DATA 602 Final Project

Title: Predicting Movie Success: Analysis and Modeling
Subtitle: Analyzing Gross Revenue, Popularity, and Award Wins
Team Members: Fomba Kassoh and Souleymane Doumbia

Abstract

This project focuses on analyzing and predicting various aspects of movies, including gross revenue, popularity, and award wins, using machine learning models. The primary goal is to derive insights into the factors contributing to a movie's success in these areas and to predict outcomes based on specific features.

For gross revenue prediction, Linear Regression, Ridge Regression, and Lasso Regression models were evaluated. The Ridge Regression model performed the best with a Mean Squared Error (MSE) of 0.7024 and an R^2 score of 0.6392.

In predicting movie popularity, Logistic Regression, Support Vector Machine (SVM), and Random Forest models were used. The Random Forest model achieved the highest accuracy at 0.79, with precision, recall, and F1-Score all at 0.79.

For award prediction, the same three models were applied. The Logistic Regression model had the highest recall at 0.55, while the Random Forest model had the highest precision at 0.43 and F1-Score at 0.44.

These results highlight the effectiveness of different machine learning models in predicting various success metrics for movies, providing valuable insights for the film industry in production, marketing, and distribution decisions.

Introduction

Research Question: How do factors such as popularity, and specific features influence the likelihood of a movie's success in terms of gross revenue, popularity, and award wins?

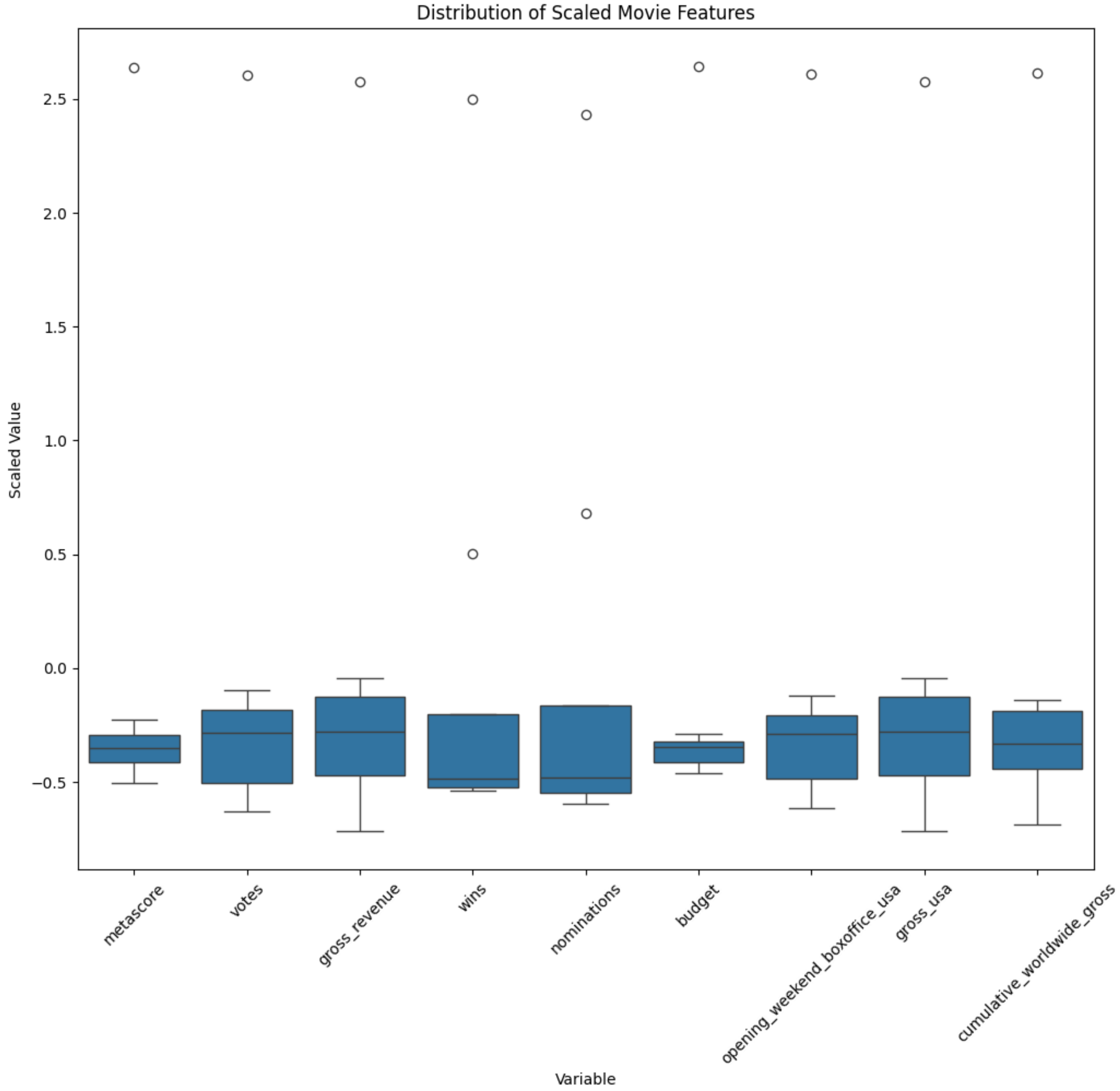
Justification: Predicting movie success is crucial for the film industry to make informed decisions about production, marketing, and distribution. This study aims to identify key factors contributing to a movie's financial and critical success. The analysis is relevant to both the film industry and marketing agencies, offering insights into what influences movie success. These insights can guide decisions on movie production, marketing strategies, and consumer preferences, ultimately enhancing the industry's ability to create and promote successful films.

