Project Report

Jie Min 301334127
Haoming Pan 301471834
Yewin Shu 301299522
Yukang Liu 301477326

CMPT 732 Big Data Lab 1, Fall 2021
Group Team Ace

December 11, 2021

1 Problem Definition

The Yelp dataset illustrates the general information regarding user evaluations on numerous businesses, and how users review every business based on their services and factors. In this project, we achieved a deep dive analysis on the dataset, to help businesses enhance future performance. We consider this issue from three perspectives:

1. The business: we analyze the business attributes to see what kinds of attributes can give the businesses better reviews and higher performance.
2. The users: we predict the stars of users' reviews which miss the star attributes by using NLP and analysis on streaming data, and we also recommend the potential users for some certain restaurants.
3. Reviews and tips on business comment on each business, and their potential effects.

However, there are several challenges we have to consider ,the quality of the business is hard to measure, and user reviews and tips can be biased and subjective. Also, the dataset can be unbalanced.

2 Methodology

In this section, a general overview of our project would be discussed. We first go through the architecture of the data-flow process in section 2.1. Then several small implementation steps would be further shown from section 2.2 to 2.6.

2.1 The general architecture

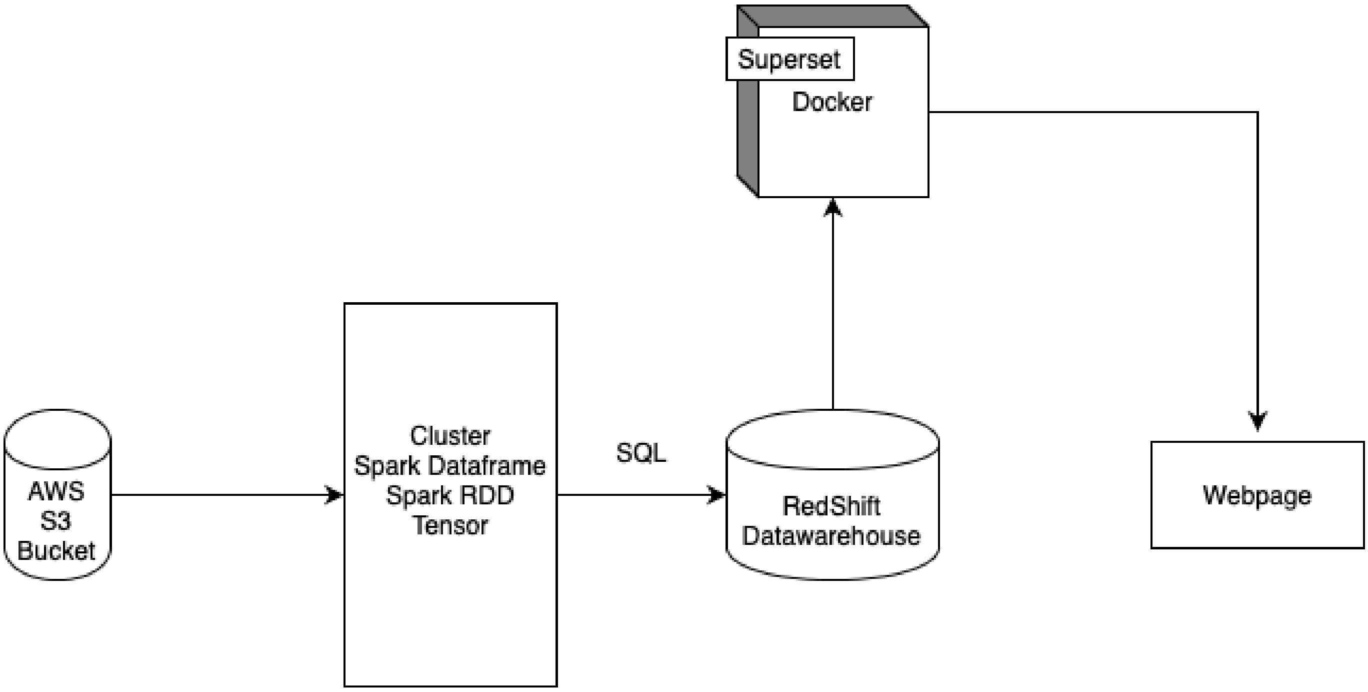

Figure 1: Project Architecture

2.2 Cleaning Data

At the beginning, we need to clean the dataset first to ensure the quality of analysis and model effect of the project. Through **PySpark**, we delete the duplicated data of all the Json files, check the legality of the data and remove the useless attributes efficiently, which makes the data more suitable for mining, analysis, and visualization.

2.3 Analysing Data

For this part, we used **PySpark** to load the cleaned data from **RedShift** Data Warehouse, analyzed things such as what is the distribution of stars in Vancouver, which business has the best good reviews ratio, and how business's attributes can affect business's stars, etc. Finally, we stored the resulting data into RedShift again for further predictions and visualizations. The reason that we chose Spark is that it has a really good API with Python. Most importantly, our dataset is very large, so parallel processing in Spark can help us to process the data faster.

