**Student number**------------23025664

**Module name**--------------- MACHINE LEARNING AND PREDICTIVE ANALYTICS

**Module number**------------ UFCFMJ-15-M

**Programme**------------------Msc Financial Technology

A bank loan can be defined as when a bank offers to lend money to consumers for a certain period (*economicshelp*). As a condition for the bank loan, the borrower will need to pay a certain amount of interest per month, or per year. To be able to access loans from Banks, a borrower is required to have a good or excellent credit score, a stable income and a low debt-to-income ratio otherwise known as debt service ratio (DSR) (Millerbernd, 2023). This is to ensure borrowers can repay their loans.

Though Financial institutions sell many products to their customers, interest income on loans is the major income line for financial institutions. However, there is an inherit risk in granting loans known as credit risk. Credit risk is the possibility that borrowers will fail to meet their loan obligations as and when due, (Sheikh, Goel, and Kumar, 2020; Nalawade et al, 2022) and predicting if a borrower will default or not is a big challenge for any Financial Institution. (Uddin *et al*,2023)

In the financial industry, loan approval prediction plays a pivotal role in mitigating risk by providing a framework for assessing the creditworthiness of applicants as well as ensuring that loan approval aligns with the risk appetite of the financial institution. (Ndayisenga,2021; Uddin *et al*,2023). When Banks and other financial lenders can't accurately predict loan defaults, it causes significant problems that can affect the financial health of the bank, (Uddin *et al*,2023).

Traditionally, loan approval processes have relied heavily on manual underwriting methods, wherein loan officers carefully evaluate numerous factors such as credit history, income level, employment level, income stability, debt-to-income ratio, etc., to make lending decisions. However, such conventional approaches are not without their limitations. The manual underwriting processes are often time-consuming, labor-intensive, and susceptible to human biases and errors, leading to inefficiencies and inconsistencies in decision-making. Moreover, the rigidity of these methods may hinder the ability to adapt to evolving market dynamics and changing borrower profiles (Nalawade et al, 2022; Uddin *et al* ,2023)

In light of these challenges, the emergence of machine learning tools has revolutionized the landscape of loan approval by harnessing the power of advanced algorithms and data-driven insights. In fact, most Challenger banks now leverage on the predictive power of machine learning tools to make loan application and approval seamless and very fast, accounting for the rapid growth being enjoyed by these online banks. (Berg, Fuster and Puri 2021; Ndayisenga,2021)

Traditional Financial institutions are also beginning to combine manual loan approval processes with Machine learning tools to optimize the loan approval process, ushering in a new era of efficiency and accuracy greater than using only manual tools. Machine learning models can analyze vast amounts of data, including borrower demographics, financial records, and macroeconomic indicators, to generate predictive models that forecast the likelihood of loan approval with unprecedented precision. (Berg, Fuster and Puri 2021; Ndayisenga,2021)

Furthermore, machine learning-based loan approval holds immense potential benefits for both borrowers and lenders. For borrowers, it promises faster access to credit, streamlined application processes, remove loan officer biases, and allows for more personalized lending solutions tailored to individual financial profiles. For lenders, it offers the opportunity to enhance risk management practices, minimize default risks, and optimize loan portfolio management strategies (Berg, Fuster and Puri 2021).

In handling this task, a dataset on loan approval prediction was downloaded from Kaggle (Archit Sharma). The loan approval dataset is made of financial records and related details used to assess whether an individual qualifies for loans. It contains factors like credit scores (cibil), income, employment status, loan duration, amount, asset value, and approval status. This study uses Logistic regression and Randon Forest Machine Learning Model to find patterns in a standard dataset of approved loans and predict future loan approval (Tumuluru et al, 2022).

The dataset was split into 80% training dataset and 20% test dataset. After training the model by using the training dataset, we used the test dataset to check the accuracy of the two models. Both Models, Logistic Regression and Random Forest have similar performance. The models are compared based on performance measures such as accuracy, precision, recall and F1\_score.

