Developing High-Performing & Ethical Credit Scoring Models in Microfinance

Sept 1, 2017

Word Count: 5,355

Abstract

With major international banks entering the microfinance (MF) sector and applying state-of-the-art credit scoring algorithms to assess customer creditworthiness, smaller Microfinance Institutions (MFIs) are facing competition in operational efficiency as they continue to rely on the Loan Officer (LO) interaction with borrowers. The following study provides practical advice to MFIs on the options and performance of various credit scoring algorithms explored in academia in the last decade. The study highlights the recent tendency away from easy-to-interpret, parametric models such as logistic regressions, to less interpretable, non-parametric models such as artificial neural networks. The primary cause for the shift is the improved performance of the non-parametric models, which tend to deliver higher out-of-sample accuracy. Nonetheless, this paper calls for MF practitioners to exercise caution when considering adoption of black-box models. Since most MFIs strive to remain financially sustainable while also pursuing a social mission of enabling bottom-of-the-pyramid communities to bootstrap themselves out of poverty, they should be careful about the possible discrimination of certain customer groups due to their race, sex, religion, age etc. The paper provides recommendations for how MFIs that decide to adopt black-box models can improve their operational efficiency and competitiveness and still act in an ethical manner towards all customer groups.

Introduction: Microfinance Institutions and Historical Developments of the Microfinance Sector

The coining of the term “microfinance” is often attributed to Mohamed Yunus who founded the Grameen Bank in Bangladesh in 1975 (Battilana & Dorado, 2010). According to Yunus, the main objective of Microfinance Institutions (MFIs) is ‘providing the poor – whom he describes as “natural entrepreneurs” – with working capital with which they can realize their entrepreneurial potential’ (Battilana & Dorado, 2010, p.1422). Thus, as a service, microfinance (MF) is tailored to the needs and realities of poor entrepreneurs operating small- or medium-sized businesses and requiring small-scale transactions of either credit loans or savings deposits (Khandker, 2005). Many MFIs strive for a double bottom line, concurrently pursuing financial sustainability and delivery of social impact through access to capital (Van Gool et al., 2009).

Over the last few decades, the financial success and fast growth in the MF sector garnered interest among major international banks. Their entry into the sector put great competitive pressure on the more socially-oriented MFIs, which now find themselves having to increase operational efficiencies, minimize transaction costs, and better control their risk exposure to survive the competition (Kiruthika & Dilsha, 2015).

Van Gool et al. (2009) further explain that strong competition, customer over-indebtedness, and economic instability in developing countries encouraged MFIs to consider adopting more efficient approaches to evaluating customer repayment capacity. ‘One of these techniques was credit scoring, which analyzes historical client data and derives a model which links repayment behavior with characteristics of the loan, lender, and borrower’. (p.1)

What is Credit Scoring and Can It Work for Microfinance?

C*redit scoring* is one of the more critical components in traditional banks and is referred to the process of 'collecting, analyzing and classifying different credit elements and variables to assess the credit decisions' (Abdou & Pointon, 2011, p.2). Hence, the aim of credit scoring is to evaluate the borrower's propensity for default.

Van Gool et al. (2009) refer to several studies that explored the applicability of credit scoring models in MF. Specifically, Capon (1982), Schreiner (2003) and Freytag (2008) warned MFIs of the vulnerabilities of applying credit scoring models in the MF realm, suggesting that incorporating variables such as customer unwillingness to pay and greater tendency for natural disasters in developing countries is hardly possible. Additionally, Bumacov et al. (2014) explain that MFIs 'face the asymmetry of information in a more intense way than traditional financial institutions because micro-borrowers, who are generally poor and illiterate, provide inadequate quality and quantity of documentation to prove their creditworthiness' (p.402).

Van Gool et al. (2009) also cites Dennis (1995), Schreiner (2003) and Kulkosky (1996) who list several advantages of the use of credit scoring in MF, including: lower default levels, opportunities to market services to different segments, and increased LO efficiency. Furthermore, Kinda and Achonu (2012) note that credit scoring removes the subjective judgement and different rules of thumb espoused by LOs, ensuring that all borrowers are evaluated with the same set of rules and that loan decisions are made more consistently. Schreiner (2000) summarizes the general sentiment on credit scoring in MF sector in the following manner:

> 'Credit scoring for microfinance can work. It is not as powerful as scoring for credit card or mortgage lenders in wealthy countries, and it will not replace the judgements of loan officers or loan groups based on informal, qualitative knowledge, but scoring does have some power to predict risk (and thus to cut costs) even after the group or loan officer makes its best judgement' (p.116).

