CMPE 255- 02: DATA MINING

Prof. Gheorghi Guzun

PROJECT REPORT

MICHELIN OR NOT ?

PREDICTING RESTAURANT SUCCESS BASED ON YELP DATASET

Report By:
Team Watson

Team Members:
Shivani Mangal (012530362)
Hemang Behl (013734214)
Prajwal Venkatesh (012557792)

1.0 Introduction

1.0.1 Motivation

The restaurant industry is the second largest employing industry in United States and it's demand is an ever-increasing one (Mandabach, 2011). However, due to it's ever-growing nature, customers have more options. Thus if restaurant owners do not keep up with growing trends, they may go out of business quickly. Today's restaurant experience is not limited to food. A millennial customer judges every aspect of the restaurant, from quality of food to location of restaurant to friendliness of the staff to the noise. A little lag on one of those might lead to an overall bad impression which may lead to negative reviews on Yelp and similar websites.

Yelp, one of the most used websites today, is a pool of data. It is a local search service and a crowd sourced forum for various businesses, from restaurants to boutiques. People from around the world, rate these businesses and share their reviews for the same. This data can be used to identify interesting patterns of user activities which can have a huge monetary impact on success of these local businesses.

With this project, we aim to help the business owners be successful in their endeavors by analyzing public dataset shared by Yelp.

1.0.2 Objective

> *"Here is a simple but powerful rule: always give people more than what they expect to get."*
>
> – Nelson Boswell

The main objective of our extensive analysis is to aid in the success of local restaurants of America. For a restaurant to be successful, it needs to pass certain benchmark. We have chosen the following criteria for a restaurant to be a success:

- should be in a location which is popular among people
- should be serving a cuisine which is popular
- should be serving a cuisine which is popular in a particular location
- should end up having a star rating of 3.5 or above
- should have low general noise level for it to be appealing to people
- should have good reviews from guests
- should have a high footfall in general as compared to other restaurants
- should not close

In order to understand and predict the success of a restaurant, we have split our analysis into multiple sections or use cases. We have applied appropriate machine learning techniques to implement classification models on this dataset which may help us in finding answers to the above mentioned.

2.0 Implementation Details

2.0.1 Algorithms considered

2.0.1.1 Use Case 1 : Location and Cuisine Analysis

Using feature selection for cuisine data and location data, we tried to predict the stars a restaurant might have using **Lasso Regression**, **Decision tree, Random Forest Classifier, Support Vector Classifier** and **Gradient boosted tree**. Along with that **RandomizedSearchCV** and **GridSearchCV** was used to train data for location and cuisine analysis. We have also used **DBScan** to cluster business within a 2km radius and TfIdf vectorizer to find similarity in various documents. Dbscan was used here as location is density based information.

2.0.1.2 Use Case 2: Noise, Footfall pattern and Business Closure analysis

Used Linear SVC, XGB, Random forests, Logistic Regression and KNN classifiers as this was a binary classification problem and these models are known to work better than the rest. SMOTE (Synthetic Minority Over-sampling Technique) was also applied to balance and make the dataset more suitable for the models to ingest it.

2.0.1.3 Use Case 3: Help people who want to start restaurant business by analysing and classifying yelp restaurant reviews as positive or negative.

Have made use of Logistic Regression, Naive Bayes, Linear SVC and Random Forest Classifier algorithms to classify the reviews as positive or negative.

2.0.2 Technologies & Tools

Python packages: Numpy, pandas, seaborn, imblearn, sklearn , NLTK, Scikit-learn (metrics and various models), xgboost

IDE and notebook used: Spyder, Jupyter notebook

2.0.3 Implementation Details

Handling the large dataset: The size of the review dataset was very large and hence we decided to choose a chunk of the data. This made it easy for us to work as our system were able to handle the load.
Converting json to csv: As we decided to go with pandas Dataframe we converted the json data into csv to reduce the load on RAM and increase the efficiency. As we wanted to classify only the restaurant reviews. First, we got all the businesses which had ‘Restaurant’ in its category list from business file and created our own csv file with business_id and categories. Later, we used this file to generate another csv file with reviews and stars to filter restaurant businesses reviews.

