PRESENTATION

ADIDAS FORECAST

MACHINE LEARNING

ANTONIO HERRERA

INTRODUCTION

PROJECT (PT. I)

ADIDAS has numerous distributors across the United States and wants to predict total sales based on multiple factors such as product type, unit price, shop location and date.

THE BENEFITS OF THIS PROJECT ARE...

- Know the total sales forecast for the future.
- See which companies will have the highest sales.
- To be able to set our objectives with greater accuracy.

DATASET

9641

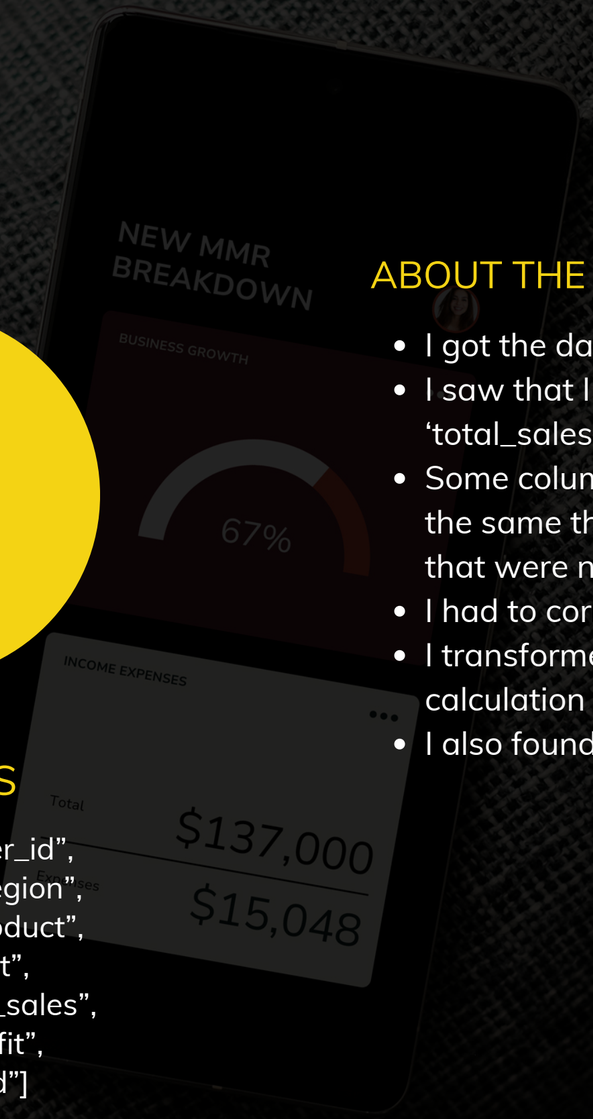
ENTRIES

We have some null and miscalculated data.

