National College of Ireland

Implementation of Machine Learning Techniques to Predict Post-Collegiate Earnings & Student-Loan Repayment Prospects

MSc Research Project
Data Analytics

Vishnu Mohan Kariyedath
x16104188

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Vishnu Mohan Kariyedath
Student ID:	x16104188
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Dr. Muhammad Iqbal
Submission Due Date:	11/12/2017
Project Title:	Implementation of Machine Learning Techniques to Predict Post-Collegiate Earnings & Student-Loan Repayment Prospects
Word Count:	5950

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	11th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Implementation of Machine Learning Techniques to Predict Post-Collegiate Earnings & Student-Loan Repayment Prospects

Vishnu Mohan Kariyedath
x16104188
MSc Research Project in Data Analytics

11th December 2017

Abstract

In this age of self education and personalized unconventional learning, majority of people still see a college degree as a gateway to job the workforce and to enter their field of interest and this outlook does not seem to be disappearing. However, the burden of rising college tuition and increasing student-loan default rates raises concerns. A common belief that only prestigious, elite and expensive institutions can bring about attractive work life, is being debunked. There exists many institutions that do not belong the ivy league and does not have huge tuitions, but, their graduates end up quite successful. Due to the lack of relevant portals and resources that provide such data, it would make more sense to introduce such a practices in the market. This research paper presents a possible solution to this gap, by filling it with a prediction model that takes data from College Scorecard (US Department of education, US Census Bureau, etc) and provides estimation of possible future earning and their chances with student-loan debt repayment. The research makes use of multiple regression, Random Forest, XGBoost and Artificial Neural Networks to achieve this goal.

1 Introduction

1.1 Domain Overview

College education is to this day, for the majority of general public, a qualifier for a successful career and there-after a wholesome life. Even-though, such norms are being challenged by many, an overwhelming majority still prefers to enroll for a bachelor's or/and post-graduate's degree, in-order to get into their desired domain of work. This is further encouraged by institutions with their hiring practices. Thus, choosing the right institution for their education is something that is not to be taken lightly.

Educational institutions choose qualified students based on many metrics of merit such as their SAT/ACT scores, respective GPAs, GMAT/GRE scores, extra curricular achievements, etc. However, the reasons why students choose their preferred institution can be quite broad and often varies person to person. According to Baum and O'Malley (2003), the two main considerations for applicants are cost of education and future job

prospects after graduating from a particular institution. Although, there are numerous online and offline sources that offer students access to information about different institutions, potential post-collegiate earnings and student-loan repayment prospects are almost always unavailable or only sparsely available, that too for a hand-full of institutions, Duncan (2015).

College Scorecard was introduced in 2010 by the US Department of Education inorder to fill the deficit of information regarding post-collegiate earnings and student-loan repayment prospects, Duncan (2015). Being a federal government initiative, it has better access to data from educational institutions across the country. However, the data depository in the form that it is, right now, has many defects and can be regarded as in its juvenile stage, intended to be useful in the following decades.

1.2 Motivation

Applicants from all around the globe applies to the thousands institutions in the United States and they make use of any & all resources available to them. The fact that College Scorecard has received so much attention, despite the presence of numerous other portals Hurwitz and Smith (2016), signals the need for better metrics to assist college selection. The trend of increased delinquency and default rates for student-loans is becoming a national crisis in the USA & UK. Student-loan borrowers (as seen in figure 1) are more likely to be delinquent on their loans than any other borrowers, claims Ryan Gorman, Business Insider UK (2015).

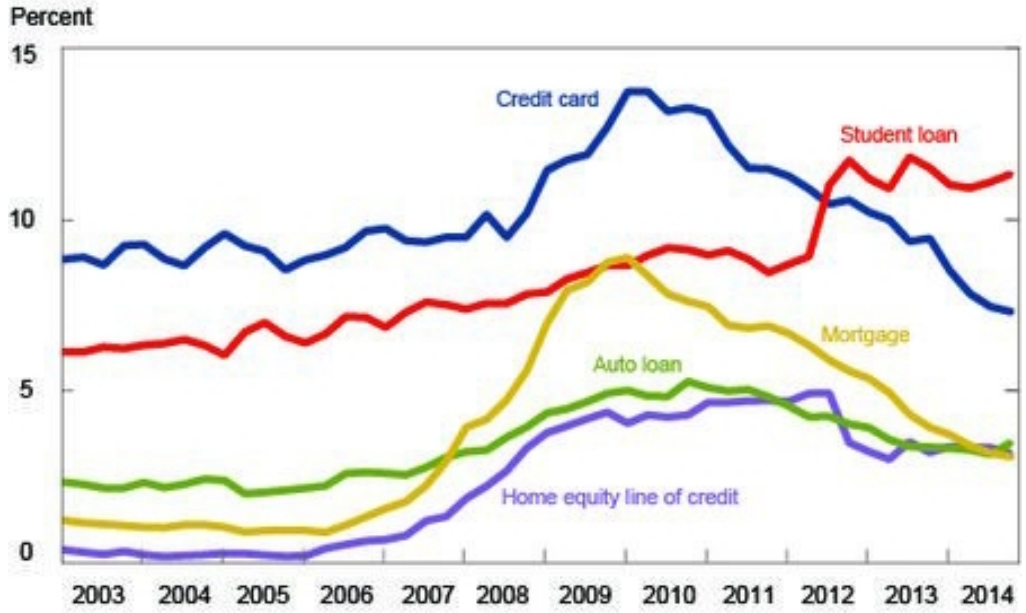


Figure 1: Percentage of over-90-days delinquent borrowers
Source: Ryan Gorman, Business Insider UK (2015)

Building a model that, gives an estimation of how much one could expect to earn & the probability for successfully repaying their student-loans after graduating from an institution, will significantly help applicants make better financial decisions, there-by, reducing their odds of failing on their education loans. Helping students make better decisions and there by make a difference in their lives, qualifies as the motivation for this project.

1.3 Research Objective

An objective could only be achieved when it is defined with crystal clarity, therefore, it is vital that the objective of the research be well declared. The research project is not intended to provide students with a portal to compare institutions to their profile, neither does it have such interfacing capabilities. The research project is supposed to provide better understanding of what could be achieved with educational institutions (College Scorecard) accumulated data and what are the best practices when dealing with such data & modelling techniques. Although, the researcher does not wish to critic on the College Scorecard depository, a few recommendations may be put forward.

How well could machine learning techniques be employed, to draw insights from the College Scorecard data and, to estimate potential post-collegiate earnings & their education loan payoff probability for an improved college selection process?

Hypothesis: The quality of an educational institution and entire college experience cannot be reduced to the numerical insights derived from a couple of attributes or so. The hiring practices of companies and remuneration offered are greatly influenced by socio-economic climate, technological circumstances, potential of the individual, etc. The insights provided my the research can only augment the statistical significance of the events and does not assure anything. For the scope of this research, reasonable economic stability, trivial turmoil & circumstantial continuity is assumed.

This report contains a modest review of literature that sheds light on various topics involved in building the models used in the research and, practices adopted while working with above mentioned data. The methodology describes the data, design and the intricate process followed while building this project, while, the implementation examines each model individually. Results & findings are discussed in the evaluation and conclusion are driven, to be noted down there-after.

