

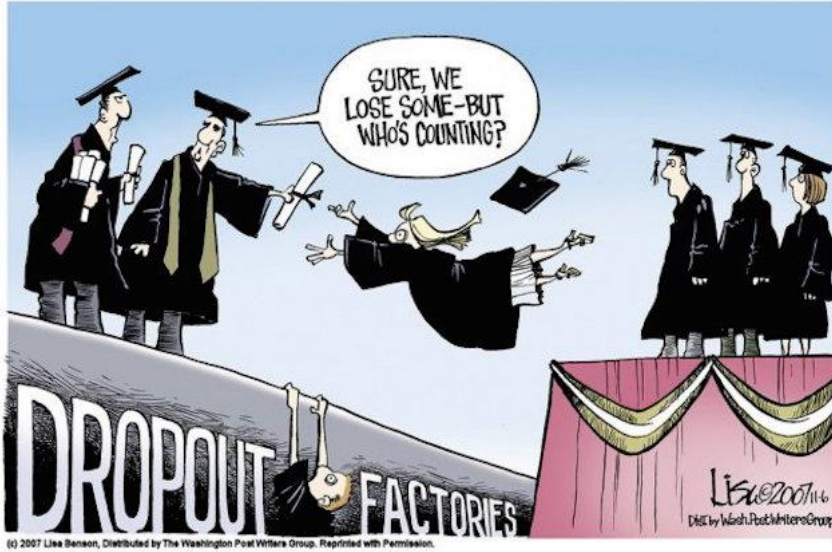
Predicting Student Dropout Rate and Academic Success

Project Description

The purpose of this project is to predict the likelihood of students dropping out of secondary education based on a various student data including social, academic and financial factors.

The goal of this project is identifying the strongest factors influencing the dropout rate and to use machine learning to model this rate.

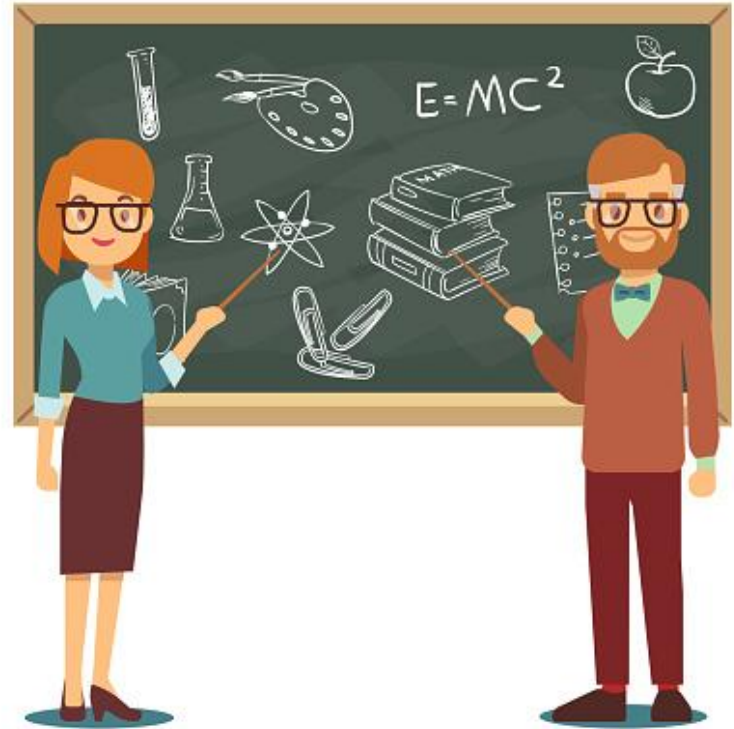
Can we identify students before they drop out?



Stakeholders

This dataset is supported by program SATDAP - grant POCI-05-5762-FSE-000191, through the National School of Public Administration Portugal.

Using models like these educators can focus on students that show a higher likelihood of dropout.



