Department of
Mathematical
Sciences

UNIVERSITY OF BATH

MA30128: Project

Modelling approaches for predicting film revenues

Michael Chan

Registration number: 149015640

Supervisor: Dr. E. Evangelou

Contents

1. Introduction

The film industry is huge. In North America alone, $11.192 billon was grossed during 2016. (Figure sourced from Forbes). [1] Despite this, cinema is currently under threat. Data from "The Numbers" website indicates that the number of tickets sold has been decreasing since 2002 [2] and the rise of streaming services like Netflix has led to more people watching films at home. Scott Mendelson from Forbes said in a recent article that Netflix has become *"the de-facto consumption method of choice".* [1]

Actor's pay has also often been discussed with Hollywood A-list stars receiving extortionate sums of money. According to Forbes, last year's highest paid actor was former wrestler Dwayne Johnson, who earned $64.5 million. [3] Some of these salaries can take up a significant chunk of a film's production budget.

There are many questions about the box office.

*"Will cinema **Die**!?"*

*"How important is the **Budget**?"*

*"Can the **Studio** make a difference?"*

*"Are **Critics** listened to or just ignored?"*

*"Which **Genre** makes the most money?"*

*"Do **Sequels** always outperform original films?"*

*"And what about audience opinions like **IMDB**?"*

*"Is there a difference between calendar **Seasons**?"*

*"Does the **MPAA Rating** matter? **PG13** or **R** Rated?"*

*"What's the impact of the **Opening Weekend** on further earnings?"*

This project seeks to answer these questions, figure out which features of a film make money and find models to predict the box office. This information is vital for the film industry and its investors because films can be very expensive to produce and carry a lot of financial risks. It will allow studio executives to make more informed decisions on which film projects should get funding.

1.1 Background

In recent years, there have been many publications on predicting box office revenue.

Yahav wrote a paper on "predicting … weekly box-office revenue". He considered various measurements for the demand dynamics of a film. He also used similarity networks which grouped films that had similar attributes such as audience appeal, genre and tone. These similarity networks came from the box office mojo website which for each film, provides a list of movies similar to it. [4]

Edwards et al also made a model for predicting the US box office. They created a deterministic model that used systems of differential equations. They also accounted for geographical effects by "introducing the concept of a region availability function". [5]

Basuroy et al have investigated the effects of critics' reviews on the box office gross. Their conclusion was that they both influenced and predicted revenue. They also found evidence of a negativity bias. This means that "negative reviews hurt performance more than positive reviews help

performance". However, this bias was only present during the film's first week of release. Finally, they analysed the effects of star actors and big production budgets as moderates for the critical reviews. They discovered that "popular stars and big budgets enhance box office revenue for films that receive more negative critical reviews", and they do not have much impact on the revenue for positively received films. [6]

Liu has estimated the box office using information on the film's "word of mouth". He used data from *Yahoo! Movies*. His results suggested that it is the volume of word of mouth that mostly explains the box office. He also found that the most significant variables were: the number of word of mouth messages, the number of screens in the first four weeks and the percentage of positive reviews. [7]

Zhang et al have built models using news data to predict the box office. They compared them to models that only used IMDb data and found that they were similar in predictive performance. However a model combining news data and IMDb yielded better results. [8]

1.2 Report Content

For this project, data will be sourced from the internet and formatted into a data frame to be analysed in R. Missing data shall be handled by using multiple imputation. The first step of the analysis will be to explore the data and get a picture of what it looks like. The report then explains the statistical modelling techniques that were used to produce the models. These models will be assessed on their residual patterns, coefficient uncertainty and prediction error. Cross-validation will be the main tool in comparing the models.

Section 2 contains details on how the data was collected. The two main sources of information were Box Office Mojo and OMDb.
Section 3 provides the descriptive analysis of the data. Its main goal was to identify correlations, important variables and possible transformations.
Section 4 reports on all the modelling used throughout the project. The modelling techniques included: standard regression, lasso regression, regression trees and random forests.
Finally, section 5 provides the project's conclusion, with a decision on the best model, and a summary on the effects of all the variables. It also gives recommendations for future researches.

2. Data Collection

2.1 Box Office Mojo

The first stage of the project was to create a database of around 1000 films released from 2009-16. The Box office mojo website was the primary source of data. This website has extensive information on the US box office and has also been used by previous researchers. Table 1 provides the details of all the variables that were extracted from it. [9]

Variable	Description
Opening Weekend Gross	The film's US box office revenue in the first weekend. Priced in USD dollars.
Number of Theatres at the Opening Weekend **(shorthand: T^o)**	The number of theatres that screened the film during the opening weekend. *In some figures, "NumOpenTheatres" is used as a label.*
Total Gross	The film's total US box office revenue throughout its run. Priced in USD dollars.
Total Number of Theatres **(shorthand: T^t)**	The number of theatres that screened the film throughout its run. *In some figures, "NumTotalTheatres" is used as a label.*
Genre	The film's genre. *This variable required category merging. - see later description*
Year	The year the film was released in.
Season	The season the film was released in: Winter, Spring, Summer, Fall and Holiday. *Definitions for these seasons are given below in Table 2*
Rating	Rating is the Motion Picture Association of America (MPAA) rating system. This consists of G, PG, PG13 and R.
Budget	The production cost of the film. Priced in millions of USD dollars. It does not include the money spent on the film's promotion and marketing. That figure is generally unavailable to the public and hence has not been included for this project. *In the case that the budget of a film was not available from box office mojo, it was sourced from the film's Wikipedia page which is usually refereed.*
Studio	The company that distributed the film. *This variable required category merging. More detail to follow.*
Sequel Indicator	Binary Variable: Seq: The film is a sequel, prequel or a remake. Ori: The film is an original project. The general principle for a film to be classified as "Seq", is that there exists another film prior to it, which is in some way related. This means that a film which was the first instalment of a franchise is not classified as "Seq". *This information comes from the film's box office mojo summary page. Some additional background research was conducted for checking.* *In some figures "Seq.Ori" is used as a label.*

Table 1: List of Variables with descriptions

Genre

There were far too many genres given by box office mojo. So it was decided to merge some of them. This resulted in 11 distinct categories: action, animation, comedy, drama, drama-comedy, family, fantasy, horror, romance, sci-fi and thriller.

The action genre is mostly made of adventure films, westerns and any other genres which centres on a violent conflict. It also contained a lot of superhero films.

The comedy genre was designed to represent films' whose primary focus was to make the audience laugh. It generally included all films with comedy as a sub-genre including action-comedy. But it excluded romantic-comedy as this was more suited for the romance category.

The horror genre consisted of films that aimed to create fear and scare an audience. It contained all films with horror as a sub-genre except horror-comedy which went into the comedy genre. This is because horror-comedies tend to be regarded as more comedic such as "Ghostbusters".

The romance genre was generally made of films that had a romantic relationship at the centre of the film's plot. So it included romantic-comedies and romantic dramas.

Sci-fi contained all films with sci-fi as a subgenre. However sci-fi horror was made an exception as it was assumed to be more suited to a horror inclined audience than a sci-fi one.

The thriller genre was predominantly made up of thriller and crime based films. The main distinction it has with the action genre is that it doesn't need to contain action or a violent conflict. This allowed for a more 'gritty' type of film to be represented here. It also included some sub genres like political thrillers and psychological thrillers.

All other categories generally merged all the films that were labelled with the category as a sub-genre.

There were also a few genres that were excluded due having a low number of observations and no suitable genre to merge with. These included musicals, documentaries and concert films. These are unlikely to affect results much.

Season

Table 2 below provides the definitions in the seasons.

Season	Shorthand	Box Office Mojo Definition	Months
Winter	Win	The first day after New Year's week or weekend through the Thursday before the first Friday in March	January, February
Spring	Spr	The first Friday in March through the Thursday before the first Friday in May	March, April
Summer	Sum	The first Friday in May through Labor Day Weekend	May-August
Fall	Fal	The day after Labor Day Weekend through the Thursday before the first Friday in November	September, October
Holiday	Hol	The first Friday in November through New Year's week or weekend	November, December

Table 2: Definitions of the Seasons, along with the range of months that they generally represent

Studio

Some of the studio categories were also merged. Merging was first done on the basis that: studios were part of the same company, or if one was a subsidiary of another. For example, "Fox

Searchlight" was merged with "Fox". After the merging, all studios with less than 25 films in the dataset (2.5% of the observations) were merged into an "other" studios category. This led to 12 studio categories: Buena Vista(Disney) , 20th Century Fox, LionsGate, Open Road Films, Paramount, Relativity, Screen Gems, Sony, Universal, Warner Bros, Weinstein and others.

Data Extraction

When gathering the data, the aim was to have a reasonably equal distribution of films across time. Films were extracted by season. For each summer the 45 highest grossing opening weekend films were initially selected. For the other seasons, the top 25 highest grossing opening weekend films were chosen. This allowed for an even spread of films throughout the year, as summer is about 4 months long and the other seasons are 2 months long. However some observations were omitted because they were either: unrated, did not have a genre or were judged not to be an appropriate film. Data was extracted from all 5 seasons in years 2009-2016. However holiday season in 2016 was an exception because some of the films were still grossing money while the data was being collected. The data was stored in an excel spreadsheet where all the category merging was conducted. The spreadsheet was then converted to a csv file and read into R. The final dataset consisted of 1002 films.

2.2 OMBb API

A second source called OMBb API also had extensive information on films. It had data on Metacritic scores, IMDb scores, number of IMDb votes, the number of awards and the number of nominations. [10]

Variables

Metacritic is a website that aggregates reviews from professional critics into a score that ranges from 0-100. It requires at least 4 reviews from critics to be published. The website assigns different weights to different critics, although they do not reveal what the weights are. [11]

IMDb is a website with a popular database for information related to films. Its film scores come from the ratings of its 75 million registered users. These scores are on a scale from 1 to 10. The number of IMDb votes is the number of users that have rated the film. [12]

Although the number of IMDb votes, awards and nominations could potentially explain some of the box office gross, it was decided that they would not be included in the model fitting. This is because the data for them is only known after the film has been in the cinema. Also, the number of IMDb votes is expected to increase over time, as it is a reflection of the number of people who have seen the film - thus they would not make suitable predictors. Nonetheless, they were still extracted for data exploration purposes.

Data Extraction

Using the film's title as an input, a function in R was written to extract the data as an XML file, read it in R as text, use the function "gsub" to take the numbers from the text and then convert them into integers. However solely inputting the film's title created some problems. This was because the titles came from the Box office mojo website, which uses the American title. This sometimes has discrepancies with the international title that OMDb API uses. Furthermore, if the same title has been used for multiple films, the Box office mojo title will also contain its release year inside a set of brackets. This is quite standard for remakes. For example, the recent "The Jungle Book" remake is

labelled as “The Jungle Book (2016)”. Films titled in this way would not be recognised by OMDb and hence would not return any data.

In order to address this, the function was edited to also take the release year as a second input. In the excel spreadsheet, the year in brackets was removed from the titles. The file was then reread into R. The function would then first look up the film using only the title (as before). If the film was found, it would then check if the inputted release year matches with the year marked on OMDb’s release date.

The reason OMDb’s release date was used instead of its release year (a separate variable it holds), is because it uses the US release date. The release year refers to the year the film was first available in the world. There are some foreign films which are sometimes released in the US in a different year from the film’s year according to OMDb. For example, “The Lobster” is an internationally co-produced film which was first released in 2015, but was only available in the US in 2016.

So if the years matched then the data is stored. If they didn’t match then the function would search the film again but use the title and release year that have been inputted. If a film was found then it is stored. The reason a year matching checker was used instead of purely searching every film using the title and year, was because it was quite slow to search using both inputs.

2.3 Missing Data

After running the function for all films, the data was inspected for NAs. There were some cases where all the variables were missing which meant the film had to be searched manually. This was usually caused by a discrepancy between the US title and the OMDb title. This was solved by finding the film’s IMDb reference number and using it to search for the film directly in the OMDb website.

There were a few films that had missing information for only some of the variables. Those observations were first checked to see if OMDb returned the correct film. There were a couple of rare cases where it returned a short film of the same name and year or a documentary about the making of the film. The prominent example for this type of error was “Terminator Genysis”. To which OMDb returned “The making of Terminator Genysis”. Again this was solved by using the film’s IMDb reference number. In the case where OMBb did return the correct film, the missing data was searched online and found from other sources.

The only variable with missing information was metacritic score. There were 6 films that didn’t have one, most likely because there were not enough critics that reviewed them. Upon closer inspection of these 6 films they all appear to be:

- Low budget
- Were only available in a small number of theatres
- Low number of IMDb votes
- Low box office gross

However there was one outlier that defied all these trends. The film “Inkheart” had a relatively large budget of $60 million and had a wide release. This is highly unusual. In fact, there appear to be no reviews for it on metacritic. The reason for this is unknown.

Multiple Imputation Method

In response to the missing information with Metacritic scores the aregimpute function from the Hmisc package in R was used to estimate five values for each NA by predictive mean matching. This

seemed appropriate as there appeared to be strong correlation between Metacritic and IMDb scores.

In this project's case where there is only one variable with NAs, the aregimpute function implements the following algorithm:

1. For the variable with NAs, give the NAs an initial value drawn from a random sample.
2. A subsample of the database where the data is not missing for that variable is then drawn.
3. From this subsample a predictive mean matching model is fitted using additive regression with the missing variable as the response. (Using the function "areg"[13]).
4. The model then predicts the values for all the observations.
5. For each NA, a value is selected by finding the observation with the closest prediction value.
6. The NA then takes this prediction value as its estimate.
7. Repeat this for 4 more iterations, each using a different subsample of the data

Please note that the algorithm written above has been simplified for the context in which the data only contains one variable with NAs. A full algorithm that explains the steps taken for multiple variables with missing data as well as alternative matching methods can be found in function's Rdocumentation page. [14]

Table 3 below displays the estimates of the missing Metacritic scores for each data frame. Figure 1 shows the estimates on a scatterplot.

Film	1st	2nd	3rd	4th	5th
Brotherly Love	53	53	55	51	39
Un Gallo con Muchos Huevos	52	52	55	52	58
Home Run	24	35	44	28	28
Preacher's Kid	16	22	22	34	32
N'Secure	11	18	22	21	26
Inkheart	51	46	39	33	39

Table 3: Multiple Imputation results

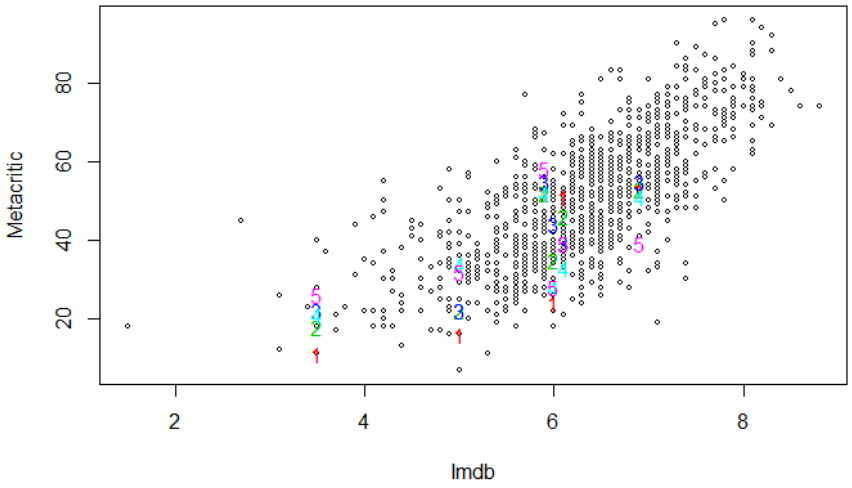

Figure 1: Scatterplot of IMDb vs Metacritic with NA estimates

3. Descriptive Analysis

3.1 Correlations

The first task for the descriptive analysis was to find correlations between the variables. Amongst the continuous variables there were three pairs that showed very high linear correlation. They were: Metacritic score with IMDb score, number of awards with number of nominations and T^o with T^t (Table 1). This is important to note in order to avoid issues with multi-collinearity. If some of the columns of the design matrix are almost collinear, then the variances of the coefficient estimates become very large.

