ML HACKATHON
LITERATURE SURVEY

BLACK FRIDAY SALES PREDICTION

SHUVAM BOSAN (MT2018116)
RAHUL RANJAN (MT2018090)
BISWAROOP BANERJEE (MT2018026)
Department of Computer Science
IIIT, BANGALORE

ABSTRACT

In today's digital age it has become all the more important for any business to understand it's customer behavior more precisely. Manufacturers aim to exploit historic customer data based on consumer's city of residence, his age group, his occupation,etc to understand customer behavior and to recommend products that might interest the consumers. They also try to identify consumers apriori and make products that would suit them.

The problem statement we worked on is similar to the above scenario where a retail company namely âĂIJABC Private LimitedâĂİ wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. The dataset contains purchase summary of various customers for selected high volume products from last month. The data set also contains customer demographics (age, gender, marital status, city-type, tenure of stay in current city), product details (product id and product category) and Total purchase amount from last month.

Now, our objective is to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

ABOUT THE DATASET

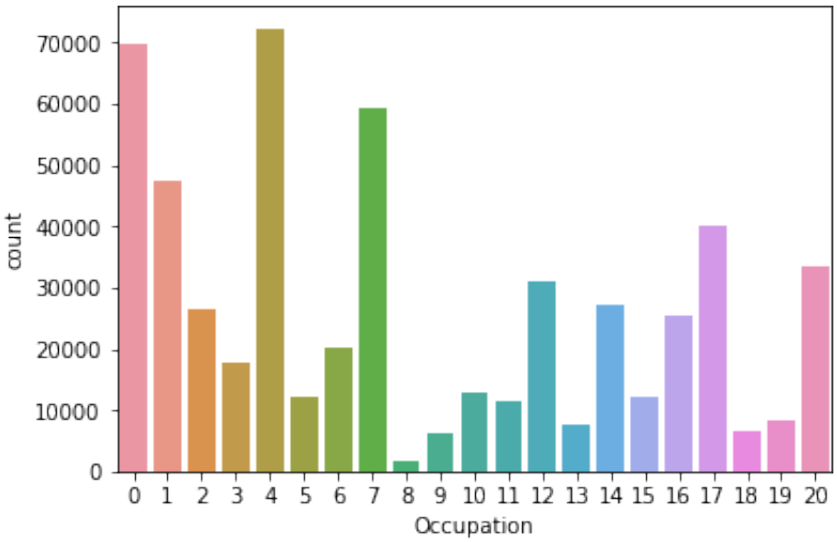
The dataset was collected from an ongoing contest on Analytics Vidhya. Both test and train datasets were taken from the contest itself. The train dataset contains of a single file with the following attributes:

Variable	Definition
User_ID	Distinct User ID
Product_ID	Distinct Product ID
Gender	Gender of User(M/F)
Age	Age group of user like 0-17,18-25,etc
Occupation	Categorical label with no identification
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years of stay in current city
Marital_Status	Marital Status of the user
Product_Category_1	Category of a product. Contains discrete categorical values
Product_Category_2	Product may belongs to other category also
Product_Category_3	Product may belongs to other category also
Purchase	Purchase Amount (Target Variable to be predicted)

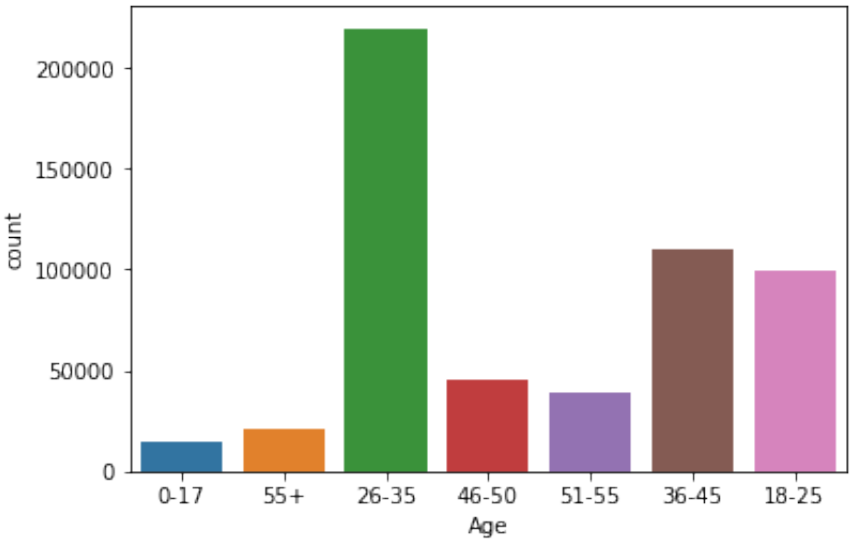
The test dataset contains of same attributes minus the target feature.

