LAUTECH Journal of Computing and Informatics (LAUJCI) – ISSN: 2714-4194
Volume 3 Issue 1, June 2023 – www.laujci.lautech.edu.ng

ASSESSING THE CREDITWORTHINESS STATUS OF MOBILE PHONE USERS USING SUPPORT VECTOR MACHINE

Rufai M. M
Department of Computer Science
Yaba College of Technology
Yaba, Nigeria
mohammed.rufai@yabatech.edu.ng

Ajala M. T.
Department of Industrial Maintenance Engineering
Yaba College of Technology
Yaba, Nigeria
mosud.ajala@yabatech.edu.ng

Lawal L. O.
Department of Computer Science
Yaba College of Technology
Yaba, Nigeria
olawal201315@gmail.com

Alao W. A
Department of Industrial Maintenance Engineering
Yaba College of Technology
Yaba, Nigeria
Olalekan.awoniran@bowen.edu.ng

ABSTRACT

Credit risk is a major concern to lenders, it is important for any lending company to be able to determine when to approve and when to decline a loan. Machine learning techniques have recently been adopted to help identify defaulting customers, and also help to speed up the decision-making process of approving a loan. In this study, relevant features that are related to customers' credit scoring are selected, and we made use of a support vector machine to build a model that could solve the underlying problems. From the test result, our developed model could predict a borrower's compliance status to loan payment. The model was able to attain a performance measure for Precision, recall, accuracy, and f1-score on test data with values of 97.2 %, 100.0 %, 99.1 %, and 98.6 % respectively. This indicates that the Support vector machine is an effective approach that could be used in credit scoring, and our developed model can be classified as a good classification.

Keywords: Credit Scoring, Machine Learning, Supervised Machine Learning, Support Vector Machine

1. INTRODUCTION

Credit scoring is employed in order to evaluate a loan applicant's creditworthiness, it is used by a lender to decide if a loan should be approved or declined. It helps lenders to avoid financial risk and bad debts. There are a set of procedures to follow to figure out the creditworthiness of a borrower, some of the procedures are paid to confirm the identities provided by the loan applicant, their consent, and their competence to pay up after being approved for a loan. To determine the loan applicant's capacity to repay, we consider their credit score, current income, and outstanding debts (Finscore, 2022). In telecommunication industries, e.g., MTN credit scoring help to determine the customers that meet up the requirement for borrowing a loan, either an airtime loan or a mobile data loan, it is the first step of the loaning process before deciding on how much should be given to the borrower and how much interest rate would be derived from the loan by using the customer's credit report history and their behavioral activities.

Credit scoring using a supervised learning approach helps to develop a risk scoring model that uses customer data to predict the creditworthiness of loan applicants, using this method helps to automate the decision-making process of who gets a loan and who does not. It also makes processing faster than manually going through each customer’s data one after the other to decide on their creditworthiness, which helps avoid human errors in making decisions about customer eligibility when working with huge data.

The objective of this study is to develop a machine learning model using an SVM (support vector machine) algorithm that can be used to predict the creditworthiness of a telecommunication loan applicant. This research work is posed to

addressed the following problems in Mobile Network services, to solve the problem of loan default, bad debts, and financial loss. It also helps to speed up the decision-making process of giving out loans. MTN "Africa's largest mobile network operator" is the scope of study of this project. Our focus is to determine if a customer is eligible for their XtraTime and Xtrabyte service. However, with little or no modification, the same study can be applied to other mobile networks in the area of credit scoring.

2. RELATED TERMS

Credit risk is the possibility that a lender won't get the principal and interest that is owed. By examining a borrower's creditworthiness variables, such as their present debt burden and income, lenders can reduce credit risk. The "5 Cs" of credit risk—credit history, repayment ability, capital, loan terms, and collateral—are used to assess creditworthiness.

Creditworthiness is what creditors take into account before approving any new credit, and it is determined by several factors, including your repayment history and credit score2.

