Project Luther Cheat Sheet

Client:¶

Fill In The Gap Productions, an up and coming movie production company, is looking to identify the best potential projects to pursue to maximize revenue over the course of a calendar year. They have approached us to test whether their business model, and its assumptions, are valid.

Client's Question/Need:¶

The client's business model is to identify the best possible projects to pursue that will yield the highest possible total gross domestic and opening weekend ticket sales using historical movie release data. Their hypothesis is that the release month, genre, rating, and number of similar projects being released that month are the best predictors of how well a project will do and want us to test this.

The "Population" of Movies:¶

The population of movies used for this project will be **all** movies currently listed on Box Office Mojo that have budget information. Currently this includes approximately 2,755 movie titles.

The Model:¶

The model used for this project was a Ridge Regression that included 595 linear features.

The Features:¶

The model used Runtime, MPAA Rating, Genre and Release Month as features. It also included interaction features between Genre and Rating, and Genre and Month. In all there were 595 features in the model.

To arrive at the final feature list, I performed some tests to determine if higher order polynomials of Runtime and Budget were helpful to the model. Neither were. I also chose a Ridge Regression model in order to correct for features that were not contributing to the model's success. This helped me to eliminate Budget as a feature.

The Outcome:¶

The end result of the project was to be able to predict the actual Domestic Total Gross +/- $52,155,519.89 with 95% confidence. We were also able to identify the specific Genre/Rating/Release Month combinations that have the highest positive effect on Domestic Total Gross.

Takeaways:

This project has demonstrated for me the difficulty in trying to build models for predicting the behavior of complex systems. Even using Ridge regression and a large feature set, the model only ever an R^2 of 0.354.