Price Prediction For Sale

Dipta Das Chowdhury, Keshav Agarwal, Nimesh Krishnani

November 18, 2018

1 Problem Statement

Setting the optimum price of a product is often a problem for retailers, espically during sale.Our solution looks into estimating the sale price Given the dataset, we estimate the price a customer would pay for an item with known Product Identification and Category as well as having customer Information.

2 Introduction

A retail company ABC Private Limited wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month. The data set also contains customer demographics (age, gender, marital status, city type, stay in current city), product details (product id and product category) and Total purchase amount from last month.

Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

3 About the Dataset

Age : Treat as numerical. It presents age groups.

City Category : Convert this to numerical as well, with dummy variables. Observed frequency of the values.

Occupation : It has at least 16 different values.

Gender : There are two gender, can make them as binary.

Product ID : Should see if the string P means something and if there are other values.

Stay In Current City Years : Should deal with the + symbol.

Product Category 2 and **Product Category 3** : had NaN values.

Marital Status : should be treated as numerical. It presents whether the customer is married or not.

Purchase : represents the amount of purchase made by the customer.

User Id : represents the identification number for each customer. Each customer has a unique ID.

VARIABLE	**DEFINITION**
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins
Occupation	Occupation (Masked)
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belongs to other category also (Masked)
Product_Category_3	Product may belongs to other category also (Masked)
Purchase	Purchase Amount (Target Variable)

Table 1: DataSet Structure

4 Data Analysis

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

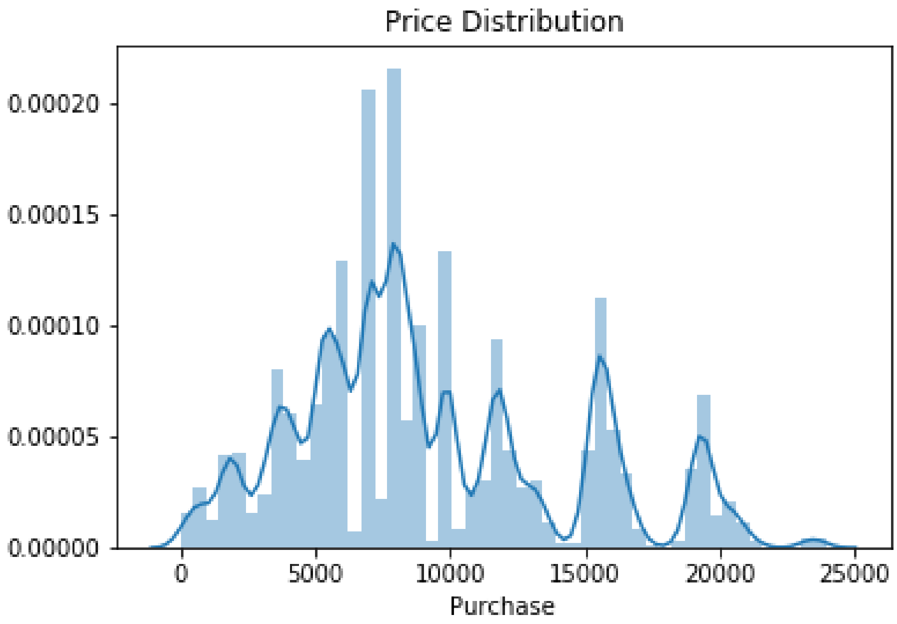

Figure 1: Figure Showing Distribution.

Here we can clearly see that the target variable i.e., purchase has an almost Gaussian Distribution.

Numerical Attributes : On observations we found out that User ID, Marital Status, Occupation, Product Category 1, Product Category 2, Product Category 3 and Purchase are our numerical predictors.

As seen in the beginning,**Occupation** has at least 20 different values. Since we do not known to each occupation each number corresponds, is difficult to make any analysis. Furthermore, it seems we have no alternative but to use since there is no way to reduce this number

Marital Status, as seen there are more single people than married people.

From the distribution for **Product Category 1**, it is clear that three products stand out, number 1, 5 and 8. Unfortunately, we do not know which product each number represents.

From the distribution for **Product Category 2**, it is clear that some products are few, number 1, 3, 7, 9, 10, 12 and 18. Unfortunately, we do not know which product each number represents.

