Group 7 Project - High Impact Keywords that Maximize Profit

1st Brown
Computer Science Dept. of
Stevens Institute of Technology
Maplewood, New Jersey
ebrown1@stevens.edu

2nd Moore
Computer Science Dept. of
Stevens Institute of Technology
Madison, Wisconsin
hmoore@stevens.edu

3rd Powell
Computer Science Dept. of
Stevens Institute of Technology
Gettysburg, Pennsylvania
tpowell2@stevens.edu

***Abstract*—This project aims to develop a movie ad campaign that maximizes profit for an advertising company by determining which keywords have the greatest impact on a movie's profitability. The study uses over 10,000 data points and limited features to prevent false positives. The research focuses on how the keywords associated with each movie correlate with profit to make predictions of which movies are most likely to be profitable. Four algorithms were used to correlate key words with profit including a decision tree, random forest, neural network, and linear regression.**

I. Introduction

This project will solve a problem for an advertising company that wants to create a new movie ad campaign that maximizes the profit for their client. We will determine which keywords most heavily impact the profit made by a movie.

Similar projects have undertaken this effort, including the Adhikari Project from 2020 and the Galvão and Henriques' Project from 2018. Both models yielded decent results with the Adhikari Project noting their findings were limited by having a sample size of only 3,000 approximately. Our data size is significantly higher at over 10,000. Additionally, we are specifically using limited features in our model so as to not over complicate the model and prevent a high rate of false positives, which may incorrectly encourage investment in poor performing projects, which would have the greatest impact on our clients' interest.

Our project will focus on how the keywords associated with each movie correlate with profit to make predictions of which movies are most likely to be profitable. In pre-processing our data, we ensured that the maximum number of data points were maintained to allow for a sufficient data size.

In this report, we explore the use of various machine learning algorithms to identify keywords associated with highly profitable movies. We consider several algorithms including decision tree, random forest, neural network, and linear regression.

These algorithms were chosen because they are effective in handling both categorical and numerical data, and can identify complex relationships between variables, which is useful in predicting the profitability of a movie. Furthermore, the feature importance output of these models provides insight into the most influential keywords in the prediction of a movie's profitability. By analyzing and comparing the results of these different algorithms, we hope to identify the most effective approach for predicting movie profitability based on its associated keywords.

II. Related Work

Efforts have been made to predict the financial success of a movies revenue. Adhikari (2020) attempted to predict the success in revenue of a new movie based on few given attributes. In his study, Adhikari split the data in two sections, featured variable and prediction/response variable, and implemented Linear Regression, Random Forest, and Gradient Boost Regressor models to identify patterns and relationships between the data.

The featured variable encompassed all variables except for revenue and log revenue, which were employed as the predicted variables. Solely numerical data was utilized for the models, and any instances of null or zero values were substituted with the median value of the columns data. Due to the limited size of the data set used, the projected revenue for the study was inaccurate. Adhikari notes that the accuracy of the model can be enhanced by incorporating additional data.

Galvão and Henriques' (2018) attempt at "Forecasting Movie Box Office Profitability" involved constructing a predictive model that utilized data mining techniques. Neural Networks, Regression, and Decision Trees were the data mining techniques implemented into the predictive model. Their focused data set contained a specific set of movies using historical data and specific variables. The study tested several factors, such as data partitioning, outlier treatment, and dimensionality reduction to increase the accuracy of the model.

Three predictive models were tested for each combination of the previously listed factors; MLP, Multiple Regression and Decision Tree. The MLP with three hidden layer neurons had the best performance. The model was trained on 70% of the data and tested on 30%. While the empirical model yielded good statistical results for binary and interval dependent variables, the multi-class prediction results were significantly different from reality, negatively impacting the model. Among the employed methodologies, neural networks demonstrated the highest predictive power, with a significant advantage in all three developed models.

Adhikari’s technique implemented three models, Linear Regression, Random Forest, and Gradient Boost Regressor, while Galvão and Henriques’ technique involved Neural Networks, Regression, and Decision trees. In comparison, Adhikari’s attempt was unsuccessful while Galvão and Henriques’ attempt was successful regarding their implementation of Neural Networks.

