Applications of Machine Learning Models for Yelp Local Business Data

Ruchi Singh Kulwinder Kaur Ovanes Mikaelian

rsingh26 calstatela.edu kkaur7@calstatela.edu omikael@calstatela.edu

**Abstract:** The aim of the project is the application of all relevant Machine Learning models to Yelp Local Business Data in order to analyze, predict and recommend. For this purpose we have used seven Machine Learning (ML) algorithms in Azure ML and have implemented four of these algorithms in Databricks Spark ML.

The analyzed Yelp business dataset contained 70 business attributes for more than 350,000 business entries. Additionally, review tips and likes from 500,000 users have been processed for the project. Comparative conclusions have been made related to efficiency of Spark ML asnd Azure ML for this dataset.

1. **Introduction**

We chose the topic of Yelp Business due to its constantly growing popularity and its importance for businesses today. According to Yelp factsheet, as of 2017, the site has about 157 million. monthly visitors with 127 million. comments. Having an insight of the businesses, users and their actions can be very beneficial for businesses in terms of gaining competitive advantage and customer satisfaction.

An added advantage of using this dataset was a good understanding of the dataset and past experience of data analysis on this data. Previously, we all analyzed Yelp to find insightful connections between U.S. states, categories, reviews, seasonal ratings, etc. This project has been the continuation of our work in terms of applying our knowledge of Machine Learning algorithms.

1. Related Works

Most of the work related to yelp business data is focused on rating prediction. Some of them have performed a detail analysis of business of a location, unlike this paper which does not take in account the location of business. For instance, Gwo-Hshiung Tzeng[3] concentrates on the criteria for a good restaurant location in Taipei. Whereas Tsung-Yu Chou[4] evaluates the importance of infrastructure cost and environmental factors responsible for setting up a hotel business. Predicting Usefulness of Yelp Reviews by Xinyue et al. uses MATLAB to perform language processing techniques.

Some of the popular machine learning projecs that contributed in selecting the right algorithms for our project are as follows: Qu et al ([2010](https://www.researchgate.net/publication/303331726_Yelp_Dataset_Challenge_Review_Rating_Prediction#pf9)) extracted feature to perform sentiment analysis on Amazon.com reviews, using unigram model. Earlier Leung et al. ([2006](https://www.researchgate.net/publication/303331726_Yelp_Dataset_Challenge_Review_Rating_Prediction#pf9)) used recommender to recommend movies to viewers based on the reviews. Fan and Khademi ([2014](https://www.researchgate.net/publication/303331726_Yelp_Dataset_Challenge_Review_Rating_Prediction#pf9)) filtered and predicted the restaurant’s average rating on Yelp. We used recommender to recommend business to users and predict the number of likes on a review tip based on the popular words. We also realised that prediction of the popularity of business will add value to all the previous works done on this dataset.

1. **Technical Specification**

For this project, we have used Microsoft Azure Machine Learning Studio and Databricks community edition to implement Spark ML using Python and R programing languages.

|  |  |  |  |
| --- | --- | --- | --- |
| Azure | | Databricks (Spark 2.1) | |
| Memory | 10 GB | Memory | 6 GB |
| No. of nodes | 1 | No. of nodes | 1 Driver (0.88 cores, 1 DBU), 0 Worker​ |
| No. of modules/ experiment | 100 | File System | Databricks file system |

Table 1. Platform specification

1. **Data Description**

The Business dataset collected from Yelp are dated between 2005 and 2016, is rich in information about the local businesses in cities from 14 states in U.S and 4 cities that include: U.K: Edinburgh, Germany: Karlsruhe, Canada: Montreal and Waterloo. The yelp data set is of size 90MB in CSV file format with 334,335 rows and 108 columns. The review tip dataset is of size 70MB in JSON file format with 591,865 rows and 7 columns.

To be prepare the data for this project we converted the JSON file to CSV. There are 452 categories of business the business dataset that has been grouped into broder categories int column Category. The columns in consideration in the business dataset are business id, category, reiew count, stars and attribute coumns describing the business. In revew tip dataset we shall take in account the business id, user id, text and likes. The two datasets are joint by the business id.

1. **Background/Existing Works**

In our project we have implemented seven algorithms in AzureML and four in SparkML. Most of our models are based on previous existing works.

**5.1 Recommender**

Our Matchbox Recommender module is based on the lab work where we constructed and evaluated a recommender using a sample of user movie rating data. Movie Ratings and Movie Titles datasets were joined in AzureML. The four score recommenders recommend different metrics: a) item recommendation, b) rating prediction, c) similar items and d) similar users. After this step, each metric is evaluated by Evaluate Recommender module and the success of the model is determined.

