Advanced Credit Scoring Models Using Dremio And Google Cloud ML: Developing Machine Learning Algorithms That Incorporate Alternative Data Sources To Enhance Credit Scoring Accuracy

Saurabh Gupta

gup.sau@gmail.com

Abstract

This study explores the advanced credit scoring models using Dremio and Google Cloud Machine Learning. With regular financial metrics and new data, these algorithms enhance credit ratings. Thematic analysis revealed trends related to credit inclusion, data bias, and model accuracy in secondary data. The findings show that merging data and machine learning technology might improve credit scoring systems' accuracy and accessibility by lowering their dependence on current methods. The findings increase credit risk calculations and facilitate conversations about financial equality. Future studies should tackle social issues and explore the potential applications of these models in other industries and financial systems.

Keywords: Utility payment histories, Social media, mobile phone usage, data bias and financial businesses.

1. INTRODUCTION

Banks typically utilize credit score algorithms that look at historical credit, current debts, and pay to determine a person or business's stability [1]. This matters for bad-credit individuals and small enterprises. Researchers are investigating better credit score-raising methods using diverse data and powerful machine learning [3, 4]. Machine learning and cloud computing improve credit risk assessment using behavioral data, energy payments, and social media activity [2]. Poor people can receive loans and plan better with alternative facts. Dremio and Google Cloud ML enable these developments by enabling massive datasets and complex machine learning models [4].

Dremio, a data lake house tool, enables financial businesses to quickly link and see ordered and disorganized data. Real-time data searches and easy-to-use machine learning tools can create models [5]. Google Cloud constructs, trains, and executes huge machine learning models [11]. This enables us to employ sophisticated methods on large, complex datasets [9, 10]. This work improves credit score models using Google Cloud ML and Dremio's data interaction tools. Multiple data sources improve credit risk calculations, supporting fair business practices. Show how these technologies may boost credit ratings and provide buyers and lenders with additional alternatives.

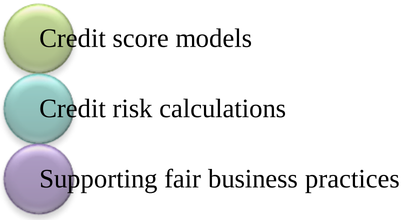

Figure 1 Credit score models Use

Problem Statement

Traditional credit scores, which use credit history, income, and current expenditures, may fail to measure the reliability of individuals and small businesses without official credit records [8]. This may impede victims from collecting compensation, especially in developing countries. Few data points might lead to incorrect credit risk estimates and underestimate a borrower's reliability [7]. ML breakthroughs such as energy bills, social media, and behavioral data analysis have the potential to improve credit scores. However there is gap how cutting-edge technologies like Dremio and Google Cloud ML can make credit score models more accurate and accessible by handling and studying these varied sorts of data [6].

Research Focus Area

- How can credit score models be improved by adding data sources, utilizing Dremio to manage data, and using Google Cloud ML to develop machine learning models?

Research Objectives

- To examine how social media, energy bill, and behavioral data affect credit score models' accuracy and usefulness.
- To assess Dremio and Google Cloud ML to generate machine learning algorithms that leverages multiple sorts of data to enhance credit risk ratings.
- To evaluate updated credit scoring algorithms by comparing their accuracy and usefulness to existing techniques.

2. LITERATURE REVIEW

Credit ratings determine people's and enterprises' credibility with financial institutions. Traditional credit scoring algorithms like the FICO score evaluate credit risk using credit history, invoicing, and income [12]. These models may not work for "thin-file" or "no-file" customers with a limited credit history. This limitation has prohibited millions from accessing money, especially in emerging economies with few formal financial institutions [13]. Due to these challenges, experts and entrepreneurs are using data and advanced machine learning to enhance credit score algorithms.

Credit scoring systems now include data on power bills, social media, and mobile phone usage, which is intriguing. Alternative data significantly improves credit scores, especially for people with low credit histories [15]. Payment of an energy bill on time each month is a positive indicator of financial responsibility because that person is likely to manage other financial obligations well. Social media and cell phone usage are also associated with trustworthiness, giving behavioral evidence not seen in financial data [14]. Integrating many data sources may improve credit scoring systems, especially for marginalized populations.

