Mining Media Data II
Summer Semester 25
Assignment-1

Prof. Dr. Rafet Sifa, Dr. Lorenz Sparrenberg, Maren Pielka

April 10, 2025

Introduction

In this assignment, we will get ourselves familiarized with some of the important tools as well as the datasets we will be utilizing during our course. This assignment has 50 Pts. in total with a bonus part of 15 for the Large Language Models (LLMs) part. All programming assignments will be implemented using the Python 3.10+ programming language.

1. Preparing a telemetric dataset for lifetime value analysis and building churn predictors (50 Pts.)

Download the two telemetry datasets from `https://tinyurl.com/pchurndatasets`[1] about two freemium mobile racing games and prepare them for further analysis. In this part we will prepare the datasets and build churn predictors[1, 2, 3] that will be evaluated using the following steps:

- Structure the data set (for the first game as the second game is in this format) in a JSONL format so that every line is a JSON file containing all the events of a particular player only. Create two disjoint splits of the dataset where 20% of the players will be saved for evaluation of the models.
- Afterwards, create a supervised data set from 80% of the players to predict churn (will be called DS1) as explained in [2] with observation period of 5 days and a prediction (churning) period of 10 days using the time information in the logs. Use the information on the occurrence in the prediction period to label each player if they churned or not.
- Extract features from DS1 considering the observation period (see [2] for examples).
- Train at least 3 different classifiers (one of the methods should be the decision tree classifier) to predict churn.
- Create a new dataset with the remaining 20% of the players and apply the same feature extraction steps as in above (the dataset will called DS2).
- Evaluate the above trained classifiers (note that training done on DS1) using DS2 and report your results. What were important features to predict churn. Which classifier worked better?
- Use an LLM of your choosing to predict churn in a zero shot manner. Namely, create a prompt for each user containing information about the user as well as other necessary information (you may include information about the game for instance using what is provided in [2]) and ask the LLM whether it considers the player to be churning or not. Evaluate this modeling approach using DS2 and compare the results to the above results.

2. Presenting the results

The results will be presented as a Jupyter Notebook or a PDF with an accompanying code base.

[1] For this assignment,we will be using only the **first** and the **second** game from [2], not the third one. The work also provides documentation for the datasets.

They will be discussed in an exercise session. Powepoint presentations are welcome!

4. Submission

All submissions will be made electronically by sending *a single .zip* file (including your Python code and PDF or Jupyter Notebook) to `amllab@bit.uni-bonn.de` by the submission deadline with the title **MMD SS2025 Assignment 1 [GroupID]**, where **[GroupID]** refers to your group id (name).

Submissions sent after the deadline and not following the title convention will not be evaluated. The submission deadline for this assignment is on 03.06.2025 at 11:59 am Germany time. Further updates about this assignment will be published on the course website.

5. A Note on Plagiarism

Work containing plagiarism will not be graded. After a second warning, a disciplinary process will be started and the corresponding students will not be allowed to attend the final examination.

References

[1] Fabian Hadiji, Rafet Sifa, Anders Drachen, Christian Thurau, Kristian Kersting, and Christian Bauckhage. Predicting player churn in the wild. In *2014 ieee conference on computational intelligence and games*, pages 1–8. Ieee, 2014.

[2] Seungwook Kim, Daeyoung Choi, Eunjung Lee, and Wonjong Rhee. Churn prediction of mobile and online casual games using play log data. *PloS one*, 12(7):e0180735, 2017.

[3] Rafet Sifa. Predicting player churn with echo state networks. 2021.