Comparison of Traditional Modelling Techniques and Machine Learning for Prediction of LGD

This paper provides model documentation for comparison of traditional modeling techniques such as logistic and linear regressions and Machine Learning (ML) methods conducted by FCG on the Global Credit Data (GCD) Large Corporates default data. The analysis covers historical averages, regression analysis and machine learning for predicting LGD and probability of Cure. We further explore the potential of the dataset by increasing the number of independent variables from known risk drivers to all applicable information provided within the GCD dataset's framework. We also explore the possibility of changing the GCD definition of "cure" in order to increase predictive power of the models.

The models are built on the data on Large Corporate defaults, provided by GCD, which includes defaults from all over the world, but is dominated by observations from North America and Europe. Results show that a Machine Learning model can perform better than the traditional regression model. This can be seen both when using an extended number of risk-drivers and when restricted to a more limited traditional set only. We could not identify any additional specifically strong risk drivers besides the original ones, but the overall predictive power for probability of cure, explained by Area under the Curve (AUC) of the ML model increased from 0.82 to 0.86. We also assessed whether altering the GCD's implemented definition of "cure" would add explanatory power to a predictive model but found no significant improvement, suggesting that the existing definition is in some way optimal.

Contents

1 Introduction

1.1 Background

Estimation of Loss Given Default (LGD) can substantially affect a bank's business as many regulatory requirements are related to this metric, including minimum capital levels and profit provisions for loan performance. This makes LGD one of the most important metrics in credit risk management and, therefore, the precision of its estimation is extremely valuable. Contemporary achievements in the field of Machine Learning (ML) allow a new level of precision within a variety of fields. However, officials are often reluctant to allow machine learning within such a sensitive industry as banking due to consequent model risk. In order to shed light on this issue, we decided to combine the power of machine learning with the large number of default observations gathered in GCD's LGD/EAD pooled database. The data provided by GCD is especially interesting for this kind of research, because it aggregates the data from various banks around the world, dramatically increasing the number of observations available for study compared to what is usually available for bank's internal models, but also increasing the variability of observations. Thus, the model development work here is also a test of whether the GCD data and data model are suitable for use in forward looking LGD estimation models, not just historical descriptions.

1.2 Purpose

The main purpose of this paper is, using a pooled data set of default data, to evaluate if ML can increase the accuracy of LGD prediction compared to traditional pooling and regression techniques. The main question of the study therefore is whether ML is worth the model risk it entails. In the development of ML "challenger" models, we also explore if ML can help in discovering additional risk drivers apart from those commonly used when estimating LGD. We also briefly address modeling the cure definition.

1.3 Scope

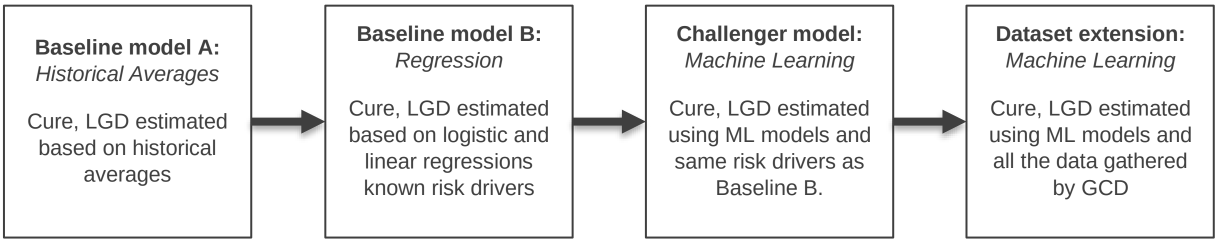
Baseline model A:
Historical Averages
Cure, LGD estimated based on historical averages
Baseline model B:
Regression
Cure, LGD estimated based on logistic and linear regressions known risk drivers
Challenger model:
Machine Learning
Cure, LGD estimated using ML models and same risk drivers as Baseline B.
Dataset extension:
Machine Learning
Cure, LGD estimated using ML models and all the data gathered by GCD

Figure 1. Description of model development

This study focuses on three main steps to compare the modeling techniques: data quality control, development of baseline models, and development of challenger models. As a part of data quality control and in order to secure the adequacy of data for modeling and comparability with results from the LGD Report 2019 - Large Corporate Borrowers (Rainone & Brumma, 2019), the paper starts with a replication of outcomes from the report and description of the data quality checks conducted by GCD and FCG to achieve the highest possible quality of the data. After that, two baseline models are designed using traditional modeling techniques, namely, historical averages and regressions. The purpose is to establish a benchmark to which ML models can be compared in order to conclude if ML delivers a significant improvement in LGD prediction accuracy. Using ML, a set of challenger models is designed and implemented. There are two different model classes implemented. The first one predicts LGD using risk drivers comparable to the baseline models. The second one explores the full potential of GCD dataset and is built on extended risk driver set. As mentioned before, the reason to add more risk drivers is to evaluate if ML techniques, having the strength to predict from datasets with few data points, can find risk drivers not previously considered relevant for traditional models.

1.4 Application

These models are intended to be a reference to any bank interested in exploring the ML approach when estimating LGD. However, a direct recreation of the model might be inadvisable due to the differences in actual loan portfolios, macroeconomic factors affecting different banks and other unique features. Therefore, it is recommended to consider the provided models as a collection of best practices and commonly used methods, applicable to estimating LGD. We note GCD's advice to always commence using GCD data sets by creating a restricted "Representative Data Set" from the master data supplied, ensuring compatibility with the member's portfolio.

2 Data and Population

2.1 Data source

The data for this study is provided by GCD, which collects the data from its member banks according to pre-determined rules. The GCD data quality standards have been developed by practitioners from the member banks over the past decade to meet the requirements of regulatory, business and accounting purposes. In order to ensure a high quality of its data, GCD establishes a number of requirements on the submissions. These standards are in compliance with international rules and regulations such as Basel, IFRS9, CECL. GCD limits data contribution to banks complying with the Basel II rules regarding Advanced Internal Rating Based approach as they must collect and maintain the data necessary to build models. Local rules and regulations are addressed indirectly, as each bank can tailor the data to its needs.

2.2 Population

This study is conducted on the defaulted exposures within GCD member banks' large corporate borrowers. The same dataset has been previously used for LGD Report 2019 - Large Corporate Borrowers (Rainone & Brumma, 2019). The original data contains information on defaults between the years 1998 and 2015. However, only the years 2000 to 2015 were chosen for this study due to data completeness reasons. This is in line with the GCD LGD Report 2019 (Rainone & Brumma, 2019) where it is explained that pre-2000 defaults were not completely collected as they occurred prior to the Basel default definition while post-2015 defaults are not yet completely collected as banks await the outcomes of sometimes long workout activities.

2.3 Data quality control

A pooled data set requires its user to filter out irrelevant observations in order to create a representative data set to ensure that the data matches the user's portfolio. This study is based on LGD Report 2019 - Large Corporate Borrowers (Rainone & Brumma, 2019) and its underlying data, the GCD Large Corporate Reference Data Sets (RDS) was replicated. This step ensures that the report results can be accurately reproduced and, consequently, that the data quality is compatible with previously reported results.

