Predicting Restaurant Failure Using Yelp Data

November 17, 2015

Introduction

Restaurants have one of the highest business failure rates of all the retail and service industries, with a 30% failure rate commonly accepted as the norm. Some of the macro factors associated with restaurant failure include the economy, federal and local legislation, climate and natural events, regional and urban planning, changing cultural factors, and new competition. There are also many business-related micro factors associated with restaurant failure, including lack of capital, lack of industry experience, poor leadership, lack of cost controls, and high fixed costs. Other micro factors that contribute to restaurant failure include location (not just the physical site, but the demographics of the surrounding area as well), failing to follow the 12 Ps of restaurant branding (Place, Product, Price, People, Promotion, Promise, Principles, Props, Production, Performance, Positioning and Press), and even the name of the restaurant (restaurants with names that are brief, descriptive and attractive are more likely to succeed). For a more detailed discussion, please see http://hospitality.ucf.edu/files/2011/08/DPI-Why-Restaurants-Fail.pdf

The primary question of interest for this project is to attempt to identify drivers of restaurant failure (i.e., predictors of restaurants that are no longer in business) using Yelp data. Given that Yelp data does not contain strong predictors such as economic information, the goal is to build a model that predicts restaurant failure using only business and review data.

Methods

Understanding the Problem and the Dataset

The data for this project was part of Round 6 of the Yelp Dataset Challenge, which consists of information about businesses from 10 cities in 4 countries and includes:

- Over 1.5M reviews by 366K users for 61K businesses, and
- Over 481K business attributes such as hours, parking, ambience, etc.

Of the five files included in the Yelp dataset (business, checkins, reviews, tips, and users), only the business and review data files will be used in this project. For the purposes of predicting restaurant failure, the target variable will be whether or not a restaurant is open for business (coded ‘Yes’ for not in business and ‘No’ for in business so the model will predict failure rather than success), and the predictor variables will be a combination of business attributes, review data, and feature-engineered variables.

Pre-Processing the Data

A total of 21,799 restaurants and 990,627 restaurant reviews were extracted from the business and review data files for further processing, based on the presence of the word ‘Restaurant’ in the business category variable. Four restaurant types are represented in the data set: fast food, fast casual, casual, and fine dining. Variables with greater than 50% missing values were removed from the dataset, and all missing values in the remaining variables were recoded to 0. All character variables were then recoded to dummy variables with the `dummyVars` function in the `caret` package using the `fullRank = TRUE` option to avoid creating linear dependencies.

Feature Engineering

Seven variables were feature-engineered for the predictive model. First, a total of 80 different restaurant categories were identified in the business category variable (from Afghan to Vietnamese). Second, a new city variable identifying the 10 cities in the dataset (Canada: Montreal, Waterloo; Germany: Karlsruhe; US: Charlotte, Las Vegas, Madison, Phoenix, Pittsburgh, Urbana-Champaign; UK: Edinburgh) was created by performing a k-means cluster analysis of the restaurants' latitude and longitude. These two variables were then converted to continuous variables using the Weight of Evidence (WOE) transformation. For the binary classification problem in this project, WOE can be defined as $WOE_i^X = ln(C_i^X/TC)/(N_i^X/TN))$ where TC and TN are the total number of closed and open restaurants, respectively, and C_i^X and N_i^X are the number of closed and open restaurants for the ith value of attribute X.

Third, a sentiment analysis of the review verbatims was performed based on Jeffrey Breen's Twitter sentiment analysis tutorial using Hu & Liu's opinion lexicon, where an overall sentiment score is computed based on the number of positive words minus the number of negative words. A normalized mean sentiment score was then computed for each restaurant by dividing the sentiment score by the review length and taking the mean across all reviews. The remaining feature-engineered variables for each restaurant were mean review length (number of words), restaurant name length (number of words), percent of review verbatims with the word 'manager' or 'management' (hypothesized to be an indicator of poor service), and percent of one-star reviews.

Selecting the Modeling Algorithm

A gradient boosted decision tree classification model was selected as the modeling algorithm for this project. Gradient boosted decision trees fit many large or small trees to reweighted versions of the training data, and classify on the target variable by weighted majority vote. Tianqi Chen's eXtreme Gradient Boosting (`XGBoost`) is a fast and efficient implementation of gradient boosting framework. `XGBoost` yields accurate predictions for most data sets, as evidenced in its use by several recent Kaggle competition winners.

Parameter Tuning Through Cross-Validation

First, the modeling dataset was split 60/20/20 into training, test, and validation sets using the `createDataPartition` function in the `caret` package in order to preserve the overall class distribution of the data. Next, a gradient boosted decision tree classification model was fit to the training data using the `xgboost` package with 10-fold cross validation repeated 5 times. The final parameters for the optimal model were `nrounds` = 150, `max_depth` = 3, and `eta` = 0.3. Predictions were computed on the test set and a confusion matrix was generated for model evaluation. A second set of predictions was then computed on the validation set.