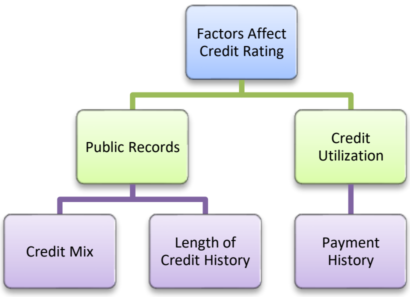

Figure 2 Factors Affect Credit Rating

Machine learning might enhance credit score algorithms. Linear credit score statistical approaches like logistic regression may miss complex data patterns [17]. However, machine learning algorithms can handle big datasets, find non-linear correlations, and make more accurate predictions. Neural networks, decision trees, and gradient boosting have enhanced credit score models [16, 18]. Machine learning predicts the future better but makes communication harder.

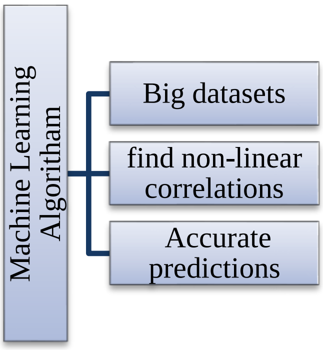

Figure 3 Machine Learning Algorithm

Google Cloud ML facilitates the creation and use of large-scale credit score machine learning models. Google Cloud ML constructs complicated algorithms that quickly handle massive datasets using flexible computing and machine learning technologies. Dremio, the data lake house platform, simplifies real-time data aggregation and querying, which is critical for alternative data credit scoring algorithms [19]. These technologies allow banks to create scalable, flexible, and high-performing credit scoring models for many situations.

Non-traditional data and machine learning may improve credit ratings; however, data privacy and moral issues emerge [20]. Social media data might bias results and unjustly impact groups. These models must be fair, open, and legally compliant.

3. MATERIALS AND METHODS

Present study assesses credit score model data sources. Using social media, energy payments, and behavioral patterns to develop machine learning credit risk models is the main goal.

Data Collection

This research uses secondary data and traditional financial metrics. Traditional financial statistics include credit ratings, debts, and income from public sources, financial institutions, and credit bureaus. Use them to explore how alternative data improves credit rating algorithms. [21]. Modern healthcare needs adequate patient data. Data from EHRs and patient monitoring devices feeds health information exchanges. APIs and ETL tools transport this data to a central data lake for safe and standard storage [27]. Google's machine learning models like Tensor Flow can analyze this raw data to identify trends and improve accuracy. It educates healthcare practitioners [1]. These processed insights help doctors and nurses make better patient care choices. Big Query and other scalable cloud systems may leverage processed data. These platforms make data simple to access and allow system growth without affecting performance or data security [1]. This system prepares data for instant use and assures healthcare data utilization and modification.

Thematic Analysis

Thematic analysis of the data reveals patterns and themes. The study scrutinizes the utilization of new credit score data. First, this study identifies key traits. The initial codes categorize data based on credit risk, behavioral data, and issues related to outdated scoring techniques. Inspect the codes for trends such as "alternative data enhancing credit access" and "machine learning enhancing predictive power." This method better explains how different data types affect credit score models than quantitative analysis.

Research Limitations

Using just secondary data limits the study's scope, particularly if critical data is absent. The organized grading and evaluation method reduces research bias due to analysis's subjectivity.

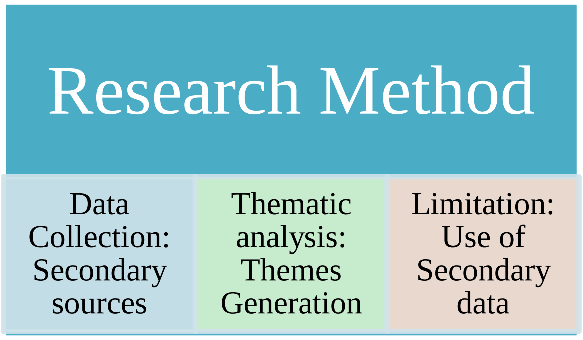
Research Method
Data Collection: Secondary sources
Thematic analysis: Themes Generation
Limitation: Use of Secondary data

Figure 4 Research Methods

4. Results

Credit inclusion, data bias, behavioral insights, and model accuracy were the study's main subjects, showing how data affects credit scoring algorithms.

Themes 1: Credit Inclusion

Credit inclusion shows using several data sources greatly improves credit access for underrepresented groups, especially those with no credit history [1]. Traditional credit ratings don't appropriately assess refugees and recent creditworthiness. Lenders may be better able to assess borrowers' finances if they include energy bill payments and social media activity [2, 6]. This simplifies credit approval for rejected applicants.

