Final Project : Yelp Dataset Challenge

INFO 7390 : Advances in Data Sciences/Architecture
Team 7

Chandni Sharma
Hari Panjwani
Sanath Shetty

Table of Contents

1. Problem Statement For Credit Card Dataset

1.1 **Project Title:** Yelp Dataset Challenge

1.2 **Domain:** Social Networking

1.3 **Description:** The sudden growth in the social media/networking domain has given rise to an enormous amount of user and business data which can be used by the business owner to improve their business based on customer needs and thereby increase their productivity.

1.4 **Problems to Address:** The project aims at providing a better experience to the user by recommending him things that he is most likely to be interested in. It would also help to predict the ratings for a given business based on numerous factors like user feedback, business amenities, location etc.

1.5 **Machine Learning Algorithms Used:** Cosine Similarity, Boosted Decision Tree Regression, Content Based Filtering (Custom), Collaborative Filtering (Custom), Text Analytics, Deep Learning, Decision Forest, Neural Network

1.6 **Tools:** Microsoft Azure Machine Learning Studio, Python, R Studio, Power BI. AWS, Spring Tool Suite

1.7 **Business Case:** Yelp could use our recommendation system to provide recommendations to users based on his interests/likes. Determining whether a given restaurant in the area provides good/bad service based on user reviews.. Using image classification, Yelp owners can classify pics based on what the user uploads

2. Data Wrangling & Cleansing

2.1 About the dataset: Owing to the sudden growth in the social media/networking there exists an enormous amount of user and business data which can be used by the business owner to improve their business based on customer needs and thereby increase their productivity. With the view to gain valuable insights from users from all sections of the society, Yelp. Inc. shares its data with the users and holds challenges to every year. The dataset provided by Yelp is in JSON format and it contains all the business details (like name, ratings, demographics, check-ins, review counts etc.), user data contains details about the reviews made by the user etc., and reviews details for each restaurant by all the users.

2.2 Data Cleaning: The initial step was reading the data from the source and creating a dataset with column names in a more readable and verbose form. Then we proceeded with the data wrangling methods taught to us. We first took the summary of the entire dataset which helped us in understanding the nature of the data. Then we started excluding all the redundant data from the dataset.

2.3 Outlier Analysis: By using the boxplot function we plotted the graph for the numeric values of latitude, longitude, check-in count, review count and star ratings in the dataset. We then remove these values from our dataset. Then we proceeded with analyzing each column in the dataset and we found that most of the data could be used as it is and there is not a lot of cleaning that is needed.

