



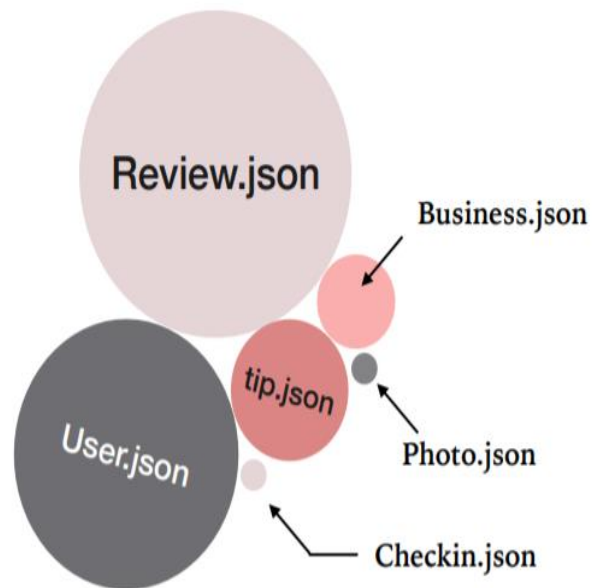
# Restaurant Closure Prediction

PRESENTED BY-  
AHNCH BALA  
SONAKSHI KARKERA  
SURBHI THAKUR  
ARUN KALAESWARAN

# PROBLEM STATEMENT



- There is an enormous amount of data available for Yelp to use to determine if a restaurant will close before it happens.
- Yelp can use this information to open up new channels of revenue, specifically in the form of helping failing restaurants come back from the brinks of failure.
- By determining whether they will close or not would give clients an opportunity to improve their user experience and try and prevent their business from closing.



## Data Source

---

- Yelp Dataset- 2019 from Yelp Dataset Challenge
- Filter out data using category as restaurant and city as Toronto from business.json
- We combine the reviews for restaurants in Toronto from reviews.json
- We have 7967 rows  $\times$  15 columns on loading the dataset.

# DATA

- **Business data**

- business\_id: ID of the business
- name: name of the business
- neighborhood
- address: address of the business
- city: city of the business
- state: state of the business
- postal\_code: postal code of the business
- latitude: latitude of the business
- longitude: longitude of the business
- stars: average rating of the business
- review\_count: number of reviews received
- is\_open: 1 if the business is open, 0 otherwise
- categories: multiple categories of the business

- **Attribues of review table**

- business\_id: ID of the business
- text: review from the user

# CONSTRAINTS OF DATA



**Validity of Reviews:** Although Yelp filters fake or bogus reviews, we are still unsure if 100% of the reviews are genuine.



**Live Dataset:** The dataset is not live, which would mean that the current situation might be different for restaurant under review.



**Missing Financial Information:** We don't know the financial side of the restaurants under review which can help us be more accurate in our predictions.



**Limited Access to Data:** We only have access to Yelp's dataset and don't have any other dataset we can use to strengthen our prediction.



# OUR APPROACH

Sentimental Analysis on Reviews

Data Pre-processing

Closure Prediction Model

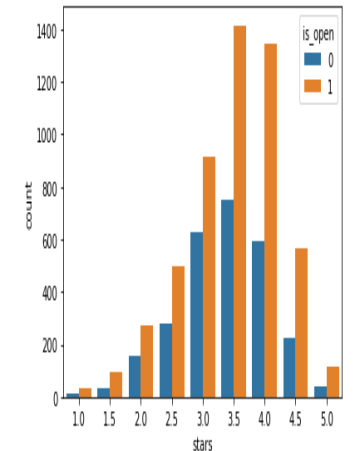
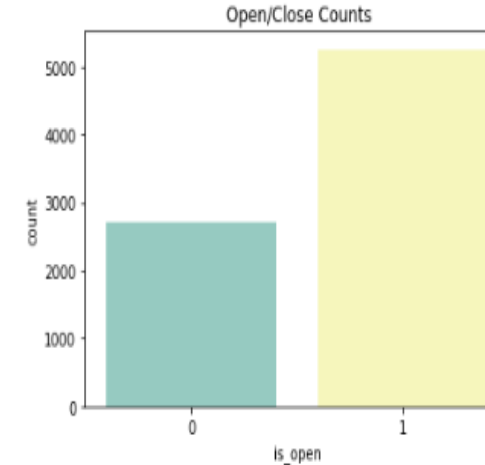
Exploratory Data Analysis

