

Predicting Credit Card Default Using Machine Learning: An Empirical Analysis

Syed Aamir Aarfi^{1*}, Nashrah Ahmed², Kashif Abdullah Syed³

¹IT Senior Product Manager., Amazon.com Inc., Seattle WA, USA

²Computer Science Engineer, Integral University, Lucknow UP, India

³IT Manager, Deloitte Consulting, Harrisburg NC, USA

Abstract Credit card lending is prevalent in the modern financial landscape, enabling consumers to access credit while exposing financial institutions to inherent risks of default. Accurate prediction of credit card default is paramount for mitigating financial losses and optimizing risk management strategies. This empirical study leverages supervised machine learning techniques to develop robust predictive models for credit card default. Utilizing a comprehensive dataset encompassing customer demographic information, credit histories, and payment patterns, we evaluate the performance of various classification algorithms, with a particular emphasis on the Sequential Minimal Optimization (SMO) algorithm. Through rigorous data preparation, feature engineering, model optimization, and performance evaluation, we provide insights into the predictive capabilities of machine learning models and contribute to the ongoing research efforts in credit risk assessment.

Keywords Credit risk assessment, Predictive model, Machine learning, Data science, Artificial intelligence

1. Introduction

The proliferation of credit card lending has facilitated widespread access to credit for consumers, enabling them to finance major purchases and manage expenses effectively. However, this convenience comes with inherent risks for financial institutions, as they must carefully assess the creditworthiness of borrowers to mitigate the potential for default. Traditional credit risk assessment methods, relying on statistical techniques and expert judgment, may not fully capture the complex interplay of factors influencing default behavior. Consequently, there is a growing need for more sophisticated Information technology (IT) and data-driven approaches to accurately predict credit card default and inform risk management strategies. In recent years, machine learning algorithms have emerged as powerful tools for predictive modeling, offering the potential to uncover intricate patterns and relationships within large and complex datasets. This study aims to leverage the capabilities of supervised machine learning techniques to develop robust predictive models for credit card default. By utilizing a comprehensive dataset encompassing customer demographic information, credit histories, and payment patterns, we endeavor to identify the most effective algorithms and feature combinations for

accurately classifying customers as defaulters or non-defaulters.

2. Literature Review

The prediction of credit card default has been an active and extensively researched area in the financial domain, attracting significant attention from researchers and practitioners alike. Numerous studies have explored a wide array of statistical, empirical, and machine learning approaches to tackle this crucial problem.

Early efforts in this field employed empirical models to predict the likelihood of an individual defaulting on a new loan. Notably, Avery et al. (2003) [1] developed such models that took into account economic conditions and situational circumstances of the area in which the borrower resides. These models aimed to capture the influence of external factors on an individual's ability to repay their loans.

As the field progressed, more advanced techniques emerged, including the adoption of Bayesian methods. In a notable contribution, Xia et al. (2017) [2] utilized the Bayesian Hyper-parameter technique to optimize credit scoring models, demonstrating the potential of probabilistic approaches in improving credit risk assessment.

With the rapid advancement of machine learning techniques, a plethora of algorithms have been explored and applied to the credit card default prediction problem. Linear regression, logistic regression, random forests, discriminant

* Corresponding author:

aamiraarfi@gmail.com (Syed Aamir Aarfi)

Received: Oct. 20, 2024; Accepted: Nov. 8, 2024; Published: Nov. 12, 2024

Published online at <http://journal.sapub.org/ajis>

analysis, and decision trees are among the widely employed methods in this domain. Researchers such as Min-Chung et al. (2009) [3] and Yashna et al. (2018) [4] have contributed significantly to the understanding and application of these algorithms in the context of credit card default prediction.

Notably, the random forest algorithm has garnered particular attention for its superior performance in this domain. Ajay et al. (2016) [5] conducted a comprehensive study demonstrating the efficacy of random forest models in accurately predicting credit card default, outperforming several other traditional and machine learning techniques. However, the latest work in the field of Neural Networks (NN), by Gui, Liyu. et. al. 2019 [6] have shown that NN is the best-in-class classifier (81.64%) when predicting customer credit or loan defaults.

