Identifying Best Predictors for Movie Revenue through Supervised Learning Models

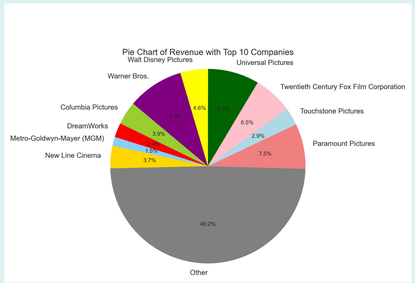
Konstantin Larin, Willie Turchetta, Connie Xu, Brandon Zhu

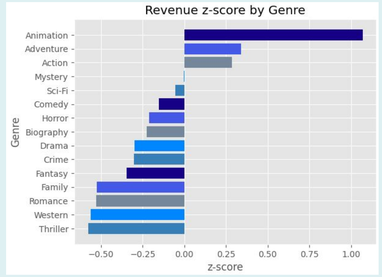
Highlights

- Budget is the most important factor in determining a movie's revenue.
- The combination of budget, company, writer, star, week of release, and rating makes for the best revenue-predicting model.
- We can explain much of the variance in movie revenues with these factors.

Background

The global movie industry is worth over $130 billion dollars. However, a large number of films lose money, so it is vitally important for movie producers to be able to estimate the revenue the movie will bring before that movie gets released.



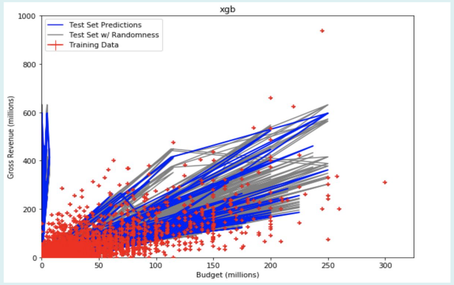


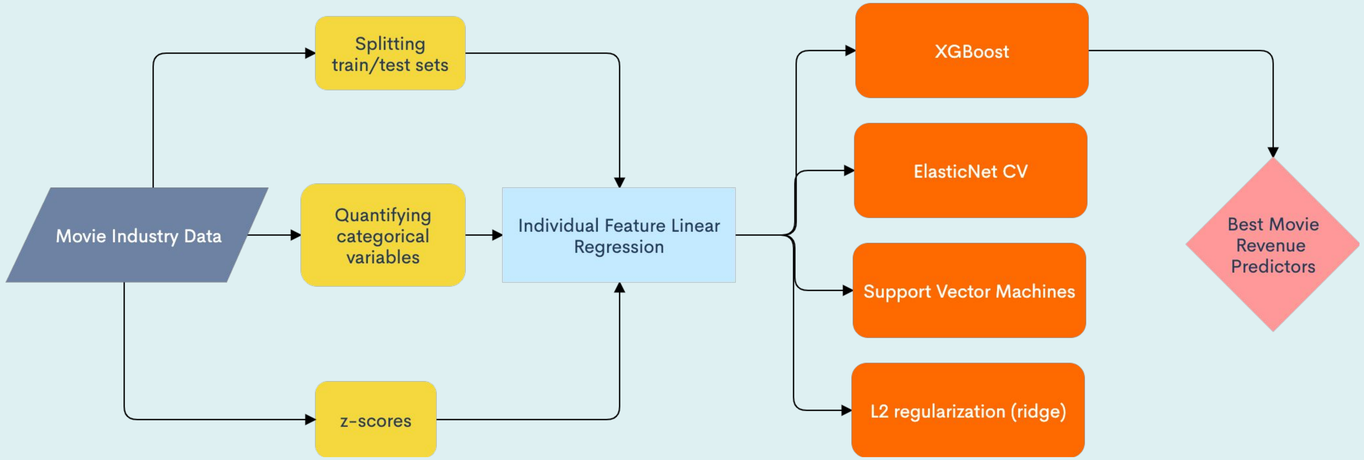
Data

We used movie_industry.csv, which provided us with the most number of features which we could use as input features for our model. We split the data into a training and test set. Then, to normalize the data, we calculated the z-score of each input feature based on its revenue.

Model

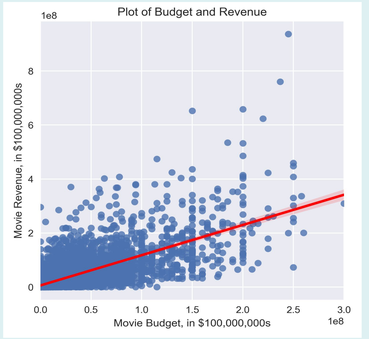
We tested a variety of supervised learning models, a combination of regressions, decision tree, and vector models. We found that a XGBoost decision tree regression yielded the best error and precision.



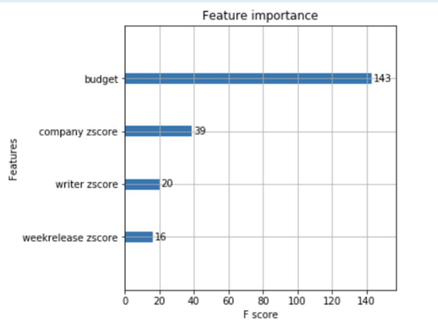


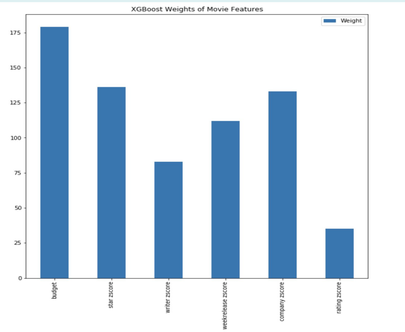
Initial exploration gives R^2 value and regression graphs - preliminary information from which we built our models.

Feature	*R^2 value*
Budget	0.43
Company	0.19
Writer	0.08
Director	0.08
Week Released	0.06
Rating	0.05
Runtime	0.05
Star	0.03
Genre	-0.0002



Our XGBoost algorithm lists the relative feature importance, and assigns budget the highest weighting to predict revenue.





Budget, followed by company, writer, and week release are the best predictors for movie revenue.

Nuance

The input features we analyzed were limited to the information that was provided on the data set, so there could be other important features which we did not cover which could better our model results even more.