1. Introduction

In general, risk is defined as the possibility of bad outcomes (such as defaults) due to some unanticipated events. According to Basel Accords, four types of risks include credit risk, market risk, liquidity risk and operational risk. In personal loans, credit risk is the main component that lenders need to consider. (Phillips, 2018) Information asymmetry is considered a main problem in the credit market. Two main aspects of information asymmetry are adverse selection and moral hazard. Adverse selection happens before giving loans when lenders are unable to distinguish between good and bad borrowers. Moral hazard happens after giving loans when borrowers are involved in risky investments when borrowers know that they would not be fully responsible for the consequences. (Mishkin & Eakins, 2006) Therefore, identifying good borrowers and weeding out bad borrowers are critical missions for lenders (usually banks).

The key variable to be estimated is probability of default (PD). PD is the probability that a borrower will default on the loan before the loan is repaid in full. One possibility is that the borrower may be facing some shocks such as layoff due to some crises. Another possibility is that the borrower defaults on the loan due to financial mismanagement. However, most modern approaches are based on the assumption that default is mainly due to poor individual decisions. (Phillips, 2018) One approach is using historical data on an individual's attributes and past behaviors to build a statistical model to explain or predict PD.

2. Data Description

The dataset used in this project is German Credit Dataset provided by Prof. Dr. Hans Hofmann available on UCI Machine Learning Repository. 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. (Dua & Graff, 2019)

20 attributes to predict PD can be classified into five categories based on the 5Cs framework in credit risk management: character, capacity, capital, collateral and condition. (Segal, 2021)

Character: previous credit history, gender, number of previous credits in this bank (including this one), occupation, number of dependents, telephone registration, foreign worker, age
Capacity: concurrent credits, instalment rate
Capital: account balance, values of saving
Collateral: guarantors, duration in current address, most valuable available assets, type of apartment
Condition: purpose, credit Amount, length of current employment

