Exploratory Data Analysis Report

Linh Vu & Duc Nguyen

2024-04-21

Research Question and Data Set

The data set we use was retrieved from UC Irvine Machine Learning Repository. Spanning 36 predictors and one response variable, the data represent 4424 student records of a school in Portugal in two years 2008-09 and 2018-19. This project aims to answer the primary research question: "What factors are the most predictive of student graduate outcomes in higher education?"

Candidate Variables

Our response variable `Target` has three levels: `Dropout`, `Enrolled`, `Graduate.` By definition, `Graduate` means that students graduated within the expected timeframe. `Enrolled` means students needed up to three additional years to complete their degree. `Dropout` means students needed more than three extra years or did not graduate at all. We first looked at the list of variables and identified the following variables that might be of interest.

- **Parents**: We believe parents' have some influence on their children's schooling. In the context of our data set, the variables are mother's and father's `qualification` and `occupation`.
- **Student's past achievement**: Intuitively, students who did well previously tend to do well in college. Thus, we are interested in `Admission grade`, `Previous qualification`, and `Scholarship holder`.
- **Student's recent performance**: Students' performance in early college might better inform us on their eventual outcome. Thus, we look into their coursework in the 1st and 2nd semesters.

Data Wrangling

We imported the data set to R using UCI's URL. We began by checking if there were any missing values in the data set, which we did not find any. Next, we used `str` to examine the data set and found out that all categorical variables are formatted as numerical. To troubleshoot, we used both `mutate` and `as.factor` to recode each predictor as factors. In addition, we renamed some columns for ease of use, e.g., `Nacionality` into `Nationality`.

We continued with data wrangling. Variables on qualifications (father's, mother's, and student's previous qualification) have over 20 levels, so we grouped them into 5 umbrella levels: `Basic_Education`, `Secondary_Education`, `Higher_Education`, `Professional_Technical`, and `Unknown_None`. This is to make sure that there are sufficient data points in each category and for model building purposes.

Another variable type that has many levels is occupation, father's and mother's. Similarly, we collapsed them into fewer levels: `High_Level_Professionals, Intermediate_Professionals, Skilled_Workers, Unskilled_Workers, Armed_Forces`, etc. We might need to collapse those factors further in the future, but we can keep them as such for now.

A variable may have levels that might be insignificant, too. For example, `Marital Status` originally had 6 levels, two of which, `Single` and `Married`, make up over 97% of the data. Meanwhile, some of the remaining levels had less than 10 counts. Therefore, we collapsed the remaining four levels into one, called `Other`. Our data set is now ready for exploratory analysis.

Exploratory Data Analysis

In choosing our variables, we began with forward selection. First, we split up the predictors into quantitative and categorical groups. We applied the five-number summary and plotted a histogram for each quantitative variable. Based on the detected skewness/outlier and our own intuition of what matters to students' dropout tendency, we selected 14 out of 20 quantitative variables. Among those selected, we created a correlation matrix to investigate possible multicollinearity despite knowing the strength of tree models. We noticed that the statistics from the first semester of college is correlated with those of the second semester. For example, `Curricular units 1st sem (enrolled)` is highly correlated with `Curricular units 2nd sem (enrolled)`. Applying field knowledge, we decided to merge the variables of first and second semester into a full year one (e.g., `Curricular units all year (enrolled)`) by taking the average. There was still some correlation between the newly-mutated variables, so we proceeded with caution.

We created boxplots for each selected quantitative variable against `Target`. From visual inspection, we dropped the variables with weak associations to the `Target` and kept 2 variables, `Curricular units all year (grade)` and `Curricular units all year (approved)`. We added back `Age at enrollment` despite weak association so that we can have more data to make predictions.

We examined the proportions of levels in each of the 16 categorical variables. Taking into account class imbalance, levels with too few counts, variables with too many levels, and our intuition of what matters to dropout rate, we omitted 7 and kept 9 predictors. Then, we visualized to examine their relationship with `Target`. Of note were 5 variables, which were the only ones we kept: `Mother's qualifcation, Father's qualification, Debtor, Tuition fees up to date`, and `Scholarship holder`.

Next Steps for Model Building

We have selected 5 categorical and 3 quantitative predictors, which will be used for our baseline model. Because our response variable has 3 classes, we primarily plan to build tree-based classification models with different complexity levels. In consultation with Prof. Wells, we will learn how to implement logistic regression, random forest, and support vector machine. We may attempt Naive Bayes with great caution of its Naive assumption. We have thought about KNN but realized it only fits pure prediction, as opposed to inference. Lastly, we may use backward selection as a benchmark for our models.