Importance of social influence of Yelp reviewers in predicting the failure of a restaurant

Maurizio Pinto - maurizio.pinto@gmail.com

11 November 2015

Introduction

According to [3] and [4], the restaurant failure rate in U.S. is 30% during the first year of operation: this implies a potential loss of $5.20 billion in the form of lost restaurant revenues to the national economy. In addition, restaurant failures also lead to the loss of nearly 40.000 jobs per year. Factors that contribute to restaurant failures can be divided into two major types: *macro factors* (economy, legislation, new competition, etc.) and *micro factors* (capital, location, entrepreneurial incompetence, experience, leadership, etc.)

In recent years, other micro factors arose that can impact the business of a restaurant: one of those are crowd-sourced local business review websites, with social networking features. In this paper, we use the Yelp Challenge Dataset round 6 and we focus on measuring the importance of the social influence of Yelp reviewers in determining the success or failure of a restaurant. The analysis consists of three main steps: pre-processing of Yelp datasets, definition of a normalized social influence score, recursive feature elimination via *caret*.

More precisely, we would like to provide an answer to the following question: in a model that predicts whether a restaurant would close within the next three months, is the social influence of reviewers (a score based on number of friends, number of fans, number of votes, number of compliments) an important feature? In other words: will the features related to social influence rank amongst the top 10 features, when recursive feature selection is applied?

Methods

Data preparation - The goal of the data preparation stage is to obtain one single tidy dataset suitable to be used for predictive modeling. This tidy dataset shall contain aggregated and normalized information about the restaurants, the social influence of the users, and the sentiment scores for reviews and tips.

Business_data - After having selected only the restaurants in the *business_data* dataset, we applied the following transformation workflow:

- all the restaurants without reviews have been deleted
- a simple *k-means* clustering algorithm on *lat* and *long* normalized the *city* variable (10 main cities only)
- all the attributes that were not related to restaurants have been deleted
- all the features with more than 60% of missing values (NA) have been deleted
- the opening hours features, *categories*, *full_address*, *neighborhoods*, and *type* features have been deleted
- the *lat* and *long* features have been replaced with the distance in meters from the center of the city (Haversine formula)
- the following attributes have been normalized/simplified:
  - *attributes.Alcohol*, *attributes.Attire*, *attributes.Wi.Fi* have been transformed to logical attributes
  - *attributes.Noise.Level* levels have been re-mapped from "*quiet*", "*average*", "*loud*", "*very_loud*" to *0*, *1*, *2*, and *3* respectively

At the end of this stage, the *business_data* dataset contains *21.799* observations of *46* variables.

Note on restaurant's lifespan - The Yelp Dataset does not contain information about the opening date (and closing date, if applicable) of a restaurant. Since previous studies (see [3] and [4]) state that the probability that a restaurant closes during the first three years is very high, we think that the lifespan of a restaurant is an important feature in a predictive model. Therefore we calculated an approximation of the lifespan of a restaurant as follows:

$$\textbf{lifespan} = \text{date of the last review - date of the first review } (\textit{measured in days})$$

Even though this is not the real lifespan of a restaurant (users could have started reviewing the restaurant months after its opening), we assume this is a good approximation for the purposes of this paper.

Review_data and tip_data - The *review_data* and *tip_data* datasets have been processed both in the same way: a simple sentiment score has been calculated for the reviews and tips text, using the approach described by Jeffrey Breen [5] and using the Hu and Liu sentiment lexicon [6]. This stage of the processing could have been improved by adding context specific terms to the lexicon, for instance positive or negative terms commonly used in restaurants' reviews.

An important note: since the goal of this paper is to implement a model that predicts the closure of a restaurant within a horizon of three months, all the reviews and tips that have less than 3 months of age have been disregarded. The age of a review is calculated as follows:

$$\textbf{age of review X for business Y} = \text{date of the last review for business Y - date of review X}$$

Checkin_data - Even though the check-ins data can potentially be important in a predictive model, this piece of information has not been used in our paper. The reason is that the check-ins data provided by Yelp are only cumulative and therefore we cannot tell the amount of check-ins at different stages of the life of a restaurant (e.g. three months before the closing date).

User_data - The goal of this stage is to calculate a social influence score for each user.

