Black Friday Project

Student: Dogan Can Demirbilek

1 Problem Statement

In retail stores, sales are highly changeable due to several reasons like discounts and promotions.

Black Friday is a day which retail stores apply significant discount on items. Discount and promotion rates differ according to product characteristics and user characteristics (generally, prime or elite customers have more discount). For example, laptops might have more discount than cell phones.

The goal of project is building a machine learning model that is able to predict the amount of purchase (in dollars) per transaction given user and product features such as age, occupation, product category etc. The data used is a sales record of an unknown retail store.

Predicting amount of purchase is crucial to have insights about revenue and profit. Knowing these concepts will be helpful to take strategic decisions by management. For instances, financial department, campaign management and customer loyalty department can take benefit from these predictions.

2 Proposed Solution

The problem can be identified as a supervised learning regression problem with numerical and categorical features.

Firstly, it is important to determine which features are related with target variable (purchase). By doing this, unrelated features will be eliminated, which may lead the model to wrong direction. This will be done by interpreting results of base-line model by using permutation importance which can be defined the decrease in a model score when a single feature value is randomly shuffled [1] and by using partial dependency plots which can be defined as the marginal effect one or two features have on the predicted outcome of a machine learning model [2].

Often the hardest part of solving a machine learning problem can be finding the right estimator for the job. For this problem, Decision Tree will be used as a base-line model because it is suitable for both categorical and numeric features moreover, it is easy to interpret and fast to learn.

After interpretation of base-line model, Random Forest, Bagging and Boosting, will be used because in general they have higher accuracy and they are more flexible than decision trees.

Comparison of these four models will be done by using test root mean squared error (RMSE) because it has the benefit of penalizing large errors [3] and it is wanted to penalize error for each prediction by squaring it. This will be done by using k-fold cross validation, data set will be splitted into 5 (k=5) equal slices then for each split, model will be learned all but k-th split will be left as an unseen data, then RMSE will be computed for k-th slice, lastly test RMSE will be averaged. Also mean absolute error (MAE, average test error) R-squared (proportions of variance for target variable that’s explained by other features) will be calculated with same method since they are easy to interpret and they show different aspects of models.

3 Experimental Evaluation

3.1 Data Description

Data has 537577 transactions and 12 features which are User_ID, Product_ID, Gender, Age (Age in bins), Occupation, City_Category, Stay_In_Current_City_Years, Marital_Status, Product_Category_1 (Cloths), Product_Category_2 (Electronics), Product_Category_3 (Home Goods), Purchase (Purchase amount in dollars).

User_ID and Product_ID are unique identifier therefore they were deleted.

To understand rest of it, descriptive data analyses were made and summary results are presented in table 3.1.

According to these results, Product_Category_3 has almost 69% missing value, it wouldn’t be statistically significant to fill these missing values by using remaining 31% so it was deleted. Product_Category_2 has almost 31% missing value, this amount can be filled by using remaining 69% (indeed most frequent observation was used to fill missing values) since 69% of these feature is a sufficient proportion to effect target variable,.

Apart from these three features (User_ID, Product_ID, Product_Category_3), rest of them will be considered for base-line model. Each unique observation from these features may increase or decrease the purchase predictions. For example men may tend to purchase more than women or 26-35 age bin may be more active or some occupation might purchase more due to high income.

Feature Name	Numerical/ Categorical	Missing Value Percentage	Number of Unique Element	Mean / Most Frequent
Gender	Categorical	-	2	M (Male)
Age (in bins)	Categorical	-	7	26-35
Occupation	Categorical	-	21	4
City_Category	Categorical	-	3	B
Stay_In_Current_City_Years	Numerical	-	5	1.86
Marital_Status	Categorical	-	2	0
Product_Category_1	Categorical	**-**	18	5
Product_Category_2	Categorical	31%	17	8
Product_Category_3	Categorical	69%	15	16
Purchase	Numerical	-	-	9334

Table 3.1: Summary Descriptive Analysis

3.2 Experimental Procedure

After descriptive analyses, train and test data were split by using k-fold (k=5) cross validation then base-line model was fitted. For base-line model, test RMSE is 3146.

It was necessary to understand which features are most useful. Hence permutation importance of features were found and plotted in Figure 3.2.

