JURNAL JTIK
(JURNAL TEKNOLOGI INFORMASI DAN KOMUNIKASI)
KITA Komunitas Informasi Teknologi Aceh
OPEN ACCESS
E-ISSN: 2580-1643
Volume 11, Nomor 1, Januari 2027 | Pages: 268-278 | DOI: https://doi.org/10.35870/jtik.v11i1.7884

Comparative Evaluation of Machine Learning Models for Credit Card Approval Prediction Using Class Imbalance Handling Methods in the PyCaret AutoML Framework

Mario Suryowisnu Wicaksono [1*], Magdalena A. Ineke Pakereng [2]

[1*,2] Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana, Kota Salatiga, Provinsi Jawa Tengah, Indonesia.

Corresponding Email: [1*]marioswicaksono@gmail.com.

Article History:
Received: 6 July 2026.
Revised: 15 July 2026.
Accepted: 25 July 2026.
Available Online: 1 November 2026.
Published: 1 January 2027.

Abstract

Credit card approval processes are often performed manually, leading to inefficiencies, higher operational costs, and inconsistent decision-making. This study comparatively evaluates several machine learning models under different class imbalance handling strategies to predict credit card approval using the PyCaret AutoML framework. Two experimental scenarios were conducted on a dataset of 25,128 credit card applications, namely PyCaret's built-in SMOTE-based imbalance handling and random undersampling. In the SMOTE-based scenario, the Random Forest model achieved high accuracy, but its recall for the minority class was only 36.11%. In contrast, random undersampling produced more balanced performance, with the Gradient Boosting model achieving 100% recall. These findings indicate that class imbalance handling plays a critical role in credit card approval prediction, and that random undersampling provides more reliable minority-class detection than PyCaret's built-in SMOTE approach for the evaluated dataset.

Keywords: Credit Card Approval Prediction; Machine Learning; Class Imbalance; PyCaret; AutoML.

Abstrak

Proses persetujuan kartu kredit sering dilakukan secara manual, sehingga menimbulkan inefisiensi, meningkatkan biaya operasional, dan menghasilkan keputusan yang tidak konsisten. Penelitian ini membandingkan secara komparatif beberapa model machine learning dengan strategi penanganan class imbalance yang berbeda untuk memprediksi persetujuan kartu kredit menggunakan framework AutoML PyCaret. Dua skenario eksperimen dilakukan pada dataset 25.128 aplikasi kartu kredit, yaitu penanganan ketidakseimbangan berbasis SMOTE bawaan PyCaret dan random undersampling. Pada skenario SMOTE, model Random Forest memiliki akurasi tinggi, tetapi recall untuk kelas minoritas hanya 36,11%. Sebaliknya, random undersampling menghasilkan performa yang lebih seimbang, dengan model Gradient Boosting mencapai recall 100%. Temuan ini menunjukkan bahwa penanganan class imbalance berperan penting dalam prediksi persetujuan kartu kredit, dan random undersampling memberikan deteksi kelas minoritas yang lebih andal dibandingkan SMOTE bawaan PyCaret pada dataset yang dievaluasi.

Kata Kunci: Prediksi Persetujuan Kartu Kredit; Machine Learning; Ketidakseimbangan Kelas; PyCaret; AutoML.

This work is licensed under a Creative Commons Attribution 4.0 International License.

Copyright © 2027 by the authors. Published by Lembaga Komunitas Informasi Teknologi Aceh (KITA). All rights reserved.

E-ISSN: 2580-1643

1. Introduction

The rapid growth of digital banking services has increased the demand for efficient and reliable credit card approval systems. Credit cards provide customers with access to short-term credit for purchasing goods and services (Peiris, 2022), requiring financial institutions to evaluate applicants' financial profiles against predefined eligibility criteria before approval (Sutedja *et al.*, 2024). The increasing number of credit card applications has made manual approval processes inefficient due to the large volume of applicants (Dalsania *et al.*, 2022; Markova, 2021; Peela *et al.*, 2022). Manual verification is labor-intensive, costly, and often relies on subjective assessments by credit analysts, potentially leading to inconsistent decisions and human errors (Duan, 2020; Kumar *et al.*, 2021). Such errors may result in approving high-risk applicants or rejecting qualified customers, thereby increasing credit risk exposure and financial losses (Chongwatpol, 2018). These challenges have encouraged financial institutions to adopt machine learning techniques that support more objective, consistent, and scalable credit approval decisions. Recent advances in machine learning have significantly improved financial risk assessment, credit scoring, and credit approval systems. Various classification algorithms have been successfully applied to classify credit applicants based on demographic, financial, and behavioral characteristics (Ayari *et al.*, 2025). Among these approaches, ensemble learning methods have consistently demonstrated superior predictive performance.

