# Part 1:

## One Acre Fund Loan Repayment Prediction Report

### 1. Executive Summary

This project aims to build a model that accurately predicts loan repayment behavior among One Acre Fund’s smallholder farmers. The primary objective is to achieve a 98% repayment rate by identifying clients at risk of default and optimizing intervention strategies. Using predictive modeling, this report evaluates key repayment predictors and presents actionable insights for better repayment targeting.

Our findings indicate that certain financial and demographic features are strong predictors of repayment behavior, including `nominal\_contract\_value`, `cumulative\_amount\_paid\_start`, and `deposit\_amount`. With these insights, we recommend focusing on clients with lower nominal contract values and higher deposit ratios, as these characteristics correlate with improved repayment likelihood.

### 2. Methodology, Solution Path, and Assumptions

#### Methodology

The solution path for this analysis involves multiple stages, including data exploration, feature engineering, model selection, and evaluation. We selected Random Forest and XGBoost models based on their strong performance in handling structured regression tasks, which aligns well with the continuous nature of our target variable, `cumulative\_amount\_paid`.

#### Solution Path

1. **Exploratory Data Analysis (EDA):**

Examined variable distributions, correlations, and checked for missing values, dropping rows where values were missing (<2% of total data).

Key variables showed positive correlations with the target, `cumulative\_amount\_paid`, indicating potential predictors of repayment patterns.

2. **Feature Engineering:**

Engineered new features to capture additional aspects of client behavior:

* days\_since\_start: Number of days since contract initiation, which represents repayment timeliness.
* deposit\_ratio: Deposit amount as a fraction of the contract value, indicating initial client commitment.
* repayment\_ratio: Proportion of the loan repaid to date, representing repayment progress.

These features aimed to improve prediction quality by capturing temporal and behavioral aspects of repayment.

### 3. Model Selection and Evaluation

We chose **Random Forest** and **XGBoost** due to their robustness in regression tasks and ability to capture complex relationships.

Hyperparameter tuning was performed for each model to optimize performance on our dataset.

#### Assumptions

Several assumptions underlie our analysis:

* Consistency in Client Behavior: Assumes consistent repayment patterns across clients, which may vary seasonally or due to external factors.
* Sufficiency of Existing Data: Assumes that existing variables, alongside engineered features, sufficiently capture repayment behavior.
* Missing Data Management: Assumes minimal bias introduced by dropping rows with missing values, given their low prevalence (<2%).

#### Techniques Attempted

##### Models Implemented

Random Forest: A highly interpretable ensemble model capable of capturing non-linear relationships, tuned with optimal parameters for depth, sample split, and estimator count.

XGBoost: A gradient-boosting model known for strong performance with tabular data and capable of capturing complex interactions, also fine-tuned for depth and learning rate.

##### Hyperparameter Tuning

For both models, hyperparameter tuning was automated to select parameters yielding the best validation results:

Random Forest: Optimal parameters included `max\_depth=20`, `min\_samples\_leaf=2`, `min\_samples\_split=5`, `n\_estimators=200`.

XGBoost: Optimal parameters included `learning\_rate=0.1`, `max\_depth=5`, `n\_estimators=200`.

#### Evaluation Metrics

RMSE (Root Mean Square Error): Chosen for its ability to quantify error magnitude in predicting `cumulative\_amount\_paid`, essential for assessing model accuracy in a monetary context.

R² (Coefficient of Determination): Measures the proportion of variance captured by the model, indicating prediction reliability.

### 4. Results and Model Evaluation

#### Model Performance

**Random Forest Results**:

* Training RMSE: 950.81, Validation RMSE: 1142.28
* Training R²: 0.9037, Validation R²: 0.8679
* Validation/Train RMSE Ratio: 1.201

**XGBoost Results**:

* Training RMSE: 938.70, Validation RMSE: 1119.97
* Training R²: 0.9062, Validation R²: 0.8730
* Validation/Train RMSE Ratio: 1.193

#### Interpretation of Results

Both models exhibit strong predictive performance, with XGBoost slightly outperforming Random Forest on validation metrics, as indicated by lower RMSE and a slightly higher R².

Given the better performance on validation data and lower RMSE, XGBoost is selected as the primary model for predicting repayment amounts.

