Predicting Movie ROI Using Advanced Modeling Techniques

Arkadiy Ekshtat, Austin Evans, Dongsuk Kim, Kenneth Mohr, Linh Nguyen

Introduction

The global film industry has exhibited consistent ~5% year-over-year growth, driven by emerging markets, streaming platforms, advanced film technology, and increasing diversity in storytelling. However, predicting a film's financial success remains a complex challenge. Investors and producers seek analytical tools to optimize investment strategies, marketing efforts, and release platforms. This project explores key factors influencing a movie's financial success using data from IMDB, TMDB, and other sources, with an innovative approach incorporating NLP techniques such as plot summary analysis to improve predictive accuracy.

Problem definition

Predicting a movie's return on investment (ROI) is challenging due to numerous factors such as budget, genre, marketing strategies, and audience reception. This project aims to develop predictive models that provide actionable insights into maximizing ROI [2], [5].

Given a dataset $D = \{(X_1, Y_1), (X_2, Y_2), \ldots\ldots (X_n, Y_n)\}$, where X_i represents features such as budget, genre, runtime, keywords, etc., and Y_i represents the Revenue of the corresponding movie, the objective is to learn a function $f(X)$ such that $Y = f(X)$ predicts Revenue with minimized error and use that predicted Revenue to calculate the predicted ROI. Additionally, for categorical success prediction, it is based on the vote average (rating) and label movies as "bad" (0) or "average" (1) or “good” (2) based on defined thresholds [8].

Proposed method

Data Preparation Steps

Following steps ensured a clean and feature-rich dataset for robust modeling

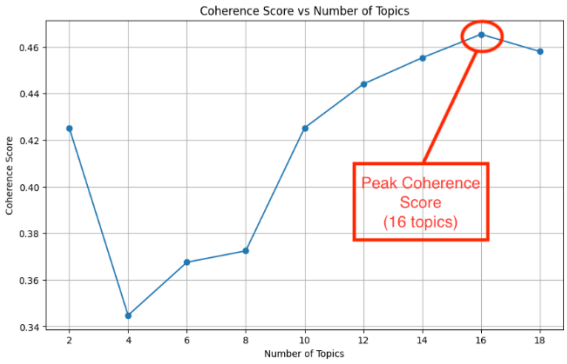
- **Data Loading and Merging:** Loaded tmdb_5000_credits.csv, tmdb_5000_movies.csv, train.csv, and test.csv. Combined tmdb_5000_credits and tmdb_5000_movies using the common id field for a unified dataset. Similarly, merged train and test datasets, ensuring consistency for modeling [12].
- **Data Cleaning** [1], [12]
  - **Handling Missing Values**: Dropped rows with NaN values for essential features (e.g., budget, revenue). For non-critical columns (e.g., tagline), fill NaN with default placeholders or left unchanged.
  - **Data Type Conversion**: Converted columns like budget and revenue to numeric data types, ensuring proper handling of invalid or missing entries.
  - **Handling Outliers** Removed the movies data if budget & revenue are below $1,000. Removed the movies data if budget is greater than $150M or revenue is greater than $200M. Removed the movies data if ROI is greater than 1000%. Removed the movies data prior to 1990
- **Data Transformation** [2], [10]
  - **Homepage Encoding**: Converted homepage column to binary (1 for presence, 0 for absence).
  - **Genre and Keywords**: One-hot encoded the top 10 genres, keywords, production companies, and production countries for easy model integration.

- **Spoken Languages**: Added columns to indicate if "English" was included and counted the total number of spoken languages.
- **Adjust Budget and revenue based on inflation**: adjust the budget and revenue based on inflation rate.
- **ROI Calculation**: Defined ROI as (revenue−budget)/budget×100 and only consider movies with ROI under 1000%.
- **Cast Count**: Parsed JSON-like data in the cast column to calculate the number of cast members, adding a new column number of cast.
- **Rating category:** movies with ratings below 6 will be considered bad movies, movies with ratings from 6 to 6.5 will be considered average movies, movies with ratings above 6.5 will be considered good movies.

Feature Extraction through NLP

In this study, the use of Natural Language Processing (NLP) techniques was instrumental in enriching a dataset of films with additional structured features. These features were aimed at facilitating predictive modeling and visualization of factors influencing film success, including thematic content, sentiment, and the presence of influential industry figures. NLP was able to offer an innovative way to extract and analyze narrative and emotional elements from movie data [7], [13].

