Majalah Bisnis & IPTEK
Vol 16, No. 1, April 2023, 8 – 15
ISSN Online: 2502-1559. DOI: https://doi.org/10.55208/bistek

Predicting Credit Paying Ability With Machine Learning Algorithms

Ama Febriyanti[1], Tomy Rizky Izzalqurny[2]
Universitas Negeri Malang[1,2]
Email: ama.febriyanti.1904226@students.um.ac.id[1], tomyrizky.izzalqurny.fe@um.ac.id[2]

ABSTRACT

Most people still have difficulty accessing finance because of a lack or even no credit history. This study aims to develop a data model that predicts a customer's ability to pay from various aspects other than credit history. This study uses the CRSIP-DM (Cross Industry Standard Process Model for Data mining) method. The data used in this study is the Home Credit Default Risk dataset collected by documentation techniques. The data were then analyzed using data modeling analysis techniques, namely logistic regressor, decision tree classifier, random forest classifier, and lgbm classifier. This study found that the best model for predicting client payment ability is the lgbm classifier or the Random Forest Classifier.

Keywords: Credit, Default, Data, Machine Learning

INTRODUCTION

Credit plays a vital role in helping people meet their economic needs. However, many individuals need help accessing credit due to a lack or absence of credit history (Beck, 2021). This condition confuses society as they are expected to have a credit history of obtaining credit, even though credit history can only be obtained when someone provides credit. Nevertheless, no one can be blamed for this issue. Creditors are simply trying to minimize risks to prevent defaults or loan failures. A high number of defaulting customers will reduce the profits and capital of the creditors (Nirwana et al., 2022).

Credit risk is among the top five risks that pose significant challenges to the banking industry or other financial institutions (Syafudin et al., 2021). Credit risk assessment is crucial in determining whether a customer can repay their debt or is at risk of default. Credit risk assessment also minimizes the possibility of defaults, enabling creditors to maintain business continuity and prevent losses (Rizki et al., 2020). The key indicator to assess credit risk is the non-performing loan (NPL) ratio. Creditors must continually monitor NPL, as a higher value indicates a more significant number of customers with high credit risk.

Predicting the status of customer default on credit (default) is necessary for NPL monitoring. Predicting default status also serves as a tool for assessing credit risk. In addition to relying on credit history, predictions need to consider other factors that can trigger defaults. Therefore, creditors need to develop predictive models of default status to assess customers' repayment abilities, thus reaching out to individuals without a credit history. Manual credit risk assessment predictions are subjective and susceptible to fraud (Syafudin et al., 2021).

Previous studies on predicting default credit status have developed models utilizing machine learning algorithms. Normah et al. (2022) used CHAID (Chi-square Automatic Interaction Detection Analysis) analysis to classify potential customers and determine credit status to reduce the number of defaulted loans. Syafudin et al. (2021) predicted bank loan status using Deep Learning Neural (DNN) and achieved an accuracy of 82.27%. The K-Nearest Neighbor (KNN) method has also been used to predict delinquent loan payments at PT FIF Group Arjawinangun Branch, resulting in an accuracy of 71% (Pratama et al., 2021). However, these previous studies have limitations, such as the absence of a test dataset, making the accuracy beyond training unknown.

Additionally, the datasets used were limited to small-sized datasets. However, large datasets (big datasets) enable companies to monitor various aspects of customers' lives comprehensively. Large datasets can also significantly improve model accuracy (Beck, 2021).

Based on this background, this study aims to develop a prediction model for default credit status using machine learning algorithms and a large dataset to predict customers' repayment abilities accurately. The next section will present the research methodology, followed by the presentation of research results, and conclude with the research findings.

METHOD

This quantitative study utilizes the CRISP-DM (Cross et al. Model for Data Mining) method. CRISP-DM is a method widely used by experts in data development modeling to solve a problem (Givari et al., 2022). The research stages are adapted to the CRISP-DM framework, as shown in Figure 1.

