Business Closing Prediction based on Yelp Dataset

# Introduction

Over time, we observe a continuous evolvement in technological advancement. Researchers are more focus on the way to make people’s lives easier. The Internet carries a vast range of information resources and services that people use and depend on day-to-day. We saw in the meantime the proliferation of [website](https://en.wikipedia.org/wiki/Website) on which reviews can be posted about people, businesses, products, or services.

Yelp.com is known as one of the top review sites where consumers can share experiences about product quality and gather information about experience goods, which have quality that is observed only after consumption. Some articles revealed that online reviews have an impact on a business. Thus, businesses should care about them for preventing bad consequences or finding new business opportunities. On December 9, 2015, The New York Times titled in his Entrepreneurship section that: “A Bad Review Is Forever: How to Counter Online Complaints”.

However, can bad reviews lead to a business closure? Or what could be the main factors which impact closing business? The purpose of this project is to predict whether a business will remain open or going to close based on the Yelp dataset.

After importing the dataset from <<https://www.yelp.com/dataset>>, I will tidy and transform it to get only the data concerning businesses (restaurants and similar food businesses), reviews and users from Toronto. Then, I will generate some visualizations to raise patterns, show findings, and build models to answer the above questions.

# Literature Review

Do online consumer reviews affect restaurant demand [1]? Michael Luca from Harvard University investigated that question in 2016 by combining two datasets from Yelp.com and the Washington State Department of Revenue. It has been finding that variation in star rating can lead to an increase or a decrease in revenue. That is mostly visible by independent restaurants but not for restaurants with chain affiliation. Changes in quality retain more the attention of consumers and the users respond more when a review contains more information. Most of time, users rely more to rating when a restaurant got an important number of reviews compare to the size of reviewers.

Yelp reviews and ratings are important source of information to make informed decisions about a venue. After inspected hundreds of reviews, five dimensions have been finding about restaurant businesses. The research has been done by analyzing business reviews and building a classifier to describe different type of businesses in the restaurant industry. Those categories are “Food”, “Service”, “Ambience”, “Deals/Discounts” and “Worthiness” [2].

Using a deep learning method to analyze photos and reviews of restaurant posted on Yelp.com from December 2004 to December 2015, it has been found that the volume and the valence of photos are strong predictors of restaurant survival. When it comes to reviews only the valence matters [3], and a restaurant closure is strongly associated with a consumer sentiment.

# Dataset

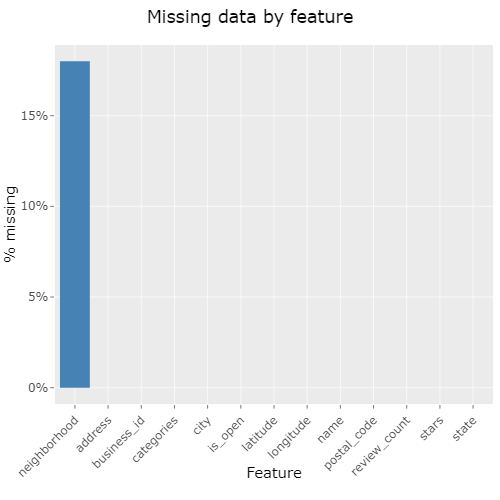
The dataset has been imported from <<https://www.yelp.com/dataset>>, it is JSON dataset which contains information for 188,593 businesses. This dataset has 5,996,996 reviews, 1,518,169 users, 280,992 photos, 157,075 check-ins, 1,185,348 tips for these businesses.

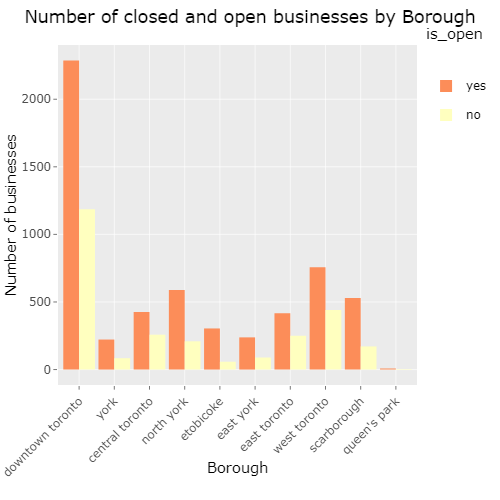
For the project, a subset of this dataset will be used by filtering the raw sets of data to get information related to businesses in the city of Toronto. This reduced the number of business to 18,233 with 474,803 reviews from 103,262 users. The goal will be to predict whether a business is closed or open from the attribute “is\_open” in the business data.

