Recommending Restaurants to investors by Predicting User Ratings and Overall trend of Business growth

-Harshal Gurjar

-hkgurjar

-200110946

Outline:

1)Problem Statement
2)Critical Thinking
3)Problem recognition
    a)Diving deep into problem statement
    b)Literature survey
4)Assumptions
5)Scope
6)Understanding datasets
7)Target question
8)Business Value
9)Project Setup
10)Data Analysis
11)Approach
12)Initial Observations
13)Model Validation
14)Results
15)Future Work
16)References

Abstract:

I am proposing a system for investors who want to invest in the restaurants. Investor will be provided with the data suggesting top 10 restaurants in US to invest in. We will be considering different factors for this recommendation system.

Objective :

- Based on Yelp Data Set of User Reviews for Restaurants, we will recommend top 10 restaurants for investment. Business value for this project are:
    - To provide recommendation on top Business in restaurant
    - To increase chances of profit for investor.
- This project involved parsing json data and converting it to csv file.
- Created csv files are then processed
- Building regression modell and prediction data as well as predicting the trend of business growth.
- Predict top 10 restaurants to invest.

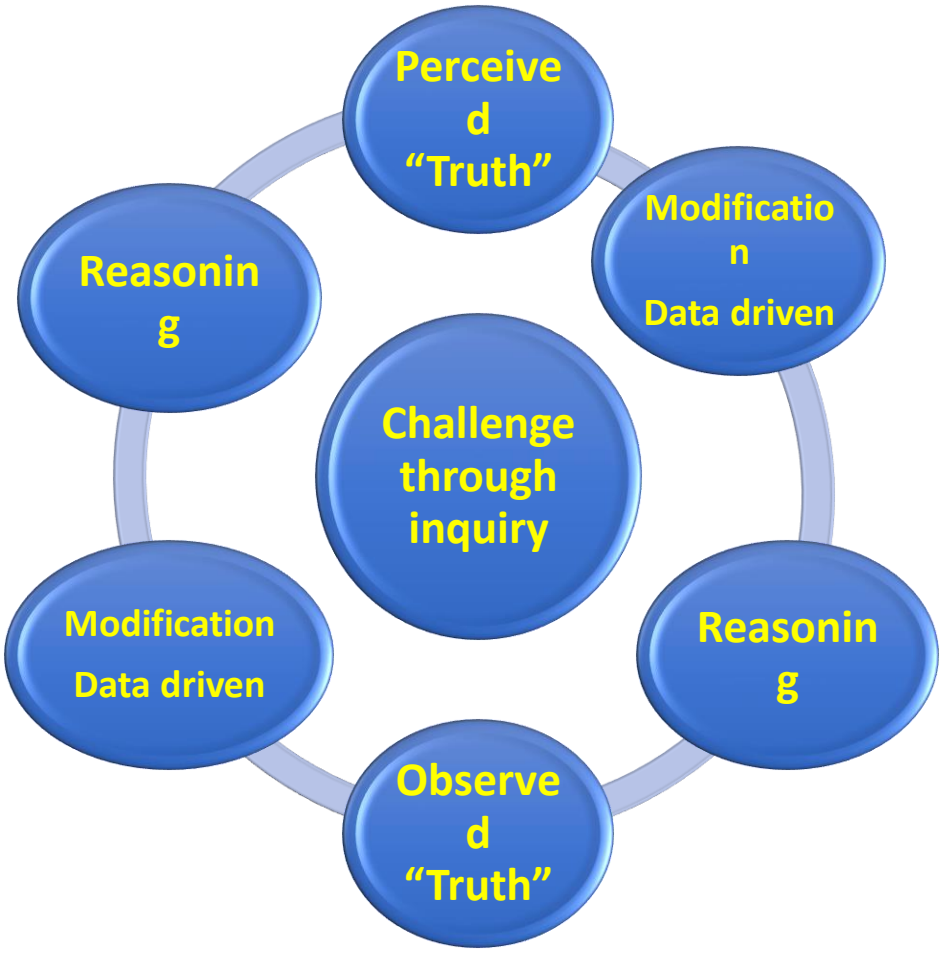
1) Problem Statement: **Recommending Restaurants to investors by Predicting User Ratings and Overall trend of Business growth.**

