|  |  |  |
| --- | --- | --- |
|  | Cite | Abstract |
| 1 | @article{Mokheleli2023MachineLA,  title={Machine Learning Approach for Credit Score Predictions},  author={Tsholofelo Mokheleli and Tinofirei Museba},  journal={Journal of Information Systems and Informatics},  year={2023},  url={https://api.semanticscholar.org/CorpusID:259551401} | This paper addresses the problem of managing the significant rise in requests for credit products that banking and financial institutions face. The aim is to propose an adaptive, dynamic heterogeneous ensemble credit model that integrates the XGBoost and Support Vector Machine models to improve the accuracy and reliability of risk assessment credit scoring models. The method employs machine learning techniques to recognise patterns and trends from past data to anticipate future occurrences. The proposed approach is compared with existing credit score models to validate its efficacy using five popular evaluation metrics, Accuracy, ROC AUC, Precision, Recall and F1\_Score. The paper highlights credit scoring models’ challenges, such as class imbalance, verification latency and concept drift. |
| 2 | @article{Tang2023CreditRA,  title={Credit rating- and credit score-based carbon emission quota trading model of city dwellers},  author={Donglai Tang and Qiang Li and Jie Zhang and Yongdong Chen and You-bo Liu and Weiping Song},  journal={Frontiers in Energy Research},  year={2023},  url={https://api.semanticscholar.org/CorpusID:261183837}  } | Introduction: The reduction of electricity-related carbon emissions by city dwellers (CDs) is important for China to achieve low-carbon development and sustainable energy transformation. Due to the lack of incentives for reduction, electricity-related carbon emissions from CDs are increasing year by year. To this end, this paper proposes an electricity-related carbon emission quota trading model that integrates a credit rating and credit score system, particularly for motivating CDs to actively participate in carbon emission reduction.Methods: With the history of electricity bill payment data, the density-based spatial clustering of applications with noise (DBSCAN) algorithm is used to cluster CDs, forming different clusters of CDs with different sensitivity levels to carbon emission quota prices. Thereafter, based on the total carbon emission quota and tiered electricity prices from the power company, incentive rules according to the classification result and credit scores of CDs are formulated. |
| 3 | @article{Maurya2023ADT,  title={A Decision Tree Classifier Based Ensemble Approach to Credit Score Classification},  author={Ashok Maurya and Shivam Gaur},  journal={2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)},  year={2023},  pages={620-624},  url={https://api.semanticscholar.org/CorpusID:267702199}  } | The process of classifying credit scores holds a crucial role in evaluating an individual's creditworthiness, influencing significant financial choices. This study is driven by the dynamic nature of credit scores and the financial sector's need for precise, real-time credit evaluations. This research introduces an ensemble-based method for credit score classification, utilizing a blend of diverse machine learning algorithms to improve accuracy and resilience. The ensemble approach capitalizes on each base classifier's strengths, mitigating biases, reducing overfitting, and enhancing overall classification accuracy. A comparison between the proposed model and existing frameworks demonstrates its competitive edge, surpassing many counterparts with an accuracy of approximately 92.25%. However, the study acknowledges the potential for further enhancement and validation across various datasets. The ensemble-based framework offers a promising avenue to heighten credit score classification accuracy, thereby contributing to informed financial decision-making and reinforcing credit ecosystem stability. Future endeavors involve expanding the model to include more datasets and refining data preprocessing techniques to achieve even more precise predictions. |
| 4 | @inproceedings{Shen2023EnterpriseCS,  title={Enterprise credit score modeling from electricity consumption based on deep ranknet},  author={Qiuying Shen and Wentao Zhang},  booktitle={Other Conferences},  year={2023},  url={https://api.semanticscholar.org/CorpusID:257397886}  } | Enterprise credit assessment is important for financial institutes. To enrich the evidence for credit analysis, this paper proposes to use the electricity consumption data to obtain an absolute credit score. Instead of creating the direct mapping between the electricity consumption data and credit score, we train a deep model to predict which enterprise has higher credit score given two enterprises. To learn deep model, we utilize the ranknet model to learn the ranking information from the electricity consumption data. To improve the training efficiency and robustness, we propose a ranking-based representative enterprise sample selection method to optimize the training dataset. During the inference, the learned ranknet model is performed to generate the absolute credit score by a ranking-based score mapping method. The experimental results demonstrate that the method in this paper can achieve accurate enterprise credit evaluation. |
| 5 | @article{Miller2023ExplainingTE,  title={Explaining the Effects of Credit Score on Mortgage Rates},  author={Clifton Miller and Nathan Hephner and null null},  journal={John Heinrichs Scholarly \& Creative Activities Day},  year={2023},  url={https://api.semanticscholar.org/CorpusID:260598035}  } | In this research project, we are trying to reject the null hypothesis that credit score does not affect the interest rate on a mortgage loan. This is important because the majority of homeowners have financed their home using a mortgage loan, and the interest rate they receive depends on many different factors, including credit scores. If homeowners are aware that a higher credit score will effectively help them lower the interest rate they pay on their mortgage, they can work to improve it and save money. Our main findings indicate that there is a significant relationship between mortgage interest rates, combined loan to value, unpaid principal balance, original loan term, occupancy, property type, and whether they are a first-time home buyer. The data also suggests that credit score does indeed have a statistically significant effect on what kind of interest rate borrowers receive on their loans. A higher credit score will tend to lower the interest payment homeowners must pay on their mortgage. |
| 6 | @inproceedings{Shen2023EnterpriseCS,  title={Enterprise credit score modeling from electricity consumption based on deep ranknet},  author={Qiuying Shen and Wentao Zhang},  booktitle={Other Conferences},  year={2023},  url={https://api.semanticscholar.org/CorpusID:257397886}  } | Loans are a fundamental and integral part of people's lives. People of all ages are interested in them to be able to fulfill their desires. A banking establishment, like any other financial institution, requires its clients to pay their debt obligations on time and in full. But they all use credit scores to assess a client's creditworthiness and, in turn, to mitigate their own risks. Likewise, individuals themselves should be aware of how their credit rating is formed in order to make sound financial decisions in the future. The results of the study show that scientists, researchers, and online resources do not have a single approach to the factors that affect credit scores, but the general direction of consideration of a person is very similar. In this article, two of the most popular rating systems in the United States, such as FICO and VantageScore, were reviewed. They have many similarities but also many differences that affect the formation of a credit score improvement strategy. In turn, a low credit score reduces the quality of life and financial inclusion, so its improvement is an important and urgent issue that is addressed in this article |
