Sowmya Parthiban/ A13790839

Niki Tran/ A13200631

Cody Smith/ A11983710

Ronald Montehermoso/ A13732474

COGS 109

Prof. Eran Mukamel

December 4, 2018

**What’s the News about my Revenue?**

**Abstract**

The main aim of our project is to predict a movie’s revenue based on certain factors that are provided by movie databases available online. Such information can help aspiring film producers and companies better understand what factors can lead to the increased reception of their movie. To predict the gross revenue, we added variables to our multiple linear regression model through Forward Stepwise Subset Selection to ultimately find which predictors were significant in predicting the gross revenue. Utilizing cross validation of the models to find which factors led to the best prediction based on the amount of testing error, we then did bootstrap resampling to find the confidence interval for our parameters. Our end result led us to find that the number of star ratings a movie received, IMDb score, and total of likes from cast Facebook pages can allow us to predict the gross revenue of a movie.

**Introduction**

Generating a large revenue is something that many big budget movie projects aim to achieve and knowing the patterns that lead to such success is often an interesting topic. What we want to know is what are these factors that lead to generating massive amounts of revenue towards a movies success and be able to predict a gross income from that information. There are various reasons for why a movie can be successful and generate tons of revenue. One reason may be the score it receives after a critics review or another is the amount of money spent on advertisements. Many other factors may be the cause of a movies success, but what truly matters is what we aim to find. In this paper, we hypothesize that various predictors such as, number of voted users, total cast facebook likes, imdb score, movie facebook likes, are linearly correlated with the total box office success.

**Materials and Methods**

The dataset that we used for this project came from a sample of movies that we found on the IMDb (Internet Movie Database) that was found in the list of Kaggle’s massive repository of databases. IMDb is a popular website that movie buffs/movie lovers use in order to find information on movie, TV and or celebrity content. The sample that we obtained from this website was a database of mainstream Hollywood movies produced from the year 1916 to 2016 that made it successful in the box-office. The dataset has 3598 observations and 27 potential predictors. We decided to focus on 4 of these as predictors; these include, the number of IMDb users that rated the movie, the total number of likes that the movie’s cast has on facebook, the movie’s IMDb score, and the number of facebook likes that the movie itself received.

The attributes that we decided to leave out include the color style of the movie, the director, the number of critic reviews, duration, the number of facebook likes the director has, actors that starred in the movie, the number of facebook likes that the actors had, the genre, budget, plot keywords, the title of the movie, content rating, and year released. We decided to omit the qualitative variables, like actors and directors, since there may be hundreds to thousands of different directors/actors and thus it might be computationally expensive to use them as qualitative predictors for the gross revenue. In addition to omitting qualitative variables, we also decided to omit some quantitative variables, like the number of likes that actors received on facebook, since they were redundant in relation to variables that we intended to use. Since our goal was to do a quantitative analysis of the revenue produced by the movie, we wanted to narrow down on different regression methods we could use to best estimate the revenue based on these attributes. As stated previously, according to our hypothesis, the predictors were believed to have a linear correlation with gross revenue, so we decided to explore a set of linear regression models in this dataset.

We first ran simple linear regression on each of the 4 predictors, and then multiple linear regressions that added a fixed predictor each time based on which had the least training MSE to check for multicollinearity.

Python was accessed through Jupyter Notebook to make computations for creating model selection and model estimation. The Pandas library was used to convert the IMDb csv file into a DataFrame. Afterwards, StatsModels library was implemented to create simple linear regressions to see which of the variables play the most significant part in accurately predicting a movie’s gross revenue. After discovering that the linear model with ‘num\_voted\_users’ had the lowest training Mean Squared Error, the next variable that has the second lowest MSE will be added to the linear model. This process of adding another variable that contributes the least change in MSE is continued until a multiple linear regression that contains all variables is produced. This method, known as Forward Stepwise Selection, was used to find which predictors accurately predict the gross revenue, because we are producing a handful of models that can be used to predict the gross revenue of a movie. For this part, we used the training MSE of each predictor to find the best model at each step since each model had the same number of predictors in a particular step, i.e., they had the same flexibility in a step. Afterwards, the StatsModels library was used again to create a multiple linear regression of the relevant variables.

