How Profitable Will Your Movie Be at the Box Office?

Nikki Fitzherbert

Abstract

The successful performance of a movie at the box office remains critically important to producers and investors in the present day, as it can not only influence the success of secondary revenue streams but also prove incredibly costly for the studio if it fails to attract a large enough audience. However, most of the previous research has focussed on forecasting gross box office revenues, which fails to take into account the size of the investment required to produce, market and distribute a completed movie. This study aimed to determine whether it was possible to construct an accurate predictive classification model using box office profitability as the measure of success rather than gross box office revenues.

The study used Hollywood movie data from the SAS Visual Analytics learning environment to perform an exploratory analysis before comparing the results from a logistic regression to a decision tree model. The primary metric used to determine the accuracy of the model predictions and perform the model comparison was the misclassification statistic. Both models identified the star value of the cast as the most important variable in classifying movies based on their profitability, but the decision tree model only performed marginally better overall.

The results provide an important contribution to the field of movie box office research by identifying some of the important predictors of profitability rather than just gross box office revenue. They would also assist movie studios and investors in avoiding movies that might be best avoided financially.

Introduction

Making movies has always been a risky and uncertain venture for both investors and producers, with only about thirty to forty per cent breaking even and just ten per cent making a profit at the box office (Ghiassi, Lio, & Moon, 2015). As a former president and CEO of the Motion Picture Association of America once said "…No one can tell you how a movie is going to do in the marketplace… not until the film opens in darkened theatre and sparks fly up between the screen and the audience." (Delen & Sharda, 2012). Despite the issues associated with trying to accurately predict such an uncertain outcome, being able to successfully and accurately estimate how well a movie will perform at the box office has always been a critical issue for the movie industry, and one that has intrigued researchers ever since Litman's (1983) seminal multiple regression paper. Even in the present day, when movies are distributed through a multitude of other channels, such as Netflix, a movie's success at the box office is still vitally important as it can influence the size of these other revenue streams (Antipov & Pokryshevskaya, 2017).

Most of the work to date in this field has used information that is available during the post-production, pre-theatrical release period. However, as Eliashberg, Hui and Zhang (2014) pointed out, most of these indicators are unknown at the point when movie studios need to decide which proposals to "green-light" or turn into movies. As a result, the common method of assessing box office potential is to compare scripts to several similar movies and use their performance at the box office as a benchmark.

Nevertheless, both these approaches fail to take into account the costs (or estimated costs) of producing, marketing and distributing a movie on its ultimate financial performance. Ghiassi et al (2015) pointed out that average production costs have just about doubled since the early 1990s, and advertising budgets have tripled in the same period.

This study aimed to contribute to the existing body of research on the financial performance of movies at the box office, by exploring whether it was possible to construct an accurate predictive

classification model using movie profitability as the measure of success rather than gross box office revenue. The hypothesis tested was:

H1: A decision tree model can more accurately predict movie box office profitability than a logistic regression model.

H0: A decision tree model is equally as accurate as a logistic regression model in predicting movie box office profitability.

Data

The dataset used for this study was the "Hollywood Movie Dataset" in the SAS Visual Analytics learning environment. Information from several different movie databases had been compiled using automatic and manual methods into a single database ("Hollywood Movie Dataset,") that contained 2,330 movies released between 2000 and 2010, with gross box office revenues ranging between about $50,000 and $2.0 million.[1] Of the 17 variables available in the dataset, 12 were used in this study and are listed in Table 1 below:

Table 1: Selected variables from the "Hollywood Movie Dataset"

Name[2]	**Possible values**	**Type**
Competition	High, Medium, Low	Nominal
MPAARating	G, PG, PG13, R, NC17, UR	Nominal
OriginalScreenPlay	Yes, No	Binary
EstimatedBudget	3,000 to 665,015,113	Interval
Genre_	Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Family, Fantasy, History, Horror, Musical, Mystery, Other, Romance, SciFi, Sport, Thriller, War, Western	Binary
GrossBoxOffice*	50,076 to 2,039,472,387	Interval
MaxScreenCount*	2 to 4,468	Interval
MovieLength	45 to 219	Interval
SpecialEffects	1 to 5 (1 = lowest, 5 = highest)	Nominal
StarValue_Cast/Director/ Producer	1 to 5 (1 = lowest, 5 = highest)	Nominal

Each movie could belong to multiple genres, so each possible value of **Genre** was represented as its own binary variable, which increased the independent feature count in the study by 19.

