Boat Listing Case Study

Ademola Olokun

Boat Listing Insights

December 2022

Introduction

- Users can advertise their used boats for sale on Nearly New Nautical. Listings require users to provide a range of information about their boats. The more views a boat receives, the more traffic the website receives, and the more potential customers there are.

- In order to boost traffic to the website, the product manager wants to eliminate boats that do not receive many views from being listed by building a model that:
  - Predicts the number of views a listing will receive based on boat features
  - On average, predicts number of views that are only 50% off of the true number of views a listing would receive

Methodology

- The dataset consists of 9,888 boat listings with 10 features, including the target label, Number of Views.

- I am choosing to use CatBoost – a gradient boosting library, because we have more categorical features with moderate cardinality and CatBoost is better equipped to handle categorical features without negatively affecting the performance of the model. A Linear Regression model will serve as the baseline model.

- For the evaluation metric, I am choosing Mean Absolute Percentage Error to evaluate the model, which measures the prediction accuracy as a loss function of our regression model.

- The MAPE of our best performing predictive model is 0.5 (50%).

Outcome

model_name	RMSE	MAPE	R2
LinearRegression	129.700000	0.800000	0.170000
CatBoostRegressor	111.460000	0.500000	0.390000

The **best performing model** for the business case is the **CatBoostRegressor model**, with a MAPE value of 0.5

Outcome

LinearRegression Predicted Value and True Value for first 100 observations of Test Data
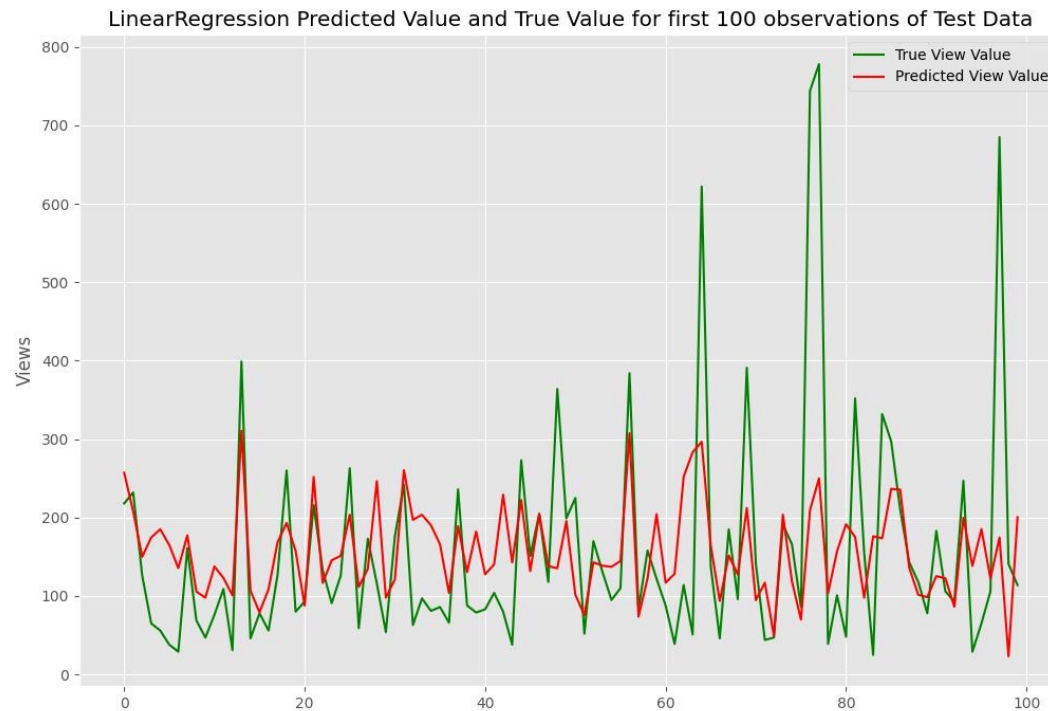
Our KPI, Mean Absolute Percentage Error for the Linear Regression model, is 0.8 (80%).

This means, on average, our predicted views will be 80% off the true number of views.

As seen from the chart, a subset of the data shows that predicted views and true views differ greatly.

