University of New Hampshire

University of New Hampshire Scholars' Repository

Honors Theses and Capstones

Student Scholarship

Spring 2020

Predicting Gross Revenue Using Online Movie Reviews

Cara J. Small
University of New Hampshire, Durham

Undergraduate Honors Thesis

Predicting Gross Revenue using Online Movie Reviews

Cara J. Small

Peter T. Paul College of Business and Economics

University of New Hampshire

Honors Advisor: Dr. Ali Hojjat

Assistant Professor of Decision Sciences

May 2020

TABLE OF CONTENTS

Abstract

Today, many people check the ratings and reviews of a movie before they watch it. A review can be easily published online and seen by thousands of people, and this can change a person's opinion on whether or not they see the film. With the increasing presence of online platforms, it has changed the way people express their thoughts and feelings. There are many different platforms people can go to find varying opinions on a particular movie. This research will consider the problem of predicting a movie's overall gross revenue. We focus on ratings, reviews and information given on the IMDb Website. (https://www.imdb.com) A list of 4,265 movies and their corresponding reviews were collected between January 2005 and December 2019. In our research, we use RStudio and sentiment analysis to determine a review's emotions and opinions. We investigate whether the sentiment in a movie's reviews can predict overall gross revenue, as well as other predictors given on the IMDb website.

Keywords: Sentiment Analysis, IMDb, Movie Reviews, Gross Revenue

1. Introduction

The topic of this thesis is to investigate how the first twenty-five reviews given on a particular movie can predict its overall gross revenue. This paper will study the relationship between a movie's gross global revenue and the sentiment in the reviews posted on the IMDb website. The use of social media and the internet as a source of information, such as business and political news, is constantly increasing. Through the vast amount of internet activity today, we are able to find the interests, concerns, and intentions of the global population on various topics. Within that, through these social networks, users are expressing concerns and opinions on various movies. Social media and the internet give the opportunity to relay information to a large amount of people almost instantly. When a user posts their opinions on a particular movie, it can change somebody else's own opinion on whether or not they see the film. In turn, this can affect a movie's overall revenue. Predicting revenue for movies has been studied in economics, marketing, statistics, and forecasting. In this paper, we use the reviews given on the IMDb website, as well as other variables to predict overall gross revenue.

The purpose of this research is to understand which variables are the best indicators in predicting a movie's overall gross revenue. We explore the effectiveness and accuracy of a linear regression in this prediction.

2. Review of the Literature

In this section we provide a brief review of the techniques used in this study, such as, sentiment analysis, linear regression, data partitioning, and feature selection. The review of these techniques will give necessary background to help understand their importance in our study. We also discuss similar studies that predict gross revenue for movies. The existing models in academic literature are useful to us in this study to help guide our model structure.

2.1 Introduction

The quick and easy access to many online platforms has had a major impact on various aspects of individuals and society. Today, many businesses prioritize these platforms as a way to spread information to a large amount of people in a short amount of time. Organizations are becoming increasingly aware that social media holds a large influence on consumer purchase decision making. This, however, holds a potential risk for filmmakers as it gives consumers equal ability to share their satisfaction or dissatisfaction with a given movie. "The internet has become a central source of information for many people when making day-to-day decisions" (Curme et al, 2014, p.11600) After a consumer watches a movie, all their thoughts and opinions can be shared to a large network of people almost instantly. This alone can strongly affect whether or not someone in that network chooses to see that movie. A bad review will only discourage a fellow consumer from seeing a movie.

2.2 Similar Studies

There have been many studies performed on predicting the gross revenue of movies. Many of these studies predict revenue based on Twitter content. The use of Twitter content was not available to us in this study as we did not have access to all of the information needed. As a result, we chose to use the reviews given on the IMDb website.

A study by Wernard Schmit and Sander Wubben was done in 2015 to explore the predictive capabilities of Twitter data by using a collection of tweets to predict rating scores of newly released movies on IMDb."(Schmidt et al, 2015) This study differed from other research because it only focused on textual data given from Twitter as opposed to other social media platforms. This study explored both regression and classification methods. Schmit and Wubben found that, “IMDb rating scores can be predicted to a certain extent using a supervised machine learning approach.” (Schmidt et al, 2015, p. 125) They concluded that their best performing model is not optimal for predicting IMDb score, but it does show textual features that can be useful for predictions similar to this.

