WORKING PAPER NO: 601

Predicting Educational Loan Defaults: Application of Artificial Intelligence Models

Jayadev M
Professor
Finance and Accounting
Indian Institute of Management Bangalore
Bannerghatta Road, Bangalore – 5600 76
jayadevm@iimb.ac.in

Neel M Shah
BITS Pilani Hyderabad
hemun95@gmail.com

Ravi Vadlamani
Professor
IDRBT, Castle Hills Road #1,
Masab Tank, Hyderabad - 500 057
vravi@idrbt.ac.in

December 20, 2019

Year of Publication – December 2019

Financial Support from the Digital Innovation Lab (DIL) of IIMB is gratefully acknowledged by the first author. The earlier version of the paper" Educational Loan Defaults: Application of Linear and Non-Linear Quantitative Models", co-authored by first author with Hemaang Kotta was presented at 4th International Conference on Business Analytics and Intelligence 2016 held at Indian Institute of Science (IISc) Bangalore during December 19-21, 2016.

Predicting Educational Loan Defaults: Application of Artificial Intelligence Models

Abstract

We show that Educational loans is a case for application of artificial intelligence models to predict potential defaulters with a reasonable accuracy. Ensemble models tend to perform better than simple artificial techniques and statistical models and that the performance can be improved significantly by model stacking. We argue here that a stacked model created using a few sparsely correlated base models is likely to be the best model for predicting Educational loan defaults given that the interaction between diverse features would create non-linearities that are impossible to model using a single model, there is little a priori knowledge of the distribution of educational loan defaults and the relationships between various factors that govern the distribution. It is evident that collateral-free loans have a considerably higher rate of default with moral hazard problem as compared to the loans with collateral. Students qualifying from well rated educational institutions are prone to strategic default or wilful default. Considering the impact of macroeconomic conditions greatly improve the classification accuracies.

Keywords: Credit Risk, Educational Loans, Statistical Techniques, Artificial Intelligence Techniques

Predicting Educational Loan Defaults: Application of Artificial Intelligence Models

1. Introduction

Khandani, Kim and Lo (2010) and Malhotra and Malhotra (2002) provide the empirical evidence on improved predictive accuracy of defaults of consumer loans by application of Machine learning and Artificial Intelligence models over the conventional statistical models. Educational loans are also similar to consumer loans in ticket size but have several unique features such as zero or low collateral, longer repayment periods with serious information asymmetry leading to very weak risk assessment by banks. It is extremely difficult to assess the repayment ability of the student loan applicant at the time of granting the loan due to limited information on recognized courses, college accreditation, employment opportunities and entry-level salaries. With the presence of uncertainty involved at multiple levels: becoming a successful student, securing a decent-paying job and the one who repays loans regularly and this to a large extent depends on his future earnings (Barr and Crawford, 2005). Another aspect of educational loans is the presence of behavioural aspects of a student and the possibility of them changing over the course of the duration of the degree which are almost impossible to map using a set of variables. These might include aspects such as motivation levels and commitment towards repaying the loan. Besides, collateral-free educational loans are prone to moral hazard. Several borrowers are known to be wilful defaulters and it is the behavioural factors which govern this difference between the ability to repay and the willingness to repay. Thus, allowing the algorithm to learn the traits of defaulters based on historic data and using the results to test it on a new observation is more likely to capture these traits. The rising educational loan default rates demands the implementation of early warning systems to detect the risks with high degree of confidence.

A consensus is the ability of humans to judge the worthiness of a loan is poor, beyond the predictability of simple statistical techniques (Glorfeld, 1996). Artificial intelligence helps to classify the borrowers as potential defaulters and non-defaulters based on a list of available variables. The variables under consideration could be monetary, such as the parental income or the loan limit; demographic and social background,; academic qualifications , the university rankings or even macroeconomic, such as the unemployment rates prevalent in the market or the GDP growth rates. Due to the diversity of these variables, lack of clarity on how they impact default rates, presence of non-linearity and an absence of credit scoring metrics based on

historical data, it is not possible to manually classify each loan applicant. Also, the task of discovering meaningful relationships or patterns from data is difficult for humans (Handzic, 2001).

There have however been fewer studies on the use of artificial intelligence techniques for problems where the outcome is significantly affected due to the presence of behavioural traits. (Dasgupta et al. 1994) used machine learning to categorise investors based upon their willingness to take financial risks. Robin and Bierlaire (2012) using logit regression, modelled the stock market behaviour and decisions of asset managers, short term and long term investors, firms and amateur investors each having a different objective, different attitudes, reactions and access to a different degree of information. Badea (2014) used artificial neural networks to capture the behavioural determinants of propensity to save; using a dataset containing variables such as age, occupation, nationality and education to classify individuals likely to have bank deposits, essentially making it a problem on consumer behaviour.

Besides, in absence of any legally bound procedure or generally accepted decision making rules for sanctioning or rejecting loans, this process includes a considerable amount of subjectivity since the decision is left to loan officers. Humans are prone to bias and the decision making process can be affected due to the presence of emotional or psychological conditions (Handzic, Tjandrawibawa and Yeo, 2003). Use of artificial intelligence makes the process less prone to biases. If an application for loan is turned down, the bank is expected to give a reason for credit denial. Machine Learning can help in this regard by suggesting, for instance how much of this credit denial was due to low parental income, how much due to the student attending a less exclusive institution or was it due to demographics or macroeconomic factors forcing the banks to be strict while giving a loan.

This study considers both, the idiosyncratic borrower specific aspects (parental income, university, geographical area, etc.) as well as the systematic (external) factors such as growth rate, inflation and unemployment levels. Both these factors contribute towards default. This study shows that complex ensemble ML Models are better at classifying defaults for such a dataset given that our ensemble models outperform all the non-ensemble models. In this sense, therefore, we are able to address the model suitability of the dataset. Although this is expected given the non-linearities and complex behavioural patterns associated with educational loans, the problem of model suitability has never been addressed in the past. Most of the previous

studies in this domain were restricted to applying a single model to the dataset and observing the results.

This paper investigates classification accuracy of multiple AI techniques to screen the educational loan borrowers. The study uses pooled data set of educational loan borrowers from four different banks and applies Artificial Intelligence techniques. The objective of this paper is to evaluate the effectiveness of AI techniques to identify probable default loans. Secondly, we compare the performance of AI techniques with the statistical logit models. Our analysis indicates that the AI techniques performance in evaluating potential loan defaulters is statistically superior to the logistic regression. AI techniques do not require any restrictive assumptions like statistical models and offers flexibility to loan officers to adopt new rules for loan evaluation.

This paper is divided in to seven sections. Section two reviews the existing literature on consumer loan defaults specifically educational loans. Section three presents an academic review of various Artificial Intelligence models, Section four describes the data used in this paper, Section five presents the methodology and results are discussed in Section six. Section seven presents conclusions of the study and summarises the paper.

2. Educational Loan Defaults: Review of Related Literature

According to Choy and Li (2006); Lochner and Monge-Naranjo (2004) the probability of default is directly proportional to the debt burden. With an increase in debt burden, the monthly payments become quite high and managing debt becomes difficult. It is almost intuitive to believe that students from low-income families are likely to incur more debt (Herr and Burt 2005); Steiner and Teszler . Knapp and Seaks (1992) and Woo (2002) concluded that lower the family income, higher the probability that the student would default. Baum and O'Malley (2003) said that a family income acts like a safety net for borrowing students.

Students' ability to repay depends more on their income after studies, than their accumulated debt. (Constantine Kapsalis, 2006). Therefore, the choice of field of study, the sector in which one attains employment and general macroeconomic trends play a crucial role in determining their ability to repay.

Chou, Looney and Watson (2017) showed that low-income students attending better institutions do better than high income students attending institutions with lesser return on

investment. Besides, it is possible that intake in higher return on investment schools is skewed in favour of rich students. Research shows that students who attend less than two-year courses have higher default rates than their peers attending four year courses. (Podgursky et al. 2002; Woo 2002).

Christman (2000) and Woo (2002) observed that students having a higher score at high school or better standardized test scores had lesser default rates. Schwartz and Finnie (2002) conducted educational loan studies in Canada and concludes that the academic branch of the student significantly affects his/her future earnings. Herr and Burt (2005), Steiner and Teszler (2005) too indicate the possibility of the choice of study affecting future salaries. However, if some fields do offer a higher salary, then in general that particular branch is likely to have a higher demand and highly meritorious students might prefer that course. Besides, it is possible that a student secures employment in a sector entirely different from their original branch. This makes it difficult to take branch into consideration. Consequently, Lochner and Monge-Naranjo (2004) show that the effects of course choice vanished after accounting for other factors such as the total debt or post-college earnings. Student's repayment pattern is also affected by his/her academic performance at graduate school. Better college grades raise the likelihood of the student securing a better job.

Researchers like Christman (2000) incorporated the borrower's age into their studies. Woo (2002); Steiner and Teszler (2005) observed the existence of a positive correlation between age and probability of default. A possible explanation for this came from Herr and Burt (2005) suggesting that older students are have a greater degree of financial responsibilities which negatively affects their ability to repay. The probability of default rises with an increase in the amount owed as the older debtors owe more than their younger counterparts (Choy and Li, 2006).

Gladieux and Perna (2005) observed that students coming from ethnic minority groups and socially backward or neglected strata of the society face difficulties in repaying their borrowings. While the exact reasons for this remain unclear, Boyd (1997) linked it to home ownership; a family belonging to a socially neglected class is less likely to have land or home ownership, reducing their ability to repay.

In case of gender, even after accounting for the gender wage gap, no statistical difference was found in the probability of default by male or female students (Harrast 2004; Volkwein and

Szelest, 1995). Choy and Li (2006) found that women take longer to repay the loans but have a slightly lesser probability of default, suggesting divergent repayment patterns.

Besides microeconomic factors, there exist some macroeconomic factors which play a role in determining the likelihood of loan defaults. Gunsel (2008) and Thiagarajan et al. (2011) found that GDP growth rates are negatively related to probability of default. Jiménez et al. (2006) studied the impact of cyclical fluctuations in real business cycles and concluded that the risk of defaults tend to rise during recessions probably due to lower employment rates.