2 Review of Literature

2.1 Domain & Related Works

As a direct result of federal educational aid programs & student-loan liberalization policies implemented during the 1980s to 1990s, higher education in the US became more accessible, Gladieux and Perna (2005). These policies along with globalization, led to an increased demand for college seats and was reflected in the trend of increasing college tuition and in the inception of many for-profit college across the nation Morey (2004). Due to this change in dynamics, many studies were conducted at the time to investigate the relationship between colleges and graduate earnings, however, most of them were highly biased and worked on limited parameters.

Brewer et al. (1999) made use of individual, family and college-specific data to draw relationship between future earnings and graduating institutions. The study was flawed by the over-generalization of college types and their respective quality into six groups, which seems insufficient to capture college diversity. With similar goals James et al. (1989) attempted at a larger study, which followed an incremental approach to collecting attributes, adding them into the model whenever they became available and thereby making the model ever more flexible. However, the study was intentional biased by taking in data on male students, thus making it obsolete in co-ed environment. Studies of Wachtel (1976) and later on Berger (1988) build regression models that took into account

additional details like tuition, study major, SAT scores, research funding received by the institution, family income, etc. The idea of calculating expenditure by time spend in college is quite unique and commendable, since, a simple money value would not be able to accommodate a student's dedication to the program. Such, a metric could not be included in this research project due to unavailability of the specific data. Berger (1988) and Wachtel (1976) were able to collect detailed data because they only took into account a few ivy-league institutions, however, this research encompasses a much broader net of colleges.

A general lack of right kind & amount of data seemed to have negatively influenced all the above mentioned studies, this issue could be reduced to an extend in this research project. Although, College Scorecard had received its fair share of attention, most of the studies that used the data had settled for knowledge discovery from mere visualizations of available data rather than building predictive models as in this research. Goodman et al. (2017) compared enrolment & graduation rates within different demographics against various other attributes to visualize & identify key drivers of improved college access. A glossary of visualizations depicting the general trends in higher education and its benefits to society are explained in the study by Ma et al. (2016). Neither of the studies nor others of similar papers provided much useful insights with handling the data, as they all resorted to elimination of entries with missing values and did no predictive analytics, but, were useful in initial data exploration.

2.2 Multiple Regression

No regression model is complete without a classic multiple regression techniques implemented as benchmark for comparison or in its optimized format, and effective prediction model of its own. A multiple regression model is highly influenced by correlations between variables, Aiken et al. (1991), as it tries to identify function based on the influence or the nearest possible relationship between the independent variables and the continuous dependent variables. Cohen et al. (2013)'s book serves as a go-to text for exploring the non-mathematically applied data analysis over a stretch of real world examples with comprehensive illustrations with the help of graphical methods. The implementation of regression models in the prediction of earnings of baseball players, in Cohen et al. (2013), makes its domain quite relatable to the research project due to the quality & features to remuneration dynamics. Though the models of both editions were well thought-out, the fact that only the peripheral relationships might have been taken into account here, seemed to be a drawback. By switching from simple linear or nearest relationship functions to more comprehensive in-depth functions could results in better prediction accuracy & lower model interpretability, as explained in Preacher et al. (2006). Optimized models such as lasso, ridge, weighted, recursive, etc. were used instead of off-the-self linear regression models and results were compared to show, how well the optimized models performed in comparison.

2.3 Random Forest

Simple decision trees design their model by branching-out according to their individually analyzed possible outcomes, however, for achieving low variance & avoiding over-fitting issues, Random Forest models, try averaging multiple decision trees, Breiman (2001). Such an ensemble technique capable of achieving far better predictive accuracy, was a

good pick for the research project. The name Random Forest, explains a lot about the model, since it is a collection of trees built on bagging up randomly selected variables from the list all available variables and allowed to branch down without pruning, in-order to fit them into regression models. Finally, all the trees or the best trees (forest) in the model, are averaged to get the optimal prediction. According to Segal (2004), Random Forests can be employed in supervised learning with both classification & regression models and, usually offer much better results for classification. However, from Liaw et al. (2002), the Random Forest model performed exceptionally better than many other models like multiple regression, support vector machines, etc. even rivaled models built on artificial neural networks, fuzzy CART & ARIMA and, is quite ideal for large datasets with higher dimensionality. The bootstrap aggregation (random sampling) implemented in the model can help reduce inverse effects of missing values within the data, Segal (2004). These qualities made Random forest suitable for the research project, but, the fact that it may behave like a black box approach and had very few options for optimization, was concerning.

2.4 Artificial Neural Networks

Artificial neural network (ANN) is what happens when biological neural networks are used as inspiration for building information processing & computational models. It is an integrated collection of nodes or neurons, which are interconnected in a layer-wise manner with each connection having a weightage linked to it, Haykin (1994). Similar to the biological neural networks, the artificial neurons ultimately detect patterns, identify relationships or arrive at predictions by learning from previous instances. The neurons are arranged within the models in multiple layers, starting from input layer and ending at the output layer, with one or more layers in the middle. The weighted sum of input signals(or output) is calculated from the net input derived when each neurons applies an input, activation and output function, Haykin (1994). A neural network has to go through a training phase in-order to be able to predict results for unknown inputs. Their parallel & distributed processing architecture enables them to train & better fit the data thought he modification of inter-neuron-connection weights. Samanta and Al-Balushi (2003) makes use of this property to perform feature selection on the implemented clustering techniques, however, the research project does not find this suitable for its dataset due dissimilarities in data in instances from reviewed literature when neural networks were used to implement feature extraction. Park et al. (1991) and Samanta and Al-Balushi (2003) discusses the use of ANN for forecasting electricity demands by analyzing historic data. They both had different sized but large data and decided to make use of three-layered (input, hidden & output) ANN. Due to the abundance of data instances, a back-propagation technique was implemented for training the ANN. Either studies reported excellent prediction accuracy relative to the field, complimenting ANN's capacity to capture non-linear relationships between features and influencing factors.

2.5 Gradient Booted Machines

Extreme Gradient Boosting (XGBoost) is a more flexible, protable and highly efficient form of more grounded gradient boosted machines that follow gradient boosting framework and has the corresponding machine learning algorithms, Chen and Guestrin (2016). XGBoost's quality to make use of parallel tree boosting to work on even huge datasets,

and produce precise results swiftly, makes them a popular implementation in MPI, Hadoop, SGE & other distributed environments. The library has seen quite a lot of hype as popularity as it has showed exceptional results with most standard classification and regression based benchmarks, Chen and Guestrin (2016). The scalability of XGBoost is due to the contributions from numerous optimization algorithms, where-as, speed & flexibility is provided by parallel & distributed computing, and its data sparsity handling ability to the use of innovative tree learning, Sheridan et al. (2016).

XGBoost won the John M Chamber's statistical software award. A quick analysis on the use of XGBoost in Kaggle, reveals that XGBoost has seen sudden influx of interest amongst data analysts,Megan Ridal, Kaggle Blog (2017), especially incorporate with LanbdaMART gradient boosting variant. Its reliability & scalability, improved prediction accuracy and ability to handle missing values, are the main reasons why, XGBoost is hailed in scoreboards of most Kaggle competitions. XGBoost is capable of providing out-of-core computation that allows individual machine to process huge datasets, Sheridan et al. (2016) and, this in extension can be used to connect multiple system connected end-to-end, creating a cluster, capable of even more data processing. Thus the decision to implement XGBoost in the research project was an easy one to take.

