**Đề tài: Credit Score Classification**

**Họ tên: Vũ Quỳnh Mai**

**Msv: 71131101167**

**Lớp: KTDL11**

|  |  |  |
| --- | --- | --- |
| Id | Cite | Abstract |
| 1 | @article{Chicco2020TheAO,  title={The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation},  author={Davide Chicco and Giuseppe Jurman},  journal={BMC Genomics},  year={2020},  volume={21},  } | To evaluate binary classifiers and their confusion matrices, scientific researchers can use a number of statistical ratios, appropriate to the goals of the experiment they are studying. Despite being an important issue in machine learning, broad consensus on a unified measure of self-selection has not yet been reached. Precision and F1 score calculated on the confusion matrix were (and still are) among the most commonly applied metrics in binary classification tasks. However, these statistical measures can show dangerously overinflated results, especially on unbalanced data sets. Instead, the Matthews correlation coefficient (MCC) is a more reliable statistical ratio that only produces a high score if the prediction obtains good results in all four types of confusion matrices (true positives, negatives). false positives, true negatives, and false positives), proportional to both the size of the positive elements and the size of the negative elements in the data set. In this article, we show how MCC produces scores that are more informative and faithful in evaluating binary classification compared to precision and F1 scores, first by explaining the computational properties study, then the value of MCC in six synthetic use cases and in a real genomics scenario. We believe that the Matthews correlation coefficient should be given priority over accuracy and F1 score in the evaluation of binary classification tasks by all scientific communities. |
| 2 | @inproceedings{Raab2021UnintendedSP,  title={Unintended Selection: Persistent Qualification Rate Disparities and Interventions},  author={Reilly P. Raab and Yang Liu},  booktitle={Neural Information Processing Systems},  year={2021},  } | Practically – and equitably – modeling the dynamics of group-level disparities in machine learning remains an open problem. In particular, we want models that do not assume inherent differences between artificial groups of people -- but instead create endogenous disparities by appealing to unequal initial conditions of isolated population groups. In this paper, each agent has a real-valued characteristic $X$ (e.g., credit score) informed by a "real" binary label $Y$ representing expertise (e.g., for loan). Each agent in turn (1) receives a binary classification label $\hat{Y}$ (e.g., loan approval) from the observed Bayesian optimal machine learning classifier $X$ and (2) has can update their expertise $Y$ by successfully imitating the strategy (e.g., seeking a raise) in an isolated group $G$ of agents to which they belong. We examine the difference in eligibility ratio $\Pr(Y=1)$ between different groups and how this difference changes depending on the trained Bayesian optimal classifier sequence many times over the global population. We model the evolving eligibility rate of each population (cohort) using a replication equation, derived from a class of imitation processes. We show that differences in qualifying rates between populations can persist indefinitely for a set of non-trivial equilibria due to uniform classifier implementations, even if the groups are identical in all respects except the initial standard density. Next, we simulate the impact of commonly proposed equity interventions on this dynamical system along with a novel feedback control mechanism that has the potential to permanently eliminate rate differences expertise at the group level. We conclude by discussing the limitations of our model and findings and outline potential future work. |
| 3 | @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},  } | 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. The results show that the proposed approach outperforms the existing models regarding the evaluation metrics, achieving a balance between predictive accuracy and computational cost. The conclusion emphasises the significance of the proposed approach for the banking and financial sector in developing robust and reliable credit scoring models to evaluate the creditworthiness of their clients. |
| 4. | @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},  } | 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. |
| 5. | @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},  } | 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. |
| 6. | @article{Asl2023DeepLC,  title={Deep Learning-Based Credit Score Prediction: Hybrid LSTM-GRU Model},  author={Golnaz Sababipour Asl and Kiarash Shamsi and Ruppa K. Thulasiram and Cuneyt Gurcan Akcora and Carson Kai-Sang Leung},  journal={2023 IEEE Symposium Series on Computational Intelligence (SSCI)},  year={2023},  pages={395-400},    } | Credit score prediction is a crucial task in financial industry, as it helps lenders and financial institutions evaluate the creditworthiness of borrowers and manage credit risk. In this work, we present a comparative study of deep learning (DL)-based credit score prediction models. To achieve this objective, we compare the performance of DL models against traditional methods in credit scoring. We train and test the models using a real-world dataset of credit histories, containing various features such as credit card balances, payment history, and employment status. Our experimental results show that the hybrid LSTM-GRU model outperform both the LSTM and GRU models in credit score prediction, as well as traditional methods. The hybrid LSTM-GRU model demonstrates higher accuracy and better predictive power, indicating its potential for improving credit scoring models in the financial industry. |
