Capstone Project 2: Proposal

Problem Statement

While the linear regression model that was found for the first capstone project looks promising, the out-of-sample predictions should be more accurate. There are many tools and strategies available to obtain a regression model with better predictive power on unseen data. In order to deliver better predictions for the optimal hiring of new actors, a more generalizable model should be found.

Possible Clients

The types of clients that would be interested in the results of a profitable actor prediction model would be anyone involved in the hiring of actors. They would include, but not be limited to casting directors, directors, and producers in the movie industry. Agents may also want to glean some insight into the competition for acting roles, as well as assess the value of their own clients.

Dataset

The dataset used for this project will be the same one that was built for the first capstone project. Just to recall, it was originally extracted through API requests from the TMDb website. The process of extraction that was used can be seen in the various Data Acquisition notebooks and detailed explanatory PDF from the first project. The dataset includes characters, billing positions, release dates, runtimes, proprietary review scores, review counts, proprietary popularity scores, genres, casts, crews, budgets, information on whether movies belong to a collection, and revenues. For actor data, they have dates of birth, locations of birth, gender, and filmographies.

Solution

Considering the significant presence of outliers seen in the residual plots of the OLS model, a tree ensemble model will be optimized to find a superior replacement. In particular, a boosting model using XGBoost will be trained to produce better results. The hyperparameter selection will be performed using the informed search algorithm, Hyperopt. Also, data transformations will be performed, such as applying the natural log function and scaling, as well as analyzing the results of outlier extraction to reign in their effects more directly. The goal of this procedure will be to find a model that produces the most accurate predictions about the revenue potentials of unseen actors.

As the scale of target values is quite broad, a custom loss function will be used (Pseudo Huber) during the model optimization process. This loss function will strike the required balance of regularization between L1 (absolute loss) for large values and L2 (squared loss) for small

values, depending on the magnitude of the prediction error. This will enable the model to make more accurate predictions on out-of-sample data over the full scale of movie revenues.

Demonstration

As a way of putting a practical spin on this project, it will be shown how to predict actor revenues from any cinematic year. An optimized subsampling of training data from the years previous to the one selected will be provided to get the best predictions for that specific year. This will test the readiness of the model to new data, and provide a way to make the best predictions for actors in the current year.

Deliverable

The final product will be available through a Github repository. It will contain a paper with all of the relevant analysis, detailed with proper visualizations and a compelling data story. To accompany this paper, a slide deck and the companion code that produced the results will be provided, as well.