Black Friday Sales Prediction

Akshay Gupta

Computer Science Department, George Washington University

akshay0712@gwu.edu

Abstract — To find the best machine learning model that predicts the purchase label accurately which means lowest root mean square error and best accuracy. Out of 11 feature columns only product category 1 gives the best solution and in order to get the better accuracy I believe we need the relationship between product category 1, product category 2 and product category 3. Since, Product category 2 has 31.06% and product category 3 has 69.44% missing values respectively. We need more data values in these columns to predict a better model.

***Keywords*— Machine learning, Root mean square error, Preprocessing, one hot encoding, cross validation**

I. INTRODUCTION

The dataset used for this problem is a sample of the transactions made in a retail store during black Friday sale. And, this store wants to learn more about the customer behavior based on the three different product categories. This a regression problem where we are trying to predict the dependent value (the amount of purchase) with 11 set of feature columns. We have a dataset of 55000 customers from black Friday sale. The dataset contains different kind of variables which can be either numerical or categorical.

II. PROBLEM STATEMENT

The objective of this machine learning project is to build an accurate regression predictive model which clearly finds the price label given a set of features.

III. RELATED WORK

This dataset has been analyzed for data research. For example, who is more likely to spend more on black Friday sale. Men or Women? Married or Unmarried? Old Resident or new Residents? The relationship of different product category in accordance to men or women.

IV. DATA DESCRIPTION

As described by the author, "The dataset is comprised of 550,000 observations about Black Friday shoppers in a retail store, it contains different kinds of variables either numerical or categorical. It contains missing values." (Mehdi Dagdoug). This dataset is taken from the Kaggle.

Following are the columns for this Data Set with some characteristics:

- **Purchase (label to predict):** Purchase amount in dollars.
- **User_ID:** has duplicate values hence very limited unique values to be related.
- **Product_ID:** No key(which can relate products to product category) given- so we can relate items with product ID
- **Gender:** Sex of shopper is categorical
- **Age:** requires to be converted from ranges to single digit and then string to numeric for one hot encoding
- **Occupation:** Occupation of shopper has no key and is categorical too.
- **City_Category:** Residence location of shopper doesn't have key either.
- **Stay_In_Current_City_Years:** Number of years stay in current city.
- **Marital_Status**: Marital status of shopper no key either.
- **Product_Category_1:** Product category of purchase is categorical.
- **Product_Category_2:** Product may belong is categorical.
- **Product_Category_3:** Product may belong is categorical

V. DATA PRE-PROCESSING

Product category 2 and product category 3 has 166986 and 373299 missing values, respectively. So, these are required to be replaced with zero. We need to drop label feature before our model training and also labels: User_ID and Product_ID has they have very limited unique values and hence cannot form a relevant relationship between product categories. Labels like age, occupation and marital status require conversion from string to numeric.

Finally, our dataset requires one hot encoding of labels; gender, city category, product category 1, product category 2, and product category 3 So that our machine learning model can understand categorical values correctly.

VI. APPROACH - Building a model for Black Friday Sales Prediction

As our label purchase is a set of continuous values, we will use a regression model for this problem. There are several models that can be used to solve this problem, But I have taken Linear Regressor, Decision Tree Regressor, Gradient Boosting Regressor, Ridge Regressor to analyze the results. It is important to compare the various aspects such as accuracy, confusion matrix, root mean square error to decide which model would to best for tour problem.

A lower rmse would mean a better fit but before find rmse we need to remove as much outliers as possible from our data set and for that we would use cross validation. Here, I will be using K-fold cross validation with 10 splits

VII. EXPERIMENTS

After model evaluation following values were gathered for each of the model.