3. Exploratory Data Analysis

3.1. Character Variable

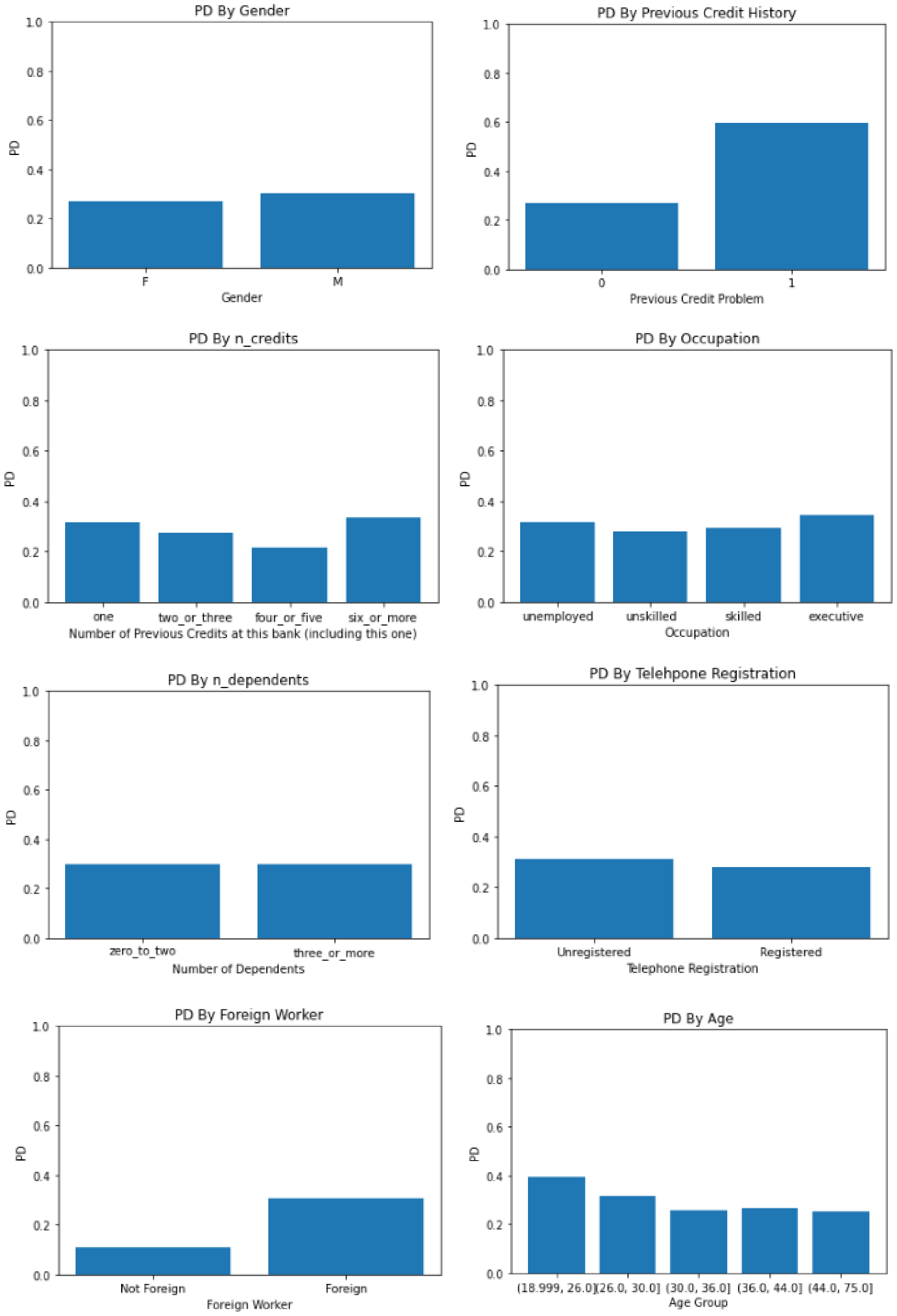

Fig 3.1: PD By Character variables

The bar graphs illustrate probability of default (PD) based on each character-related 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. The PD of male applicants is slightly higher than the PD of female applicants. The PD decreases slightly as the number of credits with this bank increases, except six or more credits. The PD of unemployed people and executives is marginally higher than the PD of unskilled and skilled workers. PD is roughly the same between two groups of applicants with different numbers of dependents. The PD of telephone-registered customers is slightly lower than the PD of customers without telephone number registration. The PD of foreign workers is much higher than the PD of non-foreign workers; however, only 3.7% of applicants are foreign workers. After dividing age into 5 groups with an equal number of applicants, the PD of older applicants is slightly lower than the PD of younger applicants, but the relationship is not strong (especially when the number of cuts increases).

3.2. Capacity Variables

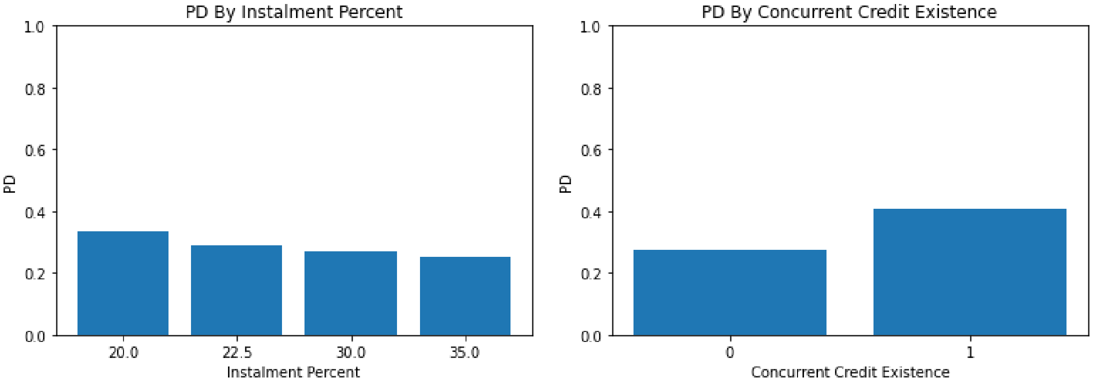

Fig 3.2: PD By Capacity Variables

The bar graphs illustrate probability of default (PD) based on each capacity-related variable. The PD of applicants with existent concurrent credit is much higher than the PD of applicants with no existent concurrent credit. After dividing instalment rate into 4 groups with an equal number of applicants, the PD of lower instalment rate of available income applicants is slightly higher than the PD of higher instalment rate applicants.

