Analysis of XGBoost Model for prediction of mycotoxin levels in corn samples-

1. Data Preprocessing-

Steps Taken:

- **Handling Missing Values:** Missing data was imputed using the median to ensure a consistent dataset.
- **Outlier Detection:** While outliers were identified, no specific action was taken.
- **Dimensionality Reduction:** Principal Component Analysis (PCA) was applied to reduce high-dimensional spectral data to lower dimensions while retaining most variance.

Rationale:

- PCA helps mitigate multicollinearity and improves computational efficiency.
- Median imputation ensures minimal data loss, but other techniques might be explored.

2. Dimensionality Reduction Insights-

- PCA significantly reduced the number of features while maintaining key variance.
- The first few principal components captured the majority of the dataset's information.
- However, PCA reduces interpretability by transforming original features into abstract components.

3. Model Selection, Training, and Evaluation-

Model Used:

- **XGBoost Regressor** was selected for its robustness in handling structured data and its effectiveness in capturing complex relationships.

Training Process:

- The model was trained on the transformed PCA features.

Evaluation Metrics:

- The notebook mentions **Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2)** but does not provide exact values.
- Model performance on unseen data was assessed using a single test set, without cross-validation.

4. Key Findings and Recommendations-

Strengths:

- PCA effectively reduces dimensionality and maintains variance.
- XGBoost is a strong choice for regression due to its handling of non-linearity.

Model Performance Summary-

The model applies XGBoost regression after PCA (Principal Component Analysis) on spectral data to predict vomitoxin (DON) concentration. While performance metrics such as MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and R^2 (Coefficient of Determination) are computed, their specific values are:-

MAE: 1525.0408073046804, **RMSE:** 2697.318992244902, **R_2 Score:** 0.9739725656547707.

Limitations:

Data quality: The model's performance is heavily reliant on the quality of the input data. The code addresses missing values by imputation with the median, but this might not be optimal. Further investigation into data quality and more robust handling of missing values and outliers are crucial.

Feature engineering: The current feature engineering relies solely on PCA. Other feature engineering techniques might reveal more informative features. Exploring spectral indices, derivatives, or other domain-specific features could improve predictive accuracy.

Interpretability: PCA reduces dimensionality but loses interpretability. Understanding the contribution of the original spectral bands to the model's prediction is difficult.

Generalization: The model's performance on unseen data needs to be evaluated rigorously. A larger, more diverse dataset is required to assess the true generalization ability.

The model provides a baseline prediction, but further improvements can be achieved by addressing these identified limitations.