MICRO CREDIT DEFAULTER - PREDICTION

Submitted by

Steffin Varghese

Batch - 26

*In partial fulfilment of Data Science – Internship*

*At*



Flip Robo Technologies

AI and Software Development company

Flip Robo Technologies | Indiranagar, Bengaluru - 560 038, Karnataka, India

Month of Submission

August 2022

Acknowledgement

I am highly indebted to FlipRobo Technologies for giving me this opportunity to work on a project and for the guidance and persistent supervision, as well as for providing necessary project information and assistance in completing the project.

I would want to convey my gratitude to the members of FlipRobo Technologies for their kind encouragement and support in completing this project.

I would like to extend my heartfelt gratitude and appreciation to SME, Ms **Khushboo Garg** for spending such close attention to me and assisting me during the project's completion, as well as towards others who have volunteered to assist me with their skills.

MACHINE LEARNING FOR PROPERTY PRICE PREDICTION AND

PRICE VALUATION

I acknowledge my gratitude towards the authors of papers: “**Prediction of Loan Defaulters in Microfinance Using Social Network Data**”, “**Predicting Credit Default among Micro Borrowers in Ghana** and “**Strategies for Reducing Microfinance Loan Default in Low-Income Markets”** for the insights I could attain through the extensive research which enhanced my knowledge in development of the project.

**CONTENTS**

**INTRODUCTION**

* Business Problem Framing
* Conceptual Background of the Domain Problem
* Review of Literature
* Motivation for the Problem Undertaken

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem
* Data Sources and their formats
* Data Pre-processing
* Data Inputs- Logic- Output Relationships
* Assumptions related to the problem under consideration
* Hardware and Software Requirements and Tools Used

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* Testing of Identified Approaches (Algorithms) and evaluation of selected models
* Key Metrics for success in solving problem under consideration
* Visualizations
* Interpretation of the Results

**CONCLUSION**

* Key Findings and Conclusions of the Study
* Learning Outcomes of the Study in respect of Data Science
* Limitations of this work and Scope for Future Work

**INTRODUCTION**

**Business Problem Framing**

A Microfinance Institution (MFI) is an organization that offers financial services to low-income populations. Microfinance services (MFS) becomes extremely useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The MFS provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on. Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services.

Though, the MFI industry is primarily focusing on low-income families and are extremely useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.

**Conceptual Background of the Domain Problem**

Today, microfinance is widely accepted as a poverty-reduction tool, representing 70 billion USD in outstanding loans and a global outreach of 200 million clients.

Many developing nations rely heavily on microfinance institutions to support their economies. However, due to the informal nature of the businesses and people they lend money to, many of these microfinance companies struggle with the issue of default.

We are working with one such client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

They understand the importance of communication and how it affects a person’s life, thus, focusing on providing their services and products to low-income families and poor customers that can help them in the need of hour. They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days.

**Review of Literature**

**Prediction of Loan Defaulters in Microfinance Using Social Network Data (2018),** David Murphy

Three main questions were the focus of this study's objectives. The first step was to introduce a relational learning approach and assess whether it was superior to conventional classification approaches in terms of default prediction. Only one of many relational learner techniques was examined, despite the fact that the results indicate that the predictions from the spreading activation model are typically worse than the classification algorithms and do not provide a major operational benefit to the organisation.

Future work may incorporate different relational learners, such as a network only link based classifier when a denser sample set of loans is available, given that network sparsity may also contribute to the subpar outcomes. Additionally, investigating different transfer functions can result in more encouraging conclusions.

**Predicting Credit Default among Micro Borrowers in Ghana (2014),**

Kwame Simpe Ofori, Eli Fianu, Kayode Omoregie, Nii Afotey Odai

One of the significant disadvantages of this laudable initiative has been recognised as micro credit default, which depletes these circulating funds and lowers investor confidence. Since effective countermeasures can be established to avoid and reduce the incidences of default, it is crucial to understand the variables that cause a loan recipient to default. Logistic regression was used in this study to determine the variables connected to the occurrence of microcredit default. Age, gender, gross monthly income, length of employment with current employer, loan amount, loan term, number of dependents, other income, and other deductions were found to be significant predictors of default.

**Strategies for Reducing Microfinance Loan Default in Low-Income Markets (2017)** Patrick Mphaka

The goal of this study was to investigate the methods employed by MFI executives to decrease loan default in the BOP market. According to the study's findings, MFI executives employed a variety of tactics to lower loan default rates, including collaborations, lending to well-established organisations, training in loans and business skills, and routine monitoring. MFI executives also need to place greater emphasis on utilising language that stresses the advantages of loan repayment for prospective borrowers as opposed to language that emphasises the drawbacks of defaulting on loans. The study's conclusions and suggestions provide a list of tactics MFI leaders in low-income economies can employ to decrease loan default and, as a result, increase the profitability and sustainability of their institutions' operations.

**Motivation for the Problem Undertaken**

In microfinance, default refers to a client's inability to pay back a loan. If loan payback rates are high and reliable, MFIs can continue to deploy loans to help achieve the goal of reducing poverty. If loan payback rates are high, MFIs can lower interest rates and processing costs, which will increase loan demand. An increase in the volume of loans disbursed to various economic sectors is sparked by a high repayment rate. A decrease in credit availability is being felt by the agricultural sector as a result of inadequate loan repayment performance. Poor management practises, loan diversion, loan refusal, and other socioeconomic issues are likely to blame for poor loan payback performance.

