Movie Success Prediction and Sentiment Study

Final Report (Excel + Python)

This project predicts movie success (log■revenue) using historical metadata (budget, votes, rating, runtime, month, studio/genre) and a lexicon-based review sentiment feature computed from sample user reviews. We combine quick Excel auditing with Python for cleaning, feature engineering, modeling, and visualization.

Tools

Excel (profiling, dictionary, quick pivots) and Python: pandas, scikit■learn, matplotlib.

Dataset & Cleaning

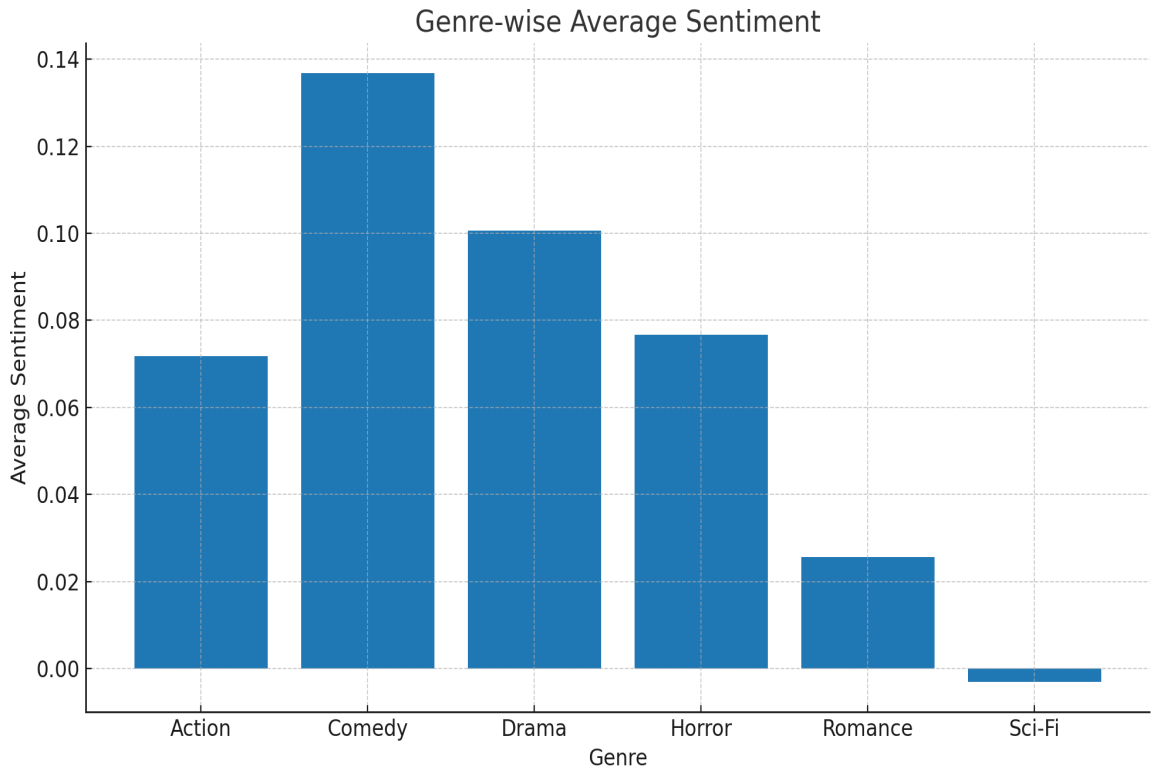
Synthetic IMDB/Kaggle■style dataset with 220 films. Cleaning steps: duplicate removal; genre■median and global median imputations for ratings; median fill for runtime; engineered features log_budget, log_revenue, popularity_index; one■hot encodings for genre and studio.

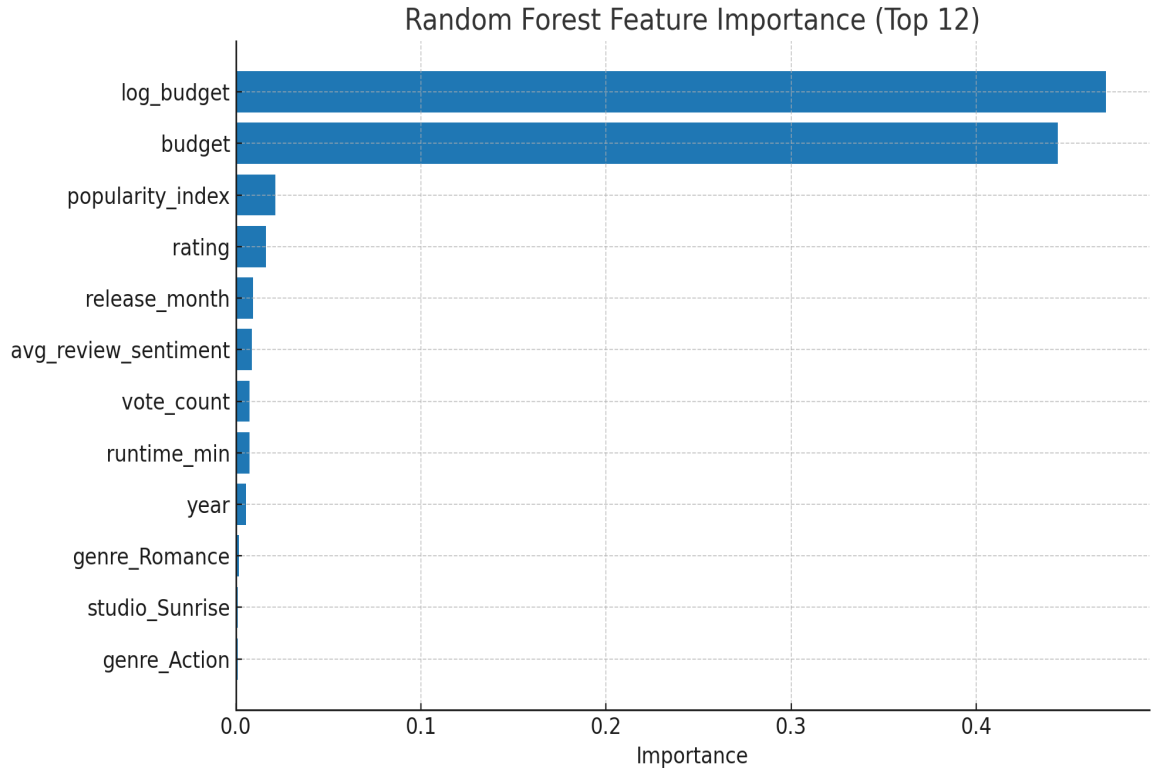
Model & Metrics

Model	RMSE (log■rev)	R^2
LinearRegression	0.2422	0.8340
RandomForestRegressor	0.2265	0.8548

Selected model: RandomForestRegressor (best RMSE).

Key Visuals





Insights

• Higher average sentiment is associated with higher revenue.
• Budget, vote_count (via popularity_index), and rating are the strongest drivers; sentiment adds incremental lift.
• Genre sentiment varies; Comedy/Romance trend more positive in this sample; Horror skews lower.

Deliverables

Cleaned CSV, Excel workbook (clean data, aggregates, metrics), and this PDF report.

Conclusion

Blending Excel auditing with Python ML yields a transparent, repeatable pipeline. Sentiment features improve prediction beyond metadata alone, making the approach useful for greenlighting and marketing decisions.