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

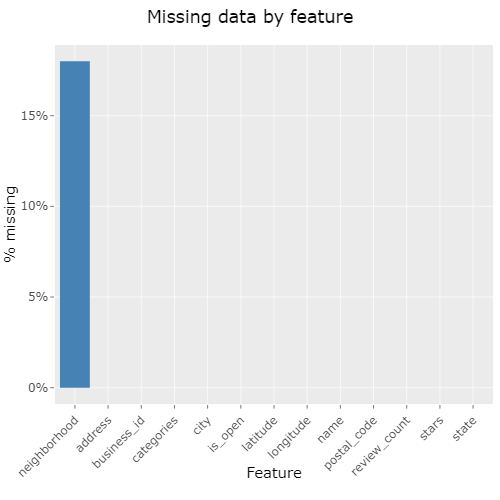
The dataset has been imported from <<https://www.yelp.com/dataset>>, it is JSON dataset which contains information for 188,593 businesses. This dataset has 5,996,996 reviews, 1,518,169 users, 280,992 photos, 157,075 check-ins, 1,185,348 tips for these businesses.

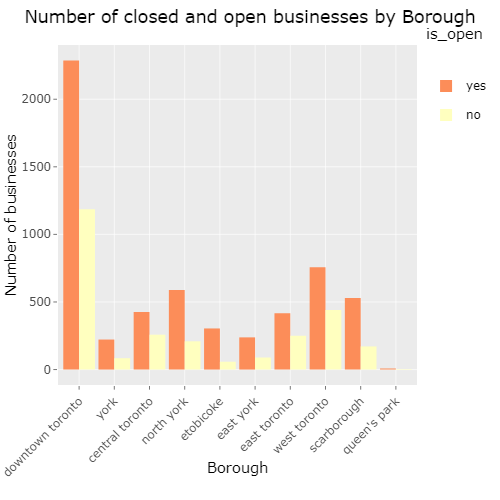
For the project, a subset of this dataset will be used by filtering the raw sets of data to get information related to businesses in the city of Toronto. This reduced the number of business to 18,233 with 474,803 reviews from 103,262 users. The goal will be to predict whether a business is closed or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

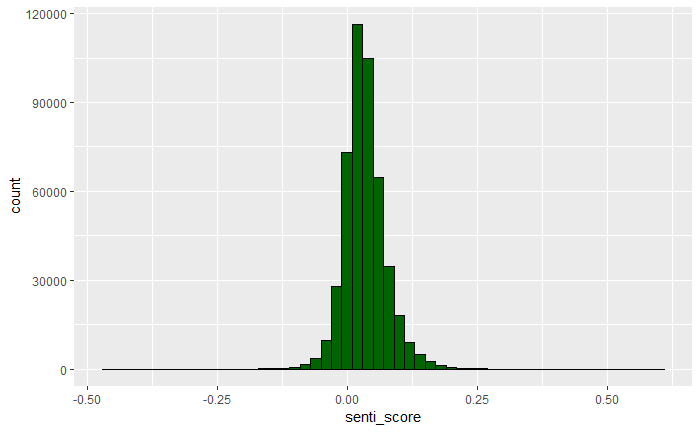
The purpose was to get a categorical feature with less levels and to eliminate missing values. The new feature “Borough” I got, has 10 levels with no missing values compare to 80 levels for the previous “neighborhood” feature.

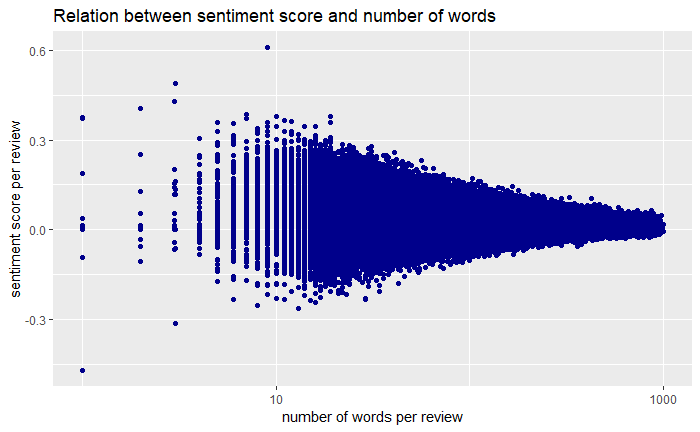
Finally, I applied a Natural Language Processing (NLP) techniques to extract keywords on reviews then assign a sentiment score to each review scaled between -1 and 1

