Melbourne Business School
Master of Business Analytics
Machine Learning and AI for Business, Winter 2023

Machine Learning – Project Specification

Due: 5 July 2023 17:00 UTC
Weight: 25% of final mark

Introduction

You are a member of a data analytics team, working at GoggleBookSoft. GoggleBookSoft is a large company which has many data analytics teams, but is currently short of revenue and is considering a round of redundancies. As part of this process, it wishes to evaluate the teams according to their performance in tackling a supervised classification task.

Your team is provided with an anonymised dataset having 36 features and 20000 instances. The dataset is anonymised, in an effort to test the pure machine learning skill of the teams. Your task is to build a model to predict whether a given instance should have class label 1 or class label 0. You will be making predictions using a supervised machine learning method or methods of your choice. You will train your model on this provided historical data (training data) and apply predictions to a different batch of data (test data) of 20000 instances for which you will not see the class labels.

We hope that you will enjoy the project. To make it more fun we will run this task as a Kaggle in-class competition. Kaggle is one of the most popular online platforms for predictive modelling and analytics tasks. You will be competing with other syndicates in the class. The following sections give more details on data format, the use of Kaggle, and marking scheme. Your assessment will be based on your team's rank, your report and your presentation.

Data Format

You have access to three files `traindata.csv`, `testdata_nolabels.csv`, and `sample_submission.csv`. These files will be available from the competition website (see the next section). File `traindata.csv` contains 36 features for 20000 samples along with the class label and instance ID. Each row corresponds to an instance, and each column has meaning as shown in the following table.

Column Name	Meaning
ID	An integer identifier which is unique within a file
f1	The first feature
f2	The second feature
…	….
f36	The thirty-sixth feature
class	The class label. Either 0 or 1.

Next, file `testdata_nolabels.csv` contains 20000 samples with the same fields as above, but excluding the `class`, because this is what you need to predict. Note that IDs are unique only within

each file. Train and test data comprise two non-overlapping sets of samples, even though in both files there will be sample 1, 2,3 and so on.

Finally, `sample_submission.csv` is an example submission file. It has the following structure.

```
Id,score
1, 0.394383
2, 0.783099
etc.
```

The first line should be a header, exactly as shown. There should be 20000 scores each with a unique ID. The IDs of predictions should match the IDs of samples in the test file. The score is a probability indicating how likely the instance is to be from class 1. Do not round your scores.

Kaggle In-class Competition

Link: https://www.kaggle.com/competitions/90542-ml-2023

First, you will need to create a Kaggle account using your university email – submissions are allowed only from accounts with `student.unimelb.edu.au` emails. Next, all members of the same syndicate should connect themselves on Kaggle into a Kaggle team[1]. For your personal account you can choose any username you like, but the Kaggle team names must be named in the format `mbusa-2023-syndicate-XX` where XX is your syndicate number. You should only make submissions using the team name, ***individual submissions are not allowed. Note that teams will be limited to 5 submissions per day and participants will need to wait until the next UTC day after submitting the maximum number of daily submissions.***

The competition has been set up as a limited participation competition. To get access to submissions for the competition, while logged into kaggle you will need to visit the following url

https://www.kaggle.com/t/0b0fb7f82b2540f49b3553d6b7d02d37

The labels for test data are hidden from you, but were made available to Kaggle. Each time a submission is made, 40% of the predictions will be used to compute your *public* performance and determine your rank in the *public leaderboard*. This information will become available from the competition page almost immediately. At the same time, the other 60% of predictions is used to compute *private* performance and rank in the *private leaderboard*, and this information will be hidden from you. At the end of the competition, only private performance and private ranks will be used for assessment. This type of scoring is a common practice and was introduced to discourage overfitting to the public leaderboard. A good model should generalize and work well on new data, which in this case is represented by the portion of data with the hidden performance.

The performance metric used in this competition is log loss[2].

Before the end of the competition each team will need to choose 2 best submissions for scoring. These do not have to be the latest submissions. Kaggle will compute private log loss for the chosen submissions only. The best out of the two will then be automatically selected for this private log loss and the corresponding private leaderboard ranking will be used for marking. If you do not make an explicit choice of your 2 best submissions, then Kaggle will automatically choose your 2 best public scoring submissions.

