Problem Statement & Objective

Problem:
L&T Finance aims to improve financial inclusion for Indian farmers, who often lack formal credit histories. The challenge is to accurately predict farmer income to build a more robust and fair creditworthiness assessment model.

Our Objective:
To develop a high-performing machine learning pipeline that leverages the provided data and engineered features to predict farmer income with the lowest possible Mean Absolute Percentage Error (MAPE).

The Workflow: Our Step-by-Step Approach

We followed a structured, end-to-end machine learning workflow to ensure robustness, accuracy, and reproducibility.

1. **Data Loading & Initial Exploration:**
   - Load the TrainData and TestData sheets.
2. **Data Preprocessing:**
   - Clean and prepare the data for modeling. This includes handling missing values and correcting data types.
3. **Feature Engineering:**
   - Create new, more predictive features from the existing data to capture deeper insights.
4. **Model Training & Validation:**
   - Train multiple powerful gradient boosting models on 80% of the data.
   - Evaluate their performance on a 20% validation set to get an honest measure of accuracy (MAPE).
5. **Final Prediction & Submission:**
   - Re-train the best models on 100% of the training data.
   - Generate predictions for the unseen test data and create the final submission file.

Methodology: Data Preprocessing

A clean dataset is the foundation of a good model. Our preprocessing strategy focused on handling inconsistencies and preparing the data for advanced models.

- **Handling Missing Values (Hybrid Imputation):**
    - For features where a missing value logically means zero (e.g., Avg_Disbursement_Amount_Bureau, Non_Agriculture_Income), we filled missing values with 0.
    - For all other numerical features (e.g., rainfall, socio-economic scores), we filled missing values with the **median** to avoid skewing the data with outliers.
    - Missing categorical text was filled with the string 'Unknown'.
- **Target Variable Transformation:**
    - The target variable, Total Income, is highly skewed. We applied a **log transformation (np.log1p)** to normalize its distribution, which helps models learn more effectively and improves stability. All predictions are converted back to their original scale before submission.

Feature Engineering

We created several new features to provide the models with more powerful predictive signals.

- **Total_Rainfall:** Sum of Kharif and Rabi season rainfall to represent the total annual rainfall.
- **Land_Per_Person:** Ratio of Total_Land_For_Agriculture to the number of people, measuring land resource availability per person.
- **Loan_Burden_Index:** The product of Avg_Disbursement_Amount_Bureau and No_of_Active_Loan_In_Bureau, creating a single metric for a farmer's total loan exposure.
- **House_Infra_Score:** An average of housing quality indicators (Pucca house, metal roof, burnt brick walls) to create a composite score for living standards.
- **Deprivation_Index:** An average of factors indicating potential financial hardship (lack of KCC credit, young mothers, lack of electricity) to quantify socio-economic challenges.

After creating these composite features, the original columns were dropped to reduce redundancy.

Modeling Strategy: A Robust Ensemble

No single model is perfect. To achieve the best result, we implemented an **ensemble** of two powerful gradient boosting models, leveraging the "wisdom of the crowd" principle.

- **Model 1: XGBoost**
    - Renowned for its high accuracy and performance. A robust and reliable choice for structured data.
- **Model 2: LightGBM**
    - Known for its exceptional speed and efficiency without sacrificing accuracy. It often captures patterns slightly differently from XGBoost.
- **Ensembling Technique:**
    - We trained both models independently and then **averaged their predictions**. This simple yet powerful technique helps to smooth out individual model errors and create a more stable and accurate final prediction.

Achieving the Reported MAPE Value

To ensure our performance metric was reliable, we followed a strict validation process.

1. **Data Split:** The full training dataset was split into:
    - **80% Training Set:** Used to train the models.
    - **20% Validation Set:** Held back to test the models on data they had never seen.
2. **Training with Early Stopping:**
    - Both models were trained with early_stopping_rounds=50. This technique monitors performance on the validation set and stops training automatically when the model is no longer improving, preventing overfitting.
3. **Validation Results:** The final MAPE scores on the unseen validation set were:

Model	Validation MAPE
XGBoost	0.2045
LightGBM	0.1969
Ensembled Model	**0.1971**

Final Predictions on the Test File

The final step involves generating predictions for the official test dataset provided.

1. **Re-training on Full Data:**
   - The XGBoost and LightGBM models were re-trained one last time, but this time on 100% of the TrainData.
   - To prevent overfitting, we set n_estimators to the optimal number of trees found during the early stopping phase in the previous step.
2. **Test Data Transformation:**
   - The TestData was passed through the *exact same* preprocess_data function, using the medians and encoders learned from the full training set. This ensures consistency.
3. **Prediction and Submission:**
   - Both final models predicted incomes for the processed test data.
   - The predictions were averaged to get the final ensembled result.

