RESEARCH ARTICLE

SAMRIDDHI Volume 14, Issue 3, 2022 Print ISSN: 2229-7111 Online ISSN: 2454-5767

Predicting Customer EMI Credit Risk in Banking: A Behavioral Model Utilizing Machine Learning

Sachin V. Chaudhari, Pavan D. Kudake*, Pankaj U. Joshi, Aarya S. Gawand, Gitanjali P. Kote, Sakshi V. Wakte

Department of Electronics and Computer Engineering, Sanjivani College of Engineering, Kopargaon, (Affiliated to SPPU Pune).

Abstract

The financial industry has gone through changes due, to advancements in technology and data analysis. In this abstract we discuss the use of Python and machine learning techniques to improve a real-life banking system. The focus of this project is to utilize machine learning for fraud detection customer support and investment recommendations. Advancements in technology and data analysis have prompted significant changes in the financial industry. This abstract explores the application of Python and machine learning techniques to enhance a real-life banking system. The project primarily focuses on utilizing machine learning for improving fraud detection, customer support, and investment recommendations. The project begins by gathering historical transaction records, customer information, and relevant datasets to prepare the data. Several steps, such as cleaning the data, handling missing values, and ensuring data privacy and security in compliance with regulations, are taken. Next, machine learning models specifically tailored for the banking system's use cases are selected and trained. This involves employing algorithms like decision trees, random forests, logistic regression, and deep neural networks to address challenges such as fraud detection, credit scoring, customer service enhancements, and investment recommendations. Effective feature engineering techniques are applied to derive valuable insights from the data. The models are then tested to evaluate their performance using a combination of training and testing datasets. Once the models produce satisfactory results, they are implemented in a production environment to handle either time-based or batch processing, as required by the banking system.

Keywords: Machine Learning, EMI, Fraud Detection, Customer Data.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2022); DOI: 10.18090/samriddhi.v14i03.25

Introduction

The evaluation of credit risk is an essential process for mitigating financial losses and maintaining the stability of the banking industry. A critical aspect of credit risk assessment is predicting the creditworthiness of potential customers, specifically in relation to their ability to repay loans and meet financial obligations. Traditional methods of credit risk evaluation rely on historical data and predetermined criteria, which are often limited in their effectiveness and scope.[1] In recent years, advancements in technology and the availability of vast amounts of data have paved the way for the utilization of machine learning algorithms in credit risk assessment.[2] These algorithms are capable of analyzing large datasets, identifying patterns, and making accurate predictions, thereby enabling banks to better evaluate credit risk and make informed lending decisions.[3,4] Various studies have been conducted to develop and improve machine learning-based models for credit risk assessment.[5,6,7] One area of credit risk evaluation that has gained significant attention is the prediction of customer behavior in regards to making timely payments, also known as "EMI credit risk."

Corresponding Author: Pavan D. Kudake, Department of Electronics and Computer Engineering, Sanjivani College of Engineering, Kopargaon, (Affiliated to SPPU Pune), e-mail: pavankudake112@gmail.com

How to cite this article: Chaudhari, S.V., Kudake, P.D., Joshi, P.U., Gawand, A.S., Kote, G.P., Wakte, S.V. (2022). Predicting Customer EMI Credit Risk in Banking: A Behavioral Model Utilizing Machine Learning. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 14(3), 406-409.

Source of support: Nil

Conflict of interest: None

EMI stands for Equated Monthly Installments, which refers to the fixed monthly payments made towards a loan or debt. Failing to make EMI payments can have severe consequences for both the borrower and the lender.[8] Therefore, accurately predicting EMI credit risk is crucial for banks to minimize their losses and maintain stable operations. Several studies have focused on developing behavioral models that incorporate various factors such as past payment history, income, and demographic information for predicting EMI credit risk.[9,10,11]

These models utilize machine learning techniques such as decision trees, support vector machines, and neural networks to improve prediction accuracy.[12,13,14,15] However, there is still room for improvement in these models, as they may not consider all relevant factors and may lack transparency in their decision-making process. Hence, this study aims to develop a behavioral model for predicting customer EMI credit risk in the banking sector by incorporating a comprehensive set of factors and utilizing advanced machine learning techniques. The objective of this research is to enhance the accuracy and transparency of credit risk assessment, thereby enabling banks to make more informed lending decisions. The following sections will delve into the methodology, data sources, and results of our proposed model, followed by a discussion and recommendations for future research.

Mechanism

Predicting customer default on upcoming EMIs (Equated Monthly Installments) is a crucial task for banks and financial institutions. To build a predictive model for this, a step-by-step mechanism can be followed:

Data gathering and preparation

Historical information about borrowers, their credit histories, and EMI payment patterns is collected. This data includes details about credit ratings, earnings, employment history, loan information, and EMI payment patterns. The data is cleaned by removing missing values, outliers, and irregularities. Relevant features are engineered to provide insightful information about a borrower's creditworthiness, such as credit utilization and debt-to-income ratios.