MaxScreenCount was converted into a binary representation (**ReleaseType**) in order to capture any impact the scale of a movie's release might have on movie profitability. As there is no clear definition distinguishing wide and more limited release movies from each other (Chen, Chen, & Weinberg, 2013), this study defined the difference empirically based on the dataset at hand. Whilst the distribution of **MaxScreenCount** was fairly uniform above 250 (see Appendix – Figure 1), this study defined a movie to be a wide release if the number of screens was greater than 1,000 movies to try and ensure there were sufficient observations in both sub-groups.

The dependent variable in this study was movie profitability in a binary representation (**ProfitabilityClass**). Following the methodology applied by Galvão and Henriques (2018), a movie

[1] To be exact, the minimum was $50,076 and the maximum was $2.04 million.

[2] Variables indicated by a * were modified to suit the objectives of the current study.

was classified as a "success" if its box office revenue (**GrossBoxOffice**) was greater than or equal to double its estimated budget (**EstimatedBudget**) and a "flop" otherwise. The doubling of the estimated budget allowed the analysis to account for other costs such as those associated with the marketing and distribution of the movie, which are generally not public knowledge (Rhee & Zulkernine, 2016).

Methods

The entirety of the analysis in this study was undertaken using SAS Visual Analytics.

The first part of the analysis involved an exploratory analysis of the 12 variables as described in the previous section of the report. This would help detect any unusual observations or potential outliers that may be candidates for removal from the dataset prior to undertaking any predictive modelling. It would also help identify whether any of the **Genre** categories could be combined together to reduce the number of variables used in either predictive model. The final purpose of the exploratory analysis was to provide some preliminary insight into the bivariate relationships between the independent variables and **ProfitabilityClass** and therefore get some sense of which variables might be important to the financial success of a movie at the box office.

The second half of the analysis used the same set of independent variables to determine if they could be used to classify a movie's financial success (**ProfitabilityClass**). As the response variable was binary in nature, this required models that were designed to deal with classification problems. Therefore, logistic regression and decision tree models were appropriate techniques. In both models, the target event level was set to "flop".

As indicated in Table 1 on the previous page, the large range of values for **EstimatedBudget** raised the possibility that this independent feature would completely overpower all other 10 variables in the dataset. With the exception of **MovieLength**, all the other independent variables were either binary or strongly ordinal. In addition, it also exhibited a significantly skewed distribution. It was therefore deemed appropriate to transform the variable using a natural logarithmic transformation to solve both issues (Appendix A – Figure 2) indicates the result more closely approximated the normal distribution.

The logistic regression model was constructed using a logit link function. It included variable selection at the 5 per cent significance level to avoid model overfitting and ensure that the final model was as parsimonious as possible. The remaining options were left at their default values. Model fit was assessed using a variety of metrics including the misclassification and Receiver-Operating Characteristic (ROC) charts, the residual plot, and the variable importance plot.

The decision tree model was constructed using the same set of independent variables. A custom growth strategy was employed to build the decision tree with a reasonably aggressive pruning strategy. The maximum number of branches was four, maximum number of levels was six and leaf size was 10. The prediction cut-off value was left at 0.5. All other options were also left at their default values. Model fit was assessed using the misclassification and ROC charts.

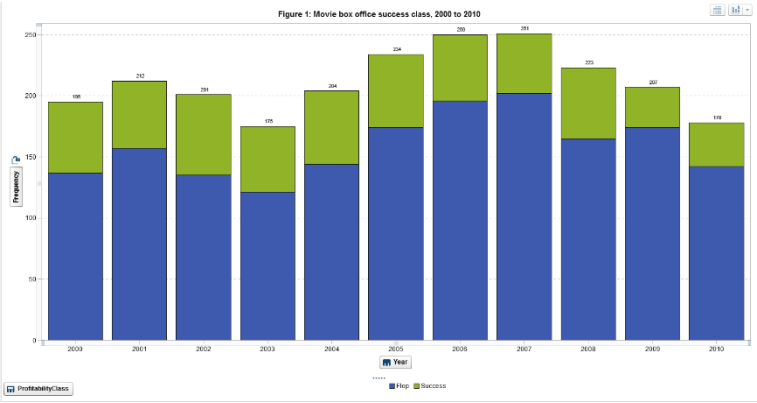
The final part of the analysis involved comparing the two models with respect to the accuracy of their predictions. The primary metric used was the misclassification statistic. The models were also compared and contrasted with respect to their most important variables. This was performed using the variable importance chart for the logistic regression model and the tree chart and node rules for the decision tree model.

