TRILYTICS'25: ANALYTICS CASE COMPETITION

- L&T Finance in collaboration with PGDBA CONCLAVE

From Field To Finance: Smart AgriForecasting

Team Name: Triolysis
Institution Name: Ajeenkya DY Patil School of Engineering, Pune-47
Date: 27 July 2025

Flow Of Presentation

- Introduction
- Problem Statement/Objective
- Dataset description
- System Architecture
- Data Flow Diagram
- Data Visualization
- Future Scope
- Conclusion

Problem statement

- India’s agricultural sector is the backbone of its economy, supporting millions of households. Yet, farmers continue to face uncertainty when it comes to their annual income. Due to limited or non-existent credit histories, many farmers struggle to secure loans from trusted financial institutions, often falling prey to exploitative lenders. Despite advancements, deserving loan applicants are sometimes rejected, simply because the existing systems fail to capture their true financial potential.

Objective

- The goal is to build a machine learning model that can predict a farmer’s annual income based on relevant factors like landholding size, crop type, crop yield, commodity price dynamics, and regional attributes. By leveraging data driven insights, this model aims to improve income transparency, support targeted interventions, and empower stakeholders across the agricultural ecosystem with actionable foresight.

DATASET DESCRIPTION

Primary Dataset	**External Dataset 1 – Agmarknet (Crop Price Data)**
Dependent Variable: Total Income **Independent Variables**: • *Demographic*: Gender, marital status, region, ownership • *Agricultural*: Landholding, cropping density, irrigated area, soil type • *Climate*: Rainfall, temperature, groundwater levels • *Socio-economic*: Electricity, sanitation, house type, night light index • *Indices*: Village Agri Score, Socio-Economic Score, Land Holding Index • *Market Access*: Distance to mandi/railway • *Granularity*: Farmer-level mapped with village-level indicators	**Columns**: District, Crop, Season, Modal_Price Captures seasonal price trends from mandis **Merge Key**: District + Season
External Dataset 2 – Crop Yield Statistics	**External Dataset 3 – Rainfall & Weather Data**
Columns: Wheat yield, maize yield, soybean area, etc. District-level productivity metrics Supports estimation of income using yield × price relationship	**Columns**: Seasonal Rainfall, Rainfall Anomaly, Temperature (min/max) Models impact of weather on yield **Merge Key**: District + Seasonal Rainfall

SYSTEM ARCHITECTURE

Model Training & Validation

Purpose: Learn patterns between features and income using supervised regression.

