Dropout factors

The relationship between Dropout of Education and its causes: Gender, Debtor, Knowledge requiremetns, and Daytime Attandence

Xiaoou Du
T00602882
DASC 5420

Aspect

Nowadays, Education is one of the most important fields in our society, and the labor force with advanced education is the basis of any industry. In other words, it is important for children (even for most adults) to be educated in today's world. However, there exist risks that students may drop out, and educators are interested in the reasons for this phenomenon. We need to know what factors influence our students and lead them to drop out. Thus the core idea of this project is to predict Student Retention and identify the risk factors for dropout. Using data from *Predict Students' Dropout and Academic Success*, we would be able to find the factors that may affect the status of students. In this project, Elastic Net Regression has been chosen to construct the model, then developed via Both-direction stepwise. After analysis, the top factors that have a negative influence on students are Debtor, Gender, Educational.special.needs, and Daytime.evening.attendance. According to the analysis results, we would be able to generalize the characteristics of dropout students. As educators, it means we need to pay more attention to the students who have these characteristics.

Introduction

Since the first Industry Revolution, the importance of Education increased significantly as the economic level increased; nowadays, it has already become one of the most essential fields in our society. There is no doubt that the labor force with advanced education is an essential and indispensable resource for any industry - whether it is the primary, secondary, or tertiary industries, those people who have higher and more suitable education experiences would be the basis of the developments. To enlarge the storage of talent, the most outstanding measure taken by many governments is Compulsory Education, which "minimizes the number of students who stop going to school because of family and economic reasons and balances the education differences between rural and urban areas". In other words, not only all children but also most adults are necessary to be educated so that they would be able to have abilities to live in our society.

However, there still exist some students who drop out of education. Educators are interested in the risks that students may drop out and the factors that lead to this phenomenon. As educators, we need to know the reasons why students may drop out of education, and what are the characteristics of them.

Therefore, the core target of this project is to predict Student Retention and identify the risk factors for dropout. In this project, we are going to find and analyze the relationship between Student status (Dropout, Enrolled, and Graduated) and other variables, then find out the factors that have a negative influence on students (in other words, the factors that may cause students to drop out). Moreover, we would like to analyze how significant these factors are: which factor has the major influence on dropout? What factors are able to reduce the risk of the dropout? And what factors actually have a slight effect on dropout that is different from our common sense?

Background

In this section, we are going to discuss the necessity of this project. Our core target is finding the relationship between Student Status and the factors that may affect it. In other words, we majorly focus on the reasons and results of dropout.

In detail, dropping out needs to be separated into two parts: **during Compulsory Education, and during Advanced Education**.

The former one is easy to understand, children have to leave school, then lack education because of subjective surrounding reasons - such as **financial problems** (mostly happen in areas where have no compulsory education, or their government is unable to bear the education), **living environment problems** (they live in somewhere fall into War, Disease, or Nature Disaster), and **Civil & Humanity problems** (the gender problems, religious persecutions, and racism). Most of these reasons are unable to be dealt with by educators, so we would talk less about them, and pay more attention to the other part.

The second part of dropping out is dropping out of Advanced Education. The most common case is first-year students giving up on their university life. According to the research of Liga Paura, *the main reasons for dropping out are students' low secondary school knowledge and low motivation to study engineering*. (Liga Paura, *Cause Analysis of Students' Dropout Rate*, 2014)
In other words, on one hand, we could say that students drop out of advanced education because of insufficient secondary education. In this case, the dropping out students do not have enough knowledge to support them to understand what they are going to learn in university, then they find it hard to keep learning and finally have to drop out.
On the other hand, low motivation is another major cause of dropping out. If the lack of secondary education knowledge could be concluded as an Objective Reason, then the low motivation would be regarded as a Subjective Reason. The low motivation could be the result of several aspects, such as **boredom from difficult and repeated work** (it is extremely easily bored with the work that cannot receive positive feedback in a short period), **the balance between inputs and relevant outputs** (according to the research of Liga, if the relevant outputs are lower than inputs, students would be easy to drop out), and **the decreased interest in majors** (the most common reason is a poor employment environment).