Current Credit Scoring Approaches in Microfinance Sector

To guard against borrower default, Van Gool et al. (2009) explain the two main methodologies outlined by Basel II framework, which in turn was established by the Basel Committee of Banking Supervision in 2006. These methodologies are:

1. ***Standardized Rating Approach***, where institutions assess the risk profile of a potential borrower using external credit assessment infrastructure such as credit bureaus.

2. ***Non-Standardized Rating Approach***, where institutions use an internally developed risk evaluation methodology, including credit scoring techniques, to identify drivers of borrower default.

Van Gool et al (2009) also refer to the work of Thomas (2000) who categorizes the latter methodology into three sub groups:

1. ***Judgemental.*** This approach, while time-consuming and expensive, still dominates across most MFIs and relies on experience and opinion of the LO who must assess the borrower's character, evaluate

her financial standing, identify whether she has loans with other institutions, validate the viability of her business proposal, assess her collateral, and monitor the repayment (Sarker, 2013).

2. ***Statistical.*** This approach primarily relies on parametric statistical models such as logistic regression to separate the future-defaulters from borrowers who repay their loans entirely and on time. From here onwards, these models are referred to as *parametric*.

3. ***Non-statistical, Non-Judgemental.*** This approach predominantly involves the use of non-parametric, black-box models such as artificial neural networks. From here onwards, these models are referred to as *non-parametric*.

Bumacov et al. (2014) conducted a survey of 405 MFIs and concluded that there is a positive relationship (correlation, not causation) between use of non-Judgemental credit scoring approaches (parametric and non-parametric) and client outreach. While this may simply imply that larger MFIs have more resources to build scoring models, it could also suggest that MFIs that use non-Judgemental scoring methods are better able to reach greater numbers of poor entrepreneurs and thus improve the MFIs' social performance.

In the subsequent sections, the paper will provide an overview of (1) the variable / feature selection process, (2) the assessment of model performance, (3) overview and performance of different parametric and non-parametric models, (4) ethical considerations of using non-parametric models, and (5) potential solutions for building non-discriminatory credit scoring models.

Variable / Feature Selection Process

In Van Gool et al. (2009) study, as well as suggested in Schreiner (2000), model features often belong to the following categories: borrower (age, net earnings of business, net earnings of household, job experience), loan (purpose, amount, duration), and lender characteristics (branch, loan officer). Meanwhile, the Consultative Group to Assist the Poor (2015) identified a number of start-up companies operating in MF space and using digital data such as mobile phone usage and social media interactions for scoring. However, it is not necessary that all available features / variables should be used in the model.

Kiruthika and Dilsha (2015) state that 'variable selection process helps to decrease the risk of over fitting the model by reducing the number of independent variables in the model' (p.123). The authors also refer to the work of Chen and Hughes (2004) to explain the concept of *parsimony*, which supposes the following: if the action of removing certain variables from the model does not impact the ability of the remaining independent variables to explain the dependent variable, then the simpler model without the excluded variables is preferred. Furthermore, Van Gool et al. (2009) suggest that including too many variables into the model may jeopardize the borrower experience by making the application forms too long.

Van Gool et al. (2009) refer to Hand and Henley (1997), Verstraeten and Van den Poel (2005) as well as Baesens, Van Gestel, and Thomas (2009) to provide three methods for feature / variable selection:

1. ***Expert Knowledge.*** Microfinance practitioners who have worked with borrowers and/or their data have strong comprehension of the borrowers' economic activities and enablers of success may often be able to identify important features that should be included in the model.

2. **Statistical Procedures.** These include forward and backward selection methods, as well as the combined forward and backward step-wise method. Kiruthika and Dilsha (2015) add that statistics such as Akaike Information Criterion or chi-square are often employed to identify which variables to retain in the models. The authors also suggest that for artificial neural networks, it is possible to remove variables that produce low connection weights, then retrain the model, and observe impact on performance.

3. **Area Under the Receiver Operating Characteristics Curve (AUROC), often referred to as AUC**. AUC graphically presents the model's ability to correctly classify 'bad borrowers' as defaulters (true positives) while minimizing incorrectly classifying 'good borrowers' as defaulters (false positives).

Assessing Performance of Credit Scoring Models

It is important to note that feature selection and model performance are intertwined; specifically, features are dropped or included with the view to improve model performance. Referring to the work of Van Gestel et al. (2006), Van Gool et al. (2009) suggest that model performance is assessed using the following criteria:

1. ***Stability***. If a model performs similarly when assessed in- and out-of-sample, it is said to be stable. Out-of-sample testing and validation is usually done by retaining a test and/or validation dataset. Alternatively, cross-validation techniques can be used.

