MARK3054

MARKETING ANALYTICS AND BIG DATA

Faiyam Islam

z5258151

Individual Report

Table of Contents

Executive Summary

The movie industry is challenging and requires managers to collect data, analyse and discover solutions to optimise movie performance. Revenue production before theatrical release is an important indicator in attracting investors (Murschetz, 2020). A myriad of factors, including genre, effectiveness of trailers and teasers, number of screens played and many more variables should be considered. The overall purpose of this report is to identify the variables which influence a movie's success, given budget constraints in order to enhance performance by utilising marketing strategies. This report critically analyses 'movie performance' as the number of sales generated after certain periods. Through quantitative analysis, this will be examined through linear regression of the necessary variables to determine the significant factors which influence sales. In essence, the insights gained from the results of this report indicates that the 3 sales variables, viewcount and screens are highly significant. Additionally, disney studio is the most influential factor in increasing sales and correspondingly increasing movie performance. Multicollinearity problems have been detected through our analysis and Principal Component Analysis (PCA) has been implemented to counteract this issue. We have integrated the 3 sales variables into 1 variable called 'revenue' and analysed a multiple linear regression to understand the ceteris paribus effect of independent variables against the dependent variable as well as identifying the significant variables to answer the research question. There are other limitations which invalidate the results, including omitted variables such as total sales, small sample size and budget constraints. Recommendations are subsequently provided in order to address these issues and allow movie managers to concentrate their marketing efforts on improving customer experience and provide an incentive for long term investment.

Introduction

The movie industry has been increasingly challenging for managers to generate and optimise movie performance. For the long-term prosperity of the movie industries, the requirement of high quality, highly usable data is key for customer retention and the creation of a compelling movie (Behrens et al., 2020). It is imperative for managers to scrutinise marketing analysis on past movie performances and based on their budget constraints, produce entertainment for viewers in order to sufficiently gain revenue using the optimal marketing strategies. The challenge for a myriad of movie producers is to identify the most significant factors which affect movie performance. In a rapidly growing industry, data analytics has allowed managers to make creative marketing decisions and accurately predict fortunes of impending movie releases (Sankaralingam, 2017). Additionally, when predicting market performance, managers should provide a personalised user experience whilst reducing churns rates in cinemas or through streaming on sites (Wallace, 2019). Leveraging analytics to film production will allow motivation for industries to develop holistic marketing strategies to gain profit and subsequently reduce churn rates. The primary challenge for managers revolves around targeting the most influential variables which affects movie performance the most. The main objective of movie managers is to identify, target, expand and enlarge the existing customer base on their movies using marketing strategies. Through the identification of these variables and how they affect sales, managers have the ability to reduce attrition rates and target the most influential variables for future investments in the movie industry.

Managerial Problem

The objective of the managerial problem is to investigate what is likely to influence a movie's success and how managers can increase movie performance. To achieve a holistic solution, the manager is required to facilitate data analysis on the relationship between movie sales and a range of key independent variables. The accumulation of qualitative and quantitative research on various genres, studios and production costs are vital in our analysis in optimising movie performance. The preceding research questions were developed in order to accomplish a solution to the managerial problem:

1. What factors are significant when determining the strategies managers can use to increase movie performance?
2. Which studio is the most influential in generating more sales?

Methodology

Initial data exploration indicates that there are 330 movies in this major movie studio. Data regarding production and advertising budget, popularity of main factors, view counts, number of sales in different periods as well as a combination of different studios which each movie was initiated in production, had been collected. The aforementioned variables will be invaluable when assessing the ceteris paribus effect of independent variables on sales. Furthermore, the inclusion of professional critic movie evaluation will allow managers to scrutinise customer feedback and determine which movie provided the most positive feedback and which one provided the most negative one. The main objective of conducting data analysis and interpretation of these key variables, is to investigate the relationship between various factors which contribute to movie performance. This report integrates multiple linear regression, binary logistic regression ANOVA tests and Principal Component Analysis (PCA), to determine a probable solution for the managerial problem. In order to precisely determine the significant factors, these techniques are essential in the manager's analysis of movie performance. It is paramount for this major movie studio to incentivise on promoting certain movies which are struggling for customer viewership whilst simultaneously developing marketing strategies to maintain the reputation of successful ones.

