**Springboard—Data Science Career Track**

**Capstone Project 1**

**Final Report**

**Black Friday Sales Prediction**

**By Sheema Murugesh Babu**

**September 2019**

1. **Introduction**

A retail company ABC Private Limited wants to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high-volume products from the 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 offers for customers against different products.

* 1. **Problem Statement**

Setting the optimum price of a product is often a challenging problem for retailers, especially during a sale like the ‘Black Friday’. The challenge here is to predict purchase prices of various products purchased by customers based on historical purchase patterns. The data contains features like age, gender, marital status, categories of products purchased, city demographics etc. My solution looks into building machine learning models to estimate the sale price. Given the dataset, I could estimate the price a customer would pay for an item with known Product Identification and Category as well as having customer Information.

1. **Approach**
   1. **Data Acquisition and Wrangling**

**About the Dataset**

The data originally was found from the “Black Friday” dataset provided by Kaggle’s website. <https://www.kaggle.com/mehdidag/black-friday> and was saved into the local as ‘**BlackFriday.csv**’.

This is the current data they have available:

|  |  |
| --- | --- |
| **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 belong to another category (Also Masked) |
| Product\_Category\_3 | Product may belong to another category (Also Masked) |
| Purchase | Purchase Amount |

**Steps on Data-Wrangling**

First, I imported the required packages that I would need for Data Wrangling. Then, I used pandas read\_csv module to import the csv file (i.e. ‘BlackFirday.csv’) which was saved in my local repository. Upon executing the read\_csv function using the info() method, the csv file has 537577 entries and 12 columns with different formats of data. Also, there were many missing values in the Product\_Category\_2 and Product\_Category\_3. Digging deeper into it using the describe() method along with the head(), there were some null/NaN values in the product categories 2 and 3. Next step would be to find the distinct values followed by cleaning the null/NaN values.

### Second, I used nunique() method to find the count of distinct values in the dataset. All the columns except 'Purchase' column have categorical values and the 'Purchase' column would be considered as non-categorical. Replacing all the NaN's with 0. Also, changing the type of Product Category 2 and 3 from float to int type. All the NaN/Null values have been replaced with '0' and the dataset is now cleaned and ready for further exploration.

### For my convenience, I pickled the notebook so that it is easier to use in the future steps. The picked file is called ‘dw\_bf.pickle’.

* 1. **Storytelling and Inferential Statistics**

### After I wrangled and cleaned the dataset, I started to explore the data in detail. Looking at the data columns, we could begin to think what are the questions that could be answered.

### The questions I considered are as follows.

### Q) Who is more likely to spend more in a black Friday sale?

### 1) Men or Women.

### 2) Married or Un-Married individuals

### 3) Old Residents or new residents/visitors

### Q) Which type of products are more likely to be sold in a sale like black Friday?

### Q) Which type of products are more frequently purchased among men and women?

### The answers to the above questions can be found using some plots.

1. Plot of 'Gender' column:

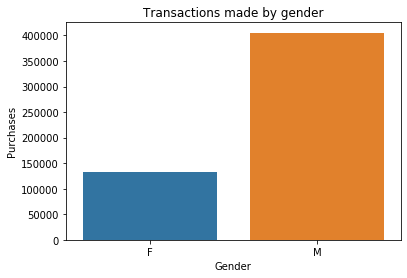


Figure: Transactions made by gender

#### Looking at above plot, it looks like a small number of females compared to males attended the Black Friday sale. But it could also mean a few number of females paid for the products and may be somebody else paid for them.

1. Plot of ‘Age’ column:

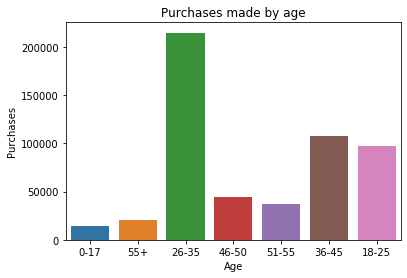


Figure: Purchases made by age

#### From the above plot for age, it seems like the largest age group is 26-35 years of age.

#### Further, we could also check among the age groups, which gender was a majority by adding a color code.

1. Money spent by Male/Female:

#### 

Figure: Money spent by Male/Female

As seen in the figure above, more males from all age groups shopped in the sale than the females.

