ISSU2028 Data Science and Big Data Analytics

UCL International Summer School for Undergraduates 2018

**Assessment I Computer Practical and Write-up (25%)**

**Practical session runs Thursday 2st August 2018**

**Upload report by Friday 3rd August 2018 12:30pm**

Complete the list of tasks below and write up the investigation in the form of a step-by-step tutorial that can be followed by a fellow student.

**Note that:**

* Your report should have a brief introduction setting out the context and overview of the work.
* For each investigation step you should include a brief comment on what you are doing, and why.
* For each result you should explain what it shows, how it can be interpreted, and discuss the relevance to the context of the work.
* Your report should have a brief summary evaluating the work done, and the wider context of this type of data analysis.

The expected word count for the write-up ~1000 words, with a word limit of 1250 words.

Marks will be given for the following components:

|  |  |
| --- | --- |
| **Component (details overleaf)** | **Marks** |
| Task completion | 40 |
| Use of Figures / Tables | 30 |
| Introduction, Commentary, Summary | 40 |
| Coding technique | 10 |
| Writing Standard | 10 |
| Presentation | 10 |
| **Total** | **140** |

**Notes:**

During the practical session you may work as usual and you may use textbooks, web resources, and request help from tutors.

However all code submitted must be your own work, and you are not permitted to transmit or use code files or sections from other students.

**Overview of tasks**

This assessment requires you to load and examine the dataset provided to complete the following tasks:

* Load in the Credit Card Default dataset provided and split the data into 50:50 training and test datasets.
* Explore the training data to investigate which variables may be suitable for predicting which customers will default on their loan, using suitable figures to illustrate your discussion.
* Perform a Logistic Regression fit to predict the probability of a customer defaulting based on their balance and display the fit.
* Calculate the misclassification rate for predicting customers who will default. We will use this measure as our indicator of model performance.
* Write a function that manually performs LOOCV to find an unbiased estimate for misclassification rate on our training data, and comment on the result.
* Write a for loop to apply a bootstrap analysis on the LOOCV misclassification rate. This will first generate a bootstrapped dataset using the training data, then calculate the associated LOOCV value using your function. Use the results to explore the generated distribution for misclassification rate.
* The code you have written forms a method for testing model performance and cross validating the performance measurement. Use this to explore manually test the performance of alternative predictor subsets, and choose an optimal model for classifying this dataset.
* Examine the performance of the model on the test dataset, and comment on your result.
* The bank wants to use the analysis to reduce the fraction of customers who default by 50%. Describe how they could use your model to do this.