EMPIRICAL ANALYSIS OF THE BIAS-VARIANCE TRADEOFF ACROSS MACHINE LEARNING MODELS

Hardev Ranglani

EXL Service Inc.

ABSTRACT

Understanding the bias-variance tradeoff is pivotal for selecting optimal machine learning models. This paper empirically examines bias, variance, and mean squared error (MSE) across regression and classification datasets, using models ranging from decision tree to ensemble methods like random forest and gradient boost. Results show that ensemble methods such as Random Forest, Gradient Boosting and XGBoost consistently achieve the best tradeoff between bias and variance, resulting in the lowest overall error while simpler models such as Decision Tree and k-NN can have either high bias or high variance. This analysis bridges the gap between the theoretical bias-variance concepts and practical model selection, and offers in- sights into algorithm performance across diverse datasets. Insights from this work can guide practitioners in model selection, balancing predictive performance and interpretability

KEYWORDS

Bias, Variance, Mean Squared Error, Model Complexity, Bias-Variance Tradeoff

1. INTRODUCTION

For machine learning models, achieving optimal model performance is often a delicate balance between Bias and Variance, a concept known as the Bias-Variance tradeoff. This tradeoff is important because it determines a model's ability to generalize to unseen data, which is the most important way to measure model performance.

The overall error of a Machine Learning model is generally measured in terms of the Mean Squared Error. It tells us the expected value of the sqaure of the difference between the predicted value and the true value. The MSE for a model's predictions can be written as :

$$MSE = \mathrm{E}[(\hat{y} - y)^2]$$

where:

- y is the true value for an observation
- $\hat{y}$ is the predicted value for an observation
- The expectation E[.] is taken over different training samples This can be further

decomposed as :

$$\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{IrreducibleError}$$

Bias represents the error introduced by oversimplifying the model, leading to systematic inaccuracies in capturing the underlying data patterns. Variance measures a model's sensitivity to fluctuations in the training data, reflecting its tendency to overfit. These two sources of error are inherently in conflict: reducing bias often increases variance and vice versa. The Irreducible

DOI:10.5121/mlaij.2024.11401

ϵ

error represents the noise in the data or errors that cannot be modeled. For example, if y is modeled as $y = f(x) + \epsilon$, where ϵ is the noise, the irreducible erros is the variance of ϵ, denoted as σ^2. Understanding and quantifying these components is essential for selecting models that strike the right balance, minimizing both sources of error to achieve optimal generalization. Thus, the MSE can be broken down into its components as:

$$\text{MSE} = \underbrace{(\mathbb{E}[\hat{y}] - y)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[(\hat{y} - \mathbb{E}[\hat{y}])^2\right]}_{\text{Variance}} + \underbrace{\sigma_\epsilon^2}_{\text{Irreducible Error}}$$

As it is not always possible to emperically separate the irreducible error from the Bias, both these components are combined into Observed Bias as:

$$\text{MSE} = \text{ObservedBias}^2 + \text{Variance}$$

The emperical evaluation of Observed Bias and Variance for various machine learning models across different datasets can help us analyze this bias-variance tradeoff. We first define these components in a more theoretical way, then provide the methodology to calculate them emperically, and then calculate and analyze these metrics across different datasets and algorithms. We use seven popular machine learning models—Decision Tree, Random Forest, AdaBoost, Gradient Boosting, XGBoost, k-Nearest Neighbors, and Support Vector Machines—on 4 distinct datasets(2 real life datasets and 2 synthetic datasets): one each for regression and one for classification.

The novel contributions of this paper are:

1. Emperical evaluation of Observed Bias and Variance of various Machine Learning Models across real-life and synthetic datasets
2. Comparing patterns of bias-variance tradeoff and how these contribute to the overall Mean Squared Error of the model
3. Insights regarding the performance of simpler models such as Decision Tree and k-NN compared to ensemble methods such as Random Forest, Gradient Boosting, etc.

