SESUG Paper 252-2018

ANALYSIS OF FACTORS INFLUENCING DROPOUTS IN SOPHOMORE ENROLLMENT

Apoorva Chandrasekaran, Oklahoma State University;

ABSTRACT

Bill Gates called U.S colleges' staggering dropout rates as 'tragic' in his blog post about 'Putting students first'. The United States is leading in terms of the number of people who start college but is lagging far behind in terms of the number of people who actually finish college. As a matter of fact, only about 50% of these students really leave the college with a diploma. This situation can be remedied. Most of the students that dropout, usually do so before entering the sophomore year. This analysis is based on data from Oklahoma State University, Institutional Research and Information Management (IRIM) Department. In this analysis, we leverage the power of predictive analytics and SAS® 9.4 to predict whether or not a student will enroll for their sophomore year.

It will take into consideration various influential factors such as financial background, scholarships, education history, family background, university services use while in college, student employment data and athletic status among others to determine students who are at-risk. Data preparation and predictive models will be handled by Base SAS® 9.4 and SAS® Enterprise Miner™ 14.2. Preliminary findings show that variables such as enrollment in the previous two semesters and college of enrollment heavily influence whether a student will enroll in their sophomore year. This study hopes to provide an opportunity for institutions to detect and help at-risk students at an early stage, by determining the course of action they can take in order to avoid dropping out of college.

INTRODUCTION

Oklahoma State University was founded on Christmas day, 1890, and has grown to become one of America's premier land-grant universities. Research is one of three essential components of Oklahoma State University's land-grant mission and brings richness and depth to their teaching and outreach missions.

To facilitate and promote the pursuit, discovery, and dissemination of new knowledge and technologies through research, scholarship, creative activities, and technology transfer for the benefit of the people of the state of Oklahoma, the nation, and the world. OSU's faculty and staff are engaged in research across the full spectrum of human endeavor and inquiry, including areas of state, regional, and national importance. Institutional Research and Information Management at OSU provides information, research, decision support, and analysis on demand to the OSU community and others and effectively manages institutional performance. IRIM continues to make consistent efforts to make sure that students are provided with all resources to succeed in their academic careers.

This paper relies on data from IRIM for Fall 2016 and Spring 2017. SAS® 9.4 was used for preparing the data that will be used by the model. Data was prepared by merging the information for these 2 terms and making sure that there were no duplicate records. Once all the information was merged with the base dataset with student information, the final data source was used in SAS® Enterprise Miner™ 14.2 for modelling phase. The best model was chosen based on the model's ability to correctly classify the students who will enroll in the upcoming Fall term. Future scope of this project is to you use this model to score the data for the fall enrollments for this year.

METHODOLOGY

DATA UNDERSTANDING:

IRIM is the Institutional Research Department at OSU and the department collects information from students, every time they sign up for events, enroll for classes, swipe their ID card to make payments

or any other reason, use university facilities such as Colvin recreational center, websites that are browsed by students etc. Below datasets were provided by the IRIM department to look into the UG retention rates:

- Admissions data – This dataset included all students that applied for all terms. This dataset had 45000+ records and around 50 variables.
- Athletes' data – This dataset included information on students who represented college in different sports. Dataset had 1000 records and 4 variables
- Colvin data – This dataset included visit information on students who accessed the university's recreational facility. This dataset had almost 330000 records and 3 variables
- Email data – This dataset included information on emails sent to students and the status of the emails. This dataset had 240000+ records and 6 variables
- Event data – This dataset included information on all events that the student registered for and whether they attended or not. This dataset had almost 54000 records and 5 variables
- Fall and Spring data – These datasets included information on all enrollments for respective terms and contained 7500+ records and 9 variables
- Interactions data – This dataset included all interactions with or without messages that were both student initiated and non-student initiated. The dataset contained about 450000+ records and 5 variables
- Student Employee Hours data – This dataset included all information regarding student employees who were working in the university services. This dataset contained around 134000 records and 7 variables
- Financial Aid – This dataset included information regarding all kinds of financial aids received by students, the source of these funds, offer amount and demographics of students. This dataset contained around 245000 records and 10 variables
- Zip code data – This dataset was downloaded from the web with information on Oklahoma zip codes and the average household income in each of the zip codes, the number of tax returns filed under each of these returns, the number of returns with High Gross Income and Low Gross Income etc.

Most of the data manipulation was carried out using SQL and Transpose procedures and Data steps. Sample of some of the codes is below:

```
proc sql;
create table irim.Enrollment as
    Select A.*, b.* from Enrolled_5 as A left join employmenthrs_2 as b
    on A.student_id = b.student_id;
quit;

proc transpose data=irim.finaid4 out=irim.finaid5 prefix=AidGiven_;
    by student_id ;
    id academic_period;
    var Fin_Aid_Type;
run;

data irim.funds;
set irim.finaid2;
    if fund_source_type_desc = "Institution" Then Institution = 1;
    else Institution = 0;
    if fund_source_type_desc = "Federal" Then Federal = 1;
    else Federal = 0;
    if fund_source_type_desc = "State" Then State = 1;
    else State = 0;
    if fund_source_type_desc = "Other" Then Other = 1;
    else Other = 0;
    if academic_period in ("201660","201720");
run;
```

After data was organized to create the final data source, this SAS® Dataset was used as the data source in the Enterprise Miner™ project for modelling purposes.

