International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 11 | November - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Credit Risk Assessment Using Machine Learning in Banking Sector

Gunjan Saini
Computer Science & Engineering
Chandigarh University
Mohali, India
gunjan.saini010@gmail.com

Sushant Singh
Computer Science & Engineering
Chandigarh University
Mohali, India
sushant.singh0630@gmail.com

Sayani Debnath
Computer Science & Engineering
Chandigarh University
Mohali, India
sayanidebnath1621@gmail.com

***Abstract*— The possibility of financial loss as a result of a borrower's failure to fulfil their financial obligations is referred to as credit risk. Even though there are many elements that affect credit risk, rigorous evaluation before loan approval (credit scoring) and continued observation of client payments and behavior can help reduce the risk of non-performing assets (NPAs) and fraud. The necessity for reliable methods of projecting loan performance has been highlighted by a noticeable rise in NPAs and fraud cases over the past few years. Researchers are looking at various ML techniques to increase the accuracy of credit risk assessment in light of the increasing expansion of artificial intelligence (AI), especially in machine learning (ML), driven by increased internet access, data availability, and computer capacity. In this model, we used Logistic Regression, Support Vector Machine, Random Forest and XGB algorithms to get the most accurate result. While customer profiles from credit rating and scoring firms are paid for by banks, continuing research is focused on utilizing ML to improve credit risk assessment. The present paper reviews 15–20 credit risk evaluation studies that were published between 2012 and March 2023. The analysis not only looks at the present limits of these studies but also identifies new research trends in the area.**

Keywords— NPA, ML, AI, XGB.

I. INTRODUCTION

In the aftermath of the global financial crisis, the banking sector has placed greater importance on risk management, resulting in a continuous focus on the identification, quantification, reporting, and mitigation of risks. Extensive research, both in academic and industry spheres, has concentrated on advancements in banking and risk management, as well as the current challenges and those on the horizon. Concurrently, the influence of machine learning in business applications has seen a significant surge, with numerous solutions already in operation and many more undergoing explorations. According to McKinsey & Co, by 2025, risk management functions within banks are expected to undergo a substantial transformation. This transformation is driven by the expansion and evolution of regulations, shifts in customer expectations, and the changing nature of risks. The development of new goods, services, and risk management techniques is becoming simpler with the use of cutting-edge analytics and developing technology. By identifying intricate, non-linear patterns in massive data sets, machine learning—which is acknowledged as a critical technology for risk management—plays a critical role in producing risk models that are more accurate. As more data is added, these models can continuously get better at predicting results. It is anticipated that machine learning will be used across a bank's risk management portfolio, potentially revolutionizing the way banks handle risk management procedures.

Machine Learning is a research study to examine how machine learning, acknowledged as a new business accelerator, has been investigated in relation to risk management in the banking industry. It also aims to identify areas that require additional research. This thorough review's main goal is to investigate and evaluate the use of machine learning methods in the field of banking risk management. It also looks for problems or areas of risk management that haven't gotten enough attention and makes suggestions for more study. This study develops a risk classification system based on an examination of banks' annual reports in order to identify dangers that are unique to banks instead of depending only on previously published research. The literature study assesses the domains of machine learning applied to banking risk management that have been investigated. It looks at the specific risk categories that machine learning has prioritized and the methods that it has utilized. It also specifies the precise machine learning techniques that are used, both more generally and in particular domains. An outline of risk management in the banking industry is provided in Section 2.1, which also provides information on key risk categories and the instruments and techniques employed in risk management. A quick overview of machine learning and its real-world uses is provided in Section 2.2. Section 3 delves into the research methodology and analyses previous studies on machine learning in banking risk management, emphasizing areas that haven't been thoroughly investigated in academic research. In Section 4, the major findings from the review are explored, along with possible roadblocks and areas that call for more research. A summary of the study's overall findings is given in Section 5, which ends with a list of other subjects or problems related to banking

risk management that could be gained from additional study using machine learning.

II. LITERATURE REVIEW

A crucial procedure in the banking industry is assessing credit risk, which entails separating trustworthy applicants from possibly dishonest ones based on credibility. Researchers have explored various algorithms to enhance the accuracy of credit risk assessment in light of the advancements in machine learning. This review highlights the use and effectiveness of algorithms like Random Forest, Support Vector Machine (SVM), and Logistic Regression while summarizing the current body of research on credit risk assessment in banking. The main goal is to comprehend the stated accuracy percentages of these models.

