Analysis of German Credit Data

Cookies & Friends

Introduction

German Credit Dataset

- The data contain the label whether the credibility of an applicant is considered good or bad with 20 other attributes as possible explanatory variables for 1000 loan applicants.

- Creditability

  1 is credit-worthy

  0 is not credit-worthy

- To Find probability of default (PD) that a borrower will default on the loan before the loan is repaid in full.

Data Description

5Cs framework

- A system used by lenders to gauge the creditworthiness of potential borrowers.
- Weighs five characteristics of the borrower and conditions of the loan, attempting to estimate the chance of default and, consequently, the risk of a financial loss for the lender.
- Capacity, capital, collateral, and conditions.

5Cs framework in German Credit Data

- 20 attributes to predict PD can be classified into five categories based on the 5Cs framework in credit risk management

  1. Character : previous credit history, gender, number of previous credits in this bank , occupation, number of dependents, telephone registration, foreign worker, age

2. **Capacity:** concurrent credits, instalment rate

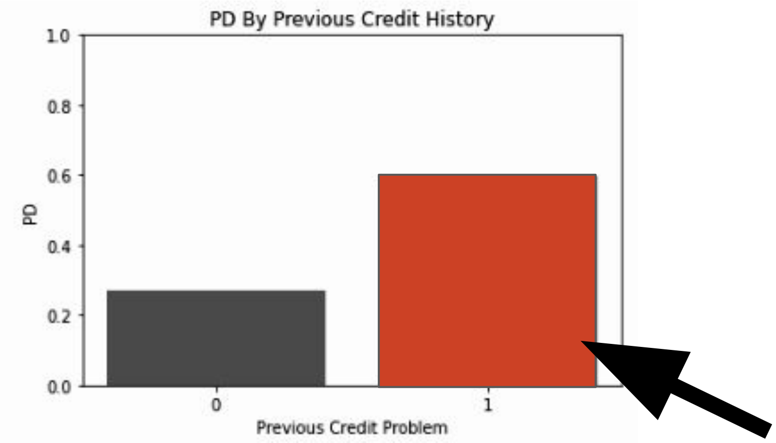
3. **Capital:** account balance, values of saving

4. **Collateral:** guarantors, duration in current address, most valuable available assets, type of apartment

