WORKING PAPER, NO: 601

Predicting Educational Loan Defaults: Application of Machine Learning and Deep Learning Models

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

Neel Shah
Columbia Business School
ns3481@columbia.edu

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

Year of Publication – December 2021

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 Machine Learning and Deep Learning Models

Abstract

Student (educational) loans are highly vulnerable to default risk and thus guaranteed by governments. We show that collateral-free educational loans are a case for the application of Machine Learning models to predict default factors with greater accuracy, helping banks in risk management and the government in designing economic policies of interest suspension and credit guarantees. We argue that heterogeneous ensembles constructed using stacking or a Hill Climb Ensemble approach are most suited for imbalanced data set since the interaction between diverse features would create non-linearities that are impossible to model using a single algorithm. Borrower/student social background emerges as an important feature explaining the loan defaults, warranting a correction in the public policy in designing the educational loan schemes for under privileged borrowers. Our paper also shows that Machine learning models are not systematically biased against underprivileged borrowers and do not lead banks to refuse credit. Ours is the first study to apply Statistical, Machine learning and Deep learning Models on a data set of student loans.

Key words: Credit Risk, Educational Loans, Statistical Techniques, Artificial Intelligence Techniques

Predicting Educational Loan Defaults:
Application of Machine Learning and Deep Learning Models

1. Introduction

Student loans are popular method of financing higher education in many countries and the governments are encouraging students to borrow with various economic incentives such as low interest rates, direct transfer of interest subsidies and extended loan repayment periods and supporting banks by under writing these loans with credit guarantee. US has an outstanding student debt of $1.6 trillion with lenders experiencing significant default rates and is a cause of worry for the federal government as it guarantees these student loans (Johnson, 2019 and Muelleer and Yannelis 2019). The scenario in Japan (Armstrong, Dearden, Kobayashi and Nagase, 2019) and UK is also similar[1]. Thus, identifying factors influencing student loan defaults is helpful for banks in risk management and for government in designing suitable public policies.

Academic research has extensively examined loan default modelling through statistical and machine learning models mainly applied to corporate firms (Altman 1968,Shin, Lee and Kim 2005, Kim and Sohn 2010, Liang, Lu, Tsai and Shih 2016) consumer (Lessmann, Baesens, Seow and Thomas, 2015) and credit card loans (Lee, Chiu, Chou and Lu 2006) and more recently to Peer-to-Peer lending datasets (Wang, Jiang, Ding, Lyu, Liu,2018, Ma, Sha, Yu, Yang, Niu, 2018 and Liang and Cai,2020). But the empirical evidence on default modelling of student loans is limited. Knapp and Seaks (1992) applies Probit model and find that parental income (also Bandyopadhyay 2016) and presence of both the parents at home have an impact on student loan defaults. A review of research on educational loans (published between 1971 and 2007) by Gross, Cekic, Hossler and Hillman (2009) summarizes that most of the studies are descriptive in nature with limited application of quantitative techniques. The review concludes that factors such as race, socioeconomic background, educational attainment, type of postsecondary institution, student debt levels, and post-school earnings are important determinants of default. However, Gross et al (2009) says that "*we are struck by the relative dearth of recent research on student loan default using large national data sets and rigorous statistical methods*". The current paper is addressing this gap.

[1] https://www.thesun.co.uk/news/8808887/student-debt-not-being-paid/

Educational loans are like consumer loans in ticket size but the two significant features are these loans have no or low collateral and longer repayment periods. Unlike consumer or credit card loans assessing a *priori* probability of a student successfully completing the academic course and securing a well-paying job is quite challenging (Barr and Crawford, 2005). Macro-economic factors and behavioral aspects of a student also influences loan repayment. Often, the governments extend incentives such as interest subsidy or waiver, elongating repayment periods and guaranteeing the loans; thus, educational loans have public policy concerns in emerging markets[2], while banks have mandated lending with social objectives. In such directed lending, banks have little choice in selection of borrowers. With high information asymmetry coupled with little or no collateral, these loans may turn to default, adversely affecting the profitability of banks and straining government finances. Thus, data-driven algorithmic risk modelling is required (Paisittanand and Olson 2006), not to discriminate or deny credit to a section of borrowers but to strengthen risk-based monitoring, to ensure efficient allocation of risk capital for the loan portfolio and in purchasing costly credit guarantee schemes.

Our paper attempts to bridge this gap by applying Statistical, Machine learning and Deep learning models on a unique dataset of collateral free educational loans of economically weak households of India. Our paper is motivated by Viera, Barboza, Sobreiro and Kimura (2019)'s work on application of ML in predicting default risk of home loans of PMCMV Programme[3] of Brazil. PMCMV is aimed at providing housing loans to low-income families, without any credit risk analysis. Like the Brazilian housing loan program, educational loans in India are also granted without much credit risk analysis, have longer repayment periods and are generally unsecured. We demonstrate that tree based, heterogenous ensemble models are better at classifying defaults with higher accuracy for an unbalanced dataset. Considering the non-linearities and complex behavioural patterns associated with this dataset, this paper explores and recommends suitable model for such datasets.

We find that, simple statistical methods such as Naïve Bayes and Logistic regression significantly underperform, while Tree-based ensemble models such as Light GBM, XGBoost,

[2] In India the Government policy directs Agriculture lending and lending to other under privileged sections. Other countries like Brazil, Belarus, China and Russia also have such directed lending programs.

[3] The Brazilian Social Housing Program (PMCMV) was created in 2008, with the enactment of the Federal Law number 11.977, and it has been the largest housing program ever implemented in Brazil.

Random Forests and Adaptive Boosting emerge as the best homogeneous models, followed by deep learning models such as CNN and LSTM. Due to the complexity of the data and presence of non-linearities, Heterogeneous ensembles are the best performers with HCES and HCES-Bag securing the top spot. Stacking and HCES do significantly better than model-agnostic techniques such as simple averaging.

Our analysis indicates that the performance of ML techniques is superior to statistical models such as logistic regression; ML techniques do not require any restrictive assumptions on specific prior knowledge and offer flexibility to loan officers to adopt new criteria in credit risk management.

This paper adds to the credit risk quantification literature by comparing various classification techniques, using a data set of economically weak students of an emerging market. It considers the idiosyncratic borrower specific aspects (parental income, academic program, geographical area, etc.), institutional factors (rating of the academic institution) as well as systematic (external) macro-economic factors such as growth rate, inflation, and unemployment s. This could allow banks to improve their risk management processes by adjusting their loan portfolios to idiosyncratic institutional and macro-economic factors, for instance, a rise in unemployment or a fall in GDP growth rate may alert the bank on possible increase of risk of default.

Besides contributing to the growing literature on the application of ML on a new dataset, this paper attempts to provide a real-world application of ML which might provide direction in designing appropriate public policies for the benefit of weaker sections of the society. Having better credit risk models could enhance the quality of the loan portfolio, reduce the guaranteed insurance burden on government, facilitate a more targeted loan monitoring process, helps in banks' risk management decisions and in designing suitable public policies on interest subsidy and other economic incentive programs.

This paper is divided into Six sections. Section two reviews relevant literature on application of statistical and ML techniques on loan defaults. Section three presents' data description, Section four outlines methodology adopted in the paper; results are discussed in Section Five and Section Six presents summary and conclusions of the paper.

2. Review of related literature:

The primary aim of credit risk quantification is to identify variables that distinguish between credit worthy borrowers and less credit worthy ones. Classical studies on this subject are primarily on application of statistical techniques such as logistic regression (Zavgren 1985) conditional logit model (Ohlson 1980), Probit analysis (Casey, Gee and Stinkey, 1986) discriminant analysis (Altman,1968), and CART and MARS (Lee, Chiu, Chou and Lu 2006). Most of these techniques are validated with higher Accuracy ratios and Power curves. Kumar and Ravi (2007) present a comprehensive review of empirical works on bankruptcy prediction and credit risk assessment published during the period (1968-2005). The review supports that ensemble classifiers outperform stand-alone models and suggests the need for research in developing new hybrid systems in various forms for different problems. A bibliometric survey by Prado, Alcântara, Carvalho, Vieira, Machado and Tonelli (2016) present review of papers published between 1968 and 2014, notices wider use of artificial intelligence and complex computing techniques with multi-disciplinary interest.

Among the recent studies, Paleologo, Elisseeff, Antonin (2010) propose subagging, an ensemble classification technique, particularly suitable for highly unbalanced data. It can build and validate robust models, with missing information, class imbalance and non-iid data points, the use of subagging improves the performance of the base classifier and subagging decision trees achieve better performance. Lessmann, Baesens, Seow and Thomas (2015) compared 41 classifiers using six performance measures across eight real-world credit scoring data sets; heterogeneous ensembles outperform the rest and provide some evidence that more accurate scorecards facilitate sizeable financial returns. Vieira et al (2019) show superiority of ensemble classifiers improving credit risk evaluation in low-income housing loans and the model is helpful in reducing defaults.

Malekipirbazari and Aksakalli (2015) use random forests (RF) for predicting borrower status on a data set of social lending platform indicating that the RF-based method outperforms the standard credit score assigned by a commercial credit scoring firm.. Wang, Jiang, Ding, Lyu, Liu (2018), propose ensemble mixture random forest (EMRF) behavioural scoring model based on a mixture survival analysis framework to predict the dynamic probability of default in peer-to-peer (P2P) lending, EMRF has a better performance in terms of predicting the monthly dynamic probability of default, while compared with standard mixture cure model and logistic regression. The model provides a meaningful output for timely post-loan risk

management. Random Forest algorithm outperforms logistic regression, decision tree and other machine learning algorithms in predicting default (Zhu, Qiu, Ergu, Ying, Liu 2019). Recent studies (Zhang, Wang, Chen,Shang,and Tian 2017) have also explored deep learning models such as Long Short Term Memory (LSTM) for overdue classification of credit evaluation with higher accuracy than well accepted data mining tool Artificial Neural Network (ANN). It is the unique ability of extracting time-series information that makes the LSTM outperform traditional approaches (Liang and Cai, 2020) and Convolutional Neural Networks (Kvamme, Sellereite, Aas and Sjurse 2018) in default prediction. However, application of LSTM requires large data, best suited for timeseries analysis and thus offer limited performance improvement in smaller datasets.

All the reviewed studies applied various statistical Machine learning and ensembling models to predict the default of corporate firms, consumer/instalment loans, small business loans, credit card loans and Peer to Peer (P2P) lending also. But our paper is on application of models on student loan data which is a first attempt.

3. Data Description

3.1 Educational Loans for Low Income Groups:

In India, educational loans, introduced in 2002 relatively a new financial product compared to other countries like US[4], UK[5], Canada[6] Australia[7]and Thailand[8]. Initially educational loan up to ₹ 0.40 million is provided without any collateral, third-party guarantee, or margin requirement, recently this has been enhanced to ₹0.75 million. To extend the outreach of educational loans, Government of India introduced the Central Sector Interest Subsidy (CSIS) Scheme in 2009 providing interest subsidy on educational loans for students of economically weaker sections (EWS), whose parental annual income is less than ₹ 0.45 million. EWS borrowers get interest moratorium not only during the study but also a year thereafter. However, if they secure a job, the interest moratorium is limited to six months.