Use Case 1: Location and Cuisine Analysis

Data Exploration

1. In order to understand the data better, we also analyzed cuisine to stars by visualizing it.
2. We tried to find if certain neighbourhoods achieve higher star rating.
3. We also find the location to stars to cuisine relation

Feature Engineering : The dataset was first analyzed and cleaned on basis of various features. A few of them were:

1. States : One of the major issues faced was that the data had rows for restaurant outside of USA as well. I reduced it down to 50 states and dropped the rest.
2. Wrong city/state combination : It was only a few rows so instead of fixing the errors, I dropped the rows
3. ReviewCount to Stars Ratio: The feature was heavily skewed. In order to make it gaussian we took a log of it
4. Bucketized review counts: The review counts were categorized into 10 quartile buckets
5. Postal code feature : Found that business with Nan address actually had address missing on Yelp
6. Latitude/Longitude feature
7. Food trucks had null addresses, marked them as NOT_AVAILABLE
8. Cuisine category had comma separated values, we used one-hot-encoding and spread them out.
9. Cleaned up the names feature and location feature by making them lowercase

Data Analysis: Our next step was to use the cleaned up data and cluster nearby restaurants based on Lat/Lon. (Fig. A)Data was transformed to new csv files to prepare them for modelling.

Data Modelling: We split the total data into 3 splits. Training data which is 60%, cross-validation set which is 36% of data and final 4% is used as a test set. The algorithms used have been mentioned above.

Use Case 2: Noise, Footfall pattern and Business Closure analysis, Gender attribute creation

The cleaned data was grouped according to the attribute business ID which is unique to each business. We then merged two files- business and reviews on the basis of the unique business ID attribute. Mean of different stars given for different noise levels was visualized.

Footfall pattern: New attributes for the total number of checkins per day were created in check-in dataset. We also took the sum of check-ins for each separate hour. We visualized newly created attributes using plots. Further analysis was done on newly created features. To create the gender attribute, we extracted and mapped the SSA dataset with the Yelp user names into a dictionary.

Business closing probability:Various algorithms mentioned in the above section were implemented on a merged file containing business, review and modified check-in.

Use Case 3: Analysing and classifying yelp restaurant reviews as positive or negative

To convert document corpus to a vector format we have used bag-of-words approach, where each unique word in a text document is represented by one number. The classification algorithm will need to be in vector format to perform the classification task. We used NLTK library available on python to remove stopwords and punctuations converting reviews into list of tokens.

To get Scikit-learn algorithms to work we need to convert each review into a vector. We used Scikit-learn's CountVectorizer to convert the text documents into a matrix of token counts. As a result of this process, we get a 2-D matrix, with each row as a unique word and column as a review. To further clean the dataset we used TfidfVectorizer on the reviews and transformed into document-term matrix and was used to train on support vector machines.

After pre-processing, we have split the dataset using train_test_split from Scikit-learn. We have used 70% of the dataset for training the model and rest for testing. Our goal was to build a model to help people who wanted to start restaurant business by looking at the positive and negative reviews given by yelp users for restaurants. We used four different models to train the dataset and classify the reviews. Logistic Regression model had the highest F1 score and hence we decided to choose it.

3.0 Experiments

3.0.1 Dataset Used

Name: Yelp dataset **Source:** https://www.yelp.com/dataset
Type: JSON files **No. of files:** 6 **Size:** 6.84 GB
File names: Yelp_academic_dataset_business.json, Yelp_academic_dataset_checkin.json, Yelp_academic_dataset_photo.json, Yelp_academic_dataset_review.json, Yelp_academic_dataset_tip.json, Yelp_academic_dataset_user.json
Pre-processing: Used json_to_csv_converter.py file to convert the 6 JSON files into separate CSV files

3.0.2 Methodology Used

Use Case 1 : Location and Cuisine Analysis

Training the data

We have used GridSearchCV and RandomizedSearchCV to train our cleaned data. Default value was chosen for cv=5. The 'liblinear' solver was chosen to implement lasso regression. Also, mean accuracy was used as an evaluation metric.

