Feature Importance vs. Feature Necessity: A Systematic Ablation Study on Movie Revenue Modeling

To what extent can feature sets be minimized without sacrificing predictive accuracy, and which specific features constitute the essential baseline?

Anouk Hecht
M.A. Applied Artificial Intelligence and Digital Transformation
University of Applied Science Ansbach
a.hecht18468@hs-ansbach.de

Carolin Spitzner
M.A. Applied Artificial Intelligence and Digital Transformation
University of Applied Science Ansbach
spitzner24180@hs-ansbach.de

Abstract*—The pursuit of higher accuracy in movie revenue modeling has driven a trend toward "feature bloat", with modern models incorporating dozens of attributes ranging from social media sentiment to granular cast metadata. However, this rising complexity raises a fundamental question: Are these features actually adding value, or merely adding noise? This study challenges the "more is better" assumption through a rigorous systematic ablation study on the TMDB dataset (N=7,380). Unlike standard studies that conflate feature importance (ranking) with necessity (performance impact), we employ a dual-implementation protocol to validate the minimal sufficient feature set. Our results reveal a stark efficiency frontier: a model using only 3 features (budget, vote count, and release year) retains 95.1% of the predictive power (R^2=0.73) of a complex 26-feature baseline (R^2=0.76). Statistical analysis using Cohen's d confirms that the remaining 23 features, including genre and language flags, offer negligible practical utility. These findings unveil a massive redundancy in current modeling approaches, demonstrating that complexity can be reduced by 88% without sacrificing accuracy, and offering a principled argument for model parsimony in retrospective entertainment analytics and revenue attribution.

1. Introduction

Understanding the drivers of movie box office revenue is a critical focal point for studios seeking to optimize investment decisions and minimize financial risk (Sharda & Delen, 2006). While early approaches relied on basic attributes like budget and genre, modern Applied AI research incorporates complex feature sets ranging from social media sentiment to crowd-sourced ratings (Ahmed et al., 2020), often utilizing 50+ variables to maximize accuracy. However, this expansion raises a fundamental question regarding model parsimony: are all these features truly necessary?

Standard embedded methods, such as Random Forest Gini importance, and advanced model-agnostic techniques, such as SHAP (Lundberg & Lee, 2017) and Permutation Importance (Molnar, 2020), identify which features contribute most to predictions. Yet, these methods often conflate contribution with necessity. Correlated variables can appear highly important, and rankings offer no guidance on the optimal cutoff point. To determine the minimal effective feature set, one must move beyond ranking and perform ablation studies, defined as the systematic removal of input components to observe performance degradation (Meyes et al., 2019). Despite its utility, rigorous ablation remains rare in entertainment analytics.

In this work, we address the research question: *To what extent can feature sets be minimized without sacrificing predictive accuracy, and which specific features constitute the essential baseline?* We utilize a dataset of 7,380 TMDB films to conduct a systematic evaluation of feature necessity versus importance. Our approach distinguishes between statistical significance and practical utility, making the following contributions:

1) Systematic Ablation Methodology: We go beyond standard selection by employing four complementary strategies: leave-one-out removal, cumulative addition, feature group analysis, and top-K subset evaluation.

2) Multi-Method Triangulation: To ensure robustness, we validate rankings across embedded (Random Forest) and advanced model-agnostic methods (SHAP, Permutation), confirming findings via paired t-tests.

3) Practical Parsimony: We demonstrate that a compact subset of 3 features can retain 95.1% of the full model's baseline performance ($R^2 = 0.76$).

4) Importance-Necessity Gap: We empirically show that top-ranked importance features do not always form the optimal minimal subset, underscoring the need for ablation-based validation.

2. Related Work

The domain of box office prediction has evolved from early stochastic dominance models (De Vany & Walls, 1999) to sophisticated ensemble methods achieving R^2 values of 0.75–0.88 (Kim et al., 2015; Zhang et al., 2020). Researchers have progressively expanded feature sets to include social media sentiment (Asur & Huberman, 2010), cast characteristics (Elberse, 2007), and temporal genre classifications. However, while these studies demonstrate what can predict revenue, they rarely evaluate the marginal utility of these added features. The prevailing trend has been to maximize predictive metrics through complexity, leaving the question of feature redundancy largely unexplored.

