Predicting Movie Revenues

By: Sean Kerzel

Dataset - TMDB Box Office Prediction

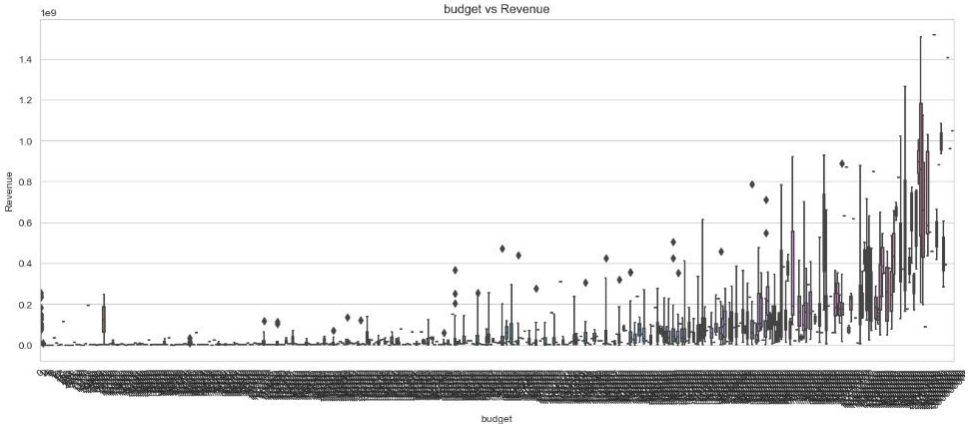
- Kaggle
  - https://www.kaggle.com/c/tmdb-box-office-prediction/data?select=train.csv
- Preprocessing
  - Feature selection and data cleaning
  - Had to account for and remove some extreme outliers
- Its origin is that this was a box office revenue prediction competition for worldwide movies

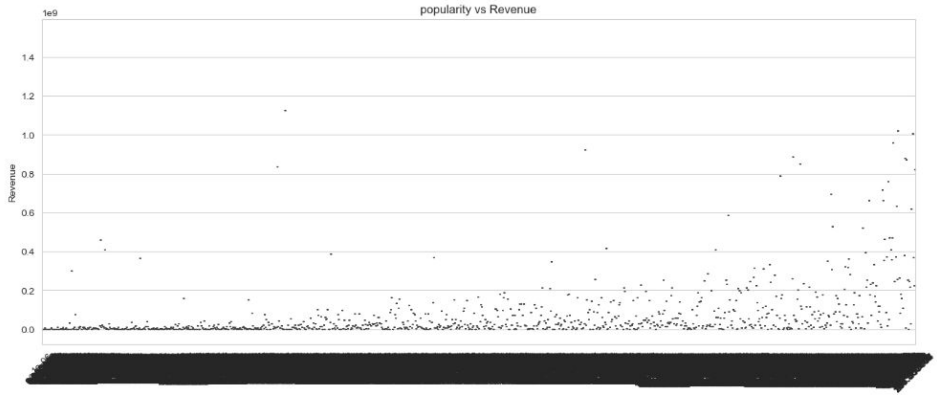
Project Goal

- Predict the overall revenue of the movie based on the most relevant features

Identifying the most important features

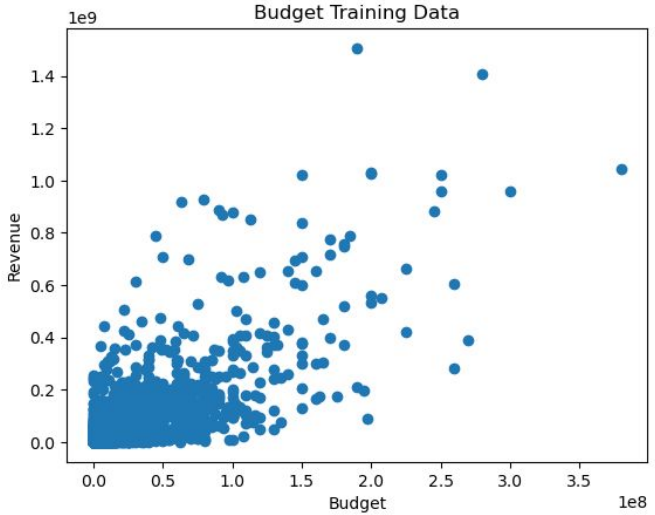
- Ran data visualization graphs for all features, these two showed the most obvious trend
- Positive correlation with revenue

