Credit Risk Analysis: Enhancing Accuracy and Reliability for Effective Lending

I. Abstract

The credit risk analysis process is crucial for lenders to determine the level of risk involved in lending to potential borrowers. This process involves evaluating various factors, including the 5 Cs of credit and financial and qualitative analysis techniques, to estimate the probability of default and the subsequent risk of credit loss. However, despite the use of these factors and techniques, lenders may still face challenges in accurately assessing credit risk, which could result in potential financial losses.

This paper explores the challenges of credit risk analysis and proposes a hypothesis that incorporating advanced data analytics techniques, such as machine learning algorithms and predictive modelling, into the credit risk analysis process can improve the accuracy and reliability of credit risk assessments. The use of these techniques can help lenders make more effective lending decisions and reduce the risk of financial loss.

Overall, this paper provides insights into the credit risk analysis process and proposes a potential solution to enhance the accuracy and reliability of credit risk assessments, ultimately helping lenders minimize the risk of financial loss.

II. Introduction

Lenders need to determine the creditworthiness of potential borrowers and their ability to honour their debt obligations to minimize the risk of financial loss. To achieve this, lenders use various factors financial and qualitative techniques to estimate the probability of default and the subsequent risk of credit loss. However, despite the use of these factors and techniques, lenders may still face challenges in accurately assessing the credit risk of borrowers, which could result in potential financial losses.

The project targets on incorporating data analytics techniques, including machine learning algorithm and predictive modelling, into credit risk analysis process, lenders can improve the accuracy and reliability of credit risk assessments, resulting in more effective lending decisions and reduced risk of financial loss.

To understand the progress of the project we first need to grasp the concept of creditworthiness. “Creditworthiness” refers to the measure of an individual or entity’s ability to repay their debts or obligations. It is a critical factor that lenders such as banking institutions consider while evaluating whether to extend credit to a borrower.

Typically credit scores, which are calculated based on credit reports, are used to assess “creditworthiness.” Credit reports contain information about a borrower’s past credit history, including their payment history, credit utilization, and duration credit. A higher credit score typically translates to more “creditworthiness” which may result in better loan terms such as lower interest rates. Other factors that can be considered are employment history, income stability, and collateral.

In the project we will examine such a dataset which includes some of the listed parameters by performing exploratory data analysis. Upon examination, the data will be pre-processed and selectively passed through supervised and unsupervised Machine Learning algorithms to classify a borrower as creditworthy or not. The machine learning techniques that have been incorporated in the project are as follows:

- Logistic Regression
- K-Nearest Neighbours
- Decision Tree
- Random Forest
- Support Vector Classification

III. Literature Review

In this section, we discuss the literature that explored Credit Risk Analysis using various machine learning methods

Qasem and Nemer (2018) studied the use of Extreme Learning Machine (ELM) for credit risk analysis. They developed an ELM model to predict the credit risk of borrowers and compared it to other machine learning techniques, including Support Vector Machine (SVM) and Artificial Neural Networks (ANN). The study found that ELM outperformed other techniques in terms of

accuracy and training time. The authors concluded that ELM is a promising technique for credit risk analysis and could be useful for financial institutions in improving their credit risk assessment process.[1]

Bequé and Lessmann (2017) evaluated the performance of Extreme Learning Machines (ELM) compared to traditional machine learning techniques, such as Support Vector Machines (SVM) and Random Forests (RF), in credit scoring. Using a dataset of 10,000 borrowers with 24 input features, they found that ELM outperformed SVM and RF in terms of prediction accuracy, sensitivity, and specificity. The study demonstrated the effectiveness of ELM in predicting credit risk and highlighted its potential to improve credit risk assessment for financial institutions. [2]

Bao, Lianju, and Yue (2019) proposed an integrated approach to credit risk assessment that combines unsupervised and supervised machine learning techniques. The authors used Principal Component Analysis (PCA) and K-Means clustering as unsupervised techniques to identify hidden patterns in the data, and Support Vector Machines (SVM) and Random Forests (RF) as supervised techniques to predict credit risk. The integrated approach outperformed SVM and RF alone, and the results could be useful for financial institutions in improving their credit risk assessment process and reducing the risk of default. [3]

Shoumo et al. (2019) explored the potential of machine learning techniques in credit risk assessment for smart banking. The authors used a dataset of 1,000 borrowers with 15 input features and five machine learning techniques to predict credit risk. They found that Artificial Neural Network (ANN) and Support Vector Machine (SVM) outperformed other techniques in terms of prediction accuracy, and credit score, age, and income were the most important features. The study highlights the advantages of machine learning techniques and could be useful for financial institutions in reducing the risk of default. [4]

