Loan Default Risk Assessment using Supervised Learning

Anushi Jain[1], Shivangi Gupta[2], Mandeep Singh Narula[3]
[1,2]Student, [3]Assistant Professor, Dept. of ECE,
Jaypee Institute of Information Technology, Noida, India

Abstract:- The goal of this research is to develop a model for forecasting loan defaults. This type of strategy is unavoidable since bad loans are a critical problem in the financial sector. To address this issue, a literature analysis has been conducted to study the significant factors that lead up to and solve this problem. Dense Neural Network with Dropout (ANN with Deep Learning), XGBoost, Random Forest, Logistics Regression, and Support Vector Classifier are the approaches utilized. We have compared the models' accuracies, performance, and confusion matrix measures during the experimental phase. The best approach has been chosen, described, and suggested based on these factors. Our final results are based on the number of defaulters predicted and actualized, while we have also suggested a model if we prefer institutional research that prioritized accuracy, performance, and speed.

Keywords:- *Credit Score, Logistic Regression, XGBoost.*

I. INTRODUCTION

Since knowing clients and their behavior has become crucial in today's financial fields, we are working on Loan Default Risk Assessment. When there were a significant number of loan applications, manual techniques were usually effective, but they were insufficient and took a long time. As a result, the machine learning model for loan prediction can be used to assess a customer's loan condition and develop plans. To that end, multiple data analyses have been conducted on databases to predict the ability of a consumer to repay what they owe. In doing so companies detect patterns in their data to prevent income loss.

Lending Club, located in San Francisco, California, was a peer-to-peer lending platform. It used to be the world's largest peer-to-peer lending marketplace. The Lending Club allowed borrowers to establish unsecured personal loans ranging from $1,000 to $40,000. Three years was essentially the normal borrowing term. On the Lending Club website, investors could search and browse loan listings and choose loans to invest in based on the information provided about the borrower, loan size, loan grade, and loan purpose. The interest determined how investors gained money. Borrowers pay an origination fee, and investors pay a service fee, which was how Lending Club made money. This was changing the banking system to make credit more accessible and investing more profitable. However, there still existed a substantial chance of debtors defaulting on their debts. As a result, using the data obtained when the loan was provided, each borrower must be classified as a defaulter or not.

Observing the previous methods used in various research papers we have concluded that currently to solve the situation of Loan Risk Assessment using predictive methods based on machine learning algorithms. The main approach has been to assign client probability to an individual based on their payment history and profile features. Summary of the past work done on this problem statement is as follows.

In [1], Bagherpour's paper was based on a dataset of loans issued between 2001-2016, at quarter frequency. To estimate the loan defaults, he relied on K-Nearest Neighbours, SVMs. Factorization Machines, Random Forest classifiers. He came to the conclusion that nonlinear and non-parametric models/algorithms provided better results than traditional logistic regression models. It was also observed that Factorization Machines predicted AUC values between 88-91% which was the highest amongst other classifiers.

One year later, in 2018 Xiaojun, M. et al. employed two relatively new machine learning algorithms namely LightGBM and XGBoost for predicting loan default of customers based on physical peer-to-peer transactions from Lending Club.[2] Also, various studies show that their application showed a significant reduction in overfitting. With an error rate of 19.9% and an accuracy of 80.1%, LightGBM was found to give better results than XGBoost.

Kvamme, H. et al. in 2018 also proposed a new approach for default mortgage prediction by taking into consideration the time series data related to customer transactions within current accounts, savings accounts, and credit cards. This algorithm was implemented via Convolutional Neural Networks. The AUC for Neural Networks was 0.918, while the AUC for a combination of Neural Networks and Random Forest Classifiers was 0.926.[3]

Koutanei, F. N. et al. in 2015 had conducted a study and proposed a new model for a hybrid credit scoring system. This works on testing four feature selection algorithms and using ensemble learning classifiers. Amongst the features, Principal Component Analysis (PCA) was regarded as the best choice while for the classification part, an ANN adaptive boosting algorithm- Artificial Neural Network-AdaBoost was chosen.[4]

