WELCOME

AGENDA

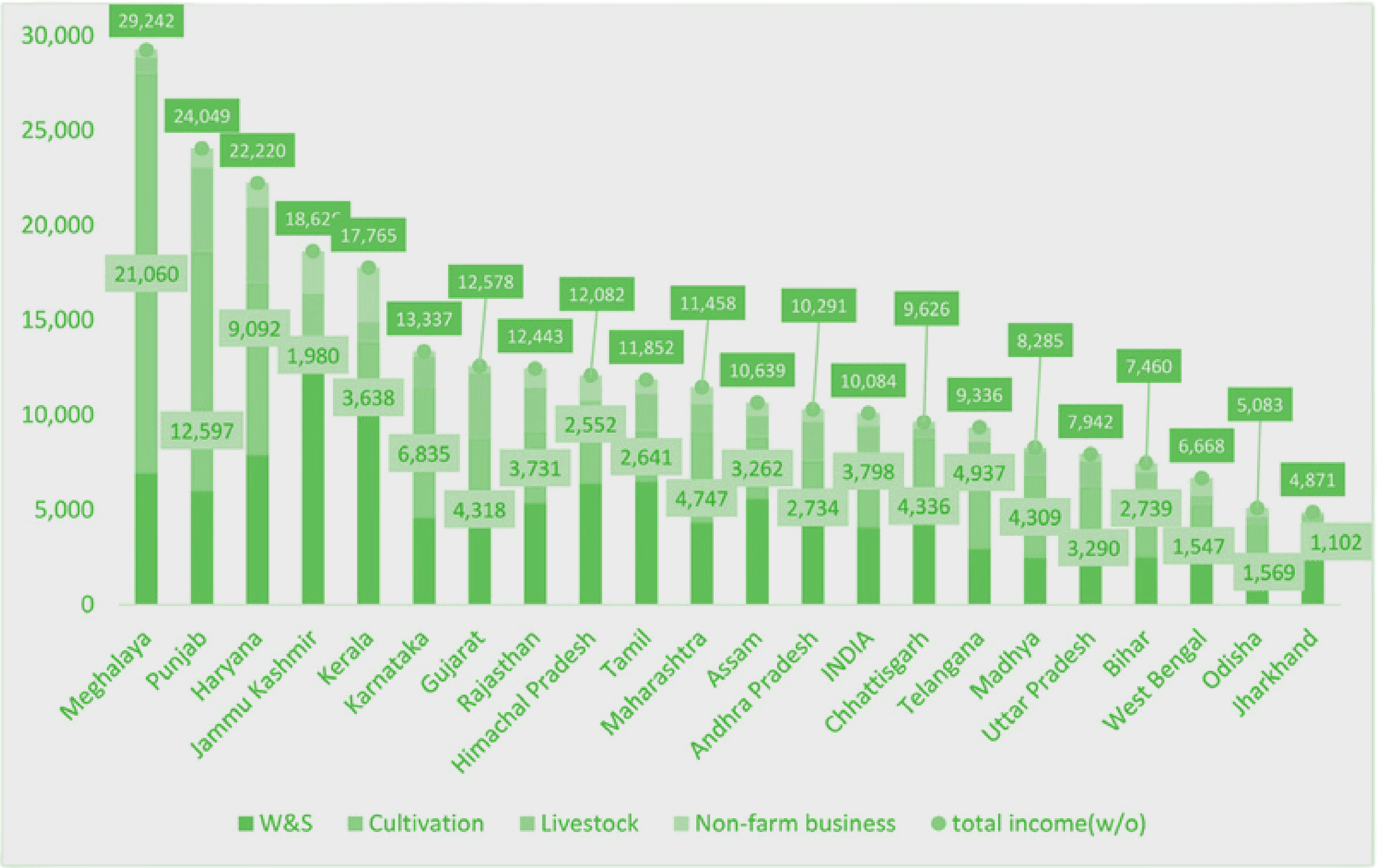
- INTRODUCTION
- EXPLORATORY DATA ANALYSIS
- PREPROCESSING
- FEATURE ENGINEERING
- MODEL SELECTION
- ENSEMBLE & PREDICTION
- CONCLUSION

INTRODUCTION : FARMER INCOME PREDICTION

LTF Challenge – Farmer Income Prediction

Many farmers in India lack formal credit histories, making it hard to access loans and forcing them to rely on risky, informal lenders - limiting their financial growth.

Our Goal: Build an accurate machine learning model to predict farmer income, aiming for a low Mean Absolute Percentage Error (MAPE) on the validation data.

