Box Office Revenue Prediction

GROUP 4

Team Member

- Nhan Nguyen
- Melissa Pollich
- Benjamin Breyfogle

Questions to answer

Despite massive profits, movie industry is one of the riskiest markets for investors due to its uncertainty and unpredictability.

Through analyzing the performance of movies on the market, we are looking for the correlation between attributes of a movie and predict the revenue on release.

Various elements like vote, popularity, budget, languages… contributes to the revenue. We want to see if a predictive model for revenue can be made from these attributes.

Data Preparation Work

We used the Movie Dataset from Kaggle.

Data points include budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.

Most of the attributes that we used in the project are adopted from previous studies that we found.

Tool Used

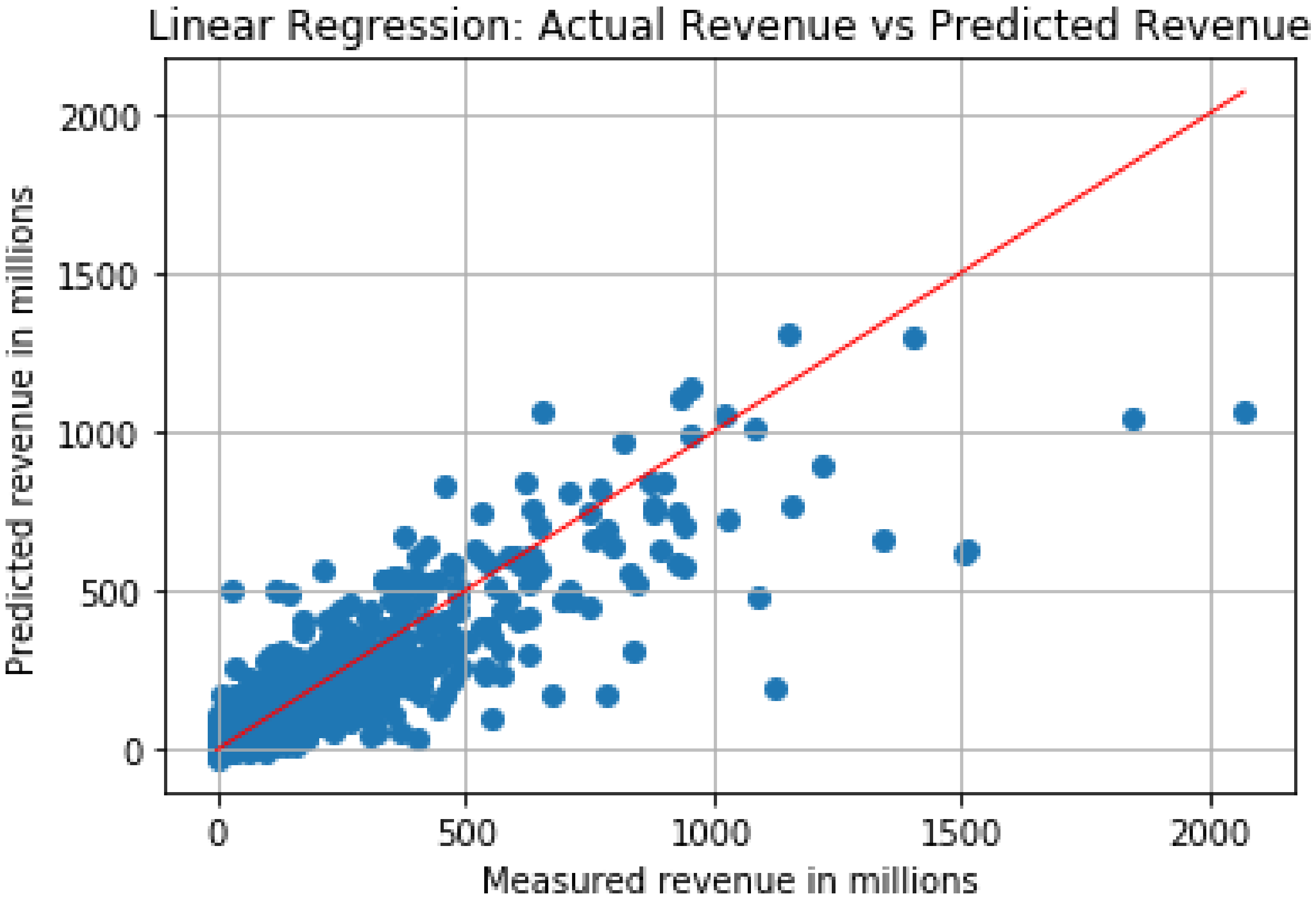
We used Slack initially but some of our members have trouble accessing the workspace.

