INSTITUTE OF MANAGEMENT, NIRMA UNIVERSITY

MINOR RESEARCH PROJECT PROPOSAL

ON

Data Mining Applications to predict students' success in completion of MBA program

PRINCIPAL INVESTIGATOR:

Dr. Sunita Guru

Institute of Management, Nirma University, Ahmedabad.

CO- PRINCIPAL INVESTIGATOR:

Dr. Mahesh K.C.

Institute of Management, Nirma University, Ahmedabad.

Data Mining Applications to predict student's success in completion of MBA program

Executive Summary

Application of data mining tools and techniques is a hot topic in recent years and has been applied in several fields including education research. Globally, India holds an important place in the education sector and a significant number of students receive MBA degrees each year. Thus in educational institutions, tremendous data is available every year from which some useful information can be extracted. Data mining tools and techniques can be adopted to extract deep insights related to education.

The main goal of this study is to apply data mining tools and techniques to predict and develop a model for student's success/failure in completing an MBA program. In this study we have used two classification models via logistic regression and decision tree to predict/classify students' success/failure. Logistic regression is used here to describe the association between a categorical response variable and a set of predictors. The ensemble method namely the random forest model is also used to estimate the success/failure (placed/unplaced) of a student.

Logistic regression model provides the percentages obtained in SSC, percentages obtained in HSC and graduation in science as significant predictors in deciding a student's success. However, using the random forest model we found that in addition to percentages in SSC, HSC and graduation, the CAT percentile, personal interview score, first year CGPA and final year CGPA were also significant.

Keywords: data mining, classification, education research, logistic regression, random forest, decision tree, MBA education.

Introduction

The capacity of data saved in the database is escalating at an incredible pace. Due to the rapid growth of technology advancements and digitalization, large amounts of data captured, generated and consumed in the universe is projected to rise speedily – approximating 59 zettabytes by 2020 (Holst, 2021). Every day around 2.5 quintillion bytes of data are produced (Marr, 2018). Hence there is need of statistical techniques that would help humans to automatically evaluate the data set for drawing useful insights. Data mining is a commonly used procedure to analyze large volumes of data to draw valuable insights and information.

In the worldwide education industry, India holds a significant place as it has the highest population of approximately 500 million in the age group of 5-24 years. This offers a great prospect for the education industry in India (IBEF, 2021). Moreover, India had around 39,931 colleges in year 2019. On an average around three and a half lakh students receive MBA degrees every year in India (India Today, 2020). Thus in educational institutions, data is increasing rapidly and hence there is the need to utilize this data and transform it into meaningful and useful information through data mining. There is hardly any study done in the Indian context to use data mining techniques to predict students' success in MBA programs. The main goal of this study is to apply data mining tools and techniques to predict and develop a model for student's success in completing an MBA program.

Literature Review

Data mining was first proposed in 1990s (Daniel, 2015). According to Shafiabadi, et al. (2021), data mining is defined as 'A complicated process to identify the correct, new and potentially useful patterns and models in a large amount of data'. A pattern is an unusual structure or relationship in the data set (Hand et al., 2000). Data mining is also called data or knowledge discovery' (Segall et.al.2008) and is mainly applied to excerpt concealed patterns and to notice associations between parameters in a huge amount of information Križanić (2020). The various definitions of data mining given by different researchers are as follows (Table 1):

Table 1: Different definitions of Data Mining

Authors	Definitions
Liu, et al. (2021)	It is the process of mining relevant information from the bases of data, data granaries or other information stored in a database, counting frequent arrangements, associations and variations in anomalous and substantial structures.
Križanić (2020)	It entails the application of data analysis methods to extract unmanifested information from data by performing the functions of pattern identification and predictive modeling.
Yoseph, et.al. (2020)	It is the process of extracting information from large data sets in order to turn it into comprehensible form for further actions.
Zheng & Cao(2020)	It is an intricate process of mining hidden information that will reveal user preferences and possibly valued information and directions for decision making from a large number of data sets.
Pascu (2018)	The process of mining information from existing data.
ELAtia et.al. (2016)	Analysis of observational data sets to find unsuspected relationship and to summarize the data in new ways that are comprehensible and of use to data owners.
Ahmad et.al. (2015)	It is the process of extracting relevant information and knowledge from a large set of warehouses.
Segall, et.al. (2008)	It is the process of examining data from diverse viewpoints and converting it into useful information
Berry & Linoff (2004)	It is the process of discovering and examining huge data sets that can reveal patterns and guidelines that can address a problem
Hand et al., (2000)	The process of pursuing interesting or valued information within big data sets
Fayyad et.al. (1996)	The course of finding interesting data arrangements unseen in huge data sets.

Data mining techniques

There are several data mining techniques. The following are the details of the commonly used techniques:

Decision Trees

It is a popular and powerful instrument for prediction and classification (Vandamme, et al., 2007) It basically contains nodes and divisions and the initial node is called 'root node'. Root node is determined by calculating the attributes which will most precisely categorize the objects conferring to the values of the decision variable. The procedure is repeated where the branches of the tree are right or left to another node. All path leads to a terminal node for any tree, confirming an important decision that is in conjunction with several tests. A decision is then made on the assignment of a class. There are several strengths and weakness of decision trees. The positives include the capacity to create comprehensive rules, handle both continuous and categorical variables, provide a clear sign of which aspects are the most significant for classification and prediction. The weakness includes the inability to predict in the presence of numerous nodes and classes.

Principle Component Analysis

It is a multivariate method to analyze a data table where observations are labelled by some inter-correlated quantitative dependent variables. The goal of this technique is to extract vital data from the table, to signify it as a set of novel orthogonal variables called principle components, and to exhibit the pattern of resemblance of the observations and of the variables as points in maps (Abdi and Williams, 2010).