5. **Condition:** purpose, credit Amount, length of current employment

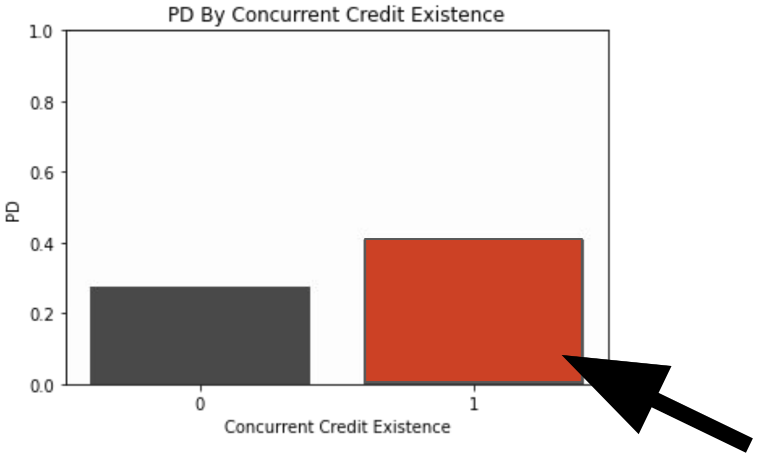
Exploratory Data Analysis

1. Character Variable



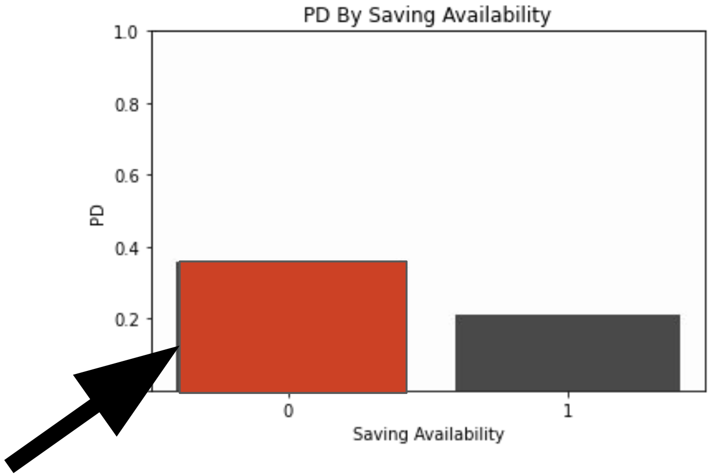
The PD of applicants with historical credit problems is much higher than the PD of applicants with no credit problems, supporting the assumption that past behaviors should reflect future behaviors.

2. Capacity Variables



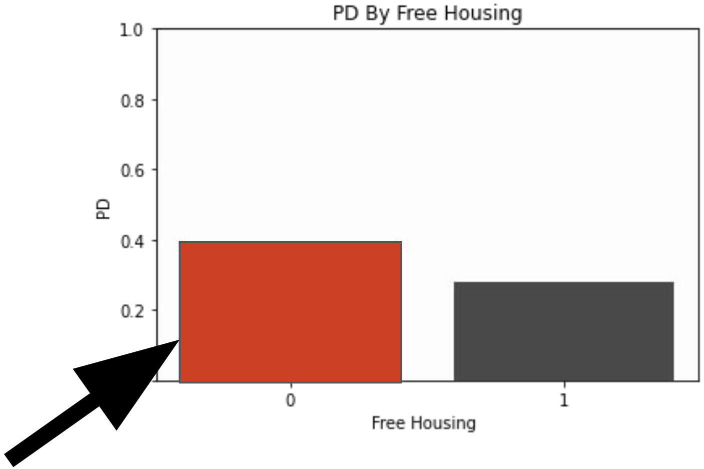
The PD of applicants with existent concurrent credit is much higher than the PD of applicants with no existent concurrent credit.

3. Capital Variables



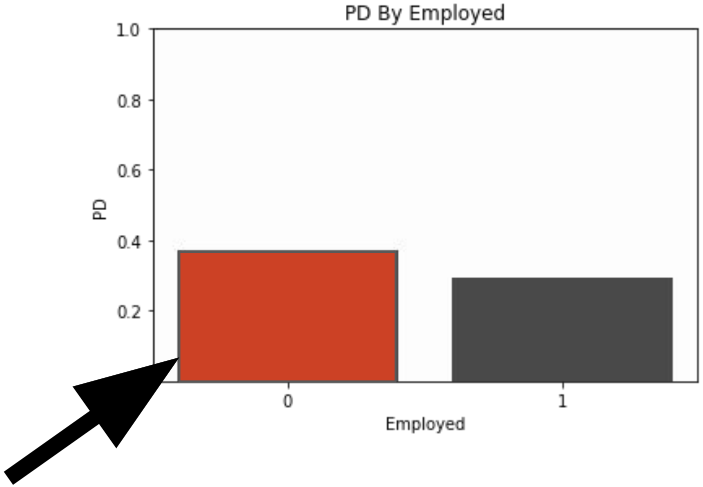
The PD of applicants with no saving availability in stocks or any values of savings is much higher than the PD of applicants who have more saving availability.

4. Collateral Variables



From the various types of apartment, such as rented flat, owner-occupied flat and free apartment. The PD of applicants with free housing is less than the PD of applicants with no free housing.

5. Condition Variables



The PD of unemployed applicants is higher than the PD of employed applicants.