3.3. Capital Variables

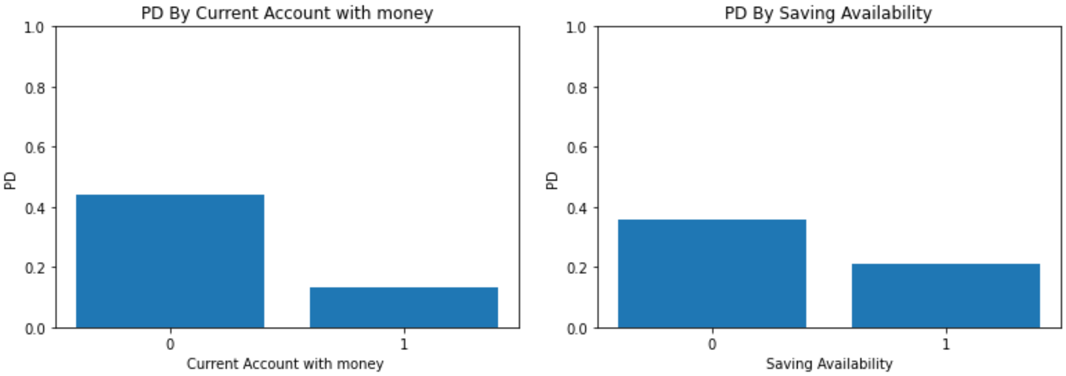

Fig 3.3: PD By Capital Variables

The bar graphs illustrate probability of default (PD) based on each capital-related variable. The PD of applicants with a current account with no money is much higher than the PD of applicants with a current account with money. Moreover, the PD of applicants with no saving availability is much higher than the PD of applicants who have more saving availability.

