Mycotoxin in Corn Report

Preprocessing Steps and Rationale

1. Dropping Non-Numeric Columns

- **Rationale**: The hsi_id column, which is a unique identifier for each sample, was dropped because it does not contribute to the predictive power of the model.

2. Standardization

- **Rationale**: The features were standardized using StandardScaler to ensure that all features have a mean of 0 and a standard deviation of 1. This step is crucial for models that are sensitive to the scale of input data, such as SVM and neural networks.

3. Log Transformation

- **Rationale**: The target variable vomitoxin_ppb was log-transformed using np.log1p to reduce skewness. This transformation helps in normalizing the distribution of the target variable, making it more suitable for linear models.

4. Feature Selection using Lasso

- **Rationale**: Lasso regression with an alpha of 0.005 was used for feature selection. Lasso helps in selecting the most important features by shrinking the coefficients of less important features to zero, thus improving model interpretability and performance.

5. Principal Component Analysis (PCA)

- **Rationale**: PCA was applied to reduce dimensionality while retaining 98% of the variance in the data. This step helps in reducing the complexity of the model and mitigating the curse of dimensionality.

6. Train-Test Split

- **Rationale**: The dataset was split into training and testing sets with an 80-20 ratio to evaluate the model's performance on unseen data.

Insights from Dimensionality Reduction

- **PCA**: By retaining 98% of the variance, PCA reduced the number of features significantly, which helped in speeding up the training process and reducing the risk of overfitting. The reduced feature set still captured the essential information needed for accurate predictions.

Model Selection, Training, and Evaluation

1. XGBoost Regressor

- **Training**: The model was trained with 2000 estimators, a learning rate of 0.003, and a max depth of 10. Early stopping was used to prevent overfitting.
- **Evaluation**: The model achieved a test MAE of 2.0458 and an R^2 of 0.1402.

2. Random Forest Regressor

- **Training**: The model was trained with 700 estimators and a max depth of 18.
- **Evaluation**: The model achieved a test MAE of 2.0772 and an R^2 of 0.1397.

3. Support Vector Regressor (SVR)

- **Training**: The model was trained with an RBF kernel, C=10, and gamma='scale'.
- **Evaluation**: The model achieved a test MAE of 1.7142 and an R^2 of 0.1061.

4. Neural Network

- **Training**: The model was trained with a architecture consisting of dense layers with 512, 256, and 128 units, using Swish activation. Batch normalization and dropout layers were used for regularization.
- **Evaluation**: The model achieved a test MAE of 1.8523 and an R^2 of 0.2095.

Key Findings

- **Performance**: XGBoost and Random Forest performed similarly, with XGBoost slightly outperforming Random Forest on the test set. SVR had the lowest test MAE but a relatively low R^2, indicating that it may not capture the variance in the data as well as the other models. The Neural Network showed a balanced performance with a decent R^2 and relatively low MAE and RMSE on the test set.
- **Overfitting**: The models showed signs of overfitting, as indicated by the significant difference between training and test performance metrics.

Suggestions for Improvement

1. **Hyperparameter Tuning**: Further tuning of hyperparameters could help in improving model performance and reducing overfitting.
2. **Feature Engineering**: Additional feature engineering and selection techniques could be applied to enhance model accuracy.
3. **Ensemble Methods**: Combining the strengths of different models using ensemble methods could lead to better predictive performance.

4. **Cross-Validation**: Implementing cross-validation during model training could provide a more robust evaluation of model performance.

Conclusion

This project demonstrated the application of various machine learning models to predict vomitoxin concentration in agricultural products. While the models showed reasonable performance, there is room for improvement through further tuning and advanced techniques. Future work could focus on enhancing model accuracy and robustness to make more reliable predictions.