Capstone Project: Predict Students' Dropout and Academic Success

Yuning Xiang (Robert) Jiaxuan Cai (Josie)

CONTENTS

01 Data understanding, cleaning, and preprocessing

Problem Understanding

- 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.

Define the model

- three category classification task.
- build classification models to predict students' dropouts and academic success

Cleaning the data

Look at the empty value

We first look at the empty value of each column and there is none

Transform the categorical value

- Drop out to 0;
- Graduate to 1;
- Enrolled to 2.

Cleaning the data

Removing outliers

We remove outliers by setting the upper bound and lower bound by three stand deviations.

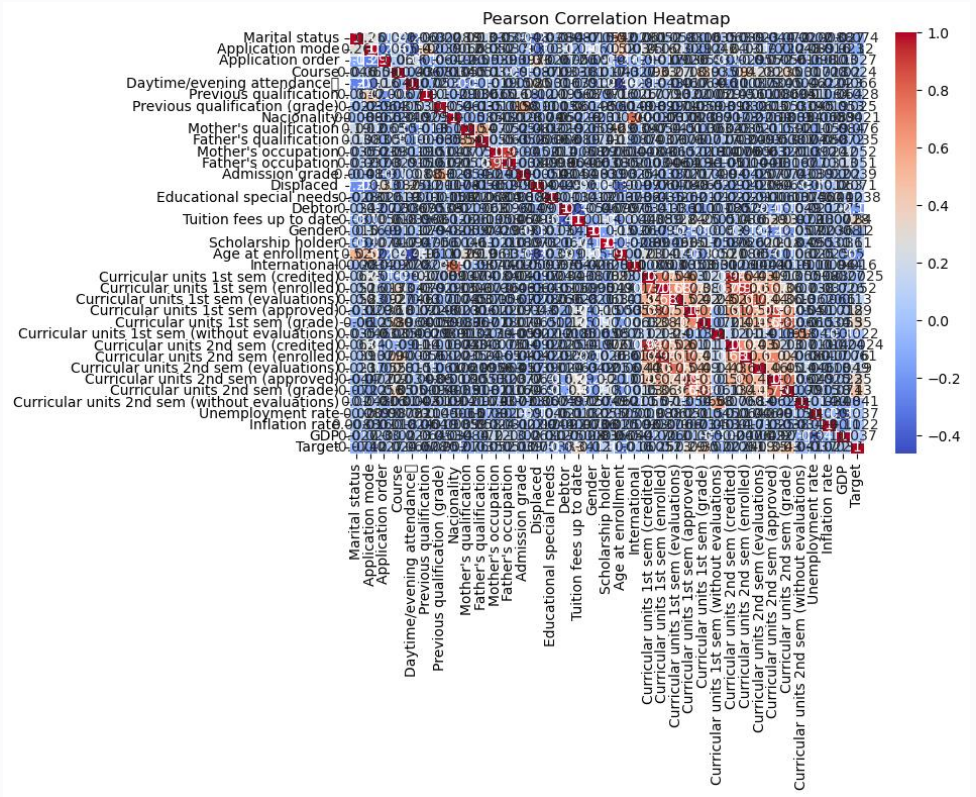
02

Exploratory Data Analysis (EDA) and initial model prototyping

Pearson Correlation Heatmap

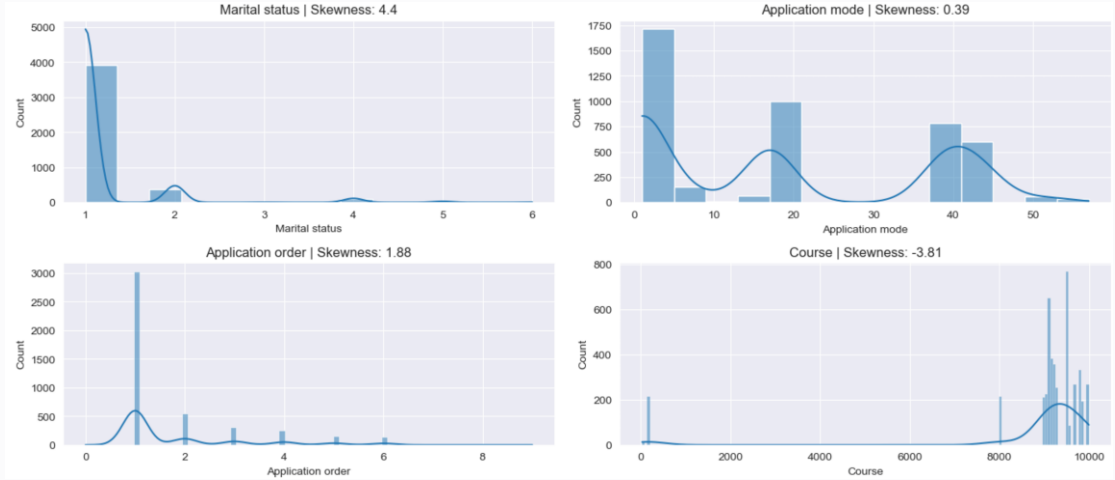
Heatmap

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.



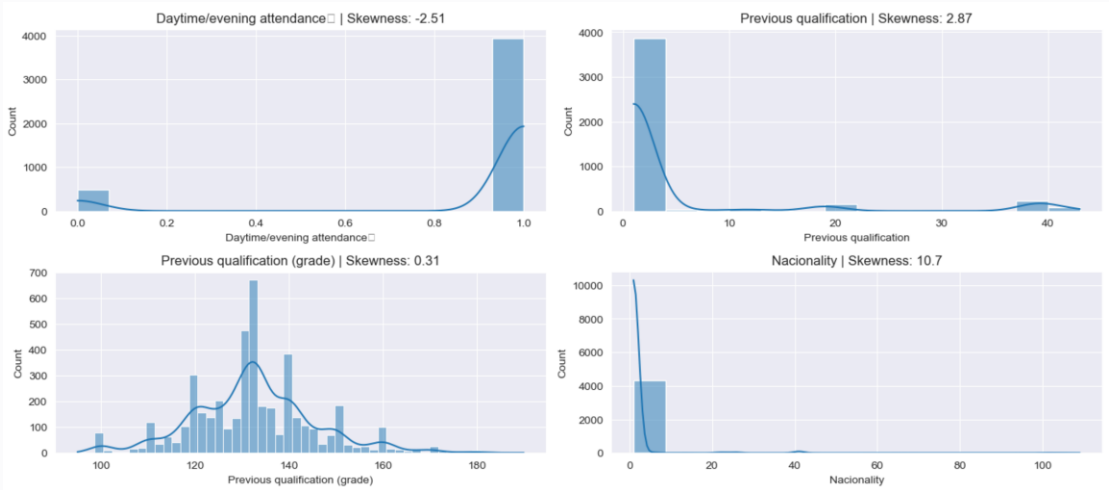
Density Plot

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.



Density Plot

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.



03

Model training, hyperparameter tuning, and addressing class imbalance.

PCA

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.

04

Final model evaluation, report preparation, and presentation.

confusion matrix for gradient boosting model

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.

```
[[ 87  93  19]
 [ 99 196  18]
 [ 51  51   3]]
```

confusion matrix for random forest 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.

```
[[132   0  67]

[157   0 156]

[ 74   0  31]]
```

F1 score for gradient boosting model

Macro F1 score	0.34692271928859886
Micro F1 score	0.46353322528363045
Weighted F1 score	0.46353322528363045

F1 score for random forest model

Macro F1 score	0.21415094651347985
Micro F1 score	0.2641815235008103 7
Weighted F1 score	0.18089810661529884

Comparision

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.

Thanks

单击此处添加副标题内容