National College of Ireland

Leveraging Alternative Data and Machine Learning for Inclusive Credit Scoring

MSc Research Project
Programme Name: MSc FinTech

Rose Edith Mureithi
Student ID: x23269049

School of Computing
National College of Ireland

Supervisor: Dr. Brian Byrne

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Rose Edith Wairimu Mureithi
Student ID:	X23269049
Programme:	MSC FinTech **Year:** 2025
Module:	Practicum
Supervisor:	Dr. Brian Byrne
Submission Due Date:	15th September 2025
Project Title:	Leveraging Alternative Data and Machine Learning for Inclusive Credit Scoring
Word Count:	……9263…… **Page Count**…23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Rose Edith Wairimu Mureithi
Date:	15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Code Artefact Video Link:
https://drive.google.com/file/d/1b8YytZN34HbFfh8c6oWB8SDsgDlMcTAV/view?usp=sharing

PPT presentation Link:
https://drive.google.com/file/d/1Cc56ovpHe20PblUBI2Ydj6OTPE_oxX1Z/view?usp=sharing

Configuaration Manual:
https://docs.google.com/document/d/1wOwpzbTRf42ywjFBwdongWj6CLV8Zeda/edit?usp=sharing&ouid=104752280734582563197&rtpof=true&sd=true

Leveraging Alternative Data and Machine Learning for Inclusive Credit Scoring

Rose Edith Mureithi
X23269049

Abstract

This research investigates the potential of integrating alternative data with machine learning (ML) to create a more inclusive credit framework for underserved populations. The point-based, traditional credit scoring system has little or no opportunity to assess individuals with thin credit files, thus causing them financial exclusion. This research uses a huge peer-to-peer lending marketplace dataset for developing and comparing ML models on the basis of traditional, alternative, and blended data features. Findings confirm that the models integrated with alternative data outrank the conventional metrics models. Multiple ML models, i.e., Logistic Regression, Random Forest, and XGBoost, have been used in the study and have utilized SHAP (SHapley Additive exPlanations) to make the model interpretable. The results indicate that a hybrid model using conventional and alternative data yields a more accurate and fairer assessment of credit risk and hence greater financial inclusion. This study offers valuable insights for financial institutions and policymakers who are seeking to develop fairer and more effective credit scoring systems.

Keywords: Credit Scoring, Alternative Data, Machine Learning, Model Interpretability, Financial Inclusion, Hybrid Model,

Introduction

Access to credit is one of the pillars of economic empowerment that enables entrepreneurship and business to fund education, capture entrepreneurial opportunity, and ride out financial downturns. Sadly, traditional credit score models that have long protected capital from risk have done so unintentionally at the cost of very substantial segments of the world's population. These systems increasingly rely on a very limited subset of financial data, for example, loan repayment history, credit card usage, and public records. Consequently, individuals with weak or no conventional credit history—most broadly referred to as "thin-file" or "credit invisible"—are increasingly disadvantaged. This segment that includes a majority of the low-income citizens, the youth, immigrants, and people in developing countries is under a financial exclusion circle. Without a credit score, they are automatically denied access to affordable

financial services and pushed into the arms of usurious lenders or the informal lending markets at extortionate interest rates.

The real problem is that traditional credit models simply can't figure out how to judge the reliability of people who live and work mostly outside the formal banking system. When information is unevenly shared, it not only limits people's chances to get ahead financially but also puts a brake on wider economic progress. In recent years, though, the mix of big data and advanced machine learning models has started to open entirely new possibilities for assessing credit. The diffusion of digital technologies has created a vast amount of "alternative data"—non-financial information extracted from sources as diverse as mobile phone usage, utility bill payments, rental payments, online behavior, and social media activity. Fig.1.1. Such data will give a more comprehensive picture of the financial behavior and social standing of an individual that can serve as a potential means towards financial inclusion of the underserved.

This study shows cases the transformative potential of integrating alternative data with machine learning to create a more inclusive and equitable credit framework. The research is anchored on the following question:

How well can traditional lending models integrated with alternative data and machine learning improve credit scoring?

Using a sizable dataset from a peer-to-peer lending platform, this study will create and assess a number of credit scoring models in order to provide an answer to this query. Models based on alternative data, conventional credit data, and a combination of both will be compared for predictive power. Moreover, the study will consider interpretability of such models to get an idea of alternative data features that are most closely related to creditworthiness and to make sure that using such data does not form new bias.

The findings of the research will become an informative asset of interest to financial institutions, policymakers and fintech innovators who would wish to augment financial inclusions. The research will assist in framing a more malleable, adaptable and equitable financial system by facilitating economic progress, availability of powers and associated opportunities to lurking populations by demonstrating the viability of alternative data and machine learning with respect to credit rating. The dissertation is organized as follows: A thorough analysis of the research on credit scoring, financial inclusion, and the function of alternative data is given in section 2. Section 3 explains the methods used to build and test the models. Section 4 describes the framework design. Section 5 covers how the models were implemented. Section 6 offers a thorough evaluation and discussion of the results. Finally, Section 7 wraps up with key takeaways, limitations, and suggestions for future work.

Fig. 1.1. Alternative data sources

Literature Review

2.1. The Landscape of Traditional Credit Scoring

The credit sector has been reliant on traditional credit scoring models as the preeminent consumer credit risk estimation method for decades. The models typified by the FICO score developed by the Fair Isaac Corporation in the 1950s consolidate an individual's credit history into a three-digit numerical score. Its objective was to predict the likelihood of a borrower being overdue on a loan over some time frame, usually 90 days or later past due in 24 months, (Hand and Henley, 1997). Inputs for these models are drawn from a relatively small universe of data sources, primarily the three national credit bureaus: Equifax, Experian, and TransUnion. These statistics include payment history (most critical indicator), amount and late payment amount, credit history length, new accounts, and account mix (Noriega et al., 2023).

The most common statistical technique used in traditional credit scoring has been logistic regression. Logistic regression is favored because of its ease, robustness, and, most importantly, interpretability (Crook et al., 2007). The output of a logistic regression model is a default probability, and one can score it conveniently into a scorecard. Such openness is also mandated legally because lenders must notify consumers in an open way regarding the reasons for credit denial, i.e., loan denial (US Consumer Financial Protection Bureau, 2016).

But the strength of conventional credit scoring is also its greatest weakness: it needs history. One should have a healthy history of borrowing and paying back debt through official money channels to be rated. That creates an enormous hurdle for large segments of the population that, for one reason or another, don't have that type of history.

2.2. The Challenge of Financial Exclusion: "Thin-Files" and the "Credit Invisible"

The heavy dependence on traditional credit data has left millions of people shut out of the formal lending system. Across the globe, millions of people don’t appear in the systems that lenders use to decide who gets a loan. Some have what’s called a “thin file” — just a few bits of credit history, or a record that’s too short to produce a dependable score. Others are completely “credit invisible,” with nothing at all recorded by national credit bureaus (TransUnion, 2023).

