Using Crowded-source based feature from social media and Conventional features to predict the movies popularity

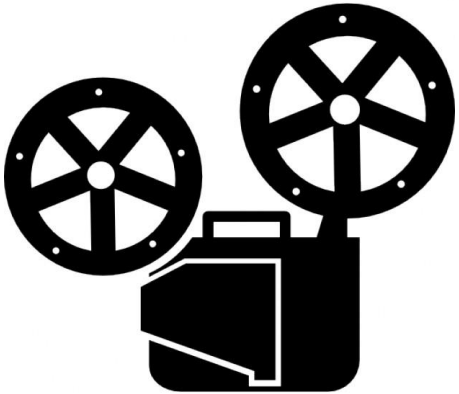
Paper reproduce and other attempts

Chong Hu(ch3467)
Wenjie Chen(wc2685)
Haoran Zhang(hz2619)

Dataset description

Movies Dataset

Dataset Description



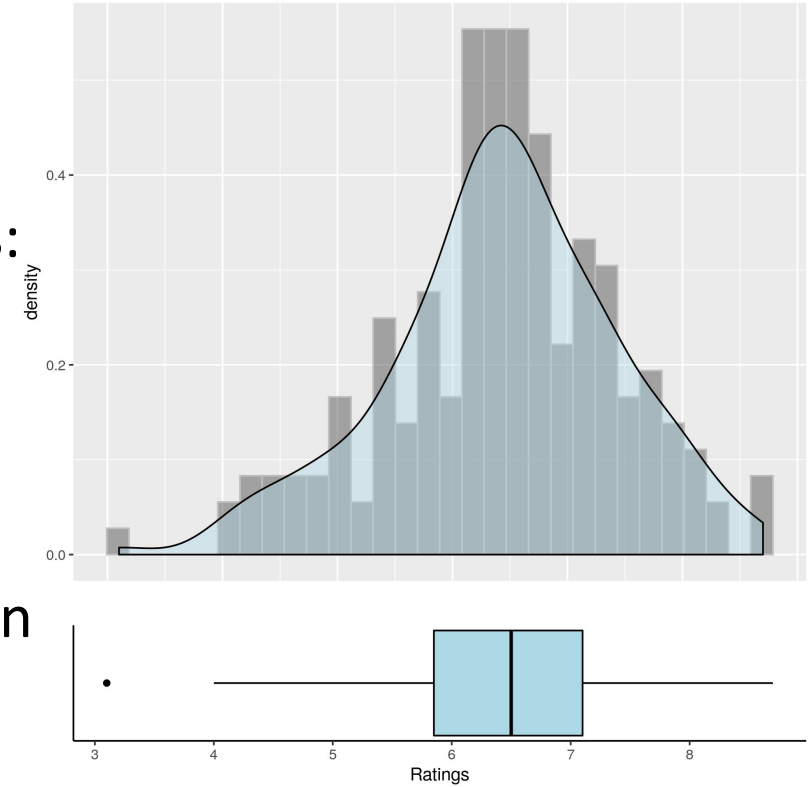
Predicting the success of movies has been of interest to economists and investors as well as predictive analysts. A number of attributes such as **cast, genre, budget, production house, PG rating** affect the popularity of a movie. Social media such as Twitter, YouTube etc. are major platforms where people can share their **views about the movies**. In this project, we collect all these features to make a prediction of movies' ratings using machine learning techniques.

Dataset Description

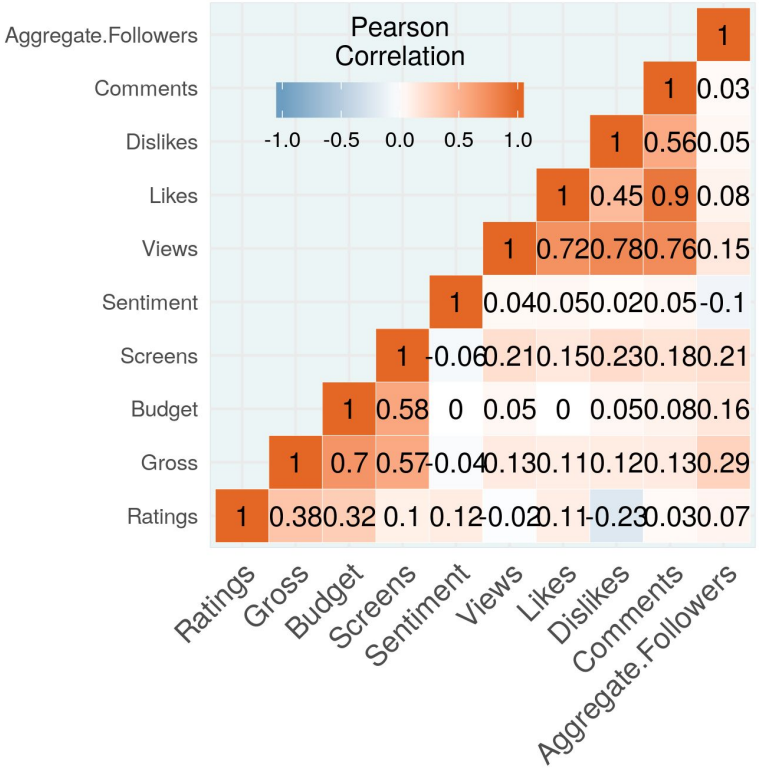
Resource: IMDB, Youtube and Twitter

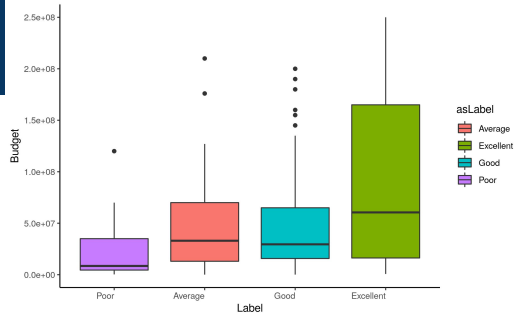
11 features are divided into two groups:
Conventional Features (5 types)&
Social Media Features (6 types)
To Predict **Rating**

Movie Ratings distribution can be seen from the graph on the right

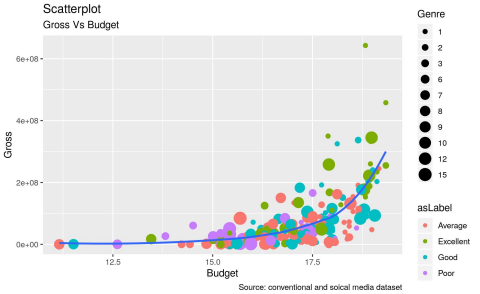


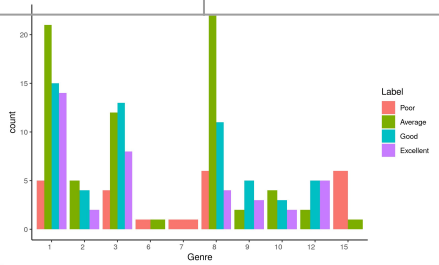
Dataset Description





Ratings	Label
0-5.2	Poor
5.2-6.4	Average
6.4-7.2	Good
7.2-10	Excellent





Methods

- Linear Regression
- Decision Tree (J48)
- Random Forest
- Support Vector Machine
- Naïve Bayesian
- LDA, QDA
- Artificial Neural Network