3 Methodology

The research projects focuses on making use of available sources of data in-order derive keys features from within and, use them in developing machine learning models that give the best prediction of post-collegiate earnings & student-loan repayment probability. A four-step methodology (seen in figure. 2), much similar to the popular CRISP-DM methodology, is followed to achieve effective implementation and proper progression of the research project.

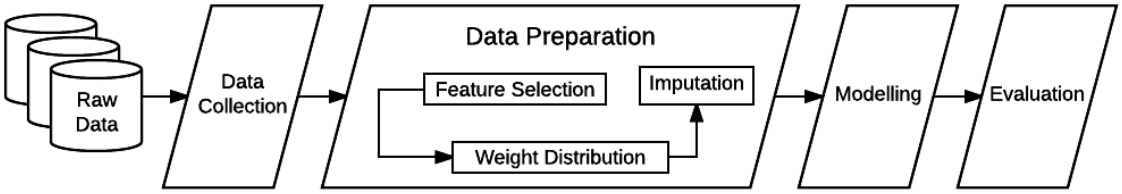


Figure 2: Followed Data Mining Methodology.

3.1 Data Collection

In this research project, a primary and multiple secondary data-sets are being worked on by as set of specialized software tools, to meet the project goals.

3.1.1 Data Overview

The public data depository, College Scorecard (Source: `https://collegescorecard.ed.gov/data/`), provides the primary the set of data. It is conceived, hosted, maintained and updated by the United States Department of Education, but, gets contributions on

applicant details, student information, college details, financial aid statements, tax returns reports, etc. from respective educational institutions and other branches of the federal government. The data-set has around 1750 attributes and 125000 unique instances, describing over 7800 educational institutions in the USA, from 1996 to 2013. The College Scorecard depository is greatly troubled by policies, reporting requirements and regulations that often changes from state to state, thus, making it gravely difficult to find conformity and uniformity within data attributes. Although, the depository is updated on a regular basis, information post 2013 is not available in the data-set. This is due to the fact that, 2013's data was reported in 2016, thus, creating a delay period.

All other secondary data-sets were taken from the US government's technology transform service, Data.gov (Source: `https://www.data.gov/`), hosted and managed by the the US General Services Administration. These data-sets were included to augment the primary data-set with characteristics which were missing or unavailable. The US national annual economic aggregates, GDP growth indicators, revenue indices, employment & unemployment rates, periods of economic unrest, etc. are some of the key aspects contained in the secondary data-sets.

3.1.2 Tools & Software Specifications

The software tools and programs used in the implementation of all stages of this project are listed below:

- Microsoft Excel 2016: Used for general data-set exploration, analysis, filtering, etc.
- PyCharm v2.4: An integrated development environment (IDE), used for running python programming language which was employed for data consolidation, cleaning, preparation, etc.
- R Studio v1.0.136: Statistical and machine learning techniques used in building the predictive models were implemented in r programming language with this IDE.
- Tableau v10.4: The visualization tool was used for a few data exploration purposes and for the graphical evaluation of results from the models.

3.2 Data Preparation

The research project has two case studies to cover and it uses two continuous numeric dependent variables to accomplish them. The annual income of alumni who were enrolled in the institution 6 years and the student-loan default rates amongst cohorts. These variables are placed in priority before while performing any data preparation. The data in its original form is not ideal to be employed into the models in this research project and, therefore, had to be converted to a more desired format. Due to the nature of the data-set, the data preparation process, without-a-doubt, consumed the significant majority of the research project's time, resources and focus. The report will try to briefly encompass the steps taken in the entirety of the data preparation process.

3.2.1 Data Consolidation & Cleaning & Normalization

The original primary data is available in 18 CSV files for data from 1996 to 2013 & another 14 CSV files for additional data from 2000 to 2013 and, 4 flat files provide the all

the secondary data. Combination and consolidation of the total 36 files individual files into a single CSV files is done with the help of some python programming. A number of JOIN and MERGE functions are used to check and combine these files, taking UNITID, Year & OPEID as the unique identifiers. The result was a 1.2 Giga-byte CSV file with around 1750 columns and 125000 rows.

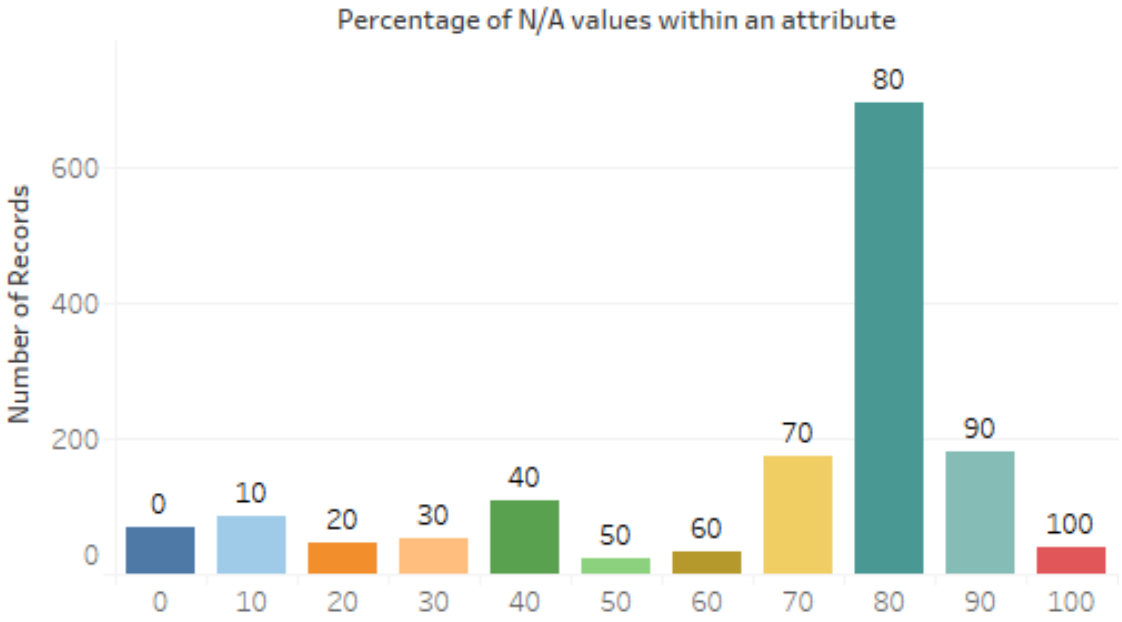


Figure 3: Percentage of N/A values in variables, prior to cleaning.

The newly created data source had no significant outliers or any duplicate values due to proper data combinations, however, it was filled with blanks (N/A) or NULL values. Figure 3, shows the spread of percentage of N/A values in the 1750 variables. All data belonging to institutions which were closed down during this (1996 to 2013) time period were removed, since, this data is likely unreliable and could not have helped further on. The research is focused only on four-year undergraduate and 2-year post-graduate students, since they are more in-line with the business question's target space. Therefore, all data on associate, certificate and non-degree programs were removed. Only data from 2002 to 2013 is considered, since, earnings reporting began only in 2002 and holds blank from 1996 to 2001. Variables like latitude, longitude, college URL, etc. served no meaningful purpose in the models, thus, they were removed in hopes of reducing time consumed in feature selection processes. Finally, all columns with more than 20 percentage NULL or Privacy-Suppressed values and institution with too much sparsity in data distribution, were removed.

