The Correlation Between Movie Rating and Movie Genre

Anonymous

1 Introduction

Nowadays, with the rapid advancement of technologies, web users highly dependent on documentation and numeric information provided on the internet. People tend to conduct a research before deciding their next step to complete a task, such as deciding which movie to watch to avoid high probability of ending up with disappointment. Movie rating is an iconic metric for viewers' reference of whether the movie is worth watching. Rating systems serve as a database by listing out details of the movie, including the title, genres it composed of, the country of the moving, the release year, and so on, and lastly the most important is to activate the rating function for users to leave comments.

Another feature for choosing a movie is according to the viewer's preferences of genres. When browsing through rating platforms, it is notable that the systems usually categorise videos based on their genres as to offer a more simple and convenient method to look up for movies promptly. Generally, users enjoy watching the movies of their preferred genres, so those genres should receive higher ratings from the watched viewers at the end. In this case, further analyse the relationship between movie genres and ratings to verify the correlation by training 2 machine learning models: decision tree classifier and random forest classifier with the training dataset, evaluate the outcome with the evaluation dataset, then make predictions on the test dataset and verify accuracy is the main purpose of this research.

2 Literature Review

Since movies are indispensable to most people's daily life, a large body of research about movie recommendation system has been implemented.

One research is reviewing the accuracy of 2 algorithms that have been used in the previous works: user-based collaborative filtering and item-based collaborative filtering with the MovieLens dataset. The proposed algorithm in the research measures the correlation between movie genres with movie rated scores and utilize the correlation value to cluster the movie. Then, the proposed algorithm makes prediction of the movie ratings using the traditional collaborative filtering algorithms. By calculating the correlation value, Pearson correlation coefficient and k-nearest are implemented in the research to obtain the correlation between each genre and best combination of movie genres are correspondent to the target users' preferences. Finally, the experimental results indicate that the proposed algorithm yields higher accuracy in movie rating predictions than existing collaborative filtering algorithms. However, there are some adjustments that could be further discussed such as the reliability of genre correlation, the lower accuracy of prediction by the proposed algorithm when the weights of genres are applied to the rating prediction, and also the inconsistency of dataset usage (Hwang, Park, Hong & Kim, 2016).

The other research is different from the first research as the concept of it is the other way around. It discussed about predicting movie genre by treating movie rating as an attribute, applying the Bernoulli event model to estimate the likelihood and evaluate the probability of the genre using the Naive Bayes rule. Similar as the previous research, the MovieLens dataset is utilized in the training process with random selection of train and test approaches, the Bernoulli model calculates the conditional probability of movie that is categorized as a certain genre, which is treated as a binary feature vector. Once making sure that the movie belongs to a particular genre, it calculates the posterior probability of a genre to provide a hint on a users' preference model towards movie genres. The result of the research appears that as the training size increases, the accuracy of predicting movie genres increases. In addition, it is unexpected that the Bernoulli multivariate models can achieve 50% prediction accuracy with only 10% of data. According to the 2 outcomes, the research proves that recommendation system probably works more efficiently based on categories than ratings. Nevertheless, the

experiment showed that the behaviour of higher ratings does not match with the lower ratings, which can be analyse in depth in the future (Makita & Lenskiy, 2016).

From the 2 related research, they primarily focus whether the correlation between genre can lead to better recommendation system than the traditional collaborative filtering algorithms implemented within the system. However, the relationship between genre and movie rating hasn't been analysed directly and explicitly. In this research, since genre correlation can predict better movie rating, assuming that genre is correlated with movie rating, training the dataset with 100k non-sparse data and 42 features to prove the assumption.

3 Method

The "TMDB_train" dataset is used for training both Decision Tree model and Random Forest model. The train dataset includes 100k rows of data and 44 features, which the target data is the "average_rate" and "rate_category", which will be the main predicting classification. After training the machine learning models, the "TMDB_eval" dataset is used for the evaluation of model accuracy performance. Lastly, making predictions on the unlabeled "TMDB_test".

3.1 Data Pre-processing

To begin with, as there are some values that are 0 in the training, evaluation and testing datasets, particularly in "budget", "revenue" and "popularity" columns, which doesn't make much sense to the dataset, so presumably they are missing values. The missing values are replaced with the mean values of the corresponding column values to align with normal distribution to avoid information bias.

