Final Report

Batch details	
Team members	Priyanka Sharma Siddhi Wagh Cherarajan K Fardin Ahmed Kshitij Sharma Aishwarya Patil
Domain of Project	*Social Media and Web Analytics*
Proposed project title	Web Analytics on Pandemic
Group Number	6
Team Leader	Cherarajan k
Mentor Name	Prachi Tare

Business Objective/ Understanding:

1. To understand the traits of Sales and Purchase of medicines during Pandemic.
   And also compare the fluctuations of Sales due to Covid.
   (#1 Pharma Sales Notebook file)

2. To understand why medicine market fluctuations are occurring.
   And to understand the spread of Covid in India.

   (#2 Covid Day wise Notebook file)

3. To analyze the spread of covid cases across each state
   and how could this helpful in understanding the trends and prepare accordingly.

   (#3 Covid State wise Notebook file)

4. Breaking down each state into districts and analyze in District level.
   Taking necessary steps to contain it and ultimately handle situation.
   Also comparing the Normal variant with Delta variant

   (#4 Covid District wise Notebook file)

5. Scraped Data from e-commerce websites for PPE and Self-test kits. Analyzing by using Sentiment analysis on the scraped data to see how good and useful this product is during covid times.

   (#5 Self-Test kits and PPE Notebook file)

Data Description and pre-processing:

1. The final Pharma Sales data set contains Date, Purchase value and Sales value

With this we'll be able to understand the sales and purchase values for top medicine manufacturing companies which contribute greatly to India's medicine market
Purchase value Sales values was in different columns we combined into single purchase and sales column for better understanding. It is a pretty cleansed dataset in it raw form.

2. Day wise data basically contained Date, Daily confirmed, Total Confirmed, Daily Recovered, Total Recovered, Daily Deceased and Total Deceased
   Column names contained different symbols and spacing issues; these were corrected.
   Null value analysis and duplicate data analysis has been done, followed by removal of negative values through capping.

3. State wise data contained Date, Name of the state, Latitude, Longitude, Total confirmed cases, Death, Cured/Discharged, New cases, New Death, New Recovered.
   Here also, we have checked on column names, spacing issues.
   Null value analysis and duplicate data analysis is done. Datatype of columns have also been done. Negative value removal through capping.
   Dropped Longitude, Latitude and New Death columns.

4. District wise data contained Sl No, State code, State, District Key, District, Confirmed, Active, Recovered, Deceased, Migrated/Other, Delta Confirmed, Delta Active, Delta Recovered, Delta Deceased, District Notes and Last Updated.
   Removed District Notes and Last updated. Same Data preprocessing just as State wise data as most of the columns are common

5. Data has been scraped from E-commerce websites such as Amazon and Flipkart using Beautiful Soup and Requests. It contained columns like Customer name, Customer rating, Review date and Customer review.
   Using libraries such as RE, NLTK special character removal has been done followed by finding sentiment polarity using Text Blob.

Data Preparation:

After pre-processing all the datasets, we extensively checked again for Null values and Duplicate values and treated them.

We are not considering outliers in this dataset as each and every entry is extremely important.

We have done descriptive analysis that helps in summarizing the datapoints.

Exploratory Data Analysis & Business Insights:

We are checking the trend of medicine sales through scatter plot. With that we are able to visualize the fluctuation in medicine market in comparison with Dates.

We have used the medicine to connect and compare with covid cases in India and relate accordingly.

In our Covid data we have set our target variable as Total confirmed case and how other features are contributing towards.

We have also checked relationships between target variable and other features and also correlation within features by using Scatter plot and heatmap.

We have also used bar plots to find how top features are contributing towards the target variable.

Basic Model:

We have used OLS as our basic model which extensively helps in understanding how each feature contributes towards our target variable.

Moreover, it helps in getting Business Insights for our objective of the Project.

For extra information: refer Notebook file attached

Feature engineering & feature extraction:

Feature Engineering and Feature extraction has been done on our covid related datasets. As this helps is reducing the variance between the features to improve the accuracy of the model through bias variance trade off.

Numerical columns which had high variance has been treated through various transformation techniques such as log transformation and square root transformation.

Hyper Parameter Tuning:

By using Random Forest and Elastic net regression models with general parameters, we were to get the best parameters for each model.

By using Grid search CV, we were able to find the best tuned hyper parameters for the model.

By applying this fine-tuned parameter, we were able to achieve better efficiency for our model.

