International Journal of Business Marketing and Management (IJBMM)
Volume 3 Issue 5 May 2018, P.P. 87-96
ISSN: 2456-4559
www.ijbmm.com

Predicting Gross Revenue of Movies with Machine Learning

Andy W. Chen[1]
[1] *University of British Columbia, Vancouver, BC, Canada*

ABSTRACT: *In this paper, I propose a number of machine learning models to predict the gross revenue of movies using features such as genre, length of movie, actors, director, production company, movie recognition, and release year. Using the sum of squared errors to evaluate the models, I find the model with the most number of the above features and transformation of these features yield the most accurate predictions for both training and test sets. I also find that for small sample size (from 0% to 20%), the prediction accuracies are low. When the sample size used for training is greater than 20%, the prediction accuracies increase significantly and stay consistently higher for both training and test sets.*

KEYWORDS: *Movie Industry, Machine Learning, Statistics, Data Science*

I. INTRODUCTION

The movie industry is often a high risk, high reward business. The gross revenue made from a movie can be unpredictable at times. There are highly anticipated movies that turn out to be mediocre or below expectation, while there are movies with low budget and relatively unknown cast but perform beyond expectation. In this paper, I propose several machine learning models that predict the gross revenue of movies using features such as movie genre, run time of movie, cast, director, production studio, awards nominated and won, and year of release. I compare the results of these machine learning models using the squared error as the performance metric.

Research in this area includes work by Nelson and Glotfelty[1] use data from IMDB and find that a top star actor would bring over $16 million revenue than an average star actor. Einav[2] find that seasonality plays an important role in affecting the demand for movies. Elliot and Simmons[3] analyze 527 movies released in the United Kingdom and find that amount of advertising affects gross revenue, while advertising itself is affected by reviews of critics. Frank[4] studies the optimal timing of movie's release in the video market depends on the movie's performance at the box office. Chisholm and Norman[5] find that significant impact of location on a movie's gross revenue at the box office.

II. METHODOLOGY

I use a database of 40,000 movies with basic movie information and gross revenue. I process the data by keeping only the movies released in or after year 2000. I also remove movies with errors in the year of release, but keep at least 90% of the data. The final dataset contains 2978 movies. Table 1 shows the variables in the dataset.

Title	Year	Rating
Release Year	Runtime	Genre
Director	Writer	Actors
Plot	Language	Country
Awards	Poster	Metascore
imdbRating	imdbVotes	imdbID
Type	tomatoMeter	tomatoImage
tomatoRating	tomatoReviews	tomatoUserMeter
tomatoUserRating	tomatoUserReviews	DVD
BoxOffice	Production	Budget
Gross Revenue	Date	Month

Table 1. Variables in the Dataset

I build several machine learning models, least squared linear regression, for predicting gross revenue of movies. I compare the performance of the models by computing the training and test root mean squared rrror (RMSE) at different training set sizes. I do several iterations of this by randomly drawing samples of 5%, 10%, 20%, 30%, . . . , 100% of the data and use it as the training set. The rest of the data is used as the test set. I compute the RMSE of the trained model's prediction on the training and test sets separately. I repeat the above for each model 20 times at each training set size and average the RMSE results for stability.

The first model I train contains only numeric variables; the second contains numeric variables and transformed numeric variables; the third contains categorical variables; the fourth contains both numeric and categorical variables; and the fifth contains all the variables plus extra ones derived from the original variables.

III. RESULTS AND DISCUSSION

The first linear regression model contains only numeric variables Year, Month, Runtime, imdbRating, tomatoRating, and tomatoUserRating. Figure 1 and 2 shows the RMSE of predictions on the training and test sets at different training sample sizes. The best mean test RMSE is 104,717,924 at 95% training set size. The results may vary for each run because the samples are drawn randomly, but the overall trend should be similar.

