International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Comparative Study of Various Classification and Regression Techniques on Movie Datasets Using R

Mareena Fernandes[1], Robin Lobo[2]

Student, Fr. Conceicao Rodrigues College of Engineering

Abstract

The entertainment industry has been exponentially growing since its inception and is primarily well-received. Creating content, in this case, Movies has been the bread and butter of the industry and has proven extremely lucrative over the years.

In this comparative study, we have picked out various Machine Learning models, viz. Classification Tree, Regression Tree, AdaBoost, SVM (Linear Kernel, Polynomial Kernel, Radial Kernel), XGBoost, Random Forest, Simple Bagging, and Naive Bayes. The Classification models are used to predict the rate of a certain movie winning the Oscar whereas the Regression model implementation estimates the average earning of a film considering the various attributes.

Keywords: Classification, Regression, Machine Learning, Classification Tree, Regression Tree, SVM (Support Vector Machine), Linear Kernel, Polynomial Kernel, Radial Kernel, AdaBoost, XGBoost, Random Forest, Simple Bagging, Naive Bayes, MAE, MSE

I. INTRODUCTION

The movie universe is vast, and numerous factors go into crafting a good film. The main goal of this study is to implement multiple Machine Learning techniques like the Classification and Regression models to evaluate the accuracy of addressing the problem. Another goal is to distinguish between various Machine Learning processes such as bagging, boosting, and support vectors. For this study, we have trained the dataset using multiple Machine Learning models, after which, testing of these models was performed using the testing dataset. These ML model implementations, testing, and evaluation aid in improving the models' accuracy and employ different principles like bagging and boosting to implement prediction models to understand the dataset and trends better. The study helps us compare various techniques and determine the accuracy of a particular prediction model results for certain analyses. Also, the comparison between the kernels associated with the Support Vector Machine (SVM), that is the linear, polynomial, and radial kernels was implemented to determine the most suitable model based on point alignment.

In this paper, we have covered the Classification and Regression strategies to understand the dataset better and draw conclusions. The implementation of Classification models was done by taking into account all features of the dataset and predicting whether or not the film would receive an Oscar. The Regression model implementation covered all the elements of the dataset to predict the average earnings of a certain movie when the release was initiated. The results generated by these 2 Machine Learning models would aid in the decision-making process of the people involved in movie production taking into account all the attributes.

II. METHODOLOGY

A. Naive Bayes

These are a family of elementary stochastic classifiers that involves Bayes' theorem and strongly infers independence between attributes or variables by measuring the conditional probabilities. The naive method presupposes the occurrence of a particular characteristic that is unrelated to other features. It is proven to outperform even the most advanced classification methods, and it is a simple-to-build model that is especially beneficial for very big data sets.

B. Decision Trees in R

Recursive segmentation is supported by the R package "rpart," which enables the creation and visualization of decision trees. It allows to regulate the growth of decision trees by specifying formulas, minimum splitting parameters, tree depth, and so on. The model to be utilized for the described data science challenge is specified by the method property.

1. **Classification Tree:**The decision tree performs classification when the method property is set to "class." The outputs are class levels, each with its unique identifier. The results, in simple terms, are fixed and discrete, such as "Yes/No" or "Summer/Winter/Monsoon," and so on.

2. **Regression Tree:**When the method parameter is set to "anova," a regression tree is constructed. The outputs are based on the independent variable's lowest and highest values in the data set. In other words, outputs are continuous numbers that fall within a defined range, such as the average temperature (28°C to 40°C) or an industry's revenue.

Overfitting contributes to poor performance by increasing cross-validated errors. A tree is pruned to its smallest cross-validated error value in this situation by selecting the complexity parameter with the lowest error.

C. Bagging

One of the ensemble learning methods for minimizing variance from imbalanced datasets. Since it combines the prediction results using aggregation and bootstrapping or sampling with replacements, it is also referred to as bootstrap aggregation. Problems involving classification and regression can both be resolved with bagging.

1. **Random Forest:**This method is contained in the R package randomForest and encompasses both classification and regression. Since this decorrelates the trees with the addition of partitioning on a random subset of properties, random forest outperforms bagging. The final output is the result that corresponds to each observation the most frequently. Each tree receives a new observation, and each categorization model receives a majority vote.

