IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Loan Status Prediction in the Banking Sector using Machine Learning

[1]R.Selvaganesh, [2] Sk Magbul Basha, [3]Sk Hasan Ali, [4]S Ashraf Ali, [5] S Sai Kiran

[1]Assistant professor, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India- 600073.

[2, 3,4,5] Student, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India- 600073.

Abstract— There have been tremendous advancements in technology that have led to an increase in the needs of people. This, in turn, has led to more loan approval requests in the banking sector. Some attributes are considered to check loan status while selecting an applicant for loan approval. Banks are facing a serious challenge when it comes to evaluating loan applications and mitigating the risks associated with borrowers potentially defaulting on their loans. This process is particularly laborious for banks, as they must thoroughly verify each individual's loan eligibility. This paper proposes utilizing machine learning models with Ensemble Learning Techniques to determine the viability of granting individual loan requests. By using this approach, it is possible to enhance the accuracy with which suitable candidates are selected from an existing list. Therefore, this process can be used to address the aforementioned concerns surrounding loan approval processes. The model is helpful to both bank staff and applicants as it drastically reduces the time taken to sanction the loan.

Keywords—Loan Status, Loan Lending, Decision Tree, Random Forest, Machine Learning

I. Introduction

Loan prediction is a problem in the banking sector. This problem arises when a bank is unable to accurately predict which loans will default. This problem can lead to a loss of revenue for the bank, and can also lead to the bank being unable to provide loans to worthy borrowers. There are several factors that can contribute to a bank's inability to predict which loans will default [16]. These factors include the borrowers' financial status, the quality of the loan, and the market conditions.

Banks can attempt to improve their loan prediction by using a variety of techniques. These techniques include reviewing past loan data, conducting research into the borrowers' backgrounds, and using machine learning algorithms. However, even with the best techniques, a bank's ability to predict which loans will default is still imperfect. One of the approaches to beat this issue is to utilise machine learning techniques. Machine learning is a subset of artificial intelligence that is utilised to create calculations dependent on information, so they can get the hang of improving after some time. The machine learns without being unequivocally modified and can improve given more information. Machine learning depends on two primary classes: supervised and unsupervised learning [8].

The table given below mentions some of the features that we have considered for the prediction in our model. For example, if a person is educated to a specific degree, he or she will be able to work and repay the debt.

Or if a person has less number of dependents then his expenditure will be much lesser than a person who has more number of dependents. Credit history lets us define if the person has been actively working or not and can pay the loan or not in that condition.

TABLE I. Data Set with Their Respective Type

Variable	*Description*
Loan_ID	Unique Loan ID
Married	Applicant married(Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/Undergraduate)
Self_Employed	Self_Employed
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of lean in months
Credit_History	Credit history meets guidelines
Property_Area	Urban/Semi Urban/Rural
Loan_Status	Loan Approved (Y/N)

In order to accurately assess a potential applicant's creditworthiness, it is necessary to utilize both specialist knowledge and statistical algorithms as part of the credit scoring process. Financial institutions have recently sought to remediate the challenges of manual credit scoring by training classifiers based on various machine learning and deep learning algorithms [18]. This allows the system to automatically predict an applicant's credit score based on their credit history and other attributes, thereby improving the accuracy of the process to select eligible candidates.

This paper aims to investigate the efficacy of various machine learning models for loan lending purposes, in order to evaluate the most efficient strategy for a financial institution to accurately identify prospective borrowers and prevent loan defaults. We will use various classifiers to work out an ensemble machine learning model to collect information about loan defaulters. We will also compare its efficiency to the results obtained from single classifiers. The classifiers will be utilised to analyse the dataset, find its patterns, and draw conclusions from them. Establish the likelihood that a new applicant will default on a loan based on that analysis.

