Using Machine Learning to Predict Movie Profitability

**Results Screenshots**

Figure 1

****

Figure 2



**Abstract**

The purpose of this project is to utilize a number of inputs provided through datasets from Kaggle to predict box office revenue for various movies. The datasets contain review and descriptive information of historical movies and their box office revenue performance. Using various fields from the dataset the goal of this study is to train a machine learning model to accurately predict whether the films are expected to achieve enough box office success to cover their cost of production.

**Research Hypothesis**

The overall hypothesis for the project is testing for a statistically significant relationship among the inputs and the revenue. The hypothesis is as follows:

H0: There is no statistically significant relationship among between the independent variables and the revenue generated by the movie.

H1: There is a statistically significant relationship between the independent variables and the total revenue a movie makes.

**Introduction**

The premise of the project is to select a US organization that would be benefitted by a utilization of one or several data analytics projects. The company chosen for this project is the entertainment provider Disney. Disney is involved in various industries from streaming services, adventure and amusement parks, character toys and merchandise, news, publishing, and even sports (Hallman, 2021). Disney as a company has created some of the most iconic characters that are known worldwide and continues to acquire more intellectual property as the days go by the most notable recent acquisition being the Star Wars franchise. Even though Disney has all these various revenue streams at the end of the day the majority of its revenue comes from generating video content. 77% of Disney’s Revenue these days comes from its television and film material whether through direct-to-consumer means, studio releases, or streaming content, while only 23% of Disney’s total revenue comes from its park operations around the world (Johnston, 2020).

The goal of this project is to train a machine learning model that can help Disney cut down on costs related to movie production which would be to help Disney save money before any money is thrown towards the movie’s budget in the first place. The aim is to see if the movie is likely to cover its budget costs given a set of inputs that are known to a studio prior to production such as genre, writers, directors, and estimated budget. If a movie is not likely to cover its budget, then Disney as a company can go ahead and either choose to not make the film or lower the films budget to the degree that it may achieve some profitability.

**Objectives**

The goal of the project is to analyze if there is a statistically significant enough relationship among the input variables and the output variable to help predict the success of the movie. To be able to accomplish this there are a few objectives that are mapped out to help break down the necessary steps.

The first objective would be to understand what relationships already exist among the various input variables and the final outcome variables. There are many different types of statistical relationships and identifying some such as correlations among variables both to each other and the final output will be beneficial in variable selection for the final model.

Another objective would be to eliminate problems that typically occur with machine learning models such as overfitting the models or inputting flawed data. Most data that comes from dataset’s isn’t as clean as it could be. By eliminating rows with missing input data, the machine learning model won’t be afflicted by many of those problems.

The final objective is to determine if there is a particular machine learning model that is a better predictor for our binary classification problem. There are a variety of machine learning models that can be used for this project such as the decision tree model, the logistic regression model, deep neural networks, and random forests. Typically, there are conditions when each of these models is the preferred model, and one of the objectives for this project will be determining which of them will be best for this scenario.

**Overview of Study**

This study will be a purely quantitative look at the statistical capability of machine learning models in prediction case scenarios with a set of input variables. There is nothing here about analyzing patterns or trying to understand about why a particular movie may outperform others more often. Instead this study simply states that it does not care so much about the why in regard to movie success. The only relevant question to this study revolves around which movies will be at the very least profitable. The ultimate goal of this project will be to save the movie studios money when it comes to producing new films by identifying which films will most likely be able to recoup their production costs.

**Literature Review**

Determining why certain movies are immediate fan favorites and why most tend to flop has been something that has been at the forefront of entertainment for quite some time. Not only this but according to data the number of big budget flops has actually increased over time according to the analysis performed by Follows (2018). This is the basis for this project that hopes to cut the amount of box office flops prior to them being produced by utilizing various information known before production begins to generate a predictive model that will reveal whether or not a movie is likely to boom or bust.

