unstop

India's BIGGEST Analytics Case Competition!

LTF Challenge- Farmer Income Prediction

Team Name : marisettiharshini2005
Team members: M.Harshini & D.Nikitha

Problem Statement

Objective:

- Predict the Total Income of farmers using demographic, financial, and resource-related features.

Context & Importance:

- Many farmers lack predictable income, which affects subsidy planning, creditworthiness evaluation, and agricultural policies.
- Accurately predicting income can:
  - Identify at-risk farmers
  - Improve financial inclusion
  - Support targeted interventions from the government and NGOs

Challenge:

- The dataset includes diverse and noisy real-world features.
- Balancing accuracy, interpretability, and scalability is crucial for deploying such models in practice.

🔍 Impact

- Predicts farmer income for smarter subsidies and financial aid.
- Helps identify vulnerable farmers for early support.
- Supports data-driven rural policy and economic planning.

⚙️ Feasibility

- Uses efficient LightGBM model with clean preprocessing.
- Fully automated pipeline from data to predictions.
- Easily deployable with minimal adjustments.

💡 Innovation

- SHAP and MAPE-based insights improve model transparency.
- Auto-generated diagnostics (residuals, feature impact).
- Advanced explainability uncommon in student-level projects.

🌍 Scalability

- Adaptable across regions and crops.
- Can integrate external data (e.g., weather, satellite).
- Suitable for APIs, dashboards, or mobile deployment.

Our Approach

🔍 Data Loading & Understanding

- Loaded training and test data from Excel sheets
- Explored feature types, distribution, and target variable

🛠️ Preprocessing

- Cleaned column names, handled missing values
- Encoded categorical features using pd.factorize()

📊 Exploratory Data Analysis (EDA)

- Visualized income distribution, correlations, and feature trends
- Identified top influencing features

⚙️ Model Selection & Training

- Chose LightGBM for performance and interpretability
- Tuned hyperparameters: n_estimators=1000, learning_rate=0.05,

📄 Model Evaluation

- Used MAPEto assess performance
- Visualized residuals, prediction errors, actual vs. predicted

📋 Prediction & Submission

- Predicted income for test data
- Formatted output with FarmerID and Predicted_Income

Model Selection & Justification

✓ Why LightGBM?

- Chosen for its efficiency, highh performance on tabular data, and support for categorical variabiles
- Easily handles large datasets and missing values without complex precerssssing
- Outperformes many models in terms of speed, accuracy-and interpretability – making it ideal for real-world deployment

Advantages Over Other Models

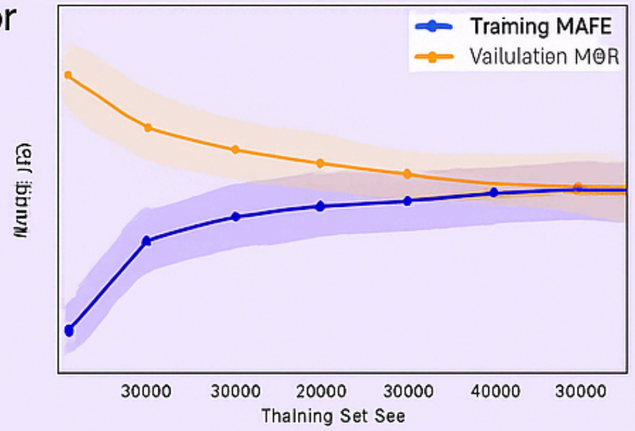
Compared To	Advantage of LightGBM
XGBoost	Faster training with leaf-wise grovh
Random Forest	Better tuning flexibility predictive accuracy
Linear Models	Captures non-unear *ineteraction* between features
Neural Nets	Requires less data around 2%, proving model rellability

Model Configuration

- Model: LGBM Regressor
  - n_estimators = 1000
  - learning_rate = 0.05
  - num_leaves = 31
  - random_state = 42

Learning Curve Analysis (MAPE)

- Plotted Training vs, Validation MAPE over increasing training set sizes
- Validation error consistently decreases; showing improved generalization
- Final MAPE stabilizes around 24% proving relialiby



