Vomitoxin Prediction from Spectral Data: Analysis Report

1. Preprocessing Steps and Rationale

Data Standardization

- Standard Scaling: Applied to normalize the spectral bands, ensuring all features contribute equally to the model regardless of their original scale.
- Rationale: Spectral data typically varies across different wavelengths, and standardization prevents features with larger magnitudes from dominating the model training process.

Target Variable Treatment

- Distribution Analysis: Initial analysis revealed a skewed distribution of the vomitoxin_ppb variable.
- Log Transformation: Applied log transformation (np.log1p) to address the skewness and make the target distribution more normal.
- Outlier Handling:
    - Identified outliers using IQR (Interquartile Range) and Z-score methods
    - Applied capping at 1st and 99th percentiles to mitigate outlier influence without removing data points
    - Implemented Robust Scaling to further reduce sensitivity to extreme values
- Rationale: These treatments improve model performance by creating a more balanced target distribution that better satisfies regression model assumptions.

Feature Engineering

- Label Encoding: Applied to transform categorical 'hsi_id' values into numerical representations.
- Rationale: Converting text-based identifiers into numerical values allows machine learning algorithms to process this information.

2. Insights from Dimensionality Reduction

Principal Component Analysis (PCA)

- Implementation: Reduced the high-dimensional spectral data to 2 principal components for visualization.
- Visualization: The PCA scatter plot revealed patterns in the data with color coding based on vomitoxin levels.
- Insights:
    - The data points showed some clustering patterns related to vomitoxin levels
    - This suggests that the spectral data contains meaningful information for predicting vomitoxin concentration
    - The visualization helped identify potential structure in the high-dimensional spectral space

3. Model Selection, Training, and Evaluation

Model Selection

Three regression models were evaluated:

- Random Forest Regressor: Ensemble method using multiple decision trees
- Gradient Boosting Regressor: Sequential ensemble technique building trees to correct errors
- XGBoost Regressor: Optimized implementation of gradient boosting

Hyperparameter Optimization

- Grid Search Cross Validation: Implemented to systematically explore hyperparameter combinations
- Parameters Tuned:
    - Tree count (n_estimators): [100, 200]
    - Tree depth (max_depth): Various options per model
    - Learning rate: [0.01, 0.1] for boosting models
    - Minimum samples parameters for tree construction

Evaluation Metrics

Models were assessed using:

- Root Mean Squared Error (RMSE): Measure of prediction error magnitude
- Mean Absolute Error (MAE): Average absolute differences between predictions and actual values
- R^2 Score: Proportion of variance explained by the model

4. Key Findings and Suggestions for Improvement

Performance Comparison

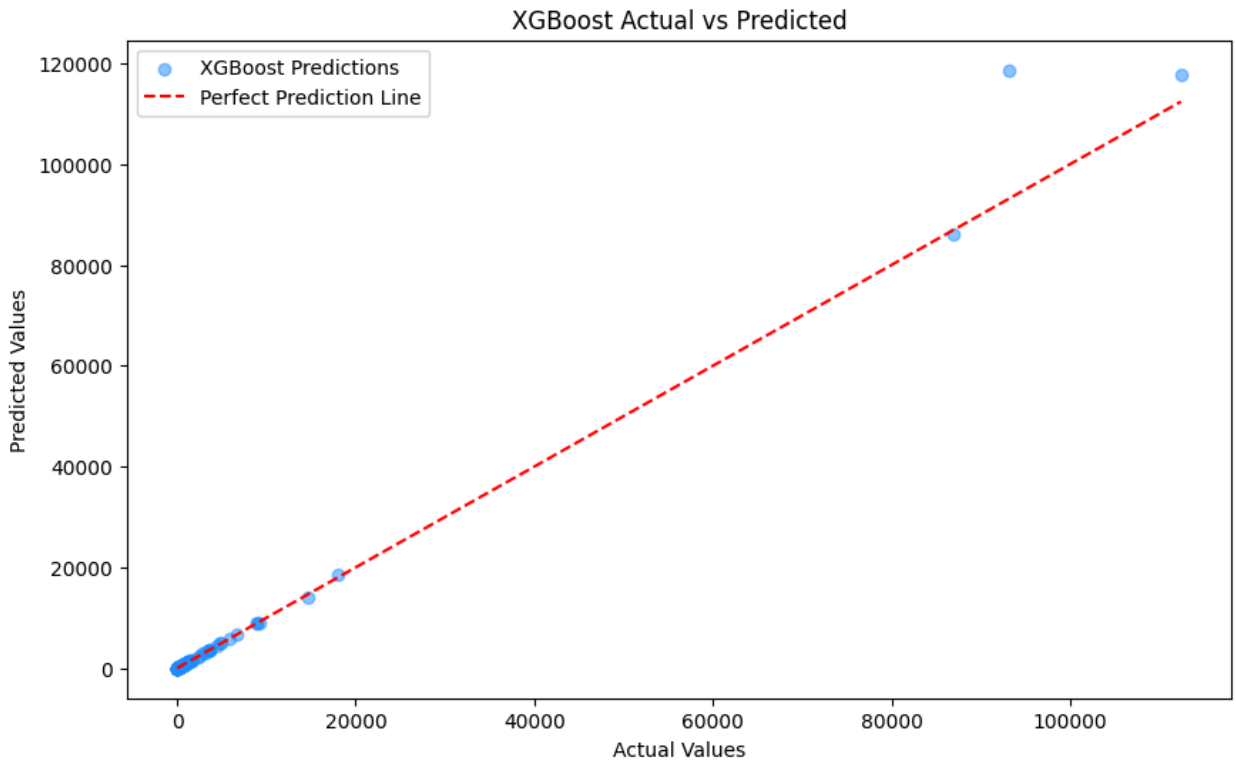
- Based on the model training results, the XGBoost model demonstrated the best overall performance with:
    - Higher R^2 score
    - Lower prediction errors (RMSE and MAE)
    - Better prediction alignment in actual vs. predicted plots

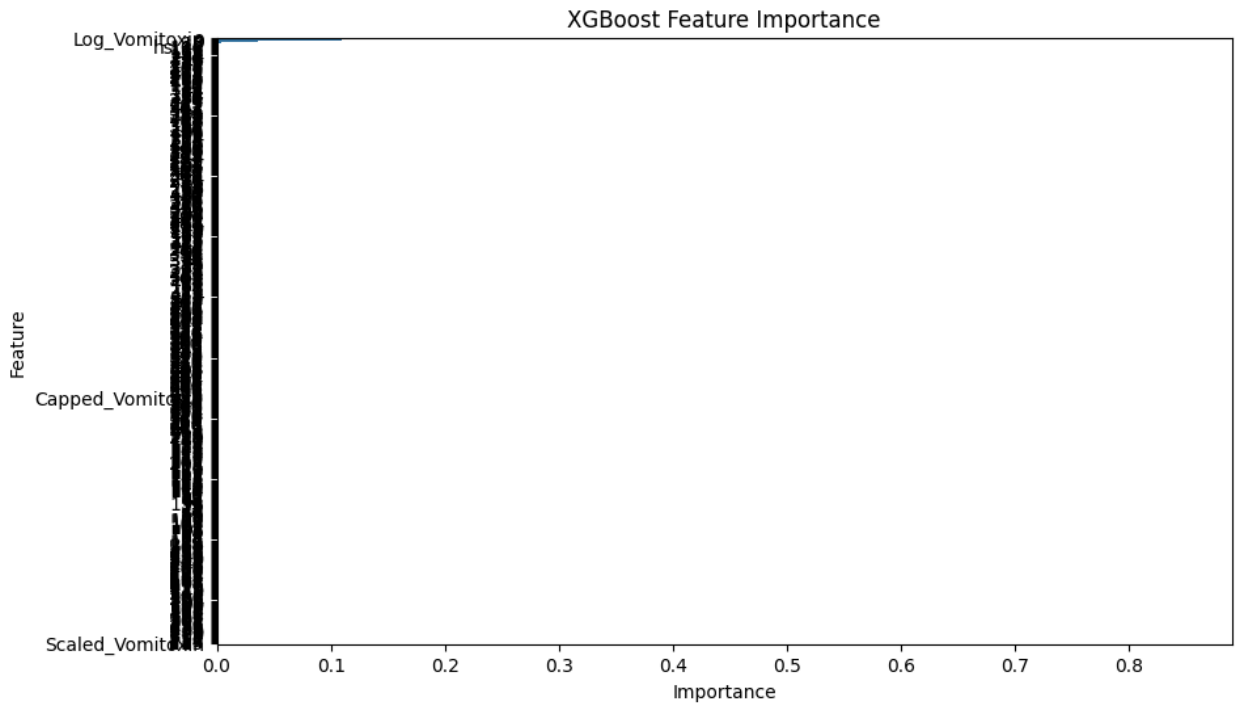
Feature Importance

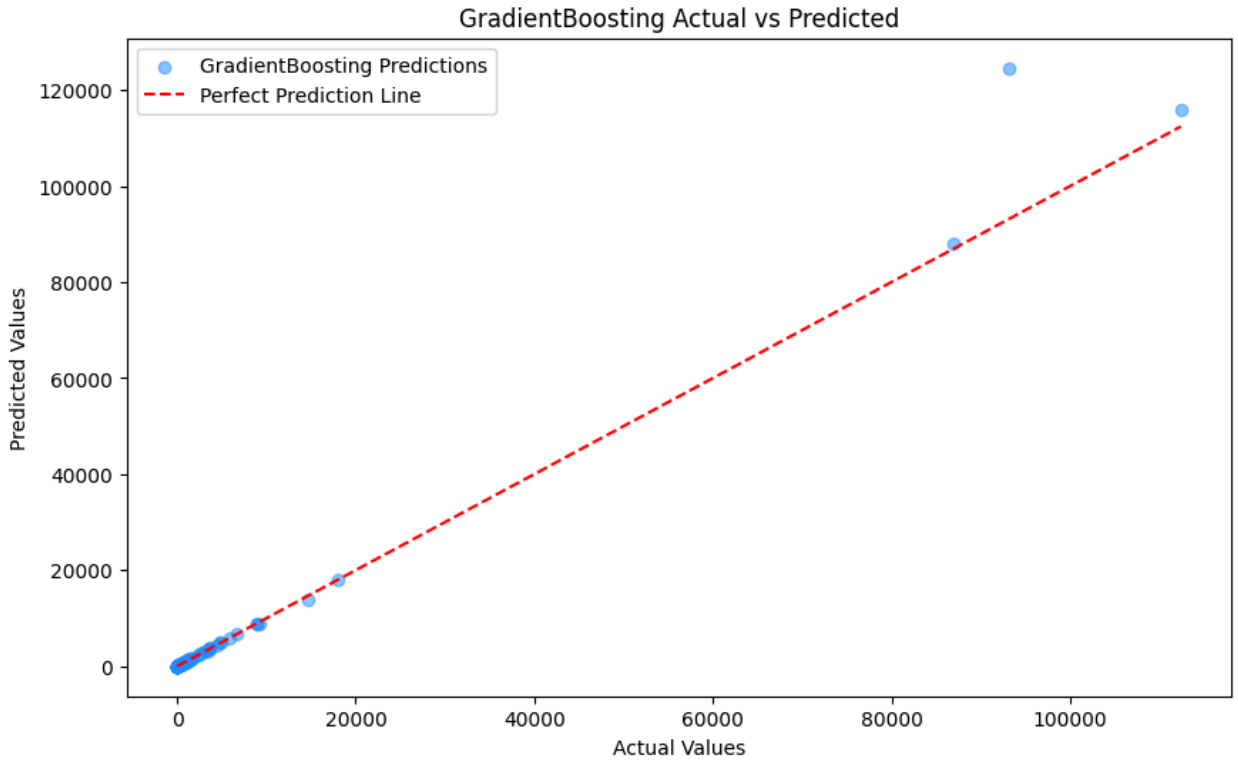
- Analysis of feature importance across models revealed specific spectral bands that consistently contribute to vomitoxin prediction.
- These high-importance wavelengths could be prioritized in future data collection or analysis.

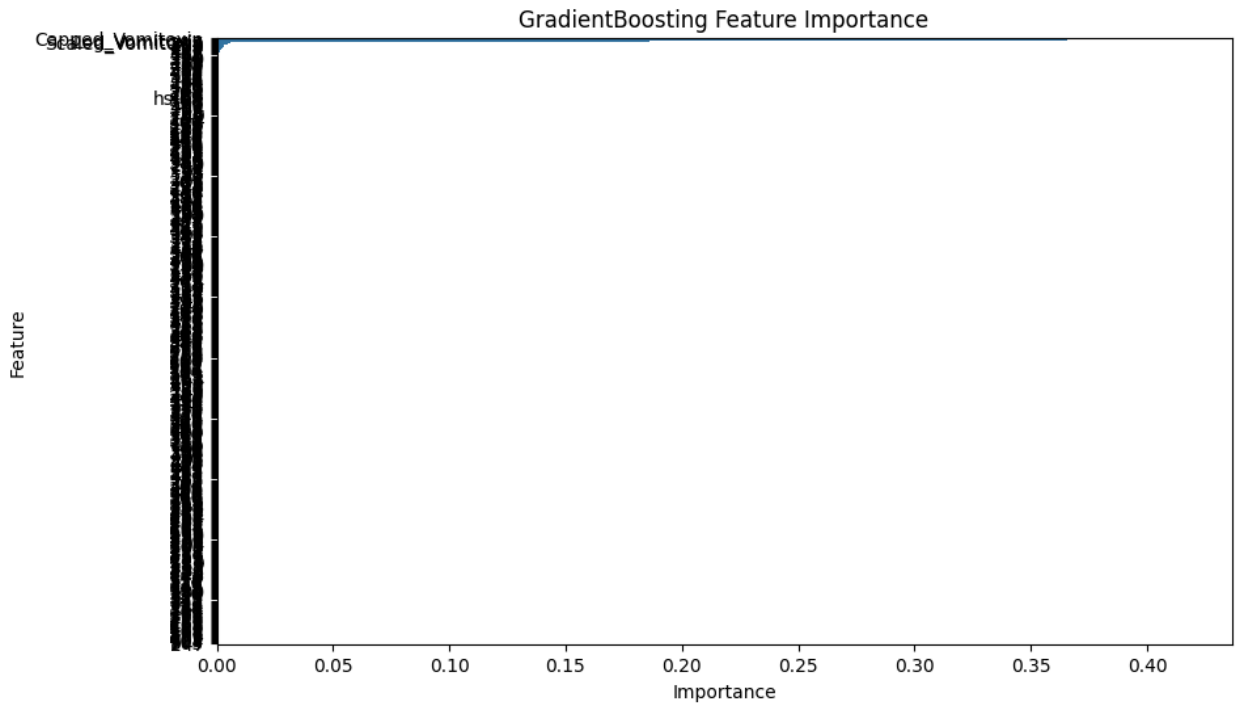
Results and Discussion (Graph):

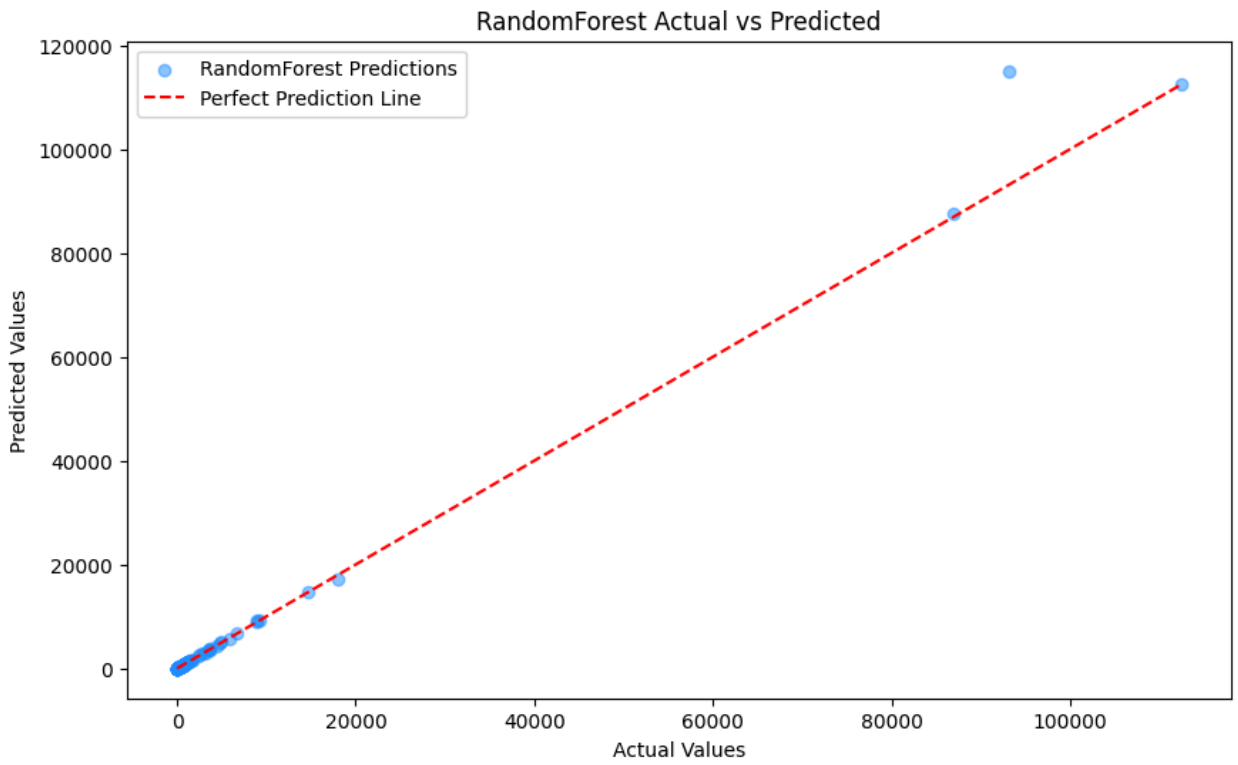
Model	RMSE	MAE	R Squared
XGBoost	2612.54	350.33	0.9756
Gradient Boosting	3170.04	391.03	0.9641
Random Forest	2198.96	256.76	0.9827
Attention Model	17318.59	—	-0.0722
Transformer Model	17291.26	—	-0.0692

