Final Project

TA email: `html_ta@csie.ntu.edu.tw`

RELEASE DATE: 10/16/2024

REPORT DUE DATE: **12/23/2024 13:00**

Unless granted by the instructor in advance, no late submissions will be allowed. That is, you will not be allowed to submit your report after the deadline and will get zero point for the final project. The gold medals cannot be used for the final project.

Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.

You need to write your report in English with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.

Introduction

Major League Baseball (MLB) is the premier professional baseball league in North America, boasting influential commercial value with billions in annual revenue and a global fanbase of billions, influencing sports culture and media worldwide. In MLB, each game stirs excitement and passion. For many fans, the game's result goes beyond entertainment—it becomes an emotional investment tied to their sense of identity, pride, and even national loyalty. Various surrounding activities, such as the virtual games on the famous Fantasy Baseball platform, are attracting the fans' attention too. Predicting the winner in the virtual games can be as important as predicting the winner for the physical games, as both an intellectualism and a business. In the past, professionals have conducted in-depth studies on predicting game outcomes, yet it is a complex field involving multiple factors and stakeholders.

The rise of AI offers a glimmer of hope in tackling this complex problem. Imagine today you are working for a "Game Prediction Company," and your task is to develop a machine learning model to predict game outcomes. Because your companies have not signed any contracts with MLB nor Fantasy Baseball to use their data officially, your boss decided to set up a virtual platform of "Hyper Thrill Machine Learning Baseball" (HTMLB), and collect the data on the platform instead. Each prediction can significantly impact the commercial value of future baseball teams, national identity, and even local public safety, making it an INCREDIBLY CRITICAL job. Please collaborate with your colleagues (team members) to write a report to convince the company's executives (who may not all have CS backgrounds) that your work is worth pursuing.

Data Set

The dataset is available on Kaggle. It is a pseudo dataset collected on HTMLB, which has been verified to share some distributional similarity with the MLB historic game data. The training set contains HTMLB game information from January to July, from 2016 to 2023. We will consider two stages of tasks. In the first stage, you are asked to predict the game results from August to December, from 2016 to 2023. In the second stage, you are asked to predict the results of all games in 2024. The following links will work only after you sign up for the competitions via the links provided in the course announcement.

- First stage: `https://www.kaggle.com/competitions/html-2024-fall-final-project-stage-1`
- Second stage: `https://www.kaggle.com/competitions/html-2024-fall-final-project-stage-2`

To ensure fairness, you are strongly prohibited to acquire any external data from MLB, Fantasy Baseball, or any other sources. Your model can only use the given data during training and testing.

Evaluation

For each stage, you are asked to predict whether the home team wins or not, which is a binary classification problem. For both stages, the deadline of submission is **2024/12/15 23:59, UTC+8** to qualify for the award ceremony on 2024/12/16. The submission site will remain open until the report deadline.

Let $w_{i,t}$ be the ground-truth of win-or-lose indicator of home-team i on date t , and $\hat{w}_{i,t}$ be your predicted result (same format as $w_{i,t}$). The point-wise error is defined as the 0/1 error of $w_{i,t}$

$$\text{err}(\hat{w}_{i,t}, w_{i,t}) = [\![\hat{w}_{i,t} \neq w_{i,t}]\!] .$$

Then, for any set, we can calculate its average err over all examples.

Survey Report

You need to study at least FOUR machine learning approaches using the data. Then, you should make a comparison of those approaches according to some different perspectives, such as (but not limited to) accuracy, stability across the two stages, efficiency, scalability, and interpretability. Then, you need to recommend THE BEST ONE of those approaches as your final recommendation and provide the "cons and pros" of the choice. The report from the study is considered the most important and incredibly exciting part of our final project.

The survey report should be no more than SEVEN A4 pages with readable font sizes. The most important criterion for evaluating your report is reproducibility. Thus, in addition to the outlines above, you should also describe how you pre-process your data, such as the features you build; introduce the approaches you tried and provide specific references, especially for those approaches that we did not cover in class; list your experimental settings and the parameters you used (or chose) clearly. Other criteria for evaluating your survey report would include, but are not limited to, clarity, strength of your reasoning, "correctness" in using machine learning techniques, the work loads of team members, and proper citations.

Our sincere suggestion: *Think of your TAs as your boss who wants to be convinced by your report.*

For grading purposes, a minor but required part in your survey report for a two-, three- or four-people team (see the rules below) is how you balance your work loads.

Competition

The submission site will be based on Kaggle. The maximum team size is 4 people. Please follow the steps to sign up for your team:

(1) Submit your team information by `https://forms.gle/9RWAn3UwZsKBTyeR6` **before 2024/10/28 23:59.**

(2) Join the competitions with the link in the NTU COOL announcement (not the links above)

(3) Form your team on Kaggle

(4) Start submitting

Each team is allowed up to five submissions per day per stage. The performance of these submissions will first be evaluated on approximately 50% of the test set (the public test set). Before the submission deadline of 2024/12/15 23:59 UTC+8, teams must select two final submissions for each stage, which will then be evaluated on the remaining 50% of the test set (the private test set) after the competition ends.

Please ensure that you form a single team on Kaggle and refrain from using multiple team accounts as this is an unfair advantage over the other teams. In addition, as mentioned, **you are strongly**

prohibited to acquire any external data from any external sources—which, if proven true, is a violation of the fairness policy. If anything suspicious is discovered and confirmed, you will be kicked out of your company (and hence fail the final project and very possibly the class). We reserve the right to request a physical demonstration of your training and submission process up to one month after the report deadline if needed.

Misc Rules

Report: Please upload one report per team electronically on Gradescope. You do not need to submit a hard-copy. Your team name on your report, your submitted form, and the Kaggle website must be the same. The report is due at 13:00 on 2024/12/23. **The gold medals cannot be used for the final project.**

Teams: By default, you are asked to work as a team of size FOUR. A one-person, two-people or three-people team is allowed only if you are willing to be as good as a four-people team. It is expected that all team members share balanced work loads. Any form of unfairness, such as the intention to cover other members' work or any unethical trading, is considered a violation of the honesty policy and will cause some or all members to receive lower, zero or even negative scores.

Algorithms: You can use any algorithms, regardless of whether they were taught in class.

Packages: You can use any software packages for the purpose of experiments, but please provide proper references in your report for reproducibility.

Source Code: You do not need to upload your source code for the final project. Nevertheless, please keep your source code until 2025/01/31 for the graders' possible inspections or demo requests.

Grade: The final project is worth 800 points. That is, it is equivalent to 4 usual homework sets. At least 720 of them would be reserved for the report. The other 80 may depend on some minor criteria such as your competition results, your work loads, etc.

Collaboration: The general collaboration policy applies. In addition to the competitions, we still encourage collaborations and discussions between different teams.

Deadlines

- Team sign-up deadline: 2024/10/28 23:59 UTC+8
- First stage submission deadline: 2024/12/15 23:59 UTC+8
- Second stage submission deadline: 2024/12/15 23:59 UTC+8
- Report deadline: 2024/12/23 13:00 UTC+8
- Please keep your source code until 2025/01/31