Pandey and Bandhu (2022) investigated the use of optimized decision tree and KNN algorithms with Bayesian optimization for credit risk assessment. They used a dataset of 1,000 borrowers with 14 input features and found that the optimized decision tree algorithm outperformed the KNN algorithm in predicting credit risk. Credit score, income, and loan amount were identified as the most important input features. The authors suggested that their model could be useful for financial institutions in managing credit risk more effectively and developing automated credit risk assessment systems. However, further research is needed to validate the results in larger datasets. [5]

Attigeri et al. (2017) evaluated the use of four machine learning algorithms, including decision tree, SVM, KNN, and ANNs, for credit risk assessment using a dataset of 1,000 borrowers. The study found that all four algorithms performed well, with SVM and KNN performing the best. Feature selection analysis revealed income, loan amount, and credit score as the most important features. The authors suggest that their findings can improve the accuracy and efficiency of credit risk assessment and can be useful in developing automated credit risk assessment systems for financial institutions. [6]

IV. Data Processing

1. Data Source:

The dataset has been collected from *Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.* The raw data contains 20 attributes comprising of 7 numerical and 13 categorical attributes with a cumulative of 1000 instances. Data file originally contain categorical or symbolic attributes. For readability of the dataset, the symbolic attributes have been replaced by their corresponding attribute descriptions.

The dataset contains the following attributes represented by feature columns:

S. No.	Attribute	Type	Variable Type
1	Status of existing checking account	Qualitative	Categorical
2	Duration in month	Numerical	Continuous
3	Credit History	Qualitative	Categorical
4	Purpose	Qualitative	Categorical
5	Credit amount	Numerical	Continuous
6	Savings account/bonds	Qualitative	Categorical

7	Present employment since	Qualitative	Categorical
8	Instalment rate in percentage of disposable income	Numerical	Continuous
9	Personal status and Sex	Qualitative	Categorical
10	Other debtors/ guarantors	Qualitative	Categorical
11	Present residence since	Numerical	Continuous
12	Property	Qualitative	Categorical
13	Age in years	Numerical	Continuous
14	Other instalment plans	Qualitative	Categorical
15	Housing	Qualitative	Categorical
16	Number of existing credits at this bank	Qualitative	Categorical
17	Job	Qualitative	Categorical
18	Number of people liable to provide maintenance for	Numerical	Continuous
19	Telephone	Qualitative	Categorical
20	Foreign Worker	Qualitative	Categorical
21	Risk	Qualitative	Categorical

Table 1 : Features

2. Data Understanding:

We visualise the dataset, with the aim of finding the following using pandas data-frame functionalities:

- Data-frame Column information : <pd_dataframe>.info()
- Feature Datatypes : <pd_dataframe>.dtypes()
- Unique values: <pd_dataframe>.nunique()
- Presence of null or NA data:
- Data-frame shape: <pd_dataframe>.shape()
- Statistical analysis of numerical features: <pd_dataframe>.describe()

Upon collecting the mentioned data, we can infer that the dataset has the following characteristics:

- No missing values.
- 7 int64 type features
- 13 object type features
- Shape of the dataset is (1000,21)

3. Exploratory Data Analysis(EDA):

To identify any anomalies, missing data, or any outliers we conduct EDA to comment on the quality of the Dataset. We also get an idea about the distribution of continuous data upon which we can make observations about statistical skewness.

From our initial understanding, we can eliminate the possibility of any missing data. We capture a heatmap of the initial impression of the data to understand the correlation among numerical features.

Figure 1: Initial Heat Map

The heat map indicates relatively higher correlation between "Credit amount" and "Duration in month". Presence of any outliers can be contributing to high correlation.

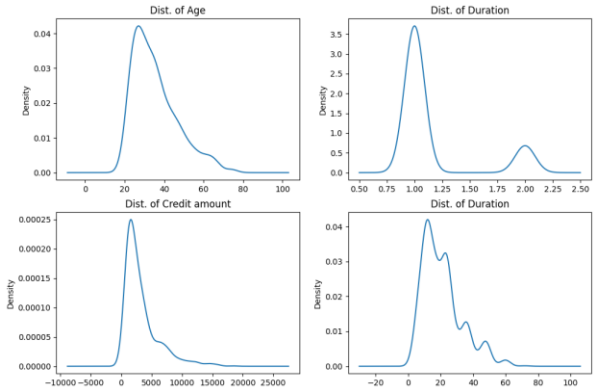

Figure 2: Density plot of Distribution of Continuous Variables

In we can observe skewness in the distribution of continuous variables "Credit amount", "Age in years", and "Duration in month" under consideration. This shall be dealt with when data will be normalised in data pre-processing phase.