Conclusion



# Exploratory Data Analysis

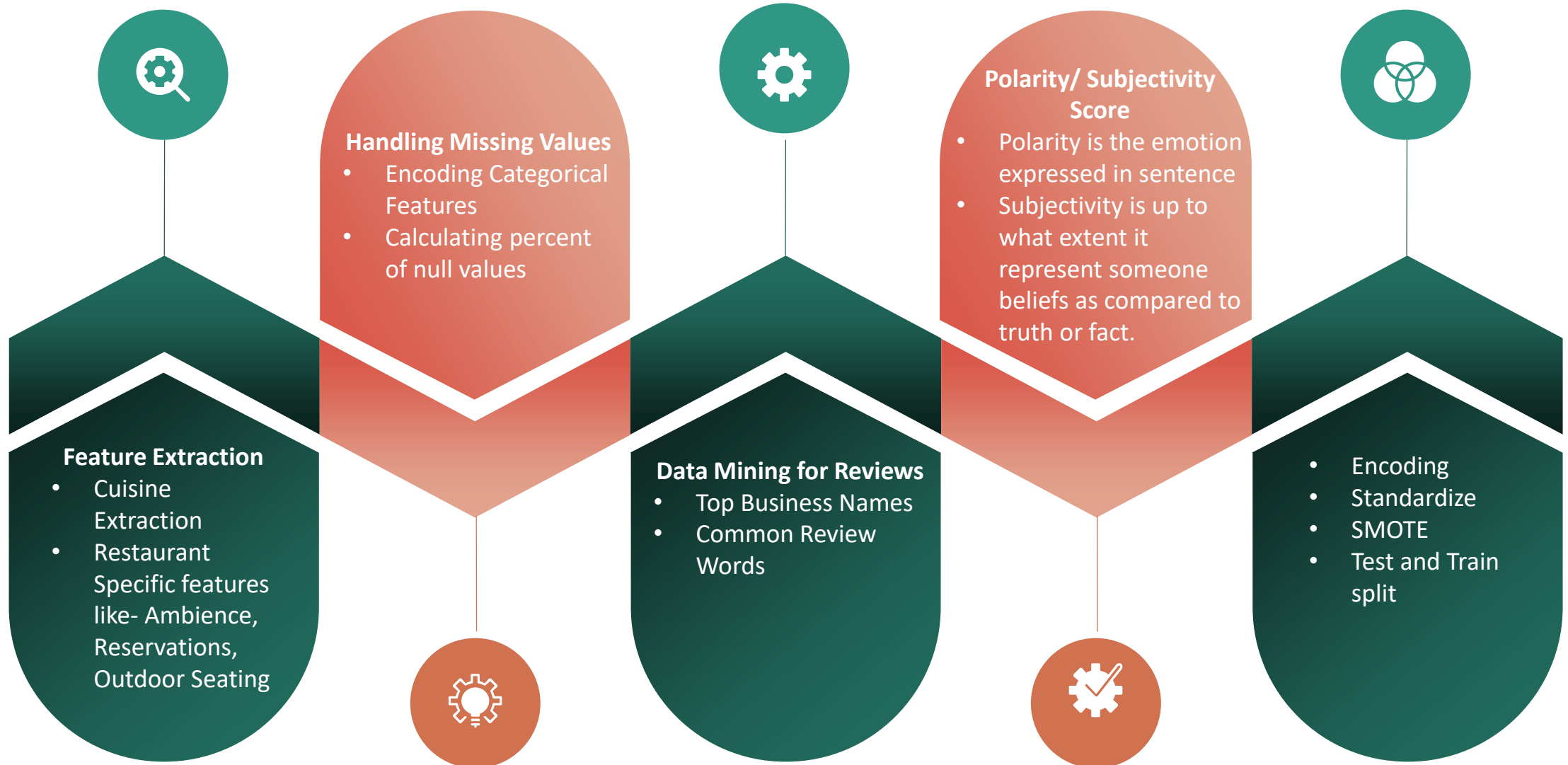
- Analysing the number of open and closed restaurants.
  - ✓ Proportion of the variable 'is\_open' in our dataset is imbalanced.
  - ✓ The businesses that are open are close to 84% and the businesses are closed are 16%.
- Distribution plot of number of open and closed restaurants based on the star rating of the restaurant.
- Missing values
- Attributes and category has sub attributes like whether a business has bike parking or free WiFi, different type of cuisine and how those attributes correlate with a restaurant closure.



```
hours          1900
attributes      350
text            0
categories      0
is_open         0
review_count    0
stars           0
longitude       0
latitude        0
postal_code     0
state           0
city            0
address         0
name            0
business_id     0
dtype: int64
```

