Statistical Learning project

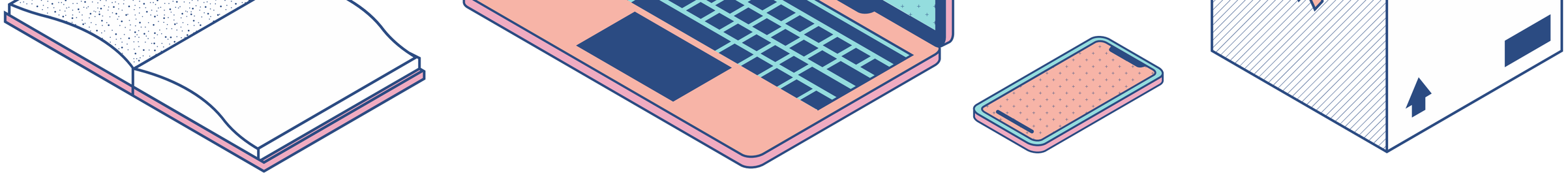
Di Pasquale Giuseppe,
Gorni Silvestrini Matteo

Overview

1. Introduction
2. Data Preprocessing
3. Exploratory Data Analysis
4. Model building
5. Model Evaluation
6. Results and Conclusions

Introduction

- The dataset comes from Polytechnic Institute of Portalegre in Portugal
- 4424 students involved, 35 features for each
- The data was originally collected to train machine learning models



Original task

Students labeled in three possible ways:

- Enrolled
- Dropout
- Graduate

Curricular data and academic performance is used when building models

Our work

Students labeled in two ways:

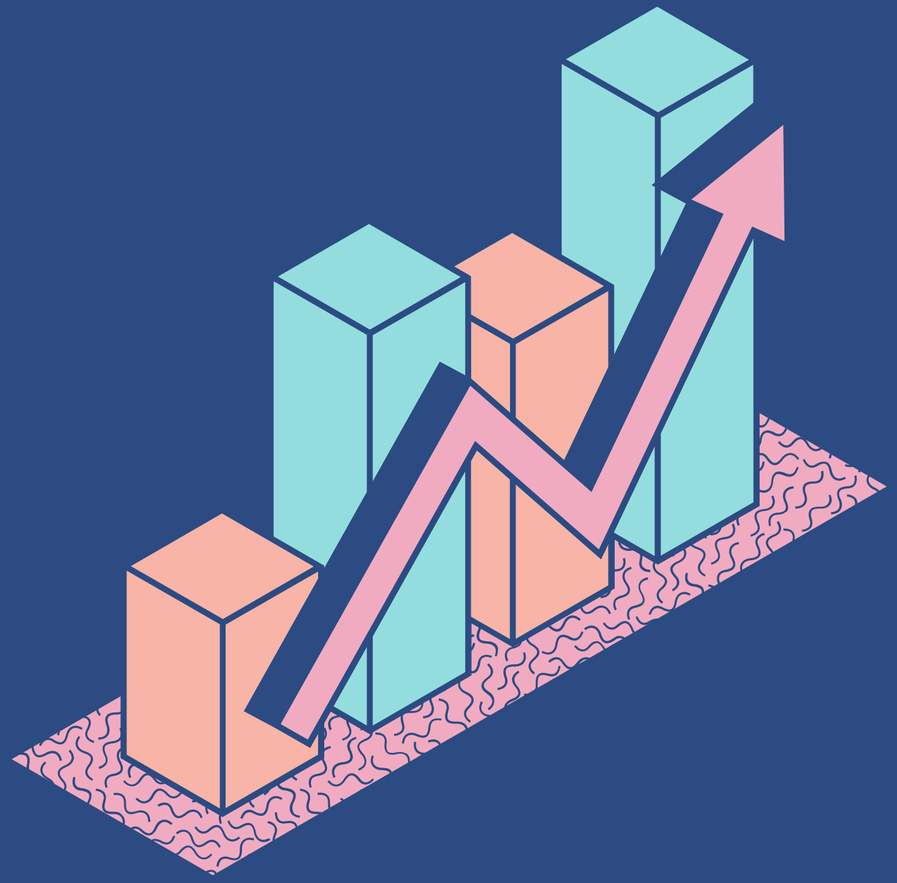
- Finished in time: yes/no

Research question: can we estimate wheter a student will finish their course in time using enrollment data only? If so, which variables are significant for this task?

