University of St Andrews

School of Computer Science

ID5059 KDDM

Assignment: **P02**

Deadline: 16th April 2021 *Credits*: 20% of the module

You are expected to have read and understood all the information in this specification and any accompanying documents at least a week before the deadline. You must contact the lecturer regarding any queries well in advance of the deadline.

Aim / Learning objectives

This group project tests your ability to collaborate towards the construction and evaluation of machine learning models. The data is based on actual insurance data, but has been cleaned and scaled. Your group will develop a range of classifier models, and report on your assessment of their strengths and weaknesses as potential models for use with real and new data instances.

Data

You will use the Kaggle March 2021 TPS data. The task is to predict a binary class from several categorical and numeric attributes. You can assess your predictions by submitting prediction to Kaggle, who will compare against known class labels.

You can obtain the data, data description and Kaggle submission details at:

https://www.kaggle.com/c/tabular-playground-series-mar-2021/overview

The data is also available as test.csv, train.csv & sample_submission.csv on Moodle.

Instructions

1. Before making predictions, you should assess three imputation methods by removing (at random) a small percentage of the data entries, imputing the missing values, and comparing your predictions to the actual values.
2. Using the full provided data, develop and assess a range of classifiers.
3. Write a report – aimed at an imaginary client – that details your imputation results, selects a specific prediction model for use with new instances, and justifies your choice of that model. Your report should also include any variable importance insights that might improve the client's data collection strategies.

Testing and Training

The initial test and train split has been performed by Kaggle. You are free to use (and document) other approaches. The hold-out data labels are known only by Kaggle.

Key points

- You can use any combination of programming languages you like to solve the problem, but your code must be presentable and understandable.
- Presentation counts. R and Jupyter notebooks are acceptable, but must contain information easily accessible and understandable by the imaginary client (who you can assume has an educated but non-technical background).
- Group task division is to be agreed by each group. Kaggle allows a maximum group size of three for this competition, so choosing one or two members to interact with Kaggle is probably sensible.

Submission

Upload three things via Moodle:

1. The code of your solution, preferably in a Jupyter notebook with markdown annotations, or something similar built to be read with a web browser or PDF reader. Each group member submits the same code.
2. A brief, clear and concise summary describing your model, your measure of its performance, and your result, in a PDF file of at most 6 pages. Each group member submits the same report.
3. A brief (one or two paragraphs) statement of your individual contribution. If you feel that a group member or members did not contribute equally, you should note this in this document.

Assessment Criteria

Marking will follow the guidelines given in the CS school student handbook (see below).

Policies and Guidelines

Marking
See the standard mark descriptors in the School Student Handbook:
http://info.cs.st-andrews.ac.uk/student-handbook/learning-teaching/feedback.html#Mark_Descriptors

Lateness penalty
The standard penalty for late submission applies (Scheme B: 1 mark per 8 hour period, or part thereof):
http://info.cs.st-andrews.ac.uk/student-handbook/learning-teaching/assessment.html#lateness-penalties

Good academic practice
The University policy on Good Academic Practice applies:
https://www.st-andrews.ac.uk/students/rules/academicpractice/

Tom Kelsey - February 2021