Do loan default risks change during stress periods? P2P lending during Covid-19

P. Raghavendra Rau* Adam Shuaib†

December 2022

Abstract

Using a unique peer-to-peer loan dataset, we compare different machine learning approaches to predicting loan default over 2017-2021, a period that covered the Covid-19 crisis. We find that the out-of-sample default predictability of short-maturity loans is considerably lower than for long-maturity loans, particularly during the the Covid-19 crisis. Higher loan repayment-to-income ratios render short-maturity loans more susceptible to Covid-driven income shocks not captured at loan origination. We show that P2P loan default factors appear stable over time, with total borrowing and account age the most important predictors across both pre-Covid and Covid sample periods. We also document a structural break in the relation between default risk and payment holiday adoption rates for borrowers that are highly uncertain in their ability to repay a loan, consistent with the hypothesis that high degrees of financial uncertainty led to precautionary borrowing and subsequent precautionary payment holiday behaviour during the Covid-19 crisis.

Keywords: P2P Lending, Credit default, Machine learning, Payment holidays, Default factors

JEL codes: G17, G51, G41, G23, C45

*Judge Business School, University of Cambridge, Trumpington Street, Cambridge, CB2 1AG, UK. E-mail: r.rau@jbs.cam.ac.uk

†Judge Business School, University of Cambridge, Trumpington Street, Cambridge, CB2 1AG, UK. E-mail: as2633@cam.ac.uk

1 Introduction

The academic literature has devoted a large amount of attention to the determinants of default risk. The literature typically focuses on the types of information used by lenders such as banks and non-bank financial intermediaries to assess borrowers' default risks in the cross-section. In addition, there is some evidence on the time series variation of default risk for bank lenders. However, little is known empirically about the determinants of time series variation in default risk for non-bank lenders. In this paper, we draw on a unique dataset of 591,400 loans originated by one of the UK's leading peer-to-peer (P2P) loan platforms from 2017 to 2021, and examine the time variation in P2P loan default factors.

Loan default risk arises because of problems of asymmetric information, such as adverse selection and moral hazard, between borrowers and lenders. Extant research on financial intermediation typically examines the role of banks in information production (Diamond (1984)). Several papers focus on the informational advantages that arise from a banking relationship (e.g., Boot (2000)). Using a sample of checking account activity and associated credit line usage of borrowers from a large German universal bank, Norden and Weber (2010) investigate whether the combined information on credit line usage and checking account activity is helpful for bank monitoring of potential borrower default. They find that credit line usage, limit violations, and cash inflows exhibit abnormal patterns approximately 12 months before default events. Using a dataset of loans made by German banks, Puri et al. (2017) similarly evaluate the impact of prior customer–bank relationships on loan default rates.

However, the motivations of financing facilitators such as P2P lending platforms are different from banks. Fintech platforms rose in prominence after the Global Financial Crisis, offering borrowers the opportunity to raise capital by connecting borrowers directly with lenders. They do not take deposits and earn fees through matching borrowers with lenders. There is no deposit insurance on investments. Hence, unlike banks, these platforms typically do not keep the risk from the loans they arrange on their own books. As a service to their lenders, fintech platforms typically rate the borrowers on the probability that they are likely to default and provide this

information to potential lenders. Because they do not have skin in the game themselves, it is an open question whether the default factors for these loans are different from those documented for regulated banks. Since these lenders do not have access to a captive pool of deposits, allowing a higher proportion of defaults (by being more lenient on lending criteria for example) might result in lenders withdrawing from the pool. So a second issue is whether the platforms dynamically adjust their lending criteria over time to maintain stable levels of default. Finally, the Covid-19 pandemic was an exogenous event that significantly changed the likely default probabilities of the borrowers with little time for the platform to adjust its algorithms. Hence, the third issue is whether the default probability of these loans change during a crisis period.

In our paper, we examine if the default risk factors for loans originated by a P2P lending platform are stable over time. Our database is appropriate for answering this research question for at least three reasons. First, for each loan in our dataset, we have access to rich origination data (provided by the borrowers themselves and supplemented with credit rating agency (CRA) data) with monthly loan performance data on the post-origination health of the loan, allowing us to examine if the relative importance of origination and credit rating agency data varies over the sample period. The platform used only quantitative factors that banks also use to judge the default risk of loans listed on the platform. Since it did not fund the loans itself, it did not use alternative data that fintech lenders that fund and keep loans on their books have been documented to have used.[1] Second, the period of analysis encompasses the Covid-19 crisis period, allowing us to examine the stability of loan default factors during a crisis period. Finally, during the Covid-19 crisis, governments allowed borrowers to strategically change their loan maturities during the crisis. Specifically, the government allowed borrowers to take payment holidays during the crisis, though the loans needed to be eventually repaid. Our database allows us to examine the types of borrowers that take payment holidays and relate the propensity to strategically modify loan maturities to the ex ante default risk of the

[1]For example, Jagtiani and Lemieux (2019) compare loans made by a U.S. fintech platform, LendingClub, to similar loans that were originated by banks. They show that the correlations between the rating grades (assigned by LendingClub) and the borrowers' FICO scores declined from about 80% (for loans originated in 2007) to about 35% for recent vintages (originated in 2014–2015) and argue that fintech lenders use more expansive datasets not already accounted for in FICO scores.

borrowers.

We begin by running a horse race over a selection of linear and non-linear models including logistic regressions, k-nearest neighbour methods, naive Bayes classifiers, random forest models, neural networks, and a decision-tree-based ensemble machine learning algorithm called XGBoost, to determine the class of model that delivers the best out-of-sample default prediction performance over time. We show that, for overall model performance, the random forest and XGBoost models outpeform all other models across the majority of the out-of-sample quarters in our dataset. In addition, both tree-based models benefit from fast training times and higher model transparency. In the remainder of the paper, we use the XGBoost model as our main model due to its average outperformance over all out-of-sample periods, though we obtain qualitatively similar results if we use the random forest model.

Using the XGBoost model, we then reduce the dimensionality of our model from the original 45 factors provided in the dataset to 10 factors using a recursive feature elimination (RFE) approach standard in the machine learning literature. The 10-factor model outperforms the 45-factor model across a majority of out-of-sample periods. Importantly, we find that short-maturity loan defaults are harder to predict, particularly during the Covid period. There is also a greater degree of instability in short-maturity model performance. This effect is not explained by differences in the rate of defaults by maturity.

We then examine the time variation in P2P default factors. We find that both the most important and least important default risk factors are relatively stable over time. Total borrowing (*Total Debt*) is the most important feature in three-fourths of the out-of-sample quarters, and ranks as either the second or third most important feature across the remaining quarters. Geographic information (specifically, postcode-level variables such as *Delinquent Accounts* (*Postcode*) and *Healthy Accounts* (*Postcode*)) are consistently ranked as either the least important or second-least-important feature. These relative importance of these factors are unrelated to the Covid crisis - they are stable both in the pre-Covid and Covid periods.

Next, we examine how financial uncertainty and investor behavioural biases might affect

the decision to take a payment holiday during the Covid crisis. A payment holiday is a feature offered by certain loans and mortgages that allows a borrower to miss occasional monthly payments according to a schedule agreed in advance with the lender. However, any missed payments during this period are usually treated as arrears and interest/charges continue to accrue. While payment holidays have existed for many years, they assumed greater importance in March 2020 as one of the economic support measures announced in the midst of the Covid-19 crisis. In non-crisis times, payment holidays are only allowed subject to strict conditions and many credit agreements do not permit such measures - lenders often record a payment holiday on borrower credit reports as an "Arrangement to Pay", which has consequences for the credit score of the borrower in question. However, payment holidays offered as part of the Covid relief measures were introduced with the understanding that credit scores would not be seriously impacted upon utilisation. In spite of this, there were reports in the financial press of lenders recording coronavirus related payment holidays as adverse factors affecting borrower credit scores. [2] Specifically, we examine whether borrowers actively engaged in precautionary payment holiday behaviour during the Covid crisis, or whether payment holiday adoption rates were in line with implied default probabilities.

We hypothesize that if implied default probabilities are very high or very low, there is a low degree of financial uncertainty as the borrower in question is almost certain to either default or not default respectively. However, borrowers with implied default probabilities close to 50% are highly uncertain. They are unsure if they will or will not default - both are equally likely, as there is 50% chance of default and a corresponding 50% chance of no-default. This interpretation rests on the (plausible) assumption that borrowers are aware and conscious of their own degree of financial uncertainty. We show that payment holiday adoption rates increase linearly with implied default probabilities in the range of 10% to 40%. However, there is a sharp structural break in the 50% implied default probability region, with a sharp rise in payment holiday adoption rates from the trend witnessed in the 10%-

[2]While a payment holiday ensures no payments are recorded as missed, prospective lenders are able to observe historical loan balances. If payments are recorded as being met but overall loan balances are not decreasing, the lender can infer that a borrower is likely to have utilised a payment holiday.

40% region. After this 50% implied default probability range, we again document a linear trend. These observations suggest precautionary payment holiday behaviour around the high-financial-uncertainty implied default probability region, i.e. financially uncertain borrowers appear to take precautionary payment holidays.

Our findings contribute to the existing body of literature surrounding non-bank originated loans in three areas. First, we analyse default feature importance over time and show that P2P loan default factors are relatively stable over time. Total borrowing and account age appear to be the most important predictors, and this importance is maintained in both the pre-Covid and Covid sample periods. Postcode-level variables are relatively insignificant across all sample periods. Overall feature importance rankings are congruent across loan maturities.

Second, we document that the out-of-sample predictability of short-maturity loan defaults is lower than long-maturity loans. This maturity effect increases in strength in the Covid sample period. Examining average monthly loan repayments across maturities, our evidence suggests that higher monthly loan repayment-to-income ratios render short-maturity loans more susceptible to income shocks not captured in loan origination data. We add to the body of prior literature supporting the proposition that increased sensitivity to income shocks, which were more prevalent during the Covid period, result in poor default predictability for short-maturity loans in stress periods.

Third, we examine Covid payment holiday adoption rates and find evidence consistent with precautionary behaviour from borrowers with the highest levels of financial uncertainty. Using a combination of logit models and drawing on the findings in prior literature, we demonstrate the existence of a structural break in the dependency between default risk and payment holiday adoption rates for borrowers that are highly uncertain about their ability to repay, and conclude that high degrees of financial uncertainty led to precautionary borrowing and subsequent precautionary payment holiday behaviour during the Covid period. While previous literature has focused heavily on the key drivers of historical P2P loan defaults, we are unaware of any study explicitly examining the *change* in non-bank default factor significance over time. To the best of our knowledge, this is also the first paper to examine why borrowers might choose

to take payment holidays in a non-mortgage setting. Finally, our dataset enables us to uncover an income shock-driven maturity effect previously undocumented in the literature.

The paper is organized as follows. In Section 2, we summarise the academic literature on non-bank loan risk pricing/default factors. Section 3 focuses on our data, forecasting methodology and estimation strategy. Section 4 examines out-of-sample performance across a range of models, subsequently focusing on P2P loan default factor importance/stability over time, while Section 5 analyzes how precautionary borrower behaviour can explain observed Covid payment holiday adoption rates. Section 6 concludes.

