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 better. 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 I built for the first capstone project. Just to recall, it was originally extracted through API requests from the TMDb website. It includes characters, billing positions, release dates, runtimes, proprietary review scores, review counts, 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, proprietary popularity scores, and filmographies.

Solution

Considering the significant presence of outliers seen in the residual plots of the OLS model, different ensemble tree models will be compared and contrasted to find a superior replacement. In particular, I will observe whether a bagging model like Random Forest or a boosting model such as XGBoost will produce better results. This selection will be performed using an informed search algorithm. The goal of this procedure will be to find the candidates that best maximize the model's precision and generalizability.
In addition, refined feature engineering will be performed to exclude redundant data that may be diminishing accuracy or leading to overfitting. I will use automated feature selection methods to find the best set of independent variables to deliver a more accurate model.
I will find the model that gives the strongest predictions on out-of-sample data. After consideration of the type of movie that needs to be cast, this model can be used to suggest which actors would be best to hire to give that movie the highest chance of success.

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, I will provide a slide deck and the companion code that produced the results, as well.