One of the biggest stages comes in the form of data preparation. Franzese & Juliano (2018) provide an explanation of correlation analysis which will be used with the quantitative input variables to determine if any can be removed for providing no benefit to the model. Trochim (2020), provides a look at the use cases for Dummy variables and how they can be incorporated for use with nominal and ordinal data to help provide quantitative measures to qualitative values.

The next main phase is the model selection and creation. There are many articles on various machine learning models that can be useful in determining how to create a predictive model for this experiment. Both Ray (2020) and Shin (2020) provide an excellent basis for understanding the use cases of various popular machine learning models. Models such as logistic regressions and random forest classifiers for classification are mentioned as simple and easy to use baseline models that will be a foundational model for this project. Finally, once the models are created and established Mishra (2020), offers multiple ways to check how effective the model is using a variety of statistical techniques and measures.

**Research Design**

As stated previously, this project will primarily involve quantitative analysis to determine statistical significance and predictive analytics to see if revenue can be predicted. However, there are a few steps before we reach that phase. The first stage of the project will involve removing flawed data from the assessment. For the first stage that would mean removing movies form the data source that lack information regarding genre, directors, actors, and budget since those are all crucial input variables.

Many of the useful fields for this project are not quantitative fields but instead are either nominal or ordinal fields. To be able to use these fields properly in the machine learning model they must first be converted to dummy variables. “A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study.” (Trochim, 2020). These will help turn the non-quantitative data into quantitative inputs that can be used in the machine learning model.

However, this will create many new columns many of which may be minimally or not at all impactful to the outcome. This will create a problem of overfitting where too much useless data causes the model to overcompensate and provide worse results. To reduce the problems of overfitting and lack of useful variables the project will also include a correlation analysis between the variables and the output to see if any of them are right off the bat completely not correlated with the final output. This would reduce the number of fields and dimensions and reduce the overfitting problem.

The dataset will then be split into a training, validation, and testing sets and the various inputs will be fed into one or more machine learning models to try and see which model parameters provide the most accurate predictions. For the project since we are trying to determine simply whether a movie will make enough to cover its costs, we are testing for a binary outcome. Some possible models in this case are the logistic regression model and the random forest model.

**Findings**

The initial imdb data set for the project contained 85,000+ rows and 22 columns of movie data to begin with. Once all rows that lacked budget or worldwide income data were eliminated there were 12,762 rows of movie data left. This accomplished the first stage of our research design of eliminating any data that would not be useful in the scope of the project.

In the next stage, the dummy variables were created from the columns of nominal data which in this case are the directors, genre, language, and writers. Originally the data had 22 columns of information but after the creation of all the dummy variables the total number of columns is 20,255. This is a problem because it can easily lead to problems of overfitting especially since the number of columns far outnumbers the rows of data.

To mitigate this problem the correlation analysis was conducted but it was found that the majority of columns had a very weak correlation with the data with only 4 columns having a correlation coefficient greater than 0.1. Therefore, the limit was set to all variables with a correlation coefficient greater than 0.025 or less than -0.025. This led to the final data frame having 12,762 rows and 200 columns.

The findings for the both machine learning models prove that there is a statistically significant relationship between the input variables and the revenue. The p-value was determined by running 1000 permutations on our machine learning models. The p-value both models is 0.0009 which is much less than the alpha value of 0.05 at the 95% confidence interval, and therefore the null hypothesis that there is no statistically significant relationship between the input variables and the revenue can be rejected. Figure 1 shows the model parameters and scores for the random forest model while figure 2 shows the parameters and scores for the logistic regression model.

The random forest classifier model that was utilized in the project came in with an accuracy of 68.76% when testing with the test data set with the hyperparameters tuned. The logistic regression model had an accuracy of 64.78% with the hyperparameters tuned. Overall, this result is not bad considering that the only inputs were variables that are publicly available before the production of a movie so there is room for much growth. The area under the ROC curve for the random forest model is at 0.69 and the logistic regression model is at 0.65, “*AUC* has a range of [0, 1]. The greater the value, the better is the performance of our model.” (Mishra, 2020). The F1 score, which is on a scale from 0 to 1, was 0.69 for the random forest model and 0.68 for the logistic regression model. The F1 score “tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances)” (Mishra, 2020).