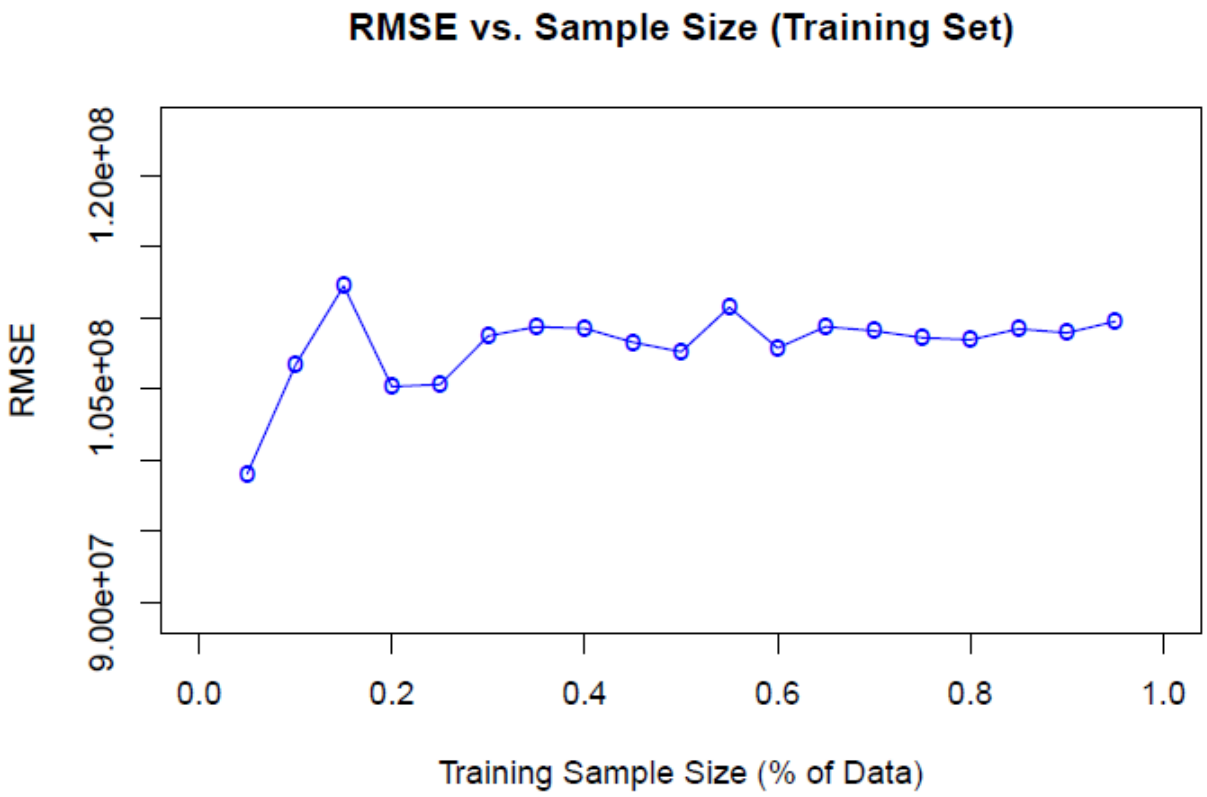

Figure 1. Model 1 RMSE vs. Training Sample Size for Training Set Predictions

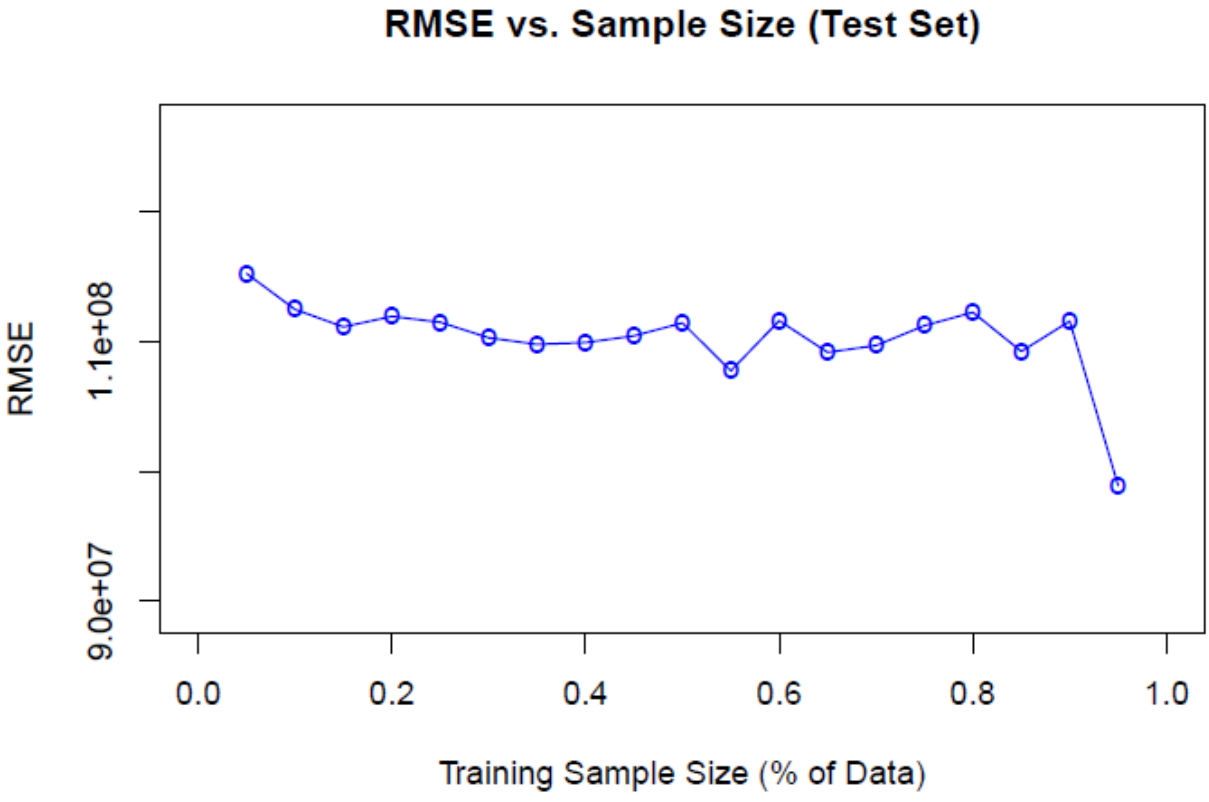

Figure 2. Model 1 RMSE vs. Training Sample Size for Test Set Predictions

For the second model, I add some transformations of the numeric variables. I include quadratic and cubic Budget terms because the scatter plot of Budget vs. Gross shows that a linear relationship may not be enough to capture the relationship between the two variables. I create 2 bins for Year (before and after 2010) because the movies released after 2010 had higher mean gross revenues than the ones before 2010. I also create 4 bins for Month to represent the 4 seasons because the movie industry is highly seasonal and movies released in certain seasons (for example, summer and winter) may have higher gross revenue because more people (especially students) are on vacation. The best mean test RMSE is 92,742,444 at 95% training set size. This is smaller than the RMSE in model 1. The reduction is likely due to the variable transformations and extra variables added to provide more predictive power for the model.