2 Literature Review

Within the existing body of literature on P2P lending, a large number of papers focus on the determinants of default risk. Using data from Lending Club, Jin and Zhu (2015) explore traditional credit risk factors in a machine learning setting and find that loan terms, annual incomes, the amount of the loans, debt-to-income ratios, credit grades, and revolving line utilization play important roles in loan defaults. Their findings are echoed by Emekter et al. (2015), who find that credit grade, debt-to-income ratio, FICO score and revolving line utilization are significant in predicting loan defaults. Loans with lower credit grades and longer durations are associated with high mortality rates. Using a logistic regression approach, Serrano-Cinca et al. (2015) document similar significant default factors. The grade assigned by the P2P lending site is the factor that is most likely to predict default, but the accuracy of the model is improved by adding other information, especially the borrower's debt level. Nigmonov et al. (2022) utilise a probit regression analysis to empirically investigate the key macroeconomic factors that influence default risk in the peer-to-peer (P2P) lending market. By aggregating the United States (U.S.) state-level data with Lending Club's loan book, they show that higher macroeconomic interest and inflation rates increase the probability of default in the P2P lending market, with the impact of interest rates on the probability of default being significantly higher for loans with lower ratings. Meanwhile, Xu et al. (2021) use a binary logistic regression model along

with survival analysis to evaluate default risk and loan performance in UK P2P lending. Their empirical results indicate that credit group, loan purpose for capital needs, sector type, loan amount, interest rate, loan term, and the age of the company all have a significant impact on the probability of loan default. Among them, the interest rate, loan term, and loan purpose for capital needs are the three most important determinants of the probability of loan defaults and survival time of loans.

When examining the pricing of P2P loans, Michels (2012) uses data from Prosper.com to demonstrate an economically large effect of voluntary, unverifiable disclosures in reducing the cost of debt. His results show that an additional unverifiable disclosure is associated with a 1.27% reduction in interest rates and an 8% increase in bidding activity. However, Mild et al. (2015) show that market develop for P2P capital do not adequately reflect associated default risk and present a decision-support-tool to assist users in the estimation of default risk using public data. Going beyond traditional P2P credit scoring systems, Serrano-Cinca and Gutiérrez-Nieto (2016) instead propose a profit scoring model. They analyze factors that determine loan profitability, and observe that these factors differ from factors determining the probability of default. Specifically, a lender selecting loans by applying a profit scoring system using a multivariate regression outperforms the results obtained by using a traditional credit scoring system, based on logistic regressions.

More closely related to our paper, Malekipirbazari and Aksakalli (2015) use Lending Club data to examine the predictive performance of several simple machine learning methods. Their results indicate that the random forest-based method outperforms both the FICO credit scores as well as Lending Club proprietary credit grades in identification of good borrowers. Byanjankar et al. (2015) obtain good predictive accuracy in predicting P2P default rates using neural networks. Ariza-Garzón et al. (2020) assess the logistic regression model and several machine learning algorithms for P2P default classification. Their comparison reveals that machine learning alternatives are superior in terms of both classification performance and explainability. More precisely, SHAP values reveal that machine learning algorithms can reflect dispersion, non-linearity and structural breaks in the relationships between each feature and

the target variable. Turiel and Aste (2019) use several machine learning techniques applied to lending data in order to replicate lender acceptance of loans and predict the likelihood of default of issued loans. They propose a two-stage model, where the first stage predicts loan rejection, while the second one predicts default risk for approved loans. In the first stage, they find that a logistic regression to the best performer, while a Deep Neural Network model performed best during the second stage. When investigating the role of local housing prices in the default of P2P loans, Mo and Yae (2022) find that borrowers who reside at a postcode where median home price is one standard deviation higher than the cross-sectional average show a 0.75% lower default probability for a given level of risk. This postcode effect is three times stronger for homeowner borrowers with mortgages, demonstrating the importance of both postcode-level data and home-ownership data when evaluating default probability - an avenue we explore in our paper.

3 Research Design

In this section, we outline the research design for our empirical analysis. We begin with an overview of our data collection procedure, as well as outlining the various modelling frameworks used in our main empirical analysis. We conclude with a short discussion on our estimation strategy, as well as highlighting our chosen method for gauging the statistical performance of our P2P default model.

3.1 Data

Our sample consists of a unique dataset of P2P loans originated in the UK over the 2017-2021 period. The data is provided by a leading UK P2P loan platform and encompasses 591,400 loans, with loan sizes up to £25,000 and loan maturities ranging from 1-5 years. Our data provider operates as a pure middleman. Specifically, the P2P site receives a loan request, performs various credit checks, and uses a model to assign prospective borrowers to different risk categories. On the other side of the agreement, prospective investors/lenders decide how

much to lend and notify the platform as to the risk level they are comfortable with. These preferences are used to match borrowers and lenders. Upon successful lending, the P2P lender levies a"go-between fee" on each party. Our data provider bears no risk in the event of non-payment; all risk is borne by the lender.

For each loan, we have access to rich origination data (provided by the borrowers themselves and supplemented with credit rating agency (CRA) data) along with monthly loan performance data concerning the health of the loan post-origination. As a result, for a given x-month default window, we use monthly loan performance data to create a binary default variable for each loan, and splice this together with loan origination data to create a single dataset for our research. For our analysis, we classify a borrower as having defaulted if they are more than 3 months late with a due loan repayment. This assumption is in line with the commercial default flag adopted by our data provider; 3 or more consecutively missed monthly repayments constitutes a default. At the onset of the Covid-19 crisis, the introduction of borrower payment holidays granted borrowers the ability to pause loan repayments and temporarily avoid default. While the specifics of payment holidays will be covered in subsequent sections, we emphasise here that payment holidays do not change our definition of borrower default.

When used in a commercial setting, raw borrower data is often enriched with CRA-derived metrics to provide additional explanatory power. CRAs have thousands of available datapoints available for each borrower, and carry out a dimensionality-reduction approach to condense these datapoints into a handful of summary scores (e.g.,"Risk Navigator Score", "Over-Indebtedness Score"). While these scores are useful for enhancing model performance, they are opaque in nature, often highly correlated with traditional credit metrics and ultimately do not allow us to identify the core borrower characteristics responsible for driving default behaviour. As a result, while we choose to omit these CRA-derived summary scores from our model, we make use of them in future sections as a robustness check when assessing the ability of our model to adequately capture borrower credit risk.

Before deciding on the final set of explanatory variables in our model, we first clean the data as follows. To account for missing data, we use a combination of mean-imputation and

zero-imputation.[3] Missing values are represented by various error codes; a missing value can either represent a zero (i.e. a missing value for previous credit searches indicates this value is zero), or a value that is non-zero but not available due to user omission (i.e. a missing disposable income figure does not indicate income is zero). As a result, we use zero-imputation when a missing value indicates a zero and mean imputation when a missing value indicates data omission. We choose to delete loans where more than 5 explanatory variables are missing (imputing large numbers of variables associated with a particular loan implies low additional value when included in a training dataset). We also remove loans with negative CRA scores and negative borrower incomes.

All categorical variables are converted to dummy variables via a one-hot-encoding (OHE) methodology. A categorical variable with n categories is converted into $n-1$ dummy variables. A 1 is recorded in the relevant dummy category column, and a 0 is recorded in all other dummy category columns.

The final dataset contains 45 explanatory variables available for each originated loan. Table 1 presents descriptive statistics for each available default factor in our 45-factor model. We sort the variables based on 7 categories. *Limits* indicates variables related to borrower credit limits. *Borrowing* relates to prior borrowing activity. *Leverage* relates to various credit limit/income and borrowing/income ratios. *History* encompasses variables capturing prior loan applications and credit searches. *Postcode* variables represent geographical information on other borrowers located within the current borrower's residential postcode. *Personal* captures personal borrower characteristics, while *Income* is related to various income metrics. Finally, *Loan* variables cover loan-specific terms. In the table, bolded terms marked with an asterisk are variables that eventually appear in our reduced set of explanatory variables.

Table 2 offers a more in-depth perspective on borrower-level details in order to provide a view of the typical borrower operating on the P2P platform. Homeowners between the ages of 25 and 40 represent the most common demographic, with most borrowers opting for a loan

[3]Mean-imputation implies the mean of observed values for each variable are computed and any missing values for each variable are filled by the corresponding mean. With zero-imputation, missing values are filled with a value of zero.

amount less than £10,000. The majority of borrowers earn between £20,000 and £60,000 annually and over 30% are repeat customers. Finally, loan uses are evenly spread across home improvements, car purchases and debt consolidations.

3.1.1 Default Horizon

Our first research question focuses on time-variation in P2P loan default factors, with an emphasis on variation in the pre-Covid and Covid periods. As a result, it is important to select a default window that allows sufficient Covid-period default factor analysis. We define the beginning of Covid as March 2020, the month in which lenders initiated borrower payment holidays. To maximise the size of the Covid sample we can analyse, we opt for a 9-month default window allowing four quarters of Covid analysis[4]. Panel A of Figure 1 shows the cumulative distribution function (CDF) of all borrower defaults by month-on-book (MoB), i.e. the number of months since loan origination. We observe that around 97% of all defaults occur within 36 months of origination, 75% occur within 24 months of origination and roughly 40% occur within 12 months of origination. Our chosen default window of 9 months captures approximately 30% of all defaults. While this figure is relatively modest, we highlight two important reasons to justify our chosen default window.

First, adopting a longer default window would limit the degree to which the impact of Covid can be detected in defaults; long default windows would not isolate time periods where Covid-based shocks were at their peak. Second, given our borrower-level explanatory variables are determined at loan origination, long default windows will likely dampen the explanatory power of loan origination variables. However, only capturing 30% of all defaults is potentially pernicious in that model performance may be affected due to the large class imbalance between defaulting and non-defaulting borrowers. In other words, the smaller the default window, the lower the default rate will be and the harder it is to predict these defaults. We demonstrate in subsequent sections that such a class imbalance does not hinder the overall performance of our model - i.e. the default rate is high enough to train an accurate model.

[4] A 9-month default window gives us the ability to analyse all loans originated on or before March 2021.

To further justify our choice of a 9-month default window, consider Panel B of Figure 1. One potential criticism of a short default window is loan maturity bias - i.e. the risk that longer maturity loans are under-represented in the group of loans that default within 9 months of origination. Breaking down default CDFs by maturity, we see that a shorter default window does not bias the analysis against long-maturity loans; for both 4-year and 5-year loans, approximately 25% of all defaults occur within the first 9 months - a figure close to the 30% figure witnessed across all maturities. As a result, we opt for a 9-month default window when creating our binary default target variable. However, for completeness (and to ensure the robustness of our results), we demonstrate in the Appendix that all our key results also hold for both 6-month and 12-month default windows.

Figure 2 illustrates 9-month default rates over time for loans grouped by maturity. Considering all maturities collectively, we observe distinct periods of high and low default rates - of particular interest to us are pre-Covid and Covid period default patterns. From July 2019, we begin to see sharp declines in 9-month default rates. The reason for this observed decline stems from a large increase in borrower payment holiday adoption rates at the beginning of the Covid-19 crisis. As highlighted earlier, payment holidays allow borrowers to defer mandatory loan repayments and "kick the can down the road" when it comes to potential default. Given that July 2019 lies 9 months prior to the introduction of payment holidays and the onset of Covid-19, payment holidays began to reduce observed 9-month default rates from July 2019 onwards.

Since a large portion of payment holidays were granted during the immediate onset of the Covid crisis, loans originated after March 2020 had considerably lower payment holiday adoption rates. Hence we note a gradual uptick in default rates post-March 2020. However, in tandem with payment holidays, lenders introduced stricter lending criteria at the onset of Covid-19 (Ennis and Jarque, 2021). As a result, while default rates began to climb subsequent to March 2020, these rates did not return to their pre-Covid peak.

