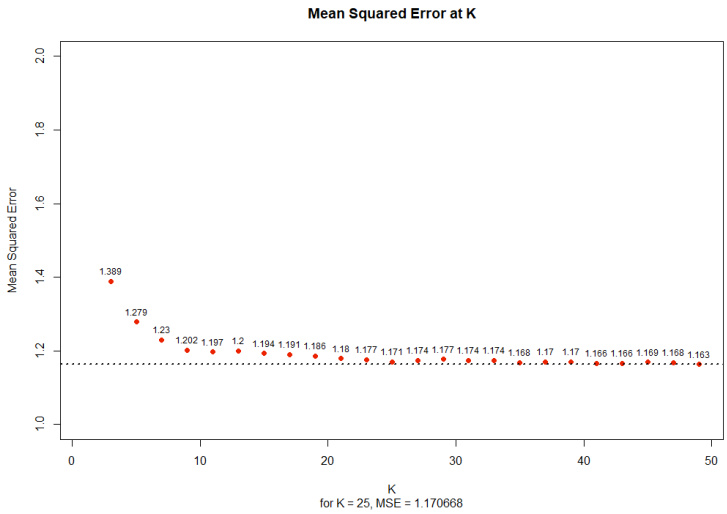
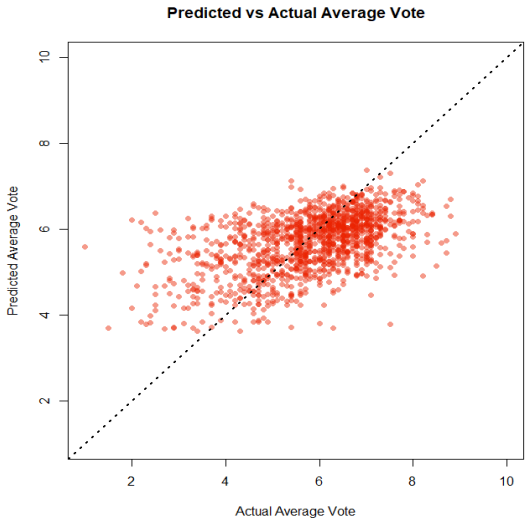
## Algorithms

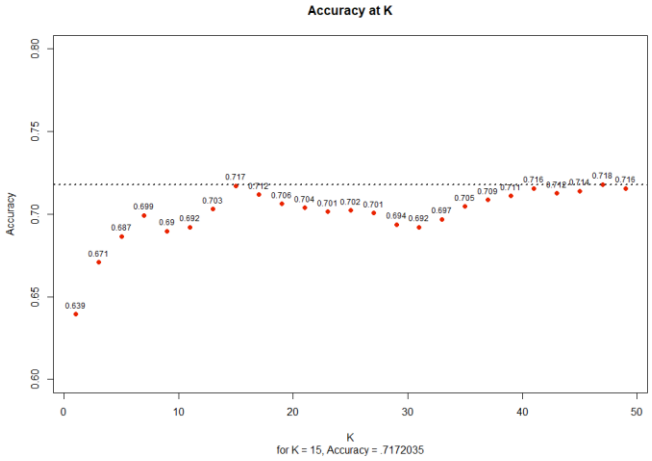
### 1. K Nearest Neighbor Algorithm:




Plotting predicted and actual avg_vote revealed issues with our model that our evaluation metrics did not immediately make clear

### KNN Classification Model:




| | Predicted No | Predicted Yes | |
|---|---|---|---|
| Actual No | 669 | 416 | 1,085 |
| Actual Yes | 370 | 1,090 | 1,460 |
| | 1,039 | 1,506 | 2,545 |

Testing on Unseen Data Results:

Accuracy: 69.11%

Sensitivity: 74.66%

Precision: 72.38%

2. **Random Forest (BEST MODEL):**

Random Forest Model Evaluation:

OOB estimate of error rate
27.54%

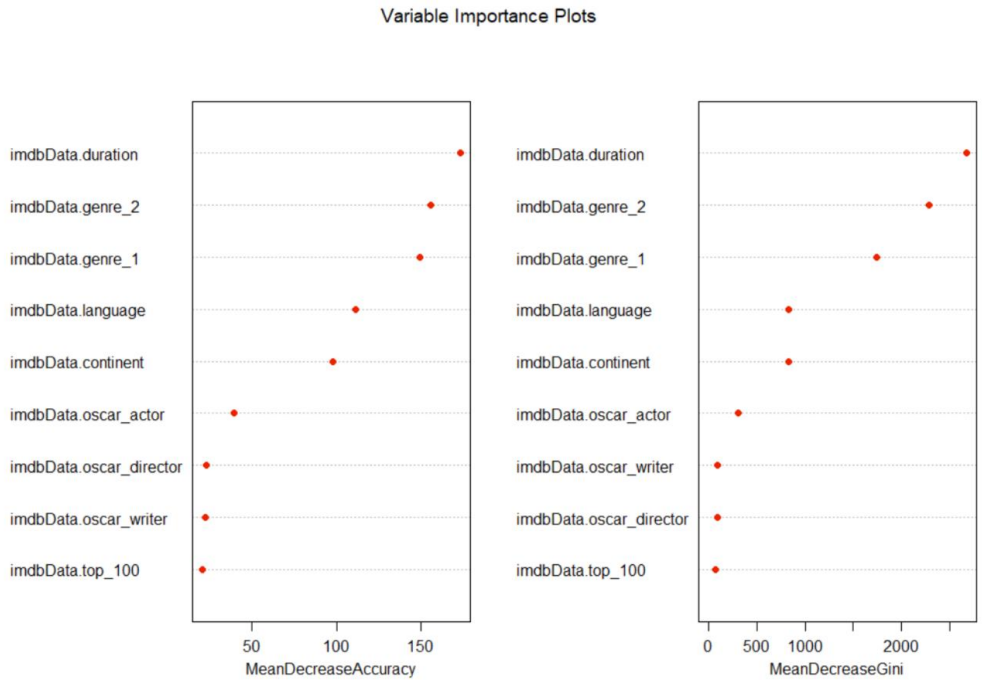
Accuracy
72.4609991%

Sensitivity
80.8940558%

Precision
73.4319545%

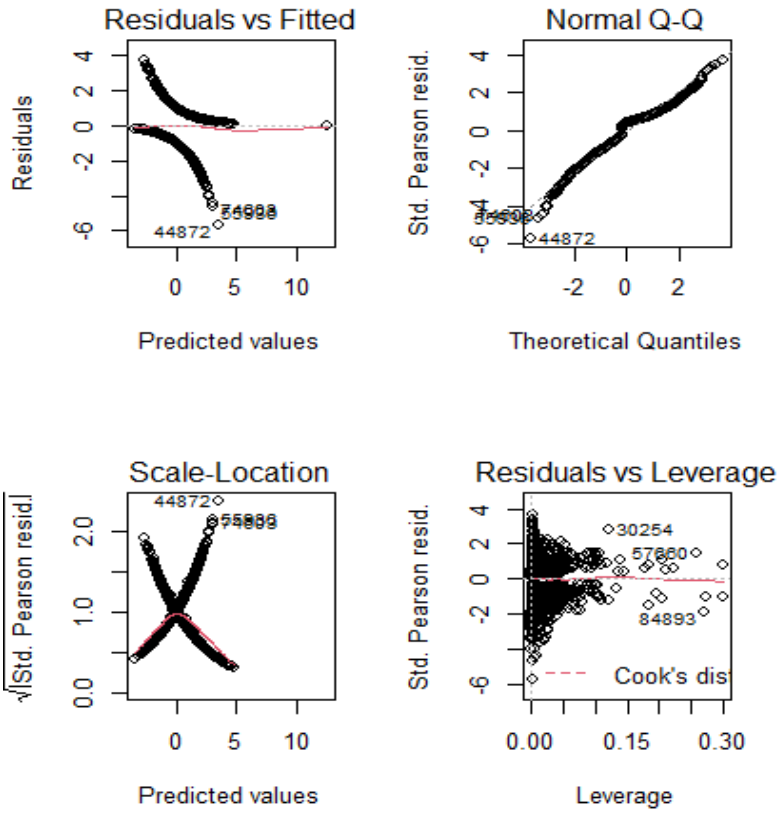
| | Predicted No | Predicted Yes | |
|---|---|---|---|
| Actual No | 13,407 | 8,459 | 21,866 |
| Actual Yes | 5,522 | 23,380 | 28,902 |
| | 18,929 | 31,839 | 50,768 |