Metacritic and IMDb

The Metacritic and IMDb scores have a correlation coefficient of 0.73. This positive correlation between them could suggest some agreement between professional critics and IMDb users on the quality of a film. Whilst it may imply that critic's and public audience's opinions seem to align (as IMDb is a public service). However, the IMDb user community is probably not a good proxy for the public audience's opinion, because its members are generally people who regard film as a major interest and would watch and rate many more films than the average person.

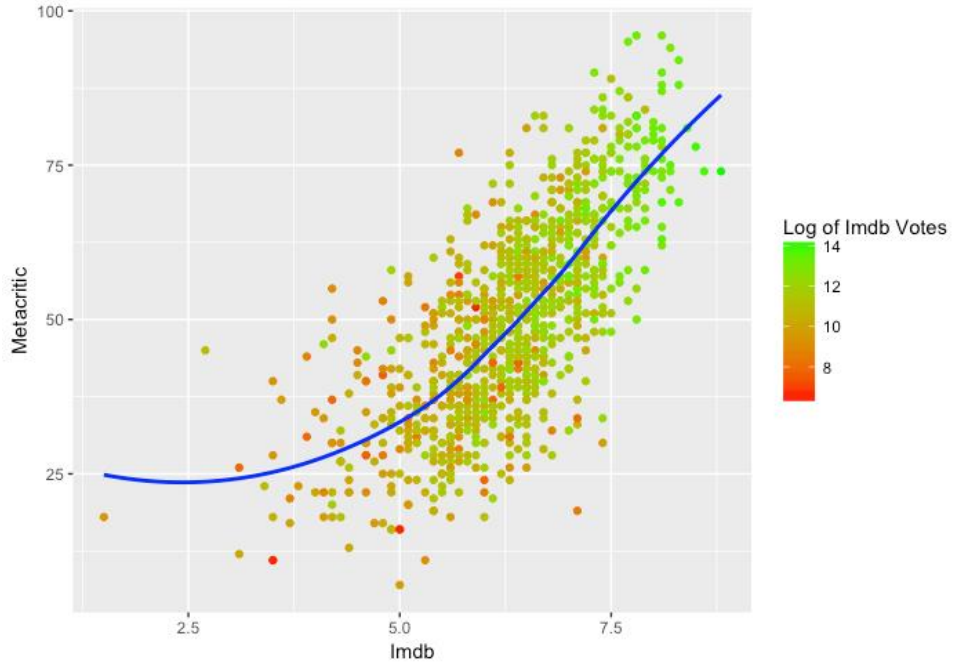

Figure 2: Scatterplot of IMDb vs Metacritic

Figure 2 shows that there seems to be a strong linear relationship between the two scores. There also appears to be some correlation with the number of IMDb votes with the two scores.

Number of theatres

The incredibly high correlation coefficient of 0.93 between T^o and T^t is pretty intuitive. For wide releases T^t tends to equal or only be slightly higher than T^o. The scatter plot clearly shows this with a straight line closely resembling $y = x$. The films which don't follow this trend are likely to be films that start with a limited release followed by a wide release at a much later date. This is common for Oscar-bait films which strategically get a small release just before the end of the year and then get a wide release around February when award season is happening. This theory is shown in figures 3

and 4. The cluster of points that don't follow the linear trend all seem to have a high number of award nominations and are either released in the holiday or fall seasons.

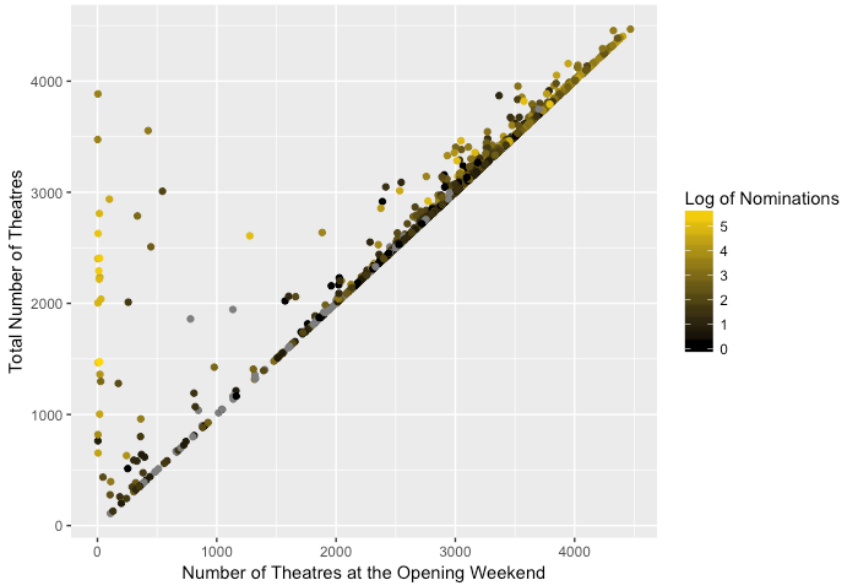

Figure 3: Scatterplot of the Number of Theatres coloured by the log of the number of nominations

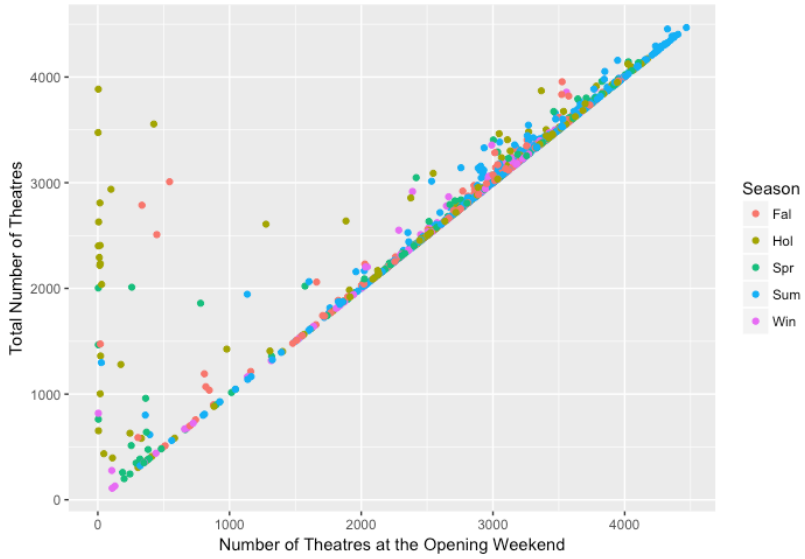

Figure 4: Scatterplot of the Number of Theatres coloured by Season

3.2 Categorical Variables

Studio and genre both seem to have plenty of observations in each category. Genre's smallest category, fantasy, had 31 films. The smallest Studio category was "Open Road Films", which had 26 observations. Both made up at least 2.5% of the data set.

However in the rating's category there is a very low number of G rated films -Only 15 films in the data set. From figure 5, G rated films are mostly either Family or Animation with only one being a drama. Family and animation are also the only two genres that are dominated by PG rated films. This makes the merging of G and PG films a suitable option worth considering.

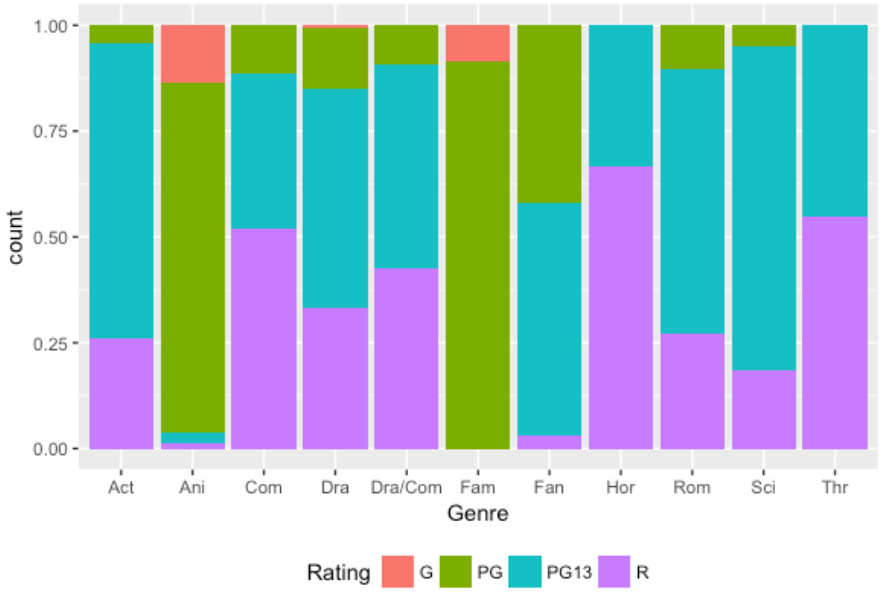

Figure 5: Mosaic Plot of Genres and Ratings

3.3 Response Variables

This project focuses on modelling two response variables: the "Opening Weekend Gross" **(shorthand: G^o)** and the "Total Gross after the Opening Weekend" **(shorthand: G^t).** The "Total Gross after the Opening Weekend" is the "Total Gross" minus the "Opening Weekend Gross". It was also allowed to include G^o as a predictor variable.

This approach was used because some of the variables, particularly the critic scores, are only available just before the release of the film. So the idea was to build a model to predict G^o just before the release of the film. Then G^t can be predicted after the first weekend, or it can be predicted before the release of the film using an estimate for G^o. This is quite practical, as it is common practice for companies to track the opening weekend box office. It's also quite possible that the two responses behave differently. The explanatory variables may differ in their importance.

3.3.1 Opening Weekend Gross Visualisations

The following graphs illustrate the relationships between the response G^o and other variables.

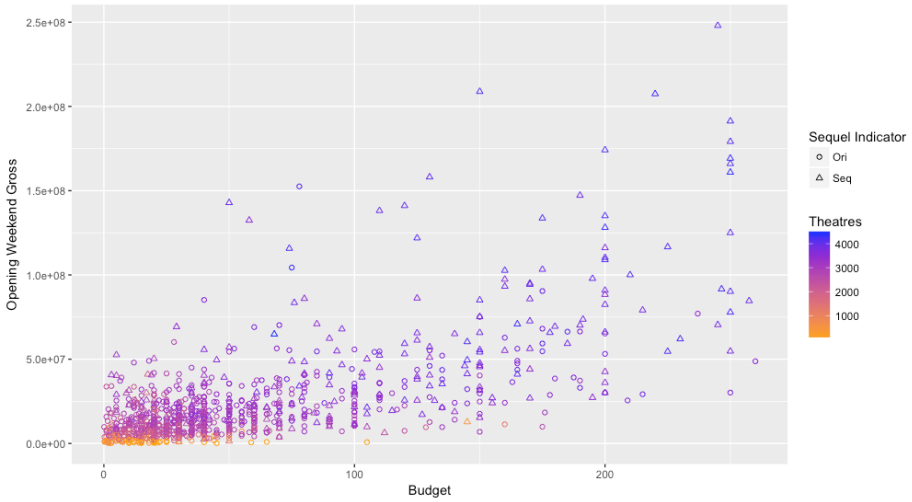

Figure 6: Scatterplot of Budget vs G^o with T^o and Sequel Indicator

Figure 6 shows that budget and $T^O$are potentially important variables for G^O. An increase in either seems to increase the Open Weekend Gross. The graph also shows that sequels gross more money than original films and tend to have larger budgets. There is also a dense cluster of low budget, low grossing, original films, located in the bottom left corner of the plot.

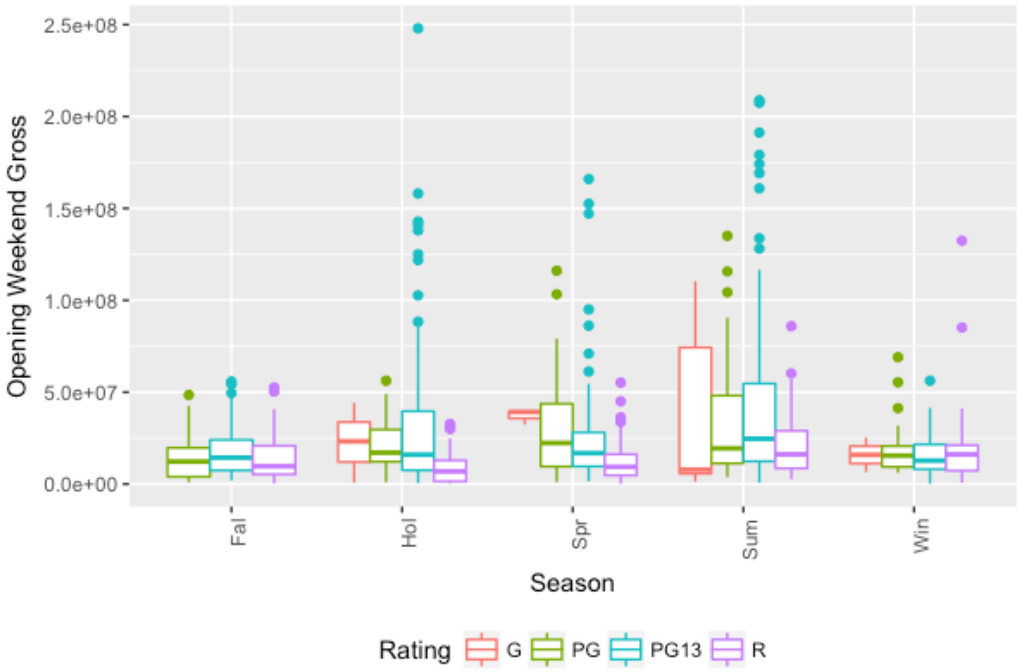

Figure 7: Boxplot of G^O by Seasons coloured by Rating

Figure 7 shows some variation between the different seasons and ratings. There are a lot of large outliers, most of which are PG13 rated. The outliers are also predominantly in the holiday, spring and summer seasons. G rated films show a lot of variability across the seasons. But this may be caused by their low sample size. R rated films look to have the poorest performing rating. However in winter they have the highest grossing rating.

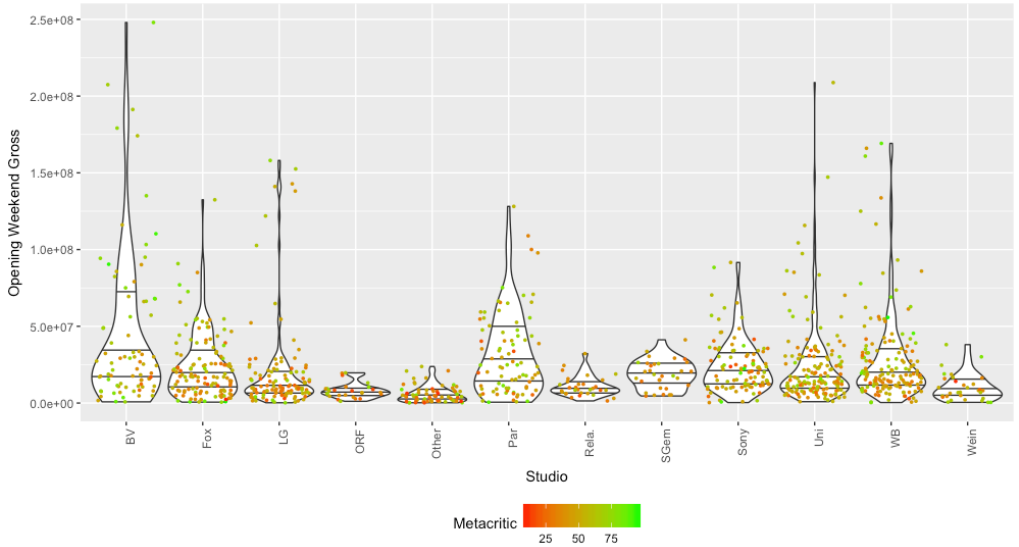

Figure 8: Violin plot of Studios vs G^O coloured by Metacritic Scores, lines represent quartiles

Figure 8 shows the differences in G^o by studio. Within each studio the observations are scattered randomly by "jitter". The violins indicate asymmetrical distributions. They also show many large outliers. The 6 highest performing studios which have very fat and dense upper quartiles are: Buena Vista(Disney), 20th Century Fox, Paramount, Sony, Universal and Warner Bros. These are commonly referred to as the "Big Six" major Hollywood studios. There also appears to be a lot of low grossing and poorly scored films in the "Open Road Films" and "Other" categories.