2.4 Snapshot of Outlier Analysis:

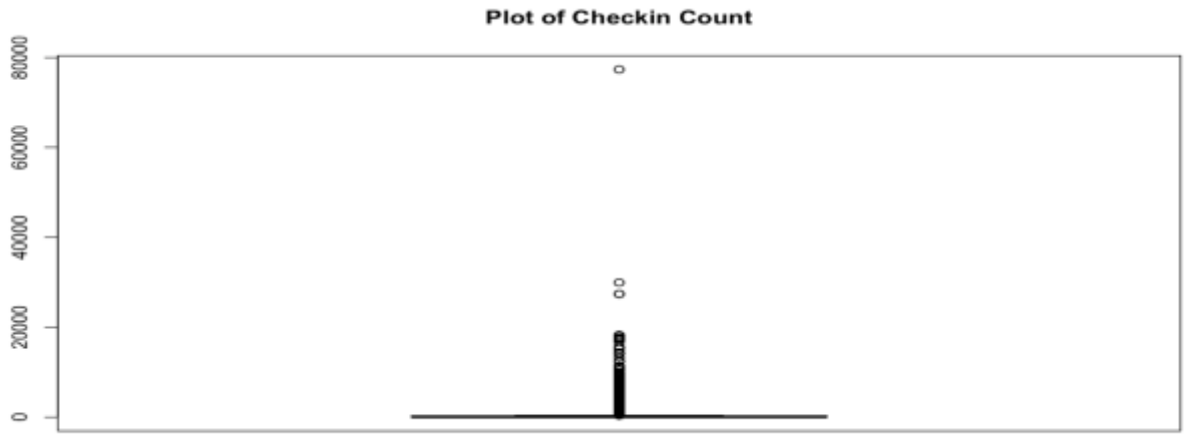


Fig : Box Plot for Check-in count

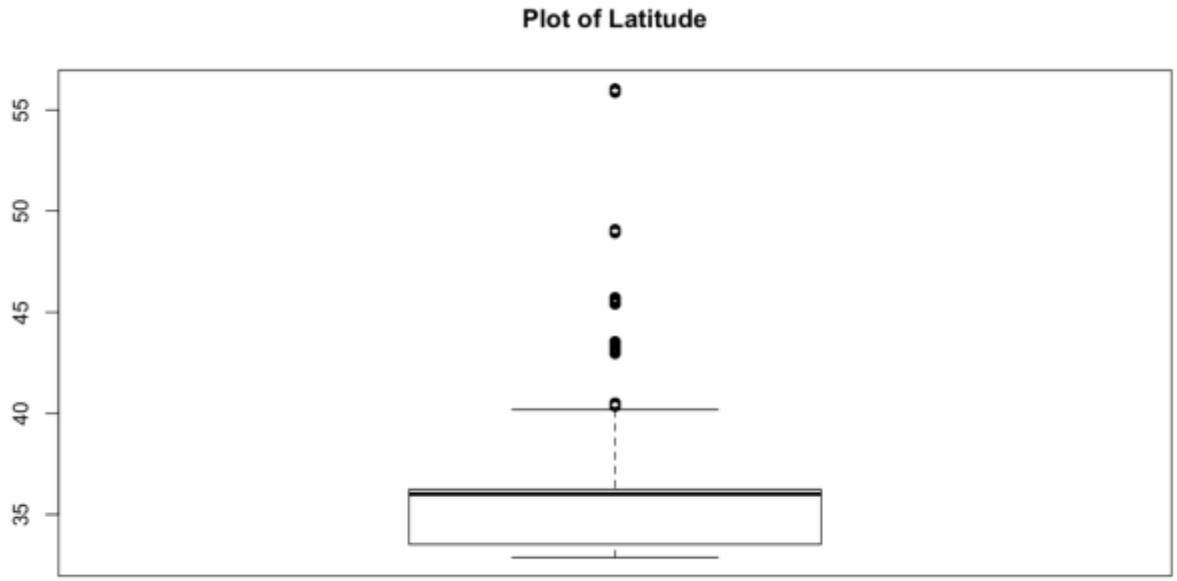


Fig: Box Plot for Latitude

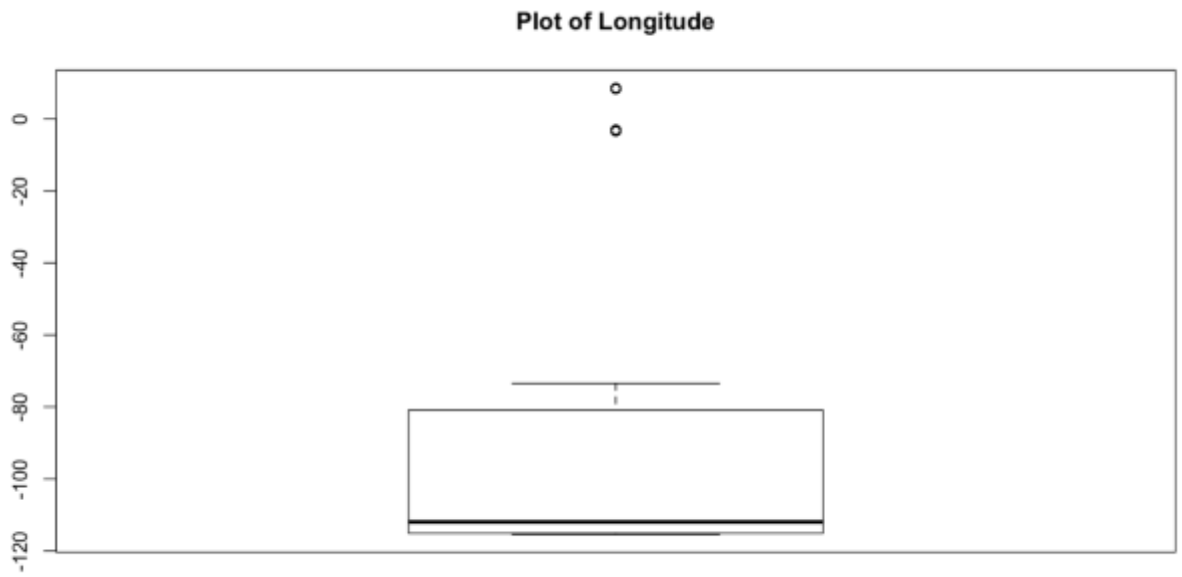


Fig: Box Plot for Longitude

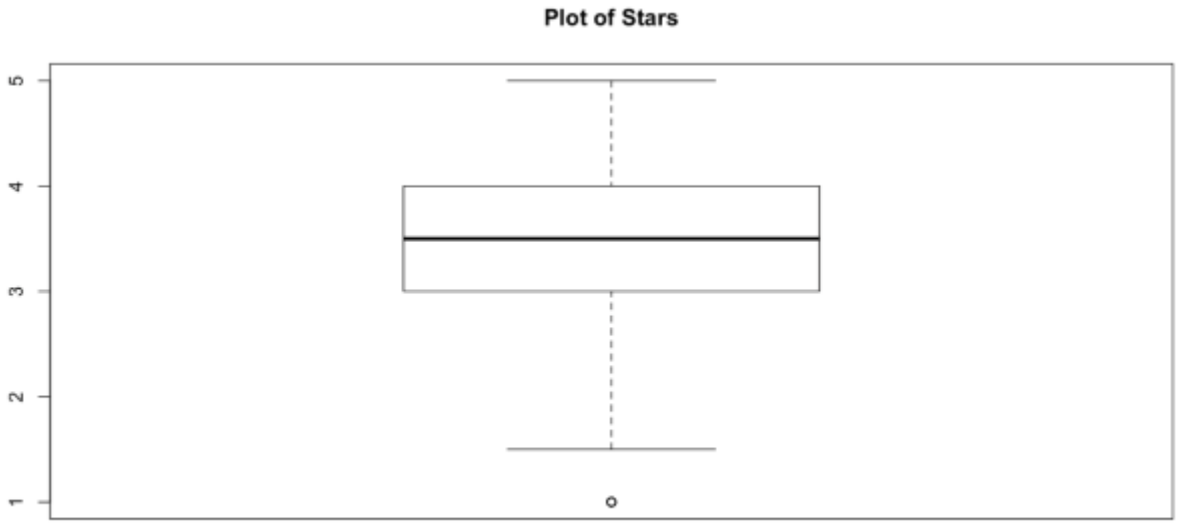


Fig : Box Plot for Stars

2.5 Feature Engineering: We have used feature engineering principle and have come up with important features to analyze the dataset.

i. **Category Map:** There were a total of 580 categories that were mentioned in the dataset. We normalized these categories and came up with frequencies of their occurrence of each category in the dataset. As you can see that after drawing out the frequencies for different categories there are some categories that