A Capstone Project Report on

ANALYSIS OF E-COMMERCE DATASET AND BUILDING A RECOMMENDATION ENGINE ON IT

Submitted to

Praxis Business School, Kolkata

(in fulfillment of the requirements for the award of the degree)

Post Graduate Program

In

Data Science

by

Ishaan Nirmal (A21015)

Pawan Thakur (A21020)

Rahul Lamge (A21025)

Shayantam Das (A21029)

Soumyadip Ghatak (A21032)

Under the guidance of

Prof. Dr. Subhashish Dasgupta

Praxis Business School

CELEBRATE YOUR WORTH

Department of Data Science

Academic Year: 2021-2022

Acknowledgment

We are profoundly grateful to **Prof. Dr. Subhasis Dasgupta** Head of Machine Learning & Analytics for his expert guidance and continuous encouragement through out to see that this project meets its target since its commencement until completion.

We would like to express deepest appreciation towards **Prof. Dr. Prithwis Mukerjee** and **Prof. Charanpreet Singh**, Founders & Directors, Praxis Business School, Kolkata. **Prof. Dr. Sourav Saha**, Academics Dean.

Lastly, we would like to express our sincere heartfelt gratitude to all the staff members of Data Science Department who helped us directly or indirectly during this course of work.

Amrita Ghosh

Koushik

Date: 13 May 2022

Contents

ABSTRACT

With the proliferation of Internet services in India as well as around the world, the benefits of conducting commerce over the world wide web has increased significantly, especially over traditional brick and mortar stores. It has reduced the need to rent up physical space for selling goods, levelled the playing field for small business who can compete in the marketplace as well as reducing other operational expenses.

In our Project, we have been given a dataset of an ecommerce site. The dataset contains user id of the user, the number of sessions the customer is logging into the website, as well as the brands the customer is viewing and the thereafter either putting it in cart to purchase or purchase directly. Our job is to gather some insights about the data, see what different factors are influencing the purchasing behaviour of the customer, use time series analysis to forecast trends or systemic patterns within the data and use recommendation engines to help the customer find the most relevant items.

Chapter 1-

Problem Statement

Revenue is one of the most important indicators of how well or poor a company is performing. The revenue of the company depends upon various factors, one of the major factors being sales of the products of the company.

Since sales of the company depends upon various factors like season of the year, discounts/offers given by the company etc. To increase the revenue we need to increase the sales of the company. Nowadays, most of the e-commerce companies have a dedicated team of analysts whose major role is to identify and enhance the sales of products so that revenue targets can be achieved. Various methods can be used for doing this and one of the major methods is using a recommendation engine. One of the main benefits of using the recommendation engine is that it recommends products to each and every user according to their buying behaviour.

Hence our objective in this project was to analyse the dataset of the company and build a recommendation engine on the basis of it. This would help increase the revenue of the company and help decide whether the company is achieving its yearly target or not, and if not then what measures can be taken to achieve its sales/revenue target.

Chapter 2-

Road Map of our Project

For the implementation of or Project, we went through a string of methods which are listed below-

2.1. Method 1-Data Imputation

a.) After getting the dataset we analysed it in google colab and found some null values in them.
b.) Null values need to be imputed accordingly so that there is no major information loss and the accuracy of the recommendation engine is good.
c.) We imputed the null values by searching about the company and its products and then imputing them in the dataset.

2.2. Method 2- EDA

a.) After imputing the null values we analysed the dataset and dropped the rest of the null values for the time being.
b.) Then we performed some EDA on the dataset to gain some insights.

2.3. Method 3- Time Series Forecasting

a.) We performed time series forecasting and analysis on the dataset.
b.) The forecasting was done on a weekly basis and daily basis to forecast the sales on the basis of category of products as well as brands.

2.4. Method 4- Recommendation Engine & Sequential modelling

We have built a recommendation on the dataset on the basis of which the most relevant items on the dataset are recommended to the users or customers and through that revenue or sales can be increased.

Chapter 3-

Data Inspection

3.1. About the Data

We have been given a dataset named “Electronics_item_online_behaviour” which contains 885129 rows and 9 columns or features. The features are **event_time**, **event_type**, **product_id**, **category_id**, **category_code**, **brand**, **price**, **user_id** and **user_session**.

