Basic Applied Machine Learning and Predictive Modelling Techniques

Exploring Movie Metrics for Data-based Decisions

Viktor Huber, Remigiusz Trydulski, Carlo Scherrer

2025-01-23

Contents

1 Introduction **2**

2 Hypotheses & Research questions **2**

3 Analysis **3**

3.1 Fundamental insights 3

3.2 Linear Model 5

3.3 Generalised Linear Model with family set to Poisson 7

3.4 Generalised Linear Model with family set to Binomial 13

3.5 Generalized Additive Model 18

3.6 Neural Network 22

3.7 Support Vector Machine 23

4 Generative AI Reflection **28**

5 Conclusion **28**

6 Limitations **29**

7 Recommendations & Next Steps **29**

1 Introduction

In this report, the project group will investigate publicly available film data to define models that provide a comprehensive insight into the behavior and impact of attributes of films. The client of this study is active in the investment business and is interested in data-based decision-making support so that important key figures on potentially successful and popular films can be identified at the planning stage and its business activities can be adjusted on this basis.

Together with the client, the project group defined key questions (research questions) for the project and then tested models for their usability. The report begins with the identification of the research questions, followed by the basic insights into the data set before addressing and applying models.

In summary, the analysis of the models provided valuable insights for data-based decisions for the client. Further and just as important is the knowledge that the results of the individual models are limited with the provided data.

2 Hypotheses & Research questions

The client of this study is active in the investment sector of film projects. The commissioned study is intended to provide added value for its business activities in order to better manage its future project decisions by means of data-based decisions. Therefore research questions/hypotheses were formulated that address various aspects of the available film data:

1. How does the budget, the genres and the popularity influence the revenue of a film?

2. How do various factors influence the number of votes/ratings a movie receives?

3. Which factors contribute a movie achieving a return on investment (ROI) of at least 150%?

4. How do various factors influence the financial performance of films?

5. Can we create a model that can reliably predict popularity?

6. Is it possible to classify films on the basis of certain characteristics to decide whether a movie can achieve financial success?

The distribution of work by model is organized among the project members as follows:

Trydulski Remigiusz Piotr: Generalised Linear Model Poisson & Generalised Linear Model Binomial

Huber Viktor: Neural Network & Linear Model

Scherrer Carlo: Generalized Additive Model & Support Vector Machine

3 Analysis

The TMBD Movie Dataset contains information about 1287 different movies that were released between 1961 and 2015. They are categorized in 24 variables allowing an in-depth and representative analysis of the movie industry. The data set can be accessed on kaggle and is publicly available[1] .

3.1 Fundamental insights

In the following, a further basic analysis of the data set is carried out, which is supported by various visual representations to give the client an initial insight into the data.

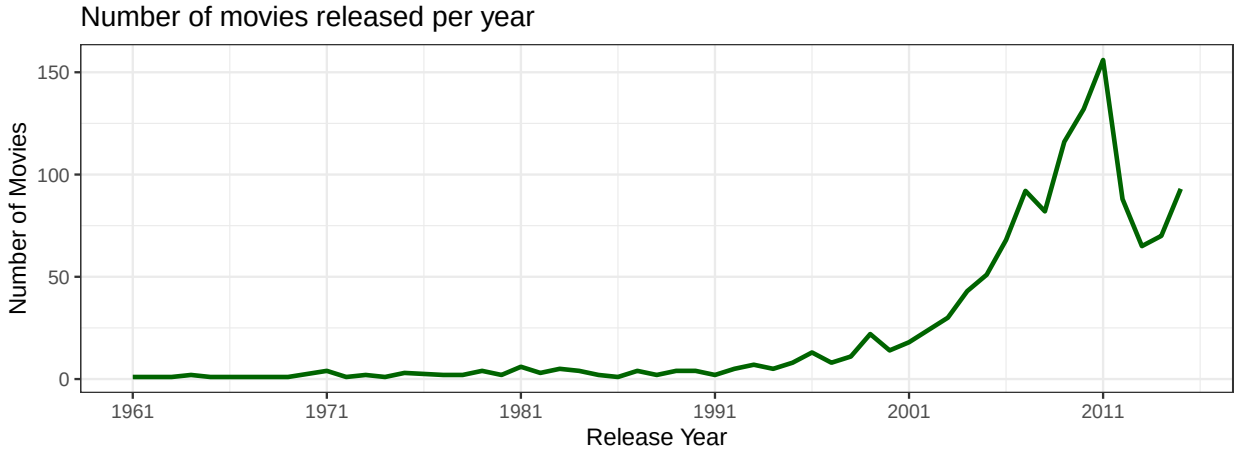


Figure 1: Number of movies released

Based on the data, there have been a total of 1287 movies released over the years. The year with the most movie releases was 2011 with a total of 156 movies released. The year with the greatest increase in the number of movies released was 2009, with an increase of 34 movies compared to the previous year.

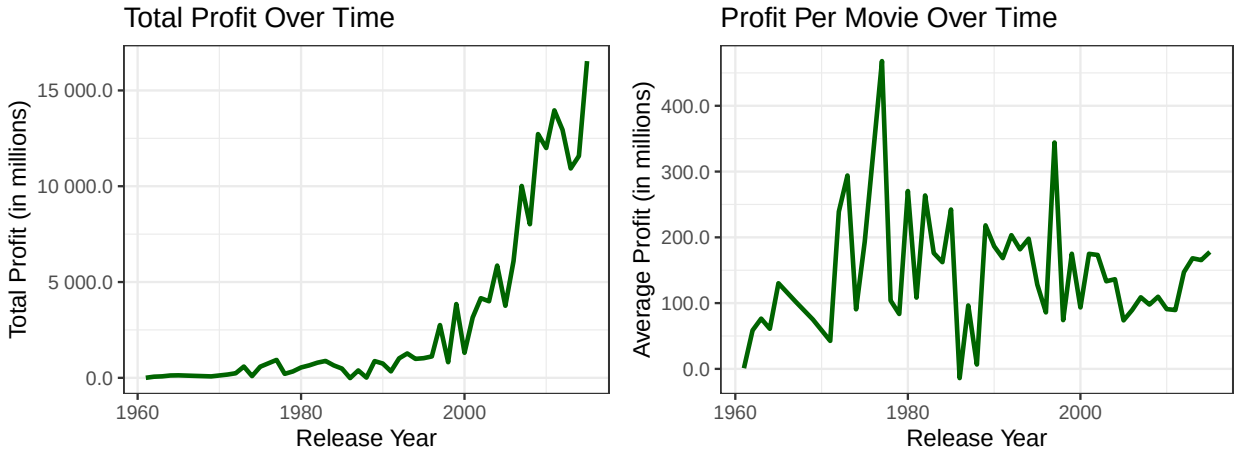


Figure 2: Profit over time

The average profit of the movies in the dataset amounts to $124.24 million, whereas the most profitable genre appears to be Fantasy.

[1] https://www.kaggle.com/datasets/successikuku/tmbd-movie-dataset/data

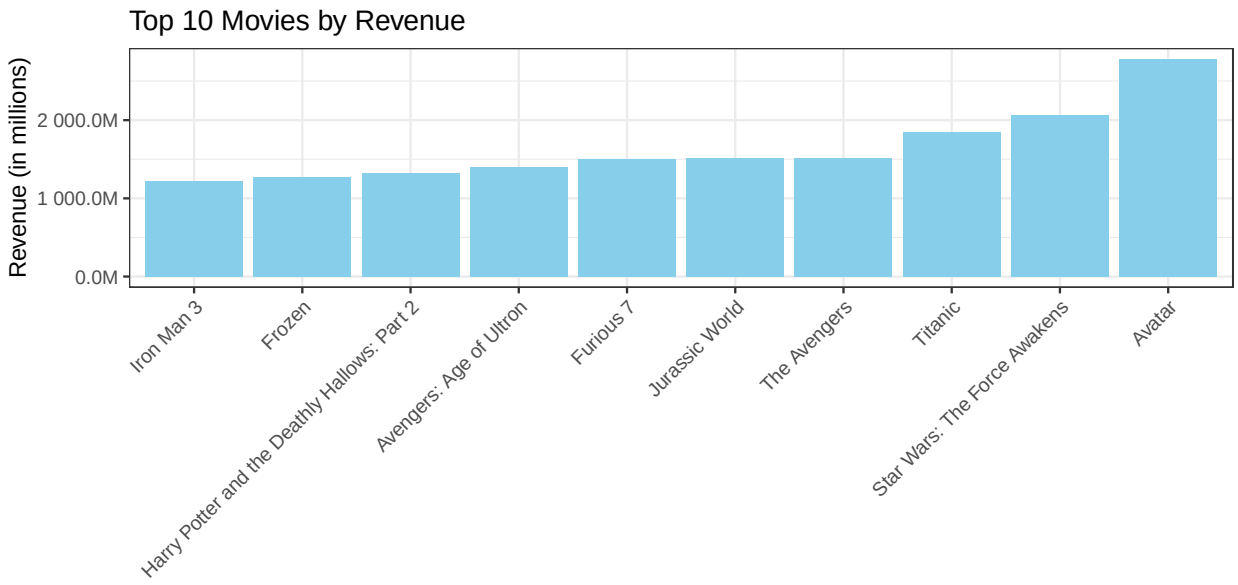


Figure 3: Top performing movies

The top 10 movies by revenue were released between the years 1997 and 2015, with a median release year of 2013. It is clear to see that newer films generate higher revenues. Possibly higher budget values are available and these could have an influence on the revenue. The following chart also provides an interesting initial insight, showing the distribution of revenues per genre and the number of movies per genre

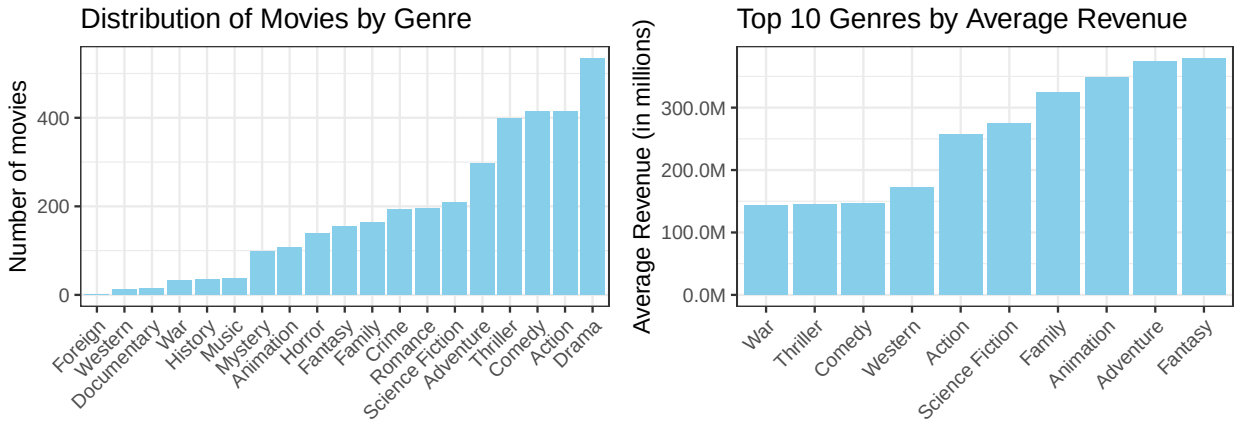


Figure 4: genre insights

It can be seen that the number of films per genre does not directly result in higher revenues. For example, most films belong to the drama genre, but in terms of average revenues, the drama genre is not represented in the top 10. Based on the initial basic findings of the project groups and the requirements of the client, the defined hypotheses are examined using suitable models in the following chapters.