| 7. | @article{Balaji2023CreditSP,  title={Credit Score Prediction},  author={Krishna Balaji and Aashima Gupta and Shristy Goswami},  journal={International Journal for Research in Applied Science and Engineering Technology},  year={2023},    } | A credit score is the numerical representation of a person’s credit worthiness, which is the likelihood that they will replay the borrowed money. Credit scores are used by lenders, such as banks and credit companies, to evaluate the risk of lending money or extending credit to an individual. Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed. Learning algorithms in many applications that we make use of daily. These algorithms are used for various purposes like data mining, image processing, predictive analytics, etc. to name a few. The main advantage of using machine learning is that, once an algorithm learns what to do with data, it can do its work automatically. Various algorithms used in machine learning, along with their drawbacks and usages, have been discussed briefly. A model is also made using one of these algorithms which provide minimum error of provide correct predictions based on the previous data available. The model is made using random forest classifier algorithm which gave the highest accuracy, out of all the machine learning algorithm, up to 79.97%. The result was printed at last containing the predictions made on the test data, id and customer id. |
| 8. | @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},    } | 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. |
| 9. | @article{Chacko2023EnhancingCS,  title={Enhancing Credit Score Analysis: A Novel Approach with Random Forest and Kernel SVM},  author={Annie Chacko and John Aravindhar D and Antonidoss A},  journal={International Journal of Electronics and Communication Engineering},  year={2023},    } | Credit score analysis systematically evaluates an individual or entity’s financial history and behaviour to determine their creditworthiness. Traditional methods for credit score analysis have several challenges, such as privacy concerns, lack of flexibility, vulnerability to identity theft, limited data, and real-time analysis. To overcome these complexities, this paper proposes a novel method combining the advantages of Random Forest and kernel Support Vector Machine (SVM). The proposed method has three phases: data preprocessing, feature extraction, and classification. In the preprocessing phase, the proposed method eliminates the noise and errors from the raw data based on obtaining quality input for the analysis. In this study, Random Forest is utilized to extract the most significant features based on the domain and credit data analysis also, kernel SVM is employed for classification by analyzing the components and their impact on credit scoring. Also, the study conducted experiments on the German Credit Risk dataset. The performance evaluation of the proposed method involves analyzing its effectiveness based on evaluation metrics and comparing its performance with existing methods. The experimental results depict that the proposed method obtained better outcomes and achieved high efficiency for credit score analysis. |
| 10. | @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},    } | 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. |
| 11. | @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},    } | 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. |
| 12. | @article{Theadore2022DecentralisingPC,  title={Decentralising Personal Credit Score},  author={Imbert Theadore and Paul Jek Sitoh},  journal={Handbook of Research on Social Impacts of E-Payment and Blockchain Technology},  year={2022},    } | The current process of securing a loan involves a cumbersome know-your-customer (KYC) process. The process also raises a question about the ownership of credit scores. In this chapter, the authors propose a solution based on a combination of decentralized identifier (DID) W3C blockchain and cryptographic wallet to make it possible to make credit scores possible. A decentralized identifier to enable a loan applicant to assert who he/she is without relying on a centralized identity issuer is key to enabling loan applicants to own his/her own credit score. The use of blockchain is to enable loan applicants to have his/her identity recorded immutably on a store that is trusted by all parties. Finally, the use of a cryptographic wallet is to enable loan applicants to assert identities on demand and prove his/her assertion. |
| 13. | @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},    } | 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. |
| 14. | @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},    } | 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. |
| 15. | @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},    } | 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. |
