IÉSEG SCHOOL OF MANAGEMENT

BIG DATA TOOLS GROUP PROJECT REPORT

Prepared for Yelp Data Analysis

April 2022

Presented by

Hina Hussain
Enita Omuvwie
Dilda Zhaksybek

Presented to

Steven Hoornaert

Executive Summary

The purpose of this report is to help Yelp identify and predict what factors lead some businesses to start doing delivery or takeout for the first time after the first Covid-19 lockdown. Moving towards delivery and take out shows a business's digital transformation and previous reports show Yelp has a 33% higher likelihood of selling ads to digitally transformed businesses. This report will not only help to identify the characteristics of such potential businesses but also propose strategies to reach out to them. This is crucial as 51%[1] of Yelp's revenue growth is tied to business partnerships which serve as its sustenance.

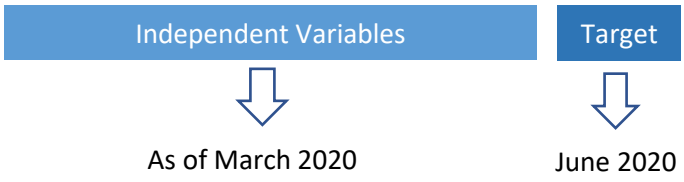
The first section of this report focuses on the technical aspect of the project. It is meant for a data science audience and contains the timeline of the data used, the preprocessing steps applied to the data tables, creation of the basetable, validation techniques used, the machine learning algorithms used, and the evaluation metrics applied. Extensive feature engineering was used to develop a robust classification model that has high predictive power and can successfully be used to identify potential businesses which will start doing delivery/takeout.

The second section of the report is targeted towards business development executives. It uses the results from the technical modeling to propose strategies for business development. The proposed solution aligns with Yelp's goal to expand its business model to include the factors that help businesses during the covid lockdown. The section outlines the profile of businesses likely to start delivery/takeout. Yelp can use this information to capture the gap in the market using two digital strategies - monthly reports & digital transformation. Implementing those strategies will benefit and help promote growth of both businesses – Yelp's and its clients.

[1] Wairimu, Sg, & Olagbaju, K. (2019, November 28). *Yelp business model: How does yelp make money?* Business Strategy Hub.

Technical Section

The aim of this section is to walk through the data science pipeline followed in this project from start to end. This project uses data from Yelp to identify and predict what factors lead some businesses to start doing delivery or takeout for the first time after the first Covid-19 lockdown. As displayed in the figure below, the drop date of the data is March 2020 where the independent variables are the status before the drop date and the target variable is binary indictor of whether the business is doing takeout/delivery in June post first lockdown.



This project uses the Apache Spark platform, Databricks, to run the entire analysis. The data set provided by Yelp includes six tables containing a range of information about the attributes of a business, its reviews, its users, and check-ins. Each table was processed separately to create useful features.

Business Table

- Filtered to only keep businesses that are in the food industry using categories “Restaurants”, “Food”, “Café”. This was done because delivery/takeout is only relevant to food businesses. Using all the businesses resulted in overfitting down the line during the modeling stage.
- Dropped variables descriptive of non-food industry businesses
- Missing categories for business attributes were labelled “Missing”
- Dictionaries were exploded to create dummy variables
- Opening and closing hours were used to obtain the number of operating hours per day and aggregated by weekend and weekdays.
- Main Variables per business: dummy variables for attributes such as already delivering and doing takeout, ratings, casual, upscale, vegetarian, accepts credit card, availability hours, city etc.

Review Table

- A discount factor was computed based on the recency (from 1 March 2020) of the reviews to adjust the ratings i.e., older reviews have less weight and recent reviews have more weight.
- Based on the original ratings, it was assumed that if the rating is higher than 3, the review is positive and if it is 3 or less then the review is negative.
- Main variables per business: average adjusted rating, count of total reviews, and count of positive and negative reviews.

Other Tables

The main output from preprocessing the users, check-in, and tips table is summarized below.

- Count of check-ins and the number of days since the last checkin from 1 March 2020
- Count of tips per business
- Average number of fans of the reviewers of a business

After preprocessing, each table had one row per unique business. All the tables were combined to form a basetable. Special attention was given to the type of joins. Null values and "missing" categories were dealt with. The target variable is a variable indicating whether a business is doing takeout or delivery and is obtained from the covid table. The dimensions of the final basetable are (7780, 52).

Feature Selection

Chi square test was used to measure the relationship between the independent and dependent variables and the result was used to only select 90% of the top features. This helped create a more robust model.

Validation Technique

The dataset was split into train and test sets using a 70:30 split. The data was converted into labels and features format using RFormula. Two different classification models were run on the train set and predictions were made using the test set

Modeling

Two classification algorithms including Logistic Regression and Random Forest were run on the train set using a spark pipeline. Predictions were run and evaluated on the test set. F1 measure and area under the curve (AUC) were used to evaluate the models. Cross validation or hyper parameter tuning could not be used given the slow speed of the Databricks platform.