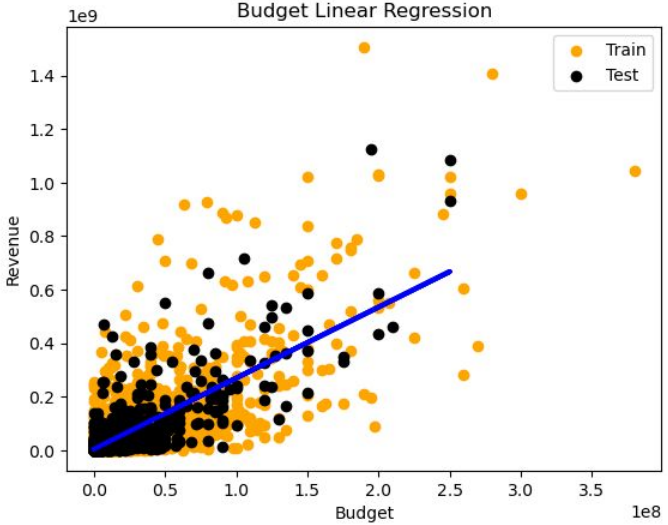




Budget vs Revenue

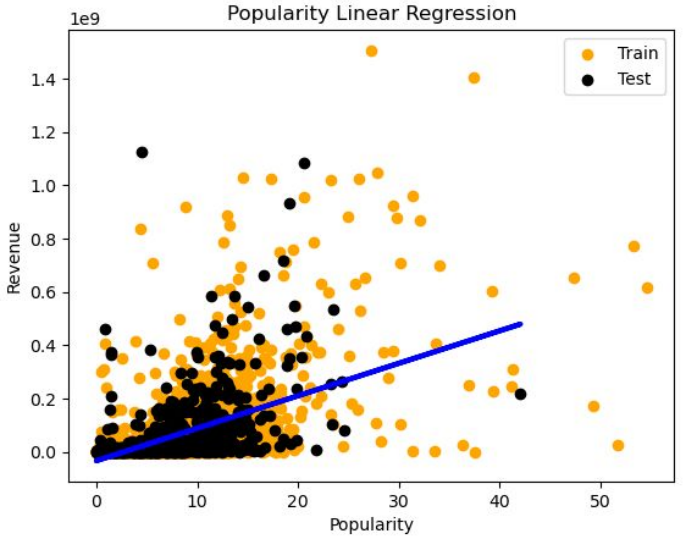
- R^2 on testing data 0.622
- Ran a train test split (both shown in linear regression in below graph)
- This feature was the strongest feature towards predicting revenue





Popularity vs Revenue

- R^2 on testing data 0.191
- Second strongest feature to predicting revenue by a large margin to remaining features despite its low R^2

