Applications of non-linear machine learning tree-based methods for prepayments forecasting of fixed-rate institutional loans

R. M. Glawion[a], J. Heuel[a,*], A. Horovitz[a], A. Szimayer[a]

[a] *University of Hamburg, Faculty of Economics and Social Sciences, Department of Social Sciences, Von-Melle-Park 5, 20146 Hamburg, Germany*

Abstract

This paper aims to enhance the econometric models mainly used by the financial services firms to predict prepayments of fixed-rate institutional loans. Upon deploying several model types for prepayment prediction on Euro-currency, fixed-rate institutional loans between 2012 and 2020, we found that tree-based machine learning methods significantly outperform logistic regressions in predictive powers. We recovered the expected inverse relationships between changes in interest rates and subsequent prepayments. We also identified other driving features that correlate with subsequent prepayments, like loan costs and changes in business conditions. We ascertained the directional covariance of significant features of non-linear inference models by means of a Shapley plot. Further, we also draw inferences on the prepayment volumes in Euro amounts and timing of prepayments.

Keywords:

Machine Learning, Institutional Loans, Prepayment, Liquidity Risk Management, Banking Regulation

*Corresponding author

Email addresses: rene.glawion@uni-hamburg.de (R. M. Glawion), johannes.heuel@uni-hamburg.de (J. Heuel), andre.peter.horovitz@uni-hamburg.de (A. Horovitz), alexander.szimayer@uni-hamburg.de (A. Szimayer)

1. Introduction

While the academic literature is rich in research publications on residential mortgages and consumer loans prepayment models, relatively little has been published on institutional loan prepayments. Some of the few publications that address institutional prepayment behaviors can be attributed to Cossin & Lu (2004) and McGuire (2008). During the periods of positive and relatively stable interest rates, the focus on residential and consumer loans was aligned with the economic priorities of most lenders. Indeed, early amortizations of institutional loans were hardly conducive to adverse effects on lenders because unexpected inflows of liquidity from institutional borrowers could be mostly deployed at attractive returns. The last decade marked by very low/negative interest rates in the OECD has challenged this premise. Unexpected early prepayments of institutional loans had the effect of converting the economic value of such products from "value creators" to "value destroyers". The negative economic profitability resulted from the unexpected inflows of liquidity that had to be deployed at unattractive rates while the tenors of the loans with attractive locked in margins became unexpectedly shorter.

Our research is motivated by the emerging interest in modeling institutional loans prepayment behaviors also echoed by the recent regulatory developments under the Basel III/IV IRRBB (Interest Rate Risk in Banking Books) recommendations published by the Basel Committee for Banking Supervision in 2016. In this paper we strive to explore the prepayment behavior of institutional obligors while concentrating on fixed-rate loans. We aim to test the results obtained by previous researchers with respect to the prepayment sensitivities to changes in the term structure of rates and other driving features on institutional loan data containing Euro-currency loans between 2012 and 2020. We commence our analysis by using traditional econometric models such as logistic regressions as employed by many previous researchers[1]. More importantly, we attempt to test

[1] See for example, Jacobs et al. (2005), Elul et al. (2010), and Agarwal et al. (2011).

the tenet of insufficient predictive powers of such models on our dataset, which is the main reason for the very modest set of behavioral models for institutional obligors in practical applications.

Above and beyond, we strive to deploy non-linear tree-based ML inference methods in an attempt to improve the predictive abilities of statistical models with respect to identifying early prepayments of fixed-rate institutional loans.

We decided to separate our research endeavors on institutional prepayments between fixed and variable-rate loans as the sensitivity to the changes in interest rates is naturally different between the two cohorts. This paper focuses strictly on the analysis of fixed-rate institutional loans. [2].

Motivated by the increased level of data availability associated with institutional loans (associated with the growing level of loan securitization), we also strived to deploy tree-based Machine Learning (ML) inference methods in an effort to ascertain relationships that could be conducive to modeling the prepayment behavior of institutional borrowers. In an analogy to the behavior of mortgagors, who exhibit lagged prepayment behaviors to changes in interest rates, we expected similar lagged prepayment behaviors of institutional obligors. Our aim is to improve the predictive accuracy of institutional fixed-rate prepayment models in an effort to enable the building of effective early warning mechanisms to predict prepayments subsequent to changes in market rates. Lastly, due to the cumbersome explanatory abilities, especially concerning directional sensitivities of analyzed features of traditional ML inference methods, we augment this deficiency by means of a dedicated robustness analysis employing Shapley plots in an attempt to validate the derived relationships. We did not employ neural networks for the classification task, since as far as we know, they do not have a feature to analyze these relationships.

While we recovered the negative directional relationships between changes in interest rates and prepayment probabilities, we also determined a direct re-

[2] A behavioral analysis of variable-rate institutional loans prepayments is discussed in Horovitz (2021)

lationship between the cost of the loan to the customers and the prepayment probabilities. Obligors belonging to the manufacturing industrial sectors are more likely to prepay than borrowers belonging to other economic sectors, with the least propensity of prepayment attributed to obligors belonging to the agricultural sector. Perhaps more importantly, the analyses supported by non-linear tree based ML methods yield far superior predictive powers than the traditional linear logistic regression methods.

From a practical implementation perspective, we envisage our analyses as conducive to the potential development of early warning systems capable of identifying the most likely institutional fixed-rate loans to be prepaid upon negative changes in interest rates. Hence, banks can inform their customer relationship management to get in touch with the borrower who will potentially prepay the loan.

We deem our results to contribute to an enhanced ability to perform liquidity risk management for commercial banks endeavoring similar analyses and consequently, optimize economic capital in the financial system as mandated by the Basel III/IV IRRBB recommendations, which were coined into financial regulation in most OECD jurisdictions.

The paper is organized as follows: Section 2 presents an overview of the relevant literature and exhibits the gaps we aim to fill via this research effort; Section 3 introduces the dataset and the features and presents the data transformations; Section 4 explains the methodology applied; Section 5 presents the key findings; Section 6 discusses the multiple robustness analyses deployed to validate the models and Section 7 concludes the paper.

2. Literature review

We concentrate our research on the exploration of behavioral features impacting prepayments of fixed-rate institutional loans as we conject that the driving factors of early institutional loan prepayments may differ from the well researched factors driving mortgage prepayments. In addition, we notice that

the academic literature is currently heavily imbalanced between publications on mortgage vs. institutional loan prepayments, in favor of the former. This result can be attributed to two main factors: for one, portfolios of institutional loans tend to exhibit lumpier distributions (in terms of size and prepayment behavior) in comparison with portfolios of mortgage and consumer loans, making the inference from the distribution of institutional loan prepayments less reliable and more difficult to analyze; secondly, for most of the period between the 1970s (when the behavioral modeling of prepayments emerged as an essential branch of quantitative asset liability management) and the early 2000s, institutional loan prepayments were by and large tiny in volume terms as compared to the loan sizes and the impact of the liquidity overflow resulting from them were rather modest. Indeed, the overflow of unexpected liquidity resulting from institutional loan prepayments was typically deployed towards lucrative investments at attractive returns. Even during periods of economic contraction, institutions found means to deploy the unexpected liquidity injections at sufficient high return. The recent decade marked by negative central bank deposit rates in OECD economies and especially in the Euro-currency economic space has changed this situation, whereby the unexpected liquidity injected from early prepayments had to be deployed at unattractive rates (sometimes negative rates) with adverse effects on lenders' profit and loss statements. We will succinctly present an evolutionary taxonomy of the existing literature to date (with no pretense of exhausting the vast volume of publications on the topic) while pointing to the gaps we identified which we attempt to address in this paper.

One of the first models for rational mortgage prepayments is attributed to Richard & Roll (1989). Given the apparent optionality features inherent in the prepayment profile, several authors attempted to value the prepayment option as a function of the interest rate levels and the property values. Hilliard et al. (1998) as well as Chen et al. (2009) conducted an approach based on a bivariate partial differential equation with reference rates following a Cox Ingersoll Ross equilibrium model and property values following a geometric Brownian motion model. Other authors reverted to traditional statistical models (especially ap-

plied to consumer loan prepayments – for the reasons articulated above). For example, Schwartz & Torous (1989) as well as Schwartz & Torous (1992) applied a log-logistic hazard rate model on aggregated Government National Mortgage Association (GNMA) pools and argued that the prepayment experience is consistent with the log-logistic representation. Their prepayment estimates were subsequently integrated in a valuation model for mortgage-backed securities. A simplified application of their model is the conditional prepayment rate/single monthly mortality model developed by the Public Securities Association in the US. The model was successfully used for decades to ascertain the prepayment behaviors of GNMA and Federal National Mortgage Association securitized tranches and their respective bonds. Giliberto & Thibodeau (1989) also used a log-logistic hazard rate model similar to Schwartz and Torous. Their contribution relates to the analysis of individual data as they added idiosyncratic obligor variables to their model. These authors provided a theoretical framework for analyzing borrowers' decisions. They show that an obligor's wealth depends on the economic gain from exercising the prepayment option. In particular, they found that an increase in interest rate volatility slows down prepayments because the obligor's economic benefit depends on the value of continuing to hold the option. However, their work falls short of representing adequately the behavior of obligors who prepay but do not move. Other authors attacked the prepayment behavior (again, given the approximately Gaussian distribution of prepayments – focusing on residential mortgages) from a valuation perspective. Peters et al. (1984) examined the prepayment experience of a US nationwide sample of conventional fixed-rate residential mortgages. Some 500 thousand mortgages were classified in 921 cohorts and used in an OLS linear regression. They regressed the conditional prepayment rate on several variables, of which the main ones included refinancing costs, the differential between contract and market rates, obligor's age and property size. The study found that refinancing costs had the dominant impact on prepayments.

Contemporaneously, Green & Shoven (1986) applied a proportional hazard rate model to mortgage prepayments. Their model represented an improvement

to the statistical estimation techniques exhibited above. The main value-added represented the enhanced explanatory power of their stochastic variable used, which they termed "lockin" – defined as the difference between the face and the market value of the mortgage expressed as a fraction of the initial principal amount. To account for the property appreciation over time (the data contained mortgages during a period of economic expansion in the US where residential dwelling values increased almost at a constant rate), they adjusted the initial principal amounts by a regional property value appreciation index. While Green & Shoven (1986) obtained very sturdy estimates of their "lockin" variable, their model was still limited (especially for practical implementation purposes) as they strictly focused on the evolution of reference rates without including additional explanatory variables. As a result, prepayments attributable to regional mobility, preference to deleveraging, divorce, or change in family size were not accounted for. A year later, Quigley (1987) expanded the Green & Shoven (1986) model by including household mobility factors. Quigley showed that mobility (and henceforth, related prepayment) is positively correlated with household size and education of the obligors. The study was criticized due to the instability of the coefficient's sign relating to the borrower's income. However, a critical contribution of Quigley's work is connected with his analysis of the significance of the assumption relating to the proportional hazard rate.

