


[The dataset](#) used for this final project is from the kaggle website and titled “Predict Students' Dropout and Academic Success.” There are three questions that I want to answer using this dataset. The primary objective is to understand which specific predictive factors largely affect student graduation. The analysis aims to provide insights into the following key aspects: 1) how demographic factors such as gender, age at enrollment, marital status, and nationality correlate with student graduation rates. 2) how students' socioeconomic status, represented by debt, tuition fees up to date and scholarship holder, impacts their likelihood of graduation. In the notebook, I did data preprocessing, exploratory data analysis to identify and analyze the risk factors on students' graduation, and trained three models that help confirm hypotheses.

The whole dataset contains 4424 rows (4424 subjects) and 35 columns (35 variables). All the data have been encoded by the producers, and no missing data were found. The primary variable “Target” is in the last column, giving information of whether a subject is graduated or dropout. According to this [article](#), all the variables can be classified into six categories – demographic data, socioeconomic data, macroeconomic data, academic data at enrollment, academic data at the end of 1st semester, and academic data at the end of 2nd semester (See  Table 1).

Firstly, let's find the correlation between the independent variables, their relationships with the target, and remove unwanted or irrelevant features from the data. Columns with an absolute correlation coefficient less than 0.01 were removed, which are course, nationality, father's qualification, educational special needs, international, and unemployment rate. These columns have very low absolute correlation values and may not provide significant predictive power for 'Target.' I then excluded highly imbalanced predictors which are the previous qualification, nationality, educational special needs, and international. Finally, I summed up curricular units that are highly correlated with each other.

The correlation plot below helped to answer the primary research question. Specifically, curricular units 1st sem (grade), curricular units 1st sem (approved), curricular units 2nd sem

(grade), curricular units 2nd sem (approved), and tuition fee up to date, are highly correlated with the target (Figure 1). Plots on the right show that students who graduate have higher approved units and grades overall