The focus will be on restaurant and similar services. That led to a data set with 8,559 observations and 13 variables. The variables are:

* address: character class representing the full address of the business
* business\_id: character class representing an identification for each business
* categories: character class containing a set of categories a business belongs to
* city: character class representing the city where a business is located
* **is\_open (target variable)**: integer class with 2 unique values (0 or 1) indicating if a business is closed or not
* latitude: numeric class, geographic coordinate
* longitude: numeric class, geographic coordinate
* name: character class, the business’s name
* neighborhood: character class, the neighborhood where the business is located
* postal\_code: character class, the business’s postal code
* review\_count: integer class representing the number of reviews received by a business
* stars: numeric class, the business’s star rating
* state: character class representing the state code

**Initial Analysis Plotting**





From the graph above, it could be gathered that "closed" businesses received less reviews, and that does not relate to rating. “Closed” and “Open” businesses have the same distribution of rating.

Some additional information will be retrieve from the reviews data set to get more insights about each business. For that purpose, feature engineering will be applied to specific attributes like date of a review, the stars obtained by review and its text content.

# Approach

**Import the dataset**

**Tidy and transform**

**Visualize**

**Modelling**

**Conclusions**

## Step 1: Data loading

In this step, I first subscribed to the Yelp challenge round 12 at <<https://www.yelp.com/dataset/challenge>> to get the permission to download the dataset. After downloaded the data sets and stored in a directory, I imported into RStudio.

## Step 2: Data wrangling and Initial feature engineering

This is the part where I tidied and transformed the data to get in a form that’s natural to work. Firstly, as it was a zip file containing 6 JSON files I ran a script to unzip it and then convert the JSON formats to R objects.

<<https://github.com/PatKakou/capstone-project/blob/master/Yelp_Dataset_JSON_to_RDS.Rmd>>

Secondly, I made some sub-setting to get Toronto’s observations from the business data set and the review data set.

<<https://github.com/PatKakou/capstone-project/blob/master/Yelp_Toronto.Rmd>>

Then I checked which variables had the most missing values. Only the "neighborhood' variable had missing value, and it was about 18%. To fill that missing values, I create a new feature to get the borough instead of neighborhood by joining the business data to a Toronto FSA data. I got that data from Wikipedia at the following link: <<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>>.

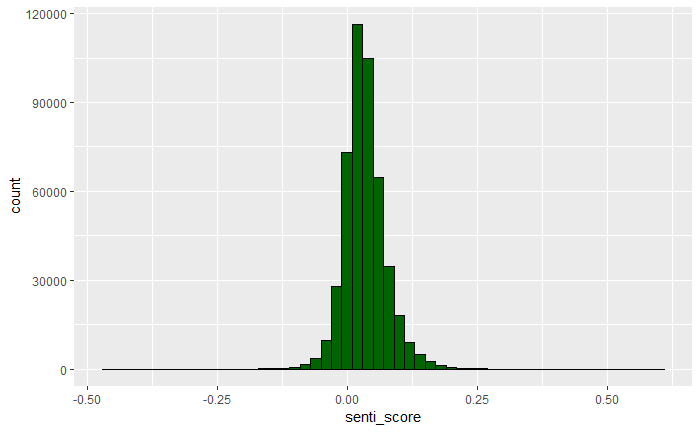
The purpose was to get a categorical feature with less levels and to eliminate missing values. The new feature “Borough” I got, has 10 levels with no missing values compare to 80 levels for the previous “neighborhood” feature.