Financial institutions use credit risk evaluation to determine how likely it is that borrowers will default on their loans during the lending process. Banks have traditionally employed credit scoring systems and traditional statistical techniques. However, new developments in machine learning have made more accurate and useful substitutes available.

Numerous studies have demonstrated how well machine learning algorithms work to improve the accuracy of credit risk assessment. These algorithms predict whether an applicant is likely to default on their loan based on historical data. The Random Forest, Support Vector Machine, and Logistic Regression algorithms are some of the frequently used ones. In particular, Random Forest is an ensemble learning technique that creates predictions by combining several decision trees. Because of its ability to handle complex datasets and identify non-linear relationships between various factors, it has gained popularity in the credit risk assessment field. Previous studies have consistently demonstrated favorable results when using Random Forest to evaluate credit risk.

The Support Vector Machine (SVM) is another well-known machine learning algorithm used in credit risk assessment. SVM can be tuned to handle imbalanced datasets, a common scenario in credit risk evaluation, and is particularly good at identifying decision boundaries in high- dimensional spaces. Several research works have highlighted how well SVM predicts credit risk with high accuracy rates. In contrast, a popular statistical technique with a lengthy history in credit risk assessment is logistic regression. Its simplicity and interpretability make it stronger even though it may not be as complex as Random Forest or SVM. As we mentioned in our research paper, Logistic Regression has shown to be able to produce high accuracy rates in certain situations.

To determine how well different machine learning algorithms perform in the evaluation of credit risk, researchers have conducted comparative analyses. SVM, ROC-AUC, recall, accuracy, precision, and F1-score for models like Random Forest, SVM, and Logistic Regression are often examined in these assessments. Logistic Regression yielded the highest accuracy rate in our dataset, according to our analysis.

The literature review indicates that machine learning algorithms such as Random Forest, Support Vector Machines, and Logistic Regression are useful instruments for evaluating credit risk in banking industry. Every algorithm has advantages and disadvantages of its own, and the dataset and specific problem it is meant to solve determine how well it works. Our research's finding that Logistic Regression provides the highest accuracy rate highlights the importance of choosing the most suitable algorithm for a given credit risk assessment task. Further research may be necessary to explore the generalizability of these results to a broader range of datasets and banking contexts.

III. DESIGN AND METHODOLOGIES

Component 1:

- Collection of Data

Component 2:

- Cleaning of Data

Component 3:

- Exploratory data analysis on dataset

Component 4:

- Applying different models and comparison

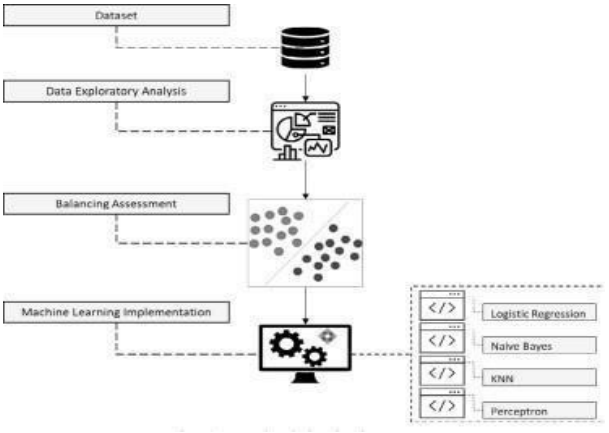


Fig.1 Implementation

IV. IMPLEMENTATION

1. ER-DIAGRAM

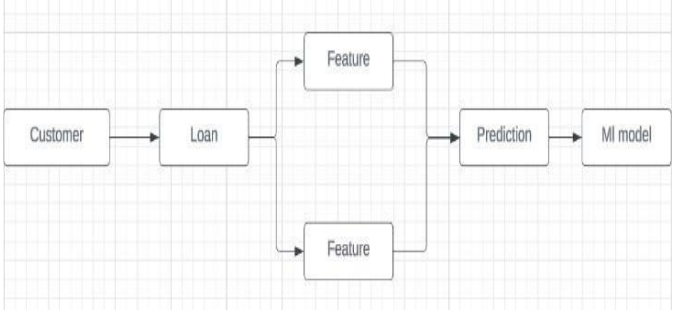


Fig.2 ER- Diagram

2. ARCHITECTURE DIAGRAM

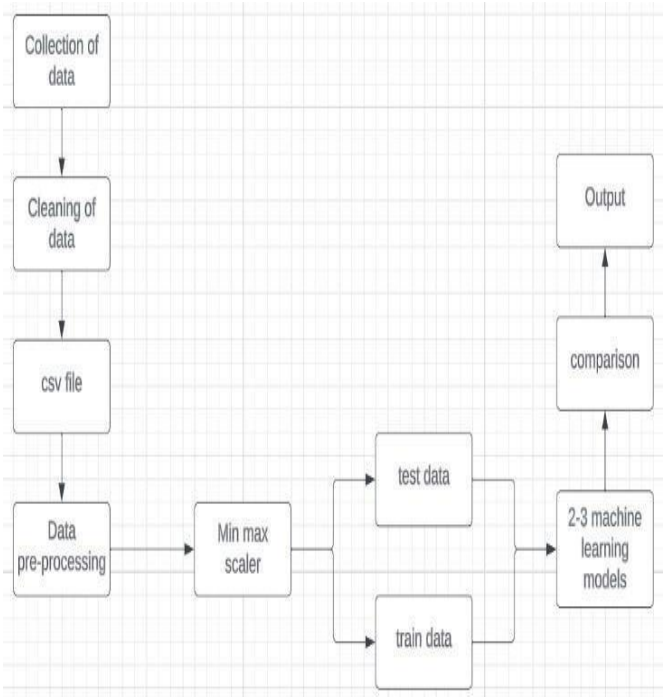


Fig.3 Architecture Diagram

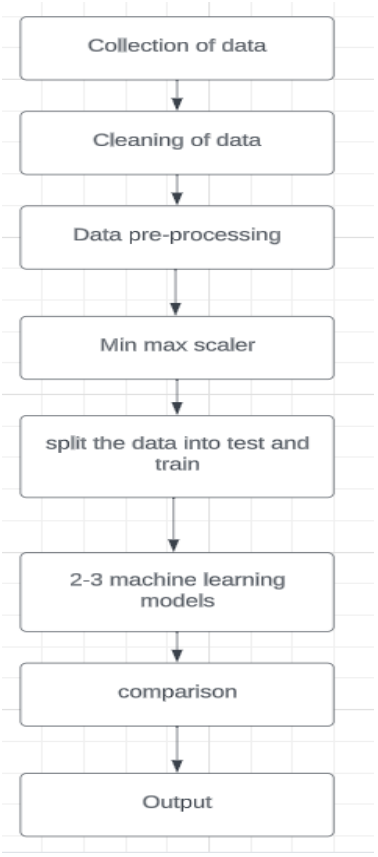
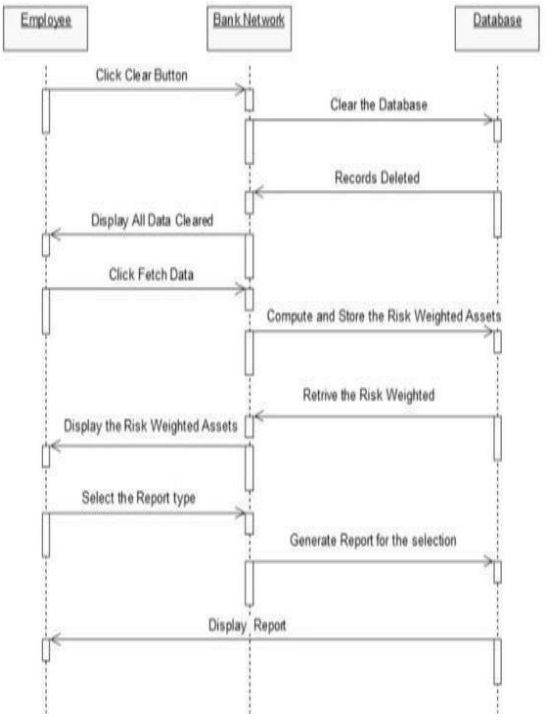


Fig.4 Data Flow Diagram

4. SEQUENCE DIAGRAM



V. DESIGN CONSTRAINTS

In this context, we are elucidating the process of interaction among different modules and components to facilitate the functioning of the system within our model. This system design has been created to fulfill the user's requirements through the application of our algorithm and statistical data. The design also encompasses the essential functions needed to comprehend the system's construction process, capturing the pivotal elements in building a comprehensive understanding of how the system operates.

