College Motivation: Predicting Students Going to College
By: Rebekah Sander

ABSTRACT

The goal of this project was to use predictive analytics to identify whether a student will make the decision to go to college. This "college motivation" dataset collected 1000 high school students in Mexico who graduated in 2014. A regression was used to make predictions of a student going to college based on variables that have to do with their academic and personal backgrounds.

METHODS–Preparing the Data

1.) Checking for missing data:
   - There is no missing data

2.) Recoding Variables:
   - "will_go_to_college" and "parent_was_in_college": Originally character values "True" and "False", changed to 1 and 0 for binary.

3.) Data Split:
   - 6:2:2 data split ratio.
   - 60% of data into new dataframe, "college_train", 20% into new dataframe "college_validation", 20% into new dataframe, "college_test".

METHODS-Building the Best Model

Regression Model:

- Logistic regression model based on data frame, "college_train" with will_go_to_college as the binary dependent variable, all other variables are used as independent variables.

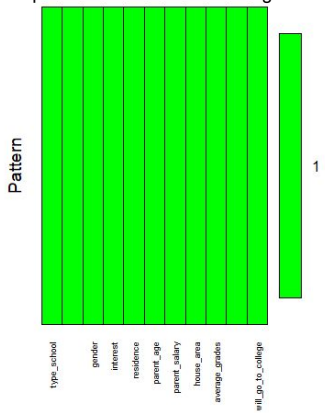
Choosing the best threshold for binary prediction:

- Best threshold determined through "college_validation" dataframe.

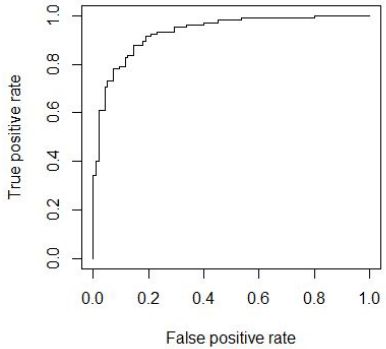
Performance and Evaluation Metrics

- PMSE
- AUC-ROC
- Accuracy

Graph 1: Visualization of Missing Data



Graph 2: AUC-ROC curve



DESCRIPTION

will_go_to_college: Student goes to college
type_school: Type of high school the student went to
school_accreditation: Student's high school's accreditation
gender: Gender of the student
interest: Student's interest in going to college
residence: Where the student lives
parent_age: Age of student's parent
parent_salary: How much the student's parent makes
house_area: Area of student's house
average_grades: Average grades of student
parent_was_in_college: Student's parent went to college

RESULTS

Table 1: Confusion Matrix

		REFERENCE	
		0	1
PREDICTION	0	86	22
	1	9	83

Best Threshold = 0.64
Accuracy = 0.86
PMSE = 0.33
AUC-ROC = 0.92

CONCLUSIONS

During this predictive analysis, a logistic regression model was utilized to identify high school students in Mexico who are likely to attend college based on their personal and academic life. The primary objective was to pinpoint and be able to reach out to students who may not have plans to pursue higher education.

The model achieved an AUC-ROC evaluation metric of 0.92. This high value indicates the model's strong ability to correctly predict whether a student is college-bound. However, when it comes to generalizability, it is important to acknowledge the dataset is based in Mexico. This model may not be the case when looking at different regions. Since this dataset was a collection of students in Mexico, this model may not be extended to other countries. To improve generalizability, more data could be collected from various other regions and countries.

Overall, this analysis provides valuable insights into identifying students at risk of not attending college so that intervention may occur.