Revisiting Nighttime Light Features in Inclusive Credit Scoring: A Fairness Analysis for Emerging Markets

Jiayi Yin
University of Wisconsin–Madison
Madison, Wisconsin, USA
jyin66@wisc.edu

Yixi Zhou
ShanghaiTech University
Shanghai, China
zhouyx2022@shanghaitech.edu.cn

Abstract

Artificial intelligence is increasingly used to expand credit access in emerging markets, often relying on alternative data sources such as satellite imagery. Among these, nighttime light (NTL) intensity is a popular proxy for local economic activity, yet its use in microcredit scoring may inadvertently encode geographic or gender bias. In this paper we evaluate the predictive and fairness implications of NTL features in loan-approval models for Peru and the Philippines using Kiva data. We design baseline and class-balanced training tasks, fit logistic classifiers with and without NTL, and audit outcomes using demographic parity, equal opportunity, and average odds across gender and rural–urban groups. We further test robustness with province and sector fixed effects and conduct threshold-sensitivity analysis. Across countries, tasks, and specifications, adding NTL yields negligible accuracy gains ($\Delta\text{AUC} \leq 0.01$) and produces minimal, non-systematic changes in fairness metrics. Observed disparities mainly reflect data imbalance rather than the satellite proxy itself. These findings highlight the limits of remote-sensing features in fairness-aware credit scoring and underscore the need for careful AI auditing when deploying alternative data in inclusive finance.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Economics**.

Keywords

financial inclusion, credit scoring, nightlights, fairness, emerging markets, gender bias, remote sensing

ACM Reference Format:
Jiayi Yin and Yixi Zhou. 2025. Revisiting Nighttime Light Features in Inclusive Credit Scoring: A Fairness Analysis for Emerging Markets. In . ACM, New York, NY, USA, 8 pages.

1 Introduction

Artificial intelligence (AI) and machine learning (ML) are transforming credit access in emerging economies. Yet more than 1.3 billion people worldwide remain unbanked, with financial exclusion concentrated in low-income and middle-income countries [5, 18]. In the absence of traditional credit histories, lenders increasingly turn to alternative data such as mobile phone metadata, utility payments, social networks, and satellite imagery. Among these, nighttime light (NTL) intensity captured by sensors such as the Visible Infrared Imaging Radiometer Suite (VIIRS) has become a widely used proxy for economic activity and wealth. Prior literature demonstrates that aggregate nightlight measures correlate strongly with GDP and household wealth indices in urban areas [3, 15], but their predictive power diminishes in rural, agricultural regions and may encode spatial or socioeconomic biases [11]. Since women are disproportionately represented in rural and informal economies, where home-based enterprises emit little detectable light, reliance on NTL features may inadvertently reinforce gender gaps in credit access.

This study re-examines the role of NTL features in credit-scoring models for low-income and middle-income countries. We focus on Peru and the Philippines, two diverse emerging markets, using loan data from the Kiva microfinance platform. Our preliminary analysis indicates that regional NTL intensity exhibits a weak correlation with the proportion of female borrowers. Furthermore, regression models with year fixed effects do not find a significant association between the two. Building on this foundation, we extend our experiments to predictive tasks that mirror operational credit scoring. Specifically, we compare logistic regression models trained on baseline features (loan amount, sector, region, and socioeconomic controls) with and without NTL; evaluate fairness across gender (female-only vs. others) and rural communities (low-NTL quantiles); incorporate province- and sector-fixed effects; and analyze performance–fairness trade-offs across decision thresholds.

We frame our analysis around three guiding research questions: first, we ask to what extent the incorporation of nighttime light data improves the predictive accuracy of credit scoring models; second, we investigate whether the inclusion of such data amplifies or mitigates disparities across gender and rural-urban groups; and third, we assess the robustness of these effects when evaluated under alternative fixed-effect specifications and thresholding strategies.

Our contributions are fourfold: (i) we construct a novel ADM1-year panel combining Kiva microcredit data with VIIRS NTL and socioeconomic controls for two emerging markets; (ii) we implement a fairness auditing framework for credit scoring, measuring demographic parity, equal opportunity, and average odds gaps; (iii) we assess robustness under alternative fixed-effect structures and threshold sensitivities; and (iv) we discuss policy implications for the responsible use of remote-sensing proxies in inclusive finance.

We conduct our work in the context of the ICAIF 2025 FinREM Workshop on *AI for Financial Inclusion, Risk Modeling and Resilience*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
FinREM@ICAIF '25, Singapore
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

in Emerging Markets. Our findings caution against indiscriminate use of remote-sensing proxies and underscore the need for rigorous fairness auditing in emerging market credit scoring.

2 Related Work

Satellite Nighttime Lights as Proxies for Economic Activity. Satellite-derived nighttime lights (NTL) have been widely used as proxies for economic activity, urbanization, and infrastructure in regions with limited data [6]. Early studies with DMSP-OLS showed strong correlations between aggregate luminosity and GDP at national and regional levels [3]. With the higher-resolution VIIRS sensor since 2012, researchers have refined these estimates. For instance, Gu et al. [12] combined annual VIIRS radiance with survey data to improve regional economic measurement in China. NTL provides a globally consistent, fine-grained proxy for socioeconomic progress where ground truth is scarce.