EXPLORATORY DATA ANALYSIS

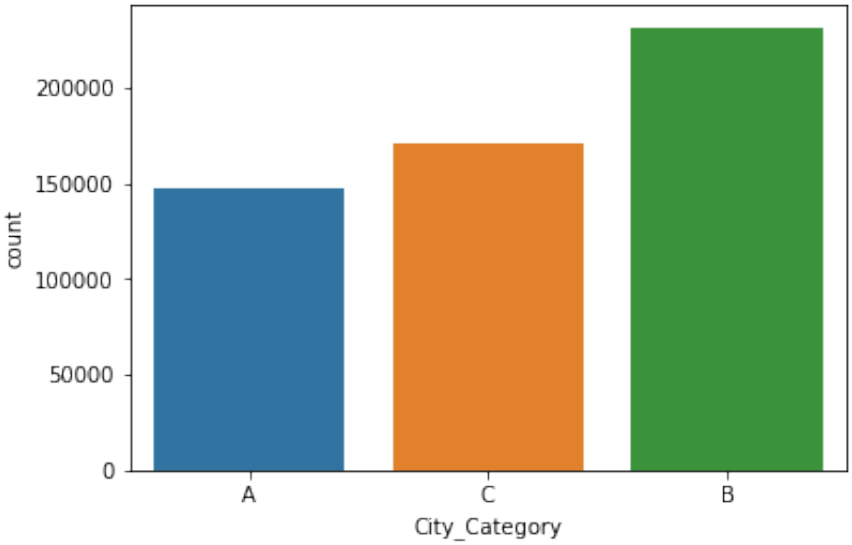
The dataset contained of 11 decision attributes and 1 target feature. Out of these 11 decision attributes, all except User_ID and Product_ID were found to be discrete values. We plotted the histogram of each attribute to visualize the frequency of each attributes along it's categories.



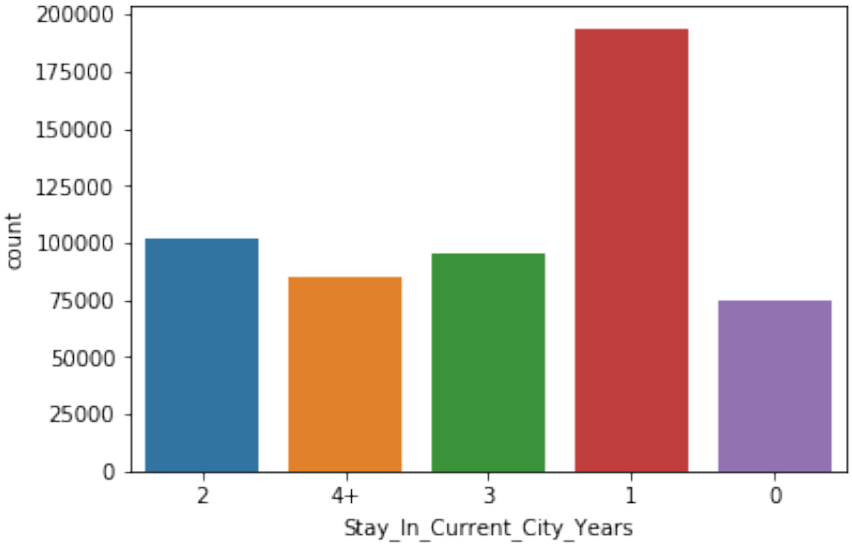
Category 8 shows least purchase count while categories 0,4,7 show overwhelming purchases



