University of Warsaw
Faculty of Economics Science

Artur Skowroński
Index no: 411423

Determinants of the probability of default – Binary choice model

A credit-based paper
for the
Advanced Econometrics course
conducted by Dr Rafał Woźniak.

Warsaw, June 2022

Abstract

This paper presents an analysis of Binary Choice Models that estimate the probability that bank customers will default on a loan. The data came from the "Give me some credit!" competition available on kaggle.com. My study found that significant variables that affect the probability of falling into default include earnings, past default and age.

Keywords: econometric model, credit risk, probability of default

Table of contents

INTRODUCTION

Credit risk is the risk to a bank that a customer will fail to repay all or part of an obligation. It is important for financial institutions to assess this risk in the best possible way in order to limit potential financial losses. There are many types of risk, the most important of which focus on 3 areas: market, operational and insolvency.

In this paper, I will focus on the latter aspect, and in particular on the so-called Probability of Default (PD), which is nothing other than the probability that a customer will not repay his loan. It is also worth remembering that in the Revised international capital framework Basel III, which is the most important regulatory document on credit risk, other risk parameters such as Loss Given Default (LGD) and Exposure of Default (EAD) were distinguished[1]. Multiplying all 3 parameters, gives the expected financial loss.

Nevertheless, the aim of this paper will be mainly to focus on the properties and the diagnosis of the model estimating the probability of falling into default, rather than on the correct estimation of the results on the test set. In the first part of the paper, the literature that has addressed the topic of credit risk estimation will be introduced, where it explains which variables may be important to include. A detailed review of the literature, will provide some baseline for my results, which will help me define whether my results are in line with other researchers.

In the next chapter, an empirical study will be conducted, where the data analysis and data processing will be made. The best possible data set for modelling will be presented step by step. Chapter 3, is strichte oriented to the modelling process itself and the selection of a suitable model for the processed data. In the last chapter, the diagnostics of the selected model will be conducted and the obtained results interpretation.

[1] European Banking Authority, (2019). Policy advice on the Basel III reforms: credit risk. Standardised approach and IRB approach.

1. LITERATURE REVIEW

Until half a century ago, when people did not have access to computers or other devices capable of processing complex operations and calculations, financial institutions granted credit on the basis of subjective assessments by bank experts[2]. However, one of the first attempts to construct such a "scoring card" dates back to 1941 and took into account 6 variables: the applicant job or position, the number of years spent in the current position the number of years spent living at the current address, details on bank accounts, life insurance policies, gender and the amount of the monthly installment (Durand, 1941).

Obviously, nowadays risk assessment on the basis of gender or place of residence is regarded as discriminatory. Secondly, it is now a much more complex and regulated process. Nevertheless, it is worth noting that already back then, so-called behavioural variables were used in the analysis, which were based, for example, on a customer's credit history[3].

In order to examine whether a customer is a healthy customer, primarily current financial data such as earnings or savings were used or other financial indexes (Beaver, 1966), as well as purely abstract variables such as reputation, which would currently be difficult to measure for individual customers (nowadays for corporate customers, Moody's ratings can be used[4]). Undoubtedly a breakthrough in the literature on risk estimation is the definition of the model calculating the Z-score (Altman, 1968), which with an accuracy close to 94% was able to predict whether a given company will go bankrupt within 24 months. This model was based on 5 financial variables, and the Z-score itself was divided into 1 final buckets, to which we can briefly assign values in the form of adjectives: safe, at risk, dangerous. In later years, this model lived to see a few of its versions, so that in an even more efficient way to calculate the result for clients.

There are currently many models or algorithms in the literature for estimating credit risk, ranging from simple statistical models to models based on ensembling or neural networks. In all likelihood, the latter methods are best able to estimate probability of default, but are difficult to explain ('black box models')[5]. Therefore, logit or probit models still remain as a kind

[2] Altman, E.I., Saunders, A., 1998. Credit risk measurement: Developments over the last 20 years. Journal of Banking & Finance 1721-1742.

[3] Anderson, R., 2007. The Credit Scoring Toolkit : Theory and Practice for Retail Credit Risk Management and Decision Automation. Oxford University Press.

[4] https://www.moodys.com/sites/products/productattachments/ap075378_1_1408_ki.pdf

[5] Gouvêa, M.A., Gonçalves E.B., 2007. Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models. POMS 18th Annual Conference.

of "best practice" due to their actually simplicity of operation, which makes it possible to explain exactly what influenced the credit decision for a given customer. Interestingly, one may even be tempted to use panel models for the determination of risk, but one has to reckon with the fact that the estimated parameters may be biased and inconsistent, which will definitely affect the correct interpretation of the model[6].

One of the more interesting studies on the use of logistic regression in estimating the probability of default concerns clients of a Portuguese financial institution (Silva, Lopes, Correia, Faria, 2020). For the analysis, the researchers used 3221 individual customers, 10% of whom defaulted. The Hosmer-Lemeshow test showed that the model had the correct functional form, and the model correctly predicted 89.79% of the cases. The most important factors influencing the form of the model were interest rate spread, age of borrower (in years), length of loan, number of credit cards and salary[7].