There are three key functionalities:

- **Adding New Data and Editing Existing Data:**
  - Users have the capability to either add new data to our model or edit the existing data of users whose information is already stored in the database.
  - Editing existing data may involve scenarios such as an organization taking another loan from a bank and updating its record in the dataset to reflect this change.
  - Adding new data encompasses situations where a new organization seeks a loan from a bank. In this case, the bank collects all relevant historical information, including Non-Performing Assets (NPAs) and other relevant data, to assess whether the organization presents a credit risk or potential fraud.
- **Credit Risk Analysis:**
  - Users input data into our model, which then employs specific algorithms to predict the credit risk associated with the provided information.
  - This data can be in the form of a dataset related to

an organization, which includes past records of the company's loan history and whether it has been a source of debt or loss for the bank.

- **Credit Risk Detection:**
- Utilizing historical data and the information provided by the user, our model assesses whether the user or organization is a suitable candidate for a loan.
- The selection of the appropriate model for this task depends on the company's requirements. Our project involved comparing three different models to determine the most effective one for credit risk prediction.

VI. INPUT AND OUTPUT

In this study, we utilized historical credit card transaction data obtained from a prominent credit card company. This dataset encompassed a wide range of information, including details about customer demographics, credit bureau records, and credit card transaction history.

	A	B	C	D	E
1	age	ed	employ	address	income
2	41	3	17	12	176
3	27	1	10	6	31
4	40	1	15	14	55
5	41	1	15	14	120
6	24	2	2	0	28
7	41	2	5	5	25
8	39	1	20	9	67
9	43	1	12	11	38
10	24	1	3	4	19
11	36	1	0	13	25
12	27	1	0	1	16
13	25	1	4	0	23
14	52	1	24	14	64
15	37	1	6	9	29
16	48	1	22	15	100
17	36	2	9	6	49
18	36	2	13	6	41
19	43	1	23	19	72
20	39	1	6	9	61
21	41	3	0	21	26
22	39	1	22	3	52
23	47	1	17	21	43
24	28	1	3	6	26
25	29	1	8	6	27
26	21	2	1	2	16
27	25	4	0	2	32
28	45	2	9	26	69
29	43	1	25	21	64

Fig.6 Dataset

```
[4] df.isnull().sum()

age          0
ed           0
employ       0
address      0
income       0
debtinc      0
creddebt     0
othdebt      0
default    450
dtype: int64
```

```
[5] df.value_counts()

age  ed  employ  address  income  debtinc  creddebt  othdebt   default
20   1   4       0        14      9.7      0.200984  1.157016  1.0       1
39   1   10      4        31      4.8      0.184512  1.303488  0.0       1
             0       8        39      7.9      1.066026  2.014974  0.0       1
             2       15       22      23.1     1.915914  3.166086  1.0       1
             4       9        38      6.5      1.178190  1.291810  0.0       1
                                                                          ..
30   2   8       4        56      6.4      0.333312  3.250688  0.0       1
             10      4        22      16.1     1.409716  2.132284  0.0       1
             12      9        68      20.1     2.856612  10.811388 0.0       1
                              98      7.2      2.935296  4.120704  0.0       1
56   1   11      20       59      15.0     4.672800  4.177200  0.0       1
Length: 700, dtype: int64

[6] df = df.dropna()
```

Fig.7 Removing Null Values

VII. SUPPORT VECTOR MACHINE

```
sv.score(xtest,ytest)

0.7928571428571428

[19] model = GridSearchCV(sv,{
         'C':[0.1,0.2,0.4,0.8,1.2,1.8,4.0,7.0],
         'gamma':[0.1,0.4,0.8,1.0,2.0,3.0],
         'kernel':['rbf','linear']
     },scoring='accuracy',cv=10)

model.fit(xtrain,ytrain)

▸ GridSearchCV
▸ estimator: SVC
    ▸ SVC

[21] model.best_params_

{'C': 0.1, 'gamma': 0.1, 'kernel': 'linear'}

[22] model2 = SVC(C=0.1,gamma=0.1,kernel='linear')
     model2.fit(xtrain,ytrain)
     model2.score(xtest,ytest)

0.8214285714285714
```

Fig.8 SVM

In our research, we harnessed the power of the Support Vector Machine (SVM) model to carry out data classification and training according to our specified model. Our objective was to refine the model to achieve higher accuracy. We are pleased to report that through our efforts, we attained an impressive accuracy rate of 82.14%.