3.2 Forecasting Methods

Before discussing our chosen estimation procedure, we first run a horse race between a selection of linear and non-linear models (see, e.g., Holopainen and Sarlin, 2017). Our goal is to determine the class of model which delivers the best out-of-sample default prediction performance over time, and hence the model most suited to analysing time-variation in P2P loan default factors. We provide a brief summary of the relatively new XGBoost machine learning framework, leaving technical details on more widely-known algorithmic procedures to the Appendix.

3.2.1 XGBoost

XGBoost is a decision-tree-based ensemble machine learning algorithm implemented via a gradient boosting framework, developed by Chen and Guestrin (2016) who extended the existing principle of gradient-boosted trees (see Mason et al., 1999). While the random forest algorithm relies on the principle of bagging, i.e. multiple trees operating in parallel to reduce over-fitting, XGBoost adds additional trees in a sequential manner. The model is initialised with a naive prediction for each datapoint (usually the log-odds), and a weak learner/pruned decision tree is fit on the residuals of the initial model, derived via the gradient of the chosen loss function. The goal of fitting a weak learner to the residuals is to predict in advance the resulting residual error for a given input feature vector. For each terminal node in the tree, the weighted average residual is calculated, scaled by a learning rate, and added to the initial naïve prediction of all the training instances that fall within the particular terminal node. The resulting model is a combination of a naïve prediction and a scaled residual-correction model.

The XGBoost algorithm repeats this step many times, sequentially adding pruned decision trees to correct the residuals of the model in the previous iteration. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best in class. Due to the presence of both hardware optimisation and parallelisation capabilities, training times for

XGBoost are often very fast.

3.3 Estimation Strategy

3.3.1 Pre-processing

Default prediction is a classic example of an "imbalanced" classification task; that is, the ratio of defaults to non-defaults in a typical dataset is very low - in the order of 1-3%. In light of this unbalanced nature, models often struggle to learn the true dependence between input variables and a binary default response variable. To overcome this concern, we adopt two data pre-processing steps to maximise the chances of our model learning the true relationship present within our data and ultimately performing well out-of-sample.

The first pre-processing method we use is a data augmentation/over-sampling technique known as the Synthetic Minority Oversampling Technique, or SMOTE (Chawla et al., 2002). SMOTE works by selecting examples that are close in the input feature space, drawing a hyperline between the examples in the space and selecting a new artificial sample at a point along the line. Specifically, a random example from the minority class (i.e. default class) is first chosen. Then k of the nearest neighbors for that example are found (typically $k = 5$). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space. This procedure can be used to create as many synthetic examples for the minority class as are required.

While SMOTE can deliver substantial improvements in model performance, a general downside is that synthetic examples are created without considering the majority class (non-defaults), potentially resulting in ambiguous instances if there is a strong overlap for the classes. To circumvent this issue, we utilise SMOTE alongside the edited-nearest-neighbours (ENN) technique as proposed by Wilson (1972). When used as an undersampling procedure, the ENN algorithm can be applied to each example in the majority class, allowing those examples that are misclassified as belonging to the minority class (loan defaults) to be removed, and those correctly classified to remain. This approach leads to less "fuzzy" decision boundaries when

training a machine-learning model. For each instance a in the dataset, its three nearest neighbors are computed. If a is a majority class instance and is misclassified by its three nearest neighbors, then a is removed from the dataset. Alternatively, if a is a minority class instance and is misclassified by its three nearest neighbors, then the majority class instances among a's neighbors are removed (He and Ma, 2013).

Finally, we scale all input data using a Yeo-Johnson transformer as proposed by Yeo and Johnson (2000). The goal of any power-transform is to remove any potential skew in the distribution of input features, resulting in a more Gaussian-esque distribution more amenable to training a machine-learing model. In statistical terms, these are variance-stabilising transformations (Zheng and Casari, 2018).

3.3.2 Estimation Windows

Following common practice in machine learning, we split the data into three sub-samples: a training set to train the model, a validation set to calibrate model hyper-parameters[5] and a testing set representing the out-of-sample period in a typical forecasting exercise.

The existing literature often implements a time-series split whereby the training/ validation sets increase in size temporally in an expanding-window approach as the model is shifted forward through time, while the testing/out-of-sample set remains the same size, i.e. a fixed-window (see Bianchi et al., 2021). In this paper, we follow in the spirit of Bianchi and Shuaib (2021) and opt for a fixed, rolling 3-month observation window size for each of the training, validation and testing datasets. The 3-month training and subsequent 3-month validation datasets are utilised for model hyperparameter calibration. Once optimal hyperparameters have been deduced, the optimised model is trained on the entire 6-month train+validation dataset and performance is evaluated on the 3-month out-of-sample test dataset. This evaluation procedure then rolls each of the training, validation and the testing samples forward by 3-months, and repeats until sample end.

[5]A model hyperparameter is a configuration that is external to the model and whose value cannot be estimated from data. An example is the number of trees to use in a random forest. These differ from model parameters which are estimted during model training (i.e. beta coefficients in a logit model)

To illustrate this, consider months $\boldsymbol{m}_1, ..., \boldsymbol{m}_{12}$ for a given year $\boldsymbol{y}$. Data from $\boldsymbol{m}_1$ to $\boldsymbol{m}_3$ is used for training, and $\boldsymbol{m}_4$ to $\boldsymbol{m}_6$ is used for hyperparameter validation. Subsequently, the calibrated model is trained on all data from $\boldsymbol{m}_1$ to $\boldsymbol{m}_6$ and evaluated on the $\boldsymbol{m}_7$ to $\boldsymbol{m}_9$ out-of-sample test period. This procedure then rolls forward and repeats, where $\boldsymbol{m}_4$ to $\boldsymbol{m}_6$ is the new training window, $\boldsymbol{m}_7$ to $\boldsymbol{m}_9$ is the new validation window and $\boldsymbol{m}_{10}$ to $\boldsymbol{m}_{12}$ is the new out-of-sample test window.

3.4 Statistical Performance

Model performance in an imbalanced classification setting is a contentious issue, and the "correct" evaluation metric varies depending on the problem at hand. A simple accuracy score is not fit for evaluation in this particular setting: if 98% of all instances belong to the "no-default" class and a model predicts "no-default" for every borrower, our model will have an accuracy score of 98%. Ostensibly this is a very strong performance, but the model misses *all* the customers who ultimately default. As a result, any chosen evaluation metric needs to separately assess the minority (default) and majority (no-default) classes.

The two metrics most suited to our problem are the receiver-operating-characteristic (ROC) area-under-curve (AUC) and the precision-recall (PR) AUC. The ROC-AUC plots the model true-positive rate (TPR) against the false-positive rate (FPR) for varying probability thresholds and calculates the AUC. Meanwhile, the PR-AUC plots model precision against the TPR for varying probability thresholds and calculates the AUC. Both PR-AUC and ROC-AUC values lie in the range [0,1], with 1 representing a perfect classifier score.

The PR-AUC metric is often cited as the appropriate approach for unbalanced classification tasks (Saito and Rehmsmeier, 2015), due in part to the stricter way in which false positives (FP) are accounted for. The precision metric in the PR-AUC considers true positives (TP) as a percentage of all positive predictions (TP + FP) made by the model. Conversely, the ROC-AUC considers FPs as a proportion of all negative classes (TN + FP) in the model, otherwise known as the FPR. If the ratio of negative:positive classes in the dataset is very high, the FPR can give an unrealistically high assessment of model performance relative to the precision score

(Sofaer et al., 2019).[6]

However, the PR-AUC makes the implicit assumption that FPs are as deleterious as false negatives (FN). In practice, this is not the case. For a default-based model, the business cost of turning away a good borrower (i.e. a FP) is considerably lower than the cost of lending to a bad borrower (i.e a FN) (Dudík et al., 2020; Mahmoudi and Duman, 2015). Consider a model with a very high PR-AUC but low ROC-AUC. This particular model only predicts a default if it is almost certain a customer will default. As a result, there are very few FPs and few good borrowers are turned away. The downside is that a large number of bad borrowers receive a loan and default as a consequence. Hence, while model PR-AUC is very high, overall model performance is sub-optimal in a default prediction setting. Conversely, consider a model with very high ROC-AUC and low PR-AUC. Such a model turns away many good borrowers (high FP rate), but rarely lends to bad borrowers (low FN rate). Furthermore, while the FP rate may be high, this simply indicates that a given borrower does not default within the specified default prediction window (9-months in our case). It does not rule out cases where customers default outside of the initial 9-month post-origination window. In addition, even though a predicted-bad borrower may not ultimately default (i.e. a FP), the borrower in question may still be under financial stress, restructure the loan, create additional paperwork for the bank etc. A FP does not necessarily indicate a "good" borrower.

The ultimate takeaway is that FPs in our particular setting are not as pernicious as FNs. As a result, we follow in the spirit of Blöchlinger and Leippold (2006) and adopt the ROC-AUC as our chosen metric of statistical performance.

4 P2P Default Factors

Using the pre-processing and estimation strategy highlighted in Section 3, we compare the performance of each model and select the best-performing framework on average across all out-of-sample windows as our model of choice. This choice model is used to rank default

[6]Note that both PR-AUC and ROC-AUC share the TPR metric in common: TPR $= \frac{TP}{TP+FN}$

factors over time via a permutation feature importance, with a focus on observing how default factors vary in the pre-Covid and Covid periods. We also examine model stability via partial dependency plots, with a view to assessing how the marginal impact of individual explanatory variables on our target default variable vary in the pre-Covid and Covid periods.

4.1 Out-of-Sample 45-Factor Model Performance

We start by using all 45 available default factors in our analysis. Each model is trained and validated via a 3-month training window and a subsequent 3-month validation window. Once optimal hyper-parameters are calibrated, the optimised model is trained on the entire 6-month train+validation dataset and tested on the 3-month out-of-sample window occurring immediately after the validation window. Table 3 summarises the performance of each model class across all out-of-sample quarterly windows in our dataset. For each out-of-sample quarter, we provide the ROC-AUC score of each model. Given any ROC-AUC score above 0.50 indicates model skill, the first takeaway from this table is that both linear and non-linear models possess a high degree of skill. With the exception of the kNN algorithm (which performs relatively poorly), all other models maintain an ROC-AUC well above 0.70 for the duration of our pre-Covid sample. Tree-based non-linear models (i.e. XGBoost and random forest) both achieve ROC-AUC scores close to 0.80 in several out-of-sample pre-Covid quarters. Before diving into the details of individual model performance, we perform a feature selection exercise detailed in the following section.

4.1.1 Feature Selection

While our 45-factor model ostensibly indicates a broad spectrum of features, high correlation between the variables drastically reduces the explanatory value addition of a significant number of our explanatory variables. Furthermore, the high correlation impacts our ability to deduce the standalone importance of individual explanatory variables (Tolo*c*si and Lengauer, 2011; Vettoretti and Di Camillo, 2021) – a key research question in our paper.