2.4 Prediction and Recommendation

At this stage, we took business attributes to predict business stars. For categorical features, we used **indexer** and **one hot encoder** to do feature engineering. For prediction, we used **Random Forest Decision Tree Regressor** from the **pyspark.ml** package to do prediction.

Derived from the analysis session and star missing issue in the dataset, A deep learning algorithm is conducted to predict the review attitude label based on the text meaning. We introduced the **Pytorch** framework due to its high flexibility to design the Neural Network, and **LSTM architecture** to avoid the gradient vanishing problem. Several **Pyspark** scripts with **nltk and scikit-learn**

packages are also involved to construct the balanced dataset after performing the general **NLP pipeline**(sentence splitting, tokenization, etc.)

At the same time, we also created a function to perform sentiment analysis on streaming data by using **PySpark**. There are lots of data generated from thousands of data sources every second and need to be processed and analyzed as quickly as possible. Through the analysis and training of existing review text data, we can score new comments (missing or no ratings) entered in real time, so that we can help merchants better understand how users are thinking about their restaurants.

In the end, because we hope that we can help the restaurants to target potential users who have never been to the restaurants before, we proposed a recommendation system based on the **Latent Factor Model**, which is connecting user interests and items through latent factors.

2.5 Visualizing Data

Matplotlib, and **SuperSet** from **Dockers** are the two visualization tools we implement when it comes to graph generating. **SuperSet** can easily connect with Redshift data warehouse, and generate graphs according to the dataset from Redshift for analyzing existing data. However, when it comes to data prediction, we use Matplotlib to illustrate the future data by converting Pyspark dataframe to pandas dataframe generated from testing sets.

2.6 Frontend

After completing all the above tasks, it is necessary to establish a platform to display the result in both descriptive and visual ways. First, we develop a simple dashboard style web page with an introductory landing page using **Bootstrap, HTML,CSS, Javascript(Jquery)**. Upon completion, we upload the project repository, which is independent of our 732 Project folder, to **GitHub** for the deployment. Lastly, we deploy the Git repository via **Vercel**(A deployment platform) to deploy our project. For the tools selection, We use Bootstrap because it has a grid system that allows us to save time on website styling, and HTML,CSS,Js are the standard bundle for web development. Also, we use Github because it is necessary to conveniently deploy on Versel. The website is yelp-dataset-review

3 Problems

3.1 Connect Redshift with Docker superset

We can connect to Redshift with amancevice/superset in docker, but we can not upload datasets to superset even though we have Redshift drive already. We did a lot of research, but none of them worked. So we had to choose another container instead, which was cloned from the git repo and installed the new version of Redshift drive, then the problem was solved.

3.2 Prediction Challenge: Feature Engineering on stars prediction

We used business attributes to do predictions on stars. However, some of the attributes such as wifi, states, and noiselevel are categorical features, which can not be used in machine algorithms directly. Therefore, we need to do some feature engineering on the categorical features. For ordinal categorical features(Categorical Features which have some order associated with themselves), we applied an indexer on it to convert it into numerical features. For example, We changed the noise level feature from very loud, loud, average, quiet to 1,2,3,4. For nominal categorical features(Categorical

Features that are only labeled without any order of precedence), we applied an indexer and an one hot encoder to convert them into numerical features. For example, we changed states FL, MA,.., OR, TX to (1,0,...,0,0), (0, 1,...,0,0), (0,0,...,1,0), (0,0,. . . ,0,1).

3.3 Data transfer Challenge: Data type conversion, manipulation on SUPER type

When we import data from s3 to Redshift, the data type and value can be unpredictable. For example, there is one field with property "state", which denotes the states in the USA, however, it also contains provinces in Canada, which have completely different abbreviations. We use stl_loaderror_detail command to derive the errors log to discover all these error and adjust the table schema in Redshift accordingly.

3.4 Review Prediction: NLP Pipline

The text contains much more information compared with pure integer regression. Apparently, we need to customize a more complicated NN model. We construct 50 thousand reviews with labels ranging from 1 to 5 to represent the attitude towards the business, 10 thousand for each. Then we used the nltk package to do the general natural language processing pipeline since the package provides solutions to process common text issues such as stemming and lemma, stopwords, etc.

