Statistical learning final project

Di Pasquale Giuseppe, Gorni Selvestrini Matteo

2024-07-11

Contents

1 Introduction

The data was originally 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 and evaluating how much it may effect the scholarship award.

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

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 rename the 'Output' variable into 'is_Graduated' and we make it binary, 'Yes' labels will correspond to the old 'Graduate' labels, while all the 'Enrolled' and 'Dropout' labels will be converted into 'No'.

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

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

# Rename '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 and data balence check:

There were no missing values and the dataset is balanced.

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

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

```r
table(data$is_Graduated)
```

```
##
##   No  Yes
## 2215 2209
```

2.1 Relabelling variables

We changed the labeling criteria in 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 with 6 values indicating the marital status of the individual. Given that the majority of the vast majority of students is single and only a few of them belong to one of the other categories we decided to create a binary variable indicating if a student is single or not.

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: Parental occupation

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

Mother/Father/Student qualification:

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

Course:

Categorical variable representing the course chosen at enrollment, we originally had 17 different courses, we decided to relabel the courses based on whether they belong to the STEM area or not.

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

2.2 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.

- 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
- 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.
- 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.
- Tuition fees up do date.

We also removed the 'International' variable since it was rendundant.

```
table(data$Nationality, data$International)
```

```
##
##                  0    1
##   Portoguese 4314    0
##   Others        0  110
```

```
data <- data[, -which(colnames(data) == 'International')]
```

3 Exploratory Data Analysis

Most of our dataset's features are categorical, therefore we will plot the data in the form of contingency tables. This procedure allows us to make some interesting observations:

- Most people who enroll are single and at the same time are more likely to graduate than others (51.4% vs 38.4%); (table 6)

- People who enroll after completing secondary school tend to graduate in time more often than those who do not (note: Portalegre's university allows some students to attend courses without having an high-school diploma); (table 7)
- 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%); (table 8)
- Mothers generally have a higher level of education compared to fathers. Specifically, 3076 fathers do not have an high school diploma, the same can be said only for 234 mothers, this does not seem to affect graduation rates of their children; (table 10)
- Students who left their parents home to study (i.e. the 'displaced' variable) have a higher graduation rate compared to those who did not (54.6% vs 44.3%); (table 14)
- Students who receive a scholarship have a higher graduation rate than those who do not (76.0% vs 41.3%); (table 16)
- Women seem graduate on time much more than men do (57.9% vs 35.2%). (table 17)
- From the figures below we can see that 'Age at enrollment' have not a normal distribution, we can see that the distribution of the age at enrollment is different for students who graduate and those who do not. Both have an high right tail with more outliers, but the distribution of graduate on time is more concentrated around the mean value (20 years).