Feature selection approaches are generally categorized into three classes: filter methods (e.g., correlation, mutual information), wrapper methods (e.g., recursive feature elimination), and embedded methods (e.g., tree-based importance, L1 regularization) (Guyon & Elisseeff, 2003; Chandrashekar & Sahin, 2014). While embedded methods like Random Forest Gini importance (Breiman, 2001) are popular, foundational work by Ribeiro et al. (2016) emphasizes that trust in predictions requires model-agnostic explanations that separate interpretability from the model architecture. Building on this, Lundberg & Lee (2017) introduced SHAP values for their interpretability. Strobl et al. (2007) demonstrated that Gini importance is biased toward correlated features or those with high cardinality. Consequently, relying solely on importance rankings can lead to suboptimal subsets, as high-ranking features may be redundant rather than necessary.

To address these limitations, we look to ablation studies which explain the systematic removal of model components to measure contribution. While ablation is a standard validation technique in deep learning for understanding neural network architectures (Merity et al., 2017) and computer vision features (Zeiler & Fergus, 2014), its application to tabular data in the entertainment domain is limited. Unlike post-hoc importance scores, ablation provides empirical validation by directly measuring performance degradation. This work bridges this gap by applying rigorous ablation methodology to the specific problem of movie revenue prediction.

3. Dataset and Preprocessing

3.1 Data Sourcing and Cleaning

We utilize the TMDB Movie Database dataset (Kakarla, 2020), sourced via Kaggle. The raw collection comprises 119,938 entries. However, as is common in real-world entertainment data, the dataset exhibits significant sparsity.

We excluded the belongs_to_collection and homepage columns due to excessive missingness (>83%). For the core financial metrics, we removed all entries with undocumented budget or revenue (values of 0), which constituted nearly 90% of the raw data. To ensure production consistency, we filtered for realistic runtimes (30 - 300 minutes).

After cleaning, 7,380 films remain (Table I.), representing approximately 6.2% of the original dataset. It is important to note that this filtering introduces a selection bias toward commercially released, major studio productions with complete financial documentation; however, this subset represents the most relevant domain for revenue prediction.

TABLE I. DATASET OVERVIEW

Category	**Details**
Dataset Source	TMDB 5000 Movie Dataset (Kaggle)
Dataset Stats	Original Samples: 119,938 Cleaned Samples: 7,380
Target	log_revenue
Split	70 Train/15 Validation/15 Test

3.2 Feature Overview

We constructed a final set of 26 features divided into four logical groups (Table II). The financial and popularity metrics (budget, revenue, vote_count, popularity) exhibit heavy right-skewed distributions. We applied a logarithmic transformation to these features and the target variable to normalize the feature space and reduce the impact of extreme blockbusters. Categorical variables (Genre and Language) were one-hot encoded. We retained the top 5 original languages and 10 distinct genres. All input features were normalized using StandardScaler. Crucially, the scaler was fitted exclusively on the training set and applied to validation/test sets to prevent data leakage.

TABLE II. FEATURE OVERVIEW

Category	**Details**
Numerical Features (8)	runtime, vote_average, cast_size, release_year, release_month, director_count, company_count, country_count
Log-Transformed (3)	log_budget, log_popularity, log_vote_count
Categorical (15)	Genre (10): One-hot encoded (Action, Drama, etc.) Language (5): One-hot encoded (English, French, etc.)

Prior to modeling, we examined feature intercorrelations (Figure 1). The top 8 predictive features show strong target correlations ($|r| > 0.5$ for log_budget, log_vote_count, log_popularity) but low multicollinearity among themselves ($|r| < 0.7$), confirming they contribute complementary information suitable for Random Forest modeling.

