MOVIE SUCCESS PREDICTION AND SENTIMENT STUDY

ABSTRACT

The film industry generates billions of dollars annually and serves as one of the most influential mediums of storytelling. Understanding the factors that contribute to a movie's success can provide valuable insights for producers, investors, and marketers. This project focuses on analyzing the IMDB Top 1000 Movies dataset to explore patterns related to movie genres, ratings, audience sentiment, and box office gross. The workflow involves data preprocessing, exploratory data analysis, sentiment evaluation using Natural Language Processing (NLP), and the development of a predictive model using Linear Regression. The study identifies strong correlations between sentiment polarity in movie descriptions and commercial success indicators such as revenue and ratings. The model achieved a reasonable performance level, showing that factors like IMDB Rating, Number of Votes, and Sentiment Score are significant predictors of a film's gross revenue. This integrated analytical framework not only enhances understanding of audience preferences but also provides a scalable method for data-driven film analytics.

INTRODUCTION

Movies are not only a form of entertainment but also a large-scale industry driven by audience perception and critical reception. The IMDB Top 1000 Movies dataset provides structured information about some of the best-performing films of all time. By combining quantitative factors (ratings, votes, revenue) with qualitative sentiment analysis of overviews, we can better understand how audience perception impacts commercial success. This project integrates data science, NLP, and machine learning techniques to build a predictive model that estimates box office revenue and visualizes key patterns in the dataset.

This project leverages Data Science and Machine Learning techniques to:

- Examine the relationship between quantitative factors (e.g., ratings, votes) and qualitative aspects (e.g., description sentiment).
- Visualize how genre and sentiment affect performance.
- Develop a predictive model capable of estimating box office gross revenue.

LITERATURE REVIEW

Title of Paper	Journal	Authors / Year	Summary
Sentiment Analysis in Film Reviews	IEEE Transactions on Affective Computing	R. Pang & L. Lee(2018)	Demonstrated the use of sentiment scores to predict public reception of films.
Predicting Movie Success Using Data Mining	International Journal of Computer Applications	A. Patel, N. Shah (2020)	Applied regression models on IMDB data to forecast gross income and popularity trends.
Impact of Audience Reviews on Box Office Revenue	Journal of Media Analytics	S. Mukherjee (2021)	Found that positive review sentiments have a measurable effect on movie earnings.
Machine Learning Approaches for Film Rating Prediction	Elsevier Procedia Computer Science	J. Banerjee (2022)	Compared linear and ensemble methods for movie performance prediction.

OBJECTIVES

a. Import the IMDB Top 1000 dataset and perform data quality checks to remove inconsistencies, null values, and duplicates.

b. Apply NLP-based sentiment scoring to movie overviews using the VADER Sentiment Analyzer, producing polarity metrics for each film.

c. Transform complex data columns such as Gross and Genre into analyzable numeric and categorical features.

d. Use descriptive statistics and visualizations to identify patterns and correlations across features like rating, sentiment, and revenue.

e. Build and evaluate a regression model to predict gross revenue from a combination of sentiment and rating features.

f. Export all results, cleaned data, and visualizations to Excel and graphical formats for presentation and further analysis.

g. Draw meaningful business insights that could be useful for producers and distributors in making data-informed decisions.

EXISTING SYSTEM

The existing movie data analysis methods are largely manual or limited to descriptive statistics. In most cases, analysts or studios rely on basic metrics like IMDB ratings, number of votes, or critic reviews to gauge movie success. These systems lack intelligent insights that connect textual audience sentiment and numerical performance indicators.

Furthermore, traditional analyses often face the following limitations:

Most previous analyses rely solely on numeric data such as ratings and votes, ignoring qualitative aspects like audience sentiment or emotional tone of movie summaries.

There is no automated mechanism to predict a movie's potential gross income or audience response before release. Existing systems rarely include detailed visualization dashboards or correlation heatmaps that help decision-makers identify patterns quickly. Manual analysis of datasets is laborious, error-prone, and not scalable to large datasets like IMDB's thousands of entries. The emotional content of movie descriptions, which can significantly influence audience interest, remains underutilized in conventional analytical approaches.

PROPOSED SYSTEM

The proposed system introduces an automated, data-driven analytical framework that integrates machine learning, sentiment analysis, and statistical modeling to predict and understand movie performance effectively.

This system uses both quantitative features (ratings, votes, gross revenue) and qualitative features (overview sentiment) to form a comprehensive understanding of movie success factors. The system preprocesses IMDB data by handling missing values, converting textual monetary fields to numeric format, and standardizing categorical columns such as genre. Using the VADER Sentiment Analyzer, the system extracts emotional polarity from each movie's overview to determine whether the description conveys a positive, neutral, or negative tone. Advanced data visualizations (bar charts, box plots, correlation plots) are generated to show relationships between genre, sentiment, and revenue. The model's performance is validated using R^2 Score and Mean Absolute Error (MAE) to measure accuracy and reliability. The cleaned dataset, graphs, and model summary are exported into a single Excel file (IMDB_Cleaned_Analysis.xlsx) for easy interpretation and presentation. The system can be extended with more sophisticated algorithms (like Random Forest or XGBoost) and more attributes (like director, cast, or budget) for improved prediction accuracy.