While all contributions mentioned above concentrated on prepayments of fixed-rate loans (also the focus of this research, albeit concentrating on institutional borrowers), it must be mentioned that a few authors attacked the problematic issues of finding explanatory models for variable-rate loans prepayments. Some prominent contributions (also addressing residential mortgages, with adjustable rates) are of Cunningham & Capone Jr (1990), McConnell & Singh (1991), Sanyal (1994) and Daniel (2008). While different in their respective approaches, these authors found that full and partial prepayments are strongly affected by the amount by which the mortgagor's highest interest rate level attained over a prior period exceeds the current variable-rate applied to the loan.

It was only by the turn of the millennium when researchers attacked the more convoluted issue of prepayments of institutional borrowers. Cossin & Lu (2004) published an article focusing strictly on corporate loans (but not loans to financial institutions) while applying a binomial tree framework to derive rational prepayment behaviors. McGuire (2008) has published a comprehensive analysis of commercial borrowers' prepayment behavior in the US. He found the main drivers of commercial prepayments to be "refi ability" (obviously, strictly applicable to fixed-rate loans), seasonality and "improved business conditions". It is precisely the sparse contributions to institutional fixed-rate prepayments that inspired us into exploring more advanced inference means of modeling. We noticed that most authors who attacked the subject, reverted to traditional econometric models, mainly logistic regressions (Jacobs et al., 2005),(Elul et al., 2010),(Agarwal et al., 2011). We further explored the use of non-linear tree-based ML models in an attempt to improve the predictive powers of such models to a level that may be conducive to developing early warning systems.

As such, we also scanned the contributions of ML applications to this subject. While the literature is not very rich in applications of ML methods for prepayment forecasting, we were able to identify a few benchmark studies that inspired our research ambitions. Quinlan (1986) developed the iterative dichotomiser 3 (ID3) algorithm, used to generate decision trees processing large datasets including many attributes. The trees are useful for classification and regression tasks. Further, new methods like Random Forests were developed. Random Forests are a learning method consisting of multiple decision trees relying on the bagging (bagging stands for bootstrap aggregation) concept. The method was found to be suitable for classification tasks by Breiman (2001). Instead of treating each tree in a forest independently, as it is the case for bagging, more recently developed tree-based methods boost a series of trees by updating the version of the previous one. One example for a tree-based gradient boosting algorithm is the LightGBM which we use in our application, developed by Ke et al. (2017). Previously, Liang & Lin (2014) used Random Forests to segment mortgagors into different groups before a proportional hazard model is

employed for predicting the prepayment timing. The results indicated that the two-stage process, which includes the Random Forests, predicts mortgage prepayments more accurately than the process without the segmentation. Another study by Guth & Sapsis (2019) compared the prediction performance of traditional statistical techniques with ML and deep learning methods in the context of bankruptcy and default events. Traditional methods were outperformed by approximately 10% in classification accuracy. Sirignano et al. (2016) unveiled a highly non-linear relationship between macroeconomic data and the behavior of 120 million mortgages across the US ranging from 1995 to 2014, the prepayment forecast being improved.

As briefly described above, the literature on prepayment behavior of institutional obligors is at this stage rather skimpy with no contributions we could identify on ML applications to prepayment modeling of fixed-rate institutional loans. In this paper we aim to explore this subject and discuss our analyses results.

3. Dataset and Features

We conducted our research based on a dataset obtained from the European Data Warehouse (EDW)[3]. The database we use collects information on loans that entered the securitization pool between Q1 2012 and Q3 2020. In addition we collected a set of macroeconomic variables. Our macroeconomic data stems from the Federal Reserve Bank of St. Louis Economic Dataset (FRED) which includes the European Consumer Price Index (Core CPI) - seasonally adjusted and the Euro area Gross Domestic Product (GDP) from Q1 2011 through Q3 2020. For the same time period we also collected interest rate data from Reuters Refinitiv. Furthermore, we obtained zero rate curves from the European Central

[3]The EDW commercializes pan-European loan data in an effort to spearhead loan securitization in the Euro-currency zone. The project was initiated by the European Central Bank (also the designer of the data model) in 2011 and was transferred for commercialization purposes to the EDW in 2012.

Bank which we used for modelling robustness tests.

3.1. Data Processing

Since we aim to analyze mainly small and medium-sized enterprise (SME) loans, we excluded loans belonging to retail customers or financial institutions such as special purpose vehicles except investment and pension funds. We found that the loan prepayment behavior for borrowers who belong to commercial retailing segments as well as the repair of motor vehicle industry is heavily idiosyncratic[4]. After further data cleaning steps due to data quality issues[5], we were left with a pool of 275,078 fixed-rate institutional, mainly SME loans. These loans represent € 52.160 bn of notional assets. Within this filtered loan pool, we observe 2,017 partial prepayments summing up to an amount of € 88 million and 1,101 prepayments at the last payment summing up to € 288 million. We define "partial prepayments" the payments during the lifetime of loans exceeding those scheduled by the loan amortization indentures and lower than the remaining outstanding loan amounts. We termed "full prepayments" the prepayments of the full outstanding loan amount at a minimum of two weeks ahead of the contracted maturities.

To compute relative GDP growth rates, we lag up to four quarters to account for expected delays in prepayment decisions,

$$\Delta^{\mathrm{rel},s}\mathrm{GDP}_t = \frac{\mathrm{GDP}_{t-s+1} - \mathrm{GDP}_{t-s}}{\mathrm{GDP}_{t-s}},$$

for $s \in \{1, 2, 3, 4\}$ quarters. Further, we calculated lagged relative CPI changes similarly for $1, 3, 6, 9$, and 12 months and in order to ascertain the overall struc-

[4]Companies belonging to this industry segment primarily finance the purchasing of motor vehicles with up to 2-years loans (in most cases fixed-rate) and prepay the loans when they sell the vehicles - as such, the prepayment decision is conditioned upon exogenous business factors. As our analysis tries to depict fixed-rate loans prepayment behaviors that are not linked to idiosyncratic business considerations, we opted to eliminate this sector from our analysis pool.

[5]See `https://www.ecb.europa.eu/paym/coll/loanlevel/faq/html/index.en.html` for additional known data quality issues.

ture of the yield curve, we included the 3-months Euribor rates as well as the 5-years EUR interest rate swap rates (5-years IRS) plus their 1-month absolute differences for the respective periods

$$\Delta^{\mathrm{abs},s} i_t = i_{t-s+1} - i_{t-s},$$

lagged for $s \in \{1, 3, 6, 9, 12\}$ months, where i_t is the respective variable in period t. In these cases we use absolute differences, since for some periods the interest rates change their signs. We purposely limited the universe of reference interest rates to one reference short term rate and one long term rate in order to avoid overfitting and minimize multicollinearity effects, in full knowledge that changes across the term structure ladder tend to be heavily intercorrelated.

3.2. Descriptive statistics

Table 1 introduces the non-metric scaled features used for our analysis. We use six different groups of features, all extracted from the EDW database. The borrowers Basel III segmentation classifies the obligors businesses form by number of employees, turnover or size of the balance sheet (among others)[6]. Additionally, the "Nomenclature statistique des activités économiques dans la Communauté européenne" (NACE) code indicates the business area in which the respective debtor operates. For our analysis, we use the letter granularity of the NACE classifications to determine the broader industry segments. Features defining the loan terms are the interest rate type and the principal payment frequency, showing different fixed-rate loan types and the timing of the principal repayment. Moreover, we extracted the quarters out of the timestamp to control for time and we segmented the original loan size (OLA) of each credit in deciles.

Table 2 summarizes the relevant metric scaled variables in the dataset we used. Listed are the central tendency parameters: the mean and the median of these variables for the samples that feature a prepayment and samples that

[6] https://www.ecb.europa.eu/paym/coll/loanlevel/transmission/html/index.en.html

Table 1: **Non-metric variables description.**
The table introduces the prepayment distribution among six non-metric features used for our analysis on the test set. All features are taken from the EDW database. Furthermore, the data shows the subcategories of each variable. Additional information on how prepayments are distributed for the original loan size and NACE codes, are reported in Table A.2 and Table A.1.

Field Name	Source	Field Definition	Prepayment	No Prepayment
Borrowers Basel III Segment	EDW	Corporate	399	42,049
		SME treated as Corporate	523	69,552
		Other	12	511,608
		No Data	7	680
Principal Payment Frequency	EDW	Monthly	668	262,316
		Quarterly	106	13,333
		Semi annually	24	4,743
		Annual	23	318,358
		Bullet	116	24,318
		Other	5	821
		No Data	0	0
Interest Rate Type	EDW	Fixed-rate loan (for life)	835	614,960
		Fixed with future periodic resets	60	6,420
		Fixed-rate loan with compulsory switch to floating	35	1,044
		Capped	9	1,593
		Switch Optionality	2	142
Quarters	EDW	Q1	324	154,813
		Q2	240	189,020
		Q3	261	167,960
		Q4	116	112,096
Original Loan Size	EDW	Dummy variable splitting the OLA into ten deciles.		
NACE Code	EDW	See United Nations (2008)		

do not feature a prepayment. For each variable in the dataset, we test whether the mean or median differs between the two samples. We note that the two sub-samples differ significantly for most of the variables suggesting that there may be subtle patterns we can identify in our subsequent analyses. For example, for prepayments we observe a higher mean and median of the contracted interest rate compared to non-prepaid loans. In addition we observe smaller absolute differences for the majority of the used market rate data. This is a first indication that loans that pay high-interest rates (mainly during periods of decreasing market rates) tend to exhibit a higher likelihood of prepayment. In

Table 2: **Central tendency tests for metric scaled features on the test set.**
We performed the Kruskal-Wallis test to investigate if the prepayment and non-prepayment samples come from populations with the same median. Further, we used a t-test for the means of two independent samples of scores. ***, **, * indicate the significance of the differences between the prepayment and non-prepayment samples at the 1%, 5% and 10% level. These statistics apply to the test sample, which includes 941 prepayments and 623,889 non-prepayments as defined in Section 4.3.

