Predicting Movie ROI Using Advanced Modeling Techniques

CSE 6242 Final Project Report Poster

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 & OBJECTIVE

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 to provide actionable insights for maximizing ROI**

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.

PROPOSED METHOD

DATA PREPERATION

 # Movie final data:
- 2183 records/movies

- **Data Source :** Loaded, merged 5000 TMDB movie and credit data and prepared train and test dataset
- **Data Cleansing** : Performed basic data cleaning (non-value data) and removed any outliers (Extreme values from budget, revenue or ROI data, Movies prior to 1990)
- **Data Transformation :** Converted and added new calculated columns for further analysis (Homepage - Y/N, English movie –Y/N, Adjusted ROI with inflations, Number of casts, Rating Category, Genre and Keywords)

FEATURE ENRICHMENT THROUGH NLP

Additional data sources used:
- 100_Greatest_Actor
- 25_Greated_Directors

- **Topic Modeling:** Latent Dirichlet Allocation (LDA) extracted themes, assigned dominant topics, and uncovered latent structures in film data, using Jelodar et al.'s framework, showcasing LDA's effectiveness in uncovering latent structures
- **Sentiment Analysis**: Analyzed movie overviews using the transformer-based model distilbert-base-uncased-finetuned-sst-2-english to classify sentiment as positive or negative, with confidence scores. This highlights sentiment's role in predicting audience engagement and box office success
- **NER:** Named Entity Recognition identified top actors and directors using IMDB’s influential lists, quantifying their impact. This adapts Wang et al.'s approach by focusing on key industry figures.

MODELING

- **Revenue Prediction:** Log-transformed revenue, training (80%) & testing (20%) data split, using linear regression, Random Forest, Gradient Boosting, and XGBoost for prediction. Each model enhanced accuracy & minimized errors, with XGBoost excelling through regularization and efficiency.
- Additionally, we implemented a success prediction model using multinomial logistic regression based on vote ratings for additional validation.

EXPERIEMNT & EVALUATION

MODEL EVALUATION – Revenue Prediction

Evaluation with NLP enriched data

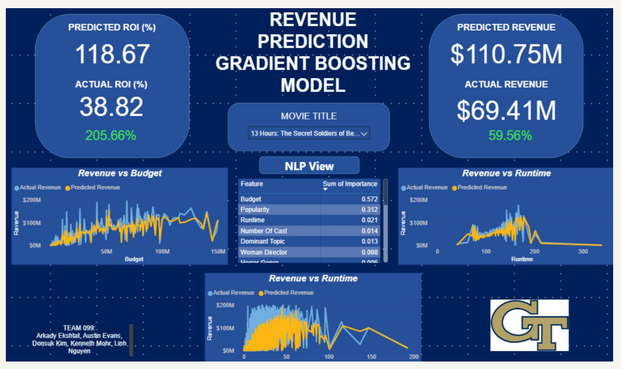
Metric	Linear Regression	Random Forest	Gradient Boosting	XGBoost
Model Score (%)	50.284537	57.763161	62.248227	52.737469
R² Score	0.502845	0.577632	0.622482	0.527375
Mean Absolute Error (MAE)	0.892349	0.774530	0.750376	0.828862
Mean Squared Error (MSE)	1.435112	1.219230	1.089762	1.364304
Root Mean Square 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

Evaluation without NLP enriched data

Metric	Linear Regression	Random Forest	Gradient Boosting	XGBoost
Model Score (%)	50.409933	56.420643	62.282561	48.001606
R² 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

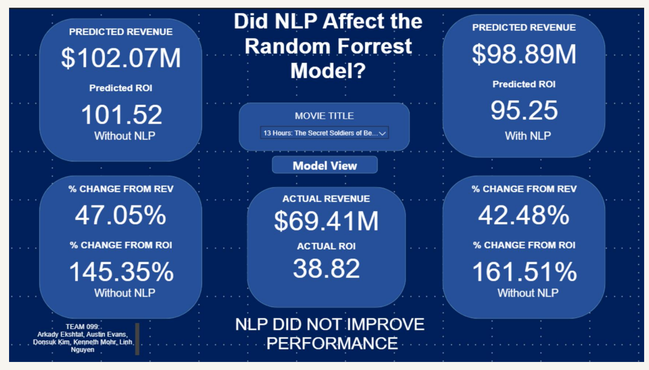
- The results indicate that **Gradient Boosting** achieved the highest performance among all models, with the best R^2 score, lowest MAPE, and smallest maximum error, **followed by Random Forest** as the second-best model. Linear Regression and XGBoost demonstrated lower R^2 scores and higher prediction errors, highlighting comparatively weaker performance.
- Additionally, **the integration of NLP had a limited overall impact on model performance**, with notable improvements observed only in XGBoost, while other models exhibited minimal differences between assessments with and without NLP.

MODEL VISUALIZATION



Revenue prediction with Gradient Boosting model (with NLP)

Revenue prediction with Random Forrest model (with & without NLP)



Illustrative Purposes Only – Values based on a single movie example

Based on the model analysis results, **interactive BI reports** were created utilizing the top-performing models, Gradient Boosting and Random Forest. These reports dynamically update based on movie selection, **providing insights into prediction accuracy** by comparing actual and predicted values for revenue and ROI **for specific movies.**

CONCLUSION

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.