Final Report Proposal

TOPIC: Predicting Movie Success Using Machine Learning

WHY THIS PROJECT?

- This project will allow me to practice machine learning techniques, I covered only theoretically before, in a domain I am passionate about: movies and TV shows.
- By building and tuning ML models, I will strengthen my understanding of traditional ML methods, covered in ML course I completed previous semester as well as first weeks this course.
- Collecting my own data (*see below*) ensures the flexibility to explore creative solutions and necessary features, as opposed to using pre-existing datasets (*i.e. Kaggle*) where much of the work is already done.

OBJECTIVE: The objective of this project is to collect data from IMDB (*via API or web scraping using BeautifulSoup*) to create a dataset of popular movies and TV shows. The goal is to build a machine learning model that can predict the success of a movie, either through its IMDB rating or box office performance, based on various features (*e.g., year, director, actors, genre, budget, country, duration,…*). The project will focus on using basic machine learning techniques like Decision Trees, and advanced techniques like Boosting (*LightGBM, AdaBoost, CatBoost*, …) or Bagging (*Random Forests, ensembles*, …) to train and compare models.

MOTIVATION: Having recently completed a theoretical machine learning course as well as graduated with a Master's degree in Data Science (*few years before*), I aim to apply the knowledge learned in a practical setting. My past experience includes using CNNs for my thesis work (https://www.hse.ru/en/edu/vkr/837853291), and while deep learning techniques could be explored, I prefer to focus now on basic machine learning models in this project. I believe this will help deepen my understanding of traditional ML methods and their practical applications, which I haven’t explored as much.

PLAN:

1. **Data Collection:**
    - Use the IMDB API or web scraping tools to gather data on famous movies, TV shows, actors, and their corresponding IMDB scores, connections (Movie-to-Actor) or box office performance.

   - Gather relevant features for prediction such as movie year, director, actors, genre, budget, country, duration, ….
   - I could start with top 10 most famous movies, get top 10 actors from them and continue that as needed (*10 top movies from each actor and top 10 actors from those movies, …)*

2. **Dataset Creation:**
   - Structure the collected data into a clean and comprehensive dataset suitable for machine learning tasks (*saved as pickle object of pandas DataFrames or as simple serverless database, like DuckDB, or flat files, like parquets which could load data directly into pandas*).
   - The dataset will include features that could influence a movie's success, and this feature set will allow flexibility for future exploration *(e.g., switching between predicting IMDB ratings and box office numbers*).

3. **Model Building and Comparison:**
   - Use machine learning models to predict movie success:
     - **Simple ML models (*ski-learn*):** Decision Trees, Random Forest.
     - **Advanced techniques (*as separate libraries*):** Boosting methods like LightGBM, AdaBoost, CatBoost.
     - **Maybe to try some ensembles as well (*optional*)**
   - Evaluate model performance using metrics like balanced accuracy, MAE (Mean Absolute Error), R-squared, f1-score, etc. depending on the prediction task (*classification for ratings, regression for box office*).

4. **Tools & Techniques:**
   - I will focus on using traditional ML models rather than Neural Networks since I have already worked with CNNs extensively for my thesis and I haven't completed DL, RL, etc. UT Austin courses yet. I would like to gain a deeper practical understanding of ML fundamentals.
   - I intend to collect and preprocess my own dataset instead of using a pre-existing Kaggle dataset. This allows me to explore creative approaches, contribute unique insights, and make the research more personalized.

Expected Outcomes:

- A comprehensive dataset of movies and TV shows with extracted features (*as well as functions to extract it*).
- A comparison of different machine learning models for predicting movie success based on multiple features.

- Insights into which features most strongly influence a movie's rating or box office performance (*also with applications of identicality reductions, like PCM or t-SNE*).

CONCLUSION**:** This project will enhance my understanding of practical machine learning applications, data collection, and model building. It will allow me to explore multiple modeling techniques and better understand their strengths and weaknesses when applied to real-world problems.

Thanks for your time reading this and thanks in advance for any comments! 😊