An increase in unemployment will reduce the incomes, raise the debt burdens, increasing the probability of default (Vogiazas et al. 2011; Bofondi et al. 2011). An increase in money supply will have a two-fold result. Firstly, it will spur investment and consumption, raising income levels and consequently reduce the probability of default. Secondly, it reduces interest rates improves the access to cheaper funds. Vogiazas *et al*. (2011) found a negative relation between money supply and credit risk. Financial market conditions such as the risk-free interest Rate, stock market return rate etc. (Figlewski et al. 2012).

An increase in interest rate will raise probability of default by increasing debt burdens (Fofack 2005). The relationship between inflation and the ability to repay a loan is non-linear. Gunsel (2008) concluded that inflation is positively related to credit risk whereas, Aver (2008) and Bofondi et al. (2011) said that there exists no correlation between them. Vogiazas et al. (2011) in fact obtained results which show a negative relation between inflation and credit risk. An issue with educational loans is that students are not liable to repay the loan immediately after it is sanctioned, but only after graduating and so some sort of prediction or forecasting of macro variables becomes imperative (Berhani and Ryskulov 2014).

3. Default Prediction Models

3.1 Statistical Techniques

3.1.1 Logistic Regression (Logit Model)

Binary logistic regression is useful for estimating the probability of an event based on the maximum likelihood method. The logit model uses the logistic cumulative density function which is sigmoid in nature and is of the form $f(z) = \frac{1}{1+e^{-z}}$. It estimates the probability of occurrence of each event as:

$$P(y|X) = \frac{1}{1-e^{-(\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn)}}$$

Where $y_i \in \{0,1\}$ is the binary endogenous variable, β_i is the coefficient of the corresponding exogenous variable X_i (i = 1,2,3,…n). The probability of the occurrence of the binary endogenous variable is related to the exogenous variables as (Hosmer and Lemeshow 1989):

$$\log[p/(1-p)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Where p is the probability of the occurrence of the endogenous variable. (Lee *et al.* 2006) used logistic regression for credit risk modelling for personal loans and credit card loans.

3.1.2 Naïve Bayes

Naïve Bayes is based on the concept of conditional probability using Bayes' theorem. The classifier analyses this to create mapping function $f:(x_1,\ldots,x_n)$ over a training set $T = \{a_1,\ldots,a_n\}$ to test it on an unknown sample $X = (x_1,\ldots,x_n)$. Naïve Bayes classifier then chooses the class with the highest posteriori probability $P(c_j \mid x_1,\ldots,x_n)$ as per the minimum error probability criterion (Zhong and Li 2012). Thus, if $P(c_i|x) = \max_{i=1,\ldots,l} P(c_i|x)$, we can conclude that observation x belongs to class c_i. Sun & Shenoy (2007) used the Bayesian classifier to give early warning predictions for bank failures.

3.1.3 Multivariate Adaptive Regression Splines (MARS)

MARS (Friedman 1991) is a non-linear, non-parametric regression technique with a strong generalization ability. It can be viewed as a combination of the recursive partitioning used for creating classification and regression tree (Breiman et al., 1984) and the generalized additive modelling (Hastie & Tibshirani, 1990). It approximates the non-linearity of a dataset by using piecewise linear regression.

MARS builds an optimal model in two phases: Forward pass involves the creation of several basis functions to fit the data and the backward pass, which prunes the model to enhance its generalization ability. Known to have a shorter training time and strong intelligibility, it has been extensively used for forecasting and classification problems (De Gooijer, Ray and Krager, 1998; Lewis and Stevens, 1991).

3.2 Classical Machine Learning Techniques

Artificial Intelligence techniques such as Decision Trees (Lee et al. 2006); Neural Networks (West 2000; Malhotra, R. and Malhotra, D. K. 2002), Support Vector Machines, Random Forest, K nearest Neighbour (Henly and Hand 1996) are known to provide better results than statistical techniques.

3.2.1 Decision Trees

Decision tree produces a rooted tree consisting of nodes by repeated segmentation of data points by applying a series of rules based on inductive reasoning. Decision rules can be obtained by navigating from the root of the tree down to a leaf, as per the outcome of the tests along the path. Decision trees are mathematically suitable for problems such as credit risk modelling and are thus widely used (Lee and Chen 2005).

This paper uses the CART algorithm which can be explained using a three-step process (Chang and Chen 2008):

1) Recursive partitioning is used to choose variables and split points using a splitting criterion. The best predictor is chosen based on the impurity or diversity measures such as least squared deviation. (Breiman *et al.* 1984).

2) Once the tree is constructed, CART prunes it to create a nested subset of trees beginning from the largest tree constructed and till only a single node of the tree remains.

3) The optimal tree is selected from those constructed based on a suitable metric such as least cross-validated error.

3.2.2 K- nearest neighbour

K-nearest neighbour (KNN) is an instance-based clustering model. An observation with n parameters: $X_i = \{a_1,\ldots,a_n\}$ is represented by a point in an n-dimensional space. When fed an unknown sample with the same parameters, predictions are made for a new instance $Y_i = \{a_1,\ldots,a_n\}$ by searching through the entire training set for the K most similar instances and summarizing the output variable for those K instances using majority vote. The unknown sample is now assigned the most common class among its k-nearest neighbours with similarity defined in terms of Euclidean Distance.

3.2.3 Artificial Neural Network (ANN)

Artificial Neural Network uses a dense network of simple nodes called neurons organized in layers linked by weighted connections to transform inputs into outputs using a non-linear activation function, typically a sigmoid or a hyperbolic tangent (Felea et al., 2012). A training algorithm such as backpropagation or feedforward determines the weights for each node. The input layer receives the input which is then passed onto a sequence of hidden layers for processing until we obtain the desired output at the output layer.

Pang et al. (2002) used the MLP to discriminate between creditworthy and risky companies when they applied for loans. Desai, Crook & Overstreet (1996) concluded that neural networks are better at correctly classifying bad loans than logistic regression by comparing the two models in the credit union environment.

3.2.4 Support Vector Machine (SVM)

SVM classifies observations into classes by creating a hyperplane in the feature space such that the distance from the hyperplane to the data points is maximized which is essentially a quadratic optimization problem and is based on the structural risk minimization principle. In case of a linear relationship between observations, SVM uses linear kernels to find an optimal hyperplane which separates the data into two parts while simultaneously maximizing the distance between the hyperplane and the closest training points called support vectors. If the data is not linearly separated, SVM uses non-linear machines such as radial and polynomial kernels to find an optimal hyperplane. For instance, Yang (2007) used support vector machines to create an adaptive scoring system which could be adjusted using an on-line update procedure.

3.3 Ensemble Classifiers

Ensemble learning (Doumpos and Zopounidis 2007) is a machine learning technique where a group of weak models, called weak learners is combined to form a powerful model with a higher classification accuracy. Several classifiers are individually created, trained on different samples and their classification results are obtained by testing them on the same test sets.

3.3.1 Random Forest Algorithm

Random forest (Breiman , 2001) is a combination of decision trees, each of which individually classifies an observation. The algorithm then uses majority voting to select the class to which the observation belongs. It is robust against overfitting and performs better than many other classifiers. (Liaw & Wiener, 2002).

The algorithm can be explained in three steps as:
1) Suppose the training set as N observations and M variables. Then, for constructing each tree, a sample of n<N observations is taken at random with replacement and, m<M variables are selected randomly to split the nodes and is much smaller than the total number of descriptors available for analysis. Each tree is thus grown on a bootstrap sample of the training set. (Lariviere & Van den Poel, 2005).
2) Each tree is grown to its maximum capacity without any initial pruning.
3) Classes for the test data are predicted by aggregating the predictions of all the independently constructed trees.

Decision Trees choose which variable to split on using a greedy algorithm that minimizes error and are thus structurally similar and might have highly correlated predictions. Ensembles, however give better results if the predictions from their sub-models are uncorrelated or weakly correlated. In case of random forests, the learning algorithm is limited to a random sample of the variables from which it is allowed to search and so the resulting predictions from all of the subtrees have less correlation.

3.3.2 Adaptive Boosting (AdaBoost)

Adaptive Boosting (Freund and Schapire 1996) is a version of boosting which creates a series of decision trees. (Random forest, on the other hand produces decision trees in parallel) AdaBoost is a sequential procedure where if an observation is misclassified in the previous trees, its weight will be increased in subsequent trees, until it is correctly classified. For the

first classifier, all observations are equally weighted. For the second classifier, the weights on misclassified observations is increased and those on the correctly classified observations are reduced. AdaBoost then continues to emphasize the misclassified data while training subsequent classifiers until the entire training data has been trained. If ε_k is the sum of the misclassified instance probabilities of the classifier under consideration C_k then the probabilities for the next classifier are generated by weighing up the probabilities of C_k's wrongly classified instances by factor $\beta_k = \frac{1-\varepsilon_k}{\varepsilon_k}$ and then renormalizing them such that their sum equals 1. The algorithm then combines these classifiers $C_1,\ldots,C_k$ using weighted voting where C_k has a weight of $\log(\beta_k)$.

3.3.3 Extreme Gradient Boosting (XGBoost)

XGBoost (Chen and Guestrin 2016), an open source implementation of the Gradient Boosting Machine is a scalable and high performance machine learning system used for supervised learning problems and uses classification and regression trees as its constituent elements. The algorithm is similar to the random forest algorithm, except the trees are trained sequentially. A new tree is trained only after all the previous trees have been optimized. The objective function includes a loss function such as the mean square error (MSE) which evaluates how well the model is predicting when using the current sample and a regularization term to avoid overfitting. The regularization term helps to smooth the final weights to avoid over fitting by reducing the coefficients of the noise terms to zero. Consider an objective function of the form (Bhatia *et al 2017)*:

$$obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i)$$

Where l is a differentiable convex loss function while Ω is the regularization term. Let $\hat{y}_i^{(t)}$ be the prediction value at step t. Now:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{n} f_k(x_i) = \hat{y}_i^{(t-1)} f_t(x_i)$$

$$y^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + c$$

The loss function is expanded to the second order using Taylor expansion. This gives:

$$y^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i)) + \frac{1}{2} h_i f_t^2(x_i) + \Omega(f_t) + c$$

Where: $g_i = \delta_{\hat{y}(t-1)}\, l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \delta^2_{\hat{y}(t-1)}\, l(y_i, \hat{y}_i^{(t-1)})$

3.4 Model Stacking

In this two-step process, we first use several base classifiers to predict the class of each observation and then use a new learner to combine predictions of some of those classifiers to reduce the generalization error. Usually, the models with minimum correlation between them are aggregated to construct the ensemble.

