PREDICTING RESTAURANT SUCCESS BASED ON YELP DATASET

Team #04 Watson
Shivani Mangal (012530362)
Hemang Behl (013734214)
Prajwal Venkatesh (012557792)

Objective

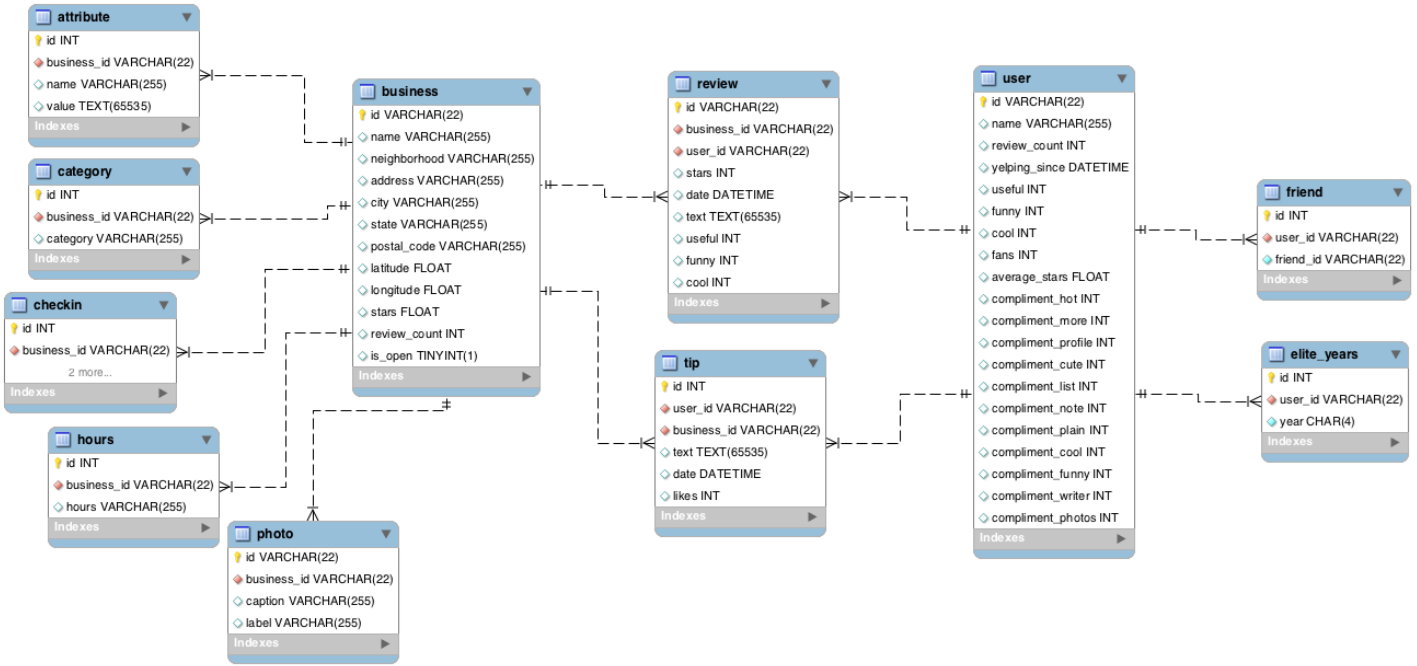
- The restaurant industry is the second largest employing industry in United States and it’s demand is an ever-increasing one

- Yelp, one of the most used websites today, is a pool of data.

- The main objective of our extensive analysis on the yelp dataset is to aid in the success of local restaurants.

- In order to understand and predict the success of a restaurant, we have taken three approaches