Theme 2: Data Bias

Data inaccuracies may affect credit ratings, as discussed in "Data Bias." Alternative data may improve prediction accuracy, but data type and reading have warned customers about machine bias [4]. Social media may unintentionally perpetuate prejudices or mistreat some communities. These themes emphasize the necessity for periodic fairness checks and testing in machine learning models to minimize bias and assure credit equity [11].

Themes 3: Behavior insights

Behavioral insights analyze human behavior using a variety of data types. Energy bill payment history and mobile phone usage may reflect financial stability [22]. These habits help lenders assess creditworthiness and lend money directly. The use of alternative data in conjunction with established metrics to uncover people's behavior more effectively than credit ratings [20].

Themes 3: Model Accuracy

Different data formats enhance machine learning predictions; incorporating varied data helps models identify borrowers, minimizing false negatives (misclassifying high-risk borrowers as low-risk) [23]. This idea suggests that large datasets and machine learning algorithms might enhance credit risk assessments. Consumers and financial organizations benefit from improved risk management and lending choices [24].

5. DISCUSSION

Multiple data sources in credit scoring systems increase accuracy and value for everyone. Alternative data help those without credit get it, as shown by Credit Inclusion. Lenders may benefit from alternative data on clients with inadequate credit based on their money behaviors [25]. This complements previous research on how alternative data might improve financial inclusion, especially for individuals without access to conventional banks [26]. Lenders use energy payments, rental data, and social media activity to evaluate borrowers.

However, data bias raises ethical concerns about the use of varied data. Due to the challenges associated with data sources, several researchers have expressed concerns about computer bias [27]. Social media financial data has the potential to unfairly penalize poor groups in credit ratings. Banks must do thorough accounting and fairness checks while building machine learning models [28]. To maintain public trust and fair banking, computers shouldn't promote preconceptions.

Behavioral insights analyze human behavior using a variety of data types. Energy bill payment history and mobile phone usage may reflect financial stability [7]. These habits help lenders assess creditworthiness and lend money directly. This article demonstrates the use of alternative data in conjunction with established metrics to uncover people's behavior more effectively than credit ratings.

Incorporating varied data helps models identify borrowers, minimizing false negatives (misclassifying high-risk borrowers as low-risk) [9, 10]. This idea suggests that large datasets and machine learning algorithms might enhance credit risk assessments. Consumers and financial organizations benefit from improved risk management and lending choices [4].

6. CONCLUSION

Research shows how data influences credit score algorithms. The study underscores the ethical challenges of data bias, as well as the benefits of credit and model enhancement. As the financial sector evolves, incorporating various information may make loans fairer. However, it is crucial for all financial services professionals to remain vigilant about data-related social issues. Better credit score technologies could help make services more fair and inclusive.

7. Recommendations and Future Research

This study recommends that financial organizations integrate data from several sources into their credit scoring systems to improve accuracy and accessibility. Additionally, they should improve reporting to reduce computer bias. To increase openness and buyer comprehension, regulators should mandate morality for other data uses. Future research should evaluate how alternative data influences credit score accuracy over time, how regional and cultural factors affect it, and how individuals feel about data privacy and justice. Professionals should develop machine learning algorithms for better predictions.