Data Overview

For what purpose was the dataset created? The dataset was created in a project that aims to contribute to the reduction of academic dropout and failure in higher education, by using machine learning techniques to identify students at risk at an early stage of their academic path, so that strategies to support them can be put into place. The dataset includes information known at the time of student enrollment – academic path, demographics, and social-economic factors. The problem is formulated as a two category classification task (dropout, and graduate).

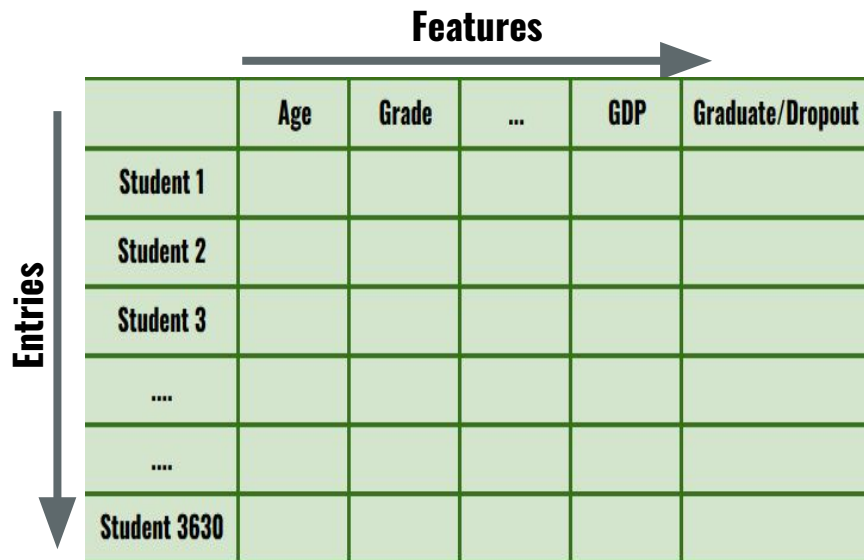
Who funded the creation of the dataset? This dataset is supported by program SATDAP - Capacitação da Administração Pública under grant POCI-05-5762-FSE-000191, Portugal.

Citation Requests/Acknowledgements: [M.V.Martins, D. Tolledo, J. Machado, L. M.T. Baptista, V.Realinho. \(2021\) "Early prediction of student's performance in higher education: a case study" Trends and Applications in Information Systems and Technologies, vol.1, in Advances in Intelligent Systems and Computing series. Springer. DOI: 10.1007/978-3-030-72657-7_16](#)

License: This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. This allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given.

Data overview

- There are 3630 entries described by 37 features.
- 1 Target column ['Graduate', 'Dropout']
- Social Features
 - Marital status, Nationality, Mother's/Father's qualifications, Mother's/Father's occupation, Displaced, Gender, International
- Academic Features
 - Application mode, Application order, Course, Daytime/Evening attendance, Previous qualification, Previous qualification (grade), Admission grade, Educational special needs, Curricular units 1st/2nd semester credited/enrolled/evaluations/.../ approved/grade/without evaluation
- Financial Features
 - Debtor, Tuition fees up to date, Unemployment rate, Inflation rate, GDP
- The data was screened for duplicate and missing values.

