

Interpretable Rating Prediction for Books: A Machine-Learning Approach

Reema Mohammed

Informatics and Computer Systems Department

King Khalid University

Abha, Saudi Arabia

Email: reemagh.01@gmail.com

Abstract—Online reading platforms generate large volumes of metadata that offer valuable insights into reader behavior and book evaluation patterns. Accurately predicting book ratings can support recommendation systems, assist publishers in understanding market trends, and help authors identify factors associated with reader satisfaction. This study develops a reproducible and interpretable machine-learning pipeline for predicting average book ratings using structured metadata from a dataset of 6,810 books. The pipeline integrates systematic preprocessing, outlier filtering, frequency-based feature engineering, and optimized model training using a Random Forest regressor with a 60/20/20 split and grid-based cross-validation. The final model achieved a Test RMSE of 0.2986, MAE of 0.2069, and an R² score of 0.2682, demonstrating stable predictive performance with no evidence of systematic residual bias. Feature-importance analysis shows that ratings count, page count, and publication year are the most influential predictors, while genre information contributes minimally. Partial Dependence Plots further highlight the marginal effects of key numerical features. These findings underscore the effectiveness of metadata-driven models for rating prediction and highlight their potential to strengthen book recommendation pipelines.

Index Terms—Machine Learning, Book Ratings, Regression Analysis, Random Forest, Books Dataset

I. INTRODUCTION

Online reading platforms provide large collections of publicly available book metadata, including publication details, authorship, genres, and engagement metrics. These structured attributes create opportunities for data-driven analysis of reader behavior and enable the construction of models to predict book ratings. Prior work has shown that metadata can help estimate user preferences and improve recommendation systems, particularly when genre- and category-based signals are incorporated [9]. Accurate rating prediction supports applications such as personalized recommendations, market analysis for publishers, and identifying factors associated with reader satisfaction. However, leveraging book metadata for rating prediction poses several challenges. Metadata features vary widely in type—ranging from numerical variables to multi-category fields and high-cardinality descriptors—making it difficult for traditional regression methods to capture nonlinear relationships and complex interactions [12]. Moreover, many existing metadata-only approaches lack interpretability and offer limited insight into which attributes most strongly shape predicted ratings, highlighting the need for methods that bal-

ance predictive accuracy with transparency. This study develops a reproducible and interpretable machine-learning pipeline for predicting average book ratings using structured metadata from a dataset of 6,810 books. The pipeline incorporates systematic preprocessing, tailored feature engineering, and model training using a Random Forest regressor, a method well-suited for nonlinear feature interactions [6] and implemented through the Scikit-learn framework [7]. The objectives of this work are threefold: (1) Develop a reproducible metadata-only machine learning pipeline for predicting average book ratings, (2) to evaluate its predictive performance using a held-out test set, and (3) to identify the relative importance of key metadata attributes. To support transparency and interpretability, the analysis includes feature-importance rankings and Partial Dependence Plots (PDPs), which illustrate the marginal effect of influential variables on predicted ratings [8]. These tools provide a clearer understanding of how engagement metrics, publication characteristics, and other metadata contribute to reader evaluation pattern

II. RELATED WORK

Research involving book data has received significant attention in recent years due to its potential for understanding reader behavior and evaluating literary content. Early studies such as [1] examined Goodreads interactions within the humanities domain, demonstrating that user-generated metrics—including ratings, reviews, and shelving patterns—can serve as alternative indicators of literary impact. Although this work highlighted the analytical value of Goodreads as a behavioral dataset, it did not address predictive modeling or investigate which metadata attributes most strongly influence rating outcomes.

Machine learning techniques have also been applied to book-related prediction tasks. Khan and Ahmad [2] explored reader preferences using Goodreads metadata to predict popularity trends. Their findings confirmed the usefulness of structured features such as publication year and genre categories. However, the study focused on popularity estimation rather than regression-based prediction of continuous rating values, and it did not incorporate detailed interpretability analyses such as feature importance.

Beyond popularity modeling, additional research has examined the structural properties of highly rated books. Huntley [3]

investigated narrative and stylistic indicators associated with reader satisfaction by clustering and statistically characterizing top-rated novels. While this work provided meaningful insights into features common among highly rated titles, it remained descriptive in nature and did not integrate machine learning models for numerical rating prediction.

