Big Data Tools

IESEG School of Management

A Group project by

Mui Han Ma, Inderpreet Rana & Vinay Rajagopalan

3rd April 2022

Table of Contents

1) Executive Summary

1.1) Problem Description

The COVID lockdown hit hard on a lot of restaurant businesses forcing some of them to start delivery and takeout for the first time. Yelp wants to help some of the businesses that still have not started delivery and takeout services by identifying corelated features with the businesses that already deliver.

1.2) Project Summary

To identify the factors/variables that lead some businesses to start doing delivery or takeout for the first time after the first lockdown and build a model to predict which businesses will start doing delivery/takeout after the first lockdown.

Build a predictive model to identify the businesses that will start the delivery and takeout services post COVID lockdown.

1.3) Algorithm's Used

Our team selected the following 3 models that we think will perform best for this use case based on the data distribution and the variable categories.

1) Logistic regression – Since it's a regression problem, we chose to test a simple and basic model that usually performs well and is easy to implement and explain.

2) GBT Classifier – We saw some features that could be important but had less weights based on the values of those variables, hence we also chose GBT classifier to take advantage of its Stochastic Gradient Boosting technique.

3) Random Forest – we chose random forest because it improves upon the decision tree algorithm by creating multiple trees and taking the majority voting and the data for the categorical variables can be easily split in if else logics, and regressed upon, so we also experimented with the decision tree which takes advantage of this type of categorical data.

We got the best performance from the logistic regression model, with the AUC of 82% and Accuracy score of 86.3% without an overfitting or underfitting.

The main reasons lead the business to start doing delivery or takeout service is to boost the business that has been impacted by the lockdown and increase customer base.

Possible business actions are suggested, they are **subscription-based plans** and **virtual services.**

For the first method, it is related to the group of factors of the time availability of the restaurants, whereas for the second method is more related to atmosphere of the restaurants

Section 2 shows the details for the insights and suggestions for business actions, while section 3 shows the technical used for building predictive models and its evaluation.

2) Insights & Recommended Actions

2.1) Interpretation of restaurant features:

There are mainly 4 groups of the final factors for the features from the final predictive model which are decisive for the restaurant to start delivery or takeout service:

1) Reviews, checking or stars from Yelp
   (Number of Users Reviews, Check in Count, Review Avg Stars, Review Cool Count, Review Count, Review Funny Count, Review Useful Count, Stars, Tip Compliments Count, Tip Count)
2) Atmosphere of the restaurants
   Ambiance upscale, Attire Casual, Attire Dressy, Attire Formal
3) Available timeslot and food provided
   (Lunch, Dinner, Dessert)
4) Location:
   (State category)

2.2) Recommendations

Before the pandemic, the customers enjoy eating in restaurants as they can enjoy the food and the atmosphere provided by the restaurants. After the lockdown, it is hard for the customers to maintain the same level of customer experience when they choose to order by delivery or takeout. With the top 20 features, it is reasonable that restaurants start to provide delivery or takeout service based on the reviews, check in or stars from Yelp. However, if Yelp can focus on the other two groups of factors to launch new services, it can help the restaurants start to provide delivery or takeout service. Therefore, there are two suggestions to Yelp related to these 2 groups of factors.

A) **Subscription Based Plans:**
   When the lockdown started, it is difficult for catering business to maintain the same level of business as there are no customers to dine in the restaurant, the revenue dropped a lot. As the predictive model showed that the available timeslot of the restaurant affects a restaurant to start doing delivery and takeout service, in which, open during daytime on weekday are more decisive, it is possible that during the lockdown period, people must work from home which led a great drop of the revenue for the restaurant as they cannot serve them anymore during the lunch time.

   Therefore, Yelp can launch the function of subscription plan. Once the restaurant chooses to activate this function, they can customize the number of plans, the length of the plan and dishes offered in the offered to the customers. Hence, the restaurant can promote their business, uniqueness, and variety and quality of food.

B) **Increase the varieties of Virtual Services:**
Businesses could provide more virtual services combined with the delivery or takeout service to be sustainable in the market, it is better a restaurant combine other services when they choose to start the delivery or takeout service. A few examples of these services could be:

a. Virtual food, beer, or wine tasting
b. Virtual cooking lessons by restaurant chefs (which can be combined with the first suggestion)
c. Virtual tour of the restaurant (including AR/VR experience when someone wants to experience restaurant dining)
d. Online celebration session for catering services for birthday or other celebration events
e. Videos of introducing the philosophy of the restaurant, dishes, and interior design of the restaurant

Yelp can help these businesses on their digital transformation journey by assisting them in offering and promoting these virtual services via Yelp's platform.

3) Technical Overview

3.1) Data pre-processing

The raw data was in json format with multiple dictionaries and lists contained in columns as well. Hence, we created the schema for each json file manually and loaded the data.

Overview of data cleaning and processing steps:

1. Our final base table had a total of 83 features including the new features we created.
2. The base table has the data for only restaurants and food service industry.
   - The end-to-end pre-processing and models can also be executed for all businesses by removing this filter in our code.
3. We dropped the columns where there were more than 70% nulls.
4. We filled the nulls values with 0 or mode based on the column description and types.