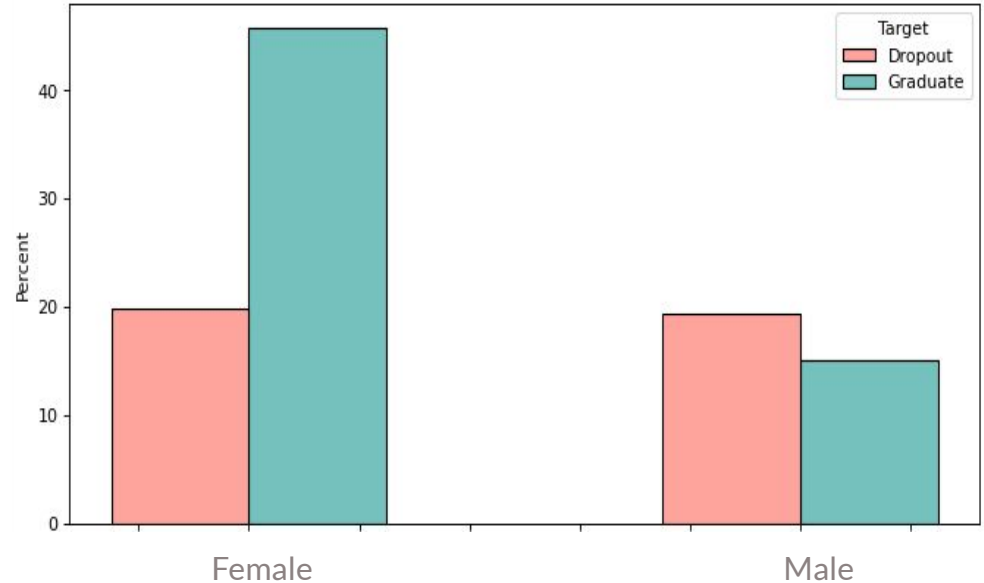


The diagram illustrates a data table structure. A horizontal arrow labeled 'Features' points to the right, indicating the columns. A vertical arrow labeled 'Entries' points downwards, indicating the rows. The table has 6 columns: an empty header cell, 'Age', 'Grade', '...', 'GDP', and 'Graduate/Dropout'. The rows are labeled 'Student 1', 'Student 2', 'Student 3', '...', '...', and 'Student 3630'.

	Age	Grade	...	GDP	Graduate/Dropout
Student 1					
Student 2					
Student 3					
...					
...					
Student 3630					

Key Findings: Social Features

- Graduates have a 3:1 female to male ratio
- Dropouts have a even distribution of female to male students
- Approximately 65% of students surveyed are Female.

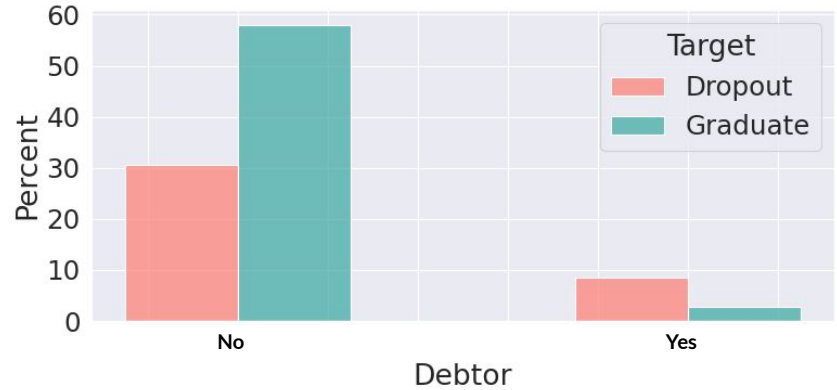


Key Findings: Financial Features

Financial Features

Students with lower debt and up to date tuition fees tend to Graduate.

	Debtor	Tuition fees up to date
Graduates Subtotal	4.6%	98.6%
Dropout Subtotal	22%	67.8%



Key Findings: Academic Features

- Good Students graduate!
- Low semester grades show strong correlation to dropouts.
- Graduates tend to have higher median grades
- ...all very intuitive, let's look at some numbers...



