Professor Cristian CIUREA, PhD
E-mail: cristian.ciurea@ie.ase.ro
Department of Economic Informatics and Cybernetics
The Bucharest University of Economic Studies
Professor Nora CHIRIȚĂ, PhD
E-mail: nora.chirita@csie.ase.ro
Department of Economic Informatics and Cybernetics
The Bucharest University of Economic Studies
Assistant Professor Ionuț NICA, PhD
E-mail: ionut.nica@csie.ase.ro
Department of Economic Informatics and Cybernetics
The Bucharest University of Economic Studies

A PRACTICAL APPROACH TO DEVELOPMENT AND VALIDATION OF CREDIT RISK MODELS BASED ON DATA ANALYSIS

Abstract. *The main objective of this research is to define how the performance of the models used by commercial banks in granting loans and for calculating the ECL according to IFRS9 is developed and validated. The development and validation of a high-performance model are two of the fundamental processes that can avoid the risk of non-payment in case of granting loans by a banking institution. Once a model has been developed, it is validated to assess the predictive power of risk estimators and rating models. Thus, in this research, several techniques for validating the performance of credit risk models will be presented and, with the help of the Python programming language, we will test these techniques on a data set consisting of observations regarding the clients of a credit portfolio. The case study illustrates how a credit risk model has been developed for default probability and how its performance has been validated in terms of power of discrimination, stability and accuracy.*

Keywords*: Credit Risk Models, Predictive Models, Python, Data Analysis.*

JEL Classification: G11, G17, G41, E47

1. Introduction

Banks play a key role in any type of economy, whether it is a developed or a developing economy, and their proper functioning contributes to accelerating

DOI: 10.24818/18423264/56.3.22.04

economic growth. Two of the most important roles of a bank are to take deposits and lend. Banks are also the main providers of payment services in any economy.

Lending is not the only service offered by banks. At the same time, they use the funds available for investment and allocate money to the most profitable projects in the economy. Thus, the allocated money is used by companies and households also for investment purposes and ends up being deposited in the bank's accounts in the form of deposits when they will profit from those investments. One of the most important sources of income for banks is the difference between the interest rate on deposits and the interest rate on loans. So, one of the crucial indicators for the banking sector is the ratio between loans and deposits **(Chirita & Nica, 2019).** Another approach related to an analysis of investment decisions from IC&T industry in the context of behavioral economy was presented in **(Chirita et al, 2021)**. The development of an index score for the internal auditor profile in Romania based on real data analysis was described in **(Mocanu & Ciurea, 2019).**

So, among the many roles of the bank we can list: it acts as an intermediary between those people who need funds and those who have excess money; facilitates commercial transactions; helps to develop the national economy by providing credit to businesses in all industries; by granting consumer loans with advantageous interest rates, it helps to raise the standard of living.

In general, banking is considered a risky business with many sources of risk, so the risk management process begins by identifying and investigating those activities that generate risks or losses and continues by assessing, analyzing and measuring them. The next step is to establish measures that can be taken to prevent or address these possible situations. Once a series of measures have been implemented, the final step is to monitor and review those risk management practices, as appropriate. In the context of the COVID-19 pandemic, many banking institutions have encountered difficulties in adapting credit risk models because various laws and regulations have been issued that have had an impact on the scoring system, and therefore on credit risk models **(Chirita & Nica, 2020).**

Credit risk together with the development, validation and management of risk models are the most critical elements in any economic entity.

2. Conceptual Dimensions of Risk in the Banking field

Since the onset of the credit crunch in the United States and in Europe in 2006, risk managers have learned valuable lessons about quantifying, assessing, and the importance of properly measuring bank risk. The field of risk management has undergone enormous change over the last 50 years, and the pace of change is accelerating, largely due to the recent crises and speculative bubbles that have taken place so far. The events of the last decade have also changed the way we think, good practices and definitions of risk management. Below we will mention some of the most significant definitions of risk management **(Bessis, 2002).**

Risk management is a discipline that clearly indicates the management of risks and the returns of each major strategic decision at both institutional and transactional level. Discipline risk management shows how to change strategy to align compensation with long-term and short-term risk profitability in an enterprise. This discipline includes several subdomains that are inseparable and can overlap. We mention some of the most important: credit risk, market risk, asset and liability management, liquidity risk, methodology for calculating and allocating capital, operational risk and performance measurement.

As a rule, the financial assets managed by a Romanian commercial bank are subject to several general accounting standards and regulations such as IFRS9, IFRS7, IAS18, EBA GL 2017/06, European Regulation 575/2013 or NBR Regulation 5/2013.

