SUMMARY

ON

CORONA VIRUS TWITTER SENTIMENT ANALYSIS

ALMABETTER CAPSTONE PROJECT

-By Yamini Peddireddi

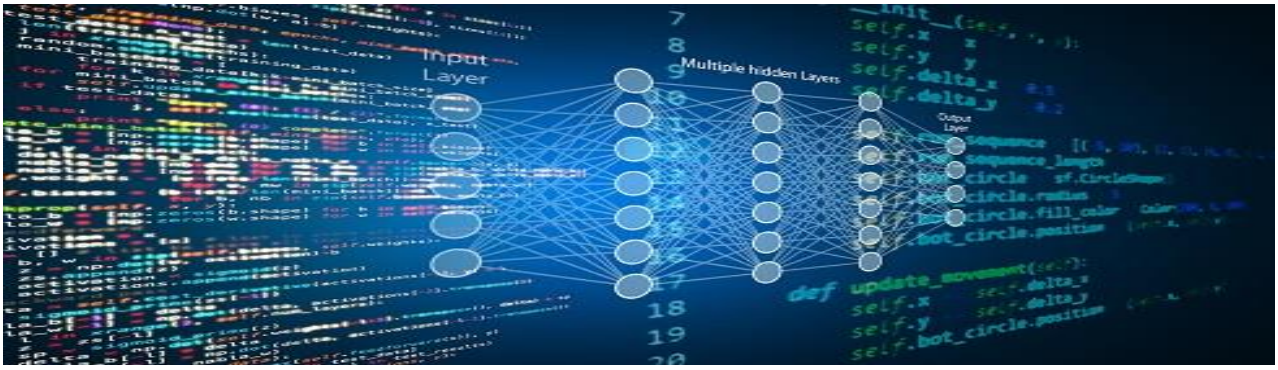
1: Introduction

1.1. Data Science

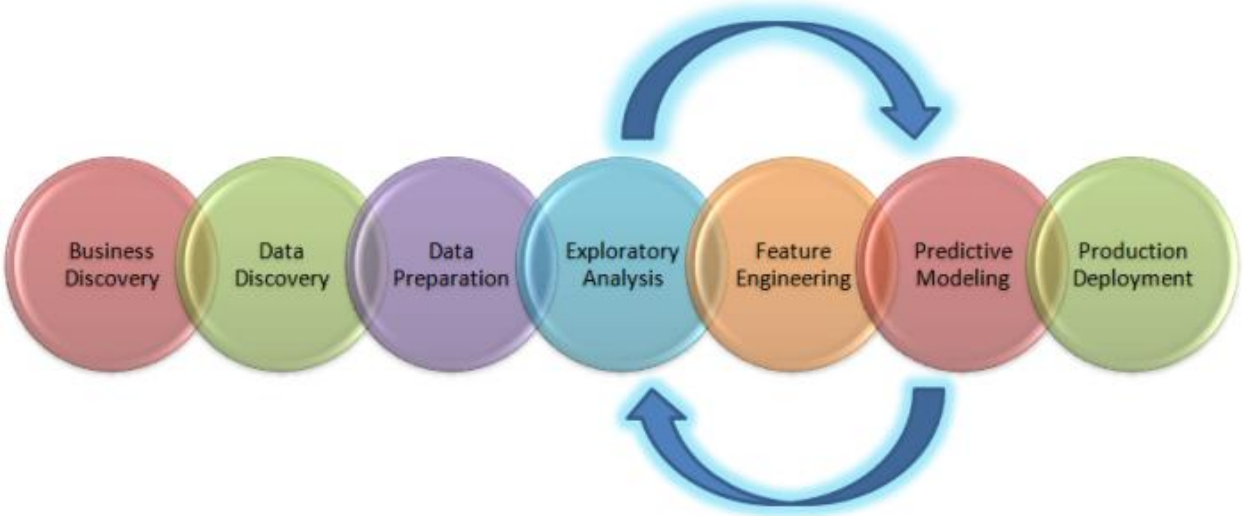
Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.

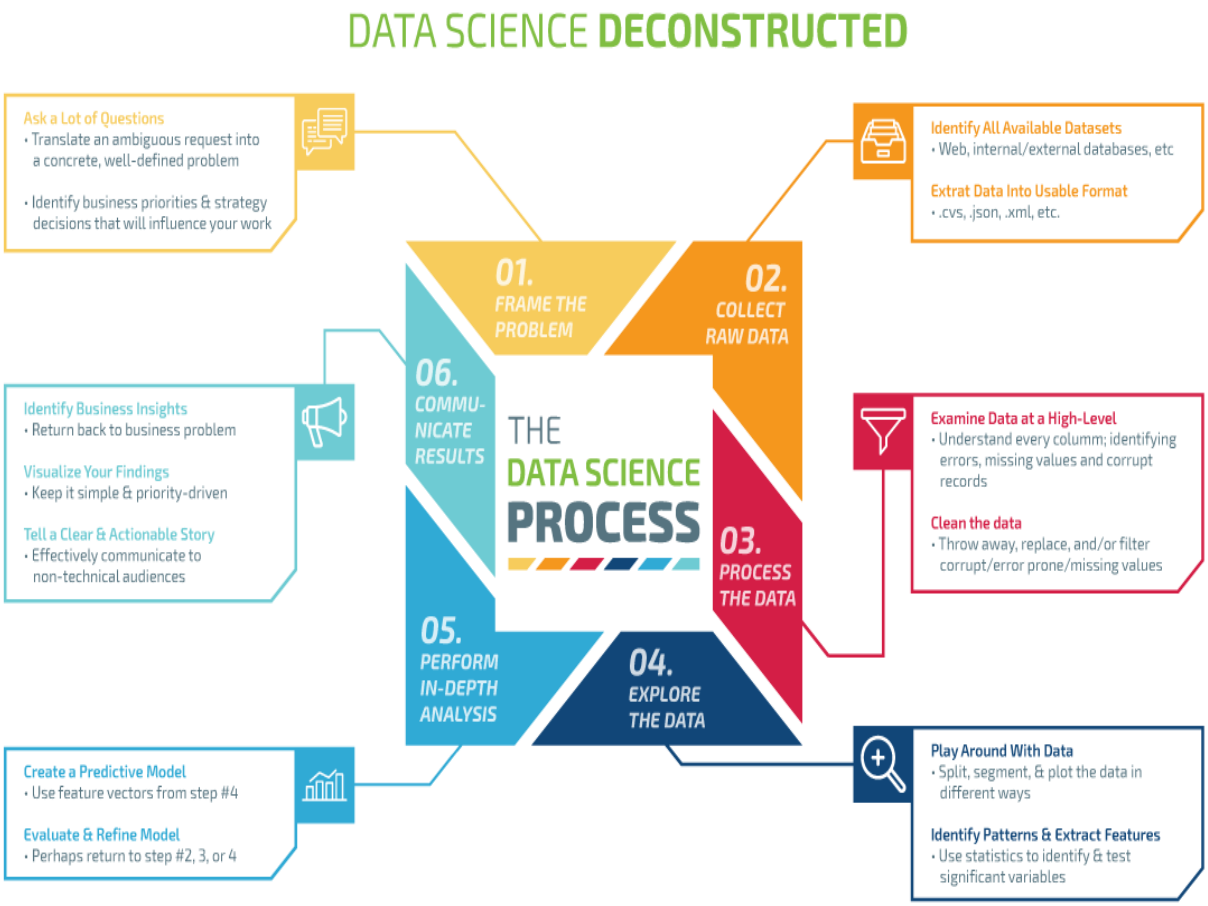
Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science. Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

Big data is very quickly becoming a vital tool for businesses and companies of all sizes. The availability and interpretation of big data has altered the business models of old industries and enabled the creation of new ones. Data-driven businesses are worth $1.2 trillion collectively in 2020, an increase from $333 billion in the year 2015. Data scientists are responsible for breaking down big data into usable information and creating software and algorithms that help companies and organizations determine optimal operations. As big data continues to have a major impact on the world, data science does as well due to the close relationship between the two.