From the distribution for **Product Category 3**, it is clear that some products are few, number 1, 2, 3, 4, 10 and 11. Unfortunately, we do not know which product each number represents.

From the **collinearity matrix representation** of the dataset we came to know that there seems to be no multicollinearity with our predictors which is a good thing, although there is some correlation among the product categories.

Looking for **Missing Values** :- The only predictors having missing value are Product Category 1 and Product Category 2.

5 Approach

We spent the first few hours just exploring the data, summarizing variables, plotting graphs, playing around with pivots and in parallel building base models (of course, XGBoost).

On the first attempt, we were able to go below par score with an optimized XGBoost model on raw features.

Usually ensembles performs well, but since we couldn't get any model close to the performance of XGB, so we decided to challenge myself to build a single powerful model. Which means, feature engineering.

6 Visualizations

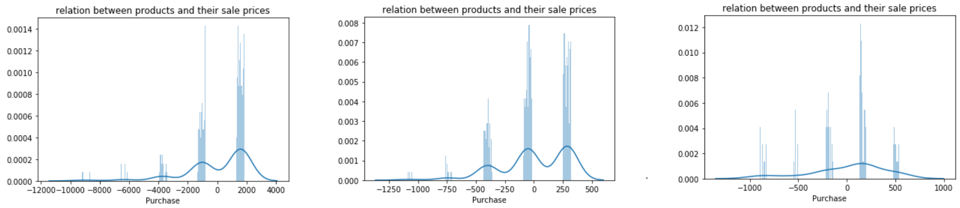

Figure 2: Figure Showing Distribution.

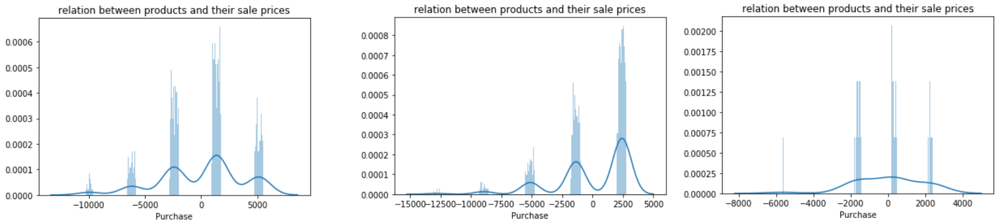

Figure 3: Figure Showing Distribution.

In our feature engineering, we generate a product count because as seen by the above few plots of product and their purchases, we concluded that the price follows a normal like distribution (biased by other attributes).

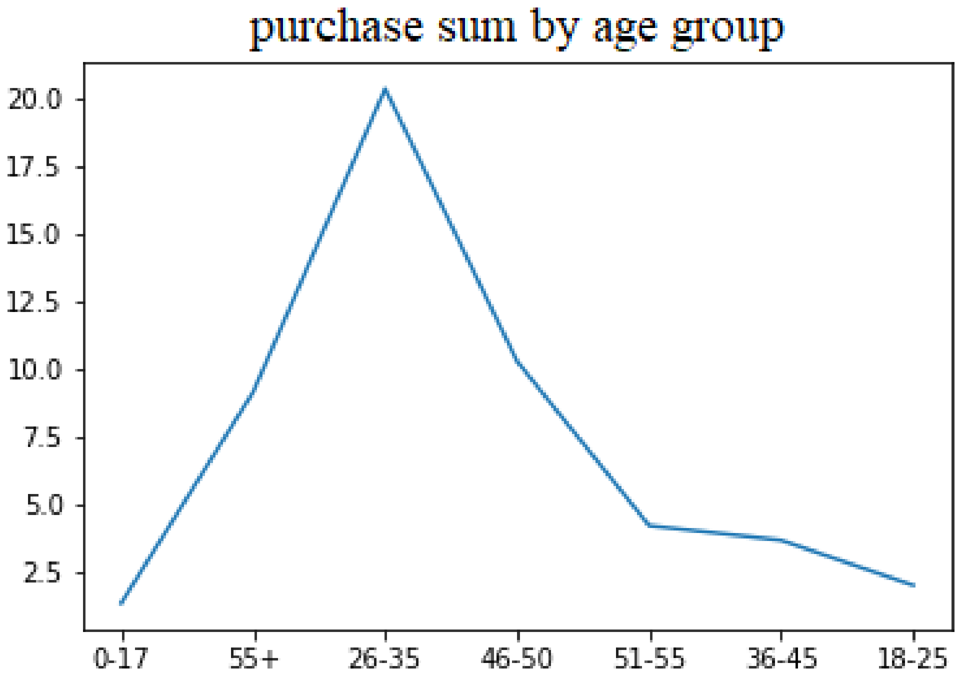

Figure 4: Figure Showing Distribution.

