Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)
IEEE Xplore Part Number: CFP21K25-ART

2021 5th International Conference on Computing Methodologies and Communication (ICCMC) | 978-1-6654-0360-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCMC51019.2021.9418277

Swindle: Predicting the Probability of Loan Defaults using CatBoost Algorithm

1st Sujoy Barua
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
sujoy.barua@spit.ac.in

2nd Divya Gavandi
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
divya.gavandi@spit.ac.in

3rd Pooja Sangle
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
pooja.sangle@spit.ac.in

4th Leena Shinde
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
leena.shinde@spit.ac.in

5th Jyoti Ramteke
Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
jyoti_ramteke@spit.ac.in

***Abstract*—Predicting the probability of loan defaults is essential for financial institutes and banks, as a major part of their income is dependent on the interest & EMIs generated on the repayment of the loans issued by them to their customers. Most of the loans issued have a high interest rate associated with them due to lack of securities and uncertainty possessed by the customers. Hence, having a model that could predict loan defaulters would be very beneficial for the financial institutes and banks for notifying them to approve a customer's loan or not. Such a model will evaluate their customer's data based on certain parameters and generate an accurate result based on that evaluation. Swindle implements CatBoost algorithm is used for predicting loan defaults along with a document verification module using Tesseract and Camelot and also a personalized loan module, thereby mitigating the risk of the financial institutes in issuing loans to defaulters and unauthorized customers.**

***Keywords*—Loan, loan defaults, loan approval, CatBoost algorithm, credit risk prediction, pytesseract, machine learning, personalized loan, credit score, optical character recognition, document verification, banking, finance.**

I. INTRODUCTION

In India, the banking sector contributes approximately 7.7% to the Indian GDP. It is very important for banking sectors to manage and mitigate the risk, as a huge amount of money is at stake. The banking sector in India knows about the risks and have made strict laws in order to mitigate or avoid any risks. According to the RBI (Reserve bank of India) there has been Rs. 68,600 crore loans of wilful defaulters written off. As AI and ML are taking over in every sector in the industry, banking is no exception. By using ML models banks can predict the probability of a loan default, thus helping them in mitigating the risk of wilful defaulters. Granting loans to customers who will default would be expensive for the banks (False Positive), but it is also expensive to not grant loans to customers who would pay them back on time (False Negative) as banks earn from the interest paid by such customers. Also the rise of Peer-to-Peer lending in India has made it very essential for such a model to exist. People are tending towards Peer-to-Peer lenders as they offer a low interest rate compared to banks.

One of the most important factors for banks to verify is the documents of their customers. It is essential to provide the banks with authentic original documents verification of documents in banks is done by professionals employed by the bank as humans tend to make errors they can miss out on disparities. Employing ML models to verify forged documents will help the banks in determining the authenticity of its customers and maintaining their records. Identity theft is a crime and there are a lot of cases where individuals impersonates somebody else to access one's bank account or to approve loans. Hence it is fair to say that document verification is essential before any activity or service provided by the banks.

Individuals sometimes find it difficult to find appropriate loans for them, there are various parameters one considers before applying for a loan. Some may prefer a short term loan while some may prefer a low cost low EMI loan. To summarize, individuals are not aware of all the rates provided by the banks they find it difficult to determine which bank has the most appropriate loan for them. By personalizing loans the users will get insights of which banks provide loans which would be best according to their requirements (at what rate, at what terms). So having a model to personalize loans according to the requirements of the users will serve them as a benefit in order to comfortably repay their loans back at the best interest rate possible considering the salary of the users.

This paper is further organized as follows: Section II includes the Literature Survey of various Loan Default prediction research papers. Section III includes the Methodology used to generate results for this research. Section IV includes the implementation phase of the project. The results of the research is observed in Section V. The paper is concluded in Section VI. Section VII describes the future work on this project.