**Dataset Description and Analysis:**

The dataset consists of 4269 rows and 13 columns, 12 columns are independent variables and 1 column, the loan status is dependent variable, with a mix of categorical and numeric variables. Out of a total of 4269 loans, 2656 were approved, and 1615 were rejected. This shows that more loans were approved than rejected. Below is the breakdown and analysis of the dataset:

The dataset is made up of a diverse set of variables related to loan applications and creditworthiness, including demographic information, financial details, and asset values.

The presence of categorical variables like Education and Employment status suggests that factors beyond numerical metrics were considered to influence loan approval decisions.

The numeric source variables provide insights into the financial profiles of loan applicants, such as income, loan amount, and credit score.

The distribution of loan statuses (approved vs. rejected) indicates that 2656 loan applications were approved, while 1613 loan applications were rejected.

2144 applications were from graduates while 2125 were received from non-graduates.

2150 applications were from those that are self-employed and 2119 from those not self-employed.

The numerical variables are not normally distributed.

42% of loan applications came from those with poor credit score rating.

A table with numbers and letters

Description automatically generated with medium confidence

A graph of different colored bars

Description automatically generated with medium confidence

A pie chart with numbers and text

Description automatically generated with medium confidence

A blue and red squares

Description automatically generated

**Bivariate and Multivariate Analysis**

Bivariate analysis shares valuable insights into the association between two variables in a dataset but does not account for the potential influence of other variables. It shows how changes in one variable relate to changes in another variable. Multivariate analysis on the other hand examines the relationships between multiple variables simultaneously (Sandilands,2014).

This task requires multivariate analysis because the models will have to compare more than two variables in the dataset to make a prediction. The Algorithms consider not only income but also factors such as education level, employment status, credit store and number of dependents of applicants to predict loan approval. Multivariate analysis enables a more comprehensive understanding of complex relationships among multiple variables and can reveal hidden patterns or dependencies that may not be apparent in bivariate analysis alone (Shossain, 2019).

In plotting the bivariate analysis, datasets were grouped by variables and loan status, and the number of occurrences were counted, While the visualization of the multivariate analysis share insight into the pairwise relationships between the variables in the dataset, as well as how these relationships vary with the loan\_status categories.

Please see below A sample of the Bivariate and Multivariate analyses of the dataset;

A graph of a graph with blue and orange squares

Description automatically generated

A graph of different colored shapes

Description automatically generated with medium confidence

**Processing/Cleaning**

Data preprocessing is a very important aspect of data analysis and modeling tasks for ensuring data quality, consistency, and suitability. It is crucial for effective data-driven decision-making and enables the extraction of actionable insights from the data.

Cibil score was categorized as Cibil status, white spaces on either side of the loan status column were also removed. Removing whitespaces is a common data cleaning step that improves data consistency, integrity, and usability for subsequent analysis. The value of the target variable, which is the loan status, was mapped to convert to numerical value, 1 for approved and 0 for rejected. The dataset is then split into training and test datasets, ensuring that the data is randomized and stratified, which is essential for training and evaluating machine learning models. Randomized sampling ensures data is selected randomly, while stratified sampling ensures that the resulting dataset maintains the same distribution of classes or categories as the original dataset.

The categorical and numeric variables were then selected for preprocessing. Kindly see below the size and the variables of the processed training and test dataset;

**Size of Training dataset**

|  |  |  |
| --- | --- | --- |
|  | Row | Column |
| X\_train | 3415 | 11 |
| y\_train | 3415 |  |

**Size of Test dataset**

|  |  |  |
| --- | --- | --- |
|  | Row | Column |
| X\_test | 854 | 11 |
| y\_test | 854 |  |

A screenshot of a computer

Description automatically generated

**Choice of Learning Algorithm:**

Logistic regression and random forests are chosen for this task as they are supervised machine learning model ideal for developing a loan approval classification model due to their effectiveness in handling binary outcomes, interpreting feature importance, and managing complex interactions within the data (Chugh,2023).

**Logistic Regression**

Logistic regression is a statistical method used for supervised classification problems like loan approval prediction, where the target variable has two possible outcomes (e.g., yes/no, 1/0, true/false, approved or rejected). Despite its name, logistic regression is a classification algorithm rather than a regression algorithm. Logistic regression works by predicting the class of a given input e.g. whether a loan is approved or rejected. Once the model has been trained, it uses a decision boundary to make new predictions or classification either to positive class or negative class.