We could check further, how many of the males from the age range 26-35 were married? For this I created a column called 'Combined\_G\_M' that represents ‘Gender’ concatenated with ‘Married\_Status’ column and then use it as hue. The plot of the 'Combined\_G\_M' column is as follows.

1. Transactions made by single vs married people:

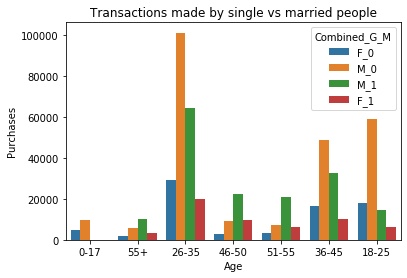
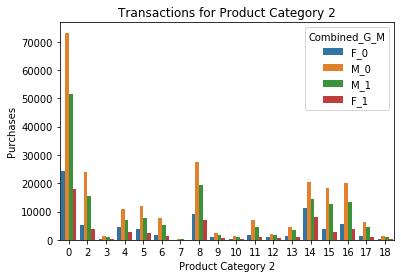
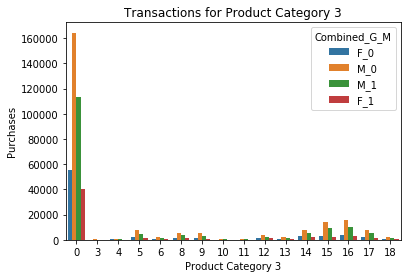


Figure: Transactions made by single vs married people

As we see above, there are no bars for married status (green or red) in the 0-17 age range which makes sense. And then if we look at the 46 and above age groups, unmarried females are very less. But on the other hand, married males paying in the age range 46-55 are also comparatively more than married females. So, it could also imply that though ladies do shop a lot, their spouses are possibly paying for it and hence data reflects that men shopped more. If we had more categorical data defining what kind of products were purchased by men, we could dig in this statement further. However, since in this dataset we don't know if there is a category that implies feminine products/clothes we cannot further explore this case.

1. Transactions for Product Category 2 and Product Category 3:



Figure: Transactions for Product Category 2 and Product Category 3

The above plots show that in both the Product Category 2 and 3, Product - 0 has been purchased the most by all people and unmarried males have the highest number of purchases in both the categories.

1. Transactions for the Stay in current city in years:

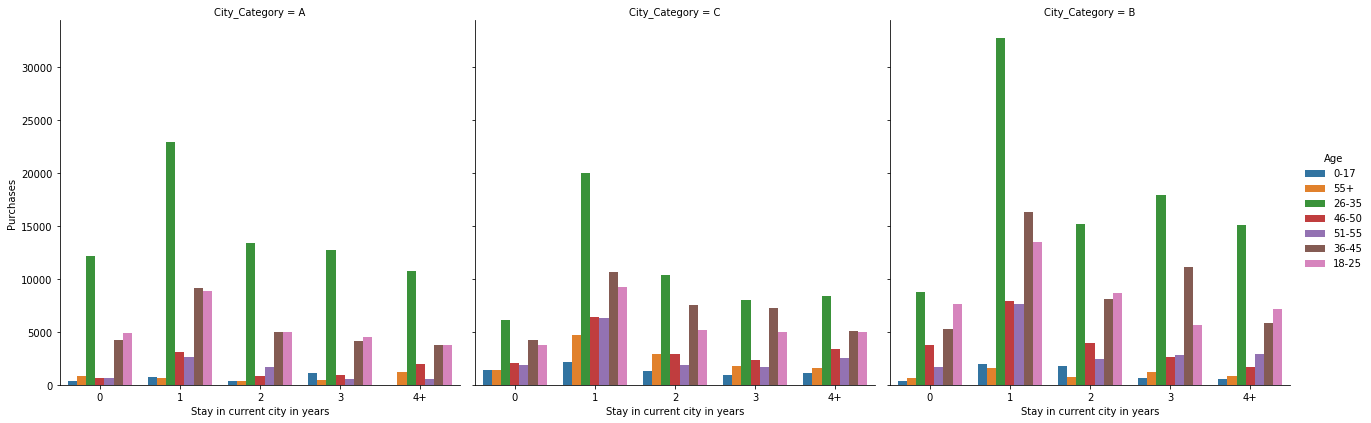


Figure: Transactions for the Stay in current city in years

The above plot shows us the number of years people with different age groups have stayed in a city with category A, B and C.

