ML to predict whether a business will start doing delivery or takeout after a lockdown - Spark

Mario Cortez | Nour Azar | Tatiane Dutra

Individual Project for the Module Big Data Tools
Professor Steven Hoornaert

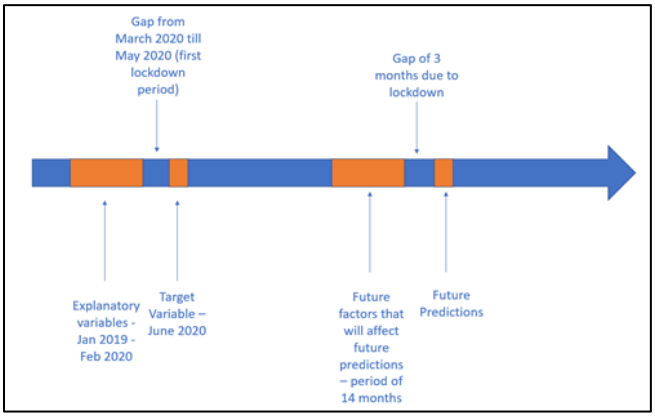
Master of Science in Big Data and Analytics for Business
IESEG Business School

Executive Management Summary

Following the covid pandemic, businesses had to make drastic changes in the way they operate for them to cope with the new pandemic Era. Some businesses adapted to these changes by adding the option of delivery and take-outs to their services after the first lockdown started. In this report, we want to create a model that predicts whether businesses will start delivery and take-outs after a lockdown.

In order to create this model, we will use data from 19018 businesses that is available in Yelp's platform, like general information, reviews, tips and check-in. As per the following timeline, we will consider information from a period before the first lockdown (January 2019 till February 2020) as explanatory variables to identify business that are doing delivery or takeout after the lockdown (June 2020).

Once the model is built, we understand what are the factor that cause businesses to start delivery or takeout, and in case of a future pandemic or global lockdown, the model could help Yelp to predict which businesses will start delivery or takeout, so that Yelp could target and do business with them such as sell ads in its platform.



To build the models, we need to train it with a table that has information about each business as the low-level analysis. Since the review, check-in and tips are data series information, we had to aggregate them, by count, take the mean or sum their variables (the details of these transformation are available on the technical session).

We also need to check the performance of our model with an unseen data. Therefore, we split the data to train the model with a subset of it and to calculate the performance with the other subset.

For this analysis, we applied two models: the Logistic Regression and the Random Forest. The Logistic Regression had a performance of 86,5 AUC while the Random Forest had a performance of 85,9 AUC. Since the Logistic Regression model has the highest AUC, we will consider this model as the final model of this report.

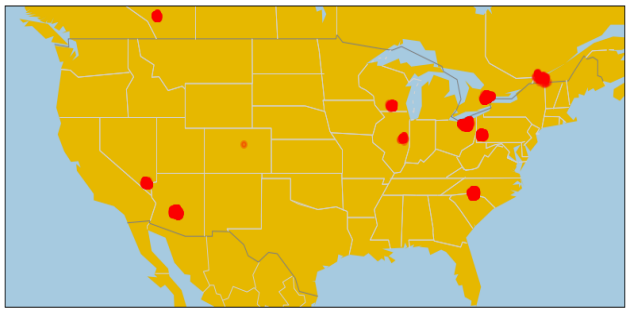
Finally, we also want to provide insights on which criteria are important for determining whether a business will start doing delivery or takeout. Therefore, we performed a feature importance analysis and identified that business within the category Food&Beverage is more likely to do delivery or take-out, which is obvious. On the other hand, business within the category Beauty and lifestyle has less probability. However, we can see that business that receives more tips in Yelp and has usually more check-ins in the weekend, has more probability to do it. In addition, there are some cities which has much more probability to do deliver-takeout, as Mississauga, while others have much less, as Beachwood.