Obviously, there are also some improvements to logistic regression that positively affect the efficiency of default probability estimation. One such technique is the use of random coefficients in the regression, where the coefficients assume a multivariate normal distribution (Dong, Kai, Yen, 2012)[8]. In their empirical study the authors used the German Credit Data Set from University of California (UCL) and compared two regressions: fixed coefficients and random. The PCC measure was used to test the predictive power of the models, where the second model was found to be 3 percentage points better overall (71% vs 74%). Thus, we see that some variation can definitely have an impact on the final model performance.

In the case of the probit model, one of the most interesting studies is undoubtedly based on corporate clients from US banks, where only financial variables such as asset quality, profitability, capital adequacy were used for risk estimation (Gurný, Gurný, 2013)[9]. Based on the likelihood-ratio test the model is significant, which was later confirmed by the Wald test. The probit model had an AUC_ROC of 82%.

[6] Louzis, D.P., Vouldis, A.T., Metaxas, V.L., 2010. Macroeconomic and bank-specific determinants of non-performing loans in Greece: a comparative study of mortgage, business and consumer loan portfolios. Bank of Greece.

[7] Silva, E.C., Lopes, I.C., Correia, A., Faria, S., 2020. A logistic regression model for consumer default risk. Journal of Applied Statistics, 47:13-15.

[8] Dong, G., Lai, K.K., Yen, J., 2010. Credit scorecard based on logistic regression with random coefficients. International Conference on Computational Science.

[9] Gurný, P., Gurný, M., 2013. Comparison of credit scoring models on probability of default estimation for US banks. Prague Economic Papers 2: 163-181.

2. EXPLORATORY DATA ANALYSIS

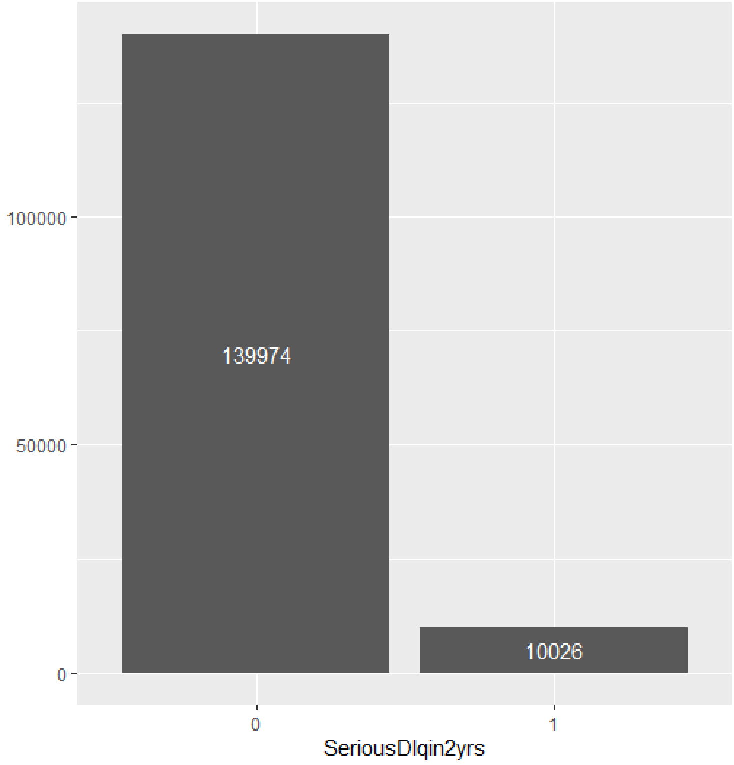
2.1. Dataset

The data for the model was obtained from a competition called "Give Me Some Credit!", from the website Kaggle.com, which is the largest online platform for users who are interested in econometrics and data science . For research purposes, only the training dataset was used, whose size is 150000 rows x 10 columns.

Dependent variable:

- **SeriousDlqin2yrs** – Binary feature. Person experienced 90 days past due delinquency or worse. Default rate is equal to 0.06684, so my target variable is definitely unbalanced, which follows the logic.

Figure 1. Distribution of dependent variable - SeriousDlqin2yrs.



Independent variables:

- **RevolvingUtilizationOfUnsecuredLines** – Numercial feature saved in percentage. Ratio of money owed to credit limit. Contains 125728 unique values from 0 to 50708.

- **Age** – Numerical feature. Age of borrower in years. The oldest borrower was 109 years old, while the youngest 0 years old.

- **NumberOfTime30-59DaysPastDueNotWorse** – Numerical feature. Number of times borrower has been 30-59 days past due but no worse in the last 2 years. The record holder, was in default 98 times.

- **DebtRatio** – Numerical feature saved in percentage. Monthly debt payments, alimony, living costs divided by monthy gross income. Contains 114194 unique values from 0 to 329664.

- **MonthlyIncome** – Numerical feature. Defines monthly income of each borrower. The richest person was earning 3008750.

