Nile University

CSCI322 – Introduction to Data Analysis

How much will a customer spend on a Black Friday?

Made by:

Dalia Mohamed Zahran, 1510074

Ehab Ibrahim Hassan, 1610235

Mennat allah Raafat, 1711034

Submitted to:

Dr. Mustafa El-Attar

Abstract

Good pricing for different products in different markets can be tricky for stores and businesses. They need to put the perfect prices that encourage customers to buy their products while getting maximum profit. This challenge arises mainly in times of big sales seasons like Black Friday and Cyber Monday! In this project, the features of customers were visualized and correlation was investigated to get insights of the data. Also, different Machine Learning techniques are used to predict the purchase amount by customers on different product categories in a retail store based on purchase history by those customers. To get the purchase predictions, we use different regression models to find the inner correlations between different parameters which include Linear Regression, Ridge Regression, Decision Tree, and Random Decision Forest. The accuracy of each model was calculated to figure out which model is the best. Depending on our results, the managers can price the products with highest sales and trending categories so that they gain capital benefits.

Table of Contents

Introduction

Black Friday and holiday seasons in the United States represent a good percentage of the annual retail sales; up to 20 percent of retail sales happens in those seasons [2]. During the sales days, people focus on the discounts rates and the money they will save for buying the items they want [3]. However, for the economy, it is a season and using strategies to get people out of their homes spending money on purchases is the radical part [4]. This behaves the way for our question: How can stores and businesses to price their data according to consumer spending trends which helps accordingly.

This is important because predicting the proper prices which can combine between both the highest benefit for the company and the most amount of customers who buy the products. This can improve the business processes, enhancing decision making and gaining the ability to direct, optimize, and automate decisions, on demand, to meet defined business goals[1]. Also, studying different attributes related to the customers and cities and visualizing them make the business owners more familiar with their market and how to take decisions that are beneficial to them.

There are different hypothesis for different perspectives:

Customer hypotheses:

Income: People with higher income should spend more on products.

Gender: Men will spend more on products.

Age: People range from 18-25 will spend more on products.

Martial Status: Married people will spend more on products.

City Hypotheses:

City Type : Tier 1 cities should have higher sales because of the higher income levels of people there.

Years individuals stay in the city: The more years the customer spend in the city, the more stable they become, so the more they spend on products.