A. Linear Regressor

Train Accuracy – 62.78%

Test Accuracy – 62.59%

RMSE = 3047.85

B. Decision Tree Regressor

Train Accuracy – 79.8276%

Test Accuracy – 55.27%

RMSE = 3332.54

C. Gradient Boosting Regressor

Train Accuracy – 63.46%

Test Accuracy – 63.24%

RMSE = 3021.32

D. Ridge Regressor

Train Accuracy –62.26%

Test Accuracy – 62.49%

RMSE = 3051.98

E. Ridge Regressor with K-Fold cross validation

Train Accuracy –64.06%

Test Accuracy – 63.89%

RMSE = 2980.37

VIII. DATASET VISUALIZATION

A. Male vs Female purchase

From the particular store we can see that males are buying 75.4% more than the females.

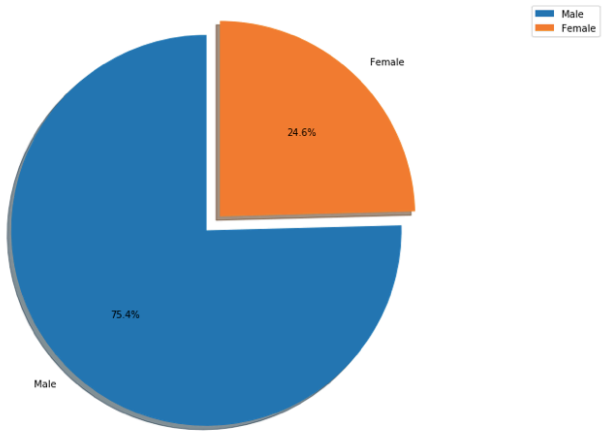

Fig. 1 Goods purchased by male vs female

B. Purchase based on city category with corresponding age group

Out of three cities we can see that city B has the highest purchase.

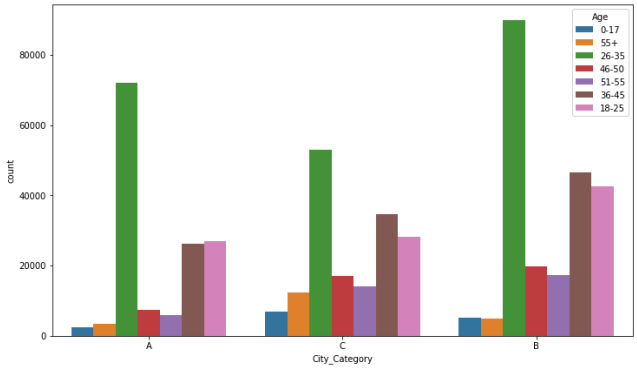

Fig. 2 City Category vs Count

C. Mean purchases per age group

We can see that in each of the city age ranging from 26-35 is buying the most.

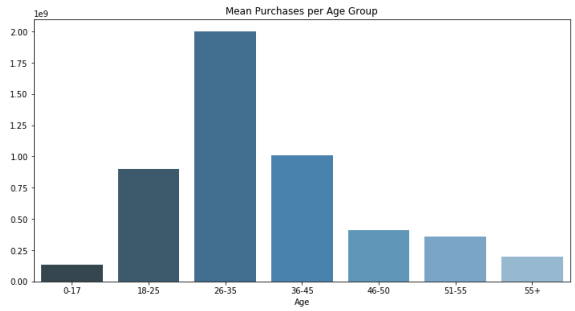

Fig. 3 Mean purchases per age group

D. Purchases based on job category

It is clear that category 7, 0 and 4 are the biggest buyers from all the 20 categories of employees.

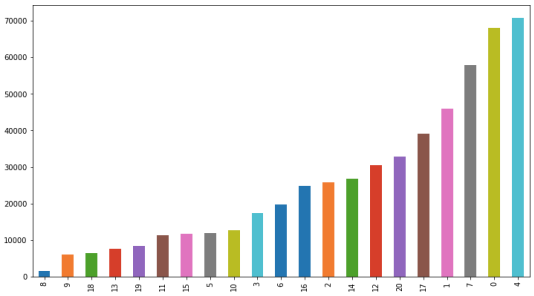

Fig. 3 number of purchases vs Job category

D. Feature Importance

We can that the prediction results have decreased less than 1% since we have 70 % and 30% missing values in product category 3 and 2 respectively. Had there been more values the overall performance for purchase prediction would have increased drastically.

