Internal ratings-based model introduction to a retail portfolio and its impact on a bank's capital requirement

Aleksandr Peussa

University of Helsinki

Faculty of Social Sciences

Economics

Master’s Thesis

December 2019

Tiedekunta - Fakultet - Faculty	Laitos - Institution - Department
Faculty of Social Sciences	Department of Political Science and Economics

Tekijä - Författare - Author

Aleksandr Peussa

Työn nimi - Arbetets titel - Title

Internal ratings-based model introduction to a retail portfolio and its impact on a bank's capital requirement

Oppiaine - Läroämne - Subject

Economics

Työn laji/ Ohjaaja - Arbetets art/Handledare - Level/Instructor	Aika - Datum - Month and year	Sivumäärä - Sidoantal - Number of pages
Master's Thesis	October 2019	26 pages

Tiivistelmä - Referat - Abstract

The bank as a business tries to maximize the profit measured by the return on allocated capital (ROAC). This measure is calculated by dividing net income on capital allocated, thus the bank should both maximize net income and minimize capital allocated on the portfolio.

The idea of the research is to check whether the higher return on allocated capital may be achieved if a bank calculates a capital requirement based on an internal ratings-based (IRB) approach instead of a standardized approach. The research question hypothesizes that the introduction of the IRB approach may lower the capital allocated on the retail portfolio, as the bank currently uses a standardized approach for the capital requirement calculation. In case the hypothesis is valid, the introduction of the IRB model lowers minimum capital requirements and through it, the higher ROAC is achieved.

The capital requirement under the IRB approach is calculated through the estimation of the probability of default (PD) and loss given default (LGD) based on the bank's historical data. Probability of default is estimated using logistic regression and for loss given default estimation, the simple linear formula is applied due to unavailability of relevant data.

The research proves that the application of the IRB approach significantly lowers bank's capital requirement and as a result increases its return on allocated capital.

Avainsanat – Nyckelord - Keywords

Credit risk, banking, capital requirement, Basel Accords

Muita tietoja - Övriga uppgifter - Additional information

Contents

1. Introduction

Banks and other financial institutions receive cash inflow from various sources. For instance, there are bank deposits like saving accounts or bonds issued. A bank uses the money obtained in order to make different investments, which always contain a risk associated with the fact that obligor could default and not pay the loan due to various reasons. As the government provides the banks' customers with deposit insurance and as financial institutions play a big role in the economic system, the government aims to ensure that financial institution is well protected against the risks it is exposed to. Hence, a bank is supposed to hold enough shareholder capital to withstand an unexpected loss in order to avoid insolvency. (Baesens 2016, 5.)

According to the Basel Accord, the financial institution has a choice between two broad methodologies for calculating their minimum capital requirements for credit risk: standardized approach and internal ratings-based approach. The approaches differ in terms of their complicity and level of flexibility related to banks' ability to use own credit risk estimates. (Joseph 2013, 289.)

The standardized approach refers to the fact that the banks are required to use measures defined by the regulator in order to estimate a minimum capital required. The values of measures depend on the structure of the product and the type of borrower. (Witzany 2017, 112.)

The internal ratings-based (IRB) approach is more complicated. Under the approach, the banks are allowed to use their quantitative models in order to estimate credit risk measures. The appeal of the internal ratings-based approach is that it may allow banks to obtain a lower level of capital requirement compared to the standardized approach. However, banks can use IRB only after approval from their local regulators - central banks. (Joseph 2013, 291.)

The bank as a business tries to maximize the profit measured by the return on allocated capital (ROAC). This measure is calculated by dividing net income on capital allocated, thus, the bank should both maximize net income and minimize capital allocated on the portfolio.

The idea of the research is to check whether the higher return on allocated capital may be achieved if the bank calculates a capital requirement based on the IRB approach instead of a standardized approach. The research question hypothesizes that the introduction of the IRB approach may lower capital allocated on the retail portfolio, as the bank currently uses a standardized approach for the capital requirement calculation. In case the hypothesis is valid, the introduction of the IRB model lowers minimum capital requirements and through it, the higher ROAC is achieved. (Balin 2008, 8.)

The first part of the thesis focuses on the theoretical aspects of the research question. Firstly, the idea behind Basel Accords and methodologies proposed for minimum capital requirements calculation are explained. Then the credit risk measures, the data and econometric methods for these measures estimation and validation are reviewed. Next, paper shows how credit risk measures are used in the internal rating-based model.

The second part contains empirical analysis, which is based on the theoretical part and the data from a bank. The bank operates in a private segment and provides secured products like cars and other asset financings. The bank does not provide mortgages or revolving products, thus the data contains non-mortgage and non-revolving products for private customers only. Based on the empirical data from the bank, the purpose of the research question is to prove whether the introduction of the IRB approach would lower the bank's capital requirement and through this would increase portfolio profitability measured by ROAC.

2. Literature review

There are no straightforward references to similar researches. The assumption is that due to compliance and corporate security reasons there is no easy access to unique bank's data in order to cover similar research questions. However, based on multiple pieces of evidence IRB approach is widely used in the banking industry. Based on European Banking Authority at least 102 financial institutions have the IRB approach applied to their portfolios, which cover 64% of EU institutions' total credit risk-weighted exposures (European Banking Authority 2017, 13.)

This is a strong signal that the IRB approach has a good potential to be applied for a bank's retail portfolio. However, it is not inevitable that the introduction of the IRB approach would necessarily decrease the capital requirement. The capital requirement under the IRB approach depends on portfolio agreements' probability of default (PD) and loss given default (LGD). Hence, in order to calculate capital requirements under the IRB approach, the PD and LGD estimations are required.

The PD estimation part is mainly covered in my previous research (Peussa, 2016), which shows that customers' probability of default can be estimated by logistic regression and shows what type of data is required for PD model estimation. However, the method of PD model estimation in the current research is enriched by multicollinearity analysis through correlation matrix and economic power considerations by using Kullback-Leibler information divergence.

The focus of previous research is to show that the statistical model can be used for underwriting purposes and prove that the credit policy based on the statistical model is more efficient compared to manual underwriting policy. The previous research is based on the data from another bank specialized in unsecured products. The generalized result of that research is that every bank might benefit from using statistical models for underwriting purposes.

The purpose of current research is to check whether the introduction of the IRB approach would increase the bank's profitability. While this research partly utilizes findings of my previous research, the purpose and contribution of new research differ remarkably from the previous one. The PD model estimation is only one part needed for capital requirement calculation under the IRB approach. This research covers the loss given default (LGD) part

also using the collateral data, which is unavailable in the previous research. Hence, the outcome of current research depends on the portfolio's PD and LGD distributions and the bank would not necessarily benefit from the PD model estimation in terms of capital requirement regardless of the contribution level of PD model itself to underwriting methods.

3. Basel Accords

The banks traditionally have a high level of leverage, as it is obvious that banks funded completely by equity are not feasible. The leverage means that a bank is funded mainly by deposits, bonds, and other liabilities. Hence, equity capital represents just a small part of banks' assets, which is the main reason why banking is the most regulated private business in the world. (de Servigny 2004, 387.)

3.1. Value at risk

The main guiding principle of the Basel regulation is presented in Figure 3.1. Banks and other financial institutions regularly suffer from expected and unexpected credit losses. The expected part of the losses – expected loss (EL) is considered as the cost of doing business and should be covered not from the capital but from annual revenues, which means that EL is factored into the products' pricing. In other words, expected losses are provisioned for, where provisions are losses in the P&L statement. This explains why expected loss is not part of the capital requirement calculation and is covered by another banking regulation – IFRS9, which is out of the research question scope.

