Predicting Movie Revenues

Sean Kerzel
Univeristy of Tennessee Knoxville
EECS Department
Knoxville, TN
skerzel@vols.utk.edu

***Abstract*—The goal of this project was to create a regression model to make predictions on the movie's total revenue based on a variety of features. Using a dataset from Kaggle, I identified the most important features towards predicting revenue and created a model to predict the revenue and compare it with the movie's actual revenue. The model accurately predicted movies with revenues not near 0 but less so for those close to 0, leading to an R squared of 0.64 overall.**

I. Introduction

Recently with all of these strikes and problems movie industries are having with profits, it is important to understand when these movies might be a hit or fail from the get go. Although a movies performance is a very complex problem, some features can be very important in the movie's success. The objective in this project is to create a predictive model to understand why some of these movies fail and why others are hugely profitable. I created this model based on a dataset from Kaggle, with box office revenues of worldwide movies with a variety of revenues. I used these features to make a regression model to predict revenue and compare accurately with the movie's actual revenue.

II. Motivation

As briefly mentioned in the Introduction, with the current state of the movie market having movies constantly losing money after release, it is important to have an idea of how these may fail in advance and work to avoid those major problems. The market is complex so the model may not perfectly predict all movies but taking the major features into account may help in maximizing the movie's performance. This motivation to predict a movies performance before release will remain no matter the state of the market so this problem is always relevant for these industries producing the movies.

III. Objective

With the current state of the movie industry, knowing whether a movie will flop in advance due to a variety of features, may help lessen the volatile state that the market is currently in. The goal of this project is to create a predictive model for a movie's revenue based on a variety of the movie's most important features for example being: production company, cast, etc. I aim to get an accurate ratio between actual and predictive revenue to have a rather reliable model for use on a movie's current state. Hopefully with this model, predicting recent movies in particular yield high accuracy as this model would be most useful in current day for predicting revenues.

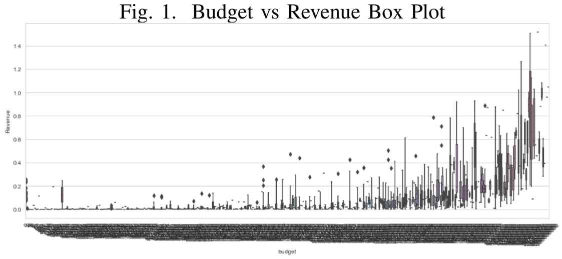
IV. Data

The dataset in use was grabbed from Kaggle for use and preprocessing is the box office statistics of worldwide movies. This dataset have 23 features in the one of which being the movies actual revenue that I later use to compare to the predicted ones. Initially, a similar but different dataset was used but deemed insufficient of information and so this new dataset was used which held more important features, most notably the popularity feature.

After reading in the dataset and parsing over the current data, some preprocessing was required before running a model on it. Running the data as initially grabbed, the data was extremely skewed towards movies with higher popularity. The majority of the data was rated with popularity between 0-60 and a few outliers (12 out of 3000) had popularities just above or very far above 60 and so those were ignored moving forward. Other than dropping the nan values (non-number in number features or corrupted/unusable data), this was the only major preprocessing worth noting.

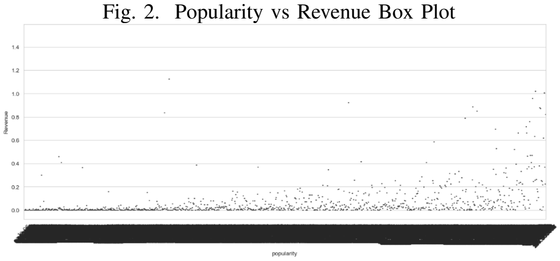
Firstly, all of the features as related to the actual movie revenue were visualized as means of determining the most important features for the predictive model. Only two showed an obvious positive trend being budget and popularity. The following two figures will show box plots of these features, mostly used to get a general idea of which features will matter most later.

Fig. 1. Budget vs Revenue Box Plot



In figure 1, there is positive trend with revenue with high budget movies typically having much higher revenues. While this is rather intuitive, the biggest range of the box plots exist in the same area, high budget. This means that although high budget has good chances of making high revenue, there are also quite a lot of examples of high budget losing a lot of money and having below average revenue.

