CMPT 732 –PROGRAMMING FOR BIG DATA 1

REVIEW USEFULNESS CLASSIFICATION AND ANALYSIS OF YELP DATASET

Table of Contents

1. INTRODUCTION

Yelp organizes a data challenge with its datasets, which contain information on Yelp's restaurants, users, ratings, and check-in(s) [1]. There is documentation also available for the data [2] which provides information on the structure of the dataset. From the dataset, some interesting analysis that can be performed on the data is to find the users who are most active on Yelp, find business ratings trend over the last few years and look for ways to improve them, find user demographics by region and to predict the usefulness of reviews. With 6.5 GB of data containing information on 1.5M+ users, 190K+ restaurants and 6M+ reviews we would face challenges in processing, analysing, and visualizing results.

2. METHODOLOYGY

❖ *Data Processing:*

The source datasets are provided in JSON format. The below Yelp datasets are relevant to the project:

- Users data (user.json): user-specific information such as user ID, first name, votes and compliments received.
- Businesses data (business.json): Business-specific information such as restaurant ID, restaurant name, address, state, latitude, longitude, review counts, rating.
- Review data (review.json): Information on the reviews, the users write for businesses.

We convert these JSON files to Parquet files using PySpark. The parallel processing abilities provided by Spark combined with quick processing of parquet files in the cluster would be excellent tools for Data Analysis and Machine learning.

❖ *Data Analysis:*

We get the top 10 categories by count and top 10 businesses by review count from the business data and focus our analysis on these areas. For these businesses, we identify the user demographics from the business and user data. Each user has attributes such as votes, fans, reviews given, compliments received. We take 3 such attributes – votes, review count, fans, rescaled between 0 and 1 (to give equal weightage to each attribute), and define a new metric - score which would be the summation of these attributes. This score is flexible and can be calculated with any of the user attributes to identify users who are most active on YELP.

Also, for each business we can find its rating trend by considering the average rating it receives every month and plot the ratings over the last few years. We check the reviews of businesses with a rating downtrend and create word clouds for low-rated reviews. We create bigram clouds as well as they encode more information than word clouds with single words. We examine the word clouds to get insights on areas the businesses can be improved.

❖ *Machine Learning:*

Processing Data: For finding the usefulness of a review, the data requires cleaning and feature extraction. The target category (categorical variable) is created based on the number of votes a review has received (***NOT USEFUL*** - 0 useful votes, ***USEFUL*** - more than 0 useful votes). Features such as *friends count, review text count, and number of months* (yelping period) are created as continuous values from *"friends", "text" and "yelping since"* fields respectively. The data is processed accordingly and stored as train and test data.

Classification and Predictions: We use *average stars, compliments, fans, review count, useful, friends count, review text count, number of months* features from user and review data to predict the target *useful category* [3]. For this, we use Spark MLlib library which provides common learning algorithms and utilities for classification, regression, and clustering. Models are built using algorithms such as Logistic Regression, Decision Trees, Naïve Bayes, Gradient Boost and the model with the highest accuracy score (AUC – Area Under the Curve for precision and recall) on validation data is selected. The model is then evaluated on the test data to get precision score, recall score and AUC. These determine how well the model generalizes and classifies the test data.

❖ *Visualization:*

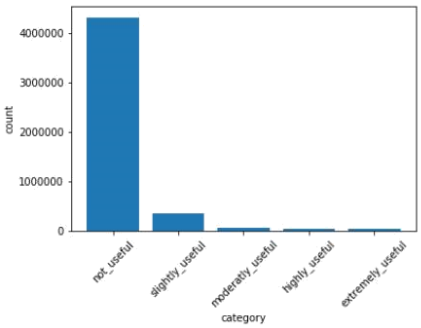
Data Analysis results are visualized using Tableau.

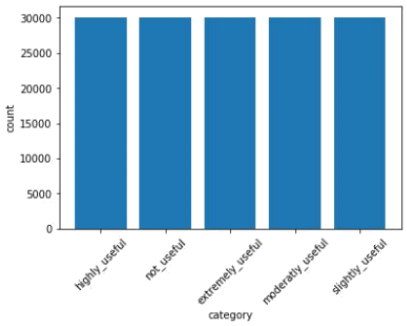
3. PROBLEMS

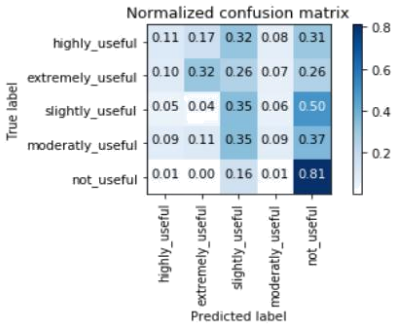
- ***Data Volume***: At first, we tried processing the data directly from JSON files but had to be repartitioned each time to support parallel processing. Later, the JSONs were converted and stored as partitioned Parquet files. This resulted in improved data processing times.

