Yelp Data Analysis

Arpita Sheth (110422828)

Introduction

Analysis of data is a process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information. In this assignment, we use Data Analysis to classify businesses, predict ratings, categorise users and how they talk about different businesses using the reviews they provide for each business. Yelp provides Academic Dataset consisting of user reviews for businesses in over 16 states of USA.

Problem Statement

The Yelp Academic Dataset is used for data analysis. The data has granular details about users like name, average number of stars awarded by user, count of reviews written so far etc. It also contains data about reviews and businesses like number of reviews a particular business gets and number of user review mappings etc. The data is provided by Yelp and the objective is to find trends, classify businesses, predict which user likes what service etc.

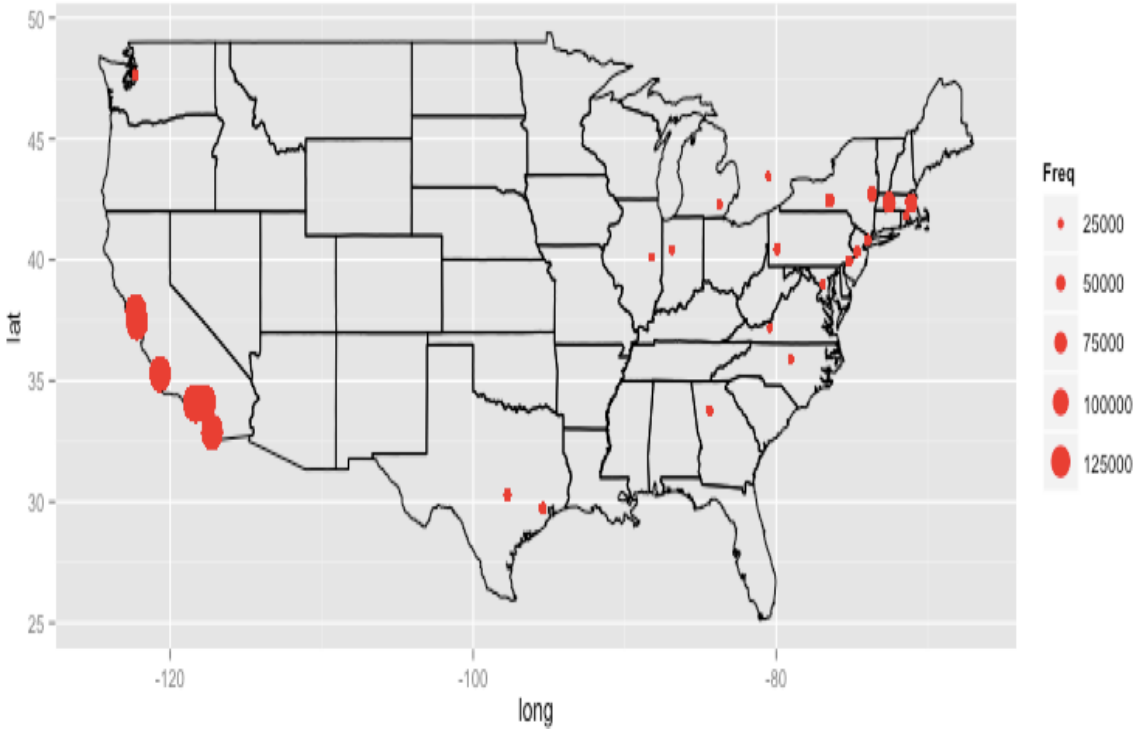
Data Acquisition and Pre - Processing

The data can be obtained using Yelp API and contains a json file containing User objects, Review objects and Business objects and type filed distinguishing each object. As a prerequisite the data is split into User objects, Review Objects and Business Objects.

Observation, Techniques Used and Results

1. **What does the data look like?**

The given dataset is categorical spatial data and text data. Spatial Data Analysis and Text Analysis can be used. The following heat map shows the number of reviews across businesses quantified state wise. It can be seen that California businesses have the largest number of reviews with total of 130182 reviews across all businesses.



