# Capstone Two Report

Springboard Data Science Course

Bence Takacs

## Problem Statement

The goal of this project was to predict a student's academic status at the end of the school year (graduate, dropout) based on student data such as admission grade, parent's level of education, occupation, etc. A model which successfully predicts the student's status could be used at the beginning or throughout the semester in order to identify which students are likely to dropout or fail the course and thus let the institution intervene with helpful resources to increase the student's chance of passing. Such a tool would be hugely useful to universities and academic institutions as a very important statistic is graduation rate, which would be increased by strategic use of the model developed in this project.

## Data Wrangling

The data set contains 4424 rows with 37 features, including the target column of graduate, dropout, or enrolled. The split between numerical and categorical features was about half and half. I checked the data set for missing or NaN values, and outliers, but did not find

any. The data set was very clean and did not need much wrangling. Because some features had a numerical value which corresponded to a descriptive string, I created a function which creates a dictionary that relates the numeric value to its descriptive string.

## Exploratory Data Analysis

The most likely features to determine the target status at the end of the year were the number of credits earned in a semester, but I wanted to explore other likely factors as well, like admission grade, whether tuition had been paid, age at enrollment, and gender. Therefore, I explored trends mainly related to these features. Because the goal with this dataset was categorization, I performed some inferential statistics to explore this data set. The first was a t-test to see how admission grades were related to gender, but I got a *p*-value of 0.58, indicating no statistically significant difference between the genders in relation to admission grades. The next test was a $\chi^2$ to see if the target feature, graduate or dropout, was related to a student's status as a scholarship holder. This yielded a *p*-value of $10^{-90}$, indicating a strong correlation and that the two features are not independent of each other. Finally, I ran a correlation test between age at enrollment and admission grade, where I used a Spearman correlation due to the non-normal distribution of ages at enrollment. This test yielded a *p*-value of $10^{-11}$, a strong correlation between age at

enrollment and admission grade. Other interesting statistics I found were that the gender of students who dropped out were fairly equal, while the gender of students who passed were very skewed towards females in an approximately 1:3 ratio of male to female. Because of these tests I expected age at enrollment and admission grade have a large feature importance when modeling, alongside the curricular units, which I did see.

## Pre-Processing

To pre-process the data for modeling, I removed the 'Enrolled' values from the target feature, as I wanted to focus on students who either dropped or passed the course. I also removed the 'Displaced', 'Unemployment rate', 'Inflation rate', and 'GDP' features as I could not find sufficient information on what they represented and did not want to use data I did not understand. I then split the data into numerical and categorical, scaling the numerical data and one-hot encoding the categorical data. It may have been more prudent to use target encoding instead, as many categorical features had a lot of unique values, but for simplicity of the project I chose one-hot encoding. Finally, I concatenated the features and performed a test-train split on the data with a train size of 75%.

## Modeling

I created five models to evaluate this data set. The first model was a very basic random

forest classifier, which scored 90.6% on the test set and had an AUC of 0.94. The feature importances were topped by number of curricular units, as well as whether tuition fees were up to date, age at enrollment, and admission grade.