- ***Persisting in Memory***: The user and review data (6GB in size) was not cached which resulted in delayed outputs while finding most active users. The issue was resolved by properly persisting the required Data Frames to memory.

- ***Class Imbalance in Machine learning data***: Initially, we went with multiclass classification of useful votes instead of binary classification. Categories were "*not useful"* for reviews less than 4 useful votes, *"slightly useful"* for reviews between 4 and 7 useful votes, *"moderately useful"* for reviews between 8 and 10 useful votes, *"highly useful"* for useful votes between 11 and 16, and *"extremely useful"* for votes more than 16. However, records for one class (*not useful*) accounted for 99 percent of all training data (skewed data bar plot). This was balanced accordingly (balanced data bar plot) which resulted in a Test score (f1 score) of 0.80 on Naïve Bayes classifier model.





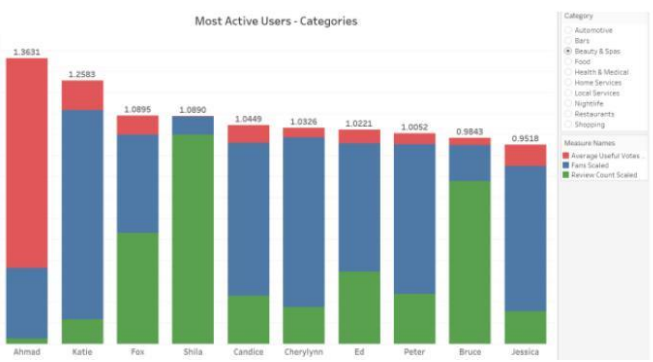


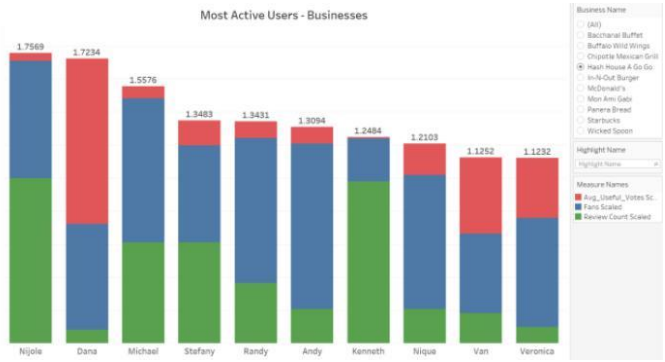
However, the model prediction results were not good for classes with a smaller number of records (from normalized confusion matrix). The problem was then done as a binary classification problem taking categories *"not useful"* - 0 useful votes and *"useful"* - more than 0 useful votes. Though the model had lesser test score (0.72) its predictions were not biased towards a particular class (refer 4.2 Machine Learning).

4. RESULTS

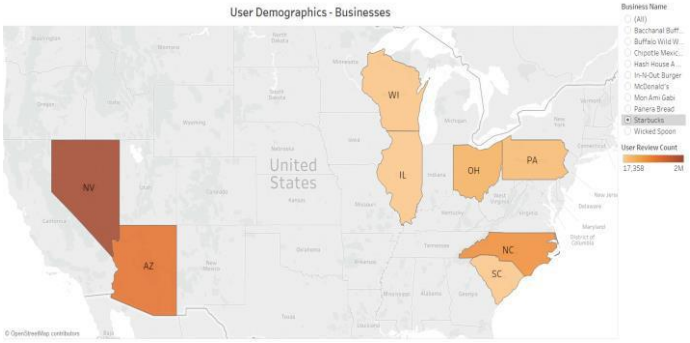
4.1 Data Analysis:

- The most active users were identified for Businesses and Categories. These users can be rewarded by YELP with credits.

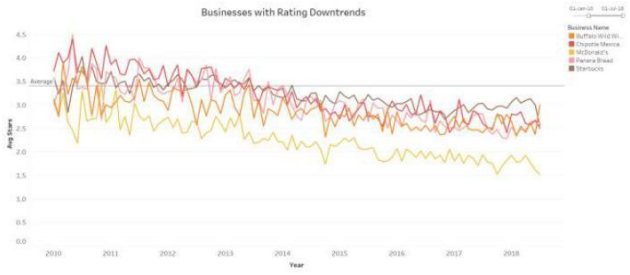




- User demographics are observed below which can be shown for each business.