Many developing nations depend heavily on microfinance institutions to support their economies. However, due to the informal nature of the businesses and people they lend money to, many of these microfinance companies struggle with the issue of default. So, it would be beneficial for MFIs to identify the possible defaulters based on the historical data.

Data Science and Machine Learning techniques would be apt in this case, as it can identify the complex patterns and analyse the historical data to build models that can predict whether a customer will be a defaulter or not based on various factors and information about the loan and the customer’s loan and repayment history.

With the help of statistics and technology or in short, with the help of data science MFIs can build a predictive model that can identify whether a customer will be a defaulter or not based on his previous credit history, repayment patterns and other crucial information about the loan. Thus, MFIs can reduce the occurrence of defaulters up to an extent. This will be helpful for MFis as well as the economy as a whole as it will help the country to maintain a smooth economic flow and proper control over the economic structure.

Since data science and machine learning allows an option for recalibration, MFIs can always reengineer the machine learning model to perform well with the changing trends in this sector. This will help the MFIs to identify the new changing patterns in credit defaulters better than heuristic approaches.

**Analytical Problem Framing**

**Mathematical/ Analytical Modelling of the Problem**

As our client is from Telecom Industry. They understand the importance of communication and how it affects a person’s life, thus, focusing on providing their services and products to low-income families and poor customers that can help them in the need of hour.

They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah).

**Data Sources and their formats**

The sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

**Table

Description automatically generated with medium confidence**

Dataset

* There are no null values in the dataset.
* There may be some customers with no loan history.
* The dataset is imbalanced. Label ‘1’ has approximately 87.5% records, while label ‘0’ has approximately 12.5% records.
* For some features, there may be values which might not be realistic. You may have to observe them and treat them with a suitable explanation.
* You might come across outliers in some features which you need to handle as per your understanding. Keep in mind that data is expensive, and we cannot lose more than 7-8% of the data.
* We have 209593 records in this dataset. We have 209593 rows and 37 columns in the dataset.
* We have string, integer, and float type of data in the dataset.
* We have 209593 non null values in the dataset.

**Description of the dataset**

**Features in Dataset (Independent Variable)**

msisdn - mobile number of users

aon - age on cellular network in days

daily\_decr30 - Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)

daily\_decr90 - Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)

rental30 - Average main account balance over last 30 days

rental90 - Average main account balance over last 90 days

last\_rech\_date\_ma - Number of days till last recharge of main account

last\_rech\_date\_da - Number of days till last recharge of data account

last\_rech\_amt\_ma - Amount of last recharge of main account (in Indonesian Rupiah)

cnt\_ma\_rech30 - Number of times main account got recharged in last 30 days

fr\_ma\_rech30 - Frequency of main account recharged in last 30 days

sumamnt\_ma\_rech30 - Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)

medianamnt\_ma\_rech30 - Median of number of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)

medianmarechprebal30 - Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)

cnt\_ma\_rech90 - Number of times main account got recharged in last 90 days

fr\_ma\_rech90 - Frequency of main account recharged in last 90 days

sumamnt\_ma\_rech90 - Total amount of recharge in main account over last 90 days (in Indonasian Rupiah)

medianamnt\_ma\_rech90 - Median of amount of recharges done in main account over last 90 days at user level (in Indonasian Rupiah)

medianmarechprebal90 - Median of main account balance just before recharge in last 90 days at user level (in Indonasian Rupiah)

cnt\_da\_rech30 - Number of times data account got recharged in last 30 days

fr\_da\_rech30 - Frequency of data account recharged in last 30 days

cnt\_da\_rech90 - Number of times data account got recharged in last 90 days

fr\_da\_rech90 - Frequency of data account recharged in last 90 days

cnt\_loans30 - Number of loans taken by user in last 30 days

amnt\_loans30 - Total amount of loans taken by user in last 30 days

maxamnt\_loans30 - maximum amount of loan taken by the user in last 30 days

medianamnt\_loans30 - Median of amounts of loan taken by the user in last 30 days

cnt\_loans90 - Number of loans taken by user in last 90 days

amnt\_loans90 - Total amount of loans taken by user in last 90 days

maxamnt\_loans90 - maximum amount of loan taken by the user in last 90 days

medianamnt\_loans90 - Median of amounts of loan taken by the user in last 90 days

payback30 - Average payback time in days over last 30 days

payback90 - Average payback time in days over last 90 days

pcircle - telecom circle

pdate - date

**Target in dataset (Dependent Variable)**

label - Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan. Label ‘1’ indicates that the loan has been paid i.e., non-defaulter, while Label ‘0’ indicates that the loan has not been paid i.e., defaulter.

**Data Pre-processing**

* The column 'pcirlce' has only one value for all the records. So, we dropped this column from the dataset.
* The first column is just indexed to identify each records. So, we dropped this column as it is not making any impact on the output.
* We have also dropped the column ‘msisdn’ as it contains the user mobile number.
* As a feature engineering, we have extracted the day and month from the column ‘pdate’ and added them to the dataset as new columns.

**Checking for Missing Values**

Shape, square

Description automatically generated

Heatmap of null values

**Data Cleansing**

**Checking whether the data is balanced or not**

A picture containing logo

Description automatically generated**A picture containing chart

Description automatically generated**

Unbalanced data Balanced Data

The data is not balanced and the records for non-defaulters is provided more in the datatset. So, we used SMOTE oversampling technique to balance the dataset.