NTL for Poverty Mapping and Economic Prediction. Building on this foundation, recent work has used NTL in machine learning frameworks for poverty and wealth prediction [1]. Jean et al. [16] pioneered a transfer-learning approach, predicting village-level outcomes from daytime imagery and nightlight intensity, substantially improving poverty mapping in Africa. Bruederle and Hodler [2] confirmed strong associations between NTL and local development indicators, while Chi et al. [4] produced global relative wealth indices by integrating NTL with multi-source features. Collectively, these advances demonstrate NTL's potential in low-data environments but also highlight its *context dependence*—accurate in urban and electrified areas, but weaker in rural or off-grid regions [9].

Limitations and Biases of NTL Data. Despite its promise, NTL data face well-documented limitations. Older sensors like DMSP-OLS suffered from saturation and blooming effects, while even VIIRS fails to capture very low levels of light. This leads to systematic underrepresentation of sparsely electrified or rural populations. It is widely recognized that nearly one quarter of the global population, especially in rural or conflict-affected areas, lacks detectable nightlight [2, 8]. In addition, non-linear factors such as gas flaring or cultural lighting further complicate interpretation. These biases create an inherent *urban skew*, motivating caution in downstream applications such as credit scoring.

Inclusive Credit Scoring in Emerging Markets. In emerging markets, where formal credit histories are scarce, fintech lenders increasingly turn to alternative data such as mobile phone usage, utility payments, and geospatial signals like NTL. Mobile-based scoring has shown promise for expanding credit access [10]. Yet biases also emerge: gender-blind models have disadvantaged women despite equal or superior repayment rates [20], and NTL-based features risk undervaluing rural borrowers if calibrated to urban data. These concerns align with the broader fair finance community's emphasis on *fairness auditing* and inclusive model design.

Fairness and Bias in Algorithmic Credit Scoring. Our study also builds on the fair ML literature, which formalizes criteria for detecting and mitigating bias in predictive models. Demographic parity (equal approval rates) and equal opportunity (equal true positive rates) are particularly relevant in lending contexts [13]. Prior work shows that such fairness metrics often conflict with accuracy, requiring trade-offs such as threshold adjustments [19]. These frameworks have been applied to audit lending models for disparate impact across race or gender and to design interventions such as reweighting or threshold calibration. By situating NTL-driven credit scoring within this literature, we ask whether incorporating remote sensing features introduces measurable fairness disparities, especially along gender and rural–urban dimensions in emerging markets.

3 Data and Panel Construction

3.1 Loan and borrower information

We build our dataset from the *Kiva* open microfinance release, which provides anonymized loan-level records with borrower and transaction details. Each record includes a unique loan identifier, the geographic coordinates of the borrower (latitude and longitude), gender information (labeled as "female," "male," or mixed group), loan sector, loan amount, the dates when the loan was posted and funded, and the repayment status.

A key challenge is the heterogeneity of location reporting: borrower locations may appear as free-text province names, coordinates, or partner-specified regions. To achieve harmonized regional identifiers, we implement a three-stage mapping procedure: (i) *Text normalization:* We normalize free-text administrative labels (e.g., "Lima," "Lima Region," "Province de Lima") through lowercasing, accent removal, and fuzzy string matching against official FAO GAUL ADM1 names; (ii) *Spatial join:* For records with valid coordinates, we perform a point-in-polygon assignment using GAUL ADM1 shapefiles. The implementation relies on shapely's STRtree spatial index, which allows efficient batch matching of tens of thousands of points to province boundaries; and (iii) *Partner fallback:* For residual unmatched loans (often due to missing coordinates or ambiguous text), we impute ADM1 as the modal region of other loans from the same microfinance partner, effectively leveraging partner-level clustering of lending activity.

Table 1 summarizes the dataset characteristics for both countries. After applying this procedure, our ADM1 coverage reaches **99.8% for Peru** (about 22,000 loans) and **97% for the Philippines** (about 160,000 loans), ensuring nearly all loans can be included in regional aggregation and fairness analysis. This high coverage minimizes bias from excluding hard-to-map rural borrowers.

Table 1: Dataset overview by country. Rural loans are defined as those in regions with nighttime light intensity below the 40th percentile of log(NTL).

Metric	Peru	Philippines
Total loans	22,233	160,441
Time period	2013–2017	2013–2017
ADM1 regions	23	19
ADM1 mapping coverage	99.8%	97.0%
Female-only loans (%)	50.2%	94.7%
Approval rate (%)	97.2%	98.2%
Mean log(NTL)	0.327	0.258
Rural loans (≤40th pctl, %)	49.4%	38.1%

3.2 Nighttime light and contextual features

To capture regional economic activity, we extract annual mean radiance values from the VIIRS nighttime lights (NTL) composites curated by the Earth Observation Group [8]. These images provide cloud-free, monthly-averaged radiance at ~500m resolution. For each ADM1 region, we calculate the mean annual radiance. Because raw NTL values are highly skewed (urban centers are extremely bright), we apply a 1–99 percentile winsorization and then take the natural logarithm (with a small positive offset for zero values), yielding log_ntl. This transformation compresses the scale while limiting the influence of outliers.

Recognizing that nightlights alone are imperfect proxies, we incorporate additional covariates as controls: (i) *Population density:* WorldPop gridded estimates for 2015 and 2020, averaged spatially to ADM1; (ii) *Cropland share:* Derived from the GHS-SMOD dataset, indicating the proportion of land used for agriculture; (iii) *Built-up share:* From GHS-SMOD, capturing the proportion of urbanized surface area; (iv) *Partner rating:* Microfinance partner institution rating where available.

