**Mile stone Report for my Capstone Project**

**Project Title:** Groceries product sale forecasting

**Problem statement:** Analyze and predict customer purchasing behavior. We need to predict which previously purchased product customer will buy again.

**Client:** The client for this project is online groceries shopping application owner like Instacart. They provide web application and mobile phone apps to their customers to shop for the groceries product from various retailers. They provide same day delivery by their personal shoppers.

This project will help them give insight on their customer’s orders. They will come to know what category of customer buying which product. What time of the day the maximum shopping happens. What day of the week more orders comes from the customer? The project will predict which previously purchased products will be in a user’s next order. It will help decide the product inventory in different retail store. It will also help them target customers with the product options to buy.

**Data:** The data for this project is from the Kaggle competition - Instacart Market Basket Analysis. The data has been made sure that not actual customer’s information is present. The data has sample of 3 million grocery order details. The data is transactional. The granularity is at the order id level. There are more than 200,000 users order details with each users have 4 to 100 order details provided. It gives details what hour of the day, what day of the week and how many days back the same order was done.

Bring the dataset in the Python notebook and check the structure, shape and few top rows of the data. Do some preliminary analysis on the data like how have rows are blank or Null. Try cleaning all those data. In the string data remove unwanted spaces so that data is consistent throughout. Check how many prior, test and training data is there. If there are not enough test data try to get more test data from the training and prior dataset. This step is important as we need to create robust machine learning models.

**Problem solving approach:** I will use Python and Machine learning algorithms on the data. I will understand, wrangle and clean the data as much as possible. I will use matplotlib and seaborn modules to create statistical chart to show order history. I will use variables like day of the week & time of the day to find the relationship between orders. I will use jupyter notebook for Python coding. I will take regular help from my Mentor.

**List of datasets:** Below are the list of dataset I am using for this project.

**Orders**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **order\_id** | **user\_id** | **eval\_set** | **order\_number** | **order\_dow** | **order\_hour\_of\_day** | **days\_since\_prior\_order** |
| 2539329 | 1 | prior | 1 | 2 | 8 | NaN |
| 2398795 | 1 | prior | 2 | 3 | 7 | 15 |
| 473747 | 1 | prior | 3 | 3 | 12 | 21 |
| 2254736 | 1 | prior | 4 | 4 | 7 | 29 |
| 431534 | 1 | prior | 5 | 4 | 15 | 28 |

**Notes:** the above tables has orders information. Which day of the week the order was made, what hour of the day the order was done. After how many days the order was done. It also gives details whether orders are from Prior Set, Training set or test set.

**Department**

|  |  |
| --- | --- |
| **department\_id** | **department** |
| 1 | frozen |
| 2 | other |
| 3 | bakery |
| 4 | produce |
| 5 | alcohol |

**Notes:** The department table has information of all the different department the order was made.

**Aisle**

|  |  |
| --- | --- |
| **aisle\_id** | **aisle** |
| 1 | prepared soups salads |
| 2 | specialty cheeses |
| 3 | energy granola bars |
| 4 | instant foods |
| 5 | marinades meat preparation |

**Notes:** The Aisle table has information of all the different Aisle the order was made.

**Product**

|  |  |  |  |
| --- | --- | --- | --- |
| **product\_id** | **product\_name** | **aisle\_id** | **department\_id** |
| 1 | Chocolate Sandwich Cookies | 61 | 19 |
| 2 | All-Seasons Salt | 104 | 13 |
| 3 | Robust Golden Unsweetened Oolong Tea | 94 | 7 |
| 4 | Smart Ones Classic Favorites Mini Rigatoni Wit... | 38 | 1 |
| 5 | Green Chile Anytime Sauce | 5 | 13 |

**Notes:** The Product table has information of all the different Product in different department & aisle the order was made.

**Other Potential DataSet:** Prioror training Data Set for different Order & Product. Data Structure can be.

|  |  |  |  |
| --- | --- | --- | --- |
| **order\_id** | **product\_id** | **add\_to\_cart\_order** | **reordered** |
| 2 | 33120 | 1 | 1 |
| 2 | 28985 | 2 | 1 |
| 2 | 9327 | 3 | 0 |
| 2 | 45918 | 4 | 1 |
| 2 | 30035 | 5 | 0 |

**Notes:** Above table gives information on order & product information and when that product is reordered.

**Data Analysis:** Initial wrangling and cleaning of data showed the data structure, shape and different datatypes of the data. After cleaning and making data ready for analysis, we see from the order table that there are close to 400K orders information. Out of that 200K has prior orders information, 130k of training orders and 75k orders as test set.

|  |  |  |
| --- | --- | --- |
|  | **Eval\_Set** | **Count\_Orders** |
| **0** | prior | 3214874 |
| **1** | test | 75000 |
| **2** | train | 131209 |

I created Histogram chart on the Order table and found that Monday & Sunday are 2 days of the week where maximum order were made. Afternoon is when maximum order was made. The charts are available in the python notebook.

Using the Heat map from seaborn libraries we could see Monday morning & Sunday afternoon that the maximum orders were made.

Produce is the department where maximum orders were made.

We found from the data that re-ordering of the products happened mostly on Sunday afternoon.

Alcohol products are bought more on Fridays which make sense as Saturday & Sunday are holidays.

We also found from the re-ordering chart that, re-ordering happen mostly after 7 days or after 30 days. Most people buy weekly or monthly groceries.

Mostly users re-order organic & fruits product.