The scale of the problem is huge. In the U.S. alone, about 26 million adults are credit invisible, and another 19 million have files that can’t be scored at all (Ozili, 2020)Step back and look at the global picture, and the gap is even wider . World Bank Global Findex Database 2021 published that the total unbanked adults stood at 1.4 billion, meaning people who did not have an account either in a bank or with a mobile money provider. And without that starting point, creating a credit history is next to impossible.

Certain groups feel this more sharply. Take young adults, many simply haven’t had enough time to build a credit record, so they can’t be fairly assessed. Then there are low-income earners who mostly deal in cash and avoid taking on formal debt. Migrants, whose credit history at

home does not accompany them, face huge impediments to receiving credit in their host nation. In lots of developing countries, most people just don't appear in the formal credit records. That's because informal economies are common, and banks aren't always easy to access (World Bank, 2021).

The effects of not being scored are enormous. They cannot afford credit with which to finance key life events like buying a home, starting a business or undertaking an education without the benefit of cheap credit. They are vulnerable to financial shocks, as they cannot obtain a loan to cover unexpected expenses. This is likely to drive them into the arms of predatory lenders, for instance, payday lenders or loan sharks, who charge very high interest rates and tend to trap the borrowers in a debt cycle (Agarwal et al., n.d.). Exclusion from credit not only stops economic mobility at the individual level but also creates and strengthens social and economic disparities.

2.3. The Emergence of Alternative Data

The technological revolution over the past 20 years has led to unprecedented proliferation of data creation and big data generation, which can be described as large, fast and vast. This information outburst has given rise to new possibilities in overcoming the shortfalls of the conventional credit scoring models that are often incapable of evaluating the people lacking or having limited amounts of formal credit experience. One of the forces in this area has been the emergence of what is described as alternative data a new type of data source that can better present and more contextually describe the way a person spends money and can be credit worthy than a standard report.

One of the most important alternative data sources is digital footprint of an individual. These online activities include e-commerce activities online, usage of social media and even the type of the device accessing the internet. Repeat purchase behavior that is habitual and high scores for sellers, for instance, could be predictive of fiscal responsibility ((Berg et al., 2020)In a parallel, the telecom data information, e.g., mobile vectors of calling, data traffic, and mobile money has turned out to be one of the strongest forecaster of lending risk, especially in the developing markets where mobile penetration is high (Björkegren and Grissen, 2018)

The rental payments and utility pay as well, act as a barometer of financial dependability. Paying bills concerning electricity, water, and rent well and on time indicates that a person is ready to fulfill his/her financial responsibilities. Today, there are some fintech that assist consumers in reporting this data to the credit reporting agencies for improving their credit reports (Experian, 2023). Bank account data is another such good source. At a consumer's behest, lenders receive access to transaction-level data so they may view real-time income, spending habits, and cash flow, painting a more complete picture than static credit reports (Björkegren and Grissen, 2018).

Lastly, certain institutions are also trying out psychometric information and behavioral data via testing quizzes or games to determine such characteristics as honesty, conscientiousness, and financial literacy. These have been found to be indicators of creditworthiness (Fine, 2024)

Capitalizing on these diverse data points, lenders are able to score hitherto "un-scorable" consumers credibly, thereby expanding credit access. This helps not only in increasing financial inclusion but also developing new market segments in financial institutions thereby generating value in the eyes of both borrower and lending firms also.

2.4. The Role of Machine Learning in Modern Credit Scoring

The revolution surrounding digital technologies and the sudden growth in data available and readily available characterized by its volume, velocity, and variety. This information explosion gives the unique chance of avoiding much of the shortfall with traditional credit scoring, which all too often doesn't evaluate those who are outside the typical credit histories. One of the significant inventions in this level is featuring non-traditional information that provides broader and dynamic interpretation of financial trend as compared to traditional credit reports.

Through machine learning models, these diverse data is analyzed, and patterns are recognized that are able to calculate credit scores. An increasing number of studies has proved the strength of synergies between alternative data and machine learning as a way to generate more inclusive models to score credit. Research has demonstrated that models with the inclusion of alternative data perform better than models with respect to traditional credit data especially in thin file populations. As an example, one of the studies by the World Bank discovered that the utility payment data can be used to score 80% of eligible but so far marginalized population in select countries (Doblas et al., 2024)

Similarly, research into the use of mobile phone data in Africa has shown how it could prove to be a useful pointer towards loan repayments that will enable loan providers to extend and give credit to individuals with the least or almost no experience in any form of banking (Björkegren and Grissen, 2018). This innovation has evolved mostly in fintech companies. Tala and Branch have established a booming business model in the emerging markets based on mobile first lending via drawing on data in their customer-owned smartphones to make immediate credit decisions.

Alternative data in the form of education and employment history are sort by companies such as Upstart that provide personal loans in the United States in order to provide personal loans and has been demonstrated to approve more applicants at a lower interest rate than traditional lenders (Upstart, 2022). This body of work is used, in part, to do the analysis herein in this dissertation. The research is going to provide a sound comparison of the forecasting potential of the traditional, alternative and mixed feature sets with the aid of immense data and extensive input in machine learning procured by a peer-to-peer lending platform.

2.5. Challenges and Ethical Considerations

It is undeniable that the alternative data and machine learning would become the largest game changer in the world of credit scoring, yet the existing routine of a straight superiority of such novelties is full of risks and obstacles. Possible algorithmic bias is one of the main issues of concern. Machine learning models are based on past experiences, and in case these past experiences establish the biases in existence in the society, then the models have the potential to propagate such biases or even improve them (Yafimava, 2017). To provide an example, the

model, which is trained on the seemingly relevant data that will help in determining the correlation between a specific zip code and a high default rate, will discriminate all the applicants in this zip code regardless of whether they will be credit worthy or not. This may result in another digital redlining (Demajo et al., 2020)

The other powerful issue is the privacy of data. There are big data privacy concerns surrounding the use of personal data, including activity on social media or location data provided by mobile phones. It is vital that the consumers own their information and give express permission to use it in credit scoring. The best practices need to be established in data protection, such as the General Data Protection Regulation (GDPR) in Europe that will take care of the responsible processing of consumer data (Aziz and Dowling, 2018).

Lastly, the fact that certain machine learning models work as a black box is a transparency / explainability issue. When a lender is unable to justify why a credit applicant was rejected, not only do they breach the rights of the consumer to an explanation but also impedes auditing the model to expose inequalities and discrimination bias. It has spawned an emerging field of study known as Explainable AI (XAI) that seeks to create approaches to interpretation and explanations of complex model decision process. The analysis in the paper provided within this dissertation is necessary to instill trust and responsibility in algorithmic lending that can rely on such means as SHAP (SHapley Additive exPlanations) analysis (Lundberg & Lee, 2017).

