Stat_text

2024-05-23

1. Obtaining the Data

We originally found the dataset on Kaggle. The data was collected by the Polytechnic Institute of Portalegre in Portugal to build machine learning models that predict a student's outcome based on various socioeconomic factors and academic performance. This was done to develop an analytics tool for the tutoring program to direct their efforts more effectively.

The dataset was created from several disjoint databases and includes students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and technologies, it consists of 4424 records with 35 features.

The prediction problem is formulated as a three-category classification task, which assesses, based on socio-economic data and performance metrics during the academic years, whether a student will:

- Graduate within the three years of planned course activities ('Graduate')
- Change course or stop studying altogether ('Dropout')
- Fail to graduate in time ('Enrolled')

According to the literature in the field, there is no agreed-upon definition of what constitutes a dropout. In this work, the authors defined dropouts from a micro-perspective, considering field and institution changes as dropouts regardless of when they occur. This approach results in much higher dropout rates than the macro-perspective, which considers only students who leave the higher education system without a degree.

Given that the number of libraries we can use is limited to those covered during the course, we decided to employ this dataset to answer a different research question. Our analysis will focus on building a model to assign a probability score to new students. This score will quantify the likelihood that a student will finish their course within a three-year timeframe based on data collected at enrollment.

A detailed description of the original dataset can be found in Appendix A.

Step 2. Data Preprocessing

Let's start by loading our data and changing the graduated column in order to represent our research question correctly: we are going to convert all the 'Enrolled' labels into 'Dropouts' since they did not manage to complete their course in the scheduled timeframe

```r
data <- read.csv('~/Downloads/Dropout and Success/student_data.csv',
                 sep = ';')
str(data)

# Rename Colums
names(data)[names(data) == 'Nacionality'] <- 'Nationality'
names(data)[names(data) == 'Marital.status'] <- 'Married'
names(data)[names(data) == 'Output'] <- 'is_Graduated'

# Remove 'Enrolled' to have a binomial problem
data$is_Graduated[which(data$is_Graduated == 'Enrolled' |
                       data$is_Graduated == 'Dropout')] <- 'No'
data$is_Graduated[which(data$is_Graduated == 'Graduate')] <- 'Yes'
```

Missing values:

there were no missing values in the dataset.

```
sum(is.na(data))
```

```
## [1] 0
```

Data labeling

We changed the label criteria for some of our variables in order to increase model explainability, in the following section we will discuss and showcase our changes.

Marital status

Categorical variable indicating the marital status of the individual. We only have 4 widowers and 6 legally separated instances, therefore we collapsed them under the variable 'Others', we also decided to merge 'divorced' and 'legally separated' since the difference between the two instances is not relevant for our analysis.

Table 1: Marital status

Status	Freq
No	3919
Yes	505

Mother/Father occupation

Categorical variables indicating the mother and father occupation respectively, while the original dataset had 32 labels for all kinds of different jobs, we decided to merge some of the labels and make this a 3-label categorical variable. Jobs were split based on a White/Blue collar distinction, as shown in the table below.

Table 2: Parent's occupation

Occupation	Mother	Father
Blue Collar	2004	2567
White Collar	2189	1645
Others	231	212

Mother/Father/Student education

Categorical variable indicating the level of each parents' qualification as well as the student's. Again, we are dealing with many categorical variables, so we decided to merge them based on whether they have an completed an higher education cycle and if they have finished high school or not

Table 3: Qualification

Qualification	Mother	Father	Student
No Secondary	234	3076	232
Secondary	3599	933	3988
Higher	591	415	204

Courses

Categorical variable representing the course chosen at enrollment, we originally had 17 different courses, we decided to relabel the courses using whether they are STEM subjects or not as a splitting criterion

Table 4: Course

Course	Freq
Stem	1722
No Stem	2702

Nationality:

categorical variable representing a students nationality: looking at the data we saw that while we had a lot of labels (one for each nationality), the vast majority of people were Portuguese, therefore we labeled data in the following way

Table 5: Nationality

Nationality	Freq
Portoguese	4314
Others	110

FEATURE REMOVAL

Given that to build our model we are going to use only information that was present at enrollment time there are some features which are not useful and will therefore be removed.

1 Curricular data: we removed all information concerning student performance over the span of three years since it doesn't help us answer our research question

2 Application mode/ application order: the original paper didn't offer a clear indication about the various labels of this feature, furthermore we don't believe them to be of any interest as far as our research question goes.

3 Macroeconomics data (Inflation rate, GDP, Unemployment rate): these economic indicators were taken over the course of the three years data collection period, we also don't believe them to be of any use in answering our research question

4 Tuition fees up do date

3. Exploratory Data Analysis

AGGIUNGERE LINK TABELLE DI RIFERIMENTO

To perform EDA we plot contingency tables for our categorical variables, looking at them gives us some insights: the majority of single people manages to graduate in time, the same can't be said for the other two categories.