Data preprocessing

STEPS	FEATURES INVOLVED
Check an handle missing values	None
Merge labels	marital status\nationality\parents and student education\ parents occupation\course
Remove unusable features	application mode\macroeconomic data\curricular performance data
Remove outliers	Age at enrollment

EXPLORATORY DATA ANALYSIS



EDA- Categorical features 1/2

Table 14: Displaced VS is Graduated

	No	Yes	Sum
0	1113	885	1998
1	1102	1324	2426

Table 17: Gender VS is Graduated

	No	Yes	Sum
Female	1207	1661	2868
Male	1008	548	1556

Table 16: Scholarship VS is Graduated

	No	Yes	Sum
0	1951	1374	3325
1	264	835	1099

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

EDA- Categorical features 2/2

Table 6: Marital status VS is Graduated

	No	Yes	Sum
No	1904	2015	3919
Yes	311	194	505

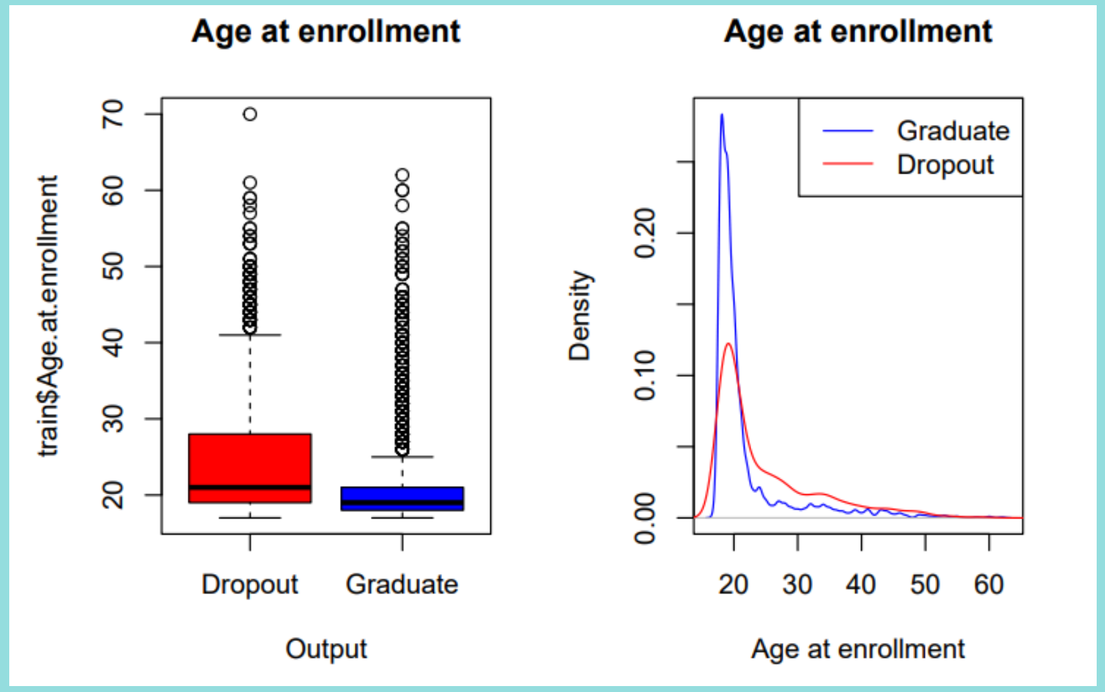
Table 8: Course VS is Graduated

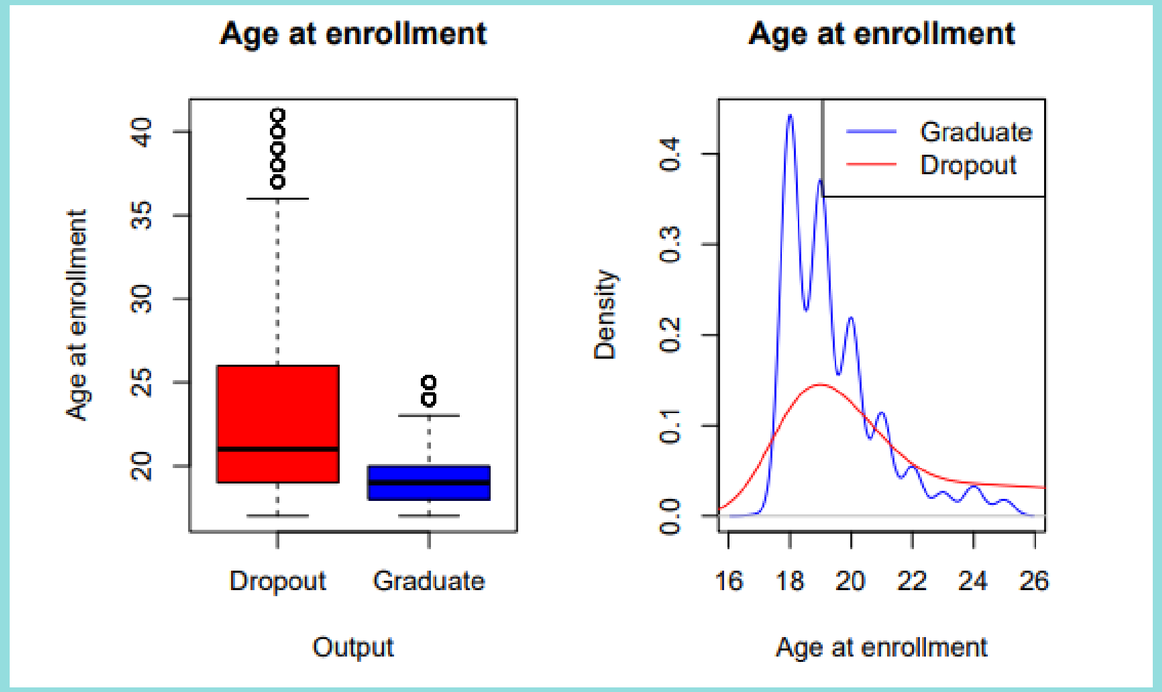
	No	Yes	Sum
