Introduction:

Microfinance Institutions (MFIs) cater to low-income populations, particularly unbanked families in remote areas. Offering services like Group Loans, Agricultural Loans, and Individual Business Loans, MFIs aim to alleviate financial constraints. The emergence of mobile financial services (MFS) has gained traction due to convenience and efficiency, challenging traditional high-touch models. Despite challenges, MFIs play a crucial role in poverty reduction, extending $70 billion in loans to 200 million global clients. Collaborating with a telecom client, a fixed wireless telecommunications network provider, we're exploring the MFS potential. Their budget operator approach, focusing on disruptive innovation, aligns with delivering cost-effective solutions to value-conscious customers.

Problem Statement:

Telecom industries recognize communication's impact on people's lives and aim to serve low-income families in times of need. Collaborating with a Microfinance Institution (MFI), they offer micro-credit on mobile balances, repayable in 5 days. Deferring from repayment within this timeframe designates a consumer as a defaulter. Loan amounts of 5 Indonesian Rupiah require a 6 Rupiah payback, while 10 Rupiah loans necessitate a 12 Rupiah repayment. Leveraging sample data from the client's database, the objective is to build a predictive model that assesses the probability of loan repayment within 5 days. This model aims to enhance customer selection and inform investment decisions for further credit improvement, distinguishing between non-defaulters (Label '1') and defaulters (Label '0'). Basically this is a classification type of problem

Data Collection and Preprocessing:

The entire datasset is provided in a single csv file Micro Credit Data file. The dataset initally consists of 209593 entries with 37 variables. Basically this is a classification type of problem. The given dataset is dirty as it contains many missing value , significant outliers are present .

Exploratory Data Analysis (EDA):

For that extensive data analysis is performed to remove the null values, checking for multi-collinearity , finding out outliers. The columns exhibiting high multicollinearity is subsequently removed . By the help of skewness we understand the direction of outliers present in our dataset and by checking correlation matrix we understood which variables are affecting the price of our target feature positively and negatively.

Based on the descriptive statistics of the dataset, we identify potential outliers by examining the range of values for each column. In the provided statistics, we looked for columns where the difference between the maximum value and the 75th percentile (75% quartile) is much larger compared to the difference between the 75th percentile and the median (50% quartile). This is a preliminary analysis. To confirm the analysis and identify potential outliers in the dataset, we plot box plots and scatter plots.

Box plots will give us a visual representation of the distribution of each column, and scatter plots can help identify extreme data points that might be outliers.

For removing the outliers we replaced them with upper and lower quartile accordingly based on the interquartile range. We then checked again the skewness of the dataset and there is 1 feature that is still exhibiting high skewness is removed. Also the features which have very little or less significance is dropped subsequently.

Feature Engineering:

Our target variable is highly imbalanced as it contains label 1 value way more than label 0. Since this is could make our model more biased to label 1 so I resampled the target feature(\'label\') with the help of random oversampler.

I applied Standardization to transform features to have a mean of 0 and a standard deviation of 1, ensuring all features contribute equally in machine learning models and preventing features with larger scales from dominating the learning process.

Model Selection and Training:

then we split our dataset 80% into training and 20% for testing using train test split . Finally our dataset is ready to be trained on different ML models

ML models used & Model evaluation:

For this problem we applied Logistic Regression, KNN , Decision Tree , Random Forest and Xgb classifier and evaluated our performance for each of these algorithym by the help of evaluation metrics like precision, recall, F1 score, confusion matrix and accuracy,

Conclusion:

Out of all the machine learning models Random Forest algorothym gives the best result with an accuracy of 97.3% in detecting customers who are probable defaultor for loan payment.

Business Implications:

This predictive model holds significant business implications for the telecom industry's collaboration with the Microfinance Institution. By accurately assessing the likelihood of loan repayment within 5 days, the model empowers the business to make informed decisions on customer credit offerings. This, in turn, minimizes default risk, enhances resource allocation, and ensures effective investments in low-income families' financial needs. With precise predictions, the business can tailor their credit strategies, optimize resources, and foster financial inclusion, ultimately contributing to improved financial stability and customer satisfaction.