The evaluation metrics obtained from the models are summarized in the table below.

Model	AUC (Train)	AUC (Test)	F1 Measure (Test)
Logistic Regression	0.985	0.962	0.973
Random Forest	0.987	0.965	0.977

The results show that the models have high predictive power and are successfully able to classify whether a business will start doing takeout/delivery or not. A high train AUC can mean that the model is overfitting on the data but trial and error with different features and business categories showed that the current way of building the model provides the best results. Predictors that have the most importance using the random forest model were extracted to create business profiles. These profiles and suggested business development strategies to reach such businesses is explained in the next section.

Business Section

Yelp helps authentic businesses have an online presence, increase their reach, and expand their customer base. Businesses can also view the reviewer engagements on Yelp to know their service ratings from their customers and what their potential market, customer engagement/experience and brand exposure looks like.

According to the Yelp report in December 2018, $907 million was generated from ads selling alone, this makes up 97% of its total revenue generation. Yelp has made Ad selling its primary revenue generator and this encapsulates everything a business needs for product exposure. Yelp has a large team of sales executives who sell these ads to local businesses. These advertisements include brand profile, Yelp verified license, enhanced profile, and other search ads/ads resales. The company has a plan to acquire more business partnerships and customers, develop new products (delivery services), and grow the audience for local businesses. There are several transactional services it provides to customers and businesses on its platform that also include deals, discounts, partnerships with restaurants, reservations, gift certificates and so on. These services will help improve any business on the platform to get through the covid lockdown without closing completely.

Below are suggestions on how Yelp can further increase revenue from advertising or sales of its services to businesses that are predicted to offer delivery and take outs. Moving towards delivery and take out shows a business's digital transformation and previous reports show Yelp has a 33% higher likelihood of selling ads to digitally transformed businesses

Business Profile

The model identified the following business characteristics which show the business will start doing delivery/take out after lockdown:

- Businesses that have high number of reviews on Yelp platform, signifying popularity.
- Businesses that have recent check-ins, signifying recency of the visits.
- Businesses rated as good for lunch or casual outing – identifying preferred category.
- Businesses have higher number of positive than negative reviews – signifying sentiment of overall experience.
- Businesses that have longer availability during workdays.

Yelp can target those businesses that fit the criteria to promote advertising and bring value in time of the pandemic and global lockdowns.

Monthly Summary Report

For instance, Yelp can introduce automatic monthly summary reports generated for those businesses that are predicted to make deliveries and take outs. The report will show summary statistics per business that already exists within Activity section such as reviews, visits, tips, and ratings. By adding summary on categorial reviews relevant for users during lockdowns, such as preferred working hours, dietary restrictions, credit cards acceptance and others that can be used both to improve what is already preferred by clients and also used as search keywords. In this report, Yelp will highlight features that users identify as top qualities for why a client chosen that particular business. That analysis of the features can promote businesses to improve on those valued features in time of pandemic, which in turn generates loyalty towards Yelp as a platform, as well as proves value and incentive into keep investing in Yelp ads.

Besides, Yelp introduced Keyword Boosting feature that allows businesses to specify for which search keywords show advertising. The monthly generated report should suggest keywords that would likely be

very effective if chosen as advertising parameters, further increasing engagement and interest with Yelp's advertising tools.

Thus, via introducing a monthly summary report per business, Yelp produces more value for existing advertisement clients, retains them and attracts new businesses into advertising with Yelp, not to mention support for businesses Yelp offers in time of pandemic.

Digital Transformation

As previous reports suggest, businesses that are advanced in digital transformation are 33% more likely to post advertisement on Yelp. There are several variables that suggest whether business is digitally advanced. Businesses may accept bitcoins to suit a wider variety of digitally invested clientele, have Wi-Fi at their locations, have website with menu and all relevant information, take orders by online forms, accept online reservations, collect, and analyze customer data, and much more. All this signifies that a business is directed towards serving technically advanced client base, which is even more relevant and even necessary in time of global pandemic which shifted practically all business online.

Yelp can increase advertising revenues by promoting and supporting digital advancement in the business. It is essential to explain the benefits of technological transformations to business, as well as how to achieve them via:

- Publishing series of blogposts or personalized emails.
- Holding online workshops or conferences for small and medium business.

Onboarding businesses who do not fall into the digitally transformed category should include, but not limit to introducing wider variety of payment methods, implementing social media advertisement for targeted audiences, online reservations, monthly subscription plans for customers, personalized deals and offers, and lot more.

Additionally, during these promotional activities, Yelp's Reservations platform should be offered as one of the steps towards digitalization, as it brings indisputable value for businesses and clients in terms of ease of management of reservations, further driving revenues for both Yelp and signed businesses, as well as increasing specter of data available for the analysis.

Promoting digital transformation can help businesses grow while making Yelp benefit from digitally advanced businesses via increased advertisement and reservations tool revenues.

To conclude, strategies proposed above aim at increasing the marketability of Yelp's advertising tools and services, as well as support businesses in time of a global pandemic.