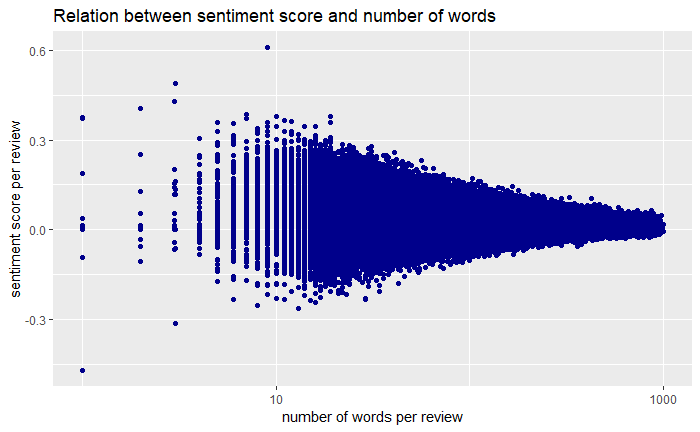
Finally, I applied a Natural Language Processing (NLP) techniques to extract keywords on reviews then assign a sentiment score to each review scaled between -1 and 1

< <https://github.com/PatKakou/capstone-project/blob/master/Yelp_reviews_sentiment.ipynb>>.

The sentiment scores generated from the reviews has been done with Python language using the NLTK package with the SentWordNet version 3.0. The SentiWordNet is a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. The purpose here was to gather some important insights from the review data set to create new variables and added them to the business data set.

<<https://github.com/PatKakou/capstone-project/blob/master/reviews_EDA.Rmd>>





## Step 3: Exploratory Data Analysis and more Feature Engineering

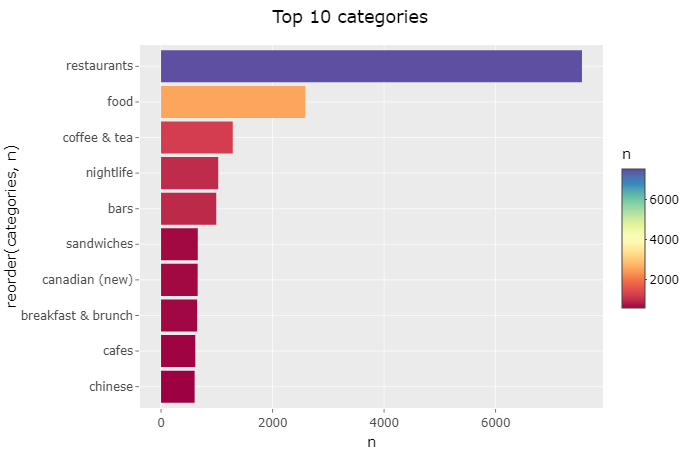
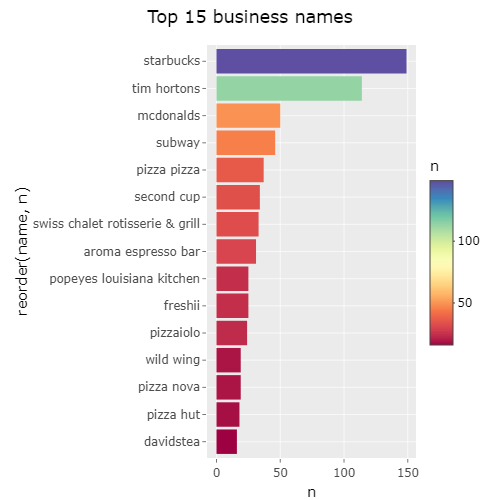
A good visualization will show you things that you did not expect or raise new questions about the data [9]. In that third step, I did some generate leads that I explored in depth later. I iteratively generated questions about the data, search for answers by visualizing and transforming the data, used what I learned to refine my questions and/or generate new questions.

My goal was to develop a good understanding of that data set.

