Different machine-learning methods to predict popularity of movies using conventional and social media features

Paper Recurrence and Other Techniques

Chong Hu ch3467*, Wenjie Chen wc2685† and Haoran Zhang hz2619‡
Electrical Engineering
Columbia University
New York, USA
Email: {*ch3467, †wc2685, ‡hz2619}@columbia.edu

Abstract

This is the EECS 6690 final project of Statistical Learning. Our goal is to do a comprehensive review of a machine learning paper, try to reproduce the results of the original paper by recreating all the model using R code, think about new models or methods that can be applied to the dataset, and document the project and our results, including the comparison and reasons behind the results between ours and original paper's. The original paper [1] is about predicting the popularity of movies. A number of attributes such as cast, genre, budget, production house, and rating affect the popularity of a movie. Social media such as Twitter, YouTube etc. are major platforms where people can share their views about the movies. The original paper uses Linear Regression and J48 tree to do the prediction based on the above two kinds of features (conventional features and social media features). We reproduce the results and try other methods that also successfully predict the popularity of movies, including Support Vector Machine, Naive Bayesian Model, LDA & QDA, Artificial Neural Network, Random Forest. At last, we make a comparison between all of these methods, and discuss about the profound nature behind the results.

Index Terms

Movie popularity, Conventional feature, Social media feature, Linear Regression, J48 Tree, Support Vector Machine, Naive Bayesian Model, LDA & QDA, Artificial Neural Network, Random Forest

I. Introduction

Statistical learning has been of great interest to more and more people, especially to economists and financial decision makers. With large amount of data, smart data processing, and comprehensive prediction process, phenomenon can be understood deeply and used to infer future events. In the original paper, such studies have been performed in predicting the popularity of movies as well, where popularity is measure in terms of Ratings (represented by either a positive numeric number smaller than 10 or a label). Conventional features and social media features are two kinds of features that will be used to predict movie popularity.

Predicting the movie popularity is a hard thing, especially when the dataset is not big enough, because the number of movies cannot compete with other dataset. Using the dataset same as the original paper, we first implemented Linear Regression and J48 Decision Tree. Same result has been produced by us, but we can't get the accuracy as high as that of the original paper.

Thus, we start to analyze the difference, and try to use different methods to gain a better prediction, including Support Vector Machine, Naive Bayesian Model, LDA & QDA, Artificial Neural Network, Random Forest. Some of them can successfully predict movie popularity, and even performs better than the methods implemented by the original paper. At last, we do a comprehensive review of all the methods used in prediction and show which kind of features performs better in each specific method.

*†‡ Equal contribution

II. Literature Review

Nowadays an increasing number of researchers are investigating predictive models for movies performance. They have tried some methods using conventional features to do the prediction. However, most of the studies were using a few specific methods or showed conventional features performed better in prediction. In this section, we briefly describe what some of the researchers have done.

In the original paper, both conventional features and social media features are used separately to predict movie popularity. Conventional attributes are mostly collected from online movies databases, while social media attributes are collected from forums such as YouTube and Twitter. Anyone can review, rate, comment and share their opinions on a movie online. The original paper pays high attention to social media for predicting movie popularity.

The conventional features considered by the original paper are Genre, Budget, Number of Screens, Sequel, Gross Income, while the social media features collected from YouTube and Twitter are Aggregate Actor Followers, Number of views, Number of Likes, Number of dislikes, Number of comments, Sentiment Score. In the experiment, Linear Regression and J48 Decision Tree were performed. And the original paper found that sentiment score are the most discriminating feature, and some social media features such as Aggregate Actor Followers do have better performance than traditionally used Top Actor followers. The best accuracy gained by the paper is 76.2% for conventional features and 77% percent for social media features.

Before this original paper, we have also seen some studies reported in movie prediction based on the conventional features, such as budget, genre, but very few studies have did prediction based on social media features. And the studies based on the social media features reported that it didn't perform well. For example, one study was conducted on a dataset from 1998 to 2002, in which the authors uses K-means clustering, Polynomial and Linear Regression to do the prediction on conventional features, and achieved accuracy of 36.9% [2] [3]. And another researcher proposed that a higher accuracy of 70% can be achieved through Linear Regression, by integrating conventional features and social media features, and increasing the dataset [4]. An author used the dataset of 2009 and 2012 movies. They implemented sentiment analysis on social media features, including tweets from twitter, and reached accuracy of 64.4% [5].

In short, movies popularity has been predicted by many researchers using different methods and different features. We try to reproduce methods in the original papers and use some new methods to see their performance, and do a comprehensive analysis on the result.

III. Dataset

A number of attributes such as cast, genre, budget, production house, PG rating affect the popularity of a movie. Social media such as Twitter, YouTube etc. are major platforms where people can share their views about the movies. In order to predict the movies popularity, we are going to use the CSM (Conventional and Social Media Movies) Dataset 2014 and 2015 Dataset [1] in our project.

A. Description

The dataset retrieved information about movies from diverse sources including movies web site, i.e. IMDB, generic web resource i.e. Wikipedia, and social media including YouTube and Twitter. Beyond that, it also used sentiment analysis libraries to get the sentiment score for different movies. The total dataset contains twelve features and can be split in to two sub-dataset, the conventional features and social media features.

1) Conventional Features: Conventional Features contain six features in total and those features are typically available on movie resource websites, such as IMDB. Those features and corresponding data types are listed in table I.

TABLE I
CONVENTIONAL FEATURES DESCRIPTION

	Feature Name	Type
1.	Genre	Factor
2.	Budget	Numerical
3.	Number of Screens	Numerical
4.	Sequel	Factor
5.	Ratings	Numerical
6.	Gross Income	Numerical

- Genre: There are 19 different types of genre in the dataset, such as Action, Adventure and Drama etc. They were already mapped on to integer value from 1-19 and in our project, they are treated as factor variables to represent different genre.
- Sequel: This variable in integer represents whether the movie is sequel or individual. 1 shows that movie is first release; other n larger than 1 shows that movie is 2nd. e.g. Pirates of Caribbean: Dead Man's Chest is 2nd in sequel, therefore it is assigned the value of 2.
- Ratings: The value of Ratings ranges between 1 to 10 with 1 being lowest and 10 the highest. These values are collected from IMDB.
- Gross Income, Budget and Number of Screens: Gross world-wide income and Budget for each movie is collected from IMDB. The unit of gross income and budget is USD and they are already converted into USD if they are represented in other currencies. Number of screens on which movie was initially launched in US is also considered.

2) Social Media Features: Social Media Features also contains six features and those features are collected for each movie. Those features and corresponding data types are listed in table II.

TABLE II
SOCIAL MEDIA FEATURES DESCRIPTION

	Feature Name	Type
1.	Aggregate Actor Followers	Numerical
2.	Number of views	Numerical
3.	Number of likes of Screens	Numerical
4.	Number of dislikes	Numerical
5.	Number of comments	Numerical
6.	Sentiment Score	Numerical

- Aggregate Actor Followers: Number of followers of actors in one movie on twitter is used. Only the top 3 in cast are considered.
- Number of Views and Comments: Those variables represent the number of views and comments of trailer of movies on YouTube.
- Number of likes and dislikes: Number of Likes and Dislikes of trailers on YouTube are considered.
- Sentiment Score: A signed integer value is used to represent sentiment score. 0 represents neutral sentiment; "+"sign shows the positive sentiment and the value shows the magnitude; "–"sign shows negative sentiment and the value shows the magnitude. The sentiment score is calculated by Ahmed et al. [1] through analysing the sentiments of tweets about one movie.

B. Exploration

Before starting re-implementation of methods in the paper, we did a series of data exploration of the original dataset.