Business Section

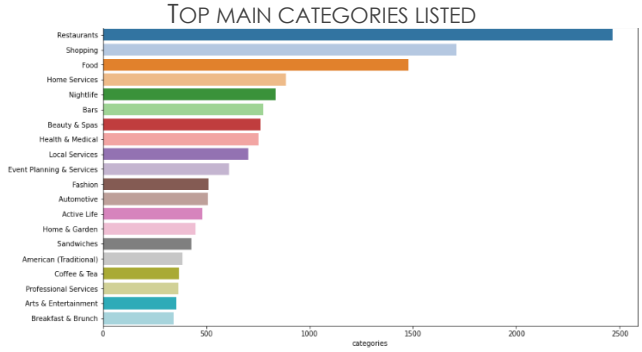
In this section we will show some insights from the raw data. And show the profile of the business that will likely do delivery on the lockdown.

Initial Insights Raw Data

THE DISTRIBUTION OF BUSINESSES ACROSS THE WORLD

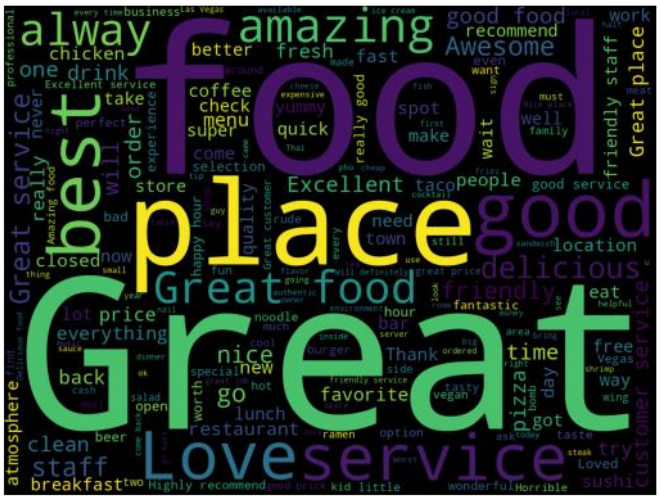


TOP MAIN CATEGORIES LISTED



As shown on the map above, the businesses in the dataset provided are located in Canada and USA, more specifically in the following states: Nevada, Arizona, Colorado, Illinois, Wisconsin, North Caroline, South Caroline, Pennsylvania, Ohio, Massachusetts, New York and Alberta.

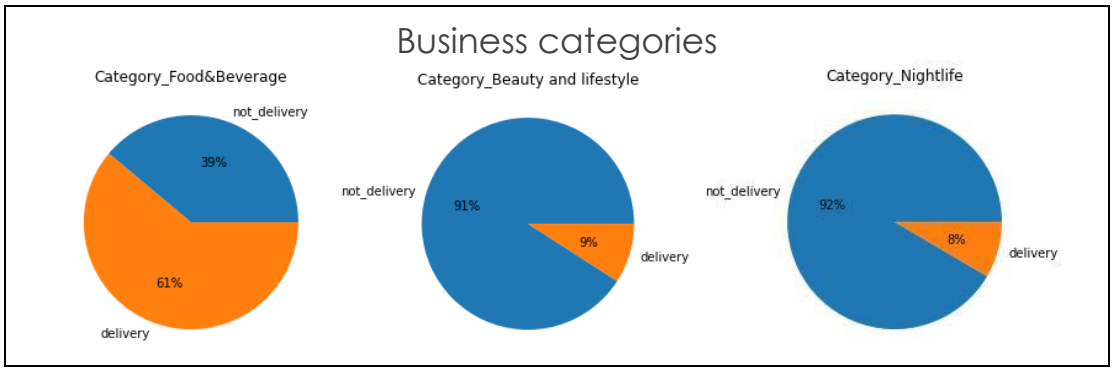
In addition, we conclude from the above bar chart that the top category of businesses on Yelp is Restaurants (including food) with about 2,500 businesses (The third category Food is probably included the restaurants category so we won't take into consideration its rank), followed by shops with about 1,700 businesses, home services with about 900 businesses followed by bars and nightlife with 1,600 users. Beauty & Spas and Health come in 7th and 8th position respectively with over 700 businesses each. The rest of the categories have lower than 500 businesses.



