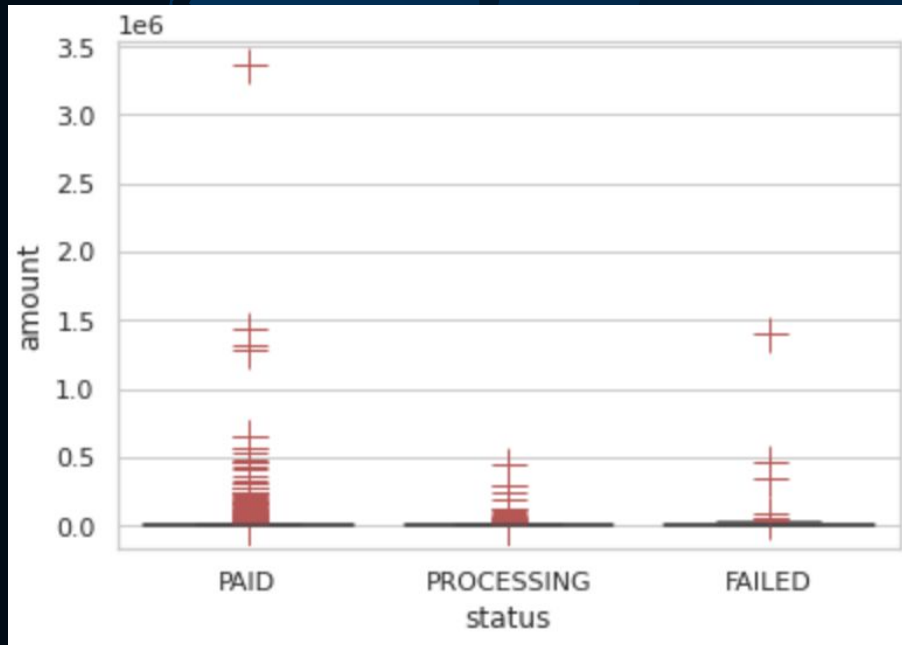


El monto máximo (3370741.92) está completamente financiado por Xepelin.  
Puede ser un error.

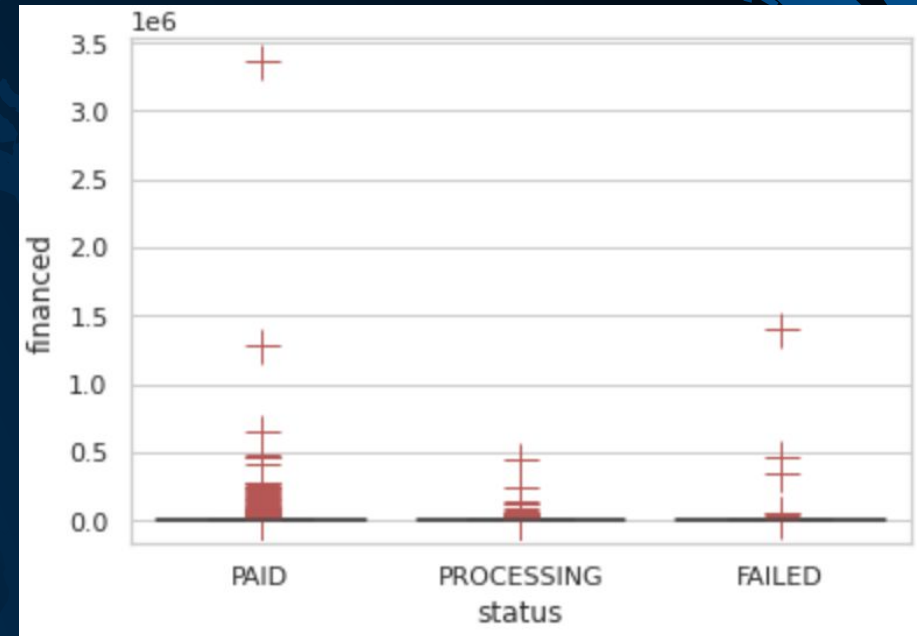
Los montos están distribuidos sobre un extremo.