Building upon comparative study of I-Cheng et. al [7], and this existing body of knowledge in predictive analysis, the present study aims to further explore the performance of various non-linear classifier algorithms and compare their effectiveness against the widely-used random forest model. Through rigorous empirical analysis and model optimization, we strive to contribute new insights to the field of credit risk assessment and provide practical guidance for financial institutions in their lending practices.

By leveraging the latest advancements in machine learning techniques and drawing upon the wealth of previous research, this study seeks to advance the understanding of credit card default prediction and develop robust Artificial Intelligence (AI) models that can aid financial institutions in mitigating risks and fostering sustainable lending practices.

3. Data Collection and Preparation

The dataset employed in this empirical study is derived from real-world credit card records of customers in Taiwan. This dataset was donated to esteemed University of California, Irvine (UCI) Machine Learning Repository [8], which serves as a valuable resource for researchers and practitioners in the field of machine learning. The dataset encompasses a comprehensive set of 18 features, meticulously curated to capture various aspects of customer behavior and credit histories.

At the core of this dataset lies a binary target variable, **CLASS**, which serves as a crucial indicator, classifying customers as either defaulters or non-defaulters on their credit card payments. This variable forms the foundation for the supervised learning task, enabling the development of predictive models capable of distinguishing between these two distinct customer groups.

In addition to the target variable, the dataset incorporates a diverse array of customer demographic information, including **AGE**, **GENDER**, **EDUCATION**, and **MARRIAGE** status. These features provide valuable insights into the potential influence of socioeconomic factors on credit card default behavior, allowing for a more holistic understanding of the underlying dynamics.

Furthermore, the dataset encompasses essential financial details, such as the **LIMIT_BALANCE**, which represents the maximum amount of credit extended to each customer. This feature is particularly relevant in assessing the risk associated with lending practices and determining appropriate credit limits.

To capture the temporal aspects of customer payment behavior, the dataset includes a comprehensive record of payment and billing histories spanning a six-month period, from April to September 2005. Specifically, the dataset incorporates variables such as **Pay_AMTx** and **Bill_AMTx**, where **x** represents the corresponding month, providing a detailed account of past bill payments and outstanding balances, respectively.

Moreover, the dataset encompasses information on payment delays, represented by the **Pay_x** variables, where **x** denotes the number of months of delayed payments. These variables range from -1, indicating timely payment, to 9, signifying a nine-month delay in payment, offering valuable insights into customer payment patterns and potential delinquency risks.

The intricate details of the dataset's features are meticulously presented in Table 1, which outlines the data type and description for each variable, ensuring transparency and facilitating reproducibility in the research process.

Table 1. Data: Feature details

Field	Data type	Description
CLASS	Nominal	Default or Not
MARRIAGE	Numeric	Marital status (1 = married, 2 = Single, 3 = others)
EDUCATION	Numeric	Education (1 = graduate, 2 = university, 3 = high school, 4 = others)
AGE	Numeric	Age (year)
SEX	Numeric	Gender (1 = male, 2 = female)
LIMIT_BALANCE	Numeric	Amount of credit given
Bill_AMT1 to 6	Numeric	Bill statement (Apr.– Sep. 2005)
Pay_AMT1 to 6	Numeric	Past bill payment (Apr.– Sep. 2005)
Pay_0 to 6	Numeric	Payment delays (1-8 months delay)

To ensure a representative and robust analysis, a random sample of 10,000 instances was drawn from the original dataset. This sample was further subdivided, with 60% allocated for cross-validation, 30% for testing, and the remaining 10% reserved for development purposes. Data pre-processing steps were undertaken to ensure data quality and compatibility with the chosen machine learning algorithms. Notably, the **CLASS** feature was converted from its original form to a nominal variable, enabling its seamless integration into the classification tasks. Additionally, any missing or inconsistent values were handled appropriately to maintain data integrity.

The comprehensive nature of this dataset, encompassing a diverse range of customer attributes, credit histories, and payment behaviors, provides a fertile ground for exploring

the intricate relationships and patterns that govern credit card default behavior. By leveraging this rich dataset in conjunction with advanced machine learning techniques, this study aims to contribute valuable insights to the field of credit risk assessment, ultimately informing more effective lending practices and risk mitigation strategies within the financial industry.