We notice that the reviews in Yelp are mostly related to food due to the high occurrence of words like 'food', 'pizza', 'dessert', 'eat', 'meat', 'appetizer' etc. and the most occurring adjectives are positive, such as 'great', 'amazing', 'good', 'nice', 'delicious', 'happy', 'kind', 'awesome' etc. In addition to that, we noticed that the most occurring words indicate that most of the reviews are related to visits and not delivery services, for example 'place', 'location', 'table', 'ambiance', 'went', 'come', 'server', 'waitress'. This makes sense because the data is related to the period before Covid when visiting restaurants was so common.

Business Profiling (Delivery or Takeout)

In the following section, we explore the characteristics of the businesses that start doing deliveries/take-outs after the first lockdown.



Companies that are in the food and beverage sector are more likely to do deliveries after the first lockdown since this category have the highest percentage of companies (61%) who do deliveries compared to other businesses. In contrast, companies that are in the beauty and lifestyle category and in the nightlife category are less likely to do deliveries after the first lockdown with 91% and 92% of the companies doing deliveries respectively.

	Tips Mean
Not Delivery_Takeout	2.221239
Delivery_Takeout	3.994965

	Mississauga	Beachwood
Prob	High	Low

	N° check-in Mean
Not Delivery_Takeout	45.5486
Delivery_Takeout	86.9139

The mean of the tips (or comments) received are 3.99 for the companies who do delivery/take-outs and 2.22 for the ones who don't do delivery and take-outs, which means that the one who do delivery and take-outs get 79% more tips than the ones who don't.

Businesses in Mississauga have more probability of starting deliveries after the first lockdown whereas businesses in Beachwood have the lowest probabilities of starting deliveries after the first lockdown.

The companies who make deliveries after the first lockdown have a higher check-in mean than companies who do not (91% higher).

	~~Mississauga~~	~~Beachwood~~
~~Prob~~	~~High~~	~~Low~~

	~~N° check-in Mean~~
~~Not Delivery_Takeout~~	~~45.5486~~
~~Delivery_Takeout~~	~~86.9139~~

How could Yelp target these businesses?

Previous reports showed that yelp has 33% higher likelihood of selling ads to businesses that are more advanced in their digital transformation. We suggest two ways on how yelp could target the businesses that are more likely to do deliveries after the first lockdown:

- Since restaurants are more likely to make deliveries after the first lockdown, we suggest that Yelp will specifically target them by offering free delivery fees if they click on the ads shown on yelp website.
- Since our model has a higher AUC, Yelp could invest more in marketing campaigns and create them with more quality by selecting few companies, but with a higher probability to be converted as a client.

Technical Section

In this session we will discuss about the technical part of this report. On the bullets points bellow we can see a summary of the technical configuration of this project:

- **Granularity** (Row Level): Business ID
- **Target Variable:** delivery/takeaway from Covid-19 dataset (period March 2020)
- **Input Variables:** Variables from the datasets Business, Check-In, Review, Tip, User (period from Jan 2015 to March 2020)
- **Timeline:** Inputs from Jan 2019 to March 2020 and Target in June 2020
- **Data Science Approach:** Supervised Learning/Binary Classification Prediction
- **Models/Algorithms:** Logistic Regression, Random Forest
- **Performance Metrics:** AUC and Precision
- **Performance Method**: Holdout cross-validation (split 80%-20%)
- **Tool:** Spark (SQL + DF API, ML, pipelines)

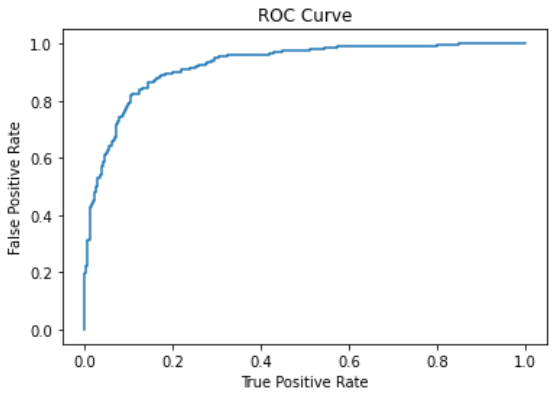
For the input variables, we receive data from Jan 2015 to March 2020, however, we preferred to use only one year of data, in order to reflect more recent information. Then, we have a gap of 3 months in evaluate the predictions, the covid dataset.