In order to reduce the dimensionality of our model, we therefore employ a recursive feature

elimination (RFE) approach standard in the machine learning literature. Our XGBoost model is used to iteratively eliminate all factors from our 45-factor model that do not result in a ROC-AUC drop when removed from our model during the first training/validation period. Using this approach, a total of 35 factors are dropped during the first training/validation stage. The final 10-factor model is used for all subsequent training, testing, and validation procedures. Given the RFE procedure is not carried out using any out-of-sample data (we choose to utilise a very small portion of in-sample data), this approach is unlikely to be biased. An alternative approach is to repeat the RFE procedure for each training/validation window. However, we choose not to adopt this methodology for one key reason: our research question involves ranking default factors over time. To accomplish this, we need a constant set of factors that are utilised over all sliding windows in our analysis. Repeating the RFE approach in each sliding window would imply certain factors being dropped from the model if they subsequently became irrelevant - this would prevent us from observing changes in their explanatory power over time (if a factor is dropped, we cannot explicitly observe the change in ranking).

A potential criticism of such an approach is the risk of omitting factors that are irrelevant in the first training/validation window but important in future sample periods. To mitigate such concerns, we use the aforementioned credit rating agency "Risk Navigator Score" (RNS) as a robustness check to demonstrate the power of our reduced-dimension default model. By construction, the RNS is intended to capture credit risk across several hundred borrower-level characteristics in a single score. Table 4 demonstrates results from a regression of the RNS on all variables in our 45-factor/10-factor model respectively. Looking at the R^2 metrics, we see only a modest drop of 6.9% when comparing the R^2 across our 45-factor and 10-factor models. As a result, our reduced-dimension 10-factor model does not appear to result in a significantly reduced ability to capture borrower credit risk. We undertake further robustness checks in subsequent sections.

Figures 3 and 4 provide additional descriptive statistics for the remaining variables in our 10-factor model. Figure 3 depicts a right-skewed distribution for the majority of factors, with the exception of loan term, which is left-skewed. Both oldest account age and healthy borrower

postcode index variables display more Gaussian characteristics. Figure 5 demonstrates the cross-correlations between all remaining variables in our 10-factor model. Most of the ten factors have low correlations with each other with one exception - both postcode-level variables in our 10-factor dataset have an observed correlation of -57%. However, given that these postcode variables both lie within the "postcode" category of default factors, this correlation is unlikely to be hazardous to our goal of ranking default factors over time.

4.2 Out-of-Sample 10-Factor Model Performance

Comparing Table 3 and Table 5, the 10-factor model marginally outperforms the 45-factor model across a majority of out-of-sample periods. With the exception of kNN, most models experience an average ROC-AUC performance increase of $\sim$0.01 when the input dimension is reduced from 45 to 10, with the neural network recording a $\sim$0.05 improvement in performance. This average outperformance is observed both in the pre-Covid period and (to a lesser degree) in the Covid period. While this performance boost is relatively minor, the ultimate purpose of this comparison is to demonstrate that using our aforementioned dimensionality reduction approach, the 10-factor model does not underperform the initial 45-factor model. In other words, we do not lose explanatory power by reducing the dimensionality of our model from 45 to 10.

Examining the performance of each individual 10-factor model, we highlight the comparison-of-means statistics detailed in Table 6. To calculate these statistics, we use a repeated bootstrap sampling approach to generate probability distributions for the out-of-sample ROC-AUC of each model during the pre-Covid and Covid periods. For each rolling test period, a 10% sub-sample of the test data is taken and the ROC-AUC is calculated for a given model. The average ROC-AUCs for the pre-Covid and Covid periods are then calculated. Repeating this process 100 times (each time with a new random 10% sample of the test data), we calculate the mean and standard deviation of the ROC-AUC for each model across the the pre-Covid and Covid periods. These figures allow us to compare pairwise model performance by way of an independent t-test.

When it comes to overall model performance, Table 6 indicates the random forest and XGBoost models notably outperform all other models across the majority of out-of-sample quarters in our dataset. In addition, both tree-based models benefit from fast training times and higher model transparency. While the XGBoost model does slightly outperform the random forest model, this outperformance is statistically insignificant (t-statistic of 0.615 overall, with a t-statistic of 0.432 during the pre-Covid period rising to a significant 2.175 during the Covid period). As a result, our top-performing models of choice are a tie between the random forest and XGBoost models. However, in light of its superior performance during the Covid period, we report results for the XGBoost model for the remainder of this paper.

While pre-Covid performance is strong for all models bar the kNN, performance begins to tail off slightly during the Covid period, suggesting that Covid had a non-negligible impact on P2P lending markets. On average, our selection of models experiences a 0.08 drop in ROC-AUC when transitioning from the pre-Covid to the Covid period. Using the above comparison-of-means approach, we record a highly significant t-statistic of 23.69 when comparing the average pre-Covid ROC-AUC of the XGBoost model to the average Covid ROC-AUC. Similarly, t-statistics of 25.04 and 26.02 are observed respectively for the neural network and logistic regression models. Consequently, the Covid period introduced a material shift in the predictability of 9-month P2P loan defaults. However, while most models saw a notable drop in performance during the Covid period, the ROC-AUC remains well above 0.5 and hence is still indicative of model skill.

Table 7 Panel A dives deeper into the performance of our top-performing XGBoost model across different loan maturities. For clarity, the training and validation dataset remains the same (i.e. all maturities) to ensure a like-for-like comparison - we simply evaluate the model on different maturities within each test dataset. Two points are immediately clear - first, short-maturity loans have a consistently lower ROC-AUC relative to long-maturity loans (average ROC-AUC difference of 0.067 during the pre-Covid period, rising to a difference of 0.134 during the Covid period). Second, the drop-off in ROC-AUC during the Covid period is considerably greater for short-maturity loans than long-maturity loans. To demonstrate this

significance, we again employ a set of comparison-of-means tests. Comparing the average ROC-AUC outperformance of our XGBoost model on long-maturity loans relative to short-maturity loans across the entire sample, we observe a highly significant t-statistic of 38.97. Furthermore, comparing the average performance drop when predicting pre-Covid vs. Covid loan defaults, a t-statistic of 21.69 is observed for short-maturity loans whilst a less significant t-statistic of 12.19 is observed for long-maturity loans. Our findings imply that short-maturity loan defaults are harder to predict, particularly during the Covid period. There is a also a greater degree of instability in short-maturity model performance during both the pre-Covid and Covid periods, with a Covid out-of-sample ROC-AUC standard deviation of 5.4% vs. 4.1% for long-maturity loans and a pre-Covid out-of-sample ROC-AUC standard deviation of 1.3% vs. 0.8% for long-maturity loans. This effect is not explained by an imbalance in the rate of defaults by maturity (1.3% for long-maturity loans compared to 1.5% for short-maturity loans).

As far as we are aware, we are the first to explicitly compare the predictability of short-maturity loan defaults vs long-maturity loan defaults. Our results are consistent with a hypothesis posed in previous research that larger monthly repayments drive significant variation in the default behaviour of short-maturity loans relative to long-maturity loans (Gaudêncio et al., 2019) - i.e. higher quarterly payment installments (relative to income) for short-maturity loans imply short-maturity borrowers have more difficulty repaying in the event of a temporary shock to income. For our dataset, we observe monthly repayment-to-income ratios 25% higher in 1-year loans vs. 5-year loans.

This theory is consistent with Hertzberg et al. (2018), who find that short-maturity loan repayments are more susceptible to income shocks, and borrowers who foresee future income shocks self-select away from short-maturity loans. Given we only use loan origination data in our analysis, this suggests short-maturity loans are more susceptible to income shocks not captured in our explanatory variables, and explains why ROC-AUC standard deviation is higher for short-term loans. This also explains why short-maturity loan predictive performance drops off significantly in the Covid period relative to long-maturity loans; the Covid period in the UK saw a considerable increase in net household income shocks (Brewer and Tasseva, 2020;

Adams-Prassl et al., 2020). As a result, we conclude that using origination-only data compromises the ability of credit models to predict defaults for short-maturity loans, particularly during periods when income shocks are more prevalent. Table A.2 demonstrates our findings are robust to the choice of default window size.

As an additional robustness check, we examine the out-of-sample performance of our top-performing 10-factor XGBoost model both including and excluding the credit rating agency "Risk Navigator Score" (RNS). Panel B of Table 7 shows the results of our analysis. RNS inclusion improves ROC-AUC by an average of only 0.016 during both the pre-Covid and and Covid periods. When carrying out a comparison-of-means test, a t-statistic of 1.59 is observed during the Covid period indicating an insignificant performance boost from the inclusion of the RNS in our model. As a result, we find that including the RNS in our 10-factor XGBoost model does not have a material impact on out-of-sample performance; our chosen 10 factors alone suitably capture borrower credit risk. In light of these findings, we opt to report our 10-factor XGBoost default model results for all subsequent analyses in this paper.

4.3 Feature Importance Ranking Methodology

In order to rank P2P loan default factors by importance, we use the permutation feature importance (PFI) methodology. First proposed by Breiman (2001) for random forests and later generalised to a model-agnostic version by Fisher et al. (2019), this measure calculates the importance of a feature by calculating the increase in the model's prediction error (or decrease in the model's chosen performance metric) after shuffling the values of a given feature across all instances in a given test dataset. A feature is deemed important if shuffling its values increases the model error/decreases the model performance; in this case the model successfully incorporates the feature when making its prediction. A feature is less important if shuffling its values leaves the model error unchanged or results in a decrease in error.

Our PFI strategy operates in three steps. Firstly, the model is trained using the training + validation datasets and the overall model ROC-AUC is calculated on the original test dataset. Subsequently, a given feature is shuffled across all instances in the test dataset. The fitted

model in step 1 is evaluated on the shuffled test dataset, and the ROC-AUC drop is calculated relative to the non-shuffled ROC-AUC. This shuffling is performed using the default number of recommended repetitions (5), with the average ROC-AUC drop recorded. This step then repeats for each feature. Finally, features are ranked by the drop in ROC-AUC associated with their shuffling.

A point of dispute in the literature is whether the training or test dataset should be used for PFI. Model performance scores based on training data are often unreliable, particularly if a model is overfit. Given that feature importance relies on model performance, PFI scores based on training data can mistakenly indicate the importance of a particular feature when in reality the model is simply overfitting and the feature in question is not relevant. As a result, we opt to perform PFI on each test dataset in our sliding window approach.

4.4 Time Variation in P2P Default Factors

For each out-of-sample quarter, we first calculate variable rankings based on PFI, then normalise the ROC-AUC drops for each variable to lie in the interval [0,1]. As a result, for each out-of-sample quarter, the most important feature is given a score of 1 and the least important feature a score of zero. Table 8 displays the results of our P2P loan default factor rankings over time. Important factors are coloured blue, while the least important factors are coloured red. Panel A shows that both the most important and least important features are stable over time. Total borrowing (*Total Debt*) remains the most important feature for 10/14 out-of-sample quarters, and ranks as either the second or third most important feature across the remaining quarters. Interestingly, both postcode-level variables (*Delinquent Accounts* (*Postcode*), *Healthy Accounts* (*Postcode*)) are consistently ranked as either the least important or second-least-important feature. These results are also observed both in the pre-Covid and Covid periods. In summary, for both the most important and least important model features when predicting 9-month defaults across all loan maturities, there is a considerable degree of feature stability in both pre-Covid and Covid periods.

