Value Added Analysis of 4-Year Undergraduate Universities

Sepehr Piri

Robert Bridge

STAT 625

Virginia Commonwealth University

May 2, 2016

Contents

1 Abstract

The U.S. Department of Education's new College Scorecard database (released in 2015) presents a dataset which, for the first time, consists of median salary information for college graduates 10 years after initial admission to their respective undergraduate institution. We use this dataset to build a regression model that will attempt to reveal the value of various college quality indicators and variables in predicting median salary. Then, an attempt is made to create a classification of college "value added", which is based upon residual median salary of our model divided by average institutional tuition (See A.12), to find the best predictor of economic outcomes of a university based upon its particular qualities and tuition.

2 Introduction

Choosing an undergraduate university is often the largest investment decision a person will make in their lifetime, but classifying school quality can be a difficult endeavor. Often, universities are ranked by selectivity of the incoming class, and by various quality indicators which can be difficult to assess. Furthermore, few if any undergraduate ranking systems take into account salary after graduation. While this may not be every prospective undergraduate student's most important decision factor, it is certainly a useful tool.

While simply ranking universities by median salary after graduation is simple and somewhat useful, much is lost in such a simple analysis. Much of what determines the after graduation salary are factors outside of what differentiates a school itself. The percentage of STEM graduates tends to be the largest of such factors, as well as median salary of the parents of incoming students and school location.

In order to rank schools for value while making all other contributors to median salary equal, we built a regression model which attempts to use specific college variables to predict median salary as closely as possible. Using only these indicators, we then can compare the model's predicted salary to the actual salary after graduation, and define "value added" as any additional salary someone makes that is above that predicted by our model. Using this, we can divide these values by tuition to create a more useful indicator of how much value a school contributes to median salary in itself while factoring in the costs of attendance.

Best Prediction Variables Found for Median Salary

count_wne_p102011	Number of non-enrolled non-working students with 10-year post-enrollment earnings data ending in 2011
hbcu	Historically black college or university
catholic	Catholic affiliated college or university
non_religious	College or university does not indicate religious affiliation

pct_Treasury_Cohort2001_2002	Freshman 2001 & 2002 cohort/Number of non-enrolled non-working students with 10-year post-enrollment earnings data ending in 2011
PCT_SubBA_awards2006	For sub-graduate awards granted in 2006, percent granted for sub-bachelor's degree awards
real_fam_inc2001	average family income of students entering in 2001-2002

3 Methods

3.1 Multivariate Regression Analysis

To begin the initial multivariate regression, a dataset of 1220 undergraduate universities was used, which included over 70 variables specific to each institution. We reviewed the literature to bring this to 15 most relevant variables for our initial analysis. Looking over the data initially proved difficult with such a large collection ($n = 1220$), but no glaring errors were found. Some missing values were located, and we decided to use an average of the respective column to fill in these missing values for our regression analysis. Additionally, background research into the types of variables was performed to ensure a thorough understanding of all variables contributing to our dependent variable (median salary 10 years after undergraduate admission).

An initial regression analysis of only 15 variables was performed (see A.1 for estimated model and parameter estimates). The global F-test was significant (F value of 41.36 with p-value of <.0001), but further normality tests failed to meet the assumption. Additionally, individual t tests initially showed that at least 6 of the variables were likely not significant in prediction. One of these, tuition, we knew would likely not contribute to our predicted median salary, as an initial scatter plot of tuition and salary seemed to be random.

As a result of the low p-values in the individual t tests, we decided to perform model selection to reduce the variables needed. We chose to use the PRESS statistic, which is based on the prediction capabilities of the model, to best create a new model. The PRESS statistic uses individual residual values to analyze the prediction capabilities of the model. We used a PRESS macro in SAS called "PRESSALL" to compute the PRESS for all candidate models in a stepwise procedure. The best PRESS model was then found to include only 7 variables (stated in the introduction, results in A.2 including parameter estimates of new model).

