Alternative-Data Credit Scoring for Migrant Workers: A Synthetic Data Simulation Study

Hrithik Manikandan[1]

Received May 5, 2026
Accepted July 25, 2026
Electronic access September 15, 2026

Credit invisibility is a major challenge to financial inclusion, particularly for blue-collar migrant workers who have regular but non-registered income. These groups are excluded by traditional credit-scoring systems, which rely on bank records. This paper examines the potential of linking machine learning models that are developed using alternative data to enhance accuracy and fairness in credit assessment among migrant workers. The synthetic dataset consisted of 10,000 profiles, with a rough 70:30 ratio of approved to rejected applicants, and included demographic, financial and behavioral variables, while respecting the privacy of the applicants. The four supervised learning models were trained and assessed by precision, recall, F1-score and AUC-ROC. Fairness was measured by Demographic Parity Difference and Equalized Odds Difference, with a cutoff of 0.05 for the parity differences. Within the synthetic dataset, all models had good predictive performance with Logistic Regression achieving an F1 score of 0.98. The results presented here are for synthetic data generation and do not constitute evidence of predictive validity in the real world. The Random Forest model had a remarkably low recall (0.64), which may be due to its sensitivity to class imbalance. In the simulated data, the most important behavioral variables were consistently paying utility bills and the frequency of mobile top up. There were minimal differences found between the nationality groups in terms of fairness indicators, which was not surprising as they are constructed in a fair manner but do not reflect actual fairness in the real world. The findings indicate that alternative data machine learning models have potential to be further explored as an avenue to expand credit access, provided for practical testing.

Keywords: credit scoring, alternative data, migrant workers, financial inclusion, machine learning, algorithmic fairness, synthetic data, binary classification

Introduction

Formal credit plays a crucial role in economic activity and enables people to invest in education, housing, health care and entrepreneurship. These investments help to become more financially stable personally and enhance the development of the whole economy[1,2]. Nevertheless, credit access has been unequally spread, especially in the developing economies as well as among people who are in informal or non-permanent jobs like the migrant workers. Conventional credit-scoring models are based on official financial documents such as bank account history, credit card usage, and loan repayment patterns, which are systematically discriminating against people who do not have a banking history[3]. This has resulted in many parts of the world population being rendered invisible in credit.

The key issue that the present paper will address is that of excluding migrant workers by formal credit systems since they do not have the traditional financial records yet they can have a steady income and a steady financial behavior. Traditional credit-scoring models do not capture a significant portion of migrant workers who rely on remittances, cash or informal financial systems. This marginalization denies them access to cheap credit and curtails chances of economic mobility. It is necessary to address this gap in order to decrease inequality and increase financial inclusion, especially in areas where informal lending is the norm and consumer protection is scarce. This research will contribute to the nascent field of AI-driven financial inclusion by examining the possibility of being associated with improved financial inclusion for under-served populations using alternative data. The goal of this study is to provide an example of a more comprehensive approach to measuring creditworthiness, while avoiding loss of predictive power or unfairness in the measurement, by using non-traditional measures such as utility payment patterns, mobile phone use, and employment stability. The practical implications of the findings are in terms of the financial institutions, policymakers and the fintech providers who want to develop responsible and inclusive lending systems. The technical contribution of this study is empirical and exploratory; it investigates the potential of alternative-data machine learning models to predict well, be fair, and be explainable in a controlled environment of synthetic data. The credit approval problem is a

[1] *Brighton College Abu Dhabi, Abu Dhabi, UAE*

binary classification problem (0 = rejected, 1 = approved).

The goals of the study are: to evaluate the predictive accuracy of machine-learning models for predicting creditworthiness from conventional and non-conventional data, by demographic group, while taking nationality as a protected attribute; to assess and compare the fairness of models by demographic group; and to make sure that the predictions made by the models are transparent and explainable to institutions and applicants. The goals are translated into three hypotheses that can be tested. H1: The machine learning models trained on the alternative and traditional data can obtain a high predictive ability on the synthetic data (F1 score $>$ 0.90, AUC-ROC $>$ 0.90). H2: The low disparity of fairness between nationality groups (Demographic Parity Difference and Equalized Odds Difference are both below 0.05) is possible with these models without using post-processing debiasing. H3: With proper data preparation, simpler models such as Logistic Regression can be as effective or more effective compared to ensemble methods.

This research is concerned with the creation and assessment of credit-scoring models on a synthetically created dataset that is aimed at replicating real-world demographic, financial and behavioral trends. Four trained machine-learning models are analyzed, and the fairness is tested with nationality as the primary factor. While synthetic data can be used to guarantee privacy and explore the use of controls, it may not be as representative of financial behavior in the real world and may have some limitations in terms of extrapolation of results.

The study is grounded in the concept of financial inclusion and the algorithmic fairness framework. Financial inclusion is about equal access to financial services, particularly for the unbanked, while fairness in machine learning is about minimizing bias and having the same results for different demographic groups[4,5]. The structures shed light on the assessment of the predictive performance as well as the ethical dimension of the proposed credit-scoring model.

In the quest to answer the research questions, this paper constructs an artificial dataset and uses various supervised machine-learning algorithms to forecast credit risk. The standard classification measures are used to evaluate the model performance, the fairness and interpretability measures are undertaken to promote ethical and transparent decision-making.

Literature Review

One of the most significant aspects of credit scoring is that lenders can gauge the dependability of the borrower and foretell how he or she will pay back the loan. Official financial coefficients such as repayment history, income, outstanding payments and credit utilization ratios are used to base traditional credit-scoring systems.[6,7]. Logistic regression was the most popular model previously because of its simplicity, interpretability and its performance in structured linearly separable data sets.

These conventional models, however, have a number of limitations. They exclude migrants, low-income and informal-sector workers, who lack a financial history, from the formal banking system, systematically. According to the World Bank (2021), it is estimated that 1.4 billion adults are still unbanked, mainly in areas like Sub-Saharan Africa, South Asia, and Latin America. Limited access to credit decreases mobility in the economy and opportunity to start businesses, finance education, and accumulate wealth in the long term[1,2]. Logistic regression has been shown to be less powerful in high dimensional or non-linear settings, and is structurally limited in populations with informal financial histories as the financial variables used are formally documented.[7]. While ensemble methods help to address some of these limitations in prediction, there are also advantages and disadvantages to them, including interpretability and regulatory transparency[8].

