Data Science Assignment

This case consists of a supervised learning example, similar to what we will be working on at Navi.

Your task is to predict the probability of default for the data points where the default variable is not set (label = 'oot'). Please filter the data by label = 'modeling' and use it to build your machine learning model.

The answer should contain the resulting predictions in a csv file with two columns, decisionID and predicted PD (probability of default==1).

The model should be developed in high-level programming languages like Python, Spark Java etc. (We don't accept any solutions developed in statistical languages like R and SAS).

Besides the actual results, the following aspects are important:

- Code quality (e.g. Use pre-processing pipelines, etc.)
- Feature Engineering and selection (Variable creation, filling missing values, etc.)
- Algorithm selection (Please explain the reason behind choosing a particular algorithm, selection of cost function and evaluation metrics)
- Model validation (metrics used for validating the model, etc)

PS: Please don't spend too much time on the prediction results, we evaluate the overall solution.

All the best.

Dataset

The data is located in the attached file dataset.csv. This is a simple semicolon-separated CSV file containing a unique id, the target variable default, label of data (modelling and oot) and a number of features with different data types and meanings. Some variables might have missing values for some rows and its upto the candidate to determine the usage of such variables in the model. Please refer to data_dictonary.csv file for the feature variables and their descriptions.

Need to submit the below as a part of the solution:

- One page write-up explaining the solution and thought process.
- Neatly documented code (preferably notebooks)
- Predictions(Follow template.csv for the submission)