DATA CLEANING AND VALIDATION:

All the SAS codes that were used to prepare the data for modelling can be found in APPENDIX A.

All datasets were received as SAS datasets. Many variables were declared as different data types across different datasets which created issue while trying to append the data for different terms. Most of the information that came in for Colvin recreational center, events, interactions, athletes, etc., were highly skewed as the number of students enrolled in a term was far higher than the number of students taking part in extracurricular activities. Therefore, all this information was converted into nominal and binary variables such as students using Colvin more than average was coded as “Above average” and students using Colvin less than average was coded as “Below Average”. Similarly, students who initiated interactions were coded as 1 to indicate that they have initiated at least 1 interaction and 0 for students who did not initiate any interaction. Same logic was applied to students who attended events.

Spring 2017 and Fall 2016 data was appended and demographic information from applications data was used to finally arrive on data that included all students who enrolled in Spring 2017 and Fall 2016. Target variable was created by adding a flag on whether a student was enrolled in the upcoming Fall term (Fall 2017) or not. This made the target variable binary (1 – Enrolled in Fall 2017, 0 – Did not enroll). Another variable was created to indicate students who enrolled only in Spring 2017/Fall 2016 and students who enrolled in both the terms. This variable was created to see how much did the enrollments influence the Target variable.

Financial aid information was further used to create summary columns such as Aids that were given for the two terms and the source of funds for these aids. Upon exploratory analysis, it was seen that these two variables were moderately correlated (Appendix B). Therefore, fund source was dropped from the final data source. Aid given for the two terms shows whether the aid was loan, scholarship, grant or work or any combination of these options. There were many variables with more than 50% missing values. All these variables were dropped from the final dataset in order to avoid inducing bias in the data. Data was initially partitioned using 70% training and 30% validation ratio. Missing data was imputed using mean for interval variables and count for class variables.

EXPLORATORY ANALYSIS:

All visualizations for the exploratory analysis can be found in Appendix B.

After preparing the final dataset, initial exploration was carried out to see how the variables were related to the target variable. Although seemingly obvious, it was interesting to note that students who enrolled in both the semesters were more likely to enroll in the third semester as opposed to students who enrolled in only one semester in the first year. According to reports from National Student Clearinghouse (NSC) a third of the students who dropped out of school, dropped out before the start of their sophomore year. An initial model showed that the prior enrollments of these students heavily influenced whether a student will enroll in the upcoming fall or not. Further analysis on demographic information of students showed that the legacy status, gender and race did not have drastic effects on the target variable. However, the number of hours a student enrolled in the first two terms seemed to have an impact on the target variable. In the sense that, it showed that more hours enrolled in the first two terms meant that the students are more likely to enroll for the third semester.

Additionally, information on Income Tax Returns based on zip codes was pulled from IRS.gov website to see whether the number of high income returns filed, the number of returns were education expenses were claimed etc. influenced whether a student enrolled in school. A brief analysis showed that the income groups did not necessarily mean enrollments will be high. It also showed that zip codes with more number of returns that claimed educator expenses and student loan deductions also had more enrollments. Reports suggest that tax credits provide a benefit for students and families already enrolled. However, it did not necessarily increase education in new students. Oklahoma provides multiple

scholarship programs that provide financial aids to students in need. One of them being Oklahoma's promise program that encourages students to study for free in public schools in Oklahoma by covering most of their tuition amount if the family AGI is less than $100,000 in a year.

ANALYSIS AND RESULTS

Analysis on the final data source was carried out in 3 steps as laid out below in Figure 1.

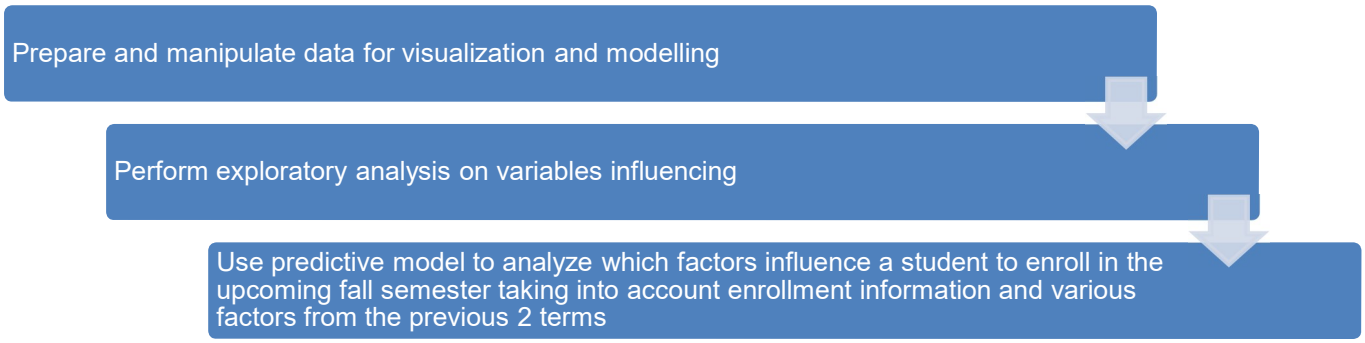

Figure 1. Steps in creating the model

Table 1 describes the data used in the modelling phase:

Variable	Description
AidGiven_201660	Aid Given – Scholarship, Grant, Loan, Work
AidGiven_201720	Aid Given – Scholarship, Grant, Loan, Work
Application_Student_Type	Indicates whether a student was enrolled as a freshman, readmit, concurrent or transfer
Avg_hh_income	Avg. Income per household in the zip code
Banner_Admitted_Residency	Residency status of the student; In-State, Out of state, International
Educator_Expenses	Number of returns in each zip code with educator expenses deductions on the returns (Only for Oklahoma as most applicants where from Oklahoma)
First_Generation	Indicates whether the student has family who graduated from OSU; 0-No, 1- Yes
Fundsource_201660	Source of funds – Federal, Institute, State and Other
Fundsource_201720	Source of funds – Federal, Institute, State and Other
Gender	Student's gender; M, F
HighAGIReturns	Number of returns filed with Adjusted Gross Income higher than $75000 per annum
Hispanic	Yes, No
LowAGIReturns	Number of returns filed with Adjusted Gross Income lower than $75000 per annum
OSU_Legacy	Whether student's parents graduated from OSU

Variable	Description
Postal_New	Postal Code of the student's home
Race	American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White
STUDENT_ID	Unique ID
colvin_flag	Indicates usage of Colvin; Do not use, Below Average, Above Average
enrolled_status	Indicates which terms the students were enrolled in Spring 2017, Fall 2016 or in both terms
event_flag	Indicates whether the student attended events organized by the college; 0 – Did not attend any events, 1 – Attended at least one event
College	College under which student is getting their degree; AG = Agriculture and Natural Resources, AS = Arts and Sciences, ED = Education, Health and Aviation, EN = Engineering, Architecture and Technology, HS = Honors College, GR = Graduate College, SB = Spears School of Business, UC = University College, VM = Center for Veterinary Health Services
irim_total_hrs	Hours enrolled in courses
term_residency	Residency for the term; I = International, N= Non-Resident, R= Resident
fall_status	Status of enrollment for Fall 2017
hs_gpa	High school GPA
interactions_flag	Indicates whether the student initiated any interaction
student_loan_deduction	Number of returns in each zip code where student loan deduction was claimed
total_hours_worked	Number of student worker hours
tuition_fees_deduction	Number of returns in each zip code where tuition fees deductions were claimed

Table 1. Data Description of Final Data Source

This data was put into Enterprise Miner, partitioned, imputed and modelled. These models were later compared and the best model was picked using the Model Comparison Node and Misclassification Rate as the selection criterion.

The model diagram is shown below in Figure 2

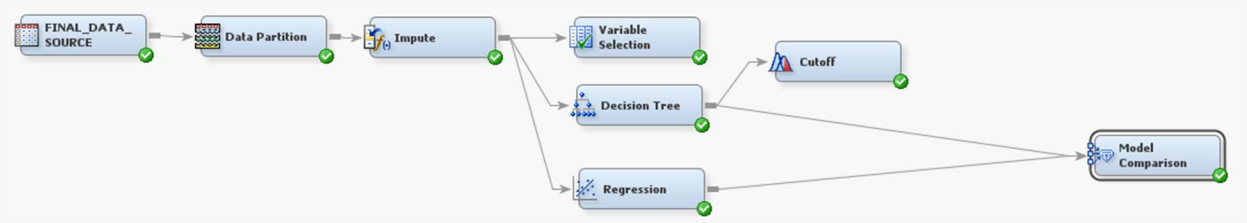

Figure 2. Enterprise Miner Model Diagram

Data partition node was used to split the data into training and validation datasets in the ratio of 7:3. Impute node was used to impute missing values for class and interval variables using count and mean options respectively. All variables except postal code was imputed.

Variable Selection node was used to understand which variables were selected by the software before incorporating business sense. The top 3 variables from the node were the college of student, number of hours enrolled and type of application. These were followed by Aid given, high school GPA, Colvin, Events, Banner Student Type and First Generation. Results of the variable selection node (Figure in Appendix B, Figure B-5). Banner Student Type was ignored from the final data source during modelling as the variable was highly correlated to Application type variable. Number of hours enrolled was also ignored as we could see in Figure B-4 that when a student enrolled for more hours in the first two terms, they were likely to continue school than the students who enrolled for fewer hours. In this case, average hours enrolled by students who did not enroll for next fall was around 11 to 11.5 while the average hours enrolled by students who enrolled for next fall was around 13 to 13.5.

MODEL OUTPUTS

Decision Tree Output:

Decision Tree settings were changed to make sure that same variable is not used as splitting criterion more than once and the assessment measure was changed to misclassification rate.

Decision Tree splitting criterion: The first splitting rule was made on College variable which was also our most important variable according to variable selection node. The split can be seen below in Figure 3. The model split Arts and Sciences and Graduate College students under one node. It showed that 42% of the students, graduating from these colleges were not enrolling for the next fall while 80% of the students graduating from the other colleges were enrolling for next fall.