Paleologo, Elisseeff & Antonini (2010) proposed credit evolution models based on K-means, SVM, decision trees and adaptive boosting algorithms. Yu *et al*. (2008) constructed ensembles using a series of artificial neural networks with different initial conditions and used maximizing decorrelation to choose the ensemble members. An ideal ensemble consists of highly correct classifiers that disagree to a great extent (Krogh and Vedelsby 1995). Logistic regression is the most widely used algorithm to combine the weak learners.

Suppose we have N different learning algorithms L1, . . . , LN on a single dataset S, which consists of examples si = (xi , yi), i.e., pairs of feature vectors (xi) and their classifications (yi) In the first step, a set of base-level classifiers C1,C2, . . . ,CN is generated, where Ci = Li (S). In the second step, a meta-level classifier is learned that combines the outputs of the base-level classifiers.

We use the learned classifiers to generate predictions for si: $\hat{y}k = Ci,k (xi)$. The training dataset consists of examples of the form $((\hat{y}i,1 \ldots \hat{y}i,n), yi)$, where the features are the predictions of the base-level classifiers. We use stacking with probability distributions and multi-response linear regression as proposed by Ting and Witten (1999) where the final predictions are probability distributions over the set of class values rather than single class values. The second-level probabilities are thus the probabilities of each of the class values returns by base-level classifiers. This allows us to use not just the predictions, but also the confidence of the base-level classifiers. The prediction of the base-level classifier C applied to example x is a PD:

$$pC(x) = (pC (C1|x), pC (C2|x), \ldots pC (Cm|x))$$

Where {C1, C2, . . . Cm} is the set of possible class values and pC (Ci|x) denotes the probability that example x belongs to class Ci as estimated (and predicted) by the classifier. The class Cj with the highest class probability pC (Cj|x) is predicted by the algorithm. The second-level attributes are the probabilities predicted for each possible class by each of the base-level classifiers, i.e.

$$pCj (Cj|x) \text{ for } i = 1, \ldots ,m \text{ and } j = 1, \ldots , N.$$

4. Data Description

We received borrower-wise educational loan data from four Indian public sector banks. It contains loans sanctioned from the year 2000 till 2011 and covered information on quantitative variables such as loan limit (the amount of loan initially sanctioned by the bank, parental income, interest rate that the bank would charge, loan liability (the amount that the student owed at the time of recording whether the student defaulted or not, i.e. at the time of writing off the loan) and categorical variables such as the type of course student is enrolled in, whether the student was an undergraduate or a postgraduate, the degree college in which the applicant had secured admission, gender, caste, religion, district to which the applicant belonged and the year of loan sanction.

Based on the type of course the student was enrolled in, we calculated the course duration [Appendix B] and added it to the year of sanctioning the loan, to obtain the year in which the student is expected to pass out. We then obtain the data on macroeconomic factors such as money supply growth rate, inflation rate (Consumer Price Index), GDP growth rates, unemployment rates and gross capital formation rates for the year in which the student would pass out and try to obtain a job/ start a venture. It is believed that graduate students with a debt would be unlikely to directly go for postgraduate studies (citation needed). Besides, the first job that the student obtains post graduating is expected to play a crucial role in determining the student's ability to repay (citation needed). Thus, we consider the impact of the strength of the economy at the time of the student passing out. Besides, the money supply growth rates were considered with a lag period of one year since its effects are not seen immediately. The data for unemployment levels was obtained from International Labour Organization's estimates, whereas that for Consumer Price Index was obtained from International Monetary Fund estimates. The rest of the macroeconomic variable data was obtained from Central Statistical Organization, Government of India's repository.

The district of the applicant was used to categorize him/her into to one of the following: rural, semi-urban, urban or metropolitan. The colleges were categorized into four tiers based on the State Bank of India's Scholars' List rankings provided by the National Institute of Ranking Framework (NIRF), Government of India [Appendix C]. We the data for 29,247 students, which after cleaning was reduced to 25,944 observations.

Table: Data Description		
Variable	Notes	Variable Type
Loan Limit		Quantitative
Loan Liability	Mean = Rs 1,82,700	Quantitative
Parental Income	Mean = Rs 1,19,000	Quantitative
Interest Rate	Mean = 13%	Quantitative
Degree Type	Undergraduate / Post-Graduate	Categorical
Quality of Institution	Classified as Tier 1 to 4	Categorical
Gender	Male / Female	Categorical
Caste	General/ Scheduled Caste/ Scheduled Tribe/ Other Backward Caste	Categorical
Area Type	Rural / urban/ Metropolitan/ semi-urban	Categorical
Courses	Engineering/ Medicine/ Law/ Management, etc.	Categorical
Unemployment Rate	Mean = 3.68%	Quantitative
Real GDP Growth Rate	Mean = 7.45%	Quantitative
Money Supply Growth Rate (M3) lagged by a year	Mean = 16.34%	Quantitative
Inflation (CPI)	Mean = 9.45%	Quantitative
Gross Capital Formation as % of GDP lagged by a year	Mean = 35.96%	Quantitative

A detailed summary of the dataset used has been presented in Appendix A.

5. Methodology

Given that our dataset consists of both categorical and continuous quantitative variables, to avoid bias against the categorical variables, we to scale the quantitative variables by using min-max scaling. We begin by performing probit regression (Results in Section 7.1) on the variables to understand their behaviour and then perform feature selection to discard variables that may not be significant with respect to the problem at hand or are heavily correlated with the existing variables.

As a part of this study, we conduct two sets of experiments. Firstly we construct models using only those variables that would be available to the banker at the time of sanctioning the loan (Experiment I). These include loan limit, parental income, interest rate, branch, institution tier, type of location, gender, under-graduate or postgraduate and caste. In the second experiment, we construct models using all the above mentioned variables along with the macro-economic variables for the Indian Economy (Experiment II). As mentioned above, we shall construct statistical models, classical machine learning models and ensemble models. To construct these models we divide our dataset into training and testing sets by first reserving 30% of the total observation points as our testing dataset.

Table1a: Distribution of Observations into Test and Training Set			
	Default	Non-Default	Total
Training Set	5012 (19.32%)	13149 (50.68%)	18161 (70%)
Testing Set	2229 (8.59%)	5554 (21.41%)	7783 (30%)
Total	7241 (27.91%)	18703 (72.09%)	25944 (100%)

Given the imbalance in our dataset, we perform experiments using three separate training sets. Training set I (training set as mentioned above) has the natural distribution of the dataset, training set II (obtained by under-sampling the majority class i.e. non-defaults) and training set III (obtained by performing Synthetic Minority Oversampling Technique i.e. SMOTE using 5 nearest neighbours to generate synthetic instances of the minority class i.e. the default class). In all cases, we preserve our testing dataset by separating it first. Furthermore, the feature selection as described in Section 6.1 below was performed using only the training dataset in order to avoid any bias in our testing dataset.

Table1b: Training Sets			
	Default	Non-Default	Total
Training Set I	5012 (27.60%)	13149 (72.40%)	18161 (100%)
Training Set II (minority under-sampling)	5012 (50%)	5012 (50%)	10024 (100%
Training Set II (SMOTE)	13149 (50%)	13149 (50%)	26298 (100%)

Highly parameterized models such as Random Forest, Extreme Gradient Boosting, Neural Networks and Support Vector Machines were fine-tuned by using 10-fold cross validation on this training dataset, while leaving our testing dataset untouched to avoid polluting it. Here, we perform stratified random Sampling. Each fold of our cross validation consists of 10% of the default observations and 10% of the total non-default observations of the entire training set. This ensures that the distribution of our original dataset is preserved. The specific parameters used to tune the models are mentioned in appendix G. Once we obtain the final parameters, we train the model once again on the complete training set and report the results obtained (Section 7.2) on the testing set. All models were trained using the same training set and tested on the same testing dataset to facilitate the comparison among them. Further, the training and testing sets for both the models are identical.

Once we obtain the probabilities of default, we choose the least correlated of the models (in terms of results obtained on the testing dataset), stack them to create an ensemble using logistic regression. We perform 10-fold cross validation and report the average results in Section 6.4.

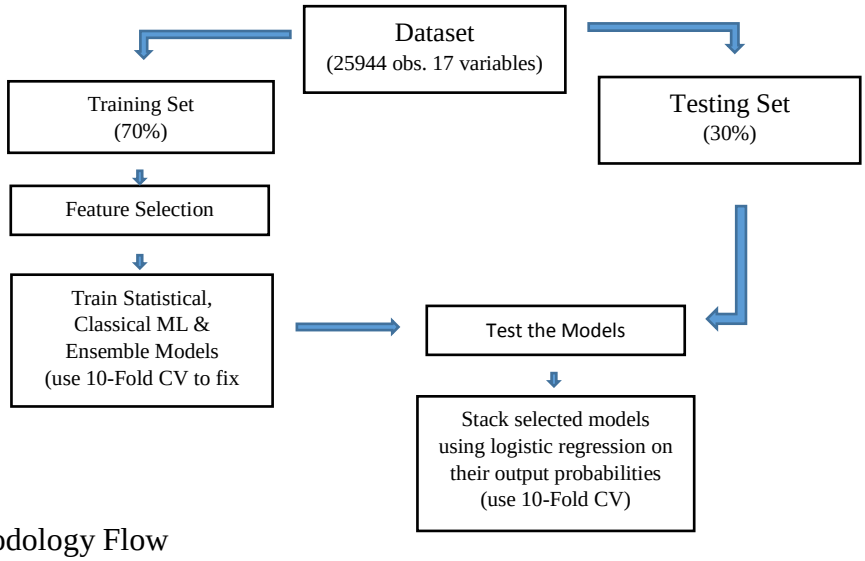


Figure 1: Methodology Flow

A QUMU virtual CPU version 1.5.3 @2.0 GHz and having 21.0 GB RAM was used to construct neural networks and Random Forests while a dell Inspiron 3537 i5-4200U CPU @1.60GHz and 4GB RAM and 500GB Hard Disk was used for the rest of the computation. R version 3.3.2 was used for constructing models while STATA 13 was used to obtain the regression results.

5.1 Feature Selection

Feature selection is a crucial step in credit risk modelling as it reduces the computation complexity and improves the performance of models by discarding irrelevant variables. Many studies have been conducted to compare feature selection techniques (Lin, McClean 2001; Ryu, Yue 2005; Tsai 2009).