Paper detail and reproduce

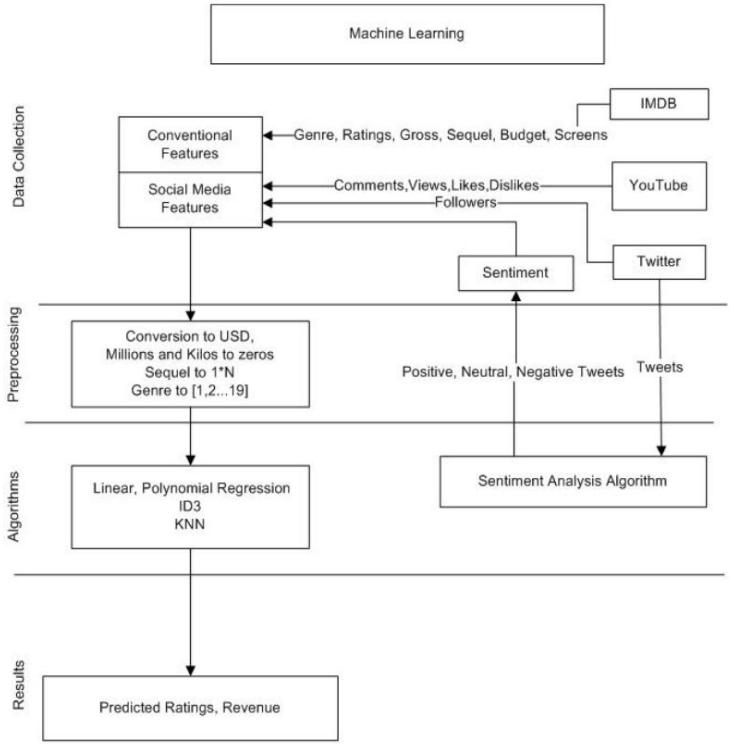
Section Subtitle Section Subtitle

First, values of Ratings are predicted using all other attributes except Gross Income (as gross income is not available before release). As Rating is a continuous numeric attributes, we have performed Linear Regression in order to predict the values.

Second, apply linear regression.

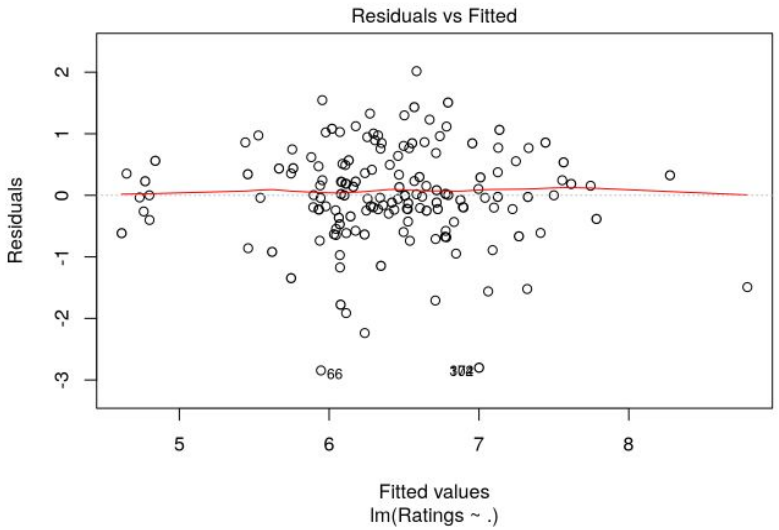
$$Accuracy_1 = \frac{\text{Number of Movies with exact prediction of Rating}}{\text{Total Number of Predictions}}$$

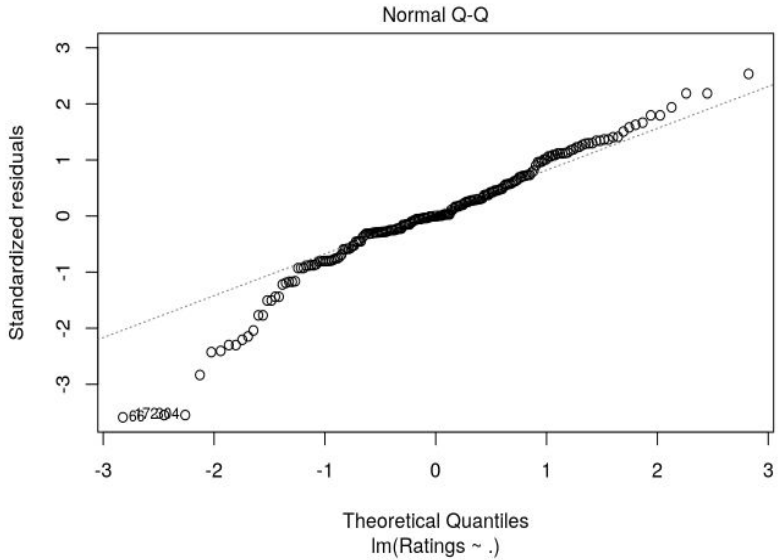
$$Accuracy_2 = \frac{\text{Number of Movies with approx prediction of Rating}}{\text{Total Number of Predictions}}$$



Linear Regression

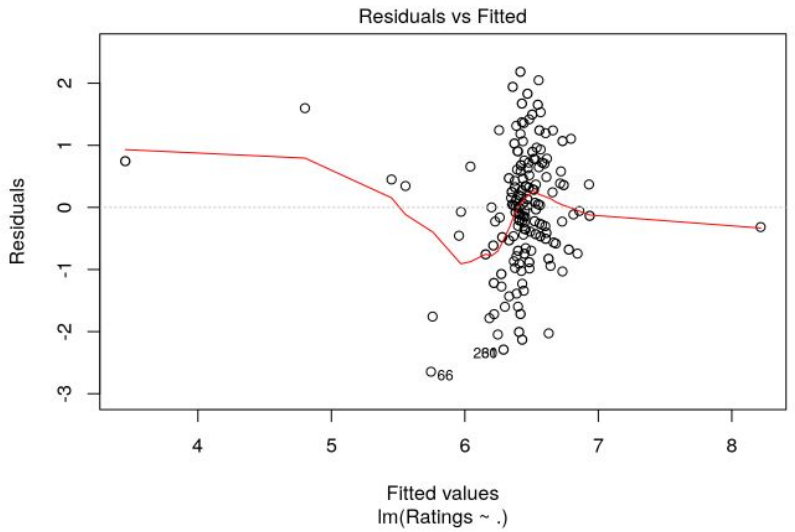
conventional features
formula = Ratings ~ ., data = conventional_train.df

