Black Friday Sales Purchase Prediction

Naomi Elvin Machado machado.n@husky.neu.edu
Suhani Ladani ladani.s@husky.neu.edu
Naga Sai Anirudh Upadhyayula upadhyayula.n@husky.neu.edu

Abstract

Black Friday Sales Purchase Prediction is the task of predicting the purchase amount of customers during the annual Black Friday sale based on different factors including but not limited to Gender, Age, Occupation, City Category, Duration of Stay in the Current City, Marital Status, and Product Categories. In the world of internet, e-commerce has become one of the mainstays of trade and millions of people across the world buy and sell items of everyday use on the internet. Hence, it becomes very important to analyze the data of millions of customers, possibly hundreds of billions of transactions on a daily basis, build models and establish trends. In this report, we have documented the procedure we followed to predict the purchase amount of a customer from the given dataset. The first step was Data Understanding, followed by Data Preprocessing, which included handling missing values and Feature Engineering. We added a few features like User Score, Product Score, and Category Count. User Score is the total number of purchases of a customer when compared to the maximum number of purchases of a customer. Similarly, Product Score is the number of times a product was bought when compared to the maximum number of times a particular product was bought. These features are calculated relative to the values present in the dataset. Category Count is a cumulative score of how many categories were present in each transaction completed by a customer. These three new features turned out to be very useful in predicting the purchase amount as they had a high correlation with the purchase amount. We have attempted to predict the purchase amounts of customers using both linear regression techniques and decision tree based regression techniques. The linear regression models we studied are Linear Regression, Ridge Regression, Lasso Regression, and Elastic Net Regression. The decision tree based regressors we learned are Random Forest Regression, Extra Trees Regression, Extreme Gradient Boosting. We used modeled our data using H2O, an open source, in-memory, distributed, fast and scalable learning and predictive analytics platform that allows building machine learning models, which generated admirable results. The final step was to compare the results obtained from all the aforementioned models and rank them accordingly. We concluded by highlighting the model which works best for the given dataset followed by future work, areas which can be improved and other interesting observations.

1. Introduction

1.1 Background Study and Literature Survey

There has been an annual average increase of 2.4% every year since the year 2002. In 2008, the year of financial crisis, there was a steep decrease of 4.6%. It is interesting how the economic patterns affect the shopping habit of the people. According to an article on the balance “On average, shoppers expect to spend $1,007.24 each. Of that, they'll spend $637.67 on gifts. Another $215.04 will go for food, decorations, flowers, and greeting cards. They'll also spend $154.53 to take advantage of the seasonal deals and promotions.”

The analysis not only helps the stores in luring the customers; but the customers could benefit by analyzing the discounts over the years to calculate the best time to buy the toys, electronics, clothes, etc. Moreover, the employment rate increases over the month of November and December, benefitting the overall economy.

“Consumer spending makes up about 70 percent of the gross domestic product, and a solid chunk of it takes place in November and December, mainly in the form of gift purchases. A fifth of all retail sales occurs in the year’s last two months, according to the National Retail Federation. ”

1.2 Why is it important to solve this problem?

In the world of internet, e-commerce has become one of the mainstays of trade and millions of people across the world buy and sell items of everyday use on the internet. Hence, it becomes very important to analyze the data of millions of customers, possibly hundreds of billions of transactions on a daily basis, build models and establish trends. The conclusions drawn from this analysis can help the e-commerce companies and other stakeholders involved understand its customer base and come up with better ways of serving them better.

By predicting the amount a customer might spend during Black Friday Sale, companies could predict which products would be sold the most, products based on the geographical locations and by identifying these trends, they can come up with better offers for sets of customers. At the end of the day, since profits are a major factor for companies to strive and thrive, the above trends can prove to be a gamechanger and may lead to added sales and revenue, making it a win-win situation for everyone.

1.3 Data

This dataset has been taken from AnalyticsVidhya.com, where it is an ongoing competition. The dataset has 13 features and there are about 5.5 lakh rows of customer data. Many of the customer details like User ID, Current City, Product Category 1, Product Category 2 and Product Category 3 are masked for obvious reasons. The target variable is the Purchase Amount. The test dataset contains about 2 lakh rows with the same features, except the target variable.

2. Methodology

In this section, we document the steps we followed to predict the target variable, the purchase amount of customers.

2.1 Data Understanding

2.1.1 About the Dataset

This dataset consists of 11 features as described in the table below. The dataset also has missing values observed in Product Category 2 and Product Category 3. The target variable for this dataset is the purchase amount of the customer.

Features of the Original Dataset:

Feature	**Description**
User_ID	The ID of the customer
Product_ID	The ID of the product purchased by the customer
Gender	Customer's gender
Age	Customer's age
Occupation	Customer's occupation
City	Customer's current city
Duration of stay in the city	Duration of the customer's stay in the current city

Marital Status	Customer's marital status
Product Categories(1,2,3)	Category of the product
Purchase(Target variable)	Purchase amount of the customer

