Abstract

New York is one of the biggest cities in the world and the stage can not get any bigger for the restaurants and cafeterias to perform at optimum levels. While we try our maximum to try every cuisine from every corner, we can not ignore the necessary safety and hygiene standards. For this purpose, the Department of Health and Mental Hygiene in New York conducts regular inspections in restaurant and school/college cafeterias which is a key in promoting public health and safety.

The goal of our research is to build a classifier which predicts the outcome in terms of whether the facility remains open or closed. We will be using maximum insights and knowledge from the predictor variables in our inspection data. The toolkit employed for this project will be a combination of Natural Language processing and Supervised machine Learning algorithms. The data-set is hosted under the link "https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j".

- Languages I will be using to analyze our data :

1) Python

- Techniques and algorithms:

1) Classification under ML Supervised Learning using algorithms like Logistic Regression, Naive Bayes, Decision Trees Classifier, Random Forest Classifier

2) Sentiment Analysis using algorithms like TF-IDF(Term Frequency and Inverse Document Frequency). Also we will be utilizing stemming and tokenization techniques

3) Exploratory Data Analysis comprising grouping and aggregation.

4) Data Visualization will be implemented using Python libraries like Matplotlib and Seaborn

The Dataset

The data set is provided by the Department of Health and Mental Hygiene (DOHMH) under the link: https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

Attribute	**Attribute Description**	**Attribute Type**
CAMIS	index variabel; unique for each food establishments	Integer
BORO	New York Boro in which in each restaurant(facility) is located	Text; categorical
CUISINE DESCRIPTION	The type of cuisine the restaurant hosts eg American, Chinese, Italian	Text; categorical
INSPECTION DATE	Date of Inspection	Date datatype
VIOLATION DESCRIPTION	This field is the brief description of the type of violation associated with the establishment. Primary focus for our sentiment analysis	Text
CRITICAL FLAG	This field contains Yes or No categories to describe whether the violation is critical or not	Binary
SCORE	Score given to the violation of each restaurant facility.	Integer

GRADE	Grading scheme based on scores. Eg Grade A, Grade B, Grade C, Grade Pending, Not Applicable.	Text; Ordinal
ACTION	Dependent variable; 0 for store open;1 for store closed	Binary

DESCRIPTIVE STATISTICS

The study only has a single numerical column (Score)