Data Description

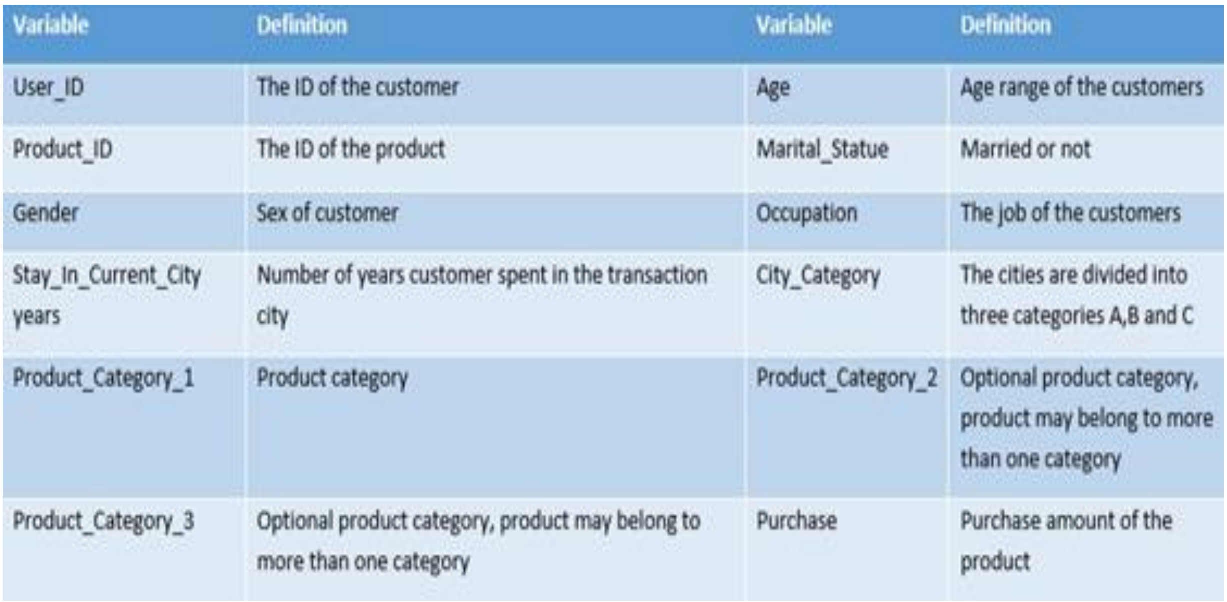
- **Data Set Collection**

  Since we do not have time to obtain data from a local store, we searched over the internet to get proper data set with detailed features and enough data records. It should also be in a suitable format to be able to import and work on it such as csv, xlsx, xml, or json file formats. The data is from the data is collected from A retail company "ABC Private Limited" milted.

- **Data Set description**

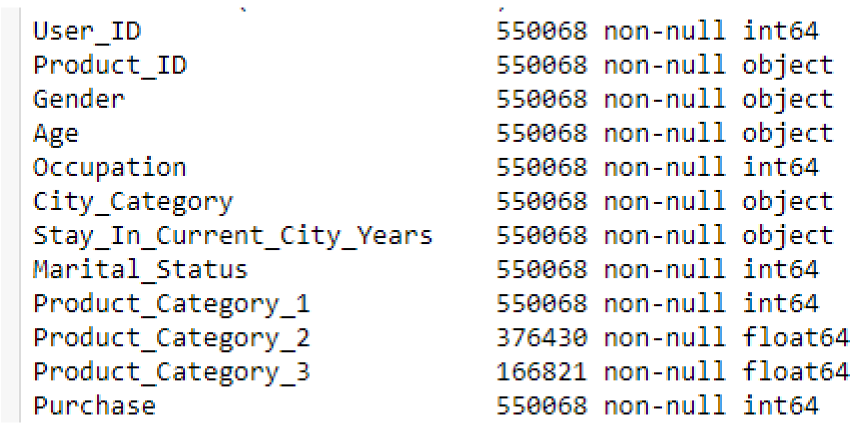
  It contains 12 columns and 550068 raw. The attributes of the data are as follows:

Table 1: Dataset attributes definition



Variable	Definition	Variable	Definition
User_ID	The ID of the customer	Age	Age range of the customers
Product_ID	The ID of the product	Marital_Statue	Married or not
Gender	Sex of customer	Occupation	The job of the customers
Stay_In_Current_City years	Number of years customer spent in the transaction city	City_Category	The cities are divided into three categories A,B and C
Product_Category_1	Product category	Product_Category_2	Optional product category, product may belong to more than one category
Product_Category_3	Optional product category, product may belong to more than one category	Purchase	Purchase amount of the product

The description of the attributes is as follows:



```
User_ID                       550068 non-null int64
Product_ID                    550068 non-null object
Gender                        550068 non-null object
Age                           550068 non-null object
Occupation                    550068 non-null int64
City_Category                 550068 non-null object
Stay_In_Current_City_Years    550068 non-null object
Marital_Status                550068 non-null int64
Product_Category_1            550068 non-null int64
Product_Category_2            376430 non-null float64
Product_Category_3            166821 non-null float64
Purchase                      550068 non-null int64
```

Figure 1: Data Description

Data Analysis Plan

1. **Data Cleaning and Pre-processing**
   a. Check for incomplete, noisy, or inconsistent data and handle them.
   b. Encoding categorical data into numerical data in order to be able to use them later in the models.
2. **Data Exploration and Visualization**
   a. Explore / Visualize the data to find patterns and trends.
3. **Data Reduction and Modelling**
   a. Find correlations between our target features and the other independent features.
   b. Do dimensionality reduction if necessary.
   c. Do a Principal Component Analysis (PCA) in order to decrease complexity.
   d. Use different models to predict our target features. These models are:
      i. Linear Regression
      ii. Ridge Regression
      iii. Decision Tree
      iv. Random Decision Forests
4. **Hyperparameter tuning**
   a. Do grid search to find the best set of parameters for each of the models in order to increase the accuracy of the predictions
5. **Data Interpretation**
   a. Get insights from the results and relate them to our hypotheses.