II. LITERATURE SURVEY

Bhoomi Patel, Harshal Patil, Jovita Hembram and Shree Jaswal [1], used four data mining classification models viz. Logistic Regression, Gradient Boosting, CatBoost Classifier and Random Forest for forecasting loan defaulters. They evaluated the models based on their accuracies and the derived results proves to be that CatBoost Classifier has the highest accuracy for their dataset, along with Gradient Boosting which has an almost identical accuracy, Random Forest is the next high performing model and Logistic Regression model is the poorest performing model in terms of accuracy calculated for their dataset. The authors even calculated the precision values and F1-scores of all the models which were all almost identical, except the Logistic Regression model which was extremely low.

Authors in [2] focus on predicting loan borrowers' cred- it worthiness using the Lending Club dataset, their approach includes dataset preparation before applying the data to the models. For dataset preparation they have used SMOTE and Bragging methods to handle missing and imbalanced data, deleted unnecessary and unique features from the dataset and preprocessed the data using data discretization. This paper has used two machine learning algorithms and compared the performance of both these algorithms which are two class Decision Jungle and two class Decision Forest using the AzureML platform.

Z. Ereiz [3] has used OptiML algorithm from BigML to determine the best machine learning algorithm for the given data. The algorithms implemented and compared in this paper are Decision forest, Neural networks and Logistic Regression using which the author has calculated the precision, recall and the F1 score of each algorithm along with its accuracy and have concluded that Logistic Regression turns out to give the most accurate results compared to the other algorithms.

A. Al-qerem, G. Al-Naymat and M. Alhasan [4] focused majorly on data preprocessing and features selection algorithms. They have used three different machine learning algorithms, one with two variations which are Naive Bayes, C4.5 decision tree (unpruned & pruned) and Random Forest and calculated the precision, recall and F1 score for all algorithms. All the algorithms were run four times, in the first iteration the models used unprocessed data and the other three iterationsused processed data with three different feature selection algorithms. The three feature selection algorithms implemented in this paper are Information Gain, Particle Swarm Optimization (PSO) and Genetic Algorithm. The results compared were all the machine learning models using unprocessed data and then all the models with each of the feature selection algorithms. The results show that there has been a significant growth inthe scores when feature selection algorithms were applied. It also turns out that the C4.5 Decision tree (Pruned) algorithm gives the best results.

In [5], the authors have used machine learning algorithms with variations and combinations. They have divided their data in three parts two for training and one for testing with a ratio of 40%, 40% & 20% respectively. They have used an ensemble learning framework in which the Gradient Boosting Decision Trees algorithm is fed all the data, then the results of which are encoded using one-hot encoding which increases the dimensions of the output. In order to reduce the dimensions of the output auto-encoder is used, the result of which serves as an input for the final model: Logistic Regression, the final result is the output generated from the Logistic Regression model.

Authors in [6] have used the Lending Club dataset, they have cleaned the dataset by deleting the empty rows and columns. The clean dataset is the input to the data preprocessing model where they create dummy variables and binary output labels. Then the approach is divided into two parts in which they use RFECV for feature selection, split the training and testing data (70:30), use SMOTE for oversampling and perform Feature Scaling all these processes are applied butthe sequence is changed. The output is then applied to the classifier algorithm and then by using hold out test set and cross validation with grid search simultaneously the resultsand analysis are generated.

Data processing is done using the XGBoost model in [7], the authors have focused on handling the imbalance loan prediction dataset for which they used a hybrid under sampling method Diversified Sensitivity Under sampling (DSUS) and then compared this model to nine resampling methods.

Ajay Byanjankar in [8] proposes the application of survival analysis to predict the survival times of loans in P2P lending. To predict the survival time the Kaplan Meier estimator and the Cox Proportional-Hazards Model are used. The results of the survival tests are then added to the original dataset and were then applied to the classification models. The classification algorithms implemented are Neural Networks and Logistic Regression and the results show that the Logistic Regression model does not improve the accuracy of the Neural Network model.

Authors in [9] analyses microfinance cases of financial institutes with data mining technologies. They have built the Logistic regression model and calculated the accuracy rate, precision rate of prepayment and precision rate of default for both training and testing data. The insights from the results were that the competing risk model predicted the incompatible behaviour.