However, rather than using a binary categorization of excellent or bad payers, Kruppa, J. et al. (2013) employs machine learning methods to assess the chance of default. They argue that these algorithms' probability estimation is based on nonparametric regression and it compares several approaches based on random forests (RF), k-nearest neighbors (KNN), and bagged k-nearest neighbors (b-kNN). Finally, they discovered that the random forests model beats the other three approaches on the test data in terms of AUC scores.[5]

Khandani, A.E. et al. (2010) had proposed to use a set of features consisting of standard credit scoring, debt-to-income ratio, and consumer banking transactions to be utilized as input for the model. He corroborated that the transactional features increase the predictive power of the model greatly.[6]

While in 2011, Khashman A. recommended a new approach for predicting the credit risk by employing an emotional neural network which would account for the negative and positive confidence during the learning process and then the results were to be compared to a traditional neural network.[7] The author resolved that the emotional neural network had a better performance index than other neural network models in terms of speed, accuracy, and clarity.

Beque, A., Lessmann, S. worked with a new type of feed-forward neural network which compared its performance to that of traditional methods such as artificial neural networks, decision trees, regularized logistic regression, forests, and support vector machines called Extreme Learning Machine(ELM)[8]. They claimed that this new method marks a substantial step forward.

Harris, T. (2013) conducted a study on credit risk prediction using a support vector machine algorithm applied for two definitions of default: on one hand, a broader rule was considered for up to 90 days payment overdue; on the other hand, with more than 90 days late payment. He believes that the model employed for the broader definition is more accurate than the other one and that it is a more dependable.[9]

Zhang, T. et al. (2018) present a new methodology for developing a credit score model that includes socio-demographic, loan application data and applicant's transaction history data. This approach allows for the extraction of dynamic features from transactional data, and the results indicated that all classifiers used with newly added data had an impactful boost in accuracy.[10]

Papouskova, M., and Hajek, P. (2019) present a novel two-stage credit risk model: the first stage consists of a model that predicts the probability of default (PD) using ensemble classifiers; the second stage performs an in-depth analysis on customers with a predicted probability of default and uses a regression ensemble to determine the exposure at default (EAD). The two models are then merged to predict the projected loss in Predictive Models for Loan Default Risk Assessment (EL). [11]

A Study of Classification Based Credit Risk Analysis Algorithm by Ketaki Chopde, et. al. From this, they discuss credit score modeling, which divides loan applicants into two groups: Good Credit borrowers and Bad Credit borrowers. Financial organizations can enhance the amount of credit they issue while lowering possible losses by accurately judging applicants' credit qualifying.[12] Then, talk about how to use decision trees for credit risk analysis in different ways.

In [13], the authors do research about behaviors of default prediction models based on credit scoring methods with machine learning algorithms. The authors compare the prediction performance of different models with the data of the "My Home, My Life" program, and the results indicate that : the accuracy of models improves with the number of days overdue increasing, traditional ensemble techniques, bagging, random forest, and boosting.

II. PROBLEM STATEMENT

Developing a solution to assess loan default risk, using machine learning models and relevant datasets, and comparing their performance. So that banks can predict the capability of each candidate for paying back the loan beforehand. This can save a lot of time for both the bank and the applicant.

III. PROPOSED SOLUTION

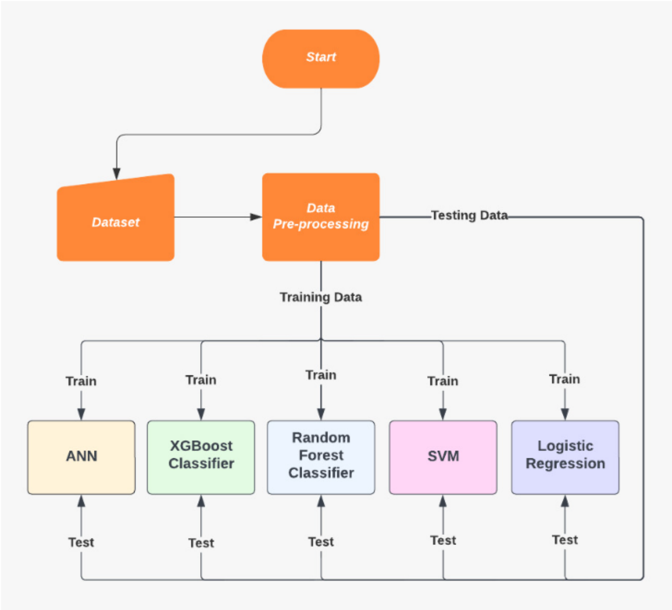


Fig. 1: Proposed solution flowchart

A. *Preprocessing Data*

We started data preprocessing by deleting columns based on how useful they were to us such as emp_length, emp_title, title, grade, issue_date.