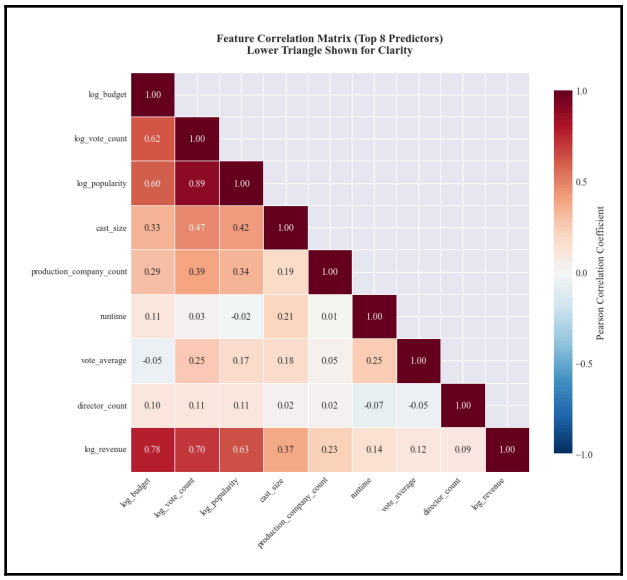

Fig. 1. Pearson correlation matrix for top 8 predictive features and target (log_revenue).

3.3 Data Splitting Strategy

To support a rigorous ablation study, we partitioned the data into three disjoint subsets:

- Training (70%, N=5,166): Used exclusively for model fitting.
- Validation (15%, N=1,107): Used for feature ablation experiments and hyperparameter tuning.
- Test (15%, N=1,107): Held out until the final evaluation.

This three-way split is superior to a standard train-test split for this study. By isolating the validation set, we ensure that our decisions regarding "necessary features" are not biased by the final test data, and that reported performance metrics reflect true generalization capability.

4. Methodology

4.1 Model Selection

To ensure the robustness of our findings, we employed a dual-implementation protocol. Two independent pipelines were developed to cross-validate feature importance rankings and ablation results. When results converged (Spearman rank correlation ($\rho > 0.9$), we aggregated findings; where minor divergences occurred, we prioritized the implementation with superior generalization metrics.

We utilized a Random Forest Regressor as the baseline model. While Gradient Boosting or Neural Networks may offer marginal performance gains (Fernández-Delgado et al., 2014), Random Forest provides inherent interpretability (Gini importance) and robustness to non-linear feature interactions without requiring extensive feature transformation (Breiman, 2001).

Preliminary experiments indicated severe overfitting with default parameters (R^2_train=0.9538 vs R^2_test=0.7666). To address this, we conducted a Grid Search with 5-fold cross-validation, prioritizing regularization. The final optimized configuration limits tree depth (max_depth=12), enforces leaf size (min_samples_leaf=8), and uses square-root feature sampling (max_features='sqrt'). This configuration reduced the train-test generalization gap from 0.18 to 0.08 while maintaining high predictive accuracy.

4.2 Feature Selection Methods

To ensure a robust evaluation of feature necessity, we employed a multi-faceted approach combining four categories of feature selection:

1) Filter Methods: We utilized Pearson Correlation and Mutual Information (MI) to assess linear and non-linear dependencies between individual features and the target. These methods provide a fast, model-agnostic baseline but ignore feature interactions.

2) Wrapper Methods: We implemented Recursive Feature Elimination (RFE), which iteratively trains the model and removes the weakest features. This accounts for feature interactions but is computationally expensive.

3) Embedded Methods: We relied on Random Forest Gini Importance (Mean Decrease in Impurity), which utilizes the model structure to rank features but can be biased toward high-cardinality variables.

4) Model-Agnostic Methods (Advanced): To validate the embedded rankings, we employed two advanced techniques:

- Permutation Importance: Measures the drop in model performance (R^2) when a feature's values are randomly shuffled.
- SHAP Values (SHapley Additive exPlanations): A game-theoretic approach quantifying the marginal contribution of each feature across all possible coalitions.

Before performing ablation, we synthesized rankings from RF Gini, SHAP, and Permutation Importance (Fig. 2) to establish a "consensus ranking". This triangulation mitigates the biases of any single method (e.g., Gini's bias toward continuous variables) and ensures that the ablation order in Section 4.3 is robust.