When the reasons for dropping out are clear, we are going to understand the results of dropping out. In most people's minds, dropping out is something personal - no matter whether someone is a dropout or not, this is just his own business, with no influence on others.
It cannot be said that this kind of thinking is wrong, it is just relatively one-sided. In fact, the most outstanding result of dropping out is **the decrease in talent storage** - there is less fresh blood in some industries, and this phenomenon would cause an increase in labor costs.
Moreover, another result of dropping out is **Class Solidification** - people find it hard to choose a work that is not related to their families' industries. It means that people who drop out find it hard to develop their social status by working in a more "successful" or "respected" industry.

However, those results are most social problems, which are hard to be influenced by educators (in my opinion, the government should be the major force to solve them). Therefore, we would not talk too much about the result of dropping out in this project and would focus on the factors that affect students dropping out of education.

Data

In this project, we would use data from *Predict students' dropout and academic success* (Valentim Realinho, 2021) to find the factors that influence students dropping out of education.

This dataset can be used to understand and predict student dropouts and academic outcomes. It includes demographic data, social-economic factors, and academic performance information that can be used to analyze the possible predictors of student dropout and academic success. This dataset contains multiple disjoint databases consisting of relevant information available at the time of enrollment, such as application mode, marital status, the course chosen, and more. (Valentim Realinho, 2021)

The research idea would be "*Which specific predictive factors are linked with student dropout or completion?*"

In detail, this data set includes 35 variables in total, most of them are categorical and the rest of them are numerical. (The original data description is in the section Cited Reference). The size of this data set is more than 4000, which we believe is large enough to be reliable.

Because of the research target, we would not use all of them in our research, which means there would be some variables removed during Pre-Processing. Moreover, it is worth noticing that there are some variables that are Binary - either 1 or 0. For example, Debater (0 means not, and 1 means yes) and Gendar (0 for males, 1 for females).

In Pre-Processing, we remove the data that lacked elements (with N/A elements) at **first**. Fortunately, the data set is complete, there is no missing value in the data, so the omitted result is the same as an original data set.

The **second** step is removing the useless variables - the grade, approval, and evaluation in the first & second semesters. We choose to remove these variables because they are dependent variables and the outcome of another research (the prediction of study grades). If we include these variables in this project, the direct result is the number of variables in the final model increase, which causes the result of the analysis to have a very high error rate and be unreliable.

Thirdly, we need to modify data in this step. The outcome variable Student Status (which is named Target) in the data set is neither numeric nor categorical, it is a character variable, which is not able to be used in a mathematical calculation. In order to construct a model with the outcome, we need to convert it from characters to factors. During this process, we record the status Dropout as "1", Enrolled as "2", and Graduated as "3". In the section Result, we would explain it in detail.

The **fourth** step is an optional step, normalizing the omitted data set. In my opinion, we'd better not do the scale of the data set in this project, because there are many Categorical and Binary variables. According to the result of the analysis (both the result based on Normalized data and the result bases on Omitted data would be shown in the Method section), we could see that if we use normalized data to do analysis, the characteristic of the model is not significant enough to identify which factor definitely influences dropping out significantly or slightly.

Method

In this project, we would choose the best model from the methods we are familiar with - Simple Linear Regression, Ridge Regression, LASSO Regression, and Elastic

