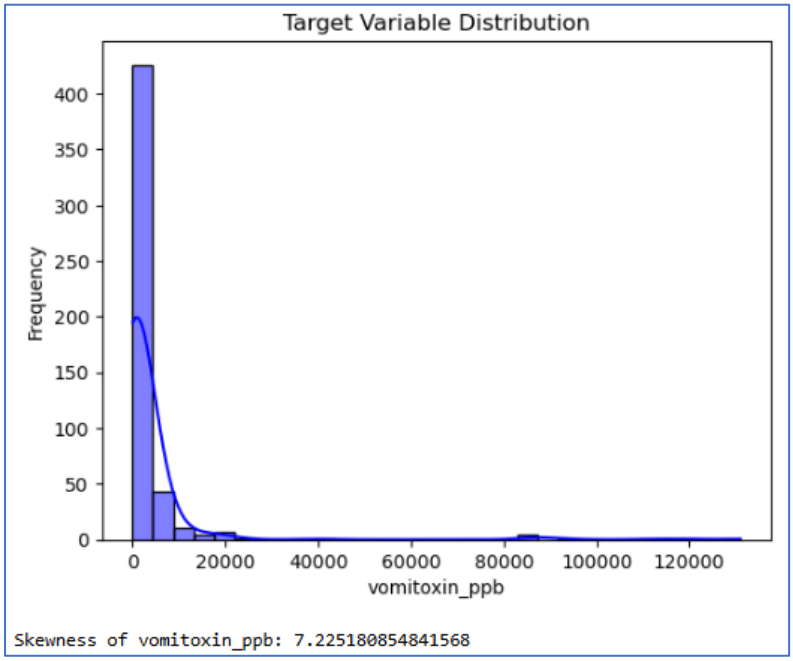
Corn Hyperspectral Reflectance for DON Prediction

Dataset Overview

The dataset consisted of 448 features along with a unique sample ID and a target column representing the Vomitoxin level in corn samples. There were 500 samples in total. On closer examination, it was found that 80 out of 500 target values were zero and 88 were outliers. Removing them was not considered meaningful as they were part of the natural data distribution. To reduce dimensionality and standardize the features, PCA (Principal Component Analysis) and StandardScaler were applied.

The target variable showed a skewness of 7.225, indicating a right-skewed distribution. A target variable distribution plot confirmed this skewness.