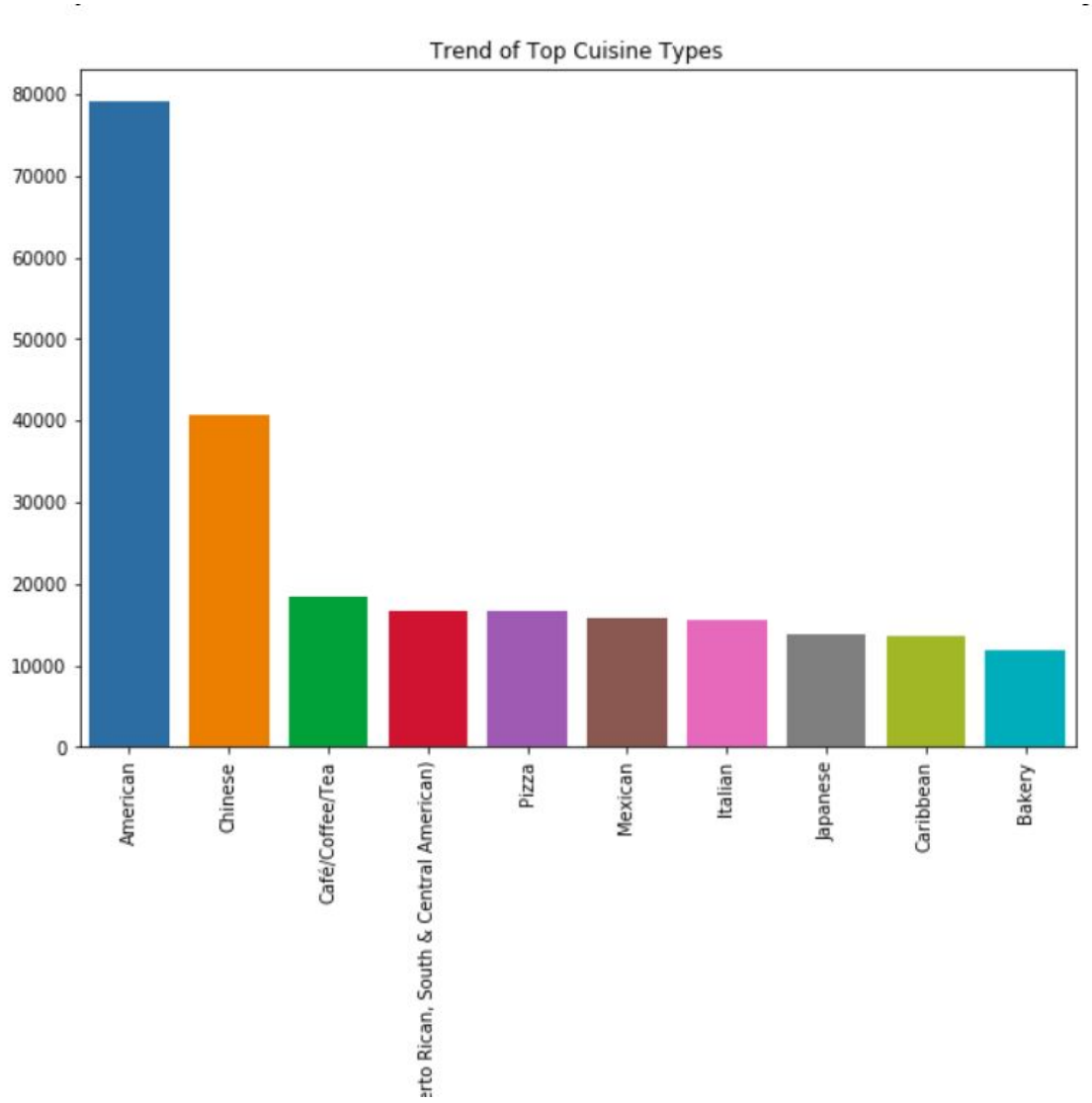
count	375270.000000
mean	20.261998
standard deviation	14.724669
minimum	-1.000000
25% quartile	11.000000
50% quartile	15.000000
75% quartile	26.000000
maximum	164.000000

Data Exploration

The complete code is available on my github account:

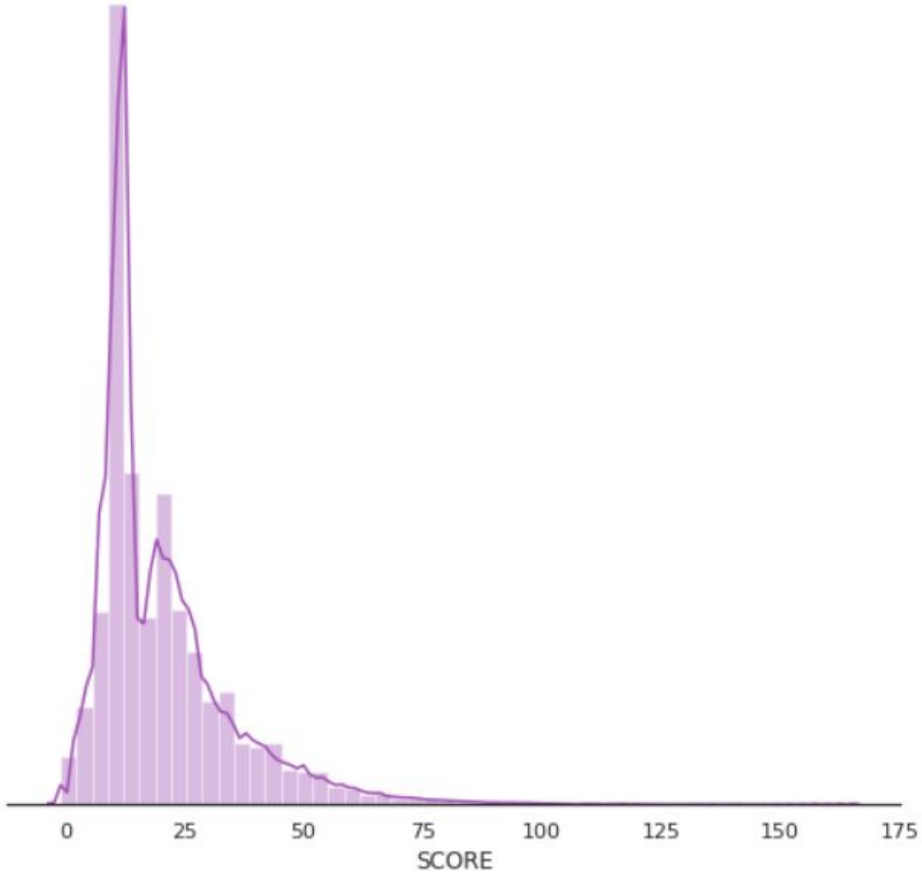
1) Different cuisine types

On exploratory data analysis we can find over 60 variations of cuisines which New York city restaurants share amongst themselves. Out of those, we can see the top 10 .