Results

1. ***What factors are significant when determining the strategies managers can use to increase movie performance?***

Factors influencing movie sales

The managers in this major movie studio define movie performance in terms of the cumulative sales. Although the data comprises 3 segments of sales: weekendsales, weeklysales and openingsales, these variables are treated independently. Initially, we shall conduct a multiple linear regression of weekendsales against the remaining variables, as shown in figure 1. According to these results, openingsales, weeklysales and viewercount are highly significant, as the p-value is less than the critical value of the 5% level of significance. Additionally, the number of screens and the dummy variable disney are also statistically significant. In terms of goodness of fit test, this regression model has an adjusted r-squared of 0.9863, which means 98% of the variation in the dependent variable (weekendsales) is explained by the independent variables, whereas the 2% is left unexplained. A correlation test between the sales variables have been conducted in order to address any potential multicollinearity issues. In figure 2, we interpret that weekendsales, openingsales and weeklysales are all highly correlated as the correlation value is greater than 0.7. The issue of multicollinearity gives the manager an incentive to utilise variation reduction techniques, such as Principle Component Analysis.

Principle Component Analysis

In addressing the multicollinearity issues encountered aforementioned as well as dealing with many variables which measure the same dimension of the same concept (sales), PCA shall be utilised. In figure 3, we observe that PC1 has the greatest proportion of variance of 0.4294, which means 43% of the variance in the 3 new variables is explained by PC1. Moreover, we take PC1 and rename it as “revenue”, indicating the total sales generated by the movie studio and simulate a multiple linear regression with other variables, displayed in figure 4. In comparison to the regression obtained in figure 1, we observe that star_power, sequel and critic_rating are now statistically significant, as their respective p-values are less than the 5% level of significance. These findings suggest that the managers should concentrate their marketing strategies on the highly significant variables. The adjusted r-squared value is 0.7224, conveying that 72% of the variation in the dependent variable (revenue) is explained by the independent variables and the remaining 28% is left unexplained. This contrasts with the MLR model in figure 1, which comprises a 0.9863 r-squared value, thus having a greater goodness-of-fit. The dichotomy between goodness-of-fit and the trade off for simplicity in the smaller dataset will allow the manager to explore and visualise what affects the movie performance. Additionally, the ANOVA test in figure 5 confirms the significance of production, viewcount, disney, star_power and critic_rating. Viewcount has the greatest mean squared value of 12.651, coupled with the statistical significance, it is highly recommended that managers promote marketing strategies to obtain greater viewership and thus generate more revenue.

2. ***Which studio is the most influential in generating more sales?***

The most significant studio

In order to provide a holistic solution for research question 2, a multiple regression of revenue against the various studio variables have been conducted, displayed in figure 6. The incentive of this analysis is to interpret the most influential studio which generates the most sales. In accordance with the results showcased in figure 6, fox studios, warnerbros, disney and universal are highly statistically significant with corresponding p-values being less than the critical value in the 5% level of significance. Disney contains the largest estimated coefficient of 2.5545, which means that a 1 unit increase if the movie was produced by Disney, on average, increases the opening sales by $2.5545, ceteris paribus. When testing for goodness of fit, the results exemplifies a low adjusted r-squared value of 0.1734. This suggests that only 17% of the variation of the dependent variable (openingsales) is explained by the independent variables and the 83% is left unexplained. The ANOVA tests in figure 7 highlights a similar pattern, where Warnerbros, Disney and Universal are significant, but interestingly Fox studios is not significant from the ANOVA tests compared to the MLR model in figure 6. Furthemore, Disney studio contains the highest mean square value of 108.766. Despite the discrepancy created between the multiple regression analysis and ANOVA test, overall the manager should prioritise their marketing strategies by utilising Disney as their main studio movie producers. It is also crucial for the movie managers to investigate the lower movie sales corresponding to the insignificant variables in order to strike a balance and avoid overloading a particular studio.