Data Analysis Pipeline Formalization

1. Data Cleaning and Pre-processing

After collecting the data, we should check it for any problems. It's a challenging and time-consuming step as it needs many decisions in the way of representing the data. But before processing the data, we tried to get a sense of the correlation of the data so that we make the process faster by eliminating the features with the least importance. We can see from the heatmap in fig. (2) that there are many uncorrelated attributes like occupation and product categories.

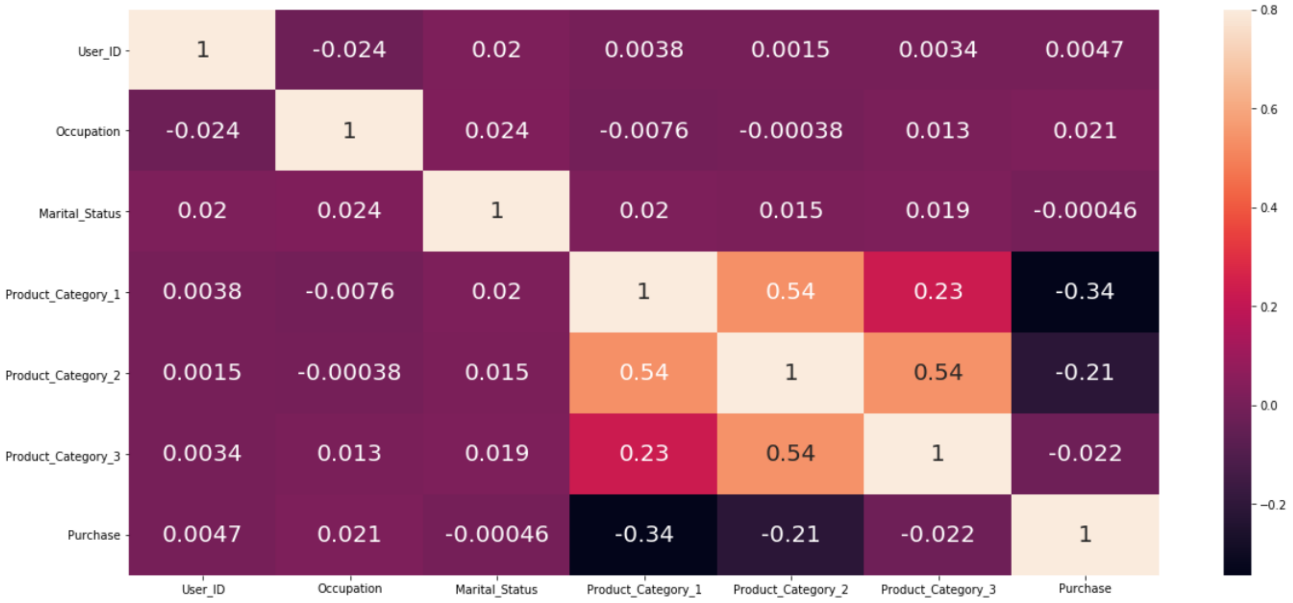


Figure 2: Attributes heat map before cleaning

According to the heat map:

- Product category 1 correlates to product category 2 but not logical to analyse.
- Product category 1 correlates to product category 3 but not logical to analyse.
- Product category 2 correlates to product category 3 but not logical to analyse.
- Product category 1 negatively correlates to Purchase so it is logical to analyse it.
- Product category 2 negatively correlates to Purchase so it is logical to analyse it.

Then, the following pre-processing was done:

a. Firstly, checking the age column we found that it is grouped into ranges, but it is better to have it in integers forms as numeric data types is much faster in calculation than other data types.
b. The city category column is divided into letter categories, converting it to digit categories would be more usable later in the models.
c. The gender also is either female or male, it could be binary.
d. Stay_in_current_city_years column has '+' symbol. This symbol converts the type to string which will increase the running time, so it was removed.
e. Product categories columns (last three columns) have NULL values, so we filled them with 0. The reason we filled the null values with zeros is that it is simply a classification of the product, so we do not need any weight for the categories that the product does not belong to.