There is more variation when examining other default factors across all maturities in Panel

A. First, the loan term increases in importance as we enter the Covid period. This is unsurprising given our conclusions in the previous section; short-maturity loans are more sensitive to income shocks, with this effect particularly notable during the Covid period. This effect reduces as we move towards the end of the Covid window. We also note an increase in the significance of a borrower's oldest active account age during the Covid period, potentially an indication that having been an active/healthy borrower over an extended number of years is suggestive of resilience during shock periods. This observation is consistent with Moffatt (2005), who show that default probability is decreasing with borrower age (see Capozza et al., 1997; Jacobson and Roszbach, 2003; Cairney and Boyle, 2004 for similar findings). We also note a decrease in the importance of short-term indebtedness (captured by *ex.Mortgage Balance to Limits*) and an increase in the importance of long-term secured leverage (captured by *Secured Debt-to-Income*) in the Covid period. The reason for this appears straightforward - long-term secured leverage captures mortgage-based borrowing and hence is indicative of whether a borrower has taken a mortgage. Previous literature has shown that borrowers with mortgages are less likely to be facing both wealth and credit constraints (Barakova et al., 2003; Bostic et al., 2005; Rosenthal, 2002), hence homeownership (serving as a proxy for credit quality) is likely to be a more important factor during the Covid period where wealth constraints are more prevalent due to the presence of income shocks.

Panel B of Table 8 reports results for short-maturity loans. During the pre-Covid period, total borrowing (*Total Debt*) still appears to be the most significant variable for the majority of quarters, and both postcode-level variables ranking as the least important across most pre-Covid quarters. *Term* ranks as the least important factor in some quarters, which is unsurprising given that short-maturity loans only encompass 1 and 2 year maturities - i.e. we observe low variance in loan term for this particular sub-sample. Both oldest active account age (*Oldest Account Age*) and soft credit checks (*# Soft Credit Checks*) rank highly (as with all-maturity loans in Panel A), short-term indebtedness (*ex.Mortgage Balance to Limits*) drops in importance during the Covid period and long-term secured leverage (*Secured Debt-to-Income*) rises in importance during the Covid period, much like Panel A. Overall, while short-maturity

feature ranking trends are broadly in line with all-maturity loans, we observe more instability on a quarter-to-quarter basis. Table 7 showed that short-maturity loans have lower default predictability relative to long-maturity loans, with this effect stronger in the Covid period. As a result, we would expect more feature instability when using loan origination variables for default prediction in short-maturity loans.

Table 8 Panel C reports results for longer-maturity loans. This is more consistency in quarter-on-quarter feature importance variation, with this effect driven by strong predictability/model performance across both pre-Covid and Covid periods. *Total Debt* remains the most important variable for all pre-Covid periods, while the *Delinquent Accounts (Postcode)* postcode variable remains unimportant. Oldest active account age (*Oldest Account Age*) remains highly important, with this importance increasing during the Covid period. We also observe a shift away from the importance of short-term, unsecured indebtedness factors (*ex.Mortgage Balance to Limits, Revolving Balance to Limits*) towards mortgage-based factors (*Secured Debt-to-Income*). Crucially, loan term appears relatively unimportant for the majority of both pre-Covid and Covid quarters. Whilst minor disparity is observed between 1-year and 2-year loans in Panel B (i.e. loan term retains some importance even within the short-maturity sub-sample), no such disparity is observed for 4-year vs 5-year loans in the long-maturity sub-sample. This result is unsurprising given our prior findings - lower average monthly repayment-to-income ratios imply less sensitivity to income shocks for long-maturity loans.

As an additional robustness check, we demonstrate that model accuracy and feature importance rankings are not contingent on our choice of testing, training, or validation window size. In Table 9, we demonstrate that the out-of-sample ROC-AUC for our XGBoost model is consistently high across a range of different train-validation-test window permutations. We subsequently show that default factor stability is consistent with our above findings across all training-validation-testing window permutations. Finally, Table A.1 demonstrates that factor stability over time is not influenced by our choice of default window.

4.5 Model Stability

To further examine the stability of default factors during the pre-Covid and Covid periods, we utilise an "explainable AI" technique known as the partial dependency plot (PDP). First proposed by Friedman (2001), a PDP indicates the marginal effect a given feature has on the predicted outcome of a machine learning model and can show whether the relationship between the target and a feature variable is linear, monotonic, or more complex. The partial dependence function is defined as:

$$f_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C) \tag{1}$$

The x_S terms are the features for which the partial dependence function should be plotted (in our case we examine each feature individually so x_S is a single feature) and X_C are all other features used in the model $\hat{f}$. By marginalizing over all other features, we derive a function that is dependent only on the feature in S (Molnar, 2020).

The partial function $\hat{f}_S$ can be estimated by calculating averages across all instances in our test data:

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)}) \tag{2}$$

This estimated partial function tells us the average marginal effect on the prediction for a given feature S. A key assumption of the PDP is that features in C are uncorrelated with features in S.

For our analysis, we compare PDPs for each explanatory variable during the *in-sample* 6-month pre-Covid and initial 6-month Covid periods.[7] PDPs display the marginal dependency between the explanatory variables and target variable learned via the training dataset,

[7]To capture the period of time in which Covid shocks were at their strongest, we chose to use 6 months as the window for Covid PDP analysis. To ensure a like-for-like comparison, this also implies selecting a 6-month pre-Covid PDP window

regardless of whether the actual PDP plots themselves are generated via in-sample or out-of-sample data. In order to capture this dependency as accurately as possible during the Covid window (and evaluate whether a dependency shift has occurred), we need to maximise the number of "peak" Covid months in the training sample. As a result, we choose to carry out in-sample PDP analysis for our Covid window using the first 6 months of Covid data for both model training and PDP generation. Similarly, we also carry out in-sample PDP analysis for our pre-Covid window using 6 months of pre-Covid data for both model training and PDP generation.[8]

In light of its superior performance, we use the 10-factor XGBoost model. Given the 6-month in-sample window size, we adopt a training window of 3-months along with a 3-month validation window for model hyperparameter optimisation. Once optimal hyperparamaters are deduced, the optimised model is trained on the entire 6-month training+validation dataset and PDPs are calculated for this same 6-month train+validation period. Our findings are summarised in Figure 6.

For both total borrowing (*Total Debt*) and long-term secured leverage (*Secured Debt-to-Income*), we observe a step-like decreasing function – i.e. default probability decreases in the level of total debt and the secured leverage ratio. While this may initially seem counter-intuitive, we note that both the total borrowing and secured leverage metrics employed here include mortgage borrowing – i.e. borrowers with high *Total Debt* and *Secured Debt-to-Income* values indicate home ownership. As a result, these partial dependencies suggest that homeowners have lower P2P loan default rates (all else being equal). We note the observed step function for both these variables remains fairly constant in both the pre-Covid and Covid periods – i.e. no structural change is graphically observed. For soft credit checks (*# Soft Credit Checks*), we observe an increasing function as expected – higher numbers of soft credit checks lead to higher implied default probabilities. This relationship is also observed in both the pre-Covid and Covid periods for hard credit checks (*# Hard Credit Checks*). For both short-term indebt-

[8] Given the goal of any PDP-based analysis is to examine the structural relationship between the explanatory and dependent variables and not to demonstrate predictive power, using an in-sample approach to PDPs is unlikely to be contentious.

edness metrics (*ex.Mortgage Balance to Limits, Revolving Balance to Limits*) we observe an increasing response function during the pre-Covid period, ie. higher short-term leverage results in higher implied default probabilities as expected, given our short-term leverage metrics do not include mortgage borrowing. However, during the Covid period we see a breakdown in the structural dependency of both variables - a flat dependency is observed. This is in line with our findings in 4.4, where non-mortgage, short-term indebtedness factors lose their explanatory power during the Covid period. With regards to loan term, in the pre-Covid period, we observe an increase in implied default probability when transitioning from 1 to 2 year maturities, with default probability decreasing for maturities greater than 2 years. For the Covid window, there is a similar increase in implied default probability when transitioning from 1 to 2 year maturities, but there is no maturity effect on default probability for maturities greater than 2 years. With regards to oldest account age (*Oldest Account Age*), we observe a fairly linear decreasing relationship in both the pre-Covid and Covid periods – i.e. older account age is associated with lower implied default probability as hypothesised in Section 4.4. Finally, we observe both postcode-level variables possessing low explanatory power in both periods, with no observed structural dependency changes.

Overall, our PDP-based analysis emphasises the relative stability of P2P loan default factors when comparing default behaviour across pre-Covid and Covid periods.

5 Payment holidays and precautionary behavior

In this section, we begin with a brief introduction to Covid payment holidays, subsequently exploring in-depth the relationship between P2P loan payment holidays and implied default probabilities. Our ultimate goal is to analyse how financial uncertainty and behavioural biases affected rational decision making during the onset of the Covid crisis.

5.1 Payment Holidays

Given the relatively recent nature of the Covid crisis, obtaining payment holiday data from lenders is not an easy task. As a result, the existing body of literature on payment holidays is sparse. The European Data Warehouse investigated some of the key characteristics of mortgages with signs of payment holidays (EDW, 2021). They find that high Current Loan to Value (CLTV) loans were more likely to have been modified than low CLTV loans, with this effect more visible in the U.K. where approximately 40% of loans with a CLTV greater than 90% showed signs of payment holidays. Overall, payment holiday adoption rates appear to be correlated with borrower default risk. Similar results are obtained by Gaffney et al. (2020), who examine payment holiday behaviour in Irish mortgages and find a close relationship between payment breaks and high loan-to-income ratios at origination, especially among more recent vintages of lending.

In contrast, the Bank of England (BoE) published initial findings on payment holiday behaviour within the U.K. market, with survey evidence suggesting that many payment holidays at the onset of the crisis may have been taken for precautionary reasons - about a third of households who took a payment holiday did not end up experiencing a fall in income (Franklin et al., 2021). Payment holidays helped these households to manage the uncertainty around their future financial situations, but the vast majority of those that took out a payment holiday have since resumed full or partial repayments. In a similar vein, Moody's reported that almost all UK borrowers who took payment holidays on their home loans during the first wave of the coronavirus crisis had resumed payments by the end of 2020 (Manchester, 2021). Both the Moody's and BoE reports indicate the potential presence of precautionary payment holiday behaviour – i.e. observed payment holiday adoption rates are potentially incongruent with borrower implied default probabilities.

In light of the above, a natural question is whether borrowers were actively engaging in precautionary payment holiday behaviour during the onset of the Covid crisis, or whether payment holiday adoption rates were in line with implied default probabilities. To the best of

our knowledge, we are the first to examine the relationship between default probability and payment holiday utilisation rates, the first to use lender data to assess precautionary borrower behaviour during the Covid crisis, and the first to research payment holidays in a non-mortgage setting.

Given our ultimate aim of comparing payment holiday adoption rates to implied default probabilities, we use a 9-month payment holiday indicator to ensure a like-for-like comparison – i.e. we examine if a borrower adopted a payment holiday within 9 months of loan origination.[9] Figure 7 illustrates payment holiday adoption rates for borrowers across our sample period. Across all maturities, there is a sharp rise in 9-month payment holiday adoption rates for loans that originated in June 2019. As the majority of payment holidays were made during the initial onset of the Covid crisis, we would expect 9-month payment holiday adoption rates to increase rapidly 9-months prior to the Covid crisis (i.e. June 2019), and this is indeed our observation. Payment holiday adoption rates peak at $\sim$ 14% in Oct-2019, before a steep decline at the beginning of 2020. While modest levels of payment holiday adoption are observed post-March 2020, we choose to focus our attention on the 9-month pre-Covid period where payment holiday activity is at its highest level.