In this paper, authors' model loan level LGD, but it is also possible to model a borrower level LGD. Then, the pool of observations was reduced to Large Corporates senior non-syndicated loans only, which were previously included in the RDS. A conservative assumption was made regarding missing values of syndication indicator, so they were also excluded from the dataset. The labels for Securitization, Seniority, and Syndication were replicated manually based on the corresponding features (indicators). To do so, exposures were divided based on underlying facility type. Then the debt instruments were assigned Seniority labels (Senior or Subordinated) and other exposures were assigned the Other/Unknown label. In other words, specific filters were applied to the data following the logic and descriptions from the original report. In the end, the graphs have been generated and the results corresponded to the ones reported by Brumma and Rainone (2019).

Another data quality control performed is analysis of descriptive statistics of the variables. The checks included calculating extrema, mean and median, identifying outliers, wrong values, counting the number of missing values in respect to the meaning of the variables. The controls were executed in relation to consequent variable selection and feature transformation and the variables presented in this study showed a good quality.

2.4 Data processing and transformation

After recreating the RDS, the dataset is prepared for modeling. First, variables are chosen as risk drivers for the models. The historical averages model explores five variables in line with the GCD LGD report. The number of independent variables was increased to cover 25 important and known risk drivers for the regression and challenger models. In the end, additional variables provided within the GCD data are included.

After the risk drivers are chosen, the selected data are formatted to be consistent with the requirements of the different models, i.e. type conversion, recoding missing values, etc. Missing data are imputed with median values and outliers winsorized (capped) to a boundary percentile value. Lastly, feature engineering is used for the data transformation and additional features such as ratios, log values, and encodings are created from the risk drivers. In the end, the data is normalized or Weight-Of-Evidence (WOE) transformed.

2.4.1 Risk-driver selection and feature transformation

This section is focused on feature engineering and transformations of variables that are included in the models. Numeric type features can be calculated in such a way that the feature is transformed either to a log value or ratios of two features. Feature transformation is presented in three sets. First list contains seven variables that are processed in order to create the groupings criteria for the historical averages model. Second part lists the popular risk drivers, created to compare the regression and ML models. The last part contains the rest of the variables used to extend the dataset.

Within the framework of the historical averages model, the observations are sorted into different groups according to the following features:

1. **Collateral Label** – a dummy variable based on whether the loan is secured or not. In case there is no information on collateral behind the loan, it is treated as non-secured.
2. **Collateral Type** – Collateral types securing the loan which the lender can usually get control of and sell if necessary.
3. **Seniority Code** – is a more detailed equivalent of Seniority Label provided within GCD RDS, consisting of five values: Super senior, Pari-passu, Junior, Equity and Unknown.

4. **Country of Residence** – is a variable describing the borrower's country of residence.
5. **Downturn Flag** – accounts for economic downturn and marks all observations from default year 2001, 2002, 2008, 2009 and European observations in default year 2012.

During the trials, other grouping variations were explored. **Seniority Label**, which is an ordinal variable that includes values "Senior", which is assigned to the super senior and pari-passu loans, "Subordinated" which corresponds to junior or subordinated loans, and "Other/Unknown" when the seniority of loan cannot be determined or is unknown. Country of Jurisdiction (the country of the court specified in the loan documentation), as provided by the dataset was tried instead of the country of residence. However, the model performance is marginally worse while using these variables.

A larger number of features is used to compare the regression and ML models. Besides previously described features, the following variables are included:

1. **Default Lender Borrower Risk Rating** – information about borrower's internal default rating.
2. **Initial Lender Borrower Risk Rating** – information about borrower's original internal default.
3. **Log** (**EAD 1)**– is common logarithm of default amount.
4. **Default Loan/Limit 1** – is a ratio of default amount to Lender Limit at the point of default.
5. **Log** (**EAD 2)** – is common logarithm of default amount.
6. **Default Loan/Limit 2** – is a ratio of default amount to Lender Limit at the point of default.
7. **Default LTV** – is a ratio of default amount and the total of collateral values at the point of default.
8. **Default LTV 1 Flag** – is a dummy variable indicating that Default LTV equals or exceeds 1.
9. **Initial Loan Amount log** – a common logarithm of the initial loan amount.
10. **Initial LTV –** is a ratio of initial loan amount to collateral value at the origination.
11. **Initial LTV 1 Flag** – is a dummy variable indicating that Initial LTV equals or exceeds 1.
12. **EAD 1/Initial Loan Amount** – is a ratio of default amount represented by the variable Default Amount 1 to Lender Outstanding Amount at the time of loan origination.
13. **EAD 2/Initial Loan Amount –** is a ratio of default amount represented by the variable Default Amount 2 to Lender Outstanding Amount at the time of loan origination.
14. **Initial Share Real Estate –** is a ratio of Collateral Value for Real Estate collateral at the time of loan origination to the Initial Loan Amount.
15. **Default Share Real Estate** – is a ratio of the Collateral Value at the time of default to the default amount.
16. **Initial Share Other –** is a ratio of other than real estate collateral to the loan amount at the origination of the loan.
17. **Default Share Other –** is a ratio of other than real estate collateral to the loan amount at the loan's default.
18. **Initial Loan/Limit –** is a ratio of initial loan amount to initial lender limit.
19. **Mean Entity Assets log** – this variable describes the size of the company represented by the common logarithm of the average assets per entity as recorded in the Financial table: $\log_{10}(\frac{\sum entity\ assets}{n(DA\ Entity\ ID)})$. In case when all records for an entity are missing, the variable is assumed to be equal to zero.

20. **Mean Entity Sales log** – this variable describes the size of the company represented by the common logarithm of the average sales per entity as recorded in the Financial table: $\log_{10}(\frac{\sum entity\ sales}{n(DA\ Entity\ ID)})$. In case when all records for an entity are missing, the variable is assumed to be equal to zero.
21. **Mean Guarantee Percentage** – is an average Guarantee Percentage per loan ID as recorded in the Guarantor table.
22. **Primary Industry Code** – is a categorical variable describing the industry that accounts for the largest percentage of the Entity's revenues, as recorded in the Entity table.

As the analysis was extended across the dataset, some additional risk drivers were included in the model:

1. **Loan Spread** – is a numeric column, represented by the Total Spread and, in case the total spread is missing, the Spread column as recorded in the pricing column. The variable is transformed to non-negative.
2. **Base Rate** – the numerical categorical variables, describing the base rate type (LIBOR, EURIBOR, etc.) as recorded in the pricing table.
3. **Total Rate** – a sum of Loan Spread and Base Rate
4. **US segment** – a field calculated by GCD which segments the data between shared segments from US members as recorded in the Loan table.
5. **Facility Type** – facility type as recorded in the loan table.
6. **Nature of Default** – is a categorical variable that indicates the first reason why the lender has put the Borrower in default (Basel II guidelines) at the Event Date (Default Date).
7. **Rank of Security** – the rank of collateral security aggregated on the loan level. For the loans with several different collaterals a "Subsequent Charge" category was assigned.
8. **Committed Indicator** - the contractual obligation for the bank to "make the funds" when the facility is drawn by the client, as recorded in the loan table.
9. **Leveraged Finance Indicator** – indicates acquisition finance or leveraged buyout at the time of default.
10. **Financial Currency** – the currency denomination of the Entity Financials same currency for all financial figures, as recorded in the financial table.
11. **Public-Private Indicator** – a categorical variable that provides further information on the ownership of the company (publicly traded/privately owned/SPV).