To analyse numerical data, we have utilized Pandas crosstab function which will provide cross tabulation of the categorical feature with respect to target variable "Risk". This will help in visualizing presence of any outliers.

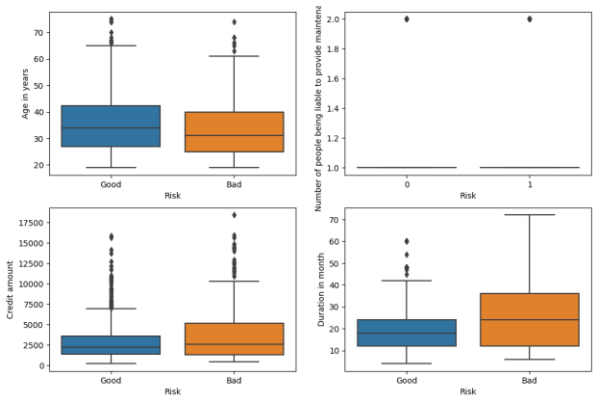

Figure 3: Box Plot of Continuous Variables

It is observable that the features "Credit amount", "Age in Years", and "Duration in month" contain significant amount of outliers. We will address these in the coming stages of the project.

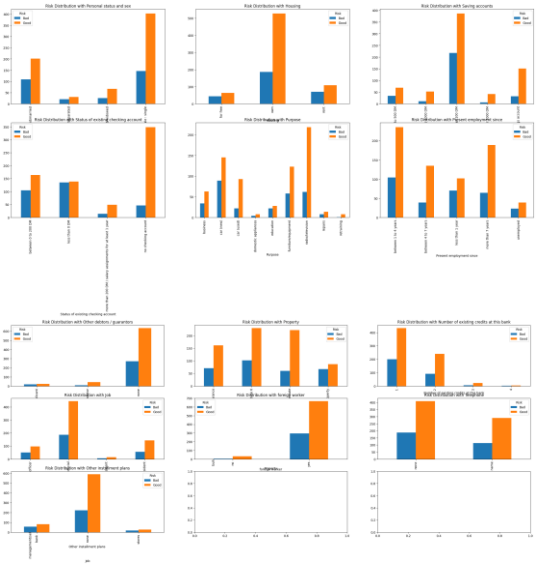

Figure 4: Cross tabulation of Categorical features with Risk

The graphs in Figure 4 depict the proportion of good and bad borrowers based on ground truth available in the dataset. The features are cross tabulated with target variable "Risk"

4. Data Pre-processing:

Before we apply pre-processing methods to dataset under consideration, we employ cleaning methods to mitigate any anomalies or missing data.

Outliers are values that are more than three standard deviations away from the mean. They are indicators of issues such as human error or poor sampling. To treat such values, we have employed the following method:

- Interquartile Range(IQR) Method : IQR is calculated using the below formula –
  - IQR = Q3 – Q1; where Q1 is First Quartile and Q3 is Third Quartile.
  - Lower Whisker = Q1 – 1.5*IQR
  - Upper Whisker = Q3 + 1.5*IQR

  Data is the rescaled between the range [Lower Whisker, Upper Whisker]

In the given dataset, feature "Credit amount" contains relatively far more outliers than other continuous variables. Therefore, while other continuous variables have been treated once for outliers, "Credit amount" has been treated with an iteration value of 4.

Outliers have not been completely irradicated based on the assumption that they indicate a mistake in data collection, whereas other times they can influence the dataset. Thus, it is important to keep them to better understand the dataset in the big picture.