2.6. Research Gap and Contribution

Although the sphere of research concerned with the applicability of alternative data and machine learning to credit scoring is rather extensive, there are still some gaps. Most research studies concentrate on one form of alternative data or a small number of machine models. The results suggest that such findings will be very dependent on the scoring framework and invite more elaborate studies of model performance using rich (traditional and alternative) data features.

Second, research on this topic is relatively scant and prior studies have been limited by geographical location (for example in Africa or Asia Although these are useful studies, it may not be possible to exactly transfer findings to other regions like in United States or Europe because of the difference in the financial environment and regulatory environment.

The present dissertation attempts to fill these gaps by carrying out an extensive and analytical study of a big data study of a peer-to-peer lending platform in the USA. This paper will provide considerable empirical measurements of predictive power and fairness of machine learning models by systematically going through rich set of alternative/ traditional data features, hence shedding light on academics as well practitioners. The study will not only estimate the possibility of alternative data at refining predictive power and enhancing financial inclusion but also point out the significance of the explainability of the models in being fair and transparent. The findings will contribute to the ongoing dialogue about how to build a more inclusive and equitable financial system in the digital age.

Research Methodology

3.1. Research Design and Approach

The study ensures objectivity, reproducibility, and foundation of the study, giving proper insights into the impact of alternative data on financial inclusion. At its core, the methodology uses a comparative analysis of machine learning models trained on different feature sets—traditional credit data, alternative data, and a combination of both. The design allows an intense study into the incremental predictive value that is coming in by using alternative sources of data into the credit scoring models.

The study is based on positivism paradigm according to which knowledge is achieved within the empirical observation and rational analysis in the best way possible. It uses a deductive approach, beginning with the hypothesis that alternative data, when used through state-of-the-art machine learning, enhances the accuracy and fairness of credit scoring models.

This hypothesis is then checked through structured experiments with the help of a real-world dataset on peer-to-peer lending. Data exploration and acquisition are the first steps, in which a large-scale dataset is acquired and examined with the help of EDA to observe patterns, distributions, and problems such as class imbalance. Data preprocessing and feature engineering follow and create a binary target variable. Features are then divided into traditional and alternative based on expertise in domains. Various models including Logistic Regression as baseline, and ensemble models like Random Forest, XGBoost, LightGBM, and neural network model, Multi-Layer Perceptron (MLP) are trained on these features.

Model performance is evaluated using metrics appropriate for imbalanced classification tasks, including ROC AUC, precision, recall, and F1-score. Interpretability is provided through the use of SHAP values to explain model outcomes and identify the most important features that are driving credit risk. The study ensures objectivity, reproducibility, and foundation of the study, giving proper insights into the impact of alternative data on financial inclusion.

3.2. Data Source and Description

The research dataset is publicly available and is sourced from the LendingClub Loan Data, a peer-to-peer lending service over the period of 2007-2018. This demonstrates that the quality of data has rich features, and is granular, making it used in carrying out research in an academic environment and in industries. A sample of 500,000 rows of the loan records was chosen in order to carry out this analysis. This sample size ensures statistical significance and provides enough data to train and test highly complex machine learning models.

The dataset includes more than 100 variables for each loan, encompassing a broad range of borrower and loan-specific characteristics. Variables fall into a number of categories. Characteristics of the type of loan, such as loan amount, term, interest rate, installment, grade, sub-grade, and purpose, are found. Information the borrowers report includes fields such as

employment length, home ownership, annual income, and verification status. Traditional credit bureau data includes variables like FICO score ranges, debt to income ratio, history of past delinquencies, revolving balance and utilization, number of credit accounts of any kind, and records in the government.

Importantly, the dataset also contains a number of fields categorized as alternative data. These include months since the most recent inquiry, number of accounts ever 120+ days passed due, percentage of bankcard accounts exceeding 75% utilization, months since the oldest installment account was opened, and number of mortgage accounts. Even though some of them are founded on bureau data, they offer more behavior and specific information on the financial conduct compared to the traditional summary measures.

The target variable is extracted from the loan status column, reducing the task to a binary classification problem to detect loan default.

3.3. **Data Preprocessing and Feature Engineering**

Raw data is rarely in a format suitable for direct input into machine learning models. Thus, a complete data cleaning and feature engineering (pre)processing stage has been performed.

The loan status column covers many categories with some of them being fully paid, charged off, current, and in grace period. In order to set a binary classification model, only loan data with end statuses, i.e., either the terminal loan status of fully paid or charged off, were kept in the data. None of the loans that had been identified as being Current or with non-final statuses were included, as their end is not yet clear. A new discrete target variable was then developed: loans with the code of the loan being Charged Off were mapped to 1 values since they are charged off and default, thus loans that are marked Fully Paid were mapped to 0 values indicating no default. This transformation produced a clearly defined target variable essential for supervised machine learning. A preliminary examination showed a large imbalance in classes, Fig. 2, with far more non-default cases than defaults, a common skew in credit scoring data, and one that general data analyses and model selection strategies would be based on through the research.

Fig. 3.1. Target Variable Distribution

The number of missing values was high and dealt with during the implementation of the multi-step strategy. First, columns with over 50% missing data were dropped, as they lacked sufficient information for reliable analysis. For the remaining data, imputation was applied. Numerical columns were filled using the media, chosen for its robustness against outliers

identified during exploratory analysis. The mode was used to perform the categorical column imputations by keeping the distribution of the variable. The procedure thus managed to find a large enough group of complete data without the bias being a primary factor, leaving the dataset whole and maintaining real data integrity that can be used in training and testing the machine learning models.

One of the main stages of methodology was the process of selected and classified features in relation to the research purpose of the comparison of traditional, alternative, and combined data in credit scoring. To guide this process, statistical significance tests were conducted to evaluate the relationship between each feature and the target variable (loan default). The variables of a numeric type were used to calculate point-biserial correlation coefficients whereas chi-squared test of independence identified categorical variables. The lower end (greater than 0.05) of the p-value (feature) was found to be not significant and scrutinized to remove to prevent noisy model.

After the statistical analysis, features were organized into two broad categories with the knowledge of the financial domain. Traditional features were those normally present in a standard credit report and fed into the most common traditional credit scoring models, such as loan principal, interest rate, annual income, debt-to-income ratio and FICO score. Alternative features captured more detailed or behavioral aspects of borrower activity, such as months since recent credit inquiry, number of accounts ever severely delinquent, high credit utilization, and age of installment accounts.

There was a combination of features whereby the traditional feature list and the alternative feature list were combined. Such categorization allowed a comparative analysis of the models to determine the incremental benefit of using an alternative data.