5.2 Financial Uncertainty and Payment Holiday Adoption Rates

Given our period of interest is the 9-month pre-Covid window where 9-month payment holiday adoption rates are at their highest, our first task is to generate model-implied default probabilities for all loans originated in this pre-Covid window. To accomplish this, we use an identical methodology to that described in Section 3.3.2 – i.e. a 3-month training period and 3-month validation period are used for hyperparameter optimisation in our XGBoost model, with the optimised model trained on the entire 6-month training+validation dataset and used to generate out-of-sample default probabilities for the subsequent 3-month out-of-sample period. Rolling this approach forward quarter-by-quarter allows us to generate out-of-sample 9-month implied default probabilities for each loan originated in our 9-month pre-Covid window of interest.

[9]Since our default indicator of choice covers a 9-month post-origination window, we choose to examine payment holiday behaviour over this same 9-month period.

Once these out-of-sample model-implied default probabilities have been generated, our next step is to examine how payment holiday adoption rates vary according to a borrower's model-implied default probability. Given the granular nature of these default probabilities (usually to 4 d.p.), some form of bucketing is required – i.e. for a given implied default probability bucket (e.g. 5%-10%), we calculate the payment holiday adoption rates for all borrowers lying in the bucket. The bucket size is a balancing act between granularity and data availability – too wide a bucket and the true relationship across implied default probabilities is obfuscated, too narrow a bucket and small borrower numbers within each bucket can result in misleading adoption percentage rates. To emphasise the reliability of our results, we explore a range of bucket widths from 3% to 5%. For each bucket, we calculate the 9-month payment holiday adoption rate, i.e. what percentage of borrowers in each implied default probability bucket adopt a payment holiday within 9-months of loan origination. As our default probability window matches our chosen payment holiday adoption window (both 9-months), our goal in this exercise is to observe whether model-implied default risk over the 9-month post-origination period is correlated with payment holiday adoption rates over this same window. In the absence of precautionary/strategic behaviour, we would expect to see payment holiday adoption rates increasing monotonically with model-implied default probability.

We hypothesize that if implied default probabilities are very high or very low, there is a low degree of financial uncertainty as the borrower in question is almost certain to either default or not default respectively. Meanwhile, borrowers with implied default probabilities close to 50% are highly uncertain – they are unsure if they will or will not default - both are equally likely, as there is 50% chance of default and a corresponding 50% chance of no-default. This interpretation rests on the assumption that borrowers are aware and conscious of their own degree of financial uncertainty – however, we do not believe this to be an extreme assumption given the asymmetric nature of a borrower-lender relationship implies any given borrower will possess at least as much information about their risk type as a prospective lender. This interpretation also rests on the accuracy of our implied default probabilities. As we note in Section 4.2, the strong out-of-sample performance of our XGBoost model provides strong

evidence on model accuracy.

Looking at each panel of Figure 8, we observe a fairly linear trend from an implied default probability of 10% to 40% - in this region, payment holiday adoption rates are increasing linearly with implied default probabilities. However, around the 50% implied default probability region a structural break is observed, with a sharp rise in payment holiday adoption rates incongruent with the trend witnessed in the 10%-40% region. Post-50% implied default probability, we see a return to a linearity. These observations suggest precautionary payment holiday behaviour around the high-financial-uncertainty implied default probability region, i.e. financially uncertain borrowers appear to take precautionary payment holidays. We hypothesise that around the 50% implied default probability region, borrower uncertainty is at its peak; borrowers in this region are as likely to default as they are to remain solvent. Due to this uncertainty, precautionary payment holidays are adopted as a way of hedging the fear of the unknown.

One could argue that in the absence of any consequences, all borrowers (regardless of financial health or uncertainty) would adopt a payment holiday. If a lender offers borrowers the option to delay repayments at no additional cost, a rational borrower would likely accept such a proposition. However, payment holidays are not without consequence. As previously highlighted, payment holiday adoption has the potential to negatively impact a borrower's credit score. In light of this, healthy borrowers are unlikely to adopt precautionary payment holidays: their default risk is low, so the prospect of a reduced credit score is unjustified. In the high-financial-uncertainty region, this trade-off changes. The prospect of potential default is more pernicious than a lower credit score, so borrowers are willing to accept a credit score hit if future default can be prevented. Conversely, while demand for payment holidays may be very high among high-risk borrowers (if a borrower knows they are going to default, there is no downside to taking a payment holiday and prolonging the inevitable default), lender discretion[10] concerning payment holidays implies 100% adoption rates are not observed.

[10] A borrower must apply for a payment holiday - it is not an automatic process. The lender then has discretion as to whether the payment holiday is approved or not.

In previous literature, it has already been established that financial uncertainty leads to precautionary borrowing. Druedahl and Jørgensen (2018) observe precautionary credit card borrowing for intermediate liquid net worth values. Alan et al. (2012) observe that constrained economic agents borrow pre-emptively in good times to ensure access to credit and thus to enable higher consumption in bad times. Our results in Figure 9 show a large spike in debt-income ratios around the 50% implied default probability area, consistent with this prior literature and also consistent with the view that 50% implied default probability borrowers are financially uncertain.

To investigate a structural break in payment holiday behaviour, we run a logit model where a variety of debt-income, credit check and short-term indebtedness variables are interacted with both mid implied default probability and high implied default probability dummy variables (given we have a total of three implied default probability regions, we only require two dummy variables). We define the mid region to represent implied default probabilities between 40% and 60%, and the high region to represent implied default probabilities greater than 60%. Our results are displayed in Table 10. For all four logit models, we observe no significant dummy variable interactions for *ex.Mortgage Balance to Limits* and *# Soft Credit Checks*. We conclude there is no evidence of any structural break driven by our credit check and short-term indebtedness metrics. When looking at our debt-income dummy interactions, we see a different story. Across all four logit model permutations, this interaction is highly significant with the mid implied default probability dummy and insignificant with the high implied default probability dummy. Examining the coefficient signs, we observe that, ceterus paribus, the effect of debt-income ratios on payment holiday adoption rates is less negative for borrowers lying in the mid implied default probability region - i.e. the marginal impact of debt-income is higher in the mid implied default probability region relative to both the low and high implied default probability regions. These results are suggestive of a structural break in payment holiday adoption behaviour driven primarily by the debt-income ratio. Table A.3 demonstrates our findings also hold for a longer 12-month default window.

As a result, our findings are consistent with high-uncertainty borrowers (possessing implied

default probabilities close to 50%) engaging in precautionary borrowing prior to P2P loan origination, with this precautionary borrowing a factor in driving precautionary payment holiday behaviour during the onset of Covid-19.

6 Conclusions

In this paper, we use a unique dataset from one of the UK's largest P2P loan providers to examine the effects of Covid-19 on the P2P loan market. After demonstrating the strong out-of-sample performance of the XGBoost machine learning model, we document a maturity effect whereby the out-of-sample predictability of short-maturity loan defaults is lower than long-maturity loans, with this maturity effect considerably stronger in the Covid sample period. Examining average monthly loan repayments across maturities, our evidence suggests that higher monthly loan repayment-to-income ratios renders short-maturity loans more susceptible to income shocks not captured in loan origination data. We provide evidence in favour of prior literature supporting the proposition that income shocks were more prevalent during the immediate Covid period, and conclude that increased sensitivity to income shocks resulted in poor default predictability for short-maturity loans during the Covid crisis.

We subsequently analyse default feature importance over time and note the relative temporal stability of our chosen set of P2P loan default factors. Total borrowing and account age are the most important predictors, with this importance maintained in both the pre-Covid and Covid sample periods. Postcode-level variables are insignificant across all sample periods. We also note that while long-maturity loan default factors are more stable relative to short-maturity loan default factors, overall feature importance rankings are congruent across maturities.

Finally, we examine Covid payment holiday adoption rates and find evidence consistent with precautionary behaviour from borrowers with the highest levels of financial uncertainty. Using a combination of logit models and prior literature findings, we show a structural break in the dependency between default risk and payment holiday adoption rates for borrowers that are

highly uncertain, and conclude that high degrees of financial uncertainty led to precautionary payment holiday uptakes by borrowers.

References

Adams-Prassl, A., Boneva, T., Golin, M., and Rauh, C. (2020). Inequality in the impact of the coronavirus shock: new survey evidence for the uk.

Alan, S., Crossley, T., and Low, H. (2012). Saving on a rainy day, borrowing for a rainy day. Technical report, working paper.

Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., and Segovia-Vargas, M.-J. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. *Ieee Access*, 8:64873–64890.

Barakova, I., Bostic, R. W., Calem, P. S., and Wachter, S. M. (2003). Does credit quality matter for homeownership? *Journal of housing Economics*, 12(4):318–336.

Bianchi, D., Büchner, M., and Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2):1046–1089.

Bianchi, D. and Shuaib, A. (2021). Sovereign credit default swaps and macroeconomic fundamentals. *Available at SSRN 3890744*.

Blöchlinger, A. and Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking & Finance*, 30(3):851–873.

Boot, A. W. (2000). Relationship banking: What do we know? *Journal of Financial Intermediation*, 9(1):7–25.

Bostic, R. W., Calem, P. S., and Wachter, S. M. (2005). Hitting the wall: Credit as an impediment to homeownership. *Building assets, building credit: creating wealth in low-income communities*, pages 155–172.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis.

Brewer, M. and Tasseva, I. (2020). Did the UK policy response to covid-19 protect household incomes? *Available at SSRN 3628464*.

Byanjankar, A., Heikkilä, M., and Mezei, J. (2015). Predicting credit risk in peer-to-peer lending: A neural network approach. In *2015 IEEE symposium series on computational intelligence*, pages 719–725. IEEE.

Cairney, J. and Boyle, M. H. (2004). Home ownership, mortgages and psychological distress. *Housing studies*, 19(2):161–174.

Capozza, D. R., Kazarian, D., and Thomson, T. A. (1997). Mortgage default in local markets. *Real Estate Economics*, 25(4):631–655.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Coomans, D. and Massart, D. L. (1982). Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-nearest neighbour classification by using alternative voting rules. *Analytica Chimica Acta*, 136:15–27.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Diamond, D. W. (1984). Financial intermediation and delegated monitoring. *Review of Economic Studies*, 51(3):393–414.

Druedahl, J. and Jørgensen, C. N. (2018). Precautionary borrowing and the credit card debt puzzle. *Quantitative Economics*, 9(2):785–823.

Dudík, M., Chen, W., Barocas, S., Inchiosa, M., Lewins, N., Oprescu, M., Qiao, J., Sameki, M., Schlener, M., Tuo, J., and Wallach, H. (2020). Assessing and mitigating unfairness in credit models with the fairlearn toolkit. Technical Report MSR-TR-2020-34, Microsoft.

EDW (2021). Monitoring the impact of covid-19: Q1 2021 rmbs report. https://eurodw.eu/wp-content/uploads/European-DataWarehouse-COVID19_Q12021_RMBS-REPORT.pdf.

Emekter, R., Tu, Y., Jirasakuldech, B., and Lu, M. (2015). Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. *Applied Economics*, 47(1):54–70.

Ennis, H. M. and Jarque, A. (2021). Bank lending in the time of covid.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81.

Fix, E. and Hodges Jr, J. L. (1952). Discriminatory analysis-nonparametric discrimination: Small sample performance. Technical report, California Univ Berkeley.

Franklin, J., Green, G., Rice-Jones, L., Venables, S., and Wukovits-Votzi, T. (2021). How have payment holidays supported mortgage borrowers during the covid crisis? *Bank Overground*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gaffney, E., Greaney, D., et al. (2020). Covid-19 payment breaks on residenfial mortgages. Technical report, Central Bank of Ireland.