Another study by P. Thomas Barthelemy, Devin Guillory, and Chip Mandal hypothesized that an increased number of tweets about a movie before its release will result in increased box office revenue. (Barthelemy et al, 2012) This study used regression to create a model for predicting box office revenue. Th study leveraged data from both Twitter and IMDb. Twitter was used to gather data relating to the number of tweets about a given movie, while IMDb was used to identify the general attributes for each movie. Researchers found that their study had room for improvement as they recognized the revenue of a movie may also be determined by other factors outside of what was included in their study.

2.3 Sentiment Analysis

Sentiment analysis is the analysis of emotions and opinions in text. “Sentiment analysis finds and justifies the sentiment of the person with respect to a given source of content.” (Amolik et al, 2016, p. 2038). Today, because of the use of the internet, there is a vast amount of sentiment data through the form of tweets, blogs, status updates, posts, and reviews given online. Millions of users actively express themselves digitally, therefore generating a large amount of data every day. “The age of the internet has changed the way people express their thoughts and feelings.” (Amolik et al, 2016, p. 2038) Because of this new age of sharing thoughts and opinions on various topics, people now check the reviews or ratings of virtually everything before completing a purchase.

Sentiment analysis is classified into two different types, a feature or aspect-based sentiment analysis and an objectivity-based sentiment analysis. Feature or aspect-based sentiment analysis breaks down text into different attributes and then allocates each into a sentiment level, positive, negative, or neutral. Objectivity-based sentiment analysis does the exploration of the text, which relates to different emotions like anger, joy, and trust. In this study we use objectivity-based sentiment analysis. A list of sentiments used in this study can be seen in Figure 2 in the Appendix.

2.4 Linear Regression

A linear regression model examines the relationship between a dependent and independent variable(s) by fitting the linear regression equation against the given data. Linear regression is classified into two types, simple linear regression and multiple linear regression. Simple linear regression is used to show or predict the relationship between two variables or factors. The factor that is being predicted or solved for is the dependent variable, and the factor that is being used to predict the value of the dependent value is called an independent variable. When two or more independent variables are used to predict the dependent variable, it is considered a multiple linear regression. In this particular study, we use multiple linear regression, which can be given by the below equation:

$$y \ = b_0 + b_1 X_1 + \ b_2 X_2 + \ \ldots . + b_n X_n$$

y is the dependent variable that is being predicted, b_1, b_2 ... b_n are the coefficients and X_1,X_2...X_n are independent variables.

2.5 Data Partitioning

In a regression analysis it is important to partition the dataset used in order to avoid overfit issues. The goal of a linear regression model is to be able to produce an accurate prediction on a new dataset that the model has not seen. It is not sufficient to test the accuracy of a model on the same dataset used to fit the model. When a model performs well on the dataset used to fit the model, but not the new data set, it is considered to be overfit.

To avoid overfit issues and validate the data, it is common to partition the data into training and testing sets. The training set is used to fit the linear regression model and estimate the coefficients. A model will see and learn from the training dataset. A testing dataset is a sample of the data that is used to provide an unbiased evaluation of the final model fit on the training data. (Shah, 2010) The testing dataset confirms the model's accuracy when it is tested on a new dataset. It is expected that a model will perform better on the training dataset, as it was fitted specifically on that data. However, we do not want the results on the testing dataset to be drastically worse, as that would show overfit issues. In this study, the training dataset includes 25% of the data and the testing dataset includes 75% of the data. This can be seen in Figure 5 in the Appendix.

2.6 Feature Selection

In a multiple linear regression, there are many techniques used in order to build an effective model. Feature selection is often used as it is an easy way to exclude variables that do not contribute to predicting the dependent variable. There are many advantages to feature selection, the biggest one being that it creates a simpler and smaller model to work with. Not all of the independent variables in a study will be significant in predicting the dependent variable, especially when a large number of independent variables are considered. There are several methodologies used when performing feature selection.