From the perspective of a preliminary analysis of the risks to which commercial banks are subject, they may be: *financial risks* (credit risk, liquidity risk, interest rate risk, excessive leverage risk, currency risk), *operational risks* (model risk , strategic risk and ICT, conduct risk, legal risk), *human risks* (psycho-social risks, risks of attracting and retaining employees), *strategic risks* (risks generated by legislative changes, risks generated by technological progress), *accidental risks* (risk generated fires, seismic risk) and *risks generated by financial contagion*.

Risk management is done by the following actors:

- Institutions of supervision and regulation of the bank: These cannot prevent bank failures. Their main role is to act as intermediaries in the risk management process and to improve and monitor the statutory framework for risk management. By creating the right environment, banking regulators and supervisors play a crucial role in influencing other key players.
- Shareholders: They are able to appoint the persons responsible for the corporate governance process and a careful examination must be carried out by the regulators in order to ensure and verify that they do not intend to use the bank only to finance their own businesses or their associates.
- Executive management: It must be chosen appropriately and appropriately for these responsibilities. Executive managers are those who have the competence to define ethical behavior and the management of the bank. He must have knowledge of both the management and the financial risks that a bank manages.
- The audit committee and the internal auditors: These should be seen as an extension of the risk management policy function of the board of directors. Internal auditors are responsible for independently assessing the bank's compliance with internal control systems, accounting practices, internal rules and procedures, and information systems. The audit committee has an important role to play in identifying and addressing risk areas, but it is not their responsibility to manage them, but they need to be integrated into all levels of management.

- External auditors: Play an important role in evaluating the risk-based financial reporting process. Appropriate liaison mechanisms are needed between banking supervisors and external auditors.
- Consumers as market participants: They must take responsibility for their own investment decisions. They need to be well informed and can be assisted by financial media, financial analysts, brokers and rating agencies.

The banking system is one of the most important systems, in our opinion, because its effects can spread beyond it, to influence the behavior of markets such as foreign exchange or high impact in an economy, both micro economically and at macroeconomic level. Therefore, we consider an analysis to be relevant both internally and externally in order to be able to make certain strategic decisions. Even if in this article we will focus on the credit risk analysis and the validation of the related models, it is important to know the whole range of existing risks that can be formed at the level of the banking system. The bank's core business, lending, is focused on borrowers and borrowers, mainly dealing with two components: money management and risk quantification **(Bouteille & Coogan-Pushner, 2013).**

In order to be able to manage the credit risk and the credit risk model within the banking institutions, it is necessary to know several fundamental notions. For example, inability to pay is a risk that a borrower can no longer repay the amount owed to the bank. The default concept practically indicates a borrower or a credit account holder who no longer pays his obligation to the bank according to the credit agreement. Another type of credit risk is the probability of default that is estimated as a measure of the probability of a debtor to repay his loan or the inability to fulfill his contractual obligations (**Bandyopadhyay, 2006**).

3. Credit Risk Models

To ensure that losses are lower when it comes to lending, banks need to take a number of steps. The best way to reduce losses is to carefully monitor and measure risk, build and implement credit risk models with which banks can make a forecast of expected losses as close to reality as possible. These losses are the amount that the lender could lose by lending and can be calculated as the product of three components:

$$ECL = PD * LGD * EAD$$

The information we have available for model development may vary depending on the type of debtor. For individuals, we may have demographic and socio-demographic information, such as age, gender, occupation, level of education, income level, zip code, etc. Other information is related to the characteristics of the product that a customer is requesting, the purpose of the loan and the interest rate. In addition, we may have data provided by external agencies, such as credit ratings.

In the case of mortgages, the ratio between the mortgage loan and the value of the mortgaged property is very important.

$$LTV = \frac{Mortgage\ credit}{Property\ Price}$$

LTV or Loan to Value is an indicator used by banks that reflects the value of the loan in the total market value of the guarantee. In analyzing a company's credit demand, banks can consider the industry in which they operate, its size, financial situation, and can calculate ratios such as return on assets (ROA), return on equity (ROE), etc.

A variety of risk modeling information may be available at the time of loan application and may be used by the bank to build models based on credit applications, by estimating credit risk, so that it can decide whether or not to grant the loan and what interest to pay. apply. The riskier a loan is, the higher the interest rate.

Many of the customer's information will be known to the bank only after the loan is granted and after the debtor's behavior has been sufficiently observed and analyzed. This additional data can be used to develop behavioral models. Behavioral models are used by banks to calculate PD and ECL and to decide whether to grant another loan to an existing customer.

3.1. Probability of Default (PD)

Probability of Default is a key risk parameter used in the context of credit risk management. It is a measure that assigns a numerical value between 0 and 1 to the probability of a properly defined credit event (such as default, bankruptcy) within a specified time horizon. In the internal valuation approach, the probability of default (abbreviated PD) of a counterparty is estimated over a period of one year.