The concept is measuring the importance of a feature by calculating the increase in the model’s prediction error after permuting the feature. A feature is “important” if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction. A feature is “unimportant” if shuffling its values leaves the model error unchanged, because in this case the model ignored the feature for the prediction [4].

After analyzing permutation importance, to understand how features affect the model, partial dependency plot were used for the most important 6 features in Figure 3.3.

A partial dependence plot can show whether the relationship between the target and a feature is linear, monotonic or more complex. For example, when applied to a linear regression model, partial dependence plots always show a linear relationship [2]. Y axis of the plots shows the expected value for target variable which is purchase. X axis shows the values of features.

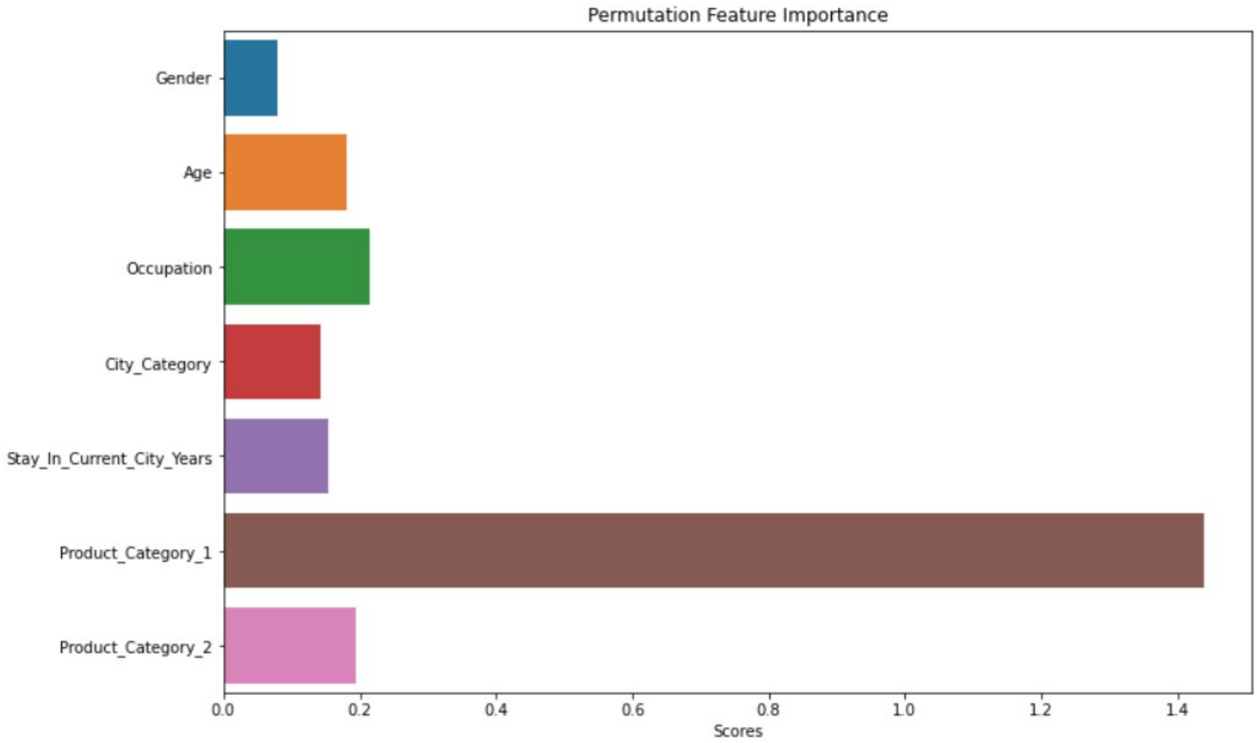

Figure 3.2: Permutation Importance

Higher the permutation importance score is more important the feature for base-line model. Product_Category_1, Occupation, Age, Product_Category_2 and Stay_In_Current_City_Years are the most important features for our base-line model. On the other hand, Gender, Marital_Status and City_Category has no significant impact on base-line model.