12

COLUMNS

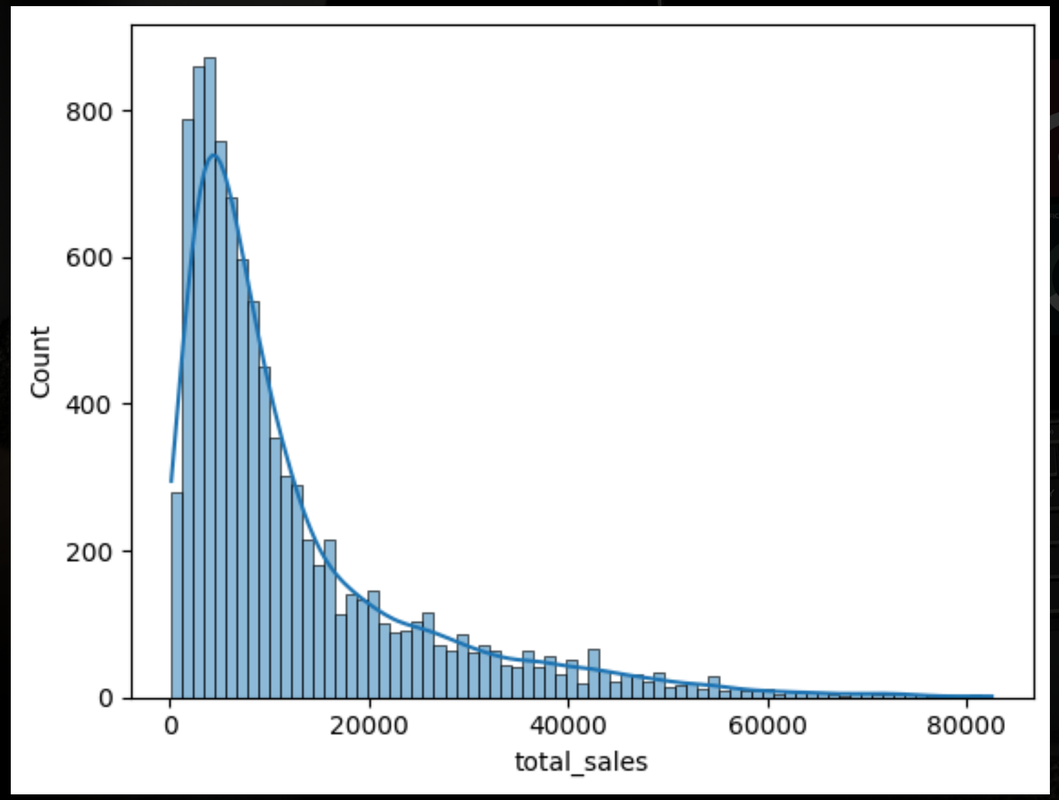
[”retailer”, “retailer_id”, “invoice_date”, “region”, “state”, “city”, “product”, “price_per_unit”, “units_sold”, “total_sales”, “operating_profit”, “sales_method”]



ABOUT THE DATASET:

- I got the data from the kaggle platform.
- I saw that I didn't have many nulls but I had my target ‘total_sales’ miscalculated.
- Some columns like ‘retailer’ and ‘retailer_id’ explained the same thing and I decided to delete those columns that were not going to be useful for this project.
- I had to correct symbols and spacing.
- I transformed the columns needed to do the target calculation correctly into numerical columns.
- I also found spelling mistakes.

