Big Data Tools 2
Group Project
IESEG MBD 2021-22

Yelp Business Analysis
Machine Learning and Data analytics using pyspark

Submitted by:
ALIPOUR MOTLAGH Mohammad Hadi
SRIPADA Saisumanth
NOOR Ayesha
April 2022

Executive summary

In this report, we will briefly explain our insights and analysis of the Yelp data set. First, we will discuss the business objectives, KPIs, and insights we get from the data, and then we will explain how we approached the data, and what we did to prepare the base table and Machin learning models we applied to the data and the results of our predictive models.

At this initial stage we would like to present the important highlights of what we have identified in the survey. Based on which please find our assumptions, our business insights, and data specification in the following paragraphs discussed in detail.

Assumptions: The data belong to before the start of the first covid pandemic then we assume that we do not know about what Yelp's actions in the upcoming months could be to deal with this situation. We are preparing this report to suggest an action plan for Yelp. Also, we do not want to pay complete attention to only food industry, as the company needs to decide the action plan to provide good services for all the customers and not just the food sector, hence we have not dropped any rows from the table, and we have two main reasons for this decision:

1. The data have wide diversity and we have all the features that we need to have a good prediction model, (however we know that the limit of our resources is the main hinder to having cross-validation inappropriate time but we manage to handle this and keep all data rows, we know that in real case the company will provide better calculation resources to analyse the data and then if the data is limited, the results of the analysis will not be efficient for making the decision.)
2. It is not acceptable to just provide insight for a specific part of the data when the business wants us to analyse the entire dataset.

Business Insights: However, we assume that we do not know about the actions Yelp will take during pandemic, but we are aware that (based on the data and articles on the internet) there were multiple actions in this respect like Covid Banner, Yelp Connect, Contactless payment, Call to Action, Yelp reservation. So, we do not want to discuss these actions, but we want to help Yelp with better approach.

Marketing SPT: (Segmentation, Positioning, and Targeting) Based on our findings the best variables that can help the business to have the meaningful segmentation of customers are **Restaurants Attire**, **HasTV**, business type **Restaurants**, **Number of check-ins in the morning**, **Good for Kids**, **Number of tips** that users give to the business. **Beauty Spa**, **extensions**, **Lunch**. There are other variables that we can also mention if the active hour of business is on **weekends or weekdays** and business categories like **Active life** and **Shopping, event Planning, Nightlife Art & entertainment**.

As we discuss in assumptions and you based on the results, it shows that starting the delivery is not just dependent on the state of belonging to the food and restaurant sector. however, the main proportion belongs to this sector but as a data analyst, we suggest Yelp to create their marketing segmentation based on the all-important factors and business types.

For the **positioning,** we can say that based on the situation of covid we suggest to Yelp to take a supportive position with reducing the cost of services and giving a bonus for businesses and some payment and credit allowance since the businesses are facing a challenging time and they must manage their costs. This can be the main approach and focal point in the Yelp marketing campaign. We will provide details of the **targeting** in the following section of reports about the segments in the market that are more important and what we can be suggested to each of the segment.

Data and Models: We have 128 features in our final base table out of which we decided to use 120 in our final models, we decided to not use binary features like takeout, and delivery in the modeling base on two main factors:

1. We do not want our model to overfit plus we want to prevent the data leakage.
2. As we discuss earlier these two variables just belong to restaurants and we want to explore all business types.

Business Insights:

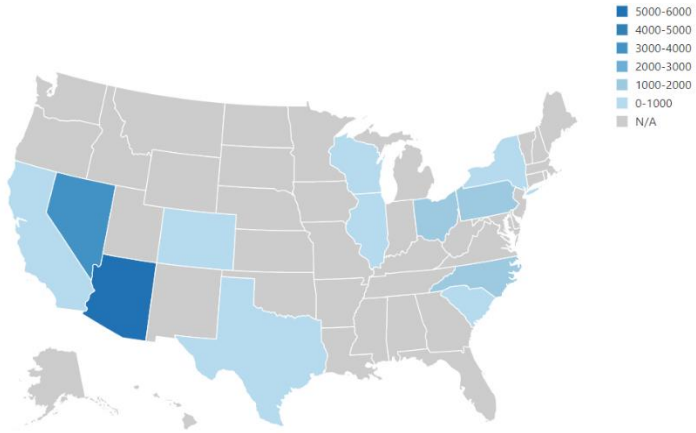
Our data is about nearly 20K businesses registered on the Yelp website. This data is diverse, and it includes a variety of businesses in the USA. We identified the list of categories and subcategories of Yelp and there are around 700 subcategories in the list that we use in the analysis and will discuss details about it in the Data and Model section.