2. ***Readability***. If a model is relatively easy to interpret (ex. years of experience decreases the risk of default by x %), it is said to be readable.

3. ***Discriminatory Power*.** Also frequently referred to as model performance, discriminatory power is the accuracy with which a model discriminates among 'good' and 'bad' borrowers. Lantz (2015) describes several measures used to assess the performance of credit scoring models:

   i. *Specificity*: number of correctly classified good borrowers / total number of good borrowers
   ii. *Sensitivity*: number of correctly classified bad borrowers / total number of bad borrowers
   iii. *Precision*: number of correctly classified bad borrowers / total number of borrowers
   iv. *False Positive Rate*: number of good borrowers that were incorrectly classified as bad / total number of good borrowers
   v. *Accuracy*: number of all correctly classified borrowers / total number of borrowers
   vi. *Error rate*: number of all incorrectly classified borrowers / total number of borrowers
   vii. *AUC*: as described above, AUC is a graph that plots the values of False Positive Rate (1-Specificity) on the x-axis and Sensitivity on the y-axis. The closer the curve passes the top left corner, the better the overall discriminatory power of the model.

Performance of Parametric & Non-Parametric Scoring Models in MF

The following section provides an overview and performance of multiple parametric and non-parametric models developed in academia in the past decade. Due to word count limitations, this paper does not explain the mathematical basis behind each model.

Sathe and Desai (2006)

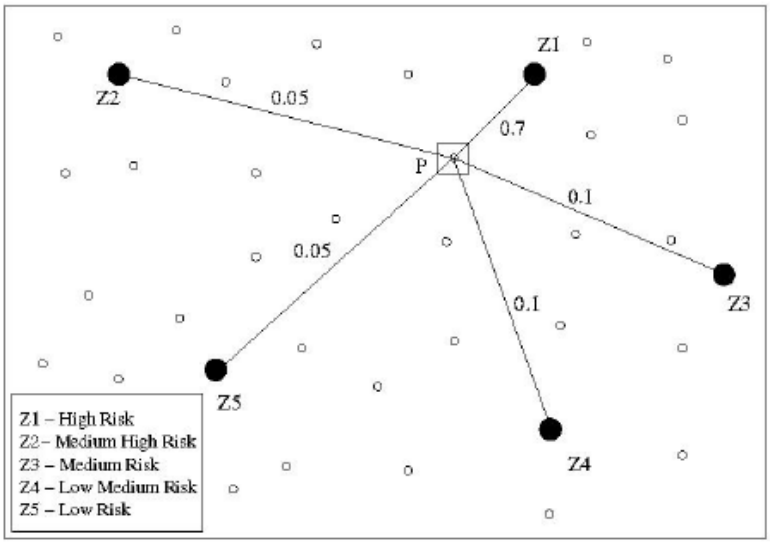
Sathe and Desai (2006) studied a women's village savings group in Maharashtra, India, prefacing the description of their model with the following observations:

1. Unlike in formal MFIs where LOs make a decision about granting a loan, in the village banking model, 10-20 women that participate in the savings groups decide whether to grant a loan to another.
2. Besides relying on several objective variables (age, number of dependents), the group members often base the decision on subjective aspects. These variables include moral character, attitude, monthly income, and social status of the borrower.

Considering the nature of these groups, an algorithm must be able to decipher patterns from subjective and sparse data. It must be quick to train the model and quick to produce a decision. For these reasons, the authors selected the Fuzzy Symbolic C-Means algorithm, which uses symbolic, as opposed to numeric features to categorize a borrower into a cluster.

The authors explain that 'unlike hard clustering, where each instance has to exclusively belong to a cluster, in fuzzy clustering each instance belongs to a cluster with a degree of association or confidence' (p.235-236). The authors obtain five cluster centers that they label as: High Risk, Medium High Risk, Medium Risk, Medium Low Risk, and Low Risk. Each new instance is processed to produce the level of a fuzzy association with each of the five clusters. Geometrically, this association may look in the following manner, implying that the instance being assessed is most likely belonging to the High Risk category:

Figure 1: *Geometrical interpretation of fuzzy clustering (Sathe and Desai, 2006, p.236)*



From above, the *readability* of output is relatively simple. In considering *Stability* & *Discriminatory Power,* it must be noted that the authors built the algorithm using 19 instances only and validated it on 9 instances. In the validation set, 1 out of 9 instances was classified incorrectly, giving 11.1% Error Rate. However, both training and validation sets are too small to conclude that these results would scale and produce consistent outcomes.

