OLIST STORE: PRICE PREDICTION PROJECT

PROJECT DOCUMENTATION

Submitted by

JOESON MISIANI AMORO

ZALEGO INSTITUTE

AUGUST 2021

ABSTRACT

The importance of predicting price of products is to help the store know which products sell the most and sell the least, using this trend the store can know what to do with their products.

With 9 datasets that has information of 100,000 orders made by several customers from multiple marketplaces, I will build machine learning model that will predict the prices of products sold by the store. Since the problem can be solved with a regression algorithm, I have created linear regression model and XGBRegressor model to come up with a solution, from the two model a will finally choose the best model based on their accuracy score.

My analysis shows that August is when the store gets more review scores from the customers, more products are being purchased in the month of May and August, most product are being ordered from 11am to 4pm, most customers prefer using credit cards while making payments, most selling products are Furniture decor, bed bath table, computers accessories, housewares, sports leisure and least selling products are fashion children's clothes, music, small appliances home oven and coffee, CDs DVDs musicals, arts and craftmanship

My study revealed that customers were purchasing products and also make review of products at high peak in specific months, this will help the store to know the best time to stock and what to stock which will help them know when to make big sells.

INTRODUCTION

Background of the study

This project is to do analysis on Olist store based on the information of 100,000 orders made by several customers from multiple marketplaces, and help the store owner(s) know the trend of their sells.

Most stores tend to get loss in terms of profit because they don't know the trend of the products they sell, which products sells the most or the least so that they can focus on products that generates profit for them.

Some products are getting more sells based on the environment they are in, the people leaving around the store or even influence of word-of-mouth from the customers that have already bought the product.

Statement of the problem

The store faces issues on generating profit from the products they sell because some products are getting more sells than others which shows that the effort, they put on selling some products doesn't reflect in terms of the money they bring in the store.

The analysis done, shows that some products are likely to get more sells on specific customer cities than others because of the customers wants who leave

Research questions

1. Which products that are mostly sold in the store?
2. Which products that are least sold in olist store?
3. Which month does the store gets more review scores for their products?
4. Which month does customers purchase more products from the store?
5. When is the best time that customers purchase products from the store?

Objectives or purpose

1. Most selling products are Furniture decor, bed bath table, computers accessories, housewares, sports leisure.
2. Least selling products are fashion children's clothes, music, small appliances home oven and coffee, CDs DVDs musicals, arts and craftmanship.
3. Review scores are high in the month of August
4. Most products are sold on the month of May and August
5. Most product are being ordered from 11am to 4pm

Significance of the study

The store should concentrate on the products which are most selling so that they can generate a reasonable profit, the store should introduce new stock in the store specifically on the month of August which has a high review score making it easier to convince customers when they make decisions, also on the month of May and August the store should put more effort on their stoke to bring more profit in the store also to consider the best time when most product are being ordered that is from 11am to 4pm.

RESEARCH METHODOLOGY

Methodological approach

The store faced issues on generating profit from the products they sold because some products were getting more sells than others which showed that the effort they used when selling some products didn't reflect in terms of the money they brought in the store.

A survey was done in the marketplaces where Olist stores were located to be familiar on how the customers did their review on the products, what were the best time customers made purchase orders, which products got most sells and least sell from the customers and which month the stores got more sells from their products.

Methods of data collection

The survey was conducted anonymously which was to monitor how customers made purchases in Olist stores in various marketplaces.

DATA ANALYSIS

Methods of analysis

Before analysis the gathered data was prepared in form of spreadsheets. The dataset was checked to see if it had any missing values in any column. I found 9 columns that had missing values and after weighing them, I found that "product_weight_g" column was useful so I used the mean value of the same column to fill the missing values of that column and the rest the rest of the columns were dropped since they were not useful in the model.

Then on dealing with duplicated data I used Key Features based strategy, I created a group of critical features as unique identifiers in the dataset to check if there were duplicates based on them, I created another dataset by removing duplicated value in the original dataset, after comparing the shapes of the two datasets and found out that there were 5,561 duplicated rows based on the unique identifiers I created earlier which I removed when creating the new dataset.