Chang *et al.* (2024) reported that XGBoost achieved an accuracy of 99.40% in credit risk prediction, while Araya *et al.* (2025) highlighted the effectiveness of CatBoost for modeling complex financial data. Likewise, previous studies have shown that machine learning models can effectively support credit card approval decisions using applicant demographic, financial, and employment information (Dalsania *et al.*, 2022; Lotfi, 2024; Peela *et al.*, 2022). Despite these promising results, several challenges remain. Credit card approval datasets are typically characterized by severe class imbalance, where approved applications greatly outnumber rejected applications. Under such conditions, classification models often achieve high overall accuracy while performing poorly in identifying minority-class instances associated with higher financial risk (Haibo He & Garcia, 2009). Previous studies have shown that imbalance handling techniques, including the Synthetic Minority Oversampling Technique (SMOTE) and random undersampling, can improve minority-class detection, although their effectiveness depends on dataset characteristics and the classification algorithms employed (Araya *et al.*, 2025; Chawla *et al.*, 2002; Haixiang *et al.*, 2017). Consequently, selecting an appropriate class imbalance handling strategy remains an important issue in credit approval prediction. Another challenge concerns the complexity of machine learning development. Conventional workflows require manual model selection, hyperparameter tuning, and performance evaluation, making the development process time-consuming and resource-intensive (Hutter *et al.*, 2019). (Dalsania *et al.*, 2022; Lotfi, 2024; Peela *et al.*, 2022)To address these limitations, Automated Machine Learning (AutoML) frameworks have been introduced to automate these processes while maintaining competitive predictive performance. (Dissanayake *et al.*, 2025)(Moez Ali, 2020)Among them, PyCaret provides a low-code framework supporting data preprocessing, model comparison, hyperparameter optimization, ensemble learning, and performance evaluation through a standardized workflow. Although AutoML has demonstrated considerable potential in financial prediction tasks (Yang *et al.*, 2025), its application to credit card approval prediction remains limited. Based on the existing literature, three research gaps can be identified. First, previous studies have primarily emphasized improving predictive accuracy or comparing classification algorithms, whereas limited attention has been given to systematically evaluating different class imbalance handling strategies in credit card approval prediction.

Second, although class imbalance has been recognized as a critical issue in financial classification, its influence on machine learning model performance under different balancing strategies remains insufficiently investigated. Third, despite the growing popularity of AutoML, few studies have integrated AutoML frameworks with class imbalance analysis within a unified credit card approval prediction workflow. To address these gaps, this study presents a comparative evaluation of machine learning models for credit card approval prediction using the PyCaret AutoML framework under different class imbalance handling strategies. Two experimental scenarios were designed using PyCaret's built-in SMOTE-based imbalance handling mechanism and random undersampling to evaluate their influence on classification performance. This study contributes by systematically comparing machine learning models under different imbalance conditions, evaluating the effectiveness of the PyCaret AutoML framework for automated model development, and identifying the most influential variables affecting credit approval decisions through feature

E-ISSN: 2580-1643

importance analysis. The findings provide both theoretical insights and practical guidance for developing more reliable and efficient machine learning-based credit approval decision support systems.

2. Research Methodology

This study employed a machine learning-based classification approach to evaluate credit card approval prediction using the PyCaret Automated Machine Learning (AutoML) framework. The methodology consisted of four phases: data acquisition and preparation, SMOTE-based classification, random undersampling-based classification, and analysis and evaluation. The overall research workflow is illustrated in Figure 1.



Figure 1. Research workflow of the proposed study

Data Acquisition and Preparation

The dataset was obtained from Kaggle under the title *Credit Card Approval Prediction (Cleaned Version)*. It contains 25,128 credit card application records with 20 predictor variables and one target variable (Status), representing approved and rejected applications. The predictor variables consist of numerical and categorical attributes describing applicants' demographic and financial characteristics. Data preparation included dataset inspection, exploratory data analysis (EDA), and class distribution analysis to identify class imbalance, which served as the basis for designing the two experimental scenarios.

