DS 310 Machine Learning

Project #2: TMDB Box Office Prediction

Fall 2020 / Amulya Yadav

Task: In a world, where movies made an estimated $41.7 billion in 2018, the film industry is more popular than ever. But what movies make the most money at the box office? How much does a director matter? Or the budget?

In this project, you're presented with metadata on over 7,000 past films from The Movie Database to try and predict their overall worldwide box office revenue. Data points provided include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries.

Dataset: You can download the dataset on Canvas.

The dataset consists of the following items, after unzipping:

- `train.csv` : a training data in the CSV format
- `test.csv` : a test data in the CSV format
- `sample_submission.csv` : a submission sample in the CSV format

In this dataset, you are provided with 7398 movies and a variety of metadata obtained from The Movie Database (TMDB). Movies are labeled with `id`. Data points include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries.

You are predicting the worldwide revenue for 4398 movies in the `test` file.

Kaggle

After teams generate the prediction results for the test data, you can post it to the Kaggle to get the feedback on the accuracy of their model:

https://www.kaggle.com/c/tmdb-box-office-prediction/overview

Submissions are evaluated on Root-Mean-Squared-Logarithmic-Error (RMSLE) between the predicted value and the actual revenue. Logs are taken to not overweight blockbuster revenue movies.

Deliverables: Each team needs to accomplish the following tasks:

1. Kaggle Leaderboard: At the end, we take the best ranked entry (performance) of each team's submission against `test.csv`, in the leaderboard as the team's final submission result.

2. Jupyter notebook: Each team posts its final model in Jupyter notebook, named as `submission.ipynb`.

3. Project Report: Project report in PDF that contains all the details of the major steps of the project.

Grading Rubric (100 points)

- 80 points: performance (65 points) and ranking (15 points) in the Kaggle Leaderboard, using each team's best `submission.csv` file. 50 points for baseline performance and 15 points for better performance.
- 10 points: submitted Jupyter notebook, `submission.ipynb`.
- 10 points: quality of report.

Grading Exceptions

The goal of this project is to encourage you to emerge yourself in a real-life data science process to learn the materials better in a hands-on fashion. **DO NOT CHEAT IN ANY WAYS.** If you do not attempt to get your hands dirty in this project, you will not learn much.

All team members need to participate in the project and contribute as much as he/she could. Each is expected to bring different skills to the table, i.e, some are good in programming, while others in analytics, etc. Find out what you can contribute within the team, and play the team work. While both members get the same points by default, I will intervene in an exceptional situation (e.g., one member does not participate at all).