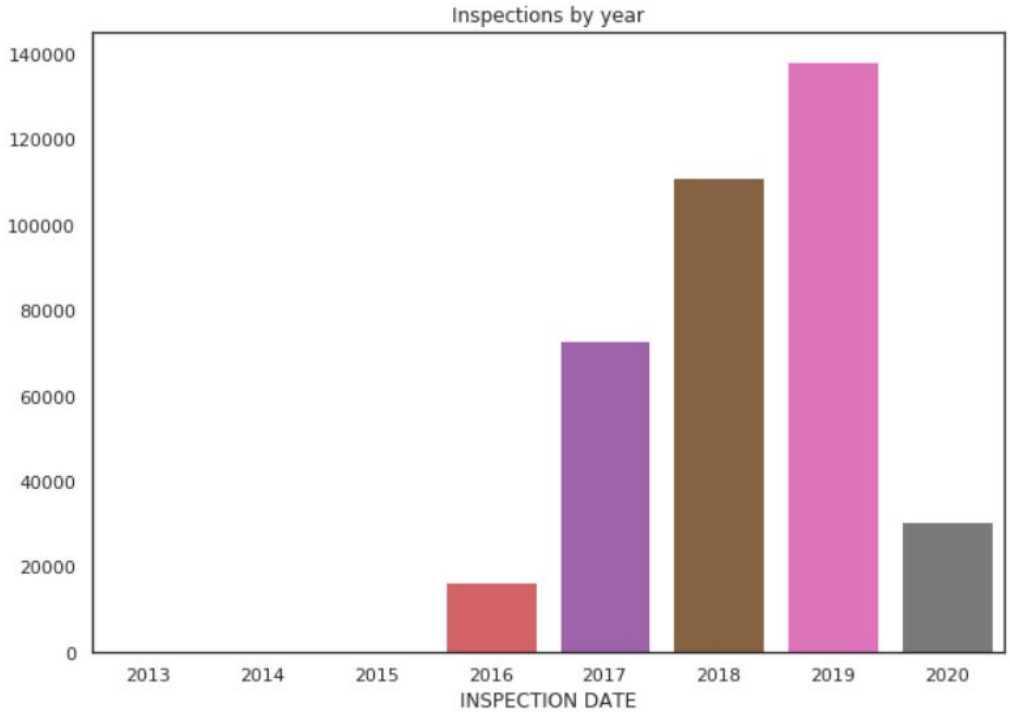
2) Distribution of Scoring

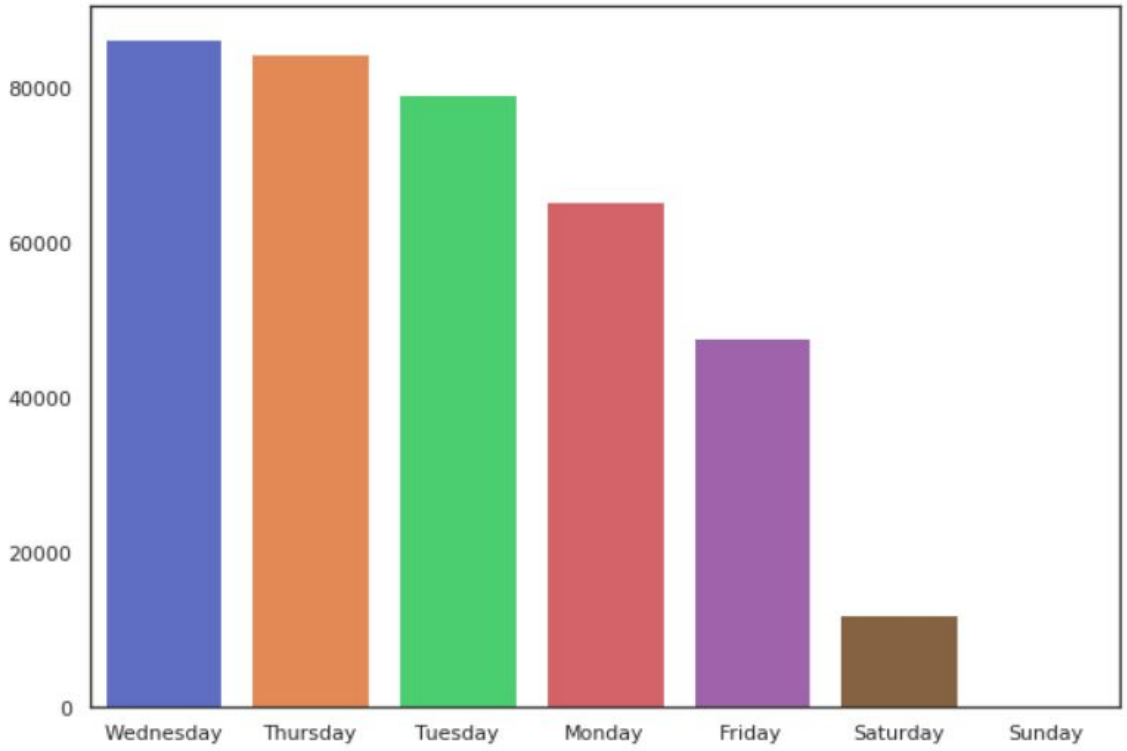
We take a look at how the score column, which is our only numerical column, is distributed. This also helps us to decide whether the distribution is normal or not.