Results and Discussion

Exploratory analysis of the data revealed that most (75 per cent) of movies were not profitable at the box office. It was also apparent that this result was very consistent across the time period for which data was available in this sample (see Figure 1 below). The highly unbalanced nature of the sample

likely influenced the predictive model results, which will be discussed in more detail later in this section.

Figure 1: Movie box office success class, 2000 to 2010



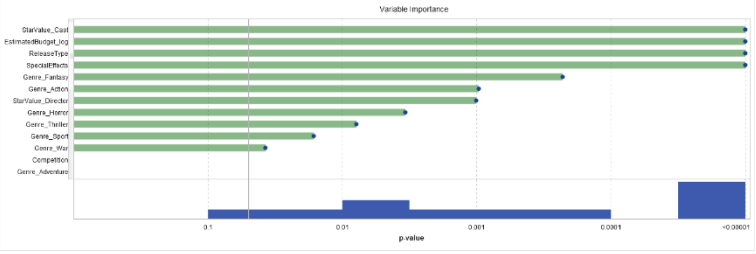
Visualisations of each of the categorical independent variables and ProfitabilityClass also indicated bivariate relationships that were for the most part consistent with the current literature on the performance of movies at the box office. The visualisations suggested that movies were more likely to be profitable if they had lower MPAA ratings, were released on more screens, had more special and technical effects, had less competition, and had high star power with respect to their casts, directors and producers. However, a movie appeared to be equally profitable (or not) irrespective of whether it was (or was not) an original screenplay.

This study could not find a consistent relationship between the profitability of a movie and movie genre, but Ghiassi et al (2015) argued that this variable is an important attribute in determining prospective audience demographics and might be better considered in conjunction with a movie's classification rating and the timing of a movie's release.

Visualisations of the numeric independent variables and Profit (the numeric form of ProfitabilityClass) indicated that there appeared to be only a slightly positive relationship with the estimated budget of a movie and a movie's runtime length (see Appendix – Figures 2 and 3).

The final logistic regression model produced indicated that the most important variables in determining whether a movie would be profitability or not were the star value of the cast, the log-transformed estimated budget, release type and the level of special effects (see Figure 2 below).

Figure 2: Variable importance in the logistic regression model



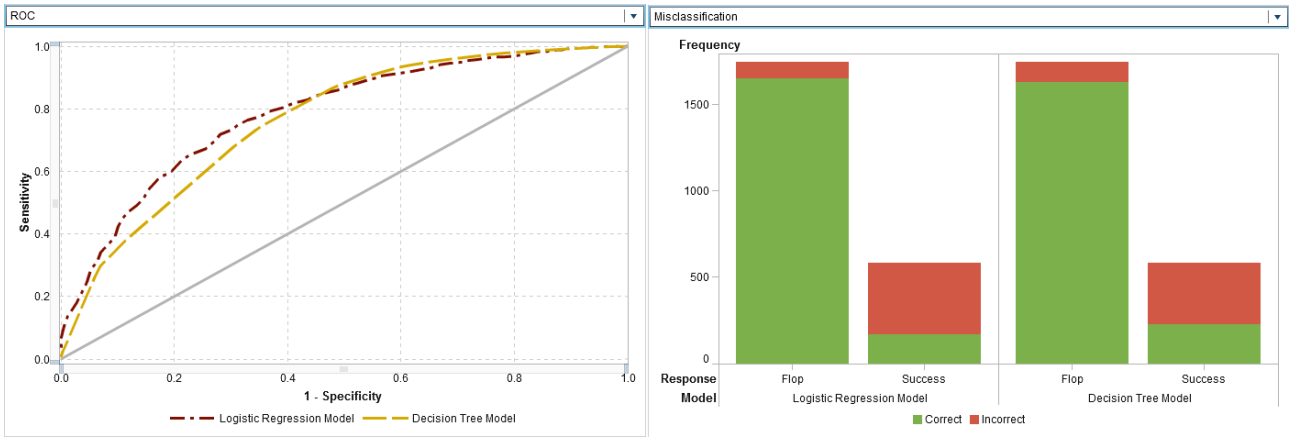
A comparison of the predicted classifications against the actual classifications of the respondents revealed that whilst the model was excellent at classifying unprofitable movies (94.8 per cent) correctly classified, it was unable to correctly classify (70.3 per cent) of profitable movies. This is likely due to the highly unbalanced sample in which far more movies were classified in the "flop" category.