Modelling the data

The data was modelled separately for cuisine analysis and location analysis.

- **Decision tree** - The index maximum depth was set from 2 to 25 with a gap of 3. The evaluation metric was chosen to be accuracy again. (Fig B)
- **Gradient Boosted trees**- The n_estimators ranged from 3 to 40. The max_depth ranged from 3 to 40. The learning_rate ranged from 0.05 to 0.4. To train the model, cv was set to 20 over 50 iterations.
- **Random Forest Classifier**- gbtree booster was used and cv was set to 20 over 50 iterations. The best model predicted had gamma = 0.440899963585365 and learning_rate=0.13793184335330955

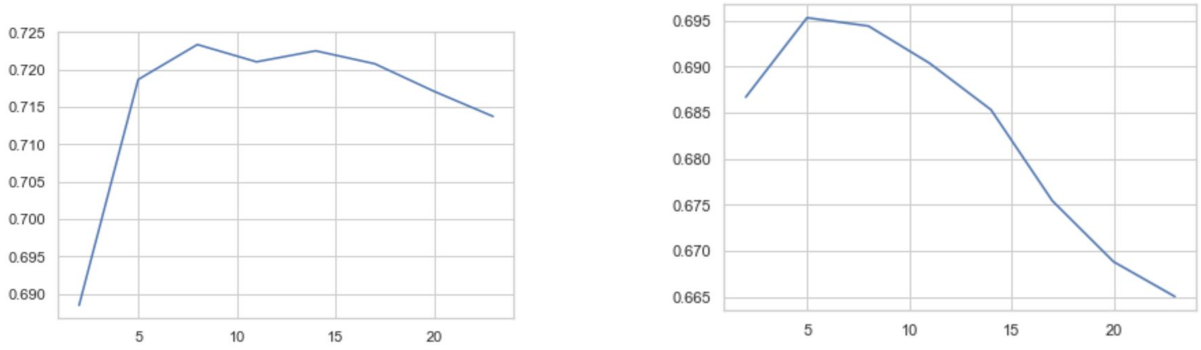
Use Case 2: Noise, Footfall pattern and Business Closure analysis

The dataset was split into train and test in the ratio of 70-30. SMOTE (Synthetic Minority Over-sampling Technique), a technique in which synthetic data samples are added to the minority class, was applied on the dataset to balance the imbalanced dataset. When the Linear SVC model was applied on the SMOTE data, it performed badly as compared to the other models.The following parameters were selected for the models:

XGBoost classifier	max_depth=7,min_child_weight=1
Random Forest classifier	min_samples_split=2, n_estimators=10, min_samples_leaf=1
Linear SVC	loss ='hinge', penalty='l2', tol=0.0001
Logistic Regression	default values
KNeighbors Classifier	n_neighbors=1, metric='minkowski'