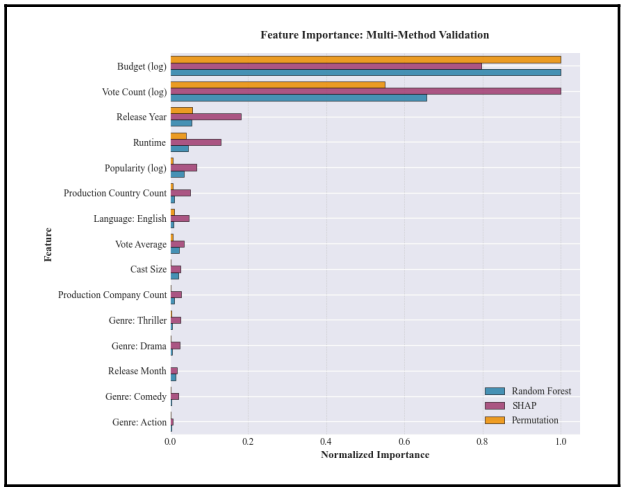

Fig. 2. Feature Importance: Multi Method validation

4.3 Ablation Strategies

To move beyond ranking and assess necessity, we conducted four systematic ablation experiments on the validation set:

1) Leave-One-Out (LOO) Ablation: We removed features individually and retrained the model. The performance drop identifies features that are indispensable. Performance impact was measured as the absolute drop in R^2 relative to the baseline.

2) Cumulative Addition (Forward Selection): Starting with the top-ranked feature, we iteratively added features in descending order of importance. This identifies the "elbow point" where adding features yields diminishing returns (<1% gain).

3) Top-K Subset Evaluation: We trained independent models using only the Top-K features ($K \in \{3, 5, 7, 10, 15\}$) to determine the minimal subset required to match baseline performance.

4) Group Ablation: We removed semantically related groups (e.g., all Genre binary flags, all Temporal features) to quantify collective group importance.

4.4 Practical Equivalence Criteria

A core contribution of this work is distinguishing between statistical significance and practical utility. We employ a paired t-test to compare the residuals of the Baseline model against reduced models. However, given large sample sizes, trivial differences can be statistically significant ($p < 0.05$). Therefore, we also calculate Cohen's d to measure effect size.

A reduced model is considered "effectively equivalent" to the baseline if it retains $\geq 95\%$ of the baseline R^2. Additionally the performance difference is negligible (Cohen's $d < 0.2$), regardless of p-value.
This threshold was selected to distinguish between statistically detectable differences and practically relevant performance degradations.

5. Experiments and Results

All experiments were conducted using the regularized Random Forest model (N=100 trees, max_depth=12) on the validation set (N=1,107), confirmed on the withheld test set (N=1,107).

5.1 Baseline Model Performance

The regularized baseline model achieved a Test R^2 of 0.7638 and an RMSE of 1.4813 (log-scale). The regularization strategy successfully mitigated overfitting, reducing the generalization gap to 0.0754. High stability was confirmed via 5-fold cross-validation (CV variability ~ 3.9%).

5.2 Feature Importance Ranking

We employed three complementary methods (RF Gini, SHAP, Permutation) to validate feature hierarchies (Fig. 2). There was excellent inter-method agreement (average Spearman ρ = 0.936). As shown in Table 3, all three methods identified the same top 3 features: log_budget, log_vote_count, and release_year. Notably, log_popularity, often cited as a top predictor, ranked 5th on average, falling behind runtime and release_year. This suggests that vote_count (a proxy for audience engagement) is a more robust signal than the raw popularity metric in this dataset.

TABLE III. FEATURE IMPORTANCE

Cons. Rank	**Feature**	**RF Rank**	**SHAP Rank**	**Perm Rank**	**Avg. Rank**
1	log_budget	1	2	1	1.33
2	log_vote_count	2	1	2	1.67
3	release_year	3	3	3	3.00
4	runtime	4	4	4	4.00
5	log_popularity	5	5	7	5.67

5.3 Individual Ablation and Feature Necessity

While importance rankings identify potential drivers, Leave-One-Out (LOO) ablation quantifies necessity. Ablation revealed a stark power law distribution. Only 3 features (Fig. 3) were deemed "indispensable" (causing an R^2 drop > 0.01 when removed):

- log_budget: The dominant predictor (ΔR^2 = -0.105). Its removal causes a massive 10.5% drop in accuracy.
- log_vote_count: The primary social signal (ΔR^2 ~ -0.06).
- release_year: A critical control for inflation and industry trends.