There are around 236219 and 212364 null values in **category_code** and **brand**. These null values are imputed as follows.

3.2. Imputation from various sources

For the null values present, for each respective brand, we find all the products which are sold for that brand and compare their prices with the prices shown either in Amazon, Flipkart, Alibaba or from the webpage of the brand and product itself. We then make an intuitive guess and impute the null values and repeat the process for other brands.

3.3. Challenges from Imputation

We tried using KNN Imputer and Cat-Boost for Imputation of missing or null values. However when comparing with manual imputation, we find that it gives inefficient and wildly inaccurate values, which cannot be used for the project.

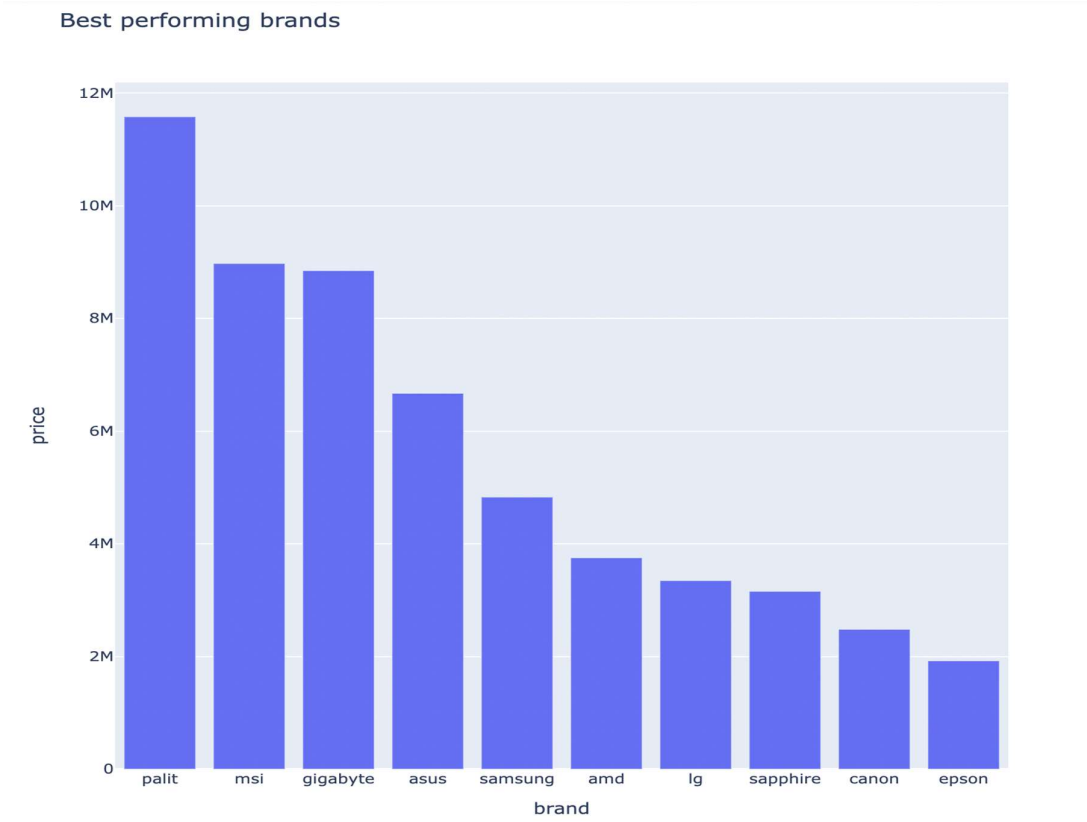
Chapter 4-

EDA

4.1. Insights gained from EDA

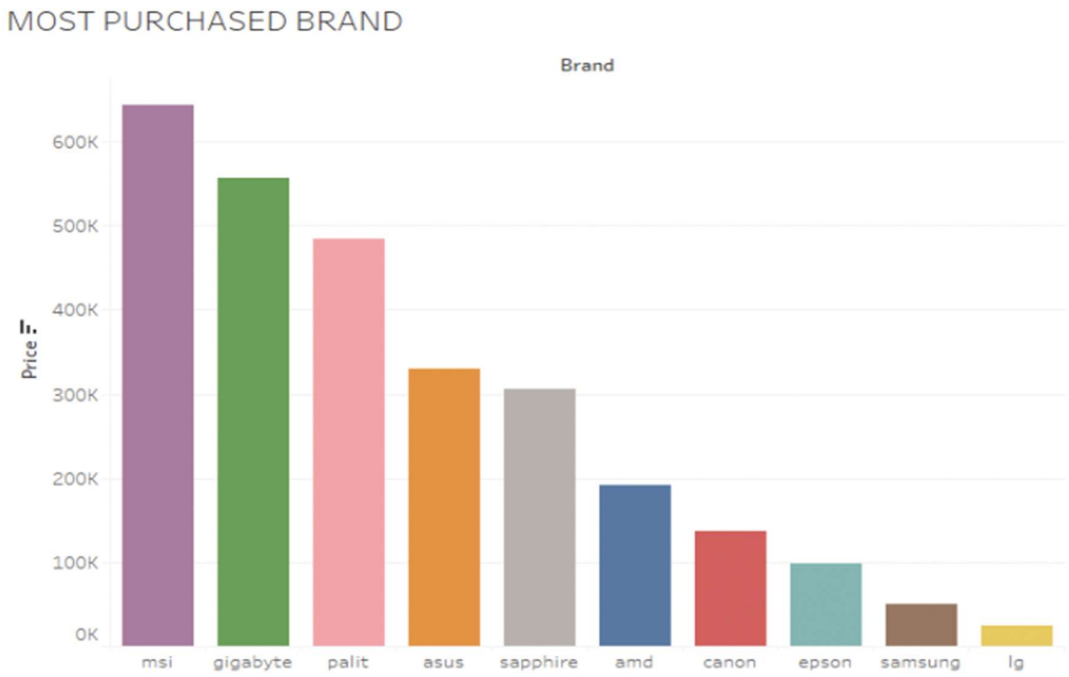
Exploratory Data Analysis was mainly done using Tableau and Python. Some interesting insights were obtained which are discussed below.

- Most Expensive brand



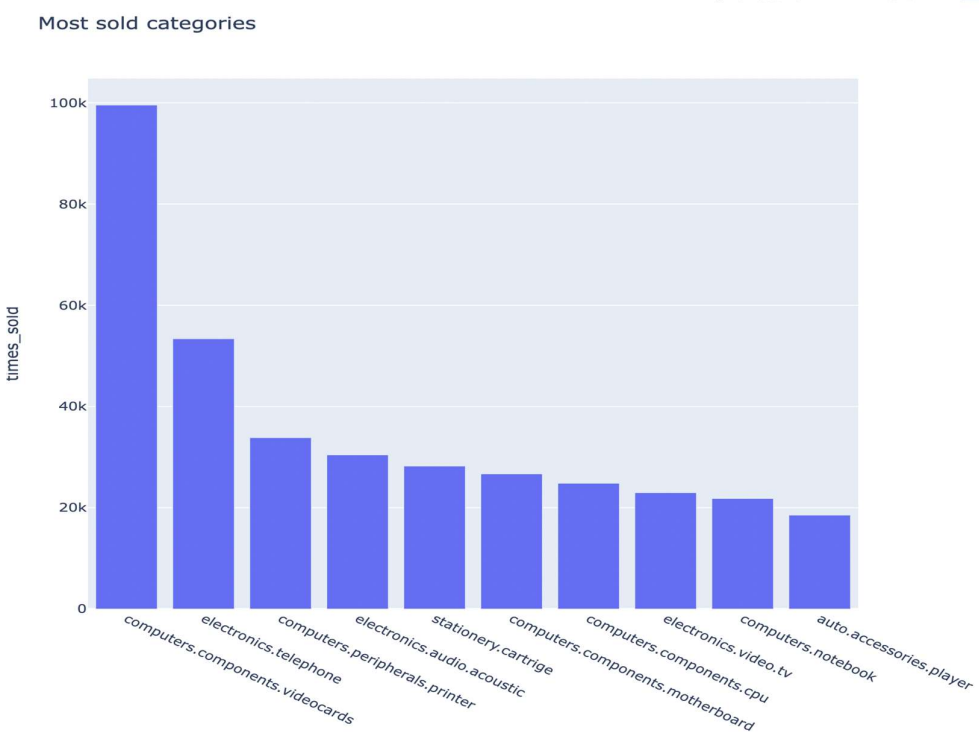
After grouping “brand” and “price” together, we sort them in descending order in order to find out the most expensive brands. We can see that the top 3 brands are “palit”, “msi” and “gigabyte” respectively.