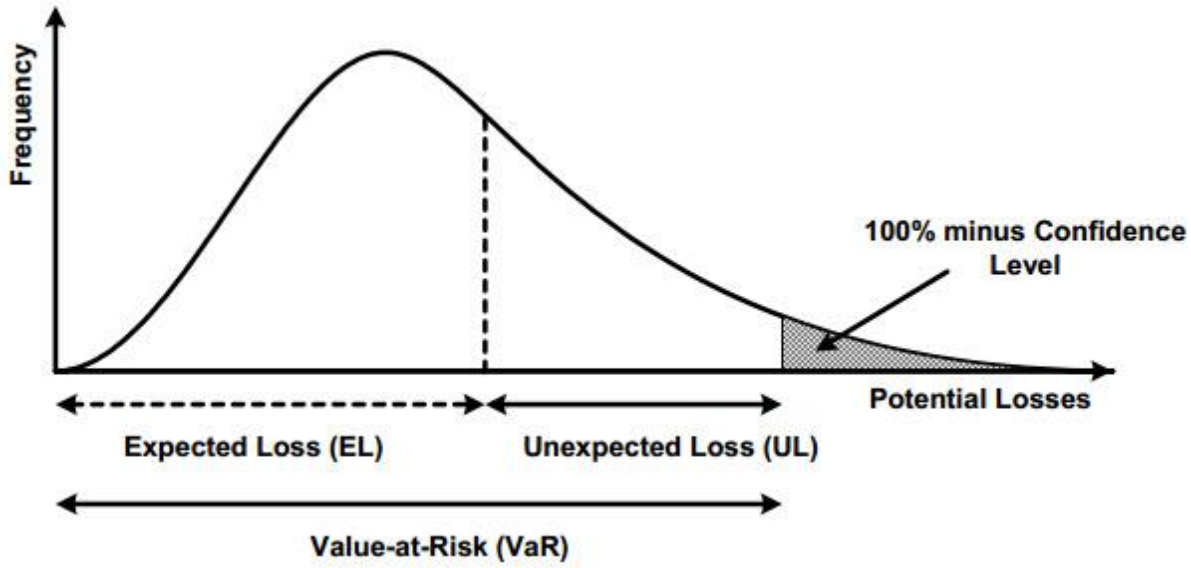
The expected loss is calculated according to the formula shown below:

$$EL = PD \times LGD \times EAD \quad (3.1)$$

where PD is the probability of default, LGD is loss given default and EAD means exposure at default. The next section focuses on a detailed description of these parameters.

The unexpected loss (UL) is the difference between the value at risk (VaR) and the expected loss (EL). The bank or financial institution calculates the value at risk in a similar way as an expected loss but uses downturn estimates for PD, LGD, and EAD, which means that these measures are stressed in order to reflect economic downturn – adverse economic scenario. (Baesens 2016, 11.)

Figure 3.1 Value-at-Risk model



The unexpected loss is not provisioned through profit and loss statement, therefore it must be charged against the capital. Banks and financial institutions have to ensure that they have sufficient capital to cover UL in order to protect the depositors and other creditors. Otherwise, bank insolvency or failure happens, which means that the bank is not able to fulfill its obligations anymore and the government should cover depositors' losses by deposit insurance. Hence, the goal of the regulation is to establish a procedure estimating the potential unexpected loss on a regulatory probability level. (Witzany 2017, 15.)

3.2. Capital requirements

The pillar of Basel regulation, which reflects the minimum capital requirement, is called the capital adequacy ratio (CAR) and is defined as the regulatory capital divided by the risk-weighted assets (RWA):

$$CAR = \frac{Regulatory\ Capital}{RWA} \geq 8\% \quad (3.2)$$

The regulatory capital is an amount of capital a bank should have according to regulation in order to protect itself against unexpected loss (UL).

Risk-weighted assets (RWA) is a sum of agreements' exposure at defaults (EAD) multiplied by the agreement's specific risk weights (RW):

$$RWA = \sum_{i=1}^{n} RW_i \times EAD_i \quad (3.3)$$

where n refers to a number of agreements in the portfolio.

The regulatory capital can be derived from the formula(3.2):

$$Regulatory\ Capital \geq RWA \times 0.08 \quad (3.4)$$

Hence, the research focuses on RWA estimation, which can be calculated using several approaches: standardized (SA) and internal rating-based (IRB).

3.3. Standardized approach

Under the standardized approach, the risk weights are set by regulation based on the product and the type of borrower. For instance, the government and central bank obligations have a risk weight equal to 0%, while retail exposures have a risk weight equal to 75% (non-mortgage). The risk weights provided by the regulation do not have a direct connection to the IRB approach, thus the regulation does not provide information under what credit parameters or other assumptions these risk weights are estimated.

Although the standardized approach looks simple and transparent, it suffers in terms of coverage of various types of exposures. The retail exposures are discriminated only in terms of mortgage or non-mortgage. (Baesens 2016, 10.)

As the research covers non-mortgage and non-revolving retail portfolios, the straightforward conclusion is that under the standardized approach the risk weight is equal to 75%.

3.4. Internal rating based approach

Under IRB banks are supposed to use their quantitative models to estimate PD (probability of default), EAD (exposure at default), LGD (loss given default) and other parameters required for calculating the RWA (risk-weighted asset). These parameters should be in line with the value-at-risk model, thus, PD and LGD should be stressed against the economic downturn.

The internal models have to satisfy minimum standards set by the local regulator in terms of methodology, data quality, length of the observation period, etc. (Witzany 2017, 111.)

For instance, the minimum length of the observation period is 5 years for PD estimation and 7 years for LGD estimation. In addition, the parametric models are preferred to non-parametric, as the main advantage of parametric models is their interpretability.

Now, the total required capital is calculated as a fixed percentage of the estimated RWA according to the formula (3.4).

IRB benefits banks and financial institutions to hold lower capital requirements as having a low default portfolio, which means the portfolio of agreements with lower probability of default compared to the assumption made in the standardized approach. (de Servigny 2004, 400.)

The internal rating-based approach can be divided into two sub approaches – foundation (F-IRB) and advanced (A-IRB). In the foundation IRB, only PD is calculated internally, while LGD is defined by Basel Accords or local regulation. In the advanced approach, the bank estimates all parameters internally and then utilizes these parameters in risk weight calculation. For retail portfolio the foundation approach is not permitted, hence the research focuses on the advanced IRB approach (A-IRB). (Baesens 2016, 11.)

4. Credit measures and risk weight formula

This section contains the information regarding credit risk measures needed for the risk weight estimation under the A-IRB approach. According to the formula (3.3), RWA is a sum of agreements' exposure at defaults (EAD) multiplied by the agreements' specific risk weights (RW) and reciprocal of the minimum capital ratio. Hence, the risk weight is estimated on an agreement level; accordingly, all credit measures required for risk weight estimation are estimated on an agreement level as well.

4.1. Probability of default

A probability of default (PD) describes the likelihood of a default event. The default definition used in the research is Basel's definition, which is based on payment delinquency of 90 days or more within 12 months after loan origination. In other words, a defaulter is defined as an obligor who has days past due (DPD) more than 90 days within the next 12 months after loan origination. (Baesens 2016, 138.)

According to Basel Accords, the bank should stress PD estimate via the concept of a worst-case default rate given a virtual macroeconomic shock based on a confidence level of 99.9% and a sensitivity to the microeconomic conditions that are based on the asset correlation. This logic is included in the risk weight formula, which means that no other PD adjustments are required. (Joseph 2013, 293.)