People who enroll after completing secondary score tend to graduate in time more compared to oter categories (portalegre university allows some students to enroll courses without having having an high-school diploma).

Although there are more students enrolled in non-STEM courses (2,702 vs 1,722), the graduation rate is higher for those in STEM courses (52.3% vs 48.4%).

It's worth noting that mothers generally have a higher level of education compared to fathers. Specifically, 234 mothers versus 3,076 fathers did not complete high school. Interestingly, regardless of parents' educational background, students with parents who graduated from college have a lower graduation rate compared to those whose parents have lower levels of education.

There is no significant difference in graduation rates based on the parents' professions, except for those marked 'Other,' who have a lower on-time graduation rate.

Students who left their parents home to study appear (i.e. the 'displaced' variable) have a higher graduation rate compared to those who are not.

Students who receive a scholarship have a higher graduation rate than those who do not (76.0% vs 41.3%).

Finally, it is evident that women are significantly more likely than men to graduate on time (57.9% vs 35.2%). add info about age at enrollment

Da valutare di togliere le tabelle prevedenti e tenere solo queste che contengono le stesse informazioni più la differenza tra i due gruppi, e sopra fare solo una descrizione di come sono stati raggruppati i gruppi.

Table 6: Marital status VS is Graduated

	No	Yes	Sum
No	1904	2015	3919
Yes	311	194	505

Table 7: Previous qualification VS is Graduated

	No	Yes	Sum
No Secondary	169	63	232
Secondary	1922	2066	3988
Higher	124	80	204

Table 8: Course VS is Graduated

	No	Yes	Sum
Stem	822	900	1722
No Stem	1393	1309	2702

Table 9: Nationality VS is Graduated

	No	Yes	Sum
Portoguese	2159	2155	4314
Others	56	54	110

Table 10: Mother's qualification VS is Graduated

	No	Yes	Sum
No Secondary	158	76	234
Secondary	1738	1861	3599
Higher	319	272	591

Table 11: Father's qualification VS is Graduated

	No	Yes	Sum
No Secondary	1502	1574	3076
Secondary	476	457	933
Higher	237	178	415

Table 12: Mother's occupation VS is Graduated

	No	Yes	Sum
Blue Collar	961	1043	2004
White Collar	1088	1101	2189
Others	166	65	231

Table 13: Father's occupation VS is Graduated

	No	Yes	Sum
Blue Collar	1220	1347	2567
White Collar	849	796	1645
Others	146	66	212