We see that the purchases depend on the age group and as expected max around 26-35.

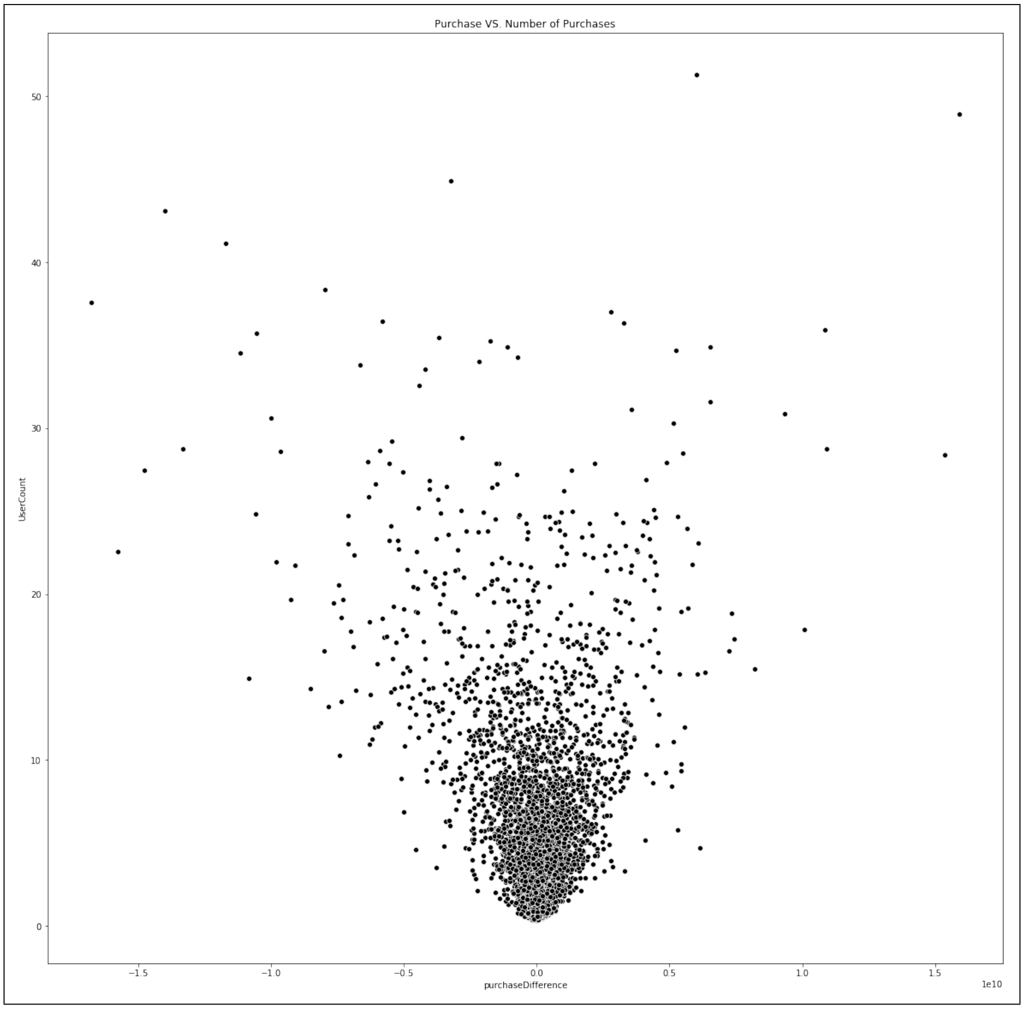

Figure 5: User Count And Price Scatter Plot.