The data used in this study is the Home Credit Default Risk dataset. Data collection is conducted using documentation techniques, which involve collecting and analyzing the necessary information and datasets from the Kaggle dataset repository (Montoya et al., 2018). The dataset is then analyzed using data modeling analysis techniques such as logistic regressor, decision tree classifier, random forest classifier, and lgbm classifier. Testing in this research is conducted with the assistance of the Google Colaboratory IDE.

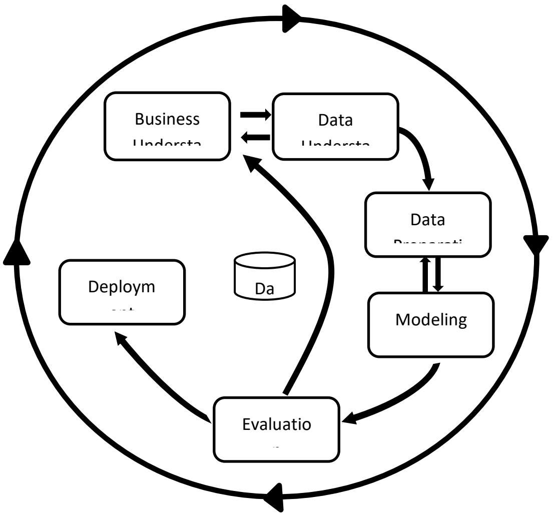


Figure 1. The CRISP-DM Framework

RESULT AND DISCUSSION

Research Results:

A. Business Understanding

Home Credit is a company that was established in 1997. Home Credit provides financing for international consumers and operates in eight countries (Home Credit, 2022). Home Credit's business objective is to develop financial services for underserved individuals accessing banking financing. Many individuals need help obtaining loans or credit due to a lack or absence of credit history (Givari et al., 2022). As a result, many credit applications are rejected. Home Credit utilizes alternative data, including telecommunications and other information, to address this situation to predict customers' repayment capacity. In 2018, Home Credit released the Home Credit Default Risk dataset containing information about its customers on the Kaggle dataset repository. Home Credit allows everyone to use this data to develop predictive models to assist the company. The goal of mining the Home Credit Default Risk dataset is to help the company assess customers' repayment capacity and make accurate and effective decisions regarding credit disbursement. This condition ensures the company does not mistakenly grant credit to customers likely to default.

B. Data Understanding

The dataset released by Home Credit is a large dataset consisting of 9 datasets in CSV file format. Each file contains specific information related to Home Credit's customers. In this study, the dataset used is the combined train dataset and the bureau dataset. The bureau dataset is used because it contains information about customers' credit history, which can be helpful during the modeling process. The details of the dataset are as follows:

Table 1. Dataset Details

Dataset	Jumlah data	Jumlah Fitur (Kolom)
Train	307.511	122

Bureau	1.048.575	17
Join train and bureau	307.513	137

Of the 137 data features resulting from the train and bureau datasets joining, 77 features are of float type, 41 features are of integer type, and 19 features are of object type, also known as categorical data.

C. Data Preparation

The first step in this data preparation stage is to import the required libraries into the Google Colaboratory IDE. After that, we load the datasets using the pandas' library. Next, we merge the datasets, namely train, and bureau, using the left join method, where we merge the data based on the training dataset. The result of the dataset merge yields 307,513 data entries with 137 features. After merging the data, the next step is data cleansing, which involves dropping irrelevant columns and handling outliers, missing values, and duplicates. These steps ensure the machine learning algorithm can process the data and produce the best results (Zuama et al., 2022). The columns or features 'SK_ID_BUREAU' and 'SK_ID_CURR' are dropped as they only contain transaction identification numbers irrelevant to the prediction model. Outliers in this dataset are found in the 'DAYS_EMPLOYED' feature and are handled by replacing them with NaN values.