Methodologies that are frequently used in feature selection are the best subsets method, forward selection, sequential replacement, and backward elimination. The best subset regression approach tests all possible combinations of variables and returns the best model with any given number of predictors. Forward selection involves starting with an empty model and adding variables one at a time. The variables that are added are the most impactful to the model. Lastly, sequential replacement replaces one variable at a time to see if a given variable enhances the performance of the model.

In this study we use backward elimination in order to find out which independent variables are significant in predicting the dependent variable. Backward elimination takes each independent variable and enters it into the linear regression equation. It then removes each independent variable one at a time if it does not contribute to the regression equation. Variables that impact the R- square least are dropped. Variables are removed until the model is reduced to the desired number of predictors.

We found that backward elimination is the best feature selection method to use because of the large amount of variables considered in this study. The best subsets regression approach would not be practical, as all possible combinations of the variables would have to be considered in order to return the best model. Through research of other similar studies, we found backward elimination down to twenty variables was the optimal number for our study. Considering the twenty most significant variables will give us a more accurate and simpler model.

2.7 Determining an Effective Model

Backward elimination determines which independent variables are significant in predicting the dependent variable. This, however, does not mean the overall linear regression is an effective model. There are many different ways to determine the best fit of a linear regression model. The parameters discussed are also used to determine which model is the most effective in predicting the dependent variable. One parameter used when determining an effective model is R squared. R squared shows how well the model fits, and ranges between zero and one. The closer the model is to one, the better the model. R squared explains the extent of variation of the dependent variable that's explained by the independent variables. R squared is given by the below equation:

$$R^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

The mean absolute percentage error, or MAPE, is useful in predicting the accuracy of the model. MAPE measures the size of the error in percentage terms. MAPE is given by the below equation:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Another parameter used in determining an effective model is mean squared error, or MSE. MSE is the averaged squared difference between the estimated values and the actual values. MSE is given by the below equation:

$$MSE = \frac{1}{n}\sum (y - \hat{y})^2$$

Another parameter used in this study for measuring our model is Root Mean Squared Error, or RMSE. RMSE is the standard deviation of the residuals. RMSE tells us how spread out the residuals are. RMSE is given by the below equation:

$$RMSE = \sqrt{\sum \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

The mean absolute deviation, or MAD, is the average distance between each data value and the mean. MAD is used to describe the variation in the mode. MAD is given by the below equation:

$$MAD = \frac{\sum_{i=1}^{n} xi - \bar{x}}{n}$$

3. Objective

The purpose of this study is to create an accurate predictive model that will predict the overall gross revenue of a movie based on the sentiments in a movie's online reviews. In order to create such a model, a few questions were considered:

1. Which variables are considered to be most useful when predicting a movie's overall gross revenue?
2. With what accuracy can we produce such a model?

4. Data Collection and Methodologies

In this section we discuss how the data was gathered and how we used R Studio to transform our raw data into a workable dataset for the regression analysis. We explain how sentiment analysis was performed and how the dataset was partitioned. The variables used in the regression analysis are explained, as well as the different models and datasets tested in the study.

4.1 Data Collection

The analysis of this study is conducted by collecting information on 4,265 different movies through the IMDb website. Before the regression analysis could be performed, the raw data needed to be cleaned and transformed into a workable dataset. Data was collected through the IMDb website by scraping the raw HTML data. (https://www.imdb.com) The data was collected during a 14-year period between January 2005 and December 2019. Our study includes movies starting in 2005 because we found the use of the internet and giving online reviews was not as popular before this. Only movies with "English" listed as the first language were used.

The first twenty-five reviews of each movie were collected, and if a movie had less then 10 reviews, it was not used in the analysis.

The information that was gathered from the IMDb website to be used in the study included: the title of the movie with a corresponding movieID, the date the movie was released, the set of genres it belongs to, the overall and initial rating, the count of reviews, the run time, the language(s), the country of origin, the budget of the film, the US gross revenue, and the open week revenue. A complete list of genres used can be seen in Figure 1 in the Appendix. If a movie belonged to a particular genre, a “one” was assigned, otherwise, a “zero” was assigned. In total, 94,521 reviews were collected and used.