Clustering

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering (Acharya & Sinha, 2017). Jain et, al., (1999) defines cluster analysis "as a technique for producing organized collections of arrangements into groups based on their resemblance of some property or action". The main aim in grouping is to locate information points that logically cluster together, splitting the entire data set into a set of groups (Baker, 2010). These methods are beneficial in instances where the most shared categories within the data set are unknown in prior.

Classification

In the classification method, features of new objects are examined and then assigned a set of predefined classes (Pascu, 2018). It is basically a data extracting method used to forecast group membership for data instances (Phyu,2009).

Regression Analysis

It is believed to be one of the most important data extracting techniques (Feng & Wang, 2004). It is a statistical method for approximating the relationship among variables with reason and result relation (Uyanık and Güler,2013). Regression models with a single dependent and a single independent variable is known as univariate regression. Multilinear regression is a regression model with a single dependent and more than one independent variable.

Data mining methods adopted by several researchers and their objectives are as follows (Table 2):

Table 2: Studies conducted on data mining

Author	Objective	Data mining method adopted
Martín- García et al. (2019)	To examine the stages of acceptance of blended learning in higher education	Cluster technique, Classification tree
Gharoun et al. (2019)	To develop a model to detect fault in aircraft turbofan aircraft engine	MLP, RBF, ANFIS
Janeja et al. (2018)	To predict clinical trial results with great precision	Classification, Class association rules, Clustering
Ahmad (2017)	To predict customer satisfaction for specific brands	Regression and attribute selection models
Hsu et al. (2017)	Developing index to measure driving performance	M5 model tree
Shortridge et al., (2015)	To predict the percentage of a country's population suffering from undernourishment	Different models- Linear , Generalized additive, Gaussian mixture , tree, ensemble and null
Wicker & Breuer (2013)	Explore the critical factors related to problems of organization	Decision tree

Cărbureanu (2012)	To predict economic indicators	PCA, Simple linear regression, Decision Trees
Chye et al. (2004)	Construct a model for credit scoring	Predictive modelling technique
Feng & Wang (2004)	To develop a model to predict the performance of the knurling process	Regression analysis and artificial neural networks
Safer (2003)	To predict abnormal stock market returns	Neural networks and Multivariate Adaptive Regression Splines

Application of data mining

Data mining is also referred to as Knowledge Discovery in Databases (KDD). It is the arena of determining innovative and possibly beneficial data from huge amounts of information (Baker, 2010). KDD is valued in various disciplines like lead management in telecommunications sector (Espadinha-Cruz et al., 2021), text mining (Shen & Qin, 2021), market segmentation (Yoseph, et.al., 2020), assessing effective disassembly time of industrial products (Marconiet al., 2019), assessment of environmental stress factors on plants (Segall et.al, 2008), an anomaly detection and dynamic energy performance evaluation technique for HVAC systems (Xu et al., 2021), development of policy initiatives to decrease severe vehicle-bicycle crashes (Zhu,2021), safety driven inferences (Singh, & Maiti, 2020), smoking status (Groenhof, et.al., 2020), credit decision making (Li & Liao, 2011), detecting fraud in financial statements (Kopun, 2020) and detecting adverse drug reaction (Karimi, et al., 2015).

Data mining techniques have also been applied in the education sector. According to Krizanic (2020) educational data mining is the discovery of information with the help of data mining methods in education. Educational Data Mining is also referred to as ‘the application of data extracting techniques to education related data for its analysis’ (Romero & Ventura, 2007). Data mining methods adopted by several researchers in education sector are as follows (Table 3):

Table 3: Studies conducted on data mining tools and techniques

Author	**Objective**	**Data mining method adopted**
Prenkaj et al. (2020)	To explore student dropout prediction in online courses	Deep Learning
Martín et al. (2019)	To examine the different stages of embracing blended learning and to find the relationship between them	Clustering analysis and decision tree analysis
Márquez- Vera et al. (2016)	To predict early dropout in school education	Classification
Chareonrat (2016)	To explore student dropout rates	Classification
Daniel (2015)	To understand challenges and opportunities affecting institutions of higher education	Big data and analytics
Jantawan & Tsai (2013)	To estimate Graduate Employment	Bayesian method and tree method
Djulovic & Li (2013)	To forecast retaining of students in university	Decision trees, Neural networks & rule Induction
Parack, Zahid, & Merchant (2012)	Profile and group student	A priori algorithm and K-means clustering
Hung , Hsu & Rice (2012)	To forecast a course enables students to achieve their goals	Clustering analysis and decision tree analysis
Vialardi et al. (2011)	To streamline the registration process for students based on scholastic performance	CRISP-DM

	To help students in the enrollment process	
Ramaswami and Bhaskaran (2010)	To predict students' performance	Simple regression
Antunes (2010)	To forecast why UG students' fail	Class Association Rules
Kovacic(2010)	To recognize students at risk of opting out in higher education	Regression, Classification, Exhaustive CHAID
Zhang et al.(2010)	To identify students at peril in higher education	Naïve Bayes, support vector
Delen (2010)	To forecast student retention in a university	Artificial neural networks, decision trees, support vector machines and logistic regression
Dekker et al. (2009)	To predict whether students would drop out after the first year of college	CRISP-DM
Lykourentzou et al. (2009)	To forecast student retention in a university	Artificial neutral networks, decision trees, support vector machines and logistic regression
Cortez and Silva (2008)	To build model of the student's performance in secondary schools	Decision trees, random trees, neural networks and support vector machines
Vandamme , Meskens and Superby (2007)	To predict student's academic success	Decision Tree, Neural networks, Linear Discriminant analysis

AI- Radaideh et al. (2006)	To predict future performance of students enrolled in C++ courses	Classification
Luan (2002)	To predict the probability of students dropping out	Artificial Neural Network , Decision Trees

After extensive and intensive literature review it was found that there is hardly any study done in the Indian context to use data mining techniques to predict students' success in MBA programs Thus the main purpose of this research is to advance a model that would classify/predict whether or not students will get placements upon completion of the MBA program (placed or not)

Research Methods and Techniques

The research design for this study is quantitative in nature. Primary data pertaining to admission, placement and academic grades was collected from the admission office, the placement cell and the examination cell of the institute for the students pursuing MBA. The total sample size was 244. Data was analyzed using different data mining techniques and tools: Logistic Regression and Decision Tree.