These variables control for structural regional differences beyond nighttime illumination.

Table 2: Variables and data sources used in this study. ADM1–year indicates region-year aggregates; loan-level indicates features attached to each loan.

Variable	Description	Source
log_ntl	Winsorized (1–99%) log(1 + radiance)	VIIRS EOG
pop_density	Population per km^2 (2015/2020)	WorldPop
cropland_share	Share of cropland area	GHS-SMOD
builtup_share	Share of built-up/urban area	GHS-SMOD
loan_amount	Requested loan amount (USD)	Kiva
sector	Economic sector (one-hot)	Kiva
gender_group	Female-only vs. other (male/mixed)	Kiva
approved	Fully funded (=1) vs. not (=0)	Kiva
ADM1_NAME	Province/region identifier	GAUL + mapping
rural_flag	1 if log(NTL) $\leq$ 40th pct (train)	Derived

3.3 Panel construction

We construct a balanced panel of ADM1–year cells covering 2014–2017 for both Peru and the Philippines. Each cell aggregates loan activity and contextual features for a given region and year. The following statistics are computed: (i) *Loan volume:* number of loans and total borrowers; (ii) *Gender composition:* number of female borrowers and the *female ratio* (fraction of borrowers who are female); (iii) *Approval rate:* proportion of loans that were fully funded (funded amount $\geq$ requested amount); (iv) *Average* `log_ntl`*:* mean nighttime light radiance for the region-year; and (v) *Sector distribution:* proportion of loans by sector category.

For predictive tasks, we retain loan-level records annotated with borrower gender, sector, and mapped ADM1 region. Each loan is assigned the `log_ntl` value corresponding to its region and year, enabling both individual-level classification and regional aggregate analysis.

To evaluate fairness by rurality, we define rural cells using a *quantile-based threshold* on the training distribution of `log_ntl`. Specifically, ADM1–year cells with radiance below the 40th percentile are labeled as rural. This data-driven definition reflects the relative scarcity of nighttime illumination and ensures approximately 40% of loans in each country are labeled rural, providing sufficient sample size in both groups for statistical comparisons. The binary rural label serves as a proxy for remoteness or level of development and is treated as a sensitive attribute in fairness analysis.

3.4 Challenges and limitations

Despite the high mapping coverage, several challenges arose during data processing: (i) *Sparse loan counts in some regions:* A number of ADM1–year cells contain only a handful of loans (or even a single borrower), particularly in remote provinces. These small samples inflate variance in computed female ratios or approval rates, and make regression coefficients sensitive to outliers; (ii) *Ambiguity in free-text locations:* Province names were often misspelled, abbreviated, or reported in local dialects, requiring extensive normalization and fuzzy string matching. While spatial joins solved many cases, a residual fraction still required partner-based imputation, which may bias against smaller partners concentrated in rural areas; (iii) *Incomplete or noisy metadata:* Borrower gender was occasionally missing or reported as "mixed group," complicating the construction of clean female-only vs. male-only samples. Loan timestamps were not always consistent, and some repayment outcomes were unreported; and (iv) *Scale mismatch between NTL and credit data:* Nightlight radiance captures broad economic activity, but Kiva loans are issued to individuals or small groups. This mismatch means that the signal-to-noise ratio of NTL at the ADM1 scale may be limited, especially for sparsely populated or agricultural regions.

These challenges underscore why careful preprocessing and robustness checks are necessary. They also highlight that fairness disparities observed in the modeling stage may partly reflect underlying data sparsity and noise, rather than purely algorithmic bias.

3.5 Alignment with research objectives

This data pipeline is tightly aligned with our research goals. By harmonizing free-text and coordinate-based locations into ADM1 regions, we ensure that female participation and loan approval rates can be systematically compared against contextual radiance levels. By incorporating both baseline socio-economic features and NTL, our panel supports a head-to-head evaluation of whether satellite proxies meaningfully improve predictive accuracy and whether they introduce fairness disparities across gender or rural–urban groups. The high coverage rates (near-complete for both countries) are particularly crucial, as they allow us to focus on substantive fairness issues rather than artifacts of missing data.

4 Methodology

4.1 Descriptive analysis and linear models

Exploratory correlation analysis. Following prior literature [17], we begin with correlation and regression analysis to establish whether nighttime light intensity is systematically associated with the gender composition of borrowers. We first explore the association between a region's nighttime light intensity and the gender composition of its borrowers at an exploratory level. Using the aggregated ADM1-year data, we compute the Pearson correlation

between log_ntl and the fraction of loans given to female borrowers in that region-year.

In Peru, we observe a weak negative correlation (around −0.1 to −0.2 depending on year), suggesting that regions with higher night light (more urbanized regions) have a slightly lower proportion of female borrowers. In the Philippines, by contrast, the correlation is mildly positive (approximately +0.1), indicating that brighter regions see a higher fraction of female borrowers, although the relationship is quite weak in magnitude in both cases.

Figure 1 illustrates these patterns at the region level: the scatterplots of female-loan fraction vs. log_ntl show a slight downward trend in Peru and a slight upward trend in the Philippines. This exploratory finding hints that gender and local development (as proxied by NTL) are not strongly coupled, and that any gender-NTL relationship may differ by context. Before drawing conclusions, we assess the relationship more formally with regression, accounting for potential confounders like time trends.

