Title: Using Machine Learning Models to Forecast Loan Repayment Behavior in Microfinance

1. Introduction

By providing small credit loans, Microfinance Institutions (MFIs) are empowering financially underprivileged individuals in the quickly changing fintech sector of today. Accurately anticipating a customer's loan repayment status is essential to lowering default rates and enhancing service reliability as the use of mobile financial services (MFS) increases.

The goal of this study is to forecast the likelihood that a borrower will return a microloan within the allotted five days. Our goal is to improve credit risk assessment techniques and identify high-risk borrowers by utilizing machine learning algorithms.

1. Problem Statement

Low-income households can get mobile credit services from our client, a telecom-based MFI service provider. The business aims to develop a trustworthy machine learning algorithm that can determine whether a borrower will make timely loan repayments.

The primary goal is to:

* Estimate the likelihood that a specific client will pay back the microloan within five days.
* Utilize these forecasts to minimize defaults and maximize credit distribution.
* Examine aspects to comprehend repayment trends and consumer behavior.

1. Data Collection and Preprocessing

The given dataset, Micro-credit-Data-file.csv, was utilized.

Actions made:

* Missing values were eliminated (df.dropna()).
* LabelEncoder was used to encode categorical characteristics.
* Continuous characteristics were normalized with StandardScaler.
* Using stratified sampling, divide the data into training and testing sets.
* Label in the target column: 1 if repaid, 0 for defaulted

Transformation of features:

* Object column one-hot encoding
* Numerical features using standard scaling

1. Exploratory Data Analysis (EDA)

Important observations:

* There is a correlation between repayment behavior and certain categorical factors, such as gender and occupation.
* As the loan amount rises, the possibility of repayment declines.
* Temporal trends indicate that the time of day and date have an impact on repayment as well.

Visualizations employed:

* Bar charts, box plots, and histograms for both univariate and bivariate research
* Correlation matrix heatmaps

1. Feature Engineering

Developed fresh features:

* Ratio of Loan Amount to Age
* Weekday from the timestamp
* The binary indicator IsWeekend

These characteristics aided in identifying behavior-based patterns and non-linear correlations.

1. Model Selection and Training

We tested more than forty-five models, such as:

* Regression using Logistic
* Trees of Decisions
* Forests at Random
* XGBoost
* LightGBM
* Classifier with HistGradient Boosting
* Uninformed Bayes
* KNN
* Group models (voting, bagging)

Training the model:

* Used stratification and train\_test\_split.
* Scaled data was used to train all models.
* Utilized classification measures to assess the performance of the model

1. Hyperparameter Tuning

Utilized:

* For hyperparameter tweaking, GridSearchCV and RandomizedSearchCV
* Adjusting the hyperparameters significantly improved the performance of XGBoost and LightGBM.

For instance, XGBClassifier parameters adjusted:

* n\_estimators = 100
* max\_depth = 3
* learning\_rate = 0.1

1. Model Evaluation

Measures employed:

* Accuracy
* Recall
* Loss of Logs
* Time of Execution

Top models based on CSV summary performance:

| **Model** | **Accuracy** | **Recall** | **Log Loss** | **Duration (s)** |
| --- | --- | --- | --- | --- |
| HisGradientBoosting | 0.9335 | 0.9791 | 0.17988 | 2.78 |
| XGBClassifier | 0.9347 | 0.9775 | 0.10961 | 1.36 |
| Classifier for LGBM | 0.9283 | 0.9841 | 0.207 | 1.67 |
| DecisionTreeClassifier | 0.9228 | 0.9812 | 0.2201 | 1.76 |

The optimal tradeoff between high precision, high recall, low log loss, and manageable training time was found in the chosen model, XGBClassifier.

1. Feature Importance Analysis

Using the feature importance of XGBoost:

Highlights:

* loan\_amount
* transaction\_time
* type of device
* usage\_category
* Ratio of Loan Amount to Age (engineered)

Meaning:

* Repayment was negatively impacted by larger loan amounts.
* Defaults are more likely to occur with specific device kinds and transaction timings.

1. Business Implications

* Loan approvals can be automated with XGBClassifier.
* It is possible to flag customers for human assessment if their estimated payback possibilities are low.
* Enhances the MFI's return on investment and helps lower risk exposure.
* Increases access to credit in a responsible manner, hence improving financial inclusion.

1. Conclusion and Future Steps

To predict loan repayment behavior, we were able to create a predictive machine learning model.

Highlights:

* 45+ models were trained and assessed.
* The final model (XGBoost) has a recall of 97% and a precision of over 93%.
* The robustness of the model was enhanced by feature engineering.

Restrictions:

* Unbalanced data could be further rectified.
* Behavioral patterns (such as call logs and app usage) that could improve predictions might not be included in the dataset.

Upcoming Projects:

* Implement the model as an API for real-time scoring.
* More behavioral data from telecom sources should be included.
* For a deeper explainability, use SHAP values.

1. Appendices

* Every model training script
* Notebook for feature engineering
* CSV summary of model metrics
* The finished model (XGBClassifier.pkl) was saved