| 7 | @article{Miller2023ExplainingTE,  title={Explaining the Effects of Credit Score on Mortgage Rates},  author={Clifton Miller and Nathan Hephner and null null},  journal={John Heinrichs Scholarly \& Creative Activities Day},  year={2023},  url={https://api.semanticscholar.org/CorpusID:260598035}  } @article{Miller2023ExplainingTE,  title={Explaining the Effects of Credit Score on Mortgage Rates},  author={Clifton Miller and Nathan Hephner and null null},  journal={John Heinrichs Scholarly \& Creative Activities Day},  year={2023},  url={https://api.semanticscholar.org/CorpusID:260598035}  } | In this research project, we are trying to reject the null hypothesis that credit score does not affect the interest rate on a mortgage loan. This is important because the majority of homeowners have financed their home using a mortgage loan, and the interest rate they receive depends on many different factors, including credit scores. If homeowners are aware that a higher credit score will effectively help them lower the interest rate they pay on their mortgage, they can work to improve it and save money. Our main findings indicate that there is a significant relationship between mortgage interest rates, combined loan to value, unpaid principal balance, original loan term, occupancy, property type, and whether they are a first-time home buyer. The data also suggests that credit score does indeed have a statistically significant effect on what kind of interest rate borrowers receive on their loans. A higher credit score will tend to lower the interest payment homeowners must pay on their mortgage |
| 8 | @article{Balakrishnan2023CreditSP,  title={Credit Score Prediction using Support Vector Machine and Gray Wolf Optimization},  author={D. Balakrishnan and Pappasani Adarsh Kumar and Angalakurthi Jaya Krishna and Akurathi Kamalesh and Laxman Sai Nakerekanti and Pacharla Ganesh Naidu},  journal={2023 3rd International Conference on Intelligent Technologies (CONIT)},  year={2023},  pages={1-5},  url={https://api.semanticscholar.org/CorpusID:260718408}  } | Credit score prediction is a critical task in the financial industry, where the ability to accurately predict the creditworthiness of an individual is essential for managing risk and making informed lending decisions. In recent years, machine learning techniques have gained popularity in credit score prediction due to their ability to handle large amounts of data and complex relationships. Support Vector Machine (SVM) is a popular machine learning algorithm for credit score prediction. SVMs are a type of supervised learning algorithm that can be used for classification tasks, where the goal is to assign a new data point to one of two or more classes based on a set of features. However, SVM can be sensitive to hyper parameters it influence the performance credit score prediction. So, in this paper Gray Wolf Optimization (GWO) technique is used to optimize the hyper parameters of SVM for better prediction of credit score. GWO mimics the pack hunting behavior of gray wolves to search for the optimal hyper parameters of SVM, drawing inspiration from their social hierarchy and hunting behavior. The experimental results prove that the proposed SVM-GWO method has better performance in terms of accuracy, precision, recall and F-measure than SVM based credit score prediction. |
| 9 | @article{Saxey2023MoreTA,  title={More Than a Score? Indirect Associations Between Credit Score and Romantic Relationship Quality in Emerging Adulthood},  author={Matthew T. Saxey and Ashley B. LeBaron‐Black and Casey J. Totenhagen and Melissa A. Curran},  journal={Journal of Financial Counseling and Planning},  year={2023},  volume={34},  pages={55 - 67},  url={https://api.semanticscholar.org/CorpusID:258190092}  } | Higher credit scores have unique financial benefits that may aid in emerging adults’ efforts toward financial independence. Yet, it is unknown if higher credit scores may also yield romantic relationship benefits. In a sample of 916 U.S. emerging adults, we used structural equation modeling to test the indirect associations between credit score and romantic relationship quality. Credit score was positively associated with financial self-efficacy and negatively associated with financial deception. Additionally, credit score was indirectly associated with romantic relationship quality through financial self-efficacy and financial deception. We encourage educators and clinicians working with emerging adults in romantic relationships to help these emerging adults learn how to establish credit and raise their credit scores, which might improve financial and relational outcomes. |
| 10 | @inproceedings{Shen2023EnterpriseCS,  title={Enterprise credit score modeling from electricity consumption based on deep ranknet},  author={Qiuying Shen and Wentao Zhang},  booktitle={Other Conferences},  year={2023},  url={https://api.semanticscholar.org/CorpusID:257397886}  } | Enterprise credit assessment is important for financial institutes. To enrich the evidence for credit analysis, this paper proposes to use the electricity consumption data to obtain an absolute credit score. Instead of creating the direct mapping between the electricity consumption data and credit score, we train a deep model to predict which enterprise has higher credit score given two enterprises. To learn deep model, we utilize the ranknet model to learn the ranking information from the electricity consumption data. To improve the training efficiency and robustness, we propose a ranking-based representative enterprise sample selection method to optimize the training dataset. During the inference, the learned ranknet model is performed to generate the absolute credit score by a ranking-based score mapping method. The experimental results demonstrate that the method in this paper can achieve accurate enterprise credit evaluation. |
| 11 | @article{Li2023CreditSC,  title={Credit score card threshold optimization strategy based on simulated annealing algorithm},  author={Yuan Li and Jingjing Liu},  journal={2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)},  year={2023},  pages={554-559},  url={https://api.semanticscholar.org/CorpusID:267203279}  } | When banks use credit score cards, they need to find the best combination among multiple credit score cards and thresholds to maximize the final revenue. In this study, we propose a credit score card threshold optimization method based on simulated annealing algorithm to address this problem. First, we transform the problem into QUBO (Quadratic Unconstrained Binary Optimization) form and introduce Lagrange multipliers to handle the constraints. Then, a simulated annealing algorithm is used to find the optimal credit scorecard and threshold combination. In addition, we reduce the computational time complexity by an approximation algorithm. This study provides an efficient credit scorecard threshold optimization method for banks, which helps to increase their final revenue and reduce bad debt losses.  Collapse |