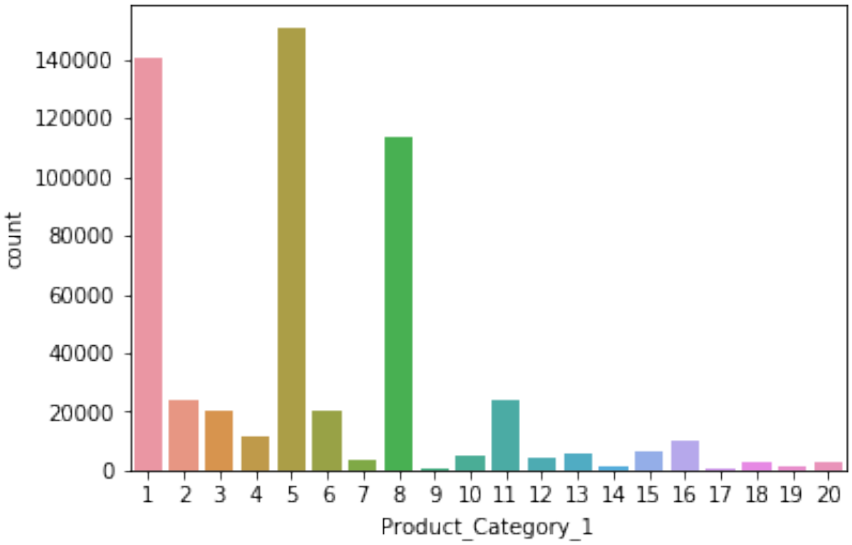
26-35 age group shows maximum interest in shopping.



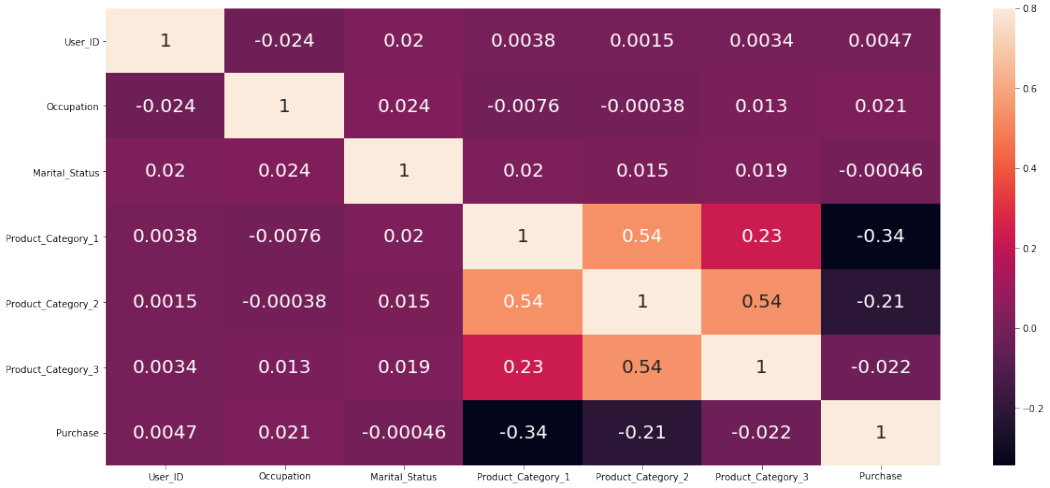
Middle tier cities show maximum sales (assuming A to be max and C to be min tier cities or vice-versa)



People new to any city make good amount of purchase. This could be attributed to furnitures, electrical or electronics appliances that people need when they relocate.

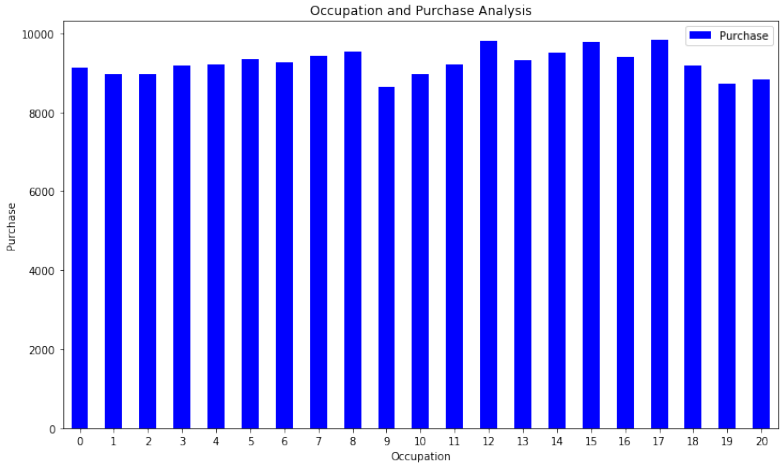


Product categories 1,5,8 show overwhelming sales. They might be day-to-day essentials.

