Black Friday Sales Prediction

Abstract: - The main aim was to use machine learning to help the company have a greater insight into the purchases of their customers to offer them better deals and understand their customers better. To solve this problem, I have done analysis on the fields provided to extract results and for understanding the data. Further, to predict the purchase amount of the customers I used prediction models like Linear regression, Random Forest Regression and XGBoost Regression.

I. Introduction

Black Friday in USA is celebrated by many companies in US by offering exclusive deals to the customers. Therefore, by considering the past sales data of a company. I am trying to solve the problem of a company trying to boost their sales by offering various deals to the customers. The prediction can help the company to give personalized offers to the customers based on their tendencies and purchases.

This problem is important as it will help the company to know which products are helping their customers which products they need to work on. Also, analyzing the data will help to understand their customers purchase behavior which can help them to adjust their production and distribution cycles accordingly.

To better understand the problem and data. I performed several analyses on the fields to extract results. The info() function showed that two fields in the data have large number of missing values. After addressing this problem by replacing the missing value by mode of the column. I also drew multiple count plots, box plots and distribution plots to get insights from the data. To solve the problem of predicting I used regression models like Linear regression, Random Forest and XGBoost. Further, I introduced two new fields to increase the accuracy of prediction.

II. Data Analysis and Modeling

A. Data Analysis

Some of the data insights that I obtained were results I anticipated like the age range of the customers did not have much effect on the purchase amount of the customer. I also saw that the Purchase was evenly distributed and was not biased at any one side. Some surprising results were Male's spending more compared to Females. Also, the count plots helped to see which kind of customer's visit the store the most. The customers having occupation 4 visit the store the most. Further, analysis also shows that customers of city category C spend more compared to A and B.

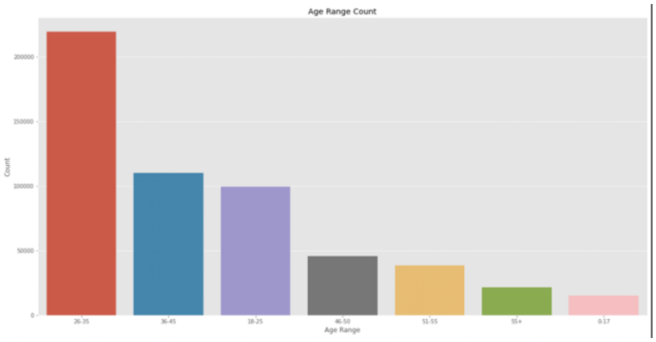


Figure 1 This figure shows the count plot of age range which tells us that age group 26-35 visits the store most.

The methods I used to explore the data include info() method to see the null values, countplot() to see the customer frequency, boxplot() to see relation of fields with purchase amount, distplot() to see distributrion of purchase amount of the customers.

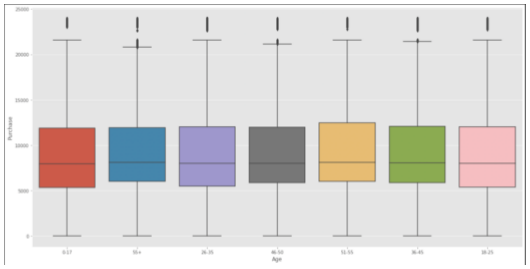


Figure 2 The figure shows box plot of Age vs Purchase graph that tells us age is not an important factor to predict purchase amount

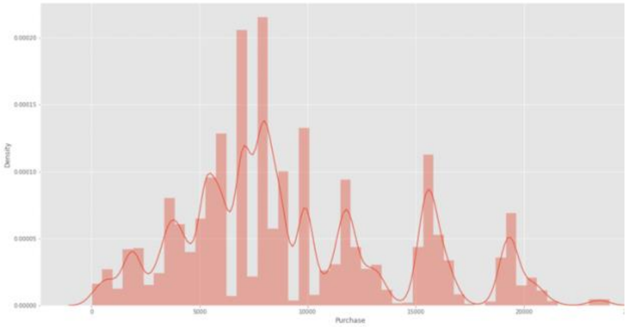


Figure 3 This figure shows the distribution plot of the Purchase amount variable

Further, steps included Label encoding of categorical variables which helps to increase the prediction accuracy of the model. Next step was to scale the data as different category can range in large area which makes prediction less accurate. Therefore, I used the standard scaler to scale all the fields in the range of -4 to 4. Lastly, I introduced two new features to help with the prediction. The two features calculated average purchase per product and average purchase per user. On seeing the correlation graph these two features were highly correlated with the Purchase amount.