| 12 | @article{Miller2023ExplainingTE,  title={Explaining the Effects of Credit Score on Mortgage Rates},  author={Clifton Miller and Nathan Hephner and null null},  journal={John Heinrichs Scholarly \& Creative Activities Day},  year={2023},  url={https://api.semanticscholar.org/CorpusID:260598035}  } | In this research project, we are trying to reject the null hypothesis that credit score does not affect the interest rate on a mortgage loan. This is important because the majority of homeowners have financed their home using a mortgage loan, and the interest rate they receive depends on many different factors, including credit scores. If homeowners are aware that a higher credit score will effectively help them lower the interest rate they pay on their mortgage, they can work to improve it and save money. Our main findings indicate that there is a significant relationship between mortgage interest rates, combined loan to value, unpaid principal balance, original loan term, occupancy, property type, and whether they are a first-time home buyer. The data also suggests that credit score does indeed have a statistically significant effect on what kind of interest rate borrowers receive on their loans. A higher credit score will tend to lower the interest payment homeowners must pay on their mortgage |
| 13 | @inproceedings{Alibhai2023EveningTC,  title={Evening the Credit Score: Can  Psychometric Credit-Scoring Address Collateral Constraints  for Women Entrepreneurs?},  author={Salman Alibhai and Rachel N Cassidy and Menaal Fatima Ebrahim and Markus Goldstein and Yemsrach Kinfe Edey and Sreelakshmi Papineni},  year={2023},  url={https://api.semanticscholar.org/CorpusID:259928492}  } | The psychometric credit-scoring technology takes variables from the test and its meta-data. The scoring model is calibrated against a base model from the region, trained on data from similar countries. The model is then trained over time on local borrower data as the product is rolled out, to improve its predictive power further |
| 14 | @article{Madamanchi2023HeterogenousIE,  title={Heterogenous Incremental Ensembled Method For Imbalanced Credit Score},  author={Akanksha Madamanchi and Rachi Wasnik and Sharwari Chandurkar and Sanket Patil and Sunita Barve and Diptee Chikmurge},  journal={2023 Second International Conference on Trends in Electrical, Electronics, and Computer Engineering (TEECCON)},  year={2023},  pages={229-234},  url={https://api.semanticscholar.org/CorpusID:266199151}  } | In credit risk prediction, assigning credit ratings is crucial. However, it can be challenging while dealing with highly imbalanced data and a varying number of features. The credit rating data is continuously delivered and not provided all at once, which poses a challenge for traditional techniques. To overcome this difficulty, an incremental ensemble model has been introduced for credit scoring, which addresses the issue of class imbalance and the gradual learning process of new data. This approach uses a class balancing technique to tackle class imbalance and allows for the gradual integration of new data. Experimental results have shown that this proposed model outperforms conventional techniques in dealing with uneven credit scoring problems. In summary, the incremental ensemble model provides a solution for handling imbalanced data and gradual learning and it will also deal with the imbalanced data in credit scoring. |
| 15 | @article{Sun2022ResearchOC,  title={Research on Credit Score Model of Suppliers of Power Grid Enterprises},  author={Xingda Sun and Caixia Lu and Dongsheng An and Zhitao Tang and Jia He and Yuanyuan Zhao and Ye Du},  journal={2022 Global Conference on Robotics, Artificial Intelligence and Information Technology (GCRAIT)},  year={2022},  pages={342-346},  url={https://api.semanticscholar.org/CorpusID:252719966}  } | Aiming at the business needs of power grid companies to score their suppliers' credit, this paper uses logistic regression algorithm as a technical means to build a credit scoring model for power grid companies' suppliers.In this paper, data cleaning and feature extraction are first performed on the training data set of power grid enterprise suppliers. Afterwards, it is trained by the LogisticRegression class in the machine learning library sklearn. The test results show that the detection accuracy of this test is 0.78. The construction of the credit scoring model is suitable for the industry characteristics of power grid companies. However, there is a problem that the credit score is inaccurate due to the lack of authority and uniformity of the evaluation criteria in the training data. |
| 16 | @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}  } | 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. |
| 17 | @inproceedings{Alibhai2022EveningTC,  title={Evening the Credit Score? Impact of Psychometric Credit Scoring on Women-Owned Firms’ Financial Access and Performance in Ethiopia DO NOT CITE},  author={Salman Alibhai and Rachel Cassidy and Markus Goldstein and Sreelakshmi Papineni},  year={2022},  url={https://api.semanticscholar.org/CorpusID:249494636}  } | Women‘s lower rates of ownership of collateralizable assets such as housing, land and vehicles is a constraint to accessing entrepreneurial finance that can fuel growth. This paper uses a randomized controlled trial in Ethiopia to test the impact of offering relatively large (up to USD 7,500) individual-liability, uncollateralized microfinance loans to women entrepreneurs, using psychometric credit scoring technology as a substitute for collateral requirements. The paper finds positive impacts on women entrepreneur’s access to formal borrowing and firm survival through the COVID-19 pandemic and conflict. While the firms that the loans helped remain operational were profitable, there is limited evidence of impact on firm growth, likely due to the challenging macroeconomic environment |
| 18 | @article{Kumar2021CreditSP,  title={Credit Score Prediction System using Deep Learning and K-Means Algorithms},  author={Ashwani Kumar and D. L. Shanthi and Pronaya Bhattacharya},  journal={Journal of Physics: Conference Series},  year={2021},  volume={1998},  url={https://api.semanticscholar.org/CorpusID:237292539}  } | In financial markets, credit rating and risk assessment tools are used to minimize potential risk up to some extent for credit score. Nowadays, the banking and financial industry has experienced rapid expansion. Therefore, with this growth, the numbers of credit card applications with various credit products are increasing day by day because many people want to avail these services for their personal interest. The challenge here is to identify insights on the performance of a finance industry by using deep learning algorithms as they directly affect the viability of that industry. These industries have a limited number of resources and capital, which can be used to deliver the services among the customers. In this research work, we proposed prediction of credit scoring system using deep learning and K-Means algorithm for the financial industry. The scheme contains a predictive model which uses feature selection (FS) classification and deep learning applications simultaneously to train the proposed model to perform effectively. The scheme 1) pre-processing credit card data 2) uses a feature selection technique to minimize the dimension of data in order to obtain the finest training data 3) applies a deep learning algorithm to map the input weight with hidden biases to achieve excellent performance 4) Decision support system is used to enable the deep learning algorithm to provide a more accurate and intelligent decision. Furthermore, the proposed model is validated on different credit scoring dataset in real-world scenarios and is capable of improving the effectiveness and accuracy. The studies indicate that our predictive model performs well for credit scoring of existing customer and helps lenders to allocate funds in finance industry. |