- Models: XGBoost, LightGBM, CatBoost, Random Forest

Feature Encoding & Selection

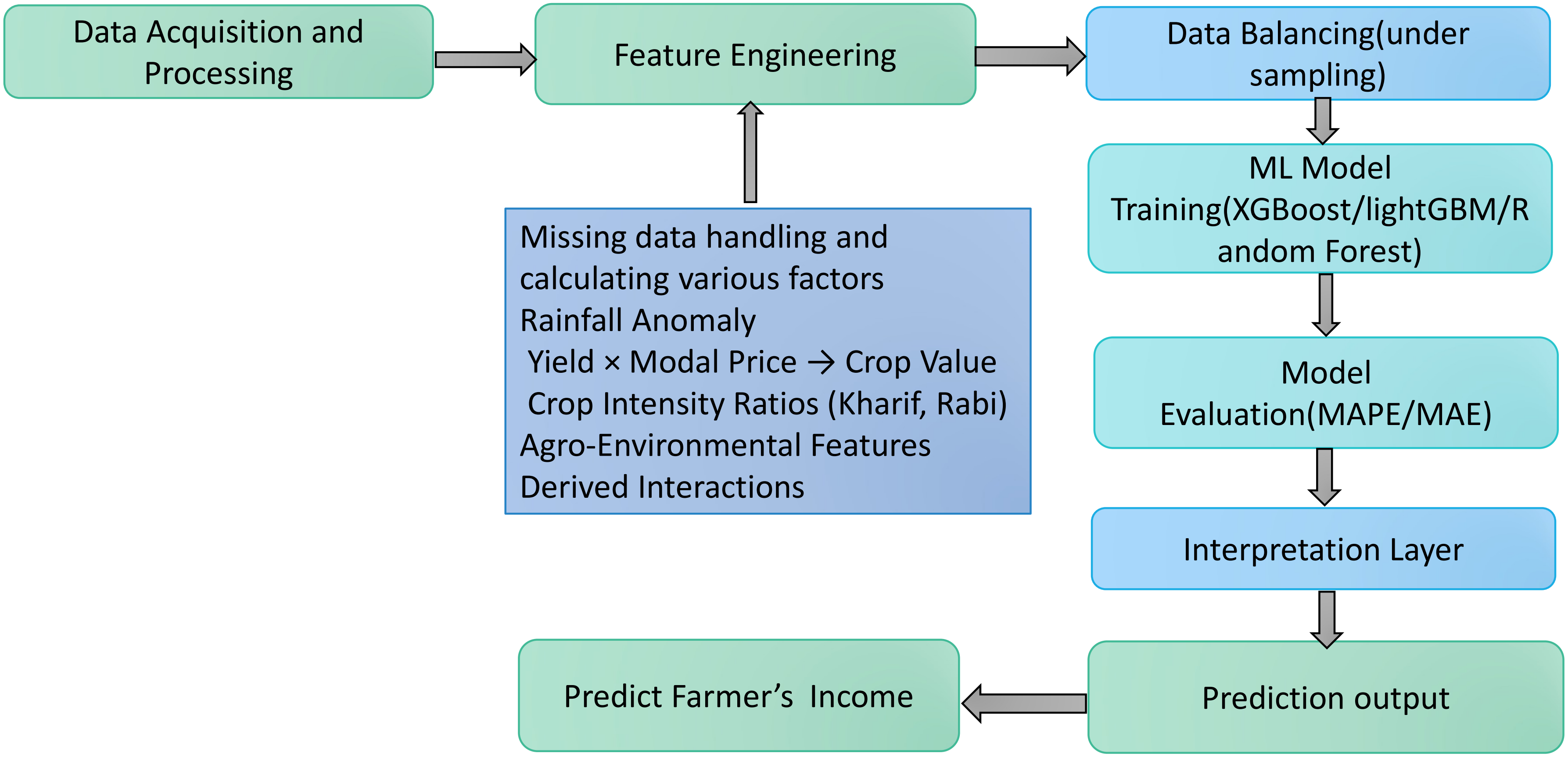
- Label Encoding for categorical variables
- XGBoost used to drop low-impact features

Training Details

- Log-transform of income to reduce skew
- 10-Fold Cross-Validation
- Metrics Used: MAPE, RMSE, MAE, R²

Model	MAPE	R^2	RMSE
Optuna-Tuned XGBoost (final)	20.53	0.7446	INR.7,95,271.55
Stratified XGBoost (10-Fold CV)	20.72	0.7492	INR.10,38,634.2
XGBoost	20.94	0.7806	INR.9,42,894.45
LightGBM	21.17	0.8947	INR.6,58,349.40
CatBoost	21.36	0.8439	INR.8,06,115.47
Random Forest	23.00	0.715	INR.11,80,000

Data Flow Diagram


Data Acquisition and Processing
Feature Engineering
Data Balancing(under sampling)
Missing data handling and calculating various factors
Rainfall Anomaly
Yield × Modal Price → Crop Value
Crop Intensity Ratios (Kharif, Rabi)
Agro-Environmental Features
Derived Interactions
ML Model Training(XGBoost/lightGBM/Random Forest)
Model Evaluation(MAPE/MAE)
Interpretation Layer
Prediction output
Predict Farmer’s Income