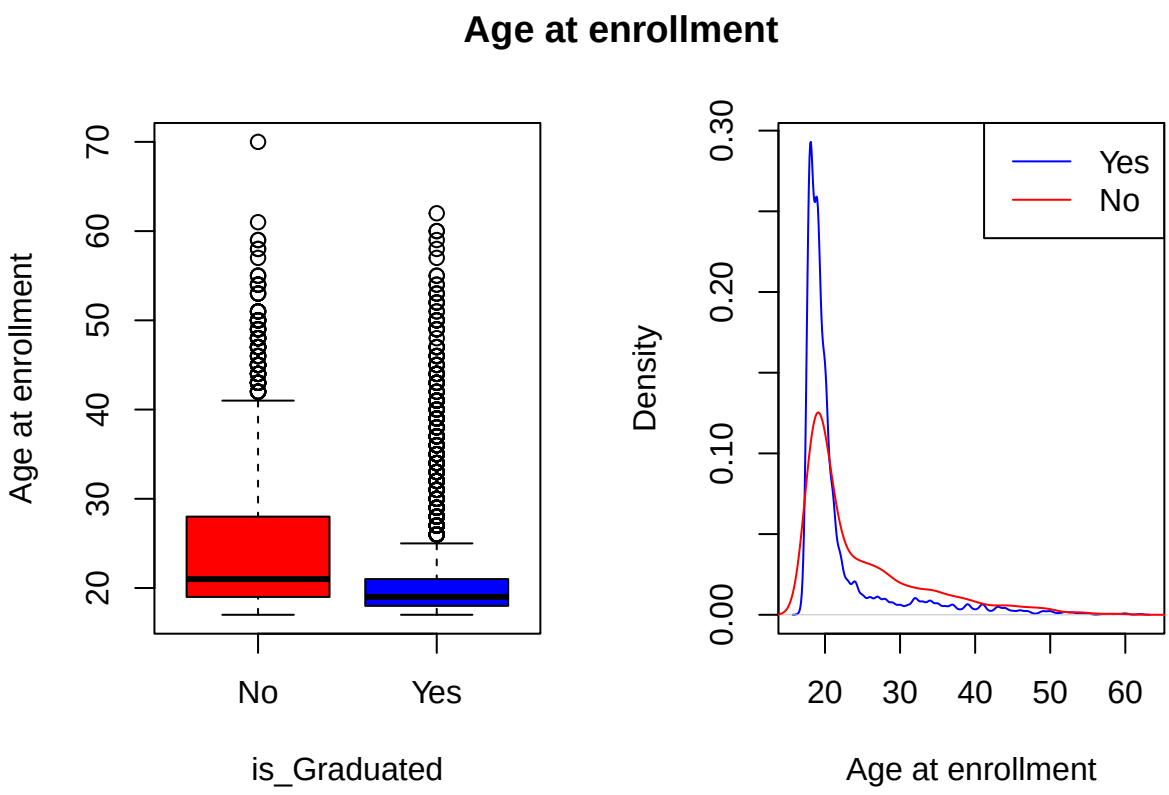


Table 6: Marital status VS is Graduated

	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

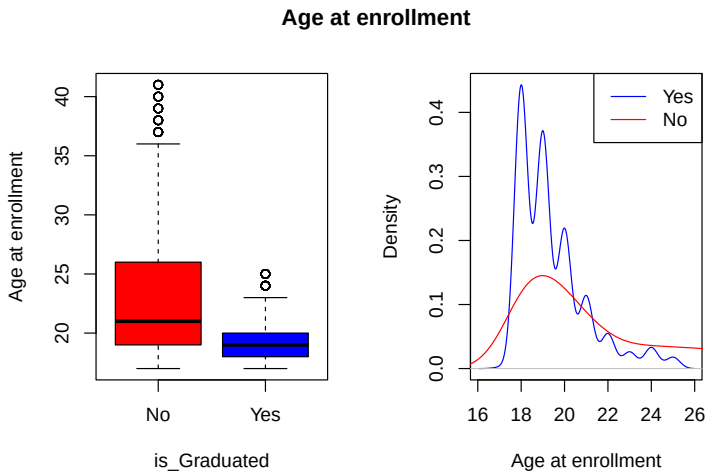
Table 17: Gender VS is Graduated

	No	Yes	Sum
Female	1207	1661	2868
Male	1008	548	1556

3.1 Outliers detection and removal

As we have seen in the previous section, the distribution of the age at enrollment contains a lot of outliers. We will now remove these outliers from the training set only, in order to preserve test set to simulate a real-world scenario. This will allow us to build a more robust model that generalizes better to unseen data.

We can see that there are more outliers left, but we decided to keep them to stay more faithful to the original data distribution.



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. We will then evaluate the performance of these models using the test set working with the confusion matrix and the ROC curve and finally we compare the relative coefficients.

4.1 Logistic Regression

We start our analysis by considering the *Logistic Regression model*, this seems like a very robust 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 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.

```
fit_logit_1 <- glm(is_Graduated ~ .,
            family = binomial,
            data = train)
summary(fit_logit_1)

# Keep only significant variables
fit_logit_2 <- step(fit_logit_1)
summary(fit_logit_2)

# Removing outliers
index_outliers <- which(abs(scale(fit_logit_2$residuals)) > 2)
train <- train[-index_outliers, ]
```

```
# Final model
fit_logit_3 <- update(fit_logit_2,
               .~.)
summary(fit_logit_3)
```

```
##
```

```
## 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
```

One very important feature of logistic regression is that it is possible to look is certain features have a significant influence over the outcome variable. Looking at the model we just built we can see that many of the predictions we made during exploratory data analysis hold true:

- being in a non-STEM course is associated with a decrease of 0.50 in the log-odds of graduating, compared to being in a STEM course (reference group);
- having a mother with secondary education is associated with an increase of 0.96 in the log-odds of graduating compared to having a mother with no education (reference group);
- having a mother with higher education is associated with an increase of 0.84 in the log-odds of graduating compared to having a mother with no education (reference group);
- being male is associated with a decrease of 0.94 in the log-odds of graduating compared to being female (reference group);
- being a scholarship holder is associated with a decrease of 2.18 in the log-odds of graduating compared to not being a scholarship holder;
- each additional year of age at enrollment is associated with a decrease of 0.37 in the log-odds of graduating.

4.1.1 Logistic Regression with interactions

In logistic regression, interactions occur when the effect of one predictor variable on the outcome depends on the value of another independent variable. We will now try to build a logistic regression model with interactions to see if we can improve the model's performance using only significant variables find previously.

```
fit_int_1 <- update(fit_logit_3,
                    .~.*.)
summary(fit_int_1)
fit_int_2 <- step(fit_int_1)
```

```r
summary(fit_int_2)
```

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

We find that the single variables are like the simple model, but from the significant interactions we have that:

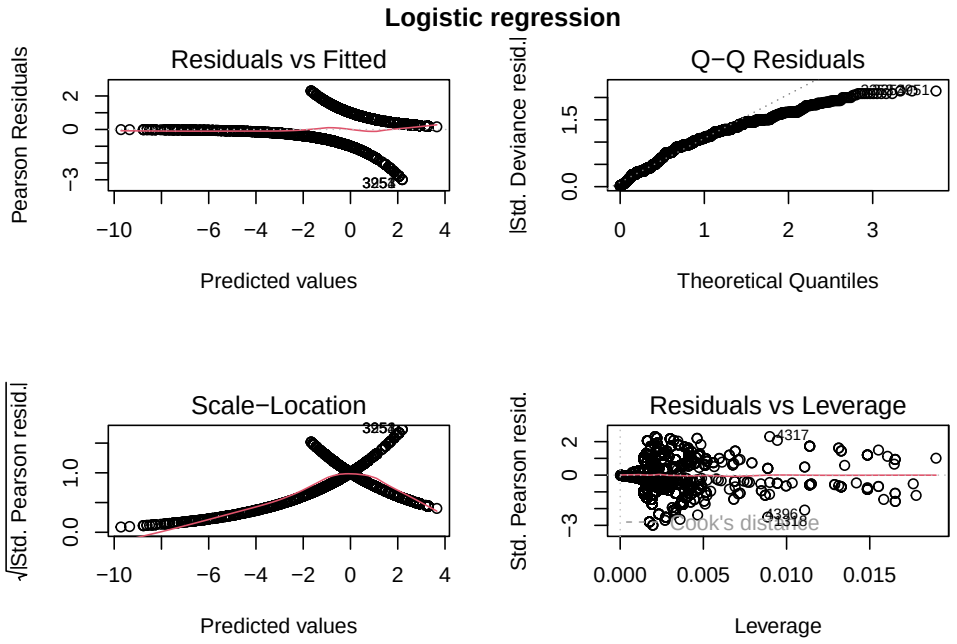
- being male in a non-STEM course is associated with an increase of 0.63 in the log-odds of graduating;
- having a scholarship and a mother with secondary or higher education is associated with an increase of, respectively, 1.36 and 1.67 in the log-odds of graduating;
- being male with a scholarship decrease the log-odds of graduating by 0.86;