Anatomy of the Yelp Dataset

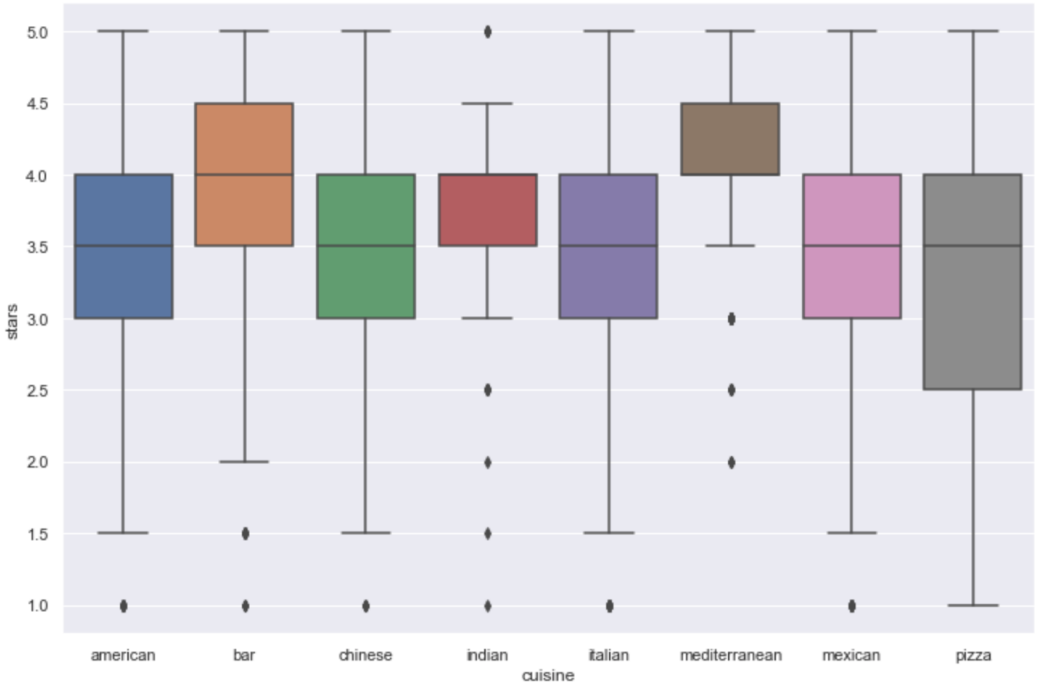


Approach 1 : Where should my restaurant be?

- **Attributes Addressed** : Stars, Review Count, Geo-Location Features, User data
- **Problems Faced** :
  - Wrong state/city combination
  - Null Address
- **Feature Engineering**
  - Clustering Latitude and Longitude to identify closeby business using DBSCAN
  - Count of nearby restaurants with similar categories
  - Count of nearby restaurant with similar user rating
  - Count of nearby restaurants with similar users visiting
  - Bucketizing review counts into 10 quartile buckets
- **Model Tuning** : GridSearchCV and RandomizedSearchCV for Cross Validation
- **Models Used : Lasso Regression, Decision Tree, Random Forests, XGBoost Classifier, SVC**
- **Best Model : XGBoost Classifier with an accuracy of 74.29%**

