**Course 2 Task 3**

Cleaning and Pre-processing:

This dataset was pretty clean. It didn’t have missing values.

One-Hot Encoding:

We did need to do one hot encoding. We use the “pd.get\_dummies” command so it would delete the existing columns. The columns that needed this were sex, education, and default payment next month.

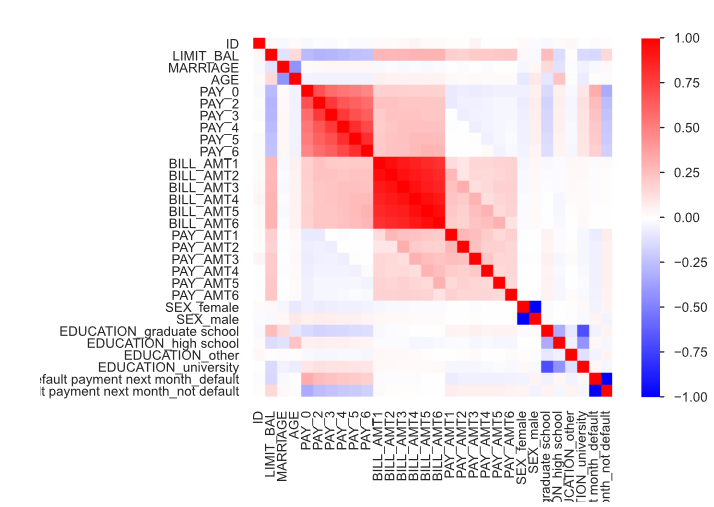
Models:

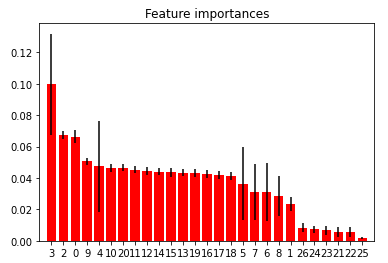
We used three models, the Random Forest Regressor, Linear Regression, and the Support Vector Regression. From these, the best one was Random Forest Regressor.



Attributes statistically significant:

Based on the following matrix and feature importances, we can tell that the factors most statistically significant are LIMIT BALANCE, AGE, and PAY\_0.





(this set for the feature importance removed the ID column. It started on Limit\_Bal)

Conclusion:

At the moment, I was not able to determine whether a customer would default accurately with this information.