Outcome

CatBoostRegressor Predicted Value and True Value for first 100 observations of Test Data
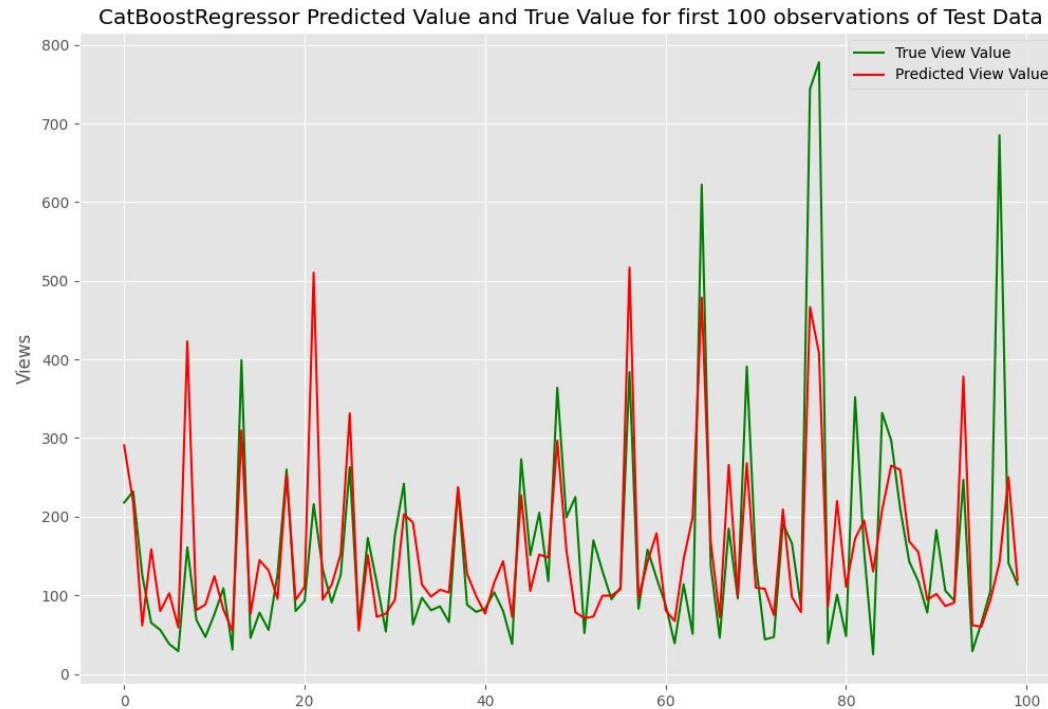
Our KPI, Mean Absolute Percentage Error for the CatBoostRegressor model, is 0.5 (50%).

This means, on average, our predicted views will be 50% off the true number of views.

As seen from the chart, a subset of the data shows that predicted views and true views are mostly overlapping.

Outcome

Important Model Features
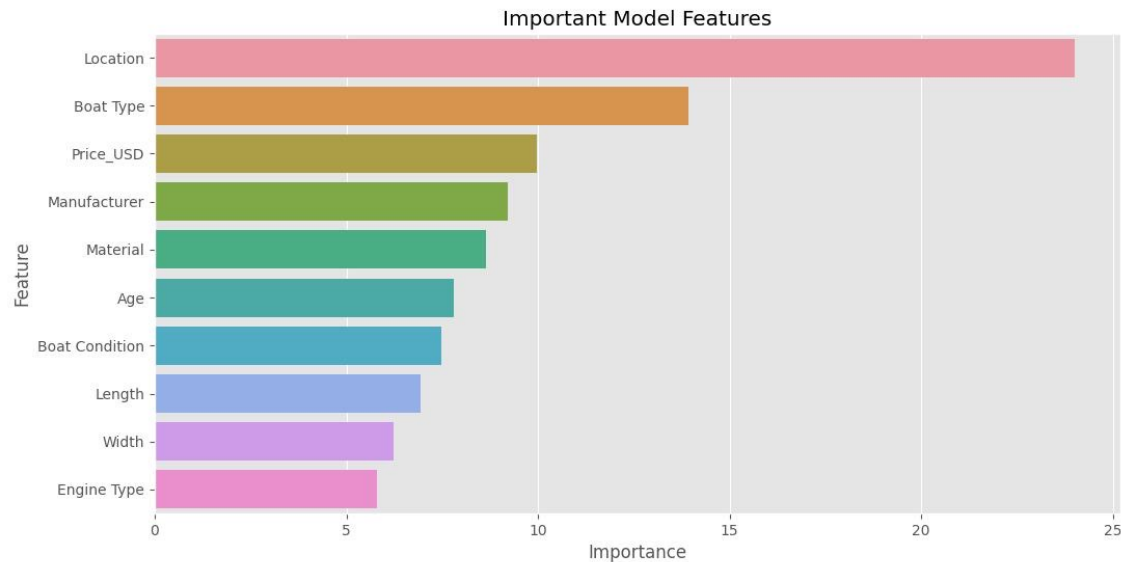
From the chart, we can see the top features that impact our CatBoostRegressor model greatly:
- **Location**
- **Boat Type**
- **Price**
- Manufacturer
- Material
- Age
- Boat Condition
- Length
- Width
- Engine Type

Recommendations

To help increase traffic to the website, we can deploy this model to filter out boat listings that do not meet a threshold of number of views.

To better improve the model, I would recommend changing the input for the Location field from country to regions, to reduce the number of unique values, and potentially improving the performance of the model.

I would also recommend that the data fields are validated to minimize the occurrence of missing values, improving data analysis and the performance of the model.

Thank you