3.2.2 Data Imputation

A direct & easy way to deal with missing values is to have them completely removed from the data, however, removing an complete entry for a missing value will quickly lead to the loss of information within the data and could inversely affect the underlying relationships, Williams-Johnson et al. (2010). Linear interpolation function was mostly used to perform multiple imputation, but, in certain cases functions like mean, neighbourhood, front-fill, back-fill, etc. were used instead. The decision, of which function to use, was taken based on the nature of the variable and its distribution within. Imputation was performed

over variables without too many missing values to avoid trend flattening, Horton and Kleinman (2007) and multiple times after grouping data based on respective institutions for maintaining mean & progression rates, Lee et al. (2011), especially in the first & last indexes. Apart from N/A values, special attention was given to make sure that the numerous 'Privacy-Suppressed' values also get replaced via multiple imputation for improving uniformity within the variable, Horton and Kleinman (2007), which came handy when running models on them. For this reason, imputations were done together with the feature selection process and not necessarily after.

3.2.3 Feature Engineering

Feature engineering is termed as the science & art of making it possible to extract more information from a given set of data, not by the addition of data, but, improving the usability of existing data. All data in its original form is a combination of noise & useful information and there exists techniques & practices that could be adopted to improve the information-to-noise for the better, Turner et al. (1999). Feature engineering can be taken as a collection of such practices, and when done well, it can do more good to a project than the best algorithms. It can be bifurcated into feature creation and feature transformation.

Data transformation is the replacement or modelling of a variable by allowing a function to act on it. This may alter the original magnitudes or scale of values but it manages to contain the relationships within them intact. These practices helps emphasize linear relationships over non-linear ones and converts skewed distributions into symmetric ones, Turner et al. (1999). Both normalization and principal component analysis were conducted for data transformation in the research project.

Normalization of data was done in R studio and the normalize function (BBmisc) in R programming. All continuous numerical predictor variables were normalized with scalabilty rather than simple mean distance, and the categorical variables were tokenized with the one-hot function. Normalization was performed to improve the efficiency of the data processing within the models, especially for the ANN. Before using PCA & Normalization, a few variables were replaced by combining them with others based on different functions to get a more desirable features.

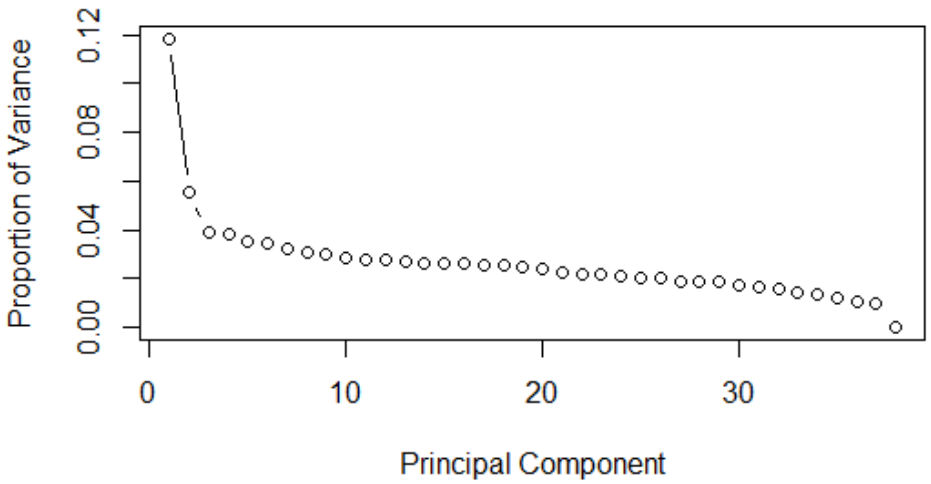


Figure 4: Proportion of variance contributed per principal component.

Principal component analysis (PCA) is essentially a multivariate variation emphas-

izing technique which is usually employed on a data containing multiple inter-correlated quantitative features, Kambhatla and Leen (2006). PCA brings out the strong patterns & extracts most important information into a new set of orthogonal features known as principal components. For n number of variables undergoing PCA, it produces n number of principal components sorted in the order of their proportion of variance. This means that those components at the lower end of the spectrum will have very little influence over the table's variance, thus they are most suitable for dimension deduction. Figure 4, refers to one of many PCAs conducted. Here, out of the 38 variable 4 principal components contributed to around 94 percentage of variance in the group, so, only those 4 were kept and the rest were removed. Removing components this way will cause some data loss, but, its a small price to pay for dropping the number of variables from 118 to 56. Since, these components only had very little influence to begin with, their elimination from the data pool will only lead to slight drop in the quality of results, Kambhatla and Leen (2006).

3.2.4 Feature Selection

The idea of feature selection is to start with a super-set of features and arrive at a much smaller subset of variables, which would help make the model more efficient at estimating the dependent variable, by reducing the total number of dimensions going into the model. The results achieved from using only the selected subset of variables can be much better or similar to the super-set. Pearson correlation test, a filter method of feature selection, was conducted to identify the features with the highest correlation to the dependent variables, however the results were less than impressive as none of the variables had satisfactory correlation coefficients.

Embedded and subset methods for offers robust feature selection Sutter and Kalivas (1993), but, due to the large size & number of dimensions of the data, they were impractical to implement. This meant that wrapper methods for feature selection would have to be used. Both forward selection and backward elimination were employed on the dataset in its original as well as normalized form. Backward elimination starts with all the variables available selected and then removes one variable per iteration, such that, the AIC improves after elimination, Sutter and Kalivas (1993). Forward selection has the very opposite methodology, where, it starts with the most correlated variable in the feature pool and starts to add the next best one. Both models continue in their direction until the addition or elimination, respectively, of features fail to improve the overall quality of the subset. In comparison, forward selection gave the lowest Akaike information criterion (AIC) at 267003 with 87 features using the non-normalized data, but, backward elimination conducted on normalized data was significantly faster and gave an AIC of 273684 with only 32 features. In the interest of efficient processing of models, and taking into account the slight difference in AIC, the 32 features selected with backward elimination was finalized.

