Business Analytics - Assignment 3 **Nicolas Tachet - nt2479**
Prof. Jacob Leshno
Due Date: 10/23/17

I submitted three files (my two R codes and this pdf with my answers). Don't hesitate to refer to my R code for more clarity.

1 DOGBARK

a) Should DogBark send pamphlets to all potential customers?

If DogBark send pamphlet to every customers (i.e. to 3293 customers), they will lose money. Indeed, only 1148 customers among the dataset are dog owners. They will bring $1148 * 0.05 * 30 - 1148$ but false positive will make the company lose 1$ each. Thus, the company will lose 1571$.

b) Load the data from dog.csv and split your sample into training (75%) and validation (25%). We will not use a test dataset for this exercise. Use the command set.seed(4650) to set the randomizer?s seed. Print the summary of the training data.

Let's print a summary of our dataset.

```
       X            dog           pub_dist          supermaket_dist
 Min.   :   1    No :2145    Min.   :   7.782    Min.   :   1.502
 1st Qu.: 824    Yes:1148    1st Qu.: 171.052    1st Qu.: 323.220
 Median :1647                Median : 409.629    Median : 546.558
 Mean   :1647                Mean   : 672.693    Mean   : 575.754
 3rd Qu.:2470                3rd Qu.: 993.679    3rd Qu.: 799.349
 Max.   :3293                Max.   :2000.000    Max.   :1746.658
  laundry_dist        park_dist        neigh_density_score
 Min.   :   1.33   Min.   :   1.018   Min.   :3.001
 1st Qu.: 196.06   1st Qu.: 630.331   1st Qu.:4.731
 Median : 537.95   Median : 919.435   Median :6.551
 Mean   : 589.30   Mean   : 962.016   Mean   :6.501
 3rd Qu.: 888.14   3rd Qu.:1257.471   3rd Qu.:8.266
 Max.   :2000.00   Max.   :2000.000   Max.   :9.995
   tree_score
 Min.   :  1.329
 1st Qu.: 30.121
 Median : 50.849
 Mean   : 63.737
 3rd Qu.:103.116
 Max.   :149.994
```

c) Find the optimal threshold (that maximize profits) for the training data. What is the confusion matrix for the training data? What would have

been be DogBar kinc.?s profit if it were to use this method to target potential customers within the training data?

We define the variable BenchmarkProfit $= 858 * 0.5 + 1611 * (-1) = -1182$ which corresponds to the situation when pamphlets are sent to every customers (where 858 is the number of dog owners).
In our code, we first try with a threshold equal to 100 for the treescore. In this scenario, the profit is equal to -298$
Then, we do the same for a threshold from 0 to 125. Using this threshold, the company is always in deficit on the train data, the best outcome is 0$ (i.e when we send nothing to the customer (no gain, no loss)).

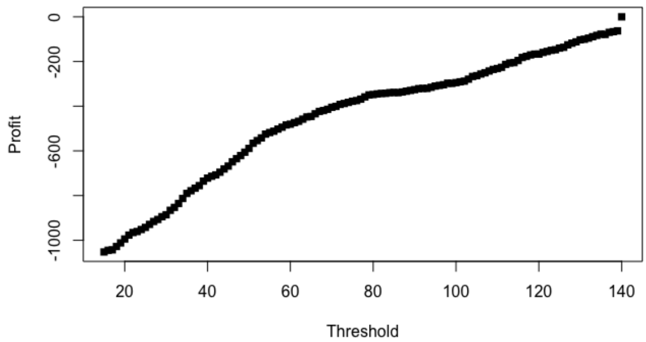

Figure 1: Profit for different values of the Threshold (test set)

d) Evaluate the classifier from the previous question on the validation data. Is the performance better or worse? Explain why.