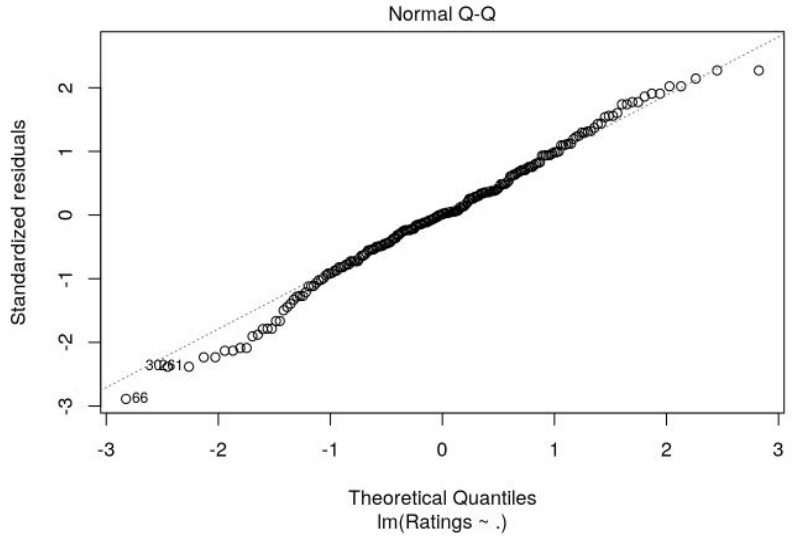




Linear Regression

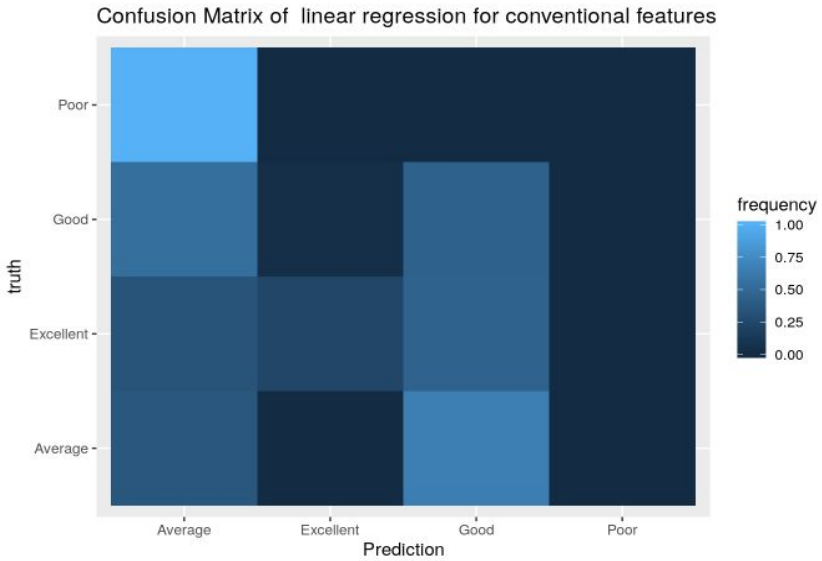
social media
formula = Ratings ~ ., data = socialmedia_train.df



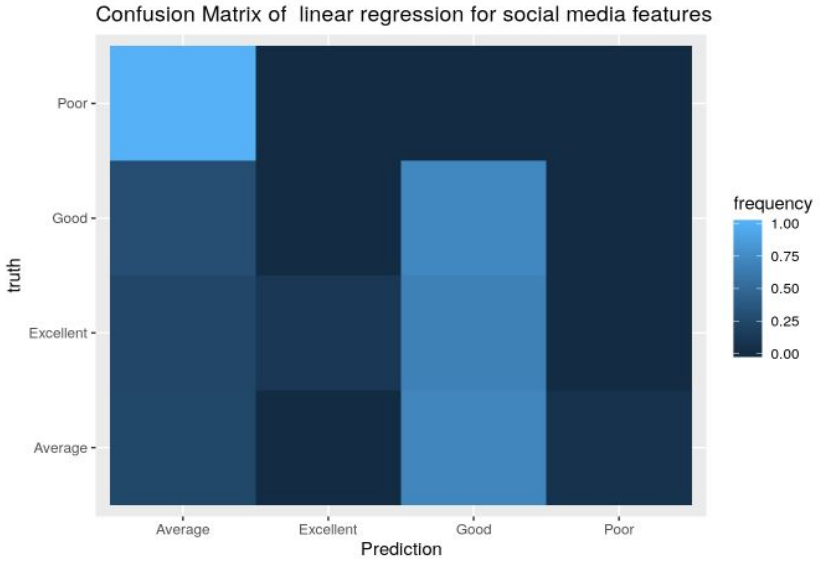


Linear Regression

convention
soft accuracy: 78.57%
classification accuracy: 35.71%
MSE: 0.7018



social media
soft accuracy: 75.0%
classification accuracy: 44.64%
MSE: 0.7167



Generalised Linear Regression

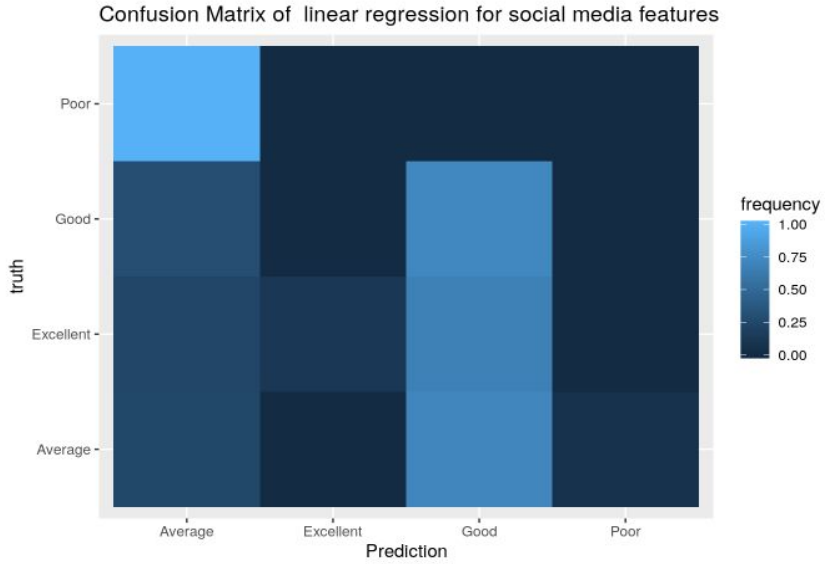
convention

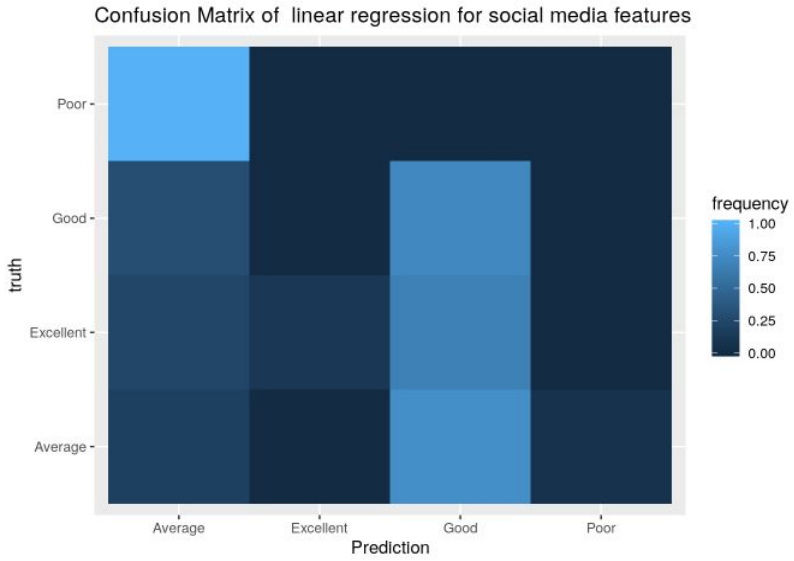
soft accuracy: 76.78%
classification accuracy: 34.71%
MSE: 0.7018