VIII. RANDOM FOREST

A Random Forest is a supervised machine learning algorithm capable of addressing both classification and regression problems. Its operation involves the creation of numerous decision trees during the training process, and when making predictions, it combines the outputs of all these trees within the "forest" to arrive at the final prediction. This ensemble approach is key to its effectiveness in various tasks.

```
[13] #random forest
     rfc = RandomForestClassifier(n_estimators=200)

[14] rfc.fit(xtrain,ytrain)

     RandomForestClassifier
     RandomForestClassifier(n_estimators=200)

[15] rfc.score(xtest,ytest)

     0.8

[16] rfc2 = cross_val_score(estimator=rfc,X=xtrain,y=y
     rfc2.mean()

     0.7785714285714286
```

Fig.9 Random Forest

IX. LOGISTIC REGRESION

Binary classification problems are frequently addressed through the statistical technique of logistic regression. Its principal goal is to model, by considering one or more predictor variables, the probability of a binary outcome, such as 1/0, Yes/No, or True/False. It uses the sigmoid function, sometimes referred to as the logistic function, to accomplish this. Any real value entered into this function is converted to a value between 0 and 1, which allows it to be understood as probability. Based on the given predictor variables, logistic regression essentially provides a means of estimating the probability of a given binary outcome.

```
[23] #logistic regression
     lr = LogisticRegression()
     lr.fit(xtrain,ytrain)
     lr.score(xtest,ytest)

     0.8357142857142857

[24] yp = lr.predict(xtest)
     c= confusion_matrix(ytest,yp)
     fig ,ax = plt.subplots(figsize=(20,10))
     sns.heatmap(c,ax=ax)

     <Axes: >
```

Fig.10 Logistic Regression

X. DATASET AND STANDARDIZATION

1. Splitting the data set

In this step, For the purpose of training and testing our model, we are partitioning the entire dataset into predetermined proportions. The purpose of this partitioning is to make sure that part of the data is used to train the model and part of it is set aside for testing and performance evaluation.

```
[11] xtrain,xtest,ytrain,ytest = train_test_split(x,y,test_size=0.2,random_state=42)

[12] sc = StandardScaler()
     xtrain=sc.fit_transform(xtrain)
     xtest=sc.fit_transform(xtest)
```

2. Data Standardization

Data standardization is the process of transforming data into a consistent and standardized format that is easily understandable by computers and machine learning models. This ensures that the data is in a uniform and compatible structure, making it easier for algorithms to process and derive meaningful insights or predictions.

```
[12] sc = StandardScaler()
     xtrain=sc.fit_transform(xtrain)
     xtest=sc.fit_transform(xtest)
```

Fig.12 Data Standardization

XI. MODEL TRAINING

In the realm of machine learning, training a model is the process of developing a system capable of analyzing data and making predictions or decisions by leveraging the patterns and information it has acquired from the training data. The trained model utilizes specific algorithms and techniques to provide the best possible output or results in accordance with what it has cleaned from the data during the training phase.

In this context, we are utilizing the Random Forest algorithm to train our model with the aim of achieving higher accuracy.

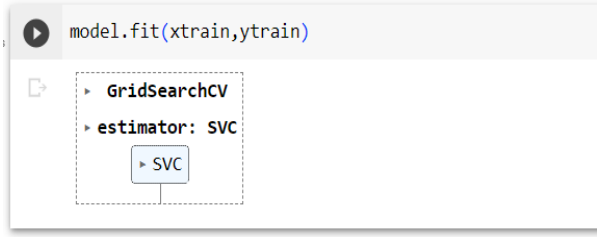
```
[13] #random forest
     rfc = RandomForestClassifier(n_estimators=200)

[14] rfc.fit(xtrain,ytrain)

     RandomForestClassifier
     RandomForestClassifier(n_estimators=200)

[15] rfc.score(xtest,ytest)

     0.8

[16] rfc2 = cross_val_score(estimator=rfc,X=xtrain,y=ytrain,cv=10)
     rfc2.mean()

     0.7785714285714286
```

Fig.13 Random Forest

Our model has obtained an accuracy of 77.85% by using the Random Forest algorithm.

