BIG DATA TOOLS 2

YELP

and its predictions for delivery or takeout

IÉSEG School of Management – April 4th, 2022

Ashwani Nitharwal

Tristan Helle

Maria Arques Rubio

Table of Contents

I - Executive Management Summary

The goal of this project is to create a model that will accurately predict if a business will be doing takeaway and deliveries after the first COVID crisis. To answer this problem, we decided to use two different algorithms, decision tree classifier and logistic regression. To fit the two models, we did pre-processing to avoid null values bias, deleted unnecessary information, feature engineered new columns to determine if businesses were open on each day of the week, the number of hours it is open, etc. After fitting the models, we have checked what variables were important for the model. We applied feature reduction to make the model go from 6500 variables to 10 variables. The model we have selected as our best performing model has an AUC of 0.804, this means that the model has a change of 80,4% of distinguishing accurately between a business that will do takeaway and delivery and a business that won't.

Thanks to the models we were able to understand some of the characteristics that businesses that have take out or delivery option contain such as few numbers of reviews or a good rating. From there we were able to build a profiling for future potential customers for Yelp. Yelp can use these criteria to target existent businesses on their website or increase its client portfolio. Later, we gave some recommendations to Yelp on how to approach potential customers to convince them to include the delivery option.

II - Business Insights

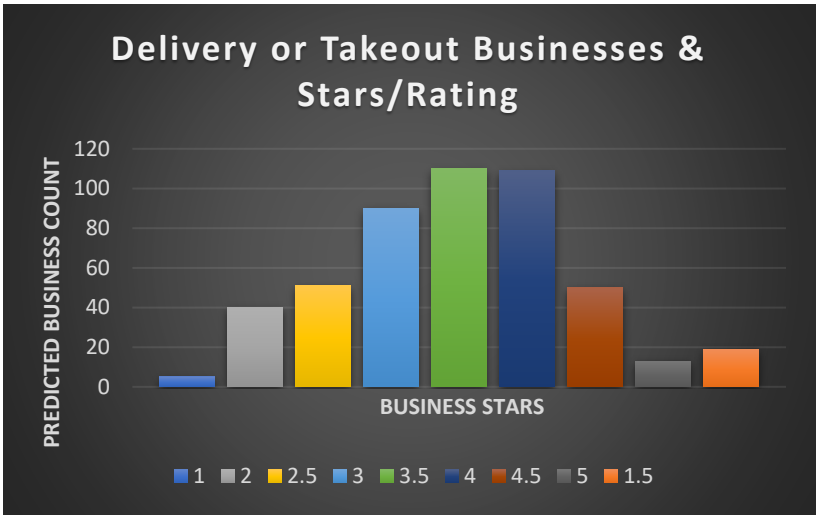
The COVID-19 pandemic has led, with instantaneous speed, to a major economic recession of absolute magnitude. What appeared to be a health crisis has brought with it economic and social consequences in different areas of our lives: beyond health, it has affected education, employment, businesses among others.

Companies that were affected by the recession were worried about the drop in the level of orders, which consequently caused a drop on the revenues. Businesses mostly from the leisure industry had to find new ways of reaching costumers to be able to maintain their gains and activity such as expand their way of selling to delivery and takeout options. Nowadays, companies are considering using Yelp as a review website to get advertised and to find a place where people can share their experiences. Reviews are useful for businesses so that they can be used for future potential clients when making decisions.

Moreover, for Yelp it is important to find the right businesses to sell advertisement; potential businesses for Yelp are the ones that are more advanced when it comes to digital transformation. Therefore, the data science team of IESEG, was hired by Yelp to achieve a prediction model that is more accurate for targeting potential businesses and to increase the response of the businesses that they are targeting.

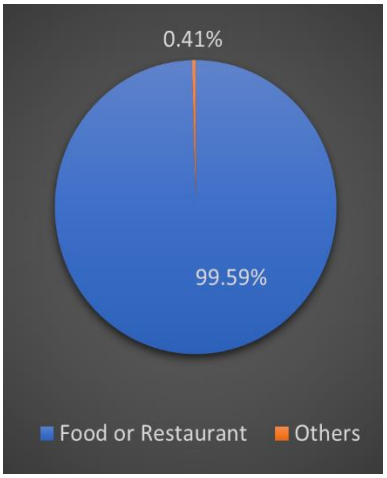
For Yelp it is necessary to comprehend the importance of predicting whether a business will start doing delivery or takeout since those businesses are more advanced in their digital transformation. Building an accurate prediction model could impact the company's performance to an important level and directly impact its revenues. The objective of our team is to provide not only a prediction model but also the profile of how does a business that does delivery looks like to be able to better target new potential clients for Yelp. Later on, some recommendations will be given on how to approach these potential customers.

