Restaurant Feasibility Predictor

Author: Anshum Saini

Problem Statement

The goal of this project is to create a model that can help the investors/lenders as well as the restaurant owners in their decision-making process. This project analyzes restaurant's data and its consumers' reviews to predict the feasibility and sustainability of the business.

Background

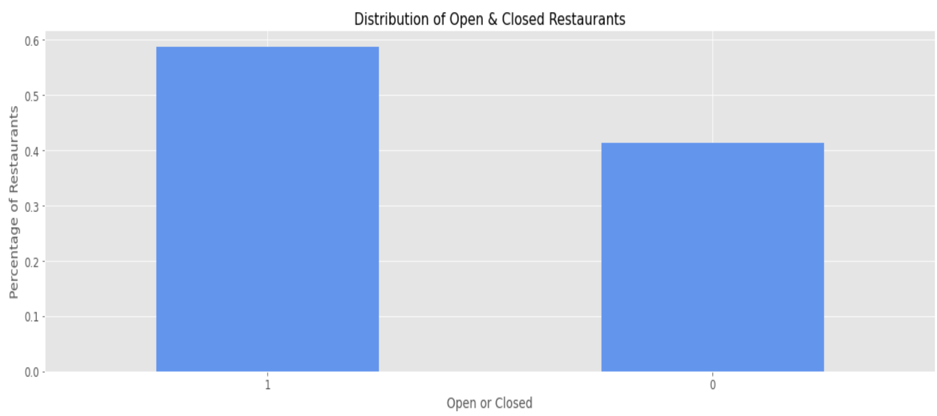
The restaurant industry in Canada generated close to $79 billion[1] in 2019 which accounted for 1.5%[1] of the country's GDP constituted by 97,000[1] businesses, which employ 6.4% (approximately 1.2 million people)[1] of the country's labor force in 2019. Given the significance of this industry, it's size and employment numbers, I decided to create a model that could help business owners and investors determine the essential features that predict a restaurant's success or failure. Based on this analysis, investors can decide whether they should invest at a particular business based on the likelihood that it is going to run out of business in the future or not; existing businesses could intervene and improve upon those parameters whereas new businesses could analyze the potential before entering the market.

Data Source and Preprocessing

In this project, I used the yelp dataset[2] which is publicly available for educational and academic purposes. This dataset contains information about 160,585 businesses across 8 Metropolitan areas for the years2005 to 2021. The whole dataset shared is 11 GB in size with 5 .json files about business, reviews, users, check-in and tips data. For this project, I have focused on restaurants business in Vancouver, BC. After filtering the data, I had 4,749 restaurants and 322,251 reviews related to these restaurants. Restaurant data has 14 features which have details about location, name, categories, attributes, hours of operation, review counts, stars and whether it is open or closed. Review dataset has 9 features with specific details about the review text, stars given by user, reactions received to review, and date of review. To map this dataset to business and user dataset, the common field available was business id and user id.

Feature engineering and EDA

Distribution of target variable (is_open) is 59% for open restaurants and 41% for closed restaurants.



Combined dataset had 22 features, but predictive ability using only original features was very low and may not have provided meaningful insights. Therefore, new features were added in the dataset. Some of the new features added were:

1. **Chain Restaurant:** During EDA of ‘name’ field, it was noted that there were 956 restaurants which were a part of 282 chains. Therefore, ‘name’ field was further analyzed and it was noted that 28% chain restaurants were closed whereas 45% of independent restaurants were closed since 2005.
2. **Number of Attributes:** There are total of 35 unique attributes that can be listed for a restaurant and every restaurant has listed different number of attributes.
3. **Top Attributes:** I had also extracted the top 7 attributes that were listed by majority of restaurants.
4. **Top categories:** Top 25 categories of restaurants (e.g., Japanese, cafes, etc.) were also extracted to check if any category has any impact on business success
5. **Repeat users:** ‘User_id’ field was used to users writing multiple reviews as repeat customers play a major role in a business’ revenue generation.
6. **Age:** This field is extracted from reviews dataset using review date.
7. **Sentiment Score:** I used Vader’s Sentiment Analysis compound score to study the sentiment behind reviews. I used compound score for this which is combination of positive (1), negative (-1) and neutral score (0).
8. **Review Length:** Negative reviews tend to be largely worded and therefore, this feature was used to determine if it helps in predicting a restaurant’s future.
9. **Density of restaurants:** Using Haversine formula and latitude/ longitude details, total of restaurants within 1 km radius were calculated for each restaurant. I also calculated the number of restaurants falling within the same category in its vicinity to further analyze if this impacts a restaurant’s success.
10. **Relative fields:** Relative values of different fields were extracted using z-score to see the position of a restaurant compared to mean performance of restaurants within its 1 km radius.