Diallo (2006)

Diallo (2006) studied the repayment behavior in a Malian MFI serving and proposed two credit scoring models – Logistic Regression (LR) and Multivariate Discriminant Analysis (MDA). He used the data on 269 loans, of which 91 were classified as defaults ('default' was defined as 30 or more days of delay in repayment).

Both LR & MDA showed similar *Discriminatory Power*, with in-sample Accuracy levels of 74.7% (AUC results were not reported). However, the author noted that the LR provides better *Readability*. Based on the translation of the article from French, it did not appear that the author tested the models out-of-sample, hence little can be said about their *Stability*.

When changing the probability of default threshold from 0.5 to 0.4 (if probability of default is above the threshold, the borrower is classified as a defaulter), the overall accuracy level for both models dropped by ~3%, but sensitivity (ability to correctly identify bad loans) increased from 25.3% to 28.6%.

Diallo's (2006) exercise also provided a valuable lesson in understanding how various variables contribute to the probability of default; this was obtainable from the magnitude and sign of the coefficients in LR. However, the overall *Discriminatory Power* would be too poor to apply this model into operations.

Dinh and Kleimeier (2007)

Dinh and Kleimeier (2007) proposed an LR credit scoring model for a retail bank in Vietnam. While this study did not focus on MF borrowers, the authors acknowledged that the Vietnam market is a developing one and its banks often engage in MF activities. The data comprised 56,000 small business loans, credit card loans, consumer loans, and mortgages.

Although the authors did not provide in-sample performance metrics, the out-of-sample accuracy boasted 97.98%. AUC results were not reported, but sensitivity was 50.25%. While these results seem impressive, the reader must keep in mind that they were obtained from a bank that engages in some MF activities, but those are not core to the overall operations.

Van Gool et al. (2009)

Van Gool et al. (2009) used the data of 6,722 loans from a Bosnian MFI, applying 70% / 30% split for train and validation sets. The authors developed two variations of LR models:

1. ***Binary LR with dummy coding*** for categorical variables. The advantage of this approach is that it allows representation of non-linear behavior. The disadvantages are: a) the resulting model may overfit the data, b) with too many categories in a given variable, the degrees of freedom are quickly lost, and 3) in some cases, near-singularity problems may occur.

2. ***Binary LR with weight of evidence coding*** for categorical variables. Here, categories are grouped based on how often they are observed in defaulted / non-defaulted loans. This approach reduces the pressure on degrees of freedom, but it may lead to overfitting as the coding is based on the dependent variable.

The authors concluded that the LR with dummy coding provided greater *Stability* and *Readability*. The *Discriminatory Power* of the dummy coding model – out-of-sample AUC of 0.707 – was also superior to that of the weight of evidence coding model (out-of-sample AUC of 0.681)

The authors also reported that such performance is not competitive in comparison to the predictive power of models used in traditional banks, such as seen above in Dinh and Kleimeier (2007). In an MFI setting, this model would at best complement the LOs, but not replace them.

Kinda and Achonu (2012)

Kinda and Achonu (2012) conducted a similar study to Diallo (2006), where an LR (selected for *Readability*) model was built using 30 loans from an MFI in Senegal. Of those, 15 were considered bad based on at least a 5-day delay in payment rule. However, the authors' aim was to better understand the relationships between independent variables such as age, gender, guarantee, etc and the probability of default than to build and apply a credit scoring model into action. Though the *Discriminatory Power* appears high with accuracy of 90%, and the AUC stands at an impressive 0.956, the readers should remember that no testing was done out-of-sample (hence no evidence of model *Stability*) and that the sample of 30 is unlikely to be representative.

Blanco et al. (2012)

Blanco et al. (2012) conducted a comprehensive exercise in comparing the performance of parametric and non-parametric models on a dataset from a Peruvian MFI. Reviewing the models that had been developed for MFIs at that time, the authors noted that nearly all had been parametric, predominantly using LR & Linear Discriminant Analysis (LDA) techniques. They also observed that 'the strict assumptions (linearity, normality and independence among predictor variables) of these traditional statistical models, together with the pre-existing functional form relating response variables to predictor variables, limit their application in the real world' (p. 357).

The authors referred to a long list of studies which explored the use of non-parametric techniques such as decision trees, k-nearest neighbours, and neural networks, suggesting that 'due to the non-linear and non-parametric adaptive-learning properties' (p. 357), Artificial Neural Networks (ANNs) demonstrated better predictive performance in credit scoring tasks. Recognizing the challenges of using ANNs (it takes long to train, and the resulting solution is a black-box), the authors aimed to develop a solution that is suitable for MFIs using a multilayer perceptron neural network (MLP) and compare its performance against LDA, LR, and Quadratic Discriminant Analysis (QDA).

