CCMVI2085U Machine Learning for Predictive Analytics in Business
International Summer University 2019-2020
Midway Mock Exam Briefing Document

Overview

This midway mock exam is not included in your course assessment. It aims to: (i) help you familiarise with the examination process, format and techniques; and (ii) give you a clearer idea of what good answers are like and how they are assessed.

You will need to create a Jupyter notebook answersheet (based on Python 3), in which you should answer the exam questions. You will need to submit:

1. Your Jupyter notebook answersheet and ensure your Python codes are clearly presented and can be executed!

2. The PDF print of your Jupyter notebook answersheet.

It should be noted that the course is assessed in terms of a 4 hour exam. Therefore, you should set the time limit of 4 hours to complete this mock exam. Each question in this briefing document has a weight and your overall grade of midway mock exam is then marked based on Danish 7-point grading scale system. For more details, please see the attached Criterion Reference Grid (CRG) document.

If you are unsure about any aspect of this midway mock exam, please seek the advice of the course coordinator Dr Bowei Chen (Email: bc.acc@cbs.dk).

Please use the following structure to create your Jupyter notebook and answer the questions.

Section 1 (15%)

Download "customer_churn.csv" dataset, and insert your Python codes below to:

1. Read the dataset into a data frame and name it df (1%)
2. Show the last 10 observations of df (2%)
3. Delete the column User_ID from df (2%)
4. Show the shape of df (2%)
5. Show the data types of each column in df and convert the object columns to categorical variables (5%)
6. Summarise the numerical variables in df (e.g., number of observations, mean, standard deviations, min value, max value, etc.) and check if the numerical variables are linearly correlated (3%)

Section 2 (15%)

Insert your Python codes below to conduct the following data pre-processing operations:

1. Create another data frame df2 which is a copy of df (2%)
2. Encode the categorical columns of df2 into binary variables (5%)
3. Create a dictionary and name it cc. The dictionary has two keys: 'data' and 'target', and the corresponding key values are NumPy arrays. For 'target', the key value is an array of values from column 'Attrition' of df2. For 'data', the key value is an array whose sub arrays of all other columns of df2. Each sub array is an observation of all other columns of df2. (8%)

Section 3 (70%)

Insert your Python codes below to:

1. Split the data (the first 80% samples for training and the rest 20% samples for testing) (5%)
2. Explain how logistic regression and artificial neural network can be used to predict customer churn (i.e., the target variable is Attrition, 500 words maximum). (10%)
3. Train both models using the training data and make predictions using testing data. You can either write your own functions or use third-party libraries such as sklearn to train and test your models. Use your allocated random seed if applicable. (15%)
4. Report the model performance in the test set by:
    - Calculating (macro) accuracy, recall, precision. (10%)
    - Draw a grouped bar chart of the above metrics. (10%)
    - Draw a ROC/AUC graph of both models. (10%)
5. Discuss what you can further do to improve your models performance (500 words maximum) (10%)