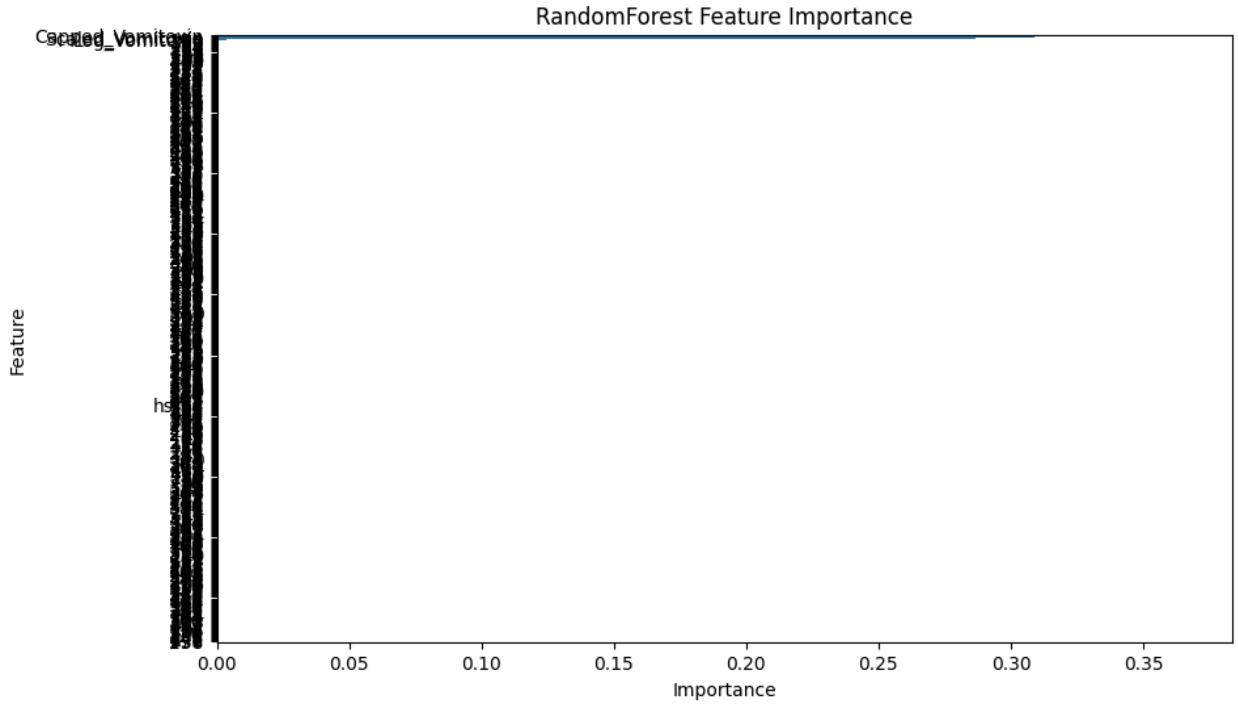



XGBoost Feature Importance
Feature
Importance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


GradientBoosting Actual vs Predicted
GradientBoosting Predictions
Perfect Prediction Line
Predicted Values
Actual Values
0
20000
40000
60000
80000
100000
120000


GradientBoosting Feature Importance
Feature
Importance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40





Discussion about Attention and Transformer Model

Attention and transformer models are underperforming. we might have to fine-tune the architecture or hyperparameters for better results