For Null value imputation:

- For mortgage_acc, we found the highest correlated column total_acc and used it to fill null values.
- Since revol_util & pub_rec_bankruptcies had very few missing values we simply dropped those rows

Our dataset consists of 27 columns after cleaning, of which it is the loan_status column that classifies loans as either charged-off or fully paid. Following data preprocessing, we performed EDA where we used Pearson correlation to find which attributes were highly related to one another. We further explored the correlated attributes such as loan_amt and installment, grade and sub_grade, term, home_ownership, verification_status and purpose, emp_title, and emp_length, int_rate and annual income, isse_d and earliest_cr_line, etcetera for more inferences. We found that loans with subgrades F, G have lower chances of being repaid:

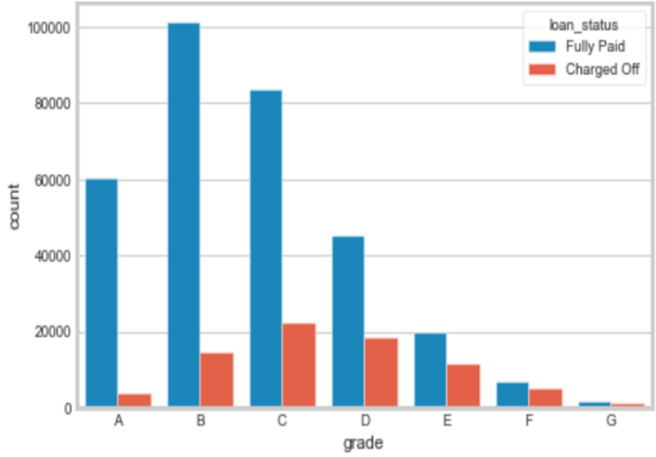


Fig. 2: Graphical Representation of Grade vs Count

We noticed that the higher the interest rate the lower the repayment chances, i.e they tend to be charged off.

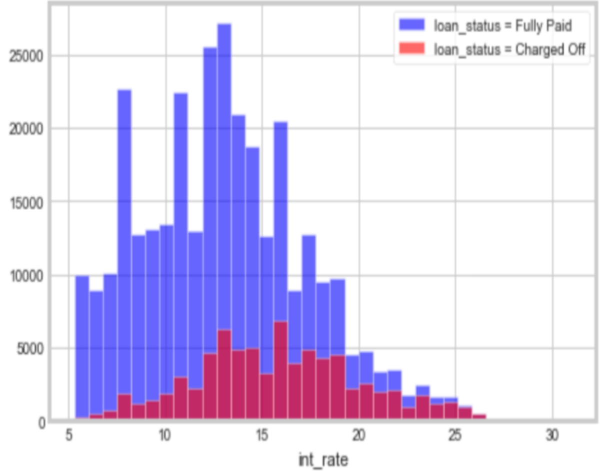


Fig. 3: Graphical Representation of Interest rate vs Count

In addition, we observed that the smaller the dti (debt to income ratio) the more likely it is that the loan will not be paid. On concluding the analysis we found the variables to be of two categories:

- Based on loan characteristics (int_rate, loan_amt, etc)
- Based on individual characteristics (annual_income, employment details, etc.)

We have performed feature engineering to remove redundant data, and extract information. We have created a new column called 'zip_code' from the address in the dataset. We have converted strings to categorical integers for normalization in 'earliest_cr_line'. We've extracted the year and converted this to a numeric feature. We have split the testing and training data with 70% for training and 30% for testing for the bias-variance tradeoff, based on positive and negative values.

B. *Building A Model and Evaluation*

We have created four different classification prediction models: Artificial Neural Networks (dense neural network with dropout), XGBoost Classifier, Random Forest Classifier, Logistic Regression, and Support Vector Machine. Compared performance of deep learning and machine learning.