The research design for this study is quantitative in nature. Primary data was collected through survey from 244 students of a management college of Western region in India. Data was analyzed using different data mining techniques and tools: Logistic Regression and Decision Tree.

Logistic Regression Model

In classical statistics, a regression model is generally used to predict a future value based on a suitable mathematical model which minimizes the errors. Two types of variables are involved in regression modelling-the variable being predicted that is referred to as response variable (say Y) and the variable being used to predict the response variable usually referred to as predictor (say X). If the response variable is continuous, then one can use a linear regression model for prediction. But when the response variable is categorical in nature, linear regression will not be a suitable technique for prediction as the model violates the basic assumptions of linear regression like linearity, normality etc.

An alternative approach to describe the relationship between a categorical response variable and a set of predictors is logistic regression which assumes a non-linear relationship between the response and the predictor.

Let Y be a dichotomous response variable defined by $Y = \begin{cases} 1 & if\ the\ response\ is\ positive \\ 0 & otherwise \end{cases}$ and $X_1, X_2, \dots, X_p$ be a set of p predictors. Let $\pi(x) = P(Y = 1|X_1 = x_1, X_2 = x_2 \dots, X_n = x_n)$ is the conditional probability of $Y = 1$ for a given values of $X_1 = x_1, X_2 = x_2 \dots, X_n = x_n$. Then the multiple logistic regression model is defined as (see James et.al, 2015 and Hastie et.al. 2013):

$$\pi(x) = P(Y = 1|X_1 = x_1, X_2 = x_2 \dots, X_n = x_n) = \frac{\exp(g(x))}{1+\exp(g(x))}; 0 \leq \pi(x) \leq 1 \text{ where}$$

$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots, +\beta_p x_p$ is the logit function. The overall significance of the model achieved using the likelihood ratio (deviance) test and the individual predictor's significance is through the Wald's test.

Deviance Test: Typically used to test the overall significance of the fitted model in logistic regression. Analogues to ANOVA in multiple linear regression. Deviance (D) refers to the degree to which a particular model deviates from another model. It is similar to the concept of SSE in linear regression. We compare the deviance of the current model with the deviance of a naïve model where a naïve model is a model in which no predictors exist and each record is classified as belonging to the majority class. The Deviance test statistic G is defined as:

$$G = D(model\ without\ predictor) - D(model\ with\ predictor).$$

It can be shown that $G = -2ln\left[\frac{likelihood\ without\ predictor}{likelihood\ with\ predictor}\right] \sim \chi_p^2$ where p is the number of predictors added to the model (see Larose and Larose, 2016). For a given level of significance α, the null hypothesis of no significant difference is rejected if the $p - value\ < \alpha$.

The Wald Test: Typically used to test the individual predictor significance of the fitted model in logistic regression. Analogues to t-test in multiple linear regression. The test statistic in this case will be: $Z_{Wald} = \frac{\widehat{\beta_r}}{SE(\widehat{\beta_r})} \sim N(0,1), r = 1,2,..,p$ where $\widehat{\beta_r}$ is the estimated value of β (see Larose & Larose, 2016).

For a given level of significance α, the null hypothesis of no significant difference is rejected if the $p - value\ < \alpha$.

We try to build a logistic regression model to predict/classify a student's success/failure in the program in terms of whether he/she is placed or not. Based on the admission data for the batch 2019-21, we have initially selected 14 variables which are given below in Table 1.

Table 1: Data Description of initial 14 variables

Variable	Description	Type of the variable
GENDER	Gender of the student	Categorical (Male & Female)
SSC_B	Board under which the student has passed matriculation	Categorical (CBSE, ICSE and State)
SSC_P	Percentage of marks SSC	Continuous
HSC_B	Board under which the student has passed higher secondary	Categorical (CBSE, ICSE and State)
HSC_P	Percentage of marks HSC	Continuous
GRAD	Stream in Graduation	Categorical (Commerce, Science, Management and Others) Commerce: B. Com Science: B. Tech & B.Sc. Management: BBA Others: BCA, BDS, BHM, B. Pharm, BMS and BA
GRAD_P	Percentage of marks in graduation	Continuous
QE	Qualifying Exam	Categorical (CAT and others)
EXP	Prior Work Experience (in months)	Integer
PIS	Personal Interview Score	Continuous
CAT_P	Qualifying Exam percentile	Continuous
FYCGPA	First year CGPA	Continuous
FCGPA	Final CGPA	Continuous

Placement	Whether student is placed or not placed/opted out	Categorical (YES, NO)

The categorical variables are converted to the corresponding indicator (dummy) variables. It should be noted that a categorical variable with k levels require only $k-1$ indicator variables.

