QBUS6850 Group Project

Due dates: Monday 29 October 2018

Value: 20%

Rationale

This assignment has been designed to help students develop valuable communication and collaboration skills and to allow students to contextualise their machine learning skills on a real data from business.

Notes

1. The assignment will be done in groups of 5 (or 4 or 6 depending on the total of students in the class) without exception. The group can be formed freely or assigned by the Coordinator. Please get close contact with your members in earlier stage. A group leader for each group shall be automatically assigned on Canvas.
2. The assignment is due at **Monday 17:00pm 29 October 2018**. The late penalty for the assignment is 10% of the assigned mark per day, starting after 17:00 pm on the due date. The closing date, 5 November 2018, 17:00pm is the last date on which an assessment will be accepted for marking.
3. Your answers shall be provided as a word-processed. Prepare one single report. Do not have separate report for each question. Add your Python code as appendix to the report. At the same, we will ask you to upload your python code to your Canvas folder.
4. Your report should include the Group ID and SIDs of all group members. **No names**! You may stay with the report cover sheet provided.
5. You need to provide full explanation and interpretation of any results you obtain. Output without explanation will receive zero marks.
6. Be warned that plagiarism between individuals is always obvious to the markers of the assignment and can be easily detected by Turnitin.
7. The data sets for this assignment can be downloaded from Canvas.
8. Presentation of the assignment is part of the assignment. Markers will assign up to 10% of the mark for clarity of writing and presentation. It is recommended that you should include your Python code as appendix to your report, however you may insert small section of your code into the report for better interpretation when necessary. Think about the best and most structured way to present your work, summarise the procedures implemented, support your results/findings and prove the originality of your work.
9. Numbers with decimals should be reported to the third decimal point.
10. The report should be NOT more than 25 pages, with font size no smaller than 11pt, including everything like text, figure, tables, small sections of inserted codes etc but excluding the appendix containing Python code and the meeting minutes.

Meeting Minutes

1. Your group is required to submit meeting minutes, which are to be attached to the report as the second appendix. Your group should use the attached templates for preparing agendas and meetings minutes.
2. You should document at least 3 meeting minutes for this group assignment, using the template provided/or a template you choose.
3. In case of a problem within a group we will request minutes of the previous meetings. We can make an individual adjustment to the group mark if there is sufficient evidence that a student has done very little. If the student has truly done Nothing, we will award a mark of 0.

Peer Assessment, Marks and Feedback

1. We may ask for peer assessment from each student. The instruction how to do this will be released later on.
2. Each group will be awarded a group mark per the marking criteria. In some cases, individual marks may be applied if there is dispute in a group and the quality or quantity of contributions made by individuals are significantly different, in which cases the unit coordinator will seek peer assessments reports from individuals in a group.
3. We will allocate 15% marks for competition among the groups. The group with the highest test score will secure full 15% marks while other groups will secure a mark according to their test score against the best test score.
4. Feedback will be provided on the marked submission.

Background and Dataset

On housing market, many people struggle to get loans due to insufficient or non-existent credit histories. Home Credit **http://www.homecredit.net/** focuses on responsible lending to people with little or no credit history, who may be underserved by traditional banks. Home Credit tries to provide simple, easy, safe and fast borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

You are provided with a full set of data containing detailed information on a number of their existing customers. As part of the contract, you are asked to write a report according to the instructions given below. Home Credit will use your model to rank their clients' repayment abilities.

The dataset has been organized in four data files: `Project_Train.csv`, `Project_Test.csv`, `ProjectTrain_Bureau.csv` and `ProjectTest_Bureau.csv`. Only `Project_Train.csv` contains `TARGET` values, where 0 means "not-default" and 1 means "default". Both `ProjectTrain_Bureau.csv` and `ProjectTest_Bureau.csv` contain clients' previous credits provided by other financial institutions that were reported to Credit Bureau. This information may be helpful to build a more accurate prediction model. Please note that the previous credit has been simplified by only keeping one credit for each client.