3.0.3 Graphs and Plots

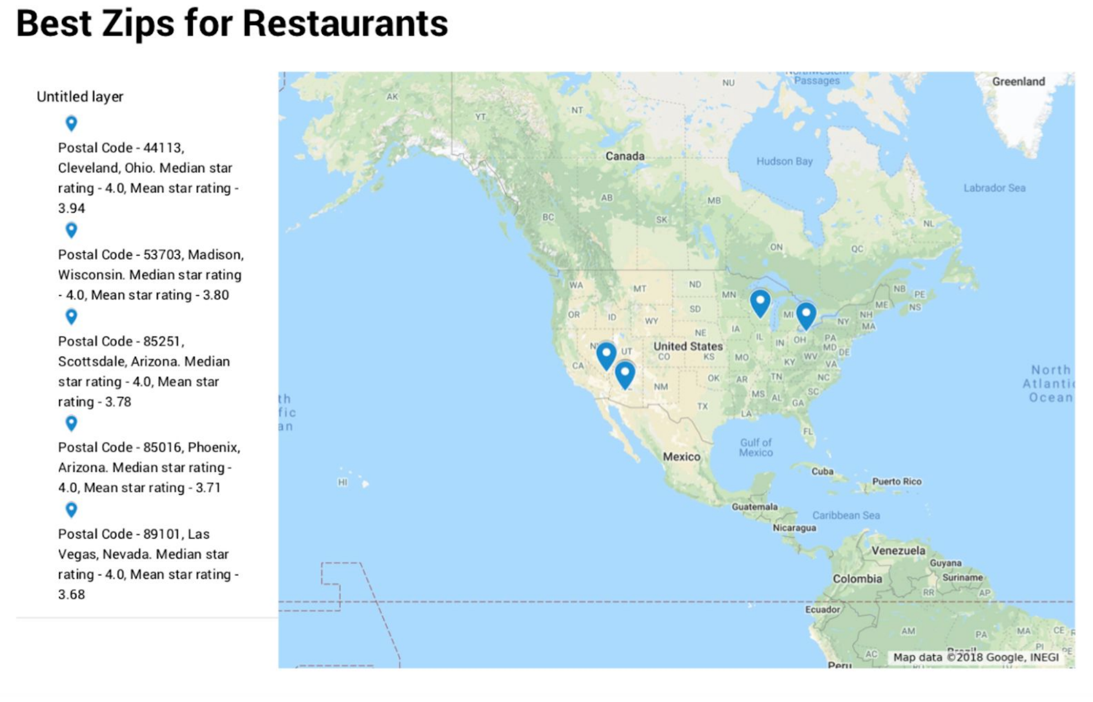
Below is a list of consolidated graphs and plots. However, due to space issues, we are sharing only the important ones.



(a) **(b)**

Fig 1: **(a)** Mean Accuracy score vs depth of tree for decision tree algorithm for cuisine to rating analysis

(b) Mean Accuracy score vs depth of tree for decision tree algorithm for geo data to rating analysis



Postal Code - 44113, Cleveland, Ohio. Median star rating - 4.0, Mean star rating - 3.94

Postal Code - 53703, Madison, Wisconsin. Median star rating - 4.0, Mean star rating - 3.80

Postal Code - 85251, Scottsdale, Arizona. Median star rating - 4.0, Mean star rating - 3.78

Postal Code - 85016, Phoenix, Arizona. Median star rating - 4.0, Mean star rating - 3.71

Postal Code - 89101, Las Vegas, Nevada. Median star rating - 4.0, Mean star rating - 3.68