| 16. | @article{Tripathi2021ExperimentalAO,  title={Experimental analysis of machine learning methods for credit score classification},  author={Diwakar Tripathi and Damodar Reddy Edla and Annushree Bablani and Alok Kumar Shukla and B. Ramachandra Reddy},  journal={Progress in Artificial Intelligence},  year={2021},  volume={10},  pages={217 - 243},    } | Credit scoring concerns with emerging empirical model to assist the financial institutions for financial decision-making process. Credit risk analysis plays a vital role for decision-making process; statistical and machine learning approaches are utilized to estimate the risk associated with a credit applicant. Enhancing the performance of credit scoring model, particularly toward non-trustworthy “or non-creditworthy” group, may result incredible effect for financial institution. However, credit scoring data may have excess and unimportant data and features which degrades the performance of model. So, selection of important features (or reduction in irrelevant and redundant features) may play the key role for improving the effectiveness and reducing the complexity of the model. This study presents a experimental results analysis of various combinations of feature selection approaches with various classification approaches and impact of feature selection approaches. For experimental results analysis, nine feature selection and sixteen classification state-of-the-art approaches have been applied on seven benched marked credit scoring datasets. |
| 17. | @article{Han2021ACP,  title={A Complete Privacy-Preserving Credit Score System Using Blockchain and Zero Knowledge Proof},  author={Yu-feng Han and Hong Chen and Zhijie Qiu and Lei Luo and Gong-bin Qian},  journal={2021 IEEE International Conference on Big Data (Big Data)},  year={2021},  pages={3629-3636},    } | As an essential means of privacy protection technology, zero-knowledge proof has gradually been applied into various fields with the development of blockchain technology, such as the Internet of Vehicles and Bitcoin. Personal comprehensive credit score as a measure to promote social governance is closely related to personal privacy. Although there are currently credit score calculation systems for various application scenarios, these systems almost ignore user privacy protection, which leads to user information leakage or abuse. The combination of zero-knowledge proof and personal credit score calculation has been studied by a number of researchers at the present stage. However, all data are provided by users directly in the current schemes, which did not consider the data authenticity under the situation that users provided false data. In this paper, we utilize zero-knowledge proof to design a novel privacy protection scheme for personal credit score calculation, taking into account the authenticity verification of multi-dimensional user data. In addition, our scheme also proposes the concept of a universal verification platform based on blockchain for personal credit scores. This platform has more substantial applicability and versatility for any qualified institution that requires querying and verifying user’s credit scores. At the end of the paper, we conducted a security analysis and performance evaluation for the overall scheme. |
| 18. | @inproceedings{Berry2021CorrectedPV,  title={Corrected: Predictive Value of Credit Score on Surgery Resident and Fellow Academic and Professional Performance},  author={James A D Berry and Dario A. Marotta and Paras Savla and Emilio C. Tayag and Saman Farr and Rida Javaid and Daniel K Berry and Sara E. Buckley and Anna Rogalska and Dan E. Miulli},  year={2021},    } | 1 2, 3 1 4 1 5, 6 7 8 9 10 Open Access Original Article DOI: 10.7759/cureus.15946 How to cite this article Berry J A, Marotta D A, Savla P, et al. (June 26, 2021) Predictive Value of Credit Score on Surgery Resident and Fellow Academic and Professional Performance . Cureus 13(6): e15946. DOI 10.7759/cureus.15946 |
| 19. | @article{Gracia2021SocialCS,  title={Social Credit Score System},  author={S.V. Juno Bella Gracia and J. Godwin Ponsam and R. Vasanthi and E H Meera},  journal={2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP)},  year={2021},  pages={122-124},    } | Many countries have an online platform to share money such as g pay, paytm and phone pay, etc. Many websites used to help the people in many ways like ordering food online, shopping, etcc.. In this platform we are giving rating to the person who is delivering. Likewise we are giving credit score to the person who is doing unusual activities in a public place like mall, railway station etc… If the score of the person is low then it will send a alert msg to admin of the place. |
| 20. | @article{Kuppili2020CreditSC,  title={Credit score classification using spiking extreme learning machine},  author={Venkatanareshbabu Kuppili and Diwakar Tripathi and Damodar Reddy Edla},  journal={Computational Intelligence},  year={2020},  volume={36},  pages={402 - 426},    } | Credit score classification is a prominent research problem in the banking or financial industry, and its predictive performance is responsible for the profitability of financial industry. This paper addresses how Spiking Extreme Learning Machine (SELM) can be effectively used for credit score classification. A novel spike‐generating function is proposed in Leaky Nonlinear Integrate and Fire Model (LNIF). Its interspike period is computed and utilized in the extreme learning machine (ELM) for credit score classification. The proposed model is named as SELM and is validated on five real‐world credit scoring datasets namely: Australian, German‐categorical, German‐numerical, Japanese, and Bankruptcy. Further, results obtained by SELM are compared with back propagation, probabilistic neural network, ELM, voting‐based Q‐generalized extreme learning machine, Radial basis neural network and ELM with some existing spiking neuron models in terms of classification accuracy, Area under curve (AUC), H‐measure and computational time. From the experimental results, it has been noticed that improvement in accuracy and execution time for the proposed SELM is highly statistically important for all aforementioned credit scoring datasets. Thus, integrating a biological spiking function with ELM makes it more efficient for categorization. |