DATA VISUALIZATION

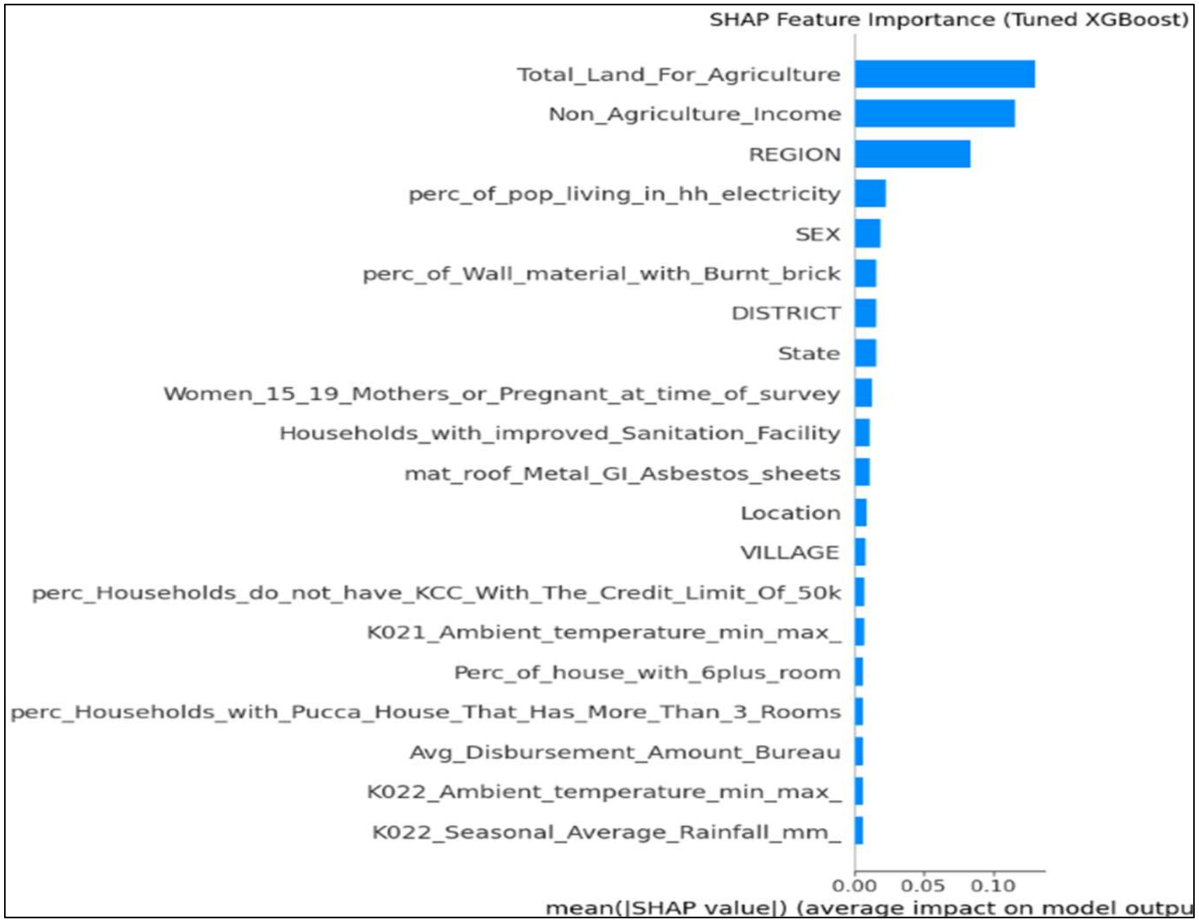


Fig.1.

SHAP Feature Bar Plot

- **Top Features**:
  - Total_Land_For_Agriculture: Most influential on income prediction.
  - Non_Agriculture_Income, REGION, and Electricity access follow.
- **Insight**: Agricultural land size and additional non-farm income significantly affect total income.
- **Interpretation**: These features should be prioritized for policymaking and model refinement.