1.2. Problem solving





2: Methodology/Approach in solving

2.1. Problem Statement

This challenge asks you to build a classification model to predict the sentiment of COVID-19 tweets. The tweets have been pulled from Twitter and manual tagging has been done then. The names and usernames have been given codes to avoid any privacy concerns.

Attributes in the data:

1. Location - Name of the place from where tweet was made
2. Tweet At - Tweeted date, month and year
3. Original Tweet - The tweet made by user
4. Label - Sentiment of the tweet

2.2. Steps to problem solving

According to the problem statement, it is clear that the problem here is to help in making prediction on the Twitter Sentiment. So as there are many client data given from which there is a need to create or design a system through which the client is able to easily make predictions based on tweet made on Twitter, Here the goal is to create a model to predict the sentiment of the tweet whether it is positive, negative, neutral, extremely positive and extremely negative.

The steps in create such a model or the steps in solving this problem are the following:

- Problem statement
- Missing value analysis
- Data Cleaning
- Data exploration and Data Analysis
- Data Visualization
- Feature Engineering and Selection
- Splitting the data into train and test set
- Dealing with imbalanced set
- Model building and hyperparameter tuning
- Model Validation using evaluation metrics
- Model Selection

2.3. Challenges

The challenges in creating the model are

- Data collection

- Data limitation
- Privacy and security
- Data manipulation
- Missing values or vague data
- Chance of misinterpretations
- Feature selection and the good number of features unknown as there are many tweets with many words in it.
- Memory crash problem
- Hyperparameter tuning takes long time and sometimes kernel dies.
- Computation time and cost

3: Data understanding and analysis

3.1. Description of the data

	UserName	ScreenName
count	41157.000000	41157.000000
mean	24377.000000	69329.000000
std	11881.146851	11881.146851
min	3799.000000	48751.000000
25%	14088.000000	59040.000000
50%	24377.000000	69329.000000
75%	34666.000000	79618.000000
max	44955.000000	89907.000000

3.2. Features in the data

The training data contains 41157 rows and 6 columns. This dataset contains 2 numerical columns and 4 categorical variables. The first two numerical variables cannot be considered for analysis as this is unique identifier number for each user using Twitter. User name describes the ID of the twitter which is in encoded format for security and privacy reasons and the same goes

for Screen name which is the different for every user. So, these variables cannot be used for analysis.

The other variables like Tweet at, Tweet, Location and Sentiment are the categorical variables which are used for further analysis. Sentiment is the target variable or controlled variable or dependent variable.

3.3. Data cleaning

- The data is further explored to check if there are any missing values as those rows may not be useful in the prediction.
- In the given dataset there are missing values in the Location variable. The tweet column being the text data it has been cleaned by removing punctuations and stop words from it so that tweet column is good for further processing after removing the redundant words from the covid tweets.
- After data cleaning I have created new features from cleaned tweet column using Tfidf vectorizer which extracts important words from the text columns based on the parameter given to it like min_df and max_df which helps in finding the words which are few in number in all the documents and most repeated words in all the documents.
- Now all the features can be used for further processing and analysis.

3.4. Data Exploration and analysis

```
tweet_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41157 entries, 0 to 41156
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype         
---  ------          --------------  -----         
 0   UserName        41157 non-null  int64         
 1   ScreenName      41157 non-null  int64         
 2   Location        41157 non-null  object        
 3   TweetAt         41157 non-null  datetime64[ns]
 4   OriginalTweet   41157 non-null  object        
 5   Sentiment       41157 non-null  object        
 6   Year            41157 non-null  int64         
 7   Month           41157 non-null  int64         
 8   Day             41157 non-null  int64         
dtypes: datetime64[ns](1), int64(5), object(3)
memory usage: 2.8+ MB
```

```
tweet_df.describe()
```

	UserName	ScreenName	Year	Month	Day	length
count	41157.000000	41157.000000	41157.0	41157.000000	41157.000000	41157.000000
mean	24377.000000	69329.000000	2020.0	4.333673	15.080399	204.200160
std	11881.146851	11881.146851	0.0	2.488591	8.460537	68.655129
min	3799.000000	48751.000000	2020.0	1.000000	4.000000	11.000000
25%	14088.000000	59040.000000	2020.0	3.000000	4.000000	151.000000
50%	24377.000000	69329.000000	2020.0	3.000000	18.000000	215.000000
75%	34666.000000	79618.000000	2020.0	5.000000	22.000000	259.000000
max	44955.000000	89907.000000	2020.0	12.000000	31.000000	355.000000

3.5. Feature engineering and selection

- There are different types of encoding which can be done to the data. They are Label Encoding or Ordinal Encoding, one hot Encoding, Dummy Encoding, Effect Encoding, Binary Encoding, BaseN Encoding, Hash Encoding and Target Encoding etc.
- Here I am using Hash encoding
- As One hot encoding is giving almost 10000 columns having high cardinality so I'm taking Hash encoding values for further processing.
- After using all the types of feature engineering techniques, I have created good number of features which can be used for analysis and model building.

4: Handling imbalanced data

4.1. Techniques of handling data

Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes which have number of instances. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class.

Evaluation of a classification algorithm performance is measured by the Confusion Matrix which contains information about the actual and the predicted class.

However, while working in an imbalanced domain accuracy is not an appropriate measure to evaluate model performance**.** For eg: A classifier which achieves an accuracy of 98 % with an event rate of 2 % is not accurate, if it classifies all instances as the majority class. And eliminates the 2 % minority class observations as noise.

Examples of imbalanced data

Thus, to sum it up, while trying to resolve specific business challenges with imbalanced data sets, the classifiers produced by standard machine learning algorithms might not give accurate results. Apart from fraudulent transactions, other examples of a common business problem with imbalanced dataset are:

- Datasets to identify customer churn where a vast majority of customers will continue using the service. Specifically, Telecommunication companies where Churn Rate is lower than 2 %.
- Data sets to identify rare diseases in medical diagnostics etc.
- Natural Disaster like Earthquakes

Resampling Techniques

Dealing with imbalanced datasets entails strategies such as improving classification algorithms or balancing classes in the training data (data preprocessing) before providing the data as input to the machine learning algorithm. The later technique is preferred as it has wider application.

The main objective of balancing classes is to either increasing the frequency of the minority class or decreasing the frequency of the majority class. This is done in order to obtain approximately the same number of instances for both the classes. Let us look at a few resampling techniques:

- Random Under Sampling
- Random Over Sampling
- Cluster-based Over Sampling
- Informed over sampling: SMOTE
- MSMOTE

4.2. Applied method and balanced

For this covid-19 twitter dataset I have applied SMOTE (Synthetic minority oversampling technique). This technique is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset. The new dataset is used as a sample to train the classification models.