don't have much impact on the star ratings and have very less frequency too. So we have decided to ignore these categories having low frequencies and have made a consolidated list of categories which have a frequency of 25 or more in the dataset and have used it for model generation in our model.

ii. **Attribute Set:** There were lot of attributes associated with the business. Based on the category map and important attributes affecting the ratings we have also created an attribute set associated with a particular business

iii. **Revie Classification:** We classified the reviews as cool, funny and useful

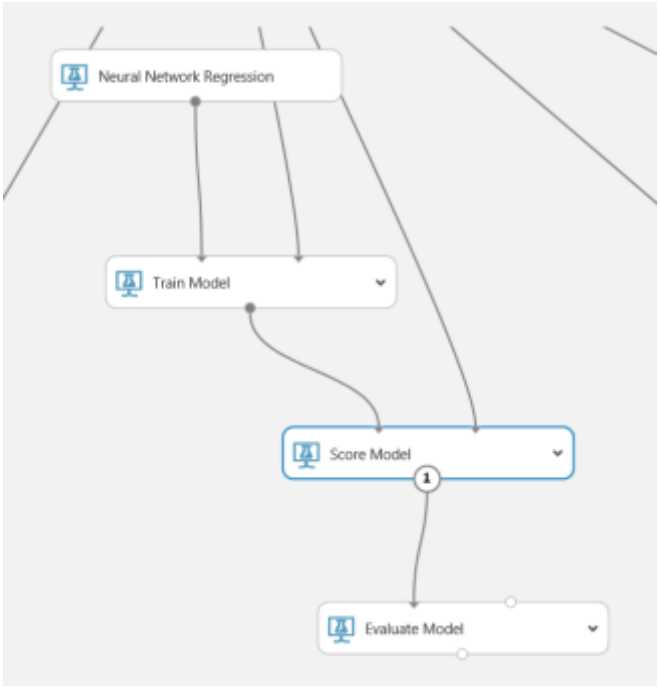
This was followed by some minor data cleaning activities like dropping tables and removing unwanted data from the dataset.

