Credit Risk Prediction Using Alternative Data

Sanket Lodha, Nishikant Patil, Aditya Gaikwad, Ashutosh Misar

D. Y. Patil Institute of Master of Computer Applications and Management, Akurdi, Pune.

Students, Department of Master of Computer Application

Abstract:

Credit risk prediction is a critical component of banking and the financial sector. It allows predicting whether the borrower is likely to repay the credit. Conventional credit risk assessments rely on credit scores, credit history, income, and already-existing credits. But a considerable number of people including students, entrepreneurs, and those who have a limited credit history might lack the financial data required for proving their creditworthiness.

Alternative data collection can resolve this problem to some extent by collecting more data on the borrower's financial behaviour. Such type of data can include payment of bills or rent, payment transaction history, online payments, and any other financial activities of a client. The machine learning approach helps analyse all these data types in order to find patterns useful in improving the prediction.

This research paper analyses the application of alternative data for the prediction of credit risk, main studies in this area, possible pros and cons of using this approach. It will also provide an overview of the most critical issues related to the use of alternative data, such as privacy issues, algorithmic discrimination, quality problems, and lack of interpretability.

Keywords: Credit Risk Prediction, Alternative Data, Machine Learning, Financial Inclusion, Credit Scoring, Data Privacy.

Introduction:

Credit risk is the probability of default on a loan or other financial commitment. Forecasting this risk is important for banks, financial institutions, and lending companies since mistakes in credit decisions can lead to losses.

Traditionally, borrowers were assessed based on their credit score, income information, payment history, job position, and existing debt. Even though all this information is useful, it may not apply to all people. For instance, people who do not have any credit history but still have an ability to pay back a loan may be self-employed or working in the informal sector.

The growth of the digital financial services sector has given lenders more sources of information that can be helpful while assessing borrowers. This information

is usually referred to as alternative data and includes utility payments, rental history, bank transaction pattern, digital payment history, etc.

Machine learning techniques can process big data volumes and find connections between the borrowers' characteristics and outcomes. For instance, logistic regression, decision trees, random forests, and gradient boosting techniques can be used for building a credit risk prediction model.

But increasing the amount of data used will not necessarily result in improved predictions. The alternative data may include mistakes, disadvantage certain social or economic groups, or provide information of a confidential nature. Thus, the collection of such data should be done in conjunction with protecting privacy. This paper explores the application of alternative data for credit risk prediction and analyses the possibilities and constraints of using such data.

Problem Statement:

The conventional approaches to credit risk forecasting rely heavily on the credit history and financial documentation of people. This makes it hard for those who do not have sufficient credit history to access loans, despite the fact that they have the capability of repaying the money.

The use of alternative data gives a chance to learn about such borrowers using financial data. Nonetheless, it is not always clear whether the data sources are reliable and suitable. This is because the use of inaccurate data, biased algorithms, privacy issues, and opaque models may result in wrong lending decisions.

As such, the key challenge tackled by the research is that of using alternative data to enhance credit risk prediction while still maintaining reliable and fair predictions, data privacy, and responsible lending.

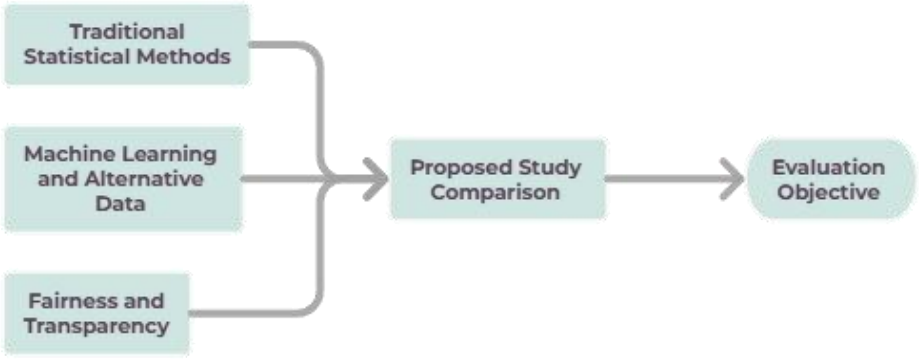
Objectives:

The following are the key goals of this study:

1. To comprehend the importance of alternative data in predicting credit risk.
2. To find out about alternative data sources that can prove to be useful in assessing borrowers.
3. To conduct a literature review on the use of machine learning and alternative data in credit analysis.
4. To explore the key risks of using alternative data for credit scoring.
5. To comprehend the issues related to data quality, privacy, fairness, and model interpretability.
6. To recommend appropriate principles for building responsible credit risk prediction systems.

Literature Review:

Prediction of credit risk has long been an area of study for researchers and many methods have been used to classify borrowers based on the likelihood of default.



The first credit scoring models were based primarily on statistical techniques and information about the borrower's past financial transactions. Logistic regression was popular as it was able to provide estimates of the probability of the borrower defaulting and yielded interpretable results.

Due to the increase in computing capabilities and the emergence of digital financial services, researchers started testing more sophisticated machine learning techniques. Decision trees, random forests, support vector machines, and boosting models could detect complex dependencies in financial data. The success of these techniques depended on the quality of the data, selected features, and evaluation metrics.

The most recent studies focused on finding out whether digital footprints and alternative financial information could help in making credit decisions. This information could be particularly useful for borrowers with no previous credit history. At the same time, researchers also pointed out the potential problems related to privacy, discrimination, and the reliability of the digital footprint as an indicator of the ability to repay a loan.

The following studies provide a foundation for understanding these developments.

Comparative Review of Key Studies

Study 1: Hand and Henley (1997) - Statistical Classification Methods in Consumer Credit Scoring

Hand and Henley have considered statistical techniques used in consumer credit scoring. Their study involved different classification techniques along with the significance of using proper techniques for credit assessment.

Relevance: The current research lays a basis for the comprehension of conventional credit scoring and the methodologies which can be used for comparison with alternative approaches.

Study 2: Lessmann et al. (2015) - Benchmarking State of the Art Classification Algorithms for Credit Scoring

Lessmann et al. compared various classification methods that can be applied to the problem of credit scoring. They found out that model comparison is important, as the assumption that a particular method would work best at all times is false.

Relevance: This research indicates the appropriateness of comparative evaluation while conducting tests on models based on different data. This research also highlights the limitations of the accuracy criterion in choosing a model.

Study 3: Jagtiani and Lemieux (2019) - The Roles of Alternative Data and Machine Learning in Fintech Lending

According to Jagtiani and Lemieux, fintech lending is influenced by alternative information and machine learning. They explain the extent to which the new forms of lending would assess the borrower based on something other than the standard credit score.

Relevance: This research is very relevant to the current study since it associates alternative information with credit scoring and financial inclusion.

Study 4: Berg et al. (2020) - On the Rise of FinTechs: Credit Scoring Using Digital Footprints

Berg et al. examined the potential of using digital footprints in credit scoring. This research was concerned with whether data from digital footprints could help predict creditworthiness.