3.4. Collateral Variables

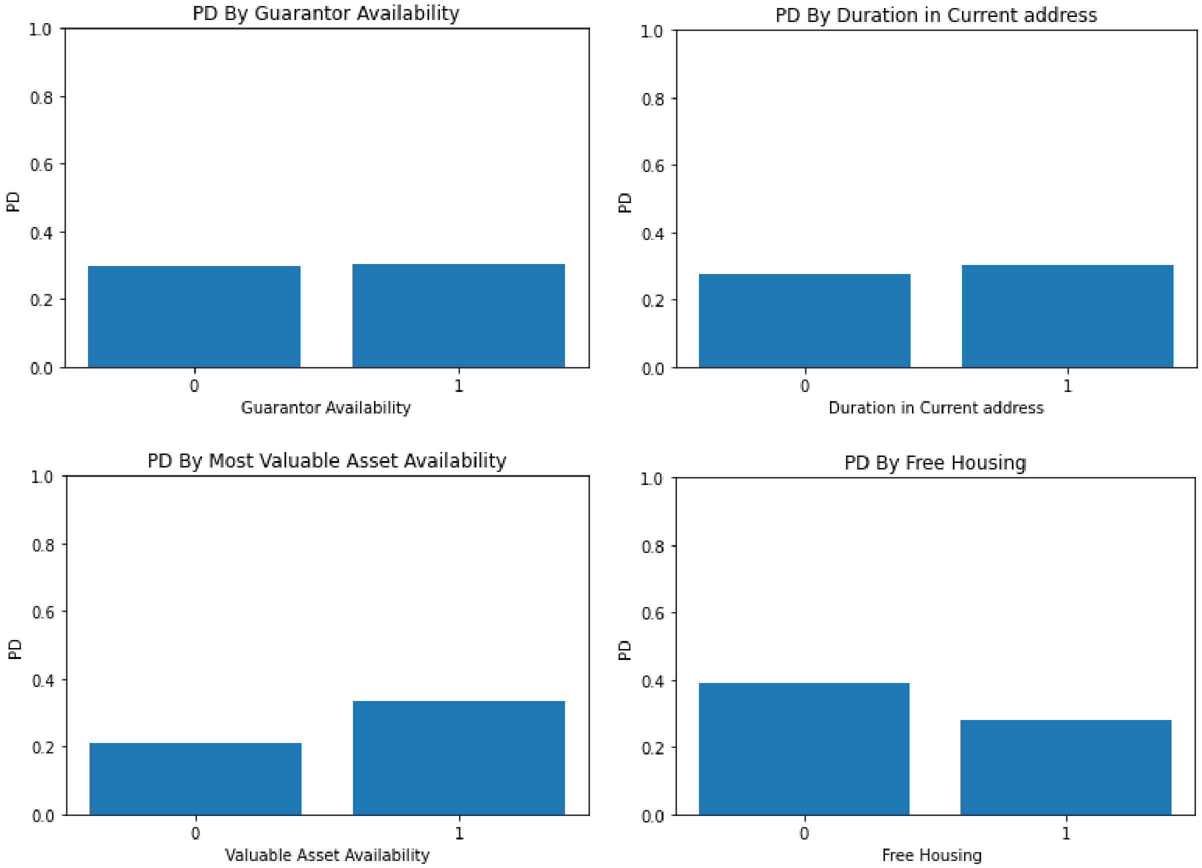

Fig 3.4: PD By Collateral Variables

The bar graphs illustrate probability of default (PD) based on each collateral-related variable. The PD of applicants with guarantors have an equal number of the PD of applicants with no guarantor. Additionally, The PD of applicants with duration in current address which is less than one year, is less than the PD of applicants with more than one year duration in current address. Moreover, The PD of applicants with more valuable asset availability is higher than the PD of applicants with no valuable asset availability. 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.