In most circumstances, ANN is used when something that happened in the past is repeated in a similar way. It is a type of machine learning technology with a large memory capacity. Based on the patterns in our dataset, we can utilize this particular feature offered by ANNs.

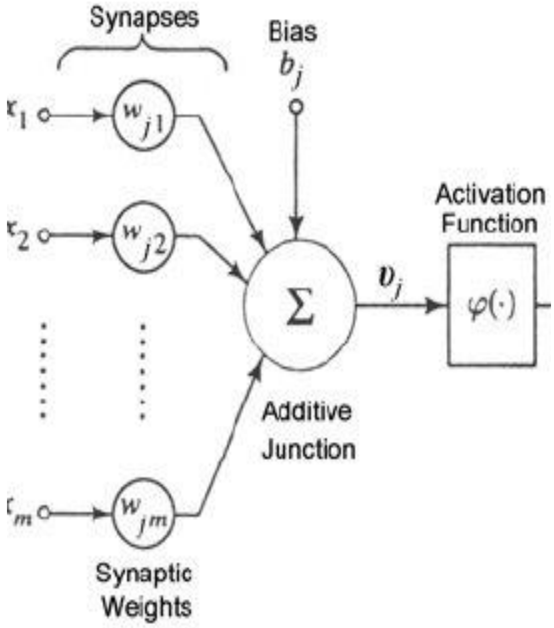


Fig. 4: ANN model from [14]

In ANN we adjusted the default batch size from 32 to 65536. Our model has 59413 parameters, of which 58357 are trainable. We implemented four drop-out layers within the neural network, which has 15 layers. We confined the model to 30 epochs because it was overfitting 150 epochs, giving us an accuracy of 22%. After reducing the number of epochs, we were able to achieve an accuracy of 88.74%. The validation accuracy is rather good, and the losses are also quite low.

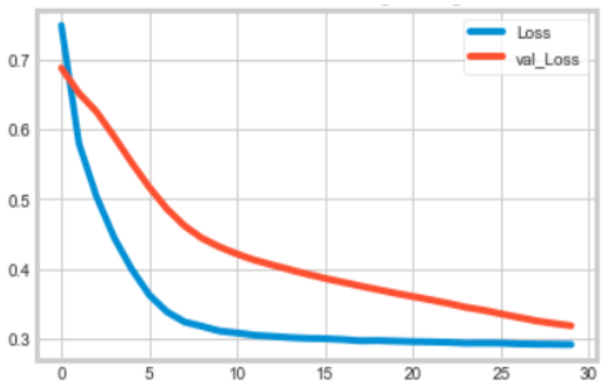


Fig. 5: Loss Evolution during training in ANN.

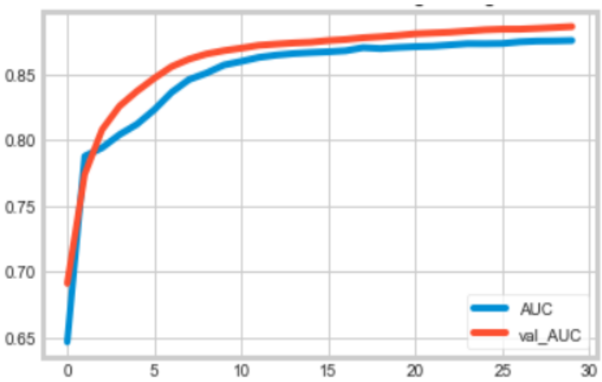


Fig. 6: AUC score during training in ANN.

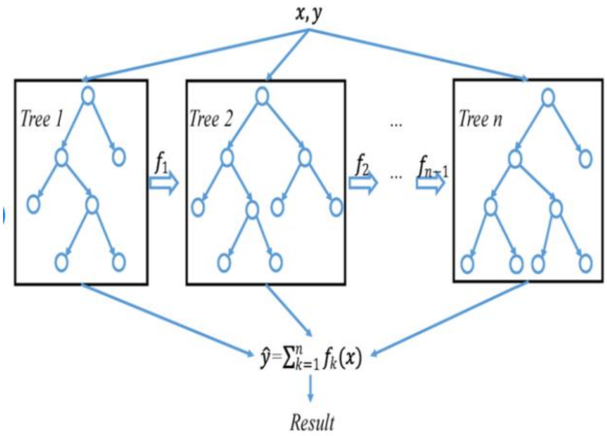


Fig. 7: General architecture of XGBoost from [18]

XGBoost is a self-contained and principled tool for boosting trees. It was created with careful consideration of both system optimization and machine learning techniques. The purpose of this library is to push systems to their boundaries in terms of computing in order to create a scalable, portable, and accurate library. It has a better confusion matrix than ANN.

