BÁO CÁO CHẤM ĐIỂM TÍN DỤNG CỦA KHÁCH HÀNG

# GIỚI THIỆU

**Phân Loại Điểm Tín Dụng: Đánh Giá Khả Năng Tài Chính và Rủi Ro**

Trong thế giới tài chính, điểm tín dụng là một yếu tố quan trọng quyết định khả năng một người hoặc tổ chức có thể tiếp cận các sản phẩm và dịch vụ tài chính hay không. Điểm tín dụng là một con số hoặc hệ thống được sử dụng để đo lường khả năng trả nợ của một cá nhân hoặc tổ chức. Dựa vào điểm này, ngân hàng và tổ chức tài chính có thể đưa ra quyết định về việc cho vay tiền, phát hành thẻ tín dụng, hay cung cấp các sản phẩm tài chính khác.

***1. Điểm Tín Dụng là gì?***

Điểm tín dụng là một con số hoặc hệ thống được tính toán dựa trên lịch sử tín dụng và các yếu tố khác để đánh giá khả năng trả nợ của một cá nhân hoặc tổ chức. Hệ thống này thường dựa trên thông tin từ các báo cáo tín dụng, như báo cáo tín dụng của FICO (Fair Isaac Corporation) ở Hoa Kỳ hoặc các cơ quan tín dụng tương tự ở các quốc gia khác. Xếp hạng được sử dụng để đo đạc tính cách của một người hoặc tổ chức trong việc quản lý và trả nợ. Điểm tín dụng thường được các tổ chức tín dụng, như ngân hàng hoặc công ty thẻ tín dụng, sử dụng để đánh giá khả năng của một cá nhân hoặc doanh nghiệp trong việc trả nợ và quản lý tài chính. Các yếu tố như lịch sử trả nợ, tỷ lệ nợ trên thu nhập, và số lượng tín dụng mở cùng với các yếu tố khác được sử dụng để tính toán điểm tín dụng của một người hoặc tổ chức. Điểm tín dụng cao thường đi kèm với nhiều lợi ích như có thể được vay tiền với lãi suất thấp hơn hoặc dễ dàng hơn.

***2. Mục đích của Điểm Tín Dụng:***

**Giúp các tổ chức tài chính đánh giá rủi ro cho vay:** Điểm tín dụng là một thước đo khách quan giúp các tổ chức tài chính đánh giá khả năng trả nợ của bạn. Dựa vào điểm tín dụng, họ có thể quyết định có cho bạn vay hay không, cũng như mức lãi suất áp dụng.

* **Bảo vệ người cho vay:** Điểm tín dụng giúp bảo vệ người cho vay khỏi những khoản vay rủi ro cao. Khi cho vay, người cho vay muốn đảm bảo rằng họ sẽ nhận lại được tiền của mình. Điểm tín dụng giúp họ đánh giá khả năng trả nợ của người vay và giảm thiểu rủi ro cho vay.
* **Thúc đẩy sử dụng tín dụng có trách nhiệm:** Điểm tín dụng khuyến khích người vay sử dụng tín dụng một cách có trách nhiệm. Khi biết rằng điểm tín dụng của mình sẽ ảnh hưởng đến khả năng tiếp cận các dịch vụ tài chính trong tương lai, người vay sẽ có xu hướng thanh toán các khoản vay và hóa đơn đúng hạn, cũng như sử dụng tín dụng một cách hợp lý.

**Ngoài ra, điểm tín dụng còn có thể được sử dụng cho các mục đích khác như:**

* **Thuê nhà:** Một số chủ nhà có thể yêu cầu người thuê nhà cung cấp điểm tín dụng để đánh giá khả năng thanh toán tiền thuê nhà của họ.
* **Mua bảo hiểm:** Một số công ty bảo hiểm có thể sử dụng điểm tín dụng để đánh giá rủi ro và mức phí bảo hiểm.
* **Xin việc:** Một số công ty có thể yêu cầu ứng viên cung cấp điểm tín dụng để đánh giá tính trung thực và trách nhiệm của họ.

Tóm lại, điểm tín dụng là một công cụ quan trọng giúp đánh giá mức độ uy tín của bạn trong việc sử dụng tín dụng. Điểm tín dụng cao giúp bạn dễ dàng tiếp cận các dịch vụ tài chính với lãi suất ưu đãi.

***3. Cách Chấm Điểm Tín Dụng của Ngân Hàng:***

Cách chấm điểm tín dụng thường bao gồm nhiều yếu tố, và mỗi ngân hàng có thể sử dụng các mô hình khác nhau. Dưới đây là một số yếu tố phổ biến mà các ngân hàng thường xem xét:

- Lịch sử thanh toán: Đánh giá xem khách hàng có đóng các khoản nợ đúng hạn không.

- Dư nợ hiện tại: Xem xét tỷ lệ giữa số tiền đang nợ và hạn mức tín dụng.

- Lịch sử tín dụng: Thời gian sử dụng các sản phẩm tín dụng và lịch sử về các khoản vay trước đó.

- Thu nhập và nợ nần: Xem xét thu nhập của khách hàng và các khoản nợ khác ngoài khoản vay hiện tại.

- Số lượng và loại hồ sơ tín dụng: Xem xét số lượng và loại hồ sơ tín dụng mà người dùng đã mở.

*Kết Luận:*

Trong tổng thể, điểm tín dụng đóng vai trò quan trọng trong việc đảm bảo tính ổn định của hệ thống tài chính và giúp tối ưu hóa quản lý rủi ro cho các tổ chức tài chính. Qua điều này, nó cũng ảnh hưởng trực tiếp đến việc cung cấp dịch vụ tín dụng cho các cá nhân và doanh nghiệp.

# NGHIÊN CỨU LIÊN QUAN

Trong lĩnh vực chấm điểm tín dụng, đã có nhiều nghiên cứu được tiến hành để đề xuất và thử nghiệm các phương pháp khác nhau nhằm cải thiện độ chính xác và độ tin cậy của các mô hình chấm điểm tín dụng.

Trong [1], tác giả đã đề xuất một mô hình đánh giá rủi ro tín dụng sử dụng kỹ thuật học máy SVM và mô hình XGBoost. Phương pháp này nhằm tăng cường khả năng dự đoán bằng cách tích hợp các mô hình khác nhau, dựa trên dữ liệu quá khứ để dự đoán tương lai. Nghiên cứu này đã thể hiện hiệu quả của mô hình ADHE (Adaptive and Dynamic Heterogeneous Ensemble) trong việc cải thiện độ chính xác của chấm điểm tín dụng so với các mô hình đồng nhất khác.

Trong [2], các nhà nghiên cứu đã thử nghiệm nhiều thuật toán học máy như LR, DT, RF, MLP, XGBoost và LightGBM trên một bộ dữ liệu chứa lịch sử giao dịch thẻ tín dụng và hồ sơ khách hàng. Kết quả thử nghiệm cho thấy MLP vượt trội hơn các thuật toán khác với độ chính xác và recall cao, đặc biệt là đạt được diện tích dưới đường cong (AUC) là 86,7% và chỉ số accuracy là 91,6%.

Trong [3], nghiên cứu đã sử dụng các thuật toán ML khác nhau như SVM, NNs, DT, RF và LR để phát triển một mô hình mới dự đoán điểm tín dụng. Kết quả cho thấy mô hình DT với lựa chọn tính năng tối ưu hóa tham số đã vượt trội hơn các mô hình khác với độ chính xác cao.

Nghiên cứu [4] tiếp tục phân tích độ chính xác của thuật toán RF và thuật toán LR trong việc phê duyệt các khoản vay ngân hàng. Kết quả cho thấy rừng ngẫu nhiên đạt được độ chính xác trung bình cao hơn so với hồi quy tuyến tính.

Cuối cùng, trong [5], tác giả đã thực hiện phân tích toàn diện các thuật toán học máy khác nhau, đánh giá khả năng ứng dụng của chúng trong phân loại điểm tín dụng. Kết quả nghiên cứu này cung cấp bằng chứng thuyết phục về hiệu quả của các phương pháp này trong chấm điểm tín dụng, với độ chính xác đáng chú ý.

**TÀI LIỆU THAM KHẢO**

[1] Almajid, Adi Sakti. "Multilayer Perceptron Optimization on Imbalanced Data Using SVM-SMOTE and One-Hot Encoding for Credit Card Default Prediction." Journal of Advances in Information Systems and Technology (2022).

[2] Moscato, Vincenzo, Antonio Picariello và Giancarlo Sperlí. "A benchmark of machine learning approaches for credit score prediction." Expert Syst. Appl. 165 (2021): 113986.

[3] Hjelkrem, Lars Ole, Petter Eilif de Lange và Erik Nesset. "The Value of Open Banking Data for Application Credit Scoring: Case Study of a Norwegian Bank." Journal of Risk and Financial Management (2022).

[4] Mokheleli, Tsholofelo và Tinofirei Museba. "Machine Learning Approach for Credit Score Predictions." Journal of Information Systems and Informatics (2023).

[5] Amato, Flora, Antonino Ferraro, Antonio Galli, Francesco Moscato, Vincenzo Moscato và Giancarlo Sperlí. "Credit Score Prediction Relying on Machine Learning." Trong Sistemi Evoluti per Basi di Dati (2022).