- Most purchased brand



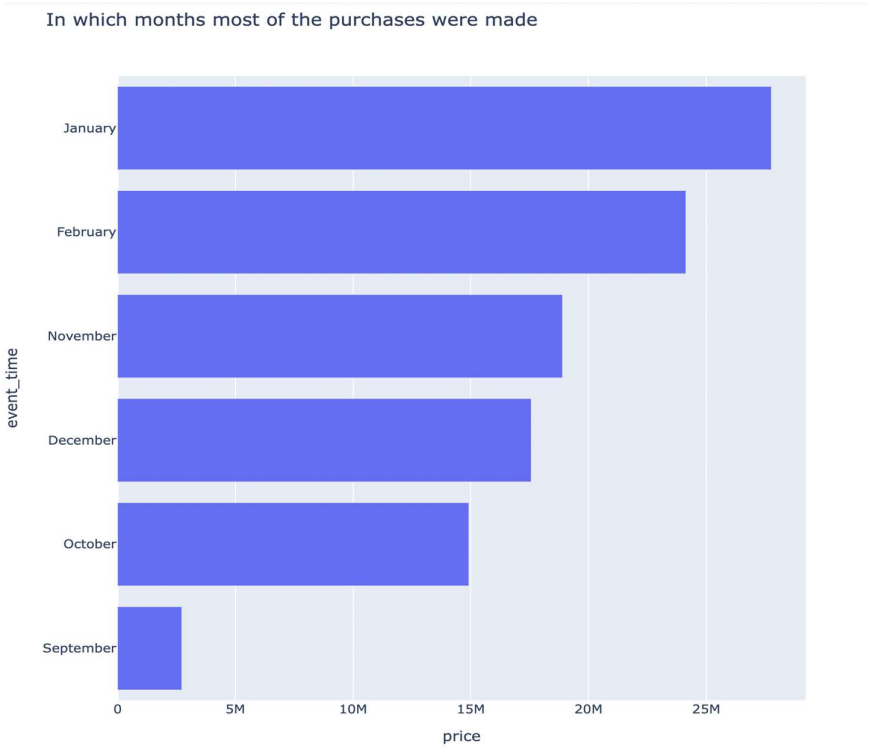
Based on the EDA done, we can clearly see that the top 3 most purchased brands are “msi”, “gigabyte” and “palit” respectively.

- Most sold categories



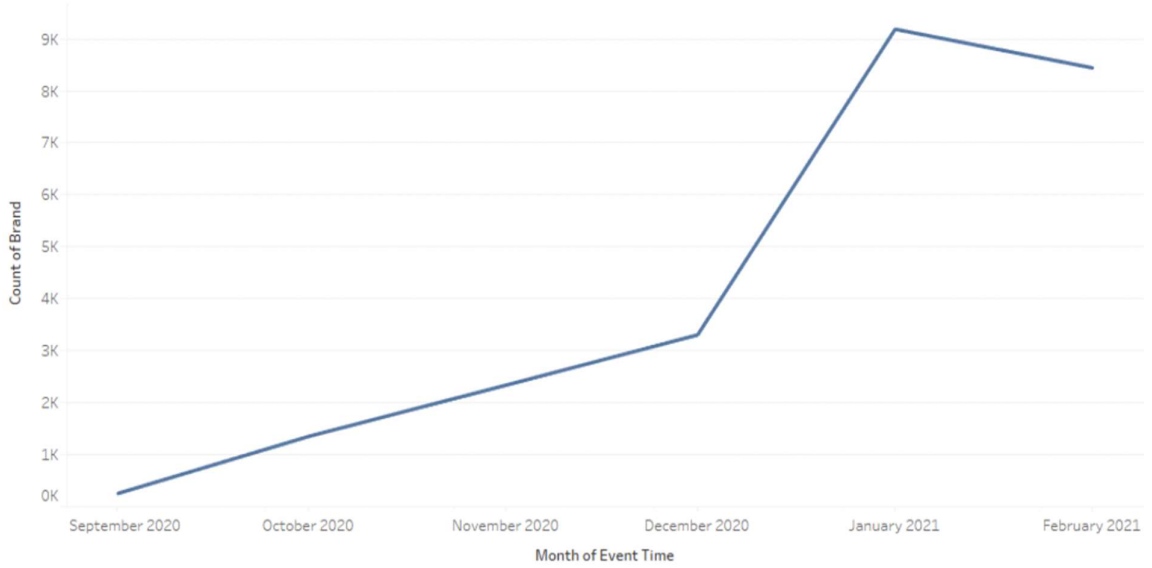
Based on the EDA done, we can see that the most sold items in our dataset is video cards followed by telephone and printer. Concurrently, the least sold item in our dataset are notebook and player (auto accessories).