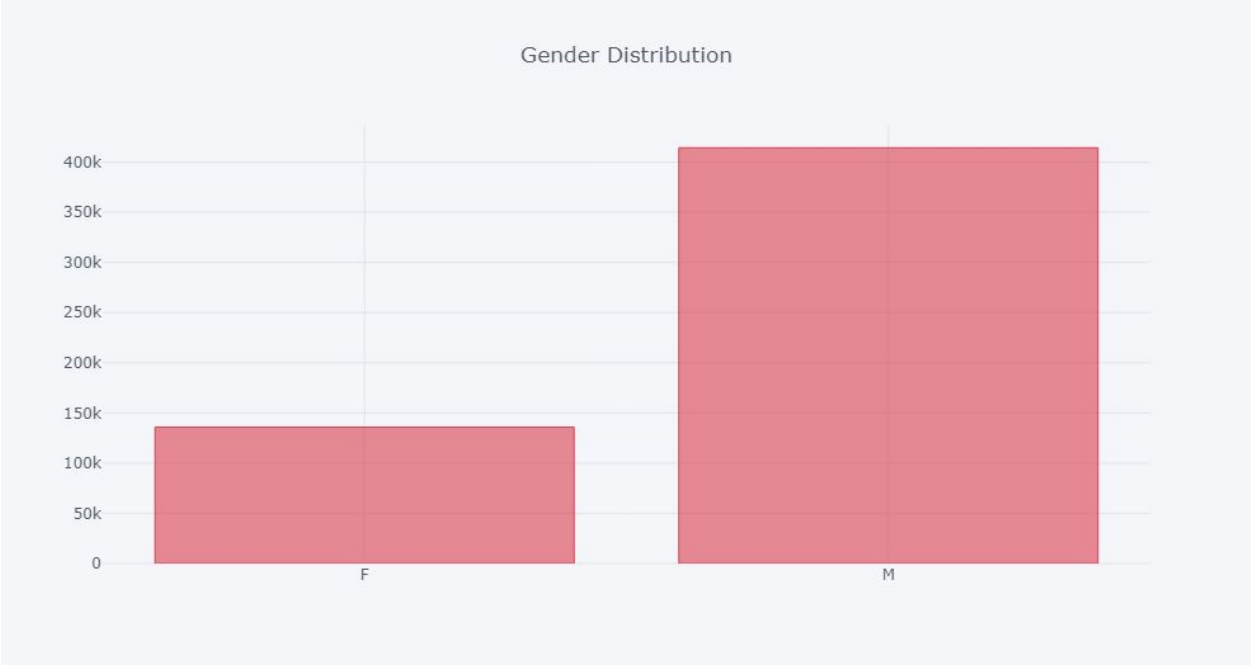
2.1.2 Exploratory Data Analysis

We explored the dataset to observe the different feature values and understand the data better.

Distribution of the dataset based on the feature:

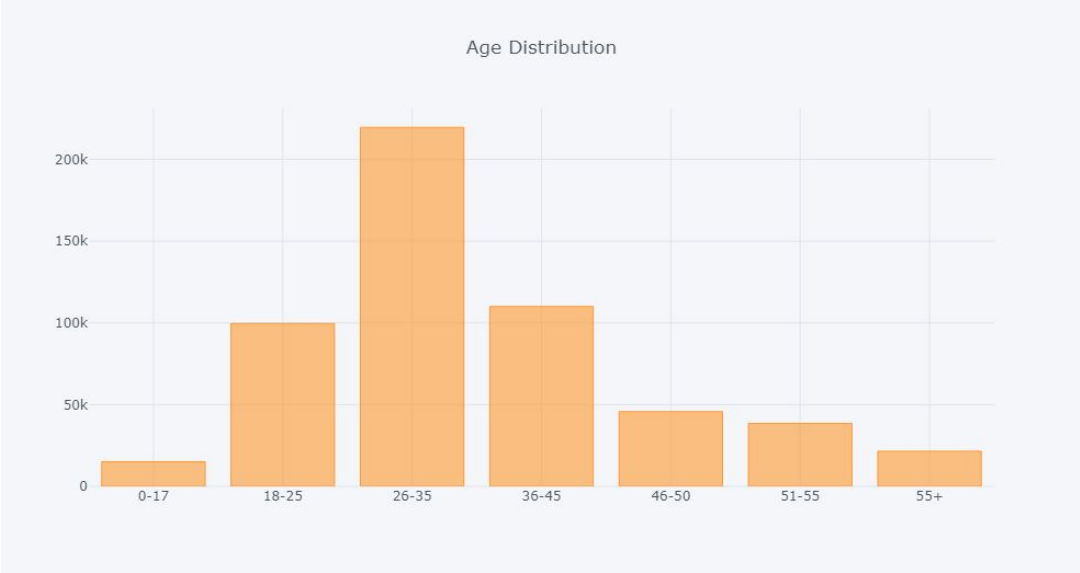
1. **Gender Distribution of the data:**

   As observed in the plot, a majority of the buyer were males during Black Friday.



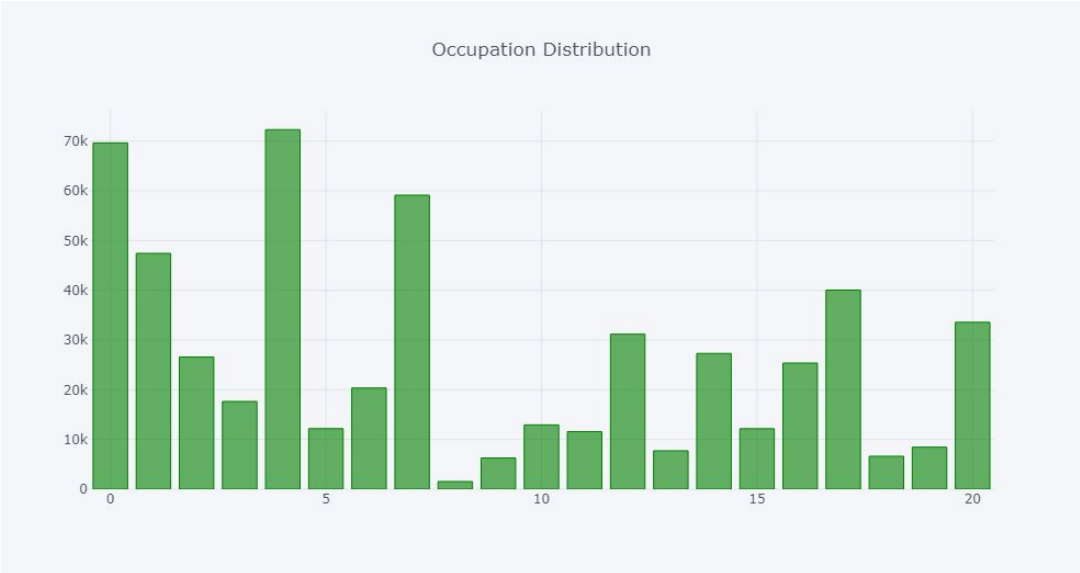
2. Age distribution of the data:

For this dataset, Age was divided into 7 distinct range categories as observed in the plot. The majority of the buyers were in the age group of 26-35.



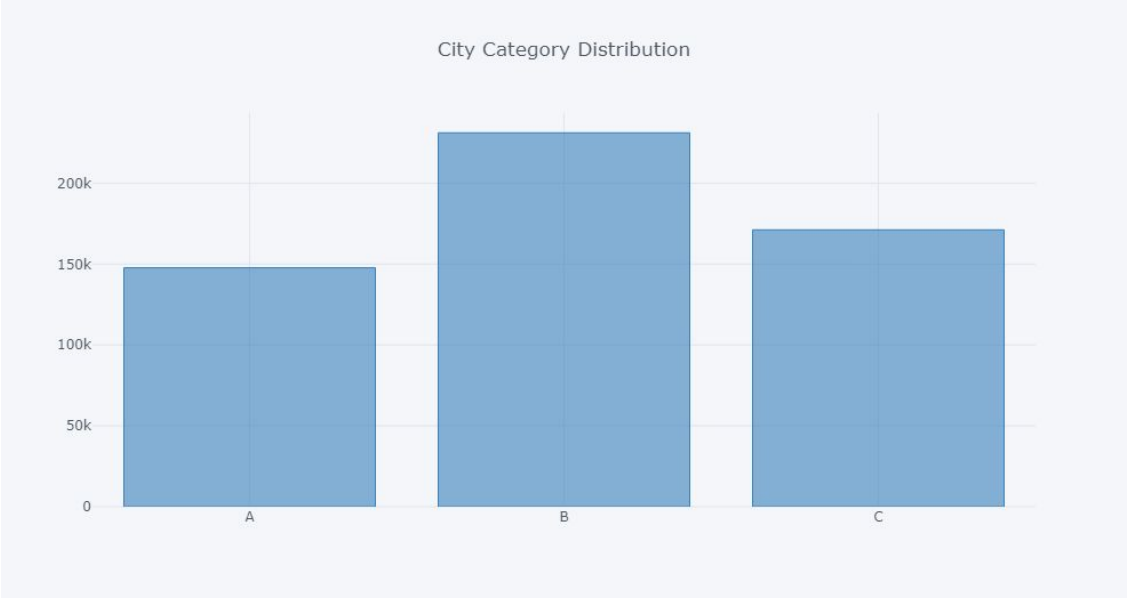
3. Occupation Distribution of data:

There are 20 masked occupation categories in this data set, with a majority of the customers belonging to categories 0, 4 and 7.



4. City Distribution of data:

There are 3 masked city categories in this data set, with a majority of the customers belonging to category B.



5. Stay In Current City(Years) Distribution of data:

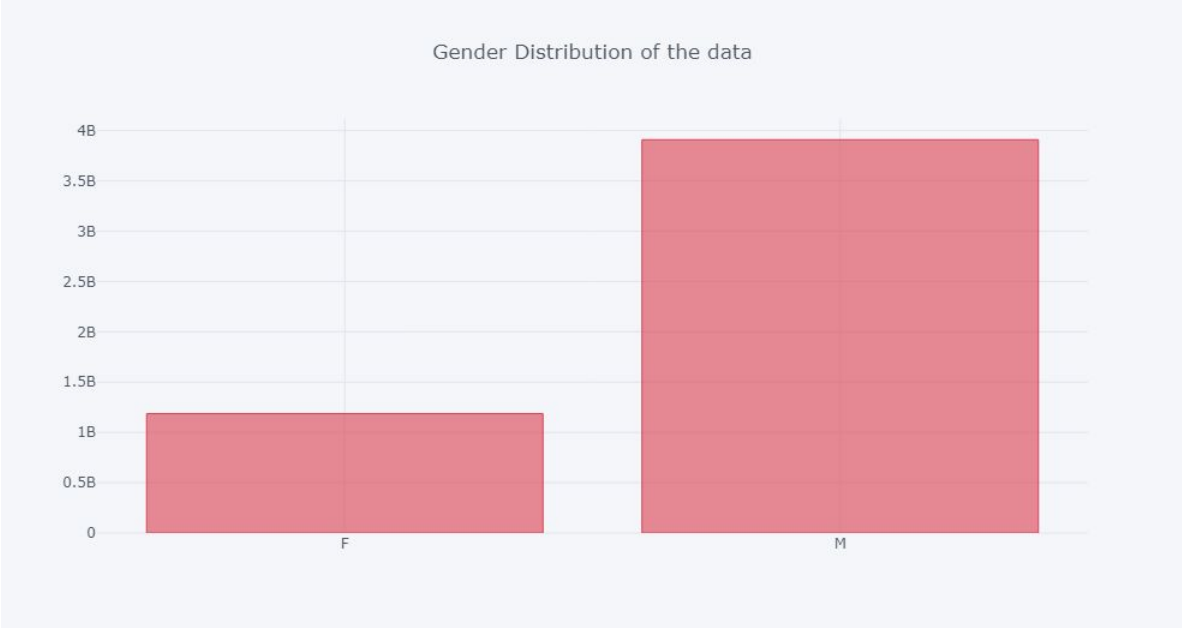
The dataset also has a feature describing the number of years a customer has lived in the city.