- **Advantages**
  - Mitigates the problem of overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances
  - No loss of useful information

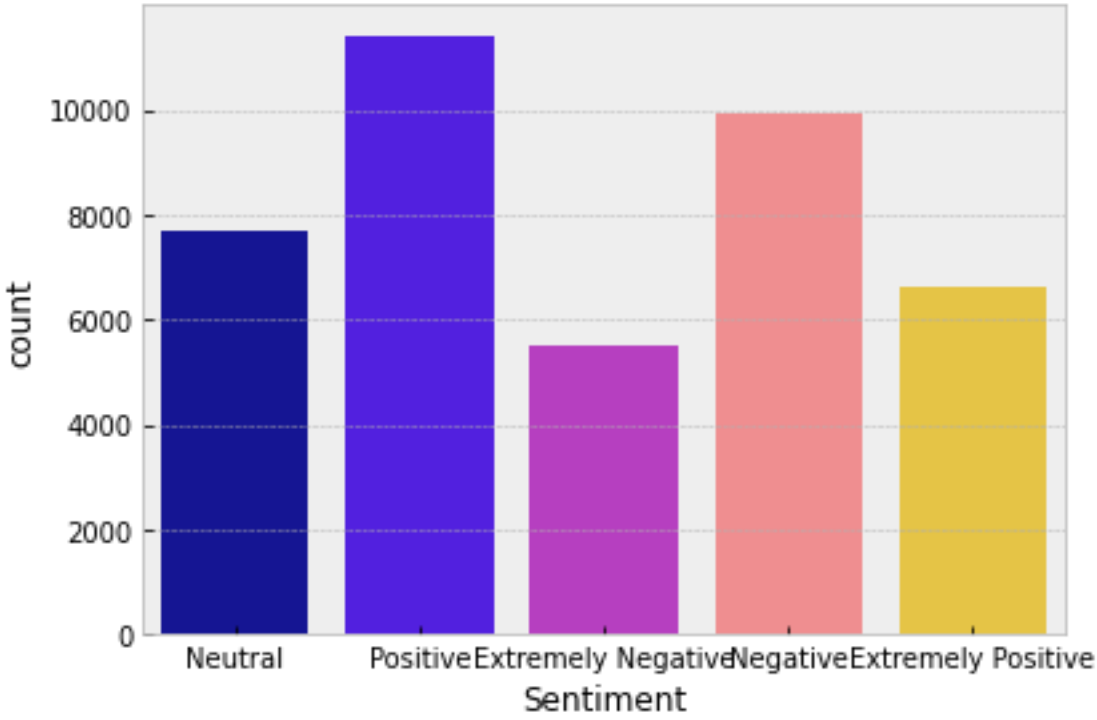
- **Disadvantages**
  - While generating synthetic examples SMOTE does not take into consideration neighboring examples from other classes. This can result in increase in overlapping of classes and can introduce additional noise
  - SMOTE is not very effective for high dimensional data

4.3. Balancing the dataset

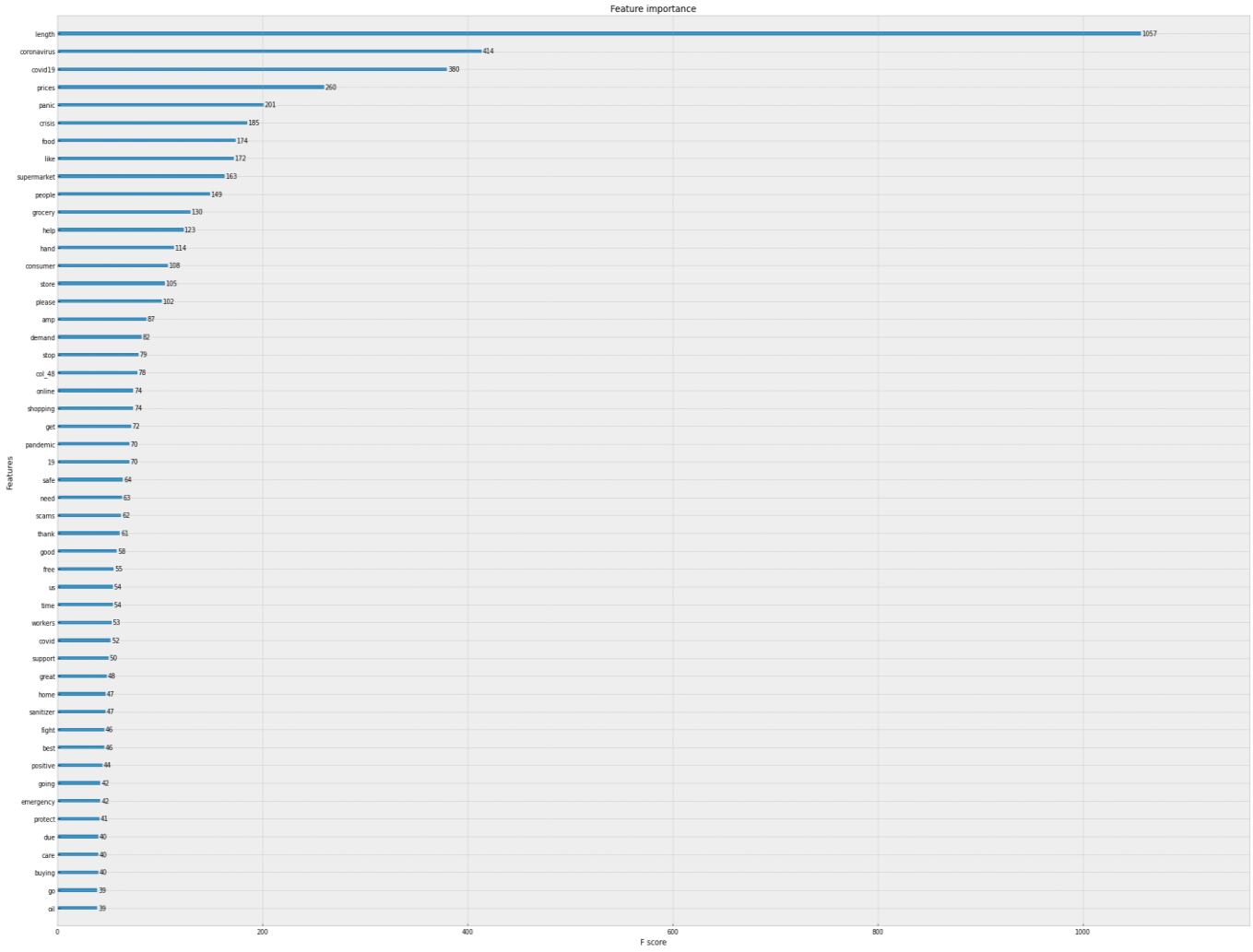
- In the target column Sentiment, there are five different types of classes which has to be classified or identified when this dataset with independent variables is passed.
- As there are more than two classes it is called multi-class classification problem.
- It can be seen even though there are different classes in the target label, they are not equally distributed or there is an imbalance in the dataset.
- It's not advisable to pass the same dataset as model may give more importance to majority classes and ignore the minority classes.
- So, I have used SMOTE to solve the imbalance data problem which generates synthetic values to the minority classes and it is one of the oversampling methods used to balance the target class.