We used Discord for general project updates as well as sharing link. We use this frequently.

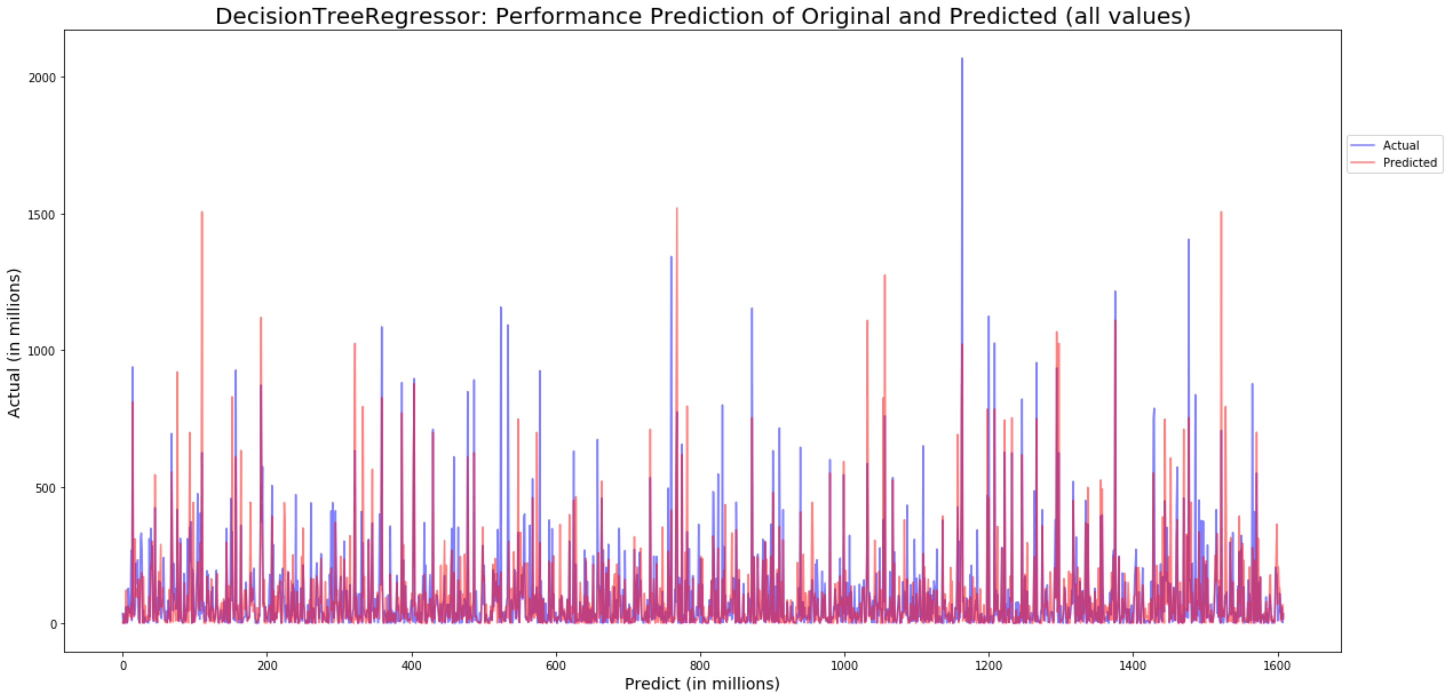
Python is our main language due to its simplicity and GitHub as our repository.

