Analytics of Bank Customers Credit Data.

Abstract, Literature Review, Data Description, and Approach

Professor: Ceni Babaoglu
Student: Zain ul Ebad Jaferi
Student Number: 501086726
Date of Submission: June 05, 2023

Abstract:

Topic: German Credit Risk

A person may borrow money from a bank or other financial organisation using bank credit. Based on the borrower's credit score, income, assets, debts, and other factors, the bank makes a decision. When a borrower forgets to repay the money they have borrowed, this process for banks always runs the risk of failing. Credit scoring is a complex management issue as a result.

Banks now employ data mining methods to create trained models by learning from samples, and the taught models are then utilised to make decisions in novel circumstances. Numerous indicators, such as the existence of a checking account, credit history, credit limit, savings account, job, debts, age, home ownership, the number of individuals who perform maintenance, etc., are used in this study to estimate a customer's creditworthiness.

This study determines whether a consumer is excellent or bad for a loan, which is a standard classification challenge. The study will employ algorithms such as decision tree, gradient boosting and random forest as well as learning techniques such as logistic regression analysis, neural network, svm, knn, and naive bayes. The data collection which I will use contains historical information about 1000 German bank customers.

I'll try to respond to these queries, which is more accurate at forecasting consumer credit: traditional models or ensemble models, which model performs better at predicting credit customers and If there exists a meaningful relationship between the independent variables.

Literature Review:

Maes (2002) By applying machine learning methods such as artificial neural networks and bayesian neural networks to the issue of researched the identification of credit card fraud. They showed that both ANN and BNN performed well in terms of fraud prediction, however BNN performed better in terms of fraud detection and had a shorter training period. However, ANN performed better than BNN in terms of speed.

Ng (2002) evaluated logistic regression and naive bayes on datasets from the UCI Machine Learning repository over the course of 15 experiments. They demonstrate that while the generative naive bayes classifier has a lower asymptotic error than the discriminative logistic regression technique, it converges more quickly to a greater asymptotic error. The generative naive Bayes perform better as the number of training cases increases; however the discriminative logistic regression outperforms naive bayes in terms of performance. They discover that there are a few instances in which the performance of logistic regression was inferior to that of naive bayes, although this is only seen in small datasets.

Bhattacharvva (2011) compared support vector machines and random forests with logistic regression as two advanced data mining techniques for spotting credit card fraud. They examine the effectiveness of the three strategies using data under sampling at various data undersampling densities. They employ a test dataset for performance evaluation that has a significantly lower fraud rate (0.5%) than the training datasets, which have varying degrees of undersampling. This method gives an idea of the performance that might be anticipated when models are used for fraud detection in situations when there aren't many fraudulent transactions. All methods demonstrated good capacity to identify fraud in the sample of data. Performance with various undersampling levels varied by method and according to various performance metrics. With smaller percentages of fraud in the training data, the sensitivity, G-mean, and weighted accuracy declined, while precision and specificity increased. While the logistic regression maintained the same performance on the F-measure and AUC, the random forest and SVM displayed a falling trend on AUC and an increasing trend on F. They demonstrate that random forests performed far better than other approaches. While SVM performance at higher file depths increased with a reduced proportion of fraud in the training data, logistic regression performed similarly across a range of undersampling levels. They made the case that SVM models were inferior to logistic regression, a widely used technique in data mining applications. They demonstrate how variable selection and exploratory data analysis affect the effectiveness of logistic regression. They compared the models using the same derived attributes. They made the case that random forests and SVM have the capacity for natural variable selection and have been shown to be effective when dealing with high-dimensional data.

Alborzi (2016) Studies on artificial neural networks for customer credit demonstrate that ANNs are highly accurate at predicting client credit. Utilising data mining and neural network

approaches, they employed a novel hybrid model of behavioural scoring and credit score to research credit clients. They used approaches for classification and grouping. They claimed that the model could successfully categorise and segment bank customers.

Lee (2006) investigated multivariate adaptive regression splines (MARS) and credit customer risks categorization and regression tree (CART). According to their findings, support vector machine, logistic regression, neural networks, and discriminant analysis all had worse average correct classification rates than cart and mars. They claimed credit scoring challenges have been studied using modelling techniques such as conventional statistical analysis and artificial intelligence methodologies. They demonstrate how frequently discriminant analysis and logistic regression are employed by customers, but that these methods have the problem of strong assumptions. Because of its memory feature, generalisation ability, and excellent credit scoring capability, they demonstrate that the artificial neural networks approach is a very well-known alternative in credit scoring tasks. However, it has some interpretive challenges and is unable to determine the relative importance of potential input variables. Since classification and regression tree (CART) and multivariate adaptive regression splines (MARS) can solve credit scoring problems without the limitations of discriminant analysis, logistic regression, and neural networks, they are used to examine the performance of credit scoring. On one bank, they employed their model.

