Film Forecast: Predicting IMDb Ratings with Machine Learning

Jiyeon (Jenna) Woo, Mahnoor Shahid, and Kaylyn Nguyen

01

EXECUTIVE SUMMARY

OVERVIEW

Project Overview

- Developed a predictive model for IMDb movie ratings using machine learning.
- Compared three advanced models: Random Forest, CatBoost, and LightGBM.
- Identified key factors influencing movie ratings.

Key Findings

- Best-performing model: CatBoost
- Top influential features: Number of voted users, Genres, Movie duration

Potential Applications

- Box office prediction
- Marketing strategy optimization
- Supporting production decisions

02

BACKGROUND

MOTIVATION

Motivation

- Data-driven decision-making is shaping the entertainment industry.
- IMDb ratings are key to measuring audience reception and movie success.
- With 500+ movies released yearly in the U.S., accurate predictions offer valuable insights.

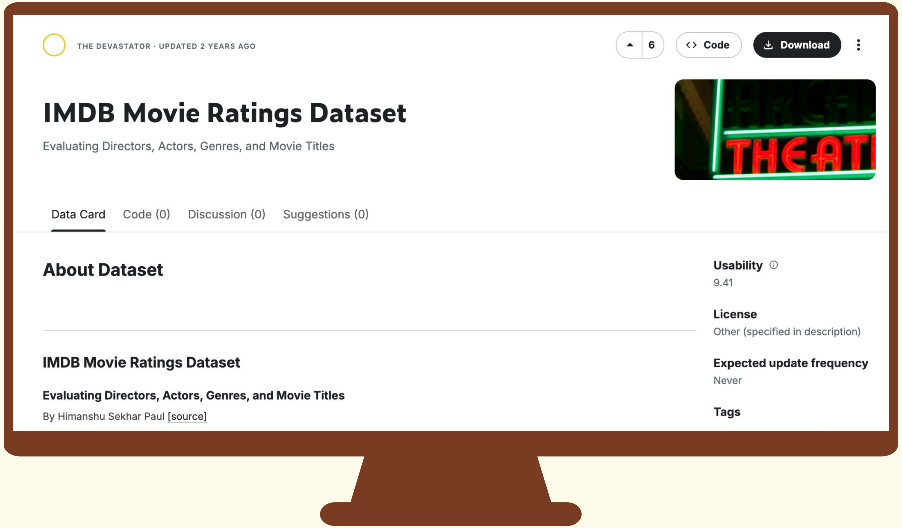
Limitations of Traditional Methods

- Subjective evaluations and historical comparisons lack precision.
- Cannot fully capture complex interactions affecting ratings.

Project Goals

- Use machine learning to predict IMDb ratings with greater accuracy.
- Analyze key factors: genre, duration, audience engagement.
- Support better decision-making in casting, production, marketing, and distribution.

DATASET



Kaggle IMDb Movie Ratings Dataset

Key features: Director name, Movie duration, Lead actors, Genres, Audience engagement (votes, reviews) Release year, language, country

03

ANALYSES

METHODOLOGY

STEP 1

Import libraries: Pandas, NumPy, Matplotlib, Seaborn for analysis; LightGBM, CatBoost, sklearn for modeling.

STEP 2

Clean the Kaggle dataset by removing unnecessary columns, handling missing values, encoding categories, and splitting into train/test sets.

STEP 3

Compare Random Forest, CatBoost, and LightGBM to select the best model.

STEP 4

Perform exploratory data analysis on the impact of certain features on the IMDb score.

STEP 5

Address cold-start issues with pre-release (cast, director, genre) and post-release (votes, reviews) factors.

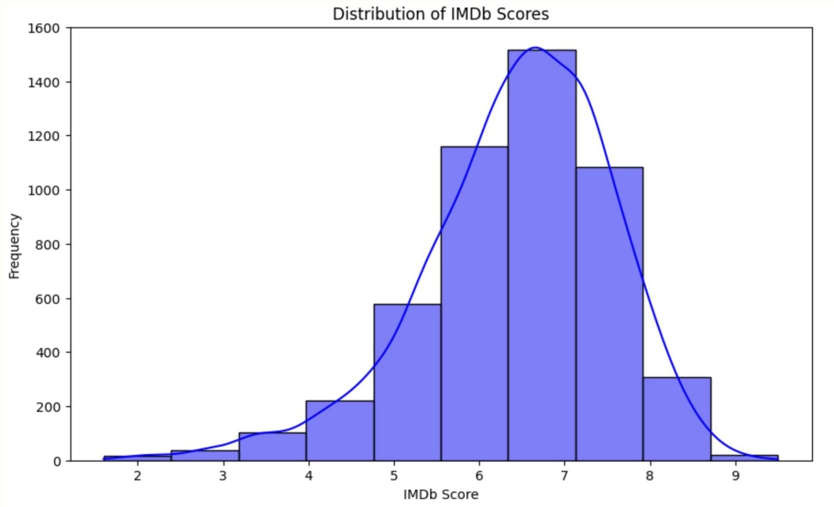
STEP 6

Build IMDb Score Prediction model based on the most optimal model.

