DataFest 2017

Chapel-Boys.R

4/1/2017

Introduction

We seek to use [OUTSIDE_DATA] to predict the propensity to book. In order, to measure such user behavior we consider a number of classification models and then use a validation procedure to select that one that performs most optimally. Then, we attempt to demonstrate the [OUTSIDE_DATA]'s importance to our model and explanatory power.

PCA

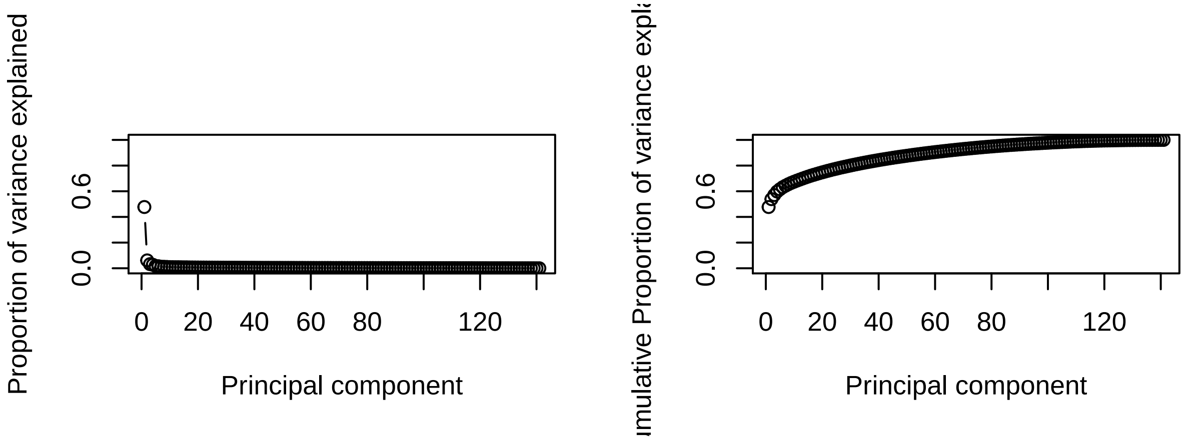
Evidently, the dest.txt file contains valuable information in relation to characteristics unique to each destination. However, it is prudent to begin our analysis by shrinking the number of explanatory variables from 144 to a smaller number of principal components.

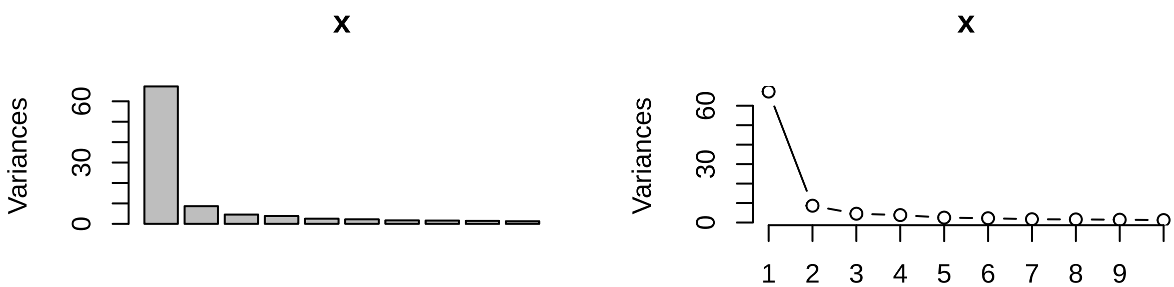
PC1 & PC2

Hence, we can visualize the first two orthogonal principal components:

```
## [1] "proportions of variance:"
##   [1] 4.771562e-01 6.126125e-02 3.209145e-02 2.714611e-02 1.787877e-02
##   [6] 1.573802e-02 1.204289e-02 1.131904e-02 1.024248e-02 8.960971e-03
##  [11] 8.581326e-03 8.188352e-03 8.157166e-03 7.779775e-03 7.476885e-03
##  [16] 7.374436e-03 6.674175e-03 6.565446e-03 6.434203e-03 6.180513e-03
##  [21] 5.997108e-03 5.964673e-03 5.754990e-03 5.442507e-03 5.379644e-03
##  [26] 5.202697e-03 5.034313e-03 4.901531e-03 4.830162e-03 4.709635e-03
##  [31] 4.666782e-03 4.555803e-03 4.482508e-03 4.381521e-03 4.292873e-03
##  [36] 4.154891e-03 3.979310e-03 3.946951e-03 3.902934e-03 3.837276e-03
##  [41] 3.702640e-03 3.571196e-03 3.549857e-03 3.517695e-03 3.448714e-03
##  [46] 3.347091e-03 3.287979e-03 3.217305e-03 3.163817e-03 3.112056e-03
##  [51] 3.042571e-03 3.023634e-03 2.984280e-03 2.950643e-03 2.910524e-03
##  [56] 2.829435e-03 2.773014e-03 2.714277e-03 2.700811e-03 2.629109e-03
##  [61] 2.526390e-03 2.512256e-03 2.492451e-03 2.479818e-03 2.395653e-03
##  [66] 2.358522e-03 2.318624e-03 2.258816e-03 2.247351e-03 2.236672e-03
##  [71] 2.193223e-03 2.102080e-03 2.095728e-03 2.035019e-03 2.007236e-03
##  [76] 1.967381e-03 1.954696e-03 1.889360e-03 1.879168e-03 1.845364e-03
##  [81] 1.834354e-03 1.751805e-03 1.686270e-03 1.667721e-03 1.611682e-03
##  [86] 1.583223e-03 1.529530e-03 1.499480e-03 1.490735e-03 1.457974e-03
##  [91] 1.433010e-03 1.416783e-03 1.350490e-03 1.323685e-03 1.298411e-03
##  [96] 1.264812e-03 1.248867e-03 1.199642e-03 1.190069e-03 1.169254e-03
## [101] 1.153888e-03 1.121171e-03 1.101499e-03 1.089197e-03 1.063745e-03
## [106] 1.052140e-03 1.011105e-03 1.004594e-03 9.708558e-04 9.183640e-04
## [111] 8.775247e-04 7.765277e-04 7.269405e-04 7.157605e-04 7.054761e-04
## [116] 6.885212e-04 6.574301e-04 6.016756e-04 5.244489e-04 5.043516e-04
## [121] 4.672321e-04 4.475665e-04 4.118352e-04 3.880592e-04 3.758124e-04
## [126] 3.538485e-04 3.409018e-04 3.175722e-04 2.609329e-04 2.562486e-04
## [131] 2.070598e-04 1.744732e-04 1.689585e-04 1.608975e-04 1.406534e-04
```