3.3.2 Total Gross after the Opening Weekend Visualisations

The following graphs illustrate the relationships between the response G^t and other variables.

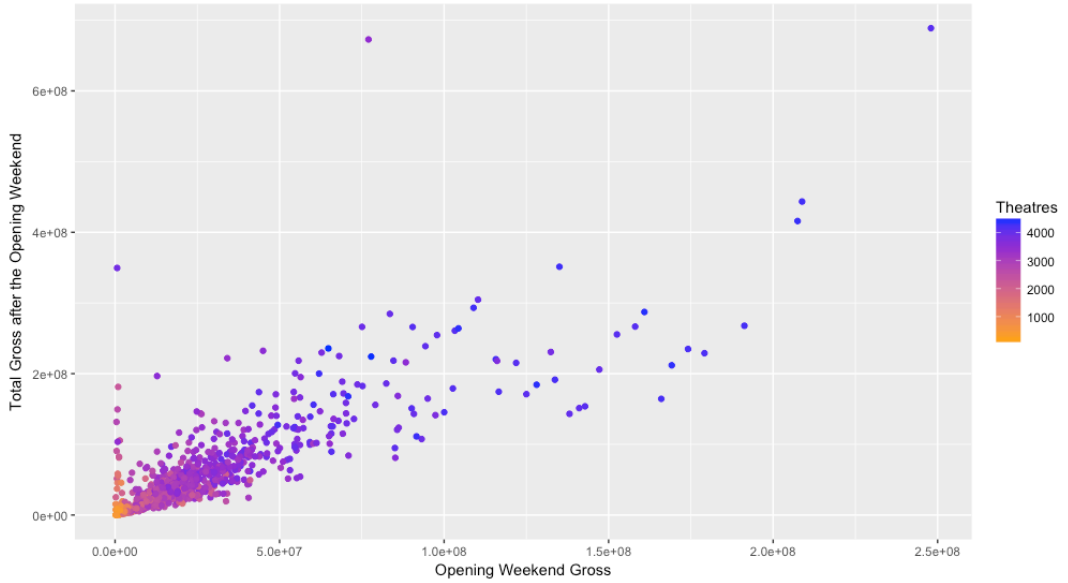

Figure 9: Scatterplot of G^o vs G^t coloured by T^t

Figure 9 shows the importance of $G^o$and T^t. They both have a positive influence on G^t. Like figure 6, there is a dense cluster of observations in the bottom left corner of the graph.

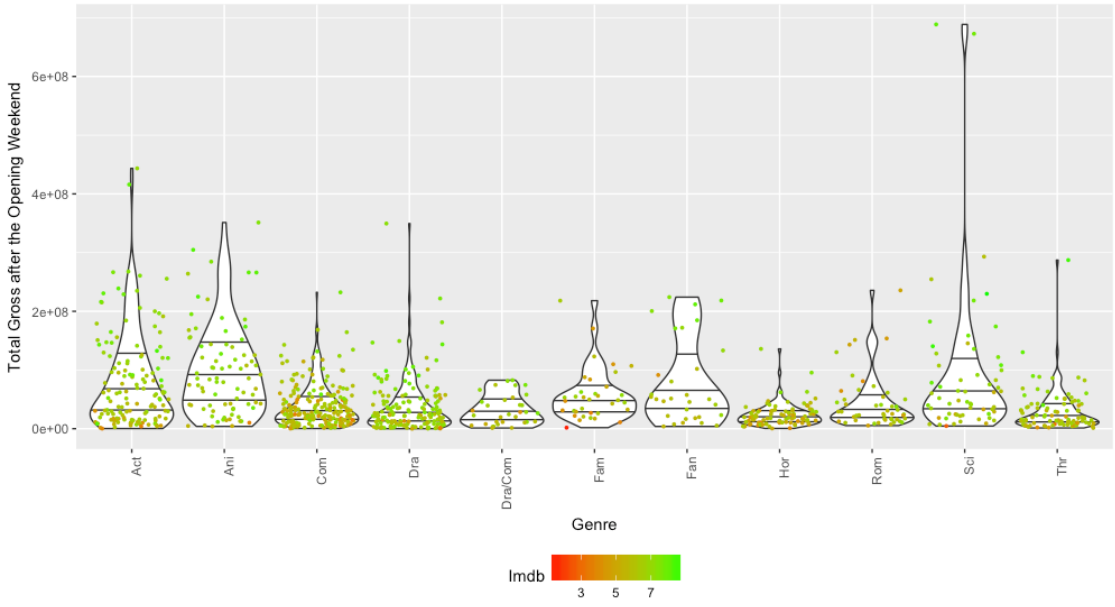

Figure 10: Violin plot of Genre vs G^t coloured by IMDb Scores, lines represent quartiles.

Figure 10 shows the variation among genres. Like figure 8, most of the violins show asymmetric distributions with large outliers. Sci-Fi has two very extreme outliers. Those observations refer to the films “Avatar” and “Star Wars: The Force Awakens”. The graph also shows some relationships with IMDb scores and genre. Dramas and animations look very green, suggesting that they are generally given quite high scores. However, comedies and horror films score much lower.

3.4 Transformations

As seen from the descriptive analysis, both responses seem to have asymmetric distributions. The scatter plots show signs of exponential trends. Therefore, a log transformation might be appropriate. The histogram of the G^o(figure 11) showed a right skew. After using a log transformation, the data looked more symmetrical as seen in figure 12.

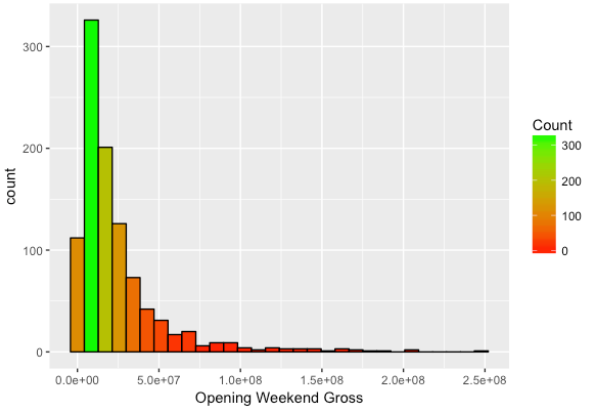

Figure 11: Histogram of G^o

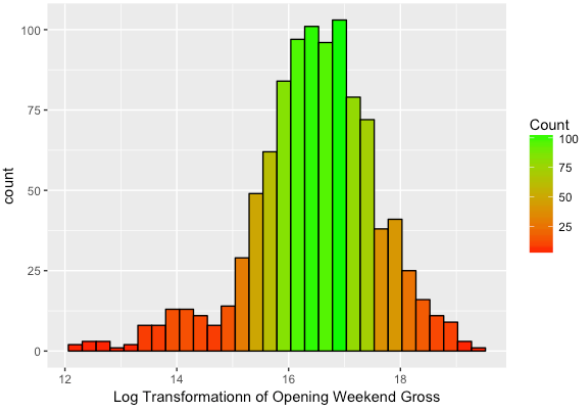

Figure 12: Histogram of the Log of G^o

G^t also seems to look more symmetrical after a log transformation as demonstrated in figures 13 and 14.

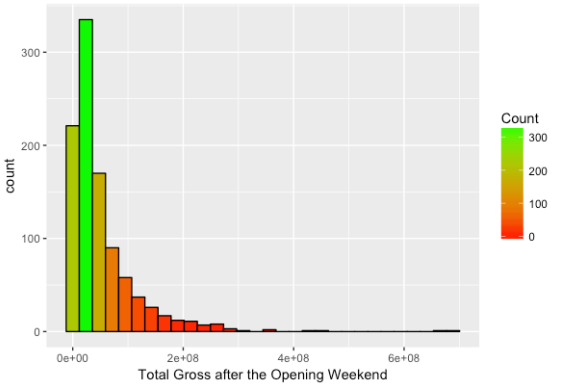

Figure 13: Histogram of G^t

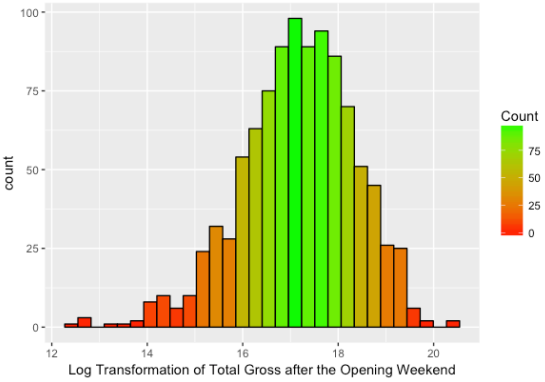

Figure 14: Histogram of the Log of G^t

Based on these graphs, the log transformation of the responses was of particular interest in regards to the modelling.

Other Transformations

Logit transformations were assumed for Metacritic and IMDb. This was so that they matched to the real numbers. The transformations are given below:

$$MetaT = logit\left(\frac{Metacritic}{100}\right), \quad ImdbT = logit(\frac{Imdb}{10})$$

4. Modelling

This project looked at four different approaches to modelling: GLMs, Lasso Regression, Regression Trees and Random Forests. All the models used the log transformations of G^o and G^t as the responses. This was because models using the untransformed response were generally less favourable and were more exposed to extreme outliers. In fact the highest grossing film of all time "Avatar" often caused problems with a large cook's distance.

Model selection was generally done using cross-validation. The GLMs used leave-one-out cross validation. While the other techniques used a 10 folds cross-validation method. The final models from each technique were then compared using 10 folds. Cross-validation errors for the Regression Trees and Random Forests were calculated using the "train" function in the "caret" package. [15]

The main models will also include a sensitivity analysis on the multiple imputation method.

4.1 GLMs

For GLMs, both Gaussian and Gamma models were considered due to the responses being both scaled and positive. An identity link was generally used for all the models as the response had been pre-transformed. A Gamma GLM with a log link on the untransformed response was considered, but was found to give poorer results.

All suitable predictor variables were used in the initial fittings. However in order to avoid issues with collinearity, the models were designed to include either T^o or T^t and either the Metacritic or IMDb score. Additional analysis was also conducted for modelling the gross per theatre.

4.1.1 Opening Weekend Gross Models

When fitting a model for opening weekend it was assumed that T^o was a more suitable predictor to include than T^t. There were two models initially considered. One contained the IMDb score. The other contained the Metacritic score. Both were Gaussian. After they were fitted, the "step" function in R was used to remove variables that reduced the models' AIC. This resulted in genre being removed from both models. The two models were then compared by their AIC. The model with Metacritic score had the lowest AIC. This suggested that Metacritic was a better predictor than IMDb. Hence, the IMDb model was rejected. On closer inspection of the model with the Metacritic score, it appeared that the rating parameters were insignificant, as they all had high p-values (Table 4).

Coefficient	P-value
PG Rating	0.49
PG13 Rating	0.26
R Rating	0.49

Table 4: P-values of the Rating Parameters

Based on this, it was decided to consider a model that excluded the rating variable and compare it using an F-test. The residual deviances for the full model and the model without rating were 257.77 and 265.2 respectively. The dispersion of the full model was estimated to equal 0.2635648; and there were 978 degrees of freedom. The test statistic was significant and so the reduced model was rejected in favour of the full model. The calculations of the F-test are provided below.

$$\frac{265.2-257.77}{3\times 0.2635628} = 9.397 \sim F_{3,978} \qquad F_{3,978;\,0.95} = 2.614$$

Two more models were fitted with the exact same initial formulas. The only difference was they were Gamma GLMs. The same approach to the analysis was carried out. This led to similar results

with the Gaussian models. As per the earlier models, the genre was removed by AIC and the Metacritic model was preferred over the IMDb one.

There were not many noticeable differences in the residual plots between the Gaussian model and the Gamma model. However the leave-one-out cross-validation prediction error of the Gaussian was slightly lower than the Gamma. On that basis, it was decided to reject the Gamma model and focus more on the Gaussian model. Plus the histogram of the residuals seem to look more normal and didn't show much of a right-skew.

Model Interpretation

So the final model was as follows:

$$\log(G^{o}{}_{ijkmp}) \sim Gaussian(\mu_{ijkmp}, \sigma^2)$$

$$\eta_{ijkmp} = \mu_{ijkmpq}$$

$$\eta_{ijkmp} = \beta_0 + \beta_1 Year_i + \beta_2 T_i^o + \beta_3 MetaT_i +$$

$$\beta_2 Budget_i + \alpha_j^{Studio} + \alpha_k^{Season} + \alpha_m^{Rating} + \alpha_p^{Sequel}$$

$$i = 1, \dots, n, \quad j\epsilon\{Studio\}, k\epsilon\{Season\}, m\epsilon\{Rating\}, p\epsilon\{Seq, Ori\},$$

$$\alpha_{BV}^{Studio} = \alpha_{Fall}^{Season} = \alpha_{G}^{Rating} = \alpha_{ori}^{Sequel} = 0$$

With dispersion estimated at 0.2636

Variable	Coefficient Range		Percentage Range	Standard Errors
	Minimum	Maximum		
Intercept	6.655×10^{1}	6.684×10^{1}	0.447%	1.477×10^{1}
Studio-Fox	-1.349×10^{-1}	-1.343×10^{-1}	0.449%	7.714×10^{-2}
Studio-LG	-3.627×10^{-2}	-3.515×10^{-2}	**3.136%**	8.364×10^{-2}
Studio-ORF	-3.011×10^{-1}	-2.997×10^{-1}	0.475%	1.251×10^{-1}
Studio-Other	-2.053×10^{-1}	-2.000×10^{-1}	**2.594%**	9.670×10^{-2}
Studio-Par	-3.671×10^{-2}	-3.612×10^{-2}	1.638%	8.492×10^{-2}
Studio-Rela	-2.049×10^{-1}	-2.039×10^{-1}	0.519%	1.164×10^{-1}
Studio-SGem	3.183×10^{-1}	3.201×10^{-1}	0.548%	1.162×10^{-1}
Studio-Sony	2.876×10^{-2}	2.951×10^{-2}	2.565%	8.145×10^{-2}
Studio-Uni	6.245×10^{-2}	6.326×10^{-2}	1.289%	7.864×10^{-2}
Studio-WB	-1.109×10^{-1}	-1.091×10^{-1}	1.212%	7.776×10^{-2}
Studio-Wein	-2.890×10^{-1}	-2.881×10^{-1}	0.285%	1.158×10^{-1}
T^o	9.204×10^{-4}	9.215×10^{-4}	0.127%	2.386×10^{-5}
Year	-2.646×10^{-2}	-2.631×10^{-2}	0.561%	7.348×10^{-3}
Season-Hol	5.374×10^{-2}	5.537×10^{-2}	**2.982%**	5.983×10^{-2}
Season-Spr	1.626×10^{-1}	1.648×10^{-1}	1.367%	5.555×10^{-2}
Season-Sum	1.387×10^{-1}	1.398×10^{-1}	0.760%	4.978×10^{-2}
Season-Win	1.981×10^{-1}	1.996×10^{-1}	0.784%	5.665×10^{-2}
Budget	1.620×10^{-3}	1.632×10^{-3}	0.713%	4.267×10^{-4}
Rating-PG	-9.749×10^{-2}	-9.679×10^{-2}	0.717%	1.417×10^{-1}
Rating-PG13	1.578×10^{-1}	1.596×10^{-1}	1.108%	1.415×10^{-1}
Rating-R	9.770×10^{-2}	9.876×10^{-2}	1.084%	1.429×10^{-1}
Sequel	2.001×10^{-1}	2.005×10^{-1}	0.226%	4.312×10^{-2}
MetaT	2.015×10^{-1}	2.038×10^{-1}	1.110%	2.384×10^{-2}

Table 5: Coefficient table for the model. Colours explained in Table 6

Colour	Code
Green	Positive Coefficients
Red	Negative Coefficients
Purple	The 3 largest percentage ranges
Blue	Significant Parameters (P-value<0.05)

Table 6: Colour Code for Coefficient tables

Sensitivity Analysis

The coefficient range in Table 6 displays the maximum and minimum coefficient estimates across the 5 data frames. The percentage range is a measure of the sensitivity of the coefficient. It was calculated by dividing the range of the coefficient estimates by its midpoint. Formulation is given below:

$$Percentage\ Range = \frac{coef_{max} - coef_{min}}{0.5 \times (coef_{max} + coef_{min})} \times 100$$

All the percentage ranges were fairly small. The largest was the studio parameter for "Open Road Films" at 3.136%. So they don't appear to be too sensitive to the Metacritic NAs.