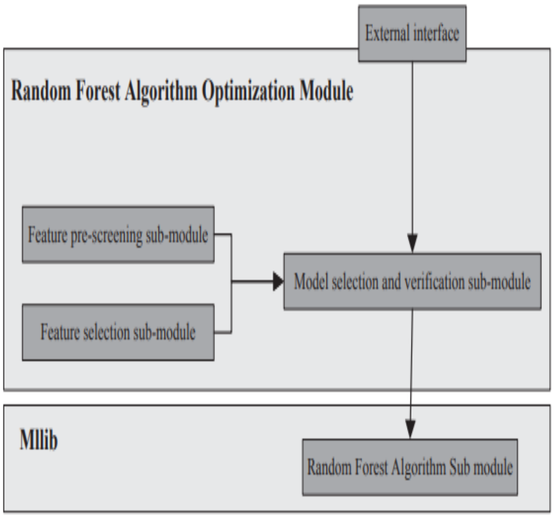

External interface
Random Forest Algorithm Optimization Module
Feature pre-screening sub-module
Feature selection sub-module
Model selection and verification sub-module
Mllib
Random Forest Algorithm Sub module

Fig. 8: Logical architecture view of random forest algorithm from [15]

For Random Forest Classifier, we normalized our data fairly well, then used the default 100 trees in the forest to model the forecast, which gave us an accuracy of 88.86%.In terms of accuracy, it is comparable to XG Boost, but it is slower than XGBoost and ANN. Since we have a large dataset and ease of understanding isn't a major concern, Random Forest seems like an appropriate choice as it is an ensemble method.

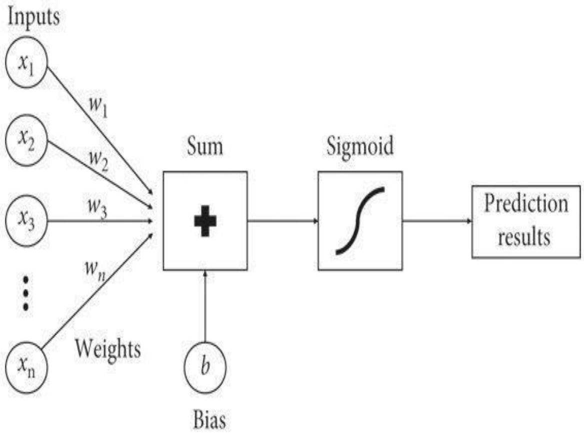


Fig. 9: Flowchart of logistic regression from [17]

For prediction of categorical dependent variables, we can also use Logistic regression. It is also valuable for predicting the likelihood of an event which is something we need for loan approval prediction. We only reached a 70% accuracy in Logistic Regression owing to overfitting when utilizing 500 max iterations. As a result, we reduced the number of iterations to 200, resulting in an accuracy of 88.89 percent. However, Logistic Regression was the most expensive in terms of time complexity, as it required the longest to run, analyze, and train the datasets. In addition, a convergence error was discovered where the maximum iterations were done before convergence.