2. **How do people talk about different businesses and the service offered?**

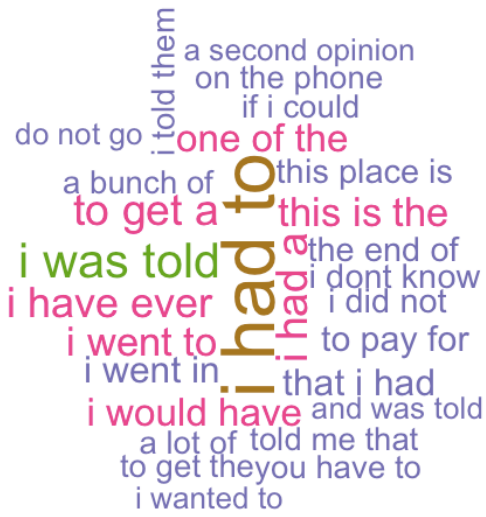
Applying **Sentiment analysis** or **Opinion Mining** using Natural Language Processing and Text Analysis to Review data helps in identifying and extracting subjective information about what people talk about different businesses and how it impacts a business' ratings.

1-star user reviews

5-star user reviews

Since California has maximum business-reviews ratio, let us consider California businesses for analysis. The above word clouds indicate that the reviews for 1 star rated businesses and services contain more negative feedback with words such as "never", "dont", "time", "bad" etc. being most dominant in the reviews. On the other hand, reviews for 5-star rated businesses and services tend to be more positive with feedbacks containing mainly of words such as "best", "great", "good", "friendly" etc.

However, a single word does not give the sense of an entire sentence, so we use n-grams. An n-gram is a contiguous sequence of n items in the given text. For analysing the review text data, we tokenize data into trigrams i.e 3-grams.



3-grams for 1-star businesses

3-grams for 5-star businesses

As it can be illustrated from the 3-grams representation, 1-star business reviews contain more of complaining feedback whereas, 5-star business reviews have phrases likely "i highly recommend" occurring repetitively.

3. **How do people rate different businesses?**

We use a **Linear Regression** model to predict how a user may rate a business or service. The features used to construct the model are the number of reviews given so far by a user, average rating made by user, average rating of a given business, number of reviews about a business. We randomly split the data into train set, test set with a 80-20% weightage and validation set contains 5% of train set. We predict the star rating using this model.

```
PROGRESS: Creating a validation set from 5 percent of training data. This may take a while.
          You can set ``validation_set=None`` to disable validation tracking.

PROGRESS: Linear regression:
PROGRESS: --------------------------------------------------------
PROGRESS: Number of examples          : 251041
PROGRESS: Number of features          : 4
PROGRESS: Number of unpacked features : 4
PROGRESS: Number of coefficients    : 5
PROGRESS: Starting Newton Method
PROGRESS: --------------------------------------------------------
PROGRESS: +-----------+----------+--------------+--------------------+----------------------+---------------+--------
---------+
PROGRESS: | Iteration | Passes   | Elapsed Time | Training-max_error | Validation-max_error | Training-rmse | Validat
ion-rmse |
PROGRESS: +-----------+----------+--------------+--------------------+----------------------+---------------+--------
---------+
PROGRESS: | 1         | 2        | 0.094297     | 4.271612           | 3.993659             | 0.972112      | 0.97484
4        |
PROGRESS: +-----------+----------+--------------+--------------------+----------------------+---------------+--------
---------+
```

Using the model to predict on the test set gives the following results. Showing the first five predictions:

```
dtype: float
Rows: 5
[3.5085189786454256, 4.301810596875041, 1.3471779588465 15, 2.729123907864972, 4.120282144746131]
```

Since the model uses linear regression it is hard to limit the predictions between 1 to 5. We evaluate the model by comparing the predictions with actual ratings. The root mean square error and max error of the model predictions are :

```
{'max_error': 3.9448841188878454, 'rmse': 0.972253261683395}
```

As it can be seen, the linear regression model can still fairly predict the user star rating for a particular business. For each feature used for training, the model learns weights for each feature. The magnitude of the coefficient for each feature indicates the strength of the feature's association to the target variable, *holding all other features constant*. The sign on the coefficient (positive or negative) gives the direction of the association. α_0 is the intercept and X_j are training features. α_j are coefficients learnt by the model.