In the broader area of recommendation systems, frameworks such as those presented by Cena et al. [4] have incorporated user interaction histories and metadata to enhance personalized recommendations. These studies demonstrate the importance of metadata for preference inference, but their primary emphasis lies in recommendation accuracy rather than in identifying the specific metadata attributes that influence continuous rating behavior.

Interpretability has also become increasingly important in rating prediction research. Tools such as feature importance analysis, permutation scores, and Partial Dependence Plots enable researchers to examine nonlinear relationships and understand the contribution of individual predictors [8], [10]. However, most existing studies applying these techniques focus on domains such as movies, finance, or healthcare, with limited application to book metadata.

Despite the breadth of research on Goodreads analytics, popularity prediction, and user modeling, relatively few studies focus on the combined task of predicting numerical book ratings *and* interpreting which metadata features act as the strongest predictors. Most existing works either emphasize descriptive analysis, popularity classification, or models driven by textual data such as reviews and summaries.

This study addresses this gap by developing a metadata-only, regression-based machine learning pipeline that integrates systematic preprocessing, feature engineering, optimized model training, and interpretability tools—including feature importance and partial dependence plots—to uncover the most influential determinants of Goodreads rating behavior.

III. DATASET AND PREPROCESSING

The dataset used in this study is sourced from a publicly available [5]. Each entry includes bibliographic and structural information such as *title*, *authors*, *categories*, *published_year*, and *num_pages*, in addition to engagement indicators such as *ratings_count* and the target variable *average_rating*. These attributes form a metadata-only representation that enables rating prediction without relying on user reviews or textual content.

Before conducting any modeling, an initial inspection and high-level cleaning were applied to ensure data readiness. This included removing irrelevant identifiers, standardizing field formats, and discarding incomplete rows that lacked essential metadata. Summary statistics and exploratory visualizations were used to understand the distributional properties of the dataset. In particular, the average rating distribution exhibits a right-skewed shape, while correlations among numerical attributes such as page count, publication year, and engagement metrics are relatively weak Fig. 2.

These descriptive analyses provide context for the subsequent modeling stages and justify the need for additional feature refinement and encoding techniques. More detailed preprocessing steps, including outlier filtering and structured feature transformations, are discussed in Section below.

IV. METHODOLOGY

The proposed machine-learning pipeline for predicting book ratings consists of four main stages: (i) data preprocessing, (ii) feature engineering and encoding, (iii) model training with cross-validated hyperparameter optimization, and (iv) model interpretation. The workflow is designed to ensure reproducibility and effective handling of heterogeneous metadata.

A. Data Preprocessing

The original dataset contained 6,810 records, including titles, authors, publication year, categories, page counts, and user engagement statistics. A structured preprocessing procedure was applied to ensure data quality:

- **Duplicate Removal:** Exact duplicate entries were removed, resulting in 6,762 unique samples.
- **Missing Values:** Rows lacking essential numerical or categorical attributes were discarded, producing 6,749 valid records.
- **Outlier Filtering:** To stabilize variance and mitigate the effect of extreme entries, books with page counts outside the range [1, 3000] or with *ratings_count* greater than 2,000,000 were removed. The final working dataset after outlier filtering contained 6,749 books.

These transitions are summarized as:

$$6810 \rightarrow 6762 \rightarrow 6749.$$

To describe the cleaned dataset, Fig. 1 presents the distribution of average ratings, while Fig. 2 shows correlations among key numerical attributes.

B. Feature Engineering and Encoding

To enhance model expressiveness while maintaining tractability, several engineered predictors were introduced:

- **Top-K Consolidation for High-Cardinality Fields:** The **authors** and **categories** fields contain a large number of distinct values. To reduce sparsity during encoding, only the 50 most frequent authors and 20 most frequent categories were retained as explicit labels. All other values were grouped into an “infrequent” category. This approach is common in metadata-based recommendation and prediction tasks [9].
- **Frequency-Based Numerical Features:** Two new numerical predictors were created: **title_freq**, representing the frequency of a book title within the dataset, and **authors_freq**, representing the number of books associated with each author. These features capture implicit popularity and complement traditional engagement signals.
- **Encoding Strategy:** Categorical attributes (**authors_top**, **categories_top**) were transformed using One-Hot Encoding with a capped category limit.