D. Boosting

1. **AdaBoost***:* The weights are redistributed to each instance, with higher weights given to instances that were mistakenly classified, hence the name "adaptive boosting." It operates under the notion that

learners advance in stages. Each learner after the first is developed from a prior one, except for the first. Simply said, weak learners are transformed into strong ones.

2. **XGBoost**: Here, the decision trees are generated sequentially in this approach. Each independent variable is given weight before being fed into the decision tree that forecasts outcomes. Variables that the tree incorrectly predicted are given more weight before being placed into the second decision tree. These different predictors are put together to make a strong and accurate model. This model can be used to solve various problems like figuring out values, sorting things into groups, deciding orders, and making personalized predictions.

E. Support Vector Machine

The aim of the SVM method is to discover a special line in a space with many dimensions that sorts the data points. The size of this line is influenced by how many different things we're looking at. When there are only two things to consider, this line acts like a regular line. The hyperplane transforms into a 2-D plane when there are three input features, and so on.

1. **Linear Kernel:** When the data can be split using a single line, or when it is linearly separable, a linear kernel is used. It is typically employed when a given data set has a significant number of features.

2. **Polynomial Kernel:** A kernel function called the polynomial kernel, which is frequently used with support vector machines (SVMs) and other kernelized models, reflects how similar vectors (training samples) in feature space are to polynomials of the original variables. This enables the learning of non-linear models. The polynomial kernel examines combinations of these properties in addition to the supplied features of input samples to assess how similar they are.

3. **Radial Kernel:** A well-liked kernel function utilized in many kernelized learning techniques is the radial basis function kernel or RBF kernel. This is not feasible with our dataset because of limited data.