Our SparkML recommender is based on Collabarative Filtering project, that uses movie titles and rating datasets. Similar to AzureML, the datasets were joined in Spark ML.We have used ALS (Alternating Least Squares) algorithm to build the recommender. Additianaly, we’ve defined parameters and used fit method to train the model. Then we test the model to see the recommended movie for each user. The evaluation is conducted to show us the RMSE.

**5.2 Classification**

The Two class Logistic Regression and Two class Boosted Decision Tree are used in the project based on the study of the prediction of flights delay. In the prediction of flight delay all the flights with delay time more than 15 minutes were classified as delayed and flights delayed les than 15 minute were classified as not delayed.The logistic regression is considerd to be the best in fast training and linear classification model where as boosted decision tree is known for its accuracy, fast training and large memory footprint, apt for big data. The implementaion of two class logistic regression in Spark ML using Train Validation Split and Binary Classification Evaluator was earlier tested on the flight dataset. The regularization para meters used to avaoid the imbalances in data are 0.01,0.5 and 2.0. The PramGridBuilder is used to generate all possible combinations of regularization parameter, max iterator and threshhold.

**5.3 Clustering**

Our K-Means Clustering model in AzureML is based on the lab work where we trained and evaluated a k-means clustering model for the Forest Fire dataset to cluster high priority/big forst fire and sperate them from the smaller fires. After the data was cleaned and normalized, two-cluster and three-cluster models were created and trained. A comparison was made to see which one was a better choice. In this case, the two-clusters worked better with clear seperation in the clusters and the three clusters didn’t have distinct separation between the clusters.

Our SparkML work was based on clustering model which clustered customers into 5 clusters. The customer dataset had several customer attributes varying from age to numbers of cars. The model used the Income as the basis of the clustering.

**5.4 Text Analysis**

The text analysis part of our projectaimed at finding the most frequently used word in the review tip and predict the likes based on the text. Each word in the text precessed to have a vector representation. For training a classifier the term-frequency (TF), is multiplied by the inverse document frequency, and the TF-IDF scores are used as feature values. N gram and uni-gram vector representaion of data are tried to find out which works best with our data. Using R script we created the word cloud of the most frequently used words in the text. In spark we used the text to predict the number of likes customer will give to the text using the multiclass logstic regression as the likes are from 1-10.

1. **Our Work**

Using the above mentioned related works as templates for our models, we have conducted a thorough analysis of the Yelp datasets, cleaned and prepared our data and build our models.

**6.1 Preprocessing**

Dataset for this module consisted of Yelp Local Business file and Tip file which were initially converted from Jason to csv format. The both csv files were joined in AzureML using business\_id as the common field. In the next step, we prepared and transformed the data by removing duplicate rows, cleaning missing values, selecting our target columns.

**6.2 Matchbox Recommender**

The goal of the recommender is to provide Yelp users with recommendations for business categories based on their previous business ratings, as well as the business ratings of other users. Moreover, the model has a feature to predict the future ratings by user for a category.

SQL transformation was conducted to select the average number of stars that each user has given to a category. After column were selected, the dataset was split into training and testing fractions by .75 to .25 ratio. After the split, the training fraction is connected to Train Matchbox Recommender module and test fraction to four Score Recommender modules. Each of the four score recommenders represent different metrics: a) item recommendation b) rating prediction, c) similar items and d) similar users.

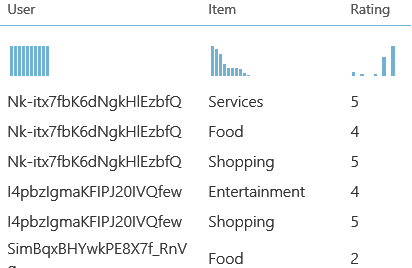


Figure 1. Rating Prediction

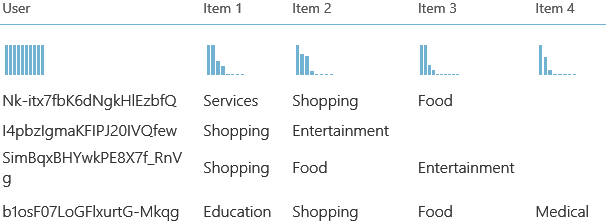


Figure 2. Item Recommendation

The Figure 1 and Figure 2 depict the visualizations of the model for Rating Prediction and Item Recommendation respectively.