The dataset had columns containing date and time that were in object data type so I converted them to be in datetime data type. Also, there were columns that were found to be unnecessary and cannot add value in the model so I decided to remove them.

Data preprocessing

Before creating the model, I realized that some columns contained categorical values that had to be transformed to numerical first, so I did data preprocessing for categorical variables and I used Label encoding to change values of the columns from text to strings

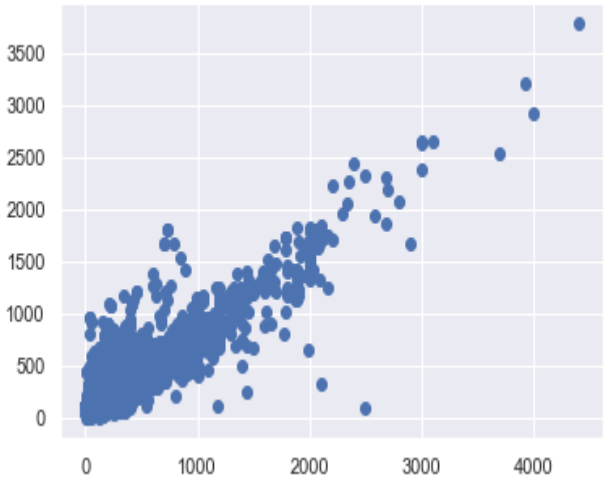
Feature scaling

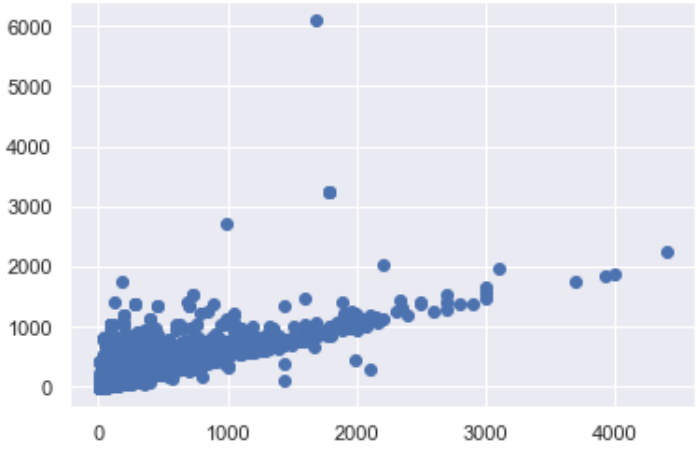
Then I split the dataset into target and features then I used train_test_split from sklearn library to split the target and features into train set and test set which can therefore be used by the model, Using MinMaxScaler from sklearn library to scale the features of the dataset, I created a scale fitted it with the train set so that it learns the parameters to ensure that there won't be overfitting or underfitting the model.

Building Models

When building the Model, I decided to create two models and weigh them in terms of their accuracy score, I created XGBRegressor Model and Linear Regression Model so that when I am evaluating them, I choose the best one to be my final Model.

