Coursera Data Science Capstone Project

David Doyle

November 14, 2015

Exploring the influence of ambience on the business score

Introduction

state the primary question, hypothesis or prediction task of interest clearly here

The Yelp datasets offer many oppportunites 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 ambiance is not a good predictor of the score (number of stars) assigned to a business.

Methods and Data

describe the (or multiple) statistical model, prediction algorithm or statistical inference described in the method

needs some exploratory data analysis with plots/summary tables that interogate the question of interest - has to be relevant to the question

The code needed to reproduce the results for this report is located on GitHub in the following repository: **put repository here**

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.

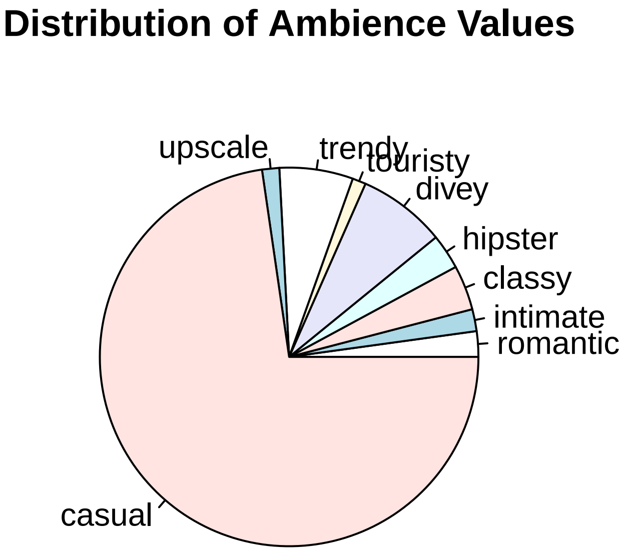
The next task is to explore the data by profiling the fields of interest - in this case to understand the makeup of the data related to ambience. As can be seen in the summary below, it is obvious that compared to the 61,184 rows in the dataset the ambiance data is very sparsely populated.

```
##   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 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.

```
##    1  1.5    2  2.5    3  3.5    4  4.5    5
##    1   24  130  594 1769 3531 3726 1174   64
```

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 summar 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 methods presented in the results section introduced in the methods section

the primary statistical model, statistical inference or prediction output in the results should be summarized and interpreted

include at least one plot or table here

description of how the results relate to the primary questions of interest, or is it otherwise clear? In other words, do not give a point if the results seem unrelated to the question of interest and there is no apparent relationship. The confusion matrix using the random forest approach is below:

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction  1.5    2  2.5    3  3.5    4  4.5    5
##        1.5    0    0    0    0    0    9    0    0
##        2      0    0    0    1    0   50    0    0
##        2.5    0    0    1    5    2  219    0    0
##        3      0    0    5   13    4  674    0    0
##        3.5    0    0    1   16    8 1369    0    0
##        4      0    0    1    8   12 1450    0    0
##        4.5    0    0    0    1    6  453    0    0
##        5      0    0    0    0    0   24    0    0
## 
## Overall Statistics
## 
##                Accuracy : 0.3398
##                  95% CI : (0.3257, 0.3541)
##     No Information Rate : 0.9806
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0041
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1.5 Class: 2 Class: 2.5 Class: 3 Class: 3.5
## Sensitivity                  NA       NA  0.1250000 0.295455   0.250000
## Specificity            0.997922  0.98823  0.9477336 0.840718   0.677674
## Pos Pred Value               NA       NA  0.0044053 0.018678   0.005739
## Neg Pred Value               NA       NA  0.9982948 0.991474   0.991831
## Prevalence             0.000000  0.00000  0.0018467 0.010157   0.007387
## Detection Rate         0.000000  0.00000  0.0002308 0.003001   0.001847
## Detection Prevalence   0.002078  0.01177  0.0524007 0.160665   0.321791
## Balanced Accuracy            NA       NA  0.5363668 0.568086   0.463837
##                      Class: 4 Class: 4.5 Class: 5
## Sensitivity           0.34134         NA       NA
## Specificity           0.75000     0.8938  0.99446
## Pos Pred Value        0.98572         NA       NA
## Neg Pred Value        0.02202         NA       NA
## Prevalence            0.98061     0.0000  0.00000
## Detection Rate        0.33472     0.0000  0.00000
## Detection Prevalence  0.33957     0.1062  0.00554
## Balanced Accuracy     0.54567         NA       NA
```

The confusion matrix using the Naive Bayes approach is below:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1.5    2  2.5    3  3.5    4  4.5    5
##        1.5    0    0    0    9    0    0    0    0
##        2      0    0    0   50    1    0    0    0
##        2.5    0    0    0  216    8    3    0    0
##        3      0    0    0  656   25   15    0    0
##        3.5    0    0    0 1264   85   45    0    0
##        4      0    0    0 1251  156   64    0    0
##        4.5    0    0    0  380   53   27    0    0
##        5      0    0    0   20    2    2    0    0
##
## Overall Statistics
##
##                Accuracy : 0.1858
##                  95% CI : (0.1743, 0.1977)
##     No Information Rate : 0.8878
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0079
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
```