4 Implementation

4.1 Architectural Design

Figure 5, show the design followed in the project from data preparation to building prediction models and finally visualizing the findings. The data preparation is the most time

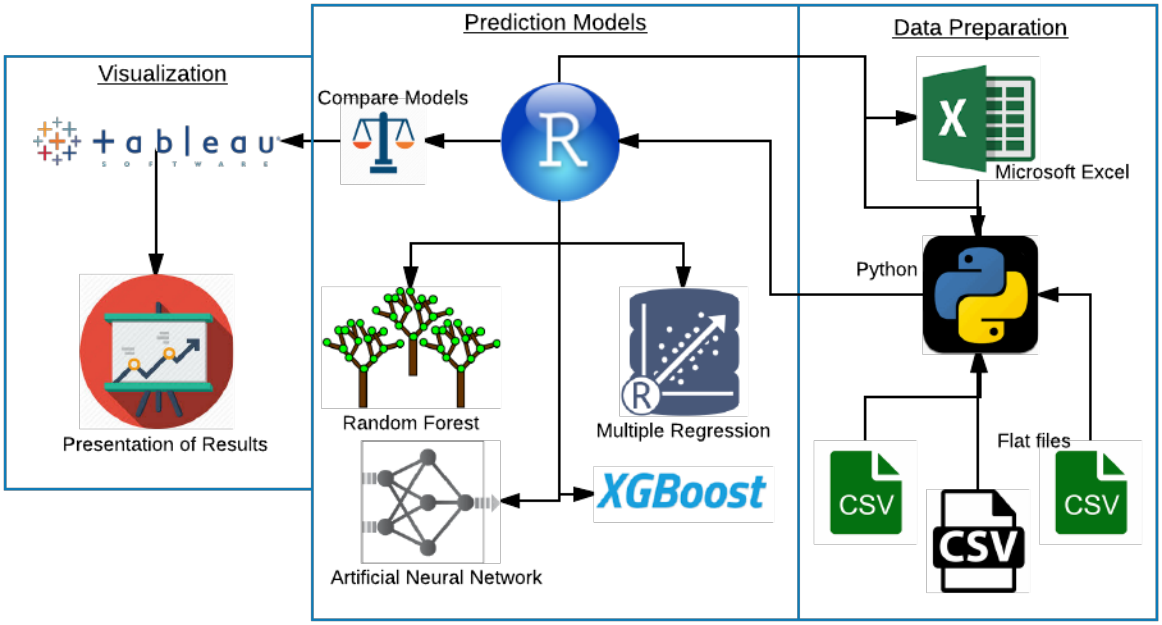


Figure 5: Architectural Design.

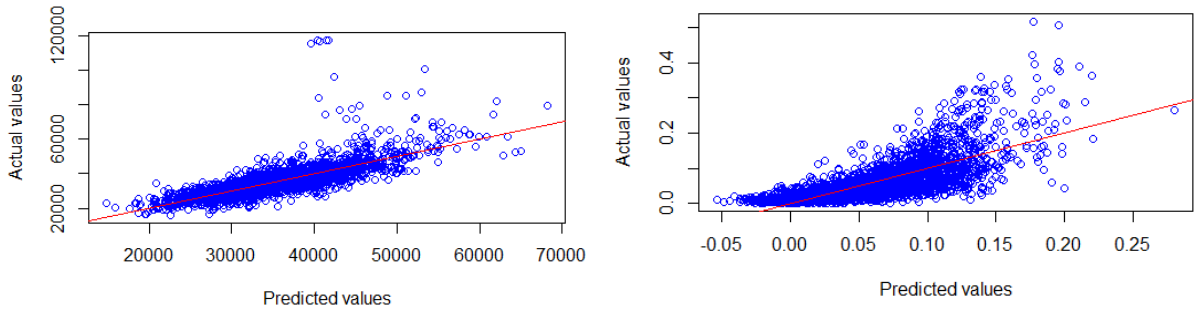
and resource consuming phase of the entire project and is perform right at the beginning. As explained in the methodology, the raw data from multiple sources are aggregated with the help of Python and Microsoft Excel. Once the desirable dataset is acquired, it is partitioned into training and testing subsets of the original dataset. The partitioning was done in a 4:1 proportion right after it got shuffled. The its allowed to be modified for the sole intention of enabling different prediction models to work on. All models are implemented within R and therefore all data to & from the models go through R studio's frameworks. Thus, on completion of processing by the models, their results are collected and compared with help of functions in R. After, acquiring satisfactory results from these models as well their complimentary functions, the results are passed over to visualization packages within R or to the dedicated visualization tool Tableau. Visualizations complimenting the case studies are put together for effect knowledge discovery and information exchange.

The following sections examine the building, optimization, processing and result extraction from the four predictive models implemented in the research project. Rather than explaining the code & algorithm as it is, a much more peripheral description is given, without, forgetting to emphasize cardinal aspects of the models. All models were subjected to the same partitions of data, as for, setting a level playing field for each model. All prediction processes for each dependent variable is performed exclusively.

4.2 Multiple Regression Model

Being one of the most popular techniques in use, it was the first model implemented on the dataset and also the simplest among. The model carries out multiple linear regressions on all the chosen features in the data, looking for any and all linear relationships that may exist between them. One of the important requirements for multiple linear regression is

that all the variables involved should be numeric & continuous in nature. The feature selection & normalization conducted had left the model with only numeric continuous variables. With all conditions met, the multiple regression model begins the process of trying to fit linear equation that captures the relationship between response variables and independent variables.



(a) PCAE (b) CDR

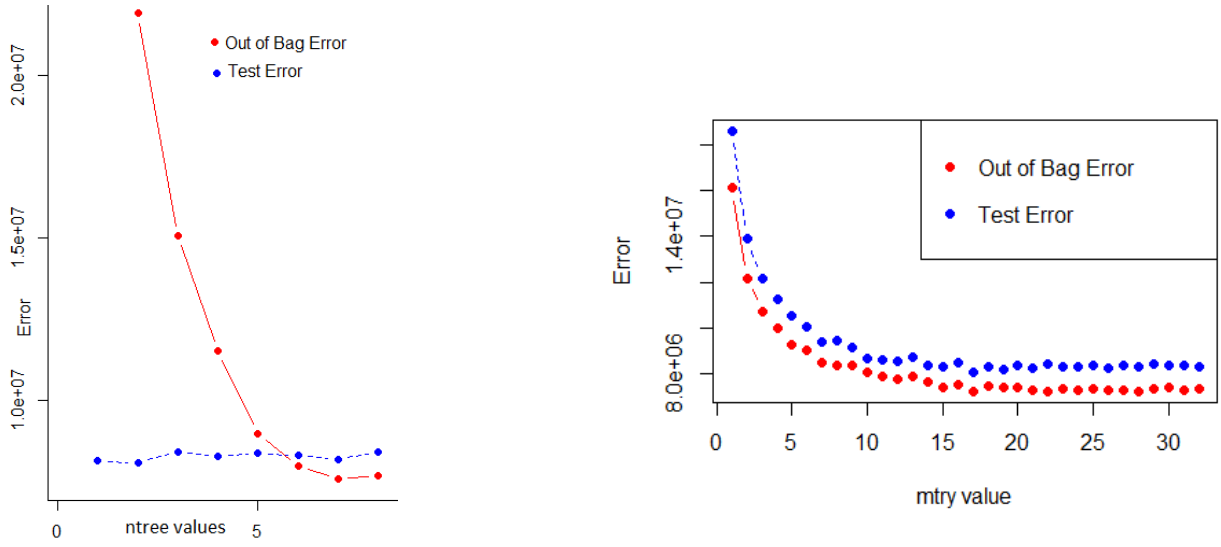
Figure 6: Predictions vs Observation

The model is reproduced in R with a simple linear model (lm) function. Training and development of the model is done on dataframe from the training data. Bigger training data can help increase training prediction accuracy, but may back fire. The model runs for both post-collegiate annual earnings (PCAE) as well as cohorts default rates(CDR), separately. At first, the training data set is used to create a model for linear regression, then this model is allowed to predict the values for the test data. The predicted values are compared to their respective observed values to calculate the quality of prediction, the model was able to achieve. For PCAE the model was able to produce root mean square error (RMSE) of 5401.6 and R-squared value of 0.5914, where-as, CDR produced a RMSE of 0.0389 & R-squared value of 0.5465. The predicted values are plotted against their corresponding actual values with a plot function in R, as seen in figure 6.a for PCAE and 6.b CDR. Ideally, all the blue-dots should be on the red line, instead, they are very much conically scattered. The figures along with the respective RMSE & R-squared values suggest that the multiple regression model was not able to deliver quality prediction.

