**Farmers’ Credit Worthiness Classification: A Machine Learning Approach**

**Data Description and Missing Values**

| **0** |
| --- |
| **farmer\_id** | 0 |
| **age** | 0 |
| **gender** | 0 |
| **location\_state** | 0 |
| **education\_level** | 425 |
| **marital\_status** | 0 |
| **dependants** | 0 |
| **farm\_type** | 155 |
| **farm\_size\_acres** | 0 |
| **years\_of\_experience** | 0 |
| **access\_to\_extension\_services** | 0 |
| **access\_to\_irrigation** | 0 |
| **mobile\_phone\_access** | 0 |
| **financial\_literacy\_level** | 0 |
| **has\_bank\_account** | 0 |
| **previous\_loan\_history** | 0 |
| **previous\_loan\_amount** | 155 |
| **previous\_loan\_repaid** | 0 |
| **total\_annual\_income** | 0 |
| **other\_income\_sources** | 1885 |
| **group\_membership** | 0 |
| **gps\_latitude** | 0 |
| **gps\_longitude** | 0 |
| **credit\_score** | 0 |
| **creditworthy** | 0 |

**Data Cleaning Steps Applied:**

1. Missing value imputation for 'previous\_loan\_amount' based on 'previous\_loan\_history':

- If 'previous\_loan\_history' is True, missing values are filled with the median of 'previous\_loan\_amount' for entries where 'previous\_loan\_history' is True.

- If 'previous\_loan\_history' is False, missing values are filled with 0.

2. Missing value imputation for categorical features ('education\_level', 'farm\_type', 'other\_income\_sources'):

- Missing values are filled with the mode of each respective column.

3. Outlier treatment using Winsorization (IQR method):

- The following numerical features were winsorized: 'age', 'dependants', 'farm\_size\_acres', 'years\_of\_experience', 'financial\_literacy\_level', 'previous\_loan\_amount', 'total\_annual\_income', 'gps\_latitude', 'gps\_longitude', 'credit\_score'.

4. Data consistency check for 'age':

- Rows where 'age' is less than or equal to 0 were removed.

**Data Distribution After Cleaning**

A group of blue boxes

Description automatically generated

Logistic Regression Metrics:

Accuracy: **0.7000**

Precision**: 0.7119**

Recall: **0.9633**

F1-score: **0.8187**

AUC-ROC: 0.8764

Random Forest Classifier Metrics:

Accuracy**: 0.9645**

Precision**: 0.9814**

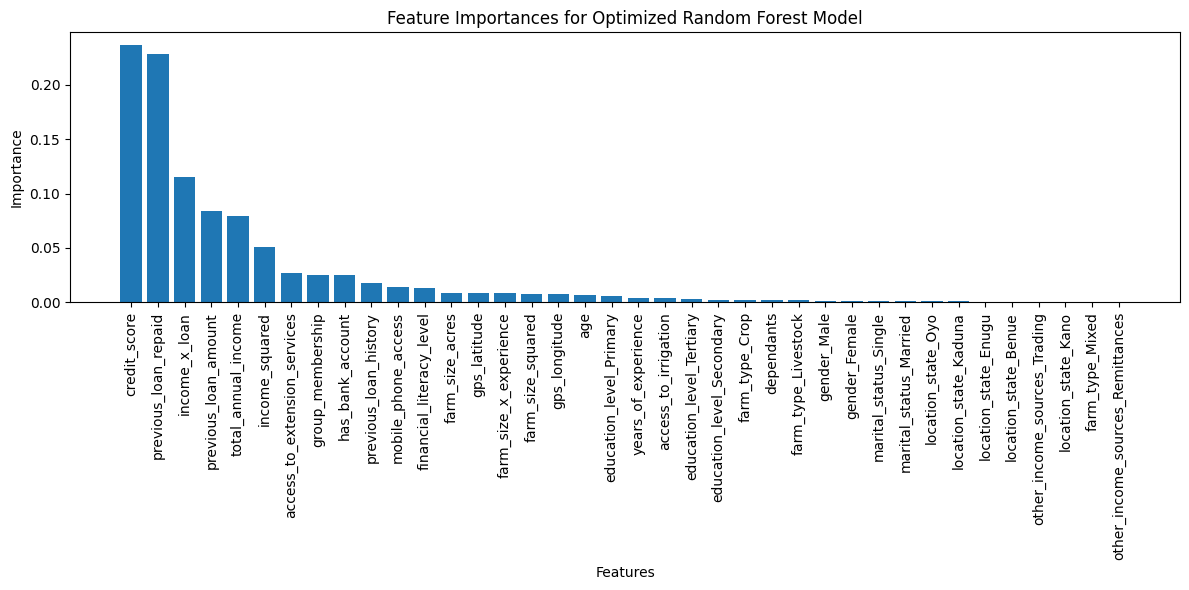
Recall**: 0.9679**

F1-score: **0.9746**

AUC-ROC**: 0.9974**

**It can be observed that** Random Forest Classifier significantly outperformed the Logistic Regression model across all evaluation metrics on the validation set. The optimized Random Forest also performed well on the test set.

Also the best hyperparameters were: {'n\_estimators': 50, 'min\_samples\_split': 10, 'min\_samples\_leaf': 4, 'max\_features': 'log2', 'max\_depth': 10}.



The key features were visualized using the feature importance bar chart, and it was discovered that Financial Literacy, gps\_longitude,gps\_latitude, location, gender and other income sources contributed little to nothing to the model so they were dropped and new models **(Logistic regression, Random Forest and XGBoost) were trained on the new data**

**For the new data**

**Data Analysis Key Findings**

**The dataset had no missing values.**

Outliers in numerical features were handled using the IQR method.

Feature engineering included interaction terms, **polynomial features, scaling of numerical features, and one-hot encoding of categorical features.**

The Random Forest Classifier consistently outperformed the Logistic Regression model.

Hyperparameter tuning of the Random Forest model using RandomizedSearchCV resulted in a model with a high F1 score (approximately 0.97) on the validation set and a slightly lower, but still excellent score, on the test set.

Feature importance analysis showed that certain features (not explicitly listed here due to formatting constraints) were more influential in the model's predictions than others.

**Insights or Next Steps**

Further Feature Engineering: Explore more complex feature interactions or transformations to potentially improve model performance.

Model Explainability: Use techniques like SHAP values to better understand the model's decision-making process and identify the most influential features for individual predictions.

**Feature Engineering Steps**

\* Interaction features

farm size irrigation = farm size acres x access to irrigation

income\_experience = total annual income x years of experience

\* Numerical feature transformations

log\_income = log(total annual income)

square root of loan amount = square root(previous\_loan\_amount)

\* Polynomial features

age\_squared = age\*\*2

experience\_cubed = years of experience\*\*2

**Model Evaluation**

Logistic Regression:

Accuracy: 0.9633, Precision: 0.9653, Recall: 0.9633, F1-score: 0.9637, AUC-ROC: 0.9953

Random Forest:

Accuracy: 0.9567, Precision: 0.9565, Recall: 0.9567, F1-score: 0.9566, AUC-ROC: 0.9908

XGBoost:

Accuracy: 0.9567, Precision: 0.9588, Recall: 0.9567, F1-score: 0.9571, AUC-ROC: 0.9968

Therefore, XGBoost was utilized as our best performing model

**The model was then deployed on streamlit**