The cleaned dataset contained personal, financial, economic and business-specific information for 5,451 borrowers. The authors qualified a default as any repayment that was late by 15 days or more. Variables for the LDA, LR, and QDA were selected using stepwise approach. The best performing parametric model in terms of AUC was LR (0.932). However, when looking at misclassification costs (the authors weighed the error of classifying a bad borrower as a good borrower five times more severely than classifying a good borrower as a bad borrower), QDA performed better. The authors developed 14 versions of MLP, the best of which produced better AUC (0.954) than any of the parametric models and simultaneously lower misclassification rate. The model was described as:

> 'It is a three-layer perceptron, with 20 input nodes, 3 hidden nodes and one output node. The training has been performed with R, using a BFGS quasi-Newton learning rule, and both the size of the hidden layer and the regularization parameter are selected by 10-fold cross-validation, the value of this latter parameter being 0.2.' (p. 362)

The authors concluded that despite the black-box (*low Readability*) nature of the ANNs, the MFIs should consider using them because of their improved *Discriminatory Power* and potential to prevent large losses.

Kiruthika and Dilsha (2015)

Kiruthika and Dilsha (2015) used the data of 520 loans disbursed within savings groups in two villages in Kerala, India. Of those loans, 48 were in default (3 weeks of delay in payment). The authors compared the performance of two models – LR and ANN – using 60% of data for training and 40% for validation. ANN approach was explained as following:

> 'The ANN model built in this study applied multilayer perceptron architecture with one hidden layer and back propagation learning algorithm. The weight decay was set at 0.01, while learning rate and momentum was set at 0.1 and 0.01' (p. 131).

Both LR & ANN had 2 variations: with variable selection and without. The resulting four models were compared using Error Rate and AUC (in- and out-of-sample).

In terms of *Stability*, both ANN models were more stable than LR models. Looking at *Discriminatory Power*, both ANN models showed a lower Error Rate than the LR in the validation set. The AUC was also slightly higher for ANN models. Despite ANN being a black-box with limited *Readability*, the authors concluded that in this study the ANN model was more effective than LR. Finally, within the two ANN models, the model with variable selection was better performing (8.64% Error Rate & 0.868 AUC out-of-sample).

What is gleaned in the above studies is in line with Abdou and Pointon (2011) who stipulated that 'there is no overall best statistical technique/method used for building credit scoring models, and the best technique for all data sets does not exist yet' (p. 24). The authors referred to Hand & Henley (1997) who in turn suggested that the best model is contingent on the specifics of the situation, the data available, the variables, and ultimately their ability to discriminate one class of borrowers from the other.

Having said that, it does appear that in the case of MF where the borrower data is sparse, the quality of that data is questionable, and the macroeconomic conditions are uncertain, the non-parametric, black-box models have been emerging as having greater *Discriminatory Power* and *Stability*, which albeit comes at the cost of lower *Readability*.

While some would suggest that this improvement is a reason to celebrate and jump on board, it is important to note that none of the studies above considered the discriminatory potential of credit scoring models based on sex, age, race, religion, ethnic group etc. The following section will explore the ethical considerations of using non-parametric credit scoring models as such considerations have not yet been explored in the MF literature.

Ethical Considerations of using Non-Parametric Credit Scoring Models

Maes and Reed (2012) expressed concern over the declining reputation of MFIs, as the perception of the sector is shifting from having both social and financial orientation to the financial orientation only. As credit scoring promises cost reduction and greater operational efficiencies, it is critical that the MF sector does not lose sight of the original mission of helping marginalized communities bootstrap themselves out of poverty.

Meanwhile, Dinh and Kleimeier (2007) explained that developing countries, for example Vietnam, unlike their developed counterparts, often do not have strict regulations on gender or religion in credit scoring models. Hence, the authors included gender into their model and determined that women were less likely to default. Their finding is in part corroborated by Schreiner (2004) who also suggested that women are generally less risky, though specifying that when accounting for other variables, the risk gap between genders disappears.

While the seemingly positive discrimination towards women does not offend many, and in fact seems to support the overall mission of MF, it does suggest that credit scoring is capable of both positive and negative discrimination. Schreiner (2003) summarises this phenomenon: 'Statistical scoring does discriminate: it assumes that each applicant is another instance of the same old thing, not a unique individual who might differ from other apparently similar cases in the database' (p.28). He then juxtaposes statistically rooted discrimination against subjective discrimination, suggesting that the latter may result from prejudices of an LO and could potentially be worse.