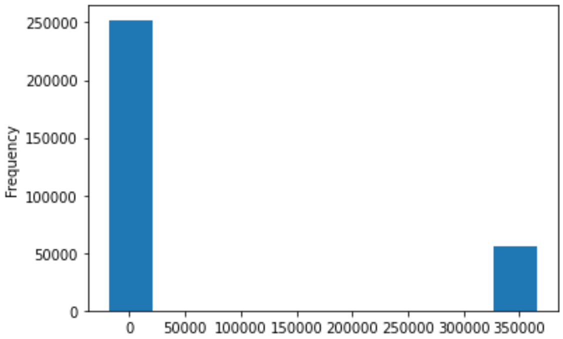
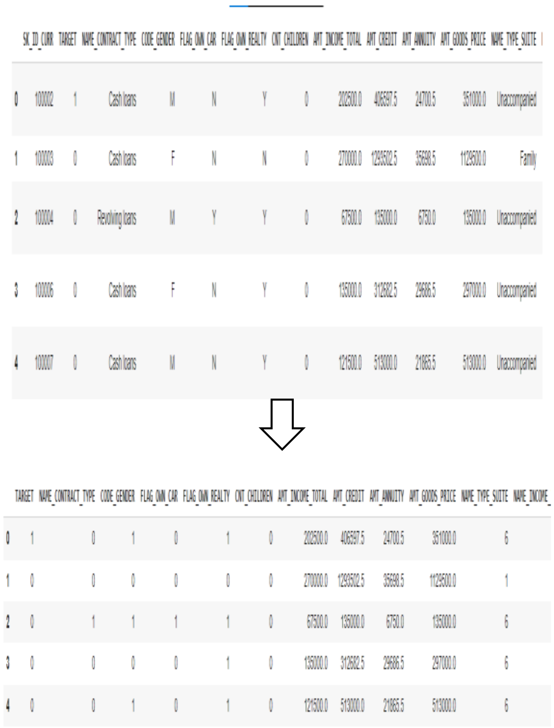


Figure 2. Histplot of days employed.

The plot diagram shows that the number of features with negative values is greater than the number of features with positive values. Upon inspection, we found 55,374 outliers in the 'DAYS_EMPLOYED' feature. We handle missing values using the fillna method, replacing them with the mean value for each numeric feature. We use the fillna method with a new value, 'Data_Not_Available,' to indicate data unavailability for categorical features. In the joined dataset of train and bureau, we did not find any duplicate data, so there is no need for duplicate handling.

After the data is cleaned, data transformation is necessary. Because the model's performance relies on numerical data, we must transform non-numerical or categorical data. In this case, we utilize label encoding to convert the data into dummy values automatically.



	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5	351000.0	Unaccompanied
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	1129500.0	Family
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	135000.0	Unaccompanied
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	29686.5	297000.0	Unaccompanied
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0	21865.5	513000.0	Unaccompanied

	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAME_INCOME_
0	1	0	1	0	1	0	202500.0	406597.5	24700.5	351000.0	6	
1	0	0	0	0	0	0	270000.0	1293502.5	35698.5	1129500.0	1	
2	0	1	1	1	1	0	67500.0	135000.0	6750.0	135000.0	6	
3	0	0	0	0	1	0	135000.0	312682.5	29686.5	297000.0	6	
4	0	0	1	0	1	0	121500.0	513000.0	21865.5	513000.0	6	

Figure 3. Encoding data

Next step is Exploratory Data Analysis (EDA) which generates a seaborn heatmap visualization depicting the 5 variables with the strongest correlation to the target, as follows:

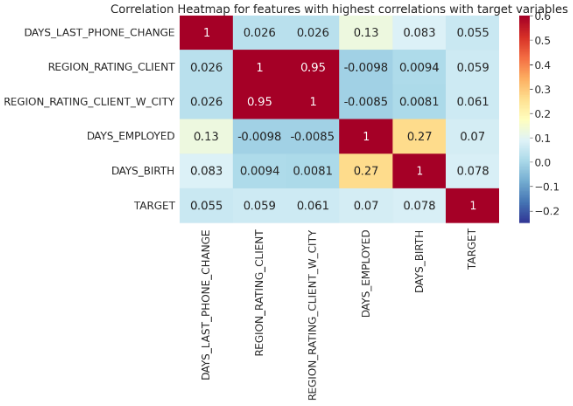


Figure 4. Heatmap of Features with the Strongest Positive Correlation

D. Modeling

The dependent variable in the advanced data modeling is 'TARGET,' while the independent variables are all features except 'TARGET' and the previously dropped features. There are 4 machine learning algorithms used to process these variables: logistic regressor, decision tree classifier, random forest classifier, and lgbm classifier. The best model will be determined based on their output.