SMOTE-Based Classification Experiment

The first experimental scenario evaluated machine learning models using the original imbalanced dataset. Class imbalance was handled using PyCaret's built-in imbalance handling mechanism by enabling the fix_imbalance=True parameter, which automatically applies the Synthetic Minority Oversampling Technique (SMOTE) during model training. The dataset was split into training and testing subsets using a 70:30 ratio. PyCaret automatically performed data preprocessing, while SMOTE was applied only to the training data, leaving the holdout set unchanged for unbiased model evaluation. Model development employed five-fold Stratified K-Fold cross-validation. Models were compared using compare_models() with Recall as the primary optimization metric because correctly identifying high-risk applicants was prioritized over maximizing overall accuracy. The selected model was then optimized using tune_model() and evaluated on the holdout dataset using predict_model(). The resulting performance served as the baseline for comparison with the random undersampling experiment.

Random Undersampling-Based Classification Experiment

The second experimental scenario evaluated machine learning models using a balanced dataset generated through random undersampling by reducing the majority class until both target classes contained an equal number of samples. To ensure a fair comparison, the same workflow as the first experiment was applied. The balanced dataset was split into training and testing subsets using a 70:30 ratio and preprocessed automatically by PyCaret. Since the dataset had already been balanced, no additional imbalance handling was applied during model training. Model development employed three-fold Stratified K-Fold cross-validation. Models were compared using compare_models() with Recall as the primary optimization metric. The selected model was optimized using tune_model() and evaluated on the holdout dataset using predict_model(). The resulting performance was compared with the SMOTE-based experiment to evaluate the effectiveness of both class imbalance handling strategies.

Analysis and Evaluation

After completing both experimental scenarios, the resulting models were evaluated through comparative performance, feature importance, and misclassification analyses. Comparative performance analysis emphasized Recall, Cohen's Kappa, and Matthews Correlation Coefficient (MCC) to assess classification performance under imbalanced data conditions. Feature importance analysis identified the variables contributing most to the prediction results, while misclassification analysis examined prediction errors to better understand the limitations of the proposed models.

Tools and Technologies

All experiments were conducted using Python 3.11.9 and PyCaret 3.3.2 within the Jupyter Notebook environment. PyCaret was employed as the AutoML framework to automate data preprocessing, model comparison, hyperparameter optimization, and performance evaluation. The hardware and software specifications used in this study are summarized in Table 1.

E-ISSN: 2580-1643

Table 1. Experimental Specifications

Name	Specifications
OS	Windows 11 Home Single Language 25H2 (64-bit)
CPU	Intel® Core™ Ultra 7 (3.80 GHz)
GPU	Intel® Arc™ Graphics (128 MB)
RAM	LPDDR5 (16 GB)

Evaluation Metrics

Model performance was evaluated using Accuracy, Area Under the Receiver Operating Characteristic Curve (AUC), Precision, Recall, F1-score, Cohen's Kappa, and Matthews Correlation Coefficient (MCC). The mathematical formulations of Accuracy, Precision, Recall, F1-score, Cohen's Kappa, and MCC are presented in Equations (1)–(6). Because this study compared different class imbalance handling strategies, multiple evaluation metrics were employed to provide a more comprehensive assessment than overall accuracy alone.

Accuracy measures the proportion of correctly classified observations:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

AUC measures the model's ability to distinguish between positive and negative classes across different classification thresholds, where higher values indicate better discriminative performance.

Precision measures the proportion of positive predictions that are correctly classified:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall measures the model's ability to correctly identify positive observations:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-score represents the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Cohen's Kappa measures the agreement between predicted and actual classifications after accounting for agreement expected by chance:

$$\kappa = \frac{Po-Pe}{1-Pe} \quad (5)$$

where Po is the observed agreement:

$$Po = \frac{TP+TN}{TP+TN+FP+FN}$$

and Pe is the expected agreement:

$$Pe = \frac{(TP+FP)(TP+FN)+(FN+TN)(FP+TN)}{(TP+TN+FP+FN)^2}$$

Matthews Correlation Coefficient (MCC) evaluates the overall quality of binary classification by considering all elements of the confusion matrix:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (6)$$

Recall was selected as the primary optimization criterion because minimizing false-negative predictions is particularly important in credit approval, where incorrectly approving high-risk applicants may increase financial risk. In addition, Cohen's Kappa and MCC were used to evaluate model reliability under imbalanced data conditions, providing a more balanced assessment than accuracy alone.

3. Hasil dan Pembahasan

3.1 Hasil

Class Distribution Analysis

Prior to model development, the distribution of the target variable was analyzed to assess the severity of class imbalance. As shown in Figure 2, the dataset was highly imbalanced, consisting of 25,007 approved applications (99.52%) and only 121 rejected applications (0.48%).