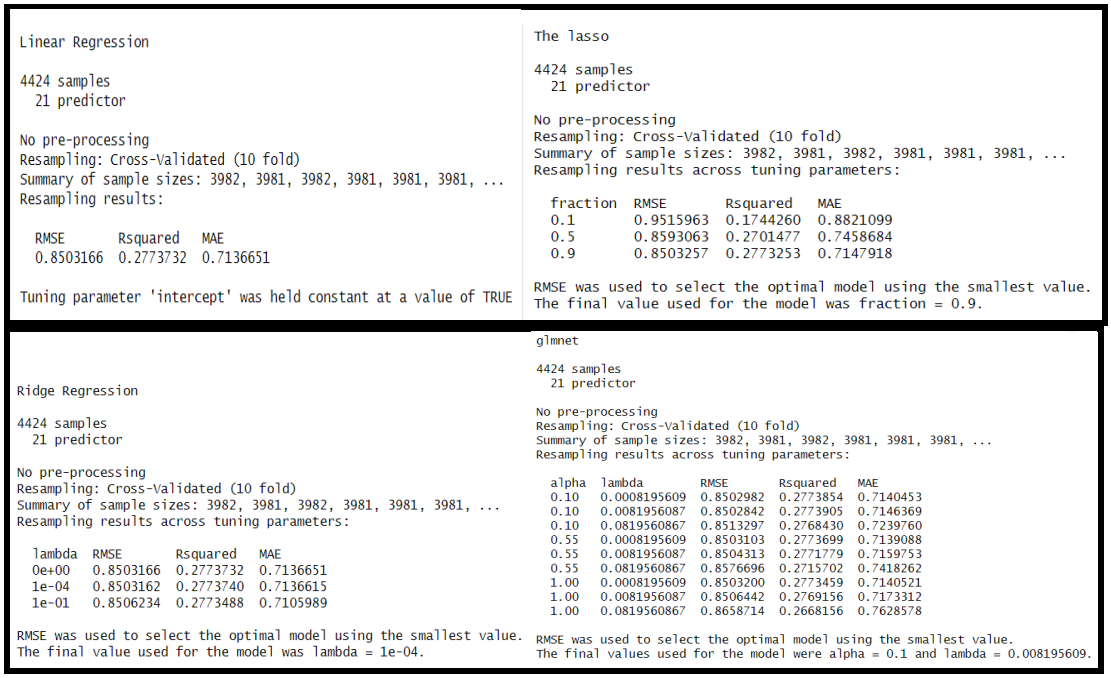
Net Regression. Even though we know that Elastic Net Regression is better than Ridge and LASSO Regression because it is the combination of these two regressions, we still need to use Cross Validation to identify which method among these four methods is the most suitable one - in this case, we would like to choose the method with smallest RMSE, which means this method has the highest accuracy.

The prototype formula we use to construct the basic model is the general formula that includes all variables. The model development would be done after model choosing via Cross-Validation.

The first step of Cross Validation is setting the train control method. We get used to using K-fold to do the cross-validation, where K = 5 or 10 in default. In this project, we choose K = 10 since the data size is large.

After setting, we are going to test methods one by one, from Simple Linear Regression to Elastic Net Regression. In this process, since K-fold cross-validation is a random process (the data would be separated into K subsets randomly, and use one of them as a test set, and others as training sets), the result would be different if we do not set a random seed - it means, there is a possibility that getting different results when repeating the experience. Thus, we need to set the random seed before doing cross-validation. We use seed 2023 here.

We could get the result of the best model of Simple Linear Regression, Ridge Regression, LASSO Regression, and Elastic Net Regression after cross-validation. As we mentioned before, we would like to use RMSE to be the standard to compare the results of models.
Here are the results of 4 methods in cross-validation.



```
Linear Regression

4424 samples
  21 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3982, 3981, 3982, 3981, 3981, 3981, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.8503166  0.2773732  0.7136651

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
The lasso

4424 samples
  21 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3982, 3981, 3982, 3981, 3981, 3981, ...
Resampling results across tuning parameters:

  fraction  RMSE       Rsquared   MAE
  0.1       0.9515963  0.1744260  0.8821099
  0.5       0.8593063  0.2701477  0.7458684
  0.9       0.8503257  0.2773253  0.7147918

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was fraction = 0.9.
```

```
Ridge Regression

4424 samples
  21 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3982, 3981, 3982, 3981, 3981, 3981, ...
Resampling results across tuning parameters:

  lambda  RMSE       Rsquared   MAE
  0e+00   0.8503166  0.2773732  0.7136651
  1e-04   0.8503162  0.2773740  0.7136615
  1e-01   0.8506234  0.2773488  0.7105989

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was lambda = 1e-04.
```

```
glmnet

4424 samples
  21 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3982, 3981, 3982, 3981, 3981, 3981, ...
Resampling results across tuning parameters:

  alpha  lambda       RMSE       Rsquared   MAE
  0.10   0.0008195609  0.8502982  0.2773854  0.7140453
  0.10   0.0081956087  0.8502842  0.2773905  0.7146369
  0.10   0.0819560867  0.8513297  0.2768430  0.7239760
  0.55   0.0008195609  0.8503103  0.2773699  0.7139088
  0.55   0.0081956087  0.8504313  0.2771779  0.7159753
  0.55   0.0819560867  0.8576696  0.2715702  0.7418262
  1.00   0.0008195609  0.8503200  0.2773459  0.7140521
  1.00   0.0081956087  0.8506442  0.2769156  0.7173312
  1.00   0.0819560867  0.8658714  0.2668156  0.7628578

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0.1 and lambda = 0.008195609.
```