Evaluation Metrics

Evalution Metrics Used

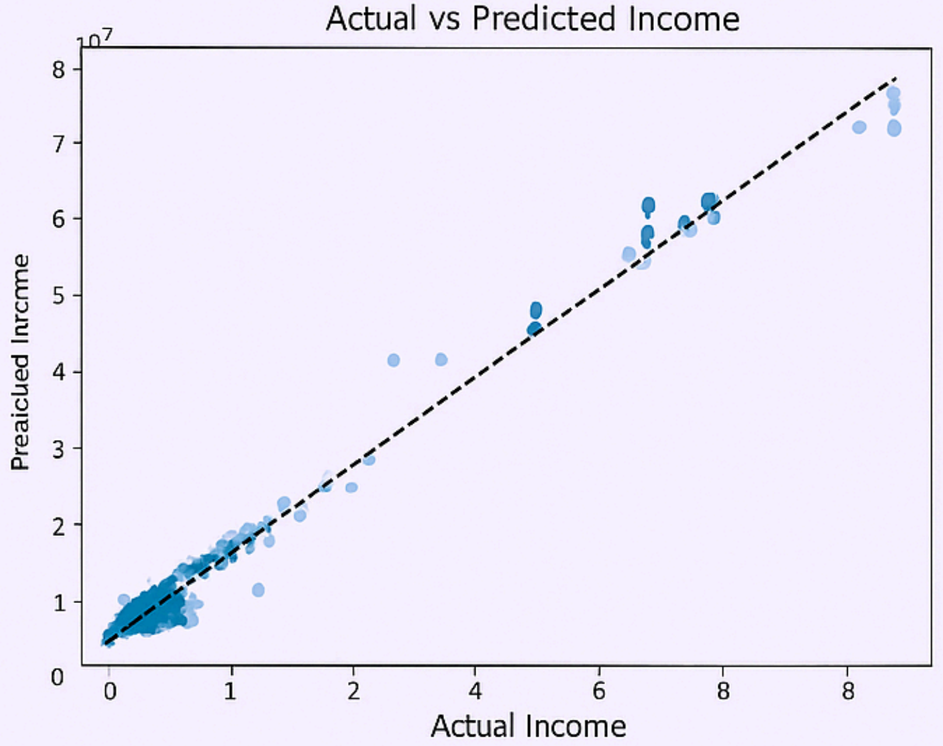
- MAPE (Mean Absolute Percentage Error)
- MAE (Mean Absolute Error)

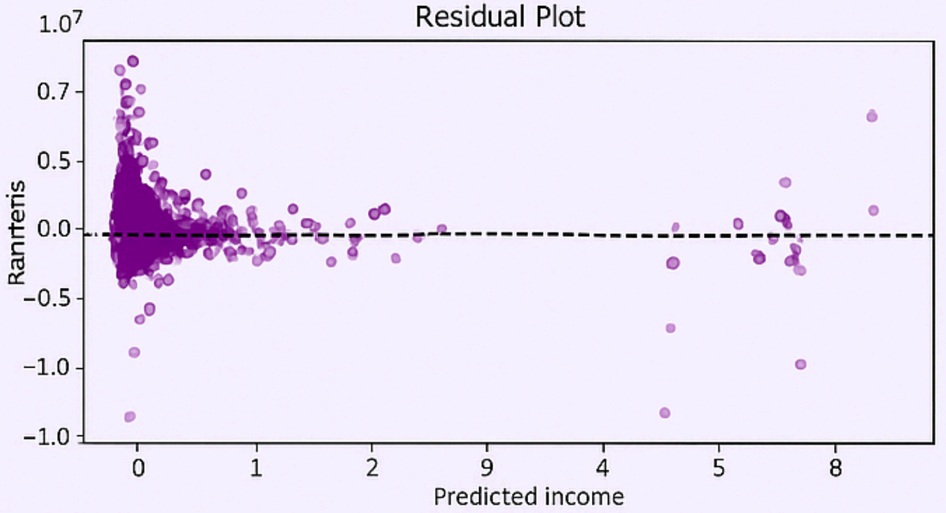
Why MAPE?