Recommendations

1. ***Use predictive analytics to determine movie performance***

Predictive analytics can be used to optimise marketing campaigns, especially in the movie industry. Instead of relying on previous historical data on certain movies, information like the budget, genre, top actor and revenue can allow the studio to concentrate whether the movie has earned more or less than the expectation (Joseph, 2019). Comparison between expected revenue and the actual revenue will allow the data to be added to the predictive model to accurately predict in the future.

2. ***Improve customer personalisation***

It is imperative that this movie industry aims to create a more personalised experience for their viewers. Personalisation has increasingly become more of a competitive advantage for movie creators. A study from Deloitte exhibits that 36% of customers are interested in personalised products and 48% would be willing to wait longer to receive it (piesync, 2021). Tailoring a more personalised experience for viewers depending on the genre and longevity of the movie will be extremely effective in legitimising greater movie performance in creating more positive feedback from viewers. In relation to customer personalisation,

3. ***Include a greater sample size***

A major disadvantage in the analysis of movie performance and the identification of key significant variables was due to the small sample of 330 movies. Movies span from 2015 to 2018 as this relatively minor sample size is a possible causation for a lower adjusted r-squared value when determining the relationship between revenue and other independent variables. For future reference, a more accurate analysis should be implemented by balancing the tradeoff between number of variables (in order to reduce multicollinearity issues) and goodness-of-fit. The manager should provide marketing strategies to incorporate more significant variables to mitigate sampling bias and multicollinearity issues.

Conclusion

To improve movie performance will require strict management in marketing strategies to uplift the sales achieved. This research report integrates analytical tools to provide insights to the manager regarding which variables are important to improve and expand their research and efforts. We have developed two research questions corresponding to the managerial problem. Our main data findings suggest that there are multicollinearity issues with the different sales variables provided. To counteract this problem, PCA was implemented to create a new variable named ‘revenue’. Through multiple linear regression analysis, it was confirmed that the number of screens, viewcount, sequel, fox, paramount and disney studio as well as star_power were all significant variables which the movie industry should primarily focus their efforts on. Through the aforementioned analyses, it was also deduced that Disney studio is the most significant studio variable which movie managers should target for potential improvement in movie performances. It is highly recommended for the manager’s to incentivise on improving customer viewership and construct trailers which will engage customers to their movies which will correspondingly increase revenue. To conclude, if managers can adopt a customer-centric and personalised experience to attain more views, in conjunction in redistributing costs to poise viewership and revenue will undoubtedly increase movie performance.

Reference

Behrens, R., Foutz, N., Franklin, M., Funk, J., Gutierrez-Navratil, F., Hofmann, J. and Leibfried, U., 2020. Leveraging analytics to produce compelling and profitable film content. *Journal of Cultural Economics*, 45(2), pp.171-211.

Joseph, Sarah E., "What Makes a Movie Successful : Using Analytics to Study Box Office Hits" (2019). Chancellor's Honors Program Projects. https://trace.tennessee.edu/utk_chanhonoproj/2252

Murschetz, P., 2020. Movie Industry Economics: How Data Analytics Can Help Predict Movies' Financial Success. *ResearchGate*, [online] pp.2-3. Available at: <http://file:///C:/Users/User/Downloads/5871-ArticleText-19335-3-10-20201017.pdf> [Accessed 10 August 2021].

piesync, 2021. Why marketing personalization is so important - and how to use it. Available at: <https://www.piesync.com/blog/marketing-personalization/> [Accessed 12 August 2021].

Sankaralingam, G., 2017. *Data Analytics | Predictive Analytics | Movie Success Prediction*. [online] LatentView Analytics. Available at: <https://www.latentview.com/blog/using-analytics-to-predict-movie-success/> [Accessed 9 August 2021].

