Project Proposal – Predicting Movies' IMDb Scores to Avoid Bad Ratings

This project revolves around predicting a movie's popularity on the Internet Movie Database (IMDb) before it is released. Hollywood is a multibillion-dollar industry that employs thousands of actors and staff each year, but still suffers from poorly rated movies each year. A movie assigned a rating worse than expected results in hundreds of millions of dollars in lost revenue. By predicting a movie's rating before it is released to the public, Hollywood can focus solely on the movies predicted to be good.

To accomplish this, the IMDB 5000 movie dataset (available for free on Kaggle) will be used to train a machine learning algorithm. This dataset is in CSV format where each row is one movie and each column represents a feature of that movie, such as director, genre, runtime, actors, etc. The algorithm trained and tested on this dataset will most likely be Random Forest regression because the target variable, IMDb score, is numerical instead of categorical and the dataset is very large.

Before the random forest regression model can be trained on the data, some exploratory data analysis (EDA) will be run to better understand the data. A correlation matrix will be generated, histograms and bar charts will be constructed for the key continuous and categorical variables, and visualizations of the top scoring film directors and movie score versus budget will be made. Once insight is gained from the EDA step, the data will be split into training and testing sets, categorical variables will be converted into numerical variables via one-hot encoding, and the model will be trained on the data using IMDb score as the target variable. The testing dataset will only be used to validate the results of the training step.

This project isn't without its concerns though. For one, if Hollywood can predict which movies always perform well, it will have no reason to ever produce differing movies, leading to

monotony in the movie industry. Furthermore, this might leave out many demographics of actors and directors from starring in movies, an issue that is already present in many film genres. Great care must be taken to avoid this bias.

References

Chakraborty, P., Rahman, S., & Zahid, Z. (2019). *Movie Success Prediction using Historical and Current Data Mining.* International Journal of Computer Applications. Retrieved on March 20, 2022 from https://www.researchgate.net/publication/335878983_Movie_Success_Prediction_using_Historical_and_Current_Data_Mining#:~:text=Identifying%20the%20right%20factors%20can,views%2C%20trailer%20views%20etc...

Mayo, A. (2021). *Hollywood is losing out on $10 billion in revenue every year by underfunding work by Black creators.* Business Insider. Retrieved on March 20, 2022 from https://www.businessinsider.com/hollywood-could-boost-revenue-by-10-billion-2021-3.

Raj, A. (2020). *A Quick and Dirty Guide to Random Forest Regression.* Towards Data Science. Retrieved on March 20, 2022 from https://towardsdatascience.com/a-quick-and-dirty-guide-to-random-forest-regression-52ca0af157f8.