ORIGINAL TARGET



160

FROM

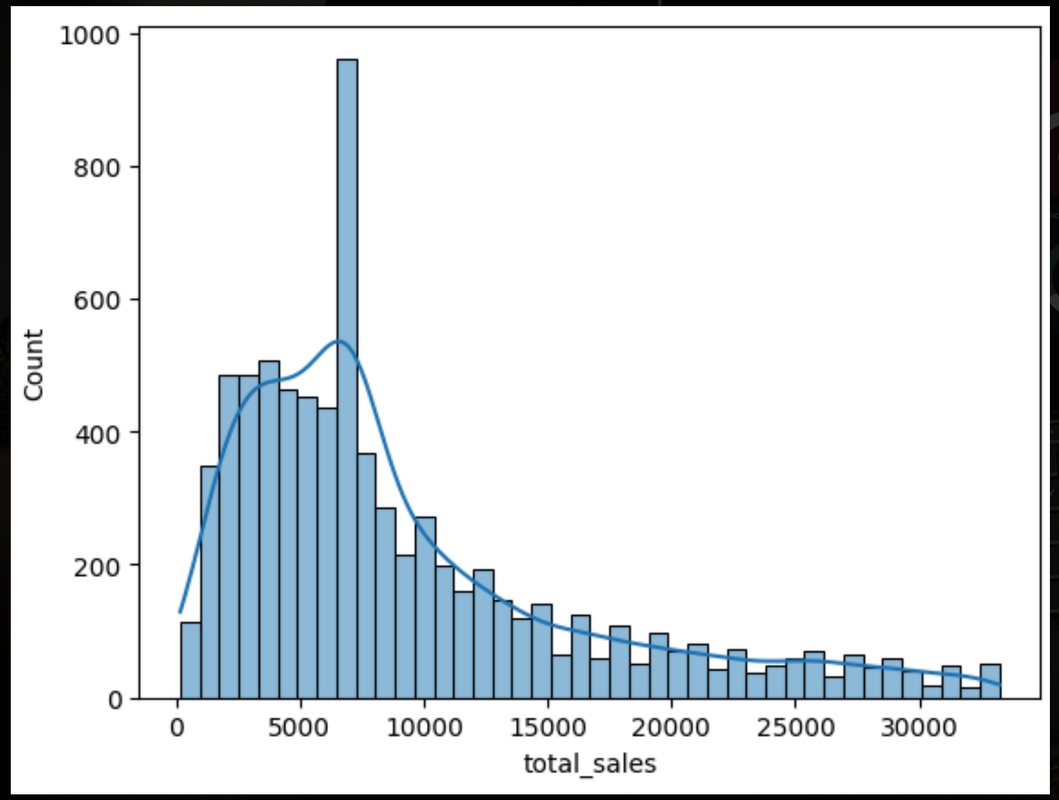
160 DOLLARS

82500

UP TO

82500 DOLLARS

TARGET IMPUTED



160

FROM

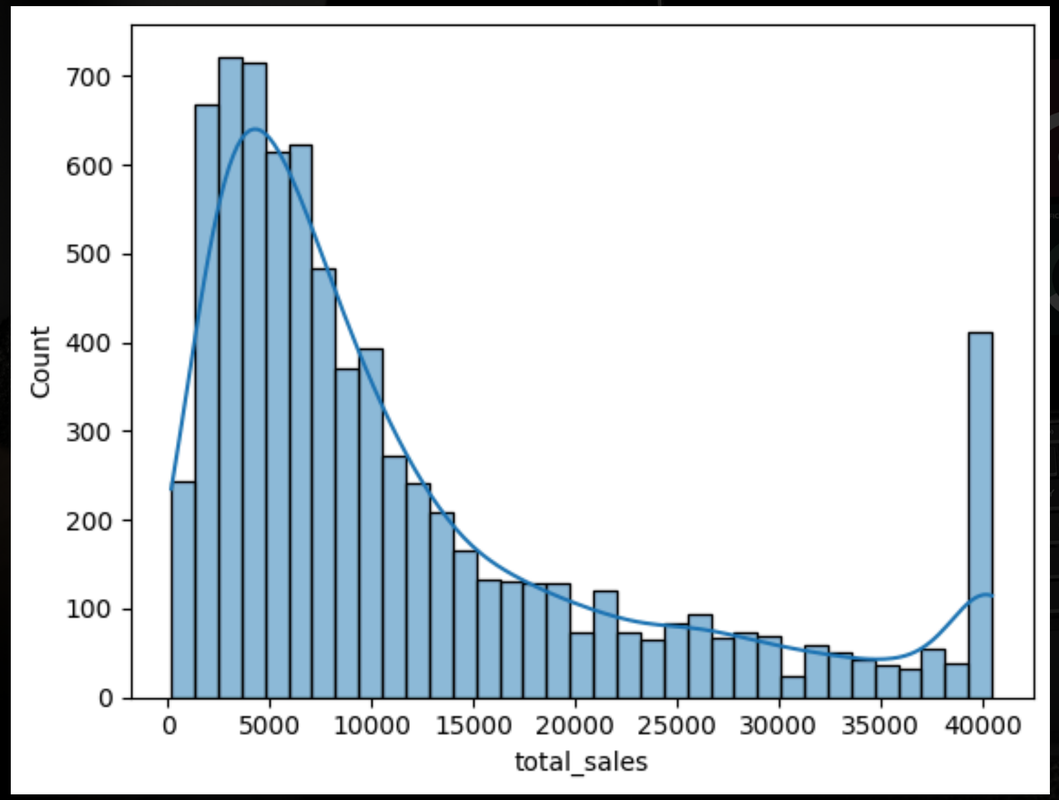
160 DOLLARS

33250

UP TO

33250 DOLLARS

TARGET WINDSORIZING



160