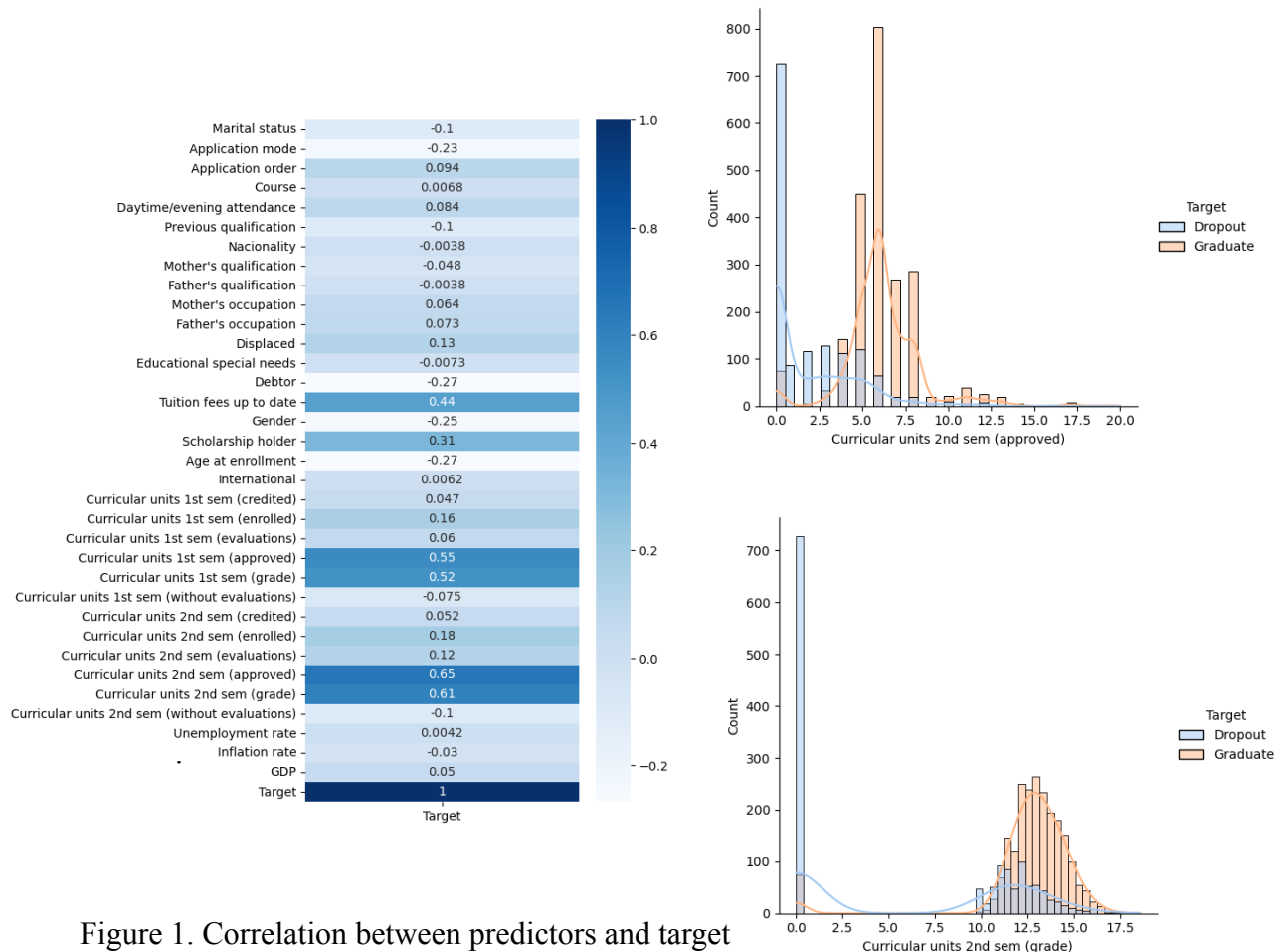
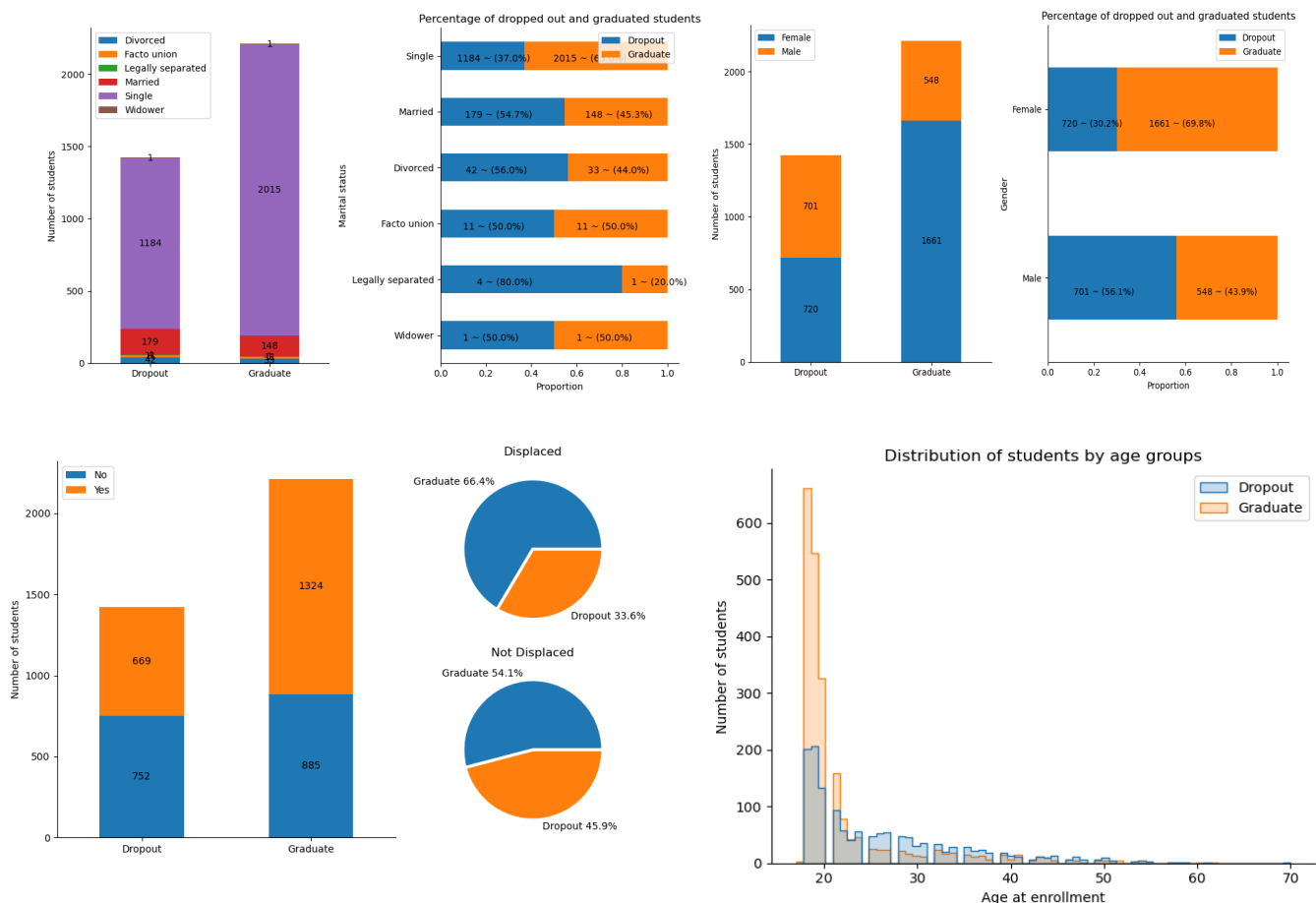