II. LITERATURE SURVEY

In the paper [1], researchers employed multiple machine learning methods to create a machine learning model. They developed a distinct model for each machine learning technique applied in the creation of a protection model. They employed a variety of machine algorithms, including the decision tree algorithm, the Bayes algorithm, and the random forest method. Out of all the models examined, the model developed with the decision tree method achieved the greatest accuracy rate of 81%. It is followed by the random forest algorithm, which has a 77% accuracy rate, and the Bayes algorithm, which has a 69% accuracy rate.

Tijo and Abdulazeez [2] has done a complete analysis of the most recent and efficient methods to decision trees in many domains of machine learning that have been undertaken by researchers in the last three years is performed.

This paper employed an ensemble model to predict loan applicants' credit risk, utilizing nine features and eleven machine learning models [3]. To assess the effectiveness of the model, several measures, including Accuracy, Gini, Auc, Roc, etc., were evaluated. The main goal of this work is to evaluate models and create a new model called the ensemble model, which combines the findings of three different models to predict consumer loans.

This paper [4] presents a Bank Loan Prediction System using Machine Learning. The system is built on IBM Cloud Watson Studio which uses IBM Watson Machine Learning (WML) API for predictive analysis and IBM Cloud Object Storage to store data. The model uses two of the most commonly used Machine Learning algorithms, namely Logistic Regression and Support Vector Machines (SVM). The dataset used was from a Kaggle dataset which contains various financial and loan related details of the customers. The performance of both the algorithms has been tested on various performance criteria like precision, recall, accuracy etc. and it was observed that logistic regression performed better in terms of accuracy, precision and recall.

This research paper [6] focuses on the use of an Ensemble Classifier for Corporate Default Prediction, specifically using Adaboost. Adaboost is a powerful machine learning technique that enables a combination of weak classifiers to achieve improved accuracy and performance in classification tasks, and is applied in this study to the task of predicting whether or not a corporation will default on its debt. The experiment was done by training Adaboost with different combinations of attributes and parameters, with the best results achieved by combining three attributes (Total Asset, Current Ratio and Equity) and using a cost-sensitive structure. The results revealed that Adaboost was able to achieve an accuracy of 96.4% for default prediction and outperform individual classifiers such as Naïve Bayes and Decision Tree.

This paper [13] focuses on the use of AzureML based analysis and prediction to determine the creditworthiness of loan borrowers. The methods proposed by the authors, include the integration of datasets from multiple sources, structured feature engineering and selection, and advanced analytics such as supervised learning algorithms. The results of their experiments demonstrate that the proposed framework accurately identifies the creditworthiness of loan borrowers with an accuracy rate of 96%. Furthermore, the paper discusses the potential applications of this framework and its implications for the banking industry.

III. METHODOLOGY

The paper presents three distinct algorithms and ensemble methods to train a large dataset of borrowers. These algorithms and methods rely on a combination of pre-processing techniques, machine learning approaches, and text mining to ensure high accuracy, scalability, and robustness.

A. Machine Learning

Machine Learning allows systems to develop an understanding of the interactions and relationships between people and objects that occur in real life. Through this system, machines can become increasingly intelligent, similar to humans, while also having the capability to constantly refine their skills by taking in new data. ML has now attracted so much attention and interest from academics and developers all over the world that they are looking to use Machine Learning techniques and algorithms in areas that will make tough tasks simpler and improve the daily lives of people. As for our problem, it is now being used in financial institutions as well to deploy efficient and accurate models to help identify problems with the existing models. Previous studies have demonstrated the wide range of issues that machine learning can be used to and how it has greatly aided in the creation of better solutions in today's fast paced technological advancements.