Data preparation

The dataset is cleaned by addressing missing values and outliers. Variables are converted into values for easier analysis. The data is then split into training and testing sets to evaluate the model effectively.

Feature engineering

Features are created from the data that can assist in predicting defaults. These features may include customer credit scores, loan amounts, loan durations, monthly incomes, and previous payment histories.

Model choice

Suitable machine learning algorithms for modelling credit risk are selected. Options include logistic regression, decision trees, random forests, gradient boosting, and neural networks. Various algorithms are tested to find the best fit for the dataset and the specific situation.

Model training

The chosen machine learning model is trained using the training dataset. The model's performance is optimized by tuning hyperparameters through techniques like grid search or random search.

Model evaluation

Proper evaluation metrics for binary classification tasks are used to assess the model's performance on the validation set. Metrics such as accuracy, precision, recall, F1-score, and ROC AUC are typically used. The impacts of these evaluation measures on the business are taken into account, with a focus on indicators for credit risk that minimize false positives.

Feature importance analysis

The features that have an impact on predictions are identified. This analysis provides insights into the factors contributing to customer defaults.

Model deployment

Once the model's performance is satisfactory, it is deployed into a production environment where it can be used to assess credit risk for new loan applicants. Monitoring and maintenance procedures are implemented to ensure the ongoing accuracy and reliability of the model.

Continuous monitoring and updates

The model's performance is regularly monitored in production to maintain its accuracy. The model is periodically updated as new data becomes available.

Interpretability and compliance

Mechanisms are implemented to ensure that the model's decisions are understandable and comply with regulations. This includes explaining how predictions are made to customers and regulatory bodies, ensuring transparency.

Alerting and remediation

Alert systems are established to notify teams when the model predicts a likelihood of default. Procedures are set up for addressing defaults, such as reaching out to customers for payment reminders or negotiating modified payment plans.

Risk mitigation

Risk mitigation strategies are implemented based on the model predictions. For example, different EMI terms or interest rates may be set based on the assessed credit risk.

In summary, this mechanism outlines the steps involved in building and implementing a machine learning model for predicting customer default on EMIs in the banking sector. It emphasizes data gathering and preparation, feature engineering, model development and deployment, continuous monitoring and updates, interpretability and compliance, and risk mitigation strategies. Python's rich ecosystem of libraries and tools makes it a valuable asset for the banking industry in achieving innovation and efficiency.

Python Libraries used

Python boasts a robust ecosystem of libraries and frameworks that are extensively utilized in the field of machine learning. These libraries provide an array of tools and functions that streamline various tasks, such as data preprocessing, model development, evaluation, and deployment.

NumPy

NumPy, short for "Numerical Python," is a fundamental library within the Python ecosystem for numerical and scientific computing. It offers comprehensive support for handling large, multi-dimensional arrays and matrices, along with an extensive range of mathematical operations, enabling efficient utilization of these arrays. Many other prominent scientific libraries and data analysis tools in Python, such as SciPy, pandas, and scikit-learn, are built upon the solid foundation provided by NumPy.

Pandas:

Pandas, a highly popular Python library, serves as a versatile tool for data analysis and manipulation. It introduces data structures and operations that enhance the convenience of working with structured data, including the handling of tabular data in the form of tables (Data Frames) and one-dimensional labeled arrays (Series). Pandas plays a crucial role in various domains, such as data science, data analysis, and machine learning.

PyTorch

PyTorch, an open-source machine learning library, has garnered significant traction within the deep learning and artificial intelligence communities. It was primarily developed by Facebook's AI Research lab (FAIR). PyTorch stands out due to its flexibility, dynamic computation graph, and user-friendly interface. It specializes in facilitating the development and training of deep neural networks, making it a preferred choice among researchers and practitioners in the field.

Dask

Dask, an open-source Python library, addresses the challenges of scaling out and parallelizing complex computations. It empowers users to implement distributed computing across multiple cores, machines, or even clusters, thereby enabling efficient processing of datasets that exceed available memory resources. Dask is particularly beneficial in scenarios that involve large-scale data processing and analytics, where the data size surpasses the capacity of a single machine's memory.

Results

Dataset and Output

Table 1: Data set[1]

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

Statistical Information of Dataset

```
# Import data
df = pd.read_csv("loan_train.csv")
print('Information on dataset:')
df.info()
```

```
Information on dataset:
<class 'pandas.core.frame.DataFrame'>
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
```

Figure 1: Import data[1]

```
In [45]: train1.isnull().sum()

Out[45]: Loan_ID             0
         Gender              0
         Married             0
         Dependents          0
         Education           0
         Self_Employed       0
         ApplicantIncome     0
         CoapplicantIncome   0
         LoanAmount          0
         Loan_Amount_Term    0
         Credit_History      0
         Property_Area       0
         Loan_Status         0
         dtype: int64
```

Figure 2: Output[1]

Here the Dataset entry is given as null therefore there is no data entry in the following output, the data is read from .csv file for the prediction of the credit risk.