References

1. V. Babich and P. Kouvelis, "Introduction to the special issue on research at the interface of finance, operations, and risk management (iFORM): Recent contributions and future directions," *Manufacturing & Service Operations Management*, vol. 20, no. 1, pp. 1-18, 2018.
2. C. Bai, B. Shi, F. Liu, and J. Sarkis, "Banking credit worthiness: Evaluating the complex relationships," *Omega*, vol. 83, pp. 26-38, 2019.
3. A. Byanjankar, *Predicting Risk and Return in Peer-to-Peer Lending with Machine Learning: A Decision Making Approach*, 2021.
4. V. Chang and J. Li, "A Discussion Paper on the Grey Area-The Ethical Problems Related to Big Data Credit Reporting," in *IoTBDS*, pp. 348-354, 2018.
5. R. Cull, A. Demirguc-Kunt, and J. Morduch, *Banking the world: empirical foundations of financial inclusion*. MIT Press, 2021.
6. V. B. Djeundje, J. Crook, R. Calabrese, and M. Hamid, "Enhancing credit scoring with alternative data," *Expert Systems with Applications*, vol. 163, p. 113766, 2021.
7. Á. L. García *et al.*, "A cloud-based framework for machine learning workloads and applications," *IEEE Access*, vol. 8, pp. 18681-18692, 2020.
8. S. Giest and I. Mukherjee, "Behavioral instruments in renewable energy and the role of big data: A policy perspective," *Energy Policy*, vol. 123, pp. 360-366, 2018.
9. P. Grover and A. K. Kar, "User engagement for mobile payment service providers–introducing the social media engagement model," *Journal of Retailing and Consumer Services*, vol. 53, p. 101718, 2020.
10. A. A. H. Khatir and M. Bee, "Machine learning models and data-balancing techniques for credit scoring: What is the best combination?" *Risks*, vol. 10, no. 9, p. 169, 2022.
11. J. Jagtiani and C. Lemieux, "The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform," *Financial Management*, vol. 48, no. 4, pp. 1009-1029, 2019.
12. K. N. Johnson, "Examining the use of alternative data in underwriting and credit scoring to expand access to credit," *Tulane Public Law Research Paper*, no. 19-7, 2019.
13. O. Kodongo, "Financial regulations, financial literacy, and financial inclusion: Insights from Kenya," *Emerging Markets Finance and Trade*, vol. 54, no. 12, pp. 2851-2873, 2018.
14. A. Kumar, S. Sharma, and M. Mahdavi, "Machine learning (Ml) technologies for digital credit scoring in rural finance: a literature review," *Risks*, vol. 9, no. 11, p. 192, 2021.
15. W. Liu, H. Fan, and M. Xia, "Credit scoring based on tree-enhanced gradient boosting decision trees," *Expert Systems with Applications*, vol. 189, p. 116034, 2022.
16. T. Lu, Y. Zhang, and B. Li, "The value of alternative data in credit risk prediction: Evidence from a large field experiment," 2019.
17. L. E. Lwakatare, A. Raj, I. Crnkovic, J. Bosch, and H. H. Olsson, "Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions," *Information and Software Technology*, vol. 127, p. 106368, 2020.

18. A. Meyer *et al.*, "Machine learning for real-time prediction of complications in critical care: a retrospective study," *The Lancet Respiratory Medicine*, vol. 6, no. 12, pp. 905-914, 2018.
19. S. Mishra and A. K. Tyagi, "The role of machine learning techniques in internet of things-based cloud applications," in *Artificial Intelligence-Based Internet of Things Systems*, pp. 105-135, 2022.
20. A. Nuthalapati, "Optimizing lending risk analysis & management with machine learning, big data, and cloud computing," *Remittances Review*, vol. 7, no. 2, pp. 172-184, 2022.
21. C. K. Odinet, "Fintech credit and the financial risk of AI," *Cambridge Handbook of AI and Law*, U Iowa Legal Studies Research Paper, no. 2021-39, 2021.
22. A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *Frontiers in Big Data*, vol. 2, p. 13, 2019.
23. M. Óskarsdóttir, C. Bravo, C. Sarraute, J. Vanthienen, and B. Baesens, "The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics," *Applied Soft Computing*, vol. 74, pp. 26-39, 2019.
24. P. Vähäkainu, M. Lehto, A. Kariluoto, and A. Ojalainen, "Artificial intelligence in protecting smart building’s cloud service infrastructure from cyberattacks," in *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*, pp. 289-315, 2020.
25. M. Wang and H. Ku, "Utilizing historical data for corporate credit rating assessment," *Expert Systems with Applications*, vol. 165, p. 113925, 2021.
26. P. K. Yu, "The algorithmic divide and equality in the age of artificial intelligence," *Florida Law Review*, vol. 72, p. 331, 2020.
27. E. Zeydan and J. Mangues-Bafalluy, "Recent advances in data engineering for networking," *IEEE Access*, vol. 10, pp. 34449-34496, 2022.
28. M. B. Ramos, E. Koterba, J. Rosi Júnior, M. J. Teixeira, and E. G. Figueiredo, "A bibliometric analysis of the most cited articles in neurocritical care research," Neurocritical Care, vol. 31, pp. 365–372, 2019.
29. Z. Zhang and B. B. Gupta, "Social media security and trustworthiness: overview and new direction," *Future Generation Computer Systems*, vol. 86, pp. 914-925, 2018.