Building the Model

In order to account for class imbalance in the dataset (i.e., the low proportion of restaurants that failed compared to those that did not), an alternate threshold for the predicted probability was computed using Youden's J index, which measures the proportions of correctly predicted samples for both the event and nonevent groups and can be defined as $J = Sensitivity + Specificity - 1$. A new set of predictions was then computed using the alternate threshold and a confusion matrix was generated to evaluate the final model.

Results

Exploratory Data Analysis

The overall prevalence of restaurant failure in the Yelp dataset was 19.8%, but varies widely by city, with only 2.1% failure in Karlsruhe but 24.6% failure in Urbana-Champaign. As a result, city should be a important predictor in the final model.

```
##
## Failed? Charlotte Edinburgh Karlsruhe Las Vegas Madison Montreal Phoenix
##     Yes     0.185     0.142     0.021     0.226   0.215    0.093   0.236
##     No      0.815     0.858     0.979     0.774   0.785    0.907   0.764
##
## Failed? Pittsburgh Urbana-Champaign Waterloo
##     Yes      0.180            0.246    0.066
##     No       0.820            0.754    0.934
```

Restaurant failure also varies widely by restaurant category, ranging from 0% for bistros to 58.1% for Cuban restaurants. The top five restaurant categories with the highest failure percentage are shown below. Restaurant category should also be an important predictor in the final model.

```
##     Category Failed?
## 45     Cuban   0.581
## 48   Russian   0.467
## 39     Cajun   0.422
## 47 Soul Food   0.419
## 2      Irish   0.400
```

Finally, there does appear to be a linear relationship between name length and restaurant failure: restaurants with 6 or more words in the name have a failure rate nearly twice as high as those with only one-word names (25.8% vs. 13.6%).

```
##
## Failed?     1     2     3     4     5     6
##     Yes 0.136 0.192 0.214 0.210 0.258 0.258
##     No  0.864 0.808 0.786 0.790 0.742 0.742
```

Model Building

The confusion matrix and evaluation diagnostics for the predictions of the tuned model on the test set are shown below. The overall accuracy was 85.3%, which is slightly better than the no-information rate of 80.2%. A kappa value of 0.457 suggests moderate agreement. However, the model has low sensitivity (43.2%), which suggests that the model has trouble predicting restaurant failure due to the class imbalance of the target variable and lack of strong predictors in the model.

```
##           Reference
## Prediction  Yes   No
##        Yes  373  149
##        No   491 3347
```

```
##     Accuracy        Kappa AccuracyNull
##    0.8532110    0.4572203    0.8018349
```

```
## Sensitivity Specificity
##   0.4317130   0.9573799
```