Data Sources: The dataset was obtained from IMDb (Internet Movie Database), a comprehensive source of movie data. We collected this data by web scraping IMDb's website, considering the nature of the fields present. The specific URL is provided below.

IMDb Top 1000 (Sorted by User rating Ascending)

Exploratory Data Analysis (EDA)

A. Distribution of numerical features



Below is an analysis of each feature's distribution based on the boxplot above:

1. **Metascore:**
   - **Distribution:** The metascores are fairly symmetrically distributed.
   - **Variability:** Moderate variability, with the interquartile range (IQR) covering scores from around 48 to 71.
   - **Outliers:** There are a few outliers on both ends, indicating some movies have exceptionally low or high metascores compared to the rest.

2. **Votes:**
    - **Distribution:** The distribution is right-skewed, indicating a larger number of movies received fewer votes.
    - **Variability:** High variability, with the IQR showing a wide range.
    - **Outliers:** Several outliers, especially on the higher end, indicate some movies received an exceptionally high number of votes.
3. **Gross Revenue:**
    - **Distribution:** Right-skewed distribution with most movies having lower gross revenues.
    - **Variability:** High variability, with significant differences in gross revenues among movies.
    - **Outliers:** Numerous outliers, particularly on the high end, reflecting movies that earned significantly more than the majority.
4. **Wins:**
    - **Distribution:** Right-skewed, indicating most movies have fewer wins.
    - **Variability:** High variability, with a wide range of wins among movies.
    - **Outliers:** A few extreme outliers on the high end, representing movies with a significantly high number of wins.
5. **Nominations:**
    - **Distribution:** Right-skewed, with most movies receiving fewer nominations.
    - **Variability:** High variability, with the IQR covering a wide range of nominations.
    - **Outliers:** Several high-end outliers, indicating some movies have an exceptionally high number of nominations.
6. **Budget:**
    - **Distribution:** Right-skewed, showing that most movies have lower budgets.
    - **Variability:** High variability, with budgets varying significantly among movies.
    - **Outliers:** Several high-end outliers, reflecting movies with exceptionally large budgets.
7. **Opening Weekend Box Office (USA):**
    - **Distribution:** Right-skewed, indicating most movies have lower opening weekend earnings.
    - **Variability:** High variability, with a wide range of opening weekend earnings.
    - **Outliers:** A few high-end outliers, indicating some movies earned significantly more than others on their opening weekend.
8. **Gross USA:**
    - **Distribution:** Right-skewed, with most movies having lower domestic gross earnings.

- **Variability:** High variability, with the IQR covering a wide range of earnings.
- **Outliers:** Numerous high-end outliers, reflecting movies with exceptionally high domestic gross earnings.

9. **Cumulative Worldwide Gross:**

    - **Distribution:** Right-skewed, indicating most movies have lower worldwide gross earnings.
    - **Variability:** High variability, with a significant range in earnings among movies.
    - **Outliers:** Several high-end outliers, representing movies with exceptionally high worldwide gross earnings.

Overall, the distributions indicate that most movie features, such as votes, gross revenues, wins, nominations, budget, and earnings, are right-skewed with high variability. There are several outliers in each feature, mainly on the higher end, indicating some movies perform exceptionally well compared to the majority.

