# Interpretability of Machine Learning Models in Credit Risk Assessment : A Comparison of Methods

*Isha Choudhary | shahooda637@gmail.com*

## ABSTRACT

Financial sector and financial firms are working with cash and money all the time. The complete industry and the business models are based on the continuous cash flows among them and the customers. With the involvement of cash and money rotation, comes the credit risk. It is really important to perform the safe and informed credit risk assessment from time to time to maintain a healthy credit flow throughout the institute and among the customers. With the advancement of technology like artificial intelligence and machine learning, the process of credit risk assessment and beforehand prediction of potential credit risks are now possible. The possibility of knowing the possible credit risks approaching the firms or companies is introducing a new way of processing the financial tasks. The automation and enables the real time assessment of the transactional data of the customers and understanding of potential risks in near future. This research project aims to understand the dataset collected by the financial institutes and the importance of financial features in calculating and assessing the credit risks. In addition to this, this research project presents a machine learning based solution for automated credit risk assessment for the financial institutes and firm which is backed by an effective explainable AI model enabling the interpretability of the machine learning models and the predictions providing by them. The best predictive performance of the machine learning models SVM, Logistic Regression and Decision Tree is recorded at 99% with the value of f1-score as 1.00. The feature influencing this prediction most as recorded by the explainable AI are the basic financial details of the customers. This study aims to explain the importance of the domain knowledge and human-in-the-loop decision making process in finance sector when using the machine learning model for credit risk assessment.

## I. INTRODUCTION

Financial Sector institutes like banks and insurance companies work with money rotation directly. Growth and development of an economy is dependent of the finance and banking system (G. Attigeri et al.,2019) . The cash flow is the main business model in the finance industry introducing the presence of an all-time credit risk. There are several types of credit risks present in the finance sector as default risk, Institutional risk, concentration risk or more, and evaluating these risks beforehand is an important task for safe functioning of all the financial institutes. There has been increased concern about the continued deterioration in the asset quality of the public sector financial institutes (D. Kadanda et al., 2018). A systematic identification, awareness and assessment of parameters is essential for early prediction of default behaviour(G. Attigeri et al.,2019). In last decade, technology advancement have equipped several sectors with automation through artificial intelligence and machine learning. Incorporating this technology into finance sector is the urgent need for the safe functioning of banks and other finance institutes. Non-recovery or partial recovery of loans has an impact on the bank's balance sheet and income statement items in the form of reduction in interest earned on loan assets, increase in provision on NPAs, increase in capital requirement and lower profits. Hence, rising NPAs are a concern for a bank and determinants of NPAs should be identified prior to loans turning into NPAs (J. K. Bawa et al., 2016). This research project proposes a machine learning based solution for performing credit risk assessment. Along with presenting the machine learning based solution for the credit risk assessment issues in the financial sector, this study presents the interpretability of these machine learning models and how the domain knowledge and human involvement combines with the interpretability of the machine learning models and evaluate the predictions presented by the machine learning models. In this project, we are using the dataset from the world bank to perform the model training and visualize the explainable AI graphs for the predictions presented by the machine learning models.

### A. Problem Statement

Credit risk assessment is an important task in the finance sector and with the advancement of technology across all industries including the finance, it is attempted to make the task of credit risk assessment automated with the help of artificial intelligence and machine learning models. Many studies have been presented machine learning based models delivering high accuracies in predicting the NPA loans beforehand and suggesting possible credit risks in near future by analysing some important financial parameters. However, these models' processing and the way of evaluating the risks using these parameters is done by the black-box machine learning models which do not provide the interpretability of the prediction presented by these models. This research project presents a machine learning based solution for credit risk assessment along with most efficient interpretability models for analysing how the machine learning models are processing the parameters to evaluate the result. Through the research, we aim to explain the importance of domain knowledge and human-in-the-loop decision making system to use these smart credit risk assessment systems effectively in analysing different types of credit risks in the finance sector.

### B. Aim and Objective

This research project aims to present a machine learning based solution for the credit risk assessment in the financial sector to create a predictive model to predict the potential credit risks for the financial institutes like banks. The solution presented in this research project is aimed to address the current challenges in the credit risk assessment predictive modelling like interpretability and explainability of the finance or credit risk predictive models and creating a human-in-the-loop solution to ensure the trust and authenticity on the predictive system. The potential objectives for this research study on the application of interpretability methods in credit risk assessment in finance can be listed below:

- To identify the most effective interpretability methods for credit risk assessment in finance, in terms of balancing interpretability and accuracy.
- TO understand how the interpretability and accuracy trade-offs differ across different credit risk assessment tasks in finance.
- To investigate the impact of domain knowledge on the interpretability of credit risk assessment models in finance.
- To develop a framework for incorporating human-in-the-loop decision making into credit risk assessment using interpretable models in finance.

### C. Research Questions

The research questions this research project aims to answer are given below:

RQ.1 What are the most effective interpretability methods for credit risk assessment in finance, in terms of balancing interpretability and accuracy?
RQ.2 How do the interpretability and accuracy trade-offs differ across different credit risk assessment tasks in finance?
RQ.3 How does domain knowledge impact the interpretability of credit risk assessment models in finance?
RQ.4 How can human-in-the-loop decision making be effectively incorporated into credit risk assessment using interpretable models in finance?

### D. Scope

The scope of this project includes the exploratory data analysis and data pre-processing of the global findex dataset for 123 countries provided by the world bank. Along with analysing and pre-processing this dataset, this research project includes the training of classification machine learning models using this pre-processed dataset and analysing the predictive power and test accuracies of different machine learning models. The interpretability of the processing of these classification machine learning models is an important aspect of this research project through which we try to explain the need and importance of domain knowledge and human-in-the-loop decision making process in the credit risk assessment in financial sector.

## II. LITERATURE REVIEW

Credit risk management and risk prediction is a crucial task for financial institutes like a bank and doing this using Artificial intelligence and machine learning techniques is a new and modern way to provide predictive

security against credit risks. The main concern in using the AI and machine learning techniques in predicting the risks and using these smart systems in the banks for automating various tasks is the interpretability and explain ability of these systems. Humans have a tendency to look for "Why" every time they experience an unexpected "What" (Christoph, 2022). The machine learning based models and predictive systems' results are trustworthy if they provide the reasoning of the presented results. The reasoning is not necessary for all the machine learning based models but in financial sector, results of these machines and algorithms are directly impactful on human life and thus human interpretability is important in these systems.

## A. Overview of Interpretability Methods

Machine learning models take on real world tasks and problems that require safety measure and testing(Christoph, 2022). These machine learning models are predicting various things and based on these predictions, the actions are taken by the machines. In order to assure the actions to be accurate and correct as per the real-world rules and understandings we must assure that the predictions are correct and unbiased. It might happen that the machine learning model we have trained for automatic approval or rejection of credit applications discriminated against the minority that has been historically disenfranchised (Christoph, 2022). Our main goal is to grant loans only to people who will eventually repay the loan amount. The incomplete problem formulation in this case lies in the fact that we want to minimize the loan defaults and at the same time are also not obliged to discriminate on the basis of certain demographics. This extended constraint of our problem formulation is missed by the machine learning model and thus the predictions are biased. In such cases, the interpretability of the machine learning models is the key to understand the reason behind the results presented by the machine learning model and use it to optimize the algorithm if and when needed.

Methods of machine learning interpretability can be classified according to various criteria, and there are various interpretation methods roughly differentiated according to their results as Feature summary statistic, Feature summary visualization, Model internals ( learned weight etc.), Data point, and intrinsically interpretable model (Christoph, 2022). Other than these classifications we have the model specific and model-agnostic interpretation tools to interpret the machine learning models. Model specific tools are limited to specific model classes whereas model-agnostic tools can be used on any machine learning model and are applied after the model has been trained (Christoph, 2022).

Talking about the Explainable AI and its opportunities, Alejandro et al. (2019) says that artificial intelligence lies at the core of many activity sectors that have embraced new information technology, including the financial sector. The sophistication of AI powered systems has increased to such an extent that almost no human intervention is required for their design and deployment and when decisions derived from such systems ultimately affect humans' lives such as in finance or healthcare, there is an emerging need for understanding how such decisions are furnished by AI methods (Alejandro et al., 2019). As black box machine learning models are increasingly being employed to make important predictions in critical contexts, the demand for transparency is increasing from the various stakeholders in AI (Alejandro et al., 2019). Explanations supporting the model outcomes are crucial in the sectors where the ML models' predictions are impacting the human lives directly as in healthcare and finance, says Alejandro et al. (2019). When developing an ML model, the consideration of interpretability as an additional design driver can improve its implement ability for 3 reasons (Alejandro et al. 2019):

- Interpretability helps ensure impartiality in decision making, that is to detect and consequently correct from bias in the training dataset.
- Interpretability facilitates the provision of robustness by highlighting potential adversarial perturbations that could change the prediction.
- Interpretability can act as an insurance that only meaningful variables infer the output, i.e., guaranteeing that an underlying truthful causality exists in the model reasoning.

Explaining the current state of explainable AI, Frank et al. (2020) says that it is important to know how explainable AI is currently defined. In simple words, one would like to have explanations of internal decisions within an AI system that lead to an external result. Such explanations should provide insight into the rationale the AI used to draw conclusions. Frank et al. (2020) believes that the goals of explainable AI are trustworthiness, causality, transferability, informativeness, confidence, fairness, accessibility, interactivity and privacy

awareness. This kind of interpretability is very important in the areas where the model predictions and actions are directly connected to and impacting the lives of human beings. Finance sector being an influencing sector to humans, requires the interpretability of AI and ML systems used to ensure the confidence and fairness of the predictions made. Interpretability in the context of finance is a concern for the stakeholders.

### B. Interpretability in Context of Finance

Finance sector involves a continuous money rotation and public dealing. The use of machine learning and artificial intelligence models in this sector for predicting the credit risk or approving or rejecting the loans requires high level of confidence and trust on the processing of the model as the results and decisions of the machine learning models impact the lives of the customers directly. Finance sector provides great opportunities to enhance customer experience, democratize financial services, ensure consumer protection, and significantly improve the credit risk management (Branka et al., 2021). Today, it is easier than ever to run state-of-art machine learning models, designing and implementing systems that support real-world finance applications have been challenging because they lack transparency and explain ability which are important factors in establishing reliable technology in credit risk management (Branka et al., 2021). In finance, the models are created directly from data by an algorithm it is often very difficult to trace back the steps the algorithm took to arrive at its decision which is a crucial part to check if the presented result is acceptable or not. This is making the use of current state-of-art machine learning and artificial intelligence models in the finance sector. It is crucial to have explanations supporting the output of a model and improvement in the understanding of the system (Branka et al., 2021). Thus, interpretability and explainability are two important components of the machine learning models used in credit risk management. Branka et al. (2021) says that there is a fine line between interpretability and explainability, interpretability is the extent to which cause, and effect can be observed within a system. It is about being able to discern the mechanics without necessarily knowing the "why", whereas explainability is the extent to which the internal mechanics of a ML/DL system can be explained in human terms.