6. Marital Status Distribution of data:

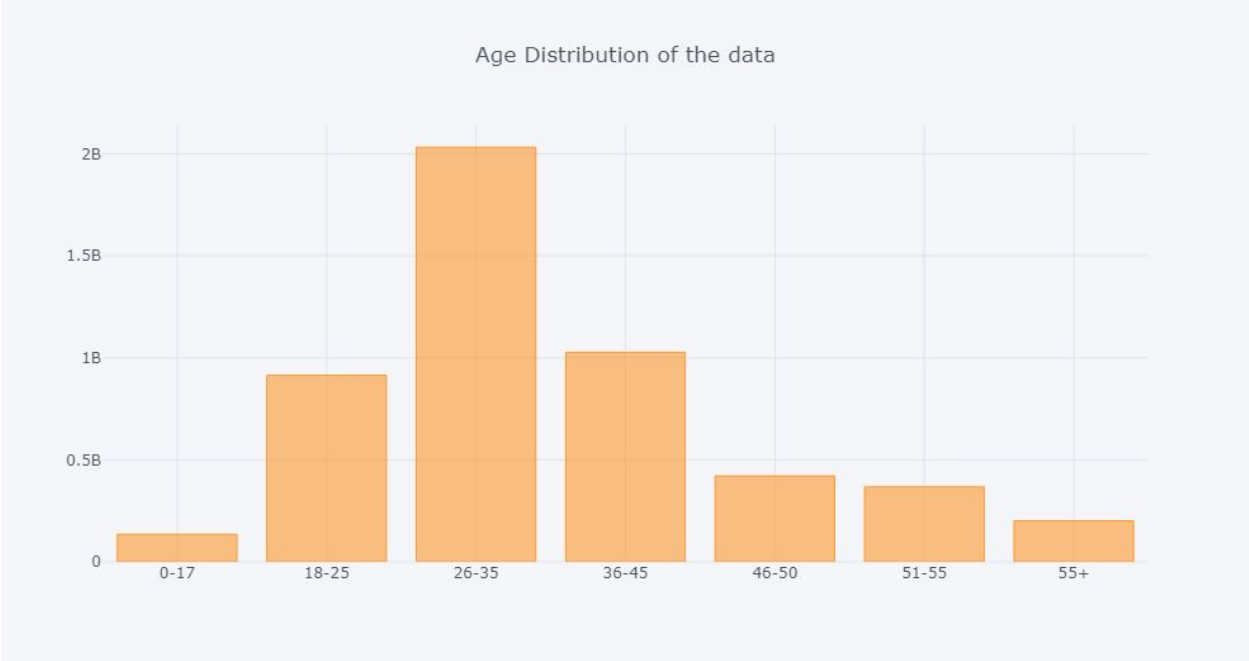
The dataset also has a distinguishing feature stating the marital status of the customer.

Purchase Amounts observed based on dataset feature:

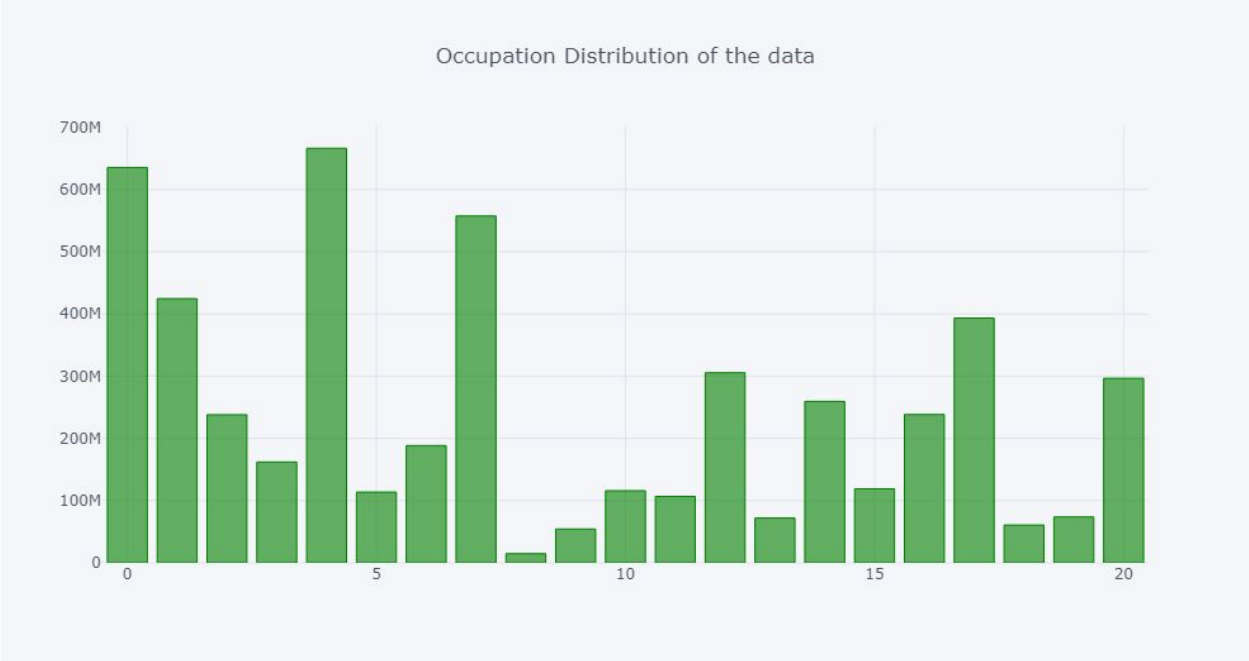
1. Purchase amounts by Gender



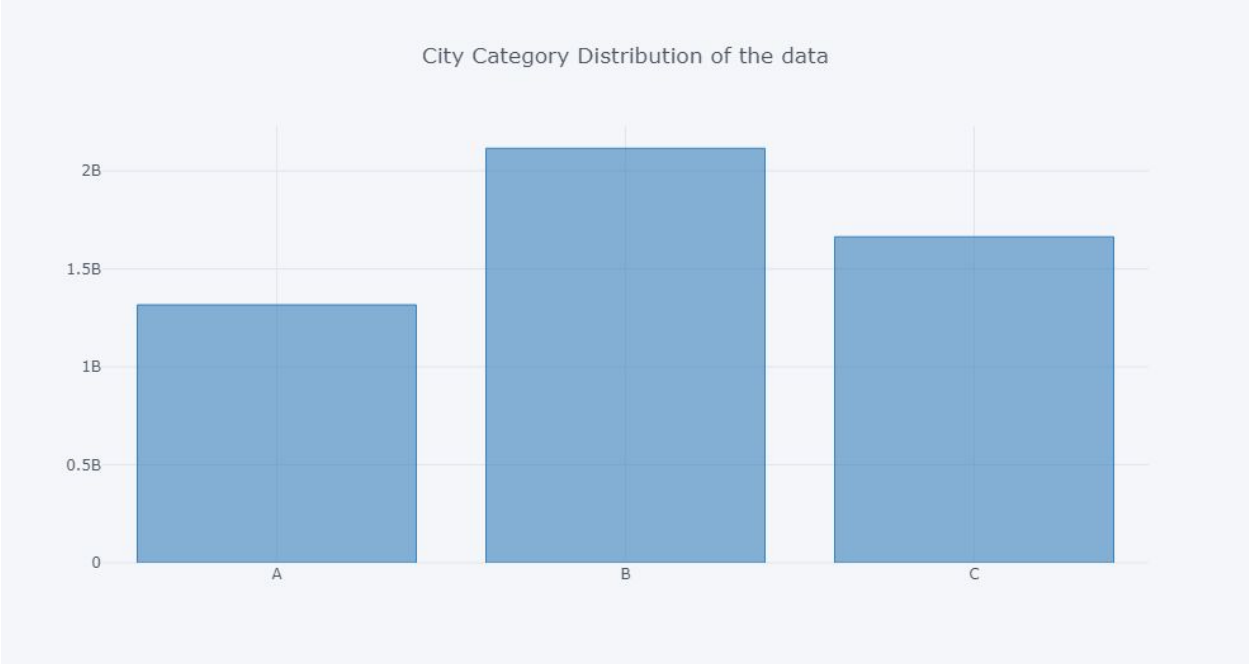
2. Purchase amounts by Age Group



3. Purchase amounts by Occupation



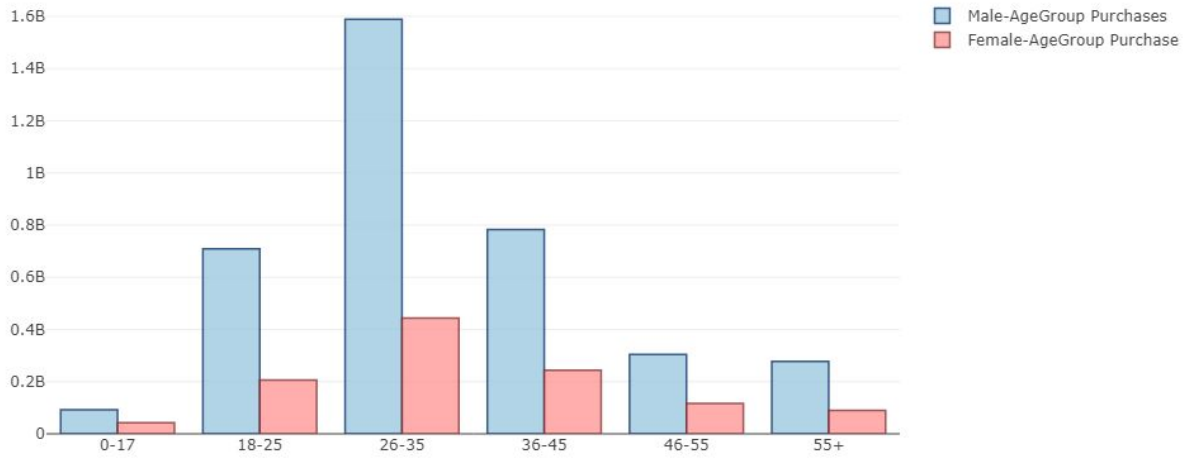
4. Purchase amounts by City Category



5. Purchase amounts by Stay Duration in the City(Years)

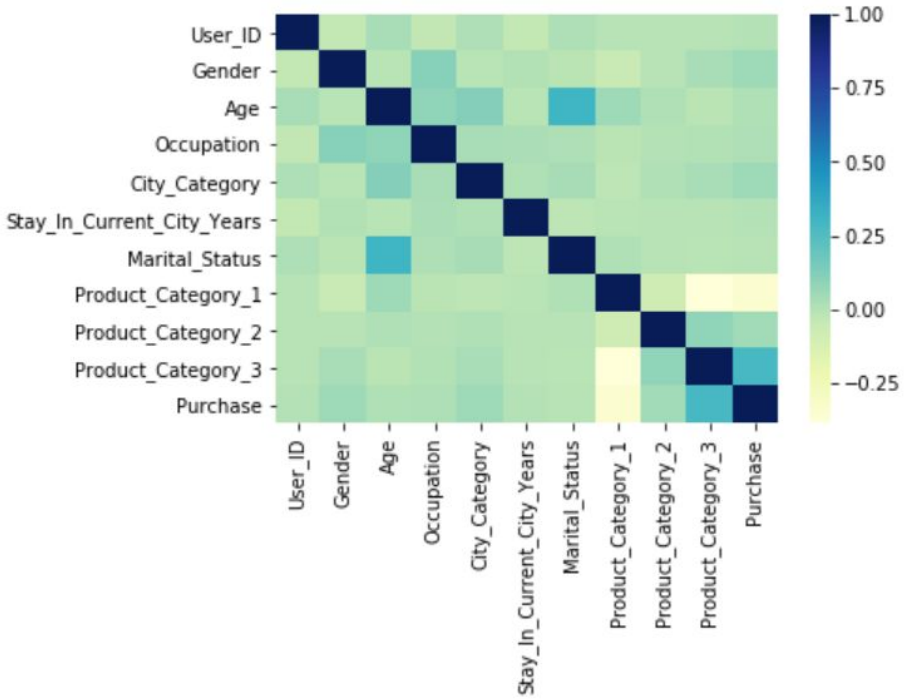
6. Purchase amounts by Marital Status

7. Comparison of the purchase amounts of Gender-AgeGroup values



HeatMaps:

Original Dataset:



2.2 Data Preprocessing

2.2.1 What is data preprocessing and why do it?

In Data mining, data preprocessing is a crucial step that involves transforming the raw data into an understandable format. It is a method of cleansing the Real-world data which is likely to contain many errors, is often incomplete, inconsistent or lacking correlation and proper behavior. It may contain outliers which introduces bias that could affect the data model and result accuracy.

2.2.2 How to do data preprocessing?

There are multiple techniques to handle different kinds of raw data.

- Data cleaning: data cleaning includes filling in the missing values, removing the outliers, smooth the noisy data, and resolve inconsistency if any. The data could have a lot of missing values which needs be to either removed or replaced with some meaningful value.
- Data integration: Data integration is bringing together all the required data from different sources.
- Data transformation: Normalizing the data if the data and aggregation of the required data.