Stem	822	900	1722
No Stem	1393	1309	2702

Table 7: Previous qualification VS is Graduated

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

EDA-Outliers removal

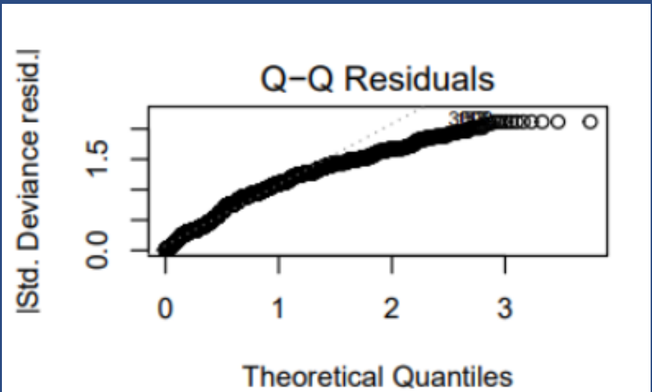




Model building

Logistic regression

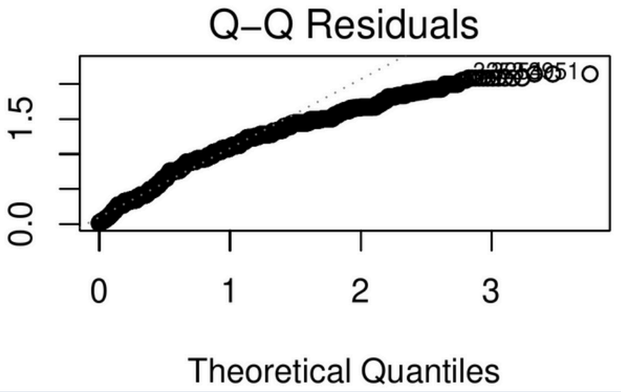
```
glm(formula = is_Graduated ~ ., family = binomial, data = train)
```




```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        0.61742    0.49234   1.254   0.2098
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 ***
```

Logistic regression stepwise

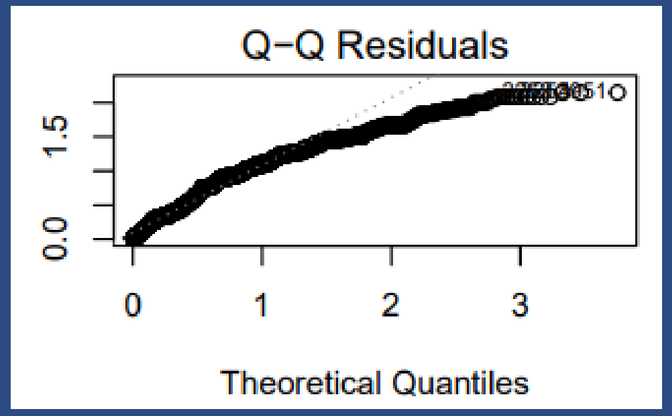
```
glm(formula = is_Graduated ~ Course + Mother.s.qualification +
    Gender + Scholarship.holder + Age.at.enrollment, family = binomial,
    data = train)
```




```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      0.51575    0.26261   1.964 0.049535 *
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 ***
```