4 The first federal student loans, however, provided under the National Defence Education Act of 1958, were direct loans capitalized with U.S. Treasury funds, following a recommendation of economist Milton Friedman
5 In UK, Educational loans started in 1989
6 The Canada Student Loans Plan established in 1964.
In Australia Higher Education Contribution Scheme (HECS) in 1989 and a grand experiment was embarked upon.
8 Thailand established student loan Fund in 1996

At the end of December 31, 2019, the amount of educational loans is ₹716.85 billion (Indian Rupees) of which ₹65.97 billion are NPAs. A staggering 86% of NPAs originate from unsecured loans not backed by any guarantor or collateral, the ticket size of these loans is less than ₹0.4 million[9]. As the government policy directs banks to lend collateral free loans, banks have little choice in credit decision. Thus, two important aspects of public policy are providing collateral free loans and waiver of interest payment during study period on loans given to EWS borrowers.

3.2 Data

We received borrower-wise educational loan data from four Indian public sector banks. It contains loans sanctioned from the year 2000 till 2011 and has information on variables such as loan limit (the amount initially sanctioned by the bank), parental income, interest rate, academic discipline of the student, nature of the academic program (undergraduate or a postgraduate), the academic institution in which the student had secured admission, gender, caste (social background), religion, geographical location of the borrower (rural, urban etc..) and the year of loan sanction.

Based on the academic discipline[10] of the student, we calculated the course duration [Appendix B] and added it to the year of sanctioning the loan, which gives the year in which the student is expected to pass out. To assess the employment potential of the student borrower we have considered macro-economic factors and unemployment rate at the time of graduating.

The data of Money supply growth rates were considered with a lag of one year since its effects are not seen immediately. Unemployment data was obtained from International Labour Organization (ILO)'s estimates and Consumer Price Index data from International Monetary Fund. The data of macroeconomic variables money supply, inflation, and GDP growth rate are sourced from Central Statistical Organization, Government of India.

The data base categorises the loan sanctioning bank branch as rural, semi-urban, urban or metropolitan based on location of the branch. We assumed that a student a student of rural area may request the branch of a bank very nearer to her geographical area. Thus, she is categorised

[9] Indian Banks Association
[10] In Indian system an undergraduate program in Medicine, Pharmacy and Law is for five years, Engineering four years and all other disciplines Humanities, Social sciences, Natural sciences, and commerce are three years.

as a rural student although she may be studying in a university located in metropolitan area. Academic Institutions were categorized into four tiers based on State Bank of India's Scholars' List and rankings of the National Institute of Ranking Framework (NIRF) [see Appendix C]. We started with a data for 29,247 students, which after cleaning is reduced to 25,944 observations. Details of the dataset is presented in Appendix A.

4. Methodology

Since our dataset consists of both categorical and continuous quantitative variables in different orders of magnitude, we scale the quantitative variables using min-max scaling. We perform logit regression using the full set of variables to understand their behaviour and perform feature selection and discarded variables that might be statistically irrelevant for predicting default or strongly correlated(linearly) with existing variables.

We follow Baesens, Gestel, Viaene, Stepanova, Suykens, and Vanthienen (2003) and Lessmann et al (2015) to perform a comparative analysis of multiple classification algorithms. Although model comparison is not the focus of our study, we use several of those algorithms to find the most suitable model for our dataset.

We conduct two sets of experiments. First, using only those variables that would be available to the bank at the time of sanctioning the loan (Experiment I). These include loan limit, parental income, interest rate, branch, institution tier, geographical location, gender, academic program (under-graduate or postgraduate) and social background(caste). While we recognize the implication of including variables such as caste and rural/urban background of the borrower, we feel it is necessary to include these as several public policies are targeted to these groups.

In the second experiment, for model construction we also consider macro-economic variables relevant at the time of student pass out. Macro-economic variables influences the individual loan defaults, these are unemployment rate, real GDP growth rate, inflation and gross capital formation (Louzis, Vouldis, Metaxas, 2012) and money supply (Song and Zhang, 2020).

We divide our dataset into training and testing sets in 80:20 ratio. Since parental income and loan limit are highly skewed, we take logarithmic value. The base categories we chose are Tier

3 for ranking of academic institution, the general category for caste, other courses for academic discipline, and semi-urban area for location.

We use linear models such as Logistic Regression, Naïve Bayes, Multivariate Adaptive regression Spline (Friedman 1991) and KNN, an instance-based classifier. We then use non-linear algorithms such as multi-layer perceptron, Radial Support Vector Machines, Decision Trees and a series of tree-based ensembles such as Random Forests, Adaptive Boosting, XGBoost and lightGBM besides using deep learning methods LSTM and One-Dimensional Convolutional Neural Networks (CNN). Following Lessmann et al 2015, we apply Heterogeneous ensembling techniques such as simple averaging, weighted averaging, Stacking and Hill-Climbing Ensemble (HCES). The HCES algorithm starts with an ensemble of a few base models and sequentially adds more base models to it until we see improvement in performance. To reduce overfitting, we use bagging for improving ensemble selection (HCES-Bag).

The presence of class imbalance could affect the predictive performance of classifiers, given the imbalance in our dataset (which is an intrinsic data characteristic), we test whether this imbalance, significantly alters our results? By following Brown and Mues (2012) and Garcia, Marques and Sanchez (2019) we perform experiments using three separate training sets (Table 1): Training set I has the natural distribution of the dataset, training set II is obtained by under-sampling the majority class i.e. non-defaults and training set III is obtained by performing Synthetic Minority Oversampling Technique (SMOTE) using 5 nearest neighbours to generate synthetic instances of the minority class i.e. the default class (Zhu, Qiu , Ergu, Ying, Liu 2019).

In all the three cases, we first preserve our testing dataset by separating so that the models are tested on the dataset that is reflective of the real-world distribution of default and non-default borrowers. Feature selection was performed using only the training dataset to avoid any bias in our testing dataset.

Highly parameterized models such as Support Vector Machines, Neural Networks and Tree-based ensembles - Random Forests and boosting were fine-tuned by performing randomized grid search over 10-fold cross validation on the training set, while leaving our testing set untouched to avoid bias. Folds were obtained by performing stratified random sampling such that each fold consists of a tenth of total default and non-default observations of our training

set. The parameters used to construct tuning grids can be found in Appendix G. All models were trained on identical training sets and tested on the same testing set to facilitate comparison among them.

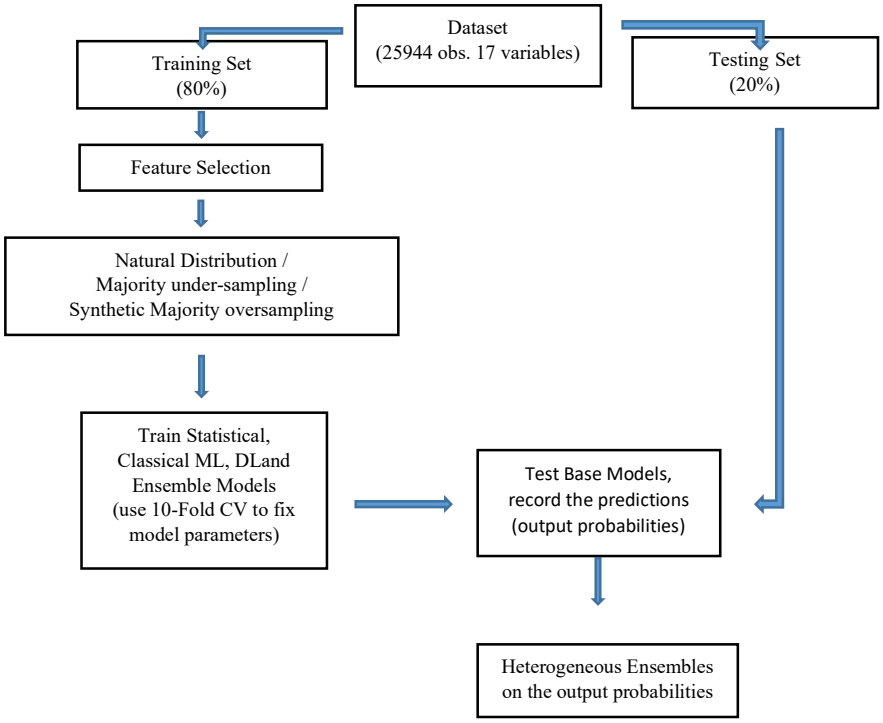


Figure 1: Methodology Flow

Fusster, Pinkham, Ramadorai and Walther (2020) argue that while Machine learning models improve the overall credit provision, they increase rate disparity between and within groups; effects mainly arise from flexibility to uncover structural relationships between default and observables, rather than from triangulation of excluded characteristics. We test whether using more sophisticated algorithms like tree-based ensembles and deep-learning methods could negatively impact the under privileged borrowers compared to using logistic regression. Here, we define privilege as borrowers having high parental income, belonging to other than under privileged social category (General) and from urban area.

4.2 Feature Selection

Feature selection is a crucial step in credit risk modelling (Ryu and Yue 2005; Tsai 2009), it reduces the computation complexity and improves the performance of models by discarding irrelevant variables. We perform an individual and grouped Wald's Chi Squared test to check the association between independent variables and the dependent variable (Table 2). We discard gender with 99% confidence.

It is interesting to note that quality of academic institution is significant as a standalone variable, but its impact on default is insignificant. This merits more explanation since it has traditionally been a key variable in determining the interest rate charged by the banks and is often incorporated in the decision process. Based on Wald's test, we also drop the money supply growth rate among macro factors.

4.3 Model Evaluation Criteria

Accuracy ratio (the percentage of correctly classified observations) is the popularly used evaluation criteria of prediction models, but Accuracy ratio is often heavily impacted by the cut-off score, especially in unbalanced datasets. Therefore, to measure overall performance and to facilitate comparison among models we choose the Area under the ROC Curve (AUC) which is immune to imbalance data set. Since AUC is an overall scalar evaluation measure, it might make misleading conclusions when the cost of misclassifying observations in one class is different from the cost of misclassifying it in the other. We therefore report directional measures sensitivity and specificity.

Type I error is the proportion of defaulters that a model classifies as non-defaulters. It is potential credit loss for the banking industry, higher the error, more is the likelihood of default. Type II error is the proportion of non-defaulters classified as defaulters and results in credit denial. Sensitivity (1 – Type II error) and specificity (1 – Type I error) are derivatives of Type-II and Type -I errors. In this sense, high specificity should not compromise the correct classification of the majority class. To attain balance between these conflicting goals, we report balanced accuracy, the arithmetic means of specificity and sensitivity. At the same time, the percentage of correctly identified safe loans out of the total loans identified as safe is also important not only in business, but also false positives could spoil customer relationships. We

thus report precision (the positive predictive value) along with the F-measure, the Harmonic Mean of precision and sensitivity.

5. Analysis of Results

Data Description:

Our data consists of individual institutional and macro-economic variables (Appendix -A), these are quantitative and categorical variables. Although counter-intuitive, loans with a higher interest rate and higher loan amount have less defaults, probably due to quality collateral associated with loan amount exceeding ₹0.75 million (Figure 2). Loans of less than ₹0.4 million are generally without collateral and have lower interest rates and higher defaults. These small-ticket loans account for higher proportion in volume due to large number of such loans, number of defaults are also high for these loans. Until 2012, loans under ₹0.4 million were collateral free while those exceeding ₹0.4 million but not ₹0.75 million required a third-party guarantee and those exceeding ₹0.75 million needed a collateral generally in the form of property. It is evident from the Figure 2 that despite smaller loan ticket size is, collateral-free loans have a considerably higher rate of default as compared to the loans with collateral. Geography (rural or urban) of the borrower is also not closely associated with loan defaults (Figure 3). Loan defaults are not very closely associated with quality or ranking of educational institutes (Figure 4).