The idea of the research is to apply logistic regression to a sample of historical loan agreements in order to estimate the probability of default for an existing portfolio. The decision to use the logistic regression is based on a publication of Desai, Crook and Overstreet (1996), where the authors have proved that logistic regression is a most appropriate approach compared to other parametric and non-parametric (machine learning) approaches for binary response variables.

In binary logistic regression, the response variable $y \in \{0,1\}$ follows a Bernoulli distribution. The response variable reflects the default defined by Basel Accords on an agreement level:

$$y = \begin{cases} 1, if\ DPD > 90\ within\ 12\ months\ after\ origination \\ 0, otherwise \end{cases}$$

The vector of agreements $\boldsymbol{y} = (y_1, y_2, \dots\ y_n)^{\,T}$ should be independent, where n indicates the number of agreements in the sample. This assumption holds for the financial institution's dataset, as each customer has only one loan agreement and customers' payment behavior is expected to be independent on a portfolio level.

The logistic regression has the logit link function, which is an inverse of a standard cumulative distribution function of the logistic distribution. The logistic regression model has a linear form for the logit:

$$logit(PD_i) = \log\left(\frac{PD_i}{1 - PD_i}\right) = \boldsymbol{\beta}^T \boldsymbol{x_i} \ (4.1)$$

where $PD_i = P(y_i = 1)$ is a probability of default for observation i, $\boldsymbol{\beta}$ is a regression coefficient vector and $\boldsymbol{x_i}$ is an $(m + 1)$ vector containing m explanatory variables and a constant term. (Agresti 2015, 2.)

The possible explanatory variables for the probability of default estimation can be the duration of living at current address, employment length, customer's age, education, payment behavior, and other variables gathered from various sources of the data discussed more in detail in section 6.

The probability of default PD_i can be derived from equation (4.1) by using the exponential function:

$$PD_i = \frac{\exp\{\boldsymbol{\beta}^T \boldsymbol{x_i}\}}{1 + \exp\{\boldsymbol{\beta}^T \boldsymbol{x_i}\}} \ (4.2)$$

The standard way to estimate a logistic regression model is a maximum likelihood estimation. (Agresti 2007, 6.). The likelihood function $L(\boldsymbol{\beta})$ is the probability for the occurrence of a sample $\boldsymbol{y} = (y_1, y_2, \dots\ y_n)^{\,T}$ given the Bernoulli probability density:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} PD_i{}^{y_i}(1 - PD_i)^{1-y_i} \ (4.3)$$

The log-likelihood function $l(\boldsymbol{\beta})$ is equal to:

$$l(\boldsymbol{\beta}) = \log\big(L(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} y_i \log(PD_i) + (1 - y_i)\log(1 - PD_i) \ (4.4)$$

Its first derivative with respect to β_j where $j \in \{1, \dots, m + 1\}$ is equal to:

$$\frac{dl(\beta)}{d\beta_j} = \sum_{i=1}^{n} (y_i - PD_i \log(PD_i)) x_{ij} \quad (4.5)$$

The maximum likelihood estimates for $\boldsymbol{\beta}$ can be found by setting each of the $(m+1)$ equations defined in (4.5) equal to zero and solving for each β_j.

4.2. Loss given default

The loss given default (LGD) means a ratio of the loss on exposure due to the default of an obligor to the amount outstanding at default. LGD is calculated as below:

$$LGD = 1 - RR \quad (4.6)$$

where RR means Recovery Rate. Recovery rate varies from 0% to 100% and refers to the recoverable part of the credit asset. The recovery rate is equal to the present value of all cash flows to the bank (payments from obligor or collateral realization) after the date of default divided by exposure at default, thus, LGD can be defined as:

$$LGD = \frac{EAD - \sum_{t=1}^{T} (CF_t / (1 + r_t)^t}{EAD} \quad (4.7)$$

where CF_t is the cash flow and r_t is a discount rate for time t.

While downturn PD is extrapolated from bank-reported PD via mapping function, which is part of the risk-weight formula, in case of LGD, the bank is asked to provide downturn LGD by own mapping function or based on their internal assessment of LGD during adverse conditions. (Bank for International Settlements 2005, 7.)

The U.S. Department of the Treasury, Federal Reserve System, and Federal Deposit Insurance Corporation (2006) proposed a linear relationship between the downturn LGD (DLGD) and the expected LGD. The formula implies a floor of 8% and a cap of 100%: (Baesens 2016, 460.)

$$DLGD = 0.08 + 0.92 \times LGD \quad (4.8)$$

In order to estimate expected LGD on the agreement level, the bank utilizes marginal LGD modeling. This means that LGD is estimated based on the data from defaulted obligors only and do not consider those cases in which no default happened.

Unlike the probability of default (PD) which follows Bernoulli distribution, LGD is a variable which primarily follows a Beta distribution**.** (Yang, Tkachenko, 2012)

The beta distribution for the LGD has two parameters α and β and has the form:

$$f(LGD) = \frac{1}{B(\alpha,\beta)} LGD^{\alpha-1}(1-LGD)^{\beta-1} \quad (4.9)$$

where $B(\alpha,\beta)$ is the beta function.

The standard model for LGD estimation is the beta regression model, which is closely related to the beta distribution. Both parameters of the beta distribution are transformed into a location (mean) parameter μ and a shape parameter δ such that parameter $\alpha = \mu\delta$ and parameter $\beta = (1-\alpha)\delta$. (Baesens 2016, 296.)

A regression model is then applied to the location and the shape parameter transformations:

$$\text{logit}(\mu) = \beta_\mu^T X \iff \mu = \frac{\exp\{\beta_\mu^T X\}}{1+\exp\{\beta_\mu^T X\}} \quad (4.10)$$

$$\log(\delta) = \beta_\delta^T X \Leftrightarrow \delta = \exp\{\beta_\delta^T X\} \quad (4.11)$$

where β_μ and β_δ are regression coefficient vectors and X is an $n \times (m+1)$ model matrix containing n observations of m explanatory variables and a constant term.

Both coefficient vectors β_μ and β_δ are estimated by maximum likelihood, similar to **the** one described in **Section 4.1. (Section with capital "S")**

The possible explanatory variables for LGD estimation can be collateral type, the size of down payment, age at default, and other variables gathered from various sources of the data discussed more in detail in the next section.

4.3. Exposure at default

Exposure at default (EAD) represents the expected level of usage of the facility utilization when default occurs. Theoretically, as maturity increases, risk increases, so the probability of credit loss increases as well. As the research concentrates on a retail portfolio, which does not include revolving products, the exposure at default is equal to exposure drawn amount and no further modelling **is** required. In other words, EAD is defined as the nominal outstanding balance on an agreement level.

4.4. Risk weight under A-IRB

According to Basel II, the formula for the agreement specific risk-weight (RW) calculation under advanced IRB for retail exposures is given below:

$$RW = 12.5 \times \left(LGD \times N\left(\frac{1}{\sqrt{1-R}} \times G(PD) + \sqrt{\frac{R}{1-R}} \times G(0.999) \right) - LGD \times PD \right) \quad (4.12)$$

where 12.5 refers to reciprocal of the minimum capital ratio of 8%, N refers to cumulative normal distribution and G to the inverse of cumulative normal distribution. Accordingly, $G(0.999)$ is an inverse of cumulative normal distribution variable for 99.9% confidence interval, which means that the institution is expected to suffer losses that exceed regulatory capital on average once in a thousand years. (Bank for International Settlements 2005, 11.)