Text

Description automatically generated with medium confidence**Removing the Skewness**

Text

Description automatically generated

Before reducing skewness of data

After adjusting skewness using power transform

The columns ['cnt\_da\_rech30', 'fr\_ma\_rech30','fr\_da\_rech30', 'last\_rech\_date\_da', 'last\_rech\_date\_ma', 'fr\_da\_rech90', 'maxamnt\_loans30'] are having high skewness. So, we have dropped these columns from the dataset as these columns can have biased impact on the output.

**Removing the Outliers**

We tried three outlier removal methods on the datasets:

* 1. Using Zscore method
  2. Using Inter-quartile range (IQR)
  3. Using percentile method

Out of which the percentile method was effective as it was not losing any data from the dataset after removing outliers. Since the data is important and we do not want to lose any crucial information from the dataset, we have proceeded with the percentile outlier removal method.

After removing outliers using percentile method, we were able to reduce the outliers up to an extent.

A picture containing application

Description automatically generated

Outliers in dataset using boxplot after using percentile method

**Checked and Removed Multicollinearity from the Datasets.**

Graphical user interface, text, application

Description automatically generated

These columns were having high multicollinearity, so we dropped the columns ['daily\_decr30', 'sumamnt\_ma\_rech30','rental30', 'amnt\_loans30'] so as to reduce the multicollinearity.

Columns with high VIF

**Final Dataset**

Table

Description automatically generated with low confidence

Final Dataset after EDA and pre-processing

**Data Inputs- Logic- Output Relationships**

**Statistical Summary**

**Describe of the data**

Text, table

Description automatically generated with medium confidence

Describe of the data

Table

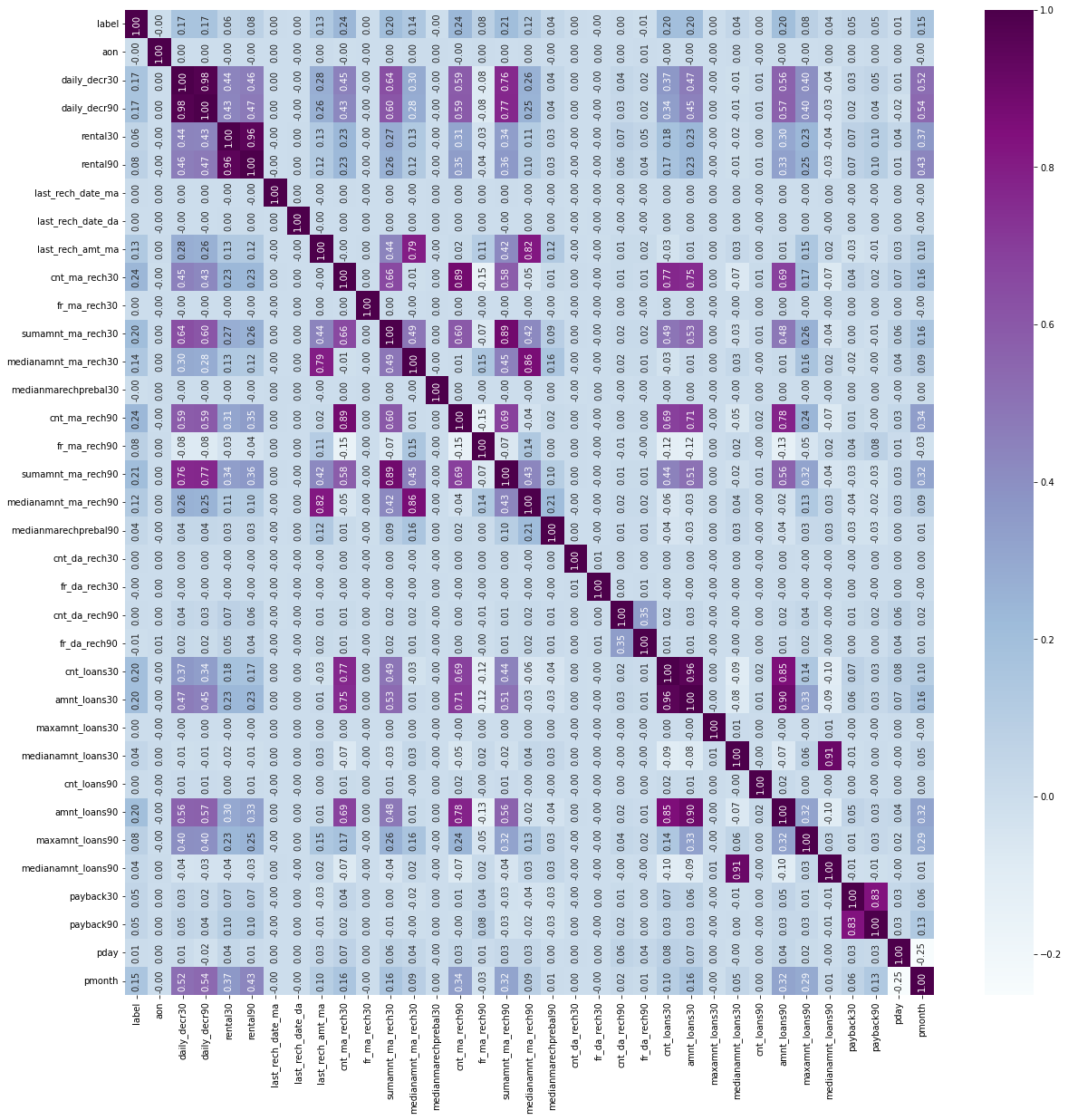
Description automatically generated with medium confidence

Heatmap of describe of data

Observations:

* All the columns except the column 'label' are having higher mean value than the median value. That means the data in these columns are skewed and data is not normally distributed.
* Except the columns ['label', 'medianamnt\_loans30', 'maxamnt\_loans90', 'medianamnt\_loans90', 'pday', 'pmonth'], rest of the columns are having huge difference between the 75% and the maximum value. That means possible outliers are present in the data.

