IÉSEG SCHOOL OF MANAGEMENT

YELP BUSINESS ANALYSIS

REPORT

3rd APRIL 2022

DATA SCIENCE TEAM

PEDRO ROMAN
DATA ENGINEER

KAMALAKANNAN THAYANIDHI
DATA ENGINEER

AZAD GHOSLYA
DATA ANALYST

TABLE OF CONTENTS

EXECUTIVE MANAGEMENT SUMMARY

OBJECTIVE

The Main Objective of this cusiness report is to analyze Pre-Covid Customer Food/Restaurant Review Data from the YELP website to build a prediction model. The main aim is to predict which of these businesses would adopt a Delivery or a Take – out Form of service model due to the New Covid Health Measures.

PROJECTLINE SETUP

The Project is set up using 6 distinct segments of Data which comprises of Customer Profile, Customer Review, Business, Check-in & Tips. By Pre-Processing all the various data and combining them in a pipeline format to create our basetable with all our dependent & Independent Variables.

As a next step, we start to train and build Machine Learning models on this base table such that the models can predict which businesses would adopt the Delivery/Take-out Model or not. Since it is a classification type Business problem, we would be running qualitative models.

Further, we would be implementing cross-validation techniques and checking the AUC score of different models to determine which of these models would give us the best prediction.

CONCLUSIONS

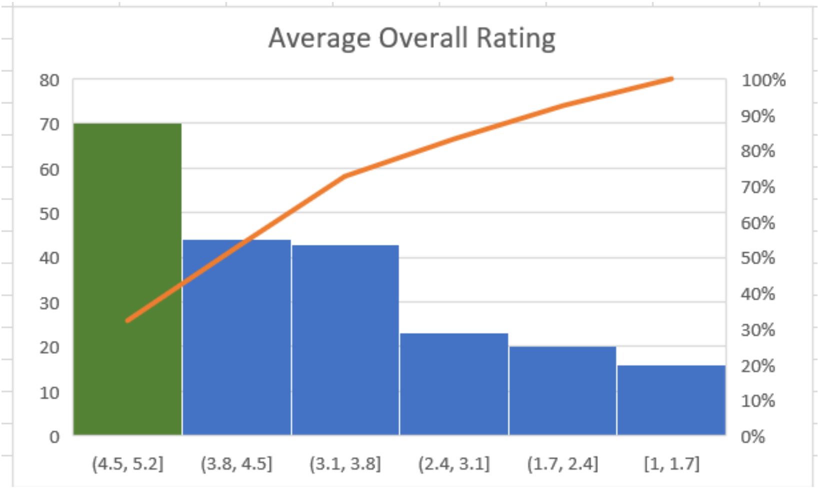
The final conclusions according to our predictions with an accuracy of 95.8% we predicted the restaurants that would adopt a Delivery/Take-out Model. The Machine Learning Algorithm that gave us this is output was Gradient Boosting with a 0.86 AUC Score. Based on the profile analysis performed to the predictions, we could determine similar characteristics among businesses that are more likely to include delivery or takeout in a new wave. Among these qualities it is noticeable the allowance for credit card payments, as well as the stars of the businesses on Yelp.

BUSINESS SUMMARY

BUSINESS ACTIONS

The Variables which had the most importance for the prediction models were mostly related to the user activity per business/restaurants. The Most Important variable was the Total Reviews Count followed by Average Overall Rating, Is Leisure and finally the total number of Tips and recency of Tips for a particular restaurant.

In the graph below, we can see that the average overall rating for the top 10% of businesses that are more likely to include credit card or takeout is condensed between ratings of 3.8 and 5:



Incentive Based Promotion -

While we cannot control the rating a particular customer gives to a particular restaurant, we can try to increase the business value from other variables. We can increase the total number of reviews a restaurant receives in a few diverse ways. Yelp can set up a business tie-up with these restaurants & chains and can provide vouchers or flat discount offers for customers who give reviews about those restaurants so that the customer has an incentive to dine again.

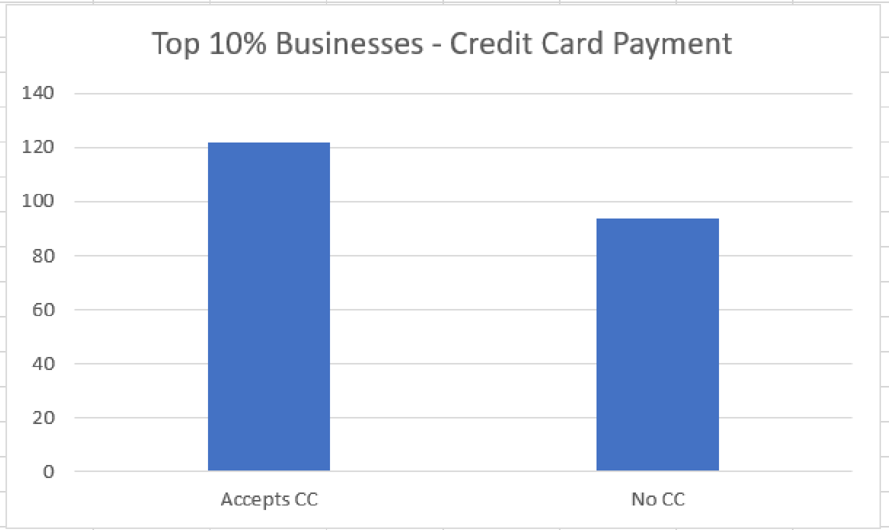
This action would increase the total number of reviews a restaurant receives and would attract more customers for a particular restaurant as they would eventually trust a restaurant with an average rating & high number of reviews than one with high rating & less number of reviews.