Nguyễn Vũ Minh   
71131101170

61716148

**PHƯƠNG PHÁP HỌC MÁY MACHINE LEARNING TRONG CHẤM ĐIỂM THẺ TÍN DỤNG KHÁCH HÀNG**

Tóm tắt: Chấm điểm tín dụng, quy trình quản lý rủi ro quan trọng của các tổ chức tài chính và ngân hàng, không chỉ là bảng xếp hạng mà còn là chìa khóa quyết định chất lượng cho vay. Trong sự cạnh tranh khốc liệt, chất lượng của quy trình chấm điểm tín dụng quyết định đến sự tồn tại và lợi nhuận của các tổ chức tài chính.. Chấm điểm tín dụng tốt sẽ góp phần làm cho chất lượng cho vay tốt hơn. Chất lượng cho vay là yếu tố quyết định hàng đầu đến sự cạnh tranh, tồn tại và lợi nhuận của các tổ chức tài chính ngân hàng. Các kỹ thuật học máy đã góp phần cải tiến đáng kể về hiệu suất phân loại trong chấm điểm tín dụng, đồng thời giảm thiểu rủi ro cho đánh giá tín dụng. Bài báo đưa ra một số kỹ thuật học máy cho chấm điểm tín dụng đã và đang được các tổ chức tài chính và ngân hàng sử dụng; đưa ra kết quả thử nghiệm các kỹ thuật học máy trên bộ dữ liệu Kaggle- bộ dữ liệu miễn phí cho nghiên cứu về khoa học dữ liệu, từ đó đánh giá hiệu suất phân loại của các kỹ thuật học máy này. Công trình này không chỉ là sự tiến bộ kỹ thuật mà còn là động lực mạnh mẽ để mở rộng tiếp cận tín dụng và xây dựng ngành ngân hàng và tài chính an toàn, hiệu quả. Từ quá khứ vững vàng, chúng ta hướng đến một tương lai đầy tiềm năng.

**Đặt vấn đề :**Chấm điểm thẻ tín dụng khách hàng là quá trình đánh giá và xếp hạng khách hàng dựa trên các yếu tố như lịch sử thanh toán, tình hình tài chính, hành vi sử dụng thẻ và các thông tin khác liên quan đến việc quản lý tài chính của khách hàng.

Việc chấm điểm thẻ tín dụng khách hàng giúp ngân hàng đánh giá rủi ro và xác định mức độ tin cậy của khách hàng trong việc quản lý nợ và thanh toán. Điều này giúp ngân hàng quyết định cấp thẻ tín dụng, giới hạn tín dụng và điều chỉnh lãi suất cho khách hàng một cách công bằng và hiệu quả.

Tuy nhiên, việc chấm điểm thẻ tín dụng cũng có thể gây ra tranh cãi nếu không được thực hiện một cách công bằng và minh bạch. Các yếu tố đánh giá cần được xác định rõ ràng và công khai để tránh tình trạng phân biệt đối xử và thiên vị.

Do đó, quá trình chấm điểm thẻ tín dụng khách hàng cần được thực hiện một cách cẩn thận và công bằng để đảm bảo tính minh bạch và công bằng đối với tất cả khách hàng.