|  |  |  |
| --- | --- | --- |
| **id** | **Cite** | **Abstract** |
|  | @article{Roy2021ModellingAS,    title={Modelling a sustainable credit score system (SCSS) using BWM and fuzzy TOPSIS},    author={Pranith Kumar Roy and Krishnendu Shaw},    journal={International Journal of Sustainable Development \& World Ecology},    year={2021},    volume={29},    pages={195 - 208},    url={https://api.semanticscholar.org/CorpusID:237886994}  } | ***Modelling a sustainable credit score system (SCSS) using BWM and fuzzy TOPSIS***  Sustainable development has emerged as a critical agenda for all organisations around the world. Despite the fact that profitability and sustainability are inorganically linked to financial institutions, sustainable lending has been a constant focus of attention. Due to regulatory pressure and stakeholder concerns, financial institutions are forced to implement a variety of sustainable measures; they are also gradually thinking to give more support to socially impactful and sustainable projects. Financial institutions can play a major role in establishing sustainable development by adopting the green lending policy. However, there have been few studies on sustainability credit score systems (SCSS) that take into account social and environmental factors. To fill the gaps of existing literature, this study proposes a multi-criteria SCSS that takes into account the environment and social aspects in addition to financial and managerial aspects. A combined Best-Worst Method (BWM) and the fuzzy-Technique for Order Preferences by Similarity to an Ideal Solution (TOPSIS) method are used in this study to create a credit scoring system. The BWM is used to determine the weight of factors, and the fuzzy-TOPSIS is used to evaluate applicants. The ambiguity while evaluating borrowers has been captured by applying fuzzy set theory. A real-life case study is used to demonstrate the efficacy of the proposed model. The model is unique in terms of the number of social and environmental factors considered. This research will assist financial institutions in identifying borrowers who engage in sustainable business practices. Borrowers can be holistically prioritised by applying the model.  IMG_257 |
|  | @article{Sadgali2021HumanBS,  title={Human behavior scoring in credit card fraud detection},  author={Imane Sadgali and Nawal Sael and Faouzia Benabbou},  journal={IAES International Journal of Artificial Intelligence},  year={2021},  volume={10},  pages={698-706},  url={https://api.semanticscholar.org/CorpusID:237157434}  } | ***Human behavior scoring in credit card fraud detection*** Now days, the analysis of the behavior of cardholders is one of the important fields in electronic payment. This kind of analysis helps to extract behavioral and transaction profile patterns that can help financial systems to better protect their customers. In this paper, we propose an intelligent machine learning (ML) system for rules generation. It is based on a hybrid approach using rough set theory for feature selection, fuzzy logic and association rules for rules generation. A score function is defined and computed for each transaction based on the number of rules, that make this transaction suspicious. This score is kind of risk factor used to measure the level of awareness of the transaction and to improve a card fraud detection system in general. The behavior analysis level is a part of a whole financial fraud detection system where it is combined to intelligent classification to improve the fraud detection. In this work, we also propose an implementation of this system integrating the behavioral layer. The system results obtained are very convincing and the consumed time by our system, per transaction was 6 ms, which prove that our system is able to handle real time process. |
|  | @article{Hayashi2022EmergingTI,  title={Emerging Trends in Deep Learning for Credit Scoring: A Review},  author={Yoichi Hayashi},  journal={Electronics},  year={2022}, | ***Emerging Trends in Deep Learning for Credit Scoring: A Review***  This systematic review aims to provide deep insights on emerging trends in, and the potential of, advanced deep learning techniques, such as machine learning algorithms being partially replaced by deep learning (DL) algorithms for credit scoring owing to the higher accuracy of the latter. This review also seeks to explain the reasons that deep belief networks (DBNs) can achieve higher accuracy than shallower networks, discusses the potential classification capabilities of DL-based classifiers, and bridges DL and explainable credit scoring. The theoretical characteristics of DBNs are also presented along with the reasons for their higher accuracy compared to that of shallower networks. Studies published between 2019 and 2022 were analysed to review and compare the most recent DL techniques that have been found to achieve higher accuracies than ensemble classifiers, their hybrids, rule extraction methods, and rule-based classifiers. The models reviewed in this study were evaluated and compared according to their accuracy and area under the receiver operating characteristic curve for the Australian, German (categorical), German (numerical), Japanese, and Taiwanese datasets, which are commonly used in the credit scoring community. This review paper also explains how tabular datasets are converted into images for the application of a two-dimensional convolutional neural network (CNN) and how “black box” models using local and global rule extraction and rule-based methods are applied in credit scoring. Finally, a new insight on the design of DL-based classifiers for credit scoring datasets is provided, along with a discussion on promising future research directions. |
|  | @article{Atif2022TheME,    title={The Most Effective Strategy for Incorporating Feature Selection into Credit Risk Assessment},    author={Dalia Atif and Mabrouka Salmi},    journal={SN Computer Science},    year={2022},    volume={4},    pages={1-18},    url={https://api.semanticscholar.org/CorpusID:254851653}  } | ***The Most Effective Strategy for Incorporating Feature Selection into Credit Risk Assessment***  This paper aims to identify the most effective strategy for incorporating feature selection (FS) into credit risk classification, employing three classifiers: Logistic regression (Logreg), Random Forests (RF), and Support Vector Machine (SVM) with the linear kernel through various embedded and wrapper strategies existing in the literature. We performed a comparative analysis on the German Credit dataset using three criteria: classification error rate, stability of selection, and calculation time. According to the Welsh t-test, RFE-RF (Recursive Feature Elimination for RF) outperformed RFE-SVM and penalized Logistic regression, with no significant difference in F1-score for RFE-SVM and suffers from the long-running computation. Conversely, RFE-SVM offers the best stability of 71% with a significantly shorter computation time. Furthermore, the paper intends to introduce a new classification of feature selection strategies in credit risk assessment in light of recent developments. Based on this new classification, a comparison with related literature reveals that the one-stage FS (RFE-RF and RFE-SVM) provides roughly the same accuracy as the two-stage FS and the two-stage classification model and, in some cases, outperforms. |
|  | @article{Li2021CreditSB,  title={Credit scoring by one-class classification driven dynamical ensemble learning},  author={Hao Li and Hao Qiu and Shu Sun and Jun Chang and Wenting Tu},  journal={Journal of the Operational Research Society},  year={2021},  volume={73},  pages={181 - 190},  url={https://api.semanticscholar.org/CorpusID:237749067}  } | ***Credit scoring by one-class classification driven dynamical ensemble learning*** It is very useful to endow machines with the ability to measure credit scores of loan applicants. Conventional methodologies always train Credit Scoring (CS) models by using data from clients who passed previous credit examination (i.e. who were considered adequately creditworthy and took out a loan). However, the CS models trained on data from the applicants who with good credit background may not work well for new applicants with plain or ambiguous credit backgrounds. Previous work always alleviates this by techniques of rejected inference and semisupervised learning. In this article, we propose a novel approach called as “One-class Classification Driven Dynamical Ensemble Learning” (abbreviated as OCDDEL). Different from rejected inference or semisupervised learning, OCDDEL does not use inferred labels of past rejected applications. Instead, OCDDEL only relies on past accepted applications and their true labels. It builds a dynamical ensemble model which deal with different test applications in different ways. To determine the ensemble weights for a specific test case, OCDDEL will learn a one-class classifier to separate test applications into groups, according to their similarities with training applicants. An experimental evaluation with 2 real-world datasets demonstrates the effectiveness of our approach. |
|  | @article{Schmitt2022DeepLV,  title={Deep Learning vs. Gradient Boosting: Benchmarking state-of-the-art machine learning algorithms for credit scoring},  author={Marc Schmitt},  journal={ArXiv},  year={2022},  volume={abs/2205.10535}, | ***Deep Learning vs. Gradient Boosting: Benchmarking state-of-the-art machu=ine learning algorithms for credit scoring***  Artificial intelligence (AI) and machine learning (ML) have become vital to remain competitive for financial services companies around the globe. The two models currently competing for the pole position in credit risk management are deep learning (DL) and gradient boosting machines (GBM). This paper benchmarked those two algorithms in the context of credit scoring using three distinct datasets with different features to account for the reality that model choice/power is often dependent on the underlying characteristics of the dataset. The experiment has shown that GBM tends to be more powerful than DL and has also the advantage of speed due to lower computational requirements. This makes GBM the winner and choice for credit scoring. However, it was also shown that the outperformance of GBM is not always guaranteed and ultimately the concrete problem scenario or dataset will determine the final model choice. Overall, based on this study both algorithms can be considered state-of-the-art for binary classification tasks on structured datasets, while GBM should be the go-to solution for most problem scenarios due to easier use, significantly faster training time, and superior accuracy. |
|  | @article{Melyanti2021OnlineDO,    title={Online Determination of Credit Score (PAK) Application Functional Teachers},    author={Rika Melyanti and M. Giatman and Riri Mayliza},    journal={International Journal of Management and Humanities},    year={2021},    url={https://api.semanticscholar.org/CorpusID:235551539}  } | ***Online Determination of Credit Score (PAK) Application Functional Teachers***  Determination of Teacher Credit Numbers (PAK) is proposed by the teacher, which are evaluated and evaluated by the Assessment Team. Calculation of credit numbers in PAK still uses manual methods using Microsoft Excel for inputting and Microsoft results as reports, errors in input such as typos and risk of accidental deletion of data still occur frequently. The DUPAK report that will be input into the PAK system is also still waiting for the Assessment Team to send the file to the Pelalawan Regency Education Office so that it takes more time to complete the functional teacher promotion report. To overcome this problem, the credit score calculation process is fast and accurate. A new web-based system, which includes all the elements that are valued by credit numbers. In the old system procedure and the new system it is not much different, the fundamental difference between the new system uses a web-based computer technology in data management that can shorten the data entry process and can overcome the obstacles of the old system. |
|  | @article{Adisa2022CreditSP,  title={Credit Score Prediction using Genetic Algorithm-LSTM Technique},  author={Juliana Adeola Adisa and Samuel Olusegun Ojo and Pius Adewale Owolawi and Agnieta Pretorius and S. Ojo},  journal={2022 Conference on Information Communications Technology and Society (ICTAS)},  year={2022},  pages={1-6},  url={https://api.semanticscholar.org/CorpusID:247961496}  } | ***Credit Score Prediction using Genetic Algorithm-LSTM Technique*** In data mining, the goal of prediction is to develop a more effective model that can provide accurate results. Prior literature has studied different classification techniques and found that combining multiple classifiers into ensembles outperformed most single classifier approaches. The performance of an ensemble classifier can be affected by some factors. How to determine the best classification technique? Which combination method to employ? This paper applies Long Short-Term Memory (LSTM), one of the most advanced deep learning algorithms which are inherently appropriate for the financial domain but rarely applied to credit scoring prediction. The research presents an optimization approach to determine the optimal parameters for a deep learning algorithm. The LSTM parameters are determined using an optimization algorithm. The LSTM parameters include epochs, batch size, number of neurons, learning rate and dropout. The results show that the optimized LSTM model outperforms both single classifiers and ensemble models. |
|  | @article{Jammalamadaka2022ResponsibleAI,  title={Responsible AI in automated credit scoring systems},  author={Krishna Ravali Jammalamadaka and Srikanth Itapu},  journal={AI and Ethics},  year={2022},  pages={1-11}, | ***Responsible AI in automated credit scoring systems***  In the last few years, Artificial Intelligence (AI) has achieved a notable momentum that, may deliver the expectations over many application sectors across the field. For this to occur, expert systems and rule-based models need to overcome the limitation of fairness and interpretability. Paradigms underlying this problem fall within the so-called explainable AI (XAI) field. This report presents the work on German credit card dataset to overcome the challenges of fairness, bias and in return, deem the machine learning models giving a responsible expectation. This is defined as responsible AI in practice. Since the dataset we dealt with, is to classify credit score of a user as good or bad, using fair ML modelling approach, the key metric of interest is the F1-score to reduce share of misclassifications. It is observed that hyper parameter tuned XGBoost model (GC2) gives optimal performance in terms of both F1-score, accuracy and fairness for the case of both gender and age as protected variable through Disparate Impact Remover, a pre-processing bias mitigation technique. The same is deployed using both Heroku through Flask API (for age). The Disparate Impact Analysis (DIA) using H2O.AI helped to identify optimum threshold levels at which the fairness metrics are observed at legally acceptable/permissible levels for both age and gender. Overall, fairness, bias responsibility and explainability have been established for the dataset considered. |
|  | @article{Jakka2023ANC,    title={A novel credit scoring system in financial institutions using artificial intelligence technology},    author={Geethamanikanta Jakka and Amrutanshu Panigrahi and Abhilash Pati and Manmath Nath Das and Jyotsnarani Tripathy},    journal={Journal of Autonomous Intelligence},    year={2023},    url={https://api.semanticscholar.org/CorpusID:261147882}  } | ***A novel credit scoring system in financial institutions using artificial intelligence technology***  In order to evaluate a person’s or a company’s creditworthiness, financial institutions must use credit scoring. Traditional credit scoring algorithms frequently rely on manual and rule-based methods, which can be tedious and inaccurate. Recent developments in artificial intelligence (AI) technology have opened up possibilities for creating more reliable and effective credit rating systems. The data are pre-processed, including scaling using the 0–1 normalization method and resolving missing values by imputation. Information gain (IG), gain ratio (GR), and chi-square are three feature selection methodologies covered in the study. While GR normalizes IG by dividing it by the total entropy of the feature, IG quantifies the reduction in total entropy by adding a new feature. Based on chi-squared statistics, the most vital traits are determined using chi-square. This research employs different ML models to develop a hybrid model for credit score prediction. The ML algorithms support vector machine (SVM), neural networks (NNs), decision trees (DTs), random forest (RF), and logistic regression (LR) classifiers are employed here for experiments along with IG, GR, and chi-square feature selection methodologies for credit prediction over Australian and German datasets. The study offers an understanding of the decision-making process for informative characteristics and the functionality of machine learning (ML) in credit prediction tasks. The empirical analysis shows that in the case of the German dataset, the DT with GR feature selection and hyperparameter optimization outperforms SVM and NN with an accuracy of 99.78%. For the Australian dataset, SVM with GR feature selection outperforms NN and DT with an accuracy of 99.98%. |