Although Adhikari’s attempt at the problem was overall unsuccessful, there was some success in his approach when comparing the results of his models, such as the Gradient Boost Regressor Model having an R square of 0.67, the highest R squared value of the three implemented models. A con from his approach was that the data set used only consisted of 3000 rows, leading to limitations in his predictive model, which in turn negatively affected the R squared predictions for the Random Forest and Linear Regression models, being 0.56 and 0.62 respectively.

Some pros from Galvão and Henriques’ study were that they used previous work to positively influence their study, used variable elimination to improve prediction accuracy, and implemented the MLP Neural Network which is known to be highly successful in predictive and classification problems.

A few cons that Galvão and Henriques encountered were that some of their predictive models exhibited error rates higher than they desired and that they had a high number of variables that were insignificant to the model to start. They concluded that their study would have had more success had the variables in the data set been evaluated for correlation to movie revenue, and substituted with variables that had higher predictive power prior to their work.

III. Our Solution

Our solution to predicting movie profitability based on keywords involved using four different machine learning models: Decision Tree, Random Forest, Neural Network, and Linear Regression. We chose these models because they are commonly used in predictive modeling and can handle both classification and regression tasks. Additionally, the models have different strengths and weaknesses, allowing us to compare their performance and choose the best approach for our specific problem.

A. Description of Dataset

This data set was sourced from Kaggle.com, and contains information about about 10,866 movies. The original features most relevant to our calculations are ‘budget_adj’ which represents the budget of the movie, ‘revenue_adj’ which represents the revenue of the movie, and ‘keywords’ which holds various words or phrases describing the movie.

The data was pre-processed in a number of ways. First, any movies with 0 in the ‘budject_adj’ and ‘revenue_adj’ features were replaced with the mean of that feature. Then, the feature ‘profit’ was added to represent the difference between ‘revenue_adj’ and ‘budget_adj’. Next, the ‘keywords’ feature was parsed into a list and placed into a new feature named ‘keyword_list’. The independent variables were set to ‘keyword_list’ and ‘original_title’, and the independent variable is ‘profit’. The data points that did not include any keywords were given the keyword ’no_value’ which was excluded from the findings.

Of particular interest at first inspection, the top twenty most common keywords are drastically different from the keywords found in the top twenty films with the highest profit. Additionally the keywords from the top 20 with the highest profit only occur 1 or 2 times in that list.

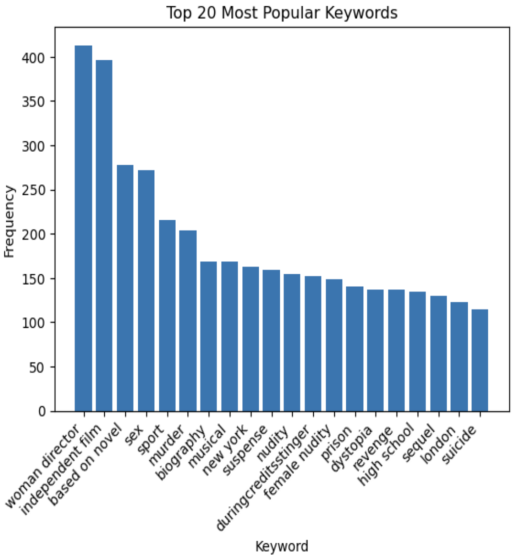


Fig. 1. This is a bar graph from our project which shows the 20 most popular keywords from our data.

B. Machine Learning Algorithms

For this project we have decided to implement four models; decision tree, random forest, neural network, and linear regression.

Decision tree and random forest are good algorithms for this problem because they can handle both categorical and numerical data, which is useful for analyzing the keywords associated with movies. Additionally, these algorithms can handle complex relationships between variables, which is important when considering how different keywords might interact to influence a movie’s profitability. Finally, decision tree and random forest models can output feature importances, which is valuable in identifying which keywords are most important in predicting high profitability for a movie.

