Crop Yield Prediction with **XYieldBoost**

A machine learning model for improving precision agriculture in Bihar, India

Oxford International Development Group

Oxford International Development Group

We are a team of researchers using machine learning to advance international development

Jessica Rapson

Jessica has a Master of Public Policy from the University of Toronto and a Master of Statistical Science from the University of Oxford. Her work involves developing machine learning tools for governments.

Shaw Chifamba

Shaw is a Machine Learning Solutions Architect for UNECE working to improve agricultural supply chains using ML. He is also studying for a Master in Precision Cancer Medicine at the University of Oxford.

Juliette Zaccour

Juliette is a Social Data Science DPhil student at the University of Oxford with a background in human-centred data science. Her research focuses on algorithmic justice and auditing methods for public sector systems.

Digital Green

Problem:

Smallholder farmers in Bihar, India struggle with resource constraints and unpredictable weather that limits food security. Developing a machine learning (ML) model to accurately predict crop yield can empower these farmers to make informed agricultural decisions, reducing poverty and malnutrition.

Introducing XYieldBoost

XYieldBoost *leverages agricultural knowledge and tree-based ML to improve crop yield predictions*

Easy to Implement and Run

Smart feature engineering is directly applied to Digital Green survey data, enabling **fast training** and **program integration**

Improves Crop Yield Prediction Accuracy

Prediction generated by XYieldBoost **reduce crop yield prediction error by 72%** compared to benchmark estimates

Provides Insights for Crop Yield Improvements

Estimates of **variable importance** and quantifications of **variable impacts** for specific predictions enables actionable insights on improving yield

Digital Green

Bihari Agricultural Context

XYieldBoost *operates by leveraging domain knowledge about agricultural practices in Bihar, India*

Bihar-Specific Agricultural Practices

The Ahar Pyne System

South Bihar is more agricultural productive than North Bihar. It employs the Ahar Pyne agricultural system, using channels and retention ponds to manage water resources and adapt to Bihar's unpredictable weather

Monsoon Cycles

Kharif crops, such as rice, are sown during the monsoon season from June to September and are watered by monsoon rainfall. These crops do well with high rain in winter. Rabi crops, such as wheat, are sown in mid-November after the monsoon.

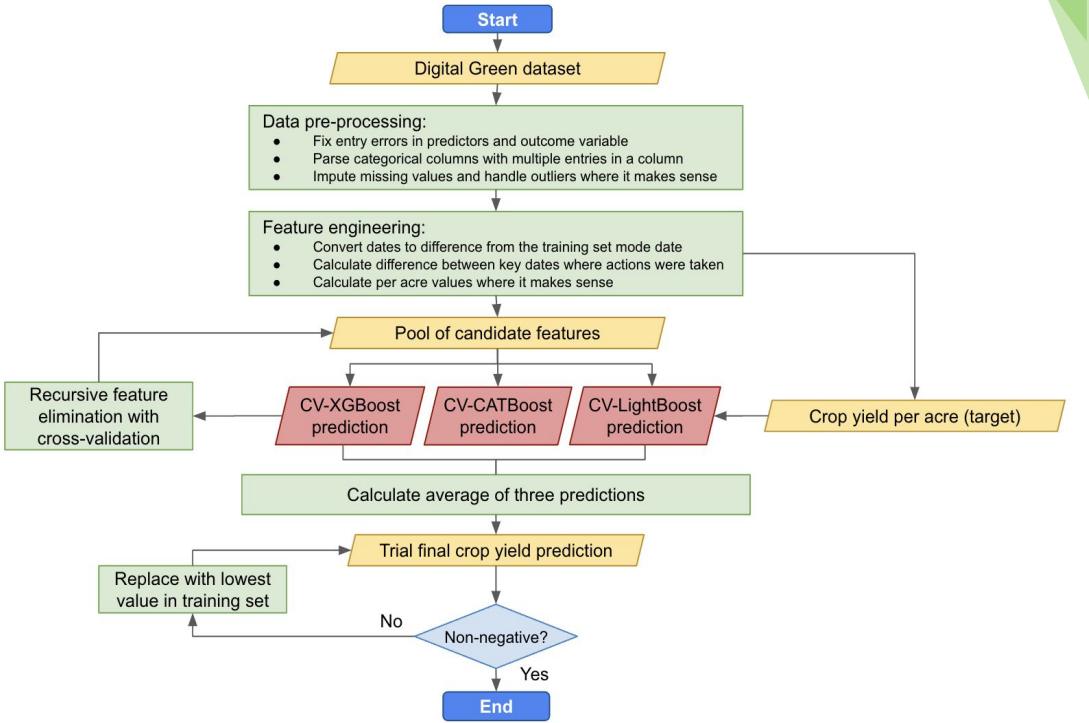
General Agricultural Practices

Domain knowledge of nitrogen cycles, fertilizer application methods, and irrigation techniques was also applied

XYieldBoost Architecture

Feature were engineered to reflect the importance of regionality and the monsoon cycle to crop yield

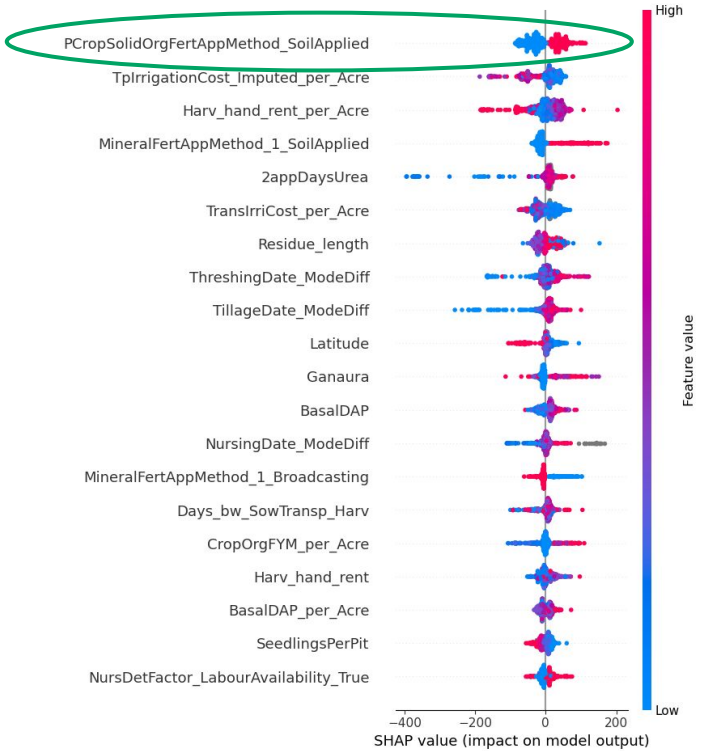
- We used an **ensemble method** that took the average of three predictions made using tree-based methods, specifically:
  - **XGBoost**,
  - **CatBoost**,
  - and **LightGBM**
- These are all **tree-based methods**, which are ideal for handling tabular data with non-linear relationships and missing values
- **Recursive feature elimination with cross-validation** was used to select the final features for our model
- We also used **cross-validation** to train and test our model to reduce overfitting



Digital Green

XYieldBoost Variable Importance

Additional steps were taken to ensure Digital Green can obtain impactful insights from the results



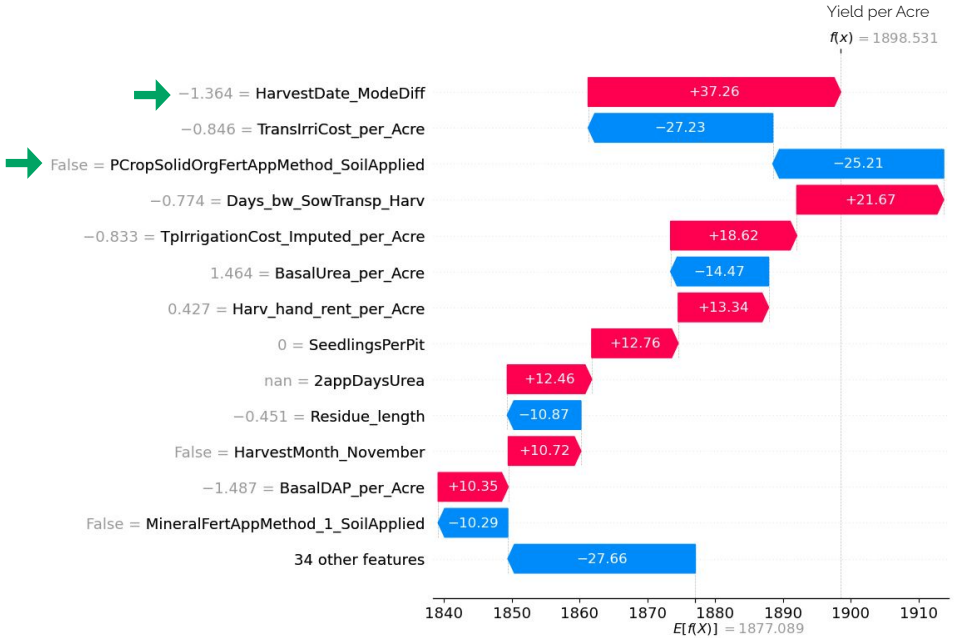
- We calculated **Shapley values** – the average expected marginal contribution of one variable after all possible combinations have been considered – for each feature
- These can be seen similarly to the **variable importance** indicators that can be extracted from random forest models
- This information can be leveraged to improve crop yield for Bihar farmers; for example, as seen in the plot, **applying organic fertiliser directly to soil in the previous crop cycle was consistently associated with higher yield per acre**
- Using this insight, Digital Green can **spread the information to farmers**, or possibly even **invest in directly providing fertilisation tools that enable direct soil application**

Digital Green

XYieldBoost Explainability Insights

Digital Green can also determine exactly which factors drive yield predictions for a given plot of land

- The plot shows which variables influenced the prediction for a **0.375 acre farm in the small town of Noorsarai in the Nalanda district**
- This farmer harvested his crop on **October 20th, 2022** – earlier than other farmers
- Typically, Kharif crops are harvested at the end of the monsoon; however, the 2022 monsoon was drier than normal, meaning that an early harvest **may have prevented crops from drying out from the lack of accumulated rain**
- As can be seen in the plot, **harvesting earlier was associated with an approximate increased yield of 37 units per acre**
- However, **not applying organic fertilizer to soil in the previous crop cycle reduced yield by 25 units per acre**, offsetting this gain



Digital Green

Thank You

*We are happy to answer any additional questions to ensure **XYieldBoost** can have an impact on reducing poverty and malnutrition in Bihar.*

Please contact jess.rapson@mail.utoronto.ca