Data Analysis

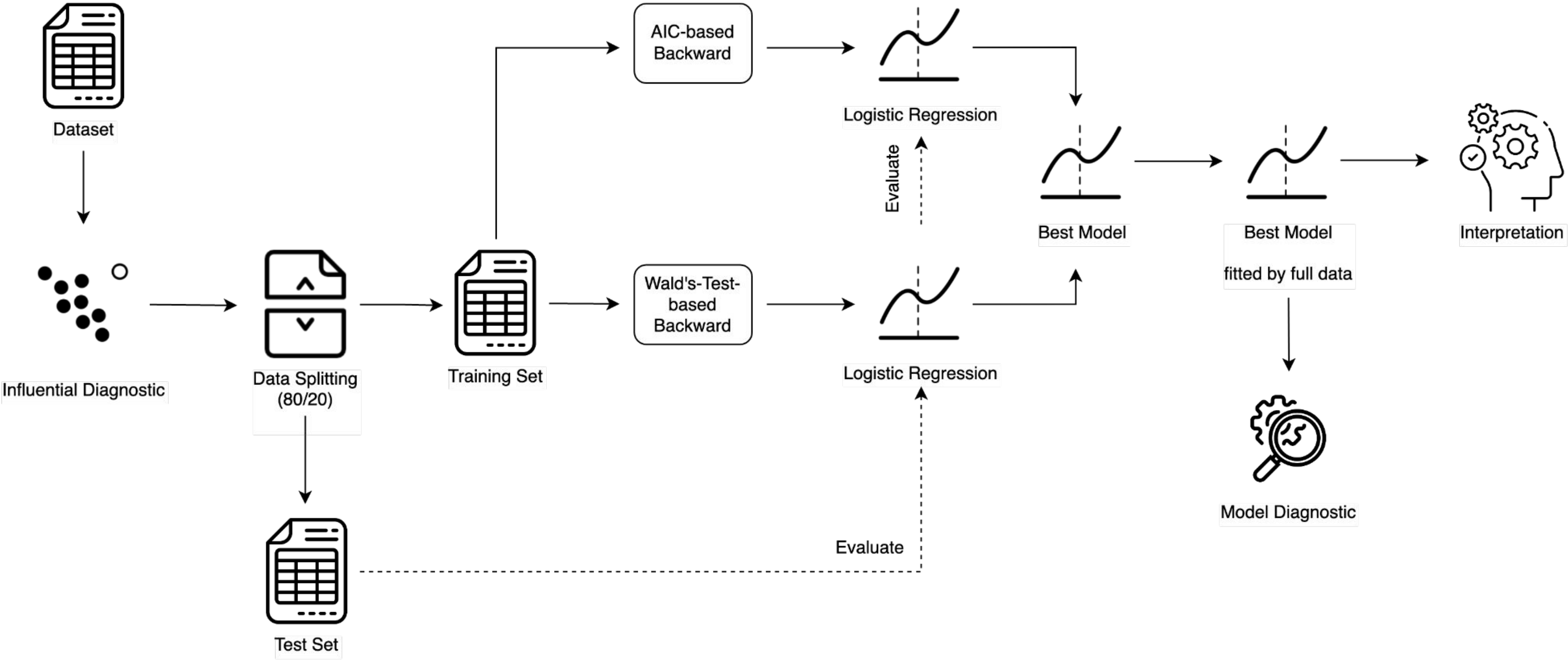
Methodology

- As our target variable is a binary response, the logistic regression model is proposed to identify the relationship between predictors and target variable (i.e. probability of default (PD)).
- The logistic regression model is defined as below:

$$logit(\pi) = \log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + ... + \beta_{p-1} x_{p-1} + \beta_p x_p$$

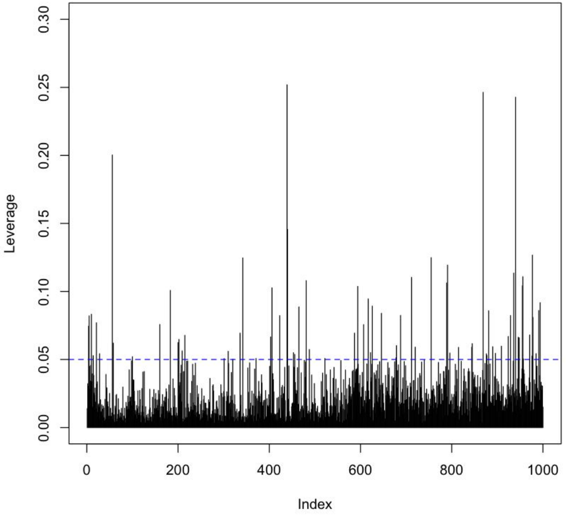
where p is the number of predictors and π is the probability of being bad credit.