Weighted least squares regression. To quantify the gender–NTL association while controlling for time, we perform a weighted least squares (WLS) regression on the ADM1-year panel. Specifically, we estimate regressions of the form:

$$\text{female ratio}_{r,t} = \beta_0 + \beta_1 \log(\text{NTL}_{r,t}) + \gamma_t + \varepsilon_{r,t}, \tag{1}$$

where γ_t are year fixed effects and weights are

$$w_{r,t} = \sqrt{\text{loan count}_{r,t}}, \tag{2}$$

to reflect sampling variation. All regressions are estimated with heteroskedasticity-robust standard errors clustered at the ADM1 level.

We weight observations by the square root of loan count because region-years vary substantially in sample size—some contain only a handful of loans while others have hundreds. The square-root weighting is a compromise between equal weighting (which ignores precision differences) and full frequency weighting (which would allow a few populous regions to dominate), giving more credibility to rates computed on larger samples while moderating the influence of outliers.

The WLS results indicate no strong or significant linear relationship between log_ntl and female-borrower share once year effects are controlled. In both countries the coefficient on log_ntl is small in magnitude and not statistically distinguishable from zero, consistent with the weak raw correlations observed in Figure 1. Thus, the inclusion of NTL as a predictive feature is unlikely to directly encode gender disparities, though indirect effects may still emerge in classification models and are evaluated through fairness metrics in Section 5.

4.2 Predictive modeling

4.2.1 Classification task and imbalanced data. We frame loan approval as a binary classification task. The Kiva dataset is highly imbalanced (roughly ∼97–98% approved, ∼2–3% rejected), posing challenges for model training—a naive classifier could achieve high accuracy by predicting all loans as approved.

To address this, we experiment with two training strategies, constructing two task variants: (i) *Approval baseline (full dataset):* uses the full loan sample with the original imbalance. Since more than 97% of Kiva loans are fully funded, this task is highly imbalanced and serves as a ceiling benchmark; and (ii) *Approval balanced:* downsamples the majority class in the training set to a roughly 50/50 split, forcing the classifier to learn minority-class signals rather than defaulting to the majority.

We still evaluate performance on the original distribution (and also report metrics like area under the precision–recall curve that are sensitive to imbalance), to ensure that the results are meaningful for the real class proportions. The combination of full-data training and balanced-training allows us to compare model behavior in a standard scenario versus one explicitly focused on minority-class prediction.

4.2.2 Model features – Baseline vs PlusNTL. For each task, we estimate two logistic regression models to isolate the impact of the NTL proxy: (i) Baseline model: uses non-NTL features including loan amount, sector, ADM1 dummies, socio-economic covariates (population density, cropland share, built-up area), and partner rating where available – essentially all available features that might be used in a typical creditworthiness model, aside from NTL; and (ii) PlusNTL model: additionally includes the log(ntl) feature for the loan's region and year.

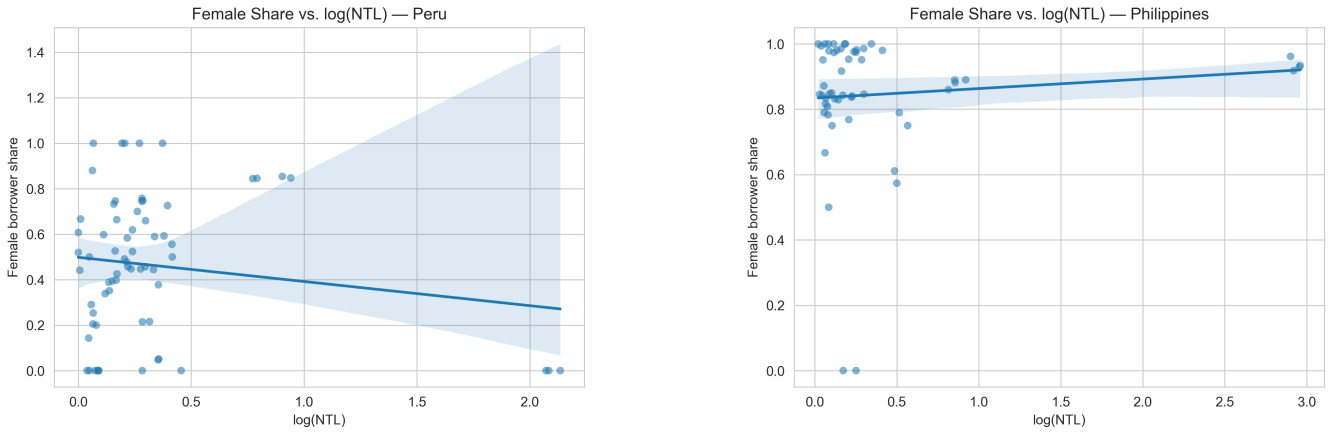
By comparing these models, we isolate NTL's incremental contribution to predictive performance and fairness. Models are trained separately for each country, with standard preprocessing (normalization, one-hot encoding). We chose logistic regression for its interpretability and as a common baseline in fair-ML studies; it also allows us to easily incorporate fixed effects as additional dummy variables in later analyses.

4.2.3 Training procedure and evaluation metrics. Models are trained on 70% of the sample with the remaining 30% held out for testing. We randomly split each country's dataset (stratified by class to ensure the rare negative cases appear in both sets). We repeat all experiments with five random seeds and report average performance to ensure robustness.

We evaluate the classification performance using several standard metrics: (i) *AUC (Area Under ROC Curve):* measures overall ranking performance, threshold-independent; (ii) *Precision–Recall AUC (PR AUC):* more informative on imbalanced data, focusing on the detection ability of minority classes; and (iii) *F1 score:* harmonic mean of precision and recall for the positive "approved" class, computed at a decision threshold (we use the default 0.5 threshold or adjust if using the balanced training where 0.5 corresponds to equalized classes).