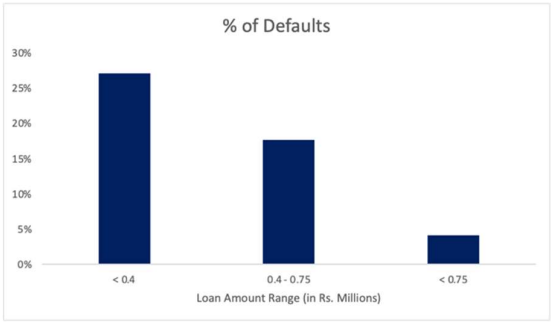


Figure 2: Impact of collateralization on default

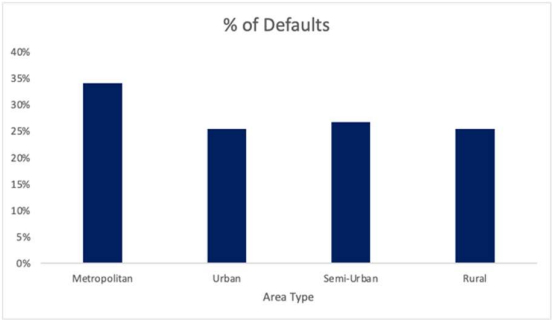


Figure 3: Default Rate and Geographical Area

This may be the moral hazard problem where students on the cusp of lucrative careers avoid paying their debt.

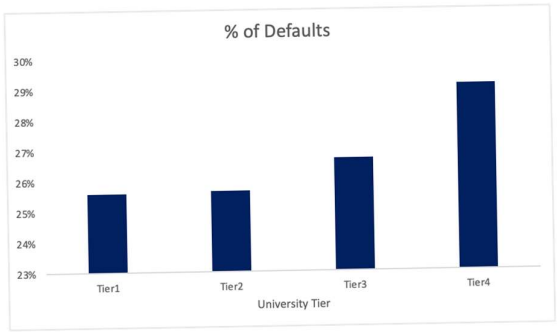


Figure 4: Ranking of Academic Institution and Default rate

5.1 Regression Analysis

Table 3 presents logit regression results. Among the borrower specific factors, loan defaults are significantly influenced by interest rate on loan, parental income, loan limit, academic program, and academic discipline. As expected, an increase in loan interest rate and loan limit are negatively correlated and significant. Students with low parental income are likely to default, results are significant. Students with Engineering and Management courses are positively associated with loan defaults, while students with Pharmacy and Nursing are negatively associated. Among the four social categories two of them are positively associated with loan defaults which warrants the need for policy intervention to provide the educational loans to the under privileged. Students with rural background are less likely to default, contrary to the imperfections noticed by Hoff and Stiglitz (1990) that loan defaults among rural borrowers () is high. While the defaults of students from metropolitan area are significant. Defaults are not associated with ranking of academic institutions, this is contrary to the US report (2018)[11] on student loans that default rate depend more on student and institutional factors than on average levels of debt. Among the macroeconomic factors, except money supply, GDP growth rate other macro factors have expected relationship and significant with loan defaults.

[11] Clayton Judith Scott (2018), The looming student loan default crisis is worse than we thought, Brookings Institution's Evidence Speaks Reports, Vol 2, #34 January 10, 2018

One of the reasons why we expect Machine Learning algorithms to work better than linear methods is presence of non-linearities, as our data set contains multiple categorical variables. We use the Box-Tidwell (1962) procedure to formally show this by testing whether the logit transformation is a linear function of the predictor. We do this individually for all the continuous variables (Table 3) and find that the interaction term is significant for all except for GDP growth rate and Gross Capital formation.

5.2 Results of Machine Learning Models

Table 4 presents the results for Experiment I and Table 5 for Experiment II; both the experiments are done on the same testing set. For brevity, we report the AUC, F-measure and balanced accuracy, and the rest in appendix E.

For experiment I (Table 4), Tree-based ensemble models Light GBM, XGBoost, Random Forests and Adaptive Boosting emerge as the best single models with AUC nearly 80%, balanced accuracies exceeding 70% and F-scores are around 0.8. This is in line with the findings of Doumpos and Zompunidis (2007), Alfaro et al (2008) and Sun and Li (2012). Tree-based ensembles are often credited as strong classifiers because of their ability to successfully capture non-linear patterns in noisy datasets and robustness to overfitting (Kruppa, Schwarz, Arminger and Ziegler, 2013). These Models occupy top four positions in terms of all the metrics in Tables 4 and 5 except for F-measure, where SVM and Decision Trees outperform. Outperformance of SVM may be due to high sensitivity, shadowed by a very low specificity and may be due to high precision for Decision Trees.

In terms of performance, homogeneous ensembles are followed by deep learning models such as CNN and LSTM, Decision Trees and then multi-layer perceptron. Simple statistical methods such as Naïve Bayes and Logistic regression significantly underperform due to the complexity of the data and presence of non-linearities. Decision Trees seem to have a better predictive power probably most of our predictors are binary. Neural Networks with two hidden layers perform better than the one with a single layer, but the difference is unlikely to be significant. Deep learning models do not perform exceptionally well despite the complexity probably they require large amounts of data, and the given data set is not of time series nature. In this sense, we agree with Lessman et al (2015) that the complexity or recency of classifiers is an inappropriate indicator of its predictive ability. On the back of our randomized grid search, we

use a two layered CNN with a Rectified Linear Unit activation and a single layered LSTM with a hyperbolic tangent activation. SVMs with a radial kernel gave a superior performance compared to linear and polynomial kernel.

We observe only a marginal change in performance in terms of all major metrics by under-sampling the majority class or over-sampling the minority class. Barring the improvement in MARS, SVM and to some extent in decision trees, data balancing worsens most models in both experiment I and II although we find little evidence to claim that the outperformance or underperformance is statistically significant due to conflicting results among similar models, among multiple metrics and balancing techniques. Besides, with a class ratio of approximately 2:3 default versus non-default, using the original class distribution is unlikely to create severe data imbalance issues.

Using 10-fold cross validation (Table 7) over the entire dataset, we perform pairwise t-test on the AUC with degrees of freedom (10+10 –2 =18) to check which classifiers are statistically equivalent to the best performing classifier (Light GBM for experiment I and XGBoost for Experiment II) in terms of the average Area Under the ROC Curve (with 99% confidence without any data balancing). In both experiments, the performance of tree-based ensembles is statistically like one another because of their structural similarity.

We report results for balanced accuracy and F-measure in Table 7 (for sensitivity, specificity and recall in Appendix J). Light GBM has the highest balanced accuracy and like AUC, this is not statistically different from the other tree-based ensembles. Support Vector Machine has the highest F-Measure (due to considerably higher sensitivity than the others). High sensitivity implies lower type II error, so fewer genuine borrowers marked as potential defaulters which might be important in case of social lending. While support vector machine has high sensitivity, has low specificity, Naïve Bayes, Decision trees and tree-based ensembles outperform the rest in terms of specificity implying they are more accurate at identifying defaulters, important from credit risk perspective. For most models, precision exceeds 75% with tree-based ensembles having the highest among the group. We could perform multi-criteria ranking of models to identify individual rankings although this is not the central focus of the paper. In general, tree-based ensembles as a group outperform the rest.

We observe a significant improvement in performance (nearly 4% on an average in terms of AUC with an increase in specificity, precision, and sensitivity) by including macro-economic variables which clearly highlights the need to incorporate the broader economic conditions in the decision-making process. The most prominent exception to this is the LSTM which saw a significant reduction in the specificity and precision and 2.5% reduction in AUC; others are reduction in specificity in Naïve Bayes, compensated by a larger increase in sensitivity and in Decision Trees model reduction in F -measure is due to sensitivity. The ranking of models remains unchanged even after including macro variables, with an exception to LSTM. While the improvement in terms of AUC, Balanced Accuracy and F-measure was observed in Naïve Bayes, Logistic Regression models and an increase in specificity is noticed in SVM (Table 7 and Appendix J).

We compare the performance of models in the two experiments by conducting, one-sided t-test with 18 degrees of freedom on the AUC obtained by 10-fold Cross-Validation (Table 8). For all models except LSTM, adding macro-economic variables significantly improves the model, in terms of average AUC, although the evidence is weaker in case of Adaptive Boosting. The largest improvement was observed in simpler linear models (over 6% on an average), whereas the tree-based ensembles improved by about 4 % in terms of average AUC (Table 8).

Overall, heterogeneous ensembles (Table 9) are the best performers with HCES and HCES-Bag securing the top spot. In our case, bagging does not significantly improve the HCES algorithm, probably we have fewer base models to choose. Stacking and HCES do significantly better than agnostic techniques such as simple averaging. We find weak evidence to suggest that simple or weighted averaging alone could also enhance the predictive ability of the model, but the improvement is marginal. For Robustness, we perform a pairwise t-test with 18 degrees of freedom on 10-Fold Cross Validated AUC to test whether the performance of HCES-Bag is statistically different from the other ensembles (Table 10). Hill Climbing ensemble clearly outperforms all other algorithms although the improvement from bagging is very marginal.

We do Shap analysis of our Random Forests Model. Shapley value has become the basis for several methods that attribute the prediction of a machine-learning model. The use of the Shapley value is justified by citing the uniqueness of the result that satisfies certain good properties (Lundberg, Erion and Lee (2018), Sundararajan, and Najmi (2020)) We choose Shapely analysis of Random Forests, because tree-based ensembles are the best homogeneous

ensemble models for this dataset, with little disagreement among the methods. We find that loan limit, parental income, and interest rate are the most significant features. followed by macroeconomic variables, ranking of institution and duration of the academic program (UG or PG). Specifically, the caste (Social background) of the borrower is not a very important predictor (Figure 5)

5.3 Model Complexity and social impacts

Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2020), raise an important question, does a more sophisticated statistical technology (in the sense of reducing predictive mean squared error) produces predictions with greater variance than a more primitive technology? In other words, will a set of borrowers be considered less risky by the new technology, or "winners", while other borrowers will be deemed to be riskier ("losers"), relative to their position under the pre-existing technology. The key question is then how these winners and losers are distributed across societally important categories such as social background (caste), income, or gender.

In the case of educational loans, we ask whether better performing models have a bias against students coming from rural areas, under-privileged social status (castes) or low parental income groups compared to simpler models? Does adopting more complex models result in more loans being denied in case of social lending? We consider two types of 'flips': Default (non-default) flips are observations that are classified as non-defaults(defaults) by logistic regression and reclassified as defaults(non-default) by complex models. In Table 11, we present whether the percentage of such flips for the entire test set is different from the percentage of corresponding flips for the under-privileged groups using 10-Fold CV. For both, experiments I and II, the percentage of default flips (adverse flips) for all groups are less than the number for the entire set, significant reduction is observed for the low-income group, particularly for experiment I. For rural areas, flips in either direction is not statistically different. We do however see a significant increase in non-default flips (non-adverse flips) for the low parental income group which is not particularly alarming. One reason for the absence of systematic bias is that the predictive power comes from non-discriminating variables as seen above, so the presence of bias is likely to be low.