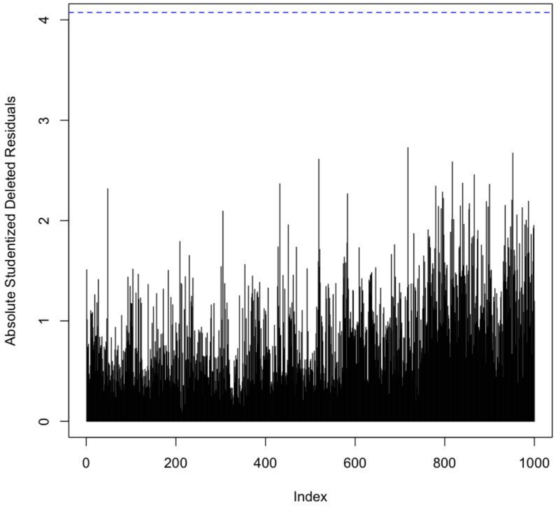
Overview of Analysis



Influential Diagnostic

- With leverage values, 74 outlying X samples are detected
- With external deleted residual, there are no outlying Y samples.
- Across 74 outliers, we found that 28 samples are influential cases by using Cook's distance computed from the full model.





Data Splitting

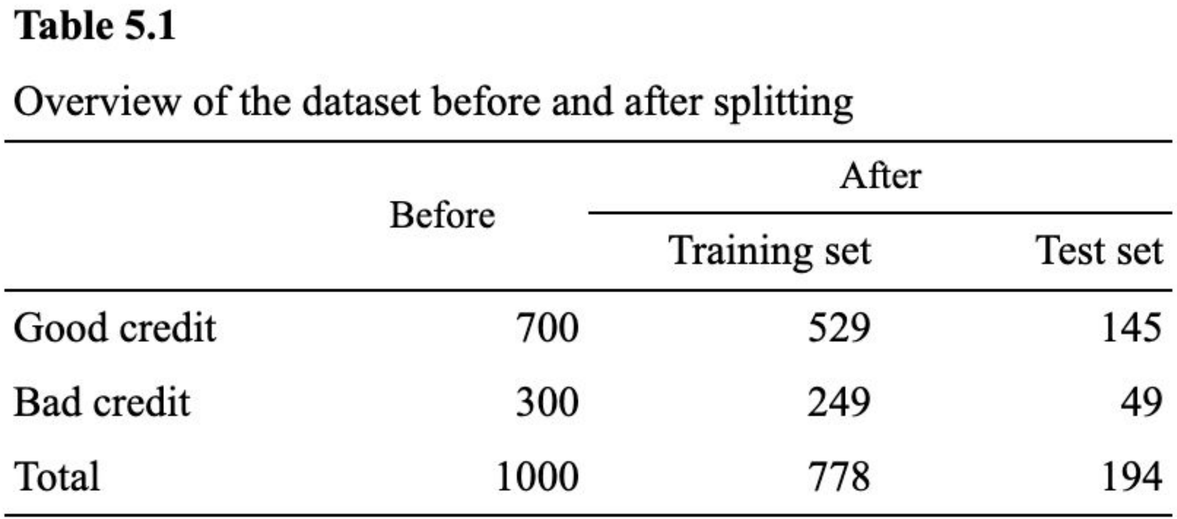


Table 5.1

Overview of the dataset before and after splitting

	Before	After	
		Training set	Test set
Good credit	700	529	145
Bad credit	300	249	49
Total	1000	778	194

Model Selection

- Model 1: Backward Stepwise regression based on AIC

  This candidate removes one variable based on AIC (Akaike information criterion) with a p-value cutoff = 0.1 at a time until no further deletion significantly improves the fit.

- Model 2: Backward Elimination based on Wald’s test

  This candidate removes one variable with the highest p-value based on Wald’s test with a p-value cutoff = 0.05 at a time until no p-value of any variable exceeds the cutoff.