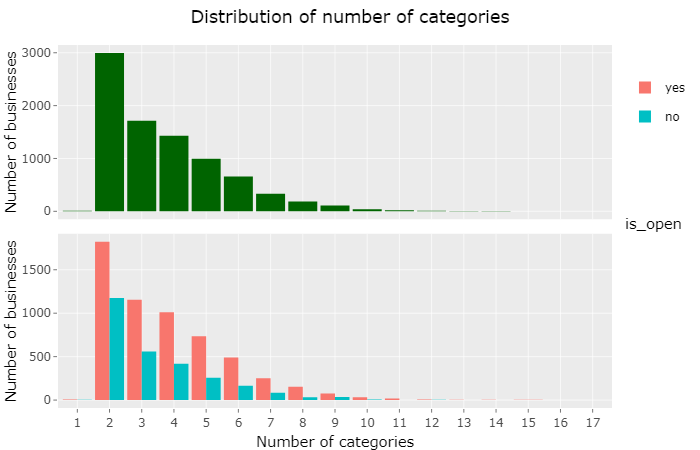
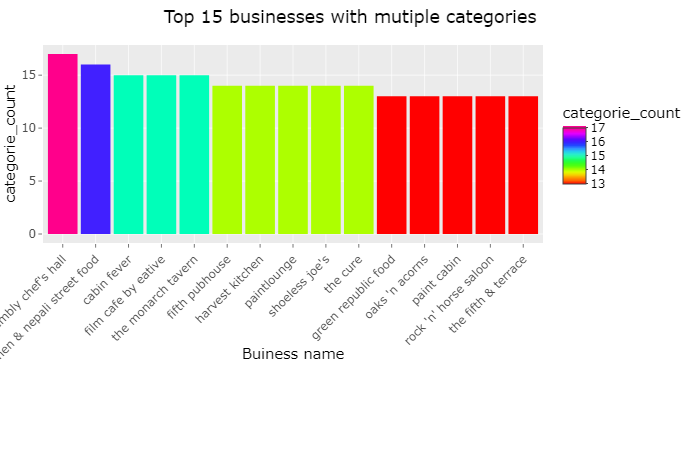
< <https://github.com/PatKakou/capstone-project/blob/master/reviews_EDA.Rmd>>

It is very important to generate meaningful features because sometimes there are keys in building a model. I was asking my self some question and tries to answer some of them by creating new feature and visualize how they behave with the other features in the data set.

How many categories a business can have? As a business is linked to multiple categories in our dataset, I flattened the categories and created a new feature which has the information about the number of categories each business has.



Is the business part of a **chain**? From the graph above, I concluded that it was important to add another feature to classify each business as part of a chain or not. I decided to choose 2 as my threshold that means if there are more than 2 businesses with the same name, they are considered part of a chain.



## Step 4: Model building

The purpose of this step is to create classification model to predict whether a business will survive. I first used the initial data without feature engineering to train a logistic regression and a random forest model. And then I used the final data with the new features created after feature engineering. For that final data set I trained three more models: Support Vector Machine (SVM), Generalized Boosted Regression Models (GBM) and Naïve Bayes.

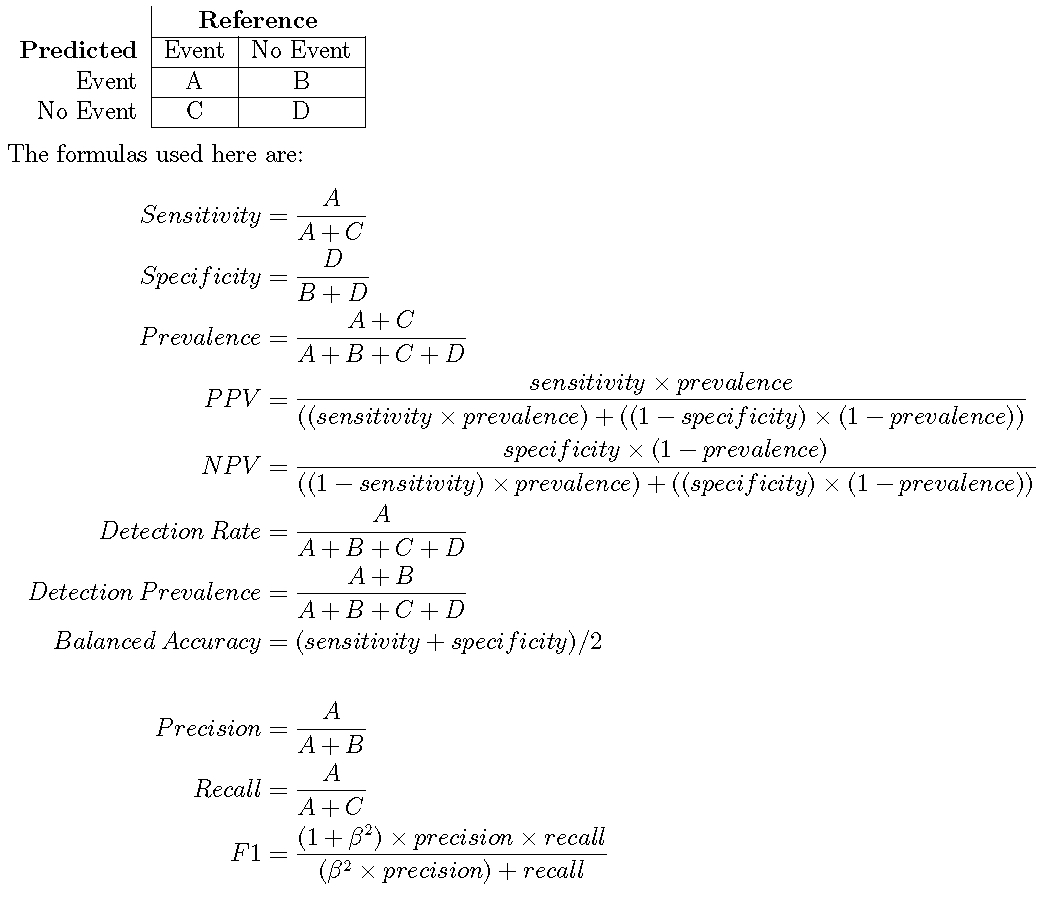
Before building and training the different models, each datasets has been split in 80% for the training set and 20% for the testing set using stratified sampling. Then, I applied a cross validation on the training set with 5 folds.