By systematically quantifying and comparing these metrics, this study aims to provide actionable insights into the selection and application of machine learning models. The findings underscore the importance of matching algorithms to task-specific requirements, ensuring robust and interpretable results in real-world applications. This work also highlights the power of ensemble methods, which often achieve a favorable bias-variance tradeoff through aggregation and boosting techniques.

This study is particularly relevant for practitioners and researchers seeking to understand the nuances of model selection and performance evaluation across regression and classification problems, guiding informed decisions based on data characteristics and predictive goals. The rest of the paper is organized as follows: the Literature review section discusses the previous work done on this topic, the Bias-Variance trade-off section defines these concepts in a more theoretical way, the Methodology section describes how these metrics are calculated and which datasets are used, and the Results section discusses the results and compares the performance of various algorithms. The Conclusion and Future work section summarizes the paper and discusses the future steps to be taken.

2. LITERATURE REVIEW

The bias–variance tradeoff describes the relationship between a model's complexity, the accuracy of its predictions, and and has been extensively studied over the decades. It plays a crucial role in determining how well a model performs on unseend data. Bias refers to the systematic error introduced by the model's inability to capture the complexity of the data, often due to oversimplification. Variance, on the other hand, measures the sensitivity of the model to fluctuations in the training data, reflecting its tendency to overfit. The conflict between these two forms of error is encapsulated in the decomposition of the Mean Squared Error (MSE), which can be expressed as the sum of Bias2, Variance, and Irreducible Error (Geman et al., 1992). Understanding this tradeoff is critical for designing models that strike the right balance between underfitting and overfitting.

Researchers have explored the decomposition of bias variation in a variety of algorithms and data sets. Linear models, such as linear regression and logistic regression, have been widely analyzed for their simplicity and low variance, though they are often limited by high bias in complex, nonlinear datasets (Harrell et al., 2001). Tree-based models such as Decision Trees, Random Forests, and Gradient Boosting have been shown to be capable of adapting to data complexity. Breiman (2001) demonstrated that Random Forests reduce variance through bootstrapping and aggregation, while Friedman et al. (2001) highlighted how boosting algorithms iteratively reduce bias by focusing on difficult-to-predict instances.

Support Vector Machines (SVMs) introduced by Vapnik (1995) provide a robust framework for handling high-dimensional data, using kernel methods to map features into nonlinear spaces. Similarly, k-Nearest Neighbors (k-NN), first analyzed by Cover and Hart (1967), is recognized for its low bias and high variance making it sensitive to noise but still effective for local patterns.

The choice of evaluation metrics has been studied extensively. Metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are commonly used for regression tasks, offering intuitive measures of the average error magnitude (Willmott & Matsuura, 2005). For classification, cross-entropy loss and probabilistic measures have been emphasized as critical for evaluating models that output probabilities (Murphy, 2012; Berrar, 2019).

Although the theoretical aspects of the bias-variance tradeoff have been well-established, empiri- cal studies have provided valuable insights into how these concepts perform on real-world data. Domingos (2000) evaluated the decomposition across zero-one and squared loss functions, and highlighted the tradeoff's universal relevance. Dietterich (2000) compared ensemble methods, such as bagging and boosting, demonstrating their ability to mitigate variance while maintaining low bias. However, most of these studies focus on regression or classification tasks, leading to the gap of evaluating both comprehensively.

There is also a lack of systematic comparisons across a wide range of algorithms using consistent evaluation metrics. Although data sets like Boston Housing and MNIST have been extensively studied, there is still a need for unified analyzes that investigate the behavior of bias and variance across both regression and classification tasks. Furthermore, metrics like Mean Absolute Percentage Error (MAPE) for classification and MAE for regression, which are more interpretable in real-world applications, are underexplored in the context of bias-variance analysis.

This study addresses these gaps by performing a comprehensive evaluation of seven machine learning algorithms. Decision Tree, Random Forest, AdaBoost, Gradient Boosting, XGBoost, k-NN, and SVM on regression and classification datasets. By comparing bias, variance, and additional metrics like MAE and MAPE across these tasks, this work provides a comprehensive

framework to understand how different models behave in diverse scenarios. This comparison offers actionable insights for practitioners, bridging theoretical concepts with practical applications, and establishing guidelines for model selection based on dataset characteristics such as complexity, noise, and feature interactions.

