CAPSTONE PROJECT IN PYTHON - STUDENT EARLY ATTRITION IN CLEAR WATER STATE UNIVERSITY

Project Report

Parthasarathy R
Jig19073

Contents

- Problem Statement
- Understanding the data
- Data Analysis
  - Dependent variable
  - Independent variables
    - Demographic data
    - Performance data
    - Financial data
- Model Building
- Model Validation
- Key drivers of Attrition
- Recommended interventions

Problem statement

- Clearwater State University offers a wide variety of degree programs.

- By leveraging data on student profiles, the following business questions are to be answered

  - Identify key drivers of early student attrition
  - Build a predictive model to identify students with higher early attrition risk
  - Recommend appropriate interventions based on the analysis

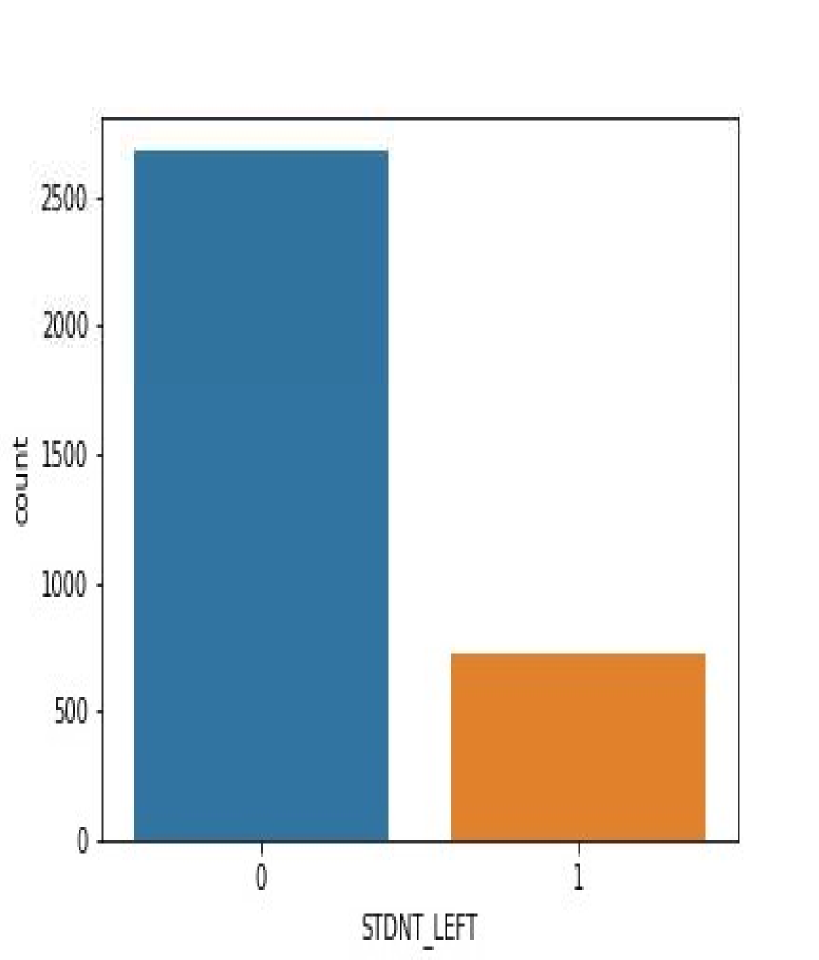
Understanding the data

- The data containing the student demographic profile, course preferences, performance record, grades, financial background, financial aid and other application information is read into jupyter notebook and basic sanity checks are made.

- Variables having more than 20% missing values are removed from the dataset.

Data Analysis – Dependent variable

Dependent Variable



- The variable 'RETURNED_2ND_YR' gives information on whether the student returned for the second year or not.

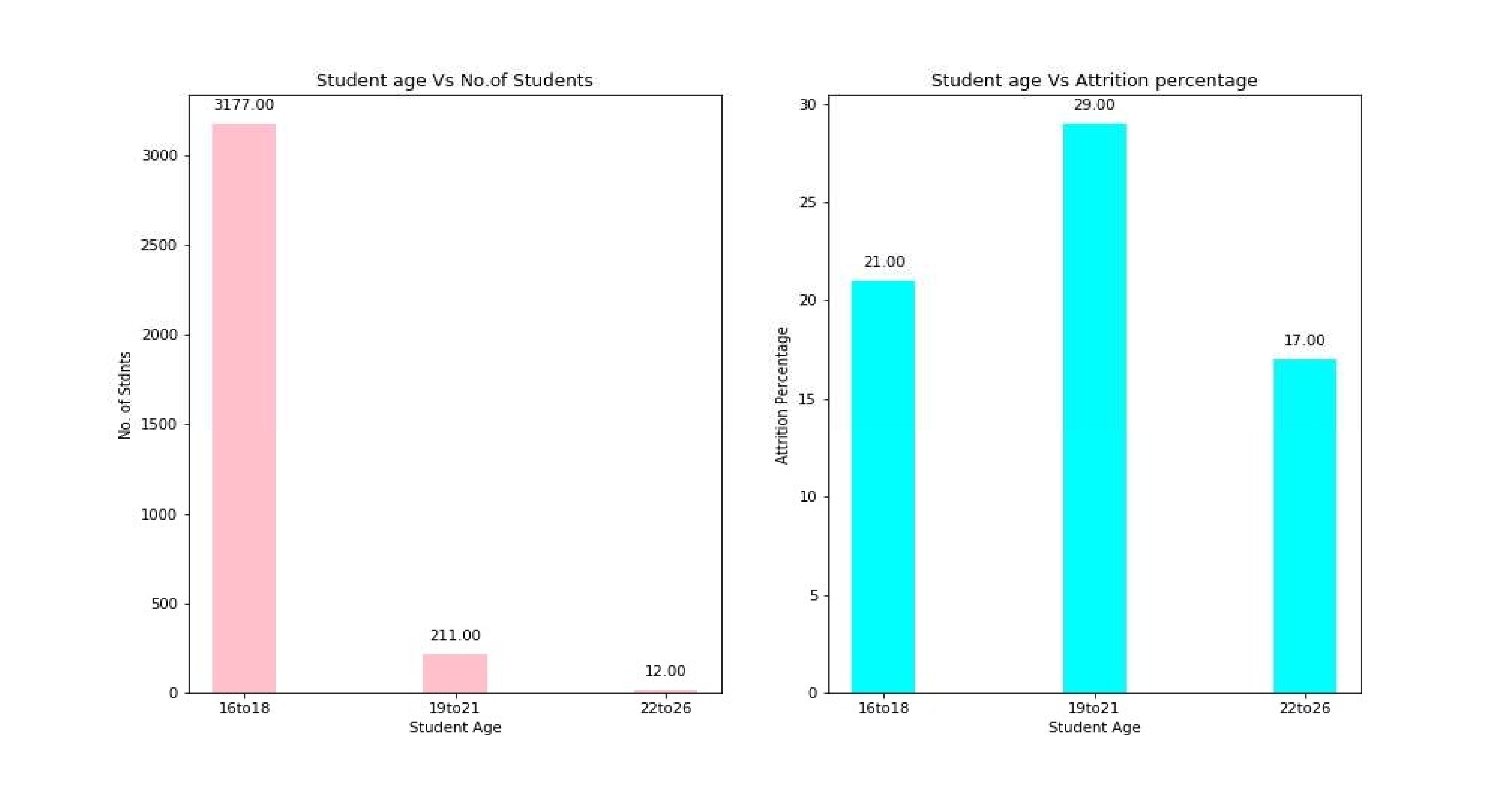
- This values of this variable are reversed to form a new variable 'STDNT_LEFT'. This variable is our target/dependent variable.

- The number of students left after first year is 723 (21.26%)

- The number of students continuing studies is 2677(78.74%)