$$GENDER = \begin{cases} 1 & if\ male \\ 0 & if\ female \end{cases},\ SSC_{CBSE} = \begin{cases} 1 & if\ the\ board\ is\ CBSE \\ 0 & otherwise \end{cases},$$

$$QE = \begin{cases} 1 & if\ CAT\ passed \\ 0 & otherwise \end{cases},\ SSC_{ICSE} = \begin{cases} 1 & if\ the\ board\ is\ ICSE \\ 0 & otherwise \end{cases},$$

$$GRAD_C = \begin{cases} 1 & if\ the\ student\ is\ Commerce\ graduate \\ 0 & otherwise \end{cases}$$

$$GRAD_S = \begin{cases} 1 & if\ the\ student\ is\ Science\ graduate \\ 0 & otherwise \end{cases},$$

$$GRAD_M = \begin{cases} 1 & if\ the\ student\ is\ Management\ graduate \\ 0 & otherwise \end{cases},\ Placement = \begin{cases} 1 & if\ placed \\ 0 & otherwise \end{cases}$$

The categorical variables $SSC_{State}, HSC_{State}\ and\ GRAD_{Others}$ are taken as reference variables and have been removed from the analysis. Since all the students admitted to the program qualified the CAT exam, the variable QE is also removed from the analysis. After this we have finally the following set of 17 variables given in Table 2.

Table 2: Data Description of final list of 17 variables

Variable	Description	Type of the variable
GENDER	Gender of the student	Categorical
SSC_CBSE	Board under which the student has passed matriculation	Categorical
SSC_ICSE	Board under which the student has passed matriculation	Categorical
SSC_P	Percentage of marks SSC	Continuous
HSC_CBSE	Board under which the student has passed higher secondary	Categorical
HSC_ICSE	Board under which the student has passed higher secondary	Categorical
HSC_P	Percentage of marks HSC	Continuous
GRAD_S	Science graduate	Categorical

GRAD_C	Commerce graduate	Categorical
GRAD_M	Management graduate	Categorical
GRAD_P	Percentage of marks in graduation	Continuous
EXP	Prior Work Experience (in months)	Integer
PIS	Personal Interview Score	Continuous
CAT_P	Qualifying Exam percentile	Continuous
FYCGPA	First Year CGPA	Continuous
FCGPA	Final CGPA	Continuous
Placement	Whether student is placed or not placed/opted out	Categorical (YES, NO)

There were two missing values corresponding to the variables PIS and CAT_P. These missing values are replaced by median imputation. We first checked the multicollinearity in the data-a condition where two or more predictors are correlated using Generalized Variance Inflation Factor (gVIF). A gVIF value greater than 5 indicate moderate multicollinearity while a gVIF value greater than 10 indicates severe multicollinearity. Multicollinearity produces incoherent results and hence must be eliminated before building the model. It has been found that the predictors $FYCGPA \text{ and } GRAD\_C$ has VIF value greater than 5 and the predictor CAT_P has VIF value greater than 10. We have removed these variables from the model building process.

Using the Wald's test we found that the predictors $SSC_P, HSC_P \text{ and } GRAD_S$ significantly contribute in predicting a student's success or failure at 10% level. Based on the deviance test the overall model is also significant $(p - value = 0.017)$ at 10% level. The R-output of the model adjusted to three decimals is summarized in the following Table 3.

Table 3: Model Summary

Predictor	Estimates	p-value
Intercept	4.169	
SSC_P	-0.095	0.1
HSC_P	0.079	0.04
GRAD_S	1.416	0.06

The logit function and the corresponding estimated logistic regression model will be:

$$\widehat{g(x)} = 4.169 - 0.095SSC_P + 0.079HSC_P + 1.416GRAD_S$$

$$\widehat{\pi(x)} = P(Placement = 1|SSC_P, HSC_P, GRAD_S) = \frac{\exp\left(\widehat{g(x)}\right)}{1 + \exp\left(\widehat{g(x)}\right)}$$

The above model can be used to predict the probability of getting placed (success) given the predictor values. For example, the probability of the candidate getting placed when he/she scored 75% in SSC, 60% in HSC and 0.96 in a undergraduate degree in the science stream.

Now we try to build a classification model to classify a student into placed (success) or not placed (failure) based on the predictors $SSC_P, HSC_P\ and\ GRAD_S$. It has been observed that the response variable "Placement" is highly imbalanced in the sense that almost 95% (232 are placed and 12 are not placed) are placed. In this case the classification model simply classifies (or predicts) "placed" for all students. So balancing the data is required. We used a technique called "oversampling" – duplicating samples from the lower frequency class-technique to balance the data to an extent and develop a new model based on this data. Based on the deviance test the overall model is significant (p-value =0) at 5% level. The individual predictors are also significant using Wald's test at 5% level. The R-output of the model adjusted is summarized in the following Table 4.

Table 4: Model Summary

Predictor	Estimates	p-value
Intercept	2.622	
SSC_P	-0.102	0.00029
HSC_P	0.083	0.00001
GRAD_S	1.957	0.00000009

The new logit function and the corresponding estimated logistic regression model will be:

$$\widehat{g(x)} = 2.622 - 0.102SSC_P + 0.083HSC_P + 1.957GRAD_S\ and\ \widehat{\pi(x)} = \frac{\exp\left(\widehat{g(x)}\right)}{1 + \exp\left(\widehat{g(x)}\right)}.$$

Building a Classification Model

One of the primary objectives of logistic regression is to classify observations based on the predicted probabilities of the class $P(Y = 1)$. This will help the decision maker to classify the observation as belonging to either class 1 (positive) or class 0 (negative). This can be achieved by deciding a cut-off probability P_c such that if the predicted probability is less than P_c then the observation is classified as negative ($Y_i = 0$) otherwise the observation is classified as positive ($Y_i = 1$). That is, $Y_i = \begin{cases} 1 & if\ P(Y_i = 1) \geq P_c \\ 0 & if\ P(Y_i = 1) < P_c \end{cases}$. For a logistic regression, the default cut-off probability is 0.5.