4.3 Random Forest Model

With the help of randomForest function in R, a Random Forest (RF) was created, which during its training time creates a preset number of decision trees and their mean predictions are used to give output. The higher the number of trees better will be the training prediction accuracy but runs higher risk for over-fitting data. This will lead to poor prediction accuracy with test data. Such generalization based errors in RF can be calculated by figuring out the out-of-bag (OOB) error and resembles N-fold cross validation. The randomForest function offers two crucial parameters, mtry, the number of variables that will be held by a decision tree and mtree, the number of decision tree that will be initiated. The challenge therefore was to calculate the optimized value for mtry & mtree. This was achieved by keeping mtree constant at 500 and then allotting a range of values for mtry, while modeling a RF for each value of mtry. Each model will test its OOB & test error and finally, plotting the values across a graph, figure 7, to pick the one with the least error. Once mtry value is optimized the same looped approach is implemented for

mtree, figure 7. From the data seen in the graphs (figure 7), optimized values for mtry & mtree were selected at 17 & 500 respectively.

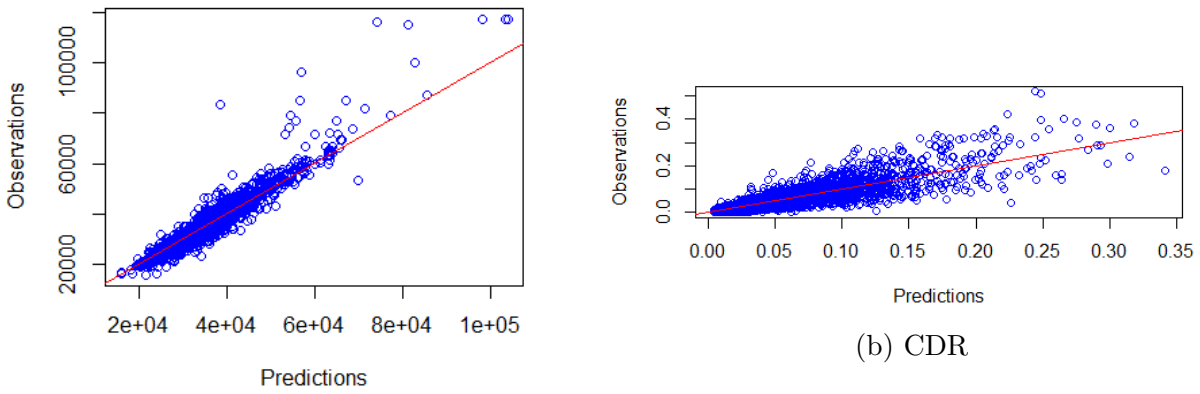


(a) PCAE

(b) CDR

Figure 7: Predictions vs Observation

Once optimized, the values were used to perform RF analysis on training data of datasets of both PCAE and CDR. The resultant model was used to predict values which were tested for prediction accuracy with the test data. When PCAE, got an RMSE of 2885.21 and R-squared value of 0.8966, CDR got an RMSE of 0.03056 & R-squared value of 0.7225. These were certainly decent results and are far better than those produced by the multiple regression model. The results for PCAE & CDR are plot graphical in figure 8.



(a) PCAE

(b) CDR

Figure 8: Predictions vs Observation

4.4 XGBoost Model

The XGBoost model is without doubt the most hyped and is at the moment in popular use. The integration of tree based learning as well as linear models and the option to be

implemented into a distributed parallel computing environment, explains its popularity. The use of XGBoost to predict PCAE & CDR values was made possible by the use of xgboost package in R. XGBoost offers parameters which when optimized could give the best results from the model. The nrounds parameter determines the number of iterations the model can perform, if the best training error value is achieved the model quits, irrespective of nround. If the XGBoost model is allowed to continue iterating without any interruptions, it will continue on until the best train error is reached, but, by this stage the model would have highly over-fit the data and would give poor results with test error. To avoid any over-fitting and to get the best prediction possible, the models were optimized by looping them and checking the test error from graphs, such as in figure 9. In the end of all the tuning test, optimized values for nrounds, eta & early-stopping were gathered as 2250, 0.09 & 10 respectively.

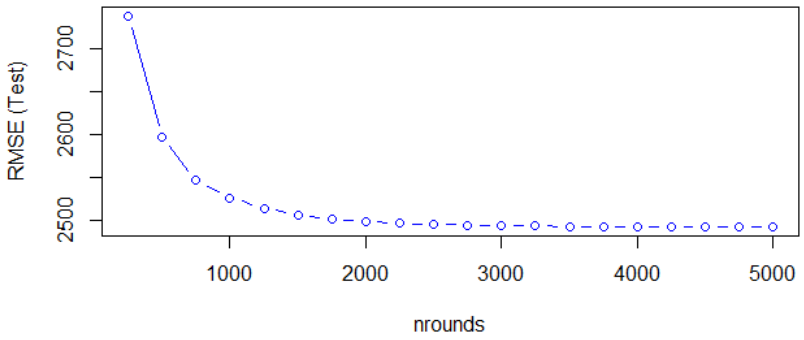
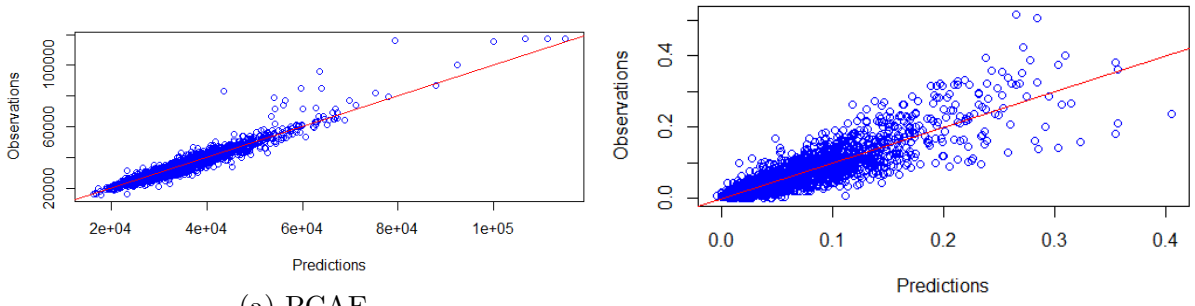


Figure 9: RMSE vs nrounds

Finally, the model is run on PCAE & CDR prediction datasets with the derived optimization values introduced into the parameters. With PCAE, a RMSE score of 2495.98 & R-squared value of 0.9153779 was acquired, when, CDR based data got RMSE of 0.02732296 & R-squared value of 0.7769. XGBoost model had given better results than random forest and multiple regression models. The final prediction results on PCAE & CDR were ploted to a graphical representation, seen in figure 10.