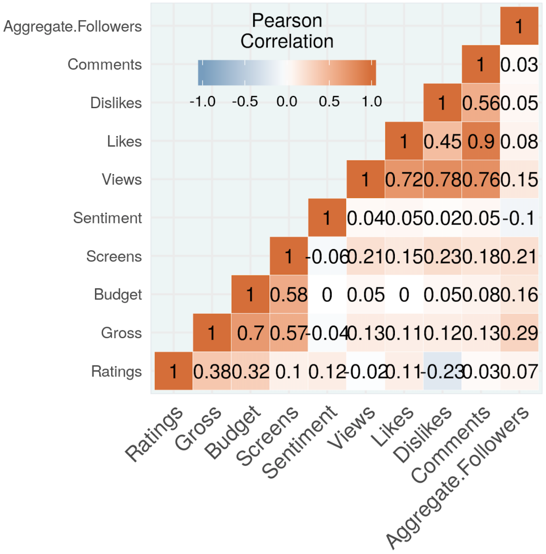


Fig. 1. Correlation Matrix on Numerical Data in CSM.

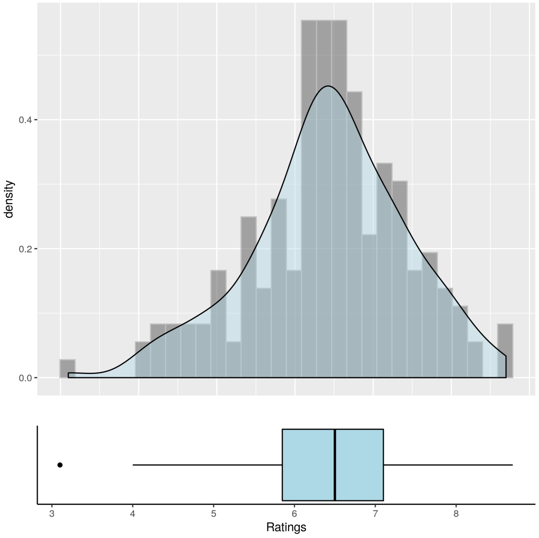


Fig. 2. Histogram and Box-plot of Rating.

From Figure 1, we plot the correlation matrix to show relationships between ratings and other variables and relationships among other variables. We use Pearson Correlation to evaluate the relationships. In the correlation matrix, we use red to show the positive relationship and blue to show the negative relationship. The color of one block also represents the magnitude of correlation. From the plot, we can see that there is no obvious direct relation between Ratings and other variables. But we draw some intuitive conclusion from this figure. The correlations among Likes, Dislikes and Comments are relative high since when people comment for one movie, they are also easily tend to evaluate the movie as well. Also Dislike has a slight negative relationship with ratings, which also match common sense that people won't give high rating for one movie if they dislike that movie. Since we are more interested in the dependent variable Rating, we plot the histogram and boxplot of Ratings. From those plots, we realize that, the method to map numerical variable Rating to four categories in original paper cannot be applied on this dataset. Since in the original paper, it maps Ratings to Excellent if rating is larger than 9.2. But in this dataset, the highest ratings is 8.9. If we use the original method to map rating to four categories, it would miss some labels and has very high bias on different categories. Therefore, we decide to re-design the method to map numerical Rating to category variables.

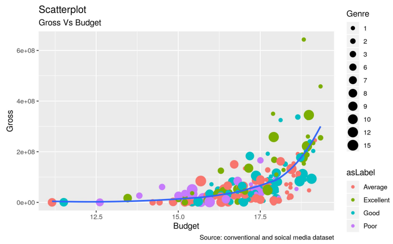
C. Data Preprocess

Table III shows that how we map the Rating to category variable. For methods exclude regression methods, we will use this categorical variable as dependent variable in our models and predictions.

TABLE III
MAPPING THE RATINGS OF MOVIES TO CATEGORICAL VARIABLE

	Ratings	Label
1.	0-5.2	Poor
2.	5.2-6.4	Average
3.	6.4-7.2	Good
4.	7.2-10	Excellent

After we re-designed the mapping method, we run a series of visualization to make sure our labels are appropriate in the later experiment.



Source: conventional and soical media dataset

Fig. 3. Relationships among Label, Genre, Gross and Income

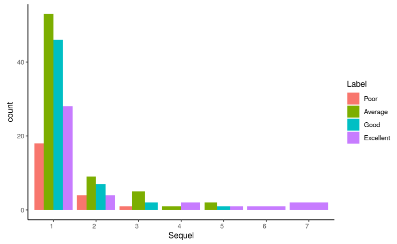


Fig. 4. Bar-plot for Sequel versus rating

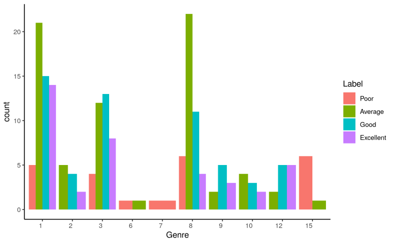


Fig. 5. Bar-plot for Genre versus Label.

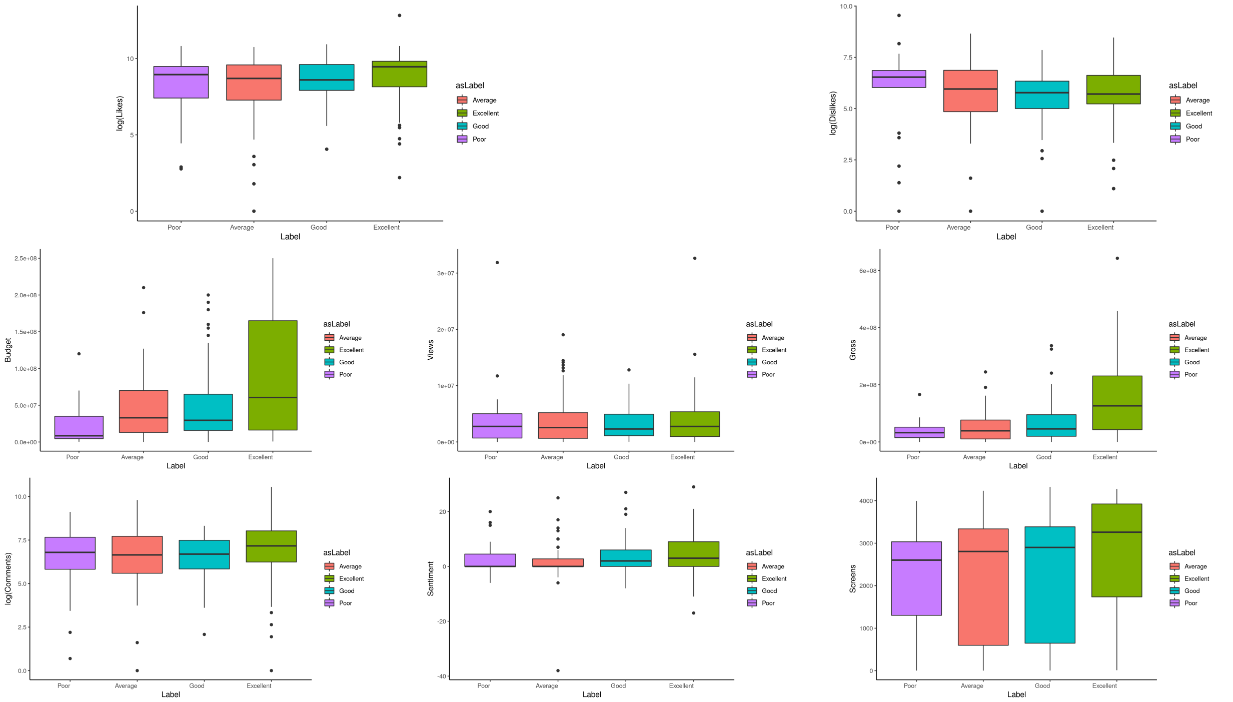


Fig. 6. Box-plots for Different Numerical Variables versus Ratings.