Preliminary steps:

1. sum up all the votes.* variables (*votes.funny*, *votes.useful*, *votes.cool*) in the *votes* variable
2. sum up all the compliments.* variables in the *compliments* variable
3. count the number of years the user have been using Yelp
4. count the numbers of years the user have been part of the Yelp Elite

Influence score is defined as:

$$\textbf{score} = \text{review count} + \text{fans} + \text{votes} + \text{compliments} + \text{friends num} + \text{years elite} + \text{yelping since}$$

All the components of the *influence.score* variable have been first normalized in the 0-1 interval, before being summed up. The influence score for the users in the Yelp dataset ranges from *0* to *2.022*. All the users with an influence score greater than *0.75* have been marked as "*influential*" (see *Appendix* - figure 1, left)

At the end of this preparation stage, the *user_data* dataset contains *269.231* observations of *3* variables (*user_id*, *user.influence.score*, *user.influential*)

Final dataset - The *user_data*, *business_data*, *review_data*, and *tip_data* datasets have been merged into one single dataset with *15.892* restaurants (*64* features). After the removal of all those features with near zero variance and a correlation greater than *0.7*, we obtained a dataset with *43* features that has been used for the modeling stage. Two periods of the story of a restaurant have been taken into consideration: **A)** from the opening date to 6 months before the last review or closing date, **B)** from 6 months before the last review or closing date to 3 months before the last review or closing date (the reason is that the model wants to predict the failure 90 days in advance).

- **attributes.*** - various business attributes (23 variables)
- Business lifespan in days (1 variable)
- Distance from city center in meters (1 variable)
- Reviews sentiment scores from the opening date and in the last 90 days (mean and sd) (4 variables)
- Tips sentiment scores from the opening date and in the last 90 days (mean and sd) (3 variables)
- Stars from the opening and in the last 90 days (mean and sd) (4 variables)
- Users' influence score from opening and in the last 90 days (mean and sd) (4 variables)
- Count of influential users from opening and in the last 90 days (2 variables)
- Tips count in the last 90 days (1 variable)

The final dataset contains *13434* open restaurants and *2548* closed restaurants. Since the two classes are slightly unbalanced, the *caret* **upSample** function has been used (randomly sample, with replacement, the minority class to be the same size as the majority class).

Predictive modeling and Recursive Feature Elimination

Three different predictive models have been developed and, as a final check, a Recursive Feature Elimination via *caret* has been executed.

Models - A *70%/30%* split has been used for the train and test sets. The three models reached the following accuracies:

1. **C5.0 Decision Trees** - accuracy : **0.9759** - 95% CI : (0.9723, 0.9792)
2. **Random forests** - accuracy **0.9801** - 95% CI : (0.9769, 0.9831)
3. **SVM tuned** ($\gamma = 0.1$, cost=10) - accuracy **0.9784** - 95% CI : (0.975, 0.9815)

The confusion matrices are summarized in the table below.

Outcome	C50	RF	SVM
TP	3887	3923	3920
TN	3979	3977	3966
FP	143	107	110
FN	51	53	64

See *Appendix* for complete results of the Random Forest and SVM tuned classifiers.

The Random Forest model shall be preferred. The top 5 important variables in the Random Forest model are: *business.lifespan*, *user.influence.score.from.opening.mean*, *user.influence.score.last.90.days.mean*, *review.sentiment.score.from.opening.mean*, *user.influence.score.from.opening.sd.* This is already a good indication of how social influence factors impact the success/failure of a restaurant.

Recursive Feature Elimination - The Recursive Feature Elimination via *caret* is implemented as the last step of our analysis: since the Random Forest method provides a better accuracy, the *rfFuncs* functions is used for the *caret* rfe execution.

```
rfProfile <- rfe(train, trainClass,
                 sizes = c(1:10, 15, 20, 25, 30),
                 rfeControl = rfeControl(functions = rfFuncs, method = "cv", repeats = 10))
```

The predictions performances obtained are: **Accuracy**: *0.9801489* , **Kappa**: *0.9602978*

Results

The Recursive Feature Elimination via *caret* shows that the *43* variables are all useful to improve the accuracy of a predictive model (see fig.2 - *Appendix*). The top *15* important variables are:

	Overall	var
1468	74.50775	business.lifespan
1469	58.94616	user.influence.score.from.opening.mean
1470	55.81359	business.distance.from.center
1471	53.23315	review.stars.from.opening.mean
1472	50.55651	review.stars.from.opening.sd
1473	48.83719	user.influence.score.last.90.days.mean
1474	47.17406	user.influence.score.from.opening.sd
1475	44.98086	user.influential.from.opening.count
1476	42.31487	review.sentiment.score.from.opening.mean
1477	40.17764	attributes.Good.For.dinner
1478	38.23755	attributes.Wheelchair.Accessible
1479	38.06063	review.sentiment.score.from.opening.sd
1480	32.13103	user.influence.score.last.90.days.sd
1481	31.16207	review.stars.last.90.days.mean
1482	30.03229	tip.sentiment.score.from.opening.mean

Discussion

The top 15 features after the Recursive Feature Elimination stage can be grouped as follows:

- business lifespan
- influence score related features
- location (distance from center)
- number of stars
- sentiment score for reviews and tips
- wheelchair accessibility
- restaurant is good for dinner

The variable with the second highest importance is the mean user influence score, measured from the opening date of the restaurant. Other interesting variables ranked respectively 6th, 7th, 8th, and 12th: the average user influence score in the last 90 days, the user influence score sd from the opening, the count of influential reviewers from the opening, the user influence score sd in the last 90 days.

The calculation of variable importance in the RF model and in the RFE profile answers the central question of our paper: social influence of reviewers is an important factor for the success of a restaurant.

This finding can be explained by the fact that reviews of influential users reach a broader audience (more friends, more fans). Or other subtle mechanisms could explain this phenomenon: for instance, reviews of influential users can be more effective in triggering a spontaneous word-of-mouth marketing that can strongly influence the success or failure of a restaurant.

Limitations and future work - The Random Forest predictive model described in this paper does not make full use of the information available in the Yelp Dataset and could be improved in many ways: for instance sentiment analysis could be based on n-grams, or other external features could be used like competition factors (i.e. similar restaurants in the same area) or factors related to entrepreneurial competence (e.g. turnover rate of employees, costs/revenue ratio).

Appendix

Plots

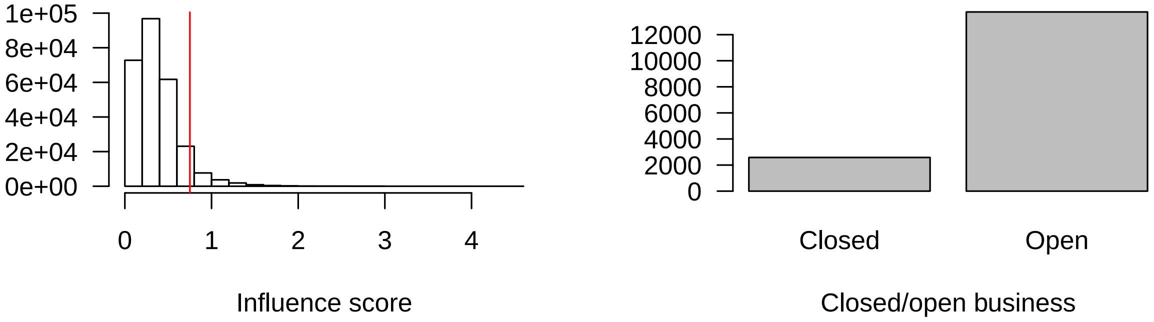

Figure 1: Influence score (left) - Majority and minority class (right)