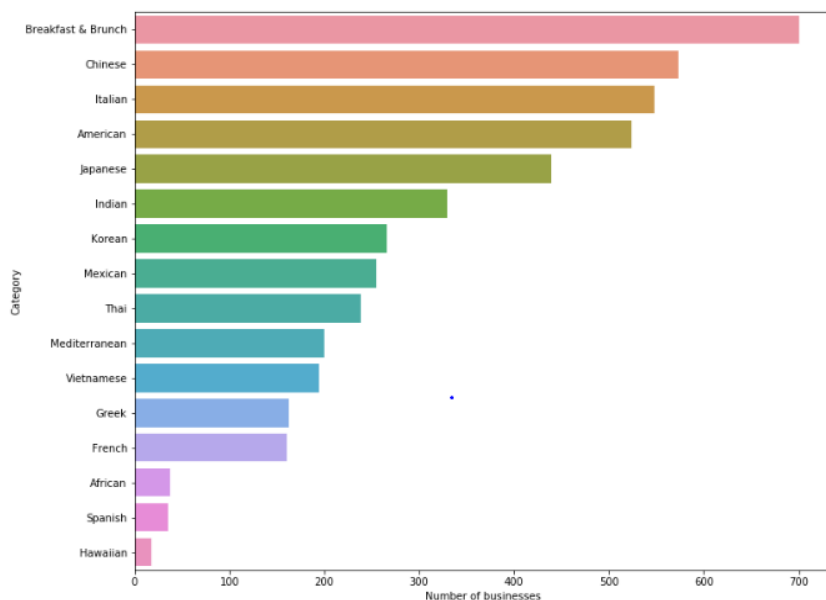
attributes	categories
{ "WiFi": "no", "BikeParking": "True", "Rest..." }	Juice Bars & Smoothies, Food, Restaurants, Fas...
{ "BusinessParking": "garage": False, "street..." }	Restaurants, Nightlife, Breakfast & Brunch, Ve...

# Data Pre-Processing





BusinessAcceptsBitcoin	4681
AgesAllowed	4681
RestaurantsCounterService	4681
DietaryRestrictions	4677
ByAppointmentOnly	4659
DriveThru	4587
BestNights	4526
BusinessAcceptsCreditCards	4516
Smoking	4499
CoatCheck	4496
Music	4447
GoodForDancing	4437
HappyHour	4436
DogsAllowed	4270
WheelchairAccessible	4019
RestaurantsTableService	3086
GoodForMeal	2255
Caters	1963
BikeParking	1657
WiFi	1377
NoiseLevel	1202
Alcohol	1097
hours	1069
HasTV	1025
Ambience	1006
BusinessParking	929
OutdoorSeating	809
RestaurantsAttire	777
RestaurantsDelivery	769
RestaurantsReservations	670
GoodForKids	655
RestaurantsGoodForGroups	576
RestaurantsPriceRange2	561
RestaurantsTakeOut	504



# Feature Extraction

- In Feature Selection we extracted different cuisines in the Toronto Region.
- We extracted restaurant specific features like Happy Hour, Ambience, Reservations etc.
- Drop values having more than 50% data missing and irrelevant features.