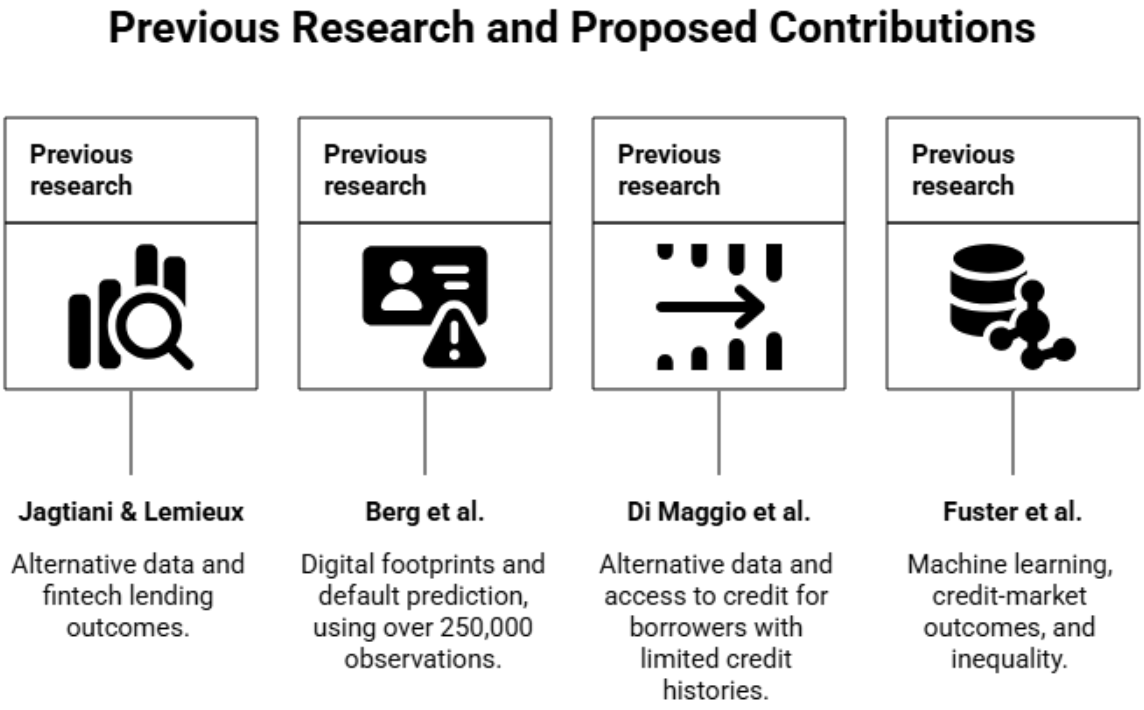
Relevance: This study explains the reasons for which researchers look to alternative ways of collecting information. At the same time, digital footprints should be analysed critically because the predictive power of such data does not justify its ethical appropriateness.

Study 5: Fuster et al. (2022) - Predictably Unequal? The Effects of Machine Learning on Credit Markets

Fuster et al. investigate the impact of machine learning on mortgage credit market and inequalities in the lending process. The authors draw attention to the need to pay attention not only to prediction accuracy but also to inequality in credit access.

Relevance: This study is significant in the context of determining whether models based on alternative data could enhance predictions, but inadvertently generate inequality.

In general, the literature suggests that there has been development in credit risk modeling from classical statistical techniques to more flexible machine learning models. Literature regarding fintech lending and digital footprint indicates that alternative data can also help in gaining some predictive insights. On the other hand, literature on credit market inequality emphasizes that model accuracy should also be evaluated in terms of fairness.



Results and Discussion

This paper applies a literature review method to explore the potential advantages, disadvantages, and difficulties of using alternative data for credit risk prediction. No results obtained from a novel dataset or a separate prediction model trained by the author are presented.

According to the analyzed studies, alternative data could be used to gain more insight into those borrowers who lack a conventional credit record. For instance, having a reliable record of timely payments for utilities or rent could serve as an indicator of financial responsibility, given that the data is correct and relevant.

Machine learning algorithms could be helpful in recognizing patterns in different datasets. The effectiveness of machine learning algorithms should be compared to that of an adequate baseline model using the same test data. Precision, recall, area under the ROC curve (AUC), and calibration could be used to assess the quality of predictions.

An example of a practical research design would be to run a traditional model with conventional credit variables and a new model with alternative data. This will enable an evaluation of which is more effective and whether there are differences depending on certain borrower segments.

The important factor here is that alternative data must be seen as extra proof instead of being seen as an alternative to traditional credit data. Alternative data is useful depending on the kind of data, the target population, quality of the dataset, and measures taken while building the model.