To overcome these constraints, researchers and fintech developers have been looking for other data sources as a supplementary input to credit analysis. Alternative data refers to a wide variety of behavioural and digital signals, such as paying utility bills, mobile phone activity, remittance frequency, employment record, e-commerce activity, applications, geolocation, and even psychometric indicators[9–11]. The non-conventional financial data are financial discipline, consistency of behaviour and social economic background. For instance, mobile phones can be used to identify a regular pattern of top-ups, which can be associated with a regular pattern of income; and geospatial mobility patterns can be used to infer employment or income status[12]. Consequently, there is a great potential in alternative data to widen credit access to historically underserved and credit-invisible populations. There are, however, not all signs of success. On the LendingClub dataset, the alternative data proved to be more predictive than the traditional data, but Jagtiani and Lemieux (2019) also observed that some of the alternative data may be proxies for protected attributes like race or nationality, which could reintroduce the very biases that they seek to avoid[10]. Li et al. (2020) further cautioned that the socioeconomic disadvantage can be captured in the behavioral and digital footprint data in a manner that is not obvious from the traditional notions of fairness[11].

The merging of various data types and finding non-linear relationships in credit data are well suited to machine-learning (ML) techniques. Some of the popular decision tree-based algorithms such as Random Forest, XGBoost and LightGBM are effective in high-dimensional setting, are good at handling multicollinearity and missing data.[8,13,14]. Random Forest minimizes overfitting by pooling predictions using a lot of trees. Gradient-boosting models (XGBoost, LightGBM) first improve the prediction by focusing on the samples that

have been misclassified and they are very accurate even in the case of imbalanced data. Logistic regression is a relatively simple model, but in financial services it is a norm due to its transparency and regulators' acceptance[7,15]. Recent research also suggests that properly designed logistic regression can perform as well as the complex ensembles and maintain high interpretability, a crucial part of consumer trust and adherence[16]. Credit-scoring applications have also shown good performance when using hybrid models combining linear and gradient-boosting models[17].

One of the key gaps in previous research is the lack of inclusion of fairness assessment along with predictive assessment. Researchers currently focus on measures of accuracy and do not systematically analyze if models are equitable to demographic groups. This study fills that gap by using Demographic Parity Difference and Equalized Odds Difference as key evaluation metrics, in addition to the standard classification metrics. With the increase in the use of AI-based credit scoring, issues of fairness and bias in the algorithm have emerged. Discrimination may happen when some people are discriminated against on a consistent basis because of their nationality, gender, age or socioeconomic status. Popular fairness measures include Demographic Parity (the rate of approval between groups) and Equalized Odds (the true-positive and false-positive rates[4,5,18]. Modern approaches towards fairness include counterfactual fairness and intersectional analysis, determining whether the applicants of similar financial statuses get equal treatment irrespective of demographic characteristics[19,20]. The research suggests that bias-reduction techniques, such as preprocessing, feature selection, and balanced sampling, can help reduce the differences without compromising predictive performance significantly[21,22]. However, there are some assumptions and known trade-offs for these fairness metrics. Demographic Parity assumes equal approval rates across groups is the correct measure of fairness, which may be different from Equalized Odds, which focuses on equal error rates. Kleinberg et al. (2018) showed that these two notions of fairness can only be met simultaneously under very special circumstances, as proved mathematically[20]. This intrinsic conflict implies that there is no single fairness measure that can always be used, and the researcher needs to clearly state which fairness measure they are optimizing for and why.

However, there are new ethical concerns with alternative data, including privacy concerns, consent concerns, and the potential to inadvertently create proxies of protected attributes. The regulatory frameworks such as GDPR, PSD2, open-banking focus on transparency, minimal data, and responsible model governance[23–25]. Explainability is also essential: models that are not transparent or very complex can lead to a drop in user trust and the ability of institutions to offer a rationale for loan decisions. The models can be used to explain the importance of features, and can be interpreted using techniques such as SHAP to provide explanations of the model results[16,26]. Although AI-based credit scoring is of great interest, there are a number of gaps in the existing literature. Most research is based on predictive accuracy and misses fairness, interpretability, or practical applicability to the real world[1]. Fairness analyses typically consider only one of the factors being protected, and do not account for intersectional or context-related biases. Not many studies discuss the possibility of integrating alternative data into regulated financial systems, without infringing on privacy standards or operational requirements[9,10]. Current evidence, therefore, offers little advice on how to institute systems that are both precise, just, interpretable, and in line with the data-protection laws.

The study extends the state of the art by creating a synthetic dataset that combines both classical and alternative data sources and by evaluating the performance of different machine-learning models from three angles: predictive performance, fairness, and interpretability. The study provides a more holistic assessment of algorithmic fairness than most existing studies, by treating nationality as a protected property. The results are expected to guide the design of transparent, inclusive, and ethically-based credit-scoring systems that can be used in the real-world financial settings.

Research Methodology

The paper proposes a systematic quantitative research design, training and testing of AI-based credit scoring models on conventional and innovative data. The framework focuses on predictive accuracy, the fairness of the decisions made with regards to the demographic groups that fall under the protection and the interpretability of the decisions made in accordance with the responsible financial practices[4,5].

Research Design

A cross-sectional, experimental research design was used. The experiment uses a single synthetic dataset, which consisted of migrant workers at a single time, whereas machine-learning models were trained experimentally under controlled conditions[7,15]. The design allows for a methodical comparison of different algorithms and allows for separating the effects of data preprocessing, feature engineering, and fairness interventions. Since the aim is to recreate real world lending conditions without having access to sensitive information about borrowers, the research uses probabilistic simulation instead of field data based on the observational approach. This ensures that the work is reproducible and ethical[1,3]. Probabilistic simulation is known to enable controlled experimentation, but does not adequately represent the complexity of lending environments in the real world and results should be interpreted accordingly.