Coefficient Remarks

- The three most important variables are: Budget, Metacritic score and T^o.
- Sequels gross 22% more than original films.
- PG13 is the highest performing rating.
- Winter is the highest performing season.
- Screen Gems is the highest performing Studio.
- An increase of $10m on budget, increase the gross by 1.6% .
- According to the year coefficient, The box office falls 2.6% annually.

Uncertainty

Figure 15 shows the uncertainty of all the model parameters. Although the intercept was not included (due to scaling) it was found to be very significant. All of the season parameters were significant except holiday. Most of the studio parameters were insignificant. It should be noted that Budget and T^o may look insignificant from the plot as they appear to be touching the red zero line. However this is deceptive as it was principally caused by scaling. Confirmation of their significance is clearly indicated by the small size of their P-value markers, with their standard errors noted in Table 6.

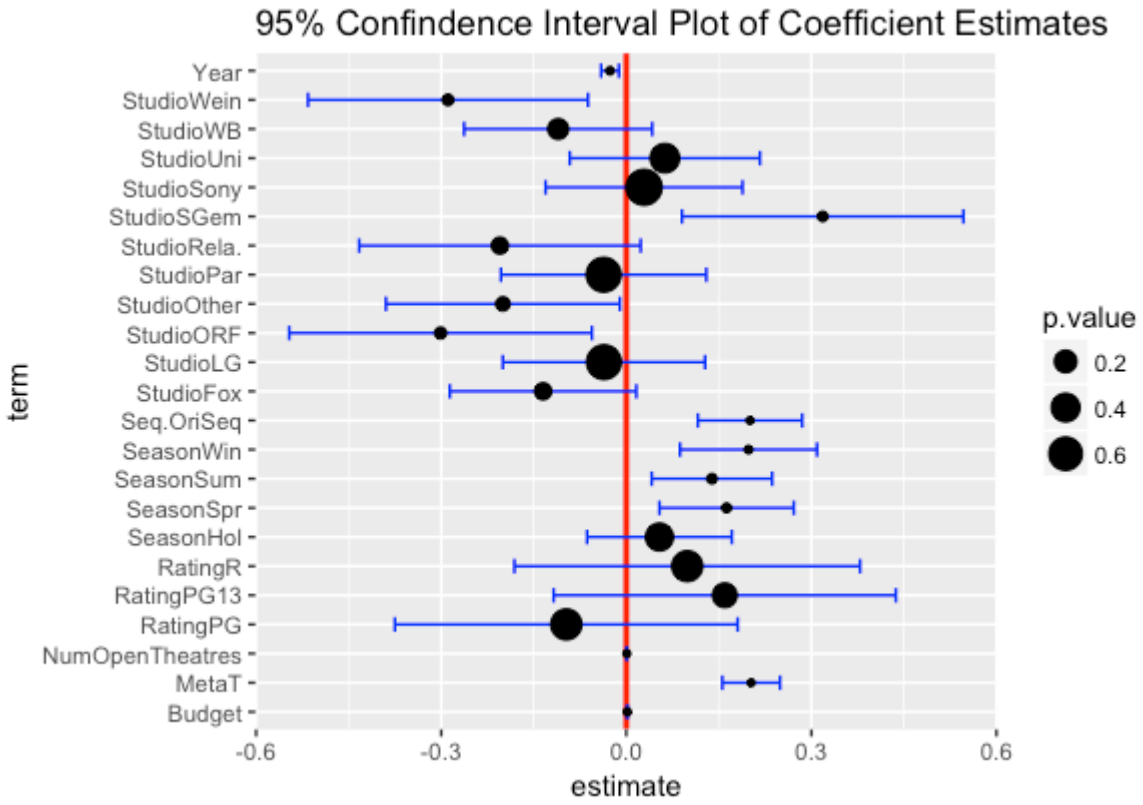

Figure 15: Confidence interval plot of the Coefficients

Residuals

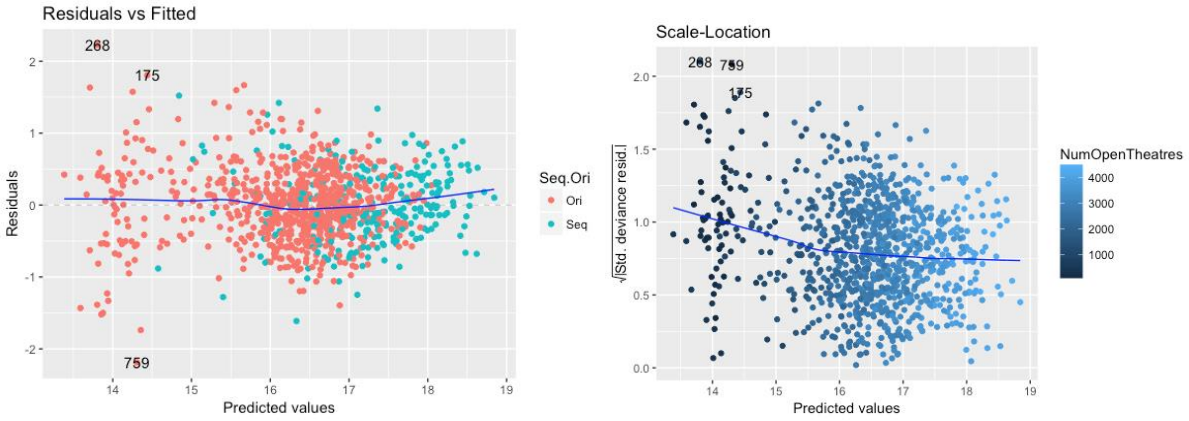

Figure 16: Residuals vs Fitted, coloured by Sequel Indicator

Figure 17: Scale-Location plot, coloured by T^o

Looking at figure 16, the fit looks ok. However there is a slight positive trend for large fitted values. The variance also seems to be larger in small fitted values. This is supported by figure 17, which shows a negative trend for small fitted values. This could be potentially caused by T^o. Perhaps films with limited releases behave very differently from wide releases. Figure 16 also demonstrates the effect that the Sequel Indicator has on the prediction value.

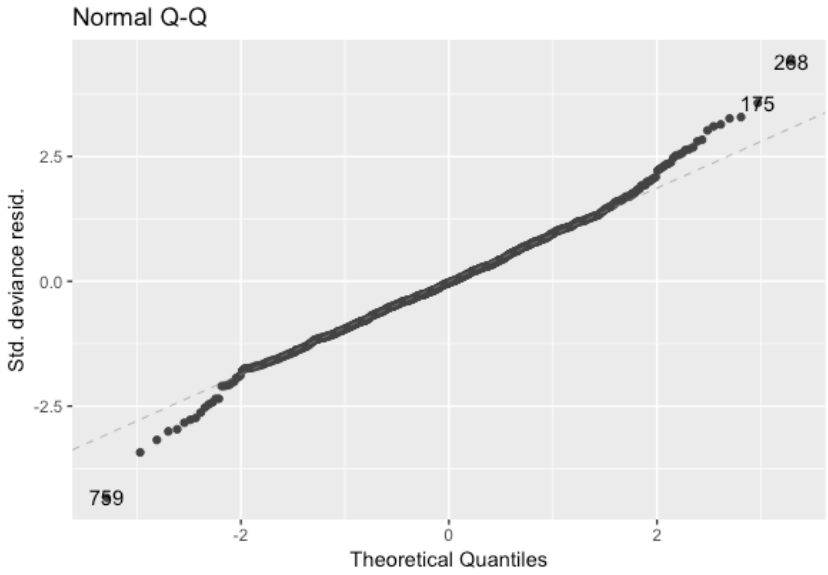

Figure 18: Normal QQplot

Figure 18 appears to be light-tailed, suggesting that the residuals follow a t-distribution rather than a normal one. So there may be some over-dispersion. Overall the model is ok, but is more suited to predicting films with wide releases rather than limited ones.

Per Theatre Models

The same analysis for all four models was repeated. However instead of having T^o as a variable, its log transformation $\log(T^o)$ was used as an offset term. This meant that they were effectively modelling the log transformations of the "Opening Weekend Gross per theatre". The following equations demonstrate this.

$$\log(G^o) = \log(T^o) + \beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}$$

$$\log(G^o) - \log(T^o) = \beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}$$

$$\log\left(\frac{G^o}{T^o}\right) = \beta_0 + \boldsymbol{X}^T\boldsymbol{\beta} = \log(\text{Opening Weekend Gross per theatre})$$

$$\boldsymbol{X} = design\ matrix,\ \beta_0 = intercept, \boldsymbol{\beta} = coefficient\ vector$$

Like the previous models, "per theatre" was better modelled as a Gaussian GLM. Metacritic was also preferred instead of IMDb. The main difference between them was that the year variable was removed alongside genre by the AIC. However, the previous G^o model is far superior as it had a lower leave-one-out cross validation error. Also figure 19 showed a clear negative trend for small fitted values.

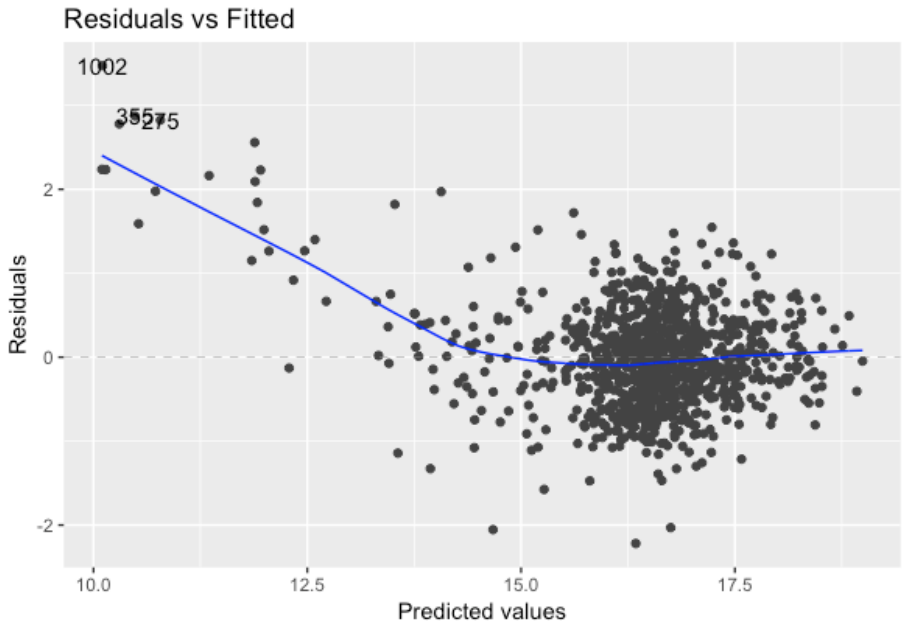

Figure 19: Residuals vs Fitted for G^o per theatre Model

4.1.2 Total Gross after the Opening Weekend Models

The modelling for G^t was done with a similar approach to G^o. The main differences were that G^o was included as a predictor variable and T^t was used instead of T^o. There was one observation that had a $G^t = 0$. In other words, it only grossed revenue during its opening weekend. This film "The Informers" was therefore excluded as its log transformation would be $-\infty$. This obviously could not be used in a regression model. So four models were fitted. They represented the four possible combinations of including either the IMDb or Metacritic score and choosing between a Gaussian and Gamma GLM.

For both the Gaussian and Gamma models, the "step" function removed the rating variable. In the IMDb model, the sequel indicator was also removed. But unlike the G^o models, IMDb was preferred over Metacritic, as it produced lower AICs. Again there weren't many differences between the Gaussian and Gamma fits. The Gaussian GLM was chosen because the errors looked more normal than Gamma and it had a lower leave-one-out cross-validation error.

Residuals

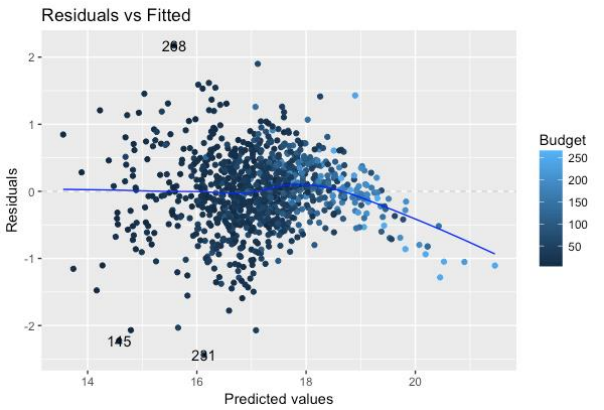

Figure 20: Residuals vs Fitted, coloured by Budget

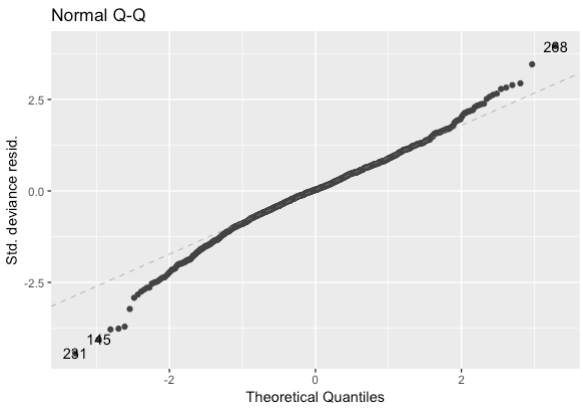

Figure 21: Normal QQplot

The main issue with all the models were the residual plots, particularly with figure 20. All the models showed a negative trend in the residuals for large prediction values. Although this trend does only concern a small number of observations. In any case, the models were likely to over-predict the box office for high grossing films. This suggested that a non-linear model might be better for G^o. As with the previous QQplots, figure 21 showed some over-dispersion.

In an attempt to improve the model, a variety of transformations were tried on some of the variables, particularly G^o and T^t. Amongst them were: log, squared and square root transformations. However none of them made much improvement. Some did reduce the magnitude of the negative residual trend, but resulted in creating an additional negative trend in the small prediction values. In many cases, they also increased the tails of the QQplot.

A further investigation was conducted on the residual trend for large prediction values. By examining the 20 observations that were predicted greater than 19.5, some patterns were discovered. They all had big budgets, with an average budget of $193.3 million. Over half of the films were from the action genre, most of which were superhero films. Nearly all of them were sequels from big franchises. The only two that were original films were "The Hunger Games" and "Inside Out". "Hunger Games" was the first film of its franchise. Whilst "Inside Out" is a Pixar Studies film and Pixar is a widely well-known and successful filmmaker.

The most important distinction these observations have (compared with the rest of the data), is their G^o is it's one of the major positive coefficient variables. The density plot (Figure 22) shows that it is much higher for this group. It's likely that the effect of G^o is non-linear and levels out when large.