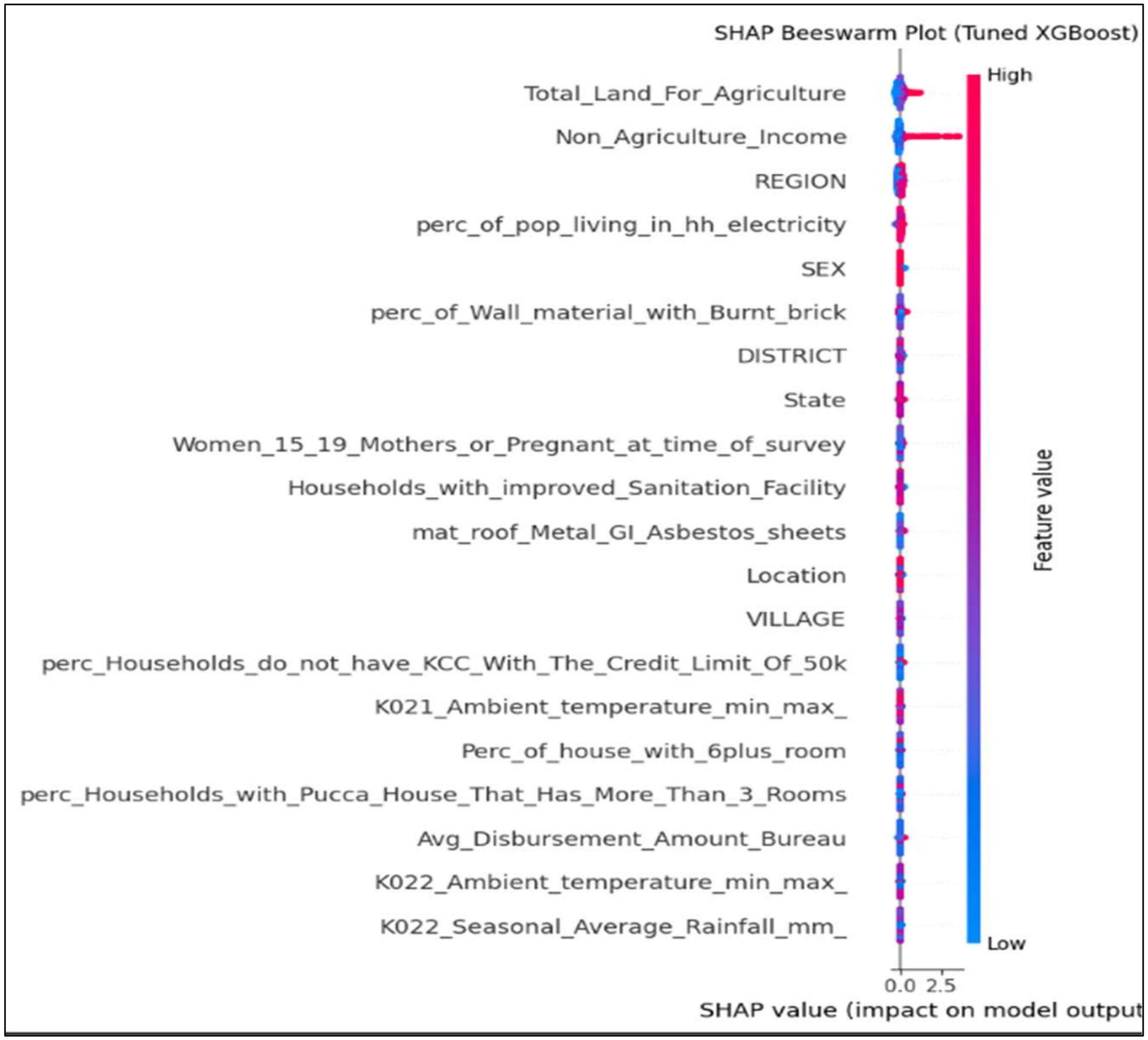


Fig.2.

SHAP Beeswarm Plot

- **Description**: Shows how each feature's value (high or low) impacts predictions.

- **Observations**:
  - High land area : strong positive SHAP values = higher predicted income.
  - REGION and Non_Agriculture_Income also drive income significantly.
  - Low sanitation or poor roofing materials slightly reduce predictions.

- **Insight**: Income isn't just land-based, regional and household conditions matter.