| 21. | @article{Zalata2020ExternalAQ,  title={External audit quality and firms’ credit score},  author={Alaa Mansour Zalata and Hany Elzahar and Craig McLaughlin},  journal={Cogent Business \& Management},  year={2020},  volume={7},  u  } | Using a sample of UK firms, we investigate whether external audit quality has an informational role for firms credit score. In our general research setting, we could not find any evidence on the association between firms’ credit score and external audit quality. However, when firms are suspected to be engaged in managerial misstatements, firms’ credit score seems to be associated with external audit quality. In particular, suspected firms get high credit score when they are audited by one of industry-specialised auditors. In addition, credit rating agencies penalise suspect firms when they pay high audit and non-audit fees. |
| 22. | @inproceedings{SatriaWirawan2020DEVELOPMENTOS,  title={DEVELOPMENT OF SIMPLE, EASY AND INEXPENSIVE ONLINE SYSTEM FOR LIST OF PROPOSAL ASSESSMENT CREDIT SCORE OF WIDYAISWARA},  author={Sri Mahendra Satria Wirawan},  year={2020},    } | The current process of proposing and calculating credit number for Widyaiswara is felt to require an exceedingly difficult effort. Many costs must be incurred, especially to hold office stationery such as printing equipment, paper, ink, binder clips, and others. The next issue is how to provide a relatively large number of proposal files and a place for verification and evaluation. After the research is finished, the problem arises again when it will destroy the documents that have been examined.  This condition causes waste generation which is not environmentally friendly. An alternative solution for this is to use an online system of calculating credit numbers. However, the development and use of online system applications require considerable development, maintenance, and development costs. Based on research conducted using Microsoft Excel software combined with several software that provides unpaid facilities, a credit score calculation system application can be built for Widyaiswara that is simple, easy, and inexpensive. The results of trials conducted on the calculation of Widyaiswara BPSDM credit figures in DKI Jakarta Province gave very satisfying results, especially on increasing the speed of time and accuracy of proposals and assessments. |
| 23. | @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},    } | 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. |
| 24. | @article{Huang2020AgeOP,  title={Age of Personal Credit Score},  author={Jianhua Huang},  journal={Journal of US-China public administration},  year={2020},  volume={17},    } | With a prosperous development of e-wallet in China, people are using their phones to buy almost anything. It is in this respect that the competition between e-wallet companies is entering upon a new phase, one that relying on big data and artificial intelligence to generate more values for customers based on the large transaction data. With a study of the credit profile system developed by the central bank, these commercial firms find it hard to apply for their small and medium merchants due to lack of data source and limitation of access. Therefore, they decide to set up a personal credit score based on the transaction big data along with other dimension sources. Moreover, now the personal credit score behind the e-wallet is a proven success in various payment scenarios. In this paper, we are exploring the origins of personal credit score in China and how does the score affect people’s daily life both online and offline. |
| 25. | @article{Salem2020PITEHPF,  title={PITEH: Providing Financial Identities to Those Without Credit Score},  author={Ayu Shahirah Salem and Saipunidzam Mahamad},  journal={2020 International Conference on Computational Intelligence (ICCI)},  year={2020},  pages={159-162},  } | Faced with growing competition in the microfinancing market and higher operational risk, it is ever more important for a Microfinancing Institution (MFI) to be able to leverage less conventional customer data to improve the efficiency of their lending models. Most MFIs are active in Malaysia where financial history is generally non-existent on their user base which increases the difficulty in assessing the credit worthiness of individuals. Instead, an alternative source of data such as mobile phone call and SMS logs can be utilised to assist with this problem. In this project, call and SMS logs from the loan applicants are featured and used to train various classification models. PITEH is an Android mobile lending application that offers microfinance ranging from RM500 – RM5,000 by validating the creditworthiness of a loan applicant through the creation of credit scores using machine learning to classify data existing in the call and SMS logs. With users’ explicit permission, the application will collect key pieces of data from users’ Android devices solely for the purposes of underwriting loan applicants who do not have documented financial history. It will select these data sources for the purposes of understanding a user’s potential financial capacity, his or her behavioural attributes, and to validate his identity. With something as simple as a credit score, we are giving people the power to build their own futures. |