4 Results

There are 9 analysis problems we generate from the Yelp dataset to help us achieve our goals. The Visual demonstration(Graph, chart) will be displayed in Web front-end: as described in Task 5: Front-end

Problem	Result
The numbers of reviews in each state	For this problem, we count the number of reviews in each state. We found that the dataset is mainly from states Washington, Oregon, Colorado, Texas, Ohio, Georgia, Florida, and Massachusetts.
which business has the most absolute relative good/bad reviews.	We defined the 4-5 stars reviews as good reviews, and 1-2 stars reviews as bad reviews. The business 'Mike's Pastry' from Massachusett has the most number of good reviews. The business 'voodoo Doughnut - Old Town from Oregon has the most number of bad reviews; at the same time, it has the second most number of good reviews. Therefore, we decided to show the relative bad reviews ratio(#bad reviews/#total reviews). The business Suncoast Energy From Florida has the highest bad reviews ratio 99.64%. The high gas price at this gas station is the main reason for such poor reviews' stars. This inspired us to do one of our main analyze on how reviews text can affect reviews stars further.
How attributes affect businesses' stars	We would like further to see what factors can affect reviews stars. Therefore, we do an analysis on the correlation of business attributes and its stars. We can see the quiet businesses tend to have better reviews' stars. The businesses that have paid wifi tend to have lower reviews' stars. This gives us an idea that uses the attributes to do star prediction.
What categories are most popular	We addressed this issue from 3 perspectives, the general review, good review and good review rates. Some categories with shopping, Food, Beauty gain most popularity in general, but do not show a high good review rate.

Problem	Result
Use attributes to predict businesses' stars	For predicting stars, we use attributes such as states, price range, and noise level to predict stars. The final prediction is not good and not bad. We have a validation r2 score of 0.5. However, it is a meaningful result, because we found that some features can affect the prediction in the right way. One main problem in the Yelp dataset is that lots of businesses do not have attributes such as WiFi, noise level. This problem makes our prediction dataset small. If these kinds of attributes can be filled up. We can collect more data and add more other features to do prediction, which can make our prediction much better.
What is the ratio of the following user group on review and tips(writes both review and tips, only tips, and only reviews)	This information is calculated on the number of tips and reviews written on the same business. Most people have an equal number of tips and reviews. However, some people prefer to give more tips than reviews on the same business, which reflect the potential improvement of each business. On the other hand, some people tend to have more reviews than tips, which means that they either they have a attitude turning on a specific business or they just simply adore.
What was star distribution of the Top 100 Business in a specific region(Vancouver)	To derive the desired Top 100 Business data from the original dataset, we have to first define what factors of the business make them Top 100. In this case there are three factors that affect the eligibility. 1. The number of reviews need to be more than 50. 2. The Average stars need to be more than 3.5 / 5 3. The region must be Vancouver. Therefore, we have to filter these conditions and order them by ASCN order. Most Top 100 businesses have 4.5 / 3.5 stars, and after eliminating the un-qualified business, there are no 4 star businesses.
The differences of number of reviews and stars of restaurants from 2008 to 2020	Total restaurants from 2008 to 2020(every 4 years): 24206, 55069, 2020, 69752. Total high rate restaurants with more than 3 stars from 2008 to 2020: 11028, 27668, 1042, 42105. How many restaurants survived from 2008 to 2020: 13154. We counted how many high-scoring restaurants survived from 2008 to 2020, and compared the data every four years (the number of high-scoring restaurants, the number of reviews, etc.) to discuss usage of Yelp from 2008 to 2020.
Where are those restaurants rated with 5 stars located	We count the number of businesses with 5 stars in each state. We found that most of the business are located in the state of Texas, Oregon, Florida, and Massachusetts
Predict review attitude	There is a serious problem: Some labels of reviews are missing in the review dataset. To fix the problem, we select 50 thousand reviews with 10 thousand reviews for each label. And we build a recurrent neural network based text prediction pipeline to fill out the missing labels.
Review from streaming data	Usiing the previous review-star data to train the model through pyspark.streaming and pyspark.ml.feature, then we be scoring the unstarred reviews in real-time.
Recommend potential users for restaurants	We use LFM model to find out the potential users for the certain restaurants with the recommendation scores, like ($StWbrmWLsA3SvlgOwtyHYw$ 0.952757), which $StWbrmWLsA3SvlgOwtyHYw$ is the potential user's id and 0.952757 is the recommendation score.

5 Summary

We did a deep analysis on what can affect reviews and businesses performance. Through visualizing the results of our proposing questions, we validated that business performance correlates with their attributes, as well as customer reviews. Finally, based on what we have found, we designed and implemented corresponding solutions, including business stars predictions, reviews stars predictions, and the Yelp recommendation algorithm. As a result, the solutions optimize the future data prediction and simplify the workflow of the review system to help starters better initialize their business.

Getting the data: Acquiring/gathering/downloading	1
ETL: Extract-Transform-Load work and cleaning the data set.	2
Problem: Work on defining the problem itself and motivation for the analysis.	4
Algorithmic work: Work on the algorithms needed to work with the data, including integrating data mining and machine learning techniques.	3
Bigness/parallelization: Efficiency of the analysis on a cluster, and scalability to larger data sets.	1
UI: User interface to the results, possibly including web or data exploration frontends.	3
Visualization: Visualization of analysis results.	3
Technologies: New technologies learned as part of doing the project.	3