3. BIAS-VARIANCE TRADEOFF: UNDERSTANDING AND QUANTIFYING ERROR

The bias-variance tradeoff is a critical concept in machine learning that affects model selection and performance optimization. It describes the interaction between two sources of error: bias, which arises from over-simplifications in the model, and variance, which reflects the model's sensitivity to variations in the training data. Striking the right balance between bias and variance is essential to minimize the total prediction error, enabling a model to generalize effectively to unseen data. Figure 1 shows how the Mean Squared Error is influenced by the balance between $Bias^2$ and Variance as the model complexity increases as a U-curve. The curve for MSE is an addition of $Bias^2$ and Variance, with its minimum point indicating the optimal model complexity that balances bias and variance for the best generalization.

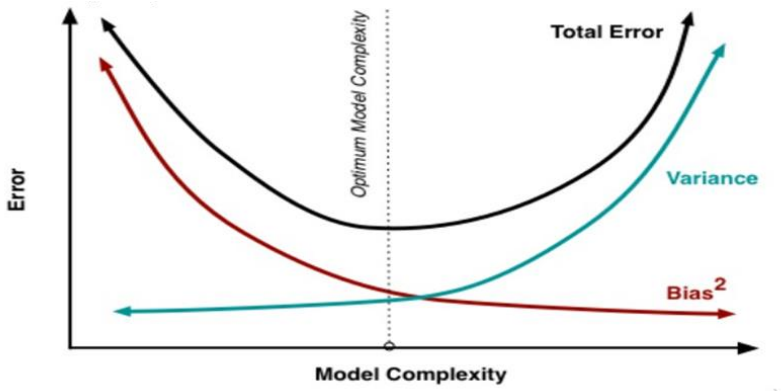

Figure 1: Effect of model complexity on Bias, Variance and Total Error.

The total error of a model, often measured using the Mean Squared Error (MSE), can be decomposed into three components: bias, variance, and irreducible error. Mathematically, this is expressed as follows:

$$\text{MSE} = \text{Bias}^2 + \text{Variance} + \text{IrreducibleError}$$

Each of these components contributes uniquely to the performance of the model. Bias measures the systematic error introduced when the model oversimplifies the true underlying function $f(x)$. It can be quantified as:

$$\text{Bias}^2 = \left(\mathrm{E}[\hat{f}(x)] - f(x)\right)^2$$

where $\mathrm{E}[\hat{f}(x)]$ is the expected prediction of the model, averaged over all possible training datasets. Models with high bias, such as linear regression applied to nonlinear data, fail to capture the complexity of the underlying relationships, leading to underfitting.

Variance, on the other hand, measures the variability of the model predictions when trained on different samples of the data. It is calculated as follows:

$$\text{Variance} = \mathrm{E}\left[\hat{f}(x) - \mathrm{E}[\hat{f}(x)]\right]^2$$

High-variance models, such as unregularized decision trees, tend to overfit, as they capture noise in the training data in addition to the true patterns. This makes them highly sensitive to fluctuations in the data.

The irreducible error accounts for the inherent noise in the data, represented as ϵ in the equation $y = f(x) + \epsilon$, where $\epsilon \sim \mathrm{N}(0, \sigma^2)$. This component reflects random factors or measurement errors that cannot be modeled or predicted, regardless of the complexity of the model.

In real-world datasets, the true function $f(x)$ is unknown, making it difficult to directly measure bias and variance. Instead, the squared observed bias often includes both the true bias and the irreducible error. As a result, the estimated bias squared is given by:

$$ObservedBias^2 = TrueBias^2 + \sigma^2$$

.

This inherent challenge with the irreducible error means that, while variance can be accurately computed, bias estimates inherently combine systematic error or true bias and noise.

