Predicting NPA Loans Using Machine Learning Techniques on Loan Details of Customers

Isha Choudhary | shahooda637@gmail.com

ABSTRACT

I. INTRODUCTION

II. AIM AND OBJECTIVES

This research project aims to create a machine learning model for predicting the NPA loans beforehand to avoid bad loans and non-performing assets leading to huge financial losses to the banks and financial institutes providing loans to their customers. The objectives of this research projects can be listed in 5 points as below:

i) Data Collection
ii) Data Pre-Processing
iii) Data Analysis
iv) Model Training
v) Explainable AI Graphs for ML models' performance

The main objective for this research project is to collect a dataset of financial transactions of the customers and the loan details including the interest rates and more of the details related to the accounts of the customers of the bank. This collected data is pre-processed and analysed to understand the patterns in the data and make it ready for training the machine learning model which is then used to predict the NPA loans beforehand to avoid bad loans and nonperforming assets. At the end of this research, we also aim to analyse the explainable AI for the models we trained for the prediction and understand how the models are able to predict the NPA loans with a certain level of accuracy and what features are influencing the prediction most.

III. SCOPE

A. In – Scope

The scope of this research project includes collecting the transactional and loan data of multiple customers of a bank or a financial institute. Along with collecting the dataset, the scope of this research project also includes the pre-processing of the dataset and analysing the characteristics of different features of the dataset in comparison to other features and how they can be used to train the machine learning model to predict the NPA loans. The data analysis and pre-processing is done with a view to understand the data and make it ready for model training. At the end the data is used to train multiple machine learning models and their accuracies in predicting the NPA loans is noted. This research project also shares the explainable AI graphs for the model showing best accuracy in prediction in comparison with all the other models.

B. Out – Scope

This research project is only presenting the machine learning model for predicting the NPA loans before they get NPA using the financial transaction details of the customer and the loan details. The development of a system using this machine learning model which can be used by the banks or financial institutes is out of the scope for this research project.

IV. RELATED WORK

The financial risk management is studied by many researchers in the world, and multiple studies, experiments and system architectures are presented by many researchers and scientists uplifting the financial industry by incorporating technology to optimize the functioning and predictive systems. This section of the research paper presents a review of existing studies in the financial sector, specifically focussing on the studies presented on the NPA and risk management topics.

Presenting a framework to predict NPA or Wilful defaults in corporate loans, [2] says that the bad loan which are Non-Performing Assets (NPA) are the measure for assessing the financial health of the bank. It is very important and crucial to controls the NPA rate as it affects the profitability and deteriorates the quality of assess of the bank. [2] believes that a systematic identification, awareness and assessment of the parameters is essential for early prediction of the wilful default behaviour. [2] aims to identify exhaustive list of parameters essential for predicting whether the loan will become NPA and thereby wilful default. The process presented by [2] includes understanding the existing system to check the NPAs and identify the critical parameters and also propose a framework for NPA/Wilful default identification. In order to select the best classification model in the framework, [2] conducted an experiment on loan dataset on a big data platform. Since, the loan data is structures, the unstructured component is incorporated by generating synthetic data. The results presented by

[2] indicate that neural network model gives the best accuracy and hence they considered it to be in the framework for predicting the NPAs or Wilful defaults beforehand and prevent them.

Further presenting their study on the methods to reduce the Gross NPA and classifying the Defaulters using the Shannon Entropy, [1] says that Non-Performing Assets (NPA) causes a huge loss to the banks and hence it becomes an extremely critical step in deciding which loans have the capability to become an NPA, and thus deciding which loan to grant and which ones to reject. Thus, to solve this problem, [1] proposed an algorithm designed to handle the financial data very meticulously to predict whether a particular loan would be an NPA in future or not with a very high accuracy. The main ideology used by [1] in their work was around the central concept of the entropy. They believe that if the sample of the data that is completely processed and homogenous then the value of entropy will be zero, while on the other hand if the sample data is equally divided then the value of entropy will be one. [1] used the local entropy and global entropy for determining the output. The entropy classifier model by [1] is then compared with the existing classifiers used to predict NPAs to assess the performance of the proposed algorithm in comparison to the currently used systems.

