İsmail Sarp Erdem 32105

DSA210 Spring25’ Term Project

Movie Revenue Analysis & Prediction Final Report

Summary

This report presents an end-to-end analysis of factors driving box-office revenue using TMDB data and predictive modeling techniques. We explore key relationships via exploratory data analysis, data visualization, testing hypotheses statistically, and build regression models (Linear, Decision Tree, Random Forest) to forecast revenue.

1. Motivation and Purpose

Box-office revenue is a critical metric for studios, distributors, and marketers. Understanding which factors most influence a film’s financial success enables better budgeting, marketing strategies, and resource allocation. This project:

- Quantifies relationships between production/popularity features and revenue.
- Demonstrates a full data science workflow: cleaning, analysis, statistical testing, and predictive modeling.
- Provides recommendations for stakeholders seeking to optimize box-office returns.

2. Data Source and Processing

Used datasets are : `tmdb_5000_movies.csv` (movie metadata) and `tmdb_5000_credits.csv` (cast/crew info).
Origin: Publicly available TMDB 5000 dataset on Kaggle (sourced via TMDB API).
Processing: Merged on the `id` column, parsed JSON-encoded fields (e.g., genres), and created a `genre_count` feature.

3. Data Analysis

3.1 Data Loading & Cleaning

Merged two CSV files into a single DataFrame with ~4895 entries and 20+ columns.
Dropped rows with missing critical values in features used for modeling (budget, revenue).
Parsed and counted genres per movie to create a numeric `genre_count` field.

3.2 Exploratory Data Analysis and Visualization

- Missing Values & Duplicates: Identified sparsity in `runtime` and `vote_average`, but full coverage for budget and revenue.
- Distributions: Plotted histograms with KDE overlays, box plots to reveal heavy right skew in budget/revenue and outliers in vote counts, correlation graphs and heat-maps for relationships, bar charts for the categorical features.
- Correlations: Computed Pearson's r for all pairs; resulting in the conclusion that budget↔revenue correlation was strongest. ($r \approx 0.78$, $p \ll 0.001$).
- Hypothesis Testing: Rejected H_0 of no correlation (budget vs. revenue) and difference in means between high- and low-rated films (t-test $p \ll 0.001$).

4. Predictive Modeling (Machine Learning Application)

4.1 Feature Selection & Preparation

Predictors: `budget`, `popularity`, `vote_count`, `vote_average`, `runtime`, `genre_count`.
Target: `revenue`.
Train/test split (80/20), followed by standard scaling of predictors.

4.2 Models & Evaluation

Model	R^2	RMSE (≈)
Linear Regression	0.76	$79 M
Random Forest	0.73	$84 M
Decision Tree (d=5)	0.41	$124 M

5. Findings

- Top Predictors: Budget and popularity together explain ~76 % of revenue variance.
- Model Performance: Simple linear models are highly competitive; random forest and decision trees add resilience but require tuning.
- Feature Impact: `vote_count` contributed meaningfully; `vote_average`, `runtime`, and `genre_count` had minimal standalone impact on revenue.

6. Limitations & Future Work

Limitations

- Scope: Only numeric TMDB features used—omitting textual or external factors like marketing spend, reviews sentiment.
- Validation: Single train/test split without k-fold cross-validation may overestimate performance stability.
- Tuning: Hyperparameters (tree depth, number of estimators) were not optimized.

Future Work

- Feature Expansion: Incorporate one-hot encoded genres, release date features, and cast/crew popularity metrics.
- Model Tuning & CV: Apply GridSearchCV or RandomizedSearchCV with k-fold cross-validation.
- Advanced Algorithms: Experiment with boosting (XGBoost, LightGBM) and model stacking.