Shen (2007) examine the use of classification models for detecting credit card fraud. They research three categorization techniques for business information analysis of credit card history, and they develop models for spotting fraud. In order to reduce the risk to the bank, they provide data mining techniques for credit card fraud detection, such as decision trees, logistic regression, and neural networks. They demonstrate that, when applied to the same data, a neural network model performs somewhat better than a logistic regression model in terms of accuracy.

Chen (2007) investigated using a hybrid support vector machine technology to mine customer credit. They demonstrate the necessity of strong fundamental assumptions for statistical categorization models. They contend that understanding the fundamental connections between input and output variables is not necessary for the use of artificial intelligence approaches. SVM is a cutting-edge data mining method that works well for classification and regression issues.

SVM, CART, and MARS were coupled to create the credit scoring models. They selected kernels to determine the parameters of the kernel and used SVM for classification and regression.

The performance of credit scoring was investigated using a hybrid modelling strategy that combined the svm approach with the cart, mars technique on a single credit card dataset provided by a local bank in China. In comparison to CART, MARS, and SVM, they demonstrate that the hybrid SVM technique has the best classification rate and the lowest Type II error. They also demonstrate that SVM has a greater ability to capture nonlinear relationships between

variables. According to their findings, the hybrid SVM credit scoring method they used in this study will have a greater level of credit scoring accuracy and a lower level of Type II error.

Li (2010) They employed logistic regression, decision trees, Random Forest, neural networks, and support vector machines to forecast client credit card segmentation. Their findings demonstrate the effectiveness of the tenfold cross-validation approach on synthetic minority oversampling technique (SMOTE) data with neural network method. They created a binary classifier for an unbalanced data set and used it to detect fraudulent transactions. They used the under-sampling strategy to balance their data set because it contained a small number of fraud transactions, which could result in a large variation of error. They display the typical F1 and area under curve (AUC) results. First, because to the nonlinear nature of a neural network, any neural network model performs better than random forest.

Sahin (2011) look into credit card fraud. They expound on the benefits of using data mining techniques, such as logistic regression and artificial neural networks, to address the issue of detecting credit card fraud. Their findings demonstrate that the suggested ANN classifiers perform better than LR classifiers in resolving the problem under study.

All models perform worse at detecting fraudulent transactions as the distribution of the training data sets becomes increasingly skewed.

Sherly (2012) For a system to detect credit card fraud, research decision trees, neural networks, and naive bayes classifiers. They demonstrate the effectiveness of decision trees using the BOAT algorithm as a classification and prediction tool. For identifying credit card fraud in situations where the data set changes dynamically, they employ the BOAT algorithm, which builds decision trees progressively. They could accurately predict fraud using their methodology.

Patil (2013) employs predictive modelling for data analytics-based credit card detection. The performance of their meta-classification approach, which combines tree, naive Bayesian, and k-nearest neighbour algorithms, has improved, according to the results. They demonstrate that the random forest decision tree performs optimally in terms of recall, accuracy, and precision. The overfitting of trees in memory as data volume rises is the only downside of random forest.

Afsar (2014) examined Customer Credit for Facility Granting. Their study's objective was to rank the customer segments and identify each one's strongest suit. They divide the customers into 10 groupings using the neural network. The clusters were ranked using the suggested methodology. Facilities grant operations were carried out for the individuals who were a part of the top clusters. In the realm of banking, Fahmi et al. (2016) examine the accuracy, sensitivity, specificity, and MCC metrics of naive bayes, k-nearest neighbour, and logistic regression approaches. The findings demonstrate that LR performs better. They contend that it may be observed that neural network models are utilised more frequently than other time-consuming techniques, and that it is crucial to carefully select any prospective input variables.

The Goal of research

My objective is to create models that will categorise and forecast customer credit for my dataset. To discover the best model for predicting consumer credit in the dataset, I'd want to compare performance of the neural network, logistic regression, decision tree, nave Bayes, knn, random forest, and xgboost algorithms using various metrics.