Schreiner (2003) recognizes that it may be unfair to assess one person based on experience with another. However, he also acknowledges that the alternative to discriminating – approving all potential borrowers – is not viable. Consequently, he suggests that fair discrimination rooted in statistics, unlike subjective discrimination rooted in prejudice, is informative in linking true risk to borrower characteristics and has the potential to 'decrease prejudice and correct mistaken inferences' (p.28).

Accepting the argument above, it should be noted that parametric models such as LR explicitly reveal the impact of a feature's value on the probability of default through the sign and the magnitude of the coefficient. For example, a negative coefficient on the Female dummy indicates a decreased probability of default by women as compared to men. However, the non-parametric models with learning capabilities often do not provide such revelations. Mittelstadt et al. (2016) propose that such algorithms, depending on their design, may not necessarily be ethically neutral. Referring to the works of Barocas (2014), Leese (2014), Macnish (2012), Mittelstadt et al. (2016) warn that 'discriminatory analytics can contribute to self-fulfilling prophecies and stigmatisation in targeted groups, undermining their autonomy and participation in society' (p.9).

A legitimate question arises: how can MFIs ensure that their credit scoring models are not inhibiting their ability to pursue a social mission of supporting and enabling marginalized groups? The following section will provide several recommendations.

Recommendations for Building Non-Discriminatory Credit Scoring Models

Mittelstadt et al. (2016) recognize that algorithms are not meant to be perfect. Consumer Federation of America National Credit Reporting Association (2002) further state that these algorithms 'are predictive

of repayment behavior for large populations' and 'work well on average' (p.5). They are replacing human effort that would have been unaffordable to commission at the necessary scale. However, Mittelstadt et al. (2016) acknowledge that organizations must be mindful of the impact of the errors and discriminatory potential of their algorithms.

Some organizations deliberately exclude features such as gender, race, age, and marital status from the datasets. Unfortunately, there is no guarantee that other proxy features (ex. income level, postal code, home ownership) would not 'give away' the marginalized nature of the person's identity (Mittelstadt et al., 2016; Romei & Ruggieri, 2014).

Romei and Ruggieri (2014) aggregate four strategies for guarding against discriminatory tendencies when designing and using a classifier (in this case, a credit scoring model):

1. ***Controlled Distortion of Data During Data Pre-Processing Stage.*** For the main techniques, Romei and Ruggieri (2014) refer to the work of Kamiran and Calders (2012) who in turn proposed the following options:

    a. *Massage the data:* Use two different thresholds, ex. probability of default of 0.6 & 0.5, to categorize marginalized and non-marginalized groups, respectively, as potential defaulters. This allows one to change several class labels of marginalized individuals from 'defaulter' to 'non-defaulter' (in the paper, the authors use 'C+' and 'C-'). This group of reclassified borrowers is then known as the *promotion group*. The reverse logic should then be applied to the non-marginalized borrowers, who would be reclassified from 'non-defaulter' to 'defaulter' and be referred to as the *demotion group*. The number of reclassifications should be such that the probabilities of default are the same for both marginalized and non-marginalized groups. Then a model can be trained on a discrimination-free set. Kamiran and Calders (2012) do acknowledge that this option is rather invasive.

    b. *Reweigh the records:* This option is significantly less intrusive. Each record, depending on whether it belongs to a marginalized or non-marginalized group, and whether it is classified as a 'defaulter' or a 'non-defaulter', is assigned to one of the four quadrants and is given a weight. The weight is determined using conditional probabilities. For example, assume that a record is for a defaulter from a marginalized group (Quadrant 1). Now, assume that 50% of the borrowers in the dataset are marginalized, while 60% of all borrowers are defaulters. Therefore, the expected probability of a marginalized borrower given a default should be 0.5 * 0.6 = 30%. But assume that in reality this number is 40%, meaning that marginalized borrowers are over-represented among defaulters. The weight for all marginalized defaulters should then be calculated using: 0.5 * 0.6 / 0.4 = 0.75. Similar calculations would be completed for the other three quadrants, thus rebalancing the dataset. The weights can then be incorporated into the scoring model.

    c. *Sampling:* The authors recognize that not all models can incorporate the use of weights. Hence, as an alternative, sampling techniques can be used. First, the dataset must be segmented into the following:

        i. Marginalized Borrowers who are Defaulters (MD)
        ii. Marginalized Borrowers who are Non-Defaulters (MN)
        iii. Non-Marginalized Borrowers who are Defaulters (ND)

iv. Non-Marginalized Borrowers who are Non-Defaulters (NN)

If discrimination exists, the ratio of MN:MD will be lower than the ratio of NN:ND. To align these ratios and thus remove discrimination, a decreased count of MD observations and/or an increased count of ND observations is required (alternatively, there could an increased count of MN or a decreased count of NN). With the necessary count of observations allowing for equal ratio between MN:MD and NN:ND being established, sampling with replacement can be used within each group until the required group size is reached. With this approach, NDs & MNs will be over-sampled whereas MDs & NNs will be under-sampled. The resulting dataset can be used to train the model.