At the first step of subset selection, which was essentially a simple linear regression of all four predictors against the gross revenue; we found the number of star ratings a movie received did best. Next, we did multiple linear regression with two predictors: number of star ratings a movie received as a fixed predictor and one of the other three predictors used as the other predictor in each model. In this step, we found that imdb score along with the fixed predictor did best. We then ran a multiple linear regression with three predictors: a fixed imdb score, number of star ratings a movie received, and the other two predictors as the third. The total of likes from cast Facebook pages did best. For the last multiple regression, we included all the predictors.

We proceeded with the next step of model selection which was to find the best model among models with different number of predictors. We used 10-fold cross validation using the SciKit-Learn Library from Python for this part, because, unlike stepwise regression, this part requires the comparison of the testing MSE of different models in order to detect overfitting as the number of predictors of a model increases.

The four models whose testing MSE we compared against each other were the best models in the section mentioned above in each step:

1. Revenue = 𝛽0 + 𝛽1\* X1
2. Revenue= 𝛽0 + 𝛽1\*X1 + 𝛽2\* X2
3. Revenue= 𝛽0 + 𝛽1\* X1+ 𝛽2\*X2 + 𝛽3\* X3
4. Revenue= 𝛽0 + 𝛽1\*X1 + 𝛽2\*X2 + 𝛽3\*X3 + 𝛽4 \*X4

Where:

X1= number of star ratings a movie received

X2= IMDb score

X3= total number of likes for the cast on their facebook pages

X4= movie Facebook likes

After creating these linear regressions, we ran 10-fold Cross Validation for the second part of model selection. Since we are using cross-validation here to estimate the test Mean Square Error of the models, we chose not to use Leave One Out Cross Validation due to the high variance in test MSE it may produce in each iteration, whereas, k-fold CV with the right k-value, is said to give more accurate measures of test error than LOOCV. We found that the model with star ratings, IMDB score, and cast Facebook likes as predictors produced the lowest Mean Squared Error.

After finding the model with the minimum MSE of the four, we decided to use Bootstrap resampling for model estimation. Bootstrap resampling allows us to use data already available to us to create a ‘new’ sample of movie data that we can analyze rather than having to collect new movie data with the same predictors. With these resamples, we then ran the multiple linear regression using the best model, 1000 times and found the mean and the confidence interval of each parameter. This gives us a good approximation of the range in which parameters for the best model to predict the revenue of the movie fall in.

**Results**

For the first step of the Forward Stepwise Subset Selection, we found that the linear regression comparing number of star ratings a movie received against the gross revenue contributed the least training mean squared error of 3054038625513336.5 dollar2 compared to the other 3 variables and had a p-value very close to 0, which indicates that it was statistically significant in predicting the gross revenue of a movie.

For the second step of Forward Stepwise Subset Selection, we found from our analysis that the linear regression comparing the star ratings variable and IMDb score against gross revenue contributed the least training mean squared error of 3023809265855595.5 dollar2 and also had a p-value very close to 0, indicating that the IMDb score is statistically significant in predicting gross revenue while also considering a star rating variable.

For the third step of Forward Stepwise Subset Selection, we found that the linear regression comparing the star ratings variable, IMDb score, and total cast Facebook likes variable against the gross revenue contributed the least mean squared error of 2997328339443963.5 dollar2 with a p-value also very close to 0. This indicates that when calculating predicted gross value with star rating variable and IMDb score, the total cast Facebook likes variable is also statistically significant when making these calculations.

For the final step of Forward Stepwise Subset, is a simple multilinear regression containing all 4 variables. This regression contributed to the least mean squared error of 2985565129697062 dollar 2 with a p-value also very close to 0. This shows that all 4 variables are statistically significant in our calculations.

The next step we took was cross validating all of our models to compare the testing error for each, listed in Table 1.