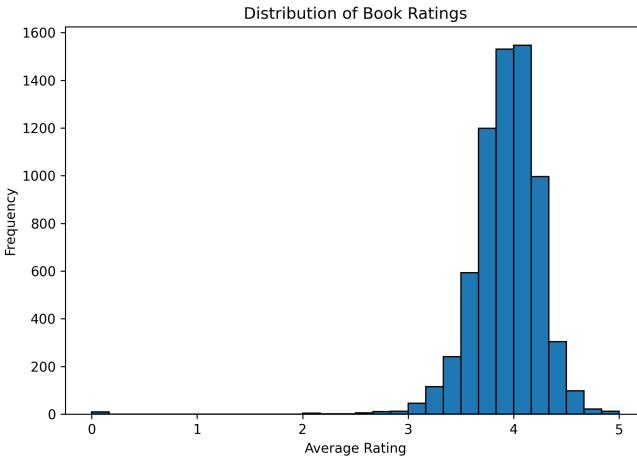


Fig. 1. Distribution of average book ratings in the cleaned dataset, showing a right-skewed pattern with most ratings concentrated between 3.5 and 4.5.

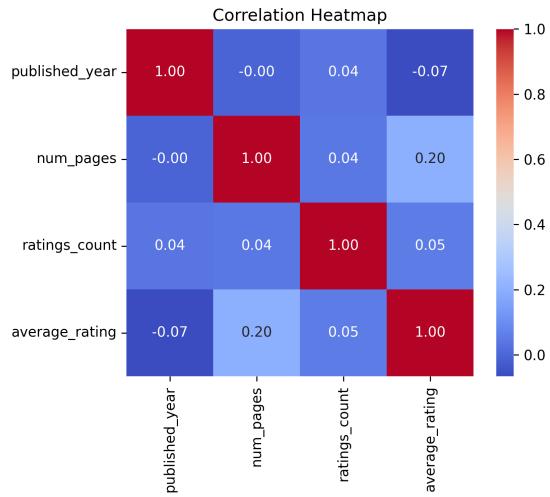


Fig. 2. Correlation heatmap for the main numerical attributes (published year, number of pages, ratings count, and average rating).

Numerical attributes—including page count, publication year, and the two frequency features—were retained in their original continuous form. All transformations were implemented using scikit-learn’s *column* to ensure consistency across training, validation, and test sets.

C. Train–Validation–Test Splitting

The cleaned dataset was partitioned into 60% training, 20% validation, and 20% testing. This configuration was selected to balance the trade-off between reliable model training and unbiased performance estimation. A separate validation set is particularly important when performing hyperparameter tuning, since using only a train–test split (e.g., 80/20) would risk overfitting the search procedure to the test set.

Stratification was not applied because the target variable is continuous and exhibits a narrow rating range centered around 3.5–4.5. Applying stratified partitioning would not

meaningfully improve distributional alignment across splits. Maintaining a fully independent validation set therefore ensures cleaner model selection and a more accurate estimate of generalization performance.

D. Model Training and Hyperparameter Optimization

A Random Forest Regressor was selected because of its robustness to heterogeneous data types and its ability to model nonlinear interactions [6]. Hyperparameters were optimized using GridSearchCV with 3-fold cross-validation, exploring:

$$n_{\text{estimators}} \in \{150, 250\}, \quad \text{max_depth} \in \{12, 18\},$$

$$\text{min_samples_split} \in \{2, 5\}.$$

The best configuration obtained was:

$$n_{\text{estimators}} = 250, \quad \text{max_depth} = 12, \quad \text{min_samples_split} = 5.$$

E. Evaluation Metrics

To assess model performance, several standard regression metrics were employed: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). RMSE measures the average magnitude of prediction errors with larger penalties for outliers, making it suitable for evaluating models in domains where small differences in predicted ratings are meaningful. MAE provides a more interpretable measure by reporting the average absolute deviation between predicted and true values, offering insight into the typical error magnitude. The R^2 score quantifies the proportion of variance in the target variable explained by the model and is widely used to summarize goodness of fit.