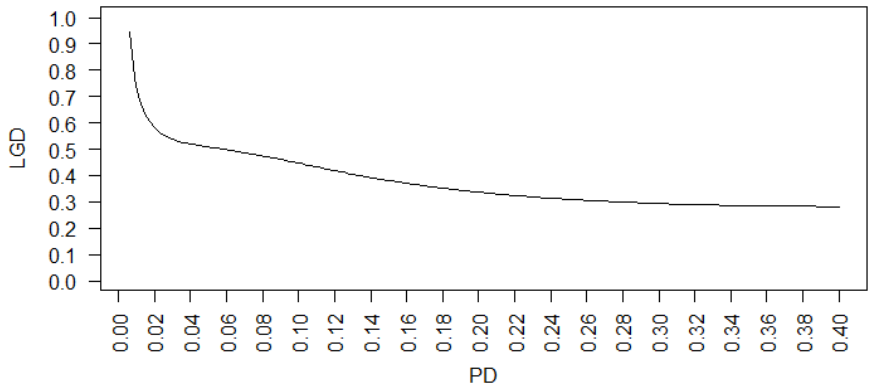
The correlation factor R depends on PD and asset class. The correlation factor for non-revolving and non-mortgage exposures uses the next formula:

$$R = 0.03 \times \frac{1 - \exp\{-35 \times PD\}}{1 - \exp\{-35\}} + 0.16 \times \left(1 - \frac{1 - \exp\{-35 \times PD\}}{1 - \exp\{-35\}}\right) \quad (4.13)$$

This formula gives a correlation between 3% and 16% depending on the probability of default value.

In order to illustrate the dependency of risk weight from PD and LGD under the IRB approach for a retail portfolio, the next figure is presented based on a formula (4.12).

Figure 4.1 Combinations of PD and LGD values satisfying 75% risk weight



This figure shows what PD and LGD levels are required to have a 75% risk weight on the portfolio level. For instance, if PD is equal to 4% (0.04) then LGD should be equal approximately to 55% (0.55) in order to reach risk weight equal to 75%.

The risk weight equal to 75% is selected because it is a level defined by a standardized approach. In other words, the curve shows what combinations of PD and LGD levels within the portfolio are expected by Basel accords under the standardized approach.

If the bank's average PD and LGD combination lies below the line, the introduction of the IRB approach would decrease capital requirement. Otherwise, the introduction of the IRB approach would increase capital requirement compared to a standardized approach. Hence, the IRB approach is not necessarily beneficial for a bank but the outcome depends on the portfolio's PD and LGD levels.

5. Research question

The scope of the research question is a non-revolving non-mortgage portfolio and a private customer segment. The scope is defined by the customer segment and the product space the bank operates with. The idea of the research is to check whether the higher return on the allocated capital may be achieved if the bank calculates its capital requirement based on the IRB approach instead of the standardized approach. The research question hypothesizes that the introduction of the IRB approach may lower the capital allocated on the retail portfolio, as the bank currently uses the standardized approach for the capital requirement calculation.

5.1. Return on allocated capital (ROAC)

The bank maximizes the profit measured by the return on allocated capital (ROAC). This measure is equal to net income divided by regular capital allocated:

$$ROAC = \frac{net\ income}{regular\ capital\ allocated} \quad (5.1)$$

By inserting formulas (3.3) and (3.4) into (5.1), the return on allocated capital can be expressed as:

$$ROAC = \frac{net\ income}{0.08\ \times \sum_{i=1}^{n} RW_i\ \times EAD_i} \quad (5.2)$$

where risk weight (RW) is applied to agreement's exposure at default and either estimated on an agreement level under A-IRB or defined by regulation under the standardized approach.

The bank currently uses the standardized approach, where the risk weight for retail exposures is equal to 75%, thus the bank calculates return on allocated capital for a retail portfolio as:

$$ROAC_{Standardized} = \frac{net\ income}{0.08\ \times \sum_{i=1}^{n} 0.75\ \times EAD_i} \quad (5.3)$$

5.2. The research question hypothesis

The hypothesis of the research question is that ROAC under A-IRB approach $ROAC_{IRB}$, where agreement based risk weights (RW) are estimated internally by the formula (4.12), is higher compared to ROAC under bank's current (standardized approach) $ROAC_{Standardized}$, where the risk weight for retail exposures is equal to 75%.

Mathematically, the expected result of the research question can be expressed as:

$$ROAC_{IRB} > ROAC_{Standardized}$$

$$\Leftrightarrow$$

$$\frac{net\ income}{0.08 \times \sum_{i=1}^{n} RW_i \times EAD_i} > \frac{net\ income}{0.08 \times \sum_{i=1}^{n} 0.75 \times EAD_i}$$

which means that

$$\sum_{i=1}^{n} RW_i \times EAD_i < \sum_{i=1}^{n} 0.75 \times EAD_i$$

Hence, the research question aims to estimate agreement based risk weights under A-IRB for the bank's retail portfolio and to test the hypothesis. If the research proves the hypothesis validity, the bank has incentive to initiate the transition from a standardized approach to A-IRB in order to increase its portfolio's profitability and attractiveness from the investors' perspective.

6. Data sources

The financial intermediary has to collect all data, which is relevant for risk management. The data available for customers' credit assessment is of three types: those derived from the application form, those available from a credit bureau search, and those describing the transaction history of the borrower. (Peussa 2016, 5.)

However, these types mentioned above cover only the data valid for the probability of default (PD) estimation. In order to estimate the loss given default (LGD), the collateral data is required.

6.1. Application data

Application data is information provided by the potential customer. The application form could contain such characteristics as salary, occupation, number of children, other loans and so on. When designing an application form, a financial institution faces a trade-off between the simplicity of the form and the quantity of the information. (Peussa 2016, 5.)

A detailed application form, with a great variety of variables, is attractive from the risk management perspective. Clearly, the more variables available a lender has, the better the model can be constructed. The representative application form contains the next variables: age, gender, education, employment type and length, accommodation type, income, loan obligations, etc.

However, a long or too detailed application form decreases the probability of its completion, which cuts sales. Hence, there is a pressure on the lender to make the form as simple as possible. Furthermore, some questions are not permitted for legal reasons. For instance, the U.S. Equal Credit Opportunity Acts of 1975 and 1976 made it illegal to discriminate in the granting of credit on the grounds of race, color, religion, national origin, sex, marital status, or age. (Thomas 2002, 124.)

The application data is a significant instrument from the risk modeling perspective but it contains problems as well. First, the data should be carefully looked through and validated. The most frequently arising problems with the application data are frauds and errors. An applicant could unintentionally violate information, putting the wrong income, for example.

This problem could be partly solved by revealing and eliminating impossible or inconsistent answers. (Peussa 2016, 6.)

The frauds are a more serious problem. Here, an applicant intentionally violates application data in order to receive a positive loan decision or better offer. Moreover, if the intermediary has economic incentives for sending good customers, they might be tempted to advise the applicant on suitable answers. Therefore, from a modeling perspective, the usage of only application data could lead to major problems in the end. This is the reason why credit bureau data is a significant part of IRB modeling.

6.2. Credit bureau data

A credit bureau or credit reference agency is an organization which collects data from various sources and provides consumer credit information on individual consumers. The data provided by the credit bureau includes its estimate of customers' probability of default, socio-demographics, applicant's borrowing, and bill-paying habits. The purpose of the credit bureau is to reduce the impact of asymmetric information between borrowers and lenders. (Peussa 2016, 6.)

From the IRB modeling perspective, the position of a credit agency is very important. It has two advantages compared to a bank: the credit bureau validates the data using municipality information and its customer base, which contains millions of applications and historical records and is much larger than the bank's portfolio. Hence, a lender protects himself against violated information, which could happen if the bank uses application data only.

6.3. Behavioral data

Behavioral data is a history of existing customers' transactions and cash flows. In other words, the data is a conglomeration of customers' payment characteristics and habits. The most common ones are minimum, maximum or average balance, total value or regularity of both debit and credit transactions, payment defaults and other delinquency indicators. (Peussa 2016, 10.)