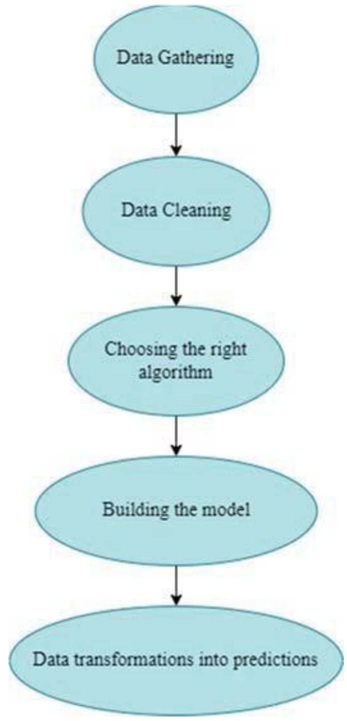

Fig. 1. Process of machine learning modelling

Financial institutions and banking authorities have increasingly been leveraging the power of Machine Learning (ML) models to identify patterns and draw conclusions related to topics such as credit card fraud and loan default prediction. By using these advanced models, they are able to detect irregularities more quickly and efficiently. It has helped to perform credit assessment with better efficiency and with more precise results to avoid loss for the financial bodies through the world. There is still room for improvement in existing machine learning models, which is what we are attempting here.

The aforementioned models were developed using a variety of machine learning techniques. For default prediction, we compare a range of classifiers, such as Decision Tree (DT), Random Forest (RF), and support vector machines (SVM), with ensemble approaches such as stacking ensemble, which employs many classifiers and feeds the resulting findings to a meta learner for prediction [12]. Stacking ensemble delivers better outcomes than bagging and boosting approaches, according to research, however this relies on the task at hand.

B. Decision Trees

Decision trees are a practical and versatile tool used for classification and regression tasks, such as calculating the optimal price or preferences. This technique can efficiently provide insights that can guide decisions. The tree is a graph of nodes with parents, branches and leaves that represent decisions. They have some advantages over other decision methods such as linear regression in that they are easily interpretable, less computationally expensive, and they can be trained with a minimum amount of data [11].

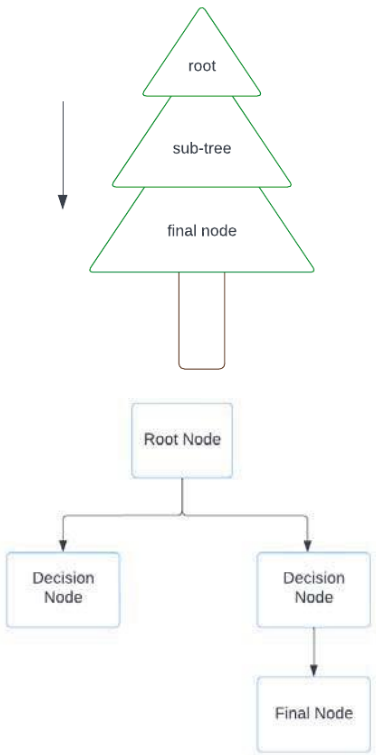

Fig. 2. Structure of a decision tree in resemblance to a tree

C. Random Forest

Random forest is an ensemble of decision trees developed to provide a "forest" structure for finding patterns in high-dimensional data. Like naive Bayes, random forests use "atree" of decision trees that grow in increasing levels of complexity from the parent atree. The output of each atree of trees is combined to get final prediction. Random forest can also be used as a feature selection method, in which the number of features used is adjusted depending on the available data. It was originally invented to overcome the failure of the bagging technique in predicting the output values from a set of input variables. The main advantage of random forests over other algorithms such as neural networks is that they can be constructed relatively easily and have high predictive accuracy [19].