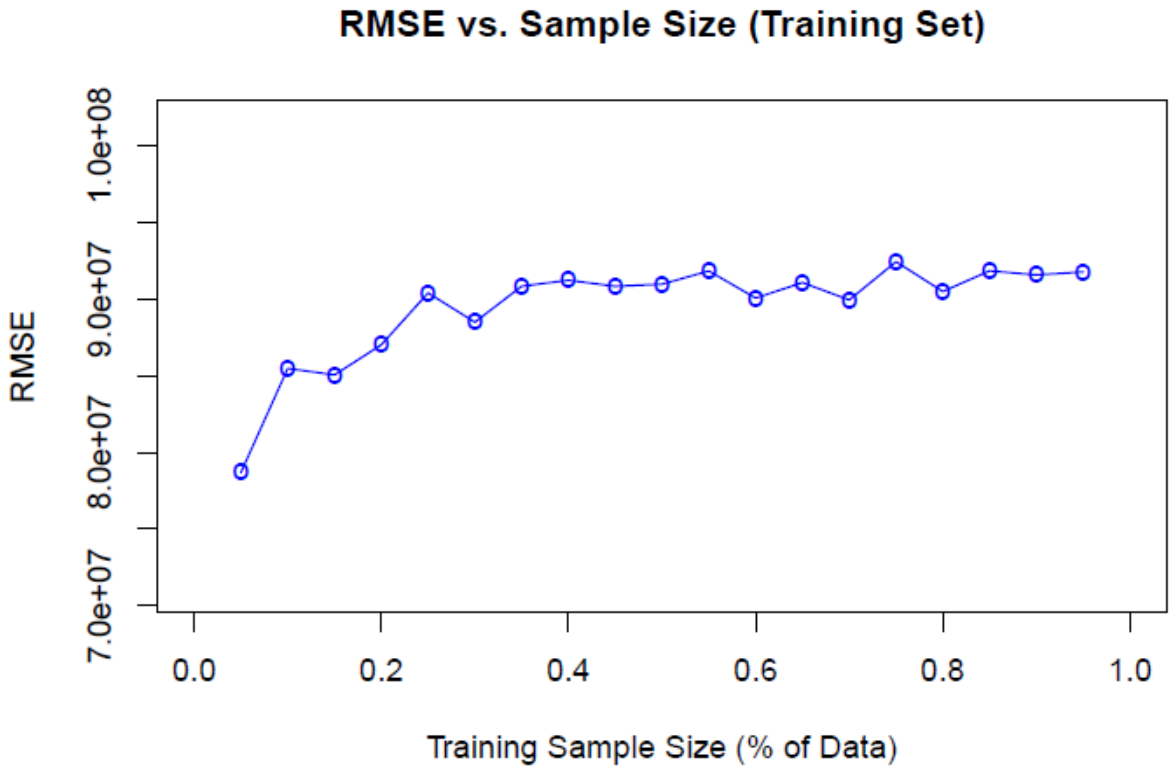

Figure 3. Model 2 RMSE vs. Training Sample Size for Training Set Predictions

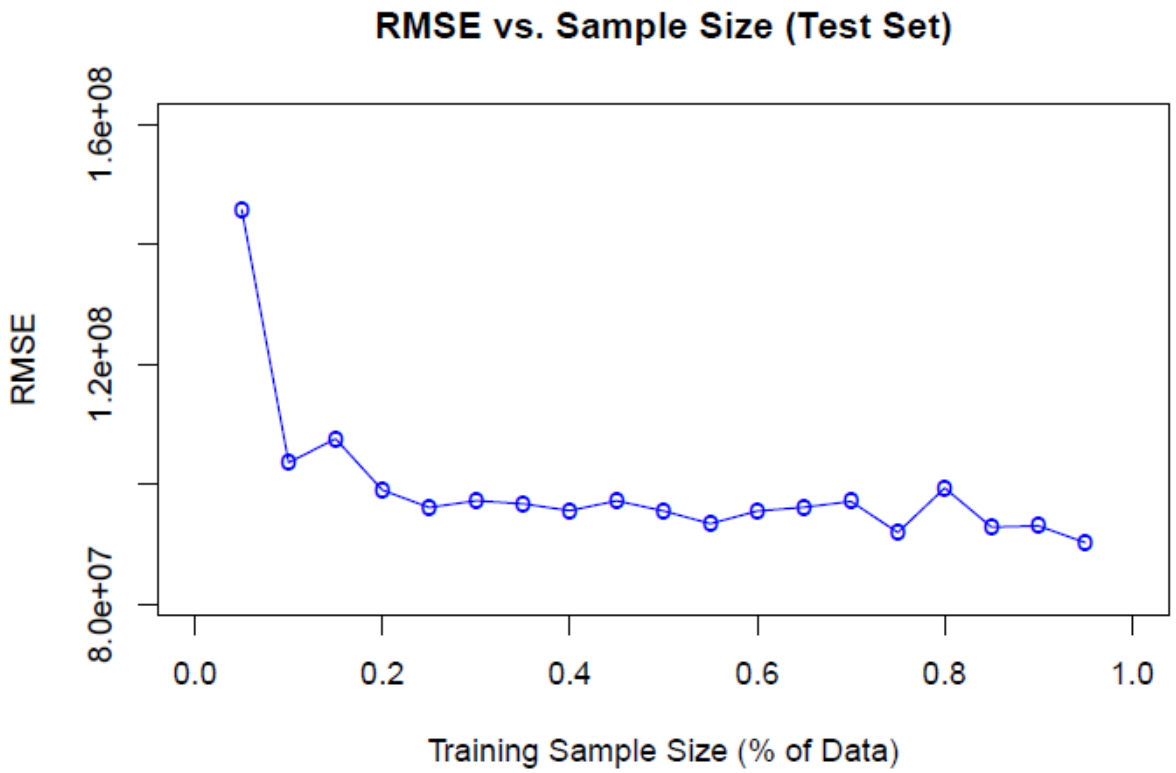

Figure 4. Model 2 RMSE vs. Training Sample Size for Test Set Predictions

For model 3, I use Awards, Rated, Genre, Language, Country, Director, Actor, and Production. I convert awards into number of nomination and wins as in project 1. For Rated and Genre, I use all the unique values and created 1 binary column for each value. For the other categorical variables, I use the top 10 most frequent categories for each variable to create binary columns. This is to keep the dimension of the model tractable because too many variables in the regression would need a lot more data than what we have to avoid the curse of dimensionality. Figure 5 and 6 show the RMSE versus training sample size for training and test set predictions. The best mean test RMSE is 134,526,546 at 95% training set size. This is higher than the RMSE in Task2, possibly because we are predicting a numerical variable Gross with only categorical variables.