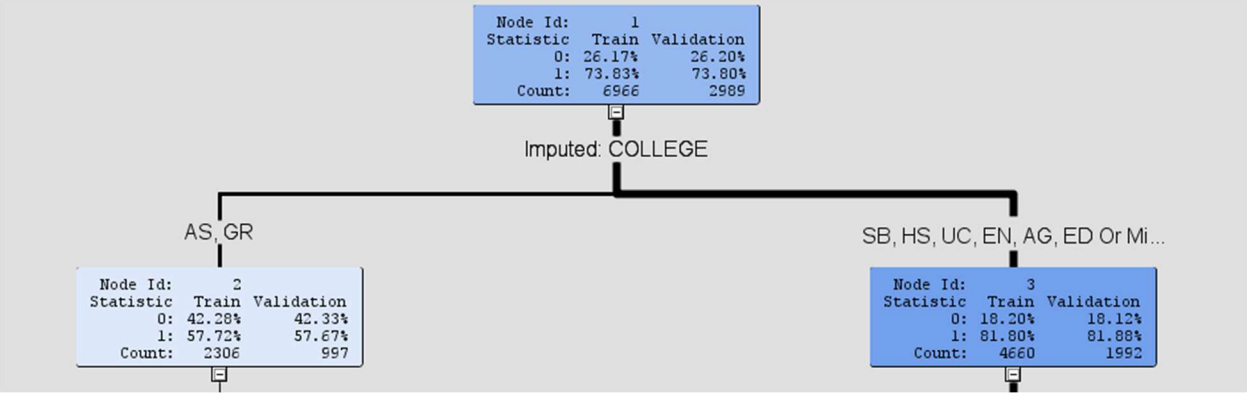

Figure 3. First Splitting Rule – Decision Tree Output

The left hand side node was further split with the variable Aid Given. This variable had 15 combinations of different kinds of aids that a student received in the first two terms (Scholarship, Grant, Loan, Work). Model combined all students who did not receive any aid along with students that received both grant and loan, and students that received grant, loan and work aids. Model showed that 55% of these students did not enroll for the next fall. In simple terms, it means that students that did receive some aid in the first two terms were more likely to enroll for the next fall than students who receive no aid at all.

The right hand side node was further split with the variable application student type. In this node, model shows that 85% of the transfer and freshman students enrolled for the next fall than the readmits or non-degree seeking students. Figure 4 shows the second splitting rule from the Decision Tree.

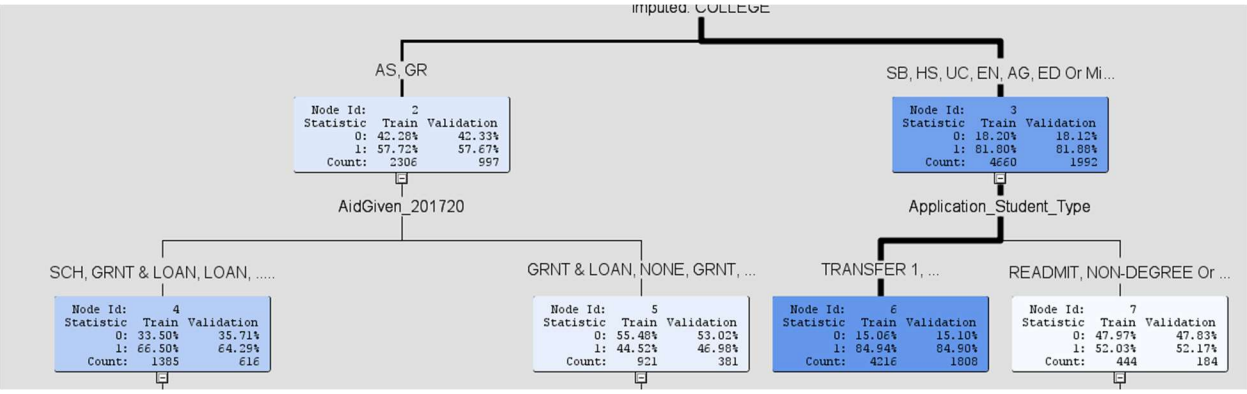

Figure 4. Second Splitting Rule – Decision Tree Output

Model Fit Statistics:

The Decision Tree model had a training misclassification rate of 23% and validation misclassification rate of 24%, sensitivity of 93%, specificity of 27% and accuracy of 74% at probability of 0.5. Since we are trying to predict at risk students, increasing specificity will make more business sense. Hence, by increasing the probability cut off ratio to 0.7, we get specificity of 65% and sensitivity of 70%. This was identified using the cutoff node.

Logistic Regression Output:

Model properties were changed to select model based on validation misclassification rate using the stepwise method for selection.

Selected model effects: The final model selected based on the misclassification rate on the validation data had the following variables - Aid given, application student type, college, high school GPA and Colvin usage.

Odds ratio interpretation: The odds ratio interpretation for some of the levels are below:

Colvin usage – Odds of fall enrollment for students who use Colvin (Above Average) is 48% higher than those who do not use. Four percent higher for students who use Colvin (Below Average) in comparison to the ones who do not use.

College – AG college student has 18% lesser odds for fall enrollment when compared with a VM college student while Arts and Sciences student has 80% lesser odds and Honors college students have 28% lesser odds of fall enrollment.

Aid Given – Without any aids, the odds of a student enrolling for next fall is 30% lower than that of a student that receives scholarship, grant or work aid.

HS GPA – For each unit increase in GPA, the odds of a student enrolling in the next fall semester goes up by 110%.

Application Student Type – A freshman student has 2.6 times more probability of enrolling for next fall in comparison to a transfer student.

Model Comparison Node:

A model comparison node was added to select the best model based on validation misclassification rate. Regression was chosen as the best model based on this selection criterion. However, the misclassification rates for both the models were extremely close (24%) so we can choose the final model based on other factors. Based on cumulative lift chart, we can see that the regression model had a higher lift value than Decision Tree. At depth of 40, regression model was 1.3 times better than the baseline while Decision Tree was about 1.1 times better than the baseline model in predicting the target variable. Cumulative lift chart can be seen below in Figure 5.