Method of Research

This study uses supervised machine learning, which makes it easier for it to draw appropriate conclusions about fresh data sets since it has a pre-defined collection of "training examples" that have been labelled. I first take a data set from the UCI website, convert it, and then load it into Python. The data set was preprocessed in several stages, including data preparation, handling missing data, data visualisation, handling outliers, handling correlations and heatmaps, handling multicollinearity (dropping two independent variables, Duration in month and Age in years that had high correlation with other variables), and handling an unbalanced data set with the undersampling method (randomly eliminating samples from the majority class until the class fractions are equal, or at least less unbalanced). To boost the accuracy of consumer credit prediction, I completed data analysis using a cross-validation method and various categorization models.

When the output variable is presented as a list of categories, such as sick or healthy, a classification model is used to forecast the results of a particular sample. Using labelled training data, supervised learning algorithms can learn the mapping function that converts input variables (x) into output variables (y). This enables us to generate outputs with high accuracy from new inputs. Below, I've attempted to define a few of the terms I used:

A sort of generalised linear model, logistic regression uses independent variables to forecast the likelihood that a binary (nominal or ordinal) variable will change its probability value from 0 to 1.

Using one or more variables, the logistic regression calculates the likelihood of a binary response. It determines the parameters that fit a sigmoid-shaped nonlinear function the best.

A technique for mining large datasets for knowledge is the decision tree. One of the most popular methods for categorization and prediction is the decision tree. The drawback is that a small alteration in the sample could have a significant impact on how it is classified.

The random forest is a collection of decision trees that is more practical and optimised for decision trees. It is often trained using the bagging approach. In a random subset of features, the random forest looks for the best feature (Geron, 2019).

A classification technique based on the Bayes theorem is known as the naive bayes. It posits the independence of the features. It selects the one with the highest probability.

According to Geron (2019), the k-nearest neighbour algorithm performs classification based on similarity measures such the Euclidean, Manhattan, or Minkowski distance functions (both for continuous variables).

Another kind of supervised learning is the use of ensemble methods. It entails merging the predictions of various weak machine learning models to obtain a stronger prediction on a fresh sample. In my research, I employed xgboosting, decision trees, and random forests as ensemble approaches.

Gradient boosting is one of the most potent ensemble machine learning techniques, according to Burkov (2019), and it can handle large datasets rapidly and easily. It also produces incredibly precise models. The number of trees and the depth of the trees are its key hyperparameters. Most machine learning algorithms have their limits, however an ensemble learning strategy could improve performance prediction. The two main ensemble learning techniques are bagging and boosting.

Machine learning metrics

I use the following formulas to measure correctness and precision in my study. A machine learning metric called the Mathews correlation coefficients (MCC) is used to assess the balance of binary (two class) classifiers. Since it considers both true and false values, it is typically viewed as a balanced measure that can be applied across all classes.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} y$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_\beta = \frac{(1 + \beta^2) \times \text{recall} \times \text{precision}}{\text{recall} + \beta^2 \times \text{precision}}$$

$$F_1 = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}.$$

Dataset

1000 elements make up the customer credit dataset from a German bank. A technique called under-sampling will be used to balance the dataset. The dataset was taken from the machine learning repository at UC Irvine; the link is in the appendix. This dataset, which includes data on 1000 clients, categorises individuals according to whether they are excellent or bad credit risks. The data set is very distorted and unbalanced. We have 21 qualities, 20 of which are independent variables, and just one of which is dependent (the cost matrix). The appendix contains the customer credit attributes from a German bank dataset.

A brief descriptive statistics of the selected dataset

There are 21 features in the dataset for 1,000 cases. Thirteen nominal variables and seven quantitative ones. The Cost Matrix, which displays whether a customer was good or bad at repaying a loan, is the dependent variable.

Quantitative Variables:

	Duration in month	Credit amount	Installment rate in percentage of disposable income	Present residence since	Age in years	Number of existing credits at this bank	Number of people being liable to provide maintenance for
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	20.903000	3271.258000	2.973000	2.845000	35.546000	1.407000	1.155000
std	12.058814	2822.736876	1.118715	1.103718	11.375469	0.577654	0.362086
min	4.000000	250.000000	1.000000	1.000000	19.000000	1.000000	1.000000
25%	12.000000	1365.500000	2.000000	2.000000	27.000000	1.000000	1.000000
50%	18.000000	2319.500000	3.000000	3.000000	33.000000	1.000000	1.000000
75%	24.000000	3972.250000	4.000000	4.000000	42.000000	2.000000	1.000000
max	72.000000	18424.000000	4.000000	4.000000	75.000000	4.000000	2.000000