B. Distribution of Categorical Variables

The categorical data shows a wide variety of unique values in most categories, suggesting a diverse dataset in terms of movies, certificates, runtimes, genres, directors, stars, countries of origin, languages, and production companies. The presence of frequent categories in some fields like certificates and countries of origin highlights certain common trends within the dataset.

- **Certificate:** There are 11 unique certificates, with "PG-13" being the most common, accounting for nearly half of the movies.

- **Runtime:** The dataset includes 107 unique runtimes, with "115 min" being the most common, but only appearing 26 times, suggesting diverse movie lengths.

- **Genre:** With 159 unique genres, "Animation, Adventure, Comedy" is the most common, appearing 106 times, indicating a variety of genre combinations.

- **Directors and Stars:** There are 988 unique directors and stars, with no single director being significantly more common than others.

- **Countries of Origin:** The dataset includes movies from 206 different countries, with the United States being the most common country of origin, accounting for 422 movies.

- **Languages:** There are 326 unique languages, with English being the most common, spoken in 410 movies.

- **Production Companies:** There are 821 unique production companies, with "Walt Disney Pictures, Pixar Animation Studios" being the top, but only producing 10 movies, indicating a wide range of production companies involved.

The categorical data shows a wide variety of unique values in most categories, suggesting a diverse dataset in terms of movies, certificates, runtimes, genres, directors, stars, countries of origin, languages, and production companies. The presence of frequent categories in some fields like certificates and countries of origin highlights certain common trends within the dataset.

Data Wrangling

Below is a summary of the data wrangling steps performed. The data wrangling process involves comprehensive steps to ensure the data is clean, consistent, and ready for analysis and modeling. Key steps include converting data types, handling missing values through imputation, transforming skewed data, extracting and processing date and runtime information, and preparing the data for machine learning models using appropriate preprocessing techniques.

1. Initial Data Inspection:
    - The dataset initially contains 22 columns with a mix of object, float64, and int64 data types.
    - Key columns like certificate, runtime, directors, stars, countries_of_origin, languages, and production_companies are stored as objects.

2. Apply Conversions:
    - Columns like votes, gross_revenue, opening_weekend_boxoffice_usa, gross_usa, cumulative_worldwide_gross, and ranking are converted to appropriate numerical types using the helper functions that converts strings with commas and decimals to integers and floats.

3. Date and Other String Conversions:
    - Extract and convert integer parts from columns like release_year and release_date.
    - Ensuring that the release_date column is in the correct datetime format.
    - Ensure that categorical columns like title, certificate, genre, etc., are set correctly using the .astype method.

4. Missing Values Handling
    - Calculate Missing Values:
        - Identifies columns with missing values (metascore, gross_revenue, wins, nominations, budget, opening_weekend_boxoffice_usa, gross_usa, and production_companies) and calculates the percentage of missing values.

    - Imputation:
        - Uses IterativeImputer with RandomForestRegressor to impute missing values for numerical columns.

5. Preprocessing for Modeling
    - Log Transformation: Applies a log transformation to the gross_revenue column to handle skewness.
    - Extract Integer from Runtime: Extracts numeric values from the runtime column and converts them to float.

- Date Transformation: Transforms the release_date column to extract year, month, and day of the week.
- Drop Unnecessary Columns: Drops columns such as title, web_link, and stars that are not needed for further analysis.

- Feature Selection and Preprocessing:
  - Defines the target variable (gross_revenue).
  - Separates categorical and numeric features.
  - Sets up preprocessing steps using ColumnTransformer to apply StandardScaler to numerical features and OneHotEncoder to categorical features.

Data Analysis and Machine Learning Modeling

A. Gross Revenue Prediction Comparison

Three models were used: Linear Regression, Ridge Regression, and Lasso Regression.

Model	Mean Squared Error (MSE)	R^2 Score	Best Parameters	Cross-validated MSE
Linear Regression with Polynomial Features	0.7057	0.6375	N/A	N/A
Lasso Regression with Cross-Validation	0.7498	0.6149	{'alpha': 0.000212}	0.4344
Ridge Regression with Cross-Validation	0.7024	0.6392	{'alpha': 0.126}	0.4332

Below, we will compare the performance of the models. We'll compare the five models:

1. Baseline Linear Model: This model serves as a baseline to compare improvements made by other models. With an MSE of 0.767 and R^2 of 0.605, it shows the basic performance without any feature engineering or cross-validation.

