Predicting Open Restaurants from Yelp Attributes

Zeydy Ortiz, Ph. D.

November 22, 2015

INTRODUCTION

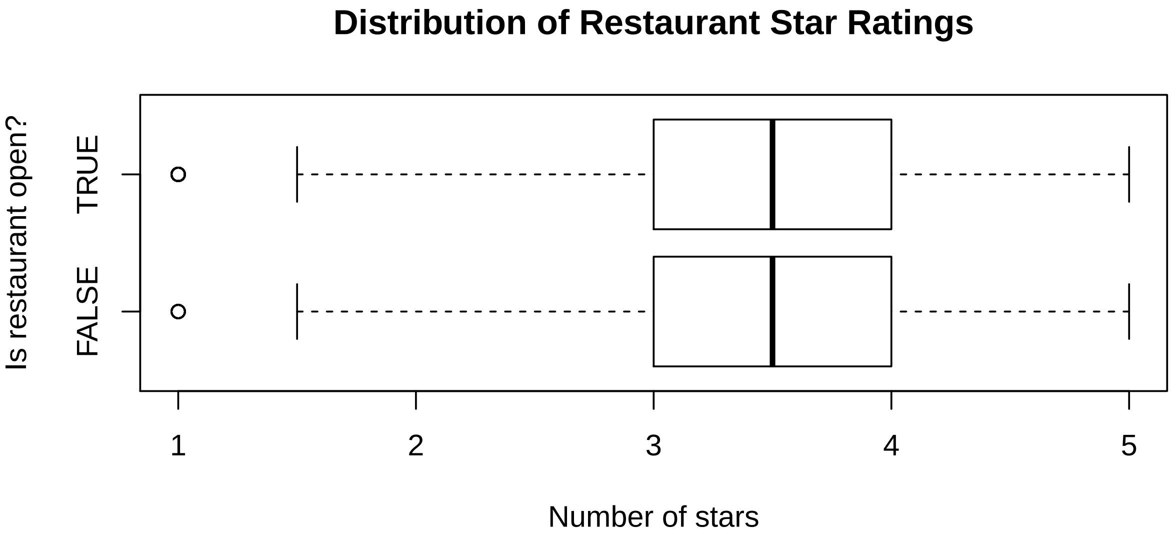
What are the factors that indicate a restaurants is in operation (open)? Many businesses fail and restaurants are believed to be among the businesses with the highest failure rates. I used the Yelp dataset to gain insights on the attributes that predict restaurants are open.

In this project I created a random forest prediction model and looked at the top importance factors. Producing a predictive model is an interesting academic exercise. However, looking at the important factors that influence the model give us insights into what patrons value. This could be of interest to restaurant owners to help them beat the odds of failure.

METHODS AND DATA

The data used in this report is part of the Round 6 Yelp Dataset Challenge provided for the Data Science Capstone project. The business information is from the `yelp_academic_dataset_business.json` file. To prepare the data for prediction, I first selected all businesses in the 'Restaurants' category. There were a total of 21892 restaurants and 80.2028138% are classified as `open`. This is a higher percentage than expected indicating that there may be a confounding variable related to the fact that the business is listed in Yelp in the first place. In the data cleaning process, I eliminated variables that had more than 90% `NAs`. For the variables left, `NA` was coded as `Unk` - unknown - in order to use them in the analysis.

I looked at the distribution of the star ratings to see if there was a difference that could explain why some of the restaurants were no longer open.



At first glance there seems to be no difference in the ratings. However, the result of the t-test below reveals that there is a difference in the mean values. So, we expect the star rating to be a factor in a predictive model.

```r
t.test(stars ~ open, data=rest_df)
```

```
##
##  Welch Two Sample t-test
##
## data:  stars by open
## t = -5.751, df = 6803.6, p-value = 9.254e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09317971 -0.04580507
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            3.425012            3.494504
```

RESULTS

The baseline accuracy of a model predicting all restaurants are open would be 80.2028138% for all restaurants. I created a random forest model as follows:

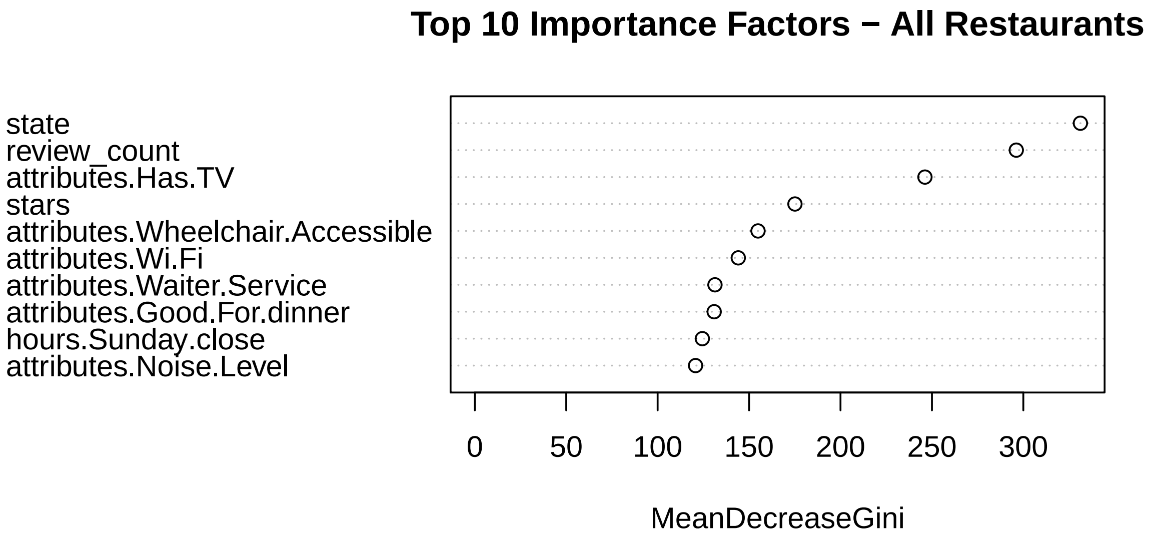
```r
# Split data set
set.seed(123)

split <- sample.split(rest_df$open, SplitRatio = 0.7)
train <- subset(rest_df, split==TRUE)
test <- subset(rest_df, split==FALSE)

# Build random forest
set.seed(456)
restRF <- randomForest(open~., data=train)
predRF <- predict(restRF, newdata=test)
table(predRF, test$open)
```

```
##
## predRF  FALSE TRUE
##   FALSE   423   87
##   TRUE    877 5180
```

The accuracy of the model is 85.3205421%, higher than the baseline model. The importance plot for this model reveal some interesting insights.



The most important factor in the model is the `state` in which the restaurant is located. We will look at two different `state` values in the discussion section below to determine how the factors may differ.

The second most important factor is `review_count`. My initial assumption was that restaurants that failed may have more people complaining and more reviews. On the other hand, restaurants that close may have been open for a shorter period of time resulting in less reviews. The summaries below show that the second assumption was correct. It may be interesting to find out how long the restaurants have been in business to normalize this factor.

```
summary(rest_df$review_count[rest_df$open==TRUE])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    7.00   18.00   55.23   54.00 4578.00
```

```
summary(rest_df$review_count[rest_df$open==FALSE])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    6.00   11.00   26.29   27.00 1352.00
```

The star rating is the 4th most significant factor when all restaurants are used in the model. The model indicates that whether a restaurant has a TV or not is the 3rd most significant factor for patrons.
More restaurants that are still open have a TV. However, for the majority of restaurants that are closed it was `unknown` whether they had a TV or not.

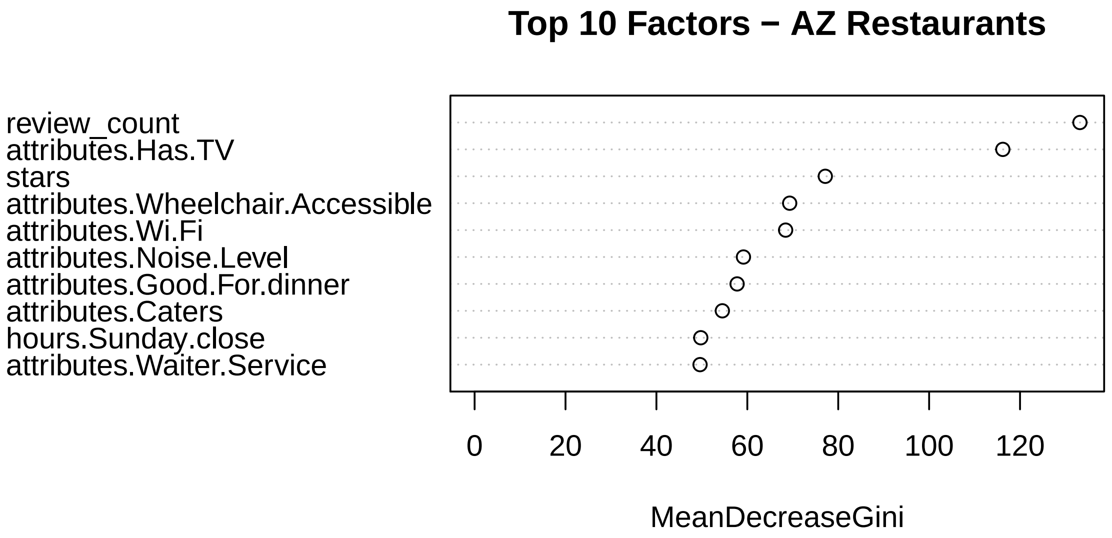
Rounding out the Top 10, the other importance factors were: Wheelchair Accessible, Wi-Fi, Waiter Service, Good for Dinner, hour they close on Sunday and the Noise Level.

