# Digital Green Crop Yield Estimate Project

DTSA 5509

November 10, 2024

# Data Challenge Introduction

The contest is sponsored by Digital Green (https://digitalgreen.org/).

The data challenge is to create a machine learning solution to predict the crop yield per acre of rice or wheat crops in India.

The data was collected through a survey conducted across multiple regions in India, which consists a variety of features, e.g., the type and amount of fertilizers used, the quantity of seedlings planted, irrigation methods, and etc.

- 5000 data records
- 43 features
- training/test data ratio: 75:25

The data challenge is available: *https://zindi.africa/competitions/digital-green-crop-yield-estimate-challenge*

# Analysis Proposal

1. Predict the crop yields per acre using different models
   - Linear Regression
   - Tree-based Regression:

     *Extra Trees Algorithm*

     *Lightweight Gradient Boosting Machine (LightGBM)*

     (and Ensemble model: the mixture of Extra Tree and LightGBM)

2. Evaluate the root mean square error (RMSE) across different models

# Exploartory Data Analysis

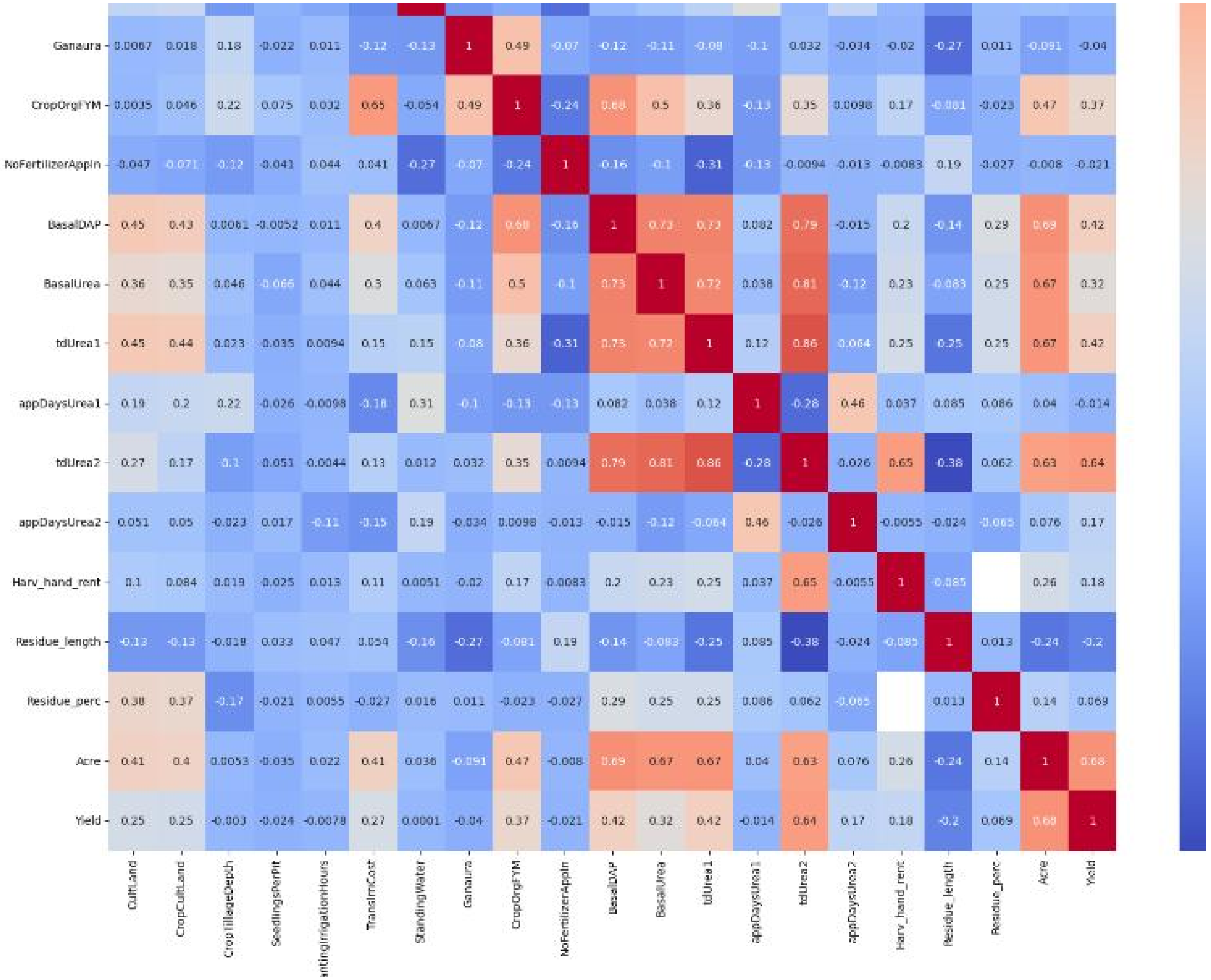
**Response : Yield**
**Factors : all other variables collected**