- **Topic Modeling**: Thematic content was extracted using Latent Dirichlet Allocation (LDA), a probabilistic model that uncovers hidden topics within a corpus [15]. The keyword data was first preprocessed using the nltk package to remove English stop words. Additionally, a list of movie industry stop words, such as actor, director, credits, etc., was manually created and removed from the keywords data as well. Following the preprocessing of keywords data, the optimal number of topics was determined through a coherence-driven approach (see figure 1). Each film was assigned a dominant topic based on its Bag of Words representation. This approach aligns with the framework outlined by Jelodar et al. [15], who highlight LDA's versatility in text mining and its capacity to uncover latent structures in textual data. This study builds on those principles by demonstrating their applicability to film-related data.



- **Sentiment Analysis:** Sentiment analysis was conducted on the movie overviews, employing the transformer-based model *distilbert-base-uncased-finetuned-sst-2-english*. This analysis categorized each film as having a positive or negative sentiment and provided a confidence score for the classification. This work contributes to understanding the emotional tone of films, complementing prior research that identifies sentiment as a significant predictor of audience engagement and box office performance [6].

- **NER :** Named Entity Recognition (NER) was applied to identify the presence of high-value actors and directors using predefined lists of influential names from IMDB's Top 100 Actors and Top 25 Directors lists. This feature extraction aims to quantify the impact of key industry figures, reflecting a simplified adaption of the methodologies described by Wang et al. who leveraged a heterogeneous network embedding model to capture the interplay among actors, directors, and production companies [3].

The NLP-derived features—dominant topics, sentiment scores, and counts of high-value collaborators—were incorporated into the dataset to improve predictive modeling. By integrating LDA-based topic modeling, sentiment analysis, and NER, this study adds to the growing literature advocating for the inclusion of free text data in revenue prediction models, beyond traditional movie metadata. This approach highlights the value of combining diverse data sources and methods to gain deeper insights into the complexities of the film industry.

Modeling

1) Revenue Prediction:

Applied a log transformation to revenue to reduce skewness. Selected a list of numerical and categorical features encoded as predictors and "log_adjusted_revenue" as the target variable. Divided the dataset into training (80%) and test (20%) sets to evaluate model performance [4], [9].

- **Linear regression:** employed as a baseline model. It fits a regression line that minimizes residual errors, defined as the cumulative difference between predicted and actual values. As emphasized by Memon and Hussain (2024), this supervised learning method relies on selecting input features strongly correlated with the target variable to enhance model validity and reduce irrelevant influences [5], [11].
- **Random Forest:** used to enhance predictive accuracy and reduce overfitting. It aggregates the predictions of multiple decision trees trained on bootstrapped subsets of the data. This model is beneficial for predicting revenue as it can capture complex non-linear relationship between the features and the target variable.
- **Gradient Boosting:** applied to leverage iterative improvement in model performance. By optimizing weak learners (typically decision trees), it minimizes prediction errors through gradient descent, adjusting subsequent trees to address residual errors from prior iterations. This model is well-suited in this case because it can iteratively refine predictions, providing improved accuracy.
- **XGBoost:** incorporated as a high-performance implementation of Gradient Boosting, optimized for speed and accuracy. It integrates regularization techniques to mitigate overfitting while efficiently handling large datasets and high-dimensional feature spaces. This model is efficient because of the ability to manage both regularization and feature importance, especially when dealing with a large set of features in our case.

2) Success Prediction

The dataset was split into training (80%) and test (20%) sets, with features standardized using "StandardScaler". The model was trained using a multinomial logistic regression approach to predict the "movie_rating_encode" target variable.

- **Multinomial Logistics Regression:** Used as a baseline model for classification, multinomial logistic regression predicts the target variable by modeling the relationship between features and multiple possible outcomes. It is suitable for classifying movie success, where the target variable has more than two categories [14].

