Paper 4003-2019

Factors Influencing Academic Performance

Onkar Mayekar, Oklahoma State University; Dr. Miriam Mcgaugh, Oklahoma State University

ABSTRACT

There are some patterns and aspects that make some students perform better than their peers in terms of academics. At Oklahoma State University, the administration wanted to identify patterns and factors leading to high academic performance. For this, data consisting of almost 22,000 records and 35 variables were obtained from Institutional Research and Information Management at Oklahoma State University. The dataset consists of student demographics, admission data, email interactions before admission including the interaction messages with the admissions office, athletes' data, gymnasium/recreation center check-in data, student employment data and participation data in various departmental events.

This paper attempts to identify factors and predict the performance of the students based on the attributes of the students including demographics, participation in extracurricular activities and exercise routine of the student. The research paper also talks about the performance of various predictive models, such as logistic regression, ensemble models, neural networks, and decision trees, that will be conducted using SAS Enterprise Miner. This analysis will help university personnel to identify the features that help students perform better and help identify areas in which proactive measures should be taken to boost the performance of the students.

INTRODUCTION

As many students ponder upon their academic performance, there are so many external factors responsible for the performance of the student. The external factors could be participation in extracurricular activities, athletic events, part-time work, or certain workout patterns. If we are able to evaluate certain factors, then students might be able to change their approach towards academic success.

DATA UNDERSTANDING

DATA COLLECTION

The data was collected by the IRIM department at Oklahoma state university. At Oklahoma State University, Institutional Research and Information Management (IRIM) is the department responsible for the collection of data on students. Generally, the student data has many attributes on the students ranging from their academic data, extracurricular participation data, data on sports participation, workout schedules, on campus part time information, high school academic information, and assessment examination information.

DATA CLEANING

The first step was the cleaning of the data using SAS Enterprise Guide and SAS Enterprise Miner. Here, the missing values were treated by either imputing them by two methods:

- Mean, Median and Mode: Imputing the missing values with the overall mean, median or mode is a rudimentary imputation technique. This technique is the only one which doesn't take advantage of the relationship between the variable or the time series characteristics.

- Linear Regression: To start, a number of predictors of the variable with missing values are determined with the help of a correlation matrix. The best predictors are tabbed and used as independent variables in the equation of regression. Here, the variable with missing data is used as the dependent variable and cases with complete data are employed to develop the regression equation; the equation is next applied to predict missing values for incomplete cases.

Outliers in data can falsify estimates and disturb the accuracy, if you don't spot and handle them suitably, particularly in regression models. Once the outliers are identified, they were handled by the approach of prediction. The outliers were replaced with NA values and then were predicted by considering them as a response variable and selecting the best predictors as independent variables in a regression equation.

DATA DESCRIPTION

A crucial component of an analytics project would be to have a sense of data being used for the purpose of generating a model which begets useful patterns and observations. The datasets for the fall 2016 semester were combined with the spring 2017, summer 2017 and fall 2017 semester. Also, various other dataset mentioned above (academic data, extracurricular participation data, data on sports participation, workout schedules, on campus part time information, high school academic information, assessment examination information, ping data) were being used to comprise a whole dataset of variables to be considered for analysis. The major variables being used for the analysis are given below and the variables which have been created from the given data are being highlighted. This was done to draw out more meaning from the data, helping us detect certain patterns through the modeling.

In the below table, the variables highlighted are derived variables from dataset.

Variable	Description
STUDENT_ID	Student's unique ID
COLLEGE	Registered college while enrollment
DEGREE	Registered degree while enrollment
Latest_Major	Registered major while enrollment
Mean_credit_hours	Mean of credit hours taken by the student
credits_10	To identify students taking more than 10 credit hours
credits_11	To identify students taking more than 11 credit hours
credits_12	To identify students taking more than 12 credit hours
credits_9	To identify students taking more than 9 credit hours
Application Entry Term	Application term for the attending the school
Residency_calc	In state, Out of State or International student
OK County of Residence	County of residence if the student is from OK

