Movie Rating Prediction Based on Movie's Attributes

Kerem Kazandır
Department of Computer Engineering
TOBB University of Economics and Technology
Ankara,Türkiye
kkazandir@etu.edu.tr

Abstract - Accurate prediction of movie ratings is essential for understanding user preferences and improving user engagement. This paper presents a novel approach to movie rating prediction using a range of machine learning algorithms, including Support Vector Regression, Multi Layer Perceptron , Random Forest , LightGBM, and Linear Regression. By leveraging these diverse models, we aim to enhance prediction accuracy and robustness. The effectiveness of each algorithm is evaluated using standard benchmark datasets, and comparative results are provided to illustrate the strengths and limitations of each approach. Our findings demonstrate that integrating multiple machine learning techniques can lead to significant improvements in rating prediction performance.

I.INTRODUCTION

Motivation

In the realm of movie production and distribution, understanding what drives a film's success is a challenging yet crucial task. To address this challenge, we propose a machine learning model designed to predict movie ratings based on various attributes. By analyzing a rich dataset of movies from TMBD 5000 , our model aims to uncover the key factors that influence movie ratings. This can provide valuable insights for filmmakers and industry professionals, enabling them to make better decisions that enhance their chances of success.

II. RELATED WORKS

The paper[1] proposes a regression model using generative convolutional neural networks (GCNNs) to predict movie ratings before release. Unlike traditional methods that depend on after production factors, this model uses intrinsic attributes available before production, such as genre, budget, cast, director, and plot. Experiments demonstrate the model's superior performance over baseline and state of the art methods.

The study[2] focuses on predicting IMDB movie ratings by enriching the original dataset with features such as genre, actor, writer, and director ratings. Ratings are categorized into low, medium, and high, and predicted using machine learning models with both soft and hard voting ensembles. The training set is balanced using SMOTE to address data imbalances. The study evaluates model performance based on classification accuracy, precision, recall, and F1 score, highlighting its unique use of ensemble algorithms and SMOTE in movie rating prediction.

The paper[3] evaluates movie recommendation models using the MovieLens 100k and 1M datasets. It proposes a system leveraging machine learning techniques Knearest Neighbors, Support Vector Machine, and Random Forest and compares their performance based on accuracy, precision, recall, and F1 score. The study also reviews past machine learning approaches for item recommendations.

III. PROPOSED SYSTEM

TMDB 5000 Movie Dataset Overview

https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata

The TMDB 5000 Movie Dataset is a comprehensive collection of movie data provided by The Movie Database (TMDB). This dataset contains information on over 5,000 movies, making it a valuable resource for various film related analyses and applications.

Key Features

1. **Genres and Keywords**: Each entry in the dataset includes essential details about the movies, such as keywords and genres. This allows for depth exploration of film characteristics and trends.

2. **Ratings**: It includes data in vote_count and vote_average which is our target.

3. **Cast and Crew**: The dataset provides information on the cast and crew involved in each movie, including actors, directors, and producers. This is useful for exploring relationships and collaborations within the film industry.
4. **Budget and Revenue**: Financial data such as the budget and revenue of movies is also included. This allows for the analysis of economic aspects of the film industry.

Data Preprocessing

At the outset of our data preprocessing phase, we concentrated on enhancing the quality and relevance of our dataset by eliminating features that were unlikely to contribute meaningful insights to our analysis. The initial version of the dataset included several complex attributes, such as genres, keywords, cast, and crew, which were stored in a dictionary format.

To address this, our first step was to simplify these dictionary based attributes by converting them into strings. This conversion process involved extracting only the names associated with the genres, keywords, cast, and crew, thereby distilling the information into a more concise and usable form. This transformation allowed us to maintain the essential elements of each attribute while reducing the complexity of the dataset.

Subsequently, we applied the one hot encoding technique to these stringified attributes. One hot encoding enabled us to convert categorical variables into a binary matrix, where each column represented the presence or absence of a specific genre, keyword, cast member, or crew member. This method not only facilitated the handling of categorical data but also enhanced the interpretability and accessibility of the dataset for further analysis.

Through these systematic preprocessing steps, we were able to create four distinct datasets, each varying in size and tailored to meet different analytical needs. These datasets offer flexibility in analysis, allowing for targeted exploration depending on the scope and depth of the inquiry.

IV. METHOD

Models

1. MLP: A multilayer perceptron is a name for a modern feedforward artificial neural network, consisting of fully connected neurons with a nonlinear activation function, organized in at least three layers, notable for being able to distinguish data that is not linearly separable

2. Linear Regression: Linear Regression is a statistical method that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal is to find the best fitting line that minimizes the difference between the predicted values and the actual values.

3. LightGBM: LightGBM is a gradient boosting framework that uses tree based learning algorithms. It is designed for efficiency and scalability, handling large datasets and high dimensional features well. LightGBM builds trees in a leaf wise manner and supports parallel and distributed learning, making it faster and more accurate than some other boosting methods.