The AUC metrics are threshold-independent, giving a sense of inherent model discriminative ability, while F1 is computed at a decision threshold. We also report accuracy, though it is less meaningful under imbalance. All logistic models are fit with L2 regularization (with strength tuned via default or simple validation) to prevent overfitting, and we ensure convergence with a high iteration limit. No class-weighting is applied in the full-data model (since we compare to the undersampled case), meaning its loss is dominated by the positive class—a contrast to the balanced model.

By comparing the Baseline vs. PlusNTL model performances (particularly any increase in AUC or recall for the minority class when log_ntl is added), we assess the predictive value of the NTL proxy. More importantly, we will examine how adding NTL affects



(a) Peru: female share vs. log(NTL).

(b) Philippines: female share vs. log(NTL).

Figure 1: Scatter of female borrower share against log(NTL) at the ADM1–year level.

the fairness of the predictions with respect to borrower gender and rurality, as described next.

4.3 Fairness evaluation

After training the predictive models, we evaluate their fairness across sensitive groups of interest. In this study, we focus on two attributes that are pertinent to loan fairness: borrower gender and community rurality. We assess group fairness along two sensitive dimensions:

4.3.1 Definition of protected groups. Each loan in the dataset is labeled for these attributes so that we can identify which group it falls into:

Gender groups. We define the protected gender group as female borrowers (female-only borrowers vs. all other loans). In practice, Kiva loans can be to individuals or groups; we categorize a loan as "female-only" if all the borrowers listed are female, and as "other" if there is at least one male borrower (including male-only or mixed-gender groups). The female-only group is treated as the disadvantaged class for fairness evaluation, reflecting common concerns that female borrowers might be underserved or face bias (even though in microfinance, females often have high participation, it is still valuable to check for model bias). The "other" (male or mixed) group serves as the reference group.

Rurality groups. Using the earlier rural label based on `log_ntl`, ADM1–year cells below the 40th percentile of the training distribution of log(ntl) radiance are labeled "rural." We designate loans in low-NTL regions as the protected rural group, and loans in high-NTL regions as the urban reference group. Here the intuition is that rural communities might be disadvantaged or less served by credit, so we examine if the model's errors or decisions disproportionately affect rural borrowers versus urban. We note that these group labels are binary and applied per loan based on the loan's mapped region.

4.3.2 Group fairness metrics. For each sensitive attribute, we compute a suite of group fairness metrics on the model predictions. We focus on group fairness criteria that compare classification outcomes between the protected vs. reference groups. For both gender and rurality, we compute:

(i) Demographic parity (DP) difference [7] the difference in the positive prediction rate between the two groups. This measures whether one group is selected (approved by the model) at a higher rate than the other, regardless of actual repayment. A value of 0 indicates parity (both groups are approved at equal rates). We compute DP difference as the protected group's approval rate minus the reference group's rate:

$$\text{DP difference} = \text{ApprovalRate}_{\text{protected}} - \text{ApprovalRate}_{\text{reference}} \tag{3}$$

(ii) Equal opportunity (EO) difference [14] the difference in true positive rates (TPR) between groups. Here we consider "approved=1" as the favorable outcome, so EO specifically checks if qualified borrowers in each group (those who indeed got approved in reality) have an equal chance of being correctly predicted as approved by the model. It is essentially the gap in recall (sensitivity) between protected and reference groups.

$$\text{EO difference} = \text{TPR}_{\text{protected}} - \text{TPR}_{\text{reference}} \tag{4}$$

(iii) Average odds difference (AOD) [14] the average of the differences in true positive rate and false positive rate (FPR) between the protected and reference groups. This metric corresponds to the Equalized Odds criterion (which requires both TPR and FPR to be equal across groups). AOD = 0 means the model has equal accuracy for both groups in terms of both misses and false alarms. We compute it as:

$$\text{AOD} = \frac{1}{2}\Big[(\text{TPR}_{\text{prot}} - \text{TPR}_{\text{ref}}) + (\text{FPR}_{\text{prot}} - \text{FPR}_{\text{ref}})\Big] \tag{5}$$

All gaps are reported as (protected – reference), so positive values imply potential disadvantage relative to the reference group. Like the other metrics, an ideal fair model would have DP, EO, and AOD differences all close to 0. In our context, we are especially attentive to negative gaps that would signal bias against females or

rural borrowers (e.g., a negative DP difference would mean a lower approval rate for female borrowers than for males).

For each trained model (Baseline and PlusNTL), we evaluate these metrics on the holdout test set, separately for gender and for rurality. The fairness metrics are computed at the default 0.5 decision threshold (unless otherwise noted), since metrics like TPR and FPR require binary decisions. We also report these values averaged over the multiple random runs to ensure robustness. By comparing the Baseline vs. PlusNTL models, we can assess how the inclusion of the NTL proxy affects fairness. For instance, does adding `log_ntl` worsen the DP or EO gaps for female borrowers? This forms the crux of our analysis on the fairness impact of NTL as a feature.