We perform variable-wise Wald's Chi Squared test to check the association between the independent variables and the dependent variable (Default or non-default in our case) and consequently discard the independent variables which do not significantly explain the dependent variable. Based on the results below, we conclude that we can discard the variables gender and Loan Liability at the 99% confidence level.

Table 2: Wald's Chi Squared Test for Feature Selection	
Variable	Wald's Statistic (p-value in Bracket)
Loan Limit	399.0315 (0.0000)***
Loan Liability	0.0202 (0.8870)
Parental Income	126.9396 (0.0000)***
Interest Rate	393.9255 (0.0000)***
Degree Type	24.2287 (0.0000)***
Quality of Institution	14.054 (0.0002)***
Gender	1.1594 (0.2816)
Caste	49.6237 (0.0000)***
Area Type	36.6606 (0.0000)***
Courses	24.2287 (0.0000)***
Unemployment Rate	534.9756 (0.0000)***
Real GDP Growth Rate	328.566 (0.0000)***
Money Supply Growth Rate (M3) lagged by a year	310.5008 (0.0000)***
Inflation (CPI)	289.982 (0.0000)***
Gross Capital Formation as % of GDP lagged by a year	378.9105 (0.0000)***

5.2 Model Performance Criteria

Generally used performance measures of default prediction systems are Accuracy, Specificity, Sensitivity and the area under the receiver operating characteristic curve (Verikas *et al*. 2010).

Table 3: Structure of a Confusion Matrix			
	Observed results		
Predicted results		Non-default	Default
	Non-default	True Positive (TP)	False Positive (FP)
	Default	False Negative (FN)	True Negative (TN)

5.2.1 Overall Classification Accuracy

Accuracy is the percentage of correctly classified observations and is the most widely used metric to evaluate credit risk models.

$$\text{Overall Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

5.2.2 Specificity and Sensitivity

Type I error is the proportion of defaulters that the model classifies as non-defaulters. Type I error a potential credit loss for the banking industry which if higher is likely to raise the NPA burden. Type II error is the proportion of non-defaulters classified as defaulters. It results in credit denial and is also important in the case of student loans as it not just results in loss of business for the banks but also the denies education opportunity which the student probably deserved to get.

$$\text{Type I error} = \frac{FP}{FP+TN}$$

$$\text{Type II error} = \frac{FN}{FN+TP}$$

Sensitivity of the model = 1 – Type II error whereas the specificity of the model = 1 – Type I error.

5.2.3 Cohen's Kappa

Kappa (κ) (Cohen 1960) measures the inter-rater agreement for categorical variables and is considered a more robust measure than accuracy as it takes into consideration the possibility of the observation being classified correctly by chance.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

$$\text{Where } P_e = \text{probability of chance agreement} = \frac{(TP+FN)*(TP+FP)}{TP+FP+FN+TN} + \frac{(FN+TN)*(FP+TP)}{TP+FP+FN+TN}$$

$$P_o = \text{overall classification accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

If the raters are in complete agreement then κ takes value 1 and if there is no agreement among the raters other than what would be expected by chance, $\kappa \leq 0$.

5.2.4 Area under the Receiver Operating Characteristic (ROC) Curve

The receiver operating characteristic (ROC) curve shows the trade-off between sensitivity and specificity graphically. Area under the ROC curve (AUC) measures the discrimination power of a model and is immune to imbalance in data. The idea behind the curve is to evaluate the different class probabilities for the possible thresholds. The area ranges from 0 (worst classifier) to 1 (perfect classifier) while a completely random model will achieve an AUC of 0.5 (Altman and Bland 1994).

6. Analysis of Results

6.1 Regression

We use the probit regression to compute the regression results. Since the parental income and loan limit are highly skewed, we take their logarithms. The loan liability was not very skewed and roughly followed the normal distribution and so we normalized it. For colleges, tier 3 was chosen as the base category, while for castes, Scheduled Tribe was selected as the base. For branches and location, other courses and semi-urban area were the chosen base categories.

Table 4: Regression Results

Exogenous Variables	Coefficients (Standard errors)	Probabilities
Intercept	10.3919 *** (0.5531)	0.7101
Interest Rate	-0.3519 *** (0.0098)	-0.0241
log(parental Income)	-0.04338 *** (0.0053)	-0.0029
log(Loan Limit)	-0.7359 *** (0.0197)	-0.0503
Loan Liability (scaled)	0.3982 *** (0.0128)	0.0272
Undergraduate = 0; Postgraduate = 1	-0.0749 ** (0.0261)	-0.00512
Male = 0; Female = 1	-0.0053 (0.0194)	-0.0004
log(Unemployment Rate)	0.8144 ** (0.3008)	0.0557
log(GDP Growth Rate)	1.2059 *** (0.0790)	0.0824
Money Supply Growth Rate$_{t-1}$	-0.0082 (0.005927)	-0.0006
Inflation (Consumer Price Index)	-0.0760 *** (0.0094)	-0.0052

College Tier (Base = Tier 3)		
Tier 1	-0.0886 *** (0.0076)	-0.0061
Tier 2	-0.1127 $^{+}$ (0.07552)	-0.0077
Tier 4	0.0059 ** (0.0203)	0.0014
Caste (Base = Scheduled Tribe)		
General category	-0.2741 ** (0.0922)	-0.0111
Other Backward Castes	-0.2355 * (0.0929)	-0.0095
Scheduled Caste	-0.1208 ** (0.0417)	-0.0049
Area Type (Base = Semi-Urban)		
Metropolitan	0.1649 *** (0.0337)	0.0066
Urban	0.0395 ** (0.0129)	0.0015
Rural	0.0103 (0.0231)	0.0007
Course (Base = Other courses)		
Engineering	0.1129 *** (0.0244)	0.0077
Medicine	-0.1889 *** (0.0445)	-0.0129
Management	0.143578 *** (0.0321)	0.0098
Law	-0.135501 (0.2323)	-0.0093
Nursing	-1.491563 *** (0.3147)	-0.1019
Pharmacy	-1.063679 ** (0.3564)	-0.0727
***Significant at 99.9% **Significant at 99% *Significant at 95% $^{+}$ Significant at 90%		

6.2 Results of the Default Prediction Models

As discussed, we shall perform two sets of experiments: One using variables available to the banker at the time of sanctioning the loan and the other, using all earlier variables along with macro-economic data for the Indian economy. The results presented are as follows: Table 5 (Section 7.2.1) presents the results for Experiment I, while table 6 (Section 7.2) presents the results for Experiment II, both on the testing set.

6.2.1 Models using Variables available at the Time of Sanctioning the Loan (Experiment I)

Table 5: Model Results – I

	Training set with original data distribution				Training set obtained by under-sampling majority class				Training set obtained by Synthetic Minority oversampling Technique			
	Sensitivity	Specificity	AUC	Kappa	Sensitivity	Specificity	AUC	Kappa	Sensitivity	Specificity	AUC	Kappa
Logit	56.17%	69.10%	67.64%	0.2299	62.18%	66.80%	69.85%	0.2555	63.84%	67.59%	70.87%	0.2773
Naïve Bayes	57.87%	63.09%	65.21%	0.1824	60.70%	63.36%	67.19%	0.2082	61.78%	66.22%	69.34%	0.2461
MARS	66.44%	68.49%	74.28%	0.3079	64.60%	68.46%	73.88%	0.2927	65.46%	69.03%	73.91%	0.3058
Neural Network (1 hidden layer)	67.25%	66.64%	74.00%	0.2945	68.46%	65.57%	73.63%	0.2928	69.27%	66.64%	74.05%	0.3104
Neural Network (2 hidden layer)	65.59%	64.85%	71.37%	0.2626	67.74%	65.16%	72.22%	0.2828	68.51%	66.42%	73.51%	0.3021
K- Nearest Neighbour	59.58%	62.73%	62.23%	0.1928	61.60%	61.29%	64.01%	0.1950	61.87%	62.05%	65.89%	0.2046
Decision Tree	67.79%	66.91%	73.10%	0.3016	68.55%	67.30%	74.54%	0.3118	69.31%	67.99%	75.81%	0.3252
Support Vector Machine	58.55%	65.14%	67.32%	0.2085	59.44%	67.05%	69.12%	0.2356	60.16%	68.38%	70.02%	0.2556
Random Forest	64.69%	68.96%	73.47%	0.2989	68.69%	66.94%	74.57%	0.3091	70.88%	68.28%	76.98%	0.3407
Extreme Gradient Boosting	65.63%	69.84%	75.28%	0.3162	69.22%	69.72%	77.20%	0.3435	70.97%	69.66%	78.18%	0.3566
Adaptive Boosting	68.37%	69.54%	76.57%	0.3347	70.08%	69.66%	77.93%	0.3496	71.42%	70.42%	79.14%	0.3686

6.2.2 Models using all Variables including Macro-Economic Variables (Experiment II)

Table 6: Model Results – II

	Training set with original data distribution				Training set obtained by under-sampling majority class				Training set obtained by Synthetic Minority oversampling Technique			
	Sensitivity	Specificity	AUC	Kappa	Sensitivity	Specificity	AUC	Kappa	Sensitivity	Specificity	AUC	Kappa
Logit	64.96%	76.16%	75.41%	0.3833	65.37%	75.21%	76.32%	0.3752	65.86%	75.75%	77.18%	0.3857
Naïve Bayes	58.10%	72.22%	69.54%	0.2802	61.78%	73.78%	71.25%	0.3289	62.14%	75.35%	73.09%	0.3502
MARS	71.47%	79.47%	80.30%	0.4771	71.33%	79.33%	80.04%	0.4742	71.24%	78.99%	79.97%	0.4692
Neural Network (1 hidden layer)	63.39%	72.15%	73.56%	0.3236	68.42%	69.43%	74.42%	0.3339	69.40%	72.09%	74.88%	0.3716
Neural Network (2 hidden layer)	61.87%	72.20%	72.69%	0.3116	67.74%	69.07%	73.91%	0.3246	69.00%	71.50%	74.25%	0.3617
K- Nearest Neighbour	57.11%	72.51%	62.19%	0.2751	61.78%	69.84%	65.55%	0.2847	68.51%	70.42%	69.81%	0.3456
Decision Tree	69.85%	76.38%	77.83%	0.4257	70.88%	74.61%	79.01%	0.4128	72.27%	75.73%	78.82%	0.4372
Support Vector Machine	58.41%	75.33%	70.37%	0.3185	65.95%	71.98%	72.89%	0.3427	65.81%	72.04%	74.63%	0.3422
Random Forest	73.76%	80.93%	83.32%	0.5141	74.43%	79.35%	82.10%	0.4991	75.06%	80.88%	84.44%	0.5238
Extreme Gradient Boosting	63.44%	73.10%	74.50%	0.3348	70.44%	72.34%	77.96%	0.3827	70.88%	74.31%	78.18%	0.4092
Adaptive Boosting	74.56%	79.82%	83.81%	0.5061	74.34%	80.32%	84.38%	0.5108	74.92%	81.24%	85.10%	0.5274

There are a few observations worth making. Neural Network with a single hidden layer performs better than that with two hidden layers. Ensemble Models such as Random Forest and Boosting perform better, in general as compared to statistical and some machine learning

techniques. Among statistical techniques Multivariate Adaptive Regression Spline (MARS) performs exceptionally well with accuracies comparable to those of Neural Network and Adaptive Boosting. Decision Trees, too give fairly good results for this dataset. This is not very surprising considering the fact that decision trees are known to perform better when most of the predictors are binary or categorical. On the other hand, support vector machine performs rather poorly compared to other models. A part of this can be attributed to the fact that it generates a binary output unlike other models which generate numerical values. In general, Adaptive Boosting and Random Forest are the best models for this dataset.