From those plots from those figures, we can learn that some features could be helpful to distinguish the levels of ratings, such as Gross, Income, the higher Gross and Income are higher, the higher probability that a movie might locate inside a higher level; for other features, we need to investigate how to make use of those features to predict the label of ratings for a movie. Also, to evaluate different methods, we split dataset into two parts, the train dataset and the test dataset. The ratio of two dataset is 8:2. For all methods mentioned in our paper, we get a model based on the train dataset and use the test dataset to evaluate the model performance.

IV. Original Paper Details and Reproduction

A. Linear Regression

In the dataset, Ratings are predicted using other attribute except Gross Income. Since Ratings are numeric, Linear Regression are suitable for this problem. We try normal linear regression, as shown by the original paper. But we found that the prediction performance is not good. Then we try to improve the method, by implementing Generalised Linear Regression and Generalised Addictive Model.

1) Normal Linear Regression: In this experiment, the model is $formula = Ratings\ ., data = conventional\_train.df$. After we build the model, we first take a look at the residual distribution. Residuals are estimates of experimental error obtained by subtracting the observed responses from the predicted responses, and can be thought of as elements of variation unexplained by the fitted model. The overall pattern of the residuals should be similar to the bell-shaped pattern observed when plotting a histogram of normally distributed data. In Figure. 7 and Figure. 8, both of them have four plots. In the subplots with the red line (which means residual equals 0), we can see that along the residual axis, the points are centered at red line, and becomes sparser outward. It scatters randomly about 0, regardless of the size of the fitted value. What's more, we also show a Q-Q plot, which provides an indication the distribution of the residuals. Most of the points are on the line, while few points are not at ends of the line. Thus, our training model is ready to predict the movie popularity, and we test our models.

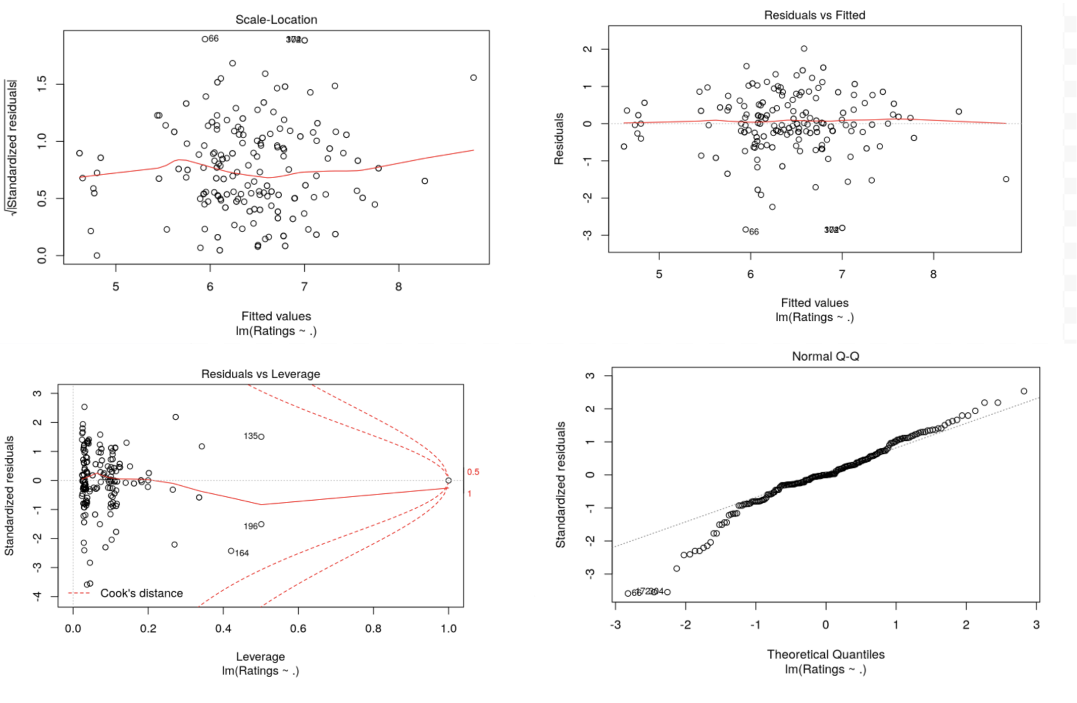


Fig. 7. Linear Regression Result on Conventional Features.

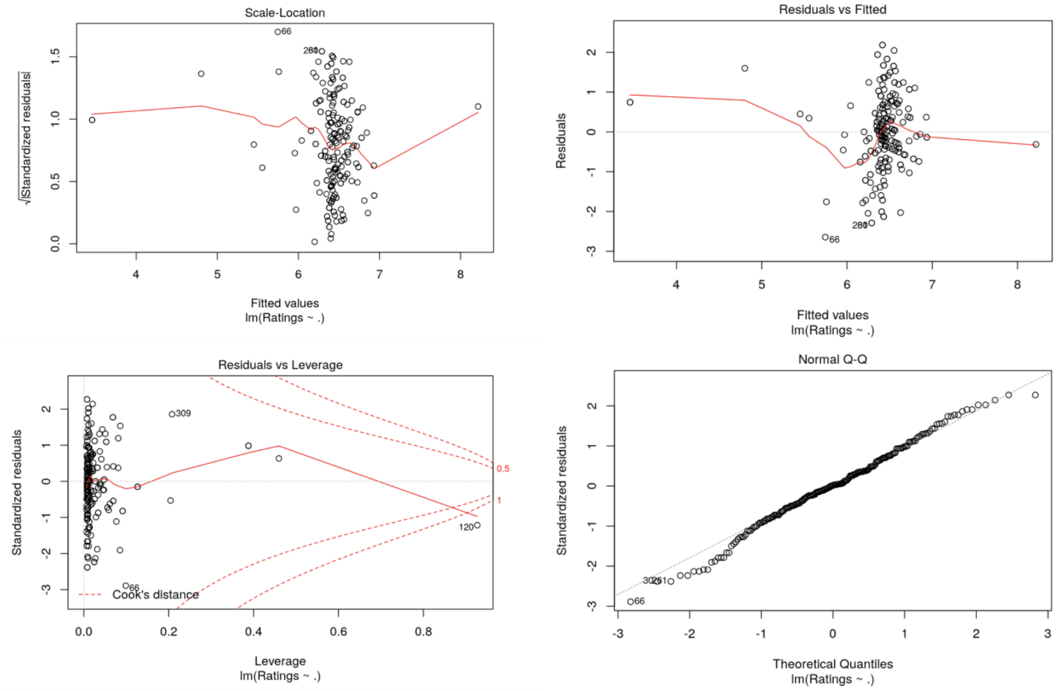


Fig. 8. Linear Regression Result on Social Media Features.

In the original paper, the Linear Regression Model gets 36.4% and 43% test accuracy respectively using conventional features and social media features. In our implementation, we get 35.71% and 44.64% test accuracy respectively using conventional features and social media features. The result is close.