Presenting a unique approach for Prediction of Loan Approval using Machine Learning Algorithm, [8] says that a bank's profit or loss depends on the loans to a large extent, whether the customers are paying back the loan or defaulting, are the defaulter are for a genuine reason or a wilful default and more. By predicting the loan defaulters, bank can reduce its non-Performing assets, and this makes this study very important. A very important approach in predictive analysis is used by [8] to study the problem of predicting the loan defaulters, the logistic regression model. [8] collected the dataset from Kaggle for studying the data and using it to train the predictive model. The final results from [8] shows that the model is marginally better because of the variables the used dataset includes. Some of the features used by [8] in the dataset are personal attributes and financial transaction of the customers like, age, purpose, credit history, credit amount, credit duration and more such features which shows the wealth of the customer. The accuracy of the model presented by [8] was recorded to be 81.1% and the study concludes that most of the time the applicants who have high income and demands for lower amount of loan are more likely to get the approval as they are more likely to pay back the loan amount.

Along with all these, presenting a solution for NPA management in Indian Banking Sector, [6] also presents a solution based on Artificial Intelligence. [8] believes that there is a close relationship between the banking sector and the economic development where the growth of the overall economy is intrinsically correlated to the health of the banking and financial industry. It also says that the key challenge for the Indian banks is to expand the credit portfolio and effectively manage NPAs while maintaining profitability. In order to overcome the perceived risks, [6] present well structured and effective credit appraisal and monitoring system powered by Artificial Intelligence technologies. According to [6], the bank specific factors affecting the NPAs are:

i) Improper Due Diligence
ii) Lenient Credit Terms
iii) Loose Monitoring
iv) Collateral Free Loans
v) Frauds
vi) Wilful Defaults by Customers

Along with these bank specific factors, [6] also shares the external factors affecting NPAs in the form of a graph as below.

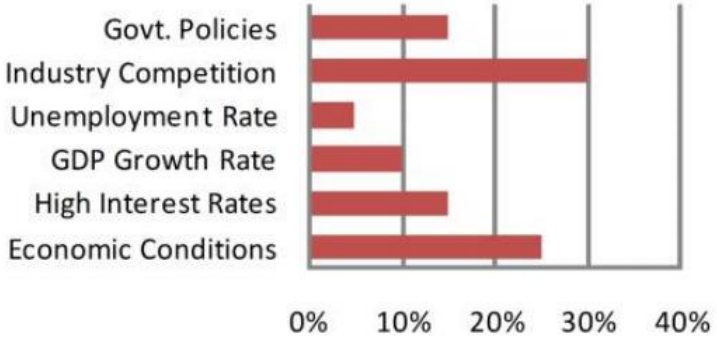

Fig. 01. External Factors affecting NPAs of a Bank [6]

V. DATA, PROCESSING AND ANALYSIS

This section of the research paper explains about the dataset used for training the machine learning models for predicting the NPA loans beforehand to prevent the financial losses to the banks and financial institutes.

A. DATASET

The collected dataset contains a total of 16 features and around 24000 records of data. All the features of the dataset along with their datatypes are listed below.

1) TICKETNO – int64
2) ADV_DATE – object
3) BRANCHID – int64
4) ADV_AMOUNT – float64
5) NETWEIGHT – float64
6) CAP_AMOUNT – float64
7) ACCR_INTEREST – float64
8) CAPITAL_PAY – float64
9) AVG_KARATAGE – float64
10) PRODUCTINTEREST – float64
11) PRODUCTNAME – object
12) GENDER – object
13) SECTOR_DESC – object
14) SUBSECTOR_DESC – object
15) TICKET_AGE – int64
16) CKASSIFI_BUCKET – object

The dataset collected contains the details of the customers the loans have been sanctioned. One of the 16 features is the remark of the account being an NPA or no. This dataset is used to train the machine learning model and get the model with the best accuracy as compared to this target feature in the collected dataset. Before training the machine learning model, this dataset is pre-processed against the missing and invalid values.

B. DATA PRE-PROCESSING

The dataset collected contains a total of 23,718 records for 16 features. The names and datatypes of all the columns are listed in the previous section. This section of the research paper explains the pre-processing of the dataset. Checking the null values, we found that there are no null values in the dataset, but there are certain invalid values in the "PRODUCTINTEREST" column. The interest can never be zero yet some of the cells contains "0" and this is an invalid value here. We replaced that value with the median of this column, thus making the values valid as per the feature. After changing the invalid value of the column "PRODUCTINTEREST", the descriptive statistics of the dataset was also improved.

Checking for the outliers in all the features of the dataset, it was found that the columns "CAP_AMOUNT" and "ACCR_INTEREST" had the outliers. These outliers were removed by keeping the difference of 0.75 and 0.25 quartile of both the features. The boxplot for both the features below and after the outlier removal can be seen below.