DISTRIBUTION OF IMDb SCORES

The distribution of IMDb scores reveals critical insights into movie ratings:

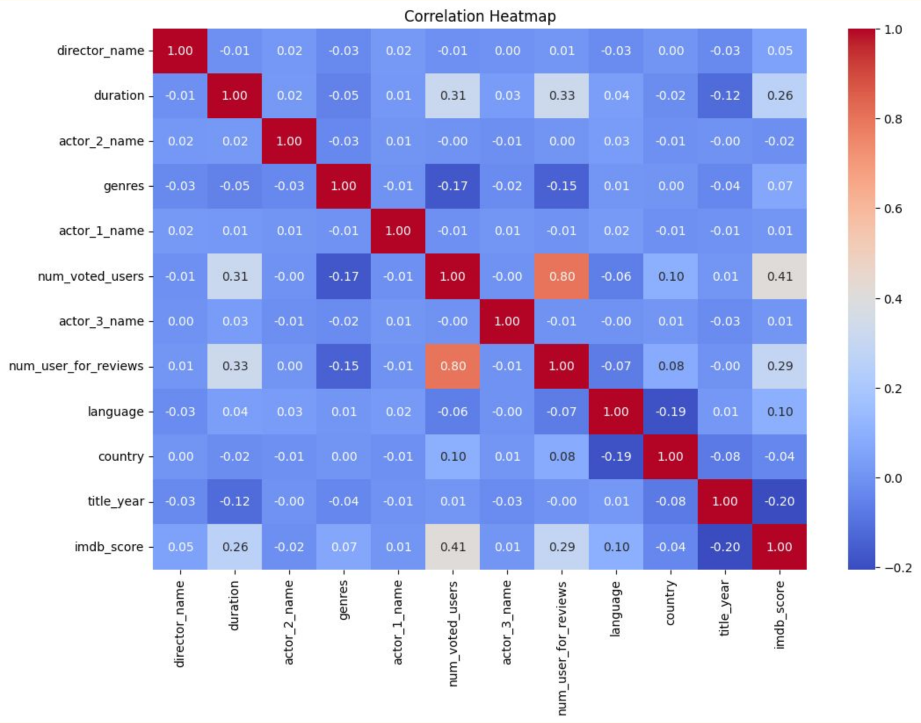
- The histogram shows a bell-shaped curve centered around 6-7, indicating most movies receive moderate ratings
- Few movies receive extremely low (<4) or extremely high (>8.5) ratings
- Most films cluster around average ratings



CORRELATION HEATMAP INSIGHTS

The correlation matrix highlights key relationships:

- Highest correlation between number of user votes and number of reviews
- Weak correlations between most features and IMDb score
- Suggests that predicting movie ratings is complex and requires sophisticated modeling techniques



MODEL PERFORMANCE COMPARISON

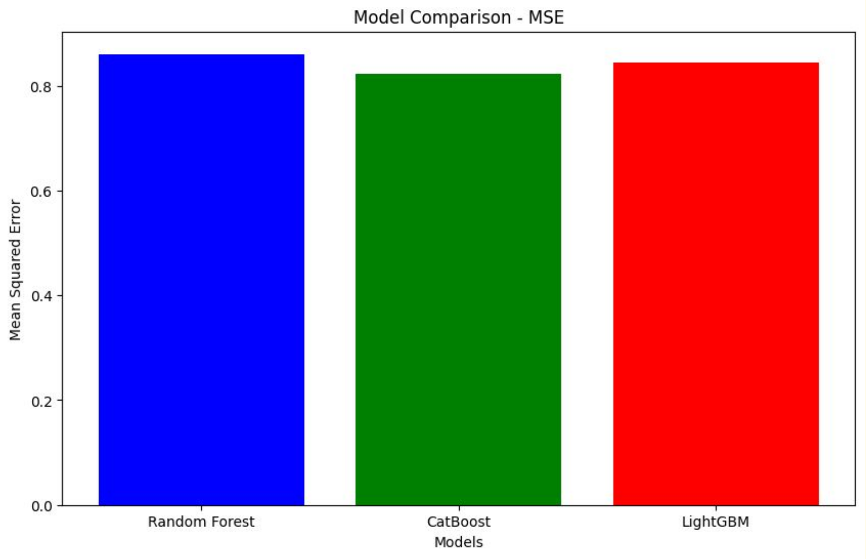
Best Model: CatBoost (lowest MSE, highest R^2)