# Distribuciones

Amount



Financiado



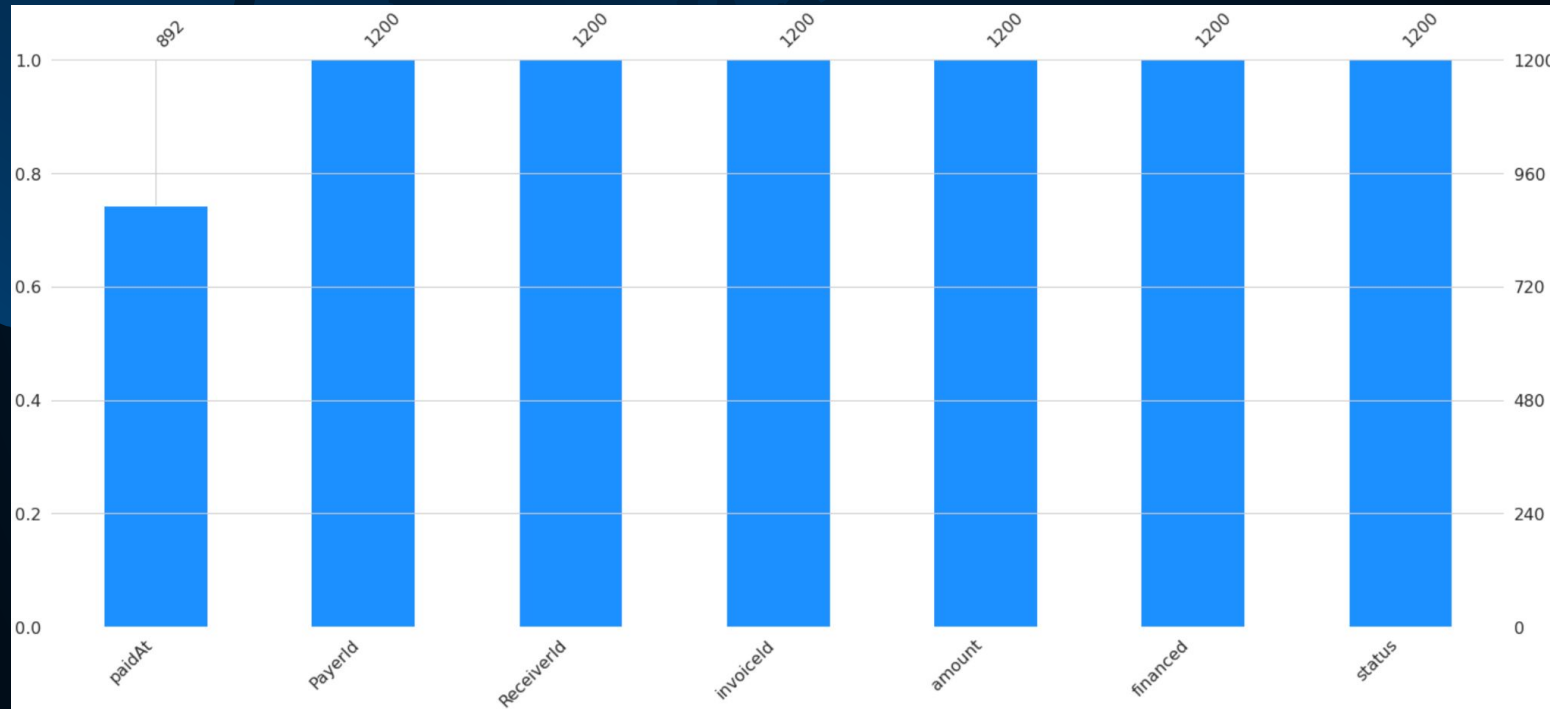
**Montos pagados:**

- Valores hacia un extremo.
- Outlier in PAID

**Financiados:**

- Valores hacia un extremo.
- Outlier in PAID

# Nulos



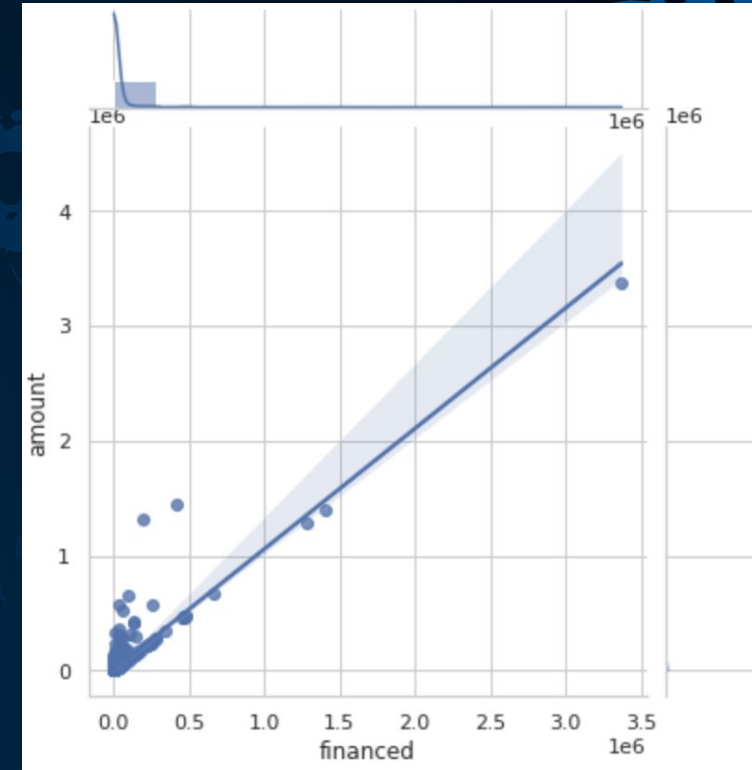
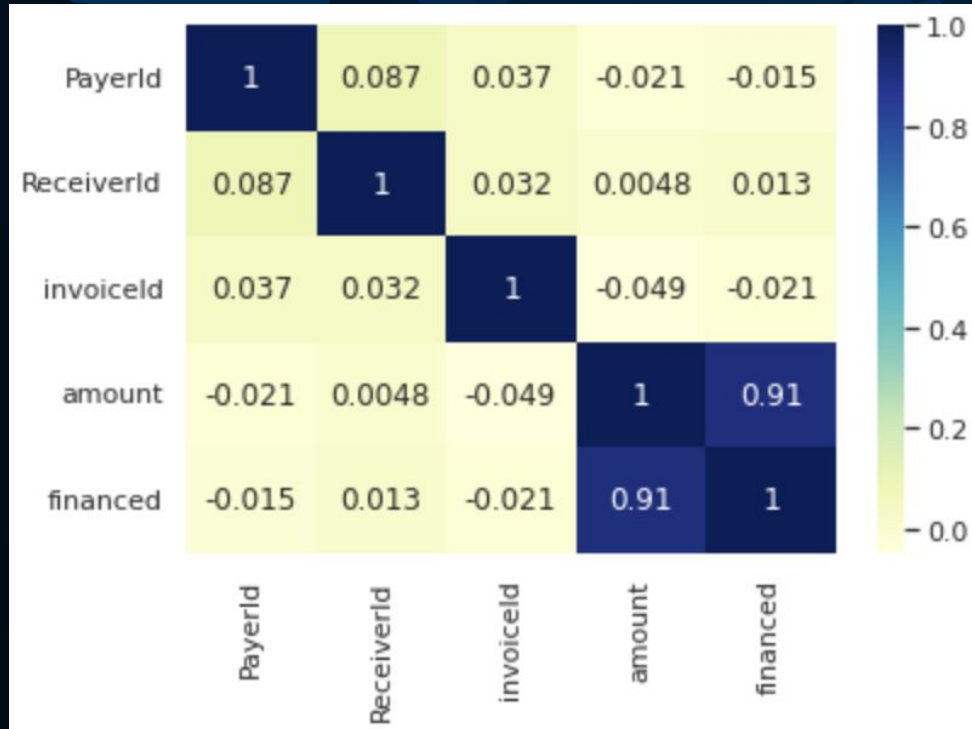
PAID = 892 (Sin nulos)  
PROCESSING = 269  
FAILED = 39

Nulls = PROCESSING + FAILED = **308**

**PaidAt = 308 nulos**  
(sin información dates  
transacciones no pagas)



# Correlaciones



- La máxima correlación es entre monto pagado y financiado. Puede ser debido a una definición de negocio.
- No hay una excelente correlación entre montos y features. Hay que trabajar las variables.



# Ingeniería de Variables (FE)

- ❑ **Básico:** tiempo, nulos, formato variables.
- ❑ **Outliers:** eliminar outliers.
- ❑ **Escalado:** escalar columnas numéricas.
- ❑ 3 Data sets para entrenar.