The new model now was then tested for multivariate normality, and after a Box-Cox transformation (lambda of 0.4 was found, but 0.5 was within the 95% confidence interval and was used) of the dependent variable, is found to be normal so the assumption is met. The normal probability plot did have some gaps but was overall a fairly straight line and is presented in A.3 along with Box-Cox transformation results..

Next, residual analysis was performed to check for multicollinearity between the variables and any high influence points amongst the observations. Variance inflation factors (VIFS) were found to all be 1.6 or less, indicating multicollinearity was not likely (see A.4). Additionally, the largest condition index was 2.1, which indicated that no likely collinearity exists.

Residual analysis was then continued to detect possible outliers and high leverage points. No large DFFITS were detected (over 2) indicating that no single observation was having a large impact on the fit of the regression model. No large DFBETAS were detected (over 2), indicating no observation was having a large impact on the regression coefficients. As a result, all of the Cook's D values were below 1 and we could conclude that there were no high leverage points, outliers, or high influence points in the dataset. Further COVRATIO analysis was deemed not necessary because of this. We resulted with a usable prediction model that had been chosen for its prediction capabilities that can be used to find predicted median salary values for further classification analysis.

3.2 Classification

For the classification part of this project, Logistic regression and CART were used. These two techniques are considered to be one of the best choices in the classification literature for their ease of interpretation and robustness. In order to evaluate how colleges stand in economics value, we created a new variable named Value_added. First, we divided the data set into half and ran the regression model on the first half and found the residuals. Then taking the average of those residuals and dividing it by average tuition in 2001 we created the cut-off for the classification. After that, in the second half of the original data set, Value_added was created (which was the residual of each observation divided by average tuition in 2011). The colleges which had a "Value_added" above the cutoff were classified as group 1 (good value added) and those below it as group 0 (bad value added).

Classification and regression trees (CART) are machine-learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are used for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values.

Classification results

Logistic Regression Results:

First a stepwise logistic regression was fit to the variables that were selected as significant in the initial regression analysis and looking at the Residual Chi-Square Test, the model appeared to be appropriate (See A.5). The model chose all variables except **count_wne_p102011**, which is the number of students graduated without a job in 2011. Looking at the odds ratio estimates, we realized that **hbcu** which was related to whether the college was historically black or not, had the smallest value among all variables and had a negative role in classifying the colleges as good. For example, for every one unit change in **hbcu**, the college will be less likely (0.09) to be classified as good. On the other hand, **PCT_SubBA_awards2006**, which was the indicator of undergraduate awards granted after graduation (or effectively the graduation rate) had the highest odds ratio estimate (6.224), so one unit changes in this would make the college be six times as likely to be classified as good. (See A.6 and A.7)

CART Results:

Before fitting the CART model, variables were all centered and scaled, since variables were measured in different scales. CART's results were pretty interesting compared to those of the logistic regression. When fitted to the data set, the most important variable was **count_wne_p102011**, number of graduate students without a job in 2011, which was not considered as significant in the logistic model. Also, looking at important variables in splitting the classes, we realized that none of religious or race related variables were considered as important. The second and third important splitters were **real_fam_inc2001,** average family income of students entering in 2001-2002, and **PCT_SubBA_awards2006** which was the indicator of undergraduate awards granted. (See A.9 and A.10)

Comparison of models:

Looking at classification criteria listed below for two methods, one can easily conclude that the CART model seems to be a better method for the classification task. Also the ease of interpretation and graphical illustration of it are its biggest pros. (See A.8 and A.11)

Method	Accuracy	Sensitivity	Specificity
Logistic Regression	71.5	72.7	70.5
CART	79	84	71

4 Discussion

Our regression and CART analysis provide a useful classification tool to find value added of undergraduate public institutions in the U.S. An additional web-based tool could be created with our results for prospective undergraduates to use to determine the value added of universities in their decision making process.

Limitations in our results point to the need for further work and analysis. Initially, our regression model explained roughly 34% of the variation in median income, and after model selection procedures to improve prediction capabilities of the model, this figure was reduced to 25%. While this is still a useful prediction model, a more thorough multivariate model could potentially explain more of the variation in median income and improve our confidence in the classification of value added.