5: Data visualization

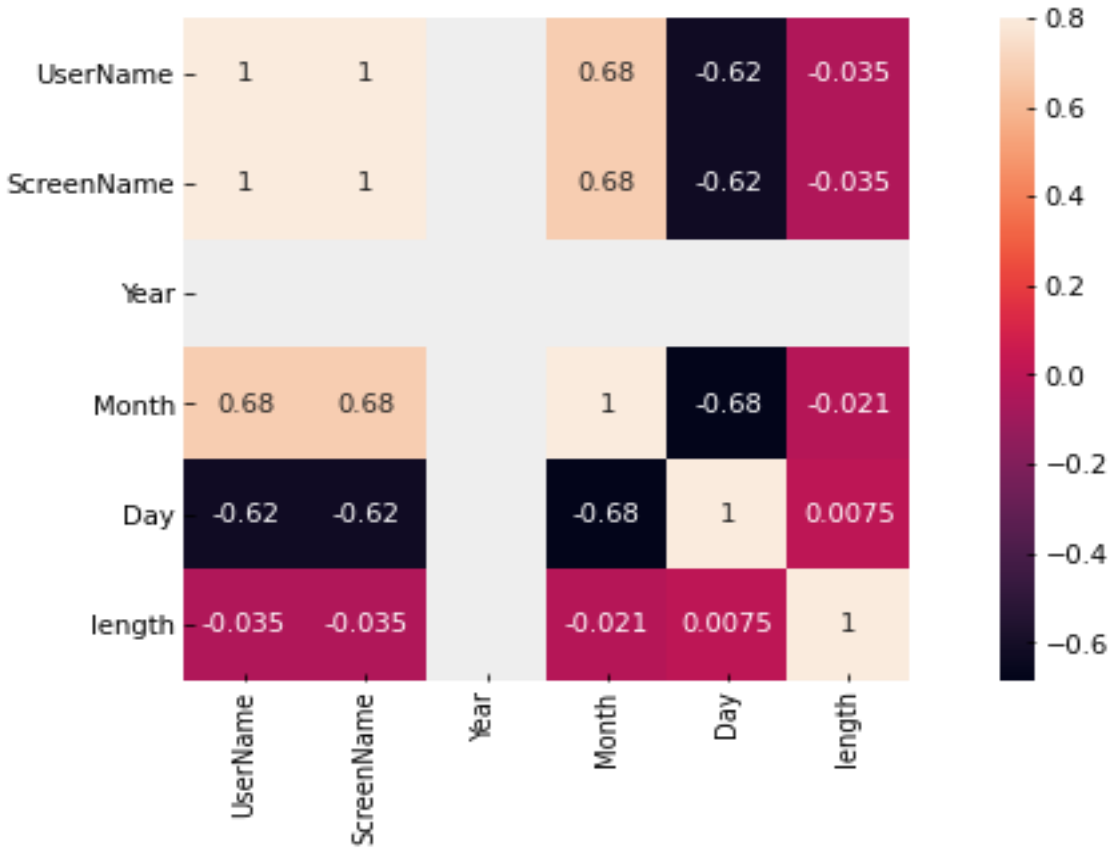
5.1. Plots and insights



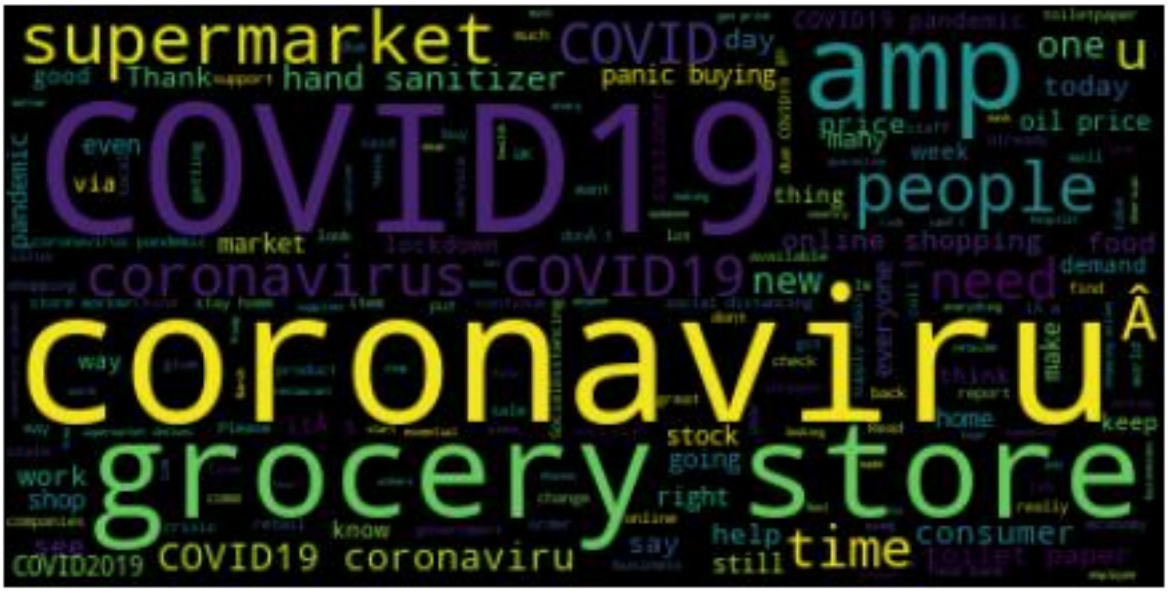
- From the above plot it can be seen that there are more Positive Sentiment values which says the data is imbalanced and the data cannot be used as it is because there is a chance of the predictions more on majority class itself and also overfitting and under fitting can happen.
- It also says that almost the Positive Sentiment classes are double in the number compared to minority class Extremely Negative Sentiment tweets.
- Hence this type of data in target column is called imbalanced dataset which should be converted to balanced dataset for accurate predictions otherwise there is a chance of misclassification.



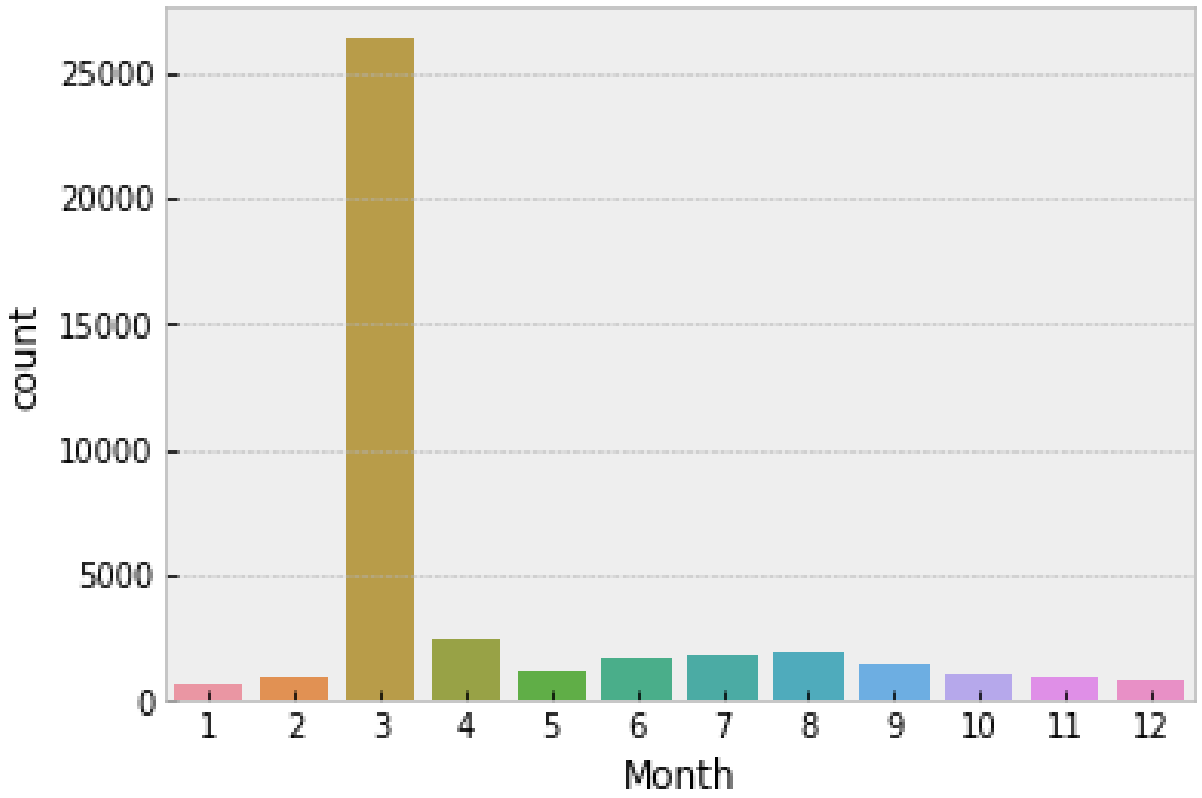
- The above plot can be used for knowing the top 50 features out of the 1500+ features obtained after splitting the data into train and test split.
- This plot is obtained using the XGBoost Classifier which is the one of good performing model for this dataset in making the accurate predictions then not the best model.



