Analyzing and Predicting Movie Earnings

Nick Borchich, Kurt Brown, and Matt Latzke

Executive Summary

Introduction

For our DATA-151 project, we researched what factors are most important in determining how much money a movie will earn. This topic is important because movies are a major part of modern culture. By looking into what leads to a movie earning more money, we can get a better understanding of what modern culture values, what stories appeal to people, and what people want out of entertainment. It is also possible that the results of this research may be used to find the most desirable movie to produce next, with the desire that such a movie earn as much money as possible (assuming all of the factors could reasonably be assembled).

Processes

We found and used a dataset containing the top 1000 ranked movies by IMDB rating on Kaggle. This dataset contains information about movies spanning from 1925 to 2019. It also includes a number of characteristics about each of these movies such as the release year, runtime, genres, director, and domestic earnings. We believed that this dataset gave a good overview of the film industry and provided sufficient information to determine what factors are most important in predicting how much money a movie will earn. To analyze and model this data, we used Python in the Google Colaboratory IDE.

We thoroughly cleaned the data, removing unnecessary columns, as well as data points that did not have enough information to properly use. We also calculated the inflation-adjusted domestic earnings for each remaining movie in our dataset indexed to 1983 dollars to provide a more common ground of analysis. We conducted exploratory data analysis to examine the data and the relationships between our explanatory and response variables. Finally, we developed a series of machine learning models, using both regression algorithms such as linear and lasso regression, and classification models such as decision tree and random forest classification, to try to predict earnings for a given movie.

Results

The coefficient of determination of the explanatory variables and the response variables for each explanatory variable revealed that many of the predictor variables did not seem to have a large impact on domestic earnings individually. The only variable with an R^2 value greater than .10 was the number of votes on a film's IMDB rating. There did seem to be a significant difference in earnings depending on whether a film was directed by a famous director. The average inflation-adjusted gross of a film made by a famous director was $69,344,856.78, while the average inflation-adjusted gross of a film made by a non-famous director was $52,474,195.54. This difference was significant at the 5% level. Furthermore, we found large

differences in film earnings based on a film's genre. Despite these findings, our best performing regression model only explained about 19% of the variance in movie earnings. The classification models were a bit more successful in predicting blockbuster films.

Background

Throughout our work on the project, we manipulated our data, performed calculations using the data, and created machine learning algorithms based on the data. We started by manipulating the data in order to create new variables, better organize our dataframe, and categorize our data by changing some of the variable types. We then calculated values like r, R^2, adjusted R^2, RMSE, and averages using the work from our data manipulation in order to help us begin to find relationships in the data. After playing around with different combinations of one or many variables and what their R^2 and RMSE values come out to be, we decided to move more into machine learning algorithms like linear regression, decision trees, and random forest classifications. At the start, we tried using various combos of variables for our linear regression models in order to try and improve the results. Later, we were able to run wider tests encapsulating all of our variables to see if we could find better relationships or better predict which factors affect how much a movie makes.

Now, as we didn't really explain in specifics what each model really does or what exactly our value types are, we will now walk through the values, each type of model, and how they function. The value r is a correlation coefficient that determines the closeness of the association of points in a scatter plot. R^2 is simply the proportion of the variation in the dependent variable that is predicted by the independent variable. Adjusted R^2 is simply an R^2 value that slightly decreases depending on how many explanatory variables there are. RMSE (Root Mean Squared Error) is a measure of how far predictions are from the true values. A linear regression is a type of modeling method that assumes a linear relationship between the independent and dependent variables, and aims to find the best-fitting line that describes the relationship by finding the line that most reduces the SSR (Sum of Squared Residuals). A decision tree is a non-parametric supervised learning algorithm that functions similar to a flowchart of decisions that put together create a model. Finally, random forest classification is an estimation technique that fits many decision trees on various samples of the dataset and uses averages to improve the predictive accuracy and control over-fitting.