3.5. Condition Variables

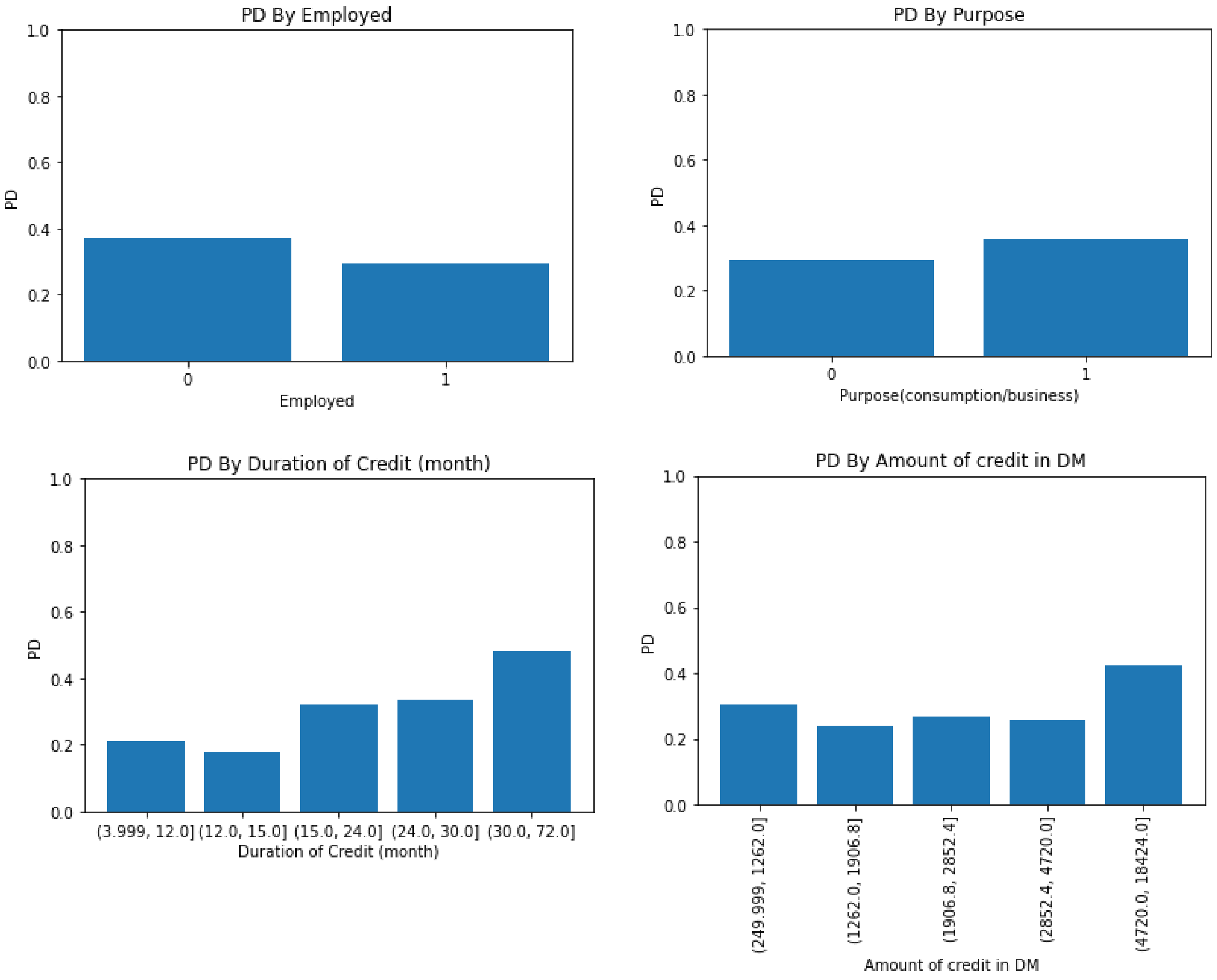

Fig 3.5: PD By Conditional Variables

The bar graphs illustrate probability of default (PD) based on each condition-related variable. The PD of applicants with the purpose for business is higher than the PD of applicants with the purpose of consumption.On the other hand, the PD of unemployed applicants is higher than the PD of employed applicants. After dividing the duration of credit into 5 groups with an equal number of applicants, the PD of lower duration of credit applicants is inconstancy lower than the PD of higher duration of credit applicants. According to the PD of applicants with the amount of credit in Deutsche Mark, the PD fluctuates from a lower amount of credit to a higher amount of credit.

4. Method

The methodology used in this project to do modeling for the prediction and investigate the relationship between the independent variables and the dependent variable (i.e. PD) is generalized linear models (GLMs). As the target variable is a binary response (i.e. 0 = good credit and 1 = bad credit), the logistic regression model is proposed to handle these tasks. 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 independent variables and π is the probability of being bad credit.

5. Experimental Setting

To get prepared for the experiment of data analysis, we do influence diagnostic and data splitting on the dataset.

5.1. Influence Diagnostic

We detect 74 outlying X samples by leverage values and no outlying Y samples by external deleted residual (see Fig 5.1). Across 74 outliers, we found that 28 samples are influential cases by using Cook's distance computed from the full model. Cook's distance for the i^{th} sample (D_i) is calculated as

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y} - \hat{Y}_{j(i)})^2}{(p+1)MSE}$$

where p is the number of independent variables and $\hat{Y}_{i(j)}$ is the j^{th} fitted value when the i^{th} sample is deleted in the fitted model. D_i that is more than 4/n is considered as an influential case. We exclude the 28 influential ones from the dataset. The dataset, thus, remains 972 samples which are 674 good loan applicants and 298 bad applicants.

