Unveiling the Impact of Key Determinants on Microloan Allocation Using Machine Learning

Sheraz Ali

Viewpoint School, USA

ABSTRACT

Microfinance initiatives have emerged as crucial tools in poverty alleviation, particularly in developing countries. These initiatives extend small loans to underserved populations and marginalized communities, as a means to improve their financial standing. This research leverages machine learning models to analyze a large microfinance initiative and find the key variables that drive micro-loan allocation. By conducting an in-depth feature analysis of a comprehensive microloan dataset, we aimed to pinpoint key variables that impact the allocation and size of microloans. Our study finds that the repayment period, lender networks, and the geographic location of borrowers significantly influence the loan allocation and size. The insights from our study can provide strategic guidance to microfinance programs in their loan allocation decisions. More specifically, lending decisions can be targeted more effectively to maximize impact of microfinance programs. In addition, this research also demonstrates the application of machine learning models in the realm of microfinance.

Introduction

Microfinance programs have proven to be an effective tool in the battle against poverty, providing much-needed financial services to underprivileged individuals, in both developing and developed countries (Mustafa et al., 2018). Even though these initiatives are increasing in number, the underlying factors that determine distribution of microcredit is an area that has seen inadequate research. With the growth of data science tools and increased data availability, machine learning offers possibilities to enhance the efficiency and impact of microfinance programs. Our research leverages machine learning techniques to determine which factors drive microloan decisions. This can provide insights into building a more precise, data-oriented methodology for distributing microloans. Through an in-depth examination of determining factors, we identify key variables that most impact the allocation of microfinance capital. Our results show that the repayment term, the number of lenders, and the geographic location of the borrower are the most critical determinant of loan allocation and size of loans. This analysis of a specific, large-scale microloan program provides vital guideposts for microfinance programs in general. The findings carry important implications for a wide array of stakeholders such as policymakers, microfinance establishments, and social entrepreneurs who are all working towards poverty reduction in developing countries.

Previous research indicates that microfinance can be a highly effective tool for poverty alleviation when used strategically (Banerjee et al., 2015). In their study, they found that microloans can lead to significant business growth and income generation, particularly when targeted towards individuals with specific entrepreneurial characteristics. While this study and others have underscored the value of microloan initiatives, less research has been conducted on improving the efficacy of those initiatives. This underscores the potential value of our machine learning approach in predicting successful loan allocation in helping to improve the efficiency of the microloan programs.

Dataset

The data utilized for this study was derived from the "Data Science for Good: Kiva Crowdfunding" dataset available on Kaggle (Kiva, 2018). This dataset consists of 671,206 unique microloan transactions distributed by Kiva, a global non-profit organization that facilitates crowdfunding to fund micro-entrepreneurs. The Kiva dataset comprises a large set of information on loan transactions. Its features include detailed information about borrower demographics, loan amounts, uses of the loans, countries of origin, sectors of investment, and more.

In the preprocessing stage, we performed thorough data cleaning and preparation to ensure the validity of our subsequent analysis. This involved loading the dataset, selecting relevant columns for our research ('lender_count', 'loan_amount', 'term_in_months', 'sector', 'region', 'activity', 'country', 'repayment_interval', and 'borrower_genders'), and handling missing data by dropping rows with incomplete information. We also transformed the categorical variables into numerical counterparts, a necessary step in building machine learning models. The dataset was then partitioned into features and then the target variable: 'loan_amount', which was the focus of our study. Lastly, we normalized the input features, effectively scaling the data from 0 to 1 to ensure that the differing scales of the variables did not influence the model performance. This preprocessing work allowed us to create a relatively clean dataset that was optimized for machine learning analysis and enabled us to accurately explore the key determinants of microloan allocation.

Methodology

Neural Network

Our study employed a layered architecture for the neural network that was designed to optimally leverage the large number of features in our dataset. The neural network begins with an input layer equipped with nodes equivalent to the total number of dataset features. This is then followed by two hidden layers, comprising of 64 and 32 nodes, respectively, as shown in Figure 1. These hidden layers adopt the Rectified Linear Unit activation function for operation as shown in Equation 1. The architecture concludes with a singular node in the output layer, mirroring our objective of predicting a singular continuous variable - the loan amount. The essence of the selected architecture lies in its potential to discern and model intricate, non-linear relationships between the input features and the target variable. The incorporation of multiple layers and nodes endows the neural network with an extensive capacity to learn nuanced patterns within the data. To discern the importance of each feature in our neural network, we employed Permutation Importance as our feature analysis method.