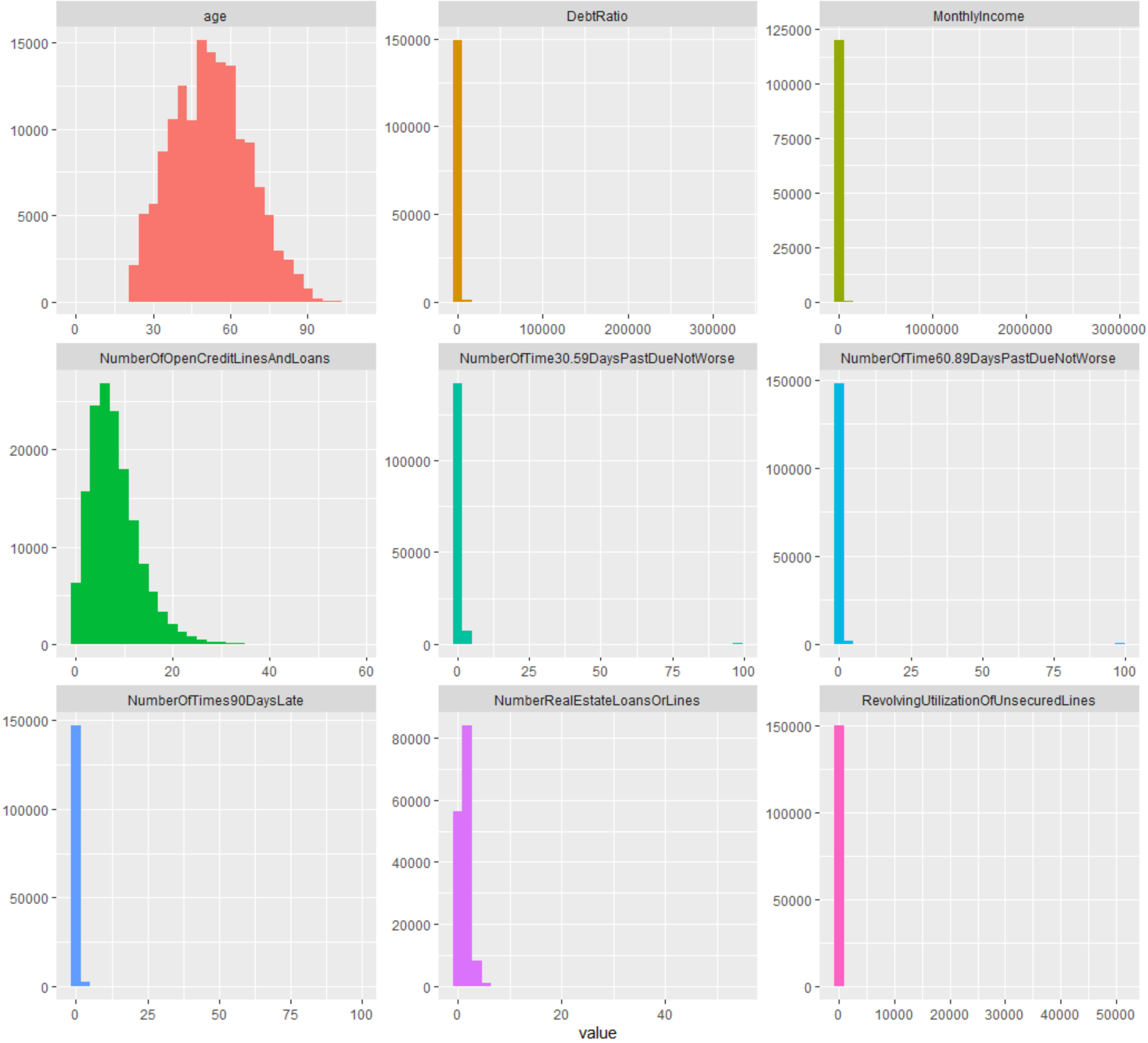
- **NumberOfOpenCreditLinesAndLoans**. Numerical Feature. Number of Open loans (installment like car loan or mortgage) and lines of credit (e.g. credit cards). The median open loans per customer is 8.

- **NumberOfTimes90DaysLate** – Numerical feature. Number of times borrower has been 90 days or more past due. Again, the record holder was in default 98 times.

- **NumberRealEstateLoansOrLines** – Numerical feature. Number of mortgage and real estate loans including home equity lines of credit. Range of this feature is between 0-54.