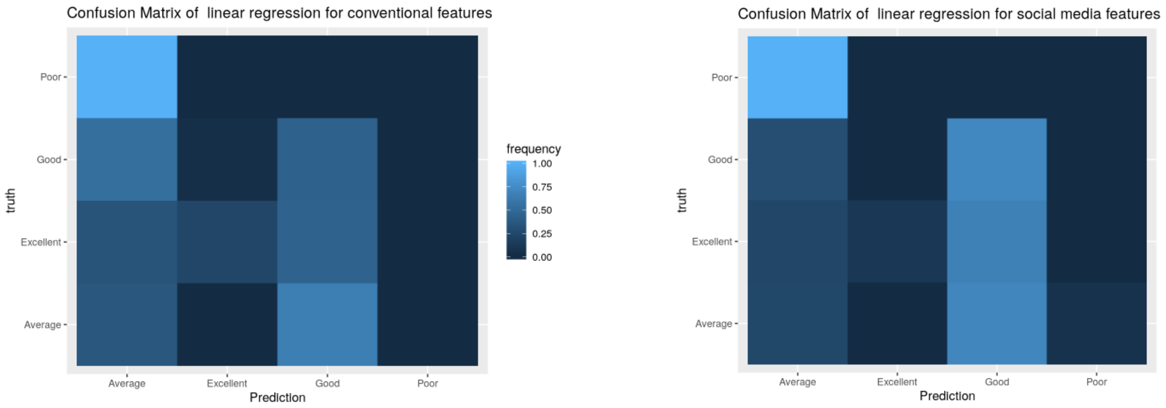


Fig. 9. Confusion Matrix for Linear Regression Using Conventional Features (Left) and Social Media Features (Right).

From Figure 9, we use confusion matrix to visualize our prediction. The horizontal axis is the prediction and the vertical axis is the truth. The brightness represents the frequency of the hit. We can see that the test model is not good. So we try to improve linear regression model.

2) Generalized Linear Regression: Generalized Linear Regression is a flexible generalization of normal linear regression that allows for response variable which have error distribution models other than a normal distribution. It uses a link function and allows the magnitude of the variance of each measurement to be a function of its predicted value. Here we choose log as the link function, and then do the prediction below.

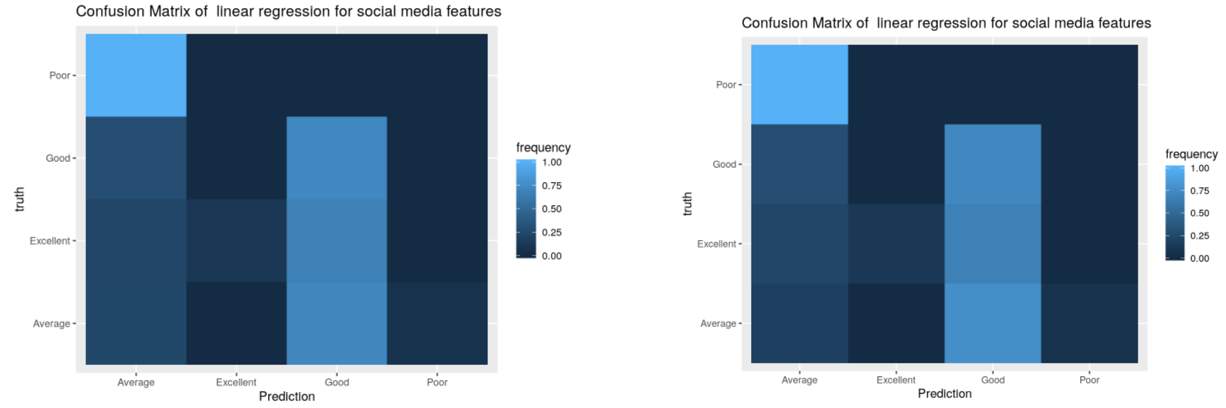


Fig. 10. Confusion Matrix for GLM Using Conventional Features (Left) and Social Media Features (Right).

As shown in the confusion matrix in Figure 10, our accuracy isn't improved too much. Only the third box along the diagonal from down to top becomes a little brighter, while others are still dim. Therefore, Generalized Linear Regression fail.

3) Generalized Addictive Model: Generalized Addictive Model is to blend properties of generalized linear models with addictive models. We use the inverse link function here and test the model.

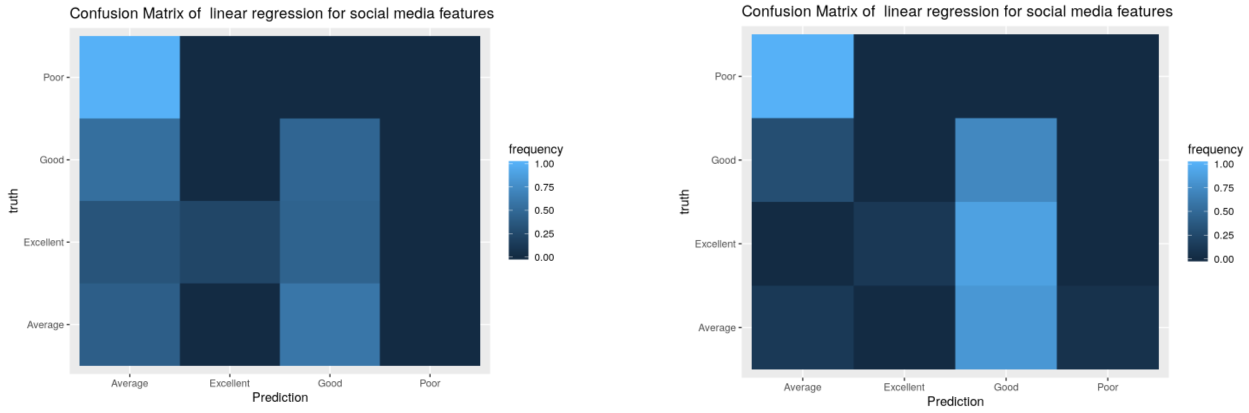


Fig. 11. Confusion Matrix for GAM Using Conventional Features (Left) and Social Media Features (Right).

We see that some difference occurs during our prediction, but the accuracy of prediction isn't improved. The accuracy using Linear Regression is no more than 50%, which is the same as that of the original paper. And we tried to improve the model, which didn't work.

Thus, we think Linear Regression Model is too simple to predict movie popularity. Although suggestion of movie popularity can still be given by Linear Regression Model, the accuracy is not good enough.

B. J48 Tree

In this experiment, the original paper converted the values of Rating into four bands. As mentioned before, we have made some changes to a more reasonable one. The original paper implemented J48 Decision Tree as the classification model. J48 Decision Tree is based on C4.5 classification algorithm, using the concept of information entropy. Every node is split depending on the normalized information gain (the highest gain is chosen to make decisions). We build two train model and visualize it below using conventional features and social media features separately.

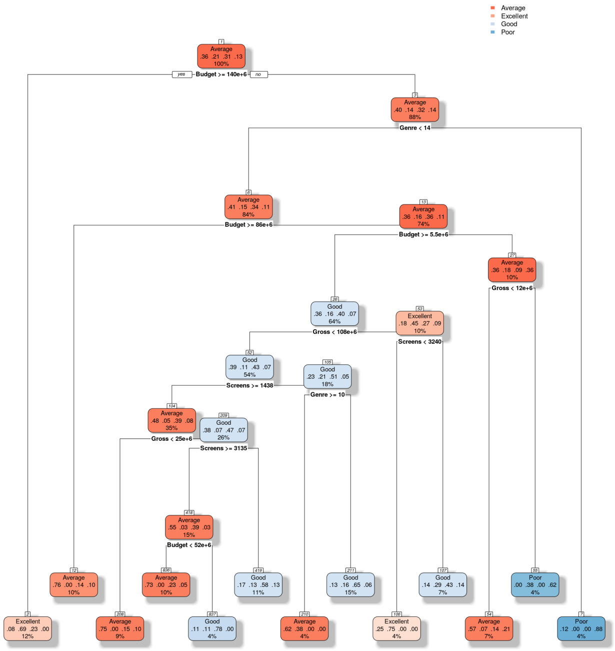


Fig. 12. J48 Tree Structure Using Conventional Features.