2. Data Exploration and Visualization

First, we tried to get sense of the data and what classes of each attribute have more weight and how they relate to each other. The histograms shown in fig. (3) gives us an idea on the distribution of the data.

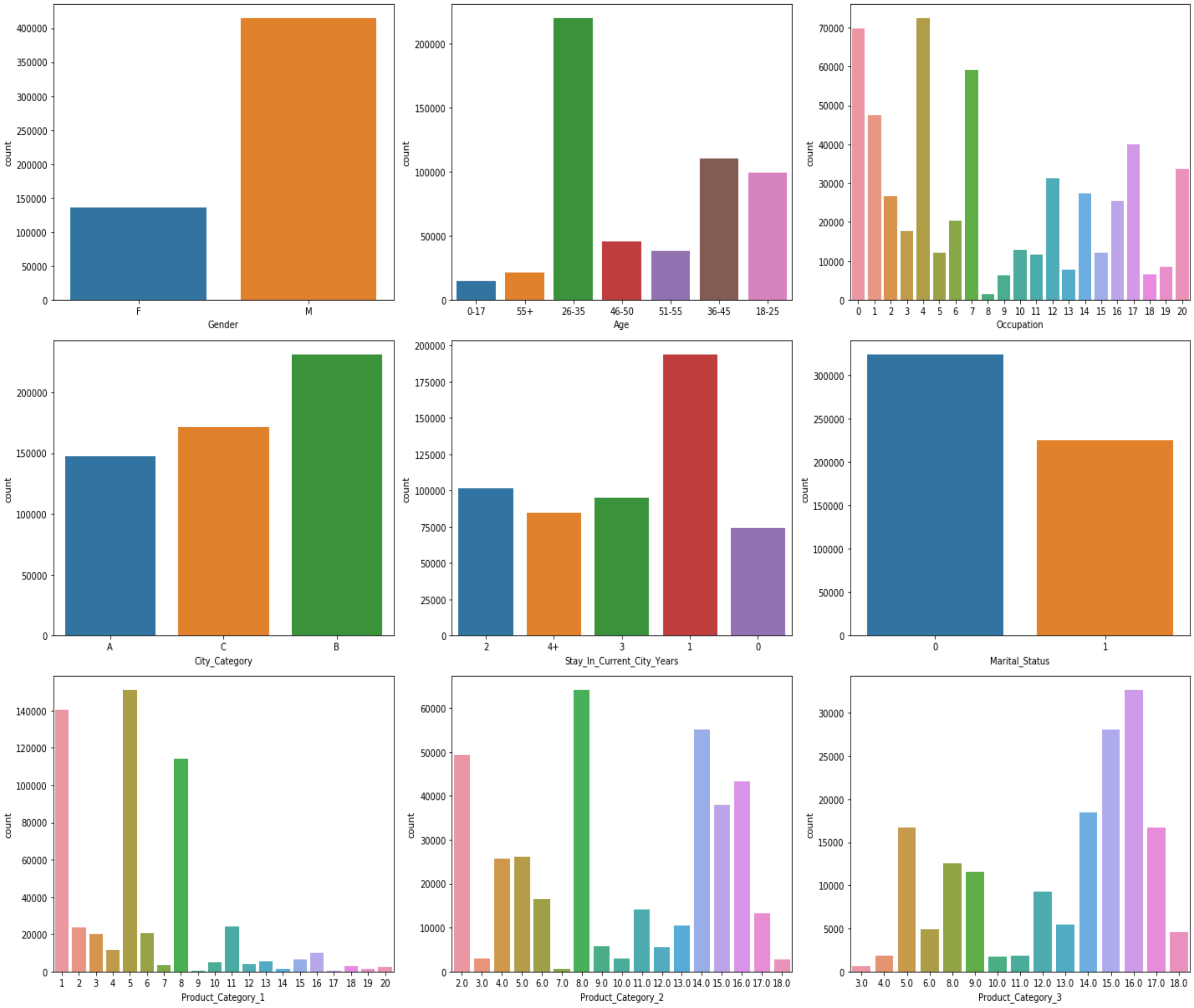


Figure 3: Data Distrbution

Then, to know how each class of each attribute relate to the purchase, we plotted bar plots for each of the attributes versus the purchase as shown in fig. (4) and fig. (5). Here, we can do it by summing the values of each class and relate it to the purchase or use the mean of each class. The difference is that there may be a class with high number of people but they contribute to the output "purchase" less than another class of a small number of people. The importance of the average plots come into place here. It might be needed from a business wise they might need to target the customers whose number is a lot not their purchase.