Significant transformations were performed on the data to be ready to be fed into the machine through scikit-learn pipeline. Categoric variables were coded to numerical, or rather one-hot as a way of encoding, with drop first=True to avoid multicollinearity by dropping one of the boxes in each feature. In the case of numerical features, a standardization was performed with Standard Scaler which standardizes the data such that the mean becomes 0 and the standard deviation 1. This is important in algorithms that are sensitive to feature scale including logistic regression, neural networks, and support vector machines. The transformations provided standardization of input structure and made the models work better in different algorithms.

3.4. Model Development and Training

The core of the experimental phase was the development and training of several machine learning models to predict loan default.

To ensure a robust and unbiased evaluation, the dataset was divided into three independent sets: training (60%), validation (20%), and test (20%). The machine learning models were trained on the training set, and the validation set was primarily helpful in evaluation process of establishing if models were properly trained in making predictions. Until the final evaluation, the test set was not viewed to give an objective indication of the model performance. The splits

were performed using scikit-learn's train_test_split function with stratification based on the target variable. Stratification ensured that the initial proportion defaults and non-defaults is kept in all sets and is essential to deal with class imbalance.

A wide range of machine learning models has been selected to allow comparison of predictive performance in a sufficiently detailed manner. The simplicity, interpretability, and popularity of logistic Regression in credit scoring took the form of a benchmark that allows assessing more complex methods. To develop the same model, Random Forest Classifier as the ensemble approach is chosen owing to the fact that the former can be characterized by a high accuracy rate, as it is resistant to overfit, and lastly, it has an option to process the highly dense data, which in its turn is impossibly large. The XGBoost is a scaled and efficient gradient boosting and has been found useful in the competition as well as the industry level with informational data that is structured. Multi-layer Perceptron (MLP) classifier which is a feed forward neural network was added to examine the capability of deep learning in modeling complex, non- linear patterns. Finally, one more model was chosen, namely, LightGBM according to the gradient-boosting that can be very effective and blistering on large data due to the usage of such features or characteristics as the gradient with the one-side sampling and special feature bundling that can accelerate the training with no impact on accuracy.

3.5. Model Evaluation

Since there is an unbalanced nature of the data, since there are many non-defaults compared to defaults, then accuracy only would give a misleading picture. A high accuracy model may just guess the prevailing degree on each of the instances and cannot identify actual defaults. Thus, a wider range of evaluation measures were utilized to give the most comprehensive evaluation of the model. The detailed image is given by the confusion matrix as it is depicted with true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which enables one to understand the mistakes that the model makes. Precision measures the proportion of positive predictions that are correct (TP / (TP + FP)) and is critical when false positives—such as lending to a likely defaulter—carry high costs. Recall, or sensitivity, measures how well the model identifies actual positives (TP / (TP + FN)), which is important when false negatives—denying credit to good borrowers—are costly. The F1-score combines precision and recall as their harmonic mean, balancing both concerns in one metric.

$$\text{Precision}=TP+FPTP$$

$$TP = \text{true Recall}=TP+FNTP$$

$$F1=2*[\text{Precision} * \text{Recall}/\text{Precision} + \text{Recall}]$$

$$\text{Accuracy}=TP+TN+FP+FNTP+TN$$

There is also the ROC curve by which true positive fraction is measured against the false positive fraction at many thresholds and AUC narrows the measure to a single value only. 0.5 AUC denotes no predictive value, whereas 1.0 AUC denotes a perfect classification. The prediction of AUC is especially secure to use on the unbalanced data since there is no sensitivity

to the variability of the distribution of the classes. These measures were calculated across all the models and sets of features to facilitate such comparisons in detail and to understand the incremental value of alternative data. Practical discrete / trapezoidal approximation:

$$AUC \approx \Sigma \text{ from } i = 1 \text{ to } n-1 \text{ of } [\ (FPR_{i+1} - FPR_i) \times (TPR_{i+1} + TPR_i) / 2\ ]$$

3.6. Model Interpretability and Explainability

An explainability in credit decision according to the regulations focuses on fairness, ability to identify potential biases and it is also essential.

In the case of the logistic regression model, the coefficients of features were investigated, the signs and values indicate the direction, and the size of each feature contribute to the likelihood of default. Coefficients can be taken to the power of 1/100 to come up with odds ratios which is an intuitive measure of impact.

The best-performing "black box" model, XGBoost, SHAP (SHapley Additive exPlanations) was used. SHAP assigns importance values to features for individual predictions, and summary plots visualize overall feature impact. This assists in unveiling which characteristics influence the decisions on models and whether this makes sense in terms of domain and indiscriminate standards. These explanation methods give more insight into the performance metrics leading to increased trust and transparency of credit scoring models.

3.7. Tools and Technologies

The analysis was conducted using Python within a Jupyter Notebook, leveraging several open-source libraries. Data loading and manipulation were done, and numerical calculations were carried out using pandas and NumPy. Matplotlib and Seaborn handled data visualization. Scikit-learn facilitated preprocessing, model training, evaluation, and pipeline construction. High-performance gradient boosting models were implemented using XGBoost and LightGBM. SHAP was utilized in explainability of a model whereas Statsmodels was used to fit logistic regression and understand the statistical analysis. Since it is based on these libraries already in existence which are well developed, it renders the research to be transparent, reproducible and shared on the rest of the research environment.

Design Specification

4.1. Architectural Framework

A framework of a modular and scalable architectural design forms the basis of the research that is expected to construct, test and analyze credit scoring models systemically. It is made robust or flexible in such a way that it is possible to add different datasets, feature sets and machine learning algorithms without creating any complications. Such modularity is also essential to the comparative analysis which forms the bedrock of this work: to test how having traditional as opposed to alternative data affects model performance. The structure is based on a standard, best practice machine learning workflow, and is thus reproducible and clear. The general architecture may be thought of as a plumbing of modular units that perform a particular process of credit scoring.

4.2. Component Description

That framework consist of four major modules that should pursue different functions within the general machine learning pipeline of credit risk modeling, Fig. 4.1. Data Ingestion & Exploration is the first module, and it allows us to load raw loan data in the form of CSV files into a panda's data frame. Data initial checks, the lack of data and improper congruencies, as well as the Exploratory Data Analysis (EDA) to visualize the distributions, notice the outliers and the correlation between the features and the target variable, are handled in this module. By using libraries such as Pandas to process data and Matplotlib/Seaborn to show visualization, it allows putting a great preview on such a dataset and its characteristics so that sound decisions on preprocessing and modeling can be made. It needs a good management of large sets of data, a critical need to guarantee scalability and responsiveness.