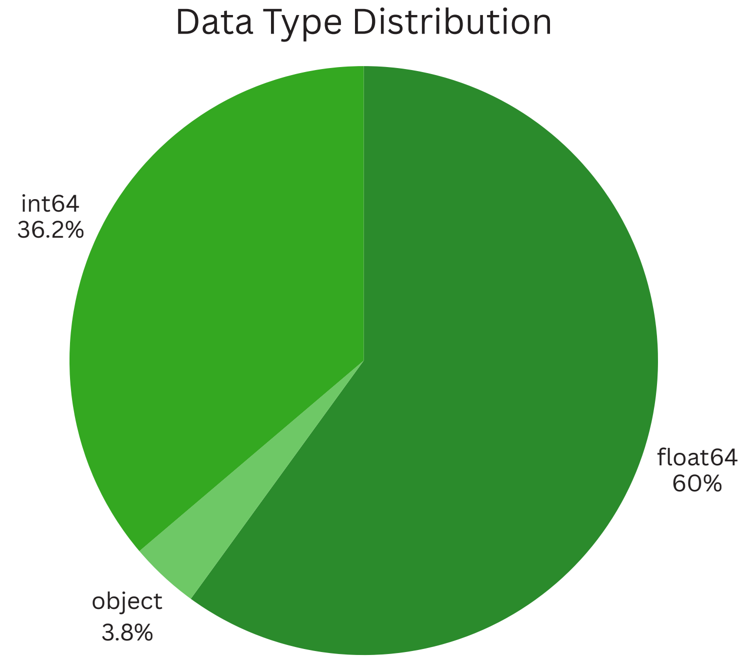


Source: NSSO 2018-19

EXPLORATORY DATA ANALYSIS: UNDERSTANDING THE DATASET

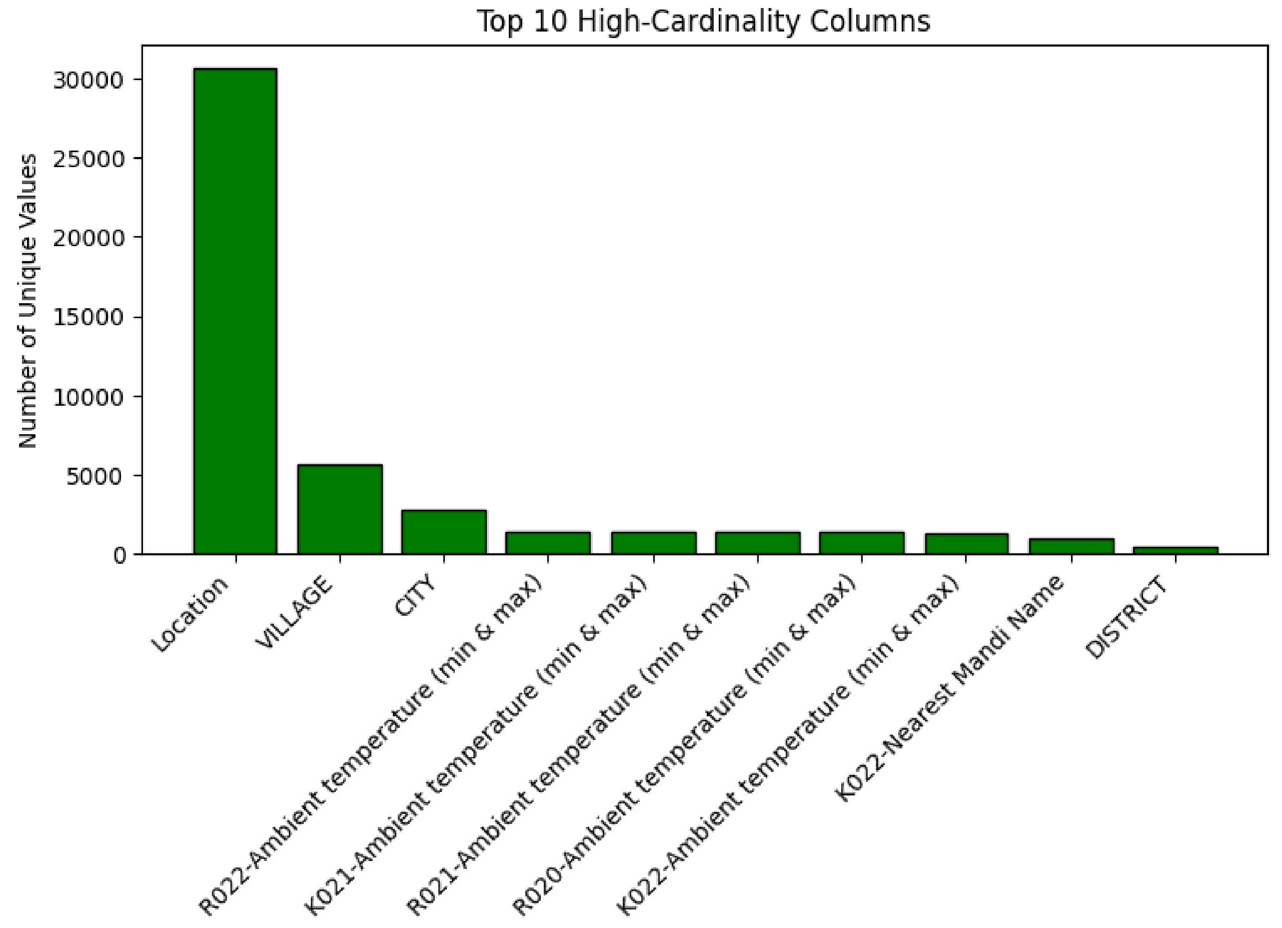
DATA DISTRIBUTION

- The LTF dataset contains 47,970 records.
- It features 105 variables spanning demographic, agricultural, environmental, infrastructural, and socio-economic domains.
- Data types include 62 float, 5 integer, and 38 object columns.
- Missing values are observed, indicating a need for data preprocessing.



CARDINALITY TEST

- Identified top-10 categorical features by unique-value count (e.g., City, Crop_Type, etc.)
- Dropped any with extremely high cardinality that added noise or inflated dimensionality