- Data reduction: Data reduction technique is useful when handling the entire data is time-consuming and expensive. It generally works when a sample of data is representative of the entire data.
- Data discretization: Sometimes the data is in a continuous form and needs to be broken down into discrete chunks before it can be processed. This is called data discretization.
- Feature Engineering: This another important aspect of data pre-processing where the features are added, subtracted based on its correlation with the target variable.

2.2.3 Data preprocessing on our data

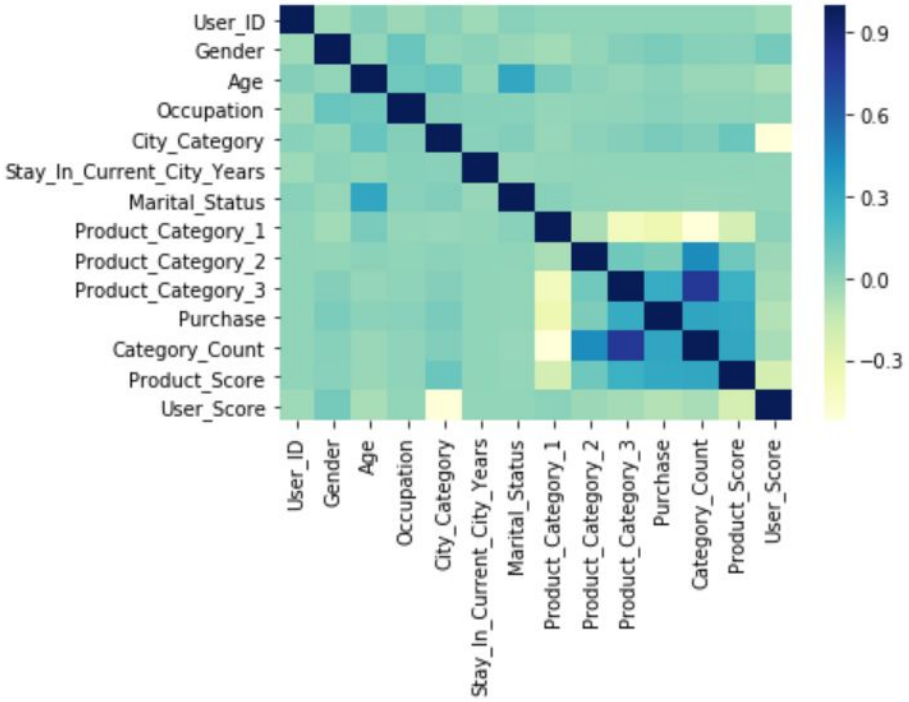
Our data only had the missing values in the product category. Some of the products only belonged to either one or two product categories(of the given three). Rest of the fields had null values. There was a huge chunk data that had product category as null, so deleting the rows with null values is not possible since that would affect the quality of the data immensely. The product categories being an ID could not be imputed by the mean/median/mode of the variable. Thus, for the starters, we replaced all the missing values with the common category (let's say category 1). We computed the model with this processing but, the model accuracy was not getting any better.

The heat map showed that the correlation of the variables with the target variables is very less. For more accuracy and fit, it is necessary that some of the variables have a better correlation with the target variable. Thus, now we processed the null values differently. We replaced the null values with zero which we further used in feature engineering.