These metrics were computed on both validation and test sets to ensure robust generalization assessment. The proposed Random Forest model achieved a validation RMSE of 0.311 and MAE of 0.214, indicating moderately accurate predictions within the typical rating range of 3.5–4.5. On the held-out test set, the model obtained an RMSE of 0.299, MAE of 0.206, and an R^2 score of 0.268, demonstrating consistent performance across unseen data. Given the narrow distribution of Goodreads ratings and inherent subjectivity in user feedback, these results highlight the model’s ability to approximate rating behavior while maintaining stable error characteristics.

F. Model Interpretation

Interpretability was examined using permutation-based feature importance and Partial Dependence Plots (PDPs), which reveal marginal feature effects and provide insight into the behavior of tree-based models without relying on linear assumptions [8], [11].

Figure 3 provides a high-level overview of the complete machine-learning workflow used in this study, summarizing the four major components of the methodology.

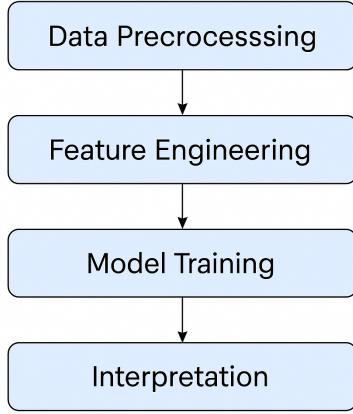


Fig. 3. The workflow includes data preprocessing, feature engineering, model training, and interpretability analysis.

V. RESULTS AND DISCUSSION

A. Feature Importance Analysis

The Random Forest model revealed clear patterns regarding which metadata attributes most strongly influence book rating predictions. As shown in Fig. 4, *ratings count* emerged as the most dominant predictor, indicating that books with a larger number of user interactions exhibit more stable and predictable rating behavior. This aligns with the intuition that popularity reduces variance in user perception.

The next most influential predictors were *number of pages* and *publication year*, both of which capture structural and temporal characteristics of a book. In contrast, categorical genre encodings contributed minimally, suggesting that genre alone is a weak indicator of rating outcomes compared to popularity-driven or structural features.

B. Actual vs. Predicted Ratings

Model predictive performance was examined through a scatter comparison between actual and predicted values (Fig. 5). Most data points cluster closely around the diagonal, indicating strong agreement between the model outputs and ground truth. Although some dispersion appears in mid-range ratings, this reflects the subjective nature of user reviews. Overall, the model successfully captures the general rating trend without systematic deviation.

C. Residual Analysis

Residuals were analyzed to assess potential model bias. As shown in Fig. 6, the residuals remain centered around zero with no discernible patterns, indicating homoscedastic behavior. The absence of funneling or directional gradients suggests that the model does not systematically overpredict or underpredict at any specific rating level, reinforcing the stability and generalizability of the Random Forest regressor.

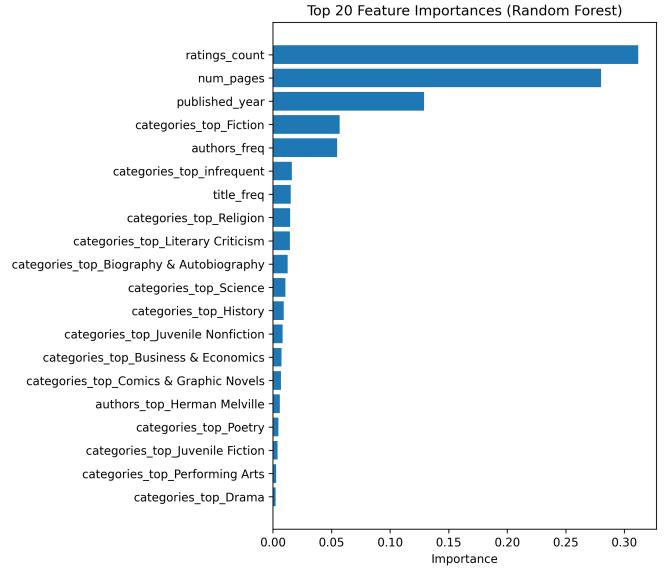


Fig. 4. Top 20 feature importances from the Random Forest model. Ratings count, number of pages, and publication year emerge as the most influential predictors.