**Correlation**



Correlation of all variables

**Correlation with the Target column**

A picture containing music

Description automatically generated

Correlation of variables with target variable

Observations:

* Most of the columns are having positive correlation to the target variable.
* The columns ['fr\_da\_rech30', 'aon', 'medianmarechprebal30', 'fr\_da\_rech90'] are having negative correlation to the target variable 'label'. Rest of the variables are having positive correlation to the target variable.
* The column 'cnt\_ma\_rech30' is having highest positive correlation to the target variable 'label', while the column 'maxamnt\_loans30' is having least positive correlation to the target variable 'label'.
* The column 'fr\_da\_rech30' is having the least negative correlation to the target variable 'label', while the column 'fr\_da\_rech90' is having the highest negative correlation to the target variable 'label'.

**Assumptions Related to the Problem Under Consideration**

**Distribution of data in columns (Checking skewness of data)**

A picture containing window, shoji, building

Description automatically generated

Skewness of Data

Observations:

* None of the columns are having normal distribution of data.
* All the columns except the columns ['pday', 'pmonth'] are having skewness and they are right skewed.

**Presence of Outliers in Data**

A picture containing shoji

Description automatically generated

Outliers in dataset using boxplot

Observations:

* All the columns except the columns ['pday', 'pmonth'] are having extreme outliers. We can handle these outliers in later step.

**Hardware and Software Requirements and Tools Used**

**Hardware Requirement:**

**Graphical user interface, text, application

Description automatically generated**

Hardware Configuration

**Software Requirements:**

* Windows Version: Windows 10 Pro
* Anaconda Navigator: 2.0.3

Anaconda offers the easiest way to perform Python/R data science and machine learning on a single machine.

* Jupyter Notebook: 6.3.0

Anaconda offers the easiest way to perform Python/R data science and machine learning on a single machine.

* Python3: Python 3.9.9

Python3 is used as the base environment for performing the machine learning and data analysis.

Python Libraries Used:

* Pandas: Data manipulation and analysis
* NumPy: Adding support for large, multi-dimensional arrays and matrices, along with an enormous collection of high-level mathematical functions to operate on these arrays.
* Matplotlib, Seaborn: For visualization of variable relations and data distribution, and analysis.
* Sklearn: Simple and efficient tools for predictive data analysis.
* SciPy: SciPy provides algorithms for optimization, integration, interpolation, eigenvalue problems, algebraic equations, differential equations, statistics, and many other classes of problems.
* Statsmodels: Provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration.
* Xgboost, catboost, lightgbm: Gradient boosting framework that uses tree-based learning algorithms.
* Pickle: Implements binary protocols for serializing and de-serializing

**Model/s Development and Evaluation**

**Identification of possible problem-solving approaches (methods)**

Text

Description automatically generated

Importing the required libraries for model building

Text

Description automatically generated

Creating the instances of algorithms

We created three functions for testing the model and for cross validations:

* best\_ran: Finding the best random state for the selected model
* mod\_test: Training the model with the train data using the best random state.
* cross\_val: Finding the best cross validation mean score for each model.

**Testing of Identified Approaches (Algorithms) and evaluation of selected models**

Graphical user interface, text, application

Description automatically generated

Code Snippet for function to find best random state

Graphical user interface, text, application

Description automatically generated

Code Snippet for function to test the model

* Logistic Regression

Table

Description automatically generated with medium confidence

Testing the model performance

* DecisionTree Classifier

Table

Description automatically generated with low confidence

Testing the model performance

* XGB Classifier

Table

Description automatically generated with medium confidence

Testing the model performance

* LGBM Classifier

Table

Description automatically generated with medium confidence

Testing the model performance

* CatBoost Classifier

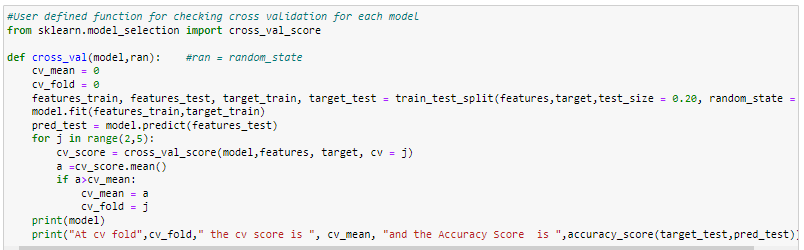
Table

Description automatically generated with medium confidence

Testing the model performance

After testing the models, the xgb classifier(xgb) is performing well by providing the maximum accuracy score of 93.55%. After testing the models, the xgb classifier(xgb) is performing well by providing the maximum accuracy score of 93.55%.