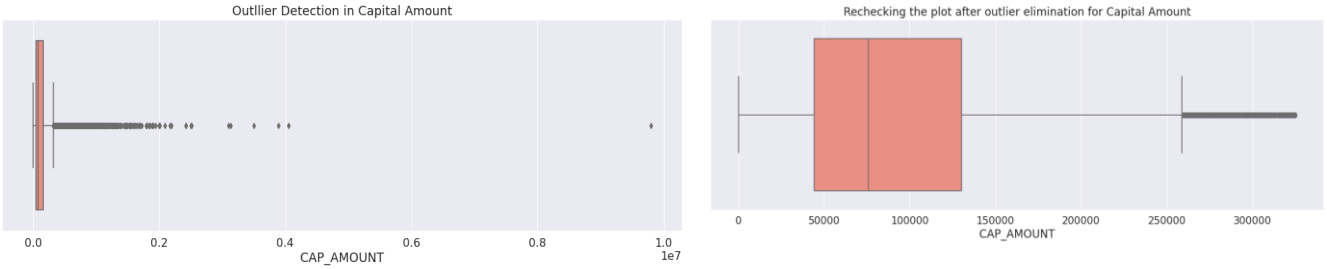

Fig.02. Outlier Analysis for "CAP_AMOUNT" before and after outlier removal

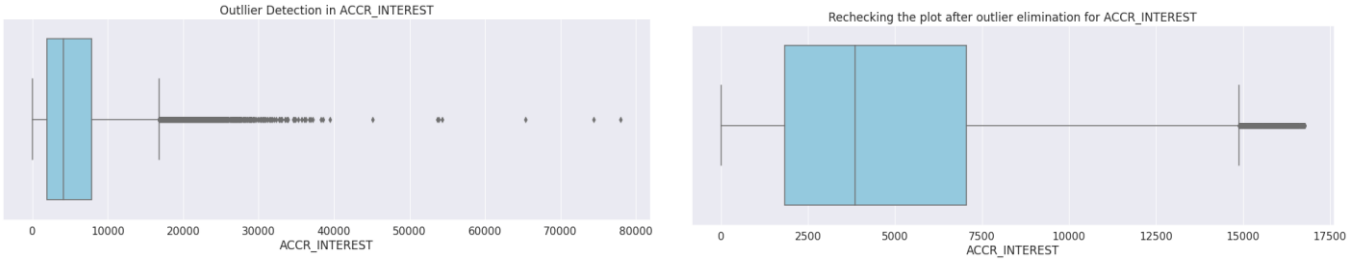

Fig.03. Outlier Analysis for "ACCR_INTEREST" before and after outlier removal

C. DATA ANALYSIS

Data Analysis is the main part of this project as it allows us to understand the data better and make an informed decision regarding the training of the machine learning models. This research project is aiming to create a predictive model. During analysis, multiple graphs were visualized to understand the data features their relationship with other data features and how they might influence the predictive power of the machine learning model. Some of the graphs from the data analysis section of the project can be seen below.