We could see that the minimum RMSE among these 4 models is 0.8502842 - which is the result of **Elastic Net Regression** with alpha = 0.1 & lambda = 0.008195609. This is the same as our common sense, Elastic Net Regression gives us the most accurate result. Hence we are going to use Elastic Net Regression as our model to test the relationship between the outcome Target and other variables.

Now we can go to the next step with the Elastic Net Regression model we got from cross-validation. The next step is model development, we need to remove the unrelated variables and the variables which are not statistics significantly. The method of development we use is Both Direction Stepwise. (The developed model has been shown in the Cited Reference)
Here we could use either Omitted Data or Normalized Data to construct and develop the Elastic Net Model. The result would give us the same variables, but different coefficients.
However, as we mentioned above, the characteristics of Normalized Data are not significant. The coefficients in this model are too close to 0. In this case, it would be hard to analyze and compare which factor has the weighted influence on Dropping out, because the coefficients are too close to each other.

Description: df [16 × 1]

	Coefficient <dbl>
(Intercept)	-4.463757e-16
Application.mode	-9.270774e-02
Course	-3.967735e-02
Daytime.evenin...	-3.131328e-02
Previous.qualifi...	3.200342e-02
Nacionality	-6.217508e-02
Mother.s.qualifi...	-3.017249e-02
Mother.s.occup...	5.577634e-02
Educational.spe...	-2.342315e-02
Debtor	-8.822989e-02

1–10 of 16 rows

Description: df [16 × 1]

	Coefficient <dbl>
(Intercept)	1.947030609
Application.mode	-0.015540215
Course	-0.008135928
Daytime.evenin...	-0.089176375
Previous.qualifi...	0.007171778
Nacionality	-0.031586083
Mother.s.qualifi...	-0.002969175
Mother.s.occup...	0.012392470
Educational.spe...	-0.194880528
Debtor	-0.246848976

1–10 of 16 rows

(The left-hand side is Normalized data, and the right-hand side is Omitted Data)

Therefore, we would like to use the Omitted Data rather than Normalized data to do the analysis.

Result
According to the final developed model, there are 15 variables that we think influence students' status. Since we are curious about dropping out (which is recorded as “1” in the data set), we would like to focus on the variables that have Negative Influences on the outcome.

Gender	-0.257560770
Debtor	-0.246848976
Educational.special.needs	-0.194880528
Daytime.evening.attendance	-0.089176375

The top 4 variables are Gender, Debtor, Educational Needs, and Daytime attendance. It is interesting that these variables are all Binary variables, so the influences could be concluded as a series of “if” sentences.
This result shows us, under the same conditions:

1. Females have a higher possibility than Males to drop out of education. This is a typical gender problem across the world. According to the research from Shobhit, gender problems & religious problems are two major reasons cause females to drop out of education. (Shobhit, *Gender, caste, and Education in India*, 2018)
2. People who are debtors would have a higher possibility to drop out of education than those who are not debtors. This is a typical economic problem, debtors are put at a high risk that unable to bear their tuition fees. The result of it would be dropping out.
3. If there are special needs in education, the possibility of dropping out would increase. This phenomenon has been proved in Liga's research. They mentioned that *the main reasons for dropping out are students' low secondary school knowledge and low motivation to study engineering*.
4. Daytime attendance is the fourth factor, but it is much weaker than the third one. In order to develop the accuracy of prediction, we include this variable in our analysis. This factor relates to the economical problem. Day attendance education means Full-time education in most cases. Students are to stay in school or in front of their laptops to take online courses. The daytime education causes them to have no time to work. This is different from the debtor, it means decreasing income.