# Ingeniería de variables:

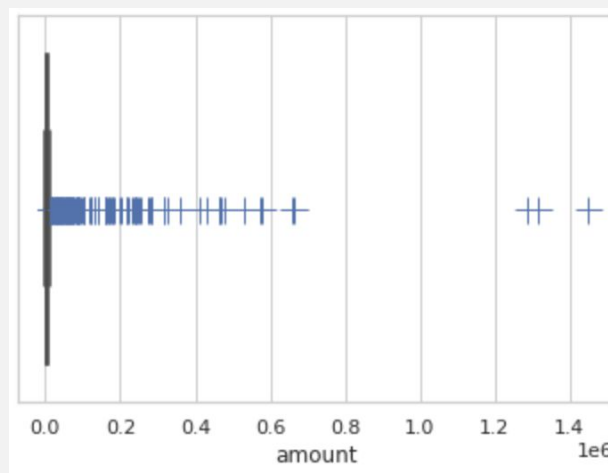
## Básica:

- **Tiempo:** se transformó el tiempo a cantidad de días (variable numérica).
- **Nulos:** se eliminaron filas sin datos que aporten al target.
- **Formato:** int o float.

days	amount	amountfinancedByXepelin
0	1490.46	0.00
0	920.26	0.00
7	4035.26	0.00
10	27979.20	10520.15
22	1477.46	0.00

## Outlier:

- Se eliminó el outlier para probar el comportamiento del modelo sin ese extremo.



## Escalado:

- Dado el rango de datos, para mejorar el modelo se escalaron las variables numéricas.

	amount	amountfinancedByXepelin
9	-0.189905	-0.140745
8	-0.193606	-0.140745
15	-0.173390	-0.140745
2	-0.017997	-0.060917
12	-0.189990	-0.140745



# Modelos: hipótesis. Regresión.

- ❑ Linear Regression.
- ❑ Cat Boost regressor.
- ❑ Autogluon: XGB, KNN, ExtraTress, CB, lightGBM, NeuralNetTorch, RF, Ensemble.
- ❑ Target x 2: monto pagado, financiado.



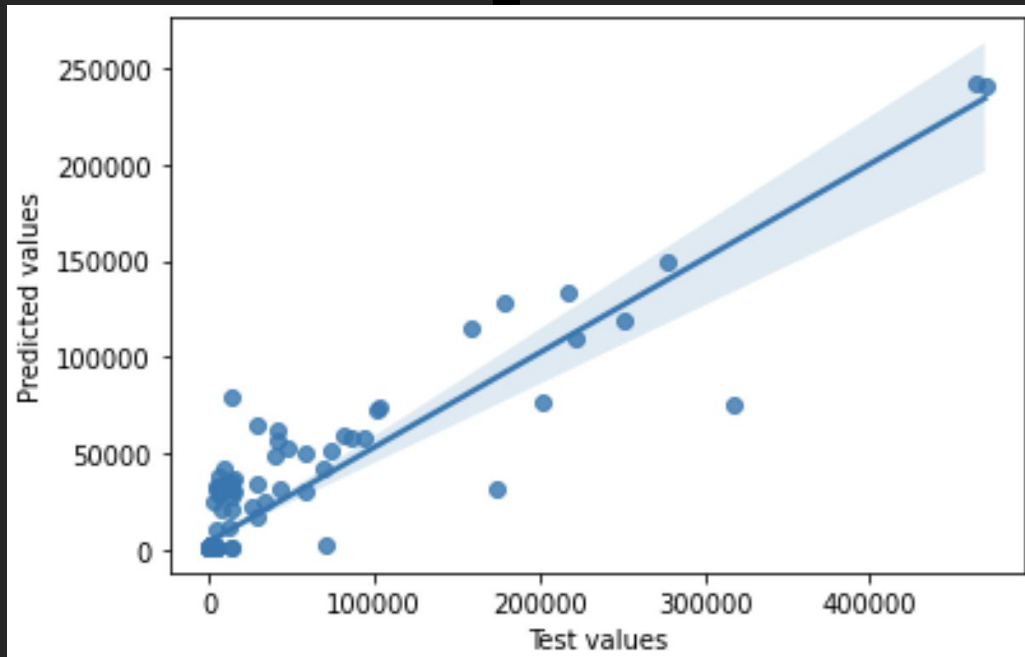


# Básico

Mejores Resultados

Nota: se presentan los mejores valores, en el código de GitHub están todos los demás.