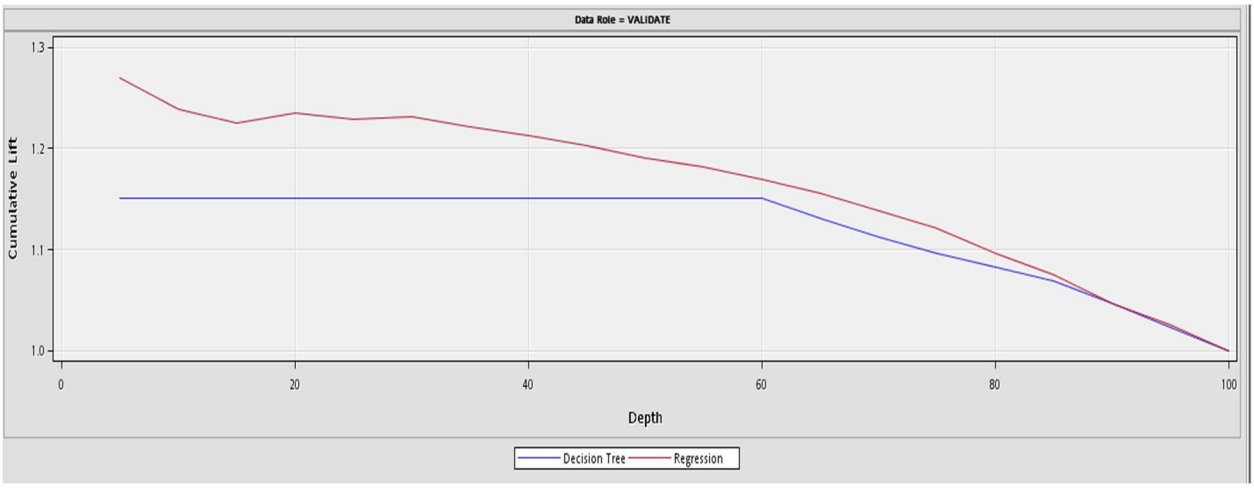

Figure 5. Cumulative Lift Chart for validation dataset

Based on the sensitivity and specificity as well, we can see that the regression model is performing slightly better than the Decision Tree. Figure 6 shows the ROC chart for training and validation datasets.

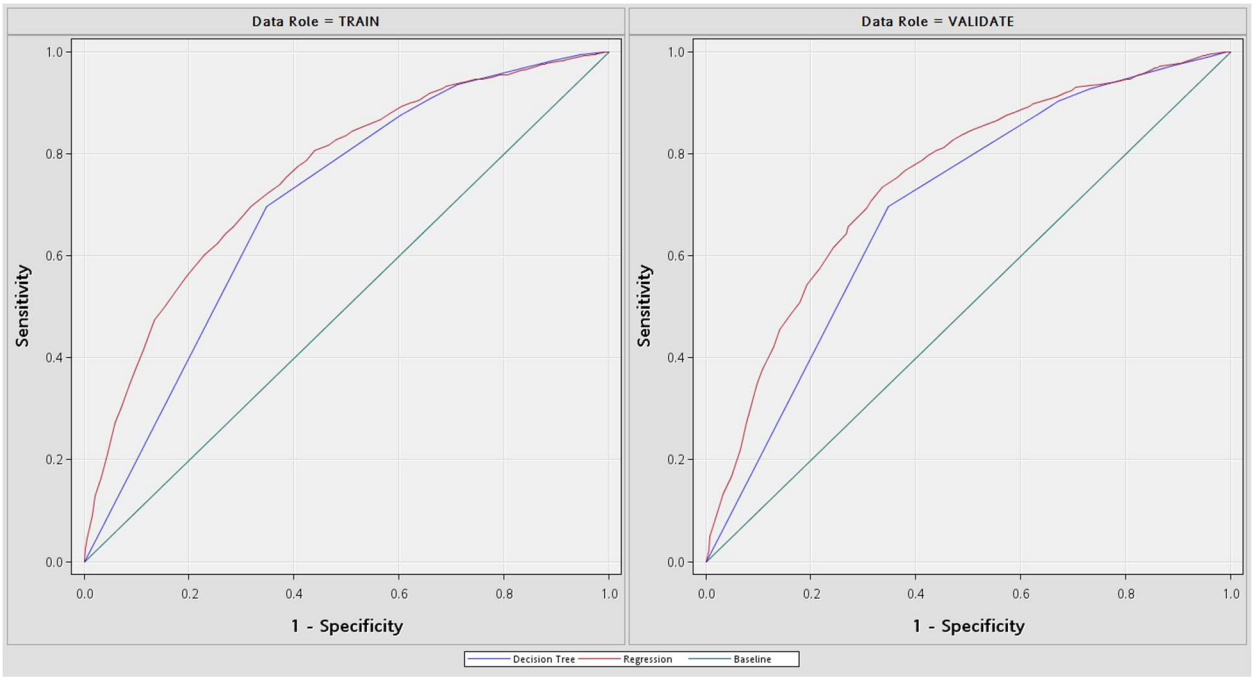

Figure 6. ROC Chart for training and validation datasets – Model Comparison

CONCLUSION

This paper is a result of stressing on the importance of student retention in undergraduate colleges. This paper strives to equip universities with information on which students require attention. Future scope for this project is to focus on what pages the students browse on university website and use this information as well to predict if they are planning to change majors or drop out. These models can also be used to help students figure out their career path. However, the main focus of the paper is to provide help to IRIM to identify at risk students and start coming up with measures to prevent them from dropping out.