2. Polynomial Interaction Features Only: Adding polynomial interaction features improves the model's performance compared to the baseline. The MSE decreases to 0.706, and the R^2 increases to 0.638. This indicates that polynomial interaction features capture more complex relationships in the data, leading to better predictions.

3. Polynomial Interaction Features with Cross-Validation: The cross-validation process helps find the best model by evaluating performance across different splits of the data. The average cross-validated MSE of 0.434 shows that this model performs well on unseen data. However, the MSE on the test set remains the same as the model with polynomial interaction features only, at 0.706. This suggests that while cross-validation helps in model selection, the final test set

performance does not improve further in this case. The R^2 score also remains the same at 0.638.

4. Lasso Regression with Polynomial Interaction Features and Cross-Validation: Did not perform as well as expected, with higher MSE and lower R^2 compared to other models. Regularization did not provide the expected benefits in this specific case.
5. Ridge Regression with Polynomial Interaction Features and Cross-Validation: Performed well, with comparable MSE and R^2 scores to the polynomial interaction model without cross-validation, indicating effective regularization.

Linear Model Selection

The best model based on the test set performance is the Ridge Regression with Polynomial Interaction Features and Cross-Validation, which achieved a lower MSE and higher R^2 on the test set compared to the other models, indicating better generalization and robustness.

B. Popularity Prediction Comparison

Models used: Logistic Regression, SVM, and Random Forest.

The models classify movies as 'Hit' or 'Flop' based on metrics like votes, metascore, and revenue.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.75	0.75	0.75	0.75
SVM	0.78	0.78	0.78	0.78
Random Forest	0.79	0.79	0.79	0.79

Based on the results:

- Random Forest has the highest accuracy (0.79), precision (0.79), recall (0.79), and F1-score (0.79), making it the best-performing model among the three.
- SVM has slightly lower metrics compared to Random Forest but still performs well.
- Logistic Regression has the lowest metrics among the three models.

Popularity Prediction Model Selection

The Random Forest model is recommended for classifying movies as 'Hit' or 'Flop' based on the given features, as it consistently outperforms the other models in terms of accuracy, precision, recall, and F1-score.

C. Award Prediction Comparison

Models used: Logistic Regression, SVM, and Random Forest.

To predict the number of awards a movie will win, we will create bins based on the number of wins and classify movies into these bins. By leveraging features such as genre, directors, votes, metascore, and release year

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.55	0.3346	0.55	0.4159
SVM	0.545	0.3878	0.545	0.4292
Random Forest	0.53	0.4252	0.53	0.4448

Based on the results:

- Random Forest has the highest metrics in most cases, making it a strong candidate for predicting the bin of the number of wins a movie will fall into.
- SVM also performs well, with metrics close to Random Forest.
- Logistic Regression has the lowest metrics among the three models.

Award Prediction Model Selection

The Random Forest model is recommended for predicting the bin of the number of wins a movie will fall into based on the given features, as it consistently outperforms the other models in terms of accuracy, precision, recall, and F1-score.

Conclusion:

- The study demonstrates that machine learning models can effectively predict various success metrics for movies, aiding movie studios, producers, and marketers in making informed decisions about future projects.
- Ridge Regression with polynomial features and cross-validation emerged as the best model for predicting gross revenue, offering a balance between complexity and accuracy. Random Forest was the most effective for classifying movies as 'Hit' or 'Flop' due to its ability to handle complex feature interactions, and it also performed well in predicting the number of award wins, making it versatile for various classification tasks.
- Insights from the data distribution revealed that most movies have budgets below $100 million, indicating a trend towards managing financial risk by limiting production costs.
- The majority of movies do not achieve blockbuster status, highlighting the need for realistic revenue expectations and diversified investments. The critical role of opening weekend performance underscores the importance of effective pre-release marketing, while focusing on high-quality production that can compete in award circuits can enhance a movie's marketability and longevity.

Future Work:

- Explore additional features like marketing spend, social media engagement, and critical reviews to improve model accuracy
- Validate models with new movies to ensure relevance. Develop genre-specific models for more precise predictions. Extend revenue predictions to include streaming, DVD sales, and international box office for a comprehensive financial overview.

Appendix

1. **Python Code:** The complete Python code used for data wrangling, EDA, model training, and evaluation, Final Project_v2.ipynb - Colab (google.com)
2. **Python Code:** The complete Python code used for data scraping, DATA602Project/Top100Movies.py at main · hawa1983/DATA602Project (github.com)
3. Project Presentation Slides, Final Presentation.pptx (live.com)