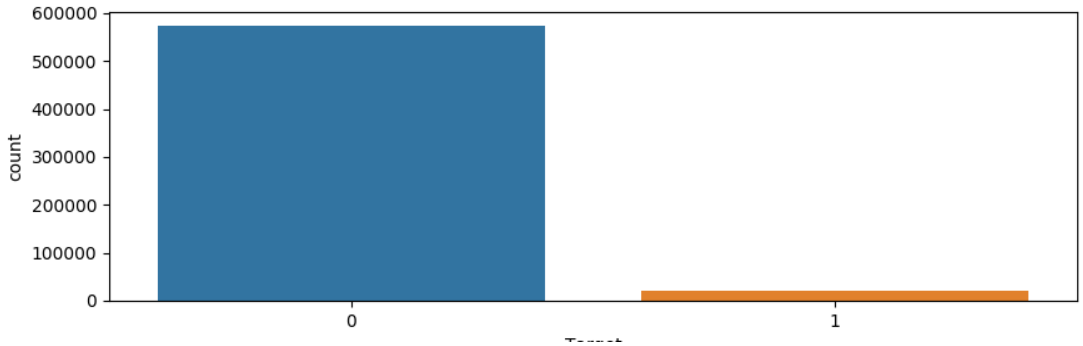
Porto Seguro's Safe Driver Prediction

MMNRie team : Camille Chanial, Benoit Guillard, Arnaud Stiegler, Paul Vardon

December 22, 2017

1 Data preprocessing

1.1 Target repartition

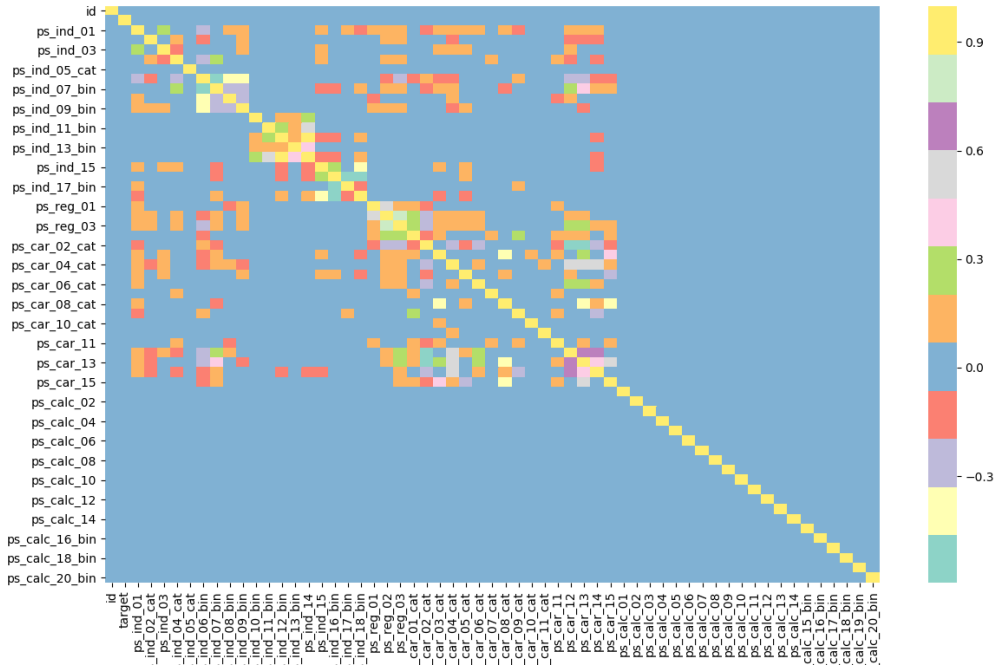


The target variable is unbalanced.

1.2 Missing values

We handled missing values by replacing them by the mean or the mode of their columns.

1.3 Correlation matrix



We can observe that all ps_calc_* variables are not correlated with the rest of the variables. Thus, it is safe to remove them from the data set.

1.4 Categorical variables

We turned the categorical variables into dummy variables.

2 Model tuning

We trained three different models :

2.1 XGBoost

We trained an XGBoost algorithm. This algorithm is based on the (extreme) gradient boosting framework. Very popular in machine learnin competition, it provided us with an effective result, we then used in our final result.

2.2 Random forests

We trained a Random Forest algorithm, with parameters chosen through a Grid-Search algorithm. This algorithm is an ensemble learning method which constructs several decision tree in order to separate the data. Random forests aim to correct the tendency decision trees have to overfit.

2.3 Logistic regression

This Logistic regression model was tuned with the following parameters, run into a Grid-Search algorithm:

```
params = {
'C': [0.001, 0.003, 0.005, 0.01, 0.03, 0.05, 0.1, 0.3, 0.5, 1]
        }
```

The best parameters were:

```
params = {
'C': 0.005
        }
```

with a roc_auc score of: 0.6347.

2.4 Final result

We finally average the results provided by the 3 models. We obtained a Gini score of 0.28672 that ranked us 498 out of 5169 teams (top 10%).