Classification Models for Movie Success Prediction

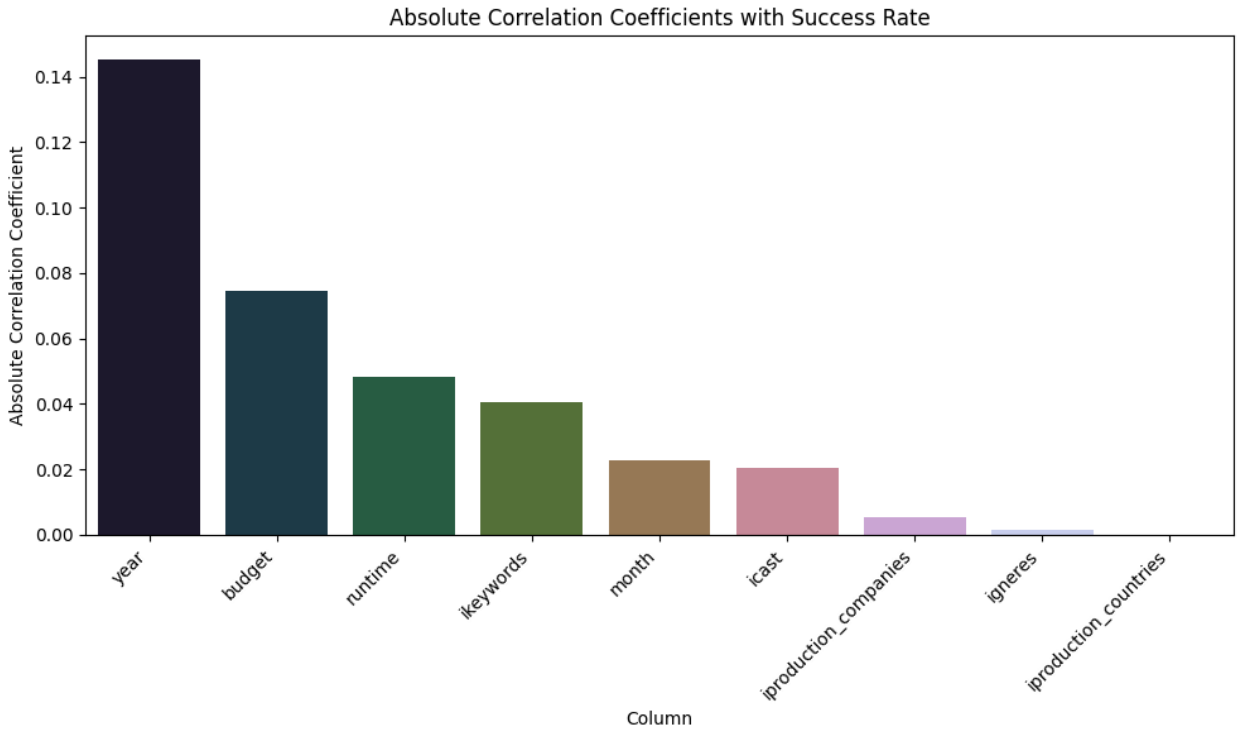
Introduction:

The primary objective of this notebook is to analyze and predict the success of movies based on various features using classification models. The dataset used contains information about 4803 movies, including budget, runtime, release year, genres, keywords, production companies, production countries, cast, and success rate.

Data Preprocessing:

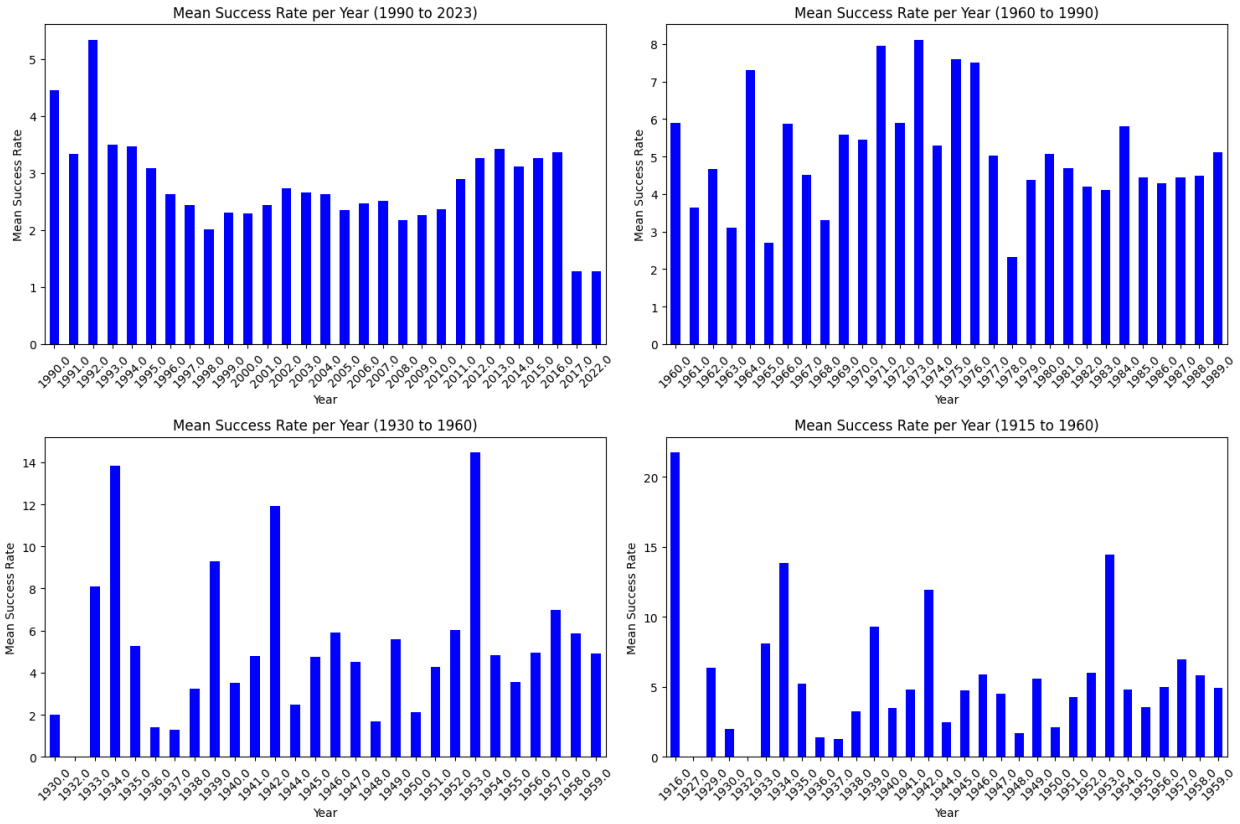
Before building the classification models, thorough data preprocessing was conducted. Missing values were handled, and irrelevant columns such as 'homepage', 'title_y', 'status', and 'original_title' were dropped. Additionally, categorical variables like genres, keywords, production companies, production countries, and cast were encoded for model compatibility.

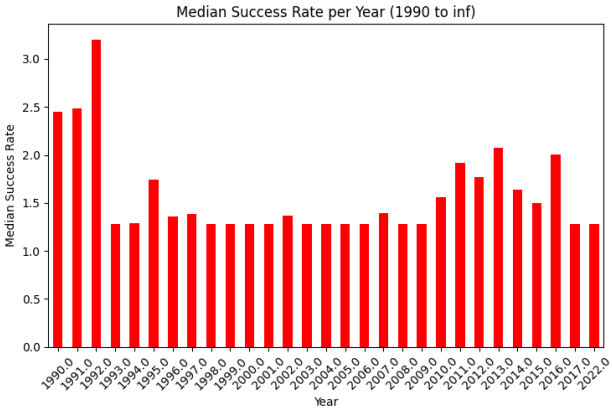
Exploratory Data Analysis (EDA):