Model was built using random forest classification using duration, genre_1, genre_2, continent, language, oscar_actor, oscar_director, oscar_writer, and top_100 as predictors




Important variable that we found was duration

### 3. Logistic Regression:




Model_2 took in less predictors than model_1, resulting in a better AIC score.

| | Predicted No | Predicted Yes | |
|---|---|---|---|
| Actual No | 253 | 200 | 453 |
| Actual Yes | 99 | 448 | 547 |
| | 352 | 648 | 1,000 |

Results on Unseen Data (using probability threshold of 50%):

Accuracy: 70%

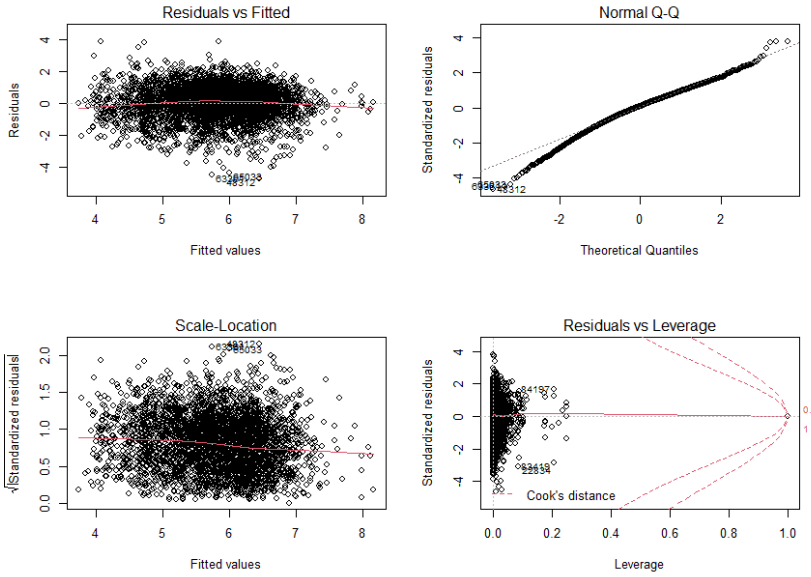
Sensitivity: 56%

Precision: 72%

## 4. Linear Regression:

Model Output:

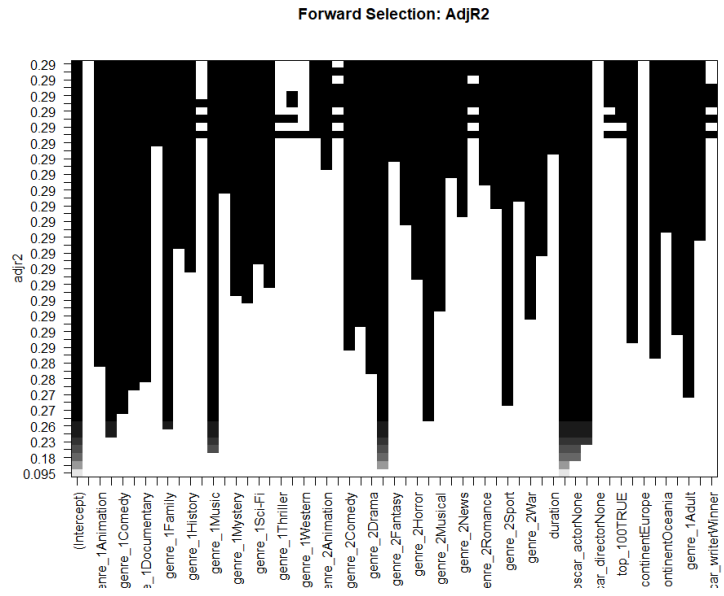
```
Residuals:
  Min      -4.7429
  1Q       -0.5849
  Median    0.0886
  3Q        0.6857
  Max       3.8662

Residual standard error:
  1.02 on 3951 degrees of freedom

Multiple R^2    0.3164
Adjusted R^2    0.3081

F-statistic:
  38.10 on 48 and 3951 DF
  p-value    < 2.2e-16
```




For the most part, our residual vs fitted had no funneling shape, which lead to no indication of heteroscedasticity. Also, it had a random scattering around the 0 line, indicating a linear relationship.

Our Q-Q graph showed us a normally distributed graph because it did not have an S pattern.

**Forward Selection:**




Forward Selection to figure out our optimal set of predictors based on lowest RSS