FROM

160 DOLLARS

40450

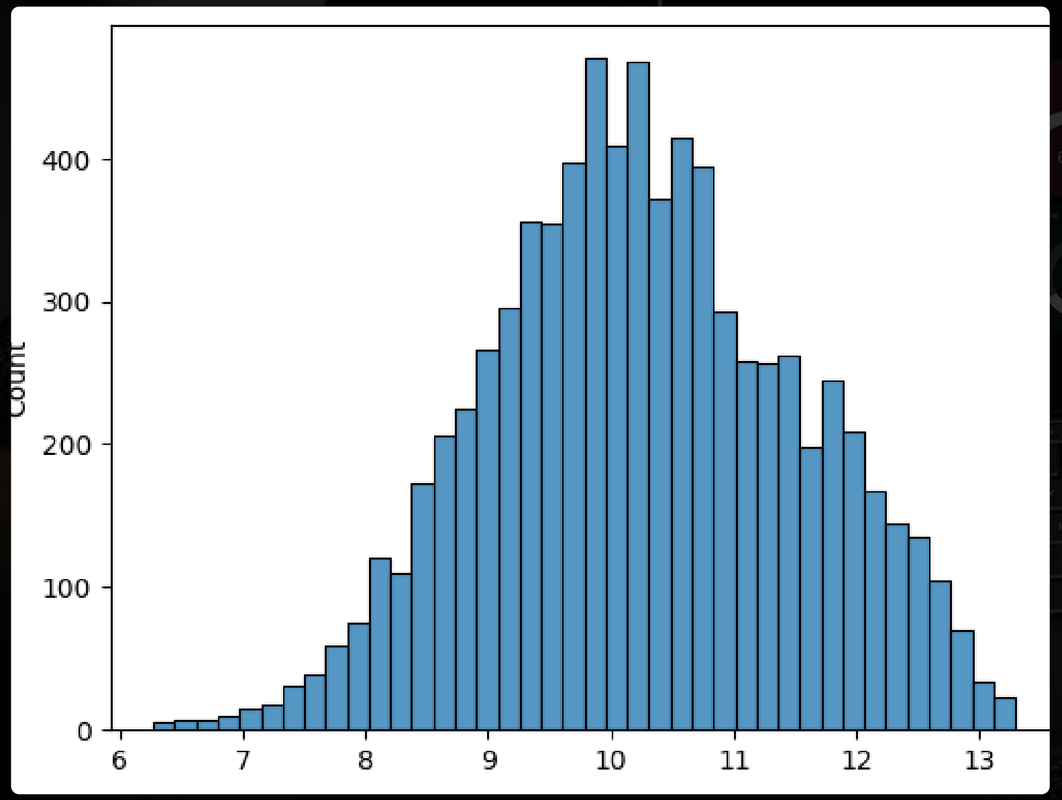
UP TO

40450 DOLLARS

TARGET FINAL

```
# Eliminamos outliers extremos
mask = (y_train >= 300) & (y_train <= 80000)

X_train_filtered = X_train[mask]
y_train_filtered = y_train[mask]
```