Table 6.1

Comparison between model candidates in the experiment phase

Coefficient	Estimate		Std. Error		z-value		Pr(>|z|)	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
(Intercept)	-0.607	0.616	0.780	0.507	-0.779	1.215	0.436	0.224
Previous credit problem	0.937	0.966	0.301	0.299	3.118	3.227	0.002 (**)	0.001(**)
Telephone registered	-0.476	-0.427	0.193	0.192	-2.466	-2.229	0.014 (*)	0.026 (*)
Foreign worker	1.106	-	0.624	-	1.773	-	0.076 (.)	-
Current account with money	-1.345	-1.336	0.191	0.19	-7.039	-7.019	> 0.001 (***)	> 0.001 (***)
Credit amount	> 0.001	> 0.001	> 0.001	> 0.001	2.051	2.125	0.04 (*)	0.034 (*)
Duration of credit	0.024	0.027	0.009	0.009	2.578	2.973	0.010 (**)	0.003 (**)
Saving availability	-0.687	-0.675	0.188	0.187	-3.659	-3.612	> 0.001(***)	> 0.001 (***)
Instalment percent	-0.044	-0.045	0.017	0.017	-2.537	-2.622	0.011 (*)	0.009 (**)
Valuable asset availability	0.344	-	0.210	-	1.638	-	0.101	-
Concurrent credit existence	0.489	0.505	0.216	0.216	2.264	2.339	0.024 (*)	0.019 (*)
Free housing	-0.613	-0.583	0.224	0.222	-2.739	-2.629	0.006 (**)	0.009 (**)

Model 1: Backward stepwise regression based on AIC; Model 2: Backward elimination based on Wald's test

Signif. codes: > 0.001 (***) 0.001 (**) 0.01 (*) 0.05 (.)

Model Selection (cont.)

- Both models fit data sufficiently well.
- To evaluate model performance, we use 5 different types of metrics: balanced accuracy, sensitivity, specificity, F1, and AUC.

● The best model is from **backward elimination based on Wald's Test (Model 1)**.

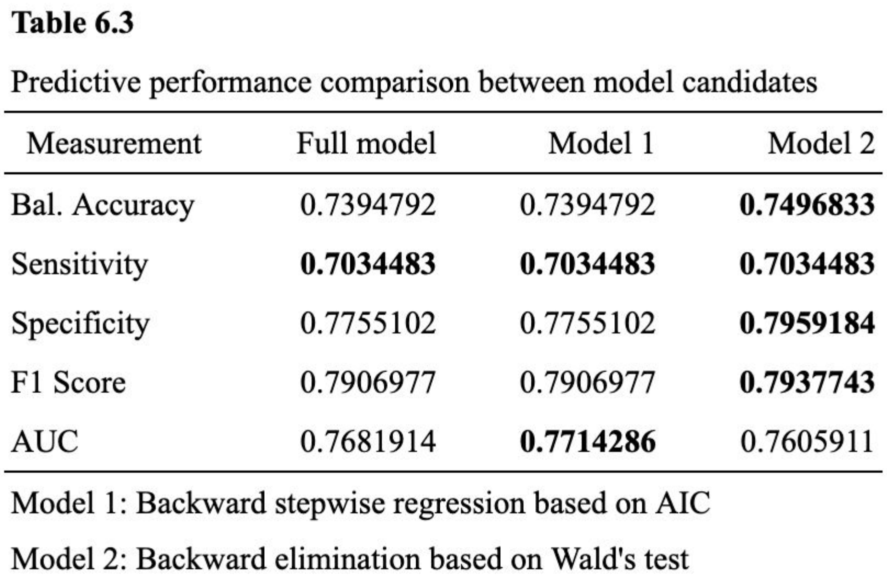


Table 6.3

Predictive performance comparison between model candidates

Measurement	Full model	Model 1	Model 2
Bal. Accuracy	0.7394792	0.7394792	**0.7496833**
Sensitivity	**0.7034483**	**0.7034483**	**0.7034483**
Specificity	0.7755102	0.7755102	**0.7959184**
F1 Score	0.7906977	0.7906977	**0.7937743**
AUC	0.7681914	**0.7714286**	0.7605911

Model 1: Backward stepwise regression based on AIC

Model 2: Backward elimination based on Wald's test

Empirical Results

- There are 9 variables in the model. All variables are separated into 2 groups.