6. Conclusion

This paper demonstrates that unsecured collateral free educational loans is a case for application of Machine Learning models to examine the factors determining defaults, and to predict potential defaulters with a reasonable accuracy. We show that, Tree-based ensemble models tend to perform better than statistical models and the model performance can be improved significantly by using heterogeneous ensembling approaches such as HCES and Stacking. Our argument is that an ensemble model created using multiple weak learners is likely to be the best model for predicting educational loan defaults given that the interaction between diverse features would create non-linearities.

We also tried deep learning models, but our results are more in favour of ensemble models. By following the direction of Lessmann et al (2015) we link the algorithm's characteristics to that of a dataset and find evidence that tree-based ensembles are best suited for our dataset with multiple categorical features, dynamic behavioural patterns, noisy data and unpredictable macroeconomic regimes. Besides, they are robust against overfitting and can handle correlated features. HCES, which is our best performing classifier also comprised most of those algorithms as base classifiers.

The classification accuracies are greatly improved by considering the macroeconomic factors in the model building. Macroeconomic factors directly impact the likelihood of a graduate securing a job and the amount of entry-level salary. Of course, borrower level repayment data may be studied to decipher the patterns of and traits specific to defaulters and to construct more sophisticated and accurate models.

A limitation of this paper is that the data is for approved loans only, which is a recurrent issue across all credit datasets. Notwithstanding the advantages of objective decision making, a certain amount of subjectivity might help in social lending. The models might identify the social (caste) or economic background of a borrower as discriminatory variables which might aggravate the prejudice against a group of borrowers making them less likely to receive funding and the opportunity to rise socially (Viera *et al*, 2019). Decisions made by models might turn out to be discriminatory even if the computing process itself is fair (Žliobait, 2017, Robb and Robinson, 2018). It would be interesting to study the impact of other variables such as the high school grades of the student and her performance throughout graduate school.

In the context of Covid-19, many governments have announced easy and flexible loans to economically weaker sections, low-income groups, and micro business units. For example, In US no interest is accruing on student loans and monthly payments have been suspended for loans in repayment through September 2021[12]. Machine Learning Models applied in this paper may provide direction for such policy decisions. Understanding the loan defaults have impact on normative public policy concerns such as providing employment, employment insurance, counselling socially disadvantaged groups to avoid loan defaults and extension of interest subsidy to economically weaker sections of the society. Further studies are needed to recognise the influences of behavioural biases on loan defaults by culling out big data from social media sources.

[12] Consumer Financial Protection Bureau

References

Alfaro, E., García, N., Gámez, M., & Elizondo, D2008. Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. Decision Support Systems, 45(1), 110-122.

Altman Edward I 1968., Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23 (4) , 589-609

Armstrong, S., Dearden, L., Kobayashi, M., Nagase, N. 2019. Student loans in Japan: Current problems and possible solutions. *Economics of Education Review*, *71*, 120-134.

Baesens, B., Gestel V, Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J. 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the operational research society, 54(6), 627-635.

Bandyopadhyay, A. 2016. Studying borrower level risk characteristics of education loan in India. IIMB Management Review, 28(3), 126-135.

Barr Nicholas and Crawford 2005. Financing Higher Education: Answers from the UK, Routledge, Taylor and Francis Group, London and New York.

Box GEP, Tidwell PW 1962. Transformation of the independent variables. Technometrics 4, 531–550.

Brown, I., Mues, C.2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446-3453.

Casey M, Gee Mc, Stinkey C (1986) Discriminating between reorganized and liquidated firms in bankruptcy. Account Rev 61(2):249–262

Doumpos, M., Zopounidis, C 2007 Model combination for credit risk assessment: A stacked generalization approach. *Ann Oper Res* **151,** 289–306, https://doi.org/10.1007/s10479-006-0120-x

Friedman J H (1991), Multivariate Adaptive Regression Splines, The Annals of Statistics. 19 (1), 1-67

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A. 2020. Predictably unequal? the effects of machine learning on credit markets. The Effects of Machine Learning on Credit Markets The Journal of Finance, forthcoming

García, V., Marqués, A. I., Sánchez, J. S. 2019. Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. Information Fusion, 47(May), 88-101.

Gross, J. P., Cekic, O., Hossler, D., Hillman, N. 2009. What Matters in Student Loan Default: A Review of the Research Literature. Journal of Student Financial Aid, 39(1), 19-29.

Guo, Y., Zhou, W., Luo, C., Liu, C., and Xiong, H. 2016. Instance-based credit risk assessment for investment decisions in P2P lending. European Journal of Operational Research, 249(2), 417-426.

Hoff Karla and Stiglitz Joseph (1990) Introduction: Imperfect Information and Rural Credit Markets: Puzzles and Policy Perspectives, *The World Bank Economic Review,* 4 (3,) pp 235-250

Johnson Daniel M (2019) What Will It Take to Solve the Student Loan Crisis? September 23, 2019, https://hbr.org/2019/09/what-will-it-take-to-solve-the-student-loan-crisis

Kim Hong Sik, Sohn So Young (2010), Support vector machines for default prediction of SMEs based on technology credit, European Journal of Operational Research 201 (3), 838-846

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.

Kruppa, J., Schwarz, A., Arminger, G., and Ziegler, A. 2013 Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, 40(13), 5125-5131.

Kumar, P. R., Ravi, V. 2007. Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. European journal of operational research, 180(1), 1-28.

Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. 2018. Predicting mortgage default using convolutional neural networks. Expert Systems with Applications, 102, 207-217.

Lee, T.S. , Chiu . C, Chou, Y., Lud C., 2006. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics and Data Analysis, 50(4), 1113–1130.

Lessmann, S., Baesens, B., Seow, H. V., Thomas, L. C. 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136.

Liang, L., Cai, X. 2020 Forecasting peer-to-peer platform default rate with LSTM neural network. Electronic Commerce Research and Applications, 43, 100997. https://doi.org/10.1016/j.elerap.2020.100997

Liang, Lu, Tsai, Shih, 2016, Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study, European Journal of Operational Research, 252 (2) 561-572

Louzis Dimitrios P, Voluldis Angelos T, Metaxas Vasilios L 2012 Macroeconomic and bank-specific determinants of non-performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios Journal of Banking & Finance, 36(4) 1012-1027

Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee 2018 “Consistent individualized feature attribution for tree ensembles.” arXiv preprint arXiv:1802.03888.

Ma X, Sha J, Wang D, Yu Y, Yang Q, Niu X 2018, Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning, Electronic Commerce Research and Applications, .31, 24-39

Malekipirbazari, M., Aksakalli, V. 2015. Risk assessment in social lending via random forests. Expert Systems with Applications, 42(10), 4621-4631.

Mankiw, N. G. 1986. The allocation of credit and financial collapse. The Quarterly Journal of Economics, 101(3), 455-470.

Mueller Holger M and Yannelis Constantine (2019), The rise in Student Loan defaults, Journal of Financial Economics, 131(1) 1-19

Ohlson James A 1980Financial Ratios and the Probabilistic Prediction of Bankruptcy, Journal of Accounting Research,. 18 (1) 109-131

Paisittanand Sineenad , Olson David L (2006), A simulation study of IT outsourcing in the credit card business, European Journal of Operational Research, 175 (2) 1248-1261

Prado J W, Alcantara VC, Carvalho F M, Vieira K C Machado L K C, Tonelli DF (2016), Multivariate analysis of credit risk and bankruptcy research data: a bibliometric study involving different knowledge fields (1968–2014), Scientometrics 106, 1007–1029

Robb, A., Robinson, D. T. 2018. Testing for racial bias in business credit scores. Small Business Economics, 50(3), 429-443.

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.

Song Wenda and Zhang Haiyang 2020 Monetary Policy and Borrowers' Loan Defaults: Research Based on Data from Renrendai, *China & World Economy 28(1) 94–121,*

Sun, J., & Li, H. 2012. Financial distress prediction using support vector machines: Ensemble vs. individual. Applied Soft Computing, 12(8), 2254-2265.

Sundararajan, Mukund, and Amir Najmi 2019. "The many Shapley values for model explanation." arXiv preprint arXiv:1908.08474 .

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

Vieira José RC, Flavio B, Vinicius AS, HK 2019, Machine learning models for credit analysis improvements: Predicting low-income families' default, Applied Soft Computing 83, 1-14

Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. 2018. A novel behavioural scoring model for estimating probability of default over time in peer-to-peer lending. Electronic Commerce Research and Applications, 27, 74-82.

Yu, L., Wang, S., & Lai, K. K. 2008. Credit risk assessment with a multistage neural network ensemble learning approach. Expert systems with applications, 34(2), 1434-1444.

Zavgren Christine V 1985, Assessing the Vulnerability to failure of American Industrial Firms: A Logistic Analysis, Journal of Business Finance and Accounting 12(1), pp 306-686

Zhang, Y., Wang, D., Chen, Y., Shang, H., & Tian, Q. 2017,. Credit risk assessment based on long short-term memory model. In International conference on intelligent computing, Springer, Cham,700-712

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. 2019. A study on predicting loan default based on the random forest algorithm. Procedia Computer Science, 162, 503-513.

Zliobaite, I. 2017. Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 31(4), 1060-1089.

Tables

Table 1: Distribution of Observations into Test and Training Sets			
	Default	Non-Default	Total
Training Set	5784 (22.29%)	14971 (57.71%)	20755 (80%)
Testing Set	1457 (5.62%)	3732 (14.38%)	5189 (20%)
Training Set I	5784 (27.87%)	14971 (72.13%)	20755 (100%)
Training Set II (minority under-sampling)	5784 (50%)	5784 (50%)	11568 (100%
Training Set II (SMOTE)	14971 (50%)	14971 (50%)	29942 (100%)

Table 2: Wald's Chi Squared Test for Feature Selection		
Variable	Wald's Statistic for single variables (p-value in Bracket)	Wald's Statistic for group (p-value in Bracket)
Loan Limit	812.5355 (0.00)***	570.0976 (0.00)***
Parental Income	48.3495 (0.00)***	63.73547 (0.00)***
Interest Rate	607.1732 (0.00)***	932.2489 (0.00)***
Degree Type	88.7456 (0.00)***	28.513 (0.00)***
Quality of Institution	20.1376 (0.00)***	1.7616 -0.62
Gender	0.5583 -0.45	0.50637 -0.48
Caste	74.4104 (0.00)***	33.3368 (0.00)***
Area Type	53.8003 (0.00)***	26.0905 (0.00)***
Courses	195.7074 (0.00)***	69.0871 (0.00)***
Unemployment Rate	780.3602 (0.00)***	4.0718 (0.04)*
Real GDP Growth Rate	528.9579 (0.00)***	175.0264 (0.00)***
Money Supply Growtht-1 Rate (M3)	495.4205 (0.00.00)***	0.0593 -0.81
Inflation (CPI)	419.1581 (0.00)***	26.5923 (0.00)***
Gross Capital Formationt-1 (% of GDP)	533.927 (0.00)***	102.6093 (0.00)***
***Significant at 99.9% **Significant at 99% *Significant at 95% + Significant at 90%		

