

Eric Young

Data Science – Final Project Paper

Predicting Box Office Results and Movie Sentiment Using Supervised Learning Methods

Project Introduction and Objective

Box Office success can be extremely influential in the production and funding of future works and may serve as an indication of positive or negative sentiment regarding a theatrical movie release. Conversely, movie sentiment may be one of many factors affecting the Box Office success of a movie. This paper explores the use of supervised learning methods to predict Box Office results and movie sentiment based on a defined feature space of attributes obtained from three data sources – Box Office Mojo, The Open Movie Database (OMDb)¹, and Rotten Tomatoes. Select variables from those data sources are provided in Table 1.




| | | |
|---|----------------|--|
|  | Title | ▪ Title of the movie |
| | Total Gross | ▪ Total domestic gross earnings |
| | Total Theatres | ▪ No. of locations playing movie |
| | Opening Date | ▪ Opening date of movie |
|  | Title | ▪ Title of the movie |
| | Year | ▪ Year of movie release |
| | Rating | ▪ Film rating (e.g. PG, PG-13, R) |
| | Runtime | ▪ Length of the movie |
| | Genre | ▪ Genre(s) of the movie |
| | Director | ▪ Director(s) of the movie |
| | Cast | ▪ Notable cast members in the movie |
| | Metacritic | ▪ “Metascore” from critics (1 – 100) |
| | IMDb Rating | ▪ Wtd. avg. of IMDb user votes (1 – 10) |
| | IMDb Votes | ▪ No. of IMDb registered user votes |
|  | Plot | ▪ Plot summary |
| | Sentiment | ▪ Fresh or rotten based on T Rating |
| | T Rating | ▪ % of positive critic reviews (1 – 100) |
| | User Meter | ▪ % positive user ratings (1 – 100) |
| | User Rating | ▪ Avg. user rating (1 – 5) |
| | User Reviews | ▪ No. of user reviews |

Table 1. Select variables in data obtained from Box Office Mojo, OMDb, and Rotten Tomatoes.

¹ Information provided by OMDb represents data sourced from IMDb.

Data Pre-Processing Steps

Two datasets comprised of one target and multiple feature variables were compiled from these data sources, requiring the data to be cleaned and prepped for analysis. Pre-processing of the data involved:

1. Web-scraping data from BoxOfficeMojo.com using web data extraction platform, Import.io.
2. Removing extraneous characters, including dollar signs, asterisks, and additional spaces from numbers and strings.
3. Converting certain string formats to numeric and date formats.
4. Replacing ampersand (“&”) with “and” in movie titles.
5. Decoding strings from Latin to UTF-8 equivalents.
6. Lowercasing all strings.
7. Imputing and removing certain null values.
8. Filtering movie results for the time period from 2000 – 2015, given increasing quality of Box Office tracking and relevance of Rotten Tomatoes, which was officially launched in 2000.
9. Creating combined movie title and year (Title_Year) feature in Box Office Mojo and OMDb data.
10. Creating two new datasets: OMDb and Rotten Tomatoes data joined on ID, and Box Office Mojo, OMDb, and Rotten Tomatoes data joined on Title_Year and ID (referred to as BOT).
11. Mapping numbers to categorical variables such as Rating and Sentiment (i.e. Fresh or Rotten).

Predicting Box Office Results

Of the two new datasets created during pre-processing, certain variables were selected from the BOT dataset to create a new dataset comprised of the target, Total Gross, and its feature space. Features were chosen to understand the impact of movie attributes, ratings, and reviews on the target variable. The feature space included: Total Theatres, Runtime, Metacritic, IMDb Rating, IMDb Votes, T Rating, Meter, Reviews, Fresh, Rotten, User Meter, User Rating, and User Reviews.

Visualizations were then used to explore the relationship between the target and its features, as well as between features. In doing so, three principal relationships were identified. These key variable relationships are shown in Figure 1.

The first scatter plot between Total Gross and Total Theatres exhibits a linear-to-exponential relationship, similar to the relationships between Total Gross and IMDb Rating, IMDb Votes, Reviews, and User Meter. The second scatter plot between Total Gross and Runtime exhibits no relationship, similar to the relationships between Total Gross and Metacritic and T Rating. The third scatter plot between two features, User Meter and User Rating, demonstrates collinearity, which can also be seen in the relationships between IMDb Rating and Metacritic.