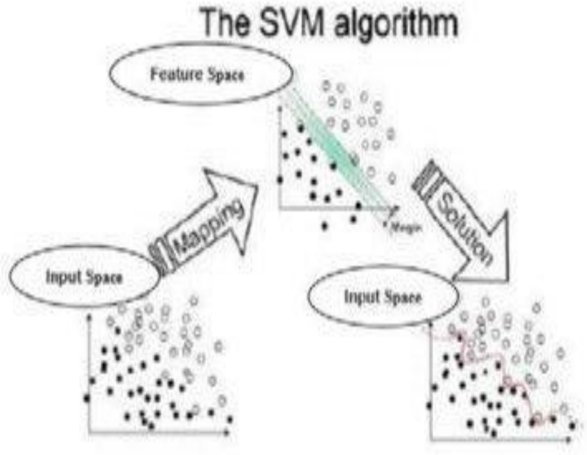


Fig. 10: Flow of SVM algorithm from [16]

SVC is more effective in high-dimensional spaces, especially when they are greater than the number of samples. In the Support Vector Classifier, the same issues as the Logistic regression model were observed but when the max number of iterations was changed from 500 to 200, only 2%

accuracy change was observed. This means that there was no overfitting due to the margin difference in the accuracy. The Support Vector Classifier also did not converge which was causing a bottleneck.

Both the Logistic Regression Classifier and the Support Vector Classifier are Euclidean models, which means they must attain convergence. However, they are approaching their maximum iterations before reaching convergence, which is nearly impossible given the data has 27 dimensions and almost 400k samples.

C. Synthetic Data Generation

Synthetic data is data generated by various algorithms that matches the statistical properties of the actual data but does not reveal any information about real people. We generated synthetic data with the same format and statistical properties by using the GaussianCopula class from SDV. We split our processed dataset into defaulter and paid dataset and then split these dataset into train and test data.We concatenated the sample synthetic data to our processed data and applied XGBoost to test the precision of our model with the synthetic data. The model gave us an accuracy of 88.45 percent along with a precision of 0.88.

Though the change in precision was insignificant , the accuracy of the model did not decrease after using data generation technique.

IV. EXPERIMENTAL RESULTS

For evaluation, we will consider the confusion matrix metric.

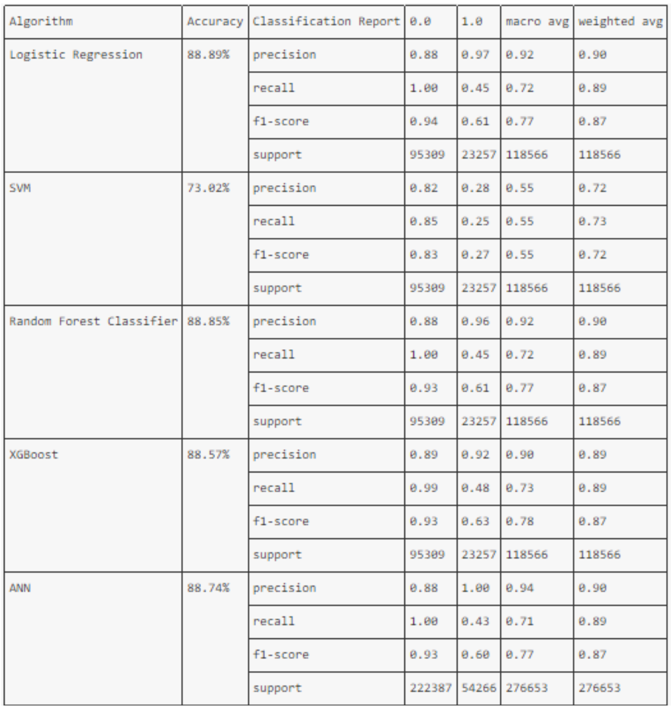
True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)

Fig. 11: Confusion Matrix for Loan Default Prediction

Where,

TP= Paid loan debt and was predicted to pay off.
FN= Paid loan debt but the prediction defaulter.
FP= Did not pay the debt but was predicted to pay off.
TN= Did not pay the debt and was predicted, defaulter.

In light of our current situation, we are concerned about defaulters if our models are used commercially. As a result, we must reduce False Positives, or circumstances in which the defaulter was expected to have paid off his obligations while maximizing True Negatives, or cases in which the defaulter was anticipated to have not paid off his debts.



