Credit Card Customer Churn Prediction With Binary Logistic Regression

Enny[1], Renaldo Pratama[2] , Alex Fernando Hasahatan[3], Jerry Heikal[4]
[1,2,3,4] Bakrie University, Indonesia
E-mail: ennymd.hery@gmail.com [1], renaldopratama53@gmail.com[2]

Article History:
Received: 04 September 2025
Revised: 20 September 2025
Accepted: 28 September 2025

Keywords: *credit card churn, binary logistic regression, predictive analytics, banking, customer retention*

Abstract: *The increasing prevalence of credit card usage in Indonesia has brought significant benefits to the national economy but also presents challenges for the banking industry, particularly regarding customer churn. This research aims to analyze the factors influencing credit card customer churn using the Binary Logistic Regression method. Utilizing a dataset of 5,000 entries from Kaggle, the research incorporates demographic and behavioral variables such as age, marital status, product ownership, inactivity periods, and transaction patterns. Data preprocessing steps included handling missing values, encoding categorical variables, and feature scaling. The model was trained with 80% of the data and tested on 20%, with variable selection based on p-values (< 0.05). The results indicate that eight key factors significantly affect churn likelihood, including number of dependents, marital status, number of products, inactive months, number of contacts, total transactions, transaction count, and the ratio of Q4 to Q1 transaction amounts. The model achieved an accuracy of 87.10%, demonstrating strong predictive performance, especially for non-churn cases. These findings suggest that classical statistical approaches remain effective for churn prediction when supported by comprehensive data processing and relevant variable selection. The study contributes valuable insights for financial institutions to develop targeted retention strategies and enhance customer loyalty.*

INTRODUCTION

The increasing use of credit cards in Indonesia has made a significant contribution to national economic growth and has facilitated financial transactions for the public. However, alongside these conveniences, the banking industry faces a major challenge in the form of customer churn, the phenomenon where customers discontinue the use of credit card services. Churn has become a primary concern for banks, as it can lead to decreased revenue, increased costs for acquiring new customers, and a decline in customer loyalty and trust toward financial institutions (Maulana, 2025).

Credit card customer churn is a complex issue, involving various factors such as demographics, transaction behavior, and credit card usage patterns. Banks are required to identify customers who are at risk of churning as early as possible so that they can design effective and efficient retention strategies. In recent years, churn prediction has become a major topic in data mining and artificial intelligence (AI), especially in the context of the financial industry. Data-driven approaches and predictive analytics are considered capable of providing more accurate and measurable solutions in identifying customer behavioral patterns associated with churn (Siddiqui, 2023).

Research highlights that credit churn is influenced by a combination of demographic, behavioral, and transactional factors. Key determinants include, demographic such as age, marital status, education level, and income significantly affect churn probability. For instance, customers aged 40–55 have been found more likely to churn, possibly due to changing financial priorities or dissatisfaction with services. Behavioral patterns such as transaction count, total transaction value, product ownership, and inactivity periods are strong predictors of churn. Customers with lower engagement or usage are at higher risk. The last is service experience such as customer satisfaction, perceived value, and quality of service also play a vital role (Amin, 2017).

Based on data from AKKI (Indonesian Credit Card Association), credit card users are increasing, which in 2024 amounted to 18.5 million and continued to grow and became as many as 18.6 million in March 2025 (Akki.or.id, 2025). However, with the increasing use of credit cards, it cannot be separated from an obstacle, especially in maintaining customer loyalty. Despite this growth, the credit card industry faces a rising Non-Performing Loan (NPL) ratio, which exceeded 2% in March 2025. One contributing factor is the declining repayment ability of customers amid economic pressures, which is closely related to the risk of customer churn (Mediatama, 2025). In 2017 BRI, one of the largest banks in Indonesia, lost 17 thousand customers, which shows that churn is a real challenge for the banking industry in Indonesia (Sunjaya, n.d.)