Credit Scoring is used to assess a person's or a small, independently owned business' creditworthiness. It is used by lenders and financial institutions to statistically analyze the possibility of a user paying back a loan and it is used to decide whether or not to grant credit.

3. RELATED WORKS

A study on credit risk and loan decision-making in financial institutions was done by Ziemba et al. in 2021. To develop decision models for credit evaluation, they investigated several data science methodologies. The results showed that the correlation-based feature selection method and random forest classifier performed well in identifying credit risk and approving or rejecting loans. However, a weakness was found in the absence of a particular approach adapted to the needs of stakeholders. Future studies should focus on creating tailored strategies that improve credit decision-making and fit with stakeholders' specific needs in order to overcome this.

Seo (2020) conducted his study on the financial institution's credit risk analysis and what enables it to only give credit to consumers with strong credit. Seo stated that various studies have examined and drawn findings regarding the ability to classify good credit and bad credit using machine learning techniques. The German dataset, which is frequently used in credit risk analysis, was used in the Seo study to compare the algorithms of earlier papers. Seo compared Logistic Regression, Decision Tree, Naive Bayes, support vector machines, linear discriminant analysis, k-nearest neighbor, and ensemble methods like boosting; bagging, and random forest. Seo's result found that logistic regression, Bagging, and support vector machine came out good, but the algorithm that was chosen as the most accurate credit risk analysis algorithm was SVM.

Niu et al. (2019) tested the reliability of social network information in predicting a loan applicant's ability to default using LightGBM, random forest, and, AdaBoost techniques. The research was limited by the lack of further social network data, which may have included the volume of calls - incoming or outgoing - and the strength of social network links. They were able to derive a result that shows the impact of a loan applicant's social network information on loan default, which can be used to improve the prediction of default on a loan. Making the appropriate feature selection options can be challenging when creating a credit scoring model, and the size of the data may be insufficient compared to the number of features needed to construct the models, according to Laborda and Ryoo (2021). The authors went on to say that the model's performance might suffer if the data contained characteristics that were not significantly related to credit risk. To have a model with good accuracy, the feature selections from the dataset, the classification, and the method to be used -either a hybrid or a single approach - must all be carefully taken into account during the model's training.

According to Natasha et al. (2019), reviewed the crucial role that classification plays in consumer risk. They opined that employing the incorrect classification technique could result in concerns with loss, loss of money, and bad credit. They

suggested DNN (deep neural network), which they claimed was the best way to categorize client loans, for evaluating the dataset, with the value of AUC as 0.638 and the number of neurons h1=10 and h2=3. Weng and Huang (2021) conducted research on the performance of using hybrid approaches to build a model over the single approach of machine learning algorithm techniques. In their study, the authors proposed a new approach that integrates instance selection, Decision tree technique, and feature selection for predicting a loan applicant's credit approval. Their findings revealed that using the hybrid method is superior to the other four machine learning approaches.

Prastyo et al. (2021) examined that presently no accurate model to determine creditors that are eligible for loans, credit score models depend on the type of loans complemented by several credit factors. the authors used Information Gain, the Weka application, and the confusion matrix. They were able to obtain performance measures with an average accuracy of 86.29%, recall of 86.29%, f-measure of 86.30%, precision of 86.33%, and 91.52% of the ROC area for their proposed method. The study was unable to investigate alternative feature selection methods and machine learning algorithms. Simumba et al. (2021) argued that because financial institutions lack access to financial data due to their geographic distance, credit evaluation of financially disadvantaged individuals hinders data collection. They looked into the likelihood of getting over the obstacle that prevents people who are financially excluded from getting credit by remotely collecting alternative data. The authors went on to say that earlier studies looked into alternative data sources, such as mobile phones, and that the research suggested combining geospatial data from the public domain, mobile phones, and satellites to improve credit evaluations in cases where financial data are lacking. Their findings showed that adding more data sources, like statistical data and other relevant satellite data types, improved performance as indicated by F1 and accuracy AUC values.