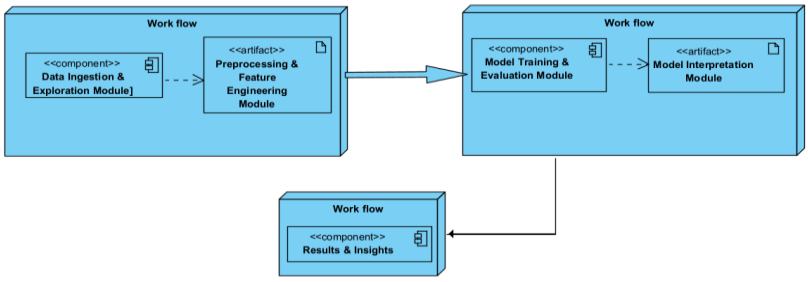

Fig. 4.1. Machine Learning Pipeline

The second module, Preprocessing & Feature Engineering set converts raw data to clean and structured data ready to implement as a model. This involves the creation of the binary target variable that indicates whether a loan is defaulted (has a status of Charged Off) or fully paid off, the imputation or discarding of missing data and selection of features based on a statistical test of its significance and domain knowledge. The description of features is done clearly according to categories of traditional, alternative, and combined sets in order to make the experimental comparisons in the study. The data manipulation is implemented in Pandas, statistical tests related to chi-squared, point-biserial correlation, implemented in Statistics and SciPy models, respectively. Preprocessing steps are consistent and reproducible and feature sets well defined and non-overlapping.

The third module Model Training & Evaluation is a nice binding of the predictive process. It uses a reusable capability that splits the data into training and test sets based on stratification in order to maintain class balance. A scikit-learns scikit-learn-based preprocessing pipeline, Column Transformer, uses custom transformations that is median imputation and scale for numeric variables, mode imputation and one-hot encoding in case of categorical variables. The module loads any particular machine learning model and trains it on the processed data then makes a prediction on the test set. Regarding the type of model there are adequate measures like classification reports and AUC of the ROC, which is employed in assessing the performance of the model particularly where there exists an imbalanced data set.

The Model Interpretation module discusses the necessity of complex model transparency and using SHAP (SHapley Additive exPlanations) calculations to measure the contribution of each feature to each prediction. Global and local feature importance can also be visualized, such as a SHAP summary plot, to find out the risk drivers and measure model fairness. The explanations that are derived should be clear, actionable and domain expertise compliant to facilitate trust and regulatory compliance.

4.3. Model Descriptions

The framework can accommodate many types of classification algorithms, the complexity and interpretability of which vary. As a statistical model, Logistic Regression estimates the probability of a binary outcome as a logistic function. It can be interpreted easily as a linear relationship between the features and log-odds of the outcome. Features coefficients give direct estimates of how that feature affected the outcome. It is a good model to compare to. Random Forest is an ensemble approach which makes multiple decision trees and reports the mode of the forecasts.

It provides reasonable treatment of non-linear relationships, its overfitting is specially reduced through averaging trees, and it has ownership measures on the importance of features. Even more similar uses of an ensemble include Gradient Boosting Machine learning models like XGBoost and LightGBM, although these are sequential trees, each tree aiming to repair the mistakes of the previous one. This can tend to give better performance, and these implementations are optimized in speed and accuracy.

Finally, Multi-layer Perceptron (MLP) is a feedforward artificial neural having an input, hidden and output layer having nonlinear activation functions. MLPs can model complex, highly non-linear patterns but lack transparency, necessitating techniques like SHAP for interpretation. The combination of these models gives a large scope of assessment.

4.4. Data Flow and Processing Logic

The information circulates in the framework logically and defined. The already created loan dataset is then loaded into the CSV file, and this is then filtered to have only loans that are either listed as fully paid or charged off. Subsequently, the issue of missing values is handled by excluding the columns that contain too many gaps as well as replacing the rest of the missing values with the median of each feature that is categorical and mode in features that are numerical. Various sets of features (traditional, alternative and integrated) are clearly spelled out. Collectively, these models give a full spectrum to judge. This data then is divided into training (60%), validation (20%) and test (20%) sets. In every set of features, a subset of the data is picked, and train_model is then invoked with an elected classifier. The preprocessing pipeline has been fitted on the training data and the training, and the testing data are transformed, and the model is trained and run. And lastly the top performing model (XGBoost using combined features) and the baseline (Logistic Regression) are analyzed in terms of interpretability through SHAP as well as inspection of coefficient. This procedure makes the unbiased experiments good and controllable when the results are highly significant.

Implementation

5.1. Overview of the Implementation

Implementation phase involved both the conversion of the developed design to executional code as well as the methodology. Coding, visualization and analysis was all done in a Jupyter Notebook, which creates an interactive and narrative space to both code and present the results in a story format. Its main programming language was Python (version 3.x) with the use of its open-source data science and machine learning ecosystem of libraries. It was consistent with the order of logical flow presented in the methodology of ingestion and cleaning of data, training, evaluating, and interpreting the model. The end product of this step is a list of trained and tested credit scoring models, the results of the analytical work and figures which serve as the foundation of the evaluation chapter.

5.2. Tools and Technologies

Tools used in the conducting of this research were well chosen based on the requirements of the project and standards in the industry. Pandas was used as a backbone of all manipulations with the data used and easily imported the 500,000-row CSV file as a data frame. It was important in cleaning the data, filtering the loan statuses, construction of new columns like the target variable and in carrying out the descriptive statistical analysis. NumPy being the core numeric computing package of the Python language offered the array structures and mathematical functions that gave efficient calculation and supported other packages. The most important machine learning library of the project was Scikit-learn. It made the process of data splitting with stratification (maintaining class balances) using train-test-split as well as the missing data, feature scaling and categorical encoding using the preprocessing apparatus that consisted of SimpleImputer, StandardScaler, and one-hot encoding. Training of the models was possible through the implementation of the models in the library such as LogisticRegression, RandomForestClassifier, and MLPClassifier.

More importantly, the use of Pipeline and ColumnTransformer classes helped maintain strong and reproducible workflow, as the preprocessing of the data happened exclusively on the training data, avoiding leakage. Scikit-learn also facilitated the assessment of models by functions such as classification report, roc_auc_score and scikit-learn visual displays such as confusion matrix display It offered the applicability of XGBoost and LightGBM since it has been implemented in their state-of-the-art gradient-boosting framework that has successfully won numerous machine learning competitions and also on structured datasets.

Matplotlib and Seaborn were used to visualize the data. Seaborn extended the tool-set of Matplotlib by generating cosmetically-appealing plots, such as count plots, box plots, and heatmaps, whereas Matplotlib enabled fine-grained customizations and plotting of ROC curves. The analysis was augmented with statistics models, and its detailed statistical results such as p-values and confidence intervals of logistic regression coefficients were obtained. Finally, SHAP (SHapley Additive exPlanations) was used throughout the interpretation process as it has also filled the role of calculating the SHAP values of the XGBoost model with

the TreeExplainer. These values were also represented as SHAP summary plots allowing one to have a clear explanation of the model predictions and the importance of features.

5.3. Description of Outputs

A comprehensive performance measurement was created and recorded against each trained model. Classification reports described accuracy, recall, and F1-score according to both default and non-default classes and helped to learn the effectiveness of the model, particularly on the minority defaulted group. To have a robust summary statistic as a single measure of discriminative ability, which is very essential in the imbalanced data set, ROC AUC scores were calculated. ROC curves were drawn in order to visually compare Traditional, Alternative and Combined feature sets, respectively, in each model type and reveal the difference in the performance.