2.4.2 Outliers and missing values

In order to prepare the data for further transformations, the data quality had been studied. Variables are dropped if over 50% of observations are missing. In some cases, a null value may indicate a 0, "No" or "N/A" instead of a missing value. This requires a detailed understanding of data.

The initially engineered features $\frac{EAD_1}{Initial\ Loan\ Amount}, \frac{EAD_2}{Initial\ Loan\ Amount}$, and $\frac{Initial\ Loan}{Limit}$ were dropped after this exercise due to the large number of missing values. Many GCD member banks do not report data from loan origination. During trials, the model with a threshold of 40% was investigated, but its predictive power is lower.

The chosen method of data transformation is robust to outliers because it categorizes data and an outlier ends up in one of the categories. Therefore, the final model is not affected by outliers even when they exist. However, an alternative where for all numeric type features, outliers are floored and capped to a boundary percentile value which are 3% and 97% respectively was also tried. Among variables that have been truncated are Default LTV, Default Share Real Estate, Initial Share Real Estate, Default Share Other, Initial Share Other.

2.4.3 Weight of Evidence transformation

Weight-Of-Evidence (WOE) method was chosen to transform the variables and prepare them for modelling. WoE is a robust to outliers method which splits observations of a given explanatory variable into groups depending on their informational value about the dependent variable. For the P(Cure), which is a discrete variable, transformation is done according to the following formulas:

$$WoE_i = \ln\left(\frac{p_{cure\ i}}{p_{not\ cure\ i}}\right) \tag{1}$$

$$IV = \sum_{i=1}^{n}(p_{cure\ i} - p_{not\ cure\ i}) \cdot WoE_i \tag{2}$$

The estimated LGD given non-cure is a continuous variable, and research suggests using a modified WoE, where percentages are used instead of the number of observations, to transform its risk drivers. However, we do not have the opportunity to transform variables once again after the prediction on Cure. Therefore, the WoE is based on $P(Cure)$.

2.5 Training and validation dataset split

The transformed dataset containing 16674 defaults between the years 2000 and 2015 is split into a training and a validation set. A random split where 80% of the dataset or 13 339 observations are randomly selected into the training set and the rest 20% (3 335 observations) are used for validation purposes is chosen. This approach allows conserving the variables' probability distributions for which the model is based on, which is in line with the purpose of the study.

Another approach where the training dataset contains loans defaulted from 2000 to 2011 and the validation set contains the loans defaulted in 2012 to 2015 was also studied within the framework of the Historical Averages modeling approach. This was an interesting exercise because such an approach to splitting datasets is widely used. Such a split provides newer data but does not cover the whole business cycle. Therefore, the default distribution within the validation dataset might differ substantially from the one in the training dataset.

3 Methodology

3.1 Assumptions

2019 Corporate LGD Report (Brumma & Rainone, 2019) assumes that $E(LGD|Cure) = 0$. According to expert judgment from member banks however, this assumption is oversimplified and cannot be applied to real-world problems. Therefore, it is assumed that the estimated loss given default for the cured loans is equal to the actual mean of LGD among cured loans:

$$E(LGD|Cure) = a$$

$$where\ a = (LGD|Cure)_{Mean} > 0$$

3.2 Important definitions

This section presents the definitions of the core concepts/dependent variables used in this paper. Note, that all the variables used in the model follow definitions assigned by GCD. Due to the variability of use of the data and requirements for its quality in different jurisdictions, GCD standards focus on the global rules and regulations, such as IFRS9, CECL, TRIM, and Basel. The definitions within the dataset are therefore compliant with those provided by the regulations.

3.2.1 Definition of Default

A default event is defined in accordance with the Basel accord. The nature of default differs from loan to loan and can imply 90 days past due, sale at material credit loss, distressed restructuring, non-accrual status, charge off or specific provision, obligor’s unlikeliness to pay or bankruptcy. Banks providing the data used their internal definitions of “unlikeliness to pay”, which is an inevitable source of difference between data from each bank.

3.2.2 Definition of Cure

A cure event is defined in accordance with CGD internal guidelines. GCD member banks have agreed on the following definition of cure: A default having time to resolution < 1 year, no write-off and no collateral sale or guarantee call. All these items are collected separately as inputs in the data template and the cure marker is calculated by GCD.

Within this research, other definitions of cure were explored within the framework of the cure predictions, with main focus on the time to resolution which was shrunk to 30 days and stretched up to 5 years. We have also explored a possibility to define a Cure as a function of LGD. The reason behind this exercise is to explore if there are more predictable definitions. For more information on this see Section 4.6.

3.2.3 Definition of Loss Given Default

The Loss Given Default (LGD) is calculated by GCD based on the loan information provided by member banks. Being a dependent variable in the models described in this paper, LGD is therefore defined in accordance with GCD internal methodologies as economic LGD where Principal Advance and Financial Claim are parts of the recovered amount. GCD uses a risk-free discount rate in the calculation of LGD and therefore the absolute levels of LGD are generally lower than the ones calculated by banks. The value of LGD for each loan is floored at 0% and capped at 150%. GCD employs several methods to calculate LGD, within the chosen framework, the amount is not aggregated on the borrower level.

3.3 Model design

Theoretical approaches to LGD estimation modeling suggest single-stage and multistage models (Tanoue et al., 2016; Kawada & Yamashita, 2013). For the ML model we employ the same structure as for the models developed using the traditional techniques, predicting the LGD in two steps, where the first step is to predict probability of Cure and the second is to predict the probability of LGD based on the previous predictions. Literature provides information on more complex multi-stage models that allow for a better precision, but they usually need to be tailored to the specific needs of a bank, while we are looking for an overall comparison between traditional modeling techniques and ML (Tanoue et al., 2016; Kawada &

Yamashita, 2013). Furthermore, such models are very useful when developed at the time of loan origination, then the first step is to predict the probability of default. The latter is out of scope for this research.

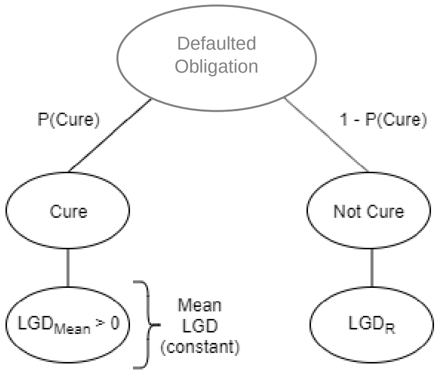

Figure 2. Model structure.

To adjust predictions to the LGD distribution, the expected LGD was broken into two separate models as shown in Figure 2. As $LGD_{Cure\ Mean}$ is assumed to be a constant, developed models are focused on separate prediction of $P(Cure)$ and $E(LGD\mid Not\ Cure)$ where $E(LGD\mid Not\ Cure)$ and $P(Cure)$ modeled in different ways. $E(LGD\mid Not\ Cure)$ is sometimes referred as Loss Given Loss, but we prefer to avoid this term as in many cases there can also be a zero loss. The overall structure is presented in Figure 2.

$$E(LGD) = P(Cure) \times E(LGD|Cure) + \big(1 - P(Cure)\big) \times E(LGD|Not\ Cure) =$$
$$= P(Cure) \times LGD_{Cure\ Mean} + \big(1 - P(Cure)\big) \times E(LGD|Not\ Cure)$$

For the historical averages model, the historical averages of cure and LGD are used for modeling. For the regression model, logistic and linear regressions are developed. The Challenger model is developed using ML, where a classifier method is used to predict the probability of cure and LGD in the case of non-cure is estimated using a regression technique.