3.2 Linear Model

3.2.1 Defintion of the model

A linear model is used to analyse the linear relationships between data. This means how a variable behaves when another variable changes. One speaks here of dependent and independent variables. This makes it relatively easy to establish correlations and thus to make predictions.[2]

3.2.2 Hypothesis

The aim of this chapter is to find out whether the revenue of the films is in any way linearly related to the budget. The objective is also to find out whether it is possible to make predictions based on the budget of a movie. In addition, the genres and popularity should be visualized and it should be determined whether these also have an influence. How does the budget, the genres and the popularity influence the revenue of a film?

3.2.3 Fitting a linear model

In the first step we try to fit a simple linear model, using just the budget and the revenue. The line has a positive slope, which means that there seems to be a linear relationship between the two values. In the summary it can be seen, that the p-value is significant, so you can deduce a positive influence of the budget on revenue here.

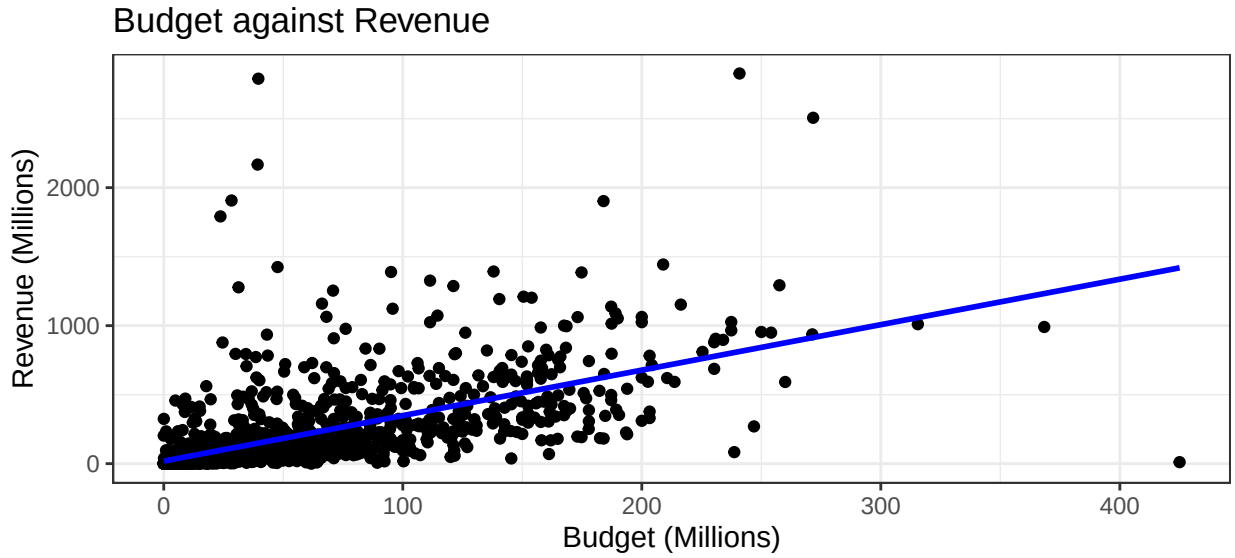
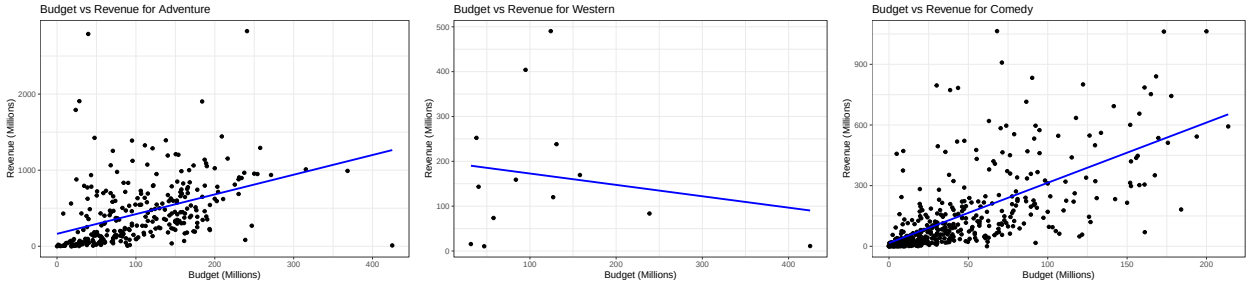


Figure 5: Linear Model

3.2.4 Fitting a linear model for each genre

In order to analyse the data in more depth, a separate linear model was calculated for each genre. This allows us to find out whether the genre could have an influence. As you can see from the individual plots, the higher the budget, the greater the positive effect for all genres. Just the western genre shows a negative trend. Only 3 Plots are shown below to demonstrate the effects.

[2]Linear model



When all genres are analysed separately, a linear relationship can be seen for each model. However, this is slightly different if all genres are included in the same model. There you can see the clear correlation that the budget has on the revenue. The genres do not all have a significant correlation with a p-value of less than 0.05. You can see that adventure, western and comedy are the only genres that are significant. Adventure is even the only one that has a positive influence. The other two have a negative influence.

Observations	1287
Dependent variable	revenue_millions
Type	OLS linear regression

F(20,1266)	44.84
R^2	0.41
Adj. R^2	0.41

	Est.	S.E.	t val.	p
(Intercept)	65.99	20.77	3.18	0.00
budget_millions	2.84	0.15	19.02	0.00
Action	-28.88	17.92	-1.61	0.11
Adventure	111.57	19.17	5.82	0.00
‘Science Fiction‘	12.14	19.39	0.63	0.53
Thriller	-12.15	16.94	-0.72	0.47
Fantasy	6.09	22.01	0.28	0.78
Crime	-14.15	20.26	-0.70	0.49
Western	-255.65	64.77	-3.95	0.00
Drama	-33.23	16.65	-2.00	0.05
Family	9.02	28.46	0.32	0.75
Animation	-7.43	32.60	-0.23	0.82
Comedy	-50.58	16.84	-3.00	0.00
Mystery	-30.04	24.95	-1.20	0.23
War	-8.22	43.05	-0.19	0.85
Romance	9.76	19.99	0.49	0.63
History	-59.40	41.69	-1.42	0.15
Horror	1.53	23.65	0.06	0.95
Music	7.35	38.54	0.19	0.85
Documentary	-56.29	64.32	-0.88	0.38
Foreign	-13.77	237.04	-0.06	0.95

Standard errors: OLS

3.2.5 Fitting a linear model for seperate popularity levels

For the third part, we try to fit a linear model for each popularity level to find out, if they differ from another. You can see here at all levels that the revenue also increases as the budget rises. However, at the high level, the gradient is steeper than at the other levels. it can therefore be deduced that a very popular film with a high budget should also generate a correspondingly higher revenue.

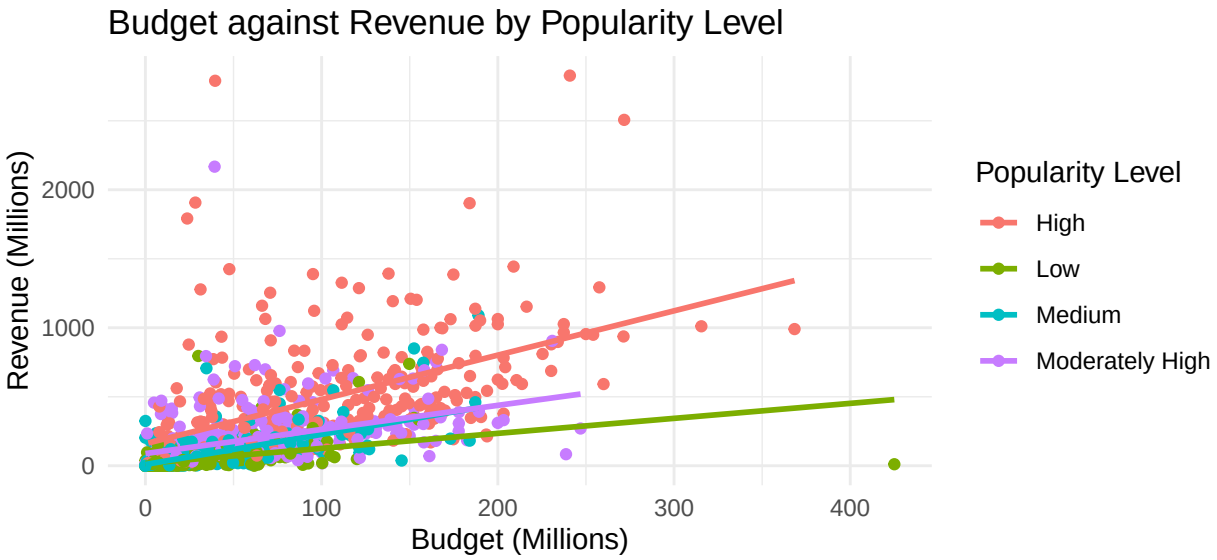


Figure 6: Linear Model on popularity levels

3.2.6 Conclusion

A comparison of the 3 models shows that they can only predict the desired values to a limited extent. The model fit is relatively low for all models and lies between 0.38 and 0.46, which is less than 50%. This means that with caution, these models can be used to make predictions, but you must be aware of the limitations. Nevertheless, a linear relationship can be recognised in all three models.