3.2) Feature Selection

For feature selection, we chose a combination of Logistic Regression feature importance and stepwise feature selection. We ran the logistic regression model with all features and selected the top 50 features.

The following is the extract of the feature importance:

	feature_index	feature_name	coef	mean	std	std_coef	feature_importance
0	63	review_count	0.005292	78.436837	185.596151	0.982136	0.982136
1	58	number_of_users_reviews	-0.006554	54.509053	133.173615	-0.872844	0.872844
2	10	Caters	1.841256	0.325494	0.468608	0.862827	0.862827
3	59	number_of_users_tip	0.018670	12.233715	34.693731	0.647743	0.647743
4	3	Attire_casual	1.446323	0.774194	0.418156	0.604789	0.604789
5	16	GoodForKids	1.197438	0.699688	0.458441	0.548955	0.548955
6	1	Alcohol	-0.872507	0.283039	0.450522	-0.393084	0.393084
7	49	is_open	0.816322	0.671384	0.469759	0.383475	0.383475
8	64	review_funny_count	-0.006183	20.112591	59.320511	-0.366789	0.366789
9	43	classy	-1.057576	0.097607	0.296813	-0.313902	0.313902
10	42	checkin_count	-0.000313	205.507180	824.303418	-0.258029	0.258029
11	29	Thursday_DT	-0.303317	1.564828	0.828190	-0.251204	0.251204
12	33	Weekday	1.139297	0.050572	0.219146	0.249672	0.249672
13	7	BusinessAcceptsCreditCards	0.489682	0.571696	0.494884	0.242336	0.242336
14	32	Wednesday_DT	0.271208	1.576067	0.833250	0.225984	0.225984
15	56	lunch	0.490406	0.256191	0.436574	0.214099	0.214099
16	14	Friday_DT	-0.240010	1.557752	0.826702	-0.198417	0.198417
17	66	stars	-0.239575	3.424766	0.816864	-0.195700	0.195700
18	41	casual	0.413596	0.336524	0.472570	0.195453	0.195453
19	34	Weekend	-0.986356	0.039542	0.194901	-0.192242	0.192242

For stepwise feature selection our approach was:

1. Fit the model with 2 features and monitor the AUC.
2. Add one more feature and monitor the change in AUC
3. If AUC improves and there is no sign of overfitting, we keep that feature
4. Add one more feature and repeat steps 1 to 4

We did multiple iterations of the stepwise and feature importance methods to select the best features for the optimal performance of the models.

While performing the feature selection, we noticed that **'Grub hub Enabled', 'Restaurant Delivers', 'Restaurant offers takeout'** were some of the features that were leading to overfitting, hence we did not include them in our final features.

The final features we used for our best performing model based on the combination of logistic regression coefficients and feature importance values plus stepwise feature selection are:

1) Ambiance upscale, 2) Lunch, 3) Dessert, 4) Dinner, 5) Attire Casual, 6) Attire Dressy, 7) Attire Formal, 8) Number of Users Reviews, 9) Check in Count, 10) Review Avg Stars, 11) Review Cool Count, 12) Review Count, 13) Review Funny Count, 14) Review Useful Count, 15) Stars, 16) State category, 17) Tip Compliments Count, 18) Tip Count

For the predictive models, logistic regression, random forest, and decision tree are chosen and evaluate the performance of the models by AUC and Accuracy scores. The parameters of each of our best models are as follows:

Logistic Regression: LogisticRegression(maxIter=1000)

Random Forest: RandomForestClassifier(numTrees=20, maxdepth=6)

GBT Classifier: GBTClassifier(numTrees=10, maxIter=10, maxbins=20, maxdepth=6)

After performing the cross-validation, the table below is the summary of the AUC and Accuracy scores on both train and test data set of each of the predictive models.

	Train AUC	Train Accuracy	Test AUC	Test Accuracy
Logistic Regression	80.7%	85.8%	82.3%	86.3%
Random Forest	85.8%	87.2%	81.1%	86.3%
GBT Classifier	87.7%	86.03%	83.2%	82.02%

According to the result above, the predictive model built by [logistic regression] have the best results from the AUC and Accuracy score without having overfitting issue.

4) References

https://databricks.com/blog/2015/01/21/random-forests-and-boosting-in-mllib.html

https://spark.apache.org/docs/latest/api/python/reference/

https://spark.apache.org/docs/latest/api/python/reference/pyspark.ml.html#vector-and-matrix

https://spark.apache.org/docs/latest/api/python/reference/pyspark.ml.html#recommendation

https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.evaluation.BinaryClassificationEvaluator.html#pyspark.ml.evaluation.BinaryClassificationEvaluator

https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.mllib.regression.LinearRegressionModel.html#pyspark.mllib.regression.LinearRegressionModel

https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.mllib.tree.RandomForest.html#pyspark.mllib.tree.RandomForest

https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.mllib.tree.DecisionTree.html#pyspark.mllib.tree.DecisionTree

https://docs.azuredatabricks.net/_static/notebooks/binary-classification.html

https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.classification.GBTClassifier.html