Talking about the ever-growing achievements in AI and recently boosted enthusiasm in financial technology, Lara Marie et al. (2020) shares their opinion on applications of machine learning and artificial intelligence in credit scoring and other financial processes. Machine learning models for Credit Scoring are the decision models that help lenders decide whether to accept a loan application based on the model's expectation of the applicant being capable or not of repaying the financial obligations (Lara et al., 2020). Lara (et al. 2020) believes that such models are beneficial since they reduce the time needed for the loan approval process, allow loan officers to concentrate on a selected loan application, leading to cost saving and reduced human subjectivity and decrease default risk. Using such predictive models in finance sector to automate the tasks using the machine learning methods which might be exceptional but are also known as black-box methods. Thus, it is highly unlikely that any financial expert is ready to trust the predictions of a model without any sort of justification (Lara et al., 2020). Model explainability has thus gained attention and is an emerging area of explainable AI (XAI), a concept focusing on opening the black-box models in order to improve the understanding of the logic behind the predictions(Lara et al., 2020). But in this also, there are number of challenges posed when working with XAI in finance like "who are the explanations for, experts or users?", or "what is the best form of representation for the explanations?" and "how can we evaluate the results?" (Lara et al., 2020). Explaining the evaluation process of XAI part of a ML model project, Lara (et al., 2020) says that Application-grounded analysis requires human domain experts to quantify the correctness and quality of the explanations provided by performing real tasks. In credit scoring loan officers are considered experts in the area since they have comprehensive knowledge of loan requirements and banking regulations.

The credit risk assessment using the machine learning and deep learning models and predictive models are created. These models are further used as the smart systems to enable the automation in the financial processing and predictions. To create the predictive models for financial processing we train the classification models such as Decision Trees and Deep learning models. Following section discuss the currently available literature on the deep learning and decision tree models for credit risk assessment and the challenges in training and implementation of these models.

## C. Deep Learning and Decision Trees Model for Credit Risk Assessment

The financial sector is associated with big data due to massive number of financial transactions. Due to advanced technology associated with big data, data availability and computing power, most banks or lending institutions in financial sector are renewing their business models (Peter et al. 2018). Credit risk predictions, monitoring, model reliability and effective loan processing are key to decision making and transparency, says Peter (et al. 2018). Presenting the binary classification model based on machine and deep learning models on real data, Peter (et al. 2018) observes that the tree-based models are more stable than the models based on multilayer artificial neural networks.

In this new era of digital and big data technology, transparency is necessary, and it should be one that does not stand in the way of innovation but allows the transformation and progress in the world. The incorporation of the technology and big data in the financial sector, the predictive models are created to automate the processing of the loan applications and other credit risk assessment tasks. The models used by Peter (et al. 2018) in their research project are Linear regression, Logistic regression, and Multinomial regression all with specific parameters, stopping criteria and activation functions. Along with these regression models, Peter (et al., 2018) also trained the classification models like Random Forest and Gradient Boosting. Along with these machine learning models for creating the predictive models, Peter (et al., 2018) also talks about the deep learning approaches stating that deep learning approaches consists of adding multiple layers to a neural network. The four architectures presented in this paper are convolutional neural networks, recurrent neural networks, recursive neural networks and the standard deep neural networks. All these 4 deep learning models presented by Peter (et al., 2018) have individual parameters and model architecture. The model accuracies in the form of AUC and RMSE values can be seen in the table below.

Sharing the statistics about the adaption of machine learning models in the corporate sector, Jacky (2018) says that more than 40% of large corporations are already using machine learning to boost their marketing and they can attribute approximately 38% of their sales improvement to machine learning, and around 76% of these corporations believe that machine learning will be a key component of their future sales growth. The

*TABLE 01 – Models' performances on the test dataset using AUC and RMSE values (Peter et al. 2018)*

| MODELS | AUC | RMSE |
|---|---|---|
| M1 | 0.876280 | 0.245231 |
| M2 | 0.993066 | 0.096683 |
| M3 | 0.994803 | 0.044277 |
| D1 | 0.904914 | 0.114487 |
| D2 | 0.841172 | 0.116625 |
| D3 | 0.975266 | 0.323504 |
| D4 | 0.897737 | 0.113269 |

applications of machine learning to business are broad, aside from targeted sales and market segmentation, it can be used for inventory optimization based on demand forecasting, personalized customer service and customer segmentation and many more including the finance sector (Jacky, 2018). Presenting the credit risk analysis using machine learning models, Jacky (2018) says in his study that, with an increasing number of companies making expansion overseas to capitalize on foreign resources, a multinational corporate bankruptcy can disrupt the world's financial ecosystem. In recent years, machine learning has become a popular field in big data analytics because of its success in learning complicated models.

Methods such as support vector machines, adaptive boosting, artificial neural networks, and Gaussian processes can be used for recognizing patterns in the data that might not be apparent to human analysts (Jacky, 2018). Results from this study shows that predictions with accuracy greater than 95% were achievable using any machine learning technique when informative features like experts' assessment were used. However, Jacky (2018) believes that using purely financial factors to predict whether a company will go bankrupt, the correlation is not as strong. More features are required to better describe the data, but these results in higher dimensional problem where the thousands of published companies' data is insufficient to populate this space with high enough density. Due to this "curse of dimensionality", flexible non-linear models tend to over-fit to the training samples and thus fail to generalize to unseen data (Jacky, 2018).

Presenting their study on credit scoring using machine learning model, specifically the Gradient Boosting method Yao (et al. 2022) says that credit scoring is an effective tool for banks and lending companies to manage the potential credit risk of borrowers. Machine learning algorithms have made great progress in automatic and accurate discrimination of good or bad borrowers (Yao et al., 2022). Random Forest, Decision Tree and Gradient Boosting machine learning techniques have become the mainstream ensemble methods for precise credit scoring in the financial sector. In their study, Yao (et al., 2022) incorporated the advantages of the bagging ensemble training strategy and boosting ensemble optimization pattern to enhance the diversity of base learners. The graph shared by Yao (et al., 2022) visualizing the average rank of credit scoring model for the Nemenyi test is given below.

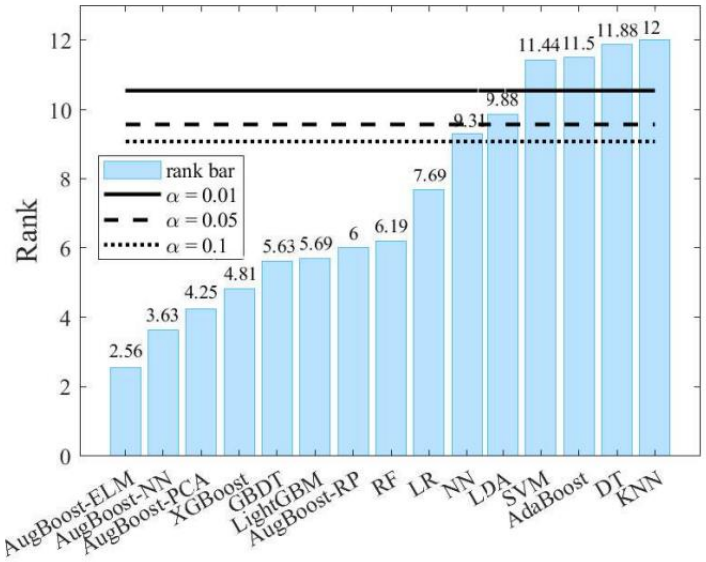



*Fig.01. Average ranks of credit scoring models for Nemenyi test (Yao et al., 2022)*

### D. Research Gap

From the complete study of the available literature and studies on credit risk assessment and predicting the potential credit risk beforehand using the financial data of the customers to automate the loan approval or rejection process in a financial institute like a bank, we understood that many authors and researchers have analysed the financial datasets and tried using them to create an AI based or machine learning based system. Most of these studies represents the potential use case of machine learning and deep learning models in the finance sector and credit risk assessment and shared the challenges in creating and using such systems in the sensitive sector like finance. The importance of interpretability and explainability is being notified by few of the authors and there are very less number of studies present which addresses this challenge in the finance sector using financial technology through machine learning, deep learning and artificial intelligence models. Along with this, keeping humans in the loop of prediction and decision making in credit risk assessment in the finance sectors is also a major challenge which has not been addressed by the researchers completely yet. This study presents a new approach towards credit risk assessment using machine learning techniques addressing the interpretability and explainability of the black-box model of machine learning and incorporates the idea of human-in-the-loop in the processing and decision-making task of the credit risk assessment.

## III. METHODOLOGY

This section of the research paper explains the dataset used to perform the analysis and train the proposed machine learning model for credit risk assessment. Along with the details about the dataset, the section lists the steps taken during data pre-processing and reasoning for them followed by the data analysis part sharing the graphs generated to understand the data and relation between the features of the data set and how they can be used to perform the prediction by the machine learning model.

## A. Dataset

The dataset used in this research project in the open dataset provided by the world bank on their website. This open dataset has the following columns which can be used as the independent features for training the predictive model for credit risk assessment. The dataset is created using the demographic and income information of the customers, constructed variables and the responses for the survey questionnaire asked to these customers from a total of 123 countries all around the world. The raw data had a sum of 127 variables including the demographic, financial and survey responses by the customers from around the world. The dataset after treating the null values and dropping the duplicate columns contains the features listed in the table below with their description and datatype.