The following observations can be made:

1. In all the city categories we can see that most purchases are made by people who have stayed for 1 year in their chosen cities.
2. Age group of 26-35 have made most purchases in all the cities.

3) City B is the most popular city for any length of stay.

4) Old residents are the people who have stayed in a city for 4+ years and new residents/visitors are categorized below 1 years. The total transactions of Old and New residents/visitors are more in City B than City A and C.

1. Occupation breakdown:

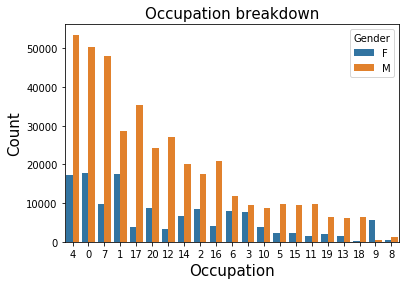


Figure: Occupation breakdown

Occupation 4 has the most people employed for both genders. There are more men working in most occupations except for occupation 9.

1. Purchases made in each product category:

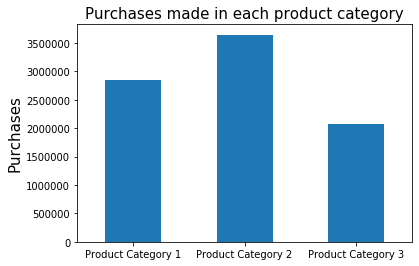


Figure: Purchases made in each product category

From the above figure we can see that Product category 2 has the most purchases.

The next step would be to provide steps on Inferential Statistics. I have performed Hypothesis testing using the one – way ANOVA tests.

**Hypothesis Test for Each Variable to Target Variable**

Interpretation:

The null hypothesis is that the means of the samples are equal. Rejecting the null hypothesis would imply that at least one of the means is different. The decision to reject the null hypothesis and accept the alternative hypothesis is based on the significance level of the test (α) and the probability of observing the effect given that the null hypothesis is true (p-value). If p-value ≤ α the null hypothesis is ruled out. We typically use a value of α=0.05, which corresponds to 95% confidence.

I used the One-way ANOVA test using statsmodels. Imported the ‘ols’ from the statsmodels.formula.api and computed the OLS for categorical and numerical variables. Here we are only interested in the F-statistic and corresponding p-value.

The categorical variables considered here are the columns, Gender, Age, City\_Category and Stay\_In\_Current\_City\_Years. After plotting the OLS for these variables we found out that the R-squared and the Adj. R-squared is considerably low i.e. 0.009, overall model was significant, p-value was less than 0.05 and hence we reject the null hypothesis. To better understand the variables, I have constructed an ANOVA table which gives us an overview of the variables. To test between groups, I performed some post-hoc testing where we can compare all groups against each other.

The overall model was significant, now to test which groups differ. There are a few different techniques that can be used like Fisher’s Least Significant Difference (LSD), Bonferroni correction, Tukey’s HSD etc. The technique here in consideration is the Tukey’s HSD.

In the Tukey’s HSD Post-Hoc Comparison, we take each separate column and compare it with the target variable. The Tukey HSD post-hoc comparison test controls type I error and maintains familywise error rate at 0.05 (FWER= 0.05 top of the table). Group1 and group2 columns are the groups being compared, the meandiff column is the difference in means of the two groups being calculated as group2 – group1, the lower/upper columns are the lower/upper boundaries of the 95% confidence interval, and the reject column states whether or not the null hypothesis should be rejected. Also, If the reject column says 'True' we reject the null hypothesis and the means are NOT equal, if the reject column says 'False' we accept the null hypothesis and the means are equal.

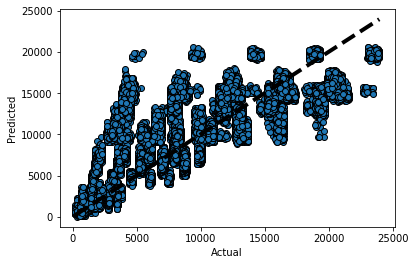
The numerical variables considered here are the columns, Occupation, Marital\_Status, Product\_Category\_1, Product\_Category\_2 and Product\_Category\_3. Here again we found that the R-squared and the Adj. R-squared is high i.e. 0.130, overall model was significant, p-value was less than 0.05 and hence we reject the null hypothesis. I followed the same steps here which I conducted for the categorical variables.