In addition, the data in "title", "overview", "tagline", "production_companies" and "original_language" are raw text values, so they are converted to numeric values by TF-IDF vectorization beforehand for the model training. Then, drop the original columns and combine the dataset into separate train, evaluation, and test data frames for the experiment.

3.2 Feature Engineering

In this research, the logistic regression classifier is set as the baseline. The logistic regression classifier is an efficient model that aids in detect the existence of meaningless data in the dataset if the results are poor because it is sensitive with little to no relation variables. It works more efficiently when removing the unrelated attributes to the output variable. During the training process with the baseline logistic regression classifier, after trying different combination of attributes, I discovered that "budget" will increase noise for model training, so it is necessary to extract the column from datasets to enhance the performance of both decision tree and random forest classifiers.

3.3 Model Training

While the topic of the research is whether genres of movie correlates with movie rating, each model runs 2 experiments, one with the genre columns and the other without genre. Decision tree classifier and random forest classifier are suitable for training because both are rooted in decision-making processes. Decision trees split data based on feature values at each node, aiming to minimize impurity or maximize information gain. As utilized in the datasets, it helps clearly generalize and categorise attributes to different "rate_catagory". For random forest classifier, it generates multiple decision trees by bootstrapping the data and select random subsets of attributes at each split, which enhances the learning curve of the training process and should provide a higher quality of outcome.

3.4 Evaluation

To evaluate the performances of machine learning models, compare the predicted values with the classification of "rate_category" column in the evaluation dataset. Accuracy, precision, recall, and f-1 score are calculated for more detailed comparison.

4 Results

The accuracy of the baseline logistic regression model is 0.272 with genre columns and 0.271 without genre columns. The accuracy is relatively low, but it proves that there is a slight difference when genre is considered as an attribute.

Both decision tree and random forest training models, however, perform significant improvements of the accuracy. Based on the information provided in Table 1, the accuracy has increased by approximately 40 points comparing with the logistic model, especially the performance

of random forest classifier with genre columns training has increased to 0.688. The gap of accuracy between with genre and without genre has enlarged as well, 0.003 and 0.015 points for decision tree and random forest classifiers respectively. The results can be evident that there is a direct correlation between genre and movie rating, since the results indicate that the prediction has higher accuracy when movie genres are considered as one of the features.

Accuracy	Decision Tree	Random Forest
With genre	0.655	0.688
Without genre	0.648	0.673

Table 1- Accuracy of training models with and without genre.

To further analyse the classifiers, weighted-average of precision value, recall value and f-1 score are calculated in Tables 2, 3 and 4. Despite discernible disparities in accuracy, the differentials observed in all three evaluation metrics between classifiers trained with and without genre information are marginal. This finding underscores the necessity for deeper investigation and analysis, which will be discussed in the next part.

Precision	Decision Tree	Random Forest
With genre	0.66	0.72
Without genre	0.65	0.72

Table 2- Precision of training classifiers with and without genre.

Recall	Decision Tree	Random Forest
With genre	0.66	0.69
Without genre	0.65	0.67

Table 3- Recall of training classifiers with and without genre.

f-1 score	Decision Tree	Random Forest
With genre	0.66	0.69
Without genre	0.65	0.67

Table 4- f-1 score of training classifiers with and without genre.

5 Discussion/ Critical Analysis

According to the results, I have condensed to 3 sections for deeper analysis and discussions.

5.1 Baseline Model Performance

The accuracy of the logistic regression classifier is only 0.271 and 0.272, which is consider as underperforming. One of the reasons is not implementing data pre-processing effectively. However, as the baseline model is aid to inspect the quality of data to diminish the outliers and unrelated data in the datasets, it proves that training with all instances is inefficient. Therefore, I selected distinct combinations of features to train, then result in the budget of movie being the noisy feature. Additionally, the low accuracy of logistic regression classifier indicates that attributes and target can be non-linear related, so it does not apply to the sigmoid function that logistic regression is based on. For training models, decision tree and random forest classifiers are selected to increase the flexibility of data utilization as the assumption of independence among features isn't required in both classifiers, more suitable for the training circumstance. The models have built-in bagging to minimise overall model variance without introducing model bias.