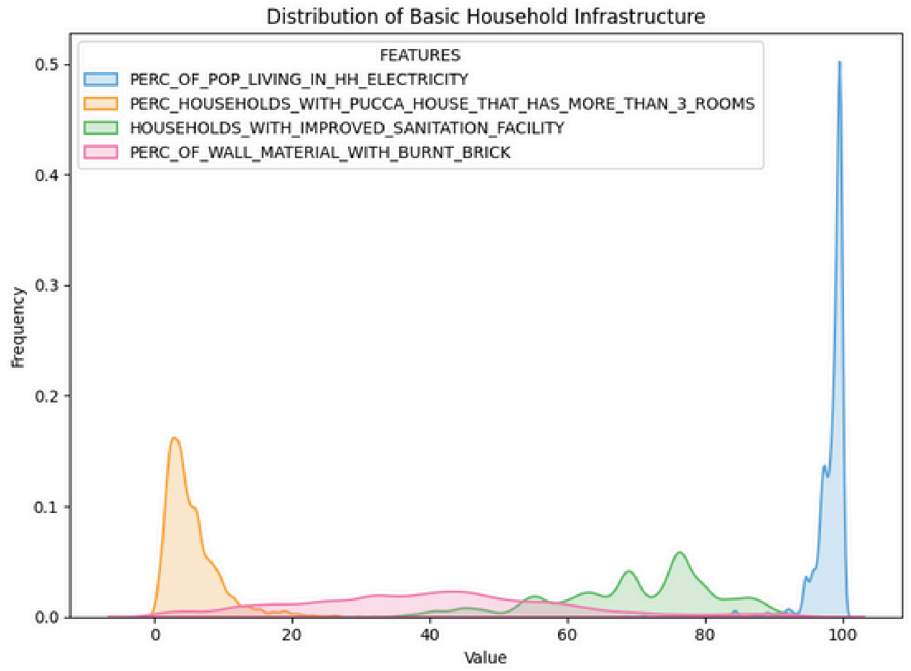


DATA PREPROCESSING: IMPUTING MISSING VALUES AND ENCODING

- INTRODUCTION
- EXPLORATORY DATA ANALYSIS
- PREPROCESSING
- FEATURE ENGINEERING
- MODEL SELECTION
- ENSEMBLE & PREDICTION
- CONCLUSION

Missing-Value Imputation

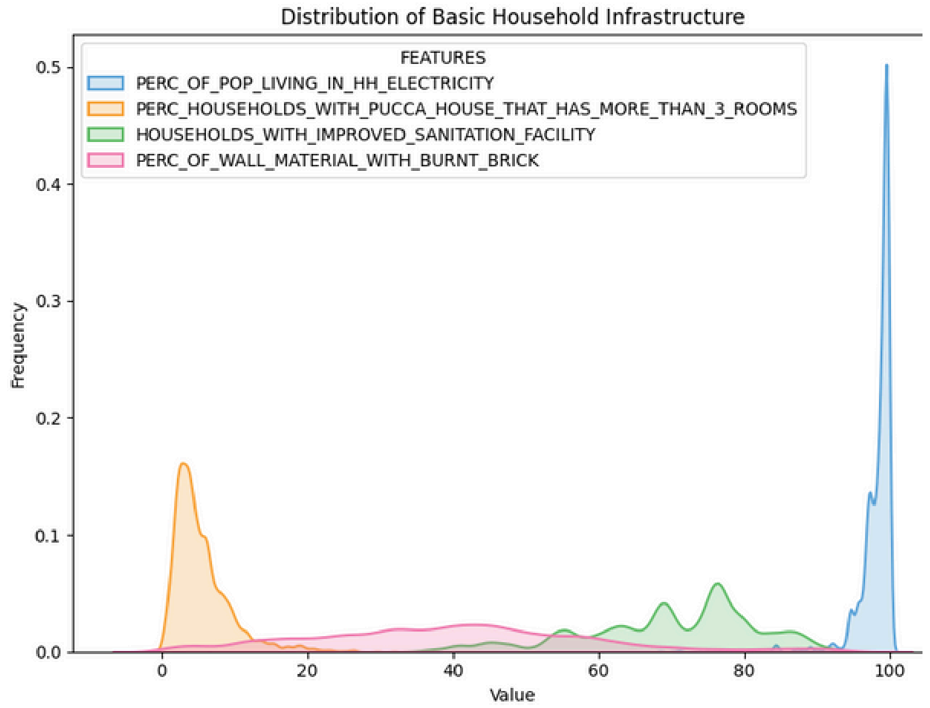
- Numerical columns: filled nulls with that column's median
- Categorical columns: filled nulls with the string "Unknown"



PRE-IMPUTATION

Categorical Encoding

- High-cardinality categorical features: Applied frequency encoding to efficiently convert them into numeric form without increasing feature space.
- Nominal categories: applied label encoding (For categorical features with ≤20 unique values).