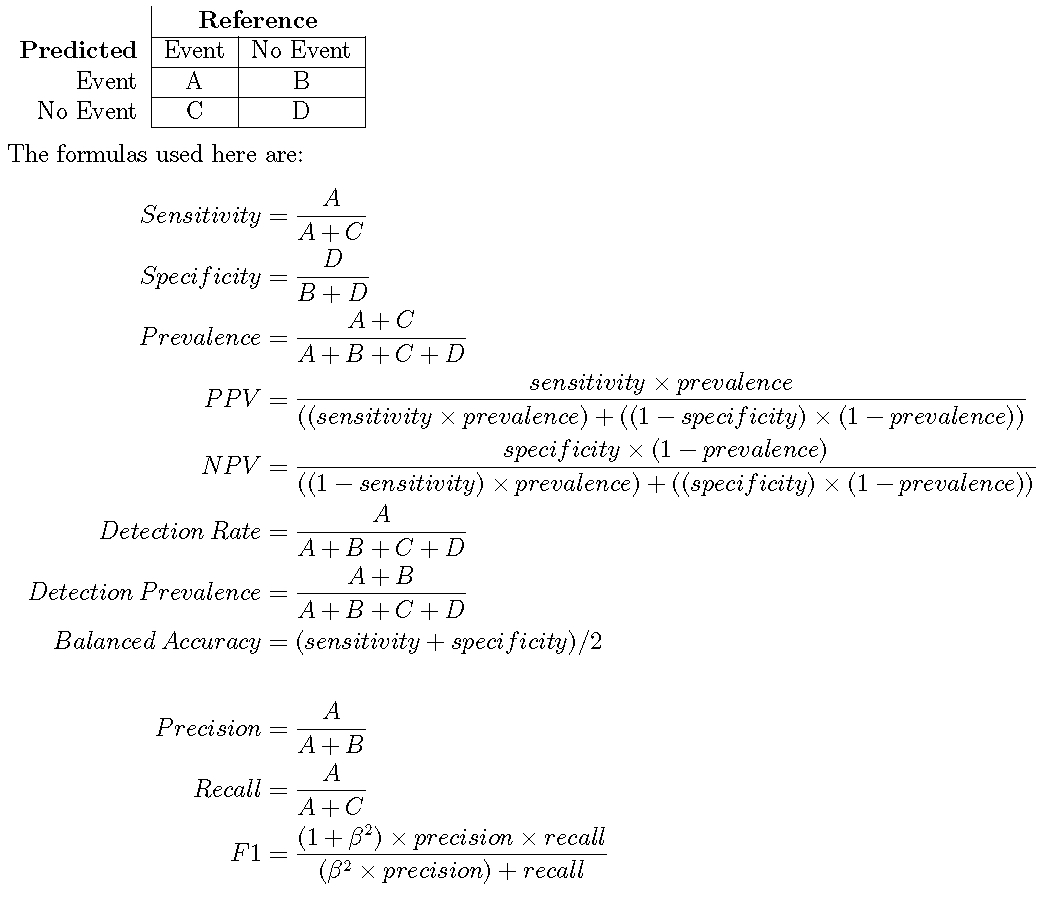
## **Initial Results**

I got the result below:

|  |  |
| --- | --- |
| **Logistic Regression** | **Random Forest** |
| **Confusion Matrix and Statistics**  **Reference**  **Prediction yes no**  **yes 1732 822**  **no 0 3**    **Accuracy : 0.6785**    **Sensitivity : 1.000000**  **Specificity : 0.003636**    **'Positive' Class : yes** | **Confusion Matrix and Statistics**  **Reference**  **Prediction yes no**  **yes 1414 559**  **no 318 266**    **Accuracy : 0.657**    **Sensitivity : 0.8164**  **Specificity : 0.3224**  **'Positive' Class : yes** |

**Performance measurement**





Looking at the confusion matrix above, the predictive ability of the models is very poor in the case of closed restaurants. The precision of closed businesses can be further improved but there is always a trade-off with the precision of open businesses. I decided to focus my attention on improving the specificity or recall of closed restaurants. I also decided to choose the Area under the Curve (AUC) of the Receiver Operating Characteristic (ROC) as a metric to compare the different models.

## Step 5: Conclusions

Finally, the different inferences gather from the previous steps will be share here and some solutions will be discussed for the next steps.

I decided to choose the Area under the Curve (AUC) of the Receiver Operating Characteristic (ROC) as a metric to compare the different models.

# Results

Explain your results here. Consider that you need to communicate your results to executives in an organization. For example:

1. Insert tables and/or charts showing the results
2. Write description of the tables and charts, such that they show the usefulness for an organization
3. Identify the evaluation measures, such as accuracy, precision, recall, etc.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **glmnet** | | **rf** | | **svm** | | **gbm** | | **nb** | |
| Classes | open | closed | open | closed | open | closed | open | closed | open | closed |
| Precision |  |  |  |  |  |  |  |  |  |  |
| Recall |  |  |  |  |  |  |  |  |  |  |
| F1 score |  |  |  |  |  |  |  |  |  |  |

# Conclusions

Give a short summary (one to two paragraphs) of your analysis and conclude the discussion by defining the usefulness of your analysis.

**Chain**

We also find that chain and mainstream restaurants have a greater chance of survival

than independent and niche restaurants

**Closed date**

To identify a restaurant’s close time, we wrote a computer program using Python. For

each closed restaurant, we aimed to find the earliest review that mentioned its closure. We then

used the date of that review to approximate the close date of this restaurant.

To identify the earliest review mentioning the closure of a restaurant, we used keywords match.

read all reviews of 200 randomly chosen closed restaurants. The research assistant was instructed to identify words and phrases representing permanent closure status of a restaurant.

This dictionary of keywords was then used to identify close dates for the 3,711 closed restaurants in our data set.If a closed restaurant had no reviews mentioning its closure, we used the last review date of the restaurant to approximate its closure date

**Time-Variant Hazard Model**

To examine the relationships between photos, reviews, and restaurant characteristics on

restaurant survival, we employed a discrete-time proportional hazard model (Cox 1972; Fahrmeir

and Tutz 1994; Kalbfleisch and Prentice 1980, 2002; Kiefer 1988; Lunde et al.1999).

In Eq. (1), 𝑖 is the index of the restaurant, and 𝑡 is the index for 𝑡th month. The hazard

function ℎ& 𝑡 is the likelihood of restaurant closure for restaurant 𝑖 at time 𝑡 conditional on the

restaurant remaining open in the previous period. ℎ\* 𝑡 is the baseline hazard, which reflects the

likelihood of closure at time 𝑡 holding all covariates (i.e., the *X* and control variables in Eq. 1) at

zero. Covariates shift the baseline up or down so that different restaurants have a different hazard

rate of survival at time t. Since our dependent variable is the hazard rate of restaurant closure, a

negative sign of coefficients reflects a positive correlation with survival chances.