On the validation data, our threshold strategy performs as badly as previously

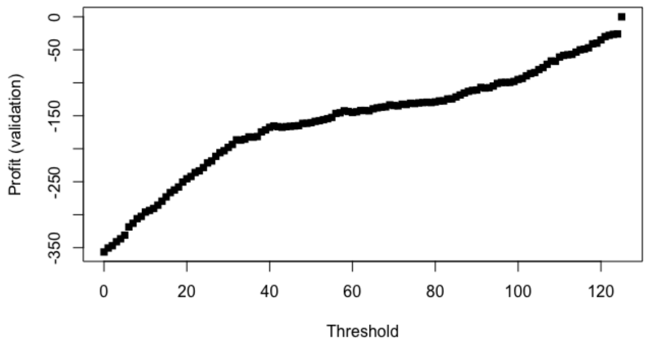

Figure 2: Profit for different values of the Threshold (validation set)

Finally, we can definitely exclude this strategy. We see that the attribute "Treescore" is not relevant to decide if a customer is a dog-owner. Indeed, having trees around your house is not an incentive to have a dog.

e) How much profits would DogBark inc. make if it had a perfect classifier?

In our dataset, there are 1148 dog owner in the dataset. Thus the max Profit is equal to $1148 * (5\%) * (30\$) - 1148\$ = 574\$$. In this case, we send pamphlet only to dog-owner.

f) Fit a logistic regression to the training data using all the covariates to predict who is a dog owner. Print the estimated coefficients and interpret them.

We fit a logistic regression on the training data. The coefficients are the following:

```
Call:
glm(formula = train$dog ~ . - X, family = binomial, data = train)

Deviance Residuals:
    Min        1Q    Median       3Q      Max
-1.3880   -0.9511   -0.7306    1.2266   2.2229

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.715e-01  2.144e-01   2.199   0.0279 *
pub_dist            -5.525e-04  7.422e-05  -7.444 9.76e-14 ***
supermaket_dist      1.801e-04  1.349e-04   1.335   0.1820
laundry_dist        -1.170e-04  9.811e-05  -1.192   0.2331
park_dist           -8.172e-04  9.179e-05  -8.903  < 2e-16 ***
neigh_density_score -9.564e-03  2.125e-02  -0.450   0.6527
```

```
tree_score              6.214e-04   1.046e-03    0.594    0.5526
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3189.4  on 2468  degrees of freedom
Residual deviance: 3037.4  on 2462  degrees of freedom
AIC: 3051.4
```

Two coefficients seem to be relevant here. The distance to the park which is not so surprising: before adopting a dog you want to make sure you're living in an environnement conducive to raise him. Surprisingly, the distance to the pub is also relevant here, this is more difficult to interpret. In addition, the residual deviance is rather important which means that this regression may not be so effective.

g) Use use the output of the logistic regression to create a classifier. What is the threshold that maximizes DogBark inc.?s profits? What is the confusion matrix?

We use a logistic regression to define a probability p "probability that the customer is a dog-owner". Then, we transform this predicted probability into a decision and test it for different threshold of probability. The maximum profit is 26.5$ and is reached for $p^* = 0.56$

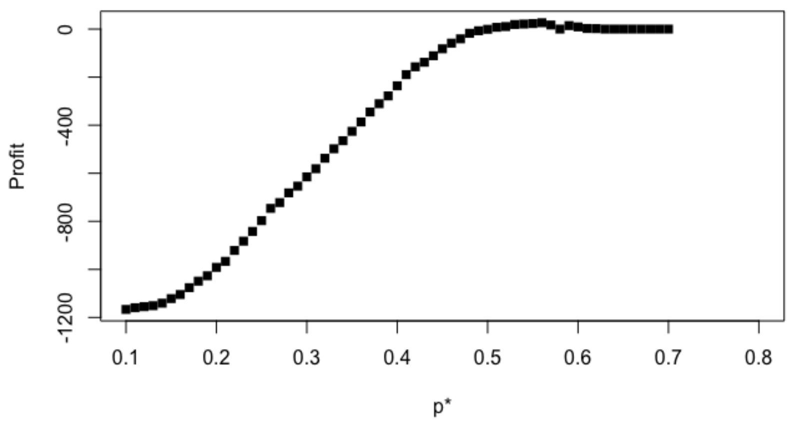

Figure 3: Profit using a logistic regression (training set)

The confusion matrix is:

```
             predict
```

truth	0	1
No	1596	15
Yes	775	83

This strategy seems to perform best than the previous one.