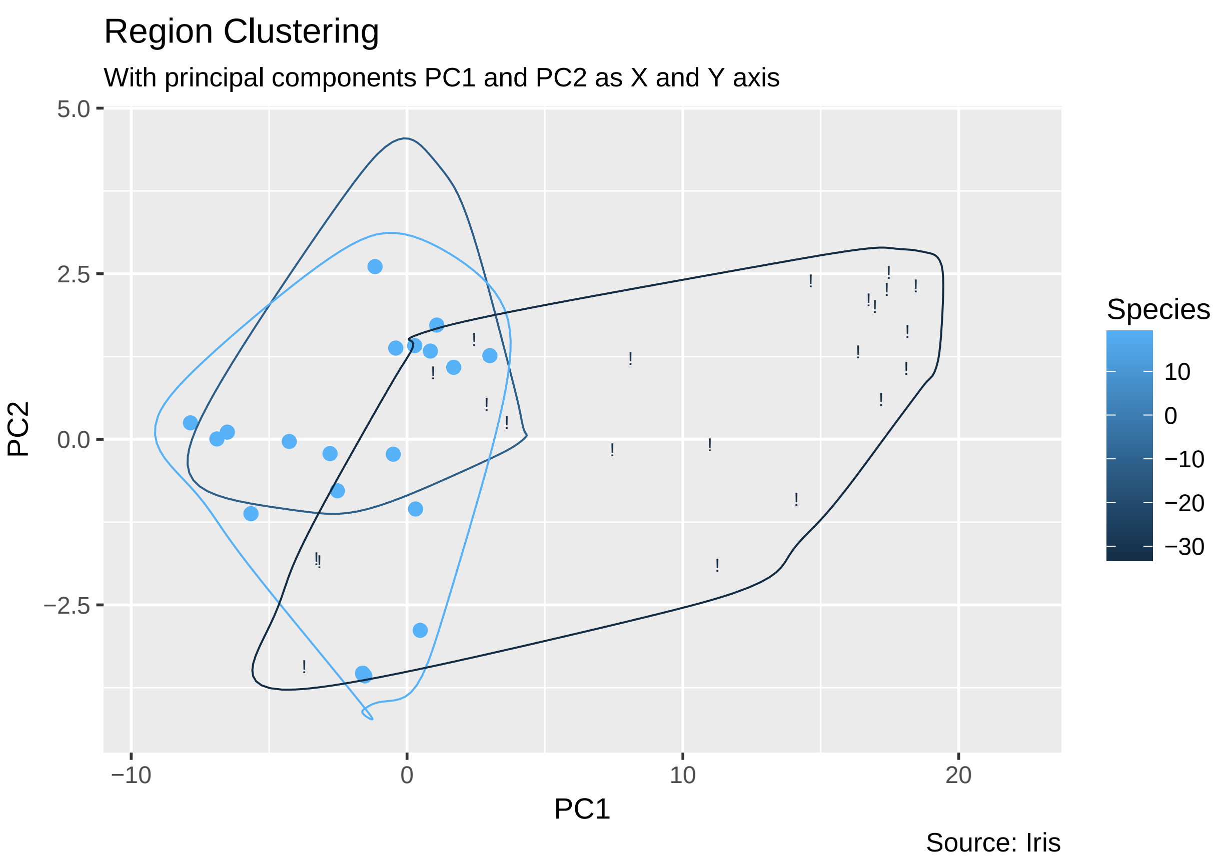
```
## [136] 1.078456e-04 8.317020e-05 5.670833e-05 4.787102e-05 1.686628e-05
## [141] 2.472614e-06
```






```
## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## width unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## metrics unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## width unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## metrics unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## width unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## metrics unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## width unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## metrics unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## width unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## metrics unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## width unknown for character 0xc
```

```
## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## metrics unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## width unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## metrics unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## width unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## metrics unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## width unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## metrics unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## width unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## metrics unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## width unknown for character 0xc

## Warning in grid.Call.graphics(L_points, x$x, x$y, x$pch, x$size): font
## metrics unknown for character 0xc
```



Lat/Long

```
load(file="int.RData")
df_5 <- df_pc[c("srch_destination_id", "PC1", "PC2", "PC3", "PC4", "PC5")]
int_datasub$srch_destination_id = as.character(int_datasub$srch_destination_id)
int_comb <- left_join(int_datasub, df_5)
```

```
## Joining, by = "srch_destination_id"
```