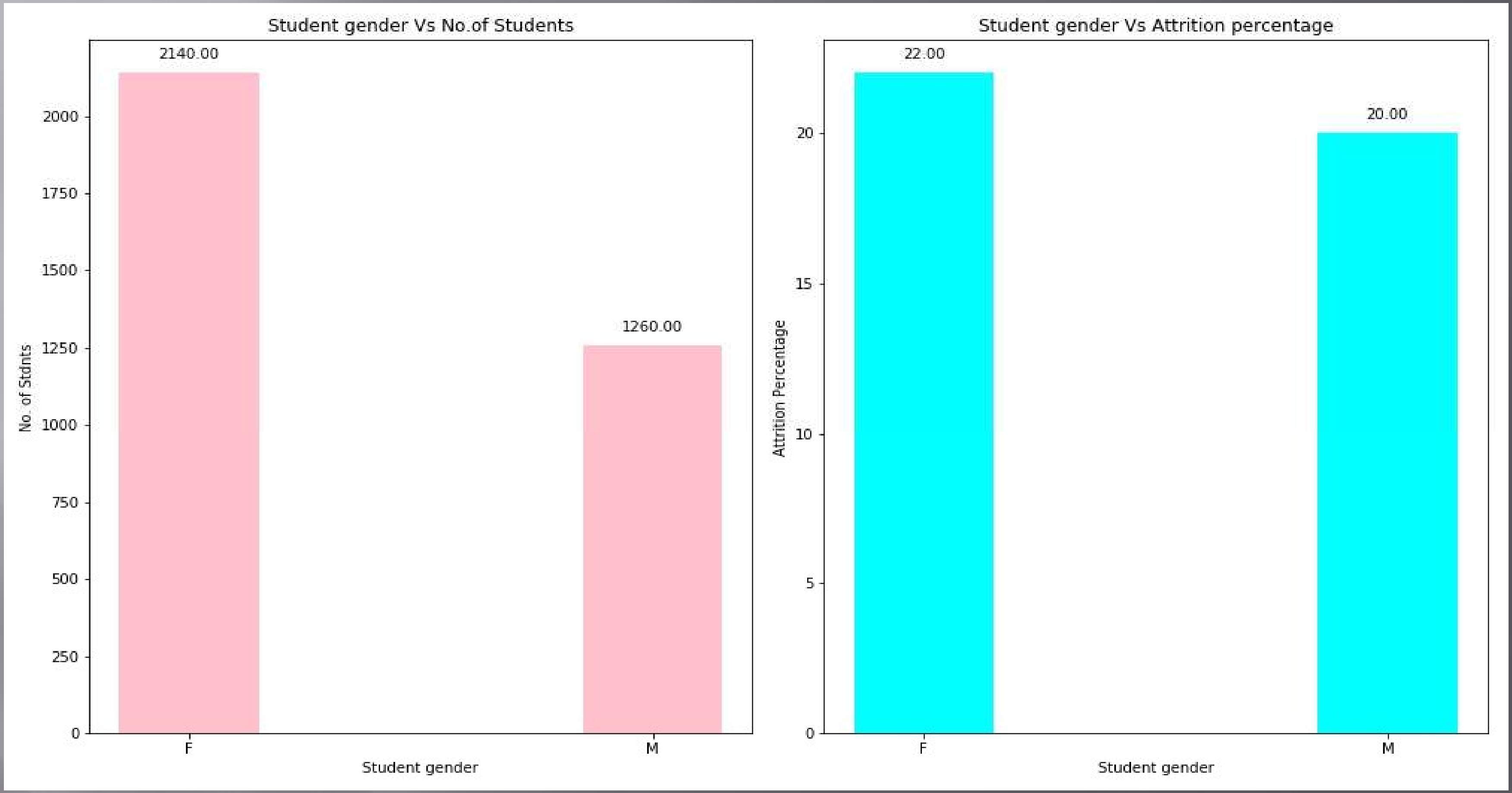
Data Analysis – Independent variables – Demographic data

Student Age



- From the above charts, it is seen that most of the students' age are from 16 to 18 and attrition rate is high for students in the higher age range.

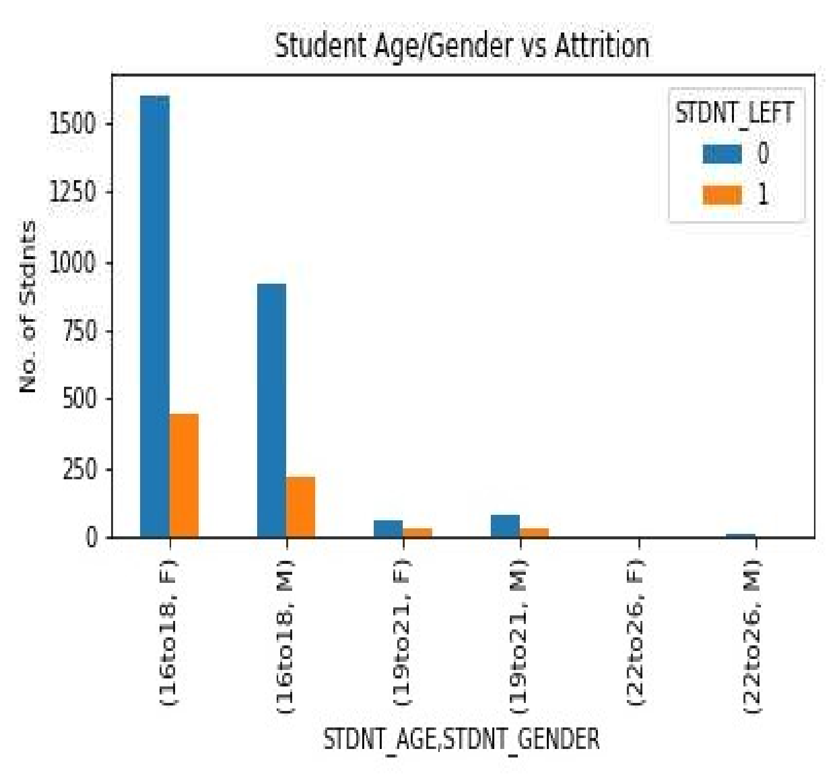
Student Gender



- Female students are more in number and their attrition rate is also slightly more than the male students.