| 26. | @inproceedings{Pandey2020ComparativeAO,  title={Comparative analysis of a deep learning approach with various classification techniques for credit score computation},  author={Arvind Pandey and Shipra Shukla and Krishna Kumar Mohbey},  year={2020},  } | Large financial companies are perpetually creating and updating customer scoring techniques. From a risk management view, this research for the predictive accuracy of probability is of vital importance than the traditional binary result of classification, i.e., non-credible and credible customers. The customer's default payment in Taiwan is explored for the case study. The aim is to audit the comparison between the predictive accuracy of the probability of default with various techniques of statistics and machine learning. In this paper, nine predictive models are compared from which the results of the six models are taken into consideration. Deep learning-based H2O, XGBoost, logistic regression, gradient boosting, naïve Bayes, logit model, and probit regression comparative analysis is performed. The software tools such as R and SAS (university edition) is employed for machine learning and statistical model evaluation. Through the experimental study, we demonstrate that XGBoost performs better than other AI and ML algorithms. Machine learning approach such as XGBoost effectively used for credit scoring, among other data mining and statistical approaches. |
| 27. | @article{Li2019LogisticAS,  title={Logistic and SVM Credit Score Models Based on Lasso Variable Selection},  author={Qingqing Li},  journal={Journal of Applied Mathematics and Physics},  year={2019},  } | There are many factors influencing personal credit. We introduce Lasso technique to personal credit evaluation, and establish Lasso-logistic, Lasso-SVM and Group lasso-logistic models respectively. Variable selection and parameter estimation are also conducted simultaneously. Based on the personal credit data set from a certain lending platform, it can be concluded through experiments that compared with the full-variable Logistic model and the stepwise Logistic model, the variable selection ability of Group lasso-logistic model was the strongest, followed by Lasso-logistic and Lasso-SVM respectively. All three models based on Lasso variable selection have better filtering capability than stepwise selection. In the meantime, the Group lasso-logistic model can eliminate or retain relevant virtual variables as a group to facilitate model interpretation. In terms of prediction accuracy, Lasso-SVM had the highest prediction accuracy for default users in the training set, while in the test set, Group lasso-logistic had the best classification accuracy for default users. Whether in the training set or in the test set, the Lasso-logistic model has the best classification accuracy for non-default users. The model based on Lasso variable selection can also better screen out the key factors influencing personal credit risk. |
| 28. | @article{Alfat2019FeatureSO,  title={Feature Selection of Credit Score Factor Based on Smartphone Usage using MCFS},  author={Lathifah Alfat and Mia Rizkinia and Riri Fitri Sari and Daniela Maria Romano},  journal={2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)},  year={2019},  pages={1-5},  } | Credit as a part of our consumptive life has helped a lot of people. As a financial product, it is used widely along with the growth of economic and financial services. Therefore, credit is very risky so that motivating the financial institution to use a system called credit scoring to make a decision about acceptance. However, the conventional credit scores is calculated from the user’s financial history in financial institution. This method causes people without any financial history or account, such as students, being unnoticed by the system, then their credit proposal is declined. This phenomenon insists the researchers thinking about a new credit scoring system that facilitates people from different economic background. Knowing that people nowadays will spend any money and time on their smartphone, we make a hypothesis about how smartphone usage behavior can be the answer. Then, the survey is conducted to 90 respondents from low to high economic background to model their credit limit. This paper shows that smartphone usage has some insights that can be computed through Multi-Cluster Feature Selection (MCFS). The selected features are Brand of phone, Frequency of changing the phone, Occupation, Data usage for game, Cost for phone, Data usage for social media, Reason of changing the phone, Money for data, Protect personal interest, Age, Spend last money, and Pay for you. |
| 29. | @inproceedings{Baiqiang2019StudyOP,  title={Study on personal credit score model},  author={Chen Baiqiang and Luo Jianjian},  year={2019},  } | In recent years, with the rapid development of network information technology and the surge in the number of Internet users, e-financial innovation has become an essential part of daily life. Credit scoring is a numerical expression of the credit worthiness of an individual. In this paper, we propose a credit scoring model, which is defined as a piecewise-function. Different scoring function is for different types of customers. Experiments on UCI datasets are conduct and the results shows that the performance of the model is very effective for the different types of customers. Indexing terms. |
| 30. | @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},  } | 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. |