III. RESULTS & DISCUSSION

A. Analysis of the dataset

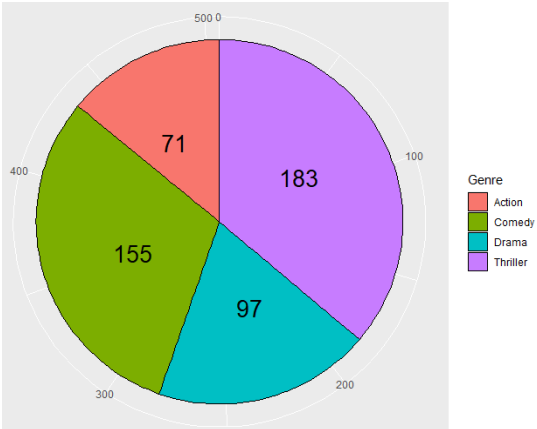

Fig. 1: Pie Chart for Movie Genres

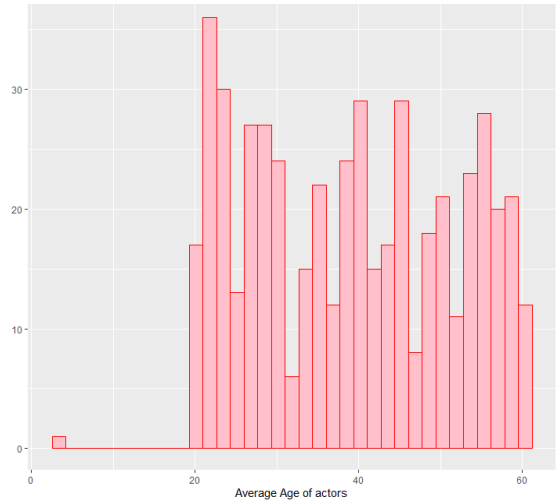

Fig. 2: Bargraph for actors' average age

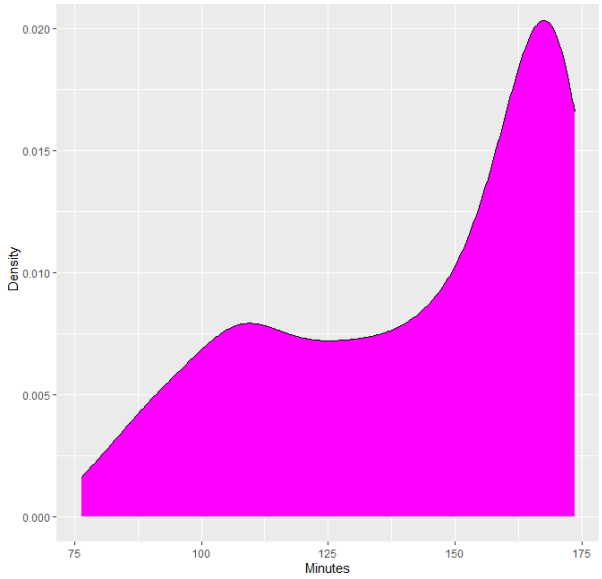

Fig. 3: Distribution of Movies Length in minutes

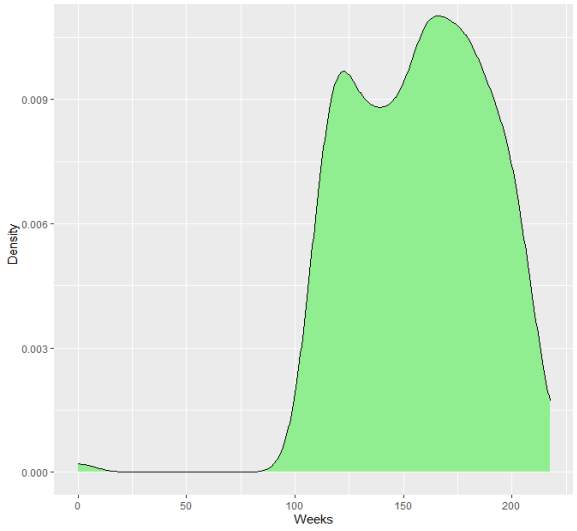

Fig. 4: Distribution of Time taken for movie in weeks

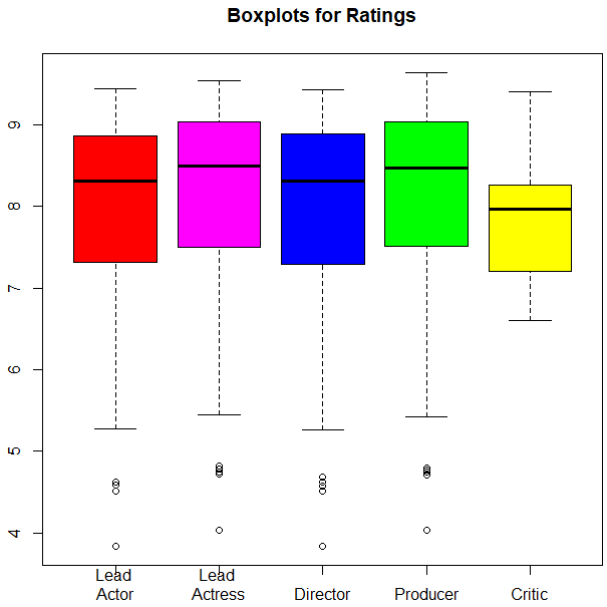

Fig. 5: Boxplots for Ratings

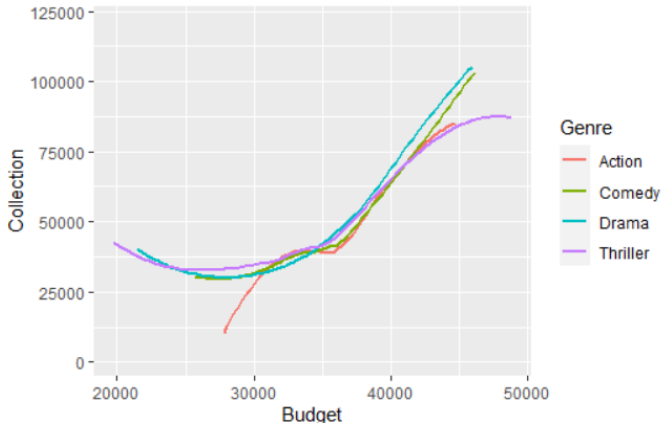

Fig. 6: Line graph for movie Budget vs Collection

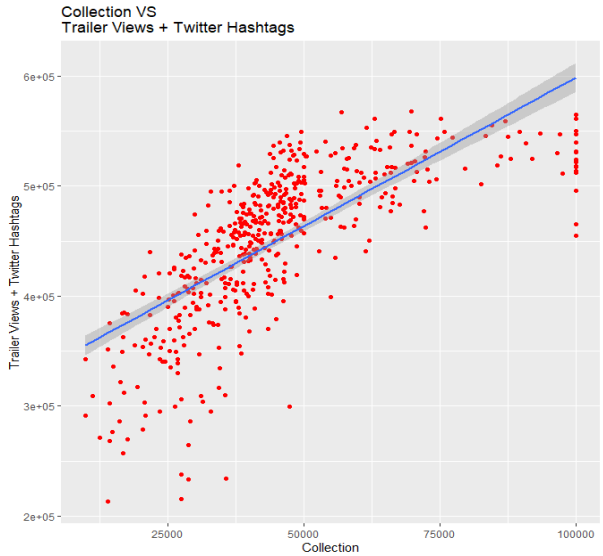