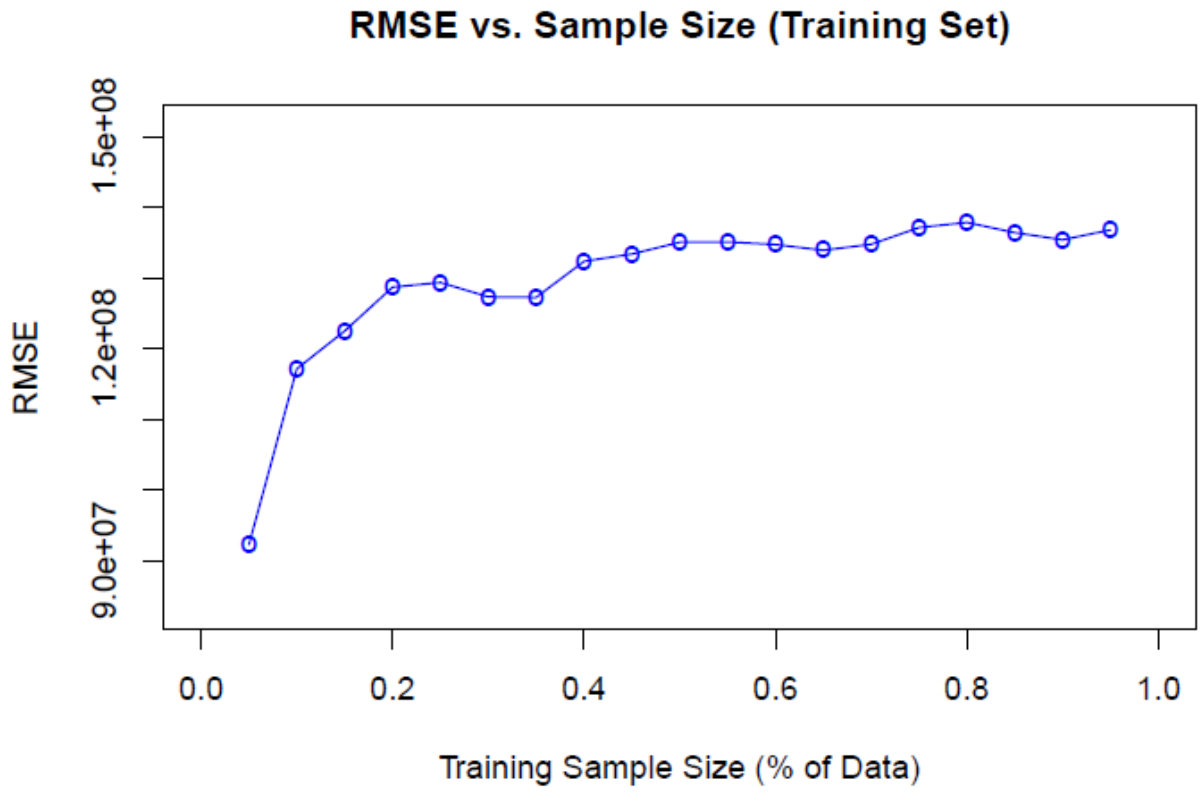

Figure 5. Model 3 RMSE vs. Training Sample Size for Training Set Predictions

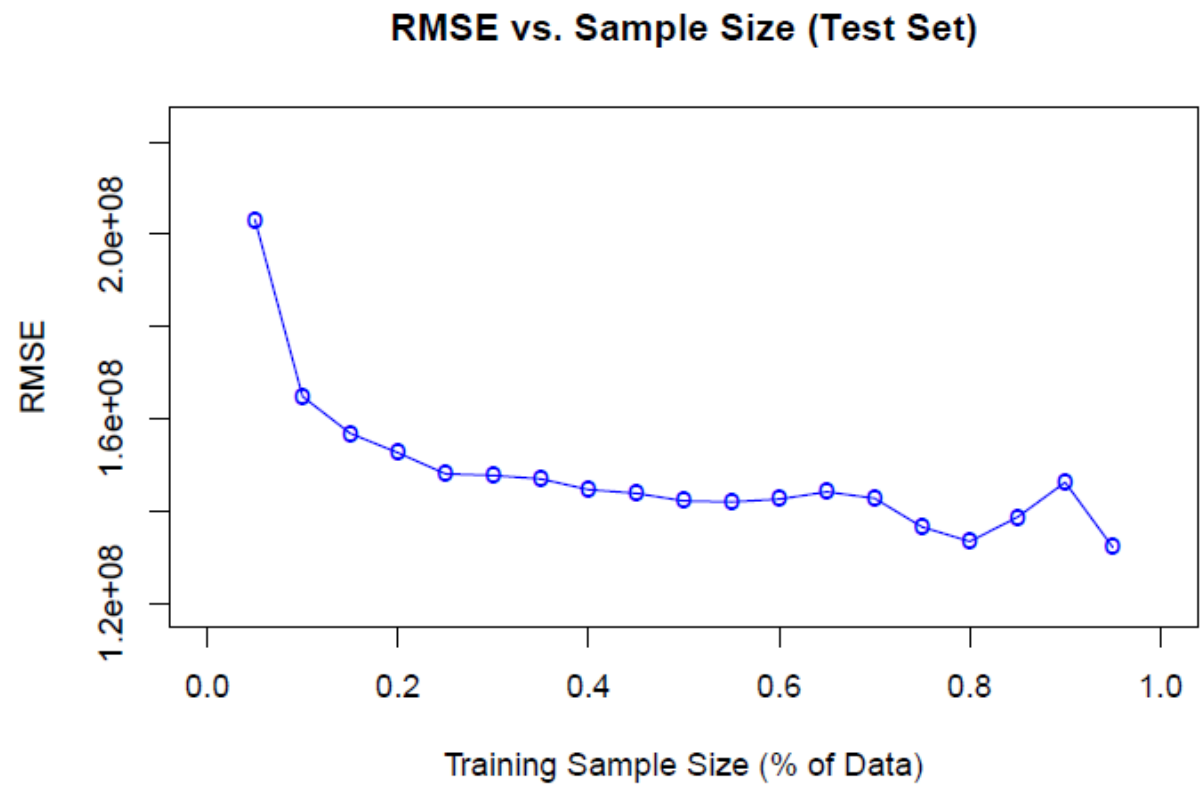

Figure 6. Model 3 RMSE vs. Training Sample Size for Test Set Predictions

For model 4, I include both numeric and categorical variables in model 2 and 3. Figure 7 shows the RMSE versus training sample size for training and test set predictions. In model 2, the best mean test RMSE is 92,742,444 at 95% training set size. In model 3, the best mean test RMSE is 134,526,546 at 95% training set size. For model 4, the best mean test RMSE is 91,747,668 at 90% training set size. This RMSE is lower than the ones in both model 2 and 3. The difference between model 3 and 4 is greater than the difference between model 2 and 4.