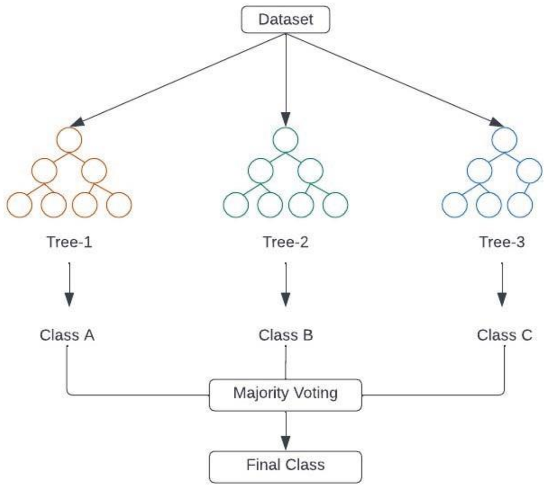

Fig. 3. Visualization of Random Forest Algorithm

D. Support Vector Machines

The support vector machine, or SVM, is a classification algorithm. It can also be used for regression. On this model, each feature is represented as a distinct point in an n-dimensional plane, where n is the total number of features. Then, we perform the process of classification by finding the hyperplane that differentiates our two classes. SVM is different from other classification algorithms such as decision tree and logistic regression in that it doesn't make any assumption about the input or target values. Another important distinction is its ability to handle missing data due to machine failure, i.e., by "filling in" new data from other sources.

It is a powerful machine learning algorithms used in a variety of tasks such as loan prediction. SVMs work by constructing hyperplanes in a high-dimensional space to separate data sets into different classes. The hyperplanes are determined by the data points closest to them, known as support vectors. SVMs use these support vectors to make predictions about new data points that lie outside of the original data set. SVMs are especially useful in loan prediction tasks because they are able to accurately classify complex data sets and make predictions based on the data. By using support vectors, SVMs can accurately classify data points and make accurate predictions about loan applicants.

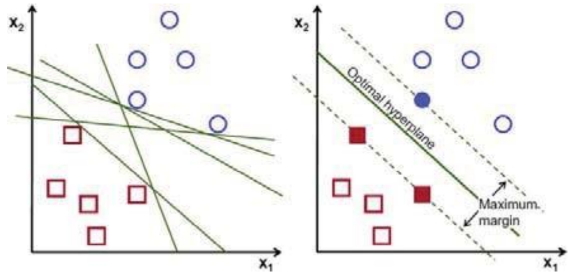

Fig. 4. Possible positions of hyperplanes and optimal hyperplane

E. Ensemble Methods

Ensemble methods are a way to combine multiple machine learning models to reach the best possible decision when faced with an unknown data distribution [14]. This method is used to reduce the variance in the results and increase their overall accuracy. Ensemble methods include bagging and boosting. Bagging uses many randomly selected decision rules, whereas boosting uses successive re-applications of those strategies on each instance.

Stacking is a technique in machine learning which uses a meta-learning algorithm to combine the predictive power of two or more underlying machine learning algorithms. Also known as Stacked Generalization, it can lead to significantly better results than any single model alone. The main benefit we see of using stacking is that it may combine the talents of many high-performing models on a classification or regression problem to provide results that surpass any one model in the ensemble. Stacking with more than two levels is also conceivable. The first layer consists of base learners or 0-level learning models. In the second layer, a meta learner or a level-1 learner receives the judgments of the basic models and produces the final prediction.

AdaBoost, as opposed to stacking, works by picking a base approach and then iteratively improving it for the provided classified training dataset. Before picking a base algorithm, this method applies equal weights to all training data. The fundamental method is applied on the data set at each level, and the weights of incorrectly classified data are increased. Its weights do not change if it is erroneously classified. The algorithm is responsible for minimising errors throughout the learning process.

AdaBoost and XGBoost are both methods that have been accepted as effective for classification problems. In low noise datasets, AdaBoost is generally resistant to overfitting and has just a few hyperparameters that must be modified to optimise model performance [7]. However, given noisy data, its efficacy is debatable, resulting in poor performance as a result of the algorithm spending too much time learning extreme instances and skewing outcomes. Furthermore, AdaBoost is not tuned for speed, therefore it is substantially slower than XGBoost. We try to explore XGBoost furthermore in this study.

In this paper, we employ the Meta learner XGBoost Classifier which is itself a gradient boosting ensemble. We use an ensemble as the meta learner of stacking ensemble. Meta learner XGBoost Classifier can achieve better results than other optimization algorithms, like boosting in some way or another.