Gaudêncio, J., Mazany, A., and Schwarz, C. (2019). The impact of lending standards on default rates of residential real-estate loans. *ECB Occasional Paper*, (220).

Gudivada, V., Irfan, M., Fathi, E., and Rao, D. (2016). Chapter 5 - cognitive analytics: Going beyond big data analytics and machine learning. In Gudivada, V. N., Raghavan, V. V., Govindaraju, V., and Rao, C., editors, *Cognitive Computing: Theory and Applications*, volume 35 of *Handbook of Statistics*, pages 169–205. Elsevier.

He, H. and Ma, Y. (2013). Imbalanced learning: foundations, algorithms, and applications.

Hertzberg, A., Liberman, A., and Paravisini, D. (2018). Screening on loan terms: evidence from maturity choice in consumer credit. *The Review of Financial Studies*, 31(9):3532–3567.

Holopainen, M. and Sarlin, P. (2017). Toward robust early-warning models: A horse race, ensembles and model uncertainty. *Quantitative Finance*, 17(12):1933–1963.

Jacobson, T. and Roszbach, K. (2003). Bank lending policy, credit scoring and value-at-risk. *Journal of banking & finance*, 27(4):615–633.

Jagtiani, J. and Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management*, 48(4):1009–1029.

Jin, Y. and Zhu, Y. (2015). A data-driven approach to predict default risk of loan for online peer-to-peer (p2p) lending. In *2015 Fifth International Conference on Communication Systems and Network Technologies*, pages 609–613. IEEE.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mahmoudi, N. and Duman, E. (2015). Detecting credit card fraud by modified fisher discriminant analysis. *Expert Systems with Applications*, 42(5):2510–2516.

Malekipirbazari, M. and Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10):4621–4631.

Manchester, E. (2021). Moody's - uk home loan arrears stay low after payment holidays, credit positive for covered bonds and rmbs. https://www.moodys.com/research/Moodys-UK-home-loan-arrears-stay-low-after-payment-holidays--PBS_1273351.

Mason, L., Baxter, J., Bartlett, P., and Frean, M. (1999). Boosting algorithms as gradient descent in function space. In *Proc. NIPS*, volume 12, pages 512–518.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Michels, J. (2012). Do unverifiable disclosures matter? Evidence from peer-to-peer lending. *The Accounting Review*, 87(4):1385–1413.

Mild, A., Waitz, M., and Wöckl, J. (2015). How low can you go?—overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending markets. *Journal of Business Research*, 68(6):1291–1305.

Mitchell, T. M. et al. (1997). Machine learning.

Mo, L. and Yae, J. (2022). Lending club meets zillow: local housing prices and default risk of peer-to-peer loans. *Applied Economics*, pages 1–12.

Moffatt, P. G. (2005). Hurdle models of loan default. *Journal of the operational research society*, 56(9):1063–1071.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Nigmonov, A., Shams, S., and Alam, K. (2022). Macroeconomic determinants of loan defaults: evidence from the us peer-to-peer lending market. *Research in International Business and Finance*, 59:101516.

Norden, L. and Weber, M. (2010). Credit line usage, checking account activity, and default risk of bank borrowers. *The Review of Financial Studies*, 23(10):3665–3699.

Puri, M., Rocholl, J., and Steffen, S. (2017). What do a million observations have to say about loan defaults? opening the black box of relationships. *Journal of Financial Intermediation*, 31(1):1–15.

Rosenthal, S. S. (2002). Eliminating credit barriers: how far can we go. *Low-income home-ownership*, pages 111–145.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432.

Serrano-Cinca, C. and Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (p2p) lending. *Decision Support Systems*, 89:113–122.

Serrano-Cinca, C., Gutiérrez-Nieto, B., and López-Palacios, L. (2015). Determinants of default in p2p lending. *PloS one*, 10(10):e0139427.

Sofaer, H. R., Hoeting, J. A., and Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10(4):565–577.

Tolo***c***si, L. and Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994.

Turiel, J. D. and Aste, T. (2019). P2P loan acceptance and default prediction with artificial intelligence. *arXiv preprint arXiv:1907.01800*.

Vettoretti, M. and Di Camillo, B. (2021). A variable ranking method for machine learning models with correlated features: In-silico validation and application for diabetes prediction. *Applied Sciences*, 11(16):7740.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):408–421.

Xu, B., Su, Z., and Celler, J. (2021). Evaluating default risk and loan performance in uk peer-to-peer lending: Evidence from funding circle. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 25(5):530–538.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Zheng, A. and Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists*. "O'Reilly Media, Inc.".

Table 1: **Descriptive Statistics**

This table reports descriptive statistics for all loan origination factors used in the main empirical analysis. We include a brief summary of each variable, as well as the variable shorthand code used for regressions in subsequent tables. LTM indicates variables calculated over the last twelve months prior to loan origination. CRA indicates credit rating agency summary score variables. (*) indicates variables represented in our final, reduced-dimension 10-factor model

Variable	Description	Category	Min	Max	Avg	Std
Risk Navigator Score	Risk Navigator score (CRA)	CRA	0	918	462	79
Total Credit Limit	Total credit limit	Limits	1	204,950	12,867	13,277
Prior Month Credit Limit	Credit limit over last 1 month	Limits	0	192,950	11,942	11,966
Total Revolving Limit	Total limits on active revolving accounts	Limits	0	194,950	12,485	12,729
Total Debt*	Total balances on all active accounts	Borrowing	0	5,684,424	94,593	127,998
Prior Month Total Balance	Total balance over last 1 month	Borrowing	0	5,515,891	87,125	121,435
3+ Payment Status (LTM)	# accounts with payment status of 3+ (LTM)	Borrowing	0	15	0	0
Credit Card Repayment Balance	Total credit card repayment amount	Borrowing	0	115,988	416	1,145
Current Limit to Income	% of current limits to gross annual income	Leverage	0	982	41	49
ex.Mortgage Balance to Limits*	Balance on ex.mortgage accounts (% of limits)	Leverage	0	9,991	179	447
Revolving Balance to Limits*	Balance on revolving credit accounts (% of limits)	Leverage	0	3,151	41	33
ex.Mortgage Balance to Limits (12m Ago)	Historic ex.mortgage balance (% of limits) 12m ago	Leverage	0	9,991	184	462
Secured Debt-to-Income*	Ratio of total secured debt-to-income (%)	Leverage	0	10,000	26	33
Balance-to-Limit Ratio	Total value of credit cards over the limit	Leverage	0	9	0	0
Balance-to-Limit Ratio (3m Ago)	Total value of credit cards over the limit (3m ago)	Leverage	0	19	0	0
Debt-to-Income	Debt to income ratio declared	Leverage	0	326	43	25
Revolving Balance to Limits (3m Ago)	Balances on revolving accounts (% of limits) 3m ago	Leverage	0	2,124	41	33
ex.Mortgage Balance to Limits (3m Ago)	Historic ex.mortgage balance (% of limits) 3m ago	Leverage	0	9,899	176	475
# Soft Credit Checks*	# checking credit application searches (LTM)	History	0	154	1	2
# Active Healthy Accounts	# active accounts with <1 payment status (LTM)	History	0	68	8	3
Oldest Account Age*	Age of oldest active account in months	History	1	769	171	88
# Loan Companies Used	# different short term loans companies used	History	0	16	0	0
# Loan Accounts	# short term loan accounts	History	0	97	0	3
Existing Borrower?	Prior approved Zopa loan	History	0	1	0	0
# Hard Credit Checks*	Number of credit searches (LTM)	History	0	4	1	1
Healthy Accounts (Postcode)*	Individuals with account status of 0 (index)	Postcode	0	134	100	15
Delinquent Accounts (Postcode)*	Average # of delinquencies per household (index)	Postcode	0	2,188	110	88
Sub-Healthy Accounts (Postcode)	Individuals with account status of 2 (LTM) (index)	Postcode	0	9,597	93	96
Gone Away Marker	Gone away marker	Personal	0	1	0	0
Residential Status	Residential status (non-numeric)	Personal	-	-	-	-
Borrower Age	Customer age	Personal	19	117	40	11
3m Income Change	% change in income over the past 3 months	Income	0	8,075	104	50
Annual Income	Annual income declared	Income	12,000	7,640,195	39,873	45,382
Disposable Income	Disposable income declared	Income	0	336,296	1,388	3,058
Loan Purpose	Loan purpose (non-numeric)	Loan	-	-	-	-
Loan Amount	Loan amount	Loan	250	25,000	7,393	5,594
Term*	Loan term (in months)	Loan	12	60	42	16

Table 2: **Detailed Borrower Statistics**

This table reports detailed borrower statistics for all loan origination factors used in the main empirical analysis. The aim is to provide the reader with a more in-depth view of the prototypical borrower in our sample. We observe that homeowners between the ages of 25 and 40 represent the most common demographic, with most borrowers opting for a <£10,000 loan amount. The majority of borrowers earn between £20,000 and £60,000 annually and over 30% are repeat customers. Finally, loan uses are evenly spread across home improvements, car purchases and debt consolidations

Variable	Values	# Borrowers	% Borrowers
Loan Maturity	12	34,074	8%
	24	76,475	18%
	36	94,771	22%
	48	80,500	19%
	60	148,413	34%
Loan Amount	<£5,000	162,685	37%
	£5,000-£10,000	134,680	31%
	£10,000-£15,000	80,397	19%
	£15,000-£20,000	33,883	8%
	£20,000-£25,000	22,588	5%
Loan Purpose	Home Improvements	91,902	21%
	Other	102,468	24%
	Car	110,117	25%
	Debt Consolidation	129,746	30%
Previous Borrower?	Yes	142,598	33%
	No	291,635	67%
Residential Status	Renting	72,200	17%
	Owner with Mortgage	254,144	59%
	Owner without Mortgage	24,312	6%
	Council Housing	599	0%
	Living with Parents	1,697	0%
	Living with Partner	12	0%
	Other	81,269	19%
Borrower Age	18-25	20,587	5%
	25-40	201,277	46%
	40-55	166,602	38%
	55+	45,767	11%
Borrower Income	<£20,000	38,786	9%
	£20,000-£40,000	240,001	55%
	£40,000-£60,000	100,876	23%
	£60,000+	54,570	13%

Table 3: **Out-of-Sample Model Performance over Time [45-Factor Model]**

The table below indicates out-of-sample ROC-AUC scores by quarter for each model in our chosen horse race. To begin, we highlight model performance utilising all 45 explanatory variables available in our dataset (excluding CRA scores, as discussed in the main text). In accordance with Section 3.3.2, each model is trained on 3 months of data and validated on the subsequent 3-month validation window. Once optimal model hyperparameters are obtained, the optimised model is trained on the entire 6-month train+validation data window and evaluated on the 3-month out-of sample quarter immediately following the validation window

	Pre-Covid										Covid			
Model	Dec-17	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
Logistic Regression	0.747	0.761	0.776	0.767	0.780	0.748	0.740	0.734	0.757	0.740	0.673	0.659	0.676	0.679
Random Forest	0.791	0.809	0.779	0.771	0.792	0.762	0.758	0.754	0.783	0.764	0.677	0.711	0.695	0.703
XGBoost	0.803	0.814	0.777	0.783	0.809	0.771	0.777	0.756	0.784	0.772	0.729	0.682	0.688	0.706
kNN	0.690	0.703	0.713	0.685	0.715	0.705	0.707	0.701	0.702	0.699	0.616	0.595	0.591	0.596
Naïve Bayes	0.746	0.77	0.734	0.725	0.746	0.707	0.707	0.699	0.726	0.714	0.635	0.654	0.632	0.676
Neural Network	0.740	0.750	0.734	0.725	0.724	0.721	0.741	0.736	0.748	0.748	0.670	0.692	0.647	0.634

