Food Poisoning & Restaurant Business Success Prediction

By:

Rohan B Patel (010745904)

Introduction

- Being one of the largest and fastest-growing industries, Restaurant businesses have pressure to succeed and maintain their brand name.

- Yelp.com, is a crowd-sourced local business review and social networking site, where Yelp users can submit reviews of restaurants and their services using a one to five star rating scale.

- The main objective of this project is twofold -
  - to identify businesses that are likely to cause food poisoning to their customers
  - to predict if a business will be successful based on its current circumstances.

  - **Identifying food poisoning**: To identify restaurants that would lead to incidences of food poisoning, analyzed all the customer reviews for mention of any symptoms of or episodes of nausea and stomach infections.
  - **Restaurant success**: To identify the success of the restaurant, analyzed the location (address, longitude and latitude), cuisine type (categories), number of stars and count of reviews.

- Moreover, also applied appropriate data mining and machine learning techniques to achieve the apt objective of this project.

FOOD POISONING PREDICTION

Data PreProcessing

- **Subset of Data**
  - Businesses with category as “Restaurant” or “Food” were selected.

- **Feature Selection**
  - Selected relevant features from review and business files and merged them into one.

- **Stop word Removal**
  - Removing stopwords from the text part of the review.

- **Lemmatization**
  - The text is lemmatized using WordLemmatizer, Whitespace Lemmatizer, pos_tag and word_tokenize.

Data Labelling

➔ Since the dataset is unlabelled, a variable is assigned to all the review records based on the occurrence of words related to any food poisoning or their symptoms and the rating given by the user (stars).

Algorithms

- ★ Naive Bayes
  - Computing time : 10 secs
  - Accuracy : 97.4%

- ★ KNN
  - Computing time : 15-20 mins
  - Accuracy : 98.1 %

Data Aggregation

- ➔ Upon the classification of data by the model, the probability of a business that a customer will acquire food poisoning from that specific restaurant is calculated.
- ➔ The set threshold is 30%.
- ➔ Around 35 restaurants were classified to have above 30% probability.

Experimenting with Review.csv

- ➔ Prior to deducing this model, experimented using attributes and categories of business as features for training the model. But most of these features were found to be irrelevant to food poisoning.
- ➔ Here, Decision tree was employed and aggregate methods such as XGBoost and Random Forest were applied.
- ➔ Highest accuracy achieved was 50% with XGBoost model. Other 2 models were accurate just upto 20-30%.

RESTAURANT BUSINESS SUCCESS PREDICTION

Data Pre-processing

- ➔ Explored the stars, is_open, state, city, review count, name, postal code, categories, address, longitude, and latitude of the yelp_business dataset.

- ➔ Examined cuisine to stars by displaying it in order to better grasp the data.

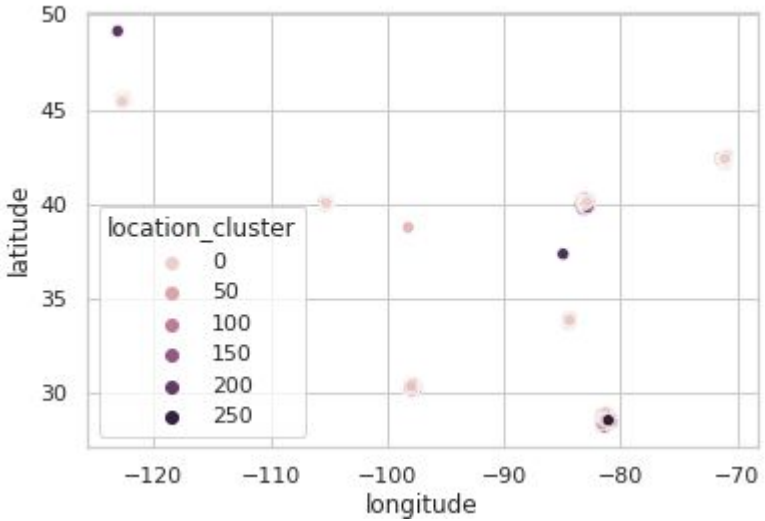
- ➔ Discover the relationship between location, stars, and cuisine.