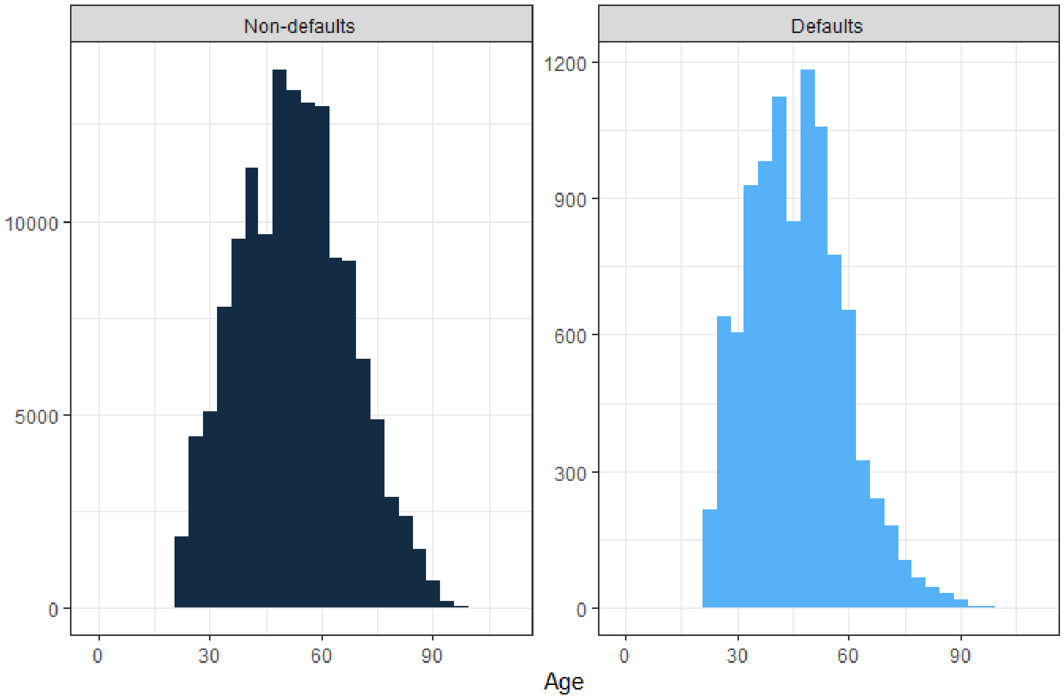
- **NumberOfTime60-89DaysPastDueNotWorse** - Numerical feature. Number of times borrower has been 60-89 days past due but no worse in the last 2 years. As above, the record holder has been defaulter 98 times.
- **NumberOfDependents** – Numerical feature. Number of dependents in family excluding themselves (spouse, children etc.). As many as 60% of clients were single or had no children or simply did not share this information.

Figure 2. Distributions of independent variables.



Next, the age distribution in relation to defaults was checked, because there was a risk that there would be a leftward asymmetry for defaults, i.e. a tendency to become insolvent only in elderly people, which would certainly be wrong. If such a situation were to occur, it could indicate that the data are highly unrealistic, because one could immediately ask: "Why then are we giving credit to the elderly if they are not repaying it anyway"?

Figure 3. Distribution of age in relation to SeriousDlqin2yrs variable.



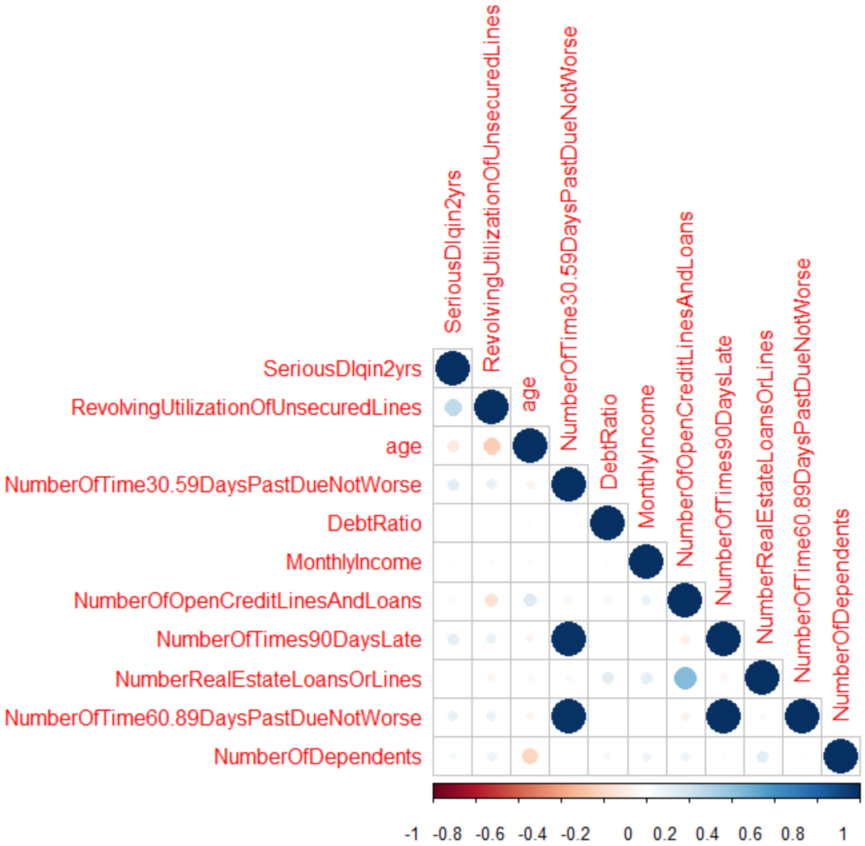
The age distribution is similar in both cases. This is a good sign because it should be remembered that, in most cases, younger people have a better chance or possibility of repaying the loan because, for example, it can be spread over a greater number of years (which means more instalments but smaller amounts).