Algorithm	Accuracy	Classification Report	0.0	1.0	macro avg	weighted avg
Logistic Regression	88.89%	precision	0.88	0.97	0.92	0.90
		recall	1.00	0.45	0.72	0.89
		f1-score	0.94	0.61	0.77	0.87
		support	95309	23257	118566	118566
SVM	73.02%	precision	0.82	0.28	0.55	0.72
		recall	0.85	0.25	0.55	0.73
		f1-score	0.83	0.27	0.55	0.72
		support	95309	23257	118566	118566
Random Forest Classifier	88.85%	precision	0.88	0.96	0.92	0.90
		recall	1.00	0.45	0.72	0.89
		f1-score	0.93	0.61	0.77	0.87
		support	95309	23257	118566	118566
XGBoost	88.57%	precision	0.89	0.92	0.90	0.89
		recall	0.99	0.48	0.73	0.89
		f1-score	0.93	0.63	0.78	0.87
		support	95309	23257	118566	118566
ANN	88.74%	precision	0.88	1.00	0.94	0.90
		recall	1.00	0.43	0.71	0.89
		f1-score	0.93	0.60	0.77	0.87
		support	222387	54266	276653	276653

Fig 12: Accuracy and Confusion Matrix of ML Models on Testing Data

In ANN, the notable points are that there are zero False Negatives i.e. no one has been mislabeled as a defaulter who had paid their debts. We should also acknowledge that 13,662 people were expected to repay their loans but ended up defaulting. While 9595 persons were expected to default, they really did.

We observe that XGBoost has a superior confusion matrix metric than ANN, since the False Positive of XGBoost is fewer than ANN's, with a value of 12158. Similarly, the True Negative, with a value of 11099, is greater than ANN. Comparatively, XGBoost is preferred if the concerned output is prediction and actuality of the number of defaulters.

The False Positives in Random Forest are 12837, while the True Negatives are 10420. Although Random Forest is not as excellent as XGBoost Classifier with respect to confusion matrix metrics. Also, it is clearly superior to Artificial Neural Network in terms of overall accuracy. Random Forest Classifier also takes longer to compute than ANN during execution.

We achieve an accuracy of 88.89 percent with Logistic Regression, but the performance of this classification model is 0.722 at all possible thresholds as calculated by AUC-ROC, which is very low when compared to ANN, XGBoost, and Random Classifier, making it a less preferred model.

Support Vector Classifiers, like Logistic Regression, have an AUC-ROC performance score of 0.501, which is quite volatile. Both these models suffer because of the convergence problem .

The GaussianCopula class from sdv was used to generate data similar to our processed data (with similar statistical properties).

We observe no significant change in accuracy and precision of XGBoost on synthetic data compared to that of XGBoost on our processed dataset.

V. CONCLUSIONS

Consequently, it can be concluded that the best model in terms of overall accuracy is Artificial Neural Network, more precisely Dense Neural Network with Dropout (Deep Learning). While in terms of speed and confusion matrix metrics XGBoost is superior. After the observations and execution of our models through the process we inferred from our data that most people return their debts, our data is skewed toward completely paid. And it's the incidents when folks don't pay in whole that are our subject of interest. As a result, rather than the more typical situation, we must forecast the rare occurrence of defaulters. This is known as outlier detection or anomaly detection. The better overall accuracy of ANN is due to the fact that it predicts positive outcomes with greater precision. The ANN model has one of the parameters, False Negative valued as 0, which is the strength for this model as no one is mislabeled a defaulter if they have paid off their debts.

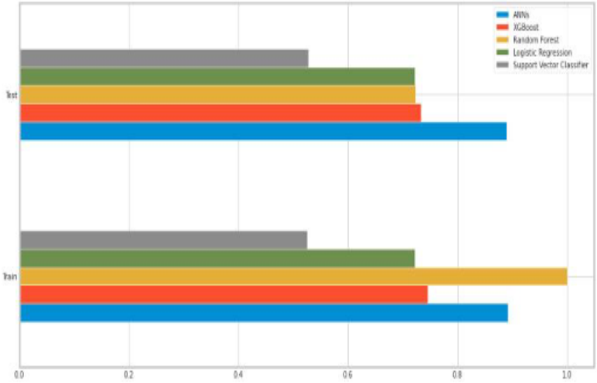


Fig.13: Graphical Comparison of results acquired via all the models employed.