As the model structure is set, several models with different features are tried within the given framework. Thus, different types of groupings are tried for historical averages model, while the regression models investigate the effect of different ways of handling missing data as well as the effect of using WOE instead of actual variables. While building the challenger models, different ML methods are explored. All the models are validated, then the development process and the best model's characteristics are described in this document.

3.4 Traditional modeling techniques

This section focuses on providing short explanations regarding the modeling techniques used for the final model, predicting LGD using the traditional model.

3.4.1 Historical averages

Historical averages are a straightforward modeling technique where LGD is predicted based on the previous observations about LGD within specified groups of the dataset. This technique was chosen because it is sometimes used in large banks (Severeijns, 2018) and is the most simple and transparent way to start an analysis.

Estimated LGD is predicted for each group based on the combination of the probability of the cure and LGD in the case of non-cure, which are given as follows:

$$P(Cure) = \frac{\#\, cured\ loans}{\#\, loans}$$

$$E(LGD \mid Not\ Cure) = (LGD_{mean}\ |Cure = N)$$

Groups are determined by the chosen risk-drivers, so each group has the same features. For more information on the risk-drivers see Section 2.4.1.

3.4.2 Logistic regression

Logistic regression is one of the most popular tools to model a binary outcome. This type of regression uses a logistic function to model a binary dependent variable. The function gives a sigmoid 'S' shaped curve, restricted between 0 and 1 modelling the probability of the event represented by the dependent variable. The coefficients are more difficult to understand than those of linear model, but easier than those of more advanced models. They represent the relationship between the explanatory variable and the logistically transformed dependent variable. According to the logistic regression, the probability of cure is given by:

$$P(Cure) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i \cdot x_i)}}$$

Where α is the intercept of the model β_i, is a slope coefficient and x_i is the value of the models i risk-driver.

In Python, logistic regression is provided within Scikit-learn package and is called via function:

```
LogisticRegression().fit(X train, y train)
```

The model performance is subsequently verified on the test dataset via function:

```
LogisticRegression().predict(X test, y test)
```

3.4.3 Linear regression

Linear regression is often used to predict a continuous outcome based on the given inputs. This is a simple and straightforward method allowing an intuitive understanding of the relationship between the dependent and the explanatory variables. As the $E(LGD|Cure)$ is assumed to be a constant, linear regression is used to model the LGD in case the loan has not been cured as follows:

$$E(LGD|Not\ Cure) = \ \alpha + \ \sum \beta_i x_i$$

Where α is the models' intercept, β_i is a slope coefficient and x_i is the value of the models i risk-driver.

In Python, linear regression is provided within Scikit-learn package and is called via function:

```
LinearRegression().fit(X train, y train)
```

The model performance is subsequently verified on the test dataset via function:

```
LinearRegression().predict(X test, y test)
```

3.5 Machine Learning

3.5.1 Decision trees

The decision tree method creates a simple model for each feature and then achieves precision by splitting the data with each step and predicting an outcome for each feature in a stepwise fashion. This method is robust to outliers and missing data and holds no assumption about the data distribution or multicollinearity within the data. It is also easy to interpret the model results. The disadvantage of the decision tree technique is that it chooses the best split for each feature within the given dataset which may lead to overlooking the best model in general and to overfitting.

3.5.2 Gradient boosting decision trees and random forest

To overcome the disadvantages of decision trees, the gradient boosting decision trees (GBDT) classifier was chosen for modeling the probability of cure. The python option XGBoost() runs an improved algorithm called extreme gradient boosting (XGBC for classifier, XGBR for regressor), which creates a sequence of models and then assembles them to create a more powerful prediction model. This technique allows for dealing with non-linear relations between the dependent and independent variables. XGBC/XGBR helps to improve predictions, i.e. choosing better decision trees, and to address the problem of overfitting, although does not completely solve it.

Another way to ensemble the decision trees is a random forest (RF). This technique builds separate decision trees on a multitude of combinations of randomly selected samples and features. Unlike the XGBC/XGBR, which ensembles a number of different decision trees on build on all given data, RF uses only a sub-sample of the data for each tree. This technique inherits all the advantages of the decision trees modeling and addresses its weak points. The resulting overfitting problem is solved in a better manner compared to GBDT. However, the interpretation of relations between dependent and independent variables is unclear.

XGBC classifier and RF classifier as well as their average results are investigated with the GridSearchCV function to find the best model for predicting Cure. XGBR regressor and RF regressor as well as their average results are investigated with the GridSearchCV function to find the best model for estimating LGD. The mean results of XGBC and Random Forest Classifier (RFC) and gradient boosting decision trees are proved to model the probability of cure in the best way and Random Forest Regressor (RFR) estimates LGD.

Neural networks (Keras DNN, MLPC) and Support Vector Machine were investigated along with the chosen modeling ways. However, there are not enough data to produce meaningful results with these methods.

3.6 Measures of predictive power

3.6.1 Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC)

This study uses ROC AUC analysis to measure the model performance when predicting the probability of cure. This measure is chosen because it is a commonly used metric for model comparison where the dependent variable is 0/1. The ROC area under the curve graphically shows the model performance power and has a direct explanation, representing a percentage of cases properly predicted by the model.

3.6.2 Mean Absolute Error (MAE)

To measure predictive power when estimating LGD in the case of non-cure, this study uses Mean Absolute Error (MAE) which is an average of absolute errors:

$$MAE = \frac{\sum_{i=1}^{n} |E(LGD|Not\ Cure)_i - (LGD|Not\ Cure)_i|}{n}$$

This is a widely used measure for assessing average model performance for non 0/1 dependent variable models. Among its advantages is that it does not penalize huge standalone errors and therefore is robust to outliers. The main disadvantage of MAE is that it, as well as most of the other error measures, does not properly reflect the bimodal distribution of the dependent variable, LGD. Thus, it chooses the model with the smallest possible error i.e. the one which prediction are mathematically closer to the true values, but another model might perform better in classifying loans by those in low and in high risk of large LGD. This disadvantage is partially addressed in the model structure attempting to first separate the "cure" cases to reach a normal distribution of the LGD when it is modelled. Although a complete separation is difficult to achieve and a bimodal distribution still exists within the framework of the presented models, a simple and explainable error measure that can be used for historical average, regression and ML models is preferred. Given this, MAE is deemed to be a good enough measure for model comparison.

3.6.3 Shapley Additive Explanations (SHAP)

Shapley Additive Explanations (SHAP) by Lundberg and Lee (2017) are used to assess individual risk drivers' power of prediction. Shapley values consider all possible predictions for an instance using all possible combinations of inputs, giving a summary scoring and ranking for all model features. The ranking shows what features contribute the most to the predictions and to what extent.