The probability of default is usually modeled by a logistic regression, being very easy to interpret. In terms of probabilities, the logistic regression model estimates the probability of an event occurring. First, for PD modeling, it must be determined which is the dependent variable or, in other words, what is to be estimated. According to **Řezáč M. and Řezáč F. (2011)**, “In the case of credit scores, it is necessary to accurately define good customers and bad customers. This definition is usually based on the number of days that have elapsed since the due date. " Usually, a borrower can be considered a bad customer if he has exceeded the due date by more than 90 days.

3.2. Loss Given Default (LGD)

The percentage of exposure that will be lost and cannot be recovered is counted by the Loss Given Default indicator **(Frye, 2004).**

As there is a direct link between the LGD and the so-called recovery rate (RR), in practice the loss is modeled in the event of non-repayment by RR, which can be calculated as that proportion of the amount financed which was recovered by the time of entry into non-refund status:

$$RR = \frac{AR}{AL}, RR \in [0,1]$$

RR means Recovery Rate; AR means Amount Recovered and AL represent Amount Loaned. Thus, with the help of the recovery rate, LGD can be calculated with the following formula:

$$LGD \ = 1 - RR, LGD \ \in [0,1]$$

An LGD value of 100% and a recovery rate of 0% are interpreted by the fact that the bank fails to recover the loan.

As in the case of PD, long-term historical data on debtors' debt is essential. When collecting this data, it must be considered that the state of the economy may change the value of collateral and other factors, as well as the ability of borrowers to repay their loans.

Many other variables that could affect this component must be considered when creating a database for estimating LGD, such as: information about loans and collateral, collateral deposits, and other characteristics of the borrower, such as creditworthiness, geographic area and the industry in which it operates.

As real estate collateral continues to be a major form of collateral, it is very important to monitor real estate market trends to see what steps need to be taken if the real estate market is declining, to find out how long it takes to sell the property in real estate. case of non-payment etc. **(Spuchlakowa & Cug, 2015).**

Financial-banking institutions often suffer additional losses due to collateral that can no longer cover loans. If the value of real estate collateral may decrease during periods of default and strongly reflect the creditworthiness of borrowers, these additional losses should be included in the LGD estimate.

3.3. Exposure at Default (EAD) and Credit Conversion Factor (CCF)

Non-repayment exposure or exposure at default (EAD) is the total amount that a creditor is exposed to when a debtor goes into default. That is, it is the maximum amount that a bank can lose when a borrower can no longer pay its debts. In many cases the bank grants a loan, but the lender does not use the full amount or may have paid off a large portion of the loan at the time of default.

Therefore, the dependent variable for the EAD model will be that proportion of the borrowed amount that has not been repaid. This ratio is called the Credit Conversion Factor (CCF) and takes values between 0 (if all the loan has been repaid) and 1 (if no part of the money has been repaid) and can be calculated so:

$$CCF = \frac{AL - AR}{AL}$$

Then, Exposure at Default can be easily calculated that way:

$$EAD = CCF * AL$$

3.4. Scorecards

For credit sales agents and employees who do not have extensive experience in statistics, the models for estimating credit risk are simplified in the form of scorecards. The score of the PD model in the form of scores is represented in a scorecard.

Depending on the information used to develop the model, scorecards can be of two types: Application Scorecards and Behavioral Scorecards.

Whether we are talking about a credit application from a new client or from an existing one, each borrower is represented by his credit score. The lower this score, the higher the risk.

This score was developed and provided by a company called Fair Isaac Corporation and can take values between 300 and 850. A credit score of 300 shows that the person being assessed has the highest possible level of credit risk and will not may honor contractual obligations. Instead, a credit score of 850 shows that the person being assessed has the lowest possible level of credit risk and the highest willingness to repay the borrowed amount **(Bouteille & Coogan-Pushner, 2013).**

4. Case Study: Development and Validation of Credit Risk Models

In the following, the components of credit risk will be modeled and scorecards will be built using Python programming language, together with *pandas*, which is a quick, powerful, versatile, and easy-to-use open source data analysis and processing tool (https://pandas.pydata.org). For this, a 7-year database will be used, which contains a series of information about people who have applied for a peer-to-peer loan. The source of the data is www.kaggle.com.

PD, LGD and EAD modeling involves steps: preprocessing data, estimating models, testing and evaluating their performance.

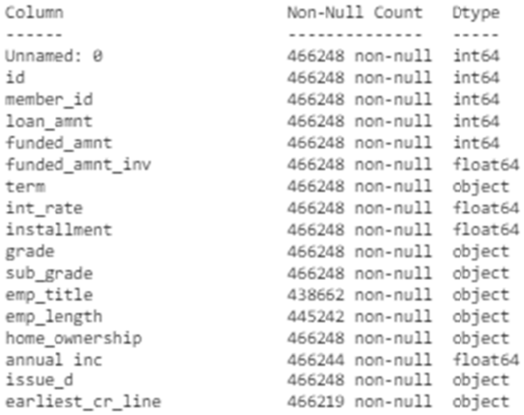
The database contains a number of dependent and independent variables. Independent variables or predictors are those that contain the information needed to make a prediction of dependent variables.

