Student Grade Prediction

Objective

The objective of this hackathon is to utilise a given dataset containing student logs to create appropriate features, train a suitable machine learning model, and predict the final grades of students based on their logs.

Dataset

1. The dataset will be provided to participants at the beginning of the hackathon.
2. The dataset comprises comprehensive student logs, encompassing diverse attributes, including the timestamp of their access to specific resources on Moodle.
3. The dataset will be provided in a structured format, specifically as CSV files, where each row corresponds to a student's specific log entry.

Tasks

Task 1: Data Exploration and Feature Engineering

1. Participants should perform data exploration to understand the dataset, its structure, and the meaning of each attribute.
2. Participants should create appropriate features that might help in predicting the final student's grade.

Participants are encouraged to choose a suitable time unit (e.g., weeks or months) according to their analysis requirements. Additionally, participants are welcome to utilize the `melt`, `dcast`, and/or `pivot_table` functions to reshape their data. You can refer to the R documentation and Pandas documentation for more information on how to use these functions effectively. Below are a few examples demonstrating the usage of these functions in both R and Python programming languages.

R

```
> library(data.table)
> dt
    metric    id   year   value
1: tuition   id1   2015       1
2: tuition   id2   2016       2
3: tuition   id3   2017       3
4: admitsize id1   2015       4
5: admitsize id2   2016       5
6: admitsize id3   2017       6
7: avgfinaid id1   2015       7
8: avgfinaid id2   2016       8
9: avgfinaid id3   2017       9

> dcast(dt, id+year ~ metric, value.var = "value", fun=max)

   id   year tuition admitsize avgfinaid
1: id1  2015       1         4         7
2: id2  2016       2         5         8
3: id3  2017       3         6         9
```

Python

```
>>> import pandas as pd
>>> df

   school  year     metric  values
1     id1  2015    tuition       1
2     id1  2015  admitsize       2
3     id1  2015  avgfinaid       3
4     id1  2016    tuition       4
5     id1  2016  admitsize       5
6     id1  2016  avgfinaid       6
7     id2  2015    tuition       7
8     id2  2015  admitsize       8
9     id2  2015  avgfinaid       9
10    id2  2016    tuition      10
11    id2  2016  admitsize      11
12    id2  2016  avgfinaid      12

>>> df2 = ( df.pivot_table(index=['school', 'year'], columns='metric', values='values').reset_index() )
>>> df2

   metric school  year  admitsize  avgfinaid  tuition
0           id1  2015          2          3        1
1           id1  2016          5          6        4
2           id2  2015          8          9        7
3           id2  2016         11         12       10
```

Task 2: Model Training and Evaluation

1. Participants should split the preprocessed dataset into training and test sets.
2. Participants should select an appropriate machine learning model for grade prediction (e.g., classification, or ensemble models).
3. Participants should fine-tune the model if necessary.
4. Participants should test their model.

Task 3: Grade Prediction

1. Participants should generate predictions for the provided unseen dataset.
2. Participants should submit their predictions of the unseen data in a specified format (i.e., a single CSV file per group).
3. Participants should submit their notebooks in a specified format (i.e., a single PDF file per group).

Please refer to the submission folder on UniHub for more details.

Evaluation Criteria

The accuracy of the grade predictions will be a significant factor in evaluating the submissions.

Discussion

The top four groups achieving the highest accuracy will be selected to present a concise explanation (5 min each) of their approach and the insights gained from their predictions.

During these presentations, groups will have the opportunity to discuss the following points:

1. *Feature Engineering*: The originality and effectiveness of the features created by the group.
2. *Model Performance*: The evaluation metrics used to assess the performance of their machine learning model.
3. *Machine Learning Model Pipeline*: The proposed machine learning models and the rationale behind their selection.
4. *Documentation*: The clarity and comprehensiveness of the approach, covering data preprocessing, model selection, training, and prediction methodologies.

Grand Finale

Two groups will be chosen based on their exceptional discussion skills. They will engage in a *challenging debate*, presenting compelling arguments to support why their approach was superior. The participant that convincingly justifies their methodology as the best will be crowned the winner and receive the ultimate recognition and award.

Prizes and Recognition

The winner will receive a nice price from us...

Timeline

- 1:45 - 2:00 pm: Meet in front of H207 & H208 rooms.
- 2:00 - 2:20 pm: Explanation of the competition and introduction to the challenge.
- 2:20 - 2:30 pm: Formation of groups with 4-5 students each.
- 2:30 - 4:50 pm: You work on the hackathon.
- 4:50 - 5:00 pm: End of the hackathon, you submit your predictions.
- 5:00 - 5:30 pm: The teams with the most accurate predictions will discuss their solutions.
- 5:30 - 6:00 pm: Announcement of the competition winner.