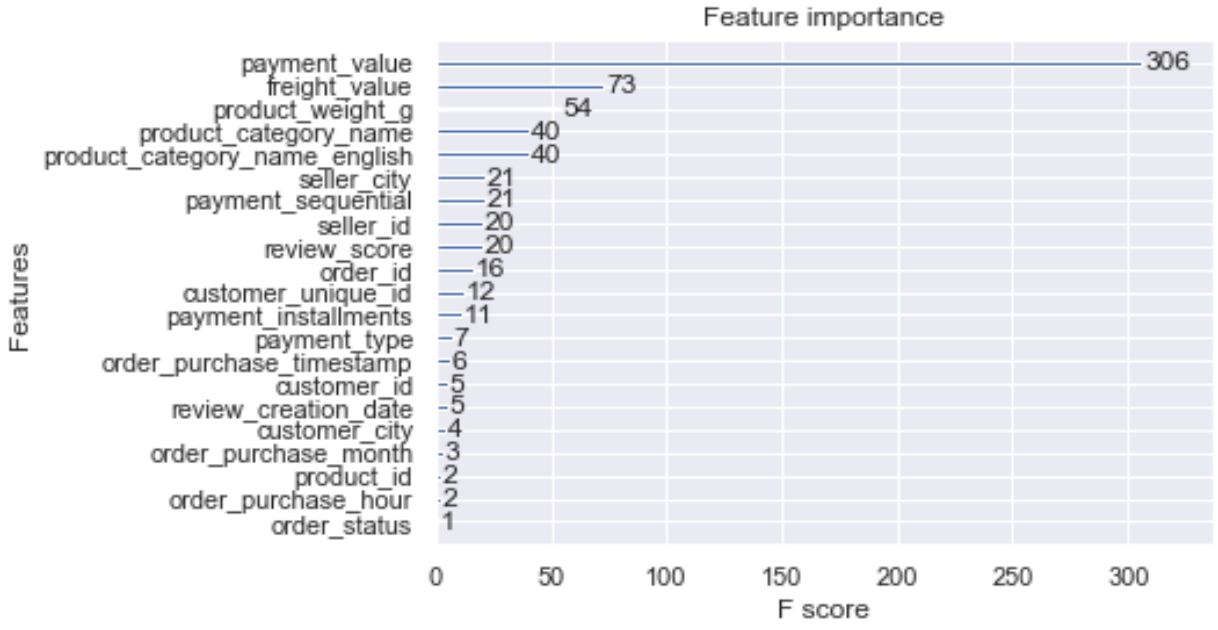
I started with building XGBRegressor Model, so I created the model trained it with train set then I did Model evaluation with both train set and test set on it and found that its training Accuracy was 83.76% and Model Accuracy was 80.36% also in mathematical evaluation I found the model had R-squared of 0.8036(4dp), MAE of 38.4976(4dp) and RMSE of 78.8427(4dp).





Finally, I built Model of Linear Regression and trained it with train set then I did Model evaluation with both train set and test set on it and found that its training Accuracy was 58.8% and Model Accuracy was 60.78% also in mathematical evaluation I found the model had R-squared of 0.6078(4dp), MAE of 52.2444(4dp) and RMSE of 111.4236(4dp).

Predictor variables are also known as x-axis variables, predictor variable with highest correlation is good predictor. I found these predictor variables being of more importance than the



Creating the Final Model

Know from the two Models that I have created and seen their accuracy scores, I choose to create a final Model using the model which had the highest Model Accuracy that is XGBRegressor. Also, I knew the feature to use in the final Model because not all of them were important in the previous Models I created.

Setting target and features, I used the first five features because they are important in the Model and the target remained constant then I used train_test_split from sklearn library to split the target and features into train set and test set which can therefore be used by the model.

Using MinMaxScaler from sklearn library to scale the features of the dataset, I created a scale fitted it with the train set so that it learns the parameters to ensure that there won't be overfitting or underfitting the model.

Then I created the model trained it with train set then I did Model evaluation with both train set and test set on it and found that its training Accuracy was 81.86% and Model Accuracy was 79.17% also in mathematical evaluation I found the

model had R-squared of 0.7917(4dp), MAE of 39.3163(4dp) and RMSE of 81.2056(4dp).