Additionally, further questions we should consider in the classification of value added include how private and public universities compare (our analysis only used 4-year public universities), how tuition changes over time factor into the classification analysis (we only used tuition during the year of admission in real 2014 USD), and how quality of life (or reported life happiness) paired with median salary compare and could be distinguished by university.

References

College Scorecard Data. (2015). Retrieved April 02, 2016, from https://collegescorecard.ed.gov/data/

U.S. Department of Education, National Center for Education Statistics. (2015). *Digest of Education Statistics, 2013* (NCES 2015-011),Chapter 3.

Appendices

Multivariate Regression

A.1

College Data

The REG Procedure
Model: MODEL1
Dependent Variable: real_earn50_p10_2011

Number of Observations Read	1220
Number of Observations Used	1220

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	45516623645	3034441576	41.36	<.0001
Error	1204	88335535293	73368385		
Corrected Total	1219	1.338522E11			

Root MSE	8565.53471	R-Square	0.3401
Dependent Mean	42707	Adj R-Sq	0.3318
Coeff Var	20.05642		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	19725	2466.37064	8.00	<.0001
count_wne_p102011	1	0.26753	0.04387	6.10	<.0001
hbcu	1	-776.65710	1773.04070	-0.44	0.6614
catholic	1	5103.24480	1274.94369	4.00	<.0001
non_religious	1	770.87097	799.81020	0.96	0.3353
pct_Treasury_Cohort2001_2002	1	0.67167	11.34244	0.06	0.9528
Prime_SubBA2006	1	-2632.90822	3119.74124	-0.84	0.3989
PCT_SubBA_awards2006	1	-2919.84920	3202.33607	-0.91	0.3621
real_fam_inc2001	1	0.06580	0.01735	3.79	0.0002
first_gen2001	1	12341	3673.59800	3.36	0.0008
pct_fedloan2001	1	-15.40299	15.27064	-1.01	0.3133
imputed_standard_score2001	1	1997.04207	553.85718	3.61	0.0003
grad_rate2001_2002	1	10989	2313.41915	4.75	<.0001
pct_STEM2006	1	12052	1407.70207	8.56	<.0001
avgfacsal2001	1	1.15180	0.19418	5.93	<.0001
INEXPFTE1	1	-0.01734	0.02411	-0.72	0.4722

A.2

MODELS SORTED BY PRESS

MSE	PRESS	RSQUARE	PREABSUM	CP
458.300	564317.11	0.24910	16440.13	202.985

The REG Procedure
Model: MODEL1
Dependent Variable: YT

Number of Observations Read	1220
Number of Observations Used	1220

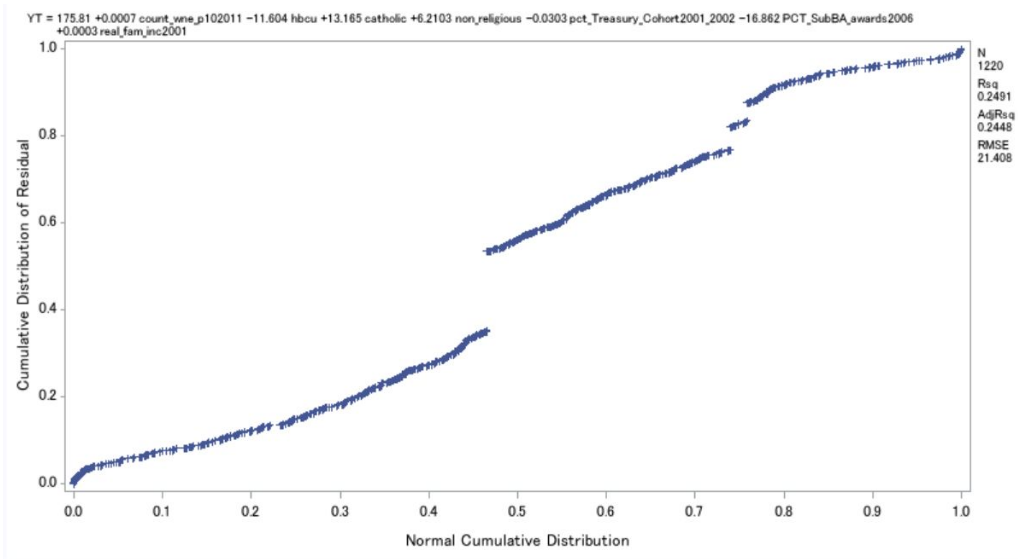
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	184264	26323	57.44	<.0001
Error	1212	555459	458.29986		
Corrected Total	1219	739724			