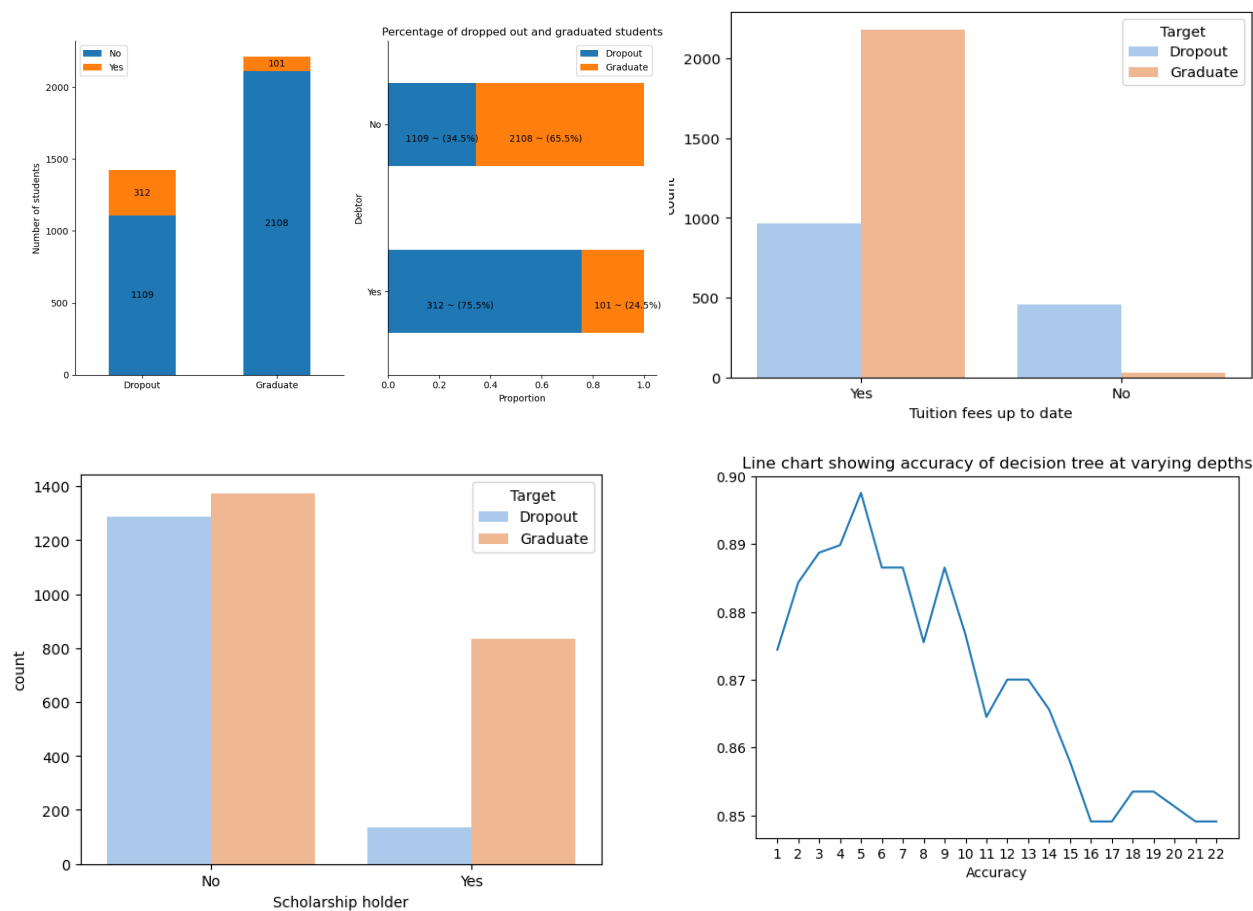
Figure 1. Correlation between predictors and target