This would begin a positive feedback loop for both yelp platform as well as the partnered restaurants to increase the customer/user base and both would benefit from it.

Improve customer experience -

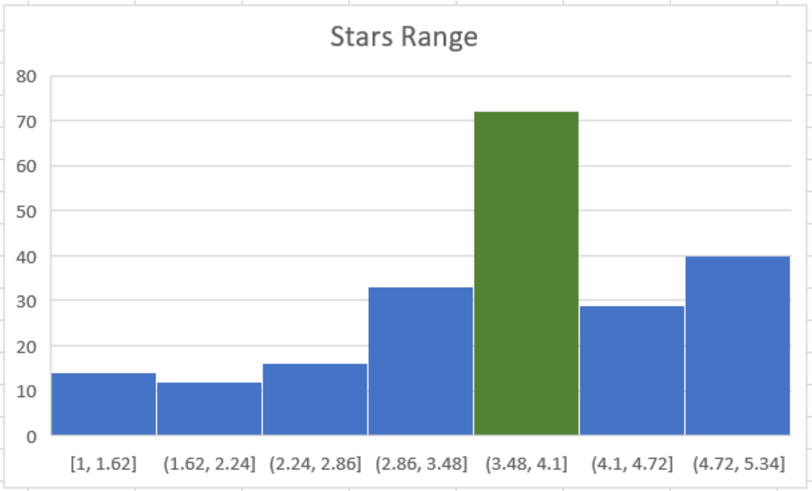
Avg overall rating: Yelp can provide the details of the customers reviews who are giving lower ratings to the restaurant so that restaurant can make improvements to enhance the customer experience in the future.

Credit card payment: Yelp can create shared promotions with payment gateway companies to increase the usage of credit cards among its users. This measure gives a win-win result since it helps the businesses' clients on their experience and at the same time enhances sells for the businesses. A graph below can demonstrate how important is the credit card payment when a business puts to disposition the delivery or takeout service:



Key Insights –

- Businesses that are open on weekends are more likely to include delivery or takeout during the first wave of the pandemic
- Among the the high likely new delivery or takeout business, leisure is the main category, comprised of mostly restaurants.
- The stars range is a strong indicator of a business including delivery or takeout.



METHODOLOGY

DATA PRE-PROCESSING

We begin the Data Preprocessing step with the Business Table as our Final Base Tables Granularity should be per business on each row (I.e., Our predictions model is for each individual business). As a First Step, we filter out all the Restaurants which do not have Take-out or Delivery before COVID, then only include businesses that were open for the timeline we have considered.

As a Next Major Step in Preprocessing, we start by dropping the columns which have > 75% Null Values and columns from which we cannot extract any relevant business information for our base table. Next, we transform & aggregate variables from the rest of the tables and merge them with Business Table. Next, we create dummy encoding for all the categorical variables present and drop the original columns from the base table. Further, we create individual columns that represent whether a particular column has a NULL value to account for all the missing data. Finally, we add our dependent/target variable column which takes value 1 for business where delivery/takeout is present.

IMPORTANT VARIABLES

The Top 5 Variables by Feature importance -

Review_Count – Total Number of Reviews a Particular Restaurant Received, **IsLeisure** – whether a particular restaurant is considered leisure or not, **Avg_Overall_Rating** – Average Overall Rating a Restaurant Received, **User_Tip_Count** – Total Number of Tips Users have given about a Restaurant, **Recency_Tips** – The Recency of the Tip Users have given about a Restaurant.

CROSS VALIDATION

As part of the validation process, the Cross Validator splits the dataset into folds, which are used as separate training and test datasets. It is repeated model fitting. Each fit is done on the major portion of data and then tested on the left-out portion.

For this data pipeline, we have defined the CV Folds to be 10 (I.e. which means it generates 10 pairs of train, test datasets). To evaluate a model the Cross Validator computes the average evaluation metric for all the Models produced by fitting the Estimator on all different (training, test) dataset pairs.

EVALUATION METRICS

The two different evaluation metrics which we are going to consider for comparing the models are AUC & F-Score. The AUC Curve is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

ALGORITHMS & PARAMETER TUNING

The Different Algorithms we are testing for this project is Logistic Regression, Random Forest & Gradient Boosting Classifier.

	LOGISTIC REGRESSION	RANDOM FOREST	GBM
AUC TRAIN MEAN	0.970	0.980	0.947
AUC TEST MEAN	0.866	0.884	0.854
Accuracy Train	0.975	0.958	0.968
Accuracy Test	0.957	0.959	0.957

From the above comparison, we can see that Random Forest is performing the Best of the Three models with the highest AUC Test Mean Score.

The Parameters that gave us the best AUC Score for Random Forest are

Max depth– 10, Max Bins – 20, Number of Trees – 20.

REFERENCES

1) https://george-jen.gitbook.io/data-science-and-apache-spark/cross-validation
2) https://knowledge.informatica.com/s/article/588927?language=en_US
3) https://medium.com/@aieeshashafique/gradient-boost-model-using-pyspark-mllib-solving-a-chronic-kidney-disease-problem-13039b6dc099
4) https://spark.apache.org/docs/1.6.0/mllib-evaluation-metrics.html#binary-classification
5)