3. Model Selection

3.1 Neural Network

For this assignment we have used Microsoft Azure Learning Studio to clean the dataset and to run the machine learning algorithms on our dataset. Once done with the cleaning using the Execute R-script option in Azure, we partition and sample the dataset in the ratio 0:70 to 0:30 in order to train our data and test the prediction. We use the train dataset to train our multi-class decision jungle model and the test dataset to predict/classify the crime category in the test dataset based on the model created on the train dataset.

3.1.1 Snapshot of Multi-Class Decision Jungle (Azure)

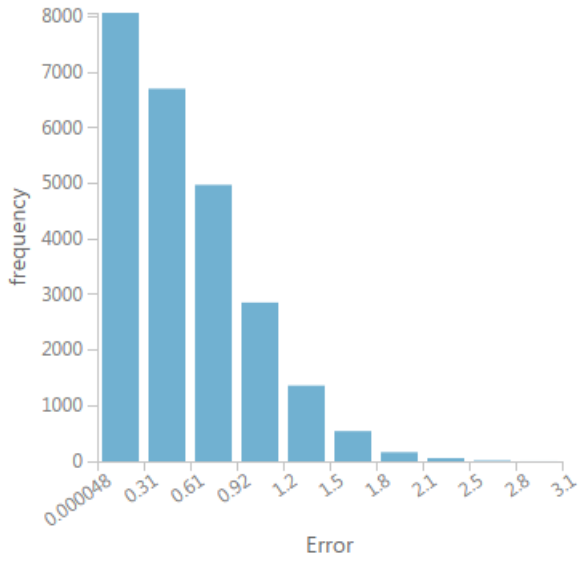


3.1.2 Performance Metrics

Metrics

Mean Absolute Error	0.5768
Root Mean Squared Error	0.719334
Relative Absolute Error	0.920941
Relative Squared Error	0.852594
Coefficient of Determination	0.147406

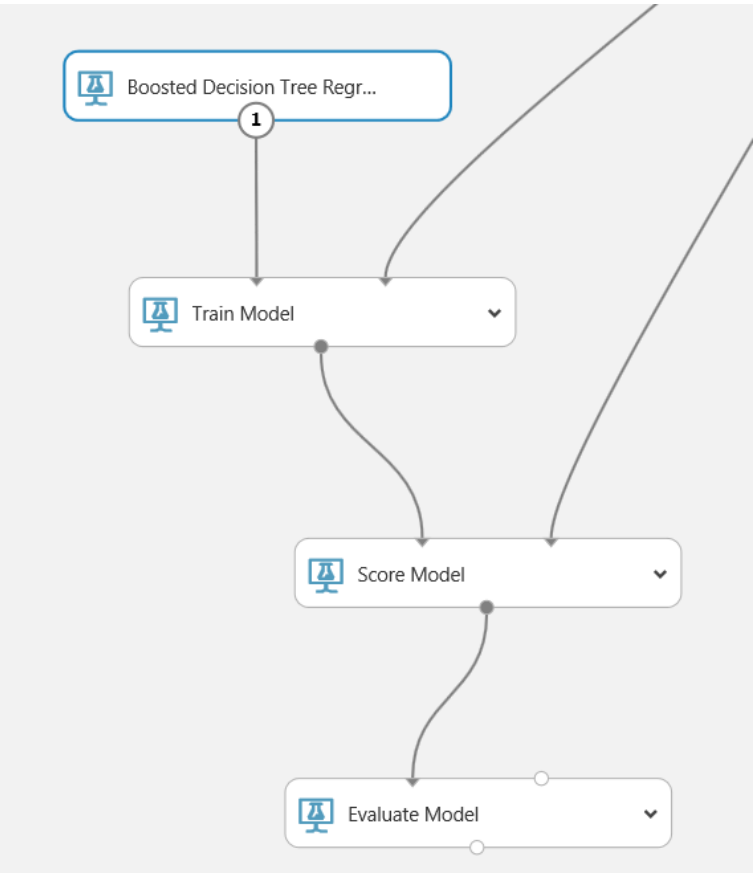
Error Histogram



3.2 Bayesian Linear Regression

For this assignment we have used Microsoft Azure Learning Studio to clean the dataset and to run the machine learning algorithms on our dataset. Once done with the cleaning using the Execute R-script option in Azure, we partition and sample the dataset in the ratio 0:70 to 0:30 in order to train our data and test the prediction. We use the train dataset to train our multi-class decision jungle model and the test dataset to predict/classify the crime category in the test dataset based on the model created on the train dataset.