For the PD model, the dependent variable will be the status of the loan. For the LGD model, the dependent variable will be how much could be recovered from the loan until the borrower went into default, and in the case of EAD, the total exposure will be used from the time of default. compared to the total past exposure.

In order to be able to present the model in a much simpler form by means of scorecards, its independent variables must be categorical. Therefore, the step of data preprocessing for the PD model will be to transform all independent variables into categorical variables or so-called dummy variables. For those variables that are already discrete, preprocessing will consist of creating dummy variables for each of their original categories. If the discrete variables have too many categories or the information, they hold is negligible, they will be combined into a new category, with the ultimate goal of reducing the total number of variables. Continuous variables, such as annual income or the number of months since the loan was granted, will also be converted to dummy variables, following a different procedure called Fine Classing. This procedure involves dividing each variable into several equal size ranges. Once these intervals are created, it is analyzed how well they differentiate between repaid and non-repaid loans.

If two common categories make the difference equally well, they will be interclassed, otherwise they will be part of another variable, and the intervals do not need to be equal. This procedure is called Coarse Classing.

In the figure below, you can see that a number of variables, such as loan repayment term or seniority, are time constraints. In order for all continuous variables to be of integer type, they will be transformed with the help of functions specific to the Python programming environment.



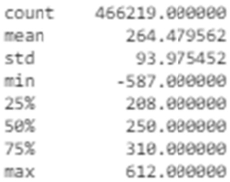
```
Column                Non-Null Count   Dtype
------                --------------   -----
Unnamed: 0            466248 non-null  int64
id                    466248 non-null  int64
member_id             466248 non-null  int64
loan_amnt             466248 non-null  int64
funded_amnt           466248 non-null  int64
funded_amnt_inv       466248 non-null  float64
term                  466248 non-null  object
int_rate              466248 non-null  float64
installment           466248 non-null  float64
grade                 466248 non-null  object
sub_grade             466248 non-null  object
emp_title             438662 non-null  object
emp_length            445242 non-null  object
home_ownership        466248 non-null  object
annual inc            466244 non-null  float64
issue_d               466248 non-null  object
earliest_cr_line      466219 non-null  object
```

Figure 1. Type of variables

Source: our own computation

The variable that describes the length of service of the debtor is stored as a string, because there is a string after each digit, which will be deleted. Similarly, for the variable describing the duration of the loan, the excess text is removed and the remainder is converted into a numeric variable. In the case of variables, the date of the first loan and the date of granting the loan, it can be seen that they are string variables, but it is assumed that they should be data variables. What can be done is to calculate the time that has elapsed since each of these events.

By examining the oldest loan granted, the difference between days and months is converted and the *pandas.to_datetime()* function is used, which will convert the variable from string to data, saving the result to a new variable. By generating descriptive statistics, the following are obtained:

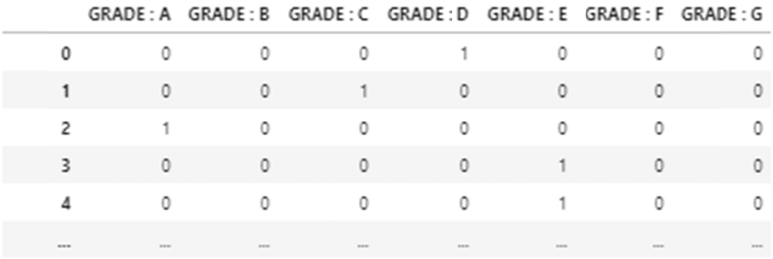


```
count    466219.000000
mean        264.479562
std          93.975452
min        -587.000000
25%         208.000000
50%         250.000000
75%         310.000000
max         612.000000
```

Figure 2. Descriptive statistics

Source: our own computation

The mean, standard deviation, quartiles, and maximum value appear to be correct, but the minimum value is negative. This is a consequence of the transformation of string variables into data variables. The function used misinterpreted the years before 2000, turning them into 2080-type observations instead of 1980. In order not to eliminate all these negative observations from the analysis, they are replaced by the maximum value of the months since the first request. credit. Dummy variables must be used to preprocess discrete variables. If that variable has k categories, k-1 dummy variables are constructed to structure the information. In this case, the *pandas.get_dummies()* function is used for the categories that are in the variable degree of risk.