Shema (2019) conducted a study to demonstrate that just airtime recharge data can be used to train accurate credit scoring models, Shema asserts that using this procedure makes the borrower privacy less invasive than the usual models employed by lenders. Shema explained that they had a partnership in Africa with a lender that is into airtime lending, which gave them the opportunity to do a comparison of their airtime-only model side by side with models that incorporated past loan data, also with the model that the lender uses. Several tests show that their model using limited data performed no less than the alternative models.

4. METHODOLOGY

The methodology adopted in this research is an 8-step approach. it consists of the following steps as shown in Figure 1.

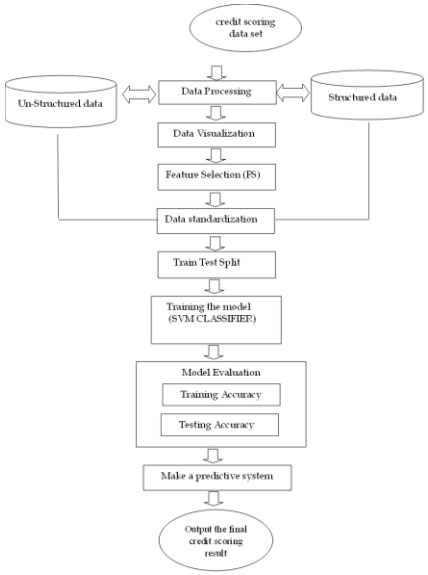


Figure 1: Block diagram of the proposed credit scoring model

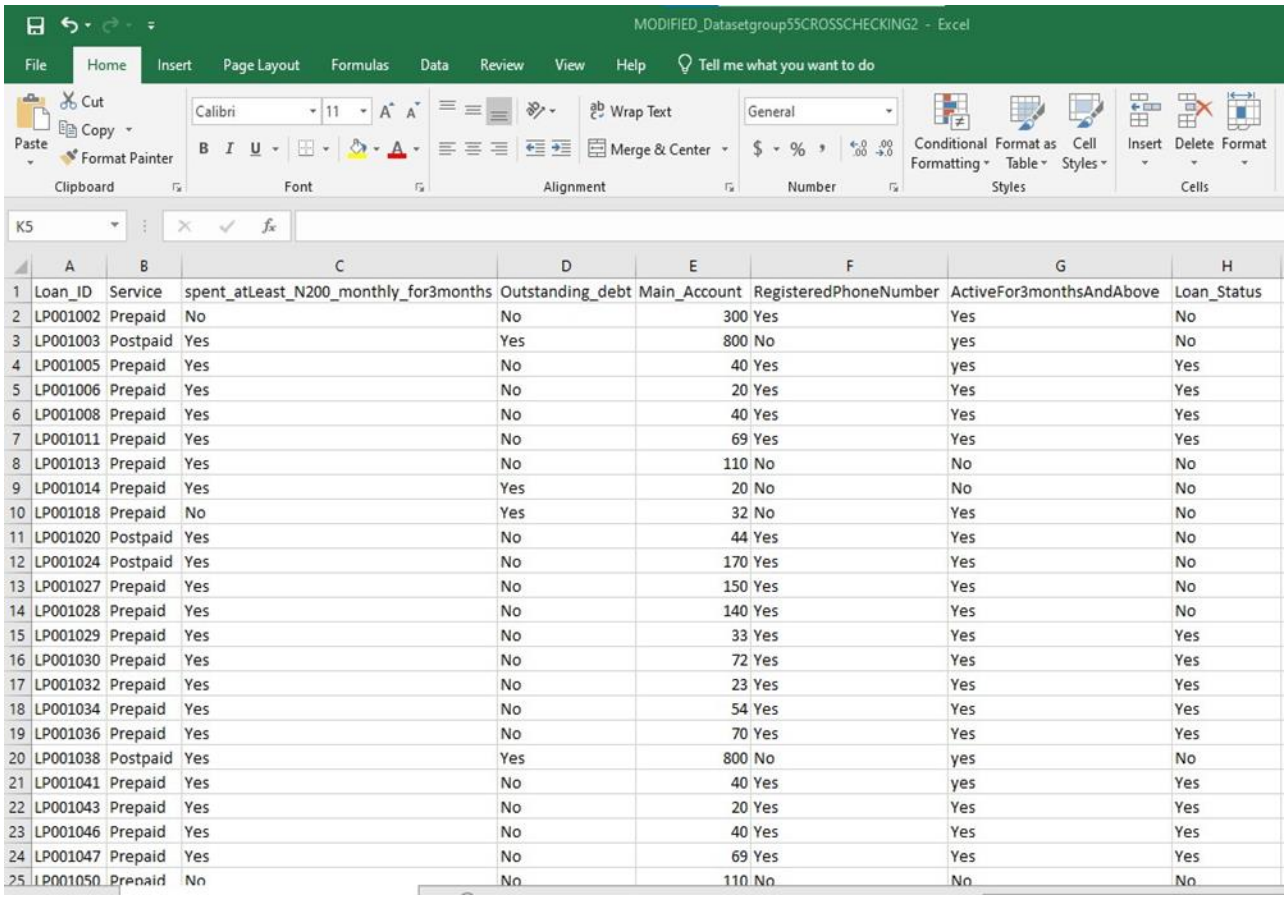
4.1 Description of the System

MTN is a network provider, it is one of the biggest telecommunication companies in Africa, they offer airtime and mobile data services and also provide loan services to their customers Their loan service is only applicable to customers who meet the conditions provided by the company. One of the loan services MTN offers is airtime lending which is called Xtratime, and a mobile data loan service called Xtrabyte. MTN Xtratime, allows their customer to borrow airtime on credit when they run out of airtime and pay back on their next recharge, while Xtrabyte allows their customers to borrow data when they run out of data and pay back on their next recharge.

For a customer to be eligible for MTN loan services he/she must meet the conditions of the company (MTN), which say "for a customer to be eligible for a loan, he/she must be on a prepaid plan, must have a registered phone number, must have been on the network for nothing less than 3 months, has spent a minimum amount of ₦200 monthly, the main account balance must be within the range of ₦0 to ₦75, and must have paid up previous outstanding debt if available".

4.2 Data Collection