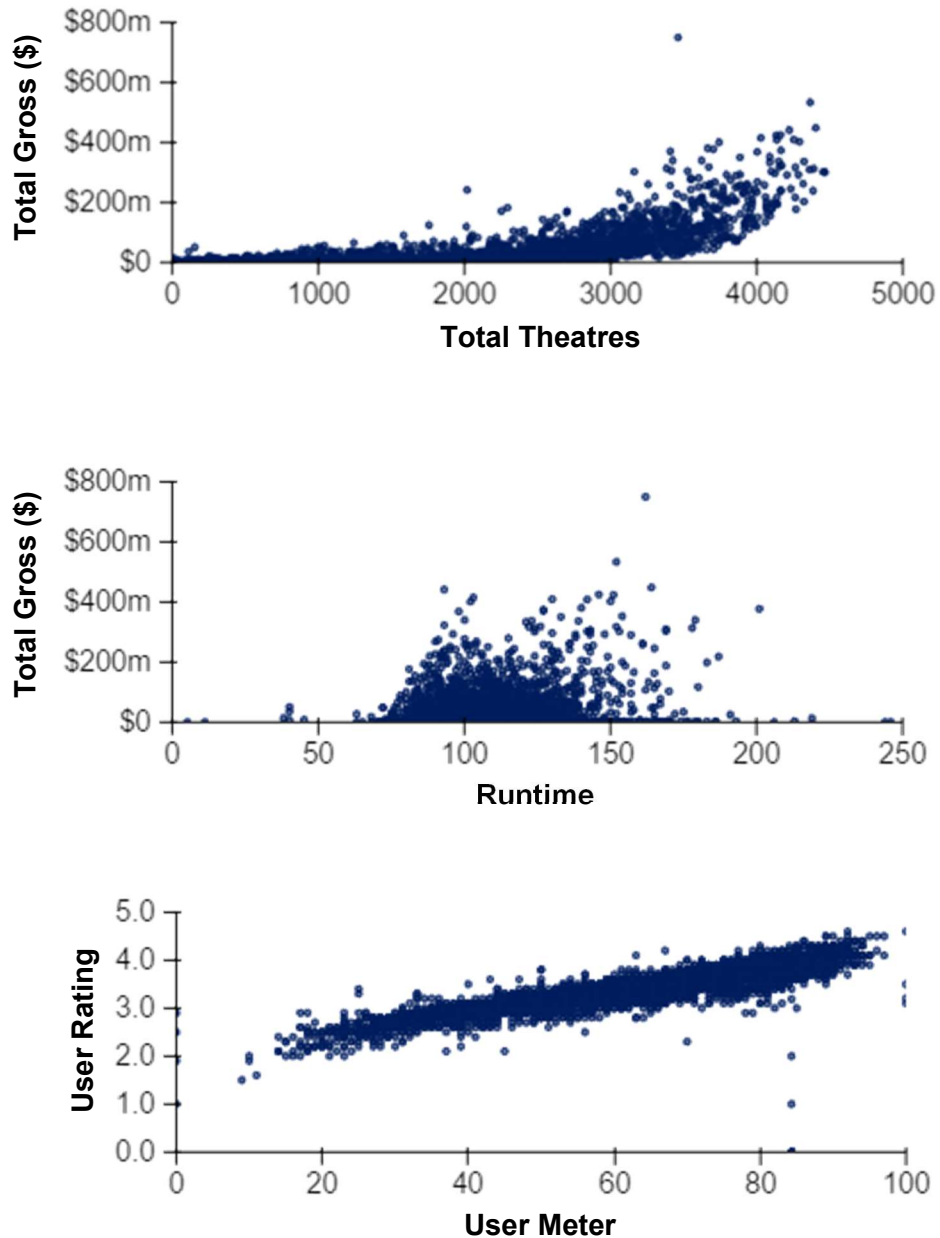


Figure 1. Scatter plot of key variable relationships.

Given the growing maturity of ratings and reviews being employed by IMDb, Metacritic, and Rotten Tomatoes, the dataset was segmented into three five-year time periods – 2000 – 2004, 2005 – 2009, 2010 – 2014 – with 2015 reserved for out-of-sample (OOS) testing. Each dataset was then trained and tested using the ordinary least squares (OLS) method of Linear Regression, producing summary statistics and predictions. The summary statistics for each time period are shown in Figure 2.