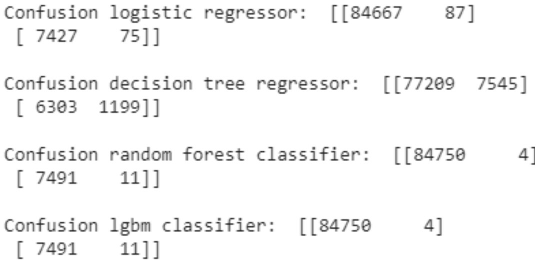
E. Evaluation

The evaluation of the four data modeling models utilizes the following assessments: Accuracy Score, Recall Score, AUC Performance Score, F1 Score, Precision Score, and Confusion Matrix.

Table 2. Model Evaluation Score

Model	**Skor**				
	Accuracy	***Recall***	***AUC Performance***	***F1***	***Precision***
Logistic Regressor	0,92	0,01	0,50	0,02	0,46
Decision Tree Classifier	0,85	0,16	0,54	0,02	0,14
Random Fores	0,92	0,00	0,50	0,02	0,73

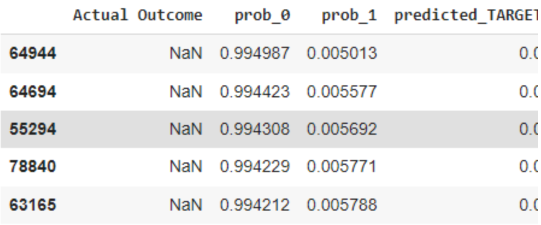
t Classifier					
Lgbm Classifier	0,92	0,00	0,50	0,02	0,73



```
Confusion logistic regressor:  [[84667    87]
 [ 7427    75]]

Confusion decision tree regressor:  [[77209  7545]
 [ 6303  1199]]

Confusion random forest classifier:  [[84750     4]
 [ 7491    11]]

Confusion lgbm classifier:  [[84750     4]
 [ 7491    11]]
```

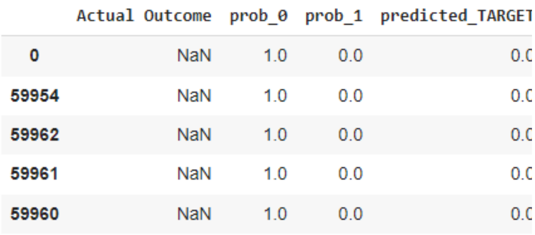
Figure 5. Confusion Matrix Score

From the above evaluation results, it is known that out of the four developed models, the best models to use are lgbm classifier and Random Forest Classifier. The following are the prediction outputs using these two best models:



	Actual Outcome	prob_0	prob_1	predicted_TARGET
64944	NaN	0.994987	0.005013	0.(
64694	NaN	0.994423	0.005577	0.(
55294	NaN	0.994308	0.005692	0.(
78840	NaN	0.994229	0.005771	0.(
63165	NaN	0.994212	0.005788	0.(

Figure 6. Output predictions with the lgbm classifier model



	Actual Outcome	prob_0	prob_1	predicted_TARGET
0	NaN	1.0	0.0	0.(
59954	NaN	1.0	0.0	0.(
59962	NaN	1.0	0.0	0.(
59961	NaN	1.0	0.0	0.(
59960	NaN	1.0	0.0	0.(

Figure 7. Prediction output with the Random Forest Classifier model

F. Deployment

Deployment is the process of reporting the results of the modeling and evaluation in data development. This report can be used to make accurate credit decisions by predicting and identifying all relevant aspects and preventing potential defaults. (Givari et al., 2022). Deployment can be done using Power BI or Google Data Studio software. The deployment of this research is as follows:

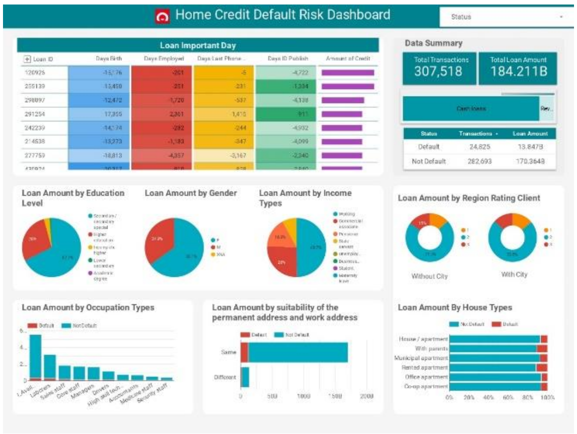


Figure 8. Dashboard Home Credit Default Risk

The dashboard summarizes the data, data distribution, and descriptive statistics of the Home Credit Default Risk dataset.

Discussion

Based on the results of the developed prediction model, it is known that the best model for predicting the large-sized Home Credit Default Risk dataset is either the lgbm classifier or Random Forest Classifier. Both models produce the same score, which is 0.92 or 91.88%. This means that out of the entire data, 91.88% has been correctly classified and predicted not to result in default or loan default. This result can be considered good because the score is close to 100%.

CONCLUSION

Based on the conducted research to predict customers' credit payment ability, it can be concluded that the developed models, namely logistic regressor, decision tree classifier, random forest classifier, and lgbm classifier, produce accuracy values higher than 50%. However, considering other evaluation scores, the lgbm classifier or Random Forest Classifier is the best model to use. The limitation of this research is the utilization and merging of limited datasets, specifically the train and bureau datasets. Therefore, future research should explore other datasets to obtain additional features supporting the developed prediction model.

REFERENCES

Beck, P. J. (2021). Summer 2021 CS 687 Capstone Project Progress Report Predicting Loan Default Likelihood Using Machine Learning.

Givari, M. R., Sulaeman, M. R., & Umaidah, Y. (2022). Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit. Nuansa Informatika, 16(1), 141–149. https://doi.org/10.25134/nuansa.v16i1.5406

Home Credit. (2022). About Us. Home Credit International a.S. https://www.homecredit.net/about-us.aspx/

Montoya, A., Inversion, KirillOdintsov, & Kotek, M. (2018). Home Credit Default Risk. Kaggle. https://kaggle.com/competitions/home-credit-default-risk

Nirwana, A., Siregar, A., & Rahmat, R. (2022). Klasifikasi Permasalahan Kredit Macet Pada Bank Menggunakan Algoritma Decision Tree C4. 5. Scientific Student Journal for Information, Technology and Science, 3(1), 43–50.

Pratama, N. S. H., Afandi, D. T., Mulyawan, M., Iin, I., & Nuris, N. D. (2021). Menurunkan Presentase Kredit Macet Nasabah Dengan Menggunakan Algoritma K-Nearest Neighbor. INFORMATION SYSTEM FOR EDUCATORS AND PROFESSIONALS : Journal of Information System, 5(2), 131. https://doi.org/10.51211/isbi.v5i2.1537

Rizki, M., Hadiyul Umam, M. I., & Hamzah, M. L. (2020). Aplikasi Data Mining Dengan Metode CHAID Dalam Menentukan Status Kredit. Jurnal Sains, Teknologi Dan Industri, 18(1), 29. https://doi.org/10.24014/sitekin.v18i1.11421

Syafudin, S., Nugraha, R. A., Handayani, K., Gata, W., & Linawati, S. (2021). Prediksi

Status Pinjaman Bank dengan Deep Learning Neural Network (DNN) Sukri. Jurnal Teknik Komputer AMIK BSI, 7(2), 130–135. https://doi.org/10.31294/jtk.v4i2

Zuama, R. A., Rahmatullah, S., & Yuliani, Y. (2022). Analisis Performa Algoritma Machine Learning pada Prediksi Penyakit Cerebrovascular Accidents. Jurnal Media Informatika Budidarma, 6(1), 531. https://doi.org/10.30865/mib.v6i1.3488