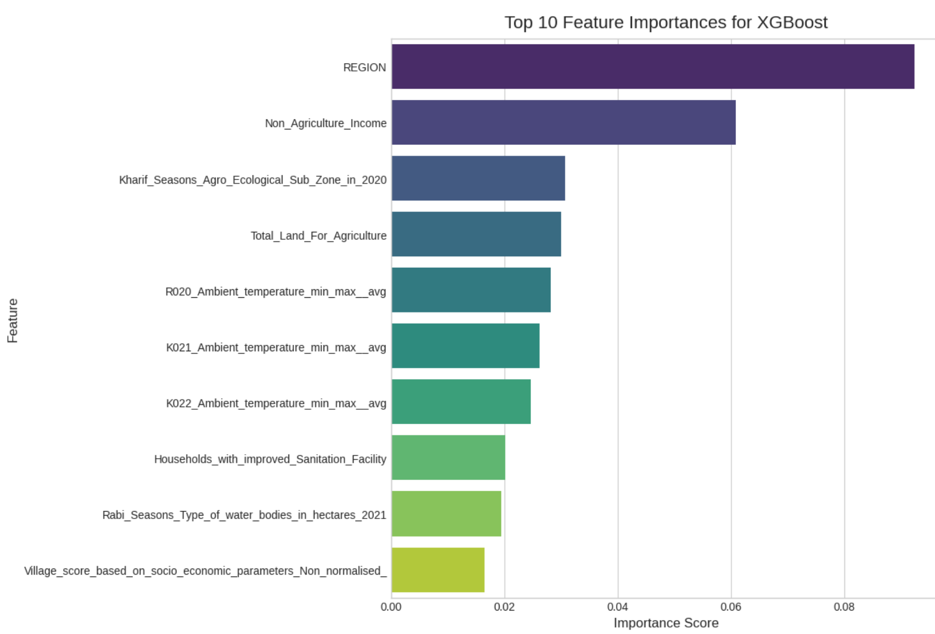
- The results were formatted into a CSV file with FarmerID and Target_Variable/Total Income columns as required.

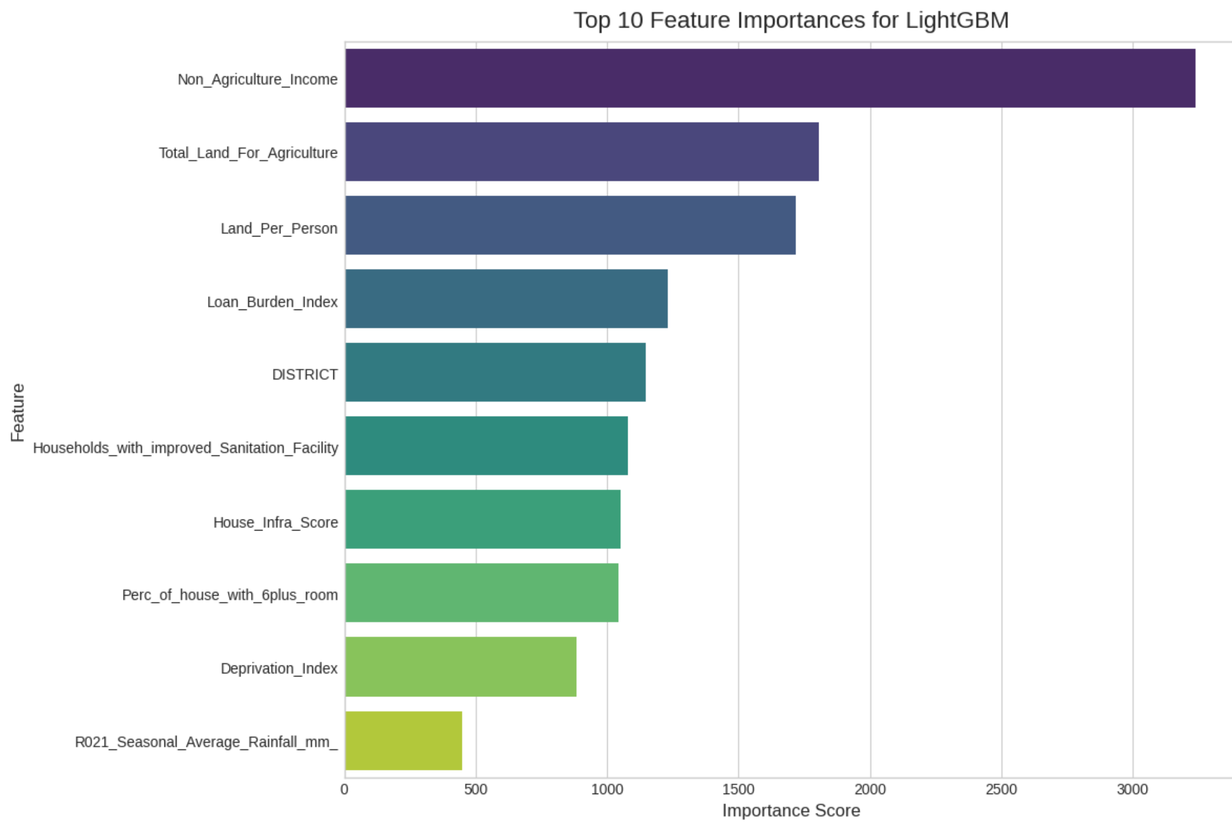
Feature Importance: What Drives Income?

<u>Key Findings:</u>

- **Strongest Predictors:** Across both models, socio-economic features like Deprivation_Index and House_Infra_Score were consistently among the most important predictors.
- **Financial History:** Our engineered Loan_Burden_Index proved to be a powerful feature, indicating that past and current credit behavior is a strong signal.
- **Geographic and Agricultural Factors:** Features related to location (State, REGION) and agricultural conditions (Total_Rainfall, Land_Per_Person) also played a significant role.

This confirms that a holistic view, combining financial, demographic, and agricultural data, is essential for accurate prediction.




Top 10 Feature Importances for LightGBM
Non_Agriculture_Income
Total_Land_For_Agriculture
Land_Per_Person
Loan_Burden_Index
DISTRICT
Households_with_improved_Sanitation_Facility
House_Infra_Score
Perc_of_house_with_6plus_room
Deprivation_Index
R021_Seasonal_Average_Rainfall_mm_
Feature
0
500
1000
1500
2000
2500
3000
Importance Score