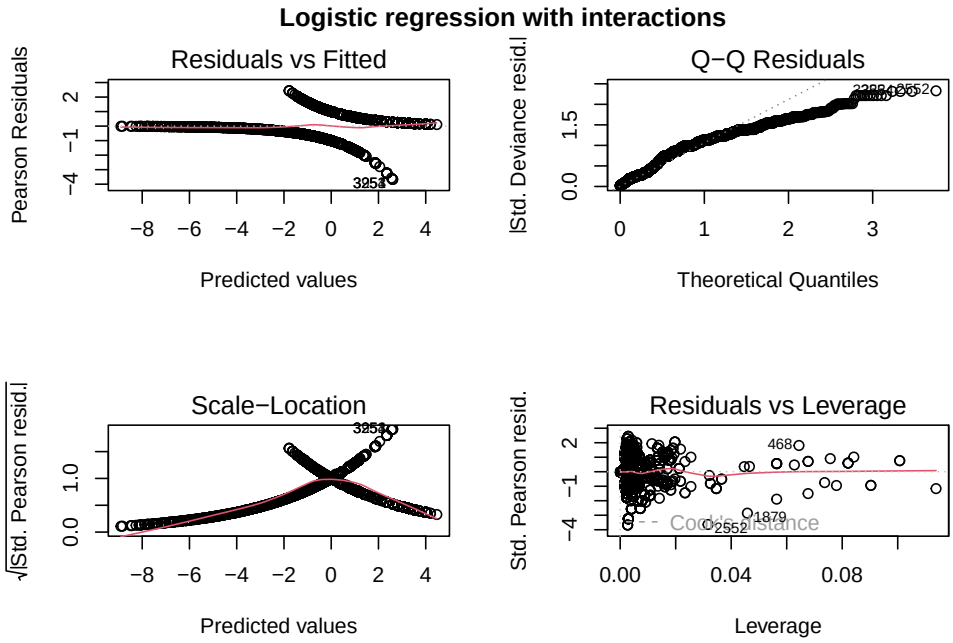
- being a scholarship holder decreases log-odds of graduating by 0.12 for each year of age at enrollment time. Recalling that the coefficient for scholarship holder is 3.64 it means that if you enroll at 20 years old and you are a scholarship holder, so the log-odds of graduating are 3.64 - 0.12 * 20 = 1.24 compared to those who did not get the scholarship.

We then plot residuals for both models in order to gain even more insights about the data we are working with: by looking Q-Q residuals plot, it is evident that the residuals are not normally distributed, suggesting that the variables included in the model are not sufficient to explain the data variability.

Furthermore, the residuals vs leverage plot reveals the presence of points with very high leverage, indicating that there are still outliers in the dataset.

The plots for the logistic regression with interactions are more similar to the ones of the logistic regression, suggesting that even if the variables are significant they do not particularly improve the model's performance. As we can see from the ANCOVA test, it is significant, but not improve particularly the residual deviance (from 2694 to 2670).






```
anova(fit_logit_3, fit_int_2)
```

```
## Analysis of Deviance Table
##
## Model 1: is_Graduated ~ Course + Mother.s.qualification + Gender + Scholarship.holder +
##     Age.at.enrollment
## Model 2: 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
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      2852     2694.3
## 2      2847     2670.8  5   23.506 0.0002701 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we have assumed previously, from the two confusion matrices below (table 18 and table 19), we can see that there are no much difference between the two models, the only little difference is that the model with interactions predict just a little more zero's than the simple model.

Table 18: Test vs Logistic predictions

	0	1
No	392	177
Yes	179	358

Table 19: Test vs Logistic with interactions predictions

	0	1
No	403	166
Yes	193	344

4.2 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.

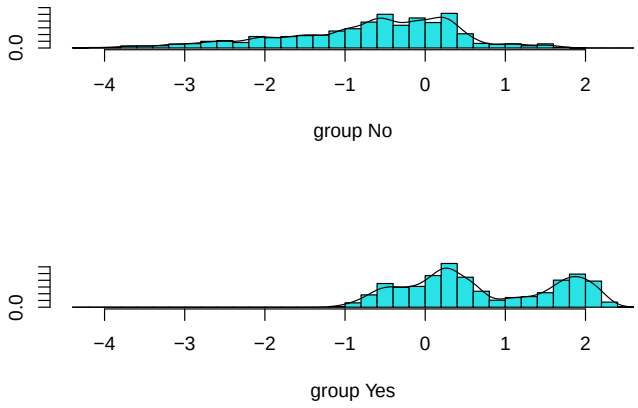
At its core, LDA operates under the assumption that the data can be represented as multivariate Gaussian distributions and that the classes share the same covariance matrix. It seeks to find a projection of the data onto a lower-dimensional space while maximizing the separation between classes and minimizing the variance within each class. Even if the assumptions of LDA are not met, it is still possible to apply it when the variables are discrete. In such cases, LDA can still provide good separation between classes, especially if the differences between classes are sufficiently pronounced. Although the optimal performance of LDA is achieved when its assumptions are satisfied, it can still be a useful tool for exploring the data and gaining preliminary insights into the structure of the classes, even when the assumptions are violated.

From now on we will use only confusion matrix on test set to evaluate the models, we will comparate all the models at the end.

DA VEDERE SE CAMBIARE ADESSO CI CONCENTRIAMO SOLO SUI RISULTATI PIÙ IMPORTANTI, QUALI LE CONFUSION MATRIX E LE ALTRE COSE... I COEFFICIENTI DELLE VARIABILI VERRANNO CONFRONTATI ALLA FINE CON GLI ALTRI MODELLI