**Cross Validation**



Code Snippet for function to find the cross validation mean score

* LogisticRegression

At cv fold 4 the cv score is 0.7816590794547977 and the Accuracy Score is 0.7859839450479059

* DecisionTree Classifier

At cv fold 4 the cv score is 0.8985968400822768 and the Accuracy Score is 0.9066005206274788

* XGB Classifier

At cv fold 4 the cv score is 0.9258223768870628 and the Accuracy Score is 0.9355212407833945

* LGBM Classifier

At cv fold 3 the cv score is 0.9197056224400632 and the Accuracy Score is 0.9272211849045289

* CatBoost Classifier

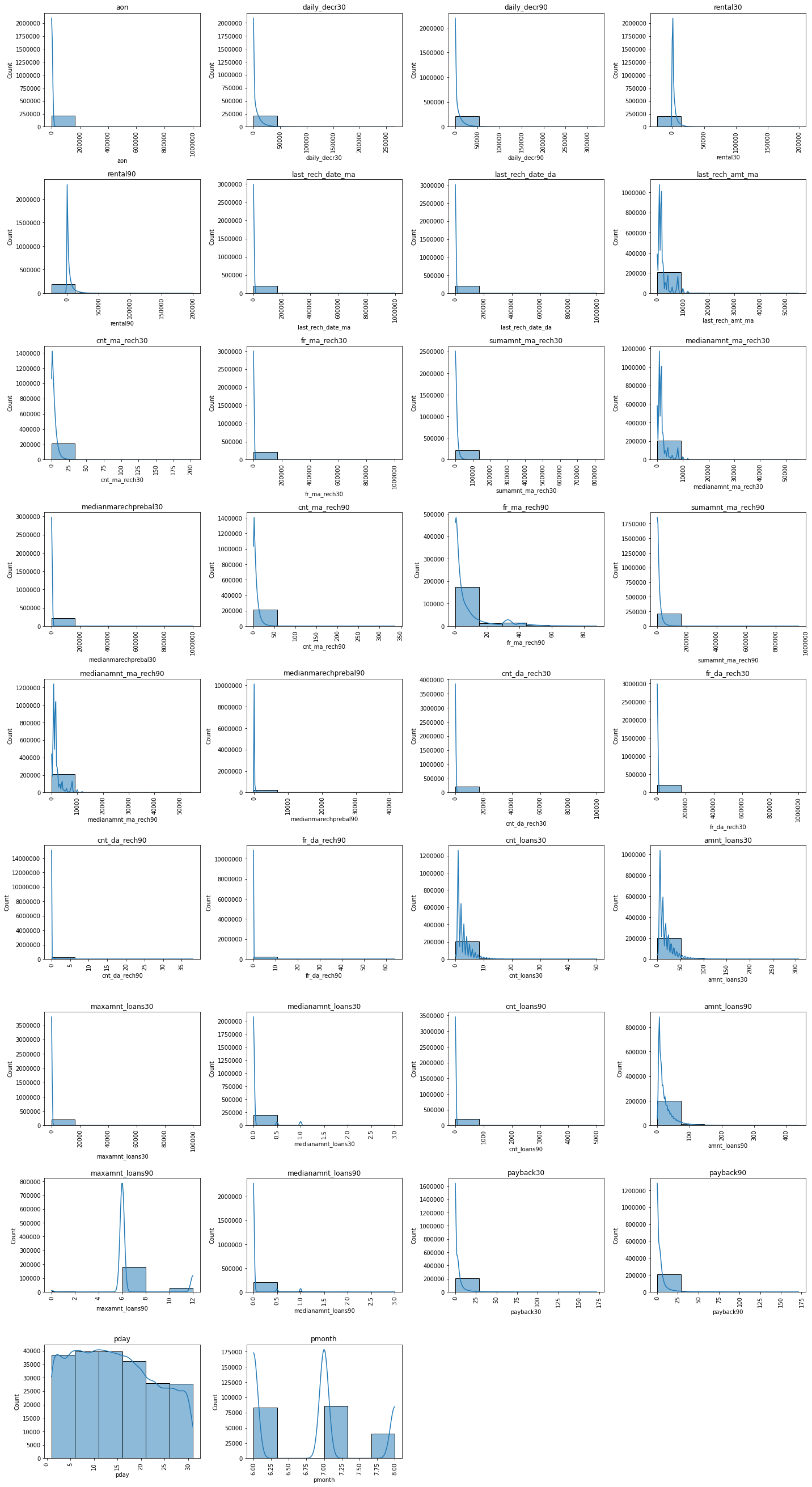
At cv fold 4 the cv score is 0.9243395300600212 and the Accuracy Score is 0.9334632630531667

Key Metrics for success in solving problem under consideration

* Accuracy Score: Metric that summarizes the performance of a classification model as the number of correct predictions divided by the total number of predictions.
* Confusion Matrix: Performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.
* Classification Report: Displays your model’s precision, recall, F1 score and support. It provides a better understanding of the overall performance of our trained model.

