# Financing the Un-financed

**Supervisor:**  Dr Prof. Carey Pridgeon

**Student Name:** Ravi Srikanth Reddy

**SID:** 9607655

We acknowledge that all the work is done on our own: **Yes**

# Abstract

Financing or loans are essential for any community, today Banks are desperate to provide loans to individuals having good tracking financial/credit history. But many people are not part of this system, they do not have trackable financial history and thus Banks are reluctant to offer loans to them. And these people are often exploited by local untrustworthy lenders. The idea is to predict borrowers’ repayment capability by using other easily available dataset like telecommunication and transactional information. The idea is to predict loan pay-ability using various statistical and machine learning methods

# Research Question, Problem Statement or Topic for In- vestigation

We are here dealing with section of people who are underprivileged, and there are not enough data about them in the system. I can clearly say that though they are in the system, but when it comes to delivering them something related to financial or social, they are virtually out of it. Our intention is to bring financial inclusion for these section by providing loan to those who have the potential to use the loan for entrepreneurial endeavors and predicting their loan repayment capability. We will design a model which will enable us to predict the loan repayment capability and identify those who deserves the loan. The problem that we have faced is data collection. There is no reliable and proper data set available for set of people who belong to underprivileged section. Clearly the data is not consistent, number of missing data points are huge. We have to sit and pre-process the data which is a herculean task. Obviously, since the size of dataset is large, the CPU would find it difficult to handle or it may take lot of time to process the data. For the purpose of the data processing, the important factor is to identify the feature vectors which has more impact on the outcome. We will be using statistical techniques like adjusted R-Square, hypothesis testing and t-statistic and z-score to arrive at the set of features.

We intend to use regularizing techniques like ridge regression and Lasso regression. The tuning parameter *λ* serves to control the relative impacts of the two terms on the regression coefficient estimates.

We will use models like GAM (Generalized additive model) which provide a general framework for extending a standard linear model by allowing non-linear functions of each of the variables, while maintaining additivity. GAMs can be applied to both quantitative and qualitative response.

Finally we will use tree based methods like bagging, random forest and boosting. The decision tress suffers from high variance. Bootstrap aggregation or bagging is general purpose procedure for reducing the variance of a statistical learning method. Given a set of n independent observations Z1, Z2, ..., Zn, each with variance σ2, the variance of the mean Z¯ of the observations is given by σ2 . The averaging a set of observations reduces variance. We will use PCA(Principal Component Analysis) for deriving a low-dimensional set of features from a large set of variables. PCA looks to find a low-dimensional representation of the observations that explain a good fraction of the variance. We are going to use unsupervised learning techniques like k-means clustering to find the homogeneous subgroups among the observations. Boosting works by growing trees : sequentially each tree is grown using information from previously grown trees. Boosting does not involve bootstrap sampling; instead each tree is fit on the modified version of the original data set. Like bagging, boosting involves combining a large number of decision trees, fˆ1, ...,fˆB .The idea behind the procedure is unlike fitting a single large decision tree to the data, which amounts to fitting the data hard and potentially over-fitting, the boosting approach instead learns slowly. Given the current model, we fit a decision tree to the residuals from the model. That is , we fit a tree using the current residuals, rather than the outcome Y, as the response. We then add this new decision tree into the fitted function in order to update the residuals. Each of these trees can be rather small, with just a few terminal nodes, determined by the parameter d in the algorithm. By fitting small trees to the residuals, we slowly improve fˆ in areas where it does not perform well. The shrinkage parameter slows the process down even further, allowing more and different shaped trees to attack the residuals. In general, statistical learning approaches that learn slowly tend to perform well. Boosting unlike in bagging, the construction of each tree depends strongly on the trees that have already been grown.

*n*

# Intended user or group of users and their requirements

The model can be used by any lending agency, or any government or Non-Governmental Organization which is working in the area of the financial inclusion. The banks can identify the people using these models and help in establishing a business by giving loan and by knowing the loan repayment capability. The government can use the model to identify the potential person and help him/her in achieving their potential. The non-governmental organization can also identify such person and assist them in getting loan from the concerned lending agency. The model has huge potential to transform the lives of the underprivileged. The model also has the huge potential in saving the billions by helping the banks to identify the potential customer and their repayment capability.