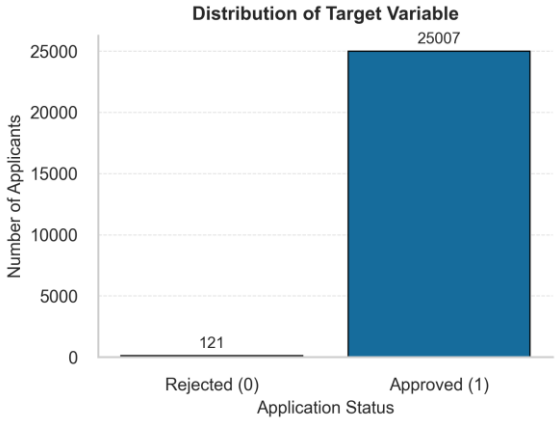


Figure 2. Class Distribution of the Original Dataset

E-ISSN: 2580-1643

Such an extreme imbalance can bias machine learning models toward the majority class, resulting in high overall accuracy while failing to correctly identify minority-class observations. In credit card approval prediction, this limitation is particularly important because misclassifying rejected applicants as approved may increase financial risk. Therefore, model performance should be evaluated using metrics beyond accuracy, including Recall, Cohen's Kappa, and Matthews Correlation Coefficient (MCC). Based on this class distribution, two experimental scenarios were designed: PyCaret's built-in SMOTE-based imbalance handling and random undersampling. Comparing these approaches enabled a systematic evaluation of their effectiveness in handling highly imbalanced credit card approval data.

SMOTE-Based Classification Results

The first experimental scenario evaluated machine learning models using the original imbalanced dataset with PyCaret's built-in SMOTE-based imbalance handling mechanism. As shown in Table 2, Random Forest was selected as the best-performing model based on the predefined optimization criterion (Recall), achieving an average cross-validation accuracy of 99.64% and an AUC of 0.9983. Other ensemble models, including Gradient Boosting, AdaBoost, and LightGBM, also demonstrated competitive performance, confirming the effectiveness of ensemble learning for credit card approval prediction.

Table 2. Performance Comparison on the Original Dataset

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
rf	Random Forest Classifier	0.9964	0.9983	1.0000	0.9964	0.9982	0.4092	0.5071
et	Extra Trees Classifier	0.9959	0.9929	0.9995	0.9964	0.9979	0.3669	0.4266
lightgbm	Light Gradient Boosting Machine	0.9992	0.9999	0.9994	0.9998	0.9996	0.9197	0.9211
ada	Ada Boost Classifier	0.9990	0.9992	0.9993	0.9997	0.9995	0.9057	0.9073
gbc	Gradient Boosting Classifier	0.9991	0.9998	0.9991	0.9999	0.9995	0.9160	0.9196

Based on the comparison results, Random Forest was selected for further evaluation because Recall was prioritized to minimize the misclassification of high-risk applicants. Hyperparameter tuning did not significantly improve performance; therefore, the original model was evaluated on the holdout dataset, as presented in Table 3.

Table 3. Holdout Performance on the Original Dataset

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
rf	Random Forest Classifier	0.9968	0.9976	0.9999	0.9969	0.9984	0.5187	0.5780

The holdout evaluation showed that the Random Forest model achieved an accuracy of 99.68%, an AUC of 0.9976, and an F1-score of 0.9984. However, the confusion matrix (Figure 3) revealed that only 13 of 36 rejected applications were correctly classified, while 23 were incorrectly predicted as approved, resulting in a Recall of only 36.11% for the Rejected class.

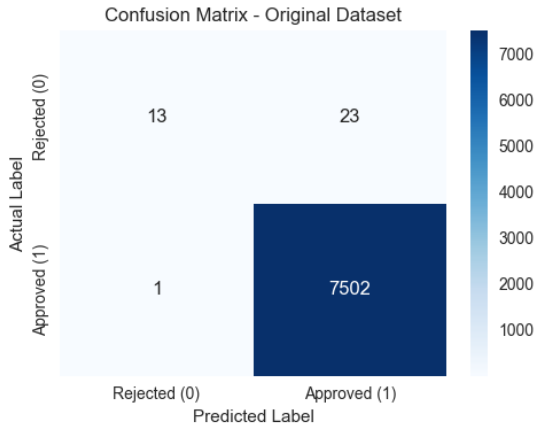


Figure 3. Confusion Matrix of the Original Dataset

E-ISSN: 2580-1643

These results indicate that, despite its high overall accuracy, the model remained strongly biased toward the majority class. The moderate Cohen's Kappa (0.5187) and Matthews Correlation Coefficient (MCC) (0.5780) further confirmed that the classification performance was not well balanced across both classes, indicating that PyCaret's built-in SMOTE-based imbalance handling was insufficient to address the severe class imbalance in the evaluated dataset. These findings are consistent with previous studies reporting the strong predictive capability of ensemble models such as Random Forest (Araya *et al.*, 2025; Chang *et al.*, 2024). However, unlike previous studies that primarily emphasized overall predictive accuracy, the present study demonstrates that high accuracy alone does not necessarily translate into effective minority-class detection under severe class imbalance. This limitation motivated the evaluation of random undersampling in the second experimental scenario.