A) Description:

Investors are always curious to invest in the restaurants which are famous. Although the industry is competitive, it continues to grow. The restaurant industry is very romanticized. And it attracts people who aren't strategic thinkers, people who have an abstract relationship with money. I don't believe the restaurant industry is riskier than any other industry if you remove the romanticized players that aren't real business people.
However I believe that, rather that the revenue of the restaurant, the customer feedback and liking of the restaurant decide whether it has a good future or not. If number of customer liking the restaurant is increasing day by day then restaurant is sure to progress in the near future. If ratings given by customer is increasing day by day then restaurant is sure to progress in the near future. If number of customer feedback for the the restaurant is improving day by day then restaurant is sure to progress in the near future.

So I have decided to consider this factors of user reviews, prediction of user ratings for future and user rating forecasting for identifying the trend of business growth. Top 10 business with highest predicted values are recommended to the users.

2)Critical Thinking[1]:



Critical thinking is the intellectually disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information gathered from, or generated by, observation, experience, reflection, reasoning, or communication, as a guide to belief and action. In its exemplary form, it is based on universal intellectual values that transcend subject matter divisions: clarity, accuracy, precision, consistency, relevance, sound evidence, good reasons, depth, breadth, and fairness.
It entails the examination of those structures or elements of thought implicit in all reasoning: purpose, problem, or question-at-issue; assumptions; concepts; empirical grounding; reasoning leading to conclusions; implications and consequences; objections from alternative viewpoints; and frame of reference. Critical thinking — in being responsive to variable subject matter, issues, and purposes — is incorporated in a family of interwoven modes of thinking, among them: scientific thinking, mathematical thinking, historical thinking, anthropological thinking, economic thinking, moral thinking, and philosophical thinking.

- **Steps in critical thinking[1]:**
- Recognize problems, to find workable means for meeting those problems
- Gather pertinent (relevant) information
- Recognize unstated assumptions and values
- Comprehend and use language with accuracy, clarity, and discernment
- Interpret data, to appraise evidence and evaluate arguments
- Recognize the existence (or non-existence) of logical relationships between propositions
- Draw warranted conclusions and generalizations
- Put to test the conclusions and generalizations at which one arrives
- Reconstruct one's patterns of beliefs on the basis of wider experience
- Render accurate judgments about specific things and qualities in everyday life

3)Problem recognition:

How well is the Business?

How does the future of this business looks like?

a)Diving deep into problem statement:

Factors considered:

- Investor will like to invest in the restaurants which are popular among the customers
- Such restaurants are likely to succeed in the future
- So to identifying which restaurants to invest in, we will consider
  - which restaurants have good feedback from user
  - which restaurants have good ratings
  - which restaurants are likely to receive good feedback from user
  - which restaurants are likely to receive good ratings
  - Trend of business in future
- Why do I need to understand the customer's feedback?
  - Customer satisfaction is most important factor in any business.
  - Major reason for business downfall is dissatisfaction among customers
  - Happy customers = growing business

B) Literature survey:

i) Sentiment analysis:

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. A common use case for this technology is to discover how people feel about a particular topic.
A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral.
Sentiment analysis – otherwise known as opinion mining – is a much bandied about but often misunderstood term. In essence, it is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention.
Thus, Sentiment analysis (sometimes known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

B) Generalized regression model:

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').
In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression.
A regression models the past relationship between variables to predict their future behavior. ... When one independent variable is used in a regression, it is called a simple regression; when two or more independent variables are used, it is called a
multiple regression. Regression models can be either linear or nonlinear.