- Month were most purchases were made



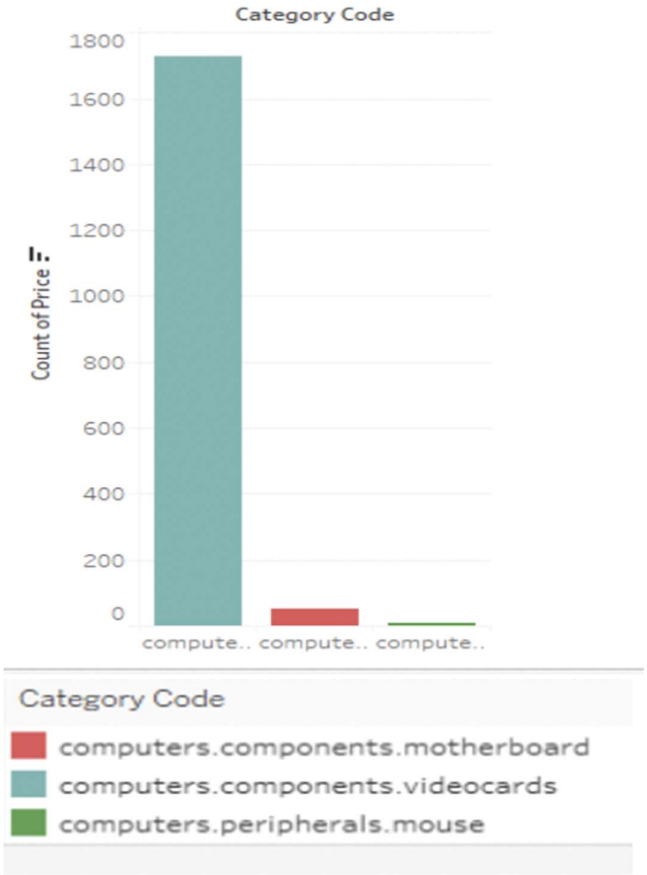
As can be seen from the graph above, we can see that January is the month were most purchases were made but since our dataset consists of only 25 weeks, we can't conclusively say that January is the most purchased in the entire year.

- Most when most purchases were made of top selling brand



From the graph above, we see that January was the month when most purchases were made for brand “msi”.

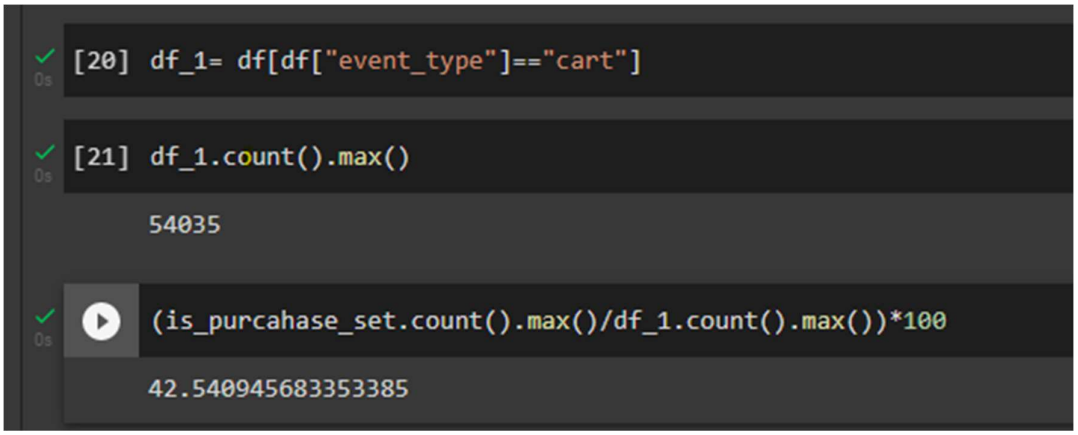
- Most popular category of top selling brand