To empirically measure bias, variance, and MSE for a machine learning model, a practical approach involves bootstrap sampling. By generating multiple training data sets through resampling and training the model on each, we can compute key metrics for a given test set. For each test data point x_i, the average prediction is :

$$\mathrm{E}[\hat{f}(x_i)] = \frac{1}{N}\sum_{j=1}^{N} \hat{f}_j(x_i)$$

where $\hat{f}_j(x_i)$ is the prediction from the j-th bootstrap sample, n is the total number of observations in the dataset and N is the total number of bootstrap samples. Using these predictions, bias squared can be estimated as:

$$\text{Observed Bias}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[\hat{f}(x_i)] - y_i\right)^2$$

and the variance is calculated as:

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N}\sum_{j=1}^{N}\left(\hat{f}_j(x_i) - \mathbb{E}[\hat{f}(x_i)]\right)^2$$

Finally, the total MSE for the test set is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbb{E}[\hat{f}(x_i)]\right)^2$$

In practical applications, these computations provide valuable insights into the behavior of different machine learning models. High-bias models tend to underfit, resulting in poor performance on complex datasets, while high-variance models overfit and fail to generalize. Ensemble methods, such as Random Forests and Gradient Boosting, achieve a favorable trade-off by reducing variance without significantly increasing bias, making them well-suited for a wide range of

datasets.

This study applies the above framework to analyze seven machine learning models in regression and classification tasks. By measuring bias, variance, and additional metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), this work bridges the- oretical insights with practical applications, guiding model selection for diverse real-world scenarios.

4. METHODOLOGY

This study analyzes the performance of machine learning models in both real-world and synthetic datasets to understand the bias-variance trade-off in regression and classification tasks. Four data sets, two real and two synthetic, were used to evaluate and compare the predictive performance of various machine learning algorithms.

4.1. Real-World Datasets

The Allstate Claims Severity dataset from Kaggle (DanaFerguson et al., 2016) was selected for the regression task. This dataset, used for regression modeling, contains 188,318 observations with 14 categorical features and 118 continuous features. The target variable, loss, represents the monetary amount of insurance claims. As it contains a high number of relevant features set and it has a real-world relevance, this dataset offers an opportunity to analyze model behavior in a highly nonlinear and noisy context. To prepare the data for modeling, categorical features were one-hot encoded, while continuous features were standardized. Based on the correlation of the target variable with the input variables, 19 variables were selected for modeling. The data were then divided into 80 % training sets and 20 % testing sets.

For the classification task, the Bank Marketing dataset(Moro, Rita, & Cortez, 2014) from the UCI Machine Learning Repository was chosen. This data set contains 45,211 observations and 17 characteristics, capturing demographic and campaign-related information. The target variable, y, indicates whether a client subscribed to a term deposit (yes or no). The preprocessing steps included one-hot encoding categorical features and scaling numerical features. The data was split into 70 % training and 30 % testing sets, ensuring a balanced representation of the two classes.

4.2. Synthetic Datasets

In addition to real-world data, two synthetic datasets were created to explore model performance in controlled environments. These data sets were designed to include linear and non-linear relationships, interaction terms, and random noise to mimic real world complexity.

For the regression task, a synthetic dataset with 10,000 samples and 10 features was generated. The target variable was a combination of linear terms, non-linear terms, interaction terms, and Gaussian noise. This allowed for a comprehensive evaluation of how well the models capture varying degrees of complexity and noise.

For the classification task, a similar synthetic data set was created. The target variable combined linear, non-linear and interaction terms, with the addition of a threshold to classify the target into binary classes(0 or 1). The target class was determined on the basis of whether the value of the continuous target variable exceeded its median, so as to maintain a balance between the two classes. This dataset facilitated a structured comparison of model performance in distinguishing non-linear decision boundaries under controlled noise conditions.