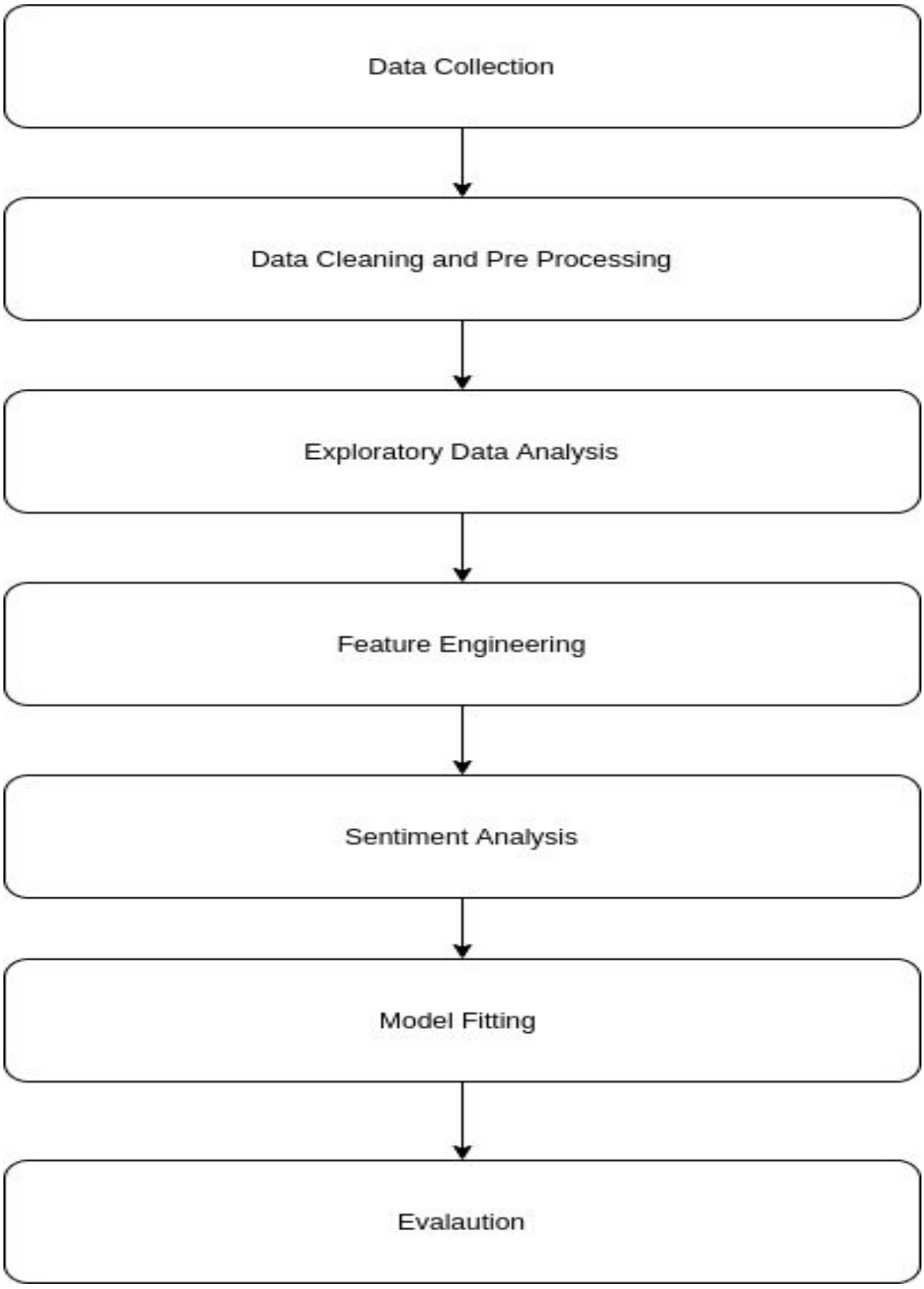
From the above graph, we can see that our scoring column is not normally distributed

3) Next we try to study the trend of inspections by grouping them by the days of a week as well as by the year. By this graph we can see that inspections mostly occur during mid week. Wednesdays and Thursdays accounting for the most common days for an inspection.


Inspections by year
140000
120000
100000
80000
60000
40000
20000
0
2013
2014
2015
2016
2017
2018
2019
2020
INSPECTION DATE


80000
60000
40000
20000
0
Wednesday
Thursday
Tuesday
Monday
Friday
Saturday
Sunday

APPROACH:


Data Collection
Data Cleaning and Pre Processing
Exploratory Data Analysis
Feature Engineering
Sentiment Analysis
Model Fitting
Evalaution

2) Data Collection:

Data collection is defined as the procedure of collecting,measuring and analyzing accurate insights for research using standard validated techniques (www.questionpro.com). For our capstone project, we have used restaurant inspection data from open data portal of new york. The dataset is hosted under the link:
"https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j".