Furthermore, we see a slight improvement in the classification performance for all classifiers when we under-sample the majority class in the training data to have a 50-50 distribution of both classes. The improvement in sensitivity however, comes at the cost of a reduced specificity. The relatively small gain in performance could also be coming from the fact that we are now training our models over a relatively smaller dataset. In case of SMOTE however, we see a larger improvement in both sensitivity and specificity as well as in the Area under the curve and Kappa. Thus shows that training our models over a balanced data tends to improve their predictive ability, but the gain is not very significant.

6.2.3 Model Comparison and Robustness Checks

Next, we perform a 10-fold cross validation over the entire dataset and perform pairwise t-test on the AUC thus recorded with 10 + 10 -2 = 18 degrees of freedom to check how many classifiers if any perform statistically as well as the best performing classifier (Adaptive Boosting in our case as determined on the basis of the average ROC) in terms of the average Area Under the ROC Curve with 99% confidence. The results shown in table 7 and 8 are on the basis of a 10 fold cross validation over the entire data and no data balancing procedure was performed in this case.

Table 7: 10 Fold CV to Compare Classifiers in Each Experiment with the Best Performing Classifier				
Classifier	Variables known at the time of Sanctioning the Loan (Experiment I)		All Variables (Experiment II)	
	Mean ROC	T-test	Mean ROC	T-test
Logit	67.24%	-15.6600 (0.0000)	74.82%	-8.1009 (0.0000)
Naïve Bayes	64.12%	-20.4600 (0.0000)	68.07%	-15.466 (0.0000)
MARS	74.06%	-5.3141 (0.0003)	79.68%	-5.2067 (0.0000)
Neural Network (1 Hidden Layer)	73.29%	-5.3700 (0.0000)	75.39%	-3.9128 (0.0010)
Neural Network (2 Hidden Layers)	72.18%	-3.4300 (0.0030)	74.33%	-11.1735 (0.0000)
Support Vector Machine	66.11%	-19.1700 (0.0000)	67.10%	-11.2190 (0.0000)
Decision Tree	71.73%	-6.3600 (0.0000)	76.68%	-3.4518 (0.0036)
K-nearest Neighbour	62.42%	-24.4000 (0.0000)	63.91%	-13.7662 (0.00)
Random Forests	73.75%	-4.4000 (0.0004)	81.83%	-2.2892 (0.0356)***
Extreme Gradient Boosting	75.33%	-2.0800 (0.0527)***	77.53%	-8.1293 (0.0000)
Adaptive Boosting	76.54%	-	82.84%	-
***Mean AUC is not statistically different from the mean AUC of Adaptive Boosting with 99% confidence. Figures in bracket indicate p-values				

We conclude that the AUC of Extreme Gradient Boosting (In Experiment I) and Random Forests (In Experiment II) is on an average quite similar to the AUC of Adaptive Boosting.

We now perform a pairwise, one-sided t-test with 10 + 10 -2 = 18 degrees of freedom on the AUC obtained by the cross validation process mentioned above to compare the performance of various models in Experiment I and II. The Null Hypothesis here is that the performance of the classifier in Experiment I is better than or equal to its performance in experiment II.

Table 8: 10 Fold CV to Compare Classifiers in Each Experiment with the Best Performing Classifier			
Classifier	Mean ROC: Variables Known at the time of Sanctioning the Loan (Experiment I)	Mean ROC : All Variables (Experiment II)	Pairwise T-test.
Logit	67.24%	74.82%	4.7556 (0.0010)***
Naïve Bayes	64.12%	68.07%	3.1307 (0.0121)**
MARS	74.06%	79.68%	5.5668 (0.0003)***
Neural Network (1 Hidden Layer)	73.29%	75.39%	4.0328 (0.0030)***
Neural Network (2 Hidden Layers)	72.18%	74.33%	4.449 (0.0016)***
Support Vector Machine	66.11%	67.10%	1.3618 (0.2063)
Decision Tree	71.73%	76.68%	11.7560 (0.0000)***
K-nearest Neighbour	62.42%	63.91%	0.7930 (0.4482)
Random Forests	73.75%	81.83%	14.6410 (0.0000)***
Extreme Gradient Boosting	75.33%	77.53%	2.7315 (0.0418)**
Adaptive Boosting	76.54%	82.84%	7.9127 (0.0000)***
***Significant at 99% confidence ** Significant at 95% Confidence. Figures in bracket indicate p-values			

Based on the above table, we can conclude that for all models except K-nearest Neighbour and Support Vector machines, considering macro-economic variables significantly improves the model.

6.3 Model Stacking

Since our dataset consists of diverse features along with non-linearities and involves behavioural aspects, it is likely that the real underlying distribution would be a function of several distributions a scenario in which a combination of models would yield improved

results. The fact that our ensemble classifiers have outperformed other models is an indication that combining several models might improve the results further.

The simplest of such models would use an unweighted average of the predictions of the input models to create the ensemble. More, generally, we might consider using a weighted average. However, we believe that each of our input models are fundamentally different and as such explain different aspects of our overall problem. Weighing them simply on the basis of the performance we obtained by using the models individually would be inappropriate.

A better approach might be to use model stacking which estimates these weights by using another layer of learning algorithm which is trained to optimally combine the input models and form the final set of predictions. For instance, using linear regression as the second-layer modelling would estimate the weighs by minimizing the least squares. Since its introduction by Woolpert (1992), stacking has been widely used to solve various problems.

It is common knowledge that ensembles of diverse base-level classifiers (classifiers constructed using different training algorithms, different hyper parameters with weakly correlated predictions) yield good performance. We therefore use a correlation matrix and select Naïve Bayes, Neural Network (With one hidden layer) Extreme Gradient Boosting, Random Forest and MARS for stacking since they belong to different families of algorithms and are therefore least correlated among our set of models. A detailed correlation matrix for all the models has been provided in Appendix D. To test our stacking framework, we perform 10-fold cross validation on our original testing set using the probabilities obtained by training the algorithms on the original training set (with the original distribution and without performing an data balancing procedures) and report the average performance.

6.4 Results: 10 Fold Cross Validation using Model Stacking

Table 9: Results: Model Stacking

	Average Sensitivity	Average Specificity	Average Area under ROC Curve	Average Cohen's Kappa
Experiment I	71.30%	72.79%	80.03%	0.3949
Experiment II	77.58%	81.98%	86.48%	0.5582

7. Discussion and Conclusion

We show that Educational loans is a case for application of artificial intelligence and the credit risk models can predict potential defaulters with a reasonable accuracy. Ensemble models tend to perform better than simple artificial techniques and statistical models and that the performance can be improved significantly by model stacking. We argue here that a stacked model created using a few sparsely correlated base models is likely to be the best model for predicting Educational loan defaults given that the interaction between diverse features would create non-linearities that are impossible to model using a single model. It is well known that statistical methods, simple machine learning models and ensemble classifiers have inherently different mathematical foundations and are suitable for different kinds of problems. Combining them would therefore yield a more robust model. Besides, due to lack of prior research in this sector and the current research by no means being exhaustive, there is little a priori knowledge of the distribution of educational loan defaults and the relationships between various factors that govern the distribution. Several known disadvantages of model stacking such as it being time-consuming, occupying excess memory, not being dynamic enough or having a lower throughput are of little concern here. What matters more is the performance of our algorithm and whether it can model the underlying data, something that a stacked model is likely to be better at.

A notable outcome of this study is the impact of collateralization on default rates and Moral Hazard. Until 2012, loans under 4 lakhs were collateral free while those exceeding 4 lakhs but not 7.5 lakhs required a third-party guarantee and those exceeding 7.5 lakhs needed a collateral in the form of property. It is evident from the figure below that collateral-free loans have a considerably higher rate of default as compared to the loans with collateral.

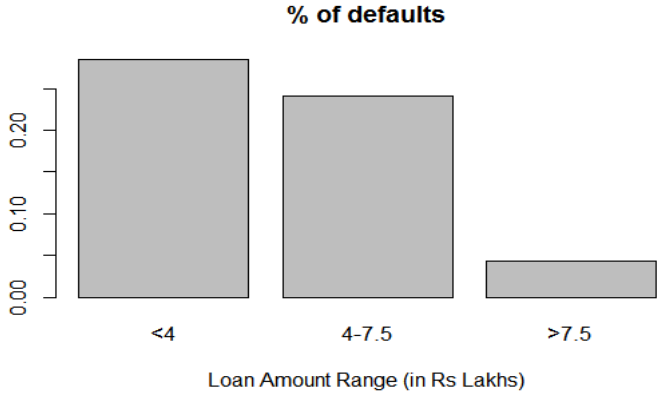


Fig 2: Impact of collateralization on default

This highlights the moral hazard problem where students on the cusp of lucrative careers declare bankruptcy to avoid paying their debt. Another aspect is the methodology employed by most public sector banks to determine the sanction amount and interest rate. Traditionally, banks have relied upon the reputation and exclusivity of the college in which the student has secured admission to determine the creditworthiness of the student. This results in banks giving students of top colleges collateral-free loans with lesser interest rates with the hope that they will be unlikely to default. As we can see from Figure 2, this is not the case and the default rates of Tier 3 and Tier 4 colleges are just faintly greater than those of tier 1 colleges.