The final decision tree model produced indicated that the most important variables in determining a movie's profitability class were very similar to that of the logistic regression model. They were the star value of the cast, the horror genre, the log-transformed estimated budget, the level of special effects and the release type (see Appendix – Figure 4).

The decision tree model also performed equally well at predicting movies that would be a “flop” (93.5 per cent) and poorly and predicting movies that would be a “success” (39.8 per cent).

The decision tree model was marginally better at classifying the movies in the sample based on the misclassification statistic (0.20 per cent compared to 0.22 per cent) and the False Positive Rate score (0.6 per cent compared to 0.7 per cent) with a 0.5 prediction cut-off value. However, when these results are combined with the equally poor ability of both models to classify a profitable movie and very similar ROC curves (see Figure 3 below and Appendix – Table 1), this study was unable to confidently conclude that a decision tree model could more accurately predict movie box office profitability.

Figure 3: Model comparison using the ROC chart and misclassification statistics



The two primary limitations of this analysis was the lack of a holdout sample for additional model assessment and the fact that there were comparatively few observations in the “success” class of the response variable. Follow-on research should correct these issues in order to increase the robustness of the findings.

Areas of possible additional research could examine in more detail the relationship between movie genre and movie profitability, and also likely interactions with other independent variables such as the MPAA rating and release timing. Another area of further research could also look at controlling for the effect of inflation on the financial variables to enable better comparability of movies across the range of different time periods in the sample.

Conclusions

The main goal of the study was to determine if it was possible to develop a model capable of predicting the box office financial performance of a sample of Hollywood movies released between 2000 and 2010 through a set of variables previously used in previous box office research. The primary difference between this study and previous papers was the modelling of box office profitability rather than gross box office revenue.

The study found that the decision tree model was only marginally better at predicting box office profitability than the logistic model based on the misclassification statistic. In addition, both models were excellent at classifying “flops” but relatively poor at predicting successes. This was likely due to the unbalanced sample used in the study where over 75 per cent had been classified as “flops”. Further research should use a more balanced sample with respect to movie profitability and incorporate the use of a holdout sample to increase model result robustness.

Nevertheless, the study still provided an important contribution to the literature on box office performance and may assist movie studios and investors in determining how to avoid costly “flops” at the box office though examination of the characteristics of such movies. It also provides a stepping stone for further research into the relationship between movie profitability and pre-production features of movies.

References:

Antipov, E. A., & Pokryshevskaya, E. B. (2017). Are box office revenues equally unpredictable for all movies? Evidence from a random forest-based model. *Journal of Revenue and Pricing Management, 16*(3), 295-307. doi:http://dx.doi.org/10.1057/s41272-016-0072-y

Chen, X., Chen, Y., & Weinberg, C. (2013). Learning about movies: The impact of movie release types on the nationwide box office. *Journal of Cultural Economics, 37*(3), 359-386. Retrieved from https://EconPapers.repec.org/RePEc:kap:jculte:v:37:y:2013:i:3:p:359-386

Delen, D., & Sharda, R. (2012). Forecasting financial success of hollywood movies: A comparative analysis of machine learning methods. *ICINCO 2012 - Proceedings of the 9th International Conference on Informatics in Control, Automation and Robotics, 1*, 653-656.

Eliashberg, J., Hui, S., & Zhang, Z. (2014). Assessing box office performance using movie scripts: A kernel-based approach. *Knowledge and Data Engineering, IEEE Transactions on, 26*, 2639-2648. doi:10.1109/TKDE.2014.2306681

Galvão, M., & Henriques, R. (2018). Forecasting Movie Box Office Profitability. *Journal of Information Systems Engineering & Management, 3*(3). doi:doi.org/10.20897/jisem/2658

Ghiassi, M., Lio, D., & Moon, B. (2015). Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Systems with Applications, 42*(6), 3176-3193. doi:10.1016/j.eswa.2014.11.022

Hollywood Movie Dataset. Retrieved from https://www.teradatauniversitynetwork.com/getmedia/29791abf-b7f5-4178-a881-868356047e37/hollywood_movie_dataset

Litman, B. R. (1983). Predicting success of theatrical movies: An empirical study. *The Journal of Popular Culture, 16*(4), 159-175. doi:10.1111/j.0022-3840.1983.1604_159.x

Rhee, T. G., & Zulkernine, F. (2016, 18-20 Dec. 2016). *Predicting movie box office profitability: A neural network approach.* Paper presented at the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). doi:10.1109/ICMLA.2016.0117