|  | @article{en2020AHB,  title={A Hybrid Bi-level Metaheuristic for Credit Scoring},  author={Doruk Şen and Cem Çağrı D{\"o}nmez and Umman Mahir Yildirim},  journal={Information Systems Frontiers},  year={2020},  volume={22},  pages={1009 - 1019},  url={https://api.semanticscholar.org/CorpusID:220508755}  } | ***A Hybrid Bi-level Metaheuristic for Credit ScoringA Hybrid Bi-level Metaheuristic for Credit Scoring***   This research aims to propose a framework for evaluating credit applications by assigning a binary score to the applicant. The score is targeted to determine whether the credit application is ‘good’ or ‘bad’ in small business purpose loans. Even tiny performance improvements in small businesses may yield a positive impact on the economy as they generate more than 60% of the value. The method presented in this paper hybridizes the Genetic Algorithm (GA) and the Support Vector Machine (SVM) in a bi-level feeding mechanism for increased prediction accuracy. The first level is to determine the parameters of SVM and the second is to find a feature set that increases classification accuracy. To test the proposed approach, we have investigated three different data sets; UCI Australian data set for preliminary works, Lending Club data set for large training and testing, and UCI German and Australian datasets for benchmarking against some other notable methods that use GA. Our computational results show that our proposed method using a feedback mechanism under the hybrid bi-level GA-SVM structure outperforms other classification algorithms in the literature, namely Decision Tree, Random Forests, Logistic Regression, SVM and Artificial Neural Networks, effectively improves the classification accuracy. |
|  | @article{Aftab2023FraudDO,  title={Fraud Detection of Credit Cards Using Supervised Machine Learning Techniques},  author={Ammar Aftab and Iqra Shahzad and Amna Sajid and Maira Anwar and Nosheen Anwar},  journal={Pakistan Journal of Emerging Science and Technologies (PJEST)},  year={2023}, | ***[Fraud Detection of Credit Cards Using Supervised Machine Learning Techniques](https://www.semanticscholar.org/paper/Fraud-Detection-of-Credit-Cards-Using-Supervised-Aftab-Shahzad/79feac25a4cecda6b6cc0476e8518dc7ba850b40)***  Credit card fraud encompasses illicit activities aimed at unlawfully obtaining confidential information to enable unauthorized individuals to engage in illegal transactions. As technology advances, fraudsters have honed their skills in evading security measures, presenting a formidable challenge in fraud detection. To address this issue, an array of algorithms and analytical techniques has emerged to identify and mitigate instances of fraud. This research aimed to identify the most appropriate supervised machine learning algorithm for credit card fraud detection. Logistic Regression, Random Forest, Support Vector Machine, and Decision Trees were implemented and compared. Due to the imbalanced nature of the dataset, the SMOTE (Synthetic Minority Oversampling Technique) technique was employed to rectify the data imbalance by oversampling the minority class. The performance of the trained models was evaluated using various metrics, including the confusion matrix, accuracy, precision, recall, f1-score, Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC). The results of the analysis revealed that Random Forests exhibited exceptional performance, achieving an impressive recall score of 84% and surpassing other algorithms. This research provides the groundwork for future investigations involving diverse deep-learning techniques applied to real-time and dynamic datasets, enabling continuous enhancements in fraud detection and prevention mechanisms. |
|  | @article{Salekshahrezaee2023TheEO,    title={The effect of feature extraction and data sampling on credit card fraud detection},    author={Zahra Salekshahrezaee and Joffrey L. Leevy and Taghi M. Khoshgoftaar},    journal={Journal of Big Data},    year={2023},    volume={10},    pages={1-17},    url={https://api.semanticscholar.org/CorpusID:256134459}  } | ***The effect of feature extraction and data sampling on credit card fraud detection***  Training a machine learning algorithm on a class-imbalanced dataset can be a difficult task, a process that could prove even more challenging under conditions of high dimensionality. Feature extraction and data sampling are among the most popular preprocessing techniques. Feature extraction is used to derive a richer set of reduced dataset features, while data sampling is used to mitigate class imbalance. In this paper, we investigate these two preprocessing techniques, using a credit card fraud dataset and four ensemble classifiers (Random Forest, CatBoost, LightGBM, and XGBoost). Within the context of feature extraction, the Principal Component Analysis (PCA) and Convolutional Autoencoder (CAE) methods are evaluated. With regard to data sampling, the Random Undersampling (RUS), Synthetic Minority Oversampling Technique (SMOTE), and SMOTE Tomek methods are evaluated. The F1 score and Area Under the Receiver Operating Characteristic Curve (AUC) metrics serve as measures of classification performance. Our results show that the implementation of the RUS method followed by the CAE method leads to the best performance for credit card fraud detection. |
|  | @article{Prassanna2020SecrecyProtectorAN,  title={SecrecyProtector: A Novel Data Analytics based credit score management system},  author={J. Prassanna and Abdul Quadir and J ChristyJackson and R. Prabakaran and Sakkaravarthi Ramanathan},  journal={International Journal of Scientific \& Technology Research},  year={2020},  volume={9},  pages={1215-1218},  url={https://api.semanticscholar.org/CorpusID:216291604}  } | ***SecrecyProtector: A Novel Data Analytics based credit score management system***   This work gives an account of the Credit Score web service application and the primary purpose of a credit score is to help lenders assess individuals' risk of not repaying a loan. Credit scoring assessment, despite the fact that a moderately new idea in the Indian money related business sector, have increased wide acknowledgment among financial specialists. In the meantime, easy-going and narrative confirmation recommends that there are worries among speculators and controllers about the execution of rating offices in India. This paper looks at financial specialists' mindfulness, discernment, understanding level and use of Credit scoring assessment through a poll-based example overview covering individual and additionally institutional speculators. We find high dissemination of rating use among all class of financial specialists, however, there is a recognizable upsetting with the dependability of appraisals, inclination of ensuing minimizing and opportuneness of rating reconnaissance. The review additionally uncovers that the institutional financial specialists have predominant information and comprehension about evaluations than individual speculators. In this way, the review underlines the requirement for rating offices to take a shot at instructing the basic speculators to engender appropriate comprehension and use of Credit score.. |
|  | @article{Kozodoi2021FairnessIC,  title={Fairness in Credit Scoring: Assessment, Implementation and Profit Implications},  author={Nikita Kozodoi and Johannes Jacob and Stefan Lessmann},  journal={ArXiv},  year={2021},  volume={abs/2103.01907}, | ***Fairness in Credit Scoring: Assessment, Implementation and Profit Implications*** The rise of algorithmic decision-making has spawned much research on fair machine learning (ML). Financial institutions use ML for building risk scorecards that support a range of credit-related decisions. Yet, the literature on fair ML in credit scoring is scarce. The paper makes three contributions. First, we revisit statistical fairness criteria and examine their adequacy for credit scoring. Second, we catalog algorithmic options for incorporating fairness goals in the ML model development pipeline. Last, we empirically compare different fairness processors in a profit-oriented credit scoring context using real-world data. The empirical results substantiate the evaluation of fairness measures, identify suitable options to implement fair credit scoring, and clarify the profit-fairness trade-off in lending decisions. We find that multiple fairness criteria can be approximately satisfied at once and recommend separation as a proper criterion for measuring the fairness of a scorecard. We also find fair in-processors to deliver a good balance between profit and fairness and show that algorithmic discrimination can be reduced to a reasonable level at a relatively low cost. The codes corresponding to the paper are available on GitHub1. |
|  | @article{Du2022ExplorationOF,    title={Exploration of Financial Market Credit Scoring and Risk Management and Prediction Using Deep Learning and Bionic Algorithm},    author={Peng Du and Hong Shu},    journal={J. Glob. Inf. Manag.},    year={2022},    volume={30},    pages={1-29},    url={https://api.semanticscholar.org/CorpusID:245301019}  } | ***Exploration of Financial Market Credit Scoring and Risk Management and Prediction Using Deep Learning and Bionic Algorithm***  The purpose is to effectively manage the financial market, comprehensive assess personal credit, reduce the risk of financial enterprises. Given the systemic risk problem caused by the lack of credit scoring in the existing financial market, a credit scoring model is put forward based on the deep learning network. The proposed model uses RNN (Recurrent Neural Network) and BRNN (Bidirectional Recurrent Neural Network) to avoid the limitations of shallow models. Afterward, to optimize path analysis, bionic optimization algorithms are introduced, and an integrated deep learning model is proposed. Finally, a financial credit risk management system using the integrated deep learning model is proposed. The probability of default or overdue customers is predicted through verification on three real credit data sets, thus realizing the credit risk management for credit customers. |
|  | @article{Sandeep2022ANA,  title={A Novel Approach for Bank Loan Approval by Verifying Background Information of Customers through Credit Score and Analyze the Prediction Accuracy using Random Forest over Linear Regression Algorithm},  author={Ch.Venkata Sandeep and T. Kalavathi Devi},  journal={Journal of Pharmaceutical Negative Results},  year={2022},  url={https://api.semanticscholar.org/CorpusID:252893023}  } | ***A Novel Approach for Bank Loan Approval by Verifying Background Information of Customers through Credit Score and Analyze the Prediction Accuracy using Random Forest over Linear Regression Algorithm*** Aim: To analyze the accuracy of Novel Random Forest (RF) and Linear Regression Algorithm (LR) algorithms used to approve bank loans. Materials and Methods: The existing model uses Linear Regression Algorithm (LR) and the proposed model employs a Novel Random Forest (RF). The Random Forest is a supervised learning model, it constructs solutions for different regression problems. It provides a high rate of accuracy by cross-validation. The 20 sample values are used to find out the Mean, Std. Deviation and Std. error means. The sample size was measured as 40 per group using G power (80%). Results: The resultant graph explains the comparison of the mean accuracy values of algorithms Novel Random Forest (RF) and Linear Regression (LR) where the mean accuracy of the Novel random forest is about 70.5% and the mean accuracy value of the Linear Regression is about 69.5%. The significance obtained is p=1.0 that is p>0.05, it shows insignificance between the groups based on independent sample T-Test. Conclusion : The mean accuracy rate of the Novel Random Forest algorithm has been improved to 70.5% compared to Linear Regression which is having around 69.5% mean accuracy. |
|  | @article{Laborda2021FeatureSI,  title={Feature Selection in a Credit Scoring Model},  author={Juan Laborda and Seyong Ryoo},  journal={Mathematics},  year={2021}, | ***[Feature Selection in a Credit Scoring Model](https://www.semanticscholar.org/paper/Feature-Selection-in-a-Credit-Scoring-Model-Laborda-Ryoo/5a673d120540c69807ae7355ed1b6c2c99b8f29e)***  This paper proposes different classification algorithms—logistic regression, support vector machine, K-nearest neighbors, and random forest—in order to identify which candidates are likely to default for a credit scoring model. Three different feature selection methods are used in order to mitigate the overfitting in the curse of dimensionality of these classification algorithms: one filter method (Chi-squared test and correlation coefficients) and two wrapper methods (forward stepwise selection and backward stepwise selection). The performances of these three methods are discussed using two measures, the mean absolute error and the number of selected features. The methodology is applied for a valuable database of Taiwan. The results suggest that forward stepwise selection yields superior performance in each one of the classification algorithms used. The conclusions obtained are related to those in the literature, and their managerial implications are analyzed. |
|  | @article{Xia2022CreditCF,    title={Credit Card Fraud Detection Based on Support Vector Machine},    author={Jia-zhi Xia},    journal={Highlights in Science, Engineering and Technology},    year={2022},    url={https://api.semanticscholar.org/CorpusID:254709190}  } | ***Credit Card Fraud Detection Based on Support Vector Machine***  Due to the increasing popularity cashless transactions, credit card fraud has become one of the most common frauds and caused huge harm to the financial institutions and individuals in real life. In this academic paper, the algorithm Support Vector Machine (SVM) is used to build models to deal with the credit card fraud detection problem with the performance metrics AUC and F1-score. The experiment dataset is named Credit Card Transactions Fraud Detection Dataset from the Kaggle website. After the step of preprocessing, the dataset is split into the training, testing and validation dataset with 11 numerical features and a label feature called “is\_fraud”. The inner parameter “class\_weight” of the SVM algorithm in Python is set as “balanced” to deal with the imbalanced datasets. The main method to find the optimized models is using the GridSearchCV function in Python library sklearn. After tuning the hyperparameters and handling the overfitting phenomenon, the optimized models for the two metrics are found. The parameter values of the best model for AUC are C=10, class\_weight= “balanced”, g =0.01, kernel = “rbf”. The training AUC is 0.87 and testing AUC is 0.90. The parameter values of the final optimized model for F1-score are C=0.8, class\_weight= “balanced”, g =0.06, kernel = “rbf”. The final training F-score is 0.305 and testing F-score is 0.260. |
|  | @article{Strelcenia2023ImprovingCP,  title={Improving Classification Performance in Credit Card Fraud Detection by Using New Data Augmentation},  author={Emilija Strelcenia and Simant Prakoonwit},  journal={AI},  year={2023},  url={https://api.semanticscholar.org/CorpusID:256534483}  } | ***Improving Classification Performance in Credit Card Fraud Detection by Using New Data Augmentation***   In many industrialized and developing nations, credit cards are one of the most widely used methods of payment for online transactions. Credit card invention has streamlined, facilitated, and enhanced internet transactions. It has, however, also given criminals more opportunities to commit fraud, which has raised the rate of fraud. Credit card fraud has a concerning global impact; many businesses and ordinary users have lost millions of US dollars as a result. Since there is a large number of transactions, many businesses and organizations rely heavily on applying machine learning techniques to automatically classify or identify fraudulent transactions. As the performance of machine learning techniques greatly depends on the quality of the training data, the imbalance in the data is not a trivial issue. In general, only a small percentage of fraudulent transactions are presented in the data. This greatly affects the performance of machine learning classifiers. In order to deal with the rarity of fraudulent occurrences, this paper investigates a variety of data augmentation techniques to address the imbalanced data problem and introduces a new data augmentation model, K-CGAN, for credit card fraud detection. A number of the main classification techniques are then used to evaluate the performance of the augmentation techniques. These results show that B-SMOTE, K-CGAN, and SMOTE have the highest Precision and Recall compared with other augmentation methods. Among those, K-CGAN has the highest F1 Score and Accuracy. |