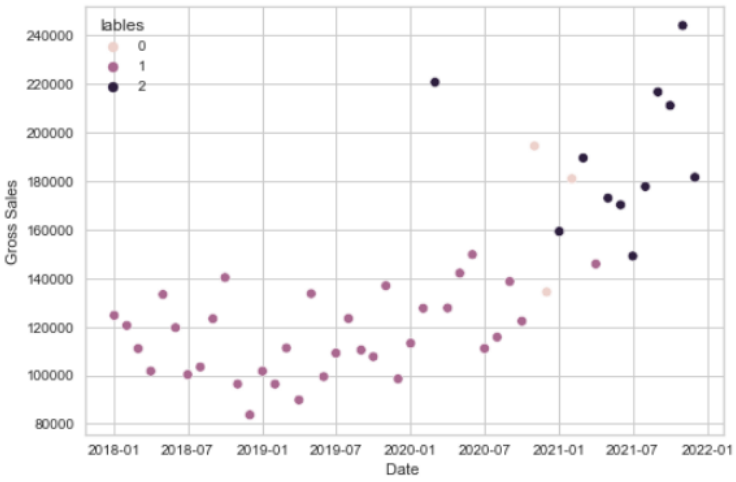
Comparison and Selection of model:

As our objective of the project is more inclined Inferential statistical analysis, we prefer OLS model for our inferences.

By comparing Random Forest and Elastic net model in Covid dataset, we are getting high bias and low variance in Elastic net whereas low bias and low variance with Random Forest Regressor. So, we prefer Random Forest.

Inferences, Results and Discussion:

1. Pharma sales dataset:

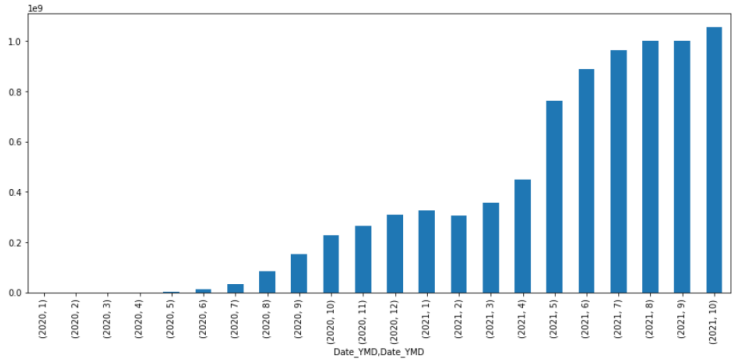


By finding the cluster we were able to observe the patterns in the dataset. Here, we can see that the medicine sales is quite uniform till 7th month of 2020, and sales increases rapidly after that.

We can say that the reason for the fluctuation is Covid.

Let's relate covid dataset along this one.

2. Covid Day wise dataset



We can see that covid started by 5th month of 2020 and when we see that dates on medicine dataset, it is very clear that sales started spiking after covid.

We did some statistical analysis on daily cases and found out the average cases ranged between 57k to 60k cases per day during peak time in India.

3. Covid State wise dataset

We found which states are most affected and least affected by covid through OLS model and various plots.

Top contributing States were found.

[Maharashtra, Tamilnadu, Andhra Pradesh, Karnataka, Delhi]

By knowing which state is most affected →

The central government can concentrate more on the state with high cases by implement laws regarding Curfew, Workforce Distribution and also maintain supply and demand of medicines within states.

4. Covid District wise dataset

Now let's break down States into districts, we are considering the top contributing state of Maharashtra.

Top contributing Districts are

[Mumbai, Pune, Thane, Raigad, Nashik]

By knowing the top contributing Districts, the state government can take various actions and get better prepared to face the pandemic and ultimately containing the cases in India.

5. Comparing Delta Deceased vs Delta confirmed and Normal Deceased vs Normal Confirmed, we are able to see that Delta cases are more deadly than Normal.

6. Sentiment Analysis on PPE and Self-test kits:

   During this pandemic, common people believe in products like these to test themselves at home for covid. And we know False Negative is very dangerous in this scenario.

   We can see that Self-test and PPE kits are having average rating (through Polarity) i.e.) not good up to the point.

   Government can implement strict laws to maintain quality of such products as these will greatly help in reducing the spread of covid cases.

Description Criterion

We can see, how web analytics can be greatly useful during situation like Covid. If we start collecting and organizing data from start and analyze properly, we can get some more information also.

REFERENCES

- Blogs, articles and social media news relevant to project.
- Kaggle and Github for Covid related data
- PPE and Self-test kits – Scraped from E-commerce websites like Amazon
- PharmaSales – Moneycontrol and Statista websites