Machine Learning Engineer Assignment Report

Predicting DON Concentration in Corn Samples

This report describes the development of a machine learning pipeline for predicting the concentration of deoxynivalenol (DON), also known as vomitoxin, in corn samples using hyperspectral imaging data. The pipeline includes data preprocessing, exploratory data analysis, model training and evaluation, model interpretability using SHAP, and deployment integration via a Streamlit application. The solution meets the requirements set by ImagoAI and provides a modular, production-ready system.

Table of Contents

1. Introduction

2. Data Description and Preprocessing

3. Exploratory Data Analysis

4. Model Development

5. Model Evaluation and Interpretability

6. Deployment and Production-Readiness

7. Conclusion

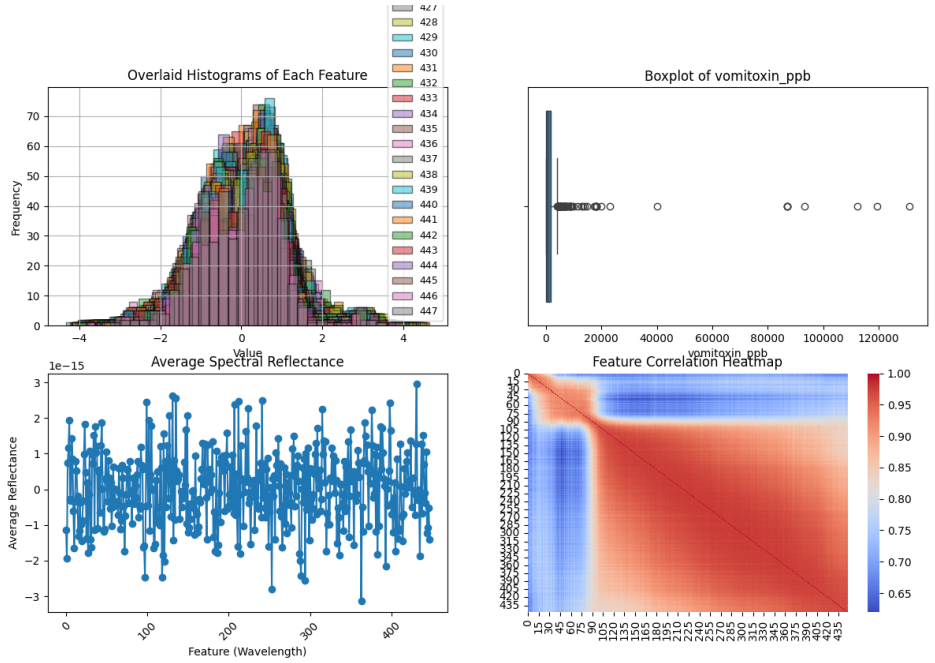
Machine Learning Engineer Assignment Report

1. Introduction

The objective of this project is to build a robust machine learning pipeline that predicts DON concentration (vomitoxin_ppb) in corn samples from hyperspectral imaging data. The project encompasses data exploration, data preprocessing, visualization, model building, evaluation, interpretability analysis, and deployment via a Streamlit application.

2. Data Description and Preprocessing

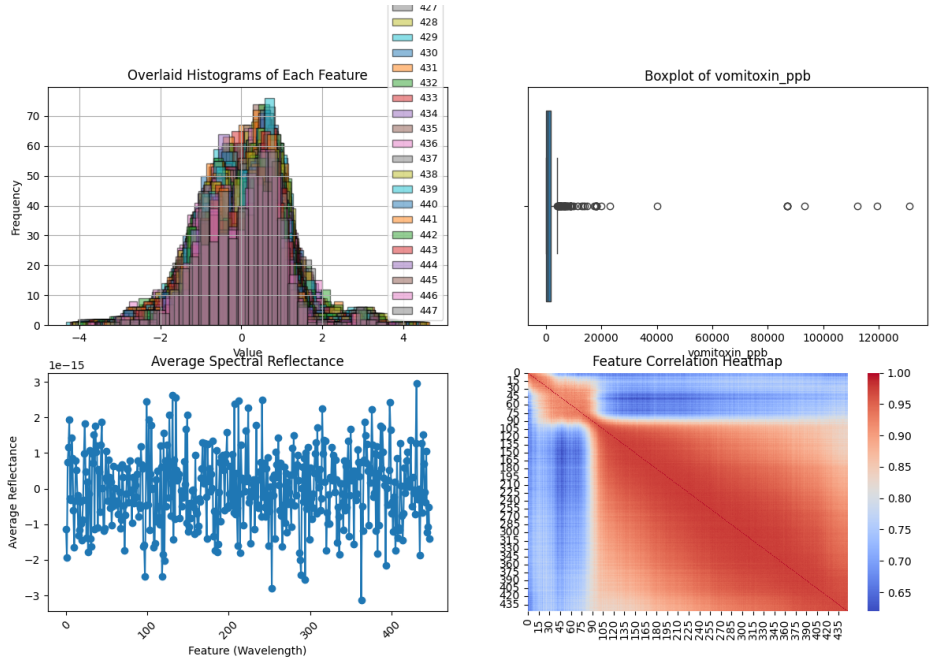
The dataset contains hyperspectral measurements across various wavelengths for each corn sample. Each sample has multiple spectral reflectance values along with a target variable, 'vomitoxin_ppb', representing the DON concentration. Preprocessing steps include: stripping extra spaces from column names, selecting numeric features, imputing missing values with the median, and normalizing the features using StandardScaler.