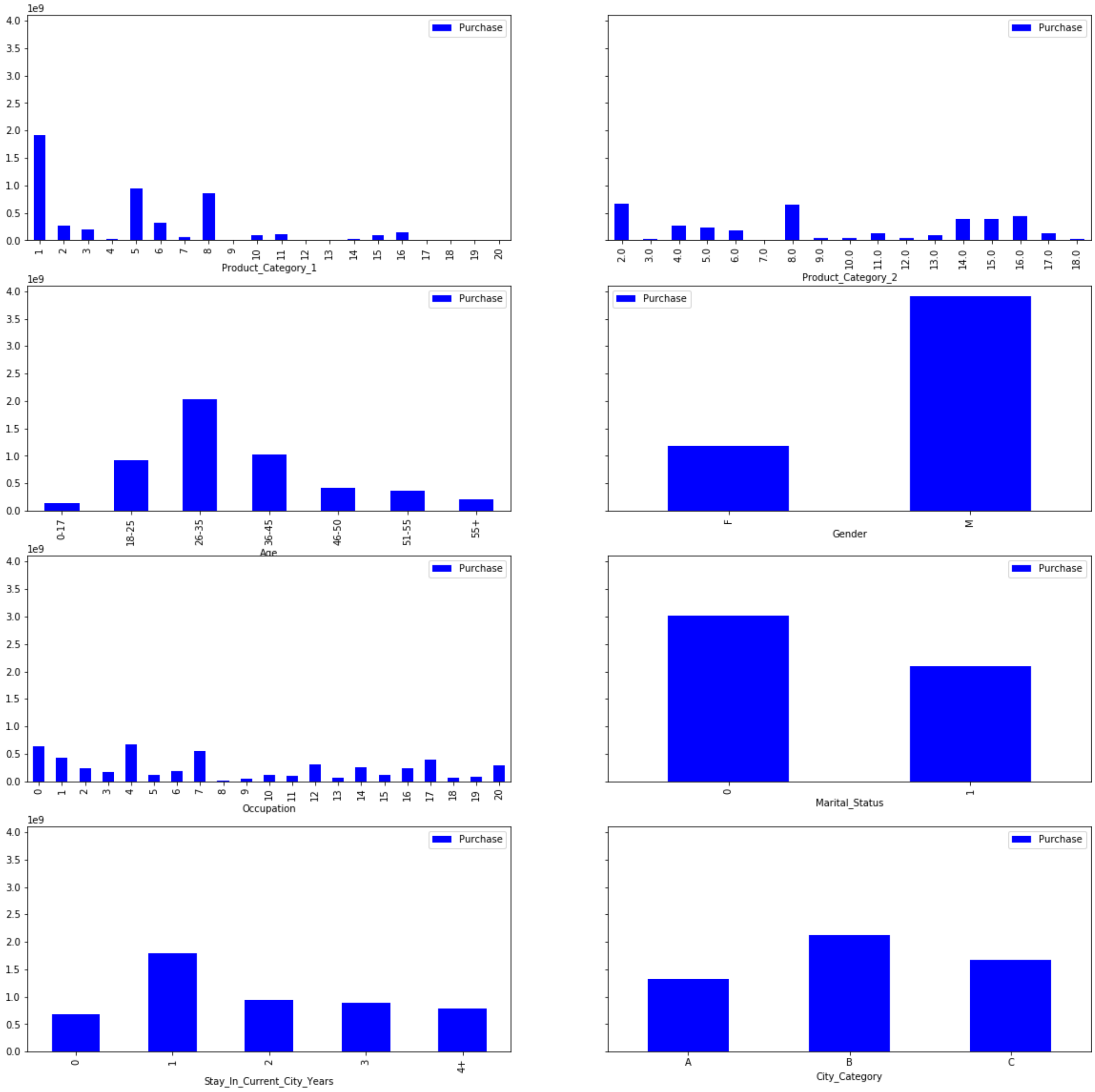


Figure 4: bar plot of each attribute and sum of purchases

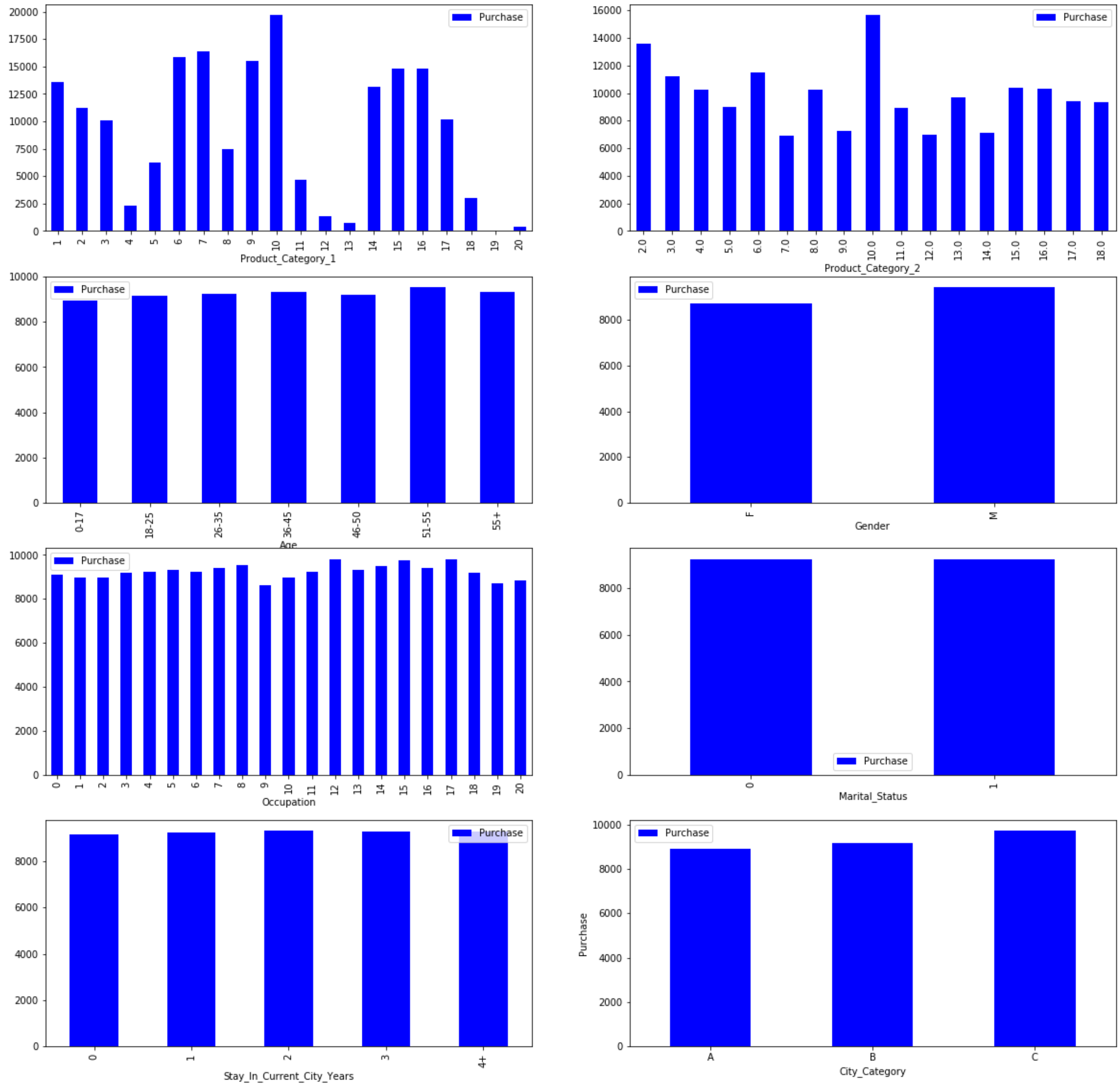


Figure 5: bar plot of each attribute and mean of purchases

3. Data Reduction and Modelling

Correlation:

The data now is organized and cleaned, so we can start checking the correlations between the different features. The correlations will show us which attributes have the most impact on the results, so we can decide which features to eliminate in order to reduce the noise in the model. This dimensionality reduction allows the model to be simplified which will decrease the computational overhead. Also, giving weight to each attribute makes the data set more reliable, this step is implicitly included in the models. So, first, we plotted the heat map again to show correlations after the cleaning and pre-processing done and the results can be noticed in fig. (6).