The loan default was predicted by considering a very different approaches for retrieving the SMS related to bank transactions which has important parameters related to credit score like transaction amount, account number and account balance. Andthe other data which was not being captured from transactions was retrieved from social media platforms like Educational details, Family members, Sentiment of content, Followers count, Professional background [10]. This invaded the user's privacy since all the SMS from the user's device were tracked. The other drawback is that people post pictures with other people's assets so social media can't be the correct parameter and this may result in wrong predictions. The credit score calculated for inactive users might also be wrong.

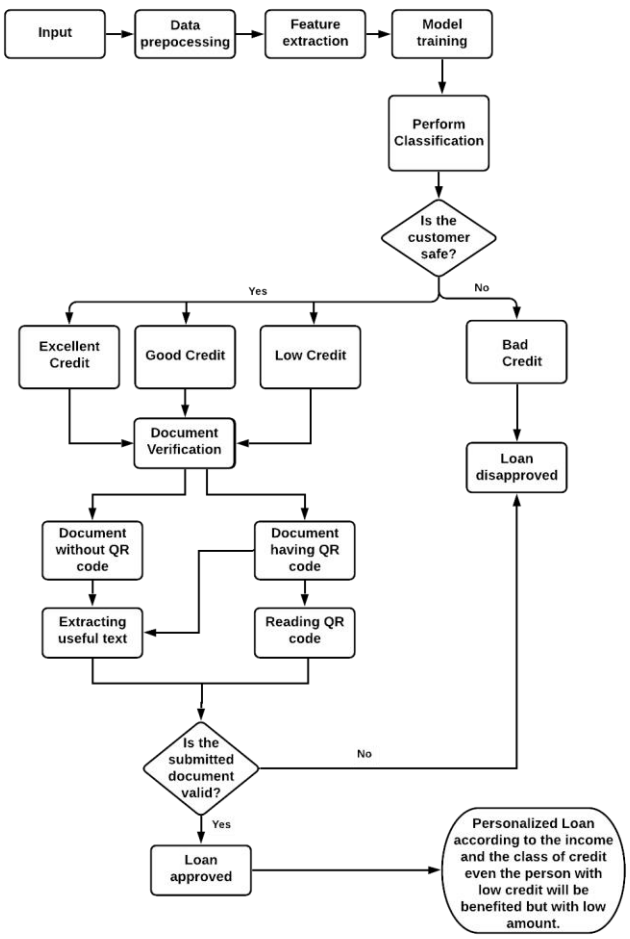

Fig. 1. Swindle Architecture.

The Naive Bayesian algorithm is used to predict the loan and classify the results into two categories: good or bad i.e. 1 or 0 [11]. When the credit scoring model is constructed, the roles of variables and their measurement levels are stated properly. There are two types of variables independent (input) or dependent (target). Here all the variables are nominal, except for the target variable and gender variable. The target variable is payment status which is denoted by 0 or 1 i.e. bad or good (binary variable). If the borrower does not default for three consecutive months in the monthly payment, then the loan is said to be good. Gender, Age, Status, ValueOfCars, ValueOfHouse, ValueOfLand and EduLevel were the variables used. But the main limitation of Naive Bayes is that it assumes all the columns as independent predictor features which is much unjustified for the real-life data so for a customer's bank related dataset it's nearly impossible, as here we have many attributes which are interdependent to one another.

Document forgery is a major issue of concern these days. People submit fake documents like academic certificates, wills, case files, birth and marriage certificates, national identity documents, insurance documents, passports, driver's licenses, etc. to create a false identity of themselves. Optical Character Recognition (OCR) is used to extract the text information from these documents. So to overcome the issue of document forgery OCR approach is implemented. Four techniques are employed; OCR, cryptographic hashing, digital signatures and 2D barcodes [12]. Tesseract was used to validate the documents.