|  | @inproceedings{Kumar2020ReviewOM,  title={Review of Machine Learning models for Credit Scoring Analysis},  author={Madapuri Rudra Kumar and Vinit Kumar Gunjan},  year={2020}, | ***Review of Machine Learning models for Credit Scoring Analysis***  Increase in computing power and the deeper usage of the robust computing systems in the financial system is propelling the business growth, improving the operational efficiency of the financial institutions, and increasing the effectiveness of the transaction processing solutions used by the organizations. Problem: Despite that the financial institutions are relying on the credit scoring patterns for analyzing the credit worthiness of the clients, still there are many factors that are imminent for improvement in the credit score evaluation patterns. There is need for improving the pattern to enhance the quality of analysis. Objective: Machine learning is offering immense potential in Fintech space and determining a personal credit score. Organizations by applying deep learning and machine learning techniques can tap individuals who are not being serviced by traditional financial institutions. Methodology: One of the major insights into the system is that the traditional models of banking intelligence solutions are predominantly the programmed models that can align with the information and banking systems that are used by the banks. But in the case of the machine-learning models that rely on algorithmic systems require more integral computation which is intrinsic. Hence, it can be advocated that the models usually need to have some decision lines wherein the dynamic calibration model must be streamlined. Such structure demands the dynamic calibration to have a decision tree system to empower with more integrated model changes. Results: The test analysis of the proposed machine learning model indicates effective and enhanced analysis process compared to the non-machine learning solutions. The model in terms of using various classifiers indicate potential ways in which the solution can be significant. Conclusion: If the systems can be developed to align with more pragmatic terms for analysis, it can help in improving the process conditions of customer profile analysis, wherein the process models have to be developed for comprehensive analysis and the ones that can make a sustainable solution for the credit system management. Originality: The proposed solution is effective and the one conceptualized to improve the credit scoring system patterns. If the model can be improved with more effective parameters and learning metrics, it can be sustainable outcome. Limitations: The model is tested in isolation and not in comparison to any of the existing credit scoring patterns. Only the inputs in terms of shortcomings from the existing models are taken in to account and accordingly the proposed solution is developed. |
|  | @article{Yarahmadi2023MultiAgentCA,    title={Multi-Agent Credit Assignment and Bankruptcy Game for Improving Resource Allocation in Smart Cities},    author={Hossein Yarahmadi and Mohammad Ebrahim Shiri and Moharram Challenger and Hamidreza Navidi and Arash Sharifi},    journal={Sensors (Basel, Switzerland)},    year={2023},    volume={23},    url={https://api.semanticscholar.org/CorpusID:256658447}  } | ***Multi-Agent Credit Assignment and Bankruptcy Game for Improving Resource Allocation in Smart Cities***  In recent years, the development of smart cities has accelerated. There are several issues to handle in smart cities, one of the most important of which is efficient resource allocation. For the modeling of smart cities, multi-agent systems (MASs) can be used. In this paper, an efficient approach is proposed for resource allocation in smart cities based on the multi-agent credit assignment problem (MCA) and bankruptcy game. To this end, the resource allocation problem is mapped to MCA and the bankruptcy game. To solve this problem, first, a task start threshold (TST) constraint is introduced. The MCA turns into a bankruptcy problem upon introducing such a constraint. Therefore, based on the concept of bankruptcy, three methods of TS-Only, TS + MAS, and TS + ExAg are presented to solve the MCA. In addition, this work introduces a multi-score problem (MSP) in which a different reward is offered for solving each part of the problem, and we used it in our experiments to examine the proposed methods. The proposed approach is evaluated based on the learning rate, confidence, expertness, efficiency, certainty, and correctness parameters. The results reveal the better performance of the proposed approach compared to the existing methods in five parameters. |
|  | @inproceedings{Dushimimana2020UseOM,  title={Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans},  author={Bernard Dushimimana and Yvonne Wambui and Timothy Lubega and Patrick E. McSharry},  year={2020},  url={https://api.semanticscholar.org/CorpusID:225423168}  } | ***Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans*** Airtime lending default rates are typically lower than those experienced by banks and microfinance institutions (MFIs) but are likely to grow as the service is offered more widely. In this paper, credit scoring techniques are reviewed, and that knowledge is built upon to create an appropriate machine learning model for airtime lending. Over three million loans belonging to more than 41 thousand customers with a repayment period of three months are analysed. Logistic Regression, Decision Trees and Random Forest are evaluated for their ability to classify defaulters using several cross-validation approaches and the latter model performed best. When the default rate is below 2%, it is better to offer everyone a loan. For higher default rates, the model substantially enhances profitability. The model quadruples the tolerable level of default rate for breaking even from 8% to 32%. Nonlinear classification models offer considerable potential for credit scoring, coping with higher levels of default and therefore allowing for larger volumes of customers. |
|  | @article{Hooman2016StatisticalAD,  title={Statistical and data mining methods in credit scoring},  author={Alireza Hooman and Govindan Marthandan and Wan Fauziah Wan Yusoff and Mohana Omid and Sasan Karamizadeh},  journal={The Journal of Developing Areas},  year={2016},  volume={50},  pages={371 - 381}, | ***[Statistical and data mining methods in credit scoring](https://www.semanticscholar.org/paper/Statistical-and-data-mining-methods-in-credit-Hooman-Marthandan/8c939302afce90f75a3068cc3f22ca7502eb0ed9)***  The growing interest in the credit industry resulted in credit scoring being developed as an essential component, especially in the credit department of banks that deals with huge sums of credit data. When a bank or a credit corporation is assessing a credit application request, they will have to decide whether to approve or deny it. This necessitates the utilization of credit scoring. Although pioneers attempt to compensate for risks via interest rates, current investigations on financial conditions of different sections of society confirmed that interest could not replace risk assessment, which means that credit risk requires its own specialized assessment. With the assistance of sorting methods, credit scoring simplifies the decision-making process. It is almost impossible to analyze this large amount of data in the context of manpower and economy, although the data mining technique helps alleviate this complexity. Nowadays, there are a lot of data mining methodologies being utilized in the management of credit scoring. However, each method has its advantages and limitations, and there has not been a comprehensive approach in determining the most utilized data mining technique in the context of credit scoring. The major goal of this paper is to provide a complete literature survey on applied data mining methods, such as discriminant analysis, logistic regression, K-nearest neighbor, Bayesian classifier, decision tree, neural network, survival analysis, fuzzy rule-based system, support vector machine, and hybrid methods. These findings will assist researchers in realizing the most suitable approach in evaluating credit scores, pinpoint limitations, enhance them, and propose new approaches with improved capabilities. Finally, the limitations of the new approaches are discussed, and further suitable methods are recommended. |
|  | @article{Leevy2023ThresholdOA,    title={Threshold optimization and random undersampling for imbalanced credit card data},    author={Joffrey L. Leevy and Justin M. Johnson and John T. Hancock and Taghi M. Khoshgoftaar},    journal={Journal of Big Data},    year={2023},    volume={10},    pages={1-22},    url={https://api.semanticscholar.org/CorpusID:258509782}  } | ***Threshold optimization and random undersampling for imbalanced credit card data***  utput thresholding is well-suited for addressing class imbalance, since the technique does not increase dataset size, run the risk of discarding important instances, or modify an existing learner. Through the use of the Credit Card Fraud Detection Dataset, this study proposes a threshold optimization approach that factors in the constraint True Positive Rate (TPR) ≥  True Negative Rate (TNR). Our findings indicate that an increase of the Area Under the Precision–Recall Curve (AUPRC) score is associated with an improvement in threshold-based classification scores, while an increase of positive class prior probability causes optimal thresholds to increase. In addition, we discovered that best overall results for the selection of an optimal threshold are obtained without the use of Random Undersampling (RUS). Furthermore, with the exception of AUPRC, we established that the default threshold yields good performance scores at a balanced class ratio. Our evaluation of four threshold optimization techniques, eight threshold-dependent metrics, and two threshold-agnostic metrics defines the uniqueness of this research. |
|  | @article{Kurniawan2021PREDICTIONFC,  title={PREDICTION FOR COOPERATIVE CREDIT ELIGIBILITY USING DATA MINING CLASSIFICATION WITH C4.5 ALGORITHM},  author={Yogiek Indra Kurniawan and Annastalia Fatikasari and Muhammad Luthfi Hidayat and Mohamad Waluyo},  journal={Jurnal Teknik Informatika (Jutif)},  year={2021},  url={https://api.semanticscholar.org/CorpusID:234893051}  } | ***Prediction For Cooperative Credit Eligibility Using Data Mining Classification With C4.5 Algorithm***  BMT Artha Mandiri is a cooperative that provides savings and loans services. In providing credit, BMT Artha Mandiri still uses the manual method, namely by looking at the ledger and history of each customer, to find out whether the applicant is worthy or not worthy of credit so that it is not effective and efficient. The purpose of this research is to make an application that can predict whether a prospective customer is eligible or not to be given credit. Predictions are made using the data mining classification method, namely the C4.5 algorithm based on the supporting data each customer has to classify which factors have the most influence on the level of credit payments in the cooperative. In a built application, the C4.5 algorithm produces a decision tree that is easy to interpret based on the existing variables. In the application, there are features that can be used to make decisions about customers who will apply for credit at the cooperative. The blackbox test results on the application show that the application has been able to run as expected, while the results of the algorithm test also show that the application has been able to implement the C4.5 algorithm correctly. In addition, the results of testing for accuracy show that the maximum average value of Accuracy is 79.19%. |
|  | @article{Arram2023CreditCS,  title={Credit card score prediction using machine learning models: A new dataset},  author={Anas Arram and Masri Ayob and Musatafa Abbas Abbood Albadr and Alaa Sulaiman and Dheeb Albashish},  journal={ArXiv},  year={2023},  volume={abs/2310.02956}, | ***[Credit card score prediction using machine learning models: A new dataset](https://www.semanticscholar.org/paper/Credit-card-score-prediction-using-machine-learning-Arram-Ayob/6d9343d5af5ea0dab0cc99ff76bf422a90262940)***  The use of credit cards has recently increased, creating an essential need for credit card assessment methods to minimize potential risks. This study investigates the utilization of machine learning (ML) models for credit card default prediction system. The main goal here is to investigate the best-performing ML model for new proposed credit card scoring dataset. This new dataset includes credit card transaction histories and customer profiles, is proposed and tested using a variety of machine learning algorithms, including logistic regression, decision trees, random forests, multi-layer perceptron (MLP) neural network, XGBoost, and LightGBM. To prepare the data for machine learning models, we perform data pre-processing, feature extraction, feature selection, and data balancing techniques. Experimental results demonstrate that MLP outperforms logistic regression, decision trees, random forests, LightGBM, and XGBoost in terms of predictive performance in true positive rate, achieving an impressive area under the curve (AUC) of 86.7% and an accuracy rate of 91.6%, with a recall rate exceeding 80%. These results indicate the superiority of MLP in predicting the default customers and assessing the potential risks. Furthermore, they help banks and other financial institutions in predicting loan defaults at an earlier stage. |
|  | @article{Li2023CreditRP,    title={Credit Risk Prediction Model for Listed Companies Based on CNN-LSTM and Attention Mechanism},    author={Jingyuan Li and Caosen Xu and Bing Feng and Hanyu Zhao},    journal={Electronics},    year={2023},    url={https://api.semanticscholar.org/CorpusID:257899057}  } | ***Credit Risk Prediction Model for Listed Companies Based on CNN-LSTM and Attention Mechanism***  The financial market has been developing rapidly in recent years, and the issue of credit risk concerning listed companies has become increasingly prominent. Therefore, predicting the credit risk of listed companies is an urgent concern for banks, regulators and investors. The commonly used models are the Z-score, Logit (logistic regression model), the kernel-based virtual machine (KVM) and neural network models. However, the results achieved could be more satisfactory. This paper proposes a credit-risk-prediction model for listed companies based on a CNN-LSTM and an attention mechanism, Our approach is based on the benefits of the long short-term memory network (LSTM) model for long-term time-series prediction combined with a convolutional neural network (CNN) model. Furthermore, the advantages of being integrated into a CNN-LSTM model include reducing the complexity of the data, improving the calculation speed and training speed of the model and solving the possible lack of historical data in the long-term sequence prediction of the LSTM model, resulting in prediction accuracy. To reduce problems, we introduced an attention mechanism to assign weights independently and optimize the model. The results show that our model has distinct advantages compared with other CNNs, LSTMs, CNN-LSTMs and other models. The research on the credit-risk prediction of the listing formula has significant meaning. |
|  | @article{Tian2021ANN,  title={A new non-kernel quadratic surface approach for imbalanced data classification in online credit scoring},  author={Ye Tian and Bo Bian and Xiaofei Tang and Jing Zhou},  journal={Inf. Sci.},  year={2021},  volume={563},  pages={150-165},  url={https://api.semanticscholar.org/CorpusID:233933973}  } | ***A new non-kernel quadratic surface approach for imbalanced data classification in online credit scoring***   Imbalanced data sets are very common in online credit scoring. Their imbalanced structures may cause statistical bias and poor performance of those traditional models. Hence, how to properly deal with them and dig useful information from them are very crucial for the risk management. This paper proposes a new approach to address the imbalanced data classification in this area by directly generating two quadratic surfaces in the original space. In this way, it avoids the time-consuming task for searching a proper kernel function and its corresponding parameters in the traditional support vector machine (SVM) models, hence significantly improve the total efficiency of the approach. Moreover, the homocentric structure and maximization margin principle are applied to obtain a good performance of our model on the issue. Besides, fuzzy weight is also incorporated to further increase the classification accuracy and robustness. It is worth noting that the linear programming structure of our model not only guarantees the global optimality of its solution, but also leads to a much higher efficiency than those benchmark models. In addition, the fewer parameters in our model further saves more time in the tuning process. Hence it is quite suitable for handling those huge-sized problems in this big data era. Finally, we conduct a comprehensive experiment to compare the performances and efficiencies of different methods. The numerical results on various data sets strongly verify the superiority of our method to some benchmark methods in handling the imbalanced data classification in online credit scoring. |
|  | @article{Hjelkrem2022TheVO,    title={The Value of Open Banking Data for Application Credit Scoring: Case Study of a Norwegian Bank},    author={Lars Ole Hjelkrem and Petter Eilif de Lange and Erik Nesset},    journal={Journal of Risk and Financial Management},    year={2022},    url={https://api.semanticscholar.org/CorpusID:254613848}  } | ***The Value of Open Banking Data for Application Credit Scoring: Case Study of a Norwegian Bank***  Banks generally use credit scoring models to assess the creditworthiness of customers when they apply for loans or credit. These models perform significantly worse when used on potential new customers than existing customers, due to the lack of financial behavioral data for new bank customers. Access to such data could therefore increase banks’ profitability when recruiting new customers. If allowed by the customer, Open Banking APIs can provide access to balances and transactions from the past 90 days before the score date. In this study, we compare the performance of conventional application credit scoring models currently in use by a Norwegian bank with a deep learning model trained solely on transaction data available through Open Banking APIs. We evaluate the performance in terms of the AUC and Brier score and find that the models based on Open Banking data alone are surprisingly effective in predicting default compared to the conventional credit scoring models. Furthermore, an ensemble model trained on both traditional credit scoring data and features extracted from the deep learning model further outperforms the conventional application credit scoring model for new customers and narrows the performance gap between application credit scoring models for existing and new customers. Therefore, we argue that banks can increase their profitability by utilizing data available through Open Banking APIs when recruiting new customers. |