The data collection process involves establishing contact with MTN. During a conversation with MTN, a dataset to work with on this research was requested but was only provided with a list of conditions used in determining the eligibility of a customer for a Credit loan. The combination of these conditions that make the customer eligible is not stated in the feedback. Working with the conditions, a questionnaire was designed to capture the required Data from MTN users. There were five hundred and sixty-five (565) respondents and their responses were used as the dataset. Each respondent had the experience of having obtained a loan from MTN at one point or the order and they were questioned on which of the conditions they satisfy that give them access to the sought loan. Their responses constitute the dataset with the conditions as features that were used for this research work. The dataset consists of 565 rows and 8 columns which was used to carry out the analysis, and the SVM model was formulated based on MTN loan conditions. The image in Figure 2 displays the appearance of the data set.



	A	B	C	D	E	F	G	H
1	Loan_ID	Service	spent_atLeast_N200_monthly_for3months	Outstanding_debt	Main_Account	RegisteredPhoneNumber	ActiveFor3monthsAndAbove	Loan_Status
2	LP001002	Prepaid	No	No	300	Yes	Yes	No
3	LP001003	Postpaid	Yes	Yes	800	No	yes	No
4	LP001005	Prepaid	Yes	No	40	Yes	yes	Yes
5	LP001006	Prepaid	Yes	No	20	Yes	Yes	Yes
6	LP001008	Prepaid	Yes	No	40	Yes	Yes	Yes
7	LP001011	Prepaid	Yes	No	69	Yes	Yes	Yes
8	LP001013	Prepaid	Yes	No	110	No	No	No
9	LP001014	Prepaid	Yes	Yes	20	No	No	No
10	LP001018	Prepaid	No	Yes	32	No	Yes	No
11	LP001020	Postpaid	Yes	No	44	Yes	Yes	No
12	LP001024	Postpaid	Yes	No	170	Yes	Yes	No
13	LP001027	Prepaid	Yes	No	150	Yes	Yes	No
14	LP001028	Prepaid	Yes	No	140	Yes	Yes	No
15	LP001029	Prepaid	Yes	No	33	Yes	Yes	Yes
16	LP001030	Prepaid	Yes	No	72	Yes	Yes	Yes
17	LP001032	Prepaid	Yes	No	23	Yes	Yes	Yes
18	LP001034	Prepaid	Yes	No	54	Yes	Yes	Yes
19	LP001036	Prepaid	Yes	No	70	Yes	Yes	Yes
20	LP001038	Postpaid	Yes	Yes	800	No	yes	No
21	LP001041	Prepaid	Yes	No	40	Yes	yes	Yes
22	LP001043	Prepaid	Yes	No	20	Yes	Yes	Yes
23	LP001046	Prepaid	Yes	No	40	Yes	Yes	Yes
24	LP001047	Prepaid	Yes	No	69	Yes	Yes	Yes
25	LP001050	Prepaid	No	No	110	No	No	No