Problem statement:

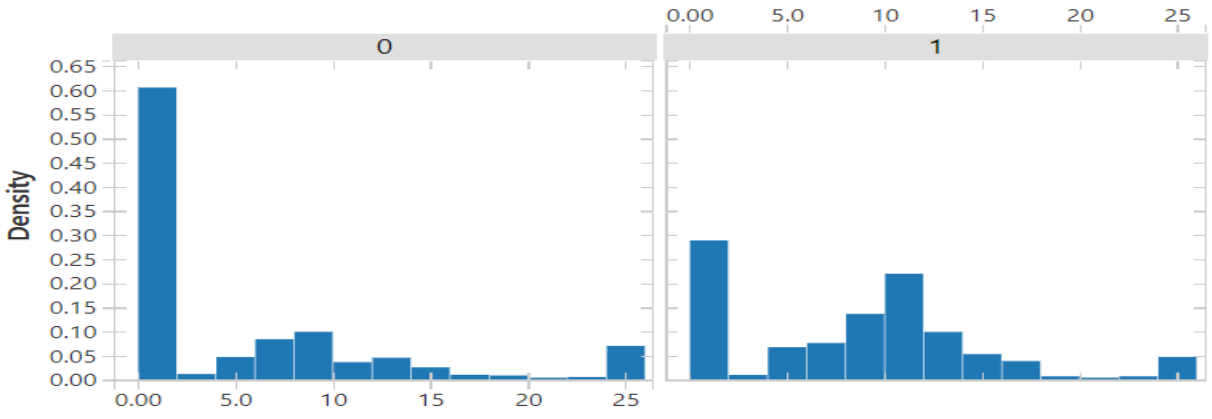
Through this project, we aimed to study the data provided by Yelp and help them identify patterns of behavior that are more likely relatable to the business capability to adopt innovative and digital transformations. To build a prediction model for their Data Science team to identify and predict what factors lead some businesses to start doing delivery or takeout for the first time after the first lockdown.

Further, for their Business Development team think and produce creative ways of how Yelp could target their communications to these selected set of businesses. All this can help them use our findings as a foundation to build their advertising solutions strategy. We found the details which have the highest impact on the restaurant business and here we want to discuss more of these findings and say how these variables will help Yelp to target their customers.

We briefly will discuss these features and at the end, we will provide some actions plan for the steps that could be taken to target the customers and provide them with needful services.