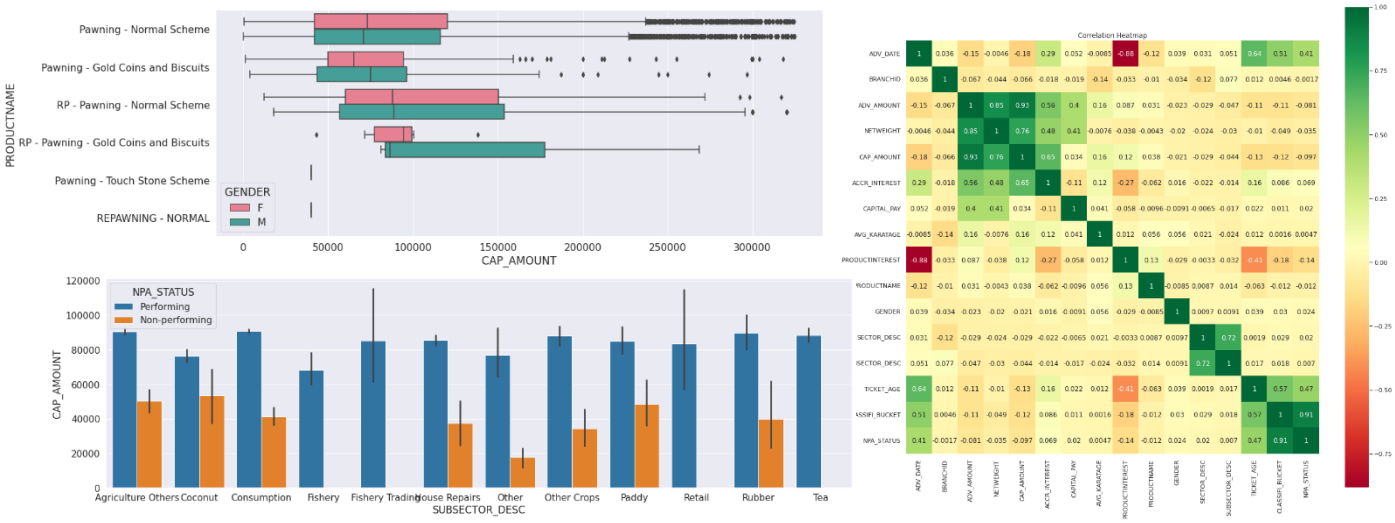

Fig. 04. Box Plot for CAP_AMOUNT against multiple features, Correlation matrix for the dataset and the NPA_STATUS for SUBSECTOR_DESC

Apart from these graphs, there are many other graphs in the data analysis section of this project which helped us understand the data and relation between different features of the dataset and reason behind the found relations. One main graph for this analysis is the Transaction class distribution showing the number of records for each class of the classification namely, "NPA" or "Not NPA". This graph shows that the collected dataset is skewed toward one type of class and this skewness can hamper the ML model performance. The correlation analysis of the dataset showed that there are few features whose correlation are greater than 85%. Further, visualizing the optimal number of clusters for the problem statement using the elbow method.

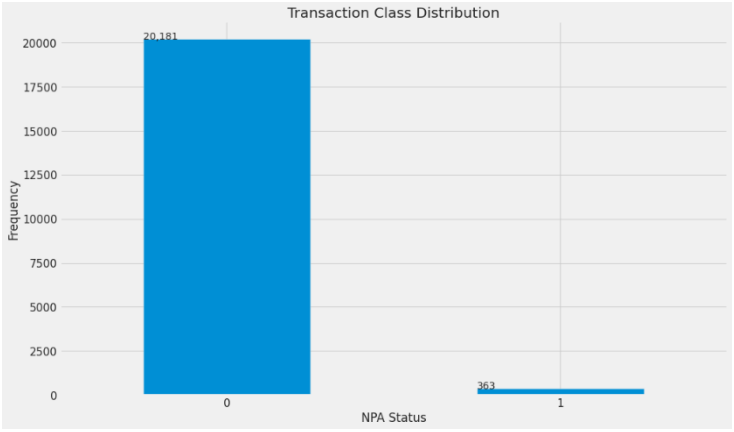

Fig. 05. Transaction Class Distribution calculating the number of NPA and Non-NPA records

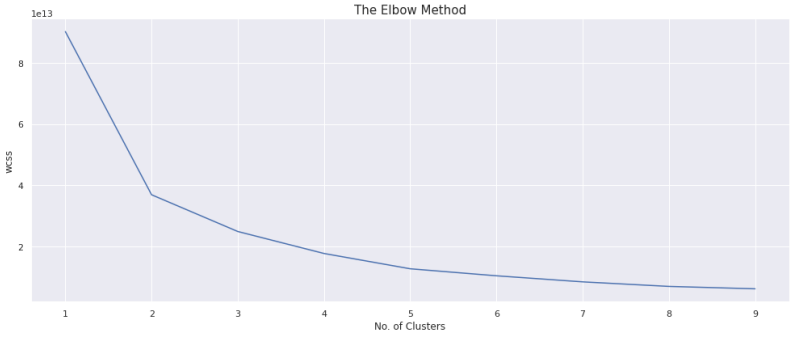

Fig. 06. Cluster Analysis, Elbow Method Graph to visualize optimal number of clusters

VI. RESULTS

After performing the cluster analysis and handling the imbalanced data using the random under sampler method, the final dataset is created and ready for training the machine learning models for predicting the NPA status of the loans listed in the dataset using the features. The problem statement for this project is a binary classification problem statement and thus we used the classification machine learning models for training them on this dataset and getting the accuracies. Previously, we had around 24k records in our dataset, but after random under sampling the shape of the dataset is (726, 12). Some of the features were dropped due to high correlation coefficient and some were not relevant for the prediction model. The machine learning models trained on this dataset are:

i) SVM Classifier
ii) PCA and SVM
iii) Logistic Regression on PCA selected features
iv) Logistic Regression with RFECV feature selection
v) SVM Classifier on PCA with 7 Components

A. MODEL PERFORMANCE

This section of the research project talks about the accuracies of the classification models trained using the dataset. The first model is the Support vector Machine using all the features in the dataset. The confusion matrix for this model shows that the model predicted all the records correctly with 119 true positive values and 99 true negative values and 0 false positive and false negative values giving us the accuracy of 100%.