Figure 2: A representation of the first 25 rows out of the 565 rows of the dataset in ".csv format"

In the diagram (Figure 2), customers who have "Yes" as a value on the "Loan_status" column are eligible for a loan, they meet up the conditions for qualifying for a loan, while the customers that have a "No" in the Loan-Status do not meet up the conditions of qualifying for a loan, it is either they failed to meet one of the conditions or more than one of the conditions that determine their eligibility.

4.3 Preprocessing the data

The preprocessing process was done with pandas and scikit-learn, which are Python libraries. Pandas were used to produce descriptive statistics of the dataset, which produced a result of the mean, max, std (standard deviation), and percentiles (see Figure 3). Then we checked for missing values that may be present in the dataset that can affect the result of the model by producing noise, we did label encoding by replacing the categorical values with numerical values, then we visualized the Outstanding debt column against the *Loan_Status* column (see Figure 4), and *ActiveFor3MonthsAbove* column against *Loan_Status* column in the dataset, in other to see the relationship between the features by using a seaborn count plot to visualize it. Also, data labeling was done on the data to split the dataset into two variables "features and target", thereafter we made use of the sklearn preprocessing package (scikit-learn StandardScaler) to standardize the features in the dataset, and then applied the fit and transform method on it.

	Main_Account
count	565.000000
mean	167.945133
std	567.695755
min	0.000000
25%	38.000000
50%	71.000000
75%	140.000000
max	7900.000000

Figure 3: Pandas descriptive statistics of the dataset

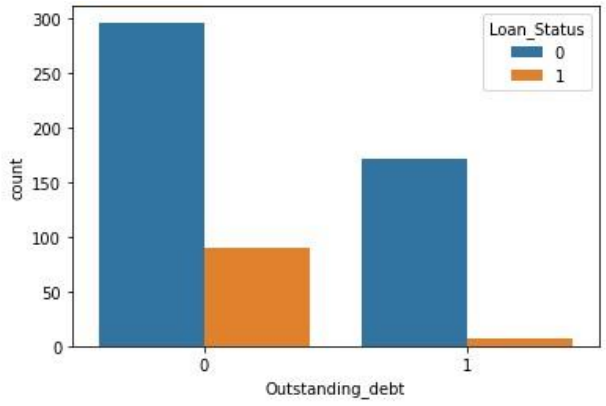


Figure 4: Count Plot Diagram

Figure 5 shows the relationship between the "Loan_Status" column and the "ActiveFor3monthsAbove" column in the dataset. It displays the customers that are eligible for the loan while being active for three months and above, and those not eligible for the loan but were also active for three months and above. For better understanding "0" means no and "1" means yes.

