**Capstone Project: Predict Students' Dropout and Academic Success**

**Problem Understanding**

Nowadays, there are thousands of students who enroll in different major in different universities. However, our target is to predict students’ dropout and academic success. Incorporating many factors like academic path, demographics, and social-economic factors, we are going to build classification models to predict students' dropouts and academic success and we classified the problem as three category classification task. Our metric includes several steps: data understanding, cleaning, and preprocessing; Exploratory Data Analysis (EDA) and initial model prototyping;model training, hyperparameter tuning, and addressing class imbalance and final model evaluation.

* Target Variable: Target
* Features:Marital Status,Application mode,Application order,Course,Daytime/evening attendance,Previous qualification,Previous qualification (grade),Nationality,Mother's qualification,Father's qualification,Mother's occupation,Father's occupation,Admission grade,Displaced,Educational special needs,Debtor,Tuition fees up to date and so on.

**About Dataset**

This dataset contains continuous and categorical data aiming to explain the relationship of students’ dropouts and academic success between many other variables (academic path, demographics, and social-economic factors) . This dataset is collected from a higher education institution (acquired from several disjoint databases) related to students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. In this capstone project, we will start by cleaning the data, exploratory data analysis, model training and bench marking, conclusions and findings.

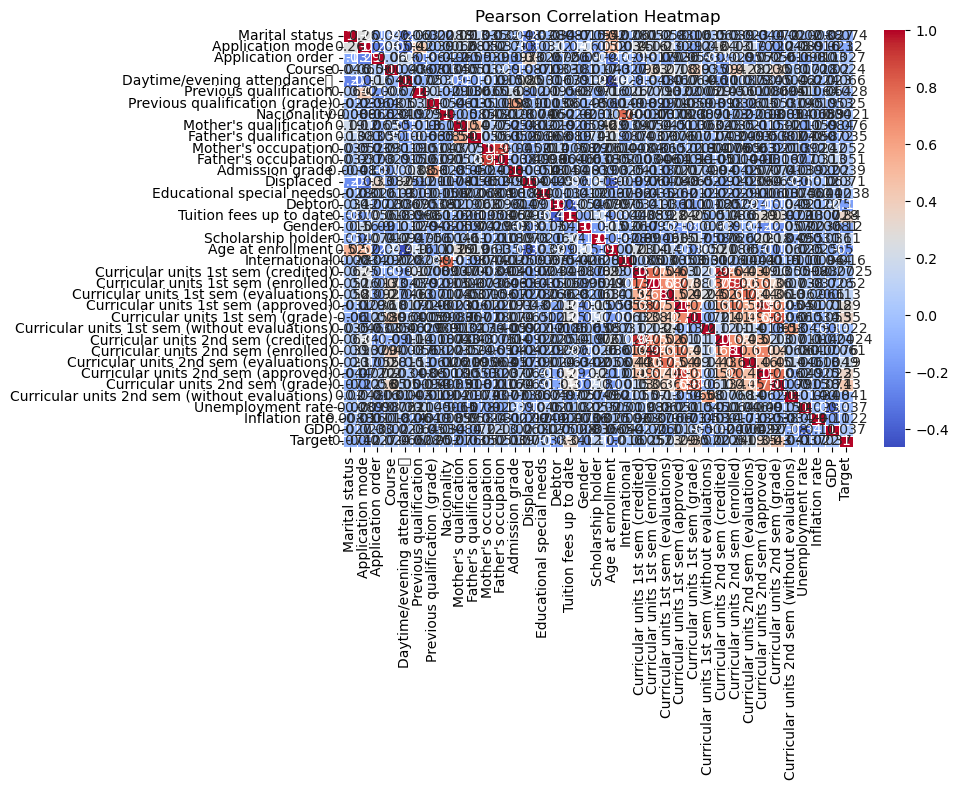
**Cleaning the data**

We start our project from a csv file. We first look at the empty value of each column and there is none. Secondly, because we want to predict the target variable, the target variable has to be numerical not text. So we developed mapping rules:Drop out to 0; Graduate to 1; Enrolled to 2.Thirdly, we started to deal with outliers. We remove outliers by setting the upper bound and lower bound by three stand deviations.