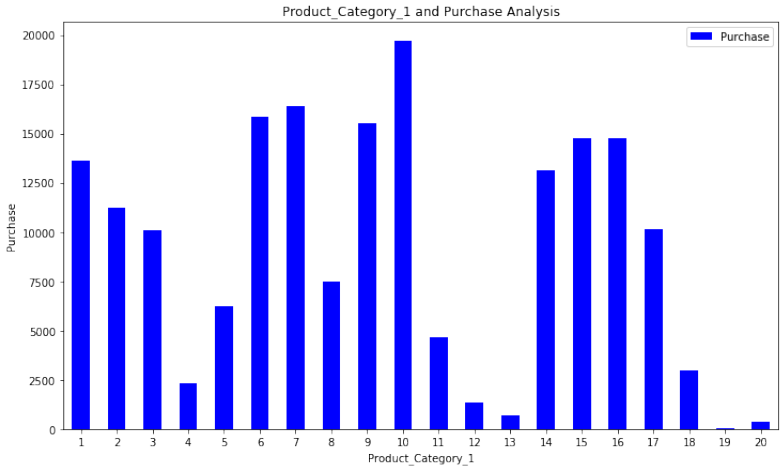


Attributes appear to be linearly independent which is a good thing, although there is some correlation among the product categories. This could be attributed to the fact that a single item can belong to multiple categories like a computer mouse can belong to categories computer accessories as well as gaming accessories.

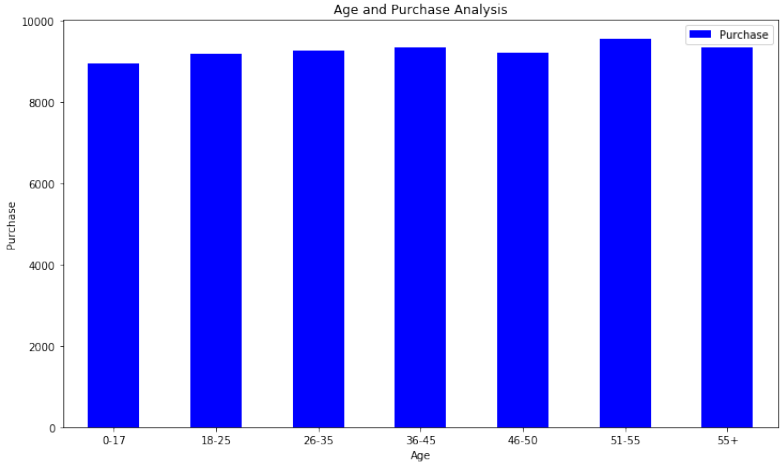
Now, let's analyze the relation of decision attributes with the target feature.



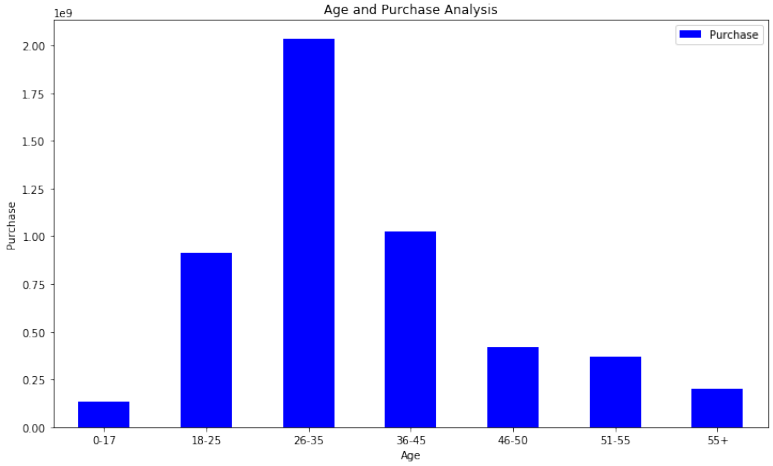
Apart from categories 9,19 and 20 , all categories show very similar purchase trends. This could be attributed to the fact that their are products targeted to audience depending upon financial constraints.



