Data

For this challenge, the dataset on predicting students' dropout and academic success is used. The dataset was compiled from various databases within a higher education institution related to students enrolled in different undergraduate degrees. It includes a wide range of information available at the time of students' enrollment, such as academic path, demographics, social-economic factors. The primary objective involves constructing classification models to anticipate students' outcomes, specifically predicting dropout and academic success.

Research question

These predictors can be put into different categories, one being the individual predictors, which provide information on academic path, demographics, and social-economic situations of the students, another being the macroeconomic factors, which are not specific to individual students or programs but may still have an impact on student outcomes. Here I'd focus on the effect of the factors pertaining to the individuals on whether or not the students dropped out. One reason for this is that, according to chapter 14 in the book "Data Science in Education Using R", we shouldn't have more than 20 predictors because it'd negatively influence the accuracy of the model. The research question is thus, **"How do factors pertaining to the individual affect students' dropout?"**. Our predictors are as follows.

```
 [1] "Marital status"                "Course"                        "Previous qualification"
 [4] "Previous qualification (grade)" "Nacionality"                   "Mother's qualification"
 [7] "Father's qualification"         "Mother's occupation"           "Father's occupation"
[10] "Admission grade"                "Displaced"                     "Educational special needs"
[13] "Debtor"                         "Tuition fees up to date"       "Gender"
[16] "Scholarship holder"             "Age at enrollment"             "International"
```

Methods

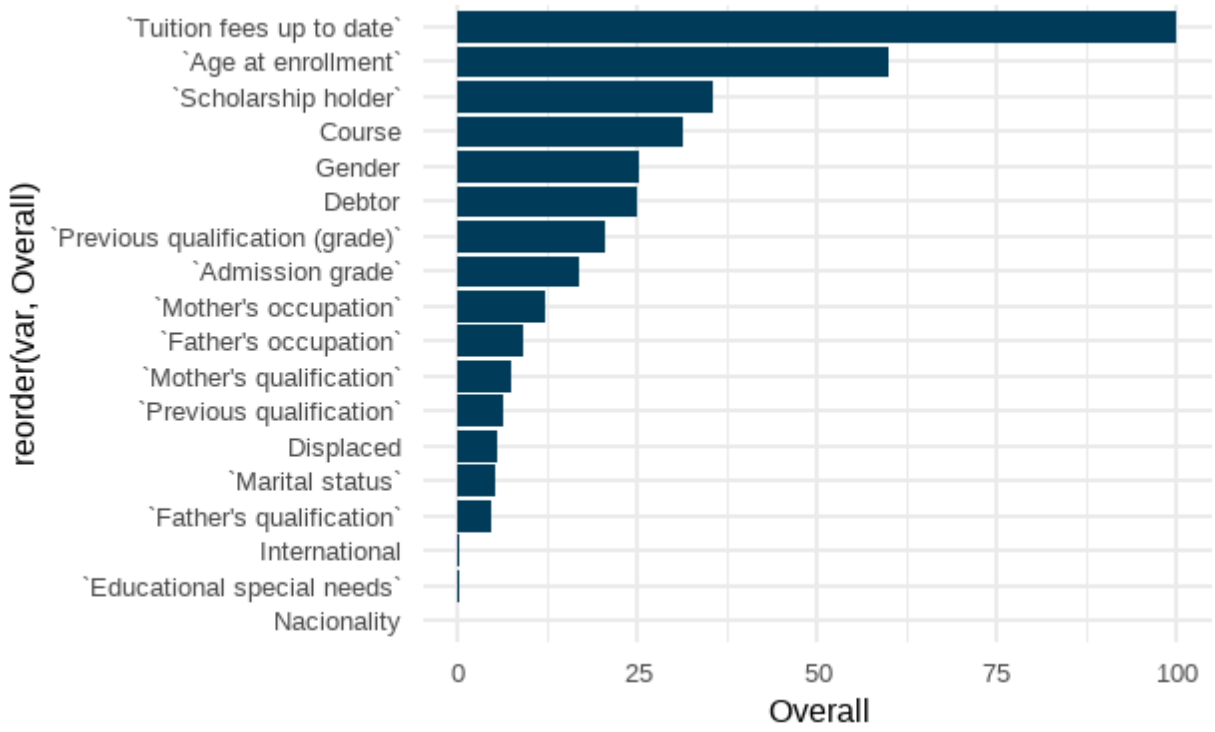
We use both predictive and multilevel models for this question. A random forest model is able to explore unexpected patterns, while a multilevel logistic regression model can provide us with coefficients that make it easier to interpret the relationships. We run a multilevel multinomial logistic regression model, with "Course" passed as an argument for grouping. Multinomial