- Group 1: 2 variables significant at 0.1, e.g., Telephone registered and Concurrent credit existence.

- Group 2: 7 variables significant at 0.05, e.g., Previous credit problem, Current account with money, Credit amount, Duration of credit, Saving availability, Instalment percent and free housing.

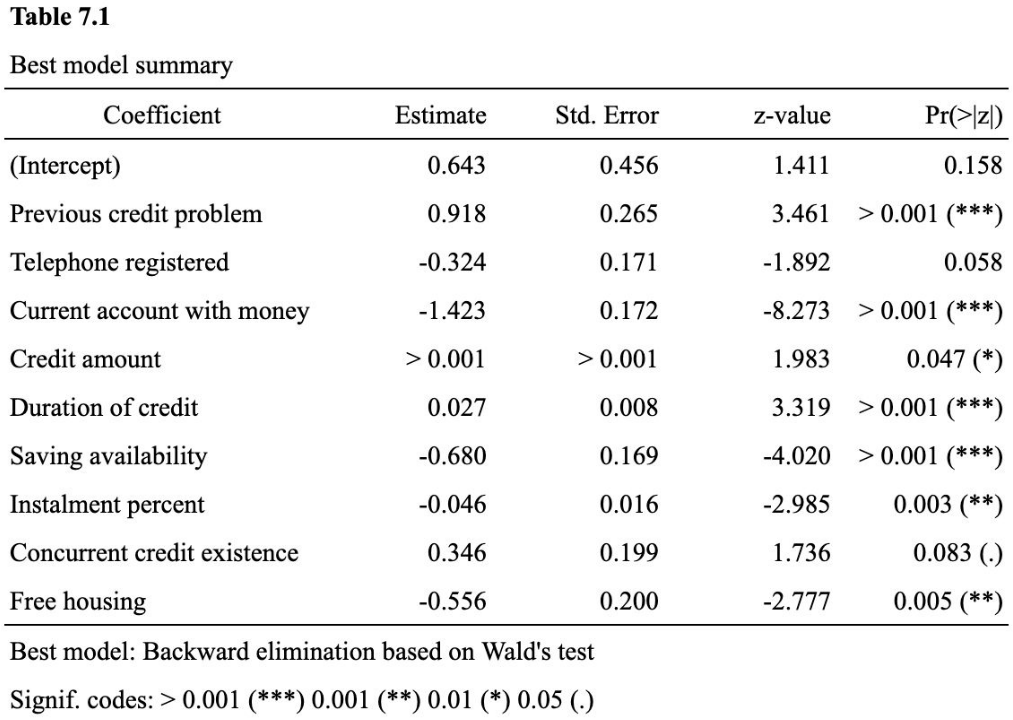


Table 7.1

Best model summary

Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	0.643	0.456	1.411	0.158
Previous credit problem	0.918	0.265	3.461	> 0.001 (***)
Telephone registered	-0.324	0.171	-1.892	0.058
Current account with money	-1.423	0.172	-8.273	> 0.001 (***)
Credit amount	> 0.001	> 0.001	1.983	0.047 (*)
Duration of credit	0.027	0.008	3.319	> 0.001 (***)
Saving availability	-0.680	0.169	-4.020	> 0.001 (***)
Instalment percent	-0.046	0.016	-2.985	0.003 (**)
Concurrent credit existence	0.346	0.199	1.736	0.083 (.)
Free housing	-0.556	0.200	-2.777	0.005 (**)

Best model: Backward elimination based on Wald's test

Signif. codes: > 0.001 (***) 0.001 (**) 0.01 (*) 0.05 (.)

Empirical Results (cont.)

- The Fitted Model : The fit of the logistic regression where π represents the probability of being bad credit (i.e.PD) is

$$logit(\hat{\pi}) = 0.643 + 0.918PCP - 0.324TEL - 1.423CAM + 0.00008CRA + 0.027DUR - 0.680SAV - 0.046INS + 0.346CCE - 0.556FRE$$

where PCP is Previous credit problem, TEL is Telephone registered, CAM is Current account with money, CRA is Credit amount, DUR is Duration of credit, SAV is Saving availability, INS is Instalment percent, CCE is Concurrent credit existence, and FRE is Free housing.