```
library(MASS)

fit_lda <- lda(is_Graduated ~ .,
               data = train)
```

We can observe the density of the LDA results respect to the "is_Graduated" variable on train set in the plot below. The coefficients represent the weights of each feature in the linear combination that best separates the classes, suggesting that the features included in the model are useful for predicting the outcome variable, although it is not linearly separable.



4.3 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.

```
fit_qda <- qda(is_Graduated ~ .,
               data = train)
```

Differently from LDA, QDA does not plot the density of the results, but we can still see the probability of the train set to be graduated on time in the tables below. We can see that QDA is really able to predict correctly who graduated on time, but it have not a good performance for predicting who did not graduate on time (89% vs 48%).

Table 20: Distribution of the results respect to NOT graduating on time in the train set

0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.48	0.05	0.01	0.01	0.01	0	0.01	0.01	0.02	0	0.41

Table 21: Distribution of the results respect to graduating on time in the train set

0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.06	0	0	0	0.01	0	0	0	0.01	0.02	0.89

We see that the LDA model is more balanced in predicting who graduated on time and who did not, unlike the QDA model which is very prone to graduate on time (table 22 and table 23).

Table 22: Test vs LDA predictions

	0	1
No	388	181
Yes	183	354

Table 23: Test vs QDA predictions

	0	1
No	265	304
Yes	123	414

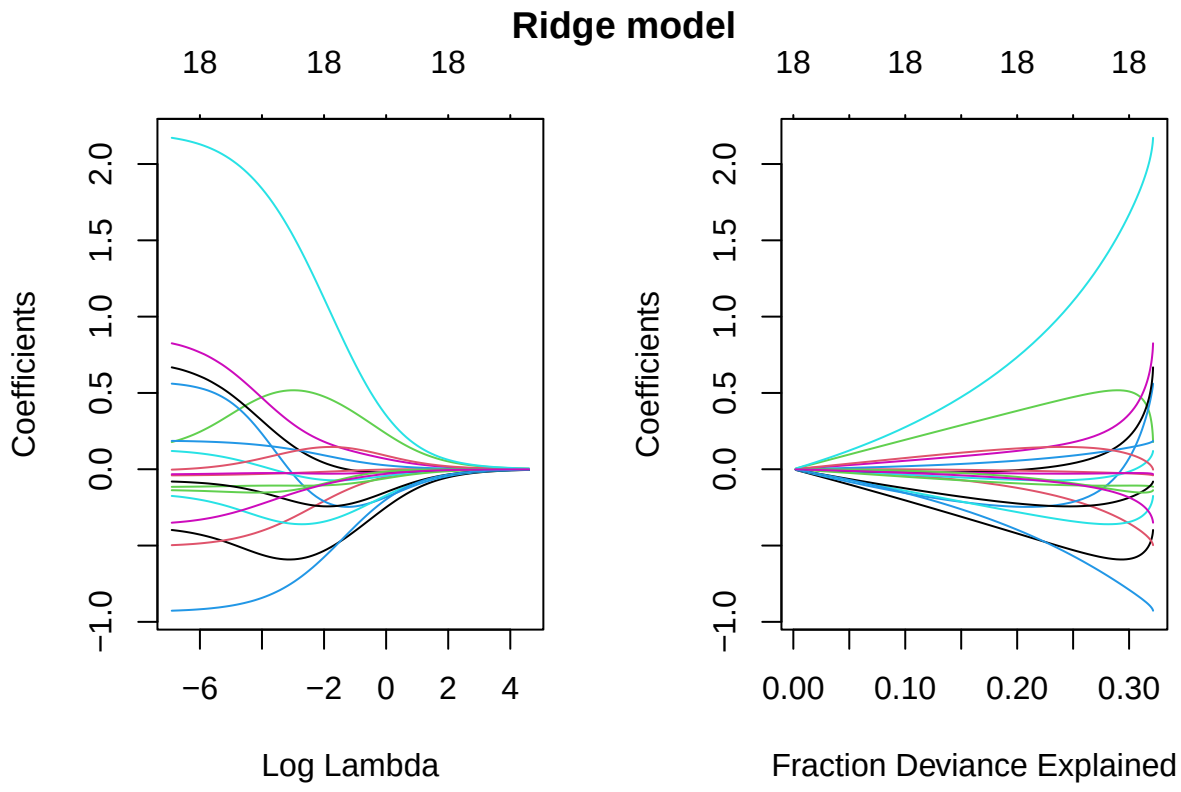
4.4 Ridge Regression

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.