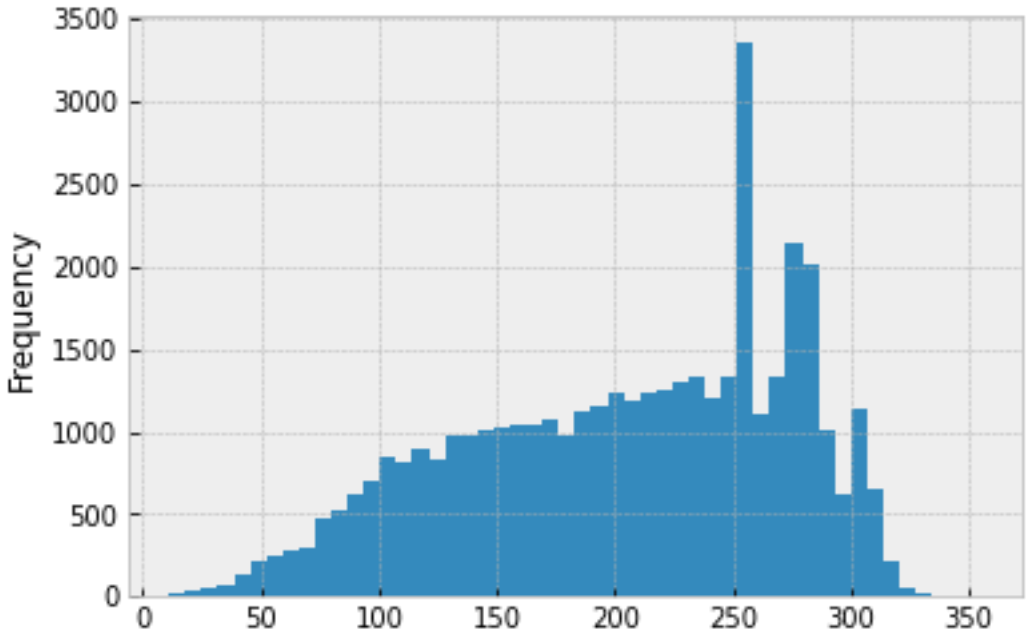
- From the above it can be seen that there is very less correlation or no correlation between the independent features. So all the variables can be used for model building.



- The important terms from tweets which are used multiple times by different users are Corona virus, Covid19, grocery store, amp, people, time, panic buying, hand sanitizer, work, market, price, thing, one, right, stock, going, work ..etc.
- There are different sentiments involved in these tweets even though the words are almost the same and people yelling on the pandemic and about provisions provided due to this hard situation. So, there are a mix of extremely negative to extremely positive tweets happened mostly in the month of March.



- There are more tweets in the month of March than on any other month may due to the newly aroused covid crisis and when user's situations have turned bad unexpectedly.



- From the histogram of length of tweets in the data, it can be seen that there are more tweets of length 250-255 and then from 260-275. Also, we can say that this follows left tail distribution.

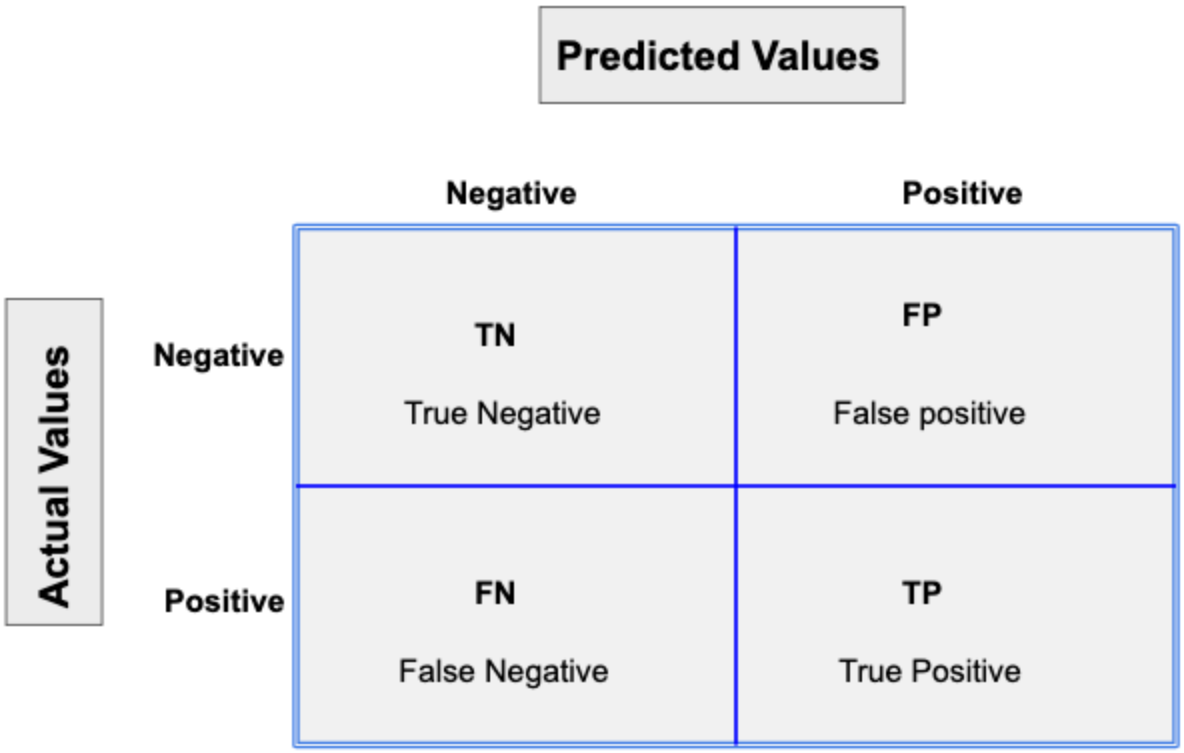
6: Model evaluation and selection

6.1. Model evaluation metrics

The evaluation metrics for evaluating the performance of a machine learning model, which is an integral component of any data science project. It aims to estimate the generalization accuracy of a model on the future (unseen/out-of-sample) data.

A confusion matrix is a matrix representation of the prediction results of any binary testing that is often used to describe the performance of the classification model (or “classifier”) on a set of test data for which the true values are known.

A confusion matrix is a performance measurement method for Machine learning classification. It helps you to know the performance of the classification model on a set of test data for that the true values and false are known. It helps us find out, how many times our model has given correct or wrong output and of what type. Hence, it is a very important tool for evaluating classification models for balanced datasets.

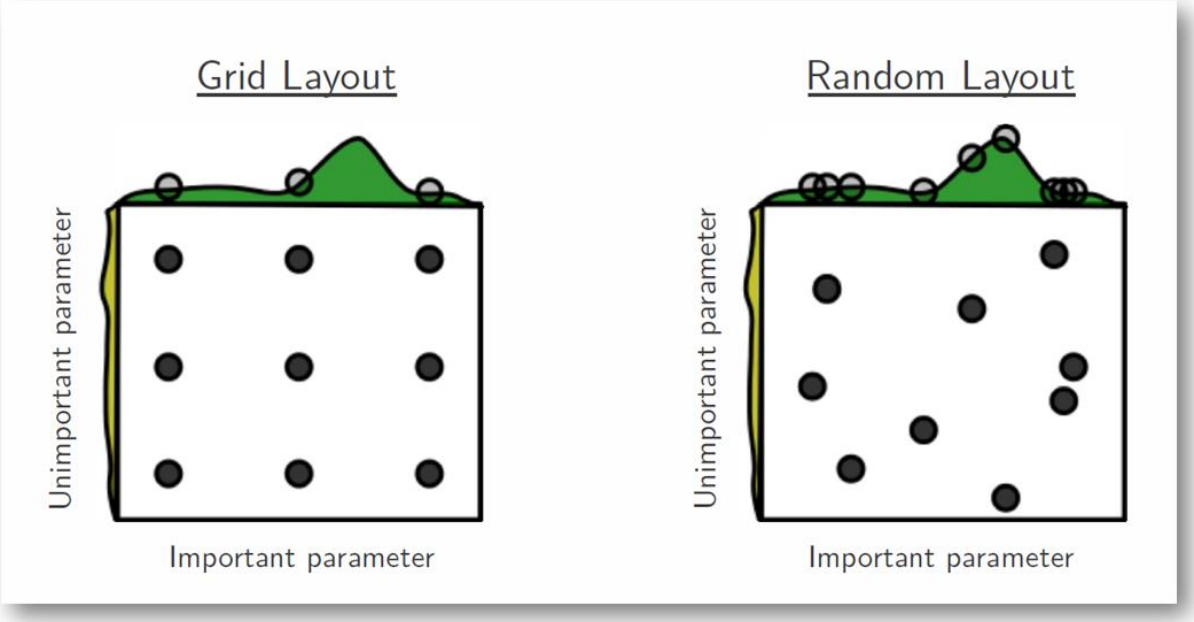


Understanding how well a machine learning model is going to perform on unseen data is the ultimate purpose behind working with these evaluation metrics. Metrics like accuracy, precision, recall are good ways to evaluate classification models for balanced datasets, but if the data is imbalanced and there's a class disparity, then other methods like ROC/AUC, Gini coefficient perform, F1-score are better in evaluating the model performance.