Fig. 2. Popularity vs Revenue Box Plot



In figure 2, there is a slight positive trend but not as much as budget. Again with very high popularities the revenue goes middle to very high but also has the chance of making almost no revenue. The average of the top third most popular movies is higher than the rest though so although slight, it still shows popularity's relevance in predicting revenue. // The main point of these two box plots were to show the positive trend of these two features. The dimensionality of the features are high as not a lot of movies share an exact budget/popularity but the purpose was in showing their importance. These also show the range of revenue at a given particular budget/popularity value. Two functions were then made. For both of these functions the data was split into training and testing data. The first was a simple scatter plot graphing function of a particular feature and their correlation to the actual movie price. After having looked through the features it was clear that the only two features with major positive correlation with actual price were budget and popularity yet again. The following two figures, 3 and 4, will be budget and popularities correlations with revenue respectively.

Fig. 3. Budget vs Revenue

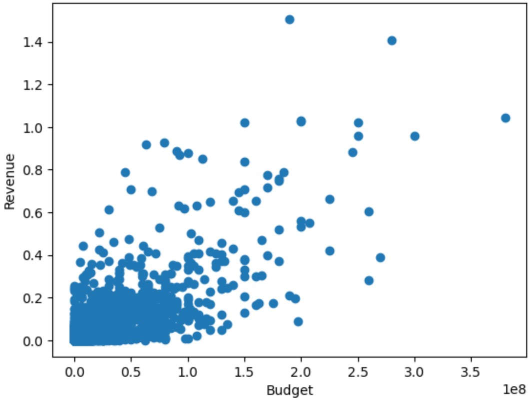
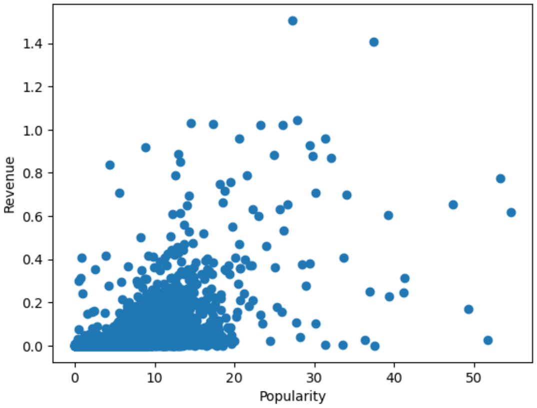


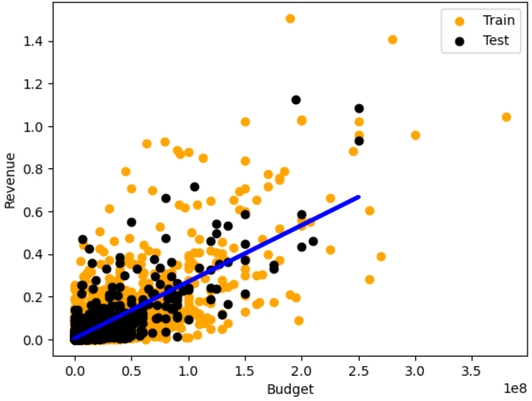
Figure 3 shows sharp edges in the bottom left of the scatter plot. For small budgets it is difficult for a lot of movies to get past 0 revenue and likewise for 0 budget movies gaining a fair bit of revenue. There is a large clump of data in this section which implies a unbalanced dataset with most of the data being in this area. The same problem occurs in Figure 4. A difference between these two figures is that past a certain budget, the movie is almost guaranteed to pass 0 revenue but that is not the case for popularity as there are a couple movies that sit at 0 revenue despite high popularity.

Fig. 4. Popularity vs Revenue



Both of these scatter plots have an obvious positive correlation with price with budget looking stronger with a more centralized points along the main diagonal. Of course just looking at the data visually wasn't a good enough indicator so then I moved to creating the second function which was a linear regression on these scatter plots for both training and testing data. The following two figures 5 and 6 are those two important features and their linear regressions.

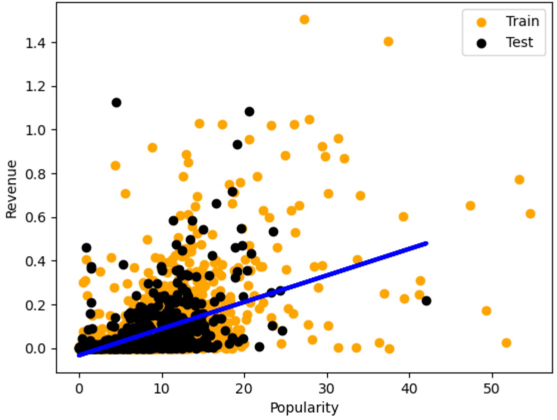
Fig. 5. Budget vs Revenue Linear Regression



In figure 5, the linear fit may be slightly biased derived from how the data was split. The test data picked very few outliers with high budget which means the line was favored to predict along the main diagonal as there weren't the outliers that there were in the training data. Figure 6 seemed to have the opposite effect which is most likely why the R squared was much lower in this one for testing. The testing split had more movie points with high revenue and lower popularities meaning that popularity is likely more important than the linear regression R squared presents it to be.