**Visualization**

**Univariate Analysis**



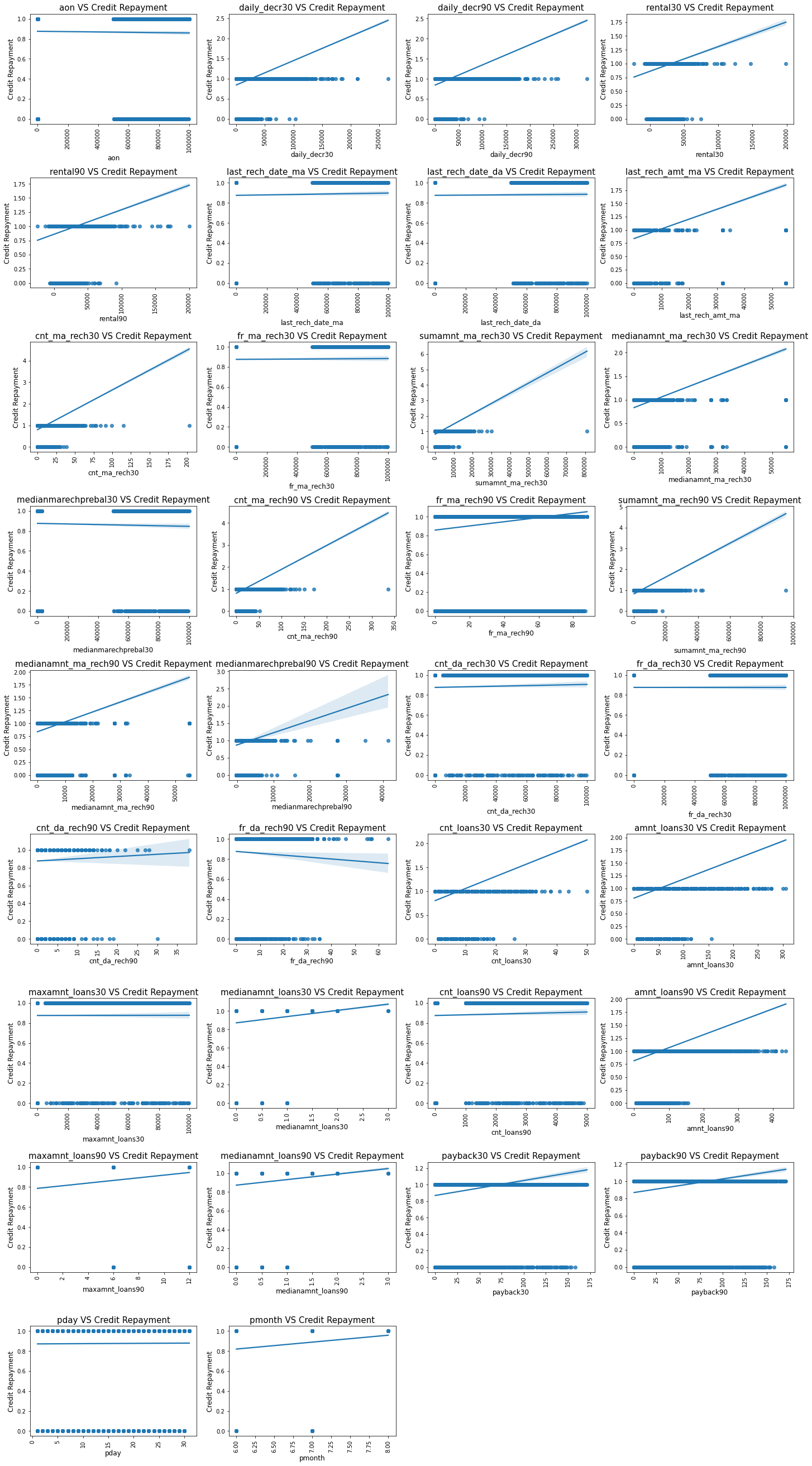
A picture containing chart

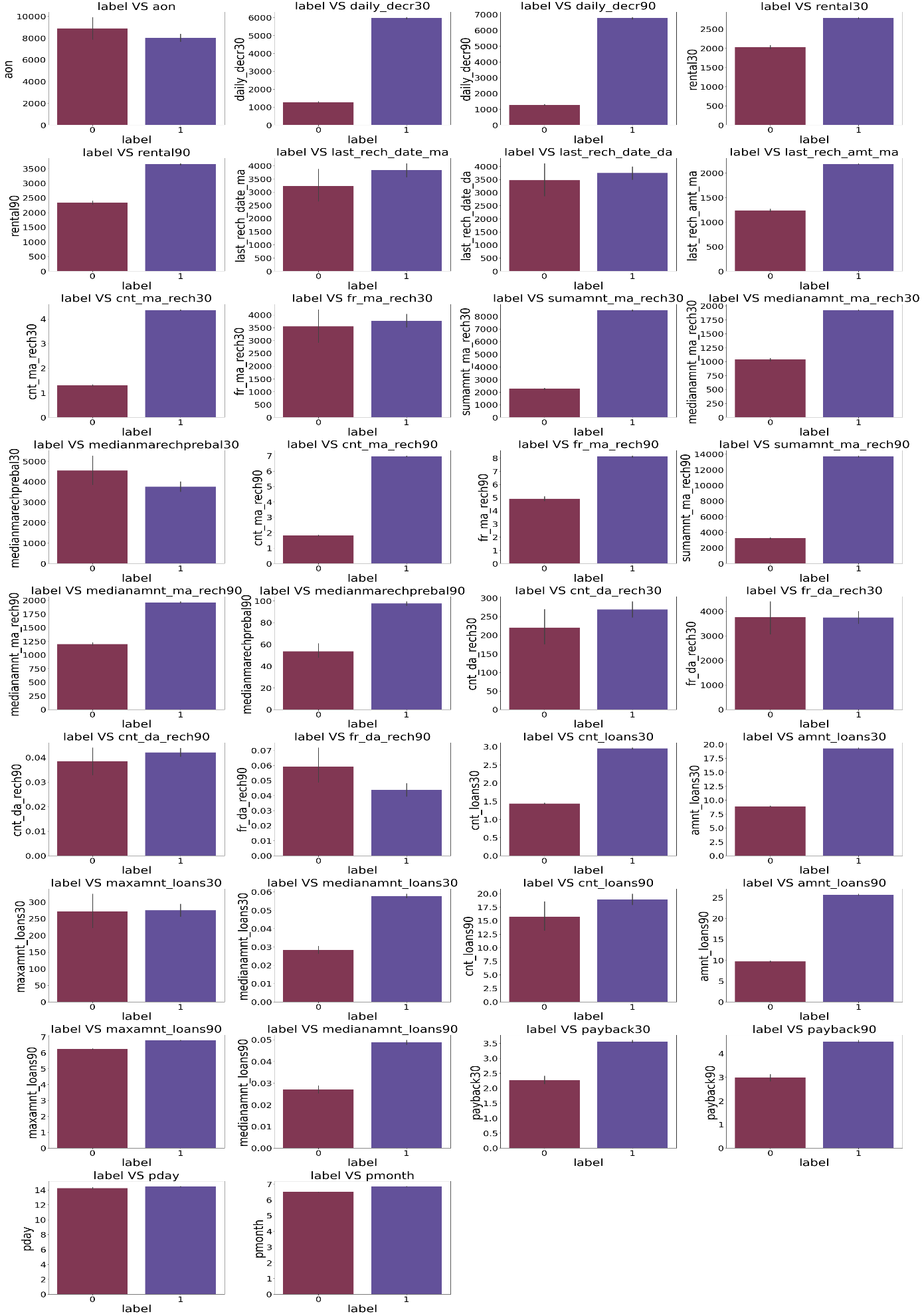
Description automatically generated

Observations:

* The average age on cellular network in days for customers ranging between -48 to 1.7 lakh.
* Most customers have spent less than Rs. 44,000 from main account averaged over last 30 and spent more than Rs. 53,000 over last 90 days.
* The average main account balance of most of customers are ranging between Rs. -23000 to Rs. 14000 over last 30 days and Rs. -25000 to Rs. 13000 over last 90 days.
* The average number of days from recharge of main account and data account for most of customers ranging between -29 to 1.67 lakh.
* Most customers have recharged their main account for an amount ranging between Rs. 0 to Rs. 9000.
* The number of times customers recharged their main account in last 30 days is ranging between 0 to 34 times and in last 90 days is ranging between 0 to 56 times
* Frequency of main account recharge for most of the customers in last 30 days is ranging between 0 to 1.67 lakh and in last 90 days is ranging between 0 to 14.67.
* Total amount of recharge for most of the customers for their main account in last 30 days is between the range Rs. 0 to Rs. 1.35 lakh and the median amount is between the range Rs. 0 to Rs. 9000. While the total amount of recharge for their main account in last 90 days is ranging between Rs. 0 to Rs. 1.59 lakh and the median amount of recharge is between Rs. 0 to Rs. 9000.
* The median main balance of most of customers just before recharge in last 30 days is between Rs-200 to Rs. 1.67 Lakh and in last 90 days is between Rs. 200 to Rs. 6700.
* Most of the customers have recharged their data account between 0 to 16000 times in last 30 days and 0 to 6.33 times in last 90 days.
* The frequency of recharge for most of customers for their data account is ranging from 0 to 1.67 lakh times in last 30 days and 0 to 11 times in last 90 days.
* Most customers have recharged their data account between 0 to 1.67 lakh times in last 30 days and 0 to 6 times in last 90 days.
* Number of loans taken by most customers is between 0 to 8 times in last 30 days and 0 to 832 times in last 90 days
* Total amount of loan taken by most customers is ranging between Rs. 0 to Rs. 51 in last 30 days and Rs. 73 in last 90 days.
* Maximum amount of loan taken by most customers is between Rs. 0 to Rs. 17000 in last 30 days and Rs. 2 in last 90 days.
* Median amount of loan taken by most customers is between Rs. 0 to Rs. 0.5 in last 30 days and Rs. 0.5 in last 90 days
* Average payback time for most customers is between the range 0 to 28 days over last 30 and 90 days.