```
library(glmnet)

design_matrix_train <- model.matrix(is_Graduated ~ .,
                                    data = train)[, -1]

design_matrix_test <- model.matrix(is_Graduated ~ .,
                                   data = test)[, -1]

# Creating a grid for lamda
grid <- 10^seq(2, -3, length = 100)

# Ridge Regression with Binomial distribution
ridge_model <- glmnet(x = design_matrix_train,
                      y = train$is_Graduated,
                      alpha = 0,
                      lambda = grid,
                      intercept = T,
                      family = "binomial")

# Find the best lambda
predictions <- predict(ridge_model,
```

```
                        newx = design_matrix_test,
                        type = 'response')

accuracies <- c()

for (i in 1:ncol(predictions)) {
  predicted_classes <- round(predictions[, i])
  confusion_matrix <- table(predicted_classes, test$is_Graduated)
  accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
  accuracies <- c(accuracies, accuracy)
}

best_lambda_index_r <- which.max(accuracies)
pred_ridge <- matrix(predictions[, best_lambda_index_r],
                     ncol = 1)
best_accuracy <- accuracies[best_lambda_index_r]

colnames(pred_ridge) <- 'pred_ridge'
```

In the figures below we can see the evolution of the coefficients of the ridge model with respect to the lambda value, left figure. We can see that the coefficients are shrinking to zero as the lambda value increases, this is a clear sign that the model is working correctly and that the regularization is working as expected. The right figure show the evolution of the deviance of the model with respect to coefficients value. We can also see that with the ridge model the coefficients can be change direction over the lambda value.



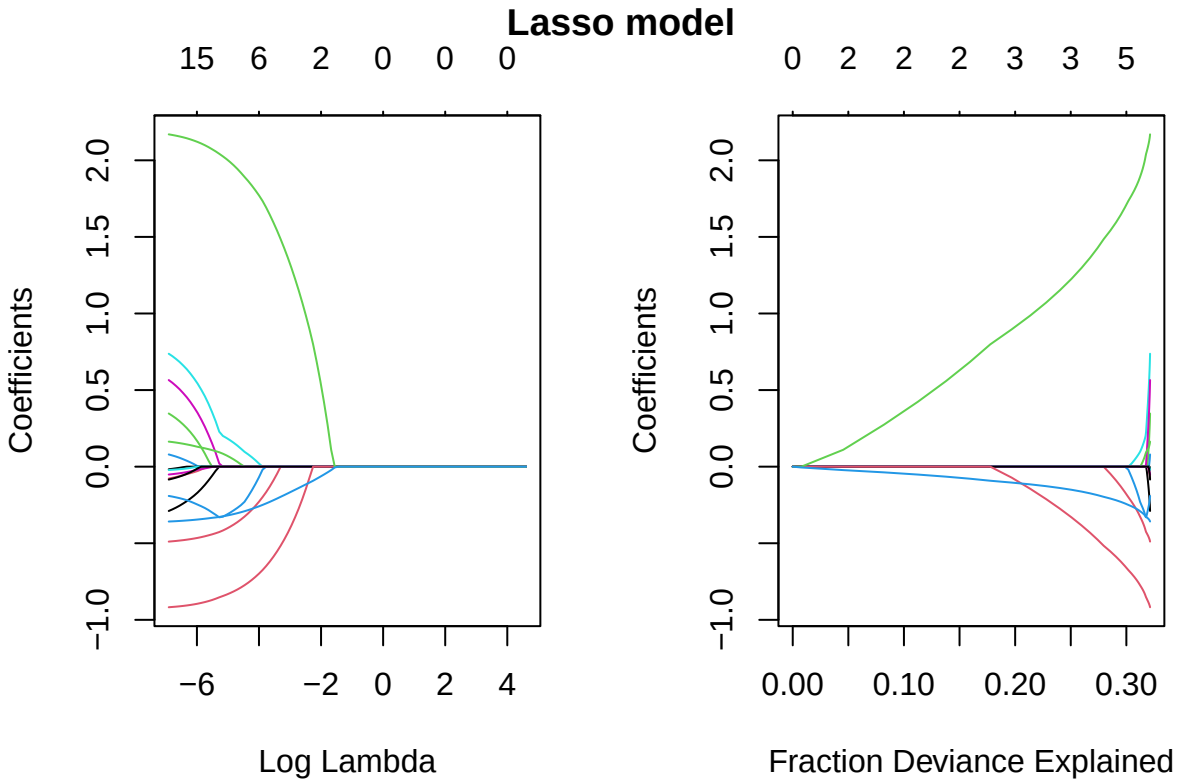
4.5 Lasso Regression

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.

```
# Lasso Regression with Binomial distribution
lasso_model <- glmnet(x = design_matrix_train,
                      y = train$is_Graduated,
                      alpha = 1,
                      lambda = grid,
                      intercept = T,
                      family = "binomial")

# Find the best lambda
predictions <- predict(lasso_model,
                       newx = design_matrix_test,
                       type = 'response')

accuracies <- c()

for (i in 1:ncol(predictions)) {
  predicted_classes <- round(predictions[, i])
  confusion_matrix <- table(predicted_classes, test$is_Graduated)
  accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
  accuracies <- c(accuracies, accuracy)
}

best_lambda_index_l <- which.max(accuracies)
pred_lasso <- matrix(predictions[, best_lambda_index_l],
                     ncol = 1)
best_accuracy <- accuracies[best_lambda_index_l]