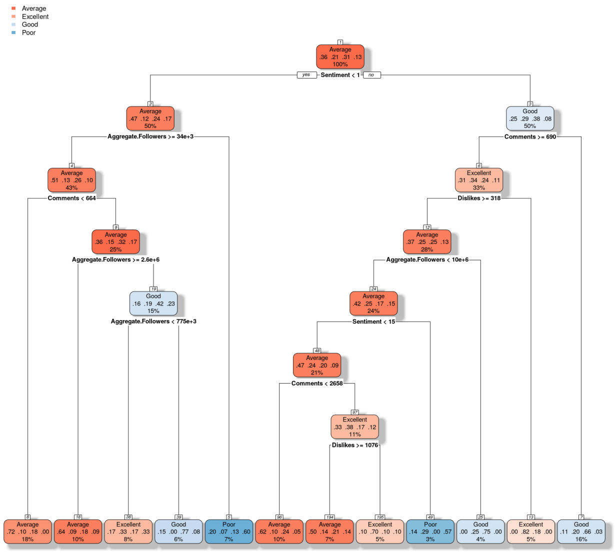


Fig. 13. J48 Tree Structure Using Social Media Features.

As shown in Figure 12 and Figure 13, the classification is easy to understand and analysis. We can modify and prune the tree depending on the figure. After the modification, we test our model.

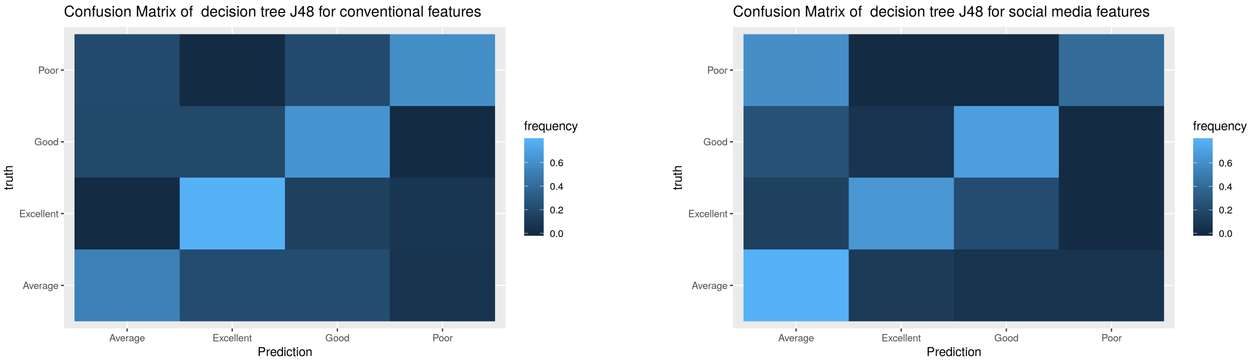


Fig. 14. Confusion Matrix for J48 Tree Using Conventional Features (Left) and Social Media Features (Right).

In the original paper, prediction based on social media features performs better than that based on conventional features. And they have 47% for conventional features and 77% for social media feature. We have the same result, but our accuracy is lower. After many times of trying and trial, our accuracy is 62.96% and 68.52% respectively. Model using social media features performs better than that using conventional features.

According to the original paper, among social media features Sentiment was the best while Likes/Dislikes were the worst. Among conventional features, No of Screens, Genre and Budget were the best three features. In our reproduction, sentiment indeed is the best among the social media features. Budget and Genre are the best among the conventional features, while Gross is the third best.

Since we have tried every way to improve the performance of our J48 Tree Model, we think the reason why our accuracy still can't reach 80% is that the original paper might use more internal data.

In summary, we have finished reproducing the original paper, and have tried some improvement. The results are generally consistent. In the next part, we try other methods to get better prediction.

V. Other methods Beyond the Original Paper

Beyond Linear Regression and Decision Tree(J48) implemented in the original paper, we also tried other machine learning methods such as Support Vector Machine, Naive Bayesian Classifier, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Artificial Neural Network and Random Forest. Some methods do not perform very well in our project. We will discuss reasons behind their poor performance. While, some methods like Neural Network and Random Forest have an excellent performance compared with others. In this part, we will elaborate on details about implementing these methods.

A. *Support Vector Machine*

Originally, Support Vector Machine were intended for the binary classification setting. With a further extension of the classifier to accommodate more classes and can be applied into more features. General Support Vector Machine is a discriminating classifier by means of using hyperplane to separating target classification space.

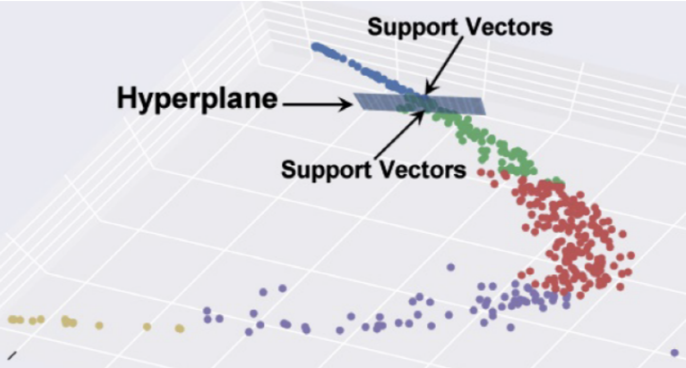


Fig. 15. SVM structure.

In our project, we have two target classification space, one is 5-dimensional target classification space (conventional features) the other one is 6-dimensional target classification space. Accordingly, we need to use 4-dimensional hyperplane and 5-dimensional hyperplane to separate target classification space. We adopt one-vs-all classification when we use SVM in our project, because we need to do 4 binary classification among total 4 classes.

Intuitively speaking, a good classification has the best hyperplanes which can have the longest distance to nearest data points in training data-set, which is also known as functional margin. while the larger the functional margin is, the better model classify the target space, and the higher the accuracy is.

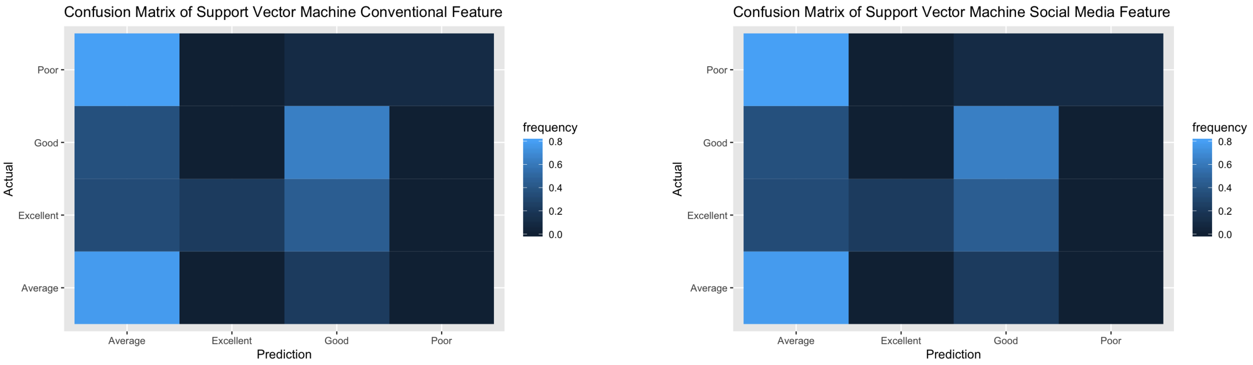


Fig. 16. Confusion Matrix for SVM Using Conventional Features (Left) and Social Media Features (Right).