(a) PCAE

(b) CDR

Figure 10: Predictions vs Observation

4.5 Artificial Neural Network Model

ANN is always a prime choice when it is desirable to incorporate the relationships, especially in cases when non-linear relationships are followed, within their prediction. ANN

in R was made possible with the h2o package and the activation parameter for the model is take as Rectifier due to the regression name of the research project. Optimization of the neural network can be done with parameters such as hidden layer count, node count and epochs. Similar to the techniques adopted in the previous model, the ANN model too was optimized by the loop based iterations technique. Figure 11, shows the graphical representation of results from the optimization test conducted on epochs. Leaving the number of hidden layers as default (2), number nodes as 200 and epochs value at 12, gave the best optimization test results.

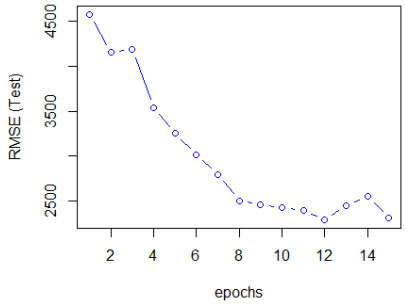
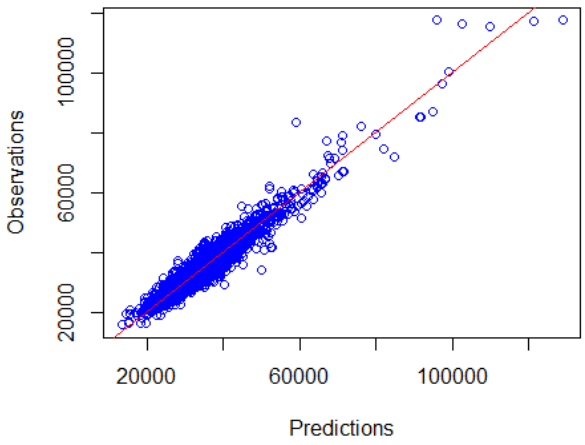
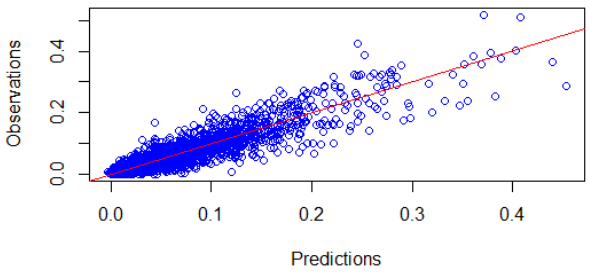


Figure 11: RMSE vs epochs

The ANN with the optimized values were able to produce exceptionally great results. On predicting and comparing with test data, on PCAE, RMSE came down to 2330.25 & R-squared of 0.9254, where as the CDR had, RMSE as 0.02414 and R-squared value of 0.82579481. For a better understanding of the prediction accuracy, the results are plotted, as seen in figure 12.



(a) PCAE



(b) CDR

Figure 12: Predictions vs Observation

5 Evaluation

The research project, on the use of predictive analytics models to estimate the post-collegiate earnings and probability of student-loan repayment prospects, has been implemented with use of multiple regression model, random forest, XGBoost & artificial neural networks. The quality of the prediction is primarily measured with the root mean square

error calculated from comparing the predicted values to the respective test data. RMSE was chosen for quantifying the quality of the prediction because error values are provides in the similar units if not scale of the dependent variable.

Model		RMSE before tuning	R-squared before tuning	Train RMSE	Test RMSE	R-squared	Speed
PCAE	MLR	10772.72	0.5914025	10772.72	5401.6	0.5914025	1
	RF	2958.580895	0.8946224	1098.51175	2885.2123	0.8966375	4
	ANN	2486.497455	0.9149663	1961.96542	2330.2552	0.9254225	2
	XGBoost	3043.636794	0.8746938	93.2891898	2495.9807	0.9153779	3

Figure 13: PCAE: Tabulated Final Quality Measures

Figure 14, shows a table containing the collection of all final prediction quality metrics that were acquired from individual models that was running on the post-collegiate earnings case study. It very clear from the table that the ANN came out having the best scores in comparison to the other three. It is also work noting that the ANN had the second most quickest processing rate, even after scaling up to huge numbers. The R-squared value helps double check the trend and results observed from the RMSE values listed.

Model		RMSE before tuning	R-squared before tuning	Train RMSE	Test RMSE	R-squared	Speed
CDR	MLR	0.03896794	0.54647482	0.07224294	0.0389679	0.5464748	1
	RF	0.03060019	0.72322256	0.03132565	0.0305677	0.7225394	4
	ANN	0.02542287	0.81245599	0.0206006	0.024147	0.8257948	2
	XGBoost	0.02990199	0.73300345	0.001446	0.027323	0.7769316	3

Figure 14: CDR: Tabulated Final Quality Measures

It is worth noting that the trend is the same with cohort's default data (CDR), shown in figure 14, which when subtracted from 1 gives the probability of a student to pay their student-loans successfully. This is because of the delinquency and default rate info gone into making CDR variable. There is no question that ANN did the best prediction in comparison & was also surprising fast, compared to the Random forest. Random forest was hands-down the most time consuming for all the other. It gave significantly worse results that XGBoost & ANN. Contrary to popular believes, ANN performed better than the new and advanced XGBoost. This is mainly due to the in fact that in the XGBoost model the prediction quality plateaus after a certain point, while ANN keep on improving the quality.

In this figure 15, The models and their pre-optimization & post-optimization RHSE as well as R-squared values into a section wise tabulation.All models showed significant improvement due to optimization. Multiple regression did not have any easily accessible methods for any kind of optimization, thus it RMSE value remain where it is. The prediction accuracy of the best model ANN is satisfactory, at PCAE domain, since a value of 2330. could be the variation of 2330 from he estimated annual earnings. such differences will not have any huge issue on the assessment of value of the institution. Similarly, with cohort default rates , a 2 to 3 percentage of difference would be quite agreeable to most applicants when imagining real world scenario. However, if the results of the 4 models were only as good as the multiple regression model, then such high error values could not be accommodated.