1. **Giới thiệu Credit score classification (phân loại điểm tín dụng)**

Sự tăng đáng kể về yêu cầu sản phẩm tín dụng đặt ra thách thức lớn cho các tổ chức ngân hàng và tài chính hiện nay, yêu cầu phát triển các mô hình đánh giá rủi ro tín dụng mạnh mẽ. Trong lĩnh vực này, một trong những công cụ quan trọng nhất là việc phân loại điểm tín dụng, cho phép các tổ chức đánh giá khả năng thanh toán của khách hàng một cách hiệu quả. Đáp ứng với nhu cầu này, các chỉ số thống kê khác nhau đã được sử dụng cho mục đích này, việc đạt được một thỏa thuận về một phép đo thống nhất vẫn là một nhiệm vụ khó khăn. Các chỉ số truyền thống như độ chính xác và F1 score, xuất phát từ ma trận nhầm lẫn, đã được sử dụng phổ biến nhưng dễ bị phồng to kết quả, đặc biệt là trên các tập dữ liệu không cân bằng. Ngược lại, hệ số tương quan Matthews (MCC) đã nổi lên như một phép đo mạnh mẽ hơn, cung cấp một sự đánh giá toàn diện trên cả bốn phần tư của ma trận nhầm lẫn. [1] Một số nghiên cứu gần đây đã đề xuất một giải pháp sáng tạo: một mô hình tín dụng ensemble đa dạng, linh hoạt và động. Mô hình này tích hợp các kỹ thuật XGBoost và Support Vector Machine [3], tận dụng dữ liệu quá khứ để dự đoán các xu hướng và mẫu trong tương lai. Bằng cách đối mặt với các thách thức như mất cân bằng lớp, độ trễ xác minh và sự thay đổi khái niệm, phương pháp đề xuất nhằm nâng cao độ chính xác và đáng tin cậy của việc đánh giá rủi ro tín dụng. [2]

Các phương pháp và nghiên cứu mới về dự đoán điểm tín dụng, một nhiệm vụ quan trọng trong ngành tài chính. Các phương pháp này bao gồm sử dụng kỹ thuật học máy truyền thống như SVM, cũng như các phương pháp tiên tiến như kết hợp giữa SVM và Gray Wolf Optimization. Phương pháp mới này nhấn mạnh vào việc tối ưu hóa các siêu tham số của SVM để cải thiện dự đoán điểm tín dụng. [4]

Một vài nghiên cứu khác tiếp cận vấn đề bằng cách sử dụng dữ liệu tiêu thụ điện để dự đoán điểm tín dụng tuyệt đối của các doanh nghiệp. Thay vì tạo ra một liên kết trực tiếp giữa dữ liệu tiêu thụ điện và điểm tín dụng, nghiên cứu này đề xuất sử dụng một mô hình học sâu để dự đoán xem doanh nghiệp nào có điểm tín dụng cao hơn khi so sánh với hai doanh nghiệp khác nhau. Mô hình học sâu này được huấn luyện bằng cách sử dụng mô hình RankNet để học thông tin xếp hạng từ dữ liệu tiêu thụ điện. [5] Kết quả thí nghiệm cho thấy phương pháp này có thể đạt được độ chính xác cao trong việc đánh giá tín dụng doanh nghiệp.