4. Methodology

The study employs a supervised machine learning approach [9] [10], to predict credit card default, treating it as a binary classification problem. The methodology encompasses several key stages:

4.1. Exploratory Data Analysis (EDA)

EDA techniques, such as statistical analysis and manual examination of instance values, were performed on the development dataset to gain insights into the data distribution, identify potential patterns, and guide feature selection.

4.2. Feature Selection

Relevant features were selected based on correlation analysis, domain knowledge, and their potential impact on predicting credit card default.

4.3. Baseline Performance

Initially, various classification algorithms were evaluated using their default parameter settings on the development dataset, with 10-fold cross-validation employed to assess their baseline performance as presented in Table 2.

Table 2. Algorithm evaluation on development dataset

Algorithm	% Accuracy	Kappa
Naïve Bias (Discrete)	80	0.3968
Logistic Regression (L2)	78.5	0.3199
SVM (Normalized)	80.73	0.3294
SMO	81.88	0.255

4.4. Error Analysis and Feature Engineering

We preferred to select Sequential Minimal Optimization (SMO) algorithm, with default parameter settings, as it predicted defaulters (CLASS) with 81.88% accuracy on the development data (1000 instances). However, there were instances of false positive and false negative as per the confusion matrix in Table 3. Around 166 instances were incorrectly predicted as defaulters.

Table 3. Confusion matrix of SMO classification - baseline

Actual (CLASS)	Predicted (CLASS)	
	0	1
0	766	16
1	166	52

182 misclassified instances were analyzed to identify potential sources of error and inform feature engineering strategies, such as converting numeric features to nominal representations or creating new features through clustering or discretization. The misclassification decreased from 182 to 172, and improved the model accuracy to 82.42% (kappa 0.35) as per Table 4. Further iterations of feature engineering didn't yield any statistical significance.

Table 4. Confusion matrix of SMO classification – Error reduction

Actual (CLASS)	Predicted (CLASS)	
	0	1
0	749	31
1	141	78

4.5. Model Optimization

The top-performing algorithm, Sequential Minimal Optimization (SMO), underwent parameter tuning to optimize its performance. A three-stage manual tuning process was employed, involving the complexity parameter (C) and the exponent value within the polynomial kernel (E). Model accuracy of the tuned model with Setting 1 over 5 folds was observed to be 82.20% (Figure 1). The t-test revealed that the tuning parameter in setting 1 (C 2, E 1) isn't statistically significant (pvalue=.0.951, t =0.0162). Hence, the tuning process indicated that neither C nor E parameters improved the performance of the model.

Data	Default	Setting 1	Setting 2	Setting 3	Setting 4	Setting 5	Optimal	Test Data	
	C1 E1	C2 E1	C1 E2	C2 E2	C1 E3	C3 E1	Optimal	% Accuracy	Kappa
Full-data	82.20	82.40	79.10	79.10	74.8	82.30			
Fold1	81.38	81.25	78.63	77.25	74.63	81.12	Setting 1	84.50	0.37
Fold2	82.00	82.13	79.00	78.88	74.8	82.13	Setting 1	81.00	0.43
Fold3	83.38	83.25	79.25	78.50	75.375	83.25	Setting 1	77.00	0.17
Fold4	81.88	81.88	78.75	79.13	76.5	81.88	Setting 1	81.50	0.35
Fold5	82.40	82.40	79.10	76.60	74.8	82.30	Setting 1	87.00	0.47
Average	82.21	82.22	78.97	78.24	75.15			82.20	

Figure 1. SMO Model tuning default and customer values of C and E parameters

4.6. Final Performance Evaluation

The optimized model was trained on the combined training and development datasets and evaluated on the holdout test dataset to obtain the final performance metrics, including accuracy, kappa statistic, and statistical significance tests. Using SMO algorithm, with parameter C = 1 and E= 1, test data instances were correctly classified with 80.99% accuracy (Kappa 0.3485). While, Naïve Bayes (Discretized) correctly classified with 78.87% (Kappa 0.377). The difference between the results is highly statistically significant (p-value = 0.000, t=-5.172) in favour of SMO.