Classification algorithm k-means was used to classify the customers into three classes which was used to predict whether the customers will default the loan or not [13]. The dataset had 73 independent attributes and 1 dependent attribute. The name of the dependent attribute is loan_status. This attribute defines three class labels such as, fully paid, current and bad. The people were classified into three income groups i.e. High, Medium and Low. But this model classified the default and no-default with accuracy 75.08%. The comparison was done by adding different levels of iterations. The iteration level 30 based K-NN model gave the best accuracy. All the existing credit score models are usually based on binary classification i.e Good and Bad or Approve and Disapprove.

One of the previous loan default prediction model used Deep Learning for a real-life data set of loan approval [14]. The paper proposed an algorithm using Deep Learning with Auto Encoders. Auto-encoders generally don't need particular labels to train on, the labels here are auto generated within the model according to the dataset. Due to the diversity of the loan applicant data, the authors believe that auto-encoders are the most fitted for this real life problem and predicts the best accuracy.

III. METHODOLOGY

A. Data Source

The dataset used for prediction is a standard Indian loan default dataset from Kaggle. The dataset was first normalized, then the unnecessary and unique columns were deleted from the dataset. At last the features not applicable for calculation and prediction of loan defaults were discarded.

B. CatBoost Algorithm

While conducting our survey machine learning algorithms such as Gradient Boosting and Decision Tree were generating the most accurate results. Hence we decided to implement CatBoost algorithm as it uses gradient boosting on decision trees and is an open source library. It is good at handling categorical features and is also faster compared to other boosting algorithms as it implements symmetric trees. With the data in the dataset changing over time CatBoost algorithm is the most competent algorithm, it works well with large datasets and has low latency requirements.