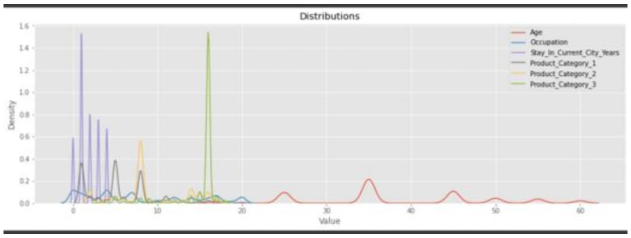


Figure 4 The figure shows the categorical data before scaling

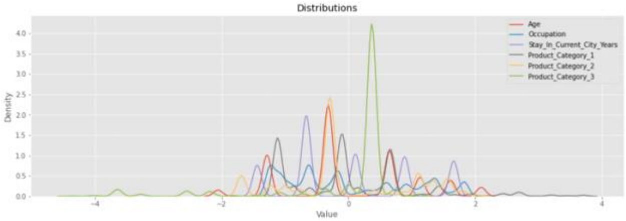


Figure 5 The figure shows categorical data after scaling has been applied

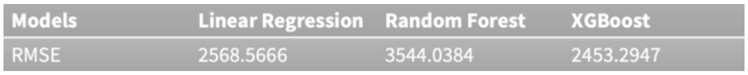
B. Model

I have mainly used three algorithms Linear Regression, Random Forest and XGBoost. Firstly, the basic concept behind linear regression is that it tries to fit a linear line between the n-dimensional space formed by the fields to predict the target variable. This is one of the most basic algorithms which creates a line which best fits the data points in the n-dimensional space and predicts the target value from that line. This algorithm can be used in when we want to predict a continuous changing variable.

Secondly, Random Forest is an extension of the Decision Tree where a decision tree is made with the fields and the tree is traversed based in the value of the fields. Random Forest on the other hand selects the field at random and forms several random trees and it combines the prediction from all the random trees and then predicts a result out of all those values.

Thirdly, XGBoost known as Extreme Gradient Boosting algorithm is a further extension of the random forest algorithm in which gradient descent is performed to get more accurate results.

To validate the model the data was split into train and test set with 80% of data being used for training and 20% for testing. The purchase amount values were predicted on the test set and then the validity of the results was calculated with the help of root mean square error value. XGBoost algorithm gave the smallest RMSE value.



Models	Linear Regression	Random Forest	XGBoost
RMSE	2568.5666	3544.0384	2453.2947

Figure 6 This figure shows the RMSE values obtained by different models on the test set

III. Conclusions

The problem was to predict purchase amount based on the data given which contained customer demographics, product categories, etc. The next step included handling the missing data which I replaced with the mode of the column. TO analyze the data, I plotted various count plots, box plots and distribution plots. Further to improve accuracy I did label encoding and scaled the categorical data. Next, I introduced two new highly correlated variables with purchase amount. After that I did the prediction using Linear regression, Random Forest and XGBoost and calculated the error by calculating the root mean square value.

References:

[1] S. Mohanapriya, S. Mohana Saranya. International Journal of Advanced Science and Technology, Vol. 29, No. 3s, (2020), pp. 1049-105

[2] Purvika Bajaj, Renesa Ray, Shivani Shedge, Shravani Vidhate, Prof. Dr. Nikhilkumar Shardoor, International Research Journal of Engineering and Technology (IRJET), Vol. 07, Issue: 06. June 2020

[3] Karandeep Singh, Booma P M, Umapathy Eaganathan, Journal of Physics: Conference Series, Vol. 1712 (2020)