Non-Resident State of Residence	State of residence if the student is out of state
Postal	This is the zip code of the residence of the student
Birth_Month	Birth month of the student
Gender	Gender of the student
Race	Race of the student
Hispanic	Indicates if the student is of Hispanic heritage
First Generation	Indicates if the student is a first generation student or not
OSU Legacy	Indicates if the student's parents or grandparents have graduated from OSU
Application Student Type	Type of application of student: Freshman, Transfer or a Readmit
Banner Student Type	In-state/Out-state/International student
HS Unweighted GPA	High school unweighted GPA
ENG-A_GPA	High school ENG GPA
MAT-B_GPA	High school MAT GPA
SCI-C_GPA	High school SCI GPA
SS-D_GPA	High school SS GPA
AMH-E_GPA	High school AMH GPA
Core GPA	Cumulative GPA in core courses in degree
Core Units	Credit hours in core courses
Sem	Semester
Per_Month_Colvin_Visits	Visits to recreation center per month
Per_Sem_Colvin_Visit	Visits to recreation center per semester
TOD	General time of day for workout
SUM_of_HOURS_WORKED	Total number of hours worked part time
AVG_of_HOURS_WORKED	Average number of hours worked on a weekly basis in part time
GPA	GPA of the student in degree
GPA_Target	Modified indicator for successful GPA (1 or 0)
worked_on_campus	Indicates whether the student worked on campus or not
WorkedOut	Indicates if a student works out at the recreation center or not

Event_Participation_per_sem	Average participation of the student in extracurriculars per semester
Athetics_participation	Indicates whether a student has participated in Athletics
Event_Participation	Indicates whether a student has participated in extracurricular activities or not

Table 1: Data Dictionary for some of the variables

DATA PREPERATION

Once new variables were created and some part of the data were cleaned up using both MS Excel and SAS Enterprise Guide, the next step was to upload the data into SAS Enterprise Miner. Here, the first act was to import the data using the file import node. Once the file is imported, the data was passed through replacement node, which at a high level allows you to replace/assign or trim values of your data. Another choice would be to replace all missing values throughout the data set with 0 by drawing on the Impute node. This particular node does not extend the level of control over the replacement criteria on a per variable basis as the Replacement node does, but it does cater to other possibly more purposeful options for the replacement value (alternative options than a default constant such as 0).

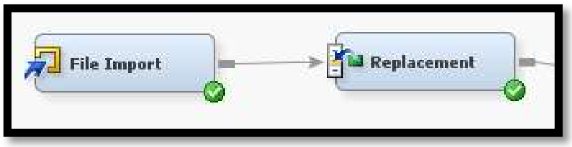


Figure 1: Replacement node for data cleaning

Exploratory analysis showed many variables with missing data and there could be multiple reasons this: data entry errors, transaction system failures, incomplete responses, or many other reasons. If the observation contains a missing value, by default, the whole observation or record is not used by Enterprise Miner for modeling with the regression, variable selection, or neural network nodes. There would be a definite loss of information if we discard incomplete records often which in turn would discard useful information. Also, this would bias the sample since the missing records might have other characteristics in common.

A better means to replace the missing values and treat the missing data would be to use the replacement node. In this project, the replacement node was used particularly for the class variables.

Then, the variables were assigned roles and levels through the metadata node and being connected to the data partition node with a stratified partitioning method and a data split of 70/30 into training and validation datasets.

PREDICTIVE MODELING

The regression and decision tree nodes were used to model the data for the target variable GPA, which is a binary since we have classified GPA into 2 categories: 1 assuming the role of a GPA of greater than or equal to 3.3 (representing the academic success of a student) and 0 for a GPA of below 3.3. This begets a logistic regression technique for modeling and we shall predict the success of a student.

The complete Process flow through the nodes is as follows:

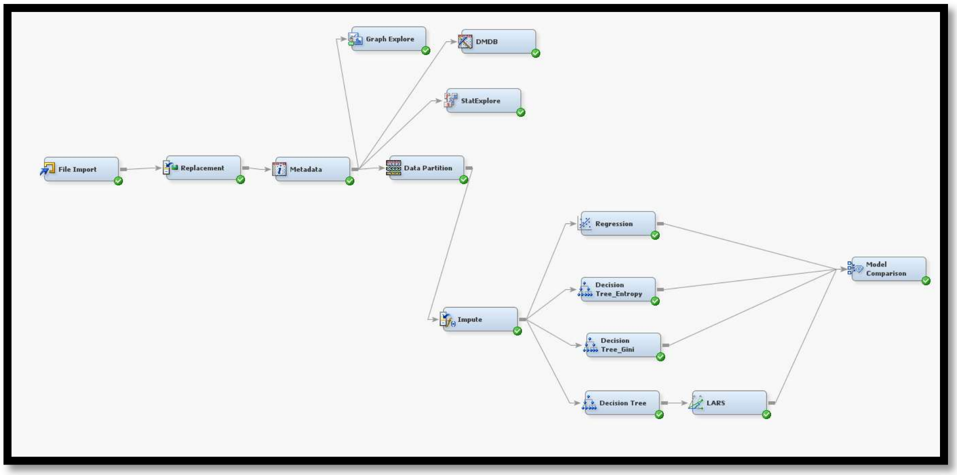


Figure 2: Predictive Model using Enterprise Miner

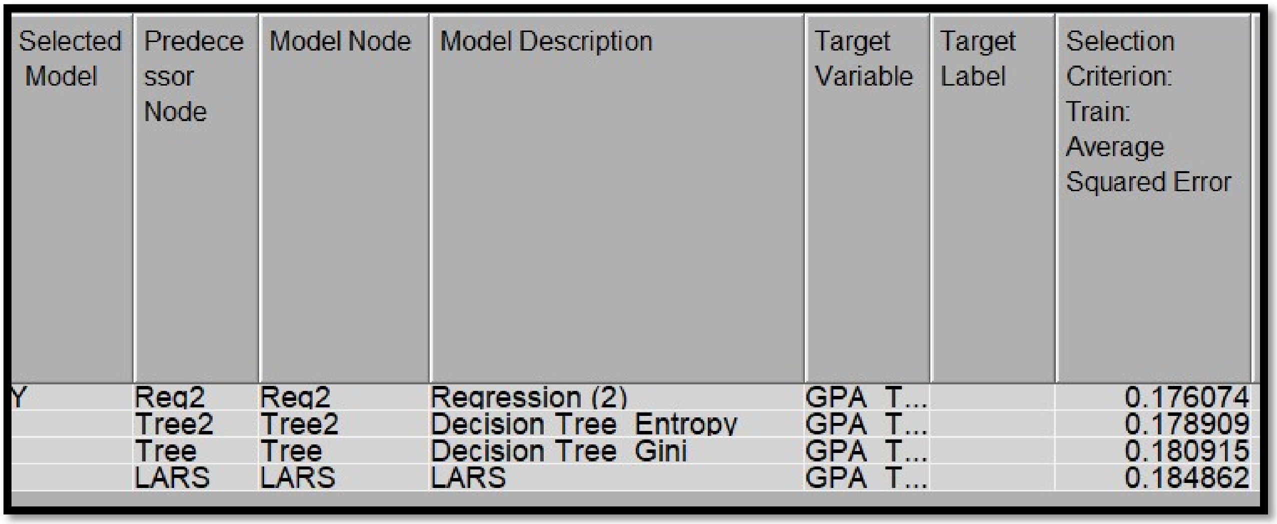
Multiple models were run and evaluated. There were 4 models used over here:

- Regression: The regression model is a statistical operation allowing a researcher to estimate the linear relationship relating to two or more variables. This linear relationship encapsulated the amount of change in one variable that is associated with change in other variables. The model can also be proved for statistical significance, to assess the observed linear relationship.
- Decision Tree with Entropy: Decision tree is an approach for approximating discrete-valued dependent variables, in which the function is defined by a decision tree. It is a decision support tool which uses an if-then structure or a tree-like model their possible results. It is one way to demonstrate an algorithm containing conditional control statements. Entropy is a measure of lack thereof or information which can be calculated by making a split, known as information gain. It is the difference in entropies and measures how you decrease the uncertainty of the label.
- Decision Tree with Gini: Here, we use Gini index as a splitting criterion for the decision tree. Gini assessment is the probability of a random sample being classified incorrectly if we randomly pick a label according to the distribution in a branch.
- Decision Tree with LARS: LASSO (Least Absolute Shrinkage and Selection Operator) penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. The assumptions of this regression are the same as least squared regression except normality is not to be assumed it shrinks coefficients to zero (exactly zero), which certainly helps in feature selection. This is a regularization method and uses ‘L1’ regularization. If a group of predictors are highly correlated, LASSO picks only one of them and shrinks the others to zero.

