



Predicting Students' Dropout

Joshua Damm | s4036018

Anete Mürk | s3790401

Hidde van Rooijen | s3377954



The dataset: Students' Dropout and Academic Success

- Classification problem
- Information known at enrollment
- Academic performance after first and second semester
- Various undergraduate degrees
- Demographic, social-economic, academic performance data.

Creators

Valentim Realinho



vrealinho@

Instituto Politécnico de Portalegre

Mónica Vieira Martins



mvmartins@ipportalegre.pt

Instituto Politécnico de Portalegre

Jorge Machado



jmachado@ipportalegre.pt

Instituto Politécnico de Portalegre

Luís Baptista



lmtb@ipportalegre.pt

Instituto Politécnico de Portalegre



Research Question

What are the key predictors of academic success in undergraduate programs at Polytechnic Institute Of Portalegre?

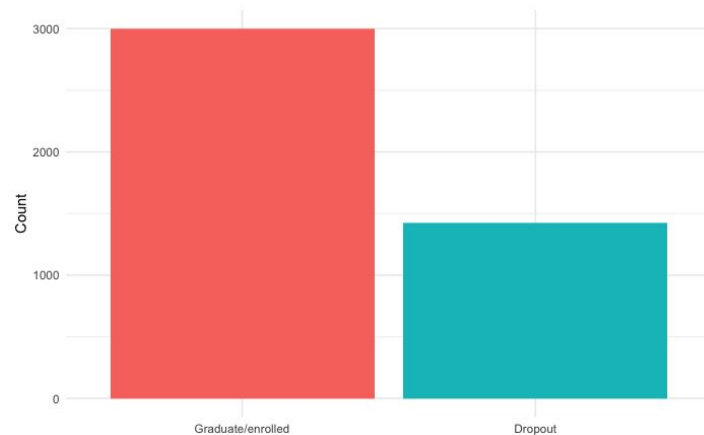
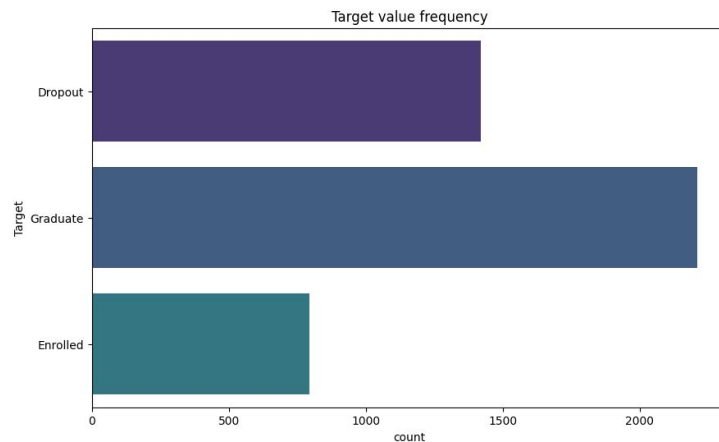


Exploratory Analysis



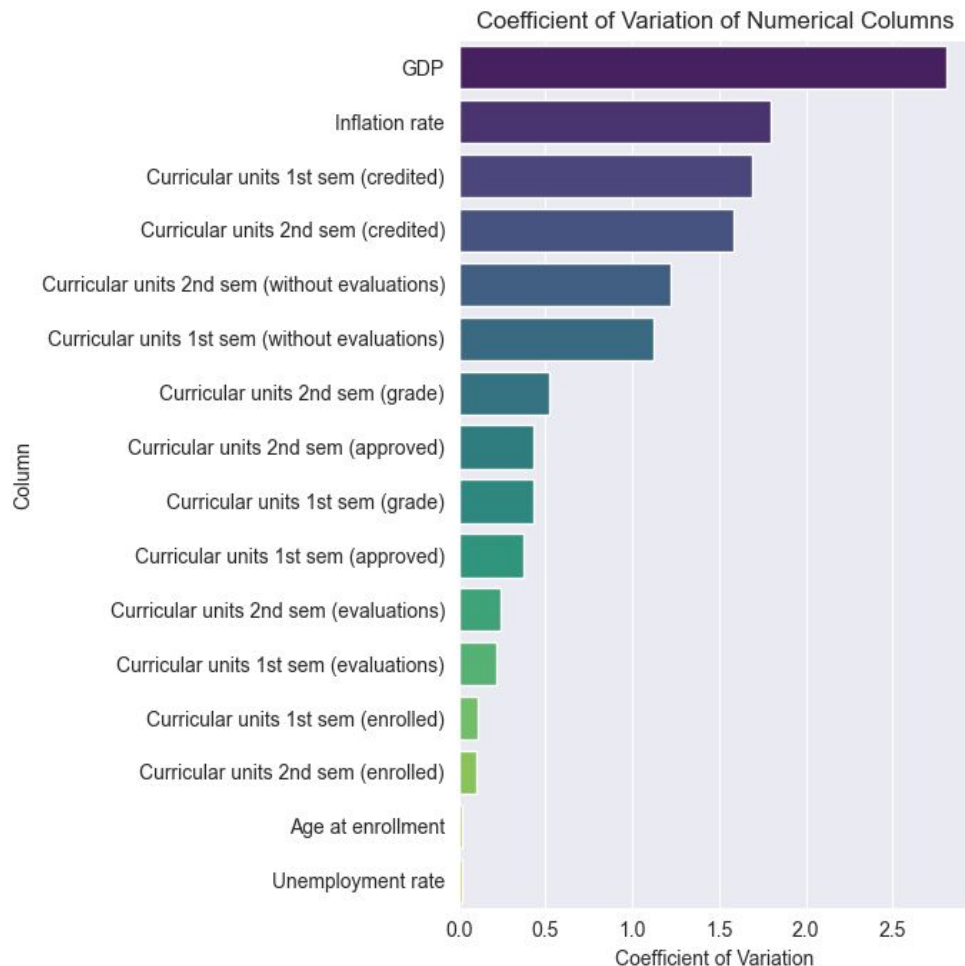
Dataset specifications

- Null values: 0
- Categorical variables: 19
- Numerical variables: 16
- N.o. students: 4424
- N.o. classes: 3