- Training MAPE ~23.5%
- Validation MAPE –24% (from learning curve)

Visual Analysis Included

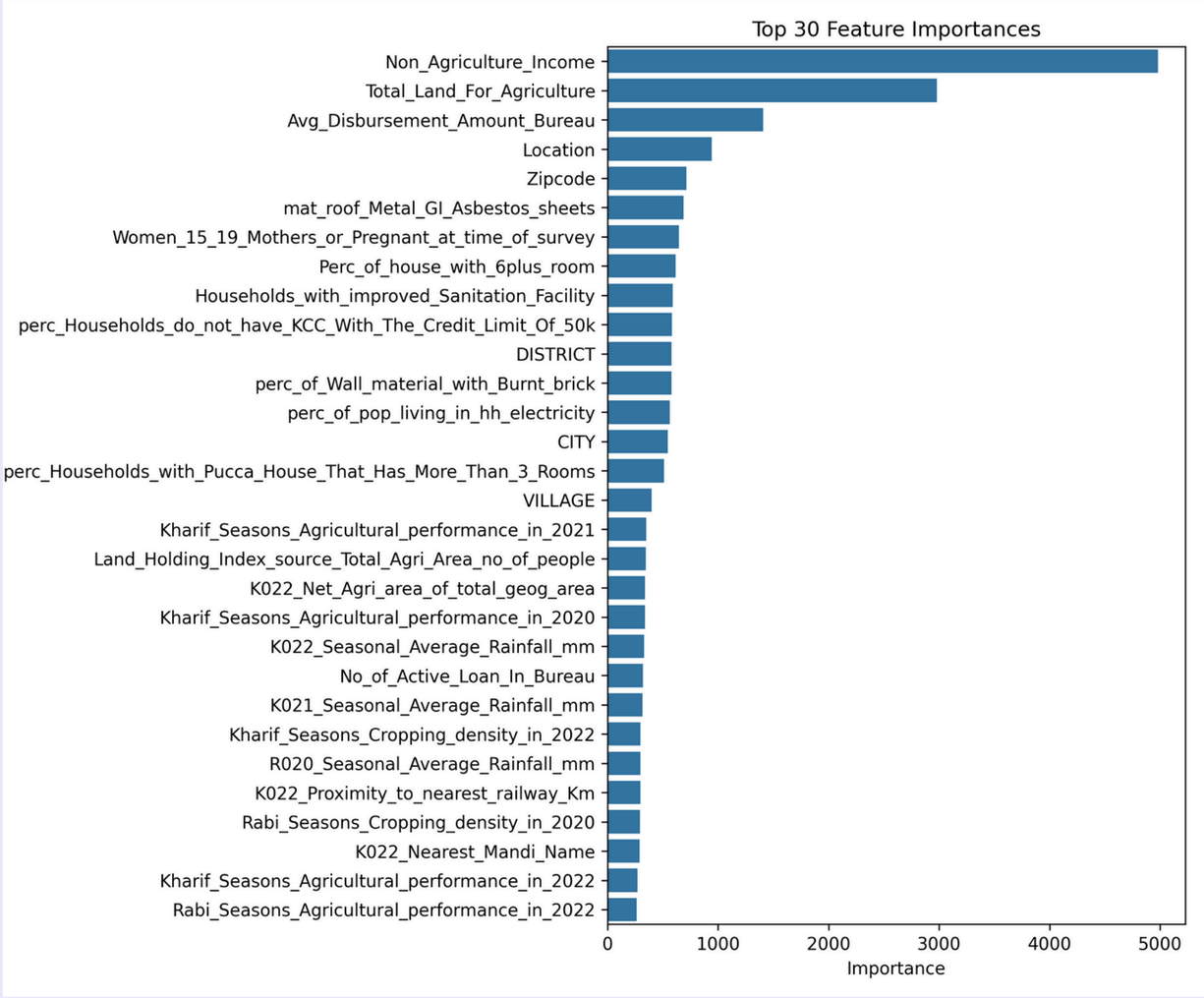
- Prediction Error Distribution
- Actual vs. Predicted Plot
- Residuals vs. Predicted Vaues