RESULTS

In linear models, commonly used metric used is the mean squared error (MSE) as the essential measure of fit. The MSE is the sum total of squared errors (SSE) by the degrees of freedom for error. Under the conventional assumptions, this process generates an unbiased estimate of the population noise variance.
For decision trees and neural networks, there is no common unbiased estimator. Additionally, for neural networks, the DFE is generally negative. Although, there are approximations available for the effective degrees of freedom, these are often exceedingly costly and are established on assumptions that might not remain. Thus, the MSE is not nearly as appropriate for neural networks as it is for linear models. One familiar key is to divide the SSE by the number of cases, not the DF which is equal to SSE/N, known as the average squared error (ASE).

The models were compared using the selection criterion Average Squared Error and the model comparison is shown below.



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Train: Average Squared Error
Y	Reg2	Reg2	Regression (2)	GPA T...		0.176074
	Tree2	Tree2	Decision Tree Entropy	GPA T...		0.178909
	Tree	Tree	Decision Tree Gini	GPA T...		0.180915
	LARS	LARS	LARS	GPA T...		0.184862

Figure 3: Model Comparison using ASE

	Condition positive(GPA>3.3)	Condition negative(GPA<3.3)
Predicted condition positive(GPA>3.3)	**TP=2480**	**FP=523**
Predicted condition negative(GPA<3.3)	**FN=547**	**TN=3070**
	Sensitivity=82%	**Specificity=86%**

Table 2. Confusion Matrix

A confusion matrix is frequently used to outline the performance of a classification model and the terms characterizing the matrix are:

- True positives (TP): These cases would be the one in which our predicted yes (they have a success GPA, i.e. GPA>3.3) and are actually true.
- true negatives (TN): These cases would be the one where we predicted no, and they don't have the success GPA (GPA<3.3).
- false positives (FP): ("Type I error.") We predicted yes, but they don't actually have GPA>3.3.
- false negatives (FN): ("Type II error.") We predicted no, but they actually have GPA>3.3.

Based on the above table we can calculate two important metrics:

Sensitivity: Sensitivity denotes how often our model was able to predict the student success when they actually succeeded. Logistic regression model’s sensitivity is 82%.

Specificity: Specificity indicates how often our model was able to predict the GPA<3.3 cases when they actually were less than 3.3. Logistic regression model’s specificity is 86%.

ROC Curve:

A Receiver Operating Characteristic (ROC) Curve is a way to analyze and compare the predictive models. It is a plot of the true positive rate against the false positive rate.
A ROC plot demonstrates:

- The relationship between sensitivity and specificity.
- Test accuracy; The nearer the curve is to the left and top sides of the graph or more the bulge on that graph, better would be the accuracy. For a flawless test, the model plot line would go straight from zero to one, up the the top-left corner and then crossways the horizontal.
- The likelihood ratio; which is found out by taking the derivative at a certain point.

Test accuracy is also shown as the area under the curve. More the area under the curve, better would be the accuracy of the model and a perfect one would have an area under the ROC curve (AUROCC) of 1. The diagonal line in a ROC curve signifies perfect chance. Alternatively, a model that trails the diagonal would have no better odds of prediction than a random flip of a coin. The area beneath the diagonal is .5 (half of the total area) and thus would not be profitable (one that has no better odds than chance alone) has a AUROCC of .5.

For the best model, which is logistic regression in this case, ROC index was found to be 0.81 which is a good metric to work with since the best it could go is 1.