Fig. 7: Support Vector - Linear Kernel Graph

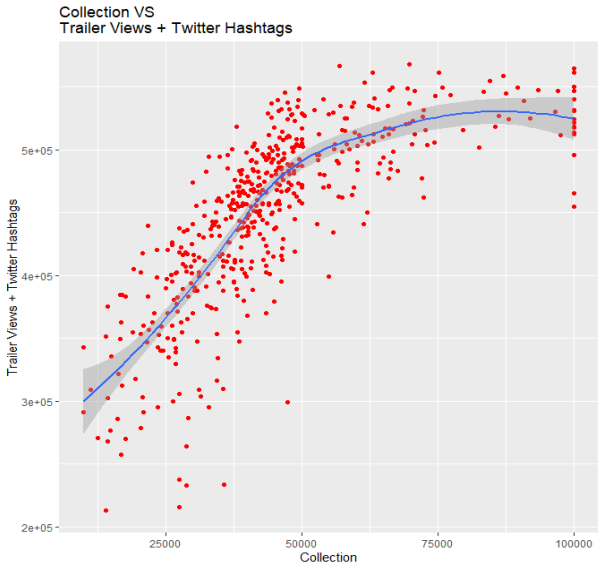

Fig. 8: Support Vector - Polynomial Kernel

B. Classification Tree

Bagging: Random Forest

The below diagram represents the unpruned classification tree up to level 3. The tree is built using R-part and plotted using R-plot libraries.

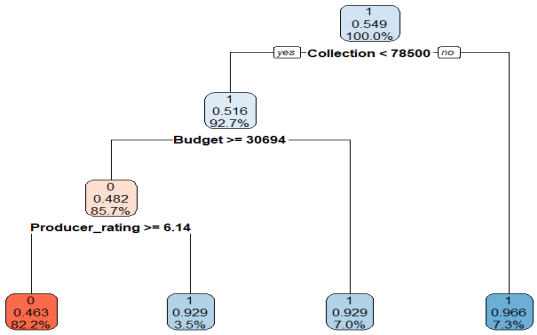

Fig. 9: Bagging Tree

Boosting: AdaBoost

The classification tree built by the AdaBoost algorithm is as follows.

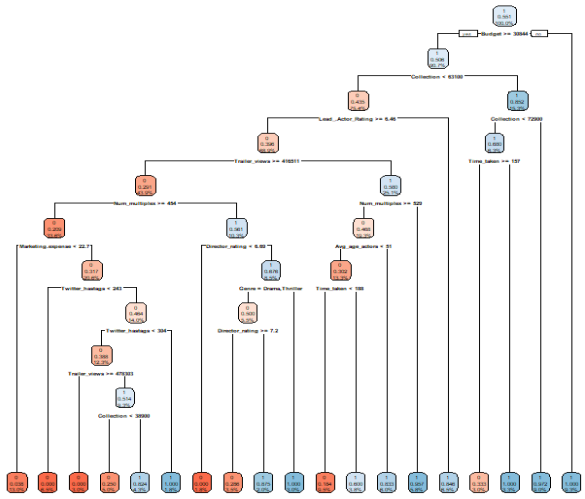

Fig. 10: Boosting Tree

C. Regression Tree

The below diagram represents the full regression tree.