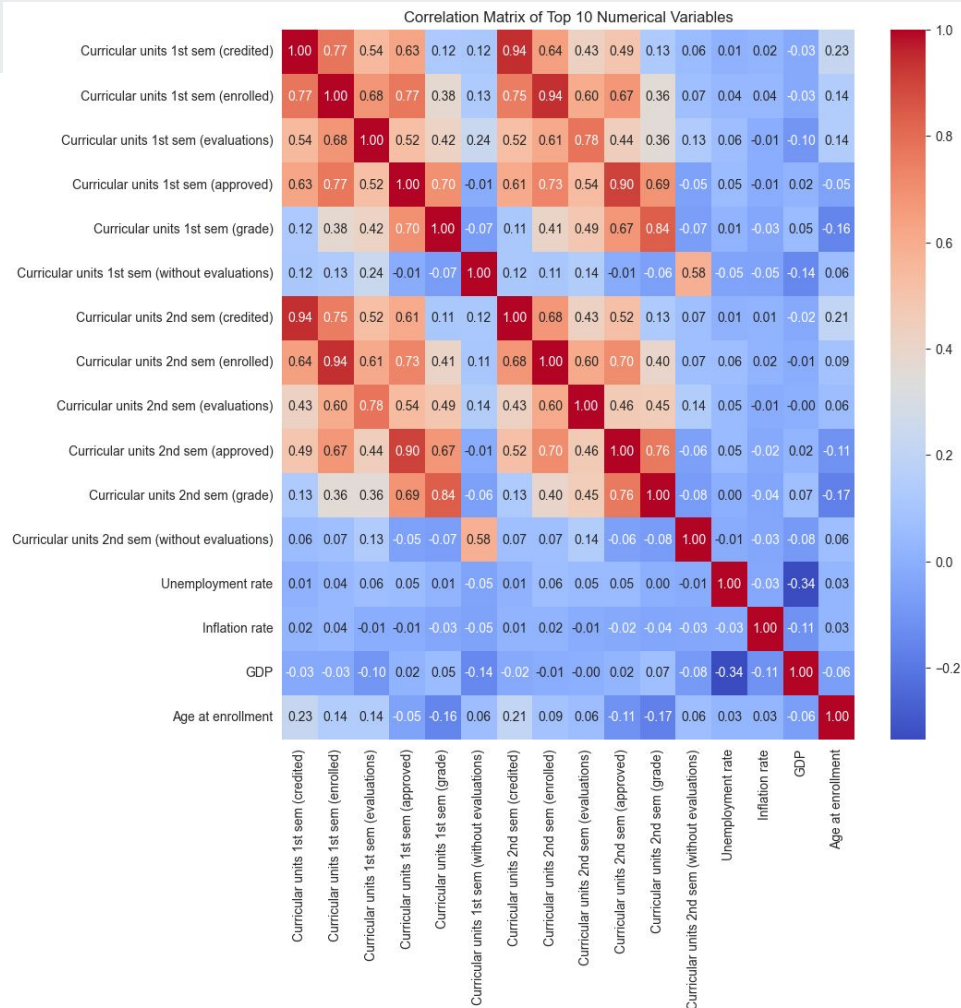
Coefficient of variation

- Log-transformed columns
- High CV in number of (credited) courses
- Extremely high CV for GDP



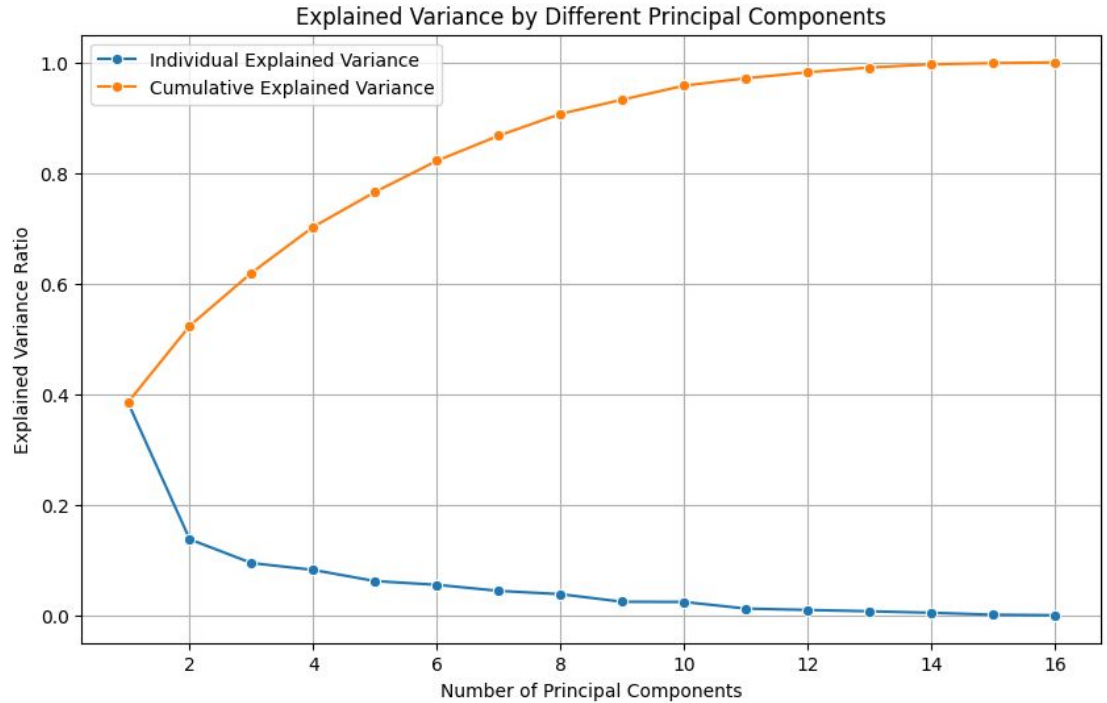
Correlation numerical columns.