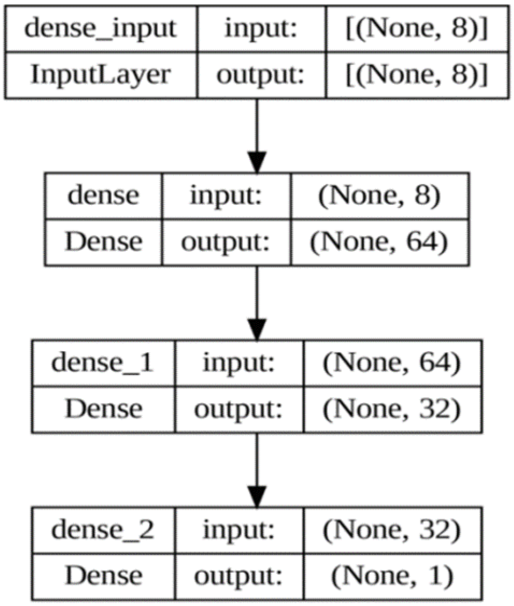
dense_input | input: | [(None, 8)]
InputLayer | output: | [(None, 8)]
dense | input: | (None, 8)
Dense | output: | (None, 64)
dense_1 | input: | (None, 64)
Dense | output: | (None, 32)
dense_2 | input: | (None, 32)
Dense | output: | (None, 1)

Figure 1. Neural Network Architecture

Equation 1. Single Neural Network Layer Equation

$$h = ReLU\ (Wx + b)$$

In Equation 1, h is the output activation vector, W is the weight matrix, x is the input vector, b is the bias vector, and $ReLU$ denotes the Rectified Linear Unit activation function utilized.

Note, that Equation 1 represents only a single layer. Our Neural Network utilized 2 hidden layers, as displayed in Figure 1.