4.3. Experimental Setup

Seven machine learning algorithms were applied to all four datasets: Decision Tree, Random Forest, AdaBoost, Gradient Boosting, XGBoost, k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM). These models were chosen for their diverse mechanisms, from simple tree-based learners to ensemble methods and distance-based classifiers, providing a spectrum of complexity and interpretability.

To analyze the bias-variance tradeoff, bootstrap sampling was employed:

4.3.1. The data was split into a random but stratified 70-30 train test split and from the training data for each dataset, 100 bootstrap samples of size 70% of the size of the original dataset were generated by sampling with replacement.
4.3.2. Each model was trained on these bootstrap samples and predictions were made on the test set.
4.3.3. For each test data point, the average prediction and its variability were calculated across bootstrap iterations.

Bias and variance were estimated using the following equations

$$\text{Bias}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[\hat{f}(x_i)] - y_i\right)^2$$

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N}\sum_{j=1}^{N}\left(\hat{f}_j(x_i) - \mathbb{E}[\hat{f}(x_i)]\right)^2$$

where $\hat{f}_j(x_i)$ represents the prediction for the test point x_i from the j-th bootstrap model, n is the number of observations in the sample, N is the number of bootstrap samples and $\mathrm{E}[\hat{f}(x_i)]$ is the mean prediction across all bootstrap samples.

The Mean Squared Error (MSE) is then decomposed into its components.

$$\text{MSE} = \text{TrueBias}^2 + \text{Variance} + \sigma^2$$

, as the irreducible error cannot be directly measured, it is combined with the True Bias to form observed bias as:

$$\text{MSE} = \text{ObservedBias}^2 + \text{Variance}$$

These metrics are calculated and compared across all ML models for a given dataset

5. RESULTS

5.1. Classification task on the Bank Marketing data

The classification task on the Bank Marketing dataset reveals clear patterns in how machine learning algorithms balance bias and variance, influencing their overall performance, as described in figure 2. Decision Trees exhibit high variance and moderate bias, leading to a high MSE due to overfitting. AdaBoost and Gradient Boosting effectively reduce variance while maintaining low bias, while Gradient Boosting achieving significantly improved MSE. XGBoost, based on gradient boost, achieves the lowest bias and variance, resulting in the smallest total MSE and making it the best performing algorithm.

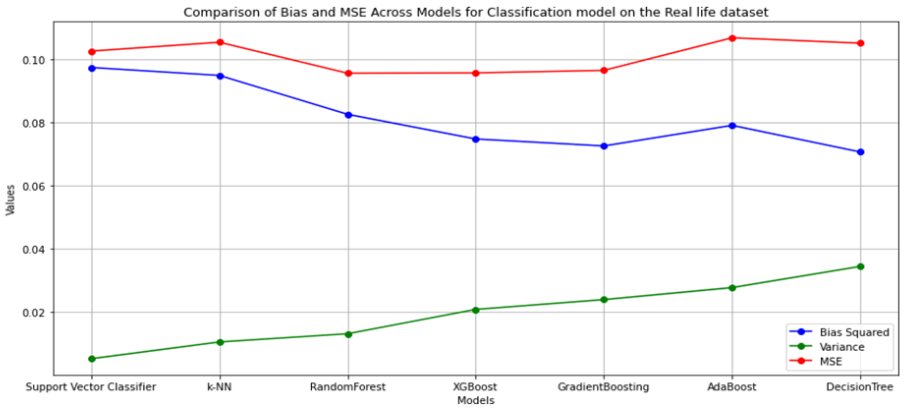

Figure 2: Bias and Variance trade-off for various ML models on the Bank Marketing Dataset

Random Forest model shows extremely low variance due to its averaging mechanism, though at the cost of slightly higher bias. It remains a strong contender with competitive MSE, making it a robust choice. Simpler models like k-Nearest Neighbors (k-NN) and Support Vector Ma- chines (SVM) struggle with high bias, resulting in moderate or high MSE. SVM, in particular, oversimplifies the classification task, leading to poor performance.

Overall, XGBoost outperforms all other models, followed by Gradient Boosting and Random Forest, which effectively balance bias and variance. These ensemble methods are better suited for complex datasets like Bank Marketing, while simpler models like k-NN and SVM require significant tuning to compete.