Random Undersampling-Based Classification Results

The second experimental scenario evaluated machine learning models using a balanced dataset generated through random undersampling. The majority class was reduced until both classes contained 121 observations, producing a balanced dataset as shown in Figure 4.

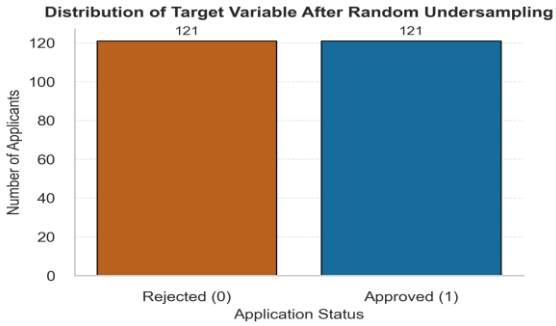


Figure 4. Class Distribution of the Balanced Dataset

Using the same evaluation workflow as the first experiment, Gradient Boosting was identified as the best-performing model. As presented in Table 4, it achieved an average cross-validation accuracy of 97.63% with an AUC of 0.9858, while LightGBM, Random Forest, and AdaBoost also demonstrated competitive performance.

Table 4. Performance Comparison on the Balanced Dataset

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
gbc	Gradient Boosting Classifier	0.9763	0.9858	0.9528	1.0000	0.9757	0.9526	0.9538
lightgbm	Light Gradient Boosting Machine	0.9763	0.9983	0.9528	1.0000	0.9757	0.9526	0.9538
rf	Random Forest Classifier	0.9588	0.9950	0.9417	0.9757	0.9582	0.9175	0.9185
ada	Ada Boost Classifier	0.9470	0.9920	0.9413	0.9543	0.9475	0.8938	0.8944
dt	Decision Tree Classifier	0.9350	0.9349	0.9294	0.9429	0.9356	0.8700	0.8710

Gradient Boosting was selected for holdout evaluation because it consistently achieved the best overall performance. Hyperparameter tuning did not significantly improve the model; therefore, the original model was retained. The holdout evaluation results are presented in Table 5.

Table 5. Holdout Performance on the Balanced Dataset

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
gbc	Gradient Boosting Classifier	0.9863	0.9857	1.0000	0.9730	0.9863	0.9726	0.9730

On the holdout dataset, the Gradient Boosting model achieved an accuracy of 98.63%, an AUC of 0.9857, 100% Recall, 97.30% Precision, a Cohen's Kappa of 0.9726, and an MCC of 0.9730. Although its overall accuracy was slightly lower than that of the SMOTE-based experiment, the model demonstrated substantially more equitable classification performance.

E-ISSN: 2580-1643

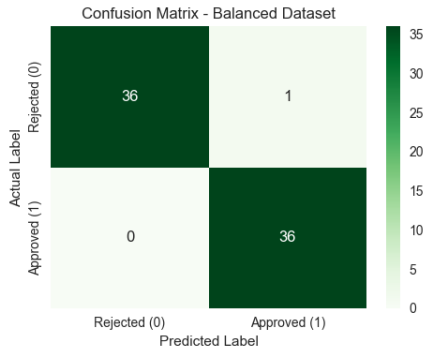


Figure 5. Confusion Matrix of the Balanced Dataset

The confusion matrix showed that all 36 rejected applications were correctly classified, while only one approved application was misclassified, resulting in perfect Recall for the minority class. This improvement indicates that balancing the dataset before training effectively reduced the model's bias toward the majority class, enabling the learning algorithm to establish more representative decision boundaries for both classes without substantially sacrificing overall predictive performance. These findings are consistent with previous studies highlighting the effectiveness of class balancing techniques for highly imbalanced financial datasets (Araya *et al.*, 2025). However, the present study further demonstrates that different imbalance handling strategies can produce substantially different outcomes even under the same AutoML workflow. For the evaluated dataset, random undersampling provided more balanced classification performance than PyCaret's built-in SMOTE-based imbalance handling, although the relative effectiveness of class balancing methods may vary depending on dataset characteristics.