- Moderate/high correlation curricular unit variables
- GDP not correlated



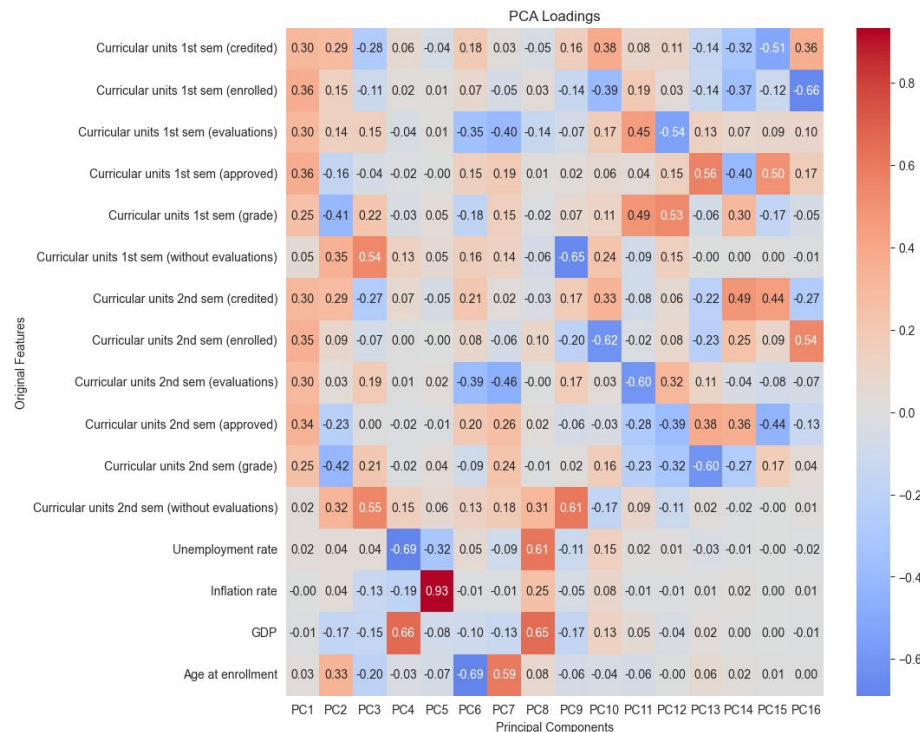
Explained variance PCA

First 4-5 components explain a significant amount of variance.



PCA loadings numerical columns

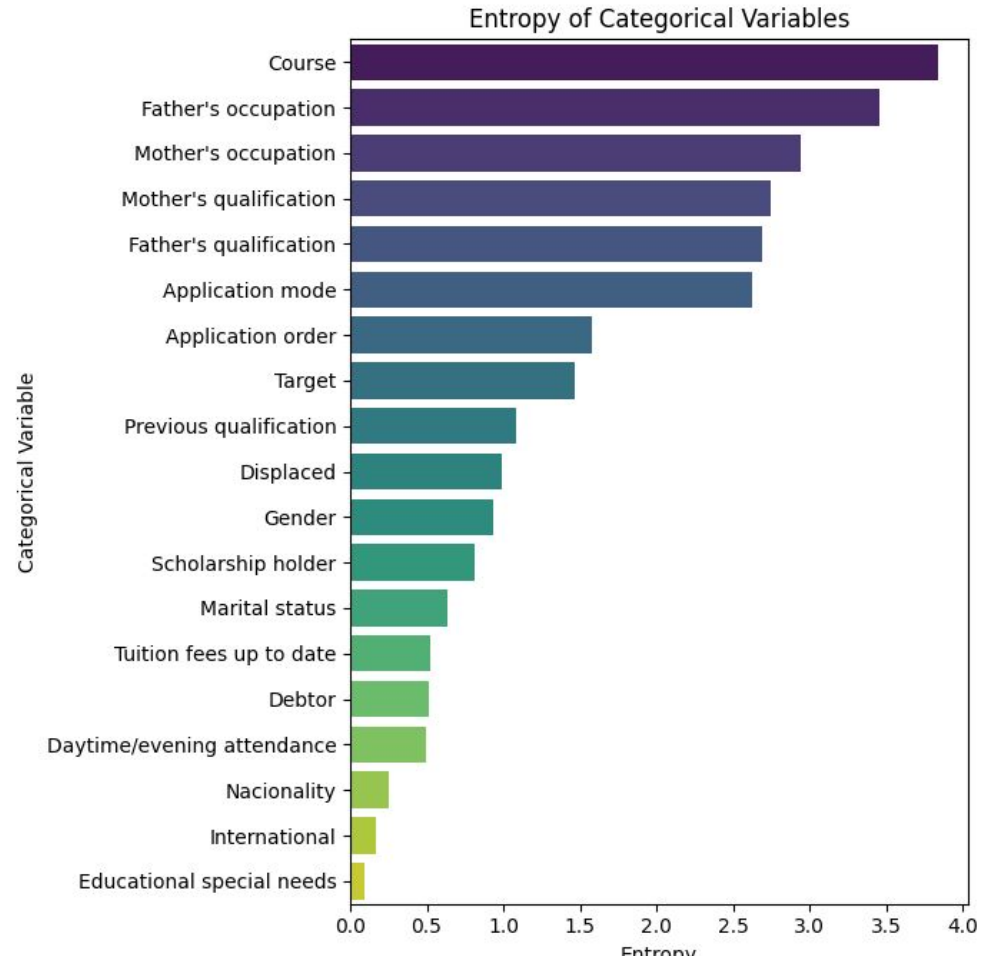
- First 4-5 PCs explain ~80% of variance
- Curricular units important for PC1
- PC2 likely resembles poor academic performance