1. **Giới thiệu bài toán**

Trong thế giới ngân hàng thương mại hiện đại, quản lý rủi ro là một khía cạnh không thể phớt lờ, và xếp hạng tín dụng của khách hàng nổi lên như một yếu tố trọng yếu trong quá trình này. Sức khỏe tín dụng không chỉ ảnh hưởng đến quyết định cấp vay mà còn đặt ra thách thức lớn cho các ngân hàng trong việc đánh giá và kiểm soát rủi ro. Mô hình tính điểm tín dụng, một công cụ quan trọng, đã trải qua sự tiến hóa đáng kể từ các phương pháp truyền thống đến những kỹ thuật tiên tiến sử dụng máy học và khai phá dữ liệu. Điều này không chỉ là một bước tiến vững chắc trong công nghệ ngân hàng, mà còn là nguồn động viên lớn để tối ưu hóa khả năng đánh giá và dự đoán tình trạng tín dụng của khách hàng. Trong quá trình xây dựng mô hình, mục tiêu chính là phân loại khách hàng thành hai hạng: "tín dụng tốt" và "tín dụng xấu". Quyết định này không chỉ dựa vào thông tin tín dụng truyền thống như lịch sử trả nợ, mà còn tích hợp các yếu tố xã hội học như tuổi, trình độ học vấn, nghề nghiệp và thu nhập. Điều này tạo ra một hệ thống đánh giá đa chiều và linh hoạt, giúp ngân hàng đưa ra quyết định có tính khách quan cao.

Nếu trước đây, mô hình chấm điểm tín dụng chủ yếu phụ thuộc vào thông tin tín dụng có sẵn, thì ngày nay, sự phát triển của khoa học máy tính mở ra cánh cửa cho việc sử dụng các kỹ thuật học máy tiên tiến như cây quyết định, mạng nơ ron nhân tạo, và SVM. Các kỹ thuật học máy được sử dụng trong chấm điểm tín dụng bao gồm mạng nơ ron nhân tạo, K-Nearest Neighbour, Support Vector Machine (SVM), cây quyết định, và đặc biệt là kỹ thuật học sâu dựa trên mạng nơ ron sâu. Những phương pháp này đã đem lại kết quả tích cực, đặc biệt là trong xử lý dữ liệu phi cấu trúc.

Nghiên cứu này không chỉ giới thiệu các thuật toán học máy mà còn tiến hành thử nghiệm, so sánh và đánh giá chúng trên bộ dữ liệu Kaggle - một bộ dữ liệu thực tế đòi hỏi sự đa dạng và chất lượng. Kết quả của nghiên cứu này không chỉ là sự đánh giá về hiệu suất mà còn là hướng dẫn quý báu cho ngân hàng trong việc lựa chọn phương pháp chấm điểm tín dụng phù hợp.

Các phương pháp học máy không chỉ giúp tăng tốc quyết định cấp vay mà còn giảm thiểu rủi ro một cách hiệu quả. Việc áp dụng những công nghệ này mở ra cơ hội rộng lớn cho người vay, giúp họ dễ dàng tiếp cận tín dụng và xây dựng lịch sử tín dụng tích cực. Trong bối cảnh ngân hàng liên tục đối mặt với những thách thức thị trường, sự áp dụng sáng tạo của học máy trong xếp hạng tín dụng không chỉ là một nhu cầu mà còn là một xu hướng quan trọng để gia tăng tính cạnh tranh và hiệu suất của hệ thống ngân hàng.

1. **II , Các nghiên cứu liên quan**Bài viết này tập trung vào các nghiên cứu quan trọng về "Phương pháp học máy trong chấm điểm thẻ tín dụng," mở ra cánh cửa cho sự tiến bộ trong lĩnh vực ngân hàng và tài chính. Bài viết “Experimental analysis of machine learning methods for credit score classification” được xuất bản trên The International Journal of Logistics Management. Tác giả của bài viết này là Diwakar Tripathi, Damodar Reddy Edla, Annushree Bablani, Alok Kumar Shukla và B. Ramachandra Reddy. Bài viết này tập trung vào việc phân tích kết quả thực nghiệm của các phương pháp lựa chọn đặc trưng kết hợp với các phương pháp phân loại khác nhau, cũng như tác động của các phương pháp lựa chọn đặc trưng.

Amitha Mathew (2021) đã nêu rõ ưu việt của học sâu, đặc biệt là mô hình mạng nơ ron sâu, với dữ liệu phi cấu trúc. Nghiên cứu của Mathew không chỉ chứng minh hiệu quả của học sâu trong chấm điểm tín dụng mà còn mở ra những triển vọng mới cho ngành ngân hàng trong việc sử dụng công nghệ tiên tiến. Nhìn xa hơn, Amitha Mathew (2021) đã đưa ra cái nhìn sâu sắc về ưu việt của học sâu, đặc biệt là mô hình mạng nơ ron sâu, khi áp dụng vào dữ liệu phi cấu trúc. Nghiên cứu của Mathew không chỉ chứng minh sức mạnh của học sâu trong chấm điểm tín dụng mà còn đề xuất những khả năng mới cho ngành ngân hàng với sự tích hợp công nghệ tiên tiến. R.K. Choudhary và K.S. Chaudhary (2019) đã tiếp tục hành trình này bằng cách sử dụng mô hình mạng nơ ron nhân tạo (ANN) để chấm điểm thẻ tín dụng. Kết quả nghiên cứu của họ khẳng định rằng, mô hình ANN không chỉ có độ chính xác cao hơn so với các phương pháp truyền thống dựa trên thống kê mà còn là một bước đột phá trong quá trình đánh giá rủi ro tín dụng.