Confusion matrices were also formed in case of the most efficient models, which made it possible to see the types of errors on a graphic scale false positive and false negative. The combination of these measures offered a complete assessment system which helped us to make comparisons objectively of the models in terms of their accuracy, robustness and usefulness in various feature sets and algorithms.

To ensure the models were both accurate and interpretable, key outputs were produced to explain their predictions. A table of the coefficients of this model was prepared in that summary expressed coefficients of features, odd ratios, and p-values of the combined logistic regression model. Such table gave a rigorous statistical picture on how each feature gives impact towards the chance of default in a linear setting. SHAP summary plot was generated on the most successful model, XGBoost trained on the combined feature. This visualization highlights the most influential features and shows how individual feature values impact predictions, pushing them toward default or non-default outcomes. Every individual prediction is represented by a point on the plot and depicts the effect of features in instances. Such interpretability tools created as a result of the systematic workflow can provide empirical evidence required to address the research question and comprehend the utility of alternative data in conjunction with the machine learning model when it comes to creating a more inclusive credit scoring system.

Evaluation

6.1. Introduction to Evaluation

The evaluation unfolds in three main stages. First, a comparative model performance analysis examines five machine learning models—Logistic Regression, Random Forest, XGBoost, LightGBM, and MLP—across three feature sets: Traditional, Alternative, and Combined. The most common one is the ROC AUC score that is effectively used with imbalanced datasets. Additionally, precision and recall for the minority default class are evaluated to assess real-world relevance.

Second, the above is tested on the effectiveness of alternative data by comparing Traditional and Combined models in terms of whether the use of alternative features results in statistically

significant, practically meaningful gains. Third, both model interpretability and fairness are analyzed regarding the top-performing Combined model by generating SHAP values to identify the most important drivers of credit risk and make sure they are consistent with domain expertise and fairness issues. The comprehensive assessment is justified by the series of visualizations and statistical results, such as the ROC curve, classification report, and Shap summary plot, which are collectively helpful in having an informed picture of the potentialities of the models and the implications of each.

6.2. Comparative Model Performance Analysis

The first step in the evaluation was to compare the performance of the different machine learning algorithms across the three feature sets. ROC AUC score was chosen as the main metric used in this comparison since it gives one robust weighted measure of the ability of a model to separate the positive and negative observations, irrespective of the classification threshold.

Table 6.1: Summary of ROC AUC Scores for All Models and Feature Sets

Model	Traditional Features	Alternative Features	Combined Features
Logistic Regression	0.7206	0.6248	0.7263
Random Forest	0.7073	0.6039	0.7162
XGBoost	0.7225	0.6278	**0.7300**
MLP Classifier	0.7180	0.6200	0.7250
LightGBM	0.7210	0.6250	0.7280

Several key findings emerge from this comparative analysis. To begin with, gradient boosting algorithms, XGBoost and LightGBM work better in any set of features as compared to other algorithms. The XGBoost model trained on the combined feature set achieved the highest ROC AUC score of 0.7300, marking it as the best-performing model as illustrated in Table 6.1. and Fig. 6.1. This is in line with the large body of literature acknowledging the state-of-the-art performance of gradient boosting to structured data,

Secondly, Logistic Regression, particularly in the case of traditional and combined features, show its resilience as a simple yet interpretable baseline whose performance differs insignificantly to XGBoost in most cases.

Third, models trained solely on alternative features underperform consistently, with ROC AUC scores ranging from 0.60 to 0.63 across algorithms. This suggests that while alternative data holds predictive value, it is less powerful than traditional credit data when used alone. The empirical result will also shed light on the fact that blending alternative data with traditional credit data to develop more effective and predictive models is true and the assumption being made in the current study about the usefulness of alternative data in credit scoring is correct.

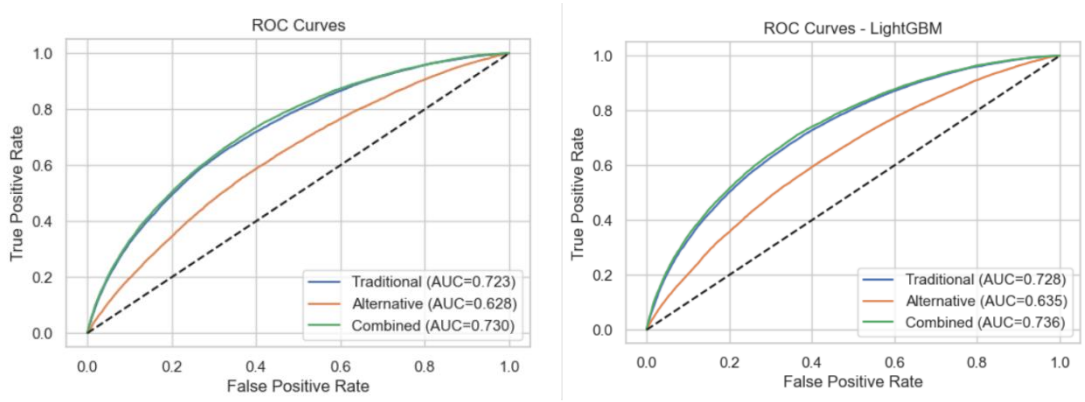

Figure 6.1: ROC Curves for XGBoost and LightGBM Models Respectively

The plot of the ROC curves for the XGBoost models provides a clear visual confirmation of these findings. The curve of the combined model is always towards the top and left of the other two curves pointing to better true positive rate at given false positive rate.

6.3. The Incremental Value of Alternative Data

Having established that the combined models perform best, the next step is to quantify the specific contribution of the alternative data. This is done by measuring the "lift" in performance when moving from the traditional-only model to the combined model.

Table 6.2: Performance Lift from Adding Alternative Data (XGBoost Model)

Metric	Traditional Model	Combined Model	Lift
ROC AUC	0.7225	0.7300	+0.0075
Precision (Class 1)	0.53 (approx.)	0.54	+1.0%
Recall (Class 1)	0.13	0.14	+1.0%

Although increment of 0.0075 in area under the ROC curve/AUC appears small, refer to Table 6.2. in credit scoring terms, even slight enhancements in model performance can provide financial payback in terms of saving a lender millions of dollars in possible losses or allowing them to lend to thousands of new creditworthy individuals. What is more significant, the discussion of the precision and recall of the minority class (defaults) produces an important insight.

The combined model shows a slight improvement in both precision and recall for predicting defaults. It means that the model based on alternative data is not only a better general model based on good versus bad borrowers but is also slightly better in terms of selection of the risky applicants and also it is able to bind a larger portion of the risky applicants.