- **Random Forest:** According to Arnab (2021), Random Forest classifiers are a type of ensemble classifier that fits a number of decision trees on different sub-samples of the dataset and uses an averaging method [12]. It is beneficial for movie success prediction as it can capture complex interactions between features, making it robust to variations in data, and is especially useful in predictive tasks with a large number of features.
- **Gradient Boosting:** Gradient boosting iteratively improves model performance by optimizing weak learners (decision trees). This model is effective in movie success classification due to its ability to minimize prediction errors and enhance accuracy through iterative adjustments to the model. This technique is widely used for classification tasks due to its efficiency in boosting model performance.
- **SVM:** SVM is a powerful supervised learning algorithm that creates a hyperplane to separate classes based on feature space. It works well for movie success prediction as it can handle high-dimensional data and non-linear relationships, offering strong classification performance when dealing with complex patterns in the dataset. As Dhir and Raj (2018) explain, SVM is particularly effective in high-dimensional spaces, which is crucial when working with datasets containing multiple features [8].

Experiments/ Evaluation

Revenue Prediction Model- Predictive Model Analysis & Evaluation

Model analysis with NLP

Metric	Linear Regression	Random Forest	Gradient Boosting	XGBoost
Model Score (%)	50.284537	57.763161	62.248227	52.737469
R^2 Score	0.502845	0.577632	0.622482	0.527375
Mean Absolute Error (MAE)	0.892349	0.774530	0.750376	0.828862
Mean Squared Error (MSE)	1.435112	1.219230	1.089762	1.364304
Root Mean Square Error (RMSE)	1.197961	1.104187	1.043917	1.168034
Normalized MSE	0.497155	0.422368	0.377518	0.472625
Explained Variance Score (EVS)	0.504049	0.577932	0.622652	0.527508
Max Error	6.172441	5.079796	4.486191	5.381898
Mean Absolute Percentage Error (MAPE)	5.635498	4.896024	4.734296	5.225527

- **Linear regression:** The linear regression model showed moderate performance with a model score of 50.28% (R^2 = 0.5028), explaining about half of the target variance. With a Mean Absolute Error (MAE) of 0.8923 and RMSE of 1.1980, its predictions were reasonable but could be improved. The MAPE of 5.64% indicates a moderate percentage error, while the maximum error of 6.1724 highlights some larger prediction discrepancies.
- **Random Forest:** The Random Forest model performed moderately well with a model score of 57.76% (R^2 = 0.5776), capturing 58% of the variance. It demonstrated good accuracy, reflected by an MAE of 0.7745 and RMSE of 1.1042. The Normalized MSE of 0.4224 and EVS of 0.5779 confirm its strong variance capture. The MAPE of 4.90% and maximum error of 5.0798 suggest relatively low prediction errors.
- **Gradient Boosting:** Gradient Boosting outperformed the other models with a model score of 62.25% (R^2 = 0.6225), explaining 62% of the variance. Its MAE of 0.7504 and RMSE of 1.0439 indicate good prediction accuracy. With a Normalized MSE of 0.3775 and EVS of 0.6227, it effectively captured variance. The MAPE of 4.73% and maximum error of 4.4862 demonstrate low prediction errors, making it the best-performing model.
- **XGBoost:** XGBoost showed moderate performance with a model score of 52.74% (R^2 = 0.5274), explaining 53% of the variance. Its MAE of 0.8289 and RMSE of 1.1680 indicate lower prediction accuracy compared to Gradient Boosting. The Normalized MSE of 0.4726 and EVS of 0.5275

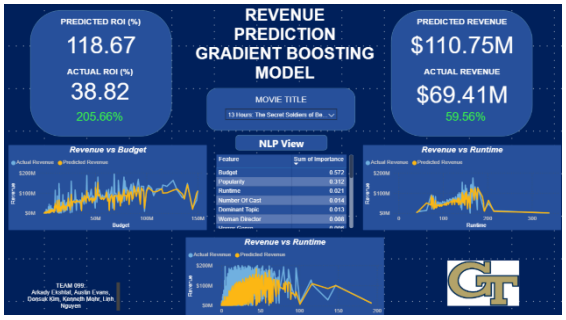
suggest weaker variance capture. The MAPE of 5.23% and maximum error of 5.3819 reflect higher prediction errors.

