## Movie Success Prediction Report

Introduction

The purpose of this project was to predict whether a movie will be successful or a flop based on features such as budget revenue or other characteristics related to its production. The goal was to explore how specific variables influence box office success and to create a model that can classify future movies based on these features.

Process

I chose 100 movies from 2023 and 10 movies from 2024 and used Google Sheets to prepare the dataset. I collected all movie movie related data from Box Office Mojo, The Numbers IMDB and Wikipedia. The columns I created for each movie were Title, Month Released, Year, IMDb rating has star?, Budget, Revenue, Profit (calculated as revenue minus budget) and success.

Text values e.g. month released, has star?, and success are manually encoded into numbers. For example, Months were assigned numeric codes like January = 1, February equals = 2 and so on. The presence of a famous star was encoded as 1 = yes or 0 = no. I also created a success label based on profit if the movie made 50 million and more in profit it was labeled as yes otherwise it was labeled as no, this yes no values were also converted into numbers with yes = 1 and no = 0.

Although I did not scale the data for models like decision trees and logistic regression since these models are not sensitive to different data magnitudes, I did apply scaling for the K nearest neighbors (KNN) model because it relies on this calculation and which was distorted by unscaled features. I split the data into two subsets subsets movies from 2023 were used as the training set and movies from 2024 were used as the testing set. I also created a new column named dataset to indicate whether each movie belonged to the training or test group.

Google colab

All data analysis and model training were completed in Google colab. I uploaded the CSV file exported from Google sheet directly into Google colab. There were no missing values in the data set, all the columns were checked to ensure proper formatting and correct data types after the upload. I used the pandas library to upload and explore the data, confirming that the structure was consistent and clean.

Method and model training

I used three different machine learning models to test using SCIKIT learn in Google collab;

1. K Nearest Neighbors

KNN classifies a new movie by comparing it to the similar ones in the training data using this between features since it relies on distance. I scaled the numerical features to make sure no single feature like budget dominated the results. I used the default value of K=5 which means the prediction is based on the 5th closest movies in the data set.

Strength

- Easy to understand

- Good for small datasets

Weaknesses

- Requires scaling

- Slow on large data set

- Sensitive to noise or irrelevant features

**Accuracy** - 60%

1. Decision tree

It splits the data based on feature values to create a flow decisions leading to a prediction. It unidentified the most useful features and decision points to classify each movie as a success or a flop.

Strength

- No need for feature scaling

- Can capture complex rules to separate the data

- Handles both numerical and categorical data well

Weaknesses

- Prone to overfitting if not probably controlled

-Small changes in data can create very different trees

-Can be biased towards features with more levels

**Accuracy** - 60%

1. logistic regression

Analyzes the relationship between features like IMDb rating, budget, genre code and the presence of a star. And predicted success based on calculated probability. It outputted either1 (successful)l or 0 (flop) for each movie in the test set.

Strength

- Simple and fast to train outputs

-Gives a clear probability for each prediction works well if the connection between features and result is mostly straight line

weaknesses

- Assumes the relationship between features and result is simple and straight

- Doesn't handle complicated patterns without extra work

- Can get confused if some features are very similar to each other

**Accuracy** - 50%

Kpi and evaluation

The key performance indicator KPI was used for accuracy, it measures the ratio of correct predictions to total predictions. Accuracy was selected due to its simplicity and appropriateness for the small balanced data sets. It was calculated using SK learn’s Accuracy score function

Dimensionality reduction

No dimensionality reduction methods were used Because their data sets were already concise and manageable.

Conclusion

This project provided practical experience in building a machine learning pipeline from data preprocessing and feature encoding to training and evaluation. It highlighted the importance of scaling, model selection and feature quality. while model accuracy was limited by data set size and simplicity, It still demonstrated the potential of basic classification models in predicting movies success.