c)Research papers:

- http://leonidzhukov.net/hse/2011/seminar/papers/chi09-tie-gilbert.pdf
- http://www.utdallas.edu/~ryoung/phdseminar/DataMiningComparison-Melody.pdf
- https://hbr.org/2015/09/can-you-predict-a-startups-success-based-on-the-concept-alone
- http://leonidzhukov.net/hse/2011/seminar/papers/chi09-tie-gilbert.pdf

4)Assumptions:

- All data provided is trustworthy
- Project calculates overall business strength
- Data is consistent throughout
- Considering only top 100000 entries due to hardware constraints
- Data from top 100000 entries signifies the data throughout

5)Scope:

- This project scopes down to only top 10 listing of the restaurant
- This project scopes down to entire data of US
- Recommendation system only considers review data, user ratings, user feedback data for recommending investors
- Many more attributes can be added to improve model
- Data from top 100000 entries signifies the data throughout

6)Understanding datasets:

- We will be using only following dataset:
  Data sets
- yelp_academic_dataset_review.json
  - stars
- yelp_academic_dataset_user.json
  - User_id
  - Text
  - Review_count
  - Average_stars

Please note that I have converted each json file to its corresponding csv file and I am importing data from csv file for majority of time.

Data Transformation:

- **Convert Json data to csv file for easy access**
  - **yelp_academic_dataset_review.csv**
  - **yelp_academic_dataset_user.csv**
- **integratedUserAndReviewData.csv:** I have created this data set which is integration of user data and review data
  - **Review length:** Calculate length of each review

7)Sample questions: Which are the top 10 restaurants to invest in today? How well is the Business?
How does the future of this business looks like? Why do I need to understand the customer's feedback?

Target Question: Which are the top 10 restaurants to invest in today?

8)Business Value:

Investor will be provided with the data suggesting top 10 restaurants in US to invest in. We will be considering different factors for this recommendation system. Investors are always curious to invest in the restaurants which are famous.
So I have decided to consider this factors of user reviews, prediction of user ratings for future and user rating forecasting for identifying the trend of business growth. Top 10 business with highest predicted values are recommended to the users.

9) Step to setup project:

Presetup:

1) Download all the source code from the zip folder provided.

2) Please download and extract Yelp Dataset Challenge data, yelp_dataset_challenge_academic_dataset from the following link: Yelp Data Set

a) Extract the zip file download.

b) Unziped file will be again the zip file. Change the extension of unzipped file to .tar

c) Unzip this file aswell. You will see 5 json files in the unzipped folder

3) Name the data folder as: yelp_dataset_challenge_round9. This folder should directly have all your Jason files

4) Put this data folder in the same directory as that of the r code.

5) run file "InstallPackages.R". This will install all the required packages:

```
install.packages("gdata")

install.packages("ggplot2")

install.packages("streamR")

install.packages("jasonlite")

install.packages("readr")

install.packages("dplyr")

install.packages("caret")

install.packages("qdap")

install.packages("quantmod")
```

6) Use "convertJsonToCsv.py" to convert all json files to csv files.

Run: python convertJsonToCsv.py filepath/JsonFileName.json
for each json file in database.

7) Now we have csv file ready to use in our folder yelp_dataset_challenge_round9

8) Now run “preprocess.R”

This will do all preprocessing work on the files and data analysis of raw data. Initial graphs will be generated

9) Now run data integration : “DataIntegration.R”