3.3 Generalised Linear Model with family set to Poisson

3.3.1 Model description and objective

The Generalized Linear Model (GLM) with a Poisson distribution is an established statistical approach used to analyze count data, where the response variable represents the number of occurrences of a specific event. In this analysis, we utilize the Poisson GLM to assess how various factors, e.g. budget, runtime, release year, influence the number of votes a movie receives. For example, the model can reveal whether higher budgets or specific genres tend to attract more votes, or how the influence of a director might affect voter engagement. Through this analytical framework, we aim to uncover patterns and trends that could support our client's strategic decisions in film production, enhancing the ability to predict and potentially increase viewer engagement.

3.3.2 Potential Count Variables

A first analysis of numeric variables showed that there are six potential count variables. They will be further inspected using histograms to analyze the distribution. The aim is to assess if they are suitable for the

poisson model.

The following could be read from the histograms:

- Right skewed distribution for variables budget, revenue and vote count.
- Left skewed distribution for release year.
- Rather central distribution for runtime.

Further, a brief variance analysis showed overdispersion in the data as it is commonly found with count data compared to the Poisson model's assumptions. There is significant or quite significant overdispersion for filtered potential count variables: budget, revenue, runtime, and vote count. As expected, release year appears to be a categorical rather than a count variable. Also, log transformation needed to be applied to normalize distributions of right-skewed variables.

3.3.3 Conclusion from Count Variable Analysis

It can be concluded that only the variable vote count from the given dataset meets the requirements for a count variable, i.e. can be used as the dependent variable in the poisson model. The other variables, even though numeric and discrete, can not be considered as count variables, and thus do not qualify as variables for the dependent role. However, they might be considered as predictors in further analysis when fitting the Poisson model.

3.3.4 Model Fitting

```
# Fit a Poisson model with 'vote count' as the dependent variable
poisson_model_1 <- glm(vote_count ~ budget + as.factor(release_year) + runtime,
                       family = poisson(link = "log"), data = d.data_raw_rt)
# summary(poisson_model_1)

# Fit a Negative Binomial model for comparison
nb_model_1 <- glm.nb(vote_count ~ budget + as.factor(release_year) + runtime,
                     data = d.data_raw_rt)
# summary(nb_model_1)
```

All predictors in the poisson model seem to affect the vote count. The significant coefficients for different levels of release year indicate that there have been year-to-year variations in vote counts that are significant.

In the negative binomial model, the variables budget and runtime also seem to have an effect on vote count, as with the poisson model. However, only a few release years seemed to be relevant.

3.3.5 Residuals vs Fitted

Poisson model: There is a visible pattern where residuals fan out as the predicted values increase. This suggests that the variance of the residuals is not constant (heteroscedasticity). Also, there are a few points with residuals significantly deviating from zero, particularly for higher predicted values. This indicates that factors other than budget, release year, and runtime might play a significant role in influencing vote counts, especially for more popular movies.

Negative binomial model: Eight rather extreme, but correct values of the predictor variables, make it challenging to plot residuals effectively. However, removing these observations from the dataset does not seem necessary due to their minimal proportion. To confirm, a brief test was conducted to assess the impact of excluding these extreme values. The pseudo R-squared value remained unchanged, indicating no impact on the model.

3.3.6 Q-Q Residuals

Poisson model: The residuals largely follow the theoretical line, but with deviations at both tails (lower left and upper right). This indicates some outliers and possible skew in the distribution of residuals. This could also signal that the Poisson assumption of the mean being equal to the variance is violated (as inspected earlier). Additional factors could enhance the accuracy.

Negative binomial model: The points largely follow the reference line, but there are deviations, especially in the upper tail (high vote counts), indicating heavy-tailed residuals. This deviation suggests that there are extreme values among the residuals that are not well modeled by the assumed distribution.

3.3.7 Residuals vs Leverage

Poisson model: Most data points are clustered to the left, suggesting low leverage. There are a few points with higher leverage and a couple with high Cook's distances. These points (movies) may influence the model heavily.

Negative binomial model: The Cook's distance lines suggest thresholds for identifying influential movies and several are close to or exceed these thresholds, particularly the ones identified with labels, indicating they might be influential. Generally, the points aren't spread too far across the leverage spectrum, but those with higher Cook's distance are of concern as those movies may disproportionately affect the model.

3.3.8 Model Comparison

Table 1: AIC Values for Initial Models

Model	AIC
Poisson Model 1:	821590.27
Negative Binomial Model 1:	19490.73

The Negative Binomial model has a significantly lower AIC value compared to the Poisson model (19490 vs 821590). This large difference in AIC values suggests that the Negative Binomial model fits the data much better than the Poisson model.

The Poisson model's higher AIC value could indicate that it is not adequately capturing the variability in the data or is too simplistic (underfitting), particularly if there is overdispersion present in the data. The Negative Binomial model, being more flexible with regard to the variance, can handle overdispersion better, which is likely reflected in the lower AIC.

3.3.9 Model Improvement

Given the model diagnostics, adding new predictors and interaction might be beneficial to improve the models. Based on the general topic understanding and the insights gained in the project so far, various models will be developed and compared. The variables revenue, director and genres will be added to both models. For practical reasons and to avoid potential issues related to high dimensionality, only the first genre will be considered for movies with multiple genres, strongly assuming it's the movies' main genre.

3.3.10 Model Fitting After Refinement

```r
# Add "revenue" and "Genre" to the model
poisson_model_2 <- glm(vote_count ~ budget + as.factor(release_year) + runtime + revenue
                       + first_genre,
                       family = poisson(link = "log"), data = d.data_genre_rt)
# summary(poisson_model_2)

# Additionally, add interaction between "budget" and "runtime"
poisson_model_3 <- glm(vote_count ~ budget * runtime + as.factor(release_year) + revenue
                       + first_genre,
                       family = poisson(link = "log"), data = d.data_genre_rt)
# summary(poisson_model_3)

# Additionally, add "director"
poisson_model_4 <- glm(vote_count ~ budget * runtime + as.factor(release_year) + revenue
                       + first_genre + director,
                       family = poisson(link = "log"), data = d.data_genre_rt)
# summary(poisson_model_4)
```

Given the lower residual deviance and AIC in Model 4, it is clear that this model is the most effective among the three at capturing the dynamics of the data. The additional variables and interactions included in model 4 significantly enhance its explanatory power, making it the preferred model for understanding and predicting the factors influencing vote counts in the given dataset:

- Poisson Model 2 -> Null deviance: 1547234; Residual deviance: 646572; AIC: 656889
- Poisson Model 3 -> Null deviance: 1547234; Residual deviance: 609898; AIC: 620218
- Poisson Model 4 -> Null deviance: 1547234; Residual deviance: 133577; AIC: 145466

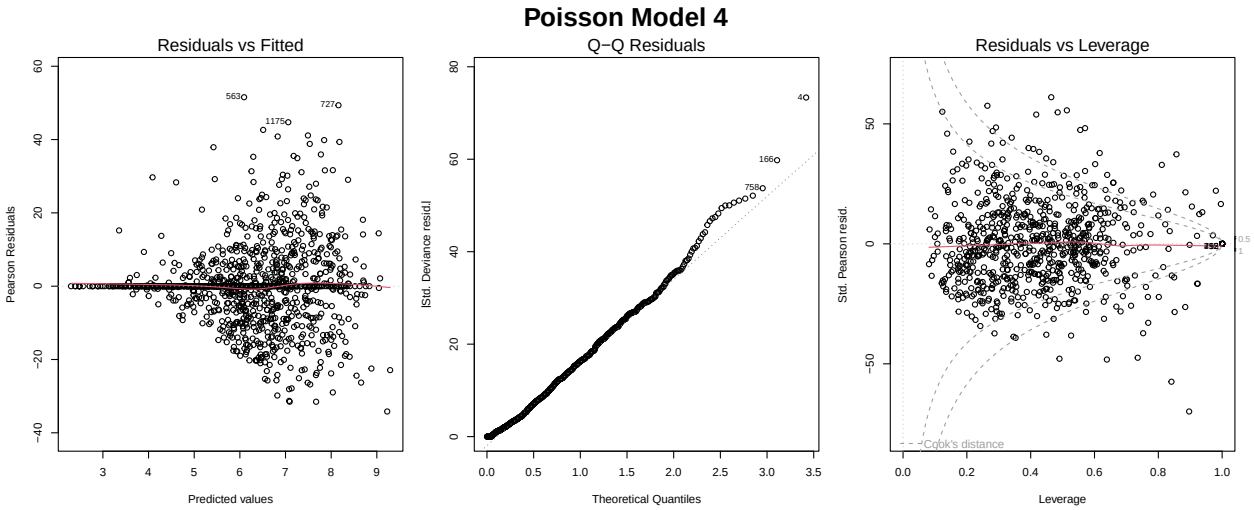
In conclusion, the Poisson Model 4 not only fits the data better than the other models, but also efficiently handles the complexity introduced by additional predictors and the interaction between budget and runtime.

```r
# same approach as for the poisson models 2-4
nb_model_2 <- glm.nb(vote_count ~ budget + as.factor(release_year) + runtime + revenue
                     + first_genre, data = d.data_genre_rt)
# summary(nb_model_2)

nb_model_3 <- glm.nb(vote_count ~ budget * runtime + as.factor(release_year) + revenue
                     + first_genre, data = d.data_genre_rt)
# summary(nb_model_3)

nb_model_4 <- glm.nb(vote_count ~ budget * runtime + as.factor(release_year) + revenue
                     + first_genre + director, data = d.data_genre_rt)
# summary(nb_model_4)
```