4.2 Sentiment Analysis Data

The sentiment analysis was calculated by using the “syuzhet” package in RStudio. The “syuzhet” package is used to pull sentiment data from your own text files. Through this package, sentiment is pulled through 10 different emotions – Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust, Negative, and Positive. The sentiments used can be seen in Figure 2 in the Appendix.

A function was created in R Studio to remove the movie title from the corresponding reviews. The movie title was removed from each review in order to avoid compromising the sentiment analysis. For example, looking at the sentiment output from the movie “The Fantastic Four”, it was showing more positive sentiment than was true. This is because every mention of the title or the word “Fantastic” showed as a positive sentiment, therefore altering the data. A linear regression was run on the data that included the movie title in the sentiment analysis, as well as without it, to see if it improved the results in the study.

We also tested the function “stopwords” found in the “TM” package in R Studio on our data. The stop words function is used to filter out commonly used words in the English language, like the word “the”. The stopwords function is used to allow more focus on other words in the given text for the sentiment analysis. The expectation of using the stopwords function was that it would bring more focus to other words, therefore strengthening the sentiment analysis scores. We ran a linear regression on the data that included the stopwords function, as well as without it, to see if it improved the results in the study. We found that the stop words function did not improve our results, so we did not use this is our final dataset.

The sentiment analysis was then applied to all 96,224 reviews. The sentiment scores for all reviews for the corresponding movies were averaged together for the final dataset. The average of the sentiment scores was taken because not all movies had twenty-five reviews. We took the square root, square and log of the sentiment scores to be used in the regression analysis.

4.3 Interaction Variables

To show the correlation between the sentiment scores and the genre of the movie, interaction variables were created between the two. All ten emotions in the sentiment analysis were multiplied by the twenty-two different genres, in return creating 220 new columns. (See Figure 1 and Figure 2 in the Appendix for the list of sentiments and genres used)

4.4 Raw and Transformed Variables

A regression analysis was run using the raw variables and the transformed variables to see which gave better results. The raw variables consisted of the year the movie was released, the open week revenue, the initial rating, the run time, whether the movie was multilingual, and whether the movie was multinational. The transformed variables included all the raw variables as well as the interaction variables, the square, square root, and log of the sentiment scores, initial rating and runtime.

4.5 Differing Models

The linear regression was run on two different models. Model 1 did not include opening week revenue in the regression analysis. Model 2 included opening week revenue as a variable in the regression analysis. Two models were also created for comparison purposes to show how the linear regression improved when the opening week revenue was included in the model.

4.6 Differing Datasets

We had three final datasets we tested to see which gave the best results. Our first dataset, referred to as MovieReview_Orig, did not remove movie titles or stop words from movie reviews. The second dataset, referred to as MovieReview_Title, only removed movie titles from the movie reviews. The final dataset, referred to as MovieReview_TitlePlus, removed movie titles and stop words from the movie reviews. Each data set was tested in the regression analysis to see which gave the more accurate results.

4.7 Data Partitioning

The final datasets were partitioned into a training dataset and a testing dataset. The training dataset consisted of 75% of the data, or 3,198 rows. The testing data set consisted of the remaining 25% or 1,067 rows. This can be referenced in Figure 4 in the Appendix.

4.8 Linear Regression

In this study, we considered one response variable. The response variable was the total revenue generated by a movie in U.S. dollars. A linear regression was calculated to determine the relationship between gross revenue and all variables collected. A regression analysis was first run on the differing datasets of MovieReview_Orig, MovieReview_Title, and

MovieReview_TitlePlus. Once the more accurate dataset was chosen based off of R square, a regression analysis was run using Model 1 and Model 2. Each model was then run using the raw and the transformed variables to see if there was improvement in the overall results.

5. Results

In this section we discuss the final results of our study and which datasets, variables and models gave the best results. We also discuss and highlight the parameters used to measure the accuracy of the study.