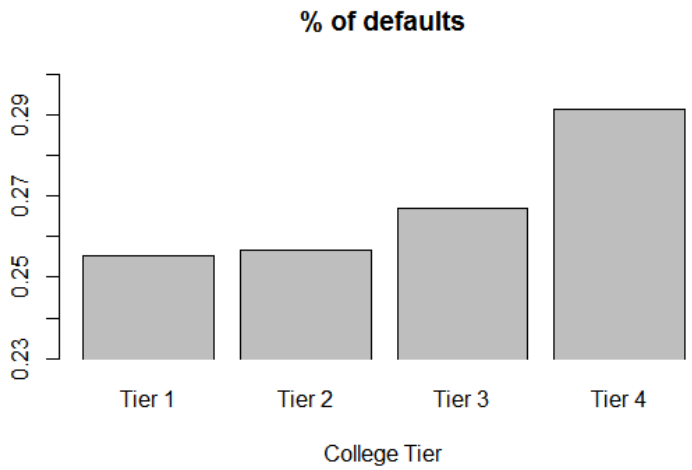


Fig 3: Impact of colleges on default rates

Considering the fact that tier 1 institutions in India are extremely selective and its graduates are likely to secure fairly good pay-scales, such a high default rate for its graduates clearly point to the possibility of these loans being prone to wilful and strategic defaults. In the Indian context where there exists a social stigma attached to non-repayment of loans and the fact that most Indian students stay with their families even after graduating.

Impact of factors such as whether the borrower belongs to rural or urban areas is unclear. For instance, a student belonging to rural area might secure employment in urban areas after completing education and so his/her ability to repay might not directly depend upon the rural-urban aspect which is recorded at the time of sanctioning the loan. However, this might indirectly affect the borrowing student's ability to repay. For instance, the quality of school education in urban areas is better when compared to that in rural areas where education spill-over effects are not significant and schools are of vernacular medium.

Further, under the Central Government Interest Subsidy scheme, loans sanctioned to students with parental incomes under 4.5 lakh per annum are insured by the government, essentially reprieving the bank of any potential downside in case of default. Thus, moral hazard in case of educational loans is actually two-staged. First, on part of the borrowers of collateral-free loans and second, on part of the bankers who make little to no effort to monitor the repayment sanctioned of educational loans. Moral Hazard is acute where costs for bad behaviour and choices are shifted to someone else. In the educational loan market, the central government directly and entirely bears the risk of students as well as the banks.

This system has huge information imbalances. Students seek loans knowing that the loan amounts are considerably larger than what they could recognize from their high school jobs. Even if they know that the job market is not great, they tend to be optimistic about it. Besides, the data that they need to make an informed decision is either unclear or unavailable. While obtaining a loan, students are generally unaware of their field of study, expected educational performance or future job prospects. Besides, universities have an incentive to exaggerate educational outcomes or job opportunities to public.

However, given the fact that loans are insured and education must be promoted and its social costs be borne, it is possible that banks might prefer to sanction a loans despite its probability of default being higher than acceptable. Nevertheless, credit risk model serve as a tool to monitor the loans by telling the banker the important variables that need to be monitored in case of each applicant. Besides, it helps the banks and the government estimate the provision needed for bad loans on the banks' balance sheet, resulting in more accurate financial reporting. Educational loans can also be divided into categories or tranches based on their credit risk.

It is important to highlight that considering the impact of macroeconomic conditions and the health of the economy greatly improve the classification accuracies. These variables directly impact the likelihood of a graduate securing a job and the entry-level salary. Although the macroeconomic scenario at the time of the student graduating is rather unknown at the time of sanctioning the loan, efforts can be made to forecast the conditions at the time of sanctioning the loans. Besides, the repayment data of each loan can be studied to decipher the patterns of and traits specific to defaulters and construct more sophisticated and accurate models.

8. References

Glorfeld, L.W. & Hardgrave, B.C. (1996). An improved method for developing neural networks: The case of evaluating commercial loan creditworthiness. Computer Operation Research, 23 (10), 933-944

Handzic, M., & Aurum, A. (2001). Knowledge discovery: Some empirical evidence and directions for future research. Proceedings of the 5th International Conference on Wirtschafts Informatics (WI'2001), 19-21.

Handzic M, Tjandrawibawa F. and Yeo J. (2003). How Neural Networks Can Help Loan Officers to Make Better Informed Application Decisions. Informing Science, 6, 97-109.

Lee, T. S.; Chiu, C. C.; Chou, Y. C.; Lu, C. J. 2006. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines, Computational Statistics and Data Analysis, 50, 1113–1130.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees (Wadsworth, Belmont, CA).

Yang, Y. (2007). Adaptive credit scoring with kernel learning methods, European Journal of Operational Research 183(3), 1521–1536.

West, D. (2000). Neural network credit scoring Models, Computers and Operational Research, 27, 1131–1152.

Malhotra, R.; Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems, European Journal of Operational Research 136(1), 190–211.

Henley, W. E.; Hand, D. J. (1996). A k-nearest neighbour classifier for assessing consumer credit risk, Statistician 44(1), 77–95.

Hosmer, D. W.; Lemeshow, S. (1989). Applied Logistic Regression. John Wiley & Sons, New York.

Lee, T. S., Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines, Expert Systems with Applications, 28, 743–752.

Chang, C. L., Chen C. H. (2008). Applying decision tree and neural network to increase quality of dermatologic diagnosis, Expert Systems with Applications 36(2): 4035–4041.

Pang, S., Wang, Y., and Bai, Y. (2002). Credit Scoring Model based on Neural Network. Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing.

L. Sun and P. Shenoy, (2007) Using Bayesian Networks for Bankruptcy Prediction: Some Methodological Issues, European Journal of Operational Research, 180(2), 738-753.

J. H. Friedman. (1991). Multivariate Adaptive Regression Splines, Annals of Statistics, 19(1), 1-141.

De Gooijer, J. G., Ray, B. K., & Krager, H. (1998). Forecasting exchange rates using TSMARS. Journal of International Money and Finance, 17(3), 513–534.

V. S. Desai, J. N. Crook and G. A. Overstreet, (1996). A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment," European Journal of Operational Research, 95(1), 24-47.

Z. Huang, H. Chen, C. J. Hsu, W. H. Chen and S. Wu,(2004). Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study, Decision Support System, 37(4), 543-558.

Hastie, T., & Tibshirani, R. (1990). Generalized additive models. London: Chapman & Hall.

Lewis, P. A. W., & Stevens, J. G. (1991). Nonlinear modelling of time series using multivariate adaptive regression splines (MARS). Journal of American Statistical Association, 86, 864–877.

G. Paleologo, A. Elisseeff and G. Antonini (2010). Subagging for Credit Scoring Models, European Journal of Operational Research, 201(1), 490-499.
L. Yu, S. Wang and K. Lai, (2008). Credit Risk Assessment with a Multistage Neural Network Ensemble Learning Approach, Expert Systems with Application, 34(2), 1434-1444.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. The Newsletter of the R. Project, 2(3), 18–22.

Lariviere, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications, 29, 472–484.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 785-794.

Bhatia S, Sharma P, Burman P, Hazari S and Hande R. (2017). Credit Scoring using Machine Learning Techniques, International Journal of Computer Applications (0975 – 8887), 161(11).

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In ICML, 96, 148-156.

Efron, B. and Tibshirani, R. J. (1993). An introduction to the bootstrap. Chapman and Hall, London.

Altman, D. G. and Bland, J. M. (1994). Statistics notes: Diagnostic tests 3: receiver operating characteristic plots. BMJ, 309(6948), 188.

Felea, I., Dan, F., Dzitac, S., (2012). Consumers Load Profile Classification Correlated to the Electric Energy Forecast. Proceedings of the Romanian Academy, Series A, 13(1), 80-88.

Badea L. M (2014). Predicting Consumer Behaviour with Artificial Neural Networks. Procedia Economics and Finance, 15, 238 – 246.

Dasgupta, C. G., Dispensa, G.S. and S. Ghose (1994). Comparing the predictive performance of a neural network model with some traditional market response models, International Journal of Forecasting, 10(2), 235-244.

Robin T., Bierlaire M. (2012). Modelling investor behaviour, Journal of Choice Modelling, 5(1), 98-130.

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In Advances in neural information processing systems, 7,231-238

Doumpos, M.; Zopounidis, C. (2007). Model combination for credit risk assessment: a stacked generalization approach, Annals of Operations Research 151(1): 289–306.

Fernández, E., & Olmeda, I. (1995). Bankruptcy prediction with artificial neural networks. In International Workshop on Artificial Neural Networks, 1142-1146

Verikas, A.; Kalsyte, Z.; Bacauskiene, M.; Gelzinis, A. (2010). Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey, Soft Computing 14(9), 995–1010.

Cohen, Jacob (1960). A coefficient of agreement for nominal scales". Educational and Psychological Measurement. 20 (1), 37–46.

Ryu, Y. U.; Yue, W. T. (2005). Firm bankruptcy prediction: experimental comparison of isotonic separation and other classification approaches, IEEE Transactions on Systems, Management and Cybernetics – Part A: Systems and Humans, 35(5), 727–737.

Tsai, C. (2009). Feature selection in bankruptcy prediction, Knowledge-Based Systems, 22(2), 120–127.

Barr, N. (2005). Financing higher education: A universal model. Financing Higher Education: Answers from the UK.

Choy, S., and Li, X. (2006). Dealing with debt: 1992–93 bachelor's degree recipients 10 years later (NCES 2006-156). National Centre for Education Statistics.

Lochner, and Monge-Naranjo, A. (2004). Education and default incentives with government student loan programs. National Bureau of Economic Research working paper.

Herr, E., and Burt, L. (2005). Predicting student loan default for the University of Texas at Austin. Journal of Student Financial Aid, 35(2), 27-49.

Steiner, M., and Teszler, N. (2005). Multivariate analysis of student loan defaulters at Texas A&M University.