4 Analysis and Results

4.1 Traditional modeling techniques

4.1.1 Historical Averages

The historical average is a naïve model that attempts to predict future LGD based on the historical average LGDs within the different subgroups of the dataset. The observations are grouped based on five risk drivers described in Section 2.4.1. In total, there were 1 183 groups created, with the maximum 762 observations in a group, and minimum of one observation per group. One of the main disadvantages of this model is 374 groups out of 1183 consist of one observation, making the model prone to overfitting. Table 1 shows the number of categories within each variable:

Grouping criteria	Number of values	Type of variable
Country of residence	93	Categorical
Collateral label	2	Dummy
Collateral type	25	Categorical
Downturn flag	2	Dummy
Seniority	5	Categorical

Table 1. Grouping criteria for historical averages model

The $P(Cure)$ is modeled for each group. Figure 3 presents the true and predicted the distribution of Cure in the validation dataset with the split 80%/20%. As it can be seen, predicted $P(Cure)$ distribution differs from the actual distribution of "cure" and "non-cure" events. However, as for any two-state event, the application of a probability will always differ from the outcome.

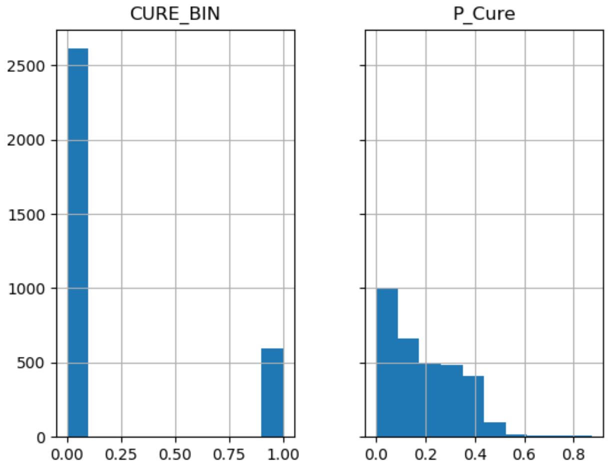

Figure 3. True (left) and predicted (right) Cure distribution

Furthermore, the AUC is 0,71 which is a typical result for an LGD model in the non-retail space. The MAE is 0,25 meaning that the predicted LGD differs from the true one. One of the reasons for this is an incorrectly predicted Cure in the previous step. As the historical averages model is also a way to understand the dataset, it was run on two different splits of the data. The results show that the Random split 80%/20% is preserving the distribution of the Cure and LGD while the split by year provides newer data that differs from the older one. Therefore, the random split is better for building a theoretical model.

Metrics	**Random split 80%/20%**	**Split by year**
AUC	0,7144	0,6305
MAE	0,2595	0,3014

Table 2. Summary results, historical averages model

4.1.2 Regression analysis

Baseline model B predicts future LGD based on the results of logistic and linear regressions. $P(Cure)$ is modeled using a logistic regression according to Lohmann and Ohliger (2019). $E(LGD \mid Not\ Cure)$ is a linear regression with known LGD risk drivers chosen according to Zhang and Thomas (2012) and Martinsson (2017). The model is developed to predict the LGD at the time of default. For this, a dataset prepared according to Section 2.3 is used: missing values are imputed, for the variables containing less than 50% missing values, otherwise, the variable is dropped, outliers are winsorized. The summary results are presented in Table 3.

Regression	**Metric**	
Logistic, predicting CURE	AUC	0,7224
Linear, predicting LGD	MAE	0,2664

Table 3. Baseline B results

After that the variables were ranked using SelectKBest() function. It allowed to create scores based on ANOVA F-value between features for Cure classification (f classif) and F-value between feature for regression model. The best 10 predictors for both Cure and LGD are presented in the table 4. As it can be seen, country of residence and country of jurisdiction are the most important factors in the estimated LGD. A possible reason for this is that a country label recognizes the different macroeconomics factors between countries (residence) while the jurisdiction label brings in the differing workout laws and practices between countries. Another important feature is, as expected, industry code, which can represent both the industry risk and the differing financial situations of companies between industries (e.g. leverage). Outside of the top three predictors, the ranking differs for Cure and LGD.

Rank	Feature Estimating Cure	Score	Feature Estimating LGD	Score
1	Country of residence	684	Country of residence	176
2	Country of jurisdiction	609	Country of jurisdiction	173
3	Industry	92.75	Industry	76.80
4	Sales log	72.08	EAD 2/Initial Loan Amount	73.41
5	Borrower Risk Rating	71.19	EAD 1/Initial Loan Amount	46.88
6	EAD 2/Initial Loan Amount	57.25	Initial Loan/Limit	44.69
7	Default Loan/Limit 2	54.26	Initial Loan Amount log	40.04
8	Initial Share Real Estate	52.80	Default Share Other	27.90
9	Default LTV	47.61	Mean Guarantee Percentage	26.52
10	Default Share Real Estate	46.52	Default Lender Borrower Risk Rating	13.65

Table 4 Ranking of best ten risk-drivers

4.2 Machine learning

The ML model is a logical continuation of the regression model. It has the same structure and the same set of risk-drivers is used. However, the challenger model uses ML to model $P(Cure)$ and $E(LGD \mid Not\ Cure)$. This approach allows for a direct comparison between traditional modeling and ML.