Neural network modelling is appropriate for our project because neural networks are able to model complex, non-linear relationships between input and output data, and handle large data sets exceptionally well. They are also capable of adaptive learning from complicated data and can be used to extract

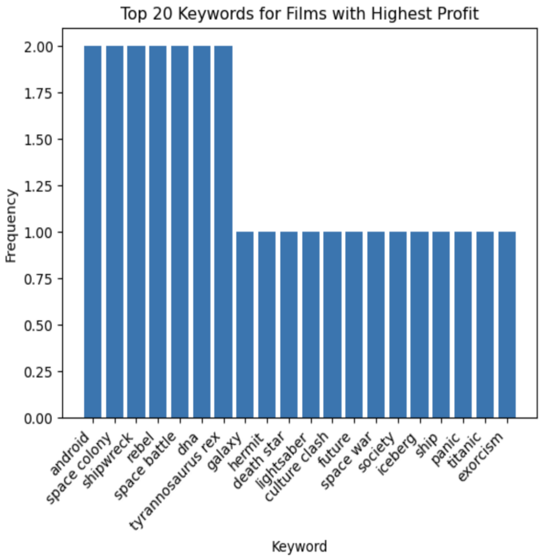


Fig. 2. This is a bar graph from our project which shows the 20 most popular keywords from the films with the highest profit in our data.

patterns and detect trends. This model will test the correlation between tagline keywords and movie profitability/success and make predictions from the data. These predictions will be tested for accuracy.

The linear regression model allows us to evaluate the impacts of one or more variables on a linear growth model. This method is appropriate for our project as we can plot the testing values against our model to fine tune our algorithm. We will use all the keywords for each movie to determine a value for the weights. Once we have created and tuned the weights, we will use them against our modeling data to create predictions of how each movie will do. Once we have the model data plot, we can analyze and adjust our weight values accordingly.

C. *Implementation Details*

Within our implementation we have split our data set into two sections. 80% of the data is used for testing and training the algorithm and 20% is used for analysis of our training's performance.

1) Decision Tree & Random Forest: Decision tree algorithms are a type of supervised learning algorithm used for classification and regression tasks. They work by recursively splitting the data into subsets based on the values of individual features until a stopping criterion is met. The resulting tree structure can be used to make predictions for new data points by traversing the tree from the root to a leaf node.

Random forest algorithms, on the other hand, are an ensemble learning method that combine multiple decision trees to improve the accuracy and reduce overfitting. They work by randomly selecting subsets of the training data and subsets of the features to use for each tree. Each tree is trained on a different subset of the data and features, and the final prediction is made by averaging the predictions of all the trees.

Both the decision tree and random forest algorithms used in the code generate feature importances as a byproduct of their training process. Feature importances represent the relative importance of each feature in the dataset for predicting the target variable. The feature importances are calculated based on how much each feature contributes to the reduction in impurity or variance when splitting the data.

The code extracts the feature importances from the trained classifier using the ’feature_importances_’ attribute. The feature importances are then stored in a dictionary called ’keyword_importance_dict’, with the corresponding keywords as keys. The code then sorts the keywords in descending order of their feature importances using the sorted() method and prints the top 30 keywords with their corresponding feature importances. The keywords are filtered to exclude any with the value ‘no_value’ using an if statement.

Finally, the code predicts and prints the accuracy of the decision tree and random forest classifiers using the predict() method on the test data and calculates the accuracy using the accuracy_score() method.

2) Neural Network: Neural network model implementation will test the correlation between tagline keywords and a binary classified profit. The input layer in the neural network is comprised of raw data from the dataset that is fed into the network. In this project, our input layer will be the ”keywords” data from the data set and be represented by ’X’. Because there are multiple keywords per movie title the keywords will be separated by their delimiter, ’—’, and placed into a new table. Label encoder will be applied to the new keywords table to convert the categorical data into numerical data, given that the network requires numerical inputs. The split keywords will subsequently be merged back into their original concatenated form.

The output layer of the neural network is influenced by both the hidden layer and the weights that are applied to the model within the output layer. In this network the output layer, ’y’, will be the ”profit” data. For improved accuracy, the ”profit” data will be converted into outputs applicable for binary classification. To accomplish this, the ”profit” data will be given a threshold that will be converted into binary numbers 0 and 1 in a new column labeled ”success”. For the specified numerical threshold, any profit greater than or equal to the threshold will be considered ”successful” and assigned a value of 1. If the profit falls below the threshold, the movie will be categorized as ”unsuccessful” and assigned a value of 0. The data will be utilized by the neural network model to generate its predictions. If the predictions exhibit high accuracy, it can be concluded that certain tagline keywords are positively correlated with movie success/profitability.