Methods

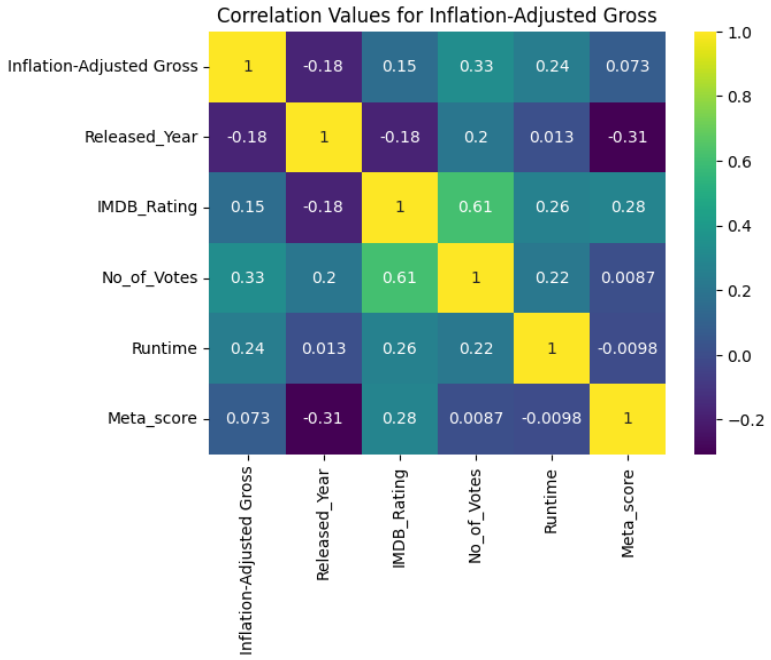
As we investigated which factors had the biggest impact on movie earnings for our dataset, we tried many different methods in order to solidify our results. We started by bringing in data about inflation in order to create a data adjusted gross value for each movie in order to more accurately compare movies over time. We then created a heatmap of correlations for our quantitative explanatory variables. After, we turned these results into R^2 values to get a more

uniform comparison of each of the variables. We also created histograms for each of these explanatory variables in order to see how our data was distributed for each of them. We then realized that after performing some preliminary calculations about the directors, the results were extremely right skewed due to outliers like James Cameron, Christopher Nolan, and others. So we created two separate directors categories and performed calculations on each of them, and our results were more correlated after splitting them apart. We then used t-tests to figure out even more how these director variables related to movie earnings. As many of the other variables hadn't given us great results, we turned to the genre variables provided by our dataset in order to see if they could help us find stronger results. We calculated the average inflation adjusted gross of each genre just to see if we could visually tell if certain genres made more than others. We also calculated the RMSE for genres as predictors. At this point though, even after calculating genre, we still hadn't revealed any strongly correlated variables, so we decided to calculate the R^2 and adjusted R^2 for all of our variables combined and other combinations of some of the variables to see if what combinations would give us the highest R^2 and adjusted R^2 values. As many of our preliminary results weren't very strong, we decided to turn to machine learning to possibly get better results.

Our hope was that machine learning would be able to perform calculations and find correlations between variables that our other models and calculations hadn't been able to thus far. We started with linear regressions of our various explanatory variables. After we received less than impressive results, we turned to decision trees and random forest classifications to try and predict blockbusters, a variable we created that contained only the top 90th percentile of movies. We ended by creating confusion matrices for the above models to get a better visual as to how accurately our models were predicting blockbusters. Through most of our models, whether they be exploratory or machine learning, we managed to only find weak relationships between our variables and movie earnings, however, our final classification models were somewhat accurate in predicting blockbuster films.

Results

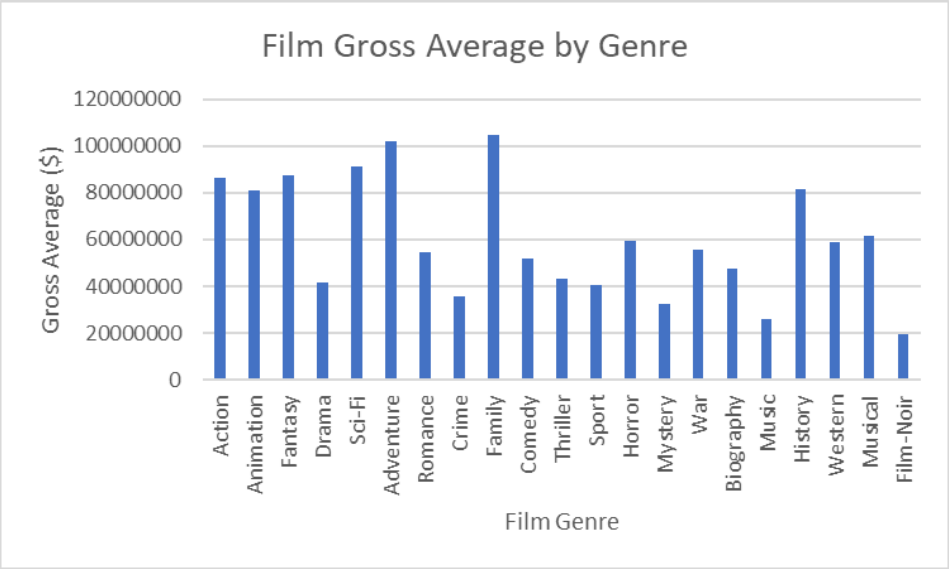
Our exploratory data analysis yielded several notable results regarding what correlates with inflation-adjusted movie earnings. First, we measured the correlation between each of our quantitative explanatory variables and inflation-adjusted gross. We found that the number of votes on a film's IMDB rating had the highest correlation.



