Building and Analyzing Binary Classification models of students success

RECOMMENDATIONS FOR UNIVERSITY OFFICIALS:

HELP YOUR FAILING STUDENTS SUCCEED

FLATIRON SCHOOL

Phase 3 Final Project

Authors: Elena Kazakova

Cohort: DS02222021

Instructor: James Irving

Summary

This Project analyses the data on Colombian engineering students performance in the professional assessment tests

- Colombia
- Data gathered and compiled between 2012 and 2018

The resulting model provides insight into

- What factors affect engineering students performance
- What Programs/Universities are the most successful helping their students achieve their potentials

Outline of the presentation

- Business Problem
- Data
- Final Model
- Results
- Conclusions

Business Problem

This project goal is to build a classification predictive model of graduating Colombian Engineering Students performance evolution based on their students socio-economic characteristics and engineering programs they attended.

This information can be helpful for university officials to identify students not fulfilling their potential and providing additional support (financial, etc.) to help them progress in their professional careers. The results of the model also can assist future engineering student in choosing professional programs and universities.

Data

The dataset used in this project has been downloaded from the Mendeley Data Repository.
The dataset has been published in ScienceDirect Elsevier publication "Data in Brief",
Volume 30, June 2020, 105537.

Quote from the paper summary:

"This data article presents data on the results in national assessments for secondary and university education in engineering students. The data contains academic, social, economic information for 12,411 students. The data were obtained by orderly crossing the databases of the Colombian Institute for the Evaluation of Education (ICFES). The structure of the data allows us to observe the influence of social variables and the evolution of students' learning skills."

The dataset has of 44 dependent and independent variables.
and each row represents a student.
The variables correspond to the student's personal information (categorical) and the result obtained in the assessments (numerical).
The academic assessment is recorded at two moments of the student life.
First, the scores of the national standardized test at the final year of the high school (Saber 11), 2012-2014 years. The second moment of academic assessment is in the final year of their professional training in Engineering, recorded on the national standardized test for higher education (SABER PRO), 2018 year.

Modifications to the Data

Adjusted target to a binary format

Removed features:

- **Test IDs**
- **Professional subject test scores**

COD_S11: S11 test student identifier; 12411 unique values, no duplicates
Cod_SPro: SABER_PRO test student identifier; 12395 unique values, 16 duplicates