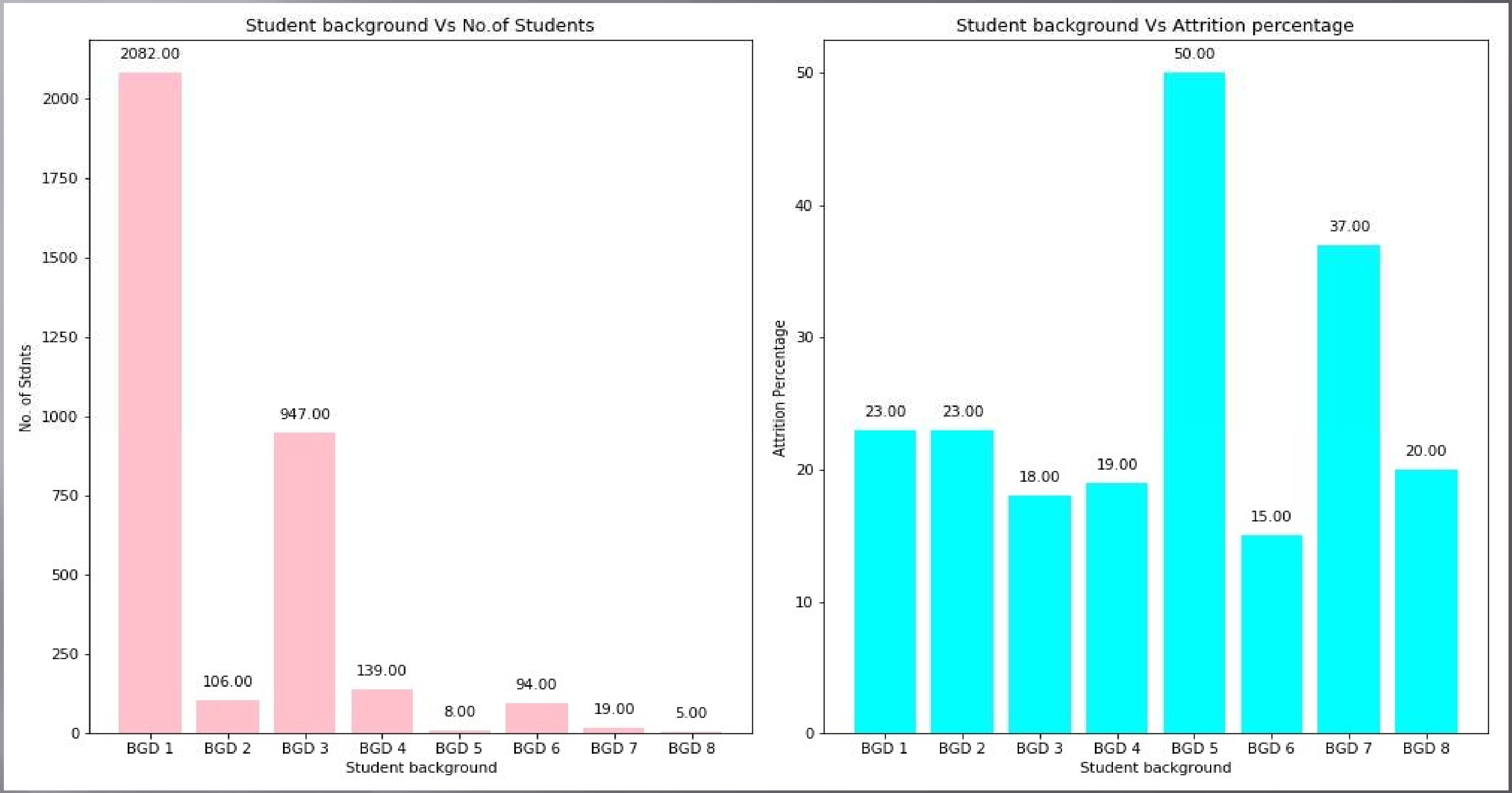
Student age & gender vs attrition

STDNT_AGE	STDNT_GENDER	STDNT_LEFT	
		0	1
16to18	F	1597	444
	M	920	216
19to21	F	66	30
	M	84	31
22to26	F	2	1
	M	8	1



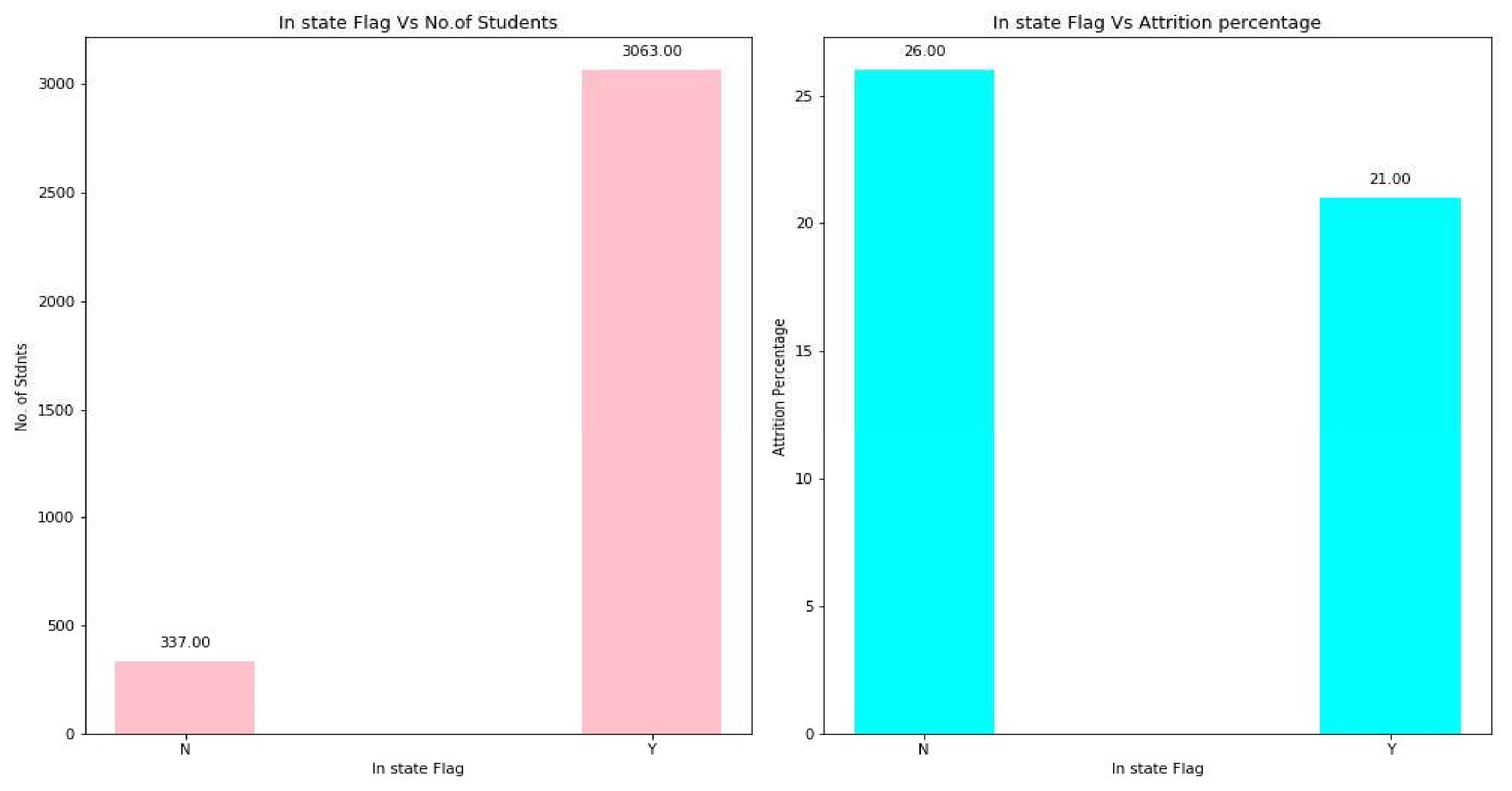
- From the above chart, it is seen that female students in the age range of 16 to 18 are having more number of attrition.

Student Background



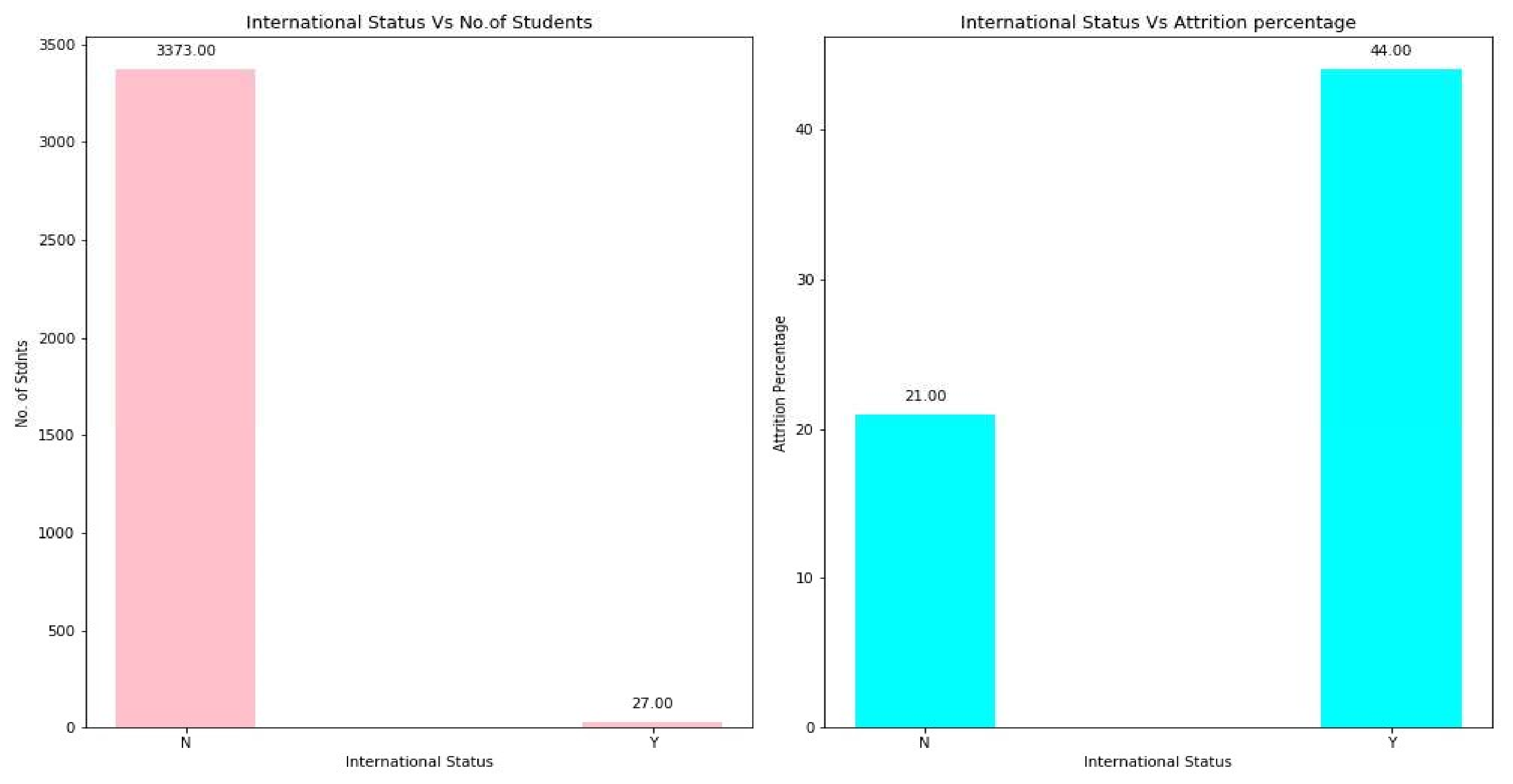
- Most of the students are from Background 1 and 3. Students are very less in number for Backgrounds 5,7 and 8. The attrition rates are also high for BGD 5,7 & 8.

In State flag



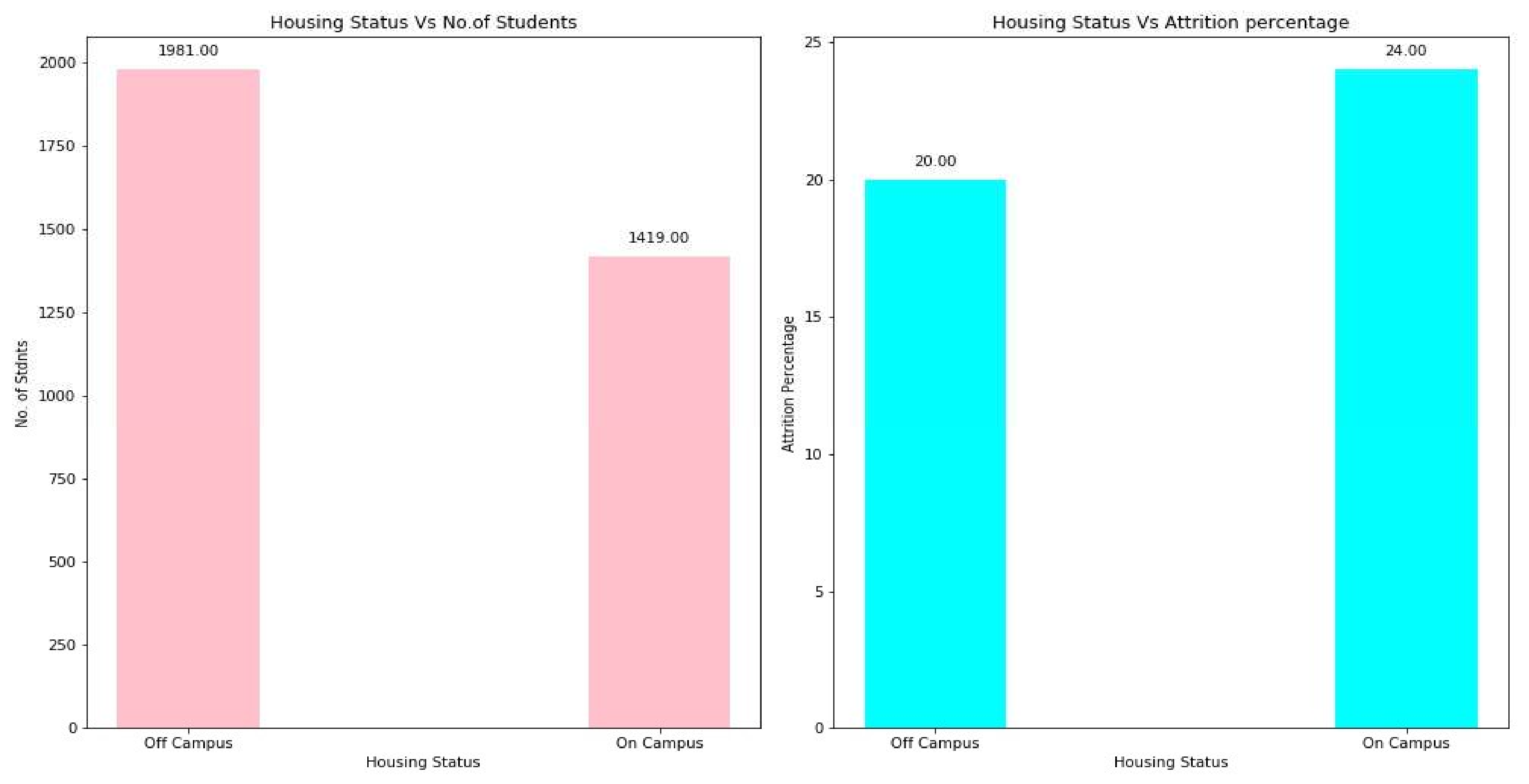
- Students from other states are very less while their attrition rate is high