Comparative Analysis of Experimental Results

The comparative evaluation demonstrates that different class imbalance handling strategies produced substantially different classification performance despite using the same AutoML workflow. Since both experiments employed identical preprocessing, model comparison, hyperparameter optimization, and evaluation procedures, the observed performance differences can be attributed primarily to the class imbalance handling strategy.

Table 6. Performance Comparison Between Experimental Scenarios

Metric	SMOTE-Based Classification Results	Random Undersampling-Based Classification Results
Best Model	Random Forest	Gradient Boosting
Accuracy (Holdout)	0.9968	0.9863
Recall (Rejected Class)	0.611	1.000
Precision	0.9969	0.9730
F1-score	0.9984	0.9863
Cohen's Kappa	0.5187	0.9726
MCC	0.5780	0.9730

As summarized in Table 6, the SMOTE-based experiment achieved slightly higher overall accuracy (99.68%), whereas the random undersampling approach substantially improved minority-class detection, increasing Recall for the Rejected class from 36.11% to 100%. Similar improvements were observed in Cohen's Kappa and Matthews Correlation Coefficient (MCC), indicating more balanced classification performance across both target classes. These findings demonstrate that class imbalance handling plays a critical role in credit card approval prediction. Although the SMOTE-based approach produced excellent overall accuracy, the model remained biased toward the majority class. In contrast, random undersampling enabled the model to learn more representative decision boundaries, resulting in substantially improved minority-class detection without a considerable reduction in overall predictive performance. Unlike previous studies that primarily emphasized predictive accuracy (Dalsania *et al.*, 2022; Lotfi, 2024; Peela *et al.*, 2022), this study evaluated the influence of different class imbalance handling strategies using complementary evaluation metrics. The results support previous findings on the effectiveness of ensemble learning for financial prediction (Araya *et al.*, 2025; Chang *et al.*, 2024), while further demonstrating that Recall, Cohen's Kappa, and MCC provide a

E-ISSN: 2580-1643

more reliable assessment than accuracy alone under severe class imbalance. For the evaluated dataset, random undersampling produced more balanced classification performance than PyCaret's built-in SMOTE-based imbalance handling, although the effectiveness of balancing techniques may vary across datasets. Beyond comparing balancing strategies, this study also demonstrates the practical applicability of the PyCaret AutoML framework for credit card approval prediction. Consistent with Yang *et al.* (2025), the results show that AutoML can simplify preprocessing, model comparison, hyperparameter optimization, and evaluation while maintaining competitive predictive performance. From a practical perspective, these findings suggest that selecting an appropriate class imbalance handling strategy is as important as selecting the classification algorithm, providing financial institutions with a more reliable and efficient approach to credit approval decision support.

Feature Importance Analysis

Feature importance analysis was performed using the final Gradient Boosting model to identify the variables that contributed most significantly to credit card approval prediction. The resulting feature importance is presented in Figure 6.

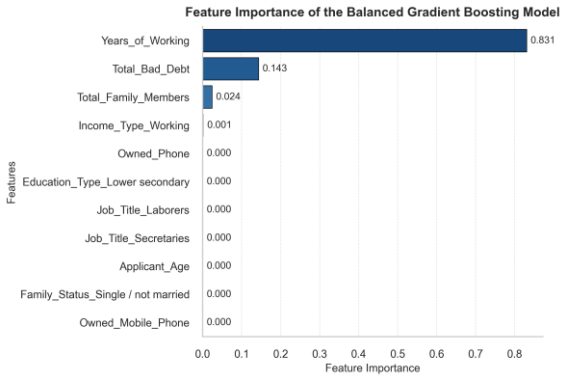


Figure 6. Feature Importance of the Final Model

The analysis showed that Years_of_Working was the most influential predictor, contributing 83.06% of the total feature importance. This finding suggests that employment stability played a dominant role in the classification process. In credit risk assessment, longer employment histories are commonly associated with stable income and lower default risk, making employment duration an important indicator of creditworthiness (Menzo *et al.*, 2024). The second most influential variable was Total_Bad_Debt, contributing 14.35% of the overall feature importance. This result is consistent with conventional credit evaluation practices, where debt-related indicators are widely used to assess applicants' financial risk (World Bank, 2019). The remaining variables contributed relatively little to the prediction, indicating that the model relied primarily on employment stability and debt-related information. While this behavior is consistent with established credit assessment practices, the strong dominance of Years_of_Working may reduce the model's sensitivity to other relevant financial risk factors, suggesting that future studies should incorporate more diverse predictive features.