Table 5. Algorithm evaluation on holdout test dataset

Algorithm	% Accuracy	Kappa
Naïve Bias (Discrete)	78.67	0.3777
SMO	80.99	0.3485

5. Results and Discussion

The empirical analysis presented in this study demonstrates the remarkable efficacy of the Sequential Minimal Optimization (SMO) algorithm in predicting credit card default behavior. When evaluated on a holdout test dataset, the optimized SMO model exhibited an impressive accuracy of 80.99% and a substantial kappa statistic of 0.3485 in classifying credit card customers as defaulters or non-defaulters. This performance was statistically significantly better than the Naive Bayes (Discretized) model, which achieved an accuracy of 78.867% and a kappa statistic of 0.377. The superior performance of the SMO algorithm can be attributed to its ability to effectively capture the intricate patterns and non-linear relationships inherent in the dataset, which encompasses a diverse range of customer demographic information, credit histories, and payment behaviors. By leveraging the power of kernel methods and optimization techniques, the SMO algorithm demonstrated its suitability for tackling complex classification tasks in the domain of credit risk assessment.

Throughout the model development process, rigorous error analysis and feature engineering steps played a pivotal role in enhancing the predictive performance of the SMO algorithm. Notably, converting the payment delay features from numeric to nominal representations significantly reduced misclassification errors, resulting in a decrease in misclassified instances from 182 to 172 and a corresponding increase in accuracy from 81.88% to 82.42% (kappa 0.35). However, techniques such as clustering or discretization of numeric features did not yield substantial improvements in the model's predictive capabilities when applied to the SMO classifier on this specific credit card dataset. This observation suggests that the scale and distribution of the loan payment and bill due amounts, represented as numeric features, may be more informative for predicting default behavior than derived or transformed representations.

In an effort to further optimize the SMO algorithm's performance, a three-stage manual tuning process was undertaken to identify the optimal values for the complexity parameter (C) and the exponent value within the polynomial kernel (E). Interestingly, the tuning process revealed that neither the C nor the E parameter significantly improved the model's accuracy beyond the default parameter settings. The baseline performance of the SMO model with default parameter values was 82.20% on the full data, and while setting 2 (C = 2, E = 1) showed a slight improvement in accuracy to 82.40%, this difference was not statistically significant. These findings suggest that the default parameter values for the SMO algorithm were adequate for this particular credit card default prediction task, and further tuning did not yield substantial performance gains. This observation highlights the robustness of the SMO algorithm and its ability to effectively handle the complexities of the dataset without extensive parameter optimization. Though parameter tuning of the machine learning models didn't yield better results, tuning has significant advantage in various other fields such natural disaster management [10], and metabolic engineering [11].

6. Conclusions

This empirical study highlights the potential of supervised machine learning techniques, particularly the SMO algorithm, in accurately predicting credit card default based on customer demographic and credit history data. The findings contribute to the ongoing research efforts in credit risk assessment and provide practical insights for financial institutions seeking to enhance their lending practices and mitigate default risks.

While the study demonstrates promising results, it is important to acknowledge the limitations imposed by the specific dataset used, which covers a six-month period for credit card customers in Taiwan from 2005.

Future research could explore the generalizability of the proposed approach to different geographical regions, time periods, and additional credit products. Furthermore, incorporating a broader range of customer information, such as longer credit history, past defaults, loan insurance amounts, spending patterns, and other relevant factors, could potentially improve the predictive accuracy of the models.

Credit card data often suffers from class imbalance when the number of defaulters is smaller than those who don't default. Class imbalance can be addressed by techniques such as resampling, ensemble methods, cost-sensitive learning, class weight adjustment, etc. Further incorporating additional feature selection methods such as Recursive Feature Elimination (RFE) or LASSO could potentially improve model interpretability and performance.