In bank and financial institution, the 5C principle (character, capacity, capital, collateral, and condition) is a foundational framework used to evaluate the creditworthiness of applicants for credit products, including credit cards. Each "C" represents a key criterion for character which refers to the borrower's reputation, integrity, and track record in repaying debts, lenders assess character by reviewing credit history, payment behaviour, and credit scores, often using credit reports from bureaus to predict future reliability. Capacity measures the applicant's ability to repay the debt, typically by analysing income, employment status, and existing debt obligations, lenders evaluate the debt-to-income ratio and the stability of the applicant's income sources to ensure they can handle new credit responsibly. Capital indicates the amount of money the applicant has invested or available as savings and assets. A higher level of capital demonstrates financial stability and a greater likelihood of meeting credit obligations, serving as a buffer in case of financial hardship. Collateral refers to assets that the borrower can pledge as security for the loan or credit. While credit cards are typically unsecured, some issuers may consider collateral for secured credit cards or in special cases to reduce the lender’s risk. Conditions encompasses the purpose of the loan or credit, the amount requested, and broader economic or market conditions. Lenders consider factors such as interest rates, economic outlook, and the applicant’s intended use of funds to assess risk (5 Cs of Credit, n.d.)

One of the most widely used methods to address churn is Binary Logistic Regression, a statistical technique used to predict the likelihood of an event occurring (in this case, churn or not-churn) based on relevant predictor variables. It is particularly suitable for cases with binary dependent variables, such as churn (1) and non-churn (0). In marketing research, this method is often used to analyze the relationship between factors such as price, promotion, product quality, and demographic and behavioral characteristics, and customers' decisions to continue or discontinue a service (Hair, 2010).

This study uses a dataset obtained from Kaggle, consisting of 5,000 credit card customer entries. The data includes demographic attributes such as age, gender, education, marital status, and annual income, as well as behavioral attributes such as the number of products owned, length of time as a customer, inactive months, number of contacts with the bank, total credit limit, total transactions, and credit card usage ratio (Maulana, 2025).

The Binary Logistic Regression model was implemented by splitting the dataset into training (80%) and testing (20%) sets. The training process was carried out systematically, ensuring reproducibility by setting a random state. Subsequently, independent variables were selected based on p-values (< 0.05), so that only statistically significant variables were included in the final model. The analysis results show that there are eight main factors that significantly influence the likelihood of customer churn, including the number of dependents, marital status, number of products, inactive months, number of contacts, total transactions, transaction count, and the ratio of fourth quarter to first quarter transaction amounts. Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score, as well as confusion matrix analysis to assess the model's ability to correctly classify churn and non-churn customers.

Efforts to predict credit card customer churn using statistical and machine learning approaches are strategic steps of great importance in today's digital era. By understanding the main factors causing churn and implementing accurate prediction models, banks can minimize the risk of losing customers, improve operational efficiency, and strengthen their competitive position in an increasingly dynamic and competitive financial services industry. The results of this study are expected to serve as a reference for financial institutions in developing more effective customer retention strategies, reducing churn rates, and increasing customer satisfaction and loyalty. Moreover, this research supports the strengthening of information technology in the banking sector through the use of data analysis and artificial intelligence for more accurate and evidence-based decision-making.

METHODE

This is a quantitative research study that applies the Binary Logistic Regression method to analyse the factors affecting customer of credit card churn prediction. Binary Logistic Regression is a statistical technique used to predict the likelihood of an event occurring based on predictor variables. This method is often applied in marketing research to analyse the relationship between factors such as price, promotion, and product quality and consumer decisions (Hair et al., 2010). This technique is particularly suitable for binary dependent variables, in this study means credit card churn or non-churn.