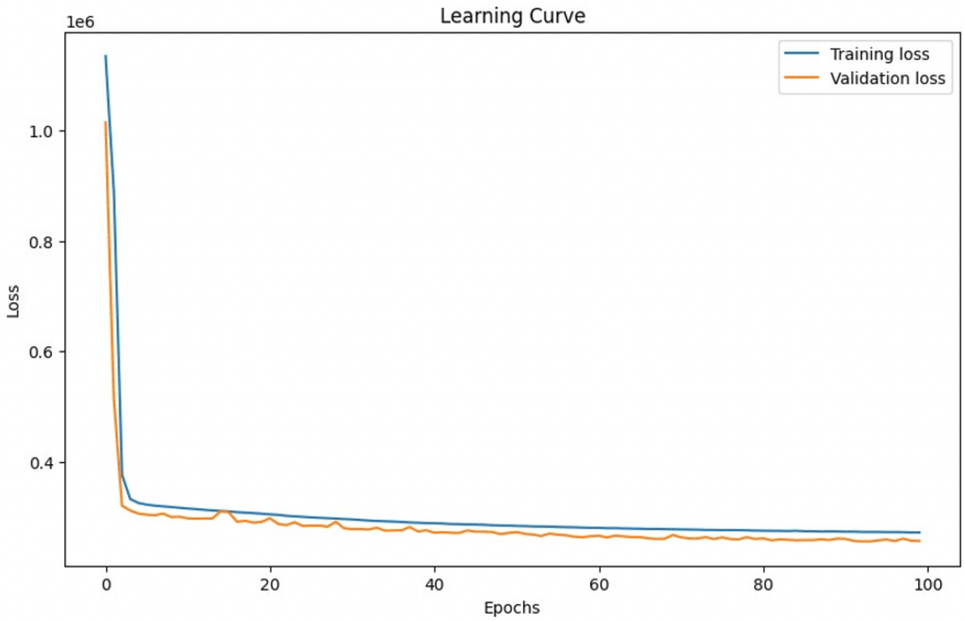

Figure 2. Neural Network Loss Curve

Figure 2 presents the learning curve of the neural network model used in this study. The curve plots both the training and validation losses across 100 epochs of training. As shown, both loss curves start at a relatively high point, indicative of the initial randomness of the model's parameters. What is noteworthy, however, is the rapid descent observed in the first few epochs, signifying the model's swift adaptation to the underlying patterns in the data. This provides a strong indication of the model's learning capability. As the epochs progress, the model starts to reach a point of diminishing returns where the loss flattens out and stabilizes. This plateau represents the point where the model has essentially 'learned' as much as it can from the training data. Further epochs do not contribute materially to improving the performance. Interestingly, the validation loss curve remains slightly below the training loss curve throughout the training process. We hypothesize that this phenomenon may be due to the way the model generalizes from the training data to unseen data in the validation set. While typically the validation loss is slightly higher due to the model's familiarity with the training data (and hence a lower training loss), a lower validation loss suggests that the model is not overfitting the training data and is able to generalize well on new, unseen data. We do not think this had any material impact on the model or the conclusions and the learning curve itself signifies a well-performing model. This underscores the viability of using this neural network architecture in our analysis.

Random Forest

The ensemble machine learning method, Random Forest, was also employed as part of our predictive modeling strategy. Known for its proficiency in handling high-dimensionality data and reducing overfitting, Random Forest creates an aggregation of decision trees, each constructed from a different subset of the training data (Brieman, 2001). The collective decision from these trees results in a robust prediction that harmonizes the individual trees' outputs. The configuration of our Random Forest model was fine-tuned with 100 estimators, a hyperparameter indicating the number of trees in the forest. This parameter was chosen to strike a balance between the model's accuracy and computational efficiency. An increasing number of trees generally improves model performance but at an increased computational cost. To further ensure consistency in the outcomes of the model, we leveraged the use of a random state seed (random_state=42). This fixed seed ensures the reproducibility of our results, generating the same sets of random numbers during each iteration.

After training the model, we utilized the built-in feature importance attribute. This attribute generates an array of scores, where each score corresponds to a feature's importance in the model.

Linear Regression

While simpler and more straightforward, the application of a Linear Regression model offered still valuable reference point due to its interpretability. In the Linear Regression model, each feature's weight (coefficient) signifies its importance in predicting the target variable. The higher the absolute value of a feature's weight, the greater its contribution to the final prediction. These weights were then analyzed to understand which factors most significantly influence the loan amount.

In this way, the linear regression model not only was useful in prediction but also served as a tool for understanding the underlying relationships within the data in a straightforward manner. However, it is important to remember that this simplicity comes at the cost of assuming a linear relationship between the features and the target variable, which may not always hold true in real-world scenarios.

Equation 2. Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + e$$

In Equation 2, y is the dependent variable, β_0 is the y-intercept, β_i represents the coefficient of the i^{th} predictor variable x_i, and e represents the error term.

XGBoost

The XGBoost algorithm is an implementation of gradient boosted decision trees, which have the capability to handle both linear and nonlinear relationships in the data, thereby making it an appropriate choice for our prediction task. The primary objective of the XGBoost model is to push the limits of computational resources for boosted tree algorithms, which it does by utilizing a more regularized model formalization to control overfitting (Chen & Guestrin, 2016). In our application of the XGBoost algorithm, we established an XGBoost regressor with hyperparameters set to optimize performance. The ‘n_estimators’ parameter, set at 100, denotes the number of gradient boosted trees we planned to construct. The ‘max_depth’ parameter, set to 5, is the maximum number of levels allowed in each decision tree. ‘Learning_rate’, set to 0.1, is a shrinkage parameter that prevents overfitting by shrinking the weights associated with features after each boosting step. Regularization parameters ‘reg_lambda’ and ‘reg_alpha’ were both utilized to avoid overfitting by adding penalties to the loss function during optimization.

Finally, we set the subsample to 0.8, which means that XGBoost would randomly sample 80% of the training data prior to growing trees and this would prevent overfitting. The model was then fit to the training data, and we made use of early stopping to avoid overfitting. This means that if the model's performance did not improve for 10 iterations (as specified by ‘early_stopping_rounds’=10), the training process would stop even if we were yet to reach the maximum number of estimators. The XGBoost model, through the application of gradient boosting mechanisms and careful tuning of hyperparameters, provided a robust and efficient method for predicting loan amounts in our dataset.