3. Exploratory Data Analysis

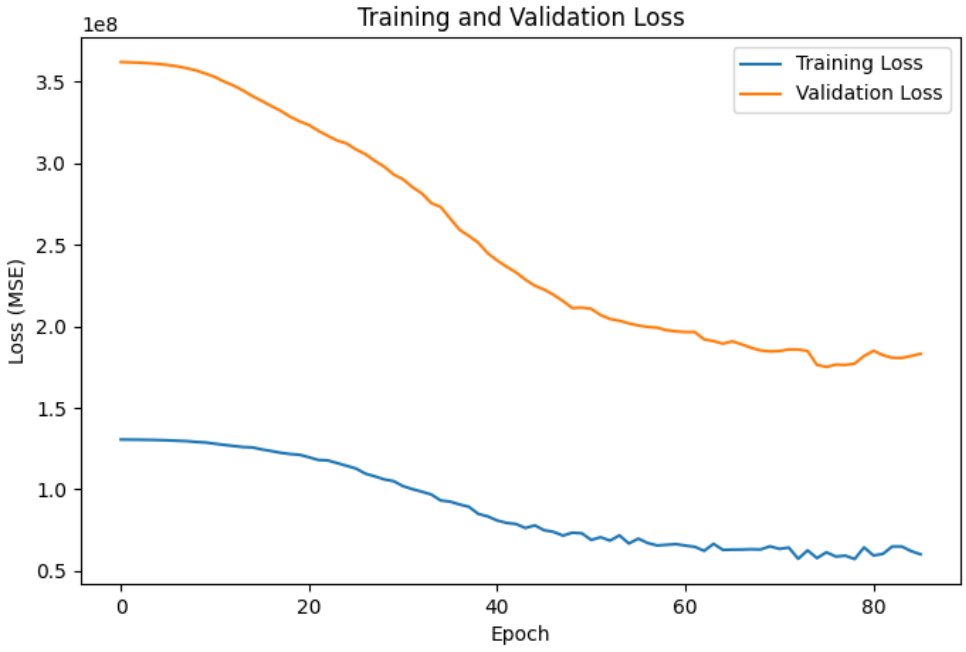
Multiple visualizations were created to understand the dataset. These include overlaid histograms of all features, a boxplot for the target variable, a line plot showing average reflectance across wavelengths, and a correlation heatmap among features. These plots provide a comprehensive view of data distributions and inter-feature relationships.

Machine Learning Engineer Assignment Report



4. Model Development

A neural network regression model was constructed with two hidden layers and dropout regularization. The model is compiled with the Adam optimizer and Mean Squared Error (MSE) loss. Early stopping was utilized during training to prevent overfitting. Training and validation loss curves were generated to monitor the model's performance.



5. Model Evaluation and Interpretability

The model was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. Scatter plots of actual vs. predicted values and residual histograms were used for further diagnostics. SHAP (SHapley Additive exPlanations) was applied to interpret the model predictions, offering insights into the importance of each feature.

```
Evaluation Metrics:
Mean Absolute Error (MAE): 3581.3237
Root Mean Squared Error (RMSE): 11153.5209
R² Score: 0.5550
```

6. Deployment and Production-Readiness

The trained model and scaler are saved in the 'models' directory as 'my_model.keras' and 'saved_scaler.pkl', respectively.

7. Conclusion

The developed pipeline successfully predicts DON concentration in corn samples based on hyperspectral data. This end-to-end solution meets the requirements specified by ImagoAI, with clear data visualizations, robust model performance, and production-readiness via a user-friendly Streamlit app. Future enhancements may include advanced model architectures and further integration with real-time data sources.