use Gamma (link = "log") as link function

social media

soft accuracy: 73.21%
classification accuracy: 42.85%
MSE: 0.7157

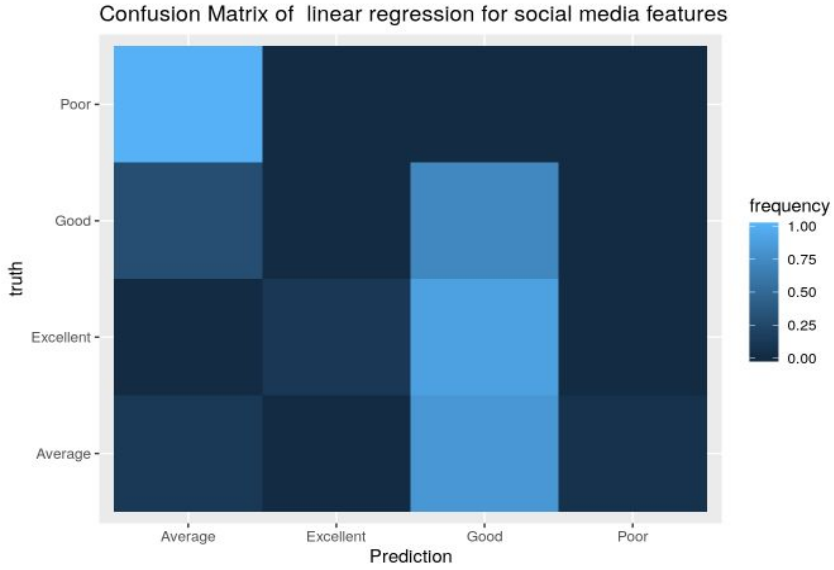




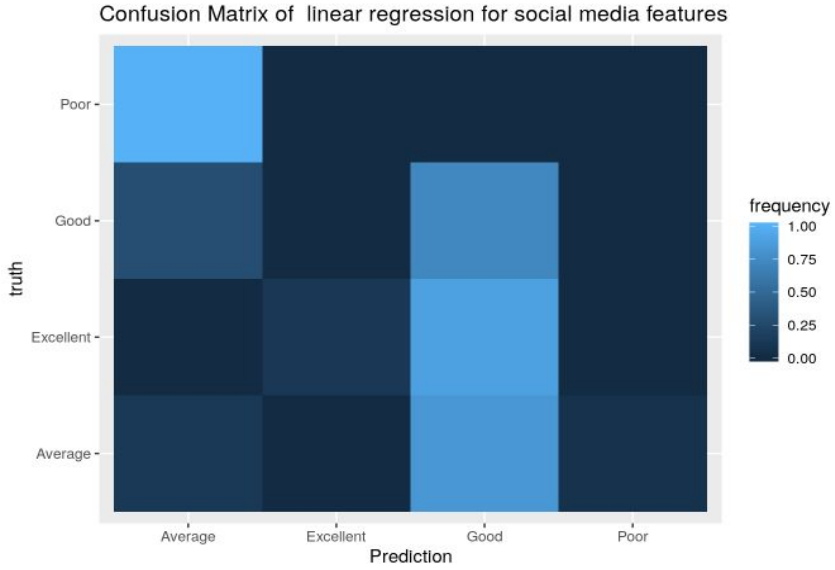
Generalised Additive Model

use ‘Gamma (link = "inverse")’ as link function

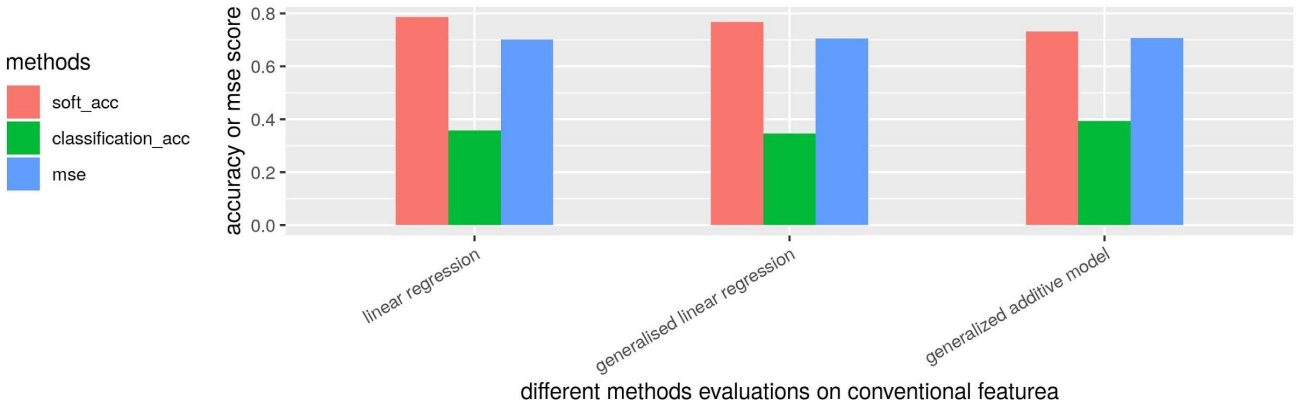
convention
soft accuracy: 73.21%
classification accuracy: 39.28%
MSE: 0.7079

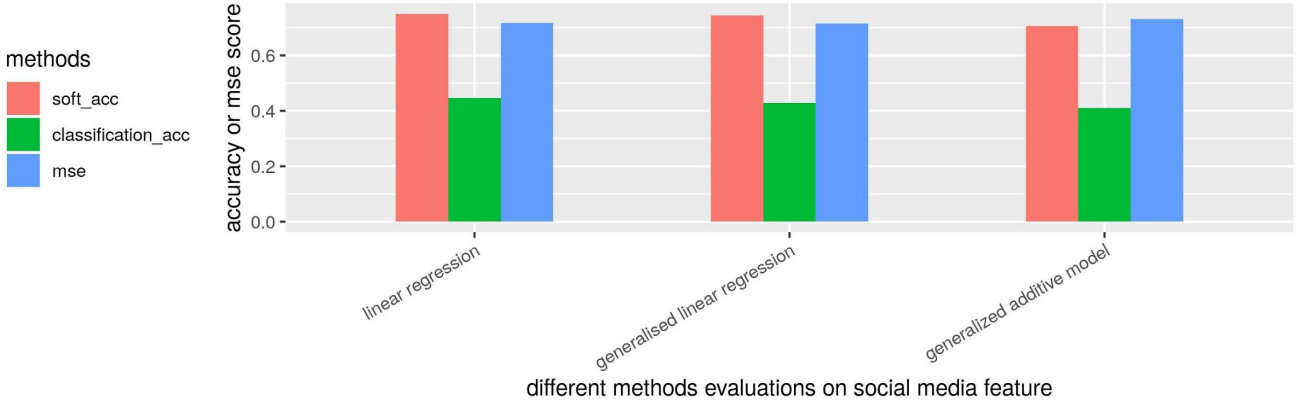


social media
soft accuracy: 70.56%
classification accuracy: 41.07%
MSE: 0.7321



Linear Regression





J48 Tree

$$Accuracy = \frac{\text{Number of Movies with correct prediction of BandofRating}}{\text{Total Number of Predictions}}$$

$$Info(D) = Entropy(D) = -\sum_{j} p(j|D) log p(j|D)$$

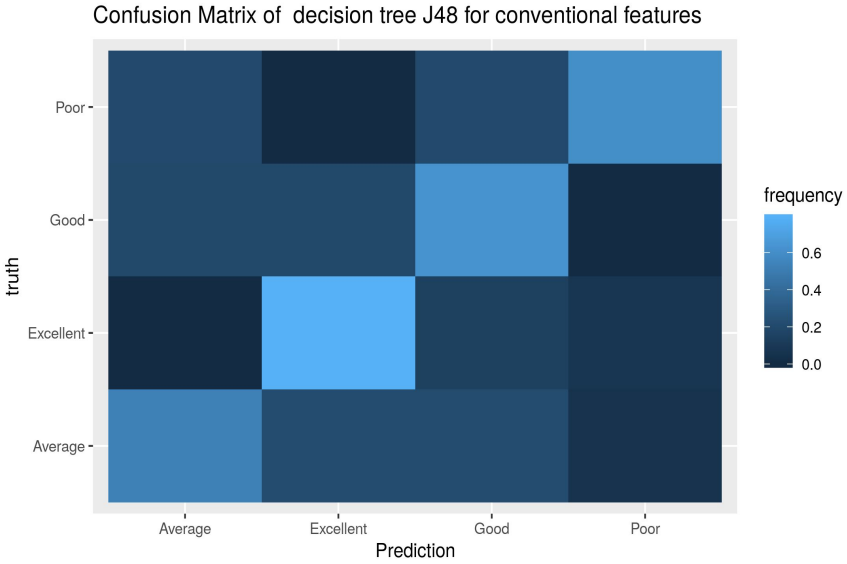
$$Info_A(D) = \sum_{i=1}^{v} \frac{n_i}{n} Info(D_i)$$