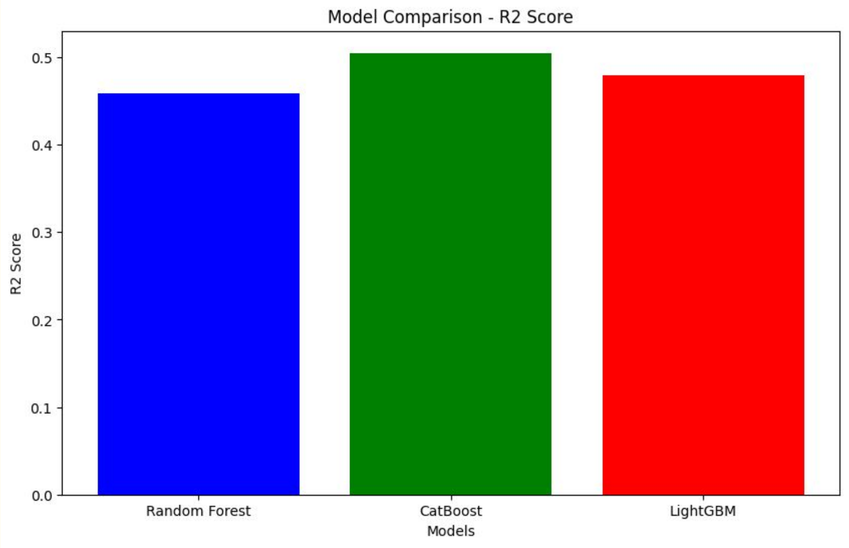
MSE:

- RF: 0.8607
- **CatBoost: 0.8236**
- LightGBM: 0.8440

R^2 :

- RF: 0.4592
- **CatBoost: 0.5049**
- LightGBM: 0.4799





FEATURE IMPORTANCE ANALYSIS

Consistent Top Features Across All Models:

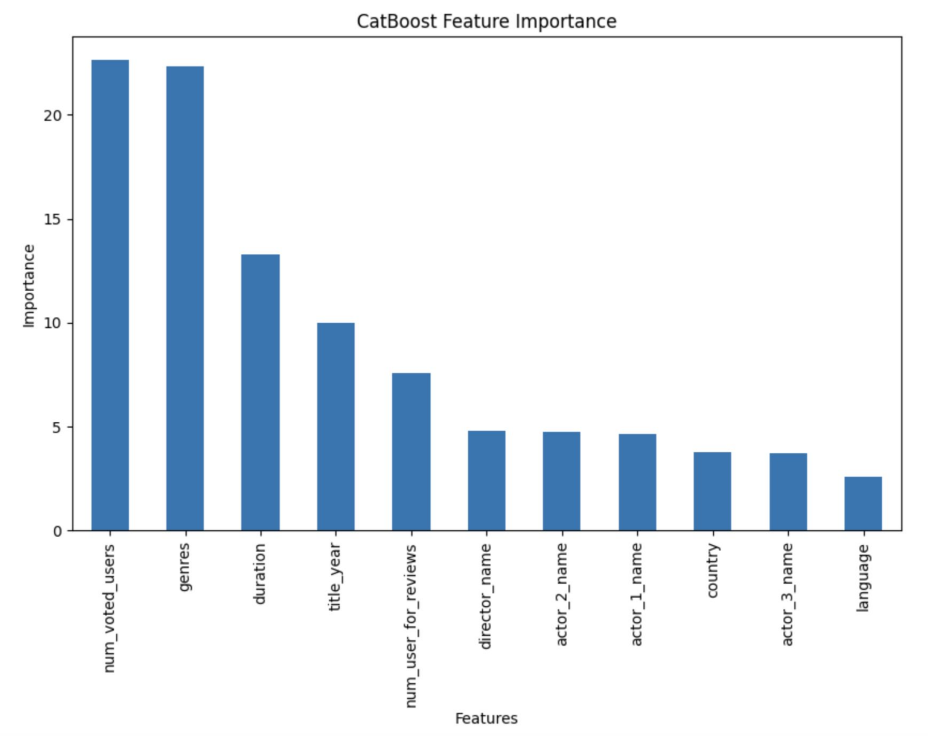
- Number of users voted
- Genres

Random Forest and CatBoost:

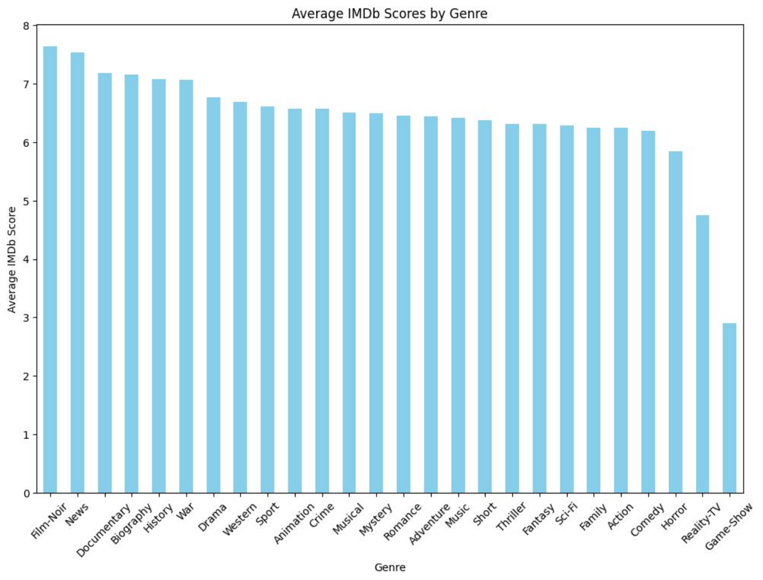
- Top features are typically numerical
- Duration
- Title year
- Number of reviews

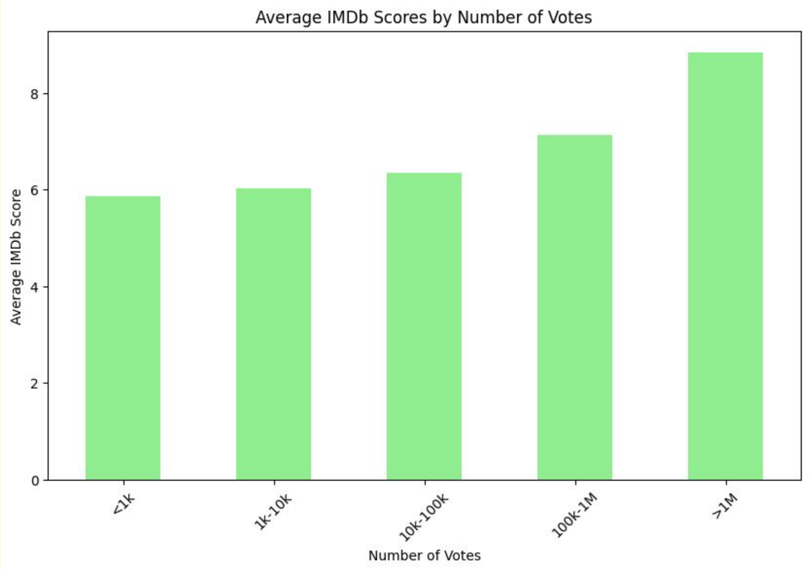
LightGBM:

- Top features a mix of numerical and categorical
- Actor names, director names



GENRE AND RATING ANALYSIS

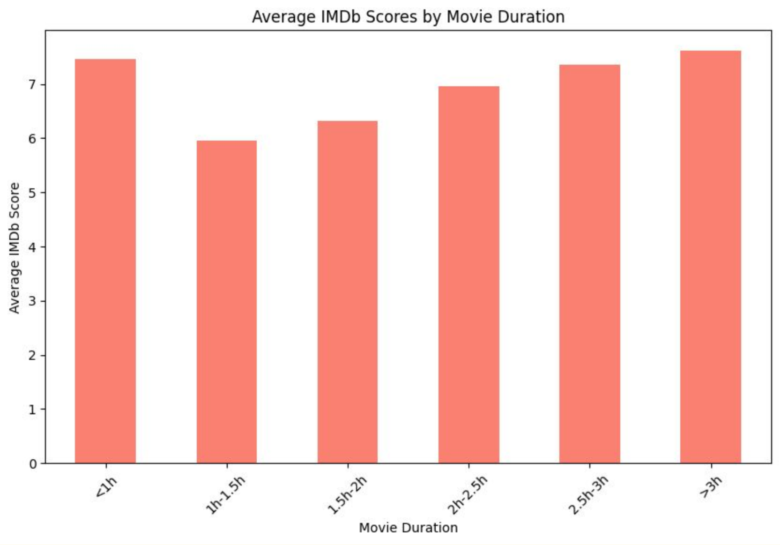




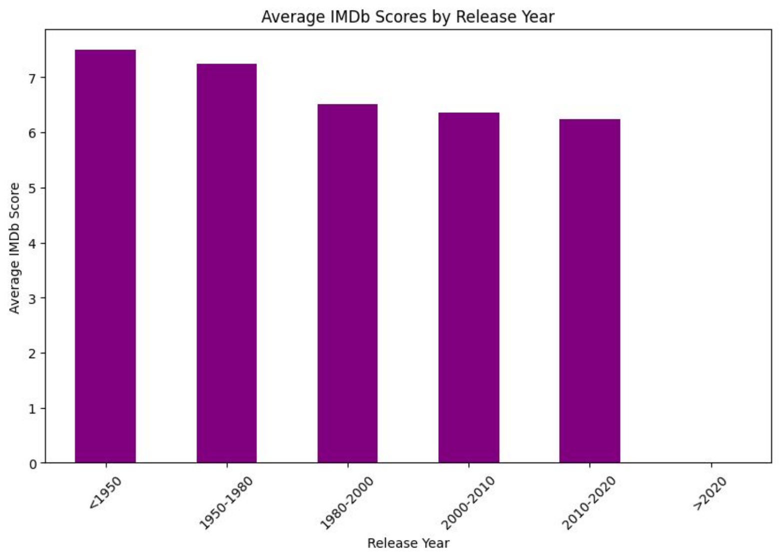
This bar graph shows significant rating variation, with top genres like Film-Noir and News, and lowest-rated genres like Reality-TV and Game Show, highlighting differing preferences.

This bar graph shows that movies with over 1M votes have the highest and most stable ratings, indicating that popular films tend to have more reliable ratings.

MOVIE DURATION AND RELEASE YEAR IMPACT



This bar graph suggests that films between 1-3 hours have lower ratings, while very short and very long films score higher, indicating an optimal length for audience satisfaction.



This bar graph shows that movies released before 1980 have the highest average ratings, suggesting a nostalgia bias or shifts in storytelling quality and audience expectations over time.

IMDb SCORE PREDICTION MODEL

Two-Model Approach

- Pre-release model: Predicts ratings using director, cast, genre, and duration.
- Post-release model: Refines predictions with votes and reviews.

Machine Learning Pipeline

- CatBoost regressor trained on historical IMDb data.
- Cold-start model uses historical averages of directors, actors, and genres.

Interactive User Interface

- Users input movie details to receive a predicted IMDb score.
- Label encoding for categorical features.

IMDb SCORE PREDICTION MODEL

Prediction Flow:

- If audience data available → Post-release model refines rating.
- If not → Pre-release model estimates rating.

Post-release model:
Minions (2015)

```
Predict the IMDb rating for a movie.

Enter the director's name: Kyle Balda
Enter the first lead actor's name: Sandra Bullock
Enter the second lead actor's name: Steve Carell
Enter the third lead actor's name: Michael Keaton
Enter the genre (e.g., Action, Comedy, Drama): Comedy
Enter the movie duration (in minutes): 90
Enter the release year: 2015
Has the movie been released? (yes/no): yes
Enter the number of votes received: 271000
Enter the number of user reviews received: 423

Predicted IMDb Rating: 7.04
```

Pre-release model:
The Minecraft Movie (2025)

```
Predict the IMDb rating for a movie.

Enter the director's name: Jared Hess
Enter the first lead actor's name: Jennifer Coolidge
Enter the second lead actor's name: Jason Momoa
Enter the third lead actor's name: Emma Myers
Enter the genre (e.g., Action, Comedy, Drama): Comedy
Enter the movie duration (in minutes): 100
Enter the release year: 2025
Has the movie been released? (yes/no): no

Predicted IMDb Rating: 6.72
```

04

RECOMMENDATIONS

RECOMMENDATIONS

Model Selection

Choose CatBoost for IMDb score predictions due to its:

- Lowest RMSE, Highest R^2 score
- Superior handling of categorical features
- Better performance on mixed data types typical in movie datasets

Two-Model Approach

Develop:

- Pre-release predictions (using movie characteristics)
- Post-release refinement (incorporating audience metrics)

Feature Engineering

Focus on:

- Audience engagement metrics
- Genre classification
- Historical performance of directors and actors
- Temporal trends in movie ratings

05

CONCLUSION

KEY TAKEAWAYS

Complex Influences:

- Ratings depend on a mix of audience engagement, genre, duration, and trends.

Best Model:

- CatBoost

Cold-Start Solution:

- Pre-release model and Post-release model

Industry Applications:

- Box office forecasting, marketing, and production decisions.

Future Directions:

- Advanced features & ensemble techniques
- Genre-specific models for deeper insights.

THANK YOU