		Median			Mean		
	Source	Prepayment	No Prepayment	Difference	Prepayment	No Prepayment	Difference
Current Interest Rate	Refinitiv	3.1250	1.1500	−1.9750***	3.2661	1.4736	−1.7925***
Quarterly GDP (lag 1)	FRED	.0049	.0053	.0004	−.0015	−.0101	−.0086***
Quarterly GDP (lag 2)	FRED	.0062	.0063	.0001***	.0029	−.0059	−.0088***
Quarterly GDP (lag 3)	FRED	.0049	.0063	.0014***	.0056	.0045	−.0011***
Quarterly GDP (lag 4)	FRED	.0049	.0065	.0016***	.0058	.0071	.0013***
Monthly CPI (lag 1)	FRED	−.1000	−.1000	.0000***	−.0536	−.0853	−.0317***
Monthly CPI (lag 3)	FRED	.0000	−.1000	−.1000***	−.0067	−.0600	−.0533***
Monthly CPI (lag 6)	FRED	.0000	−.1000	−.1000***	−.0150	−.0513	−.0362***
Monthly CPI (lag 9)	FRED	.0000	.0000	.0000***	.0470	−.0087	−.0557***
Monthly CPI (lag 12)	FRED	.0000	.0000	.0000***	−.0023	−.0346	−.0322***
3M Euribor (lag 1)	Refinitiv	−.0020	−.0070	−.0050**	−.0102	−.0079	.0023***
3M Euribor (lag 3)	Refinitiv	−.0010	−.0010	.0000***	−.0010	−.0047	−.0036***
3M Euribor (lag 6)	Refinitiv	−.0010	.0000	.0010***	−.0144	−.0025	.0119***
3M Euribor (lag 9)	Refinitiv	.0000	.0000	.0000***	−.0107	−.0065	.0042***
3M Euribor (lag 12)	Refinitiv	−.0020	.0000	.0020***	−.0502	−.0098	.0405***
5y Eurirs (lag 1)	Refinitiv	−.0630	−.0470	.0160***	−.0712	−.0222	.0491***
5y Eurirs (lag 3)	Refinitiv	−.0665	−.0630	.0035***	−.0356	−.0227	.0129***
5y Eurirs (lag 6)	Refinitiv	−.0065	−.0460	−.0395***	−.0099	−.0201	−.0103***
5y Eurirs (lag 9)	Refinitiv	.0600	−.0390	−.0990***	.0435	−.0174	−.0610***
5y Eurirs (lag 12)	Refinitiv	.0043	−.0520	−.0563***	−.0114	−.0366	−.0252***

Section 5 we will confirm this conjecture by showing that those variables exhibit high features importance in our models.

Additionally, we use cross tables (A.1 and Table A.2) to investigate the distribution of the prepayment samples and the non-prepayment samples on the test set. Each table has three indexes. The first index indicates whether the samples are prepayments or non-prepayments, the second index shows the decile of the OLA and the third index exhibits the borrowers Basel III segment the respective obligor belongs to. The columns display the NACE codes of each loan sample. The row and column margins show the respective sums as well as their relative distribution. Observing the column margins in Table A.1 we note that more than 50% of loans belong to NACE code A which represents agriculture,

forestry and the fishing sector. Further, NACE codes C, F, G and N have a share between 5% and 10% of the non-prepaid samples[7]. The share of the remaining NACE codes is smaller than 3.8%. The rows margins show that firms other than corporates and SMEs treated as corporates hold between 6% and 10% of the loans in each of the first eight OLA deciles. The distribution changes in the last two deciles representing the higher OLAs. We observe more loans taken by SMEs treated as corporate (between 4.8% and 6%) while loans take by SMEs classified as other drop almost below 3%. The loan allocation changes for prepaid loans as indicated by Table A.2. Around 28% of the prepayments occur for NACE code C which represents the manufacturing industry. Other industries for which we see a high amount of prepayments are G (Wholesale and retail trade) with more than 19% as well as F (Construction) and M (professional, scientific and technical activities) with approximately 8%. Moreover, the relative row margins show that corporates and SMEs treated as corporate exhibit a high frequency of prepayments for all OLA deciles. Especially in the deciles representing high OLAs the two Basel III segments are particular dominant. We conject that loans taken by larger firms are more frequently prepaid. We hypothesize that these firms are more often financially managed by professionals who react more sensitively to recent market condition developments relevant for their businesses.

4. Methodology

We commence the inference analysis by using a logistic regression where the prepayment is defined as a binary dependent variable whereas the features are various macro and microeconomic indicators along with obligor and industry specific data. To ensure that the coefficients represent an economically admis-

[7]This will be of importance in Section 5, where we will comment that in spite of exhibiting a large number of observed prepayments, this sector does not seem significant as the volume of prepayments is relatively modest in absolute terms and also the density of prepayments is lower than in other NACE cohorts.

sible inference explanation, we perform a coefficient t-test analysis at a 95% confidence level. As such, if a feature coefficient keeps the same sign over the entire interval, we conclude that it is likely to be representative in direction (coefficient sign) for the inferenced relationship. We perform the analyses on both the training and the test samples.

We augment our analyses by the use of non-linear tree-based ML methods in an effort to enhance the discriminatory power and accuracy measures of the logistic regression model deployed. To analyze the direction of effects in the tree-based models we use SHAP (SHapley Additive exPlanation) values. Lastly, we analyze the robustness of the parameters of all models deployed.

4.1. Models

Logistic Regression. We endeavor an attempt to model the prepayment behavior of fixed-rate institutional borrowers by means of a logistic regression. The logistic regression builds a classifier in two steps: fit a conditional probability model for $P(Y = \text{prepayment}|X = x)$, and subsequently classify as one if $\hat{P}(Y = \text{prepayment}|X = x) \geq 0.5$, and zero otherwise (Efron & Hastie, 2016, p. 109ff).

Decision Trees. Decision trees are non-linear models that break the input space into regions and have separate parameters for each region. In the classification framework, we use the Gini impurity to split the regions. Further, we tune the hyperparameters for the maximum depth of a tree during our estimation procedure, the minimum sample size for a split, the minimum samples in a leaf of a tree for a split, and the number of maximum features, where a leaf is a node in a tree with degree 1.

Random Forests. Random Forests grow many "deep" regression trees to randomized versions of the training data. Here, deep refers to the number of layers (depth) within the specific trees. Compared to decision trees the main idea is variance reduction by averaging over the trees. Each tree fits a piecewise constant surface $\hat{r}(x)$ over the domain by recursive partitioning (Efron & Hastie,

2016, p. 325). The random forest then takes the average

$$\hat{r}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{r}_b(x),$$

for any prediction point x and the number of trees B. Since the Random Forest is composed of multiple decision trees, we tune the same hyperparameters as for the decision trees.

LightGBM. Boosting methods and Random Forests have a lot in common. They both represent the fitted model by a sum of regression trees. However, there are some stark differences. In contrast to regression trees, boosting methods grow "shallow" trees (built on the residuals) while developing additive models as "sums of trees". The basic fitting mechanism is based on bias reduction compared to variance reduction in the case of Random Forests (Efron & Hastie, 2016, p. 324ff). The main idea is to fit a model generated by the exponential family of response functions of the form

$$\eta(x) = \sum_{b=1}^{B} g_b(x;\theta_b),$$

where η is the natural parameter of the conditional distribution $Y|X \sim x$, and the $g_b(x;\theta_b)$ are simple functions of the shallow trees.

Developed by Microsoft Research, Light Gradient Boosting Machines (LightGBM) is a gradient boosting framework (Ke et al., 2017). LightGBM was especially designed for higher efficiency and scalability of boosting to large datasets. One of the main differences compared to Random Forests is that LightGBM grows the tree leaf-wise instead of level-wise. It will choose the leaf with the maximum loss to grow. Holding the number of leaves fixed, leaf-wise algorithms tend to achieve higher accuracy as compared to level-wise algorithms (Shi, 2007). Here, we tune the learning rate, the maximum depth of a tree, the maximum number of leaves, as well as the number of boosted trees to fit.

4.2. Model Evaluation Criteria

The model we use for our analysis is

$$E[y_{t+1}|\mathcal{F}_t] \approx g(X_t,\theta), \tag{1}$$

where $y_{t+1} \in \{0, 1\}$ denotes whether we observe a prepayment in period $t + 1$, $\theta \in R^{p+1}$ is a vector of weights (hyperparameters) we want to optimize in the sense of some given metric, $X_t \in \mathcal{F}_t$ is the matrix of features and $\mathcal{F}_t$ is the filtration. The function $g(\cdot)$ depends on the method applied. In our analyses we use the ML models Decision Trees, Random Forests and LightGBM. For our benchmark logistic regression we only use $p = 46$ features, since we only use one lagged value $(t-1)$ to avoid multicollinearity. Institutional prepayments are rare events in the underlying dataset occurring with a probability of approximately 1 in 100. Standard econometric procedures tend to underestimate the probability of such events in favor of the majority class, which in our case is "no prepayment" (King & Zeng, 2001). He & Ma (2013) review the algorithms and applications of imbalanced learning and conclude that traditional performance measures (e.g., accuracy) do not serve as good indicators of discriminatory powers. Guth & Sapsis (2019) suggest to use the F_1 score for model evaluation since it depends on normalized quantities which take into account the highly imbalanced dataset of prepayments and Wang et al. (2015) use the F_1 score for an imbalanced credit scoring model[8]. We abide by the cited literature of imbalanced learning and maximize the F_1 score when training the different models

$$F_1 = \frac{\text{True positives}}{\text{True positives} + \frac{1}{2}\left(\text{False positives} + \text{False negatives}\right)}.$$

4.3. Splitting the Dataset

To ensure appropriate data representation, we partitioned our loan pool into a training data sample encompassing 70% of the loans and the remaining 30% into a test (validation) sample by stratified sampling. The training and the test samples are disjoint. As such, we ensured that the proportion of the original loan amounts and the number of the prepayments (associated with the last

[8] An alternative to the F_1 score proposed by Velez et al. (2007) is the balanced accuracy. However, we deem balanced accuracy as suboptimal in the underlying setting since we are more concerned about detecting positive instances which are achieved by the F_1 score rather than detecting negative instances which are achieved by the balanced accuracy.

scheduled payment) remain comparable in both datasets. To avoid overfitting in the training process, we employ a five-fold cross-validation procedure.

5. Key Findings and Analysis Results

We commence the analysis with the logistic regression, obtaining similar results to other research publications with respect to the main elasticities, like McGuire (2008), albeit parametrized for the Euro-currency zone. The key drivers to prepayments are the borrowers Basel III segment, month-to-month lagged changes in GDP, month-to-month lagged changes in interest rates (short and long term), and the cost of the loan to the borrowers. While the directional impacts of the features are unsurprising (as evidenced by the coefficients signs which are stable at a 95% confidence level test), the predictive power of the base model is weak. The analysis via tree-based non-linear ML methods exhibits superior discriminatory powers. Of the three methods endeavored (Decision Trees, Random Forests and LightGBM), the best results obtained on the validation sample are exhibited by the Random Forest algorithm.

The directional feature sensitivities of the tree methods were obtained via a SHAP plot on the test sample confirming the coefficient signs of the main features of the base logistic regression. In addition, the SHAP plots reveal some less obvious correlations between features and the target variable such as the industry classification. It appears that obligors classified as industrial and manufacturing firms exhibit a significantly higher propensity to prepay as compared to other industry classes with agricultural firms exhibiting a lower propensity to prepay. Loans of higher notional amounts, mostly belonging to obligors classified as industrial and manufacturing companies, exhibit a higher likelihood of being prepaid.

5.1. Logistic Regression Results

We obtained similar results to the previous researchers (Richard & Roll, 1989; Jacobs et al., 2005; Cossin & Lu, 2004; McGuire, 2008) with respect to the

main feature elasticities to prepayment probabilities, albeit while parameterizing our model on Euro-currency institutional loans. The main features driving prepayments are the short-term and long-term interest rates. [9]. As shown in Table 3, declining interest rates trigger higher probabilities of subsequent prepayments as evidenced by the negative coefficients of -1.016 and -5.074 for the short term and the long-term rates in our full model (1). The coefficients maintain their negative signs both at a 95% and at a 99% confidence level over various model specifications.

