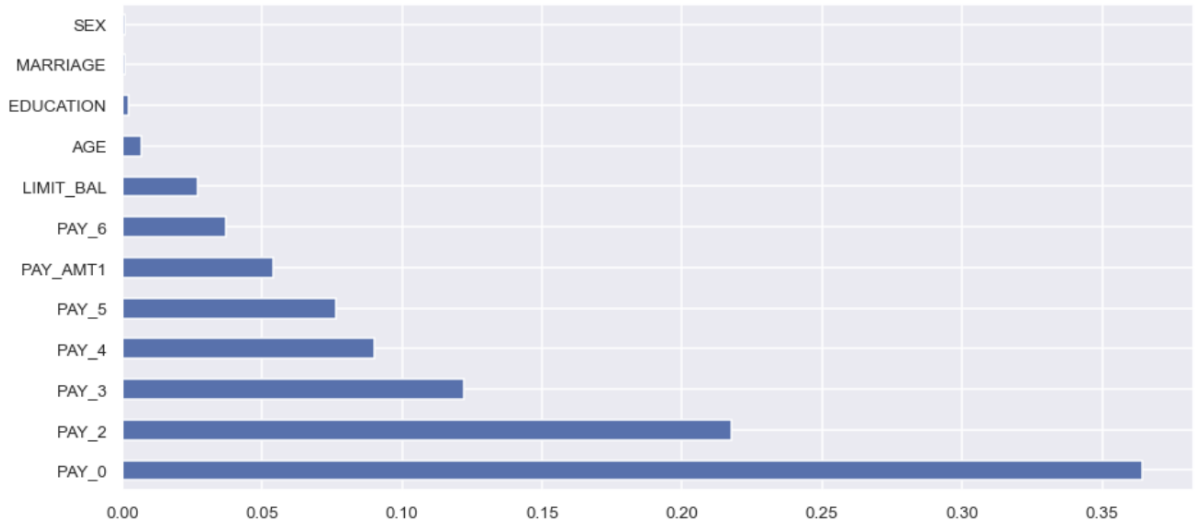
Building the model

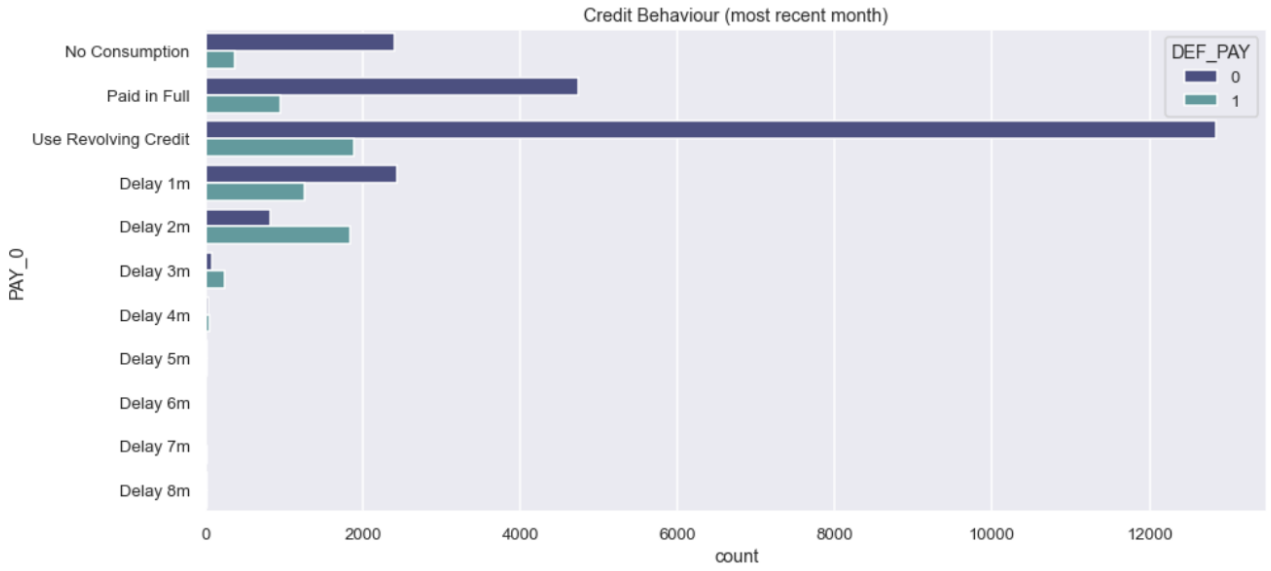
By conducting various machine learning models on the dataset, we trained several regressors that can be used for predictive analysis. Our next step is to compare the performance of different models and pick top ones to be used under business environment.



Through the Train Test Split, all the independent variables were tested, but at the end just the features below were chosen as the most important. The categorical variables SEX, MARRIAGE, and EDUCATION have not a significant relation to explain the behavior of the defaulted payments of loans.

```
Decision Tree Regressor -0.6123300344575618
Random Forest Regressor 0.13865122481251127
Linear Regression 0.11555957026486725
Support Vector Regression -0.08550379820236809
```

Through Cross-validation, the skill of the machine learning models were estimated. The model with the best score to help predicting, what type of customer will default or not the payment of the loan, was the Random Forest Regressor model; even that, the score is under 14%, which means that our model has not a quiet significant accuracy to make predictions. That can also be complemented with the result of 0.144 obtained of the R Squared of the Random Forest, which means that the model does not explain any of the variation in the response variable.



Also, detecting potential late payments for the credit account is a special task that we should focus more on how successfully the model can predict to recognize clients with late payment potential.

To conclude, the three regression models used can not give us the accurate enough data to help us predict which customers will default on their loan's payment.