DMS2051 – IT in Business Applications

Task 4 – Classification & Evaluation

Please submit in pdf or word document via Blackboard system by 23:59:59 on Apr 13, 2019

In this task, you are to apply classification tree to analyze data of bank clients. Your submission will be a report including descriptions of your **analysis procedure** and **results**. The assessment of the report is based on the completeness, correctness and innovativeness of your analysis, as well as the depth of the insights you generated from the results.

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification task is to predict if the client will subscribe a term deposit (variable y).

Part I – Analysis Procedure

Your analysis procedure should include the following steps:

1. Importing dataset.
2. Describing the predictors/attributes and the outcome variable.
3. Preparing for your training set and test set (using either holdout or k-fold cross validation).
4. Creating decision tree with training set.
5. Testing and reporting performance with test set.

Extra analysis steps:

6. *Pruning the tree with one or more parameters (using either manual pre-prune or grid search).*
7. *Finding out the most important feature(s) among the attributes.*

For each step above, you need to provide the snapshots of your codes and results from *Jupyter Notebooks*, and your textual descriptions of the procedure if needed.

Part II – Results

Your results should include:

1. Descriptions of the characteristics/uniqueness of the dataset.
2. Descriptions of the classification parameters and performance.
3. *[For extra step] Descriptions of the classification parameters and performance with pruned tree(s).*
4. *[For extra step] Descriptions of the most important features.*
5. Your insights/recommendations to the bank on locating their potential clients who are more likely to subscribe a term deposit.

Important notes

1. Please follow the above steps to complete your report, and submit in a pdf or word document.
2. Please provide your snapshot(s) for each step in Part I separately. Snapshots that are not placed for each step may not be assessed properly.
3. Please set the **random_state** to **100** for any sklearn function/class that involves the parameter.
4. For extra steps, you may be awarded a maximum of 1 bonus mark based on the completeness, correctness and innovativeness of your codes and report.
5. A good insight often comes from the domain knowledge and understanding to the data and the job context. So, you are recommended to read the data descriptions carefully, and obtain a deeper understanding of the classification job first.

Marking scheme

	PART I							PART II					Total
Step	**I-1**	**I-2**	**I-3**	**I-4**	**I-5**	***I-6**	***I-7**	**II-1**	**II-2**	***II-3**	***II-4**	**II-5**	
Mark	0.5	0.5	0.5	0.5	1	*0.2*	*0.2*	1	1	*0.3*	*0.5*	–	**6 + *1***