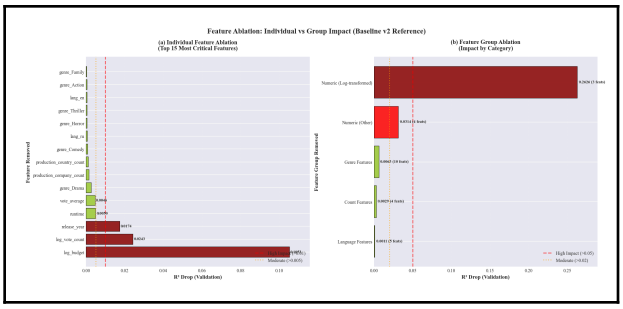

Fig. 3. Feature Ablation - Individual vs. Group Impact

The remaining 23 features (88% of the set) appear largely redundant. Removing all 10 Genre features simultaneously caused a statistically measurable drop, revealing latent feature interactions that individual ablation failed to capture. In some cases (e.g., release_month), removal slightly improved performance, indicating these features may introduce noise.

5.4 Optimal Subset Selection (Top-K)

To determine the minimal sufficient model, we evaluated the performance of Top-K subsets (Fig 4.). Contrary to complex industry models, our results indicate that a 3-feature model is optimal. While candidate subsets were evaluated on the validation set to identify the optimal K, the final metrics reported below reflect performance on the held-out test set to ensure an unbiased assessment

1) *Performance*: The Top-3 model achieved a Test R^2 of 0.7267.

2) *Efficiency*: This retains 95.1% of the baseline performance (0.7638).

3) *Parsimony*: By reducing the feature set from 26 to 3 (an 88.5% reduction), we achieve a highly interpretable model without statistically significant loss in practical utility.

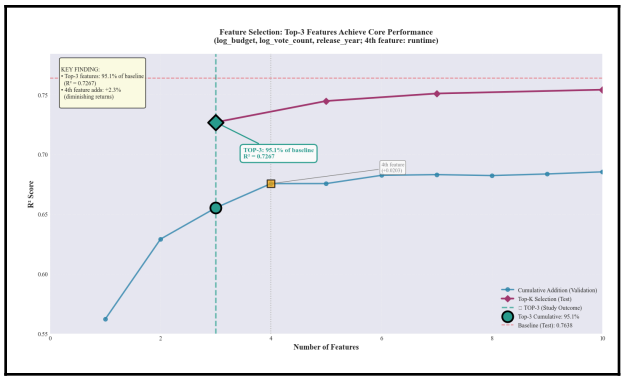

Fig. 4. Feature Ablation - Cumulative vs. Top K

5.5 Statistical Validation

The paired t-test confirmed that all Top-K models differed statistically from the baseline ($p < 0.001$), an expected result given the large test set ($n = 1,107$). However, effect size analysis (Cohen's d) revealed that these differences were negligible for all configurations ($|d| < 0.2$). Following the parsimony principle, we recommend the Top-3 model as it offers a 88% reduction in feature complexity while maintaining practical equivalence to the full baseline.

5.6 Cross-Analyst Robustness

Independent validation corroborated these findings. Both analysts identified the same Top-3 consensus features and observed the same diminishing returns curve/performance plateau, confirming that the sufficiency of the 3-feature subset is a robust property of the domain, not a modeling artifact.

6. Discussion

6.1 Interpretation of Key Predictors

Our results establish a stark Pareto efficiency: a minimal subset of 3 features (log_budget, log_vote_count, release_year) captures 95.1% of the baseline's predictive power (R^2=0.76).

1) *Budget as Capacity*: Consistently the rank-1 predictor, budget acts as a proxy for unmeasured variables like star power and production quality. The necessity of the log-transformation confirms that revenue scales exponentially with investment.