# PHƯƠNG PHÁP

## 3.1. Các kỹ thuật nền tảng

Trong phần này, nhóm chúng tôi sẽ trình bày và giải thích các kỹ thuật cơ bản để xây dựng một hệ thống chấm điểm tín dụng thường sử dụng. Các kỹ thuật này là các công cụ và phương pháp chính để xây dựng mô hình dự đoán rủi ro tín dụng và đánh giá khả năng trả nợ của khách hàng. Dưới đây là một số kỹ thuật quan trọng:

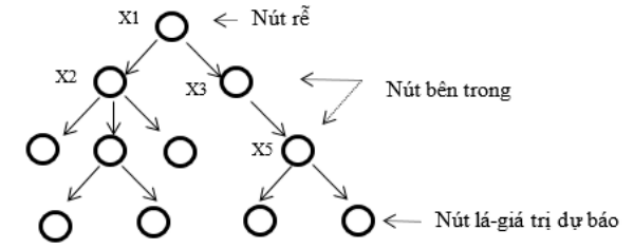
* *Phân tích và xử lý dữ liệu:* Đây là bước quan trọng nhất trong quá trình xây dựng mô hình chấm điểm tín dụng. Phân tích dữ liệu giúp hiểu rõ hơn về các đặc điểm của khách hàng và mối quan hệ giữa các yếu tố dữ liệu và khả năng trả nợ của họ. Xử lý dữ liệu bao gồm việc xử lý các giá trị bị thiếu, chọn lọc đặc trưng và chuẩn hóa dữ liệu bằng Min-Max Scaling (chuẩn hóa Min-Max) hay Standard Scaler .
* *Mô hình hóa dữ liệu:* Các kỹ thuật này bao gồm việc lựa chọn và xây dựng các mô hình phân loại và học máy để dự đoán rủi ro tín dụng. Các mô hình phổ biến bao gồm Logistic Regression, Decision Trees, Random Forests, Adaboost, và Neural Networks.
* Đánh giá mô hình: Kỹ thuật này liên quan đến việc đánh giá hiệu suất của các mô hình dự đoán. Các phương pháp đánh giá bao gồm Confusion Matrix, ROC Curve, Precision - Recall Curve và các độ đo như Accuracy, Precision, Recall và F1-score.
* Xử lý mất cân bằng dữ liệu: Trong các vấn đề chấm điểm tín dụng, thường có sự mất cân bằng giữa các nhóm dữ liệu, ví dụ như số lượng khách hàng trả nợ và không trả nợ. Các kỹ thuật như Oversampling, Undersampling, và SMOTE được sử dụng để giải quyết vấn đề này.
* Tinh chỉnh mô hình: Sau khi xây dựng mô hình, quá trình tinh chỉnh là cần thiết để điều chỉnh các tham số mô hình và tối ưu hóa hiệu suất của nó. Các kỹ thuật như Cross-Validation và Grid Search được sử dụng để tinh chỉnh mô hình.

Các kỹ thuật nền tảng này cùng với các công cụ và phương pháp khác tạo ra cơ sở cho việc phát triển các mô hình chấm điểm tín dụng hiệu quả và chính xác. Trong phần tiếp theo, chúng ta sẽ đề xuất phương pháp cụ thể để áp dụng các kỹ thuật này vào việc giải quyết vấn đề chấm điểm tín dụng khách hàng.

## 3.2. Đề xuất phương pháp

Vấn đề lớn đối với các tổ chức tài chính và ngân hàng là xác định tín dụng xấu, vì nó có thể gây ra các vấn đề nghiêm trọng trong tương lai như thất thoát vốn, giảm doanh thu và tăng tổn thất cho ngân hàng. Hiện nay, có rất nhiều kĩ thuật học máy để chấm điểm tín dụng của khách hàng. Nhóm chúng tôi đã đề xuất sử dụng một tổ hợp các mô hình học máy truyền thống và mô hình học sâu để chấm điểm tín dụng của khách hàng. Cụ thể, chúng tôi đề xuất sử dụng 7 mô hình sau đây:

* Decision Tree (Cây quyết định): Cây quyết định (Decision Tree - DT) là một thuật toán học máy được ứng dụng rộng rãi trong phân loại và dự đoán, bao gồm cả lĩnh vực đánh giá rủi ro tín dụng. Mô hình này hoạt động dựa trên hệ thống các quy tắc logic được biểu diễn dưới dạng cây, giúp phân loại khách hàng và dự đoán khả năng thanh toán khoản vay của họ. Mô hình này bao gồm một nút gốc (Root node), các nút bên trong (Internal node) và nút lá (Leaf node). Cấu trúc của cây quyết định gồm nút gốc - đại diện cho điểm khởi đầu của quá trình phân loại, nút bên trong - thể hiện các thuộc tính (biến) được sử dụng để phân chia dữ liệu và nút lá - đại diện cho kết quả dự đoán (ví dụ: mức độ rủi ro tín dụng) (Hình 1).



