Predictive Analytics of Bank Customers Credit: A Case Study of Germany

Literature Review, Data Description, and Approach

Professor: Dr Derya Kici

Student: Mohsen Selseleh

Student Number: 500726502

Date of Submission: Dec 29, 2021

Table of Contents

Abstract

Bank credit means amount of money that a person could borrow from a bank or other financial institution. The bank decides on basis borrower’s credit rating, income, assets, debts, etc. This process for banks always has risk of failure, when borrower neglects to pay the money obtained. Therefore, credit scoring is a challenging management issue. Nowadays, banks use the data mining techniques to build trained model by learning samples and trained model is used to make decision in new situations. In this study, many factors determine credit of a customer as status of checking account, credit history, credit amount, saving account, employment, debts, age, ownership a house, number of people provide maintenance, etc. This study is a typical classification problem which determine a customer is good or bad for a loan. The research will use a supervised learning (neural network,svm, logistic regression analysis, knn, and naïve bayes) and ensemble algorithms (gradient boosting , random forest, decision tree). The dataset has historical data for 1000 customers of bank in Germany. I will try to answer these questions:

Traditional models or ensemble models are better at prediction customer credit?

Which model has a better performance at prediction of credit customer?

Is there significant relation between independent variables?

Introduction

Customer credit is an ever-growing threat in the finance sector. The financial institutes need to know about credit of their customers. This kind of loan has labeled as customer credits which means customer's ability to repay his/her loan. The credit allows the customers to buy goods and services without need to pay for them, immediately. The most evident example of customer credit is credit cards. Bank credit means amount of money that a person could borrow from a bank or other financial institutes. This credit has a profitability risk, especially sometimes customers could not afford to repay their debts. Because the fund is limited, the banks as the most important financial institutes, must assigned their funds carefully to gain more profit.

The main part of the study of credit risk is finding good and bad customers which is an imperative task for the banks. Nowadays, machine learning algorithms make a great improvement at these kinds of studies because dataset is big data. The Customer credit dataset is highly imbalanced because there will be more good customer credit when compared to bad customer credit.

The detection credit customer is so critical for preventing risk in the banking sector. Because failing to repay has disadvantages for bank and customer, too. The customer should pay additional interest charges if they could not repay on time. The credit risk is one the most important hazards that threat banks and they try to with risk management knowledge and validation models study this risk and change its threat to opportunity.

As transactions become the frequent mode of payment, customer credit becomes more important, Finding the bad customer credit using traditional methods is time consuming and incorrect. Nowadays, financial institutes use machine learning techniques to study customer credit problem.

The research about Customer credit is dependent on sampling approach, selection of variables and machine learning methods. This study uses under-sampling method with naïve bayes, k-nearest neighbor, neural network, decision trees, random forest, logistic regression, and Gradient Boosting on highly imbalanced customer credit dataset.

Literature review

Ng et al (2002) compared logistic regression and naïve bayes on 15 experiments on datasets from the UCI Machine Learning repository. They show that the discriminative logistic regression algorithm has a lower asymptotic error than the generative naive bayes classifier, but it converges more quickly to higher asymptotic error. When, the number of training examples is increased, the generative naive Bayes do better, but the discriminative logistic regression overtakes than the performance of naive bayes. They find that there are a few times in which logistic regression's performance was worse than the naive bayes, but it is observed in small datasets.

Maes et al (2002) have studied credit card fraud detection by using machine learning techniques; artificial neural networks and bayesian neural networks to the problem. They presented that ANN and BNN have good results in fraud prediction, although BNN had better results for fraud detection and its training period is shorter, but ANN was more quickly than BNN.

Bhattacharvva et al (2011) compared two advanced data mining approaches, support vector machines and random forests with the logistic regression, as part for detecting fraud on credit card transactions. They use data under sampling and study the performance of the three techniques with varying levels of data under sampling. For performance assessment, they use a test dataset with much lower fraud rate (0.5%) than in the training datasets with different levels of under sampling. This method provides an indication of performance that may be expected when models are applied for fraud detection where the number of fraudulent transactions is low. All techniques showed good ability to detect fraud in the data set. Performance with different levels of under sampling was different by methods and on different performance measures. The sensitivity, G-mean and weighted-accuracy decreased with lower proportions of fraud in the training data, precision and specificity were increased. The random forest and SVM showed a decreasing trend on AUC and an increasing trend on F while logistic regression maintained the same performance on the F-measure and AUC. They show random forests showed much higher performance than other methods. The logistic regression had the same performance with different levels of under sampling, while SVM performance at the upper file depths increased with lower proportion of fraud in the training data. They presented that logistic regression is a standard technique in data mining applications better than the SVM models. They show that the performance logistic regression depends on variable selection and exploratory data analysis. they used the same derived attributes for the comparison the models. They argued that random forests and SVM carry natural variable selection ability and have been noted to perform well with high dimensional data.