This plot indicates the User count and the ratio of times when the user paid more than average

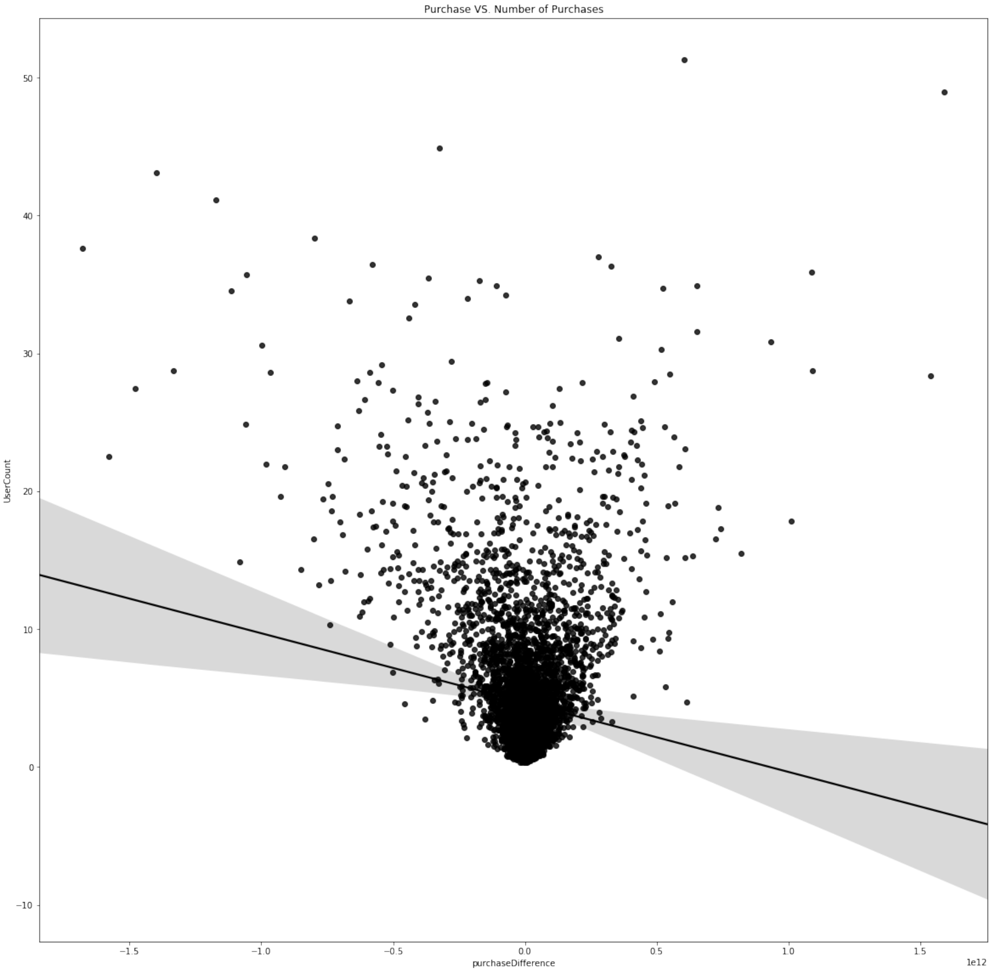

Figure 6: User Count And Price Reg Plot.

As the above scatterplot wasn't interpretable we used linear regression over it and the skewness of the plot clearly indicates co-relation.