# Target: Amount



## Básico.

Best Evaluations on test data:

"root\_mean\_squared\_error": -32700.21812527079,

"mean\_squared\_error": -1069304265.4402882,

"mean\_absolute\_error": -9565.182551713424,

"r2": 0.6873496603624003,

"pearson": 0.9170478762304354

MODELO: **KNeighborsDist**

# Target: Financiado



## Básico.

Best Evaluations on test data:

"root\_mean\_squared\_error": -18979.749368865476,

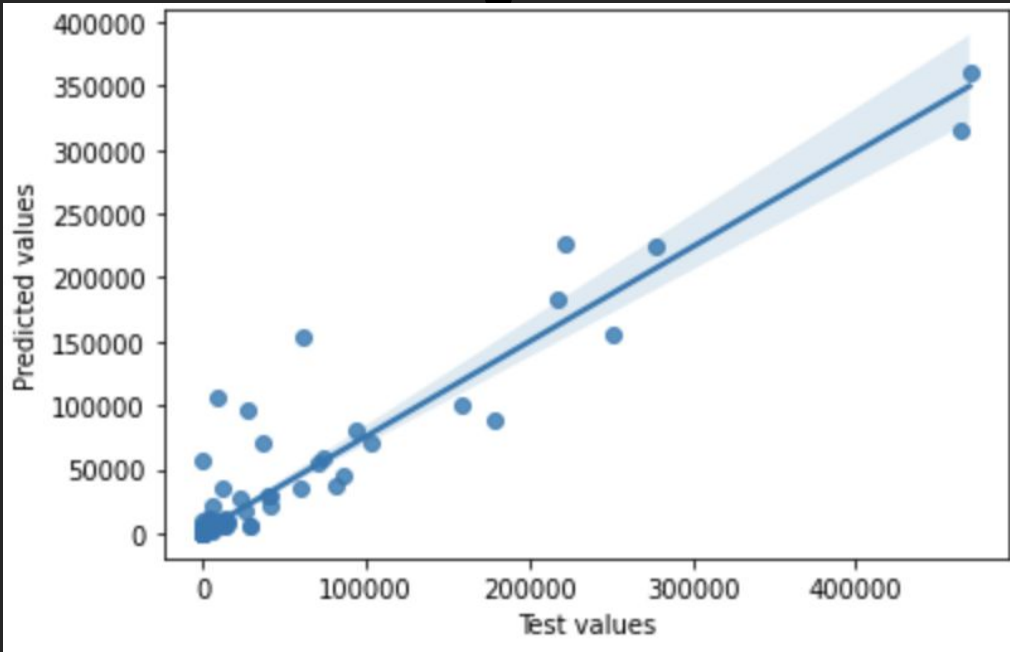
"mean\_squared\_error": -360230886.1049494,

"mean\_absolute\_error": -5218.375419505726,

"r2": 0.8733080036348294,

"pearson": 0.9504725217575404

MODELO: **KNeighborsUnif**





# Outliers

Resultados

Nota: se presentan los mejores valores, en el código de GitHub están todos los demás.

# Target: Amount



## Outliers.

Best Evaluations on test data:

"root\_mean\_squared\_error": -29026.452276982633,

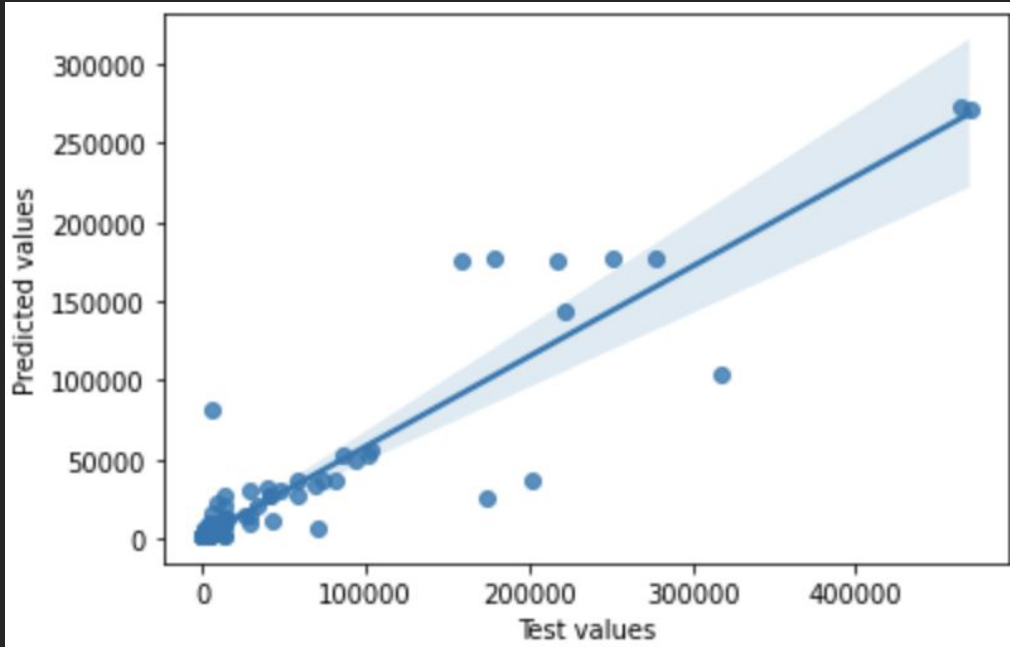
"mean\_squared\_error": -842534931.7879503,