colnames(pred_lasso) <- 'pred_lasso'
```

Like previously, we can see the evolution of the coefficients of the lasso model with respect to the lambda value, left figure. We can see that the coefficients are shrinking to zero as the lambda value increases, this is a clear sign that the model is working correctly and that the regularization is working as expected. The right figure shows the deviance of the model with respect to the lambda value, we can see that the deviance is decreasing as the lambda value increases, this is a clear sign that the model is working correctly and that the regularization is working as expected. However, differently from the ridge model, the coefficients difficultly change direction over the lambda value and they collapse to 0 more quickly than ridge model.



From the confusion matrices below (table 24 and table 25), we can see that the ridge model predictions are more similar to the lasso model predictions, likely logistic models.

Table 24: Test vs LDA predictions

	0	1
No	390	179
Yes	177	360

Table 25: Test vs QDA predictions

	0	1
No	395	174
Yes	182	355

5 Model Evaluation

We intend to evaluate our models by plotting their respective ROC curves and computing some standard evaluation metrics, such as precision, accuracy, recall and f1 score.

ROC (Receiver Operating Characteristic) curves are graphical representations used to evaluate the performance of binary classification models. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

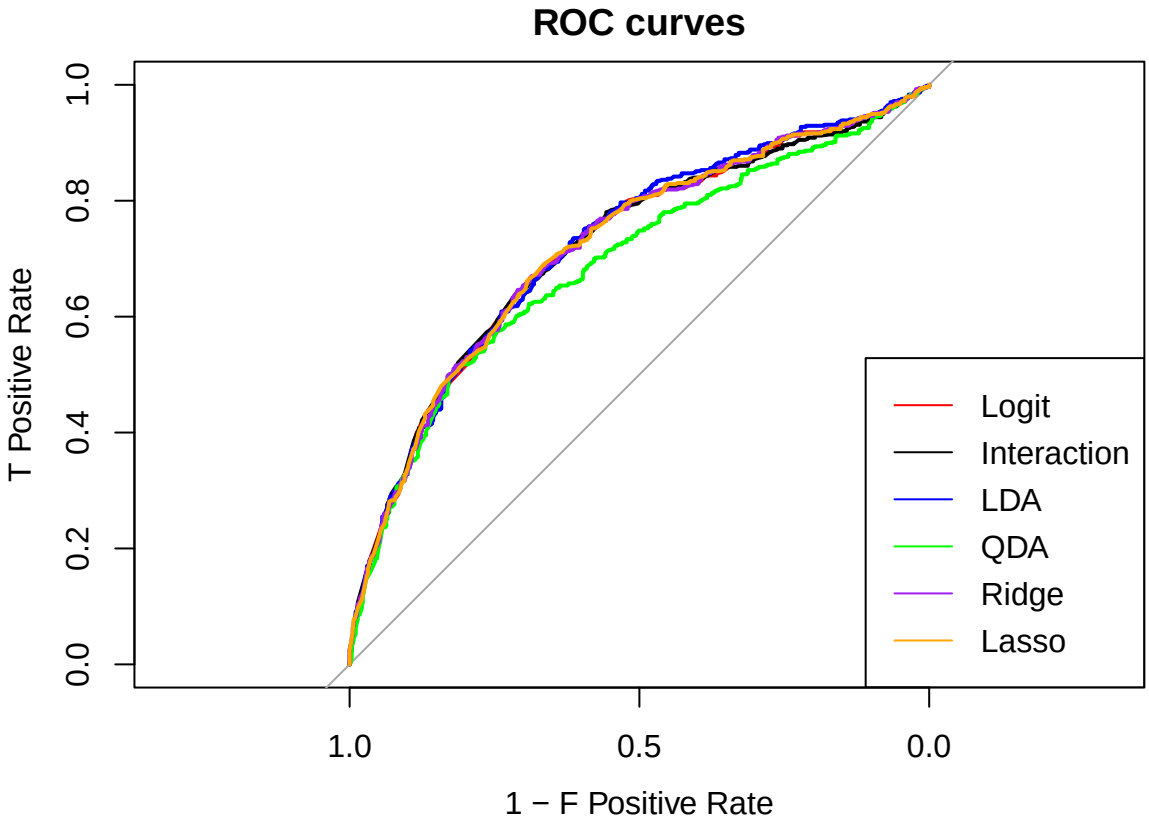


Table 26: Model evaluation

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

As we can see from the figure above and from table 26, all the models have a very similar performance, the only model that stands out is the QDA model, which has the best recall, but the worst precision since it predict more 'graduated on time' than other models. The other models have a very similar performance in terms of AUC, accuracy, precision, recall and F1 score. This tells us that there are no particular differences between the models we used and we have covered all the possible explanation from data.

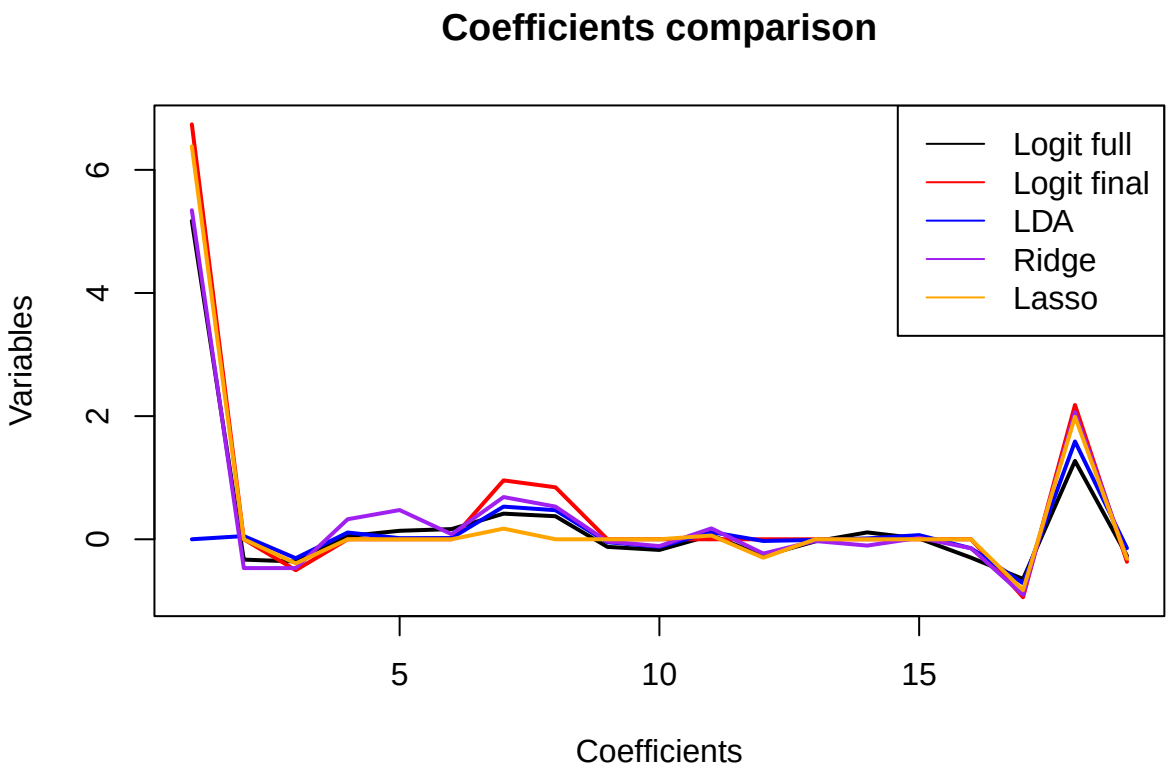
To explore this topic further we can comparate the coefficients of all models keeping out the QDA model, which has no coefficients comparable, and the Logistic with interactions model due to the fact it has more variables than the other models.

Come possiamo vedere dalla tabella precedente, a parte per la QDA i modelli risultano molto simili tra le metriche utilizzate, in compenso la QDA ha ottenuto la migliore Recall. Questo risultato porta a pensare che non ci siano particolari differenze tra i modelli usati. Date queste motivazioni, per semplicità di spiegazione si è deciso di usare il modello logistico semplice per spiegare il fenomeno. Per come è strutturato il modello non si può valutare l'incremento percentuale di ogni singola variabile, ma si può osservare di quanto è il rapporto tra le variabili. Possiamo solo valurare solo gli odds ratio. Cioè per valurare quanto influenzi il sesso basta

fare l'exp del coefficiente relativo al sesso. (exp(sex_Male) = 0.382, quindi se il sesso è maschile, la probabilità di laurearsi è 0.382 volte quella di una femmina, cioè il (100 - 38.2)% in meno rispetto ad una femmina). Mentre il rapporto tra Sposato e Single è di 1.06 quello tra Altro e Single è di 0.07, il che stà ad indicare che chi rientra nella categoria Altro ha una notevole differenza nello sposarsi in tempo. Chi sceglie materie non stem ha il 58.9% di probabilità relativa in meno rispetto a stem. Si può notare che ha una madre istruita ha più del doppio di probabilitò di laurearsi in tempo rispetto a chi ha una madre che non ha terminato le superiori, a quanto pare conoscendo il titolo di studio della madre non è influente sapere anche quello del padre. Ottimo chi decide di seguire una Scholarship, ha il 2.24 volte di probabilità di laurearsi in tempo rispetto a chi non la segue. Come ci si poteva aspettare, andando avanti con l'età la probabilità di laurearsi in tempo diminuisce, infatti per ogni anno in più di età la probabilità di laurearsi in tempo diminuisce decresce a livello esponenziale del 70% per ogni anno in più, ovviamente è solo un'approssimazione del modello