Root MSE	21.40794	R-Square	0.2491
Dependent Mean	205.18494	Adj R-Sq	0.2448
Coeff Var	10.43348		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	175.76495	2.63042	66.82	<.0001
count_wne_p102011	1	0.00067788	0.00010690	6.34	<.0001
hbcu	1	-11.60418	4.20331	-2.76	0.0059
catholic	1	13.16511	3.17113	4.15	<.0001
non_religious	1	6.21030	1.87576	3.31	0.0010
pct_Treasury_Cohort2001_2002	1	-0.03032	0.02807	-1.08	0.2803
PCT_SubBA_awards2006	1	-16.86204	1.71708	-9.82	<.0001
real_fam_inc2001	1	0.00032344	0.00002518	12.85	<.0001

A.3

Multivariate Normality (Normal Probability Plot)



WHICH LAMBDA MAXIMIZES MAX_LOG_LIKEKIHOOD?

LAMBDA	MAXIMUM OF MAX_LOG_LIKELIHOOD
0.4	−11063.38

LIMITS OF CONFIDENCE INTERVALS SUCH THAT .90=<CONF. COEFF.<=.99

IF YOU NEED AN APPROXIMATE .95 CONFIDENCE INTERVAL, CHOOSE TWO LAMBDA VALUES FOR LOWER AND UPPER LIMITS FROM BELOW SUCH THAT

THE INTERVAL CONTAINS THE LAMBDA MAXIMIZING MAX_LOG_LIKELIHOOD
AND
THE CONF. COEFF. IS NEAR .95, OR,
IF YOU WANT TO BE CONSERVATIVE, MINUMUM AMONG THOSE>=.95.

LAMBDA VALUES FOR CONF. INTERV. LIMITS	MAX_LOG_LIKELIHOOD AT GIVEN LAMBDA	DIFF. BETW. MAX. MAX_LOG_L. & MAX_LOG_L.	CONF. COEFF. OF INTERVAL W/ GIVEN DIFF.
0.31	−11064.87	1.48486	0.91516
0.49	−11065.12	1.73894	0.93781
0.28	−11066.07	2.68787	0.97958
0.52	−11066.42	3.04237	0.98636

A.4 Residual Analysis

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	175.81040	2.64618	66.44	<.0001	0
count_wne_p102011	1	0.00067788	0.00010690	6.34	<.0001	1.60456
hbcu	1	−11.60418	4.20331	−2.76	0.0059	1.07891
catholic	1	13.16511	3.17113	4.15	<.0001	1.25831
non_religious	1	6.21030	1.87576	3.31	0.0010	1.33467
pct_Treasury_Cohort2001_2002	1	−0.03032	0.02807	−1.08	0.2803	1.57163
PCT_SubBA_awards2006	1	−16.86204	1.71708	−9.82	<.0001	1.07329
real_fam_inc2001	1	0.00032344	0.00002518	12.85	<.0001	1.08873