We added a new feature category count i.e. the number of categories that a product belongs into. We noticed that the correlation of this variable with the target variable is 0.32 which is much better than the rest of the variables. We noticed that most of the given features do not have much correlation with the target variable, thus it was important to introduce a new set of variables which has a better correlation with the target variable.

Next, we calculated user score which is a relative variable that calculates the number of times a user appears in relation to the most frequent buyer in the dataset. Similarly, we also calculated a product score which is again a relative variable which calculates the number of times a feature appears in the dataset in relation to the most frequently bought product. It is noticed that the correlation of the product score with the target variable is 0.3. Thus, this gave a boost to the overall accuracy of our model.

2.2.4 HeatMap after Data preprocessing:



2.3 Data Modeling

Data Modeling is one of the most important steps in any Data Mining project, as this is the step that delivers the required results. In this section, we list all the regression models we used to train our data and predict the purchase amounts of customers. The regression models we considered are:

1. Linear Regression
2. Ridge Regression
3. Lasso Regression

4. Elastic Net Regression
5. Random Forest Regression
6. Extra Trees Regression
7. Extreme Gradient Boosting
8. Distributed Random Forest using H2O

2.3.1 Linear Regression

Linear Regression attempts to model the relationship between 2 variables by fitting a linear equation to observed data. One variable is called as an explanatory variable and the is called a dependent variable. Linear Regression can be represented as a linear equation of the form Y = a + bX; where X is the explanatory variable and Y is the dependent variable. In this context, the target variable, i.e. the purchase amount is the dependent variable as it depends on various features.

We have implemented 3 forms of Linear regression, namely Ridge Regression, Lasso Regression, and Elastic Net Regression.

2.3.2 Ridge Regression

Ridge Regression is a form of Linear Regression which is most useful when there is multicollinearity between multiple features of the dataset. In this context, multicollinearity occurs when there is a high correlation between features of the dataset. When multicollinearity occurs, least squares estimates are biased resulting in large variance between predicted and actual values. By adding a degree of bias, Ridge Regression reduces standard errors. Linear Regression is also known as the L2 regularization technique because there is more emphasis on the ridge than regression.

Ridge Regression can be represented using the expression below;

$$\hat{\beta}_{ridge} = (X'X + \lambda I_p)^{-1} X'Y$$

The hyperparameters used for Ridge Regression are alpha=0.05 and normalize=True

2.3.3 Lasso Regression

Least Absolute Shrinkage and Selection Operator or LASSO is another form of Linear Regression, in which values are shrunk towards a central point, the mean for instance. It Works well for simple models, i.e. models with fewer features. LASSO performs L1 regularization

technique, which adds a penalty equal to the absolute value of the magnitude of coefficients and the coefficients which become zero are eliminated from the model

Lasso Regression can be represented using the expression below;

$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

The hyperparameters used for Lasso Regression are alpha=0.3 and normalize=True

2.3.4 Elastic Net Regression

Elastic Net Regression is the ratio of L1 and L2 regularization techniques. Linearly combines the L1 and L2 penalties of the lasso and ridge methods.

Elastic Net Regression can be represented using the expression below;

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1)$$

The hyperparameters used for Elastic Net Regression are alpha=1, l1_ratio=0.5 and normalize=True

2.3.5 Random Forest Regression

Random Forest Regression is an ensemble supervised learning algorithm, a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. This technique produces great results even when hyperparameters are not provided.

Random Forest Regression additional randomness to the model, while growing the trees. The predictions of samples x can be made by averaging the predictions from all the individual regression trees on x; which the below expression demonstrates.

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees on x, demonstrated by the expression below;

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

The hyperparameters used for Random Forest Regression are n_estimators=20 and criterion='mse'.

The best RMSE score for Random Forest Regression was recorded when n_estimators=20, with the RMSE score being 2914.72. The RMSE score was 3009.87 when n_estimators=8 and 2971.9109 when n_estimators=10.

2.3.6 Extra Trees Regression

Extra Trees Regression is another ensemble supervised learning algorithm, a meta estimator that fits a number of extremely random decision trees, i.e. extra trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It is very similar to Random Forest Regressor in terms of optimality and performance. However, Extra Trees Regressor performs a bit worse when there is a high number of noisy features, especially in high dimensional datasets.

Extra Trees Regressor often leads to increased accuracy because of no computational burden like determining the optimal cut point; it rather randomizes the selection of cut point. One advantage of Extra Trees Regressor over Random Forest Regressor is that extra trees are computationally faster and are cheap to train.

The hyperparameters used for Random Forest Regression are n_estimators=10 and criterion='mse'. The best RMSE score for Extra Trees was recorded when n_estimators=10,, with the RMSE score being 2927.62. The RMSE score was 3033.5451 when n_estimators=6.

2.3.7 Extreme Gradient Boosting

Extreme Gradient Boosting or XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added

together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. It works best for structured or tabular datasets on classification and predictive modeling algorithms.

The hyperparameters used for Extreme Gradient Boosting are n_estimators=100 and max_depth=7.

2.3.8 Distributed Random Forest using H2O

H2O is an open source, in-memory, distributed, fast, and scalable machine learning and predictive analytics platform that allows in building machine learning models.

Distributed Random Forest (DRF) turned out to be the best performing model with our dataset using H2O. DRF generates a forest of classification or regression trees, rather than a single classification or regression tree. Each of these trees is a weak learner built on a subset of rows and columns. More trees will reduce the variance. Both classification and regression take the average prediction over all of their trees to make a final prediction, whether predicting for a class or numeric value.

