Executive Summary

When transforming categorical data into numerical data, we used LableEncoder, OrdinalEncoder and dummy variable. LableEncoder was used to convert target variable ‘Credit Rating’, and OrdinalEncoder was used to convert features that are ordinal data, such as 'Requested credit amount', 'Number of Dependents', 'Monthly income' and 'Monthly expense', since these data were represented in a scale. "Marital\_status" was transformed using dummy encoding, which added two more columns to the data set. Hence, we now have 8081 rows and 9 columns.

Standardization was performed to give every feature the same sensitivity regarding the variances of the initial variables to prevent biased results due to a different range of data.

Next, we used PCA to simplify the complexity of the potential model. The first step is calculating variance explained ratio matrix to determine the number of principal components needed. And then, we used five principal components to represent the dataset with 85% variance. Hence, we used PC value to represent the original data set in later analysis.

The dataset was splatted into 70% training data and 30% test data. MLP was used to fit the dataset. We have tested a couple of hyperparameters manually, and k-fold cross-validation scores were computed. It turns out that each layer with 3 nodes or 2 nodes have approximately same cv score. Therefore, the best hyperparameter was (2,2,2,2,2) – five hidden layers where all layers have two nodes, the following analysis will be based on MLP(2,2,2,2,2) model.

To find the threshold of approving an application or not, we have set a threshold list which has the range from zero to one with step size 0.01.We aimed to use the f1 score regarding different threshold to find the best threshold, and we would use the peak point on f1-score plot as the threshold. However, the plot shows same f1 score from threshold 0 to 0.81. Since we are building a model to predict the probability of paying back the loan, we want to set a relatively high threshold to minimize risk. We set the threshold for request high amount to 0.7, since when people requesting high amount loan, they are likely to use high value personal property as mortgage, which means even they can not pay back with cash, the bank can still sale their mortgage. Similarly, if their request amount is low, they will not need to mortgage anything. If they cannot pay back, the bank cannot receive anything back. So we set the threshold for low requesting to 0.9.

As a result, most application get approved.

Limitations: Due to time constraint, we only used f1 score to find the threshold, f1 score gives us weighted average of the true positive rate and precision. However, a confusion matrix would provide better result because we aim to minimize false positive. In confusion matrix, we can find both precision and recall.