International status



- The number of international students is very low and their attrition rate is high.

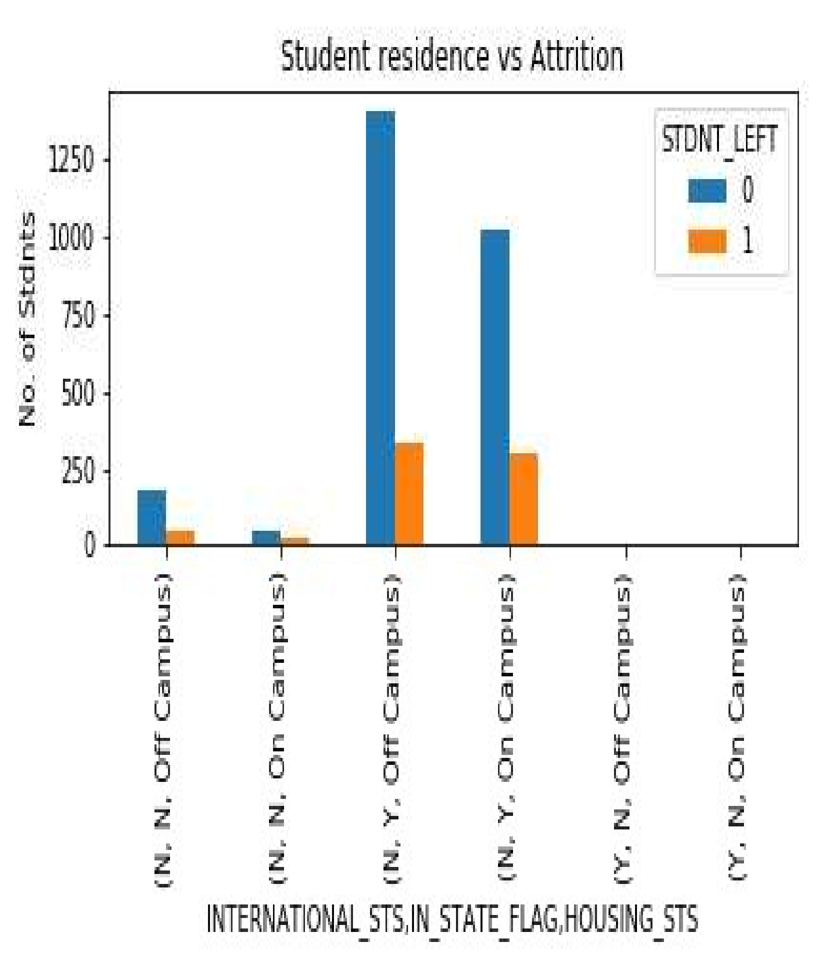
Housing status



- The number of students living inside the campus is lower. Their attrition rates seem to be slightly more than the students living off campus.

Student residence vs attrition

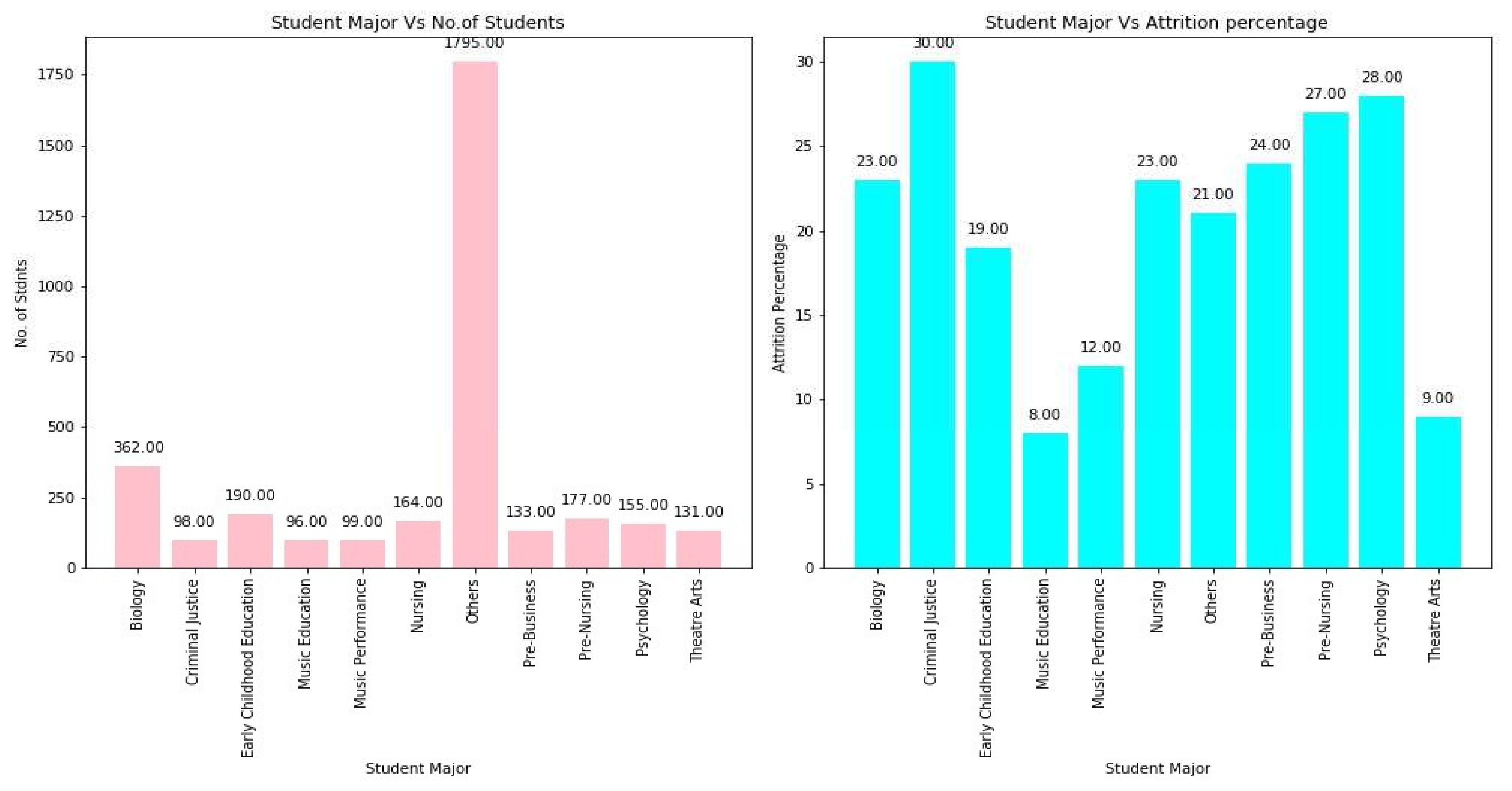
INTERNATIONAL_STS	IN_STATE_FLAG	HOUSING_STS	STDNT_LEFT	
			0	1
N	N	Off Campus	180	46
		On Campus	53	31
	Y	Off Campus	1406	338
		On Campus	1023	296
Y	N	Off Campus	6	5
		On Campus	9	7



- Form the above chart, we can see that students from the same country and state as the university and living off campus are showing more attrition.