Through this study we investigate the relationship between the gross global revenue and movie review sentiment during a 14-year period. First, we tested all three datasets using the raw variables on Model 1 and Model 2 to see which was more effective in predicting overall gross revenue. The dataset that gave the best results was the second dataset, referred to as MovieReview_Title. This dataset only removed movie titles from the movie reviews. This was determined by estimating a linear regression model using the raw variables and the training dataset and referencing the given R square. Model 1 and Model 2 were both used when estimating the linear regression. MovieReview_orig, which did not remove movie titles or stop words from the reviews, had a r squared of 0.4699 for Model 1 and 0.8397 for Model 2. MovieReview_Title, which only removed movie titles from the reviews, had an R square of 0.47 for Model 1 and 0.8396 for Model 2. The last dataset, MovieReview_TitlePlus, which removed movie titles and stop words, had a R square of 0.4633 for Model 1 and 0.8383 for Model 2. The given results can be seen below:

Dataset	**Model 1**	**Model 2**
MovieReview_orig	0.4699	0.8397
MovieReview_title.csv	0.47	0.8396
MovieReview_titlePlus	0.4633	0.8383

By comparing each datasets r squared, it can be determined that only removing the movie titles from the review provided more accurate results when predicting a movie's gross revenue. For Model 1 the R square was highest when using the MovieReview_Title dataset. Although the R square was slightly lower using the MovieReview_title for Model 2, we are focused on the results given in Model 1 in this study. The MovieReview_Title dataset was used for the remainder of the study.

Once the more accurate dataset was chosen, a linear regression was fit using each model and each set of variables. Backward elimination was then used on both models and both sets of variables to determine which 20 variables had the most significance when determining a movie's gross revenue. We compared the linear regression results when backward elimination was and was not used. The given R Square values can be seen below:

Model 1:	**Train**	**Test**
Transform Variables		
Full Regression	0.5839	0.3887
Backward Elimination	0.4894	0.4151
Raw		
Full Regression	0.47	0.443
Backward Elimination	0.4667	0.4371

Model 2:	**Train**	**Test**
Transform Variables		
Full Regression	0.8766	0.8574
Backward Elimination	0.8531	0.8682
Raw		
Full Regression	0.8396	0.8829
Backward Elimination	0.8391	0.8831

Looking at Model 1 and the training data, it can be determined that the more accurate model used transformed variables, but not backward elimination. It can also be determined that when opening week revenue is included in the analysis, the results improve immensely. In Model 2, the better model also used transformed variables, but not backward elimination.

Looking at Model 1 using transformed variables it can be seen that our model did show overfit issues. When we look at Model 1 on the "full regression", the training accuracy is 58.39% and the testing accuracy is 38.87%. This shows the model did not perform as well on the testing dataset. After feature selection was used, the accuracy outputs are much closer together. The training accuracy was 48.94% and the testing accuracy was 41.51%. Although the accuracy was lower when backward elimination was used, we are confident that the accuracy level is similar when applied to new data. Similarly, when the raw data was used, the prediction accuracy was lower, but the training and testing accuracy was much closer together. This shows the model performed similarly on a new dataset.

Looking at Model 2, the accuracy was much higher for raw and transformed variables compared to Model 1. These results were expected, as we predicted that opening week revenue can be an accurate predictor when forecasting the overall gross revenue of a movie. When transformed variables were used on the full regression our model's accuracy on the training dataset was 87.66% and 85.74% on the testing dataset. The training and testing accuracy were very close showing we did not have overfit issues on this model. When feature selection was performed on the transformed variables on Model 2, the model performed better on the testing dataset then on the training dataset. The training accuracy was 85.31% and the testing accuracy was 86.82%. Similarly, when the raw variables were used, the testing accuracy was higher than the training accuracy.