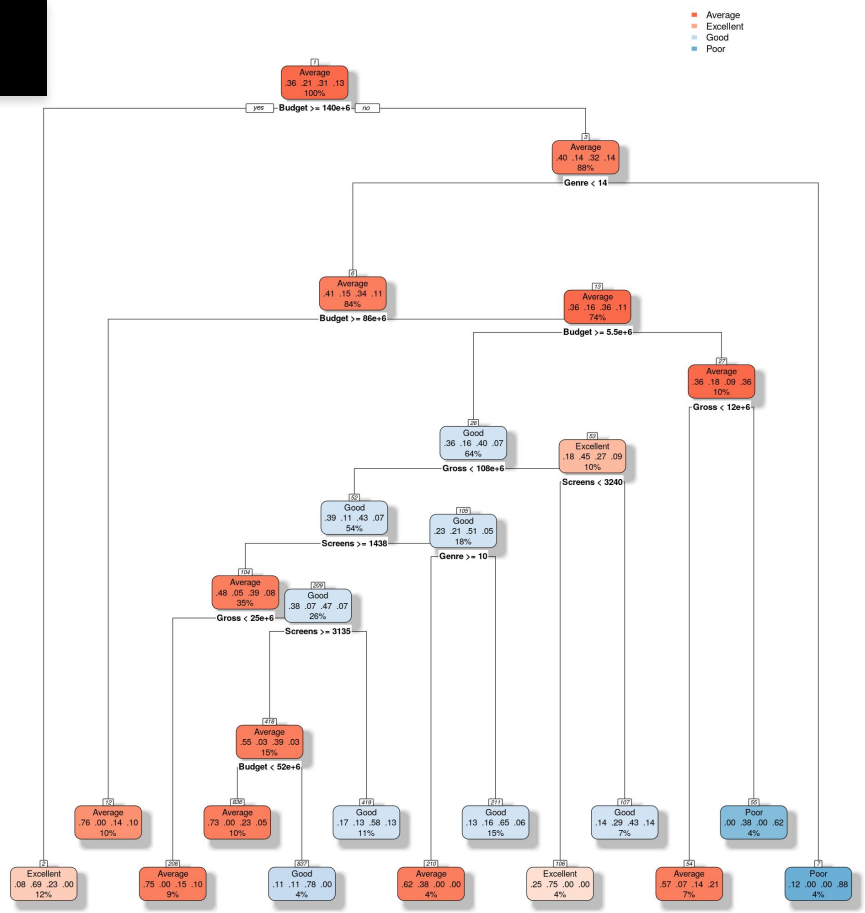
$$Gain\ (A) = Info(D) - Info_A(D)$$

J48 Tree

conventional

Accuracy:62.96%

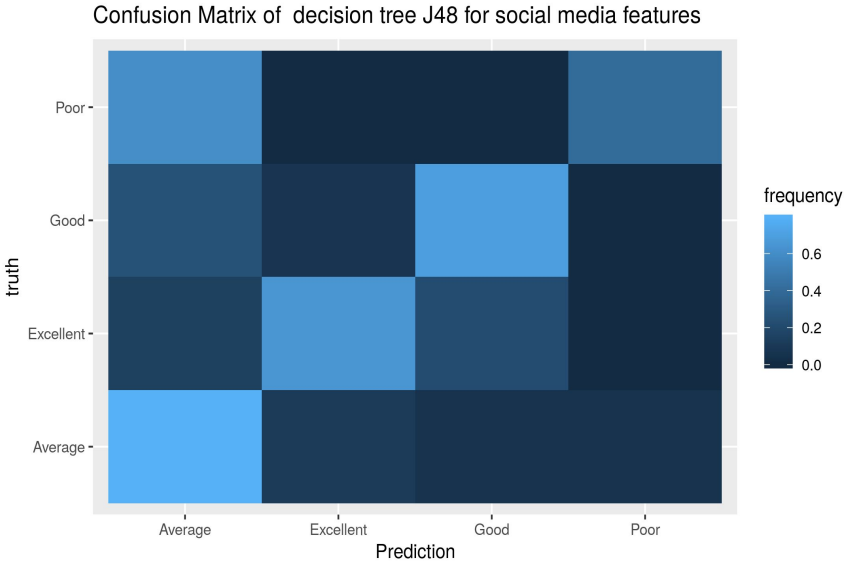


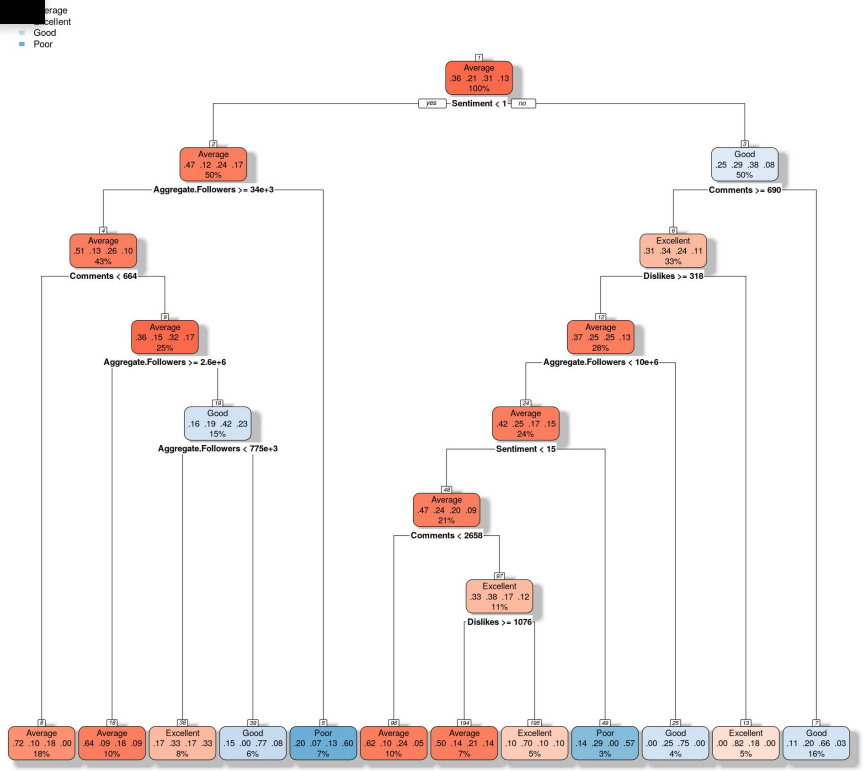


J48 Tree

social media

Accuracy:68.52%

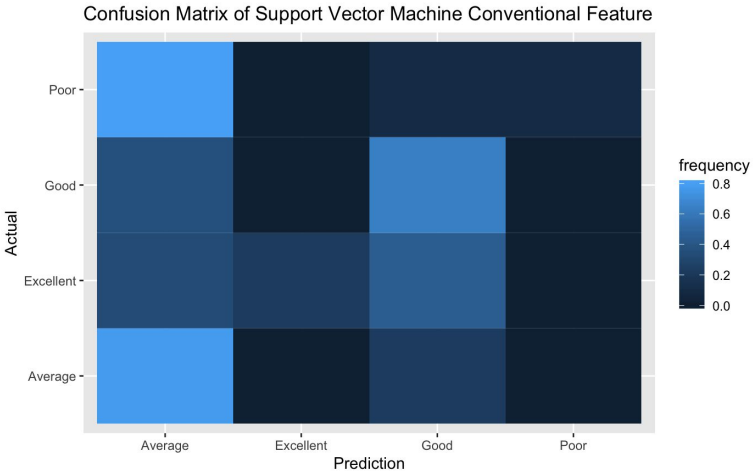