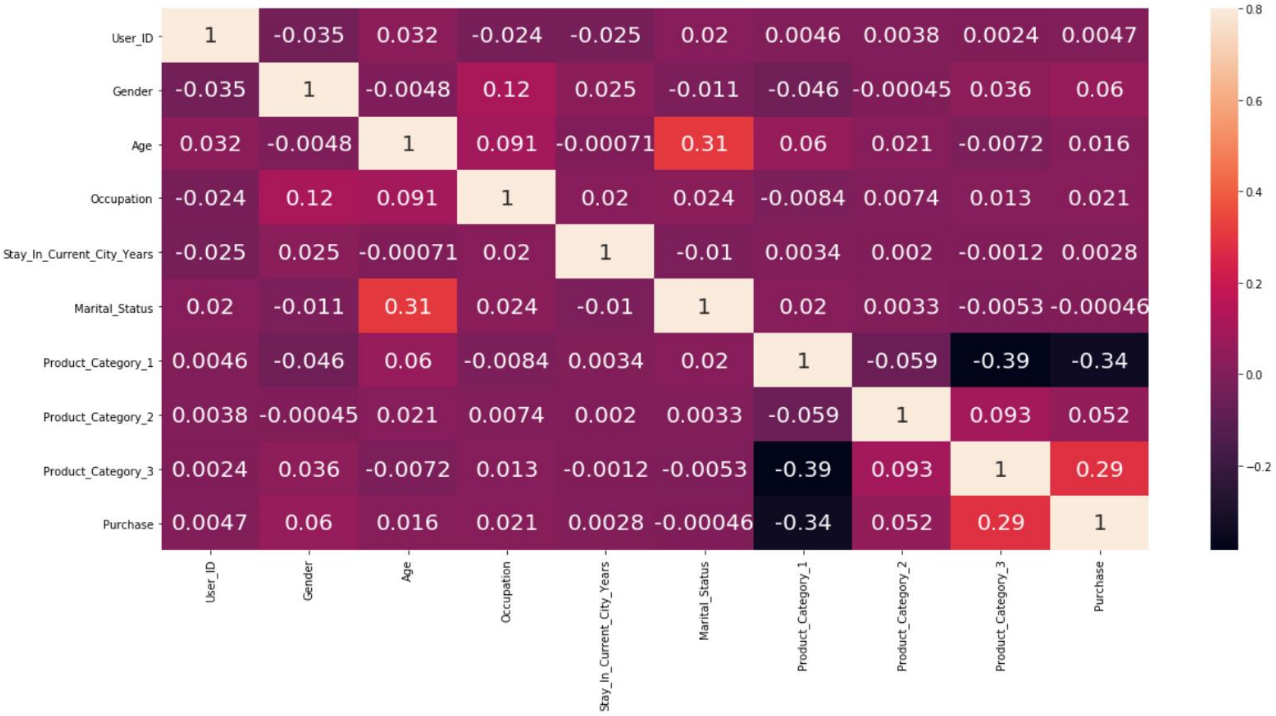


Figure 6: heat map after cleaning

Then, since user ID and Product ID does not relate to the purchase amount and will not affect our model computation, we dropped them. The next step was to choose the most important features to decrease the computational overhead. So, we applied Principal Component Analysis (PCA), to get a smaller number of features to train and test our models.

PCA Analysis:

In this analysis, we tried all the combinations of the number of features and trained our models with them and calculated the accuracy. It showed that using 8 features gives us the best accuracy, however decreasing the number from 10 to 8 is not enough. So, we decided to use 5 features since the difference in accuracy is not high. Accuracy of the 8 features is 65.25% and accuracy of 5 features is 65.07%.

