TDT4173 Machine Learning Course Project

September 17, 2024

1 Submission

- Each group submit three things:
  - Select 2 predictions on Kaggle
  - Two short Jupyter notebooks (Short_notebook_1.ipynb and Short_notebook_2.ipynb)
    - They contain only the necessary steps to reproduce your two selected Kaggle predictions, respectively.
    - We may re-run your notebook to check whether the results are reproducible. At least one of the two short notebooks should repeat your best Kaggle score.
  - A report.
    - You can use a PDF file (Report.pdf) or a long Jupyter notebook (Report.ipynb)
    - The report summarizes all steps in your group work e.g., exploratory data analysis, feature engineering, all models/algorithms no matter they are helpful or not.
    - Remember to include model interpretation in the report.
    - We will not run your long notebook. So, if you use a Jupyter notebook as the report, you should include the running result, not only the source code.
    - There is no minimum or maximum length limit to the report.
- You should submit
  - the predictions to Kaggle
  - the notebooks and report to Blackboard
  - Please include your group number in your Kaggle team name, e.g., “[34] A Fantastic Team”
    - No group number and duplicated group numbers are treated as hacking
- You can use whatever programming languages, tools, platforms, software libraries, or file formats you want during development, but you should use Jupyter notebooks for your submission.
- Do not use separate files. Put all code in the Jupyter notebooks.
- The project deadline is November 10, 2024 at 22:00 Oslo time.
  - The deadline is strict. The competition will be closed immediately after the submission deadline. Being late in any part for even one minute in Blackboard is treated as a late submission.
  - If we cannot find your predictions, notebook(s), or report after the deadline, we will ask you to submit them again and treat them as a late submission.
- Late submission deadline is November 14, 2024 at 14:00 Oslo time.
  - A late submission is not recommended. It will cause a -30 deduction in your course points.
  - If you submit even later (i.e., after the late submission deadline), the whole group fail the course.
  - Late predictions will be submitted to another Kaggle link (announced later)

- Most excuses, e.g., short-term sickness and other activities, are not accepted for a deadline extension. Start early for the project work.
- At the beginning of each notebook and the report, write your full names, student ids, and Kaggle team name. Student ID is the one on your student card, not the candidate number in the course.

2 Grading

- Project points = base points + possible deductions
- Convert points to letter grade
  - A: 89-100
  - B: 77-88
  - C: 65-76
  - D: 53-64
  - E: 41-52
  - F: 0-40
- Base points
  - proportional to the number of Virtual Teams (VTs) you defeat
  - max 100 (defeat all VTs) and min 41 (defeat 1 VT)
  - if you defeat 0 VT, you fail the course
- Virtual Teams (VTs)
  - VTs are prepared by the teachers and assistants
  - A VT uses the same data as you have
  - All VTs will be frozen after 20/October/2024
- We use the **private leaderboard** to calculate base points
  - You can see only the public leaderboard before the competition closes
  - You select two predictions before the deadline (your short notebooks should be able to reproduce the selected predictions)
  - We use the best between your selected in the private leaderboard to calculate your base points
  - Notice: Ranking in the public leaderboard has no effect in grading, and it is possible that there is (big) difference between the two leaderboards.
- More about the reproducibility
  - No external data is allowed.
  - Your short notebooks will be run in an offline setting
  - Your short notebooks should start from the given raw data
    - ∗ You can include a configuration file (e.g., requirements.txt or environment.yaml) to specify the required software packages and their versions.
    - ∗ Before running your code, we will delete everything from you except the two short notebooks and the configuration file.
    - ∗ During the running, your program can store some middle results as disk files.
  - You can use a few programming constants, but writing massive data in the code and in the configuration file is not allowed.

  - The maximum running time for a short notebook is 12 hours on a normal PC.
  - You can directly specify a few hyperparameter values in the short notebooks, but you must describe how you obtain the values in the report.
  - The short notebooks should only have the relevant parts to generate your selected Kaggle predictions.
  - A small difference (not affecting your ranking against VTs) is acceptable. Otherwise, we will run it once more and pick the best among our multiple runs for grading.

- Possible deductions

  - pass the individual assignment in the second chance (-5); the deduction applies to individuals, not the whole team.
  - late submission (between project deadline and late submission deadline) of the project (-30)
  - no exploratory data analysis (-3). To avoid the deduction, you should do at least four or more items of the following list:
    - Search domain knowledge
    - Check if the data is intuitive
    - Understand how the data was generated
    - Explore individual features
    - Explore pairs and groups of features
    - Clean up features
  - only one type of predictor is used (-3). To avoid the deduction, you should show that you have tried two or more types of predictors in the report (e.g., XGBoost and Random Forest). It is allowed to use only one type of predictor in a short notebook for Kaggle predictions.
  - no feature engineering (-3). To avoid the deduction, you should show that you have tried one or more feature engineering techniques (i.e., feature selection and/or feature extraction) in the report.
  - no model interpretation (-3). To avoid the deduction, you should show one or more model interpretation results (e.g., PDP, feature importance, LIME) in the report.
  - Please use clear section titles in the report for easy lookup of the above parts.

- All deductions are binary. That is, there are only two possibilities: full deduction and no deduction. There is no intermediate deduction.

- A grading example: a student

  - passes the individual assignment in the first chance
  - submits the project in time
  - his or her team defeats 7 of 10 VTs
  - no model interpretation in the notebooks (-3)
  - Then the student's course points
    - base points $= 41 + \frac{100-41}{10-1} \times (7-1) \approx 80$
    - course points $= 80 - 3 = 77$
    - The corresponding letter grade is B.