5.2. Classification Task on the Synthetic Data

The results for the synthetic classification dataset show a consistently low bias across all models, indicating that the relationships based on the dataset, linear, nonlinear, and interaction, are relatively simple to capture by the algorithms. The differences in performance majorly arise from the degree to which each model manages variance.

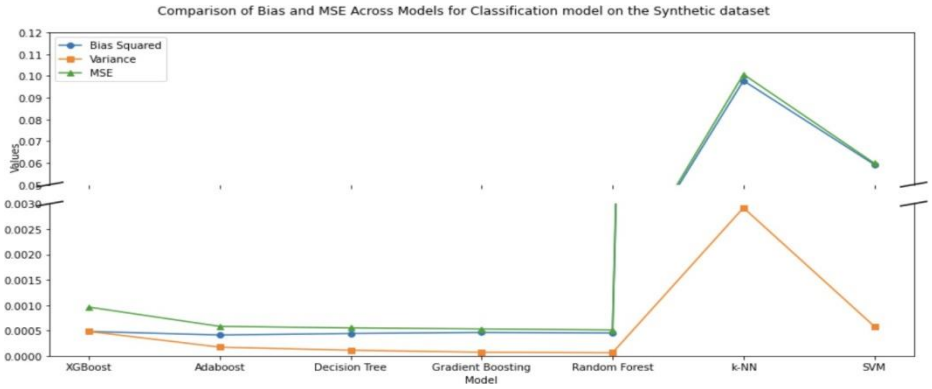

Figure 3: Bias and Variance trade-off for various ML models on the Synthetic Classification dataset

XGBoost achieves the best overall performance, as it balances low bias and well-controlled variance to deliver the lowest MSE. Gradient Boosting follows closely, offering a strong trade-off between bias and variance. Random Forest demonstrates excellent robustness with the lowest variance, though its slightly higher bias places it behind the boosting methods.

In contrast, Decision Trees exhibit high variance, resulting in the largest MSE despite their low bias. AdaBoost, while effective at reducing variance, has a slightly higher bias than XGBoost and Gradient Boosting, leading to a slightly higher MSE. Overall, boosting methods such as XGBoost and Gradient Boosting are the most effective, highlighting their ability to handle both complexity and noise while minimizing errors.

5.3. Regression Task on the All State Claims data

The analysis of the Allstate Claims Severity data set highlights significant differences in how algorithms handle the bias-variance tradeoff. XGBoost emerges as the best performing model, achieving the lowest MSE by minimizing both bias and variance. It effectively captures the complex patterns of the data set, making it highly suitable for this task. Gradient Boosting follows closely, as it balances slightly higher variance with low bias.

Random Forest shows excellent robustness with the lowest variance among all models, though its slightly higher bias places it behind the boosting methods. However, it remains a reliable choice, particularly for data sets with high noise.

In contrast, Decision Trees struggle with extreme variance, resulting in poor generalization and the highest MSE. Similarly, SVM and k-NN fail to capture the dataset's non-linear relation- ships, exhibiting high bias and moderate MSE, making them less effective for this regression problem. In general, ensemble methods such as XGBoost, Gradient Boosting, and Random Forest clearly outperform standalone models, emphasizing their suitability for complex and noisy regression tasks such as predicting claim severity.

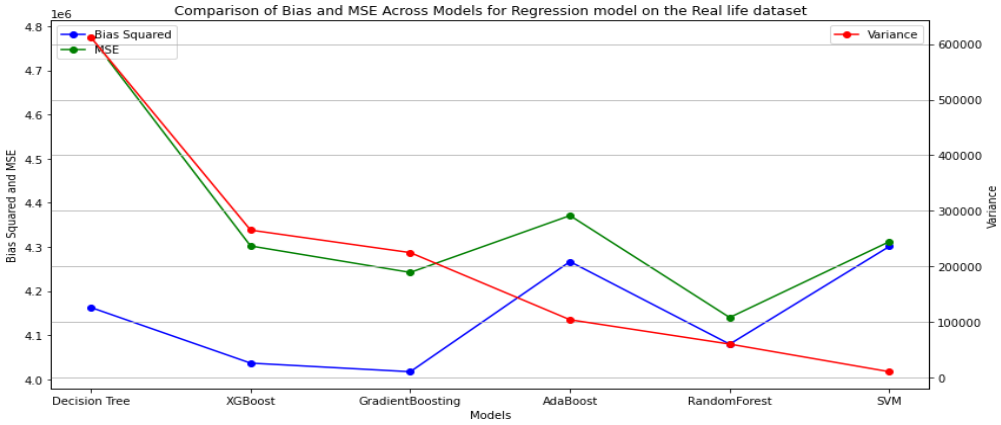

Figure 4: Bias and Variance trade-off for various ML models on the All State Claims dataset

5.4. Regression Task on the Synthetic data

The synthetic regression dataset, designed with a mix of linear, non-linear, and interaction patterns, results in consistently low bias across all models. This consistency reflects the fact that the relationships between inputs and outputs in the data are not overly complex, allowing all models to adequately capture the underlying structure. The key differences in performance arise primarily from the way each algorithm handles variance.

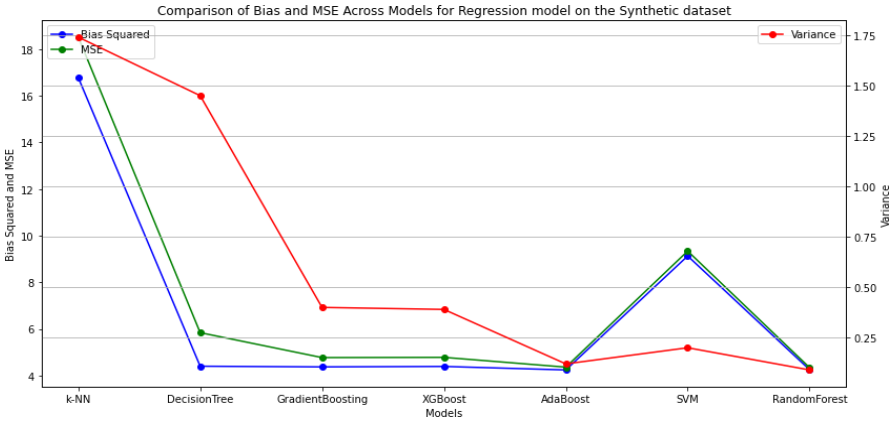

Figure 5: Bias and Variance trade-off for various ML models on the Synthetic Regression dataset

XGBoost achieves the best overall performance, combining low bias with well-controlled variance, resulting in the lowest MSE. It effectively captures the complexity of the dataset while avoiding overfitting, making it the top-performing algorithm. Gradient Boosting follows closely, balancing slightly higher variance with low bias to deliver strong results. Both boosting methods demonstrate their effectiveness in managing variance while maintaining a low bias.

Random Forest exhibits minimal variance, leveraging its averaging mechanism to achieve high robustness against overfitting. However, its slightly higher bias compared to the boosting methods leads to a marginally higher MSE. AdaBoost also performs well, offering a reasonable balance between bias and variance, though its simpler boosting mechanism prevents it from achieving the same level of performance as XGBoost and Gradient Boosting.

In contrast, Decision Trees struggle due to their high variance, which leads to poor generalization and the highest MSE among all models. Although their bias is low, their tendency to overfit

limits their effectiveness. This low and consistent bias between models highlights that the structure of the dataset is relatively simple for models to learn, leaving variance as the main differentiating factor.

Overall, XGBoost and Gradient Boosting stand out as the most effective algorithms for this synthetic dataset, with Random Forest offering a strong and robust alternative. Simpler models like Decision Trees struggle with high variance, emphasizing the importance of ensemble methods in managing complex relationships and noise.