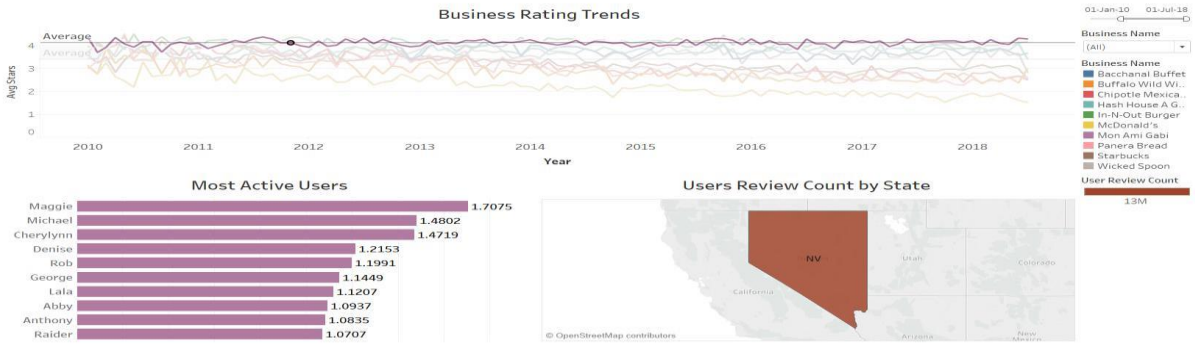
- Top 10 businesses monthly average rating trends are compared. We see that there is a downtrend for some businesses.



- Unigram and bigram clouds are created on reviews for businesses with rating downtrends.

For McDonalds, we observe there is a downward rating trend. From the bigrams we can say that "Customer Service" can be improved. Also some food orders such as "Ice Creams", "Big Mac", "Chicken Nuggets", "Happy Meals" might need improvements. Also, "Order Wrong" indicates that wrong orders are being taken. "15 minutes", "20 minutes" indicate that the service is slow. A similar inference can be made for Chipotle Mexican Grill and other businesses. We can say that improving these areas can help the businesses get higher ratings and better customer satisfaction.

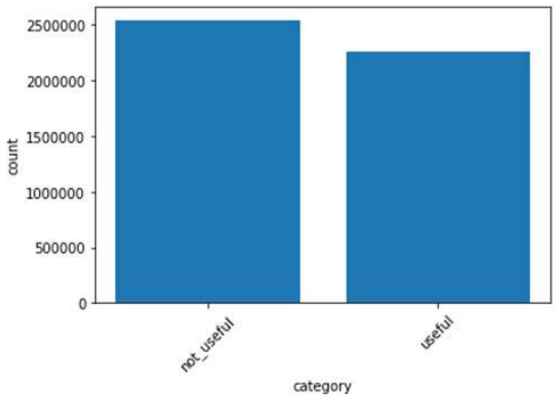
- Business Dashboard- changes dynamically on selecting a business.



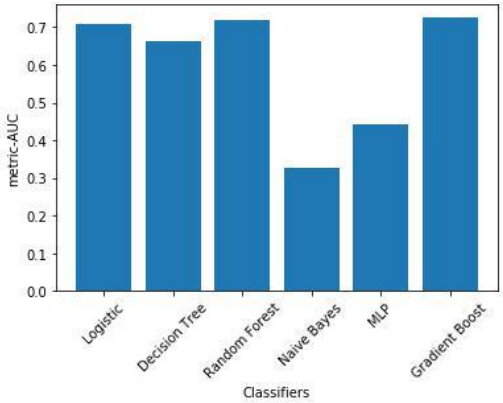
4.2 Machine Learning:

Predicting the useful category of a review-

Class count for training data



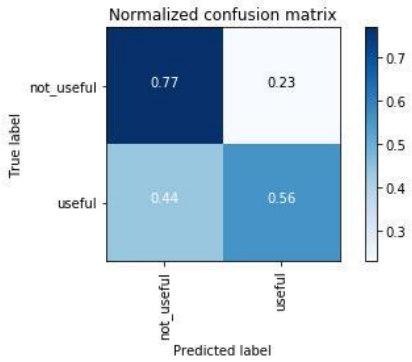
AUC Validation scores for Classifier models



Gradient Boosting Classifier model gives the best validation score.

Test score for model (Area Under Curve): 0.727331 Precision: 0.68 Recall: 0.56

Confusion matrix:



REFERENCES

1. Yelp Dataset [link]
2. Yelp Data Documentation [link]
3. Predicting the Usefulness of a Yelp Review [link]

PROJECT SUMMARY

Category	Points
Getting the data: downloading datasets from Yelp	1
ETL: PySpark jobs to clean the datasets, perform Extract-Transform-Load and Analysis	6
Problem: Finding best users, their locations and trends for business. Review usefulness prediction.	1
Algorithmic work: MLLib Classifiers to train models and predict usefulness of reviews	3
Bigness/parallelization: 6.5GB of data with 1.5M users, 6M reviews and 190K businesses	1
UI: User interface to the results, possibly including web or data exploration frontends.	0
Visualization: Visualization of analysis results in Tableau	5
New Technologies: Tableau, NLP for text processing	3
TOTAL	20