2. ***Integration of Anti-Discrimination Criteria into the Model During the Modeling Stage:*** Here, Romei and Ruggieri (2014) refer to the following works:

   a. Calders and Verwer (2010) provide recommendations for Naive Bayes approaches. Since Naive Bayes is not frequently used in credit scoring as it assumes feature independence, these recommendations will not be covered in detail. However, one of the relevant approaches by Calders and Verwer (2010) is to train two separate models, one for a marginalized group and one for a non-marginalized group, thus removing the correlation between the probability of default and the marginalized nature of some borrowers.

   b. Kamiran et al. (2010) provide a solution when constructing tree models; generally, the tree is split into leaves to achieve best possible accuracy using information gain criterion at every split. However, the authors suggest using not only accuracy but also the degree of discrimination (of marginalized groups) as a second criterion for the split.

3. ***Threshold & Label Management During the Post-Processing Stage:*** Once again, Romei and Ruggieri (2014) refer to the following studies:

   a. Kamiran et al. (2010) suggest that an already existing tree can be seen as a space divided into a number of non-overlapping regions. Thus, when a new borrower record is provided for the classification, the record falls into one of the regions and is labeled with the majority class in that region (ex. 'defaulter'). The authors suggest that in some cases, it is possible to 'relabel the leaves of the decision tree in such a way that the discrimination decreases while trading in as little accuracy as possible' (p.871).

   b. Pedreschi et al. (2009) suggest modifying the thresholds in the model, making it slightly easier for marginalized borrowers to be classified as 'good'. To ensure that both marginalized and non-marginalized communities are equally likely to be classified as bad, a threshold modification can also be applied to the non-marginalized group.

4. ***Modification of Predictions During the Post-Modelling Stage:*** The final strategy does not require modifications of datasets or change in model development approaches. Instead, Romei and Ruggieri (2014) suggest correction of actual predictions that result from the model to keep equal proportions of defaulters in both marginalized and non-marginalized groups. Specifically, the authors refer to the work of Kamiran et al. (2012) who suggest amending predictions that are near the decision boundary.

Conclusion

The above paper discussed the evolution of parametric and non-parametric credit scoring models used in the MF sector with the aim to remain competitive, scale operations, and expand outreach to communities that would otherwise lack access to capital. Questions of viability were posed by multiple researchers, inquiring whether credit scoring could work as well in the context of MF as it does in the banks operating in wealthy countries. Some suggested that credit scoring in MF will not be sufficient enough on its own, but it could complement the judgement of the LOs. However, it has become evident that with the entry of major international banks into the MF space, the development and application of credit scoring models will continue to grow within the sector, both by those banks and by the MFIs that will inevitably need to compete with them.

Assessment of the current literature suggested a shift from parametric, easy-to-interpret models such as LR towards the non-parametric, uninterpretable models such as ANNs. This shift is primarily driven by the improved *Stability* and *Discriminatory Power* of those non-parametric models, as measured by out-of-sample Accuracy, Error Rate, Sensitivity, and AUC outcomes.

As most MFIs pursue both financial sustainability and social impact on marginalized communities, they will need to be careful about the ways in which they design their credit scoring algorithms due to their potential do discriminate against already marginalized groups based on applicant gender, ethnicity, age, marital status etc. The paper provides a number of recommendations on how to avoid incorporating discriminatory bias into the training data, the models built using that data, and the resulting predictions, allowing the MFIs to design algorithms that are both effective and ethical.

References

Abdou, H. & Pointon, J. (2011) Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance & Management*, 18 (2-3), 59-88. Available from https://pdfs.semanticscholar.org/791c/f0b410e2dcd8d9be88452959c52b3b066580.pdf [Access Aug 30th, 2017]

Battilana, J., & Dorado, S. (2010) Building sustainable hybrid organizations: The case of commercial microfinance organizations. *Academy of Management,* 53(6), 1419-1440.