Alborzi et al (2016) studies credit customers with artificial neural networks and they show that ANN has good accuracy in prediction customer credit. They used a new hybrid model of behavioural scoring and credit scoring by using the data mining and neural networks techniques to study credit customers. They applied clustering and classification techniques. They presented that the model could effectively segment and classify bank customers.

Lee et al (2006) studied credit customer risks classification and regression tree (CART) and multivariate adaptive regression splines (MARS). Their results demonstrate that cart and mars both have better average correct classification rate in comparison with discriminant analysis, logistic regression, neural networks, and support vector machine. They argued modeling techniques like traditional statistical analyses and artificial intelligence techniques have been used to study the credit scoring tasks. They show that discriminant analysis and logistic regression are the used a lot for customer, but its drawback is strong assumption. They show that the artificial neural networks approach is a very famous alternative in credit scoring tasks because of its memory characteristic, generalization capability, and outstanding credit scoring capability, its drawback is inability to identify the relative importance of potential input variables, and certain interpretative difficulties. They study the performance of credit scoring using with classification and regression tree (CART) and multivariate adaptive regression splines (MARS), since cart and mars can successfully solve credit scoring problems without those drawbacks of discriminant analysis, logistic regression, and neural networks. They used their model on one bank

Shen et al (2007) study application of classification models on credit card fraud detection. They study three classification methods to analysis of the credit cards history business information and have made the fraud detecting models. They present the of the data mining techniques including neural networks, logistic regression, and decision tree to the credit card fraud detection, for the goal of decreasing the bank's risk. They show that neural networks model provides higher accuracy than a logistic regression and decision tree on the same data, while neural networks slightly better than logistic regression.

Chen et al (2007) studied mining the customer credit using hybrid support vector machine technique. They show that the statistical classification models need strong essential assumptions. They argue that the artificial intelligence techniques do not require the knowledge of the underlying relationships between input and output variables. SVM is a modern data mining technique and suitable for classification and regression problem.

They built credit scoring models combined SVM and CART, MARS. They used SVM for classification and regression, they choose kernels to determine the kernel's parameters.

They studied the performance of credit scoring with hybrid modeling procedure , integrating the svm approach with cart, mars technique on one credit card dataset provided by a local bank in China is used in this study. They show the hybrid SVM technique has the best classification rate, and it has the lowest TypeII error in comparison with CART, MARS and SVM and they show that SVM having better capability of capturing nonlinear relationship among variables. They resulted that the hybrid SVM credit scoring approach which they use in this study will have higher credit scoring accuracy and lower TypeII error.

Li et al (2010) predict customer credit card segmentation, they used logistic Regression, decision trees, Random Forest, neural network, and support vector machine. their results show that the tenfold cross-validation method on the synthetic minority oversampling technique (SMOTE) data with neural network method has produced excellent results. They built a fraud transaction detection model, as a binary classifier for imbalanced data set. Their data set that had only a small number of fraud transactions set that could lead to a big variance of error, they use the under-sampling method smote to balance the data set. They show the average results of F1 and area under curve (AUC). First, the performance of every neural network model is better than random forest because of the nonlinear characteristic of a neural network.

Sahin et al (2011) study credit card fraud. They argue the advantages of applying the data mining techniques including artificial neural network and logistic regression to the credit card fraud detection problem. Their results show that the proposed ANN classifiers outperform LR classifiers in solving the problem under investigation.

However, as the distribution of the training data sets become more biased, the performance of all models decreases in catching the fraudulent transactions.

Sherly et al (2012) study decision tree, neural network and naïve bayes classifiers for credit card fraud detection system. They show that decision trees with BOAT algorithm is powerful tool for classification and prediction. They use an algorithm BOAT for constructing decision tree incrementally for detecting the credit card fraud where the data set changes dynamically. They could predict fraud with their model with high accuracy.

Patil et al (2013) use predictive modelling for cedit card detection using data analytics. They use a meta-classification method which consists of tree, naïve bayesian and k-nearest neighbor algorithms, results show performance improves. They show that random forest decision tree performs best in terms of accuracy, precision, and recall. The only drawback with random forest is overfitting of tree in memory as data increases.