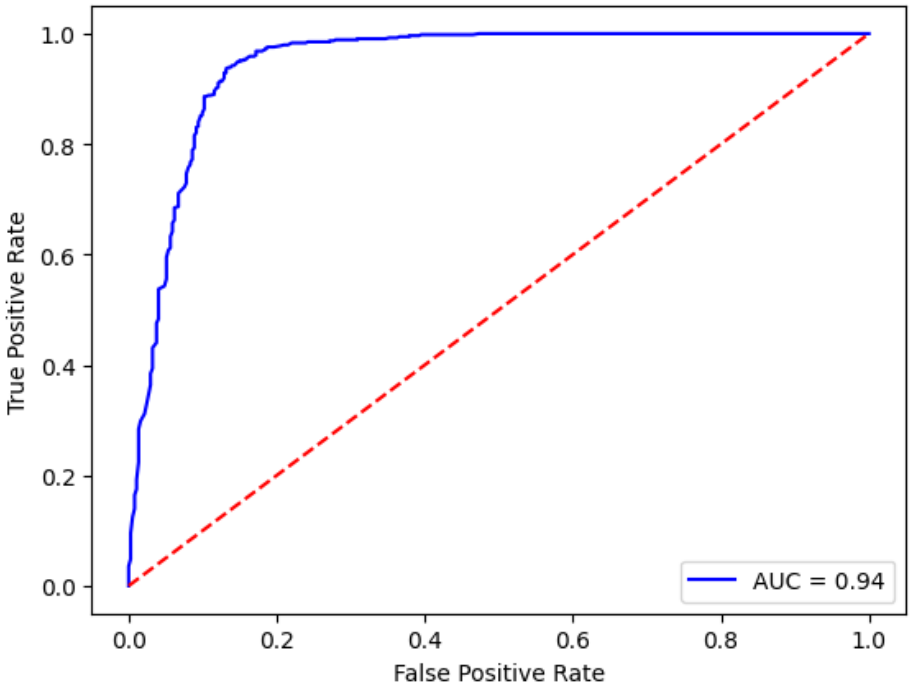


**図 1Figure 1: ROC Curve of Random Forest Model (1)**

The second model was a gradient boosted decision tree, where I first tested several models to find an ideal learning rate of 0.25; the model had a maximum of 10 features and a maximum depth of 10. It performed with an accuracy of 91.0% on the test set with an AUC of 0.95. The most important feature was by far the number of curricular units in the 2nd semester, but other higher ranked features included tuition fee status and admission grade.

The third model used Gini impurity, where multiple maximum depths were tested to find

the optimum at 5. This model performed with an accuracy of 90.3%, had an AUC of 0.93, and had one feature vastly more important than any other: approved curricular units in the second semester.

The fourth model used the Light gradient boost package, which trains for up to 100 rounds until validation scores no longer improve for 10 rounds. This model had an accuracy of 91.2% and an AUC of 0.95. The most important feature by about half was approved curricular units in the second semester, with the other important features being curricular units but also including admission grade, previous qualification grade and age at enrollment.

The final model used Bayesian optimization to optimize the model parameters, some of which were a learning rate of 0.01 and a maximum depth of 52. This model had an accuracy of 90.9% and an AUC of 0.95. The most important feature was admission grade, followed by curricular units, previous qualification, and age at enrollment. This model also included cross validation.

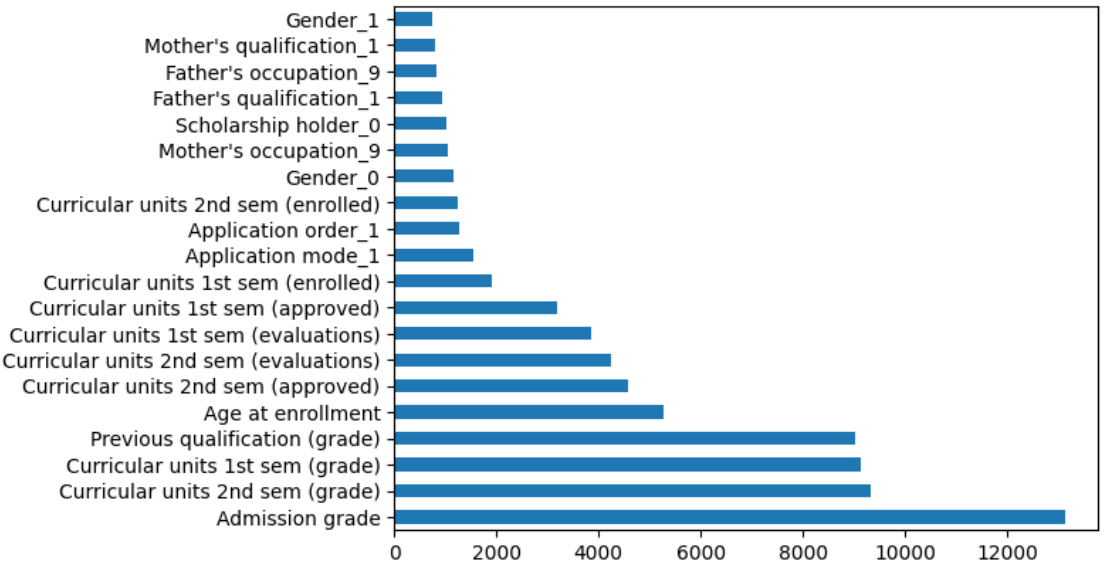


**図 2Figure 2: Feature Importance of Bayesian Optimization Model (5)**

Due to all models having approximately similar performance and feature importance, I chose to use the final model as it included cross validation, making it more accurate in its performance.

## Further Research and Recommendations

I believe further research should include an optimization in reducing the number of features used and maximizing accuracy, as less features required would require less information collected and more money saved for the institution. Further research could also look into how early potential dropout students can be detected, for instance by removing the curricular units from the modeling and looking only at the information of an incoming student. These findings can be used in various ways: to identify students which

are likely to drop out of a class, and thus intervene with strategies to help them pass; to identify which classes have large amounts of students dropping out, and thus investigate a possible issue with the curriculum or the professor; identify students whose tuitions fees are not up to date and are predicted to drop out but would otherwise be predicted to pass, allowing the institution to administer scholarships to 'at-risk' students.

## Code and Models

The finalized code and model metrics can be found in this Github repository.