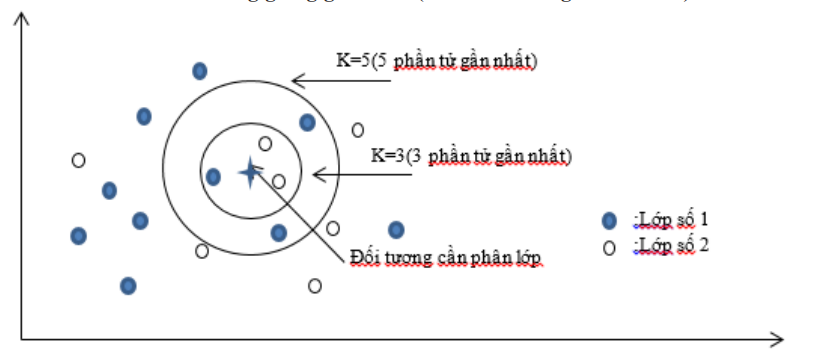
**Hình 1.** Cây quyết định (Decision tree)

*Nguồn: Abdou. (2011)*

* K-Nearest Neighbor (K-Láng giềng gần nhất): KNN là một mô hình dựa trên sự gần gũi của điểm dữ liệu, dựa vào các điểm láng giềng gần nhất để dự đoán nhãn của một điểm dữ liệu mới.

Từ Hình 2 có thể thấy, nếu chọn số điểm gần nhất để phân loại là 3 thì xung quanh điểm cần phân lớp có 3 điểm trong đó 1 điểm thuộc lớp 1 và 2 điểm thuộc lớp 2. Với số điểm thuộc lớp 2 nhiều hơn, điểm cần phân lớp được đưa vào lớp số 1. Tương tự như vậy, nếu lựa chọn số điểm gần nhất k=5 thì có năm điểm xung quanh điểm cần phân lớp, trong đó có 3 điểm thuộc lớp số 1 và 2 điểm thuộc lớp 2. Nếu theo dữ liệu này thì điểm cần phân lớp sẽ thuộc lớp số 1. Trên thực tế dữ liệu cần phân loại có nhiều thuộc tính trong đó mỗi thuộc tính tương ứng với một chiều không gian, do vậy khi tính khoảng cách gần nhất cần tính khoảng cách vector trong không gian đa chiều với công thức khoảng cách Euclidean:

trong đó: x, y là 2 điểm cần xác định khoảng cách; n là số chiều không gian (số thuộc tính của biến mục tiêu).



**Hình 2.** K láng giềng gần nhất (K-Nearest Neighbors - KNN)

*Nguồn: Marinakis và cộng sự (2008)*

* Random Forest (Rừng ngẫu nhiên): RF là một mô hình ensemble của nhiều cây quyết định, có khả năng tạo ra các dự đoán mạnh mẽ và ổn định bằng cách kết hợp kết quả từ nhiều cây quyết định. Random Forest có thể được sử dụng để dự đoán khả năng thanh toán của khách hàng dựa trên các đặc trưng tài chính và hành vi thanh toán. Điều này là do RF thích hợp với việc xử lý dữ liệu có cấu trúc và có khả năng làm việc với các đặc trưng có mức độ quan trọng khác nhau mà không cần phải tiền xử lý đặc biệt.
* XGBoost (Extreme Gradient Boosting): XGBoost là một mô hình ensemble learning, cải tiến từ Gradient Boosting, giúp cải thiện hiệu suất dự đoán bằng cách tối ưu hóa hàm mất mát và xây dựng các cây quyết định tuần tự.
* Naive Bayes: Naive Bayes là một mô hình đơn giản dựa trên xác suất, dựa trên giả định về sự độc lập giữa các biến, nhưng thường cho kết quả tốt trong nhiều bài toán phân loại.
* Convolutional Neural Network (Mạng nơ-ron tích chập sâu): CNN là một loại mạng nơ-ron nhân tạo đặc biệt phù hợp để phân tích dữ liệu dạng lưới, phổ biến nhất là hình ảnh. Nó có khả năng học một lượng lớn các dữ liệu trong khoảng thời gian ngắn hơn nhiều so với mạng nơ ron thông thường. Lý do là nó sử dụng ít trọng số hơn trong khi độ chính xác chỉ kém hơn một phần nhỏ so với kiến trúc truyền thống.
* Recurrent Neural Network (Mạng nơ-ron hồi quy): RNN là một mô hình mạng neural thích hợp cho việc xử lý dữ liệu chuỗi như văn bản hoặc chuỗi thời gian, có khả năng nhớ thông tin từ quá khứ để đưa ra dự đoán tín dụng cho tương lai.

Sử dụng một tổ hợp của các mô hình này giúp tăng cường khả năng dự đoán và đảm bảo tính đa dạng trong quá trình đánh giá tín dụng của khách hàng. Các mô hình truyền thống cung cấp một cơ sở ổn định và dễ hiểu, trong khi các mô hình học sâu có khả năng học được các biểu diễn phức tạp từ dữ liệu, giúp cải thiện hiệu suất dự đoán trong các trường hợp phức tạp.

Phương pháp được thực hiện theo tiến trình:

**Bước 1:** Khai thác dữ liệu

* Quan sát tổng quan dữ liệu (Hệ tương quan, giá trị thống kê, vấn đề mất cân bằng dữ liệu ở biến mục tiêu).

**Bước 2:** Chuẩn bị dữ liệu

* Tiến hành tiền xử lý dữ liệu (xóa trùng lặp, kiểm tra dữ liệu trống), biến đổi các trường dữ liệu định danh với phương pháp mã hóa nhãn (LabelEncoder).
* Xóa các đặc trưng không quan trọng
* Sử dụng phương pháp IQR để loại bỏ các giá trị ngoại lệ
* Sử dụng phương pháp Min-Max Scaling để chuẩn hóa

**Bước 3:** Xây dựng tập dữ liệu huấn luyện và kiểm tra

Bộ dữ liệu được chia thành 2 tập gồm tập huấn luyện 80% và tập dữ liệu kiểm tra 20%.

**Bước 4:** Xây dựng và huấn luyện các mô hình

* Với tập dữ liệu huấn luyện tiến hành xây dựng các mô hình với phương pháp Grid search CV để tìm ra các siêu tham số.

**Bước 5:** Mô hình thử nghiệm với tập dữ liệu kiểm tra đưa ra đánh giá mô hình.

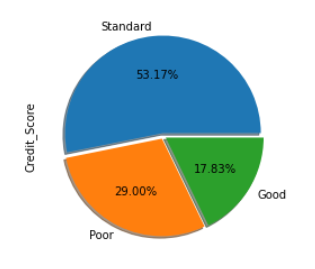
**Bước 6:** So sánh các mô hình dựa trên hiệu suất các chỉ số và đưa ra mô hình tốt nhất

**IV. THỰC NGHIỆM**

## 4.1. Miêu tả dữ liệu

Để tuân thủ các quy định pháp luật hiện hành về bảo mật thông tin cá nhân, các tổ chức tài chính phải tuân thủ nguyên tắc không tiết lộ thông tin chi tiết về tài khoản và các giao dịch của khách hàng. Trong phạm vi nghiên cứu này, chúng tôi tiến hành thực nghiệm trên một tập dữ liệu được công bố công khai có tên là “Credit score classification”, được truy cập và tải về từ trang Kaggle.com. Bộ dữ liệu này được thiết kế để mô phỏng các loại dữ liệu giao dịch tài chính và không liên quan trực tiếp đến bất kỳ ngân hàng cụ thể nào. Tập dữ liệu bao gồm 28 đặc trưng với 100,000 quan sát và chi tiết được mô tả như trong Bảng 1.

Trường dữ liệu “Credit Score”: được sử dụng để đánh giá khả năng thanh toán của khách hàng và phản ánh lịch sử tín dụng của họ. Trường dữ liệu này có ba nhóm giá trị: “Good” (tốt), “Standard” (trung bình), và “Poor” (kém). Nhãn “Good” được gán cho những khách hàng có điểm tín dụng cao, thường là do họ có lịch sử thanh toán nợ đúng hạn và duy trì một hồ sơ tín dụng tích cực, nhãn này có 17828 quan sát chiếm 17.828%. Nhãn “Standard” thường ám chỉ đến những khách hàng có lịch sử tín dụng trung bình, có thể có một số vấn đề hoặc trễ hẹn trong quá khứ, nhãn này có 53174 quan sát chiếm 53.174%. Trong khi đó, nhãn “Poor” được gán cho những khách hàng có điểm tín dụng thấp, thường xuyên trễ hẹn hoặc không thực hiện thanh toán đúng hạn, nhãn này có 28998 quan sát chiếm 28.998%. Thông qua việc phân loại khách hàng thành các nhóm dựa trên biến "Credit\_Score", các tổ chức tài chính có thể chấm điểm tín dụng khách hàng dựa trên lịch sử tín dụng và các yếu tố liên quan.



**Hình 3.** Tỷ lệ các quan sát trong trường dữ liệu “Credit\_Score”

**Bảng 1:** Mô tả dữ liệu thực nghiệm

|  |  |  |
| --- | --- | --- |
| **Trường dữ liệu** | **Mô tả** | **Kiểu dữ liệu** |
| **ID** | Định danh duy nhất của mỗi quan sát | object |
| **Customer\_ID** | Định danh của khách hàng | object |
| **Month** | Tháng | object |
| **Name** | Tên khách hàng | object |
| **Age** | Tuổi khách hàng | object |
| **SSN** | Số An sinh Xã hội | object |
| **Occupation** | Nghề nghiệp của khách hàng | object |
| **Annual\_Income** | Thu nhập hàng năm của khách hàng | object |
| **Monthly\_Inhand\_Salary** | Lương thực nhận hàng tháng | float64 |
| **Num\_Bank\_Accounts** | Số tài khoản ngân hàng | int64 |
| **Num\_Credit\_Card** | Số thẻ tín dụng | int64 |
| **Interest\_Rate** | Lãi suất | int64 |
| **Num\_of\_Loan** | Số khoản vay | object |
| **Type\_of\_Loan** | Loại khoản vay | object |
| **Delay\_from\_due\_date** | Số ngày trễ so với ngày đáo hạn | int64 |
| **Num\_of\_Delayed\_Payment** | Số lượng thanh toán bị trễ | object |
| **Changed\_Credit\_Limit** | Số lượng thay đổi hạn mức tín dụng | object |
| **Num\_Credit\_Inquiries** | Số lượng điều tra tín dụng | float64 |
| **Credit\_Mix** | Tổ hợp tín dụng | object |
| **Outstanding\_Debt** | Nợ chưa thanh toán | object |
| **Credit\_Utilization\_Ratio** | Tỷ lệ sử dụng tín dụng | float64 |
| **Credit\_History\_Age** | Lịch sử tín dụng | object |
| **Payment\_of\_Min\_Amount** | Số tiền thanh toán tối thiểu | object |
| **Total\_EMI\_per\_month** | Tổng EMI hàng tháng | float64 |
| **Amount\_invested\_monthly** | Số tiền đầu tư hàng tháng | object |
| **Payment\_Behaviour** | Hành vi thanh toán | object |
| **Monthly\_Balance** | Số dư hàng tháng | object |
| **Credit\_Score** | Điểm tín dụng | object |

## 4.2. Tiền xử lý dữ liệu

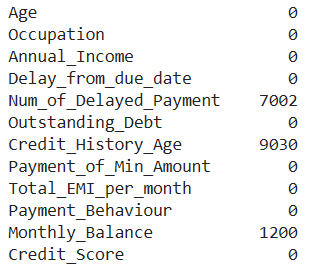
Việc tiền xử lý dữ liệu là một bước quan trọng trong quá trình chuẩn bị dữ liệu trước khi đưa vào quá trình huấn luyện mô hình học máy. Mục tiêu của tiền xử lý dữ liệu là tối ưu hóa dữ liệu để tăng cường hiệu suất của mô hình và cải thiện độ chính xác của mô hình theo mô hình học máy. Chúng tôi đã thựuc hiện tiền xử lý dữ liệu với các bước như sau:

**Bước 1:** *Loại bỏ dữ liệu không cần thiết*

Loại bỏ các cột dữ liệu không sử dụng trong mô hình học máy, 16 trường dữ liệu này bị loại bỏ đi bao gồm là 'ID', 'Customer\_ID', 'Month', 'Name', 'SSN', 'Monthly\_Inhand\_Salary', 'Num\_Bank\_Accounts', 'Num\_Credit\_Card', 'Interest\_Rate', 'Num\_of\_Loan', 'Type\_of\_Loan', 'Changed\_Credit\_Limit', 'Num\_Credit\_Inquiries', 'Credit\_Mix', 'Credit\_Utilization\_Ratio', 'Amount\_invested\_monthly'

**Bước 2:** *Loại bỏ các giá trị NaN*

Xóa các hàng chứa giá trị NaN, cụ thể số lượng hàng bị xóa do các trường dữ liệu có chứa giá trị NaN được thể hiện trong hình 1.