Fig 2: Best zip codes to open a restaurant

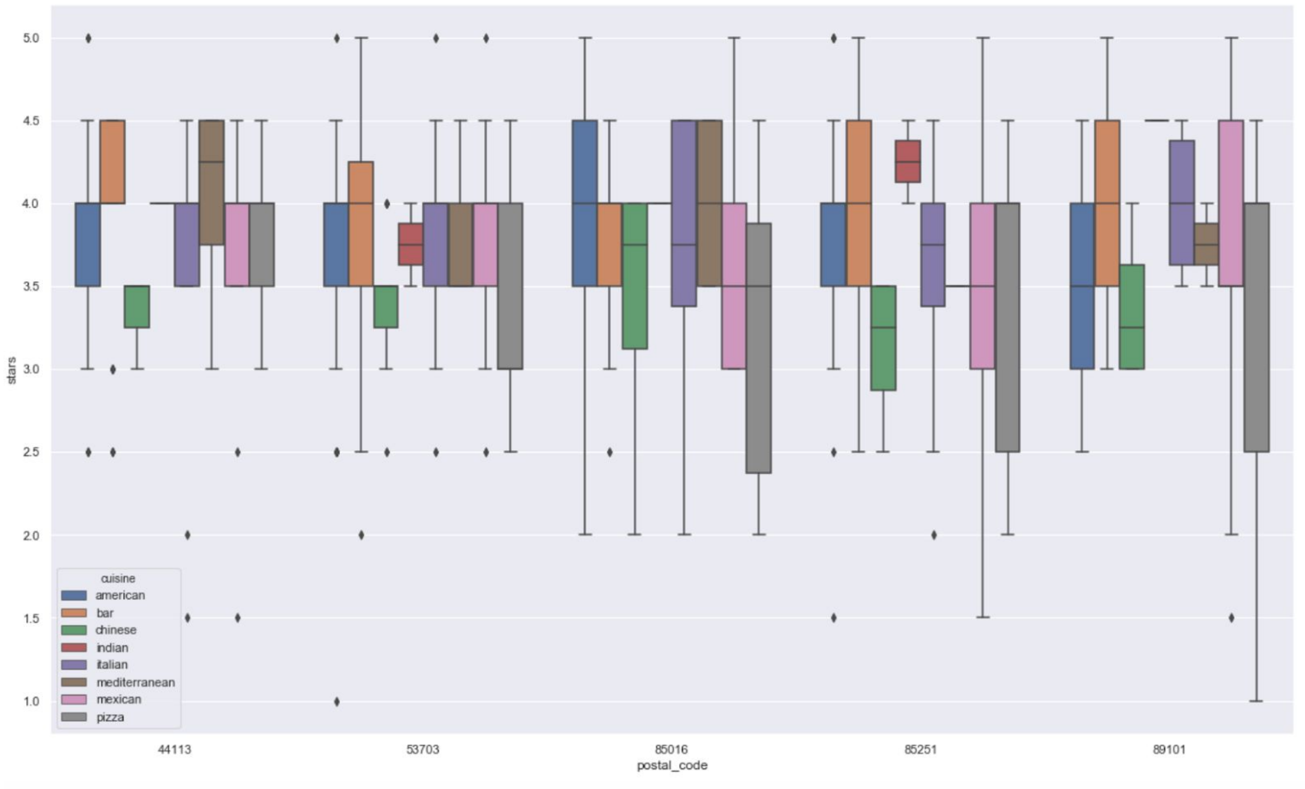
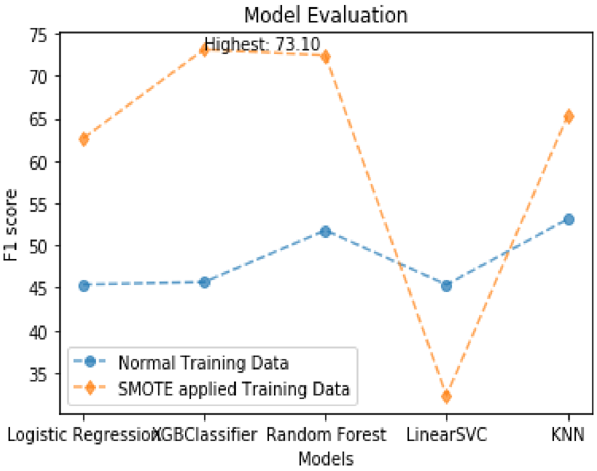
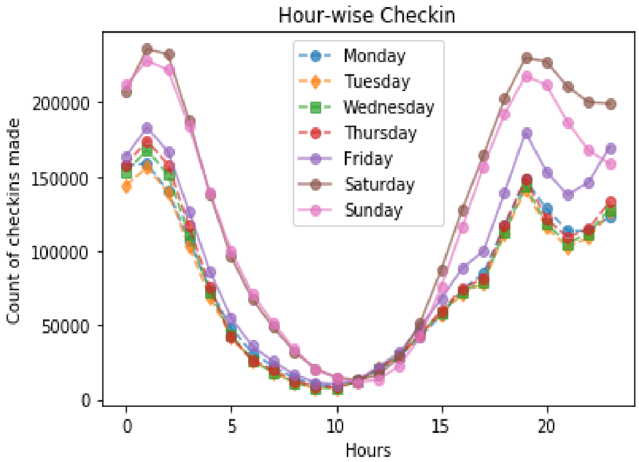