**Recommendations**

Since the project was able to establish enough credible statistics to prove that there is a statistically significant relationship between the inputs and revenue, this project can be further expanded in the future. Now that certain features and their impact have been observed, new features that may not be publicly available prior to movie production may also be included to boost accuracy of the models. Further research can be conducted to investigate what other features can be added to the model to make them more effective.

It would also prove beneficial to change certain aspects of this research methodology. For example, instead of going the feature selection route shown here with eliminating variables that lack correlation with revenue, maybe instead a dimensionality reduction route using linear discriminant analysis or principal component analysis can be used instead or in addition.

Currently there are no privacy or security concerns in this research project because there is no sensitive or private information utilized. However, any expansion that requires private data it is advised to anonymize all the personally identifying information through tokenization or another form of anonymization. It is also recommended to hold all the data in a secure and encrypted database so that even if there is a privacy breach all the stored data is still relatively safe from prying eyes.

Finally, I would also recommend that more machine learning models are tested with this data since only two are tested here. The two models chosen here are very capable of robust predictions, however, there are many other models that may work better. A KNN model may be useful here not only as a model to test as but also to identify any clusters that can be used as additional features for the machine learning models.

**Conclusion**

Overall, this project has shed some light into what can be considered a contender in Hollywood. The project has been able to successfully prove there is a significant relationship among the inputs and revenue which can be exploited to create a machine learning model that can predict a movie’s likelihood to recoup production costs prior to production. The project utilized multiple statistical methods to determine if given the limited amount of public knowledge prior to production is enough to generate a prediction model that can accurately form these predictions.

This project has proven that there is enough evidence and inputs here to generate machine learning models that can help make better decisions and should be taken advantage of. The project leaves the door open to future enhancements to current research in the form of new model testing, new process changes, and additional features.

The random forest model, with an overall accuracy of 68%, is already better at predicting this than simple chance by more than 18%. In 2019 the US produced 792 movies at a cost of $65,000 each (Mueller, 2021). However, roughly 80% of movies fail to make money entirely so that 12% can mean a difference of hundreds of thousands of dollars. Therefore, even the slightest increases in accuracy can mean large sums of money are saved, so future enhancements, no matter how small, will make a big difference.

**References**

Follows, S. (2018, December 3). *Is the number of box office flops increasing?* Stephen Follows. https://stephenfollows.com/movie-flops/.

Franzese, M., & Iuliano, A. (2018, September 6). Correlation Analysis. Encyclopedia of Bioinformatics and Computational Biology.

Hallman, C. (2020, October 20). Every Company Disney Owns: A Map of Disney's Worldwide Assets. TitleMax. <https://www.titlemax.com/discovery-center/money-finance/companies-disney-owns-worldwide/>.

Johnston, M. (2020, November 26). How Disney Makes Money: Media Networks generates the most income. Investopedia. https://www.investopedia.com/how-disney-makes-money-4799164.

Mishra, A. (2020, May 28). *Metrics to Evaluate your Machine Learning Algorithm*. Medium. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

Mueller, A. (2021, January 6). *Why Movies Cost So Much To Make*. Investopedia. https://www.investopedia.com/financial-edge/0611/why-movies-cost-so-much-to-make.aspx.

Ray, S. (2020, December 23). *Commonly Used Machine Learning Algorithms: Data Science*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/.

Shin, T. (2020, October 17). *All Machine Learning Models Explained in 6 Minutes*. Medium. https://towardsdatascience.com/all-machine-learning-models-explained-in-6-minutes-9fe30ff6776a.

Trochim, W. M. K. (2020, July 24). Dummy Variables. Research Methods Knowledge Base. https://conjointly.com/kb/dummy-variables/.