Afsar et al (2014) studied Customer credit for granting facilities. The purpose of their study was ranking the customer groups and specifying the best part of them. They use the neural network and change the customers into 10 clusters. Using the proposed model, the clusters were ranked. The top clusters were identified, and facilities grant operations were done to the members of these clusters. Fahmi et al (2016) compare naïve bayes, k-nearest neighbor and logistic regression techniques on accuracy, sensitivity, specificity and MCC metrics in the banking field. The researchers show that LR has better performance. They argue that it could see that neural network model is used more than others which is time consuming method and choosing its potential input variable is so critical

The Goal of research

My goal is to build models to classify and predict customer credit for my dataset. Meanwhile, I would like to compare performance; neural network, logistic regression and decision tree, naïve Bayes, knn, random forest and xgboost algorithms with different metrics to find the best model for prediction customer credit in the dataset.

Method of Research

This research is a supervised machine learning which there are pre-defined set(labeled) of "training examples", it facilitates its ability to receive to an accurate conclusion for new data set.

First, I extract data set from UCI website, transform and extract and load it to python software. I preprocessed the data set in a few stages as; data preparation, handling missing data, data visualization, handling outliers, correlations and heatmap, handling multicollinearity (drop two independent variables; Duration in month, Age in years that had high correlation with other variables), handling imbalanced data set with under sampling method (randomly throwing out samples from the majority class until the class fractions are equal, or at least less imbalanced). Finally, I did data analysis with different classification models with cross-validation method to increase accuracy for prediction customer credit.

A classification model is used to predict the outcome of a given sample when the output variable is in the form of categories as sick or healthy. Supervised learning algorithms use labeled training data to learn the mapping function that turns input variables(x) into the output variable(y). This allows us to accurately generate outputs when given new inputs. I tried to give definition of some of them I used, below:

The logistic regression is a type of generalized linear models, it is a function of independent variables for the prediction the probability of a binary (nominal or ordinal) variable that the probability value changes between 0 and 1.

The logistic regression estimates the probability of a binary response based on one or more variables. It finds the best-fit parameters to a nonlinear function called sigmoid.

The decision tree is a method that finds knowledge in big dataset. The decision tree is one of the most used techniques in classification and prediction. The disadvantage is that a little change in the sample may cause big difference in classifying it.

The random forest is an ensemble of decision trees, generally trained via the bagging method, it is more convenient and optimized for decision trees. The random forest searches for the best feature among a random subset of features (Geron,2019)

The naïve bayes is a classification method based on bayes' theorem. It assumes that the features are independent. It chooses the decision based on the highest probability.

The k-nearest neighbor perform classification based on a similarity measure, like Euclidean, manhatan (both for continuous variables) or mminkowski distance (categorical variables) functions (Geron,2019).

The ensemble methods are another type of supervised learning. It means combining the predictions of multiple machine learning models that are individually weak to produce a more accurate prediction on a new sample. I used xgboosting, decision trees and random forest as ensemble methods at my research.

Burkov (2019) stated that gradient boosting is one the most powerful ensemble machine learning algorithms which could manage big dataset, easily and quickly. Also, it creates very accurate models. The important hyperparameters of it are the number of trees and the depth of trees. The most machine learning algorithms have their limitations, but ensemble learning approach could boost the performance prediction. Two principal ensemble learning methods are boosting and bagging.

Machine learning metrics

In my research, I use the following formulae to evaluate, accuracy and precision. The Mathews correlation coefficients (MCC) is a machine learning measure which is used to check the balance of the binary (two class) classifiers. It considers all the true and false values that is why it is generally regarded as a balanced measure can be used even if there are different classes.

$$\begin{aligned}
\text{accuracy} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} y \\
\text{precision} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \\
\text{recall} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\
F_\beta &= \frac{(1+\beta^2)\times \text{recall} \times \text{precision}}{\text{recall} + \beta^2 \times \text{precision}} \\
F_1 &= \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} .
\end{aligned}$$

Dataset

The dataset of customer credit is from a German bank containing 1000 items. The dataset will be balanced with under-sampling technique. The dataset sourced from UCI machine learning repository; its link is found at appendix. This dataset classifies people described by a set of attributes as good or bad credit risks by a German bank in 1994, it consists of 1000 customers. The data set is highly imbalanced and skewed. We have 21 attributes; twenty attributes are independent variable, and we have one dependent variable (cost matrix). Attributes of customer credit of a german bank dataset is found at appendix.