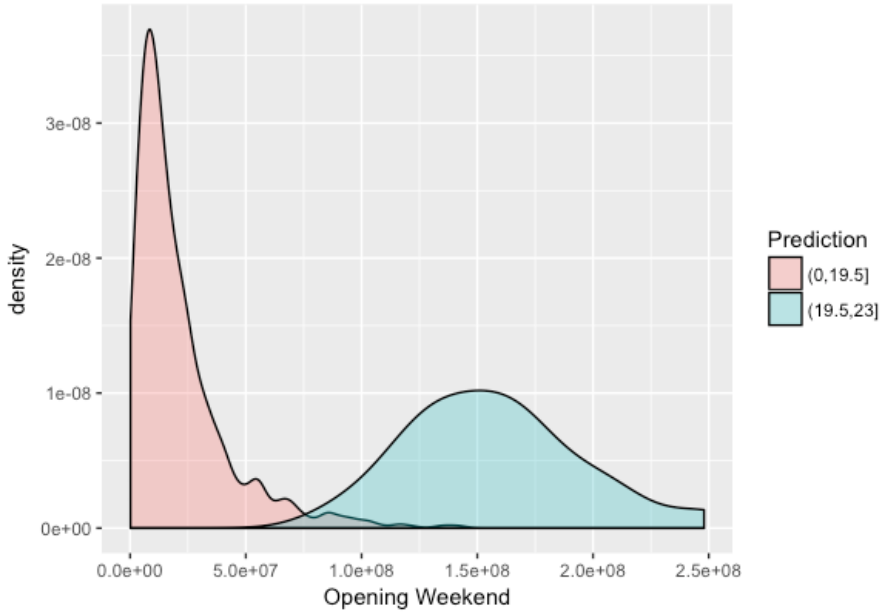

Figure 22: Density plot of G^o split by prediction=19.5

Model Interpretation

So the final model for G^t was:

$$\log({G^t}_{ijkmp}) \sim Gaussian(\mu_{ijkmp}, \sigma^2)$$

$$\eta_{ijkmp} = \mu_{ijkmpq}$$

$$\eta_{ijkmp} = \beta_0 + \beta_1 year_i + \beta_2 T_i^t + \beta_3 ImdbT_i + \beta_4 budget_i + \beta_5 G_i^o + \alpha_j^{studio} + \alpha_k^{season} + \alpha_m^{genre}$$

$$i = 1, \dots, n, \quad j\epsilon\{Sudio\}, k\epsilon\{Season\}, m\epsilon\{Genre\},$$

$$\alpha_{BV}^{studio} = \alpha_{Fall}^{season} = \alpha_{Act}^{Genre} = 0$$

Dispersion was estimated at 0.3093

Variable	Estimate	Standard Errors
Intercept	8.108×10^{1}	1.618×10^{-1}
Studio-Fox	4.320×10^{-3}	8.330×10^{-2}
Studio-LG	-6.233×10^{-2}	8.811×10^{-2}
Studio-ORF	-4.483×10^{-1}	1.331×10^{-1}
Studio-Other	-2.355×10^{-1}	1.039×10^{-1}
Studio-Par	5.631×10^{-2}	8.992×10^{-2}
Studio-Rela	-2.467×10^{-1}	1.256×10^{-1}
Studio-SGem	1.773×10^{-1}	1.241×10^{-1}
Studio-Sony	1.893×10^{-1}	8.771×10^{-2}
Studio-Uni	-4.173×10^{-2}	8.252×10^{-2}
Studio-WB	-1.274×10^{-1}	8.114×10^{-2}
Studio-Wein	-1.703×10^{-1}	1.237×10^{-1}
G^o	8.579×10^{-4}	3.252×10^{-5}
T^t	1.215×10^{-8}	9.203×10^{-10}
Year	-3.346×10^{-2}	8.042×10^{-3}
Season-Hol	4.987×10^{-1}	6.510×10^{-2}
Season-Spr	1.681×10^{-1}	6.084×10^{-2}
Season-Sum	2.021×10^{-1}	5.468×10^{-2}
Season-Win	1.111×10^{-1}	6.163×10^{-2}
Budget	-1.505×10^{-3}	5.468×10^{-4}
Genre-Ani	1.860×10^{-1}	7.980×10^{-2}
Genre-Com	2.810×10^{-1}	6.770×10^{-2}
Genre-Dra	2.817×10^{-1}	7.342×10^{-2}
Genre-Dra/Com	4.823×10^{-1}	1.150×10^{-1}
Genre-Fam	3.860×10^{-1}	1.122×10^{-1}
Genre-Fan	1.367×10^{-1}	1.127×10^{-1}
Genre-Hor	1.041×10^{-1}	8.343×10^{-2}
Genre-Rom	1.764×10^{-1}	9.886×10^{-2}
Genre-Sci	1.463×10^{-1}	8.741×10^{-2}
Genre-Thr	1.276×10^{-1}	7.946×10^{-2}
IMDbT	6.642×10^{-1}	4.839×10^{-2}

Table 7: Coefficient table for the model. Colours explained in Table 6

A sensitivity analysis was not needed for this model because Metacritic was not included.

Coefficient Remarks

- The most important variables were G^o, T^t and budget
- Drama/Comedy is the highest performing genre
- Holiday is the highest performing season
- Sony is the highest performing studio
- An increase of £10m at the opening weekend increases the gross by 13%
- According to the year coefficient, The box office falls 3.3% annually

Uncertainty

Like figure 15, figure 23 presents the confidence interval plot of the model parameters. It shows that all the season parameters were significant. Some of the studio and genre parameters were found to be insignificant.

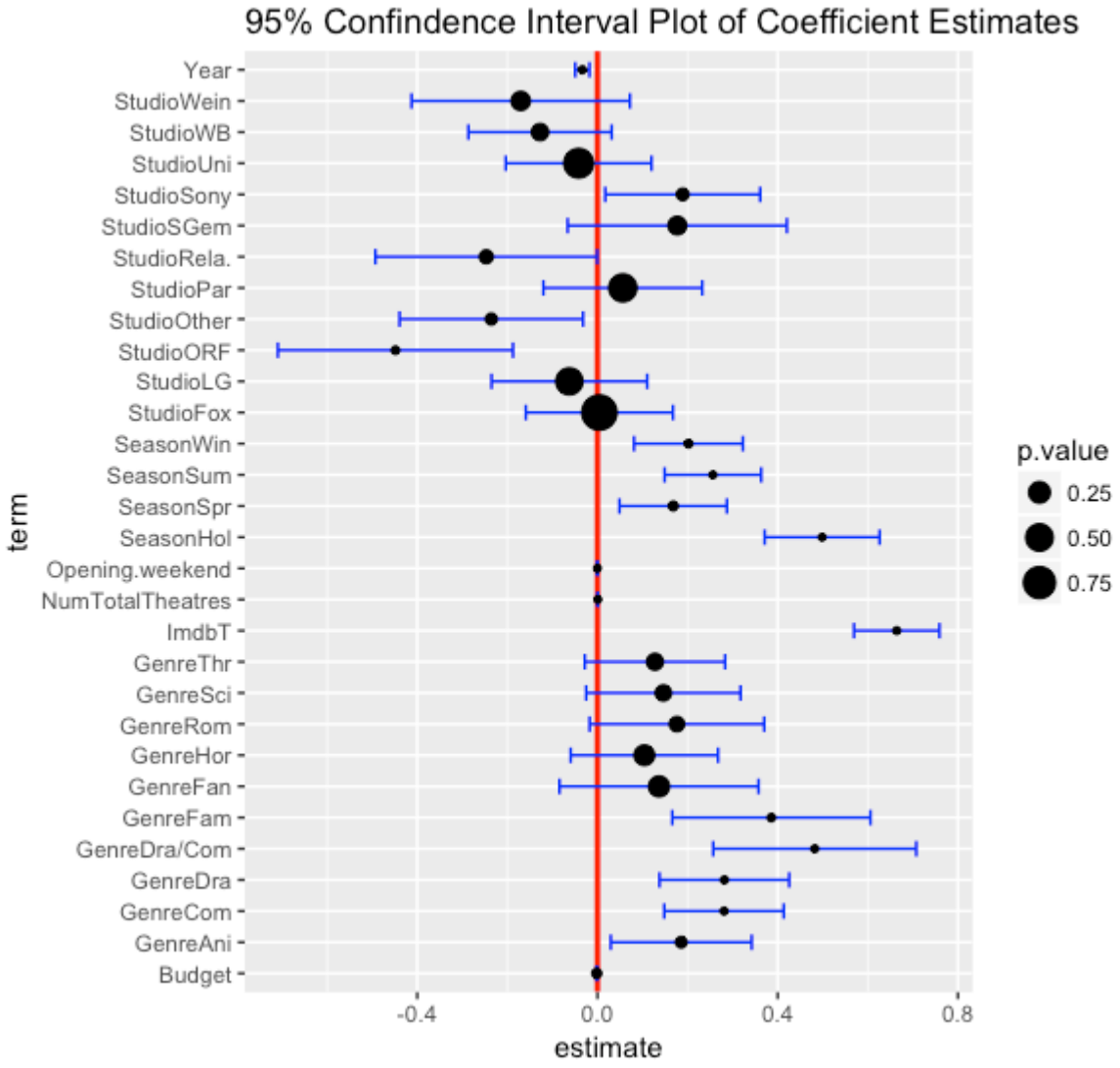

Figure 23: Confidence Interval plot of the coefficients

Per Theatre Models

As was done for G^o, four "per theatre" models were fitted but with $\log(T^t)$ as an offset term. Again IMDb was preferred against Metacritic. However there were some differences in the variable

selection. For both Gaussian and Gamma GLMs, the sequel indicator and rating variable were removed. In the IMDb models, budget was also removed.

They all suffered the same problem as the G^t models. The residual plot (Figure 24) still showed that negative trend. However for the Gaussian IMDb model, the QQplot (Figure 25) was quite different. It looked good for the positive residuals with only a few obeservations not following the line. But it had a much heavier tail for the negative residuals. Nevertheless it did have a similer leave-one-out cross validation error to the final model for G^t.

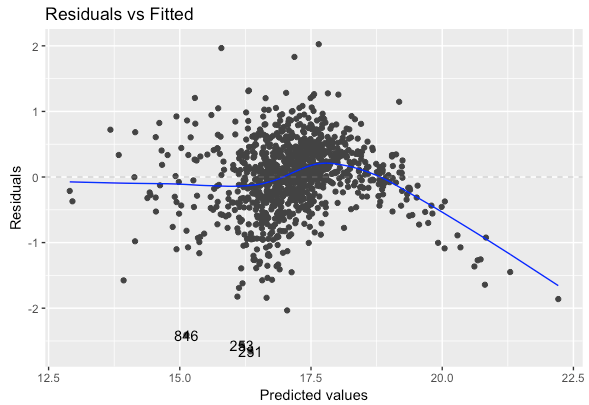

Figure 24: Residuals vs Fitted for G^t per theatre Model

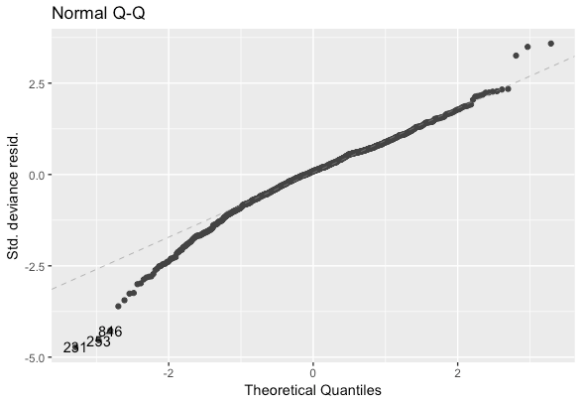

Figure 25: Normal QQplot

4.1.3 Summary and Comparison of the Response Variables

Table 8 displays some differences in the modelling of the two responses. It suggests that sequels only generate more revenue during the opening weekend. This implies that they have no effect on the rest of the gross. Although it can be argued that as $G^o$is a variable for G^t, then sequels will have an indirect effect on G^t. In fact if a model was fitted without G^o and underwent the same variable selection procedure, then the final model would include the sequel indicator. However it would not include rating. So rating is most likely insignificant for G^t.

Model	G^o	G^t
Cross Validation Error	0.271	0.32
Exclusive Variables	• T^o • Metacritic • Rating • Sequel	• Genre • IMDb • T^t • G^o
Common Variables	• Budget • Studio • Year • Season	

Table 8: Comparison table of G^o and G^t

IMDb was better for G^t , whereas Metacritic was preferred for G^o. This might mean that the critics have some influence on cinema attendance for the opening weekend. Perhaps some section of the public make their decisions based on their reviews. It further implies that tickets sold after the opening weekend might be caused by word of mouth or recommendations from friends.

The most curious comparison is that budget has contrasting effects on the two responses. For G^o it has a positive effect, whereas for G^t it has a negative effect. This suggests that big budget films peak during the opening weekend and subsequently fall sharply in the weeks after. However if G^o is removed from the G^tmodel, then budget becomes a positive coefficient. Perhaps the GLM is using budget to try and reduce the over-predicting in G^t. Or it might have correlation issues with some of the other variables. As shown in figure 26, there is some dependence between budget and genre. Action, animation, fantasy and Sci-fi all tend to have larger budgets compared with other genres. There is also a very clear trend that sequels have bigger budgets than original films.

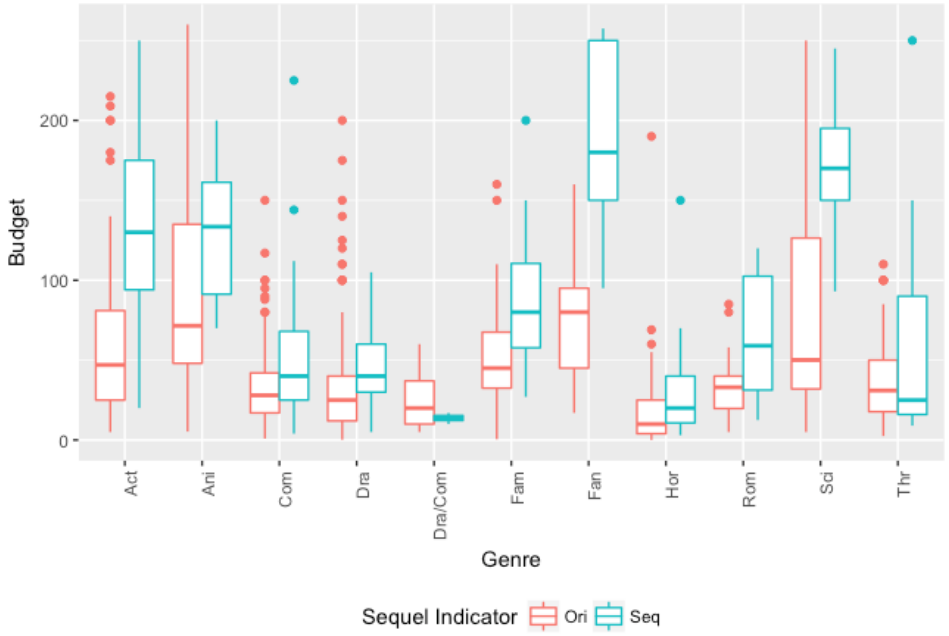

Figure 26: Boxplot of Genre vs Budget coloured by Sequel Indicator

Table 9 ranks the seasons by their coefficient estimates for the G^o and G^t models. Fall appears to be the lowest performing season for both responses. But the remaining four seasons have the complete reverse ranking between $G^o$and G^t. This is probably caused by the school holidays and the vacation timeframe. In the holiday and summer seasons, people have more free time during the week. And so are more likely to go to the cinema on a week day.

Rank	G^o	G^t
1	**Winter**	**Holiday**
2	**Spring**	**Summer**
3	**Summer**	**Spring**
4	**Holiday**	**Winter**
5	**Fall**	**Fall**

Table 9: G^o and G^t ranked by Season

4.2 Lasso Regression

For the lasso regression, two models were fitted for each response. The first model used all the suitable predictors. The second model included all the possible interaction terms.

4.2.1 Theory

Lasso Regression is an alternative regression modelling technique. [16] Unlike linear regression it doesn't minimise the residual sum of squares. Instead it minimises the RSS plus a shrinkage penalty term. The formula is provided here:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

This penalty term is the ℓ_1 norm of the coefficient vector β multiplied by a tuning parameter λ≥0. This tuning parameter is chosen separately. Although the method used for this project is cross-validation. So λ is found by minimising the mean squared error. The value of λ is key to estimating the coefficients. As λ increases, more of them shrink towards zero.

The main advantage of Lasso regression compared with other shrinkage methods, like ridge regression, is that it will set some of its coefficients to zero. So a reasonably large λ will allow the model to select the most important variables. Increasing λ will also decrease the variance but will increase the bias.

The Lasso can also be formulated as the following optimisation problem:

$$\text{minimize}_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}\quad\text{subject to}\quad\sum_{j=1}^{p}|\beta_j|\le s$$

The value for s is determined by the value of λ. The problem is essentially trying to minimise the RSS given the constraint that the sum of the absolute values of the coefficients is restricted. This also provides an intuition as to why some of the coefficients will be set to zero.