Now let's examine the demographic data. In terms of marital status, the majority of enrolled students are single. The percentage of single students who graduated (63%) is higher than the percentage of single students who dropped out (37%). Students with a marital status Legally separated have the lowest percentage of graduation (20%), followed by Divorced (44%) and Married (45.3%). In terms of gender, even though there are more female students than male students in the dataset, the number of dropout female and male students is almost equal, meaning the percentage of graduate male students is lower than the percentage of graduate female students. In terms of whether a student is displaced or not, surprisingly, students who are

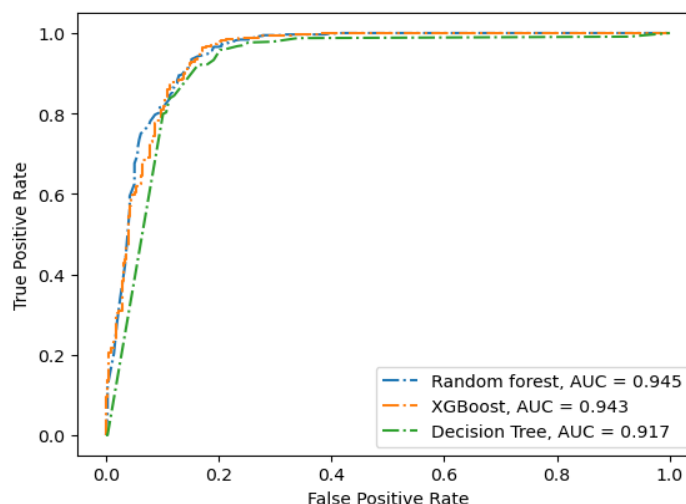
displaced have higher graduation rates than those who are not. I assumed the reasons could be displaced students are resilient to overcome challenges. To sum up, demographic data do have an impact on students' graduation rate with female students, students who enroll before 25, students who are single, and those who are displaced are much more likely to graduate.



The following graphs helped us understand how students' socioeconomic status, represented by debt, tuition fees up to date and scholarship holder, impacts their likelihood of graduation. Specifically, students without debt are almost two times more likely to graduate than to drop out. It is noteworthy that more than 90% of students who did not pay tuition on time end up not graduating and scholarship recipients are much more likely to graduate. To conclude, students are more likely to graduate if tuition is paid on time, if they are not in debt, and if they are scholarship holders.



Now, the dataset is ready for modeling. Since there are many categorical variables, the first model I used is the random tree classifier. I first plotted a line chart showing the accuracy of the decision tree at varying depths and found that the decision tree has the best performance at depth 5 (See figure above). Given that the decision tree model is a greedy optimization method, and we have a large number of datasets, I then used a random forest classifier and XGBoost classifier. Comparing the three classifiers, the decision tree model has the lowest accuracy score, and random forest (AUC: 0.945) and XGBoost (AUC: 0.943) perform equally well.



To conclude, let's see the feature importance in random forest and XGBoost models. Firstly, the approved curricular unit got the highest weight in both two models, confirming that academic data largely affects student graduation rate. Secondly, consistent with our hypothesis on socioeconomic factors, tuition fees up to date plays a crucial role in the likelihood of graduation. It weights up to 10% in random forest models and more than 25% in XGBoost. It is worth noting that results from EDA are more similar to XGBoost, i.e., socioeconomic factors have a relatively important role in whether a student graduates or not. Specifically, the scholarship holder and the debtor are the fourth and fifth important factors in XGBoost. Lastly, two models both confirm that although weak, the demographic factors, such as gender, age at enrollment and marital status impact the graduation.