Table 1 thus enlists all the metrics across all the models and datasets for comparison and the code used to build all the models and calculate the metrics can be found here

Table 1: Performance metrics of various machine learning models on real-life and synthetic datasets for regression and classification tasks.

Dataset	Task	Metric	DecisionTree	RandomForest	XGBoost	GradientBoosting	AdaBoost	SVM	k-NN
Real-Life	Regression	Bias Squared	4,162,902	4,079,684	4,036,769	4,017,277	4,266,771	4,301,220	9,908,458
		Variance	612,125	60,463	265,461	224,980	104,280	10,805	229,293
		MSE	4,775,027	4,140,147	4,302,230	4,242,257	4,371,051	4,312,025	10,137,751
	Classification	Bias Squared	0.0706	0.0825	0.0747	0.0725	0.0790	0.0973	0.0948
		Variance	0.0344	0.0131	0.0208	0.0239	0.0277	0.0052	0.0105
		MSE	0.1050	0.0955	0.0956	0.0964	0.1067	0.1025	0.1053
Synthetic	Regression	Bias Squared	4.40	4.25	4.39	4.37	4.24	9.13	16.78
		Variance	1.45	0.09	0.39	0.40	0.12	0.20	1.74
		MSE	5.84	4.34	4.78	4.77	4.36	9.34	18.52
	Classification	Bias Squared	0.00044	0.00045	0.00048	0.00046	0.00041	0.05923	0.09784
		Variance	0.00011	0.00006	0.00048	0.00007	0.00017	0.00058	0.00292
		MSE	0.00055	0.00051	0.00096	0.00053	0.00058	0.05981	0.10075

6. Conclusion & Future Work

This paper provides an in-depth analysis of the bias-variance tradeoff across multiple machine learning algorithms on both regression and classification tasks using real-world and synthetic datasets. The results highlight that ensemble methods like XGBoost, Gradient Boosting, and Random Forest consistently outperform simpler models by effectively balancing bias and variance and result in the lowest errors. Simpler models, such as Decision Trees and k-NN, show either high variance or high bias, emphasizing the importance of model complexity and regularization in achieving optimal performance.

The consistently low bias across models in synthetic datasets suggests that their manageable complexity allowed all models to adequately capture the underlying relationships, with variance being the primary differentiator. Ensemble methods, particularly XGBoost, demonstrated their superiority in managing variance and generalizing to unseen data.

Future work could extend this analysis to a broader range of datasets, including more complex real-world challenges, and explore the role of regularization and hyperparameter tuning in managing bias and variance. Further studies on deep learning models and their tradeoffs, as well as incorporating explainability techniques such as SHAP values, can provide additional insights into model performance. Developing methods to better estimate irreducible error in real-world data

also remains an open area of work. This work lays the foundation for selecting robust models and balancing predictive accuracy with generalizability in diverse machine learning tasks.

REFERENCES

[1] Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. Neural Computation, 4(1), 1-58.

[2] Harrell, F. E. (2001). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer.

[3] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[4] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

[5] Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer.

[6] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.

[7] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error over the root mean squared error. Climate Research, 30(1), 79-82.

[8] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688.

[9] Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

[10] Berrar, D. (2019). Cross-Entropy Loss Function. In Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics (pp. 546-550). Elsevier.

[11] Domingos, P. (2000). A unified bias-variance decomposition for zero-one and squared loss. In Proceedings of the Seventeenth International Conference on Machine Learning (pp. 231-238). Morgan Kaufmann.

[12] Dietterich, T. G. (2000). Ensemble methods in machine learning. In Proceedings of the First International Workshop on Multiple Classifier Systems (pp. 1-15). Springer.

[13] DanaFerguson, Meg Risdal, NoTrick, Sara R, Sillah, Tim Emmerling, and Will Cukierski. Allstate Claims Severity. https://kaggle.com/competitions/allstate-claims-severity, 2016. Kaggle.

[14] Moro, S., Rita, P., & Cortez, P. (2014). Bank Marketing [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.

[15] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.