POST-IMPUTING

DATA PREPROCESSING: TARGET VALUE DISTRIBUTION

INTRODUCTION

EXPLORATORY DATA ANALYSIS

PREPROCESSING

FEATURE ENGINEERING

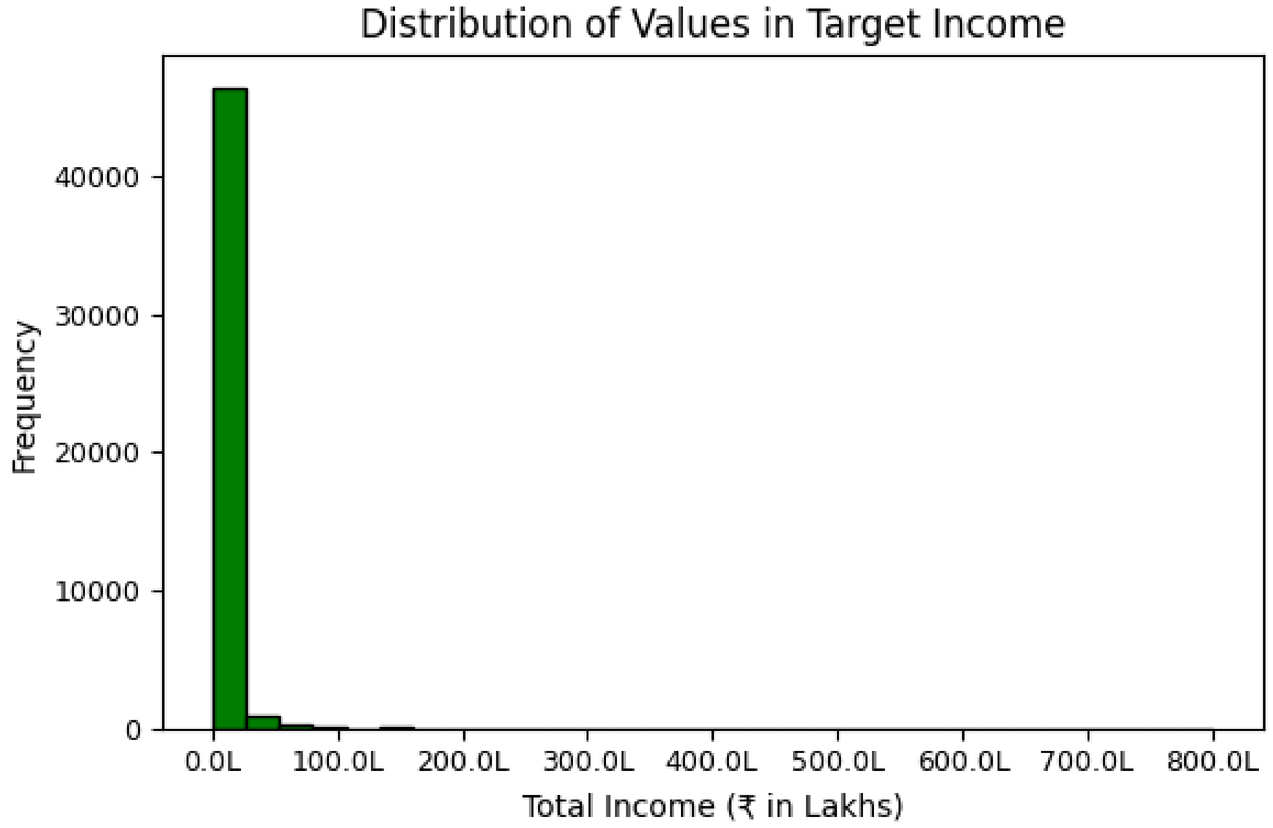
MODEL SELECTION

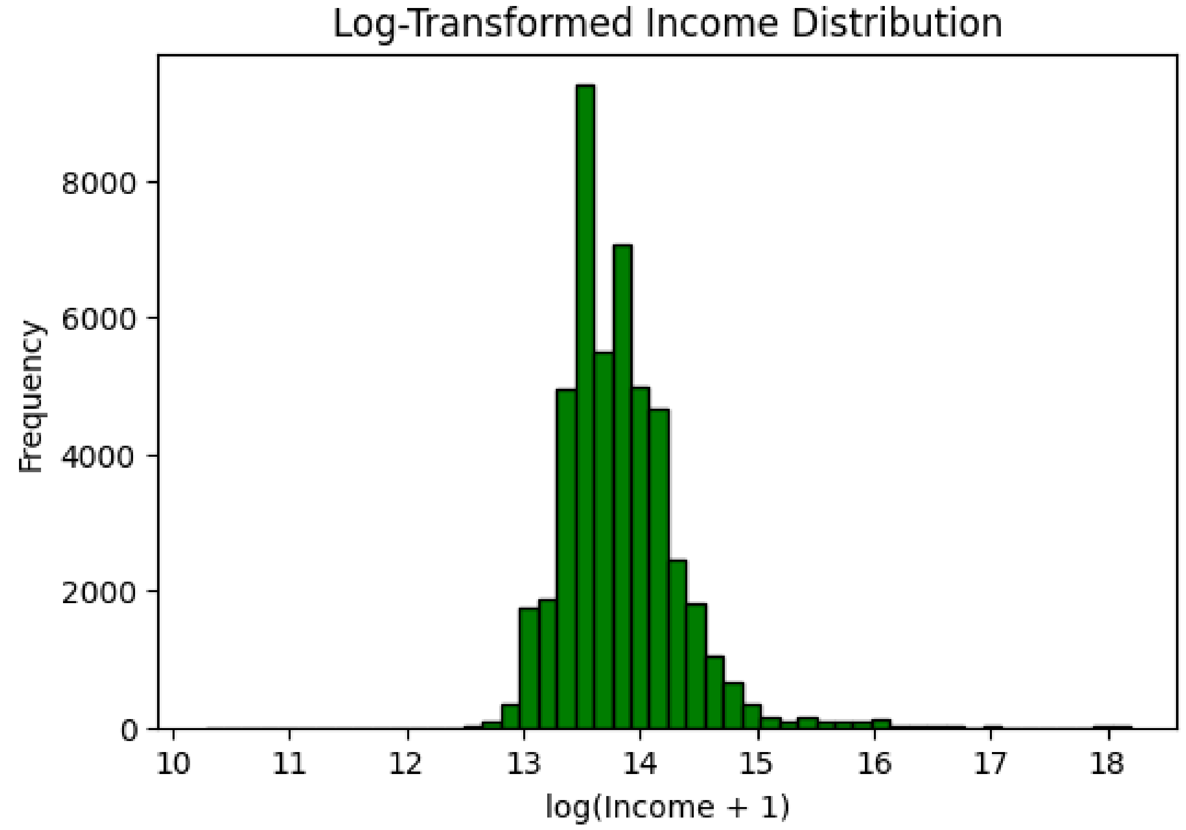
ENSEMBLE & PREDICTION

CONCLUSION

PRE-TRANSFORMATION

POST-TRANSFORMATION





Total income is highly right-skewed.