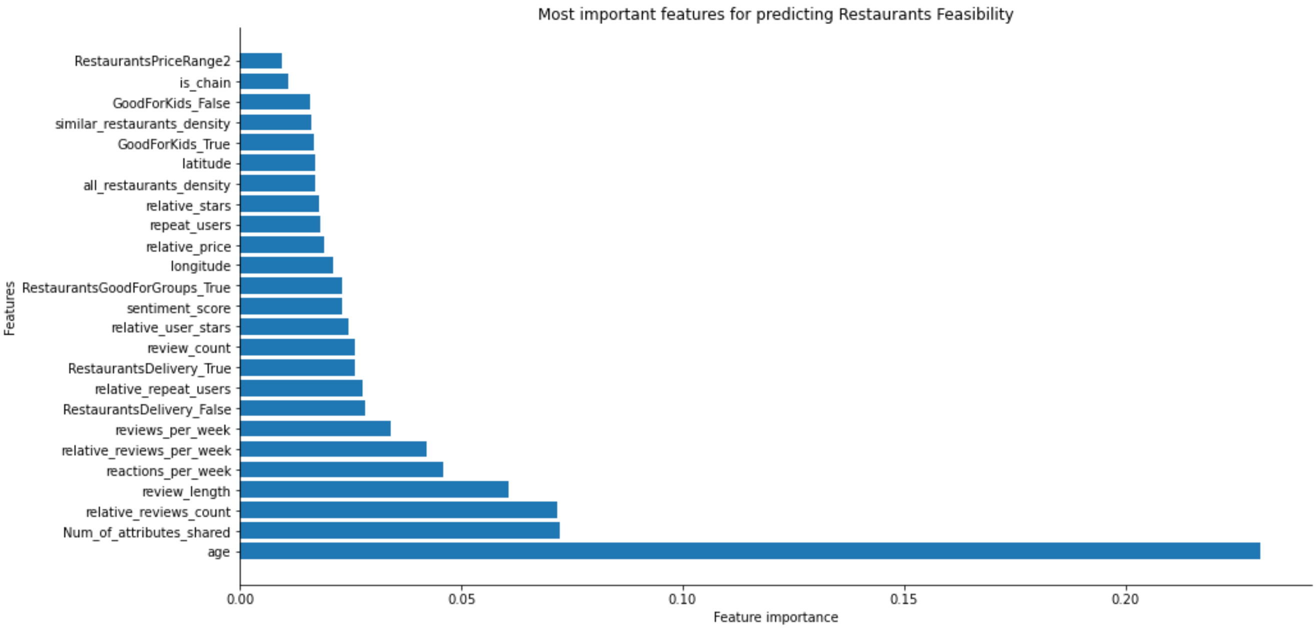
Modelling Summary

I evaluated 4 different supervised machine learning models on this dataset: Logistic Regression, Decision Tree, Random Forest and Gradient Boosting. Based on their test scores, precision, recall score and AUC, I found Logistic Regression and Gradient Boosting performed better than Decision Tree and Random Forest model. In order to select the better model between these two, I used GridSearchCV for hyperparameter tuning. Summary of the results are tabulated below:

Model	Test Score	AUC	Class	Precision Score	Recall Score
Logistic Regression	0.826	0.887	0	0.81	0.76
			1	0.84	0.87
Decision Tree	0.765	0.838	0	0.72	0.70
			1	0.79	0.81
Random Forest	0.813	0.871	0	0.85	0.67
			1	0.80	0.91
Gradient Boosting	0.839	0.913	0	0.84	0.76
			1	0.84	0.90
Gradient Boosting with hyperparameter Optimization	0.853	0.915	0	0.84	0.79
			1	0.86	0.89

Using Gradient Boosting model and optimal values for hyperparameters, test score improved by 1.4% and AUC by 0.2%. This model was able to predict 86% of open restaurants correctly and 84% of closed restaurants. Recall score improved radically by 3% for closed restaurants as compared to initial models.

Summary



Based on our analysis of the given dataset, key features for predicting a restaurant's feasibility are:

1. **Age**: Age of a restaurant plays a key in predicting a restaurant's life. Probability of a restaurant getting closed is low if it's been in business a long period of time as compared to new restaurants who are yet to establish their name and place in the market. The older the restaurant, the larger its customer base would be. More customers in turn generate larger revenues for the business and help the business live longer.
2. **Number of Attributes Shared**: Most successful restaurants share information about their business' attributes which help customers in making informed decision about whether to visit that outlet or not based on their requirements. Uninformed customer is more likely to get disappointed later if the customer does not get any particular service/feature they were looking.
3. **Review Count**: Number of reviews received convey the popularity of a restaurant and helps in its success in the digital world where new customers always check the reviews before visiting a new place. If a place is highly positively reviewed, a new customer is more likely to visit that restaurant.
4. **Review Length**: Number of words used in the reviews can help predicting the restaurant's feasibility as a customer may write about their exceptional experience or dissatisfaction in a well-written, strongly worded review. Emotions, positive or negative, are generally explained with reasons and hence, longer. This could strongly help in predicting the life/performance of a restaurant.
5. **Restaurant Delivery**: Whether a restaurant is providing delivery services or not is essential in today's times. Not everyone is able to dine-in during the weekdays. With the ongoing pandemic, people generally prefer to place an order for home delivery to avoid contact with other people. This also helps reduce the wait and travel time for them.
6. **Reviews per week**: This feature provides insight about how a restaurant is performing in terms of weekly footfall and weekly reviews written by its customers.
7. **Relative Reviews Count**: Review count is directly proportional to the footfall. This field provides information about how a restaurant is performing as compared to the mean value for restaurants in its vicinity.
8. **Reactions per week**: Reactions are votes received on reviews in terms of how other users found the review, useful, funny or cool. The more the reactions, larger gets the expected footfall.

9. **Relative Price**: Price range of a restaurant as compared to its neighboring restaurants can considerably impact the business. If the prices are set higher than the neighboring competing restaurants, it can have a negative impact on its business given the fact that similar quality and taste are offered by its competitors. Hence, a deeper analysis into the right price range may be required by the businesses for sustenance.
10. **Restaurant Density**: Number of restaurants within its vicinity can help in predicting a restaurant's success. As the number of restaurants increase, the customer base gets divided leading to a tight race for survival.
11. **Relative User Stars**: Stars received by a restaurant relative to places in its neighborhood is a strong metric. When in the area, a person definitely looks at its stars rating while deciding on a place to eat. Restaurants with higher ratings are generally preferred and this field can immensely help in measuring a restaurant's performance.

Next Steps

As a next step, I would like to analyze the restaurants opened/closed during a shorter duration of say 2 years or 5 years for focused insights as the world is fast paced and always changing.

Also, in the dataset available on Yelp, financial information such as yearly revenue, year-on-year growth and net profit was not available. These metrics could play a substantial role in predicting a restaurant’s feasibility and assist in the decision-making process.

I would also like to study the impact of public transportation and it’s close-by attraction spots to understand the geographic areas attracting footfall.

1. *https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2021025-eng.htm*
2. *https://www.yelp.com/dataset*