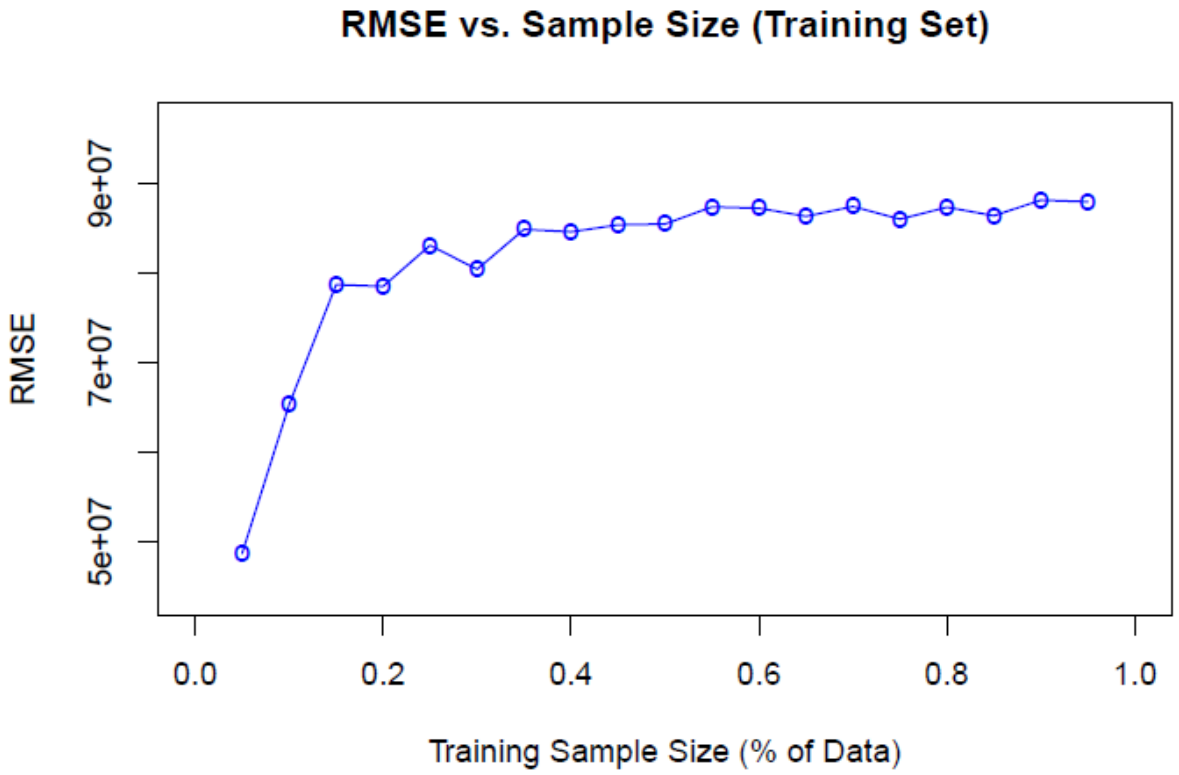

Figure 7. Model 4 RMSE vs. Training Sample Size for Training Set Predictions

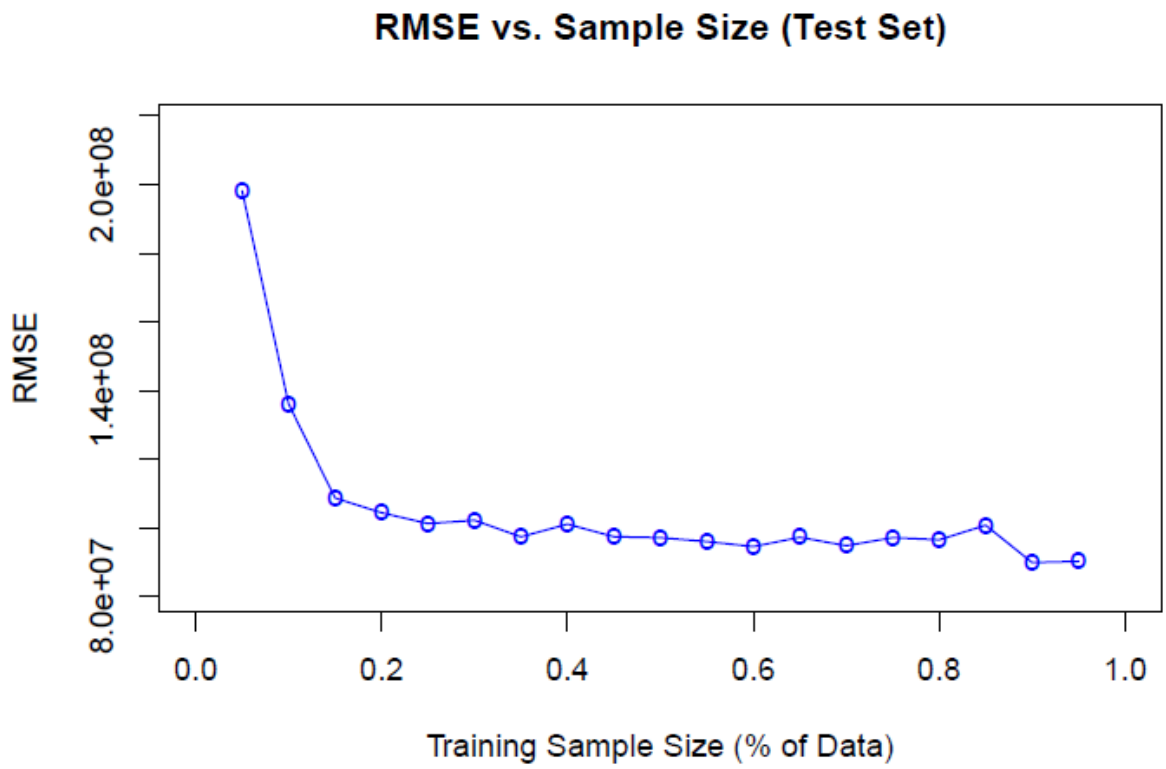

Figure 8. Model 4 RMSE vs. Training Sample Size for Test Set Predictions

For model 5, I include the interactions of Year and Budget as well as Month and Budget because it is likely that the impact of budget on gross revenue will vary by time. For example, $1 of budget may have a different impact on gross revenue in the early 2000's versus 2010's, or in the spring versus winter months. I include the length of title behind the idea that people may be more interested to see movies with more descriptive titles (i.e. Titles that convey more information about the movie). It turns out that title length is significant in predicting gross revenue with a p-value near 0. I also include number of genres because if a movie belongs to more genres, it would attract more audience. Other interactions include the ones between year, month and genre because, again, it is likely that the effect of genre will be different in different years and months. Figure 9 and 10 show the RMSE versus training sample for training and test set predictions. The final RMSE is 89,618,203 at 90% training sample size. Throughout the project I learned to use repeated estimations to obtain more stable results; how to create new variables from original variables using different methods such as creating dummy variables, binning, polynomial terms, log terms, interactions; and how to compare the results using different sample sizes for training.