We are going to pick our best subset selection model which had a lower MSE than model using best subset selection. Then, we are going to predict every return for all the data we did not use to build our model (the 25% test data). Indeed, if we were using this model to predict the all data set it would be meaningless... This would mean that we are trying to forecast the data we used to build our model.

h) Fit a decision tree to the training data to predict who is a dog owner. Plot the tree. What is the confusion matrix? What would have been Dog-Bark inc.?s profit if it were to use this method to target potential customers within the training data?

The tree is the following:

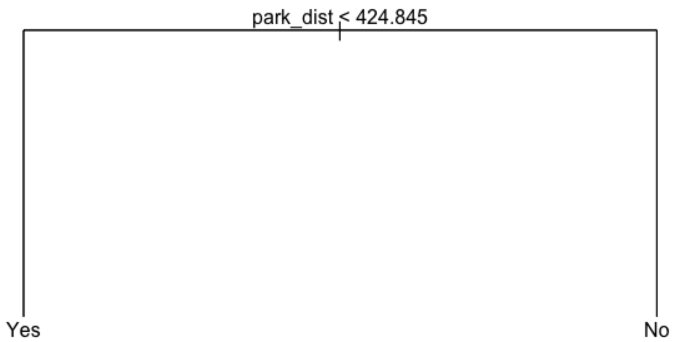

Figure 4: Profit using a logistic regression (training set)

Here is the confusion matrix:

	predict	
truth	0	1
No	1493	118
Yes	658	200

On the training data with this confusion matrix, the company's deficit is equal to 18$. This might be explain by the fact that the tree is really simple and only sort the customers according to their distance to the park. We saw that it might be a relevant coefficient. However, it cannot explain everything. This strategy does not seem appropriate.

i) Evaluate the performance of each of the classifiers on the validation data. Which method performed the best in terms of total error rate? Which method would generate the highest profits ? Which method would you recommend using?

We test each method on the validation data and compute the total error rate.

- With **the logistic regression classifier, the maximum profit is** 6$ **with a Total error rate equal to** 32.7%**.** Here is the confusion matrix:

```
confusion Matrix
       predict
truth   0    1
   No  526   8
  Yes  262   28
```

- With the **tree classifier, we have a deficit of** 4.5$ **and a total error rate of** 30.9% Here is the confusion matrix:

```
confusion Matrix
       predict
truth   0    1
   No  490   44
  Yes  211   79
```

- With the **perfect classifier, we have a profit of** 145$ **and a total error rate of** 0% **(which is logical)** Here is the confusion matrix:

```
confusion Matrix
       predict
truth   0    1
   No  534   0
  Yes  0   290
```

By analyzing these results, we see that the logistic regression might more profitable. In addition, it has a lowest total error rate. We definitely choose this strategy.

j) Suppose that DogBark inc. got stuck with a large inventory of dog toys, and wishes to change the goal of the market campaign. Instead of maximizing profits,DogBark inc. wishes to use the marketing campaign to get 1,000 purchases by sending the minimal number of pamphlets. DogBark inc. has mailing information and the indexes included in the dataset about 1,000,000 potential customers.The data in dog.csv is a representative sample of the larger dataset. Using the classifier you selected in the previous question, how many pamphlets (on average)would DogBark inc. would have to send in order

to get 1,000 purchases?