"mean\_absolute\_error": -7868.054821388937,

"r2": 0.7536539962537405,

"pearsonr": 0.9320078996635309

MODELO: **KNeighborsDist**





# Target: Financiado



## Outliers.

Best Evaluations on test data:

"root\_mean\_squared\_error": -18979.749368865476,

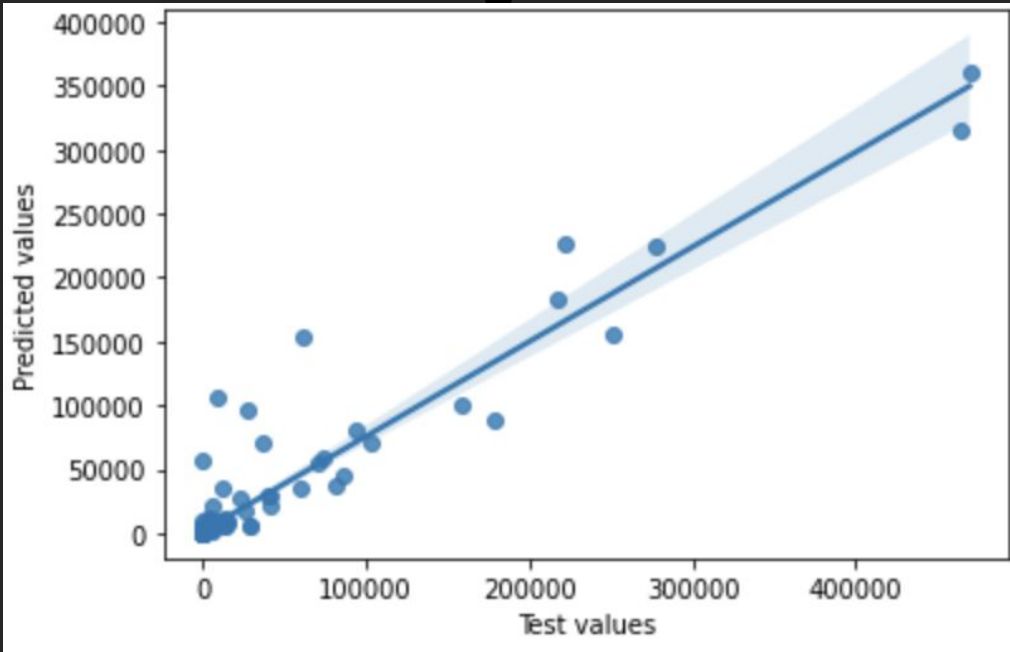
"mean\_squared\_error": -360230886.1049494,

"mean\_absolute\_error": -5218.375419505726,

"r2": 0.8733080036348294,

"pearson": 0.9504725217575404

MODELO: **KNeighborsUnif**







# Scale

Resultados

Nota: se presentan los mejores valores, en el código de GitHub están todos los demás.

# Target: Amount



## Scale.

Best Evaluations on test data:

"root\_mean\_squared\_error": -0.0708471683271134,

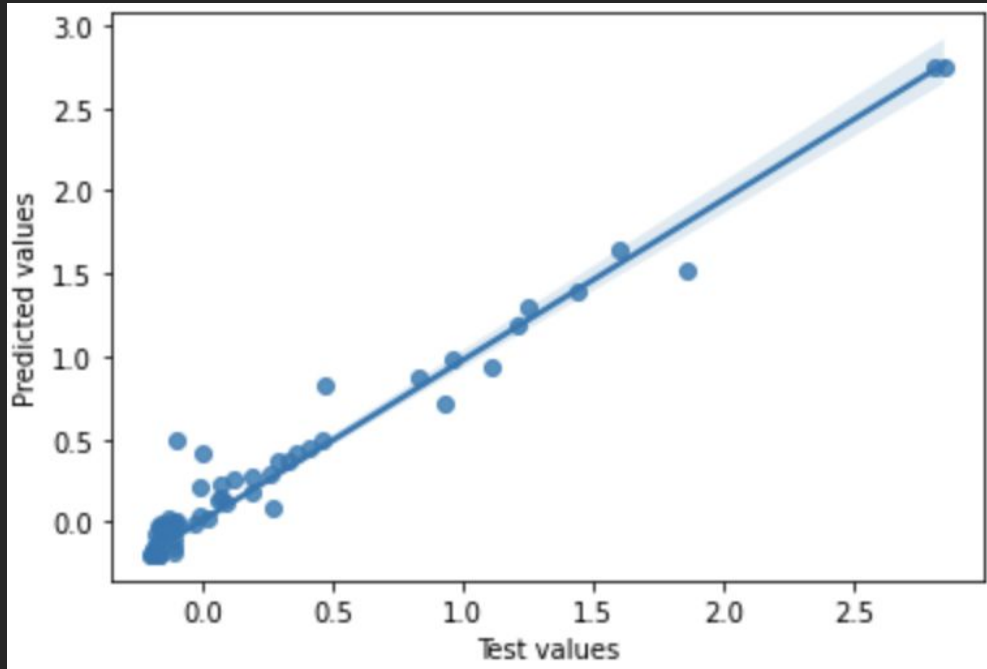
"mean\_squared\_error": -0.005019321259970339,

"mean\_absolute\_error": -0.026002506859516578,

"r2": 0.9651557850811286,

"pearson": 0.9829974167440408

MODELO: **KNeighborsDist**



# Target: Financiado



## Outliers.

Best Evaluations on test data:

"root\_mean\_squared\_error": -0.07002561129825359,

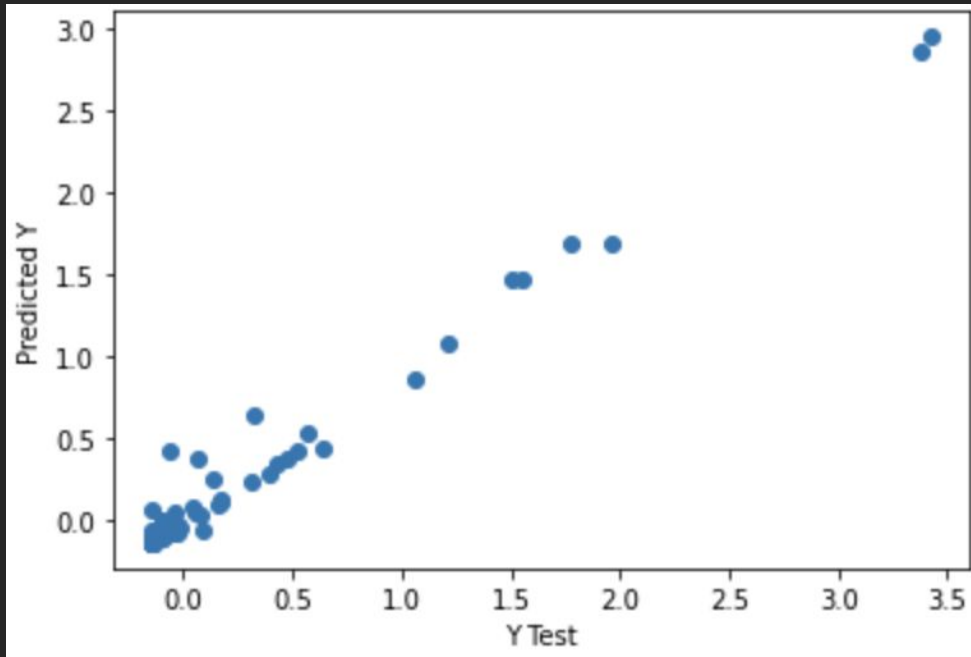
"mean\_squared\_error": -0.004903586237694101,