Applying log(Income+1) yields a much more symmetric distribution

Reduces skewness in income distribution and improves model performance on regression tasks

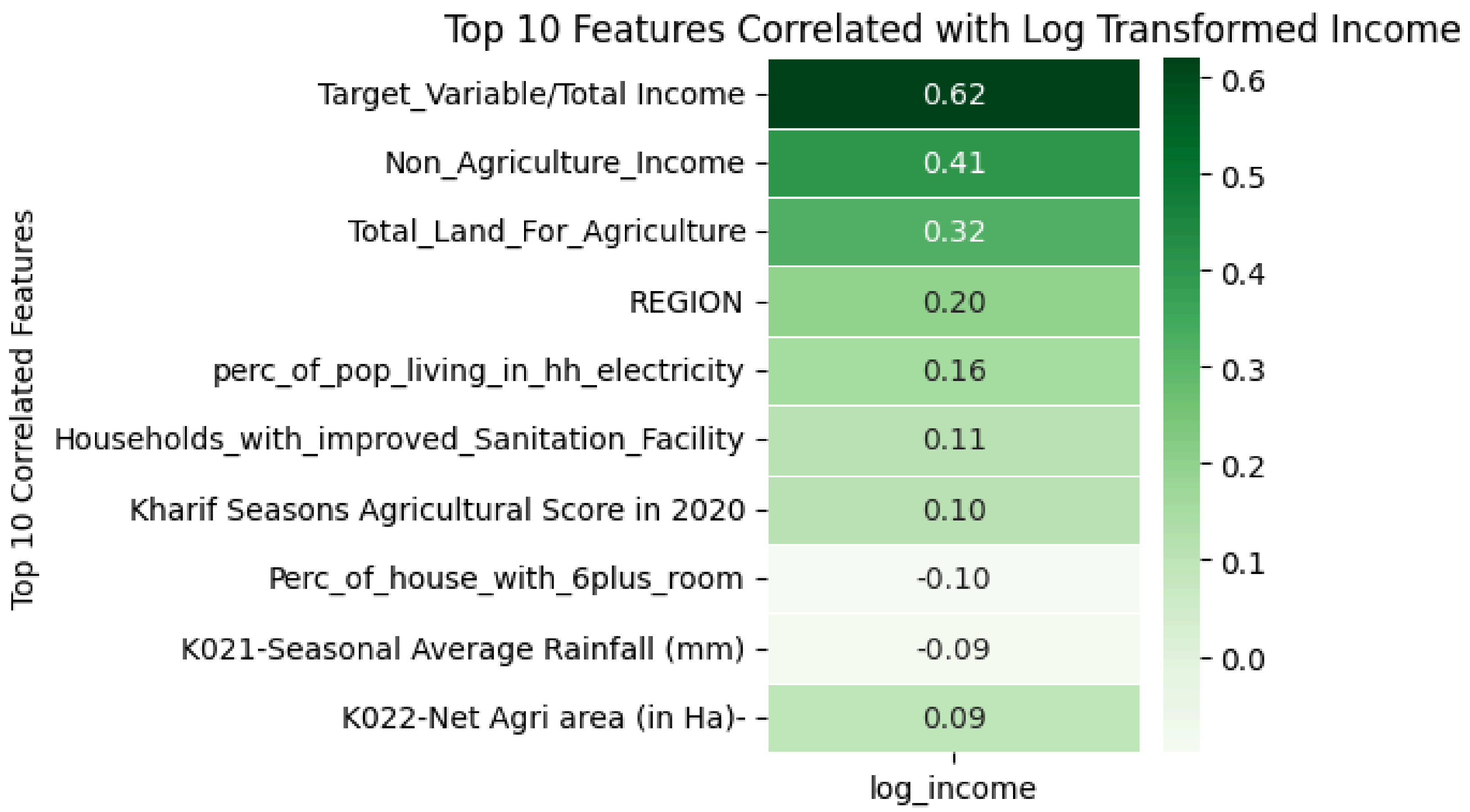
DATA PREPROCESSING: CORRELATION WITH TARGET VALUE

Objective: To identify which features have the strongest influence on the target variable.

Why it matters: Understanding these correlations helps in feature selection, improving both model accuracy and interpretability.

Key Findings

- High correlation values (positive or negative) indicate strong linear relationships with the target.
- Features like Non-Agricultural Income and Socio-Economic Score show direct impact on Total Income.
- Distance-related variables (e.g., Proximity to Mandi) have negative correlation, implying infrastructural access impacts income.