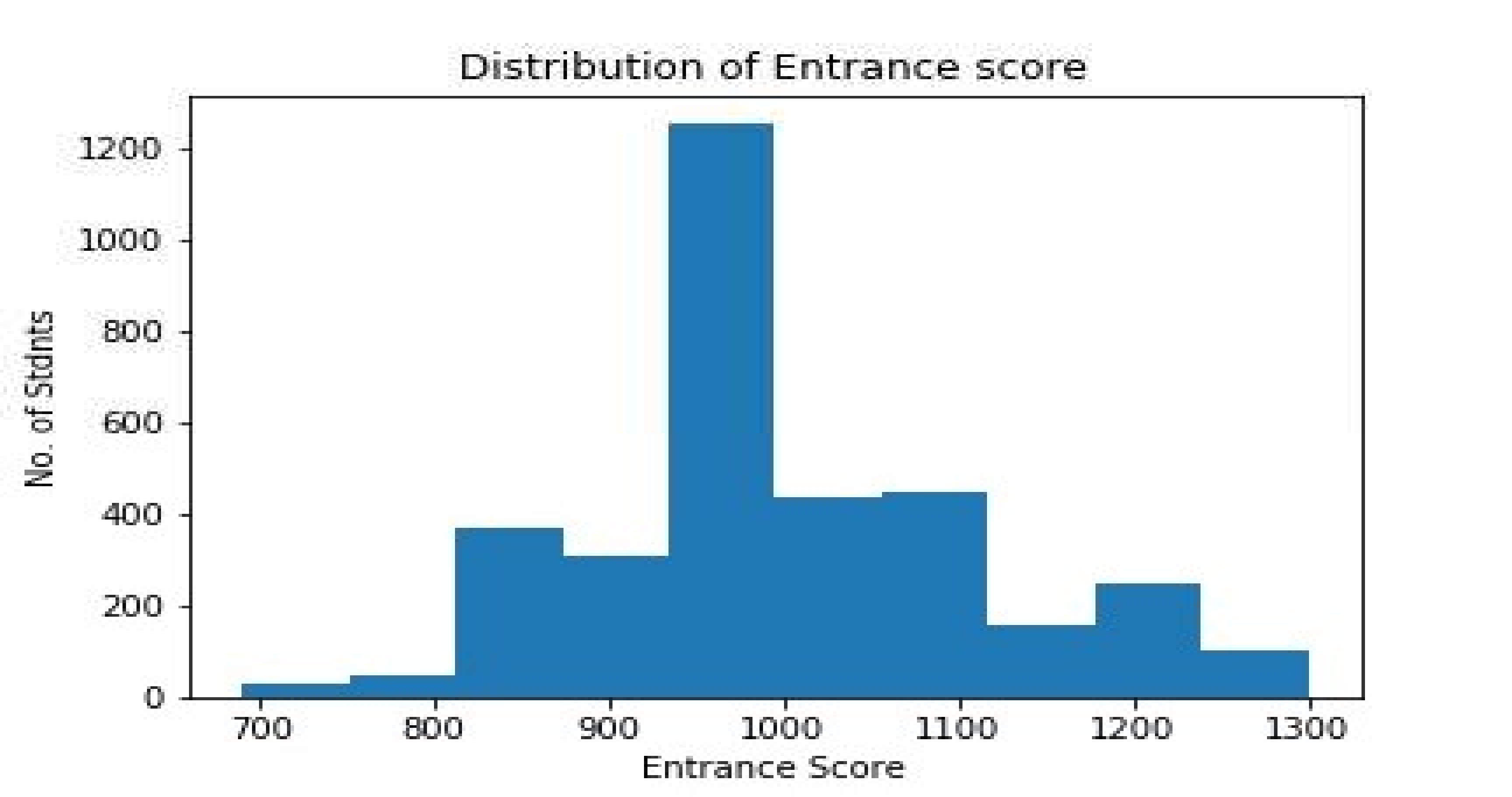
Data Analysis – Independent variables – Performance data

Student Major



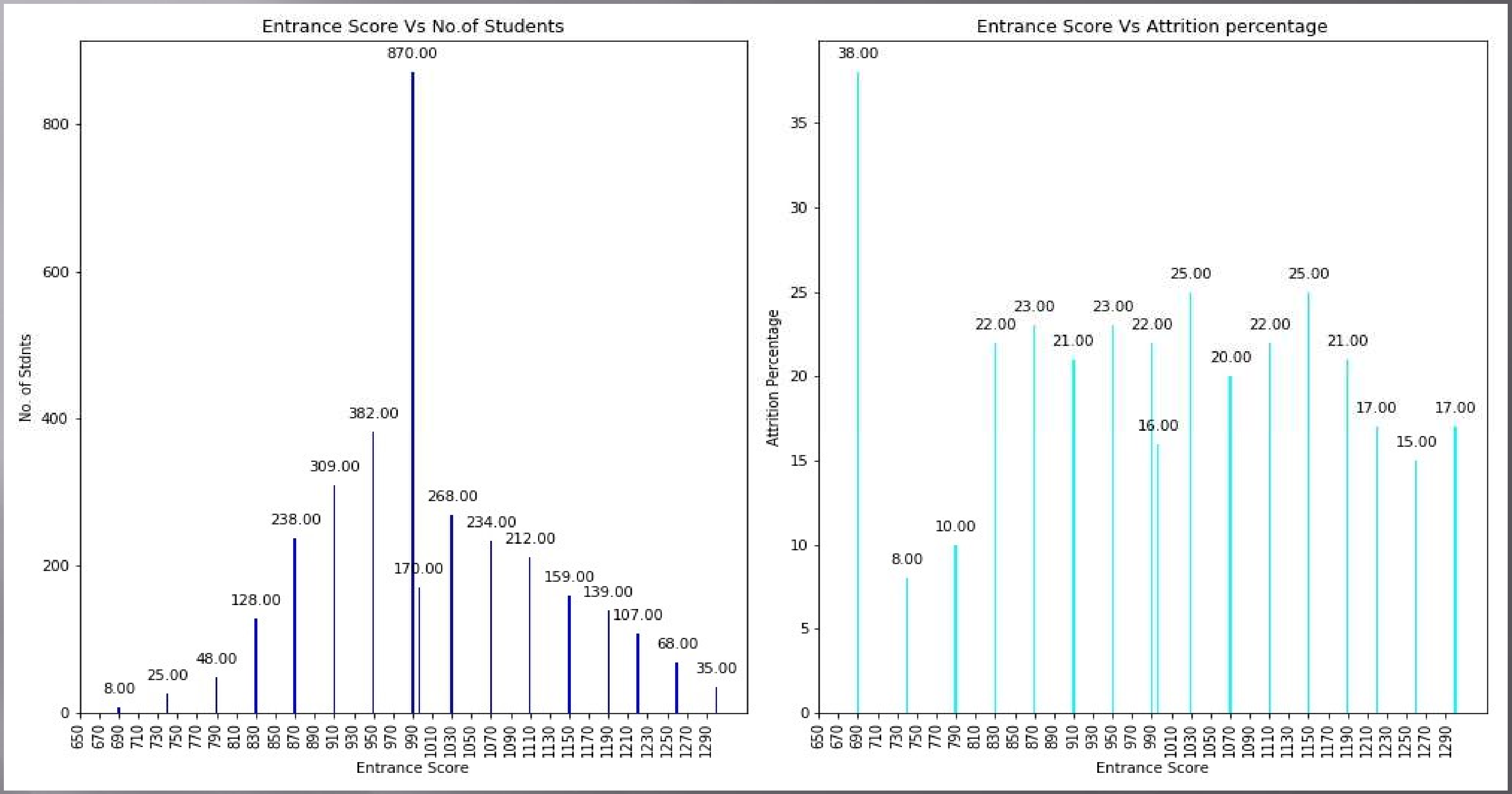
- Leaving 'others', the number of students with BIOLOGY as their major is more. The attrition rate is in the range of 20 to 30% for most of the courses(which seems very high), with 'CRIMINAL JUSTICE' having the highest attrition rate.

Student entrance tests score



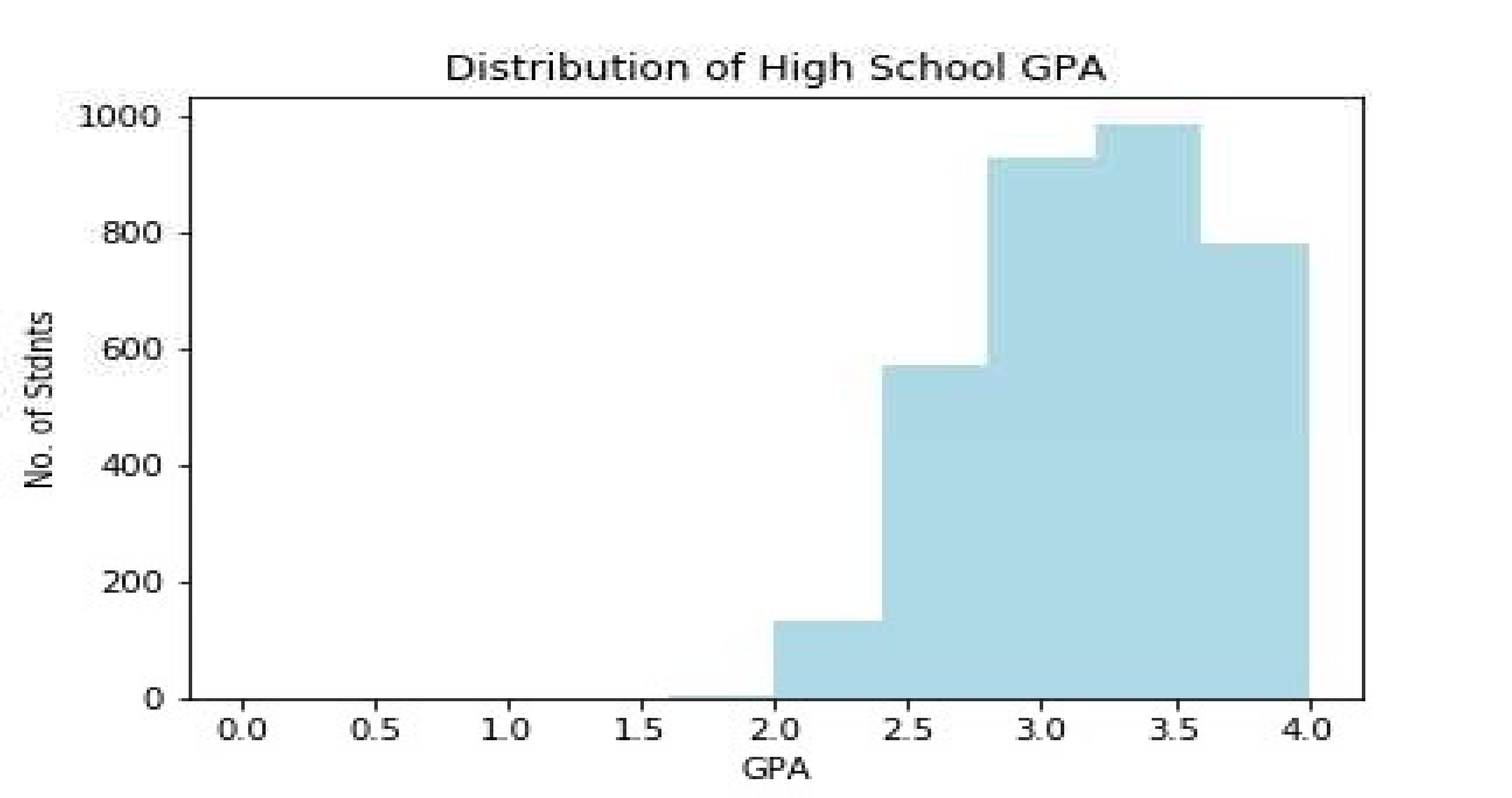
- Most of the student are in the medium range of entrance test marks.