logistic regression model is “used when the outcome variable being predicted is nominal and has more than two categories that do not have a given rank or order” NU Academic Success Center), our response variable “Target” has 3 values: Dropout, Graduate and Enrolled, therefore multinomial logistic regression model is an appropriate choice.

Results

Random forest

We use 80% of the dataset for the training model, which is the split recommended by the authors of the dataset. The resulting model is expected to make correct predictions about 62.98% of the time, and the accuracy for the test data is a bit higher than for the training data, which suggests that the model is able to handle new data. The importance of the predictor variables is visualized as follows.



From this graph, we can see that “tuition fees up to date”, “age at enrollment”, “scholarship holder”,”course”, “gender”, “debtor” are the most important predictors in the model. However,we can't see exactly how the predictors influence the likelihood of a student dropping out. We know

that "tuition fees up to date" is important, but we don't know if that is a predictor for a student being enrolled, already graduating, or dropping out (though we can certainly take a guess).

Multilevel Multinomial Logistic Regression Model

We can see the impact of each factor on students' dropout by interpreting the odds ratios.

```
          (Intercept) `Marital status`   Course `Previous qualification` `Previous qualification (grade)`
Enrolled    0.7049016         1.172393 1.000022                1.0061532                          0.9904684
Graduate    0.1371384         1.095798 1.000034                0.9998735                          0.9973304
          Nacionality `Mother's qualification` `Father's qualification` `Mother's occupation`
Enrolled    0.9814668                0.9894984                1.002482              1.009242
Graduate    0.9661596                0.9945142                1.002312              1.005980
          `Father's occupation` `Admission grade` Displaced `Educational special needs`    Debtor
Enrolled              0.9977605          1.006606 0.9047085                   0.8371480 0.8638060
Graduate              0.9974467          1.022208 1.0244107                   0.5636181 0.3961814
          `Tuition fees up to date`    Gender `Scholarship holder` `Age at enrollment` International
Enrolled                  7.738093 0.7109785             1.481725           0.9451647      3.890958
Graduate                 23.262659 0.4276575             4.271941           0.9537846      6.536192
          1 | CourseTRUE
Enrolled       0.7049016
Graduate       0.1371384
```

From this result, we can see that similar to our findings from the random forest model, tuition fees up to date is indeed the most important factor in the model. Holding all other predictors constant, a one-unit increase in "Age at enrollment" is associated with a 7.74 times increase in the odds of being "Enrolled" and 23.26 times increase in the odds of being a graduate compared to being a dropout. Being a scholarship holder will also considerably increase the likelihood of someone graduating. On the other hand, being a debtor or having educational special needs are associated with a significant decrease in the chance of graduating, which is something we don't see reflected in the random forest.

What is surprising (and confusing) is that "international", which was not an important predictor according to the random forest, is the second most important factor here in this logistic regression model. Being an international student is associated with a substantial increase in the odds of being enrolled or graduating compared to being a dropout.

On the contrary, there are a few predictors deemed quite important by the random forest but not by logistic regression, such as "Age at enrollment" and "Course". This suggests that there might be relationships between "Age at enrollment" / "Course" and our response variable "Target", which can't be modeled with linear regression.

Sources

Libguides: Statistics Resources: Multinomial logistic regression. Multinomial Logistic Regression - Statistics Resources - LibGuides at Northcentral University. (n.d.). https://resources.nu.edu/statsresources/Multinomiallogistic

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T., & Realinho, V. (n.d.). *Predict students' dropout and academic success*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Ryan A. Estrellado, E. A. B. (n.d.-a). *Data science in education using R*. 13 Walkthrough 7: The role (and usefulness) of multilevel models. https://datascienceineducation.com/c13

Ryan A. Estrellado, E. A. B. (n.d.-b). *Data science in education using R*. 14 Walkthrough 8: Predicting students' final grades using machine learning methods with online course data. https://datascienceineducation.com/c14