In the next section(s), I compared different machine learning models with each other to build the best performing model to predict the black Friday sales. I performed Linear Regression, Ridge, Lasso Regression, Random Forest Regression as well as some feature analysis.

* 1. **Baseline Modelling**

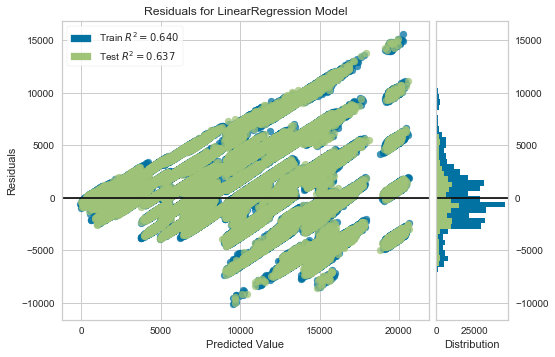
I built a Linear Regression model as a base model for my analysis. First, I dropped the extra column ‘Combined\_G\_M’ which was added in the previous steps. I used one-hot encoding using the get\_dummies method of pandas to create dummy features for all the categorical data. Then, I imported all the necessary packages that I would be using for the machine learning analysis. Then, I divided the dataset into input ‘X’ and label ‘y’. The variable ‘X’ contains all columns except the target variable and the variable ‘y’ contains only the target variable. By the help or using input ‘X’, we will find out, predict, or classify the label ‘y’. Then, I divided the X and y into training and testing data set. The training set is 75% of our total data and training set is 25% of the data.

Next step would be to instantiate a Linear Regression object. Here again I, fitted the model on the training set, predicted on the test set as well as the training set. I calculated different scoring metrics to measure the performance of my model. I stored each score to compare which method was the best. I mainly used R2 score to measure the performance of our model. Higher the R2 score, better would be our model. I then plotted a graph of the predicted versus the actual plot as shown in the figure below.

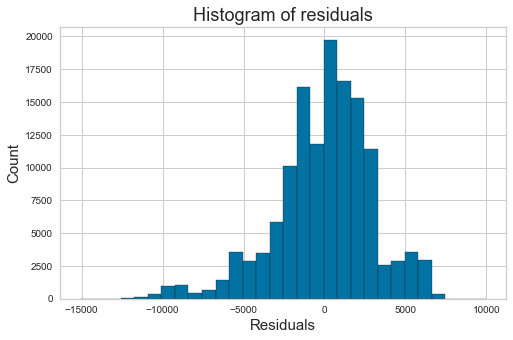


Next step would to plot a graph of predicted values vs the residuals. Residuals are calculated using the formula,

residual\_i = y\_target\_i - predicted\_i. I used the yellowbrick regressor package for this. This method gave me a graph of the predicted value vs the residuals as well as a histogram of residuals. The figure below gives us a detailed information.



In the next step, I plotted a separate histogram of residuals plot to get a better idea. I calculated the residuals manually using the above-mentioned formula. The figure is as below,



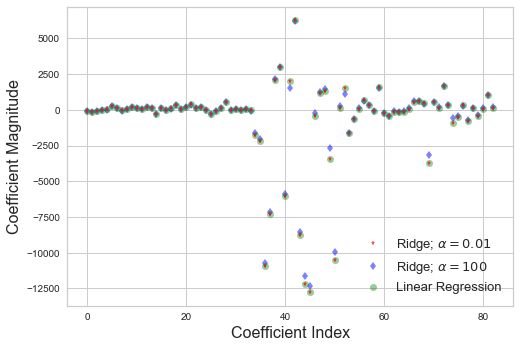
We can see that the above plot shows us a normal distribution. The Linear regression’s R2 score for the test set was 63.69%.

Just to be sure that our above model is free of overfitting and underfitting, I performed the Ridge and Lasso Regression.

* 1. **Extended Modelling**

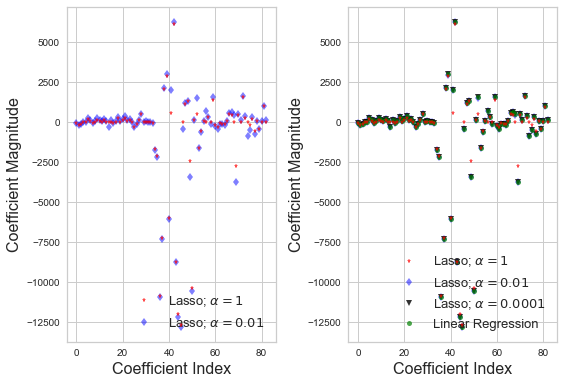
This section includes all the other machine learning models. The first point of discussion would be the Ridge Regression. Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent overfitting which may result from simple linear regression.