**Data Cleaning -** check collinearity

- Dropped the categorical variables, focus on the effects from numerical variables.

- Removed the highly correlated numerical variables

- Deleted the records with missing Yield value

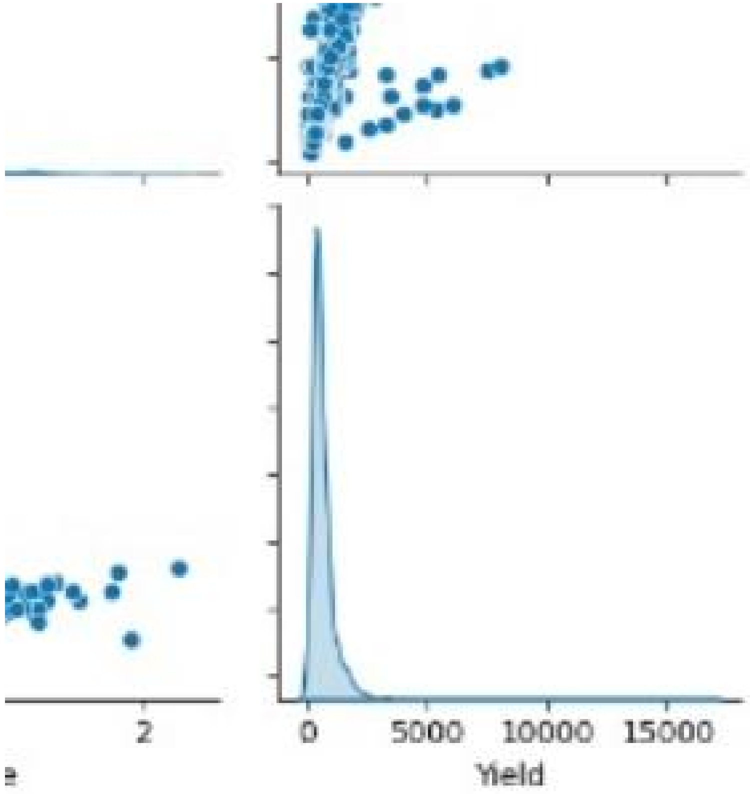


# Exploartory Data Analysis

### Response Variable Derivation

- Log transformed Yield as it is left-skewed distributed
- Derived new Yield variable :