< <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

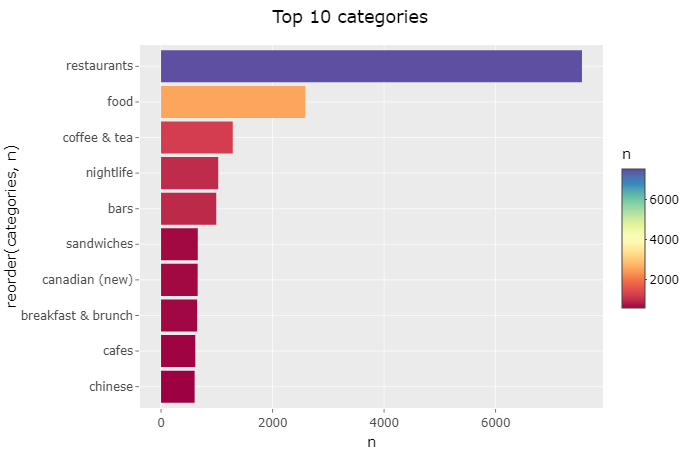
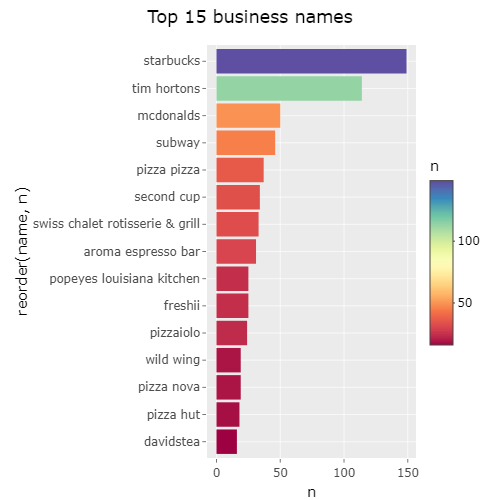
A good visualization will show you things that you did not expect or raise new questions about the data [9]. In that third step, I did some generate leads that I explored in depth later. I iteratively generated questions about the data, search for answers by visualizing and transforming the data, used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

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

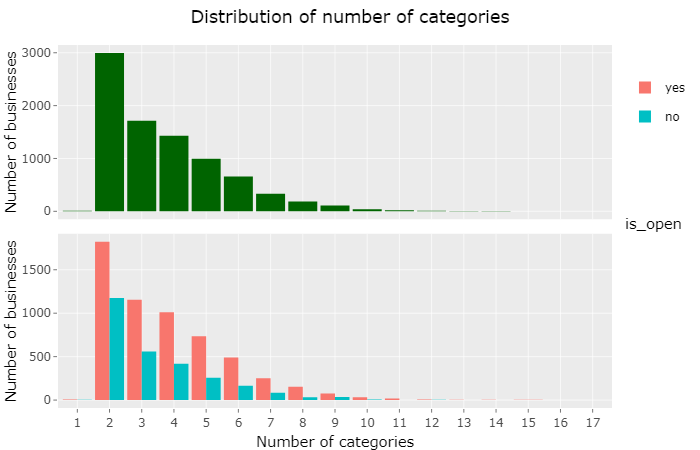
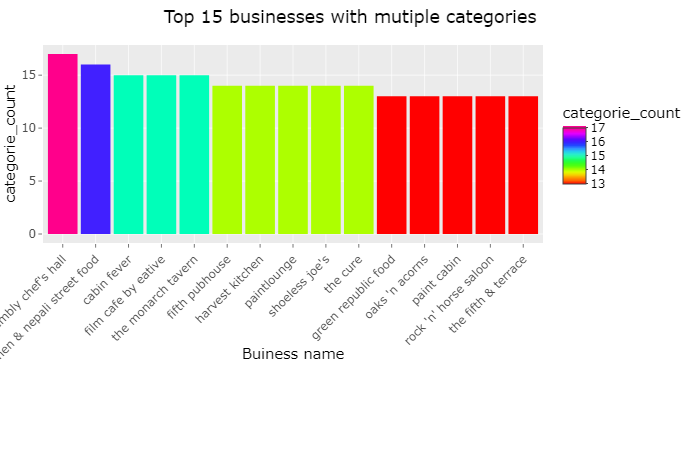
< <https://github.com/PatKakou/capstone-project/blob/master/reviews_EDA.Rmd>>