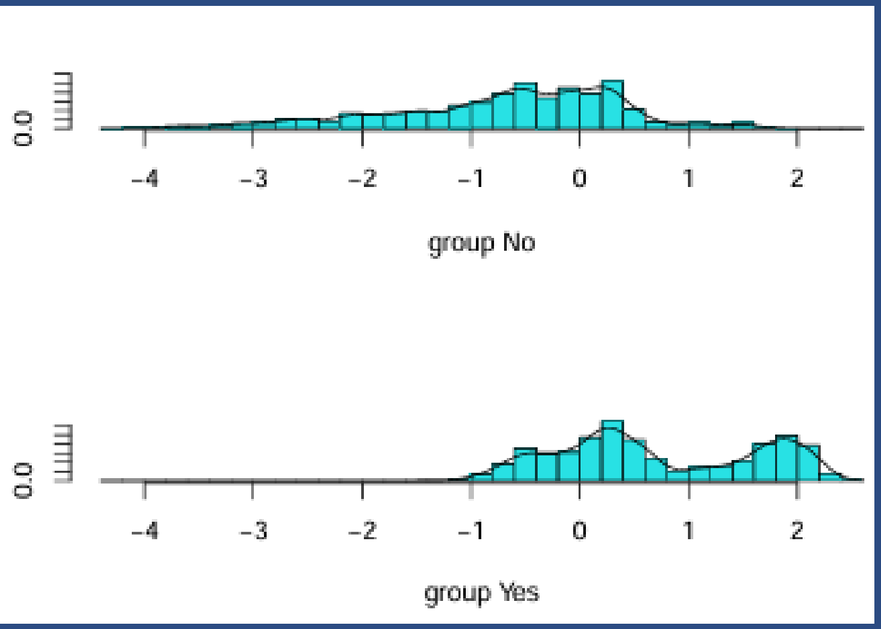
Logistic regression with interactions

```
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 Pr(>|z|)
(Intercept)                                          0.80097  0.00453 **
CourseNo Stem                                       -0.71582 5.92e-09 ***
Mother.s.qualificationSecondary                      0.70977  0.00692 **
Mother.s.qualificationHigher                         0.59889  0.03376 *
GenderMale                                          -1.20512 1.62e-13 ***
Scholarship.holder                                   1.59962  0.01210 *
Age.at.enrollment                                   -0.34760  < 2e-16 ***
CourseNo Stem:GenderMale                             0.62710  0.00221 **
Mother.s.qualificationSecondary:Scholarship.holder   1.35887  0.03260 *
Mother.s.qualificationHigher:Scholarship.holder      1.66945  0.06897 .
GenderMale:Scholarship.holder                       -0.85532  0.00335 **
Scholarship.holder:Age.at.enrollment                -0.11990  0.06573 .
```

LDA



QDA

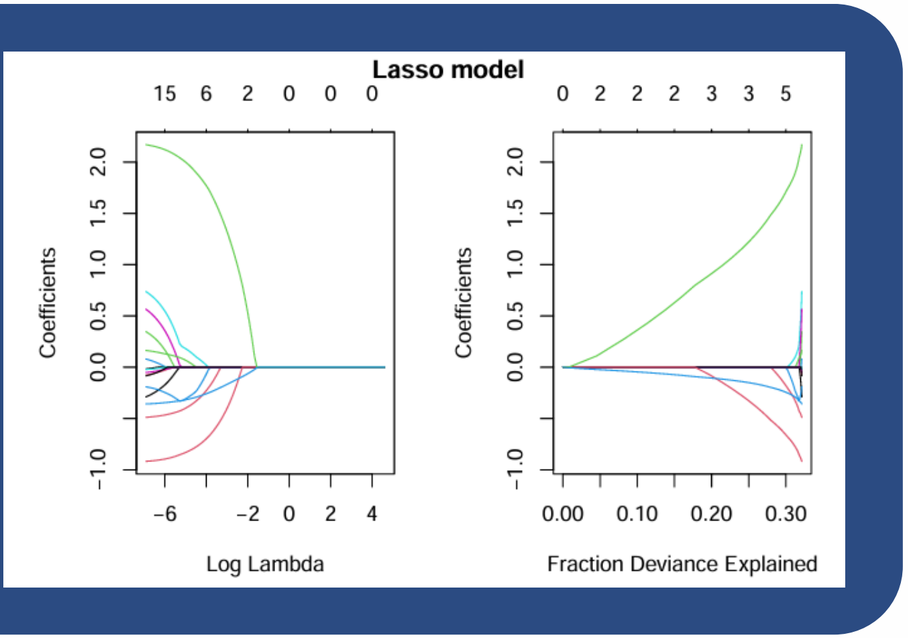
Table 20: Distribution of the results respect to NOT graduating on time in the train set