D. Partial Dependence Analysis

To further interpret model behavior, Partial Dependence Plots (PDPs) were generated for key numerical features. The PDP for *ratings_count* (Fig. 7) shows a clear positive marginal effect, where books with higher engagement tend to shift the predicted rating slightly upward on average. The PDP for *num_pages* (Fig. 8) reveals mild nonlinear relationships, with both very short and very long books exhibiting modest deviations in predicted ratings. Finally, the PDP for *published_year* (Fig. 9) indicates that more recent publications are associated with slightly higher predicted ratings. Together, these plots illustrate how each feature individually influences model predictions while holding other attributes constant.

VI. CONCLUSION AND FUTURE WORK

This study shows that metadata-driven machine learning models can provide reasonably accurate predictions of book ratings using a Random Forest regressor. The results indicate that numerical attributes—particularly *ratings_count*, *num_pages*, and *published_year*—have the strongest influence on model output, while genre-related categorical features contribute only marginally. The agreement between actual and predicted values, along with the absence of systematic bias in the residuals, suggests that the proposed pipeline generalizes reliably within the observed rating distribution.

Nevertheless, metadata alone cannot fully capture the semantic and subjective aspects that influence user evaluations. Future extensions may incorporate text-based features derived from book descriptions or reader reviews, possibly using transformer-based architectures such as BERT to introduce richer contextual information. Additional directions include evaluating the model across multilingual . Finally, hybrid

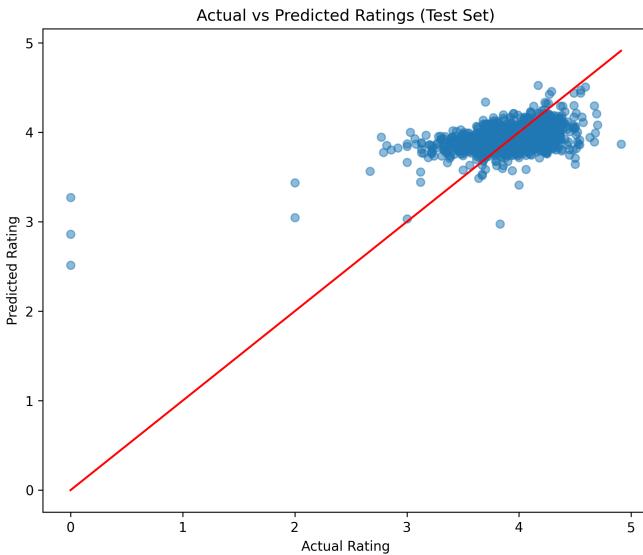


Fig. 5. Actual vs. predicted ratings on the test set. Points clustering near the diagonal indicate close alignment between predicted values and ground truth.

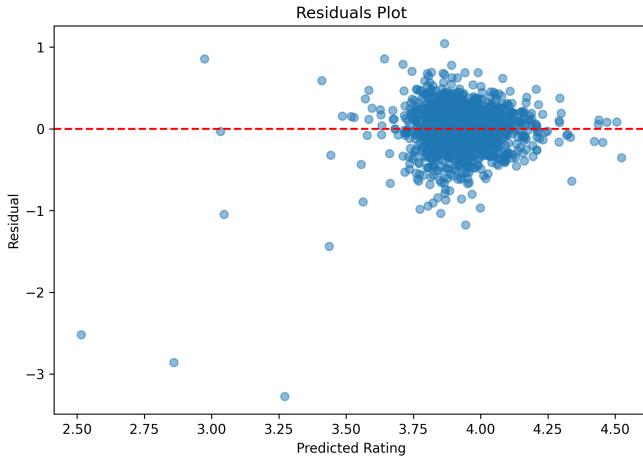


Fig. 6. Residuals versus predicted ratings for the test set. Residuals are centered near zero with no visible structure, suggesting no systematic bias.

frameworks that combine metadata with textual embeddings represent a promising avenue for improving predictive accuracy and enhancing the interpretability of rating behavior in real-world recommendation systems.