Collinearity Diagnostics (intercept adjusted)									
			Proportion of Variation						
Number	Eigenvalue	Condition Index	count_wne_p102011	hbcu	catholic	non_religious	pct_Treasury_Cohort2001_2002	PCT_SubBA_awards2006	real_fam_inc2001
1	1.76913	1.00000	0.09506	0.00032502	0.07534	0.10014	0.08889	0.01843	0.02244
2	1.46466	1.09903	0.09256	0.01081	0.10466	0.07244	0.10137	0.07535	0.04958
3	1.23379	1.19745	0.00148	0.38923	0.00052733	0.02682	0.00236	0.09755	0.22350
4	0.88265	1.41575	0.00105	0.00599	0.20332	0.03732	0.01236	0.62908	0.11441
5	0.72785	1.55905	0.00058623	0.55721	0.00868	0.00185	0.00437	0.10182	0.58992
6	0.53214	1.82333	0.00602	0.03635	0.59657	0.71525	0.03770	0.06114	0.00013514
7	0.38978	2.13043	0.80324	0.00008947	0.01091	0.04619	0.75295	0.01662	0.00001881

Logistic Regression and CART Output:

A.5 Goodness of fit

Looking at the Residual Chi-Square Test, the P-value is large enough to fail to reject the null hypothesis and conclude that the logistic model is appropriate**.**

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	842.853	742.423
SC	847.266	773.317
-2 Log L	840.853	728.423

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	112.4303	6	<.0001
Score	101.3494	6	<.0001
Wald	88.5180	6	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
0.4259	1	0.5140

A.6 The coefficient estimates:

All variables are significant at 0.1 significance level.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.7464	0.4943	30.8651	<.0001
hbcu	1	-2.4136	0.8648	7.7890	0.0053
catholic	1	-1.4348	0.8089	3.1460	0.0761
non_religious	1	-1.5581	0.3401	20.9933	<.0001
pct_Treasury_Cohort2	1	0.00419	0.00233	3.2433	0.0717
PCT_SubBA_awards2006	1	1.8284	0.2218	67.9825	<.0001
real_fam_inc2001	1	-0.00003	4.87E-6	40.6336	<.0001

A.7 Odds ratio estimates:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
hbcu	0.089	0.016	0.487
catholic	0.238	0.049	1.163
non_religious	0.211	0.108	0.410
pct_Treasury_Cohort2	1.004	1.000	1.009
PCT_SubBA_awards2006	6.224	4.030	9.613
real_fam_inc2001	1.000	1.000	1.000

A.8 Classification performance of the logistic model:

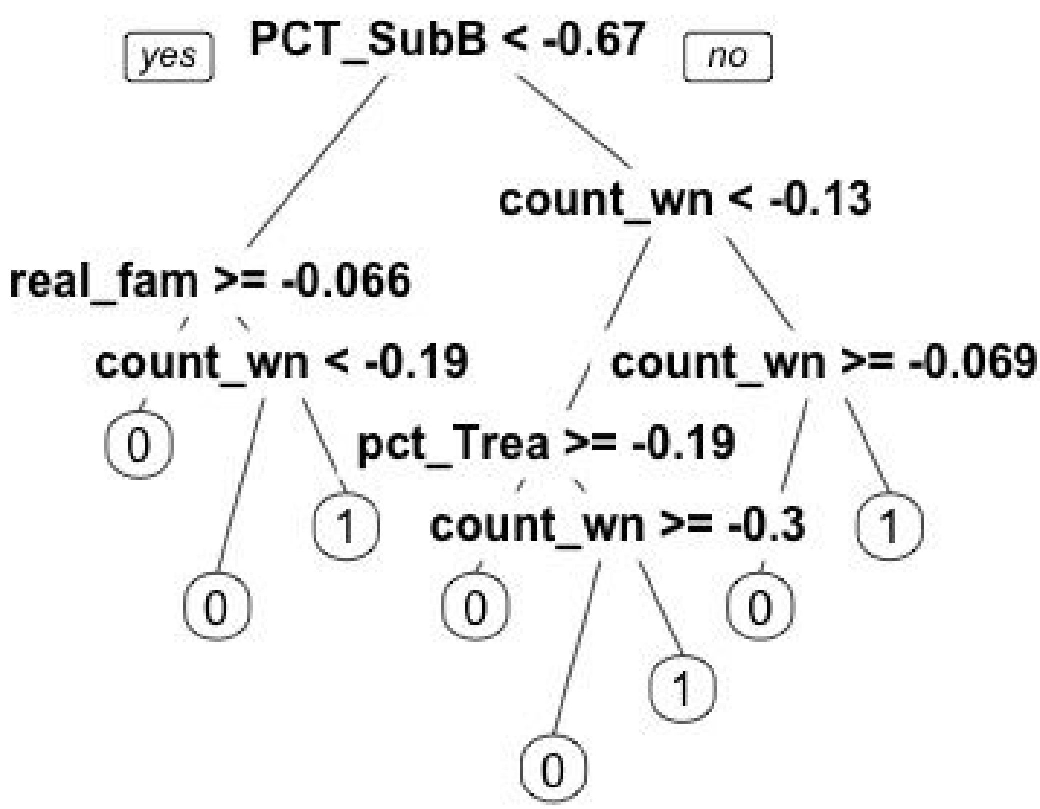
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	73.7	Somers' D	0.522
Percent Discordant	21.5	Gamma	0.548
Percent Tied	4.8	Tau-a	0.259
Pairs	92296	c	0.761