Table 2A: Box Tidwell Test	
Variable	Coefficient
Loan Limit (Scaled)	-6.7981 (0.288)***
Loan Limit (Scaled) X Log(Loan Limit (Scaled))	12.7474 (0.744)***
Parental Income (Scaled)	-0.416 (0.111)***
Parental Income (Scaled) X Log(Parental Income (Scaled))	3.6415 (0.865)***
Interest Rate on Loan (Scaled)	-2.0416 (0.100)***
Interest Rate on Loan (Scaled) X Log(Interest Rate on Loan (Scaled))	5.0046 (0.351)***
Growth Rates (Scaled)	0.9974 (0.072)***
Growth Rates (Scaled) X Log(Growth Rates (Scaled))	0.0379 (0.271)
Unemployment (Scaled)	1.4913 (0.062)***
Unemployment (Scaled) X Log(Unemployment (Scaled))	0.6705 (0.154)***
Capital Formation (Scaled)	-2.7760 (0.220)***
Capital Formation (Scaled) X Log(Capital Formation (Scaled))	0.6193 (0.420)
CPI (Scaled)	-2.1537 (0.082)***
CPI (Scaled) X Log(CPI (Scaled))	2.5057 (0.180)***
***Significant at 99% Confidence Figures in bracket indicate Standard Errors	

Table 3: Regression Results		
Variables	Coefficients (Std. Errors)	p-value
Intercept	14.17 (0.737)***	0
Interest Rate	-49.13 (1.609)***	0
log(parental Income)	-0.07 (0.009)***	0
log(Loan Limit)	-0.6 (0.025)***	0
Undergraduate = 0; Postgraduate = 1	-0.23 (0.043)***	0
Male = 0; Female = 1	0.02 (0.032)	0.477
Unemployment Rate	30.61 (15.168)*	0.044
GDP Growth Rate	-22.99 (1.738)***	0
Money Supply Growth Rate$_{t-1}$	0.26 (1.055)	0.808
Inflation (Consumer Price Index)	-8.86 (1.718)***	0
Gross Capital Formation$_{t-1}$	-9.28 (0.916)***	0

College Tier (Base: Tier 3)		
Tier 1	0.01 (0.126)	0.927
Tier 2	-0.18 (0.165)	0.274
Tier 4	0.02 (0.033)	0.525
Caste (Base: Scheduled Tribe)		
Other Backward Castes	0.05 (0.035)	0.133
Scheduled Caste	0.3 (0.061)***	0
Scheduled Tribes	0.48 (0.151)**	0.002
Area Type (Base: Semi-Urban)		
Metropolitan	0.27 (0.056)***	0
Urban	0.05 (0.038)	0.175
Rural	0.01 (0.039)	0.954
Course (Base: Other courses)		
Engineering	0.11 (0.041)**	0.006
Medicine	-0.09 (0.074)	0.234
Management	0.1436 *** (0.0321)	0.0098
Law	-0.14 (0.381)	0.709
Nursing	-3.07 (0.717)***	0
Pharmacy	-1.95 (0.736)**	0.008
***Significant at 99.9% ** Significant at 99% *Significant at 95% $^{+}$ Significant at 90%		

Table 4: **Models using variables available while sanctioning the Loan (Experiment I)**

	Training set with original data distribution			Training set obtained by under-sampling majority class			Training set obtained by Synthetic Minority oversampling Technique		
Classifier	AUC	F-Measure	Bal. Acc.	AUC	F-Measure	Bal. Acc.	AUC	F-Measure	Bal. Acc.
Logistic Regression	65.62%	0.6916	61.53%	65.57%	0.6598	61.61%	65.46%	0.653	61.22%
Naïve Bayes	61.16%	0.5221	57.16%	60.86%	0.3236	53.62%	61.38%	0.2472	52.51%
MARS	66.87%	0.6911	61.94%	67.19%	0.6813	62.22%	73.00%	0.6978	65.04%
Multi-layer perceptron (1 hidden layer)	72.10%	0.7601	66.16%	69.47%	0.7273	63.85%	73.38%	0.7131	66.84%
Multi-layer perceptron (2 hidden layer)	73.84%	0.7672	66.78%	72.84%	0.7325	66.14%	73.32%	0.7041	66.11%
K- Nearest Neighbour	55.48%	0.6826	55.48%	55.89%	0.6853	55.89%	55.74%	0.6882	55.74%
Decision Tree	74.09%	0.781	68.42%	74.54%	0.7475	68.02%	74.77%	0.7265	67.03%
Support Vector Machine	65.56%	0.8045	58.22%	67.97%	0.6909	61.70%	68.80%	0.6917	62.68%
Random Forest	78.82%	0.7879	70.49%	77.85%	0.7496	69.91%	74.67%	0.7516	67.45%
Extreme Gradient Boosting	79.73%	0.8095	71.67%	78.42%	0.7505	70.05%	76.92%	0.7732	69.20%
Adaptive Boosting	78.86%	0.7748	71.00%	78.49%	0.7837	70.68%	78.11%	0.7316	69.69%
Light Gradient Boosting Machine	79.94%	0.8087	71.51%	78.54%	0.754	70.06%	77.14%	0.7662	69.61%
LSTM	76.55%	0.7215	69.42%	76.65%	0.643	64.74%	76.60%	0.6902	67.84%
1D- Convolutional Neural Network	74.33%	0.7597	67.89%	73.27%	0.6986	65.64%	74.95%	0.7042	66.50%

Table 5: **Models using all Variables including Macro-Economic (Experiment II)**									
	Training set with original data distribution			Training set obtained by under-sampling majority class			Training set obtained by Synthetic Minority oversampling Technique		
Classifier	AUC	F-Measure	Bal. Acc.	AUC	F-Measure	Bal. Acc.	AUC	F-Measure	Bal. Acc.
Logistic Regression	70.25%	0.7275	64.59%	70.21%	0.6869	64.52%	70.14%	0.6862	64.56%
Naïve Bayes	66.44%	0.6544	61.22%	66.10%	0.5709	60.10%	65.60%	0.3474	54.59%
MARS	70.52%	0.705	64.36%	70.99%	0.6952	64.46%	75.86%	0.728	67.90%
Multi-layer perceptron (1 hidden layer)	75.07%	0.7889	68.34%	75.56%	0.7499	68.12%	75.64%	0.7272	67.72%
Multi-layer perceptron (2 hidden layer)	75.88%	0.7885	67.82%	76.14%	0.726	68.22%	75.89%	0.7225	67.64%
K- Nearest Neighbour	60.42%	0.7039	60.42%	59.97%	0.6998	59.97%	60.12%	0.6996	60.12%
Decision Tree	74.83%	0.7384	68.01%	74.96%	0.7536	68.07%	77.68%	0.6959	68.56%
Support Vector Machine	70.71%	0.8212	64.92%	73.43%	0.7235	65.86%	73.36%	0.7243	66.10%
Random Forest	81.75%	0.8069	72.96%	80.92%	0.7687	72.24%	79.13%	0.7701	70.91%
Extreme Gradient Boosting	83.68%	0.8254	73.68%	82.51%	0.7701	72.23%	81.80%	0.7883	69.99%
Adaptive Boosting	82.18%	0.7933	72.49%	80.86%	0.8003	72.34%	81.09%	0.7572	72.70%
Light Gradient Boosting Machine	84.02%	0.8165	72.79%	83.92%	0.7696	72.46%	83.61%	0.7833	71.09%
LSTM	74.24%	0.6749	60.12%	69.84%	0.5644	58.07%	73.85%	0.5601	57.84%
1D- Convolutional Neural Network	76.92%	0.7941	69.14%	76.08%	0.7409	67.98%	76.37%	0.7376	68.17%

Table 6: 10-Fold CV to Compare Classifiers in Each Experiment with the Best Performing Classifier in terms of Balanced Accuracy												
Classifier	Variables known while sanctioning the Loan (Experiment I)						All Variables (Experiment II)					
	Mean AUC	T-test	Mean Balanced Accuracy	T-test	Mean F-Measure	T-test	Mean AUC	T-test	Mean Balanced Accuracy	T-test	Mean F-Measure	T-test
Logistic Regression	65.95%	-31.8286 (0.00)	64.94%	-5.94 (0.0002)	68.95%	-13.85 (0)	72.25%	-41.568 (0.00)	64.94%	-5.94 (0.0002)	72.52%	-8.95 (0)
Naïve Bayes	62.95%	-33.9808 (0.00)	61.89%	-8.55 (0)	52.11%	-27.34 (0)	69.38%	-43.2945 (0.00)	61.89%	-8.55 (0)	65.04%	-18.06 (0)
MARS	66.75%	-18.296 (0.00)	65.41%	-5.4 (0.0004)	69.57%	-9.84 (0)	73.41%	-19.7877 (0.00)	65.41%	-5.4(0.0004)	70.02%	-12.47 (0)
Multi-layer perceptron (1 Hidden Layer)	71.92%	-10.0147 (0.00)	68.14%	-3.03 (0.0143)***	75.37%	-8.01 (0)	77.33%	-10.613 (0.00)	68.14%	-3.03 (0.0143)***	77.72%	-4.3 (0.002)
Multi-layer perceptron (2 Hidden Layers)	72.88%	-10.2483 (0.00)	67.08%	-3.7 (0.0049)	77.42%	-3.84 (0.0039)	78.46%	-11.812 (0.00)	67.08%	-3.7 (0.0049)	78.93%	-4.87 (0.0009)
Support Vector Machine	65.35%	-30.9116 (0.00)	64.53%	-6.74 (0.0001)	81.31%	-	73.21%	-24.0933 (0.00)	64.53%	-6.74 (0.0001)	82.59%	-
Decision Tree	74.58%	-10.9077 (0.00)	68.75%	-2.06 (0.0696)***	77.58%	-2.92 (0.0169)***	78.13%	-20.8611 (0.00)	68.75%	-2.06 (0.0696)***	72.97%	-9.24 (0)
K-nearest Neighbour	55.49%	-45.8375 (0.00)	60.90%	-10.53 (0)	67.99%	-15.83 (0)	61.57%	-96.6122 (0.00)	60.90%	-10.53 (0)	70.37%	-10.91 (0)
Random Forests	78.90%	-2.3367 (0.0443)***	69.82%	-1.13 (0.2894)***	76.65%	-4.74 (0.0011)	83.84%	-1.8697 (0.0943)***	69.82%	-1.13 (0.2894)***	80.12%	-2.79(0.021)***
Extreme Gradient Boosting	79.17%	-1.9909 (0.0777)***	70.24%	-0.64 (0.5383)***	79.46%	-2.01 (0.0756)***	83.90%	-	70.24%	-0.64 (0.5383)***	80.12%	-3.99 (0.0032)
Adaptive Boosting	79.79%	-0.9041 (0.3895)***	70.74%	-0.11 (0.9145)***	75.93%	-7.23 (0)	82.94%	-1.4862 (0.1714)***	70.74%	-0.11 (0.9145)***	77.95%	-3.78 (0.0044)
Light GBM	80.52%	-	70.86%	-	80.06%	-1.28 (0.2315)***	83.38%	0.7908 (0.4494)***	70.86%	-	80.07%	-2 (0.077)***
LSTM	77.65%	-5.615 (0.0003)	54.27%	-17.94 (0)	64.23%	-12.47 (0)	76.24%	-14.7758 (0.00)	54.27%	-17.94 (0)	64.86%	-30.37 (0)
1D - CNN	74.14%	-8.5114 (0.00)	66.45%	-4.32 (0.0019)	73.12%	-7.47 (0)	79.12%	-10.1797 (0.00)	66.45%	-4.32 (0.0019)	77.84%	-5.54 (0.0004)
*** Mean AUC is not statistically different from the Mean AUC ,Mean Balanced Accuracy is not statistically different from the mean Balanced Accuracy, Mean F-measure is not statistically different from the Mean F-measure, of the best classifier with 99% confidence. Figures in bracket indicate p-values. Results in table 7 are based on 10-fold CV on the dataset, results in table 4 and 5, on a randomized 80-20 split.												