Appendices:

Figure 1 – Distribution of MaxScreenCount

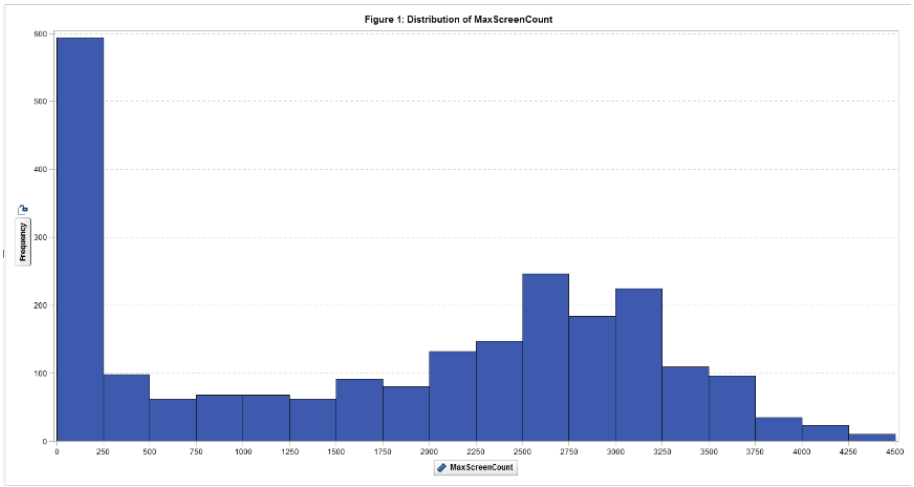


Figure 2 – Distribution of EstimatedBudget after a natural logarithm transformation