First, we import all the necessary packages. Here there are two alphas i.e. low and high are considered. Higher the alpha value, more restriction on the coefficients, low alpha means more generalization. The same are to be followed as we did in the Linear Regression Model. Instantiated a Ridge Regression object. Fitted the model on the training set for both the low alpha (i.e. 0.01) and high alpha (i.e. 100). Calculated the score for train and test set for each alpha and stored it in a variable. Next I have compared all the scores of Ridge Regression (both the alphas) and the linear regression score. The observed results were that the train and test scores for both the alphas were almost similar to the linear regression model. The Ridge Regressor did not make much of a difference. Linear regression is so far the appropriate model. Next I, plotted a graph of coefficient index vs the coefficient magnitude using the coefficient attribute (i.e coef\_) as in the figure below.



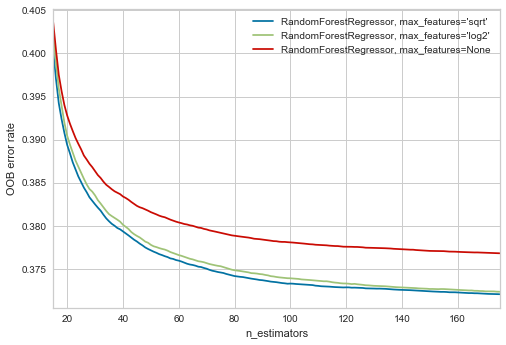
A detailed interpretation of the figure above. In X axis we plot the coefficient index. For low value of α (0.01), when the coefficients are less restricted, the magnitudes of the coefficients are almost the same as of linear regression. For higher value of α (100), we see that for coefficient indices of approx. 38,40,43,45 and so on, the magnitudes are considerably less compared to linear regression case. This is an example of shrinking coefficient magnitude using Ridge regression.

The second point of discussion is the Lasso Regression. Similar steps are to be followed here just like we performed in the Ridge Regression. Imported all the necessary packages, fitted on the train set, calculated the score and the coefficients for train and test set for each alpha (i.e. 1, 0.01 and 0.0001) and stored it in a variable. Next I, Plotted the graphs for each of the coefficients as shown below,

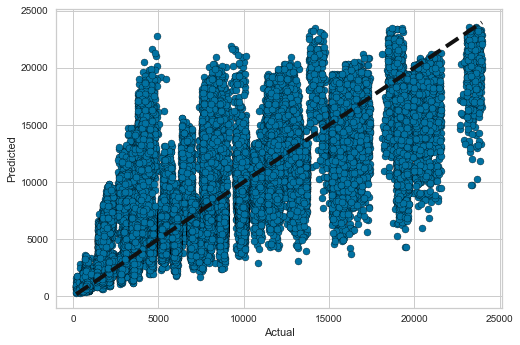


Understanding the plot above, the default value of regularization parameter in Lasso regression (given by α) is 1. The observed results were that both the training and test scores are low; we can conclude that the model is under-fitting the Black Friday dataset. Let’s try reducing this under-fitting by reducing alpha and increasing number of iterations. Now α = 0.01, training and test score increases. Comparison of coefficient magnitude for two different values of alpha are shown in the left panel of figure 1. For alpha =1, we can see most of the coefficients are zero or nearly zero, which is not the case for alpha=0.01. Further reduce α =0.0001, training and test scores are similar to basic linear regression case. In the right panel of figure, for α = 0.0001, coefficients for Lasso regression and linear regression show close resemblance.

