Executive Summary

With the technology nowadays, it becomes easier for banks to collect the information of its customers and make good of it. One of the things that banks can do to avoid losing big is to build a comprehensive credit risk management process. The objective of this project is to help making the process of managing credit risk easier for banks. I utilized several machine learning algorithms in Python and trained it on the dataset which contains 1,000 data points and is accessible on UCI[1]. To build a model which can predict the good/bad credit consumers will be my final goal. The 21 features from this dataset are listed below.

- Over Draft: <0, 0<=X<200, >=200, no checking account
- Credit Usage: Duration in month
- Credit History: No credits/all paid, All paid, Existing paid, Delayed previously, Critical/other existing credit
- Purpose: New car, Used car, Furniture/Equipment, Radio/TV, Domestic appliance, Repairs, Education, Vacation, Retraining, Business, Other
- Current Balance: Credit amount
- Average Credit Balance: <100, 100<=X<500, 500<=X<1000, >=1000, no known savings
- Employment: unemployed, <1 year, 1 year<=X<4 year, 4 year<=X<7 year, >=7 year
- Location: 1 to 4
- Personal Status: male div/sep, female div/dep/mar, male single, male mar/wid, female single
- Other Parties: none, co applicant, guarantor
- Residence Since: 1 to 4
- Property Magnitude: real estate, life insurance, car, no known property
- cc_age: cc_age in month
- Other Payment Plans: bank, stores, none
- Housing: rent, own, for free
- Existing Credits: 1 to 4

[1] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

- Job: unemp/unskilled non res, unskilled resident, skilled, high qualif/self emp/mgmt.
- Number of Dependents: 1 or 2
- Own Telephone: Yes or No
- Foreign Worker: Yes or No
- Class: Good or Bad

Since the dataset contains numerical and categorical features, I will first split the whole dataset into test set and train set and then normalize it for analytical purpose. And also, because the number of good and bad of the target variable is quite imbalanced (7:3), so I will further utilize SMOTE to perform my oversampling process. The purpose of these processes above is to get my dataset ready for building the machine learning model.

The scores of every algorithm that I tried on my train set to get my best model are listed below.

Algorithm	Accuracy Score	Precision Score	Recall Score
k-nearest neighbors	0.7939	0.7729	0.9102
Logistic Regression	0.7378	0.7128	0.8020
Decision Tree Classifier	0.7469	0.7309	0.8939
Random Forest Classifier	0.7684	0.7501	0.8143
Extra Trees Classifier	0.7582	0.7324	0.8204
Gaussian Naïve Bayes	0.7296	0.6936	0.8204
XGboost	0.7929	0.7865	0.8122

Because in this particular case, I wanted to set the probability of prediction of bad credit to be higher than the good ones, I will use recall score to evaluate all my models. After taking the scores above into consideration, I decided to use Random Forest to build my model while the recall score I got on the test set is lower than I expect. The reason could be on the oversampling process. This process might be able to solve the imbalanced issue but might also result in an overfitting model. One way to have a better model than the current best model is to add more data to lower the high variance in the current model.