**Hình 4.** Số lượng giá trị lỗi NaN trong các trường dữ liệu

**Bước 3:** *Loại bỏ các hàng chứa giá trị không chính xác và thay thế các dữ liệu sai*

Xóa các hàng có dữ liệu không chính xác trong cột "Occupation" và "Payment\_Behaviour", thay thế ký tự đặc biệt bằng chuỗi rỗng.

**Bước 4:** *Mã hóa và chuẩn hóa dữ liệu*

* Chuyển đổi dữ liệu cột "Credit\_History\_Age" từ chuỗi thành số, cụ thể thay thế chuỗi con ' Years and ' (bao gồm khoảng trắng) bằng dấu chấm ('.') và chuỗi con 'Months' (bao gồm khoảng trắng) bằng chuỗi rỗng (''). Ví dụ: '2 Years and 3 Months' thành '2.3'. Mục đích là chuyển đổi định dạng từ năm và tháng sang dạng thập phân để tính toán dễ dàng hơn.
* Chuyển đổi các nhãn trong cột “Payment\_Behaviour” thành các giá trị theo cấp độ. Sủ dụng hàm ‘.str.replace(...)’ để thay thế gắn nhẵn “Low\_spent\_Small\_value\_payments” thành ‘1’,

“Low\_spent\_Medium\_value\_payments” thành ‘2’,

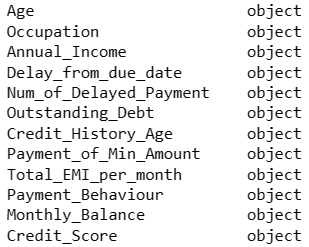
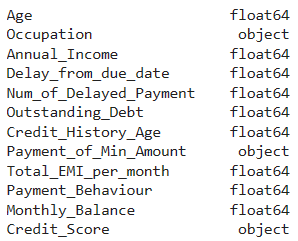
“Low\_spent\_Large\_value\_payments” thành ‘3’,

“High\_spent\_Small\_value\_payments” thành ‘4’,

“High\_spent\_Medium\_value\_payments” thành ‘5’,

“ High\_spent\_Large\_value\_payments” thành ‘6’.

* Chuyển đổi kiểu dữ liệu của các cột 'Age', 'Delay\_from\_due\_date', 'Num\_of\_Delayed\_Payment', 'Outstanding\_Debt', 'Total\_EMI\_per\_month', 'Monthly\_Balance', 'Payment\_Behaviour', 'Credit\_History\_Age', 'Annual\_Income' sang số thập phân

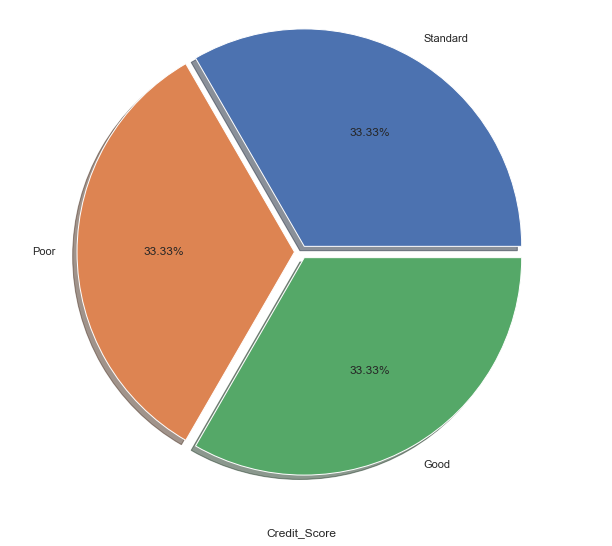
*Trước khi xử lý* *Sau khi xử lý*

**Hình 5.** Dữ liệu trước và sau khi thay đổi kiểu dữ liệu

* Tương tự đối với cột “Credit\_Score” và “Payment\_of\_Min\_Amount”, các nhãn của cột “Credit\_Score” lần lượt là ‘Good’, ‘Standard’, ‘Poor’ sẽ được thay thế thành ‘3’, ‘2’, ‘1’ và các nhãn cột “Payment\_of\_Min\_Amount” lần lượt là ‘NM’. ‘Yes’, ‘No’ được thay thế thành ‘0’, ‘1’, ‘2’, ‘3’. Cuối cùng chuyển kiểu dữ liệu 2 cột này thành số nguyên int64
* Sử dụng phương pháp IQR để loại bỏ các giá trị ngoại lệ. Đây là một phần quan trọng của tiền xử lý dữ liệu, vì ngoại lệ có thể làm méo mó các mô hình học máy và dẫn đến dự đoán không chính xác.

Phương pháp IQR (Interquartile Range) thường được sử dụng để xác định ngoại lệ. IQR là phạm vi giữa quartile thứ nhất (Q1) và quartile thứ ba (Q3) của tập dữ liệu. Các giá trị nằm ngoài khoảng từ Q1 - 1.5IQR đến Q3 + 1.5IQR thường được coi là ngoại lệ và có thể bị loại bỏ khỏi dữ liệu phân tích.

**Bước 5:** Sử dụng module Random Over Sampler xử lý sự mất cân bằng dữ liệu trong tập huấn luyện



**Hình 6.** Tỷ lệ các quan sát trong trường dữ liệu “Credit\_Score” sau khi cân bằng

**Bước 6:** Sử dụng phương pháp Min-Max Scaling để chuẩn hóa các biến số (numerical variables) . Các trường dữ liệu thuộc biến độc lập lúc này sẽ được chuẩn hóa Min - Max đưa dữ liệu về một phân bố trong phạm vi [0, 1], chuẩn hóa được xác định theo Công thức (1).

(1)

trong đó: x là giá trị ban đầu

x\_scaled là giá trị sau khi áp dụng Min-Max Scaling (thường nằm trong khoảng 0 đến 1)

x\_min, x\_max lần lượt là giá trị nhỏ nhất, giá trị lớn nhất của trường dữ liệu trong tập dữ liệu.

Bảng 2 minh họa 3 quan sát đầu của dữ liệu sau khi được chuẩn hóa.

**Bảng 2:** Dữ liệu sau chuẩn hóa

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Annual Income | Delay from due date | Num of Delayed Payment | Outstanding Debt | Credit History Age | Total EMI per month | Monthly Balance |
| 0.07031 | 0.083668 | 0.111111 | 0.002500 | 0.162020 | 0.67062 | 0.000602 | 1.0 |
| 0.10937 | 0.192389 | 0.166667 | 0.000909 | 0.121012 | 0.79228 | 0.000229 | 1.0 |
| 0.10937 | 0.192389 | 0.111111 | 0.000455 | 0.121012 | 0.79525 | 0.000229 | 1.0 |

**Bước 5:** *Mã hóa dữ liệu*

* Sử dụng phương pháp One - Hot Encoding để biến đổi các giá trị của cột "Occupation" thành các biến giả định dạng số.

Kết quả là DataFrame chứa dữ liệu cần biến đổi sẽ bao gồm các biến giả định mới được tạo ra từ One-Hot Encoding, mỗi biến giả định tương ứng với một nhóm nghề nghiệp (Occupation) và chứa các giá trị 0 hoặc 1 để biểu thị sự hiện diện của từng nhóm nghề nghiệp.

**Bước 6:** *Xác định dữ liệu tập huấn và kiểm tra*

Bộ dữ liệu được chia theo tỷ lệ 80:20 với tập huấn luyện (80%) và tập kiểm tra (20%), việc huấn luyện mô hình trên toàn bộ tập dữ liệu có thể tốn nhiều thời gian và tài nguyên, đặc biệt với các tập dữ liệu lớn. Chia dữ liệu thành hai tập nhỏ hơn giúp giảm thời gian huấn luyện và cho phép bạn thử nghiệm nhiều mô hình khác nhau một cách hiệu quả hơn. Mục tiêu là để đảm bảo rằng dữ liệu được chia dẫn tới phương sai không quá cao.

* Tập huấn luyện (80%) cung cấp dữ liệu cho mô hình học và xây dựng mối quan hệ giữa các biến.
* Tập kiểm tra (20%) là tập dữ liệu độc lập, chưa được sử dụng trong quá trình huấn luyện. Việc đánh giá hiệu suất mô hình trên tập kiểm tra giúp đảm bảo tính khách quan, tránh tình trạng "quá khớp" (overfitting) - khi mô hình học thuộc lòng tập huấn luyện nhưng không thể khái quát hóa tốt cho dữ liệu mới.

## 4.3. Các độ đo đánh giá hiệu năng

Trong bài nghiên cứu này, các phương pháp đánh giá nhóm đã sử dụng là các chỉ số đánh giá Accuracy, Precision, Recall và F1 score, dựa trên ma trận nhầm lẫn có dạng như sau:

* Phương pháp Accuracy (độ chính xác) của mô hình là thước đo đơn giản nhất để đánh giá hiệu suất mô hình, tỷ lệ giữa số lượng mẫu được phân loại chính xác trên cho tổng số mẫu của tập dữ liệu theo công thức sau:

|  |
| --- |
|  |

Accuracy cho biết mức độ tin cậy chung của mô hình chấm điểm. Nếu đạt giá trị cao: Mô hình dự đoán chính xác điểm tín dụng của khách hàng và giá trị thấp: Mô hình có thể đưa ra dự đoán sai lệch.

* Phương pháp Precision (độ chính xác dương) là tỷ lệ giữa số lượng mẫu được dự đoán khớp mẫu trên tổng dự đoán khớp mẫu và dự đoán đúng nhưng sai, được biểu diễn theo công thức dưới đây:

|  |  |
| --- | --- |
|  |  |

Precision đánh giá mức độ tin cậy của dự đoán khách hàng có điểm tín dụng cao (tốt). Nếu đạt giá trị cao: Mô hình ít dự đoán sai "tốt" cho khách hàng có điểm thấp và giá trị thấp: Mô hình có thể dự đoán sai "tốt" cho khách hàng có điểm thấp, dẫn đến việc đánh giá rủi ro sai lệch.

* Phương pháp Recall (độ nhạy) là tỷ lệ giữa số lượng mẫu được dự đoán khớp mẫu với tổng dự đoán khớp mẫu và dự đoán sai nhưng đúng được xác định theo công thức sau:

|  |
| --- |
|  |

Trong đó:

* + TP (True Positive) là tổng số trường hợp dự báo khớp mẫu đúng.
  + TN (True Negative) là tổng số trường hợp dự báo khớp với mẫu sai.
  + FP (False Positive) là tổng số trường hợp dự báo các quan sát thuộc mẫu đúng tính thành sai.
  + FN (False Negative) là tổng số trường hợp dự báo các quan sát thuộc mẫu sai tính thành đúng

Recall đánh giá khả năng xác định chính xác khách hàng có điểm tín dụng cao. Nếu đạt giá trị cao: Mô hình ít bỏ sót khách hàng có điểm cao, giúp giảm thiểu việc đánh giá thấp rủi ro và giá trị thấp: Mô hình có thể bỏ sót khách hàng có điểm cao, dẫn đến việc cấp tín dụng sai cho những người có rủi ro cao.