A confusion matrix and classification report will be added to summarize performance scores on both training and testing data. The purpose of a confusion matrix is to provide a detailed assessment of a classifier's performance. From the confusion matrix, we will be provided the ”true negative”, ”false positive”, ”false negative”, and ”true positive” classification results. In summary, the accuracy of the neural network’s classification predictions will be evaluated by comparing the predicted classes with the actual or true classes to determine how many movies were accurately or inaccurately classified as ”successful” or ”unsuccessful”. The classification report will provide the performance metrics (accuracy, precision, recall, and F1-score) of the confusion matrix.

3) Linear Regression: Our weights are based on the performance of the movie as a whole and the weights of each factor are then adjusted based on this. If a movie’s profit is ¡ 0, the items in the weight are deducted a tenth from their initial score of 1. For example if the key word ”Sharks” is in a movie that loses money, the word’s weight as a keyword then becomes .9 as we deduct .1 from its starting score of 1.0. If a movie makes up to a profit similar to the mean of the profit category, the weights make no change, as the movie performed average. If a movie performs very well and generates more than mean of the profit category, the weight will increase by .1 percent. For example, if ”Sharks” is in a movie that had initial investment of 100 million dollars and the movie makes 200 million dollars, ”Sharks” will now have a score of 1.1. This is found by adding .1 to its initial score of 1.0.

This model of training is applied to each of the keywords associated to a movie, to determine if they are performing well and what key words a director should include to maximize their chance at generating the most revenue. Once these scores are applied and we use them for our testing data, we apply the average of each key word to our model. So if a movie has keywords with the weights 1.3, .8, and 1.0, the keyword score for the movie would be 1.03. We use this weight for our initial prediction and the begin to fit the models and the weights of each key word based on the margin of error we found at the end of the plotting.

After completing the analysis of the training data set, we attempt to find a margin of weight for keywords and an algorithm of analysis the program can complete to maximize the accuracy of the reports for each film. For example if one item of the training has a margin of error of 8 million dollars, our goal is to find a weight for the keywords where the margin for error is minimized. Finding an algorithm to properly fit the keyword weights is our current highest priority to fit the data properly into the linear regression model and to get accurate predictions.

D. Comparison

In this project, we utilized four machine learning models to see the if the profitability of a movie could be predicted based on the movie’s keywords. The models used were Decision Tree, Random Forest, Neural Network, and Linear Regression.

The accuracy for both Decision Tree and Random Forest models was found to be 0.34 using the accuracy_score from sklearn.metrics. The Neural Network model had an accuracy of 0.69, with a precision of 0.79 and recall of 0.79 for class 0, and a precision of 0.38 and recall of 0.39 for class 1. The Linear Regression model had a mean squared error (MSE) of 2.47 x 101̂6.

Overall, the results of the models were not very promising as none of the models were able to accurately predict the profitability of keywords. This suggests that there is no correlation between the given keywords and their profitability. One possible reason for this could be that the dataset used in

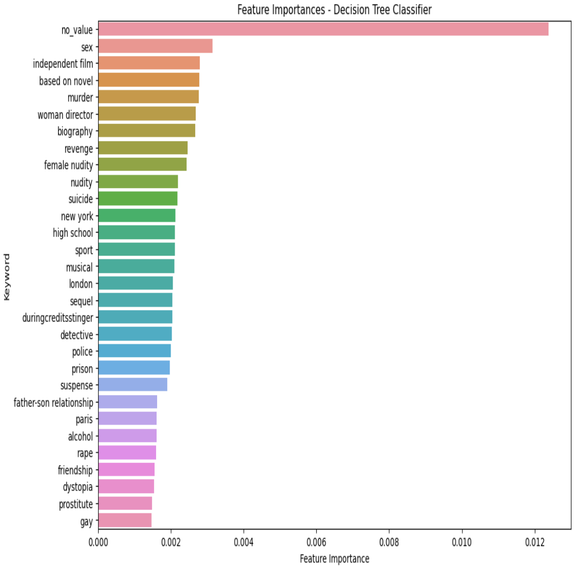


Fig. 3. This bar graph shows the feature importances for the top 30 keywords found in the Decision Tree model.