In order to get 1000 purchases, DogBark has to send on average 20,000 pamphlets to dog owners i.e. predict 'Yes' when it is effectively 'Yes' 20,000 times. First, we are going to change the threshold. Indeed, in order to have 20,000 'Yes' among 1,000,000 customers, we only need to have 66 True Positive among our 3293 customers (entire dataset). We chose $p^* = 0.578$. We have the following confusion matrix:

```
      predict
truth    0    1
  No  2134   11
  Yes 1082   66
```

Thus, to reach 66 persons we need to send 77 pamphlets. So if we want to reach 20,000 customers, we have to send $\frac{20{,}000}{66} * 77 = 23,334$ pamphlets.

2 Cuisine Preferences

a) Take the data corresponding to your section and use it as your training data. The data from the other section will serve as test data.

Here is a summary of the data:

```
> summary(cuisine)
    Italian          Mexican        Chinese...Cantonese Chinese....Sichuan
 Min.   :2.000   Min.   :1.000   Min.   :1.000       Min.   :1.000
 1st Qu.:4.000   1st Qu.:3.000   1st Qu.:3.000       1st Qu.:3.000
 Median :4.000   Median :4.000   Median :5.000       Median :4.000
 Mean   :4.172   Mean   :3.621   Mean   :4.132       Mean   :3.873
 3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:5.000       3rd Qu.:5.000
 Max.   :5.000   Max.   :5.000   Max.   :5.000       Max.   :5.000
                                 NA's   :5           NA's   :3
     Greek            Thai            Indian          French
 Min.   :2.00    Min.   :1.000   Min.   :2.000   Min.   :1.000
 1st Qu.:3.00    1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000
 Median :3.00    Median :4.000   Median :4.000   Median :4.000
 Mean   :3.34    Mean   :4.018   Mean   :3.947   Mean   :3.625
 3rd Qu.:4.00    3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:4.000
 Max.   :5.00    Max.   :5.000   Max.   :5.000   Max.   :5.000
 NA's   :11      NA's   :1       NA's   :1       NA's   :2
   Steakhouse      Ethiopian        Spanish         Carribean
 Min.   :1.000   Min.   :1.000   Min.   :2.00    Min.   :1.000
 1st Qu.:4.000   1st Qu.:3.000   1st Qu.:3.00    1st Qu.:3.000
 Median :5.000   Median :3.000   Median :4.00    Median :3.000
 Mean   :4.154   Mean   :3.261   Mean   :3.62    Mean   :3.278
 3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.00    3rd Qu.:4.000
 Max.   :5.000   Max.   :5.000   Max.   :5.00    Max.   :5.000
```

```
NA's    :6         NA's    :35        NA's    :8         NA's    :22
   Seafood            Vegan           Sushi             Pub.food
 Min.    :1.000    Min.    :1     Min.    :1.000    Min.    :1.000
 1st Qu.:4.000    1st Qu.:2      1st Qu.:4.000    1st Qu.:3.000
 Median :5.000    Median :3      Median :5.000    Median :3.000
 Mean    :4.283    Mean    :3     Mean    :4.275    Mean    :3.411
 3rd Qu.:5.000    3rd Qu.:4      3rd Qu.:5.000    3rd Qu.:4.000
 Max.    :5.000    Max.    :5     Max.    :5.000    Max.    :5.000
 NA's    :5        NA's    :6     NA's    :7        NA's    :2
   Vietnamese      Middle.Eastern
 Min.    :1.000    Min.    :2.000
 1st Qu.:3.000    1st Qu.:3.000
 Median :3.500    Median :3.000
 Mean    :3.543    Mean    :3.604
 3rd Qu.:4.000    3rd Qu.:4.250
 Max.    :5.000    Max.    :5.000
 NA's    :12       NA's    :10
```

We select Section 1 as the training set. Section 2 will serve as validation test.

b) Let us say that two students have similar cuisine preferences if their rankings agree on cuisines they both ranked. Use the Euclidian distance on common rankings to create a matrix of the similarity between each two students in your section. Add yourself to the dataset, and print the 5 closest students to you.