| 2010 - 2014 | | | | |
|----------------|--------------|---------|----------------|---------------|
| Variables | Coefficient | P-Value | Adj. R-Squared | Condition No. |
| Total Theatres | 32,280,000 | 0.000 | 0.751 | 1.82E+16 |
| Runtime | 1,451,000 | 0.194 | | |
| Metacritic | 3,569,000 | 0.206 | | |
| IMDb Rating | (7,199,000) | 0.000 | | |
| IMDb Votes | 27,280,000 | 0.000 | | |
| T Rating | 20,460 | 0.996 | | |
| Meter | 1,881,000 | 0.588 | | |
| Reviews | (3,703,000) | 0.001 | | |
| Fresh | (409,600) | 0.783 | | |
| Rotten | (6,791,000) | 0.000 | | |
| User Meter | (16,750,000) | 0.000 | | |
| User Rating | 22,870,000 | 0.000 | | |
| User Reviews | 9,925,000 | 0.000 | | |
| 2000 - 2004 | | | 2005 - 2009 | |
| Adj. R-Squared | 0.707 | | 0.708 | |
| Condition No. | 9.69E+15 | | 8.99E+15 | |

Figure 2. Summary statistics for each time period using OLS Linear Regression.

The average test set prediction error for each time period from 2000 – 2014 was 7.7%, 6.9%, and 5.9%, respectively. Additionally, adjusted R-squared of 0.751 for the time period 2010 – 2014 indicated proportionally more variance being explained by the model than those fitted on data from earlier time periods. Although condition numbers reported by each model point to significant multicollinearity between features, the model trained on data from 2010 – 2014 was selected on the basis of the above mentioned results. Using this model to predict OOS data yielded an average prediction error of 6.4%.

In reviewing p-values, features identified as not statistically significant at the alpha = .05 level and those exhibiting collinearity were removed. These features included: Runtime, Metacritic, T Rating, Meter, Fresh, and User Meter. After training and testing the model on the new feature space, the Reviews feature was removed, and a final set of summary statistics were produced. These results are shown in Figure 3.

| 2010 - 2014 | | | | |
|---------------------------|-------------|---------|-------------------|------------------|
| Variables | Coefficient | P-Value | Adj. R-Squared | Condition No. |
| Total Theatres | 31,050,000 | 0.000 | 0.748 | 4.35 |
| IMDb Rating | (6,926,000) | 0.000 | | |
| IMDb Votes | 26,160,000 | 0.000 | | |
| Rotten | (8,045,000) | 0.000 | | |
| User Rating | 9,058,000 | 0.000 | | |
| User Reviews | 9,658,000 | 0.000 | | |
| Test Sample Metrics | | | | |
| Avg Prediction (\$) | | | \$35,733,508 | |
| Avg Prediction Error (\$) | | | \$2,213,967 | |
| Avg Prediction Error (%) | | | 6.60% | |
| Min Prediction (\$) | | | (29,149,846) | |
| Max Prediction (\$) | | | \$274,829,881 | |
| Avg Actual (\$) | | | \$33,519,541 | |
| Min Actual (\$) | | | \$935 | |
| Max Actual (\$) | | | \$400,738,009 | |

Figure 3. Summary statistics for 2010 – 2014 updated OLS Linear Regression model.

Based on the above results, the condition number vastly improved, while adjusted R-squared remained relatively in-line with its previous result. Despite a slight increase in the average test set prediction error, average OOS prediction error improved to 6.2%.

Although average prediction error suggests a suitable model, further inspection points to one that does not generalize well for a significant proportion of the data. Figure 4 illustrates the distribution of prediction error for both the test set and OOS data. While Box Office success for a majority of films can be predicted adequately with less than 10% error, ~30% of films predicted by the model resulted in greater than 25% error and nearly 20% of films in greater than 100% error.

There may be a number of potential reasons for such an outcome. Linear Regression models do not capture non-linear relationships between dependent and independent variables, and it is very possible the relationship between the two can be explained by another model, perhaps exponential. Linear Regressions also tend to be high bias and low variance in nature, as the OLS method aims to minimize the overall error between the observed and predicted responses, leading to underfitting. Many features were not contemplated, including movie ratings, genre, cast, and production budget. Moving forward, more features should be added to the model to expand the hypothesis space, which may also better represent the data.