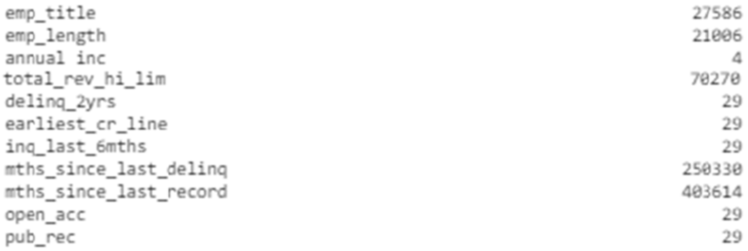
	GRADE : A	GRADE : B	GRADE : C	GRADE : D	GRADE : E	GRADE : F	GRADE : G
0	0	0	0	1	0	0	0
1	0	0	1	0	0	0	0
2	1	0	0	0	0	0	0
3	0	0	0	0	1	0	0
4	0	0	0	0	1	0	0
...	...	...	...	...	...	...	...

Figure 3. Dummy variables for the degree of risk

Source: our own computation

Seven dummy variables are obtained that match the number of categories of the initial variable. For example, the applicant with ID 3 falls into the B degree of risk. Similarly, dummy variables are constructed for all categorical variables and concatenated with the original database.

Next, we examine whether there are missing values by summing the missing values on each column using the *sum()* and *isnull()* methods.



```
emp_title                                                    27586
emp_length                                                   21006
annual inc                                                       4
total_rev_hi_lim                                             70270
delinq_2yrs                                                     29
earliest_cr_line                                                29
inq_last_6mths                                                  29
mths_since_last_delinq                                      250330
mths_since_last_record                                      403614
open_acc                                                        29
pub_rec                                                         29
```

Figure 4. Missing values

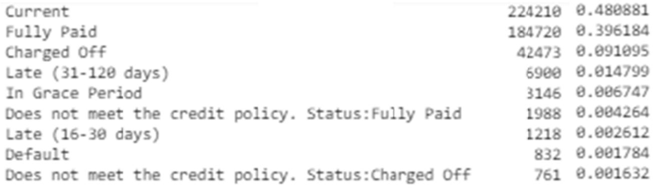
Source: our own computation

From the figure above, it can be seen that there are a number of missing observations: occupation, seniority, date of first credit, etc. One way to deal with missing values is to remove or replace them with another value. Because there are missing comments in the annual income columns and the current credit limit, they will be replaced by the average annual income, respectively the average of the amounts financed by the creditor. The missing values for the other variables that are required in the analysis will be replaced by the value 0.

Next, based on the constructed scorecard, the PD model will be formed. The statistical method that will be used to model the probability of default is a logistic regression in which the dependent variable is represented by the chances of a customer being good-paying or bad-paying, and the independent variables are represented by dummy variables.

$$\ln\left(\frac{Ndef}{Def}\right) = \varphi_0 + \varphi_0 * X_0 + \varphi_1 * X_1 + \cdots + \varphi_m * X_m$$

The only relevant information that can help build the dependent variable can be found in the column indicating the status of the loan.



```
Current                                               224210  0.480881
Fully Paid                                            184720  0.396184
Charged Off                                            42473  0.091095
Late (31-120 days)                                      6900  0.014799
In Grace Period                                         3146  0.006747
Does not meet the credit policy. Status:Fully Paid      1988  0.004264
Late (16-30 days)                                       1218  0.002612
Default                                                  832  0.001784
Does not meet the credit policy. Status:Charged Off      761  0.001632
```

Figure 5. Loan situations

Source: our own computation

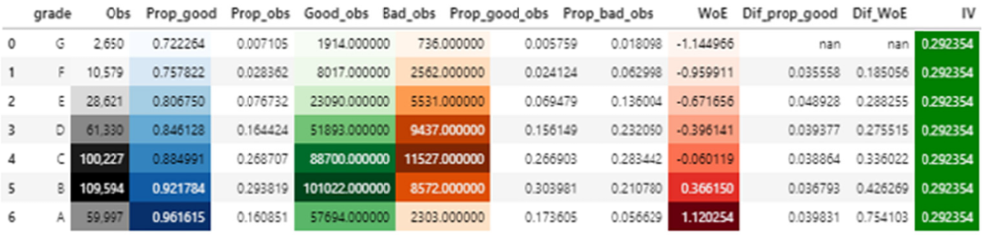
The figure above shows the statuses that a loan can have, the number of accounts that correspond to each status, and their proportion. A new dummy variable is built and using the *numpy.where()* function and the *isin()* method, it will take the value 0 if it falls into one of the variants ("Default", "Late (31-120 days)")

, "Charged Off", "Does not meet the credit policy. Status: Charged Off") and the value 1, otherwise.

Using logistic regression, the relationship between the dependent variable and the independent variables will be analyzed and the regression coefficients will be estimated.

To avoid the problem of overfitting or underfitting, a training set and a test set are used, the scikit-learning library offering a pretty good method of sharing the initial data set. In this case, the training set will contain 80% of the initial database and the test set 20%.