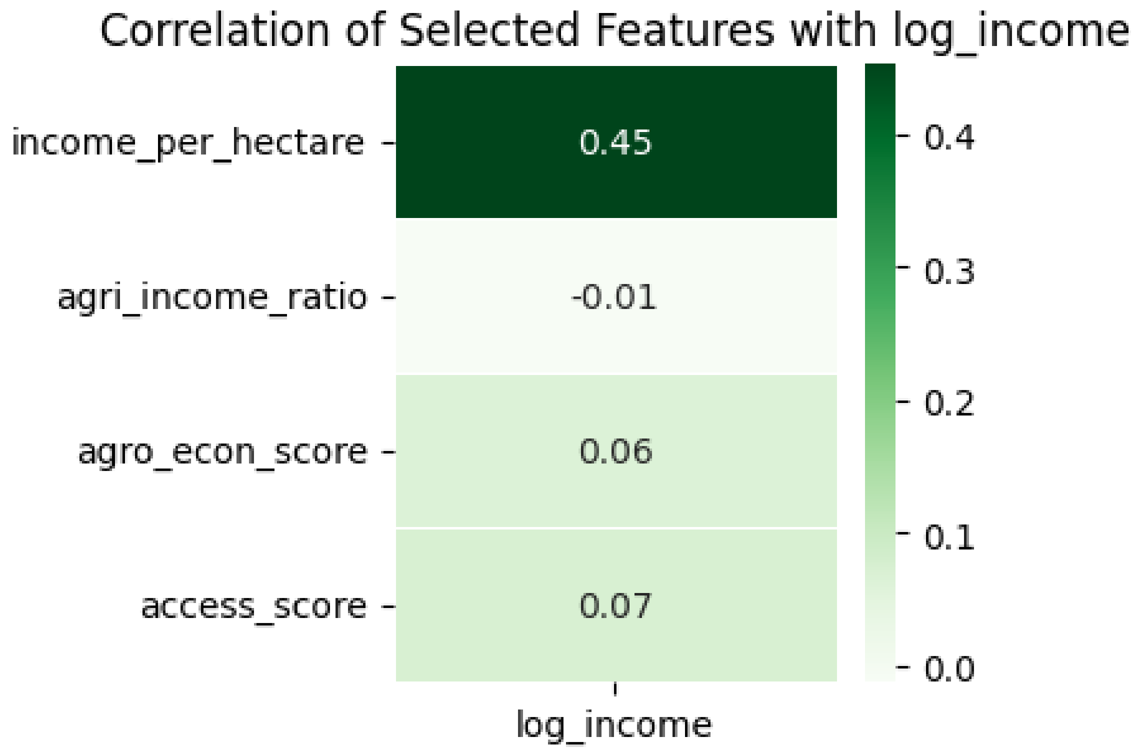


FEATURE ENGINEERING: METRIC EVALUATION AND FEATURE CREATION

INTRODUCTION
EXPLORATORY DATA ANALYSIS
PREPROCESSING
FEATURE ENGINEERING
MODEL TRAINING
ENSEMBLE & PREDICTION
CONCLUSION

AIM: Reducing feature redundancy while deriving more meaningful variables.

Impact: These new features link income to land productivity and socio-economic context, improving model fit (validated by higher accuracy).



Features Added

Income per Land

$$\text{Income per hectare} = \frac{\text{Total Income}}{\text{Total land for Agriculture}}$$

Agricultural Income Ratio

$$\text{Agricultural Income ratio} = \frac{\text{Income - Non Agriculture Income}}{\text{Income+1}}$$

Agro-Economic Score

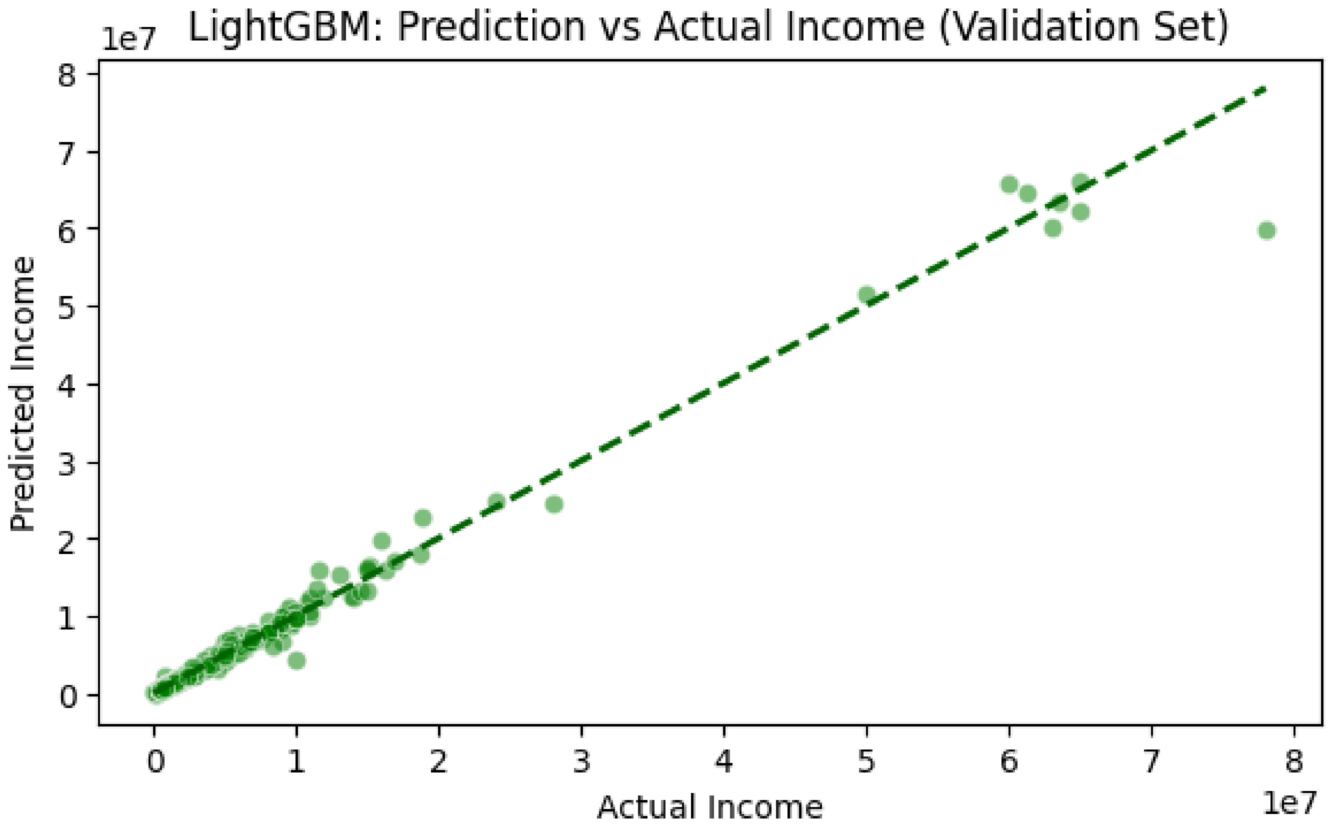
Agriculture-Economic Score = mean(Kharif Agricultural Score, Rabi Agricultural Score, Village SocioEconomic Score, Night light index)