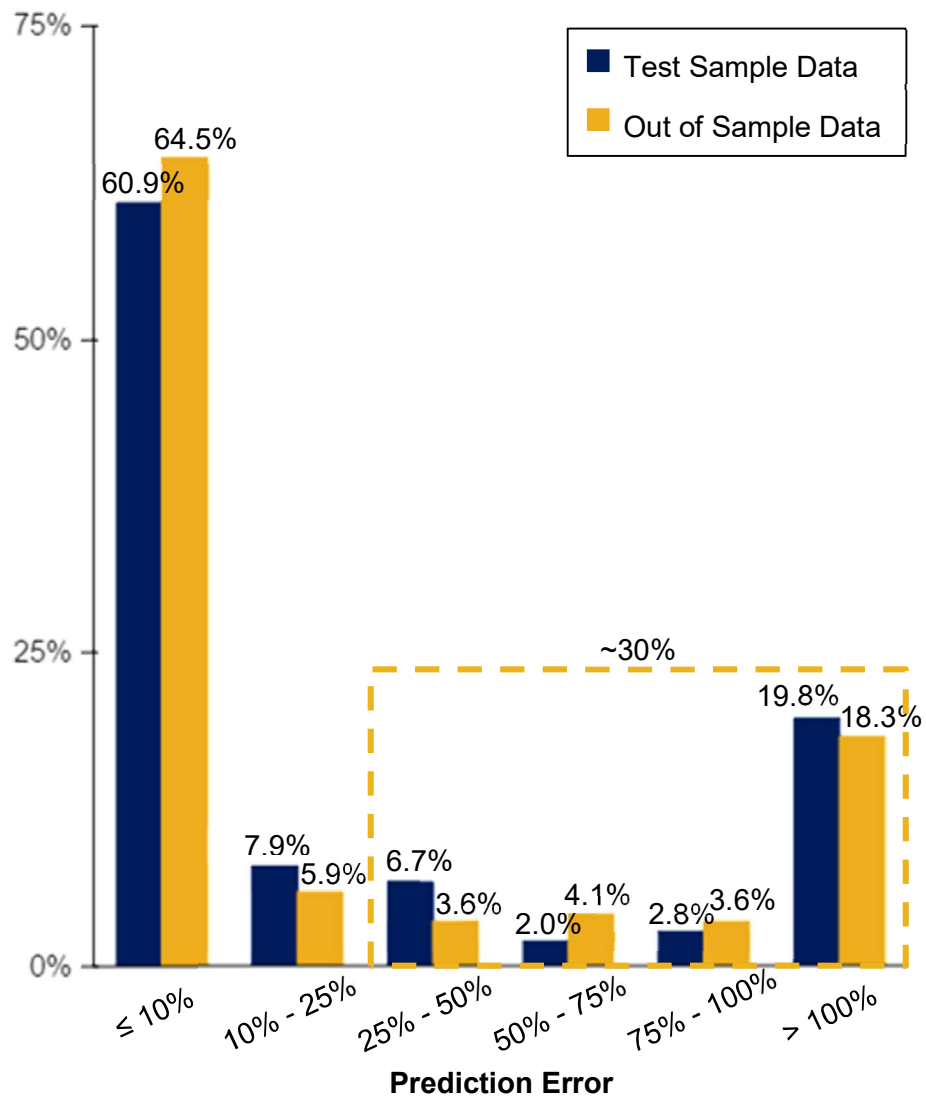


Figure 4. Distribution of prediction error for test set and OOS data.

Predicting Movie Sentiment Based on Box Office Results and Select Movie Attributes

Predicting movie sentiment is structured as a binomial classification problem, where sentiment is defined as whether or not a movie has received a fresh or rotten rating from Rotten Tomatoes, which represents a consensus view by film critics. The feature space was constructed to understand the impact Total Gross, Rating, Runtime, Genre, Director, and Cast would have on movie sentiment.

Further pre-processing of the BOT dataset was required in order to prepare it for classification. Training and testing data were comprised of movies from 2000 – 2014, with 2015 movies reserved for OOS testing. Continuous variables such as Total Gross and Runtime were standardized and scaled with zero mean and unit variance and categorical variables such as Rating, Genre, Director and Cast were one-hot encoded. Genre, Director, and Cast, listed multiple names for each movie and needed to be separated into individual features. For instance, the Genre of a movie listed as “Action, Adventure, Fantasy” would result in three distinct features “Genre_Action,” “Genre_Adventure,” and “Genre_Fantasy” coded as Class 1. These pre-processing steps resulted in a very large feature space of nearly 10,000.