Table 7: 10-Fold CV to Compare Classifiers in Each Experiment with the Best Performing Classifier				
Classifier	Mean AUC: variables known while sanctioning the Loan (Experiment I) **(%)**	Mean AUC : All Variables (Experiment II) **(%)**	Difference (AUC in Exp II , AUC in Exp I) (%)	Pairwise T-test.
Logit	65.95%	72.25%	6.30%	18.6543 (0.00)***
Naïve Bayes	62.95%	69.38%	6.43%	19.3742 (0.00)***
MARS	66.75%	73.41%	6.66%	10.2856 (0.00)***
Multi-layer perceptron (1 Hidden Layer)	71.92%	77.33%	5.40%	9.3878 (0.00)***
Multi-layer perceptron (2 Hidden Layers)	72.88%	78.46%	5.58%	9.3114 (0.00)***
Support Vector Machine	65.35%	73.21%	7.86%	4.7798 (0.00)***
Decision Tree	74.58%	78.13%	3.55%	9.3034 (0.00)***
K-nearest Neighbour	55.49%	61.57%	6.09%	5.0936 (0.00)***
Random Forests	78.90%	83.84%	4.94%	5.5256 (0.00)***
Extreme Gradient Boosting	79.17%	83.90%	4.73%	3.6317 (0.01)***
Adaptive Boosting	79.79%	82.94%	3.15%	2.79 (0.02)**
Light GBM	80.52%	83.38%	2.87%	-4.5394 (0.00)***
LSTM	77.65%	76.24%	-1.41%	4.4203 (0.00)***
1D - CNN	74.14%	79.12%	4.98%	5.9713 (0.00)***

***Significantly different at 99% Confidence; **Significantly different at 95% Confidence
Figures in bracket indicate p-values
Results in table 6 are based on 10-fold CV on the dataset, results in table 4 and 5, on a randomized 80-20 split.

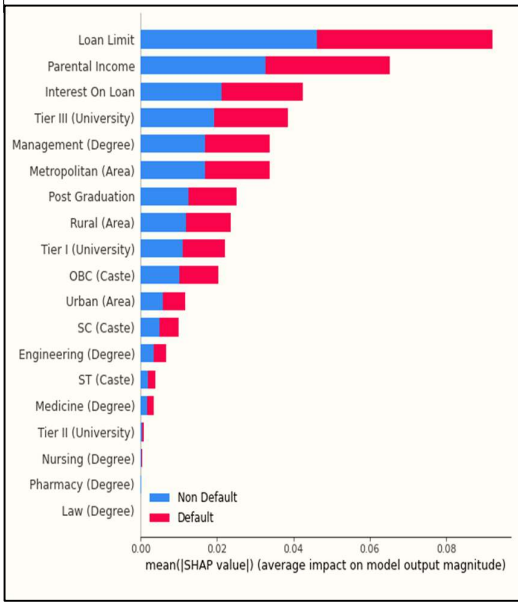
Table 8: 10-Fold CV to Compare Ensembles in with the Best Performing Heterogeneous Ensemble

Classifier	Variables known while sanctioning the Loan (Experiment I)		All Variables (Experiment II)	
Random Forests	Mean AUC (%)	T-test	Mean AUC(%)	T-test
Extreme Gradient Boosting	78.90	-5.3904 (0.0004)	81.84	-8.6883 (0.00)
Adaptive Boosting	79.17	-6.5438 (0.0001)	82.90	-8.2288 (0.00)
Light GBM	79.79	-5.3391 (0.0005)	81.94	-11.3841 (0.00)
Simple Avg.	80.52	-3.5307 (0.0064)	82.38	-8.7716 (0.00)
Weighted Avg.	81.31	-3.9982 (0.0031)	84.31	-4.8824 (0.0009)
Stacking	81.94	-5.4342 (0.0004)	84.94	-5.5292 (0.0004)
HCES	83.18	-1.2152 (0.2552)***	87.68	-1.7975 (0.1058)***
HCES-Bag	84.80	0.0559 (0.9567)***	87.48	-1.6128 (0.1412)***

***Mean AUC is not statistically different from the mean AUC of HCES-Bag with 99% confidence. Figures in bracket indicate p-values

Figure 5: Feature Importance from Shap Analysis for Tree Ensembles

Shap Summary Plot for Experiment I (For Tree Ensembles)



Shap Summary Plot for Experiment II (For Tree Ensembles)

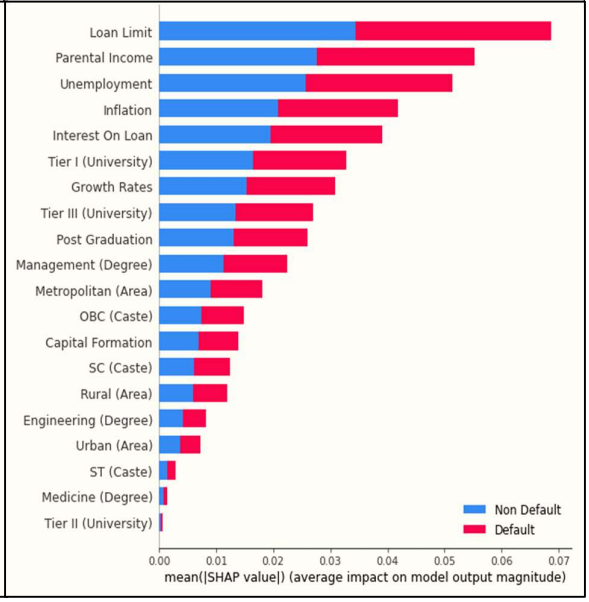


Table 9: Heterogeneous Ensembles (Experiment I and II)						
	AUC	F-Score	Bal. Acc.	Sensitivity	Specificity	Precision
Experiment I						
Simple Avg.	80.70%	0.8133	72.95%	69.11%	76.80%	86.43%
Weighted Avg.	81.64%	0.8127	72.94%	69.18%	76.69%	86.44%
Stacking	81.98%	0.8293	73.78%	66.45%	81.10%	84.85%
HCES	83.11%	0.8208	75.75%	74.12%	77.38%	87.38%
HCES-Bag	84.12%	0.8266	77.44%	77.64%	77.24%	88.89%
Experiment II						
Simple Avg.	82.31%	0.8344	75.40%	71.17%	79.64%	87.62%
Weighted Avg.	84.86%	0.8338	75.41%	71.31%	79.50%	87.65%
Stacking	84.43%	0.8453	76.11%	69.33%	82.90%	86.23%
HCES	87.67%	0.8436	77.61%	74.12%	81.10%	87.89%
HCES-Bag	87.69%	0.8559	80.38%	79.23%	81.52%	90.09%

Table 10: 10-Fold CV to Compare Ensembles in with the Best Performing Heterogeneous Ensemble

Classifier	Variables known while sanctioning the Loan (Experiment I)		All Variables (Experiment II)	
Random Forests	Mean AUC (%)	T-test	Mean AUC(%)	T-test
Extreme Gradient Boosting	78.90	-5.3904 (0.0004)	81.84	-8.6883 (0.00)
Adaptive Boosting	79.17	-6.5438 (0.0001)	82.90	-8.2288 (0.00)
Light GBM	79.79	-5.3391 (0.0005)	81.94	-11.3841 (0.00)
Simple Avg.	80.52	-3.5307 (0.0064)	82.38	-8.7716 (0.00)
Weighted Avg.	81.31	-3.9982 (0.0031)	84.31	-4.8824 (0.0009)
Stacking	81.94	-5.4342 (0.0004)	84.94	-5.5292 (0.0004)
HCES	83.18	-1.2152 (0.2552)***	87.68	-1.7975 (0.1058)***
HCES-Bag	84.80	0.0559 (0.9567)***	87.48	-1.6128 (0.1412)***

***Mean AUC is not statistically different from the mean AUC of HCES-Bag with 99% confidence. Figures in bracket indicate p-values

Table 11 A: (Experiment I) Default Flips (observations reclassified as default compared to Logistic Regression)							
	Overall	All Backward Castes		Rural Areas		Low Income	
	Mean % Reclassification (A)	Mean % Reclassification (B1)	Paired T-Test (A vs B1)	Mean % Reclassification (B2)	Paired T-Test (A vs B2)	Mean % Reclassification (B3)	Paired T-Test (A vs B3)
Multi-layer perceptron (1 Hidden Layer)	6.53%	8.40%	5.4738 (0.0004)***	6.25%	-0.7298 (0.4841)	4.45%	-4.1429 (0.0025)***
Multi-layer perceptron (2 Hidden Layers)	6.93%	7.55%	1.0647 (0.3147)	6.22%	-1.3713 (0.2035)	3.64%	-10.5621 (0.0)***
Decision Tree	7.15%	5.84%	-5.3013 (0.0005)***	8.14%	2.2544 (0.0506)	2.94%	-14.9627 (0.0)***
Random Forests	9.84%	11.51%	6.938 (0.0001)***	9.86%	0.0487 (0.9622)	7.04%	-14.9376 (0.0)***
Extreme Gradient Boosting	7.45%	7.62%	0.9267 (0.3783)	7.65%	0.7753 (0.4581)	4.10%	-10.0723 (0.0)***
Adaptive Boosting	11.38%	11.09%	-1.7424 (0.1154)	11.32%	-0.1724 (0.8669)	5.89%	-15.3442 (0.0)***
Light GBM	7.57%	7.66%	0.6014 (0.5624)	7.01%	-1.6458 (0.1342)	4.23%	-9.9165 (0.0)***
LSTM	23.62%	25.18%	1.1433 (0.2824)	22.28%	-0.6484 (0.5329)	16.31%	-2.1035 (0.0648)
1D - CNN	8.91%	7.98%	-1.0735 (0.311)	9.57%	0.9347 (0.3744)	6.19%	-3.4607 (0.0072)***
	Non-Default Flips (observations reclassified as non-default compared to Logistic Regression)						
	Overall	All Backward Castes		Rural Areas		Low Income	
	Mean % Reclassification (A)	Mean % Reclassification (B1)	Paired T-Test (A vs B1)	Mean % Reclassification (B2)	Paired T-Test (A vs B2)	Mean % Reclassification (B3)	Paired T-Test (A vs B3)
Multi-layer perceptron (1 Hidden Layer)	14.40%	11.09%	-7.6776 (0.0)***	13.61%	-1.4536 (0.18)	18.00%	5.5672 (0.0003)***
Multi-layer perceptron (2 Hidden Layers)	14.80%	11.16%	-11.0616 (0.0)***	13.62%	-2.6665 (0.0258)**	18.67%	7.8163 (0.0)***
Decision Tree	20.62%	16.93%	-11.1245 (0.0)***	15.63%	-9.0084 (0.0)***	26.96%	10.0085 (0.0)***
Random Forests	17.98%	13.96%	-14.8707 (0.0)***	16.11%	-4.3803 (0.0018)**	22.10%	6.7912 (0.0001)***
Extreme Gradient Boosting	18.11%	14.51%	-15.7191 (0.0)***	15.79%	-5.7693 (0.0003)***	23.88%	9.1663 (0.0)***
Adaptive Boosting	15.17%	12.31%	-12.4313 (0.0)***	13.87%	-3.2722 (0.0096)**	19.87%	7.3784 (0.0)***
Light GBM	17.98%	14.10%	-13.9841 (0.0)***	15.93%	-5.027 (0.0007)***	23.52%	8.3719 (0.0)***