# Systems requirements, project deliverable and final project outcome

We are required to first collect the dataset, define each data points and its purpose. We require to preprocess the data and fine tune the dataset. Finally, we would like to build a model, train the model on testing set, validate the dataset with using cross validation. We need to find hyper-parameter and use the hyper-parameter on testing data. It would not be possible for the simple CPU to perform the computation. We will be using google colab which provides GPU, to run the model and train the model. The project will finally help in building the model which will help the lending agency in bringing the underprivileged under financial umbrella. The final outcome of the project is it will give me insight into banking lending activities, the problems faced by bank in lending to underprivileged section and finally convincing the lending agency in using the model to identify people from underprivileged and lend them money in establishing entrepreneurial activities.

# Primary Research Plan

The first objective is to collect data. The data collection will happen through various resources. By physically visiting and asking questions to people. By finding data online through world bank and other agency. Once the data collection is complete, it is important to process the data and make it usable. Processing the data is most important part of any machine learning model. Prepossessing involves transforming and cleaning the data. It involves activities like filling the missing values by mean, median, mode or deleting the data points if it is not necessary. It also involves transforming the data from qualitative to quantitative. This step is important because the machine learning models cannot understand categorical data. Finally we need to normalize the data. The different feature vectors may have data ranging in different scale. So we need to normalize the data and bring it to common scale so that the training the model would give consistent and expected output.

# Initial/Mini Literature Review

Bank loaning and credit quality: the case of India [1]This paper examinations how non- performing credits (NPLs) of Indian banks carry on through the cycle. We discover that a one-percentage point increment in advance development is related with an increment in NPLs over add up to progresses (NPL proportion) of 4.3 per cent within the long run with the reaction being higher amid expansionary stages. Moreover, NPL proportions of banks are found to be delicate to the intrigued rate environment and the generally development of the economy. EFFECT OF MICROFINANCE ON POVERTY REDUCTION: A CRITICAL SCRUTINY OF THEORETICAL LITERATURE [2] The arrangement of micro-credit to the destitute more particularly to the youth and ladies within the rustic destitute is an important calculate on destitution decrease through strengthening. Most inquire about thinks about done, have shown some improvements in work creation as well as increment in pay levels through self-employment. There has been signs of independence on the side of the ladies as well as the youth, on things concerned with family use on the side of ladies and instruction consumption on the side of the youth being able to raise earnings from self employment. The analyst has too found that most analysts have utilized at slightest two to four models the popular ones being: Grameen solidarity gather show, focusing on ladies groups’ demonstrate, normal repayment schedule show and town managing an account show. The analyst has managed with different discourses, the major ones being: that micro-credit has been able to elevate at-least 55% of its clients out of destitution, they are being provided by clean drinking Doing business with the poor: the rules and impact of the micro-finance institutions[3].Of the full worldwide populace, at slightest 14.5% are living on less than 1.25 a day, 34% of the females within the slightest created nations are incapable to total their essential instruction, and a few 805 million are accepted to be nourishment uncertain. To bring these numbers into understanding with the Thousand years Advancement Objectives, there are at slightest a dozen of diverse programs working around the world. Micro-credit, being one of those programs, is considered predominant to the rest for being the as it were participatory approach and for being common sufficient to cater for a number of approach mediations. Micro-credit or credit to the destitute is given beneath two exceptionally distinctive components; the welfarist component and the institutionalist instrument. Each of these components has its advocates, its faultfinders. The current paper observationally assesses the two approaches in a precise way.

# Bibliography

1. *Pallavi Chavan and Leonardo Gambacorta Bank lending and loan quality: the case of India*.
2. *Nyarondia Samson Mecha EFFECT OF MICROFINANCE ON POVERTY REDUCTION: A CRITICAL SCRUTINY OF THEORETICAL LITERATURE .*
3. *Wajid Khan, Sun Shaorong, Ikram Ullah Doing business with the poor: the rules and impact of the microfinance institutions*
4. *T.J Hastie and R.J. Tibshirani Generalized Additive Models(Chapman & Hall/CRC Monographs on Statistics and Applied Probability)*
5. *Robert E .Schapire The Boosting Approach to Machine Learning: An Overview*
6. *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani An Introduction to Statistical Learning with Application in R.*
7. *Christopher M. Bishop Pattern Recognition and Machine Learning*
8. *Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman Mining of Massive Datasets*
9. *Trevor Hastie, Robert Tibshirani, Jerome Friedman The Elements of Statistical Learning*