Misclassification Analysis

To further evaluate the limitations of the proposed model, a misclassification analysis was conducted on the only incorrectly classified observation in the balanced dataset experiment. The characteristics of the misclassified sample are presented in Table 7.

Table 7. Characteristics of the Misclassified Sample

Variable	Value
Years_of_Working	15
Total_Bad_Debt	22
Total_Family_Members	2
Actual Status	Rejected (0)
Predicted Status	Approved (1)
Prediction Score	1.0

The applicant was predicted as Approved, whereas the actual class was Rejected. One possible explanation is the applicant's Years_of_Working value of 15 years, which strongly favored approval because employment duration was identified as the most influential predictor in the feature importance analysis. However, the applicant also had a relatively high Total_Bad_Debt value of 22, indicating elevated financial risk. These conflicting characteristics likely made the decision boundary more difficult for the model to distinguish. Despite this isolated error, the model correctly classified 71 of 72 holdout observations, indicating strong generalization performance. Nevertheless, this case suggests that dominant predictor

E-ISSN: 2580-1643

variables may outweigh other relevant risk indicators in borderline cases. Incorporating additional financial attributes or more advanced modeling techniques may further improve prediction reliability for such applicants.

3.2 Discussion

The results show that class imbalance handling strategically shapes the performance of machine learning models for credit card approval prediction. Under the original imbalanced dataset, the SMOTE-based scenario produced a Random Forest model with a holdout accuracy of 99.68% and an F1-score of 0.9984. However, the model correctly identified only 13 of 36 rejected applications, yielding a Recall of 36.11% for the Rejected class. This pattern confirms that high overall accuracy can mask poor minority-class detection under severe imbalance, a concern widely noted in the literature (Haibo He & Garcia, 2009). Consequently, accuracy alone is an insufficient criterion for evaluating credit approval models. In contrast, random undersampling produced a Gradient Boosting model that correctly classified all 36 rejected applications, achieving a Recall of 100% and an MCC of 0.9730. Although its holdout accuracy (98.63%) was slightly lower than that of the SMOTE-based model, its Cohen's Kappa (0.9726) indicates substantially more balanced agreement across both classes. This finding is consistent with Araya *et al.* (2025), who reported that class balancing techniques can improve the detection of minority-class instances in highly imbalanced financial datasets. Likewise, the superior performance of Gradient Boosting aligns with the ensemble-learning results reported by Chang *et al.* (2024), who found that boosting-based models achieve strong predictive performance in credit risk prediction.

The difference between the two strategies also reflects the mechanisms of the techniques themselves. SMOTE generates synthetic minority samples through interpolation (Chawla *et al.*, 2002), which may not fully capture the decision boundary of the rejected class when the minority class contains only 121 original observations. Random undersampling, by contrast, reduces the majority class to create a balanced training set, which allows the learning algorithm to focus on both classes equally. Haixiang *et al.* (2017) emphasized that the effectiveness of resampling methods depends on dataset characteristics and the classifier used, which explains why the relative advantage observed here may not generalize to all datasets. Accordingly, the present findings should be interpreted as specific to the evaluated dataset rather than as universal evidence that undersampling outperforms SMOTE. The feature importance analysis provides further insight into the model's decision logic. Years_of_Working contributed 83.06% of the total feature importance, and Total_Bad_Debt contributed 14.35%. The dominance of employment duration is consistent with credit risk assessment practices, in which longer employment histories are associated with income stability and lower default risk (Menzo *et al.*, 2024). Debt-related indicators also reflect conventional credit evaluation practices (World Bank, 2019). However, the strong dependence on a single predictor may limit the model's sensitivity to other relevant financial risk factors. This concern is illustrated by the only misclassified observation in the balanced experiment: an applicant with 15 years of employment but a Total_Bad_Debt value of 22 was predicted as approved, despite being actually rejected. This case suggests that dominant predictors can outweigh other risk indicators in borderline cases. From a practical perspective, these findings highlight that selecting an appropriate imbalance handling strategy is as important as selecting the classification algorithm itself. The PyCaret AutoML framework simplified preprocessing, model comparison, and hyperparameter optimization, consistent with Yang *et al.* (2025), and enabled a controlled comparison between the two strategies under identical workflows. Nevertheless, the study is limited to a single dataset with a small number of minority-class observations in the holdout set (36 rejected applications). Future research should validate these findings on larger and more diverse credit datasets and incorporate additional financial attributes to improve prediction reliability for borderline applicants.