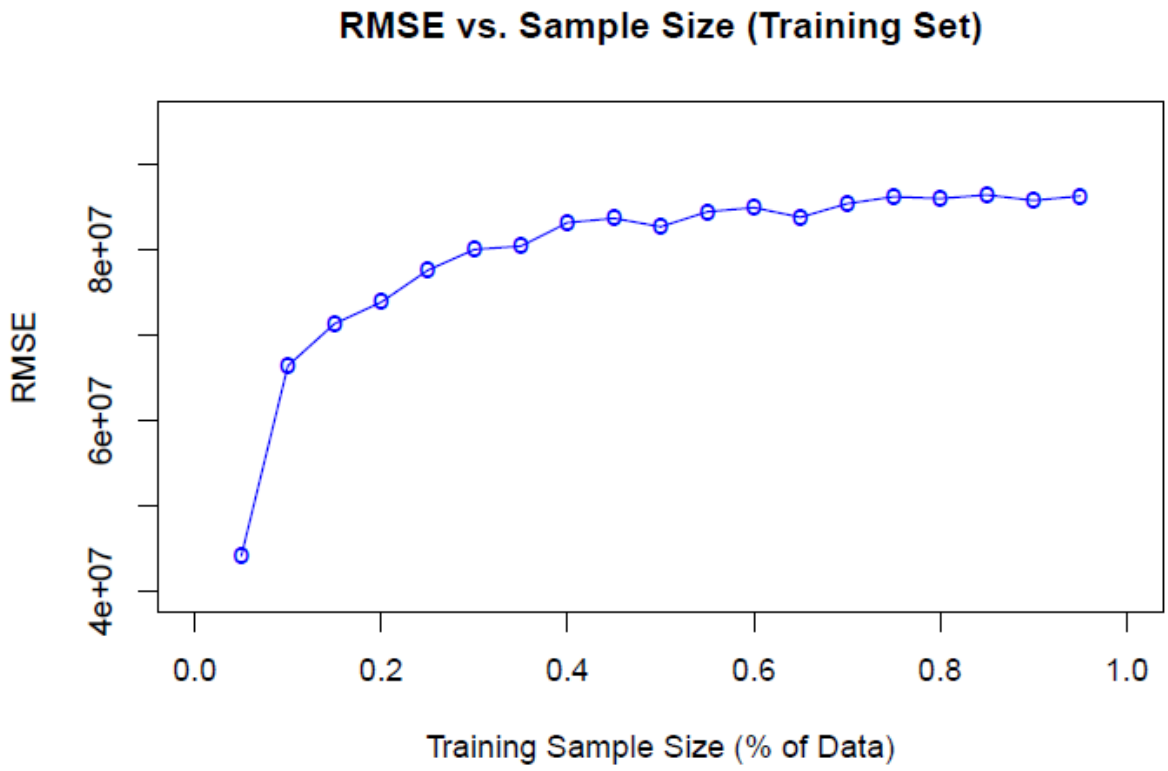

Figure 9. Model 5 RMSE vs. Training Sample Size for Training Set Predictions

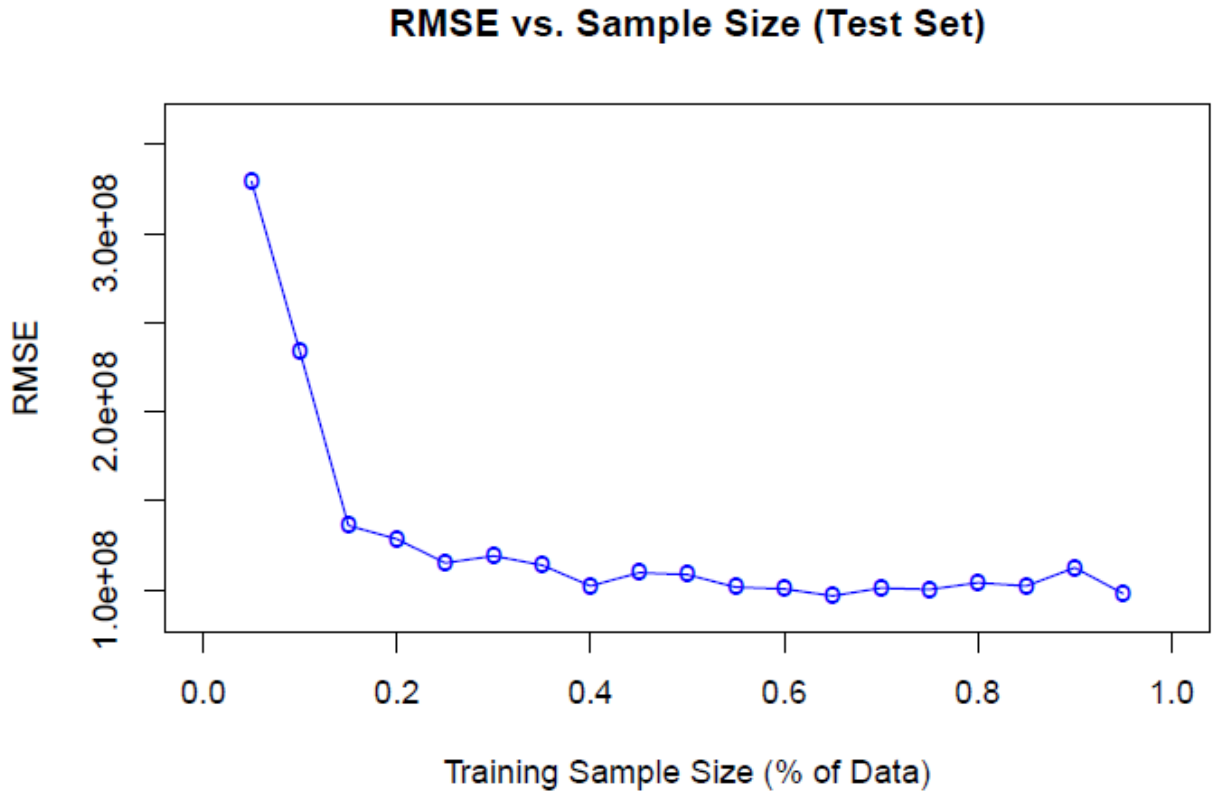

Figure 10. Model 5 RMSE vs. Training Sample Size for Test Set Predictions

Out of the 5 models, model 5 shows the lowest RMSE for test set. This is likely because it includes the most variables and therefore has the highest predictive power. However, the improvement over other models is not to a great extent. Model 2 and 4 with numeric variables also perform similarly. Model 3, on the other hand, has much higher RMSE because it only contains categorical variables, making it difficult to predict a continuous variable such as gross revenue in this case.

IV. CONCLUSION

This paper presents different least squares linear regression models for predicting gross revenue using feature variables that describe a movie's release timing, content, production team, language, etc. It seems that the models have reached a bottleneck of performance improvement. This task is to predict a continuous variable, gross revenue, so many classification models in machine learning is not suitable. However, future extensions can include more detailed features such as the plot, reputation and popularity of the actors, and history of directors. This will require more extensive gathering and processing of the data. For example, one can use natural language processing to extract features of the plot summary and include it as a feature to predict gross revenue.

REFERENCES

[1] Nelson RA, Glotfelty R. Movie stars and box office revenues: an empirical analysis. Journal of Cultural Economics. 2012;36(2):141-166.

[2] Einav Liran. Seasonality in the U.S. Motion Picture Industry. The RAND Journal of Economics. 2007;38(1):127-145.

[3] Elliot C, Simmon R. Determinants of UK Box Office Success: The Impact of Quality Signals. Review of Industrial Organization. 2008;33(2):93-111.

[4] Frank B. Optimal Timing of Movie Releases in Ancillary Markets: The Case of Video Releases. Journal of Cultural Economics. 1994;18(2):125-133.

[5] Chisholm DC, Norman G. Spatial competition and market share: an application to motion pictures. Journal of Cultural Economics. 2012;36(3):207-225.