Explanation of R-squared:
* Shows how well the data fit the regression model
* Generally, a higher r-squared indicates a better fit for the model

Explanation of MAE:
* Represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set

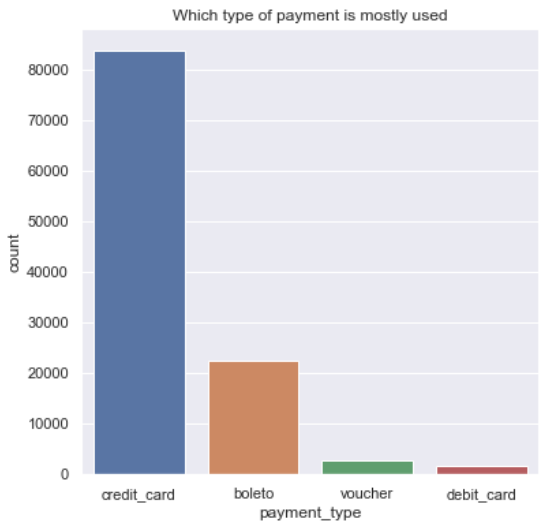
Explanation of RMSE:
* Interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable.
* Lower values of RMSE indicate better fit.

RESULT AND FINDING DISCUSSION

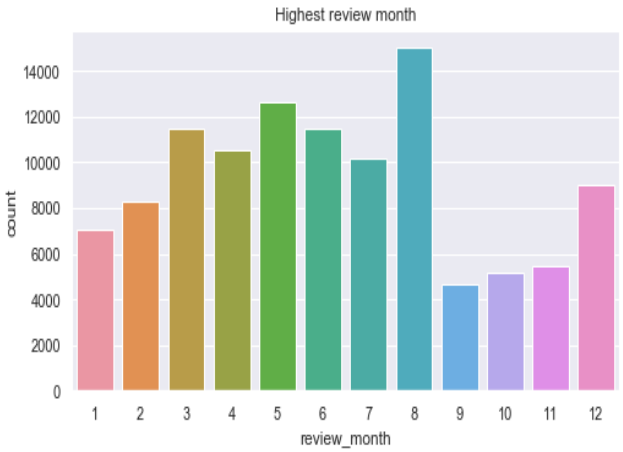
According to the analysis I found out that most sold products in Olist store are furniture decor which was bought by 3270 customers, bed bath table which was bought by 3148 customers, computers accessories which was bought by 1933 customers, housewares which was bought by 1837 customers, sports leisure which was bought by 1692 customers while least selling products are fashion children's clothes which were bought by 2 customers, music which was bought by 2 customers, small appliances home oven and coffee which was bought by 2 customers, CDs DVDs musicals which was bought by 4 customers and arts and craftmanship which was bought by 4 customers.

Products with high review scores on average were computers accessories, garden tools, sports leisure, bed bath table, toys, home comfort while products with low review scores on average were bed bath table, luggage accessories and watches gifts.

There were four types of payment method that happened in olist store, which were: credit card, boleto, voucher and debit card. And analysis showed that most preferred mode of payment by customers is credit card while the least preferred mode of payment by customers is debit card.



August is when the store got more review scores from the customers while more products were being purchased in the month of May and August



Most product were being ordered by customers from 11am to 4pm

My study revealed that customers were purchasing products and also make review of products at high peak in the month of May and August, this will help the store to know the best time to stock and what to stock which will help them know when to make big sells.

Justify your approach

There were several algorithms that I opted not to use like Random Forest, Decision Tree Regression and KNN Model because they didn't accept features that were continuous so I used Linear Regression and XGBRegressor weighed them and finally decided to use XGBRegressor as my final model because of their accuracy score.

CONCLUSION AND RECOMMENDATION

1. Most selling products are Furniture decor, bed bath table, computers accessories, housewares, sports leisure, I recommend the store to put more effort on these products because they are most selling products in the store showing that customers like them.

2. Least selling products are fashion children's clothes, music, small appliances home oven and coffee, CDs DVDs musicals, arts and craftmanship, I recommend the store to change or stop selling these products because they don't bring profit in the store which clearly shows that customers are not interested with them.

3. Review scores are high in the month of August, the store should be aware on this month so as to stoke more and even introduce new products in the store so that the customers may be convinced to try new products.

4. Most products are sold on the month of May and August, I recommend the store to stock more so that they can make bi g sells and generate profit from the products sold.

5. Most product are being ordered from 11am to 4pm, I recommend that the store to display their stock at this time to convince the customers when they are making purchase orders.