3.2.1 Snapshot of Bayesian Linear Regression (Azure)



3.2.2 Performance Metrics:

rows	columns
1	6

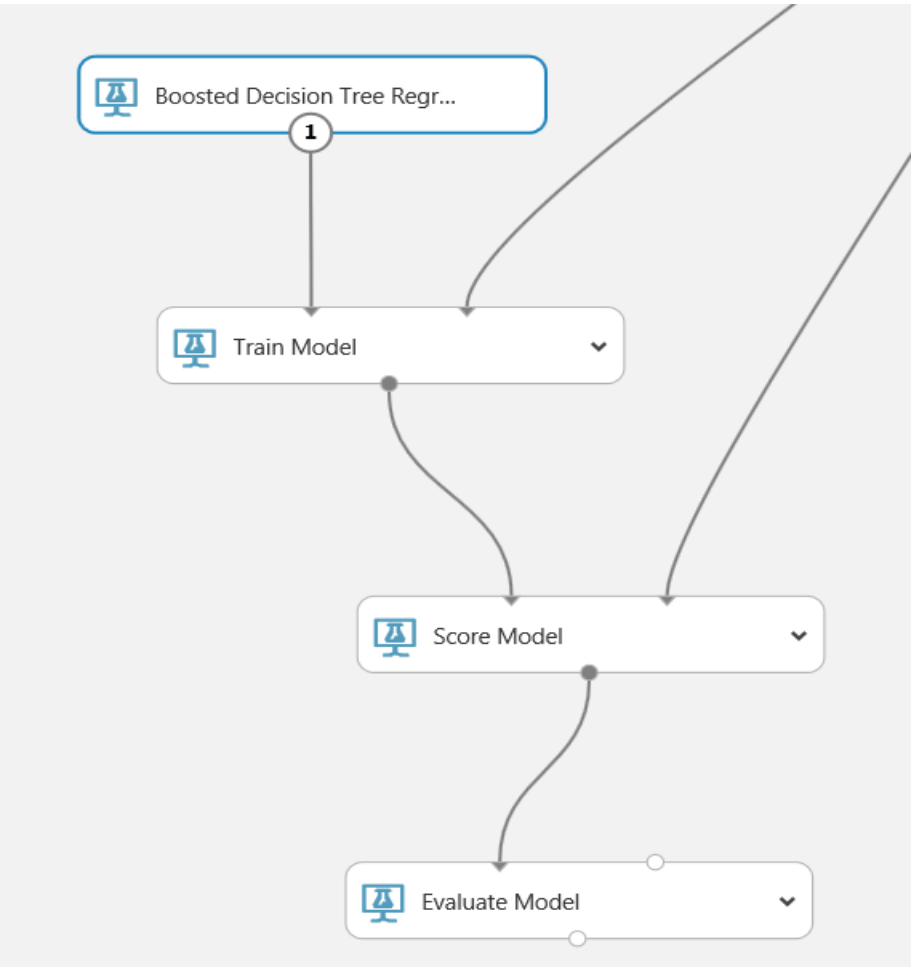
view as

Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
94830.282962	0.575407	0.726384	0.918717	0.869389	0.130611

3.3 Boosted Decision Tree Regression

For this assignment we have used Microsoft Azure Learning Studio to clean the dataset and to run the machine learning algorithms on our dataset. Once done with the cleaning using the Execute R-script option in Azure, we partition and sample the dataset in the ratio 0:70 to 0:30 in order to train our data and test the prediction. We use the train dataset to train our multi-class decision jungle model and the test dataset to predict/classify the crime category in the test dataset based on the model created on the train dataset.