3. Code

We used the libraries listed below for data manipulation & analysis and training models:

1. numpy
2. pandas
3. matplotlib
4. plotly
5. sklearn.linear_model.LinearRegression
6. sklearn.linear_model.Ridge
7. sklearn.linear_model.Lasso
8. sklearn.linear_model.ElasticNet
9. sklearn.ensemble.RandomForestRegressor
10. sklearn.ensemble.ExtraTreesRegressor
11. xgboost
12. h2o

4. Results

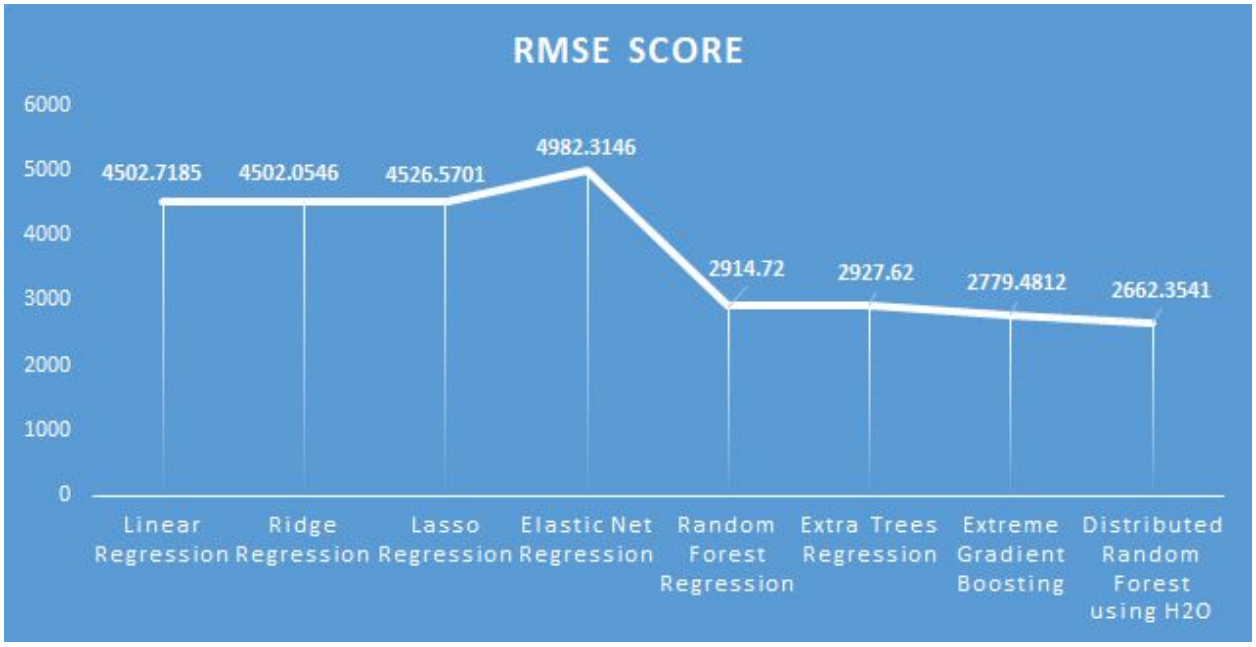
Since this is part of an ongoing competition, we were required to verify our results by uploading code and the output Excel sheet containing the User ID, Product ID and Purchase Amount on the Analytics Vidhya website. Once we upload our observations, Root Mean Square Error (RMSE) score would then be generated based on the purchase amounts (target variable) predicted and we have used the same RMSE score to compare results obtained from various models.

The table below shows the results we obtained:

Model	**RMSE Score**
Linear Regression	4502.7185
Ridge Regression	4502.0546
Lasso Regression	4526.5701
Elastic Net Regression	4982.3146
Random Forest Regression	2914.72
Extra Trees Regression	2927.62
Extreme Gradient Boosting	2779.4812
Distributed Random Forest using H2O	2662.3541

5. Discussion

The figure below shows RMSE scores obtained from different models:



Distributed Random Forest using H2O turned out to be the best model for this dataset with an RMSE score of 2662.3541. Extra Gradient Boosting had the next best RMSE score followed closely by Random Forest Regressor and Extra Trees Regressor. Another interesting observation is that all the Linear Regression models had high RMSE scores.

The ranking based on performance and RMSE scores of models on this dataset is:

1. Distributed Random Forest using H2O
2. Extreme Gradient Boosting
3. Random Forest Regression
4. Extra Trees Regression
5. Ridge Regression
6. Linear Regression
7. Lasso Regression
8. Elastic Net Regression

6. Future Work

The initial dataset had 11 features but the correlation values of each of the features with the target variable turned out to be very less. Other features generated like User Score, Product Score, and Category Count had higher correlation values and were instrumental in determining the target variable. If we could identify/generate more features with high correlation with the target variable, that could have a positive impact on the models and could result in better RMSE scores for each of the models.

It would also be interesting to see how Neural Network based modeling techniques like Tensor Flow would work on this dataset.

Additionally, mapping individual features with the target variable (purchase amount) could lead to interesting observations which could help e-commerce companies and other stakeholders involved understanding the customer spending patterns.

7. Conclusion

We would like to conclude by stating that Decision Tree-based regression models worked well for our dataset when compared to Linear regression models. There is a significant difference between the RMSE scores of both and Distributed Random Forest using H2O worked best for this dataset, closely followed by Extra Gradient Boosting, Random Forest Regressor and Extra Trees Regressor.

Secondly, this dataset is from an ongoing competition and we are currently ranked 498 out of 13,000 participants and we plan on improving our rank by the end of the competition.