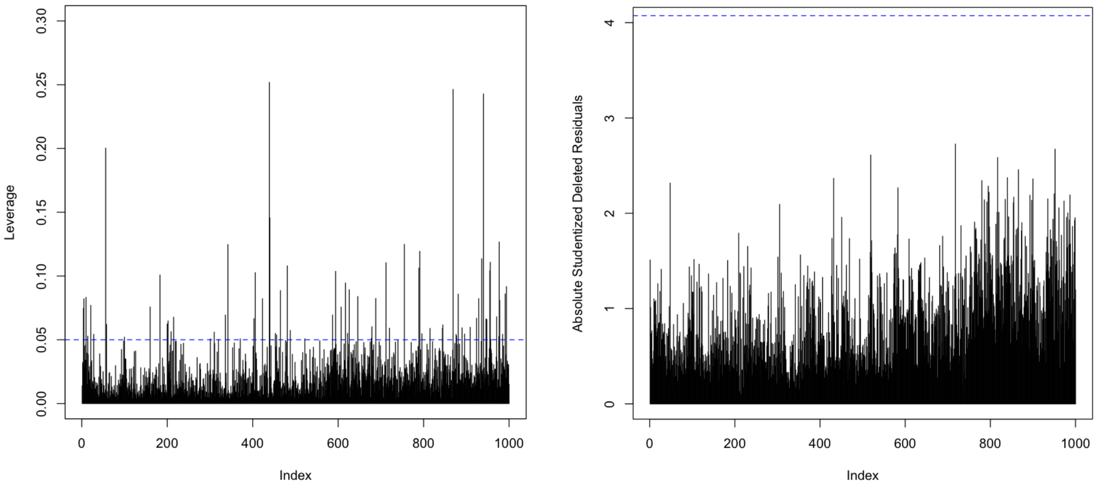

Fig 5.1: Outlying X and Y sample Detection

5.2. Data Splitting

With the remaining dataset, we do data splitting using the one split method in which 80 percent are training set and 20 percent are test set. The dataset is stratified randomly split to preserve the same proportion of the target variable in both the training and test set. Table 5.1 shows the overview of the training and test set. As can be seen, 778 samples are in the training set while 194 samples are in the test set.

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

6. Experiment

After the experimental setting, we do the experiment to find the best model. There are 2 candidates for model selection which are: 6.1.) Backward Stepwise regression based on AIC and 6.2.) Backward Elimination based on Wald's test.

6.1. 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. The AIC formula is

$$AIC = n * log(MSE) + 2(p+1)$$

where n is the number of observations and p is the number of model parameters.

6.2. 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. Wald's test is used to test the significance of individual coefficients in the model. On the basis of asymptotic normality of the MLE. Follows test statistics given by

$$z = \frac{\hat{\beta}_j}{SE[\hat{\beta}_j]}$$

The null and alternative hypotheses are stated below:

$H_0 : \beta_j = 0 \, vs \, H_1 : \beta_j \neq 0$

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 (.)

Table 6.2

Diagnostic comparison between model candidates in the experiment phase

Model	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?
Model 1	975.470	777	798.400	766	2.978	8	0.936	0.899	13	1.020	No
Model 2	975.470	777	805.370	768	15.393	8	0.052	0.523	15	1.027	No

Model 1: Backward stepwise regression based on AIC; Model 2: Backward Elimination based on Wald's Test

6.3. Comparison of model candidates

Table 6.1 shows the comparison between model candidates in the experiment phase. The table contains estimated coefficients of predictors, standard error, z-value, and p-value. As can be seen, Model 1 consists of 11 variables while Model 2 takes the same set of variables except Foreign worker and Valuable asset availability. The effect of each variable is the same direction in both models. For the significance of predictors, in Model 1, all of the predictor variables except valuable asset availability were not significant while in Model 2 all of the variables were significant.

6.4. Model Diagnostic

To diagnose these models, we use 3 different types of criteria: goodness of fit, drop in deviance, and quasi likelihood. Table 6.2 shows diagnostic comparison between model candidates in the experiment phase.

a) Goodness of fit

We use the Hosmer-Lemeshow test to test goodness-of-fit.

$$H - L\,Test\,Statistics = \sum_{i=1}^{g} \frac{(\sum_{j=1}^{m_i} y_{ij} - \sum_{j=1}^{m_i} \hat{\pi}_{ij})^2}{\sum_{j=1}^{m_i} \hat{\pi}_{ij}(1 - \sum_{j=1}^{m_i} \hat{\pi}_{ij}/m_i)} \sim \chi^2_{g-2}$$

Where y_{ij} and $\hat{\pi}_{ij}$ denote the observed and fitted values for the observation j in group i of the partition, j = 1, 2, …, m_i, i = 1, …, g.