The regression model's accuracy formula is as follows:

$$y = \frac{1}{1 + exp^{(-(c+b1x1+b2x2+b3x3\ldots)}}$$

Y : Dependent Variable
X1, X2, X3 : Independent Variable

This dataset has 5000 entries encompassing detailed attributes pertaining to client demographics and credit card utilization patterns. The data was gathered from Kaggle to guarantee that the dataset reflects a wide and varied sample of credit card consumers. The data collection method entailed gathering comprehensive metrics regarding client behaviour and demographics, which are crucial for constructing precise predictive models. The dependent variable in this study is credit card churn (binary: 1 = churn and 0 = not churn). The dimension based on "5C" with detailed independent variables as shown in table 1 below:

Table 1. Dimension and Indicator

Dimension	Indicators
Character	usia, gender, pendidikan, status nikah, jumlah tanggungan, jumlah kontak, lama nasabah
Capacity	penghasilan tahunan, total transaksi, jumlah transaksi, rasio_transaksi_Q4_Q1, rasio_jumlah_transaksi_Q4_Q1, rasio_pemakaian
Capital	total limit kredit, total limit kredit dipakai, sisa limit kredit
Collateral	tipe kartu kredit, jumlah produk
Condition	bulan nonactive

Data processing in this study analysed using the phyton binary logistic regression model to identify the factors affecting the dependent variable (Y) namely credit card churn prediction. Due to the nature of this dataset, prior to analysis, the data underwent several pre-processing stages including handling missing values, encoding categorical variables into numeric values, and scaling numeric features using MinMaxScaler from Scikit-Learn. These steps are crucial to ensure that the data used in modelling is ready and free from technical biases that could affect the results.

Table 2. Encoding Feature Category

Variable	Description
Gender	Gender (Female = 1, Male = 0)
Pendidikan	Education level (0 = Uneducated, 1 = Doctorate, 2 = Graduate, 3 = High School, 4 = College, 5 = Master, 6 = Unknown)
Status _Nikah	Marital Status (0 = Divorced, 1 = Married, 2 = Single)
Penghasilan_Tahunan	Income categories (0 = $120K+, 1 = $40K-$60K, 2 = $60K-$80K, 3 = $80K-$120K, 4 = Less than $40K)
Yipe_Kartu_Kredit	Type of credit card (0 = Blue, 1 = Gold, 2 = Platinum, 3 = Silver)

RESULT AND DISCUSSION

Data processing in this study analysed using the phyton binary logistic regression model to identify the factors affecting the dependent variable (Y) namely credit card churn prediction. The dataset split into training (80%) and testing (20%) and then we set the random state to 42 for reproducibility and we run the binary logistic regression training. The step 1 result can be seen in Figure 1.

Dep. Variable:	churn	No. Observations:	3559
Model:	Logit	Df Residuals:	3540
Method:	MLE	Df Model:	18
Date:	Sat, 21 Jun 2025	Pseudo R-squ.:	0.4573
Time:	18:20:23	Log-Likelihood:	-835.15
converged:	True	LL-Null:	-1538.9
Covariance Type:	nonrobust	LLR p-value:	3.399e-288

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.4091	3.65e+06	9.35e-07	1.000	-7.14e+06	7.14e+06
usia	-0.0106	0.012	-0.850	0.395	-0.035	0.014
jumlah_tanggungan	0.2290	0.050	4.541	0.000	0.130	0.328
pendidikan	-0.0018	0.035	-0.051	0.959	-0.070	0.066
status_nikah	0.3408	0.109	3.116	0.002	0.126	0.555
penghasilan_tahunan	0.1053	0.100	1.050	0.294	-0.091	0.302
tipe_kartu_kredit	0.1344	0.114	1.178	0.239	-0.089	0.358
lama_nasabah	0.0005	0.013	0.043	0.966	-0.024	0.026
jumlah_produk	-0.4824	0.047	-10.365	0.000	-0.574	-0.391
bulan_nonactive	0.4452	0.065	6.803	0.000	0.317	0.573
jumlah_kontak	0.4742	0.061	7.821	0.000	0.355	0.593
total_limit_kredit	-0.3250	1.02e+07	-3.18e-08	1.000	-2e+07	2e+07
total_limit_kredit_dipakai	-2.4038	7.5e+05	-3.2e-06	1.000	-1.47e+06	1.47e+06
sisa_limit_kredit	0.0044	1.08e+07	4.09e-10	1.000	-2.11e+07	2.11e+07
rasio_transaksi_Q4_Q1	-0.5985	0.321	-1.862	0.063	-1.228	0.031
total_transaksi	8.3794	0.670	12.505	0.000	7.066	9.693
jumlah_transaksi	-0.1159	0.006	-18.692	0.000	-0.128	-0.104
rasio_jumlah_transaksi_Q4_Q1	-2.6325	0.306	-8.600	0.000	-3.232	-2.033
rasio_pemakaian	-0.1848	0.417	-0.443	0.658	-1.002	0.632
male	1.2344	3.65e+06	3.38e-07	1.000	-7.15e+06	7.15e+06
female	2.1747	3.65e+06	5.96e-07	1.000	-7.15e+06	7.15e+06