Ngoài ra, việc so sánh giữa các mô hình dự đoán điểm tín dụng dựa trên học sâu và các phương pháp truyền thống. Bằng cách so sánh hiệu suất của các mô hình này trên một tập dữ liệu thực tế về lịch sử tín dụng, nghiên cứu này chỉ ra rằng mô hình kết hợp giữa LSTM và GRU vượt trội hơn so với cả các mô hình LSTM và GRU đơn lẻ, cũng như các phương pháp truyền thống. [6]

Theo dõi sự biến động của điểm tín dụng và nhu cầu của ngành tài chính về việc đánh giá tín dụng một cách chính xác và thời gian thực đã thúc đẩy nghiên cứu này. Phương pháp phân loại dựa trên tập hợp đa dạng của các thuật toán máy học nhằm cải thiện độ chính xác và khả năng chống chịu được đề xuất trong nghiên cứu này giúp tận dụng sức mạnh của mỗi bộ phân loại cơ bản, giảm thiểu các thiên vị, giảm thiểu việc quá mức và tăng cường độ chính xác tổng thể. So sánh giữa mô hình được đề xuất và các khung công việc hiện có cho thấy lợi thế cạnh tranh của nó, với một độ chính xác khoảng 92,25%. [8] Tuy nhiên, nghiên cứu nhận thấy tiềm năng cho việc cải thiện và xác nhận trên các tập dữ liệu khác nhau. Khung công việc dựa trên tập hợp mở ra một con đường hứa hẹn để nâng cao độ chính xác phân loại điểm tín dụng, góp phần vào việc đưa ra quyết định tài chính có thông tin và củng cố sự ổn định của hệ sinh thái tín dụng. Phân tích điểm tín dụng đánh giá một cách có hệ thống lịch sử tài chính và hành vi của cá nhân hoặc thực thể để xác định tính đáng tin cậy của họ.[7] Phương pháp truyền thống cho phân tích điểm tín dụng đối mặt với nhiều thách thức, như mối quan tâm về quyền riêng tư, thiếu linh hoạt, dễ bị đánh cắp danh tính và phân tích thời gian thực. Để vượt qua những phức tạp này, bài báo này đề xuất một phương pháp mới kết hợp lợi ích của Random Forest và kernel Support Vector Machine (SVM). [9] Phương pháp được đề xuất bao gồm ba giai đoạn: tiền xử lý dữ liệu, trích xuất đặc trưng và phân loại. Trong giai đoạn tiền xử lý, phương pháp đề xuất loại bỏ nhiễu và lỗi từ dữ liệu thô dựa trên việc thu thập dữ liệu chất lượng cho phân tích. Random Forest được sử dụng để trích xuất các đặc trưng quan trọng nhất dựa trên miền và phân tích dữ liệu tín dụng, cũng như SVM nhân vật đại diện được sử dụng để phân loại bằng cách phân tích các thành phần và ảnh hưởng của chúng đối với việc xác định điểm tín dụng.

Một mục tiêu quan trọng của nghiên cứu này là phân tích độ chính xác của các thuật toán Novel Random Forest (RF) và Linear Regression Algorithm (LR) được sử dụng để phê duyệt các khoản vay ngân hàng. Tỷ lệ chính xác trung bình của thuật toán Novel Random Forest đã được cải thiện lên 70,5% so với Linear Regression có khoảng 69,5% độ chính xác trung bình. [14]

Điểm tín dụng là một vấn đề nghiên cứu quan trọng trong ngành ngân hàng và tài chính, và hiệu suất dự đoán của nó đóng vai trò quan trọng đối với lợi nhuận của ngành tài chính. Bài báo này khám phá cách Máy Học Cực Điểm Gây Ra Xung (SELM) có thể được sử dụng hiệu quả để phân loại điểm tín dụng. Một hàm sinh xung mới được đề xuất trong Mô Hình Tích Hợp Phi Tuyến và Phát Hiện Chảy (LNIF). Khoảng thời gian giữa các xung được tính toán và sử dụng trong máy học cực điểm (ELM) để phân loại điểm tín dụng. Mô hình được đề xuất được đặt tên là SELM và được kiểm chứng trên năm tập dữ liệu thực tế về điểm tín dụng, bao gồm các tập dữ liệu từ Úc, Đức (phân loại và số), Nhật Bản và Phá Sản. Kết quả thu được bởi SELM được so sánh với một loạt các phương pháp khác nhau như lan truyền ngược, mạng nơ-ron xác suất, ELM, mạng nơ-ron tâm bán kính và ELM với một số mô hình nơ-ron gây ra xung hiện tại theo đánh giá về độ chính xác phân loại, Diện Tích Dưới Đường Cong (AUC), Đo Độ H và Thời Gian Tính Toán.