We also calculated the R^2 value for each of these variables, as well for the release year of the movie and the unadjusted gross. The only variable with an R^2 value greater than .10 was the number of votes.

	Release Year vs Gross	Release Year vs Adjusted Gross	IMDB Rating vs Adjusted Gross	Number of Votes vs Adjusted Gross	Runtime vs Adjusted Gross	Runtime vs Adjusted Gross
R^2 value	0.055	0.031	0.023	0.106	0.057	0.0052

In our exploratory data analysis stage, we looked to measure the relationship between our categorical variables and earnings. The average inflation-adjusted gross of a film made by a famous director was $69,344,856.78, while the average inflation-adjusted gross of a film made by a non-famous director was $52,474,195.54. Conducting a two-sample t-test yielded a test statistic of 1.999 with 748 degrees of freedom, corresponding to a p-value of .023. This meant that the difference was significant at the 5% level. We also examined whether certain genres consistently earn more money than others. Graphing the average earnings by genre showed major differences between them.

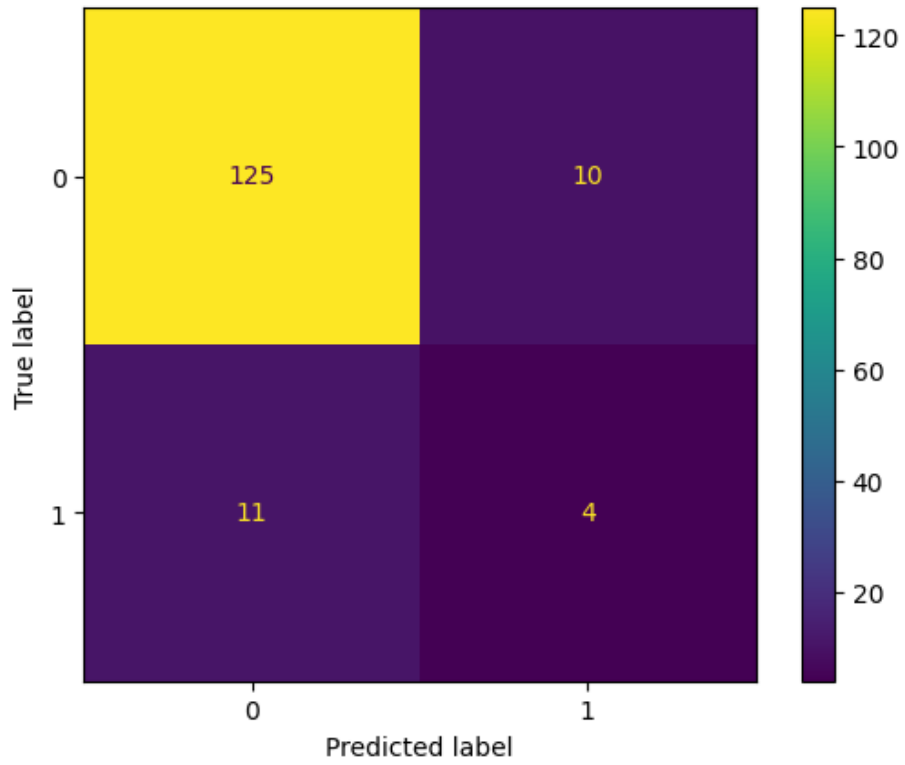


When creating a regression model to predict inflation-adjusted earnings, none of our models were particularly accurate. We tried a number of different combinations of variables to try to build an effective model, but very few even had an R^2 value greater than 0.1. The best performing linear model used the number of votes, IMDB rating, a dummy variable to indicate whether a movie was directed by a famous director, and a dummy variable to indicate whether a movie was in the Action or Adventure genres. For this model, the R^2 was 0.179 and the adjusted R^2 was 0.157. The root mean square error was quite high at 61,700,609.54. These unimpressive results suggested that linear regression was not the best choice for predicting film earnings.