Figure 1. Logistic Regression Result

In above step 1 logistic regression result we can see the significance (p-values) of each variables. There are 8 significant variables namely jumlah_tanggungan, status_nikah, jumlah_produk, bulan_nonactive, jumlah_kontak, total_transaksi, jumlah_transaksi and rasio_jumlah_transaksi_Q4_Q1.

The Step 2, we choose the independent variable with p value result < 0.05 and re-running the logistic regression model with the significant variables. Below is the result as seen in Figure2.

```
==============================================================================
Dep. Variable:                  churn   No. Observations:                 3559
Model:                          Logit   Df Residuals:                     3550
Method:                           MLE   Df Model:                            8
Date:                Sat, 21 Jun 2025   Pseudo R-squ.:                  0.3803
Time:                        18:20:24   Log-Likelihood:                -953.61
converged:                       True   LL-Null:                       -1538.9
Covariance Type:            nonrobust   LLR p-value:                2.112e-247
===============================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                         2.6727      0.394      6.782      0.000       1.900       3.445
jumlah_tanggungan             0.2330      0.046      5.090      0.000       0.143       0.323
status_nikah                  0.4016      0.103      3.891      0.000       0.199       0.604
jumlah_produk                -0.4854      0.043    -11.240      0.000      -0.570      -0.401
bulan_nonactive               0.4524      0.061      7.420      0.000       0.333       0.572
jumlah_kontak                 0.5018      0.056      8.933      0.000       0.392       0.612
total_transaksi               0.0004   3.42e-05     11.859      0.000       0.000       0.000
jumlah_transaksi             -0.1024      0.006    -18.314      0.000      -0.113      -0.091
rasio_jumlah_transaksi_Q4_Q1 -3.0321      0.277    -10.947      0.000      -3.575      -2.489
===============================================================================================
```

Figure 2. Logistic Regression Result

In this refined output, all the independent variables are significant and can be used for further final model analysis as below.

$$y = \frac{1}{1 + exp^{(-(c+b1x1+b2x2+b3x3\ldots)}}$$

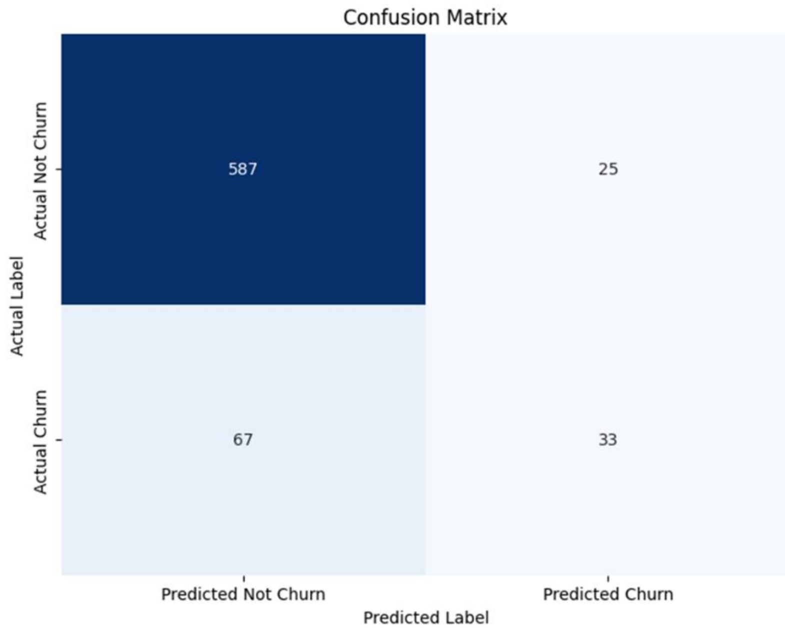
$$y = \frac{1}{1 + exp^{(-(2.6727+(0.2330.x1)+(0.4016.x2)+(-0.4854.x3)+(0.4524.x4)+(0.5018.x5)+(0.0004.x6)+(-0.1024.x7)+(-3.0321.x8))}}$$