As we can see from Figure 16. For conventional features, SVM can achieve accuracy of 44%, and for social media feature SVM can achieve accuracy of 38%. Specifically, we do kernel tricks on SVM classifier. For Conventional Features, we pick radial kernel, which can achieve the highest accuracy using SVM. For Social Media Features, we pick polynomial kernel and tune parameters by setting `gamma = 1`, `coefficient = 7` and `degree = 3`.

Generally speaking, SVM method is not ideal to our dataset, even if we try to tune the parameter to get the best performance. The reason why our accuracy is rather poor is that we use SVM to deal with multi-class problem in high dimension, while SVM method is designed to binary classification.

B. *Naive Bayesian Model*

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). The prerequisite of Naive Bayesian Classifier is that the probability of predictor (x) on a given class (c) is independent of the values of other predictors. Naive Bayesian Classification Models works well on models that follow class conditional independence assumption.

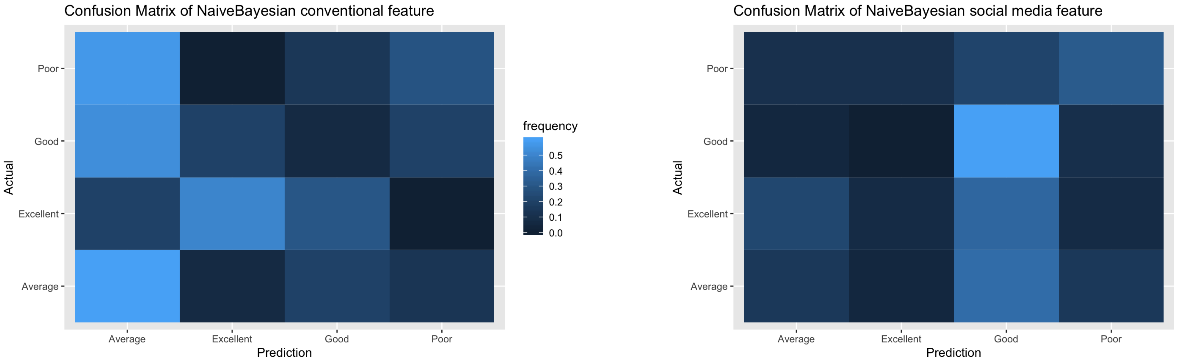


Fig. 17. Confusion Matrix for Naive Bayesian Using Conventional Features (Left) and Social Media Features (Right).

From Figure 17 above we find that using Naive Bayesian Models, for conventional features it could only achieve accuracy of 36%, while for social media features it can achieve accuracy of 40%. As we can see from confusion matrix above, using conventional features, the model always predicts target as 'Average' type. Using social media features, the model always predicts the target as 'Good' type.

The poor performance of Naive Bayesian is within our expectation. Because Naive Bayesian Method only works well when the dataset follow Bayesian Assumption and it is obvious that both conventional and social media features do not follow Bayesian Assumptions.

C. LDA & QDA

Both LDA and QDA are two classic classifiers that, as we can see from their names, a linear classification surface and a quadratic classification surface respectively. They are based on Bayes' Theorem with assumption on conditional Multivariate Normal Distribution.

LDA is an indirect approach to model the predicted probabilities given the predictors. This approach separately find predictors' probabilities distribution in different target class, then use Bayes's theorem to reflect possible estimate.

QDA estimates a separate convariance matrix for each target classes, the number of predictors are very large. Compared with LDA, QDA is better if our dataset has a very large training set, which means the variance of the classifier is not a major concern.

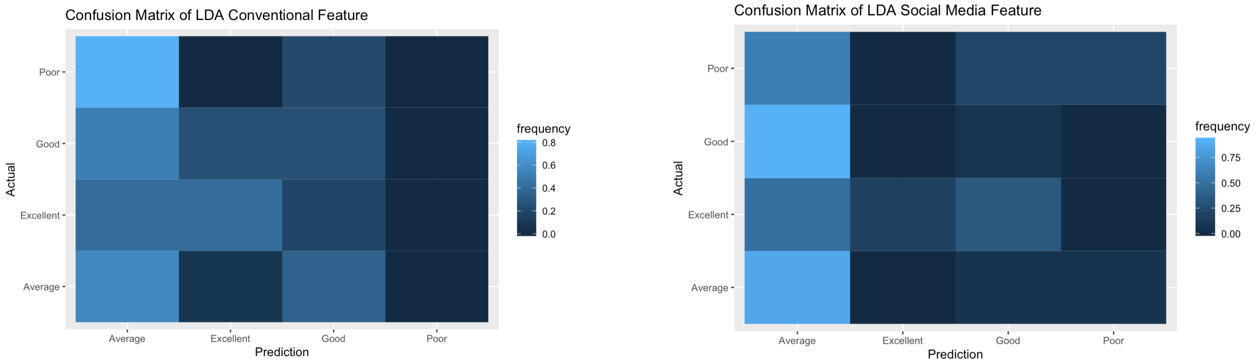
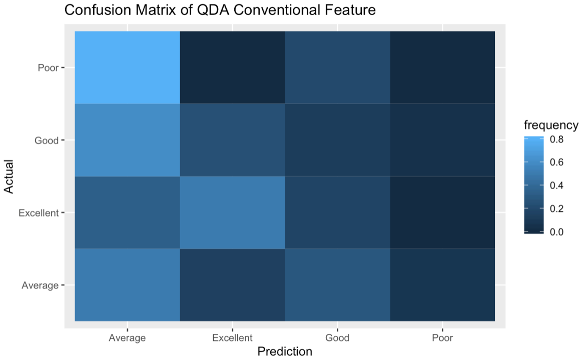


Fig. 18. Confusion Matrix for LDA Using Conventional Features and Social Media Features.



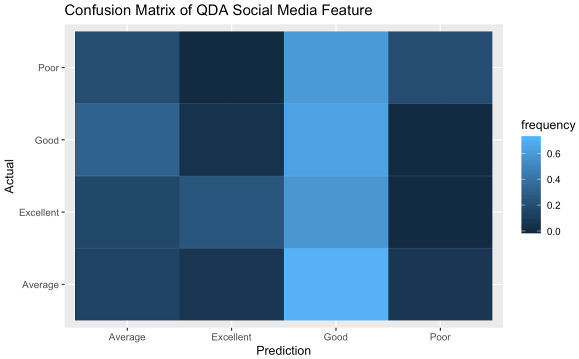


Fig. 19. Confusion Matrix for QDA Using Conventional Features (Left) Social Media Features (Right).

As we can see in Figure 18 and Figure 19, both LDA and QDA have rather poor performance in our experiment. LDA has accuracy of 30% using conventional features and accuracy of 39% using social media features. QDA has accuracy of 34% using conventional features and accuracy of 29% using social media features. Both LDA and QDA works well on dataset which follow multivariate normal distribution. Specifically, LDA needs the convariance matrix are all the same for all classes and QDA needs the convariance matrix are not the same for all classes. It is obvious that our dataset do not follow assumptions and properties,making the accuracy rather poor.

D. Artificial Neural Network

Artificial Neural Network "learn" to perform tasks without task-specific rules. Since we don't know the exact relationship between all the features and results, Artificial Neural Network might be a good choice. ANN is based on each connected node which is like neurons, and transform signal at each layer.

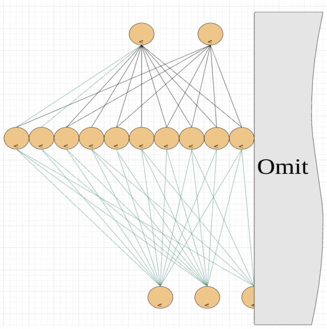


Fig. 20. ANN Models Structure.