Other Techniques

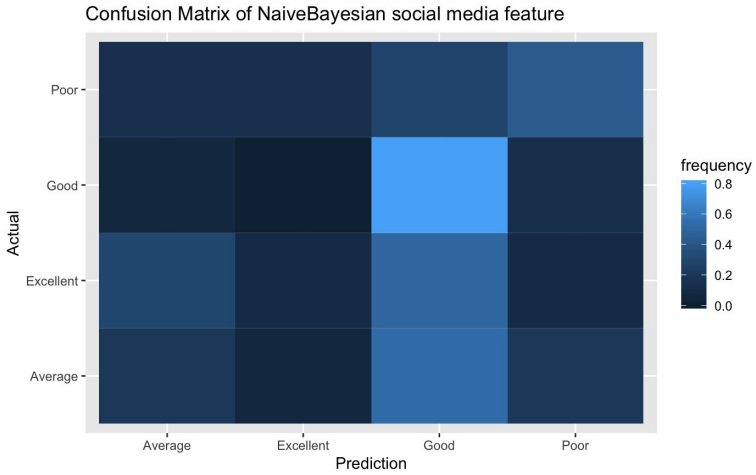
Support Vector Machine

Accuracy:0.4359



Kernel: radial

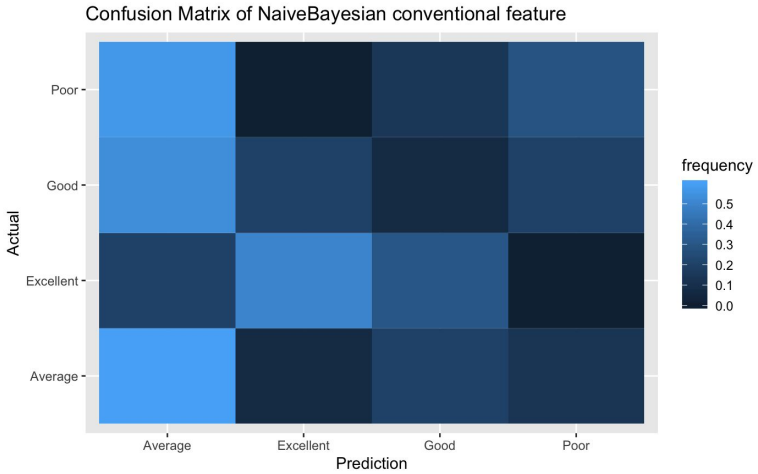
Accuracy:0.3846



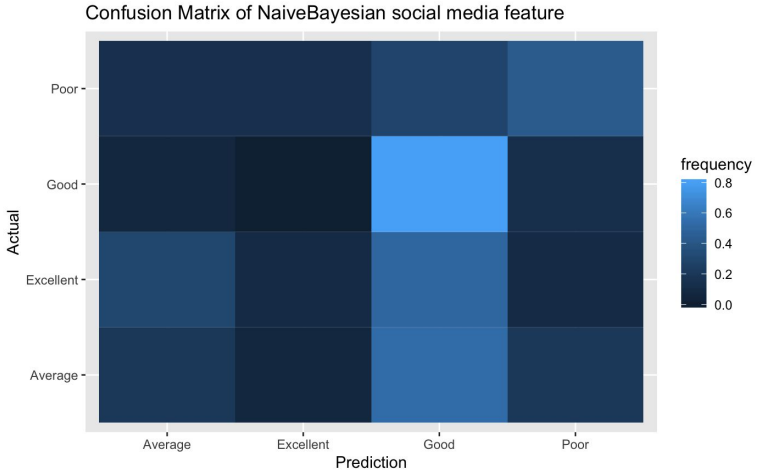
Kernel: polynomial
gamma: 1
coefficient: 7
degree: 3