4.4 Robustness checks and threshold analysis

4.4.1 Model variants with fixed effects. To test robustness, we extend the logistic regression with fixed effects (FE) for ADM1 and loan sector. In addition to the main logistic models, we conduct robustness checks by augmenting the feature set with fixed effects for region and sector. The motivation is to see whether controlling for coarse categorical differences alters the influence of NTL. We evaluate four main variants: (i) *NoFE*, with no fixed effects (the main specification described above); (ii) *ADM1FE*, which adds ADM1 region dummies (region FEs). Including region fixed effects means the model effectively compares loans within the same region, so any performance gain from `log_ntl` in the scenario would imply that NTL is capturing within-region variation (since across-region variation is largely absorbed by the region dummies); (iii) *SectorFE*, which adds loan-sector dummies (sector FEs) and controls for differences between, e.g., agriculture and retail that may correlate with rurality; and (iv) *FullFE*, which includes both ADM1 and sector fixed effects.

In exploratory analyses, we also include microfinance partner identifiers where available. These FE specifications allow us to benchmark the incremental contribution of NTL against models that already absorb substantial regional or sectoral heterogeneity. These controlled models test the justification of using NTL: if NTL's predictive power drops to zero with region FEs, it suggests NTL was mainly acting as a proxy for region-specific effects that could alternatively be captured by a categorical variable.

We evaluate these variants with the same metrics (AUC, PR AUC, F1 and fairness gaps) to see if our main conclusions hold. For example, we compare the change in AUC when adding NTL in the no-FE vs. region-FE models, or examine if the DP gap between female and male changes when sector effects are accounted for. We summarize these comparisons in our results (e.g., presenting a bar plot of the ΔAUC and ΔDP gap when NTL is added, under each specification). This helps demonstrate the robustness of NTL's effect on accuracy and fairness. Notably, we find that even with region and sector controls, the inclusion of `log_ntl` still yields some improvement in predictive performance, albeit smaller, confirming that NTL provides a signal beyond just a static region identifier. At the same time, the fairness impact (e.g., any increase in disparity for rural borrowers) is observed consistently across variants, which suggests that the proxy's effect on model bias is not an artifact of omitted categorical variables.

4.4.2 Threshold scanning and fairness–accuracy trade-offs. Finally, we conduct threshold scanning, varying the classification cutoff from 0.1 to 0.9 in increments of 0.05. We examine how the choice of decision threshold affects the fairness outcomes, a crucial consideration in deployment. Using the PlusNTL model as a case study (similar patterns hold for Baseline), we generate prediction probability scores on the test set and then vary the classification threshold. For each threshold, we recompute the model's overall accuracy (or other performance metric) and the group fairness metrics (especially DP and EO differences) for both gender and rural groups.

This produces accuracy–fairness trade-off curves. At low thresholds, nearly all applicants are approved, shrinking demographic-parity gaps but reducing accuracy. At high thresholds, nearly all are rejected, again achieving superficial parity but sacrificing utility. In the intermediate range, we typically observe true trade-offs: as the threshold increases, the overall approval rate drops and can affect one group more than the other, causing fairness gaps to widen or flip. These curves help identify whether NTL shifts the feasible accuracy–fairness frontier for policy decisions.

Threshold curves are produced for both Peru and the Philippines, and for gender and rural groups separately. We plot fairness gap versus accuracy to visualize this trade-off (for example, see Figure 2 in our results, which shows a curve where one end is high accuracy but a large DP gap, and the other end is lower accuracy but a smaller gap). Such threshold scanning helps identify if there is a "sweet spot" where we can maintain decent accuracy while also reducing disparity. If not, it at least quantifies the cost of fairness in terms of forgone accuracy. In our methodology, we do not select the threshold based on this (we keep 0.5 as default for main comparisons), but this analysis serves to underscore that mitigating bias might be achievable by adjusting decision rules post-model. It provides a more nuanced view of the fairness implications of using the NTL proxy: if adding NTL shifts the trade-off curve (e.g., making it possible to attain higher accuracy at a given fairness level), that would be an important insight for policy. We include these threshold sweep results as a set of accuracy. Fairness trade-off curves for both sensitive attributes, and also report the maximum fairness-achievable accuracy (and vice versa) to inform stakeholders of the potential operational choices.

4.5 Justification of design choices and reproducibility

Throughout the above methodology, we have made certain design decisions with justification in mind. We applied weighted least squares, using the square root of the loan count as weights, to appropriately account for heteroskedasticity in the correlation analysis. (ensuring no single region with many loans dominates the trend). We opted for a binary rural label (low vs. high NTL) because group fairness metrics are most interpretable for clearly defined groups, using a threshold on a continuous proxy creates a salient "disadvantaged group" for analysis, and the 40th percentile cutoff balances representativeness with identifying truly low-light regions. The inclusion of fixed effects in robustness checks demonstrates our commitment to disentangle the proxy's effect from latent factors, a critical step when dealing with a proxy variable that could encode

structural differences like geography or industry. Lastly, the threshold scanning exercise is included to reflect deployment considerations: in practice, decision-makers might tune model thresholds to trade off equity and efficiency, so our methodology provides insight into how the NTL-enhanced model behaves under such adjustments.

All results are aggregated across seeds and stored in reproducible tables and figures (e.g., ΔAUC by FE variant, ΔDP by threshold). This pipeline ensures replicability and facilitates systematic comparison between Baseline and PlusNTL models across countries, tasks, and fairness dimensions. All these steps ensure that our findings on the predictive power and fairness impact of nighttime lights are rigorous, context-aware, and actionable for the finance + AI community. By clearly outlining this methodology, we enable readers to understand how we arrive at our results and how each analysis component ties into the overarching question of whether satellite-derived data can improve credit models without exacerbating bias.