IV. Exploratory Data Analysis

We're not ready to go into machine learning modelling just yet. We obtain our dataset from a trusted source, and it must be processed before it can be used to train and test our technique. The first crucial step was to load the necessary libraries as well as the model's data files, such as numpy, pandas, and seaborn [5]. In Python, we utilise the scikit package to do this. The next step was to conduct an exploratory data analysis on the given data to explore its features and examine things like: features that differentiate each loan, what attributes can give us redundant results, previous loan defaulter patterns, and the most suitable method for cleaning the dataset based on our criteria and needs [9]. It is critical to ask these questions in order to improve our dataset and utilise it for modelling to get relevant outcomes.

"Fig. 5." depicts a data description with their appropriate data types to assist us in visualising the quantitative and categorical characteristics in our dataset [10]. The dataset contains features that may result in null results for a certain column. In order to drop columns that do not meet a certain percentage threshold, it is necessary to compute the proportion of null values in each column as outlined in "Fig. 6" of the data set. After this data cleaning step has taken place, exploratory data analysis should then be conducted.

```
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Loan_ID            614 non-null    object
 1   Gender             601 non-null    object
 2   Married            611 non-null    object
 3   Dependents         599 non-null    object
 4   Education          614 non-null    object
 5   Self_Employed      582 non-null    object
 6   ApplicantIncome    614 non-null    int64
 7   CoapplicantIncome  614 non-null    float64
 8   LoanAmount         592 non-null    float64
 9   Loan_Amount_Term   600 non-null    float64
 10  Credit_History     564 non-null    float64
 11  Property_Area      614 non-null    object
 12  Loan_Status        614 non-null    object
dtypes: float64(4), int64(1), object(8)
memory usage: 62.5+ KB
```

Fig. 5. Description of attributes in the given dataset against their data types

It is an important foundation during exploratory data analysis to be able to identify the type and colour of a relationshipIt was essential to gain familiarity with the different correlations present in the data before progressing to classification, for which various types of plots were utilized [15]. Analyzing these correlations helped us in understanding the outcomes of the model. Inquiring about these ties supplied us with information about relationships we may not have realised existed.

```
Loan_ID              0
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
Loan_Amount_Term    14
Credit_History      50
Property_Area        0
Loan_Status          0
dtype: int64
```

Fig. 6. Total count of null values in each attribute of the dataset

You can choose to drop all empty or missing values from your entire dataset by using the drop() function. We will drop the Dependents feature because it has many missing values. To improve the quality of our dataset, we should replace any blank fields with values calculated by statistical methods such as mean, mode or median. After cleaning, we must guarantee that there are no blank fields in the dataset.

The below figures visualizes the value counts of loan approved or not for different attributes such as gender, education, self-employed and married. In this example, it can help users to understand loan approval rate by gender, education and marital status.

“Fig. 7.” shows the value count of the number of times the loan was accepted or denied based on the Gender property.

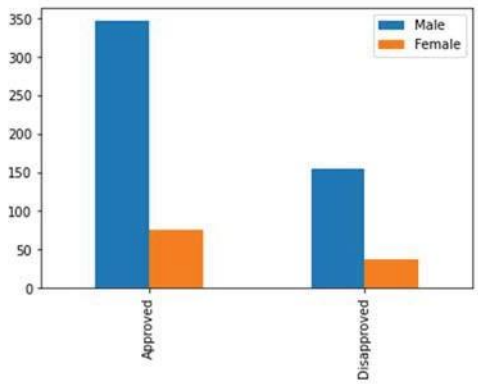

Fig. 7. Bar Graph for Gender Property

“Fig. 8.” shows the value count of the number of times the loan was accepted or denied based on the Educated property.

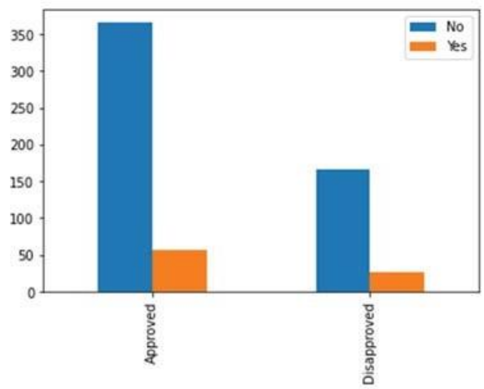

Fig. 8. Bar Graph for Educated Property

“Fig. 9.” shows the value count of the number of times the loan was accepted or denied based on the Self_Employed property.

