**Model Development & Training**

**1. Introduction**

This report outlines the development and evaluation of several machine learning models used for classification tasks, focusing on performance comparison, analysis of learning curves, and model diagnostics. The models considered include Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine (SVM), and XGBoost. The goal is to identify the best-performing model based on accuracy and other relevant metrics, while considering interpretability and generalization performance.

**2. Model Architectures**

1. Logistic Regression:

- A linear model used for binary classification that predicts the probability of a sample belonging to a particular class by applying a logistic (sigmoid) function to a linear combination of the input features.

- Advantages: Simple, interpretable, and often effective when the relationship between input features and the output class is linear.

2. Decision Tree Classifier:

- A non-linear model that splits the data into subsets based on feature value thresholds. Each node represents a decision rule, leading to leaf nodes that make final predictions.

- Advantages: Easy to interpret and visualize. However, they tend to overfit on the training data unless regularized.

3. Random Forest Classifier:

- An ensemble of decision trees trained on random subsets of data and features. It combines the predictions of multiple trees to improve accuracy and reduce overfitting.

- Advantages: Reduces variance compared to a single decision tree, robust to overfitting, and works well on a wide range of datasets.

4. Support Vector Machine (SVM):

- A classification algorithm that aims to find the hyperplane that maximizes the margin between classes. SVM can work well with non-linear data by using kernel functions.

- Advantages: Effective in high-dimensional spaces, robust with various kernel options for non-linear classification.

5. XGBoost:

- An optimized implementation of gradient boosting that iteratively builds decision trees. Each subsequent tree corrects errors from the previous trees.

- Advantages: Highly efficient and performs well on structured/tabular data, but can be harder to interpret compared to simpler models like logistic regression.

**3. Model Performance Comparison**

Metrics: Models were evaluated based on Accuracy, Precision, Recall, F1 Score, and ROC-AUC. Cross-validation was used to assess model performance.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **ROC-AUC** |
| Logistic Regression | 0.8814 | 0.9202 | 0.6901 | 0.7887 | 0.9251 |
| Decision Tree Classifier | 0.8113 | 0.6997 | 0.7218 | 0.7106 | 0.7877 |
| Random Forest Classifier | 0.8847 | 0.8957 | 0.7254 | 0.8016 | 0.9272 |
| Support Vector Machine | 0.8678 | 0.9196 | 0.6444 | 0.7578 | 0.924 |
| XGBoost | 0.8847 | 0.8611 | 0.7641 | 0.8097 | 0.9303 |

**Key Observations:**

1. XGBoost achieved the highest ROC-AUC (0.9303), making it the best model in terms of distinguishing between classes.

2. Logistic Regression and SVM had the highest Precision values (0.9202 and 0.9196, respectively), indicating they excel at minimizing false positives.

3. Random Forest offered a balanced performance with strong Accuracy (0.8847), F1 Score (0.8016), and Recall (0.7254), making it a solid all-around model.

4. Decision Tree had the lowest overall performance, especially in Accuracy (0.8113) and ROC-AUC (0.7877), reflecting overfitting on the training data.

5. XGBoost had the highest Recall (0.7641), making it the most effective model at capturing true positives, along with a high F1 Score (0.8097), indicating a good balance between precision and recall.

**4. Learning Curves Analysis**

The learning curves revealed valuable insights about model behavior:

**-** Random Forest: exhibited high training accuracy (~1.00) but validation accuracy stabilized at 0.86-0.88. While some overfitting was present, the model performed well on unseen data.

**-** XGBoos**t:** also displayed very high training accuracy (~1.00) and slightly lower validation accuracy (0.85-0.87), indicating strong generalization with a minor degree of overfitting.

**Bias-Variance Tradeoff:**

- Random Forest & XGBoost: Both models struck a balance between bias and variance, leading to optimal performance with minimal overfitting.

**5. Recommendation:**

-XGBoost is the most balanced and robust model, with excellent generalization capabilities and performance across all metrics.

- If interpretability is needed, Logistic Regression can still be a strong contender, especially in scenarios where minimizing false positives is key.

- Random Forest offers another strong option, with a good mix of accuracy, precision, and recall, and less complexity compared to XGBoost.

**6. Conclusion**

In this analysis, we built and evaluated five machine learning models for classification tasks. XGBoost emerged as the top-performing model, followed closely by Random Forest. Both models demonstrated strong generalization with minimal overfitting. For applications requiring higher interpretability, Logistic Regression remains a valid option despite slightly lower performance.