- As described evaluation of classification problems can be made based on accuracy score and from confusion matrix many of the metrics can be obtained like accuracy, precision, recall, sensitivity, specificity, f1 score etc. Classification report describes well about all these metrics.

6.2. Hyper parameter tuning

- There are many techniques through which hyperparameter tuning can be performed like Grid Search CV, Randomized Search CV and Bayesian Optimization etc
- Now I have used Grid Search CV for tuning the parameters and getting the best parameters and reconstructed the model with those parameters and have check the grid best score and cross validation score which gives best parameters.

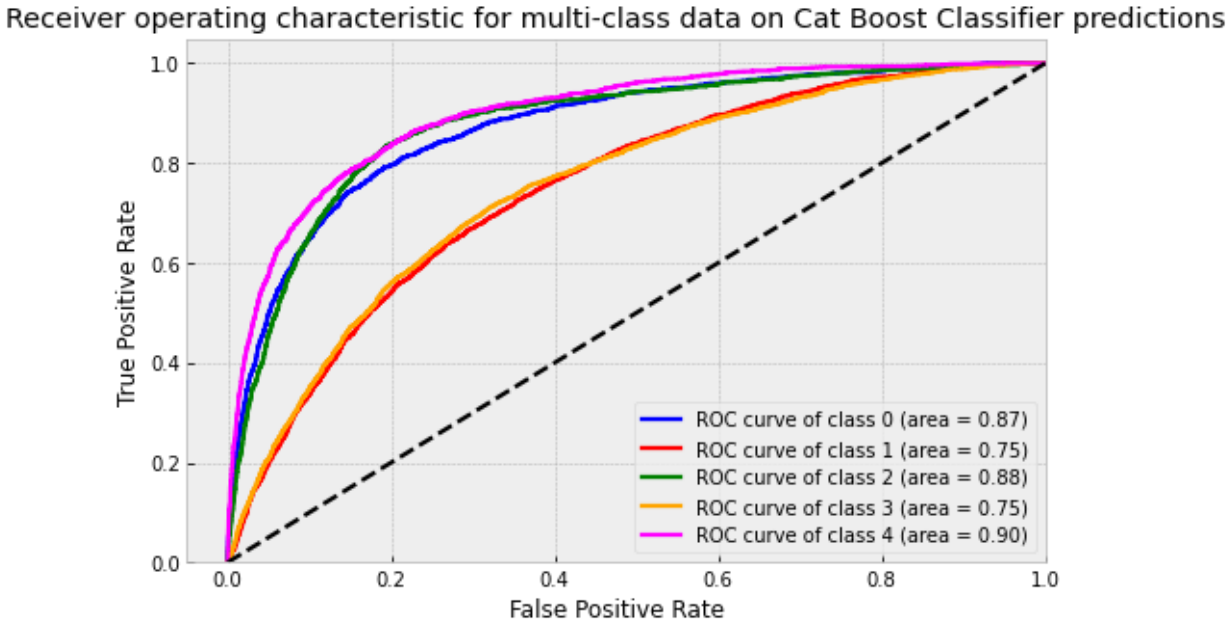


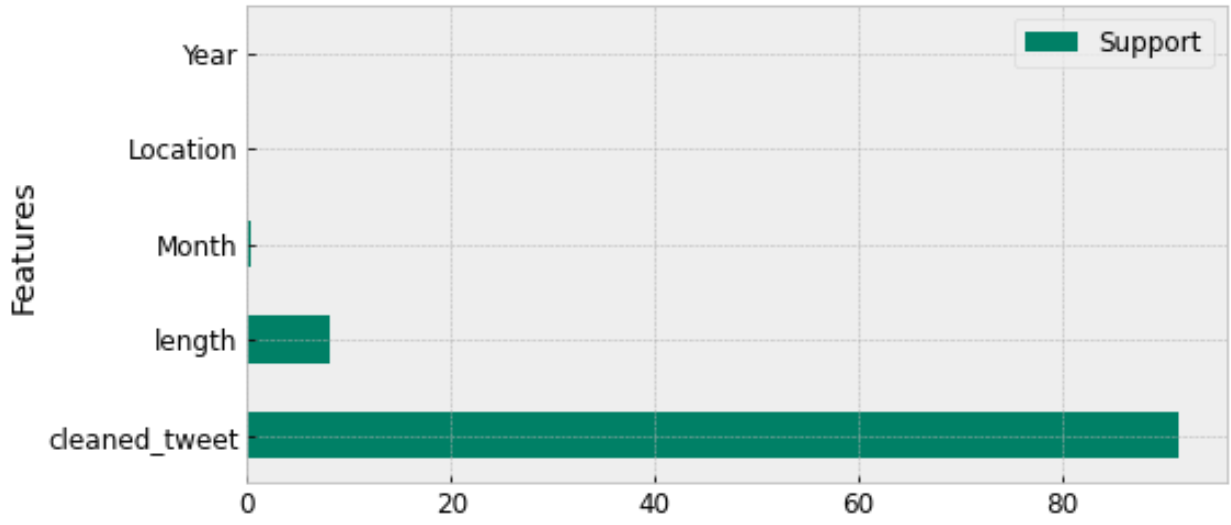
6.3. Rebuilding the model

- The models used for reconstruction with best parameters obtained after hyperparameter tuning are XG Boost, Random Forest and Cat Boost Classifiers.

6.4. Model selection

- I used some of the good performing models in grid search construction and after rebuilding with all those models with best parameters. I again calculated the evaluation metrics for all these models and found that XG Boost improves after hyper parameter tuning and then Cat Boost remains best performer for this dataset.


Receiver operating characteristic for multi-class data on Cat Boost Classifier predictions
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.87)
ROC curve of class 1 (area = 0.75)
ROC curve of class 2 (area = 0.88)
ROC curve of class 3 (area = 0.75)
ROC curve of class 4 (area = 0.90)


Support
Features
Year
Location
Month
length
cleaned_tweet
0
20
40
60
80

	Models	Accuracy	Roc_auc_score	Avg_precision	F1_score
0	Decision Tree	0.404519	0.630663	0.187193	0.408390
1	Random Forest	0.516926	0.823481	0.211746	0.523336
2	Multinomial Naive Bayes	0.469307	0.793574	0.199421	0.477305
3	KNN	0.214480	0.555297	0.189209	0.201310
4	XGB	0.536605	0.829632	0.214075	0.542512
5	Catboost	0.545999	0.835100	0.096942	0.548660

- As Cat Boost is best model for this dataset according to the above scores as it is also able distinguish all the classes and gave good predictions for this multi-class classification problem.
- Now I consider Cat Boost as the best model for this project on Corona virus twitter sentiment predictions and which helps in identification of sentiment based on covid-19 tweets.