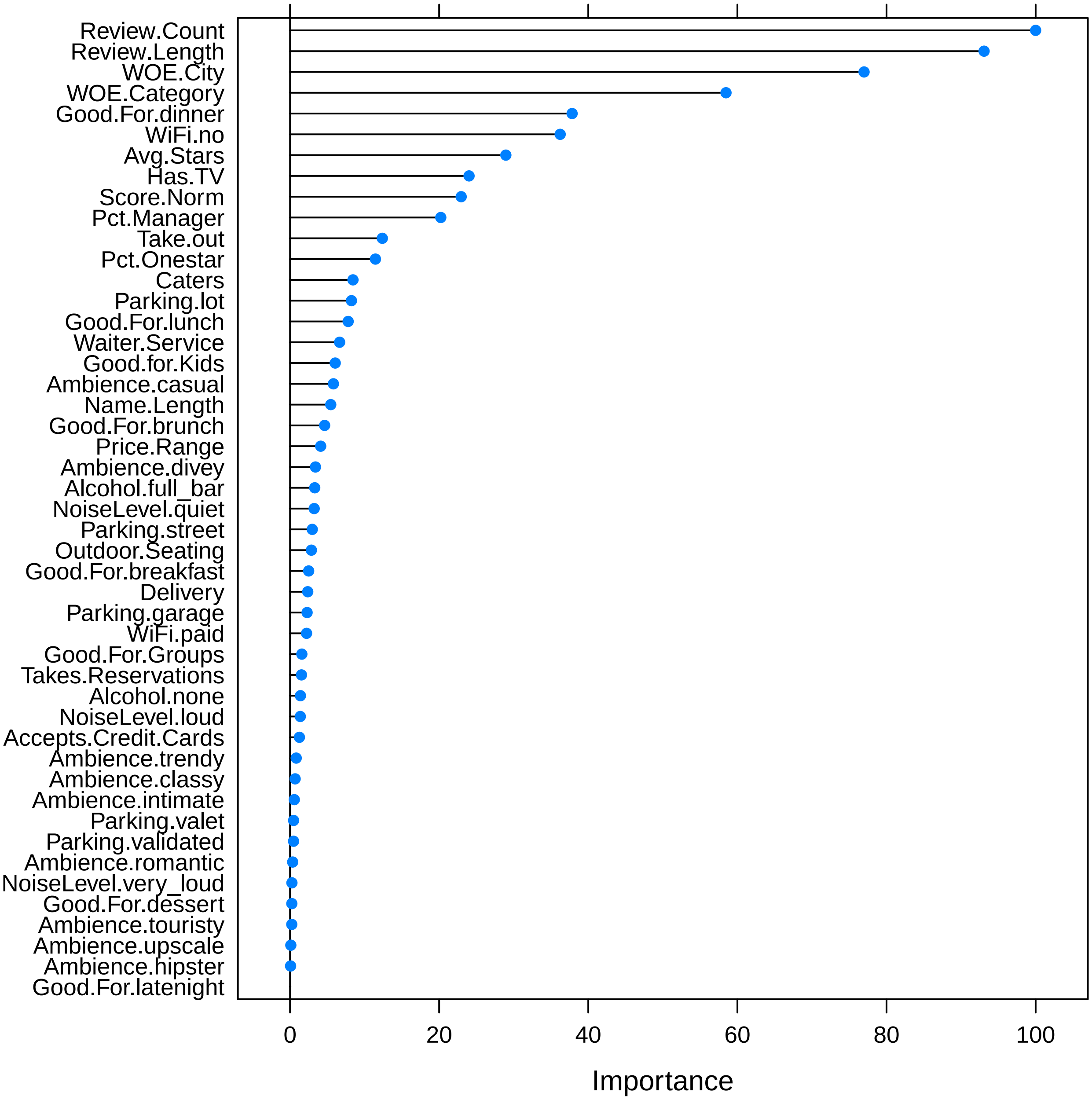
While the issue of strong predictors cannot be addressed with the data on hand, the class imbalance can be addressed by adjusting the cutoff value for the predicted probabilities, which is set to 50% by default. Using an alternate cutoff of 20.5% (i.e., probabilities greater than 0.205 are called events) based on Youden's J index increases the sensitivity of the model from 43.1% to 76.5%.

```
##   threshold specificity sensitivity
##   0.2053138   0.7854691   0.7659328
```

The confusion matrix for the model predictions on the validation set using the alternate threshold is shown below. The final model clearly does a much better of job of predicting restaurant failure while achieving a better balance between sensitivity and specificity.

```
##           Reference
## Prediction  Yes   No
##        Yes  661  750
##        No   202 2746
```

A dotplot of variable importance in the final model is shown below. The top 5 predictors are review count, review length, city, restaurant category, and the business attribute 'Good For Dinner'.



Restaurants that are no longer in business tend to have fewer but longer reviews on average.

```
##   Failed? Review.Count Review.Length
## 1     Yes     26.37069      704.6122
## 2      No     55.45143      595.0459
```

Restaurants that are flagged as 'Good For Dinner' have a failure rate nearly twice as high compared to those that are not (28.1% vs 17.0%).

```
## 
## Failed?     0     1
##     Yes 0.170 0.281
##     No  0.830 0.719
```

Discussion

The primary question of interest for this project was to identify drivers of restaurant failure using Yelp review data. Using a powerful prediction algorithm in the form of gradient boosted decision trees resulted in a predictive model with reasonably high sensitivity after the model was adjusted for class imbalance. As a result, it was possible to answer the primary question of interest despite the lack of strong predictors in the dataset. It is interesting to note that five of the seven feature-engineered variables were among the top 10 drivers in terms of variable importance. Without feature engineering it is doubtful that an accurate predictive model could have been achieved.

Review count and review length as the top two drivers of restaurant failure was unexpected. Fewer reviews on average for failed restaurants is perhaps not too surprising, as review count could be considered to be a proxy for popularity. However, the finding that reviews for failed restaurants are over 100 words longer on average compared to reviews of successful restaurants suggests that longer reviews tend to express concern rather than praise.

The city and restaurant category variables played an important role in predicting restaurant failure, as initially hypothesized. However, the prevalence of restaurant failure in the Canadian and German cities was much lower than the other cities in the dataset, and not even close to the 30% rate considered to be the industry norm. As a result, while city was one of the top five drivers of restaurant failure identified by the model, removing these three cities from the model would likely have resulted in a much lower importance score for the city variable. The restaurant category variable suggests that restaurants with non-traditional cuisines such as Cuban, Russian, Cajun and Soul Food run the highest risk of failure. The 'Good for Dinner' flag suggests that restaurants at the higher end of the price scale (i.e., casual and fine dining) run the highest risk of failure.

In terms of further analysis, it would be interesting to see if the relationship between business failure, review count and review length identified in this project generalizes to other businesses in other cities. While an analysis of Yelp reviewers themselves was not considered in this project, a follow-up analysis of review length and rating by reviewer would also be an interesting analysis.