The null, alternative hypotheses, and test criteria for this test are stated below:

H_0 : model fits the data

H_1 : not H_0

$\alpha = 0.05$

0.936 and 0.052 are the p-value from Model 1 and Model 2 respectively, which are greater than 0.05, therefore, they fail to reject H_0. This means that both models fit the data sufficiently well.

b) Drop in deviance

We use the drop in deviance test to find the model between the reduced model and the full model. The general form of the drop in deviance test statistic is

$$LRT = -2\log(LMAX_{reduced} + 2\log(LMAX_{full})) \sim \chi^2_{df}$$

where df is the difference between the number of parameters in the full and reduced models.

The null, alternative hypotheses, and test criteria for this test are stated below:

H_0 : Prefer reduced model

H_1 : Prefer full model

The p-value of Model 1 is 0.899 which is greater than 0.05, therefore, we fail to reject H_0 and it means that it prefers the reduced model (Model 1). The p-value of Model 2 is 0.532 which is greater than 0.05, therefore, we fail to reject H_0 and it means that it prefers the reduced model (Model 2).

c) Quasi likelihood approach

We consider the dispersion parameter (Φ) of both models to identify occurrence of overdispersion and underdispersion if Φ is greater than 1 we got overdispersion, in contrast if Φ is less than 1 we got underdispersion. The estimated value of the dispersion parameter is defined by:

$$\hat{\phi} = \frac{X^2}{n-(p+1)}$$

where X is Pearson's residual and p is the number of covariates in the model.

Φ of Model 1 is equal to 1.020 and that of Model 2 is 1.027. Both are likely to 1, therefore, we can conclude that they are the standard binomial model.

6.5. Prediction

To evaluate model performance, we use 5 different types of metrics: balanced accuracy, sensitivity, specificity, F1 score, and AUC.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}, Sensitivity = \frac{TP}{TP+FN}, Specificity = \frac{TN}{TN+FP}$$

$$F1\,Score = \frac{2TP}{2TP+FP+FN}$$

Table 6.3 shows that comparison between model candidates. Both Model 1 and Model 2 win the full model. Over 5 measurements, Model 2 wins Model 1 in 3 measurements which are balanced accuracy, specificity, and F1 score, both achieve the same sensitivity, while Model 1 wins Model 2 in only AUC. Therefore, Model 2 is the best model.

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

7. Empirical results

Finally, we got the best model by backward elimination (Wald's Test with p-value cutoff = 0.05). In this part, it shows the model summary, fitted model, interpretation, and model diagnostic as below.

7.1. Best model summary

For the best model, there are 9 variables. Table 7.1 shows the overall result contains variables, coefficient, standard error, z-value and p-value. There are 2 variables significant at 0.1, e.g., Telephone registered and Concurrent credit existence. And there are 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 (.)

7.2. 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.

7.3. Model Interpretation

Interpreting coefficients β:

$$\log(\pi) = \log(\omega) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

$$\omega = exp\{\beta_0 + \beta_1 x_1 + ... + \beta_9 x_9\}$$

where p is the number of independent variables of the best model (i.e. 0,...,9), π is the probability of being bad credit, ω is odds. The odds increase/decrease multiplicatively by $exp\{\beta_j\}$ for every 1-unit increase/decrease in $_1$ given fixed values of the other $_1$ given fixed values of the other.

From the fitted model with the fit of the logistic regression where π represents PD probability. We can interpret the solution as below.

a) Previous credit problem

Since the p-value (i.e. 0.001) is less than 0.05, we reject H_0. We can conclude that there is enough evidence that Previous credit problem is related to PD at the 0.05 level of significance.

Interpretation: the odds increases multiplicatively by $exp\{0.918\}$ for having Previous credit problems given fixed values of the other predictor variables.

b) Telephone registered

Since the p-value (i.e. 0.0585) is not less than 0.05, we reject H_0. We can conclude that there is enough evidence that Telephone registered is related to PD at a 0.1 level of significance.

Interpretation: the odds decreases multiplicatively by exp{-0.324} for Registered telephone, given fixed values of the other predictor variables.

c) **Current account with money**

Since the p-value (i.e. > 0.001) is less than 0.05, we reject H_0. We can conclude that there is enough evidence that Current account with money is related to PD at a 0.05 level of significance.