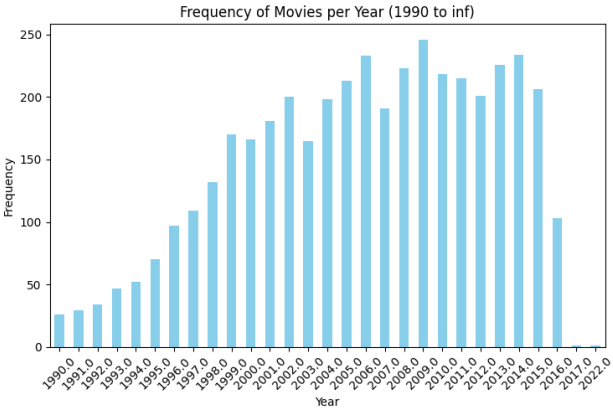


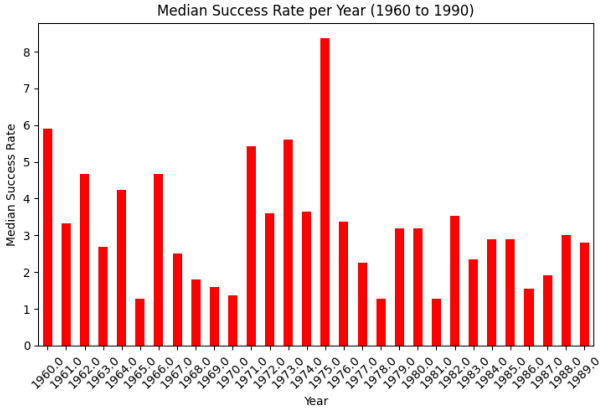
As we can see, year, budget, and runtime have the most correlation with the success of movies, but there are no strong correlations.

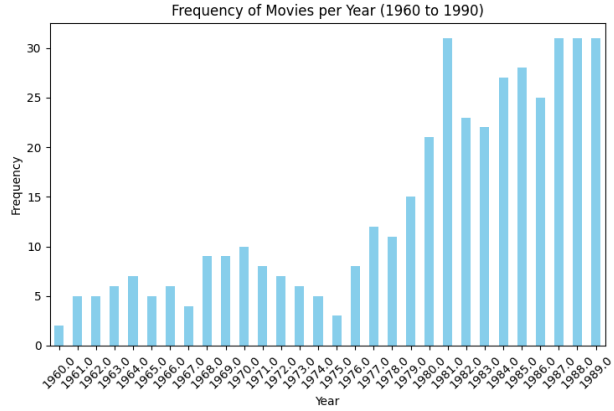
In this picture, we can see a scatter plot that shows the success rate of movies per budget. We observe that most movies with higher success rates are low-budget movies.