5 Results and Discussion

5.1 Descriptive statistics and correlations

We examine the association between nighttime lights (NTL) and borrower gender composition. Scatterplots of female-borrower share versus log(ntl) at the ADM1–year level (Figure 1) show weak trends: in Peru, brighter regions have slightly lower female participation ($r \approx -0.1$ to -0.2), whereas in the Philippines the correlation is mildly positive ($r \approx +0.1$). Weighted least squares regressions with year fixed effects confirm that these associations are small and not statistically significant. Thus, NTL intensity is not a strong linear proxy for the gender composition of microfinance borrowers in either country, suggesting that including NTL is unlikely to encode direct gender disparities.

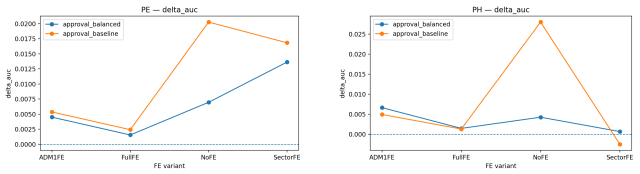
5.2 Predictive performance

We assess whether NTL improves predictive accuracy in loan-approval models (Table 3), comparing a *Baseline* model (excluding NTL) with a *PlusNTL* model (including log(ntl)). Across both Peru and the Philippines, performance gains are negligible: in Peru's baseline task, AUC increases from 0.877 to 0.880; in the balanced variant, from 0.864 to 0.866; and in the Philippines balanced task, AUC remains ~0.742. PR AUC and F1 are likewise stable, with differences below 0.01. Despite NTL's prominence in the development literature, its incremental value for predicting Kiva loan approval appears minimal.

5.3 Fairness evaluation

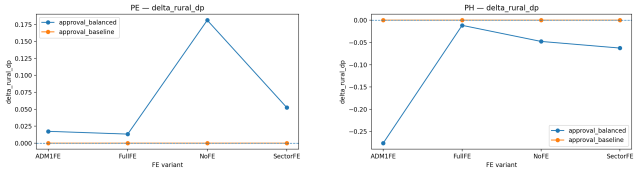
We then evaluate group fairness regarding borrower gender and rurality. For gender, we compare female-only loans to all other loans; for rurality, we classify ADM1–year cells below the 40th percentile of log(ntl) as rural.

The fairness metrics confirm that adding NTL has little effect. In Peru's balanced task, the gender demographic parity (DP) gap averages +0.469 for the Baseline model and +0.459 with NTL, while the rural DP gap shifts from −0.467 to −0.455. Equal opportunity (TPR) gaps and average odds differences show marginal changes (< 0.02). In the Philippines balanced task, the gender DP gap is



(a) Peru: ΔAUC across FE variants **(b) Philippines: ΔAUC across FE variants**

Figure 2: Change in AUC (PlusNTL − Baseline) across fixed-effect variants for each country. Each line is a task (approval_baseline / approval_balanced)



(a) Peru: ΔDP_{rural} across FE variants **(b) Philippines: ΔDP_{rural} across FE variants**

Figure 3: Rural DP gap change (PlusNTL − Baseline) across FE variants; Positive = greater disparity.

~0.55 and the rural DP gap ~0.01 in both models, with NTL again producing no difference.

Notably, fairness disparities are driven more by task design than by the inclusion of NTL. In baseline tasks (with the full imbalanced dataset), fairness gaps are nearly zero, reflecting the overwhelming prevalence of approved loans. In balanced tasks (with equalized classes), disparities emerge: female borrowers show higher approval prediction rates, while rural borrowers face systematically lower rates. However, these disparities are consistent across Baseline and PlusNTL models, implying that they stem from the data distribution and class balancing procedure rather than from the use of NTL.

5.4 Discussion

Overall, our findings indicate that nighttime lights (NTL) do not meaningfully enhance predictive performance or fairness in microfinance credit scoring once regional and socio-economic covariates are included. While NTL is widely regarded as a proxy for development, its contribution in this setting appears limited, highlighting the importance of domain-specific validation. Importantly, fairness gaps are shaped more by dataset design than by proxy choice: baseline tasks with extreme class imbalance suppress disparities, whereas balanced tasks reveal systematic differences by gender and rurality. These results underscore that fairness evaluations must be interpreted in the context of task construction and feature selection.

5.5 Threshold scanning and calibration

We perform threshold scans to explore how fairness metrics vary with the classifier threshold (from 0.1 to 0.9). Across all configurations, we observe that DP and EO gaps remain large in balanced tasks even when adjusting the threshold; there is no clear value

Table 3: Average predictive performance and demographic parity (DP) gaps under FullFE, averaged over two seeds. DP gaps are for female-only and rural (low-NTL) groups. Differences between Baseline and PlusNTL are small.

Country	Task	Model	AUC	PR AUC	F1	DP (Gender / Rural)
Peru	approval_baseline	Baseline	0.825	0.993	0.986	0.000 / 0.000
		PlusNTL	0.827	0.994	0.986	0.000 / 0.000
Peru	approval_balanced	Baseline	0.817	0.993	0.763	0.138 / -0.710
		PlusNTL	0.819	0.994	0.762	0.136 / -0.697
Philippines	approval_baseline	Baseline	0.720	0.992	0.991	0.000 / 0.000
		PlusNTL	0.721	0.993	0.991	0.000 / 0.000
Philippines	approval_balanced	Baseline	0.720	0.992	0.740	0.091 / 0.014
		PlusNTL	0.721	0.993	0.745	0.093 / 0.003

that simultaneously eliminates disparities and maintains reasonable accuracy. Baseline tasks show stable fairness across thresholds because the positive class is overwhelmingly dominant. These scans suggest that simple threshold calibration is unlikely to resolve fairness issues arising from data imbalance.