A brief descriptive statistics of the selected dataset

Dataset has 21 features for one thousand instances. seven variables are quantitative and thirteen are nominal. The dependent variable is Cost Matrix that show customer was good (1) or bad (0) in regard with repay their loan.

Sample Chart

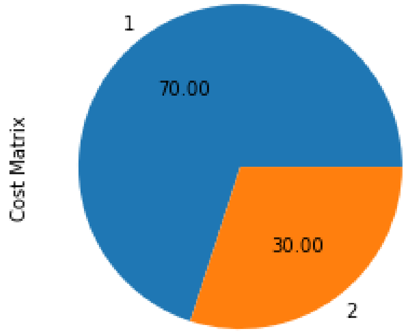


Table1. Sample Table for seven quantitative variables

Variable	mean	std	medain	min	max
Duration in month	20.903	12.058	18.0	4.0	72.0
Credit amount	3271.258	2822.736	2319.5	250.0	18424.0
Installment rate in percentage of disposable income	2.973	1.118	3.0	1.0	4.0
Present residence since	2.845	1.103	3.0	1.0	4.0

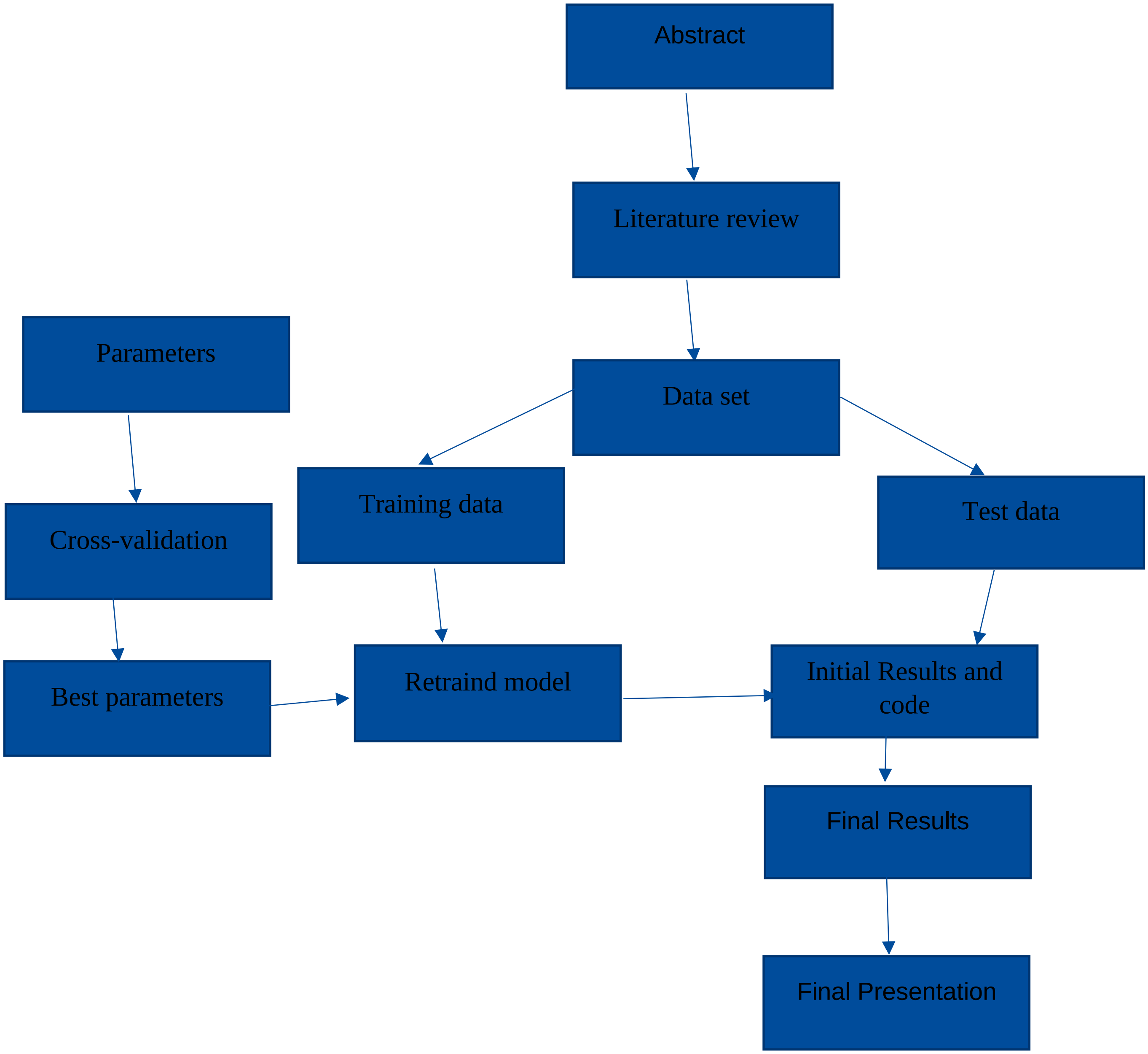
Age in years	35.546	11.375	33.0	19.0	75.0
Number of existing credits at this bank	1.407	0.577	1.0	1.0	4.0
Number of people being liable to provide maintenance for	1.155	0.362	1.0	1.0	2.0