The next model trained on the data is the PCA with SVM using 6 features. The accuracy of this model can be seen from the confusion matrix. This model scored the 97.7% of accuracy.

The third model is the logistic regression with PCA selected features, and this model achieved the accuracy of 97.2% . The dataset used for training these models is the one we got after performing the random under sampling.

The fourth model is the Logistic Regression model with RFECV feature selection. This model achieved the accuracy of only 88.5% in predicting the NPA status for the loan records listed in the dataset in 10 different features.

The fifth model is the SVM with PCA using 7 components. This model alike the first model achieved the accuracy of 100% on the dataset we used to train it.

All these accuracies are the test accuracies and we have tried best to avoid the overfitting or underfitting of the machine learning model on the training dataset. It can be observed that the first and the last model achieved the maximum accuracy of 100% on the dataset we collected and pre-processed for training the ML models. This is the highest and the best accuracy any machine learning model can achieve. To visualize the processing of these machine learning models on how these models are processing the data fed to these models.

B. EXPLAINABLE AI (XAI) FOR ML MODELS

This section of the research project shares the explainable AI graphs for the 2 models giving the maximum accuracy of 100% for predicting the NPA status for the loans. The explainable AI (XAI) is a tool we can use to study how the AI model processes the data fed to it to result the achieved accuracy of the machine learning models. In this project, we have used the SHAP explainable AI model to visualize different aspects of how the best ML models are processing our data to predict the NPA status as NPA or Non-NPA loans.

1) SVM CLASSIFIER WITH ALL THE FEATURES

Support Vector Machine is the machine learning model for classification problem statement. In this project the problem statement is the binary classification problem and SVM model trained using all the features of the data set gives us the 100% accuracy in predicting the NPA status for the loans. Following are the XAI graphs drawn using the SHAP XAI model for the SVM classifier. It can be observed that the feature "CAP_AMOUNT" is the feature influencing the prediction majorly followed by the "ACCR_INTEREST' and "CAPITAL_PAY" as the top three features influencing the prediction.

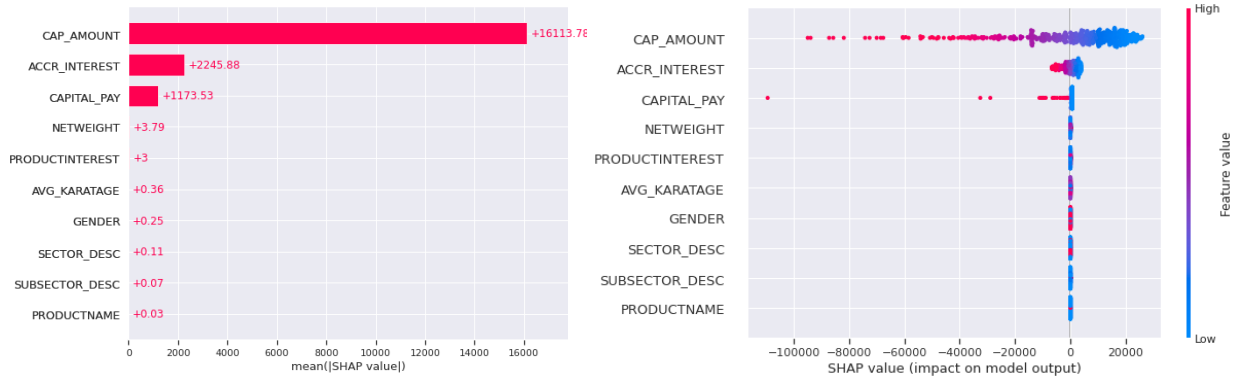

Fig. 07. Explainable AI (XAI) Graphs for Model 01 (Features influencing the model performance)