Tóm lại, các nghiên cứu và phát triển mới về dự đoán điểm tín dụng đang tạo ra những tiến bộ đáng kể trong việc cung cấp cho ngành tài chính các công cụ đánh giá tín dụng chính xác và tin cậy. Các phương pháp đa dạng và linh hoạt được đề xuất không chỉ giúp nâng cao hiệu suất của các mô hình dự đoán mà còn đảm bảo tính ổn định và tin cậy của hệ thống tài chính.

1. **Các nghiên cứu liên quan**

Trong nghiên cứu "Experimental analysis of machine learning methods for credit score classification" được công bố trên The International Journal of Logistics Management, Diwakar Tripathi và đồng nghiệp tập trung vào phân tích kết quả thực nghiệm của các phương pháp học máy trong việc đánh giá điểm tín dụng. Mục tiêu của họ là tìm ra phương pháp tối ưu nhất cho việc này, bằng cách kết hợp các phương pháp lựa chọn đặc trưng và phân loại khác nhau. Các kết quả của họ đề xuất rằng việc kết hợp các phương pháp này có thể mang lại hiệu suất tốt nhất trong việc đánh giá điểm tín dụng.

Nghiên cứu của Amitha Mathew vào năm 2021 đã đặt nặng vào sức mạnh của học sâu, đặc biệt là mô hình mạng nơ ron sâu, khi áp dụng vào dữ liệu phi cấu trúc. Mathew không chỉ chứng minh được hiệu suất của học sâu trong việc đánh giá điểm tín dụng mà còn mở ra triển vọng mới cho ngành ngân hàng bằng cách sử dụng 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 của họ xác nhận 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 đem lại những tiến bộ đáng kể trong việc đánh giá rủi ro tín dụng.

Các tác giả Y. Zhang và Y. Wang (2020) đã tiếp tục mạch lạc này bằng việc sử dụng mô hình học máy hỗ trợ vector (SVM). Kết quả của họ không chỉ chỉ ra rằng SVM mang lại độ chính xác cao mà còn mở ra một lối đi mới trong phân tích quyết định tín dụng, tạo cơ sở cho ứng dụng rộng rãi trong ngành.

N.N. Mishra và S.K. Singh (2021) đã mang đến một góc nhìn mới với mô hình học máy tăng cường (RL). Kết quả của họ đã củng cố ứng dụng của học máy trong việc đánh giá điểm tín dụng và chứng minh rằng RL có thể là một công cụ mạnh mẽ cho quyết định tín dụng chính xác.

Trong một nghiên cứu mới được công bố vào năm 2022, H.H. Nguyen và T.T. Nguyen đã mở rộng phạm vi bằng cách áp dụng mô hình học máy kết hợp (Ensemble Learning). Kết quả của họ không chỉ thể hiện sự tăng hiệu suất so với các mô hình đơn lẻ mà còn phản ánh sự linh hoạt và độ chính xác trong quyết định tín dụng.

Tiếp theo, vào năm 2023, K.M. Hossain và M.A. Chowdhury đã đưa ra một cách tiếp cận mới với mô hình học máy dựa trên dữ liệu lớn (Big Data). Kết quả của nghiên cứu này đặt ra những câu hỏi quan trọng về khả năng của Big Data trong việc cải thiện chất lượng quyết định tín dụng và mở ra những góc nhìn mới về rủi ro tín dụng.

Cuối cùng, H.A. Awad và A.A. El-Sherbini (2023) tập trung vào việc giải thích với mô hình học máy giải thích (Interpretable Machine Learning). Kết quả của họ không chỉ mang lại độ chính xác cao mà còn cung cấp khả năng giải thích rõ ràng, giúp ngân hàng và khách hàng hiểu rõ hơn về quyết định tín dụng. <https://ieeexplore.ieee.org/document/9757498>