*TABLE II – All the features in the dataset after treating the null values.*

| VARIABLE NAME | DESCRIPTION | DATA TYPE |
|---|---|---|
| economycode | Name of the economy | object |
| pop_adult | Adult (15+) population using 2020 World Development Indicators (WDI) | float64 |
| wpid_random | Individual-level identifier to merge with Gallup World Poll data | int64 |
| wgt | Weight assigned to each observation | float64 |
| female | Respondent is female or male<br>= 1 if the respondent is female<br>= 2 if the respondent is male | int64 |
| age | Respondent's age (in years) | float64 |
| educ | Respondent's education level<br>= 1 if the respondent has completed primary school or less<br>= 2 if the respondent has completed secondary school<br>= 3 if the respondent has completed tertiary education or more | int64 |
| inc_q | Respondent's within-economy household income quintile (1 to 5) | int64 |
| emp_in | Respondent is in workforce or no<br>= 1 is the respondent is in the workforce<br>= 2 if the respondent is out of the workforce | float64 |
| urbancity_f2f | Respondent lives in rural area or urban area<br>= 1 if the respondent lives in a rural area<br>= 2 if the respondent lives in an urban area | float64 |
| account | Has an account<br>= 1 if the respondent has an account at a financial institution, a mobile money account, or both<br>= 0 if the respondent does not have an account | int64 |
| saved | = 1 if the respondent personally saved or set aside money in the past year, including using an account at a financial institution, via a mobile money account, a savings club or person outside the family, or for any reason<br>= 0 if the respondent did not save | int64 |
| borrowed | = 1 if the respondent, personally or together with someone else, borrowed money in the past year, including from a bank or similar financial institution, via a mobile money account, from family or friends, or from an informal savings group, or for any other reason<br>= 0 if the respondent did not borrow | int64 |
| mobileowner | Owns a mobile phone | int64 |
| internetaccess | Respondent has internet access or not | int64 |
| anydigpayment | Made or received a digital payment.<br>= 1 if respondent used mobile money, a debit or credit card, or a mobile phone to make a payment from an account or used the internet to pay bills or to buy something online or in a store, or paid bills or sent or received remittances directly from or into a financial institution account or through a mobile money account in the past year. It also includes respondents who received payments for agricultural products, government transfers, wages, or a public sector pension into a financial institution account or through a mobile money account in the past year.<br>= 0 if the respondent did not make or receive a digital payment | int64 |
| remittances | Made or received a domestic remittance payment | int64 |
| merchantpay_dig | Made a digital merchant payment (1 for yes; 0 for no) | int64 |
| fin1_1a | Opened first account to receive a wage payment.<br>Note: Asked only of account owners (excluding mobile money accounts). | float64 |
| fin1_1b | Opened first account to receive money for the government.<br>Note: Asked only of account owners (excluding mobile money accounts). | float64 |
| fin2 | Has a debit card.<br>Note: Asked only of account owners (excluding mobile money accounts). | int64 |

| VARIABLE NAME | DESCRIPTION | DATA TYPE |
|---|---|---|
| fin4 | Used a debit card.<br>Note: Asked only of account owners (excluding mobile money accounts) that have an ATM/debit card. | float64 |
| fin4a | Used a debit card in store.<br>Note: Asked only of account owners (excluding mobile money accounts) that have an ATM/debit card. | float64 |
| fin5 | Used a mobile phone or internet to access account.<br>Note: Asked only of account owners (excluding mobile money accounts). | float64 |
| fin6 | Used a mobile phone or internet to check account balance.<br>Note: Asked only of account owners (excluding mobile money accounts) | float64 |
| fin7 | Has a Credit Card<br>Note: Asked only of account owners (excluding mobile money accounts). | float64 |
| fin8 | Used a Credit Card<br>Note: Asked only of account owners (excluding mobile money accounts) who have a credit card. | float64 |
| fin8a | Used a Credit Card in store.<br>Note: Asked only of account owners (excluding mobile money accounts) who used a credit card. | float64 |
| fin8b | Paid credit card balances in full.<br>Note: Asked only of account owners (excluding mobile money accounts) who have a credit card. | float64 |
| fin9 | Made any deposit into the account.<br>Note: Asked only of account owners (excluding mobile money accounts). This includes cash or electronic deposits, or any time money is put into their account(s) by themselves, by an employer, or another person or institution. | float64 |
| fin9a | Make any deposit into the account two or more times per month.<br>Note: Asked only of account owners (excluding mobile money accounts) | float64 |
| fin10 | Withdrew from the account.<br>Note: Asked only of account owners (excluding mobile money accounts). This includes cash withdrawals made in person using a debit card or mobile phone, electronic payments or purchases, checks, or any other sanctioned circumstance in which money is removed from the account(s) either by the account owner or by another person or institution. | float64 |
| fin10_1a | Reason for inactive account: too far<br>Note: Asked only of account owners with an inactive account in India. | float64 |
| fin10_1b | Reason for inactive account: no need<br>Note: Asked only of account owners with an inactive account in India. | float64 |
| fin10_1c | Reason for inactive account: lack money.<br>Note: Asked only of account owners with an inactive account in India. | float64 |
| fin10_1d | Reason for inactive account: not comfortable using it<br>Note: Asked only of account owners with an inactive account in India. | float64 |
| fin10_1e | Reason for inactive account: lack trust.<br>Note: Asked only of account owners with an inactive account in India. | float64 |
| fin10a | Withdrew from the account two or more times per month.<br>Note: Asked only of account owners (excluding mobile money accounts). | float64 |
| fin10b | Used account to store money.<br>Note: Asked only of account owners (excluding mobile money accounts). | float64 |
| fin11_1 | Unbanked: use account without help<br>Note: Asked only of adults without an account (excluding mobile money account owners). | float64 |
| fin11a | Reason for no account too far<br>Note: Asked only of adults without an account. The country-level averages in the report and the databank are calculated excluding mobile money accounts. | float64 |
| fin11b | Reason for no account : too expensive<br>Note: Asked only of adults without an account. The country-level averages in the report and the databank are calculated excluding mobile money accounts. | float64 |
| fin11c | Reason for no account: lack documentation.<br>Note: Asked only of adults without an account. The country-level averages in the report and the databank are calculated excluding mobile money accounts. | float64 |
| fin11d | Reason for no account: lack trust.<br>Note: Asked only of adults without an account. The country-level averages in the report and the databank are calculated excluding mobile money accounts. | float64 |
| fin11e | Reason for no account: religious reasons<br>Note: Asked only of adults without an account. The country-level averages in the report and the databank are calculated excluding mobile money accounts. | float64 |
| fin11f | Reason for no account: lack money.<br>Note: Asked only of adults without an account. The country-level averages in the report and the databank are calculated excluding mobile money accounts. | float64 |

| VARIABLE NAME | DESCRIPTION | DATA TYPE |
|---|---|---|
| fin11g | Reason for no account: family member already has one.<br>Note: Asked only of adults without an account. The country-level averages in the report and the databank are calculated excluding mobile money accounts. | float64 |
| fin11h | Reason for no account: no need for financial services<br>Note: Asked only of adults without an account. The country-level averages in the report and the databank are calculated excluding mobile money accounts. | float64 |
| fin13_1a | Reason for no mobile money account: too far<br>Note: Asked only of adults without an account in Sub-Saharan Africa. The country-level averages in the report and the databank are calculated excluding financial institution accounts. | float64 |
| fin13_1b | Reason for no mobile money account: too expensive<br>Note: Asked only of adults without an account in Sub-Saharan Africa. The country-level averages in the report and the databank are calculated excluding financial institution accounts. | float64 |
| fin13_1c | Reason for no mobile money account: lack documentation.<br>Note: Asked only of adults without an account in Sub-Saharan Africa. The country-level averages in the report and the databank are calculated excluding financial institution accounts. | float64 |
| fin13_1d | Reason for no mobile money account: lack of money<br>Note: Asked only of adults without an account in Sub-Saharan Africa. The country-level averages in the report and the databank are calculated excluding financial institution accounts. | float64 |
| fin13_1e | Reason for no mobile money account: use agent<br>Note: Asked only of adults without an account in Sub-Saharan Africa. The country-level averages in the report and the databank are calculated excluding financial institution accounts. | float64 |
| fin13_1f | Reason for no mobile money account: no mobile phone<br>Note: Asked only of adults without an account in Sub-Saharan Africa. The country-level averages in the report and the databank are calculated excluding financial institution accounts | float64 |
| fin13a | Use mobile money account two or more times a month.<br>Note: Asked only of adults who use a mobile money account. | float64 |
| fin13b | Use mobile money account to store money.<br>Note: Asked only of adults who use a mobile money account. | float64 |
| fin13c | Use mobile money account to borrow money.<br>Note: Asked only of adults who use a mobile money account. | float64 |
| fin13d | Use mobile money account without help.<br>Note: Asked only of adults who use a mobile money account. | float64 |
| fin14_2 | Paid digitally for an in-store purchase for the first time after COVID-19<br>Note: Asked only of adults who used a mobile phone or a debit or a credit card to pay for an in-store purchase. | float64 |
| fin14c | Paid online or in cash at delivery.<br>Note: Asked only of adults who used a mobile phone or the Internet to buy something online. | float64 |
| fin14c_2 | Paid online for an online purchase for the first time after COVID-19<br>Note: Asked only of adults who used a mobile phone or the Internet to buy something online and paid for it online. | float64 |
| fin24 | Main source of emergency funds in 30 days | int64 |
| fin24a | Difficulty of emergency funds in 30 days<br>Note: Asked only of adults who reported a main source of emergency funds (that it would be possible to come up with emergency funds in the next 30 days). | float64 |
| fin24b | Difficulty of emergency funds in 7 days<br>Note: Asked only of adults who reported a main source of emergency funds (that it would be possible to come up with emergency funds in the next 30 days). | float64 |
| fin26 | Sent domestic remittances | float64 |
| fin28 | Received domestic remittances | float64 |
| fin30 | Paid utility bill | int64 |
| fin31b1 | Paid a utility bill from an account or mobile phone for the first time after the start of COVID-19<br>Note: Asked only of adults who paid a utility bill using an account or a mobile phone. | float64 |
| fin32 | Received wage payments | int64 |
| fin35 | Received wage payments into an account or to a phone or a card and paid higher than expected bank fees.<br>Note: Asked only of adults who received a wage payment to an account, a mobile phone, or to a card. | float64 |
| fin37 | Received government transfer.<br>Note: This money could include payments for educational or medical expenses, unemployment benefits or subsidy payments. | int64 |
| fin38 | Received government pension | int64 |
| fin42 | Received an agricultural payment | float64 |

| VARIABLE NAME | DESCRIPTION | DATA TYPE |
|---|---|---|
| fin42a | Grow own crops or raise livestock.<br>Note: Asked only of adults who received a payment for the sale of an agricultural product. | float64 |
| fin44a | Financially worried: old age | int64 |
| fin44b | Financially worried: medical cost | int64 |
| fin44c | Financially worried: bills | int64 |
| fin44d | Financially worried: education | int64 |
| fin45 | Financially most worried<br>Note: Asked only of adults who are very or somewhat worried about two or more financial issues – old age, monthly bills, medical bills or school fees. | float64 |
| fin45_1 | Financially worried due to COVID-19 | float64 |

These are the variables present in the dataset after replacing the null values with appropriate alternative keeping the integrity of the features present in the dataset. Most of the null values were replaced with the number 0, as these were the questions asked to selective customers leaving the other cells empty. Entering 0 to those cells shows that this customer has not been asked to add a response for this particular question in the questionnaire. For the other features which were asked to all the customers, like "age", the null values were replaced with the mean of the column to keep those records in the dataset while having no null values. But the variables present in the dataset are similar and interconnected with each other, or in other words are correlated and shares similar information about an application. These variables need to be treated to create single column or feature for each unique information about the application. To get this we have the pre-process the dataset to analyse and understand different aspects of the dataset and different properties of the features to decide how to get the desired features to get the required dataset for training the machine learning model which we aim to use for credit risk assessment.