The more expensive the loan is to the customer, the higher the propensity of subsequent prepayments. This is evidenced by the positive coefficient of the current interest rate feature taking the value of 0.196. The positive sign remains stable, also at the 99% confidence level as indicated in column 1 of Table 3.

Macroeconomic changes as expressed by the month-to-month changes in GDP and CPI seem to exhibit at first sight somewhat counterintuitive elasticities to subsequent prepayments – as evidenced by the positive signs of their respective coefficients. It appears that improvements in economic conditions (higher GDP and CPI) are eliciting higher propensities of subsequent prepayments. This seems to conflict with professional managers´ traditional economic expectations that strong economic quarters align with expectations of increases in interest rates. Nevertheless, we note that our observation period was between 2012 and 2020, a period marked by very low rates in the Euro-currency economic space with very modest levels of rate changes[10]. Lastly, for the logistic regressions, we find no evidence that corporates and SME companies treated as corporates tend to exhibit a higher propensity to prepay loans as compared to institutions of other categories. As we will further show in Section 5.2, these features will exhibit significant correlations to prepayments under non-linear ML inference analyses.

[9] we used the 1 month lagged absolute changes in the 3-months Euribor rate and the 1-month lagged 5-years IRS rate for the short and long rates, respectively

[10] See Cochrane (2018) for a macroeconomic discussion of this period.

Table 3: **Logistic Regression Results.**

This table reports the regression output for the analysis of partial and full prepayments for different model specifications. Estimation of standard errors is heteroscedasticity consistent according to White (1980). ***, **, * denote significance of the estimated parameter at the 1%, 5%, and 10% level, respectively.

	Total Prepayment					
	(1)	(2)	(3)	(4)	(5)	(6)
Corporate	.1118		.4196	-.0581	.2047	.0461
	(.200)		(.299)	(.163)	(.246)	(.182)
SME treated as Corporate	-.2773		-.0271	-.5111***	-.2687	-.4432**
	(.194)		(.295)	(.156)	(.241)	(.175)
Other	-5.5413***		-4.9116***	-4.9605***	-4.8666***	-5.8588***
	(.362)		(.363)	(.284)	(.333)	(.312)
Current Interest Rate	.1959***		.1597***	0.2442***	.1911***	.2090***
	(.012)		(.012)	(.012)	(.012)	(.011)
$\Delta^{\text{rel},1}$ GDP	1.1240*	5.5239***	.3008	.9210	.2308	1.2135*
	(.618)	(.365)	(.587)	(.629)	(.618)	(.668)
$\Delta^{\text{abs},1}$ CPI	0.4227***	.7338***	.4959***	.3275***	.3933***	.3866***
	(.073)	(.062)	(.074)	(.073)	(.073)	(.069)
$\Delta^{\text{abs},1}$ Euribor 3m	-1.0161**	2.4590***	-.4655	-1.1492***	-.5061	-1.2613***
	(.454)	(.351)	(.465)	(.456)	(.471)	(.454)
$\Delta^{\text{abs},1}$ EURIRS 5y	-5.0742***	-5.800***	-3.6542***	-4.6016***	-3.7524***	-5.1526***
	(.280)	(.315)	(.231)	(.265)	(.228)	(.266)
Controls for Payment Frequency	Y	N	Y	Y	Y	Y
Controls for NACE Code	Y	N	N	Y	N	Y
Controls for Loan Size	Y	N	Y	N	N	Y
Controls for Timing	Y	N	Y	Y	Y	N
Observations	1,462,612	1,462,612	1,462,612	1,462,612	1,462,612	1,462,612
Pseudo R^2	.2772	.0218	0.2572	.2652	.2652	.2717

While the logistic regression classifies almost all samples as non-prepayments it results in a low number of false positives but also in only 15 correctly classified prepayments, shown in Table 4. In consequence, Table 5 shows a high accuracy and precision but poor recall and F_1 scores.

5.2. Results of the Tree-Based Methods

We further augment our analysis by three non-linear tree-based methods, namely a Decision Tree, a Random Forest and the gradient boosting method called LightGBM, which are all described in Section 4.1. In the following para-

graphs we exhibit the results on the test set by the use of Table 4, and Table 5. The hyperparameter range as well as the optimal parameter resulting from a grid search algorithm are reported in Appendix B. We also report the training set results, to provide evidence that we do not overfit the data. Indicated by comparable error metrics on both sets, we conclude that our models do not suffer from overfitting. Although, Table 5 shows the performance metrics that can be derived from confusion matrices given in Table 4, we report both for better interpretability. The exhibited numbers in both tables result from the model optimization with respect to the F_1 score.

Table 5 exhibits different discriminatory metrics concerning the observational errors the models produce on the training and on the test sets. Observing good results in only one measure is not sufficient which is why they must be contrasted against each other. The standard approach of measuring the performance of binary classifiers is accuracy. It is calculated by the ratio of correct predictions to total sample size. All models achieve high accuracy values of around 99% since most observations are correctly classified as "non-prepayment". Since the underlying data is highly imbalanced this can hardly be considered as an appropriate measure. Larger differences occur for the precision and recall measures.

The precision parameter quantifies the correctly predicted prepayments out of all prepayments that are classified as prepayments including the false positive forecasts. This measure aims to minimize false positive predictions. The Random Forest analysis results in a value of 85.98% on the test set. It significantly outperforms the two other ML methods. The LightGBM method exhibits the poorest results for this measure and achieves a value of 29.90%.

The recall measure quantifies the number of the correctly predicted prepayments out of all possible prepayments in the dataset. This metric aims to minimize false negative predictions. The recall parameters values range from 18.17% to 27.42% with the decision tree method exhibiting the lowest value and the LightGBM method exhibiting the highest value among all methods deployed.

Table 4: **Confusion matrices for partial and full prepayments.**
True positives denote loans that do not feature a prepayment while the models correctly predict no prepayments. False positives are loans that are not getting prepaid, while our models falsely predict a prepayment. False negatives are loans that feature a prepayment, while our models flag no prepayment. True negatives are the loans where a model correctly identifies a prepayment. Each of these instances is given for the training as well as the test set per applied method. All reported numbers result from a model optimization with respect to the F_1 score.

	True Negatives	False Positives	False Negatives	True Positives
PANEL A: Logistic Regression				
Training Sample	1,460,413	22	2,136	41
Test Sample	623,873	16	926	15
PANEL B: Decision Trees				
Training Sample	1,460,286	149	1,750	427
Test Sample	623,819	70	770	171
PANEL C: Random Forest				
Training Sample	1,460,376	59	1,510	667
Test Sample	623,852	37	714	227
PANEL D: LightGBM				
Training Sample	1,459,557	878	1,426	751
Test Sample	623,284	605	683	258

The two measures exhibit significant differences among the ML models in our setting. While the maximum difference for the recall measures between the methods is 9%, the maximum difference for precision measures is 55%. Since predicting true negatives is deemed as less important in our analysis for ascertaining the models performance, we employ the F_1 score as an error metric, see Section 4.2.

The F_1 score combines the recall and precision measures by allocating more weight to minimize false positives and false negatives predictions. As shown in Table 5, the application of the Random Forest method results in the highest F_1

Table 5: **Performance metrics for partial and full prepayments.**
The table summarizes different performance measures for classifying prepayments for partial and full prepayments. We trained the model according to Section 4.2. All reported numbers result from a model optimization with respect to the F_1 score.

	Accuracy	Precision	Recall	F1	Log Loss
PANEL A: Logistic Regression					
Training Sample	.9985	.6508	.0188	.0366	.0510
Test Sample	.9985	.4389	.0159	.0309	.0521
PANEL B: Decision Tree					
Training Sample	.9987	.7413	.1961	.3102	.0448
Test Sample	.9987	.7095	.1817	.2893	.0464
PANEL C: Random Forest					
Training Sample	.9989	.9187	.3064	.4595	.0371
Test Sample	.9988	.8598	.2412	.3768	.0415
PANEL D: LightGBM					
Training Sample	.9984	.4610	.3450	.3946	.0544
Test Sample	.9979	.2990	.2742	.2860	.0712

score of 37.68% on the test set. This is approximately 9% higher than the F_1 scores for the LightGBM method and the Decision Tree. The table also reports the log loss measure which ranges between 4.64 % and 7.12% using the F_1 score optimization.

Table 4 shows that each of the three methods is capable of predicting more than 170 prepayments correctly, whereby the LightGBM method ranks as best in true positives predictions with 258 correctly forecasted prepayments, compared to 227 for the Random Forest. We observe that the Decision Tree and the Random Forest methods generate a significant number of misclassifications by not classifying actual prepayments as true prepayments. The Decision Tree method results in 770 false negatives, the Random Forest in 714 false negatives, while the LightGBM exhibit in 683 false negative predictions. Whereby the false

positive classifications remain also over 600 for LightGBM, we observe that only 70 non-prepayments get classified as prepayments for the Decision Tree and 37 for the Random Forest, reinforcing the meaning illustrated by the differences in the F_1 scores expressed by the confusion matrix.

In view of the underlying results, we conclude that the tree-based methods yield better prepayment forecasts than the logistic regression model, as evidenced by the superior discriminant power statistics, applied to both the training and the test samples.

Each of the three methods used is capable of predicting more prepayments than the logistic regression. While each practical application endeavored on lender idiosyncratic samples will likely yield different results, we observe that the Random Forest and the LightGBM methods seem to be superior to other methods in terms of their predictive powers. Compared to the logistic regression, the Random Forest predicts correctly fifteen times more prepayments while raising slightly fewer false positive signals. While the number of false negative signals amount to 926 for the logistic regression (also higher for the LightGBM model), it correctly predicts only 15 prepayments out of 941 in the test set. The resulting F_1 score of 5.21% is considerably lower than the F_1 scores of the tree-based methods, which stands as the reason for our qualification of the logistic regression as a poor performing model in the introduction as observable from Table 4 and Table 5.

We can thus validate the widely applied industry practice of neglecting prepayments of fixed-rate institutional loans or, in selected cases (a practice mostly applied by wholesale banks), applying a slim haircut to contractual maturities for transfer pricing and economic attribution analysis purposes when using traditional statistical models - as we know from anecdotal evidence from industry experience that applications use logistic regressions in their installed models.

The observation that the Random Forest method generates fewer false positives while also generating a considerable number of true positives compared to other analyses methods would be an encouraging signal for a lender building an early warning system, as fewer false alarm predictions seem to be occurring. We

deem the Random Forest configuration to be the overall superior model within the underlying setting (see Table 4). The LightGBM configuration produces more than sixteen times of false positive signals while yielding only comparatively few additional true positive signals. The high imbalance between the considerably more false positives and the slightly more true positives results in an approximately 9% lower F_1 score compared to the Random Forest method (see Table 5).