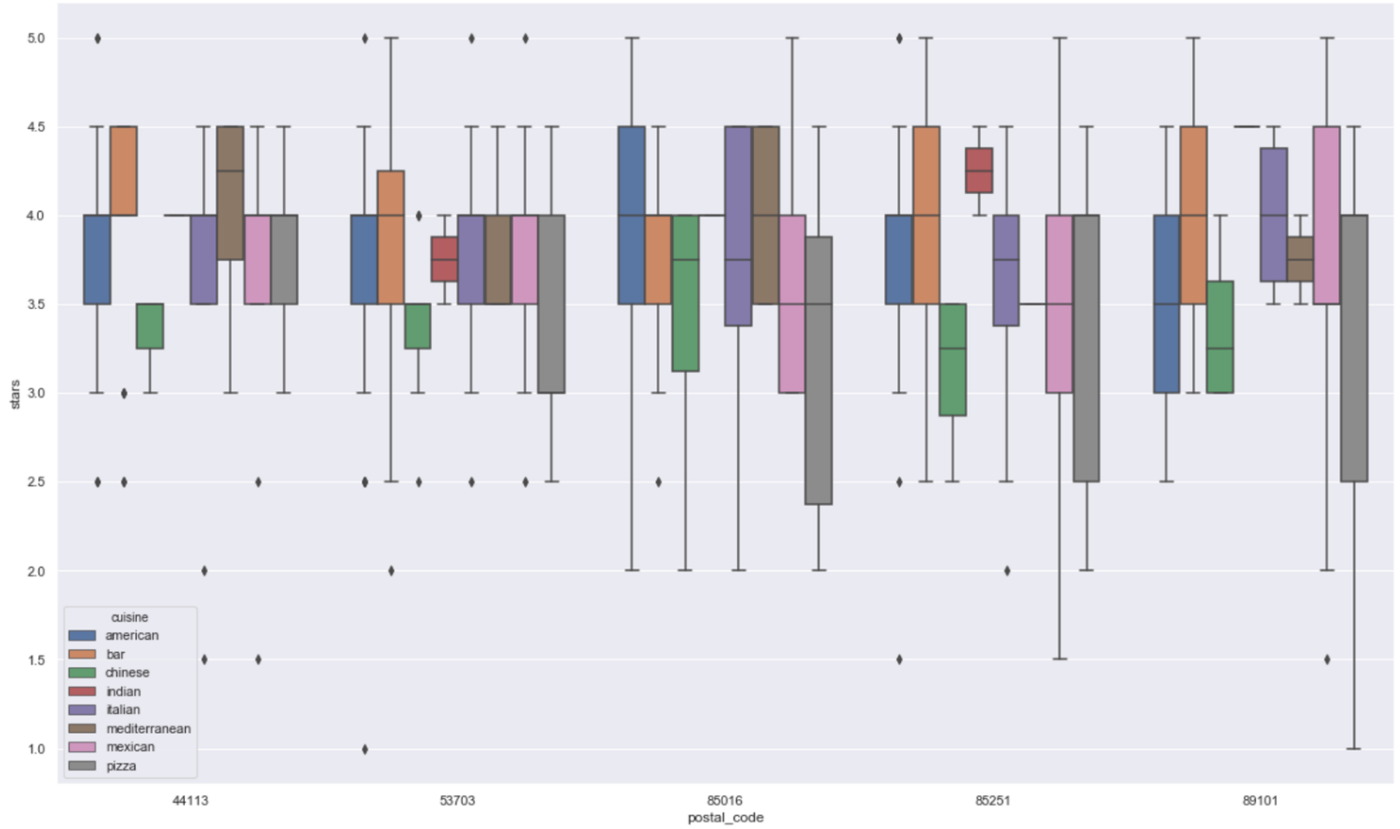
```
def get_user_sim_counts(temp):
    X = normalize(temp[['mean_user_review_counts', 'mean_months_since_yelping', 'mean_user_fans', 'mean_total_compliments']])
    cosine_sim = linear_kernel(X)
    user_sim_counts = np.apply_along_axis(count_similar, 1, cosine_sim)
    return user_sim_counts
```

Interesting Finds



- People really like their bars!
- Tough business to get into- Pizza
- A hopeful message for a business owner, there will always be someone who doesn't like your food, that's ok, they're an outlier!
- opportunity for any aspiring Indian restaurant owner to get to the top of bunch

Interesting Finds ..continued



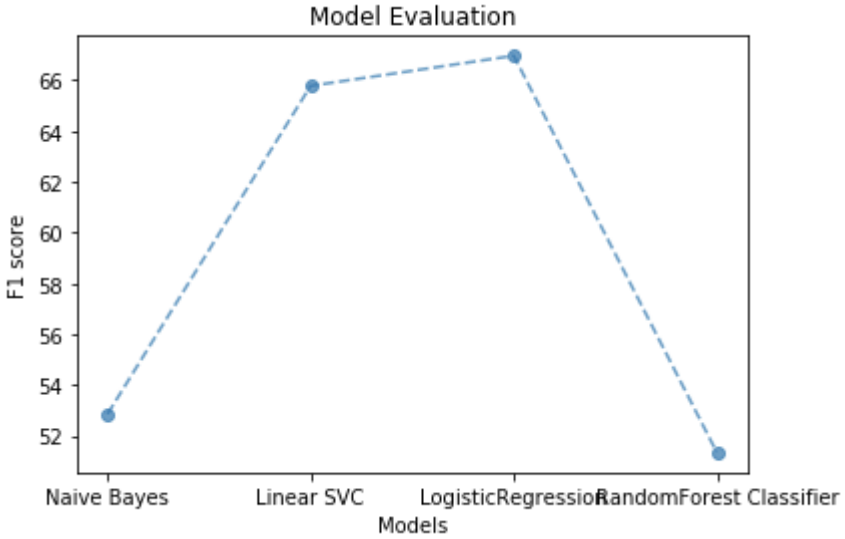
Approach 2: What about the reviews?