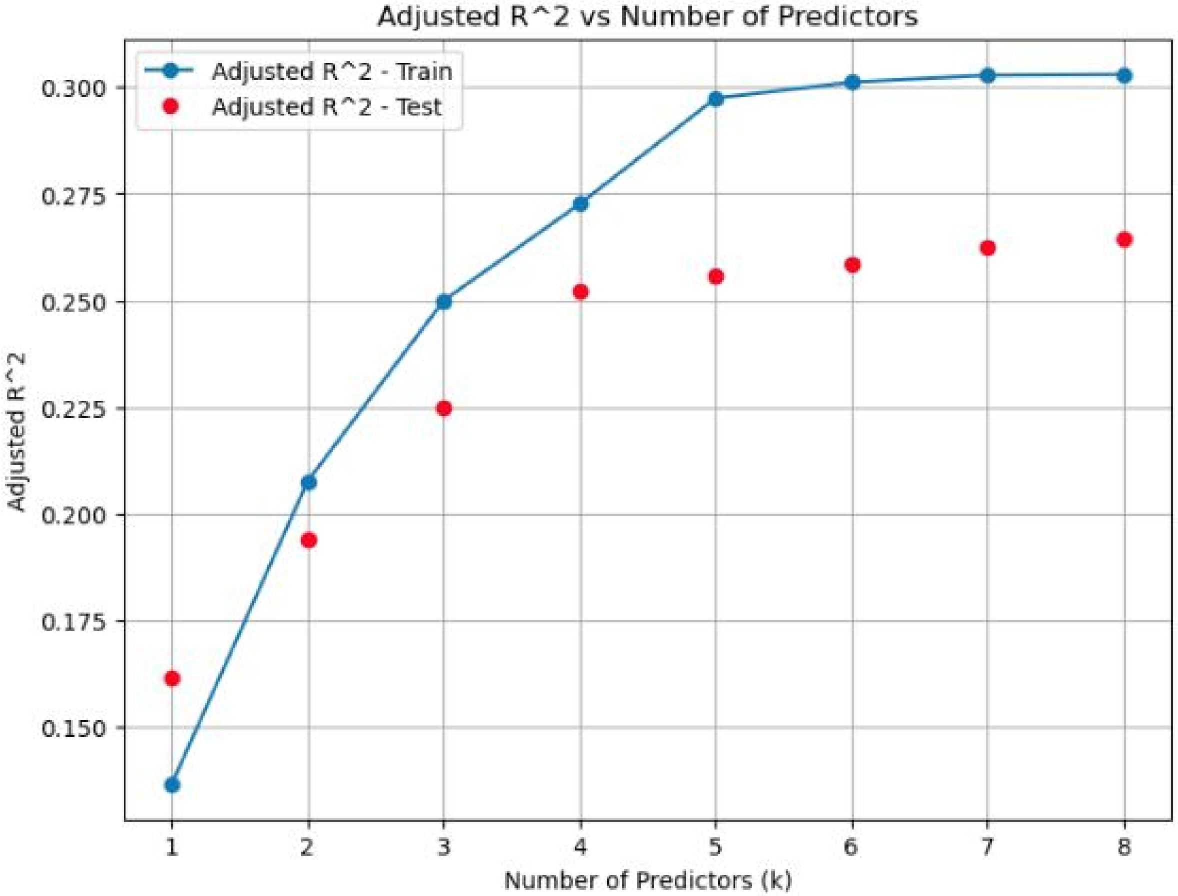
Yield per acre = Yield / Acre



# Outputs - Multiple Linear Regression

- Forward feature selection
- Good prediction with 5 or 6 features.
- Final Model:

| Variable Name | Variable Description |
|---|---|
| BasalDAP | Amount of DAP(in kgs)applied during land preparation |
| CropCultLand | Area of land under cultivation |
| TransIrriCost | Cost of irrigation during transplantation |
| Residue_length | Length of the residue left after harvesting |
| Harv_hand_rent | If labours were used or harvesting machine hired, what was the rent (in rupees) |
| TransIrriCost | Cost of irrigation during transplantation |
| CropOrgFYM | Amount of FYM (Farm yard manure) organic fertilizer used (in Quintals) |
| CropTillageDepth | Depth of the tillage |

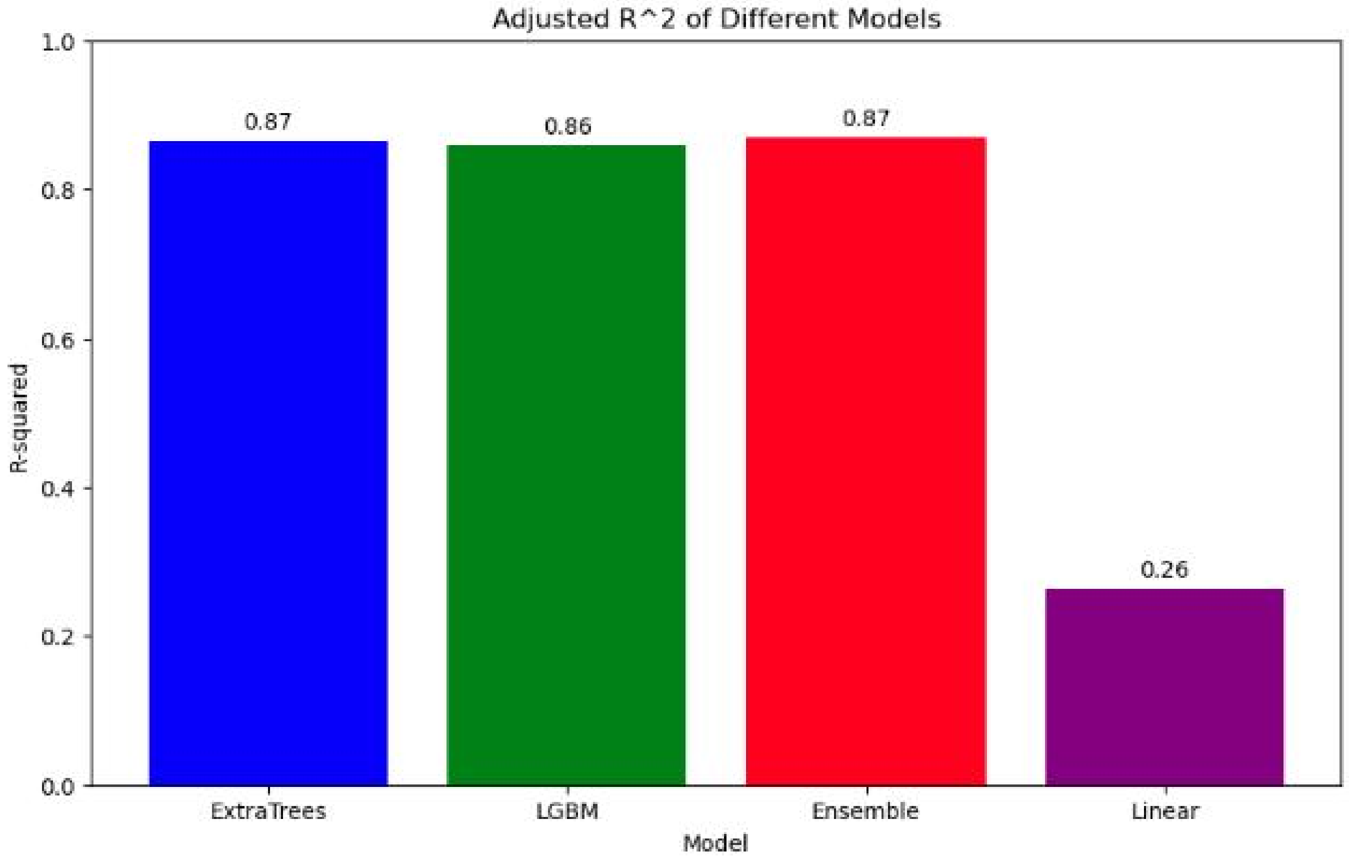


# Outputs - Tree-based Models

Compare the model performance using adjusted R^2 using testing datat

- Extra Tree
- LightGBM
- Ensemble model is the average of Extra Tree and LightGBM modele

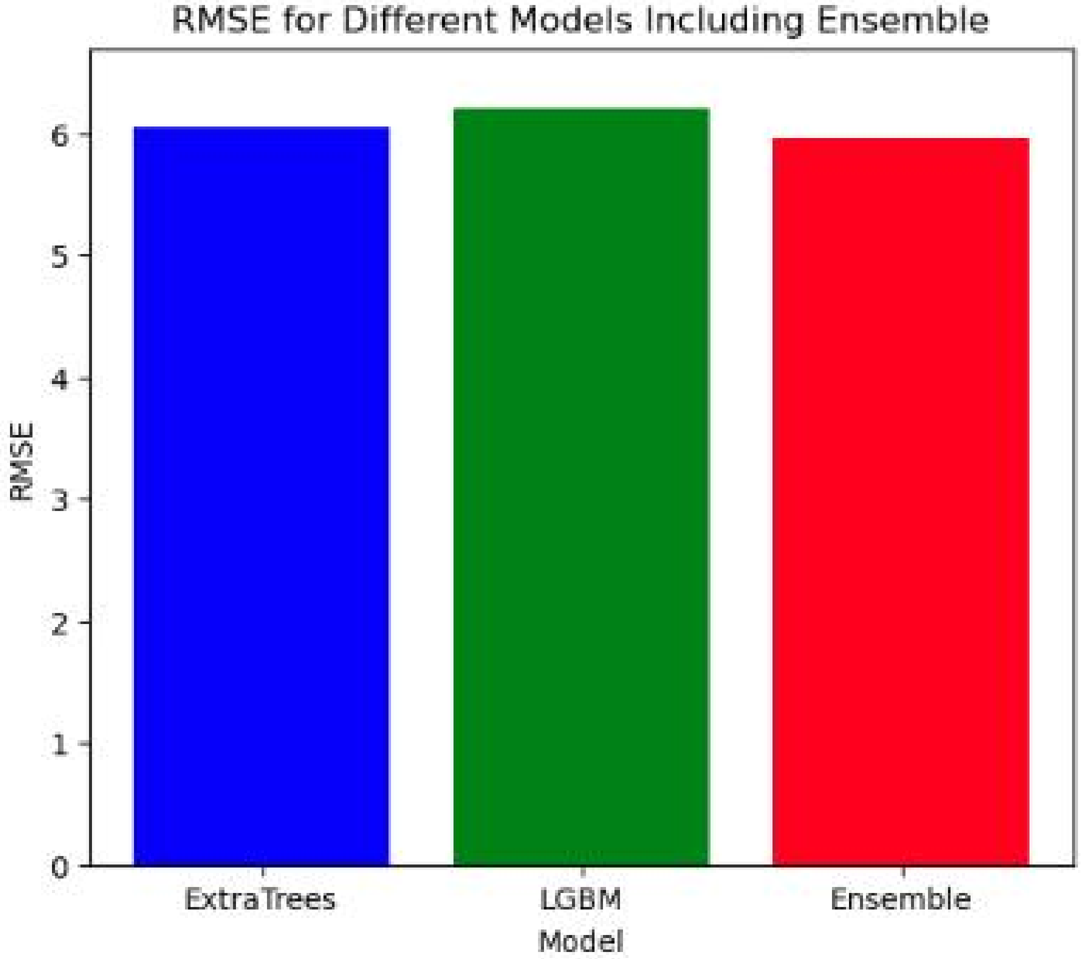
All the three tree-based models have much higher adjusted R^2 values.



# Outputs - Tree-based Models

Comparison of RMSE

The three tree-based models have similar RMSE.



# Conclusion and Discussion

- In the final prediction model, I select the ensemble model, which is a mixture of Extra Tree and LightGBM

- The multiple linear regression is not recommended for real-world data, but could be a good practice for exploratory analysis.

- Limitations:

  Categorical features are not taken into consideration.

  Overfitting vs. Underfitting

# Resources

1. Extra Tree and LightGBM Coding Resources

*https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html*

*https://lightgbm.readthedocs.io/en/stable/Python-Intro.html*

2. Data Source

*https://zindi.africa/competitions/digital-green-crop-yield-estimate-challenge*

3. My GitHub

*https://github.com/lzheng01/Digital-Green-Crop-Yield-Estimate-Project*

# Thank You!