Table 14: Displaced VS is Graduated

	No	Yes	Sum
0	1113	885	1998
1	1102	1324	2426

Table 15: Education special VS is Graduated

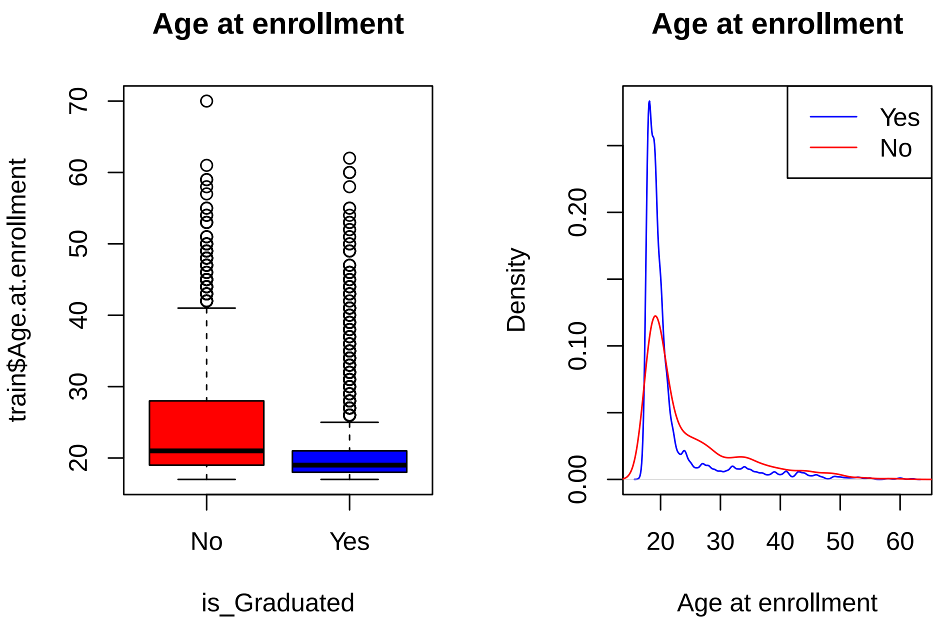
	No	Yes	Sum
0	2187	2186	4373
1	28	23	51

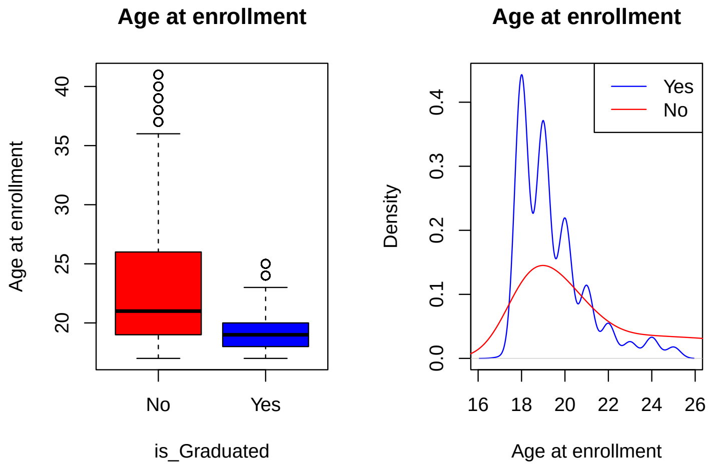
Table 16: Scholarship VS is Graduated

	No	Yes	Sum
0	1951	1374	3325
1	264	835	1099

OUTLIERS DETECTION AND REMOVAL

We looked for outliers among numerical features in our data, 'age at enrollment' was the only one that presented some. Since it have not a normal distribution, we decided to remove all values which are outliers in the boxplot of values condictioned by the graduateds. We then perform a train-test split before and remove the outliers from the training set only: keeping outliers in the test set improves the ecological validity of the model and reduces overfitting.





4. Model building

We will now try to build different models using the training set,the models we will use include logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), ridge regression, and lasso regression.

Logistic Regression

One of the first model we employ is logistic regression, this seems like an obvious choice given the fact that our research question revolves around finding a probability in a binary classification task.

The core idea behind logistic regression is to model the relationship between one or more independent variables (features) and a binary dependent variable (outcome) using the logistic function. This function maps the graduated of a linear combination of the features to a probability score between 0 and 1.

In logistic regression, the coefficients associated with each feature are estimated using maximum likelihood estimation. These coefficients represent the impact of each feature on the log-odds of the outcome variable.

Logistic Regression full

```
##
## Call:
## glm(formula = is_Graduated ~ ., family = binomial, data = train)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                      5.17175    0.70040   7.384 1.54e-13 ***
## MarriedYes                      -0.33058    0.33188  -0.996   0.3192    
## CourseNo Stem                   -0.36087    0.08898  -4.056 5.00e-05 ***
## Previous.qualificationSecondary  0.05152    0.38877   0.133   0.8946    
## Previous.qualificationHigher     0.13741    0.49686   0.277   0.7821    
## NationalityOthers                0.16523    0.27734   0.596   0.5513    
## Mother.s.qualificationSecondary  0.41600    0.24220   1.718   0.0859 .  
## Mother.s.qualificationHigher     0.37397    0.26783   1.396   0.1626    
## Father.s.qualificationSecondary -0.12368    0.10824  -1.143   0.2532    
## Father.s.qualificationHigher    -0.17350    0.16876  -1.028   0.3039    
## Mother.s.occupationWhite Collar  0.06535    0.10057   0.650   0.5158    
## Mother.s.occupationOthers       -0.26858    0.31915  -0.842   0.4000    
## Father.s.occupationWhite Collar -0.03505    0.10084  -0.348   0.7281    
## Father.s.occupationOthers        0.11142    0.31145   0.358   0.7205    
## Displaced                        0.01253    0.09236   0.136   0.8921    
## Educational.special.needs       -0.29952    0.38691  -0.774   0.4389    
## GenderMale                      -0.64493    0.09248  -6.974 3.09e-12 ***
## Scholarship.holder               1.27219    0.10682  11.909  < 2e-16 ***
## Age.at.enrollment               -0.26790    0.02030 -13.200  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4125.2  on 2978  degrees of freedom
## Residual deviance: 3209.5  on 2960  degrees of freedom
## AIC: 3247.5
##
## Number of Fisher Scoring iterations: 5
```

Logistic Regression final

```
##
## Call:
## glm(formula = is_Graduated ~ Course + Mother.s.qualification +
##     Gender + Scholarship.holder + Age.at.enrollment, family = binomial,
##     data = train)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                      6.73903    0.52050  12.947  < 2e-16 ***
## CourseNo Stem                   -0.50218    0.09805  -5.121 3.03e-07 ***
## Mother.s.qualificationSecondary  0.95651    0.24848   3.849 0.000118 ***
## Mother.s.qualificationHigher     0.84373    0.26959   3.130 0.001750 ** 
## GenderMale                      -0.94168    0.10034  -9.385  < 2e-16 ***
## Scholarship.holder               2.18108    0.13580  16.061  < 2e-16 ***
## Age.at.enrollment               -0.36608    0.02284 -16.030  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3961.4  on 2858  degrees of freedom
## Residual deviance: 2694.4  on 2852  degrees of freedom
## AIC: 2708.4
##
## Number of Fisher Scoring iterations: 6
```