- Dataset: 600,000
- Large dataset, used only a small chunk as it took more time to process.
- Dataset split - 70% for training and 30% prediction.

Pre-processing

- Used NLTK library to remove stopwords and punctuations.
- Used CountVectorizer to convert the text documents into a matrix of token counts.
- Used TfidfVectorizer on the reviews and transformed into document-term matrix.
- Reviews with 1 and 2 stars classified as Negative review and 3-5 stars as Positive reviews.

Models used

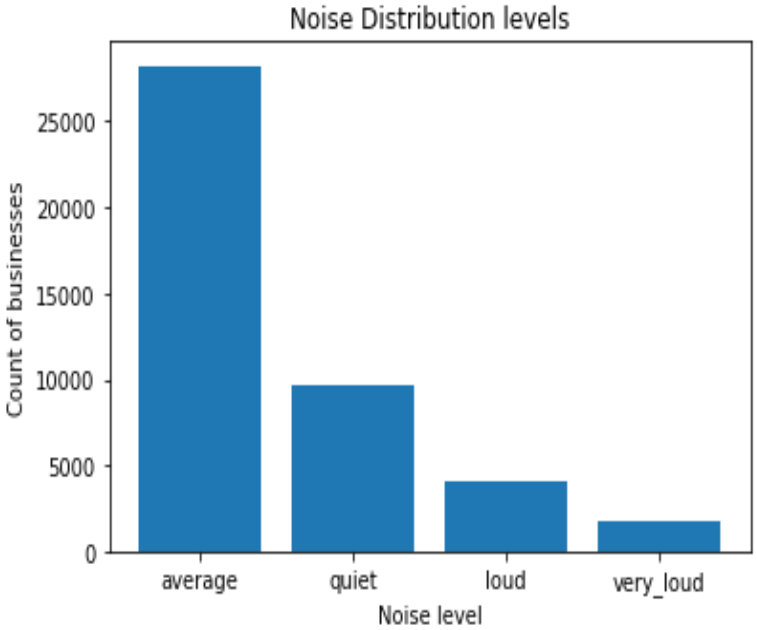


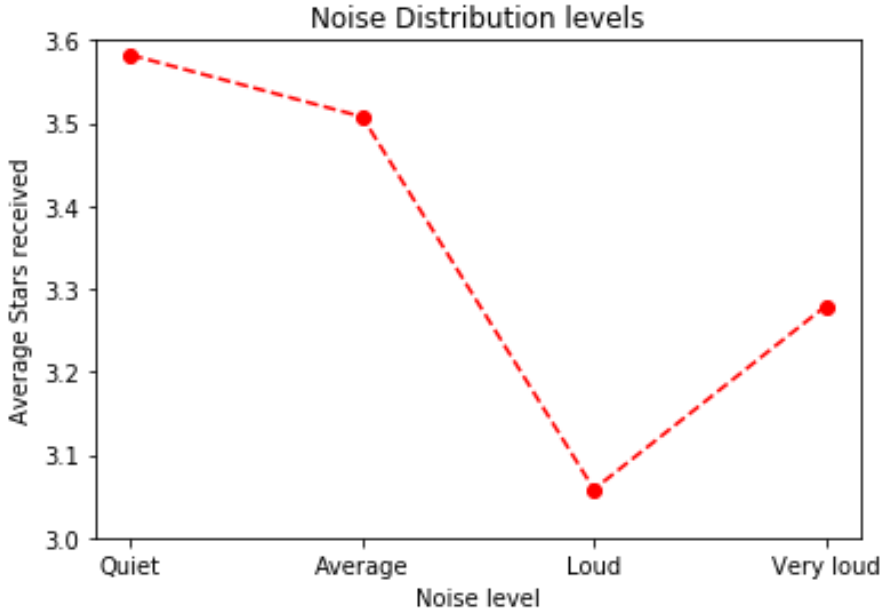
Model	F1 Score
Naive Bayes	52.85
Random Forest classifier	51.34
Linear SVC	65.78
Logistic Regression	**66.96**