3.3.1 Snapshot of Boosted Decision Tree Regression (Azure)

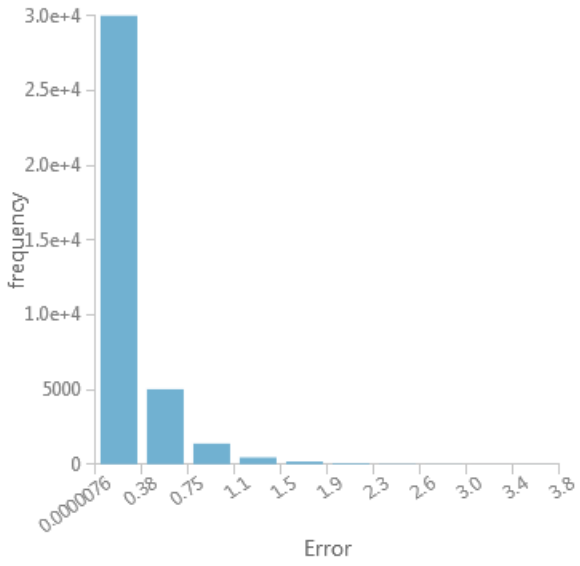


3.3.2 Performance Metrics

Metrics

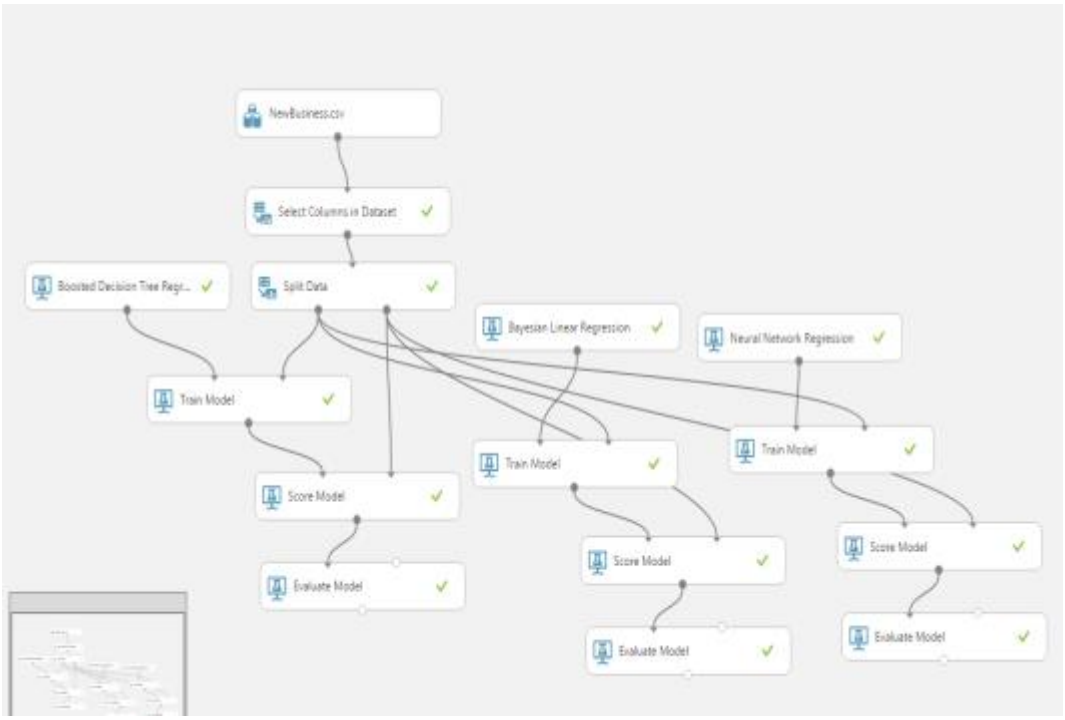
Mean Absolute Error	0.240013
Root Mean Squared Error	0.373283
Relative Absolute Error	0.383875
Relative Squared Error	0.23002
Coefficient of Determination	0.76998

Error Histogram



4. Selecting the Best Model

4.1 Snapshot of All Models:



4.2 Comparing Performance Metrics:

Metrics	B.Decision	Forest	Bayesian
Mean Absolute Error	0.24003	0.5768	0.575407
Root Mean Square Error	0.373283	0.7193	0.72638
Relative Absolute Error	0.393945	0.920941	0.918717
Relative Squared Error	0.23002	0.852594	0.8693
Coefficient of Determination	0.276586	0.147406	0.1306

Based on the above performance metrics, we have decided to select the Boosted Decision Tree Regression model as our best one as it has a good accuracy as compared to the others.