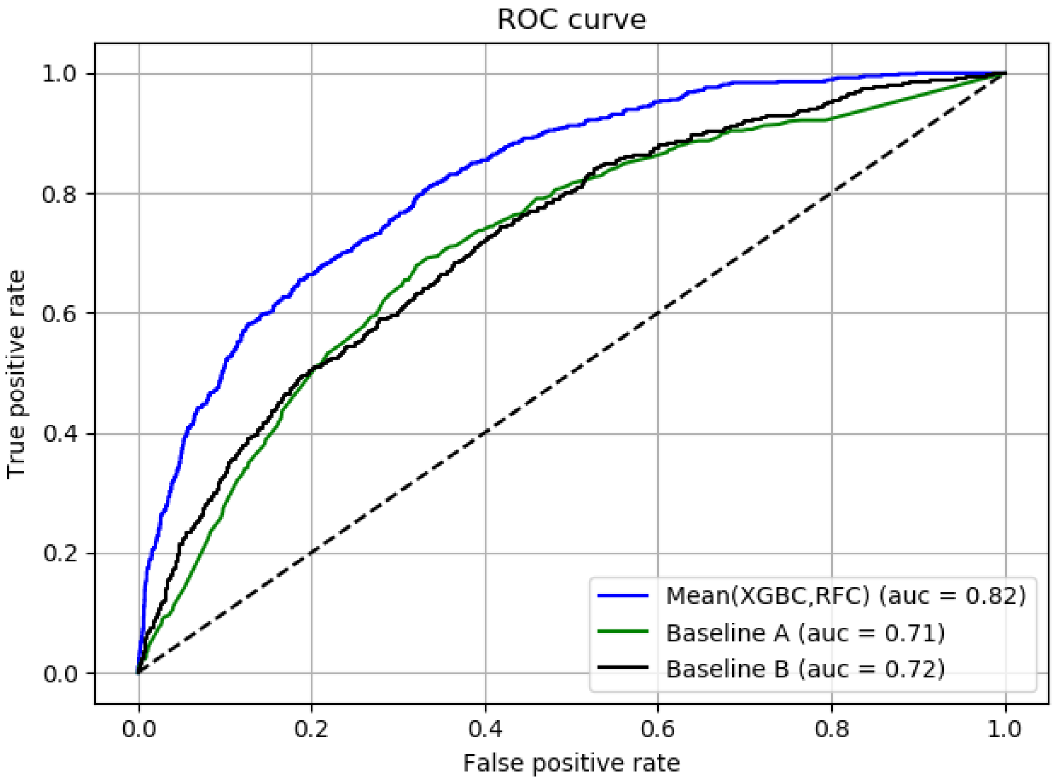

Figure 4. Comparison of the traditional modeling techniques and ML

For a better comparison between the models, each model's best hyperparameter setup has been investigated using scikit-learns GridSearchCV function. This python function searches for the best parameters using a fraction of the training data as a validation set. It repeats each unique parameter setup *k*-times (usually *k*=3) and picks the parameters with the highest average score. For classification (prediction of $P(Cure)$) the scoring is AUC, while for regression ($E(LGD \mid Not\ Cure)$), the scoring is MAE. The table below shows the results from the best performing model according to the gridSearch:

Model	Metric	
Mean (XGBC, RFC)	AUC	0,8231
RFR	MAE	0,2238

Table 5. ML model results

Machine learning shows better results compared to regression analysis or historical averages models. AUC increases by 0,1 which is substantial, but MAE decrease is not particularly large as it only decreased by 0,04 from 0,26 to 0,22.

4.3 Risk drivers' analysis

In order to assess the relevance of the risk drivers used in the models described in the previous section, they were assessed using SHAP and RFE analysis. RFE analysis recursively removes features, builds a new model using the remaining features and calculates AUC or MAE. The ranking shows what features, relative the other features, contribute the most to the predictions.

Both SHAP and RFE ranked Country of Residence and Country of Jurisdiction as the most useful features for classification of Cure and Non-Cure as well as LGD estimation. However, in a bank-developed model these features should be used separately due to high correlation.

In the case of the current model it is not a problem, because the scope of this research is a pure comparison.

Rank	XGBC (Cure)	RFR (LGD)
1	Country Of Jurisdiction	Country Of Jurisdiction
2	DA Country Of Residence	DA Country Of Residence
3	Default Share Real Estate	Primary Industry Code
4	Initial Lender Borrower Risk Rating	EAD 1/Initial Loan Amount
5	Mean Entity Sales log	Mean Entity Sales log
6	Mean Guarantee Percentage	Default Loan/Limit 2
7	Initial LTV	EAD 1 log
8	Initial Share Other	Mean Entity Assets log
9	Default Lender Borrower Risk Rating	Default Share Other
10	Primary Industry Code	Default Loan/Limit 1

Table 6 Top risk-drivers from SHAP and RFE analysis

4.4 Extending the analysis

4.4.1 Extending the number of model features

A natural question to ask is if the improved predictive power of ML-models can be extended even further by adding more model features to the dataset. To test this, all analysis was re-created using an extended dataset. The features added to the analysis are presented in table 7 and together with features presented in section 2.4.1 created the extended dataset. Baseline A models were excluded from the extension of the analysis.

Model features	Type	Model features	Type
Discount Rate	Numeric	Rank of Security	Dummy
Loan Spread	Numeric	Committed Indicator	Dummy
Base Rate	Numeric	Leveraged Finance	Dummy
Total Rate	Numeric	Collateral Label	Dummy
US segment	Dummy	Seniority Label	Dummy
Facility Type	Dummy	Financial Currency	Dummy
Nature of Default	Dummy	Public-Private Indicator	Dummy

Table 7. Extended features added to the analysis

4.4.2 Results

The result was very promising, both the Baseline B models and the Machine Learning Decision Trees models increased their predictive power both for predicting the probability of a cure event and LGD.

	AUC	MAE
XGBC + RFR	0.85	0.216
Baseline B	0.76	0.259

Table 8 Extended ML results

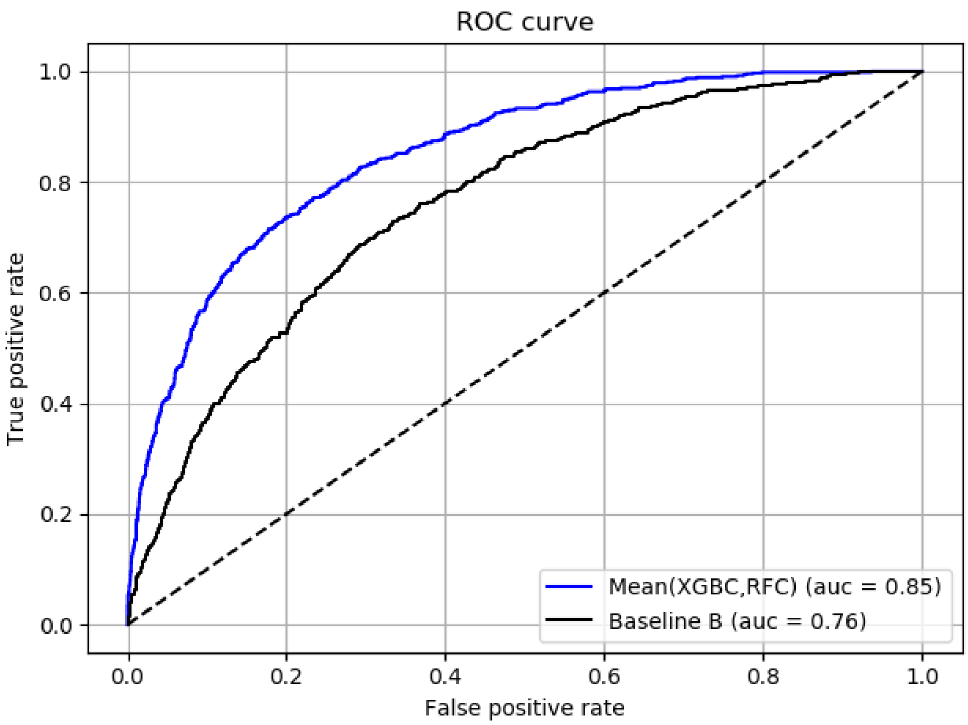

Figure 2: Best single-stage model compared to Baseline B

SHAP and RFE analysis concluded that several of the added risk drivers were significant.

Rank	XGBC (Cure)	RFR (LGD)
1	*Rank Of Security*	Country Of Jurisdiction
2	Country Of Jurisdiction	*Facility Type*
3	Country Of Residence	Industry
4	*Collateral Type*	*Nature Of Default*
5	Mean Guarantee Percentage	Country Of Residence
6	*Nature Of Default*	EAD 1/Initial Loan Amount
7	*Public Private Indicator*	*Collateral Type*
8	Mean Entity Sales log	Default Loan/Limit 2
9	*Total Rate*	Mean Entity Assets log
10	Mean Entity Assets log	NOM DEFAULT AMOUNT 1

Table 9. Top risk-drivers from SHAP and RFE analysis, underscore indicates risk-driver from the extended features.