Classifiers including Logistic Regression, Naïve Bayes, KNN, Random Forest, and SVM were all developed to determine which one would perform the best. For the Logistic Regression and Random Forest classifiers, initial parameters for the model were determined by testing a range of values and selecting the one producing the highest cross-validation score. These parameters were C-value and number of trees, respectively. Both feature coefficient and importance were examined and the initial model was then used to predict OOS data. Top features by absolute values are provided in Tables 2 and 3.

| LR Top Features / Coefficients | | RF Top Features / Importance | |
|--------------------------------|--------|------------------------------|-------|
| Documentary | 2.69 | Total Gross | 0.042 |
| Paul Giamatti | 1.49 | Runtime | 0.035 |
| PG-13 | (1.44) | Documentary | 0.029 |
| PG | (1.38) | Comedy | 0.014 |
| Michael Bay | (1.29) | Drama | 0.014 |
| Robert Rodriguez | 1.14 | PG-13 | 0.011 |
| Penelope Cruz | (1.14) | Biography | 0.011 |
| Ewan McGregor | 1.11 | Action | 0.010 |

Tables 2 and 3. LR feature coefficient and RF feature importance for top features by absolute value.

For Logistic Regression, coefficients indicate both feature importance and direction. Positive coefficients increase the likelihood the target will be categorized as Class 1 (Fresh) and negative coefficients increase the likelihood of Class 0 (Rotten), given a change in the respective feature. Absolute values of the coefficients indicate the degree or magnitude of feature importance. For instance, Documentary movies and those with Paul Giamatti increase the probability the movie will be Fresh, and those with PG-13 or PG movie ratings increase the probability the movie will be Rotten.

Random Forest feature importances do not point to the direction of classification, only the average impurity decrease from each feature. Feature selection based on impurity reduction suffers from a bias

towards preferring variables with more categories. However, top features shown in Tables 2 and 3 reflect those likely attached to more movies, and may also contribute to the impurity reduction power of other correlated features. This may partially explain Total Gross and Runtime exhibiting three to four times the importance of genres Biography and Action.

SVM initial parameters were set to linear kernel and the C-value producing the optimal test score, and KNN initial parameters were selected based on K-neighbors producing the optimal cross-validation score. No tuning was performed on the Naïve Bayes model. ROC curves and OOS predictions were made for each classifier using these initial parameters. See Figure 5 for ROC curves and OOS predictions.

