Seneca
POLYTECHNIC

Course name:	Machine Learning
Course code:	AIG100NAA.09980.2241
Professor:	Asad Norouzi
Assignment name:	Project 2
Student name:	Yuliya Dubovichenko

TABLE OF CONTENTS

OBJECTIVES

Regression Objective

Objective: Predict the box office collection (revenue) of movies based on various features.

Justification: By accurately predicting the box office collection, movie producers can better understand the potential financial success of their films and make informed decisions regarding budget allocation, marketing strategies, and distribution plans.

Classification Objective

Objective: Classify movies into "Successful" or "Unsuccessful" categories based on their box office performance.

Justification: By classifying movies into successful or unsuccessful categories, stakeholders can gain insights into the factors that contribute to a movie's commercial success. This classification can aid in identifying trends, patterns, and characteristics associated with successful films, thereby guiding future production and investment decisions.

INTRODUCTION

In the competitive landscape of the film industry, predicting the success of a movie is a complex and multifaceted task. The Technical Oscar awards serve as a prestigious recognition of excellence in filmmaking, making them a valuable metric for gauging a movie's impact and success. This project aims to develop robust predictive models to determine whether a movie will win a Technical Oscar award. Leveraging a dataset titled "Movie_classification.csv", encompassing information on 506 movies, including numerical and categorical variables, we embarked on a journey of data preprocessing, model building, and feature analysis to unravel the intricate dynamics influencing a movie's likelihood of winning such accolades.

DATA PREPROCESSING

The journey commenced with meticulous data preprocessing to ensure the integrity and quality of our analyses. Missing values in the 'Time_taken' column were imputed using the median, ensuring minimal data loss. Categorical variables underwent one-hot encoding, enabling the incorporation of qualitative information into our models. Moreover, numerical features were scaled using MinMaxScaler to mitigate the impact of varying scales on model performance.

Feature engineering endeavors led to the creation of novel features such as the marketing budget ratio and the product of lead actors' ratings, enriching our dataset with deeper insights.

OUTLIER DETECTION AND REMOVAL

The pursuit of robust predictive models necessitated the identification and elimination of outliers lurking within our dataset. Employing the interquartile range (IQR) method, outliers in numerical features were pinpointed and subsequently excised. This meticulous process not only enhanced the reliability of our models but also fortified their resilience against spurious influences.

MODEL BUILDING AND EVALUATION

In the ever-evolving landscape of the film industry, understanding the factors that contribute to the success of movies is paramount for stakeholders, including producers, investors, and distributors. Accurately predicting the box office performance of films can empower decision-makers to allocate resources effectively, devise strategic marketing plans, and optimize distribution strategies.

In this report, we delve into the realm of predictive modeling to forecast movie box office collections and classify films into successful or unsuccessful categories. By leveraging machine learning algorithms, we aim to uncover patterns, relationships, and key features that influence a movie's commercial success.

Model Performance Comparison

Logistic Regression vs. Decision Tree Classifier

Logistic Regression

Accuracy: 0.563

Precision: Class 0 - 0.65, Class 1 - 0.50

Recall: Class 0 - 0.50, Class 1 - 0.65

F1-score: Class 0 - 0.56, Class 1 - 0.56

Decision Tree Classifier:

Accuracy: 0.606

Precision: Class 0 - 0.64, Class 1 - 0.55

Recall: Class 0 - 0.68, Class 1 - 0.52

F1-score: Class 0 - 0.66, Class 1 - 0.53

Findings: Decision Tree Classifier outperforms Logistic Regression in terms of accuracy, precision, recall, and F1-score.

Linear Regression vs. Decision Tree Regression

Linear Regression:

Mean Squared Error (MSE): 0.245

Decision Tree Regression:

MSE: 0.394

Findings: Linear Regression achieved a lower MSE compared to Decision Tree Regression, indicating better performance in predicting continuous outcomes.

Logistic Regression vs. Support Vector Machine (SVM)

Logistic Regression:

Accuracy: 0.563

Precision: Class 0 - 0.65, Class 1 - 0.50

Recall: Class 0 - 0.50, Class 1 - 0.65

F1-score: Class 0 - 0.56, Class 1 - 0.56

SVM:

Accuracy: 0.606

Precision: Class 0 - 0.68, Class 1 - 0.54

Recall: Class 0 - 0.57, Class 1 - 0.65

F1-score: Class 0 - 0.62, Class 1 - 0.59

Findings: SVM outperforms Logistic Regression in terms of accuracy, precision, recall, and F1-score.

Random Forest Regressor Feature Importance

- Features are ranked based on their importance scores.
- Top features include 12, 11, 15, 10, and 3, indicating their significant impact on predicting the target variable.

Gradient Boosting Classifier Feature Importance

- ✓ Features are ranked based on their importance scores.
- ✓ Top features include 3, 0, 15, 12, and 8, suggesting their importance in predicting the target variable.

FEATURE IMPORTANCE ANALYSIS

Feature importance analysis is crucial in understanding the contribution of each feature towards predicting the target variable. Let's explore the feature importance rankings for both the Random Forest Regressor and Gradient Boosting Classifier models.

Random Forest Regressor Feature Importance

Feature 12 (0.087761): This feature holds the highest importance according to the Random Forest Regressor model, indicating its strong influence on predicting the target variable.