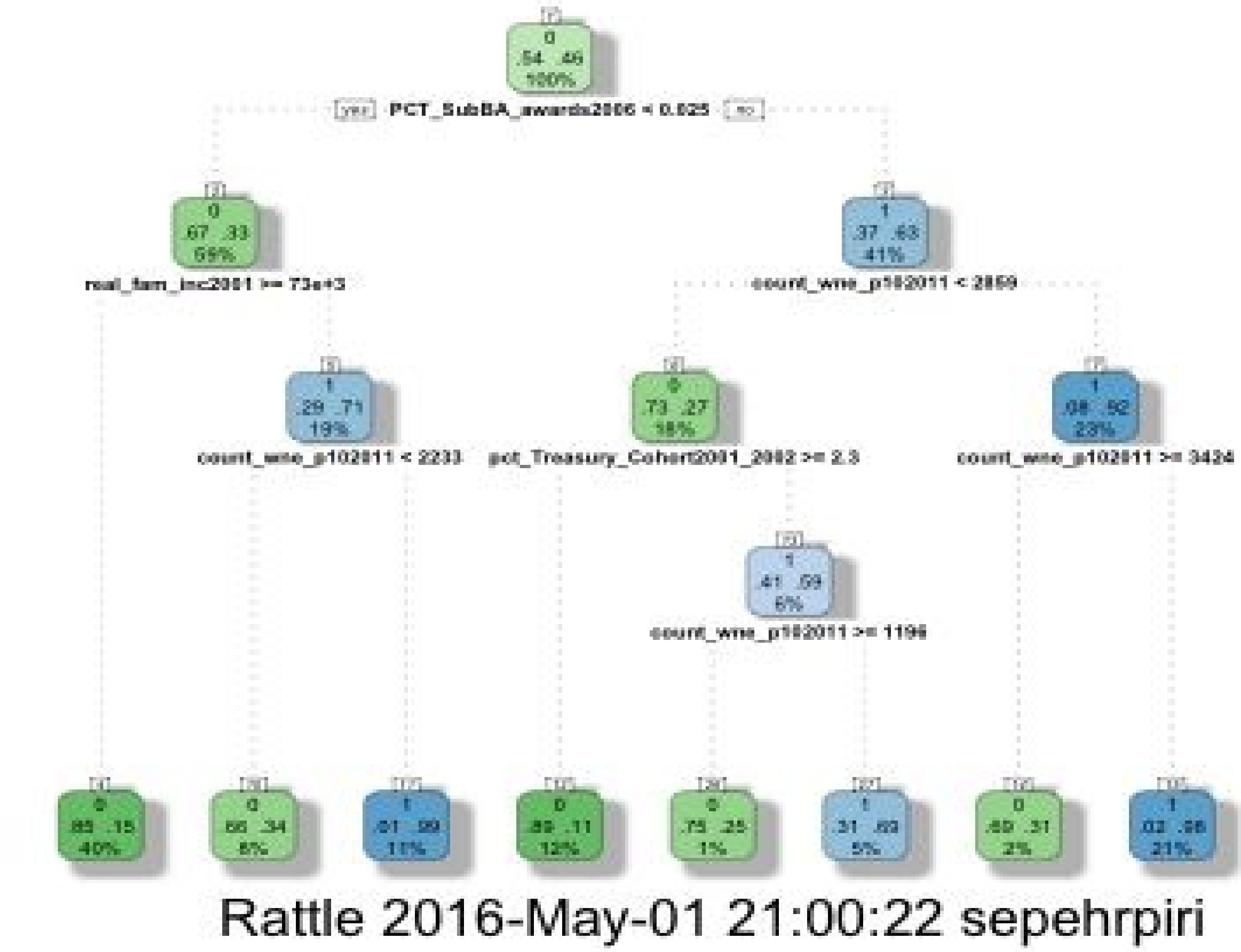
Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	202	234	98	76	71.5	72.7	70.5	32.7	24.5