**Suggest Improvement**

The results of this model are very promising, and they indicate a significant improvement for lending purposes relative to a random model. The key for further improvement, in my opinion, is adding more features, possibly through utilizing different data sources.

* One possible reason for a restaurant closure is health inspection ratings. Adding health inspection ratings as a feature in our model could increase its precision.
* Another reason for restaurant closure is high rent charges. Adding rent pricing per region could help explain more restaurant closures.
* A change in population demographics in certain areas of a city can increase or decrease traffic to some restaurants.
* New surrounding venues are another reason that can drive traffic to restaurants and lead to success that cannot be predicted from this model in its current form.
* Success of a restaurant is currently defined as the restaurant remaining open. A more accurate definition of success that would be more appropriate for lending purposes would be correlated to restaurant revenue. Even though the revenue of most restaurants is not public information, relevant metrics can be constructed. For instance, multiplying the number of weekly comments received by a restaurant with the price (i.e. general dining cost) of the restaurant can act as a useful metric.

Summary sample

* This model was built for restaurant lending purposes and identifies restaurants that remain open in a 4-year period with a precision of 91%.
* The dataset was built by pulling recent information about restaurants that used to exist in 2013 in Phoenix, AZ through the Yelp and Google Search APIs.
* Some very predictive features of this model were built using Yelp review and location metadata. This helped to construct relative metrics like restaurant density and quantities that are relative to surrounding restaurants.
* The machine learning model used was a simple logistic regression model, which was optimized for precision of open restaurants using grid search with cross-validation.
* One lesson learned is that the most important factor that defines whether a restaurant will remain open is whether it is part of a chain. Restaurants that belong to chains close less frequently.
* Another lesson learned is that building a restaurant in an area with a lot of other restaurants is generally negative, except if those restaurants offer similar food (e.g. building a Chinese restaurant in China Town).
* This model can be improved with the incorporation of further datasets such as health inspection data (not publicly available for Phoenix, AZ at the moment), and information about surrounding venues.

**References**

[1] Luca, Michael, Reviews, Reputation, and Revenue: The Case of Yelp.Com (March 15, 2016). Harvard Business School NOM Unit Working Paper No. 12-016. Available at SSRN: <https://ssrn.com/abstract=1928601> or [http://dx.doi.org/10.2139/ssrn.1928601](https://dx.doi.org/10.2139/ssrn.1928601)

[2] <https://www.ics.uci.edu/~vpsaini/files/technical_report.pdf>

[3] Zhang, Mengxia and Luo, Lan, Can User Generated Content Predict Restaurant Survival: Deep Learning of Yelp Photos and Reviews (March 2018). Available at SSRN: <https://ssrn.com/abstract=3108288> or [http://dx.doi.org/10.2139/ssrn.3108288](https://dx.doi.org/10.2139/ssrn.3108288)

[4] Luca, Michael and Zervas, Georgios, Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud (May 1, 2015). Harvard Business School NOM Unit Working Paper No. 14-006. Available at SSRN: <https://ssrn.com/abstract=2293164> or [http://dx.doi.org/10.2139/ssrn.2293164](https://dx.doi.org/10.2139/ssrn.2293164)

[5] O'Hern, Matt S. and Rindfleisch, Aric and Schweidel, David A. and Antia, Kersi, The Impact of User-Generated Content On Product Innovation (March 3, 2011). Available at SSRN: <https://ssrn.com/abstract=1843250> or [http://dx.doi.org/10.2139/ssrn.1843250](https://dx.doi.org/10.2139/ssrn.1843250)

[6] McQuarrie, Edward F. and McIntyre, Shelby H. and Shanmugam, Ravi, What Motivates Consumers to Produce Online Reviews? Solidarity, Status, and the Soapbox Effect (February 1, 2013). Available at SSRN: <https://ssrn.com/abstract=2210707> or [http://dx.doi.org/10.2139/ssrn.2210707](https://dx.doi.org/10.2139/ssrn.2210707)

[7] [https://en.wikipedia.org/wiki/Review\_site#Impact](https://en.wikipedia.org/wiki/Review_site%23Impact)

[8] The New York times, A Bad Review Is Forever: How to Counter Online Complaints

<https://www.nytimes.com/2015/12/10/business/smallbusiness/small-business-counter-bad-reviews.html>