Table2. Nominal Variables

Nominal Variables
Status of existing checking account
Credit history
Purpose
Savings account/bonds
Present employment since
Personal status and sex
Other debtors / guarantors
Property
Other installment plans
Housing

Job
Telephone
foreign worker

Graph methodology



Results

Is there significant relation between independent variables?

There was multicollinearity and I drop two independent variables; Duration in month, Age in years, that had high correlation with other variables.

Traditional models or ensemble models are better at prediction customer credit?

Which model has a better performance at prediction of credit customer?

Algorithms	Accuracy	AUC	Precision	Recall
XG	68.09	68.02	68.18	65.15
RF	75.5	71.0	58.78	58.67
KNN	62	61.8	59	70
DT	61.7	77.0	63	52.17
SVM	53	51.72	50	82.08
ANN	77	68.02	68.18	65.21
NB	54.18	68.02	68.00	65.17
LR	57.5	68.02	68.00	65.21

I could see that there is not significant difference between ensemble methods and traditional methods for prediction customer credit at my data set. All the methods had good metrics for classification and prediction. Although, random forest and neural network had the highest accuracy

Reference

Afsar, A., Houshdar M. R., &B. Minaie B. (2014). Customer credit clustering for presenting appropriate facilities. *Management Researches in Iran 17(4)*,1-24. https://www.sid.ir/en/journal/ViewPaper.aspx?ID=491564

Alborzi, M., & Khanbabaei, M. (2016). Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed RFM analysis method. *International Journal of Business Information Systems*, *23*(1), 1-22.

Awoyemi, J. O. (2017). Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis. A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)* (pp. 1-9). IEEE.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision support systems*, *50*(3), 602-613.

Burkov,A.(2019). *The hundred-page machine learning book*. Andriy Burkov

Chen, W., Ma, C., & Ma, L. (2009). Mining the customer credit using hybrid support vector machine technique. *Expert systems with applications, 36(4)*, 7611-7616. https://doi.org/10.1016/j.eswa.2008.09.054

Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. *Procedia computer science*, *165*, 631-641.

Fahmi, M., Hamdy, A., & Nagati, K. (2016). Data mining techniques for credit card fraud detection: Empirical study. *Sustainable Vital Technologies in Engineering & Informatics*, 1-9.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Kumar, A.& Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies*, *1*(1), 4-28.DOI:10.1504/IJDATS.2008.020020

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, *50*(4), 1113-1130. https://doi.org/10.1016/j.csda.2004.11.006

Li, Z., Liu, G., & Jiang, C. (2020). Deep representation learning with full center loss for credit card fraud detection. *IEEE Transactions on Computational Social Systems*, *7*(2), 569-579.

Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies, 261-270.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems* (pp. 841-848).

Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. *Procedia computer science*, *132*, 385-395.

Sahin, Y., & Duman, E. (2011). Detecting credit card fraud by ANN and logistic regression. In *2011 International Symposium on Innovations in Intelligent Systems and Applications* (pp. 315-319). IEEE.

Shen, A., Tong, R., & Deng, Y. (2007). Application of classification models on credit card fraud detection. In *2007 International conference on service systems and service management* ,(pp.1-4). IEEE.

Sherly, K. K., & Nedunchezhian, R. (2010). BOAT adaptive credit card fraud detection system. In *2010 IEEE International Conference on Computational Intelligence and Computing Research* (pp. 1-7). IEEE.

Appendix

A link to a repository on GitHub

CIND820_FINAL-PROJECT/CIND 820 Final(2).ipynb at main · Mohsenselseleh/CIND820_FINAL-PROJECT (github.com)

Dataset Statlog (German Credit Data) Data Set

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