The Confusion matrix: Consider two classes C_1 (positive class or 1) and C_2 (negative class or 0). Let n_{ij} denotes the number of records that are class C_i members and were classified as C_j members. Then a general confusion matrix will be of the form:

Predicted Class	Actual Class	
	C1 (1)	C2 (0)
C1 (1)	$number\ of\ C_1\ records\ classified\ correctly$ (or true positive TP) n_{11}	$number\ of\ C_2\ records\ classified\ incorrectly\ as\ C_1$ (or False Positive FP) n_{21}
C2 (0)	$number\ of\ C_1\ records\ classified\ incorrectly\ as\ C_2$ (or False Negative FP) n_{12}	$number\ of\ C_2\ records\ classified\ correctly$ (or true negative TN) n_{22}

Also note that $n = n_{11} + n_{12} + n_{21} + n_{22}$.

The accuracy of the model is defined as: $Accuracy = \frac{TP+TN}{n} \in [0,1]$. Higher the accuracy, better the model. It should be noted that the model selection cannot be completely based on the overall accuracy as a model with higher overall accuracy may not be the better model as the worth of a model depends to a great degree on the cut-off probability. An alternate way of measuring the model performance in logistic regression is done based on the concept of sensitivity and receiver operating characteristic (ROC) curve and its area under the curve (AUC) (see Dinesh Kumar, 2017).

Sensitivity: The ability of the model to correctly classify positives. In medical science, it is the ability of the diagnostic test to identify disease if it is present in a patient. Statistically, $Sensitivity = P(model\ classifies\ Y_i\ as\ positive | Y_i\ is\ positive = \frac{TP}{TP+FN}$.

ROC Curve and AUC: It is a plot between sensitivity and 1-specificity. As a rule of thumb, $AUC \geq 0.7$ for all practical applications.

With a default cut-off value of 0.5, the above model provides an accuracy of 80% indicating that when a new record comes, 80% of the time the model correctly classifies the record. The sensitivity of the positive class (Placement=Yes) indicates the probability that given $Placement = Yes$ the model classifies it correctly (see Table 5).

<table>
<tr><th colspan="5">Table 5: Summary of Classification Model</th></tr>
<tr><td colspan="3">Confusion Matrix</td><td>Accuracy</td><td>Sensitivity</td></tr>
<tr><td rowspan="2">Predicted</td><td colspan="2">Actual</td><td rowspan="4">0.8</td><td rowspan="4">0.97</td></tr>
<tr><td>0</td><td>1</td></tr>
<tr><td>0</td><td>14</td><td>6</td></tr>
<tr><td>1</td><td>54</td><td>226</td></tr>
</table>

It has been observed that for the above model, the exact cut-off value is 0.55 with overall accuracy of 83% (see figure 1).

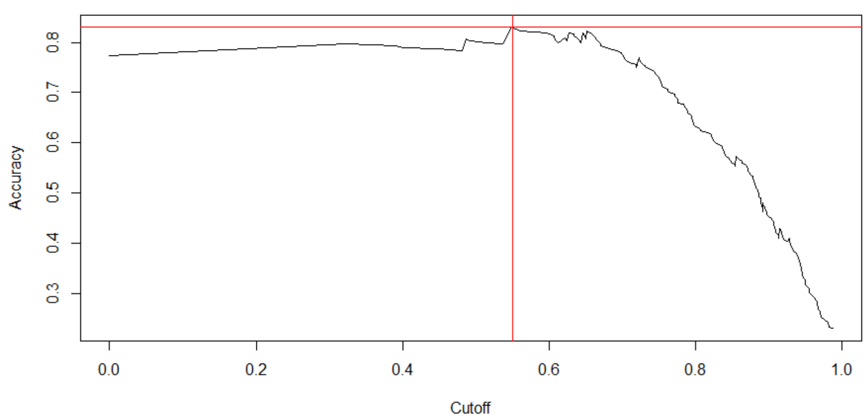

Figure 1: Cutoff probabilities against Accuracy

To check the overall worth of the logistic regression model and thereby logistic regression as a classifier, we used the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) and found that $AUC = 0.77$ indicating that the model is moderately good (see figure 2).

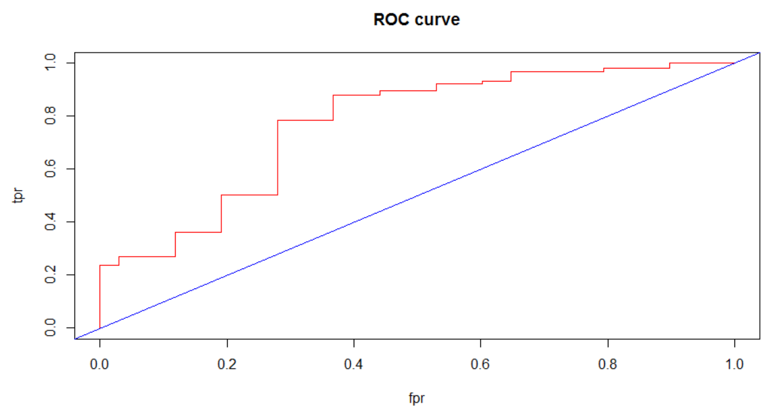

Figure 2: ROC curve and AUC

Random Forest Model

Here we use an ensemble method called random forest algorithm to predict the success (placed) of a student. The ensemble methods combine results from a set of classification models in order to increase the accuracy and reduce the variability of the classification. In any modelling, the prediction error (residual) is a function of bias, variance and noise. The noise component of the residual is an intrinsic characteristic of the prediction problem which cannot be eliminated. The random forest algorithm uses bagging (Bootstrap Aggregate Sampling)—a resampling technique which creates subsets of the same size from the original data with replacement. The algorithm will construct several trees (in R random forest will generate a default 500 trees) and combine the performance and the final model will be selected using the majority voting principle.