---

[1] See https://www.quora.com/How-do-I-form-a-team-in-Kaggle

[2] https://www.kaggle.com/code/ikennaanigbogu/understanding-logloss/notebook

Report

Each syndicate must submit a report with description, analysis, and comparative assessment of approaches for prediction. There is no fixed template for the report, but it must include the following sections

- (2 marks) Basic summary statistics about the training data (e.g. means and standard deviations of feature values, distribution of missing values). The purpose being to give the reader an overview of what the training data is like, what it contains and what is missing.
- (5 Marks) A description of pre-processing and feature selection methods applied to the training data. E.g. data cleaning, feature removal, feature transformations and synthesis of new features. Justification and description of expected and actual effectiveness of pre-processing strategies.
- (6 marks) A description of the prediction performance of at least 3 different machine learning models for prediction, along with your motivation for trying them. Reflection on why the methods performed or didn't perform well. Comparison of the performance of the different methods to each other.
- (3 marks) A description of how hyperparameter optimization was carried out and what steps were taken to achieve an unbiased performance evaluation for models? How did you perform model selection and based on what evidence? It should be clear whether overfitting (to the training data) is thought to be an issue, with reference to evidence.
- (3 marks) What would you do differently next time in terms of techniques tried, team processes, software setup? If you had a little more time, what would you try next?

Your report should be clear and concise. You should write it at a level that a postgraduate student can read and understand without difficulty. If you use any existing algorithms, you do not have to rewrite the complete description, but must provide a summary that shows your understanding and references to the relevant literature. In the report, we will be very interested in seeing evidence of your thought processes and reasoning for choosing one algorithm over another.

The report should be a PDF file and be no more than 8 A4 pages in total (single column, font size of 11 or more and margins at least 1 cm). You do not need to include a cover page. If a report is longer than 8 pages in length, we will only read and assess the report up to page eight and ignore further pages.

Submission and Assessment

By the due date, each syndicate is required to make the following submission for this project via canvas, as a single zip file, containing the following:

- One or more submission files with predictions for test data (at Kaggle);
- Report in PDF format
- Source code used in this project. You may use any coding language you like (R, Python, ….)

The project will be marked out of 25. The mark will consist of three parts.

- **Prediction performance (3 marks):** Syndicates will be ranked by Kaggle using the private leaderboard. The mark assigned for this component will be 3*(n-R)/(n-1), where there are n syndicates and R is the rank of the syndicate. Ties are handled so that you are not penalised by the tie. For example, if team A scored best, teams B and C had the same second highest score, and team D has third highest score, then $R(A) = 1$, $R(B) = R(C) = 2$, and $R(D) = 4$. Submissions from external teams or individual students will be removed before

computing the ranks. Note that invalid submissions will come last *and* will attract a mark of 0 for the score, so please ensure your output conforms to the specified requirements, and have at least some kind of valid submission early on.

- **Report (19 marks):** See above for the marking breakdown. The report body will be assessed according to two dimensions
    - Critical analysis: Is the approach well motivated and its advantages/disadvantages clearly discussed; thorough insightful analysis of why the approach works/not works for the provided training data; insightful discussion of other approaches and why they were not used. A good motivation for choosing the approach might include either reference to literature, or reference to performance of this type of approach evaluated by others in different Kaggle competitions, or it might just be purely empirically based (what worked on the data)
    - Clarity and structure: How clear and accessible is the description of all that has been done? Is it easy to find where the information is, is there a logical progression of explanations? Definitions of any unusual jargon/terminology. Could a postgraduate student read with no difficulty? Is it easy to determine what was done (i.e. how the model(s) were trained, on what training data, with what validation data, whether cross validation was used, how hyperparameter tuning was accomplished and what hyperparameters were varied and in what range). It should be clear whether any feature transformations or generation of new features was performed (and if so, how it was done). It should be clear where to find the information explaining why one approach was preferred over the other; Enough information should be provided, that a 3rd party could reproduce the final submission(s), with minimal (e.g. 5min discussion) clarification/guidance from the syndicate. The report should adhere to formatting and report length requirements

- **Presentation (3 marks)**
    - On 6th July, each syndicate must give an 8 minute presentation from 2pm in the afternoon workshop. This should cover the following items:
        - The approach taken to develop your best performing model (including any pre-processing and any feature synthesis)
        - How you estimated the performance of this best performing model?
        - What have been the challenges?
    - The presentation mark will be based on how clear the descriptions are for each item, effectiveness of visual resources, handling of questions and adherence to time

Plagiarism policy

You are reminded that all submitted project work in this subject is to be your own individual work. It is University policy that cheating by students in any form is not permitted, and that work submitted for assessment purposes must be the independent work of the student(s) concerned. For more details, please see the policy at http://academichonesty.unimelb.edu.au/policy.html

Other Rules

You are only allowed to make submissions under the team name. No individual submissions allowed.

If requested, you must be able to demonstrate how you trained your model based on data from one or more of the files *traindata.csv and testdata_nolabels.csv*