To determine WoE and IV for discrete variables, a function is built that will automate these calculations. The next step is to calculate the proportion of observations in the total number of observations, then the number of good payers and the number of bad payers and then the proportions of good and bad borrowers in the total number of borrowers, resulting in WoE by dividing the two. Calculating the difference between the two proportions and multiplying by WoE will result in IV. By applying the above function to the variable indicating the degree of risk to the debtor, the following results are obtained:



	grade	Obs	Prop_good	Prop_obs	Good_obs	Bad_obs	Prop_good_obs	Prop_bad_obs	WoE	Dif_prop_good	Dif_WoE	IV
0	G	2,650	0.722264	0.007105	1914.000000	736.000000	0.005759	0.018098	-1.144966	nan	nan	0.292354
1	F	10,579	0.757822	0.028362	8017.000000	2562.000000	0.024124	0.062998	-0.959911	0.035558	0.185056	0.292354
2	E	28,621	0.806750	0.076732	23090.000000	5531.000000	0.069479	0.136004	-0.671656	0.048928	0.288255	0.292354
3	D	61,330	0.846128	0.164424	51893.000000	9437.000000	0.156149	0.232050	-0.396141	0.039377	0.275515	0.292354
4	C	100,227	0.884991	0.268707	88700.000000	11527.000000	0.266903	0.283442	-0.060119	0.038864	0.336022	0.292354
5	B	109,594	0.921784	0.293819	101022.000000	8572.000000	0.303981	0.210780	0.366150	0.036793	0.426269	0.292354
6	A	59,997	0.961615	0.160851	57694.000000	2303.000000	0.173605	0.056629	1.120254	0.039831	0.754103	0.292354

Figure 6. WoE and IV calculation - degree of risk

Source: our own computation

In the figure above, it can be seen that IV = 0.29, which means that the variable "degrees" has a medium to high predictive power and should be included in the model.

In order to be able to decide how to organize the original categories of discrete variables into dummy variables and to be able to interpret WoE, a function is built with which a series of graphs will be made **(Costea, 2017).**

These graphs will have the independent variable categories on the abscissa and the WoE values on the ordinate.

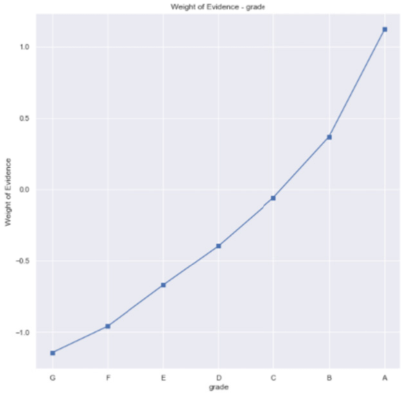
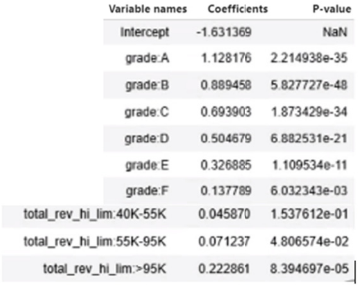


Figure 7. WoE - degree of risk

Source: our own computation

As can be seen from the chart, WoE increases almost monotonously as the risk increases from the highest (G) to the best (A). The better the grade, the higher the WoE. Therefore, because the dummy variables will be introduced in a regression model, one of them is kept as a reference category against which the impact of all other variables on the result will be evaluated. In this case, the reference category will be the one with the highest degree of risk (G).

Because the drive set is used to build the model and the test set is used to test it, the latter must have the same dummy variables as the drive set and go through the same preprocessing steps. Therefore, the entire code is run by replacing the drive set with the test set. In order to estimate the PD model, the fit () method specific to the scikit-learn library must be applied, which will receive as arguments the set consisting of all dummy variables, without reference categories, and the set containing the dependent variable. In addition, two other functions will be used to display the estimated values of the coefficients and the value of free time. It is also necessary to decide which independent variables contribute to the estimation of the good-paying or bad-paying status of each client. The method to be used is to check the statistical significance of the regression coefficients using p-value.



Variable names	Coefficients	P-value
Intercept	-1.631369	NaN
grade:A	1.128176	2.214938e-35
grade:B	0.889458	5.827727e-48
grade:C	0.693903	1.873429e-34
grade:D	0.504679	6.882531e-21
grade:E	0.326885	1.109534e-11
grade:F	0.137789	6.032343e-03
total_rev_hi_lim:40K-55K	0.045870	1.537612e-01
total_rev_hi_lim:55K-95K	0.071237	4.806574e-02
total_rev_hi_lim:>95K	0.222861	8.394697e-05

Figure 8. Intercept, regression coefficients and p-value

Source: our own computation

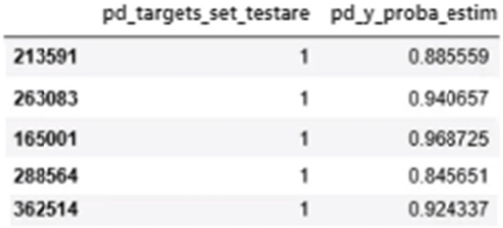
Each original independent variable is represented by several dummy variables. So, if the coefficients of these variables are statistically significant, they must all be kept. But if none of the coefficients are statistically significant, all dummy variables must be removed.