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

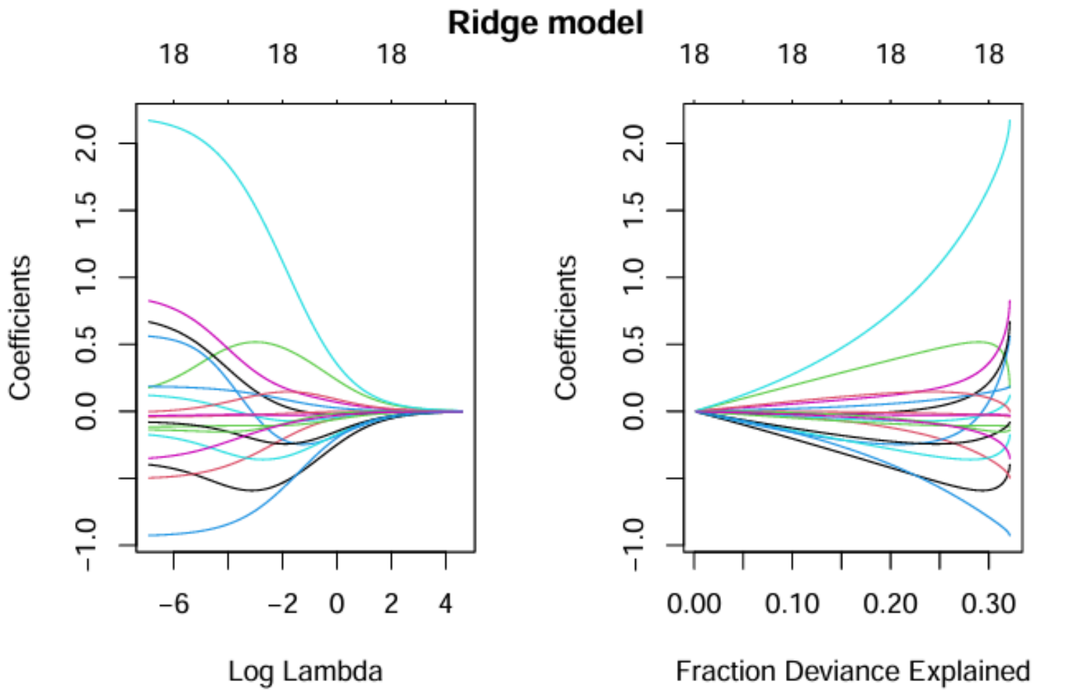
Table 21: Distribution of the results respect to graduating on time in the train set

