Coursera Data Science Capstone Project

David Doyle

November 14, 2015

Exploring the Influence of Ambience on the Business Score

Introduction

The Yelp datasets offer many oppportunities for exploring the data for useful business insights. The question I have decided to pursue is:

> Does the ambience of each business influence the review score - i.e. do certain ambiences tend to result in higher or lower scores overall.

In an ideal world one would expect that the ambience would not be the sole influence on the score - the score should be a reflection of the customer experience. I intend to use the business data set to test my hypothesis that ambience is not a good predictor of the score (number of stars) assigned to a business.

Methods and Data

The code needed to reproduce the results for this report is located on GitHub in the following repository: **https://github.com/cyberhiker1965/YelpProject**

Exploring the Data

The initial task was to read the business dataset and convert it from JSON into a data frame. As the time to extract and convert the data is significant the resulting data frame is saved so that it can be reloaded directly in the future without the conversion overhead.

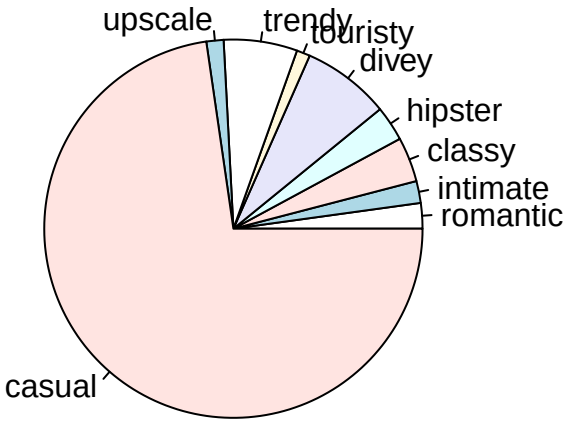
The next task is to explore the data by profiling the fields of interest - in this case to understand the makeup of the data related to ambience. As can be seen in the summary below, it can be seen that compared to the 61,184 rows in the dataset the ambience data is very sparsely populated. This makes sense as it would not apply to many categories of business.

```
##   romantic      intimate       classy       hipster        divey
##  FALSE:18281   FALSE:18316   FALSE:18092   FALSE:17965   FALSE:16685
##  TRUE :  250   TRUE :  215   TRUE :  439   TRUE :  344   TRUE :  861
##  NA's :42653   NA's :42653   NA's :42653   NA's :42875   NA's :43638
##   touristy       trendy        upscale        casual
##  FALSE:18399   FALSE:17810   FALSE:18228   FALSE:10193
##  TRUE :  132   TRUE :  721   TRUE :  168   TRUE : 8338
##  NA's :42653   NA's :42653   NA's :42788   NA's :42653
```

Removing the businesses where no ambience value is populated provides a smaller set of data for evaluation - 11,013 rows. This is 18% of the original data. I will focus on this set of data so my question was refined into "Does the ambience of each business influence the review score *when one or more of the ambience fields are populated*?"

As can be seen from the revised summary and plot of the values set to TRUE the casual ambiance setting is very common. Can we make a prediction with only this information?

```
##   romantic      intimate       classy       hipster        divey
##  FALSE:10763   FALSE:10798   FALSE:10574   FALSE:10654   FALSE:10019
##  TRUE :  250   TRUE :  215   TRUE :  439   TRUE :  344   TRUE :  861
##                                            NA's :   15   NA's :  133
##   touristy       trendy       upscale        casual
##  FALSE:10881   FALSE:10292   FALSE:10774   FALSE:2675
##  TRUE :  132   TRUE :  721   TRUE :  168   TRUE :8338
##                              NA's :   71
```



Building A Prediction Model

Two prediction models that are most suitable for use with binary predictors (the TRUE/FALSE ambience values) were attempted - Random Forest and Naive Bayes. In both cases a data split approach was used to derive and test a prediction model. The data was split into a 60% training set and a 40% testing set. The accuracy of each method was determined using a confusion matrix.

To support creation of the prediction models some additional cleansing was applied to the data:

- as there was only one single-star measurement (see the summary below) it was dropped
- the entries containing NAs were also dropped as the prediction approaches selected do not support the use of NA values

```
## Loading required package: lattice
## Loading required package: ggplot2
```

```
##   stars freq
## 1     1    1
## 2   1.5   24
## 3     2  130
## 4   2.5  594
## 5     3 1769
## 6   3.5 3531
## 7     4 3726
## 8   4.5 1174
## 9     5   64
```

```
## Loading required package: randomForest
## randomForest 4.6-12
## Type rfNews() to see new features/changes/bug fixes.
## Loading required package: klaR
## Loading required package: MASS
```

Results

The confusion matrix using the Random Forest approach yielded an accuracy of 0.3397969, while the accuracy of the Naive Bayes approach was even less impressive at 0.1858264.

(Random Forest overall results)

```
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##    0.339796861    0.004087339    0.325689047    0.354120988    0.980609418
## AccuracyPValue  McnemarPValue
##    1.000000000            NaN
```

(Naive Bayes overall results)

```
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##    0.185826408    0.007853811    0.174342898    0.197735111    0.887811634
## AccuracyPValue  McnemarPValue
##    1.000000000            NaN
```

The results of both approaches are disappointing so they were retried using repeated k-fold cross validation. For Random Forest:

```
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##    0.339796861    0.004087339    0.325689047    0.354120988    0.980609418
## AccuracyPValue  McnemarPValue
##    1.000000000            NaN
```

. . . and for the Naive Bayes model:

```
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##    0.185826408    0.007853811    0.174342898    0.197735111    0.887811634
## AccuracyPValue  McnemarPValue
##    1.000000000            NaN
```

For the Random Forest model the accuracy was unchanged at 0.3397969 and unchanged for the Naive Bayes model at 0.1858264

Discussion

As can be seen from the previous section it is *not* possible to predict the star score with any reasonable accuracy when provided the ambience values. This supports my hypothesis that ambience does not play a significant influence on the overall business score.

While it may influence the customer experience somewhat I believe that the customers most concerned with ambience are going to self-select and avoid establishments with an ambience that does not match their tastes. The remaining customers are going to rate the overall customer experience with ambience only being one factor in the ranking. Further analysis could be done on the Yelp datasets to determine the potential influence of these other factors - possibly using the content of the reviews to identify the customer sentiment and look for keywords that reflect the sentiment and are associated with high/low socres.