## B. Data Pre-Processing

The dataset used for this research project is the dataset provided by the world bank on their website with the title of "Global Findex 2021". This dataset contained the global findex survey responses and other financial and introductory information of the customers from a total of 123 countries. There were a total of 127 features in the data set with 63 features with null values. These null values were replaced with appropriate response as per the column or with 0 if the feature was asked to selected customers only, maintaining the integrity of the dataset provided by the world bank.

After removing the null values, we had a complete dataset with 0 null values and 127 columns. These columns contained the introductory data regarding the customers like, "country", "gender", "age", "employment status" and "whether he/she lives in a rural/urban area". Apart from these, the dataset had the basic financial data provided by the customers such as "ever Saved money", "ever borrowed money", "has a mobile and internet connection", "has an account" and more. Rest of the features were the questions from the survey questionnaire asked to these customers by the world bank to create this data. This questionnaire contains many such questions which were repeating thus creating multiple columns showing similar situations, These repeated columns were to be deleted. To decide which columns to keep and which should be deleted in order to keep the features which shows the maximum information regarding the customer, we performed the exploratory data analysis for these features and visualize the distribution of the similar features to compare the number of responses and recognize the one feature to keep. Below is the list of features analysed through this method.

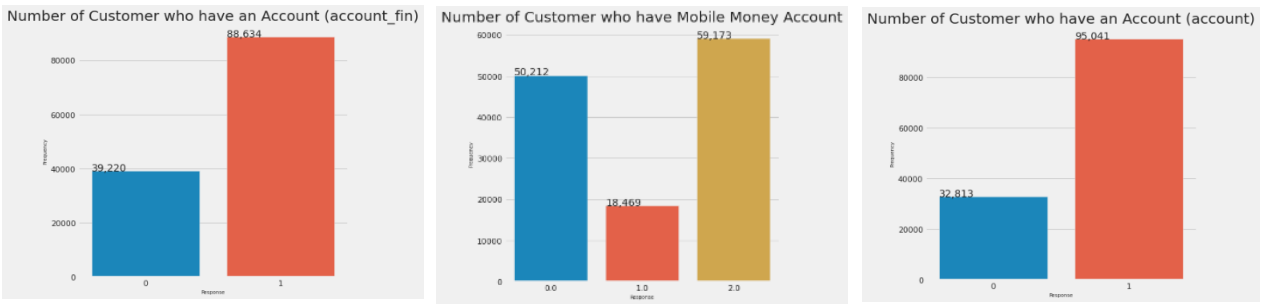



*Fig. 02. Number of customers having Account, and Mobile Money Account (account, account_fin , account_mob)*

The first graph in fig. 02 shows the number of customers who have an account in any financial institute, the second one shows the number of customers who have mobile money account, and the third graph shows the number of customers who have either the account, or the mobile money account or both. The third feature is a combination of the first two features and thus it is kept in the dataset while the other two features, "account_fin" and "account_mob" were dropped.

Further in the analysis, the first graph in fig. 03 shows the distribution of the column "borrowed" and shows the number of customers who has ever borrowed money from a financial institute or from family or friends, or for medical purposed. The second graph is for the feature "fin20", which is for customers who borrowed for medical purposes. The third graph is for "fin22a", the customers who borrowed from a financial institution and last one is for "fin22b", for customers who ever borrowed from family or friends. The feature "borrowed" is a combination for all other from "fin20", "fin22a", and "fin22b" thus keeping the column "borrowed" and dropping the 3 other features.

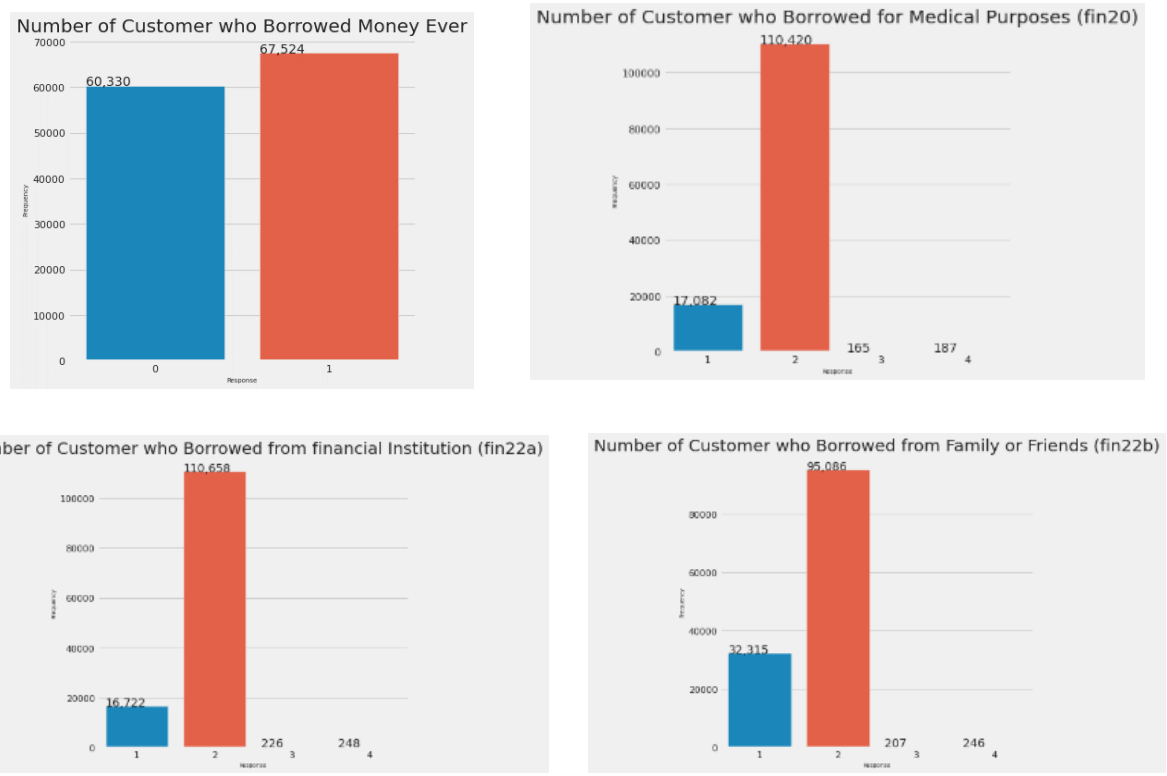
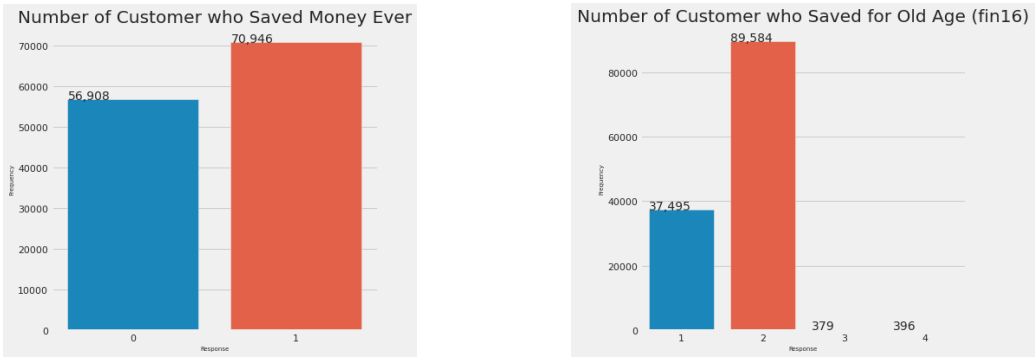



*Fig. 03. Number of customers who ever borrowed money (borrowed, fin20, fin22a, fin22b)*

Next looking at the column similar to "saved", in fig. 04, the first graph shows the data from the column "saved" showing the number of customers who ever saved money in any format. The second graph is for the column "fin16", showing the customers who has saved for old age, third graph for "fin17a", showing the customers



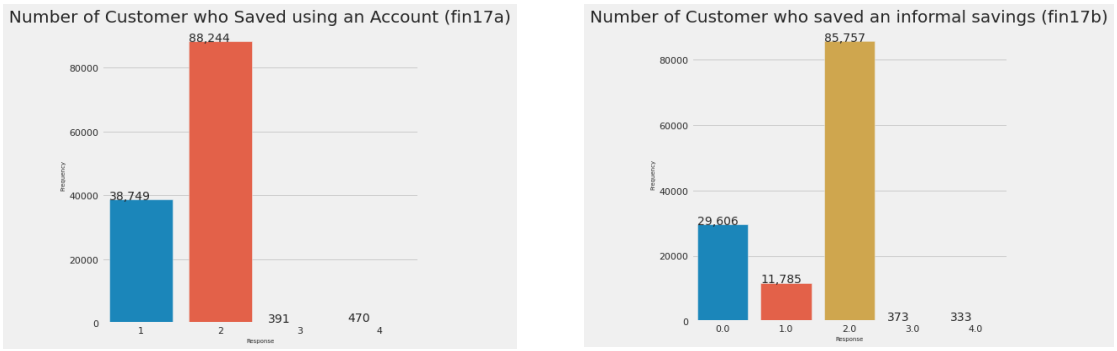



*Fig. 04. Number of customers who ever saved money (saved, fin16, fin17a, fin17b)*

who saved using an account and fourth graph for "fin17b", showing the number of customers who saved and informal savings. According to the description of the features given by the world bank, the feature "saved" is a combination of the features "fin16", "fin17a", and "fin17b". Therefore, keeping the column "saved" in the dataset and dropping "fin16", "fin17a", "fin17b". Similar process was followed for rest of the duplicate columns listed below. In each set of similar features, one feature describing the maximum information or the combination of information for all the others was kept in the dataset to remove the duplicate information. Below is the list of groups of duplicate features analysed.

i) receive_wages, fin32
ii) receive_transfers, fin37
iii) receive_pension, fin38
iv) receive_agriculture, fin42
v) pay_utilities, fin30
vi) anydigpayment, merchantpay_dig, fin14_1, fin14a, fin14a1, fin14b

From these group of similar features, one were kept in the dataset and other were dropped. The features which were kept in the dataset are listed in table 02. Moving forward with the analysis, we visualized the number of customers with mobile phone and those with the internet connection. Fig. 05 shows the graph for the customer distribution as per the responses of the customers who has the debit card and has used it.