Additionally, benchmarking advanced algorithms such as Gradient Boosting Machines or Neural Networks (NN) can help us evaluate which models work better than the Support Vector Machine (SVM) model. Machine learning (ML) techniques are a rapidly evolving field. Insights from a recent study by Nafea et al. (2024) [13] on the current state of Supervised Machine Learning (ML) and Deep Learning (DL) should help us evaluate the strengths and weaknesses of the future model selection and optimization approach for credit risk assessment.

Overall, this study serves as a foundation for continued research in the field of credit risk assessment, highlighting the value of leveraging machine learning techniques to facilitate informed decision-making and foster a more sustainable financial ecosystem.

REFERENCES

- [1] Robert B Avery, Paul S Calem, Glenn B Canner, "Consumer credit scoring: Do situational circumstances matter?", *Journal of Banking & Finance*, Volume 28, Issue 4, 2004, Pages 835-856, ISSN 0378-4266, <https://doi.org/10.1016/j.jbankfin.2003.10.009>.
- [2] Yufei Xia, Chuanzhe Liu, YuYing Li, Nana Liu, A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring, *Expert Systems with*

- Applications, Volume 78, 2017, Pages 225-241, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2017.02.017>.
- [3] Ming-Chun Tsai, Shu-Ping Lin, Ching-Chan Cheng, Yen-Ping Lin, "The consumer loan default predicting model – An application of DEA–DA and neural network, Expert Systems with Applications", Volume 36, Issue 9, 2009, Pages 11682-11690, ISSN 0957-4174 <https://doi.org/10.1016/j.eswa.2009.03.009>.
- [4] Y. Sayjadah, I. A. T. Hashem, F. Alotaibi and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, Subang Jaya, Malaysia, 2018, pp. 1-4, doi: 10.1109/ICACCAF.2018.8776802.
- [5] A, Ajay & Venkatesh, Ajay & Jacob, Shomona. (2016). Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers. *International Journal of Computer Applications*. 145. 36-41. 10.5120/ijca2016910702.
- [6] Gui, L. (2019). Application of Machine Learning Algorithms in Predicting Credit Card Default Payment. *UCLA*. ProQuest ID: Gui_ucla_0031N_17808. Merritt ID: ark:/13030/m57t2mts. Retrieved from <https://escholarship.org/uc/item/9zg7157q>.
- [7] I-Cheng Yeh, Che-hui Lien "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Systems with Applications", Volume 36, Issue 2, Part 1, 2009, Pages 2473-2480, ISSN 0957-4174 <https://doi.org/10.1016/j.eswa.2007.12.020>.
- [8] Yeh, I-Cheng. "Default of Credit Card Clients." UCI Machine Learning Repository, 2009, <https://doi.org/10.24432/C55S3H>.
- [9] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal "Data Mining Practical Machine Learning Tools and Techniques" (Fourth Edition) ISBN: 978-0-12-804291-5.
- [10] Morocco Solidarity Hackathon, "Leveraging AI for Natural Disaster Management: Takeaways From The Moroccan Earthquake," 2023. <https://arxiv.org/abs/2311.08999>.
- [11] Christopher E. Lawson, Jose Manuel Martí, Tijana Radivojevic, Sai Vamshi R. Jonnalagadda, Reinhard Gentz, Nathan J. Hillson, Sean Peisert, Joonhoon Kim, Blake A. Simmons, Christopher J. Petzold, Steven W. Singer, Aindrita Mukhopadhyay, Deepti Tanjore, Joshua G. Dunn, Hector Garcia Martin, Machine learning for metabolic engineering: A review, *Metabolic Engineering*, Volume 63, 2021, Pages 34-60, ISSN 1096 7176, <https://doi.org/10.1016/j.ymben.2020.10.005>.
- [12] Mohd Iqbal Ashraf, Advanced Techniques for Artificial Intelligence with Python, *Software Engineering*, Vol. 10 No. 1, 2023, pp. 7-10. doi: 10.5923/j.se.20231001.02.
- [13] Nafea, A. A., Alameri, S. A., Majeed, R. R., Khalaf, M. A., & AL-Ani, M. M. (2024). A Short Review on Supervised Machine Learning and Deep Learning Techniques in Computer Vision. *Babylonian Journal of Machine Learning*, 2024, 48–55. <https://doi.org/10.58496/BJML/2024/004>.