Equation 3. XGBoost Objective Function

$$\mathrm{obj}(\theta) = L(\theta) + \Omega(\theta) = \sum_i L\,(y_i, \hat{y}_i) + \sum_k \Omega\,(f_k)$$

In Equation 3 , L represents the training loss function, and Ω is the regularization term. The parameters of our model are represented by θ. y_i is the true label of the i^{th} instance, and $\hat{y}_i$ is the predicted label. f_k denotes the k^{th} base learner, and $L\,(y_i, \hat{y}_i)$ represents the loss from the i^{th} prediction.

The importance of a feature in XGBoost can be calculated by the average gain of the feature when it is used in trees. The idea is that before adding a new split on a feature to the branch, one can compute the loss function of the model. Then, after adding the split, one can calculate the loss again. The gain from the feature is the original loss minus the new loss after the split. A feature with a high importance score indicates that the model's prediction accuracy improved significantly when it was included in the model.

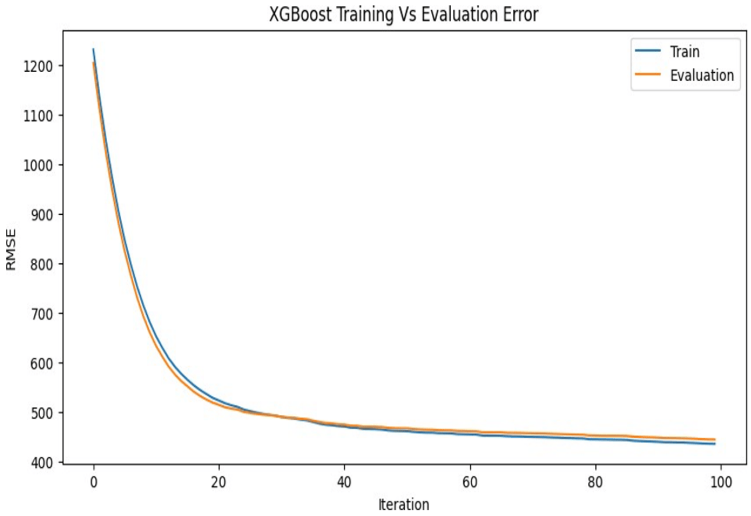

Figure 3. Learning Curve of XGBoost Model Showing Training and Evaluation RMSE Over Iterations

Figure 3 illustrates the learning curve for the XGBoost model utilized. The graph displays the training and evaluation of Root Mean Square Error (RMSE) across numerous iterations. At the start, both RMSE curves commence from a higher point, a typical result from the initial random state of the model. However, the swift drop visible in the early iterations underscores the model's capacity to rapidly comprehend and conform to the inherent patterns within the data. As we progressed through the iterations, we reach a point of diminishing returns, the point where the RMSE flattens and stabilizes. This horizontal trajectory suggests that the model has reached its learning saturation point from the training data, and further iterations do not bring meaningful improvements in the prediction performance. Interestingly, the evaluation RMSE remains closely tied with the training RMSE throughout the learning process. While one might expect the evaluation error to be slightly higher due to the model's familiarity with the training data, the observed close alignment implies that our model is proficient at avoiding overfitting to the training data and instead generalizes effectively to unseen data. The learning curve validates the strong performance of our model. The rapid decrease in RMSE during the early iterations, followed by a stable phase, indicates efficient learning and generalization.

Model Evaluation

Table 1. Test Set Performance Metrics of Machine Learning Models in Predicting Microloans Amount

Model	Mean Squared Error (MSE)	R^2
Random Forest	166298.82	0.849
XGBoost	197947.78	0.820
Neural Network	244584.84	0.779
Linear Regression	344486.35	0.689

In evaluating the predictive prowess of our four chosen models, we utilized two key performance metrics: Mean Squared Error (MSE) and R^2. Our analysis revealed that the Random Forest model outperformed the other models on both accounts. It demonstrated the lowest prediction error, as shown by its MSE value, and the highest proportion of variability explained in the loan amounts, as suggested by its R^2 value. This indicates a high degree of accuracy and a comprehensive understanding of the underlying patterns in our dataset. Next, the XGBoost model, although not as proficient as the Random Forest model, still showcased considerable predictive capabilities. It reported a higher prediction error and a slightly lower proportion of variability explained compared to the Random Forest model, but its performance was still quite commendable. In the third place was the Neural Network model, which reported a moderate performance. Although it was less accurate in its predictions and captured a lower proportion of variability compared to both the Random Forest and the XGBoost models, it still showed a reasonable ability to learn from the dataset. Lastly, the Linear Regression model displayed the least impressive performance among the four, with the highest prediction error and the lowest R^2 value. In conclusion, although all four models demonstrated their ability to predict loan amounts to varying extents, the Random Forest model emerged as the superior model, having the best balance of accuracy, and understanding of the data.