Individually, these features had very large mean squared errors, however the R squared values for the regressions proved their importance. The R squared for budget was 0.622 on the test data and for popularity was 0.191. Though the popularity value doesn't seem that large compared to budget, it was still vastly more important than the remaining features in the

Fig. 6. Popularity vs Revenue Linear Regression



dataset. Moving forward these two features will be the main building blocks of the predictive model.

V. Models and Algorithms

The main models that are to be used for predicting revenue are different types of regression curves. These are all types of supervised learning which fits our data most as a combination of these features should show an accurate fit towards actual revenue. The first model to try is Bayesian Ridge Regression. This is a type of Linear Regression that is very conditional and gets a probability of a predicted revenue. It is likely that this will perform the poorest but is still worth trying. The next is the normal Linear Regression that will act as our standard/median model and should perform average. Lastly is Random Forest Regression; this is a decision tree based regression model that creates an output based on the prediction of the majority of decision trees. This is likely to perform better than standard linear regression. All of these models were fit with the training data and then run predictions and ran the model on the testing data while calculating the Mean Squared Error(MSE) and the R squared value, which acts as an accuracy metric of the model's predictions.

VI. Results

As mentioned in the model section, our 3 models that we will use for our predictions on revenue are, in order, Bayesian Ridge Regression, Linear Regression, and Random Forest Regression. These models are a combination of important features that work together to create a predictive revenue and then these revenues are compared to the actual revenues of the movies. Budget and popularity are weighted the highest as they were found to be the most important. Shown in figure 7 is the Bayesian Ridge regression.

This model, despite having a high mean squared error like the previous linear regressions, has a relatively high R squared of 0.566. This is the worst of the 3 models and it is obvious that it has some problems predicting some movies close to 0 revenue and a couple with higher revenues. The next model, Linear Regression, should fit the data better than this one shown in figure 8.

Fig. 7. Bayesian Ridge Regression

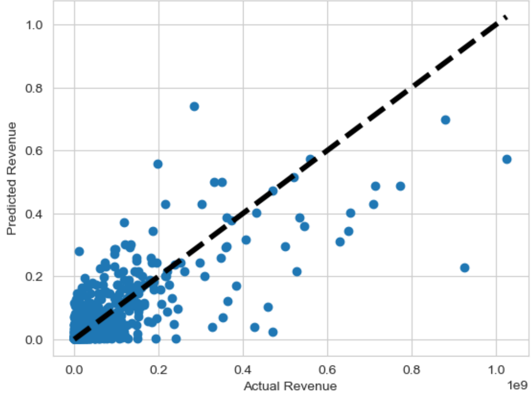
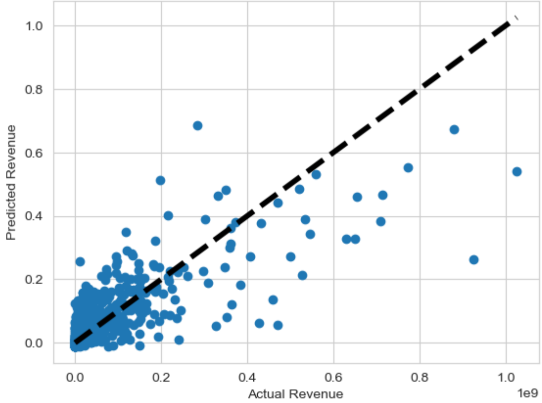
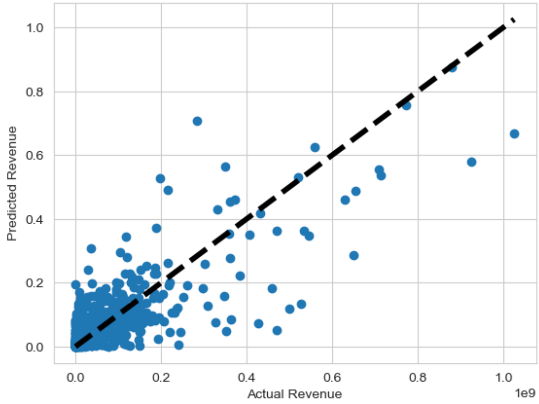


Fig. 8. Linear Regression



As shown in the figure, this model does slightly better with predicting the high revenue outliers but more so improves with revenues near 0 as the predicted values are tighter to the regression line. This is shown in the R squared value being 0.606. The Linear Regression model made a slight improvement but there is still a large variance in the start of the scatter plot with low relative revenues. The last model, figure 9, should be a further improvement on Linear Regression being Random Forest Regression.