3) Data Cleaning and PreProcessing:

Data Cleaning and preprocessing is the most crucial step of the analysis. Data cleaning accounts for almost 80% of the time. It comprises many important steps like taking care of missing values, duplicate values by case, Data Cleaning and data transformation.
Steps involved in Data Cleaning are as follows:

- Removing outliers in Grade column
- Removing missing values in the Score and Grade column
- Removing missing values from VIOLATION CODE','VIOLATION DESCRIPTION','CRITICAL FLAG' columns
- Cleaning our text data [VIOLATION DESCRIPTION] for the purpose of using tf-idf and kmeans

Data Preprocessing involves making the raw data in a suitable format so that we can use it for the modelling purposes. Major steps involved in Data Preprocessing are as follows:

- Mapping of our label column (ACTION) into 0: for establishment open and 1: for establishment closed
- Since we have used a combination of NLP(Natural Language Processing) and classification, we have used a library texthero which has prebuilt clean function which removes tags, removes non characters and symbols, converts text to lowercase, removes stop words to make it suitable for using tf-idf and k means.

4) **Exploratory Data Analysis**:

Next step is the EDA part. For our capstone project, we have tried to investigate and visualize our data based on different criterias. Some of them are as follows:

- Grouping the inspection results based on the top cuisines in new york.
- Grouping by inspection date: Day of week
- Trends grouped by inspection year

5) **Feature Engineering** :

Since our end goal is to use classifier algorithms for modeling our data, we need to convert our dataset into numerical form to be able to use it. For this purpose I have used numerous techniques which are as follows:

- Label Encoding using pandas.get.dummies() on [Cuisine Description, BORO description, Inspection Date, Grade values, Critical Grade, Violation Code, Inspection Type] to make the data in a machine readable format.
- After Label encoding deleting all the label columns
- For our text based data which is the Violation Description, I have used **texthero** which is an open source NLP library to carry out sentiment analysis and give it a numerical score using tf-idf to reflect the importance of that description.
- After using tf-idf I have applied pca on the results for dimensionality reduction and to reflect the important words in the Violation description and use their numerical significance for classification model purposes.
- For better accuracy purposes, I have also used normalization and standardization techniques on the SCORE column. Normalization has been implemented using sklearn library with MinMaxScaler and standardization using StandScaler.

6) **Model Fitting**

After we have converted our raw data into a numerical form, we are going to use this representation in a supervised learning algorithm. For this project we are going to use the following algorithms:

1. **Logistic Regression**: One of the most basic machine learning classifying algorithms is our go to method since our problem case in binary in nature. It is a predictive analysis algorithm based on the concept of probability using a complex sigmoid function. The probabilities are then transferred into binary values (0 or 1) to make an actual prediction. (machinelearningmastery)

2. **Naïve Bayes**: Naive Bayes is a classification algorithm for binary as well as multiclass classification based on calculating conditional probabilities using Bayes Theorem. The assumption is that the features are independent of each other which is why it is called naive which is a very strong assumption which doesn't happen in real life data. (machine learning from scratch, 2019)

3.**Decision Tree Classifier**: Decision Tree classifier is a supervised machine learning algorithm that makes predictions based on a tree structure. The algorithm checks conditions at node, then splits data based on the conditional statement. The algorithm classifies the dataset on the basis of minimum entropy (i.e. the level of impurity in the dataset) and maximum information gain(i.e. feature which gives maximum information about the classes will be used to split the data first.)

7) Evaluation: Based on the results, we aim to classify our dataset in terms of whether the inspection record leads to the closing of the store and lets it remain open.

Model and Evaluation:

I have uploaded my complete code on:
https://github.com/swede77/DOMHM_analysis.git

The goal of the project is to classify each inspection record into two binary classes whether they contribute to keeping the store open or it contributes to the closing of the facility. Two main techniques are used :

1) NLP using tf-idf and k means on text column (VIOLATION DESCRIPTION)
2) Using the numerical representation of the text combined with the encoded dataset in machine learning classifying algorithms.