4. SVR: Support Vector Regression is a regression technique that extends Support Vector Machines for predicting continuous values. It aims to find a function that deviates from the actual observed values by a margin less than a specified threshold, while minimizing the complexity of the function and maximizing the margin between the predicted values and the actual values.

5. Random Forest: Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and control overfitting. It builds a large number of decision trees during training, each using a random subset of features, and makes predictions by averaging the results from all the trees (for regression).

Performance Metrics

1. MSE: Mean Squared Error (MSE) measures the average squared difference between predicted and actual values, quantifying how well a regression model fits the data. Lower MSE indicates a better fit, but it can be sensitive to outliers due to the squaring of errors.

2. MAE: Mean Absolute Error calculates the average absolute difference between predicted and actual values, reflecting the average magnitude of prediction errors. It is less sensitive to outliers than MSE, as it does not square the errors.

3. R^2 Score: R squared represents the proportion of the variance in the dependent variable that is predictable from the independent variables, indicating how well the regression model explains the variability of the data. A higher R^2 value means a better fit, but it can be misleading if used alone, especially with nonlinear relationships.

4. Accuracy Within Threshold: Accuracy within Threshold measures the proportion of predictions where the absolute error between predicted and actual values is within a specified range. It assesses how often the model's predictions fall within an acceptable margin of error.

V. RESULT

Test and training performance rates of the models for four different data sets created

1. **Only_genres_keywords dataset**:

This dataset is comprised of 55 attributes, making it relatively comprehensive in terms of the variety of information it includes. This dataset, in comparison to others, contains the least amount of information and notably lacks detailed cast and crew data. Its limited scope may pose challenges for depth analysis, particularly in areas where cast and crew information could provide valuable insights.

Test result:

	Lineer reg	MLP	LightGBM	SVR	RF
MSE	0.398686	0.391482	0.329099	0.401383	0.356216
R^2	0.425448	0.435830	0.525731	0.421562	0.486653
MAE	0.482756	0.479308	0.439664	0.483969	0.454082
AWT 0.1	0.143533	0.140379	0.149842	0.149842	0.171924
AWT 0.3	0.414826	0.425868	0.424290	0.405363	0.413249
AWT 0.5	0.613565	0.613565	0.675079	0.613565	0.645110
AWT 0.7	0.757098	0.753943	0.794953	0.750789	0.782334
AWT 0.9	0.865931	0.865931	0.892744	0.859621	0.875394

Training result:

	Lineer reg	MLP	LightGBM	SVR	RF
MSE	0.380183	0.369819	0.212013	0.383918	0.146403
R^2	0.453945	0.468831	0.695487	0.448581	0.789722
MAE	0.477323	0.469541	0.355896	0.474587	0.290159
AWT 0.1	0.135071	0.139021	0.192338	0.137046	0.227883
AWT 0.3	0.396130	0.406398	0.513823	0.411137	0.615324
AWT 0.5	0.610585	0.614929	0.750395	0.620458	0.838863
AWT 0.7	0.773697	0.778831	0.884676	0.770142	0.933649
AWT 0.9	0.869273	0.877172	0.949842	0.869668	0.971169

2. **Small_sized dataset**:

This dataset comprises 123 attributes, offering a broad yet somewhat selective range of information. This dataset does not include cast and crew information, but it compensates with a broader range of genres and keywords.

Test result:

	Lineer reg	MLP	LightGBM	SVR	RF
MSE	0.389229	0.378588	0.319402	0.385906	0.324148
R^2	0.409265	0.425415	0.515242	0.414308	0.508039
MAE	0.485535	0.482719	0.437705	0.482188	0.446440
AWT 0.1	0.148265	0.140379	0.164038	0.154574	0.145110
AWT 0.3	0.400631	0.392744	0.429022	0.395899	0.419558
AWT 0.5	0.591483	0.593060	0.652997	0.597792	0.637224
AWT 0.7	0.741325	0.755521	0.804416	0.755521	0.804416
AWT 0.9	0.861199	0.870662	0.891167	0.864353	0.892744

Training result:

	Lineer reg	MLP	LightGBM	SVR	RF
MSE	0.368103	0.351133	0.213206	0.376403	0.142613
R^2	0.477514	0.501601	0.697375	0.465733	0.797575
MAE	0.468266	0.455710	0.355035	0.463890	0.284664
AWT 0.1	0.142575	0.146919	0.192733	0.136256	0.238942
AWT 0.3	0.401264	0.421011	0.528041	0.437994	0.627172
AWT 0.5	0.631122	0.644550	0.750395	0.647314	0.840837
AWT 0.7	0.781201	0.794629	0.877172	0.780806	0.939573
AWT 0.9	0.866904	0.877962	0.948262	0.864929	0.973539

3. **Medium_sized dataset**:

This dataset is composed of 209 attributes, offering a comprehensive array of information that covers various aspects of the films or shows being analyzed. While the dataset does not incorporate cast and crew information, it compensates by offering an extensive array of genres and keywords, ensuring a comprehensive representation of various thematic elements and categories.

