ID5059 KDDM

Assignment: P02

Deadline: 1 April 2022 *Credits*: 25% 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. 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 a fraud detection dataset found here – you will need a Kaggle account to access it:

General description: https://www.kaggle.com/c/ieee-fraud-detection

The data: https://www.kaggle.com/c/ieee-fraud-detection/data

The dataset is approximately 1.3GB, consisting of 871 variables, with the *isFraud* being the thing we wish to predict from the other variables. You can assess your predictions by submitting prediction to Kaggle, who will compare against known class labels.

There is extensive discussion of the input variables that might be of some use to you. However, given the number of variables, you will not, in the time available, be able to treat these in any careful individual way: https://www.kaggle.com/c/ieee-fraud-detection/discussion/101203

Instructions

1. As a group, using the full provided data, develop and assess a range of classifiers.

2. Write brief client reports that detail any imputation, models trialled, model selected, and justification of that choice. Also include variable importance insights that might improve the client's data collection strategies *if possible* (may not be).

3. Write a brief individual technical report. This will detail a model you have fitted to the data.

Group considerations

You're free to decide how your group tackles the problem e.g. divisions of labour, however every member must fit *some* predictive model of their own for their individual report. Similar models might be adopted by different members, but given a main objective is to find a good model, it would be best if a wide range of types were considered.

It is likely a good idea to collaborate on the pre-processing of the data to get an agreed dataset for all to members to use. You might then all attempt some individual modelling - although help each other, your competition chances will be improved this way. You will have to collaborate for the assessing of the competing models (perhaps even combine/average these). You will have to collaborate at the reporting level to agree the group-level submissions.

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/markdown are acceptable throughout, but for the group submissions this 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 may limit the group size 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 team's final/best solution that you intend to submit to Kaggle. This is preferably in a Jupyter notebook or other markdown with annotations, built to be read with a web browser or PDF reader. Each group member submits the same code - include your team-name in the document.

2. A brief, clear and concise summary describing your results for your imaginary client. This should include it's measured performance; high-level description of what model was settled on (i.e. for a non-technical audience), and some speculative business case/sales pitch for using your model. Feel free to make up figures about the business cost of individual frauds and therefore what savings your model provides (once your consulting or licencing fee is deducted). Limit this to a 2-page PDF. Each group member submits this same report - include your team-name in the document.

3. A brief, clear and concise summary describing (one of) your individual modelling efforts. This is for a technical audience and should describe the model, fitting process (e.g. parameter tuning) and measures of its ultimate performance. Include a brief (one or two paragraphs) statement of your individual contribution to the group work. If you feel that a group member or members did not contribute equally, you should note this in this document. This should be no more than a 4-page PDF. Each group member submits an individual report.

Marks will be weighted 30/30/40 for these parts respectively.

The competition!

The competition winner will be determined by the Kaggle platform. Your team's best score on the leader board **at midnight of the 30th of March will** be used to determine the winner. Take a screen-shot with the upload time visible and upload this with your project submission. I will check the leader board at the finishing time, but the screenshot may be referred to in event of dispute.

Create a Kaggle team to make the submissions and use your allocated/selected team name *so I can find your team*.

Policies and Guidelines

Marking

See the standard mark descriptors for modules:

https://www.st-andrews.ac.uk/mathematics-statistics/students/taught-modules/assessment/

For this sort of assessment, see the CS descriptors as a useful guide:

https://info.cs.st-andrews.ac.uk/student-handbook/learning-teaching/feedback.html#Mark_Descriptors

Lateness penalty

The standard school penalty for late submission applies.

https://www.st-andrews.ac.uk/mathematics-statistics/students/taught-modules/late-work/

Good academic practice

The University policy on Good Academic Practice applies:

https://www.st-andrews.ac.uk/students/rules/academicpractice/

Carl Donovan

March 2022