Figure 27 illustrates the optimisation problem with two variables. The blue diamond is the constraint region. $\hat{\beta}$ represents the coefficients that gives the minimal RSS. The red contour lines represent the RSS. So a solution is found when the contour lines expand outwards to first touch the constraint region. This is likely to be a corner of the region as demonstrated in the diagram. A lasso constraint will always have corners on all of the axes, hence the coefficient will be set to zero. In the particular case of the diagram β_1 will be set to zero. In the context of higher dimensions the idea is similar. If there are 3 parameters, the constraint region is a polyhedron. And when they are more than 3 parameters, the region is a polytope.

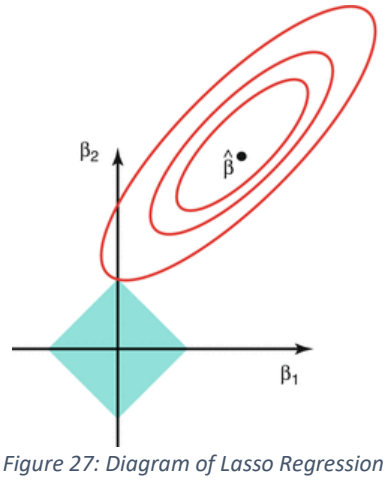

Figure 27: Diagram of Lasso Regression

The advantage Lasso regression has over standard linear regression is that it automatically performs variable selection. This is especially useful when there

are a lot of variables. It also deals with the issue of multi-collinearity by usually only selecting one of the variables from a correlated group.

A more detailed explanation of Lasso regression can be found in chapter 6 of '*An Introduction to Statistical Learning'* [17]. This chapter also contains content on other shrinkage methods like ridge regression.

4.2.2 Opening Weekend Gross Models

For modelling the opening weekend gross using lasso regression, all the available variables were used. This included both IMDb and Metacritic. However T^t was not included as it might not be known at the start of the release. This can often be the case with films that start with a limited release followed by a wide release at a much later date.

The general approach was to use cross-validation to find a value for λ. A lasso model is then fitted using λ as the tuning parameter. Figure 28 shows the mean-squared error (with confidence intervals) for different values of λ. The left most vertical line marks the λ that minimises the mean-squared error. In this case:

$$\lambda = 0.002308\,.$$

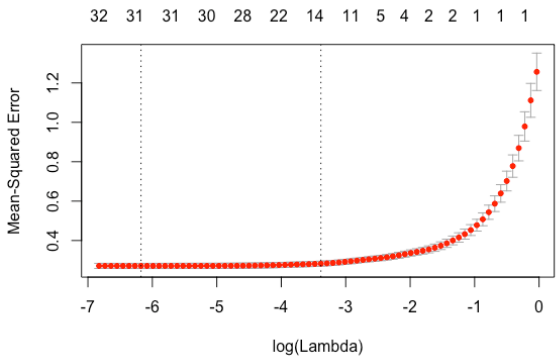

Figure 28: Cross-Validation Plot

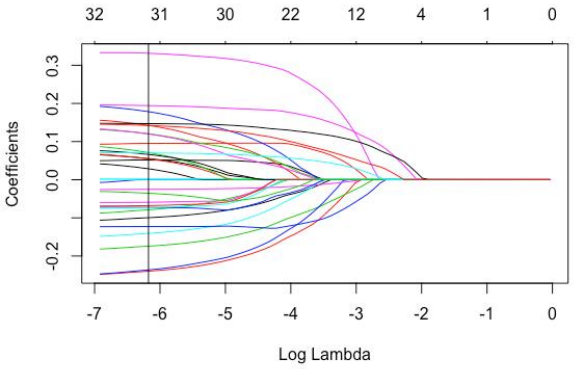

Figure 29: Lasso Coefficient Estimates by λ

Figure 29 shows how the estimates of the coefficients change as you increase λ. The vertical black line marks the value for lambda obtained from the cross validation and shows that a few of the coefficients have been set to zero. This method of finding λ is repeated for the other three models.

Model Interpretation

Variable	Coefficient Range		Percentage Range
	Minimum	Maximum	
Intercept	6.413×10^{1}	6.435×10^{1}	0.352%
Studio-Fox	-1.005×10^{-1}	-1.000×10^{-1}	0.454%
Studio-LG	0	0	NA
Studio-ORF	-2.413×10^{-1}	-2.399×10^{-1}	0.584%
Studio-Other	-1.796×10^{-1}	-1.750×10^{-1}	2.599%
Studio-Par	0	0	NA
Studio-Rela	-1.417×10^{-1}	-1.411×10^{-1}	0.487%
Studio-SGem	3.303×10^{-1}	3.313×10^{-1}	0.287%
Studio-Sony	5.112×10^{-2}	5.133×10^{-2}	0.414%
Studio-Uni	9.449×10^{-2}	9.466×10^{-2}	0.181%
Studio-WB	-7.937×10^{-2}	-7.899×10^{-2}	0.480%
Studio-Wein	-2.377×10^{-1}	-2.370×10^{-1}	0.279%
T^o	9.167×10^{-4}	9.174×10^{-4}	0.084%
Year	-2.524×10^{-2}	-2.513×10^{-2}	0.441%
Season-Hol	2.953×10^{-2}	3.086×10^{-2}	**4.421%**
Season-Spr	1.434×10^{-1}	1.451×10^{-1}	1.184%
Season-Sum	1.211×10^{-1}	1.219×10^{-1}	0.673%
Season-Win	1.790×10^{-1}	1.802×10^{-1}	0.637%
Budget	1.941×10^{-3}	1.951×10^{-3}	0.555%
Genre-Ani	-6.007×10^{-2}	-5.742×10^{-2}	**4.498%**
Genre-Com	5.714×10^{-2}	5.744×10^{-2}	0.516%
Genre-Dra	5.455×10^{-2}	5.639×10^{-2}	3.318%
Genre-Dra/Com	7.249×10^{-2}	7.424×10^{-2}	2.388%
Genre-Fam	-7.254×10^{-2}	-7.031×10^{-2}	3.120%
Genre-Fan	-7.168×10^{-2}	-6.858×10^{-2}	**4.430%**
Genre-Hor	1.218×10^{-1}	1.230×10^{-1}	0.959%
Genre-Rom	6.702×10^{-2}	6.786×10^{-2}	1.244%
Genre-Sci	-6.696×10^{-2}	-6.616×10^{-2}	1.211%
Genre-Thr	-3.391×10^{-2}	-3.359×10^{-2}	0.938%
Rating-PG	-1.232×10^{-1}	-1.214×10^{-1}	1.443%
Rating-PG13	6.985×10^{-2}	7.031×10^{-2}	0.661%
Rating-R	0	0	NA
Sequel	1.943×10^{-1}	1.945×10^{-1}	0.099%
MetaT	1.435×10^{-1}	1.481×10^{-1}	3.191%
IMDbT	1.426×10^{-1}	1.486×10^{-1}	4.120%

Table 10: Coefficient Table. Colours explained in Table 6

Sensitivity Analysis

In the sensitivity analysis λ was kept constant for all 5 data frames. From table 10, the coefficients don't look too sensitive to the NAs. All percentage ranges were below 4.5%.

Coefficient Remarks

The fitted model for G^o showed that all the variables were significant. Although the coefficient for R rating was reduced to zero. This means that it has the same effect as a G rated film. The studio coefficients for Lionsgate and Paramount were also set to zero, suggesting that they have the same

effect as Buena Vista (Disney). So the model was reduced to 31 parameters. Interestingly both Metacritic and IMDb were included after the variable selection, suggesting that the collinearity between them may only be mild. Plus their effects are quite similer in magnitude.

In comparison with the Gaussian GLM, the models were quite similar. Except that genre and IMDb were included. There were no parameters with contrasting effects. So the models were generally in agreement with each other.

Residuals

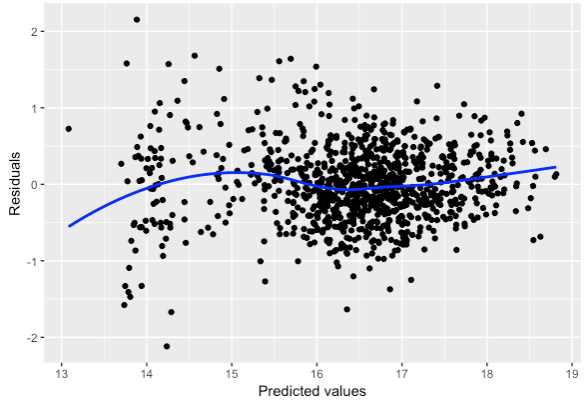

Figure 30: Residuals vs Fitted

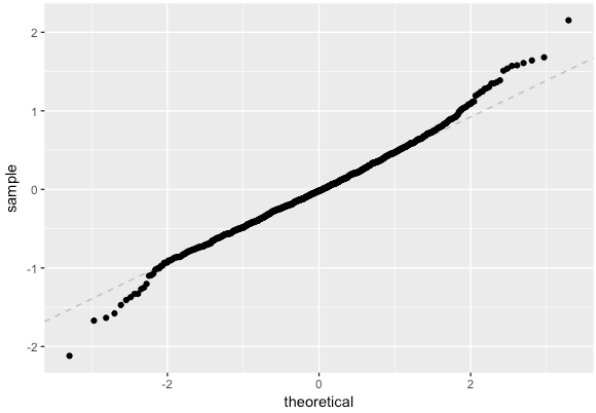

Figure 31: Normal QQplot

Figure 30 looks good as it doesn't exhibit much of a trend. However there is perhaps still a hint that the variance slightly increases for small fitted values. Like the previous GLM models, the QQplot (figure 31) is light tailed. Again suggesting that the errors follow a t-distribution and that there is some over-dispersion. The model also had mean cross-validation error of 0.2677.

Model with Interactions

A model with all the possible interaction terms was also fitted, initially consisting of 487 parameters.

$$\lambda = 0.1766$$

The variable selection reduced the model down to 109 parameters.

Interaction Coefficient Remarks

The largest Season-Genre interaction term was winter-horror. It suggested that horror films released in winter gross 23.5% more than action films in the fall. This is in line with figure 7 in the descriptive analysis.

A PG13 rated sequel is predicted to gross 13.4% more than a sequel from any other rating.

There were quite a few interaction terms with the romance genre. A romance sequel is expected to gross 11.7% more than an action sequel. They're also predicted to gross an extra 1.16% for every $20m in its budget compared with every other genre. However this is likely influenced by the very successful 'Twilight' films. This franchise made up half of the 8 romance sequels in the database and they all had very high budgets. This is further expressed by the coefficient for the studio Lionsgate interacting with romance, which increases the gross by 1.22%. Lionsgate distributed all five "Twilight" films.

Residuals

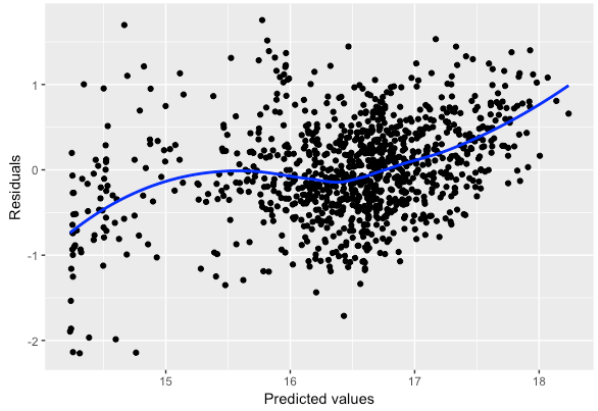

Figure 32: Residuals vs Fitted

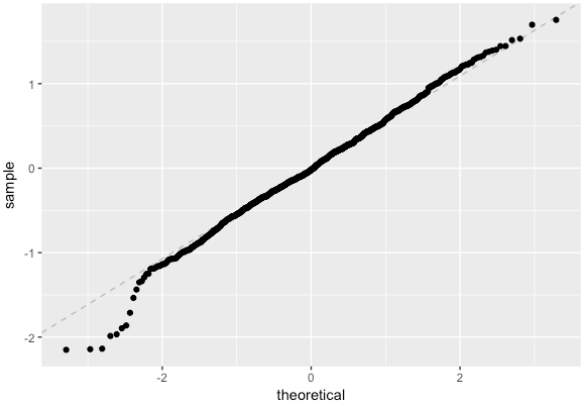

Figure 33: Normal QQplot

Figures 32 looks similar to the lasso model without interactions. However it does show a much stronger trend. It almost looks quadratic. On the other hand figure 33 shows some improvement in the QQplot, particularly with the positive residuals. This model had a mean cross-validation error of 0.2657.

4.2.3 Total Gross after the Opening Weekend Models

For G^t, all available predictor variables were used, including G^o, T^t, T^o.

$$\lambda = 0.001397$$

Variable	Coefficient Range		Percentage Range
	Minimum	Maximum	
Intercept	8.549×10^{1}	8.572×10^{1}	0.270%
Studio-Fox	2.899×10^{-2}	2.935×10^{-2}	1.235%
Studio-LG	0	0	NA
Studio-ORF	-2.982×10^{-1}	-2.970×10^{-1}	0.408%
Studio-Other	-1.430×10^{-1}	-1.387×10^{-1}	**3.033%**
Studio-Par	4.386×10^{-2}	4.411×10^{-2}	0.562%
Studio-Rela	-1.607×10^{-1}	-1.601×10^{-1}	0.315%
Studio-SGem	2.353×10^{-1}	2.363×10^{-1}	0.408%
Studio-Sony	2.080×10^{-1}	2.083×10^{-1}	0.152%
Studio-Uni	0	0	NA
Studio-WB	-6.000×10^{-2}	-5.963×10^{-2}	0.616%
Studio-Wein	-2.172×10^{-1}	-2.165×10^{-1}	0.351%
T^t	1.384×10^{-3}	1.385×10^{-3}	0.099%
G^o	1.258×10^{-8}	1.261×10^{-8}	0.236%
T^o	-5.161×10^{-4}	-5.145×10^{-4}	0.305%
Year	-3.571×10^{-2}	-3.559×10^{-2}	0.325%
Season-Hol	3.580×10^{-1}	3.593×10^{-1}	0.345%
Season-Spr	1.399×10^{-1}	1.418×10^{-1}	1.322%
Season-Sum	2.422×10^{-1}	2.430×10^{-1}	0.326%
Season-Win	1.958×10^{-1}	1.969×10^{-1}	0.557%
Budget	-1.129×10^{-3}	-1.118×10^{-3}	0.960%
Genre-Ani	1.202×10^{-1}	1.225×10^{-1}	1.882%
Genre-Com	2.549×10^{-1}	2.553×10^{-1}	0.159%
Genre-Dra	2.159×10^{-1}	2.174×10^{-1}	0.674%
Genre-Dra/Com	3.279×10^{-1}	3.298×10^{-1}	0.600%
Genre-Fam	2.832×10^{-1}	2.853×10^{-1}	0.717%
Genre-Fan	1.052×10^{-1}	1.089×10^{-1}	**3.466%**
Genre-Hor	7.660×10^{-2}	7.768×10^{-2}	1.406%
Genre-Rom	1.740×10^{-1}	1.751×10^{-1}	0.609%
Genre-Sci	9.479×10^{-2}	9.582×10^{-2}	1.081%
Genre-Thr	1.085×10^{-1}	1.089×10^{-1}	0.357%
Rating-PG	0	0	NA
Rating-PG13	0	0	NA
Rating-R	-7.551×10^{-2}	-7.496×10^{-2}	0.735%
Sequel	-5.483×10^{-2}	-5.470×10^{-2}	0.248%
MetaT	1.454×10^{-1}	1.506×10^{-1}	**3.543%**
IMDbT	3.741×10^{-1}	3.797×10^{-1}	1.499%

Table 11: Coefficient Table. Colours explained in Table 6

Sensitivity Analysis

The coefficient estimates again do not seem to be too sensitive. All the percentage ranges are low, with the largest being 3.543%.