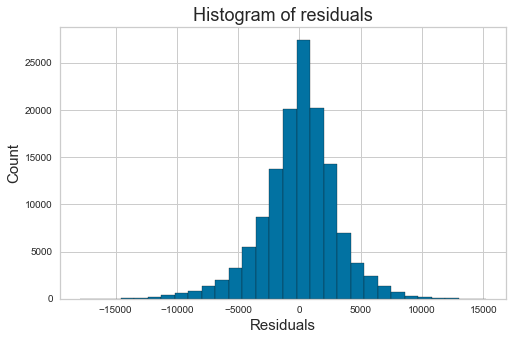
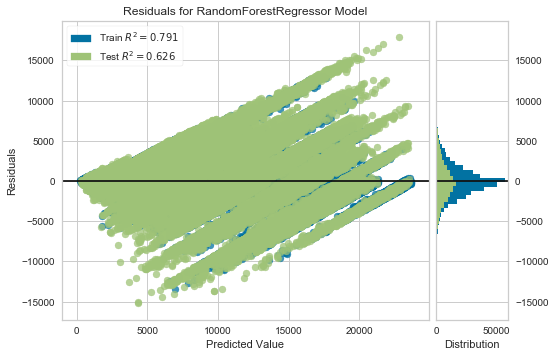
The third point of discussion is the Random Forest Regression. We start by importing all the necessary packages. To find the best max feature and the n-estimators, I have plotted a plot of OOB error rate vs n-estimators. Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests and other machine learning models utilizing bootstrap aggregating (bagging) to sub-sample data samples used for training. Here again we instantiate a Random Forest Regressor and fit it on the training set and calculate the OOB score. The resultant plot is as shown below.



Looking at the above graph, we could set the max\_features = ‘sqrt’ and the n\_estimators after 160, seems to be stable. So, let's set the n\_estimators=160 using the set\_params method. Fitted it again on the train set, predicted on the test set and calculated different scoring metrics to measure the performance of the model. I then plotted a graph of the predicted versus the actual plot as shown in the figure below.



The next steps were similar to the ones performed in the Linear Regression model. I plotted a graph of predicted values vs the residuals. Here again, I used the yellowbrick regressor package for this, also plotted a separate histogram of residuals for better understanding. The figures below give us a detailed information of the two.



We can see that the above plots show us a normal distribution. The Random Forest Regression’s R2 score for the test set was 62.64% which is less than the Linear Regression’s score of 63.69%.

1. **Findings**

So far, we saw how to predict the Black Friday Sales using different machine learning algorithms. It's time to summarize our findings. We calculated different scoring metrics to measure the performance of the model.

I used the Root mean squared error (RMSE), the Mean Absolute Percentage Error and R2 score scoring metrics. Theroot-mean-square error (RMSE) (or sometimes root-mean-squared error) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The mean absolute percentage error (MAPE) is a measure of prediction accuracy of a forecasting method in statistics, for example in [trend estimation](https://en.wikipedia.org/wiki/Trend_estimation), also used as a loss function for regression problems in machine learning. The R2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R2 of 1 indicates that the regression predictions perfectly fit the data.

Firstly, I put up a Linear regression model as my base model, performed all the necessary steps. I have considered mainly the RMSE and the R2 scores. The Root mean squared error for the Testing Set is 3005.25 and R2 score for the Testing Set is 63.69%.

Secondly, I performed Ridge and Lasso Regression with all the necessary steps. Ridge regression for the testing set with low alpha is 63.69% and with high alpha is 63.67%. Likewise, Lasso Regression for the testing set with alpha=1 is 63.65%, alpha=0.01 is 63.69% and with alpha=0.0001 is 63.69%.

At last, I performed the Random Forest Regression. The Root Mean squared error for test set is 3048.62 and the R2 score for test set is 62.64%.

|  |  |  |
| --- | --- | --- |
| **Algorithms** | **Root mean squared error** | **R2 score (Test Set)** |
| Linear Regression | 3005.25 | 63.69% |
| Ridge Regression | NIL | Low alpha = 63.69%  High alpha = 63.67% |
| Lasso Regression | NIL | alpha=1 - 63.65% alpha=0.01 - 63.69% alpha=0.0001 - 63.69% |
| Random Forest Regression | 3048.62 | 62.64% |

