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.

Introducing the variables

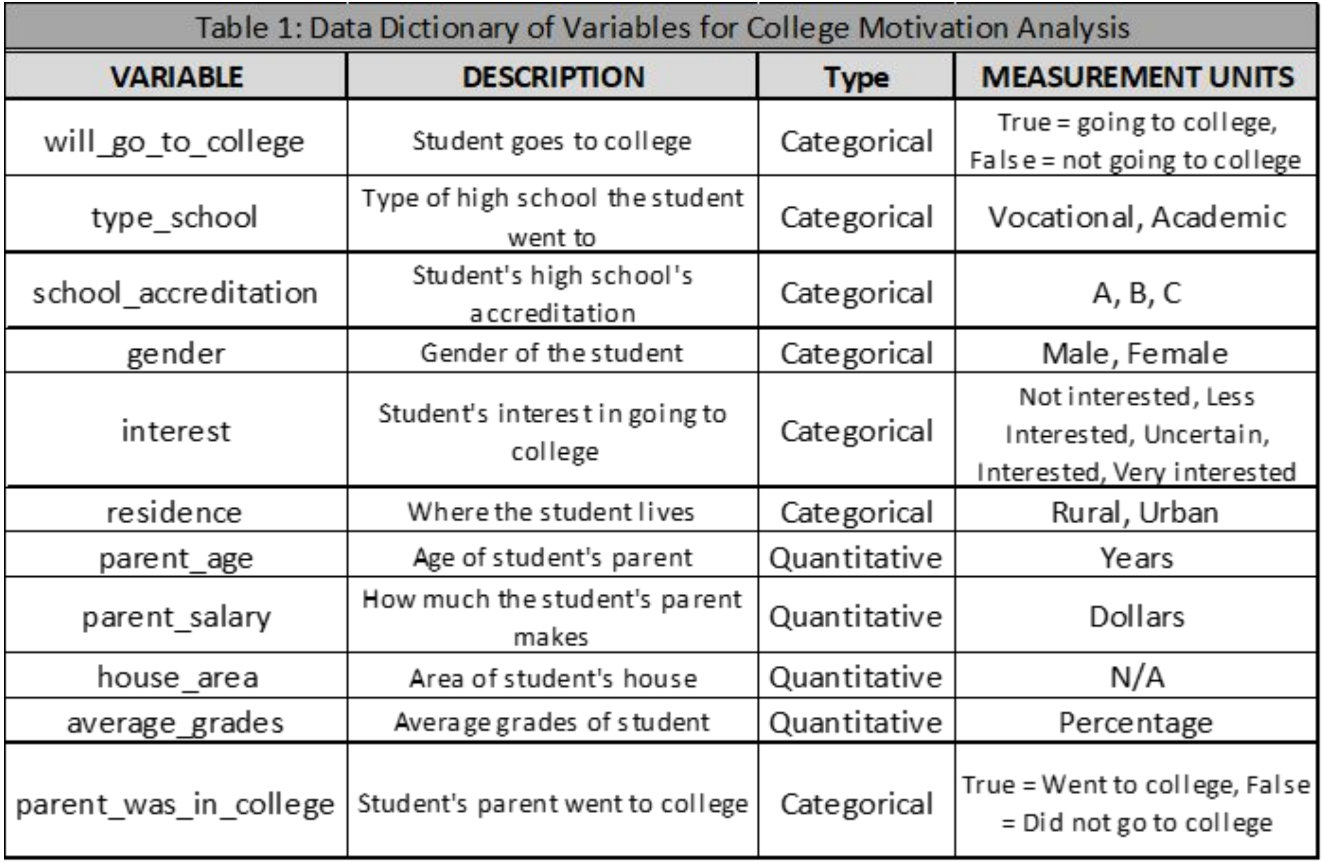


Table 1: Data Dictionary of Variables for College Motivation Analysis

VARIABLE	DESCRIPTION	Type	MEASUREMENT UNITS
will_go_to_college	Student goes to college	Categorical	True = going to college, False = not going to college
type_school	Type of high school the student went to	Categorical	Vocational, Academic
school_accreditation	Student's high school's accreditation	Categorical	A, B, C
gender	Gender of the student	Categorical	Male, Female
interest	Student's interest in going to college	Categorical	Not interested, Less Interested, Uncertain, Interested, Very interested
residence	Where the student lives	Categorical	Rural, Urban
parent_age	Age of student's parent	Quantitative	Years
parent_salary	How much the student's parent makes	Quantitative	Dollars
house_area	Area of student's house	Quantitative	N/A
average_grades	Average grades of student	Quantitative	Percentage
parent_was_in_college	Student's parent went to college	Categorical	True = Went to college, False = Did not go to college

Methods–Preparing the Data

1.) Checking for missing data:

- There is no missing data

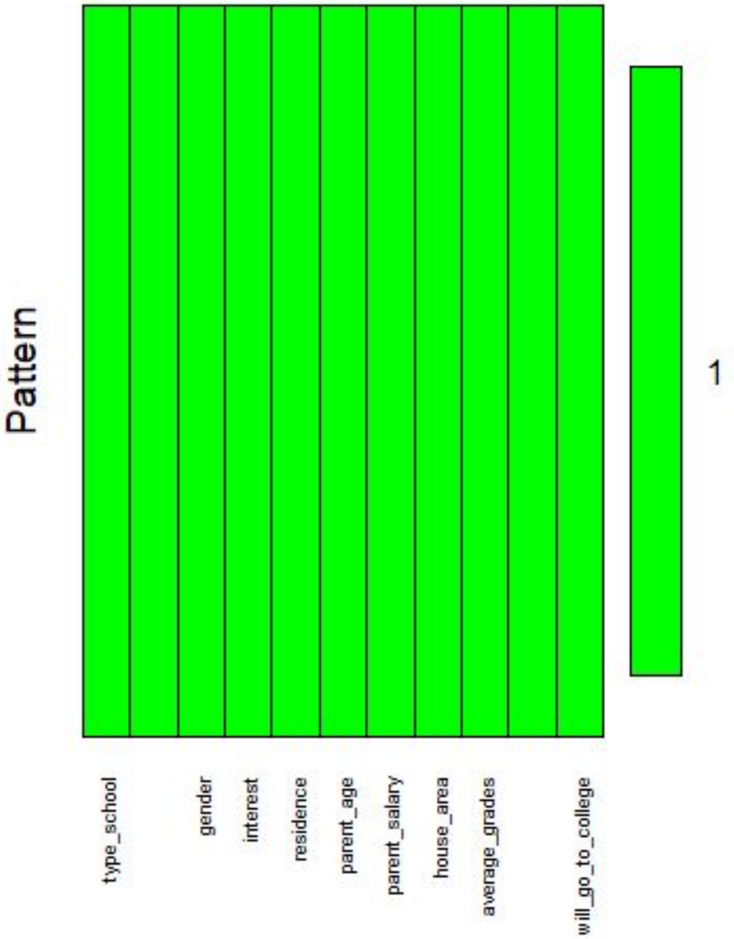
2.) Recoding Variables:

- “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”.

Graph 1: Visualization of Missing Data



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

Results

Best Threshold = 0.64

PMSE = 0.33

Accuracy = 0.86

AUC-ROC = 0.92

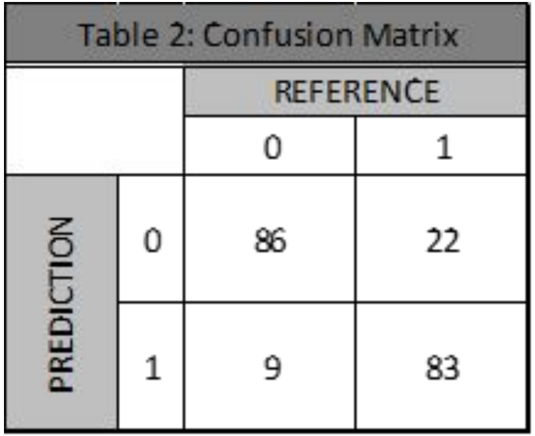
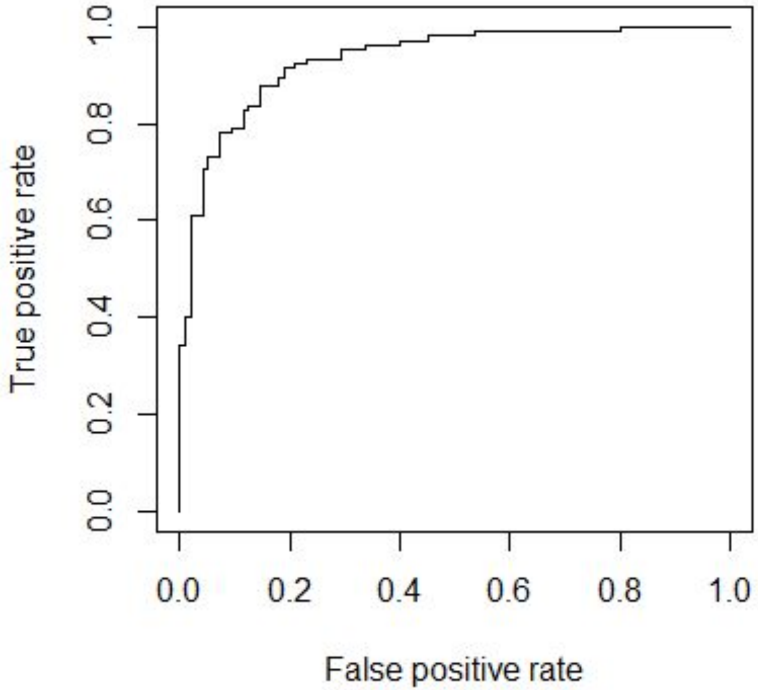


Table 2: Confusion Matrix

		REFERENCE	
		0	1
PREDICTION	0	86	22
	1	9	83

Graph 2: AUC-ROC curve



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.