Table 1 Comparative Summary of Prior Studies in AI-Based Credit Scoring

Study	Data Type	Models Used	Fairness Assessed	Key Finding
Jagtiani & Lemieux (2019)	Real-world (LendingClub)	Logistic Regression, ML	No	Alternative data improves predictive performance but may encode proxy bias
Bao & Huang (2021)	Real-world (emerging markets)	Ensemble methods	No	Behavioral data improves inclusion for underbanked populations
Li et al. (2020)	Real-world	Multiple ML models	Partial	Digital footprints predict creditworthiness but raise privacy concerns
Nguyen et al. (2020)	Real-world	Hybrid linear + ensemble	No	Hybrid models balance accuracy and interpretability
This study	Synthetic (10,000 profiles)	LR, RF, XGBoost, LightGBM	Yes (nationality)	Explores predictive performance, fairness, and interpretability in a controlled simulation

Participants / Sample

The sample size is 10,000 synthetic financial profiles, which are designed to be the credit history of migrant workers with no or minimal conventional credit history. The key demographic characteristics are age (20-55 years), nationality (categorical protected attribute), education level (primary, secondary, vocational and tertiary) and family structure (single or extended family). The financial and behavioral characteristics are based on the trends that have been reported by the World Bank Global Findex, IMF Financial Access Survey, and case studies of micro-lending among migrant communities[9,10]. A 10,000 sample size is used for training a model and statistically significant fairness assessment. The dataset is fairly balanced in terms of approved and rejected applicants, with 70% of the population being approved and 30% being rejected, as reported in micro-lending contexts[1].

Data Collection

Data collection was done purely based on synthetic data generation since no actual individuals were surveyed or interviewed. This approach allows full control of the distribution of features, reduces the ethical risk, and reproduces the realistic financial behavior. Authentic variability was created for a number of variables: salary was sampled from a truncated normal distribution with bounds set at minimum wage and median migrant income; length of employment was sampled from an exponential distribution as this represents the typical distribution of short-term contracts; the frequency with which migrant households paid their utility bills was sampled from a beta distribution that captures the range of regularity of bill payments, while the frequency with which migrant households made remittances was sampled as a categorical variable (weekly, bi-weekly, or monthly) following the typical distribution of remittance frequency in migrant households. To prevent feature correlations, such as between income and rent, or between job stability and remittance consistency, from being unrealistically independent, joint sampling was used.

Label Generation

The binary credit approval label is based on a weighted threshold function of three primary variables: salary to rent ratio, employment length and the regularity of utility payments. To avoid the problem of trivial learning of the classification task, the threshold function was modified by adding a stochastic noise term. In particular, the credit approval (1 or 0) was assigned to an applicant based on whether the weighted score was above or below a certain threshold after adding the noise term. This is recognised as a limitation, as the approval label is generated from engineered features and not observed repayment behaviour, so the models are learning the label generation function, and not the true creditworthiness signal. The outcome variable of interest would be the repayment or default of loans over a certain period of time, which is also identified as a priority for future research.

Variables and Measurements

Credit approval is the dependent variable in this study, which is a binary variable (0/1) indicating whether or not an applicant is approved for a loan. The traditional financial independent variables include salary, rent, employment length, and the applicant's profession and work sector. Other variables and behaviors are consistency of utility payments, frequency of top-ups, frequency of remittances, and overall measures of employment stability. Demographic variables include age, education level, household size and nationality (which is used as a protected attribute to assess fairness). One-hot encoding was used to transform categorical variables. All continuous variables were standardized and min-max scaled to facilitate model convergence[8,13,14]. Interaction terms were developed to capture meaningful financial interactions, such as income-to-rent ratio and normalized remittance intensity. Temporal behavioral indicators are created to measure reliability over the past six months, consistent with the way lenders measure stability. The models can be trained on these aspects of credit scoring: financial capacity and behavioral responsibility.

Procedure

The research took a multi-step systematic process. A synthetic dataset was created based on probabilistic distribution, preserving feature correlations. A realistic missingness of about 12% was added as a marker of the flaws in financial records. The data pre-processing included the imputation of missing values using median and mode, outlier clipping at 1st and 99th percentile, multicollinearity analysis by Variance Inflation Factor, and engineering financial ratio and behavioral stability features. The class balance was ensured by using a stratified train-test split (70/30), which led to approximately 7,000 training and 3,000 test profiles (70/30 ratio was held across both splits). Four algorithms were trained: Logistic Regression, Random Forest, XGBoost and LightGBM. Hyperparameter tuning was done by grid search with 5-fold cross validation, and a fixed random seed of 42 was used for reproducibility.

The hyperparameters for Logistic Regression were the regularization strength (C: 0.01, 0.1, 1, 10), while for Random Forest, LightGBM, and XGBoost the hyperparameters were the number of estimators (100, 200, 500), maximum depth (3, 6, 9), and learning rate (0.01, 0.1, 0.3). The classification metrics were used as a measure of performance. The credit score outputs were scaled to the 300-850 range, with 70% based on the model probability, 20% based on normalized income, and 10% based on normalized age. Demographic parity and equalized odds metrics were used for fairness assessment. The results from the SHAP interpretability analysis are not available and were performed in the original research. This is seen as a limitation and SHAP analysis is identified as a priority for future work. The same goes for the original analysis code, which is not available, limiting the ability to reproduce this study. This ensures a transparent, repeatable and morally sound workflow.

Data Analysis

The machine learning models employed were Logistic Regression (LR), Random Forest (RF), XGBoost (XGB) and LightGBM (LGB), which are considered as interpretable baseline model, non-linear pattern recognition model, gradient boosting model with high predictive power and fast, efficient and scalable boosting model, respectively. The accuracy, precision, recall, F1-score and AUC-ROC were used to measure the performance metrics, including correctness, sensitivity and robustness. The measures of fairness were Demographic Parity Difference (DP Difference) and Equalized Odds Difference (EOD). These assess the potential for disadvantage of protected groups. Use of a 300-850 range gives the AI system a familiar framework for lenders and makes the credit-scoring system more similar to the actual one.

Ethical Considerations

There were no real human data, all profiles were completely artificial and therefore privacy threats were not possible. The ethical guidelines were adhered to, namely: not using sensitive personal identifiers, transparency through interpretability, fairness and avoiding discrimination[4,5,18], and adhering to the responsible AI principles applied to finance. The methodology recognizes that the deployment to any actual world would need continuous monitoring, documentation that would be compliant with regulators and further testing of fairness.

Results & Analysis

This study's results provide insight into the predictive accuracy, fairness, and interpretability of traditional financial data vs. new behavioral data AI credit-scoring models. The analysis is based on the measures of model evaluation, the distribution of credit score and the outcome of fairness with regard to the attribute of nationality which is protected. The results presented in this section are based on the synthetic dataset and are not intended to be a measure of predictive validity in the real world.