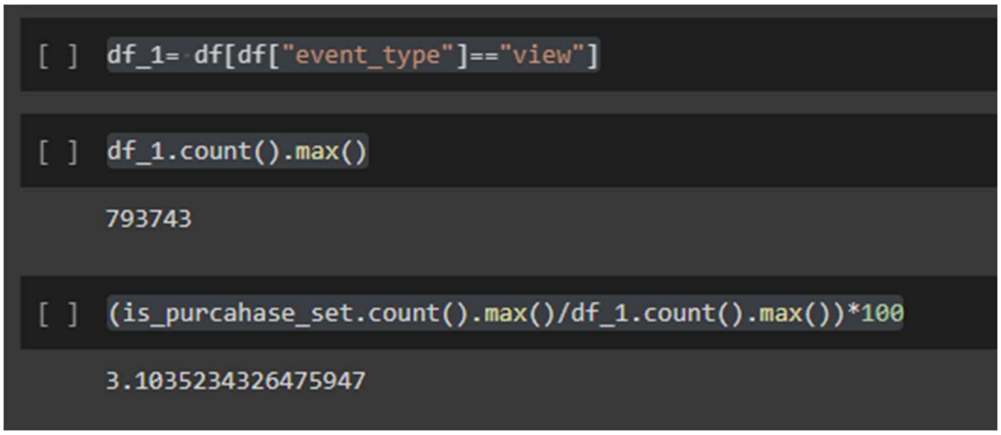
Based on the graph above, it can be inferred that video-cards were the most sold category for best-selling brand “msi”.

- Percentage of items bought from cart to purchase and view to purchase

When the item is transferred from “cart” to “purchase” section, the percentage of items bought is around 42.54%.



When item is directly bought from “view” section to “purchase” section, the percentage of items bought is around 3.10%.



4.2. Challenges faced in EDA

The purpose of doing EDA is to analyse and summarize the data to discover trends or patterns.

One of the main issues while analysing the data is to make sure if outliers which are present don't skew the outcome of the results. The skewness of the results can be also extended to factors such as Data Imputation which can lead to bias.

Chapter 5-
Time series Forecasting

5.1. Overview of Time series forecasting

Time series forecasting is the process of analyzing time series data using statistics and modeling to make predictions and inform strategic decision-making. It's not always an exact prediction, and likelihood of forecasts can vary wildly—especially when dealing with the commonly fluctuating variables in time series data as well as factors outside our control.

In this project we have used ARIMA & SARIMA model

ARIMA model :- takes past values and predicts the future values on its basis

SARIMA model :- it also uses the past values but takes into account the seasonality pattern

In this project we have performed daily time series analysis on the top 5 performing brands

1. Gigabyte
2. Sapphire
3. Palit
4. Asus
5. Msi

Reasons for performing daily analysis:- Since the dataset is very small of approx. 6 months so weekly or monthly analysis was not possible on this as the results were unsatisfactory .

5.2. Methodology of forecasting

In this project first we took the data and checked if the data was stationary or not, the result to it was that the data was not stationary

So, to make the data stationary we had to do differencing, so our flow of forecasting on the top 5 performing brands looked somewhat like this.