Where :
Y : Dependent variables : Credit Card Churn Prediction
X1 : Independent variables representing jumlah_tanggungan
X2 : Independent variables representing status_nikah
X3 : Independent variables representing jumlah_produk
X4 : Independent variables representing bulan_nonactive
X5 : Independent variables representing jumlah_kontak
X6 : Independent variables representing total_transaksi
X7 : Independent variables representing jumlah_transaksi
X8 : Independent variables representing rasio_jumlah_transaksi_Q4_Q1

We then run the training again and below is the result in table 3.

Table 3. Classification Report

	Predicted Not Churn	Predicted Churn	Percentage Correct
Actual Not Churn (0)	587	25	95.90%
Actual Churn (1)	67	33	33.00%
Overall			87.10%



Based on the table above, it can be concluded that these are 587 customers who were actually not going to churn, and the model correctly predicted that. This is the largest group in the dataset, which shows the model is quite strong at recognizing loyal or retained customers. These 25 customers were predicted to churn, but in reality, they did not churn. This is a relatively small number, which is good. However, in a real business setting, this could lead to unnecessary retention efforts or incentives being offered to people who didn't need them. These are 67 predicted not churn and customers actually churned, but the model failed to identify them. This is a concern because these customers left without the company intervening. If your goal is churn prevention, false negatives are costly. There are 33 customers were correctly predicted to churn and did actually churn. These are the valuable wins from the model and for this customers we can apply retention strategies. The accuracy of this binary logistic regression model is 87.10%.

Based on 5C framework (character, capacity, capital, collateral, and condition), the research findings indicate that 4C factors significantly influence credit card churn which are character, capacity, collateral, and condition. The character representing jumlah tanggungan, status nikah. The model confirms that personal and behavioural traits influence churn. More contacts and dependents increase churn risk, suggesting a need for better customer engagement and family centred financial planning. The capacity representing total transaksi, jumlah transaksi and rasio jumlah transaksi Q4_Q1. Behavioral indicators (like reduced transaction frequency) are key churn signals. High-value but low-frequency usage may be less predictive than consistent engagement. The collateral representing jumlah produk. This aligns with theory: greater product bundling (cross-sell) increases loyalty and reduces churn. There is opportunity to enhance by upselling more value-added services (insurance, investment, etc.). And lastly the condition which representing bulan nonactive. This is a leading indicator of disengagement. Monitoring inactivity can help identify at-risk customers before they churn.

CONCLUSION

The binary logistic regression model using python can be utilized to analyse and identify the factors affecting the credit card churn prediction. The binary logistic regression model used in this study demonstrates strong performance in predicting credit card churn, with an accuracy of

87.1%. The confusion matrix shows 587 customers were correctly predicted not to churn, representing the largest group. This indicates the model is effective at recognizing loyal or retained customers. 25 customers were predicted to churn but did not. While this is a small number, it may lead to unnecessary retention efforts, such as incentives or interventions for customers who were not at risk. 67 customers churned but were incorrectly predicted to stay. This is a critical issue, as it suggests missed opportunities for intervention, and highlights the cost of false negatives in churn prevention strategies. 33 customers were correctly predicted to churn. These are valuable insights for the business, as these individuals can be targeted for retention programs such as personalized offers or engagement strategies. There are 4 out of 5C dimension namely character, capacity, collateral and condition significantly influence churn behaviour in this study The findings highlight the importance of combining behavioural analytics with the 5C credit evaluation model to enhance churn prediction accuracy. By focusing on engagement, transaction behaviour, product bundling, and early warning signals like inactivity, financial institutions can design targeted retention strategies and reduce churn risk effectively.