Analysis Done on Geo Location & Cuisine Data

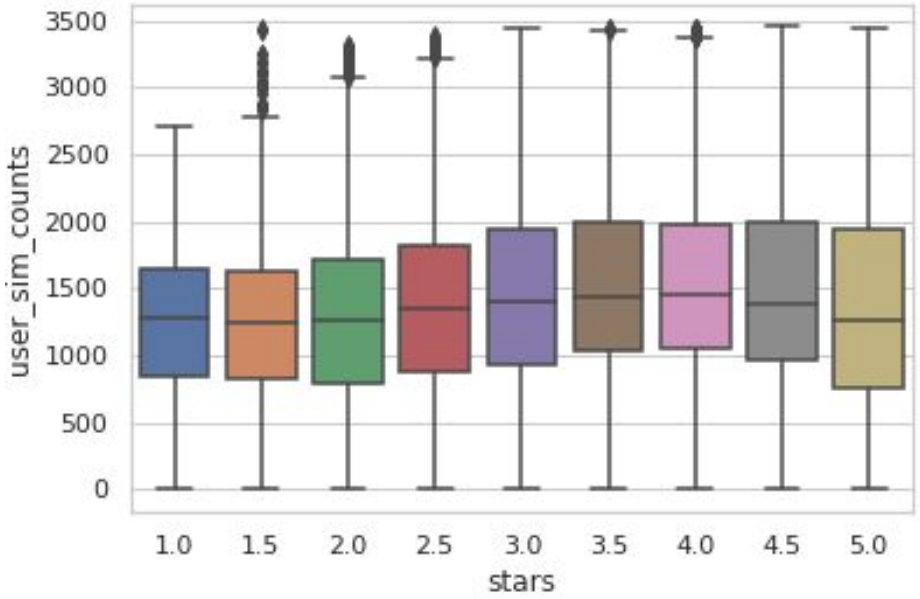
- ➔ Cuisine to Stars Correlation
- ➔ Location to Stars Correlation
- ➔ Finding best postal codes for restaurants
- ➔ Clustering Latitude & Longitude to identify closeby business
- ➔ Classification of restaurant based on location based features
- ➔ Classification of restaurant based on location features and cuisine features