4.5 Comparing all models

Comparison of all models is presented at the table 10 and the figure 6. Table 10 shows differences in AUC and MAE among the models. Figure 6 presents the absolute error for the models.

	AUC	MAE
Baseline A	0.71	0.260
Baseline B	0.72	0.266

XGBC + RFR	0.82	0.224
Baseline B extended risk drivers	0.76	0.259
XGBC + RFR extended risk drivers	0.85	0.216

Table 10: AUC and MAE for all tested models

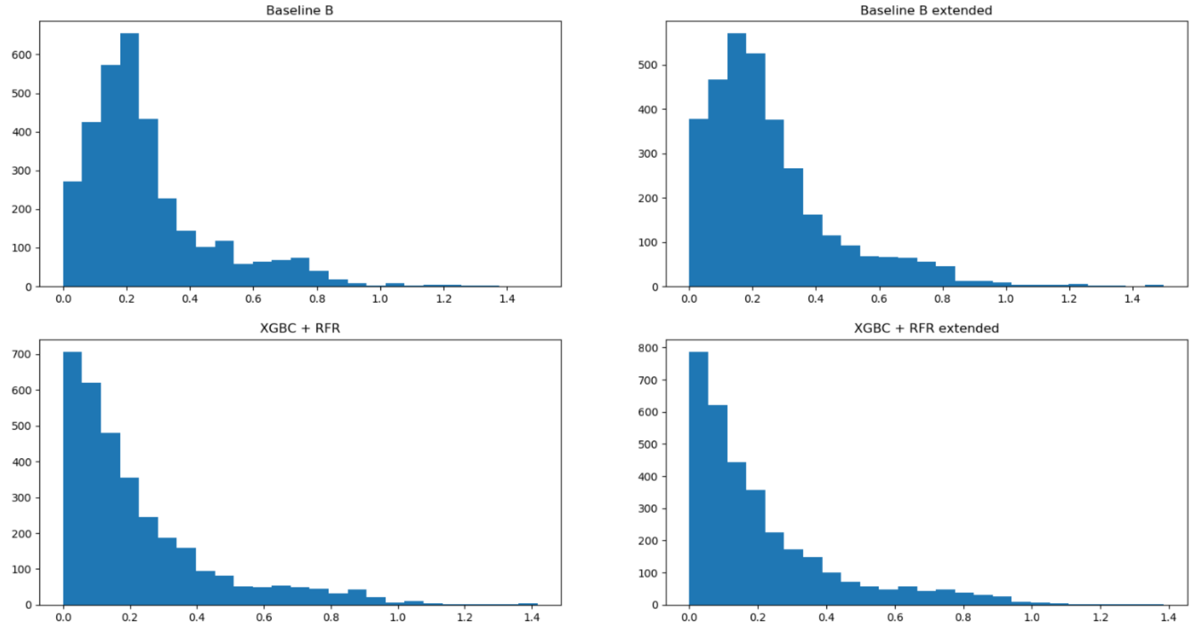

Figure 6: histogram over absolute error for selected models

4.6 Altering the cure definition

The cure definition in section 3.2.2. might at first glance seem a bit arbitrary. To test if the predictive power of the baseline models could be improved further, the cure was redefined as a function of time to resolution or LGD.

Cure Definition	MAE Baseline A	MAE Baseline B
Original Definition	**0.260**	**0.267**
TTR < 30d	0.259	0.267
TTR < 100d	0.262	0.265
TTR < 200d	0.263	0.267
TTR < 2y	0.282	0.281
TTR < 3y	0.286	0.280
TTR < 5y	0.296	0.366
LGD < 1%	0.264	0.297
LGD < 2%	0.266	0.302
LGD < 3%	0.266	0.268
LGD < 5%	0.268	0.269
LGD < 10%	0.272	0.275
LGD < 20%	0.275	0.306
LGD < 30%	0.277	0.278
LGD < 50%	0.285	0.291
LGD < 75%	0.296	0.320

Table 11 Cure definition results

The conclusion is that even though the MAE metric could be improved slightly, it is not significant. Since the original cure definition is standard practice, it would not make much sense to alter the cure definition based on the increased predictive power of LGD.

5 Conclusion

The possibility to recreate the RDS, signalizes an unchanged data quality since the LGD report issue (Rainone & Brumma, 2019). Further data quality checks allowed to understand the data and pick suitable risk-drivers to build models showing acceptable level of predictive power.

LGD is considered to be difficult to predict due to its bimodal distribution. The potential future research can be focused on increasing the number of steps in the ML models, including predicting the PD before the actual default occurs, as well as separate prediction of 0 ≤ LGD < 0,2 (this is the range the most observations of LGD fall in) and LGD ≥ 0,2 where the second-highest peak is located. The cure definition used by GCD proved to be similarly predictable compared to alternatives with shorter or longer time to resolution and different levels of actual LGD. Another way to look at the cure could be to present it as a function of time to recovery, which is another opportunity for the future research.

Historical averages model

The historical averages model performs in line with the expectations and is the best in identifying the customers whose potential LGD is likely to be high (see Figure 7). Its overall power of prediction corresponds to the models used in the banks, proving that the GCD data contains useful information that can be successfully incorporated into the models. The limitations of the model arise from the limited number of observations available for some groups, which makes the model prone to overfitting.