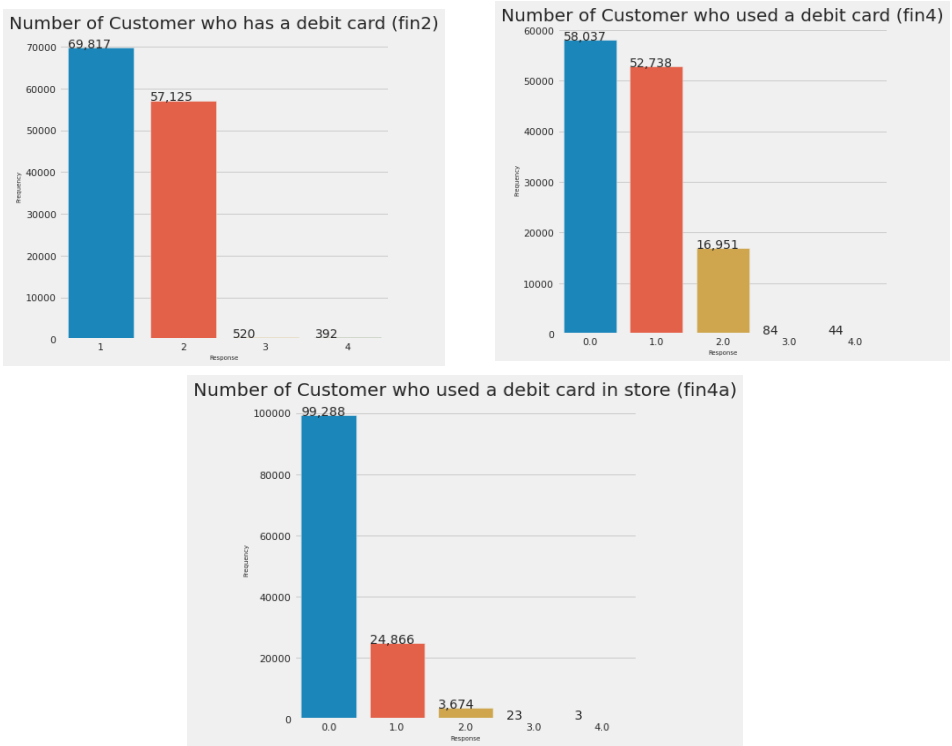



*Fig. 05. Number of customers who have a debit card and has used it (fin2, fin4, fin4a)*

Checking for the customers who has credit card and has used it and paid the credit card bill on time, we got the graphs shown in fig. 06. The features for this information are fin7, fin8a and fin8b.

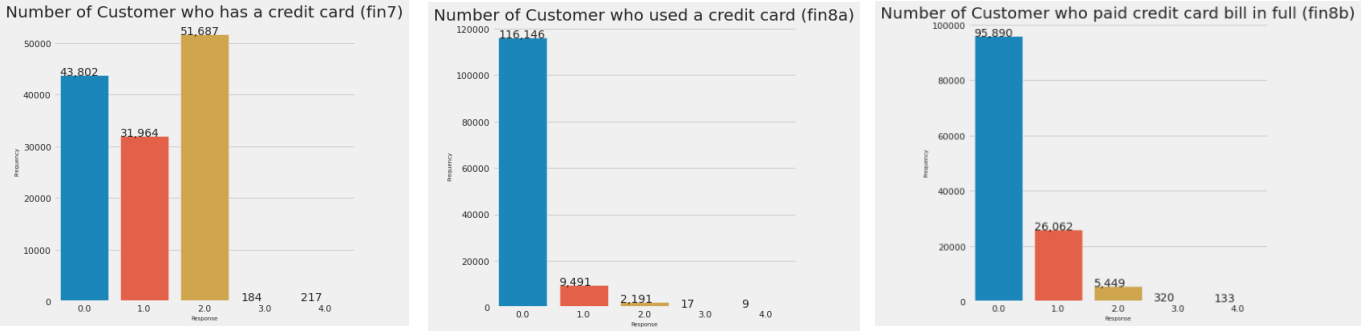



*Fig. 06. Number of customers who have a credit card and has used it and paid the bill (fin7, fin8a, fin8b)*

The responses for maximum questions in the questionnaire is in the format as mentioned below:

- 0 = It was a null cell in the raw dataset
- 1 = Yes
- 2 = No
- 3 = I do not know
- 4 = Refuse to answer

For some features like "fin24" the response definition is a bit different as compared to the standard responses for other features in the dataset as described below.

- 1 = if the respondent's main source of emergency funds is savings
- 2 = if the main source of emergency funds is family, relatives, or friends
- 3 = if the main source of emergency funds is money from working
- 4 = if the main source of emergency funds is borrowing from a bank, employer, or private lender
- 5 = if the main source of emergency funds is sale of assets
- 6 = if the main source of emergency funds is from some other source
- 7 = if the respondent could not come up with the money
- 8 = if don't know
- 9 = if refused to answer

The customer distribution for the feature "fin24" can be seen in fig. 07 showing the main source of emergency funds in 30 days. The meaning for each response in the graph below is described in the bullet points above.

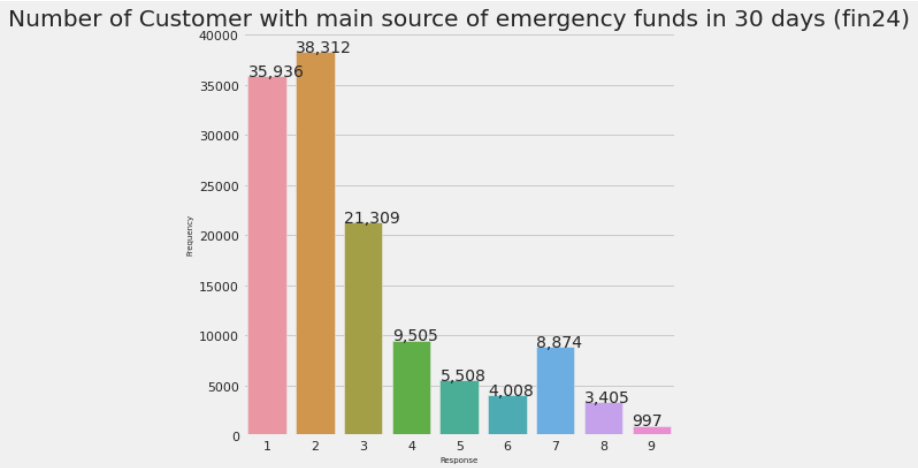



*Fig. 07. Number of customers with main source of emergency funds in 30 days (fin24)*

After removing the duplicate columns after checking for the customer distribution in all the similar features and analysing them for providing similar information, the dataset features were reduced to 80 from 127 before. The next challenge in data pre-processing was to combine all the extended features like the set of features “fin11a”, “fin11b”, “fin11b”, “fin11c”, “fin11d”, “fin11e”, “fin11f”, “fin11g” and “fin11h” and creating one feature for them which provides the information for all the above-mentioned columns. There are multiple such pairs in the dataset and to identify them we plotted the spearman correlation heatmap for the features for correlation coefficient greater than 0.85. The heatmap for the spearman correlation can be seen in fig. 08 below. In the spearman correlation heatmap, the grids in orange and red color shows the correlation more than 88% and these are the features we need to combine and create a common feature. There are 5 different groups of grids of orange/red color, and these are the 5 groups of features which will be processed to create 5 new features for each group. The dataset are performing the PCA on all the correlated features in the dataset has a total of 34 features. In the dataset, there is a categorical feature “economycode” which stored the country code for each country out of 123 countries present in the data. This is an important feature and as this is a categorical feature, it is needed to be encoded. As there are already a lot of features in the dataset, one-hot encoding for this column is not feasible. Therefore, the binary encoding is performed resulting in a total of 40 features in the dataset with all unique and numerical columns/features. All these 40 features are the independent features. The dataset does not possess the target feature. To create the target feature, the dataset is used to create an unsupervised machine learning model for classification.

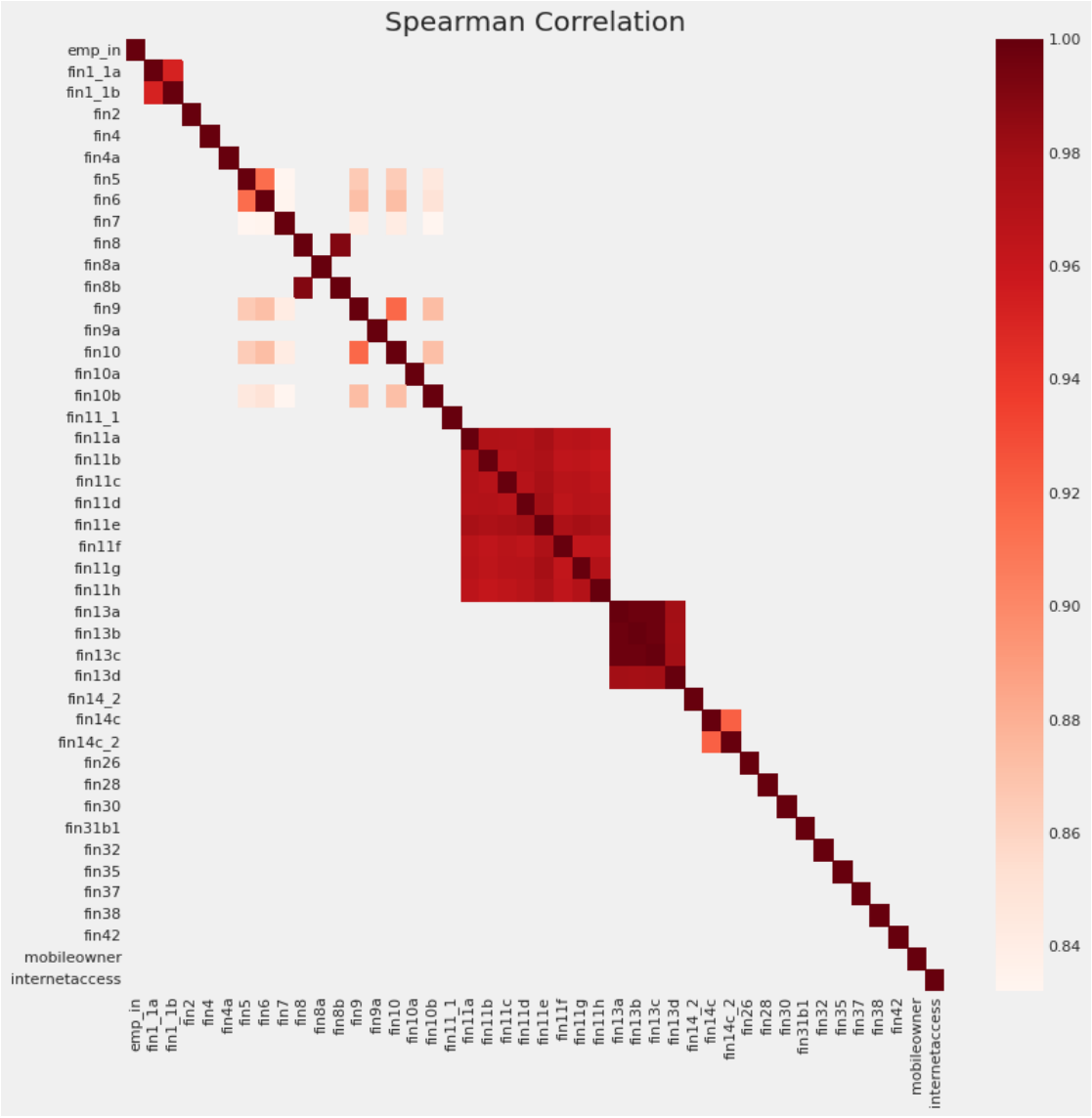



*Fig. 08. Spearman Correlation Heatmap for coefficient > 0.85*

This model is the k-means clustering, creating 2 clusters based on the dataset with 40 features. The classification performed by the unsupervised machine learning technique, k-means clustering model marked each record in the dataset with 0 or 1 for the records or application being negative or positive for credit risk, respectively. The graph in fig. 09 shows the number of records classified as negative i.e., "0" or positive i.e., "1" for the credit risk for the financial institute.