### 5. Feature Importance and Interpretation

#### Correlation Analysis

The top correlated features with `cumulative\_amount\_paid` are `nominal\_contract\_value` (0.880), `cumulative\_amount\_paid\_start` (0.845), and `deposit\_amount` (0.748).

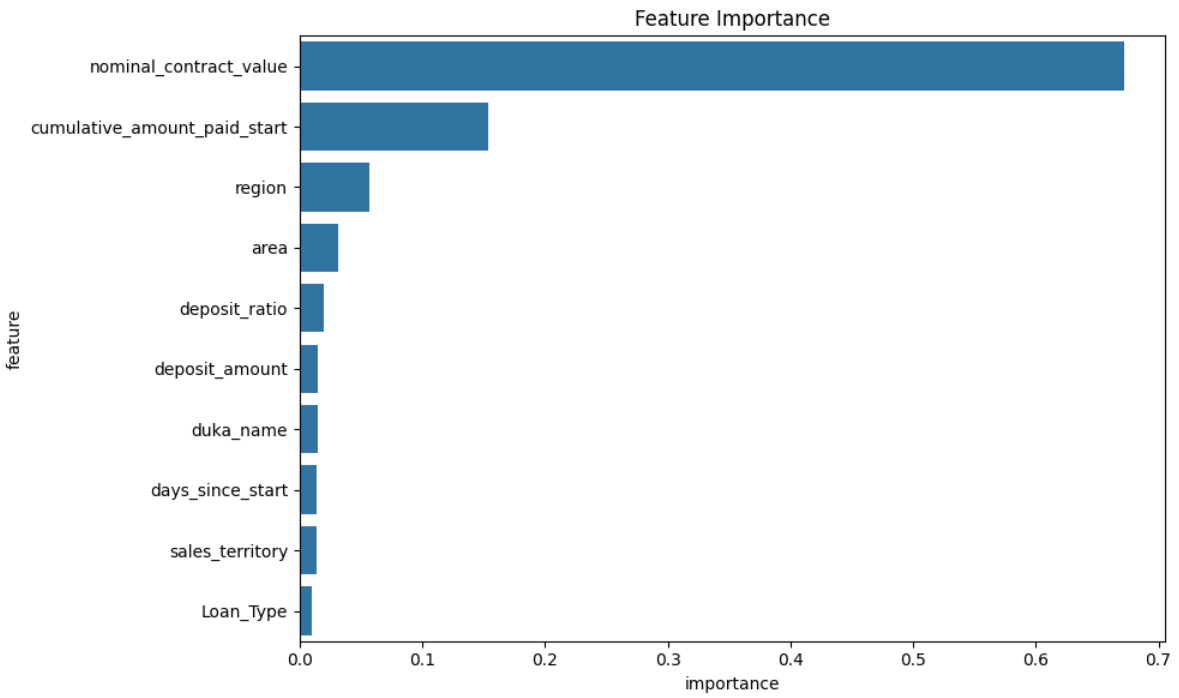
These features represent contract size, initial payment progress, and upfront payment commitment, highlighting their potential impact on repayment likelihood.

#### Feature Importance (from Model Analysis)

##### Top 10 Features in Model

1. nominal\_contract\_value (0.672): Strongest predictor, reflecting the loan size’s influence on repayment amount.
2. cumulative\_amount\_paid\_start (0.153): Indicates early repayment progress, a key predictor of final repayment.
3. region (0.056): Highlights geographical impact on repayment behavior.
4. area (0.031): Demonstrates location-related factors affecting repayment.
5. deposit\_ratio (0.019): Emphasizes the significance of initial deposit on repayment.
6. deposit\_amount (0.015): Important for understanding upfront financial commitment.
7. duka\_name (0.014): Acts as a unique identifier, indirectly capturing client behavior.
8. days\_since\_start (0.014): Reflects repayment pattern over time.
9. sales\_territory (0.013): Indicates territorial influence on repayment.
10. Loan\_Type (0.010): Shows that loan type affects repayment likelihood.

Visualizations: Feature importance plots illustrate these variables’ significance in influencing predictions, underscoring the relationship between initial contract size, deposit ratio, and repayment outcomes.



