

Machine Learning Case Study

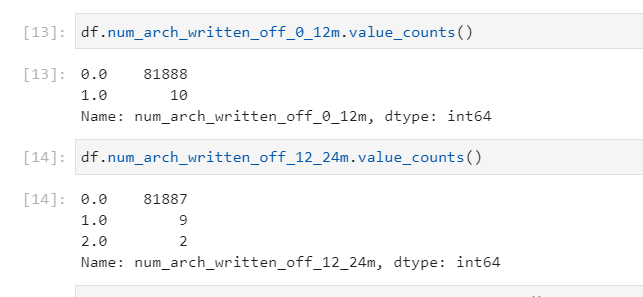
Reported by Eser İnan Arslan

**DataSet Preperation**

In this case, there is a dataset which has supervised data variance. At the end of the Project, we aim to classify default values as a probability rate of (default==1).

Our dataset has 99976 rows on 43 columns. Target variable (default) has 10 percent “Nan” value and 1 percent “1” value. There is almost %90 zero values. Not only default value, but also lots of columns have unbalanced data.

Some columns consisted of even more than 99.9% 0(zero) values.

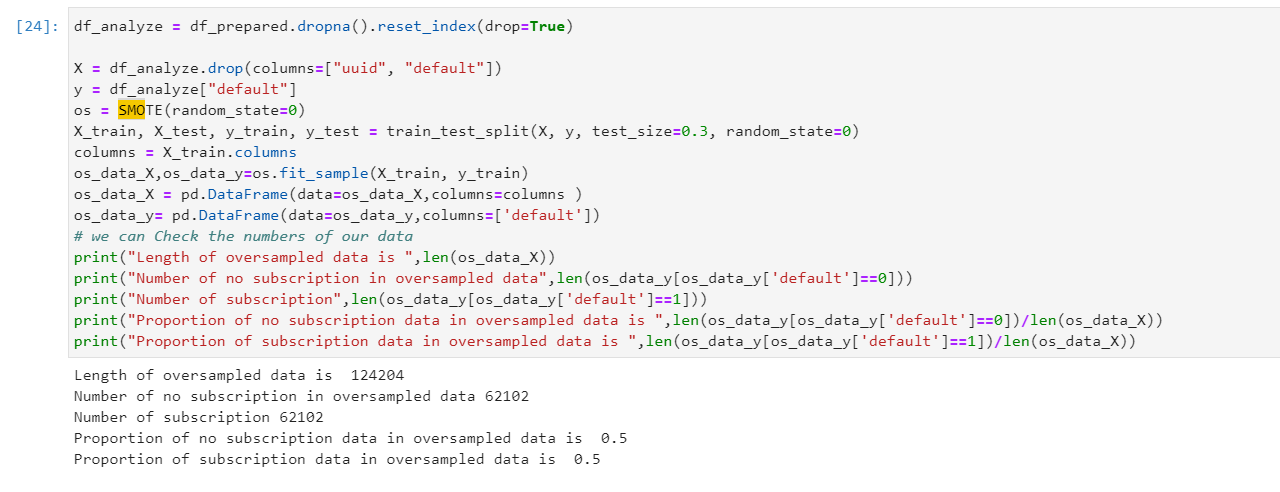


Lots of columns have no correlation with the “default” parameters, just only nine of them have upper than 0.1 correlation score(both negative and positive relation)

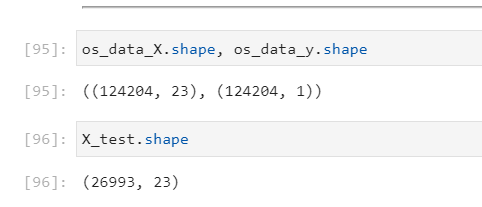


Besides these, some columns which has more correlation rate with each other are eliminated like previous ones. I chose only one of them.

Generally I can say that our dataset is unbalanced so I need to make it more balanced with over-sampling by the help of SMOTE method from imblearn. You can see results below:



Our final train and test dataset became like this:



**Model Implementation**

On model development part, at the beginning I prefer 3 different algorithm:

Random Forest,

Naïve Bias

SVM

Naïve bias has two different versions; isotonic and sigmoid. Both random forest and Naive bias have also two version, normal model creation and calibrated ones. In both types, I prefer to GridSearchCV for hyperparameter tuning and I used best parameters according to our dataset. They are also good performed and their results are acceptable, but we cannot say same thing to SVM. Its process took to much times, so I implemented it (again both normal and calibrated versions) but they didn’t finish their process, so I killed them. To store the results, I prefer SQLite. I f my data is so big, I wouldn’t prefer this database because generally it has performance problems. But in this time, my data isn’t so huge so I chose it because it doesn’t need any installation.

Here are the results:

