Crop Yield Prediction and Profit Optimization using Regression and Linear Programming

Harsh Sinha (IMT2023571)
Kartikeya Dimri (IMT2023126)
Syed Naveed (IMT2023119)
International Institute of Information Technology, Bangalore

September 25, 2025

1 Problem Statement

Agriculture is the backbone of India's economy, yet farmers often face uncertainty about which crop to grow, when to grow it, and how much profit they can expect. Traditional practices rely on experience rather than data-driven insights, which may not always be optimal in terms of yield or profitability.

The goal of this project is to build a data-driven decision support system for farmers. Using historical crop production data, rainfall, and input usage, we aim to:

1. Predict crop yield in a specific state using regression models.
2. Estimate profitability by integrating predicted yield with predicted crop prices (using historical MSP values).
3. Optimize crop selection and allocation using linear programming so that farmers maximize profit under resource constraints.

This combines predictive analytics (regression) with prescriptive analytics (optimization), providing a real-world decision support tool.

2 About the Dataset

We use crop production data from government sources, which is state-wise and annual. Each record corresponds to one crop grown in a given year, along with associated features.

Features available

- **CropName** – Name of the crop (e.g., Rice, Wheat, Arecanut).
- **CropYear** – Year of cultivation.
- **Season** – Kharif, Rabi, or Whole Year.

- **State** – State where crop was cultivated (e.g., Karnataka).
- **Area (ha)** – Land under cultivation (hectares).
- **Production (metric ton)** – Total production.
- **Annual Rainfall (mm)** – Recorded rainfall.
- **Fertilizers (kg)** – Fertilizer usage.
- **Pesticides (kg)** – Pesticide usage.
- **Yield (ton/ha)** – Production per unit area (target variable for prediction).

Sample Record

Crop Name	Crop Year	Season	State	Area (ha)	Production (ton)	Annual Rainfall (mm)	Fertilizers (kg)	Pesticides (kg)	Yield (ton/ha)
Arecanut	1997	Whole Year	Karnataka	93100	133342	1266.7	8860327	28861	1.29

3 Input Features and Output Labels

- **Input Features (X):** Season, Area, Rainfall, Fertilizer, Pesticides, CropYear, Crop Name, State.
- **Output Label (y):** Yield (production per hectare).

This makes yield the central prediction target, since it reflects efficiency of cultivation and helps compare crops fairly across varying land sizes.

4 Methodology

4.1 Easy Regression

We begin with basic models to establish benchmarks:

Ordinary Least Squares (OLS)

OLS minimizes the sum of squared residuals:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2$$

Polynomial Regression

An extension of OLS by adding polynomial terms of features:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d + \epsilon$$

4.2 Advanced Regression

Ridge Regression (L2 Regularization)

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \|\beta\|_2^2 \right)$$

Lasso Regression (L1 Regularization)

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \|\beta\|_1 \right)$$

Elastic Net

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right)$$

XGBoost Regression (Linear Booster)

XGBoost optimizes a regularized objective function:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k} \Omega(f_k)$$

where l is the loss (squared error for regression) and $\Omega(f_k)$ is the regularization term:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2$$

Here T is the number of leaves in the tree, ω are leaf weights, and λ, γ are regularization parameters. In *linear booster* mode, XGBoost reduces to a highly optimized regularized OLS solver.

5 Price Prediction

To calculate profit, we also need crop price. This will be estimated using historical Minimum Support Price (MSP) data with another regression model. While this is a separate task, we note that predicted prices will be integrated into the optimization stage.

MSP Dataset

The MSP dataset contains information for each crop in a given year and season:

- **Year** – Agricultural year (e.g., 2022–2023).
- **Crop** – Name of the crop (e.g., Bajra).
- **Season** – Kharif, Rabi, or other.
- **MSP** – Minimum Support Price (Rs/quintal).

For example, one entry from the dataset is:

Year	Crop	Season	MSP (Rs/quintal)
2022–2023	Bajra	Kharif	2350.0

In this case, the target variable y is the MSP of the crop. By combining yield prediction with MSP prediction, we can estimate potential revenue for farmers more accurately.

6 Optimization Using Linear Programming

After predicting yield and price, we solve an optimization problem. Linear Programming (LP) is the mathematical method used to solve this optimization problem. It is perfectly suited for this task because both the objective function (total profit) and the constraints (land, resources) are linear relationships.

Objective Function

$$\text{Profit} = \sum_i \Big(\text{PredictedYield}_i \times \text{PredictedPrice}_i \times \text{LandAllocated}_i - (\text{FertilizerCost}_i + \text{PesticideCost}_i + \text{MiscCost}_i) \Big)$$

where i = crop index and costs include fertilizer, pesticide, and optional miscellaneous inputs given by the farmer. Here, the decision variable that the model needs to determine is land allocated for each crop.

Constraints

- **Total Land Constraint:**

$$\sum_i \text{LandAllocated}_i \leq \text{Farmer's Land}$$

- **Fertilizer Constraint:**

$$\sum_i \text{FertilizerUsage}_i \leq \text{Available Fertilizer}$$

- **Pesticide Constraint:**

$$\sum_i \text{PesticideUsage}_i \leq \text{Available Pesticide}$$

- **Seasonal Constraint:** Each crop can only be allocated to its valid season (Rabi, Kharif, Whole Year).

7 Real-Life Application

The system works as follows:

1. Farmer inputs: state, land size, available fertilizers/pesticides, input prices (fertilizers, pesticides, miscellaneous).
2. Model predicts yields for all crops possible in that state.
3. Price model predicts MSP-based crop prices.
4. Linear programming optimizes crop allocation to maximize total profit.
5. Output: Recommendation for Rabi, Kharif, and Whole-year crops with expected profit.

8 Impact

- **For Farmers:** Provides clear, data-driven crop planning strategies, maximizing income.
- **For Policymakers:** Helps in identifying profitable crops per region, guiding subsidy and MSP policies.
- **For Researchers:** Combines predictive and prescriptive analytics for agriculture, a step beyond pure regression models.

By integrating machine learning with optimization, this project provides a real-world decision support tool that can directly impact farmer welfare.