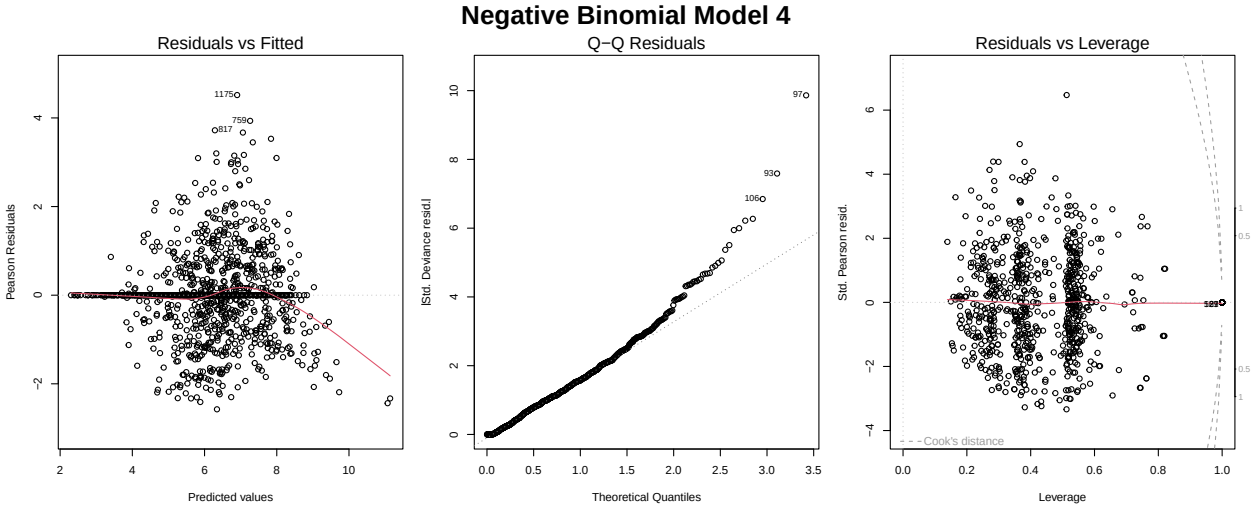
The analysis of the negative binomial models suggests a steady enhancement in model performance from model 2 to model 4. Model 4, with the lowest residual deviance and AIC, is clearly the most robust model, suggesting that the variables and interactions introduced in this model are essential in capturing the dynamics influencing the number of vote counts effectively.

- Negative binomial model 2 -> Null deviance: 3062.1; Residual deviance: 1421.7; AIC: 19160
- Negative binomial model 3 -> Null deviance: 3142.2; Residual deviance: 1418.6; AIC: 19123
- Negative binomial model 4 -> Null deviance: 13983; Residual deviance: 1305; AIC: 18570

3.3.11 Model Diagnostics After Refinement

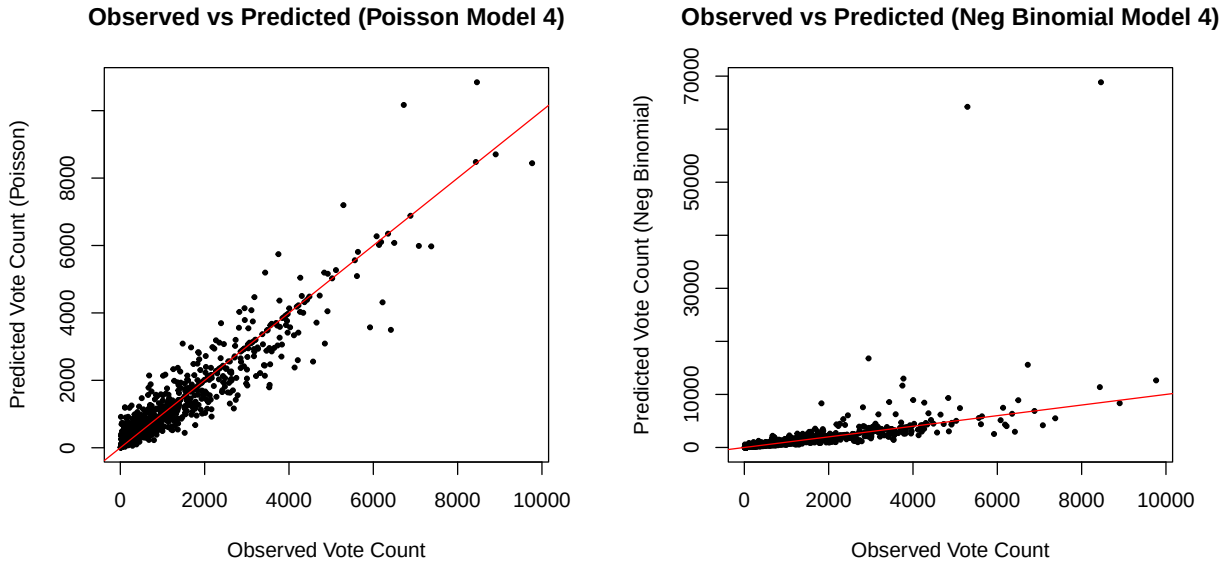


Examining poisson models, the spread of residuals around the fitted line seems to be narrower with Model 4 compared to Models 2 and 3, suggesting better consistency in variance and less apparent overdispersion. Also, Model 4 shows better adherence to the line with fewer deviations, especially in the middle quantiles, but there are still some movies with very high or low vote counts that the model does not predict accurately. With a few high leverage points, there are similar patterns at model 4 as at the previous models. However, the overall influence on the model seems to be limited.



Considering the diagnostic plots for the negative binomial models, Model 4 appears to be the best overall choice. It manages to maintain a reasonable fit across the central range of the data, showing little pattern and uniform spread in the Residuals vs. Fitted plot. Model 4 also handles outliers effectively, as indicated by the Q-Q plot, which shows fewer and less extreme deviations from normality. Furthermore, it controls for influential observations quite well, as seen in the Residuals vs. Leverage plot, where the distribution of residuals is more balanced and less affected by high leverage points than with the other models.

3.3.12 Predictive Power Check After Refinement



The plots show that the Poisson model predictions align better with the observed data along the lower range of counts, but appear to underestimate as the vote counts increase. There is increasing spread in the residuals as the observed vote count increases, which indicates that the Poisson model does not adequately handle the variance in the data.

The negative binomial model shows that it better handles the overdispersion present in the data. The spread in residuals is more consistent across different levels of observed vote counts, and it captures higher counts more accurately than the poisson model. Although there is still some underestimation for higher vote counts, the overall fit is improved compared to the poisson model.

3.3.13 Final Conclusion

The analysis showed that certain factors, such as budget, runtime, genre, and director can influence the number of votings a movie receives. Consequently, considering these factors in movie production seems beneficial to increase the vote counts. Release year of the movie, even though having somewhat an impact on the number of votes in the past, might not be a valuable factor for planning new movies.

After comprehensive diagnostics, the negative binomial Model 4 has emerged as the best model due to its superior handling of overdispersion. This is evidenced by its significantly lower residual deviance and AIC compared to the Poisson model. The negative binomial model provides a more accurate fit, especially for higher counts, and better captures the data's variability. Also, it offers a more reliable representation of the underlying relationships in the data. This model is intended to serve as a starting point for initial decision making. It is recommanded to further investigate e.g. if there are optimal ranges for budget and rutime. Also, exploring other approaches, such as GAMs, could help capture more complex relationships in movie industry/ data, and thus further enhance the model performance, especially to better explain higher vote counts.

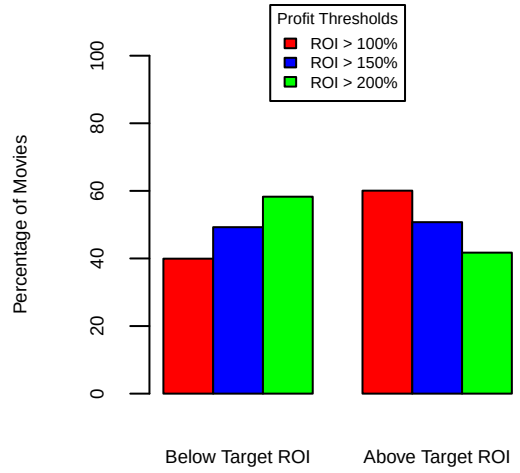
3.4 Generalised Linear Model with family set to Binomial

3.4.1 Model description and objective

In this section, we employ the Generalized Linear Model (GLM) with a binomial distribution to help our client to determine the key factors contributing to the financial success of movies. This approach is ideal for addressing binary outcomes, such as whether a movie achieves significant profitability or not. We integrate critical predictors like budget, genre or the release timing to understand their impact on a movie's financial performance. Our goal is to equip our client with the ability to predict financial success effectively, optimizing investments and maximizing returns in future projects.

3.4.2 Defining the outcome variable for the model

The first step of the analysis is to divide the movies into "profitable" and "not profitable". The calculation shows that almost 80% of the movies achieved a profit. However, from a business perspective, merely covering the costs is usually not considered financial success. For that reason, further analysis is needed to better distinguish between financially successful movies and less successful. Consequently, it is worth to analyze the return on investment (ROI) profitable movies achieved.



The barplot reveals the following:

- ROI $>$ 100%: 40% of the movies are below, 60% above
- ROI $>$ 150%: 49% of the movies are below, 51% above
- ROI $>$ 200%: 58% of the movies are below, 42% above

Considering that the focus is on the financial success of movies compared to other movies, ROI of $>$ 100% seems too low, given that 60% of movies achieved it. The ROI $>$ 150% will be taken as a threshold to distinguish between financially successful movies (above target ROI) and movies that are below this ROI. This seems to be a reasonable choice from a business perspective, but also considering statistical robustness. A balanced target variable / classes (ca. 50/50%) can prevent the model from developing a bias toward the majority class, which can distort predictive accuracy and the interpretability of model coefficients. A brief check reveals that 634 movies have a ROI $<$ 150% and 653 movies are above, which matches with the results from the barplot.

3.4.3 Selecting Predictors

Considering the given dataset, the following variables might be relevant predictors: Budget, Runtime, Genre and Release Season. This will be inspected in this section.

When examining the genre table, there are some genres represented only by a few movies. This needs to be observed when designing initial models and how each genre might influence financial success. It might be beneficial to combine sparse genres for model refinement given the unequal representation.