4. Conclusion

This study demonstrates that class imbalance handling plays a critical role in machine learning-based credit card approval prediction. For the evaluated dataset, random undersampling produced more balanced classification performance than PyCaret's built-in SMOTE-based imbalance handling by substantially improving minority-class detection while maintaining strong overall predictive performance. These findings highlight the importance of evaluating imbalanced classification models using complementary metrics beyond accuracy. The study also demonstrates the practical applicability of the PyCaret AutoML framework as an efficient and reproducible approach for developing credit card approval prediction models, providing practical value for researchers and financial

institutions implementing machine learning-based decision support systems. Nevertheless, the findings are limited to a single publicly available dataset with a relatively small minority class, while the random undersampling strategy may discard informative majority-class observations. Future research should investigate alternative class imbalance handling techniques using larger and more diverse real-world datasets to improve model robustness and generalizability.

5. Acknowledgement

The author used ChatGPT (OpenAI) as an AI-assisted writing tool for English language editing and manuscript refinement. The AI was used solely for language improvement and writing assistance. All research design, data collection, data preprocessing, model development, experimental execution, result interpretation, and the final manuscript were conducted and validated by the author, who assumes full responsibility for the accuracy, originality, and integrity of this work.

6. References

Ali, M. (2020). PyCaret: An open source, low-code machine learning library in Python. *PyCaret version*, *2*, 514.

Araya, F. T., Assefa, B. G., Mohapatra, S. K., & Nanda, C. B. (2025). Sustainable credit risk prediction: Applying machine learning for responsible lending. *Journal of Transformative Technologies and Sustainable Development, 9*(1), 12. https://doi.org/10.1007/s41314-025-00080-7.

Ayari, H., Guetari, R., & Kraïem, N. (2025). Machine learning powered financial credit scoring: A systematic literature review. *Artificial Intelligence Review, 59*(1), 13.

Chang, V., Sivakulasingam, S., Wang, H., Wong, S. T., Ganatra, M. A., & Luo, J. (2024). Credit risk prediction using machine learning and deep learning: A study on credit card customers. *Risks, 12*(11), 174.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chongwatpol, J. (2018). Credit card applications pending: Who are our best prospect cardholders? Improved decisions through business analytics and business intelligence. *Journal of Information Technology Teaching Cases, 8*(1), 29–44.

Dalsania, N., Punatar, D., & Kothari, D. (2022). Credit card approval prediction using classification algorithms. *International Journal for Research in Applied Science and Engineering Technology, 10*(11), 507–514.

Dissanayake, D., Navarathna, R., Ekanayake, P., & Rajadurai, S. (2025). A survey of evaluating AutoML and automated feature engineering tools in modern data science. In *Proceedings of the 27th International Conference on Enterprise Information Systems* (pp. 218–225).

Duan, L. (2020). Performance evaluation and practical use of supervised data mining algorithms for credit card approval. In *2020 International Conference on Computing and Data Science (CDS)* (pp. 251–254). https://doi.org/10.1109/CDS49703.2020.00057.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73*, 220–239.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2019). *Automated machine learning*. Springer International Publishing.

Kumar, A., Sharma, S., & Mahdavi, M. (2021). Machine learning (ML) technologies for digital credit scoring in rural finance: A literature review. *Risks, 9*(11), 192.

Lotfi, E. (2024). Predicting credit card approval using machine learning techniques. *International Journal of Applied Data Science in Engineering and Health, 1*(1).

Markova, M. (2021). Credit card approval model: An application of deep neural networks. In *Seventh International Conference New Trends in the Applications of Differential Equations in Sciences (NTADES 2020)*.

Menzo, N., Prince, B., Morge, D., Yeboah, A., & Samuel. (2024). *Beyond income: The complexities*

E-ISSN: 2580-1643

of credit risk in developing countries (Document No. 122365).

Peela, H. V., Gupta, T., Rathod, N., Bose, T., & Sharma, N. (2022). Prediction of credit card approval. *International Journal of Soft Computing and Engineering, 11*(2), 1–6.

Peiris, M. P. C. (2022). *Credit card approval prediction by using machine learning techniques* [Thesis, University of Colombo].

Sutedja, I., Lim, J., Setiawan, E., & Adiputra, F. R. (2024). Credit card approval prediction: A systematic literature review. *Journal of Theoretical and Applied Information Technology, 15*(3).

World Bank. (2019). *Credit scoring approaches guidelines*.

Yang, F., Qiao, Y., Qi, Y., Bo, J., & Wang, X. (2025). BACS: Blockchain and AutoML-based technology for efficient credit scoring classification. *Annals of Operations Research, 345*(2–3), 703–723.