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

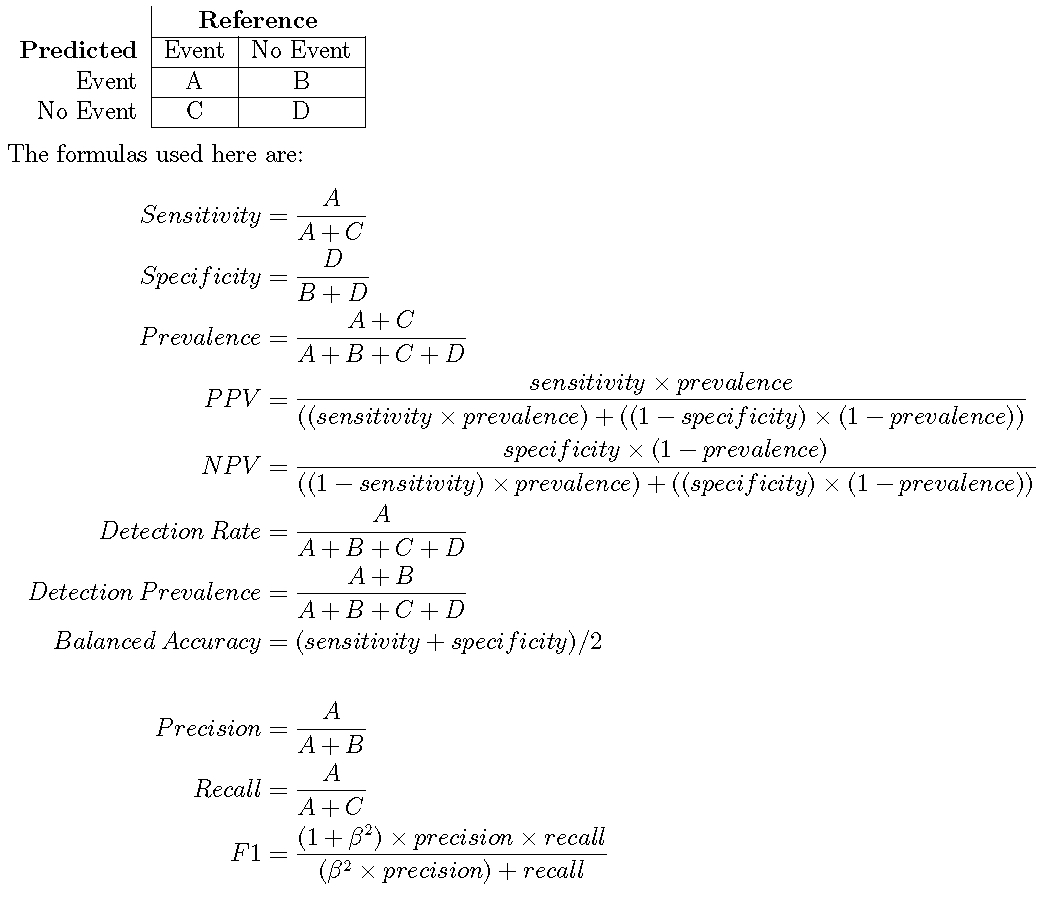
The purpose of this step is to create classification model to predict whether a business will survive. I first used the initial data without feature engineering to train a logistic regression and a random forest model. And then I used the final data with the new features created after feature engineering.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