Wallace, F., 2019. How Data Science Is Used Within the Film Industry. [Blog] *KDnuggets*, Available at: <https://www.kdnuggets.com/2019/07/data-science-film-industry.html> [Accessed 9 August 2021].

Appendix

Figure 1: Multiple Linear Regression of weekendsales against remaining variables
Dependent variable: weekendsales

	Estimate	Std. Error	t-value	p-value	significance
(Intercept)	-6.655e+08	4.539e+08	-1.466	0.143648	
screens	1.133e+03	4.333e+02	2.615	0.009359	**
year	3.286e+05	2.251e+05	1.459	0.145469	
production	-1.041e+01	1.169e+03	-0.009	0.992901	
openingsales	1.280e+00	1.105e-01	11.582	< 2e-16	***
weeklysales	3.051e-01	3.433e-02	8.889	< 2e-16	***
viewcount	6.594e-02	1.918e-02	3.437	0.000668	***
teasertrailer	-2.633e+05	5.647e+05	-0.466	0.641327	
teaser	-9.051e+04	1.375e+06	-0.066	0.947579	
numbertrailer	5.775e+04	3.599e+05	0.160	0.872641	
sequel	1.784e+05	7.200e+05	0.248	0.804497	
fox	-7.688e+05	1.020e+06	-0.754	0.451448	
paramount	5.671e+05	1.082e+06	0.524	0.600547	
warnerbros	-4.063e+04	9.297e+05	-0.044	0.965169	
columbiasony	-1.790e+05	8.570e+05	-0.209	0.834659	
disney	2.341e+06	1.067e+06	2.195	0.028890	*
universal	-8.688e+05	7.707e+05	-1.127	0.260465	
other_studios	NA	NA	NA	NA	
star_power	-3.750e+02	2.546e+03	-0.147	0.883017	
genre	-2.033e+05	6.296e+05	-0.323	0.746954	
season	-8.081e+04	1.856e+05	-0.435	0.663567	
critic_rating	1.886e+04	9.627e+03	1.959	0.051028	.
advertising	5.406e-02	3.403e-02	1.589	0.113191	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Goodness of fit:

Multiple R^2	Adjusted R^2	F-statistic	DF	p-value
0.9872	0.9863	1129	308	< 2.2e-16

Figure 2: Correlation test between weekendsales, openingsales and weeklysales

	weekendsales	openingsales	weeklysales
weekendsales	1.0000000	0.9898050	0.9891759
openingsales	0.9898050	1.0000000	0.9884102
weeklysales	0.9891759	0.9884102	1.0000000

Figure 3: Principal Component Analysis (PCA)
Importance of components:

	PC1	PC2	PC3
Standard deviation	1.1349	0.9959	0.8487
Proportion of Variance	0.4294	0.3306	0.2401
Cumulative Proportion	0.4294	0.7599	1.0000

Figure 4: Multiple Linear Regression of revenue against remaining variables
Dependent variable: revenue

	Estimate	Std. Error	t-value	p-value	significance
(Intercept)	3.099e+01	9.537e+01	0.325	0.745473	
screens	2.847e-04	8.981e-05	3.170	0.001679	**
year	-1.660e-02	4.730e-02	-0.351	0.725792	
production	4.339e-04	2.448e-04	1.773	0.077221	.
viewcount	4.182e-08	3.306e-09	12.651	< 2e-16	***
teasertrailer	-1.570e-02	1.190e-01	-0.132	0.895090	
teaser	9.124e-02	2.885e-01	0.316	0.752048	
numbertrailer	8.553e-02	7.516e-02	1.138	0.256015	
sequel	7.262e-01	1.451e-01	5.006	9.33e-07	***
fox	-4.206e-01	2.136e-01	-1.969	0.049817	*
paramount	-5.642e-01	2.259e-01	-2.498	0.013014	*