Colleges can collect more information from students on a day to day basis to understand what drives them to continue schooling. Surveys are a great way to understand what works for these students and to get a feel of their expectations. Several federal statistics revealed in the past few years have brought up the fact that almost 50% of the students do not graduate. This can be due to many reasons such as financial need, lack of time, personal situations etc. Student aids can be explored further to see if students have awareness about the various scholarships that are made available to them. It is important for colleges to figure out what can help in motivating the students to graduate as improving the number of graduates, which helps to increase employment and that in turn helps the economy.

REFERENCES

(1) Aston, Esme. January 16, 2018. "Why Students Drop Out of College, and How We Can Do Something About It" Huffingtonpost.com

(2) Oklahoma Promise Scholarship Program for College Students. Available at https://www.okhighered.org/okpromise/college-faq.shtml

(3) Oklahoma State University Research Department. More information available at: https://research.okstate.edu/about-osu-research.html

(4) Oklahoma State University Institutional Research information available at https://irim.okstate.edu/

(5) Leonhardt, David. June 13, 2015. "Bill Gates, College Dropout: Don't Be Like Me". New York Times. Available at https://www.nytimes.com/2015/06/04/upshot/bill-gates-college-dropout-dont-be-like-me.html

(6) Bureau of Labor Statistics, Research data for demographics, https://www.bls.gov/bls/blswage.htm.

(7) Census.gov, Income data tables for research, https://www.census.gov/topics/income-poverty/income/data/tables/cps.html

(8) Internal Revenue Services, individual tax information by zip code, https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi

(9) Median Household income by zip code, Oklahoma income research data , http://zipatlas.com/us/ok/zip-code-comparison/median-household-income.2.htm

ACKNOWLEDGMENTS

I would like to thank Dr. Goutam Chakraborty, SAS® Professor of Marketing Analytics and Dr. Miriam McGaugh, Clinical Professor at Oklahoma State University for their continuous guidance and motivation.

CONTACT INFORMATION

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

Apoorva Chandrasekaran
Oklahoma State University
2245092956
apchand@okstate.edu
www.linkedin.com/in/apchand

APPENDIX A

```
/*Dropping variables with more than 50% missing values in admissions
dataset*/
data admissionsnew;
set irim.admissions_data;
keep STUDENT_ID   Round 'Application Entry Term'n     'Banner Admitted
Residency'n 'OK County of Residence'n
'OK County of Residence1'n    'Non-Resident State of Residence'n  Region
      Postal      Birthdate   Gender      Race
Hispanic 'First Generation'n  'OSU Legacy'n 'Application Student Type'n
'Decision History (confirmed)'n
'Banner Student Type'n  'Application Fee Date'n 'Date Submitted'n 'Source
First'n
'HS Unweighted GPA'n ;
run;

/*Merging admissions dataset with Fall 2016*/
data fall2016;
set irim.fall2016 (rename =(term_residency = f16term_residency irim_total_hrs
= f16irim_total_hrs college = f16college
degree = f16degree major = f16major first_concentration =
f16first_concentration academic_period = f16academicperiod
academic_period_desc = f16academic_period_desc));
run;
proc sql;
create table NewFall2016 as
select a.*, b.* from admissionsnew as a left join fall2016 as b
on a.student_id = b.student_id;
quit;

data enrolled excluded;
length enrolled_status $32.;
set NewFall2017;
if f17academic_period_desc ne "" then fall_status = "1";
else fall_status = "0";
if f16academic_period_desc ne "" and s17academic_period_desc ne "" then
      do enrolled_status = "BothF16&S17";
      output enrolled;
      end;
else if f16academic_period_desc ne "" and s17academic_period_desc = "" then
      do enrolled_status = "OnlyF16";
      output enrolled;
      end;
else if f16academic_period_desc = "" and s17academic_period_desc ne "" then
      do enrolled_status = "OnlyS17";
      output enrolled;
      end;
else output excluded;
run;

/*Merge athletes information with enrolled*/
```

```
Proc sql;
Create table Enrolled_1 as
Select A.*, b.S17_Sports, b.F16_Sports from Enrolled as A left join
Athletes_3 as b
on a.student_id = b.student_id;
quit;

/*Merging Events information with enrolled*/
Proc Sql;
Create table Enrolled_2 as
Select A.*, b.Events_Attended, b.Events_NotAttended from enrolled_1 as A left
join event_3 as b
on a.student_id = b.student_id;
quit;

/*Merge interactions data with enrolled*/

Proc sql;
Create table Enrolled_3 as
select a.*, b.no_of_SIInteractions from enrolled_2 as A left join
interactions_1 as b
on a.student_id = b.student_id;
quit;

Proc sql;
Create table Enrolled_4 as
select a.*, b.no_of_SIInteractionsNM from enrolled_3 as A left join
interactions_2 as b
on a.student_id = b.student_id;
quit;
/*Ping Data*/

data ping;
set irim.ping_data;
drop timestamp ua;
Landing_Page = Scan(url, 2, '//');
Landing_Page2 = Scan(url, 3, '//');
Landing_Page1 = Scan(Landing_Page2 , 1 , '?,@,#, ,.,%');
If Landing_Page = 'fqdn3.dasnr.okstate.edu' Then Main_Page =
Scan(Landing_Page, 2, '.');
Else If Landing_Page = 'www.google.com' Then Delete;
Else if Landing_Page = 'www.bing.com' Then Delete;
Else if Landing_Page = 'humansciences2.okstate.edu' Then Main_Page =
'humansciences';
Else if Landing_Page = 'webcache.googleusercontent.com' Then Delete;
Else if Landing_Page = 'www.oklahomaproven.org' Then Main_Page = 'Oklahoma
Proven';
Else Main_Page = Scan(Landing_Page, 1, '.');
Drop landing_page2;
run;

/*Merging Colvin data with Enrolled*/

proc sql;
create table Enrolled_5 as
Select A.*, b.colvinf16_visits, b.colvins17_visits from Enrolled_4 as A left
join colvin_2 as b
```

```
on A.student_id = b.student_id;
Quit;

/*Merging Employment information with enrollment and exporting final file*/
proc sql;
create table irim.Enrollment as
Select A.*, b.* from Enrolled_5 as A left join employmenthrs_2 as b
on A.student_id = b.student_id;
Quit;
data irim.enrollment_final4(drop = OSU_Legacy
rename=(OSU_Legacy1=OSU_Legacy)) ;
set irim.enrollment_final3;
if OSU_Legacy in ("1","Yes") then OSU_Legacy1="1";
else if OSU_Legacy in ("0","No") then OSU_Legacy1="0";
else OSU_Legacy1="";
Postal_New = scan(Postal,1,'-');
run;
/*Include ZipCode financial status information*/

proc sql;
create table irim.enrollment_final7 as
select a.*, b.'Avg. Income/H/hold'n as Avg_hh_income from
irim.enrollment_final6 as a left join work.zipcode_data1
as b
on a.postal_new = b.zipcode_new1;
quit;
```