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

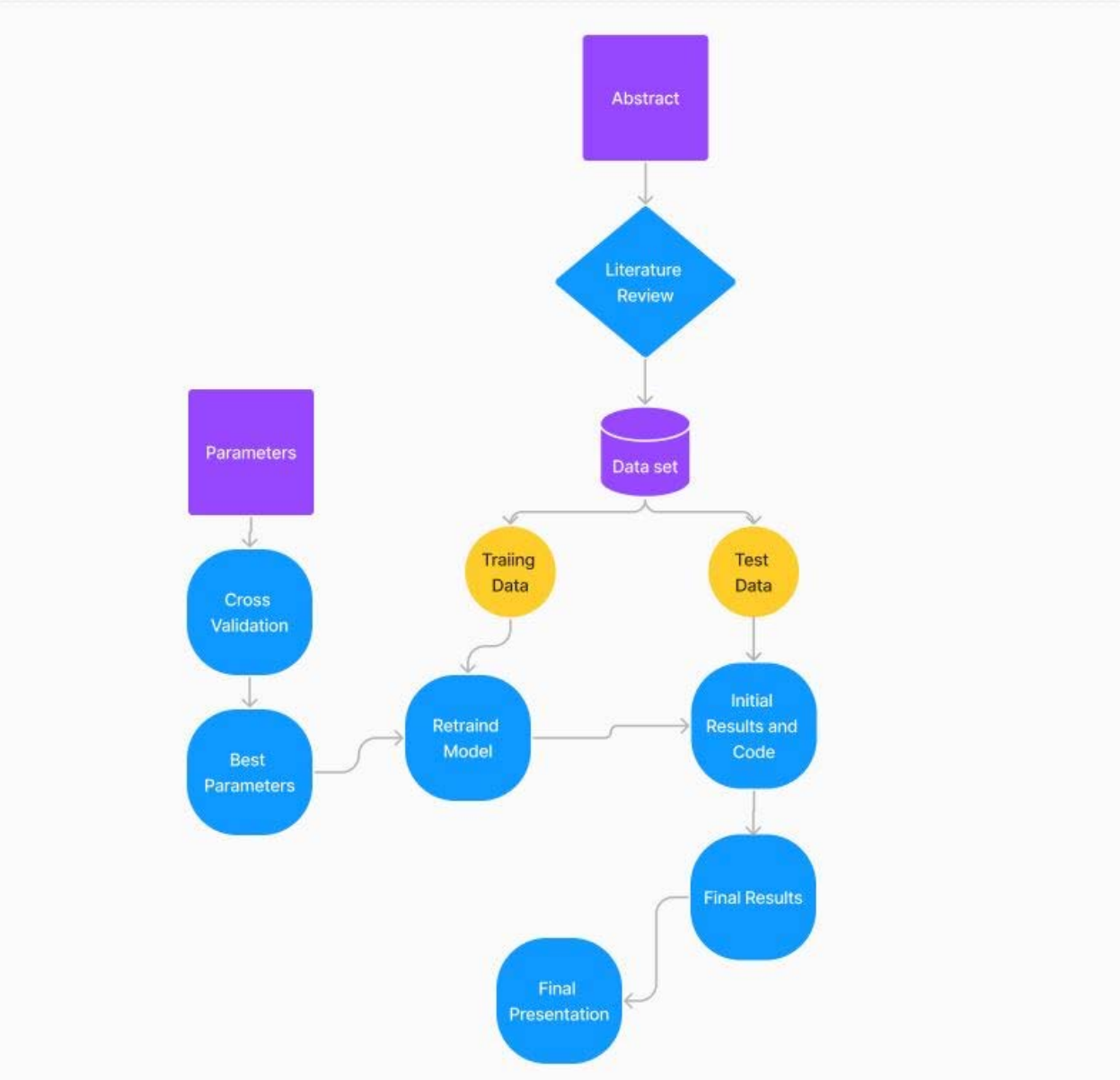
Housing

Job

Telephone

Foreign worker

Graph for tentative overall methodology



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

Below link to a repository on GitHub

https://github.com/zainirs/Analytics-of-Bank-Customers-Credit-Data.git

Dataset(German Credit Data)

Literature review pdf

German Credit Analytics.ipynb