Table: Results summarized

1. **Conclusions and Future Work**

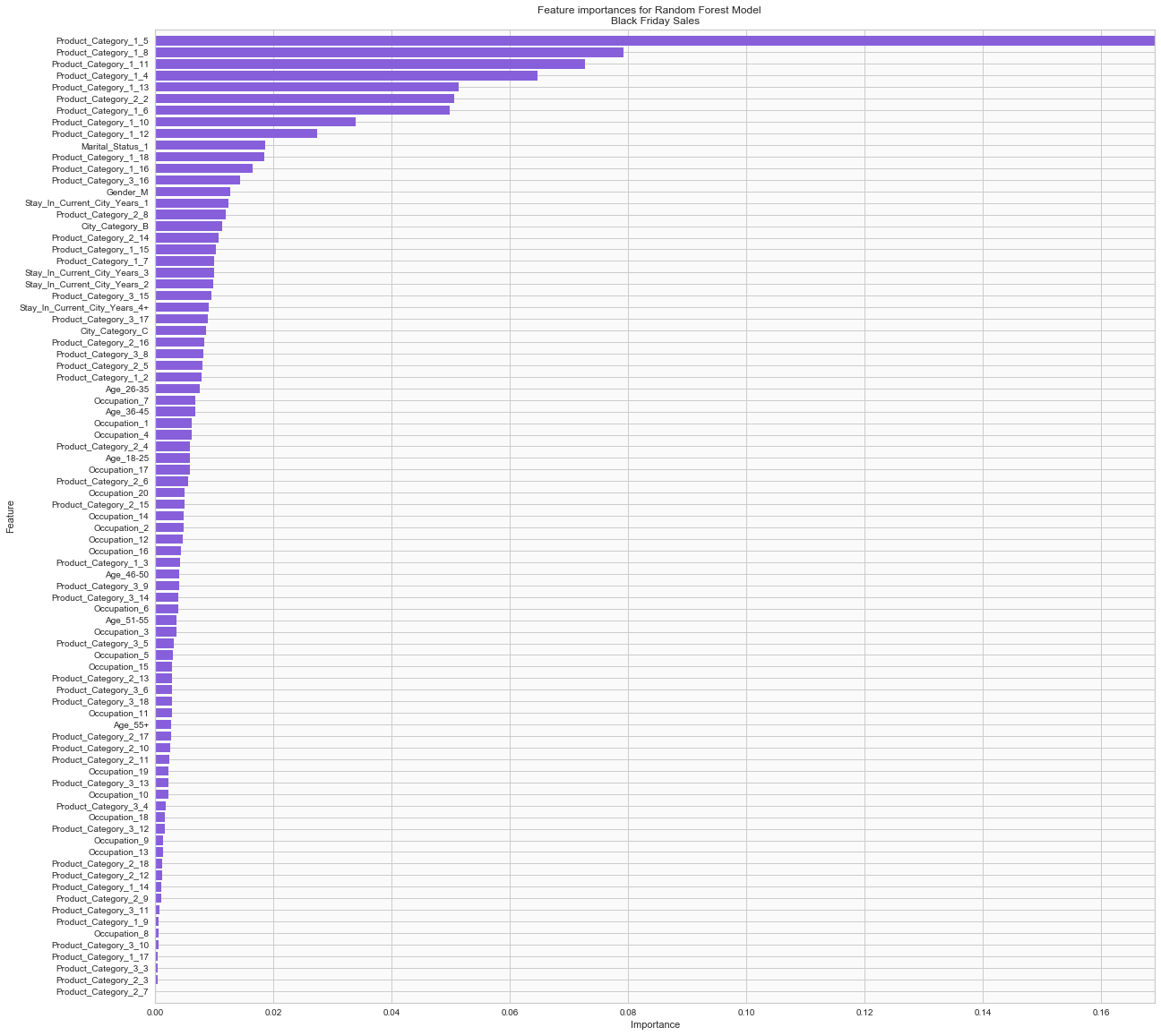
In the last section we jotted down the R2 scores of all the four machine learning models. We can conclude that Linear Regression Model performed better.

Many different approaches than the ones above can also be used in the Black Friday sales prediction. Going forward I would like to try out some different algorithms. One such algorithm which come to my mind is the XGBoost model.

1. **Recommendation for the Client**

In this section, I performed some feature analysis, to provide an idea for the clients on what product(s)/product category were purchased by the customers on Black Friday.

This is in continuation with the Random Forest Model, I used the feature\_importances\_ attribute and defined a function. The result of the function gave me a list of features (with their names) and their importance. Higher the importance value, illustrates that the product(s)/product category was purchased more at the Black Friday Sales by the customers. The figure below shows the feature importance plot for the Random Forest Model.



In the figure above we can see that the ‘Product\_Category\_1\_5’ has higher importance. ‘Product\_Category\_1\_8’, ‘Product\_Category\_1\_11’, ‘Product\_Category\_1\_4’ could also be recommended.