5. Model Building – Recommendation System

We tried using our dataset with the Azure Machine Learning module which helps us to recommend the output use its own matchbox recommendation. But the data available from Yelp is not in context with the recommendation system module provided by Azure ML. So we decided to build our own custom recommendation based on the Yelp dataset to help user predict restaurants based on his liking. We tried implement/create our own recommendation system using content based and collaborative filtering.

Algorithm for Collaborative Filtering:

1. First we sort and group the business/restaurants based on the star ratings that it has received by the user.
2. Then we group the business by means of user_id. This helps us to know which user has rated the business
3. Then we create a utility matrix. This matrix represents ratings for user for a particular business he has written a review for.
4. Then we develop a coo-matrix to remove duplicates and convert the available sparse matrix into dense matrix
5. Then we provide a user_id as an input to the model. The argmax() function recognizes the pattern/ inclination of the user for a particular restaurant.
6. Finally we applied cosine similarity tirhe users liking/ inclination towards the restaurants we have recognized in the above step

Algorithm for Content Based Filtering

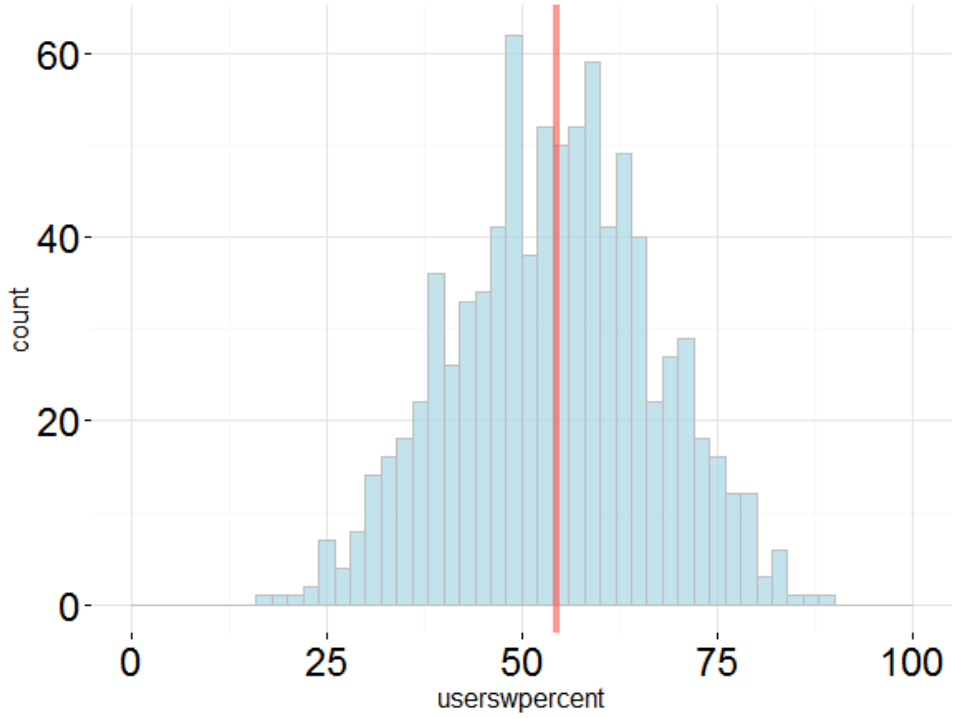
1. We start by reading the business data from the JSON with main focus on category, attributes and stars
2. Using python pandas we convert the category which is a list into a data frame. Similarly we convert the attributes which is in a dictionary into a data frame
3. We use a function which is FeatureUnion, where we do feature selection with factors like categories, attributes and ratings for that particular business. Finally the function normalizes and classifies the data using fit and transform.
4. For the next part we take the user_id as the input and analyze the reviews and star rating provided by the user
5. Finally we create matrix to represent a profile for the user with more importance for stars and features.
6. It finds out the similarities among the business for a user
7. Now based on the users liking/ inclination , we use the matrix created above to recommend restaurants for the user which has the highest value in the matrix

6. Analysis on the San Francisco Crime Dataset

6.1 Restaurants known for their food but with bad service

Observations: For each restaurant, the average rating for non-service reviews is plotted against the service-related reviews. Red points are the restaurants with significant differences between the two sets of reviews. The red points above the grey dotted line are GFBSs, those that are below the line are BFGSs. Overall at a false-discovery-rate of 5%, we have identified that are a list of 35 restaurants that serve good food but have bad service.