More applications and examples of decision trees can be found in James, et.al. (2015), Hastie, et.al (2013), Zhou, (2021) and Salcedo, (2019).

We have developed a random forest model to predict whether a student is placed or not and the accuracy of the model is found to be 1 (see Table 6).

<table>
<tr><th colspan="5">Table 6: Summary of Random Forest Model</th></tr>
<tr><td colspan="3">Confusion Matrix</td><td>Accuracy</td><td>Sensitivity</td></tr>
<tr><td rowspan="2">Predicted</td><td colspan="2">Actual</td><td rowspan="4">1</td><td rowspan="4">1</td></tr>
<tr><td>0</td><td>1</td></tr>
<tr><td>0</td><td>69</td><td>0</td></tr>
<tr><td>1</td><td>0</td><td>231</td></tr>
</table>

The variable importance plot shows that the predictors which significantly contribute towards the classification/prediction are GRAD_P, HSC_P, PIS, CAT_P, FCGPA, FYCGPA and SSC_P based on the mean decrease in Gini (see Table 7 and figure 3).

Table 7: Mean Decrease in Gini

Variables	Mean Decrease in Gini
GRAD_P	17.636
HSC_P	13.173
PIS	11.688
CAT_P	11.538
FCGPA	10.565
FYCGPA	10.308
SSC_P	10.100

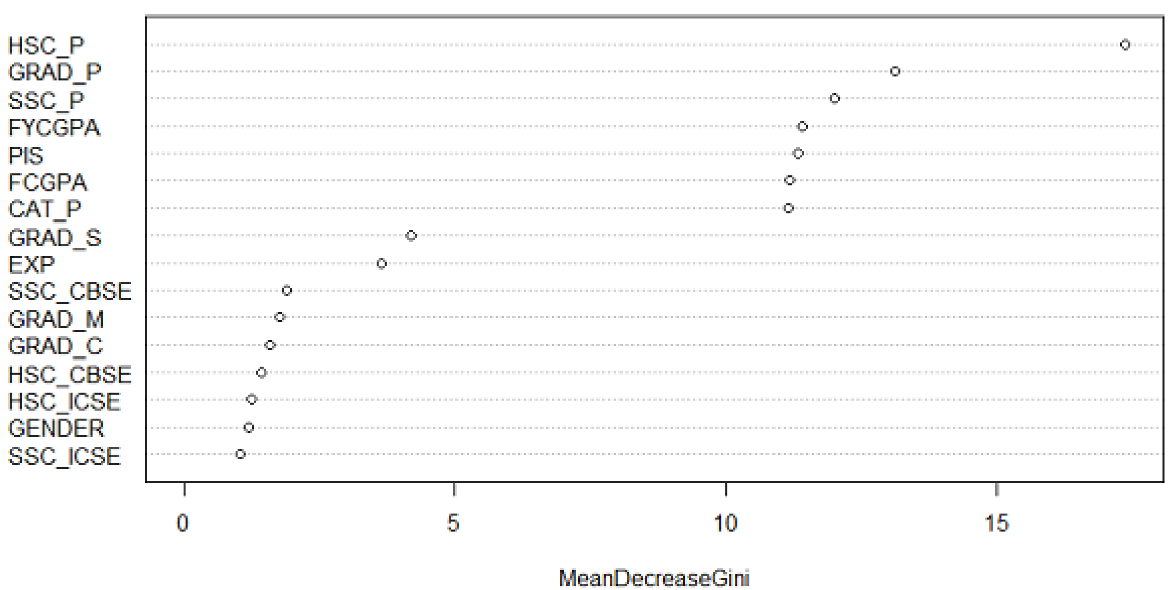

Figure 3: Variance Important Plot

Conclusion

In this study, we used two different techniques to predict/classify a student's success/failure in the program in terms of whether or not he/she will get placed. First we used logistic regression for the classification task. The model found that the percentages obtained in SSC, HSC and an undergraduate degree in science are significant predictors in deciding a student's success. Even though the model provides 80% accuracy in classifying a student's success, it ignores many prominent variables like the CAT percentile, scores obtained in personal interview etc. Secondly, we used random forest- an ensemble classification model and found that the model built is 100% accurate. Here we found that the contributing variables are CAT percentile, personal interview score, first year CGPA and final year CGPA in addition to the scores obtained in SSC, HSC and graduation.

Implication

This study will enable management institutes to take more informed decisions regarding admitting students to the MBA program.

References:

1. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.

2. Acharya, A., and Sinha, D. (2017). An Educational Data Mining Approach to Concept Map Construction for Web based Learning. Informatica Economica, 21(4), 41-58.

3. Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences*, *9*(129), 6415-6426.

4. Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006). Mining student data using decision trees. In *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*.

5. Antunes, C. (2010). Anticipating student's failure as soon as possible. *Handbook for Educational Data Mining*, 353-363.

6. Baker, R. S. J. D. (2010). Data mining for education. *International encyclopedia of education*, *7*(3), 112-118.

7. Berry, M.J.A and Linoff, G. (2004). Data mining techniques: for marketing, sales and customer relationship management, 2nd edition ed. Indianpolis, Ind: Wiley Pub.

8. IBEF (2021). Indian Education Sector in India Industry Report. Retrieved from https://www.ibef.org/industry/education-sector-india.aspx on 13-07-21.

9. Cărbureanu, M. (2012). The Annual Inflation Rate Analysis Using Data Mining Techniques. *Economic Insights-Trends & Challenges*, *64*(4).