C. pytesseract

Python-tesseract is a tool for optical character recognition, as the documents required for loan approvals are scanned is requires the tool to read the text embedded in the image. We have used it as a script so that is will print the recognized text instead of writing it to a file.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 181398 entries, 0 to 181397
Data columns (total 41 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   UniqueID                              181398 non-null  int64
 1   disbursed_amount                      181398 non-null  int64
 2   asset_cost                            181398 non-null  int64
 3   ltv                                   181398 non-null  float64
 4   branch_id                             181398 non-null  int64
 5   supplier_id                           181398 non-null  int64
 6   manufacturer_id                       181398 non-null  int64
 7   Current_pincode_ID                    181397 non-null  float64
 8   Date.of.Birth                         181397 non-null  datetime64[ns]
 9   Employment.Type                       175250 non-null  object
 10  DisbursalDate                         181397 non-null  datetime64[ns]
 11  State_ID                              181397 non-null  float64
 12  Employee_code_ID                      181397 non-null  float64
 13  MobileNo_Avl_Flag                     181397 non-null  float64
 14  Aadhar_flag                           181397 non-null  float64
 15  PAN_flag                              181397 non-null  float64
 16  VoterID_flag                          181397 non-null  float64
 17  Driving_flag                          181397 non-null  float64
 18  Passport_flag                         181397 non-null  float64
 19  PERFORM_CNS.SCORE                     181397 non-null  float64
 20  PERFORM_CNS.SCORE.DESCRIPTION         181397 non-null  object
 21  PRI.NO.OF.ACCTS                       181397 non-null  float64
 22  PRI.ACTIVE.ACCTS                      181397 non-null  float64
 23  PRI.OVERDUE.ACCTS                     181397 non-null  float64
 24  PRI.CURRENT.BALANCE                   181397 non-null  float64
 25  PRI.SANCTIONED.AMOUNT                 181397 non-null  float64
 26  PRI.DISBURSED.AMOUNT                  181397 non-null  float64
 27  SEC.NO.OF.ACCTS                       181397 non-null  float64
 28  SEC.ACTIVE.ACCTS                      181397 non-null  float64
 29  SEC.OVERDUE.ACCTS                     181397 non-null  float64
 30  SEC.CURRENT.BALANCE                   181397 non-null  float64
 31  SEC.SANCTIONED.AMOUNT                 181397 non-null  float64
 32  SEC.DISBURSED.AMOUNT                  181397 non-null  float64
 33  PRIMARY.INSTAL.AMT                    181397 non-null  float64
 34  SEC.INSTAL.AMT                        181397 non-null  float64
 35  NEW.ACCTS.IN.LAST.SIX.MONTHS          181397 non-null  float64
 36  DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS   181397 non-null  float64
 37  AVERAGE.ACCT.AGE                      181397 non-null  object
 38  CREDIT.HISTORY.LENGTH                 181397 non-null  object
 39  NO.OF_INQUIRIES                       181397 non-null  float64
 40  loan_default                          181397 non-null  float64
```