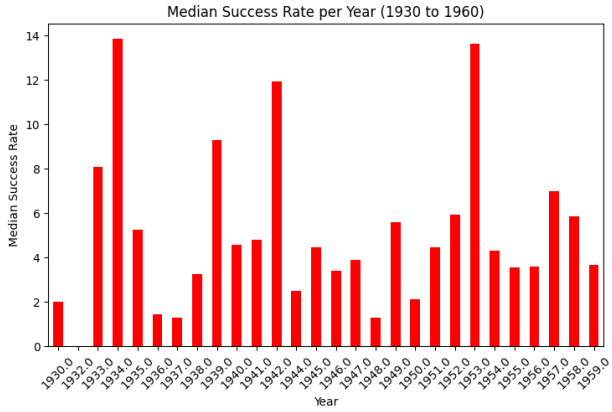


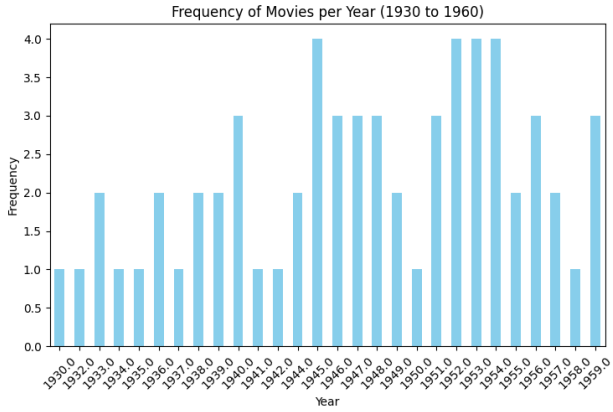

Median Success Rate per Year (1990 to inf)
Median Success Rate
Year

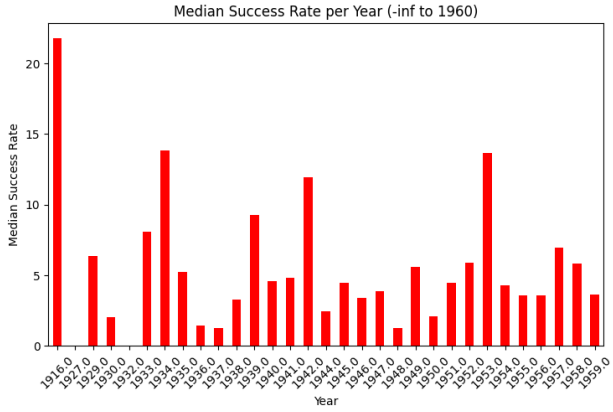

Frequency of Movies per Year (1990 to inf)
Frequency
Year

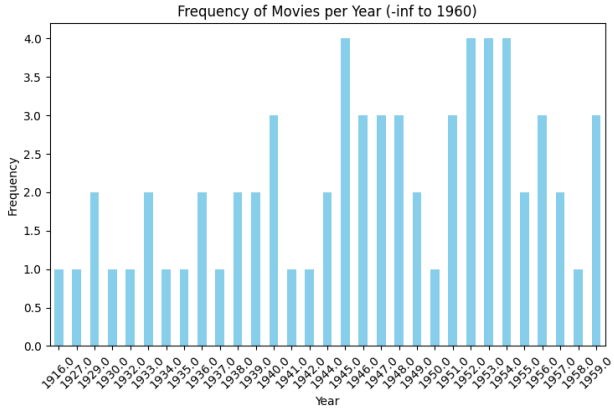

Median Success Rate per Year (1960 to 1990)
Median Success Rate
Year


Frequency of Movies per Year (1960 to 1990)
Frequency
Year

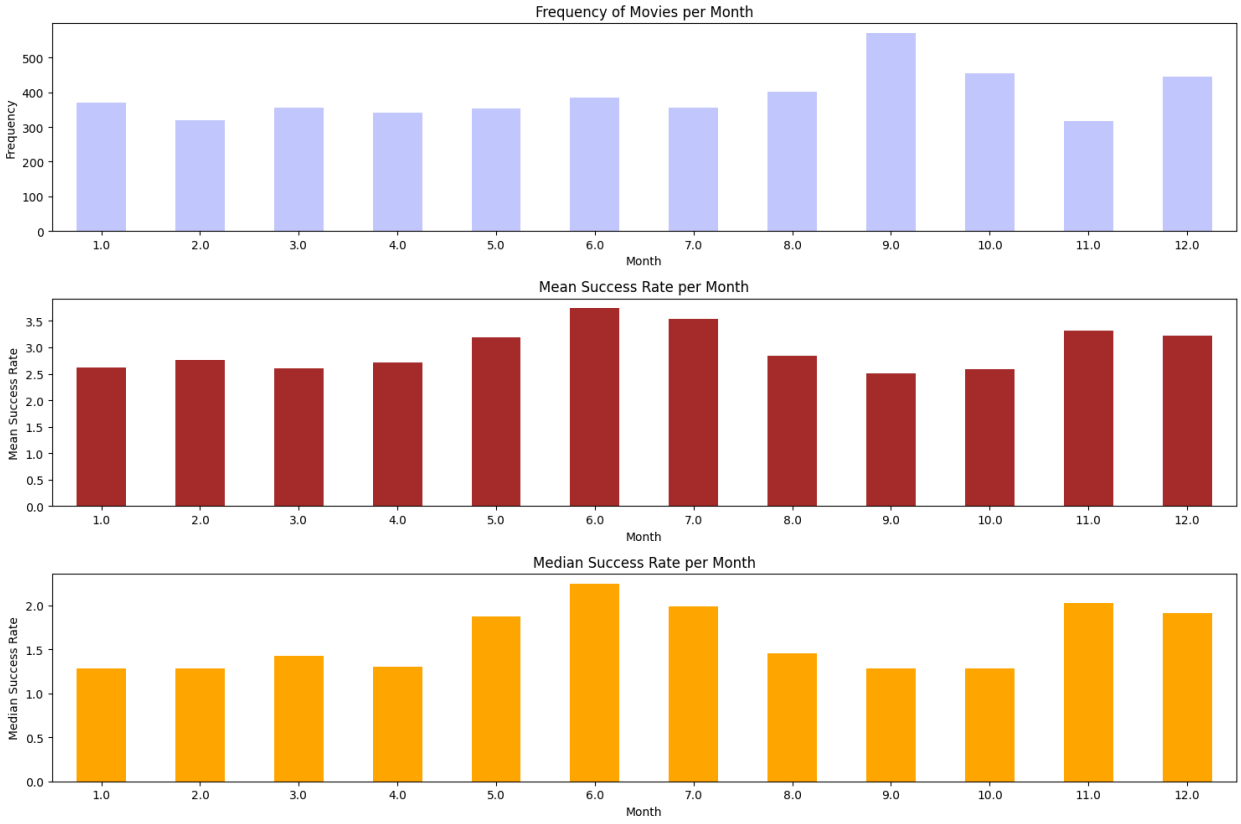

Median Success Rate per Year (1930 to 1960)
Median Success Rate
Year


Frequency of Movies per Year (1930 to 1960)
Frequency
Year

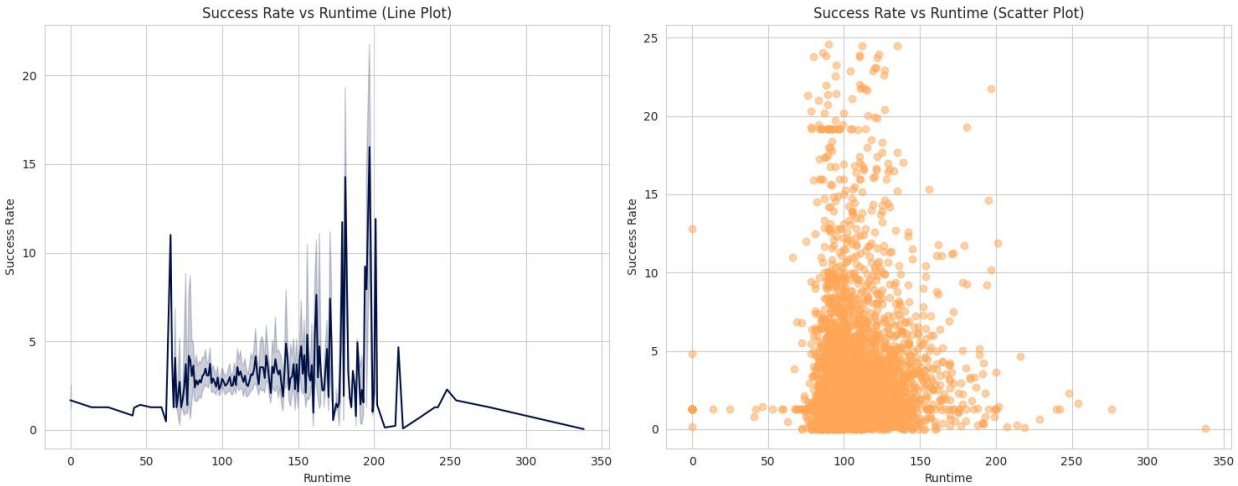

Median Success Rate per Year (-inf to 1960)
Median Success Rate
Year


Frequency of Movies per Year (-inf to 1960)
Frequency
Year

As we can see in the plots above, we have more successful movies from 1916 to 1960. However, we should consider that we have fewer movies in this period compared to 1960 to 2022. In the latter period, we have more movies, but the success rate of them decreases.



I plotted the median success rate and movie frequency per month. As we can see, most movies are released in September, October, and December. However, the most successful movies are released in June, July, November, and December, indicating that people tend to visit movie theaters more in the middle and end of the year.



In the plot above, we have success rate vs. runtime. As we can see, there is a common runtime range for movies between 75 days and 200 days. Additionally, we observe that movies with lower success rates typically have runtimes of more than 200 days.