Feature 11 (0.087716): Following closely behind, Feature 11 also exhibits significant importance in the model's decision-making process.

Feature 15 (0.084343): Feature 15 ranks third in importance, suggesting its substantial impact on the target variable prediction.

Feature 10 (0.083994): Feature 10 holds considerable importance in the model, contributing significantly to the predictive power of the Random Forest Regressor.

Feature 3 (0.077067): Feature 3 is also among the top features, indicating its relevance in predicting the target variable.

Gradient Boosting Classifier Feature Importance

Feature 3 (0.139516): According to the Gradient Boosting Classifier model, Feature 3 holds the highest importance, implying its strong influence on classification decisions.

Feature 0 (0.116669): Feature 0 ranks second in importance, suggesting its significant contribution to the model's predictive performance.

Feature 15 (0.111658): Similar to the Random Forest Regressor model, Feature 15 holds substantial importance in the Gradient Boosting Classifier as well.

Feature 12 (0.086833): Feature 12 is among the top features, indicating its relevance in predicting the target variable.

Feature 8 (0.080805): Feature 8 also exhibits significant importance, contributing notably to the model's decision-making process.

- ✓ Both models identify similar features as important, albeit with some differences in ranking.
- ✓ Features such as 3, 15, and 12 consistently appear among the top features in both models, suggesting their critical role in predicting the target variable.
- ✓ Understanding feature importance aids in feature selection, model interpretation, and identifying key drivers of the target variable, thereby guiding decision-making processes in real-world scenarios.

RESULTS AND IMPLICATIONS

The culmination of our endeavors unveils a nuanced understanding of the intricate dynamics governing success in the film industry. While budget and marketing endeavors wield significant influence, factors such as movie length and lead actor/actress ratings also exert profound impacts on a movie's likelihood of winning Technical Oscars. The predictive models developed herein serve as invaluable tools for filmmakers and production houses, facilitating informed decision-making and resource allocation strategies to optimize their chances of success.

Potential Implications in Real-World Scenarios

The results of our analysis carry significant implications for stakeholders across the film industry, offering valuable insights that can inform strategic decision-making and enhance the likelihood of success in real-world scenarios.

For Movie Producers and Studios

Budget Allocation: Accurate predictions of box office revenue enable producers to allocate budgets more effectively, ensuring optimal resource utilization and maximizing returns on investment.

Marketing Strategies: Insights into factors influencing box office performance aid in devising targeted marketing campaigns, allowing producers to reach their target audience more efficiently and drive ticket sales.

Production Decisions: Understanding the characteristics associated with successful films empowers studios to make informed decisions regarding script selection, casting choices, and genre preferences, thereby increasing the probability of producing commercially viable movies.

For Distributors and Exhibitors

Screening Prioritization: Classification of movies into successful and unsuccessful categories facilitates informed decisions regarding which films to prioritize for distribution and exhibition, optimizing theater scheduling and maximizing box office returns.

Negotiation Leverage: Insights into box office performance trends provide distributors with valuable leverage in negotiations with studios and production houses, enabling more favorable distribution deals and revenue-sharing agreements.

For Investors and Financiers

Risk Assessment: Predictive models offer investors a quantitative framework for assessing the financial viability of potential film investments, mitigating risks associated with uncertainty in box office performance and enhancing portfolio diversification strategies.

Portfolio Management: Classification of movies based on their commercial success allows investors to construct well-balanced portfolios comprising a mix of high-performing and low-risk films, optimizing returns while managing exposure to market volatility.

For Market Analysts and Researchers

Industry Insights: Analysis of feature importance and model performance provides valuable insights into the underlying factors driving box office success, enabling market analysts and researchers to identify industry trends, consumer preferences, and emerging patterns in film consumption behavior.

Forecasting Trends: Predictive models can be leveraged to forecast future box office trends, allowing industry analysts to anticipate shifts in market dynamics, adapt strategies accordingly, and capitalize on emerging opportunities.

The implications of our results extend beyond the confines of academic research, offering tangible benefits and actionable insights for stakeholders across the film industry. By leveraging predictive models and classification techniques, industry professionals can navigate the

complexities of film production, distribution, and investment with greater confidence and precision, ultimately enhancing the likelihood of success and driving sustainable growth in the ever-evolving landscape of the entertainment industry.

CONCLUSION

In conclusion, this project epitomizes the fusion of data science methodologies with the realm of filmmaking, offering unprecedented insights into the determinants of success in the film industry. The predictive models developed herein herald a new era of data-driven decision-making, empowering stakeholders to navigate the complexities of the industry with confidence and precision. As the landscape of filmmaking continues to evolve, the insights gleaned from this project serve as a beacon of guidance, illuminating pathways to success amidst an ever-changing landscape.

In summary, we have achieved both primary objectives by developing predictive regression and classification models and evaluating their performance using relevant metrics. These models provide valuable insights for stakeholders in the film industry to make informed decisions and optimize their chances of success.

REFERENCES

Python code: Provided within the report.

Documentation: Sklearn and Seaborn documentation.

Data source: Movie_classification.csv dataset.

1. Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
2. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189-1232.
3. Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning*. Packt Publishing Ltd.
4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R*. Springer.