Further, log transformation of budget was necessary due to right-skewness. Runtime was normalized due to different scale compared to budget, needed for further analysis. The correlation of 0.366 between budget and runtime suggests a moderate positive relationship. As the budget increases, the runtime of movies also tends to be longer, but the relationship is not very strong. Given the moderate correlation, both variables can be included in the models for now.

3.4.4 Model Fitting

```
# Model 1 (with main predictors)
model_1 <- glm(financial_success ~ log_budget + norm_runtime + factor(main_genre)
                  + factor(season),
                family = binomial(link = "logit"), data = d.data_raw_rt)
# summary(model_1)

# Model 2 (with interaction)
model_2 <- glm(financial_success ~ log_budget + norm_runtime + factor(main_genre)
                  + factor(season) + log_budget:factor(main_genre),
                family = binomial(link = "logit"), data = d.data_raw_rt)
# summary(model_2)

# Model 3 (with polynomial terms to capture potential non-linear effects)
model_3 <- glm(financial_success ~ poly(log_budget, 2) + poly(norm_runtime, 2)
                  + factor(main_genre) + factor(season),
                family = binomial(link = "logit"), data = d.data_raw_rt)
# summary(model_3)
```

3.4.4.1 Analysis of Coefficients **Model 1:** The coefficient for log_budget is -0.16354, indicating a slightly negative relationship between budget and the probability of financial success, i.e. ROI $> 150\%$. This suggests that, in isolation, increases in budget at a certain point/ amount could decrease the probability of financial success. The variable Runtime is significant with a coefficient of 0.39690, suggesting that longer movies tend to have a higher probability of financial success. The genres Adventure, Animation and Horror show a relevant positive effect on financial success. Particularly, movies in the Horror genre have a significantly higher likelihood of being financially successful, as indicated by a coefficient of 0.81461. Music genre shows a huge negative coefficient, but with a very large standard error, possibly due to few data points in this category as seen in frequency tables. Also, its p-value (0.96580) is not relevant. None of the seasons seem to be influential predictors.

Model 2: The coefficient for log_budget is 0.1407 with a p-value of 0.290689, suggesting that the main effect of budget alone does not significantly impact financial success when not interacting with other factors. The coefficient of runtime is 0.4260 and is highly significant, indicating a strong positive relationship between runtime and financial success, similar to Model 1. Some genres show significant impact on financial success: Horror, Drama, Comedy and Thriller. The interaction terms of Dramas and Horrors are notably significant and negative, indicating that while dramas and horrors tend to be successful, their success is less sensitive to

increases in budget compared to other genres. A decrease of residual deviance from null deviance suggests that the model explains a significant amount of variability in the data.

Model 3: The coefficient of poly(log_budget, 2)1 is -4.92298 with a p-value of 0.07725, indicating a marginal non-linear effect of budget on financial success. This suggests that the relationship between budget and success is not straightforward and may involve diminishing returns. The coefficient poly(log_budget, 2)2 is 10.67026, highly significant ($p < 0.00001$). This indicates a pronounced non-linear effect, possibly suggesting an optimal budget level. The coefficient poly(norm_runtime, 2)1 is 12.48737, also highly significant ($p < 0.00001$), indicating a strong non-linear positive relationship between runtime and financial success, possibly suggesting an optimal range of runtime for maximizing success. The coefficient poly(norm_runtime, 2)2 is 1.01164 with a p-value of 0.67891, suggesting no significant secondary curvature in the relationship between runtime and success. Horror continues to show a strong positive effect on financial success ($p = 0.00154$), consistent with previous models. Adventure, Animation, and Romance also demonstrate significant positive impacts on financial success. None of the seasons show significant effects on financial success, consistent with findings from previous models.

3.4.4.2 Comparing AIC & BIC

Table 2: AIC and BIC Values for Initial Models

Model	AIC	BIC
Model 1	1765.939	1884.621
Model 2	1770.865	1977.268
Model 3	1747.339	1876.341

Model 3 shows the lowest AIC value, suggesting a better fit to the data, likely due to the inclusion of polynomial terms which capture more complexity. Moreover, model 3 has also the lowest BIC of the three models, indicating that it is the best model among the three in terms of balancing goodness of fit with complexity. This model includes polynomial terms for log_budget and norm_runtime, suggesting that these non-linear transformations capture important patterns in the data more effectively than the linear and interaction terms used in model 1 and 2.

3.4.4.3 Cross Validation of models

Table 3: Initial Model Performance Metrics

Model	ROC	Sensitivity	Specificity
Model 1	0.5727898	0.5756696	0.5342424
Model 2	0.5837758	0.5806300	0.5342424
Model 3	0.5948908	0.6119048	0.5252448

3.4.5 Assessment of the initial models

From the results above, Model 3 has the lowest AIC and a very competitive BIC score, suggesting it's potentially the best fit among the three. Furthermore, Model 3 shows the highest ROC value in cross-validation, which suggests it generalizes better than the other models. As a conclusion, especially model 3 will be further examined and optimized. However, there are still some insignificant and/or underrepresented predictors in the models that need to be addressed.

3.4.6 Analysis for model refinement

Given the very few movies and/or very low proportion of financial success, some genres will be put in one category. The findings underline the importance of genre-specific budgeting strategies. Investing in genres that respond well to budget increases and adjusting strategies for those that do not optimize financial outcomes. Hence, adding an interaction between budget and genres as in model 2 may be helpful. The significant polynomial terms for budget and runtime suggest that both factors influence financial success in complex ways. For budget, there may be an optimal level of investment that maximizes financial returns, while for runtime, certain lengths may be more favorable than others. Specific genres like Horror, Adventure, Animation, and Romance show a propensity to be more financially successful. This emphasizes the importance of genre choice in film production. The lack of significant seasonal impact suggests that the timing of a movie release within the year may not be as critical to its financial outcome as genre, budget and runtime.

3.4.7 Model Refinement Part 1

```
# New model 1 - grouped genres
new_model_1 <- glm(financial_success ~ poly(log_budget, 2) + poly(norm_runtime, 1)
                   + factor(main_genre_new),
                 family = binomial(link = "logit"), data = d.data_raw_rt)
#summary(new_model_1)

# New model 2 - only statistically significant genres
model_2_filtered_data <- subset(d.data_raw_rt, main_genre_new
                                %in% c("Adventure", "Animation", "Horror", "Romance"))

new_model_2 <- glm(financial_success ~ poly(log_budget, 2) + poly(norm_runtime, 1)
                   + factor(main_genre_new),
                 family = binomial(link = "logit"), data = model_2_filtered_data)
#summary(new_model_2)
```

The first grouping of genres ("Mystery_Crime", "Doc_Hist_War_Mus_West" and "SciFi_Fantasy") does not seem to improve the model. Hence, further grouping might be needed in order to improve the model. However, the new model 2 where only statistically relevant genres are considered, shows a massive improvement in regards to AIC and BIC values.

3.4.8 Model Refinement Part 2

```
# New Model 3 - further grouped genres
new_model_3 <- glm(financial_success ~ poly(log_budget, 2) + poly(norm_runtime, 1)
                   + factor(main_genre_new),
                 family = binomial(link = "logit"), data = d.data_raw_rt)
# summary(new_model_3)

# New Model 4 - further grouped genres only statistically significant genres
model_4_filtered_data <- subset(d.data_raw_rt, main_genre_new
                                %in% c("Adventure", "Animation_Family", "Horror"))

new_model_4 <- glm(financial_success ~ poly(log_budget, 2) + poly(norm_runtime, 1)
                   + factor(main_genre_new),
                 family = binomial(link = "logit"), data = model_4_filtered_data)
```

```r
# summary(new_model_4)

# New Model 5 - include interaction between budget and genre
new_model_5 <- glm(financial_success ~ poly(log_budget, 2) * factor(main_genre_new)
                   + poly(norm_runtime, 1),
                 family = binomial(link = "logit"), data = model_4_filtered_data)
# summary(new_model_5)
```

3.4.8.1 Comparing AIC & BIC of the refined models

Table 4: AIC and BIC Values for Refined Models

Model	AIC	BIC
New Model 1	1738.3678	1815.7688
New Model 2	379.2307	404.8957
New Model 3	1738.2020	1800.1228
New Model 4	367.7284	389.5799
New Model 5	367.7584	404.1775

3.4.9 Cross Validation of the refined models

Table 5: New Model Performance Metrics

Model	ROC	Sensitivity	Specificity
New Model 1	0.6003227	0.6008929	0.5343124
New Model 2	0.6429763	0.3469697	0.7973856
New Model 3	0.5956874	0.6214782	0.5344755
New Model 4	0.6354027	0.3530303	0.7786765
New Model 5	0.6356618	0.4000000	0.7845588

3.4.10 Final Conclusion

Based on the evaluation of the refined models, New Model 5 emerges as the overall best fit for predicting the financial success of movies. It shows one of the lowest AIC and BIC values, indicating an optimal balance between model fit and complexity. Additionally, New Model 5 demonstrates strong performance metrics, with a high ROC of 0.6357, suggesting that it can effectively distinguish between movies achieving ROI>150% and those that do not. Its sensitivity of 0.4000 and specificity of 0.7846 further confirm its balanced approach in correctly identifying both successful and unsuccessful movies. While New Model 2 achieves the highest ROC (0.6429) and specificity (0.7974), its slightly higher AIC and BIC values compared to New Model 5 indicate a trade-off in model simplicity. New Model 3, with the highest sensitivity (0.6215), excels in identifying successful movies but lacks the overall balance in AIC, BIC, and ROC.

To sum up, certain genres like Adventure, Animation/Family, and Horror tend to have higher chances of being financially successful. Additionally, there is evidence suggesting an optimal budget level, although this ideal amount appears to vary by genre. Furthermore, a longer movie runtime generally seems to have a positive influence on financial success, but there also appears to be an optimal range for runtime.

3.5 Generalized Additive Model

Our aim is to gain a comprehensive understanding of the financial performance of films on the basis of various factors. In this chapter, the investigation is based on a generalized additive model. We begin by examining the relationship between budget and revenue in different film genres, using both linear and non-linear modeling techniques. The number of variables considered that could have a possible influence is then expanded so that a generalized additive model can be created at the end. The model is analyzed for both scenarios, with and without interactions between the selected variables.