EDA was performed to gain insights into the distribution and relationships between different features. Plots were generated to visualize the success rate distribution across different variables such as budget, runtime, release year, and month. Key observations from EDA include:

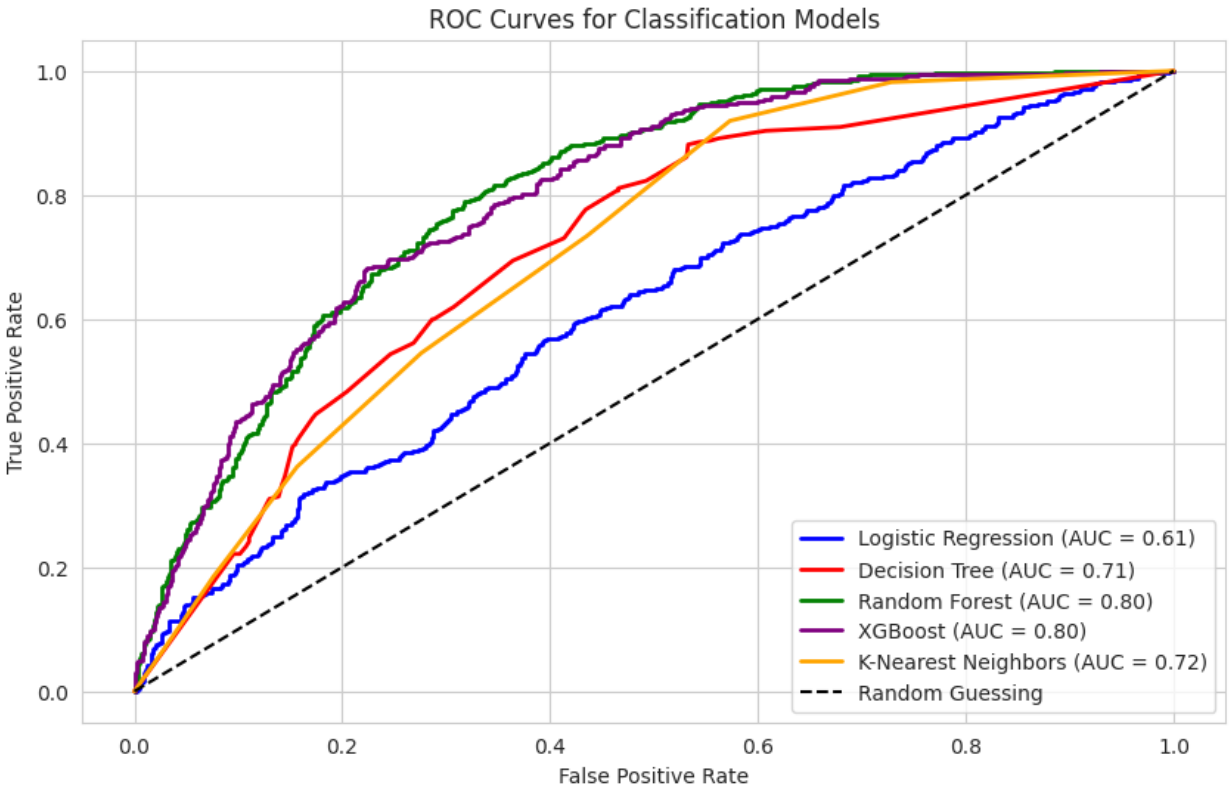
- A significant number of successful movies were observed between 1916 and 1960, indicating a period of increased interest in cinema.
- Low-budget movies showed a higher success rate compared to high-budget ones, suggesting that budget alone does not guarantee success.
- Successful movies were more likely to be released during certain months, with peaks observed in June, July, November, and December.

Model Training and Evaluation:

Five classification models were trained and evaluated using the dataset:

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. XGBoost Classifier
5. K-Neighbors Classifier

Each model was evaluated using metrics such as accuracy, precision, recall, and F1-score. Based on the evaluation results, it was observed that Random Forest and XGBoost classifiers outperformed the other models in terms of AUC, precision, recall, and F1-score. These models demonstrated better predictive performance for classifying successful and unsuccessful movies.



Conclusion:

In conclusion, this notebook presents a comprehensive analysis of classification models for predicting movie success. Through thorough data preprocessing, exploratory data analysis, and model training, valuable insights were gained into the factors influencing movie success. The evaluation of different classification models highlighted the effectiveness of Random Forest and XGBoost classifiers for this task. Future work could involve further feature engineering, hyperparameter tuning, and ensemble methods to improve model performance and enhance predictive accuracy.