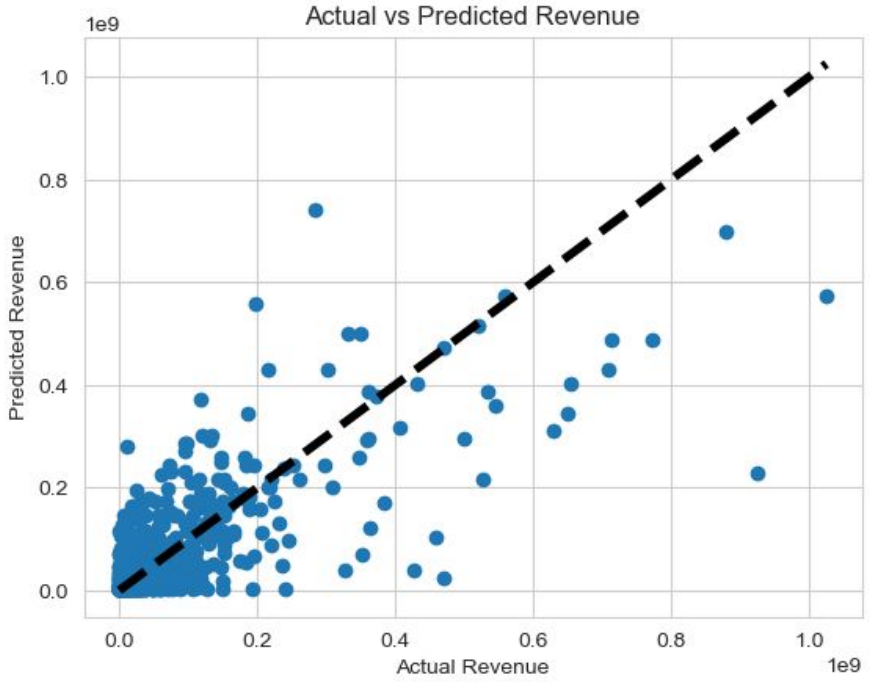


Results

Bayesian Ridge Regression Model

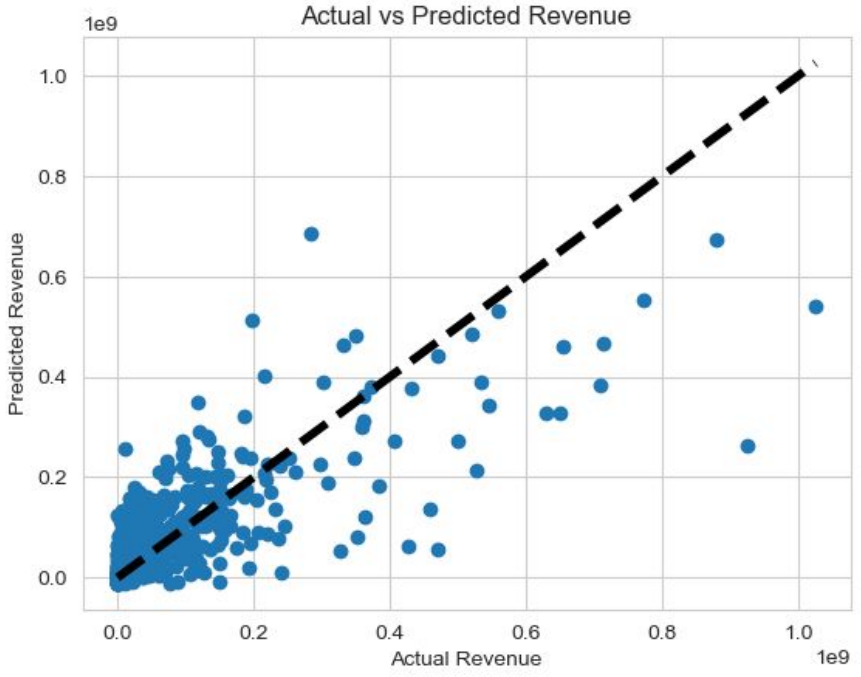
- Following regression models predict revenue based on combination of features

- R^2 of 0.566
- Bayesian is a form of linear regression that gets a probability of predicted revenue
  - Conditional model