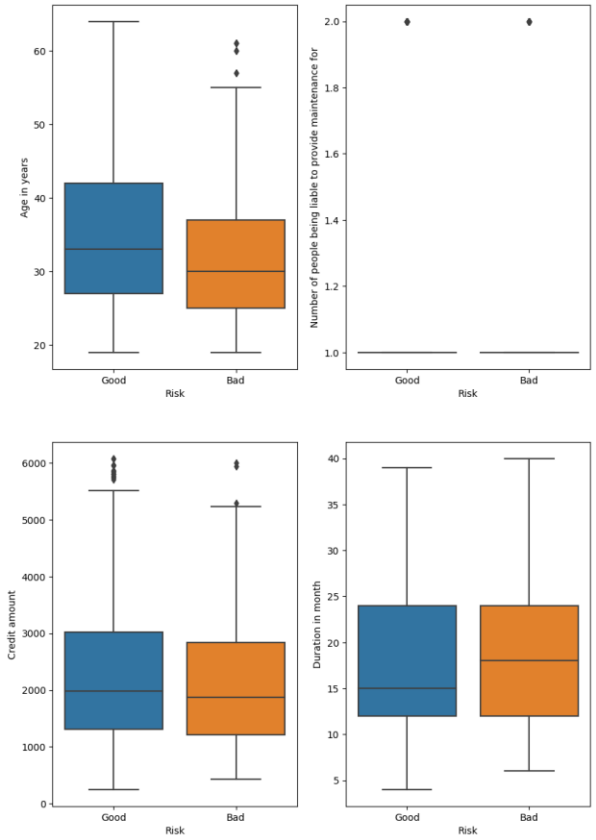

Figure 5: Boxplot of Distribution of Continuous Variables after IQ

In Figure 5, it can be observed that outliers have been significantly treated.

In next stage of Data Pre-processing, log transformation of continuous variables is conducted to reduce their skewness. Categorical values are converted to numerical values.

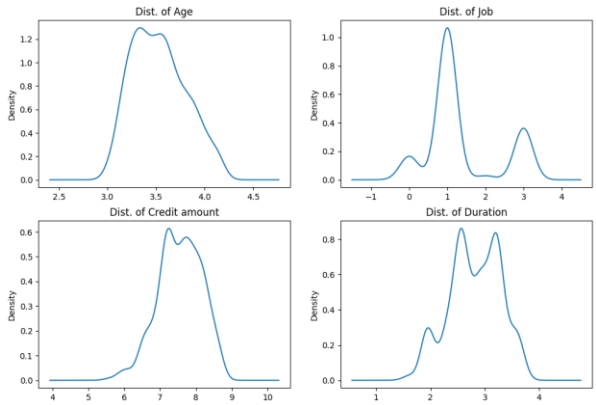

Figure 6: Distribution of Continuous Variables post Log Transformation

Post transformation all features have the same datatype.

5. Feature Engineering:

When correlation heat map is run on the encoded dataset, we get the following output:

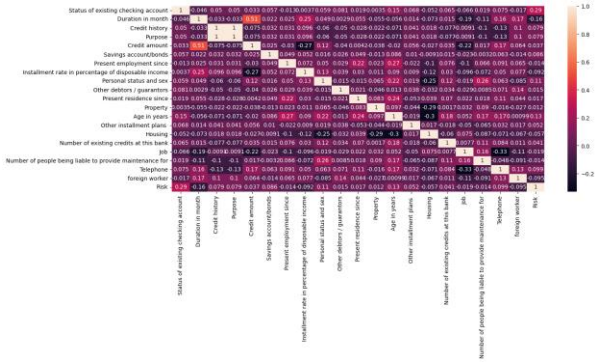

Figure 7: Correlation Heatmap after feature encoding

It can be observed that the features have significantly low correlation with target variable "Risk." For further inspection, multiple scoring methods have been adopted. They are:

- SelectKBest with chi2 test
- SelectKBest with ftest
- Extra Trees Classifier
- Mutual Information Classifier

The chi2 test and ftest measure correlation between feature variable and target variable assessing standard linear correlation and non-linear correlation respectively.

Extra Tree Classifier and Mutual Information Classifier are predictive models which produce correlation between feature variable and target variable as by-product of learning process.

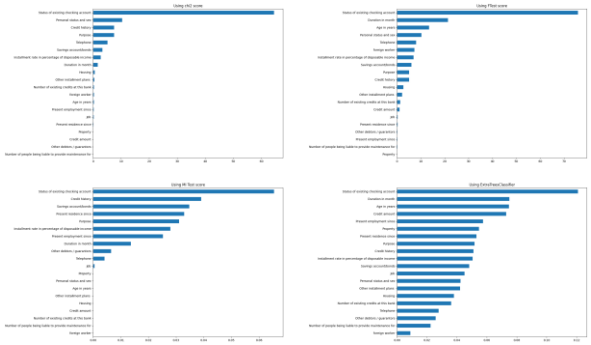

Figure 8: Feature scores via different methods

Considering the scores obtained for features from different testing methods, it can be concluded that "Status of existing checking account" gets a high score on all measures and thus is not very predictive and can be dropped. This can be corroborated by employing Feature Reduction using Principal Component Analysis (PCA).