2) *Active vs. Passive Engagement:* Crucially, vote_count (Rank 2) outperforms TMDB's popularity metric (Rank 5). This indicates that active audience participation (voting) is a stronger predictor of commercial success than passive engagement metrics (page views).

In contrast, metadata features like genre and language offered negligible marginal utility ($\Delta R^2 < 0.005$). This suggests that financial success is largely genre-agnostic at the blockbuster level; a high-budget Action film and a high-budget Drama perform similarly once financial capacity is controlled for.

6.2 Implications and Limitations

These findings challenge the industry's reliance on "feature bloat". For analysts conducting post-hoc revenue attribution, the 3-Feature Model serves as a robust heuristic, demonstrating that the computational cost of processing complex feature interactions in metadata yields diminishing returns compared to simple financial and engagement signals. For pre-release decision-making, studios would need to substitute vote_count with available leading indicators such as pre-release buzz metrics or advance booking data. However, these results must be interpreted within the context of survivorship bias; our dataset (N=7,380) overrepresents major studio releases with complete financial data and likely does not generalize to micro-budget indie films. Furthermore, while vote_count is predictive, it is not strictly causal, high marketing budgets often drive the visibility that leads to votes. Future work should integrate Prints & Advertising (P&A) data to test if marketing spend displaces organic engagement signals.

Conclusion

This study challenged the prevailing trend of "feature bloat" in movie revenue modeling. Through a systematic dual-analyst ablation study, we demonstrated that modern model complexity is largely unnecessary. A minimal subset of 3 features (log_budget, log_vote_count, release_year) is sufficient to achieve 95.1% of the baseline accuracy (R^2=0.76). The remaining 23 features, including genre, language, and cast metrics, offer negligible practical utility. We conclude that for retrospective revenue analysis, model parsimony is not just a theoretical preference but a practical operational advantage, allowing for faster, more interpretable, and robust modeling with reduced data requirements.

References

Ahmed, M., Jahangir, M., Afzal, H., Majeed, A., & Siddiqi, I. (2020). Using crowd-source based features from social media and conventional features to predict the movies popularity. *IEEE Access, 8*, 20258–20273. https://doi.org/10.1109/ACCESS.2020.2968563

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 1, 492–499. https://doi.org/10.1109/WI-IAT.2010.63

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

De Vany, A., & Walls, W. D. (1999). Uncertainty in the movie industry: Does star power really matter? *Review of Industrial Organization, 15*, 285–318. https://doi.org/10.1023/A:1007833619566

Elberse, A. (2007). The power of stars: Do star actors drive the success of movies? *Journal of Marketing, 71*(4), 102–120. https://doi.org/10.1509/jmkg.71.4.102

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*(1), 3133–3181. https://jmlr.org/papers/v15/delgado14a.html

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182. https://www.jmlr.org/papers/v3/guyon03a.html

Hur, Y., Kang, P., & Cho, S. (2016). Box office prediction using social network analysis: The case of Korean movies. *International Journal of Information Management, 36*(6), 1251–1263.

Kakarla, R. (2020). *Movie data (100K+ titles with budget, credits)* [Data set]. Kaggle. Retrieved November 08, 2025, from https://www.kaggle.com/datasets/kakarlaramcharan/tmdb-data-0920

Kim, S. H., Park, N., & Park, S. H. (2015). Box office prediction using social media: An approach of social network analysis. *Cluster Computing, 18*, 257–269.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

Merity, S., Keskar, N. S., & Socher, R. (2018). Regularizing and optimizing LSTM language models. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1708.02182

Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). *Ablation studies in artificial neural networks*. arXiv. https://doi.org/10.48550/arXiv.1901.08644

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics, 8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications, 30*(2), 243–254. https://doi.org/10.1016/j.eswa.2005.07.018

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer, Cham. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, L., Luo, J., & Yang, S. (2020). Forecasting box office revenue with social media: A systematic review. *International Journal of Forecasting, 36*(3), 856–867. https://doi.org/10.1016/j.ijforecast.2019.11.002

Claude (Anthropic, 2025) was used as code creation assistance; Gemini Writing Editor (Google, 2025) was used as a language assistance tool to improve language clarity, coherence and to refine academic phrasing.