Logistic Regression with interaction

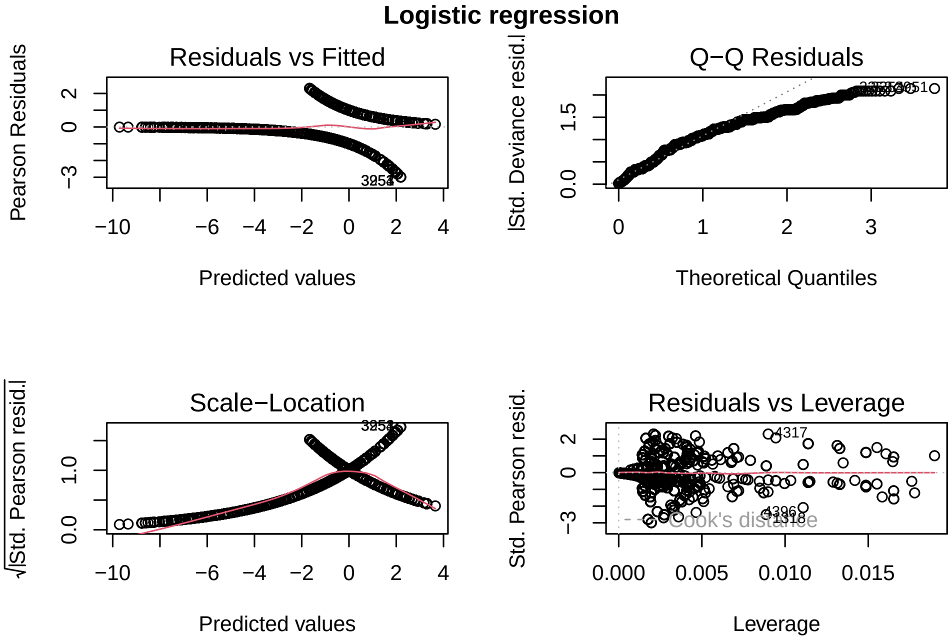
```
##
## Call:
## glm(formula = is_Graduated ~ Course + Mother.s.qualification +
##     Gender + Scholarship.holder + Age.at.enrollment + Course:Gender +
##     Mother.s.qualification:Scholarship.holder + Gender:Scholarship.holder +
##     Scholarship.holder:Age.at.enrollment, family = binomial,
##     data = train)
##
## Coefficients:
##                                            Estimate Std. Error z value
## (Intercept)                                 6.71016    0.55878  12.009
## CourseNo Stem                              -0.71582    0.12302  -5.819
## Mother.s.qualificationSecondary             0.70977    0.26283   2.700
## Mother.s.qualificationHigher                0.59889    0.28210   2.123
## GenderMale                                 -1.20512    0.16337  -7.377
## Scholarship.holder                          3.63787    1.42443   2.554
## Age.at.enrollment                          -0.34760    0.02439 -14.252
## CourseNo Stem:GenderMale                    0.62710    0.20495   3.060
## Mother.s.qualificationSecondary:Scholarship.holder  1.35887    0.63590   2.137
## Mother.s.qualificationHigher:Scholarship.holder     1.66945    0.91797   1.819
## GenderMale:Scholarship.holder              -0.85532    0.29155  -2.934
## Scholarship.holder:Age.at.enrollment       -0.11990    0.06515  -1.840
##                                            Pr(>|z|)
## (Intercept)                                 < 2e-16 ***
## CourseNo Stem                              5.92e-09 ***
## Mother.s.qualificationSecondary             0.00692 **
## Mother.s.qualificationHigher                0.03376 *
## GenderMale                                 1.62e-13 ***
## Scholarship.holder                          0.01065 *
## Age.at.enrollment                           < 2e-16 ***
## CourseNo Stem:GenderMale                    0.00221 **
## Mother.s.qualificationSecondary:Scholarship.holder  0.03260 *
## Mother.s.qualificationHigher:Scholarship.holder     0.06897 .
## GenderMale:Scholarship.holder               0.00335 **
## Scholarship.holder:Age.at.enrollment        0.06573 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3961.4  on 2858  degrees of freedom
## Residual deviance: 2670.8  on 2847  degrees of freedom
```