The meaning of information (i.e. features) presented in the above dataset has been explained in `HomeCredit_columns_description.csv`. Most features are obvious according to their title names. Obviously there is no guarantee to directly use all these features to build a good prediction model. One of your tasks is to carefully choose meaningful features for the modelling task when necessary.

Tasks

Please note most tasks are deliberately designed open. This gives more freedom for you to explore a better solution.

Data Pre-processing: Conduct initial analysis over the entire data. Write python program to clean up the data, e.g., check/delete incomplete information, check missing values and their percentage, etc. It is up to you how to clean up the data so that the resulting dataset can be well incorporated in training your chosen model(s). You **MUST** retain your python program (or code section) used for cleaning up data.

Exploratory Data Analysis (EDA): You should do a thorough EDA for the given dataset. For example, which feature is categorical and what their number of type values is, any outliers, how to deal with missing values, and feature correlation analysis etc. In you report, carefully present your analysis and findings.

Benchmark Model: Build a logistic regression model to predict `TARGET`. It is always a good idea to split the given training data into a training subset and a validation subset. Thus you may validate your model building against whatever hyperparameters in the model. You may start with all the available features in `Project_Train.csv` to build the model. Something you may need to think about include data balance issue and feature suitability etc. Document your findings and justification.

Build Advanced Models: You are requested to build at least TWO advanced models such as Random Forest, Extreme Gradient Boosting, Deep Neural Networks etc. This is your choice. In building your chosen models, you need to at least optimise models in terms of e.g. the number of trees in the random forest, and/or other parameters as well. Simply building a model without any consideration of validation and hyperparameter optimisation does not meet the minimal requirement for this task. Document your findings and justification.

Feature Selection: In this part, you may consider the feature selection strategies such as using random forest to rank features and then choosing the top 20 most important features to re-build models, or when reasonable, adding new expanded/combined features from existing features to build models. Report your setting and comparisons with the models using the original full set of features.

Use of Extra Features: Clients' previous credit is one piece of important information. Explore the way on how to incorporate this information to build one new model, for example, a random forest classifier or any model you choose (note you are in competition!). Compare the results from the models you have built.

Competent Model and Final Result(s): Finally, according to your work, decide your best model and apply the model on the test data. We only ask you to report the prediction results

for their clients whose `Index_IDs` appear in `ProjectTest_Bureau.csv`. Save your result into a csv file containing two columns, one for the Client indices (`Index_ID` from the test data csv files) and the other column for the predicted 0s or 1s. Name your file as `GroupXX_Results_Bureau.csv`. The result will be assessed by our markers in order to decide your group performance among the entire class (competition!).

Note: *Given the nature of unbalanced classes, we may use the F1-score, see Lecture 7.*

Presentation

- **Please submit your project through the electronic system on Canvas**

- The assignment material to be handed in will consist of a final report that:

  i) Takes a research article form in which you shall have a number of sections such as introduction, methodology, experiment results, findings/interpretation, and conclusion. All references should be properly cited and take a full bibliographical format. Here are couple of examples
  http://cs229.stanford.edu/proj2015/007_report.pdf
  http://cs229.stanford.edu/proj2015/188_report.pdf
  http://cs229.stanford.edu/proj2015/031_report.pdf

  ii) Details ALL steps and decisions taken by the group regarding requirements above.

  iii) Demonstrates an understanding of the relevant principles of machine learning approaches used.

  iv) Includes an executive summary and also your recommended model and justifies it.

  v) Clearly and appropriately presents any relevant graphs and tables.

- The MAXIMUM page limit is 20 pages, including any computer output, graphs, tables, etc.

- Your group is required to submit meetings minutes. Your group may use the attached templates for preparing agendas and meetings minutes. You should document at least 3 meetings during semester. Documentation should be in terms of attendance, discussion points, actions decided, etc. You may use your own form or find something online.

- You, as a member of a group, may be also required to submit your peer assessment. Please use the attached criteria sheet and assessment form for this purpose. You will be informed of how to use online form when it becomes available.