First, I create my dataframe of preferences:

```
# We create our dataframe of preference
myPref = data.frame(X = 'Nicolas Tachet', Italian = 5,
    Mexican = 4, Chinese...Cantonese = 1, Chinese....Sichuan = 1,
    Greek = 3, Thai = 5, Indian = 4, French = 5, Steakhouse = 4,
    Ethiopian = 2, Spanish = 4, Carribean = 2, Seafood= 5, Vegan = 1,
    Sushi = 5, Pub.food=2, Vietnamese = 4, Middle.Eastern = 2
                    )
```

Then, I defined a function to compute the euclidean distance between two students. As our dataframe is filled with NA values, I decided to compute the calculation $(x_i - x_j)^2$ distance only for non NA values. Then, I divide each distance by the number of differences I calculated (the more NA values for the two students, the smaller the number). I did this to have "homogenous" data (this is a sort of normalization). Afterwards, I created the similarity matrix and found my 5 nearest neighbors.
My 5 nearest neighbors are:

- number 36: Lars-Patrik Roeller
- number 6: Ling Dong
- number 27: Kevin Qiu

- number 41: Mathieu Nohet
- number 15: Zhi Li

c) Use the 3-NN method to complete the missing rankings in the data. Create a prediction matrix of how each student in your section will rank each cuisine.

I created the "Answer3NN" matrix. For each student, we create a table of rankings where the top rows are the most similar students. For each cuisine, we take the top ratings from the ordered matrix. These correspond to the ratings from the most similar students. Then, we predict with the average the top 3-nn ratings (from the most similar students who rated the cuisine). If there are less than total of 3-nn ratings for the cuisine, then we predict the average all of them.

d) Find the number of neighbors that minimizes the in training RMSE. Consider number of neighbors from 1 to 20 and plot the RMSE (Root MSE) for each.

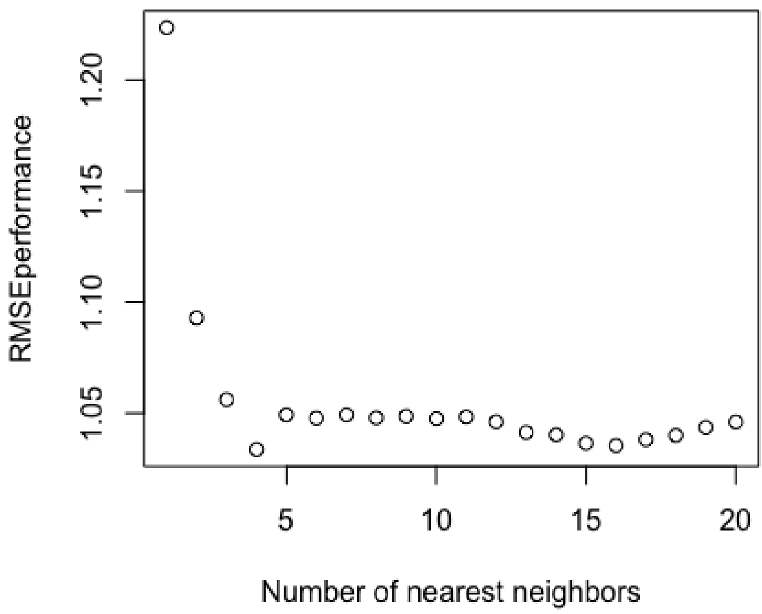

Figure 5: Evaluation of the RMSE for different values of NN

Best result is reached for **NN=4 with RMSE = 1.033670**.

e) Using the best number of neighbors you found, run K-NN to predict the cuisine choices for 3 students from the other section. What was the RMSE

of the predictions for the 3 students among all cuisines?

We want to compute the RMSE for three students in the other section. Let's say we consider Pierre Laurent, Christina Papadimitriou and Omar Abboud. The **RMSE we found was equal to** 1.247393**.**