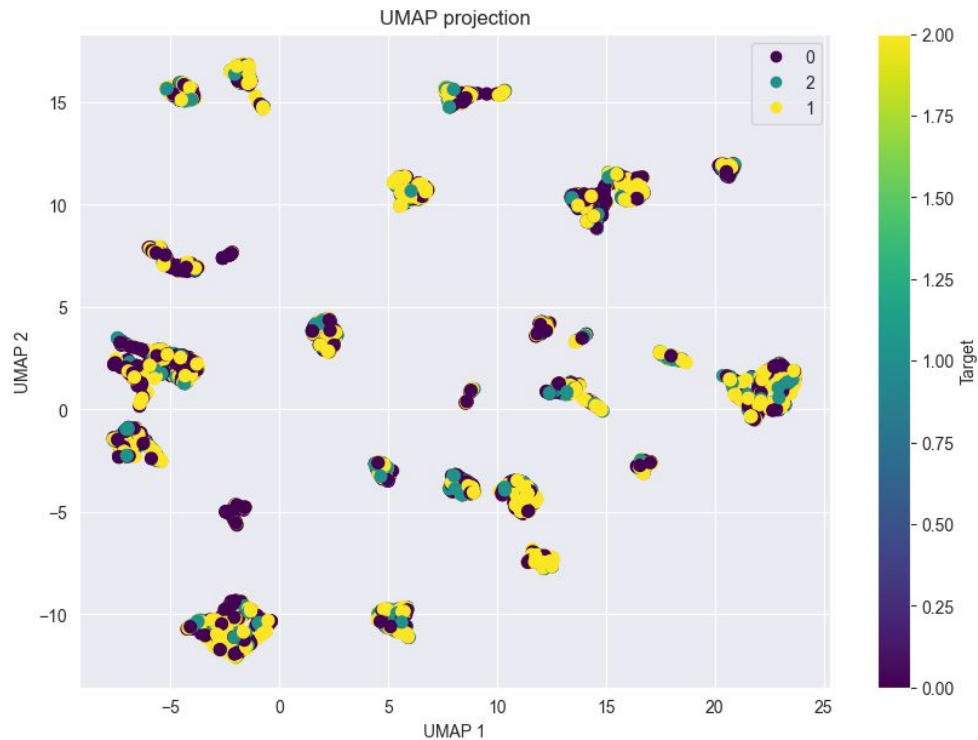
Entropy categorical variables

- Large amount of variables with low levels of variability:
 - Very few classes
 - Highly skewed distributions
- Large diversity in student background



UMAP projection categorical variables

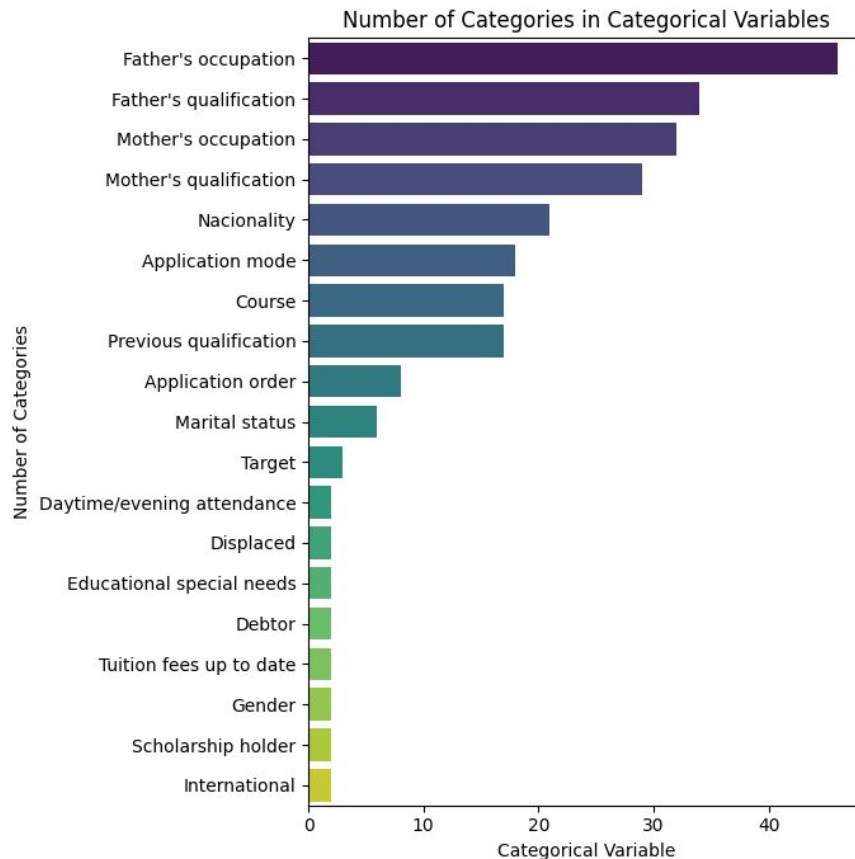
- Classes not well separable based on categorical features
- Many distinct groups of categorical variables