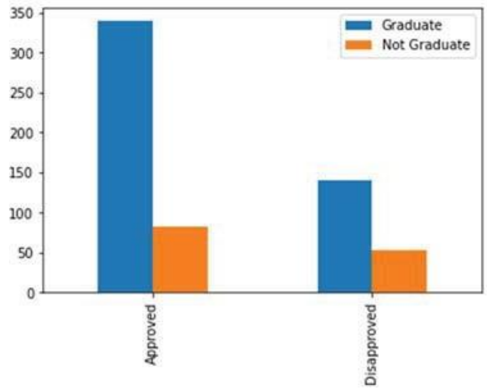

Fig. 9. Bar Graph for Self_Employed Property

“Fig. 10.” shows the value count of the number of times the loan was accepted or denied based on a person's marital status hat is the Married property.

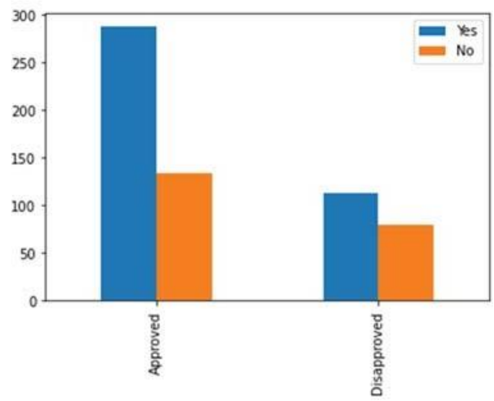

Fig. 10. Bar Graph for Married Property

Real coded genetic algorithms (RCGA) are an evolutionary computing technique that uses genetic operators such as selection, crossover, and mutation to search for solutions to a problem. This technique is used to optimize numerical parameters in a given problem domain. A key component of RCGA is k-fold cross validation, which is used to validate the performance of the algorithm by dividing the dataset into k parts, training the algorithm on k-1 parts, and then testing the performance on the remaining part [17]. This process is repeated k times, and the average performance is used to evaluate the algorithm's performance. K-fold cross validation helps to reduce the chances of overfitting the data, making RCGA a reliable and effective optimization technique.

As shown in “Fig. 11.”, K-fold cross validation entails splitting the data into k subgroups, training the model on the first k-1 subsets, then testing on the remaining subset. This procedure is performed k times, with each subset acting as the test set just once. The final prediction is then calculated as the average of the predictions made on each of the k folds.

Fig. 11. Depiction of K-Cross Validation Algorithmic method

V. Result

In this paper, we employed three machine learning algorithms as base learners: Random Forest (RF), Decision Trees (DT), and Support Vector Machine (SVM) to apply the Meta learner XGBoost Classifier, which is a stacking ensemble meta learner, to develop the model for loan approval and credit evaluation. We used the open-source scikit-learn library for implementing our model in python.

The model's output and accuracy results are provided in the figures below : “Fig. 12.” for DT, “Fig. 13.” for RF and “Fig. 14.” for XGBoost Algorithm.

```
ds = DecisionTreeClassifier(max_depth=8, max_features=0.9, max_leaf_nodes=30,
                            min_impurity_decrease=0.05, min_samples_leaf=0.02,
                            min_samples_split=10, min_weight_fraction_leaf=0.005,
                            random_state=2, splitter='random')
ds.fit(X_train, y_train)
pred4 =ds.predict(X_test)
loss(y_test, pred4)

0.7592592592592593
0.9761904761904762
0.7723577235772358
```