6.4. Collateral data

The collateral data includes information about customers’ paid down payment and the financed asset's value. The customers' paid down payment increases the level of loans' securitization and via this decreases the bank's collateral risk.

The collateral value of the asset financed is not equal to the purchase price of the asset. The reason behind this is that collateral value should cover unfavorable scenarios like poor maintenance of the asset or negative changes in the market demand for the asset. In other words, it should express the expected value of an asset if customer defaults and the bank realizes the asset. The collateral value estimate generally is based on variables like asset type, mileage, and age. The bank provides either its collateral value estimate or orders it from the third parties like asset evaluation companies.

7. Optimal model selection

Input parameters of risk weight formula - PD and LGD estimates are based on internal modeling, which is not covered by regulation, which means the better PD and LGD models perform, the more precise risk weight estimate is achieved. This section describes the methodology utilized for constructing an optimal model. The optimal model means the best model from a performance perspective with respect to the data available.

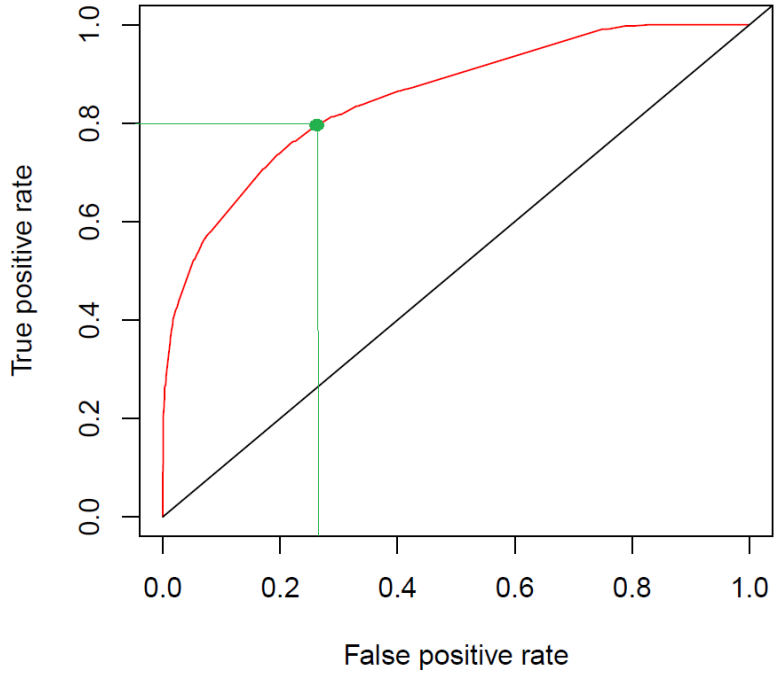
7.1. PD model selection

The response variable $PD \in \{0,1\}$ follows a Bernoulli distribution, which means that a receiver operating characteristic curve (ROC curve) may be used for the model's performance estimation.

The ROC curve is a graphical plot that illustrates the performance of a binary classifier model, thus it is mainly used for Bernoulli distributed response variables. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold points. The threshold point means the value of the model's estimate at which the agreement is regarded as positive. In terms of the research question model, positive means defaulted obligor (PD = 1), while negative means non-defaulted obligor by Basel definition.

The TPR defines how many correct positive results occur among the sample's positive results. FPR, on the other hand, defines how many incorrect positive results occur among the sample's negative results. For instance, the green point in figure 6.1 shows that at this specific threshold, the model predicts correctly 80% of positive ones and 27% is a share of false-positive ones from the sample's negative ones. The red curve in figure 6.1 represents a ROC curve for a model estimated. The diagonal line on the figure shows the "random guessing" model which is expected to have no separating power at all that turns into the equal TPR and FPR rates. Intuitively, the bigger distance is between diagonal and the ROC curve, the better is a model's predictive power.

Figure 7.1 ROC curve



Based on the ROC curve, the Gini coefficient may be calculated, which is used as a model performance indicator. The Gini coefficient is defined as the ratio of the area between the diagonal and ROC curve to the area of the triangle formed by the diagonal and axes.

If the ROC curve is approximated on each interval as a line between consecutive points, then the Gini coefficient can be approximated with trapezoids as:

$$Gini = 1 - \sum_{k=1}^{n} (FPR_k - FPR_{k-1})(TPR_k + TPR_{k-1}) \quad (7.1)$$

where FPR_k and TPR_k are indexed in increasing order such that $FPR_k > FPR_{k-1}$ and $TPR_k > TPR_{k-1}$, so that:

- FPR_k is the false positive rate for k = 0,…,n with $FPR_0 = 0$ and $FPR_n = 1$
- TPR_k is the true positive rate for k = 0,…,n with $TPR_0 = 0$ and $TPR_n = 1$

In other words, the ROC curve consists of a set of points, where every point represents the false and true positive rates of the sample picked according to a cut-off presented by this particular point. The cut-off means an estimate of the probability of default provided by the model.

The value of the Gini coefficient may fluctuate between zero and one, where one means the perfect classification model, and zero means "random guessing" model. According to the bank's internal governance, it is forbidden to use models with the Gini coefficient lower than 10% (0.1), which gives a minimum requirement for PD modeling.

The Gini coefficient for a set of candidate models is used and the model with the highest Gini coefficient is selected.

7.2. LGD model selection

In contrast to PD, LGD is not Bernoulli distributed, thus another performance measurement metric is used. Given a set of candidate models, the Bayesian information criterion (BIC) supports the selection of the most appropriate model, which has a balance between the goodness of the model, measured by the maximum value of the likelihood function, and simplicity, measured by several explanatory variables in the model. (Peussa 2016, 20.)

It is defined as

$$BIC = -2\ln(\hat{L}) + k\ln(n) \quad (7.2)$$

where $\hat{L}$ is the maximum value of the likelihood function for the model, k is the number of parameters in the model and n is the number of observations.

The most appropriate model from the selection criterion perspective minimizes the BIC value.

8. Estimation

This section contains models estimated for the probability of default (PD) and loss given default (LGD). Based on these models, the agreements' specific risk weights and through this the risk-weighted assets are calculated, which are required for capital requirement calculation under the IRB approach.

8.1. Data

The research is based on the bank's internal data, stored in the bank's data warehouse (DWH). Every type of data presented above is stored in the own SQL database, thus the aim is to prepare a final data set that includes all the data required for the research question purposes. Therefore, an SQL query is created, which combines and links all the data required in one sample. Due to the bank's operational challenges related to the data, the research is based on the constrained sample in terms of date range limitation and absence of part of LGD related data like final loss data.

The scope of the research question is a non-revolving non-mortgage portfolio and private customer segment. Hence, the data set includes household-level data from consumer secured products like car financing. Mortgages and revolving financial products like credit cards are not included in the data set as they are out of the scope of the research question. The data is based on a period from the beginning of July 2017 till March 2018. The initial dataset includes 2093 observations and 52 variables, from which

- 48 variables are potentially explanatory for PD model
- 1 variable is a response variable for the PD model, thus **it** reflects the default event defined by Basel Accords
- 1 variable, which reflects customer’s paid down payment
- 1 variable, which reflects the asset’s collateral value
- 1 variable, which reflects an agreement's exposure at default (nominal balance)

Potential explanatory variables are based on the data from the application, the credit bureau and the bank's internal data reflecting customers' past behavior. However, due to compliance and business reasons, all potential explanatory variables are not raw data but are already transformed according to the bank's internal policy standards into a rule-based format. Hence,

all potential explanatory variables in the data set are binary. Due to corporate security reasons, the exact nature of potential explanatory variables is not revealed. That is why labels of these variables are recoded into "A_" form.

