The goal of this puzzle was to predict the probability of default on labeled data. This is a supervised probabilistic classification problem.

The attached Notebook file includes the code alongside step-by-step documentation. The overall approach was started by the cleaning and exploring of the data. The models tested were Logistic Regression, Random Forest and XGBoost.

The step of data cleaning consisted of identifying missing values and discarding features with no variation and categorical variables with high cardinality, although some methods to reduce the dimension could be implemented (perhaps a cluster analysis).

In the exploratory analysis, the correlation matrix of the predictors was made showing a high correlation between two features, one being removed due to multicollinearity problem. The distribution of the variables was also shown for visualization of needs of transformation and potential outliers.

The training set was split for cross-validation and an oversampling was made to correct the imbalance of the target variable so the model wouldn't be biased by the majority class.

The algorithms chosen to be tested were the Logistic Regression and the Random Forest, as the first is a classic probability model and the latter is known for being a powerful ensemble method, using the voting system as the probability. The evaluation metrics used were the AUC and the Log Loss Score.

These models however, depends on variables having too many missing values in the test dataset. To overcome this, a XGBoost was modeled as it can handle missing data, and therefore is able to predict for every observation in the test dataset.

Although the final score isn't very great, I tried my best on getting both classes predicted, finding a tradeoff of sensitivity and specificity.

The expected performance score is of the metric given, of an AUC of 0.71.