Fig. 2. Attributes in loan data.

Once the loan default of an individual is predicted upon which the financial institutes will decide whether to approve an individual's loan or not. The documents are verified of an individual, by using pytesseract, if the documents turn out to be valid and authentic, the individuals will be recommended with personalised loans by the financial institutes on our platform Swindle, based on certain eligibility parameters of the individuals, they will be recommended loans as per the category they belong to.

IV. Experiments and Results

A. Loan Default Prediction

For predicting loan defaults we have implemented the CatBoost Classifier Algorithm with the proper hyper parameters that will result in higher accuracy and therefore when compared with other boosting algorithm gives the highest accuracy. The other models like Gradient Boosting Classifier, Random Forest Classifier and XGB Classifier were also used just to compare the performance of the loan forecasting model with Catboost Classifier. The model classifies the users into four categories as it is considered as a multi-class problem.

The system helps to decide whether a user should be granted a loan or not. We are able to successfully classify users into four categories namely:

1) Excellent Credit.

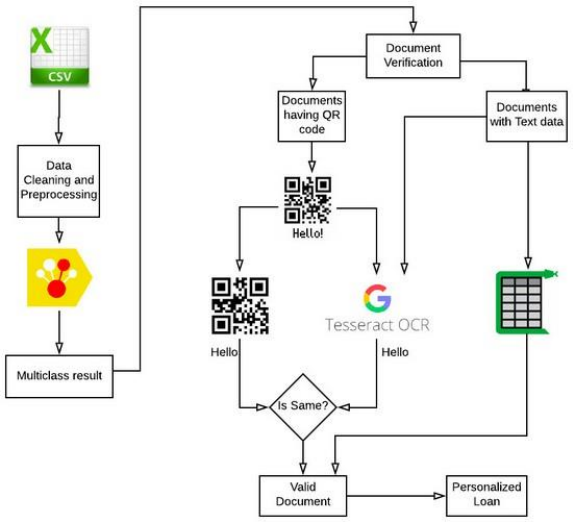

Fig. 3. System Flow Diagram.

2) Good Credit.
3) Low Credit.
4) Bad Credit.

According to the classes if the applicant belongs to the first three classes they are classified into approved applicants otherwise they are labeled as rejected applicants.

B. Document Verification

The python tool pytesseract is used for optical character recognition using this library the text was extracted from the documents uploaded by the loan applicants. Many documents which are required for applying loans are verified and validated such as

1) Aadhaar card.
2) Pan card.
3) Driving license.
4) Payment slip.
5) Bank Statement.
6) Electric Bill.
7) ITR / Form 16.
8) Marksheet.

Aadhaar card's written text was extracted and the QR code present on the card was read using pyzbar library. The two outputs were then compared and tried to match, if the name of the applicant, aadhaar number and gender matches, then the user is said to be a valid applicant. Similarly other documents like pan card, driving license were used to validate and gain more information regarding the applicant. Many important details like address, signature, type of vehicle users are also extracted. Address from various documents were compared and validated with Electric Bill. Other documents related to income for example ITR/Form 16 was used to check the annual income, last 6 months bank statement were used to check the transactions. The applicant's document is validated if the documents are found to be forged then the applicant is

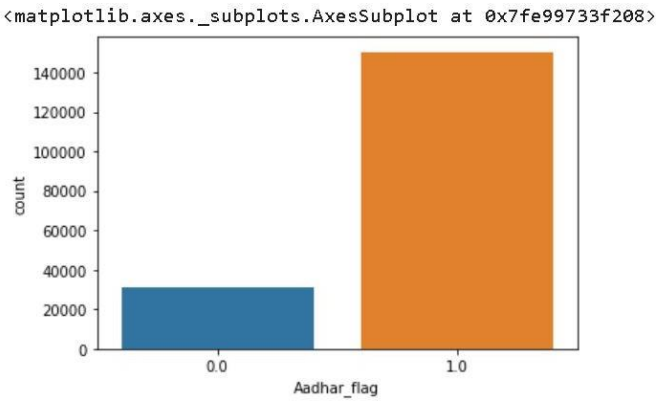

Fig. 4. Chances of defaulting when compared with Aadhaar card.

categorized into the Low class and is eliminated from the loan approval process. Payment slip was useful for extracting the salary of the applicant. Also his/her savings and deduction of previous loans were noted. The below figure is the output of the text that is extracted from an applicant's uploaded payment slip.