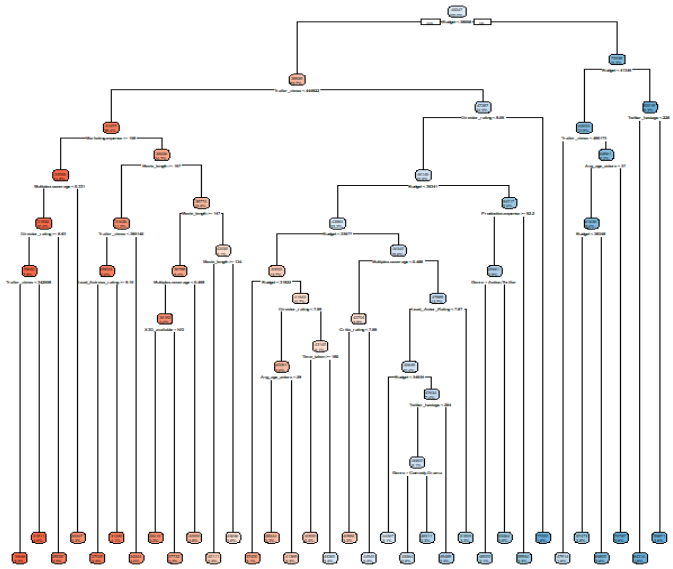

Fig. 11: Full Regression Tree

In order to prune the tree, a complexity parameter is chosen at which the cross-validation error is discovered to be on the border between increasing and decreasing values, halting the growth of the tree at that point.

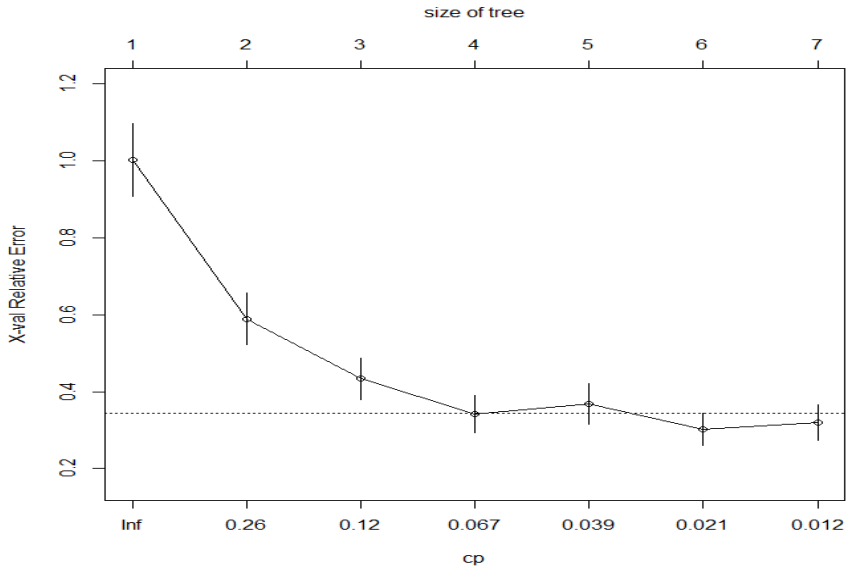

Fig. 12: Complexity Factor Graph

The below regression tree is pruned up to 3 levels to develop the resultant model as follows producing higher accuracy.