Data Preprocessing Outcomes

Preprocessing of data was effective in tackling the statistical and operational challenges before training the model. The analysis of the Variance Inflation Factor (VIF) showed that a few features were moderately multicollinear (Phone Top-Ups VIF = 7.83, Salary VIF = 6.83, and Rent VIF = 6.59). There was, however, no value that was above the widely accepted value of 10, which shows that there were no variables that needed to be dropped[27]. The effect of outliers was removed using the clipping technique, which did not alter the overall shape of the distribution, at the first and the 99th percentile.[28]. In addition, all continuous variables were normalized in order to compare features and enable all the machine-learning models to converge successfully[29].

Model Performance

The predictive performance of the four machine-learning models differed based on the measures. The best results were obtained with Logistic Regression, with 99 percent accuracy, F1-score of 0.98, precision of 1.00 and a recall of 0.96. Besides good predictive capability, it is interpretable and therefore, it is especially appropriate in regulatory compliance and real-life application in financial systems. These high metrics are likely due to the carefully controlled and internally consistent nature of the synthetic dataset, its label creation process, and not to the generalizable predictive power. The good accuracy of 1.00 in particular, indicates that the model might be

learning the label generation function instead of a real creditworthiness signal. Random Forest also achieved a high accuracy (approx 99%) and a much lower recall of 0.64. This means that there are more false negatives thus is especially critical in credit scoring which may result in missed financial opportunities of identifying creditworthy applicants [8]. The anomaly is not inherent to the model and is caused by either the class imbalance in the data or default threshold miscalibration and will not be a good measure of model performance when used in a credit scoring application. In this study, F1-score and AUC-ROC are therefore considered as the main performance metrics.

XGBoost and LightGBM were both performing equally well, where the accuracy and F1-scores are around 97% and 0.92 respectively. These models were found to be very successful in modelling complex nonlinear relationships, while having relatively low risk of overfitting [13,14]. Overall, the results show that more sophisticated ensemble methods are successful, but less sophisticated models, such as Logistic Regression, can be as or even more successful when combined with effective preprocessing and feature engineering, particularly when interpretability is important. A limitation of this study was that the confidence intervals and statistical significance tests were not computed for these indicators.

Table 2 Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Logistic Regression	0.99	1.00	0.96	0.98	0.99
Random Forest	0.99	0.98	0.64	0.78	0.95
XGBoost	0.97	0.93	0.91	0.92	0.97
LightGBM	0.97	0.92	0.90	0.92	0.97

Generation and Distribution of Credit Score

A credit-score system that could be interpreted was created using the weighted model probability and key financial indicators: 70% model probability, 20% normalized income and 10% normalized age. These scores were graphed against a normal range of credit scores (300-850). The analysis of the box-plots showed that there was an evident distinction of the approved and rejected applicants, where the median credit score of the approved ones was much higher than that of the rejected ones, and there was a minimal overlap between the approved and rejected applicants. This is a very discriminative scoring system. The score distribution was skewed to the higher end of the scores (700-850) among those who were approved, while the lower scores (600-699) were more evenly distributed among the applicants who were not approved. That the score distributions are clearly separated between the approved and rejected categories is consistent with the generation of synthetic data, where the approval labels were obtained from a subset of the same features that scored the data. That should be taken into account when considering the discriminative potential of the scoring system.

Fairness Assessment

Fairness evaluation characteristic chosen was nationality. All models had low values of the Demographic Parity Difference (DP-diff) of under 0.02, which means that the approval rates of the protected and reference group were not much different. The Equalized Odds Difference (EO-diff) was near to zero (below 0.01) as well in Logistic Regression and XGBoost. Random Forest and LightGBM, however, showed slightly higher results of 0.036 – 0.04, which were still acceptable [4,18]. These low disparity values are not surprising as the synthetic data generator intentionally did not create any disparity based on nationality. These results are not therefore a reflection of real-world fairness, and are used to confirm that there was no nationality bias introduced in this simulation when creating data and training models. The nationality was not considered as an explicit model variable, but may have been considered as such implicitly by the other variables that were correlated with it, such as the remittance frequency, sector of employment and income level, which is a possible source of proxy bias that this study can not fully rule out. These findings indicate that without any extra post-processing debiasing, fairness can be enhanced by preprocessing with care, balanced training, and model selection. However, it has limited the analysis as it does not consider multi-attribute fairness or potential for proxy variables [21,26].

Predictive Performance and Ethical Objectives Alignment

Two aspects that need to be matched in AI-driven credit systems are predictive performance and ethical fairness. Interestingly, the best predictive performance model (Logistic Regression) also showed a more equal fairness result. This suggests that there is no conflict between predictive accuracy and fairness in this data set. Unlike the well studied trade-off between fairness and performance in machine-learning systems [5], the trade-off between fairness and performance is not explored in this work. This tension can be reduced by the effective preprocessing and feature selection as indicated in the above study. The observation should be taken with a pinch of salt, however, as the data is synthetic. This accuracy-fairness relationship may not hold true in a real world setting where the data generation process is based on real-world differences in demographics.

Implications for Practice

The results are that AI-driven credit-scoring models can achieve a high accuracy, fairness, and interpretability in a

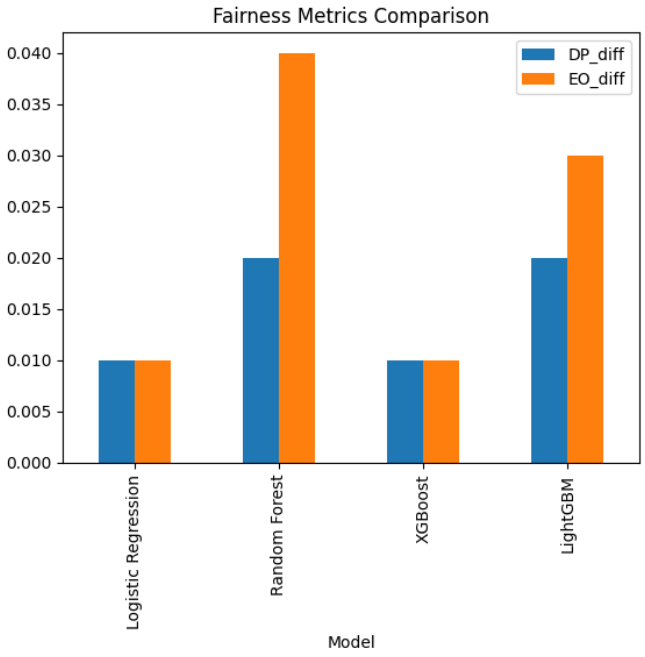

Figure. 1 Fairness Metrics Comparison

controlled simulation environment with the proper integration of alternative data sources. In the simulated data set, the behavior-based or alternative financial indicators were found to be predictive, especially for underbanked profile types that lack a traditional credit history [1,2].