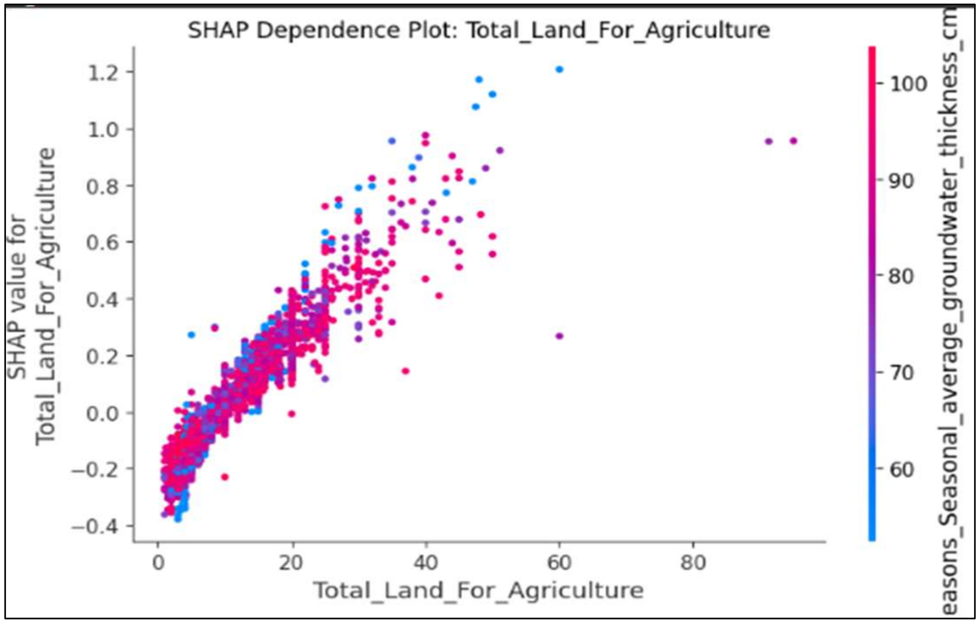


Fig.3.

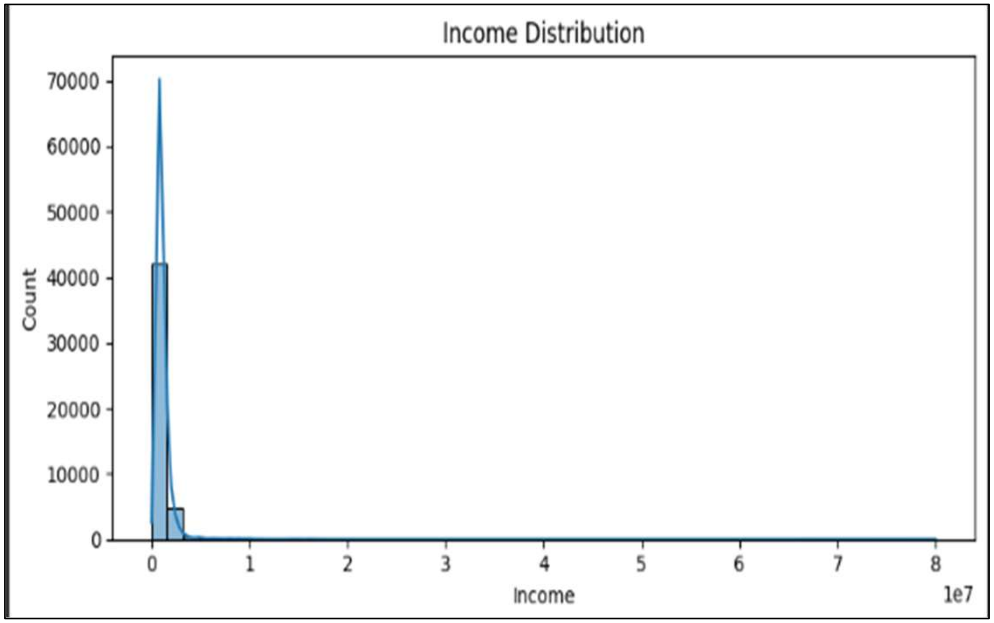


Fig.4.

SHAP Dependence Plot – Total_Land_For_Agriculture

- **Insight**: As land size increases, SHAP value increases linearly.
- **Interaction**: Colored by groundwater thickness deeper water tables slightly amplify land's effect.
- **Conclusion**: More land consistently leads to higher income, especially with better water access.

Income Distribution

- **Observation**: Highly right-skewed distribution.
- **Insight**: A few farmers earn significantly more than the average model is robust to outliers.
- **Action**: Justified use of log-transformation for model stability.

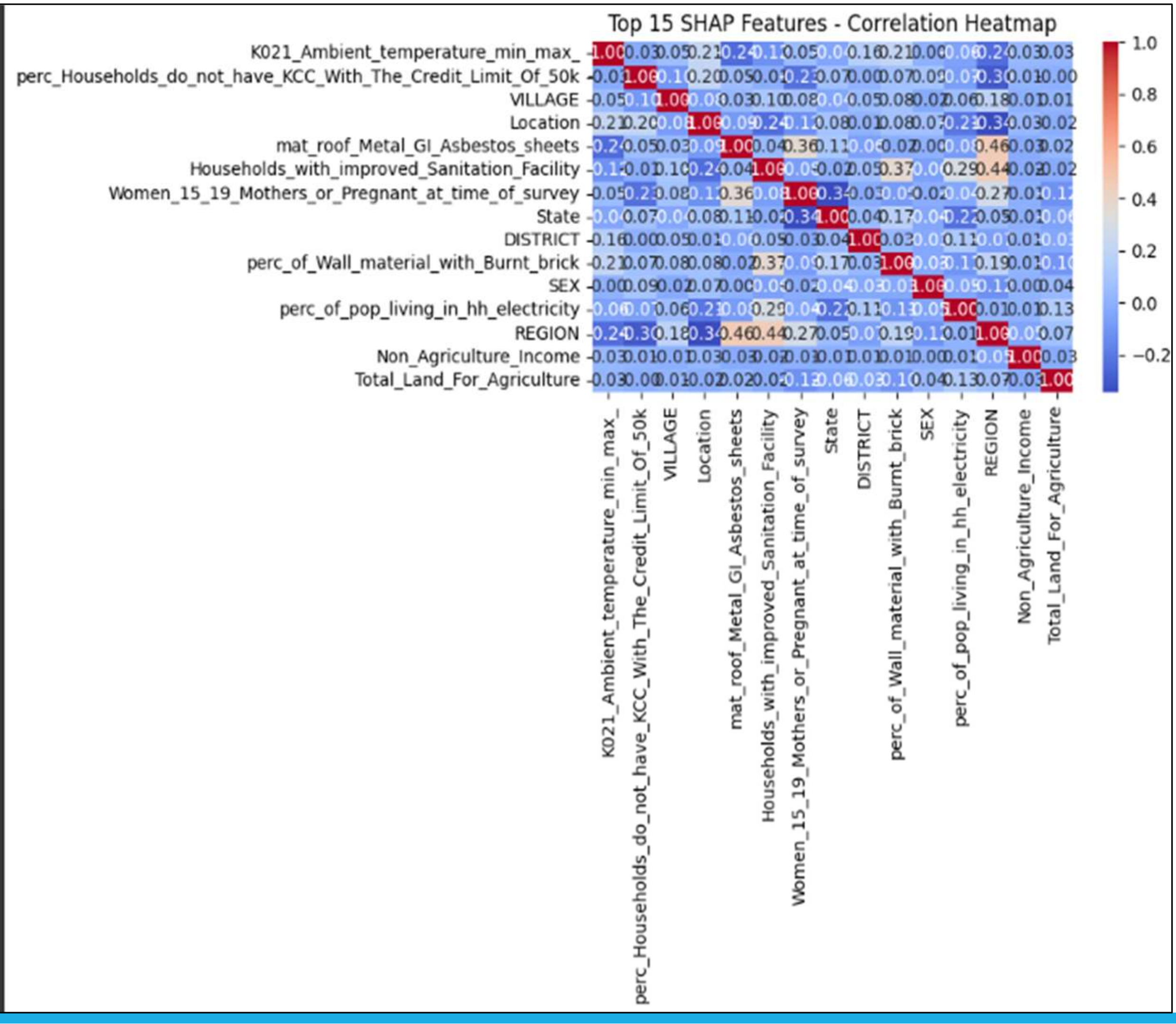


Fig.5.

Top 15 SHAP Features – Correlation Heatmap

- **Observation**: Most top features have **low inter-correlation** (blue = weak).
- Conclusion: Model benefits from **diverse, uncorrelated features**.
- **Insight**: No severe multicollinearity

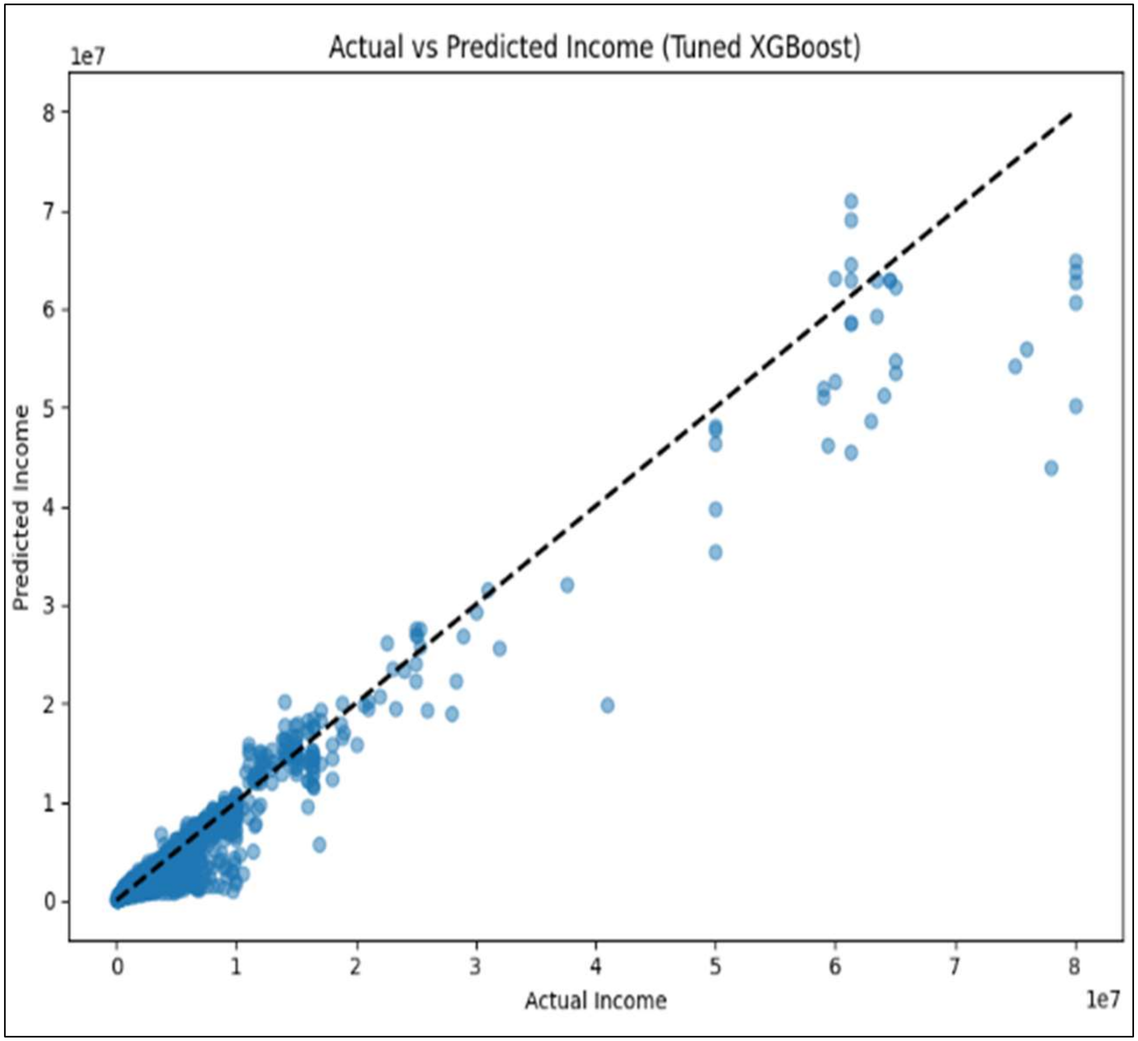


Fig.6.

Actual vs Predicted Income Plot

- **Observation**: Most points lie near the diagonal (ideal line).
- **Interpretation**: The model performs well with minor overprediction at extreme values.
- **Conclusion**: Good generalization as minimal bias and solid fit.

Future Scope

- **Farmer-Facing Platforms**

Developing user-friendly mobile/web tools to help farmers estimate future income and integrate advisory services based on predicted risk or opportunity.

- **Incorporation of Real-Time Data**

Enabling continuous income updates and real-time credit risk monitoring.

- **Integration with Government Portals , Schemes and Startups**

Sync with PM-KISAN, eNAM, and AgriStack to enrich data and deployment reach. And collaboration with drone, IoT, and sensor-based startups to enhance feature richness.

- **Automated Loan Decision Systems**

Embedding the model into digital lending platforms for real-time, bias-free loan decisions.

Conclusion

- Predicting farmer income using machine learning model and optimization offers a scalable and impactful solution to **bridge gaps in rural credit access**.

- By leveraging historical and alternative data sources such as crop type, weather, etc. models can estimate income with **greater accuracy** and reliability. This approach enhances **transparency**, supports fair lending, and reduces dependence on traditional credit history.

- The model's predictions are **interpretable and aligned with real-world rural dynamics**.
- **Land ownership, non-farm income, and regional factors** are the biggest levers to improve farmer livelihoods.

- Features like **infrastructure and gender** matter more than expected and suggest **policy opportunities beyond just agriculture.**

- Overall, farmer income prediction presents a **transformative opportunity** to promote financial inclusion, rural development, and sustainable growth through intelligent data use.