#### Target Variable Dataset

The model was able to predict the holdout dataset target variable, and it can be [found here](https://drive.google.com/file/d/1p61VFO_FyPxpspMaQ4NmQ19-uK5OhhPA/view?usp=sharing)

### 6. Recommendations for One Acre Fund

Based on our analysis, we recommend the following actions to enhance One Acre Fund’s loan repayment strategies:

#### 1. Repayment Targeting:

Focus on clients with smaller `nominal\_contract\_value` and higher `deposit\_ratio`, as these characteristics correlate positively with repayment likelihood. Early targeting for these groups could support high repayment rates.

#### 2. Enhanced Monitoring for High-Risk Clients:

Clients with low `cumulative\_amount\_paid\_start` or high contract values may require closer monitoring and targeted intervention to ensure repayment progress, potentially through tailored support or reminder systems.

#### 3. Future Data Collection:

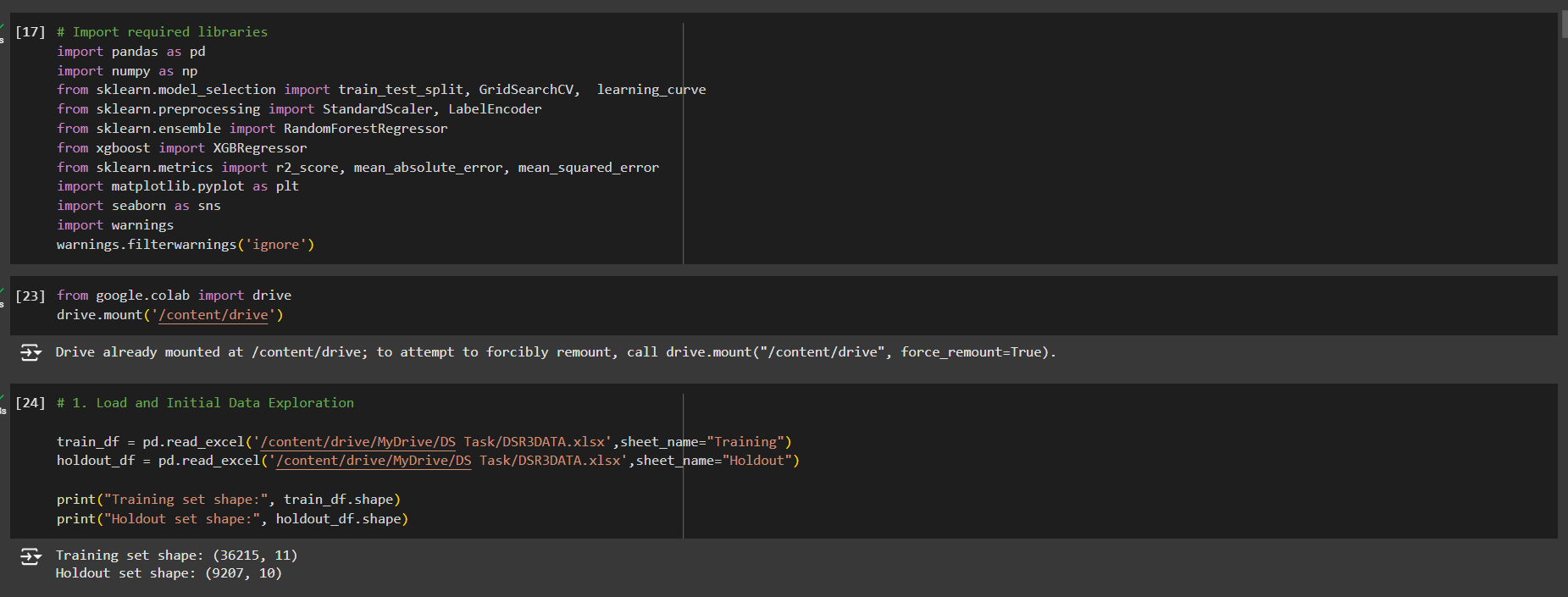
Additional demographic or socio-economic data, such as client age or income levels, may further refine predictive power and provide more comprehensive insights into repayment risks.

#### 4. Geographically-Informed Strategies:

Use ‘region’ and ‘area’ insights to tailor strategies based on regional repayment patterns.