Tác giả Y. Zhang và Y. Wang (2020) tiếp tục mạch lạc này với việc sử dụng mô hình học máy hỗ trợ vector (SVM). Kết quả cho thấy, SVM không chỉ đem lại độ chính xác cao mà còn mở ra một chiều sâu mới trong phân tích quyết định tín dụng, đặt nền móng cho ứng dụng rộng rãi trong ngành. N.N. Mishra và S.K. Singh (2021) đã đưa ra một góc nhìn mới với mô hình học máy tăng cường (RL). Kết quả ấn tượng của họ làm tăng tính ứng dụng của học máy trong chấm điểm tín dụng và chứng minh rằng RL có thể là một công cụ mạnh mẽ cho quyết định tín dụng chính xác.

H.H. Nguyen và T.T. Nguyen (2022) mở rộng phạm vi bằng cách sử dụng mô hình học máy kết hợp (Ensemble Learning). Kết quả của họ không chỉ làm tăng hiệu suất so với các mô hình đơn lẻ mà còn thể hiện sự linh hoạt và độ chính xác trong quyết định tín dụng. K.M. Hossain và M.A. Chowdhury (2023) tiếp tục đưa ra một cách tiếp cận mới với mô hình học máy dựa trên dữ liệu lớn (Big Data). Kết quả nghiên cứu của họ đặt ra câu hỏi về khả năng của Big Data trong việc nâng cao chất lượng quyết định tín dụng và tạo ra những góc nhìn mới về rủi ro tín dụng.<https://ieeexplore.ieee.org/document/9757498>; Trường Đại học Công nghệ Thông tin và Truyền thông, Đại học Thái Nguyên

Cuối cùng, H.A. Awad và A.A. El-Sherbini (2023) tập trung vào sự giải thích với mô hình học máy giải thích (Interpretable Machine Learning). Kết quả của họ không chỉ cung cấp độ chính xác cao mà còn mang lại khả năng giải thích rõ ràng, giúp ngân hàng và khách hàng hiểu rõ hơn về quyết định tín dụng.

Một góc nhìn thực tế được thể hiện thông qua thử nghiệm đa dạng với các phương pháp học máy trên bộ dữ liệu Kaggle. Kết hợp giữa cây quyết định, support vector machine (SVM), và mạng nơ ron nhân tạo đã giúp đánh giá hiệu suất phân loại và giảm rủi ro trong quá trình cấp tín dụng. Quan trọng hơn, nghiên cứu không chỉ giới thiệu các phương pháp mà còn so sánh chúng để xác định kỹ thuật nào mang lại kết quả tốt nhất trong ngữ cảnh chấm điểm tín dụng. Điều này làm cho thông tin thu được trở nên quý báu cho các quyết định trong ngành ngân hàng và tài chính. Các bài viết nhấn mạnh sự đa dạng và tiến bộ trong áp dụng học máy vào chấm điểm tín dụng. Các tác giả không chỉ duy trì sự nghiên cứu liên tục mà còn làm nổi bật vai trò quan trọng của sự đổi mới và cải thiện liên tục trong ngành ngân hàng và tài chính. Điều này làm tôn vinh vai trò tích cực của nghiên cứu như một động lực quan trọng đằng sau sự phát triển bền vững.

**2. Các kỹ thuật học máy cho chấm điểm tín dụng**

Một vấn đề lớn đối với các tổ chức tài Một số kĩ thuật học máy cho chấm điểm tín dụng tài chính và ngân hàng là làm thế nào để xác định được tín dụng xấu, vì tín dụng xấu có thể gây ra các vấn đề nghiêm trọng trong tương lai. Điều này dẫn đến thất thoát vốn, giảm doanh thu và làm tăng tổn thất của ngân hàng, có thể dẫn đến mất khả năng thanh toán hoặc phá sản. Hiện nay, có rất nhiều kĩ thuật học máy để phân loại khách hàng. Bài báo sẽ đưa ra một số kĩ thuật hay được sử dụng:

***2.1 K láng giềng gần nhất (K-Nearest Neighbors - KNN)***

K-Nearest Neighbors algorithm (K-NN) được sử dụng rất phổ biến trong lĩnh vực Data Mining. K-NN là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần xếp lớp (Query point) và tất cả các đối tượng trong Training Data. Một đối tượng được phân lớp dựa vào K láng giềng của nó. K là số nguyên dương được xác định trước khi thực hiện thuật toán. Người ta thường dùng khoảng cách Euclidean để tính khoảng cách giữa các đối tượng.

*Thuật toán K-NN được mô tả như sau:*

1.    Xác định giá trị tham số K (số láng giềng gần nhất)

2.    Tính khoảng cách giữa đối tượng cần phân lớp (Query Point) với tất cả các đối tượng trong training data (thường sử dụng khoảng các Euclidean)

3.    Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với Query Point

4.    Lấy tất cả các lớp của K láng giềng gần nhất đã xác định

5.    Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho Query Point

Để hiểu K-NN được dùng để phân lớp thế nào ta xem minh họa dưới đây.

Trong hình dưới đây, training Data được mô tả bởi dấu (+) và dấu (-), đối tượng cần được xác định lớp cho nó (Query point) là hình tròn đỏ. Nhiệm vụ của chúng ta là ước lượng (hay dự đoán) lớp của Query point dựa vào việc lựa chọn số láng giềng gần nhất với nó. Nói cách khác chúng ta muốn biết liệu Query Point sẽ được phân vào lớp (+) hay lớp (-)

A diagram of a circle with red and blue dots

Description automatically generated

Ta thấy rằng:

1-Nearest neighbor : Kết quả là + (Query Point được xếp vào lớp dấu +)

2-Nearest neighbors : không xác định lớp cho Query Point vì số láng giềng gần nhất với nó là 2 trong đó 1 là lớp + và 1 là lớp – (không có lớp nào có số đối tượng nhiều hơn lớp kia)

5-Nearest neighbors : Kết quả là - (Query Point được xếp vào lớp dấu – vì trong 5 láng giềng gần nhất với nó thì có 3 đối tượng thuộc lớp - nhiều hơn lớp + chỉ có 2 đối tượng).