While discussing these results, it is crucial to note that the relatively high MSE values could be attributed to the large range and high variability in loan amounts in our dataset. High-value loans can significantly inflate the MSE if predicted inaccurately.

Results

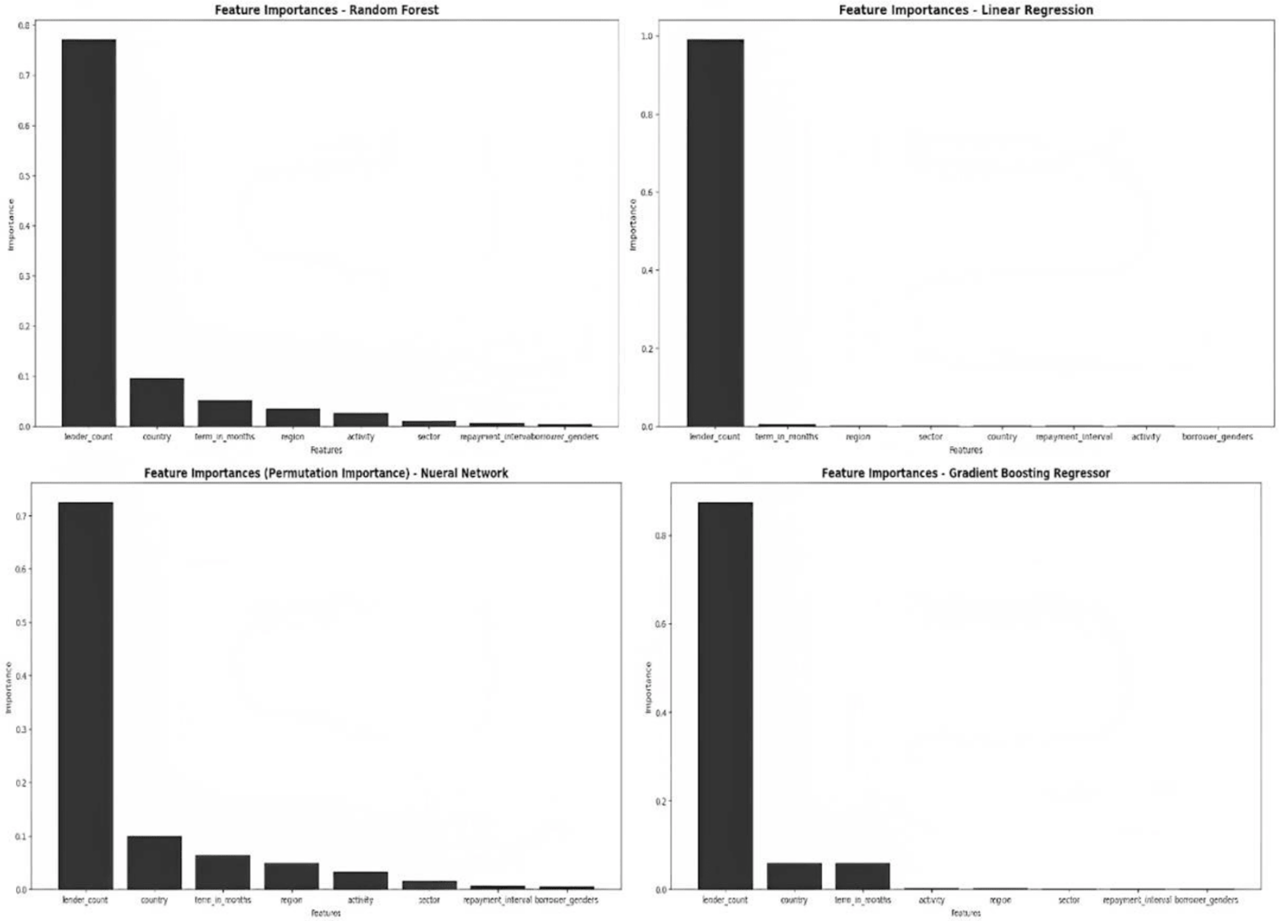

Figure 4. Comparison of Feature Importance Across Models: Random Forest, Linear Regression, Neural Network, and XGBoost