* Phương pháp F1 - Score là trung bình trọng số của Precision và Recall, thể hiện sự cân bằng, trong đó F1 - Score đạt giá trị tốt nhất là 1 và kém nhất là 0, chỉ số F1 - Score được xác định theo công thức sau:

|  |
| --- |
|  |

F1- Score cân bằng giữa việc dự đoán chính xác "tốt" và xác định chính xác khách hàng có điểm cao. Nết đạt giá trị cao: Mô hình có hiệu quả tốt trong việc phân biệt khách hàng có điểm tín dụng cao và thấp và giá trị thấp: Mô hình cần được cải thiện để nâng cao hiệu quả phân loại.

Các chỉ số đánh giá Accuracy, Precision, Recall, và F1-Score là những chỉ số quan trọng để đánh giá hiệu quả mô hình tín dụng. Việc sử dụng các chỉ số này giúp lựa chọn mô hình phù hợp và cải thiện hiệu quả đánh giá tín dụng.

## 4.4. Các tham số và môi trường cài đặt

*Môi trường Phần mềm:*

Ngôn Ngữ Lập Trình: Sử dụng Python trong môi trường Jupyter Notebook, một công cụ phổ biến trong khoa học dữ liệu cho phép thực hiện mã, ghi chú và hiển thị kết quả một cách tương tác.

*Thư Viện Sử Dụng:*

* pandas và numpy cho việc xử lý và phân tích dữ liệu.
* matplotlib và seaborn cho việc trực quan hóa dữ liệu.
* sklearn cho các mô hình học máy như Decision Tree, KNN, Random Forest, SVM và Adaboost, cũng như các công cụ tiền xử lý dữ liệu như MinMaxScaler.
* tensorflow.keras được sử dụng cho việc xây dựng và huấn luyện mô hình mạng nơ-ron tích chập (CNN) và mạng nơ-ron hồi quy (RNN).

*Môi trường Phần cứng:* Môi trường thử nghiệm của nhóm chúng tôi như sau: CPU Intel Core i5-8265U, RAM: 2.80GHz.

*Thiết Lập Thí Nghiệm:* Được thực hiện trên tập dữ liệu đã được tiền xử lý, với việc chia tách tập dữ liệu huấn luyện và kiểm thử mô hình, và sử dụng các phương pháp đánh giá như accuracy, precision, recall, và F1-score để đánh giá hiệu quả của mô hình.

## 4.5. Các phương pháp cơ sở

Trong phần nghiên cứu này, nhóm chúng tôi thực hiện so sánh 7 mô hình để đánh giá tính hiệu quả và chính xác. Tất cả các phương pháp được thực hiện dưới cùng một điều kiện thí nghiệm để đảm bảo công bằng trong so sánh:

* Phương pháp phân lớp: Decision Tree, K-Nearest Neighbor, Random Forest, XGBoost, Naive Bayes
* Mô hình học sâu: Convolutional Neural Network, Recurrent Neural Network

Các thí nghiệm được thực hiện trên cùng một tập dữ liệu, sử dụng cùng một quy trình tiền xử lý và chia dữ liệu để đảm bảo tính khách quan. Mỗi phương pháp được đánh giá dựa trên các chỉ số như accuracy, precision, recall, và F1-score để có cái nhìn toàn diện về hiệu suất của chúng.

## 4.6. Kết quả và phân tích

*4.6.1. Các kỹ thuật học máy truyền thống*

Trong bài nghiên cứu, nhóm chúng tôi đã xây dựng, thử nghiệm các mô hình học máy truyền thống sau: Decision Tree (Cây quyết định), K-Nearest Neighbor (K- láng giềng gần nhất), Random Forest (Rừng ngẫu nhiên), XGBoost, Naive Bayes và có kết quả được trình bày trong bảng dưới đây:

**Bảng 3.** Kết quả đánh giá các mô hình học máy truyền thống

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| 1 | **Decision Tree** | 0.855382 | 0.854924 | 0.855382 | 0.851688 |
| 2 | **K-Nearest Neighbor** | 0.712090 | 0.707868 | 0.712090 | 0.702121 |
| **3** | **Random Forest** | **0.885779** | **0.887773** | **0.885779** | **0.882678** |
| 4 | **XGBoost** | 0.835985 | 0.837708 | 0.835985 | 0.830886 |
| 5 | **Naive Bayes** | 0.617953 | 0.611433 | 0.617953 | 0.585627 |

Qua bảng phân tích của các mô hình học máy cổ khi sử dụng lần lượt các chỉ số đánh giá hiệu suất Accuracy, Precision, Recall, F1-score và kết quả cho thấy mô hình Random Forest (Rừng ngẫu nhiên) có hiệu suất đạt tốt nhất khi các chỉ số đều đạt trên 88%, cao hơn các mô hình còn lại khoảng từ 0-30% tùy vào từng phương pháp so sánh. Bên cạnh đó, Naive Bayes và K-Nearest Neighbor hai mô hình có hiệu suất được nhìn nhận là thấp nhất trong bảng phân tích với kết quả lần lượt là hơn 60% và hơn 70% ở cả bốn chỉ số đánh giá.

*4.6.2 Các kỹ thuật học sâu*

Nhóm chúng tôi xây dựng mô hình Convolutional Neural Network (Mạng neural tích chập sâu) và Recurrent Neural Network (Mạng nơ ron hồi quy), thu được kết quả sau:

**Bảng 4.** Kết quả đánh giá các mô hình học sâu

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **Thời gian** |
| 1 | **CNN** | 0.686358 | 0.679902 | 0.686358 | 0.677860 | **1 phút 56s** |
| 2 | **RNN** | 0.675702 | 0.671962 | 0.675702 | 0.663117 |  |

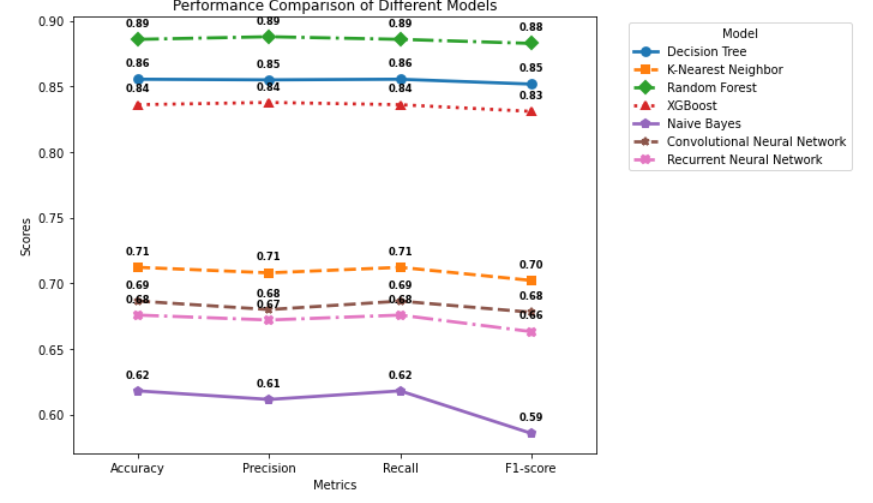
Hai mô hình ở bảng 4 đưa ra kết quả là hiệu suất ở cả 2 phương pháp đánh giá xấp xỉ nhau. Tuy nhiên, mô hình Convolutional Neural Network đã có mức % nhỉnh hơn một chút so với mô hình còn lại ví dụ như Accuracy của CNN chỉ cao hơn 1% so với RNN. Quan sát kết quả thu được có thể thấy RNN đưa ra kết quả không cao và thời gian thực nghiệm dài hơn CNN khoảng

*4.6.3. So sánh các mô hình*

**Bảng 5.** So sánh các mô hình

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| 1 | **DT** | 0.855382 | 0.854924 | 0.855382 | 0.851688 |
| 2 | **KNN** | 0.712090 | 0.707868 | 0.712090 | 0.702121 |
| 3 | **RF** | **0.885779** | **0.887773** | **0.885779** | **0.882678** |
| 4 | **XGB** | 0.835985 | 0.837708 | 0.835985 | 0.830886 |
| 5 | **NB** | 0.617953 | 0.611433 | 0.617953 | 0.585627 |
| 7 | **CNN** | 0.686358 | 0.679902 | 0.686358 | 0.677860 |
| 8 | **RNN** | 0.675702 | 0.671962 | 0.675702 | 0.663117 |

Dựa vào bảng tổng hợp các chỉ số đánh giá hiệu suất của các mô hình trên, ta nhận thấy rằng mô hình Random Forest có tỷ lệ hiệu suất cao nhất trong 7 mô hình còn lại lần lượt với tất cả các chỉ số đánh giá đều trên 88%.



**Hình 7.** Đánh giá hiệu quả của các mô hình

Thông qua biểu đồ trên ta nhận thấy rằng các phương pháp đánh giá trên mô hình RF có hiệu suất cao nhất (khoảng 89%), tiếp đó là mô hình DT với hiệu suất chỉ kém 2-3%. Ngoài ra, mô hình XGB có hiệu suất cao thứ ba (khoảng 84%). Với phương pháp đánh giá hiệu suất bằng 4 chỉ số accuracy, precision, recall, f1-score thì mô hình RNN có kết quả hiệu suất là thấp nhất (trên 60%).

# V. KẾT LUẬN

Tóm lại, bài báo này đã so sánh nhiều cách tiếp cận khác nhau của mô hình chấm điểm tín dụng, mô hình truyền thống với 5 mô hình cơ sở là DT, KNN, RF, XGB và NB; mô hình học sâu CNN và RNN. Việc đánh giá trên nhiều mô hình lựa chọn ra mô hình có khả năng chấm điểm tín dụng chính xác nhất. Các thí nghiệm đã chỉ ra rằng mô hình Random Forest được huấn luyện vượt trội hơn các mô hình khác với các chỉ số đánh giá hiệu năng accury, precision, recall, f1-score đều trên 88%. Đối với phương pháp xây dựng tính năng, mỗi phương pháp thể hiện thế mạnh của mình khi được đào tạo bởi các mô hình khác nhau. Tuy nhiên, cần nhấn mạnh rằng không có mô hình tính điểm tín dụng lý tưởng cũng như không có kỹ thuật tốt nhất được sử dụng để xây dựng mô hình chấm điểm tín dụng; và kỹ thuật tốt cho tập dữ liệu này chưa chắc tốt trên tập dữ liệu khác. Kỹ thuật này hơn kỹ thuật khác khả năng dự đoán trong các trường hợp cụ thể khác nhau. “Kĩ thuật tốt nhất” phụ thuộc vào từng yêu cầu cụ thể, cấu trúc của dữ liệu, các chức năng của ứng dụng, mức độ tách biệt các lớp bằng mục tiêu phân loại.

Trong công việc trong tương lai, để cải thiện hiệu suất của các mô hình được đề xuất, nhiều mô hình và kỹ thuật học máy sẽ được khám phá, chẳng hạn như Deep Forest và xếp chồng mô hình. Ngoài ra, chúng tôi đang có kế hoạch thử nghiệm các mô hình được đề xuất với nhiều bộ dữ liệu chấm điểm tín dụng hơn để xác minh tính mạnh mẽ của chúng.