In the next step, it was decided to analyse, if there were **missing data**. With such a large data set (as a reminder 150000 rows), it seems to be impossible that all the columns are filled. Such a case came true for two columns: MonthlyIncome - 29731 NA values, NumberOfDependents - 3924 NA. Without doing any statistical testing, it was concluded that there are certainly outliers or even wrong values in my data. Therefore, it was decided to fill in the missing data:

- median – for the MonthlyIncome variable, which is an outlier-proof measure
- mode – for the NumberOfDependents variable.

Having already a full data set, it was decided to check the relationships between all the variables. Due to the fact that they are mostly continuous variables (the exception being the dependent variable) and the interest is around linear relationships between variables, Pearson’s correlation was used.

Figure 4. Pearson correlation between all of the variables



From the above graph, it can be seen that the customer credit history variables are highly correlated with each other. Therefore, at this stage it was decided to create two additional variables that are dependent on the 3 variables with the prefix "NumberOfTime":

- default_flag – binary variable. Has the client ever been in default before?
- max_default_occurences – numeric variable. Maximum number of appearances in default.

Then the variables with the prefix "NumberOfTime" were removed. The correlations were rechecked and everything looked much better now. It is also worth noting the correlation between the variables NumberRealEstateLoansOrLines and NumberOfOpenCreditLinesAndLoans (0.43). Nevertheless, it was not considered to be crucial at this point.

2.2. Outliers

At first, the age variable was taken care of, due to the fact that there was one case where the person was a newborn baby. It was decided, that this was an incorrect value (the next value was 21 years) and therefore, this observation has been removed. In the case of persons older than 100 years, it was decided that such a situation, despite being extremely rare, could occur. From the upper limit of the variable, nothing has been changed.

Then the credit history of the clients has been analysed. To my surprise, as many as 264 customers defaulted equally 98 times. The attention was also caught by the fact that the next values in order were: 96 times (5 customers) and 17 times (1 customer, default +90 days). Therefore it was considered all values above 90 as outliers, artificially assigning them a value of 20. This allows to reduce the range of variable and not treat the outliers as extreme values, but only preserve their sense of being outliers.

The attention was also drawn to the variable RevolvingUtilizationOfUnsecuredLines, where as many as 3323 people owed more than they earned. Moreover, as many as 254 people owed 5 times what they earn. It was considered that these were certainly outliers and to any such people, it was decided to assign them a value of 5 in advance.

Then the MonthlyIncome variable has been analysed. In the case of this variable, it was interesting that 605 people earn $1 per month and 1634 earn $0 (in which case calculating the DebtRatio variable is impossible). So, it was considered that all people earning less than $100 to be currently unemployed and artificially assigned them a value of 1. Ultimately, 2277 people were considered unemployed of which DebtRatio > 0 has 2134 people.

2.3. Additional features

In the last step of data preparation, it was decided to add new variables to the analysis. Due to the right-sided asymmetry of the variables NumberOfOpenCreditLinesAndLoans and MonthlyIncome (figure 1), the logarithmic transformation was implemented. It is worth noting that in the case of the first of the mentioned variables, due to the occurrence of 0 values, the log(1+x) formula was used.

Then, additional binary variables were created:

- dangerous_clients - defining clients who owe money but are unemployed at the same time.
- open_loan_on_pension – customers aged 65 and over who have open credit accounts

Therefore, the final sample for modelling consists of 149999 observations and 12 variables. A total of 2 transformations and 3 interactions were created.

3. MODELLING PART

3.1. Research hypotheses

With reference to the literature presented above and based on my own observations, it was decided to distinguish 4 different hypotheses based on relevant aspects in risk modelling.

- Main hypothesis: Customers who have been in defaults at least once in their credit history are more likely to fall into default again.
- First side hypothesis: Earnings have a positive effect on the probability of being in default.
- Second side hypothesis: As age increases, a person is more likely to default on a loan.
- Third side hypothesis: The more lines of credit a customer has, the more likely they are to be insolvent.

3.2. Selection of optimal model

The following equation for preliminary considerations was used:

$$\begin{gathered}\text{SeriousDlqin2yrs}_i = \beta_0 + \beta_1\text{RevolvingUtilizationOfUnsecuredLines} + \beta_2\text{age} \\ + \beta_3\text{DebtRatio} + \beta_4\text{NumberRealEstateLoansOrLines} + \beta_5\text{NumberOfDependents} \\ + \beta_6\text{default\_flag} + \beta_7\ \text{max\_default\_occurences} \\ + \beta_8\log(\text{NumberOfOpenCreditLinesAndLoans}) + \beta_9\log(\text{MonthlyIncome}) \\ + \beta_{10}\text{age}^2 + \beta_{11}\text{open\_loan\_on\_pension} + \beta_{12}\text{dangerous\_clients} + \varepsilon_i\end{gathered}$$

Three models were performed in order to get estimations: logit, probit and OLS for binary dependent variable (Linear Probability Model) using White's robust matrix. Due to the limitations of the LPM model, not much emphasis was put on the interpretation of this model. However, for the logit and probit model, measures of information criteria were calculated: AIC and BIC in order to optimally select the model for later considerations. Eventually, due to lower values for both metrics, **it was decided to choose the probit model for further analysis**.

