Business Delivery or Take-out Model Prediction

Executive Summary

In 2020, the world was hit by the COVID-19 pandemic, this impacted all parts of society. Because of this, many commerce sectors had to adopt new ways of operating and developing their business. That is why we have developed a prediction model which identified the factors by which some establishments began to make home deliveries or take out after the first lockdown.

The Problem

As part of the restrictions implemented to contain the COVID-19 virus, many sectors had to completely restrict the gathering of people and as is well known, the Restaurants, Foods & Bar sectors were highly impacted by this measure. That is why those businesses had to implement new ways of operating such as home delivery and take away. Since Yelp is a platform that offers information oriented to consumers, it is necessary to be able to estimate and identify the establishments that will tune in to this new way of operating, at least in times of pandemic.

Predictive Model Execution

We were provided with the six datasets which contain Yelp user information, which contain data related to the interaction of users and their interactions when issuing a review of the places visited. In the same way, these datasets provided us with information on each of the restaurants and their characteristics that build a different experience for each client who visits them.

Making use of the different data processing techniques, we built a base table with which we applied three classification prediction models: Logistic Regression, Decision Tree and Random Forest.

By making a detailed evaluation of the results of each model, we were able to determine that the Logistic Regression model was the one with the best precision when identifying which establishment would start making home deliveries and take outs. Additionally, the different factors that drive these places to carry out the aforementioned activity can be appreciated.

Highlights

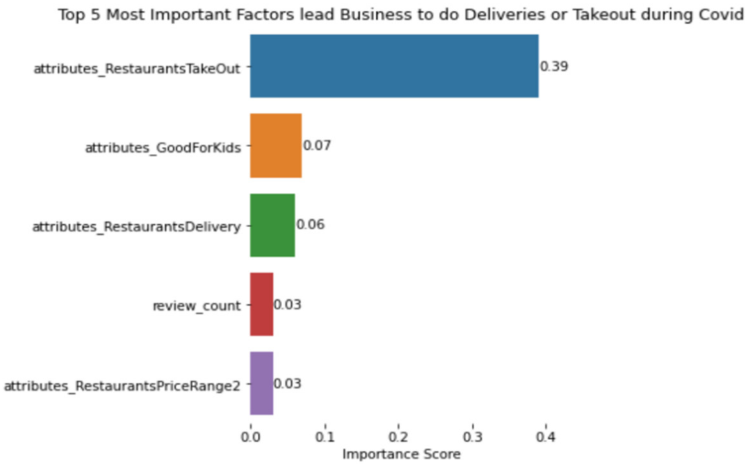
The analysis carried out is made up of two sections, Business and Technical. In which each one can find the different insights that give value to the executed model.

In the Business section you can find the different factors that lead establishments to make deliveries or take outs. Additionally, three different proposals were issued at the marketing level so that Yelp can attract these businesses and in this way be able to establish a commercial relationship with the purpose of pooling the profits of both parties.

In the technical section, you will be able to identify the different parts of the elaboration process of our model, as well as a detailed description of the provided data sets. To finish, you will have the measurements made to each model to check our results.

Business Section

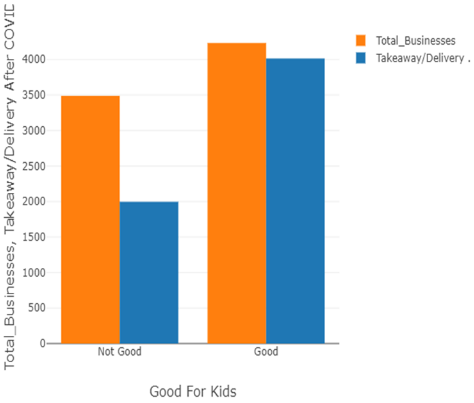
Identifying the businesses that will do delivery or takeout opens a lot of marketing opportunities for Yelp. This section explains some of the factors that were considered important in deciding whether a business will start doing delivery or takeout.



This chart on the left displays the top 5 factors identified as important during the prediction if they do delivery of takeaway.

The most important factor is identified as Restaurants with takeout availability before the COVID-19 pandemic. This factor had a massive weightage of 39% in the predictions made. The number of businesses that had takeout before COVID is 5673 and the number of the same businesses providing takeout after COVID is 5654. This proves 99.66% of the businesses that provided takeout before the pandemic have provided the same even after the pandemic.