DISCUSSION

Since `state` is a highly influential factor in the predictive model, I created separate models for restaurants in Arizona (AZ) and Quebec (QC). These `states` were selected because they had a large number of restaurants in the data set - 7985 in AZ and 2353 in QC - and their climates are very different. I wanted to see if the factors in their random forest models would differ from the general random forest model and from each other.

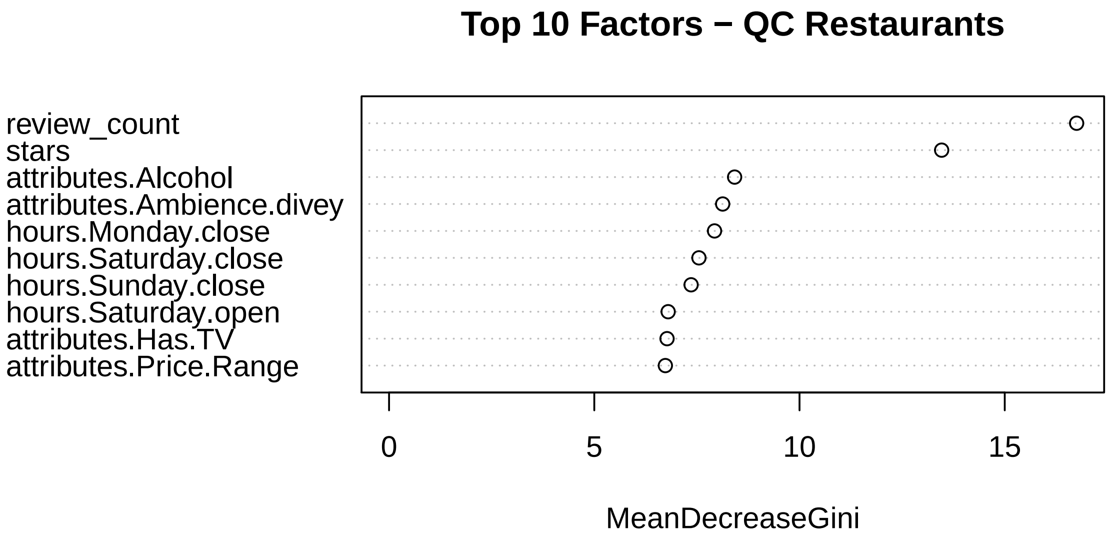
Only 75.8678377% of restaurants in AZ are open (as opposed to 80% for all restaurants). The random forest model for restaurants in AZ accurately predict 83.5215391% of the test cases - a 7.6% increase from the baseline model. Incidentally, the general model that was discussed in the previous section can accurately predict 83.5633626% of the cases.

The top 10 factors for restaurants in AZ are very similar to the factors for the general model. Since 36.4745112% of the restaurants are in AZ, it is not surprising that they highly influence the general model for all restaurants.



On the other hand, only 10.7482185% of the restaurants are in Quebec. The baseline for these restaurants is 90.0280899%. The general model accurately predicts 89.8876404% of the cases - one case less than the baseline. Tailoring a model for restaurants in QC increases the accuracy back to 90.0280899%.

The accuracy of the random forest models did not significantly increase. Looking at the plot of importance factors help us understand better what patrons in QC value. Only 5 of the top 10 factors (including `state`) for QC are in the top 10 for all restaurants. However, factors like the Ambience, whether they serve Alcohol, the closing hours and the price range are of importance.



CONCLUSION

This project demonstrated the use of a predictive model to uncover insights on the important factors influencing the model. In this case I used the business attributes for restaurants to predict whether they are in operation (open). However, the most interesting part was understanding the importance factors for the model and building separate models for restaurants in different geographical areas.