3.5.1 Linear and Non-linear Models for Predicting Movie Revenue

As a first step, we start by fitting linear models for each genre to understand how well budget predicts revenue, evaluating the effectiveness of these models using R^2 values. The basic idea is, that the linear regression highlights the differences in predictive power between genres.

In a second step, the models were then adjusted to take into account the relationship between budget and revenue across all genres in order to minimise the influence of the number of observations.

In a third step we extend the analysis to non-linear relationships using polynomial regression for quadratic effects, this analysis provides a deeper insight into the impact of budget on revenue with more complex patterns. The following analysis carried out (entire section can be seen in the code) can be summarised with the following points:

Genre-specific Analysis:

- The fit is highly dependent on which model (linear or non-linear) is used.
- The R^2 values vary significantly across genres, indicating that the budget's explanatory power on revenue could be genre-dependent.
- Some genres, like music and western show an almost perfect fit.
- For other genres like horror and Adventure the R^2 values are quite low.

This illustrates the strong effect of the number of observations, so that an overall model is better suited for general statements.

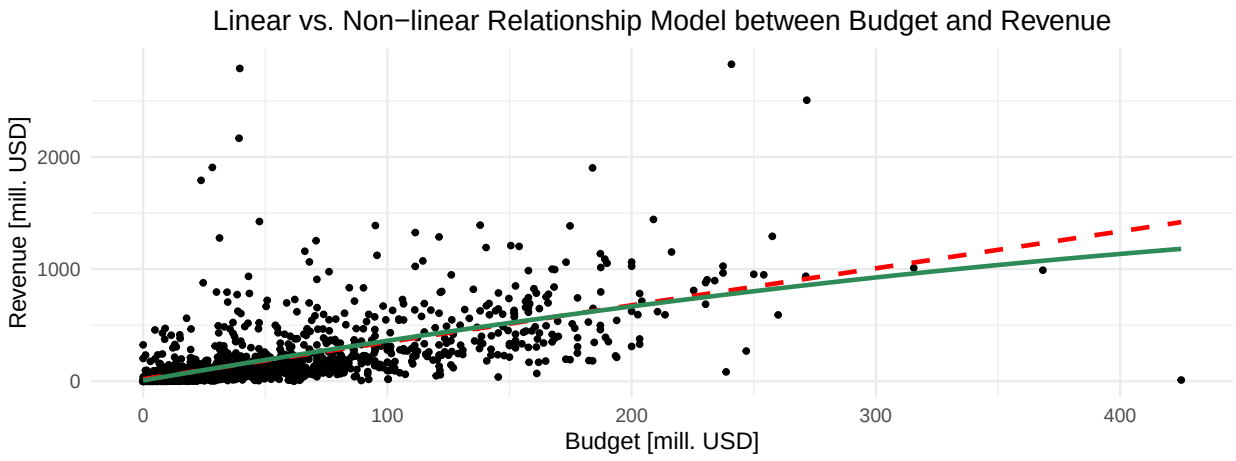


Figure 7: Grapical analysis - linearity vs. non-linearity

Linear Regression Model:

- The linear regression model indicates a positive relationship between the budget and revenue of movies.

- The coefficient for budget_adj is suggesting that for every million-dollar increase in the budget, the revenue increases by approximately 0.9 million dollars.
- Approximately 50% of the variance in revenue is explained by the budget alone and the residuals exhibit a significant spread, indicating that many factors beyond budget influence movie revenue.

Polynomial Regression Model (Quadratic Effect):

- This model has only a slightly higher R^2 value.
- Overall the model indicates that while there is a slight non-linear relationship, the improvement over the linear model is only marginal.

Graphical Analysis:

- The red dashed line represents the linear regression fit and many data points lie far away from the line. This is an indicator that the model does not capture all of the variability.
- The green solid line represents the polynomial regression fit and captures some non-linear patterns, due to the slight bend.
- However, the analysis has shown that the improvement is only marginally better than the linear model.

These findings suggest that further investigation of other potential influencing factors and more complex modelling is necessary. The next step will therefore be to examine the relationship between the running time of films and their revenues to see if there are any non-linear patterns that can be better captured by more flexible models.

3.5.2 Is there an influence of the variable run-time?

The film duration is selected as a further influencing variable. The reason for selecting this variable is a recent survey[3] which shows that society prefers films of a certain length, namely 92 minutes. Therefore, the film length should have a significant influence on the revenue and further increase the model quality.

Based on this assumption, the hypothesis is first examined using the t-test to determine whether the film length really has an influence on the revenue and can be included in the model.

1. **Null hypothesis (H0)**: There is no difference in revenue between films with a running time of more than 92 minutes and films with a running time of 92 minutes or less: $H_0 : \mu_1 = \mu_2$
2. **Alternative hypothesis (H1)**: Films with a running time of more than 90 minutes generate higher revenues than films with a running time of 90 minutes or less: $H_1 : \mu_1 > \mu_2$

Here are the key results and their interpretation:

t-value: 5.06 & **p-value:** 3.096e-07

Mean values of the samples:

- Mean value of group 1 (films > 90 minutes): 213
- Mean value of group 2 (films < 90 minutes): 129

[3]Talkerresearch.com (2024). Research reveals the perfect movie length. accessed on 14.05.2024 via https://talkerresearch.com/research-reveals-the-perfect-movie-length/

The t-value is positive and relatively high, so this could indicate a strong difference between the two groups. The very small p-value (well below 0.05) means that the probability of observing such an extreme value of the test statistic if the null hypothesis were true is extremely low. The mean values of each sample show that the average revenue of films with a running time of more than 90 minutes is around 84 millions higher than that of shorter films.

The results of the test support the alternative hypothesis that films with a running time of more than 92 minutes generate significantly higher revenues than films that last 92 minutes or less. The running time is therefore included in the model. As a next step, we create a multiple linear model with interactions with all variables mentioned above.

3.5.3 Completion to a multiple linear model

The following differences are recognizable:

```
# Multiple linear model without interactions and without the genres
lm.fit_inter_ext_1 <- lm(log(revenue_adj) ~ log(budget_adj) * runtime, data = movies)
# summary(lm.fit_inter_ext_1)

# Multiple linear model with interactions and genres-specific
lm.fit_inter_ext_2 <- lm(log(revenue_adj) ~ log(budget_adj) * runtime * First_Genre,
                         data = movies)
# summary(lm.fit_inter_ext_2)
```

- Without genre-specific interactions, 50% of the variance in revenue is explained and there are significant interactions between budget and duration, indicating that these two variables together have an impact on revenue. This model is simple and straightforward to describe.
- With genre-specific interactions, the variance of the model is around 57% and therefore higher than without interactions. There are some significant interactions between budget, running time and genre, which indicates that these variables together have an influence on revenues. But this is depending on the genre and the model is far more complex.

These observations lead to the basic consideration that a more flexible model that can better capture non-linear relationships might be more appropriate. Therefore, the application of a Generalised Additive Model (GAM) appears to be a logical next step to model the relationships between variables more comprehensively and improve the accuracy of results.

3.5.4 GAM - with and without interactions

The generalized additive model can be used to explore complex patterns that may not be fully captured by a linear modeling approach. Therefore, budget, genre, runtime and revenue are reflected in the model. The basic idea is to fit a GAM with a qualitative predictor, in this case Genre. In a first model the approach is to observe the effect of budget, runtime and genre on revenue independently. Therefore a GAM with individual smooth terms for no interactions is builded.

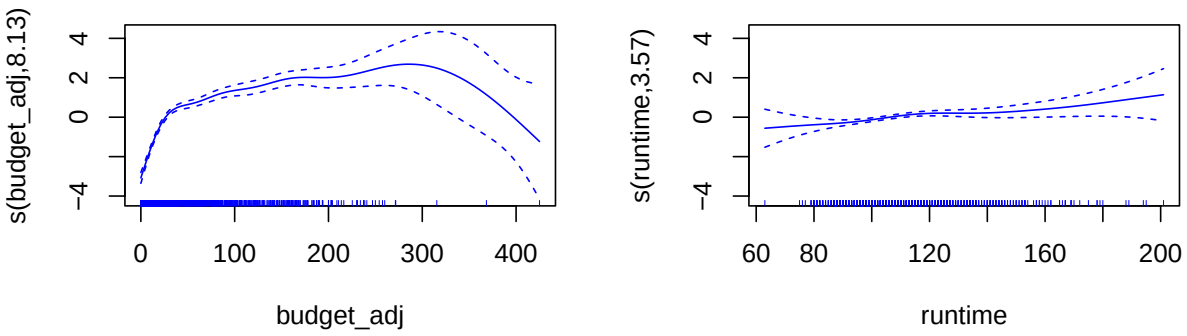


Figure 8: GAM with Individual Smooth Terms (No Interactions)

For comparison, we extend the model and add another variable, the release year, to maximize the model, using interactions between the individual variables.

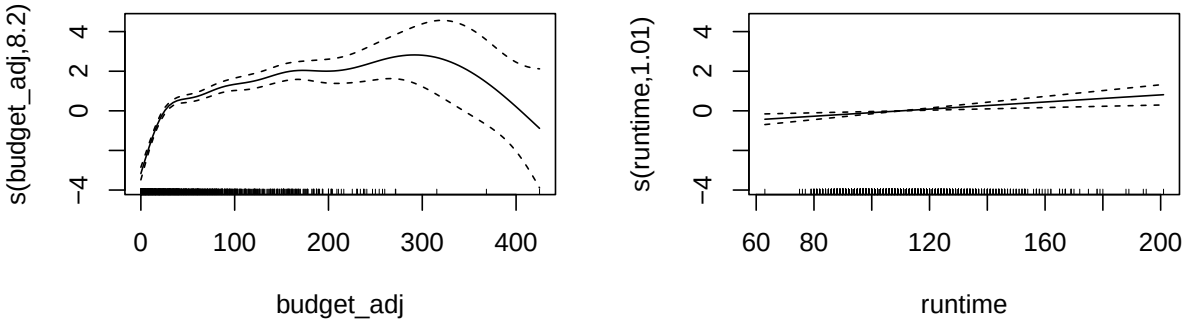


Figure 9: GAM with Interaction Terms

3.5.5 Conclusion

The first approach of using a linear and multiple linear model to examine the dependency of the revenue has not proved to be very practicable. The influence of the factors is probably not linear and a model must be used that can reflect complex and non-linear relationships.The first Generalized Additive Model focuses on the individual effects of the budget and genre on the adjusted revenue. Several genres show significant differences in revenue. The smooth term for budget is highly significant, indicating a non-linear relationship between budget and revenue. Overall the model explains about 50% of the variance in the revenue data. The second model includes interaction terms, allowing it to capture more complex relationships between budget, genre, release year on the adjusted revenue. All predictors show significant effects on revenue and the model explains 55% of the variance, which is a improvement.