6 Implications and Future Work

Our findings have several implications for the responsible use of alternative data in inclusive credit scoring. First, NTL intensity, despite its popularity as a socioeconomic proxy, does not meaningfully improve credit scoring or fairness when region-level socioeconomic controls and fixed effects are present. Policymakers and practitioners should avoid relying on such proxies without rigorous validation. Second, fairness gaps in balanced tasks highlight structural inequities in loan outcomes. These disparities are not rectified by adding NTL; instead, attention should focus on data collection, sampling strategies, and modeling choices that account for gender and rural heterogeneity. Third, threshold adjustment alone is insufficient to mitigate group-level disparities; more sophisticated fairness interventions (e.g., re-weighing, constraint optimization) and domain-specific approaches are needed.

Future work will extend our analysis to other countries (e.g., Kenya, India) and tasks such as loan repayment prediction. We will also explore combining NTL with mobile phone and transaction data to assess fairness, and evaluate algorithmic fairness interventions in microfinance while engaging local stakeholders on practical implications.

References

[1] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732. doi:10.15779/Z38BG31 CC BY-SA 4.0.

[2] Anna Bruederle and Roland Hodler. 2018. Nighttime lights as a proxy for human development at the local level. *PLOS ONE* 13, 9 (2018), e0202231. doi:10.1371/journal.pone.0202231

[3] Xi Chen and William D. Nordhaus. 2011. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences* 108, 21 (2011), 8589–8594. doi:10.1073/pnas.1017031108

[4] Guanghua Chi, Han Fang, Sourav Chatterjee, and Joshua E. Blumenstock. 2022. Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences* 119, 3 (2022), e2113658119. doi:10.1073/pnas.2113658119

[5] Asli Demirgüç-Kunt, Leora Klapper, Dorothee Singer, Saniya Ansar, and Jake Kray. 2022. *The Global Findex Database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19.* Technical Report. World Bank, Washington, DC. https://openknowledge.worldbank.org/entities/publication/b74e1909-3ecf-5009-b51c-8527fc4eefeb

[6] Dave Donaldson and Adam Storeygard. 2016. The View from Above: Applications of Satellite Data in Economics. *Journal of Economic Perspectives* 30, 4 (2016), 171–198. doi:10.1257/jep.30.4.171

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12).* ACM, 214–226. doi:10.1145/2090236.2090255

[8] Christopher D. Elvidge, Kimberly Baugh, Mikhail Zhizhin, Feng Chi Hsu, and Tilottama Ghosh. 2021. VIIRS Night-time Lights. In *Remote Sensing of Night-time Light* (1st ed.). Routledge, London. Open access chapter.

[9] Ryan Engstrom, Jonathan Hersh, and David Newhouse. 2022. Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-being. *The World Bank Economic Review* 36, 2 (2022), 382–412. doi:10.1093/wber/lhab015 First published online July 31, 2021.

[10] Carter Faust, Anthony Dukes, and D. Daniel Sokol. 2022. Fintech and Financial Inclusion: A Review of the Empirical Literature. *Southern California Law Review Postscript* 95 (2022), PS135–PS151. Review article.

[11] John Gibson and Terry McKinley. 2021. Night Lights in Economics: Sources and Uses. *Journal of Economic Surveys* 34, 5 (2021), 955–980. doi:10.1111/joes.12387 First published online 2020.

[12] Hailing Gu, Chao Chen, Ying Lu, Yanli Chu, and Yuxiang Ma. 2019. Construction of Regional Economic Development Model Based on Remote Sensing Data. *IOP Conference Series: Earth and Environmental Science* 310, 5 (2019), 052060. doi:10.1088/1755-1315/310/5/052060

[13] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, Vol. 29. 3315–3323. https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf

[14] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS 2016).* https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

[15] J. Vernon Henderson, Adam Storeygard, and David N. Weil. 2012. Measuring Economic Growth from Outer Space. *American Economic Review* 102, 2 (2012), 994–1028. doi:10.1257/aer.102.2.994

[16] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794. doi:10.1126/science.aaf7894

[17] Souknilanh Keola, Magnus Andersson, and Ola Hall. 2015. Monitoring Economic Development from Space: Using Nighttime Light and Land Cover Data to Measure Economic Growth. *World Development* 66 (2015), 322–334. doi:10.1016/j.worlddev.2014.08.017

[18] Leora Klapper, Dorothe Singer, Laura Starita, and Alexandra Norris. 2025. *The Global Findex Database 2025: Connectivity and Financial Inclusion in the Digital Economy.* World Bank, Washington, DC. doi:10.1596/978-1-4648-2204-9 CC BY 3.0 IGO.

[19] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic Fairness. In *AEA Papers and Proceedings*, Vol. 108. American Economic Association, 22–27. doi:10.1257/pandp.20181018

[20] Toan Khang Trinh and Daiyang Zhang. 2024. Algorithmic Fairness in Financial Decision-Making: Detection and Mitigation of Bias in Credit Scoring Applications. *JACS* (2024). doi:10.69987/JACS.2024.40204