A.9 Variable importance table of CART model

```
                              myrpart1.variable.importance
count_wne_p102011                                 96.80453
real_fam_inc2001                                  78.37404
PCT_SubBA_awards2006                              47.33032
pct_Treasury_Cohort2001_2002                      45.67052
catholic                                          25.42929
hbcu                                              25.41887
non_religious                                     22.32611
```

A.10 Graphical illustrations of the CART model




yes
PCT_SubBA_awards2006 < 0.025
no
real_fam_inc2001 >= 73e+3
count_wne_p102011 < 2859
count_wne_p102011 < 2233
pct_Treasury_Cohort2001_2002 >= 2.3
count_wne_p102011 >= 3424
count_wne_p102011 >= 1196
Rattle 2016-May-01 21:00:22 sepehrpiri

A.11 Classification performance of the CART model:

```
> confusionMatrix(myrpart.predict1,mydata1$group)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 320  66
         1  61 163

               Accuracy : 0.7918
                 95% CI : (0.7574, 0.8234)
    No Information Rate : 0.6246
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5541
 Mcnemar's Test P-Value : 0.7226

            Sensitivity : 0.8399
            Specificity : 0.7118
         Pos Pred Value : 0.8290
         Neg Pred Value : 0.7277
             Prevalence : 0.6246
         Detection Rate : 0.5246
   Detection Prevalence : 0.6328
      Balanced Accuracy : 0.7758
```

A.12 Average tuition of public 4 year institutions:

SOURCE: U.S. Department of Education, National Center for Education Statistics. (2015). *Digest of Education Statistics, 2013* (NCES 2015-011), Chapter 3.

Average total tuition, fees, room and board rates charged for full-time undergraduate students in degree-granting institutions, by type and control of institution: Selected years, 1982–83 to 2012–13

Year and control of institution	Constant 2012–13 dollars[1]			Current dollars		
	All institutions	4-year institutions	2-year institutions	All institutions	4-year institutions	2-year institutions
All institutions						
1982–83	$9,138	$10,385	$6,396	$3,877	$4,406	$2,713
1992–93	12,097	14,216	6,830	7,452	8,758	4,207
2001–02	14,775	17,708	7,424	11,380	13,639	5,718
2002–03	15,262	18,344	7,943	12,014	14,439	6,252
2003–04	16,104	19,276	8,336	12,953	15,505	6,705
2004–05	16,647	19,925	8,563	13,793	16,510	7,095
2005–06	17,014	20,289	8,412	14,634	17,451	7,236
2006–07	17,547	20,934	8,461	15,483	18,471	7,466
2007–08	17,737	21,160	8,346	16,231	19,363	7,637
2008–09	18,421	21,996	8,879	17,092	20,409	8,238
2009–10	18,839	22,515	9,109	17,649	21,093	8,533
2010–11	19,355	23,118	9,323	18,497	22,092	8,909
2011–12	19,741	23,409	9,461	19,418	23,025	9,306
2012–13	20,234	23,872	9,574	20,234	23,872	9,574