

BT2101 Decision Making Methods and Tools

SEMESTER I 2020-2021

Group project

Due: Friday, 20 November 2020

Instructions:

1. This is a group project. Prepare and submit only 1 report per group.
2. Download the dataset "card.csv" from Luminus Group-Project folder. You can find more information about this dataset below.
3. For each group, the project deliverables are:
 - (a) A report to be upload as pdf file into Luminus folder Project-Submission. Deadline: 11.59 pm, Friday, 20 November 2020. The report should:
 - show the names of all members of the group.
 - include description of the problem and dataset.
 - describe briefly the machine learning/statistical methods that you have tested in order to achieve a good solution to the problem. There is no requirement on the minimum or maximum number of methods that you have to test.
 - include a discussion of which method(s) you consider to be best in terms of prediction accuracy for this dataset.
 - not be more than 20 pages in length (inclusive all graphs, figures, tables, but not your code), typed with fontsize 12pt, single spacing.
 - (b) The R (or Python) code you write for this project. These are the steps on how to submit your R code using Jupyter Notebook as a pdf file:
 - Step 1: Installing Jupyter notebook by following the instructions on <https://jupyter.readthedocs.io/en/latest/install.html> See also:
 - i. [https://www.kdnuggets.com/2019/06/jupyter-notebooks-data-science-reporting.html/](https://www.kdnuggets.com/2019/06/jupyter-notebooks-data-science-reporting.html)
 - ii. <https://www.kdnuggets.com/2019/02/running-r-and-python-in-jupyter.html/>
 - Step 2: For R users, do follow the instructions on the following url so that you may use R language on Jupyter notebook <https://docs.anaconda.com/en/latest/anaconda/navigator/tutorials/create-r-environment/?highlight=create%20r%20environment%20tutorial>
 - Step 2b: For Python users, no additional steps are needed.
 - Step 3: Save your Jupyter notebook in PDF format and submit to Luminus. If in doubt, please follow the steps on using nbconvert: <http://www.blog.pythonlibrary.org/2018/10/09/how-to-export-jupyter-notebooks-into-other-formats/>

- (c) A completed peer review form to be submitted individually by each member of the group. Upload the completed form into Project-Submission folder.

Project work: to be included in your report are the following segments:

1. Brief introduction of data set and data modeling problem;
2. Exploratory data analysis: This refers to performing initial investigations on data so as to discover patterns, to spot anomalies, inconsistencies and missing values, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.
References:

- (a) <https://towardsdatascience.com/tagged/exploratory-data-analysis>
- (b) <https://www.ritchieng.com/machine-learning-project-titanic-survival/>

3. Data pre-processing: This involves transforming raw data into an understandable and usable format.

References:

- (a) Lecture 5 notes
- (b) <https://towardsdatascience.com/tagged/data-pre-processing>
- (c) <https://towardsdatascience.com/data-pre-processing-techniques-you-should-know-8954662716d6>

4. Feature selection: Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

Reference:

<https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/>

5. Model selection: See the recommended steps for parameter estimation and model selection in the following reference:

<https://www.ritchieng.com/machine-learning-project-student-intervention/>

6. Model evaluation: Evaluation metrics explain the performance of a model. An important aspect of evaluation metrics is their capability to discriminate among model results.

References:

- (a) Lecture 4 notes
- (b) <https://www.ritchieng.com/machine-learning-evaluate-classification-model/>

7. Discussion on whether there is any room for improvement.

Dataset:

1. The dataset "card.csv" has been obtained from UCI repository:

<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

It contains payment information of 30,000 credit card holders obtained from a bank in Taiwan. Each data sample is described by 23 feature attributes (columns B to X). The target feature (column Y) to be predicted is binary valued 0 (= not default) or 1 (= default).

2. Read the two 'header' lines in the data file and divide the samples for training and testing:

```
data <- read.table("card.csv", sep=",", skip=2, header=FALSE)
header <- scan("card.csv", sep=",", nlines=2, what=character())
set.seed(111)
n = length(data$V1)
index <- 1:nrow(data)
testindex <- sample(index, trunc(n)/4)
test.data <- data[testindex,]
train.data <- data[-testindex,]
```

3. Additional references for this dataset:

- <https://www.kaggle.com/search?q=credit+card+client>
- I-ChengYeh and Che-huiLien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Systems with Applications Volume 36, Issue 2, Part 1, March 2009, Pages 2473-2480.