Products of categories 19 and 20 show almost negligible average sales.

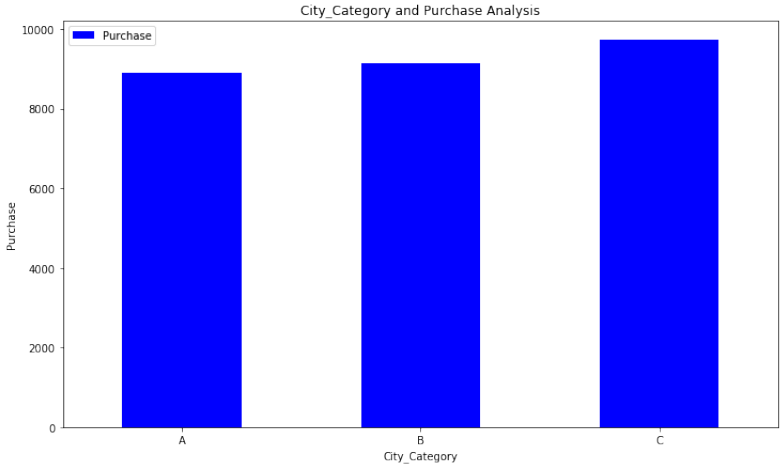


Although age group 26-35 showed max no of buyers, it is interesting to view that the average purchase amount across all age groups is nearly the same. With age people tend to spend differently.

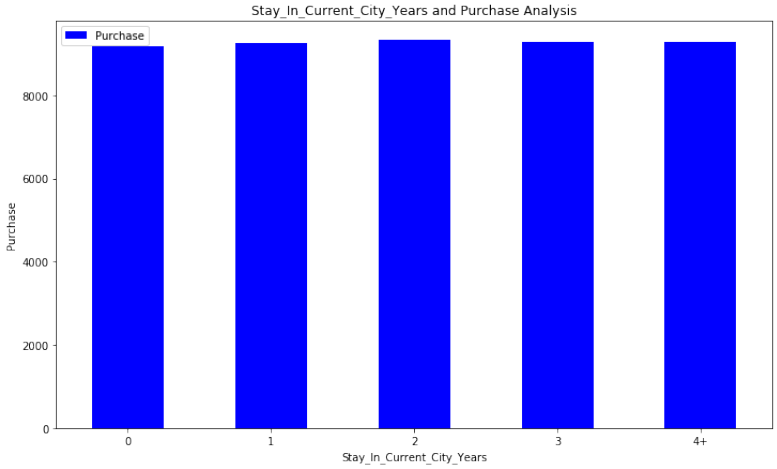


But the total purchase figure concorts with our initial observation about purchasing

frequency of the young generation. "Plots don't lie!"



Although the middle tier cities registered the maximum purchases, they did not contribute as much as the highest tier city(assuming C to be tier 1)



Although people new to the city make the most purchases, over time they tend to spend more luxuriously in terms of quality than quantity

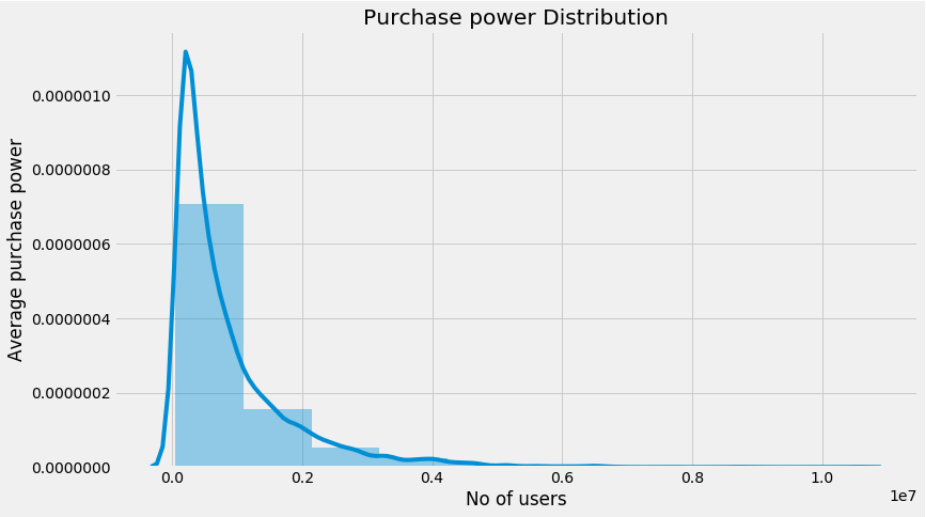
Observations while analyzing the dataset:

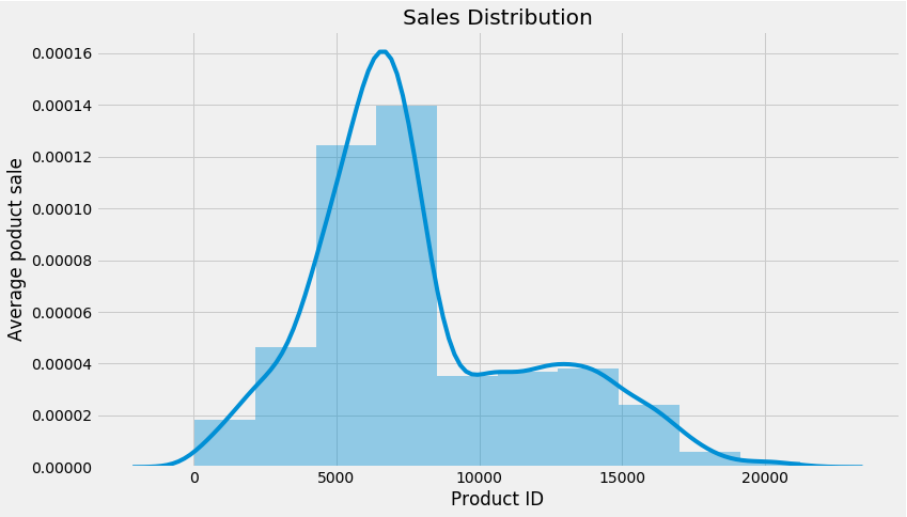
- Only product categories 2 and 3 have missing values.
- Attribute values need not be normalized, although they might be needed to label encoded.
- Attributes are linearly independent values with 0

FEATURE ENGINEERING

Through our data analysis, we found out that the columns namely **Product_Category_2** and **Product_Category_3** had some missing values. We attribute this to the fact that some products might not be classified under more than one categories. So we filled these values with 0. Remaining categorical data was label encoded so that our model could work with it.

Along with this we decided to have a look at the average sale of each product and the average purchasing power of each customer.





It is interesting to see that both these attributes follow a Gaussian distribution.Hence we decided to append two new features into our dataset: **Average_Purchase_By_Product_ID** and **User_Category**

APPROACH AND MODEL BUILDING

We first implemented a simple linear regression model with default hyper-parameters and cross validation. Our accuracy was about **21%**. We further applied regularization but there was little improvement.This gave us a score of **4400** on the contest leader-board.

Next we used a **Decision Tree Model** with cross validation that improved our accuracy. We tweaked the hyper-parameters and finally obtained **42%** accuracy and a score of **3800** on the contest leader-board.

We further implemented a **Random Forest Model** with cross validation to obtain an accuracy of **67%** and a score of **2800** on the contest leader-board.

Finally we resorted to **XGBoost** and tuned the hyper-parameters to obtain a score of **2488**. We tried to play around with XGBoost and created an ensemble of XGBoost models just by tweaking the hyper-parameters. This gave us a score of **2464** and we were within the top **10%** of the leader-board with a global rank of **136**.

VALIDATION AND TEST METRICS

We used the contest leader-board as test-metric to verify our approach. Two of our members participated actively in the contest. Currently they are ranked **136(handle: raul1rnjn535_3)** and **207(handle: banerjee13)**. **Link to the leader-board:** Analytics Vidhya Black Friday Sale Leaderboard

REFERENCES

1. Analytics Vidhya for contest and dataset.
2. google.com for everything