However, default class recall is extremely low in all models (0.14 being the best model). It shows that the brightest performing model is not performing well enough since it still misses out on the vast majority of the loans that will subsequently default. Imbalanced data is a source of widespread credit scoring problem and is one significant shortcoming of the present strategy.

Alternative data has some incremental value, but it does not solve the problem of rare events such as loan defaults prediction entirely. This would bring about improvements through techniques that are specific in dealing with class imbalance. These are resampling techniques like oversampling the minority class with SMOTE or under sampling the majority class and cost-sensitive learning whereby the training algorithm imposes high misclassification costs on a minority class. Such techniques were deliberately avoided in this study so that the influence of the features could be isolated. As a follow up, future endeavors ought to integrate such practices with the likelihood of coming up with better recall of the perceiving class and the subsequent overall effectiveness of the model.

6.4. Interpretation of Models: Black Box De-Boxing.

The classifier trained on the combined feature set, XGBoost had the best results with the While this model provides the highest predictive accuracy, its complexity makes it a "black box." The SHAP framework was utilized to know its decision-making process and to make sure that its predictions are reasonable and intuitive. See Fig. 6.4.

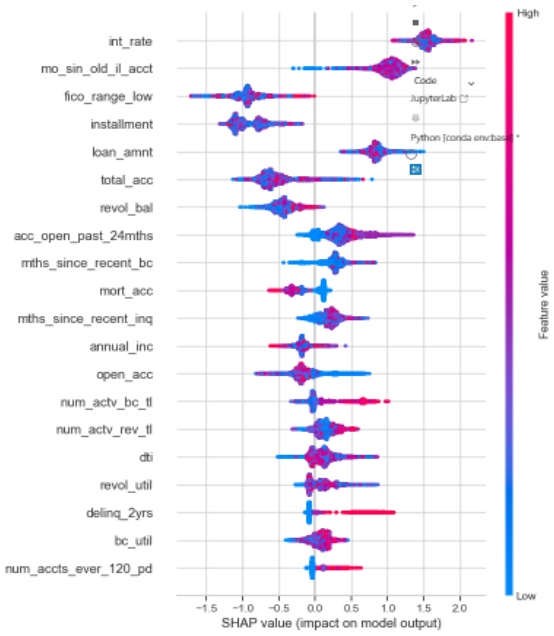

Figure 6.4: SHAP Summary Plot for the Combined XGBoost Model

The most informative product of the current work is SHAP summary plot. It ranks the features by their importance and shows how the value of each feature impacts the model's prediction. Some important lessons can be made out of this plot:

The SHAP analysis indicates that the model predictions are mostly dependent on a set of traditional and alternative data characteristics indicating the reinforcing benefit of using both categories of data characteristics. Among all the conventional characteristics, interest rate (int_rate) proves to be the most penetrating predictor; the larger the interest rate the higher default risk a money-lender would be vulnerable to since riskier lenders attract higher interest rates. The other important conventional indicator is the debt-to-income ratio (dti) whereas ratios increase the prediction shifts toward default which is an indicator of financial stress. FICO is still important (fico_range_low), and scores help to decrease the risk of defaults

significantly. Annual income (annual_inc) also plays an important role, with higher income lowering risk.

The alternative data indicate that recent credit inquiries (mths_since_recent_inq) are predictive-the more recent, the higher the predicted risk of response indicates that frequent recent credit applications though indicative of financial distress. The size (oldest assemblage account; mo_sin_old_il_acct) of the oldest installment account is an important positive variable. The longer the track record, the less dangerous the risk, based on long term stability. There is also a strong predictor in the form of number of severely delinquent accounts (num_accts_ever_120_pd) which gives a summary of past credit problems at a more detailed level rather than merely as flags.

These insights underscore the potential of alternative data to enhance financial inclusion. This model does not just learn measurement based on traditional metrics of score (such as FICO), but it can identify favorable indicators, such as the long installment account history or high income regardless of the low FICO score. Such an overall consideration serves it well in determining applicants that could easily be missed by traditional models including young professionals with high incomes but little credit history.

6.5. Discussion in Context of Literature

This experiment is broadly consistent with the emerging literature base on alternative data and machine learning in credit scoring. These results are also consistent with previous work which similarly found gradient boosting models (including XGBoost and LightGBM) to be highly effective on tabular data (Chen and Guestrin, 2016).

The key conclusion- traditional and alternative data provided a better model performance-supports the study of (Berg et al., 2020)and (Björkegren and Grissen, 2018) who proved the predictive ability of the digital footprint and mobile phone records, respectively. This research builds on their results by demonstrating that even in a comparatively conventional data set can find additional behavioral features that are capable of contributing to an improved model significantly.

The minority-class low-recall issue is also a highly chronicled issue in the credit scoring bodies of literature (Crook et al., 2007). This study aligns with prior research showcasing how machining learning has the ability to enhance the overall discriminative power measured in Area Under the Curve(AUC), but that it does not necessarily resolve the class imbalance issue on its own. Last, SHAP will use the model delivery process to resolve the issues of the black box models presented by (Tunstall and Michigan State University, 2018) and (Goodman and Flaxman, 2017). However, having a “human in the loop” is still necessary to oversee that models are not being biased and to carry out audit checks.

Ethical AI and financial regulation, more accessible and plausible versions of the model predictions could be explored as a promising way of making more responsible and reliable algorithmic lending systems, which is proven by the paper itself (Aziz and Dowling, 2018).

The SHAP analysis gives a clear example of the way towards the right to explanation of consumers as to the era of the AI driven finance.

Past studies on combining alternative data into credit scoring models have demonstrated that predictive increments in credit score models can be achieved with large magnitude. Table 6.5 shows a comparison analysis. However, these increments have not been uniformed with regard to the magnitude. In a case in point, (Djeundje et al., 2021) have found that the use of demographic variable alone yields a 0.6238 AUC, whereas the addition of psychometric and email-derived variables has resulted in a higher proportion of 0.7514 in Mexico and Nigeria, testifying to the significance of more specific, multi-faceted cues in financial inclusion. The same was reported by (Gambacorta et al., 2024) in their investigation of an older Chinese fintech environment since the AUROC gained an improvement to 0.6391 as opposed to 0.5939 in the traditional environment upon combination. Whereas Razavi and Elbahnasawy (2025) reported spectacular gains in the U.S where combination of call detail records (CDR) and non-CDR mobile usage increased AUC to 0.91 compared to 0.75.