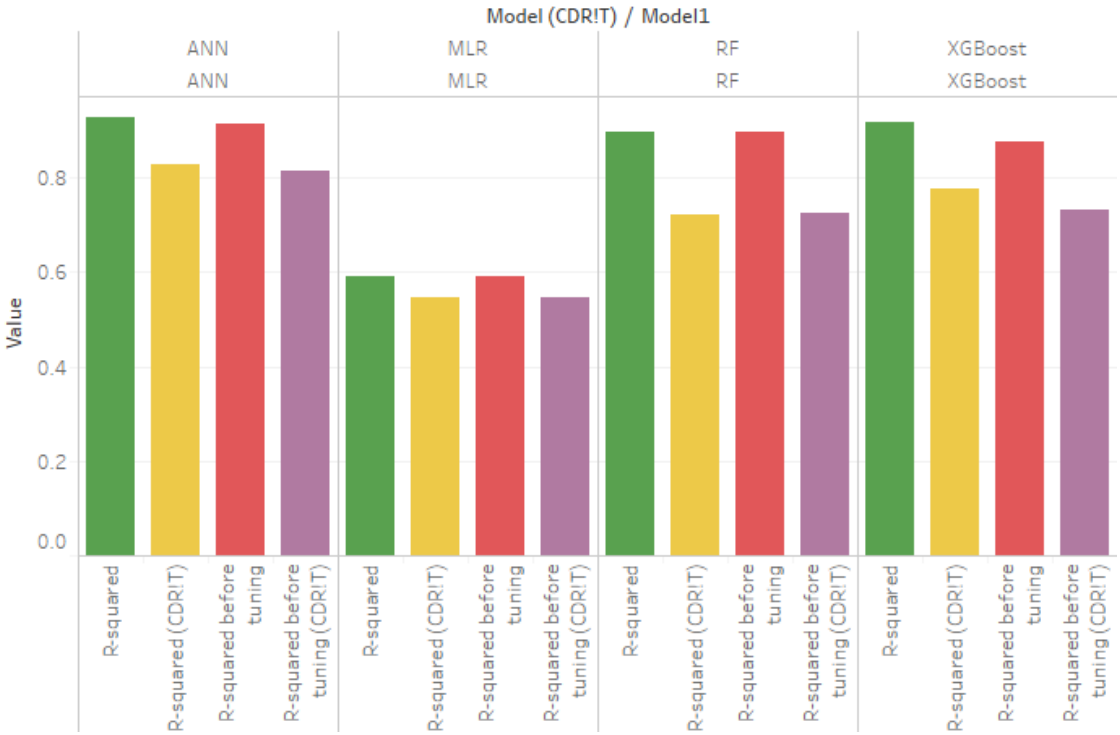


Figure 15: PCAE and CDR metrics comparison.

6 Conclusion and Future Work

College education will continue to be an integral decision and part of any student or youngster's life, thus helping them make this decision better and easier is still highly motivating to the researcher. The student debt-crisis is no longer purely an American problem since, people around the world are following their example, thus making any means to mitigate such socio-economic issues. After going through with the project, the researcher is confident that, although better quality would have been desirable, the present form of the project holds substantial value.

This is mainly due the first half of the project where qwerks & defects of the College Scorecard data can be well examined. Any researcher working on the dataset would find this research as well as the codes used to be of substantial value. It is the firm belief of the researcher that the prediction models, the quality they possess now, could serve in providing useful information. For any future works with the data, when significant number of years of data gets properly accumulated, it would serve good to approach the dataset not in a supervised learning manner, but as a time series analysis. This would make good use of the year factor in the data, which is presently useless due to short-coming in data distribution within the dataset.

Most of the issues faced during the research project was a direct result of limitation and faults of the data. These arise due to the short-comings of institutions form reporting and US Department of Education from properly rendering it into the dataset. Since, students would prefer to attend more transparent institutions, this should be used to convince institution from resorting to privacy suppression. A major design short coming of the data was that it did not have any attributes that would help quantify the quality of different institutions.

References

Aiken, L. S., West, S. G. and Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*, Sage.

Baum, S. and O'Malley, M. (2003). College on credit: How borrowers perceive their education debt, *Journal of Student Financial Aid* **33**(3): 1.

Berger, M. C. (1988). Predicted future earnings and choice of college major, *ILR Review* **41**(3): 418–429.

Breiman, L. (2001). Random forests, *Machine learning* **45**(1): 5–32.

Brewer, D. J., Eide, E. R. and Ehrenberg, R. G. (1999). Does it pay to attend an elite private college? cross-cohort evidence on the effects of college type on earnings, *Journal of Human resources* pp. 104–123.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, pp. 785–794.

Cohen, J., Cohen, P., West, S. G. and Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*, Routledge.

Duncan, A. (2015). Obama administration launches college scorecard, *Homeroom, February* .

Gladieux, L. and Perna, L. (2005). Borrowers who drop out: A neglected aspect of the college student loan trend. national center report# 05-2., *National Center for Public Policy and Higher Education* .

Goodman, J., Hurwitz, M. and Smith, J. (2017). Access to 4-year public colleges and degree completion, *Journal of Labor Economics* **35**(3): 000–000.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*, Prentice Hall PTR.

Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models, *The American Statistician* **61**(1): 79–90.

Hurwitz, M. and Smith, J. (2016). Student responsiveness to earnings data in the college scorecard.

James, E., Alsalam, N., Conaty, J. C. and To, D.-L. (1989). College quality and future earnings: Where should you send your child to college?, *The American Economic Review* **79**(2): 247–252.

Kambhatla, N. and Leen, T. K. (2006). Dimension reduction by local principal component analysis, *Dimension* **9**(7).

Lee, J. H., Huber Jr, J. et al. (2011). Multiple imputation with large proportions of missing data: How much is too much?, *United Kingdom Stata Users' Group Meetings 2011*, number 23, Stata Users Group.

Liaw, A., Wiener, M. et al. (2002). Classification and regression by randomforest, *R news* **2**(3): 18–22.

Ma, J., Pender, M. and Welch, M. (2016). Education pays 2016: The benefits of higher education for individuals and society. trends in higher education series., *College Board* .

Megan Ridal, Kaggle Blog (2017). No free hunch : Xgboost. [Online; accessed December 01, 2017].
URL: *http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting*

Morey, A. I. (2004). Globalization and the emergence of for-profit higher education, *Higher education* **48**(1): 131–150.

Park, D. C., El-Sharkawi, M., Marks, R., Atlas, L. and Damborg, M. (1991). Electric load forecasting using an artificial neural network, *IEEE transactions on Power Systems* **6**(2): 442–449.

Preacher, K. J., Curran, P. J. and Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis, *Journal of educational and behavioral statistics* **31**(4): 437–448.

Ryan Gorman, Business Insider UK (2015). How student-loan debt is dragging down the economy. [Online; accessed December 01, 2017].
URL: *http://uk.businessinsider.com/3-charts-explain-the-effect-of-student-loans-on-the-economy-2015-5?r=US&IR=T*

Samanta, B. and Al-Balushi, K. (2003). Artificial neural network based fault diagnostics of rolling element bearings using time-domain features, *Mechanical systems and signal processing* **17**(2): 317–328.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression, *Center for Bioinformatics & Molecular Biostatistics* .

Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J. and Gifford, E. M. (2016). Extreme gradient boosting as a method for quantitative structure–activity relationships, *Journal of chemical information and modeling* **56**(12): 2353–2360.

Sutter, J. M. and Kalivas, J. H. (1993). Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection, *Microchemical journal* **47**(1-2): 60–66.

Turner, C. R., Fuggetta, A., Lavazza, L. and Wolf, A. L. (1999). A conceptual basis for feature engineering, *Journal of Systems and Software* **49**(1): 3–15.

Wachtel, P. (1976). The effect on earnings of school and college investment expenditures, *The Review of Economics and Statistics* pp. 326–331.

Williams-Johnson, J., McDonald, A., Strachan, G. G. and Williams, E. (2010). Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (crash-2): a randomised, placebo-controlled trial, *West Indian Medical Journal* **59**(6): 612–624.