The data regarding the collateral value of the asset is provided by a third-party – asset evaluation company.

8.2. PD model

Based on the bank's internal recommendations the data set used for PD model estimation is split by train (2017 H2) and test (2018 Q1) samples according to a timestamp. The training sample contains 1588 observations and the test sample contains 505 observations. The training sample is used for model development and the test sample is used for model validation. The timestamp-based split is explained by the need to both ensure that there is no overfitting issue and that model is stable in time.

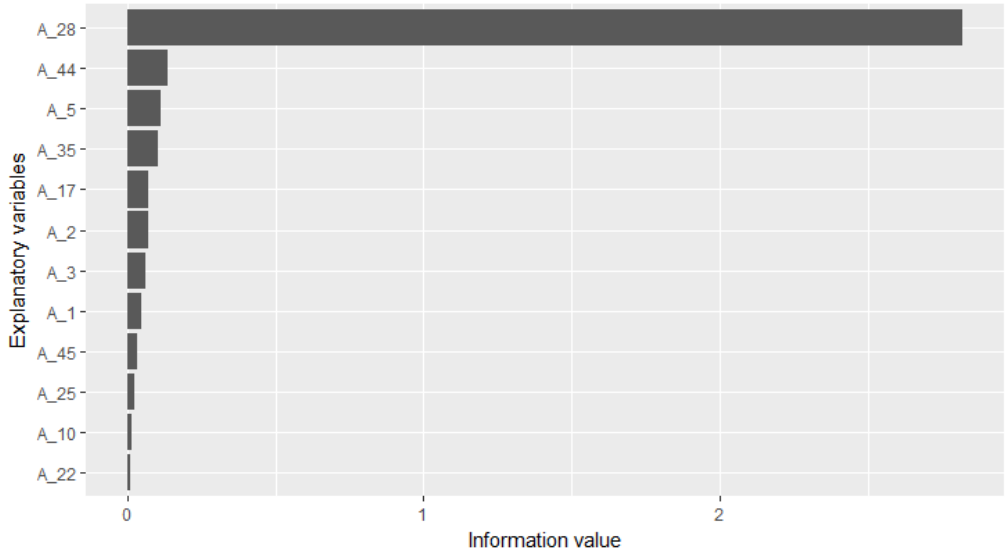
For the optimal model selection, a stepwise methodology is utilized based on backward elimination. In other words, the method involves starting with a model performance check for the model including all candidate variables and testing the model performance again after deletion of the least important variable. The importance of the variable is measured by both economic power (information value) and statistical significance (p-value).

The information value (IV) measures the economic power of the variable and is based on Kullback-Leibler information divergence:

$$IV = D_{KL}(P\|Q) + D_{KL}(Q\|P) = \sum_{I}[P(i) - Q(i)] \times \log\left(\frac{P(i)}{Q(i)}\right) \quad (8.1)$$

where P and Q are the variable's probability distributions for defaulted and non-defaulted agreements. In other words, P tells how defaulted agreements are distributed between the variable's categories and Q tells how non-defaulted agreements are distributed between the same variable's categories. The higher the information value variable receives, the higher is the difference between defaulted and non-defaulted agreements' distributions, the higher is the economic power of the variable. Based on the bank's internal recommendations, the variables with economic power below 0.02 should not be included in the model.

Figure 8.1 Economic power of candidate variables measured by information value

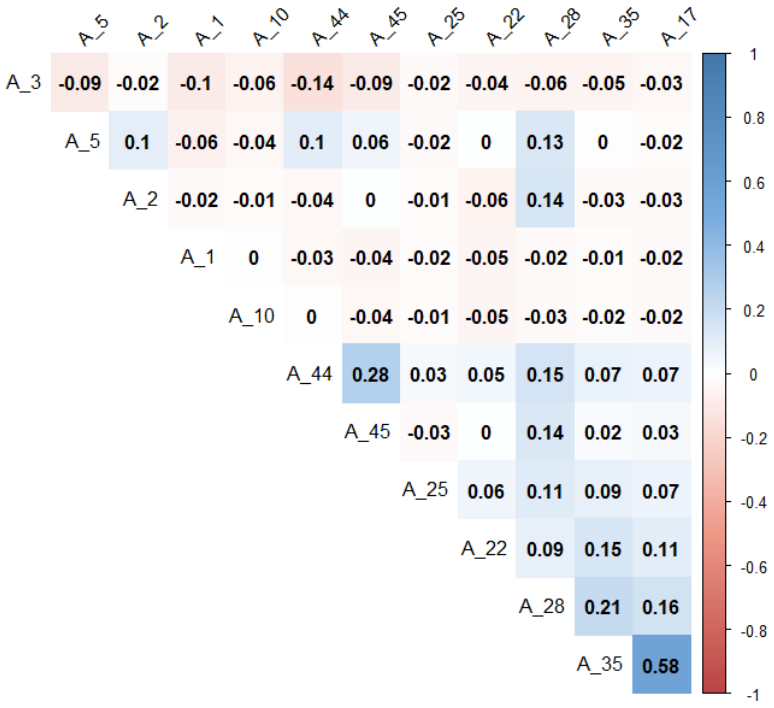


The majority of the variables are excluded from the candidate list due to their irrelevancy, economic power weakness or compliance reasons.

For instance, the U.S. Equal Credit Opportunity Acts of 1975 and 1976 made it illegal to discriminate in the granting of credit on the grounds of race, color, religion, national origin, sex, marital status, or age. (Thomas 2002, 124.)

After first exclusions based on economic power weakness, compliance considerations like discrimination risk and business reasons, the candidate list includes 12 potential explanatory variables.

Figure 8.2 Correlation matrix of candidate variables



These candidates are checked by a correlation matrix in order to verify that there is no risk of multicollinearity.

Based on Figure 8.1 the variable A_28 has exceptionally strong economic power. This is explained by the nature of the business rule behind the variable, which reveals a high credit risk level associated with the population for which this business rule is relevant. Furthermore, based on the finding the bank's credit policy is updated. According to the credit policy update, the population for which business rule under variable A_28 is relevant cannot be approved anymore. Hence, this population is out of the scope of the research question as the capital requirement is required only for the approved population.

After the exclusion of the population related to variable A_28, the dataset includes 11 potential explanatory variables and 1745 observations from which 1331 observations belong to the training sample and 414 observations belong to the test sample.

For PD model estimation, the backward elimination principle is used supported by economic power (IV) considerations and correlation matrix. The idea is to ensure that explanatory variables remaining in the model would provide the highest possible economic power keeping the correlation between explanatory variables as low as possible. As the purpose is

to use a training sample for model estimation, the model estimation is performed on the dataset including 1331 observations.

After several iterations, the final model is produced consisting of six explanatory variables and an intercept. Every explanatory variable is significant at least at the 10% significance level. The nature of variables in the final model cannot be revealed due to corporate security reasons, but what is noteworthy is that variables, based on information stated by the customer, are not part of the final model due to weak significance. This can be explained by intentional or unintentional information violation by the customer. (Peussa 2016, 30.)