Logistic regression is widely used in several sectors such as finance, healthcare, marketing, and social sciences, due to its simplicity, interpretability, and effectiveness for binary classification tasks (Edgar and Manz, 2017).

**The Random Forest Model**

The Random Forest Algorithm is a popular supervised machine learning model used for both classification and regression tasks. Numerous decision trees are developed by using diverse random subsets of both data and features. Each decision tree acts as an individual expert, offering its perspective on how data should be classified. Predictions are made by evaluating the prediction of each decision tree and then selecting the most common outcome. (Sruthi,2024; Sharfi 2023).

Random Forests are one of the most popular and widely used Machine learning Algorithms because of their versatility, ease of use and robustness. They are very effective in predictive modeling and feature selection (Sruthi,2024).

**Analytical Evaluation of Logistic Regression and Random Forest Models**

Logistic regression tends to have lower computational costs for both training and deployment compared to random forests. Random forests training time is higher as they involve creating multiple decision trees and combining their predictions. However, after deployment, they can efficiently handle prediction tasks.

|  |  |
| --- | --- |
| **Model** | **Training Time** |
| Logistic Regression | 0.05s |
| RandomForestClassifier | 0.43s |

**Evaluation Metrics**

The two models were evaluated on performance metrics such as accuracy, precision, recall and F1 score. These metrics are usually used to evaluate the performance of classification model. (Sharfi,2023)

|  |  |
| --- | --- |
| Performance Metrics | Description |
| Accuracy | measures the proportion of correctly classified instances out of all instances. |
| Precision | measures the proportion of true positives (correctly classified positive instances) out of all instances classified as positive |
| Recall | measures the proportion of true positives out of all actual positive instances |
| F1 score | measures the harmonic mean of precision and recall |

Accuracy = *𝑇𝑟𝑢𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒𝑠+True Negatives*

*False Positives +False Negatives+𝑇𝑟𝑢𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒𝑠+True Negatives*

Precision = *𝑇𝑟𝑢𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒𝑠*

*𝑇𝑟𝑢𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒𝑠 + 𝐹𝑎𝑙𝑠𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒𝑠*

Recall = *𝑇𝑟𝑢𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒𝑠*

*𝑇𝑟𝑢𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒𝑠 + 𝐹𝑎𝑙𝑠𝑒 Negatives*

F1 = 2\* Precision*\*Recall*

*Precision +Recall*

**Performance Metrics for Logistic regression**

A graph of red and blue bars

Description automatically generated

A graph showing a number of numbers and a number of numbers

Description automatically generated with medium confidence

A diagram of a matrix

Description automatically generated with medium confidence

A graph with numbers and lines

Description automatically generated

A blue and white squares

Description automatically generated with medium confidence

**Performance Metrics for Random Forest Classifier**

A graph of red and blue bars

Description automatically generated

A graph showing a number of numbers and a number of numbers

Description automatically generated with medium confidence

A diagram of a graph

Description automatically generated with medium confidence

A graph with numbers and lines

Description automatically generated with medium confidence

**Feature Importance-Logistic Regression**

A graph with different colored lines

Description automatically generated

**Feature Importance-Random Forest**

A graph with different colored lines

Description automatically generated

Analysis of the features in the dataset shows that for logistic regression, Poor credit status, Income, loan amount and loan term have highest impact on approval decision s, While the top four features with high impact for Random Forest Classifier are the categories of credit status.

K-Means clustering is an unsupervised learning algorithm. It does not take into account the labels during the training process. Therefore, data are clustered based on the features, not on the binary outcomes. This can sometimes lead to clusters that do not correspond well with the true binary labels. Since K-Means does not use the label information, the performance in terms of accuracy for binary classification might not be accurate. For this task, the datasets are grouped into two clusters (0 and 1), which have a similar distribution of loan statuses. In Cluster 0, approximately 61.87% of the loans are approved, while in Cluster 1, approximately 62.55% of the loans are approved. The proportions of approved and rejected loans are almost same between the two clusters, showing that the k-means clustering did not find distinct groups with significantly different loan approval rates, and which confirms that K-Means Clustering is not a good model for classification task (Arikuncoro, 2020).