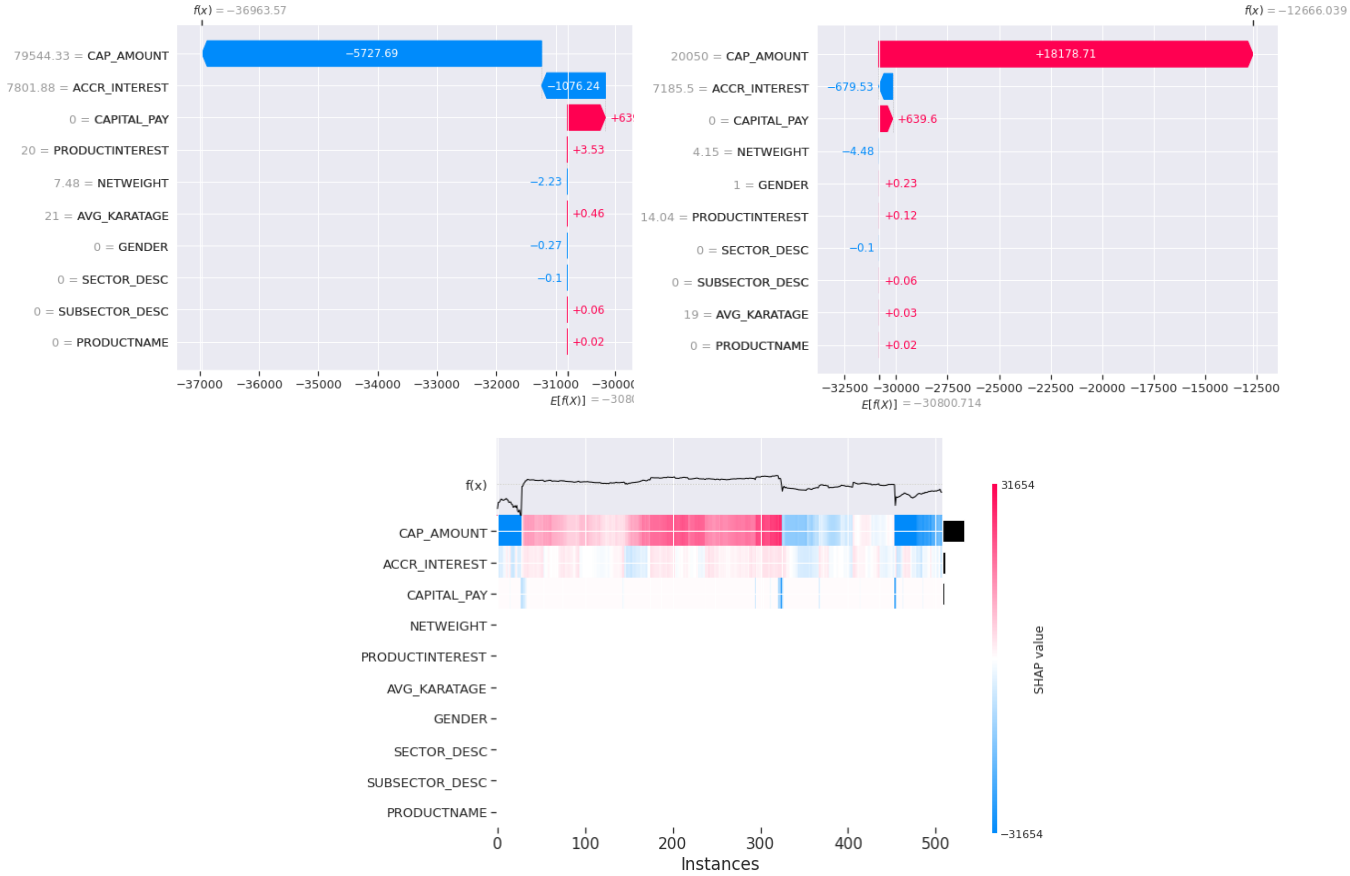

Fig. 08. Explainable AI (XAI) Graphs for Model 01(Feature influence for a random data point and the Heatmap)

The first 2 graphs shoes the Shap value for the features, the higher the shap value, the higher the influence of the feature on the predictive power of the machine learning model. The next two graphs shows the waterfall plot for the influence of features for selective data points from both the classes of classification. At the end is the heat map showing how the machine learning model is processing the features of the data to deliver the prediction.

2) SVM CLASSIFIER WITH 10 PCA COMPONENTS

Support Vector Machine is the machine learning model for classification problem statement and PCA is used for reducing the number of features in the dataset. The second model giving us the 100% accuracy is the SVM model with the PCA with 10 components. Following are the XAI graphs drawn using the SHAP XAI model for the SVM classifier with PCA and 10 components. It can be observed that PC2 is the feature influencing the prediction of this machine learning model followed by PC3 and PC1 as the first three features influencing the model performance the most.