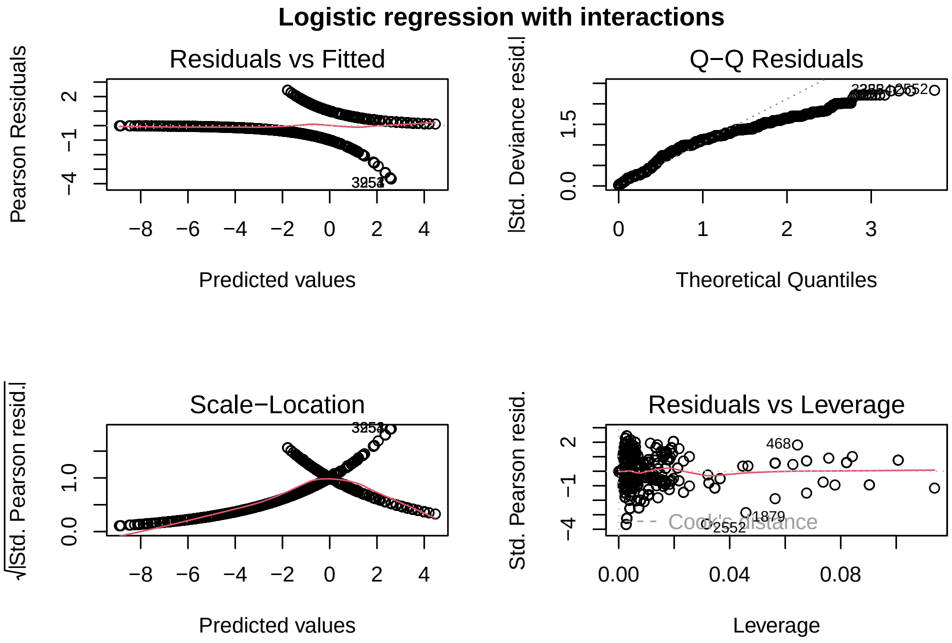
Table 17: Logistic regression coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.739	0.520	12.947	0.000
CourseNo Stem	-0.502	0.098	-5.121	0.000
Mother.s.qualificationSecondary	0.957	0.248	3.849	0.000
Mother.s.qualificationHigher	0.844	0.270	3.130	0.002
GenderMale	-0.942	0.100	-9.385	0.000
Scholarship.holder	2.181	0.136	16.061	0.000
Age.at.enrollment	-0.366	0.023	-16.030	0.000

```
## AIC: 2694.8
##
## Number of Fisher Scoring iterations: 6
```

In the Q-Q residuals plot, it is evident that the residuals are not normally distributed, suggesting that the variables included in the model are insufficient to explain the data variability. Additionally, the residuals vs leverage plot reveals the presence of points with very high leverage, indicating that there are still outliers in the dataset.

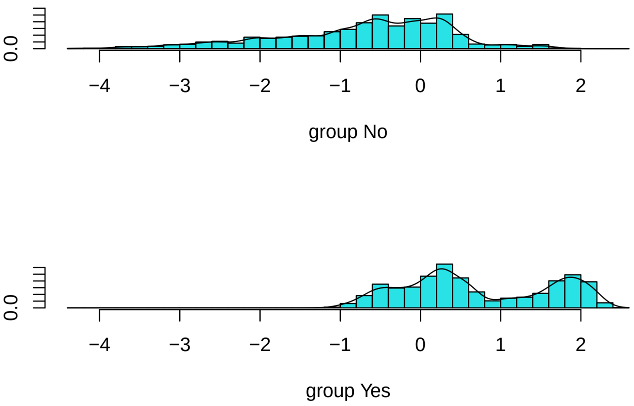




Linear Discriminant Analysis (LDA)

LDA is a classic statistical technique utilized in machine learning and pattern recognition for classification tasks. We employ it to identify which linear combination of features best separates multiple classes or categories in our dataset.

At its core, LDA operates under the assumption that the data can be represented as multivariate Gaussian distributions and that the classes share the same covariance matrix. It seeks to find a projection of the data onto a lower-dimensional space while maximizing the separation between classes and minimizing the variance within each class.



Quadratic Discriminant Analysis (QDA)

QDA is an extension of Linear Discriminant Analysis (LDA) used for classification tasks. While LDA assumes that different classes share the same covariance matrix, QDA relaxes this assumption, allowing each class to have its own covariance matrix.

Table 18: LDA means

	No	Yes
MarriedYes	0.113	0.011
CourseNo Stem	0.626	0.572
Previous.qualificationSecondary	0.879	0.986
Previous.qualificationHigher	0.057	0.009
NationalityOthers	0.027	0.021
Mother.s.qualificationSecondary	0.772	0.845
Mother.s.qualificationHigher	0.157	0.131
Father.s.qualificationSecondary	0.224	0.228
Father.s.qualificationHigher	0.114	0.078
Mother.s.occupationWhite Collar	0.508	0.542
Mother.s.occupationOthers	0.072	0.019
Father.s.occupationWhite Collar	0.399	0.373
Father.s.occupationOthers	0.062	0.018
Displaced	0.515	0.668
Educational.special.needs	0.012	0.011
GenderMale	0.473	0.222
Scholarship.holder	0.068	0.416
Age.at.enrollment	23.590	19.308

```
## Warning in styling_latex_position_right(x, table_info, hold_position,
## table.envir): Position = right is only supported for longtable in LaTeX.
## Setting back to center...
```