Figure 4 presents the normalized Feature Importance derived from the analysis of various machine learning models. The importance values range from 0 to 1, representing the relative predictive power of each feature in the microloan dataset.

As per our analysis of the different machine learning models, term_in_months, lender_count, and country emerged as the top three most significant features in predicting the loan amount across the board. The column charts and normalized feature importance effectively portray this information, providing a clear representation of the relationships between the various features and the target variable.

Term in Months

One of the prominent discoveries from our model's examination was the role played by the repayment duration, quantified in months, in determining the loan amount. This trend held true across all models used, thereby highlighting the significance of this factor within the landscape of microfinance. The association between the repayment term and the loan sum can be described as directly proportional. In other words, as the repayment term lengthens, the loan amount tends to increase. There are multiple reasons why this relationship is both plausible and expected with microloans. Microloans are frequently leveraged by borrowers to finance micro/small businesses or activities that generate income. It stands to reason that larger loans allow for more substantial investments in business operations, which could, over time, yield higher returns. This increased income takes time to realize upon the acquisition of the loan and accumulates only gradually as the business expands and becomes more profitable. Hence, borrowers who opt for larger loans require an extended period to generate the supplementary income needed for debt repayment. In addition, prolonging the repayment term for larger loans also provides benefits to lending institutions. By offering borrowers a longer duration to repay, lenders can mitigate some of the risks associated with defaults on larger loan amounts. Essentially it gives more time to repay. In addition, the longer repayment term also provides microfinance institutions with the ability to maintain a steady cash flow over an extended period, allowing them to manage their assets and match them with their liabilities. Therefore, the relationship between the repayment term and the loan amount is a logical one, grounded in the realities of the borrowers' repayment capacity and the risk management strategies of lending institutions.

Lender Count

The variable termed 'lender_count' was another key determinant in our predictive models, emphasizing the crucial role of the collective in making credit decisions. This was a function of our dataset, which utilized the Kiva model of microfinance. The Kiva model crowdsources small contributions from a multitude of online lenders, which can then collectively be offered as a loan. Consequently, the quantity of lenders participating in a loan, referred to as 'lender_count' in our dataset, directly influences the concluding loan amount. This suggests a strong correlation in our dataset.

In basic terms, an increase in the number of lenders translates to a higher capital base which in turn increases the capability to finance more extensive projects, resulting in a larger loan amount. This positive correlation between 'lender_count' and the loan sum is a function of the crowdfunding dataset used but does offer broader insight for microloan programs in general. The prominent role of 'lender_count' in forecasting loan amounts emphasizes the need for cultivating a large and engaged network of lenders within a particular microfinance program. It also suggests the collective appeal of certain loan recipients versus others, including factors that would suggest higher propensity to repay. This insight could be instrumental for microfinance institutions in strategizing their outreach and engagement efforts to expand their lender base, thereby enhancing their ability to fund larger loans.

Country

The borrower's geographic location, specifically the country, exerted a substantial effect on the loan amount.

Elements such as economic status, average income brackets, and living costs can dramatically differ between countries, factors which may sway the size of the loan. For instance, nations with steeper living costs may witness more substantial loan applications in comparison to countries with a lower cost of living. Moreover, the presence and conditions of alternative financing options can fluctuate based on the country, which might consequently influence the size of loans sought from microfinance institutions.

Conclusion

In conclusion, our extensive research and data analysis have brought forth intriguing insights about the factors influencing microloan allocation. Utilizing machine learning techniques, including linear regression, neural networks, random forests, and gradient boosting, we've explored the key determinants impacting the allocation of microfinance capital. Our research findings paint a comprehensive picture of microloan distribution dynamics, showcasing the intricate relationships between various factors and their impact on loan amounts. Particularly, three key determinants stood out as crucial influencers of microloan allocation — 'term_in_months', 'lender_count', and ‘country’. 'term_in_months' has been identified as a significant predictor, signifying that the duration of repayment directly influences the loan amount, while 'lender_count' reveals the importance of a diverse lender network in determining the loan's size. The influence of ‘country’ emphasizes the interplay between macroeconomic conditions and loan amount.