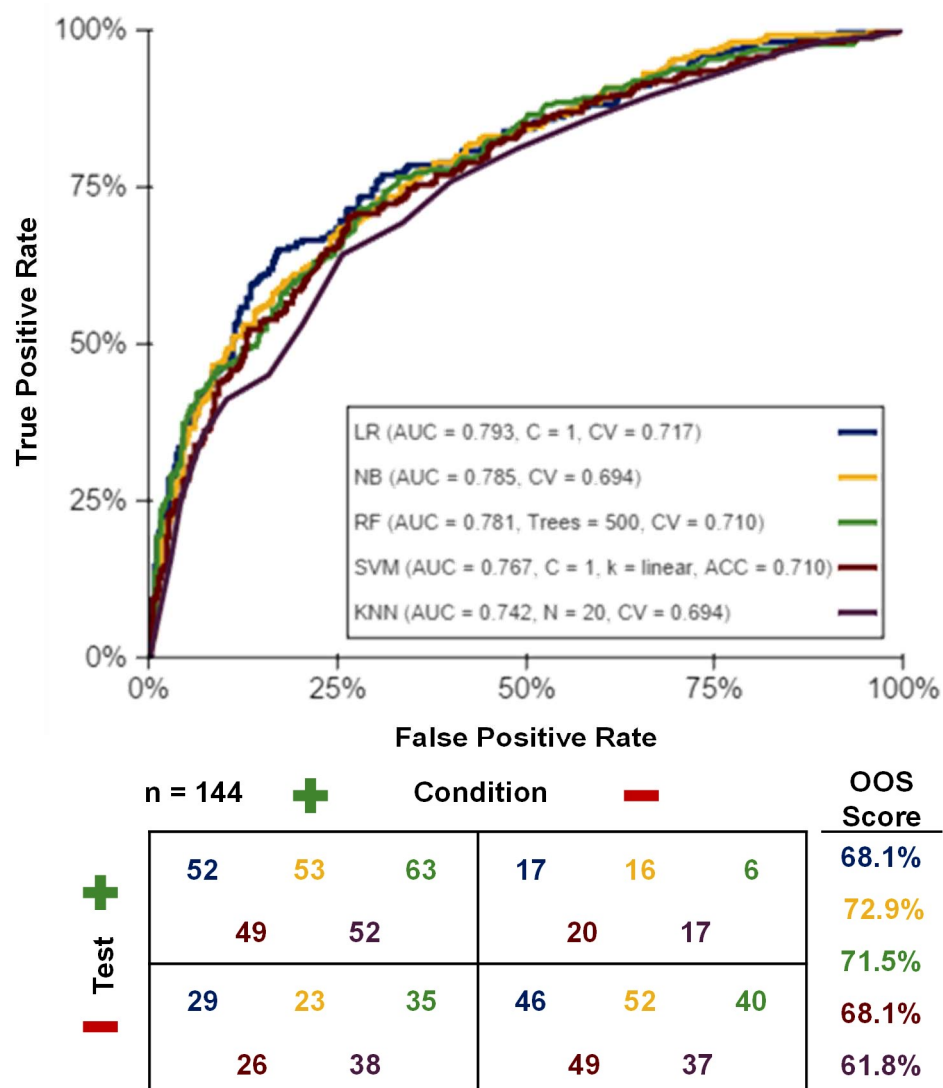


Figure 5. ROC curves and OOS predictions for movie sentiment classifiers using initial parameters.

Based on the Area Under the Curve (AUC) for each classifier, Logistic Regression was most effective but performed poorly on OOS data. Naïve Bayes and Random Forest performed the best on OOS data, and no cross-validation score was generated for SVM due to the computationally expensive performance of the model on the feature space.

Plots of Logistic Regression, Random Forest, and SVM feature coefficients and importances were used to perform feature selection, where multiple points were tested within a range to create a drop list of features for each model. See Figure 6 for feature coefficient and importance plots used in feature selection. The absolute values of feature coefficients between 0.3 and 0.8 were tested at 0.05 increments for the Logistic Regression model, and values of feature importance between 0.00025 and 0.0015 were tested at 0.00025 increments for the Random Forest model.

