CS 412: Introduction to Machine Learning

Project Final Report

Yashashvini Rachamallu (UIN: 650785365)

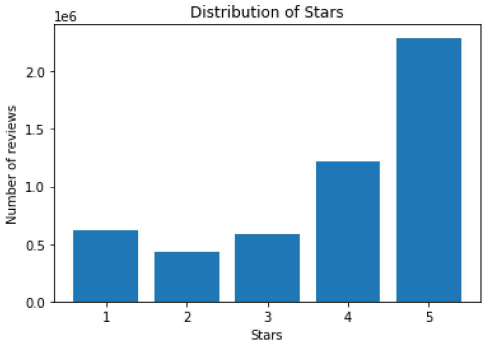
1. Project: A brief review of Machine Learning algorithms and a recommendation system
2. Dataset: Yelp Dataset
3. Steps:

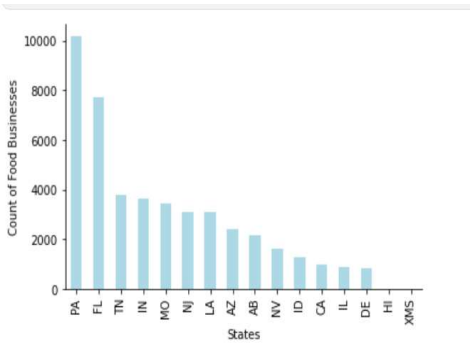
   The machine learning project has different stages. I made sure my project goes through each of those stages.

   a. **Data Reading/Loading Stage**:
      i. The Yelp dataset is a part of larger yelp business and user dataset. The data contains various information as JSON files which include:
         1. Business data including location data, attributes, and categories.
         2. Contains full review text data including the user_id that wrote the review and the business_id the review is written for.
         3. User data including the user's friend mapping and all the metadata associated with the user.
         4. Tips data which mentions the tips and suggestions given by users.
         5. Checkin data which contains the checkin information for the
      ii. Each of the dataset is read and converted into CSV for ease of access

   b. **Data Cleaning and Manipulation Stage:**
      i. All the NaN values in 5 datasets have been filled or replaced as per requirement.
      ii. Missing values in business_df like pincode have been filled using GeoPy library
      iii. Missing values in review_df where there was no data for any feature have been dropped
      iv. Missing values for hours of operation and address have been filled with “Unknown”

   c. **Data Analysis:**
      i. Some data analysis has been conducted to observe patterns in data





   d. **Classification Modelling:**
      i. **Goal**: Perform sentiment analysis using classification
      ii. **Algorithms implemented:** Logistic Regression, Naive Bayes, K-NN Classifier
      iii. **Procedure**:
         1. **Data Preparation:**
            a. The text data from both the reviews and tips is pre-processed using NLTK package
            b. Later, TF-IDF was used as vectorizer to perform text vectorization of both review and tips text data.

c. Dropped NaN values.
d. Later, some necessary pre-processing was done
e. We calculated the sentiment of existing texts by the below method:
    i. If Stars < 3.5, then it is Negative sentiment
    ii. If Stars > 3.5, it has Positive sentiment

2. **Training/Modelling:**
    a. The vectorised text data has been split with a 80-20 train test split and in a randomized way.
    b. Three classification models have been implemented for both reviews and tips data.
        i. Logistic Regression, Naive Bayes, K-NN Classifier
    c. Due to system restrictions some models were not trainable on the available resources.
    d. Later, a voting classifier was implemented and trained as a part of ensemble learning.
3. **Testing:**
    a. The performance has been tested using appropriate metrics

Model	Reviews	Tips
Logreg	69.36%	62.7%
GNB	67.66%	53%
SVC	-	-
KNN	-	51.47%
Ensemble	69.18%	51.47%

    b. Reasons for low performance:
        i. Not able to do more than 1 gram analysis due to large data
        ii. We were not able to perform bagging and boosting due to system restrictions, due to which ensemble model was not able to get higher accuracy.

4. **Recommendation System:**
    a. **Goal**: Recommend restaurants based on user preferences
    b. **Procedure**:
        i. **Data Preparation:**
            1. The business data is prepared here for recommendation
            2. The attributes of business like parking, price range, Alcohol etc are
            3. The categories like cuisines have been separated using “get_dummies” method

ii. Feature selection:

1. Based on correlations the following features have been selected to implement the recommendation system

Breakfast/Brunch/Dinner	Cuisines	Price for 2
Ratings	Alcohol preference	WiFi Preference
Parking requirement	Delivery/Pickup	Distance

iii. Recommendation system

1. Inputs

```
Please press 1 to take the current location, if not press 2 2
Please enter the address: 101 Walnut St,18054
Do you have any specifications like Cusines,Price, Parking etc.. If so type "Yes" : yes
What do you prefer among the "Breakfast","Brunch","Dinner" : Breakfast
Types of Cusines:  ['Pizza', 'Mexican', 'American', 'Italian', 'Indian', 'Pakistani', 'Thai', 'Japanes
e', 'French', 'German', 'Vietnamese', 'Chinese', 'Hungarian']
Pick a cuisine from the above list: Pizza
Please enter your specification price for two (1-5 range) : 2
Types of Parking:  ['street_park', 'validated_park', 'lot_park', 'valet_park']
Please enter your specification of parking Type from the above list:  lot_park
Any preferences of rating ranged from (0-5) : 4
If restuarent has delivery ("False" or "True"): True
If you need wifi type "Yes":
If you need alchol type "Yes":  Yes
If you recommendation of the nearest distances type "yes" :
```

2. Output:

a. Outputs the Map of top 10 restaurants