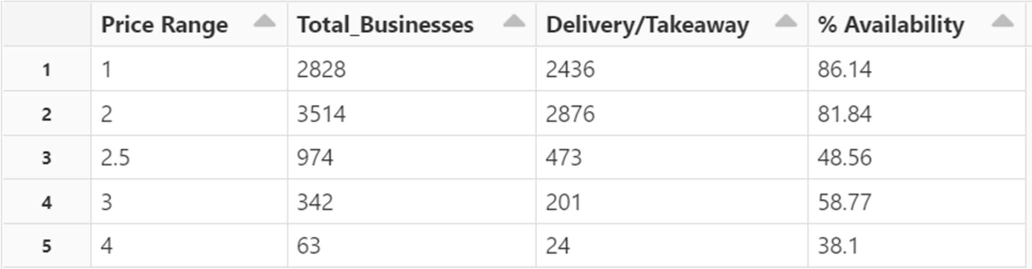
The second most important factor is if a business is good for kids. This factor had an 7% of prediction power in our model. From the Data, it is found that 94.82% of the businesses that are good for kids have delivery or takeout availability after COVID19 while only 57.2% of businesses that are not good for kids had such availability. Hence, it is more likely for a business that is good for kids to have delivery and takeout options.



Kids friendly business factor is closely followed by a factor which is the delivery availability of the restaurants before the pandemic. As per the data, 4106 businesses out of 5806 which did not do delivery before COVID19 started doing delivery after the pandemic. This shows that the businesses are changing in operations after the pandemic.

The next factor in the list is the total count of the reviews in Yelp. It is possible that higher the count of the reviews, more popular and bigger the restaurants are, such restaurants might do deliveries and takeaways. The average number of reviews of businesses that do delivery and takeaway is found to be 62 whereas the number is just 32 for businesses that do not. Hence, more the number of reviews, more is the possibility that the businesses might have this option.

The last of the top 5 important factors is price range of the restaurants. The table below shows some data on how the price range categories impacts the delivery/takeout availability.



	Price Range	Total_Businesses	Delivery/Takeaway	% Availability
1	1	2828	2436	86.14
2	2	3514	2876	81.84
3	2.5	974	473	48.56
4	3	342	201	58.77
5	4	63	24	38.1

From the data available on yelp, we can prove that lower the price range of the restaurant, the higher is the possibility of the delivery or takeout availability. As the price range increases, this availability reduces gradually.

Suggestions for Yelp to target these businesses:

1. A lot of people work from home and order food. In promoting and helping users in choosing food quickly, adding a new webpage in yelp's website called "EFH: Eat from Home" can work. In this webpage, in addition to displaying ads, Restaurants can have a toll-free number called "Call for cuisine suggestions" or an additional space in the webpage where they can input the best cuisines or must try ones in that place. "Eat from Home" can also have a subscription feature where users can subscribe on weekly or monthly basis.
2. Creating a new feature for the businesses that can automatically let businesses share "review of the day" to their social media platforms based on text mining. The algorithm will check the text of the review and emotions in it and if the review meets certain conditions, it will be shared on the restaurant's social media account as a picture, once every day.
3. Reviews drive the businesses for restaurants and other food related businesses. To motivate users in writing reviews, restaurants can promote discounts for the exchange of reviews on yelp. Restaurants can choose rules such as top 10 reviewers, or only considering reviews with pictures in it or reviews with valid invoice IDs mentioned in it. The discount codes can be sent to the users automatically using Yelp's technology and businesses can set workflows to manage these functionalities.

Technical Section

Problem Identification & Data Science Approach

In this case, we are trying to identify what are some of the most important factors lead some business to start doing a delivery or takeout for the first time after the first lockdown. We do this using 2 approaches, first we do a comprehensive, in-depth Data Analysis to understand what do these businesses have in common and try to see if there could be any link in between. Next to this, we are also developing a predictive model, that essentially tries to predict, given on certain variables, which of these businesses are going to start doing a delivery or takeout. We do this by benchmarking different models, we then evaluate and compare them and pick the best one. The focus of the technical section concerns the second approach: Predictive Modelling.