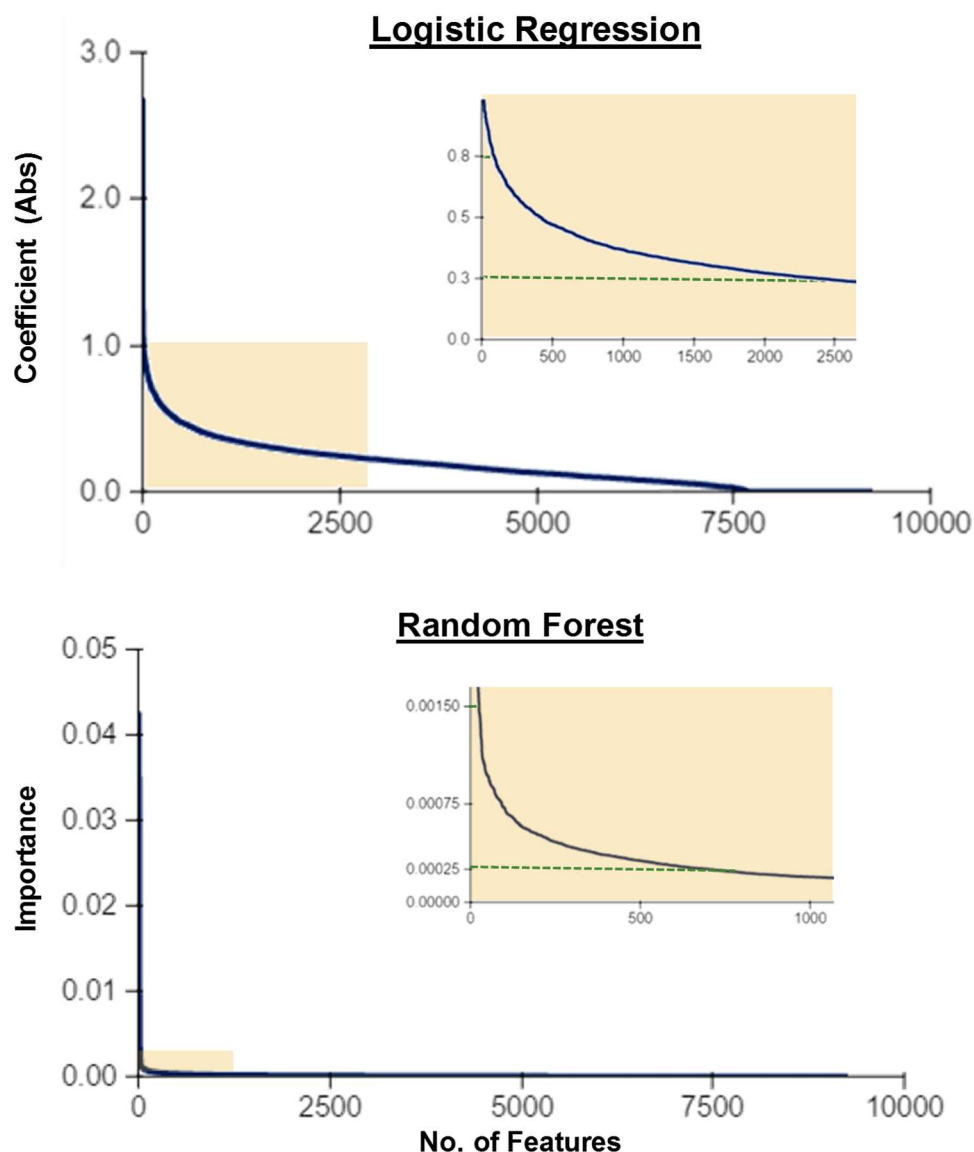


Figure 6. Feature coefficient and importance for Logistic Regression and Random Forest classifiers.

Logistic Regression, Random Forest, and SVM classifiers were trained and tested with features in the drop list removed, and the model producing the maximum cross-validation score was chosen and used to predict OOS data. ROC curves and predictions were generated using these “optimized” classifiers and are shown in Figure 7.

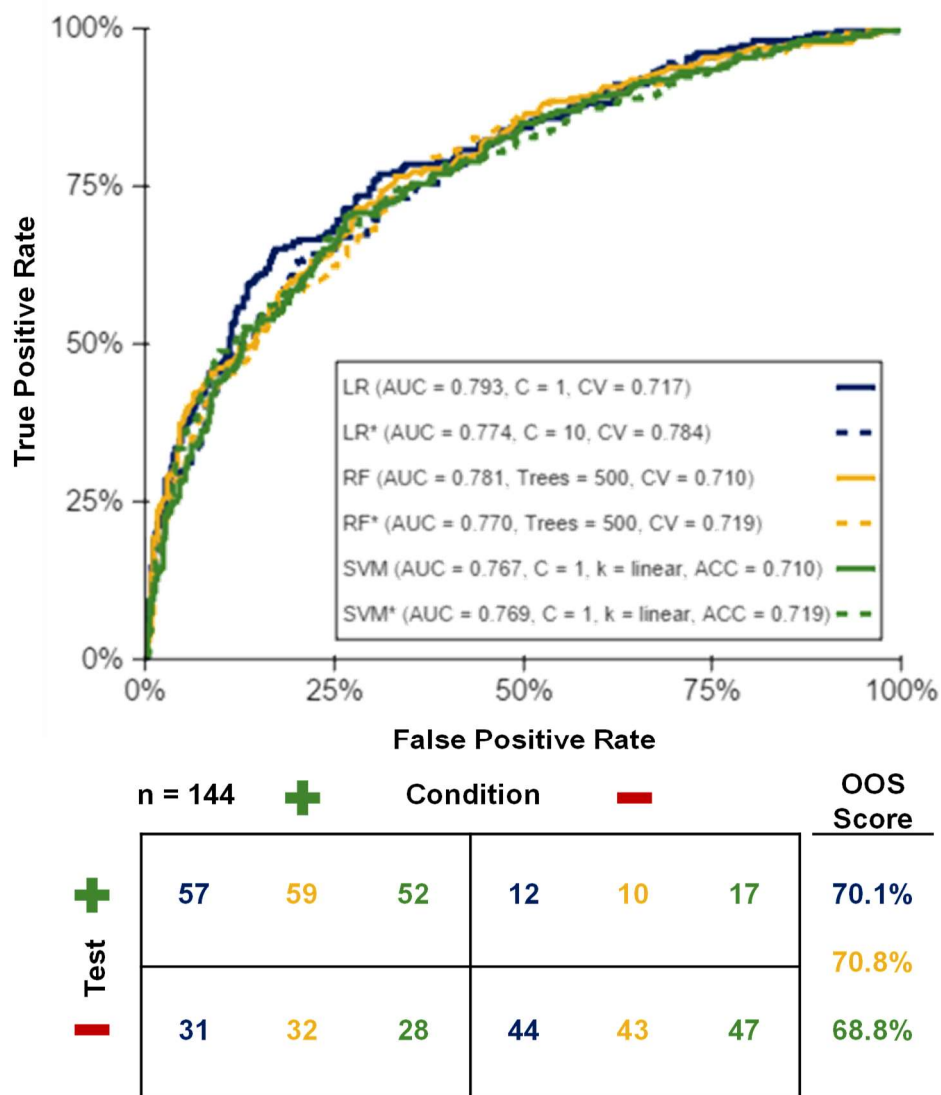


Figure 7. ROC curves and OOS predictions for movie sentiment classifiers using optimized parameters.

