Assignment 2 – Data Mining and Exploratory Data Analysis in R

Introduction:

The purpose of this assignment is mainly to introduce and practise the modelling and evaluation stages of the CRISP-DM model.

In this assignment you will be using the Higgs[1] and Titanic[2] datasets (on Moodle). The Higgs dataset was part of a competition posted by CERN on Kaggle.com with a prize of 13,000$. The beauty of this competition was that it did not require any knowledge of Physics and was therefore the most popular paid competition on Kaggle. The Titanic dataset is one of several datasets available on Kaggle for learning purposes.

Guidelines:

Points will be taken off for not adhering to the following guidelines:

- Use the code file from moodle. **DO NOT remove comments from it.**
- Questions will be answered in the forum only.
- Points appear next to each question in brackets and provide an indication of the question's difficulty (so if you get stuck try to move on). Your **final grade**, however, will also be affected by the **readability** of your code, so make sure to make it as concise as possible and to use meaningful names for variables.

Submission:

Deadline for submission: 24/01/16
Teams: 2 people
Form: .R code file from moodle filled in with code and answers.
Submission: .R file uploaded to Moodle. Language: English
Score: 1-100. A script that will fail execution will receive 0 points.
TA email: alonale1@mail.tau.ac.il

Useful Tips:

- You will likely need to install many packages along the way. In order to do so execute `install.packages` (look it up on google).
- If you want to know more about a certain function – execute the function with "?" before it, e.g.: `?nameOfFunction`
  At the bottom of the help page that will open you can usually find very good examples which you can copy and execute yourself and play around with. If the function's package is not installed – you can either type: '`??nameOfFunction`' (notice the two question marks) or simply use google.

GOOD LUCK!!

[1] Adapted from: https://www.kaggle.com/c/higgs-boson/data (took out weights variable)
[2] Taken from: https://www.kaggle.com/c/titanic/data

Preparations:

Set your working directory to be the assignment folder (not the nested folders).

 # HIGGS #

 # 1. PREPARATION

1.a. (1) Load the data in the appropriate format:

1.b (1) Set your random seed to some value so that our model comparisons will not be affected by randomness.

1.c (2) As the Higgs file is large, and will cause our models to run for a long time - take a sample of 50,000 rows

1.d. (3) Split the data into test (30%) and train (70%) sets with respect to the target variable.

 # 2. FEATURE SELECTION AND CORRELATION

2.a. (4) Display the correlation plot of the features. Make sure the plot is clearly visible.

2.b. (2) Find features that have a correlation of over 0.65

2.c (1) Save new data frames that will hold your train and test data, with the highly correlated features removed

 # 3. KNN

3.a (4) With the new train and test data frames - predict the test outcomes using knn, with k=1.

3.b (2) Display the confusion matrix:

3.c (5) Using cross-validation train a knn model. Use input parameters to center and scale beforehand and to have the model train a few different k's (but no more than 3). Your code should run in less than 1 minute.

3.d (2) Use the model you trained to predict the test data's labels:

 # 4. ROC & F-MEASURE

4.a. (6) Display the ROC the model you trained.

4.b. (6) Show the F1 measure (F-measure with alpha =0.5). Teams that will achieve this using an existing function for f-measure will gain all points.

 # TITANIC #

 # 5. PREPARATIONS

5.a load train file:

5.b (3) impute the missing values for the feature Age using the mean (disregard missing values of categoricals)

 # 6. LOF

6.a (4) plot the density of the LOF scores using only the following "5 features": Age, Fare, Pclass, SibSp and Parch.

6.b. (4) Based on the plot above - remove outliers above a certain LOF score threshold:

6.c. (4) add two new columns "male" and "female" which will be dummy variables presenting the sex of the passenger:

 # 7. SVM

7.a (1) split to train (70%) and test (30%):

7.b (4) Create an SVM model with as many features as possible. Your grade on this will be partially based on your error rate (computed below) and partially on your feature selection. Use it to predict the test labels and save the predictions with the name res (for results):

7.c (1) compute the error rate:

7.d (6) Tune the SVM model using no more than 5 different costs, 5 different gammas and 5 CV. Full points if you improve your error rate to < 18%.

7.e (3) display the best model found (its parametes) and use it to predict the test values - save the predictions with the name res2:

7.g (1) show if it improved by computing the new error rate:

 # 8. RANDOM FOREST

8.a (4) Create a random forest model with as many features as possible (but choose them wisely). Use no more than 2000 trees.

8.b (1) Use your model to predict the test outcome and save your predictions as resForest:

8.c (1) display the error rate:

8.d (3) Find a function that plots the importance of the variables to see how each variable, when taken out, affected the accuracy and gini measures:

 # 9. KMEANS

9.a (3) using the "5 features" (see LOF) run kmeans using 3 centers.

9.b (3) plot the clusters:

9.c (2) display the centers - do they seem to make sense?

 # 10. PCA (using your train and test data from the Higgs section)

10.a (3) Compute the train data's principal components

10.b (2) plot the drop in the variance explained by the PC's.

10.c (6) Using the PC's you've created above create two new data frames named train.h.pca and test.h.pca in which the features are replaced by PCs (our 'new features')

10.d (3) Using only the first 3 PC's - fit a simple knn model (like the first one we did) with k=7.

10.e (2) Show the confusion matrix