6.2 Service Cranks



Observations:

Histogram of the percent of reviews containing service words for each individual. The solid red line indicates the mean percent of reviews containing service words. We've identified restaurants that have strong differences between the ratings of their service-based reviews and the other reviews. We can also ask the same question about individuals - do certain individuals care more about service? Although there is not a significant overall difference between service-based review scores and other review scores when looking across all individuals together (p-value of 0.26), there are 12 reviewers who individually showed significant differences (at a false discovery rate of 5%). Interestingly, the majority of these individuals gave higher ratings if they mentioned service, the opposite trend than we

saw with businesses. It seems that there is a small subset of individuals who particularly notice great service and reward it!

6.3 Service & other factors that affect ratings

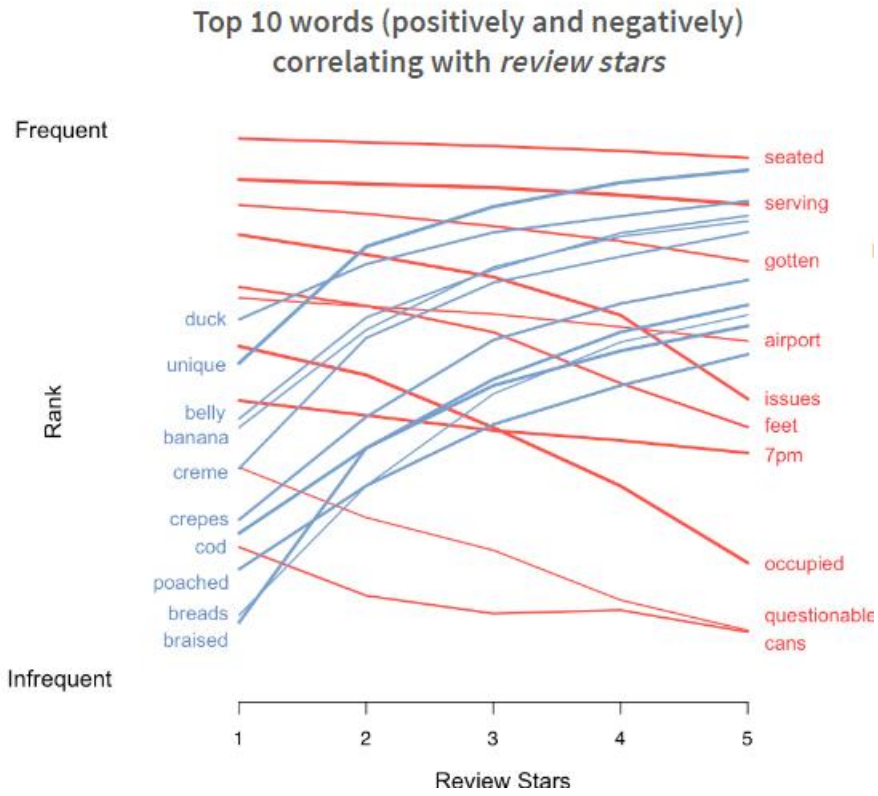


Fig: Rank of words Vs Review Stars

Observations:

As discussed above, we found more GFBSs than BFGSs. This could imply that reviews that mention service may generally have lower ratings than the reviews that don't mention service. We tested this directly by looking across all reviews and businesses jointly. Comparing the scores between service-based reviews and non-service based reviews across all businesses, we found that service-based reviews tended to have a significantly lower rating (p-value = 1.5 e-11). This makes sense because people, at least from our personal experience, enjoy griping about service and raving about food. But what

about all other types of words? Are certain words used more frequently when people are writing a good review?

We explored this question by testing how the presence or absence of a word in a review can correlate with the score. After looking at every single possible word, there were 1184 words that significantly correlated with the rating (at an FDR of 5%). The top sets of words were rather interesting:

7. References

- Kaggle: https://www.kaggle.com/c/sf-crime
- Microsoft Azure Machine Learning Studio: https://azure.microsoft.com/en-gb/
- Microsoft Power BI: https://powerbi.microsoft.com/en-us/