Student entrance tests score



- From the above charts, we can see that most of the student are in the medium range of entrance test marks. The attrition rate ranges from 15% to 25% for most of the scores.

High School GPA



- Most of the students GPA is distributed between 2.5 and 3.5.

Creating Derived Variables

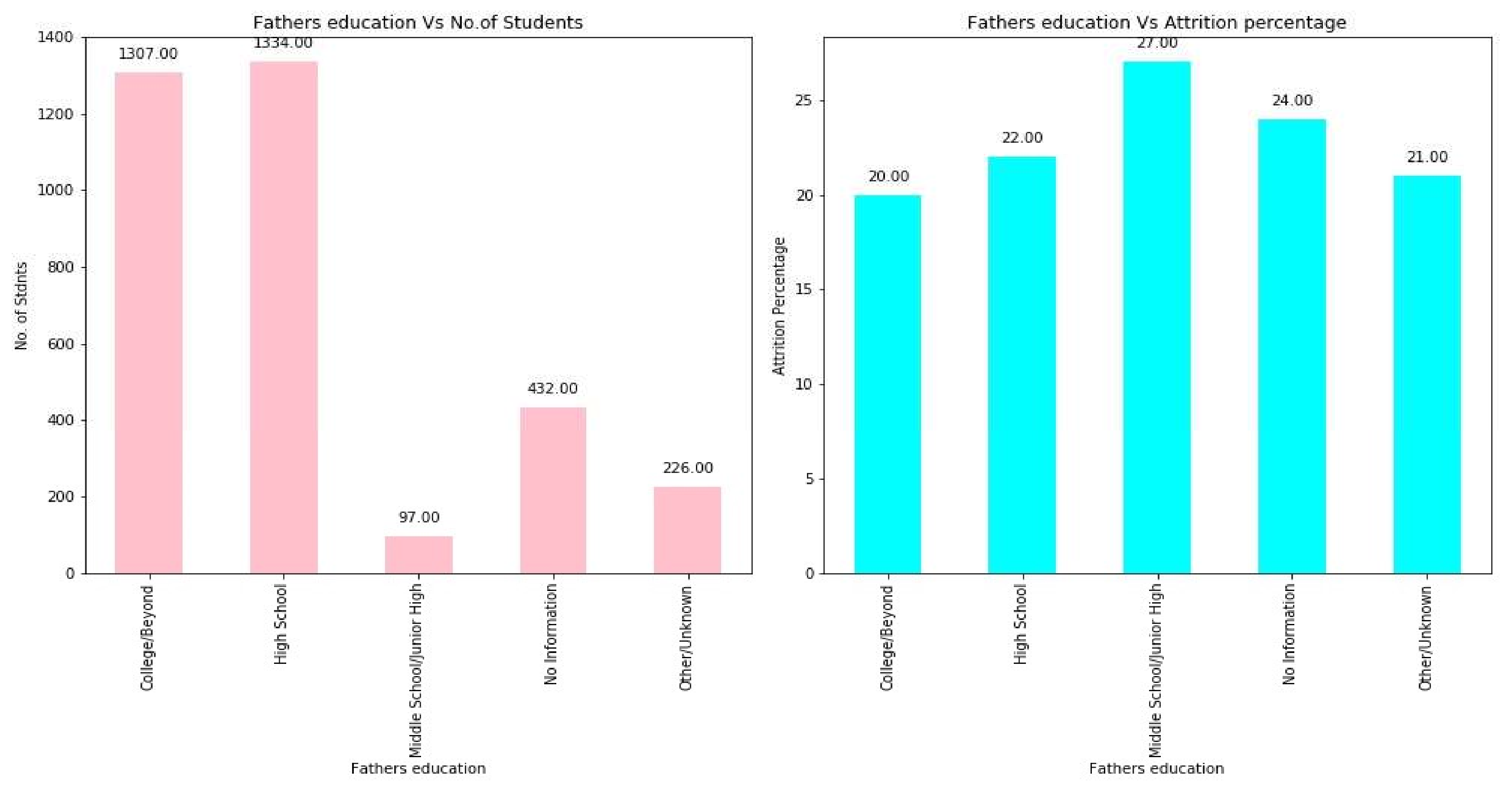
- For knowing the performance of the students in the first and second terms, derived variables 'FIRST_TERM_PERFORMANCE' and 'SECOND_TERM_PERFORMANCE' are created.

- FIRST_TERM_PERFORMANCE= FIRST_TERM_EARNED_HRS/ FIRST_TERM_ATTEMPT_HRS

- SECOND_TERM_PERFORMANCE= SECOND_TERM_EARNED_HRS/ SECOND_TERM_ATTEMPT_HRS

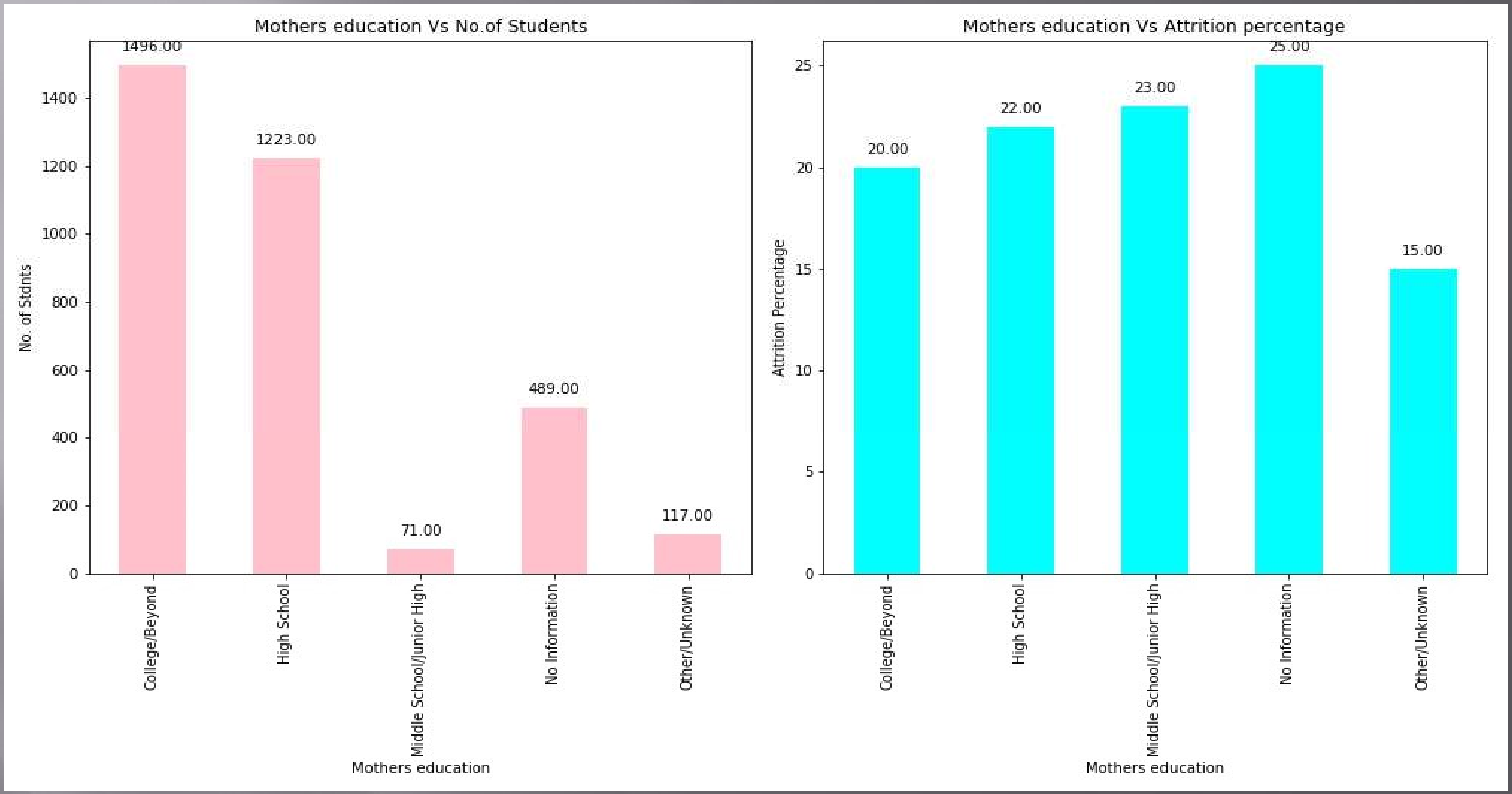
Data Analysis – Independent variables – Financial data