Taking into account that most of the movies in our dataset earned relatively little money, while a minority earned a great deal. We decided to see if we could develop a model to predict what movies would be "blockbusters" and earn in the 90th percentile or higher. Using the variable IMDB rating, and indicator variables for whether the movie was directed by a famous director, whether it was an Action or Adventure movie, and whether it was a Family movie, we developed both a Decision Tree and a Random Forest model to classify movies as blockbusters or not. The Decision Tree model had an F_1 score of 0.6.

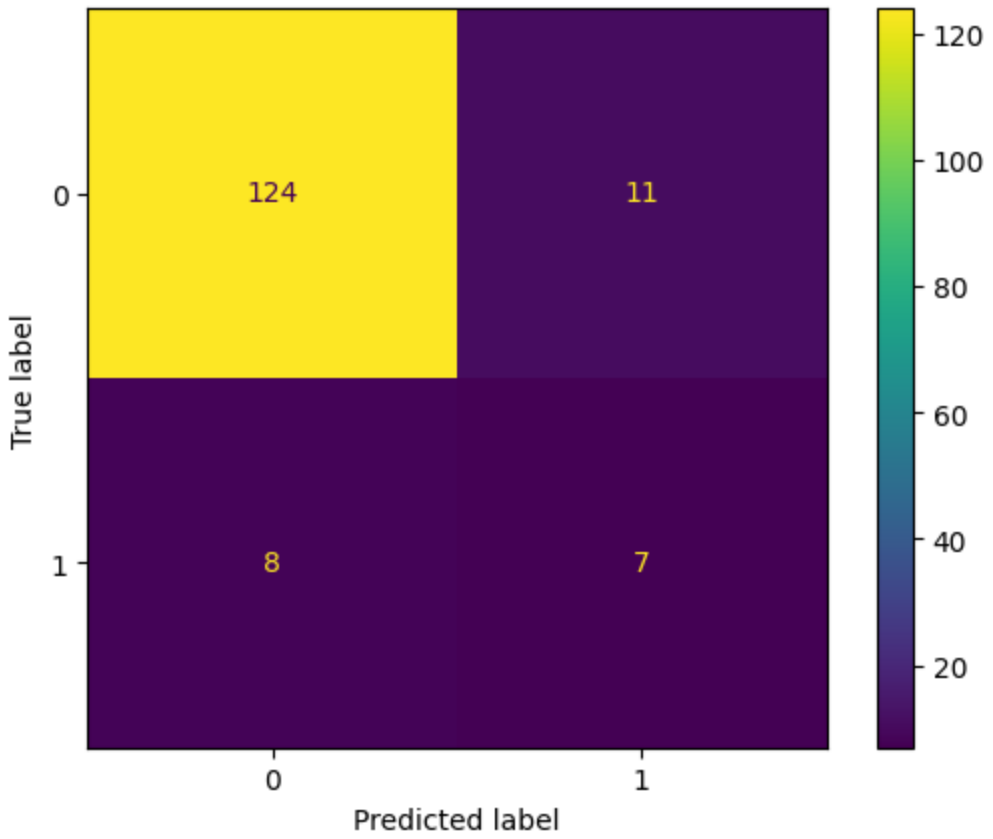
	Precision	Recall	F_1 score	Support
Not blockbuster	0.92	0.93	0.92	135
Blockbuster	0.29	0.27	0.28	15
Accuracy			0.86	150
Macro Average	0.60	0.60	0.60	150
Weighted	0.86	0.86	0.86	150

Average				



This shows that our model was very accurate at filtering out movies that were not blockbusters. It still struggled, however, to identify those that were. The Random Forest model we developed was about as accurate in correctly classifying non-blockbusters, while it was slightly better at predicting which movies were blockbusters.

	Precision	Recall	F_1 score	Support
Not blockbuster	0.94	0.92	0.93	135
Blockbuster	0.39	0.47	0.42	15
Accuracy			0.87	150
Macro Average	0.66	0.69	0.68	150
Weighted Average	0.88	0.87	0.88	150



This result was encouraging. Using our Random Forest model, we have a way to identify which movies are more likely to become big earners that is fairly accurate.

The relative success of the classification models in comparison to the regression models indicates that the top earning movies have distinctive, common characteristics that set them apart from less commercially successful films. These characteristics are related to the film's IMDB rating, whether the movie was directed by a famous director, whether it was an Action or Adventure movie, and whether it was a Family movie. We looked at the feature importances of the variables and saw that the most important variable was IMDB rating, followed by the action indicator variable, the family indicator variable, and finally the famous director indicator variable. We concluded that these variables are particularly important in determining how much money a movie will earn.