Capstone 2: Data Wrangling

The raw data used for the second capstone project was carried over from the first project. A detailed walk-through of the techniques used and choices made during data cleaning can be found in the Data Wrangling PDF from the first capstone project. The additional choices made, with respect to selecting the final form of the data, will be highlighted, thereby optimizing its use for making predictions using a boosted tree model.

The first difference between the datasets used in the two capstone projects pertains to the way that the individual movie data was aggregated over each actor. For the first capstone project, the monetary features, revenue and budget, had values that were the result of summations, as opposed to averages, over each actor. While having the data in this form aided in increasing the accuracy of the optimum linear regression model, it left a desire for more in terms of the usefulness of the predictions, as their interpretability was less informative. For example, the linear regression model made better predictions when using the total lifetime movie revenue of actors, instead of using the average movie revenue of each. Unfortunately, this resulted in a bias toward actors with longer careers. These actors were more likely to have larger movie revenue totals than did actors with fewer movies. To have a model that was more suited to making predictions on lesser known actors, the monetary data was modified, as described.

To ensure the usefulness of the final model to make predictions on unseen actors, it was important to simulate any actor's movie revenue history at some arbitrary point in that actor's career. This would allow the model to predict an actor's varying earning potential at different points in time. Training a model with this objective, would lead to better generalizability for making predictions about new actors, because actors could be observed as they once were early in a career. For example, a director may have needed more resources to guide the decision of whether or not to hire Harrison Ford for American Graffiti (1973) than the director who chose to hire him for Clear and Present Danger (1994). So, for the second capstone project, the average of the monetary movie features was taken, thereby selecting to have less bias toward actors with well established careers. Cell 128 of the Data Wrangler notebook shows the choices that were made for the aggregation by actor.