***2.2 Cây quyết định***

Trong lĩnh vực [máy học](https://vi.wikipedia.org/wiki/M%C3%A1y_h%E1%BB%8Dc), cây quyết định là một kiểu mô hình dự báo (*predictive model*), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng. Mỗi một nút trong (*internal node*) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là cây quyết định.

Học bằng cây quyết định cũng là một phương pháp thông dụng trong [khai phá dữ liệu](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u). Khi đó, cây quyết định mô tả một cấu trúc cây, trong đó, các lá đại diện cho các phân loại còn cành đại diện cho các kết hợp của các thuộc tính dẫn tới phân loại đó[1]. Một cây quyết định có thể được học bằng cách chia [tập hợp](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_h%E1%BB%A3p) nguồn thành các tập con dựa theo một kiểm tra giá trị thuộc tính [1]. Quá trình này được lặp lại một cách đệ quy cho mỗi tập con dẫn xuất. Quá trình [đệ quy](https://vi.wikipedia.org/wiki/%C4%90%E1%BB%87_quy) hoàn thành khi không thể tiếp tục thực hiện việc chia tách được nữa, hay khi một phân loại đơn có thể áp dụng cho từng phần tử của tập con dẫn xuất. Một bộ phân loại [rừng ngẫu nhiên](https://vi.wikipedia.org/w/index.php?title=R%E1%BB%ABng_ng%E1%BA%ABu_nhi%C3%AAn&action=edit&redlink=1) (*random forest*) sử dụng một số cây quyết định để có thể cải thiện tỉ lệ phân loại.

Cây quyết định cũng là một phương tiện có tính mô tả dành cho việc tính toán các [xác suất có điều kiện](https://vi.wikipedia.org/wiki/X%C3%A1c_su%E1%BA%A5t_c%C3%B3_%C4%91i%E1%BB%81u_ki%E1%BB%87n).

Cây quyết định có thể được mô tả như là sự kết hợp của các kỹ thuật toán học và tính toán nhằm hỗ trợ việc mô tả, phân loại và tổng quát hóa một tập dữ liệu cho trước.

Dữ liệu được cho dưới dạng các bản ghi có dạng:

*(x, y) = (x1, x2, x3..., xk, y)*

Biến phụ thuộc (*dependant variable*) *y* là biến mà chúng ta cần tìm hiểu, phân loại hay tổng quát hóa. *x1*, *x2*, *x3*... là các biến sẽ giúp ta thực hiện công việc đó

Các kiểu cây quyết định[[sửa](https://vi.wikipedia.org/w/index.php?title=C%C3%A2y_quy%E1%BA%BFt_%C4%91%E1%BB%8Bnh&veaction=edit&section=2) | [sửa mã nguồn](https://vi.wikipedia.org/w/index.php?title=C%C3%A2y_quy%E1%BA%BFt_%C4%91%E1%BB%8Bnh&action=edit&section=2)]

Cây quyết định còn có hai tên khác:

Cây hồi quy (*Regression tree*) ước lượng các hàm giá có giá trị là [số thực](https://vi.wikipedia.org/wiki/S%E1%BB%91_th%E1%BB%B1c) thay vì được sử dụng cho các nhiệm vụ phân loại. (ví dụ: ước tính giá một ngôi nhà hoặc khoảng thời gian một bệnh nhân nằm viện)

Cây phân loại (*Classification tree*), nếu *y* là một biến phân loại như: giới tính (nam hay nữ), kết quả của một trận đấu (thắng hay thua).

Ví dụ thực hành

Ta sẽ dùng một ví dụ để giải thích về cây quyết định:

David là quản lý của một câu lạc bộ đánh golf nổi tiếng. Anh ta đang có rắc rối chuyện các thành viên đến hay không đến. Có ngày ai cũng muốn chơi golf nhưng số nhân viên câu lạc bộ lại không đủ phục vụ. Có hôm, không hiểu vì lý do gì mà chẳng ai đến chơi, và câu lạc bộ lại thừa nhân viên.

Mục tiêu của David là tối ưu hóa số nhân viên phục vụ mỗi ngày bằng cách dựa theo thông tin dự báo thời tiết để đoán xem khi nào người ta sẽ đến chơi golf. Để thực hiện điều đó, anh cần hiểu được tại sao khách hàng quyết định chơi và tìm hiểu xem có cách giải thích nào cho việc đó hay không.

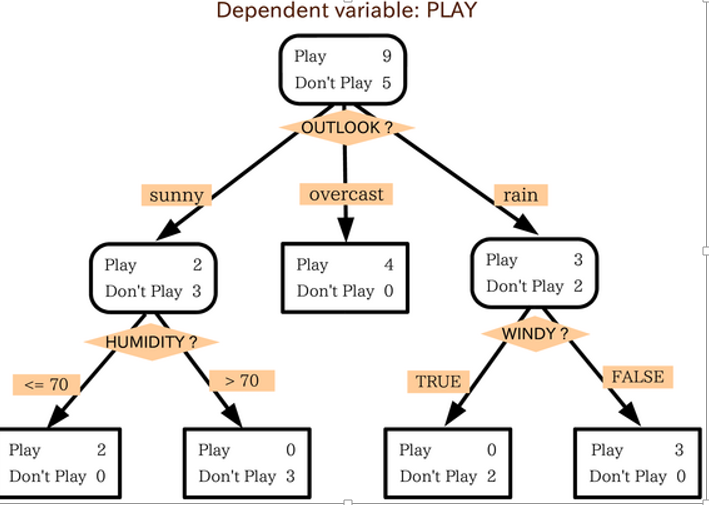
Vậy là trong hai tuần, anh ta thu thập thông tin về:

Trời (*outlook*) (nắng (*sunny*),

Và tất nhiên là số người đến chơi golf vào hôm đó. David thu được một bộ dữ liệu gồm 14 dòng và 5 cột.