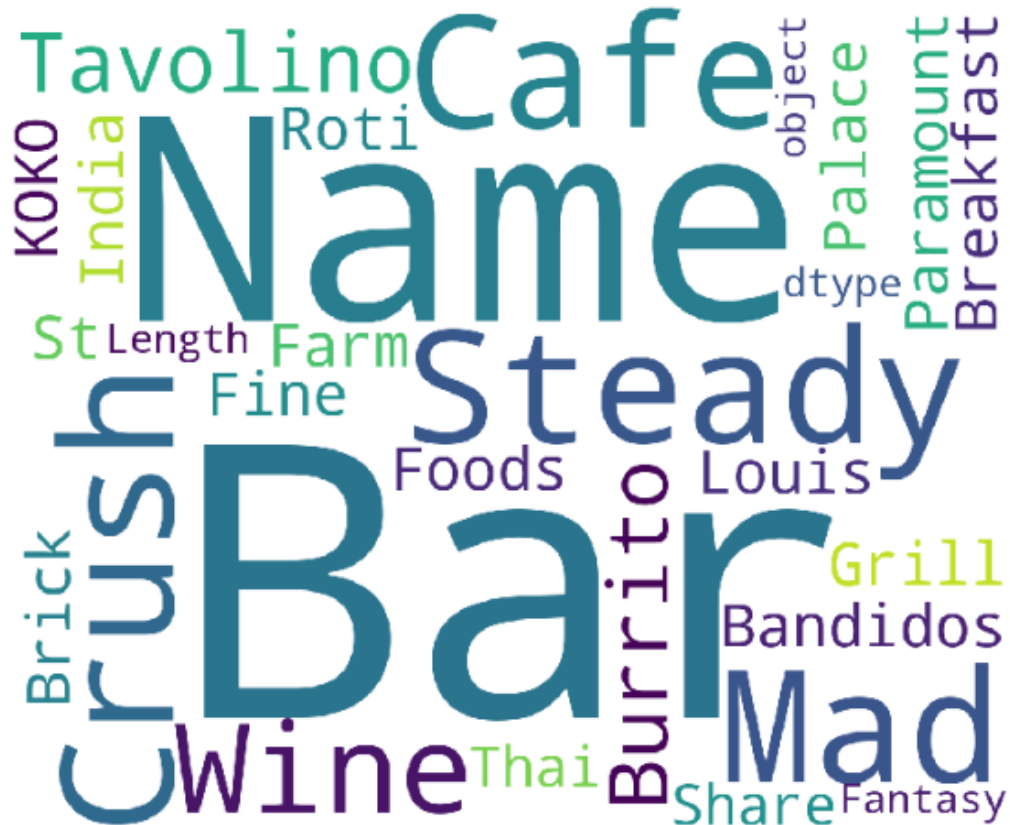


# Sentimental Analysis on Reviews

- We use the words in texts to calculate a measure of the sentiment of the people from their reviews.
- Remove unwanted characters, numbers and symbols,
- Stopwords and converting them into lowercase.
- We then get polarity and subjectivity score using Textblob function.

# Word Bags

---



A word cloud featuring restaurant names and related terms. The words are arranged in a circular pattern, with 'Name' and 'Bar' being the largest and most central. Other prominent words include 'Cafe', 'Steady', 'Crush', 'Wine', 'Burrito', 'Mad', 'Grill', 'Bandidos', 'Share', 'Fantasy', 'Louis', 'Foods', 'Fine', 'Farm', 'St', 'Length', 'Roti', 'Palace', 'Paramount', 'Breakfast', 'object', 'dtype', 'KOKO', 'India', 'Tavolino', 'Brick', and 'Crush'.

Word cloud containing restaurant names and related terms:

- Primary words: Name, Bar, Cafe, Steady, Crush, Wine, Burrito, Mad, Grill, Bandidos, Share, Fantasy, Louis, Foods, Fine, Farm, St, Length, Roti, Palace, Paramount, Breakfast, object, dtype, KOKO, India, Tavolino, Brick, Crush.