Coefficient Remarks

The resulting model had 32 parameters. All the variables were significant, although Studio had two of its parameters set to zero, Lionsgate and Universal. Rating had two of its three parameters set to zero, PG and PG13. This reflects the variable selection of the GLM model, which removed rating. It also suggests that R rated films are expected to gross 7.3% less than the other three ratings. In comparison with the GLM model, all the coefficients agree in the direction of the effects. However, unlike the GLM, this model included T^o and Metacritic. T^o appears to have a negative effect on the gross. This perhaps allows compensation for films that open with a limited release but then receives a wide release. Both Metacritic and IMDb have positive effects, however IMDb's coefficient is a lot larger, suggesting it has more influence. Interesting the sequel indicator has been included, but this time it has a negative effect.

Residuals

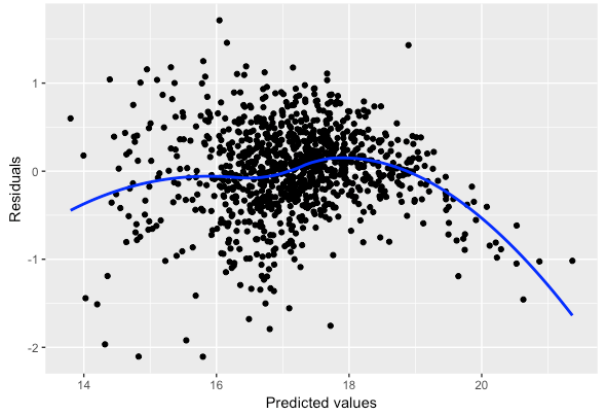

Figure 35: Residuals vs Fitted

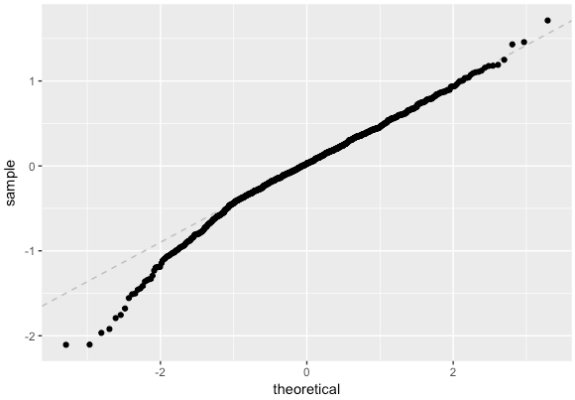

Figure 36: Normal QQplot

The residual plot (Figure 35) shows that the lasso model shares the same problem that the GLMs had, a negative residual trend for large values. This is not surprising as they are both linear models. The QQplot (Figure 34) shows a large tail for the negative residuals which is similar to the "per theatre" GLM. The cross validation error came to 0.2822.

Model with Interactions

Like with G^o, a model considering of all the interaction terms was fitted.

$$\lambda = 0.002026$$

The lasso reduced the model from 558 parameters to 315.

Interaction Coefficient Remarks

G^o had a few interesting interactions. It had a large positive interaction with the comedy genre. But had a negative interaction with romance films. The expected difference in gross between the two genres is 25% per $10m in G^o.

Residuals

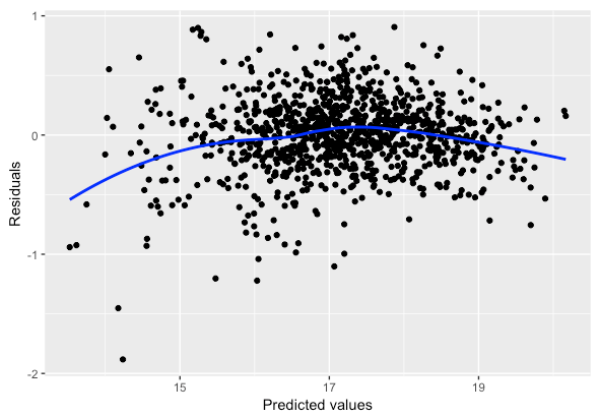

Figure 36: Residuals vs Fitted

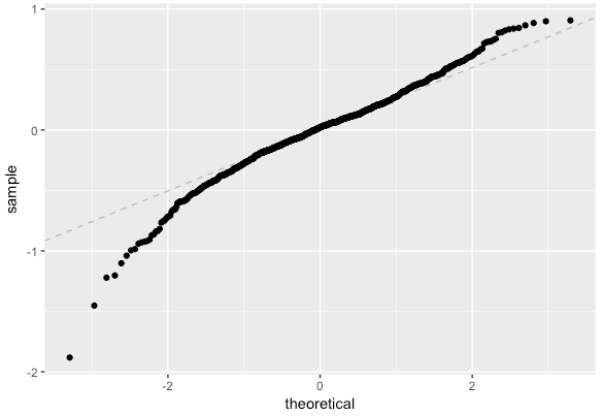

Figure 37: Normal QQplot

From the residual plots, there seems to be some improvement. The trend in figure 36 does not seems so strong, although it has become more concave. Also the QQplot (figure 37) still has the same issue of the heavy tail. The cross validation error was 0.2282.

4.2.4 Lasso Summary

Overall the Lasso models were an improvement over the GLMs.

The final lasso model for G^o was the one without interactions. Despite the interaction model having a smaller cross validation error, the difference was only slight. In fact the difference was less than 1%. And as the interaction model required over triple the number of parameters, it was regarded as unnecessarily complicated.

The model with interactions was judged to be best. It had a much lower cross validation error than the other model. The trend of the residuals was also less apparent.

Table 12 summarises the lasso models. It highlight the final models in purple. And it colours the lowest cross validation errors in red.

Model	λ	Number of parameters	Cross validation error
G^o	2.308×10^{-3}	31	0.2677
G^o with interactions	1.766×10^{-1}	109	**0.2657**
G^t	1.397×10^{-3}	32	0.2822
G^t with interactions	2.026×10^{-3}	315	**0.2282**

Table 12: Lasso Model Summary

4.3 Regression Trees

4.3.1 Theory

The third type of modelling that was considered was regression trees. [18] Regression trees work by breaking down the predictive space into a finite number of disjoint regions. The regions are defined by the explanatory variables that the model is using. From the training data, in this case the 1002 film dataset, each observation will belong to a particular region. Each region is allocated a prediction, which is usually the mean of its observations' response values. Any new observation that falls into a specific region takes that region's coefficient as its prediction value.

When constructing a regression tree the aim is to minimise the RSS. However the formula for the RSS is different from linear regression. The formula is as follows:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

$$\hat{y}_{R_j} = \frac{1}{|R_j|} \sum_{i \in R_j} y_i$$

Where $R_1 \dots R_J$ are the regions and $\hat{y}_{R_j}$ is the prediction value.

Due to the vast number of possible partitions of the predictive space, an algorithm is used to build a tree using a top down approach. The algorithm starts with a single region that contains all the observations. It then chooses one of the explanatory variables with a cut-off point and splits the region into two smaller regions: one with all the observations below the cut-off point and one with all those above it. The algorithm finds the value of the cut-off point that minimises the RSS. This binary splitting process repeats until a certain criteria is reached. A standard example for a criteria could be that all the regions have less than a certain number of observations.

The resulting tree is likely to be very large and has probably over fitted the data. This presents a problem for making predictions with new data. In order to reduce this issue, the algorithm prunes the tree. This means that it looks at the tree's subtrees and evaluates their prediction error by using cross-validation. However it would be impractical to examine every single subtree. So the algorithm uses cost complexity pruning to select a "sequence of subtrees indexed by a non-negative tuning parameter alpha".

The advantage of regression trees over linear regression are that they are very easy to explain and illustrate. They will also be more suited for non-linear data. This might be useful for G^t as figure 20 has shown signs of a non-linear structure.

A full explanation of regression trees can be found in chapter 8 of '*An Introduction to Statistical Learning*' [17]. It contains more detail on the how the cost complexity pruning works.

4.3.2 Opening Weekend Gross Model

The regression tree for G^o (figure 38) used all the available predictor variables except T^t. Unlike the linear models, the untransformed variables for Metacritic and IMDb were used. The regression tree only had one parameter, T^o. This indicated that T^o was the most important variable in reducing the RSS. The tree had six regions. The tree used a complexity parameter of 0.01. Complexity parameter in this context is the minimum amount a split has to increase R-squared in order to be executed. The value for it was selected to minimise the cross validation error.

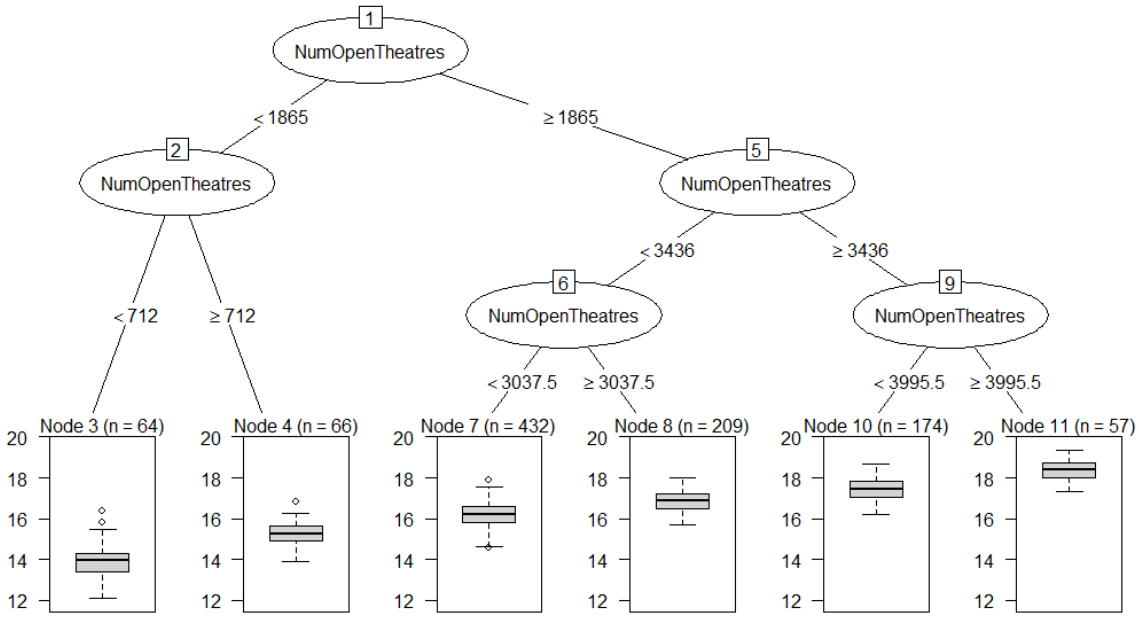

Figure 38: Illustration of the regression tree for G^o

Table 13 below provides the prediction values for all 6 regions. It also classifies them by T^o. Nodes are taken from figure 38.

Node	3	4	7	8	10	11
T^o	(0-711)	(712-1864)	(1865-3037)	(3038-3436)	(3436-3995)	(3995-∞)
Prediction	13.89	15.26	16.20	16.83	17.43	18.36

Table 13: Predictions for the G^o regression tree

Residuals

From figure 39, it appears that the residuals do not show any trend; with figure 40 shows that they follow a normal distribution. This is a definite improvement against previous models.

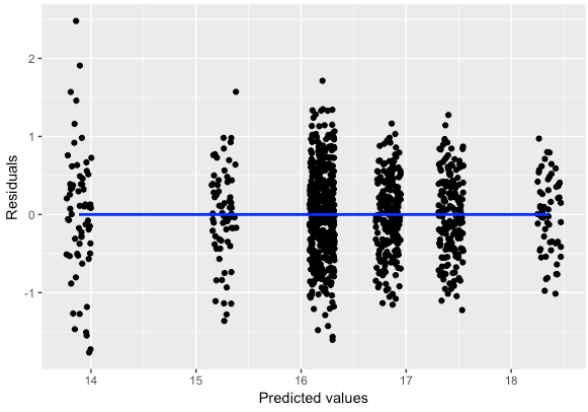

Figure 39: Residuals vs Fitted

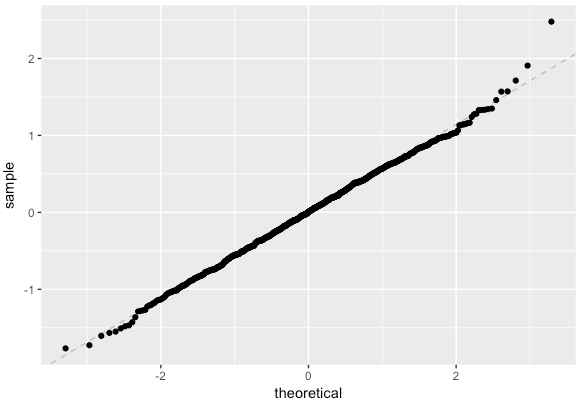

Figure 40: Normal QQplot

However it has a cross validation prediction error of 0.5938, which is much higher than previous models.

4.3.3 Total Gross after the Opening Weekend Model

For G^t, the regression tree (figure 41) is a bit more complicated. This time it uses three variables: G^o, T^t, T^o,and has splitted the data into eight regions. Four of which have resulted from purely splitting G^o. This tree also used a complexity parameter of 0.01.

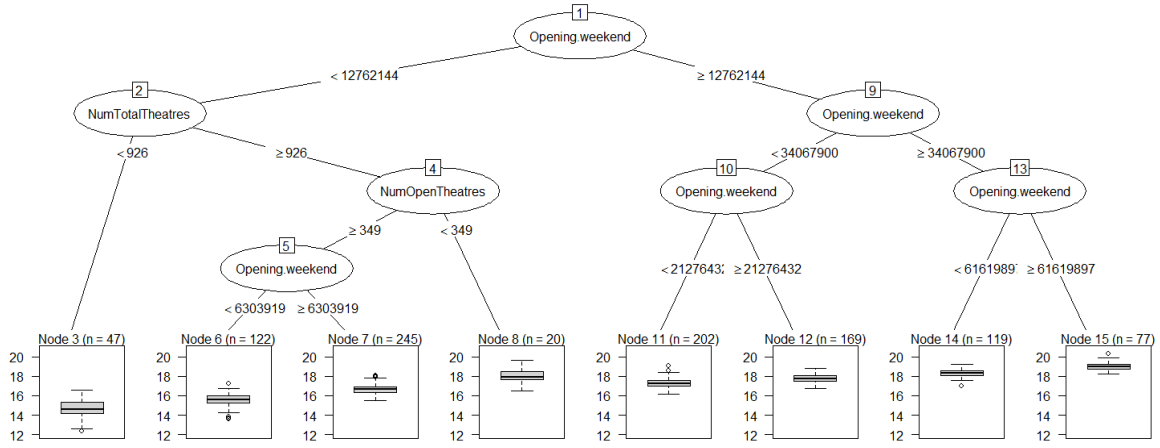

Figure 41: Illustration of the regression tree for G^t

Table 13 below provides the prediction values for all 8 regions. Nodes are taken from figure 41.

Node	3	6	7	8	11	12	14	15
Prediction	14.70	15.61	16.64	18.02	17.30	17.78	18.34	19.04

Table 14: Predictions for the G^t regression tree

Residuals

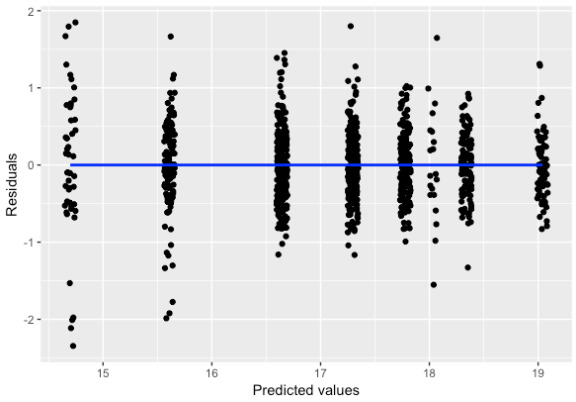

Figure 42: Residuals vs Fitted

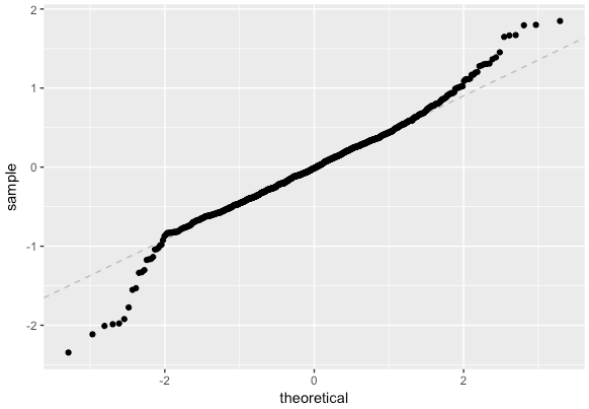

Figure 43: Normal QQplot

Figure 42 looks ok. It doesn't show any trend. Although the QQplot (figure 43) is light tailed.