10. Chareonrat, J. (2016). Student drop out factor analysis and trend prediction using decision tree. *Suranaree Journal of Science and Technology*, *23*(2), 187-193.

11. Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In: Proceedings of the 5th Future Business Technology Conference, Oporto, Portugal, pp. 5-12.

12. Chye, K. H., Chin, T. W., & Peng, G. C. (2004). Credit scoring using data mining techniques. *Singapore Management Review*, *26*(2), 25-48.

13. Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. *British journal of educational technology*, *46*(5), 904-920.
14. Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining*, Cordoba, Spain, pp. 41-50.

15. Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498-506.

16. Dinesh Kumar, U. (2017). Business Analytics, Wiley.

17. Djulovic, A., & Li, D. (2013). Towards freshman retention prediction: a comparative study. *International Journal of Information and Education Technology*, *3*(5), 494-500.

18. India Today (2020). Studying MBA: Busting the myths and facing the reality of MBA in India. Retrieved from https://www.indiatoday.in/education-today/featurephilia/story/studying-mba-busting-myths-of-mba-reality-of-mba-in-india-1696954-2020-07-04 on 13-7-21.

19. ELAtia, S., Ipperciel, D., & Zaiane, O.R. (2016). Data Mining and Learning Analytics: Applications in Educational Research, Wiley.

20. Espadinha-Cruz, P., Fernandes, A., & Grilo, A. (2021). Lead management optimization using data mining: A case in the telecommunications sector. *Computers & Industrial Engineering*, *154*, 107122.

21. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, (1996). Advances in knowledge discovery and data mining. American Association for Artificial Intelligence/ MIT Press, 1996.

22. Feng, C. X. J., & Wang, X. F. (2004). Data mining techniques applied to predictive modeling of the knurling process. *Iie Transactions*, *36*(3), 253-263.

23. Friedman, J. H. (1998). Data Mining and Statistics: What's the connection? *Computing science and statistics*, *29*(1), 3-9.

24. Gharoun, H., Keramati, A., Nasiri, M. M., & Azadeh, A. (2019). An integrated approach for aircraft turbofan engine fault detection based on data mining techniques. *Expert Systems*, *36*(2), e12370.

25. Groenhof, T. K. J., Koers, L. R., Blasse, E., de Groot, M., Grobbee, D. E., Bots, M. L., ... & Westerink, J. (2020). Data mining information from electronic health records produced high yield and accuracy for current smoking status. *Journal of clinical epidemiology*, *118*, 100-106.

26. Hand, D.J., Blunt, G., Kelly, M.G. and Adams, N.M. (2000), Data mining for fun and profit. *Statistical Science*, *15*(2), 111-131.

27. Hastie, T., Tibshirani, R., & Friedman, J. (2013). The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer.

28. Holst, A. (2021). Amount of Information globally 2010-2024. Retrieved from https://www.statista.com/statistics/871513/worldwide-data-created/ on 4-5-21.

29. Hsu, C. Y., Lim, S. S., & Yang, C. S. (2017). Data mining for enhanced driving effectiveness: an eco-driving behaviour analysis model for better driving decisions. *International Journal of Production Research*, *55*(23), 7096-7109.

30. Hung, J. L., Hsu, Y. C., & Rice, K. (2012). Integrating data mining in program evaluation of K-12 online education. *Journal of Educational Technology & Society*, *15*(3), 27-41.

31. Jain, A.K., Murty, M.N., & Flynn PJ. Data clustering: a review. ACM Comput Surv, 31, 264-323.

32. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2015). An Introduction to Statistical Learning with Applications in R, Springer.

33. Janeja, V. P., Gholap, J., Walkikar, P., Yesha, Y., Rishe, N., & Grasso, M. A. (2018). Collaborative data mining for clinical trial analytics. *Intelligent Data Analysis*, *22*(3), 491-513.

34. Jantawan, B., & Tsai, C. F. (2013). The application of data mining to build classification model for predicting graduate employment. *International Journal of Computer Science and Information Security*, 11(10), *arXiv preprint arXiv:1312.7123*.

35. Kopun, D. (2020). Application of Data Mining Techniques in the Detection of Financial Statement Fraud. *Journal of Accounting and Management*, *10*(2), 97-114.

36. Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data. Information Science & IT Education Conference, 647-665.

37. Larose, D and Larose, C. (2016). Data Mining and Predictive Analytics, Wiley.

38. Li, W. and J.Liao (2011). An empirical study on credit scoring model for credit card by using data mining technology. Proceedings of the 7th International Conference on Computational Intelligence and Security, IEEE, 1279-1282. DOI:10.1109/CIS.2011.283.

39. Liu, P., Qingqing, W., & Liu, W. (2021). Enterprise human resource management platform based on FPGA and data mining. *Microprocessors and Microsystems*, *80*, 103330.

40. Luan, J. (2002). Data Mining and Knowledge Management in Higher Education-Potential Applications. In: Proceedings of AIR Forum, Toronto, Canada, pp. 1-18.

41. Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., & Paris, C. (2015). Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR)*, *47*(4), 1-39.

42. Križanić, S. (2020). Educational data mining using cluster analysis and decision tree technique: A case study. *International Journal of Engineering Business Management*, *12*, 1847979020908675.

43. Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, *53*(3), 950-965.

44. Marconi, M., Germani, M., Mandolini, M., & Favi, C. (2019). Applying data mining technique to disassembly sequence planning: a method to assess effective disassembly time of industrial products. *International Journal of Production Research*, *57*(2), 599-623.

45. Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=15e98f0c60ba on 4-5-21.