**Table 1 : Cross Validation Results**

|  |  |  |
| --- | --- | --- |
| **Model type** | **Training MSE (Dollar2)** | **Testing MSE (Dollar2)** |
| Revenue = 𝛽0 + 𝛽1\*X1 | **3053158594095970.5** | **3071140720377484.0** |
| Revenue= 𝛽0 + 𝛽1\*X1 + 𝛽2\*X2 | **3022074087331693.0** | **3056792070068005.5** |
| Revenue= 𝛽0 + 𝛽1\* X1+ 𝛽2\*X2 + 𝛽3\*X3 | **2996251272393357.0** | **3018986101982309.5** |
| Revenue= 𝛽0 + 𝛽1\*X1 + 𝛽2\*X2 + 𝛽3\*X3 + 𝛽4\*X4 | **2983665463560067.0** | **3024235475973454.0** |

From the results, we see that model Revenue= 𝛽0 + 𝛽1\* X1+ 𝛽2\*X2 + 𝛽3\*X3 contributes to the least testing MSE and chose this as our best model. We then did bootstrap resampling in order to finding the confidence interval in which the parameters would fall under. Listed below are the parameters’ standard error, mean, and confidence interval.

**Model Estimation with Bootstrap**

**Table 2a : Standard Error through Bootstrap Resampling**

|  |  |
| --- | --- |
| **Parameter** | **Standard Error** |
| const | 5.690873e+06 |
| num\_voted\_users | 2.071754e+01 |
| cast\_total\_facebook\_likes | 1.252319e+02 |
| imdb\_score | 1.026302e+06 |

**Table 2b : Mean through Bootstrap Resampling**

|  |  |
| --- | --- |
| **Parameter** | **Mean** |
| const | 5.764245e+07 |
| num\_voted\_users | 3.008499e+02 |
| cast\_total\_facebook\_likes | 3.127116e+02 |
| imdb\_score | -6.087833e+06 |

**Table 2c : Confidence Interval through Bootstrap Resampling**

|  |  |
| --- | --- |
| **Parameter** | **Confidence Interval** |
| const | [46488336.0130260660, 68796559.10007483] |
| num\_voted\_users | [260.2434847177625, 341.45625719376915] |
| cast\_total\_facebook\_likes | [67.25714565470679, 558.1660064731798] |
| imdb\_score | [-8099385.154112216, -4076280.0835236125] |

At each step, the p-value corresponding to every parameter in each model is very close to 0, given that cutoff for significance is p=0.05. which indicates that they all have a significant effect in predicting the gross revenue.

|  |  |
| --- | --- |
|  | ***Plot 1.*** *The values in table* ***1*** *have been plotted. This plot showcases the relationship between our MSE\_train data, MSE\_test data and the number of predictors in each of the best models in each step of forward stepwise subset selection. As seen in the figure, the model with the four predictors overfits our model. MSE error is to the order of* |
|  | ***Plot 2.*** *This plot showcases the relationship between our true gross revenue and our predicted gross revenue and fits a line showing the pearson correlation between true and predicted revenue.* |

**Discussion/Conclusion**

From this project, we learned that there is a linear correlation between the number of star ratings a movie received, IMDb score, total number of likes for the cast on their facebook pages to the gross revenue that a movie will make. The correlation between our true and our predicted gross revenue turned out to be better than expected. Pearson correlation value between predicted and true revenues were estimated to be: 0.6378197955459448

Analyzing the coefficients of our model, we see that the imdb score seems to have a high negative correlation with the gross revenue, with a mean coefficient value of -6.087833e+06 . However, in the real world one would expect a movie with a high imdb score to be more successful and produce more revenue. This indicates that our model was either underfitting( that is, it was highly biased or inflexible), or there were other unknown factors/ predictors that we didn’t take into consideration.

Another issue that should be pointed out in general is that all coefficients had really large values, which implies that our target variable was highly sensitive to the predictors. To address the sensitivity of our model, ridge or lasso regression would be a better approach for model selection than forward stepwise subset selection.

Throughout the experiment there were many miscellaneous limitations that could have proven to a be a problem when trying to find a correlation between our predictors and our hypothesis. However, with the dataset we found online, we encounter the problem where a lot of features did not actually have a value in place and instead stored a “Nan” value, which did not provide anything useful so how this is was solve was to fill in every “Nan” slot with the mean of the feature we were trying to find and then using that same feature. However, a better solution is always to find a better dataset.