In the following table, we can see the features created from each dataset provided by Yelp. It is important to notice that every feature was aggregated to a business ID level, since it is the granularity of the analysis. All the original variables used in the feature engineering phase are listed on the table as well as the calculations and aggregations done with them.

Original Dataset	Original Variable	Aggregated Variables (Basetable)
business	business_id	-
	city	dummy variable
	state	dummy variable
	stars	-
	review_count	-
	is_open	-
	categories	Text Manipulation (to get the main category) + categories_dummy_vector
check-in	date	count_checkins count_checkins_weekday (dummy vector)
tips	text	tip_count
	compliment_count	tip_compliments_total (sum of compliments)
review	starts	review_count (how many times the business received reviews) star_mean

	useful	review_useful_mean
	funny	review_useful_mean
	cool	review_useful_mean
covid	delivery_takeout	-

We built two different types of models, the Logistic Regression, which identifies linear relationship and the Random Forest, a non-linear method. In order to measure the performance of them, due to computational issues, we used only the holdout cross-validation method, where we split the data into 70% for training and 30% for test. Both methods have a similar AUC, however the Logistic Regression performed a lit bit better, with 86,5 AUC, than the Random Forest, with 85,9 AUC. Therefore, the Logistic Regression was chosen as the final model of this analysis. Find bellow its ROC Curve.



After built the model, we want to understand how it looks like and one way to do that is calculating the Feature Importance. It refers to techniques that calculate a score for all the input features for a given model — the scores simply represent the "importance" of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable. The top10 features of our model in presented on the table below.

	feature_index	feature_name	coef	mean	std	std_coef	feature_importance
0	12	checkin_Wed	-0.068574	9.198216	41.944317	-2.876310	2.876310
1	11	checkin_Tue	0.047138	8.876115	41.192005	1.941693	1.941693
2	10	checkin_Mon	-0.030736	9.981169	51.148871	-1.572087	1.572087
3	15	checkin_Sat	0.023530	16.713578	66.800730	1.571854	1.571854
4	14	checkin_Fri	0.030222	11.495540	47.468141	1.434578	1.434578
5	6	Category_Food&Beverage	2.192087	0.584737	0.493012	1.080724	1.080724
6	7	Category_Beauty and lifestyle	-2.852321	0.117939	0.322695	-0.920431	0.920431
7	17	tip_count	0.152730	3.584737	5.653785	0.863504	0.863504
8	43	city_dum_Mississauga	9.618095	0.007929	0.088733	0.853444	0.853444
9	69	city_dum_Beachwood	-15.009470	0.001982	0.044499	-0.667912	0.667912

As we can see, the feature with more importance is the total of check-ins, however if we look at the coefficients, we can see that when we have a higher count of check-ins in the weekdays, it tends to have less delivery, while we have a higher count of check-ins in the weekends, it tends to have more delivery. The category of the business, the total tips are also important to identify if the business does or note delivery. In addition, we can see that the city Mississauga is more likely to have delivery while Beachwood doesn't, since they have a high coefficient but one is negative and other positive.

Sources

HOORNAERT S. (2022). Big Data Tools. [Course]. Lille: IESEG Management School. MSc in Big Data Analytics.

https://stackoverflow.com/questions/46528207/dummy-encoding-using-pyspark

https://www.youtube.com/watch?v=1a7bB1ZcZ3k

https://www.kaggle.com/code/sudhirnl7/basic-exploration-of-business-review-at-yelp-com

https://stackoverflow.com/questions/28009370/get-weekday-day-of-week-for-datetime-column-of-dataframe

https://analyticjeremy.github.io/Databricks_Examples/Write%20to%20a%20Single%20CSV%20File

https://www.kaggle.com/code/vksbhandary/exploring-yelp-reviews-dataset

https://sparkbyexamples.com/pyspark/convert-pandas-to-pyspark-dataframe/

https://towardsdatascience.com/machine-learning-with-pyspark-and-mllib-solving-a-binary-classification-problem-96396065d2aa