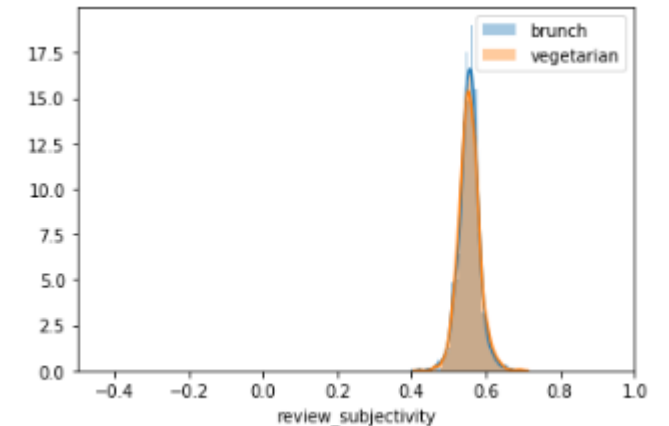
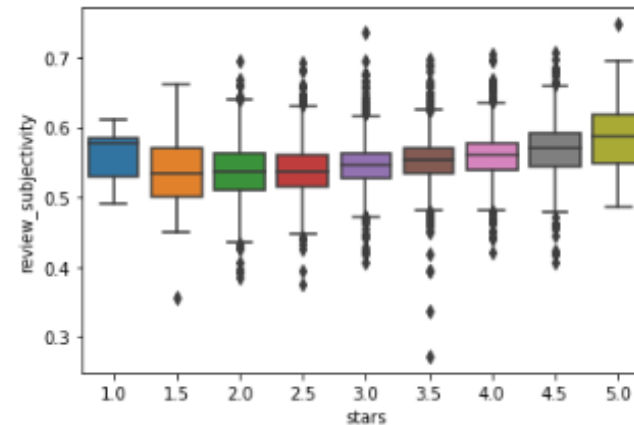
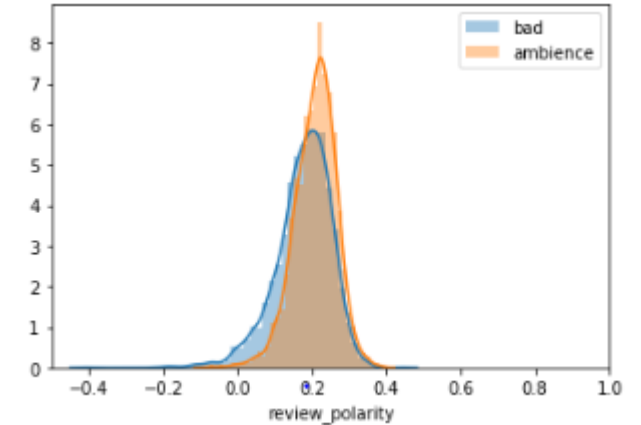
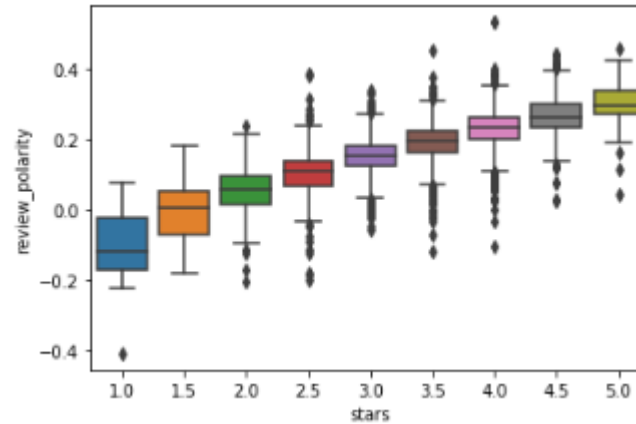
A word cloud featuring food and dining-related terms. The words are arranged in a circular pattern, with 'food', 'place', 'Great', 'Toronto', 'loving', 'Beach', 'dining', 'eaten', 'offeri', 'Uh', 'raining', 'Length', 'Woodbine', 'addition', 'vegetarian', 'OFTEN', 'sad', 'nBy', 'dtype', 'Big', 'middle', 'far', 'traditional', 'eastern', 'serve', 'little', 'text', 'came', 'bench', 'frequent', 'know', 'day', 'joint', 'object', 'Thai', 'Staff', 'nice', 'new', 'Toro', 'wa', 'eaten', 'offeri', 'Uh', 'raining', 'Length', 'Woodbine', 'addition', 'vegetarian', 'OFTEN', 'sad', 'nBy', 'dtype', 'Big', 'middle', 'far', 'traditional', 'eastern', 'serve', 'little', 'text', 'came', 'bench', 'frequent', 'know', 'day', 'joint', 'object', 'Thai', 'Staff', 'nice', 'new', 'Toro', 'wa'.