Fig. 12. Results for DecisionTreeClassifier

```
from sklearn.ensemble import RandomForestClassifier

randomized_search(params={
                            'min_samples_leaf':[1,2,4,6,8,10,20,30],
                            'min_impurity_decrease':[0.0, 0.01, 0.05, 0.10, 0.15, 0.2],
                            'max_features':['auto', 0.8, 0.7, 0.6, 0.5, 0.4],
                            'max_depth':[None,2,4,6,8,10,20],
                           }, clf=RandomForestClassifier(random_state=2))

Training score: 0.819
Test score: 0.772

RandomForestClassifier(max_depth=2, max_features=0.5,
                       min_impurity_decrease=0.01, min_samples_leaf=10,
                       random_state=2)

import joblib
joblib.dump(ds, "model.pkl")
model = joblib.load('model.pkl' )
model.predict(X_test)

array([0., 0., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 0., 1.,
       1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 0., 1., 1., 1.,
       1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 0.,
       0., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 0., 1., 1.,
       1., 1., 1., 1., 0., 1., 1., 1., 1., 1., 0., 1., 1., 0., 1., 1., 1.,
       1., 1., 0., 1., 1., 1., 1., 0., 1., 1., 1., 1., 1., 1., 1., 1., 1.,
       1., 1., 1., 1., 0., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 0., 1.,
       0., 1., 1., 1.])

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
forest = RandomForestClassifier()
forest.fit(X_train, y_train)
y_pred_test = forest.predict(X_test)
accuracy_score(y_test, y_pred_test)

0.7560975609756098
```

Fig. 13. Results for RandomForestClassifier

```
from xgboost import XGBClassifier

xgb = XGBClassifier(learning_rate =0.1,
 n_estimators=1000,
 max_depth=3,
 min_child_weight=1,
 gamma=0,
 subsample=0.8,
 colsample_bytree=0.8,
 objective= 'binary:logistic',
 nthread=4,
 scale_pos_weight=1,
 seed=27)
xgb.fit(X_train, y_train)
pred3 = xgb.predict(X_test)
loss(y_test, pred3)

0.7954545454545454
0.8333333333333334
0.7398373983739838
```

Fig. 14. Results for XGBClassifier

Following modelling, we ran several classifiers on our dataset and achieved results from both single classifiers and the ensemble technique. We studied the results further to determine what worked for our challenge and where we might enhance the model's evaluation.

DecisionTreeClassifier gave us an accuracy of 0.7723577235772358. RandomForestClassifier gave a slightly lower accuracy than Decision Trees, of about 0.7560975609756098. The SVM Classifier alone gave an accuracy of 0.6829268292682927. Our meta learner, the XGBClassifier, gave us an accuracy of 0.7398373983739838, which we may increase later by recording our process and examining the places where we can adjust or do better.

VI. Conclusion

This research has addressed the challenge of accurately predicting loan status within the banking sector using machine learning. This work has explored various machine learning algorithms and their associated performance results for the loan status prediction task. The analysis found that XGBoost Classifier outperforms Decision Tree, Random Forest and Support Vector Machine for loan status predictions in the banking sector. XGBoost Classifier achieved an accuracy of 73.98%. The slight improvement in performance due to XGBoost Classifier is due to its efficient use of time-series data, robustness to overfitting, and greater classification accuracy achieved over the other three machine learning approaches. This indicates that XGBoost Classifier is a suitable tool for predicting loan status in the banking sector. It is therefore recommended for organizations in the banking sector that are planning to use machine learning for loan status predictions.

These results demonstrate that machine learning can be used to accurately predict loan status in the banking sector. Moreover, this research has highlighted the potential benefits of machine learning in improving customer loan experience. Such an approach could reduce overhead costs associated with manual loan processing and potentially increase business efficiency. It is important to note that although the machine learning approach is more accurate than traditional analysis, there are still challenges associated with this approach such as data availability, pre-processing and feature selection. In addition, proper assessment must be done to determine the cost-benefit of using machine learning approaches as an alternative to traditional methods.

