Box-Office-Mojo Project: Linear Regression

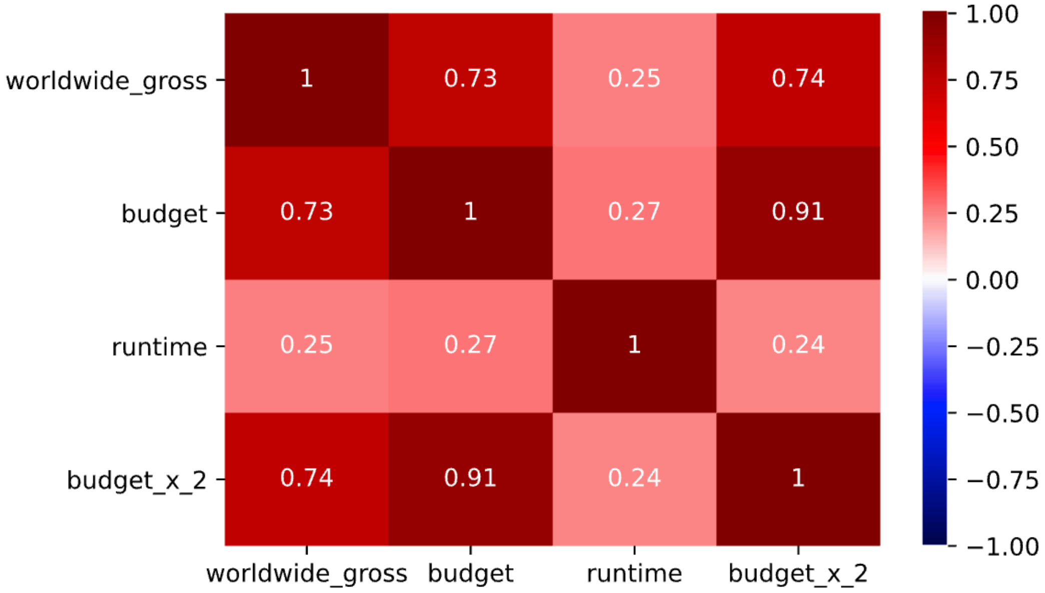
OBJECTIVE: An initial proposal was made to aid Hollywood studios in approaching movie selection from a more scientific point of view, namely linear regression modeling. Logically one would assume that a strong plot line is the pre-dominate factor in determining a movie's success, however, many times this is simply not the case. Also, identifying strong plot lines in pre-production phases is highly subjective . The number of variables studios can control, however, are very limited. Below is a list of the 12 chosen features that offer studios choice in the film making and selection process, labeled by their respective data types. Our target will be the world-wide box office gross numbers of films. Nearly 5700+ movie were scraped.

FEATURES	DATA TYPE
Budget	Quantitative
Run Time	Quantitative
Rating	Categorical
Director	Categorical
Distributor	Categorical
Writer	Categorical
Producer	Categorical
Lead actor 1	Categorical
Lead actor 2	Categorical
Lead actor 3	Categorical
Season Release	Categorical
Genre	Categorical

Up to this point 6 varying linear models have been processed, below is a summary:

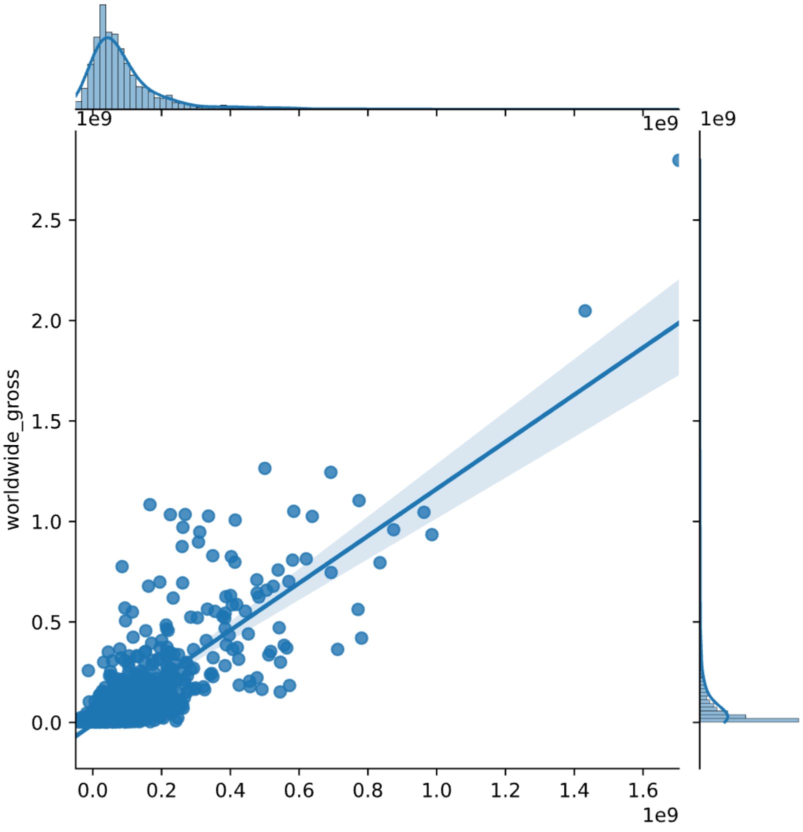
Regression Type	Description	R-square
Plain Vanilla Baseline	Both quant variables	0.536
Polynomial	add budget squared	0.543
Plain Vanilla	All categoricals added	0.62
Tranformation	Target converted to log	0.48
Regularization	Lasso	0.64
Regularization	Ridge	0.64

Below is a heat map of our baseline model, including only quantitative variables and an associated polynomial. Note the high correlation between budget and worldwide gross.

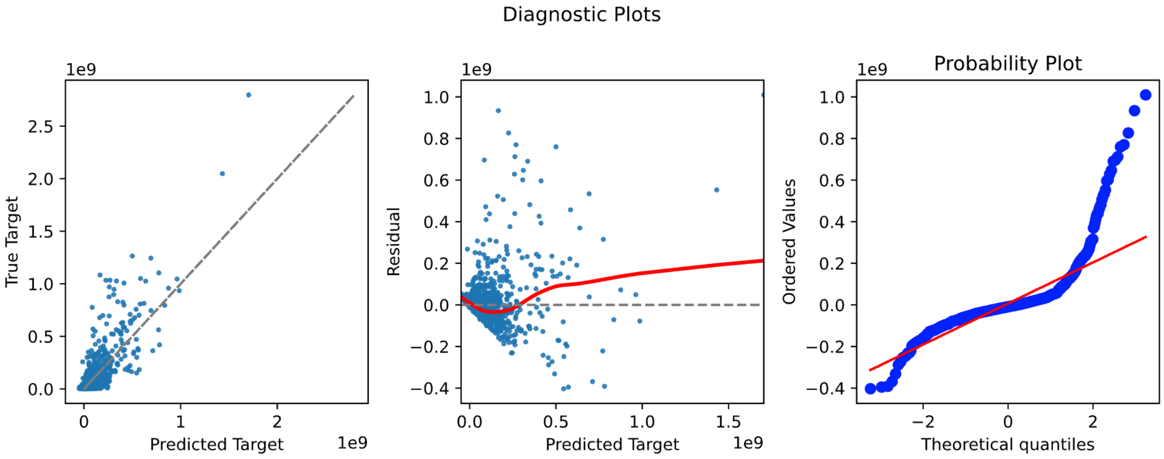


Since Ridge and Lasso produced nearly identical results a decision was made to focus on Lasso. This choice would produce an interpretability bonus to our predictive objective. The model was scored on test data producing an R-squared of 0.64, meaning 64% of the variation in world-wide box office gross is explained by model. Below is a plot of the model's predictions vs. actual.

Actual



A diagnostic plot was produced to gauge the extent to which the model met the assumptions that linear regression is based on(see below).



A few key take aways:

- Our selected mode so far explains nearly 64% of the variation in world-wide gross.
- Variance is not constant
- The model seems to over predict less than 2 standard deviations and under predict greater than two standard deviations.
- The heteroskedasticity and failures in predicting extreme values could be due to a non-normal distribution(right-skew) in the target variable.

FURTHER WORK:

The variance in this model is of concern, undoubtedly affected by extremes on both ends of the scale. World wide gross figures, in particular are heavily right skewed. In order to address this, a Box-Cox transformation will be performed in order to normalize the underlying data. Extra care will be taken to insure the interpretability of results. In another iteration, we will also evaluate the errors of our model at various bandwidths of budget. This will allow us to see where our model performs the best and worst.