Test result:

	Lineer reg	MLP	LightGBM	SVR	RF
MSE	0.385677	0.385716	0.307723	0.381288	0.340984
R^2	0.433579	0.433520	0.548065	0.440023	0.499215
MAE	0.484848	0.486147	0.433142	0.478690	0.463900
AWT 0.1	0.121451	0.118297	0.149842	0.134069	0.135647
AWT 0.3	0.411672	0.388013	0.424290	0.406940	0.386435
AWT 0.5	0.605678	0.596215	0.662461	0.605678	0.621451
AWT 0.7	0.771293	0.771293	0.802839	0.779180	0.780757
AWT 0.9	0.854890	0.856467	0.905363	0.864353	0.884858

Training result:

	Lineer reg	MLP	LightGBM	SVR	RF
MSE	0.345295	0.336660	0.212397	0.356249	0.139470
R^2	0.506422	0.518765	0.696391	0.490764	0.800636
MAE	0.453054	0.447678	0.354941	0.450024	0.280108
AWT 0.1	0.157188	0.154818	0.193523	0.135861	0.248025
AWT 0.3	0.421801	0.432859	0.513823	0.464060	0.639021
AWT 0.5	0.633096	0.641390	0.744471	0.653633	0.845577
AWT 0.7	0.788705	0.786730	0.884281	0.791864	0.940363
AWT 0.9	0.880727	0.882701	0.948262	0.868878	0.974724

4. **Big_sized dataset**:

This dataset is composed of 401 attributes, making it one of the most extensive and detailed datasets available for analysis. Among these attributes, a significant portion is dedicated to providing depth information about the cast and crew.

Test result:

	Lineer reg	MLP	LightGBM	SVR	RF
MSE	0.419966	0.405722	0.322422	0.420897	0.345458
R^2	0.405844	0.425996	0.543846	0.404527	0.511255
MAE	0.507544	0.498239	0.442757	0.506512	0.463061
AWT 0.1	0.134069	0.148265	0.143533	0.130915	0.118297
AWT 0.3	0.367508	0.359621	0.413249	0.381703	0.391167
AWT 0.5	0.578864	0.589905	0.630915	0.580442	0.621451
AWT 0.7	0.730284	0.744479	0.821767	0.742902	0.787066
AWT 0.9	0.845426	0.853312	0.902208	0.848580	0.889590

Training result:

	Lineer reg	MLP	LightGBM	SVR	RF
MSE	0.300312	0.292591	0.204576	0.315409	0.128227
R^2	0.566683	0.577824	0.704820	0.544900	0.814982
MAE	0.424073	0.418453	0.348208	0.418678	0.268329
AWT 0.1	0.150869	0.161137	0.191548	0.127962	0.258294
AWT 0.3	0.440758	0.441548	0.528436	0.513823	0.655608
AWT 0.5	0.670221	0.676145	0.759084	0.690758	0.862559
AWT 0.7	0.817141	0.823460	0.892970	0.817536	0.948657
AWT 0.9	0.902844	0.905213	0.952607	0.889021	0.980253

Summary

Upon reviewing the performance metrics across the four different datasets, it is clear that the models consistently yield better results when applied to the medium_sized dataset. This trend indicates that the medium_sized dataset offers a more optimal balance of data complexity and sample size for rating prediction tasks.

In terms of model performance, both the LightGBM and Random Forest models have consistently outperformed other models across these datasets. LightGBM, in particular, has demonstrated superior performance relative to Random Forest. LightGBM is proving to be the most effective model for rating prediction in the context of these datasets, outperforming not only Random Forest but also other models

.

VI. CONCLUSION AND FUTURE SCOPE

In this study, we have conducted a comparison of LightGBM with other machine learning models across various datasets and evaluation metrics. The experimental results demonstrate that LightGBM consistently outperforms the other models in terms of predictive accuracy.

.

The findings confirm that LightGBM is a robust and scalable solution for a wide range of machine learning tasks, particularly when dealing with large datasets. It not only delivers superior predictive performance but also significantly reduces training time, making it a preferred choice for real world applications where accuracy.

Future work would try to more complex models.

REFERENCES

[1] Ning, X., Yac, L., Wang, X., Benatallah, B., Dong, M., & Zhang, S. (2020). Rating prediction via generative convolutional neural networks based regression. Pattern Recognition Letters, 132, 12–20.

[2] A. F. Dereli, “IMDB Movie Rating Prediction with Feature Extraction and Machine Learning Methods,” 2022. Marmara Universitesi (Turkey).

[3] Siddique, N. A., Abid, N. M. K., Fuzail, N. M., & Aslam, N. N. (2024). Movies Rating Prediction using Supervised Machine Learning Techniques. International Journal of Information Systems and Computer Technologies, 3(1), 40–56.s