Fig. 9. Random Forest Regression



The final regression model ended up performing the best improving very slightly on the two problems the models had before. This model ended up with an R squared of 0.639. As it is difficult to see the slight changes in the graphs between these models I chose to visualize the frequency of my predicted revenues to actual revenues with this model. Figure 10 shows this distribution over the revenue value.

Fig. 10. Distribution of Predicted vs Actual Revenue

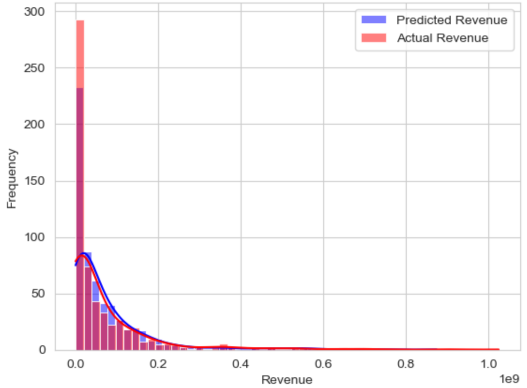
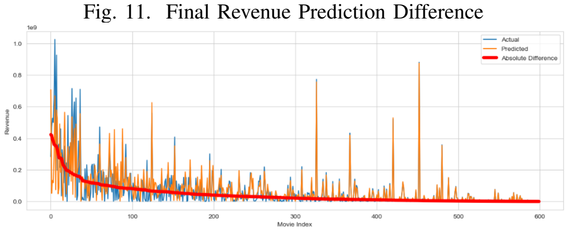


Figure 10 shows a more precise visualization of where my predictive model was off. It is obvious that it is most off with revenues just after 0 with the first group totaling almost 300 movies and the predicted revenue under-predicting around 50 of these 300 on average. Just after this section though, the model tends to under-predict most subsequent revenues but not by very much at all in each section. The predictive vs actual lines have very little difference after the initial section and the curve lines are overlapping mostly. This figure shows a more precise difference but I figured it would also be useful to see a more continuous show of the difference shown in figure 11.

Fig. 11. Final Revenue Prediction Difference



In the figure the red bar shows the absolute difference curve at all points along the movie index. The absolute difference is very thin along the entire curve and the same patterns can be seen in this graph a little differently. The under-predicting at the start can still be seen with the blue peaks (actual) reaching higher revenues than the predicted. After that, the peaks stay relatively overlapped with a low absolute difference, showing that the predictive model has high accuracy at most points. // Some of the issues I experienced with this project have been briefly mentioned or will be mentioned in the next section, but to sum it up, the main problems with this project boiled down to the dataset. The dataset in use, although an upgrade from the first dataset I had attempted to use, had a lot of bad and biased data. Trying to focus on the important aspects and avoid the bad data just didn't yield as good results as one would've gotten with a more even dataset with better values. As mentioned before, a lot of the movies were on the lower end of the range of revenues and that was after removing some movies on the high end, meaning in the future this dataset would need some major undersampling or the use of a different dataset.

VII. Future Work

More can still be done on this topic both with this dataset and with other. With this dataset a couple of other things could be looked at in the future. Diving deeper into the other features could prove useful toward the predicting model but still more information would be needed because as is there were no other features strong enough to be worth using the model. This is of course a big weakness of the dataset that implies that only two things really matter in the making of a movie's revenue being the budget put into the movie and the general audience popularity going into it. One way to make things more clear with this dataset could be dimensionality reduction or unsupervised learning but both would take a lot more catering of the dataset and still might need some more outside data to get more conclusive information.

This brings me to the conclusion that more data is needed, may it be in the form of an entire new dataset all together or just adding some more relevant features that have a strong impact on revenue. Although at first the use of a worldwide dataset of movies seemed to be good to make a general prediction, the data seemed to have too wide of a variety of movies while also not having enough samples of each category to create a conclusive trend, point being that the project could benefit would a more centralized dataset around a specific country, the US preferably. This could certainly improve predictions but at the same time the dataset would need more samples and possibly staying away from movies with near 0 revenue and towards movies done by movie industries only or just having been aired in theaters to justify relevance.

Similarly to having a more centralized dataset around where the movies come from, picking when the movies come from is also important. Picking a more narrow time frame could improve the accuracy of predictions by a lot so it is important to pick a large dataset with movies around the same time period. Along these lines, updating the dataset with movies from the last couple years would be necessary for the model as the whole purpose of the model is to predict revenues for current day as well.

VIII. Org Chart

The entirety of this final project was completed by Sean Kerzel.