**Conclusion**

The results of my analysis show that both Models, The Logistic Regression and Random Forest Classifier performed excellently in predicting loan approval, as they both have the same range of accuracy on the training datasets as well as the test datasets. I would recommend financial institutions to use machine learning techniques in their loan approval process as the integration of machine learning into the loan approval process represents a transformative step towards achieving greater efficiency, transparency, and fairness in the financial services industry (Berg, Fuster and Puri,2021; Uddin et al, 2023). This Enhances decision-making in loan processing, potentially reducing defaults, improving customer satisfaction, reducing loss and increasing revenue for Lenders. Also, the results show that credit status is one of the most important features that influenced decisions as more customers with poor credit status had their loans rejected.

**Word count**------------------2,107

**References**

1. Bank Loans. Economics Help [online]. [Accessed 10 May 2024]
2. Millerbernd, A. (2023) What Is a Bank Loan? How Do They Work?. Nerdwallet [online]. [Accessed 28 April 2024].
3. Sheikh, M.A., Goel, A.K. and Kumar, T. (2020) An Approach For Prediction of Loan Approval Using Machine Learning Algorithm. Ieee Explore [online]. [Accessed 08 April 2024].
4. Nalawade, S., Andhe, S., Parab, S. and Prof Sankhe, A. (2022) Loan Approval Prediction. International Research Journal of Engineering and Technology (Irjet) [online]. 9 [Accessed 05 May 2024].
5. Uddin, N., Ahamed, K.U., Uddin, A., Islam, M., Talukder, A. and Aryal, S. (2023) An Ensemble Machine Learning Based Bank Loan Approval Predictions System with a Smart Application. International Journal of Cognitive Computing in Engineering [online]. 4, pp. 327-339. [Accessed 20 April 2024].
6. Ndayisenga, T. (2021) Bank Loan Approval Prediction Using Machine Learning Techniques. University of Rwanda Digital Repository [online]. [Accessed 05 May 2024].
7. Kadam, A.S., Nikam, S.R., Aher, A.A., Shelke, G.V. and Chandgude, A.S. (2021) Prediction For Loan Approval Using Machine Learning Algorithm. International Research Journal of Engineering and Technology (Irjet) [online]. 8 (4) [Accessed 13 May 2024].
8. Berg, T., Fuster, A. and Puri, M. (2021) Fintech Lending. National Bureau of Economic Research [online]. [Accessed 03 May 2024].
9. Tumuluru, P., Burra, L.R., Loukya, M., Bhavana, S. and Csaibaba, H.M.H. (2022) Comparative Analysis of Customer Loan Approval Prediction Using Machine Learning Algorithms. Ieee Explore [online]. [Accessed 05 May 2024].
10. Sandilands, D. (2014) Bivariate Analysis. Encyclopedia of Quality of Life and Well-being Research [online]., p. 416–418. [Accessed 08 May 2024].
11. Shossain, P.B. (2019) What Is Univariate, Bivariate and Multivariate Analysis?. Hotcubator [online]. [Accessed 10 May 2024].
12. Edgar, T.W. and Manz, D.O. (2017) Logistic Regression. Research Methods For Cyber Security [online]. [Accessed 13 May 2024].
13. Sruthi, E.R. (2024) Understand Random Forest Algorithms with Examples (Updated 2024). Analytics Vidhya [online]. [Accessed 10 May 2024].
14. Shafi, A. (2023) Random Forest Classification with Scikit-learn. Data Camp [online]. [Accessed 12 May 2024].
15. Soni, A. (2020) Advantages and Disadvantages of Knn. Medium [online]. [Accessed 05 May 2024].
16. Harrison, O. (2019) Machine Learning Basics with the K-nearest Neighbors Algorithm. Towards Data Science [online]. [Accessed 05 May 2024].
17. Chugh, V. (2023) Logistic Regression in R Tutorial. Datacamp [online]. [Accessed 10 May 2024].
18. Arikuncoro, J. (2020) K-means and Hierarchical Clustering in Python. Catatan Si Data Scientist [online]. [Accessed 17 May 2024].