REFERENCES

Akki.or.id. (2024). Jumlah kartu kredit. *Asosiasi Kartu Kredit Indonesia*. Retrieved from https://akki.or.id/statistik/

Amin, M. (2017). Islamic banks: Contrasting the drivers of customer satisfaction on image, trust, and loyalty of Muslim and non-Muslim customers in Malaysia. *International Journal of Bank Marketing, 34*(1), 35–50.

Faisal, F., Sari, I. R., Saraswati, I., & Heikal, J. (2024). Analisis pengaruh karakteristik produk terhadap niat beli ulang pelanggan menggunakan metode regresi logistik biner. *AKADEMIK: Jurnal Mahasiswa Humanis, 4*(3), 1182–1190.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Pearson.

Hendrawan, E., Zakaria, D., Salwa, E., & Heikal, J. (2024). Customer renewal prediction for motor vehicle insurance using binary logistic regression in PT XYZ Insurance. *Innovative: Journal of Social Science Research, 4*(6), 2311–2320.

Maulana, B. A. (2025). Churn prediction in credit customers using random forest and XGBoost methods. *Indonesian Journal of Data and Science, 2*(1), 14–21.

Investopedia. (n.d.). 5 Cs of credit: What they are, how they're used, and which is most important. Retrieved June 26, 2025, from https://www.investopedia.com/terms/f/five-c-credit.asp

Mediatama, G. (2025, May 29). Dibayangi kenaikan NPL, bisnis kartu kredit perbankan masih tumbuh hingga April 2025. *Kontan.co.id*. Retrieved from https://keuangan.kontan.co.id/news/dibayangi-kenaikan-npl-bisnis-kartu-kredit-perbankan-masih-tumbuh-hingga-april-2025

Pangestuti, I., & Heikal, J. (2024). Analisis faktor-faktor yang memengaruhi keputusan pembelian kopi Torono dengan menggunakan regresi logistik biner untuk menentukan strategi pemasaran yang tepat. *Rank Research: Journal of Multidisciplinary Research and Development, 6*(5), 2173–2181.

Riyani, S., Kristianto, F., Wulandari, R., & Heikal, J. (2024). Penerapan metode regresi logistik biner dengan menggunakan Python untuk menganalisis pengaruh media sosial terhadap probabilitas pembukaan rekening pada Bank X. *Scientific Journal of Reflection: Economic, Accounting, Management and Business, 7*(2), 438–449.

Siddiqui, T. (2023). Machine learning-based customer churn prediction for banking sector: A review. *Data Science and Engineering, 8*(3), 55–63.

Suheni, S., & Heikal, J. (2024). Prediction of employee disciplinary punishment at the Department of Agriculture Payakumbuh City through approach binary logistic regression. *Journal of Business Economics and Management, 1*(2), 133–138.

Sunjaya, M. I. (n.d.). Tingkat churn tabungan pada industri perbankan. *Ekspektra: Jurnal Bisnis dan Manajemen*. Retrieved June 26, 2025, from https://ejournal.unitomo.ac.id/index.php/manajemen/article/view/1101

Watugilang, A., & Heikal, J. (2024). Pengaruh kualitas jasa servis terhadap kepuasan pelanggan perusahaan servis kalibrasi alat survey geomatika di Jakarta dengan binary logistic regression. *Indonesian Research Journal on Education, 4*(4), 79–83.

Zulfahmi, M. R. Y., & Heikal, J. (2024). Analisis prediksi financial distress perusahaan industri kimia dasar. *Jurnal Mirai Management, 9*(1), 488–505.