APPENDIX B

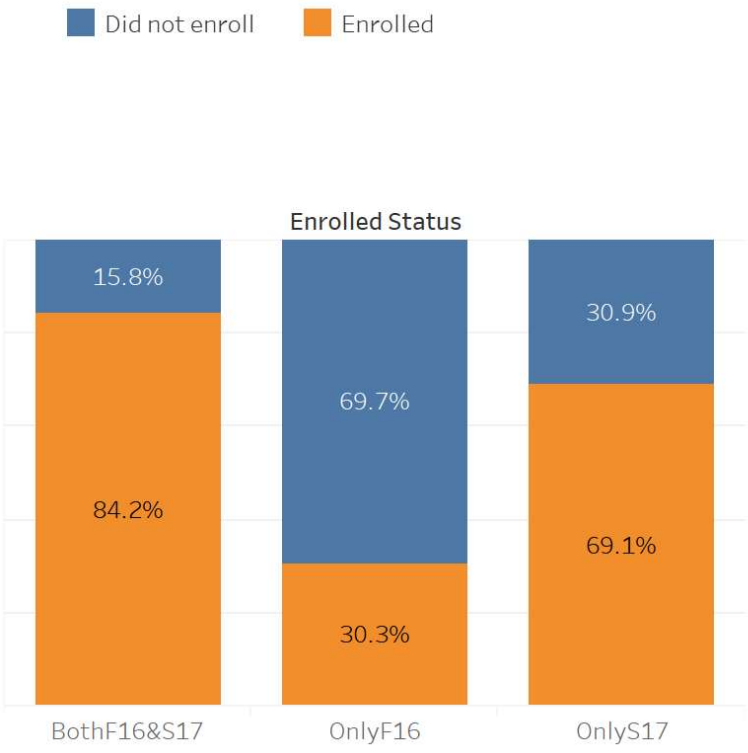

Figure B-1 Enrollment status for Fall 2017 against previous enrollments

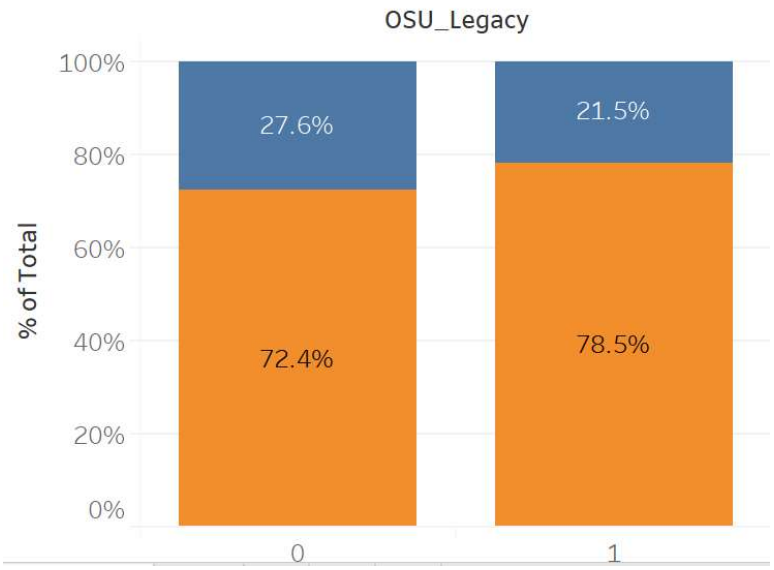

Figure B-2 Enrollment status over students' OSU Legacy

Figure B-3 Enrollment status over Gender

Fall_Enrollment	AvgF16_Hrs	AvgS17_Hrs
Did Not Enroll	11.6	10.9
Enrolled	13.2	13.6