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

Approach 3:
Noise, Footfall Pattern and Business Closure analysis

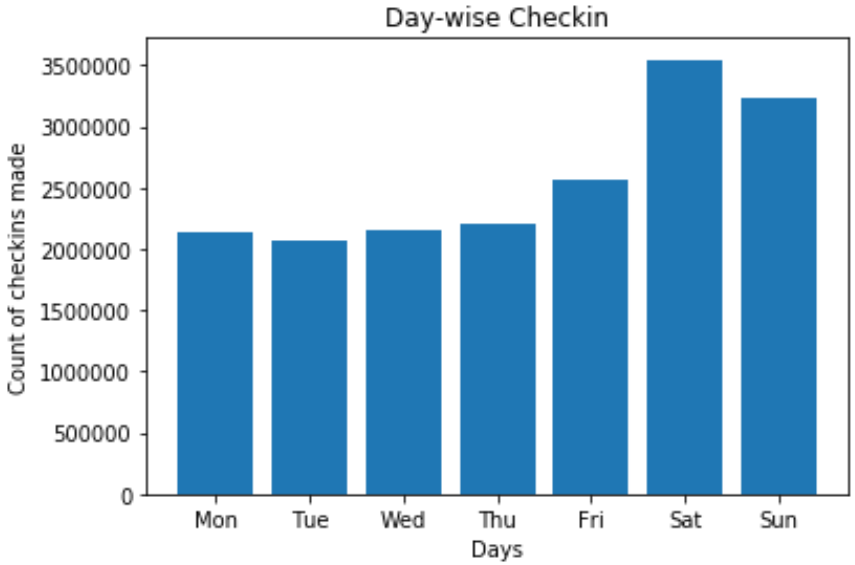
Find how noise levels effect star ratings