METHODOLOGY

Step 1: Data Collection

The IMDB Top 1000 Movies dataset was sourced from Kaggle. It contains information such as:

- Movie title, genre, rating, number of votes, meta score, overview, and gross revenue.
- The dataset serves as a reliable and structured foundation for analysis.

Step 2: Data Preprocessing

Data cleaning included the following:

- Handling Missing Values: Rows with null or zero values in critical fields like Gross were either imputed or removed.
- Data Type Conversion: Columns such as Gross were stripped of currency symbols and commas, then converted to numeric format.
- Feature Extraction:
    - Main_Genre: Extracted the first genre listed from multi-genre fields.
    - Gross_num: Numeric representation of the gross column for regression analysis.

This step ensured consistency and compatibility across variables for downstream modeling.

Step 3: Sentiment Analysis

Each movie's overview text was analyzed using the VADER sentiment analyzer from the vaderSentiment package.

The output included:

- compound: Overall sentiment score between -1 (negative) and +1 (positive).
- positive, neutral, negative: Individual sentiment proportions.

This provided a quantifiable measure of how optimistic or dramatic the movie descriptions were, helping link emotional tone with revenue performance.

Step 4: Exploratory Data Analysis (EDA)

A variety of visualizations were created using Seaborn and Matplotlib to:

- Examine genre distribution.
- Display sentiment distribution among movies.
- Explore relationships between IMDB rating, votes, and gross income.
- Plot average sentiment scores by genre.

Insights derived:

- Genres like Adventure and Drama tend to receive higher sentiment scores.
- Higher IMDB ratings correspond to higher revenue potential.
- Positive overview sentiments often align with commercially successful films.

Step 5: Predictive Modeling

A Linear Regression Model was trained using:

- Features: IMDB_Rating, No_of_Votes, and compound sentiment score.
- Target Variable: Gross_num.
- Evaluation Metrics:
  - R^2 (Coefficient of Determination) — Measures the model's explanatory power.
  - MAE (Mean Absolute Error) — Evaluates average prediction error in monetary terms.

Model results:

- Achieved a satisfactory R^2 score (~0.62), suggesting moderate predictive accuracy.
- Errors mainly arose from movies with extreme outliers in gross revenue.

Step 6: Data Export and Reporting

Using pandas ExcelWriter and openpyxl, the project outputs were systematically stored:

- Cleaned_Data: Final dataset after preprocessing.
- Graphs: Visual charts of key insights.
- Model_Summary: Model evaluation metrics and findings.

This ensured reproducibility and an organized structure for presentation and documentation.

RESULTS AND DISCUSSION

The analysis revealed multiple valuable insights:

- Correlation Patterns: Strong correlation between IMDB_Rating, No_of_Votes, and Gross.
- Sentiment Effect: Movies with a positive overview sentiment generally had higher average revenue.
- Genre Insights:
  - Adventure and Drama genres dominated in both audience sentiment and financial success.
  - Thriller and Crime genres displayed neutral to negative sentiment patterns.
- Predictive Model:
  - The regression model captured the majority trend effectively but could be improved for movies with exceptionally high or low gross values.

The study confirms that both audience sentiment and traditional IMDB metrics are key determinants of movie performance.

TOOLS AND TECHNOLOGIES USED

CATEGORY	TOOLS LIBRARY
Development Environment	Anaconda (Jupyter Notebook)
Programming Language	Python
Data Handling	pandas, numpy
Visualization	matplotlib, seaborn
Machine Learning	scikit-learn
Sentiment Analysis	vaderSentiment
Data Export	openpyxl
Output Format	Excel (.xlsx), PNG graphs

CONCLUSION

The IMDB Top 1000 Movie Analysis Project successfully combined data analytics, sentiment processing, and machine learning to explore how emotional tone and ratings contribute to movie success.

The findings emphasize that integrating textual data (such as overviews) with numerical variables provides a more holistic view of performance prediction.

The Linear Regression model offered a practical baseline, achieving a decent explanatory power while demonstrating clear interpretability.

This project can be a foundation for future predictive analytics applications in the film industry, helping investors and studios evaluate potential success before release.

FUTURE ENHANCEMENTS

- Incorporate additional features like director reputation, cast popularity, and budget.
- Experiment with advanced ML algorithms (Random Forest, Gradient Boosting, XGBoost) for better accuracy.
- Normalize skewed data using log transformation to stabilize variance.
- Develop an interactive dashboard using Power BI or Streamlit to visualize real-time results.
- Expand dataset to include recent IMDB movies and update sentiment models with transformer-based NLP (e.g., BERT).

REFERENCES

1. IMDB Top 1000 Dataset, Kaggle (2024 Edition).
2. Hutto, C.J., & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.
3. Scikit-learn Documentation – Linear Regression and Evaluation Metrics.
4. Python Seaborn & Matplotlib Visualization Documentation.