N.o. categories in categorical variables

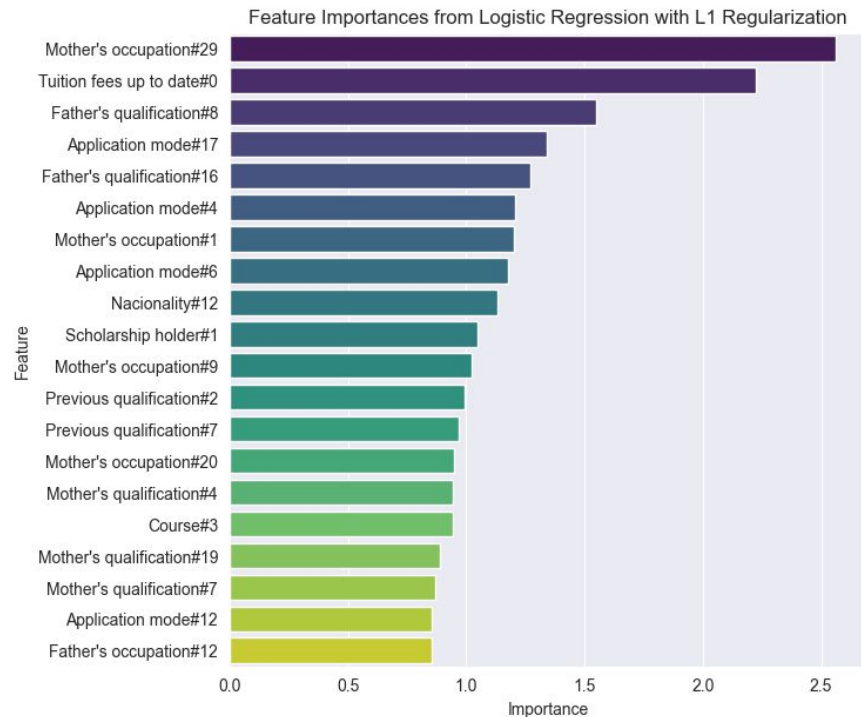
- Some categorical variables contain a large amount of classes
- Some of these have low entropy (see previous slide), meaning high risk of bias.
- One-hot encoded dataframe contains **261 columns**





Feature importances logistic regression, L1 regularization

- Many moderately important class values





EDA Conclusions

- Imbalanced target classes
- Few numerical columns explain most of the variance
- Dataset contains many categorical variables, some with a lot of classes
- High risk of overfitting on individual classes
- High dimensional



Methods

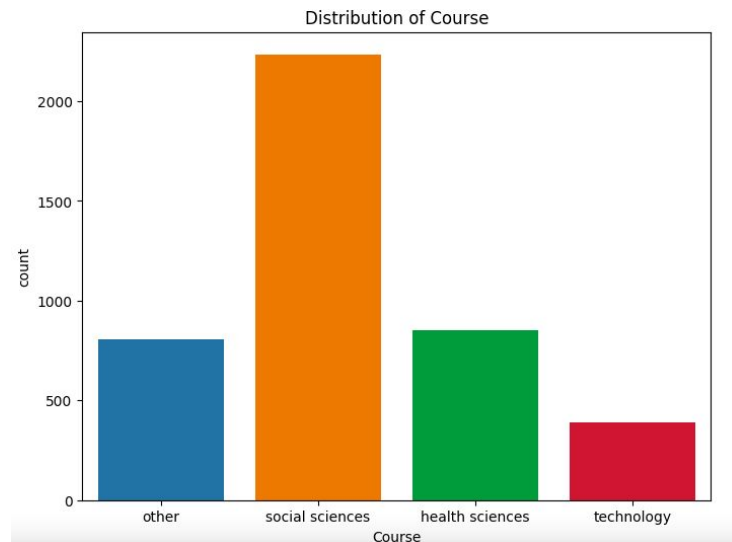
- **Random Forests:**
 - Can handle high dimensional data
 - Can handle both numerical and categorical data
 - Interpretable (for answering RQ)
 - Less sensitive to hyper-parameters
- **GBM:**
 - Can handle high dimensional data
 - Can handle imbalanced data
 - Can handle both numerical and categorical data
 - Interpretable (for answering RQ)
- **Logistic Regression with Regularization**
 - L1 or L2 regularization reduces risk of overfitting
 - Interpretable (for answering RQ)
- **SMOTE for balancing target data**