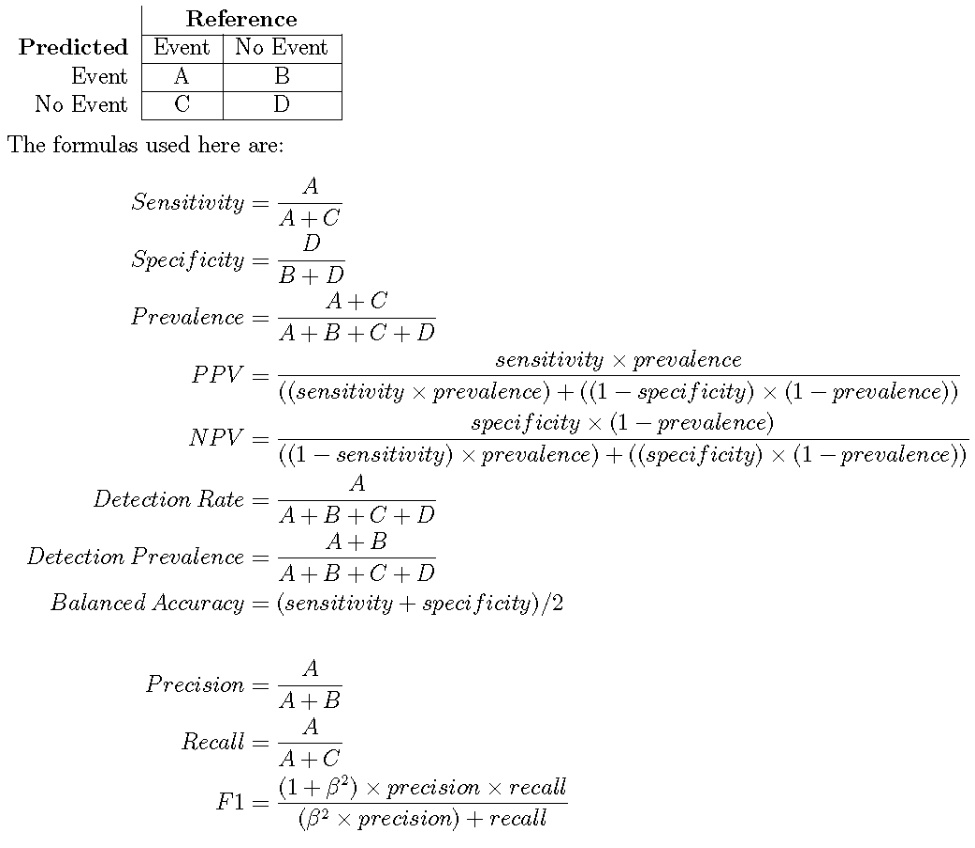
Before building and training the different models, each dataset has been split in 80% for the training set and 20% for the testing set using stratified sampling. Then, I applied a cross validation on the training set with 5 folds.

## **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:**

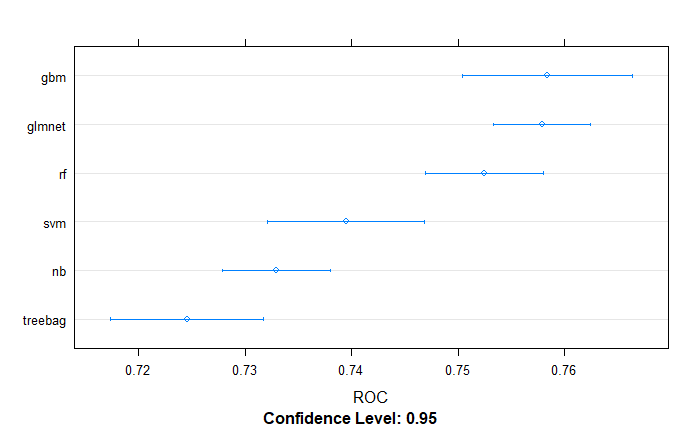


Looking at the confusion matrix above, the predictive ability of the models is very poor in the case of closed restaurants. The precision of closed businesses can be further improved but there is always a trade-off with the precision of open businesses. I decided to focus my attention on improving the 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. The AUC will measure how well a parameter can distinguish between two diagnostic groups (open/closed).

