**Gross Movie Prediction from IMDB 5000 Movie Dataset (SIRT)**

We have used IMDB 5000 movie dataset to predict movie gross. Both categorical and numerical features of the database have been used to predict gross.

**2. Preprocessing:**

Data has been preprocessed for machine learning training (using both numerical and textual data and treating textual features as categorical feature (e.g., director’s name)).

**2.1. Features:**

Following features have been used for gross prediction:

*Numerical Features*: 6 numerical features have been used. Director’s Facebook likes, budget, number of critics review, cast total Facebook like, IMDB score, and duration are the numerical features that has been used to predict gross.

*Text Features*: 1 textual feature have been used in gross prediction, which is the director’s name

**2.2. Features Preprocessing:**

Data Cleaning has been cleaned before data type specific preprocessing. Rows with nan and missing gross value has been removed. Records with missing major feature values were also removed from data.

*Numerical Features Preprocessing:* Numerical data like budget was split into two groups. The first group was classified as high budget movies while the second group was classified as low budget movies. This classification was achieved by finding the mean of the budget column and then grouping budget below the mean as low budget movies and amounts above the mean as high budget moves. The grouping was necessary to let us test and train the dataset for machine learning purposes

*Textual Features Preprocessing:*

*Motivation for using Textual Data as Categorical Data:*

Textual data has been labeled for each column separately. Each column was assigned label for each distinct feature. The focus for the textual data was the director’s name column. The goal is to find out, if movie directors (director name) with a certain number of movies has more chances of achieving good gross predictions

Motivation behind using the textual features like director names as categorical data is that they contain information that can be crucial for good gross prediction. Also, transforming director names and some of the Numerical data helps in maintaining the complex relation among the feature columns. For example, a set of features like budget, director combination could attain certain range of gross for a movie. So, using the features as category could help preserve their interrelations.

**4. Machine Learning Models:**

Random Forest and Neural Network has been used to predict gross. First, we prepared the input data and created a model. Train and fit training data to the model. Reason for using this model, neural network is effective at detecting complex, nonlinear relationships, greater tolerance to messy data and able to ignore noisy characteristics in data.

**5. Model Evaluations:**

Both model's predictive accuracy, and their output is very similar. Both the random forest and deep learning models were able to predict correctly whether a director’s influence can significantly predict the gross of a movie by over 85% of the time.

Though their predictive performance was comparable, their implementation and training times were not—Random Forest classifier was able to train on the large dataset and predict values faster, while the deep learning model required more time to train the data points.

Table: Random Forest and Deep Neural Network Performance Evaluation for gross prediction using both

numerical and categorical data

|  |  |  |
| --- | --- | --- |
| Evaluation Metrics | Deep Neural Network | Random Forest |
| Accuracy: | 0.8893 | 0.888 |

**6. Conclusion:**

In conclusion using directors name to predict gross displayed the importance of using an experienced director in the movie making business. However, adding textual data like actor’s name, genre for gross prediction could help further improve the result.