```
Employee Id + 11001 Name : Alex Jacob
Department : Information Technology Designation + Sr. Software Engineer
Days Worked 222.0 Bank Name, Branch = HSBC, Lagos
Bank Acet/Cheque Number 18005110026 Casual Leave = 0.0 (Op: 4.0 Cl: 4.0)
Earned Leave = 0.0 (Op: 5.0 Cl: 5.0) Overtime Hours 2 34.00
Earnings Amount | Deductions Amount
Basic Pay 1,368.00 | National Insurance 85.00
Medical Allowance 273.60 | Loss of Pay 0.00
Housing Allowance 136.80 | Loan Repayment 230.00
```

Fig. 5. Extracted text of payslip.

C. Personalizing Loans

According to the category in which the applicant falls, an unique scheme was designed for each individual. There were four categories in which the individual may be categorized and according to the category they fall under, Swindle recommends loan schemes to the individuals.

class	scheme_1	tenure_1	scheme_2	tenure_2	scheme_3	tenure_3	recommended scheme
Good	27221	2.0	18147	2.9	13610	3.9	Scheme 2
Good	28753	1.9	19168	2.9	14376	3.9	Scheme 2
Good	29088	2.2	19392	3.4	14544	4.5	Scheme 2
Excellent	25043	1.9	16695	2.8	12521	3.7	Scheme 1
Good	26578	1.9	17719	2.9	13289	3.9	Scheme 2

Fig. 6. Category wise personalized EMI.

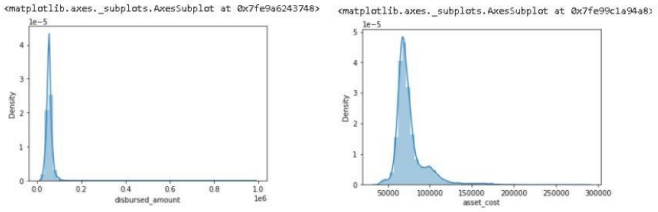

Fig. 7. Distplot of disbursed amount and asset cost.

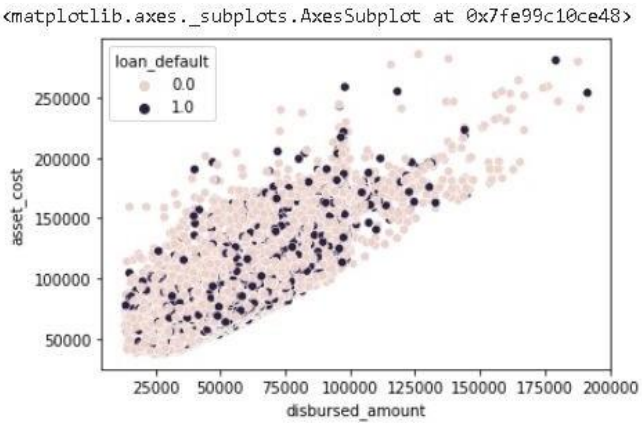

Fig. 8. Scatterplot of disbursed amount vs asset cost.

V. RESULTS

A. Case study

As observed in Table 1, above are the three cases which we will be considering.

Case 1: The "Excellent Credit" user case. In this case, considering the salary and other details of the user, the loan default probability is calculated which comes out to be 0.22107478482189. As the loan default probability is closer to 0 (that is, as the loan default probability is low), the user is classified as the "Excellent Credit" user.

Case 2: The "Good Credit" user case. In this case, considering the salary and other details of the user, the loan default probability is calculated which comes out to be 0.27064843640738. As the loan default probability is closer to 0, but more than that needed for the "Excellent" class (that is, as the loan default probability is medium), the user is classified as the "Good Credit" user.

Case 3: The "Low Credit" user case. In this case, considering the salary and other details of the user, the loan default probability is calculated which comes out to be 0.502976583014957. As the loan default probability is closer to 1 (that is, as the loan default probability is high), the user is classified as the "Low Credit" user.

As the loan default probability decreases, the chances of that applicant defaulting on the loan also decreases. So if the probability is less, the user is classified into an "Excellent class". And as the Loan default probability increases the user is classified into "Good class" and "Low class" respectively.

Cases	UniqueID	Salary	asset_cost	Employment Type	Date of Birth	Loan Default Probability
1.	763449	100172	63896	Self employed	01-06-1963	0.22107478482189
2.	655269	108887	63558	Salaried	19-01-1974	0.27064843640738
3.	748400	106663	67445	Salaried	01-12-1971	0.502976583014957

Fig. 9. Table 1.

Swindle also gives personalized loan recommendations to the approved applicants. So for an "Excellent class" user, the recommended scheme is "Scheme 1" (S1), for a "Good class" user, the recommended scheme is "Scheme 2" (S2) and for a "Low class" user, the recommended scheme is "Scheme 3" (S3).

Case	UniqueID	Class	S1's EMI	S1 Tenure period	S2's EMI	S2 Tenure period	S3's EMI	S3 Tenure period	Recommended scheme
1.	763449	Excellent	25043	1.9	16695	2.8	12521	3.7	Scheme 1
2.	655269	Good	27221	2	18147	2.9	13610	3.9	Scheme 2
3.	748400	Low	26665	2	17777	3.1	13332	4.1	Scheme 3

Case	UniqueID	Loan default probability	Class	Recommended scheme	EMI amount	Tenure period
1.	763449	0.22107478482189	Excellent	Scheme 1	25043	1.9
2.	655269	0.27064843640738	Good	Scheme 2	18147	2.9
3.	748400	0.502976583014957	Low	Scheme 3	13332	4.1

Fig. 10. End Result.

VI. CONCLUSION

In this paper, we have explored the use of CatBoost algorithm for loan default prediction. This paper has compared our algorithm with two different algorithms namely random forest and gradient boosting. CatBoost has achieved the highest accuracy amongst all other algorithms. So using the CatBoost algorithm, the loan default probability has been achieved after which the personalized loan scheme was recommended to the applicants. Also, major documents required for loan approval were verified for legality. The results in Fig. 10, Table 1 shows the loan default probability and Fig. 11 shows the final result, that is, the recommended scheme for a particular individual based on their repaying capabilities, where it will be beneficial for individuals and banks.