5.3. SHAP Plot

To achieve a more insightful economic interpretation and to increase transparency of the results generated by the ML models, we employ SHAP values (Lundberg et al., 2018). The SHAP plot represents a bee swarm plot of all SHAP values grouped by feature. The features are described on the ordinate axis. The features plotted higher in the diagram exhibit stronger contributions to the prediction outcome. The abscissa axis exhibits the impact on the model output of the respective SHAP values. The stronger the dot color to red, the higher the SHAP value corresponding to the respective feature and the stronger the color to blue, the lower the SHAP value.

In general, the SHAP Plot also reinforces an important set of features along with their directional elasticities that were obtained via the logistic regression analysis discussed in 5.1. The SHAP plot in Figure 1 reports the result of the Random Forest method. It reveals that beyond the expected inverse relationships between prepayment probabilities and changes in reference rates, we discover some other important relationships between target variables and the probability of early loan prepayments. Like we found in the logistic regression, high amongst the examined features is the Basel III segment. Referring to Table A.2 and to Section 4 we validate our conclusion that corporates and SMEs treated as corporates exhibit the highest propensity of early prepayments. We attempt to explain this result by hypothesizing that many industrial obligors are managed by professional financial managers who are more likely to seize on the opportunities presented by advantageous changes in the economic environment

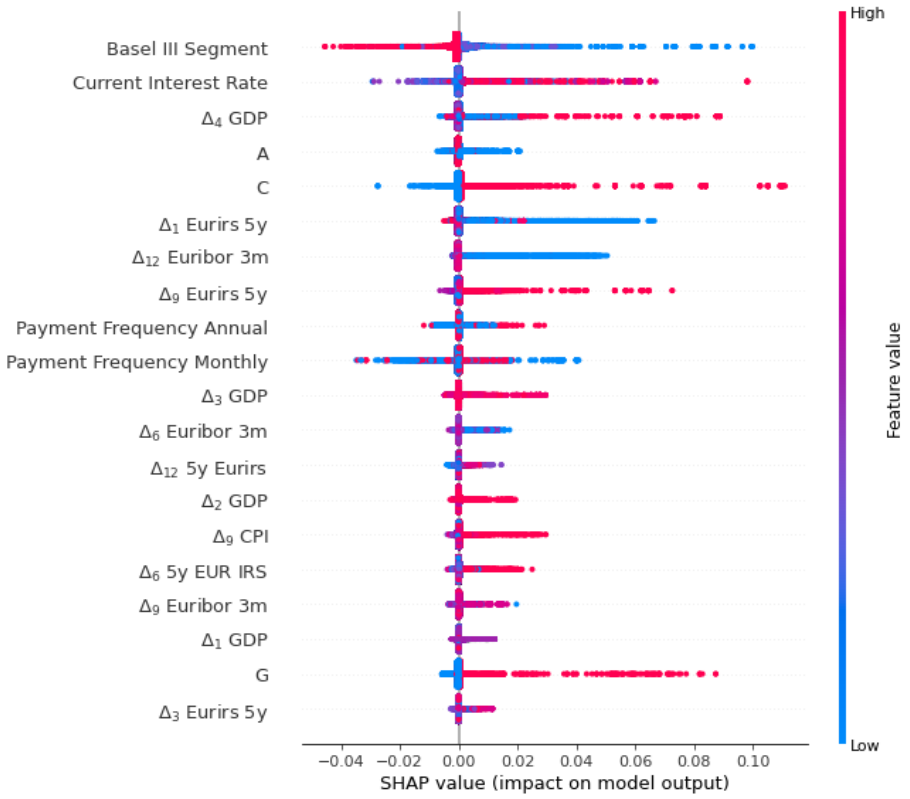

Figure 1: **SHAP value variable importance plot for Random Forest.**
The SHAP variable importance plot orders the features according to their significance from top to bottom, where the top feature is the most significant one. Additionally, the color code indicates the influence of high variable values (red) or low values (blue) having a positive relationship (right from zero) or a negative relationship (left side from zero) on the target variable.

(lower interest rates and increases in GDP growth/inflation rate).

The current interest rate is also a very important predictor. Loans with high-interest rates exhibit a large probability of a prepayment or, expressed differently, more expensive loans tend to exhibit a higher propensity of being prepaid than cheaper loans. Furthermore, we observe that the absolute changes of the lagged 5-years Euro swap interest rate as well as the absolute changes

of the lagged 3-months Euribor represent important features for prepayment prediction of institutional fixed-rate loans. Negative relative changes in long-term rates, proxied by the 5-years swap rate correlate with higher levels of subsequent prepayments. The same pattern is observable for short-term rates, proxied by the 1 month lagged 3-months Euribor rates.

Moreover, higher prepayment propensities are consequently associated with previous observed changes in macroeconomic indicators. We observe that higher levels of GDP growth rates and inflation rates represented by the European CPI index correlate with subsequent higher probabilities of prepayments. We imply that positive economic trends tend to elicit decisions to deleverage, a relationship also found in the logistic regression model analysis and aligned with the findings of McGuire (2008) albeit in the US markets.

Lastly, we notice a differentiation in prepayment patterns across various industry segments. As such, companies belonging to the manufacturing industry (NACE code C) tend to exhibit higher levels of prepayment behavior than the others. On the other side, companies belonging to agriculture, fishing or forestry industry (NACE code A) tend to exhibit very low probabilities of prepayments. This observation is also supported by A.2[11].

5.4. Materiality and Timing of Prepayments

In Panel A to Panel C of Figure 2, we show the distribution of the prepayment materiality defined by the Euro amount and in Panel D to Panel F of Figure 2 we report the distribution of the timing of the prepayments by use of histograms. From a materiality perspective, prepayments follow a bimodal distribution (as evidenced by analyzing the prepayment distribution of the test set). Partial prepayments tend to represent up to 5% of the outstanding loan amounts, while full prepayments represent over 90% of the outstanding loan

[11] While agricultural loans are higher in absolute numbers of loans, their overall size per loan is typically lower and the prepayments volumes as a proportion of the overall volume are far less significant

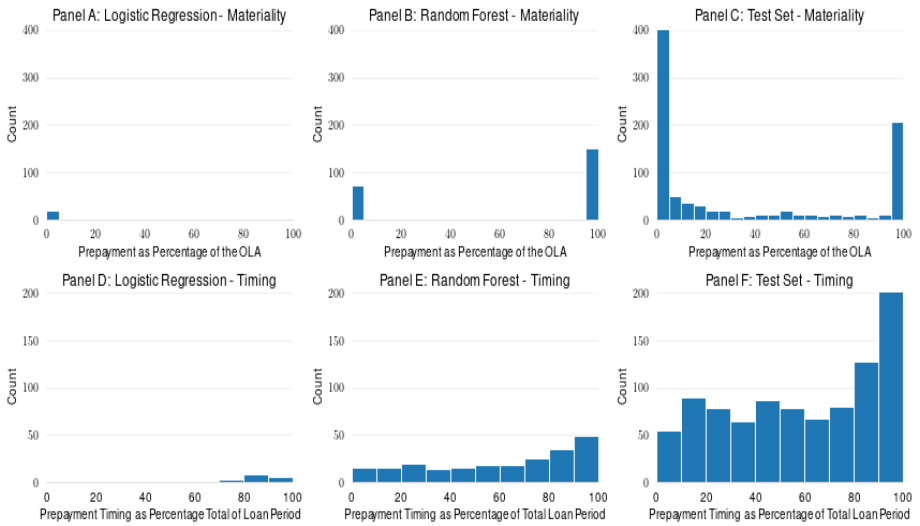

Figure 2: **Size and timing of prepayments.**
The figure shows three histograms concerning the prepayment amount as a percentage of the original loan amount in the first row. While panel C presents all available prepayments in the test set, panel A and panel B demonstrate the materiality captured by the logistic regression and the Random Forest method, respectively. The second row shows three histograms concerning the prepayment timing as a percentage of the total loan tenor. While panel F captures all available prepayments in the test set, panel D and panel E indicate the timing of the prepayments predicted by the logistic regression and the Random Forest method, respectively.

amounts, see Panel C. The Random Forest model is the best capable in predicting the tails of the prepayment distribution. The model stands out by exhibiting very good prediction performances for full prepayments in the right tail of the distribution. Approximately three-fourth of all full prepayments are correctly forecasted (Panel B and Panel C). This finding can be of importance from an economic point of view, since large prepayments lead to banks balance sheet liquidity risks. Panel A shows that the logistic regression method results in an inferior prediction performance as it correctly predicts very few partial prepayments. From a prepayment timing perspective, prepayments are rather evenly distributed across the maturities time ladder in the test set (Panel F). We find

Table 6: **Summary of predicted Euro amount on the test sample for partial and full prepayments.**

This table reports the prediction performance of each method in terms of Euro volumes. While the first four columns indicate each applied method, the last column reports the numbers for the entire test set. The first two rows, report the total predicted materiality and the average predicted materiality respectively. The average predicted materiality is calculated by the predicted materiality divided by the number of correctly predicted prepayments as shown in Table 4. The last row, shows the ratio of the total predicted materiality (as given in the first row of the table) and the total possible amount (as given in the first row of the last column).

	Logistic Regression	Decision Tree	LightGBM	Random Forest	Full Dataset
(Predicted) Materiality	4,499.08	12,400,148.30	20,777,353.28	12,180,431.22	118,866,232.90
Average (predicted) Materiality	299.94	72,515.49	80,532.38	53,658.29	126,319.06
Ratio of predicted to total Materiality	0.00	10.43	17.48	10.25	-

no evidence of a crowding pattern of prepayments around any particular tenor region.

As an in-depth analysis of the model performance regrading the materiality prediction, we show in Table 6 the Euro amount of all observed prepayments in the test set and the predicted amount of each method. The upper section of the table reports the total amount as well as the average amount. The average amount is calculated by the total prepayment materiality in the test set divided by the 941 prepayments observed in the test set (see Table 4). The lower section shows the prediction performance for each applied prediction method in absolute terms and their average predicted materiality. The average predicted materiality is derived by the ratio of the predicted materiality to the number of true positives of the respective algorithm, given in Table 4. Additionally, the last row indicates the ratio of the predicted materiality to the total amount of prepayments. We observe that LightGBM predicts more than 20 million Euro correctly. This is approximately 18% of the prepaid amount in the test set of more than 118 million Euro. Both, the Random Forest and the Decision Tree predict approximately 12 million Euro out of the 118 million correctly. This is more than 10% of the observed prepayment volume. Moreover, we observe that the LightGBM predicts only 8 million Euro more than the Decision Tree

and 27 million Euro more than the Random Forest on average. Far behind is the performance of the logistic regression which predicts less than 4,500 Euros correctly.