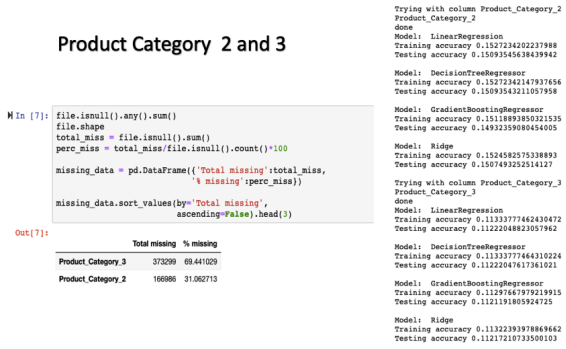

Fig. 5 Relevance of product category 2 and product category3 respectively

IX. DISCUSSION

The first model being used here is linear regressor which the most widely used linear regression model. But it has many flaws which includes multicollinearity, autocorrelation and has limited relationship between feature. Hence it would be advisable to use other regression model that find better relationship inside feature set. So linear regressor would give optimal results only incase where data has linear shape.

In order to remove the flaw of multicollinearity and be able to work on non-linear dataset. We would choose ridge regressor which takes lambda parameter performs slightly better than linear regressor. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision tree.

The intuition behind gradient boosting algorithm is to repetitively leverage the patterns in residuals and strengthen a model with weak predictions and make it better. One can use gradient boost regressor which literally boosts the optimization algorithm to work against anomalies like outliers, but it is sensitive to overfitting.

Since Decision trees implicitly perform variable screening on feature selection the top splits define the most crucial features of the dataset. And, the non-linearity in the dataset does not affect the performance of the model. It can give good results although at the cost of overfitting. This arise when the split in decision tree precisely to data set which later effect the results when someone tries to test the model. Which is evident in our model evaluation that the accuracy of the model has increased by 17% at the expense of high overfitting which is descriptive through rmse.

We can see through the Decision Tree Regressor model evaluation that the testing accuracy has gone up to 80% at the cost of overfitting as the testing accuracy reduces to 55 %.

```
model_DTR = DecisionTreeRegressor()
X_train, X_test, y_train, y_test = preprocess(X_original, Y)
model_DTR.fit(X_train, y_train)

mse = mean_squared_error(y_test, model_DTR.predict(X_test))
rmse = np.sqrt(mse)
predicted_label = model_DTR.predict(X_test)

print("predicted labels =",predicted_label)
print("RootMeanSquareError =",rmse)
model_DTR.score(X_train, y_train), model_DTR.score(X_test, y_test)
```

```
 predicted labels = [12882.51351351 15886.33333333  6767.40909091 ...  6396.48484848
  16552.38461538  8831.        ]
 RootMeanSquareError = 3332.547243042518

(0.7982761011654916, 0.5527890827739447)
```

Fig. 6 predicted labels, rmse, accuracy (training, test) for Decision Tree Regressor

Also, from the model testing for each model when the features very tested individually the accuracy of the model came out to be less than 1 % for all the features except product categories which tells that none of the feature performs well individually to predict.

Moreover, the missing value in product category does not help the model figure out its dependency with another category. Instead it should have values that relate it with other product category. The missing keys in user and product id also affects the performance of the model since they do not form relationship with products as such.

X. GOING FORWARD

The work done so far tells us a lot about the important features in the dataset and the need to find the relationship between every feature better. So that the overall performance can be reduced, simultaneously decreasing the overfitting that may require incorporating more sale data or experimenting with other useful machine learning model.

XI. CONCLUSION

Hence, we can say that Product category_1 is the fundamental feature that is helping our model to predict accurately – 0.798(best). If we could have got more values in Product category_2/_3 or relationship between categories our prediction would have been better. Plus, as it is evident through testing that all features such as age, gender, occupation, city in are not very helpful in predicting accuracy maybe taken together or alone. For better prediction: we require better relationship between features, better models and addition of more values in dataset would be required.

REFERENCES

[1] https://www.kaggle.com/mehdidag/black-friday/