Table 19: QDA coefficients

	No	Yes
MarriedYes	0.113	0.011
CourseNo Stem	0.626	0.572
Previous.qualificationSecondary	0.879	0.986
Previous.qualificationHigher	0.057	0.009
NationalityOthers	0.027	0.021
Mother.s.qualificationSecondary	0.772	0.845
Mother.s.qualificationHigher	0.157	0.131
Father.s.qualificationSecondary	0.224	0.228
Father.s.qualificationHigher	0.114	0.078
Mother.s.occupationWhite Collar	0.508	0.542
Mother.s.occupationOthers	0.072	0.019
Father.s.occupationWhite Collar	0.399	0.373
Father.s.occupationOthers	0.062	0.018
Displaced	0.515	0.668
Educational.special.needs	0.012	0.011
GenderMale	0.473	0.222
Scholarship.holder	0.068	0.416
Age.at.enrollment	23.590	19.308

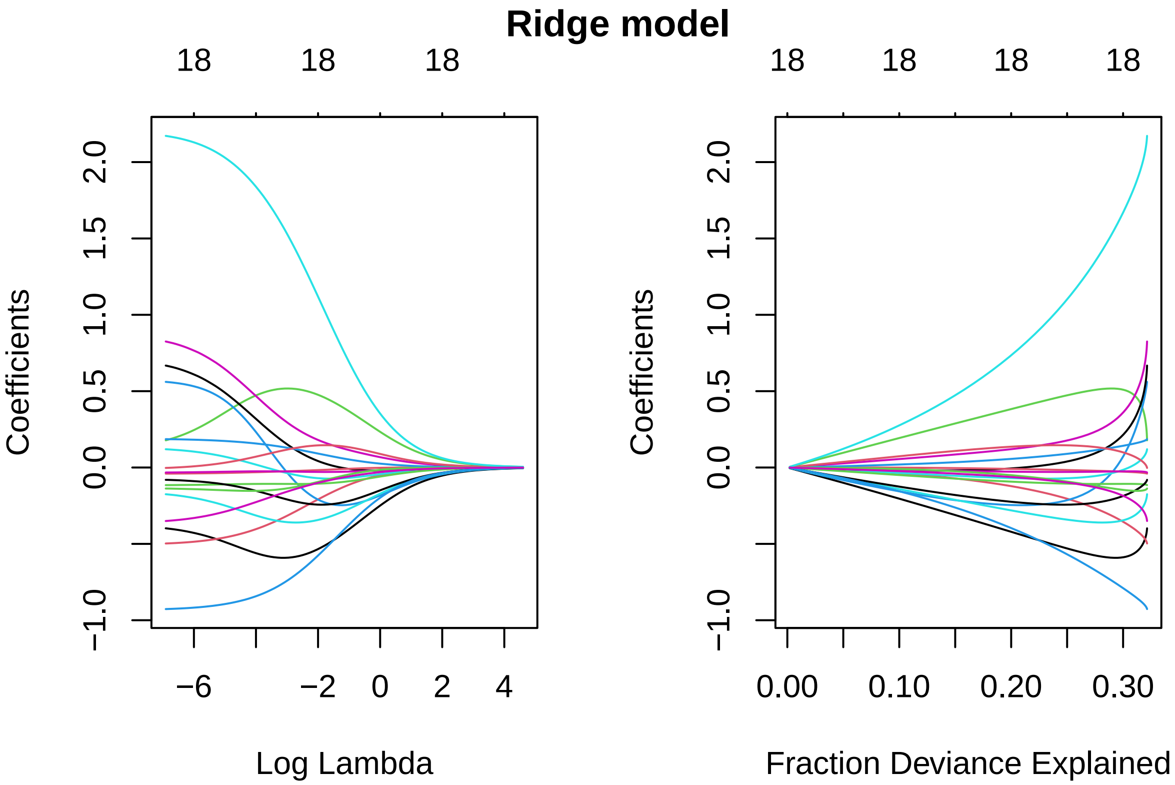
Ridge Regression

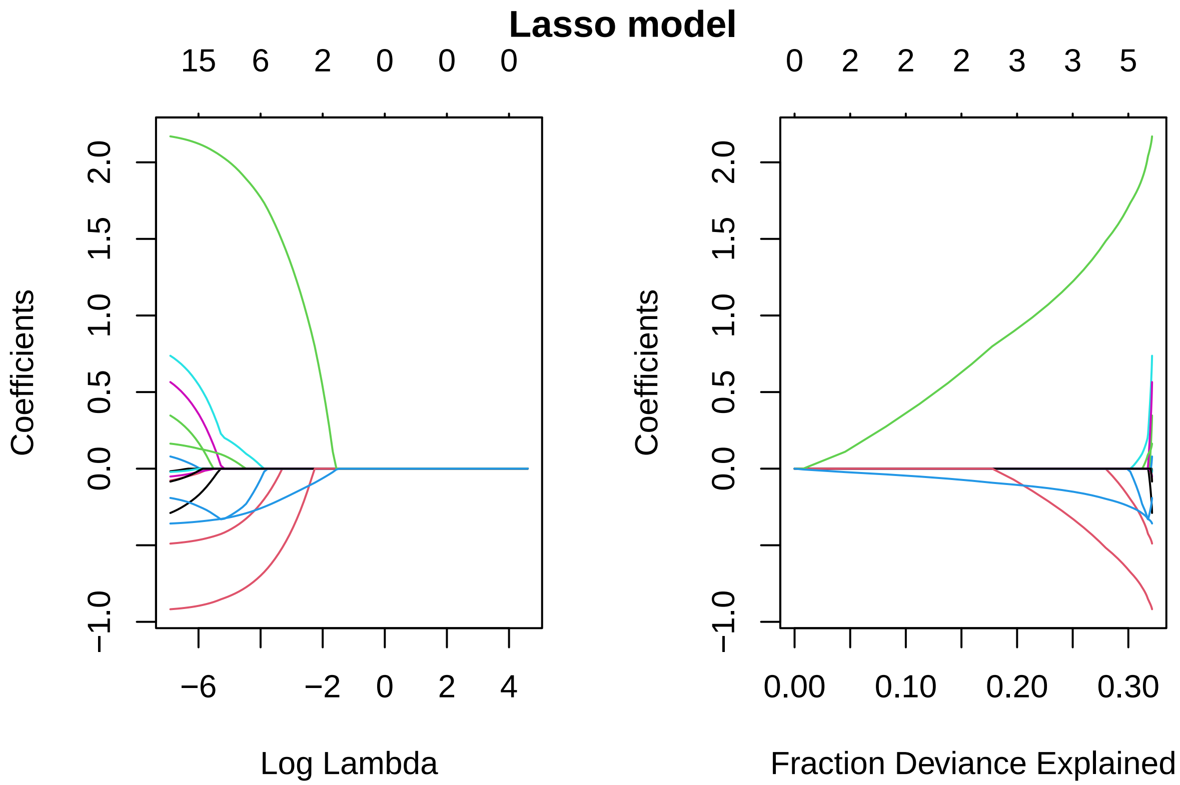
Ridge regression is a regularization technique used in linear regression to mitigate the issues of multicollinearity and overfitting: it can be used to work with high dimensional data, which is why we find it valuable in our case

Lasso Regression

This regression techniques takes a slightly different approach by adding a penalty term that penalizes the absolute values of the regression coefficients, instead of their squares (which is what Rigde regression does).

This penalty term encourages sparsity in the coefficient vector, effectively driving some coefficients to exactly zero. As a result, Lasso regression not only helps in shrinking coefficient values but also performs variable selection by automatically excluding irrelevant features from the model.





5. Model Evaluation

we are going to evaluate a our models by plotting ROC curve and by computing some standard evaluation metrics: TALK ABOUT ROC ECC ECC HERE