Father’s education



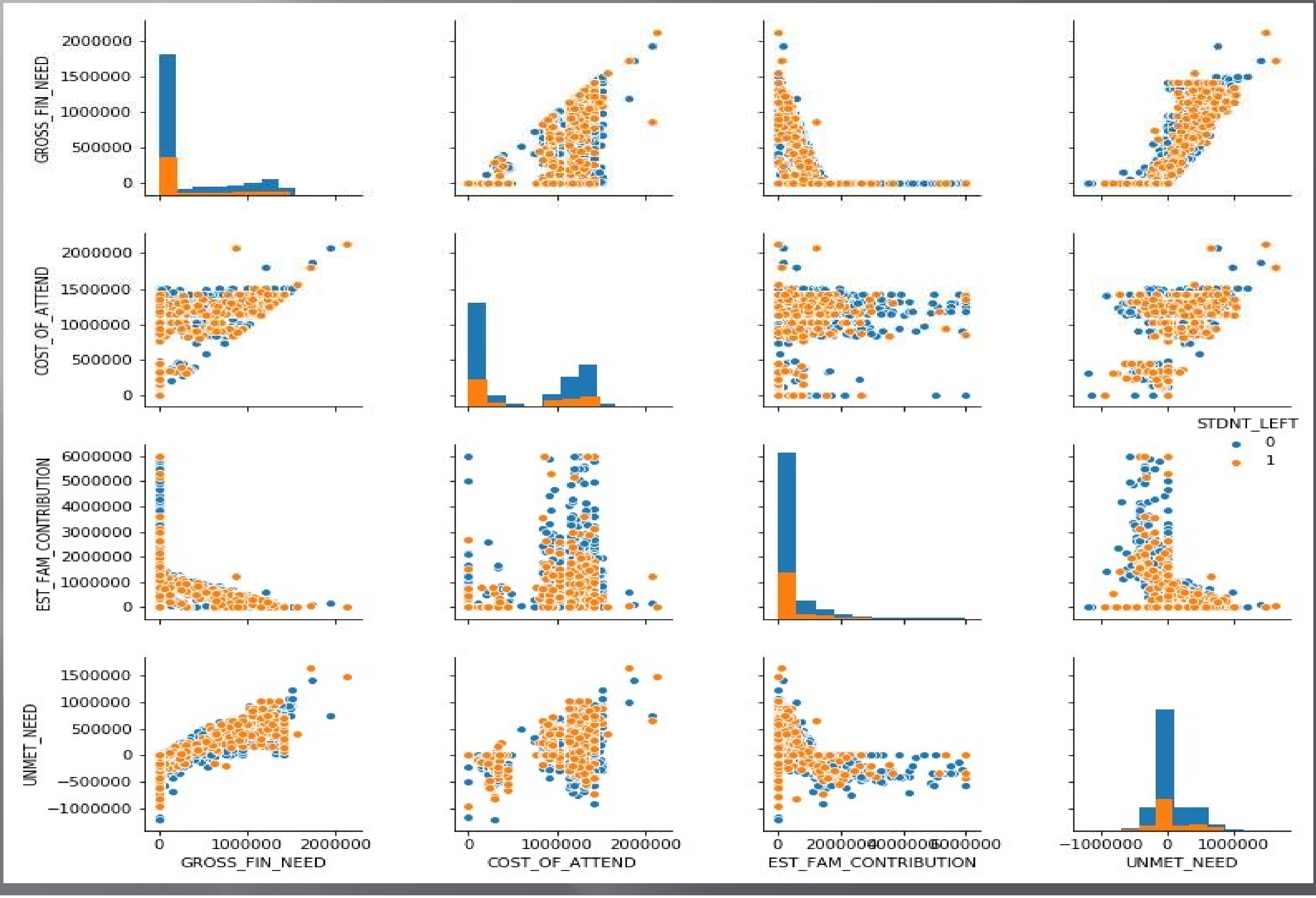
- Most of the students' fathers have education above high school level. The attrition rates for students of fathers whose education is below high school level are more than the other students.

Mother’s education



- Most of the students' mothers have education above high school level. The attrition rates for students of mothers whose education is below high school level are more than the other students.

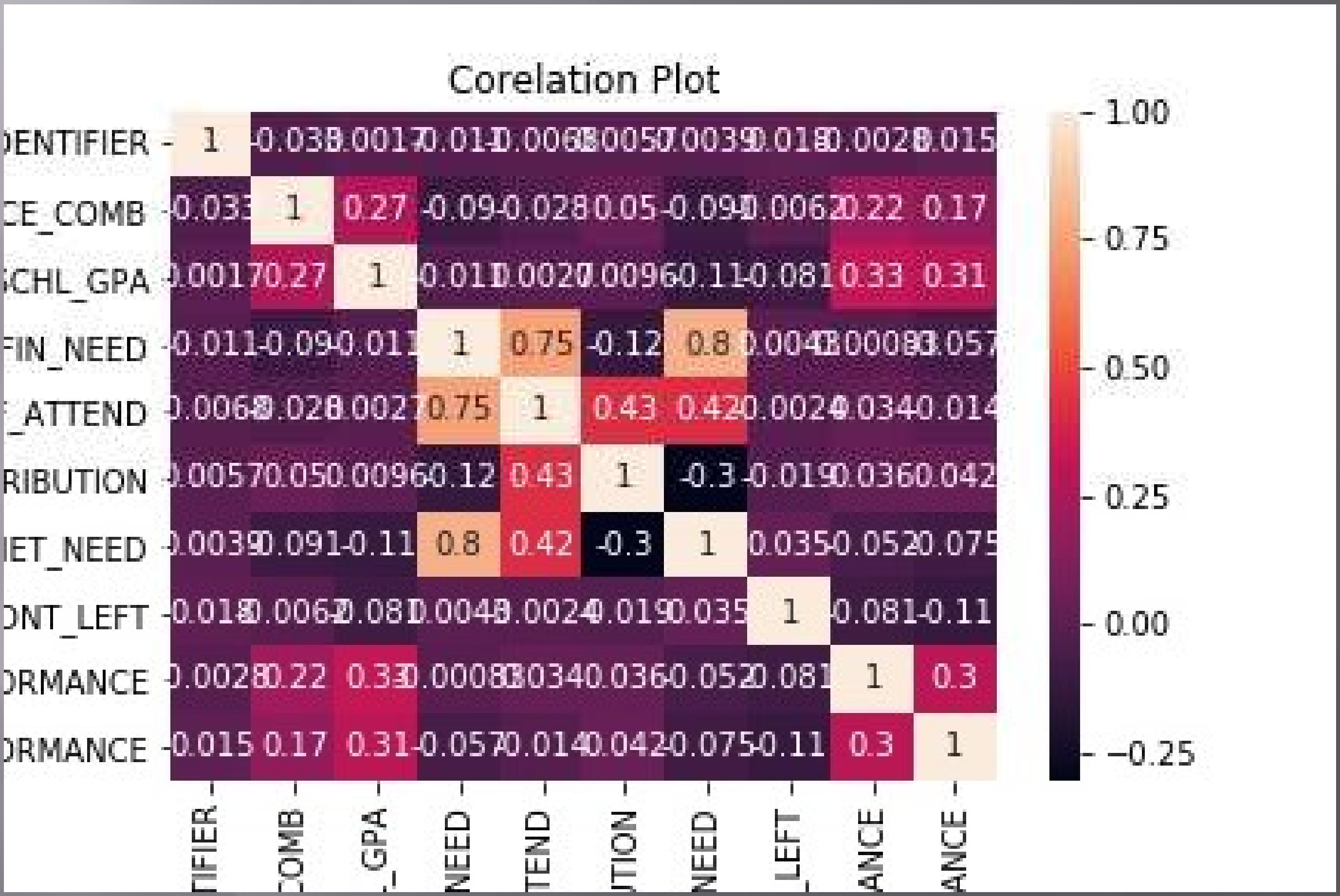
Student’s financial status variables



Student's financial status variables

- Inferences from Pair plots
  - As Gross financial need increases, the unmet need of the student increases. There is no clear difference seen between students who attrite and those who do not.
  - As Gross financial need increases, the estimated family contribution decreases. There is no clear difference seen between students who attrite and those who do not.
  - As the course fees increases, obviously the gross financial need increases. Again, there is no clear difference seen between students who attrite and those who do not.
  - With increase in the course fees, the estimated family contribution is lesser for the students who attrite than for the students who continue the course. So Higher fess and low family contribution can be a reason for attrition.
  - Course fees for most of the courses lie around 1000000.