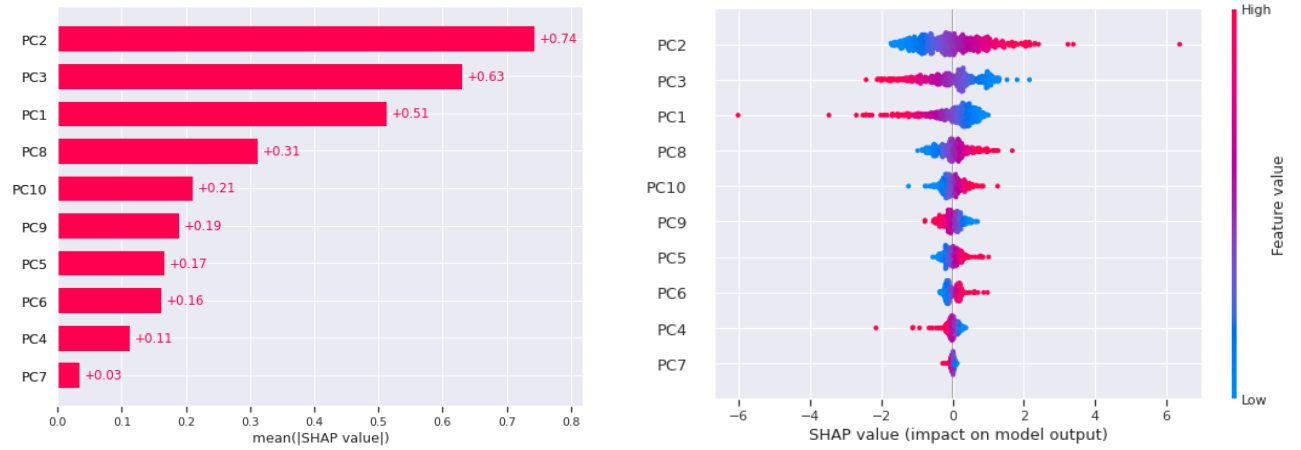

Fig. 09. Explainable AI (XAI) Graphs for Model 05 (Features influencing the model performance)