Shlens, J. 2005. A tutorial on principal component analysis. Institute for Nonlinear Science. 13

Le Roux; B. and H. Rouanet (2004). Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis. Dordrecht. Kulwer, 180.

Greenacre, Michael and Blasius, Jörg (2006). Multiple Correspondence Analysis and Related Methods. London: Chapman & Hall: Texas Guaranteed Student Loan Corporation.

Knapp, L. G., & Seaks, T. G. (1992). An analysis of the probability of default on federally guaranteed student loans. The Review of Economics and Statistics, 74(3), 404-411.

Woo, J. H. (2002). Factors affecting the probability of default: Student loans in California. The Journal of Student Financial Aid, 32(2), 5-23.

Baum, Sandy and O'Malley, Marie (2003). College on Credit: How Borrowers Perceive Their Education Debt, Journal of Student Financial Aid, 33(3).

Kapsalis, C. (2006). Factors affecting the repayment of student loans.

Chou, T., Looney, A., & Watson, T. (2017). Measuring Loan Outcomes at Postsecondary Institutions: Cohort Repayment Rates as an Indicator of Student Success and Institutional Accountability (No. w23118). National Bureau of Economic Research.

Podgursky M., Ehlert M., Watson D., & Wittstruck J., Monroe R. (2002). Student loan defaults and enrolment persistence. Journal of Student Financial Aid, 32(3), 27-42.

Christman (2000). Multiple realities: Characteristics of loan defaulters at a two-year public institution. Community College Review, 27(4), 16-32.

Schwartz, S., and Finnie, R. (2002). Student loans in Canada: An analysis of borrowing and repayment. Economics of Education Review, 21(5), 497-512.

Gladieux, L., & Perna, L. (2005). Borrowers Who Drop Out: A Neglected Aspect of the College Student Loan Trend. National Center Report# 05-2. National Center for Public Policy and Higher Education.

Boyd, L. A. (1997). Discrimination in mortgage lending: The impact on minority defaults in the Stafford Loan program. The Quarterly Review of Economics and Finance, 37(1), 23-37.

Harrast, S. A. (2004). Undergraduate borrowing: A study of debtor students and their ability to retire undergraduate loans. Journal of Student Financial Aid, 34(1), 21-37.

Volkwein, J. F., and Szelest, B. P. (1995). Individual and campus characteristics associated with student loan default. Research in Higher Education, 36(1), 41-72.

Figlewski, S., Frydman, H., & Liang, W. (2012). Modelling the effect of macroeconomic factors on corporate default and credit rating transitions. International Review of Economics & Finance, 1, 87-105.

Gunsel, N. (2008). Micro and Macro determinants of bank fragility in North Cyprus economy. African Journal of Business Management, 6(4), 1323-1329

Thiagarajan, S., Ayyappan, S., & Ramachandran, A. (2011). Credit Risk Determinants of Public and Private Sector Banks in India. European Journal of Economics, Finance & Administrative Sciences, 34, 147-154.

Jiménez, G., & Saurina, J. (2006). Credit Cycles, Credit Risk, and Prudential Regulation. International Journal of Central Banking, 2(2), 65-98.

Bofondi, M., & Ropele, T. (2011). Macroeconomic determinants of Bad loans: evidence from Italian Banks.

Vogiazas, S. D., & Nikolaidou, E. (2011). Credit risk determinants in the Bulgarian banking system and the Greek twin crises. MIBES, 177-189.

Aver, B. (2008). An empirical analysis of credit risk factors of the Slovenian banking system. Managing Global Transitions, 6(3), 317-334.

Fofack, H. (2005). Non-performing loans in Sub-Saharan Africa: causal analysis and macroeconomic implications. Policy Research Working Paper Series (3769).

Berhani, R. and Ryskulov, U. (2014) Macro-economic Determinants of Non-Performing Loans in Albanian Banking System. 2nd International Conference on Economic and Social Studies.

David H. Wolpert (1992). Stacked generalization. Neural Networks, 5, 241–259.

Ting, K. M., & Witten, I. H. (1999) Issues in stacked generalization. Journal of Artificial Intelligence Research, 10, 271–289.

Zhu, J., Zou, H., Rosset, S. and Hastie, T. (2009). Multi-Class Adaboost, Statistics and its interface, 2, pp. 349-360.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.

Lin, F. Y., & McClean, S. (2001). A data mining approach to the prediction of corporate failure. In Applications and Innovations in Intelligent Systems VIII, 93-106.

Appendix A: Data Description

Quantitative Variables	**Default**	**Non-default**	**Total**
Loan Limit			
Loan Limit < Rs.400000	6762	16991	23753
Loan Limit ≥ Rs.400000	479	1712	2191
Loan Liability			
Loan Liability < Mean Loan liability (Rs.182700)	4320	11435	15755
Loan Liability ≥ Mean Loan liability (Rs.182700)	2921	7268	10189
Parental Income			
Parental Income < Mean Parental Income (Rs.119000)	4847	11597	16444
Mean Income (Rs.119000) ≤ Parental Income < 400000	2306	6581	8887
Parental Income ≥ 400000	88	525	613
Interest Rate			
Interest Rate < Mean Interest Rate (13.04%)	2172	4600	6772
Interest Rate ≥ Mean Interest Rate (13.04%)	5069	14103	19172
Categorical Variables			
Under-Graduates	14575	5241	19816
Post-Graduates	4128	2000	6128
College Tier (Quality of Institution)**			
Tier 1	108	315	423
Tier 2	58	168	226
Tier 3	3244	8900	12144
Tier 4	3831	9320	13151
Gender			
Male	4963	12729	17692
Female	2278	5974	8252
Caste			
General Category	4653	11241	15894
Scheduled Tribes	79	144	223
Scheduled Caste	523	1153	1676
Other Backward Castes	2004	6165	169
Area Type			
Metropolitan	793	1546	2339
Urban	2216	5784	8000
Semi-Urban	2129	5442	7571
Rural	2103	5931	8034
Courses			
Engineering	3858	10804	14662
Medicine	328	1301	1629
Management	1052	1915	2967
Law	12	25	37
Nursing	2	103	105
Pharmacy	2	33	35
Others	1987	4522	6509
Macroeconomic Factors ***			
Unemployment Rate			
Unemployment Rate < Mean unemployment Rate (3.68%)	4600	15148	19748
Unemployment Rate ≥ Mean unemployment Rate (3.68%)	2641	3555	6196
GDP Growth Rate			
GDP Growth Rate < Mean GDP Growth Rate (7.45%)	3074	10067	13141
GDP Growth Rate ≥ Mean GDP Growth Rate (7.45%)	4167	8636	12803
Money Supply Growth Rate (M3) with one year lag			
M3 growth < Mean M3 growth (16.34%)	4395	12439	16834
M3 growth ≥ Mean M3 growth (16.34%)	2846	6264	9110
Consumer Price Index (CPI)			
CPI < Mean CPI (9.45%)	3801	8312	12113
CPI ≥ Mean CPI (9.45%)	3440	10391	13831

Gross Capital Formation as % of GDP with one year lag			
Capital Formation < Mean Capital Formation (35.96 %)	3161	5371	8532
Capital Formation ≥ Mean Capital Formation (35.96 %)	4080	13332	17412
** The basis for college classification have been provided in Appendix C *** The year wise Macroeconomic variables have been provided in Appendix E			

Appendix B: Rules used for Calculating the Course Duration

Program	**Undergraduate course**	**Postgraduate course**
Engineering	4 years	2 years
Management	3 years	2 years
Medicine	5 years	3 years
Law	3 years	2 years
Pharmacy	3 years	2 years
Others	3 years	2 years

Appendix C: Methodology for Classifying Colleges

College Tier	**Source**
Tier 1	State Bank of India Colleges under Scholar Loan Scheme. List A colleges.
Tier 2	State Bank of India Colleges under Scholar Loan Scheme. List B colleges.
Tier 3	Colleges not in Tier 1 or 2 and in the top 100 ranks of National Institute of Ranking Framework's list in their respective categories.
Tier 4	Remaining institutions.

Appendix D: Correlation between the Models

Set 1 (Variables known at the time of sanctioning the loan: Experiment I)

	Neural network	XG Boost	Logit	Naïve Bayes	Decision Tree	Random Forest	Adaptive Boosting	MARS
Neural network								
XG Boost	34.45%							
Logit	74.55%	32.43%						
Naïve Bayes	60.91%	26.19%	83.22%					
Decision Tree	49.12%	57.56%	42.67%	39.19%				
Random Forest	64.56%	63.03%	53.46%	41.34%	77.56%			
Adaptive Boosting	43.34%	76.13%	40.14%	34.77%	70.13%	71.56%		
MARS	55.49%	51.90%	61.45%	50.30%	78.68%	77.18%	63.89%	

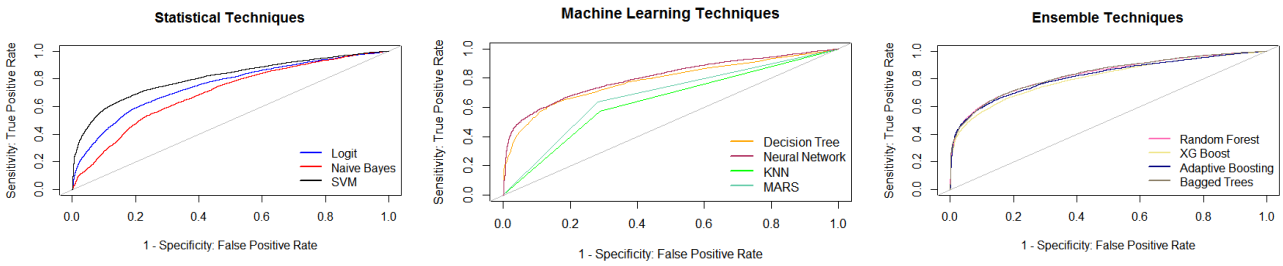
Set 2 (All Variables: Experiment II)

	Neural network	XG Boost	Logit	Naïve Bayes	Decision Tree	Random Forest	Adaptive Boosting	MARS
Neural network								
XG Boost	68.15%							
Logit	66.14%	48.15%						
Naïve Bayes	44.37%	29.73%	61.57%					
Decision Tree	71.90%	64.88%	56.79%	45.99%				
Random Forest	79.16%	75.02%	63.89%	51.40%	74.23%			
Adaptive Boosting	77.36%	80.74%	51.47%	45.78%	69.85%	85.10%		
MARS	77.18%	66.49%	71.76%	53.48%	74.70%	73.91%	71.88%	