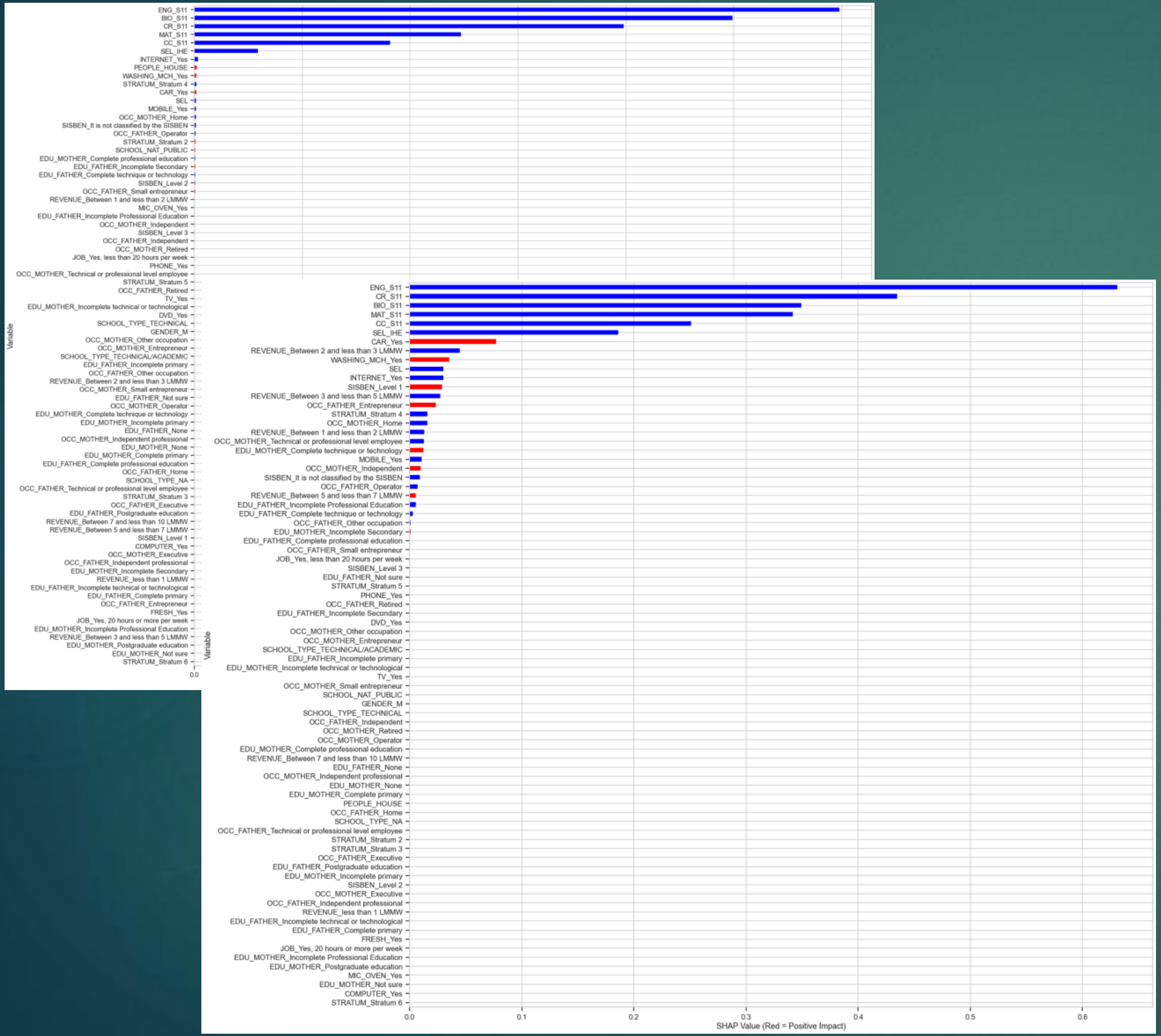
Categorical variables

GENDER: Student's gender; 2 unique values (F/M)
EDU_FATHER: Level of education of a student's father; 12 unique values
EDU_MOTHER: Level of education of a student's mother; 12 unique values
OCC_FATHER: Occupation of a student's father; 12 unique values
OCC_MOTHER: Occupation of a student's mother; 12 unique values
STRATUM: Colombian socio-economic class; 7 unique values
SISBEN: Levels of Colombian welfare system; 6 unique values
PEOPLE_HOUSE: Number of people in the household; 13 unique values
INTERNET: Does a household have an internet connection; 2 unique values (Yes/No)
TV: Does a household have a TV; 2 unique values (Yes/No)
COMPUTER: Does a household have a computer; 2 unique values (Yes/No)
WASHING_MCH: Does a household have a washing machine; 2 unique values (Yes/No)
MIC_OVEN: Does a household have a microwave oven; 2 unique values (Yes/No)
CAR: Does a household have a car; 2 unique values (Yes/No)
DVD: Does a household have a DVD player; 2 unique values (Yes/No)
FRESH: Does a household have access to fresh water; 2 unique values (Yes/No)
PHONE: Does a household have a landline phone; 2 unique values (Yes/No)
MOBILE: Does a student have a mobile phone; 2 unique values (Yes/No)
REVENUE: Household income category; 8 unique values

Final Models

- Several models were tested
- Final four models have the best predictive power and provide the best insight
- Separate models were developed for two goals of the project:
  - The first two models help university officials to flag incoming students who might be at risk of falling behind in their studies
  - The second two models generate choice recommendations for future students by ranking universities in the order of educational success of their engineering students