Table 27: Coefficients comparison

	Logit_full	Logit_final	Ridge	Lasso	LDA
(Intercept)	5.17	6.74	5.34	6.38	0.00
MarriedYes	-0.33	0.00	-0.47	0.00	0.05
CourseNo Stem	-0.36	-0.50	-0.47	-0.39	-0.31
Previous.qualificationSecondary	0.05	0.00	0.33	0.00	0.11
Previous.qualificationHigher	0.14	0.00	0.47	0.00	0.02
NationalityOthers	0.17	0.00	0.08	0.00	0.02
Mother.s.qualificationSecondary	0.42	0.96	0.69	0.17	0.53
Mother.s.qualificationHigher	0.37	0.84	0.53	0.00	0.47
Father.s.qualificationSecondary	-0.12	0.00	-0.04	0.00	-0.05
Father.s.qualificationHigher	-0.17	0.00	-0.11	0.00	-0.12
Mother.s.occupationWhite Collar	0.07	0.00	0.18	0.06	0.12
Mother.s.occupationOthers	-0.27	0.00	-0.23	-0.30	-0.03
Father.s.occupationWhite Collar	-0.04	0.00	-0.03	0.00	-0.01
Father.s.occupationOthers	0.11	0.00	-0.10	0.00	0.01
Displaced	0.01	0.00	0.02	0.00	0.07
Educational.special.needs	-0.30	0.00	-0.15	0.00	-0.15
GenderMale	-0.64	-0.94	-0.90	-0.83	-0.71
Scholarship.holder	1.27	2.18	2.06	1.99	1.59
Age.at.enrollment	-0.27	-0.37	-0.30	-0.31	-0.14



As previously assumed, from the graph we can see that the coefficients of the various models are very similar to each other. All models give a considerable importance to the intercept, i.e. the reference group, except for the LDA model which does not have a direct intercept in the Rstudio output. All models agree on assigning a negative weight to Male sex, a positive weight to Scholarship holder, which is consistent with the previous analysis and a negative weight for the age. Except for the Lasso model, they all agree with the positive importance of the mother's level of education, however Father's qualification is not significant since is high correlated with it. Finally, it can noted that only the Ridge model gives importance to the student's level of education, in contrast to what was assumed in the EDA (table 7).

5.1 Explaining the model

As we can see previously, the logistic model is the best model to explain the phenomenon, so we will use it to explain the results.

Diversamente da quanto abbiamo assunto nell'EDA, essere sposati o non avere un titolo di studio superiore sembrano non essere le variabili influenti nel determinare se uno studente si laureerà in tempo o meno. L'unica variabile familiare che risulta rilevante è il titolo di studio della madre, che ha un effetto positivo sulla probabilità di laurearsi in tempo di almeno 2.3 volte rispetto a chi non ha una madre che non ha terminato le superiori.

Non c'è una differenza significativa tra i fuori sede e non e neanche tra chi ha bisogno di aiuti speciali.

Chi scegli un corso stem ha 1.64 volte di probabilità in più di laurearsi in tempo rispetto a chi sceglie un corso non stem.

Anche essere maschi impatta negativamente sulla probabilità di laurearsi in tempo, con un decremento di circa il 38% rispetto alla probabilità che hanno le femmine.

Mentre l'aiuto economico della borsa di studio ha un impatto notevolemente positivo sulla probabilità di

laurearsi in tempo, con un incremento di 7.86 volte rispetto a chi non ha una borsa di studio, andando ad indicare che non dover pensare a come pagarsi gli studi e potersi permettere di non lavorare durante il periodo di studio ha un impatto notevolmente positivo.

Infine, l'età ha un effetto negativo sulla probabilità di laurearsi in tempo, con un decremento del 70% per ogni anno in più di età, andando ad indicare che più tardi s'inizia a fare l'università maggiori sono le probabilità probabilità di diventare fuoricorso o abbandonare gli studi.

Data Characteristics

Feature Quality: The predictors appear to have a similar level of influence on the response variable (is_Graduated). This suggests that no single model is able to extract substantially more information from the predictors than the others.

