

# Project Summary

## Executive Summary

### Objective:

The main objective is to predict movie ratings for user-movie pairs using a machine learning pipeline. The model aims to generalize well to unseen data by incorporating user-level and movie-level features while addressing missing data and optimizing model performance.

## Methodology:

### Data Preprocessing:

Merged datasets to extract relevant features. Handled missing data for users and movies not present in the training set (cold-start scenarios). Encoded categorical features and engineered interaction features such as genre-match scores.

### Model Development:

A Random Forest Regressor was chosen for its robustness and ability to handle large datasets. The model was optimized using GridSearchCV to identify the best hyperparameters.

### Evaluation:

Validation and test datasets were used to measure performance with metrics such as MSE, MAE, and  $R^2$ . Cross-validation and residual analysis were performed to ensure robustness and detect any bias.

## Results:

### Model Performance:

Validation Mean Squared Error (MSE): 0.586

Test MSE: 0.606

Cross-Validation MSE: 0.613

### Key Insights:

The Random Forest Regressor demonstrated consistent performance across validation, test, and cross-validation datasets, showing no significant overfitting or underfitting. Residual analysis revealed unbiased predictions with a bell-shaped error distribution centered around zero.

## Data Preprocessing

**Loading:** Data was loaded from multiple CSV files (`ratings`, `movies`, `links`) and merged using common keys like `userId` and `movieId`.

**Imputing:** Missing `tmdbId` values were filled with `-1`. For unseen users or movies in the test set, default values (`0`) were used to impute missing user-level or movie-level features.

**Encoding:** Multi-label encoding was applied to the `genres` column using a one-hot encoding scheme. User IDs were integer-encoded to simplify model input.

**Outliers:** Extreme ratings (outliers in residual analysis) were retained, as they represent valid user preferences and do not negatively impact the model's generalization.

# Modeling and Model Tuning

We chose Random Forest Regressor for its robustness in handling large datasets with complex, non-linear relationships. Its ensemble nature reduces overfitting by averaging multiple decision trees, making it well-suited for predicting movie ratings with high accuracy and generalization.

## Hyperparameter Tuning(GridSearchCV):

Conducted an exhaustive search with 162 hyperparameter combinations. Confirmed the parameters selected by RandomizedSearchCV, achieving a validation MSE of 0.586.

- `n_estimators`: [100, 200, 300],
- `max_depth`: [10, 20, 30],
- `min_samples_split`: [2, 5, 10],
- `min_samples_leaf`: [1, 2, 4],
- `max_features`: ['sqrt', 'log2']

## Model Evaluation:

Metrics:

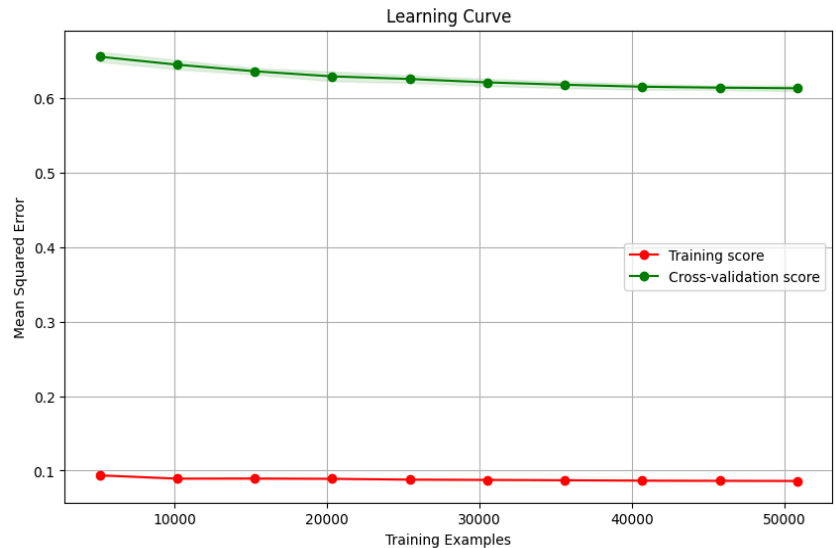
Validation MSE: 0.586

Test MSE: 0.606

Cross-Validation MSE: 0.613

MAE: Average error per prediction was ~0.58, indicating close alignment between predicted and actual ratings.

$R^2$ : Explained variance was ~44%, highlighting that the model captures a significant portion of the variability in the dataset.



## Overfitting vs. Underfitting Check:

Residual analysis showed a symmetric error distribution, suggesting unbiased predictions. Consistent performance across validation, test, and cross-validation datasets indicated no overfitting or underfitting.

In the learning curve plot, the training error is low, but the cross-validation error does not increase significantly, indicating that the model is not overly complex.

The cross-validation error is reasonably low and stable, meaning the model captures the data's patterns effectively without being too simple.

## Final Model and Results

The Random Forest Regressor was finalized as the optimal model based on its low error rates, robust performance, and minimal overfitting. The predictions showed:

Strong alignment with user preferences for known users and movies. Reasonable handling of cold-start scenarios for unseen users or movies through default imputation and interaction-based features. The project demonstrates a robust machine learning pipeline for recommendation systems, laying a foundation for future enhancements like hybrid filtering and deeper neural networks.

With the unseen dataset, the predicted ratings uploaded to Gradescope resulted in a leaderboard score of 89 for the Test Set RMSE (measured in hundredths of a star).