## Step 5: Conclusions

To improve my results, I did some feature engineering and trained the new data with three more models as I mentioned previously. I used another metric to compare the different models. That metric is the Area under the Curve (AUC) of the Receiver Operating Characteristic (ROC).

# Results

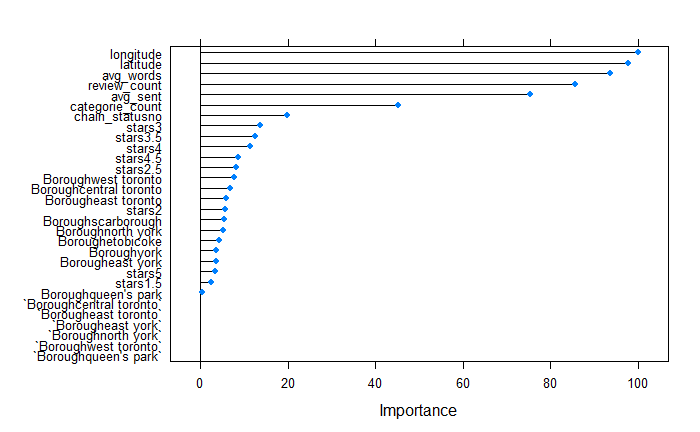
**AUC variation per model**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **glmnet** | | **rf** | | **svm** | | **gbm** | | **nb** | |
| **Accuracy** | 0.7296 | | 0.7378 | | 0.7249 | | 0.7296 | | 0.7167 | |
| **Classes** | **open** | **closed** | **open** | **closed** | **open** | **closed** | **open** | **closed** | **open** | **closed** |
| **Precision** | **0.7707** | **0.6052** | **0.7715** | **0.6284** | **0.7575** | **0.6086** | **0.7877** | **0.5892** | **0.7228** | **0.67005** |
| **Recall** | **0.8554** | **0.4655** | **0.8710** | **0.4582** | **0.8736** | **0.4127** | **0.8225** | **0.5345** | **0.9437** | **0.24000** |
| **F1 score** | **0.8108** | **0.5262** | **0.8182** | **0.5300** | **0.8114** | **0.4919** | **0.8047** | **0.5605** | **0.8186** | **0.35341** |

**Evaluation measures per models**

The graph represents the plotting of the AUC variation per model. The best model will be the one with the highest median and a low variance. Thus, the glmnet (logistic regression) model is the best model.

From the table, the random forest, glmnet and gbm models are the best models. And all the models have well predicted open businesses compare to closed business. As our goal is to predict if a business will be closed or not, we decided to prioritize the glmnet model for our final model.



**Feature Importance from gmlnet model**

* This model was built for a restaurant warning purposes and it identified restaurants that have a risk of closure based on a data in a 11-year period with a precision of 60.52%.
* Different machine learning models have been used but we decided to choose a logistic regression model by fitting a generalized linear model via penalized maximum likelihood (glmnet).
* This model can be improved with the incorporation of further datasets such as health inspection data and information about surrounding venues.

# Conclusions

The key for further improvement, in my opinion will be adding more features possibly through utilizing different data sources.

One could be by running a K-means clustering with the latitude and longitude of businesses and the latitudes & longitudes of the centres of the borough as initial cluster. Then create new features that could be interesting for prediction. One feature could be tagging a business with the number of businesses belonging the same cluster to construct relative metrics like business density.

Through the year health inspections evaluate if a restaurant setting is healthy or not, and that is sometime one reason for a restaurant closure. Definitely, adding health inspection ratings as a feature in our model could increase its performance.

Toronto is reputed to have a high rent charges and that varies from one borough to another. In our point of view, adding the rent pricing per borough could help explain more restaurant closures too.

**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>