Word cloud containing food and dining-related terms:

- Primary words: food, place, Great, Toronto, loving, Beach, dining, eaten, offeri, Uh, raining, Length, Woodbine, addition, vegetarian, OFTEN, sad, nBy, dtype, Big, middle, far, traditional, eastern, serve, little, text, came, bench, frequent, know, day, joint, object, Thai, Staff, nice, new, Toro, wa.

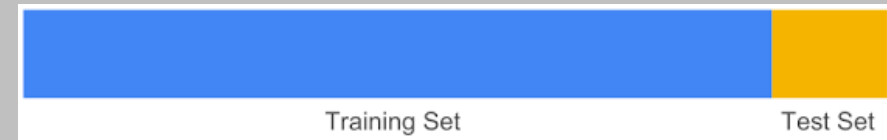
# Polarity and Subjectivity

- Polarity: Polarity score is a float within the range  $[-1.0, 1.0]$ , also known as orientation polarity is the emotion expressed in the sentence. It can be positive, negative or neutral.
- Subjectivity: It measures subjectivity of sentence or to what extent it represents someone's personal feelings, views, or beliefs compared to objective truth or facts. subjectivity is a float within the range  $[0.0, 1.0]$  where 0.0 is very objective and 1.0 is very subjective.



# Data Preparation

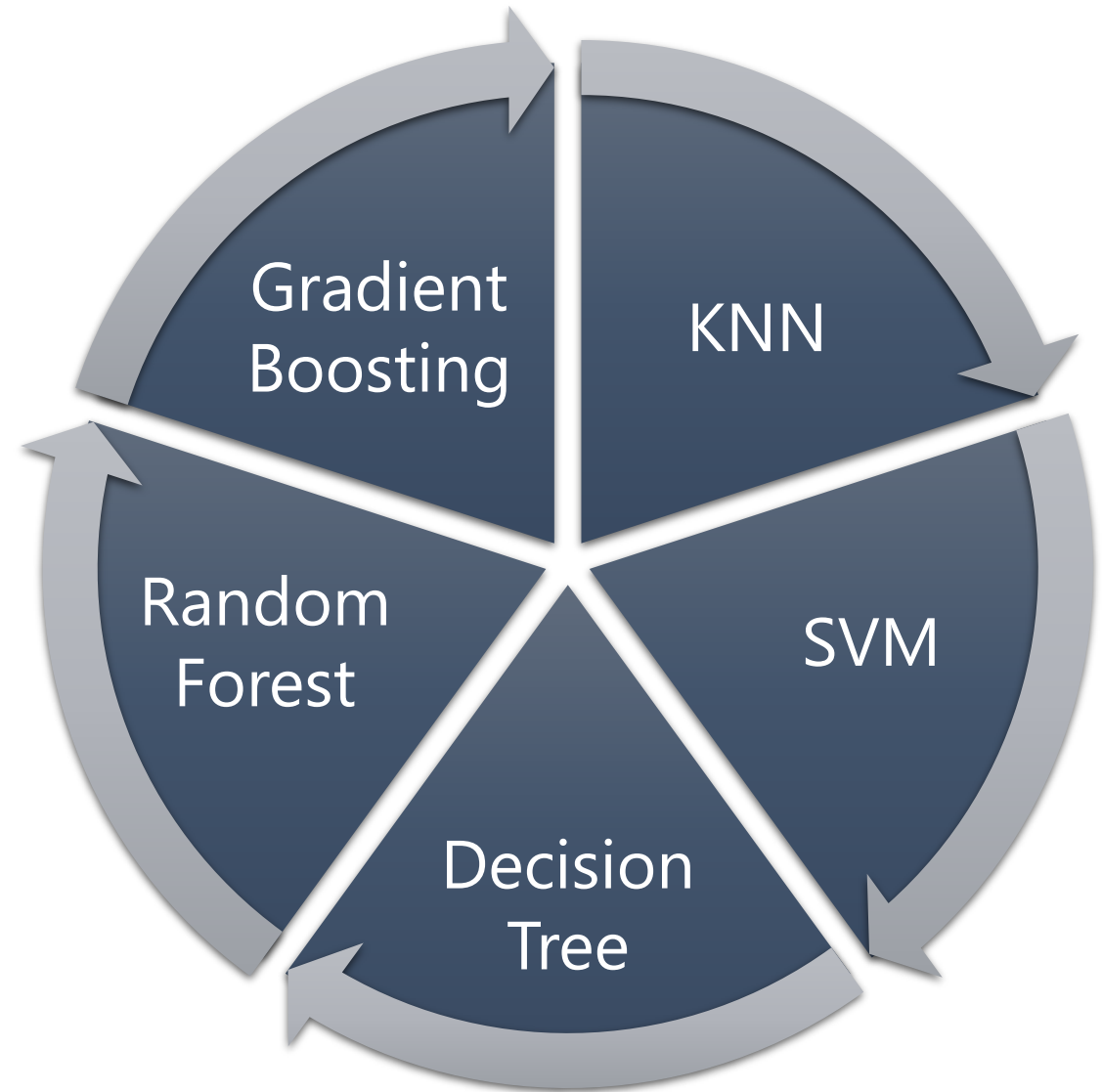
- Replaced the null values for rest with maximum value counts of the feature.
- Mapped values True and False to 0 and 1 and for rest we used get\_dummies for encoding categorical variables.
- Scaled the data using StandardScaler()
- SMOTE (Synthetic Minority Over-sampling Technique)
  - we oversampled the minority class which in our
  - case is closed restaurants.
- Split the dataset into Test and Train using 80/20 split.



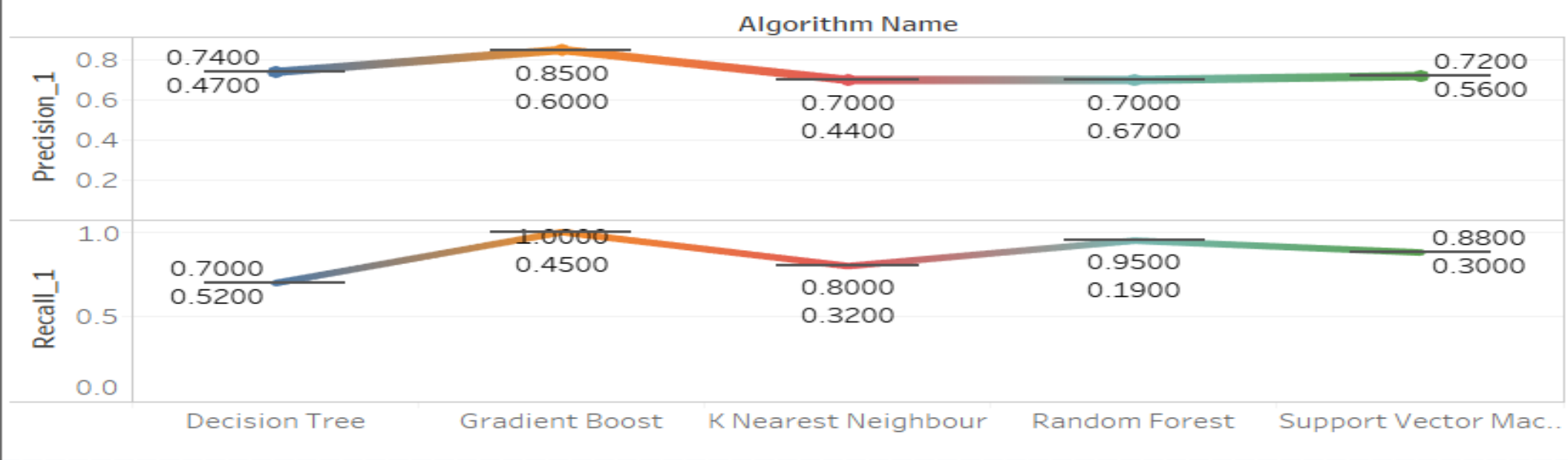
```
In [48]: from sklearn.model_selection import train_test_split
         from imblearn.over_sampling import SMOTE
         X=dataset_final.drop(['is_open'], axis = 1)
         y=dataset_final['is_open']
```