In the figure above the p-value is less than 0.05 for all the coefficients of the variables that describe the degree of risk of the client, therefore, all the degrees in the analysis are kept. Instead, for the maximum amount a customer can spend, all coefficients are insignificant, so all variables are removed. Once it has been decided which variables are not useful in the analysis and which are eliminated, the logistic regression will be applied again.

The exponent of the coefficient of a dummy variable shows how the chances of an event occurring that occur between a zero-value observation (bad-payer) and a 1-value (bad-payer) observation can change. Therefore, higher values of the manikin variables in the PD model mean a higher chance of being a good paying customer.

From the figure above, it can be seen that risk C borrowers are more likely to be good payers than risk G borrowers with $e^{0.889458} = 2.43$

After the model creation step, the test step follows. The model is used to estimate the probability of default for each observation in the test data. This will be done using the *predict_test()* method, which takes the model data and applies the prediction method, with the test set entries being the only argument the method receives. The output of this method will give each customer the probability of being good or bad.



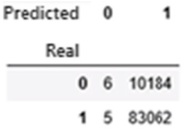
	pd_targets_set_testare	pd_y_proba_estim
213591	1	0.885559
263083	1	0.940657
165001	1	0.968725
288564	1	0.845651
362514	1	0.924337

Figure 9. Probabilities of good-bad values

Source: our own computation

An analysis of the confusion matrix can be made to evaluate the model. This shows how many good or bad-paying debtors were predicted as such and how many of them are good customers, but the model predicted them as bad and vice versa.

The *crosstab()* function is used to create the confusion matrix in Python. According to **McKinney (2013)**, "Cross-tables are a special case of pivot tables, which calculate group frequencies." The function will receive as parameters the data of the test set and the predicted data and will build us a table that will show us the number of observations of the variables that fall into each category.

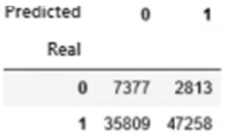


Predicted	0	1
Real		
0	6	10184
1	5	83062

Figure 10. Good-bad confusion matrix – cut-off 0.5

Source: our own computation

In the figure above, it can be seen that the use of a cut-off of 0.5 generates a multitude of false-positive observations, many bad-paying debtors being wrongly classified as good-paying. This is due to the fact that very few clients in the portfolio are late in paying their debts, and the logistic regression classified almost all applicants as good payers. Therefore, if this model with a threshold of 0.5 were used to make the decision to accept or reject credit applications, loans would be granted to a large number of bad credit applicants. Using a cut-off of 0.9, the following results are obtained:



Predicted	0	1
Real		
0	7377	2813
1	35809	47258

Figure 11. Good-bad confusion matrix – cut-off 0.9

Source: our own computation

At the threshold of 0.9 there are far fewer false-positive observations, but also fewer true-negative observations. So, if a lender uses this model to grant a loan, it will dramatically reduce the number of bad customers, but also the number of approved applications. This shows that the value of 0.9 of the cut-off is far too restrictive, the main interest of creditors when it comes to modeling credit risk is to minimize risk, but also to grant as many loans as possible.

In the case of the 0.5 cut-off, the accuracy of the model was about 0.9, and in the case of the 0.9 cut-off it was 0.59. Therefore, it can be concluded that accuracy is not a universal method of measurement for a PD model and that we are more interested in false-positive observations. One of the most common ways to approach this problem and check the performance of the model is to see the rate of true-positive and false-negative observations for different threshold values and to construct the ROC curve.

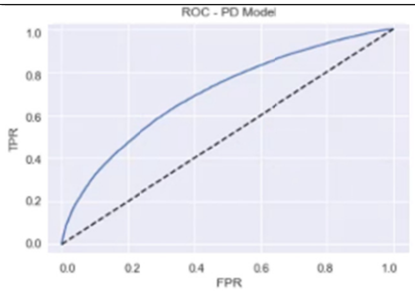


Figure 12. ROC Curve of PD Model

Source: our own computation

The graph above shows that the model is a good one, because the ROC curve is to the left of the 45-degree dotted line. To find out exactly how good its performance is, calculate the area under the curve (AUC). Its value is about 71%, so the model has a reasonable performance.

Next, the performance of the model is analyzed using the Gini coefficient and the Kolmogorov-Smirnov test. The first measures income inequality between rich and poor people in an economy. In this case, it will be used to measure the inequality between good and bad debtors.