320

FROM

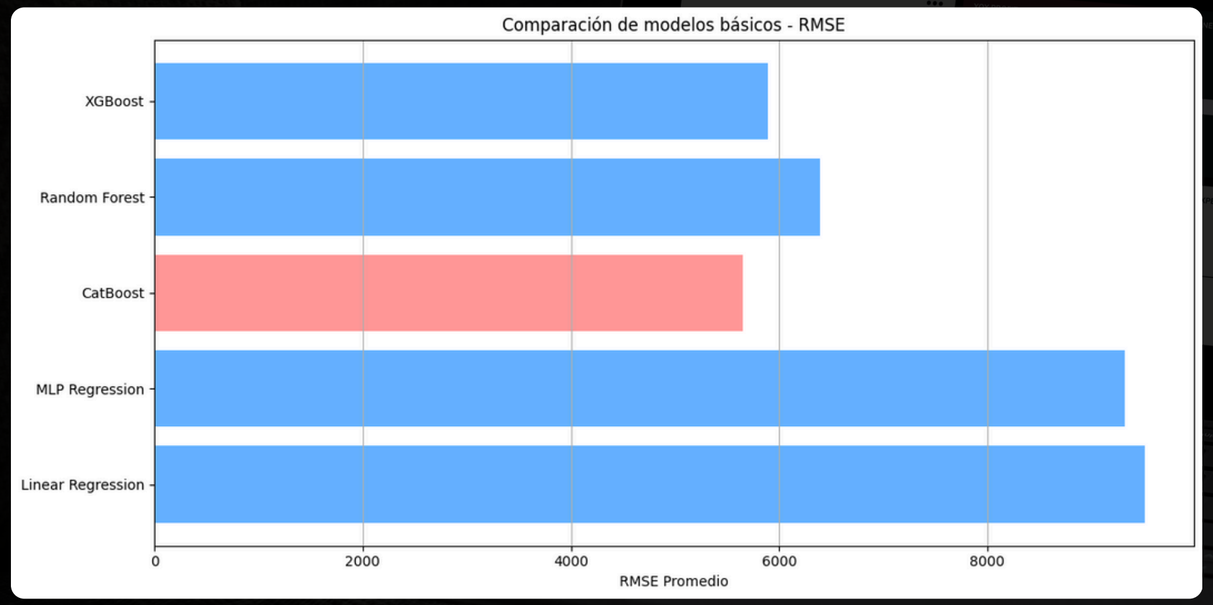
320 DOLLARS

78000

UP TO

78000 DOLLARS

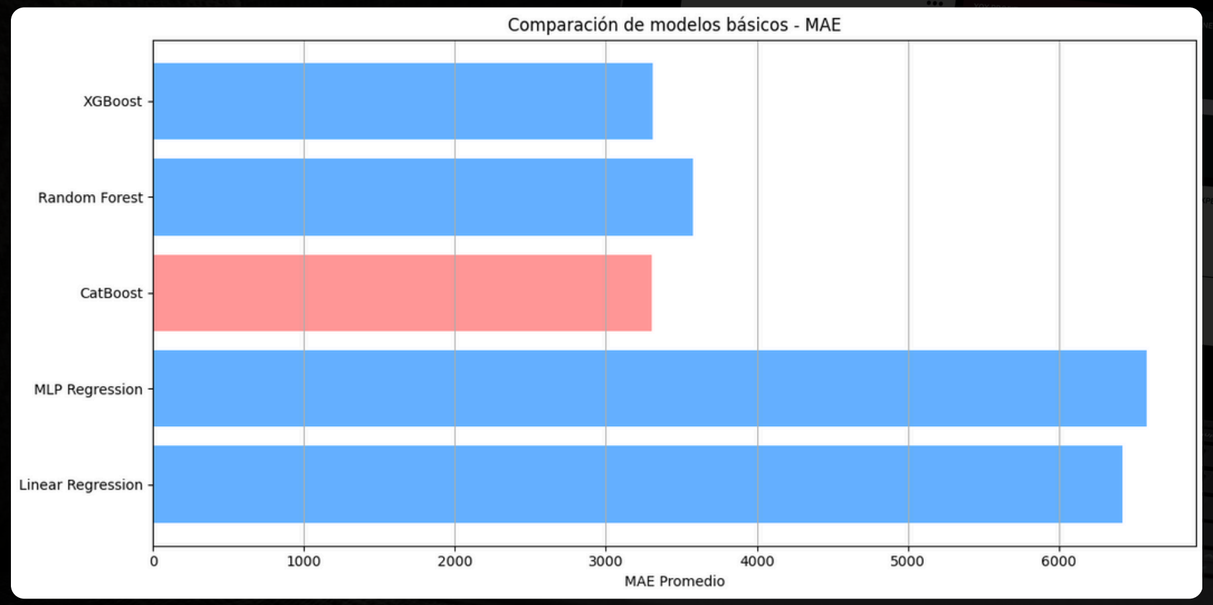
BASELINES MODELS



PROVISIONAL WINNER

CATBOOST

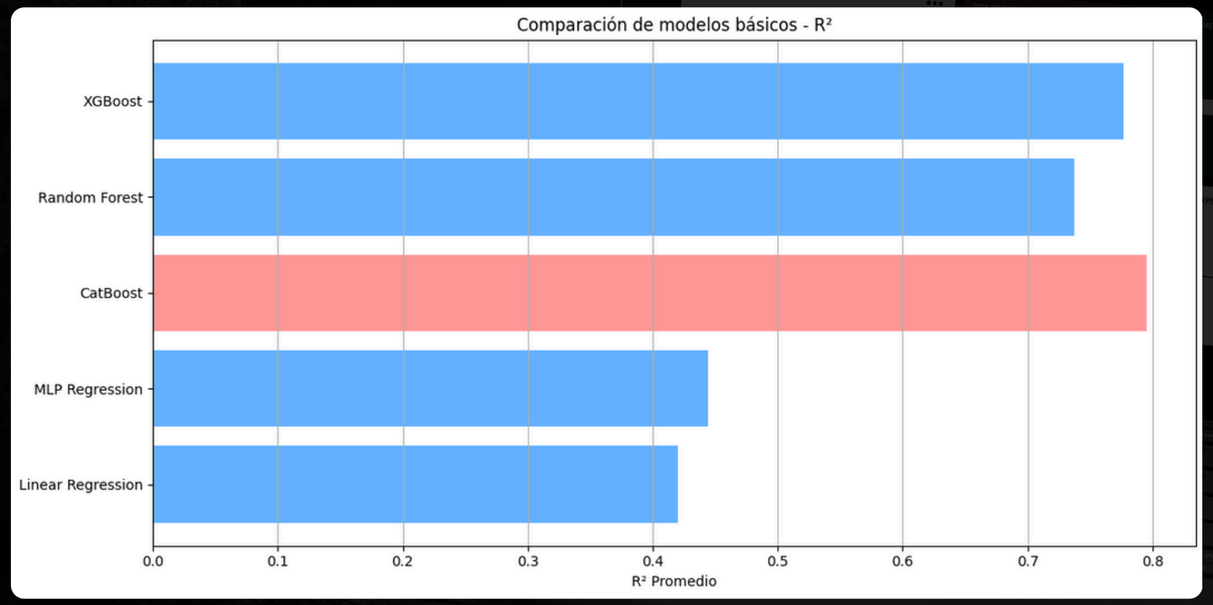
BASELINES MODELS



PROVISIONAL WINNER

CATBOOST

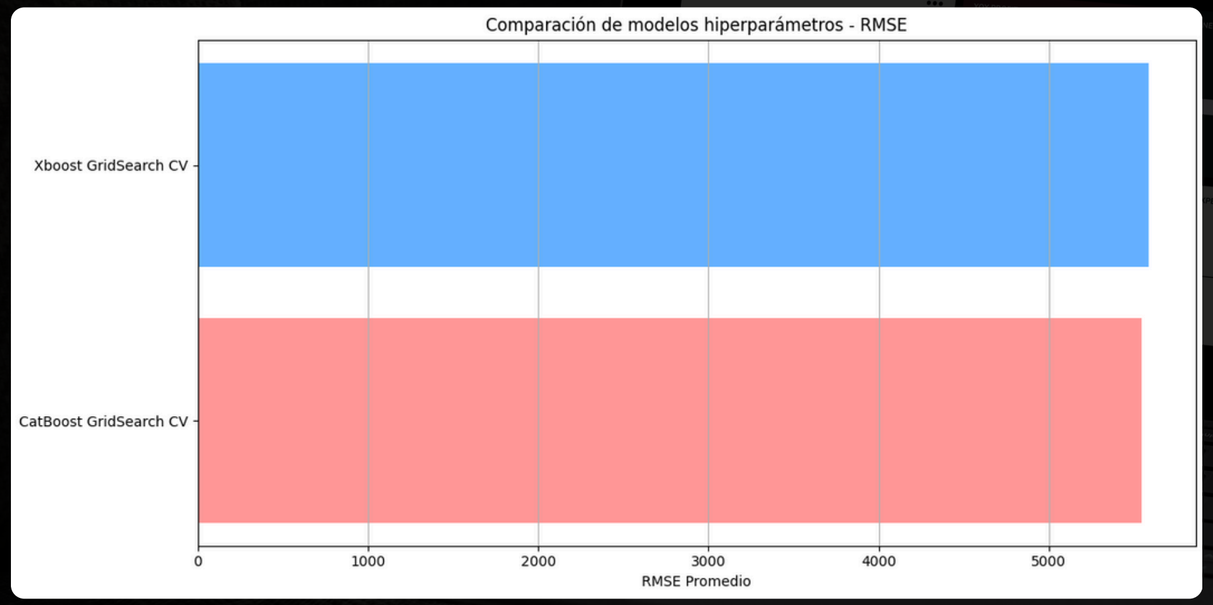
BASELINES MODELS



PROVISIONAL WINNER

CATBOOST

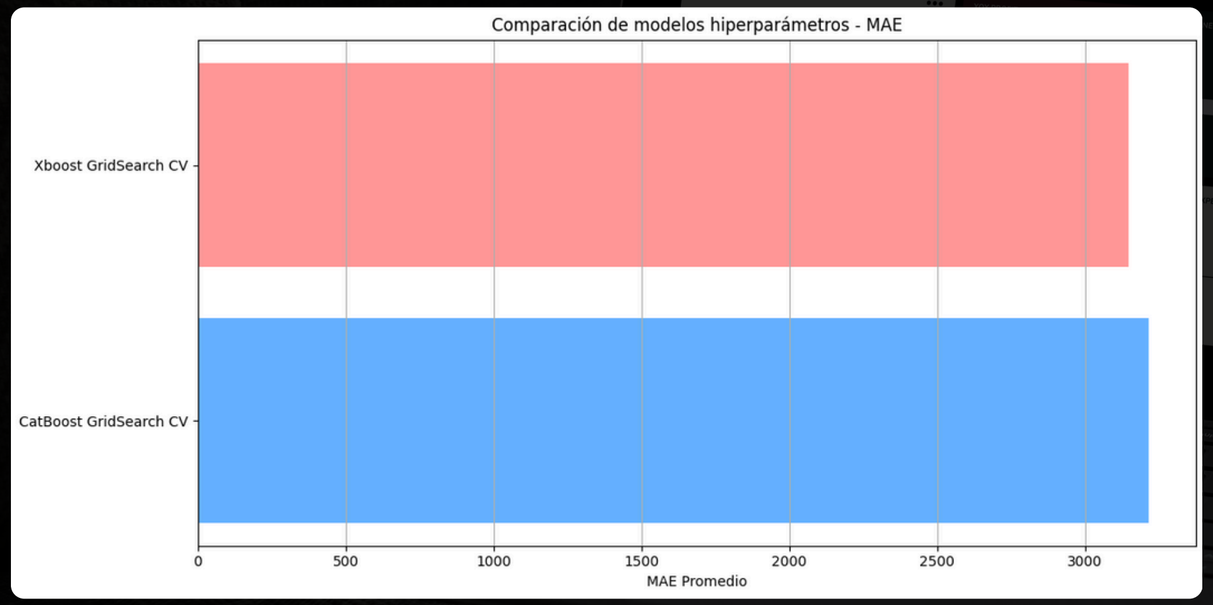
MODELS WITH HYPERPARAMETERS



PROVISIONAL WINNER

CATBOOST

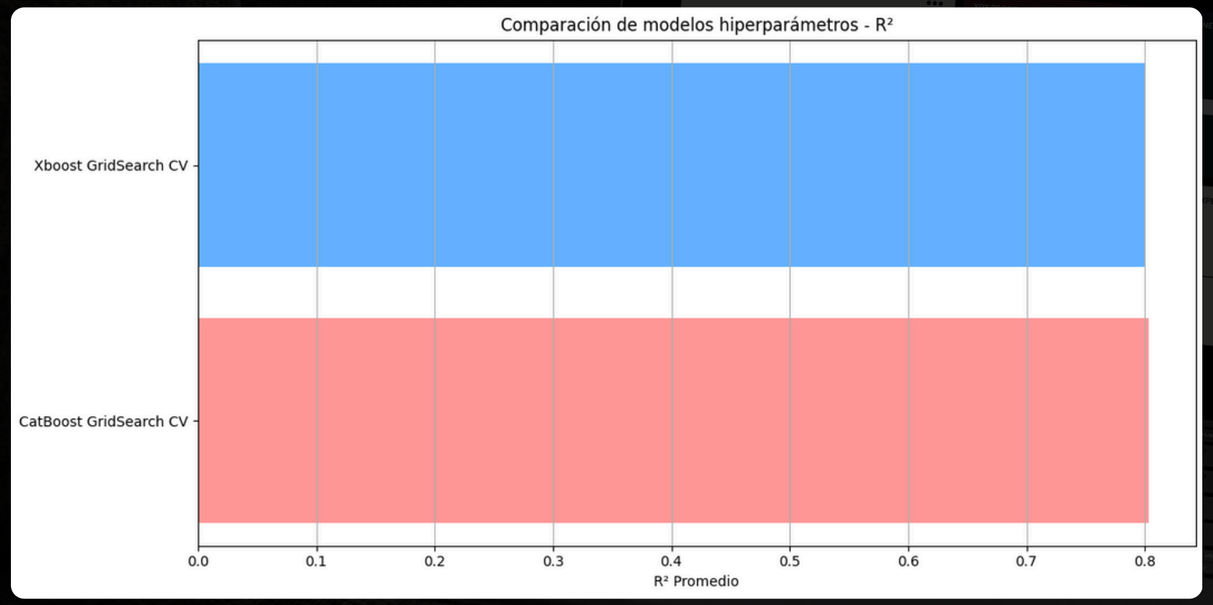
MODELS WITH HYPERPARAMETERS