Overall, this research has established that Machine Learning algorithms can be used to effectively and accurately predict loan status in the banking sector. This work has laid a foundation for further exploration of using machine learning to improve customer experiences and mitigate business risks.

VII. Future Scope

The scope of this research paper can be further extended to cover various other aspects of loan status prediction in the banking sector such as customer profiling, financial statement analysis, credit scoring and behavioral analysis. Additionally, the scope can be expanded to include big data analytics, advanced algorithms and artificial intelligence to provide more accurate loan status prediction. For example, incorporating natural language processing (NLP) technology can help identify patterns and features from the financial data that might not be visible in the traditional methods. Furthermore, cloud computing can provide faster and more efficient computing resources for analyzing large datasets. Lastly, incorporating augmented reality and virtual reality technology into loan status prediction could potentially be used as another method of user interaction with the system.

References

[1] Cowell RG AP, Lauritez SL DJ. Graphical models and Expert Systems. Berlin: Springer. This is a good introduction to probabilistic graphical models, 1999.

[2] Tijo & Abdulazeez / Journal of Applied Science and Technology Trends Vol. 02, No. 01, pp. 20 - 28 (2021).

[3] Anchal Goyal & Ranpreet Kaur "Loan Prediction Using Ensemble Technique" / International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.

[4] Gupta, A., Pant, V.K., Kumar, S., Bansal, P.K. (2020). Bank Loan Prediction System using Machine Learning. 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), 423-426.

[5] Mehul Madaan et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012042.

[6] Ramakrishnan, Suresh & Mirzaei, Maryam & Bekri, Mahmoud. (2015). Adaboost Ensemble Classifiers for Corporate Default Prediction. Research Journal of Applied Sciences, Engineering and Technology. 9. 224-230. 10.19026/rjaset.9.1398.

[7] M. Sheikh, A. Goel, T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," International Conference on Electronics and Sustainable Communication Systems (ICESC), (2020).

[8] P. Dutta, "A STUDY ON MACHINE LEARNING ALGORITHM FOR ENHANCEMENT OF LOAN PREDICTION", International Research Journal of Modernization in Engineering Technology and Science, (2021).

[9] R. Salvi, R. Ghule, T. Sanadi, M. Bhajibhakare, "HOME LOAN DATA ANALYSIS AND VISUALIZATION," International Journal of Creative Research Thoughts (IJCRT), (2021).

[10] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," Journal of Applied Science and Technology Trends, vol. 1, no. 4, pp. 140–147, 2020.

[11] Priyanka and D. Kumar, "Decision tree classifier: a detailed survey," International Journal of Information and Decision Sciences, vol. 12, no. 3, pp. 246-269, 2020.

[12] I. Ramadhan, P. Sukarno, and M. A. Nugroho, "Comparative Analysis of K-Nearest Neighbor and Decision Tree in Detecting Distributed Denial of Service," in 2020 8th International Conference on Information and Communication Technology (ICoICT), Yogyakarta, Indonesia, Jun. 2020.

[13] Alshouiliy K, Alghamdi A and Agrawal D P 2020 AzureML based analysis and prediction loan borrowers creditworthy The 3rd Int. Conf. on Information and Computer Technologies (ICICT).

[14] Short-term prediction of Mortgage default using ensembled machine learning models, Jesse C.Sealand on july 20, 2018.

[15] Li M, Mickel A and Taylor S 2018, "Should this loan be approved or denied?": a large dataset with class assignment guidelines Journal of Statistics Education.

[16] Vaidya A 2017 Predictive and probabilistic approach using logistic regression: application to prediction of loan approval The 8th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT).

[17] Shoumo S Z H, Dhruba M I M, Hossain S, Ghani N H, Arif H and Islam S 2019 Application of machine learning in credit risk assessment: a prelude to smart banking TENCON 2019 – 2019 IEEE Region 10 Conf. (TENCON).

[18] Addo P M, Guegan D and Hassani B 2018 Credit risk analysis using machine and deep learning models Risks.