Using the Area Under the ROC curve, the accompanying graph *Fig. 13* compares the performance characteristics of the models we used. We find that ANNs have the highest accuracy for training data, which they maintain for testing datasets as well, indicating that there is no overfitting. The AUC-ROC for Random Forest, on the other hand, is much lower. Also, while Logistic Regression achieved the maximum accuracy of 88.89 percent, its performance measure is considerably lower, therefore ANN, XGBoost, and Random Tree would still be favored over it. In comparison, the Support Vector Classifier still underperforms.

Even if the overall accuracy of Artificial Neural Networks is better, XGBoost is somewhat superior since ANN is a generalized model rather than a decision-based model. Even while False Negatives in confusion matrix measures are incredibly low in ANN, the minimization of False Positives, which is the statistic we genuinely care about as a corporation.

REFERENCES

[1.] Bagherpour, A. (2017), Predicting Mortgage Loan Default with Machine Learning Methods; University of California, Riverside;

[2.] Xiaojun, M.et al. (2018), Study on a Prediction Of P2P Network Loan Default Based on the Machine Learning Lightgbm and Xgboost Algorithms According to Different High Dimensional Data Cleaning; Electronic Commerce Research and Applications, 31, pp.24-39;

[3.] Kvamme, H. et al. (2018), Predicting Mortgage Default Using Convolutional Neural Networks; Expert Systems With Applications, 102, pp.207-217;

[4.] Koutanaei, F.N. et al. (2015), A Hybrid Data Mining Model of Feature Selection Algorithm and Ensemble Learning Classifiers for Credit Scoring; Journal of Retailing and Consumer Services, 27, pp.11-23;

[5.] Kruppa, J. et al. (2013), Consumer Credit Risk: Individual Probability Estimates Using Machine Learning; Expert Systems with Applications, 40, pp.5125-5131;

[6.] Khandani, A.E. et al. (2010), Consumer Credit-Risk Models via MachineLearning Algorithms; Journal of Banking & Finance, 34, pp.2767-2787;

[7.] Khashman, A. (2011), Credit Risk Evaluation Using Neural Networks: Emotional versus Conventional Models; Applied Soft Computing, 11, pp.5477-5484;

[8.] Beque, A., Lessmann, S. (2017), Extreme Learning Machines for Credit Scoring: An Empirical Evaluation; Expert Systems with Applications, 86, pp.42-53;

[9.] Harris, T. (2013), Quantitative Credit Risk Assessment Using Support Vector Machines: Broad versus Narrow Default Definitions; Expert Systems with Applications, 40, pp.4404-4413;

[10.] Zhang, T. et al. (2018), Multiple Instance Learning for Credit Risk Assessment with Transaction Data; Knowledge-Based Systems, 161, pp.65- 77;

[11.] Papouskova, M., Hajek, P. (2019), Two-stage Consumer Credit Risk Modeling Using Heterogeneous Ensemble Learning; Decision Support Systems, 118, pp.33-45;

[12.] KetakiChopde, Pratik Gosar, ParasKapadia, NiharikaMaheshwari, Pramila M. Chawan, “A Study of Classification Based Credit Risk Analysis Algorithm”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume- 1, Issue-4, April 2012.

[13.] J R de Castro Vieira, F Barboza, V A Sobreiro et al., "Machine learning models for credit analysis improvements: Predicting low-income families’ default[J]", *Applied Soft Computing*, vol. 83, pp. 105640, 2019.

[14.] Haykin, S. (1994) Neural Networks: A Comprehensive Foundation. Macmillan Publishing, New York.

[15.] Research on machine learning framework based on random forest algorithm, AIP Conference Proceedings 1820, 080020 (2017)

[16.] Chaudhari, Mohini & Govilkar, Sharvari. (2015). A Survey of Machine Learning Techniques for Sentiment Classification. International Journal on Computational Science & Applications. 5. 13-23. 10.5121/ijcsa.2015.5302.

[17.] Khan, Mohammad & Masud, Mehedi & Aljahdali, Sultan & Kaur, Manjit & Singh, Parminder. (2021). A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease. Journal of Healthcare Engineering. 2021. 10.1155/2021/9917919.

[18.] Wang, Yuanchao & Pan, Z. & Zheng, J. & Qian, L. & Mingtao, Li. (2019). A hybrid ensemble method for pulsar candidate classification. Astrophysics and Space Science. 364. 10.1007/s10509-019-3602-4.