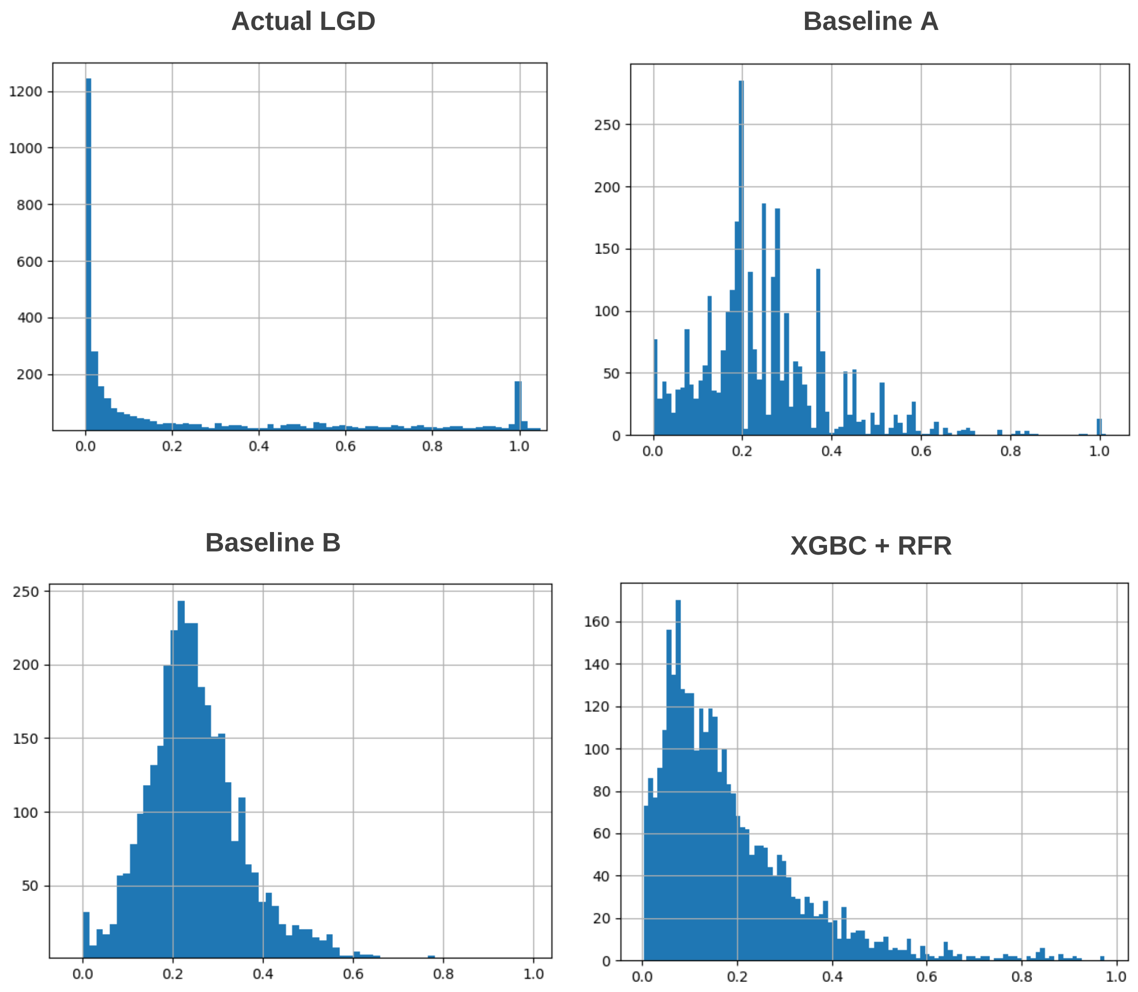

Figure 7. Actual vs Predicted LGD (x-axis) in number of observation (y-axis)

Regression model

Regression model performs marginally better than its historical averages counterpart when predicting probability of cure. However, the regression's overall MAE is worse than the one of historical averages model. The conclusion is therefore that the regression model performs in line with historical averages.

Machine Learning model

ML shows some progress compared to traditional techniques as its area under the curve score increases by 0.10 to 0.82 and MAE decreases by 0.04 to 0.22 compared to the regression model. These improvements signal a better model performance. The ML model is also the best model in picking the low-risk customers, whose potential LGD is close to zero (see Figure 7).

Dataset extension

The conclusion from extending the dataset was positive. Several of the added risk drivers contributed to the increase in model prediction ability.

Of the extended variables the discount rate had highest predictive power but was excluded from the analysis. The discount rate is defined as the 3M Euribor rate at the date of default. While using macro variables in the analysis makes sense, the discount rate for the period of data collection (2000-2015) basically divides the data into 3 different categories, high/normal rates up to 2008, low rates between 2009 and 2013, and negative rates after 2013. This raised the question if the discount rate was increasing the predictive power or if it was just a way to divide the dataset into different subsets and that different characteristics of the subsets increased the predictive power. The latter case would lead to no increase in predictive power for a loan defaulting today. The authors suggest that this variable could be used as an input to the downturn flag rather than a standalone independent variable. The loan spread and base rate looked initially like very promising features that should have a lot of predictive power but unfortunately, there are too many missing values in the dataset and these drivers scored low in the SHAP/RFR analysis.

It is the authors' firm belief that given good data quality and a sound choice of model features, the increased predictive power from Machine Learning models goes some way to offsetting the increased model risk it entails (although metrics for comparing model risk vs predictive power have not been prepared). Furthermore, the models produced using GCD's pooled data show strong predictive power and typical industry drivers, indicating that when combined with an actual bank's portfolio, the data could aid robust modelling.

6 Appendix

6.1 Reference to Python projects and tools

Following Python packages were used for the model:

1. Keras
2. logging
3. matplotlib
4. numpy
5. os
6. pandas
7. pprint
8. random
9. re
10. scipy
11. shap
12. sklearn
13. statsmodels.api
14. string
15. TensorFlow
16. traceback
17. warnings
18. xgboost

6.2 Features in regression analysis and machine learning model (not extended).

Significant coefficients[1]	CAP LGD 1	CURE (WOE)
Country of jurisdiction	Mixed*	0.05
Country of residence	Mixed*	0.08
Default Loan/Limit 1	0.07	-
Default Loan/Limit 2	-	0.14
Default LTV	-	0.09
Default Share Other	-0.08	-
Default Share Real Estate	-0.06	0.09
Downturn Flag	0.03	-
EAD 1 log	-0.31	-
EAD 2/Initial Loan Amount	-	0.11
EAD 2 log	0.30	0.37
Borrower Risk Rating	Mixed*	0.21
Initial Loan/Limit	-	-0.07
Initial Loan Amount log	-0.03	0.10
Initial LTV	-	-
Initial Share Other	-	0.10
Initial Share Real Estate	-	-
Mean Entity Assets log	-0.04	0.26
Sales log	0.05	0.13
Mean Guarantee Percentage	-	0.16
Industry	Mixed*	0.09

* These categorical variables have been one-hot encoded (made to dummy variables) with some categories (country/industry code/risk rating) being significant and others not.

[1] Coefficients with a significance level > 0.05 are excluded and marked with "-"

Model features	Type	Model features	Type
Country of jurisdiction	Dummy	Default Share Real Estate	Numeric
Country of residence	Dummy	Downturn Flag	Dummy
Default Lender Borrower Risk Rating	Dummy	EAD 1 log	Numeric
Default Loan/Limit 1	Numeric	EAD 2 log	Numeric
Default Loan/Limit 2	Numeric	Borrower Risk Rating	Dummy
Default LTV	Numeric	Initial Loan Amount log	Numeric
Default LTV 1 Flag	Dummy	Initial LTV	Numeric
Default Share Other	Numeric	Initial LTV 1 Flag	Dummy
Mean Entity Assets log	Numeric	Initial Share Other	Numeric
Sales log	Numeric	Initial Share Real Estate	Numeric
Mean Guarantee Percentage	Numeric	Industry	Dummy

7 References

Brumma N., Rainone N., (2019) LGD Report 2019 - Large Corporate Borrowers, *Global Credit Data,* https://www.globalcreditdata.org/library/lgd-report-large-corporates-2019

Konečný, Tomáš; Seidler, Jakub; Belyaeva, Aelita; Belyaev, Konstantin (2017): The time dimension of the links between loss given default and the macroeconomy*, ECB Working Paper*, No. 2037, ISBN 978-92-899-2759-8, European Central Bank (ECB), Frankfurt a. M., http://dx.doi.org/10.2866/52109

Lohmann C., Ohliger T. (26/03/2019) Factors that drive the cure of a defaulted company *presented at* GCD European Conference, Vienna, Austria

Martinsson F., (2017) Exotic approaches for modeling Loss Given Default, *Stockholm University*

Ruey-Ching Hwang & Chih-Kang Chu (2018) A logistic regression point of view toward loss given default distribution estimation, Quantitative Finance, 18:3, 419-435, DOI: 10.1080/14697688.2017.1310393

Zhang, J., Thomas, L. C., (2012), Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD, *International Journal of Forecasting*, vol 28, pp. 204–215.

Lundberg S. M., Lee Su-In. A Unified Approach to Interpreting Model Predictions. 2017. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf