Project 4 Proposal

Project Title: *Unlocking Box Office Success: Predicting Movie Popularity*

Group Members:

- Gursimran Kaur (Simran)
- Jeff Kim
- Rose Mary Rios

Tools and Technologies

- **Programming Language(s)**: Python
- **Data Storing**: CSV files
- **Libraries**: Pandas, NumPy, Seaborn, Matplotlib, Scikit-Learn, Json
- **Software**: Jupyter Notebook, Visual Studio Code

Introduction

The movie industry operates within a complex landscape, where accurately gauging a film's potential popularity is essential for filmmakers, producers, and studios. The success of a movie is influenced by various factors such as budget, genre, cast, runtime, and critical reception. With the wealth of available movie data, including detailed metadata and user ratings, the ability to predict a film's popularity has become a focus of interest for decision-makers. Leveraging advanced machine learning techniques offers the opportunity to make informed, data-driven predictions about a movie's potential success.

This project aims to develop two machine learning models—**Linear Regression** and **Random Forest**—to predict a movie's popularity score. Using a combination of features like budget, runtime, genre, and user ratings, the models aim to uncover patterns that drive audience engagement and success. The proposal also integrates data engineering practices to ensure the data is thoroughly prepared for analysis, along with exploratory data analysis (EDA) to reveal hidden insights. Through model optimization, the project aims to deliver high-performance predictions that provide stakeholders with actionable insights to inform production, release strategies, and marketing decisions.

Purpose

The goal of this project is to develop a predictive model that enables stakeholders in the movie industry to estimate a film's potential popularity prior to its release. By utilizing historical movie metadata and Oscars data, the model will provide predictions that highlight the key attributes influencing a film's success.

Research Questions

- What factors (e.g., budget, genre, runtime, award) have the most significant impact on a movie's popularity?
- Can machine learning models accurately predict a movie's popularity score based on historical metadata and user ratings?
- How do different machine learning models (Linear Regression vs. Random Forest) compare in predicting movie popularity?
- What level of predictive accuracy can be achieved, and what insights can be derived from the model's predictions?

Objectives and Goals

- **Develop Predictive Models:** Build machine learning models, specifically a Linear Regression and Random Forest Regressor to forecast movie popularity scores based on movie metadata such as budget, runtime, genre, and user ratings.
- **Analyze Movie Features:** Measure the influence of various movie features (e.g., budget, genre, runtime) on popularity scores and identify the key factors that drive higher ratings.

Data Sources

- **TMDB 5000 Movie Dataset:** https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata
- **Oscar Best Picture Movies:** https://www.kaggle.com/datasets/martinmraz07/oscar-movies

Methodology

The methodology for this project will be broken down into several phases:

I. Data Collection

- Retrieve data from various sources, including credits.csv, movies.csv, and oscars.csv.

II. ETL

- **Handle Missing Values**: Clean the datasets to manage missing data.
- **Data Normalization and Scaling**: Prepare features for modeling.

- **Feature Engineering**:
    - **Genre**: Convert genres into one-hot encoding to make them machine-readable.
    - **Weighted Rating**: Use a weighted score based on average votes and vote counts.
    - **Release Date**: Create a release_year/day/month feature to see how movies perform over time.

III. EDA

- **Correlation Analysis**: Explore relationships between features using correlation heatmap(s).
- **Visualizations**:
    - Histograms of all numerical column distributions and feature relationships.
    - Scatter plots of various features visualizing their relationship with popularity.
    - Box plot comparing vote_average across different genres.
    - Box plot comparing revenue distribution by award status.

IV. Machine Learning Model Implementation

- **Feature Selection**: Use feature importance analysis to select relevant features.
- **Model Development**:
    - **Linear Regression:** This model will serve as a baseline for predicting popularity based on movies dataset.
    - R**andom Forest Regressor:** This model will be used to capture more complex relationships between features like budget, runtime, vote_count, Oscar status, genres.

V. Model Evaluation

- Split data into training and test sets, evaluate model performance, and compare results.
- **Baseline Metric**: Calculate R-squared and Root Mean Squared Error (RMSE) for the regression model.

VI. Model Optimization and Visualizations

- Feature Engineering
    - Log Transformations
    - Interactive Terms (e.g. runtime x budget)
    - New features (e.g., Oscar Seasons)
    - One-hot encoding

VII. Presentation

Expected Outcome

The expected outcome is to provide predictive insights into movie ratings and personalized recommendations, enabling stakeholders to better understand viewer preferences and make informed decisions in the rapidly evolving film industry.