Generally, in order to reduce the rate of dropping out, as educators, we need to focus on the students who have economical problems and gender problems. Similarly, students who have weak on their secondary education also need extra attention from educators.

Reference
https://github.com/SaratogaVictorica/Dropout-of-Education

Valentim Realinho, Jorge Machado, Luís Baptista, & Mónica V. Martins. (2021). Predict students' dropout and academic success (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5777340

Wikipedia, Compulsory Education. https://en.wikipedia.org/wiki/Compulsory_education#:~:text=Nowadays%2C%20compulsory%20education%20has%20been,between%20rural%20and%20urban%20areas.

Liga Paura, Irina Arhipova, Cause Analysis of Students' Dropout Rate in Higher Education Study Program, Procedia - Social and Behavioral Sciences, Volume 109, 2014, Pages 1282-1286, ISSN 1877-0428, https://doi.org/10.1016/j.sbspro.2013.12.625.

Shobhit Goel, Zakir Husain, Gender, caste, and education in India: A cohort-wise study of drop-out from schools, Research in Social Stratification and Mobility, Volume 58, 2018, Pages 54-68, ISSN 0276-5624, https://doi.org/10.1016/j.rssm.2018.10.002.

Cited Reference:

Column name	Description
Marital status	The marital status of the student. (Categorical)
Application mode	The method of application used by the student. (Categorical)
Application order	The order in which the student applied. (Numerical)
Course	The course taken by the student. (Categorical)
Daytime/evening attendance	Whether the student attends classes during the day or in the evening. (Categorical)
Previous qualification	The qualification obtained by the student before enrolling in higher education. (Categorical)
Nacionality	The nationality of the student. (Categorical)
Mother's qualification	The qualification of the student's mother. (Categorical)
Father's qualification	The qualification of the student's father. (Categorical)

Mother's occupation	The occupation of the student's mother. (Categorical)
Father's occupation	The occupation of the student's father. (Categorical)
Displaced	Whether the student is a displaced person. (Categorical)
Educational special needs	Whether the student has any special educational needs. (Categorical)
Debtor	Whether the student is a debtor. (Categorical)
Tuition fees up to date	Whether the student's tuition fees are up to date. (Categorical)
Gender	The gender of the student. (Categorical)
Scholarship holder	Whether the student is a scholarship holder. (Categorical)
Age at enrollment	The age of the student at the time of enrollment. (Numerical)

International	Whether the student is an international student. (Categorical)
Curricular units 1st sem (credited)	The number of curricular units credited by the student in the first semester. (Numerical)
Curricular units 1st sem (enrolled)	The number of curricular units enrolled by the student in the first semester. (Numerical)
Curricular units 1st sem (evaluations)	The number of curricular units evaluated by the student in the first semester. (Numerical)
Curricular units 1st sem (approved)	The number of curricular units approved by the student in the first semester. (Numerical)

The result of developed Elastic Net Model:

	Coefficient
(Intercept)	1.947030609
Application.mode	-0.015540215
Course	-0.008135928
Daytime.evening.attendance	-0.089176375
Previous.qualification	0.007171778

Nacionality	-0.031586083
Mother.s.qualification	-0.002969175
Mother.s.occupation	0.012392470
Educational.special.needs	-0.194880528
Debtor	-0.246848976
Tuition.fees.up.to.date	0.846523770
Gender	-0.257560770
Scholarship.holder	0.405590118
Age.at.enrollment	-0.011656493
International	0.453994487
GDP	0.010232087