* The data was collected for three months: - June, July, and August. The data is uniform for most of the days, but a greater number of data is collected from the beginning of the month, and it is reducing towards the end of month.
* Most of the customers have paid back the credit amount within 5 days of issuing loan amount.

**Bivariate Analysis**

****

****

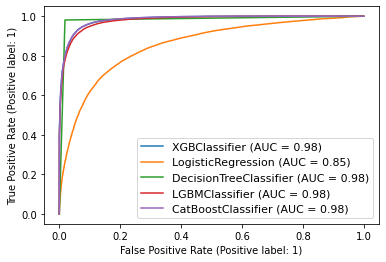
Observations:

* Age of cellular network in days is not having much impact on the loan defaulters. The number of credit defaulters are slightly higher if the age of cellular network is higher.
* Customers who have least amount of spent from main account over last 30 and 90 days are more likely to be a defaulter.
* Customers who maintain low average main account balance over last 30 and 90 days are more likely to be a defaulter.
* Customers who had a greater number of recharges of main account and data account in past 30 and 90 days are not likely to be a defaulter.
* Customers who have recharged main account for higher amount have made the repayment of credit on time.
* Customers who are having frequent recharges in last 30 and 90 days are less likely to be a defaulter.
* Customers whose total amount and median amount of recharge in last 30 and 90 days is lower are more likely to be a defaulter.
* Customers who have higher median main balance just before last recharge in 30 days are more likely to be a defaulter, while the customers who have higher median main balance just before last recharge in 90 days are less likely to be a defaulter.
* Customers who have higher number of data account recharges in last 30 and 90 days are less likely to be a defaulter.
* Customers who have higher frequency in recharge of data account in last 90 days are more likely to be a defaulter, while the frequency of recharges of data account in last 30 days is not having much impact on identifying whether the customer will be a defaulter or not.
* Most number of customers who have taken most number of loans and higher amount of loan over the last 30 days and 90 days are less likely to be a defaulter.
* Maximum amount taken by customers as loan over the last 30 and 90 days are not providing much information on whether the customer would be a defaulter or not.
* Customers who have higher median loan amount over the last 30 and 90 days are less likely to be defaulter.
* Customers who payback the amount in more days over the last 30 and 90 days are less likely to be a defaulter.
* The day and month of which the data was collected is not having much impact on the identifying whether the customer would be a defaulter or not.