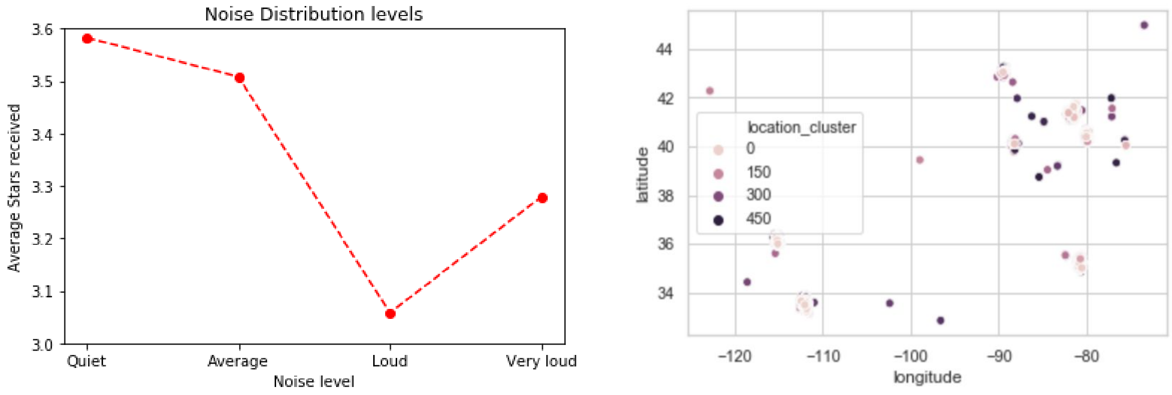
Fig 3: Cuisine to Star Correlation



(a)



(b)



(c) **(d)**

Fig 4: (a) Plot of F1 scores for predicting business closure **(b)** Plot of footfall pattern

(c) Plot of noise level and stars received **(d)** Cluster Analysis of restaurants based on latitude and longitude

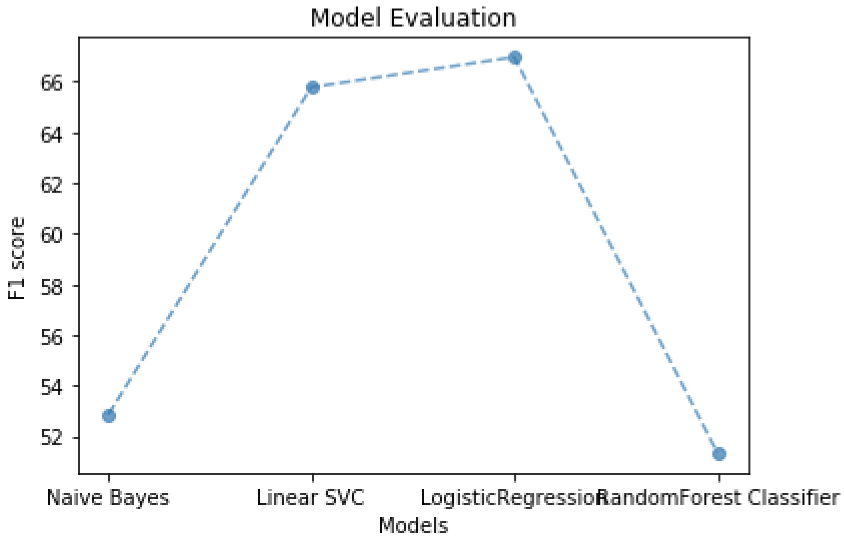


Fig 5: Plot of F1 scores for different model for classifying reviews

3.0.4 Analysis of Results

Based on our analysis we find that a higher level of noise considered as loud has a negative connection with the average rating received by a business. We were able to classify whether a business will close or not with a F1 score of 73%. We found that businesses have a peak time around 7-8 PM and then again around 1-3 AM. The overall footfall was found to be higher on weekends as expected.

We also found the gradient boosted trees perform the best when it comes to using location data to analyze star rating.