Further, the derived credit scores provide a standardized and interpretable output and can facilitate better communication between the financial institutions and the applicants, enhancing transparency, regulatory compliance and trust among users. These implications are preliminary, and must be confirmed using actual loan data before any applied conclusions can be drawn.

Discussion

This paper investigated the potential of combining conventional financial information with other forms of behavioural information to create predictive credit scores that are accurate, interpretable and fair, using AI-based credit-scoring models. The findings give an understanding of how machine-learning strategies can alleviate constraints of traditional systems of credit-scoring especially among credit-invisible individuals like migrant workers, and those in the informal sector of employment. It is important to note, however, that all findings are from a synthetic simulation environment, and the extent to which the findings would apply to real-world lending contexts is not yet verified. In this section, the results are explained in light of the research objectives, implications are drawn, limitations and future research directions are mentioned.

Interpretation of Model Performance

The results indicate that the performance of the model is not solely determined by the complexity of the algorithms, but also by the quality of the data, data preprocessing, and feature engineering. Although more complex ensemble models, such as the Random Forest, XGBoost, and LightGBM are generally expected to yield better performance than simpler models, the results show that in this experiment, the Logistic Regression was more successful than the other two. This is likely because the synthetic data is linearly separable, and not because Logistic Regression is generally superior to simpler models. Logistic Regression was structurally appropriate for the recovery of the credit approval label as it was generated by a weighted threshold function, which is a linear operation. This does not mean Logistic Regression would be more successful than ensemble methods on real credit data with more complex and non-linear relationships between variables.

Logistic Regression showed that a simple linear model can be as accurate as a more complicated model when the data is well-organized and is well processed [7,27]. This helps in the notion of predictive performance not being entirely reliant on model complexity.

Random Forest, in their turn, had the same accuracy, but a much lower recall (0.64). This would imply that false negatives are higher and therefore, some creditworthy people might have been misclassified as being high-risk. This is particularly important in a credit-scoring context, where it could result in lower levels of financial inclusion and access to opportunities. This is probably because the synthetic dataset is not sensitive to class imbalance or because the default threshold is miscalibrated due to lack of sensitivity to class imbalance. It also sheds light on the reasons why accuracy alone is a poor measure to use for credit scoring and why F1-score and AUC-ROC were used as primary measures in this study.

The capability of these machine learning models to embrace nonlinear relationships and feature interactions in heterogeneous data was the reason for their high and balanced performance (around 97% accuracy and F1-score of 0.92) [13,14]. However, the not so easy-to-understand nature of these is a disadvantage in controlled financial environments.

Table 3 Model Comparison

Model	Predictive Power	Fairness	Interpretability
Logistic Regression	High	High	Very High
Random Forest	High	Medium	Medium
XGBoost	Very High	High	Low
LightGBM	Very High	Medium	Low

The comparison also verifies that the Logistic Regression's

performance is most balanced when the three criteria are all considered. In general, the results indicate that simpler models are better suited in situations where interpretability, transparency and regulatory requirements are emphasized, while not being significantly less accurate in predicting outcomes. The results from this conclusion are only applicable to this synthetic study and cannot be generalized to other situations until they are tested in the field.

Fairness Implications

The fairness test indicates that the models were fairly consistent on the attribute that was being guarded (nationality). Demographic Parity Difference and Equalized Odds Difference were small in all of the models, indicating that there is not much difference in the rate of approval and distribution of error between groups[4,18].

The low disparity values should be taken with a grain of salt, though. The low values of both DP-diff and EO-diff should not be interpreted as a sign of the fairness of the models, since the data generator was not designed to generate such differences in the first place. The fairness results would have been much different if the generator had been programmed to reflect socioeconomic differences that are associated with nationality, as would be the case in actual lending situations.

Logistic Regression and XGBoost gave the best fairness, and some more, but still fair, disparity in Random Forest and LightGBM. These findings indicate that one can enhance fairness by preprocessing carefully, training in a balanced manner, and selectively choosing features. The limitation of the analysis is however the use of a single attribute which is being protected. In the context of a practical credit system there are a number of aspects of fairness which could be overlapping due to gender, age and socioeconomic status. As these cross-over can, they may develop latent biases that cannot be learned by taking a single attribute into account[5].

Moreover, the current study did not look at counterfactual fairness, which would account for different outcomes for individuals with comparable financial circumstances based on demographic differences[21,26]. This is also a valuable topic to be studied in the future. Policymaking-wise, the results indicate the need to incorporate fairness assessment in AI credit systems, such as ongoing auditing and regulatory controls to guarantee fair results among populations.

Interpretability and Operational Use

The proposed system has an advantage in being interpretable, whereby the output of the model is mapped to the standardized credit score scale of 300-850. This conversion makes usability more efficient to financial institutions and makes model outputs to match the pre-defined credit assessment structures.

The addition of the predicted probabilities and with income and age makes it more transparent and allows the stakeholders to have a more clear understanding of how the credit scores are derived. The clear distinction between accepted and rejected applicants (as observed in score distributions) also helps to make the model discriminative.

It is important to note, however, that this separation is partly because of the process of labelling, where the approval is given by a sub-set of the same features that are scored. The distributions of the scores would overlap to a larger extent in a real system.

In financial systems interpretability is particularly important because the actions taken by the financial system should be understandable to regulators and consumers. Non-transparent models have the potential to decrease trust, add regulatory risk, and constrain use in practical lending settings.

By contrast, explainable systems enable borrowers to see how their habits like paying utilities or charging their mobile phones are associated with their credit assessment outcomes. This openness also makes it possible to have some human checks and balances that may allow financial officers to confirm or challenge algorithmic decisions where needed, thus alleviating the risk of automation errors or model drift.