**Exploratory Data Analysis**

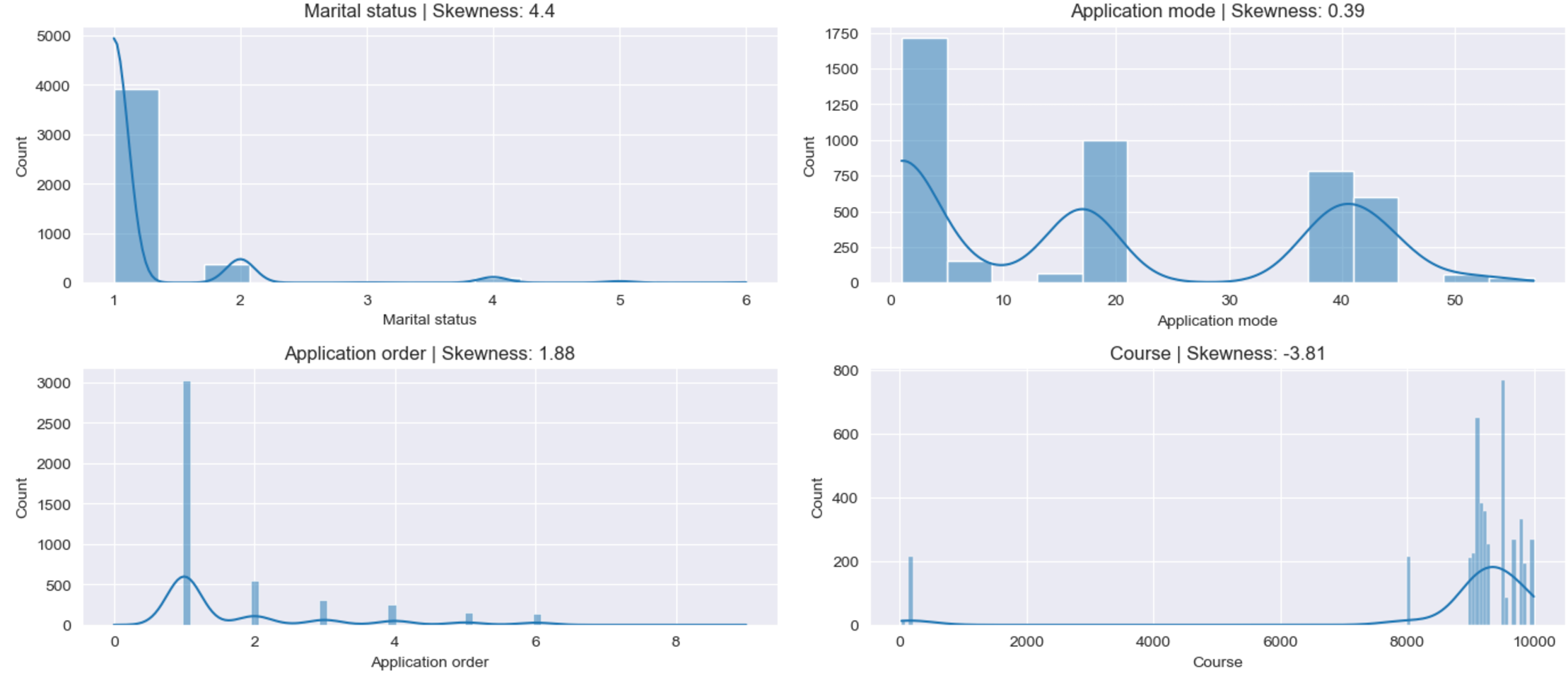
First and foremost, to demonstrate the relationship between these variables, we used Pearson Correlation Matrix as follows. This heatmap can reveal the correlation with one another without being too overwhelming or compact.