The dataset has a total of 1,27,854 records, out of which 85,576 were classified as negative for the credit risk whereas, 42,278 records were classified as positive for the credit risk. The classification mapping of "0" means the application is not risky for the financial institute and on the other hand, "1" means that the application is risky and might end up being risky for the financial institute. Observing the distribution of the classification, around 66.66% of the records were classified as "not risky" and remaining 33.33% of the records were classified as "risky" for the credit risk assessment making the dataset unbiased for any one class of the binary classification. This makes this classification column suitable for the target variable and can be used to train the supervised machine learning classification models.

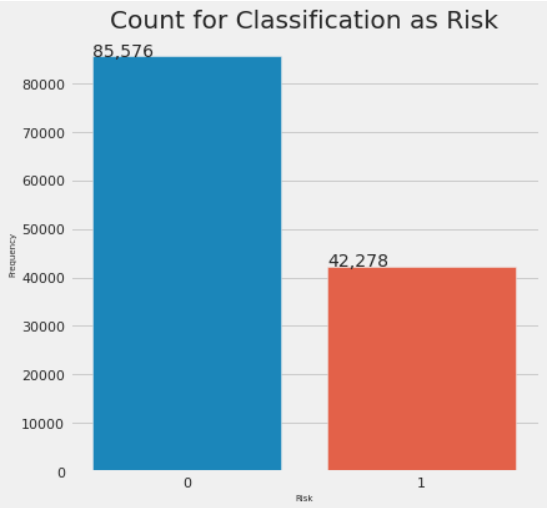



*Fig. 09. Classification results from k-means clustering (unsupervised machine learning technique)*

## IV. RESULTS

This section of the research paper displays the results of the research and the prediction of the machine learning models trained by the dataset for predicting the credit risk using the financial features of a customer of a financial institute. Along with the machine learning models accuracy, this section also displays the results of the explainable AI models for the black box machine learning models.

### A. Machine Learning Models

This research project aims to provide a machine learning based solution for credit risk assessment and introduce the automation in prediction. The first research question this study tries to answer is to identify the most effective interpretability methods for credit risk assessment in finance, in terms of balancing interpretability and accuracy. For finding the most suitable interpretability method, the first step is to train and develop a prediction model to get the insights from. The machine learning models trained on the pre-processed data are the classification machine learning models SVM, Logistic Regression and Decision Tree Classifier. The accuracy score for these machine learning models are given in Table 03. The problem statement of this research project is a binary classification problem statement where we are creating a machine learning model to assess the credit risk using a set of features related to the financial history of the customer recorded through their financial transactions and a survey questionnaire. The pre-processed dataset is divided into the independent and dependent variables as x and y, respectively. The dataset was divided into test and train dataset using the test-train split from the python module sklearn. The test size for all the models is kept as 0.30, keeping the 70% of the dataset for training the

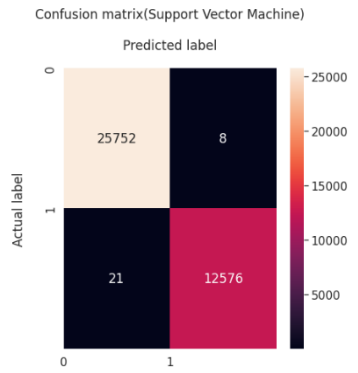
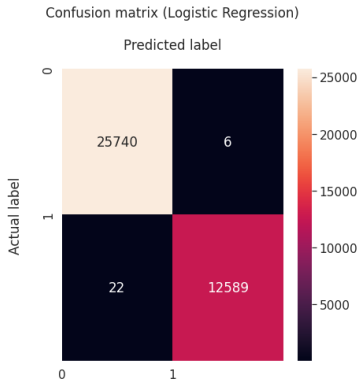
model and remaining 305 for testing. The classification models trained using this dataset are the Support Vector Machine Classifier, Logistic Regression and Decision Tree Classifier. The test accuracy for all three model is given in Table 03.

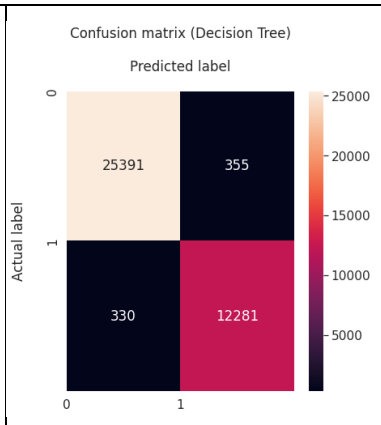
*TABLE III – Test Accuracy of Machine Learning Models for Classification.*

| MODEL | ACCURACY |
|---|---|
| Support Vector Machine (SVM) | 99.924% |
| Logistic Regression | 99.91% |
| Decision Tree Classifier | 98.12% |

The SVM classifier is a supervised machine learning algorithm used for both classification and regression problem statements. In this research project, SVM is trained for a binary classification problem statement. The goal of the SVM algorithm us to create the best decision boundary to segregate n-dimensional space into classes. The f1-scores for support vector machine classifier model for the test dataset is 1.00 for both the classes in the classification. The precision and score values for the SVM model for both the classes is also recorded as 1.00. The confusion matrix and classification report for the prediction on the test dataset by the SVM algorithm can be seen in Table IV.

*TABLE IV – Confusion Matrices and other evaluation matrices for ML models.*

| ML Model | Confusion Matrix | True Positive | False Positive | True Negative | False Negative | f1 - Score | Precision | Recall |
|---|---|---|---|---|---|---|---|---|
| Support Vector Machine |  | 25,752 | 8 | 12,576 | 21 | 1.00 | 1.00 | 1.00 |
| Logistic Regression |  | 25,740 | 6 | 12,589 | 22 | 1.00 | 1.00 | 1.00 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Decision Tree Classifier |  | 25,391 | 355 | 12,281 | 330 | 0.99 | 0.99 | 0.99 |

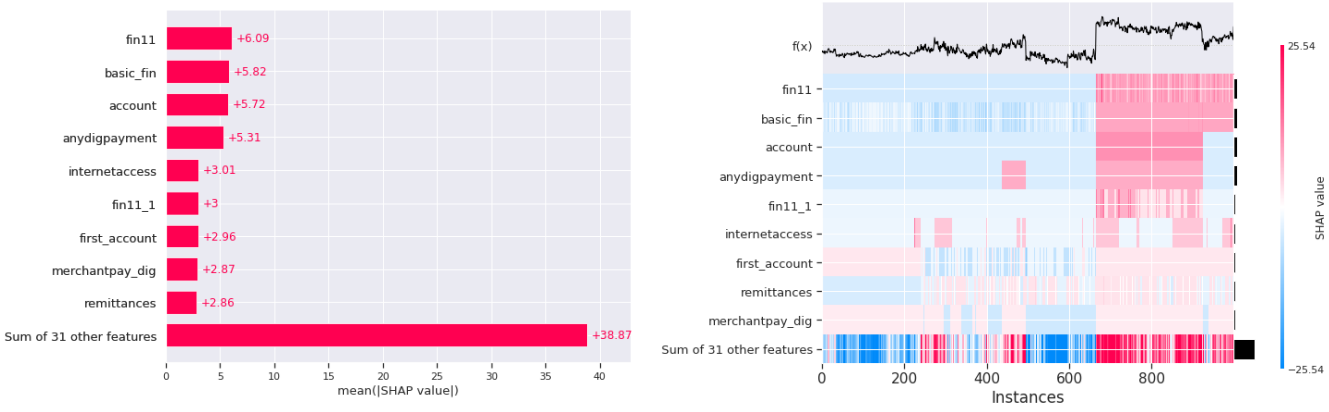
The next machine learning model trained for predicting the credit risk using the world bank data was logistic regression. Logistic regression is a supervised machine learning technique used for classification. It is a statistical method for analysing the dataset with more than one independent variable determining the outcome. The intension behind using logistic regression is to find the best fitting model to describe the relationship between the independent and dependent features in the dataset. The dataset used for this research project contains 39 independent features and 1 independent feature. The 39 independent features are used to analyse and predict the outcome through the logistic regression model.

Similar to SVM, the classification report for Logistic Regression also shows the f1-scores for both the classes in the classification as 1.00. The precision and recall values for the logistic regression model are also 1.00 for both the classes. The confusion matrix and classification report for the logistic regression model can be seen in Table IV along with the confusion matrix of SVM. The problem statement is a binary classification problem and therefore the third machine learning model trained is the decision tree classifier. The model accuracy for the decision tree classifier with the test dataset was recorded as 98.21%. The classification report for the decision tree model calculates the f1-score for class “0” as 0.99 and that for class “1” as 0.97. The precision and recall scores for “0” and “1” are also calculated as 0.99 and 0.97 respectively. The confusion matrix and classification report for the predications by decision tree classifier can also be seen in Table IV.

## B. Interpretability and Explainability of ML Models

After training the machine learning model we need to understand the processing of these machine learning models to find the model which performs the predictions most effectively. This also answers the first research question for finding the most efficient interpretability methods for credit risk assessment in finance. There are multiple interpretability methods present which can be used to understand the black – box machine learning models. The interpretability method used in this research project is Shapley Value (SHAP) to understand how the machine learning models are processing the input data to formulate and deliver the output.

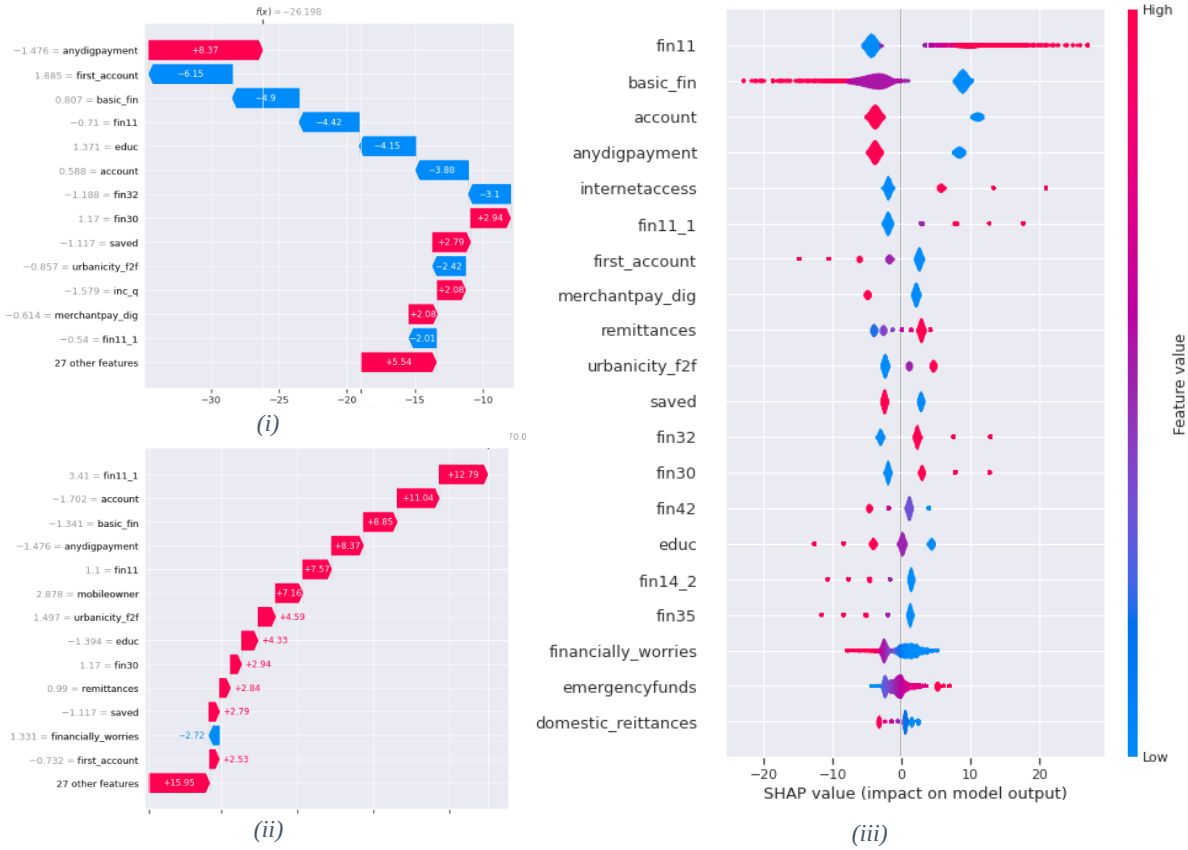
Fig. 10 shows the shap bar plots and the heatmap for Support Vector Machine algorithm. The bar plot shows the shap values for different features which describes the influence of these features on the predictions delivered by the SVM algorithm and the heatmap in fig. 16 which shows the influence of features in the prediction at different instances of the predications.




*(ii)* *(i)*

*Fig.10 . Shap Bar plot (i) and Heatmap (ii) for SVM algorithm*

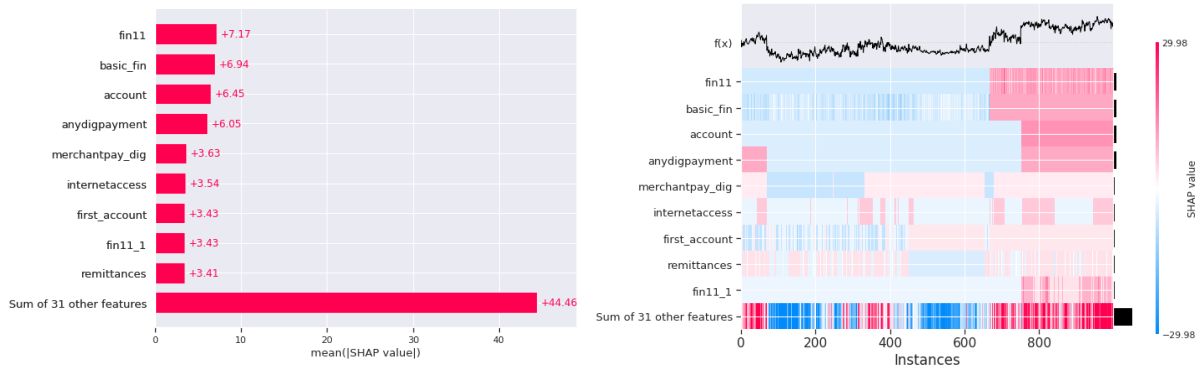
The summary and the waterfall plot from shap for the SVM algorithm is shown in the fig. 11. The summary plot shows the influence of first 20 features on the predictive power of the machine learning model.




*(i)* *(ii)* *(iii)*

*Fig. 11. Shap Waterfall plot (i, ii) and Summary plot (iii) for SVM algorithm*

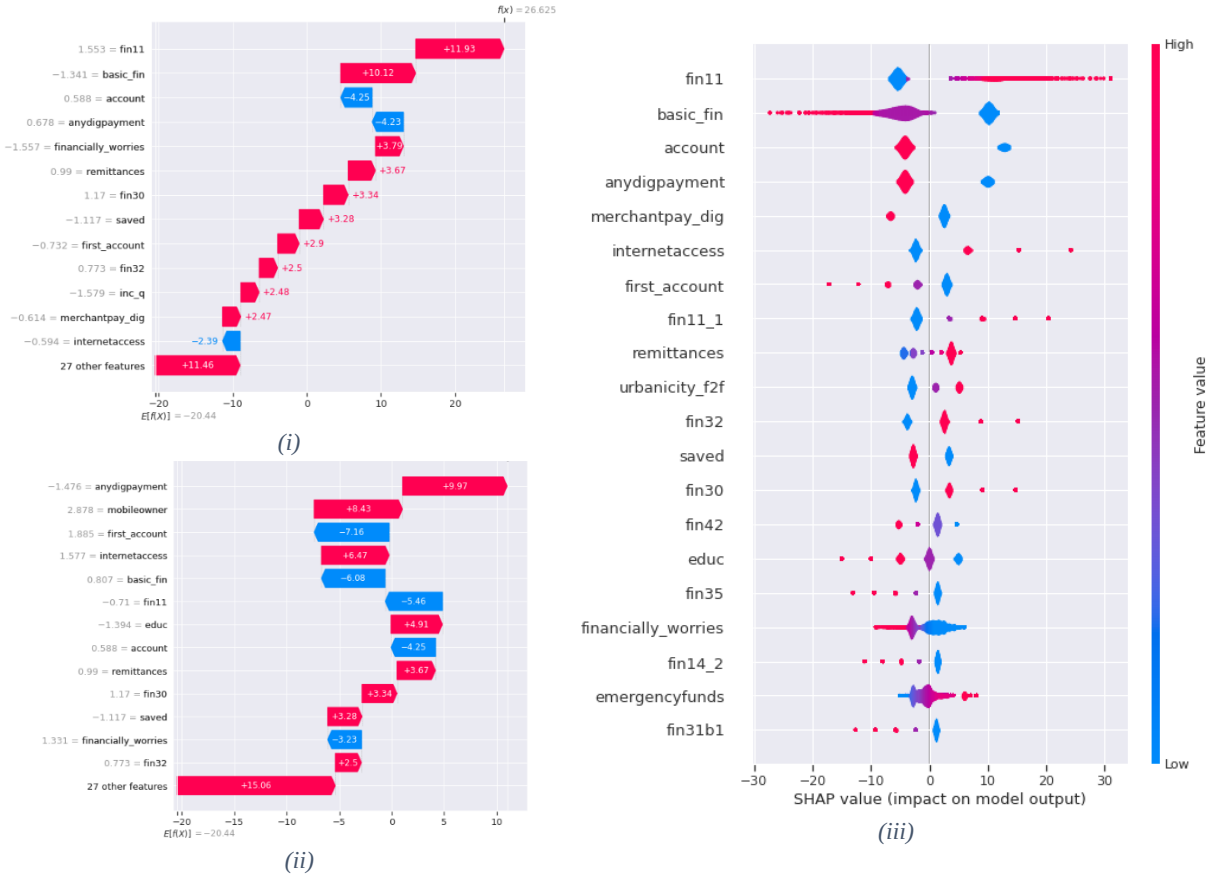
The waterfall plot is visualized for 2 different applications. This graph shows the influence of features in the final prediction for that particular data point. Fig. 12 shows the shap bar plot and heatmap for the logistic regression machine learning model. The bar plot arranges the features according to their influence on the prediction of the model.




*(i)* *(ii)*

*Fig. 12. Shap Bar plot (i) and Heatmap (ii) for Logistic Regression Model*

The summary and the waterfall plot from shap for the Logistic Regression machine learning model is shown in the fig. 13. The summary plot shows the influence of first 20 features on the predictive power of the machine learning model. The waterfall plot is visualized for 2 different applications. This graph shows the influence of features in the final prediction for that particular data point similar to that for the SVM model.




*(i)*

*(ii)*

*(iii)*

*Fig. 13. Shap Waterfall plot (i, ii) and Summary plot (iii) for Logistic Regression*

The third classification machine learning model is Decision Tree. The explainable artificial intelligence model for decision tree in the shap model is shap.TreeExplainer and plots the shap decision plot. The decision plot for the decision tree model trained on the world bank dataset can be seen in fig.14 below. The multiple lines seen in the decision plot shows the contribution of different features in the final classification through the decision tree classifier.

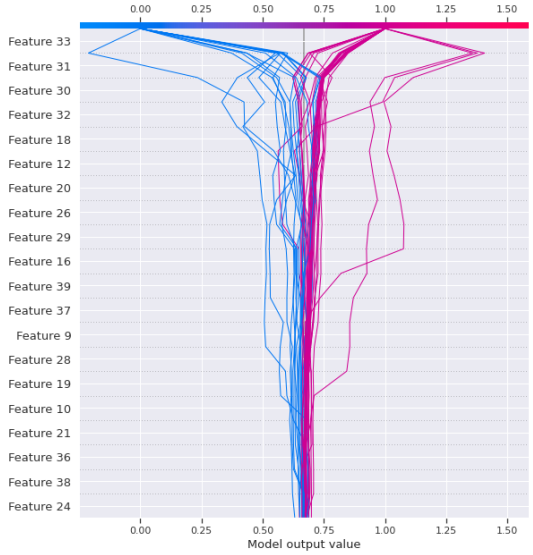



*Fig.14. Shap Decision plot for Decision Tree Classifier*

The other shap plot to visualise the processing of the decision tree classifier is the force plot which shows the contribution of difference features for a single classification. The force plot for decision tree classifier can be seen in fig.15 below which shows the contribution of different features for the final classification.