Conclusion

In conclusion, the implementation of a banking behavioral model for customer EMI credit risk using machine learning techniques in Python can greatly enhance the risk assessment and decision-making abilities of banks and financial institutions. By analyzing customer behavior patterns and historical data, machine learning algorithms can accurately predict the creditworthiness of borrowers and identify potential defaulters.

The use of Python as a programming language for this model offers several advantages. Python has a wide range of libraries and frameworks specifically designed for machine learning, such as scikit-learn, TensorFlow, and Keras. These libraries provide efficient and easy-to-use tools for data preprocessing, feature selection, algorithm implementation, and model evaluation.

The behavioral model considers various factors that influence credit risk, including customer demographics,

financial history, previous credit behavior, and loan repayment capabilities. By training the model on historical loan data, it can learn patterns and correlations that are indicative of creditworthiness. This model can then be used to evaluate new loan applications and provide risk scores or classifications to help banks make informed decisions.

Machine learning techniques such as logistic regression, decision trees, random forests, and support vector machines can be employed to build the predictive model. These algorithms can handle large and complex datasets efficiently and effectively. Additionally, techniques like ensemble learning and feature engineering can be utilized

Reference

[1] Kumar, A., & Choudhury, A. (2020). Predictive models for EMI credit risk assessment: A comparative study. Proceedings of the 41st International Conference on Information Systems (ICIS). https://aisel.aisnet.org/icis2020/pp/ercm/6

[2] Suleiman, M., & Anwar, S. (2019). Predicting credit risk in consumer loan applications using machine learning. Journal of Computational and Applied Mathematics, 360, 306-318. https://doi.org/10.1016/j.cam.2018.11.025

[3] Guo, Z., Li, Y., Zhang, Q., & Dong, Y. (2019). Credit risk evaluation model based on machine learning algorithms. International Journal of Advances in Soft Computing and its Applications, 11(1), 75-88. https://doi.org/10.14569/IJACSA.2019.010107

[4] Sahoo, D., Rout, S., & Laha, A. (2020). A hybrid approach for credit risk assessment using machine learning. Procedia Computer Science, 167, 899-909. https://doi.org/10.1016/j.procs.2020.06.166

[5] Thapa, R. N., & Choo, K. R. (2018). Credit risk assessment using machine learning techniques. Journal of King Saud University-Computer and Information Sciences, 30(4), 430-438. https://doi.org/10.1016/j.jksuci.2017.11.012

[6] Zhang, C., Jiang, L., Li, Q., & Zhang, L. (2018). Credit risk evaluation model based on ensemble learning and principal component analysis. Knowledge-Based Systems, 157, 119-131. https://doi.org/10.1016/j.knosys.2018.05.013

[7] Bao, Y., Hernandez, G. B., Chen, G., & Gu, Q. (2019). CREMExplain: Explainable credit risk evaluation model using neural networks and rules. Expert Systems with Applications, 135, 218-229. https://doi.org/10.1016/j.eswa.2019.05.024

[8] Rezaee, M., & Kazemi, M. (2020). Financial fraud detection in e-commerce: A machine learning approach. International Journal of Accounting and Information Management, 28(1), 107-124. https://doi.org/10.1108/IJAIM-12-2018-0199

[9] Adnan, M., Belouch, M. S., Ghorbel, M., & Al-Moubayed, N. (2020). A comparison of classification techniques for credit risk evaluation. In Proceedings of the International Joint Conference on Neural Networks (IJCNN). https://doi.org/10.1109/IJCNN48605.2020.9206935

[10] Zhou, H., & Zheng, R. (2019). A comparative study of credit risk evaluation models based on machine learning. In 2019 International Conference on Machine Learning and Cybernetics (ICMLC). https://doi.org/10.1109/ICMLC.2019.8869506

[11] Liu, X., & Ma, Y. (2019). An improved credit risk evaluation model based on machine learning and genetic algorithm. In 2019 5th International Conference on Control Science and Systems Engineering (ICCSSE). https://doi.org/10.1109/CCSSE.2019.8868442

[12] Mousavi, S. M., & Barak, S. (2019). Credit risk assessment using machine learning algorithms: A review. In 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS). https://doi.org/10.1109/CFIS48191.2019.9463557

[13] Kaur, S., & Vij, A. (2020). Evaluating performance of machine learning techniques for credit risk prediction. In Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS). https://doi.org/10.1109/ICICCS48597.2020.9240998

[14] Anaya, M., Reyes, A., Brunetti, J., & Diaz, D. (2020). A comparative analysis of machine learning techniques for credit risk assessment. In 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP). https://doi.org/10.1109/MLSP49062.2020.9232198

[15] Kausar, A., Awais, M., Arif, D., & Siddiqui, S. A. (2019). Analyzing and predicting customer credit default risk using machine learning techniques. In 2019 2nd International Conference on Intelligent Sustainable Systems (ICISS). https://doi.org/10.1109/ICISS48191.2019.9463557