Bumacov, V., Ashta, A. and Singh, P. (2014) The use of credit scoring in microfinance institutions and their outreach. *Strategic Change*, 23(7-8), 401-413.

Blanco, A., Pino-Mejías, R., Lara, J. & Rayo, S. (2012) Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications*, 40(1), 356-364.

Calders, T., & Verwer, S. (2010) Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2), 277-292.

Consultative Group to Assist the Poor. (2015) *The Potential of Digital Data: How Far Can It Advance Financial Inclusion?.* Available: from https://www.cgap.org/sites/default/files/Focus-Note-The-Potential-of-Digital-Data-Jan-2015_1.pdf [Accessed 25th Aug, 2017].

Consumer Federation of America & National Credit Reporting Association. (2002) *Credit score accuracy and implications for consumers*. Available from: http://www.consumerfed.org/pdfs/121702CFA_NCRA_Credit_Score_Report_Final.pdf [Accessed 29th Aug, 2017]

Diallo, B. (2006) Un modèle de "crédit scoring" pour une institution de micro-finance Africaine: le cas de Nyesigiso au Mali. Available from: https://hal.archives-ouvertes.fr/halshs-00069163/document [Accessed 26th Aug, 2017].

Dinh, T.H.T. and Kleimeier, S. (2007) A credit scoring model for Vietnam's retail banking market. *International Review of Financial Analysis*, 16(5), 471-495.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010) Discrimination aware decision tree learning. In: *Data Mining (ICDM), 2010 IEEE 10th International Conference*. pp. 869-874. IEEE. Available from: https://pure.tue.nl/ws/files/3216677/692173.pdf [Accessed 30th Aug, 2017]

Kamiran, F., & Calders, T. (2012) Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1-33.

Khandker, S. R. (2005) Microfinance and poverty: Evidence using panel data from Bangladesh. *The World Bank Economic Review*. 19(2), 263-286.

Kinda, O., & Achonu, A. (2012) Building a Credit Scoring Model for the Savings and Credit Mutual of the Potou Zone. *Consilience: The Journal of Sustainable Development*, (7), 17-32.

Kiruthika & Dilsha, M. (2015) A neural network approach for microfinance credit scoring. *Journal of Statistics & Management Systems*, 18(1-2), 121-138.

Lantz, B. (2015) *Machine Learning with R*. 2nd Edition. Birmingham, UK, Packt Publishing.

Maes, J.P. & Reed, L.R. (2012) State of the Microcredit Summit Campaign Report 2012. *Microcredit Summit Campaign, Washington, DC*. Available from: http://www.microcreditsummit.org/uploads/resource/document/socr-2011-english_41396.pdf [Accessed 26th Aug, 2017]

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1-21. Available from: doi: 10.1177/2053951716679679 [Accessed 7th Aug, 2017].

Romei, A., & Ruggieri, S. (2014) A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5), 582-638. doi:10.1017/S0269888913000039

Sarker, D. (2013) Pressure on loan officers in microfinance institutions: An ethical perspective. *Journal of Economics and Sustainable Development*, 4(12), 84-88.

Sathe S. K. & Desai U. B. (2006) Cell Phone Based Microcredit Risk Assessment using Fuzzy Clustering. In: *Proceedings of the 2006 International Conference on Information and Communication Technologies and Development*, *May 2006*, *Berkeley, CA*. IEEE. pp 233-242. Available from: doi 10.1109/ICTD.2006.301843 [Accessed 3rd Jul, 2017].

Schreiner, M. (2000) Credit scoring for microfinance: Can it work?. *Journal of Microfinance/ESR Review*, 2(2), 105-118.

Schreiner, M. (2003) Scoring: The next breakthrough in microcredit? CGAP Occasional Paper No. 7. Available from: https://www.cgap.org/sites/default/files/CGAP-Occasional-Paper-Scoring-The-Next-Breakthrough-in-Microcredit-Jan-2003.pdf [Accessed 13th Aug, 2017].

Schreiner, M. (2004) Scoring Arrears at a Microlender in Bolivia. *Center for Social Development*. Available from: http://www.microfinance.com/English/Papers/Bolivia_Scoring_Arrears.pdf [Accessed 22nd Jul, 2017].

Van Gool, J., Baesens, B., Sercu, P. & Verbeke, W. (2009) An analysis of the applicability of credit scoring for microfinance. Available from: https://www.researchgate.net/profile/Wouter_Verbeke/publication/255560672_An_Analysis_of_the_Applicability_of_Credit_Scoring_for_Micronance/links/00463534f802adfdd6000000.pdf [Accessed 22nd Jul, 2017].