LSTM	22.11%	19.37%	-2.1159 (0.0635)	19.52%	-3.0113 (0.0147)**	27.41%	1.6383 (0.1358)
1D - CNN	13.97%	13.25%	-0.64 (0.5381)	13.14%	-0.7519 (0.4713)	18.39%	7.7078 (0.0)***

***Significantly different at 99% Confidence; **Significantly different at 95% Confidence
Figures in bracket indicate p-values
All Backward castes indicate the union of Scheduled Castes(SC), scheduled tribes(ST) and Other Backward Castes (OBC); Low Income indicates those in bottom 20 Percentile of Parental Income

Table 11B: (Experiment II) :Default Flips (observations reclassified as default compared to Logistic Regression)

	Overall	All Backward Castes		Rural Areas		Low Income	
	Mean % Reclassification (A)	Mean % Reclassification (B1)	Paired T-Test (A vs B1)	Mean % Reclassification (B2)	Paired T-Test (A vs B2)	Mean % Reclassification (B3)	Paired T-Test (A vs B3)
Multi-layer perceptron (1 Hidden Layer)	6.69%	8.95%	4.7217 (0.0011)***	6.79%	0.2857 (0.7816)	8.12%	4.4678 (0.0016)***
Multi-layer perceptron (2 Hidden Layers)	7.30%	7.65%	0.8396 (0.4229)	6.48%	-1.6966, (0.124)	6.13%	-3.0533 (0.0137)**
Decision Tree	9.86%	6.93%	-6.204 (0.000)***	8.30%	-6.1688 (0.0002)***	6.04%	-7.9983 (0.0)***
Random Forests	8.78%	9.40%	2.5768 (0.0299) **	9.00%	0.8832 (0.4001)	9.73%	3.4386 (0.0074)***
Extreme Gradient Boosting	5.92%	5.34%	-5.5084 (0.0004)***	5.39%	-2.2587 (0.0503)	6.48%	1.9596 (0.0817)
Adaptive Boosting	7.81%	5.83%	-17.7843 (0.0)***	7.43%	-1.3806 (0.2007)	7.14%	-2.5087 (0.0334)**
Light GBM	7.28%	7.92%	3.1622 (0.0115)**	6.48%	-2.582 (0.0296)**	8.12%	3.1333 (0.0121)**
LSTM	26.59%	28.43%	0.8851 (0.3991)	27.47%	0.4516 (0.6622)	24.22%	-0.9414 (0.3711)
1D - CNN	3.71%	2.58%	-3.2924 (0.0093)***	2.33%	-4.3901 (0.0017)***	2.18%	-3.2421 (0.0101)**
	Non-Default Flips (observations reclassified as non-default compared to Logistic Regression)						
	Overall	All Backward Castes		Rural Areas		Low Income	
	Mean % Reclassification (A)	Mean % Reclassification (B1)	Paired T-Test (A vs B1)	Mean % Reclassification (B2)	Paired T-Test (A vs B2)	Mean % Reclassification (B3)	Paired T-Test (A vs B3)
Multi-layer perceptron (1 Hidden Layer)	14.25%	10.56%	-11.9465 (0.0)***	14.77%	1.3366, (0.2142)	15.96%	3.7304, (0.0047)***
Multi-layer perceptron (2 Hidden Layers)	13.26%	9.74%	-14.1709 (0.0)***	13.75%	1.3098, (0.2227)	16.55%	13.0131 (0.0)***
Decision Tree	14.26%	13.04%	-4.2607 (0.0021) ***	13.57%	-1.5993 (0.1442)	18.06%	(6.5881 0.0001)***
Random Forests	15.30%	11.99%	-14.0011 (0.0)***	15.40%	0.2149 (0.8346)	17.25%	4.6251 (0.0012)***

Extreme Gradient Boosting	15.31%	12.26%	-14.6225 (0.0)***	15.07%	-0.5704 (0.5824)	16.57%	3.3612 (0.0084)***
Adaptive Boosting	12.16%	10.24%	-10.0424 (0.0)***	11.86%	-0.8858 (0.3988)	14.10%	6.4831 (0.0001)***
Light GBM	15.57%	12.41%	-13.8641 (0.0)***	15.71%	0.3161 (0.7591)	18.17%	7.9184 (0.0)***
LSTM	20.24%	17.62%	-4.8796 (0.0009)***	18.16%	-1.8335 (0.0999)	22.79%	(2.3419 0.0439)
1D - CNN	15.17%	14.00%	-3.3021 (0.0092)***	16.53%	2.4632 (0.036)**	19.85%	8.4775 (0.0)***

***Significantly different at 99% Confidence; **Significantly different at 95% Confidence
Figures in bracket indicate p-values
All Backward castes indicate the union of Scheduled Castes(SC), scheduled tribes(ST) and Other Backward Castes (OBC); Low Income indicates those in bottom 20 Percentile of Parental Income

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 D			

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: Macroeconomic Variables

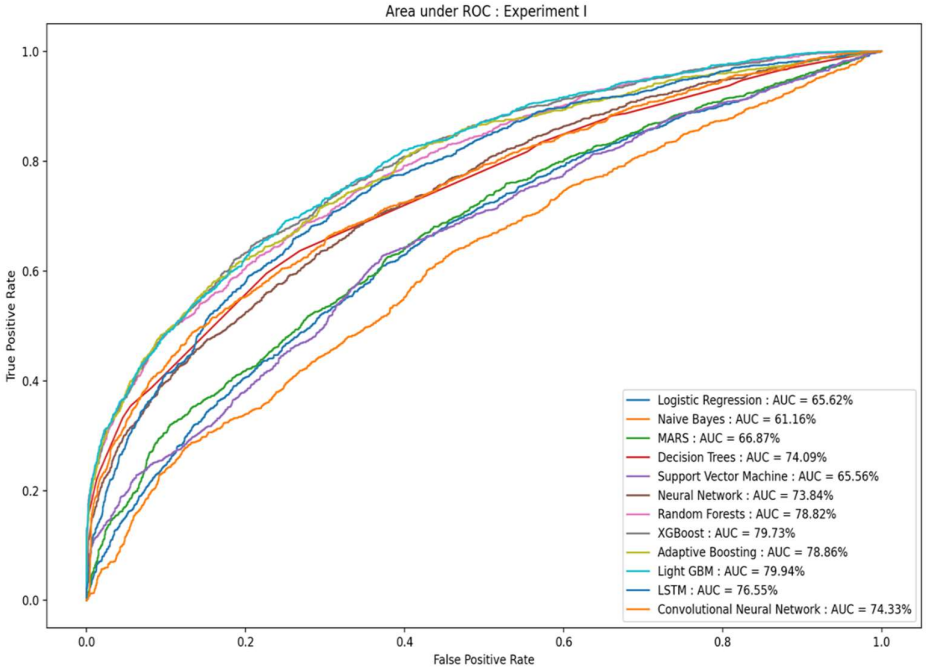
Year	GDP Growth Rate	Gross Capital Formation as a % of GDP (t-1)	CPI	Money Supply Growth Rates (t-1)	Unemployment Rates
Source	CSO	CSO	CSO	CSO	ILO
	8		3.4		4.31
2000-2001	4.15	26.97	3.7	16	3.775
2001-2002	5.39	24.21	4.3	16.1	4.316
2002-2003	3.88	25.65	4.1	13	3.929
2003-2004	7.97	25.02	3.8	14	3.889
2004-2005	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					

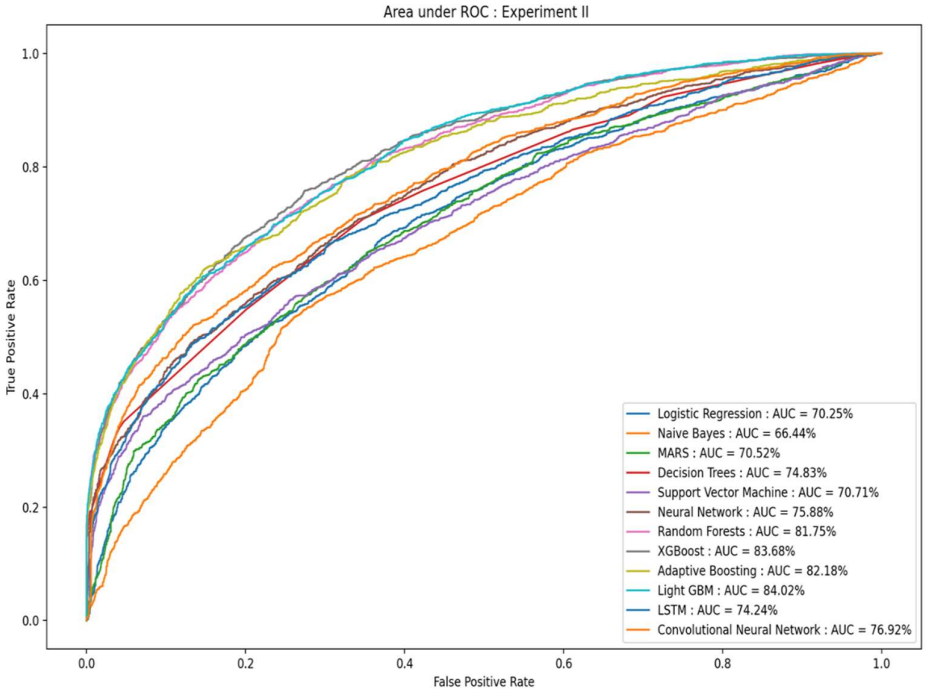
Appendix E(A): Sensitivity, Specificity and Precision for Models available while sanctioning the loan

Table EA: 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		
Classifier	Specificity	Sensitivity	Precision	Specificity	Sensitivity	Precision	Specificity	Sensitivity	Precision
Logistic Regression	62.46%	60.61%	80.53%	67.81%	55.41%	81.51%	67.88%	54.56%	81.31%
Naïve Bayes	75.63%	38.69%	80.27%	86.96%	20.28%	79.94%	90.39%	14.63%	79.59%
MARS	63.56%	60.32%	80.91%	65.89%	58.55%	81.47%	70.28%	59.81%	83.75%
Multi-layer perceptron (1 hidden layer)	61.91%	70.42%	82.56%	62.11%	65.59%	81.60%	72.27%	61.41%	85.01%
Multi-layer perceptron (2 hidden layer)	62.11%	71.44%	82.85%	67.06%	65.22%	83.53%	71.93%	60.29%	84.62%
K- Nearest Neighbour	48.80%	62.17%	75.67%	49.35%	62.43%	75.95%	48.46%	63.02%	75.80%
Decision Tree	63.69%	73.15%	83.77%	69.18%	66.85%	84.75%	70.42%	63.64%	84.64%
Support Vector Machine	31.02%	85.42%	76.03%	63.01%	60.40%	80.70%	65.34%	60.02%	81.60%
Random Forest	67.81%	73.18%	85.34%	73.71%	66.10%	86.56%	66.92%	67.98%	84.03%
Extreme Gradient Boosting	66.44%	76.90%	85.44%	73.92%	66.18%	86.67%	67.33%	71.06%	84.78%
Adaptive Boosting	71.79%	70.20%	86.44%	69.18%	72.19%	85.71%	76.39%	63.00%	87.24%
Light Gradient Boosting Machine	66.16%	76.85%	85.33%	73.30%	66.83%	86.51%	69.80%	69.43%	85.48%
LSTM	77.42%	61.41%	87.45%	78.04%	51.45%	85.71%	78.59%	57.10%	87.23%
1D- Convolutional Neural Network	66.51%	69.27%	84.12%	71.65%	59.62%	84.34%	72.89%	60.10%	85.03%