Although the prediction accuracy was higher when feature selection was not performed for Model 1, we will use the transformed variables and backward elimination results for the remainder of this analysis. This is because when feature selection was performed it solved the overfit issues. Additionally, the transformed variables will be used because the prediction accuracy is higher compared to the raw variable results. The given R Square, MAPE, MSE, MAD, and RMSE results for when feature selection and transformed variables were used can be seen below:

	Train	**Test**
Model 1:		
R Square	0.4894	0.4151
MAPE	48908.94	118823.7
MSE	17946.05	13811.46
MAD	68.4366	64.8616
RMSE	133.9629	117.5222
Model 2:		
R Square	0.8531	0.8682
MAPE	6840.823	5188.885
MSE	5161.888	3112.343
MAD	30.2485	26.7196
RMSE	71.8463	55.7884

When backward elimination was run, the variables that were found significant in predicting overall gross revenue for Model 1 were:

- InitRating
- RunTime
- Anticipation
- Action:Negative
- Action:Positive
- Adventure:Anticipation
- Adventure:Joy
- Adventure:Surprise
- Sci.Fi:Trust
- Animation:Disgust
- Animation:Joy
- Drama:Positive
- War:Anticipation
- Biography:Anticipation
- Biography:Sadness
- News:Fear

- Adventure:Negative
- Family:Anticipation
- Family:Joy
- Family:Negative
- News:Surprise
- News:Positive
- Year_Exp

When backward elimination was run, the variables that were found significant in predicting overall gross revenue for Model 2 were:

- RunTime
- OpenWeek
- Adventure:Anger
- Adventure:Anticipation
- Adventure:Joy
- Adventure:Trust
- Adventure:Positive
- Family:Anger
- Family:Anticipation
- Fantasy:Negative
- Fantasy:Positive
- Animiation:Joy
- Animation:Positive
- Musical:Anger
- Musical:Disgust
- Comedy:Disgust
- Comedy:Sadness
- Biography:Anger
- Biography:Anticipation
- News:Fear
- News:Surprise
- News:Positive
- Year_exp

6. Conclusion and Future Work

We explored the accuracy of using linear regression methodology to predict the overall gross revenue of movies. We first began by collecting a large amount of raw data from the IMDb website and transforming this into a workable dataset for linear regression. We only collected movies with "English" listed as its first language between January 2005 and December 2019. This narrowed our dataset to a list of 4,265 different movies. We then wrote a function in R Studio that removed the movie title and stop words from the associated reviews. Through R Studio, sentiment analysis was run to gather the sentiment in the given review. Interaction variables were created to show the relationship between a movie's genre and the sentiment given in the reviews. Transformed variables were created to see if they improved the regression analysis. We ran two different regression analyses, one included opening week revenue as an independent variable and one did not. In total we had three differing datasets to choose from, two different models, and two varying sets of variables.

After running the regression analysis it was found that removing movie titles from the associated reviews but not stop words improved the accuracy in the results. Predicting accuracy was improved when transformed variables were used and opening week revenue was included. When feature selection was not used, the accuracy was higher, but showed overfit issues. Backward elimination improved our overfit issues and showed confidence that our model will provide similar results when applied to a new dataset.

In this regression analysis, even though we tested various datasets, models and variables, there is still room for improvement. Revenue of a movie may also be determined by other factors then what is included in this study. For example, the revenue of the lead actor, a movie's budget, and whether the movie was part of a series can all have an effect on the financial success of a movie.

To improve the results of this study, including the previously mentioned variables could help the accuracy of the analysis. Additionally, in this study we were only able to collect at most twenty-five reviews from each movie. Including more reviews in the sentiment analysis could help improve the accuracy. In future work, using Twitter data in replacement of the reviews given on IMDb would be interesting to see if it improved the overall results. By using Twitter data, the amount of times the movie is mentioned on Twitter can be included as an independent variable. As this would show how popular the movie was during the time of release.

References

Amolik, Akshay & Jivane, Niketan & Bhandari, Mahavir & Venkatesan, M.. (2016). Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques.. International Journal of Engineering and Technology. 7. 2038-2044.