Chapter 7: Case studies

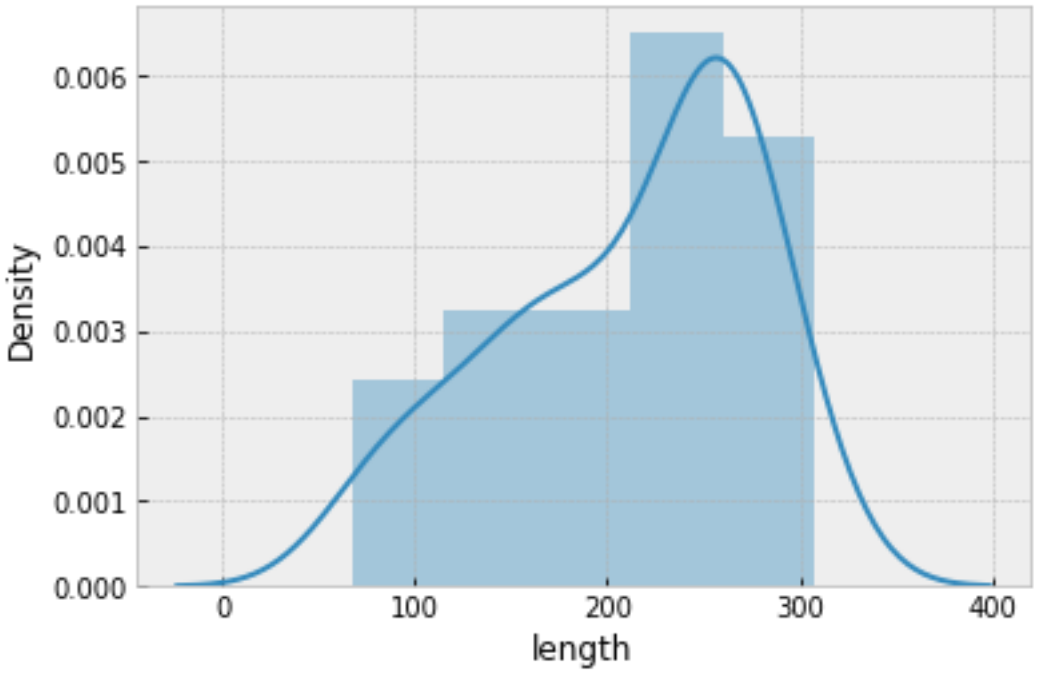
1. Finding the tweets which are positive from Canada location as they are highest in length of tweets from the sample findings

```
locCnSenP.shape
```

```
(51, 10)
```

```
locCnSenP['cleaned_tweet'].tolist()[5:10]
```

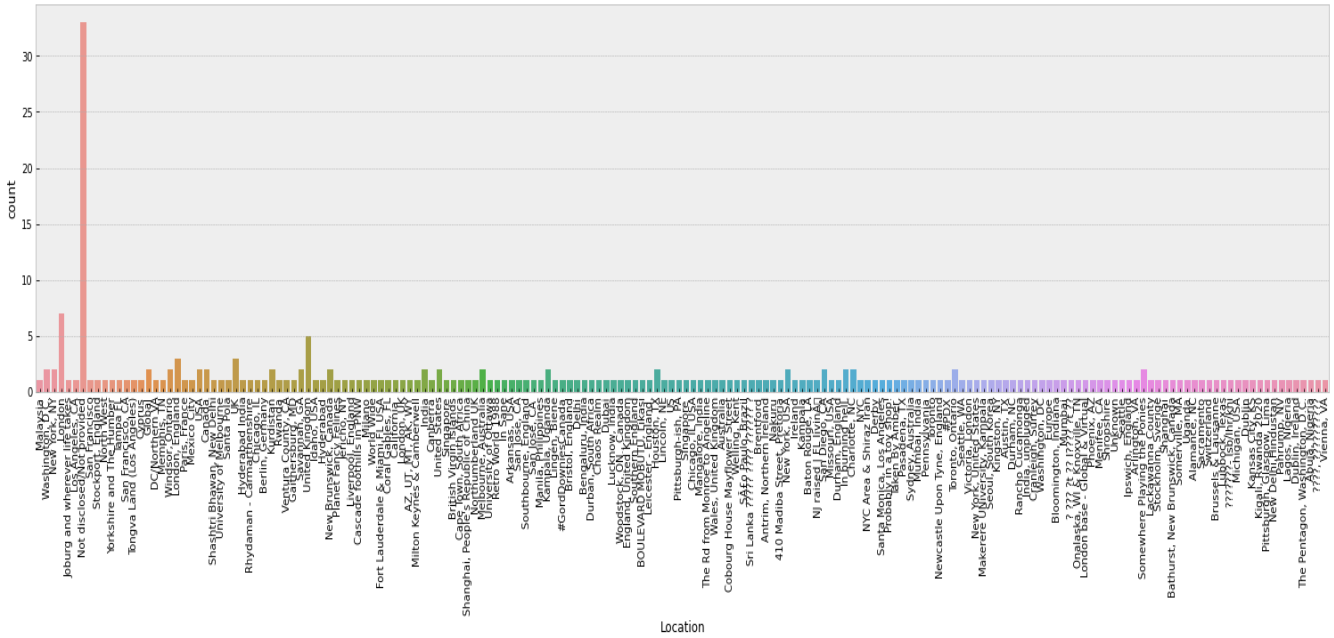
```
['MelanieMoore6 Youre welcome Melanie Keep eye httpstcoz4joiwVLi3 rotating available selections also check httpstcobTXzPMqLP
b updates supporting customers difficult',
 'Gym Closed Instead driving go long walk grocery store Great workout legs butt also work back arms shoulders carrying groce
ry bags fitness SocialDistancing Covid19',
 'grocery store amp pharmacy employees essential workers get paid like Many minimum wage workers Ontario government subsidiz
ing substantial temporary raise workers Ontario coronavirus fordnation',
 'YoniFreedhoff mattgurney need Teachersunions easily teach virtuallypost lessons onlineeven give feedback students know wor
k home paid Time step leave grocery store',
 'business international markets webinar show navigate today exporting challenges rising commodity prices due COVID 19 virus
WEBINAR Managing coronavirus impact global supply chains']
```



Observations:

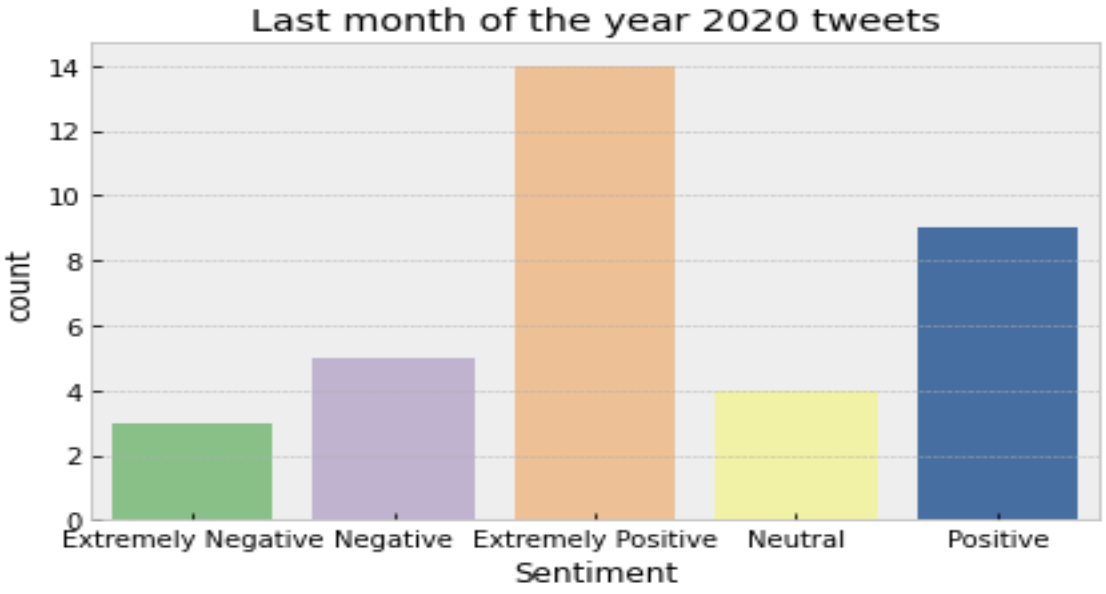
- It can be said that the sample findings have shown accurate results saying there are more tweets from Canada location.
- As it can be seen that there are 51 tweets only which are Positive and there are many other tweets which under other categories of Sentiment
- The length of these positive tweets from Canada location ranges from 100 to 310 where high number of people tweeted with the length of the tweet >210.