Interpretation: For comparing two Current account with money, the odds decreases multiplicatively by exp{-1.423} for every 1-unit increase in current account with money, given fixed values of the other predictor variables.

d) **Credit amount**

Since the p-value (i.e. 0.047) is less than 0.05,we reject H_0. We can conclude that there is enough evidence that credit amount is related to PD at a 0.05 level of significance.

Interpretation: For comparing two Credit amount, the odds increases multiplicatively by exp{ 0.00008} for every 1-unit increase in Credit amount, given fixed values of the other predictor variables.

e) **Duration of credit**

Since the p-value (i.e. > 0.001) is less than 0.05, we reject H_0. We can conclude that there is enough evidence that Duration of credit is related to PD at the 0.05 level of significance.

Interpretation: For comparing two Duration of credit, the odds increases multiplicatively by exp{0.027} for every 1-unit increase in Duration of credit, given fixed values of the other predictor variables.

f) **Saving availability**

Since the p-value (i.e. > 0.001) is less than 0.05, we reject H_0.We can conclude that there is enough evidence that Saving availability is related to PD at a 0.05 level of significance.

Interpretation: the odds decrease multiplicatively by exp{-0.680} for having saving availability given fixed values of the other predictor variables.

g) **Instalment percent**

Since the p-value (i.e. 0.003) is less than 0.05, we reject H_0. We can conclude that there is enough evidence that Instalment percent is related to PD at a 0.05 level of significance.

Interpretation: the odds decreases multiplicatively by exp{-0.046} for having instalment, given fixed values of the other predictor variables.

h) Concurrent credit existence

Since the p-value (i.e. 0.083) is not less than 0.05 but p-value is less than 0.1, we reject H_0. We can conclude that there is enough evidence that Concurrent credit existence is related to PD at a 0.1 level of significance.

Interpretation: the odds increases multiplicatively by exp{0.346} for having concurrent credit existence, given fixed values of the other predictor variables.

i) Free housing

Since the p-value (i.e. 0.005) is less than 0.05, we reject H_0.We can conclude that there is enough evidence that Free housing is related to PD at a 0.05 level of significance.

Interpretation: the odds decrease multiplicatively by exp{-0.556} for every 1-unit decrease in $x1$ given fixed values of the other predictor variables.

- **Conclusion on interpretation**

Among 9 factors, Previous credit problem and Current account with money that are the most effective factors on PD. To conclude in non-technical terms, there are 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. On the other hand, 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. Therefore, the person who would be predicted as a bad credit applicant (i.e. who is not likely to repay the loan causing approving the loan to the person results in a financial loss to the bank) the most, is a person with previous credit problems, no telephone registered, low money in current account, high credit amount, have long credit duration, low saving availability, low instalment, have other banks or dept stores, and no free housing.

7.4. The best model diagnostic

Table 7.2 shows the best model diagnostic.To diagnose the best model, we use 3 different types of criteria: goodness of fit, drop in deviance, and quasi likelihood .

a) Goodness of fit

The p-value is 0.704 which is > 0.05, therefore, this model fits data sufficiently well.

b) Drop in deviance

The p-value is 0.228 which is > 0.05, therefore, the data prefers reduced model.

c) Quasi likelihood

The dispersion parameter (Φ) for the best model is 1.028. Therefore, the best model is the standard binomial 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

8. Conclusion

Most businesses function by allowing clients to purchase goods or services on credit and pay for them later. This, however, creates the risk to lenders (e.g. banks) to lose the principal of their loan. Therefore, it is important for them to manage their credit risk. In this project, we, thus, modeled credit risk and investigated the factors for the probability of default (i.e. PD). We did exploratory data analysis and data preprocessing to prepare the data for modeling. To model the problem, we deployed a logistic regression methodology. After the experiment of analysis, we found that Wald's-test-based backward elimination approach achieves the best model measured by the classification metrics. Our model fits the data well and doesn't have any problems.

With data analysis, there are 9 factors affecting the probability of being bad credit including Previous credit problem, Telephone registered, Current account with money, Credit amount, Duration of credit, Saving availability, Instalment percent, Concurrent credit existence, and Free housing. In addition, Previous credit problems and Current account with money are the most impactful factors to being bad credit. To be concluded, 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.

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