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

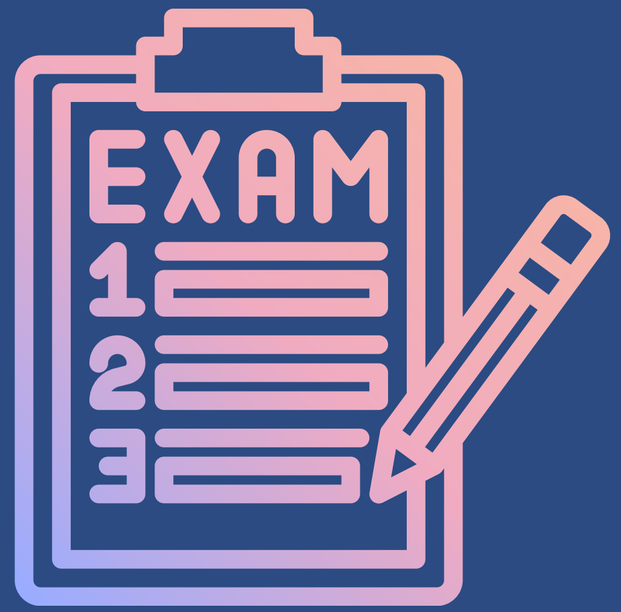
Lasso regression



Ridge regression



Models evaluation



ROC curves and precision metrics

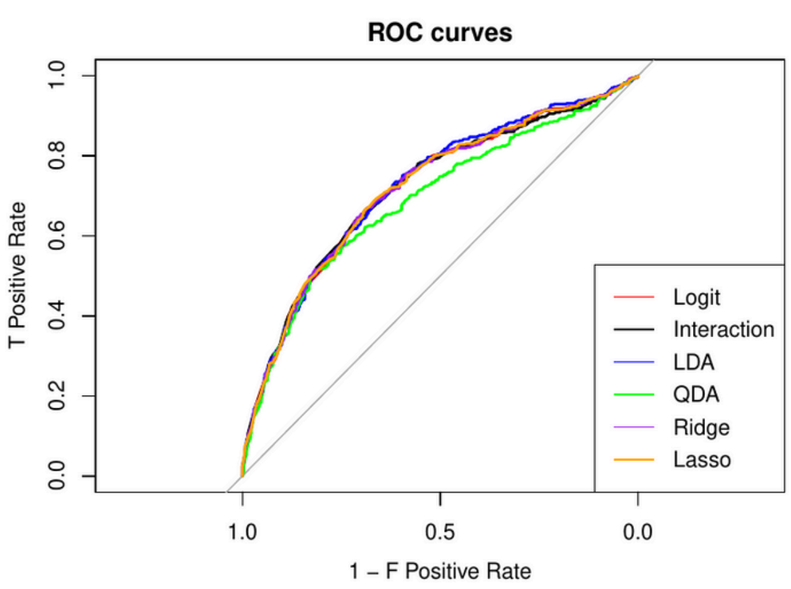
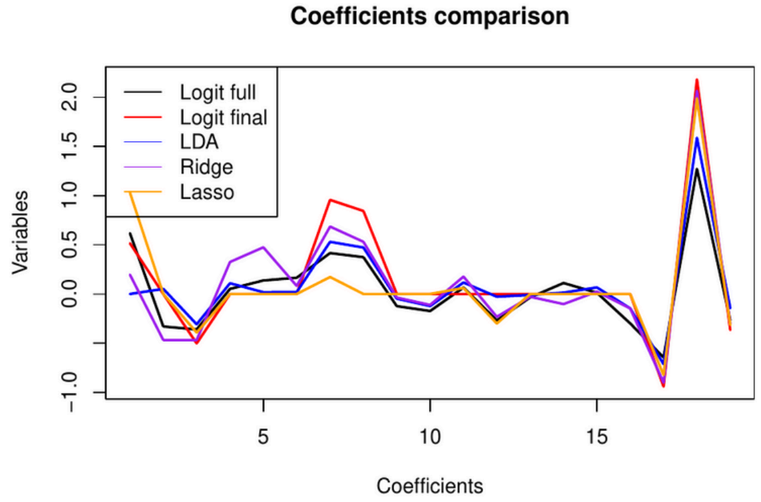


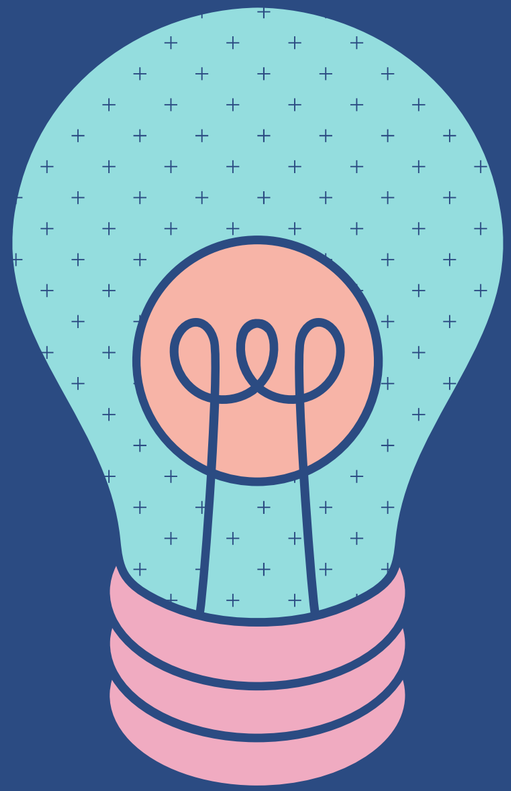
Table 26: Model evaluation

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

Coefficients comparison



Results and Conclusions



Some interesting results...

Table 27: Coefficients comparison

	Logit_full	Logit_final	Ridge	Lasso	LDA
(Intercept)	0.62	0.52	0.20	1.03	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

... and some final considerations:

- Limitations of our analysis
- Potential applications of our model

The end