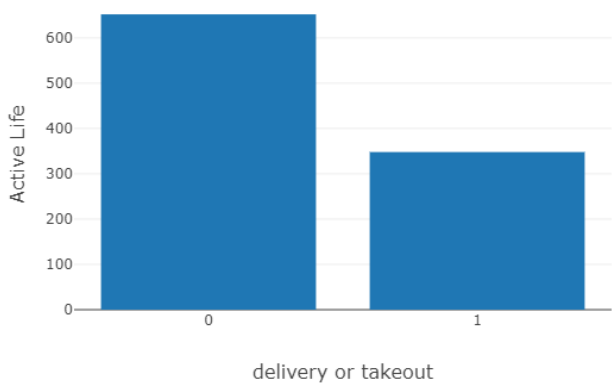


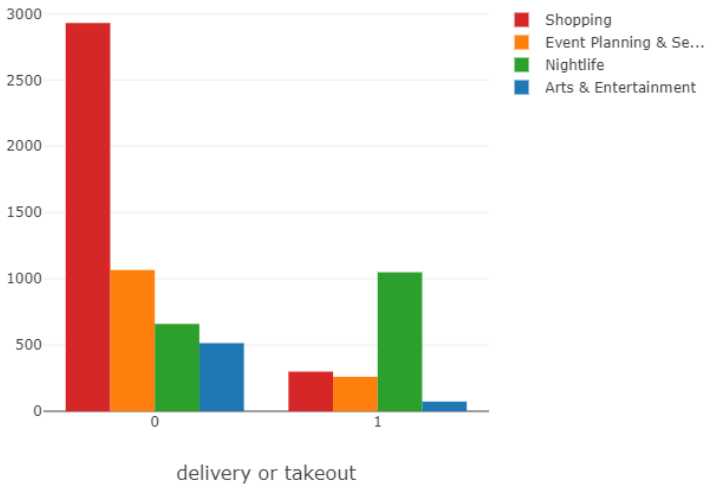
Distribution of Restaurants in the US: The first and the most important feature is if the type of business is food & restaurant or not. As we can see most of the restaurant and food categories on the west coast of the US and Arizona and Nevada are the first and second states with resp. 5466 and 3698 number of businesses. The reason for choosing a map for this feature is that we can provide better actions based on the culture and other conditions of the geographical distribution of our first target category that we will provide in the following page.



If the Businesses work on weekends or not: The second Feature is related to the activity days of the business on weekends, as you can see there is a significant difference between the activity hours of the businesses that have or have not delivered on weekends. The businesses which are more active on weekends tend to have delivery.

Active Life: as we discuss earlier Active life category, which is not related to food or restaurants is also has a high proportion of delivery and you can see in the right chart nearly 33% of businesses in this category started delivery after covid and our model in both decision tree and random forest detect this category as an important feature.





Other Nonfood Categories: Here are the other categories that started delivery, and we have provided this final colorful chart, to highlight that it is another support to our approach, to have all businesses and try to find all categories which are important in contributing to the final model. Here you can see that shopping, event planners, nightlife, and Art& entertainment as we found are among the businesses that start delivery and have a contribution to the out final model.

Length of operation: We have created multiple new features and among them "Length of operation" was a significant feature that contribute to the model and helped us to have a better view of the actions Yelp took after the pandemic to update businesses about campaigns. We can see that the main feature that leads to offering virtual services is the length of operation of that business.

Our Recommendation to Yelp: Based on the analytical result of the ML part we suggest the following steps for yelp:

1. Target the businesses that are active during the weekends and the business categories like active life, event planning, nightlife, and shopping. Specially Shopping businesses and Active life. Because their nature of business is like restaurants can send services and goods to the homes.
2. Yelp has a variety of services, like the call to action, relief, stay connected, and financial supports, as we discussed. We suggest to yelp that use the different channels to reach and inform the business about these campaigns, they can reach shopping categories customer and rest by Instagram, email and advertisement, but the one important suggestion is about not just focusing on the length of the relationship of the customer but try to reach to businesses that are newly entered into the customers of yelp.

Data and Model

In this research, we had 6 datasets and we did the main necessary steps for preparing the data and base table, since we are using Pyspark and data bricks there are additional steps to prepare the final base table for applying the Machine learning models, here we briefly discuss the steps we followed and if there is a need for an explanation, we can discuss it more deeply. The steps we took were as follows:

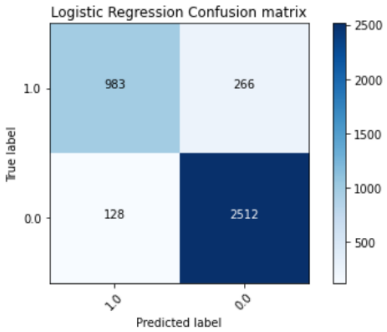
1. Preparing the business table:
   a. The business table is the main dataset that is a nested JSON that we use different schemas based on the nested columns to read it correctly.
   b. we use the yelp category file (in the reference you can find) to explode the category column.
2. Other tables and feature engineering:
   a. For the rest of the tables, we read the data and create a feature based on the recency, frequency, and other aggregations. In the following table, we have provided the details of these features.
3. Creating the base table:
   a. We have created a base table for running the models and also another separate base table for visualizations, the number of variables in the visualization base table is more since we haven not deleted anything.
   b. We have used the vector assembler to put our independent variables in the features column and after that, we have splitted the data to train and test and then used these tables to fit the model and do the predictions and analyze the results.