"mean\_absolute\_error": -0.019359172786031976,

"r2": 0.970048681878183,

"pearson": 0.9897920152931613

MODELO: **KNeighborsDist**





# CONCLUSIONES

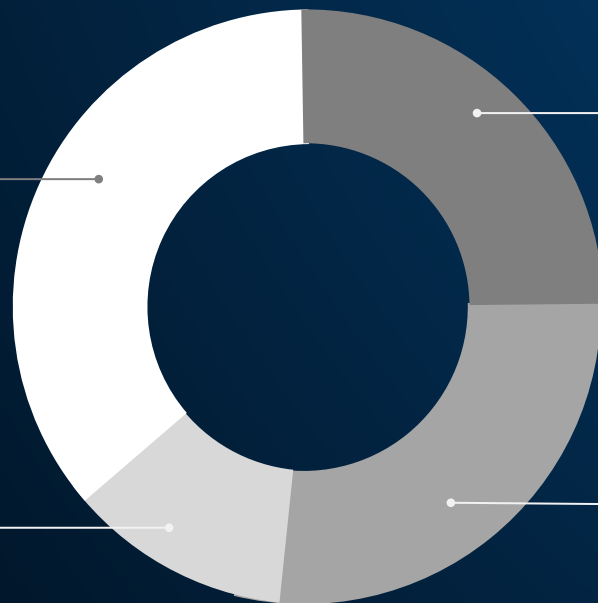




# Ventajas del Modelo.

Predicción  
de Montos

Buen resultado  
con pocas  
features

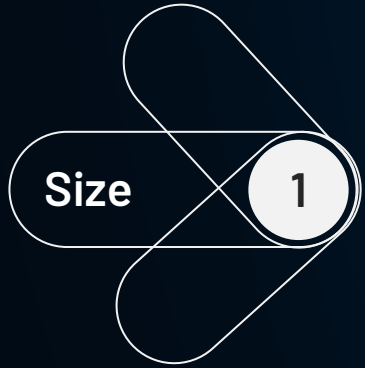


Modelo  
escalable

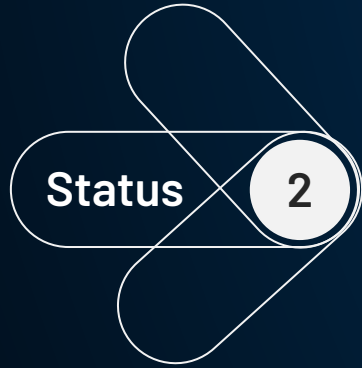
Métricas  
excelentes



# Future Adding: Información relevante.



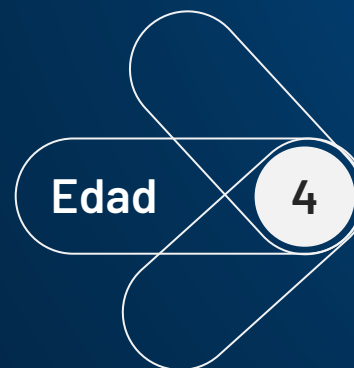
Tamaño más grande de datos disponibles.



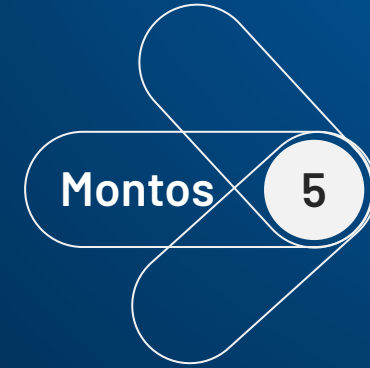
Estudiar qué sucede con las transacciones en proceso y fallidas.



Variables económicas de los clientes. Coeficiente de riesgo, mora, préstamos de otras instituciones, etc.



Historial de cuánto lleva el cliente en la empresa y en ejercicio de servicio.

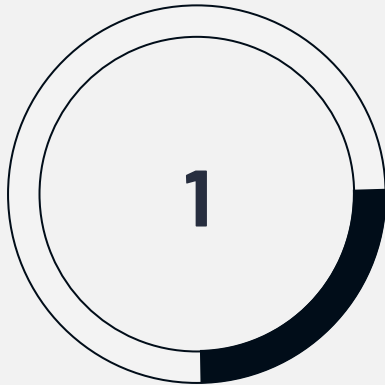


Montos con historia. Qué transacciones hizo el cliente y la relación con otros clientes.

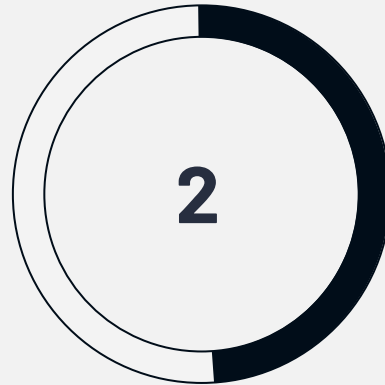




# Necesidades



Capacidad de procesamiento.



AWS: memoria.



MVP. Se tiene que probar en más datos.



Funcionamiento con todo el rango de status



# Gracias!

Guillermo J. Bergues