Moreover, the consistent agreement across our top-performing models regarding feature importance reinforces the validity of our findings and offers a robust foundation for future research and practical applications. The consistency of these findings also adds to the reliability of machine learning methods in predicting and analyzing microfinance trends, proving its capability in making these models more accessible and comprehensible to policymakers and stakeholders. Our research has indeed pulled back the curtain on the major influencing factors in microloan allocation. By identifying and examining these key determinants, our study provides an analytical basis for improving the design and implementation of microfinance initiatives. In practical terms, the insights generated by our research can guide microfinance institutions in tailoring their loan products and adjusting their strategies.

Ultimately, by leveraging machine learning techniques to optimize loan allocation, our study contributes towards a more effective, targeted, and equitable approach to poverty alleviation. It also highlights the areas where access to microfinance capital is lacking, pinpointing where efforts need to be concentrated.

In future research, we encourage an expansion of this work to explore the influence of other factors that might be significant in a microfinance context. We believe that this machine learning approach can provide a springboard for further innovative and influential research in microfinance, driving new strategies and policies that aim to enhance financial inclusivity and poverty alleviation in developing nations.

Through this comprehensive and rigorous analysis, our study underlines the critical role machine learning techniques can play in enhancing microfinance initiatives. By translating complex interactions between features into understandable insights, machine learning techniques offer an invaluable tool to analyze, predict, and strategize microloan allocation, ultimately driving more impactful poverty alleviation efforts in developing countries.

Limitations

Despite the intriguing findings of our study, it is necessary to acknowledge certain limitations inherent in our research approach. This not only ensures the transparent and comprehensive interpretation of our results but also serves as an important guide for further studies.

One notable limitation stems from the nature of the dataset we used—the “Data Science for Good:

Kiva Crowdfunding” dataset. While this dataset offers valuable insights into a wide range of microloan transactions, it is representative of a specific subset of the microfinance landscape. Consequently, our results might not be wholly applicable to other microfinance contexts with dissimilar borrower demographics, geographic distributions, or loan characteristics. Future research would benefit from the incorporation of diverse datasets, covering a wider range of microfinance situations to increase the external validity of our findings.

Also, it is important to consider the noticeable disparity in gender representation in the dataset, with females being the majority. This imbalance could have potentially introduced biases in our models, as they might overly tune to patterns associated with female borrowers, thus underplaying the impact of gender. This is not a reflection of the relative importance of gender but rather it is a limitation of the data. As such, we should interpret results with an understanding of these potential biases and be cautious when extrapolating these findings to a broader population.

Furthermore, the data preprocessing stage in our research required handling missing data, which was accomplished by eliminating rows with absent values. There was a total of 56,808 rows eliminated in our preprocessing stage. This method, although common and convenient, might have inadvertently introduced a degree of selection bias. By removing these instances, there's a potential risk that the retained observations could be systematically different from the eliminated ones. While this approach was essential to ensure the compatibility and cleanliness of our data for model building, it is crucial to keep this potential bias in mind when interpreting the results.

Acknowledgments

Special thanks to my mentor Mr. Odysseas Drosis, Cornell University alumnus and current Ph.D. candidate at École Polytechnique Fédérale de Lausanne, for his valuable guidance and mentorship throughout this research.

References

Banerjee, A., Karlan, D., & Zinman, J. (2015). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics*, *7*(1), 1-21. https://doi.org/10.1257/app.20140287

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/a:1010933404324

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* ArXiv. https://doi.org/10.48550/arXiv.1603.02754

Kiva. (2018). *Data Science for Good: Kiva Crowdfunding*. Kaggle. https://www.kaggle.com/datasets/kiva/data-science-for-good-kiva-crowdfunding

Mustafa, F., Khursheed, A., & Fatima, M. (2018). Impact of global financial crunch on financially innovative microfinance institutions in south asia. *Financial Innovation*, *4*(1). https://doi.org/10.1186/s40854-018-0099-8