$Y=\alpha_0 + \sum_j \alpha_j X_j + \epsilon$

name	value
(intercept)	-2.25438208159
user_avg_stars	0.789304824857
business_avg_stars	0.812517741351
user_review_count	2.90326745581e-05
business_review_count	1.4848374946e-05

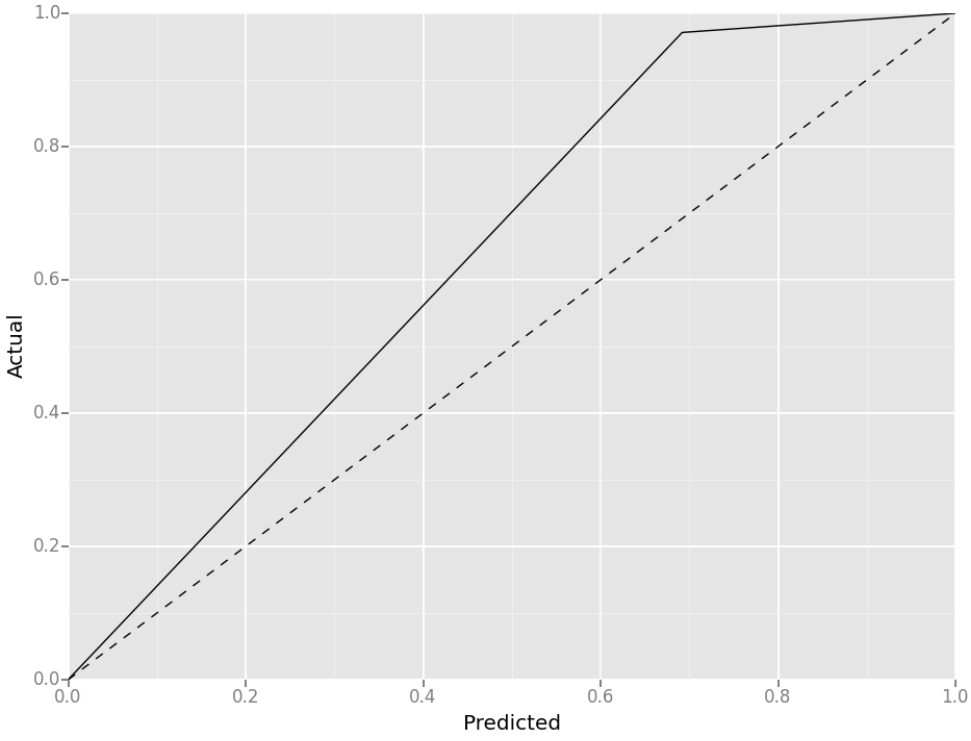
4. Is a business superior or inferior?

We use a **Logistic Regression** model for binary classification of businesses with the notion that a business is deemed superior if it has an average star rating of equal or above 3 and inferior otherwise. To perform logistic regression, we first create a binary target variable to denote if the business is superior or inferior and model the probability of the binary target being True as the logistic function. The data is split 80-20 % in train set and test set.

```
PROGRESS: Logistic regression:
PROGRESS: --------------------------------------------------------
PROGRESS: Number of examples          : 250930
PROGRESS: Number of classes           : 2
PROGRESS: Number of feature columns   : 4
PROGRESS: Number of unpacked features : 4
PROGRESS: Number of coefficients    : 5
PROGRESS: Starting Newton Method
PROGRESS: --------------------------------------------------------
PROGRESS: +-----------+----------+--------------+-------------------+---------------------+
PROGRESS: | Iteration | Passes   | Elapsed Time | Training-accuracy | Validation-accuracy |
PROGRESS: +-----------+----------+--------------+-------------------+---------------------+
PROGRESS: | 1         | 2        | 0.629864     | 0.846499          | 0.851057            |
PROGRESS: | 2         | 3        | 0.914685     | 0.849986          | 0.852896            |
PROGRESS: | 3         | 4        | 1.241481     | 0.850293          | 0.852130            |
PROGRESS: | 4         | 5        | 1.527377     | 0.850277          | 0.851900            |
PROGRESS: | 5         | 6        | 1.797984     | 0.850265          | 0.851900            |
PROGRESS: +-----------+----------+--------------+-------------------+---------------------+
```