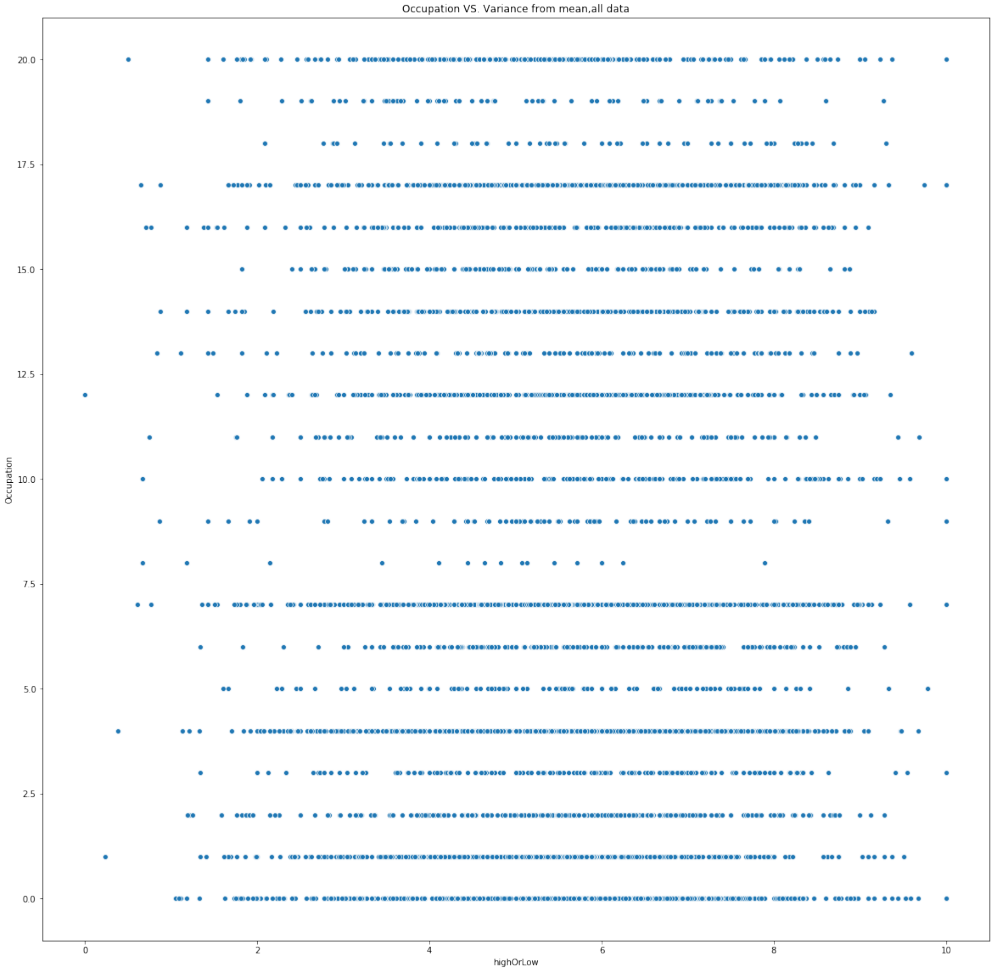

Figure 7: Occupation vs Variance From Mean Full Data.

To understand if a person paying high/low for some prices influences whether they will pay high/low in the future as well We tried to find using one of the important features corresponding to person, Occupation.

This plot is the average product price paid, greater than 0.5 indicating more price and less than 0.5 indicating less price, over the full data set.