After this step, each metric is evaluated by Evaluate Recommender module and the success of the model is determined. Figure 3 shows the evaluation result for the above-mentioned metrics.

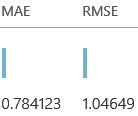
 

Figure 3 NDCG for Item Recommnedation and RMSE for Rating Prediction

The Item Recommendation is successful since the NDCG is close to 1. However, the Rating Prediction RMSE of 1.05

could be improved. Therefore, a different recommedation model was conducted in SparkML to see if the number could be improved.

**6.3 Collabarative Filtering Recommender**

Our SparkML recommender is based on Collabarative Filtering project, that uses a dataset called reviewstar, which we creted in AzureML. It represents the cleaned and transformed data that we processed and downloaded from AzureML and contains four columns: user\_id, category, review\_count, stars. User\_id and category are selected as features and stars is selected as label. The dataset is split to train and test fractions by .7 to .3 ratio. We have used ALS (Alternating Least Squares) algorithm to build the

recommender. Additianaly, we’ve defined parameters and used fit method to train the model. Then we test the model to see the recommended category for each user.

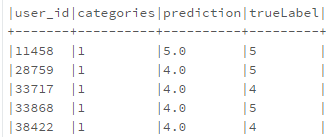




Figure 4. Rating Prediction and RMSE

The above figure depicts the output of the Prediction table and the RMSE of the model. Compared to the RMSE in AzureML (Figure 3.), the RMSE in SparkML (0.596) is much lower, thus making SparkML the better choice.

**6.4 Classification Models**

To predict the popularity of the business we defined the popular business to have stars greater than 3 and unpopular business to have stars less than 3. To select the feature columns and have the accurate prediction for the popularity of the business we chose the food category. All the attribute columns related to the food category like good for breakfast, lunch, dinner, take out, delivery, parking, alcohol, Wi-Fi, waiter service, wheelchair and noise level and considered as feature columns. We categorize all the columns for the classification model.

In Azure, we take a sample of the dataset (10%) and train it for Two Class Logistic regression and Two Class Boosted Decision Tree. The logistic regression is used to find the probability of the two states of the target variable. Whereas the boosted decision tree is an ensemble learning tree to make the prediction. The evaluation of both the models give an AUC of 0.72 and 0.73 respectively. This is a very good score with an accuracy of 0.8 and recall of 0.9 for both the models. Thus, both the models are to be suitable for the prediction of the popularity of the business.

Implementation of Two Class Logistic Regression in Spark using Binary Classification Evaluator on the complete dataset gives a value of 0.7 and AUR of 0.617. The AUR value of model has dropped in Spark ML due to training the logistic regression for the complete dataset as oppose to the sampled dataset. The dataset had a very small percentage of food business having less than 3 stars. Thus, the model does not train well to predict unpopular business. The result could improve if the dataset was balanced with popular and unpopular business.

**6.5 K-Means Clustering**

The clean food category data is used for 3 cluster and 5 cluster K-Means clustering model in Azure with the feature columns selected for classification model. Following are results of training the dataset.

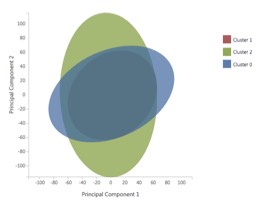
****

Figure 5. 3 cluster model

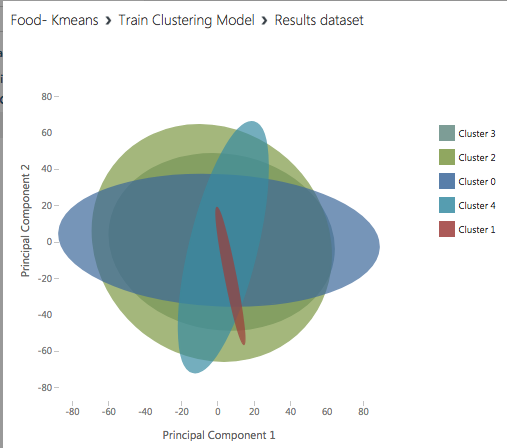
****

Figure 6. 5 cluster model

In SparkML we used Food data which we cleaned and transformed in AzureML. The table includes stars, review\_count, and categorized columns describing whether the restaurant is good for breakfast, dinner, lunch or take-out. The latter attributes describing the restaurant are chosen as features. The model clusters the restaurants based on the count of reviews each restaurant has received. For example, review\_count from 1 to 80 were put in one cluster, 81-200 in another, etc. By comparing the distance from center in SparkML and AzureML, we got a better result in SparkML (9.77 vs. 11.72).