warnerbros	-2.754e-01	1.954e-01	-1.409	0.159734	
columbiasony	-2.042e-01	1.803e-01	-1.133	0.258245	
disney	6.471e-01	2.221e-01	2.914	0.003828	**
universal	-3.512e-03	1.625e-01	-0.022	0.982767	
other_studios	NA	NA	NA	NA	
star_power	2.415e-03	5.172e-04	4.669	4.51e-06	***
genre	-7.672e-02	1.327e-01	-0.578	0.563616	
season	7.822e-02	3.862e-02	2.025	0.043686	*
critic_rating	6.752e-03	1.965e-03	3.437	0.000669	.***
advertising	3.416e-09	7.128e-09	0.479	0.632161	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Goodness of fit:

Multiple R^2	**Adjusted R^2**	**F-statistic**	**DF**	**p-value**
0.7384	0.7224	46.06	310	< 2.2e-16

Figure 5: ANOVA test of revenue against remaining variables
Dependent variable: revenue

	Df	**Sum Sq**	**Mean Sq**	**F-value**	**p-value**	**Significance**
screens	1	390.33	390.33	472.1433	< 2.2e-16	***
year	1	4.730e-02	-0.351	0.725792	0.959895	
production	1	2.448e-04	1.773	0.077221	3.298e-06	***
viewcount	1	3.306e-09	12.651	< 2e-16	< 2.2e-16	***
teasertrailer	1	1.190e-01	-0.132	0.895090	0.135690	
teaser	1	2.885e-01	0.316	0.752048	0.821666	
numbertrailer	1	7.516e-02	1.138	0.256015	0.094037	.
sequel	1	1.451e-01	5.006	9.33e-07	1.325e-06	***
fox	1	2.136e-01	-1.969	0.049817	0.153897	
paramount	1	2.259e-01	-2.498	0.013014	0.052175	.
warnerbros	1	1.954e-01	-1.409	0.159734	0.041128	*
columbiasony	1	1.803e-01	-1.133	0.258245	0.033512	*
disney	1	2.221e-01	2.914	0.003828	5.614e-06	***
universal	1	1.625e-01	-0.022	0.982767	0.533768	

star_power	1	5.172e-04	4.669	4.51e-06	1.400e-07	***
genre	1	1.327e-01	-0.578	0.563616	0.819343	
season	1	3.862e-02	2.025	0.043686	0.105686	
critic_rating	1	1.965e-03	3.437	0.000669	0.000549	***
advertising	1	7.128e-09	0.479	0.632161	0.632161	
Residuals	310	256.28	0.83			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 6: Multiple Linear Regression of revenue against studio variables
Dependent variable: revenue

	Estimate	**Std. Error**	**t-value**	**p-value**	**Significance**
(Intercept)	-0.7064	0.1341	-5.270	2.50e-07	***
fox	1.2603	0.3304	3.815	0.000163	***
paramount	0.5835	0.3677	1.587	0.113513	
warnerbros	1.3746	0.2877	4.778	2.69e-06	***
columbiasony	0.5556	0.3006	1.848	0.065493	.
disney	2.5545	0.3472	7.357	1.56e-12	***
universal	1.1102	0.2612	4.251	2.79e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Goodness of fit:

Multiple R^2	**Adjusted R^2**	**F-statistic**	**DF**	**p-value**
0.1885	0.1734	12.5	323	1.122e-12

Figure 7: ANOVA test of revenue against studio variables
Dependent variable: revenue

	Df	**Sum Sq**	**Mean Sq**	**F-value**	**p-value**	**Significance**
fox	1	9.02	9.021	3.6644	0.056469	.
paramount	1	0.12	0.122	0.0496	0.823918	
warnerbros	1	22.27	22.268	9.0452	0.002841	**
columbiasony	1	0.00	0.001	0.0003	0.987319	
disney	1	108.77	108.766	44.1804	1.276e-10	***
universal	1	44.49	44.487	18.0705	2.790e-05	***
Residuals	323	795.18	2.462			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1