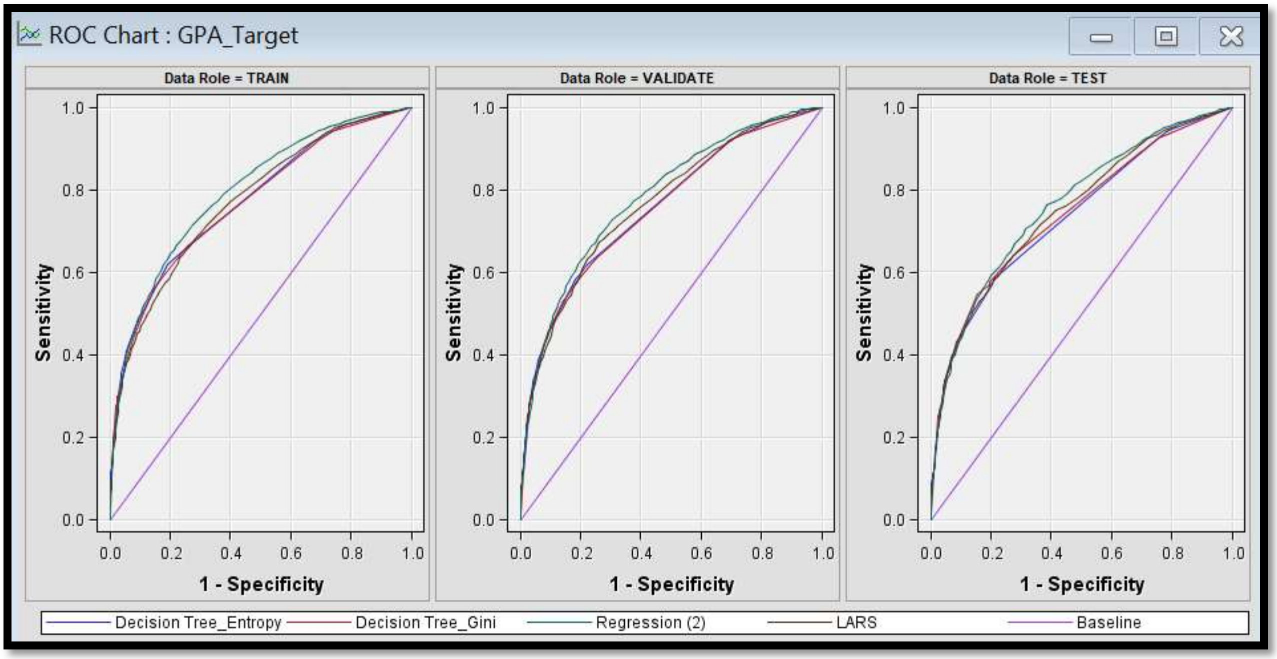


Figure 4: Model Comparison using ROC Curve

The significant variables in the order of their importance according to the Beta coefficients are as follows: High school GPA, English GPA, First Generation, Gender, time of day of workout, per semester recreation center visits, number of credits taken per semester were 10, indicator if student worked on campus.

Limitations

Data is from Oklahoma State University only: This is a drawback since the data could be and wouldn't take into account various other demographic factors in the academic performance of the students.

Detailed workout information is missing in the datasets. The actual timing of the workout sessions, instruments borrowed by students through their ID cards, the type of sports students are playing at the gym, are some of the information which could have added a lot of detail into the analysis.

Data related to breakfast timings at University Dining services could have been an added advantage.

Family conditions: The income of the family also to some extent could be a factor. If we would have had the data, we would have been able to add a lot more information to our analysis.

CONCLUSION

Logistic regression model was chosen as the best model based on the model comparison node and the smallest Average Squared Error and we can make interpretations from the significant variables. It is clear from the results that there are a lot of external factors which may affect the GPA of a student. However, some factors may or may not be in the control of the student. There are certain controllable factors such as the English GPA (signifying proficiency in English language), effective time of workout day, workout visits, part time work affects a student's GPA.

REFERENCES

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, An Introduction to Statistical
LearningJournal

Alice Zheng and Amanda Casari, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists

Elizabeth Covay,
William Carbonaro , After the Bell: Participation in Extracurricular Activities, Classroom Behavior, and Academic Achievement

Chris Schacherer, Paper 540-2013, SAS® Data Management Techniques: Cleaning and transforming data for delivery of analytic datasets

Data Mining Using SAS® Enterprise Miner™ : A Case Study Approach, Fourth Edition

https://irim.okstate.edu/Publications

https://www.statisticshowto.datasciencecentral.com/receiver-operating-characteristic-roc-curve/

https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

ACKNOWLEDGMENTS

This research utilized data from and was supported by Institutional Research and Information Management of Oklahoma State University. We are thankful to them for providing the data for our analysis purposes.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Onkar Mayekar
Master of Science in Business Analytics
+405-614-9906
https://www.linkedin.com/in/onkarm17/
onkar.mayekar@okstate.edu