Table 1. Information criteria for the logit and probit model.

	AIC	**BIC**
Logit	57566.87	57695.81
Probit	57265.94	57394.88

3.3. General to Specific procedure

In order to test the combined insignificance of the excluded variables for the probit model the Likelihood Ratio Test was calculated, which is a hypothesis test that helps to choose the "best" model between two nested models[10] . Below are the hypotheses, which were used to remove the variables with the highest pvalue. Significance level I assumed at α=0.05:

$$H_0: \beta_x = 0$$
$$H_1: all\ variables\ are\ jointly\ significant$$

Therefore, the aim is to prove that the parameter of a given variable is equal to 0. In that case, non-significant variables might be removed from the model because they do not contribute any relevant information. The variable elimination procedure was carried out only twice was.

In the first case, the significance of the age2 variable was checked, because the pvalue for it was the highest at 0.7586. Likehood Ratio Test was performed to verify the significance for this variable, comparing the base model with the model without this one variable. The pvalue for this test was 0.7539, so the null hypothesis cannot be rejected. The age^2 variable was excluded from the model.

For the new model, it was checked again what the pvalue distribution looks like for the variables and it was found that the variable DebtRatio is the new variable with the highest pvalue (0.05037). Although its pvalue is at the border of the significance level, the variable unfortunately does not meet this condition. Therefore, again the Likehood Ratio Test was conducted to identify whether the variable has a significant effect on the model. The pvalue for this test was 0.09419, so the null hypothesis cannot be rejected. The DebtRatio variable was excluded from the model. Only the significant variables remained in the model.

[10] https://www.statisticshowto.com/likelihood-ratio-tests/

3.4. Specification tests

The next step in the model was to check the correctness of the functional form of the model. For this purpose, 3 diagnostic tests were conducted: Linktest, Hosmer-Lemeshow and Osius-Rojek test. In case of the first test we expect that variable yhat is significant and variable yhat2 should be insignificant. In the case of the other two, we aim to satisfy hypothesis 0 that our model has the correct functional form. Unfortunately, in both cases this condition was not met.

Table 2. Linktest for the base probit model

	Estimate	**Std. Error**	**Z Value**	**Pr(>|Z|)**
(Intercept)	-0.197027	0.014198	-13.88	<0.0000000000000002
yhat	0.371685	0.012124	30.66	<0.0000000000000002
yhat2	-0.031740	0.002607	3.425	<0.0000000000000002

Table 3. Hosmer-Lemeshow and Osius Rojek test for the base probit model

Test	**Stat**	**Val**	**df**	**P-value**
HL	chiSq	391.41727	9	9.281274e-79
OsRo	Z	-10.39896	NA	2.506586e-25

The results obtained are far from perfect, and therefore an attempt was made to fix this error in several ways by: excluding three interactions (singly, in pairs and jointly) and excluding the variable log(NumberOfOpenCreditLinesAndLoans), which was not only a logarithmically transformed variable but also had the largest (significant) pvalue.

Thus, a total of 7 attempts were made to improve the model, but none of them gave positive results. Moreover, in some cases (e.g. removal of all three interactions) even worsened the results. For the probit model the correct functional form could not be obtained. Therefore, for the following chapter, the interpretation of the results should be taken with a pinch of salt, as they do not necessarily describe reality correctly.

4. INTERPRETATION OF MODEL

4.1. Hypothesis verification

First, it was focused on verifying the main hypothesis, i.e. people who have been in defaults at least once in history are more likely to fall into them again. To test this hypothesis, a Likelihood Ratio Test was conducted to compare the final probit model vs. the model with one explanatory variable default_flag. The pvalue for this test is equal to 0. Therefore, hypothesis 0 should be rejected. A customer who has been defaulter in the past is more likely to become insolvent again.

Three side hypotheses were then verified. In two of them we tested whether, as age increases, the chance of defaulting on a loan increases and whether an increasing number of open credit lines has a positive effect on falling into default. For these two hypotheses, it was expected that their β_{age} and $\beta_{NumberOfOpenCreditLinesAndLoans\_log}$ will be higher than 0. For the remaining hypothesis, in which, it was examined the effect of increased earnings on the decrease in the probability of default, it was expected that the $\beta_{MonthlyIncome\_log}$ parameter will be higher than 0.

To verify these 3 side hypotheses, again conducted 3 Likelihood Ratio Tests were conducted in which the final probit model vs. a model with one explanatory variable: age, NumberOfOpenCreditLinesAndLoans_log and MonthlyIncome_log, respectively were compared. For all 3 hypotheses, pvalue's obtained were 0's. Therefore, the null hypotheses for all three cases were rejected. Thus, it should be concluded that with increasing age the probability of falling into default increases, opening new lines of credit has a positive effect on the chance of defaulting on a loan, and increasing earnings offsets the possibility of falling into default. The results obtained seem to be satisfactory as they appear to be reasonable and in line with reality.