This will carry out integration work

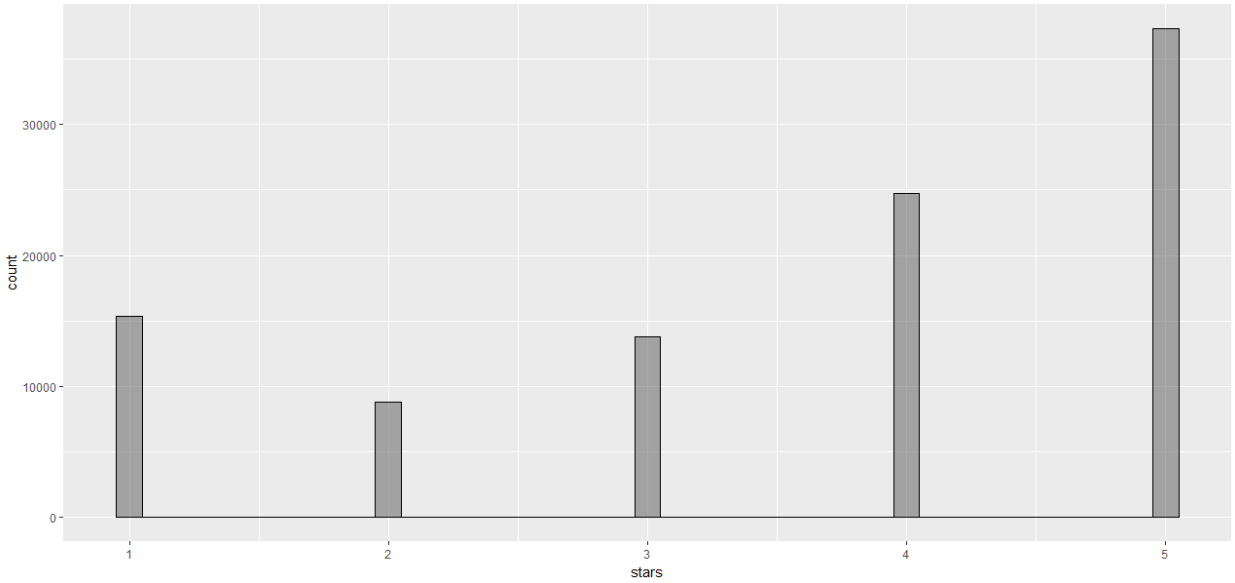
10) Now run “Analysis.R”

This will carry out all analysis wrk and output with top 10 restaurants will be displayed.

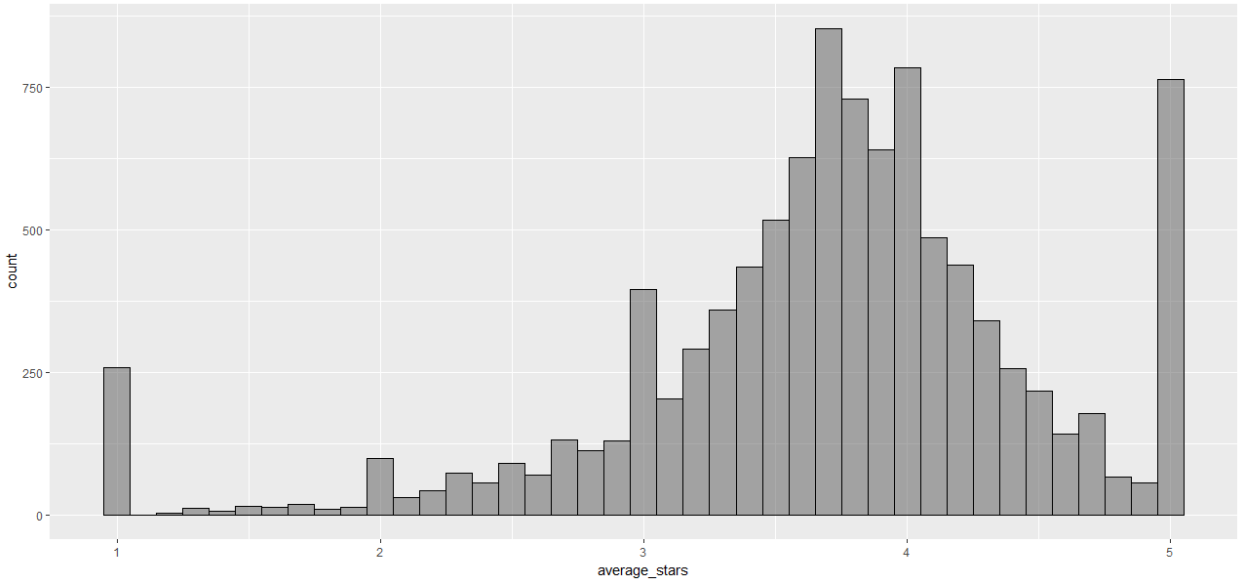
10)Data Analysis:

We will do analysis of data by plotting some graphs and getting summary of the data:

a) Understanding the distribution of ratings across the users

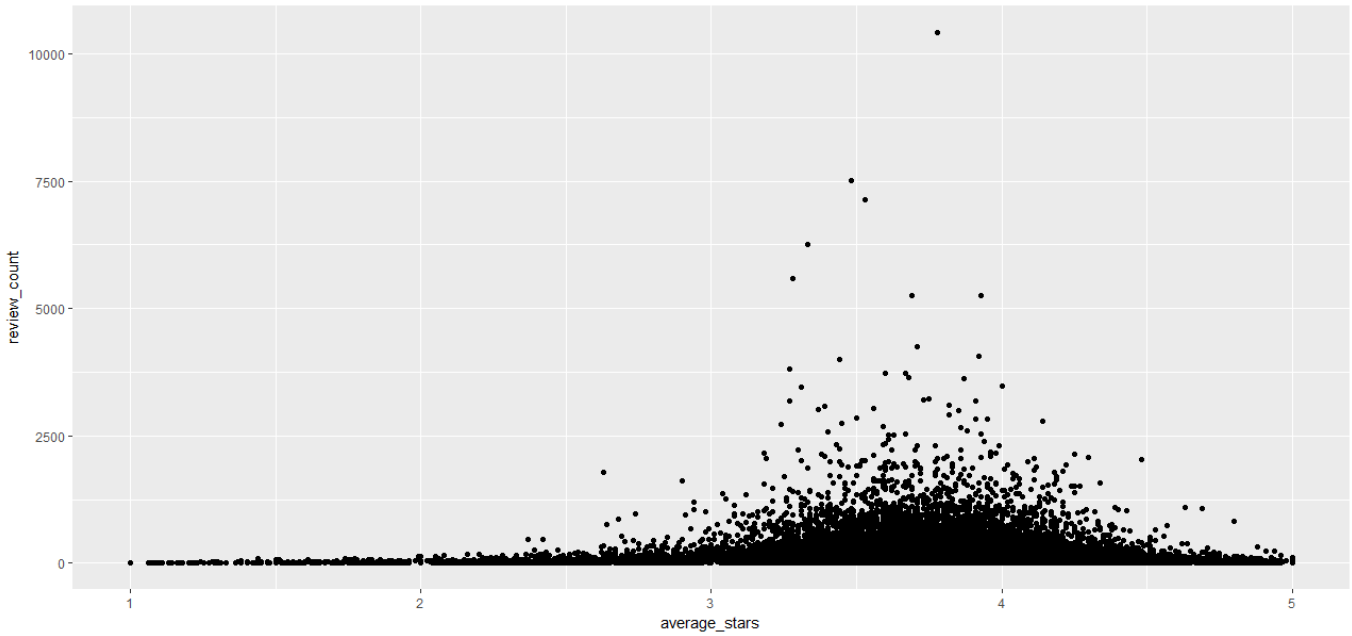


Histogram shows that there are substantial number of users with average rating of either 1 or 5 star (in fact, a large number of people have average rating of 5 star, more so than the others).