After text preprocessing and having clean reviews I input my data in three machine learning algorithms. For the following model I will be comparing the following:

Accuracy on validation score that is the score to evaluate the models performance. **TN / True Negative:** when a case was negative and predicted negative

TP / True Positive: when a case was positive and predicted positive

FN / False Negative: when a case was positive but predicted negative

FP / False Positive: when a case was negative but predicted positive

Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class it is defined as the ratio of true positives to the sum of true and false positives.

Precision – Accuracy of positive predictions.

Precision = TP/(TP + FP)

Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives

Recall = TP/(TP+FN)

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

1). **Logistic Regression:**

One of the most basic machine learning classifying algorithms is our go to method since our problem case in binary in nature. It is a predictive analysis algorithm based on the concept of probability using a complex sigmoid function. The probabilities are then transferred into binary values (0 or 1) to make an actual prediction. Results of the model on both normalized as well as denormalized data:
Accuracy of model: 0.964
Accuracy of model on normalized data: 0.955

Confusion Matrix for denormalized data:

```
[[71205   544]
 [ 2103  1202]]
```

Confusion Matrix with normalized data:

```
[[71153   596]
 [ 2095  1210]]
```

Classification report with denormalized data:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	71749
1	0.69	0.36	0.48	3305
accuracy			0.96	75054
macro avg	0.83	0.68	0.73	75054
weighted avg	0.96	0.96	0.96	75054

Classification report with normalized data:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	71749
1	0.70	0.39	0.50	3305
accuracy			0.97	75054
macro avg	0.84	0.69	0.74	75054
weighted avg	0.96	0.97	0.96	75054

2) **Naive Bayes**

For having additional information while using Naive Bayes model we have used standardization techniques using MinMaxScaler from sklearn library.
Accuracy of model: 0.225
Accuracy on standardized data model: 0.171
Classification report :

	precision	recall	f1-score	support
0	0.98	0.19	0.32	71749
1	0.05	0.93	0.10	3305
accuracy			0.23	75054
macro avg	0.52	0.56	0.21	75054
weighted avg	0.94	0.23	0.31	75054

3) Decision Tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.(scikit-learn.org/stable/modules/tree.)
We have performed modeling using decision trees on both denormalized data and normalized data as well:
Accuracy of model: 0.958
Accuracy of model on normalized data: 0.958

Confusion Matrix for denormalized data:

```
array([[70452,  1297],
       [ 1851,  1454]])
```

Confusion Matrix for normalized data:

```
[[70452  1297]
 [ 1851  1454]]
```

From the results on Decision Tree Classifier we can conclude that normalization does not have much effect on the evaluation of the results.

CONCLUSION:

This project has applied three different machine learning algorithms Logistic Regression, Naive Bayes and Decision Tree Classifier on the New York Restaurant Inspection dataset. The results from this exercise showed that in term of accuracy, classification with Logistic regression and Decision Trees classifier approach achieves much better results than the Naive Bayes algorithm. We can further improve the final model by using more training data and feature engineering techniques to give better performance and accuracy.

References:

1) David W. Nadler. Using Logistic Regression to Model New York City Restaurant Grades Over a Two-Year Period
2) Makoto Nakayama, Yun Wan. The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews
3) MachineLearningMastery.com
4) Text Mining with R - A Tidy Approach - Julia Silge and David Robinson – 2017-05-07
5) Algorithms Every Web Developer Can Use and Understand.
6) New York City Department of Health and Mental Hygiene. About_NYC_Restaurant_Inspection_Data_on_NYC_OpenData_092418.docx
7) New York City Department of Health and Mental Hygiene. https://data.cityofnewyork.us/api/views/43nn-pn8j/files/ec33d2c8-81f5-499a-a238-0213a38239cd download=true&filename=RestaurantInspectionDataDictionary_09242018.xlsx
8) https://www.kaggle.com/ambarish/extensive-modelling-chicagofoodinspections
9) https://www.apu.edu/live_data/files/288/literature_review.pdf