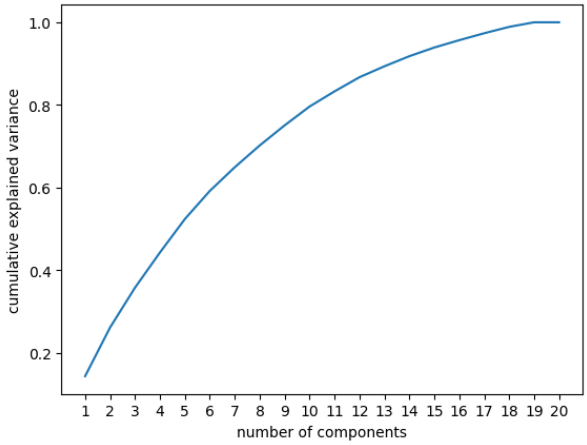

Figure 9: PCA outcome

The output graph shows plateaued progress after component 19, thus indicating a feature can be dropped.
The feature "Status of existing checking account" is dropped from training data, which will be utilized by machine learning algorithms.

V. Learning Methods

The following learning methods have been adopted in the project.

1. Logistic Regression

For problems involving binary classification with a binary output, one type of supervised learning technique is called logistic regression (e.g., 0 or 1). By analysing an instance's features, logistic regression seeks to determine the chance that it belongs to a particular class (such as 0 or 1) based on those qualities (i.e., independent variables). The relationship between the input variables and the

log-odds of the output variable is modelled by logistic regression as opposed to linear regression, which forecasts a continuous result. Regarding its interpretability, this method has advantages. In terms of how each independent variable affects the probability, the model's parameters are simple to understand. Due to this, the model can provide information about the variables that have an impact on the result variable.

In credit risk analysis, the goal is to assess the risk of default for a loan applicant based on various features such as their credit history, income, employment status, and other factors. Logistic regression can be used to model the probability of default given these features, and can provide insights into which features are most predictive of default risk

2. Support Vector Classifier

The Support Vector Classifier is a supervised learning technique used for binary classification problems. SVC's main goal is to find a hyperplane that divides the data points into their corresponding classes while minimising the distance between the hyperplane and the closest data points in each class. The Maximum Margin Hyperplane is the hyperplane that maximises the margin (MMH). If the data is not linearly separable, the approach employs a kernel function to translate it into a higher-dimensional space where it is more likely to be linearly separable.

The goal of credit risk analysis is to determine the likelihood of default for a loan application based on factors such as credit history, income, job status, and other factors. SVC can be used.

VI. Analysis, Testing, and Results

After applying the learning methods on 80% Train and 20% Test data, the following results are obtained:

	Logistic Regression	Support Vector Classifier
Training Acc.	0.764	0.753
Test Acc.	0.741	0.716
F1 score	0.837	0.835
Recall	0.923	0.991
Precision	0.766	0.720
CV acc.	0.747	0.744

Table 2: Results

Initial impression of performance metrics, Logistic Regression has higher accuracy on Training data in comparison to Support Vector Classifier. K-Fold Cross-Validation is performed with K set as 10.

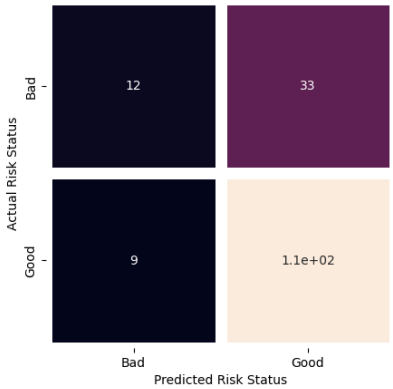

Figure 10: Logistic Regression

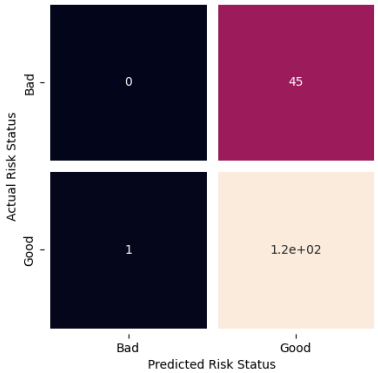

Figure 11: Support Vector Classifier

However, understanding what these metrics signify and what importance do they hold in account of Credit Risk Analysis is important.