There are users who only write 1 or 5 star reviews. These users likely write reviews for the sole purpose of either strongly complaining or complimenting about their experiences (which is why there are not that many of them)

b) Understanding the relation between average stars and review count



Plot looks at the relationship between the number of reviews and average star ratings of the users in this dataset. The plot shows that users who give average rating of 1 or 5 star write only small number of reviews, and it gets increasingly more for users that are in between. Plot is also densely populated across all average stars, meaning there are also users who write only a few reviews that average anywhere between 1 and 5 star, so not exclusively for strongly negative or positive reviews

Summary :

lm(formula = stars ~ review_count, data = df_userReview)

Residuals:

Min	1Q	Median	3Q	Max
-2.6078	-0.6076	0.3964	1.3926	1.7562

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.608e+00	4.874e-03	740.251	< 2e-16	***
review_count	-5.109e-05	9.562e-06	-5.343	9.16e-08	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.443 on 99998 degrees of freedom

Multiple R-squared: 0.0002854, Adjusted R-squared: 0.0002754

F-statistic: 28.55 on 1 and 99998 DF, p-value: 9.161e-08

c) Now lets find the trends between stars ~ average_stars + review_count + ReviewLength

```
#summary of stars ~ average_stars + review_count + review length
Call:
Call:
lm(formula = stars ~ average_stars + review_count + ReviewLength,  data = df_userReview)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8814 -0.6843  0.1354  0.8405  3.7284

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.195e-02  2.007e-02   2.090   0.0366 *
average_stars  1.011e+00  5.078e-03 199.060   <2e-16 ***
review_count   4.218e-06  8.013e-06   0.526   0.5987
ReviewLength  -1.512e-03  3.607e-05 -41.906   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.205 on 99996 degrees of freedom
Multiple R-squared:  0.3028,    Adjusted R-squared:  0.3028
F-statistic: 1.447e+04 on 3 and 99996 DF,  p-value: < 2.2e-16
```

From the above data we can say that there is substantial relation between overall credibility of restaurant and average_stars + review_count + ReviewLength. Therefore we will use this model for our future predictions.

11)Approach

- **Exploratory Data Analysis – to understand the pattern in data**
- **Generalized Linear Regression**
- **Sentiment analysis to get length of useful words**
- **Model Building to predict overall trend of business**
  - average stars, word count in review and number of reviews will be used to build the model.
    - **Annova Test to test the result**
    - nested likelihood ratio test

12)Initial Observations:

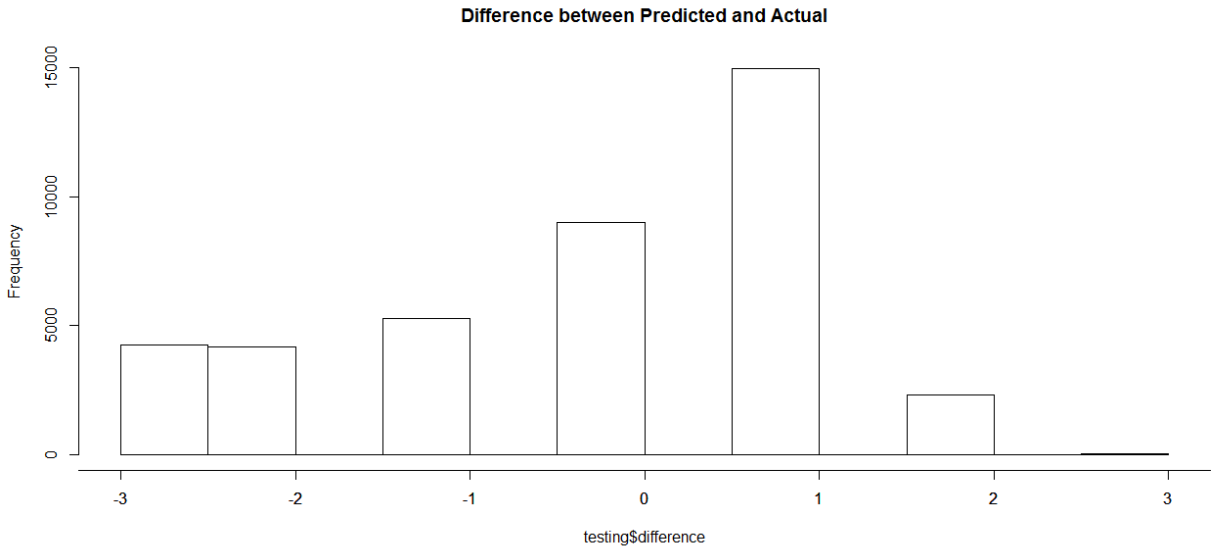
- Result shows that users average stars, review length or review counts are terrible predictors of users' ratings
- However when all this factors are considered together, it is good predictor of user ratings and overall business trend
- They will therefore be the basis for the model.