Checking the stationarity
↓
Differencing
↓
Applying ARIMA model
↓
Applying SARIMA model

5.3. Insights obtained

1. Since we performed the forecasting on the top 5 performing brands so the order of their ARIMA & SARIMA are as follows

Gigabyte :- ARIMA(0,1,2) , SARIMA(0,1,2,15)

Sapphire :- ARIMA(0,1,2) , SARIMA(0,1,2,31)

Palit :- ARIMA(0,1,1) , SARIMA(0,1,1,31)

ASUS:- ARIMA(0,1,1) , SARIMA(0,1,2,31)

MSI :- ARIMA(0,1,1) , SARIMA(0,1,1,31)

2. We concluded that ARIMA was not giving us satisfactory result as the data consisted up of seasonality & since the AIC score of SARIMA was less than ARIMA so we can say that SARIMA was a better choice of model than ARIMA .

3. In most cases we have found the seasonal order of 31 , so we can conclude that we have monthly seasonality data but since we do not have enough data to support it we are going in with an intuitive decision
4. Residuals were following normal distribution & we also performed the shapiro test & everytime the p>0.05 so we accepted the null hypothesis that the residuals followed normal distribution .

5.4. Challenges faced

1. Data was not sufficient(only 6 months) so couldn't perform weekly or monthly forecasting & in financial terms daily forecast might not be good for decision making .
2. Determining the seasonality order has been done by our own intuition again due to lack of data .

Chapter 6-
Recommendation Engine

6.1. Overview of recommendation engine

A recommendation engine is a kind of data filtering tool using ml algorithms to recommend the most relevant items to the particular user.

It operates on the principal of finding patterns in consumer behaviour data.

6.2. Various recommendation engines made

The various recommendation engine made by us in this project are :-

1. KNN basic
2. KNN with zscore
3. KNN baseline
4. SVD - RMSE (1.49)
5. Baseline only – RMSE(1.49)

The accuracy of all these engines were not at all satisfactory so we had to change our approach from recommendation engine to sequential modelling .

6.3. Insights obtained

There were no real insights obtained from our recommendation engine as the dataset was very small so building the engine was actually a tough task .

6.4. Challenges faced

1. The RMSE value was almost equal to the S.D. of user rating which gives us the idea that the model was not functioning properly .
2. User rating metric were not present there so we had to make it on our own .
3. We formed separate recommendation engines for view , cart & purchase .
4. General recommendation engine was also made for all the three event types still the RMSE was high so the model wasn't good enough .

Chapter 7-

Sequential Modelling

7.1. Overview of sequential modelling

Sequential modelling is the ability of a computer program to model , interpret , & make predictions about or generate any type of sequential data .
In our project the dataset had a time stamp & it was a sequential data so building a recommendation engine wasn't a good idea on it that's why we changed our approach to sequential modelling . We made LSTM model on our dataset .

7.2. Insights obtained

We made two sequential models as follows:-

1. The first one was that if a person is viewing some products so what will be the next event type of the user & building it was an easy task since we had only 3 event types that's why the accuracy was around 90%.

2. The second model was that which next product will the user choose for any sort of action & it was a tough task since there were over 20,000 products so the accuracy was very low approx. 10%.

7.3. Challenges faced

1. Since we know that in our second sequential model the accuracy was around 10% approx. or below that so to overcome or enhance its performance we had to add an embedding layer due to which accuracy went to about 30-40% approx. but since we had hardware constraints so we couldn't train a complex model .

2. Before adding the embedding layer the model was were able to predict only a single product that too with a very poor accuracy .

Chapter 8-
Conclusion & future scope

Conclusion :- After doing this project we concluded that to boost or to advice the committee for decision making regarding how to increase the revenue we could recommend the result obtained by these models for time being as measures and present the forecast which is prepared to give a glimpse that how their sales is gonna look like and how it will affect the revenue , but since the data is of very short duration so we cant expect any significant change in the revenue based on the results or measures recommended by our models .

Future scope:- With respect to the future scope for the time being we can deploy these models for about a year or two and collect the larger duration of data and then repeat the same process on that data & enhance the models in use & we will be able

to forecast even to a further extent resulting to which we can expect some significant changes in revenue .

References :-

Some of the references used in this project are as follows:-

1. https://medium.com/machine-learning-basics/sequence-modelling-b2cdf244c233

2. https://www.kaggle.com/competitions/store-sales-time-series-forecasting

3. https://github.com/topics/lstm-model

4. https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-recommendation-engine-python/