The GAM models have shown that they are better able to capture the non-linear and interactive effects between different factors influencing movie revenues. For the client, the results mean that future modeling

efforts should continue to explore nonlinear and interactive effects to improve the accuracy and insight of predictions in the movie industry. By integrating additional variables, the explained variance could be increased to get a more accurate picture of which factors influence the financial success of movies.

3.6 Neural Network

3.6.1 Defintion of the model

A neural network in the context of machine learning is a concept derived from the human brain. The idea is to send information via several of these so-called neurons and thereby gain insights. The parameters for an optimal model are calculated during the training process. Using this method, insights can be extracted from complex and, at first glance, unrelated data.[4]

3.6.2 Hypothesis

The aim of this chapter is to find out whether a neural network can be used to predict the popularity of a film. For this purpose, the budget, the profit and the genres are taken as input. Can we create a model that can reliably predict popularity?

3.6.3 Build the network

To create the network, the first step is to split the data into a test and a training data set. In this case, we have opted for a split of 80% training and 20% test. This will be used later to test the reliability of our predictions. Two hidden layers, each with two account points, were defined as the structure of this model for this first step. This is because it is not too complex, but offers enough possibilities.

3.6.4 Confusion Matrix

Now that the neural network has been created, we can test the model with our test set. We like to use a confusion matrix for this. This matrix makes it very easy to see how often the model was right and how often it was wrong. This is a very good way of determining reliability. If you now look at these results, you can see that the accuracy is only around 50 per cent. This means that only about half of the popularity levels can be predicted correctly. What you can see, however, is that low and high can be predicted better than the middle two levels. Also, the error is usually one level above or below. At least you can estimate the correctness relatively well with the model.

3.6.5 Cross validation

This step is about checking the parameters that we have used for the model and thus improving the model once again. The aim here is to find the optimum parameters and thus make the predictions once again. In this way, the accuracy of a model can be improved. In this graphic you can see the different configurations that were tested and you can see that the original network structure was not so far away from the optimum.

3.6.6 Conclusion

With the model now optimised by cross-validation, another confusion matrix is created. However, the data basis is now the entire data set on which the predictions are tested.

[4]Neural Network

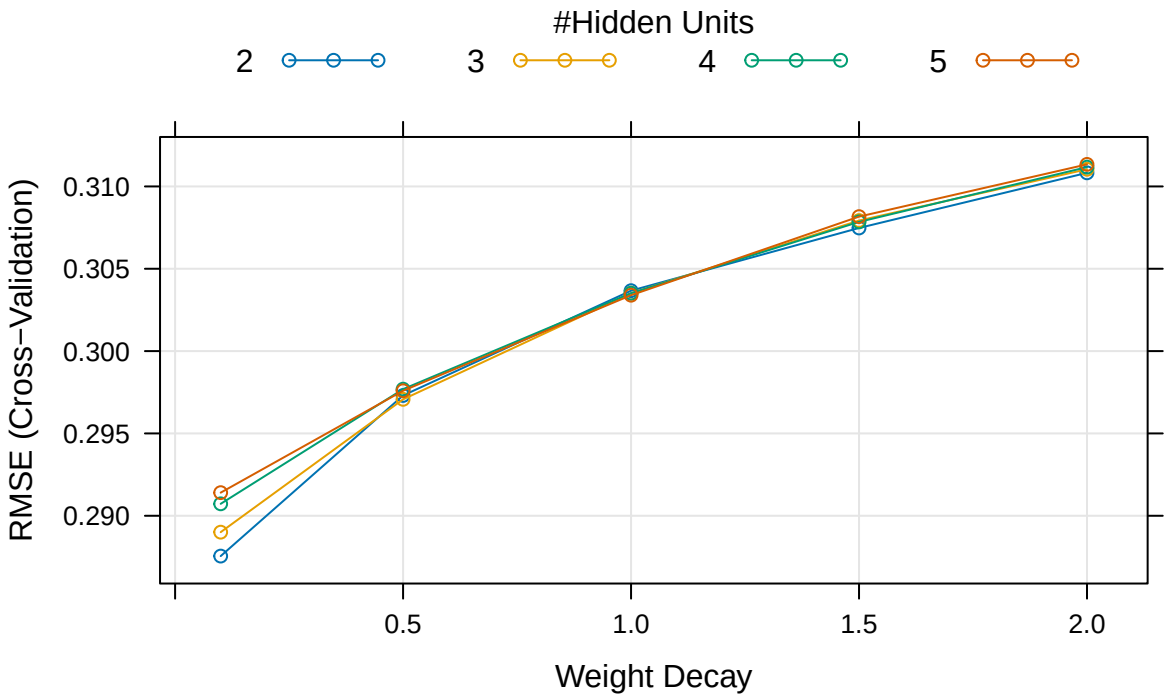


Figure 10: Cross Validation

You can see in the final model that the accuracy is approximately 46.51%, indicating that the model correctly predicted the popularity level about 46.51% of the time. There is a substantial number of misclassifications, especially between adjacent classes (e.g., “Low” and “Medium”, “Medium” and “Moderately High”). This is about the same result, as seen above in the first model. It is possible that there is an imbalance between the different classes and this would also explain the imbalance between the results.

How would you continue here? You could tweak the network structure and parameters again. You could also consider whether the popularity level factors need to be weighted to ensure better results. One should also consider using other libraries as a basis for the calculations, as this might allow a clearer separation of the predictions.

Overall, while the model shows some predictive capability, especially distinguishing the “High” class, it needs improvement to better distinguish between other popularity levels. Further tuning and exploration of data preprocessing techniques should be considered.

3.7 Support Vector Machine

The financial success of a film is of great importance to the client. This chapter therefore examines whether a film can be classified as financially successful based on certain characteristics. A support vector machine model is used to predict whether a film can make a significant profit.

To begin, the publicly available data is used to identify when a film is considered successful and how films that have already been released can be classified as successful.

3.7.1 How to classify if a movie is financial succesful?

Looking at the data set and in particular the profit of the individual films, the following can be seen: (The use of log() was not used, which leads to a highly right-skewed distribution. Reason: Profit can/must also be negative.)

1. **range:** The values range from -413,912 (The warrior's way) to 2544,506 (Avatar) million USD. The mean value is USD 124,241 million and the median is USD 45,243 million.
2. **distribution:** The histogram and skewness value indicate a right-skewed distribution of profit. as indicated by the long tail on the right side of the histogram. Therefore log() is used for further analysis.
3. **break-Even-Point:** The break-even point is reached when the revenue covers the production and marketing costs of a film. Anything above this is considered a profit. If we now assume that the budget values include these cost components, around 1022 of 1287 have reached the break-even.

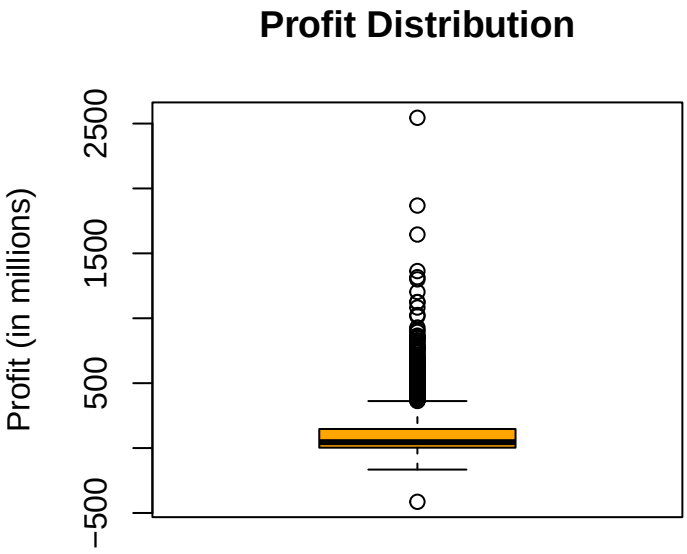


Figure 11: Profit distribution

3.7.2 Determination of a treshold value

In the opinion of the project group, however, it makes little sense to call a film successful (from a financial perspective) once it has reached break-even. For this reason, the project group assumes that a film is declared financially successful when it has made a profit of at least 25 million.

With this assumption, we will now investigate whether it is possible to predict whether a film will be financially successful based on budget and popularity.

Why these two variables? It is important for the client to understand whether a large budget is required to produce a successful film. The popularity level is also used, as a certain level of popularity on the market is a prerequisite for generating sufficient profit. The popularity level [5] describes several measures, that have to been taken into account.

The following plot shows the division of the data by profit into two groups based on the variables budget and popularity with the threshold value of USD 25 million.