VII. FUTURE WORK

For further research, attributes of the applicants such as their age, medical history and the nature of their jobs can be considered in evaluating the uncertainty parameter of repaying loans, in addition potential defaults in corporate loans can be predicted for companies and startups.

Currently the loan schemes recommended to the approved applicants are designed by Swindle and are recommended to the applicants by determining the category they belong to, subsequently it can be attempted to recommend loan schemes to the approved applicants by commercial Indian and foreign banks.

REFERENCES

[1] B. Patel, H. Patil, J. Hembram and S. Jaswal, "Loan Default Forecasting using Data Mining", *2020 International Conference for Emerging Technology (INCET)*, Belgaum, India, 1-4 June 2020.

[2] K. Alshouiliy, A. AlGhamdi and D. P. Agrawal, "AzureML Based Analysis and Prediction Loan Borrowers Creditworthy", *2020 3rd International Conference on Information and Computer Technologies (ICICT)*, San Jose, CA, USA, 302-306 March 2020.

[3] Z. Ereiz, "Predicting Default Loans Using Machine Learning (OptiML)", *2019 27th Telecommunications Forum (TELFOR)*, Belgrade, Serbia, 1-4 Nov. 2019.

[4] A. Al-qerem, G. Al-Naymat and M. Alhasan, "Loan Default Prediction Model Improvement through Comprehensive Preprocessing and Features Selection", *2019 International Arab Conference on Information Technology (ACIT)*, Al Ain, United Arab Emirates, 235-240 Dec. 2019.

[5] S. Chen, Q. Wang and S. Liu, "Credit Risk Prediction in Peer-to-Peer Lending with Ensemble Learning Framework", *2019 Chinese Control And Decision Conference (CCDC)*, Nanchang, China, 4373-4377 June 2019.

[6] S. Z. H. Shoumo, M. I. M. Dhruba, S. Hossain, N. H. Ghani, H. Arif and S. Islam, "Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking", *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, Kochi, India, 2023-2028 Oct. 2019.

[7] Y. Chen, J. Zhang and W. W. Y. Ng, "Loan Default Prediction Using Diversified Sensitivity Undersampling", *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, Chengdu, 240-245 July 2018.

[8] A. Byanjankar, "Predicting credit risk in Peer-to-Peer lending with survival analysis", *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, Honolulu, HI, 1-8 Dec. 2017.

[9] Jun-Ya Zeng, Jian-Bang Lin and Tian Wang, "A new competing risks model for predicting prepayment and default using data mining", *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdu, 985-989 June 2017.

[10] J. Lohokare, R. Dani and S. Sontakke, "Automated data collection for credit score calculation based on financial transactions and social media", *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)*, Pune, 134-138 Feb. 2017.

[11] O. J. Okesola, K. O. Okokpujie, A. A. Adewale, S. N. John and O. Omoruyi, "An Improved Bank Credit Scoring Model: A Na¨ıve Bayesian Approach", *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, 228-233 Dec. 2017.

[12] N. Dlamini, S. Mthethwa and G. Barbour, "Mitigating the Challenge of Hardcopy Document Forgery", *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, Durban, 1-6 Aug. 2018.

[13] G. Arutjothi and C. Senthamarai, "Prediction of loan status in commercial bank using machine learning classifier", *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, Palladam, 416-419 Dec. 2017.

[14] Y. Abakarim, M. Lahby and A. Attioui, "Towards An Efficient Real-time Approach To Loan Credit Approval Using Deep Learning", *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*, Rabat, Morocco, 306-313 Nov. 2018.

[15] S. Yadav and S. Thakur, "Bank loan analysis using customer usage data: A big data approach using Hadoop", *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, Noida, 1-8 Aug. 2017.

[16] Zakikhani, Kimiya, Fuzhan Nasiri, and Tarek Zayed. "A failure prediction model for corrosion in gas transmission pipelines." *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability (2020): 1748006X20976802.*

[17] Vijayakumar, T. "NEURAL NETWORK ANALYSIS FOR TUMOR INVESTIGATION AND CANCER PREDICTION." *Journal of Electronics 1*, no. 02 (2019): 89-98.

[18] Muthukumar, Vignesh, and N. Bhalaji. "MOOCVERSITY-Deep Learning Based Dropout Prediction in MOOCs over Weeks." *Journal of Soft Computing Paradigm (JSCP) 2*, no. 03 (2020): 140-152.

[19] Suma, V. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." *Journal: Journal of Soft Computing Paradigm September 2020*, no. 3 (2020): 153-159.