Model training

Dimension reduction

- Binning data to reduce dimensions
- ChatGPT 4o very useful for such tasks!
- From 261 columns down to 93 (one-hot_encoded)





Class weights

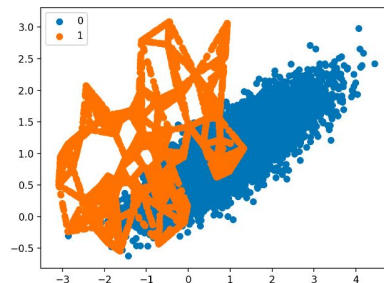
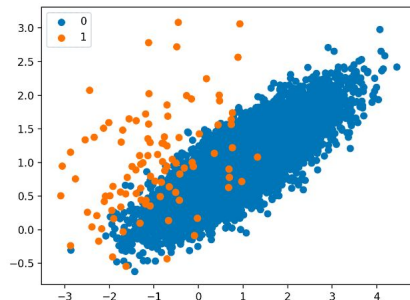
$$w_j = n_{\text{samples}} / (n_{\text{classes}} * n_{\text{samples}_j})$$

- w_j is the weight for each class (j signifies the class)
- n_{samples} is the total number of rows in the dataset
- n_{classes} is the total number of classes in the target variable
- n_{samples_j} is the total number of rows in class j

$$\log \text{loss} = \frac{1}{N} \sum_{i=1}^N [-(w_0(y_i * \log(\hat{y}_i)) + w_1((1 - y_i) * \log(1 - \hat{y}_i)))]$$

Synthetic Minority Oversampling TEchnique (SMOTE)

- Chawla et al. 2002*
- Increase the sample size of minority class with synthetic samples



<https://machinelearningmastery.com/sMOTE-oversampling-for-imbalanced-classification/>

*Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.



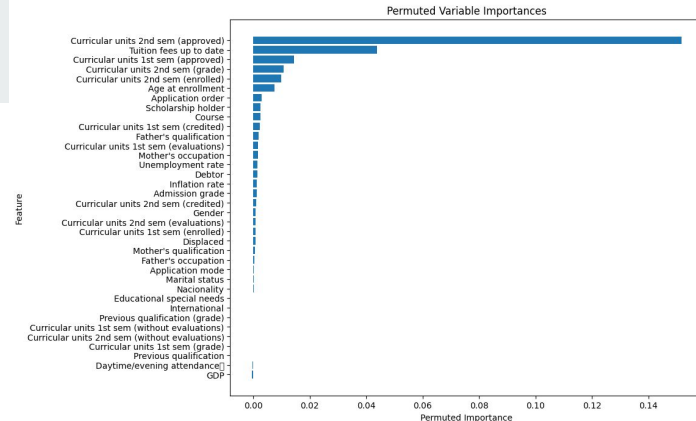
Logistic regression

- Data transformation to meet linear model assumptions. Standardizing the data to compare coefficients and enable efficient SMOTE
- Grid-search with 5-fold CV
- Scoring on ROC AUC
- Two models:

	Logistic regression (class weights)	Logistic regression (SMOTE)
k-neighbours	-	7
sampling_rate	-	0.5
penalty	L1	L1
C	0.1	0.1

GBM

- Full dataset with binned data since non-parametric models don't make assumptions
- Remove curricular units 2nd semester (approved) - heavily overpowers other predictors
- Grid search with 5-fold CV
- Two models:



	GBM (class weight)	GBM (SMOTE)
k-neighbors	-	5
sampling_strategy	-	0.6
learning_rate	0.05	0.05
max_depth	4	6
n_estimators	200	100
subsample	0.6	0.6