Variables corelation



- From the pair plots and the co-relation plot, we can see that there is a strong co-relation between Gross financial need & Course fees and Gross financial need & unmet need. Therefore, dropping variable 'GROSS_FIN_NEED'

Model Building

- For this classification problem, we are using gradient boosting classifier to predict the attrition rate.
- The dummy variables for the categorical variables are created.
- The dataset is split into training and test datasets in the ratio 80:20
- The model is built with number of estimators as 80 and is fit on the X and y training datasets.
- A grid search is conducted to find the best number of estimators to get the best accuracy score. The grid search returned a value of n_estimators=210.
- Therefore the final model is built with 210 estimators. The accuracy score obtained is 0.8382352941176471

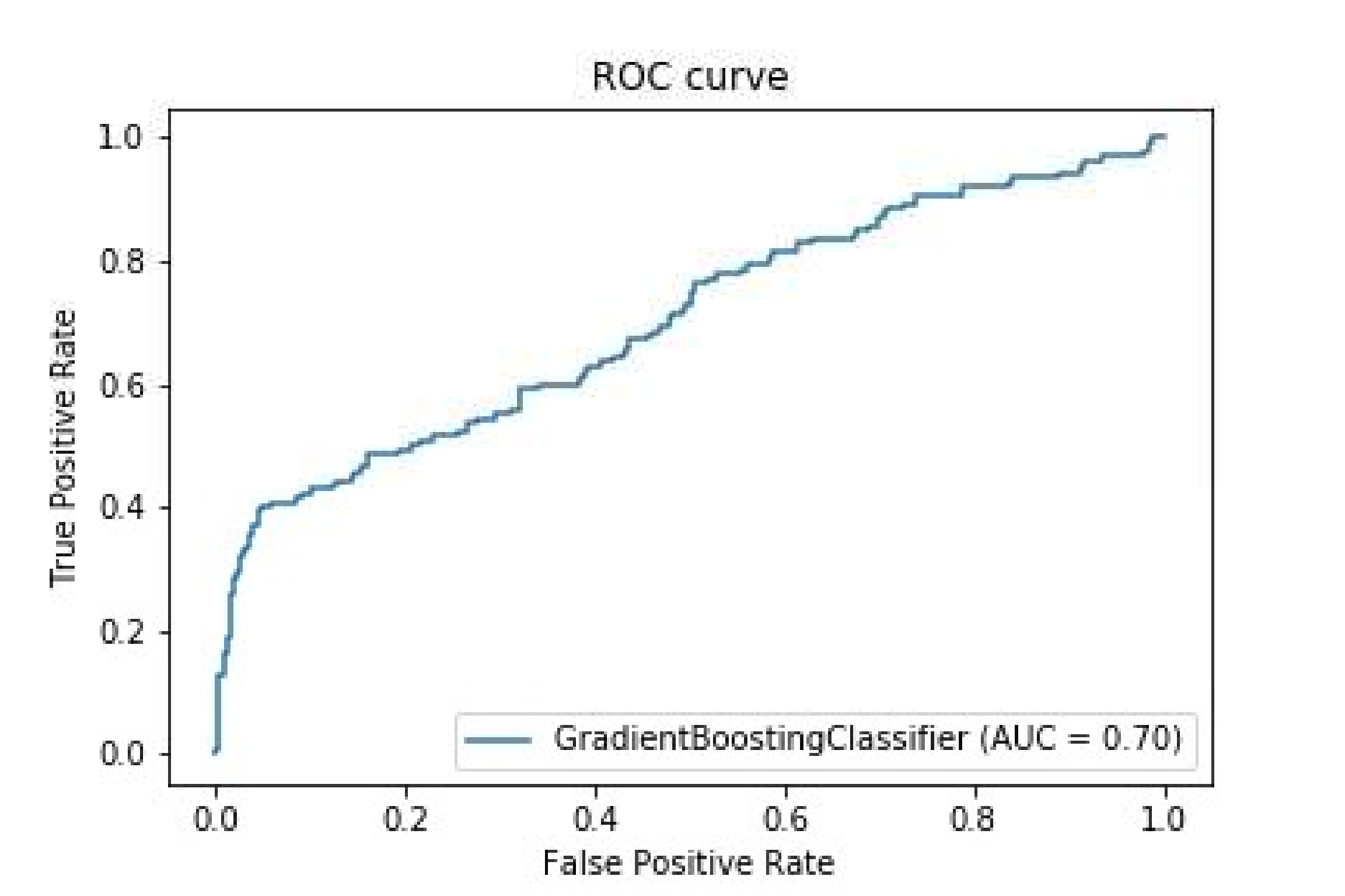
Model Validation

- To validate the model, we use confusion matrix and ROC curve

- Confusion Matrix

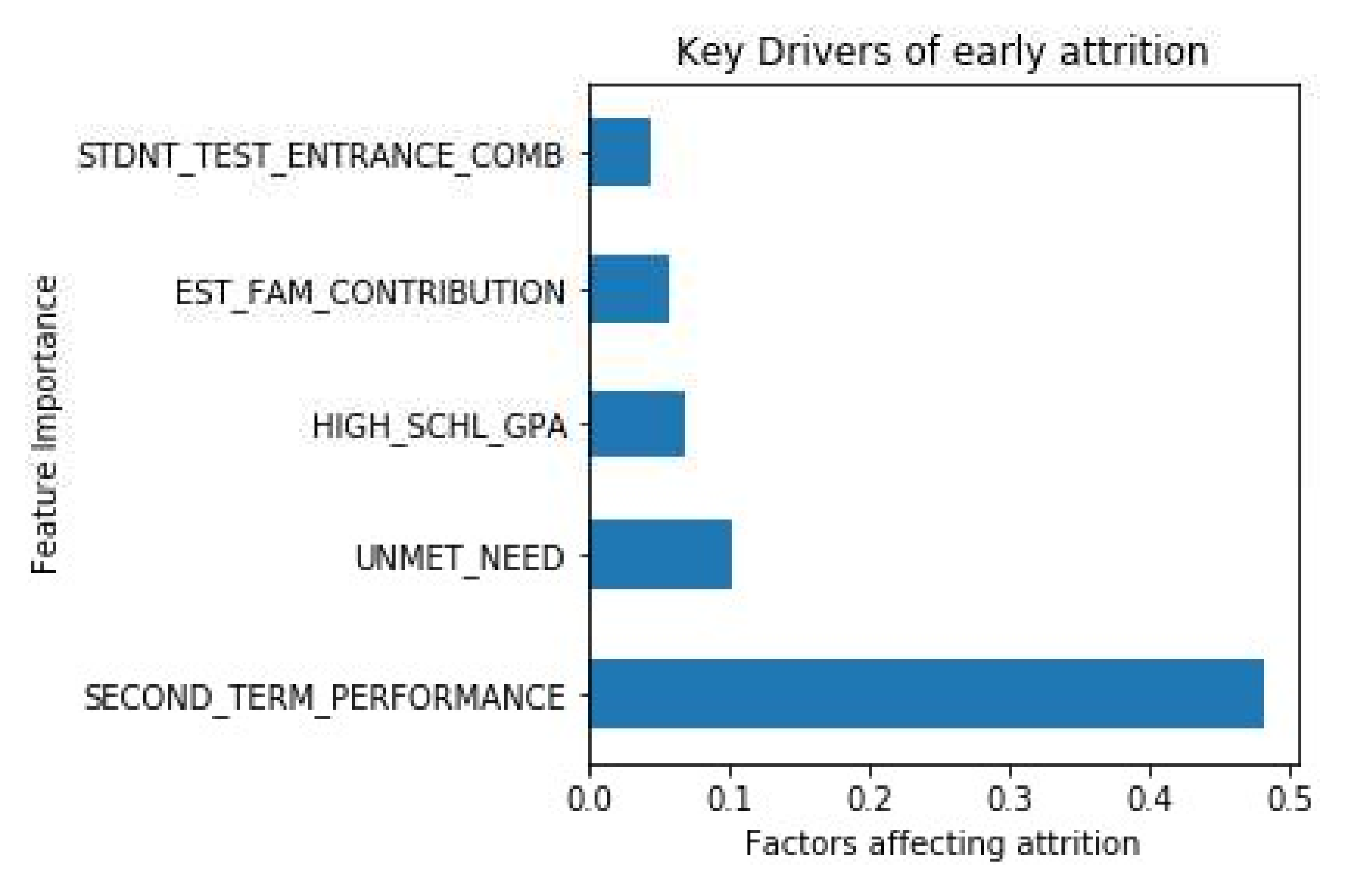
	0	1
0	524	16
1	94	46

Model Validation



- The Area Under Curve(AUC) in ROC curve for the model is 0.7004100529100528. The model performs well.

Key Drivers of early attrition



- The feature importance for each variable is computed. The variables with the top 5 feature importance are the key drivers of early attrition.

Recommended interventions

- From the model results it is seen that the performance of the students right from the high school to the second term in the university, highly affect the student attrition. The teachers and parents should encourage the students to perform well and provide the required assistance to the students, right from school, so that they do not leave the university without completing their course.

- Next to the Student performance, the financial status (unmet need and estimated family contribution) of the students determine their possibility of attrition. The University might identify student s performing well and provide them with financial assistance in the form of scholarships or sponsors, in collaboration with the government or NGO's if needed.

Recommended interventions

- Other inferences from analysing the feature importance are
  - The feature importance for the GENDER_F is more than that of GENDER_M, implying that female students are more likely to attrite.
  - The feature importance is more when the education background of the parents is 'Middle School/Junior High' than other education backgrounds.
- The university may identify these students and pay special attention in providing them required assistance.

Thank You!