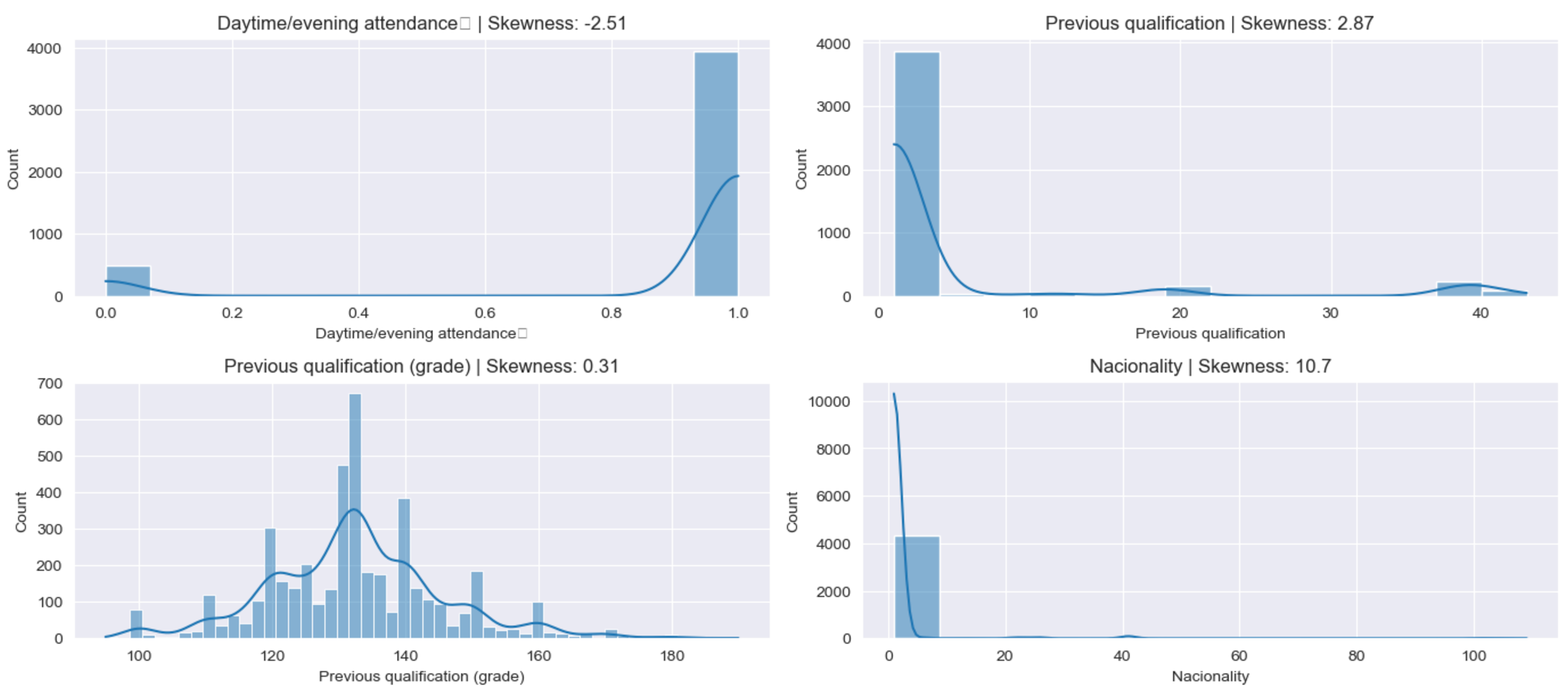
Figure 1 Pearson Correlation Heat Map



From the heat map, it’s easy to tell from the colors that -0.2 to 0.2 means little correlation, 0.2 to 0.4 means positive median correlation and 0.4 to 1 means positive strong relationship. In this case, Curricular units in 1st and 2nd term showed positive relationship with student academic success. In the plot, it’s easy to tell that there are too many dimensions in the dataset, so it’s important to use dimension reducing techniques such as PCA.