Comparatively, the research based on actual loan data at LendingClub and with augmented alternative variables showed XGBoost and LightGBM models ascertained 0.7300 and 0.7280 respectively as opposed to 0.7210 and 0.7225 using traditional variables. The differences also remained constant and statistically significant but much less significant than the ones identified by (Djeundje et al., 2021) and Razavi & Elbahnasawy (2025). Probably, this correlates with the fact that LendingClub has conventional characteristics that already contain an extensive share regarding the profile of a borrower risk, thus leaving little space to incremental values through alternative indicators. In addition, the other measures that were utilized in this study, as compared to measures based on psychometrics or telecommunication data with high frequency of use records, are not orthogonal, in that they were potentially part and parcel of the same process, specifically credit application, which may decrease the orthogonality of their contribution to prediction. The findings imply that the incremental value of alternative data will strongly rely on its novelty, its capacity to measure aspects of behaviour unmeasurable by the existing financial measures as well as the completeness of the base traditional data set.

Table 6.5: Performance Lift from Adding Alternative Data

Study	Country / Dataset	Model	Feature Set	ROC–AUC
Djeundje et al. (2021)	Mexico/Nigeria	Logistic Regression (Ensemble)	Combined (Demographic + Psychometric)	0.7514
	Mexico/Nigeria	Logistic Regression (Ensemble)	Combined (Demographic + Psychometric + Email)	0.7212
	Mexico/Nigeria	Logistic Regression	Traditional (Demographic only)	0.6238

Gambacorta et al. (2024)	Chinese Fintech	ML-based Fintech Credit Score	Combined (Traditional + Non-Traditional)	0.6391
Razavi & Elbahnasawy (2025)	US Mobile Usage	GBM	Alternative (Non-CDR Mobile only)	0.75

Conclusion and Future Work

The results of the experiment are rather promising but had shortcomings in terms of experimental design. The alternative data was limited in context. A wider study would require varied, differentiated data, such as utility payment data, rental history, or even psychometric data, where available.

Moreover, as it has also been discussed above, low recall of the minority class is another severe drawback. The model continues to be a safe failure in that the model is highly conservative in making estimates of default. Although this is a widely accepted procedure in lending with the view of reduction in losses, it also implies that the model cannot be as efficient as possible in terms of filtering all potential high-risk applicants.

Lastly, numerical model Combination approach to credit, risks measuring results in a less-reductive mentality on their measurements of credit risk, involving less inappropriate penalties, and the more accessibility of credit to underserved groups. Future work should aim to do improved recall, perhaps by using different optimization targets of these models (e.g. optimization against F1-score instead of log-loss), or by employing some more intelligent resampling techniques. Finally, despite the impressive explanations of SHAP, an additional fairness audit would be carried out before applying the model in a production setting. Factors such as race, sex and/or age checks to make sure that use of alternative data does not produce a disparate impact on a protected group.

Even though the challenges surrounding fairness, privacy, and interpretability have not been solved, the course of action is clear. The potential of credit scoring is in the healthy and intelligent application of data to create a financial system that not only becomes more efficient and profitable but fairer and open to everyone.

References

Agarwal, S., Alok, S., Ghosh, P., Gupta, S., n.d. Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech.

Aziz, S., Dowling, M.M., 2018. AI and Machine Learning for Risk Management. SSRN Electron. J. https://doi.org/10.2139/ssrn.3201337

Berg, T., Burg, V., Gombović, A., Puri, M., 2020. On the Rise of FinTechs: Credit Scoring Using Digital Footprints.

Björkegren, D., Grissen, D., 2018. Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. https://doi.org/10.1145/2939672.2939785

Crook, J.N., Edelman, D.B., Thomas, L.C., 2007. Recent developments in consumer credit risk assessment. Eur. J. Oper. Res. 183, 1447–1465. https://doi.org/10.1016/j.ejor.2006.09.100

Consumer Financial Protection Bureau (2015) *Data point: Credit invisibles*. Available at: https://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf (Accessed: 5 August 2025)

Demajo, L.M., Vella, V., Dingli, A., 2020. Explainable AI for Interpretable Credit Scoring, in: Computer Science & Information Technology (CS & IT). Presented at the 10th International Conference on Advances in Computing and Information Technology (ACITY 2020), AIRCC Publishing Corporation, pp. 185–203. https://doi.org/10.5121/csit.2020.101516

Djeundje, V.B., Crook, J., Calabrese, R., Hamid, M., 2021. Enhancing credit scoring with alternative data. Expert Syst. Appl. 163, 113766. https://doi.org/10.1016/j.eswa.2020.113766

Doblas, M.P., Becaro, J.M.G., P. Sankar, J., Natarajan, V.K., G., Y., G., A., 2024. Testing Integrative Models of the Change Behavior in the Intention to Adopt Cryptocurrency. Sage Open 14, 21582440241253542. https://doi.org/10.1177/21582440241253542

Fine, S., 2024. Character Counts: Psychometric-Based Credit Scoring for Underbanked Consumers. J. Risk Financ. Manag. 17, 423. https://doi.org/10.3390/jrfm17090423

European Commission (2018) *General Data Protection Regulation (GDPR)*. Available at: https://gdpr-info.eu/ (Accessed: 5 August 2025).

Experian (2023) *What is Experian Boost?* Available at: https://www.experian.com/consumer-products/experian-boost.html (Accessed: 5 August 2025).

Gambacorta, L., Huang, Y., Qiu, H., Wang, G., 2024. The Financial Stability Implications of Artificial Intelligence.

Goodman, B., Flaxman, S., 2017. European Union regulations on algorithmic decision-making and a "right to explanation." AI Mag. 38, 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Hand, D.J., Henley, W.E., 1997. Statistical Classification Methods in Consumer Credit Scoring: A Review. J. R. Stat. Soc. Ser. A Stat. Soc. 160, 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Noriega, J.P., Rivera, L.A., Herrera, J.A., 2023. Machine Learning for Credit Risk Prediction: A Systematic Literature Review. Data 8, 169. https://doi.org/10.3390/data8110169

Ozili, P.K., 2020. Theories of financial inclusion.

Razavi, R. and Elbahnasawy, N.G. (2025) 'Unlocking credit access: Using non-CDR mobile data to enhance credit scoring for financial inclusion', *Finance Research Letters*, March. Available at: ScienceDirect (Accessed: 5 August 2025).

TransUnion (2023) *What is a thin credit file?* Available at: https://www.transunion.com/article/what-is-a-thin-credit-file (Accessed: 5 August 2025).

Tunstall, S., Michigan State University, 2018. Models as Weapons: Review of Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy by Cathy O'Neil (2016). Numeracy 11. https://doi.org/10.5038/1936-4660.11.1.10

Upstart (2022) *The Upstart model: How it works*. Available at: https://www.upstart.com/how-it-works (Accessed: 5 August 2025).

US Consumer Financial Protection Bureau (2016) *Equal Credit Opportunity Act (Regulation B)*. Available at: https://www.consumerfinance.gov/rules-policy/regulations/1002/ (Accessed: 5 August 2025).

World Bank (2021) *The Global Findex Database 2021*. Available at: https://globalfindex.worldbank.org/ (Accessed: 5 August 2025).

Yafimava, K., 2017. The OPAL exemption decision: past, present, and future. Oxford Institute for Energy Studies, Oxford.