Navïe Bayesian

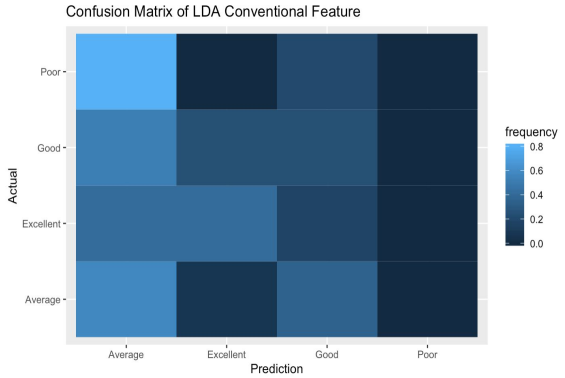
Accuracy:0.3617



Accuracy:0.4043



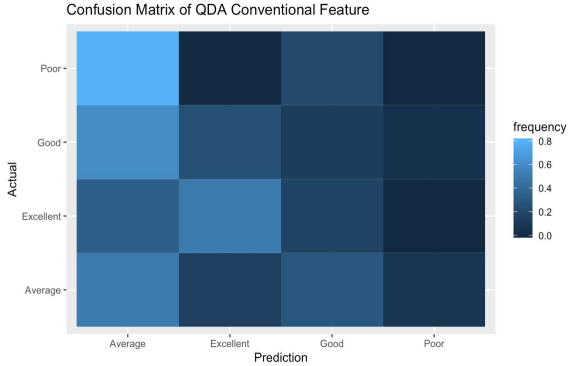
LDA & QDA

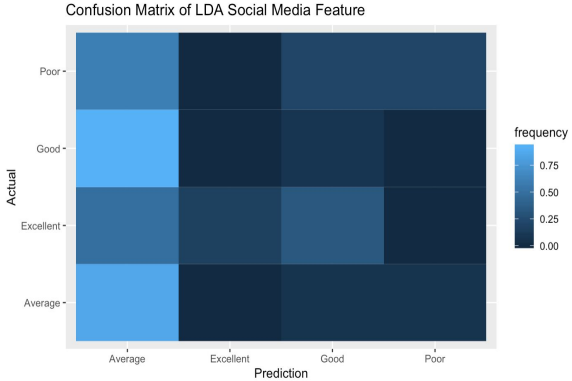


Poor Performance

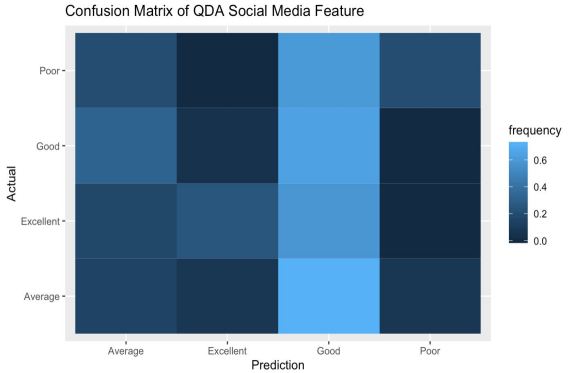
|

Curse of High Dimension



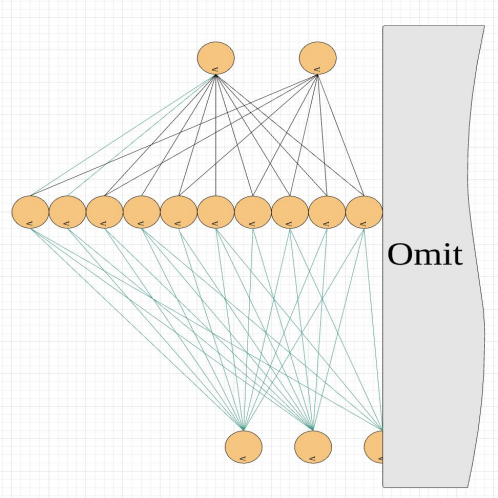


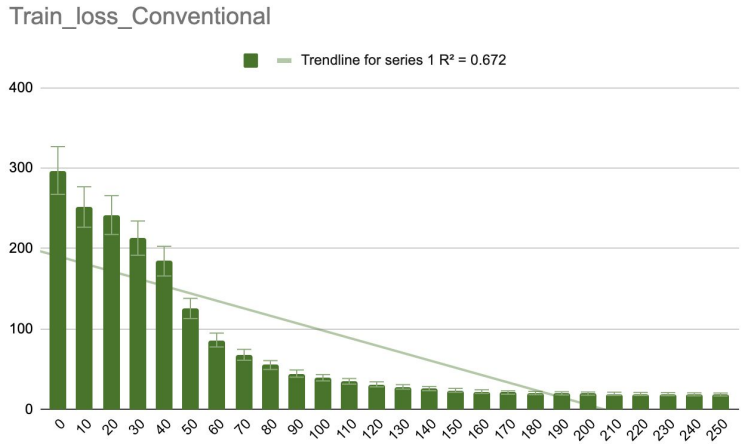
	LDA	QDA
Conventional	0.3035	0.3392
Social Media	0.3928	0.2857

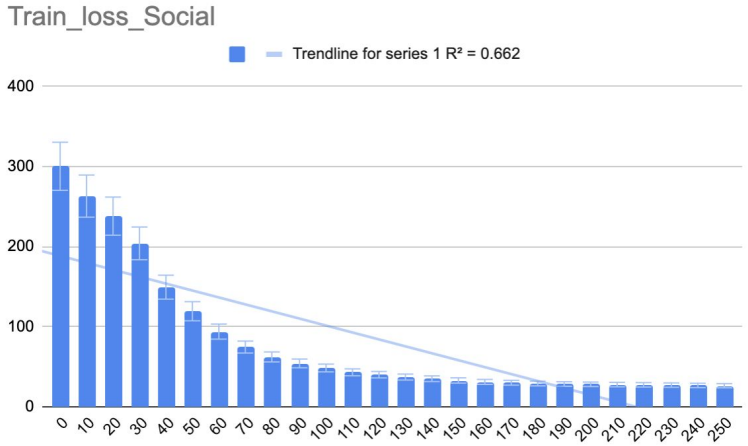


Artificial Neural Network

Architecture :
Number of units in the hidden layer = 20,
Initial random weights=0.2,
Weight decay=5e-4,
Maximum number of iterations = 250,
Skip=1.





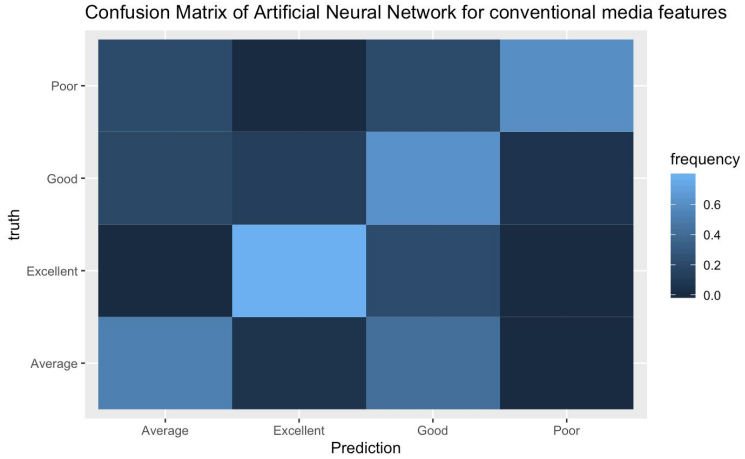


Artificial Neural Network

Test accuracy for conventional media feature and social media feature respectively

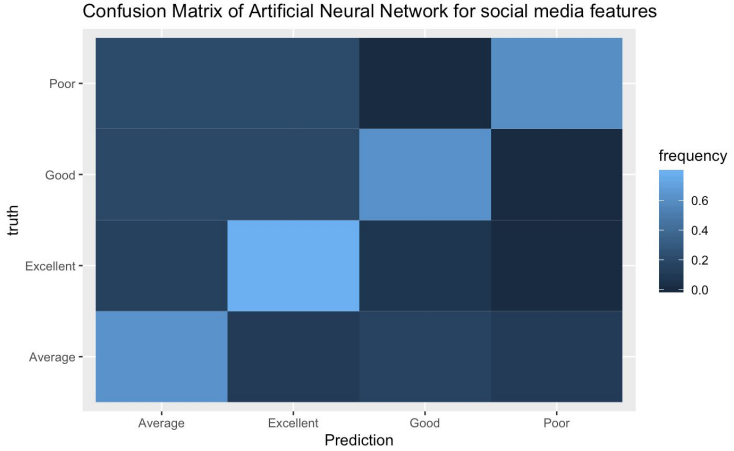
Convention:

accuracy: 66.67%



Social Media

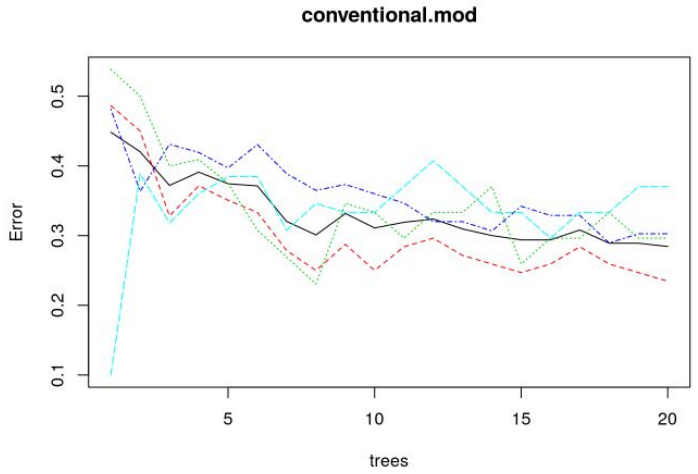
accuracy: 62.96%



Random Forest

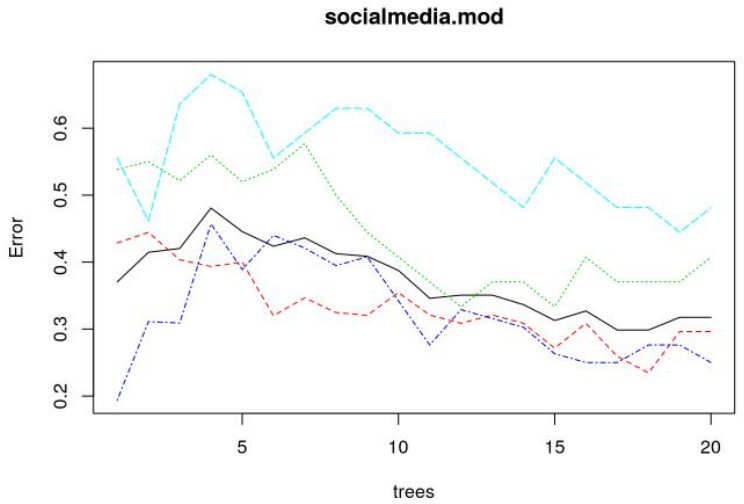
Convention:

training error for 20 subtrees



Social Media

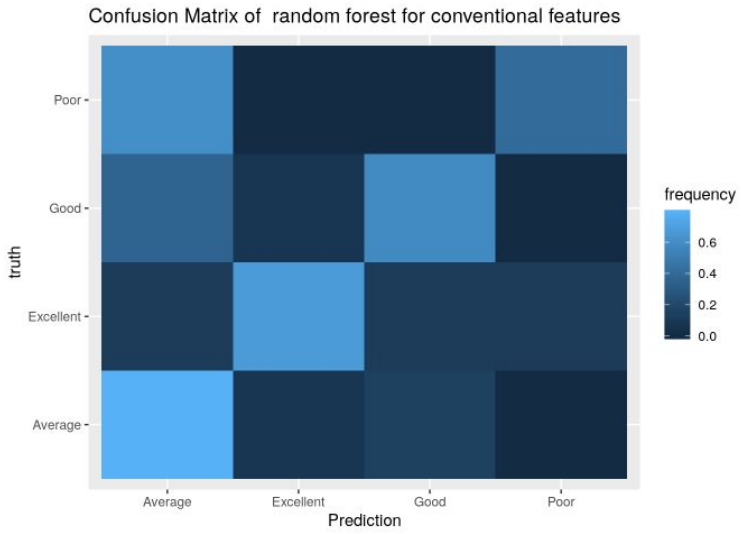
training error for 20 subtrees



Random Forest

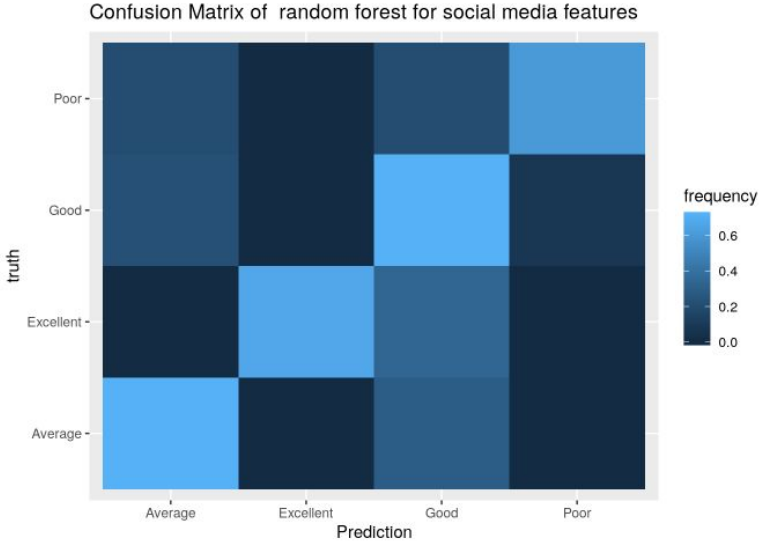
Convention:

accuracy: 62.5%

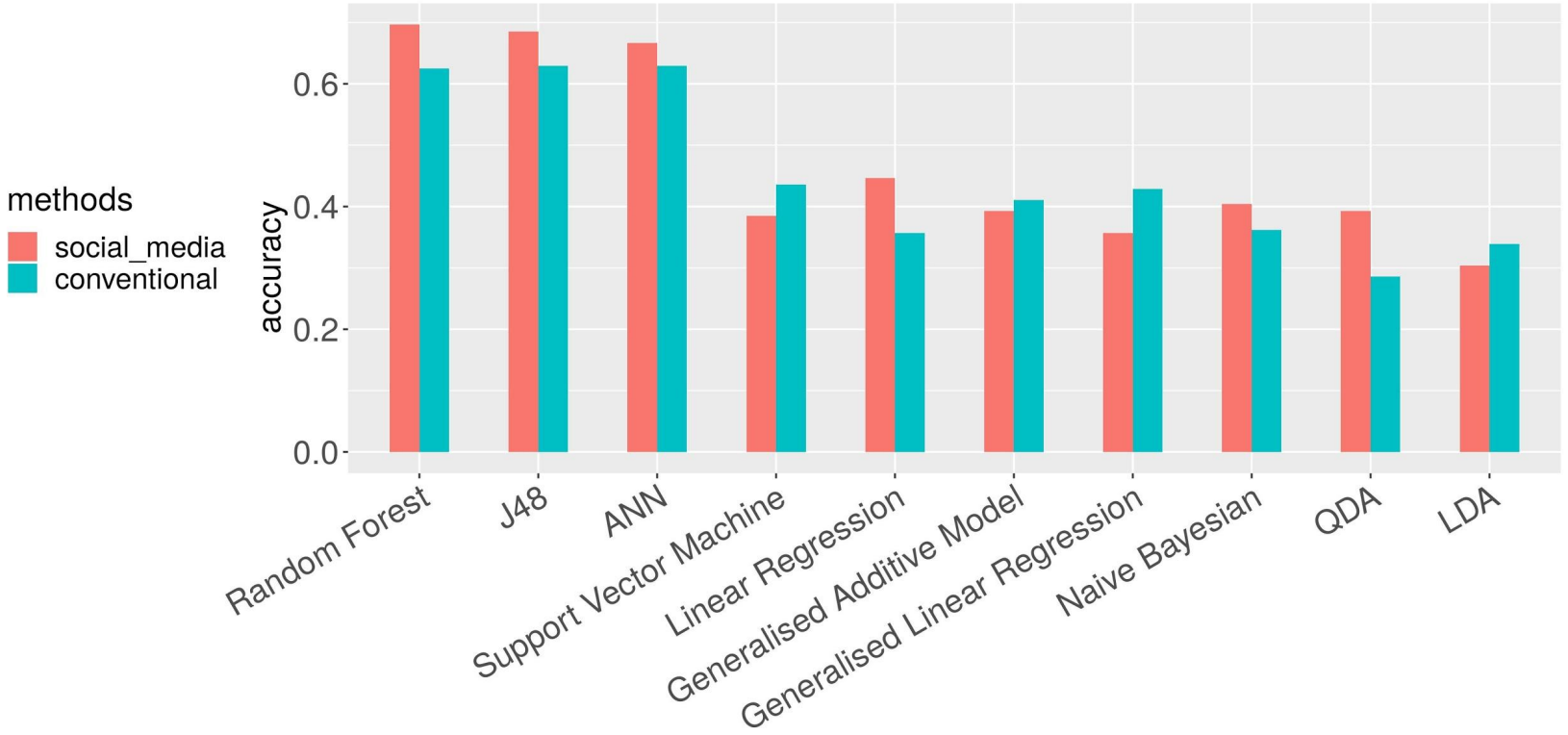


Social Media

accuracy: 69.64%



Conclusion and Discussion



Different methods evaluations on convensional features and social media features

Thank you