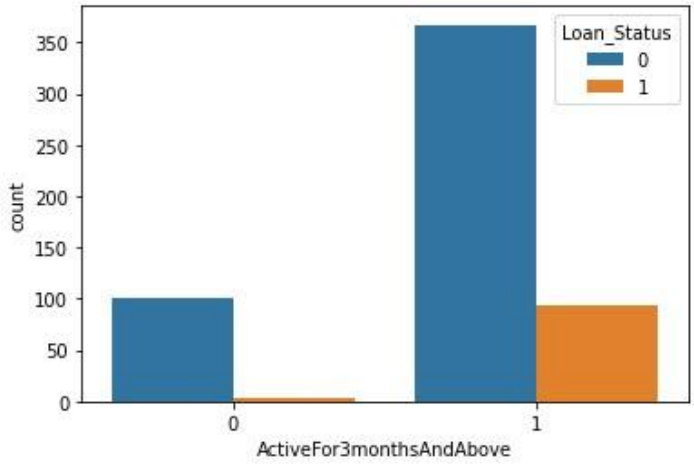


Figure 5: Count Plot Diagram

The count plot diagram above (Figure 4), shows the relationship between the "Loan_Status" column and the "Outstanding debt" column in the dataset. It displays the customers who do not have outstanding debt and was eligible for the loan, and the customers who had outstanding debt and never got a loan.

4.4 Training and Testing

The training process was done using scikit-learn train_test_split and scikit-learn Support Vector Classification, train_test_split was applied to the feature and target datapoint to split the arrays into random subsets for training and testing the data. The Overall dataset in train_test_split is measured as 1, so to split it into train and test, 20% of the dataset was assigned for testing while 80% was used to train the model. Thereafter SVM classifier was applied to train the model to be able to understand the dataset patterns, and then the Fit method was applied to fit the SVM model according to the given trained data.

4.5 Evaluation

Five model performance measures were used to evaluate the credit scoring model, the performance measures used are confusion matrix, precision, accuracy, recall, and F1_Score.

a) Confusion matrix: A confusion matrix is a matrix used to assess a classification model's performance. It displays and summarizes a classification algorithm's performance. In the confusion Matrix, TP + TN= correct predictions while FP + FN =wrong predictions.

The Confusion Matrix presented in Figure 6 has various components which include:

i. True Positives (TP): They are values that both turned out to be positive and were expected to be positive.

ii. False Positives or FPs: These are values that were actually predicted to be negative but turned out to be positive. known also as a Type I error.

iii. False Negatives or FNs: These are values that were predicted as negative but were actually positive. An error of the same type as Type II.

iv. True Negative or TN: These values are those that were both actually negative and predicted to be negative.

v. Additional evaluation metrics related to it are, accuracy, precision, recall, and F1_Score

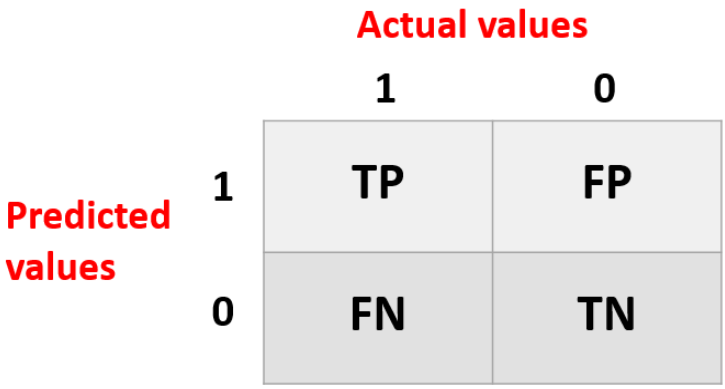


Figure 6: Confusion Matrix Diagram (depending on the built model)

b) Accuracy score: The accuracy score is the proportion of correct predictions to all the input data points. It is computed by taking the total number of correct predictions and dividing it by all of the predictions. The mathematical computation is as follows: (TP + TN) / overall datapoints

c) Precision: The precision measures how many of the total positively predicted outcomes are actually positive. The mathematical expression is given as:

$$\text{Precision} = TP/(TP + FP) \tag{1}$$

d) Recall: The recall is the percentage of correctly predicted positive outcomes out of the total number of positive outcomes. The mathematical expression is

$$\text{Recall} = TP/(TP + FN) \tag{2}$$

e) F1_Score: The F1 score is also referred to as the F Measure. The F1 score represents the balance of precision and recall. It has a mathematical formula is:

$$\text{F1\_Score} = (\text{Precision} \times \text{Recall})/ (\text{Precision} + \text{Recall}) \tag{3}$$

5. RESULT

Our proposed SVM model was able to have a confusion matrix performance measure to be True Positive (TP) = 93, False positive (FP) = 4, False negative (FN) = 0, and True negative (TN) = 16, as shown in Figure 7.