Delivery or takeout Business Profiling:



As we can see in Figure 1, businesses that are likely to have the delivery or takeout option are those that have a good rating, ranging from 3 to 4 stars. Businesses that have stars from 4 to 5 are the ones that aim for excellence or that are focused on offering exclusive and premium services which can't be achieved throughout the delivery option (for example gastronomic restaurants).

Figure 1



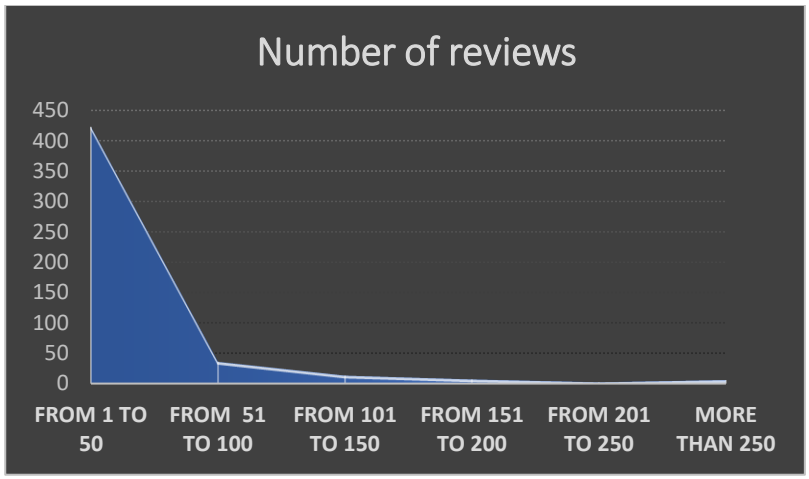


Figure 2 Figure 3

From figure 2 we can say that businesses that are likely to implement the delivery or takeout option are the ones that fall in the category Food and Restaurants. Even though we were provided with data from all type of companies from the leisure industry including recreation, restaurants, bars, and other social gathering hotspots, the ones corresponding to Food or Restaurant category are the potential clients for Yelp. If we take a look at Figure 3, potential businesses are the ones that have a smaller number of reviews, more specifically from 1 to 50 reviews. These businesses are not the most popular as people don't write many reviews and want to compensate it by trying to reach the highest number of customers offering the delivery option. Contrarily, businesses that are already settled down and that have good reputation (the ones with highest number of reviews), don't need to implement the takeout option.

Approaches for Yelp potential customers:

1-**Information campaign with interactive tutorials**: Yelp could create an informative campaign in a short video format that could be send to potential businesses with the main goal to try to persuade them to implement delivery or takeout. The video needs to explain how easy it is for a business to convert to takeaway and show the benefits of it, in terms of number of clients reached and grow in revenues. Moreover, the campaign could include short examples of how other companies implemented the option. A picture is worth a thousand words but a video clip, a million!

2- **1-month free trial**: Since March 2022, Yelp is collaborating with UberEATS "the ride sharing app's users now have access to Yelp's high quality local business content". Yelp can provide 1-month free trial sample thanks to its agreement with uber eats. Potential businesses will have the "perfect excuse" to try the delivery or takeout option and benefit from its non-charged price. Moreover, Yelp will have in their website more powerful and digital oriented customers. Lastly, UberEATS will benefit from it by adding new businesses as their clients. The best way to convince someone is by letting them try!

III - Data & Modeling Insights

In this part of the report, the authors will discuss how we worked with the given data, what variables we decided to keep, why we kept them, what variables we added to increase the accuracy of the model and the models we used.

I - Data pre-processing & featuring engineering

We received 6 different files the check in, covid, review, tip, user and business. The goal was to process each table individually and merge them on the business ID to have a unique observation for each business.

Checkin Table

For the check in table, we grouped by the different businesses and counted all of the different dates. It gave back a count of the different check-ins for each business. As there were no missing values for this table, we did not process it any further.

Business Table

The business table had many missing values for a lot of its rows, for example attributes_AgesAllowed and attributes_AcceptsInsurance had more than 18 thousand missing values out of the 20 thousand original rows. We decided to remove all the columns that had more than 30% of missing values as replacing them by the mean, mode or a given value would strongly bias and alienate the data. Another variable that contained missing values was “categories”, as there were only 39 missing values, we assumed they were restaurants (more popular category). For the missing values of the opening hours of the different days we replaced them by the most occurring value of each respective column. Once we had our basetable containing one row per business id, the authors proceed by filtering the data based on the category type. As the main goal of the project is to predict which businesses will do delivery or takeout, categories that are related to food or grocery shopping are the useful ones. Therefore, we only kept the businesses that have that contain these types of categories: Restaurants, Food, Grocery, Bars, Coffee, Tea, Bakeries, Steakhouse, Noodles, Cafes, Vegetarian, Seafood, Breakfast, Shopping and Brunch. This reduced the number of rows we were working with to 11 thousand which is half of the original dataset and eventually made all of the following modeling faster as well.