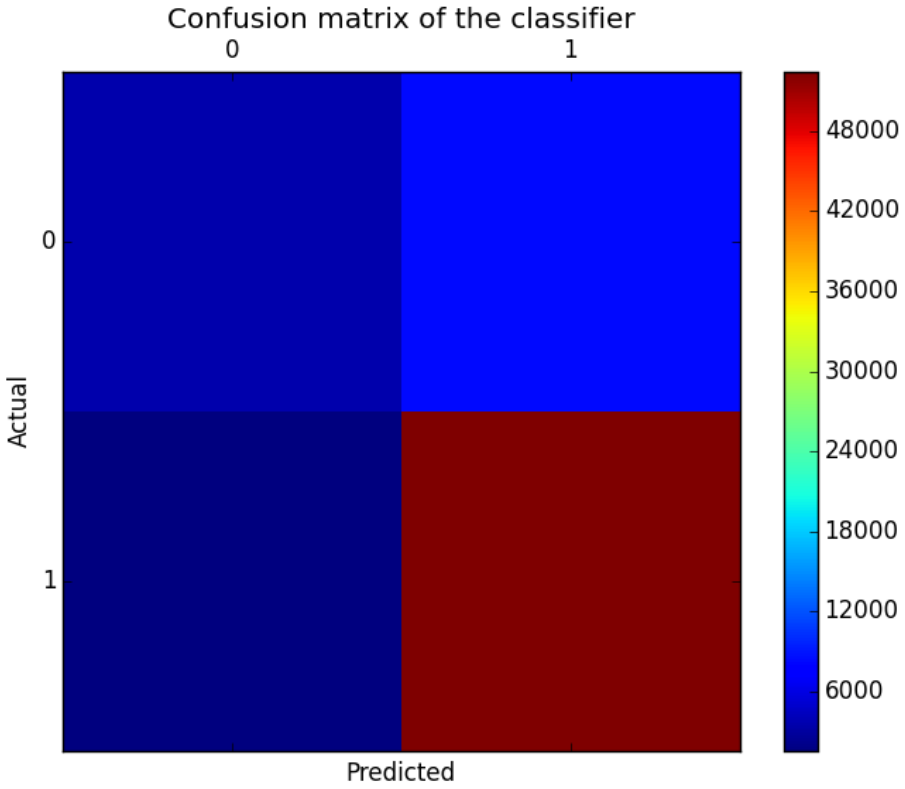
The model is used to classify businesses in test set. We evaluate the model by comparing the classification with original values and calculate the accuracy. Accuracy of the model is given as : **Accuracy : 0.849521100334**

Plotting the predicted vs. actual values gives the following curve:



The confusion matrix for the binary classification model can be given as follows:

```
Confusion Matrix :
+--------------+-----------------+-------+
| target_label | predicted_label | count |
+--------------+-----------------+-------+
|      0       |        1        |  8389 |
|      0       |        0        |  3736 |
|      1       |        1        | 52408 |
|      1       |        0        |  1556 |
+--------------+-----------------+-------+
[4 rows x 3 columns]
```



5. Own a business? Know where it stands!

Multiclass classification classifies instances into one of the more than two classes. In the yelp data, we can classify businesses in 5 classes according to ratings from 1 to 5. We develop a model to predict the rating of a business and classify it accordingly.

```
PROGRESS: --------------------------------------------------
PROGRESS: Number of examples          : 250738
PROGRESS: Number of classes           : 5
PROGRESS: Number of feature columns   : 4
PROGRESS: Number of unpacked features : 4
PROGRESS: Number of coefficients    : 20
PROGRESS: Starting Newton Method
PROGRESS: --------------------------------------------------
PROGRESS: +-----------+----------+--------------+-------------------+---------------------+
PROGRESS: | Iteration | Passes   | Elapsed Time | Training-accuracy | Validation-accuracy |
PROGRESS: +-----------+----------+--------------+-------------------+---------------------+
PROGRESS: | 1         | 2        | 0.774381     | 0.432639          | 0.436726            |
PROGRESS: | 2         | 3        | 1.273298     | 0.459651          | 0.456509            |
PROGRESS: | 3         | 4        | 1.775387     | 0.461294          | 0.455603            |
PROGRESS: | 4         | 5        | 2.217805     | 0.461717          | 0.456207            |
PROGRESS: | 5         | 6        | 2.678101     | 0.461733          | 0.456056            |
PROGRESS: | 6         | 7        | 3.140723     | 0.461733          | 0.456056            |
PROGRESS: +-----------+----------+--------------+-------------------+---------------------+
```

We use the model to predict the classes for businesses in the test set. The probabilistic estimate that a business belongs to each class can be shown as follows:

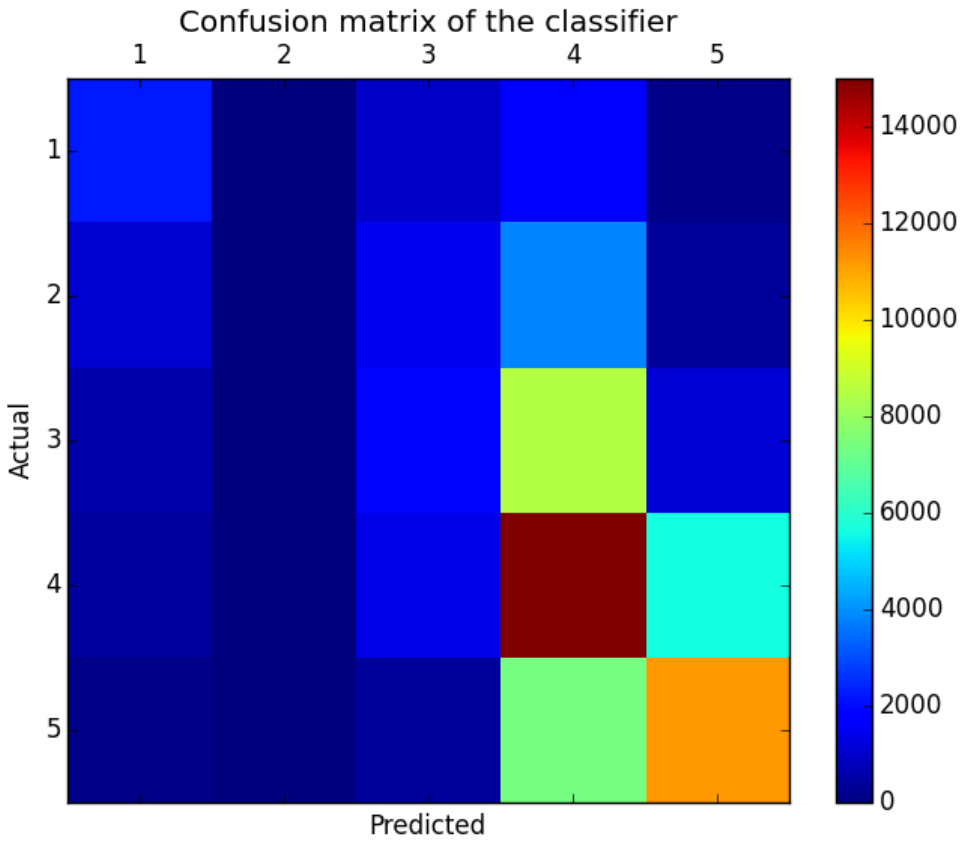
id	class	probability
0	4	0.414196043505
0	3	0.23779545812
0	5	0.185214534739
0	2	0.109089332634
0	1	0.0537046310016

[5 rows x 3 columns]

With a probability of 0.4 the above mentioned business will be classified in class 4. By evaluating the model against test set the classification accuracy of the model can be given as the fraction of test set with correct class predictions.

Classification Accuracy: 0.4615140189744133

The cross-tabulation of predicted and actual class labels can be shown using Confusion matrix.



Conclusions

The spatial data analysis performed on Yelp data indicates that businesses in California receive a larger number of reviews as compared to the other 15 states. It can be inferred using Sentiment Analysis that 1-2 star rated restaurants receive more number of complaints and 4-5 star rated businesses receive majority of cheerful reviews. Tokenizing the review data in trigrams helps to better analyse and extract the sense of reviews. Linear Regression model can fairly predict the ratings a user is likely to give to a particular business with a root mean square error of 0.97. The binary classification model can predict if a business is good or bad with an accuracy of 0.84. Multi-class classification is used to categorise businesses in 5 categories and it is more successful in classifying businesses of class 3,4,5 because the overall review data consists more of reviews for 3 - 5 star rated businesses.

References

Leman, A. Statistical Learning [PDF document]. Retrieved from Lecture Notes Online Web site: https://blackboard.stonybrook.edu/bbcswebdav/pid-3506663-dt-content-rid-22157305_1/courses/1158-CSE-590-SEC01-81617/08-StatisticalLearning-I.pdf

Data Machine Learning Graphlab Tutorial: https://dato.com

Multiclass Classification. Retrieved from: http://www.mit.edu/~9.520/spring09/Classes/multiclass.pdf

R ggmap documentation. Retrieved from http://stat405.had.co.nz/ggmap.pdf

R wordclouds. Retrieved from: http://www.r-bloggers.com/word-cloud-in-r/

RStudio Installation - R Documentation Installation link: https://support.rstudio.com/hc/en-us/articles/ 200552306-Getting-Started

The Yelp Academic Challenge Dataset : https://www.yelp.com/academic_dataset