```
##                      Class: 1.5 Class: 2 Class: 2.5 Class: 3 Class: 3.5
## Sensitivity                  NA       NA         NA   0.1706    0.25758
## Specificity            0.997922  0.98823     0.9476   0.9177    0.67291
## Pos Pred Value               NA       NA         NA   0.9425    0.06098
## Neg Pred Value               NA       NA         NA   0.1227    0.91661
## Prevalence             0.000000  0.00000     0.0000   0.8878    0.07618
## Detection Rate         0.000000  0.00000     0.0000   0.1514    0.01962
## Detection Prevalence   0.002078  0.01177     0.0524   0.1607    0.32179
## Balanced Accuracy            NA       NA         NA   0.5441    0.46524
##                      Class: 4 Class: 4.5 Class: 5
## Sensitivity           0.41026         NA       NA
## Specificity           0.66307     0.8938  0.99446
## Pos Pred Value        0.04351         NA       NA
## Neg Pred Value        0.96784         NA       NA
## Prevalence            0.03601     0.0000  0.00000
## Detection Rate        0.01477     0.0000  0.00000
## Detection Prevalence  0.33957     0.1062  0.00554
## Balanced Accuracy     0.53667         NA       NA
```

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

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1.5    2  2.5    3  3.5    4  4.5    5
##        1.5    0    0    0    0    0    9    0    0
##        2      0    0    0    1    0   50    0    0
##        2.5    0    0    1    5    2  219    0    0
##        3      0    0    5   13    4  674    0    0
##        3.5    0    0    1   16    8 1369    0    0
##        4      0    0    1    8   12 1450    0    0
##        4.5    0    0    0    1    6  453    0    0
##        5      0    0    0    0    0   24    0    0
##
## Overall Statistics
##
##                Accuracy : 0.3398
##                  95% CI : (0.3257, 0.3541)
##     No Information Rate : 0.9806
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0041
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1.5 Class: 2 Class: 2.5 Class: 3 Class: 3.5
## Sensitivity                  NA       NA  0.1250000 0.295455   0.250000
## Specificity            0.997922  0.98823  0.9477336 0.840718   0.677674
## Pos Pred Value               NA       NA  0.0044053 0.018678   0.005739
## Neg Pred Value               NA       NA  0.9982948 0.991474   0.991831
## Prevalence             0.000000  0.00000  0.0018467 0.010157   0.007387
## Detection Rate         0.000000  0.00000  0.0002308 0.003001   0.001847
## Detection Prevalence   0.002078  0.01177  0.0524007 0.160665   0.321791
## Balanced Accuracy            NA       NA  0.5363668 0.568086   0.463837
##                      Class: 4 Class: 4.5 Class: 5
## Sensitivity           0.34134         NA       NA
## Specificity           0.75000     0.8938  0.99446
## Pos Pred Value        0.98572         NA       NA
## Neg Pred Value        0.02202         NA       NA
## Prevalence            0.98061     0.0000  0.00000
## Detection Rate        0.33472     0.0000  0.00000
## Detection Prevalence  0.33957     0.1062  0.00554
## Balanced Accuracy     0.54567         NA       NA
```

…and the Naive Bayes model:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1.5    2  2.5    3  3.5    4  4.5    5
##        1.5    0    0    0    9    0    0    0    0
##        2      0    0    0   50    1    0    0    0
##        2.5    0    0    0  216    8    3    0    0
##        3      0    0    0  656   25   15    0    0
##        3.5    0    0    0 1264   85   45    0    0
##        4      0    0    0 1251  156   64    0    0
##        4.5    0    0    0  380   53   27    0    0
##        5      0    0    0   20    2    2    0    0
##
## Overall Statistics
##
##                Accuracy : 0.1858
##                  95% CI : (0.1743, 0.1977)
##     No Information Rate : 0.8878
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0079
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1.5 Class: 2 Class: 2.5 Class: 3 Class: 3.5
## Sensitivity                  NA       NA         NA   0.1706    0.25758
## Specificity            0.997922  0.98823     0.9476   0.9177    0.67291
## Pos Pred Value               NA       NA         NA   0.9425    0.06098
## Neg Pred Value               NA       NA         NA   0.1227    0.91661
## Prevalence             0.000000  0.00000     0.0000   0.8878    0.07618
## Detection Rate         0.000000  0.00000     0.0000   0.1514    0.01962
## Detection Prevalence   0.002078  0.01177     0.0524   0.1607    0.32179
## Balanced Accuracy            NA       NA         NA   0.5441    0.46524
##                      Class: 4 Class: 4.5 Class: 5
## Sensitivity           0.41026         NA       NA
## Specificity           0.66307     0.8938  0.99446
## Pos Pred Value        0.04351         NA       NA
## Neg Pred Value        0.96784         NA       NA
## Prevalence            0.03601     0.0000  0.00000
## Detection Rate        0.01477     0.0000  0.00000
## Detection Prevalence  0.33957     0.1062  0.00554
## Balanced Accuracy     0.53667         NA       NA
```

Discussion

- primary question of interest answered / refuted or was there a description of why no clear answer could be obtained*

References