Methods Applied

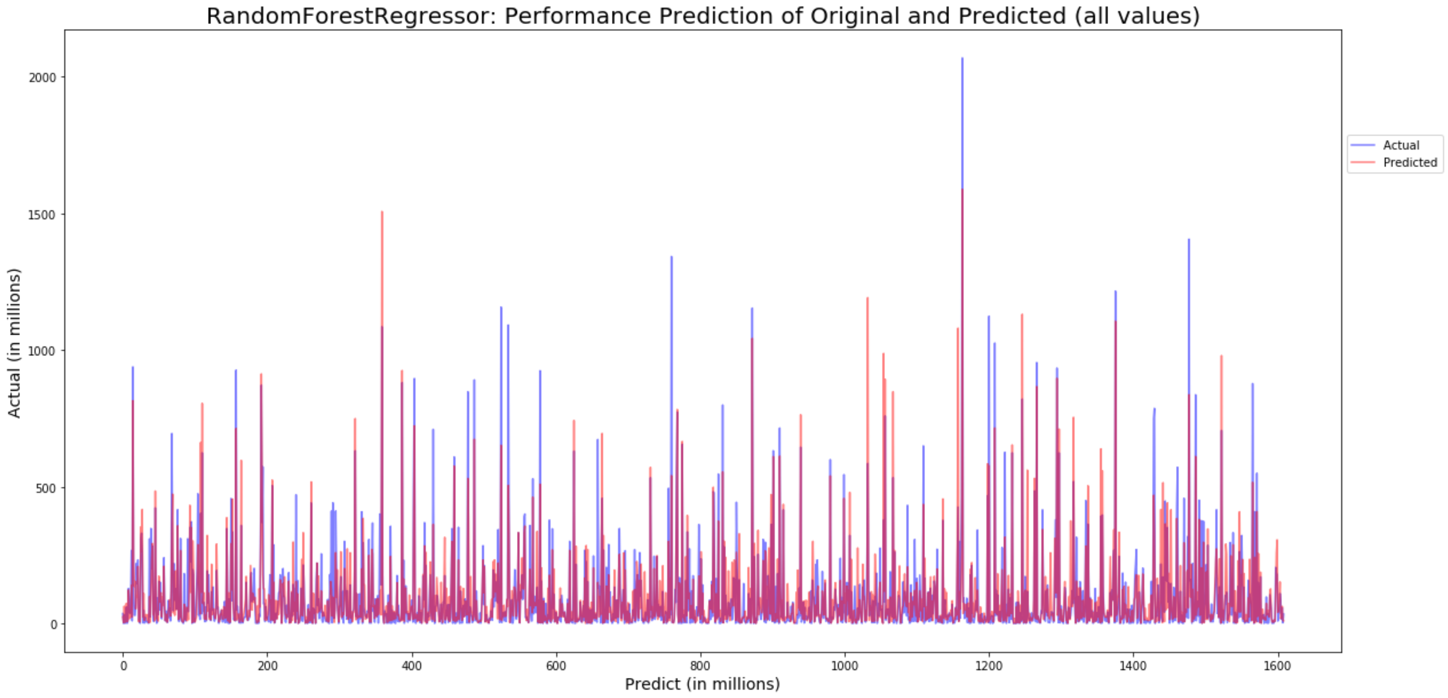
- Data cleaning to reorganize the database.
- Data processing to remove unwanted attributes.
- Mapping of budget to vote count, which are two most correlated attributes
- Binarizing of categorial features.
- Training set and testing data set for analysis.
- Predictive model based on Linear, Random Forest and Decision Tree Regression.



