
GITAM
DEEMED TO BE UNIVERSITY

A Project report on

MOVIE SUCCESS PREDICTION AND SENTIMENT STUDY

Submitted in partial fulfillment of the requirements for the internship undertaken after the Fourth Semester Bachelor of Science in Mathematics, Computer Science , Statistics (B. Sc MSCS)

Submitted by

Sushmika Golla

2023003692

B. Sc MSCS

GITAM University

Internship Mentor

Dr Mallikarjuna Reddy Doodipala

Associate Professor

GITAM University

Date of submission: 19 May 2025

Signature of the student :

MOVIE SUCCESS PREDICTION AND SENTIMENT STUDY

INTRODUCTION:

The film industry is both creative and commercial. The movie success depends on the different factors , audience response being one of it. This project explores how data can be used to predict the box office's performance and study public sentiment. The goal is to build a regression model and see how different genres effect the success of the movie.

ABSTRACT

This project analyses movie data using sentiment analysis and machine learning. Using the TMDB 5000 Movie Dataset from Kaggle, we extracted features like budget, vote average, and textual sentiment. Sentiment scores were derived from each movie's overview using VADER (from NLTK). A linear regression model was then used to predict movie revenue. We also visualized sentiment trends across different genres. The model achieved an R^2 score of ~0.50, indicating moderate predictive power.

TOOLS USED

Python: Main programming language

Pandas: Data handling and preprocessing

NLTK (VADER): Sentiment analysis of movie overviews

Scikit-learn: Regression modeling and evaluation

Matplotlib/Seaborn: Data visualization

Google Colab: Coding and experimentation platform

STEPS INVOLVED IN BUILDING THE PROJECT:

Data Import and Merging -

Used TMDB's movies and credits CSV files

Merged them on the title column

Data Preprocessing -

Selected important columns: title, overview, genres, budget, revenue, vote average

Cleaned missing and zero-value rows for better accuracy

Sentiment Analysis -

Applied VADER to each movie overview

Added sentiment score column to dataset

Score ranged from -1 (negative) to +1 (positive)

Genre-wise Sentiment Trends -

Extracted primary genre from each movie

Plotted average sentiment per genre using bar charts

Regression Modeling -

Used budget, vote average, and sentiment score as features

Predicted revenue using Linear Regression

Evaluated model using Mean Squared Error and R^2 Score

CONCLUSION:

This beginner-friendly data science project successfully used textual sentiment and numerical features to predict movie revenue. Although simple, the linear regression model explained about 50% of the variation in revenue. The sentiment analysis also showed clear differences in emotional tone across movie genres. This project demonstrates how data, text, and machine learning can be combined to gain real-world insights.

****

The complete code and analysis for this project can be accessed through the Google Colab notebook here ;

https://colab.research.google.com/drive/1sB_BE-mMrZoqwEC29nSq47wAyYHRCSRh?usp=sharing