Figure B-4 Enrollment status for Fall over the enrollment hours for the two terms

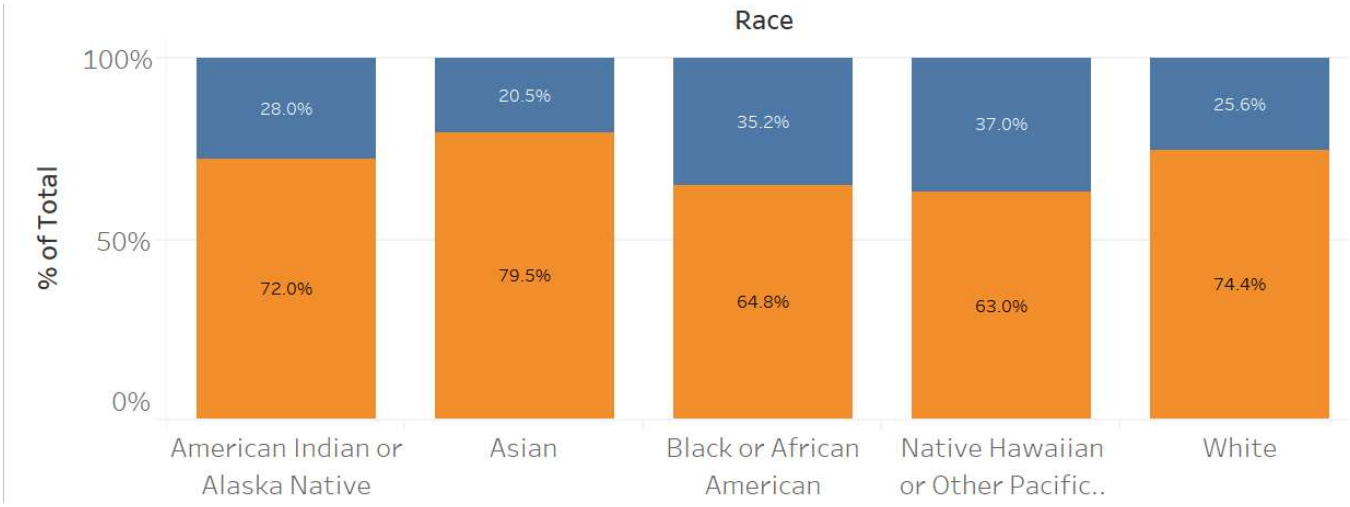

Figure B-5 Enrollment status over Race

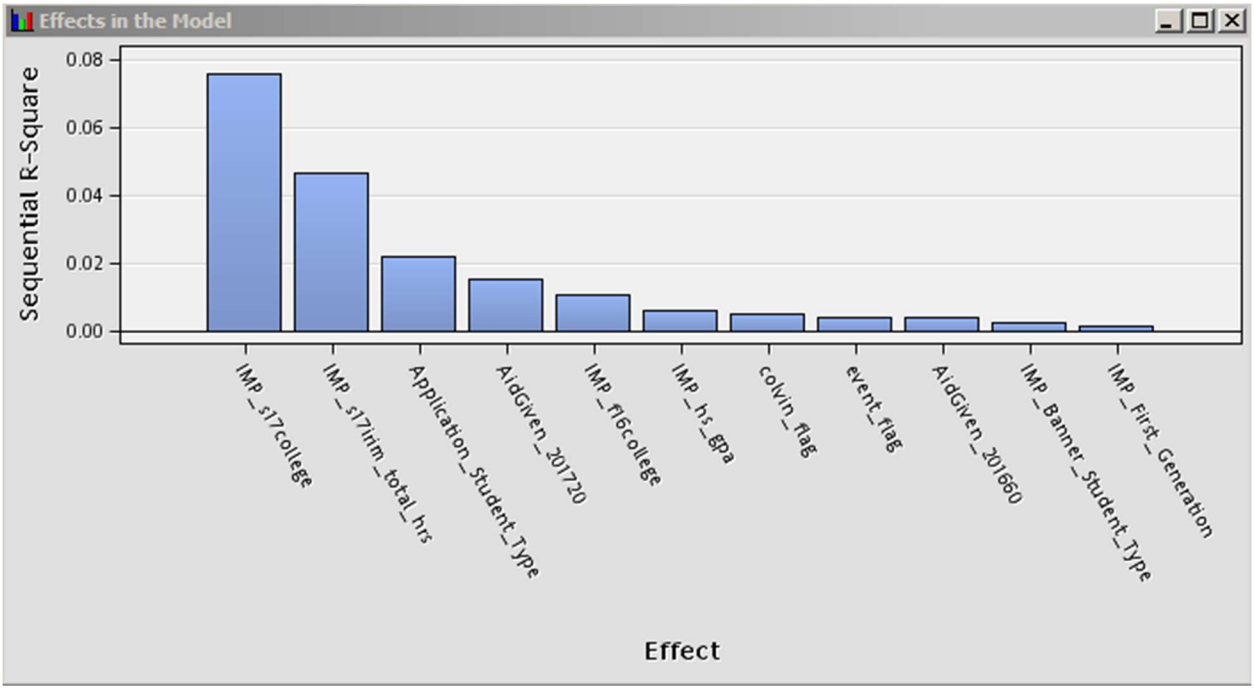

Figure B-6 Variable Selection Node Results