2. S.V.M. MODEL TRAINING

In this instance, we are utilizing the Support Vector Machine (SVM) algorithm with a linear kernel to train our model with the aim of improving accuracy.




```
[21] model.best_params_

     {'C': 0.1, 'gamma': 0.1, 'kernel': 'linear'}

[22] model2 = SVC(C=0.1,gamma=0.1,kernel='linear')
     model2.fit(xtrain,ytrain)
     model2.score(xtest,ytest)

     0.8214285714285714
```

Fig.14 SVM

By applying the Support Vector Machine (SVM) algorithm with a linear kernel, our model has achieved a notable accuracy rate of 82.14%.

3. LOGISTIC REGRESSION

Logistic regression is a statistical method primarily used for binary classification problems. It is designed to model the probability of a binary outcome (e.g., 1/0, Yes/No, True/False) based on one or more predictor variables. This technique employs the logistic function, also known as the sigmoid function, to transform any real-valued input into a value within the range of 0 to 1. This output can be interpreted as probability, indicating the likelihood of the binary outcome occurring.

```
✓ 0s [23] #logistic regression
         lr = LogisticRegression()
         lr.fit(xtrain,ytrain)
         lr.score(xtest,ytest)

         0.8357142857142857
```

By applying Logistic Regression, our model has achieved a notable accuracy rate of 83.57%.

XII. EVLAUTION OF OUR MODEL

In our analysis, we have applied an evaluation approach to assess the performance of our model. This evaluation process helps in determining how effectively our model is working. Specifically, we have utilized a confusion matrix on our highest accuracy model, which is the logistic regression model. Additionally, we have created a heatmap

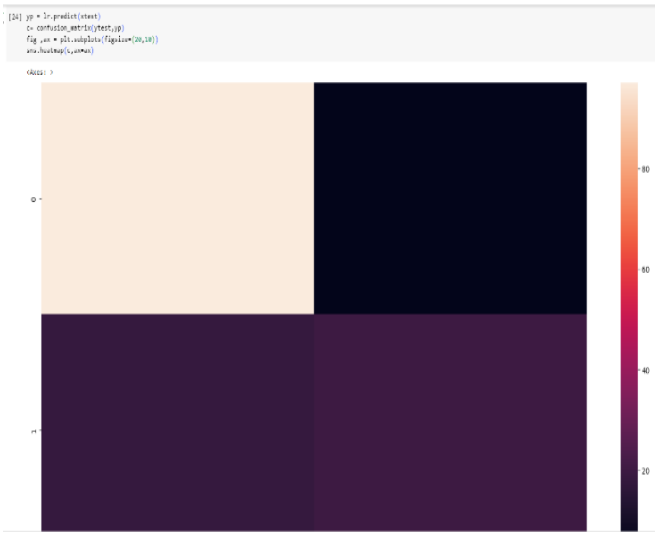


Fig.16 Heat Map

XIII. CONCLUSION AND FUTURE WORK

In this research, we have designed and implemented a machine learning model and algorithm to evaluate the accuracy of various algorithms in credit risk assessment. Our aim is to enhance the user experience and the efficiency of credit risk prediction, ultimately making it more accessible for organizations through user interfaces or device applications.

In summary, the utilization of machine learning techniques in credit risk assessment has demonstrated considerable potential in improving the precision, speed, and effectiveness of creditworthiness evaluations. Financial institutions will benefit greatly from this progress as it will enable them to minimize risks, make well-informed decisions, and maximize their lending portfolios. Credit risk has been predicted with exceptional accuracy using a variety of machine learning models, including decision trees, random forests, support vector machines, neural networks, and ensemble approaches. Furthermore, these models' performance has been greatly enhanced by the thoughtful application of feature engineering and data preprocessing techniques. It is critical to incorporate additional data sources and continuously improve the models in order to further improve predictive accuracy and create a more comprehensive credit assessment procedure.

- Incorporating Non-traditional Data Sources: Explore the integration of unconventional data sources, such as social media activity, online behavior, and transactional data, alongside traditional credit information to obtain a more holistic view of creditworthiness.