Empirical Findings

The 4 applicant’s characteristics **increasing** the probability of being a bad loaner include:

- 1 character factor which is having previous credit problems,
- 1 capacity factor which is having concurrent credit existence, and
- 2 condition factors which are higher credit amount and longer credit duration.

Empirical Findings (cont.)

The 5 applicant’s characteristics **decreasing** the probability of being a bad loaner include:

- 1 character factor which is having telephone registration,
- 1 capacity factor which is higher instalment rate,
- 2 capital factors which are higher money in current account and higher saving availability, and
- 1 collateral factor which is having free housing.

Best Model Diagnostic

- 3 types of diagnostic: Goodness of fit, Drop in deviance, and Quasi likelihood
- From Table 7.2, The best model is the standard binomial model (no overdispersion) which fits the data and is better than the full model .

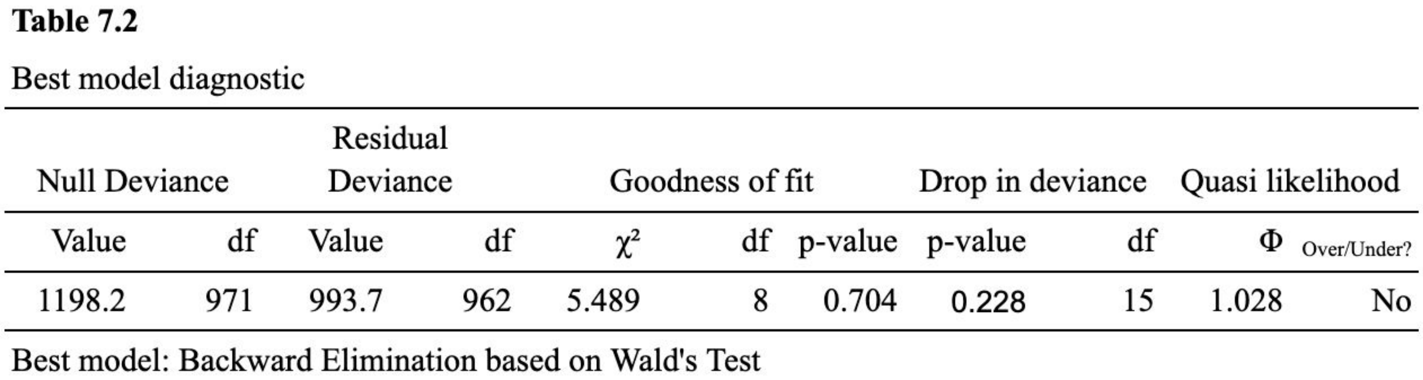


Table 7.2

Best model diagnostic

Null Deviance		Residual Deviance		Goodness of fit			Drop in deviance		Quasi likelihood	
Value	df	Value	df	χ^2	df	p-value	p-value	df	Φ	Over/Under?
1198.2	971	993.7	962	5.489	8	0.704	0.228	15	1.028	No

Best model: Backward Elimination based on Wald's Test

Conclusion

- Allowing clients to buy services on credit is common practice for most businesses. It is important to manage credit risk because that creates the chance to lenders (e.g. mostly, banks) to lose their loan.
- From our data analysis on credit risk by logistic regression with Wald’s-test-based backward elimination approach, we found that there are 9 impactful applicant’s characteristics to classify a good or bad loaner.
- With our findings, the person who is most likely to be a bad credit applicant would have previous credit problems, no registered telephone, low money in the current account, high credit amount, long credit duration, low saving availability, low instalment, other banks or dept stores, and no free housing where having previous credit problems and low money are the most impactful factors to being bad credit.

References

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Mishkin, F. S., & Eakins, S. G. (2006). Financial markets and institutions. Pearson Education India. Hofmann, Hans.

Phillips, R. L. (2018). Pricing credit products. In Pricing Credit Products. Stanford University Press.

Segal, T. (2021, September 20). 5 C's of Credit.Investopedia. https://www.investopedia.com/terms/f/five-c-credit.asp

Thank You

for

Your Attention