6. Robustness

6.1. Splitting Partial and Full Prepayments

For the analyses presented thus far, we have not distinguished between full and partial prepayments within the loan's lifetime. However, there may be differences in the reasons for and consequences of prepayments. For example, following a full prepayment, a creditor may easier refinance her loan at another bank, which is more difficult in the case of partial prepayments. To analyze the different behaviors, we run additional analyses similar to Section 5, allowing for the separation between the two types. Our results indicate that the ML methods we employed are superior to the logistic regression in predicting full prepayments and partial prepayments. While the ML methods result in an F_1 score above 57.56% for full prepayments and above 16.28% for partial prepayments, the logistic regression achieves only a score of 0% and 14.29%, respectively on the test set (see Table C.2 and Table C.4). This confirms our finding from Section 5, that the logistic regression is hardly capable of predicting full prepayments even when we split the dependent variable. Also, very few partial prepayments are being correctly predicted by the logistic regression. Since the full prepayments account for the larger amounts which may unexpectedly flow into a bank's financial statements, this result leads us to conject that the deployment of tree-based inference methods is more conducive to helping forecast future large prepayment volumes, an imperative piece of information for liquidity risk management purposes. We present the complete results of this analysis differentiating between full and partial prepayments in Appendix C.1.

6.2. Using strictly the Yield Curve Information

We investigated if the superior performance of the non-linear models compared to the logistic regression is not attributable to overfitting. We evaluate

the prediction performance on a disjoint test sample to mitigate for this risk. However, the final ML models use many variables and interactions within each tree. To further mitigate for the risk of overfitting, we reduce the model to the principal components of the yield curve. As discussed in Knez et al. (1994), Duffie & Kan (1996) or Dai & Singleton (2000) those latent factors are often called "level", "slope", and "curvature". We performed all analyses with only the three factors mentioned above and the current interest of the loan. As level we used the one year AAA-rated government bonds. Additionally, we compute the slope and the curvature between the one year and ten years AAA-rate government bond.

The resulting model performs worse than our richer ML model and, perhaps more importantly, positively tests that institutional prepayments are predictable. While the logistic regression is able to predict 15 prepayments correctly on the test set on the full model it does not predict any prepayment correctly on the reduced test set as described above. Additionally, we observe that the ML methods predict only 100 fewer prepayments on the reduced test set when compared to the full test set. We list the results in Appendix C.4.

6.3. Alternative Loss Functions

Classic regression approaches to model prepayments use logistic loss because it most closely relates to minimizing a quadratic cost function (Jacobs et al., 2005). Logistic loss is also the standard approach currently used in the banking industry to model prepayments (Sirignano et al., 2016). Replicating the approach reveals why most institutions currently refrain from modeling institutional prepayments. As shown in Appendix C.3, the ML models trained with the logistic loss exhibit poor prediction performances. Such results confirm the industry practice of expecting either no prepayments as previously mentioned or applying a haircut of a pre-determined fixed percentage value to all outstanding loans. In contrast, our analysis highlights that it is possible to identify institutional prepayments by non-linear inference models.

Further, we also ran all analyses with balanced accuracy. Some authors

like Brodersen et al. (2010) and Zhou & Wang (2012) argue for using balanced accuracy as an alternative loss function for imbalanced datasets. The results for training the ML models with logistic loss do not differ significantly from our main analysis using the F_1 score. Thus, we omit presenting those results in detail.

6.4. Oversampling the Prepayments during Training

In our primary analysis, we tackled the problem of an imbalanced dataset by adjusting the training objective. However, some authors argue that one can oversample the training dataset, e.g. Gosain & Sardana (2017), so that the model can "see" a more significant number of prepayments in the minority group, comprised of prepaid loans. Post oversampling, the resulting model should, in principle, distinguish between the majority and minority groups on a test set following the actual distribution, i.e., without oversampling.

We implemented two standard approaches: first, a naive oversampling method, which duplicates samples from the minority groups; second, the Synthetic Minority Oversampling Technique (SMOTE, Chawla et al. (2002)), which generates new synthetic samples from the minority groups. After oversampling with both procedures, we used the logistic loss objective to train the models.

Observing that the SMOTE technique results in a maximum F_1 score of 22.22% and the naive oversampling technique in a maximum F_1 score of 30.60%, we note that both are significantly lower than the results derived in the main analysis as shown in Appendix C.2.

7. Conclusion

Noticing the imbalance in the academic literature between prepayment models for consumer and institutional loans, we focused on analyzing the prepayment behaviors of institutional obligors in the European Union. Motivated by the long periods of very low interest rates in the Eurozone, we addressed the issue of modeling fixed-rate institutional loan prepayments which have exacerbated the problems of poor European commercial banks profitability during

the last decade. We examined 275,078 fixed-rate institutional loans containing some 3,118 total and partial principal prepayments. The loans were obtained from a database designed by the European Central Bank and commercialized by EDW. We identified approximately one-third of the full or partial prepayments on the test set and examined prepayment behavior against a set of 59 specific and macroeconomic variables. We found that logistic regressions do not appropriately discriminate prepayments in a way that would allow for building sturdy early warning systems. However, tree-based nonlinear methods tend to exhibit superior discriminant power statistics, enabling early warning systems capable of better-predicting prepayments of fixed-rate institutional loans. The main features driving prepayments are the lagged absolute changes in reference rates as drops in reference rates correlate with a higher probability of subsequent prepayments, the Basel III Segment where SME borrowers treated as corporates exhibit the highest propensity to prepay, GDP and CPI growth rates. In addition, we found that obligors belonging to the manufacturing segment exhibit the highest likelihood of prepayments while companies in the agricultural/forestry/fishing sector exhibit the lowest propensity to prepay. While we could find no distinguishable pattern in prepayment timing, we noticed a bimodal distribution in prepayment sizes with the overwhelming prepayment amounts being of full prepaid principal. We encourage commercial lenders to explore the deployment of nonlinear tree-based methods trained on their own specific datasets in an attempt to develop behavioral models for fixed-rate institutional prepayments as recommended by the European Banking Authority and the Basel III forum under "Interest Rate Risk in the Banking Books" and attempt to build early warning mechanisms better capable of forecasting institutional fixed-rate loan prepayments.

Perhaps more important is the observation that non-linear tree-based ML methods are capable of discovering less obvious explanatory features to institutional loan prepayments, which is conducive to conclude that institutional borrowers, especially those belonging to the manufacturing sector, tend to exhibit predictable behavioral patterns of loan prepayments. We encourage fellow

researchers to further explore such behavioral models by deploying more advanced pattern recognition algorithms, like deep learning methods.

References

Agarwal, S., Amromin, G., Ben-David, I., Chomsisengphet, S., & Evanoff, D. D. (2011). The role of securitization in mortgage renegotiation. *Journal of Financial Economics*, *102*, 559–578.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition* (pp. 3121–3124). IEEE.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, Y., Connolly, M., Tang, W., & Su, T. (2009). The value of mortgage prepayment and default options. *Journal of futures and Markets; Wiley*, .

Cochrane, J. H. (2018). Michelson-morley, fisher, and occam: The radical implications of stable quiet inflation at the zero bound. *NBER Macroeconomics Annual*, *32*, 113–226.

Cossin, D., & Lu, M. (2004). Pricing prepayment option in c&i loans at origination. *White paper, University of Lauseanne*, .

Cunningham, D. F., & Capone Jr, C. A. (1990). The relative termination experience of adjustible to fixed rate mortgages. *The Journal of Finance*, *45*, 1687–1703.

Dai, Q., & Singleton, K. J. (2000). Specification analysis of affine term structure models. *The journal of finance*, *55*, 1943–1978.

Daniel, J. (2008). A variable-rate loan-prepayment model for australian mortgages. *Australian Journal of Management*, *33*, 277–305.

Duffie, D., & Kan, R. (1996). A yield-factor model of interest rates. *Mathematical finance*, *6*, 379–406.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference* volume 5. Cambridge University Press.

Elul, R., Souleles, N. S., Chomsisengphet, S., Glennon, D., & Hunt, R. (2010). What" triggers" mortgage default? *American Economic Review - Papers & Proceedings*, *100*, 490–94.

Giliberto, S. M., & Thibodeau, T. G. (1989). Modeling conventional residential mortgage refinancings. *The Journal of Real Estate Finance and Economics*, *2*, 285–299.

Gosain, A., & Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review. In *2017 international conference on advances in computing, communications and informatics (ICACCI)* (pp. 79–85). IEEE.

Green, J., & Shoven, J. (1986). The effects of interest rates on mortgage prepayments. *Journal of Money, Credit and Banking*, *18*, 41–59.

Guth, S., & Sapsis, T. P. (2019). Machine learning predictors of extreme events occurring in complex dynamical systems. *Entropy*, *21*, 925.

He, H., & Ma, Y. (2013). *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

Hilliard, J. E., Kau, J. B., & Slawson Jr, V. C. (1998). Valuing prepayment and default in a fixed-rate mortgage: A bivariate binomial options pricing technique. *Real estate economics*, *26*, 431.

Horovitz, A. (2021). On the modeling of prepayments for variable rate institutional loans-ascertaining the inference of bank internal default probabilities on subsequent prepayments. *Available at SSRN 3893837*, .

Jacobs, J., Koning, R., & Sterken, E. (2005). Modelling prepayment risk. *Dept. Economics, University of Groningen*, .

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: a highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*, 3146–3154.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, *9*, 137–163.

Knez, P. J., Litterman, R., & Scheinkman, J. (1994). Explorations into factors explaining money market returns. *The Journal of Finance*, *49*, 1861–1882.

Liang, T.-H., & Lin, J.-B. (2014). A two-stage segment and prediction model for mortgage prepayment prediction and management. *International Journal of Forecasting*, *30*, 328–343.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, .

McConnell, J. J., & Singh, M. K. (1991). Prepayments and the valuation of adjustable - rate mortgage-backed securities. *The Journal of Fixed Income*, *1*, 21–35.

McGuire, W. J. (2008). Commercial loan prepayment behaviors. *Financial Managers Society*, .

Peters, H. F., Pinkus, S. M., & Askin, D. J. (1984). Figuring the odds: a model of prepayments. *Secondary Mortgage Markets*, *1*, 18–23.

Quigley, J. (1987). Interest rate variations, mortgage prepayments and household mobility. *The Review of Economics and Statistics*, .

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81–106.

Richard, S. F., & Roll, R. (1989). Prepayments on fixed-rate mortgage-backed securities. *Journal of Portfolio management*, *15*, 73.

Sanyal, A. (1994). Ammunition for arms: A panel data approach to prepayment modeling. *The Journal of Fixed Income*, *4*, 96–103.

Schwartz, E. S., & Torous, W. N. (1989). Prepayment and the valuation of mortgage - backed securities. *The Journal of Finance*, *44*, 375–392.

Schwartz, E. S., & Torous, W. N. (1992). An empirical investigation of mortgage prepayment and default decisions. *UCLA finance, Working paper*, .

Shi, H. (2007). *Best-first decision tree learning*. Ph.D. thesis The University of Waikato.

Sirignano, J., Sadhwani, A., & Giesecke, K. (2016). Deep learning for mortgage risk. *arXiv preprint arXiv:1607.02470*, .

United Nations (2008). International standard industrial classification of all economic activities. *Department of Economic and Social Affairs Statistics Division*, *SeriesM No.4*. URL: `https://unstats.un.org/unsd/publication/seriesM/seriesm_4rev4e.pdf`.

Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M., & Moore, J. H. (2007). A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic Epidemiology: the Official Publication of the International Genetic Epidemiology Society*, *31*, 306–315.

Wang, H., Xu, Q., & Zhou, L. (2015). Large unbalanced credit scoring using lasso-logistic regression ensemble. *PloS one*, *10*, e0117844.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, (pp. 817–838).

Zhou, L., & Wang, H. (2012). Loan default prediction on large imbalanced data using random forests. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, *10*, 1519–1525.

Appendix A. Additional Descriptive Statistics

Table A.1: Non-prepayment cross table.

This table reports the absolute distribution of the non-prepayment samples. The table has a two-fold index. The first index shows the deciles the original loan amount of the respective loan belongs to. The second index is the borrowers Basel III segmentation of the sample. The columns of the table display the NACE code segmentation.

		NACE code	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	Z	All	Relative
Prepaid	OLA Decile	Borrowers Basel III Segment																								
No	1	Corporate	1	0	26	0	0	12	14	0	1	0	0	0	3	4	0	0	8	0	0	0	0	20	89	.180
		SME treated as corporate	4	0	21	0	0	15	37	2	6	5	0	8	9	5	1	4	4	2	0	0	0	59	182	.2209
		Other	6322	0	115	4	5	182	234	24	82	14	0	105	121	672	0	26	32	16	51	0	0	0	8005	9.7157
		N/A	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	3	.0036
	2	Corporate	5	1	22	1	0	32	29	3	3	1	0	2	9	9	0	1	7	1	2	0	0	23	151	.1833
		SME treated as corporate	9	0	43	1	1	31	74	8	43	15	0	8	14	31	1	3	11	6	13	0	0	95	407	.4940
		Other	5732	3	143	2	7	295	315	42	110	15	0	110	142	593	0	18	30	25	50	0	0	0	7632	9.2630
		N/A	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	5	.0061
	3	Corporate	2	0	66	0	3	58	66	4	1	2	0	0	10	24	0	1	13	0	2	0	0	42	294	.3568
		SME treated as corporate	5	0	68	0	0	69	112	16	43	12	2	6	22	31	0	1	8	10	21	0	0	120	546	.6627
		Other	5487	0	192	8	3	397	355	68	114	15	0	122	159	514	0	29	34	20	56	0	0	0	7573	9.1914
		N/A	1	0	2	0	0	1	2	2	0	0	0	0	0	1	0	0	0	0	0	0	0	3	12	.0146
	4	Corporate	4	1	61	0	4	82	82	16	8	0	0	3	11	20	0	0	7	1	1	0	0	31	332	.4030
		SME treated as corporate	9	0	83	0	3	75	151	20	64	9	0	19	41	26	0	4	7	8	15	0	0	118	652	.7913
		Other	4713	2	198	8	8	514	417	82	113	19	0	140	193	511	0	15	65	28	41	0	0	0	7067	8.5773
		N/A	0	0	3	0	0	0	3	2	3	0	0	2	1	0	0	0	0	0	0	0	0	1	15	.0182
	5	Corporate	5	0	67	0	2	88	113	43	6	2	0	6	20	31	0	2	11	4	2	0	0	36	438	.5316
		SME treated as corporate	10	0	96	1	3	74	176	45	60	10	0	15	51	44	0	5	20	4	24	0	0	132	770	.9346
		Other	4868	0	222	8	5	462	378	101	107	27	0	134	168	469	0	20	53	23	44	1	0	0	7090	8.6052
		N/A	0	0	2	0	0	3	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	4	15	.0182
	6	Corporate	14	2	116	2	5	68	111	42	9	5	0	7	25	24	0	3	12	1	3	0	0	71	520	.6311
		SME treated as corporate	13	2	134	0	6	105	268	44	90	23	2	22	59	46	0	10	25	12	26	0	0	205	1092	1.3254
		Other	4936	5	227	12	8	353	440	120	143	30	0	187	183	299	1	14	41	33	50	0	0	2	7084	8.5079
		N/A	0	0	4	0	0	2	4	1	2	0	0	0	1	0	0	0	1	0	1	0	0	1	17	.0206
	7	Corporate	16	1	120	0	1	83	128	54	17	10	0	24	28	32	0	2	15	3	3	0	0	81	618	.7501
		SME treated as corporate	43	4	207	8	8	134	329	57	96	41	3	38	82	45	0	6	36	15	19	0	0	221	1392	1.66895
		Other	4373	4	210	8	10	237	441	106	136	22	0	226	162	206	0	8	54	20	55	0	0	1	6279	7.6209
		N/A	0	0	1	0	0	2	10	1	2	0	0	1	2	0	0	0	1	0	1	0	0	1	22	.0267
	8	Corporate	30	4	159	2	7	89	222	112	22	12	0	40	42	44	0	9	14	4	5	0	0	104	921	1.1178
		SME treated as corporate	66	2	316	12	16	124	426	143	82	48	3	55	105	62	0	10	35	16	17	0	0	301	1839	2.2320
		Other	3307	3	227	15	12	165	354	115	100	34	0	266	157	162	1	8	49	13	29	0	0	4	5021	6.0940
		N/A	0	0	8	0	0	2	6	1	4	0	0	1	0	0	0	1	1	1	1	0	0	1	27	.0328
	9	Corporate	56	20	388	8	16	114	345	102	39	35	0	140	67	69	0	24	27	10	5	0	0	173	1638	1.9881
		SME treated as corporate	135	20	891	32	30	243	1033	165	103	88	6	127	223	138	0	11	112	37	38	1	2	573	4008	4.8645
		Other	906	1	254	69	9	105	300	92	72	42	0	381	147	74	4	3	86	14	14	0	0	7	2580	3.1314
		N/A	1	0	3	0	1	3	3	0	2	0	0	0	0	2	0	0	0	0	0	0	0	0	15	.0182
	10	Corporate	38	9	514	9	32	59	284	75	41	38	0	258	130	72	0	6	40	7	5	0	3	119	1739	2.1106
		SME treated as corporate	86	28	1312	59	62	172	1508	181	108	110	5	186	510	151	1	6	66	17	28	0	0	281	4877	5.9193
		Other	22	4	164	121	10	91	163	43	20	12	0	482	111	21	97	1	35	10	8	0	0	3	1418	1.7210
		N/A	0	0	4	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	7	.0085
All			41220	116	6689	390	278	4542	8936	1933	1853	697	21	3123	3011	4432	106	251	960	361	632	2	5	2834	82392	
Relative			50.0291	.1408	8.1185	.4733	.3374	5.5127	10.8457	2.3461	2.2490	.8460	.0255	3.7904	3.6545	5.3792	.1287	.3046	1.1652	.4381	.7671	0.0024	0.0061	3.4397		

Table A.2: Prepayment cross table.

This table reports the absolute distribution of the prepayment samples. The table has a two-fold index. The first index shows the deciles the original loan amount of the respective loan belongs to. The second index is the borrowers Basel III segmentation of the sample. The columns of the table display the NACE code segmentation.

		NACE Code	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	Z	All	Relative
Prepaid	OLA Decile	Borrowers Basel III Segment																						
Yes	1	Corporate	0	0	16	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	18	1.9129
		SME treated as corporate	0	0	2	0	0	2	9	1	2	2	0	1	1	0	0	1	2	2	0	1	26	2.7630
		Other	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0.2125
		N/A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.0000
	2	Corporate	0	0	16	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	19	2.0191
		SME treated as corporate	1	0	2	0	0	1	5	2	1	1	0	3	2	2	1	0	0	0	0	0	21	2.2317
		Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.0000
		N/A	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	.1063
	3	Corporate	0	0	29	0	0	4	6	0	0	1	0	0	1	1	0	0	0	0	1	0	43	4.5696
		SME treated as corporate	5	0	5	0	0	3	8	4	1	0	1	0	3	2	0	2	2	0	1	5	42	4.4633
		Other	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	.1063
		N/A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.0000
	4	Corporate	0	0	5	0	0	0	5	0	0	0	0	1	1	4	0	0	0	0	0	0	16	1,7003
		SME treated as corporate	1	0	4	0	0	10	11	2	5	1	0	5	5	1	0	0	0	0	2	0	47	4.9947
		Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.0000
		N/A	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	.1063
	5	Corporate	0	0	14	0	0	2	3	0	0	0	0	1	0	2	0	0	0	2	0	0	24	2.5505
		SME treated as corporate	1	0	5	0	0	2	9	3	1	0	0	0	7	7	0	0	5	0	0	0	40	4.2508
		Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.0000
		N/A	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	.1063
	6	Corporate	1	0	19	0	0	1	3	1	0	0	0	1	0	1	0	0	0	1	0	0	28	2.9756
		SME treated as corporate	3	0	6	0	0	7	15	5	4	0	1	4	10	1	0	1	2	0	0	0	59	6.2699
		Other	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	.1063
		N/A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.0000
	7	Corporate	0	0	3	0	0	5	1	3	0	1	0	2	1	4	0	0	0	0	0	0	20	2.1254
		SME treated as corporate	1	0	13	1	0	9	8	3	3	1	1	1	10	3	0	0	3	0	0	0	57	6.0574
		Other	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	.1063
		N/A	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	.1063
	8	Corporate	0	0	11	0	0	8	3	4	1	0	0	4	4	0	0	0	0	1	0	1	37	3.9320
		SME treated as corporate	5	0	5	0	0	2	10	5	3	2	0	2	6	4	0	0	2	0	2	0	48	5.1010
		Other	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3	.3188
		N/A	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	.1063
	9	Corporate	1	0	47	0	2	8	26	4	3	2	0	5	5	7	0	0	0	1	1	0	112	11.9022
		SME treated as corporate	13	0	16	0	0	6	23	6	3	4	1	13	5	3	0	2	13	0	3	0	111	11.7960
		Other	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	2	.2125
		N/A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	.1063
	10	Corporate	2	0	31	2	1	3	18	3	6	2	0	2	9	3	0	0	0	0	0	0	82	8.7141
		SME treated as corporate	9	1	13	0	1	3	14	1	2	1	0	9	10	2	0	0	4	0	2	0	72	7.6514
		Other	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	.2125
		N/A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	.2125
All			47	1	264	3	4	80	182	47	38	18	4	57	81	48	1	6	33	8	12	7	941	.1063
Relative			4.9947	.1063	28.0553	.3188	.4251	8.5016	19.3411	4.9947	4.0383	1.9129	.4251	6.0574	8.6079	5.1010	.1063	.6376	3.5069	.8502	1.2752	.7439		

Appendix B. Hyperparameter Tuning

Table B.1: **Hyperparameter Tuning via Grid Search.**
The first three columns of the table show the model parameters as well as the parameter sets applied to each classifier. The last column indicates the optimal parameter chosen by the model via grid search for the main analysis.