Basetable Creation & Feature Engineering

We have a total of 6 tables in our dataset:

Business table, the business table contains detailed information about businesses. It contains some demographical information about the business (name, location), operational information (opening hours, are they open), attributes (do they have a parking, do they have a Wi-Fi, is it good for kids, etc.), and YELP related information (stars, number of reviews). The business table is the basis of our basetable since this table is already on business-level granularity. The way we process this table is we first filter some of the business in our data so that it only concerns the businesses related to food servings (Restaurants, Foods, Bar). Final step is to identify and drop variables that we think has no relation to delivery or takeout.

Check-in table, the check-in table only contains 2 information, the check-in time and the place/business that has been checked in. Check-in is the action that could be done by Yelpers if they visit a certain business, it is essentially a "prove" that someone visited a business. Some of the interesting things that we calculated from check-in table are how many times does a certain business has been checked in by people, and what was the period between our time reference (April 2020, before covid lockdown) and the last check-in. Both information could be useful to identify whether is there any some sort of link between these variables, and the business decision to start doing deliveries or takeout.

Covid table, the covid table is the only table that is recorded in June 2020, following the first lockdown because of covid. It contains information about covid-related attributes for businesses (do the businesses do deliveries or takeout, do the businesses offer virtual services, do they have covid banner, etc.). We first filter out businesses that are temporary closed during covid, since it wouldn't be fair to include these in our model (they can't opt in for delivery or takeout if they aren't operating). We then only take the information variable on whether a business do deliveries or takeout after the first covid lockdown.

Review table, the review table contains information about reviews that has been given by a user, to a business. It contains the review text itself, the stars given, the time of the review was given, and some reactions to the review (number of people who find this funny, useful, or cool). Some interesting variables we calculate from this table are average stars given to the business, the number of reviews, the number of reviewers, the average "useful" reaction count per reviews and the period between our time reference and the last review given. This information might be useful to identify whether a business that has been reviewed a lot and generally has good reviews (as identified by the stars and the "useful" reaction count) will lead to doing delivery or takeout after the first lockdown.

Tip table, the tip table contains information about tips that has been given by a user to a business. In Yelp, users can give a tip to the business, tip is the shorter version of a review, and is usually given as a key information about a business. From the table, we calculated the number of tips given to a business and the period between our time reference and the last tip given.

User table, the user table contains detailed information about Yelp users. The problem with handling the user table is that this table is on user-level granularity, so we need to merge this table with reviews table first, and then do aggregation on to the business level there. For each user, we calculated the number of reviews given, average stars given, and the length of relationship.

After we processed all the tables, we merge them into one table and form our basetable. Our basetable in the end contains 7721 different businesses and 55 variables about the businesses. In summary, our basetable contains detailed information about each business and their attributes, including their performance on Yelp (reflected by the stars, tips, & reviews) and their relationship with the Yelp users that gives them reviews (reflected by the number of reviewers, etc.).

Model Development, Evaluation, and Selection

We developed different predictive models and compare them to pick the best model. We first do ML preprocessing on the basetable: Missing values imputation and One hot encoding for nominal variables. Next step is to do stratified splitting into training and testing, we will fit our model based on the training set, and then evaluate them based on the testing set. We developed a total of 3 models:

- Logistic Regression
- Decision Tree
- Random Forest

Hyperparameter tuning is also done through grid search for each model. Best model output of the grid search is based on 10-fold cross validation. Each best model output from the grid search is tested on the testing set, and then evaluated on 2 criteria: AUC and Precision & Recall score.

Model	Train AUC	Test AUC	Precision		Recall	
			Class 0	Class 1	Class 0	Class 1
Logistic Regression	0.97	0.96	0.85	0.99	0.96	0.95
Decision Tree	0.98	0.95	0.88	0.98	0.93	0.96
Random Forest	0.96	0.9	0.88	0.95	0.84	0.97

Our Model has a great performance in terms of AUC, this is mainly because of one strong predictor, which is an identifier whether a business do takeout before covid. Precision & Recall also shows that our model predicts class 1 (do delivery or takeout) better rather than class 0. This is because we have imbalance classes in our dataset, we have more businesses with class 1 as opposed to class 0. In addition, an overfitting problem could also be seen from the Random Forest Model. Based on model evaluation, we have decided to select Logistic Regression as our best model.