### 7. Code Used to Develop the Model

The following [Python code](https://colab.research.google.com/drive/1iwbX0Jnr8Y-ZQN2BIki0786N7pLuHIYE?usp=sharing) was used to develop the models and evaluate their performance. This includes steps for data preprocessing, feature engineering, model building, and hyperparameter tuning.



The complete code includes comments for reproducibility and insights into each step, from preprocessing to model evaluation.

# Part 2:

## Memo to the Tupande Steering Committee

Date: 2024-11-01

Prepared by: Kevin Odhiambo

**Subject: Lending Strategy Recommendations to Enhance Repayment Rates and Targeting Efficiency**

### Context for the Task

In alignment with Tupande’s objectives to achieve a 98% repayment rate, reduce defaults to 1%, and expand the farmer base through accurate targeting, a predictive model was developed to identify key drivers of repayment behavior among smallholder farmers. By leveraging machine learning, this analysis provides strategic insights into which clients are more likely to meet their repayment obligations, enabling more informed decision-making and resource allocation to support Tupande’s growth and impact.

### Findings from the Model

To support these objectives, we built and evaluated predictive models using Random Forest and XGBoost algorithms. The results below highlight model performance and key findings from the analysis:

#### 1. Model Performance:

**Random Forest:**

* Training RMSE of 950.81
* Validation RMSE of 1142.28
* with a Training R² of 0.9037
* Validation R² of 0.8679

**XGBoost:**

* Training RMSE of 938.70
* Validation RMSE of 1119.97
* with a Training R² of 0.9062
* Validation R² of 0.8730.

XGBoost performed slightly better on validation data, capturing a high proportion of the repayment behavior variance (87.3%), and was selected as the primary model. Both models show consistent validation-to-training RMSE ratios (close to 1.2), indicating that the models generalize well to new data and effectively capture repayment patterns across different client profiles.

#### 2. Top Predictive Features:

* Loan and Payment Amounts: The model identified `nominal\_contract\_value` (loan amount) and `cumulative\_amount\_paid\_start` (initial repayment) as the most significant predictors of repayment likelihood.
* Geographic and Client Characteristics: Features such as `region` and `area` contribute to predicting repayment trends, indicating region-specific repayment behaviors.
* Deposit and Repayment Ratios: Higher initial `deposit\_ratio` (deposit as a percentage of loan value) correlates with higher repayment reliability, suggesting that clients who make larger initial deposits are more likely to repay on time.

### Recommendations

Based on these findings, we recommend the following actions to enhance repayment rates, reduce default risk, and expand the client base:

#### 1. Targeted Client Support Based on Loan and Early Repayment Patterns:

Focus attention on clients with higher `nominal\_contract\_value` and lower `cumulative\_amount\_paid\_start`, as these clients exhibit a higher risk of delayed payments.

Early repayment behaviors (e.g., higher `deposit\_ratio`) can serve as indicators to prioritize clients who may need additional support or interventions to stay on track with repayments.

#### 2. Region-Specific Strategies:

Tailor communication and repayment strategies based on geographic repayment trends. Regions with lower repayment rates may benefit from targeted outreach, payment reminders, or field officer support to improve repayment outcomes.

#### 3. Enhanced Lending Criteria for High-Risk Groups:

Refine eligibility and loan terms based on initial payment behavior and deposit ratios to better target clients who demonstrate a likelihood of full repayment.

### Next Steps for Implementation

To implement these recommendations, we propose the following technical, business, and operational actions:

#### 1. Technical:

Model Deployment: Integrate the XGBoost model into the client management system to enable real-time scoring of repayment likelihood for new and existing clients.

Automated Monitoring: Set up dashboards to track repayment indicators and alert staff to clients at risk, enabling timely, proactive interventions.

#### 2. Business:

Develop Tailored Client Communication: Craft targeted messaging strategies for clients based on risk factors such as loan size, region, and initial deposit ratio.

Adjust Lending Policies: Update lending criteria to focus on clients with high repayment likelihood, informed by the key predictive features identified.

#### 3. Operational:

Enhanced Client Support: Equip field officers with resources and training to assist clients flagged by the model, including tools for budget planning, repayment reminders, and agricultural advice.