Major Risk Categories

1. **Data Privacy Risk**
   The alternative data will include personal and sensitive details about the financial transactions and everyday dealings of an individual. Gathering too much information will raise privacy issues as well as the risk of misuse.
2. **Algorithmic Bias**
   There is a possibility that past data from lending reflects inequalities. The model will not be properly evaluated and may result in biased decisions made towards certain communities. Other variables could serve as proxies to sensitive attributes.
3. **Data Quality Risk**
   Alternative information may be inaccurate, out-of-date, duplicated, or irrelevant. For instance, failure to make payment on utility bills could be due to a dispute over billing issues and not because of non-payment of the loan.
4. **Model Explain ability Risk**
   There are some complex machine learning algorithms that are hard to interpret. In case the borrower gets denied, it must be possible for the lender to give understandable reasons for this decision.
5. **Cyber security Risk**
   The digital accounting system can be prone to unauthorised access, hacking, or misuse. Access control, encryption, security, and monitoring play an important role in ensuring the safety of the borrower's data.
6. **Regulatory and Compliance Risk**
   It is important for financial institutions to make sure that data collection, processing, and loaning are done in accordance with all applicable laws and regulations. It is not permissible to utilize alternative data just because it can be collected.

Mitigation Principles and Open Challenges

Some of these principles include those which can reduce risks related to credit scoring using alternative data.

1. Lenders must use data sources that clearly relate to repayment behaviour. Data must be relevant to the task of assessing creditworthiness, and there must be no need for collecting unnecessary personal data.
2. Data quality checks need to be done before the models are trained. The missing values, incorrect data, duplicates, and outliers need to be scrutinized. Also, researchers need to see whether the data set used is for the borrowers who will use the model.
3. It is imperative that models be validated using the proper techniques. Data for training and testing must be kept separate, and the test data set must not be involved in the modeling process. Temporal validation may provide information on whether a model still holds up under different financial circumstances.

4. Fairness should be considered together with performance. Researchers must consider errors and outcomes for relevant groups when it is legal and proper to do so. A model that performs well on average may perform badly for a subset of borrowers.
5. Explain ability needs to be taken into account when selecting a model. Models that can be interpreted could be used if explanation is especially crucial. In the case of complex models, one can use explanation techniques and human judgment to aid responsible decision-making.
6. Institutions should have clearly defined policies regarding data access, storage, security and borrower complaints.

Despite this, there are still a number of issues that exist. One, alternative information will not always be available for each and every single customer. Secondly, a lack of digitization does not necessarily mean high credit risk. The patterns of data will evolve from time to time.

One of the other challenges faced by researchers is that of separating payment risk from economic hardship. Irregular payments may not reflect payment risks but could simply be a part of life for people who are self-employed. Researchers need to explore how to use alternative data in ways that will not penalise people who lack digital skills.

Conclusion

Predicting credit risk is one area of the financial field where data analytics can help make accurate predictions. The use of traditional techniques of credit scoring can prove to be efficient, but they might fail to give a holistic view of those people who do not have a sufficient credit record. Information such as utility bill payments, rental bill payments, and other digital financial transactions could give valuable insight into evaluating those customers.

The literature covered in this paper gives clear indications about the possibilities offered by machine learning and alternative data. But at the same time, success depends upon several other aspects as well.

Privacy, algorithmic bias, explain-ability, cyber-security, and regulation compliance continue to be issues of significance. A model cannot be regarded as a successful one only based on its superior performance. Rather, it should ensure reliable results and maintain fairness as well as privacy of the borrower.

Alternative data can facilitate credit risk prediction through a combination of alternative and traditional data as well as proper measures. Further research may concentrate on the development and testing of such a model by utilizing a publicly available credit risk dataset. The comparison of traditional and machine learning models as well as their predictive performance and fairness will be necessary.

References

1. Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160*(3), 523–541.
2. Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.
3. Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management, 48*(4), 1009–1029.
4. Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *The Review of Financial Studies, 33*(7), 2845–2897.
5. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance, 77*(1), 5–47.