- Improving Model Interpretability: Provide techniques that will improve the interpretability and transparency of machine learning models used in credit risk assessment, fostering stakeholder confidence, and guaranteeing adherence to legal requirements.
- Real-time Risk Assessment: Focus on implementing real-time credit risk assessment systems that can adapt to changing financial behaviors and economic conditions, enabling faster decision-making for lenders.
- Addressing Class Imbalance and Bias: Create techniques to handle class imbalances and mitigate bias in credit risk assessment models, ensuring fairness and impartial lending practices.
- Multi-modal Learning: Investigate the potential of incorporating multiple data modalities, including structured, unstructured, and temporal data, for a more comprehensive understanding of credit risk, potentially using approaches such as multimodal learning and time-series analysis.
- Global Standardization and Collaboration: Foster collaboration among financial institutions, researchers, and regulatory bodies to establish standardized frameworks, datasets, and evaluation metrics, facilitating comparisons and benchmarking of various machine learning models in credit risk assessment.
- Cybersecurity and Privacy Measures: Concentrate on enhancing the security and privacy aspects of credit risk assessment systems, ensuring robust protection of sensitive financial information and compliance with data privacy regulations.

By addressing these aspects, the application of machine learning in credit risk assessment can be further refined, ultimately leading to more accurate risk predictions, and fostering a more inclusive and responsible lending environment.

XIV. ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to Mr. Adil Husain Rathar for his invaluable guidance and mentorship throughout this project and research endeavor. We are sincerely thankful for his unwavering support and the clear direction he provided, which enabled us to successfully complete our tasks within the stipulated time frame.

We also extend our deep appreciation to all the members of our group who collaborated tirelessly, contributing their expertise and effort to ensure the successful completion of this research project. Their dedication, teamwork, and valuable contributions played a significant role in finding effective solutions and achieving our research goals.

Lastly, we would like to convey our sincere thanks to Chandigarh university for providing us with a conducive platform and access to a comprehensive library, which greatly facilitated our research efforts. These resources were instrumental in our ability to conduct meaningful research and contribute to the academic community.

XV. REFERENCES

[1] Haq, I. U., Hassan, I. U., & Shah, H. A. (2023, April). Machine Learning Techniques for Result Prediction of One Day International (ODI) Cricket Match. In 2023 IEEE 8th International Conference for Convergence in Technology (I2CT) (pp. 1-5). IEEE.

[2] Haq, I. U., Rather, A. H., & Kaur, G. (2023, July). A Comparative Analysis of Machine Learning Algorithms for the Early Prediction of Cardiovascular Disease. In 2023 2nd International Conference on Edge Computing and Applications (ICECAA) (pp. 987-993). IEEE.

[3] Rather, A. H., & Haq, I. U. (2023, April). Intelligent Framework for Early Prediction of Diabetic Retinopathy: A Deep Learning Approach. In International Conference on Paradigms of Communication, Computing and Data Analytics (pp. 377-388). Singapore: Springer Nature Singapore.

[4] Shah, H. A., Hassan, I. U., & Haq, I. U. CONTROVERSIAL ISSUES AND DIGITAL MARKETING. CONTEMPORARY ISSUES IN, 23.

[5] ul Haq, I., Kumar, N., Koul, N., Verma, C., Eneacu, F. M., & Raboaca, M. S. Machine Learning Techniques for Result Prediction of One Day International (ODI) Cricket Match. In Proceedings of International Conference on Recent Innovations in Computing: ICRIC 2022, Volume 2 (p. 95). Springer Nature.

[6] Wu, C., Guo, Y., Zhang, X., Xia, H., 2010. Study of personal credit risk assessment based on support vector machine ensemble. Int. J. Innovative 6 (5), 2353–2360.

[7] P. M. Addo, D. Guegan, and B. Hassani, “Credit Risk Analysis Using Machine and Deep Learning Models,” SSRN Electronic Journal, 2018.

[8] G. V. Attigeri, M. M. M. Pai, and R. M. Pai, “Credit Risk Assessment Using Machine Learning Algorithms,” Advanced Science Letters, vol. 23, no. 4, pp. 3649–3653, Jan. 2017.

[9] A. E. Khandani, A. J. Kim, and A. W. Lo, “Consumer Credit Risk Models Via Machine-Learning Algorithms,” SSRN Electronic Journal, 2010.

[10] H. A. Bekhet and S. F. K. Eletter, “Credit risk assessment model for Jordanian commercial banks: Neural scoring approach,” Review of Development Finance, vol. 4, no. 1, pp. 20–28, 2014.

[11] M. Saar-Tsechansky and F. Provost, “Handling missing values when applying classification models”, Journal of Machine Learning Research, vol. 8(Jul), pp. 1623– 1657, 2007.