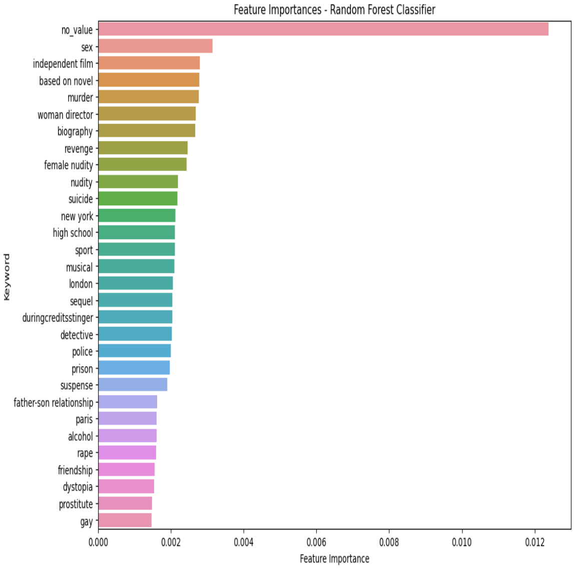
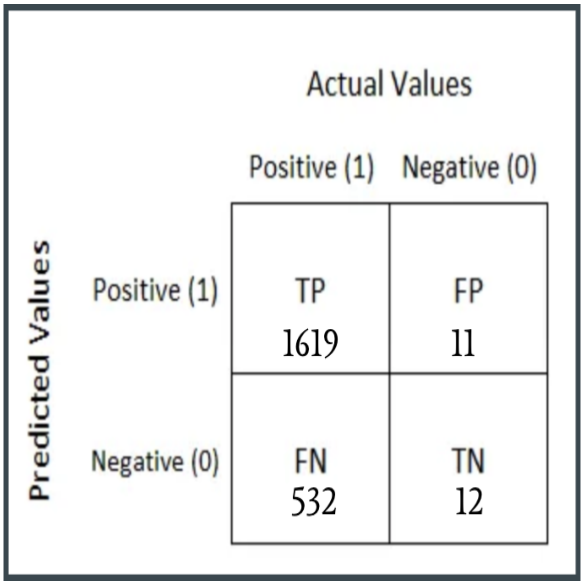


Fig. 4. This bar graph shows the feature importances for the top 30 keywords found in the Random Forest Model.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP 1619	FP 11
	Negative (0)	FN 532	TN 12

Fig. 5. This is a confusion matrix shows the accuracy of the Neural Network.

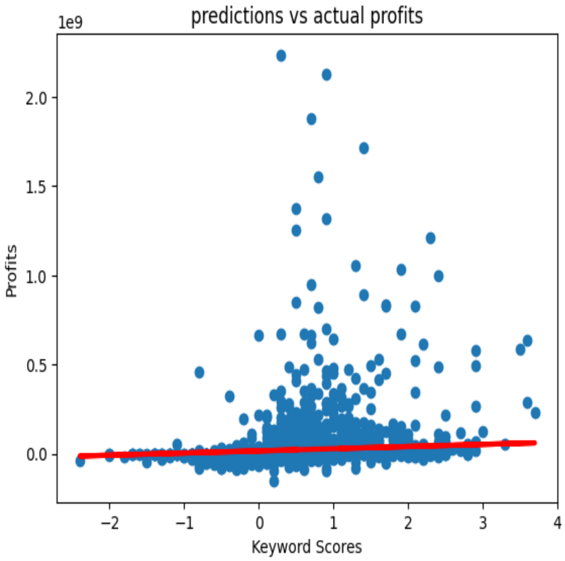


Fig. 6. This graph shows the predicts vs. profits for the Linear Regression model.

this project was not representative of the entire population. It is possible that the dataset did not include enough diverse and relevant data, leading to inaccurate results, or that there was too much inconsistency in the keywords and quality of that aspect of the data. Additionally, it could be that the features used in the models were not able to capture the complexity of the problem, resulting in poor predictions.