To calculate the Gini coefficient, we need the cumulative percentage of the population, the cumulative percentage of good customers and the cumulative percentage of bad customers.

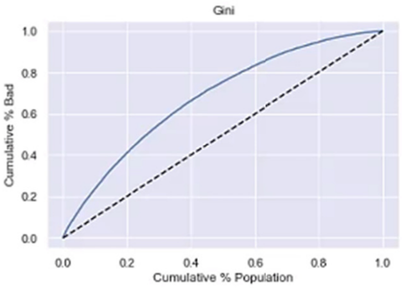


Figure 13. GINI Curve of PD Model

Source: our own computation

The Gini coefficient will be calculated using directly under the area under the curve (AUC), as follows **(Schatz, 2020)**:

$$GINI = 2 * AUC - 1 = 2 * 0.71 - 1 = 0.42$$

So, Gini is equal to 0.42, a pretty good value for a bank's client portfolio.

4. Conclusions

The case study highlighted that the modeling of credit risk components required the construction of dependent variables and the choice of a set of independent variables. In the process of estimating the regression and p-value coefficients for the PD model, we used logistic regression to model whether or not the borrowers in the loan portfolio are in default. The first step was to apply the same logistic model used in the case of PD, analyzing whether the dependent variable represented by the recovery rate has the value zero or not.

In the test step of the model, we used the *predict()* or *predict_test()* methods specific to the *sklearn* library of the Python language to determine the probabilities or predicted values of the dependent variables.

In the step of evaluating the performance of the PD model we used the confusion matrix, we calculated the accuracy for two thresholds and we noticed that this cannot be considered a good method of evaluating a model. Next, we used the ROC curve and calculated the area under the curve, resulting in a reasonable performance of the model. We also calculated the Gini coefficient using the area under the curve.

Credit risk is the most important and most difficult to quantify the risk to which a banking financial institution is exposed, the incorrect assessment of the probability of default having serious consequences on that institution and beyond.

REFERENCES

[1] **Bandyopadhyay, A. (2006),** ***Predicting Probability of Default of Indian Corporate Bonds: Logistic and Z-Score Model Approaches.*** The Journal of Risk Finance, 7(3), 255–272;

[2] **Bessis, J. (2002),** ***Risk Management in Banking****:* Second Edition*. Wiley;*

[3] **Chirita, N., Ciurea, C., Nica, I. (2021),** ***An Analysis of Investment Decisions From IC&T Industry in the Context of Behavioral Economy*;** Economic Computation and Economic Cybernetics Studies and Research, Issue 2, Vol. 55, pp. 159-176, *ASE Publishing;*

[4] **Chirita, N., Nica, I. (2019),** *Cibernetica Firmei. Aplicatii si Studii de Caz. Economica, Vol. 1, pp.89-104;*

[5] **Chirita, N., Nica, I. (2020**), ***Analysis of the Impact Generated by COVID-19 in Banking Institutions and Possible Economic Effects;*** Theoretical and Applied Economics, No.3/2020 (624), *Spring;*

[6] **Costea, A. (2017),** ***A Quantitative Approach to Credit Risk Management in the Underwriting Process for the Retail Portfolio***. The Romanian Economic Journal; Issue no. 63, pp.157-184;

[7] **Dar, A. A. & Anuradha, N. (2017),** ***Probability Default in Black Scholes Formula: A Qualitative Study.*** Journal of Business and Economic Development. Vol. 2, No. 2, pp. 99-106. doi: 10.11648/j.jbed.20170202.15;

[8] **Lessambo, F. (2019),** ***The U.S. Banking System: Laws, Regulation, and Risk Management.*** *Springer*;
[9] **Luo, S., Murphy, A. (2020),** ***Understanding the Exposure at Default Risk of Commercial Real Estate Construction and Land Development Loans****.* Federal Reserve Bank of Dallas;
[10] **McKinney, W. (2013*), Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython***. Sebastopol: O'Reilly Media;
[11] **Mocanu, M., Ciurea, C. (2019),** ***Developing an Index Score for the Internal Auditor Profile in Romania Based on Real Data Analysis*;** *Economic Computation and Economic Cybernetics Studies and Research*, Issue 2, Vol. 53, pp. 93-111; *ASE Publishing;*
[12] **Řezáč, M. & Řezáč, F. (2011)**, ***How to Measure the Quality of Credit Scoring Models.*** Czech Journal of Economics and Finance. 61, 486-507;
[13] **Schatz, I. (2020),** ***Using the Gini Coefficient to Evaluate the Performance of Credit Score Models.*** Published in Toward Data Science;
[14] **Spuchlakowa, E., Cug, J. (2015),** ***Credit Risk Modelling and LGD Modelling****.* Procedia Economics and Finance, vol.23, pp. 439 – 444, *Elsevier.*