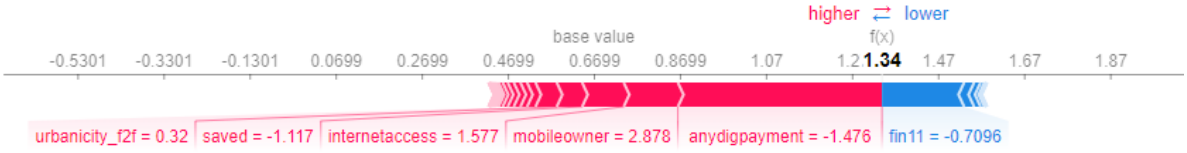



*Fig. 15. Shap Force plot for Decision Tree Classifier*

## V. DISCUSSION

This section discusses the essential interpretation based on the key finding of the research. The research questions this work tries to answer can be listed as follows:

RQ.1 What are the most effective interpretability methods for credit risk assessment in finance, in terms of balancing interpretability and accuracy?
RQ.2 How do the interpretability and accuracy trade-offs differ across different credit risk assessment tasks in finance?
RQ.3 How does domain knowledge impact the interpretability of credit risk assessment models in finance?
RQ.4 How can human-in-the-loop decision making be effectively incorporated into credit risk assessment using interpretable models in finance?

To answer the first research question and understand the reasoning behind the predictive performance of the machine learning models for classifying the dataset against the credit risk and finding the most effective interpretability methods for credit risk assessment we applied the explainable artificial techniques to visualise the processing of the black box machine learning models. There are several interpretability methods and tools for interpreting the black-box machine learning models. These tools are listed in an order from least to greatest complexity below:

- Partial Dependence Plot (PDP)
- Individual Conditional Expectation (ICE)

- Feature Importance
- Global Surrogate
- Local Surrogate (LIME)
- Shapley Value (SHAP)

For this research project, the explainable AI model "SHAP" is used to visualise and understand the processing of the machine learning models and check the features influencing the predictions delivered by the machine learning models. The SHAP model explains a prediction by assuming that each feature value of the instance is a "player" in a game. The contribution of each feature is measured by adding and removing the feature from all subsets of the rest of the features. The Shapley Value for one feature is the weighted sum of all its contributions. These values are additive and locally accurate. If we add up the Shapley Values of all the features, plus the base value or the prediction average, we will get the exact prediction value. This feature is missing in other explainable AI models.

The shap bar plot for SVM from fig. 10 shows the influence of features on the prediction results given by the model. As per the shap bar plot, "fin11" is the feature influencing the prediction followed by the feature "basic_fin", and " account", where, "basic_fin" feature is the combination of multiple features "fin2", "fin4", "fin5", "fin6", "fin7", "fin8", "fin8a", "fin8b", "fin9", "fin9a", "fin10", "fin10a", and "fin10b". The summary plot for SVM in fig. 11 (iii) shows the influence of first 20 features on the predictive power of the machine learning model. The waterfall plot in fig 11 (i) and (ii) is visualized for 2 different predictions and shows the influence of features in the final prediction for that particular record. The last shap graph for the SVM algorithm is the heatmap in fig. 10 (ii) which shows the influence of features in the prediction at different instances of the predications. Observing all the shap plots, "fin11" is the feature having the highest influence on the prediction, closely followed by "basic_fin" and "account". This explanation shows that the basic financial features and the status of financial account is an important feature in predicting the credit risk for an application.

Fig. 12 (i) shows the shap bar plot for the logistic regression machine learning model and similar to the results for SVM model, in logistic regression model the shap bar plot reads the feature "fin11" as the feature with highest influence on the prediction closely followed by "basic_fin" and "account". The summary plot for logistic regression model is shown in fig. 13 (iii) which shows the top 20 features having maximum influence on the prediction of the logistic regression model. Observing the summary plot for SVM and Logistic regression model, we can say that top 4 features for both the models are same as "fin11", "basic_fin", "account", and "anydigpayment". These results shows that the classification is majorly dependent on the financial data of the customer like account, debit card, credit card, on time payment of credit card bill, use of mobile phone and internet for digital payment and internet banking. Followed by these, the features influencing the rejection of an application or predicting it as risky is "saved", "inc_q" "financially_worries" and "emergencyfunds". Fig. 13 (i) and (ii) shows the waterfall plot for two different predictions in the dataset showing the positive and negative influence of the features on the final prediction for that single record. Fig. 12 (ii) shows the heatmap for the influence of features on the final predictions at multiple different instances in the dataset and predictions.

The third classification machine learning model is Decision Tree. The explainable artificial intelligence model for decision tree in the shap model is shap.TreeExplainer and plots the shap decision plot. The decision plot for the decision tree model trained on the world bank dataset can be seen in fig. 14. The multiple lines seen in the decision plot shows the contribution of different features in the final classification through the decision tree classifier. The other shap plot to visualise the processing of the decision tree classifier is the force plot which shows the contribution of difference features for a single classification. The force plot for decision tree classifier can be seen in fig. 15 which shows the contribution of different features for the final classification.

Comparing the results from all the three Shapley models for three machine learning models, the mains feature influencing the prediction or in the case of decision tree classifier, the root node of the decision tree is found to be the feature "fin11". "fin11" is the feature which is created by performing PCA on the group of features "fin11a" to "fin11h". This feature records the reason for having no account in any financial institute. This outcome from the explainable AI tool SHAP describes the importance of "having an account" in deciding the credit risk. This feature was closely followed by multiple other features like "basic_fin", "saved", "inc_q",

"financially_worries" and more giving us the list of top 10 features, which are contributing to the prediction power of our machine learning models. This answers the second research question of this project as the features influencing the predictions decides the actual meaning of the prediction. This subjective nature of the explanation of the black-box classification models introduces the need of domain knowledge in determining the actual meaning of the results from the machine learning model.

The third research question of this project asks how the domain knowledge impact the interpretability of credit risk assessment models in finance. The subjective nature of the interpretability answer this question. As discussed already that the influence of features on the predictions can be explained in different way for different prediction tasks and this explanation requires the domain knowledge. For example, if a predictive model is predicting the financial instability of a customer in repaying the loan amount, the meaning of the contribution of his credit score and previous bill payments is different as compared to the contribution of the same features in predicting the need of loan to a customer. To assess this difference and process the prediction accordingly the process needs the domain knowledge and human-in-the-loop decision making process in the credit risk assessment using the machine learning models like the ones presented in this research project.

## VI. CONCLUSION

The advancement in technology is leading innovation in all sectors of the various different industries across the world. Everything in the world is optimized to become automated and used least amount of human force and human intelligence. Among all these industries is the finance industry. The finance industry is the one which operates on credit and money. There is a constant flow of money from one form to another in the finance sector which brings the high probability of credit risk. Talking about the credit risk, there are multiple types of it for example loan defaults, market crash or inflation impact. All these credit risk requires specific attention and assessment beforehand to be aware of coming risks and prepare to prevent them. With the involvement of technology like artificial intelligence and machine learning, this credit risk assessment can be made an automated task for an institute based on some assumptions and rules. The idea of automating the credit risk assessment is a crucial task and implementing it using the machine learning technology involves the trust issues on the processing and predictions of these models. This research project presented a machine learning based solution for automated credit risk assessment and also solves the problem of interpretability of these machine learning models and their prediction by involving the explainable AI and human-in-the-loop decision making process. As per the study and results presented in this study, the credit risk assessment is a crucial and subjective task which involves domain knowledge in deciding the actual meaning of the predictions made and requires the human-in-the-loop decision making process to ensure the authenticity of the assessment and decision making. The best accuracy of the machine learning models presented for credit risk assessment is recorded as 99% with the value of f1-score as 1.00. The explainable AI graphs plotted using the XAI tool SHAP supports the recorded accuracy and explains the processing of the machine learning models and describes the influence of each feature on the final predictions by the model. The scope of this research project was limited to presenting a machine learning model for credit risk assessment and the most effective interpretability model to assess the processing and the results given by the machine learning models to formulate the final decision based on the domain knowledge and expert advice. The development and implementation of such a system for evaluating and predicting the credit risk is a task for future research in the domain.

## REFERENCES


Demajo, Lara & Vella, Vince & Dingli, Alexiei. (2020). Explainable AI for Interpretable Credit Scoring. 10.5121/csit.2020.101516.

Christoph Molnar. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Branks Hadji Misheva, Ali Hisra, Joerg Osterrieder, Onkar Kulkarni & Stephen Fung Lin. (2021). Explainable AI in Credit Risk Management. https://arxiv.org/pdf/2103.00949.pdf

Daniel Kaszynski, Bogumil Kaminski & Tomasz Szapiro. (2020). Credit Scoring, In Context of Interpretable Machine Learning. ISBN 978-83-8030-424-6.

Bussmann, N., Giudici, P., Marinelli, D. *et al.* Explainable Machine Learning in Credit Risk Management. *Comput Econ* **57**, 203–216 (2021). https://doi.org/10.1007/s10614-020-10042-0

Barredo Arrieta, Alejandro & Diaz Rodriguez, Natalia & Del Ser, Javier & Bennetot, Adrien & Tabik, Siham & Barbado González, Alberto & Garcia, Salvador & Gil-Lopez, Sergio & Molina, Daniel & Benjamins, V. Richard & Chatila, Raja & Herrera, Francisco. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion. 58. 10.1016/j.inffus.2019.12.012.

Jacky C. K. Chow. (2017). Analysis of Financial Credit Risk Using Machine Learning. https://arxiv.org/ftp/arxiv/papers/1802/1802.05326.pdf

Peter Martey Addo, Dominique Guegan, Bertrand Hassani. (2018). Credit Risk Analysis using Machine and Deep learning models. ISSN: 1827-3580 No. 08/WP/2018.

Addo, Peter & Guegan, Dominique & Hassani, Bertrand. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. Risks. 6. 38. 10.3390/risks6020038.

Jillian M. Cllements, Di Xu, Noosin Yousefi, Dmitry Efimov. (2020). Sequential Deep Learning for Credit Risk Monitoring with Tabular Financial Data. https://doi.org/10.48550/arXiv.2012.15330

Emmert-Streib, Frank & Yli-Harja, Olli & Dehmer, Matthias. (2020). Explainable Artificial Intelligence and Machine Learning: A reality rooted perspective.

P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, Amy N. Yates, Kristen Greene, David A. Broniatowski, Mark A. Przybocki. (2021). Four Principles of Explainable Artificial Intelligence. https://doi.org/10.6028/NIST.IR.8312

Waddah Saeed, Christian Omlin. (2021). Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities. https://arxiv.org/pdf/2111.06420.pdf

Girija Attigeri, Manihara Pai M M, Radhika M Pai. (2019). Framework to predict NPA/Willful defaults in corporate loans: a bug data approach. International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 5, October 2019, pp. 3786~3797 ISSN: 2088-8708, DOI: 10.11591/ijece.v9i5.pp3786-3797

Bawa, Jaslene & Goyal, Vinay & Mitra, Subrata & Basu, Sankarshan. (2018). An analysis of NPAs of Indian banks: Using a comprhensive framework of 31 financial ratios. IIMB Management Review. 31. 10.1016/j.iimb.2018.08.004.

Dhananjaya Kadanda, Krishna Raj. (2018). Non-Performing assets (NPAs) and its determinants: a study of Indian public sector banks. Journal of Social and Economic Development (2018) 20:193-212 https://doi.Org/10.1007M0847-018-0068-0