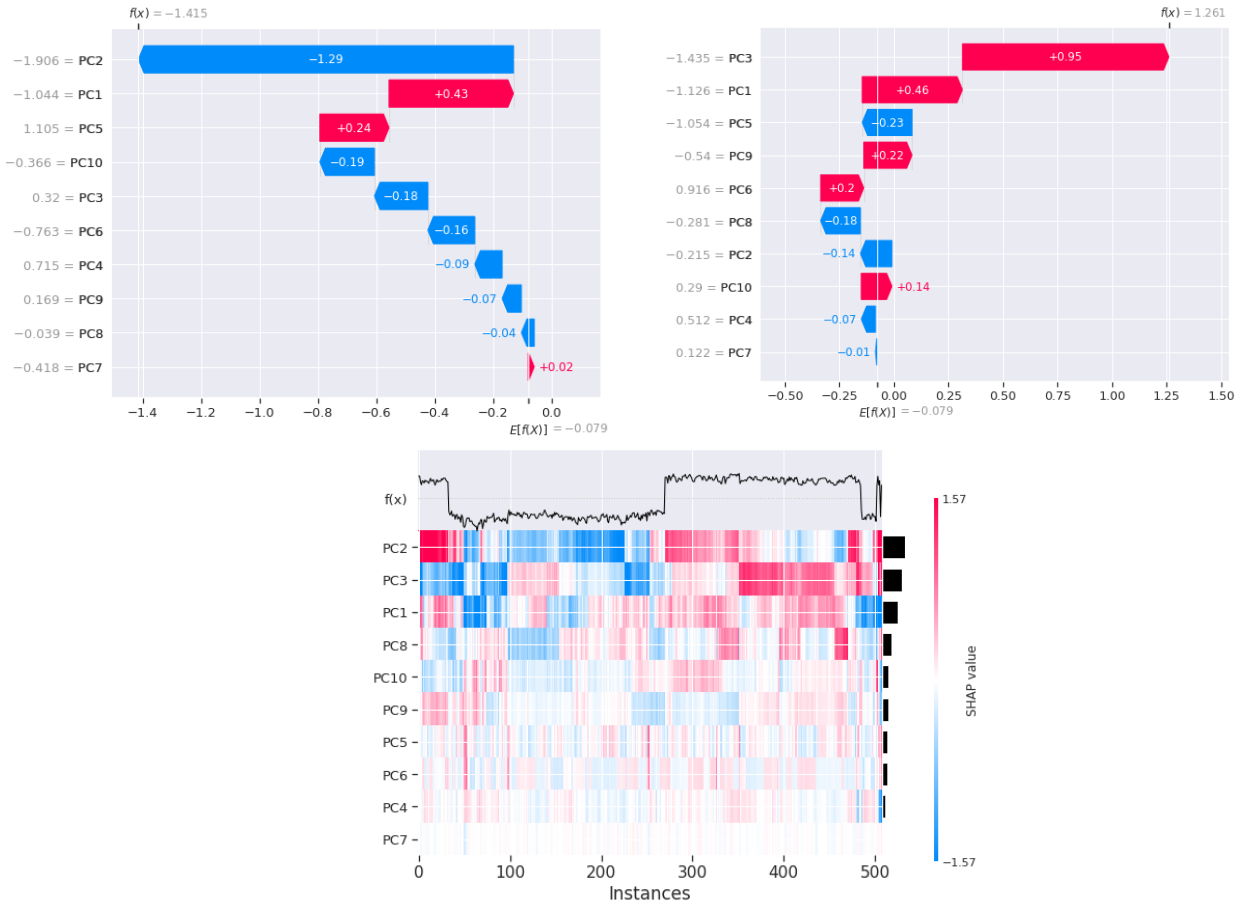

Fig. 10. Explainable AI (XAI) Graphs for Model 05(Feature influence for a random data point and the Heatmap)

VII. DISCUSSION

The research project presenting a machine learning model for predicting the NPA status as "NPA" or "Non-NPA" for each loan sanctioned by the bank using a dataset. The research project delivers a smart solution for predicting bad loans but at the same time have few problems and challenges. Many issues related to the prediction of the bad loans are addressed in the work but below is the list of some challenges this research project is not providing solutions to.

i) The project provides the machine learning models with best accuracy, but to use these machine learning models we need a system incorporated with these models and ready to use for the end user.

ii) The scope of this project was initially to provide a trained machine learning model to predict the NPA status of the loans and is very useful for the banks and financial institutes, but the ML models are trained on a dataset, and it can be possible that the dataset used does not match the real-life scenarios very much. An extended study is required to assess the dataset used resembles the real life.

VIII. FUTURE WORK

The research project was defined to present the machine learning model with best possible accuracy to predict the NPA status of the loans. Some of the components of this research project need to be addressed and considered under future enhancement,

- We need to create a system for using this machine learning model and make this model scalable for real life scenarios.
- We need to create a web application of software to use the proposed machine learning model for an end user may be a bank employee of financial professional.

IX. CONCLUSION

This research project presents a machine learning model for predicting

REFERENCES

[1] Nikhil Sonavane, Ambarish Moharil, Chirag Kedia, Mansimran Singh Anand, "*To Reduce Gross NPA and Classify Defaulters Using Shannon Entropy*", 2021

[2] Girija Attigeri, Manohara Pai M M, Radhika M Pai, *"Framework to predict NPA/Willful defaults in corporate loans: a big data approach*", International Journal of Electrical and Computer Engineering (IJECE), 2019
[3] Entropy-Based Financial Asset PricingMiha´ly Ormos*, Da´vid ZibriczkyOrmos M, Zibriczky D, "*Entropy Based Financial Asset Pricing*", PLoS ONE 9(12): e115742 doi: 10.1371/journal.pone. 0115742, 2014
[4] Bawa, J. K., Goyal, V., Mitra, S. K., & Basu, S, "*An analysis of NPAs of Indian banks: Using a comprehensive framework of 31 financial ratios*", IIMB Management Review, 31(1), 51-62, 2019
[5] Kadanda, D., & Raj, K, "*Non-performing assets (NPAs) and its determinants: a study of Indian public sector banks*", Journal of Social and Economic Development, 2018.
[6] Dr. Pratapsinh Chauhan, Shaleen Srivastava, "*Artificial Intelligence: A Way Forward for NPA Management in Indian Banking Sector (Special Reference to Global Trends*", Emerging Trends and Innovations in Modern Management, 2020.
[7] Pandey, Manish Kumar, Mamta Mittal, Karthikeyan Subbiah, "*Optimal Balancing & Efficient Feature Ranking Approach to Minimize Credit Risk*", International Journal of Information Management Data Insights, 2021
[8] Mohammad Ahmed Sheikh, Amit Kumar Goel, Tapas Kumar, *"An Approach for Prediction of Loan Approval using Machine Learning Algorithm"*, IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4, 2020.