Movie Revenue Prediction with

Regularized Linear and Polynomial Regression Models

Ashley Lopez, I Man Liu

Northeastern University

DS 4400: Machine Learning and Data Mining 1

Professor Miguel Fuentes-Cabrera

April 4th, 2025

Abstract

Accurately predicting movie revenue is crucial for stakeholders in the film industry, allowing them to make informed financial and marketing decisions. This study explores the effectiveness of various regression models, including linear regression, regularized regression (Ridge and Lasso), and polynomial regression, in forecasting box office revenue. Using a dataset sourced from Kaggle, we conducted comprehensive data preprocessing, feature encoding, and standardization to optimize model performance. Our findings indicate that polynomial regression with a degree of 2 outperformed linear and regularized models, achieving the highest R^2 score of 0.6121, effectively capturing the non-linear relationships between movie features and revenue. While Ridge and Lasso regression reduced model complexity and mitigated overfitting, their predictive performance was slightly lower than that of polynomial regression. Higher-degree polynomial models, however, suffered from overfitting, reducing their generalizability. These results highlight the potential of polynomial regression in revenue prediction and suggest further exploration of feature engineering techniques to enhance predictive accuracy.

Introduction

The movie industry is a high-stakes business where financial success is influenced by a complex interplay of factors, including budget, genre, language, and so on. Accurately predicting a movie's revenue can provide significant advantages to production studios, companies, investors, and marketers, enabling them to make data-driven decisions that minimize financial risk and maximize profitability. The data sourced was from Kaggle: Getting Started with a Movie Recommendation System which contains the movies_metadata.csv file with about 5800

rows and 24 columns such as budget, genre, original_language, popularity, runtime and so on. This file contains a sufficient number of features which can be used in regression models. Regression models are widely used in revenue prediction because they can identify and quantify the relationships between independent variables (e.g., budget, runtime, popularity, original language, etc.) and a dependent variable (box office revenue). By leveraging historical data and machine learning techniques, regression-based approaches allow stakeholders to estimate a film's financial performance with greater accuracy. This paper explores the effectiveness of different regression models in predicting movie revenue, comparing their predictive power and identifying key factors that contribute to box office success. Through this analysis, we aim to provide valuable insights into the application of data-driven methodologies in the film industry, ultimately aiding in more informed decision-making.

Methods

Data Preprocessing

To ensure data quality and consistency, the dataset underwent a comprehensive preprocessing phase before model implementation. Initially, irrelevant columns were removed, retaining only key features: **adult, genres, original_language, budget, popularity, revenue, runtime, and vote_average**. Among these, **adult, genres, and original_language** were identified as categorical features, while **budget, popularity, runtime, and vote_average** were numerical.

Invalid entries were removed, specifically movies with a budget, revenue, or runtime of zero, as well as those lacking genre information (e.g., empty list of genres). Categorical features

were encoded using the **one-hot encoding method**. However, due to the high cardinality of the *original_language* feature (40 unique languages), encoding each language separately would result in a high-dimensional dataset. To mitigate this, we retained the top five languages (**English, Hindi, French, Russian, and Japanese**), categorizing all others under the label **"other"**.

For encoding the *genres* feature, we utilized **MultiLabelBinarizer**, which transformed the genre lists into a binary matrix. This matrix was then incorporated into the original dataframe, ensuring each genre was represented as an independent binary column. After encoding, the dataset comprised **32 columns**, including numerical and categorical features: **adult, budget, popularity, revenue, runtime, vote_average, original_language_en, original_language_fr, original_language_hi, original_language_ja, original_language_other, original_language_ru,** and **20 genre-related binary columns**.

To prepare the data for regression analysis, all numerical features were standardized using **StandardScaler**. Following encoding and scaling, it was observed that the *adult* column exhibited insufficient variation in its values, leading to its removal from the dataset. Consequently, the final dataset after encoding contained **30 independent features**, with *revenue* as the target variable for prediction.

Regression Models

Given the high dimensionality of the dataset after encoding, we initially employed a **linear regression model using gradient descent**, setting a **learning rate of 0.001** and **10,000 iterations** to optimize model performance. A baseline model was first constructed using standard linear regression to establish fundamental relationships between independent variables and revenue.

To enhance generalization and prevent overfitting, we implemented **regularized regression techniques, including Lasso (L1 regularization) and Ridge (L2 regularization)**. These methods penalize large coefficients, reducing model complexity while improving stability and predictive accuracy.

Finally, to capture potential nonlinear relationships between features and revenue, we applied **polynomial regression** with degrees **2, 3, and 4**, expanding the feature space to improve model flexibility and performance.

Model Evaluation

Each model was evaluated using **mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2) scores** to assess predictive performance. Hyperparameter tuning, including **polynomial degree and regularization strength**, was conducted using **cross-validation** to optimize model accuracy and generalization. Additionally, visualization techniques such as **scatter plots and convergence plots** were employed to compare model predictions against actual revenue values.

Through this approach, we aimed to determine the most effective regression model for predicting movie revenue while balancing accuracy and computational complexity.

Results and Discussion

From the regression models, we got some important insights into the predictions of movie revenues. Starting with linear regression, both gradient descent and closed-form solutions produce similar outcomes concerning the **Mean Squared Errors (MSE)** and **R-Squared (R^2)**, showing a strong alignment between the two methods. Specifically, the gradient descent model

achieved an **MSE** of approximately **1.073×10161.073×1016** and an **R²** score of **0.6126**, suggesting that the model was able to capture a significant portion of the variance in the revenue data. The similarity in results from both approaches indicates that the hyperparameters from gradient descent like the learning rate and number of iterations were well-tuned.

Next, the regularized regression models, such as the Ridge and Lasso models, were applied to mitigate overfitting by penalizing large coefficients. Both models yielded similar results, with **MSE** values just slightly higher than the linear regression (around **1.165×10161.165×1016**) and **R²** scores of approximately **0.5848**. These values indicate a slightly lower predictive performance than the gradient descent model but still reflect the **importance of certain features such as budget, popularity, and voting average** in predicting movie revenue. The feature importance analysis revealed that these variables consistently appeared at the top of both Ridge and Lasso models, confirming their substantial impact on revenue prediction.

As for Polynomial regression, the best performance was observed with degree 2 in which it achieved a R^2 score of **0.6121**, a significant improvement compared to the linear models. This suggests that the relationship between the features and revenue is indeed non-linear, which is why this quadratic model better captures the patterns. However, the **higher polynomial degrees**, such as 3 and 4, resulted in **poor performance with R²** scores dropping significantly into negative areas, highlighting overfitting issues. The **second degree polynomial model** offered a reasonable trade-off between complexity and accuracy, **balancing both and explaining 61.21%** of the variance in movie revenue, while higher degrees failed to hold stable and introduced model generalization.

Conclusion

Overall, the linear regression model emerged as the most effective approach for predicting movie revenue, out-performing both the regularized regression models and polynomial models. It's worth noting that the linear regression model performs only slightly better than the polynomial regression model of degree 2 measured by R^2 scores. The non-linear relationships captured by the polynomial model proved to be crucial, while higher degree polynomials led to overfitting, as seen in the drop in R^2 scores. While the Ridge and Lasso regression models provided similar predictive performance, their utility lies in their ability to reduce model complexity and improve generalization. Future work could include more exploration for feature engineering or selection to improve model performance and rescue prediction error. Additionally, addressing the high variance in performance across different data splits, as indicated by the standard deviation cross-validation scores could lead to a more robust model.