4.2. Marginal effects of final probit model

In the case of models where the dependent variable is a binary variable the values of the estimated parameters cannot be directly interpreted. The only possible thing is to interpret the strength of the relationship, that is, in the case of positive parameters, it can be said that a given variable will positively affect the positive case (i.e. 1). In the case of negative parameters, the situation will be the opposite, i.e. the given variable will act in favour of the negative case (i.e. for 0). In order to verify the numerical effect of a given variable on the probability of obtaining a value (in my case - falling into default), it is necessary to calculate the so-called marginal effects or first derivatives of the variables.

For the final probit model, marginal effects for the mean values of each variable were determined.

Table 4. Marginal effects for averaged observations.

Variable	**Marginal effects (df/dx)**
RevolvingUtilizationOfUnsecuredLines	0.047009836
age	-0.000640584
NumberRealEstateLoansOrLines	0.003940963
NumberOfDependents	0.002232169
default_flag	0.098144987
max_default_occurences	0.009242558
NumberOfOpenCreditLinesAndLoans_log	0.001844109
MonthlyIncome_log	-0.008066282
open_loan_on_pension	-0.005995735
dangerous_clients	-0.035940548

For the average variable RevolvingUtilizationOfUnsecuredLines an additional unit will increase probability of defaulting by 4.70 percentage points. For age, the relationship is reversed, i.e. an increase in age by 1 will decrease the probability of defaulting by 0.06 percentage points. Additional unit of NumberOfDependents will increase probability of being in default by 0.394 percentage points.

If the family grows by 1 person, the probability of being in default increases by 0.223 percentage points. If the client has ever been in default before, the probability of falling into default again is 9.81 percentage points. In the case of an increase in the number of times a customer has been in default in the past, the probability that the customer falls back into default is 0.924 percentage points. If the customer takes out a new loan, the probability that the customer falls into default increases by 0.184 percentage points. A one per cent increase in earnings will reduce the probability of falling into default by 0.806 percentage points.

For customers who are over 65 years old and have open credit accounts the probability of falling into default decreases by 0.599 percentage points. For customers who are unemployed and also owe money, the probability of falling into default decreases by 3.59 percentage points. Especially for the last two variables it is economically incorrect. However, it should be borne in mind that the model does not have the correct functional form, which certainly influences the results obtained.

4.3. Pseudo R^2 statistics and models comparison

For classification models, the most commonly used measures seem to be AUC, PR, F-score and accuracy. However, from an econometric point of view, the focus will be on other things, not necessarily on the correct estimation results. Therefore, three other metrics were used, such as: Count R2, Adjusted Count R2, McKelvey & Zavoina.

Table 5. R^2 statistics for estimated tobit model.

Count R^2	Adjusted Count R^2	McKelvey & Zavoina
0.93223288	-0.01386395	0.25397348

Count R^2 is a measure that tells what percentage of the dependent variable was correctly estimated, assuming a treshold of 0.5. In the final probit case 93% of the dependent variable was correctly estimated. The result should not be surprising, due to the fact that the dependendt variable is heavily unbalanced in favour of negative cases (healthy clients).

Interestingly, in the case of Adjusted Count R^2, even a negative value of this statistic can be seen. The adjusted R-squared measures the percentage of correct predictions beyond the level of the random model. Unfortunately, in case of final probit, the model might performs worse than if 0-1 variables were randomly generated.

The last measure i.e. McKelvey & Zavoin what percentage of the variance in the dependent variable can be explained by the independent variables present in the model. This measure has a range from 0 to 100% where my result i.e. 25.39% should be considered low.

In summary, only Count R^2 allowed to get results that I would be happy with. Of course, the fact that the model has an incorrect functional form certainly only highlights that the results are not fantastic. In addition, below a table comparing all the estimated models is attached.

Table 6. Comparison of all models using stargazer package.