It is also worth noting that the Linear Regression model had a very high MSE, which indicates that the model was not a good fit for the data. This further supports the idea that the features used were not able to accurately capture the relationship between keywords and profitability.

Overall, the findings suggest that predicting the profitability of keywords may be a challenging task, and further research may be necessary to develop more accurate models.

IV. Future Directions

Albeit having a larger data set in comparison to the related works studied, the performance of our models were still subpar. We hypothesize that the incorporation of more relevant features to the model will improve the relationship between our selected features and movie profitability. Specifically, text mining techniques could be used to extract more nuanced information from plot summaries, reviews, and other text-based data. This could provide a more detailed understanding of how specific plot elements and themes contribute to movie profitability.

If provided an additional 3-6 months, the extended time frame would allow for more extensive research into prior works on the available features in the data set, and allow us to choose a feature more likely to have a strong correlation with movie profitability. Additionally, the extra time would provide the opportunity for our team to test the other available features in the data set to find the feature that has the best linear relationship with movie profitability and prediction accuracy in our models.

V. Conclusion

Our models were built to check if there was a correlation of keywords that are used when creating a movie and the overall profit a movie will generate. We use four models to create this prediction.

The first model we used in this prediction was a Decision Tree. The decision tree splits the keywords into subsets and creates predictions based on combinations of keywords. Using these subsets, the algorithm then combines each layer's predictions to find a final result for each combination of keywords. The decision tree algorithm yielded an accuracy score of 34% which is would imply there is no correlation. The decision tree method may be too binary to solve a problem like this where the values include outliers and dynamic data combinations.

The second model used was a random forest model. A random forest model is the combination of multiple decision trees. This allows the algorithm to dynamically combine results to increase accuracy and make the results more dynamic to each instance. This model also yielded an accuracy score of 34% which also shows there is no correlation between the two values when creating a movie. Though this algorithm is more dynamic than a lone decision tree, it still seems slightly too binary to find and handle outliers accurately at the current level of layers our decision trees were implemented with. To continue with this method, we would likely need to allow for much larger trees which harshly impact performance.

After attempting the first two models, we attempted a neural network model as it would yield binary results for if the profit was higher than the threshold of 10 million, which would be regarded as a successful movie in our project. A neural network works best with binary results because of the way it combines complex layers. In this model we input the keywords and run them through the nodes of the model as an encoded value. The output is then a binary value as either ”successful” or ”unsuccessful”. This model yielded an accuracy 73%. This model’s accuracy was the most accurate model we were able to create and suggests that with proper training it may be possible to predict if a movie will do well based on the keywords associated with the movie. This model was the best performing model we had implemented and would likely do well if the nodes were trained in a complex way to yield predicted profits as numerical data opposed to binary data, but the initial findings imply the neural network is possible of somewhat accurately making these predictions.

Finally we integrated a Linear Regression model to see if there would be a linear correlation. In this model we created an algorithm to keep scores of keywords normalized and attempted to plot the regression based on this. This data normalization allow us to process the data as keyword scores directly against profit. This model yielded a mean squared error(MSE) of 2.47e+16. Being that an MSE value is more accurate the closer it is to 0, this model was not able to predict the very well. This likely occurred because the linear data set is not capable of handling outliers in any way and with a polynomial implementation may have a chance at yielded somewhat accurate predictions.

When viewing the results as a collective the decision tree, random forest, and linear regression models imply there is no correlation between the keywords used in a movie and the profit to be made by the movie. However, when we view the neural network we can assume that there is some correlation, despite the possibility of extreme outliers to predict accurately if a movie will succeed or not based on its keywords. Though our model’s were not able to predict the correlation and predict the values of profit accurately, the neural network does show potential for being able to solve this problem in the future with one of our models.

References

Adhikari, Saphal. “How to Use Machine Learning Approach to Predict Movie Box-Office Revenue / Success?” Medium, Analytics Vidhya, 1 May 2020, https://medium.com/analytics-vidhya/how-to-use-machine-learning-approach-to-predict-movie-box-office-revenue-success-e2e688669972.

Galvão, Marta & Henriques, Roberto. (2018). Forecasting Movie Box Office Profitability. Journal of Information Systems Engineering Management. 3. 10.20897/jisem/2658.