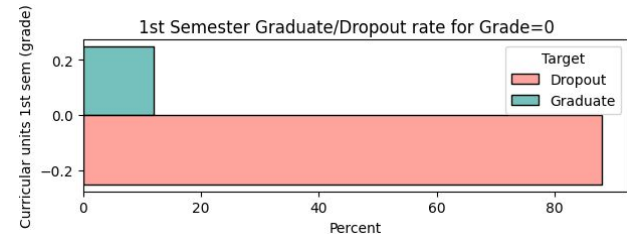
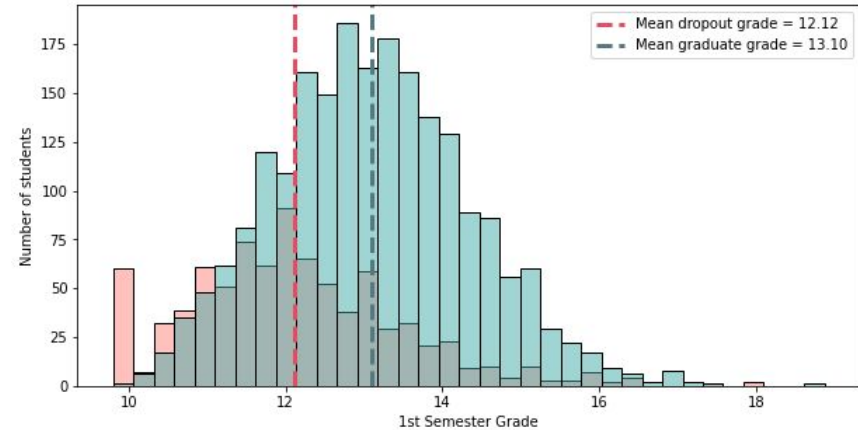
publicdomainvectors.org

1st Semester Grades a closer look...

- **Graduates** shown in blue have a 1st semester mean grade of 13.1 and a uniform distribution about the mean.
- **Dropouts** shown in red have a 1st semester mean grade of 12.1 with a wider distribution about the mean.
- Students who scored 0 in the 1st semester had a 88.1% Dropout rate.

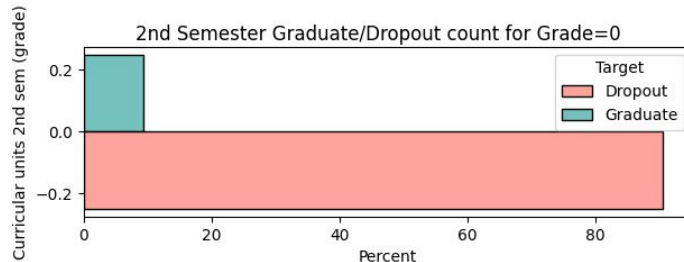
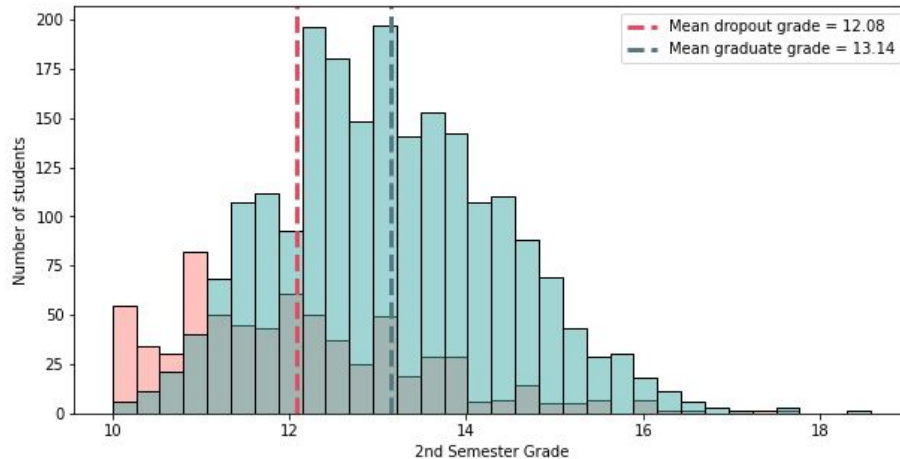
NOTE: This histogram does not include students with 0 grade score for 1st semester. Students with 0-score dropped out at a rate of 570 to 77. Students with 0-score are assumed to have not completed the semester and therefore their grade is not shown in the grade distribution histogram.

Dropout and Graduate grade distribution (1st Sem)



2nd Semester Grades a closer look...