Access Score

Access Score = - (Dist to nearest mandi + Dist to railway) + Road density

MODEL SELECTION: FINDING BEST-FIT MODEL

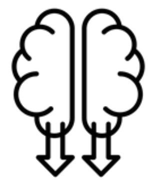
- INTRODUCTION
- EXPLORATORY DATA ANALYSIS
- PREPROCESSING
- FEATURE ENGINEERING
- **MODEL SELECTION**
- ENSEMBLE & PREDICTION
- CONCLUSION



Why LightGBM ?

Speed & Efficiency

Accuracy



Low Memory Usage

MAPE: 1.52%
R^2 Score: 0.9987
Accuracy within 10% error: 99.65%

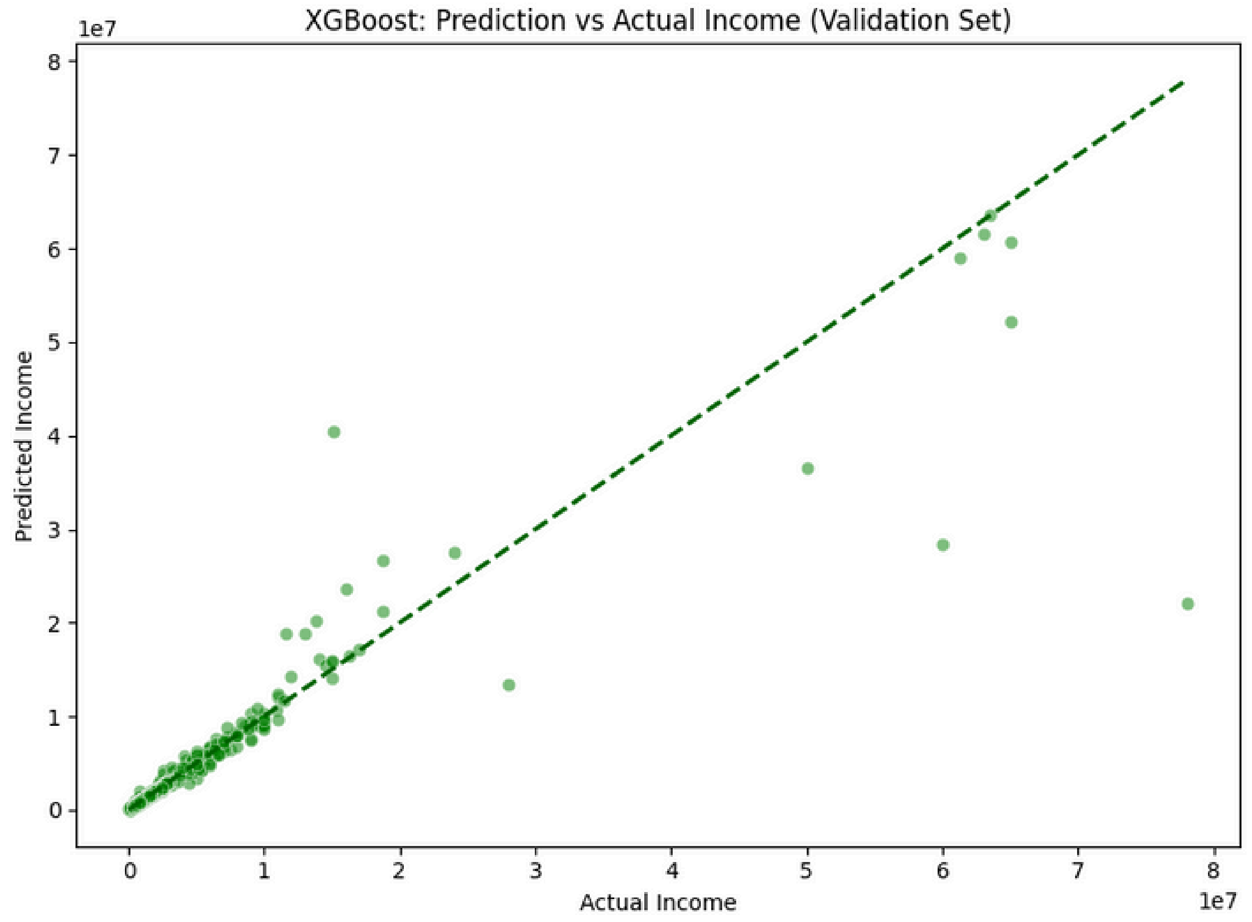
Why XGBoost ?