The cross validation for this tree was 0.5342, which again is much greater the linear models.

Regression Trees Summary

Overall, the regression trees might have produced better residual plots. However they were hugely inferior in terms of prediction accuracy. The Lasso and GLM models both had lower cross validation errors. As Metacritic wasn't used in either of the final trees, these models aren't sensitive to the NAs.

4.4 Random Forests

4.4.1 Theory

Regression trees may be easy to use, unfortunately they are generally not good at predictions. This project alone has shown this to be true. To reduce this disadvantage, random forests were also considered for modelling. Random forest [19] build multiple trees from bootstrapped training data. The key difference is that at the splitting the tree only considers a subset of the explanatory variables. The number of parameters in this subset is commonly set to around the square root of the total number of parameters. The advantage here is that it gives other variables a chance to have an effect on the model. Whereas regression trees only use the most important and influential variables. This can be rather limiting as other variables may contain important information. An observation is then predicted by taking the average prediction value of all the trees.

In this project, m refers to the tuning parameter of the random forest. In other words, it is the number of variables that are randomly chosen for the subset that the tree considers. For both responses, $m \approx \sqrt{p}$, where p is the total number of variables.

As with regression trees, the theory of random forest is available in chapter 8 of '*An Introduction to Statistical Learning*' [17].

4.4.2 Opening Weekend Gross Model

The random forest for G^oused a tuning parameter of $m = 3$. Figure 44 shows the importance of the variables. The graph on the left shows the percentage increase of the mean squared error if a variable is removed. The graph on the right shows the increase of mean squared error by Node Purity. Node Purity is the idea of predicting an observation by following the opposite branches of a certain variable. And then comparing its mean squared error with its true prediction. For example the sequel indicator in figure 44 has a 'IncNodePurity' of 64.4. This is the increase in MSE if the observations were predicted using the incorrect category. So a sequel would be predicted as an original film and vice versa. These measures were calculated by the "importance" function for which more detail can be found in its RDocumentation page. [20]

From figure 44, the three most important variables are T^o, budget and studio which is generally in line with the other models. However the relative importance of Metacritic and IMDb is mixed. Under node purity Metacritic is more important. However, the removal of IMDb causes a larger increase in the MSE than the removal of Metacritic.

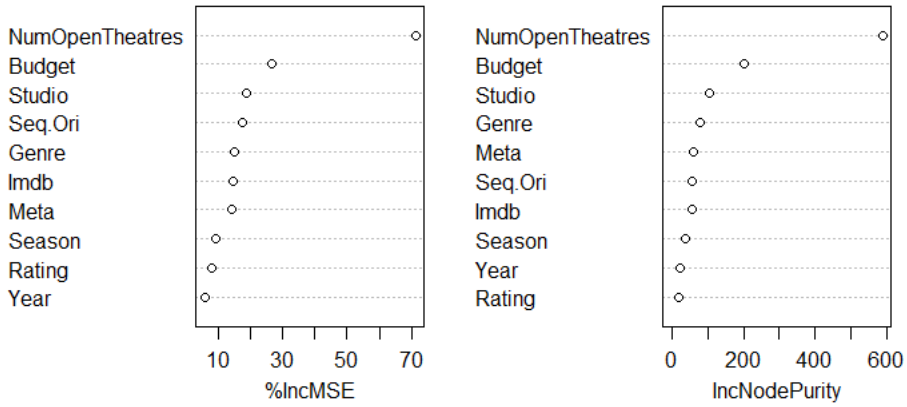

Figure 44: importance graphs of the G^o Random Forest

Residuals

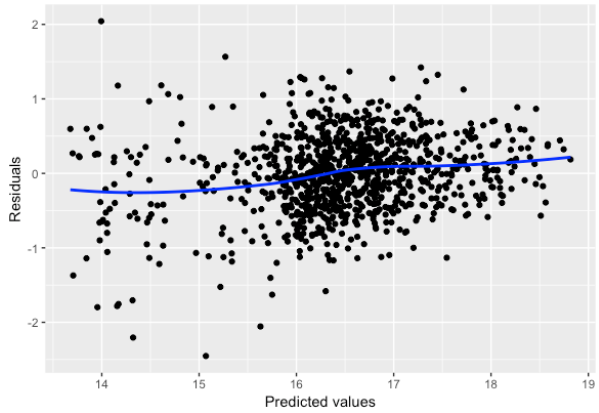

Figure 45: Residuals vs Fitted

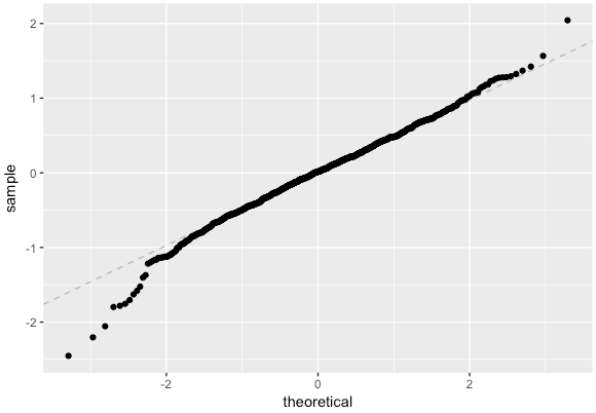

Figure 46: Normal QQplot

Figure 45 shows a slight upward trend in the residuals. Figure 46 appears to be light tailed. Although the positive residuals appears good with only one outlier, the QQplot is in fact worse when compared with the regression tree.

The model had a cross validation error of 0.5221.

4.4.3 Total Gross after the Opening Weekend Model

The random forest for G^tused a tuning parameter of $m = 4$. Like figure 44 for G^o, figure 47 shows the importance of the variables. As with the GLM and Lasso, the random forest sees IMDb more important than Metacritic. Similarly, the sequel indicator and rating are not particularly influential.

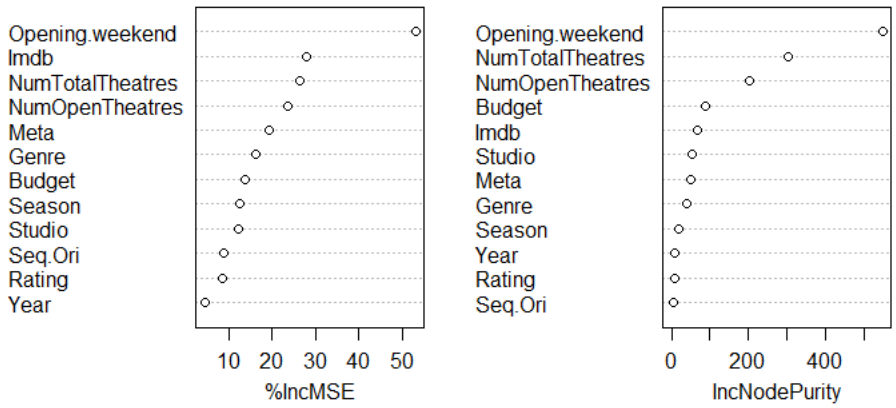

Figure 47: Importance graphs of the G^t Random Forest

Residuals

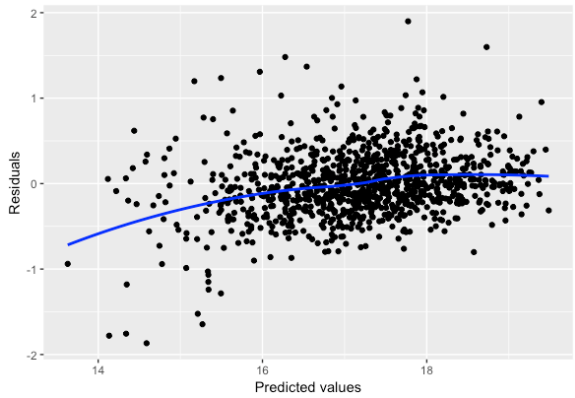

Figure 48: Residuals vs Fitted

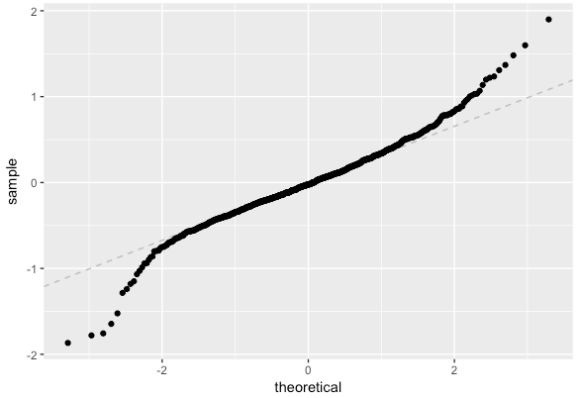

Figure 49: Normal QQplot

Figure 48 again shows a slight upward trend in the residuals. The QQplot in figure 49 shows signs of over-dispersion.

The model had a cross validation error of 0.4784.

Random Forests Summary

Overall, the random forests did make better predictions than the regression trees. However their accuracy were still inferior compared with the GLMs and Lasso models.

5. Conclusion

In this project, box office data was gathered from the internet. The data was examined with multiple graphs to identify important features. A variety of models were fitted using four different techniques. These models were then assessed on their residuals and parameter uncertainties. The best models from each technique were compared using 10 folds cross-validation.

In conclusion, predicting the box office is very difficult. Based on cross validation errors, Lasso regression was the best method to model both G^o and G^t. The regression trees and random forests were found to be the worst techniques at predicting data.

The most important variables for modelling G^o were T^o, budget and studio. In contrast the most important variables for G^t were IMDb, G^o, T^t and T^o. G^o in particular had a hugely positive impact on G^t. However, this effect appears to diminish as G^o got large. Both T^o and T^t also had positive influence on the gross.

The models suggest that sequels do make more money than original films. PG 13 is the preferred rating in maximising the gross. Both critics and audience scores are generally influential and have a positive effect. A big budget does usually mean a high grossing film. Screen Gems was the highest performing studio, whereas Open Road Films was the poorest performing one (with everything else equal). Box office does seem to be in decline as the year variable had a negative coefficient. The effects of season and genre were however quite mixed.

Tables 15 and 16, provide summaries for the final models of all four modelling techniques.

Model Method for G^o	Cross Validation Error (10 Folds)
GLM	0.2729
Lasso	**0.2677**
Regression Tree	0.5938
Random Forest	0.5221

Table 15: Summary of the G^o models

Model Method for G^t	Cross Validation Error (10 Folds)
GLM	0.3234
Lasso (with interactions)	**0.2282**
Regression Tree	0.5342
Random Forest	0.4784

Table 16: Summary of G^t models

Future Research

With regard to future research, it will be interesting to further investigate the effect of the opening weekend gross on the remaining revenue. G^o was a very significant variable in all the G^t models (and the most important one according to random forests). However the GLMs suggest that its effect might be non-linear. So perhaps some research could be done in finding a suitable transformation for this variable.

The marketing budget for a film would be a worthwhile variable to include in future analysis. Intuitively, it is likely to help explain more of the gross and lead to better predictions. Although it is recognised that the data in question is generally kept in secret and is difficult to obtain.

Another variable that deserves adding in would be the Rotten Tomato scores. The critics and audience scores used in this analysis are shown to be significant. So the Rotten Tomato scores could potentially help to explain more of the gross. It may also be worth considering these scores as categorical variables rather than continuous ones. For example the Metacritic scores can be changed to a 5 grade system as demonstrated in figure 50. [11]

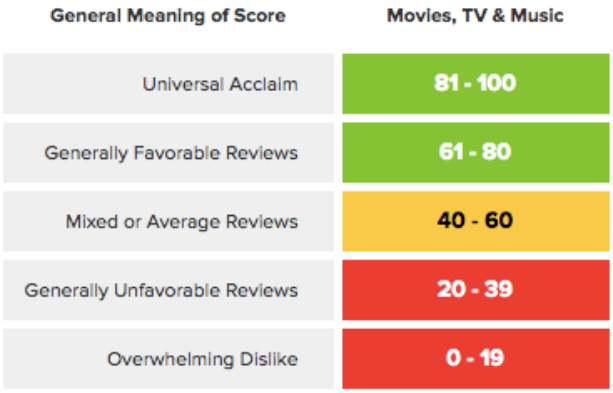

Figure 50: Metacritic scores classification

Future research should also consider splitting the data into two sets: films with wide releases and films with limited releases; and to try to fit separate models for them. It is possible that they may behave differently. The G^o GLM indicated some increase in variance for the residuals of films with low T^o

It may also be worth considering breaking down the response variable even further. So instead of just looking at G^o and G^t, consider examining the second weekend too. Or even discuss the percentage drop-off between the opening and second weekend. This figure is often used to assess a film's early performance. Only recently the film "Batman v. Superman: Dawn of Justice" had a colossal drop-off at 69% which was *"the second steepest drop in history for a marquee superhero title"* as said by the Hollywood Reporter. [21]

Finally to extend the idea of splitting the gross into G^o and G^t, it might be important to analyse the proportion of gross that is made during the opening weekend over the total gross. The formula for which is given below. This proportion could explain any potential differences in behaviour between G^o and G^t

$$Prop = \frac{G^o}{G^o + G^t}$$

6. References

1. Domestic Movie Theatrical Market Summary 1995 to 2017. *Available at: http://www.the-numbers.com/market/*
2. Mendelson, S. (2016) *'Why A Record-Breaking Year For Box Office Feels Like A Loss For Movies' Forbes*
3. Robehmed, N. (2016) '*The World's Highest-Paid Actors 2016: The Rock Leads With Knockout $64.5 Million Year' Forbes*
4. Yahav, Inbal (2016) *Network analysis: understanding consumers' choice in the film industry and predicting pre-released weekly box-office revenue.* Appl. Stoch. Models Bus. Ind. 32, no. 4, 409-422
5. Edwards, D. A., Buckmire, R. & Ortega-Gingrich, J. (2014) *A mathematical model of cinematic box-office dynamics with geographic effects.* IMA J. Manag. Math. 25, no.2, 233-257
6. Basuroy, S., Chatterjee, S. & Ravid, S. A. (2003) *How critical are critical reviews? The box office effects of film critics, star power, and budgets.* J. Mark,67, 103-117
7. Liu, Y. (2006) *Word of mouth for movies: its dynamics and impact on box office revenue*. J. Pop. Cult., 16,159-175
8. Zhang, W. Skiena, S. (2009). *'Improving Movie Gross Prediction Through News Analysis'* International Conference on Web Intelligence and Intelligent Agent Technology Vol 1, 301-304
9. Box Office Mojo. *Available at:* http://www.boxofficemojo.com/
10. OMDb API. *Available at:* http://www.OMDbapi.com/
11. Metacritic. *Available at:* http://www.metacritic.com/about-metascores
12. IMDb. *Available at:* http://www.IMDb.com/
13. Breiman. Friedman, *Journal of the American Statistical Association* (September, 1985)
14. RDocumentation aregimpute. *available at:* https://www.rdocumentation.org/packages/Hmisc/versions/4.0-2/topics/aregImpute
15. Kuhn, M. (2008). *`Building Predictive Models in R Using the caret'* Journal of Statistical Software Vol 28, Issue 5
16. Tibshirani, Robert. (1996). *Regression Shrinkage and Selection via the lasso*. Journal of the Royal Statistical Society. Series B (methodological) 58 (1). Wiley: 267–88.
17. James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning. 203-264, 303-335*
18. Rokach, Lior; Maimon, O. (2008). *Data mining with decision trees: theory and applications.* World Scientific Pub Co Inc.
19. Ho, Tin Kam. (1995). Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282*.
20. RDocumentation importance. *Available at:* https://www.rdocumentation.org/packages/randomForest/versions/4.6-12/topics/importance
21. McClintock, P. (2016). '*Box Office: Inside 'Batman v. Superman's' Historic Drop-off' The Hollywood Reporter*