Implications for Financial Inclusion

The implementation of alternative information provides a great deal of capability to the credit-scoring models to evaluate consumers who don't have any conventional credit history. Financial responsibility signals can be provided by behavioural signals such as utility payments, mobile top-ups, and remittance patterns. This strategy is a direct contribution to the goal of financial inclusion, as it proposes a possible route to expanding access to credit for traditionally underserved groups in the formal financial system, including migrant workers and informal-sector workers[1,2].

The implications are tentative and rely on validation in the real world, and the results are demonstrated to be viable in a simulation, but not necessarily in the real world.

The social impacts of increased credit availability are also larger, including increased entrepreneurship, increased financial security in households and increased investment in education and health. However, privacy and responsible handling of data must be taken into account. Consent is not the only concern with the use of behavioral data, misuse and surveillance are also concerns even with anonymized and aggregated data. Thus, to make its application responsible, good governance structures are needed, such as adherence to data protection laws and clear consent procedures[23,25].

Integration with Policy and Practice

The findings of this research have immediate financial regulatory and industry practice implications. This information can help regulators establish more precise measures of fairness, transparency, and explainability in AI-driven credit scoring. This will include ensuring that models are routinely reviewed and are not biased and that models are understandable to institutions as well as to the consumers.

Financial institutions can benefit from the benefits of transitioning to hybrid scoring systems that are not only predictive but also interpretable to enable not only good risk management practices but also inclusive lending practices. And, it is important to constantly be monitored, because model performance and equity can vary over time, depending on economic changes or policy changes, or population changes. Thus, AI governance mechanisms should incorporate real-time surveillance systems, regular retraining, and fairness audits to achieve long-term dependability and ethical adherence.

Conclusion and Future View

In general, the results indicate that AI-driven credit-scoring systems can be highly predictive and fair and interpretable when designed in a suitable manner within a controlled synthetic environment. The results suggest that the performance of the models depends more on the quality of the preprocessing, feature engineering, and evaluation strategy than on the complexity of the algorithm. It is a methodological lesson that careful data preparation can be more important than selecting the model: this is a transferable lesson, but needs further empirical testing for generalisability to real world data.

The study also suggests that predictive performance and fairness do not have to be mutually exclusive goals, especially with properly designed models. However, in a synthetic environment, as in this study, this is not the case as the data generator intentionally did not create any nationality imbalance. Whether this alignment is seen in the real data sets in which there are genuine demographic differences is unknown.

Further studies are needed on the multi-attribute and intersectional fairness and longitudinal validation with real-world data. Further research is also required to evaluate the hybrid explainability methods, like those based on SHAP explanations and counterfactual fairness analysis to enhance transparency without decreasing predictive power. SHAP interpretability analysis was performed during the original research process, but the results are not included in this manuscript; future analyses should include global feature importance plots, examples of local explanations and nationality-stratified comparisons of attribution.

These advancements will enhance the credibility, ethical soundness, and practicality of AI-based credit scoring systems, which will be beneficial in the wider financial inclusion throughout the world.

Limitations

Though this research shows that AI-based credit-scoring models with alternative data can reach high predictive performance without compromising fairness within a synthetic simulation environment, a number of limitations should be considered. These are related to the constraints of the data, methods, fairness assessment, ethics, and real world implementation. It is important to note these limitations to properly interpret the findings and inform future studies.

Data-Related Limitations

One of the primary limitations of this study is in the representativeness of the data set. Even though the synthetic dataset was created to be more realistic in terms of financial and demographic distribution, it might not be representative of the variety of the real-life applicants of credit. In particular, it may not be representative of high-risk populations, or even of the entire unbanked population. This can limit the applicability of the model to a wider range of actual lending scenarios[1]. Another and more basic restriction on data is label-generation circularity. The binary credit approval label is based on a subset of the same features that are used as the model inputs, and is generated by a weighted threshold function. This means that the models are learning a version of the label generation rule that is more or less accurate, not a true creditworthiness signal (which is most likely the case), and explains the near-perfect performance metrics seen. These performance metrics are internally valid, in that there is no external validation, and therefore they are meaningless.

Moreover, although the fairness measures using nationality showed that there were only minor inequalities, there is still a possibility of proxy bias, in which other factors like occupation, income level, or type of employment could indirectly capture demographic data. This could lead to unintentional and/or implicit discrimination that could not be identified by a direct assessment of fairness[4]. Furthermore, as the disparity in nationalities was not introduced by design in the synthetic generator, the low disparity values in fairness are not meaningful results but rather by design.

Lastly, the dataset is cross-sectional, and lacks temporal dynamics. Consequently, the models are unable to explain time-varying financial behavior, including those due to economic cycles, inflation or job insecurity. This restricts the capacity to assess the long-term predictive stability as well as equity performance.

Methodological Limitations

The models have been tested on a held-out test set, without external testing on independent datasets of other populations or economic circumstances. This hinders the extrapolation of the results and introduces doubt as to how the models would behave in the real-life deployment environment[7]. To quantify, there was no baseline model built using only traditional financial data to compare with. Therefore it is not possible to measure the effectiveness of the alternative data in comparison to the traditional financial data and credit scoring data. It is important to make this comparison in future studies. Moreover, the research was devoted to four machine-learning models such as the Logistic Regression, Random Forest, XGBoost, and LightGBM. These have seen widespread use in credit scoring, but other approaches that are fair and/or interpretable, such as adversarial debiasing or fairness constrained optimization methods, were not examined. This restricts the extent of comparison and might rule out models that could offer better fairness-performance trade-offs[21].

Computational feasibility was also a constraint to hyperparameter tuning. Even with reasonable optimization, it may be possible to improve the performance of the model by tuning it more, but this may also result in overfitting. In addition, the original analysis code and outputs have not been kept, reducing the ability to reproduce this study. There were no additional metrics of credit scoring (PR-AUC, calibration error, Brier score) reported, nor were confidence intervals and statistical significance tests calculated. These are identified weaknesses which future research should consider. A second limitation of reproducibility is that the SHAP interpretability analysis was conducted as part of the original research process, but is not included with the outputs.