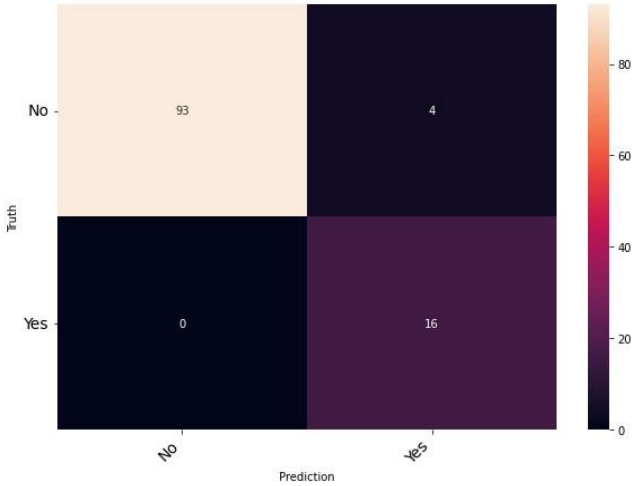


Figure 7: Visualization of the Confusion Matrix

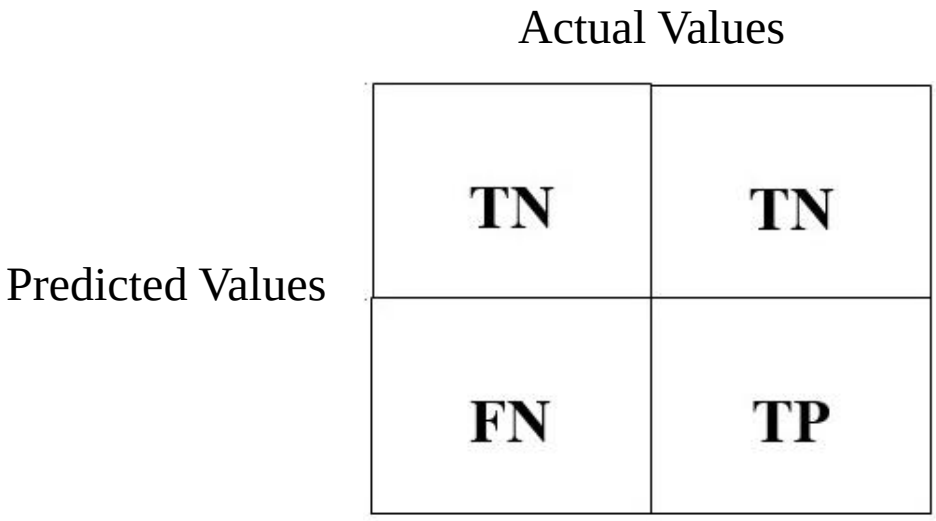


Figure 8: Arrangement of the Confusion Matrix of the Model Performance

The confusion matrix in Figure 8 shows the performance of the SVM algorithm in predicting eligibility and non-eligibility for MTN credit loan seekers. It shows that:

i. 93% of the dataset were truly predicted as true positive, i.e., eligible for loan credit

ii. 16% were truly predicted as Truly Negative, i.e., non-eligible.

iii. 4% were predicted wrongly as false meaning they were to be eligible

iv. 0% were falsely predicted as Negative i.e., no false prediction on non-eligibility

The representation of our confusion matrix in Figure 7 can be used to calculate the precision, recall, F1_Score, and accuracy of our model. For precision performance measures, the test result was 80.0 %, the accuracy score was 96%, and for F1_Score performance measures were 89 %. For recall performance measures, the result gave us 1, then after converting to percentage, it gave us 100.0 %.