High Performance

Cross-Validation Built-in

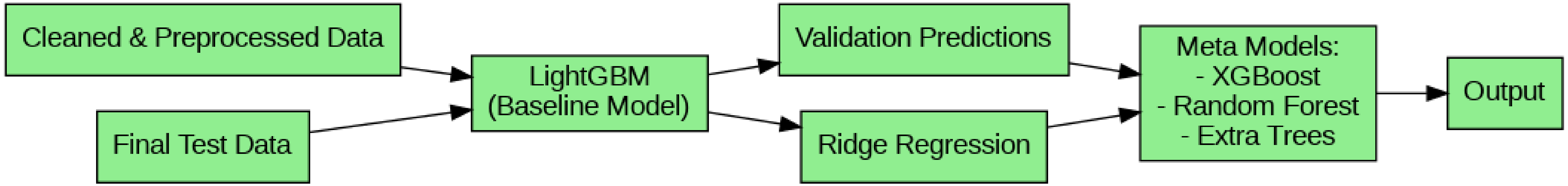
Low Memory Usage

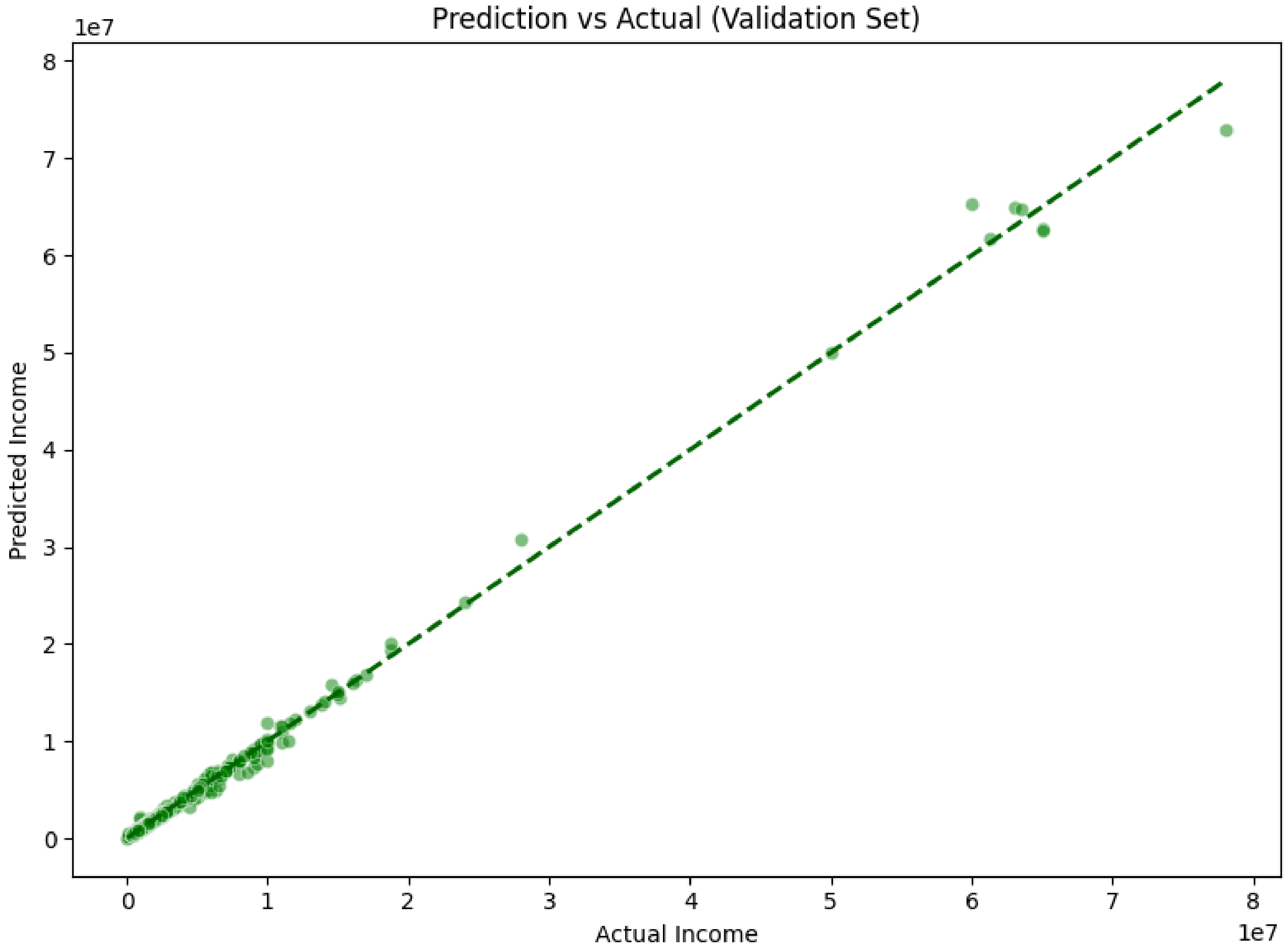
MAPE: 1.77%
R^2 Score: 0.8389
Accuracy within 10% error: 98.52%



ENSEMBLE & PREDICTION: STACKING REGRESSOR AND MAPE REDUCTION

Model Architecture





Why Ensemble?

- Ensembling combines multiple models to capture diverse patterns.
- Reduce errors to achieve higher accuracy and lower MAPE than any single model.

MAPE: 0.84%
R^2 Score: 0.9963
Accuracy within 10% error: 99.34%

CONCLUSION

INTRODUCTION
EXPLORATORY DATA ANALYSIS
PREPROCESSING
FEATURE ENGINEERING
MODEL TRAINING
ENSEMBLE & PREDICTION
CONCLUSION

Key Achievements

Achieved extremely low error:

MAPE 0.89% and 99.55% accuracy within 10% using extended stacked ensemble.

Built an end-to-end pipeline:

Data cleaning → Feature engineering → Model ensembling → Evaluation & Visualization.

Optimized for real-world heterogeneity:

Captured diverse patterns from agriculture, weather, socio-economic, and financial data.

Robust & scalable solution:

Model avoids overfitting, handles large datasets efficiently, and is ready for production deployment or further tuning.

Future Work

Integrate K-Fold Stacking

Further reduce the risk of overfitting and improve generalization.

Incorporate Temporal & Satellite Data

Use seasonal time-series & remote sensing data for richer predictions.

Deploy as an Interactive Dashboard

Build a web-based income prediction tool for decision-makers.

Expand to Risk & Loan Scoring

Utilize the same pipeline for credit risk, crop insurance, and policy planning.

THANK YOU