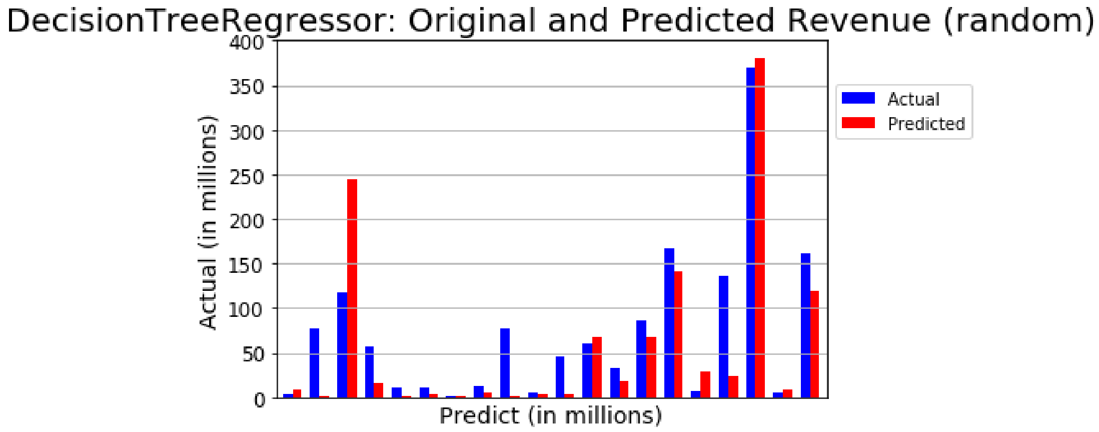
Knowledge Gained

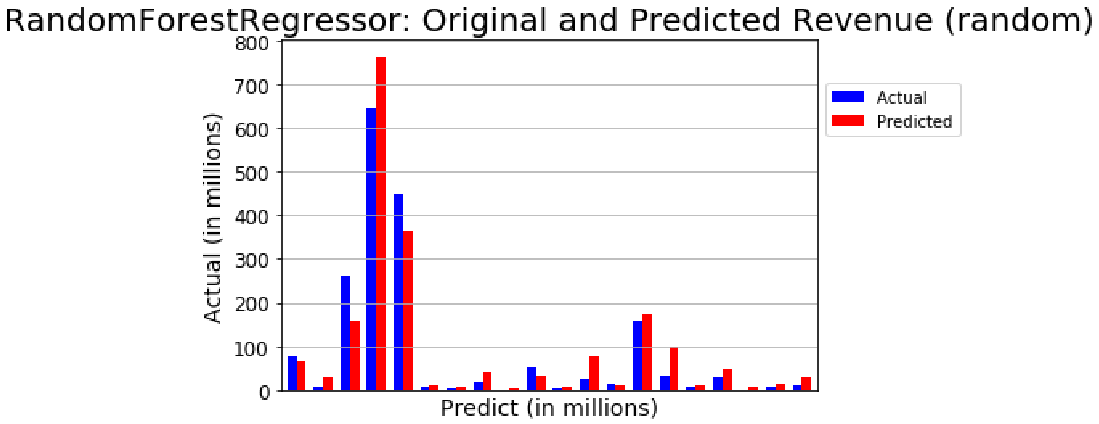


Knowledge Gained



Knowledge Gained





Knowledge Gained

Knowledge Gained

Metric	Linear	Random Forest	Decision Tree
Mean Absolute Error	39.430	44.325	59.009
Mean Squared Error	6817.850	7290.244	14342.770
Median Absolute Error	17.202	19.789	22.663
Explained Var Score	0.781	0.760	0.528
R^2 Score	0.780	0.760	0.528

Knowledge Gained

The predicted values are not close to the reported data values as we had hoped to achieve, but they are not too far off from the actual data.

The predicted values are consistent between three regression model.

Textual data like plot summary, movie name and some more categorial data like actors, directors could help improve the result.

There are many interactions outside the data like trailers, actor’s popularity that affects the revenue.

Using better technique like Deep Neural Network could help make the prediction more accurate

Applications

While these are not accurate model, we can use this as references for revenues of unreleased movies because of some of the most correlated attributes like budget and vote count.

These models could be improved in the future with better techniques like Scaling, Transforming, analyzing of Textual attributes like plot summary, preview…

Including the amount of interactions on social media, list of actors and directors… from other datasets (integration) could help further improve the predictive model.

Thank you