**Interpretation of the Results**

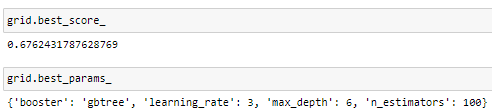
* Most of the columns are having positive correlation to the target variable.
* The columns ['fr\_da\_rech30', 'aon', 'medianmarechprebal30', 'fr\_da\_rech90'] are having negative correlation to the target variable 'label'. Rest of the variables are having positive correlation to the target variable.
* The column 'cnt\_ma\_rech30' is having highest positive correlation to the target variable 'label', while the column 'maxamnt\_loans30' is having least positive correlation to the target variable 'label'.
* The column 'fr\_da\_rech30' is having the least negative correlation to the target variable 'label', while the column 'fr\_da\_rech90' is having the highest negative correlation to the target variable 'label'.

**AUC ROC Curve**

****

**All the models except logistic regression model is providing the maximum AUC ROC score. Since the XGB model is performing well with all the training, testing, cross validations and auc score, we can consider this model as the best performing model.**

**Hyperparameter Tuning**

****

Best tuning score and best tuning parameters

After all the tests, cross validations and tuning the XGB classifier(xgb) model is performing well and providing maximum accuracy score of 93.55%.

**Finalized Model Performance**

**Chart, treemap chart

Description automatically generatedTable

Description automatically generated with low confidence**Final Model Performance

Now we have trained our model and it is ready to test with the actual data to cross verify the performance.

Table

Description automatically generatedChart, line chart

Description automatically generated

Model Predictions and Actual Label Model performance with predictions vs actual label

Our model is performing well with predictions and provided accurate results.

**Saving the best model**

Graphical user interface, text

Description automatically generated

We have saved the machine learning model for future predictions. We have serialized and saved the binary file as “micro credit defaulter prediction\_model.pkl” using the pickle library.

**CONCLUSION**

**Key Findings and Conclusions of the Study**

With the help of data science and machine learning, we were able to create a machine learning model using XGBoost algorithm, which can predict whether the customer will be a defaulter or not.

Now this model can be used to predict whether a customer will be a defaulter or non-defaulter of credit repayment based on the following features:

cnt\_ma\_rech90  amnt\_loans90  cnt\_ma\_rech30

sumamnt\_ma\_rech90  medianamnt\_ma\_rech90  cnt\_loans30

last\_rech\_amt\_ma  cnt\_loans90  medianamnt\_loans90

medianamnt\_loans30  payback30 payback90  medianamnt\_ma\_rech30

daily\_decr90  pmonth  medianmarechprebal90

medianmarechprebal30  rental90 maxamnt\_loans90  fr\_ma\_rech90

pday  cnt\_da\_rech90  aon

**Impact of Variables on Target Variable (Correlation)**

* Most of the columns are having positive correlation to the target variable.
* The columns ['fr\_da\_rech30', 'aon', 'medianmarechprebal30', 'fr\_da\_rech90'] are having negative correlation to the target variable 'label'. Rest of the variables are having positive correlation to the target variable.
* The column 'cnt\_ma\_rech30' is having highest positive correlation to the target variable 'label', while the column 'maxamnt\_loans30' is having least positive correlation to the target variable 'label'.
* The column 'fr\_da\_rech30' is having the least negative correlation to the target variable 'label', while the column 'fr\_da\_rech90' is having the highest negative correlation to the target variable 'label'.

**Learning Outcomes of the Study in respect of Data Science**

Artificial intelligence (AI) and machine learning are not new, but microcredit organisations are reluctant to incorporate these techniques into their process for determining credit risk. Instead, they continue to use traditional credit scoring techniques that are based on the linear calculation of a limited number of indicators. Results from this scoring model are inconsistent and unreliable. On the other hand, machine learning provides a far broader perspective of a client and can be used to control not only credit risk but also other company hazards.

In order to improve predictions for other risks as well as credit risks, such as the potential for early repayment leading to interest income losses, the potential for money laundering, etc., machine learning algorithms enable the creation of new models using historical anonymized data that has already been collected. Financial institutions might use a decent model to forecast the likelihood that a client will return the loan before the maturity date and then set up procedures with pre-set preventive actions to be taken, prior to this occurring.

**Limitations of this work and Scope for Future Work**

**Limitations**

* The data was vast, and the client tried to include all the important information that would benefit the model building. But the data was noisy, and we had to left out some part of data during the pre-processing.
* Most of the features were having skewed data and huge outliers and high multicollinearity between the variables. This has resulted in complex model.
* Even though the data included many crucial features that had impact on the output, there are still many other factors that influence the output such as the financial freedom of the customer, their spending behaviour, economic condition, inflation etc.

**Scope**

* Future work may incorporate different relational learners, such as a network only link based classifier when a denser sample set of loans is available, given that network sparsity may also contribute to the substandard outcomes.
* We can also consider including less noisy data as it will be helpful for the model to easily understand the patterns impacting the outcomes and also it will not lose much data during pre-processing, thus we can preserve the true nature of the data throughout the model building.
* This study has only focused on whether the customer would be a defaulter or not, but there are other important inferences which can be beneficial to the company such as predicting the time of repayment of loan for each customer, chances of early settlement of loan etc.

**Thank You**