Figure 2 Density Plot





Secondly, we employed density plot to check the distribution and skewness of our variables. There are several findings. For example, most of our data are single Portuguese. What’s more, most of our data are first choice. Most of our students’ previous education is around 130, which looks like normal distribution.Most of our students are daytime participants.More findings and plots are recorded in the jupyter notebook.

Thirdly, there are more than 10 variables such as nationality, previous qualification, daytime/evening attendance in the dataset. Not all are closely related with our target variables. Therefore, we should first reduce the dimension of the data for further studies. Principal component analysis, is a useful tool to put a large dataset into a small dataset, while still maintains its key feature of the dataset. When choosing the number of principal components (k), we choose k to be the smallest value so that for example, 99% of variance, is retained. As stated in the jupyter notebook, when k equals 1, we retained nearly 99% of the variance.

**Model Development and Evaluation**

In the model development part, we identify predicting students' dropout and academic success as a classification problem. We choose two models including gradient boosting models and random forest for comparison and evaluation.First of all, we split data for training set and testing set. Training set is used to train our models and testing set for evaluation. Secondly, to evaluate model’s accuracy, we applied F1 score and confusion matrix.

F1 score usually ranges from 0 to 1, with 0 meaning the lowest possible result and 1 denoting a flawless result, meaning that the model accurately predicted each data successfully.Looking at confusion matrix for gradient boosting model, there are 87 models in testing set who are drop are successfully predicted as drop, 93 models in testing set who are drop are predicted as enrolled, 19 models models in testing set who are drop are predicted as graduate. There are there are 99 models in testing set who are graduate are predicted as drop, 196 models in testing set who are graduate are predicted as graduate, 18 models in testing set who are graduate are predicted as enrolled. There are there are 51 models in testing set who are enrolled are predicted as drop, 196 models in testing set who are enrolled are predicted as graduate, 18 models in testing set who are enrolled are predicted as enrolled.

Figure 3 confusion matrix for gradient boosting model

[[ 87 93 19]

[ 99 196 18]

[ 51 51 3]]

On the other hand, we compare random forest model to gradient boosting model. In confusion matrix for random forest model,there are 132 models in testing set who are drop are successfully predicted as drop, 0 models in testing set who are drop are predicted as enrolled, 67 models models in testing set who are drop are predicted as graduate. There are there are 157 models in testing set who are graduate are predicted as drop, 0 models in testing set who are graduate are predicted as graduate, 156 models in testing set who are graduate are predicted as enrolled. There are there are 74 models in testing set who are enrolled are predicted as drop, 0 models in testing set who are enrolled are predicted as graduate, 31 models in testing set who are enrolled are predicted as enrolled.

Figure 4 confusion matrix for random forest model

[[132 0 67]

[157 0 156]

[ 74 0 31]]

Our another model evaluation techniques if F1 score. Usually F1 score from 0.8-0.9, means a good score, F1 score from 0.5-0.8 means an OK score. In this case, both random forest and gradient boosting model may not perform very well due to imbalanced data. This is because of the limitations of F1 score,

when one class significantly outweighs the other, the regular F1 score metric might not give a true picture of the model's performance. Because of the majority class’s influence, model who achieve high accuracy in the minority class may have a low F1 score.

Figure 5 F1 score for gradient boosting model

|  |  |
| --- | --- |
| Macro F1 score | 0.34692271928859886 |
| Micro F1 score | 0.46353322528363045 |
| Weighted F1 score | 0.46353322528363045 |

Figure 6 F1 score for random forest model

|  |  |
| --- | --- |
| Macro F1 score | 0.21415094651347985 |
| Micro F1 score | 0.26418152350081037 |
| Weighted F1 score | 0.18089810661529884 |

**Conclusion**

Now that we have completed data cleaning, model training and model evaluation steps. And in the evaluation step, we compared two models performance using F1 score and confusion matrix. Our conclusion is that gradient boosting model performs better than random forest model in F1 score and confusion matrix model in predicting students’ academic success.