Appendix E: Macroeconomic Variables

Year	**GDP Growth Rate**	**Gross Capital Formation as a % of GDP (t-1)**	**CPI**	**Money Supply Growth Rates (t-1)**	**Unemployment Rates**
Source	CSO	CSO	CSO	CSO	ILO
	8		3.4		4.31
2000-2001	4.15	26.97	3.7	16	3.775
2001-2002	5.39	24.21	4.3	16.1	4.316
2002-2003	3.88	25.65	4.1	13	3.929
2003-2004	7.97	25.02	3.8	14	3.889
2004-2005	7.05	26.17	3.9	15.9	4.4
2005-2006	9.48	32.45	4.2	20	4.331
2006-2007	9.57	34.28	6.8	22.1	3.724
2007-2008	9.32	35.87	6.2	20.5	4.154
2008-2009	6.72	38.03	9.1	19.2	3.906
2009-2010	8.59	35.53	12.3	16.2	3.55
2010-2011	8.91	36.3	10.5	15.8	3.537
2011-2012	6.69	36.53	8.4	13.4	3.623
2012-2013	4.47	36.39	10.2	17	3.574
2013-2014	4.74	34.7	9.5	14.1	3.53
2014-2015		31.4		15	
CSO = Central Statistical Organization, Government of India ILO = International Labour Organization					

Appendix F: ROC curves (For Experiment I)



Appendix G: Parameters and Model Specifications for Classifiers

1) Decision Trees

Specification	
Pruning Methodology	Minimizing the cross-validation error
Complexity Parameter for pruning	0.001

2) Support Vector Machine

Specification	
Kernel	Radial
Cost	8000
Gamma	0.6
Epsilon	0.1
Nu	0.5
Rho	0.7683
Sigma	0
Number of Support Vectors	3246

3) Neural Network

Specification	
Hidden units	6 for Single Layer; 6,3 for Double Layer
Error function	Sum of Squared Error
Activation Function	Hyperbolic Tangent
Convergence Threshold	0.01
Maximum Number of iterations allowed	1 Million

4) Random Forests

Specification	
Number of Trees	600

5) Extreme Gradient Boosting

Specification	
Evaluation Metric	Log-loss to be minimized
Learning Time of arrival	1
Maximum depth	7
Number of parallel threads used	6

6) Adaptive Boosting

Specification	
Max. depth of each Base tree used to optimize	7
Number of Trees constructed	200
Algorithm Used	SAMME (Zhu et. Al 2009)[116]

Appendix H: Rules obtained by Decision Tree

Rule number: 1954 (prob= 0.98)
If Loan Limit (Scaled)< 0.06062 OR Loan Limit (Scaled) >=0.02455 & Interest Rate (Scaled)< 0.75 OR Interest Rate (Scaled)>=0.63 & College Tier = 4 & Course = Undergraduate & Degree = Management then Prob(default) = 0.98
Rule number: 1466 (prob=0.96)
If Loan Limit (Scaled)>=0.06062 OR Loan Limit (Scaled)< 0.2986 & Interest Rate (Scaled)>=0.75 & Parental Income (Scaled)>=0.004044 & Course = Undergraduate & Area != Metropolitan & Caste = OBC then Prob(default) = 0.96
Rule number: 680 (prob=0.77)
If Loan Limit (Scaled)>=0.06062 OR Loan Limit (Scaled)< 0.1165 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)>=0.008307 & Parental Income (Scaled)< 0.04913 & Area = Rural & (Degree = Management OR Degree = Engineering) then Prob(default) = 0.77
Rule number: 662 (prob=0.74)
If Loan Limit (Scaled)>=0.2129 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)>=0.75 & Parental Income (Scaled)< 0.004044 & College Tier!= 1 then Prob(default) = 0.74
Rule number: 2018 (prob=0.71)
If Limit (Scaled)>=0.06062 & Parental Income (Scaled)< 0.008307 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled) < 0.75 & Course = Undergraduate then Prob(default) = 0.71
Rule number: 8 (prob=0.65)
If Loan Limit (Scaled)>=0.06062 & Loan Limit (Scaled)< 0.2004 & Parental Income (Scaled)< 0.04913 & Interest Rate (Scaled)>=0.664 & Course = Undergraduate & Caste != SC & Caste!= ST then Prob(default) = 0.65
Rule number: 2228 (prob=0.64)
If Loan Limit (Scaled)>=0.06062 & Loan Limit (Scaled)< 0.07315 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)< 0.04913 & Area = Metropolitan & Caste = OBC then Prob(default) = 0.64
Rule number: 16 (prob=0.62)
If Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Loan Limit (Scaled)>=0.03543 & Loan Limit (Scaled) < 0.06062 & Parental Income (Scaled) < 0.004215 & Caste != SC & Course = Undergraduate then Prob(default) = 0.62
Rule number: 121 (prob=0.61)
If Loan Limit (Scaled)>=0.03543 & Loan Limit (Scaled)< 0.06062 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled) < 0.75 & Parental Income (Scaled)< 0.004329 & Caste != SC then Prob(default) = 0.61
Rule number 1110: (prob=0.58)
If Loan Limit (Scaled)>=0.03543 & Loan Limit (Scaled)< 0.06062 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)< 0.004329 & Caste = SC then Prob(default) = 0.58
Rule number: 365 (prob=0.55)
If Loan Limit (Scaled)>=0.03543 & Loan Limit (Scaled) < 0.06062 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled) < 0.75 & Parental Income (Scaled)< 0.004329 & Area != Rural & Course = Undergraduate then Prob(default) = 0.55
Rule number: 364 (prob=0.40)
If Loan Limit (Scaled)>=0.03543 & Loan Limit (Scaled)< 0.06062 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)< 0.004329 & Area = Rural then Prob(default) = 0.40
Rule number: 1254 (prob=0.36)
If Loan Limit (Scaled)>=0.06062 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)< 0.04913 & Degree = Management & College Tier = 4 then Prob(default) = 0.36
Rule number: 2246 (prob=0.36)
If Loan Limit (Scaled)>=0.06062 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)>=0.008307 & Parental Income (Scaled)< 0.04913 then Prob(default) = 0.36

Rule number: 1442 (prob=0.35)
If Loan Limit (Scaled)>=0.03543 & Loan Limit (Scaled)< 0.06062 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Caste != SC & Parental Income (Scaled)< 0.008429 then Prob(default) = 0.35
Rule number: 630 (prob=0.34)
If Loan Limit (Scaled)>=0.06062 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)< 0.04913 & Area = Metropolitan & Course = Postgraduate then Prob(default) = 0.34
Rule number: 720 (prob=0.32)
If Loan Limit (Scaled)>=0.03543 & Loan Limit (Scaled)< 0.06062 & Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Caste != SC & Parental Income (Scaled)>=0.008429 then Prob(default) = 0.32
Rule number: 449 (prob=0.31)
If Loan Limit (Scaled)>=0.03543 & Loan Limit (Scaled)< 0.06062 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Caste != SC & Parental Income (Scaled)>=0.004215 then Prob(default) = 0.31
Rule number: 314 (prob=0.29)
If Loan Limit (Scaled)>=0.06062 OR Loan Limit (Scaled)< 0.07315 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)>=0.004215 & Area != Metropolitan then Prob(default) = 0.29
Rule number: 73 (prob=0.25)
If Loan Limit (Scaled)>=0.06062 & Interest Rate (Scaled)>=0.75 & Parental Income (Scaled)>=0.004215 & Area = Metropolitan then Prob(default) = 0.25
Rule number: 158 (prob=0.21)
If Loan Limit (Scaled)>=0.06062 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)>=0.004215 & Parental Income (Scaled) < 0.2356 then Prob(default) = 0.21
Rule number: 1274 (prob=0.21)
If Loan Limit (Scaled)>=0.06062 & Loan Limit (Scaled)>=0.07315 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)>=0.004215 & Parental Income (Scaled) < 0.2356 then Prob(default) = 0.21
Rule number: 312 (prob=0.20)
If Loan Limit (Scaled)>=0.06062 & Loan Limit (Scaled)>=0.07315 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)>=0.004215 & Parental Income (Scaled) < 0.2356 Degree = Engineering then Prob(default) = 0.20
Rule number: 626 (prob=0.18)
If Loan Limit (Scaled)>=0.06062 & Loan Limit (Scaled)>=0.07315 & Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Parental Income (Scaled)>=0.004215 & Parental Income (Scaled) < 0.2356 Degree = Management & College Tier = 2 OR College Tier = 1 then Prob(default) = 0.18
Rule number: 74 (prob=0.18)
If Loan Limit (Scaled)>=0.06062 & Loan Limit (Scaled)< 0.2129 & Interest Rate (Scaled)>=0.75 & Parental Income (Scaled) >= 0.2356 & College Tier = 2 then Prob(default) = 0.18
Rule number: 38 (prob=0.15)
If Loan Limit (Scaled)>=0.06062 & Parental Income (Scaled)>=0.04913 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Area = Urban then Prob(default) = 0.15
Rule number: 290 (prob=0.13)
If Loan Limit (Scaled)>=0.06062 & Loan Limit (Scaled)< 0.2129 & Parental Income (Scaled) >= 0.2356 & Interest Rate (Scaled)>=0.75 & Area != Metropolitan then Prob(default) = 0.13
Rule number: 582 (prob=0.09)
If Loan Limit (Scaled)>=0.06062 & Loan Limit (Scaled)< 0.2986 & Interest Rate (Scaled)>=0.75 & Parental Income (Scaled) >= 0.2356 & Area != Metropolitan & Course = Postgraduate then Prob(default) = 0.09
Rule number: 1272 (prob=0.07)
If Loan Limit (Scaled)>=0.06062 & Parental Income (Scaled) >= 0.2356 & Interest Rate (Scaled)>=0.63 & Interest Rate (Scaled)< 0.75 & Degree != Engineering then Prob(default) = 0.07
Rule number 1985: (prob=0.02)
If Loan Limit (Scaled)>=0.06062 & Loan Limit (Scaled)>=0.2986 & Interest Rate (Scaled)>=0.75 & Parental Income (Scaled)>=0.01674 & Area = Semi-Urban & College Tier = 1 & Degree!= Management then Prob(default) = 0.02