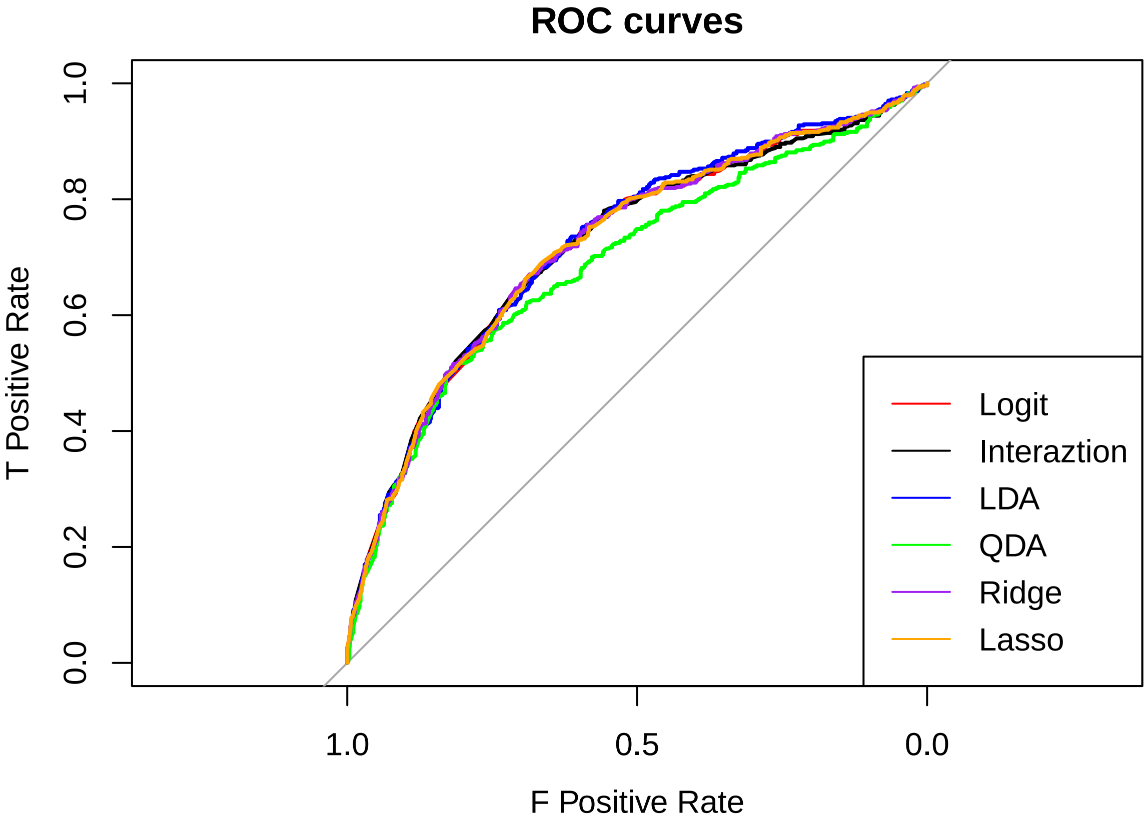


Table 20: Model evaluation

Model	AUC	Accuracy	Precision	Recall	F1
Logit	0.719	0.678	0.669	0.667	0.668
Interaction	0.718	0.675	0.675	0.641	0.657
LDA	0.722	0.671	0.662	0.659	0.660
QDA	0.690	0.614	0.577	0.771	0.660
Ridge	0.719	0.678	0.668	0.670	0.669
Lasso	0.719	0.678	0.671	0.661	0.666

APPENDIX A

Table 21: Dataset Description

Variable	Description
Marital status	Categorical variable indicating the marital status of the individual
Application mode	Categorical variable indicating the mode of application
Application order	Numeric variable indicating the order of application
Course	Categorical variable indicating the chosen course
evening attendance	Binary variable indicating whether the individual attends classes during the daytime or evening
Displaced	Binary variable indicating whether the individual has been displaced
Educational special needs	Binary variable indicating whether the individual has educational special needs

Variable	Description
Tuition fees up to date	Binary variable indicating whether the tuition fees are up to date
Gender	Binary variable indicating the gender of the individual
Scholarship holder	Binary variable indicating whether the individual holds a scholarship
Age at enrollment	Numeric variable indicating the age of the individual at the time of enrollment
International	Binary variable indicating whether the individual is international
Curricular units 1st sem (credited)	Numeric variable indicating the number of credited curricular units in the 1st semester
Curricular units 1st sem (enrolled)	Numeric variable indicating the number of enrolled curricular units in the 1st semester
Curricular units 1st sem (evaluations)	Numeric variable indicating the number of evaluations for curricular units in the 1st semester
Curricular units 1st sem (approved)	Numeric variable indicating the number of approved curricular units in the 1st semester
Curricular units 1st sem (grade)	Numeric variable indicating the average grade for curricular units in the 1st semester
Curricular units 1st sem (without evaluations)	Numeric variable indicating the number of curricular units in the 1st semester without evaluations
Curricular units 2nd sem (credited)	Numeric variable indicating the number of credited curricular units in the 2nd semester
Curricular units 2nd sem (enrolled)	Numeric variable indicating the number of enrolled curricular units in the 2nd semester
Curricular units 2nd sem (evaluations)	Numeric variable indicating the number of evaluations for curricular units in the 2nd semester
Curricular units 2nd sem (approved)	Numeric variable indicating the number of approved curricular units in the 2nd semester
Curricular units 2nd sem (grade)	Numeric variable indicating the average grade for curricular units in the 2nd semester
Curricular units 2nd sem (without evaluations)	Numeric variable indicating the number of curricular units in the 2nd semester without evaluations
Unemployment rate	Variable indicating the unemployment rate (Unemployment rate (%))
Inflation rate	Numeric variable indicating the inflation rate (Inflation rate (%))
GDP	Numeric variable indicating the Gross Domestic Product
Output	Categorical variable indicating the target variable (e.g., Dropout, Graduate, Enrolled)
Previous qualification	Numeric variable indicating the level of the previous qualification
Nationality	Categorical variable indicating the nationality of the individual
Mother's qualification	Numeric variable indicating the level of the mother's qualification
Father's qualification	Numeric variable indicating the level of the father's qualification
Mother's occupation	Categorical variable indicating the mother's occupation
Father's occupation	Categorical variable indicating the father's occupation

“‘