13)Model Validation:

- The given dataset is divided into training and testing dataset
- Training = 60%
- Testing = 40 %
- Model is trained using training set
- Model is fitted and tested for testing set
- Confusion matrix is used to determine the results

14)Results

```
[1] Oasis Springs Apartments
77734 Levels: 'do blow dry bar ... ZZYZX Plumbing
[1] Guess Outlet
77734 Levels: 'do blow dry bar ... ZZYZX Plumbing
[1] Red Talent Agency
77734 Levels: 'do blow dry bar ... ZZYZX Plumbing
[1] McDonald's
77734 Levels: 'do blow dry bar ... ZZYZX Plumbing
[1] Safe Home Security of the Carolinas
77734 Levels: 'do blow dry bar ... ZZYZX Plumbing
[1] Resortcom International
77734 Levels: 'do blow dry bar ... ZZYZX Plumbing
[1] Urology Associates
77734 Levels: 'do blow dry bar ... ZZYZX Plumbing
[1] Nice Cleaners
77734 Levels: 'do blow dry bar ... ZZYZX Plumbing
[1] Bristol West Insurance
77734 Levels: 'do blow dry bar ... ZZYZX Plumbing
[1] The Auto Clinic
77734 Levels: 'do blow dry bar ... ZZYZX Plumbing
> 
```



- Model achieves 62% of accuracy in predicting user ratings for future as well as 59% for predicting the trend in the business
- Considering the factors that customers ratings can also depend upon the external factors such as there emotional state, Financial state, Scenario before giving ratings, we can surely say that this prediction level is a good achievement

15)Future Work:

- We can identify growth of particular category of restaurant such as food, ambience, service etc
- For this we just need to tweak the algorithm to that granularity
- Extensive sentiment analysis and NLP for predicting the tone of reviewer
- Considering the revenue of restaurant for identifying customer satisfaction
- We can recommend restaurant for each location, for each particular category
- We can take feedback from investor and understand more about their factors in investment decision

16)References:

- https://en.wikipedia.org/wiki/Linear_regression
- https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&cad=rja&uact=8&ved=0ahUKEwiynIC8x8rTAhXBOyYKHe7VAM0QFgg7MAM&url=http%3A%2F%2Fcs229.stanford.edu%2Fproj2014%2FYun%2520Xu%2C%2520Xinhui%2520Wu%2C%2520Qinxia%2520Wang%2C%2520Sentiment%2520Analysis%2520of%2520Yelp%27s%2520Ratings%2520Based%2520on%2520Text%2520Reviews.pdf&usg=AFQjCNHCJ83fnobr9CpBPBtziGuFRDAVSw
- http://minimaxir.com/2014/06/reviewing-reviews/
- https://en.wikipedia.org/wiki/Generalized_linear_model
- https://www.slideshare.net/ohassta/critical-thinking
- http://leonidzhukov.net/hse/2011/seminar/papers/chi09-tie-gilbert.pdf
- http://www.utdallas.edu/~ryoung/phdseminar/DataMiningComparison-Melody.pdf
- http://www.criticalthinking.org/pages/defining-critical-thinking/766