Data	Variable Created	Significance
Business	*_DT, Duration*	Duration of business activities on each day of the week
	*_cat	Categorical numbers for state and city
	All variables	One Hot Encoding of all variables
CHECKIN	Number Of Check-ins	How Many times user Checked in during the period
	Check-in time of the Day	Check-in Morning/ Afternoon/ Night
	Operation Duration	Total Duration of the Operation of the Business
REVIEW	Number of Reviews	Total number of Reviews per Business
	Sum cool/funny/useful	The total sum, cool and funny Reviews
	Average stars	Average rating per Business
TIPS	Compliment Count	Number of Compliments per each Business
	Count of Users' tip	Number of users given Tip

Models

Based on the data, the problem description, and our target variable we need to use Unsupervised classification Machine Learning models to make a prediction. We applied Logistic Regression, Decision tree, and Random forest Classifier to this project. The steps for each of these models are as follows:

1. Define the model and parameters that we want to use.
2. Fit model on the train set.
3. Predict the train and test set.
4. Calculate the accuracy of train and test predictions.
5. Calculate the AUC of train and test predictions.



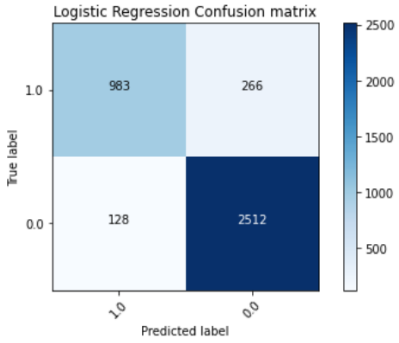
Logistic Regression

The results of Logistic regression with the maximum iteration of 5 areas are below the table, as we can see the results are acceptable and the performance of the model is good.

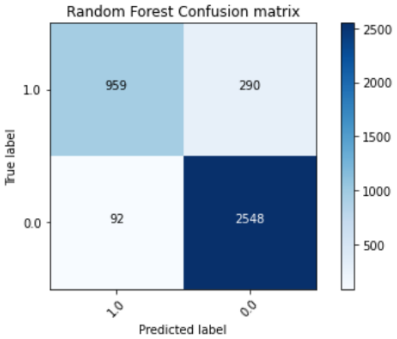
	Accuracy	AUC
Train	0.8886	0.8588
Test	0.8986	0.8692

Decision Tree

The results of the Decision tree with the maximum iteration of 5 areas are below the table, as we can see the results are acceptable and the performance of the model is good.



	Accuracy	AUC
Train	0.9069	0.8769
Test	0.9046	0.8704



Random Forest

The results of Random Forest with 10 trees for the depth of 5 and for seed of 42 is as below table, as we can see the results are acceptable and the performance of the model is good.

	Accuracy	AUC
Train	0.8972	0.8634
Test	0.9017	0.8664

In the end, we select the best Model that we will discuss in the following. We can say that all models perform well However, we have to check these results with CV and also we can do the Grid search to boost the results and improve our predictions.

Conclusions:

In conclusion, we want to mention the following issues:

1. We suggest using better resources for applying and validating the results of these data sets.
2. Because of the time limit and other resource limits we couldn't do extra feature engineering, we suggest applying sentiment analysis and checking if there is any significant difference between the two groups of target or not, however, we think that this could just help to improve the results slightly because none of the columns like funny or useful, cool or star that probably have a high correlation with sentiment analysis were among our important features. But if we had enough time and computing resources, we would have checked this as well.
3. Cross-validation is the most important part to validate this model since we don't have any validation set. We suggest creating a validation dataset or providing a solution to validate the data. However, if we had better resources we could have done that.

4. For the business we want to mention again that do not just focus on the food and restaurants. In marketing terminologies we call this as Marketing Myopia and it is better for a business like Yelp to develop a marketing strategy that is as inclusive as possible plus try to provide services to different segments based on their needs.
5. Here we can go deep and analyze more because the number of pages is limited and we just try to focus on the main concepts, we suggest conducting a deep survey to have better results.

Refrences:

1. https://blog.yelp.com/businesses/setting-up-call-to-action-business-highlights-connect/#call-to-action
2. https://blog.yelp.com/news/yelp-connect-a-new-voice-for-restaurants-to-reach-locals/?utm_source=biz_blog&utm_medium=yelp_blog&utm_content=blog_text_link
3. https://blog.yelp.com/news/yelp-covid-19-response-and-support-for-local-businesses/
4. https://docs.google.com/file/d/0B6-DBpFJUJgLQzNFQVg5SDZ1cnc/edit?resourcekey=0-vLeplmnGDU0PxyWWCBO2Hw