For University Officials



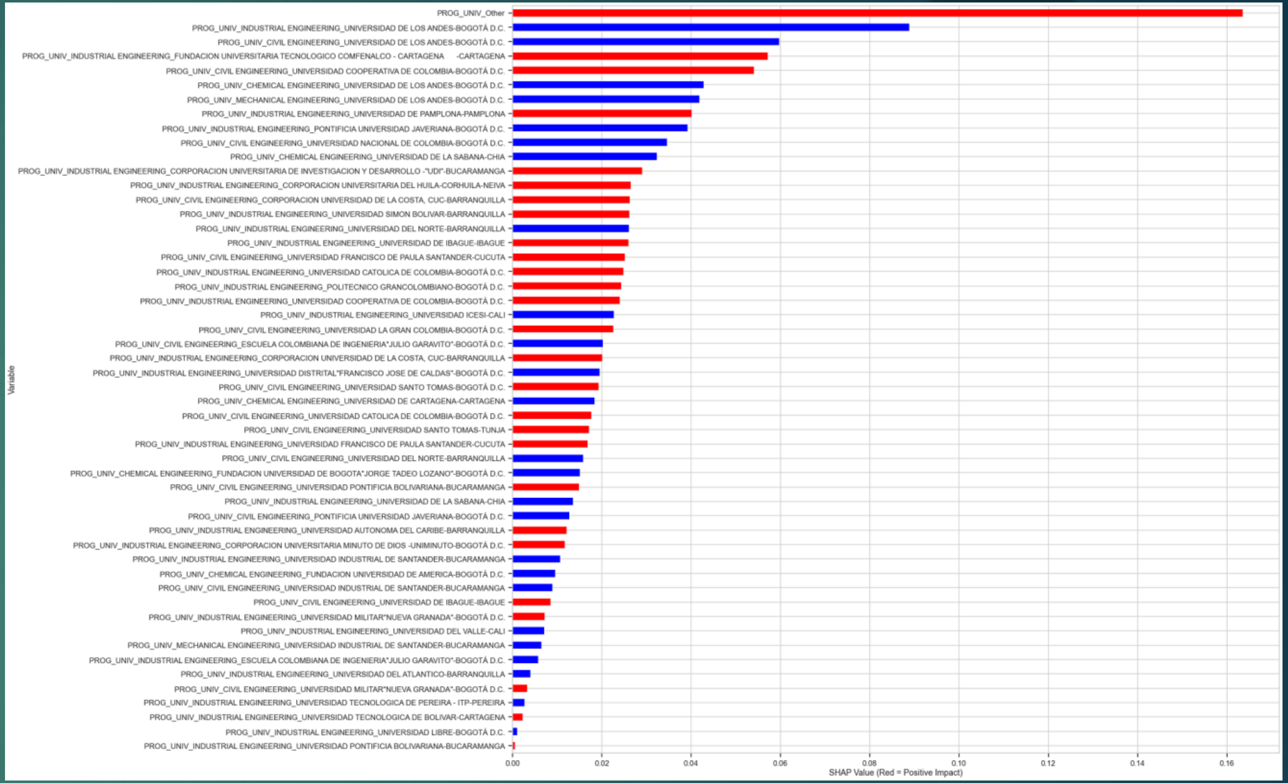
<u>Recommendations to university officials</u>:

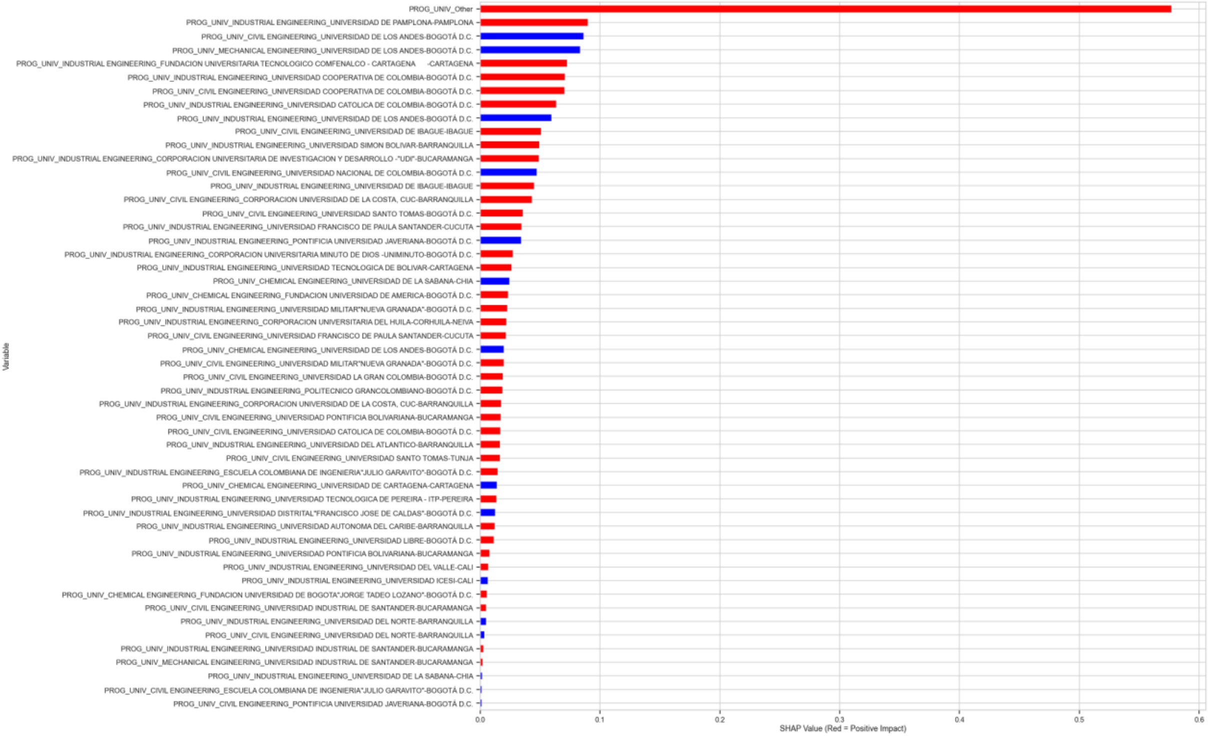
Flag incoming student with

- Lower HS subject scores in
  - ENG_S11
  - CR_S11
  - MAT_S11
- Lower Socio-Economic level
  - SHE level below average

Use our model to flag incoming student who are at risk of falling behind

For Future Engineering Students





For Future Engineering Students

Highest ranked programs:

- PROG_UNIV_CIVIL ENGINEERING_UNIVERSIDAD NACIONAL DE COLOMBIA-BOGOTÁ D.C
- PROG_UNIV_CIVIL ENGINEERING_UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C
- PROG_UNIV_CIVIL ENGINEERING_UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.
- PROG_UNIV_INDUSTRIAL ENGINEERING_UNIVERSIDAD DEL NORTE-BARRANQUILLA

Lowest Ranked programs:

- PROG_UNIV_CIVIL ENGINEERING_UNIVERSIDAD NACIONAL DE COLOMBIA-BOGOTÁ D.C
- PROG_UNIV_CIVIL ENGINEERING_UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C
- PROG_UNIV_CIVIL ENGINEERING_UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.
- PROG_UNIV_INDUSTRIAL ENGINEERING_UNIVERSIDAD DEL NORTE-BARRANQUILLA

Recommendations

Recommendations to university officials:

- Flag incoming student with

  Lower HS subject scores in

  - ENG_S11
  - CR_S11
  - MAT_S11

  Lower Socio-Economic level

  - SHE level below average

- Use our model to flag incoming student who are at risk of falling behind

Recommendations (continued)

Recommendations to future engineering students:

- Consider one of the highest ranked programs from our list for your Engineering education
- Steer away from the programs on the low ranked list
- Invest time and efforts to increase your scores in S11 MAT, CR, and ENG tests. Your scores are highly indicative of your success as engineering students

Conclusions:

Limitations of the model:

- Other important factors are not included in the dataset
- Data is limited to Colombian Higher Institutions

Additional analysis suggested:

- Add variables to the original dataset like
- Update data
- Consider regression rather than classification models

Thank you!

Email: e.v.kazakova@gmail.com

GitHub: @sealaurel

LinkedIn: https://www.linkedin.com/in/elena-v-kazakova/