Credit Card defaults report

**Validation and training set**: At the very beginning, the puzzle\_training\_data is divided into two parts: out\_of\_sample\_data (15%) and new\_training\_dataset (85%). While dividing the data, class imbalance is also kept in mind such that both datasets maintain the same frequency. The out\_of\_sample is kept aside for validating the ML model’s efficiency.

**Data Cleaning**: The main purpose of data cleaning module was to fill in missing values and to remove outliers in the data set. Rows were dropped if the default value was NULL or if a features had very less NULL values (less than 2%). Furthermore NULL in nominal categorical columns like Gender and Fb\_profile were filled with “unknown” and then converted to dummy value columns. Additionally, columns with high values of non-numeric data were dropped. (Eg: Zip, reason, Job\_name,state). It was also observed that channel feature had only one value and hence was dropped. It was also observed that n\_isssues and n\_accounts had a very high correlation (0.99) and hence n\_issues was dropped as it also contained many NULL values. Further all the non-numeric columns were mapped to numbers and the mapping function was pickled for future use: out\_of\_sample and testing data. In this way all the data contained no NULL values and was completely numeric. Furthermore data points away 2 std deviation were corrected to the value at 2 std dev. In this way outliers (only in continuous column) were corrected as dropping them would have caused huge data loss. An alternative method for correcting outliers by creating discrete bins was also experimented, which gave similar results in training phase as this current method but further increased complexity.

**Data standardization**: Since all the features in the data set were in a different scale, it became necessary to bring them on a similar scale. Standardization converted all the continuous features to assume a Gaussian distribution with mean =0 and std=1. Categorical features were rescaled to contain values between 0 and 1. Standardization mapper was pickled for the future use of testing data.

**Visualization of Features**: Various plotting techniques were used to understand the data and their inter-relationship. Histogram plots were used to understand the nature of their distribution. Box plots were used to see how the data varied from the interquartile range to detect outliers and correlation matrix was used to see how related the features were to the “default” as well as amongst each other. It was observed that income and amount\_borrowed had a slight positive correlation.

**Feature importance and selection**: Statistical scoring techniques which only consider the relationship of the feature with the target feature were used. Chi\_squared score and Mutual Information score were used to rank the features. Deliberately, 5 new random features which were synthesised randomly (having a Gaussian distribution) were introduced. The features which were ranked higher than the random variables by both techniques were selected as final features for our ML model. PCA was also used to calculate the variance score of the features transformed into the Eigen space, but since the feature space was non-linear it failed to gather enough variance across orthogonal vectors.

**Comparison and selection of Machine Learning model**: Since the data was unbalanced, roc\_auc was used as a scoring metric for the ML models. The data was divided into k-folds (stratified to take care of unbalanced dataset) and various models were tested on the same. A box plot of all the model metric scores was used to compare which model to choose. Once a model was finalized, it’s hyper parameters are tuned using grid search. The model was then saved for later use on validation and unseen data.

**Machine Learning model score**: GradientAdaboost had the highest in-sample roc\_auc score of 0.71 but the roc\_auc score for out-of-sample data was 0.51. Hence multiple other ML models were experimented in a similar fashion. It was found that Naïve Bayes had the highest out of sample roc\_auc of 0.58. The confusion matrix clearly showed that Naïve Bayes was able to predict True Negatives (when the model predicts default and someone actually defaults) of out-of-sample data much better than any other model. The out of sample as well as the puzzle\_testing\_data was pre-processed in the same way as the training data. Finally, the trained Naïve Bayes ML model was used to predict the default target for puzzle\_testing\_data which was stored as predictions.csv file in reports folder.