Regional Programs: Initiate region-specific programs to address lower repayment trends, with tailored support and community engagement in areas identified by the model.

### Conclusion

These recommendations, derived from predictive insights, can significantly enhance Tupande’s ability to meet its lending objectives. By focusing on key predictive factors, targeting high-potential clients, and adapting support strategies, Tupande can strengthen repayment rates, reduce defaults, and expand its farmer base effectively and sustainably.

Prepared by: Kevin Odhiambo

Position: Data Scientist, Tupande

# Part 3

## Subject: Project Outline for 2-Year Countrywide Scale-Up of Repayment Prediction Model

Hi Tupande,

Following the successful 6-month trial of our repayment prediction model in Lower Western, Kenya, the Steering Committee has approved scaling up the project countrywide. This initiative will be critical for advancing Tupande’s goals of serving more farmers while ensuring a 98% repayment rate. The purpose of this email is to outline your role in planning this expansion, including the project’s context, scope, and the expected final deliverable.

### Project Context

The goal of the scale-up is to expand our lending operations by targeting low-risk farmers across all regions in the country, using our predictive model to maintain a high repayment rate. During the trial, we tailored interventions based on insights from the model, which led to improved repayment rates and reduced default risk. Now, as we move forward with the nationwide plan, your role will involve designing an operationally feasible, data-driven strategy for implementing this model in all target regions over the next two years. This expansion plan will guide the team’s operational efforts, ground coordination, and resource allocation.

### Scope of the Analyst’s Role

Your primary objective is to create a 2-year strategic and operational plan for the nationwide deployment of the repayment prediction model. The plan should detail how to:

* Increase the number of low-risk farmers served each year while maintaining a 98% repayment rate.
* Leverage the model’s insights across diverse regions to identify and support low-risk clients effectively.
* Coordinate resources and roles within our current team and support staff to ensure a seamless rollout.

The plan will require close collaboration with our manager, data analyst, survey enumerators, and government liaison. Your recommendations should balance accuracy with operational feasibility, outlining how we can both deploy and monitor the model’s impact effectively.

### Expected Final Deliverable

The final deliverable will be a comprehensive scale-up plan that addresses the following elements:

#### 1. Project Overview and Goals:

Define the project’s objectives and key performance indicators (KPIs) such as the target number of farmers served, repayment rate, and projected default reduction.

#### 2. Timeline and Phasing:

Map out a 2-year timeline broken down into phases (e.g., initial deployment, mid-point evaluation, full implementation).

For each phase, identify specific goals, milestones, and any region-specific adjustments to the model implementation.

#### 3. Regional Implementation Strategy:

Detail how the predictive model will be integrated into lending workflows in each region, accounting for the unique needs and repayment trends of various locations.

Include strategies for tailoring outreach, client communication, and support based on region-specific insights.

#### 4. Team Roles and Responsibilities:

Define roles for each team member, including the manager, field analyst, survey enumerators, and government liaison, and describe how they will support the scale-up.

Include specific tasks for the field team to ensure efficient monitoring and data collection.

#### 5. Resource Requirements and Budget:

Estimate the financial and operational resources required for each phase, covering model deployment, field support, training sessions, and data monitoring tools.

Identify any additional resources or budget adjustments necessary to sustain this initiative.

#### 6. Monitoring and Evaluation Plan:

Propose metrics and data collection points to assess the model’s effectiveness continuously across regions.

Include a plan for reporting progress, identifying early signs of issues, and refining the model’s criteria if necessary.

#### 7. Risk Management and Mitigation:

Outline potential risks (e.g., client default spikes, resource limitations) and strategies to mitigate these risks, including fallback options or phased implementation adjustments.

#### 8. Deliverable Format:

Please submit the scale-up plan as a structured document or slide deck that clearly covers each section. Include visuals such as timelines, regional maps, and flowcharts to clarify the rollout phases and team responsibilities.

Please reach out if you need clarification on any part of this outline. I’ll be available for regular check-ins as you work on the plan, and we can schedule an initial meeting to discuss your ideas. Thank you for your attention to detail and dedication to this project!

Best regards,

Kevin Odhiambo

Data Scientist, Tupande