Random forest

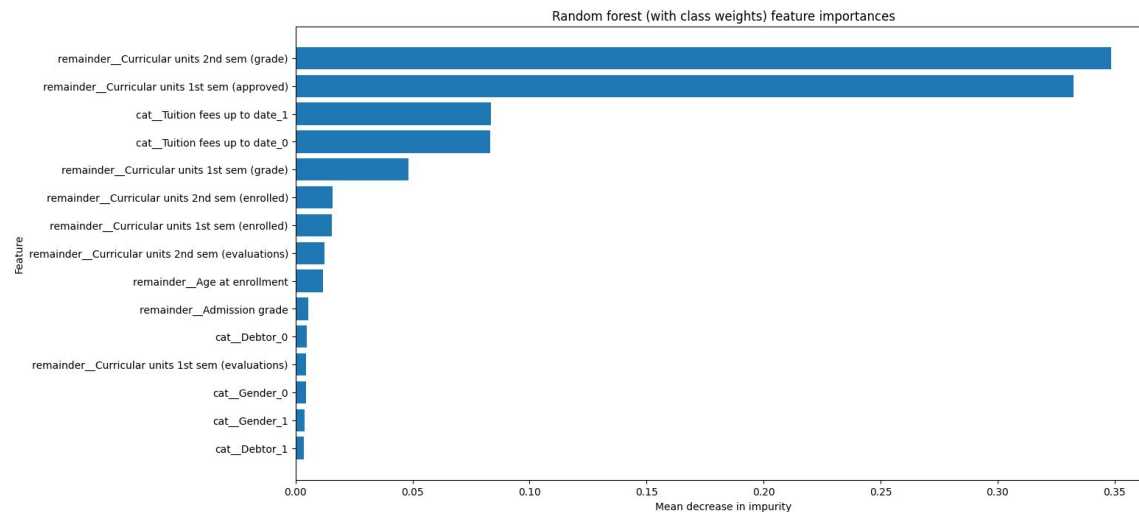
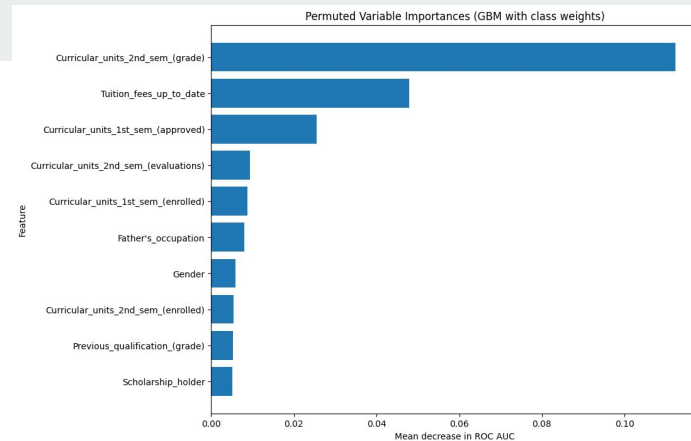
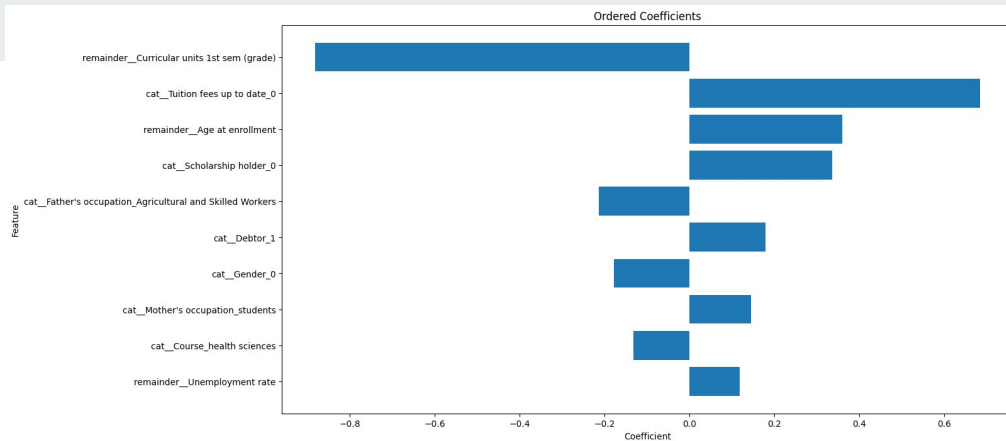
	RF (class weight)	RF (SMOTE)
k-neighbors	-	7
sampling_strategy	-	0.5
max_features	0.6	0.6
max_depth	4	4
n_estimators	300	300
min_samples_leaf	5	5