Model Testing:

The next step was to use our trained models to predict the purchase amount spent by customers for the test data. We used some of the built in regressors to do the predictions. Linear Regression, Ridge Regression, Decision Tree, and Random Decision Forests were used. We will refer to the results of each model in the results section. After testing, our best accuracy was not satisfying so we wanted to tune the parameters of each model.

Hyperparameter tuning:

To tune the models' parameters, there were two options which are using grid search to find the best parameters for each model or tune them manually until we reach a satisfying accuracy. We tried both options, however, the grid search was computationally exhausting and we could not get it to work because of the processing power limitations. So, the obvious option was to do it manually using try and error, and logic. Our final options are stated in the results section below.

Results

We visualized the distribution of the attributes of the data set (it might be needed from a business wise they might need to target the customers whose number is a lot not their purchase). Also, we tested hypotheses we made. Results of the hypotheses are as follows:

1) **Customer hypotheses:**

- **Income:** Our hypothesis about it was true as there are some occupations which have higher representations,the graphs show that the amount each user spends on average is nearly the same for all occupations. However,of course, occupations with the highest representations will have the highest amounts of purchases.
- **Gender:** Our hypothesis was true and indeed men spent more on products.
- **Age:** Our hypothesis was wrong as the data showed that people ranging from 26-35 spent more on products. However, the average amount spent by group age is almost the same for everyone. Surprisingly, on average customers with more than 50 years old spent the most on the product.
- **Martial Status:** Our hypothesis was wrong as the data showed that non-married people spent more on products. However, on average an individual customer tends to spend the same amount independently if his/her is married or not because the number of single people who buy are more than the married people.

2) **City Hypotheses:**

- **City Type:** It is indeed hard to judge this hypothesis as city type 'B' had the highest number of purchases. However, on average, City type 'C' customers spent more on the products.
- **Years individuals stay in the city:** This hypothesis was wrong as people who are relatively new in the city spent more on the products. However, on average, all

customers with different years stayed in the city tend to spend the same amount on purchases.

The training and testing accuracy of models we used to predict the purchase amount are as follows:

Model Name	**Training accuracy**	**Testing accuracy**
Linear Regression	14.73%	14.63%
Ridge Regression	14.69%	14.61%
Decision Tree Regressor * Parameters: - max_depth=15 - min_samples_leaf=100 - other parameters are set to default	66.3%	65.07%
Random Forest Regressor * Parameters: - max_depth=70 - min_samples_leaf=80 - max_leaf_nodes=250 - min_samples_split=50 - default number of estimators = 10	65.95%	65.64%

The previous table showed that Decision Tree Regressor and Random forest Regressor seemed promising, so the team thought of applying Grid Search for parameter tuning for both regressors by giving a range for each parameter, and the Grid Search combine all of them together and then getting the best combined parameter which gives the highest accuracy. However, the computational power of the team’s laptops was not enough. So, the parameters were tuned manually. Decision tree model with max_depth=70, min_samples_leaf = 80, max_leaf_nodes = 250, and min_samples_split = 50 achieved the highest accuracy from all the decision trees parameter combinations with testing accuracy of 65.4% and training accuracy of

65.67%. After increasing the number of estimators (trees) in the random forest the training accuracy reached 65.95% and testing accuracy reached 65.55%.

Conclusion

Data Analysis has a great role in understanding customer behaviour which can help to take the good decisions; these decisions will have the highest benefits to the company. Our data showed which type of customers is predicted to spend more on the purchase depending on different features: age, gender, marital status, occupation, the city type, and the number of years the customer spend in this city. All of these features effects were discussed in the results. Also, the Random Forest algorithm was the best to perform on the data to predict the purchase amount of the user with the training accuracy of 65.95 % and testing accuracy of 65.55%.

References

[1]

[2] KIMBERLY Amadeo. How much do americans spend on black friday?, Dec 2018.

[3] Jane Boyd Thomas and Cara Peters. An exploratory investigation of black friday consumption rituals. International Journal of Retail & Distribution Management, 39(7):522–537, 2011.

[4] Sharron Lennon, Minjeong Kim, Jaeha Lee, and Kim KP Johnson. Thrilled or angry: Consumer emotions on black friday. 2016.