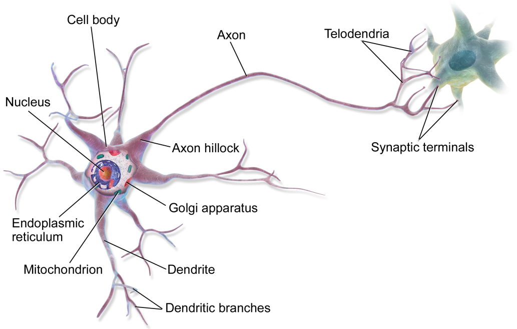


Fig. 21. Neuron Structure.

In our task, we have four units at output layer as the four label- "Poor", "Average", "Good", "Excellent", twenty units at hidden layer. At input layer, we have five and six units respectively for conventional features and social media features, as shown in Figure 20.

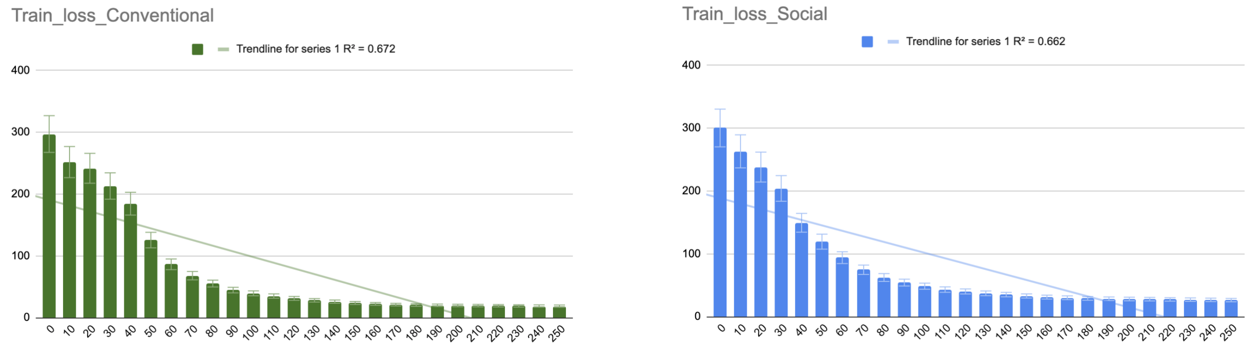


Fig. 22. Train Loss of ANN for Conventional Features (Left) Social Media Features (Right).

After training the model (process shown in Figure 22), we get all the probabilities for the four labels. We pick the label with the highest probabilities as the prediction. Then we test our ANN model.

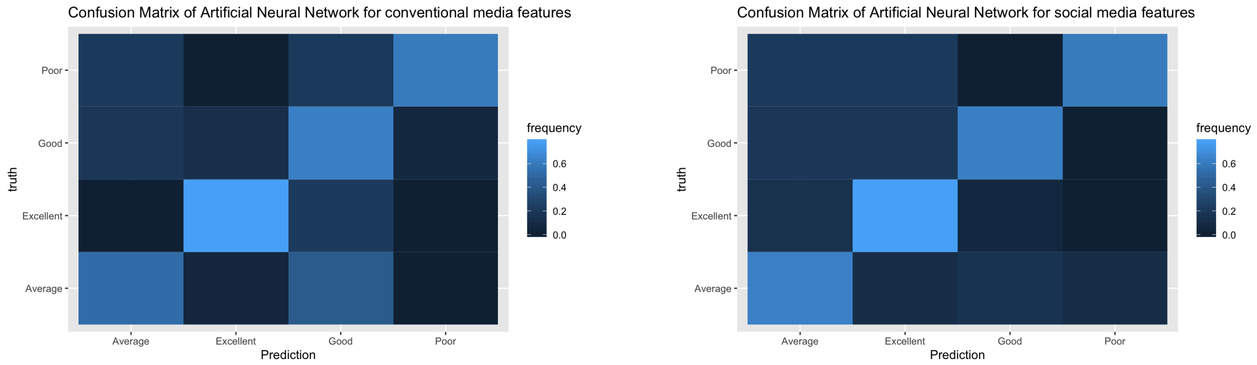


Fig. 23. Confusion Matrix for ANN Using Conventional Features (Left) and Social Media Features (Right).

We run evaluations and get the confusion matrix as shown above. The test accuracy is 62.96% and 66.67%, for conventional features and social media features respectively. And the test accuracy even reaches 74.07% using all the features. The accuracy is good enough to reach our expectation. Same as the previous experiment, the model using social media features performs better than that using conventional features.

E. Random Forest

For the Random Forest method, we set `replace = TRUE`, `importance = TRUE`, `ntree = 20`, which means we are using data with replacement, assessing importance of predictors during training and the numbers of sub-trees in the forest is 20. The metric to build a tree is based on Gini Index. Figure 24 is showing the training error for those 20 sub-trees. The black line represents the average error and the other four colorful lines represent error for four different labels. We can see that different trees get different performance on the train dataset which contributes to the robustness of the overall random forest performance.

After we get random forest model for those two sub-dataset, we run evaluations on test dataset and plot the confusion matrix in Figure 25. The test accuracy for Conventional Features and Social Media Features are 62.5% and 69.3% respectively. Also in the confusion matrix plots, the diagonal of matrix is obviously lighter than the other blocks. Only to the matrix for Conventional Features, there are mistakes between Average and Poor.

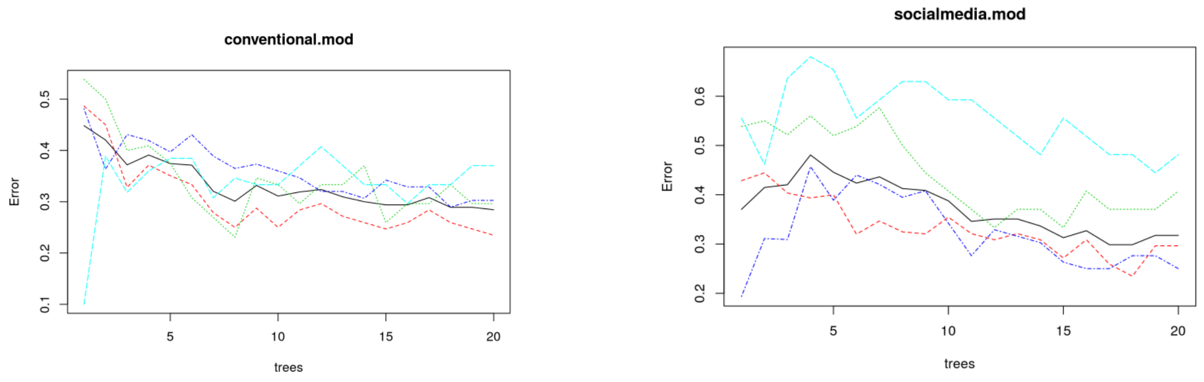


Fig. 24. Training Error of Random Forest for 20 subtrees in Conventional Features (Left) and Social Media Features (Right).