Clustering Restaurants

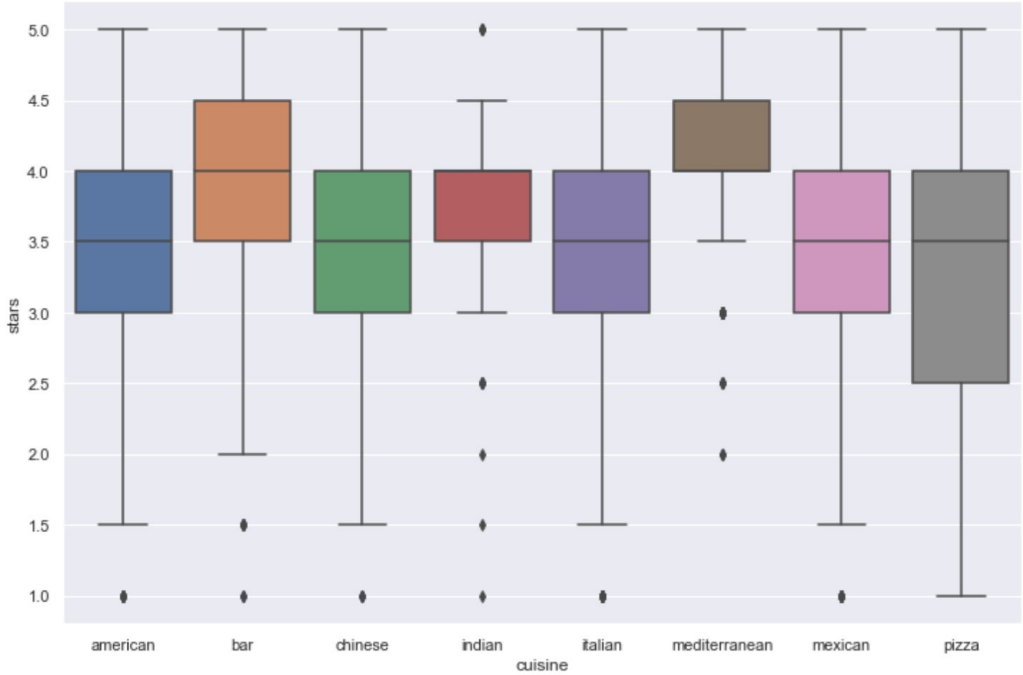
➔ Accomplished via DBScan



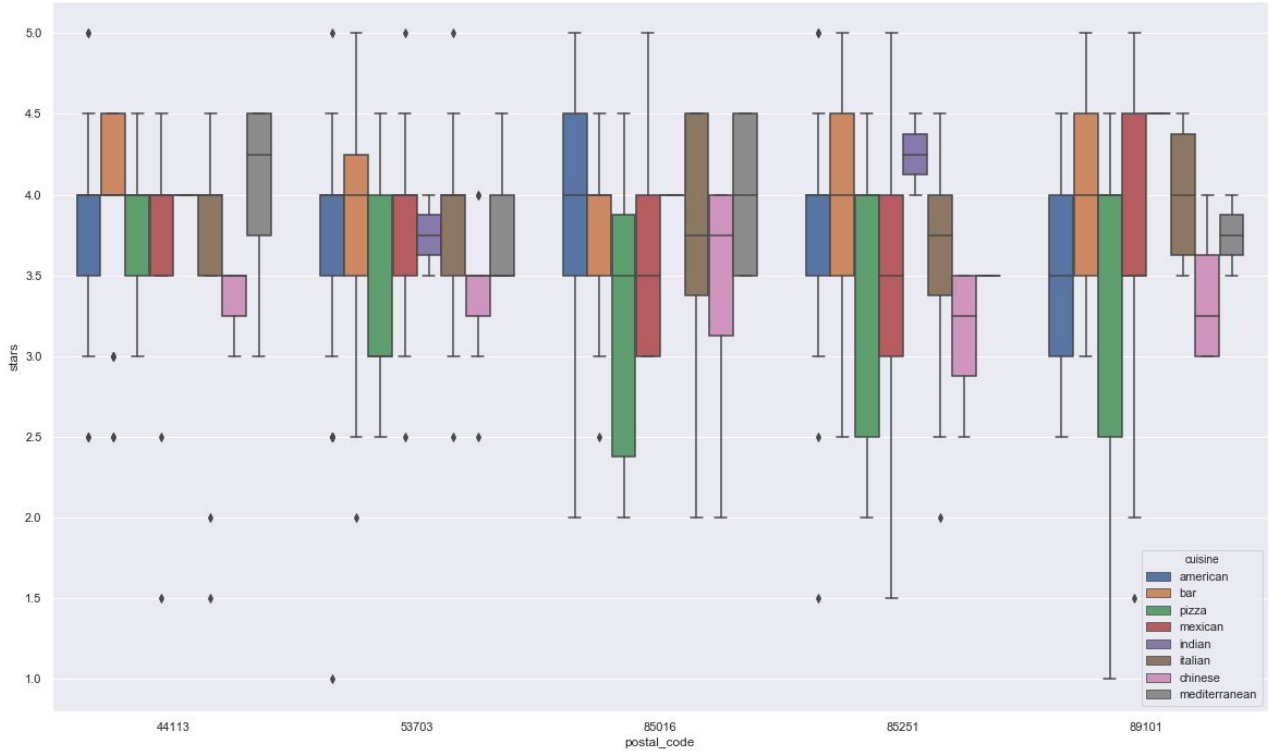
Estimate of nearby restaurant with similar users visiting



Cuisine to Stars Correlation



Cuisine to Location to Stars Correlation



Models Employed & Established Accuracy

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

➔ Best accuracy was established via employing Gradient Boosted Trees.

Results

➔ Gradient boosted trees perform the best when it comes to using location data to analyze star rating with a score of 75%

Conclusion

- ➔ If someone wishes to open a restaurant and view the Yelp dataset, the **Restaurant Business Success Prediction** will assist them in making appropriate business decisions based on client preferences.

- ➔ If a customer wishes to judge a restaurant based on the quality of food, or if someone wants to inspect a restaurant for food poisoning cases based on customer reviews, they can make use of the **Food Poisoning Prediction** section of this Project.

Thank You !