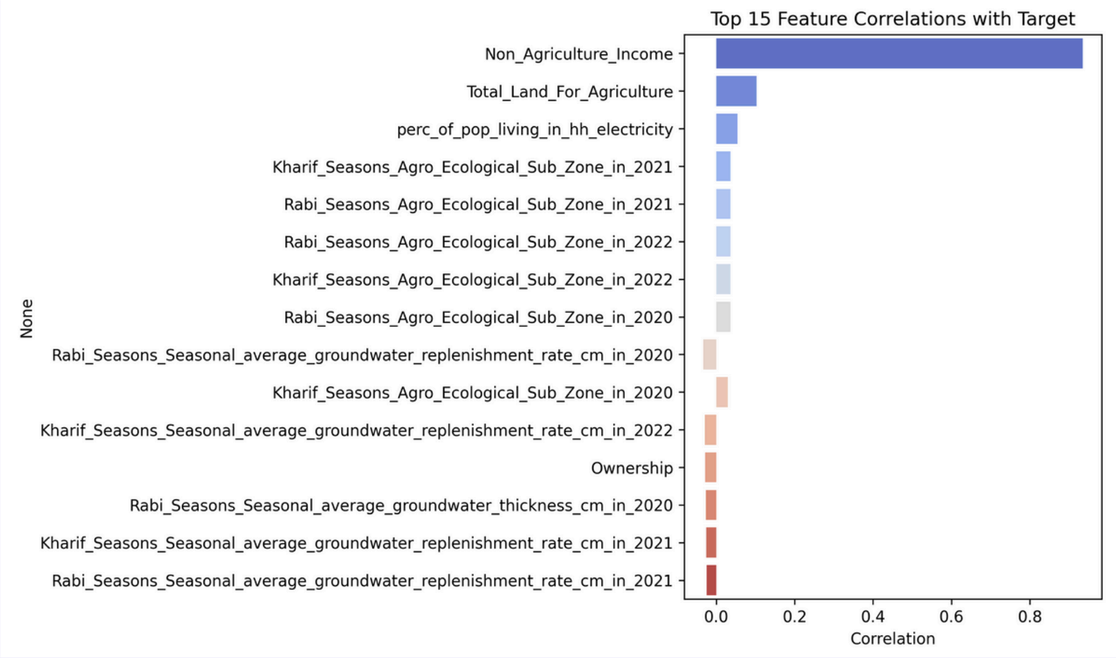


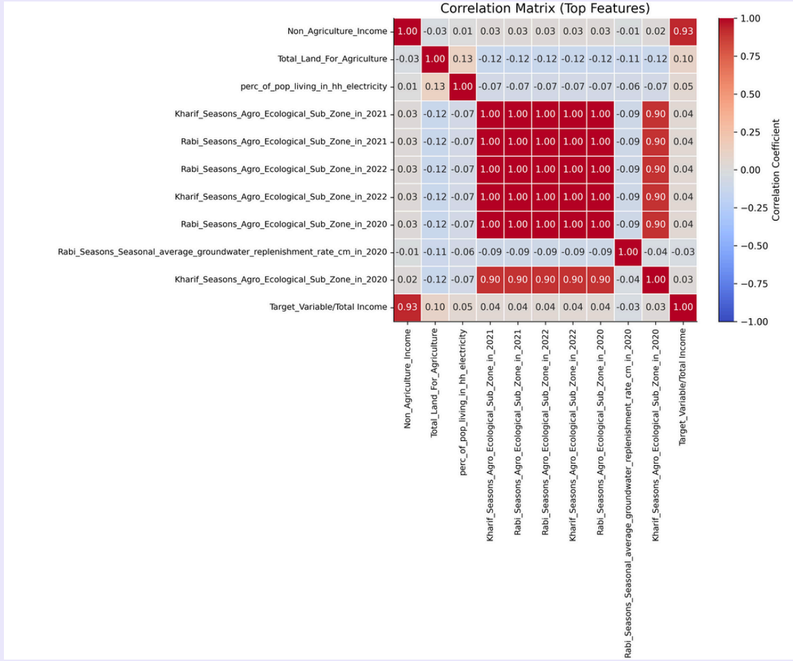


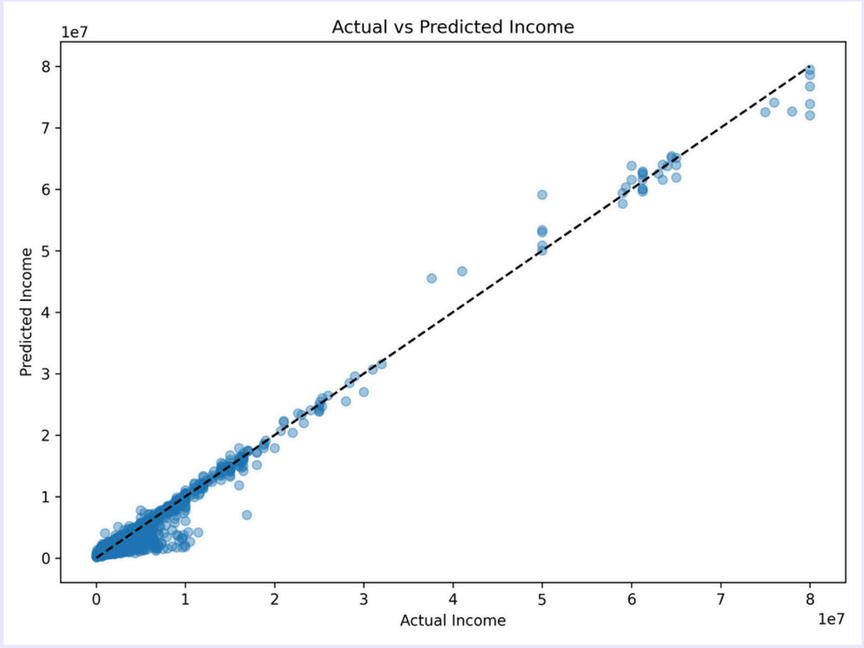
Data Analysis

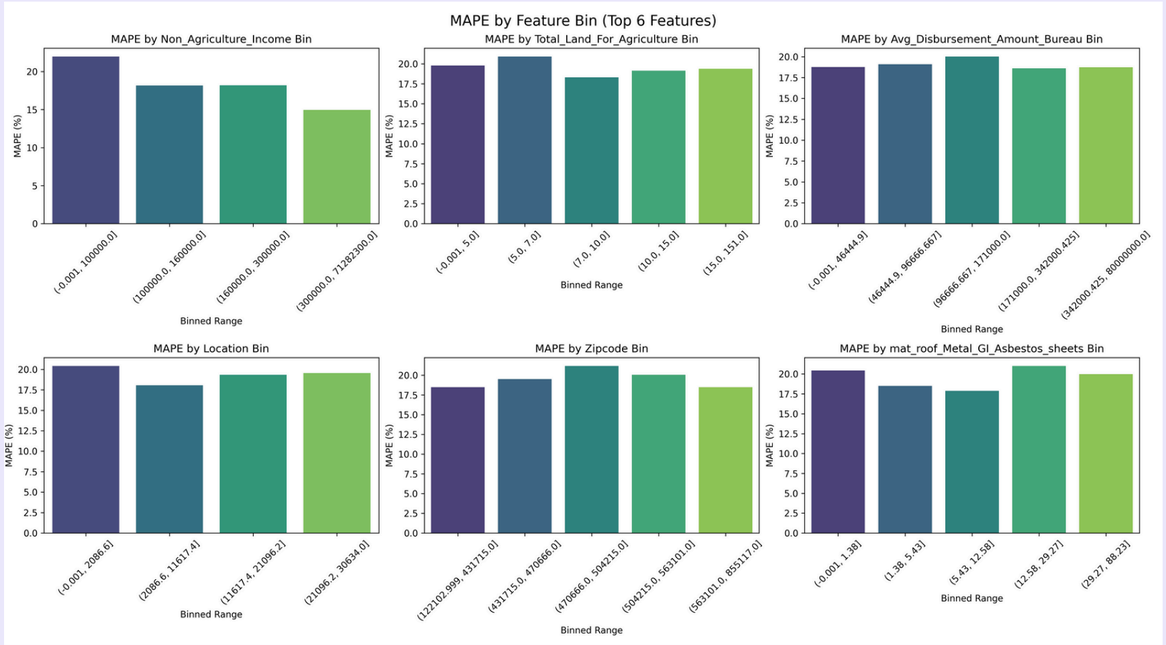
- Generated a feature importance plot to highlight the most impactful variables influencing farmer income predictions.
- Plotted Top 15 feature correlations to identify strong predictors of income.
- Created a correlation heatmap to visualize feature-target relationships.
- Used pairplots on top features to explore interactions and clustering.
- Validated key features for modeling through visual and statistical insights.