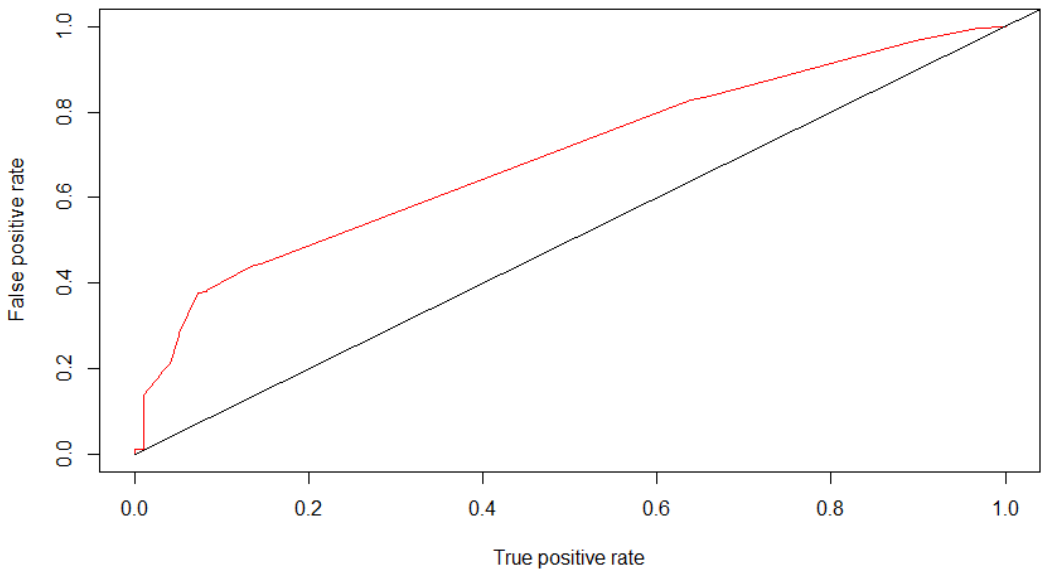
Table 8.1 Coefficient estimates for the probability of default model

Variable	Estimate	Standard error	Significance level
Intercept	-2.4135	0.1511	<0.01
A_46	-1.8227	0.5936	<0.01
A_1	-14848	0.7261	<0.05
A_13	1.0416	0.3351	<0.01
A_44	0.6002	0.2339	<0.05
A_45	-1.0390	0.4094	<0.05
A_3	-0.8550	0.4390	<0.1

The final model includes variables based on credit-bureau data and the bank's internal data. All variables are checked from a business perspective in order to verify that no illogical behavior exists. Each variable has acceptable economic power and a coefficient sign indicates what the variable's effect on credit risk is. The positive sign means that the variable has a positive correlation with credit risk, and a negative sign means that the variable has a negative correlation with credit risk.

The Gini coefficient derived from the ROC curve measures the performance of the final model based on the training sample.

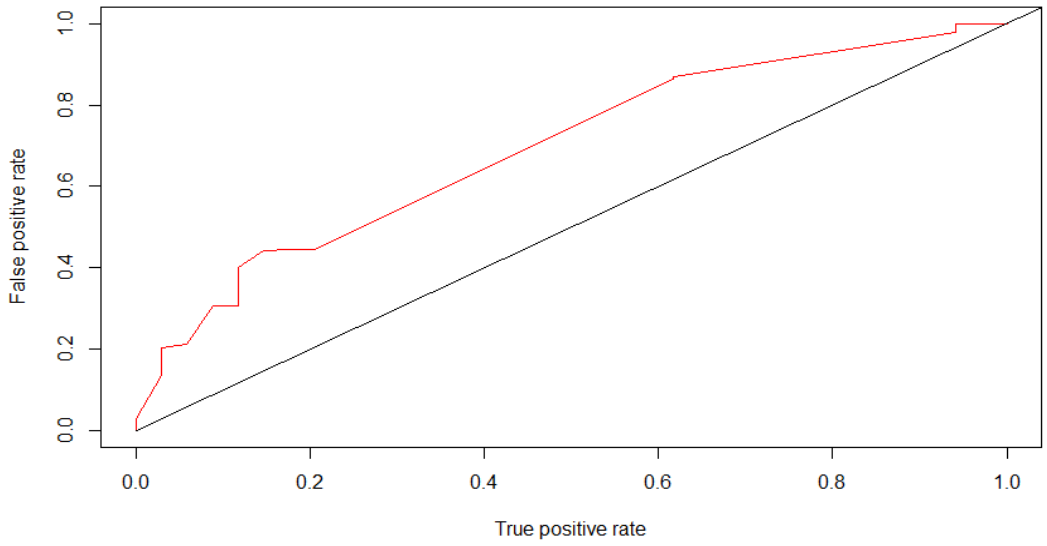
Figure 8.3 ROC curve of the final PD-model (train sample)



The Gini coefficient of the final model based on the training sample is 38.44%, which indicates that the performance of the model is on an acceptable level.

In order to verify that the overfitting issue does not appear and the model is stable in time, the Gini coefficient is derived for the test sample as well:

Figure 8.4 ROC curve of the final PD-model (test sample)



The Gini coefficient of the final model based on the test sample is 38.12%, which indicates that the performance of the model remains on an acceptable level and the model is stable in time.

8.3. LGD model

Due to sufficient data unavailability for LGD estimation, the simplified LGD model is used for risk weight estimation based on the linear relationship with securitization level (SL). Securitization level is defined as an underlying asset's collateral value plus a customer's paid down payment divided by the exposure amount:

$$SL = min\left(\frac{collateral\ value + down\ payment}{EAD}, 1\right) (8.2)$$

Fully secured exposure (securitization level equal to 100%) means that an underlying asset's collateral value is equal to or higher than the customer's exposure amount. Fully unsecured exposure has securitization level equal to 0%. Based on the bank's historical data for mortgages and unsecured credit cards aligned with Financial Supervisory Authority (FSA), the downturn LGD estimate for fully secured exposure is 10% and for fully unsecured exposure is 44%. This means that based on the bank's historical data gathered from a period of time, which reflects the economic downturn, the bank experienced a loss of 44% from the exposure at default for defaulted customers in case of credit card products and 10% in case of mortgage products.

Based on these estimates, the linear equation below is proposed for downturn LGD estimation:

$$LGD = 0.44 - 0.34 \times SL\ (8.3)$$

where LGD is downturn loss given default estimate and SL is agreement's securitization level.

The equation is derived by using DT LGD estimates mentioned above and assuming the linear relationship between the level of securitization (SL) and loss given default (LGD). Generally, the equation is calculated using two points: (0, 0.44) and (1, 0.10) where the first point refers to fully unsecured exposure and the second point refers to fully secured exposure.

8.4. Risk-weighted assets

The main component for capital requirement calculation under the IRB model is an agreement specific risk weight. Having PD and LGD models in place, agreement specific risk weight is calculated based on a formula (4.12):

$$RW_i = 12.5 \times \left(LGD_i \times N\left(\frac{1}{\sqrt{1-R_i}} \times G(PD_i) + \sqrt{\frac{R_i}{1-R_i}} \times G(0.999) \right) - LGD_i \times PD_i \right)$$

where RW_i is agreement specific risk weight, LGD_i is agreement specific loss given default estimated by formula (8.2), PD_i is agreement specific probability of default estimated by PD model provided in Section 8.2 for agreement i.

The agreement specific risk weights are calculated for the total dataset excluding variable A_28 related population**,** meaning that risk weight formula is applied to 1745 observations.

Now, based on a formula (3.3), the sample's risk-weighted assets may be calculated under the IRB approach:

$$\sum_{i=1}^{1745} RW_i \times EAD_i = 0.27 \times EAD$$

where RW_i refers to agreement specific risk weight based on formula (4.12), EAD_i indicates agreement's specific exposure at default, 1745 shows the number of agreements in the dataset and EAD is a sum of all agreements' exposures at default in the dataset.