Classifier	Parameter	Parameter Range	Optimal Parameter
Logistic Regression			
Random Forest	max tree depth	{4,10,20,40}	40
	min sample split	{2,4,10}	2
	min sample leaf	{2,4,8}	2
	max features	{auto, sqrt, log2}	auto
Decision Tree	max tree depth	{4,10,20,40}	40
	min sample split	{2,4,10}	2
	min sample leaf	{2,4,8}	8
	max features	{auto, sqrt, log2}	auto
LightGBM	max tree depth	{4,10,20,40}	40
	number of boosted trees	{50,100,200}	100
	max tree leaf	{11,31,61}	61
	learning rate	{0.01,0.1,0.3}	0.1

Appendix C. Robustness

Appendix C.1. All Tables for the Split between Full and Partial Prepayments

Table C.1: **Confusion matrices for full prepayments.**
True positives denote loans which do not feature a prepayment and where our models correctly predict no prepayments. False positives are loans which are not getting prepaid, but where our models falsely predict a prepayment. False negatives are loans which feature a prepayment, but where our models flag no prepayment. True negatives are the loans were a model correctly identifies a prepayment.

	True Negatives	False Positives	False Negatives	True Positives
PANEL A: Logistic Regression				
Training Sample	1,461,839	0	773	0
Test Sample	624,501	1	328	0
PANEL B: Decision Trees				
Training Sample	1,461,721	118	407	366
Test Sample	624,457	45	173	155
PANEL C: Random Forest				
Training Sample	1,461,819	20	456	317
Test Sample	624,491	11	191	137
PANEL D: LightGBM				
Training Sample	1,461,744	95	384	389
Test Sample	624,460	42	164	164

Table C.2: **Performance metrics for full prepayments.**

The table summarizes different performance measures for the classification of prepayments. We trained the model according to Section 4.2.

	Accuracy	Precision	Recall	F1	Log Loss
PANEL A: Logistic Regression					
Training Sample	.9995	.0000	.0000	.0000	.0183
Test Sample	.9995	.0000	.0000	.0000	.0182
PANEL B: Decision Trees					
Training Sample	.9996	.75626	.4735	.5823	.0124
Test Sample	.9997	.7750	.4726	.5871	.0121
PANEL C: Random Forest					
Training Sample	.9997	.9407	.4101	.5712	.0112
Test Sample	.9997	.9257	.4177	.5756	.0112
PANEL D: LightGBM					
Training Sample	.9997	.8037	.5032	.6189	.0113
Test Sample	.9997	.7961	.5000	.6142	.0114

Table C.3: **Confusion matrices for partial prepayments.**

True positives denote loans which do not feature a prepayment and where our models correctly predict no prepayments. False positives are loans which are not getting prepaid, but where our models falsely predict a prepayment. False negatives are loans which feature a prepayment, but where our models flag no prepayment. True negatives are the loans were a model correctly identifies a prepayment.

	True Negatives	False Positives	False Negatives	True Positives
PANEL A: Logistic Regression				
Training Sample	1,461,181	27	1,284	120
Test Sample	624,206	11	565	48
PANEL B: Decision Trees				
Training Sample	1,461,115	93	1,187	217
Test Sample	624,164	53	554	59
PANEL C: Random Forest				
Training Sample	1,461,188	20	1,117	287
Test Sample	624,191	26	543	70
PANEL D: LightGBM				
Training Sample	1,461,070	138	1,96	209
Test Sample	624,128	89	540	73

Table C.4: **Performance metrics for partial prepayments.**

The table summarizes different performance measures for the classification of prepayments. We trained the model according to Section 4.2.

	Accuracy	Precision	Recall	F1	Log Loss
PANEL A: Logistic Regression					
Training Sample	.9991	.8163	.0855	.1547	.0310
Test Sample	.9991	.8136	.0783	.1429	.0318
PANEL B: Decision Trees					
Training Sample	.9991	.7000	.1546	.2532	.0302
Test Sample	.9990	.5268	.0962	.1628	.0336
PANEL C: Random Forest					
Training Sample	.9992	.9349	.2044	.3355	.0268
Test Sample	.9991	.7292	.1142	.1975	.0315
PANEL D: LightGBM					
Training Sample	.9991	.6023	.1489	.2387	.0315
Test Sample	.9990	.4506	.1191	.1884	.0348

Appendix C.2. Using Oversampling Techniques for the Minority Class

Table C.5: **Confusion matrices for the SMOTE oversampling.**
True positives denote loans which do not feature a prepayment and where our models correctly predict no prepayments. False positives are loans which are not getting prepaid, but where our models falsely predict a prepayment. False negatives are loans which feature a prepayment, but where our models flag no prepayment. True negatives are the loans were a model correctly identifies a prepayment.

	True Negatives	False Positives	False Negatives	True Positives
PANEL A: Logistic Regression				
Training Sample	1,291,579	168,856	44,045	1,416,390
Test Sample	551,995	71,894	117	824
PANEL B: Decision Trees				
Training Sample	1,444,130	16,305	2,435	1,458,000
Test Sample	616,666	7,223	379	562
PANEL C: Random Forest				
Training Sample	1,452,660	7,775	521	1,459,914
Test Sample	620,296	3,593	422	519
PANEL D: LightGBM				
Training Sample	1,453,906	6,529	846	1,459,589
Test Sample	620,822	3,067	440	501

Table C.6: **Performance metrics for the SMOTE oversampling.**

The table summarizes different performance measures for the classification of prepayments. We trained the model according to Section 4.2.

	Accuracy	Precision	Recall	F1	Log Loss
PANEL A: Logistic Regression					
Training Sample	.9271	.8935	.9698	.9301	2.5176
Test Sample	.8848	.0113	.8757	.0224	3.9806
PANEL B: Decision Tree					
Training Sample	.9936	.9889	.9983	.9936	.2216
Test Sample	.9878	.0722	.5972	.1288	.4202
PANEL C: Random Forest					
Training Sample	.9972	.9947	.9996	.9972	.0981
Test Sample	.9936	.1262	.5515	.2054	.2219
PANEL D: LightGBM					
Training Sample	.9975	.9955	.9994	.9975	.0872
Test Sample	.9944	.1404	.5324	.2222	.1939

Table C.7: **Confusion matrices for the naive oversampling.**

True positives denote loans which do not feature a prepayment and where our models correctly predict no prepayments. False positives are loans which are not getting prepaid, but where our models falsely predict a prepayment. False negatives are loans which feature a prepayment, but where our models flag no prepayment. True negatives are the loans were a model correctly identifies a prepayment.

	True Negatives	False Positives	False Negatives	True Positives
PANEL A: Logistic Regression				
Training Sample	1,241,494	218,941	53,361	1,407,074
Test Sample	530,975	92,914	41	900
PANEL B: Decision Trees				
Training Sample	1,442,040	18,395	0	1,460,435
Test Sample	15,790	8,099	401	540
PANEL C: Random Forest				
Training Sample	1,456,325	4,110	0	1,460,435
Test Sample	21,852	2,037	403	538
PANEL D: LightGBM				
Training Sample	1,455,644	4,791	0	1,460,435
Test Sample	621,236	2,653	400	541

Table C.8: **Performance metrics for the naive oversampling.**

The table summarizes different performance measures for the classification of prepayments. We trained the model according to Section 4.2.

	Accuracy	Precision	Recall	F1	Log Loss
PANEL A: Logistic Regression					
Training Sample	.9068	.8654	.9635	.9118	3.2200
Test Sample	.8512	.0096	.9564	.0190	5.1384
PANEL B: Decision Tree					
Training Sample	.9937	.9876	1.0000	.9937	.2175
Test Sample	.9864	.0625	.5739	.1127	.4699
PANEL C: Random Forest					
Training Sample	.9986	.9972	1.0000	.9986	.0486
Test Sample	.9961	.2089	.5717	.3060	.1349
PANEL D: LightGBM					
Training Sample	.9984	.9967	1.0000	.9984	.0567
Test Sample	.9951	.1694	.5749	.2617	.1688

Appendix C.3. Using Logistic Loss as the Cost Function

Table C.9: **Confusion matrices for the logistic loss.**
True positives denote loans which do not feature a prepayment and where our models correctly predict no prepayments. False positives are loans which are not getting prepaid, but where our models falsely predict a prepayment. False negatives are loans which feature a prepayment, but where our models flag no prepayment. True negatives are the loans were a model correctly identifies a prepayment.

	True Negatives	False Positives	False Negatives	True Positives
PANEL B: Decision Trees				
Training Sample	1,460,435	0	2,177	0
Test Sample	623,889	0	941	0
PANEL C: Random Forest				
Training Sample	1,460,428	7	2,037	140
Test Sample	623,885	4	881	60
PANEL D: LightGBM				
Training Sample	1,460,389	46	1,865	312
Test Sample	623,868	21	817	124

Table C.10: **Performance metrics for the logisitc loss.**

The table summarizes different performance measures for the classification of prepayments. We trained the model according to Section 4.2.

PANEL B: Decision Tree					
Training Sample	.9985	.0000	.0000	.0000	.0514
Test Sample	.9985	.0000	.0000	.0000	.0520
PANEL C: Random Forest					
Training Sample	.9986	.9524	.0643	.1205	.0483
Test Sample	.9986	.9375	.0638	.1194	.0489
PANEL D: LightGBM					
Training Sample	.9987	.8715	.1433	.2462	.0451
Test Sample	.9987	.8552	.1318	.2284	.0463

Appendix C.4. Using the Slope and Curvature of the Yield Curve

Table C.11: **Confusion matrices for using the yield curve information.**
True positives denote loans which do not feature a prepayment and where our models correctly predict no prepayments. False positives are loans which are not getting prepaid, but where our models falsely predict a prepayment. False negatives are loans which feature a prepayment, but where our models flag no prepayment. True negatives are the loans were a model correctly identifies a prepayment.

	True Negatives	False Positives	False Negatives	True Positives
PANEL A: Logistic Regression				
Training Sample	1,460,435	0	2,177	0
Test Sample	623,889	0	941	0
PANEL B: Decision Trees				
Training Sample	1,460,351	84	1,882	295
Test Sample	623,848	41	834	107
PANEL C: Random Forest				
Training Sample	1,460,347	88	1,853	324
Test Sample	623,841	48	835	106
PANEL D: LightGBM				
Training Sample	1,459,990	445	2,005	172
Test Sample	623,704	185	860	81

Table C.12: **Performance metrics for using the yield curve information.**
The table summarizes different performance measures for the classification of prepayments. We trained the model according to Section 4.2.

	Accuracy	Precision	Recall	F1	Log Loss
PANEL A: Logistic Regression					
Training Sample	.9985	.0000	.0000	.0000	.0514
Test Sample	.9985	.0000	.0000	.0000	.0520
PANEL B: Decision Tree					
Training Sample	.9987	.7784	.1355	.2308	.0464
Test Sample	.9986	.7230	.1137	.1965	.0484
PANEL C: Random Forest					
Training Sample	.9987	.7864	.1488	.2503	.0458
Test Sample	.9986	.6883	.1126	.1936	.0488
PANEL D: LightGBM					
Training Sample	.9983	.2788	.0790	.1231	.0579
Test Sample	.9983	.3045	.0861	.1342	.0578