**6.6 Text Analysis**

The text analysis is done by comparing two algorithms, uni-gram feature extraction and n-feature extraction in Azure ML studio and logistic regression model in Spark. The tip.csv dataset has six columns including text, likes, business\_id, user\_id, date and type. The operations are done on text and likes only because we only need the text and number of likes on that text. Data has been cleaned in AzureML studio. N-gram feature extraction includes the occurrence frequency of 2-gram with hashing bits 15 in the text instance. We have considered 1000 features. The uni-gram TF-IDF identifies the words that are frequent in document but rare in the corpus. R scripts were written to remove stopwords, urls, special characters, duplicates which is called preprocessing text. The word cloud has been created using r script which represents the most frequent words, relevant words with the polarity of negative and positive. The dataset is split in 70:30 ratio. We used Tune model Hyperparameters finding the optimum settings for a model. Out of these two models n-gram feature extraction gave good score of 0.769. The accuracy was 0.931 and precision was 1. The frequency of the relevant and useful words showed the customers sentiments and business satisfaction.

2^15 = 32,768 entries

In AzureML studio, we extracted the high frequency words, so it was necessary to predict likes to a piece of text written by users. In Spark, we used the classification model to predict likes. Spark has various libraries HashingTF, Tokenizer, StopWordsRemover, pipeline, etc. Some SQL queries were used to access the dataframe. We used pipeline algorithm with Tokenizer to split the text into individual words, StopWordsRemover to remove common words such as "a" or "the" that have little predictive value. A HashingTF class to generate numeric vectors from the text values.



Figure 7. Rating Prediction and RMSE

A Logistic Regression algorithm to train a binary classification model. So, the stopwords are removed and likes are predicted with respect to relevant text. The pipeline is used as an estimator and run with fit() method on training data to train the model. Classifiers are created with confusion matrices which gave true positives 2055.0, precision 1.0 and recall 1.0. The BinaryClassificationEvaluator class evaluator is used to measure the area under a ROC curve for the model which was 1.0 and is an ideal value. So the logistic regression model gave perfect score for predicting the likes.

1. **Conclusion**

Azure ML and Spark ML are powerful platforms for machine learning. Azure ML studio gave the flexibility to try various machine learning algorithms on the sampled dataset with simple drag and drop. Spark ML could train the model with large data volumes in relatively small time.

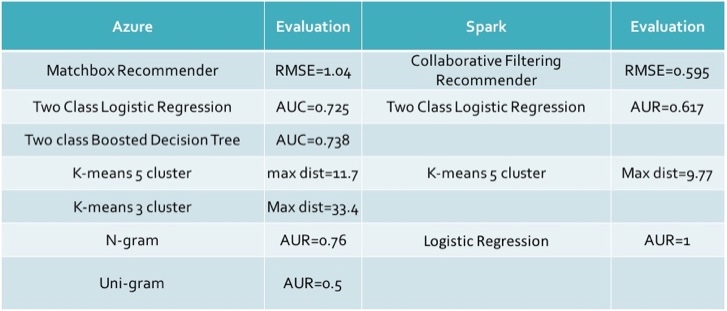


Table 2. Result Table

The Recommender built in this project can recommend business owners, their target customers to promote the business. It can help in finding customers with similar choices and profiles.

The popularity prediction using classification of business helps to understand what percentage of business in a category are popular. It can help investors invest in right business.

The business is intuitively grouped based on their range of review counts and their attributes. These businesses have similar popularity amongst customers therefore they have almost same number of review counts.

The text Analysis helps to understand customer sentiment and satisfaction of a business.

### References

[1] Data Set, “*https://www.yelp.com/dataset\_challenge/*

*dataset”*

[2] Github, “*https://github.com/rsingh26/DataScience/*

*tree/master/MachineLearning”*

[3]Chou, T. Y., Hsu, C. L., and Chen, M. C., 2008. “*A Fuzzy Multi-Criteria Decision Model for International Tourist Hotels Location Selection*” International Journal of Hospitality Management, Vol. 27, No. 2, pp. 293-301, 2008.

[4]Tzeng, G. H., Teng, M. H., Chen, J. J., and Opricovic, S., 2002. “*Multicriteria selection for a restaurant location in Taipei*” International Journal of Hospitality Management, Vol. 21, No. 2, pp. 171-187.

[5]Text analysis,”*https://gallery.cortanaintelligence.com*

*/Experiment/Text-Classification-Step-1-of-5- datapreparation-3”*