2. Find the tweets where the length of tweets is greater than 300 and Sentiment is negative from the month of March



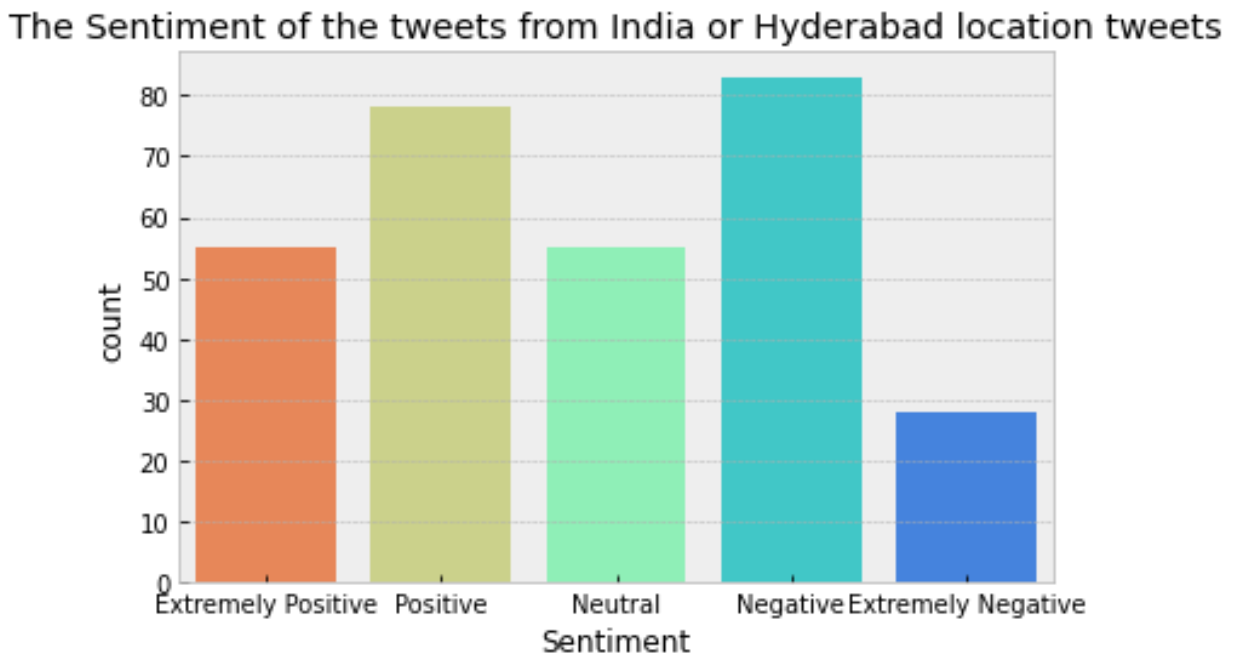
- There are 244 such rows which are having tweet length greater than 300 and resulting to negative sentiment from the month of March
- It can also be observed from the above plot that the negative tweets are more from London location then United Kingdom and their states followed by USA, Canada, Germany …etc

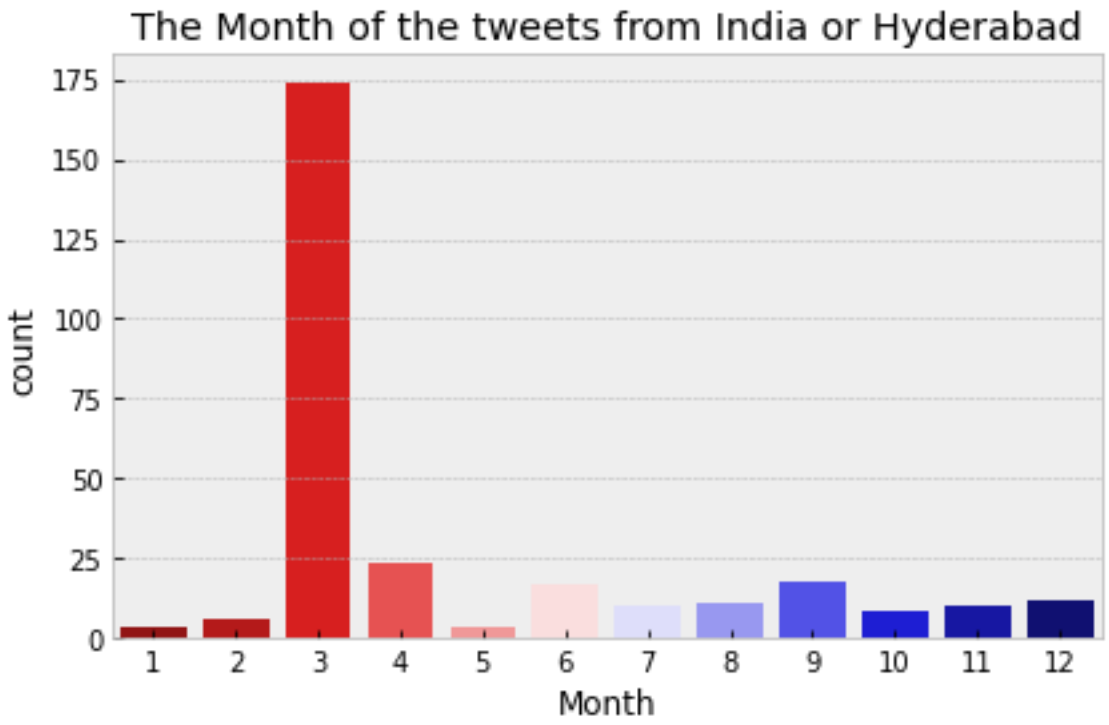
3. The number of tweets from last month and last day of the year 2020 and the length of the tweet be >300



- There are a total of 35 tweets which are from December 2020 and the length of tweet is greater than 300.
- There are more extremely positive tweets in this category indicating the users are good and may be have received some goodies/offers/benefits and are looking forward for the coming expressing the same in the twitter.
- The more about the extremely positive tweets and other tweets can be found in the cleaned tweet column.

4. Find the tweets which are from Hyderabad or India as the Location





Observation:

- After analysing the tweets which come from India or Hyderabad location, it can be seen that there are more negative tweets than the positive sentiment tweets.
- There are a total of 299 rows which are from India or Hyderabad location from which India(country) location tweet are greater than one city Hyderabad tweets. 😂
- There are high number of tweets in the month of March than any other month in a year.

8: Conclusion

- Here the most important feature that gives most of the correct predictions is due to the cleaned_tweet feature which contains the cleaned tweet of the tweets given in the data.
- The Cat Boost model which understands the language and terms used in it has given the correct predictions compared to the model which couldn't capture or understand the data and has underfit to the data.
- XGB Classifier has given good score and also precision is high compared to other models then didn't show good results on minority class.
- Cat Boost was performed very well capturing all the classes and improved performance on all the levels of Sentiment present in the data.

- This data is collected from almost all the countries as it can be seen there are many countries all over the world who are using tweeter. So, this a global reach implying this is the sentiment analysis of covid tweets all over the world.
- There are more tweets in the month March than any other month saying users have no clue of what's happening in their country and I have shared their views on the twitter.
- From the words used in the tweets it can be observed that the discussion is only about the covid and related tweets and nothing away from it. Many people have used same terminology in the tweets and says how much they are affected by it.
- Some of the twitter users have not revealed their location and they constitute 20.87% of missing values in Location variable. So it cannot be imputed with simple methods and domain knowledge would come rescue and I have filled it with Not disclosed instead of dropping it or other methods of imputation so that the other data won't get ruled out from analysis.