Equity and Ethical Restrictions

One of the key weaknesses of the fairness assessment is the consideration of a single attribute that is protected, nationality. Other important demographical factors such as gender, age, socio-economic status or disability were not taken into consideration. This may lead, as a consequence, to the research not being representative of the intersectional or multidimensional bias that is prevalent in actual credit systems[5]. The study was also not conducted with individual level counterfactual fairness testing. This restricts the possibility to decide whether people with the same financial profile but different demographic features would get the same results[26]. While the data used in this study is synthetic and anonymized, in other environments where alternative behavioral data are used, there are important issues surrounding ethics, primarily privacy, consent and potential misuse of sensitive personal data. These were explored but not measured in this study.

Practical and Deployment Limitations

Although the model has shown good theoretical performance, there are other challenges in implementing the model in real life. These include regulatory compliance, integration with the existing banking system, and explainability in high-stakes financial decision making[2]. The other critical weakness is the possibility of model drift in such a way that variations in economic conditions or population behavior over time can decrease the model accuracy and fairness. This would require constant monitoring, retraining and auditing to guarantee reliability in the long term. In addition, too much reliance on automated decision-making processes can reduce human oversight, leading to unintended errors or a lack of awareness in credit assessments.

Summary

On the whole, although the results of this paper are encouraging and reveal the potential of AI-based credit scoring in enhancing its predictive power and in increasing its fairness, they need to be viewed through a set of constraints. They include limitations in data representativeness, limitations in circularity of label generation, limitations in synthetic data overseparability, code and output unavailability, limited scope of evaluating fairness, methodological limitations, and practical deployment. These questions will be important in future research to develop more robust, generalizable and moral credit-scoring systems.

Conclusion & Recommendations

This work has developed and evaluated an AI credit-scoring simulation to investigate predictive performance and fairness by leveraging traditional financial data with other behavioral factors in a synthetic setting. The study fills the gap of conventional credit-scoring models, most of which lack the ability to capture non-linear financial behavior and are often insensitive to credit-invisible groups like migrant workers and workers in the informal sector[1]. To determine if AI systems could help make credit decisions more inclusive and transparent, the predictive performance, fairness metrics, and interpretability of the models under test – Logistic Regression, Random Forest, XGBoost, and LightGBM – were evaluated. All results are preliminary and based on simulation conditions; not proven in the real world.

Research Purpose and Scope

This study was aimed at specifying and testing an AI-based credit-scoring model that would strike a balance among three

important goals, namely, predictive performance, fairness, and interpretability. Traditional credit-scoring systems often rely heavily on historical financial data, which can limit their usefulness for those who don't have a long credit history and can contribute to demographic bias. To fill this void, the researcher applied four machine learning models: Logistic Regression, Random Forest, XGBoost and LightGBM, to a structured dataset of demographic, financial and behavioral variables. The study focused on the performance of AI models in the binary classification of credit approval, as well as on the equality of treatment across nationality groups based on Demographic Parity Difference and Equalized Odds Difference, and the interpretability of the model results in a standardized credit-score format.

Key Findings

The study findings suggest that when the conditions are well structured in the synthetic dataset, the predictive and fairness performance are good. Preprocessing of the data, such as outlier management, normalization and multicollinearity, enhanced the stability of the model and provided consistency between the algorithms. This was needed to make sound comparisons between models. Logistic Regression had the highest predictive performance with 99% accuracy and 0.98 F1 score. Random Forest demonstrated the same accuracy but much lower recall (0.64), which means more false negatives. The accuracy was 97% with the F1-scores of 0.92, which is a good performance in handling with nonlinear relationships, and the results of XGBoost, LightGBM were similar. The disparities were low across all models, as indicated by the evaluation of fairness: Demographic Parity Difference was not greater than 0.02 and Equalized Odds Difference was not greater than 0.04. The low disparity values are not a reflection of fairness in the real world, but arise from the process of generating the synthetic data. The model produced credit scores ranging from 300 to 850 and showed a substantial difference between credit scores of approved and denied applicants. Part of this separation may be due to circularity in the labeling process and not to discriminative power. Overall, it is found that it is feasible to obtain internal consistency between fairness and predictive performance by carefully designing data preprocessing and feature engineering in a simulation environment.

Knowledge and Practical Contributions

The research has a contribution to the current body of literature by showing that less complex models like Logistic Regression can compete with more complex ensemble models when trained on well-prepared data in a synthetic setting. This implies (but does not prove) that predictive performance is not solely dependent on model complexity. Based on the theory, the findings suggest that alternative behavioral data could potentially be a viable addition to traditional financial measures for underserved populations, pending empirical testing. In terms of applications, the proposed framework provides a practical example of the applicability of interpretable and efficient models in low-resource or regulated environments. Also, the fairness evaluation integration offers a systematic way of auditing bias in credit-scoring systems.

Policy and Industry Recommendations

The findings of this research have implications for policy makers and financial institutions, but these are only suggestive and based on a simulation study rather than evidence-based recommendations. Regulators ought to promote the use of explainable and transparent AI in financial decision-making. This involves setting out clear fairness standards, responsible usage of alternative data, and supportive open banking systems with robust consent and privacy rights[23]. In the industry perspective, hybrid credit-scoring systems are recommended that can provide balance between predictive accuracy and interpretability to the financial institutions. To guarantee long-term reliability, regular fairness audits, constant monitoring, and updates of the model should be implemented. Also, companies ought to invest in enhancing the consumer perception of AI-based credit decisions by making them more trustful and transparent.

Future Research Directions

Some future research is needed to expand this work in some directions. First, the research should look at models tested on a multi-country or multi-institution level, to make the models more relevant to different economic and demographic contexts. Second, longitudinal studies need to be conducted to understand the evolution of model performance and fairness in realistic settings over time. Third, future studies should also incorporate other data sources to augment predictive coverage, such as behavioral, network-based and geospatial features. Fourth, multi-attribute and intersectional frameworks such as gender, age, and socioeconomic background should be included in fairness analysis to address more complicated types of bias. Fifth, for future comparisons, a baseline model with traditional features should be included to provide a measure of the value of other data. Sixth, SHAP interpretability analysis must be conducted and communicated in full detail, including nation-wise feature importance plots, examples and explanations, and nation-wise comparisons of attribution. Seventh, other evaluation metrics such as PR-AUC, calibration error, Brier score and bootstrap confidence inter-

vals should be reported to give a more stringent assessment of the model performance. Finally, other approaches to explainability such as SHAP and counterfactual explanations should be explored further to improve their interpretability without compromising their predictive power.