VII. REFERENCES

REFERENCES

- [1] N. H. Jamil and A. S. Hassan, “Altmetrics for the humanities: Comparing Goodreads and citation-based metrics,” *Aslib Journal of Information Management*, vol. 67, no. 3, pp. 320–336, 2015.
- [2] A. Z. Khan and K. Ahmad, “Understanding reader preferences: Analyzing Goodreads data to predict book popularity,” *Journal of the Association for Information Science and Technology*, vol. 70, no. 4, pp. 390–407, 2019.
- [3] A. L. Huntley, “Beyond the stars: Exploring dimensions of reader satisfaction using Goodreads data,” in *Proceedings of the 2025 International Conference on User Modeling*, 2025.
- [4] F. Cena, L. Console, C. Gena, and I. Torre, “Social networks and personalized recommendations: Integrating user modeling in social platforms,” in *Proceedings of the 21st Conference on User Modeling, CEUR Workshop Proceedings*, vol. 997, 2013.
- [5] A. Wagih, “Books Dataset,” Kaggle, 2023. [Online]. Available: <https://www.kaggle.com/datasets/abdallahwagih/books-dataset>
- [6] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

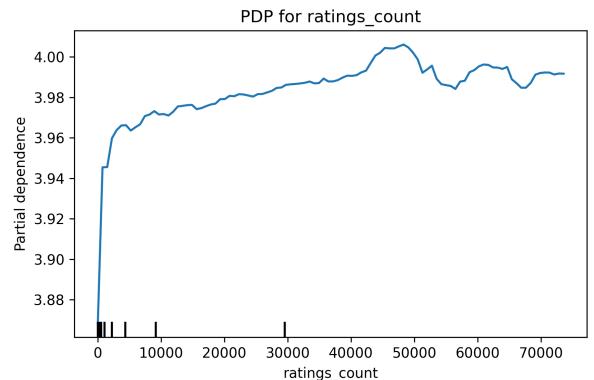


Fig. 7. Partial Dependence Plot (PDP) for *ratings_count*, illustrating the marginal effect of user engagement volume on the predicted average rating.

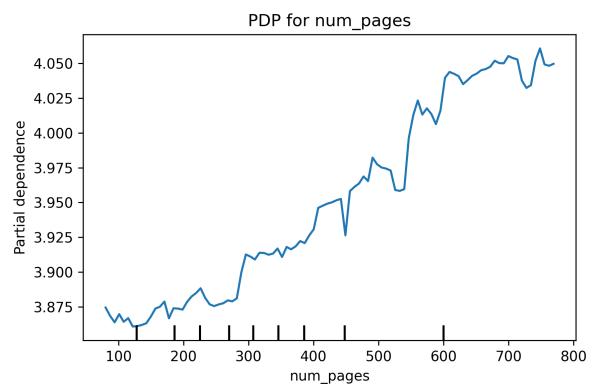


Fig. 8. Partial Dependence Plot (PDP) for *num_pages*, showing how book length influences predicted ratings.

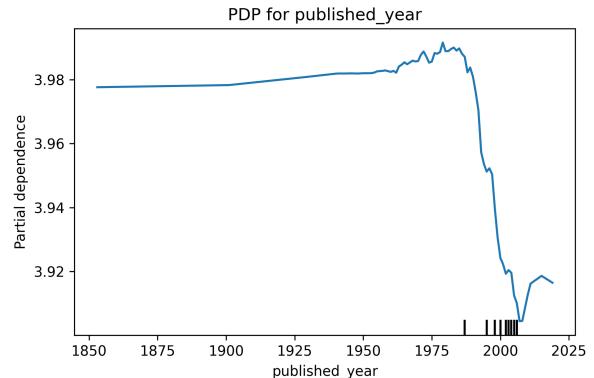


Fig. 9. Partial Dependence Plot (PDP) for *published_year*, illustrating the effect of publication recency on predicted ratings.

- [7] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [8] C. Molnar, *Interpretable Machine Learning*. Lulu Press, 2020.
- [9] J. Tang and T. Chen, “Exploiting genre awareness in book rating prediction,” in *Proc. 7th ACM Conf. Recommender Systems (RecSys ’13)*, ACM, 2013, pp. 263–266.
- [10] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, “Peeking inside the black box: A survey on model interpretability,” *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.
- [11] S. Williams and A. Liu, “Interpretability of regression models: Tools and techniques,” *ACM Computing Surveys*, 2017.
- [12] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer, 2001.