USE CASE	Algorithm	Score
Location and Cuisine Features affecting Star Rating(Analysis metric: accuracy)	Decision Tree	0.65
	Lasso Regression	0.695
	Gradient Boosted Tree	**0.75034**
	SVC	0.745
	Random Forest Classifier	0.712
Cuisine Features affecting star rating(Analysis Metric Used: Accuracy)	Decision Tree	0.746
	Lasso Regression	0.7233
	SVC	0.661
	Gradient Boosted Tree	**0.7461**
Business Closure Prediction	Logistic Regression	F1 score: 45.38, F1 score- SMOTE: 62.66
	XGBoost classifier	F1 score: 45.68, F1 score- SMOTE: 73.10
	Random Forest classifier	F1 score: 51.77, F1 score- SMOTE: 72.39
	Linear SVC	F1 score: 45.36, F1 score- SMOTE: 32.33
	KNeighbors Classifier	F1 score: 53.08, F1 score- SMOTE: 65.25
Analysing and classifying yelp restaurant reviews	Naive Bayes	52.85
	Random Forest classifier	51.34
	Linear SVC	65.78
	Logistic Regression	**66.96**

4.0 Discussion & Conclusions

4.0.1 Difficulties faced

1. Yelp dataset for reviews was really huge so we had to work on 30% of the original data for reviewing. After the reduction, the dataset was taking lot of time for performing data mining techniques.
2. As execution time was very high and when we had to make any small changes and to test it again it took lot of time to execute. We had to refresh the RAM regularly and use DEL command.

4.0.4 Conclusion

If a person is willing to start a restaurant business and decides to look at the yelp dataset, our review classification would help them by giving insight to make good business decisions according to the customer likes and dislikes. Based on our analysis, the assumptions made by us in selecting the success criteria has been proved right.

5.0 Project Plan and Task Distribution

Shivani Mangal

- Code can be found under folder ‘CMPE255_SMangal’
- Data exploration done on following features : stars, is_open, state, city, Review Count, Name, neighbourhood, postal code, categories, Latitude/Longitude, Address
- Cleaned data on above mentioned features
- Identified popular cuisine features
- **Clustering** done using **DBScan** to identify closeby business
- Analysis of transformed data conducted for modelling with **Lasso Regression**, **Decision Tree**, **Gradient boosted trees, SVC** for cuisine data
- Identified cuisine to star correlation and made inferences
- Identified Location to star correlation and made inferences
- Found best zip codes to open a restaurant
- Exploration of transformed data for Location features, review_count, categories based data using **Lasso Regression**, **Decision Tree**, **Gradient Boosted Trees, SVC** and **Random Forest Classifier**
- Worked on report and presentation

Hemang Behl

- Code can be found under CMPE255_Hemang
- Conversion of JSON files to CSV files
- Data exploration
- Perform data cleaning and merging on the datasets: business, review, checkin and user
- Implementation of classifiers [**Logistic Regression**, **Random Forests**, Linear Support Vector Classification (**SVC**), Extreme Gradient Boosting (**XGBoost**) classifier, **KNN** classifier] to predict whether a business will close or not
- Create the gender attribute by mapping the SSA baby names dataset with the Yelp user dataset
- Find whether noise leads to a poor rating for businesses
- Find the footfall pattern of businesses
- Project Report Documentation and Presentation

Prajwal Venkatesh

- Code can be found under CMPE255_Prajwal
- Data cleaning and pre - processing.
- Generated own dataset to classify the reviews.
- Classifying restaurant reviews as a positive or a negative review.
- Implementation of Linear Support Vector Classification (**SVC**), Logistic Regression, Random Forest and Naive Bayes classifiers.
- Report and presentation.

Appendix A: References

1. Mandabach, K. H., Siddiqui, M. A., Blanch, G. F., & Vanleeuwen, D. M. (2011). Restaurant viability: Operations rating of contributing success factors. *Journal of culinary science & technology*, *9*(2), 71-84.
2. https://www.yelp.com/dataset/

Appendix B

Scripts used:

Script name	Purpose
extract.py	Extracts all the files from tar file
json_to_csv_convertor.py	Converts all json files to csv
classifyingRestaurantCategoryfromBusiness.py	Generates a ‘FilteredRestaurantReviews.csv’ file with restaurant business_id’s and categories
getRestrauntReviewOnly.py	Generates a onlyRestaurants.csv file with only restaurant reviews and star ratings
predictNegativeOrPositiveReview.py	Predicts whether a given review is positive or negative.