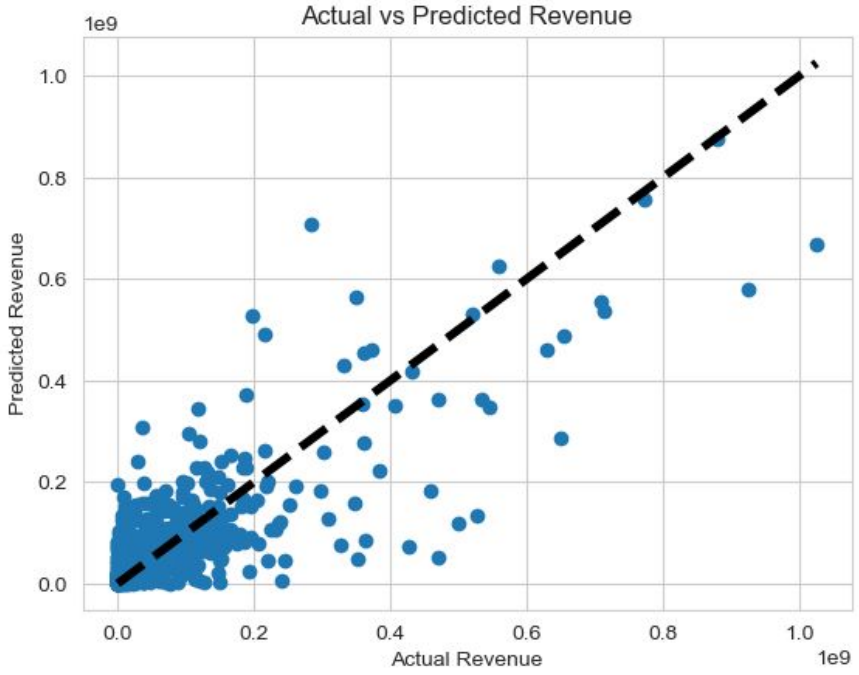
Linear Regression

- R^2 of 0.606
- Classic linear trend between predicted and actual



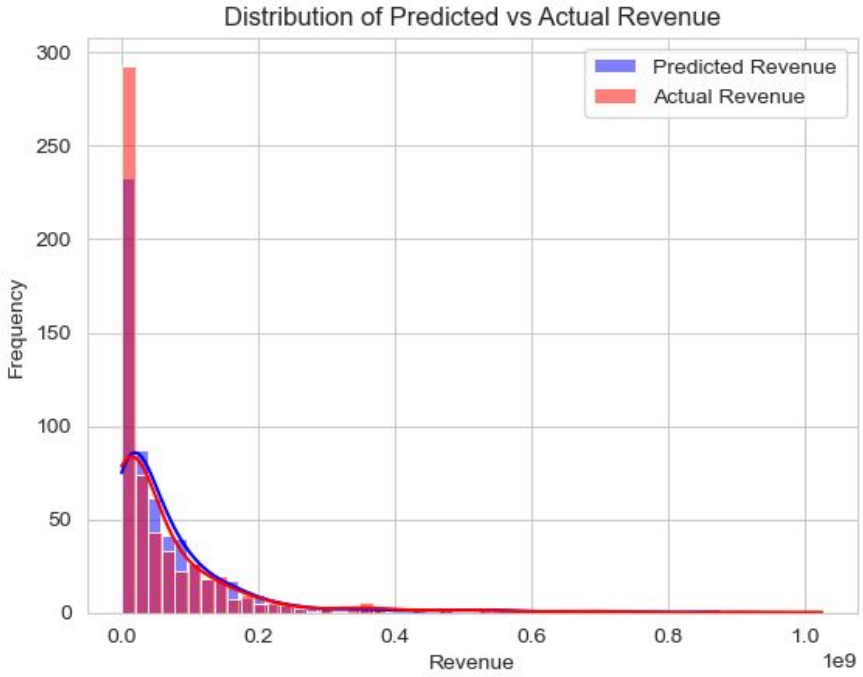
Random Forest Regression

- R^2 of 0.639
- Decision tree based regression model that predict revenue
- Output based on prediction of most trees
- Strongest regression model for predicting revenue

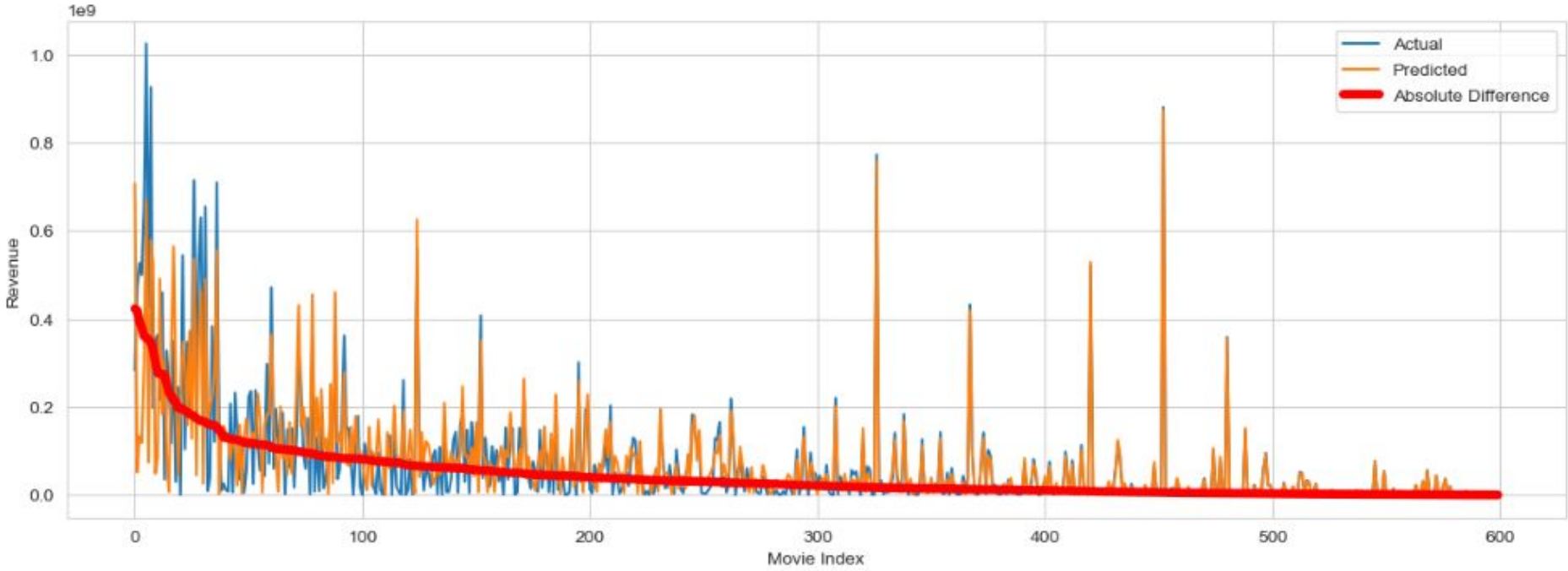


Frequency of predicted to actual revenue

- Small variation in predictions
- Scaled down (millions of dollars)
- At most, predicted incorrectly at 16% at a given revenue range
- At least, <1%



Final Revenue Prediction difference



Future Work

- Addition of new features to the data
  - Features with strong correlation to revenue
    - Cast, crew, genre, production company, release platform
- More focused region of movies
  - This is a worldwide dataset, not just US
- Bigger spread of recent movies
- Dimensionality reduction/unsupervised learning