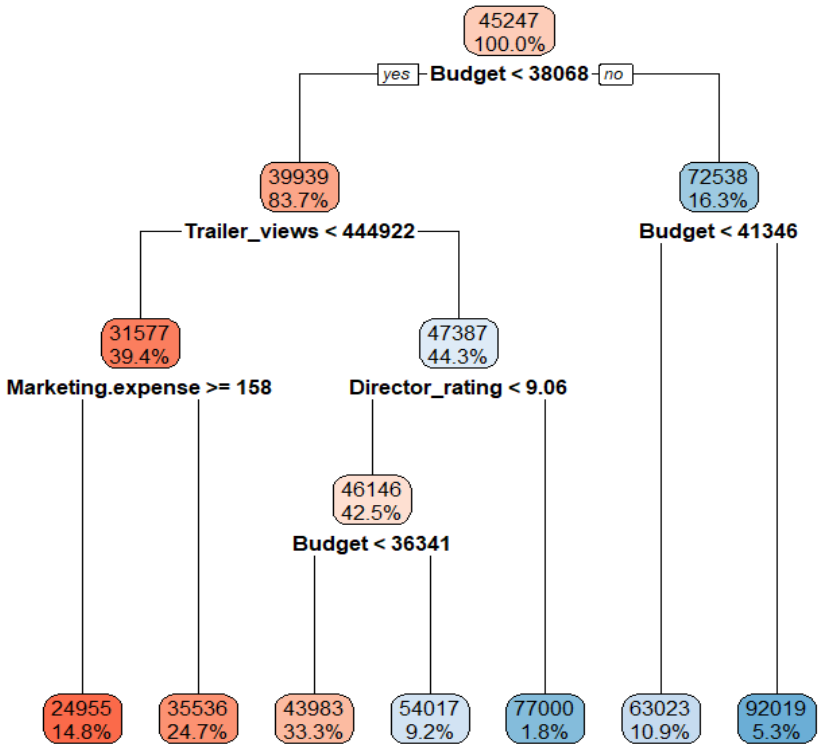

Fig. 13: Pruned Regression Tree (level 3)

D. Accuracy Report

	Methods	Accuracies
1	Clasification Tree	0.5981308
2	Naive Bayes	0.5794393
3	Bagging	0.6822430
4	Random Forest	0.6355140
5	AdaBoost	0.6355140
6	XGBoost	0.6448598
7	SVM (Linear Kernel)	0.5700935
8	SVM (Polynomial Kernel)	0.6074766
9	SVM (Radial Kernel)	0.5420561

Table 1: Classification Results

	Methods	MAE	MSE
1	Regression Tree (Full)	99037642	6779.264
2	Regression Tree (Pruned)	93731913	7035.959
3	Bagging	36577222	4320.672
4	Random Forest	37853148	4304.978
5	SVM (Linear Kernel)	102267174	7259.349
6	SVM (Polynomial Kernel)	68981513	6098.068
7	SVM (Radial Kernel)	351728647	13431.342

Table 2: Regression Results

IV. FUTURE WORK

In this study, we have only considered the Classification and Regression models but for further in-depth comparison, we can consider implementation using the Neural networks. Using Neural Networks could be an advantage with respect to distributed memory and learning.

To yield better results, we can implement sentiment analysis by collecting sentiments from social media platforms like Twitter, Reddit, Facebook, Instagram, and others. Adding a direct perspective from the audience will help create well-aimed predictions.

The dataset used in this research was retrieved from a datastore. To create a precise dataset, as a future effort, we could perform IMDB scraping which would help in producing more accurate results. Also, scraping the reviews from IMDB would cater to the possibility of implementing sentiment analysis.

V. CONCLUSION

A comparative study for multiple Classification and Regression techniques under the broader domain of Machine Learning was successfully conducted. The results of the study revealed that basic bagging with random forest produced improved classification and regression results. However, the XGBoost approach was discovered to be the greatest fit for the boosting challenge. Apart from that, the SVM (Support Vector Machine) model's linear kernel produced the best classification accuracy compared to the polynomial and radial kernels. The polynomial kernel of SVM produced the best results for regression. The outcomes of this study will assist filmmakers, directors, producers, and others working in the film industry in deciding on a variety of elements for producing films that can potentially aid them in producing successful films in the future.

VI. REFERENCE

1. Kharb, Latika & Chahal, Deepak & Vagisha,(2020). “Forecasting Movie Rating Through Data Analytics.” DOI: 10.1007/978-981-15-5830-6_21.
2. Kim, Jong-Min, Leixin Xia, Iksuk Kim, Seungjoo Lee, and Keon-Hyung Lee. 2020. "Finding Nemo: Predicting Movie Performances by Machine Learning Methods" Journal of Risk and Financial Management 13, no. 5:93. DOI: 10.3390/jrfm13050093
3. Sourabh S Kulkarni, M. Vaishnavi, Kusuma H, 2020. “Predicting the Conceptual Appeal of Movies using Data Analytics", INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 02, DOI: 10.17577/IJERTV9IS020006