Signal-to-Noise Ratio: The comparable performance across models indicates a balanced signal-to-noise ratio in the data. If the data were very noisy or the predictors weak, greater variability in model performance would be expected.

Multicollinearity: The predictors might be collinear, meaning they contain overlapping information. This can cause different models to perform similarly because they are capturing the same patterns in the data.

Model Complexity:

simpler models (e.g., Logistic Regression) and more complex models (e.g., QDA) perform similarly, suggesting that increasing model complexity does not lead to significant improvements. This implies that the underlying relationships in the data are relatively straightforward and do not require complex modeling to capture.

Data Linearity:

the similar performance of linear models (Logit, LDA) and non-linear models (QDA) suggests that the relationships between predictors and the response variable may be approximately linear. Non-linear models do not have a significant advantage, indicating no strong non-linear patterns in the data that are not being captured by the linear models.

Predictive Power:

The results suggest that the predictors sare a limited predicting power, indicating that there are be other unobserved factors influencing whether a student graduates on time or not,this is consistent with our project framework.

6 Conclusions

We were able to build several models that are able with a moderate degree of certantly whether a student will graduate or not just by looking at enrollment data. A tool like this could be useful in several real world applications, universities could use these to better understand who is struggling to graduate in time and implement policy changes to address eventual issues

DA RIVEDERE

In conclusion, we have built several models that are able to predict with a moderate degree of certainty whether a student will graduate on time or not just by looking at enrollment data and if they are scholarship holder. These models can be useful in several real-world applications, such as helping universities identify students who are at risk of dropping out and implementing interventions to support them. By understanding the factors that influence student success, universities can take proactive measures to improve graduation rates and student outcomes. The models we have developed provide a valuable tool for universities to identify at-risk students and provide them with the support they need to succeed.

- SPIEGARE IL FATTO CHE PARTE DEL SUCCESSO NON È DOVUTO ALLE VARIABILI CONSIDERATE, MA A VARIABILI NON OSSERVATE. SI POTREBBE CONSIDERARE COME VARIABILE IL VOTO DI USCITA DALLE SUPERIORI.

- RISPIEGARE IN BREVE QUALI SONO LE VARIABILI CHE INFLUENZANO DI PIÙ IL RISULTATO
- RICORDARSI CHE IL MODELLO NON SPIEGA SE LO STUDENTE SI LAUREA O MENO, MA SE LO FA IN TEMPO
- SPIEGARE CHE IL MODELLO NON È PERFETTO, MA È UN BUON INIZIO PER CAPIRE QUALI SONO LE VARIABILI CHE INFLUENZANO DI PIÙ IL RISULTATO

6.1 Limitations and Future Work:

Our models are based on enrollment data and do not take into account other factors that may influence student success, such as personal circumstances, academic performance, or external factors. Future work could include incorporating additional data sources to improve the accuracy of the models and provide a more comprehensive understanding of student success. Additionally, further research could explore the impact of different interventions on student outcomes and identify the most effective strategies for supporting at-risk students.

7 References

Realinho, V.; Machado, J.;Baptista, L.; Martins, M.V. Predicting Student Dropout and Academic Success. Data 2022, 7, 146. https://doi.org/10.3390/data7110146

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

7.1 Appendix A

Table 28: Dataset Description

Variable	Description
Married	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
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

Variable	Description
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

‘‘‘