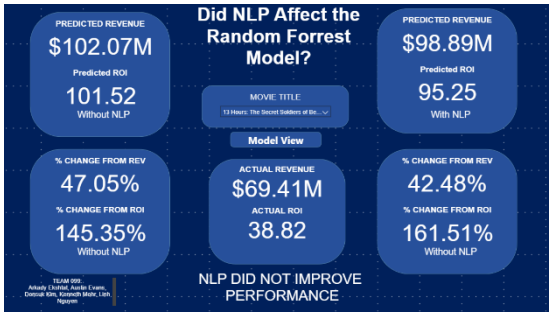
Model analysis without NLP

Metric	Linear Regression	Random Forest	Gradient Boosting	XGBoost
Model Score (%)	50.409933	56.420643	62.282561	48.001606
R^2 Score	0.504099	0.564206	0.622826	0.480016
Mean Absolute Error (MAE)	0.890917	0.785581	0.748765	0.868201
Mean Squared Error (MSE)	1.431492	1.257984	1.088771	1.501012
Root Mean Square Error (RMSE)	1.196450	1.121599	1.043442	1.225158
Normalized MSE	0.495901	0.435794	0.377174	0.519984
Explained Variance Score (EVS)	0.505124	0.564375	0.622910	0.480434
Max Error	6.213446	5.001248	4.648886	5.528892
Mean Absolute Percentage Error (MAPE)	5.630052	4.972739	4.726863	5.494726

To assess the impact of NLP, the same modeling analysis was conducted without NLP-enriched data. The results revealed that Natural Language Processing (NLP) did not significantly enhance predictive accuracy across most models. However, the XGBoost model showed measurable improvement with NLP integration. These findings suggest that while NLP had a positive impact on specific models, its contribution was insufficient to meaningfully improve the performance of the leading model, thereby limiting its overall applicability in this context.

Revenue Prediction Model- Visualization





Based on the results of the model analysis, interactive business intelligence (BI) reports were developed utilizing the two top-performing models, **Gradient Boosting** and **Random Forest**. These reports are designed to dynamically update based on the selected movie, offering comprehensive insights into the predictive accuracy of the models. By comparing the actual and predicted values for revenue and return on investment (ROI) for individual movies, these reports enable a more detailed understanding of model performance and its applicability to specific scenarios within the film industry.

Predictive Model Analysis & Comparison- Success Prediction

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	Confusion Matrix
Logistic Regression	0.674095	0.669642	0.676875	0.669631	[[86, 17, 5], [39, 60, 22], [7, 27, 96]]
Random Forest	0.643454	0.636190	0.642875	0.637210	[[76, 23, 9], [38, 57, 26], [10, 22, 98]]
Gradient Boosting	0.626741	0.635663	0.628353	0.628735	[[74, 25, 9], [40, 67, 14], [8, 38, 84]]
SVM	0.657382	0.650418	0.661059	0.651334	[[87, 15, 6], [41, 55, 25], [6, 30, 94]]

- **Multinomial Logistics Regression:** The Multinomial Logistic Regression model demonstrated solid performance, achieving an accuracy of 67.41%. It maintained a macro-averaged precision of 66.96%, recall of 67.69%, and an F1-score of 66.96%, indicating a balanced ability to predict across all classes. The confusion matrix shows relatively high precision in distinguishing true positives for each class. Despite its strong performance, there were notable misclassifications, particularly in the middle class (class 2), which could be a focus for improvement.
- **Random Forest:** The Random Forest model achieved a slightly lower accuracy of 64.35%, with macro-averaged precision at 63.62%, recall at 64.29%, and an F1-score of 63.72%. The confusion matrix suggests moderate performance, particularly in identifying the third class accurately. While its performance was decent, it struggled with distinguishing some borderline cases, making it less robust compared to logistic regression.
- **Gradient Boosting:** Gradient Boosting had the lowest accuracy among the models at 62.67%, with macro-averaged precision of 63.57%, recall of 62.83%, and an F1-score of 62.87%. The confusion matrix indicates some improvement in precision for the second class but a trade-off with the third class. Although Gradient Boosting captured patterns effectively in certain cases, its overall performance lagged behind Logistic Regression and Random Forest.
- **SVM:** The SVM model displayed competitive performance, with an accuracy of 65.74%. It achieved macro-averaged precision of 65.04%, recall of 66.11%, and an F1-score of 65.13%. The confusion matrix highlights its strength in accurately predicting the first and third classes, while misclassifications in the middle class remain a challenge. Overall, SVM provides a good balance between precision and recall, making it a reliable alternative to Logistic Regression.