Appendix E(B): Sensitivity, Specificity and Precision for Models using all variables									
	Training set with original data distribution			Training set obtained by under-sampling majority class			Training set obtained by Synthetic Minority oversampling Technique		
Classifier	Specificity	Sensitivity	Precision	Specificity	Sensitivity	Precision	Specificity	Sensitivity	Precision
Logistic Regression	63.97%	65.22%	82.26%	70.76%	58.28%	83.62%	70.97%	58.15%	83.69%
Naïve Bayes	67.67%	54.77%	81.27%	76.60%	43.60%	82.67%	87.10%	22.08%	81.42%
MARS	67.33%	61.39%	82.80%	69.25%	59.67%	83.25%	72.41%	63.40%	85.48%
Multi-layer perceptron (1 hidden layer)	61.84%	74.84%	83.40%	68.98%	67.26%	84.74%	72.07%	63.37%	85.32%
Multi-layer perceptron (2 hidden layer)	60.54%	75.11%	82.98%	73.58%	62.86%	85.90%	72.68%	62.59%	85.44%
K- Nearest Neighbour	57.52%	63.32%	79.24%	57.10%	62.83%	78.96%	57.52%	62.73%	79.09%
Decision Tree	70.83%	65.19%	85.13%	68.15%	67.98%	84.54%	79.48%	57.64%	87.80%
Support Vector Machine	45.30%	84.54%	79.83%	67.95%	63.77%	83.60%	68.43%	63.77%	83.80%
Random Forest	70.49%	75.43%	86.75%	76.25%	68.22%	88.04%	72.48%	69.35%	86.58%
Extreme Gradient Boosting	68.43%	78.94%	86.49%	75.98%	68.49%	87.96%	66.37%	73.61%	84.86%
Adaptive Boosting	72.07%	72.91%	86.99%	70.21%	74.46%	86.49%	79.62%	65.78%	89.21%
Light Gradient Boosting Machine	67.95%	77.63%	86.12%	76.66%	68.25%	88.22%	70.35%	71.84%	86.12%
LSTM	61.70%	58.55%	79.66%	72.61%	43.52%	80.28%	72.61%	43.06%	80.11%
1D- Convolutional Neural Network	62.87%	75.40%	83.87%	70.28%	65.68%	84.99%	71.38%	64.95%	85.32%

Appendix F: ROC curves (From Table 4 and 5, Training set with original distribution)





Appendix G: Hyperparameters for Randomized Grid Search tuning

Classifier	HyperParameters	Choices
Decision Trees	Max Depth	3, 4, 5, 6,... 15
	Error Criteria	Entropy, Gini
	Pruning	Min Cost Pruning
Supper Vector Machine	Kernel	Radial, Polynomial, Linear
	L2 Regularization	1, 1.5, 2, 2.5, 3, 3.5, 4
	Degree for polynomial Kernel	2, 3, 4, 5
	Max Iterations	No Limit
Neural Networks	No. of Units in 1st hidden layer	4, 5, 6, 7, 8
	No. of Units in 2nd hidden layer	2, 3, 4
	Activation Function	ReLU, tanh
	Solver	Adam, Stochastic Gradient Descent
	Learning Rate	Adaptive, Inverse Scaling, Constant (0.1, 0.01, 0.001)
	Maximum Iterations	No Limit
Random Forests	Number of Trees	50, 100, 250, 500
	Error Criteria	Entropy, Gini
	Min samples for splitting an internal node	2, 3, 4,..10
	Min samples required at leaf node	1, 2, 3, 4, 5
Extreme Gradient Boosting	Number of Trees	50, 100, 250, 500
	Error Criteria	Mean Squared Error, Friedman MSE
	Learning Rates	0.1, 0.01, 0.001
	Min samples for splitting an internal node	2, 3, 4,..10
	Min samples required at leaf node	1, 2, 3, 4, 5
	Max Depth	3, 4,..10
Adaptive Boosting	Number of Trees	50, 100, 250, 500
	Error Criteria for base estimator	Entropy, Gini
	Max Depth for Base estimator	3, 4,..10
	Learning Rates	0.1, 0.01, 0.001
Light GBM	Max Depth	3, 4,..10, No Restriction
	Number of Leaves	10, 20, 30, No Restriction
	Number of Trees	50, 100, 250, 500
	Learning Rates	0.1, 0.01, 0.001
	Feature and Observation Subsampling	75%, 80%, 90%, All
	Number of hidden layers	1 or 2
	Nodes for input or hidden layers	32, 64, 128

Convolutional Neural Network	nKernels for input or hidden layers	1, 2, 3
	Activation Function	ReLU
	Dropout (Regularization) for Input or hidden layers	0.1, 0.2,..0.99
	Units in dense Layer	5, 10, 15, 20
	Loss function	Categorical Cross Entropy
	Optimizers	Adadelta, Adam
	nEpochs	5
LSTM	Number of Nodes	5, 6, 7,…15
	Activation Function	ReLU, tanh
	Dropout (Regularization) for input or hidden layers	0.1, 0.2,..0.9
	Loss function	Mean squared Error
	Optimizers	Adadelta, Adam
	nEpochs	5

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 and 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

Appendix JA: 10-Fold CV to Compare Classifiers in Each Experiment with the Best Performing Classifier in terms of Specificity				
Classifier	Variables known while sanctioning the Loan (Experiment I)		All Variables (Experiment II)	
	Mean Specificity	T-test	Mean Specificity	T-test
Logistic Regression	61.69%	-16.53(0)	63.72%	-10.68 (0)
Naïve Bayes	75.30%	-	67.48%	-4.47 (0.0015)
MARS	63.42%	-10.23(0)	66.54%	-3.53 (0.0064)
Multi-layer perceptron (1 Hidden Layer)	61.50%	-17.4(0)	62.17%	-9.87 (0)
Multi-layer perceptron (2 Hidden Layers)	63.50%	-13.27(0)	60.95%	-9.66 (0)
Support Vector Machine	30.29%	-39.01(0)	43.61%	-25.61 (0)
Decision Tree	63.72%	-16.34(0)	72.06%	-
K-nearest Neighbour	48.12%	-21.74(0)	56.95%	-12.75 (0)
Random Forests	64.60%	-9.7(0)	68.75%	-3.71 (0.0048)
Extreme Gradient Boosting	64.14%	-12.63(0)	64.83%	-8.8 (0)
Adaptive Boosting	67.60%	-5.8(0.0003)	69.87%	-1.74 (0.1155)***
Light GBM	63.02%	-13.58(0)	66.51%	-4.13 (0.0026)
LSTM	61.84%	-10.96(0)	50.19%	-21.22 (0)
1D - CNN	60.83%	-16.6(0)	61.64%	-7.81 (0)
***Mean Specificity is not statistically different from the mean Specificity of the best classifier with 99% confidence. Figures in bracket indicate p-values. Results in table JA are based on 10-fold CV on the dataset, results in tables EA and EB, on a randomized 80-20 split.				

Appendix JB: 10-Fold CV to Compare Classifiers in Each Experiment with the Best Performing Classifier in terms of Sensitivity				
Classifier	Variables known while sanctioning the Loan (Experiment I)		All Variables (Experiment II)	
	Mean Sensitivity	T-test	Mean Sensitivity	T-test
Logistic Regression	60.65%	-29.42 (0)	66.41%	-32.45 (0)
Naïve Bayes	39.07%	-40.36 (0)	54.38%	-28.15 (0)
MARS	59.84%	-21.96 (0)	59.71%	-39.43 (0)
Multi-layer perceptron (1 Hidden Layer)	71.97%	-22.81 (0)	74.85%	-8.62 (0)
Multi-layer perceptron (2 Hidden Layers)	71.23%	-15.36 (0)	76.06%	-9.26 (0)
Support Vector Machine	86.33%	-	84.41%	-
Decision Tree	71.80%	-16.81 (0)	66.86%	-24.61 (0)
K-nearest Neighbour	62.56%	-16.52 (0)	63.72%	-32.03 (0)

Random Forests	72.27%	-21.36 (0)	75.19%	-14.95 (0)
Extreme Gradient Boosting	75.27%	-24.6 (0)	77.23%	-11.06 (0)
Adaptive Boosting	69.03%	-17.8 (0)	71.68%	-9.89 (0)
Light GBM	75.34%	-13.91 (0)	78.07%	-8.79 (0)
LSTM	53.86%	-62.32 (0)	58.63%	-22.04 (0)
1D - CNN	68.10%	-21.71 (0)	72.67%	-14.7 (0)

***Mean Sensitivity is not statistically different from the mean Sensitivity of the best classifier with 99% confidence.
Figures in bracket indicate p-values.
Results in table JB are based on 10-fold CV on the dataset, results in tables EA and EB, on a randomized 80-20 split.

Appendix JC: 10-Fold CV to Compare Classifiers in Each Experiment with the Best Performing Classifier in terms of Precision

Classifier	Variables known while sanctioning the Loan (Experiment I)		All Variables (Experiment II)	
	Mean Precision	T-test	Mean Precision	T-test
Logistic Regression	81.09%	-3.47 (0.007)	83.46%	-2.32 (0.0458)***
Naïve Bayes	79.31%	-7.94 (0)	81.28%	-4.07 (0.0028)
MARS	81.42%	-3.87 (0.0038)	81.96%	-3.32 (0.0089)
Multi-layer perceptron (1 Hidden Layer)	82.70%	-2.8 (0.0206)***	83.51%	-1.92 (0.0873)***
Multi-layer perceptron (2 Hidden Layers)	82.12%	-3.11 (0.0125)***	83.43%	-2.52 (0.0326)***
Support Vector Machine	76.47%	-8.59 (0)	79.87%	-4.33 (0.0019)
Decision Tree	83.70%	-1.34 (0.2131)***	85.29%	-0.38 (0.7154)***
K-nearest Neighbour	74.37%	-18.66 (0)	80.43%	-4.02 (0.003)
Random Forests	83.11%	-1.49 (0.1695)***	84.72%	-1.63 (0.1381)***
Extreme Gradient Boosting	83.80%	-1.57 (0.1504)***	85.78%	-
Adaptive Boosting	85.01%	-	84.57%	-1.04 (0.3257)***
Light GBM	84.85%	-0.86 (0.5194)***	85.30%	-0.57 (0.5855)***
LSTM	77.88%	-6.12 (0.0002)	74.55%	-13.99(0)
1D - CNN	82.26%	-2.93 (0.0169)***	82.22%	-2.66 (0.0259)***

***Mean Precision is not statistically different from the mean Precision of the best classifier with 99% confidence.
Figures in bracket indicate p-values.
Results in table JC are based on 10-fold CV on the dataset, results in tables EA and EB, on a randomized 80-20 split.