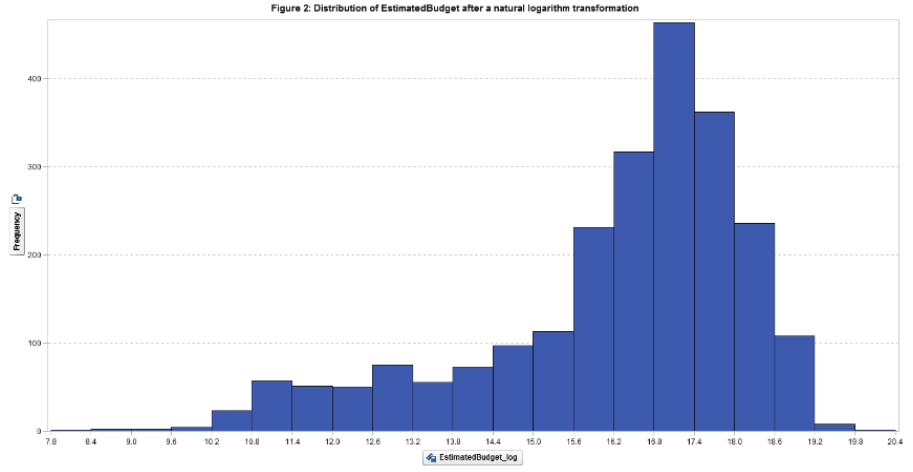


Figure 3 – Profit by EstimatedBudget

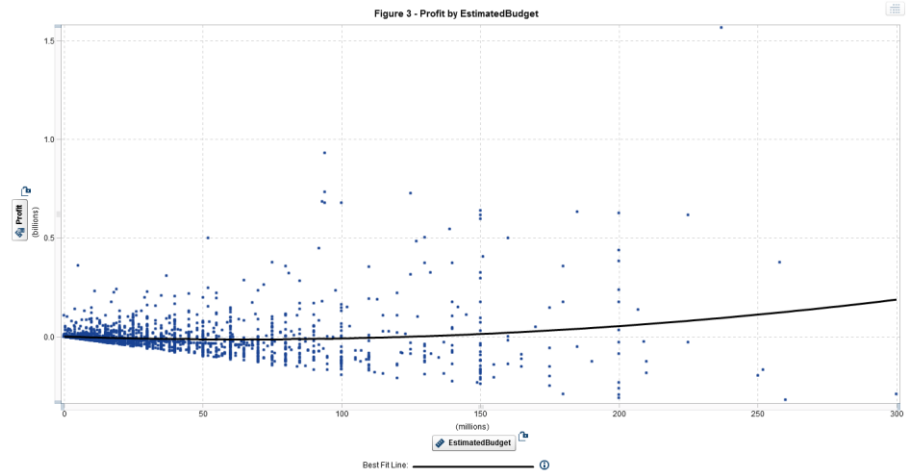


Figure 4 – Profit by MovieLength

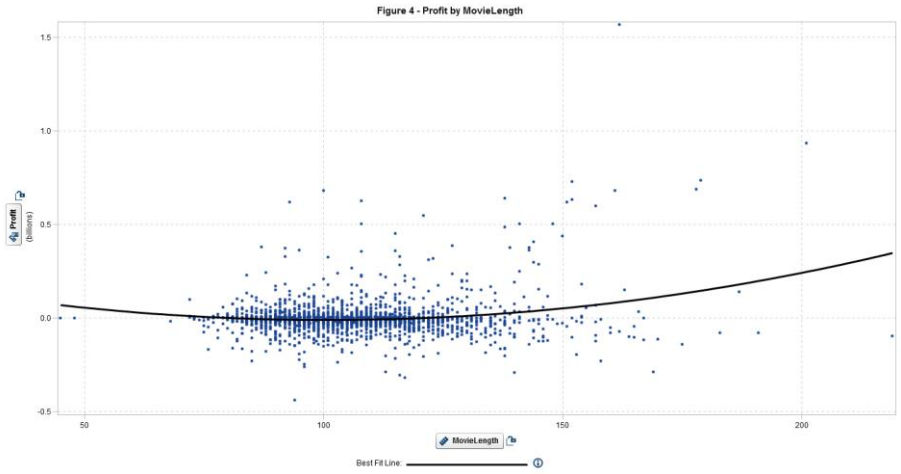


Figure 3 – Top four levels of the decision tree model

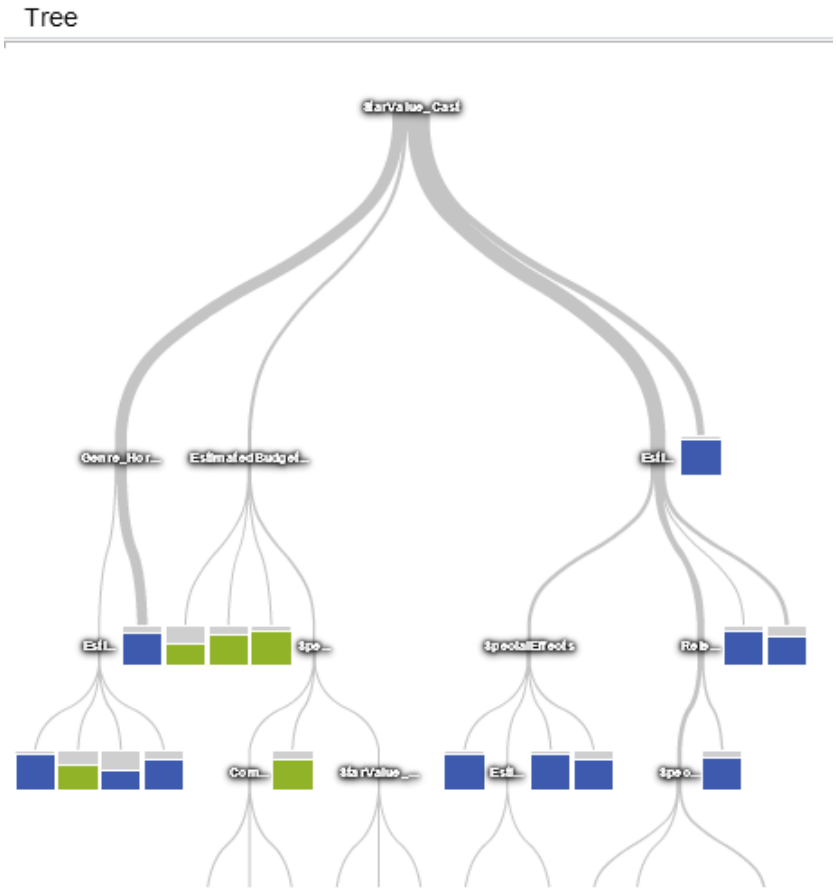


Table 1 – Confusion matrices of the logistic regression (LHS) and decision tree models (RHS)

		Actual Class		
		Flop	Success	
Prediction Outcome	Flop	1,657	173	1,830
	Success	90	410	500
		1,747	583	

		Actual Class		
		Flop	Success	
Prediction Outcome	Flop	1,633	232	1,830
	Success	114	351	500
		1,747	583	