6. CONCLUSION

Overall, this study shows how machine learning, in particular the SVM method, can be used to create credit scoring models for the telecommunications sector. Companies like MTN can effectively evaluate consumer creditworthiness, reduce financial risks, and provide quick loan services by automating the loan decision-making process. However, additional investigation is required to confirm and improve the model, taking various feature selection strategies and machine learning algorithms into account.

Not just MTN, but also other mobile network operators and financial institutions looking to improve their loan approval procedures, can benefit from the knowledge obtained from this study. These institutions can increase the effectiveness of their decision-making, improve consumer experiences, and lower the likelihood of defaults and financial losses by utilizing machine learning approaches.

7. REFERENCES

[1] AMIT, J. (2021, November 11). Essentials of Machine Learning. There are 3 types of Machine Learning… | by AMIT JAIN | Medium. https://medium.com/@amitjain2110/essentials-of-machine-learning-3d2fb04d56c2

[2] Finscore. (2022, August 22). Credit Scoring Company in the Philippines | FinScore. https://www.finscore.ph/

[3] IBM Cloud Education. (2020, July 15). What is Machine Learning? | IBM. https://www.ibm.com/cloud/learn/machine-learning

[4] J.S. Bridle, "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition," Neurocomputing—Algorithms, Architectures and Applications, F. Fogelman-Soulie and J. Herault, eds., NATO ASI Series F68, Berlin: Springer-Verlag, pp. 227-236, 1989. (Book style with paper title and editor)

[5] Knutson, M. L. (2019). CREDIT SCORING APPROACHES GUIDELINES. https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/original/CREDITSCORINGAPPROACHESGUIDELINESFINALWEB.pdf

[6] Laborda, J., & Ryoo, S. (2021). Feature selection in a credit scoring model. Mathematics, 9(7). https://doi.org/10.3390/math9070746

[7] Natasha, A., Prastyo, D. D., & Suhartono. (2019). Credit scoring to classify consumer loan using machine learning. AIP Conference Proceedings, 2194. https://doi.org/10.1063/1.5139802

[8] Niu, B., Ren, J., & Li, X. (2019). Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending. Information (Switzerland), 10(12). https://doi.org/10.3390/INFO10120397

[9] Orlova, E. v. (2020). Decision-making techniques for credit resource management using machine learning and optimization. Information (Switzerland), 11(3). https://doi.org/10.3390/info11030144

[10] Prastyo, P. H., Prasetyo, S. E., & Arti, S. (2021). A Machine Learning Framework for Improving Classification Performance on Credit Approval. IJID (International Journal on Informatics for Development), 10(1), 47–52. https://doi.org/10.14421/ijid.2021.2384

[11] Seo, J. Y. (2020). Machine Learning in Consumer Credit Risk Analysis: A Review. International Journal of Advanced Trends in Computer Science and Engineering, 9(4), 6440–6445. https://doi.org/10.30534/ijatcse/2020/328942020

[12] Shema, A. (2019, January 4). Effective credit scoring using limited mobile phone data. ACM International Conference Proceeding Series. https://doi.org/10.1145/3287098.3287116

[13] Simumba, N., Okami, S., Kodaka, A., & Kohtake, N. (2021). Spatiotemporal integration of mobile, satellite, and public geospatial data for enhanced credit scoring. Symmetry, 13(4). https://doi.org/10.3390/sym13040575

[14] Weng, C. H., & Huang, C. K. (2021). A Hybrid Machine Learning Model for Credit Approval. Applied Artificial Intelligence, 35(15), 1439–1465. https://doi.org/10.1080/08839514.2021.1982475

[15] Ziemba, P., Becker, J., Becker, A., Radomska-Zalas, A., Pawluk, M., & Wierzba, D. (2021). Credit decision support based on real set of cash loans using integrated machine learning algorithms. Electronics (Switzerland), 10(17). https://doi.org/10.3390/electronics10172099