Based on Figure 7, the AUC for Logistic Regression slightly decreased, with the cross-validation score notably higher from 71.7% to 78.4%. Similarly, the AUC for Random Forest slightly decreased, with a modest improvement in cross-validation score. AUC and the test set prediction for SVM both slightly increased.

Comparing classifiers using initial and “optimized” parameters points to an increased ability for Logistic Regression, Random Forest, and SVM to generalize. Cross-validation scores improved from 71.7% to 78.4%, 71.0% to 71.9%, and 71.0% to 71.9%, respectively. However, given the nature of the problem to

see good movies and avoid bad ones, selection of an appropriate classifier may be based on measures of Precision (True Positives / Test Positives) and Specificity (True Negatives / Test Negatives). In calculating these metrics across all classifiers, Random Forest performed significantly above the next highest classifier Logistic Regression – 91.3% and 87.0% compared to 82.6% and 78.6%.

Predicting Movie Sentiment Based on Plot Summary

Predicting movie sentiment based on plot summary required the OMDb and Rotten Tomatoes joined dataset. Movie sentiment, the target, is defined as those receiving a fresh or rotten rating from Rotten Tomatoes. Plot summary required further pre-processing. Each plot summary removed punctuation, numbers, and stop words. Stop words are those frequently occurring but without significant meaning. Remaining, meaningful words were then stemmed and tokenized using a word count limit of 5000, and the dataset separated into training and test data comprised of movies from 2000 – 2014, with 2015 movies reserved for OOS testing.

Naïve Bayes and Random Forest were both employed to generate predictions, where Naïve Bayes performed better on both test and OOS data. Cross-validation and OOS prediction scores for Naïve Bayes and Random Forest were 60.7% / 61.3% and 60.1% / 60.5%. Token ratios for Fresh:Rotten are shown in Table 4. Based on these scores, neither classifier performed particularly well. Further inspection of words after stemming or the use of fuller, descriptive plots may be helpful in developing a more predictive vocabulary moving forward.

| Top / Bottom Token Ratios (OOS) | | | |
|---------------------------------|-------|--------|-----|
| Token | Fresh | Rotten | |
| Age | 7 | 1 | 7:1 |
| Park | 6 | 1 | 6:1 |
| Border | 6 | 1 | 6:1 |
| Known | 6 | 1 | 6:1 |
| Revolut | 6 | 1 | 6:1 |
| Feel | 1 | 5 | 1:5 |
| Rescu | 1 | 6 | 1:6 |
| Maria | 1 | 6 | 1:6 |
| Assassin | 1 | 6 | 1:6 |
| Special | 1 | 6 | 1:6 |

Table 4. Top and bottom token ratios for Fresh:Rotten.

Summary of Key Learnings

Using Linear Regression to predict Box Office results may be suitable for a majority of movies but significantly underfitted a large proportion of them as well. This may be due to a number of reasons. Perhaps, the feature space was too narrow and considering one inclusive of attributes such as ratings, genre, director, cast, and production budget may be more predictive. Additionally, plots of key variables point to linear-to-exponential relationships between Box Office results and the feature space. Moving forward, exploring non-linear relationships may correct for underfitting and yield more predictive results across a greater proportion of movies.

Binary classification of movie sentiment seemed to be a less challenging problem, supported by the performance of each classifier, particularly Logistic Regression, Random Forest, and SVM. These classifiers demonstrated the best ability to generalize based on cross-validation score, though Naïve Bayes performed the best on OOS data. Further, recursively removing / selecting features from the Random Forest model had little change on its performance. Slight shifts in rank may have occurred due to order of processing and feature importance weights were simply spread across fewer variables. Given the nature of the classification problem to see good movies and avoid bad ones, Precision and Specificity may be the most appropriate measures of classifier performance. Based on these measures, the Random Forest classifier performed the best in predicting movie sentiment given this feature space.

Naïve Bayes and Random Forest classifiers both performed poorly in using the plot summary to predict movie sentiment. Much more work would need to be done in examining the stemming of words, and it may be helpful to use full plot descriptions. However, plot summaries may not include the unique vocabulary necessary to separate a good movie from a bad one. Exploring movie sentiment based on user reviews may be more interesting and predictive.

Successfully implementing these supervised learning methods can have very real-world implications, specifically for the movie industry. Accurately predicting Box Office results and sentiment may be helpful in movie investment and prioritization decisions. Studios can focus on releasing and marketing highly profitable movies consumers enjoy.