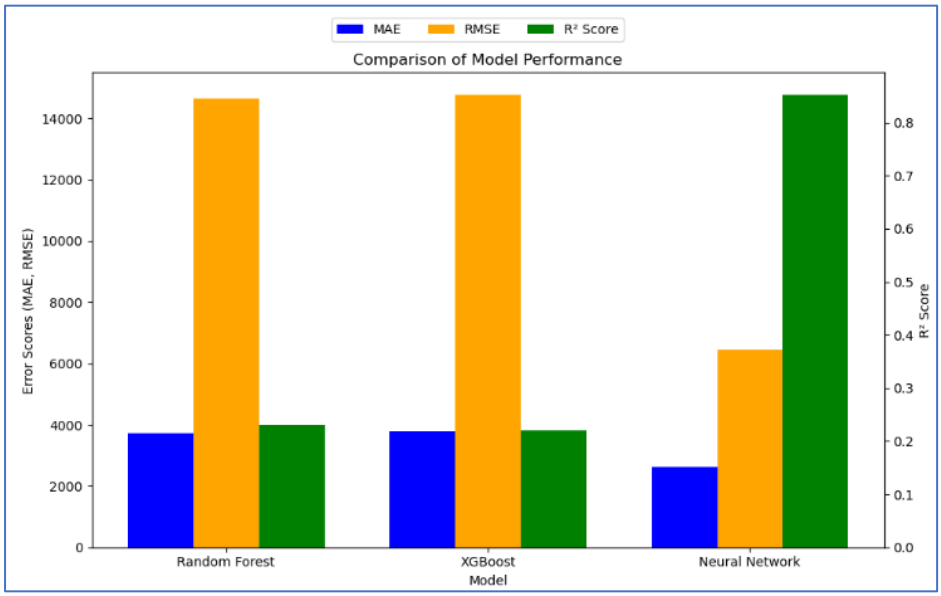
Target Variable Transformation

To address the skewness, different transformation techniques such as log1p, Box-Cox, and Yeo-Johnson were tested. Among these, Box-Cox Transformation was selected as it showed better consistency with K-Fold validation, while the model without any transformation failed due to inconsistent distribution between the training and test samples.

```
=== No Transformation ===
No Transformation - Shape after PCA: (500, 22)
No Transformation - Individual R² scores: [ 0.80265807  0.41873321 -0.01060036  0.59398903  0.9440268 ]
No Transformation - Mean R² score: 0.5498

=== Log1p Transformation ===
Log1p Transformation - Shape after PCA: (500, 22)
Log1p Transformation - Individual R² scores: [0.14029208 0.20082422 0.22459162 0.2465048  0.15132698]
Log1p Transformation - Mean R² score: 0.1927

=== Box-Cox Transformation ===
Box-Cox Transformation - Shape after PCA: (500, 22)
Box-Cox Transformation - Individual R² scores: [0.28518838 0.27815486 0.32231199 0.38652639 0.31050432]
Box-Cox Transformation - Mean R² score: 0.3165

=== Yeo-Johnson Transformation ===
Yeo-Johnson Transformation - Shape after PCA: (500, 22)
Yeo-Johnson Transformation - Individual R² scores: [0.28401241 0.27260512 0.31355155 0.38993857 0.30880839]
Yeo-Johnson Transformation - Mean R² score: 0.3138
```

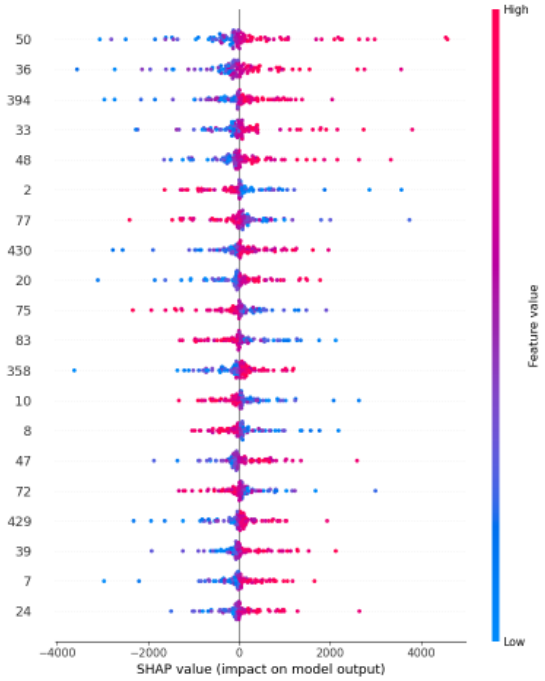
Model Training and Evaluation

Random Forest and XGBoost models were trained using RandomizedSearchCV for hyperparameter tuning. However, both models showed low R^2 and high MAE and MSE values, indicating poor performance. On the other hand, a deep learning model performed significantly better after tuning with Optuna for hyperparameter optimization. A performance comparison graph clearly showed the superiority of the deep learning model.



Model Interpretation with SHAP

SHAP (SHapley Additive exPlanations) was used to identify the most influential features affecting the model's predictions. Features with a wider spread of SHAP values showed higher variability in their impact. Positive SHAP values increased the prediction, while negative values lowered it.



Model Deployment

The best-performing model was selected for deployment. It was containerized using Docker and deployed to Azure for serving predictions.

The image below demonstrates how the features are uploaded as a JSON payload to the local host, and how the model returns predictions in response. The model is accessed through a FastAPI endpoint running on the local host, which processes the input features and generates predictions based on the trained model.