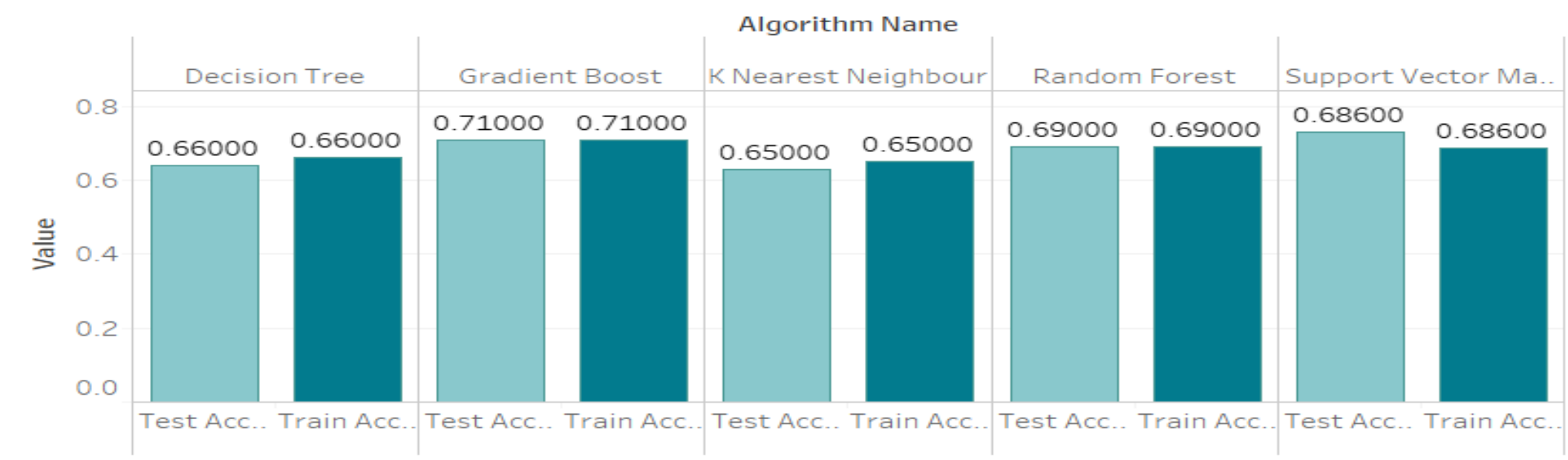
# Prediction Models



# Precision and Recall Values



# Training and Test Accuracy



# CONCLUSION

- With help of Natural Language Processing (NLP), we dealt with the text data, that helped in data mining or text mining (extracting important words from the review).
- By visualisation of dataset, we analysed the imbalance in the data, which resolved by sampling.
- Using Text blob package, sentiment orientation of reviews gives a sentiment polarity and sentiment subjectivity which helped us in labelling and training the model.
- Hence, here in this project we were able to correctly analyse and visualise the data, extracted features and dealt with reviews and were able to train several classifiers out of which Gradient Boost gave us a maximum accuracy score of 71.00 %, which correctly predicts the if the restaurant will close or not.



**THANK YOU**