Figure 12: Classification

[5] Stanford.edu. (2016).[General Machine Learning] Predicting Movie Popularities Using Their Genomes. accessed on 14.05.2024 via https://cs229.stanford.edu/proj2016/report/NgiawXuNg-PredictingMoviePopularitiesUsingTheirGenomes-report.pdf

3.7.3 Training the Support Vector Machine Model

Based on this, the SVM model was trained using budget, popularity, runtime, vote average and the first genre as features. The plot shows the decision boundary created by the SVM model, classifying areas of financial success and non-success based on budget and popularity.

```
set.seed(123)
svm_model <- svm(financial_success ~ budget_adj + popularity + runtime + vote_average +
                   First_Genre, data = train, kernel = "linear",
                 scale = TRUE, cost = 10, probability = TRUE)

#summary(svm_model)
```

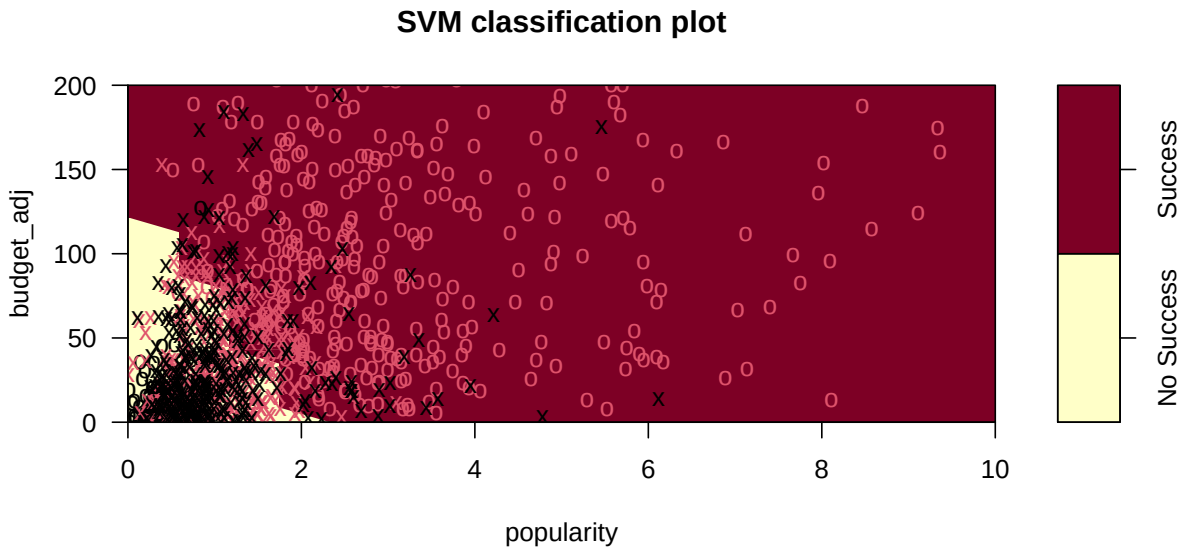


Figure 13: SVM - linear kernel

The plot shows the linear decision boundary created by the SVM model to separate the classes (financially successful vs. not successful) based on the budget_adj and popularity features. The background colors indicate the regions classified as "Successful" (red) and "Not Successful" (yellow). Only the relationship between two variables at a time are ploted (due to the two-dimensional visualizations). In this case, the plot is showing how the decision boundary separates movies based on their budget_adj and popularity values. The other variables (runtime, vote_average, and First_Genre) are still used in the model for training and predictions but they are not visualized in this particular 2D plot. The confusion matrix shows the following results for the defined model:

Table 6: Confusion Matrix

Prediction	Reference	Freq
No Success	No Success	60
Success	No Success	18
No Success	Success	29
Success	Success	86

The accuracy indicates that the model correctly classified around 76% of the movies in the test set. The sensitivity around 76% suggests that the model is moderate at identifying movies that are not financially successful and around 74% of actual successful movies correctly identified by the model.

3.7.4 Model improvment

The next step is to change the model so that non-linear relationships between the features are taken into account and therefore use the kernel radial. The intention is to include complexity in the model and achieve better performance. The cost parameter is also adjusted. A higher cost parameter makes the model more sensitive to misclassification and allows it to create more complex decision boundaries. This can lead to better performance on the training data, but can also increase the risk of overfitting.

```
# switch to radial kernel & higher costs: more complex model, maybe better results
set.seed(123)
svm_model_2 <- svm(financial_success ~ budget_adj + popularity + runtime + vote_average +
                     First_Genre, data = train, kernel = "radial",
                   scale = TRUE, cost = 100, probability = TRUE)
#summary(svm_model_2)
```

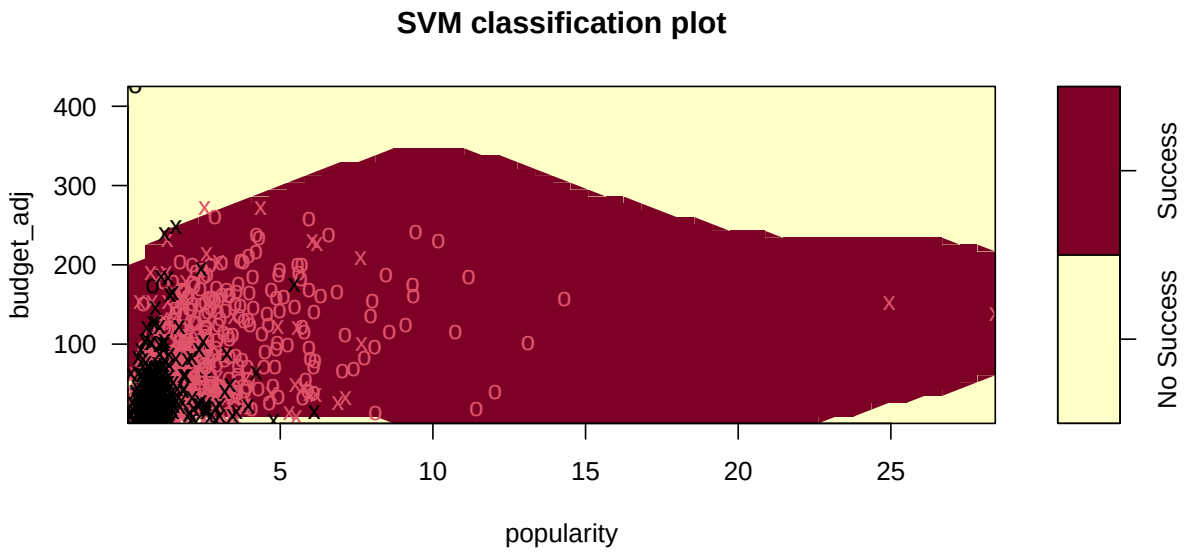


Figure 14: SVM - radial kernel

Table 7: Confusion Matrix

Prediction	Reference	Freq
No Success	No Success	52
Success	No Success	26
No Success	Success	24
Success	Success	91

The radial SVM model with a higher cost parameter achieved a slightly lower accuracy and specificity compared to the linear SVM model. The improvement overall is marginal, suggesting that while the radial kernel captures more complex relationships, the linear model's performance is also quite robust.

3.7.5 Conclusion

Both the linear and radial SVM models show reasonable performance in classifying films based on financial success using specific features (budget, popularity, running time, rating average and genre). The radial SVM model shows a slight improvement in accuracy and specificity (better at identifying successful movies) compared to the linear model, allowing more complex patterns to be captured according to the project group's interpretation. However, the improvement is marginal so it can be said that the linear model also provides robust results with and it is more reliable in general performace across both classes.

Insights for our client are:

- The linear SVM model provides a simpler, reliable and more interpretable solution with good performance.
- The radial SVM model offers a slight improvement for capturing non-linear relationships, but at the cost of increased complexity.
- Both models suggest that budget and popularity are important predictors of a movie's financial success.
- Further improvements to the models could be captured by adding more features.

4 Generative AI Reflection

Generative AI, especially Chat GPT, was a supportive tool to structure some project ideas and to get best practice advice for coding in general. The tool was also helpful in debugging the code by providing explanations for the error messages and suggesting possible solutions. Furthermore, an important benefit of using generative AI was to gain a deeper knowledge and understanding of the different machine learning models. The tools were also helpful in the final review of the programmed R codes in order to optimize and further improve the codes by making them cleaner. This also included specific and consistent formatting of the report and visualizations. The AI-tuned code was tested directly in R to verify that the result met the project team's expectations.

5 Conclusion

By creating a linear model, it was found that there is a significant positive linear relationship between budget and revenue. However, this finding is also heavily dependent on the specific movie genre.

To increase the number of votings, and thus overall popularity, factors such as a higher budget, optimal runtime, specific genres, and the choice of director should be considered. Further, it can be concluded that achieving a substantial financial success (ROI>150%) appears to be genre-dependent, and that there is an optimal budget level, which however may vary by genre. Also, even though longer movie runtime seems to influence financial success positively, there appears to exist an optimal range.

In addition, by applying a generalized additive model, it can be said that non-linear and interactive effects must be given greater consideration in future analyses in order to predict financial success in a meaningful way. If financial success is considered quantitatively with a defined threshold value, a support vector machine is a suitable solution for using the existing variables to classify financially successful and unsuccessful films. In principle, the neural network can be used to make a certain prediction as to whether a film will achieve a certain level of popularity on the market, but here too it is necessary to include further optimization and the use of other libraries in order to be able to use meaningful results.

6 Limitations

The analysis of the movie data is limited by the variables available in the public data set. With a larger data set, which would include additional factors and observations, greater insight into dependencies could be achieved. The analysis showed the client which basic models are suitable for which key indicators of movies and how these would need to be extended in the actual application in the real business activities. Consequently, some of the models presented in this report provide a basic orientation, and can support initial decision-making. However, they may overlook specific details and be overly generalistic.

7 Recommendations & Next Steps

How to estimate the revenue:

- With the budget the revenue can be estimated. The planned film genre can also be used as an additional indicator to predict the success of the film.
- Gather more data, so the models can even be fitted better.

How to increase the number of votings:

- Influential factors appear to be budget, runtime, genre, and director, which should be considered in terms of vote count and further elaborated.
- Explore the impact of additional variables, but also of other approaches that capture the apparent complex relationships more effectively. The main aim should be to gain a better explanation for the highest vote counts.
- Investigate the impact of specific budget and runtime ranges.
- Identify the most influential directors whose movies achieve the highest vote counts.

Which factors lead to a ROI>150% or a threshold value of 25 millions:

- Focus on the genres Adventure, Animation/Family, and Horror, which appear to have higher chances to achieve the financial target.
- Have an adequate budget and rather longer runtime. In this context, in-depth research is recommended to identify optimal budget levels and runtime ranges, both generally and by genre.
- Support vector machines are suitable for classifying films into unsuccessful and successful. The client must determine internally which financial threshold is relevant for its business activities

How to identify the popularity of a film:

- With the help of a neural network the popularity of a planed film can be estimated based on the genre and the budget. It is important to give enough information into the network, so that it can determine the level of popularity. So it would be advised to expand the current neural network model with more data.