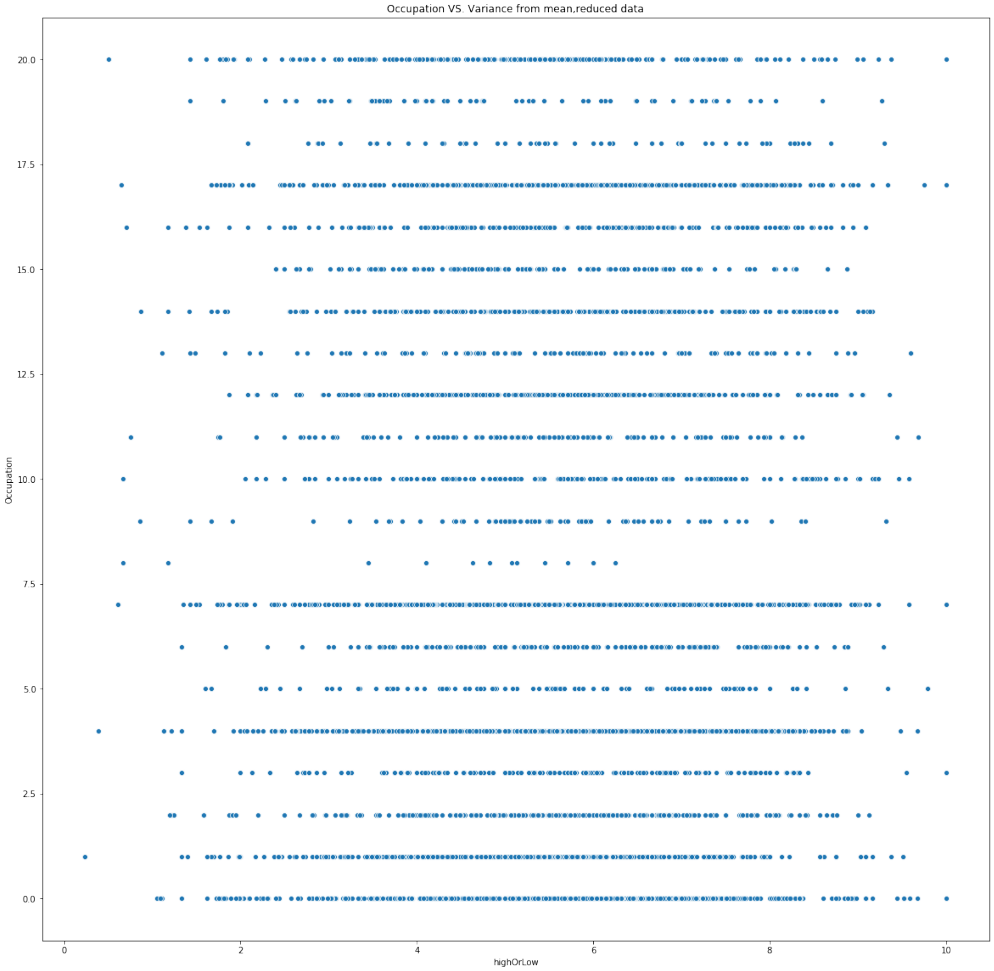

Figure 8: Occupation vs Variance From Mean Reduced Data.

We carried on the above to plot for a sample of dataset and as expected the plots were very similar, hence confirming our prediction of price being paid high/low being influenced by person.

7 Feature Engineering

1. **Age** group are replaced by their mean- just the labels doesn't give model the numerical emphasis of the buyer, while the average numeric value.

2. **User count** calculation- we predicted there would be a skewness of lower prices paid by buyers with higher count of purchases. As seen in the graph in visualization there was evidence of so.

3. Convert **gender** to binary data

4. **City-** We tried a level encoding and one-hot encoding , the latter gave better results

5. **Average Product Price-** We saw correlation between average purchase price of the product and the sale price Infact, average price is used as a central metric around which other factors modify the value to predict the price

6. **User ID**: Used as a raw feature

7. **Marital Status**: Used as raw feature

8. **Occupation**: Used as raw feature

9. **Stay In Current City**: Converted to numeric

10. **Product Category 1, 2, 3**: Used as raw feature, we also tried making a combined classifier of categories by addressing categories 1 through 20 as prime numbers and null values as 1. This didn't seem to work out as it might have introduced additional complexity.

11. **Product Count**: Number of observations of the product

12. **Product Mean**: Average purchase amount of product

13. **User High**: Proportion of times the user purchases products at a higher amount than the average purchase amount of the product

8 Model Building

We used XGBoost as our model with the following parameters :-

objective=”reg:linear” :- A learning task perimeter, corresponds to linear regression.

nrounds=500 :- Specifies the number of rounds.

max depth=10 :- Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit. 0 indicates no limit. Note that limit is required when grow policy is set of depthwise.

eta=0.1 :- Step size shrinkage used in update to prevents overfitting. After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative.

colsample bytree=0.5 :- Subsample ratio of columns when constructing each tree. Subsampling will occur once in every boosting iteration. range: (0,1]

seed=69 :- Random number seed.

metric=”rmse” :- Evaluation metrics for validation data, a default metric will be assigned according to objective (rmse for regression, and error for classification, mean average precision for ranking)

9 Results

We use root-mean-square error to measure the differences between values (sample or population values) predicted by our model and the values observed. The RMSE represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(d_i - f_i\Big)^2} \tag{1}$$

where n is the number of data points, d_i is the desired value output, f_i is the function output. RMSE is good to be used as a loss/error measure as the number of outliers in our data set are less.

We achieved a Validation score of 2425.75 while training on 80 % of the data and testing on the rest. We used the test data provided on Analytics Vidhya and got a score of 2488.79 and Public Leader-board Rank of 208.

References

[1] Black Friday Price Prediction, *https://datahack.analyticsvidhya.com/contest/black-friday/*