Top 15 Feature Correlations with Target
Non_Agriculture_Income
Total_Land_For_Agriculture
perc_of_pop_living_in_hh_electricity
Kharif_Seasons_Agro_Ecological_Sub_Zone_in_2021
Rabi_Seasons_Agro_Ecological_Sub_Zone_in_2021
Rabi_Seasons_Agro_Ecological_Sub_Zone_in_2022
Kharif_Seasons_Agro_Ecological_Sub_Zone_in_2022
Rabi_Seasons_Agro_Ecological_Sub_Zone_in_2020
Rabi_Seasons_Seasonal_average_groundwater_replenishment_rate_cm_in_2020
Kharif_Seasons_Agro_Ecological_Sub_Zone_in_2020
Kharif_Seasons_Seasonal_average_groundwater_replenishment_rate_cm_in_2022
Ownership
Rabi_Seasons_Seasonal_average_groundwater_thickness_cm_in_2020
Kharif_Seasons_Seasonal_average_groundwater_replenishment_rate_cm_in_2021
Rabi_Seasons_Seasonal_average_groundwater_replenishment_rate_cm_in_2021
None
0.0
0.2
0.4
0.6
0.8
Correlation


Correlation Matrix (Top Features)
Non_Agriculture_Income
Total_Land_For_Agriculture
perc_of_pop_living_in_hh_electricity
Kharif_Seasons_Agro_Ecological_Sub_Zone_in_2021
Rabi_Seasons_Agro_Ecological_Sub_Zone_in_2021
Rabi_Seasons_Agro_Ecological_Sub_Zone_in_2022
Kharif_Seasons_Agro_Ecological_Sub_Zone_in_2022
Rabi_Seasons_Agro_Ecological_Sub_Zone_in_2020
Rabi_Seasons_Seasonal_average_groundwater_replenishment_rate_cm_in_2020
Kharif_Seasons_Agro_Ecological_Sub_Zone_in_2020
Target_Variable/Total Income
Correlation Coefficient
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Actual vs Predicted Income
1e7
Predicted Income
Actual Income
1e7
0
1
2
3
4
5
6
7
8


MAPE by Feature Bin (Top 6 Features)
MAPE by Non_Agriculture_Income Bin
MAPE by Total_Land_For_Agriculture Bin
MAPE by Avg_Disbursement_Amount_Bureau Bin
MAPE by Location Bin
MAPE by Zipcode Bin
MAPE by mat_roof_Metal_GI_Asbestos_sheets Bin
MAPE (%)
Binned Range
(-0.001, 100000.0]
(100000.0, 160000.0]
(160000.0, 300000.0]
(300000.0, 7128200.0]
(-0.001, 5.0]
(5.0, 7.0]
(7.0, 10.0]
(10.0, 15.0]
(15.0, 151.0]
(-0.001, 46444.9]
(46444.9, 96666.667]
(96666.667, 171000.0]
(171000.0, 342000.425]
(342000.425, 80000000.0]
(-0.001, 2086.6]
(2086.6, 11617.4]
(11617.4, 21096.2]
(21096.2, 30634.0]
(122102.999, 431715.0]
(431715.0, 470666.0]
(470666.0, 504215.0]
(504215.0, 563101.0]
(563101.0, 855117.0]
(-0.001, 1.38]
(1.38, 5.43]
(5.43, 12.58]
(12.58, 29.27]
(29.27, 88.23]

KEY BUSINESS INSIGHTS

Non-Agricultural Income

Farmers with higher secondary income are more financially resillent.

Agricultural Performance

Recent seasonal yields (2020–2022) directly influence income predictions.

Landholding & Amenities

Larger land size and access to facilities (e.g, sanitation, roofiCIS, electricity) drive higher incomes

Geographic Disparities

Location-specific variables (ZIP code, district, city) reveal income gaps and intervention

REAL–WORLD IMPACT

Targeted Farmer Support

Helps identify financially vulnerable households for early intervention.

Informed Loan/Policy Decisions

Supports accurate subsidy distribution and loan disbursement strategies.

Data-Driven Agricultural Planning

Enables local governments and stakeholders to design evidence-based programs

Thank You