Barthelemy, P. Thomas & Guillory, Devin & Mandal, Chip. (2012). Using Twitter Data to Predict Box Office Revenues. CA, United States: University of Stanford

Bouzianis, Stephen, "Predicting the Outcome of NFL Games Using Logistic Regression" (2019). *Honors Theses and Capstones*. 474. https://scholars.unh.edu/honors/474

Chester Curme, Tobias Preis, H. Eugene Stanley, Helen Susannah Moat
Proceedings of the National Academy of Sciences Aug 2014, 111 (32) 11600
11605; DOI:10.1073/pnas.1324054111

Dooms, S., Pessemier, T.D., & Martens, L.R. (2013). MovieTweetings: a movie rating dataset collected from twitter. *RecSys 2013*.

Gray, Alexis A., "Brands Take a Stand for Good: The Effect of Brand Activism on Social Media Engagement" (2019). Honors Theses and Capstones. 440.

Hennig-Thurau, Thorsten & Wiertz, Caroline & Feldhaus, Fabian. (2014). Does Twitter Matter? The Impact of Microblogging Word of Mouth on Consumers' Adoption of New Movies. Journal of the Academy of Marketing Science. 43. 10.1007/s11747-014-0388-3.

Joshi, Mahesh & Das, Dipanjan & Gimpel, Kevin & Smith, Noah. (2010). Movie Reviews and Revenues: An Experiment in Text Regression.. 293-296.

Schmit, Wernard & Wubben, Sander. (2015). Predicting Ratings for New Movie Releases from Twitter Content. 122-126. 10.18653/v1/W15-2917.

Shah, T. (2017, December 10). About Train, Validation and Test Sets in Machine Learning. Retrieved from https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

Appendix

Figure 1: Movie Genres

Action	Drama
Adventure	Musical
Family	Romance
Fantasy	Comedy
Sci.Fi	Mystery
Animation	Horror
Music	Sport
Documentary	War
Biography	Western
News	History
Thriller	Crime

Figure 2: Sentiments

Anger	Sadness
Anticipation	Surprise
Disgust	Trust
Fear	Negative
Joy	Positive

Figure 3: Variables

ID	Fantasy	Crime	History	Disgust	Positive	Multilingual
Title	Sci.Fi	Thriller	Music	Fear	InitRating	Country
Year	Animation	Mystery	Documentary	Joy	Rating	Multinational
ReleaseDate	Drama	Horror	Biography	Sadness	ReviewCount	Budget
Action	Musical	Sport	News	Surprise	NumComments	OpenWeek
Adventure	Romance	War	Anger	Trust	RunTime	GrossUS
Family	Comedy	Western	Anticipation	Negative	Language	GrossGlobal

Figure 4: Training and Testing Data

	Train	Test	Total
Percentage	25%	75%	100%
Number of Rows	1,067	3,198	4,265

Figure 5: MovieReview_title Comparison of Results

		Full Regression		Feature Selection	
		Train	**Test**	**Train**	**Test**
Transform Variables	**Model 1:**				
	R Square	0.584	0.389	0.4894	0.4151
	MAPE	58330	122397	48908.94	118823.7
	MSE	14623	14434	17946.05	13811.46
	MAD	62.88	64.63	68.4366	64.8616
	RMSE	120.93	120.15	133.9629	117.5222
	Model 2:				
	R Square	0.877	0.857	0.8531	0.8682
	MAPE	17105	21741	6840.823	5188.885
	MSE	4335	3366	5161.888	3112.343
	MAD	29.12	29.46	30.2485	26.7196
	RMSE	65.84	58.02	71.8463	55.7884
		Train	**Test**	**Train**	**Test**
Raw Variables	**Model 1:**				
	R Square	0.47	0.443	0.4667	0.4371
	MAPE	61691.58	106396.7	59555.86	118913.4
	MSE	17469.48	16828.14	17576.38	17005.74
	MAD	68.6262	69.9216	68.5718	70.2736
	RMSE	132.1722	129.7233	132.5759	130.406
	Model 2:				
	R Square	0.8396	0.8829	0.8391	0.8831
	MAPE	8789.046	7246.67	8808.93	7730.855
	MSE	5285.792	3536.335	5302.115	3532.957
	MAD	29.9546	29.2531	29.9584	29.2758
	RMSE	72.7034	59.4671	72.8156	59.4387