- Recall measures the model's ability to detect borrowers who are likely to default on their loans. Recall is particularly important in credit risk analysis because it is generally more costly for lenders to miss a default than to identify a non-defaulting borrower as a default risk. If a borrower is incorrectly classified as low-risk and granted a loan, but then defaults on the loan, the lender may suffer financial losses. On the other hand, if a borrower is incorrectly classified as high-risk and denied a loan, but would have paid back the loan, the lender may miss out on potential profits. Therefore, for Credit Risk Analysis a model with high recall can help lenders identify high-risk borrowers more accurately and avoid potential losses
- Precision measures the model's ability to accurately identify borrowers who are likely to default on their loans. Precision is particularly important in credit risk analysis because it helps lenders minimize the number of false positives, i.e., borrowers who are incorrectly

identified as high-risk and denied loans. False positives can lead to missed business opportunities and may result in the lender losing potential profits. Therefore, a credit risk analysis model with high precision can help lenders make more accurate lending decisions and avoid missed business opportunities. However, high precision may come at the cost of lower recall, i.e., the proportion of actual defaults that are correctly identified by the model. Therefore, it is important to strike a balance between recall and precision to optimize the performance of the credit risk analysis model.

- F1 score provides a more comprehensive evaluation of the model's performance. The F1 score is particularly important in credit risk analysis because it helps lenders balance the trade-off between precision and recall, which are two competing objectives. A high F1 score indicates that the model has achieved a good balance between precision and recall and can accurately identify both defaulting and non-defaulting borrowers.

logistic regression and support vector classifiers have their own strengths and weaknesses for credit risk analysis. Logistic regression is simple and efficient, but may not be suitable for complex, non-linear datasets or imbalanced datasets. On the other hand, support vector classifiers can handle non-linear data and imbalanced datasets, but may require more computational resources and tuning to achieve good performance. The choice between the two algorithms ultimately depends on the specific characteristics of the dataset and the goals of the analysis.

VII. Conclusion

Based on the collected metrics, we judge the models on their F1 score, Recall and Precision rather than accuracy as the accuracy is influenced by imbalance in the dataset. Support Vector Classifier and Logistic Regression model have similar F1 score of higher order, indicating they have balanced Recall and Precision. In case of Credit Risk Analysis, a model with higher Recall score is more beneficial. Therefore, Support Vector Classifiers holds to be better than Logistic Regression for Credit Risk Analysis on dataset under consideration.

The model can be improved by incorporating additional data sources such as external credit data, macroeconomic data, or social media data can provide additional insights and improve the accuracy of the model. Other machine learning algorithms such as decision trees, neural networks, or gradient boosting can be explored as alternative models to improve the performance of the model.

There are several possible extensions and business applications of credit risk analysis using logistic regression and support vector classifiers. Few applications can be as follows:

- Portfolio optimization: Credit risk analysis can be used to optimize loan portfolios by selecting loans that maximize return while minimizing risk.
- Risk management: Credit risk analysis can be used to assess the overall risk of a portfolio and identify areas where risk can be reduced.
- Credit limit setting: Credit risk analysis can be used to set credit limits for individuals or businesses based on their creditworthiness.
- Insurance underwriting: Credit risk analysis can be used in insurance underwriting to assess the risk of insuring a particular individual or business.

VIII. References

[1] Qasem, M.H. and Nemer, L., 2018. Extreme learning machine for credit risk analysis. Journal of Intelligent Systems, 29(1), pp.640-652.

[2] Bequé, A. and Lessmann, S., 2017. Extreme learning machines for credit scoring: An empirical evaluation. Expert Systems with Applications, 86, pp.42-53.

[3] Bao, W., Lianju, N. and Yue, K., 2019. Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. Expert Systems with Applications, 128, pp.301-315.

[4] Shoumo, S.Z.H., Dhruba, M.I.M., Hossain, S., Ghani, N.H., Arif, H. and Islam, S., 2019, October. Application of machine learning in credit risk assessment: a prelude to smart banking. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 2023-2028). IEEE

[5] Pandey, P. and Bandhu, K.C., 2022. A credit risk assessment on borrowers classification using optimized decision tree and KNN with bayesian optimization. International Journal of Information Technology, 14(7), pp.3679-3689.

[6] Attigeri, G.V., Pai, M.M.M. and Pai, R.M. (2017). Credit Risk Assessment Using Machine Learning Algorithms. Advanced Science Letters, 23(4), pp.3649–3653.