9. Results

The hypothesis of the research question is that ROAC under A-IRB approach $ROAC_{IRB}$, where agreement based risk weights (RW) are estimated internally by the formula (4.12), is higher compared to ROAC under bank's current (standardized approach) $ROAC_{Standardized}$, where the risk weight for retail exposures is equal to 75%.

So, the expected result of the research question can be expressed as:

$$\sum_{i=1}^{1745} RW_i \times EAD_i < \sum_{i=1}^{1745} 0.75 \times EAD_i \quad (9.1)$$

The second part of inequality may be expressed as:

$$\sum_{i=1}^{1745} 0.75 \times EAD_i = 0.75 \times EAD \quad (9.2)$$

where EAD is a sum of all exposures at default within a dataset.

Based on agreement specific risk weights, which are estimated by applying the final PD and LGD model presented in Sections 8.1 and 8.2, the risk-weighted assets under the IRB approach are equal to:

$$\sum_{i=1}^{1745} RW_i \times EAD_i = 0.27 \times EAD \quad (9.3)$$

According to ROAC definition based on formula (5.2) and utilizing estimation results from (9.2) and (9.3) standardized and A-IRB based returns on allocated capital may be presented as:

$$ROAC_{IRB} = \frac{net\ income}{0.08 \times 0.75 \times EAD}$$

$$ROAC_{Standardized} = \frac{net\ income}{0.08 \times 0.27 \times EAD}$$

In terms of ROAC comparison, the estimated result means that return on allocated capital under A-IRB is 177.8% higher compared to that return on allocated capital under standardized model:

$$\frac{ROAC_{IRB} - ROAC_{Standardized}}{ROAC_{Standardized}} \times 100\% = 177.8\%$$

Hence, based on estimation results, the research proves the hypothesis validity and supports the portfolio transition from a standardized approach to A-IRB in order to increase the portfolio's profitability in terms of return on allocated capital (ROAC).

Figure 9.1 Combinations of PD and LGD values satisfying 75% and 27% risk weights

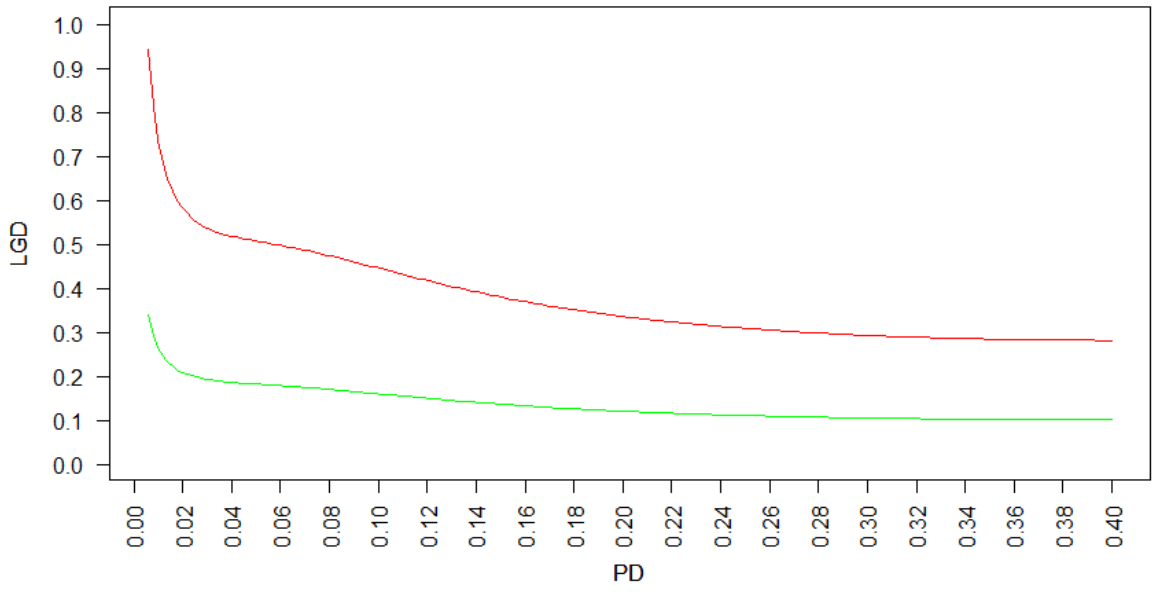


Figure 9.1 tends to illustrate the difference in PD and LGD combinations for portfolio having the same risk weight under the IRB approach as in a standardized approach (red line) compared to the bank's portfolio (green line). It is clear that the combinations of PD and LGD for the bank's portfolio are much lower than expected by Basel Accords under the standardized approach. Hence, taking into account the bank's PD and LGD on a portfolio level, it is logical to expect that it would require smaller capital requirements compared to banks, which tends to operate under the standardized approach.

10. Conclusion

The idea of the research is to check whether the higher return on allocated capital may be achieved if the bank calculates a capital requirement based on the IRB approach instead of a standardized approach. The research question hypothesized that the introduction of the IRB approach may lower the capital allocated to the retail portfolio.

IRB approach uses financial institution's own credit risk related estimates rather than whole market expectations presented by standardized approach under Basel III. If the bank's customer base has lower credit risk compared to the market expectations, there is a perfect sense to the bank to adopt the IRB approach.

The capital requirement under the IRB approach is calculated through the estimation of the probability of default (PD) and loss given default (LGD) based on the bank's historical data. The historical data for the probability of default calculation includes application, credit bureau, and behavioral data. It is noteworthy to mention that behavior and credit bureau data explain best the probability of default, while application data is proven to be irrelevant based on research findings. The probability of default is estimated using logistic regression, which is recommendation-based on the bank's compliance policy. The historical data for loss given default estimation is unavailable due to technical gaps in the bank's data warehouse; hence, the simple linear formula is applied based on historical downturn LGD high-level aggregates.

The research covers the bank's retail portfolio and provides evidence that if a bank adopts the A-IRB approach, the expected profitability increase of the retail portfolio is 177.8%. The estimate presented in this research is not final, as the research implies some assumptions to LGD estimation, which decrease the accuracy of the prediction made. However, such a remarkable increase in profitability proves the validity of the hypothesis even if some margin of conservatism is applied.

The research proves that the application of the IRB approach significantly lowers a bank's capital requirement and as a result increases a return on allocated capital.

11. Bibliography

1. Agresti A. (2015). “Foundations of linear and generalized models.” Wiley
2. Baesens B., Rösch D., Scheule H. (2016). “Credit risk analytics.” Wiley
3. Balin B. J. (2008). “Basel I, Basel II, and emerging markets: A nontechnical analysis.”
4. Bank for International Settlements (2005). “An Explanatory note on the Basel II IRB Risk Weight Functions.” Bank for International Settlements
5. Desai V.S., Crook J.N., Overstreet G.A. (1996). “A comparison of neural networks and linear scoring models in the credit environment.” European journal of operational research, 95, 24-37
6. European Banking Authority (2017). "EBA Report on IRB modeling practices". European Banking Authority
7. Joseph C. (2013). “Advanced credit risk analysis and management.” Wiley
8. Lewis E.M. (1992). "An Introduction to Credit Scoring." Athena Press
9. Peussa A. (2016). “Credit risk scorecard estimation by logistic regression.” Pro gradu. Department of mathematics and statistics. University of Helsinki
10. De Servigny A., Renault O. (2004). “Measuring and Managing credit risk.” McGraw-Hill
11. Thomas L.C. (2002). "Credit scoring and Its Applications." SIAM Monographs on mathematical modeling and computation
12. Witzany J. (2017). “Credit Risk Management.” Springer
13. Yang B.H., Tkachenko M. (2012). “Modelling of EAD and LGD: Empirical Approaches and Technical Implementation.” Journal of Credit Risk, 8