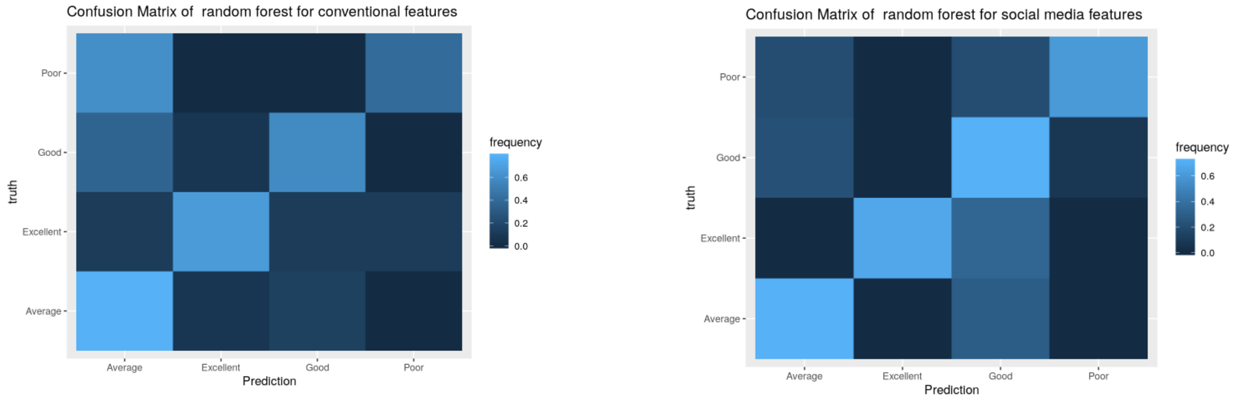


Fig. 25. Confusion Matrix for Random Forest Using Conventional Features (Left) and Social Media Features (Right).

The Random Forest method gets a good performance on this dataset. The reason why random forest works good on the dataset is because it uses a multiple-trees strategy so that the robustness is improved. A single tree may not work well on the dataset, but a combination of multiple trees could summary those sub-trees and get a better performance. Also, generating tree based on different features help the robustness of the model. Random forest method is good at extracting information from high dimension data, like this dataset with multiple features.

VI. Conclusion and Discussion

Our paper presents a comparison of different methods to predict popularity of movies based on Conventional Features and Social Media Features. Table IV shows different method's performance on those two sub-dataset and Figure 26 gives a bar-plot to visualize the comparison.

From the plot, we can see that J48, Random Forest and ANN can get a better performance than other methods. These three methods could make full use of high dimensional data and extract useful information from the data. For the other traditional methods, we can see they can not get a good performance on this data-set. They are too simple to solve high dimensional data. For linear regression, LDA & QDA and Naive Bayesian, they require a very strict normal assumption to guarantee the model performance. Linear Regression needs the residuals follow normal distribution centering at zero; QDA and Naive Bayesian need the data itself follow normal distribution; LDA even needs an equal variance assumption over it. This dataset doesn't have such kind of properties and therefore their performance is not satisfying. SVM method is designed to do binary classification, while our project is multi-class problem in high dimension.

TABLE IV
EVALUATIONS FOR DIFFERENT METHODS ON CONVENTIONAL FEATURES AND SOCIAL MEDIA FEATURES

Methods	Conventional Features	Social Media Features
Linear Regression	0.358	0.446
Generalized Additive Model	0.411	0.393
Generalised Linear Regression	0.429	0.357
LDA	0.339	0.303
QDA	0.285	0.393
Naive Bayesian	0.362	0.404
Support Vector Machine	0.436	0.385
ANN	0.630	0.667
Random Forest	0.625	0.693
J48	0.629	0.685

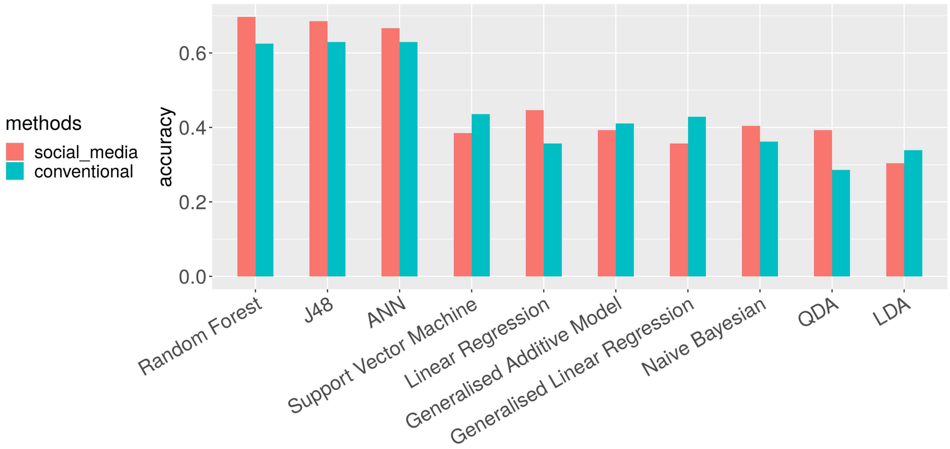


Fig. 26. Evaluations for Different Methods on Conventional Features and Social Media Features.

For GLM and GAM those models don't require very strict assumptions. However, they are not good at solving such complicate data witch such high dimension. Optimal solutions for GLM and GAM even could not fit data very well; complicate kernels in SVM are also not enough to find a good solution.

However, J48, Random Forest and ANN could provide satisfying result. J48 is an improved version of decision tree, which is enough to fit training data well. With pruned method and other constraints, the model also offers a good performance. The reason why random forest works good on the dataset is because it uses a multiple-trees strategy so that the robustness is improved. Generating tree based on different features help the robustness of the model and high performance on test data. ANN could construct a neural network to fit data which provide lower error rate on training data. The regularization and setting appropriate hyper parameter also improve the robustness of the model.

Furthermore, we can draw a conclusion that different models can get similar performance on those two sub-dataset. The difference of two accuracy provided by one specific method is often within 10%. Also, one method has better performance then another method on one sub-dataset doesn't guarantee it also wins on the other sub-dataset. If we focus on the three methods with best accuracy, we can find that social media features can provide a higher accuracy than conventional features. In our experiment, when we use social media feature with Random Forest method, we can achieve the highest accuracy 69.3%.

VII. Future Work

In this paper, we have already applied different methods and tested corresponding performances on this dataset. However, we can use other methods to predict the rating of a movie. Also, in this paper, we only use Ratings as dependent variable. In the future work, we investigate the relationship among other variables. Furthermore, the dataset we used is still not large enough to get a very robust result. We can increase the dataset size by collecting more data from IMDB, YouTube, Twitter, etc. In this way, the result will become more robust and we can reach a more general conclusion.

Acknowledgment

First, we would like to thank Professor Predrag Jelenkovic, and all TAs, including Yue Feng, Aijia Gao, Yutao Tang, Zhen Wang, Xiaosheng Zhu, and so on. Without you, we can never finish our final project from scratch. Thank you for the instructions and help of a whole semester!

Second, we would like to thank all the authors of the original paper and all the contributors of the datasets. They have done a great work before us.

References

[1] Ahmed M, Jahangir M, Afzal H, Majeed A, Siddiqi I. Using Crowd-source based features from social media and Conventional features to predict the movies popularity. InSmart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on 2015 Dec 19 (pp. 273-278). IEEE.

[2] Sharda, R., Delen, D. (2006): ”Predicting box-office success of motion pictures with neural networks”. Expert Systems with Applications, 30(2), 243-254.

[3] Nikhil Apte, Mats Forssell, and A. Sidhwa, ”Predicting Movie Revenue”. 2011.

[4] Bhave, A., Kulkarni, H., Biramane, V., Kosamkar, P. (2015). ”Role of different factors in predicting movie success”. In Pervasive Computing (ICPC), 2015 International Conference on (pp. 1-4). IEEE.

[5] Jain, Vasu. ”Prediction of Movie Success using Sentiment Analysis of Tweets.” The International Journal of Soft Computing and Software Engineering3.3 (2013): 308-313.

[6] Asur, Sitaram, and Bernardo Huberman. ”Predicting the future with social media.” Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2010.

[7] Rui, Huaxia, Yizao Liu, and Andrew Whinston. ”Whose and what chatter matters? The effect of tweets on movie sales.” Decision Support Systems 55.4 (2013): 863-870.

[8] Apala, Krushikanth R., et al. ”Prediction of movies box office performance using social media.” Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013.