PROVISIONAL WINNER

XGBOOST

MODELS WITH HYPERPARAMETERS



PROVISIONAL WINNER

CATBOOST

FINAL EVALUATION

```
---CATBOOST---
  Métrica  Entrenamiento   Validación         Test
0    RMSE    5080.578684  5541.396372  5260.050513
1     MAE    3134.637553  3214.433756  3136.047537
2       R       0.834973     0.803171     0.836928
---XGBOOST---
  Métrica  Entrenamiento   Validación         Test
0    RMSE    3979.642015  5584.249962  5185.203407
1     MAE    2327.322940  3148.272584  3067.247314
2       R       0.898745     0.799932     0.841536
```

CATBOOST

It is the chosen one. For me the strength it presents seems to be very important. Although there is not much improvement as we have done the process... I see that it is a firm and safe model.

XGBOOST

Better metrics and higher speed than CatBoost but your training may indicate to me that you are not generalising well. There may be a risk of overfitting.

CONCLUSIONS - IMPROVEMENTS

OPINION

As I was doing this project, I felt more and more important the ‘Business’ or ‘Team Leader’ part because many times there are details that we don't have and they are important. There are decisions that I have taken in this project that have been made by my own knowledge of the sector but depending on what the company ADIDAS is looking for, it could have given a different result. For example the choice of the model.

IMPROVEMENTS

Possibly we could analyse the case of XGBoost so that it does not OVERFITTING. Such trees tend to learn on training and become memorised. We must control their hyperparameters, for example reduce the depth.

I WANT TO SAY

THANK YOU

FOR YOUR ATTENTION