Dropout and Graduate grade distribution (2nd Sem)



- Graduates shown in blue have a second semester mean graduate grade of 13.14
- Dropouts shown in red have a second semester mean grade of 12.07
- Students who scored of 0 for 2nd semester had a 90.6% Dropouts rate.

NOTE: This histogram does not include students with 0 grade score for 1st semester. Students with 0-score dropped out at a rate of 727 to 75. Students with 0-score are assumed to have not completed the semester and therefore their grade is not shown in the grade distribution histogram.

Machine Learning

Machine Learning is the process of putting a bunch of data (Features) into a funnel or model and spitting out a prediction for a particular feature's value.

In this case we build a model to determine if the student will Graduate or Dropout.

How do we know if we have good model?



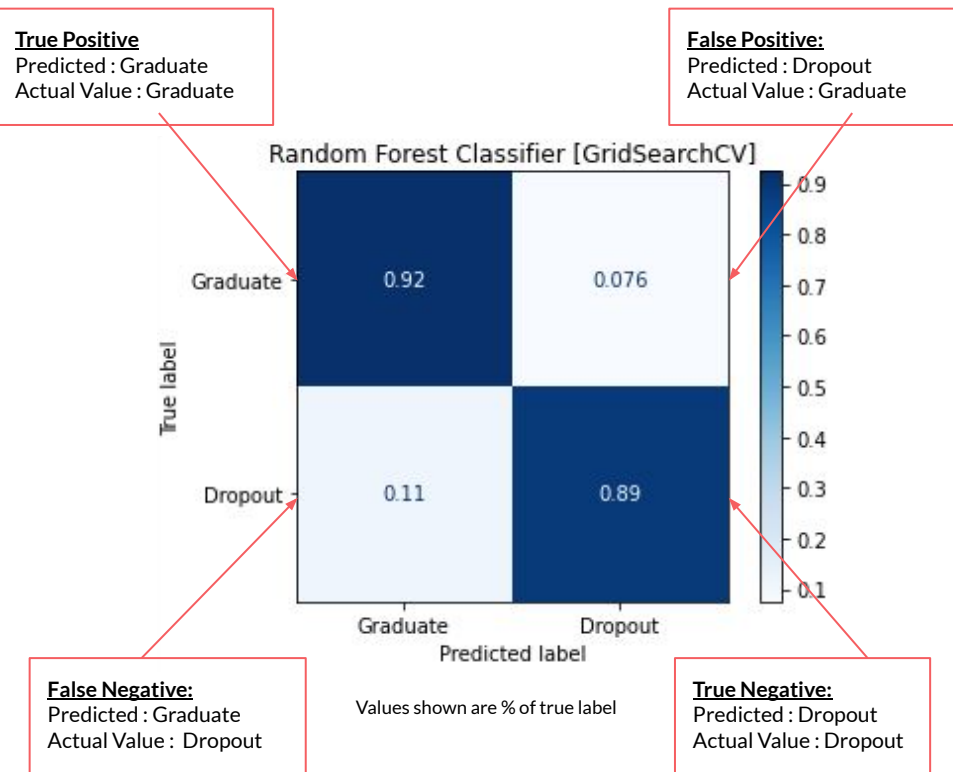
Do we have a good model?

To model this data we are looking to classify students as 'Graduate' or 'Dropout'. This is a binary classification problem.

Various modeling techniques were tested, tuned, and compared*. The model that showed the highest accuracy with the lowest False Positives and False negatives was an *Undersampled Random Forest classifier tuned with GridSearchCV with no PCA feature reduction*.

This model accurately predicts 89% of the dropouts and 92% of the graduates. It has a 11% False Negative rate and a 7.6% False Positive rate.

*A full comparison of all models used can be found in the GitHub Repository.



Recommendations

- The machine learning model can be used by educators to help predict student Dropout rates
- False Negatives (Students predicted to Graduate, but who drop out) may require additional review
- Student grades are a great metric for success. Closely monitoring grades along with predictive modeling teachers should have a good understanding on how to efficiently divide their time to minimize the Dropout rate.