Covid Table

We assigned binary values to the different variables in covid to make the processing easier for the models we will use further. We then apply a group by and aggregate descriptive information for each business individually.

User Table

The user table has been left out as it does not provide any useful information regarding the business.

Tip Table

For the Tip table we performed a group by and summed the number of tips for each business.

Feature engineering

Once all the tables were merged, we started by creating a column that gives the opening and the closing hour separately for each day of the week. From there, we calculated how many hours a day the businesses were open by simply subtracting the closing and opening hour .

We also added more specific categories such as is_drinks_restaurant if the business sells teas, juices and cocktails and is_food_restaurant when the business is only selling food.

Columns that specify what day of the week the business is open has also been added. If the opening hours of the day are above 1 then the business will be considered open that day.

Removing columns that were not providing useful information such as name, longitude, latitude, postal_code and address was the final step with some variable renaming.

II – Modeling

We used the following models after manually indexing categorical variables: Decision Tree classifier & Logistic Regression. We separated the train and test by doing an .20 .80 split. We see that the variable importance for the Decision Tree Classifier is the following:

	idx	name	score
1	26	is_food_restaurant	0.8425607104432826
2	1	review_count	0.07499445630829882
3	12	Highlights	0.020122167397760415
4	8	Grubhub enabled	0.011006417726766303
5	16	avg_stars	0.009255662177527303
6	22	hours_open_sunday	0.007380738648164724
7	5	count(date)	0.007153096210107126
8	25	hours_open_wednesday	0.004762226623844207
9	479	categories_v_Florists, Shopping, Flowers & Gifts	0.004424115128688433
10	7142	hours_Sunday_v_17:0-23:0	0.0028798999008586186
11	35	open_days_week	0.0026725280676295118
12	498	categories_v_Flowers & Gifts, Shopping, Florists	0.0022741929255572484
13	14	number_of_funny_votes	0.0019580098723555347
14	7168	hours_Sunday_v_9:0-23:0	0.0016847958324629469
15	518	categories_v_Florists, Flowers & Gifts, Shopping	0.001532455614215092
16	7105	hours_Sunday_v_9:0-17:0	0.0015030073100479886
17	20	hours_open_monday	0.0012095970079221152
18	108	city_v_Painesville	0.0011442435219170526
19	19	hours_open_friday	0.0008768807786328462
20	4	tips_count	0.0006047985039610577
21	5983	hours_Monday_v_15:30-23:0	0

First, we can see that selling food is extremely important for takeaway and delivery while selling drinks is not. Secondly, we see that the shops with the highest review count are the shops that are most likely to do delivery or takeout. Thirdly, the Highlights variable is considered as important. We can also see those other variables such as Grubhub are impactful. This is logic as Grubhub is a delivery and takeout company, companies who have Grubhub enabled are very likely to do takeout and delivery already. Finally, the opening hours of special days of the week are important to determine if a business will do delivery or takeout as we assume deliveries and takeout are more frequent on certain days.

We refit the model using feature reduction techniques. We therefore ended with 3 models for the Decision Tree Classifier. Model 1 was ran on all of the features, Model 2 was ran on the top 20 features and Model 3 was ran on the top 10 features.

Model 1 resulted in an AUC of 0.849 and an accuracy of 0.637 which is a very good result, but we had around 6500 ids for variables which made the model very complex and not very good for business explanation wise. The Model 2 with 20 features performed worst with an AUC of 0.798 and an accuracy

of 0.856 which was to be expected as we reduced the number of features. Overall, the result is still very acceptable. The last model, Model 3 that ran only on 10 features gave an AUC of 0.804 and an accuracy of 0.856 which is worse than Model 1 but better than Model 2. Considering the complexity trade-off, we will consider Model 3 to be the best model for Decision Tree Classification.

Finally, the second algorithm we decided to run was the logistic regression. The logistic regression gave an AUC of 0.76 and an accuracy of 0.969 which is worse than all of the models from Decision Tree Classifier for the AUC but better in terms of accuracy, as AUC is the best measure for performance, we will only take that into account.

We will therefore keep Model 3 (Decision Tree Classification with top 10 features) as the best performing model with an AUC of 0.804 and an accuracy of 0.856.

VI – Recommendations

Further work that can be done for this project is cross-validation. We have tried to perform 10-fold cross-validation in order to have more accurate results but the code was taking too long compared to the time restriction, the data bricks cluster was also getting disconnected after running for a few hours.

Creating more features based on the categories like the ones we created that ended up as the top variables for predicting delivery or takeout (is_food_restaurant).

Running the dataset on other categories than food related, and grocery shopping related categories could also provide exploitable results.

Improving data quality by providing advantages to businesses that provide full information on their attributes would be the best option as it would allow us to exploit all the different attribute columns.