Conclusions and discussion

Predicting movie revenue/ROI remains a complex challenge in the global film industry due to the influence of numerous factors such as budget, genre, marketing strategies, and audience reception. This study explored advanced predictive techniques, identifying Gradient Boosting as the most reliable model for ROI prediction, offering better accuracy than the other models. However, the inclusion of NLP, while enhancing the performance of the XGBoost, did not improve all other models including Gradient Boosting's results, limiting its overall contribution [6], [12].

1) Gradient Boosting was identified as the most effective model, providing robust predictions for revenue in the dynamic and unpredictable film industry.
2) The effectiveness of NLP varied across models. A selective and targeted application of NLP for prediction tasks is recommended to maximize its benefits.

In addition to revenue and ROI prediction, the success classification task identified Multinomial Logistic Regression as the most consistent and accurate model. While NLP features improved the accuracy of most models by 2%, they slightly reduced the performance of the Random Forest Classifier. Nevertheless, Logistic Regression consistently achieved the highest accuracy, irrespective of the inclusion of NLP features, reaffirming its position as the best-performing model for this task. These results underscore the robustness of Multinomial Logistic Regression in classification tasks, particularly when the dataset is well-prepared, and features are carefully engineered.

Throughout the project, our team successfully achieved clear outcomes by leveraging innovative technologies, with all team members contributing equally and collaboratively to its success.

Reference

1. Udandarao, V., & Gupta, P. (2024). *Movie Revenue Prediction Using Machine Learning Models.* arXiv. https://doi.org/10.48550/arXiv.2405.11651
2. Ahmad, Ibrahim Said, et al. *A Survey on Machine Learning Techniques in Movie Revenue Prediction.* SN Computer Science, vol. 1, 2020, https://doi.org/10.1007/s42979-020-00249-1
3. Wang, Zhaoyuan, et al. *Predicting and Ranking Box Office Revenue of Movies Based on Big Data.* Information Fusion, vol. 60, 2020, pp. 25–40. ScienceDirect, https://doi.org/10.1016/j.inffus.2020.02.002
4. Katarzyna Pekala, Katarzyna Woznica, Przemyslaw Biecek. *Triplot: Model Agnostic Measures and Visualizations for Variable Importance in Predictive Models That Take into Account the Hierarchical Correlation Structure.* https://doi.org/10.48550/arXiv.2104.03403
5. Memon, Z.A., & Hussain, S.M. *Predicting Movie Success Based on Pre-Released Features.* Multimed Tools Appl 83, 20975–20996 (2024). https://doi.org/10.1007/s11042-023-16319-4
6. Chintagunta, Pradeep K., et al. *The Effects of Online User Reviews on Movie Box-Office Performance.* SSRN, https://doi.org/10.2139/ssrn.1331124
7. *Electronic Word-of-Mouth, Box Office Revenue and Social Media.* Electronic Commerce Research and Applications, vol. 22, March–April 2017, https://doi.org/10.1016/j.elerap.2017.02.001
8. Dhir, R., & Raj, A. *Movie Success Prediction Using Machine Learning Algorithms and Their Comparison.* ICSCCC, 2018, https://doi.org/10.1109/ICSCCC.2018.8703320
9. Mahmud, Quazi Ishtiaque, et al. *A Machine Learning Approach to Predict Movie Revenue Based on Pre-Released Movie Metadata.* Journal of Computer Science (2020).
10. Quader, N., et al. *A Machine Learning Approach to Predict Movie Box-Office Success.* ICCIT (2018). https://doi.org/10.1109/iccitechn.2017.8281839
11. Agarwal, M., et al. (2021). *A Comprehensive Study on Statistical Techniques for Movie Success Prediction.* https://doi.org/10.5121/csit.2021.111802
12. Arnab Sen Sharma, et al. *A Larger Movie Dataset and Pre-Released Attributes' Effects on Revenue.* https://doi.org/10.48550/arXiv.2110.07039
13. Asur, A., & Huberman, B.A. *Social Media Metrics and Box Office Revenues.* (2010).
14. Nihalaani, R., et al. *Naïve Bayes, Logistic Regression, and SVM in Movie Prediction.*
15. Jelodar, H., et al. *Latent Dirichlet Allocation (LDA) and Topic Modeling: A Survey.* arXiv.