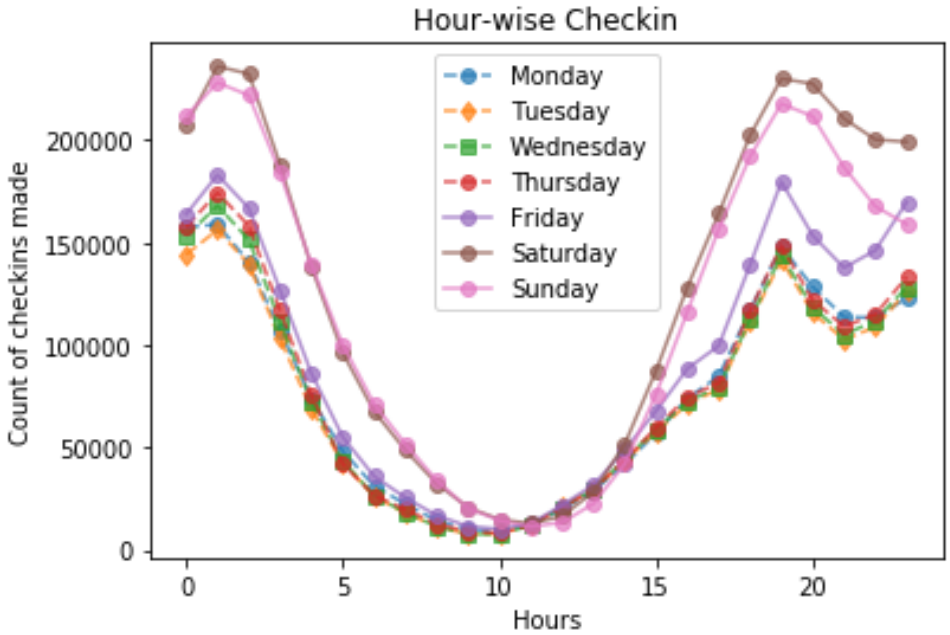




People tend to give less ratings to places with high levels of noise

Footfall Pattern Analysis

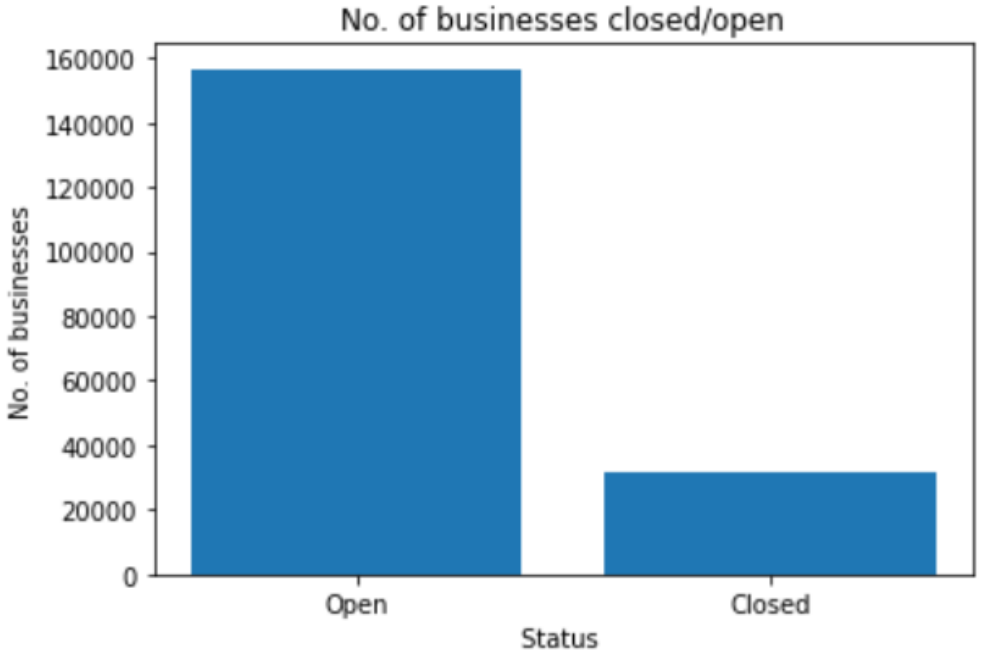




We found that businesses have a peak time around 7-8 PM and then again around 1-3 AM. The overall footfall was found to be higher on weekends as expected.

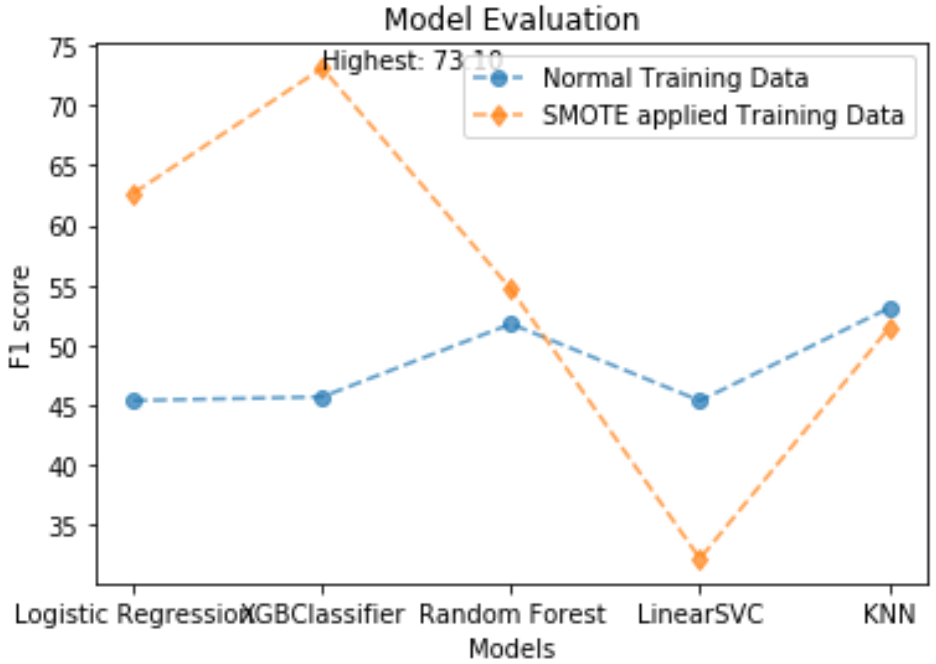
Will a business close or not

Merged datasets: review, business and checkin
-Imbalanced data



We are able to predict whether a business will close or not with a F1-score of 73.1%

MODEL	F1 score	F1- SMOTE
Logistic Regression	45.38	62.66
XGBoost classifier max_depth=7, min_child_weight=1	45.38	62.66
Random Forest classifier min_samples_split=2, n_estimators=10, min_samples_leaf=1	45.68	73.10
Linear SVC loss ='hinge', penalty='l2', tol=0.0001	51.77	54.66
KNeighbors Classifier n_neighbors=1, metric='minkowski'	45.36	32.19



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.

Thank You

Source
https://www.yelp.com/dataset/