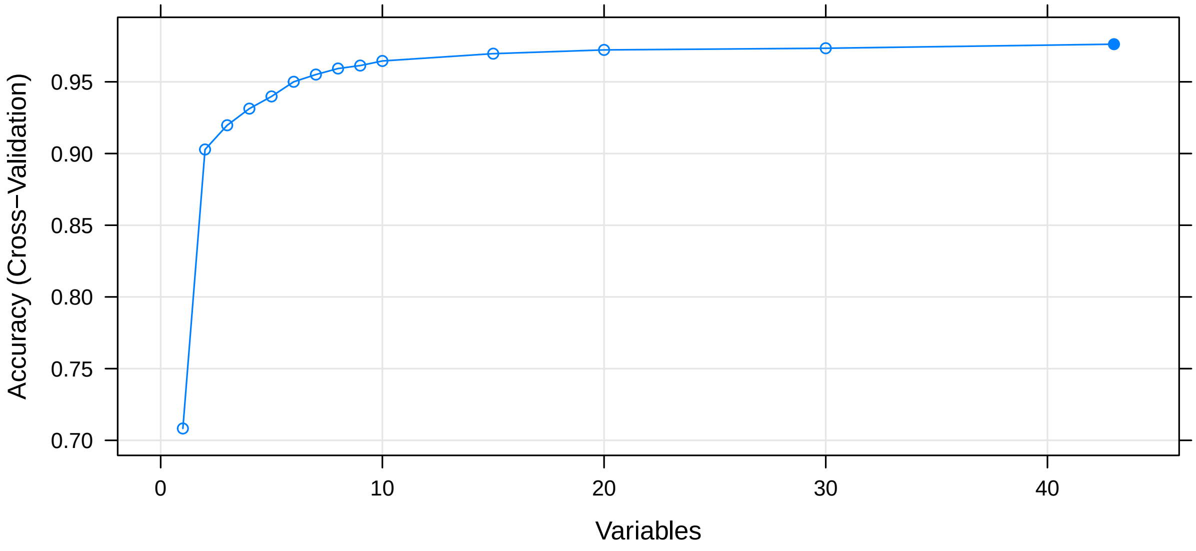

Figure 2: RFE Random Forest profile

Models evaluation

Random Forest

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3977  107
##      TRUE     53 3923
##
##                Accuracy : 0.9801
##                  95% CI : (0.9769, 0.9831)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9603
##  Mcnemar's Test P-Value : 2.789e-05
##
##             Sensitivity : 0.9868
##             Specificity : 0.9734
##          Pos Pred Value : 0.9738
##          Neg Pred Value : 0.9867
##              Prevalence : 0.5000
##          Detection Rate : 0.4934
##    Detection Prevalence : 0.5067
##       Balanced Accuracy : 0.9801
##
##        'Positive' Class : FALSE
##
```

SVM tuned ($\gamma = 0.1$, cost=10)

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3966  110
##      TRUE     64 3920
##
##                Accuracy : 0.9784
##                  95% CI : (0.975, 0.9815)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9568
##  Mcnemar's Test P-Value : 0.0006462
##
##             Sensitivity : 0.9841
##             Specificity : 0.9727
##          Pos Pred Value : 0.9730
##          Neg Pred Value : 0.9839
##              Prevalence : 0.5000
##          Detection Rate : 0.4921
##    Detection Prevalence : 0.5057
##       Balanced Accuracy : 0.9784
##
##        'Positive' Class : FALSE
##
```

References

[1] J. Mejia, S. Mankad, and A. Gopal. “More Than Just Words: Using Latent Semantic Analysis in Online Reviews to Explain Restaurant Closures”.

[2] Youn, Hyewon, and Zheng Gu. “Predict US Restaurant Firm Failures: The Artificial Neural Network Model Versus Logistic Regression Model”. Tourism and Hospitality Research 10.3 (2010): 171-187. Web..

[3] H.G. Parsa, MS, MS, Ph.D., FMP, Amy Gregory, MBA, Michael ‘Doc’ Terry, MBA. WHY DO RESTAURANTS FAIL? PART III: AN ANALYSIS OF MACRO AND MICRO FACTORS. https://hospitality.ucf.edu/files/2011/08/DPI-Why-Restaurants-Fail.pdf

[4] National Restaurant Association. Restaurant Industry Operations Report 2010 Edition. https://s3.amazonaws.com/s3.documentcloud.org/documents/291534/t288-nrarept2010.pdf

[5] Jeffrey Breen. R by example: mining Twitter for consumer attitudes towards airlines. https://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/

[6] Hu and Liu, KDD-2004. A list of positive and negative opinion words or sentiment words for English. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

[7] Recursive Feature Elimination. The caret Package website. http://topepo.github.io/caret/rfe.html

[8] Yelp Dataset Challenge. http://www.yelp.com/dataset_challenge