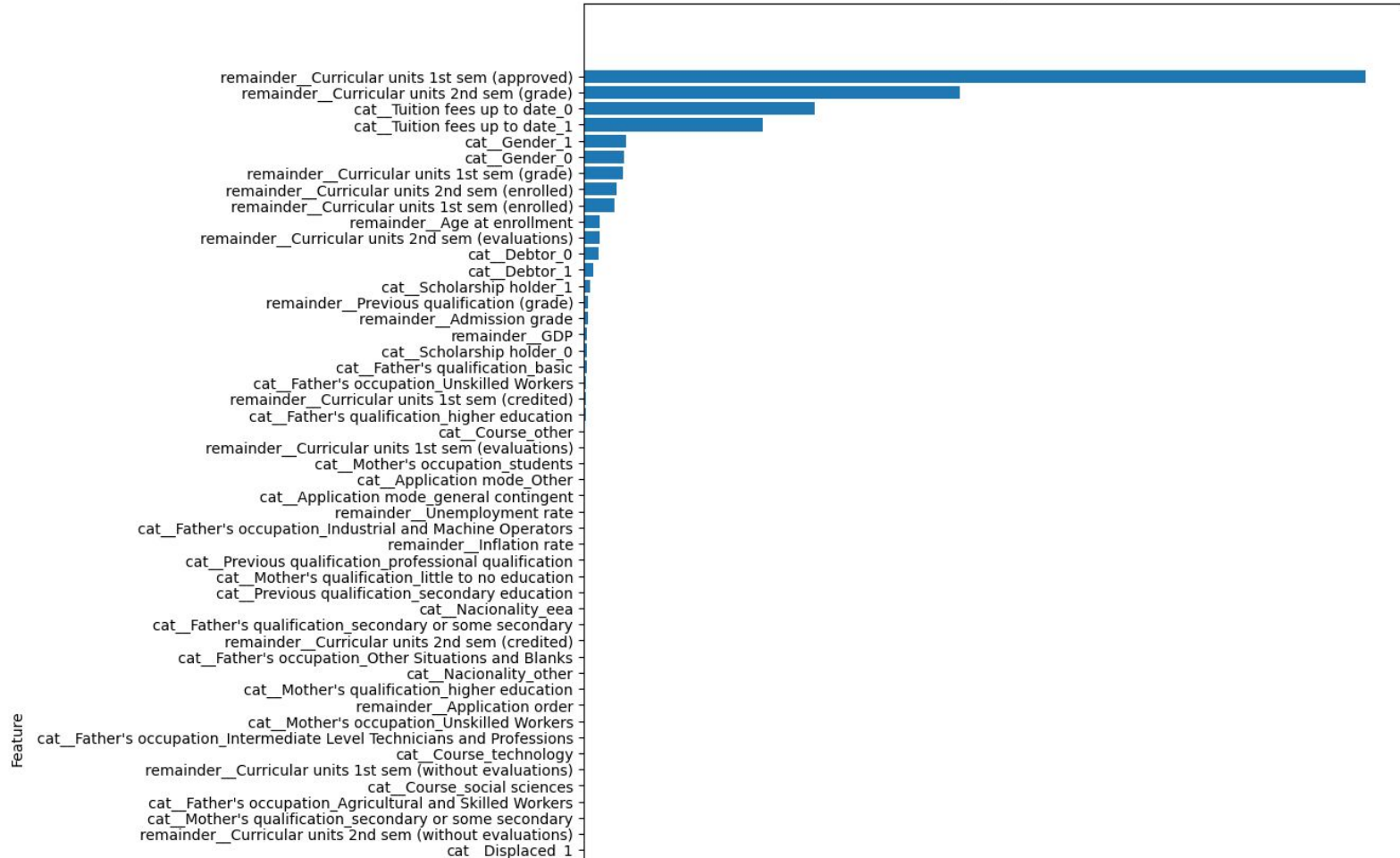
Results

<i>class weights</i>	Logistic Regression	GBM	Random Forest
F1-score no dropout	0.85	0.89	0.88
F1-score dropout	0.69	0.77	0.76
ROC AUC	0.77	0.84	0.83

<i>SMOTE</i>	Logistic Regression	GBM	Random Forest
F1-score no dropout	0.86	0.90	0.90
F1-score dropout	0.69	0.77	0.78
ROC AUC	0.77	0.83	0.83



Variable Importances





Conclusions



Research Question

What are the key predictors of academic success in undergraduate programs at Polytechnic Institute Of Portalegre?



Results recap

Best predictors of academic success across all methods (top 6 variables):

- **Students' course performance measured by curricular units**
- **Tuition fees up to date**

Ambiguities:

- **Gender:** men are less likely to graduate than women (LR, RF)
- **Occupational status:** High father's occupation makes it less likely to dropout, especially for agricultural and skilled workers (LR, GBM)
- **Age:** Older students are more likely to drop out than younger students (LR, GBM)
- **Scholarship holder:** Students holding a scholarship are less likely to dropout (LR)
- **Debtor:** Students who took on debts for their studies are more likely to dropout (LR)



Methods reflection





- Good ROC_AUC and F1 scores in overall classification across all methods used:
 - Logistic Regression (class weights, SMOTE) ~ 0.77 ROC_AUC
 - **Gradient Boosted Machines (class weights) ~ 0.84**, GBM (SMOTE) ~ 0.83
 - Random Forest (class weights, SMOTE) ~ 0.83
- **However, all models still perform worse for the prediction of the minority dropout class (0.12 - 0.17 F1 score)**
- Class weights as **imbalance technique** perform similar to synthetic minority oversampling (SMOTE), although it is way less computationally extensive
- Binary classifier performance metrics (ROC_curve, F1 score) are able to **pinpoint misclassifications properties**
- Tree ensemble and boosting methods show comparably **high predictive accuracy** in predicting student dropout, but at the expense of **less precise interpretability**

Comparing results to original paper

- Martins et al. (2021) also used LR, RF and GBM on similar data with SMOTE to predict dropout
- **3 class problem** (failure, relative success, and success)
- They did not have academic performance after enrollment (strongest predictors in our case)
- Lower F1-score in predicting student dropout for all methods used (worse performance for minority class):
 - Logistic Regression (SMOTE) ~ 0.61
 - **Gradient Boosted Machines (SMOTE) ~ 0.72**, Random Forest (class weights, SMOTE) ~ 72%
 - *Support Vector machines* ~ 60%



Early Prediction of student's Performance in Higher Education: A Case Study

Mónica V. Martins¹ , Daniel Tolledo¹, Jorge Machado¹ ,
Luís M. T. Baptista¹ , and Valentim Realinho^{1,2} 

¹ Polytechnic Institute of Portalegre (IPP), Portalegre, Portugal
mvmartins@ippportalegre.pt

² VALORIZA - Research Center for Endogenous Resource Valorization, Portalegre, Portugal



Limitations

- **cross-sectional study:** only information about predictors at the time of enrollment (and students' performance in the 1st and 2nd semester) available
- **Academic performance metrics in the 1st and 2nd semester** as predictors is self-evident and outpowers all other predictors by far
- **Lower prediction accuracies for the minority risk group**, which is our target
- **No information about students' motivation and compatibility with the program** available (e.g. unmet expectations and needs, personality, cognitive abilities and skills)



Research Question

What are the key predictors of academic success in undergraduate programs at Polytechnic Institute Of Portalegre?

Students' course performance and whether they pay their tuition fees on time predict academic success. However, the precise underpinnings of academic success are not as clear. Being younger, female, holding a scholarship, having taken on no debts, and having a father with high occupational status puts students at lower risk of dropout.



Future directions and improvements

- **Minority sampling techniques** should be further elaborated to better **investigate risk groups**
- **Remove curricular units** and look solely at **sociodemographic predictors** for better predictability in a real-world scenario
- **More nuanced information** about students' motivation and attitude towards their program could be collected
- **Longitudinal assessment** might yield valuable insights into different academic profiles and trajectories deciding about success



Thanks for your attention!