[9] Garrett, Grolemund and Hadley Wickham*,* R for Data Science (January 5, 2017). <http://r4ds.had.co.nz/introduction.html>

[10] <https://towardsdatascience.com/using-yelp-data-to-predict-restaurant-closure-8aafa4f72ad6>

[11] <https://www.kaggle.com/jessicali9530/best-las-vegas-restaurants-eda>

# Readme sample In GitHub

# Project: Restaurant Success Model

## A model that predicts if a restaurant is likely to close within the next 4 years.

### Project Description

The goal of this project was to built a model that can predict whether a restaurant is likely to close within a 4-year period. This information would be useful to restaurant lenders (such as banks) and investors.

To achieve the above goal I started by building a relevant dataset which I then used to fit a logistic regression algorithm that can separate between restaurants that are likely to close and restaurants that are likely to remain open. Building a meaningful dataset involved several steps that are breaken down below:

1. Find an initial list of restaurants that used to exist in the [past](https://www.kaggle.com/c/yelp-recsys-2013).
2. Find information about the current state of the restaurants from the past. a) Pull information from the Yelp Search API. b) Check if the information is correct (restaurant names and addresses have changed in 4 years). c) Find additional information about non-matched restaurants using the Google Search API. d) Pull new information from Yelp Business API using the web addresses found through Google. e) Match information from old and new datasets to make sure it is correct.
3. Engineer relevant features.
4. Test different Machine Learning models and optimize parameters based on above use case.

The work described in the above steps is split in five Jupyter notebooks.

### Code

All code is presented in a series of notebooks showing the steps followed for each of the individual processes and the resulting numbers and graphs.

This repository includes the code to pull data from the Yelp Search API, based on the names and addresses of the restaurants found on a [Yelp Kaggle Dataset](https://www.kaggle.com/c/yelp-recsys-2013) released in 2013.

The data from the two sources (Yelp Search API and Kaggle) are matched to guarantee the consistency between the two sets and only 65% of the entries had a matching address and name. The Google Custom Search API is used for the unmatched data to find the right restaurant urls ([file: Restaurants\_yelp\_GoogleCustomSearchAPI.ipynb](https://github.com/alifier/Restaurant_success_model/blob/master/Restaurants_yelp_GoogleCustomSearchAPI.ipynb)) and pull the remaining data using the Yelp Business API ([file: Restaurants\_yelp\_API\_for\_GoogleSearchResults.ipynb](https://github.com/alifier/Restaurant_success_model/blob/master/Restaurants_yelp_API_for_GoogleSearchResults.ipynb)).

The two datasets are then matched together with higher success rate ([file: Restaurants\_yelp\_join\_all.ipynb](https://github.com/alifier/Restaurant_success_model/blob/master/Restaurants_yelp_join_all.ipynb)) and the resulting dataframe is used to generate additional features ([file: Restaurants\_yelp\_more\_features\_final.ipynb](https://github.com/alifier/Restaurant_success_model/blob/master/Restaurants_yelp_more_features_final.ipynb)).

Different ML algorithms are fitted using the above features Logistic Regression and Gradient Boosting being the most successful ([file: Restaurants\_yelp\_ML\_final.ipynb](https://github.com/alifier/Restaurant_success_model/blob/master/Restaurants_yelp_ML_final.ipynb)).

For a more detailed discussion of the process followed to develop this model and the obtained results, see the relevant [blog post](https://medium.com/@alifier/using-yelp-data-to-predict-restaurant-closure-8aafa4f72ad6).

K-means clustering

1. City Tagging - In order to subset the data to inspect Edinburgh the dataset needed to be tagged with which city each record belonged to. In order to do this k-means clustering was run using the latitudes & longitudes of restaurants and the latitudes & longitudes of the centres of the cities as initial cluster loci;

Then data enrichment was performed - creating new fields that might be of interest:

1. Distance from Cluster centre - The (euclidian) distance from the final cluster centres was calculated for each business;
2. Tag Chains - enrich the dataset by adding a field that tags a restaurant as a chain (if there are more than 2 businesses with the same name they are considered part of a chain);
3. Analyse the number of clusters of restaurants - Using scree plot;
4. Tag Size of cluster - enrich the dataset by adding a field that tags a restaurant with the number of restaurants belonging to the same cluster;