46. Márquez- Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, *33*(1), 107-124.

47. Martín- García, A. V., Martínez- Abad, F., & Reyes- González, D. (2019). TAM and stages of adoption of blended learning in higher education by application of data mining techniques. *British Journal of Educational Technology*, *50*(5), 2484-2500.

48. Parack, S., Zahid, Z., & Merchant, F. (2012). Application of data mining in educational databases for predicting academic trends and patterns. In *2012 IEEE international conference on technology enhanced education (ICTEE)* (pp. 1-4). IEEE.

49. Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. *ACM Computing Surveys (CSUR)*, *53*(3), 1-34.

50. Phyu, T. N. (2009, March). Survey of classification techniques in data mining. In Proceedings of the international multi conference of engineers and computer scientists (Vol. 1, No. 5).

51. Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. International Journal of Computer Science Issues: 7(1), 10-18.

52. Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.

53. Shafiabadi, M., Pedram, H., Reshadi, M., & Reza, A. (2021). An accurate model to predict the performance of graphical processors using data mining and regression theory. *Computers & Electrical Engineering*, *90*, 106965.

54. Shen, X., & Qin, R. (2021). Searching and learning english translation long text information based on heterogeneous multiprocessors and data mining. *Microprocessors and Microsystems*, *82*, 103895.

55. Singh, K., & Maiti, J. (2020). A novel data mining approach for analysis of accident paths and performance assessment of risk control systems. *Reliability Engineering & System Safety*, *202*, 107041.

56. Safer, A. M. (2003). A comparison of two data mining techniques to predict abnormal stock market returns. *Intelligent Data Analysis*, *7*(1), 3-13.

57. Salcedo, J. (2019). Machine Learning for Data Mining, Packt.

58. Segall, R. S., Guha, G. S., & Nonis, S. A. (2008). Data mining of environmental stress tolerances on plants. *Kybernetes*, Vol. 37, No.1, 127-148.

59. Shortridge, J. E., Falconi, S. M., Zaitchik, B. F., & Guikema, S. D. (2015). Climate, agriculture, and hunger: statistical prediction of undernourishment using nonlinear regression and data-mining techniques. *Journal of applied statistics*, *42*(11), 2367-2390.

60. Statista.com, 2021. Volume of data/information created, captured, copied and consumed worldwide from 2010 to 2025. Retrieved from https://www.statista.com/statistics/871513/worldwide-data-created/ on 28-07-21.

61. Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, *106*, 234-240.

62. Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). Predicting academic performance by data mining methods. *Education Economics*, *15*(4), 405.

63. Vialardi, C., Chue, J., Peche, J. P., Alvarado, G., Vinatea, B., Estrella, J., & Ortigosa, Á. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *User modeling and user-adapted interaction*, *21*(1), 217-248.

64. Wicker, P., & Breuer, C. (2013). Exploring the critical determinants of organisational problems using data mining techniques: evidence from non-profit sports clubs in Germany. *Managing Leisure*, *18*(2), 118-134.

65. Xu, Y., Yan, C., Shi, J., Lu, Z., Niu, X., Jiang, Y., & Zhu, F. (2021). An anomaly detection and dynamic energy performance evaluation method for HVAC systems based on data mining. *Sustainable Energy Technologies and Assessments*, *44*, 101092.

66. Yoseph, F., Ahamed Hassain Malim, N. H., Heikkilä, M., Brezulianu, A., Geman, O., & Paskhal Rostam, N. A. (2020). The impact of big data market segmentation using data mining and clustering techniques. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-15.

67. Zhang, Y., Oussena, S., Clark, T., & Hyensook, K. (2010). Using data mining to improve student retention in HE: a case study. International Conference on Enterprise Information Systems, Portugal, 1-8.

68. Zheng, Q., Li, Y., & Cao, J. (2020). Application of data mining technology in alarm analysis of communication network. *Computer Communications*, *163*, 84-90.

69. Zhou, H. (2021). Learn Data Mining through Excel, Apress.

70. Zhu, S. (2021). Analysis of the severity of vehicle-bicycle crashes with data mining techniques. *Journal of safety research*, 76, 218-227.

The current pandemic situation did not permit us to execute the offline survey and utilize the approved budget. The data collection was done from Institute of Management , Nirma University.

Budget Details for Minor Research

Budget Head	**Amount Sanctioned**	**Revised Budget**
Books & Journals	30000	27000
Chemical, Glassware and consumables	3000	0.00
Contingencies	5000	85 (Printing cost of paper)
Travel to Field Work	5000	0.00
Total	43000	27085

Purchased Book Details:

Sr. No	**Title of the book**	**Author(s)**	**Publisher**	**Purchase Price**	**Status**
1	Machine Learning for Data Mining: Improve your data mining capabilities with advanced predictive modeling	Jesus Salcedo	Packt Publishing	1,553.26	Book Received
2	Mining Text Data	Charu C. Aggarwal, and Cheng Xiang Z	Springer	7,328.95	Book Received
3	Data Mining and Learning Analytics: Applications in Educational Research	Samira ElAtia , Donald Ipperciel, Osmar R. ZaÃ¯	Wiley	13,033.00	Book Received
	Learn Data Mining Through Excel: A Step-by-Step Approach for Understanding Machine Learning Methods	Hong Zhou	Apress 978-1484259818	591.26	Book Received
4	The Elements of Statistical Learning: Data Mining, Inference and Prediction	Trevor Hastie, Robert Tibshirani and Jerome Friedman	Springer	1919	Book Received

5	An Introduction to Statistical Learning with Applications in R	Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani	Springer	555	Book Received
6	Quick Guide to IBM® SPSS®: Statistical Analysis With Step-by-Step Examples	Alan C. Elliott and Wayne A. Woodward	Sage	1654.8	Book Received
				Total Amount: 26, 635.27	