5.2 Training Model Performance

From the tables in Results, we can notice that the accuracy of each models is higher with genre, comparing to the ones without genre, evidently shows that genre is indeed positively correlated with movie rating. Nevertheless, the other evaluation values do not show much differences between with and without genre. One rationale is that the data is imbalanced as the examples given for each classification are not equal. Based on the confusion matrixes from Figure 1 to 4, the most accurate and precise "rate_cateory" predicted is the label 3, then is label 2. It suggests that the data includes instances belonging to label 2 and 3, so the models learn and perform better by finding a certain pattern for label 2 and 3. This also indicate that there should be imbalanced classes as the data includes larger portion of features that belongs to label 2 and 3. In the next phase of the research, it is important to append instances of other labels to the dataset to increase the learning curve and help avoid overfitting so that models can be trained efficiently.

In addition, it is notable that random forest model outperforms decision tree model. One factor is that it conducts feature randomization during tree-building by considering only a random subset of features at each split. This procedure avoids overfitting, preventing

individual trees from relying on one certain pattern of features, and also encourages diversity among the trees, which leads to better performance of predicting classes.

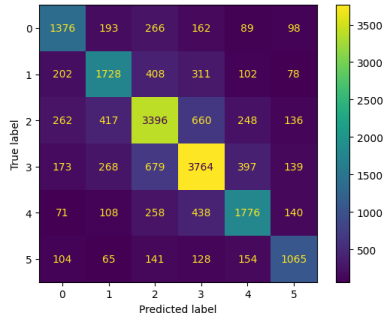

Figure 1- Confusion matrix for decision tree model training with genre

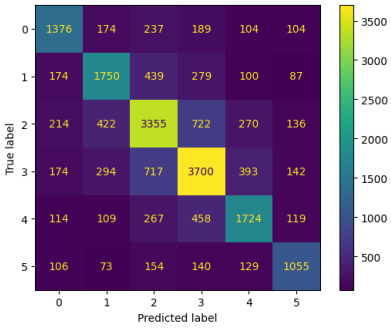

Figure 2- Confusion matrix for decision tree model training without genre

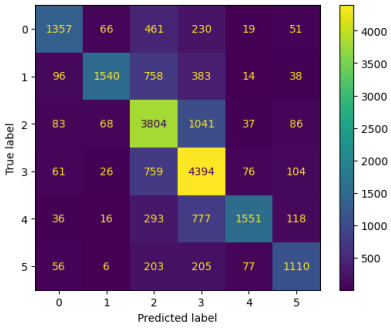

Figure 3- Confusion matrix for random forest model training with genre

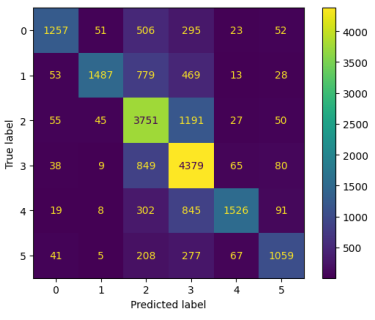

Figure 4- Confusion matrix for random forest model training without genre

Secondly, the robustness to noise of random forest classifier minimizes the impact of outliers to results. Even though feature engineering is done beforehand, it is not guaranteed that the data is completely clean for training and evaluation, random forest training will not fluctuate with unrelatable data to "rate_catagory" and provide higher accuracy predictions.

5.3 Future work

Despite the above promising results, some questions remain. Since the highest accuracy of model training is 0.688, one of the questions will be whether changing hyperparameters of the models will increase the accuracy or even increase precision of prediction. Cross validation of datasets can also be considered as one method to build a more reliable and stable model for prediction. It won't directly enhance the performance, but it provides benefits such as optimization of hyperparameters. Another key question will be the level of correlation between genre and movie rating. Since it is evident that is correlated with movie rating, it is worth further investigation about how much it affects movie ratings, and the correlation discrepancy of individual movie genre and a collection of movie genres. Once obtaining the results, it could be possible to improve the current recommendation system or the systems mentioned in the past literature on stream platforms to users.

6 Conclusions

From the experiment results, it is apparent that movie genre is positively correlated to movie rating, the predictions are more accurate when the datasets include genre features. This result corresponds with the previous related work using genre correlation and rate score to better recommend users movies. Besides adjusting the current research methods to improve the performance of prediction, the next phase of the research can further analyse the level it is correlated with movie rating.

References

Hwang, T. G., Park, C. S., Hong, J. H., & Kim, S. K. (2016). An algorithm for movie classification and recommendation using genre correlation. Multimedia Tools and Applications, 75,

12843-12858.

Makita, E., & Lenskiy, A. (2016). A movie genre prediction based on Multivariate Bernoulli model and genre correlations. arXiv preprint arXiv:1604.08608.