	Dependent variable:				
	SeriousDlqin2yrs				
	LPM	*logistic*		*probit*	
	(1)	(2)	(3)	(4)	(5)
RevolvingUtilizationOfUnsecuredLines	0.092***	1.019***	0.574***	0.574***	0.574***
	(0.002)	(0.023)	(0.012)	(0.012)	(0.012)
age	-0.002***	-0.004	-0.007**	-0.008***	-0.008***
	(0.0003)	(0.007)	(0.003)	(0.001)	(0.001)
DebtRatio	-0.00000**	-0.00002	-0.00001**	-0.00001*	
	(0.00000)	(0.00001)	(0.00001)	(0.00001)	
NumberRealEstateLoansOrLines	0.004***	0.092***	0.050***	0.050***	0.048***
	(0.001)	(0.010)	(0.005)	(0.005)	(0.005)
NumberOfDependents	0.003***	0.050***	0.026***	0.026***	0.027***
	(0.001)	(0.010)	(0.005)	(0.005)	(0.005)
default_flag	0.089***	1.631***	0.785***	0.785***	0.785***
	(0.004)	(0.029)	(0.015)	(0.015)	(0.015)
max_default_occurences	0.044***	0.204***	0.113***	0.113***	0.113***
	(0.002)	(0.009)	(0.005)	(0.005)	(0.005)
NumberOfOpenCreditLinesAndLoans_log	-0.001	0.013	0.023**	0.023**	0.023**
	(0.001)	(0.020)	(0.011)	(0.011)	(0.011)
MonthlyIncome_log	-0.010***	-0.197***	-0.099***	-0.099***	-0.098***
	(0.001)	(0.017)	(0.009)	(0.009)	(0.009)
age2	0.00001***	-0.0001*	-0.00001		
	(0.00000)	(0.0001)	(0.00004)		
open_loan_on_pension	0.0004	-0.140**	-0.072**	-0.078***	-0.076***
	(0.002)	(0.070)	(0.033)	(0.025)	(0.025)
dangerous_clients	-0.099***	-2.051***	-0.992***	-0.990***	-1.002***
	(0.009)	(0.179)	(0.088)	(0.088)	(0.088)
Constant	0.140***	-1.883***	-1.030***	-1.012***	-1.014***
	(0.010)	(0.183)	(0.091)	(0.070)	(0.070)

Observations	149,999	149,999	149,999	149,999
Log Likelihood	-28,770.440	-28,619.970	-28,620.020	-28,622.330
Akaike Inf. Crit.	57,566.870	57,265.940	57,264.040	57,266.670
Note:	*p<0.1; **p<0.05; ***p<0.01			

5. CONCLUSIONS

In this paper 3 econometric models are presented, which were based on data concerning credit risk and, more specifically, the estimation of the probability of falling into a so-called default, i.e. a situation in which the customer is insolvent. For the final analysis, a probit model was used due to its lowest information criterion. On the selected model, a "general-to-specific" procedure was carried out, which allows the removal of statistically insignificant variables. Attempts were also made to obtain the correct functional form of the model, but these failed. Therefore, the final results and interpretation should be approached with some distance.

Nevertheless, the verification of all four research hypotheses allowed to obtain satisfactory results, i.e. the influence of all examined variables on the probability of falling into default seems to be in line with reality. In addition, the R^2 statistics appropriate for models with a dependent binary variable were also checked. Marginal effects were also interpreted, which allowed for a better understanding of the model and examination of the exact strength of the dependence of all variables. It is worth noting that, especially with credit risk models, it seems to be crucial to understand which exact characteristics significantly affect the final outcome. Nevertheless, it should be considered that the aim of the paper has been achieved.

Personally, I see a few places where the work could be developed in the future. The first of these, undoubtedly, concerns getting the functional form correct, which is reflected in the final results. One could be tempted to create new interesting interactions based on the variable RevolvingUtilizationOfUnsecuredLines.

It would certainly also add value to the work to create additional columns based on the so-called Weight of Evidence (WoE) InformationValue, which are very often used in credit risk models as they help to find relationships between the dependent variable and the independent variables. Ultimately, the work could be augmented by the creation of credit ratings, i.e. breakdowns that define for us the degree of risk for a given customer.

BIBLIOGRAPHY

Altman, E.I, (1968). Financial ratio, discriminant analysis and the prediction of corporate failure. Journal of Finance, Vol.23 No. 4, pp. 589-609.

Altman, E.I., Saunders, A., (1998). Credit risk measurement: Developments over the last 20 years. Journal of Banking & Finance, 21(11-12), p. 1721-1742.

Anderson, R., (2007). The Credit Scoring Toolkit : Theory and Practice for Retail Credit Risk Management and Decision Automation. Oxford University Press.

Beaver, W., (1966). Financial ratios as predictors of failure. Empirical Research in Accounting: Selected Studies. Journal of Accounting Research, 4, 71–111.

Dong, G., Lai, K.K., Yen, J., (2010). Credit scorecard based on logistic regression with random coefficients. International Conference on Computational Science.

Durand, D., (1941). Risk elements in Consumer Instatement Financing. National Bureau of Economic Research.

European Banking Authority, (2019). Policy advice on the Basel III reforms: credit risk. Standardised approach and IRB approach.

Gouvêa, M.A., Gonçalves E.B., (2007). Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models. POMS 18th Annual Conference.

Gurný, P., Gurný, M., (2013). Comparison of credit scoring models on probability of default estimation for US banks. Prague Economic Papers 2: p. 163-181.

Louzis, D.P., Vouldis, A.T., Metaxas, V.L., (2010). Macroeconomic and bank-specific determinants of non-performing loans in Greece: a comparative study of mortgage, business and consumer loan portfolios. Bank of Greece.

Silva, E.C., Lopes, I.C., Correia, A., Faria, S., (2020). A logistic regression model for consumer default risk. Journal of Applied Statistics, 47:13-15.

https://www.kaggle.com/code/leafar/give-me-some-credit/data.

https://www.moodys.com/sites/products/productattachments/ap075378_1_1408_ki.pdf

https://www.statisticshowto.com/likelihood-ratio-tests/