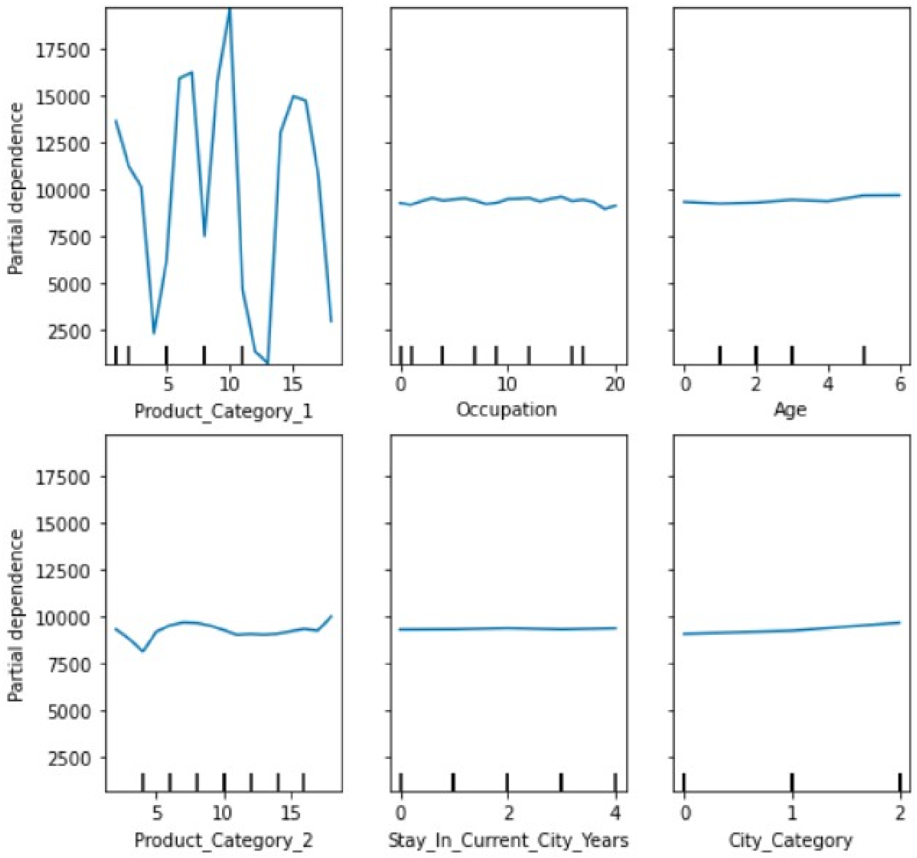

Figure 3.3: Partial Dependency Plot

It is important to point that some sort of product categories has significant impact on target variable. For instances, category 10 in Product_Category_1 feature has a increasing impact on predictions. Category 4 in Product_Category_1 has reverse impact on predictions.

In conclusion, the final predictors are: Product_Category_1, Occupation, Age, Product_Category_2, Stay_In_Current_City_Years and City_Category. Because they have higher permutation importance.

3.3 Results

Models are compared in the table provided below. From the table, it can be seen that all models have almost same R-squared value. Nearly, 63% variability can be explained by models. Random Forest and Bagging almost same for all metrics. Although, MAE of boosting is higher than other models, it has better RMSE.

Model	RMSE	MAE	R-squared
Base-line	3146	2287	0.6
Random Forest	3035	2236	0.63
Bagging	3059	2249	0.62
Boosting	3022	2307	0.63

Table 3.2: Model vs Test Metrics

Taking into account of all performance metrics and considering complexity and performance of models, boosting algorithm seem to perform best.

Nevertheless, performance of the best model isn’t satisfactory because average of target variable is 9333 so RMSE values are high for this average. It can be understand that data is not informative enough to provide better predictions on purchase.

4 References

1: Sklearn Community, Permutation feature importance, 2020,
https://scikit-learn.org/stable/modules/permutation_importance.html
2: Christoph Molnar, Partial Dependence Plot (PDP), 2020,
https://christophm.github.io/interpretable-ml-book/pdp.html
3: Radwa Elshawi, Mouaz H. Al-Mallah, Sherif Sakr, On the interpretability of machine learning-based model for predicting hypertension, 2019
4: Christoph Molnar, Permutation Feature Importance, 2020,
https://christophm.github.io/interpretable-ml-book/feature-importance.html