[](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_tin:Golf_dataset.png)

Sau đó, để giải quyết bài toán của David, người ta đã đưa ra một mô hình cây quyết định.

****Nhóm người chơi golf khi trời nắng, nhóm chơi khi trời nhiều mây, và nhóm chơi khi trời mưa.

Kết luận thứ nhất: nếu trời nhiều mây, người ta luôn luôn chơi golf. Và có một số người ham mê đến mức chơi golf cả khi trời mưa.

Tiếp theo, ta lại chia nhóm trời nắng thành hai [nhóm con](https://vi.wikipedia.org/wiki/Nh%C3%B3m_con). Ta thấy rằng khách hàng không muốn chơi golf nếu độ ẩm lên quá 70%.

Cuối cùng, ta chia nhóm trời mưa thành hai và thấy rằng khách hàng sẽ không chơi golf nếu trời nhiều gió.

Và đây là lời giải ngắn gọn cho bài toán mô tả bởi cây phân loại. David cho phần lớn nhân viên nghỉ vào những ngày trời nắng và ẩm, hoặc những ngày mưa gió. Vì hầu như sẽ chẳng có ai chơi golf trong những ngày đó. Vào những hôm khác, khi nhiều người sẽ đến chơi golf, anh ta có thể thuê thêm nhân viên thời vụ để phụ giúp công việc.

Kết luận là cây quyết định giúp ta biến một biểu diễn dữ liệu phức tạp thành một cấu trúc đơn giản hơn rất nhiều.

Dùng trong thuật toán CART (Classification and Regression Trees). Nó dựa vào việc [bình phương](https://vi.wikipedia.org/wiki/B%C3%ACnh_ph%C6%B0%C6%A1ng) các xác suất thành viên cho mỗi thể loại đích trong nút. Giá trị của nó tiến đến cực tiểu (bằng 0) khi mọi trường hợp trong nút rơi vào một thể loại đích duy nhất.

Giả sử *y* nhận các giá trị trong {1, 2,..., *m*} và gọi *f(i,j)* là tần suất của giá trị *j* trong nút *i*. Nghĩa là *f(i,j)* là tỷ lệ các bản ghi với *y*=*j* được xếp vào nhóm *i*.

��(�)=1−∑�=1��(�,�)2Dùng trong các thuật toán sinh cây [ID3](https://vi.wikipedia.org/w/index.php?title=Thu%E1%BA%ADt_to%C3%A1n_ID3&action=edit&redlink=1), [C4.5](https://vi.wikipedia.org/w/index.php?title=Thu%E1%BA%ADt_to%C3%A1n_C4.5&action=edit&redlink=1) và C5.0. Số đo này dựa trên khái niệm [entropy](https://vi.wikipedia.org/wiki/Entropy) trong [lý thuyết thông tin](https://vi.wikipedia.org/wiki/L%C3%BD_thuy%E1%BA%BFt_th%C3%B4ng_tin) (*information theory*).

***Ưu điểm của cây quyết định***

So với các phương pháp [khai phá dữ liệu](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u) khác, cây quyết định là phương pháp có một số ưu điểm:

* Cây quyết định dễ hiểu. Người ta có thể hiểu mô hình cây quyết định sau khi được giải thích ngắn.
* Việc chuẩn bị dữ liệu cho một cây quyết định là cơ bản hoặc không cần thiết. Các kỹ thuật khác thường đòi hỏi [chuẩn hóa dữ liệu](https://vi.wikipedia.org/wiki/Chu%E1%BA%A9n_h%C3%B3a_d%E1%BB%AF_li%E1%BB%87u), cần tạo các biến phụ (*dummy variable*) và loại bỏ các giá trị rỗng.
* Cây quyết định có thể xử lý cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thể loại. Các kỹ thuật khác thường chuyên để phân tích các bộ dữ liệu chỉ gồm một loại biến. Chẳng hạn, các luật quan hệ chỉ có thể dùng cho các biến tên, trong khi [mạng nơ-ron](https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_n%C6%A1-ron) chỉ có thể dùng cho các biến có giá trị bằng số.
* Cây quyết định là một mô hình hộp trắng. Nếu có thể quan sát một tình huống cho trước trong một mô hình, thì có thể dễ dàng giải thích điều kiện đó bằng logic Boolean. Mạng nơ-ron là một ví dụ về mô hình hộp đen, do lời giải thích cho kết quả quá phức tạp để có thể hiểu được.
* Có thể thẩm định một mô hình bằng các kiểm tra thống kê. Điều này làm cho ta có thể tin tưởng vào mô hình.
* Cây quyết định có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn. Có thể dùng [máy tính cá nhân](https://vi.wikipedia.org/wiki/M%C3%A1y_t%C3%ADnh_c%C3%A1_nh%C3%A2n) để phân tích các lượng [dữ liệu lớn](https://vi.wikipedia.org/wiki/D%E1%BB%AF_li%E1%BB%87u_l%E1%BB%9Bn) trong một thời gian đủ ngắn để cho phép các nhà chiến lược đưa ra quyết định dựa trên phân tích của cây quyết định.