Overall, this paper demonstrates the potential of AI credit-scoring systems in a synthetic simulation setting and the potential of responsible AI frameworks to increase financial inclusion, provided they are tested with real-world data. The methodological and exploratory contribution of this study is to offer a structure for the evaluation of the performance of alternative-data credit scoring from the predictive performance and fairness and interpretability angles, and to give clear directions for future empirical research.

Acknowledgments

The author would like to thank the reviewers for their detailed and constructive feedback, which substantially strengthened this manuscript.

References

1 M. Bazarbash, Expanding financial inclusion through digital credit scoring. International Monetary Fund, 2019., https://www.imf.org/en/Publications/WP/Issues/2019/08/23.

2 J. Manyika and M. Chui and J. Bughin and R. Dobbs and P. Bisson and A. Marrs, Digital finance for all: Powering inclusive growth in emerging economies. McKinsey Global Institute, 2016., https://www.mckinsey.com.

3 World Bank, Global Findex database 2021: Financial inclusion and access. 2021., https://globalfindex.worldbank.org.

4 S. Barocas and M. Hardt and A. Narayanan, Fairness and machine learning: Limitations and opportunities. MIT Press, 2019.

5 N. Mehrabi and F. Morstatter and N. Saxena and K. Lerman and A. Galstyan, A survey on bias and fairness in machine learning. ACM Computing Surveys. Vol. 54, pg. 1–35, 2021., https://doi.org/10.1145/3457607.

6 FICO, Understanding credit scores and credit risk. n.d., https://www.fico.com.

7 D. W. Hosmer and S. Lemeshow, Applied logistic regression. 2nd ed. Wiley, 2000.

8 L. Breiman, Random forests. Machine Learning. Vol. 45, pg. 5–32, 2001., https://doi.org/10.1023/A:1010933404324.

9 Y. Bao and J. Huang, Leveraging alternative data for inclusive credit scoring: Evidence from emerging markets. Journal of Financial Technology. Vol. 3, pg. 45–63, 2021.

10 J. Jagtiani and C. Lemieux, The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub dataset. Journal of Economics and Business. Vol. 105, pg. 1–21, 2019., https://doi.org/10.1016/j.jeconbus.2019.04.001.

11 F. Li and L. Wang and X. Li, Using alternative data in credit scoring: Behavioral, psychometric, and digital footprints. Journal of Financial Services Research. Vol. 58, pg. 367–392, 2020.

12 Deloitte, Alternative data in financial services: Expanding inclusion and managing risk. Deloitte Insights, 2021., https://www2.deloitte.com.

13 T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pg. 785–794, 2016., https://doi.org/10.1145/2939672.2939785.

14 G. Ke and Q. Meng and T. Finley and T. Wang and W. Chen and W. Ma and T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems. Vol. 30, pg. 3146–3154, 2017.

15 D. R. Cox, The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological). Vol. 20, pg. 215–242, 1958.

16 S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. Vol. 30, pg. 4765–4774, 2017.

17 H. Nguyen and T. Tran and D. Le, Hybrid credit scoring using linear and ensemble models for fintech applications. Expert Systems with Applications. Vol. 160, pg. 113653, 2020., https://doi.org/10.1016/j.eswa.2020.113653.

18 C. Dwork and M. Hardt and T. Pitassi and O. Reingold and R. Zemel, Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pg. 214–226, 2012., https://doi.org/10.1145/2090236.2090255.

19 M. J. Kusner and J. Loftus and C. Russell and R. Silva, Counterfactual fairness. Advances in Neural Information Processing Systems. Vol. 30, pg. 4066–4076, 2017.

20 J. Kleinberg and S. Mullainathan and M. Raghavan, Inherent trade-offs in the fair determination of risk scores. Proceedings of Innovations in Theoretical Computer Science. pg. 43:1–43:23, 2018., https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

21 Y. Zhang and B. Lemoine and M. Mitchell, Mitigating unwanted bias with adversarial learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pg. 335–340, 2018., https://doi.org/10.1145/3278721.3278779.

22 Fairlearn, Fairlearn: A toolkit for assessing and improving fairness in AI. n.d., https://fairlearn.org.

23 European Commission, Revised payment services directive (PSD2). 2015., https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32015L2366.

24 MAS, Guidelines on responsible use of AI and data analytics in financial services. Monetary Authority of Singapore, 2020., https://www.mas.gov.sg.

25 UK Open Banking Implementation Entity, Open banking standards and regulatory compliance. n.d., https://www.openbanking.org.uk.

26 S. Wachter and B. Mittelstadt and C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology. Vol. 31, pg. 841–887, 2017.

27 R. M. O'Brien, A caution regarding rules of thumb for variance inflation factors. Quality & Quantity. Vol. 41, pg. 673–690, 2007., https://doi.org/10.1007/s11135-006-9018-6.

28 B. Iglewicz and D. C. Hoaglin, How to detect and handle outliers. ASQC Quality Press, 1993.

29 Scikit-learn contributors, Scikit-learn: Machine learning in Python. n.d., https://scikit-learn.org.

Appendix A: Distribution Choices

The synthetic dataset was chosen to be privacy-conscious to the extent to ensure privacy violations were reduced, but this must reproduce authentic economic and demographic heterogeneity and thus allow AI credit-scoring experimentation to be rigorous. Salary was drawn from a truncated normal distribution bounded by minimum wage and average income of migrants. Duration of employment followed a skewed exponential distribution reflecting short-term employment patterns. Utility bill payments were modelled using a beta distribution that focuses on steadiness. Phone top-ups were generated using a Poisson distribution modelling digital and non-continuous periodic payments. Remittance frequency was treated as a nominal variable with values of weekly, bi-weekly, and monthly. This approach maintains reasonable variability while protecting personal privacy.

Appendix B: Preprocessing Information

Data pre-processing was done to guarantee quality, reliability and model preparedness. Non-numeric features including profession, sector, education, and nationality were one-hot encoded to reduce computational complexity and preserve interpretability. Continuous valued variables including age, salary, rent, employment duration, utility bills, and phone top-ups were normalized or min-max scaled. Missingness was synthetically generated at approximately 10%, with numeric data imputed to the median and discrete data imputed to the mode. Values above the 1st and 99th percentile were capped to prevent biased predictions. Variability inflation factor analysis indicated a low level of multicollinearity redundancy. Such protocols are consistent with the best practices in applied financial AI studies, which facilitates predictive performance as well as fairness assessment.