Table 4: **CRA Score Regressions on Default Factors**

This table reports results from regressions of the Equifax Risk Navigator Score (RNS) on default factors used in our 10-factor and 45-factor credit models. Given the RNS is a single CRA-derived score aimed at summarising borrower credit risk across several hundred borrower-level factors, we view the R^2 from a regression of the RNS score on a set of default factors as a proxy for the degree to which our chosen default factors capture borrower credit risk. (**) indicates significance at 5% level

Variable	45-Factor	10-Factor
Total Credit Limit	0.001**	
Prior Month Credit Limit	-0.002**	
Prior Month Total Balance	0.000**	
Current Limit to Income	-0.064**	
# Soft Credit Checks	-1.384**	-1.734**
Debt-to-Income	0.091**	
# Active Healthy Accounts	-0.431**	
Total Debt	0.000**	0.000**
ex.Mortgage Balance to Limits	-0.004**	-0.005**
Revolving Balance to Limits	-0.367**	-1.025**
Total Revolving Limit	0.001**	
3+ Payment Status (LTM)	-32.57**	
Oldest Account Age	0.028**	0.074**
# Loan Companies Used	-2.460**	
# Loan Accounts	0.023	
Healthy Accounts (Postcode)	0.307**	0.355**
Revolving Balance to Limits (3m Ago)	-0.234**	
ex.Mortgage Balance to Limits (3m Ago)	-0.001	
Delinquent Accounts (Postcode)	-0.072**	-0.088**
ex.Mortgage Balance to Limits (12m Ago)	-0.001**	
Credit Card Repayment Balance	-0.001**	
Sub-Healthy Accounts (Postcode)	-0.017**	
Gone Away Marker	6.814**	
Secured Debt-to-Income	0.090**	0.191**
Balance-to-Limit Ratio	-28.02**	
Balance-to-Limit Ratio (3m Ago)	-19.26**	
3m Income Change	-0.005**	
Annual Income_declared	0.000**	
Disposable Income_declared	0.000**	
Loan Amount	0.000**	
Term	0.260**	0.574**
Existing Borrower?	8.012**	
Borrower Age	0.097**	
# Hard Credit Checks	-17.52**	-18.66**
Residential Status Council Housing	34.45**	
Residential Status Living with Parents	34.99**	
Residential Status Living with Partner	55.47**	
Residential Status Owner (No Mortgage)	61.73**	
Residential Status Owner (With Mortgage)	65.74**	
Residential Status Other	38.33**	
Residential Status Renting	39.15**	
Loan Purpose Car	85.60**	
Loan Purpose Consolidate existing debts	83.55**	
Loan Purpose Home improvements	80.94**	
Loan Purpose Other	79.77**	
n	434,233	434,233
R^2	49.6%	42.7%

Table 5: **Out-of-Sample Model Performance over Time [10-Factor Model]**

Using recursive feature elimination (RFE), we eliminate all factors from our initial 45-factor model that do not result in a ROC-AUC drop when removed from our model during the first training/validation period. The 10 remaining factors constitute our explanatory variable set used for all future sample periods. The table below indicates out-of-sample ROC-AUC scores by quarter for each 10-factor model highlighted in Section 3.2. As with Table 3, each model is trained on 3 months of data and validated on the subsequent 3-month validation window. Once optimal model hyperparameters are obtained, the optimised model is trained on the entire 6-month train+validation data window and evaluated on the 3-month out-of sample quarter immediately following the validation window. We have highlighted in bold the XGBoost model, which is the best-performing out-of-sample model chosen for all subsequent analyses

	Pre-Covid										Covid			
Model	Dec-17	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
Logistic Regression	0.766	0.798	0.775	0.760	0.769	0.744	0.744	0.746	0.753	0.740	0.679	0.674	0.681	0.649
Random Forest	0.803	0.806	0.795	0.780	0.801	0.766	0.761	0.760	0.782	0.768	0.699	0.719	0.689	0.687
XGBoost*	**0.809**	**0.807**	**0.792**	**0.779**	**0.803**	**0.771**	**0.768**	**0.759**	**0.788**	**0.776**	**0.706**	**0.707**	**0.692**	**0.695**
kNN	0.661	0.639	0.665	0.686	0.694	0.677	0.676	0.655	0.660	0.657	0.594	0.575	0.595	0.578
Naïve Bayes	0.793	0.793	0.782	0.768	0.779	0.761	0.753	0.744	0.780	0.759	0.686	0.704	0.701	0.704
Neural Network	0.791	0.797	0.794	0.770	0.798	0.763	0.768	0.753	0.785	0.760	0.709	0.663	0.677	0.683

Table 6: **Comparison of Means t-Statistics**

This table reports the results of pairwise comparison of means t-statistics for ROC-AUC scores across a variety of models during both the pre-Covid and Covid period. (**) indicates a rejection of the null (H_0: the pairs have the same performance) at the standard 5% confidence level. Negative t-statistics indicate outperformance of the column model, while positive t-statistics indicate outperformance of the row model

Panel A: Overall Sample

Models	XGBoost	Random Forest	Neural-Net	Naive Bayes	kNN	Logistic Regression
XGBoost		0.615	8.740**	3.950**	60.623**	14.304**
Random Forest			8.715**	3.598**	64.229**	14.762**
Neural Network				-4.677**	52.683**	5.094**
Naïve Bayes					56.224**	9.942**
kNN						-51.119**
Logistic Regression						

Panel B: Pre-Covid

Models	XGBoost	Random Forest	Neural-Net	Naive Bayes	kNN	Logistic Regression
XGBoost		0.432	8.333**	7.754**	67.689**	15.612**
Random Forest			9.071**	8.470**	70.140**	16.615**
Neural Network				-0.575	59.520**	7.313**
Naïve Bayes					60.065**	7.883**
kNN						-52.159**
Logistic Regression						

Panel C: Covid

Models	XGBoost	Random Forest	Neural-Net	Naive Bayes	kNN	Logistic Regression
XGBoost		2.175**	6.651**	0.690	26.810**	7.920**
Random Forest			4.262**	-1.298	22.783**	5.243**
Neural Network				-5.356**	17.684**	0.665
Naïve Bayes					23.092**	6.349**
kNN						-18.300**
Logistic Regression						

Table 7: **XGBoost Model Performance: The Impact of Risk Navigator Scores**

To demonstrate the ability of our reduced-dimension 10-factor model to adequately capture borrower risk, we show below the out-of-sample performance of our XGBoost model including and excluding the Equifax Risk Navigator Score (RNS). Given this score is intended to condense hundreds of borrower-level characteristics into a single risk metric, the table demonstrates the negligible change in out-of-sample performance when including the Equifax RNS in our 10-factor model; across all maturities, RNS inclusion improves ROC-AUC by an average of only 1.6%

Panel A: XGBoost Out-of-Sample Performance by Maturity (Without Risk Navigator Score)

	Pre-Covid										Covid			
Maturity	Dec-17	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
All Maturities	0.809	0.807	0.792	0.779	0.803	0.771	0.768	0.759	0.788	0.776	0.706	0.707	0.692	0.695
1-2 years	0.798	0.755	0.729	0.751	0.796	0.727	0.705	0.727	0.709	0.698	0.607	0.588	0.659	0.584
4-5 years	0.816	0.820	0.831	0.793	0.796	0.783	0.795	0.780	0.820	0.809	0.776	0.756	0.698	0.781

Panel B: XGBoost Out-of-Sample Performance by Maturity (With Risk Navigator Score)

	Pre-Covid										Covid			
Maturity	Dec-17	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
All Maturities	0.825	0.831	0.814	0.802	0.823	0.782	0.782	0.771	0.791	0.794	0.717	0.717	0.720	0.709
1-2 years	0.834	0.797	0.726	0.774	0.817	0.764	0.719	0.741	0.718	0.733	0.649	0.604	0.696	0.616
4-5 years	0.823	0.845	0.846	0.805	0.819	0.807	0.810	0.792	0.84	0.818	0.758	0.772	0.727	0.785

Table 8: **Default Factor Rankings over Time**

The panels below indicate default factor rankings over time for each variable in our 10-factor model, firstly focusing on all maturities and subsequently looking at short/long-maturity loans. All rankings were initially determined via permutation feature importance, before being normalised to lie in the interval [0,1]

Panel A: Factor Rankings (All Maturities)

	Pre-Covid										Covid			
Factor	Dec-17	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
Total Debt	1.00	1.00	1.00	1.00	0.86	0.87	1.00	1.00	1.00	1.00	1.00	0.81	1.00	0.67
Secured Debt-to-Income	0.00	0.08	0.08	0.21	0.31	0.43	0.06	0.08	0.19	0.04	0.09	0.56	0.53	1.00
# Soft Credit Checks	0.16	0.14	0.31	0.37	1.00	0.55	0.54	0.32	0.24	0.18	0.18	0.21	0.32	0.30
# Hard Credit Checks	0.05	0.06	0.16	0.24	0.21	0.23	0.21	0.12	0.12	0.07	0.10	0.06	0.08	0.07
ex.Mortgage Balance to Limits	0.14	0.18	0.23	0.24	0.24	0.16	0.12	0.10	0.11	0.15	0.06	0.04	0.00	0.11
Revolving Balance to Limits	0.07	0.10	0.12	0.04	0.13	0.23	0.21	0.06	0.11	0.07	0.10	0.01	0.09	0.11
Term	0.06	0.15	0.16	0.25	0.48	0.17	0.09	0.07	0.01	0.07	0.15	0.17	0.11	0.00
Oldest Account Age	0.35	0.39	0.43	0.65	0.96	1.00	0.82	0.49	0.59	0.48	0.68	1.00	0.87	0.56
Delinquent Accounts (Postcode)	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.12
Healthy Accounts (Postcode)	0.03	0.00	0.07	0.13	0.12	0.10	0.04	0.03	0.03	0.04	0.03	0.01	0.11	0.06

Panel B: Factor Rankings (1-2 year Maturities)

	Pre-Covid										Covid			
Factor	Dec-17	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
Total Debt	1.00	1.00	1.00	1.00	0.13	0.44	0.82	1.00	1.00	1.00	0.47	0.94	1.00	1.00
Secured Debt-to-Income	0.05	0.17	0.21	0.47	0.19	0.24	0.16	0.06	0.31	0.00	0.23	0.66	0.67	0.58
# Soft Credit Checks	0.47	0.29	0.48	0.39	1.00	0.99	1.00	0.47	0.65	0.54	0.38	0.39	0.00	0.33
# Hard Credit Checks	0.20	0.17	0.21	0.32	0.29	0.33	0.49	0.24	0.43	0.12	0.71	0.20	0.16	0.24
ex.Mortgage Balance to Limits	0.26	0.47	0.54	0.66	0.57	0.19	0.21	0.11	0.17	0.21	0.00	0.09	0.06	0.36
Revolving Balance to Limits	0.04	0.12	0.15	0.01	0.14	0.52	0.58	0.03	0.22	0.24	1.00	0.32	0.18	0.02
Term	0.00	0.42	0.00	0.31	0.41	0.66	0.41	0.22	0.00	0.25	0.35	0.32	0.06	0.00
Oldest Account Age	0.55	0.57	0.38	0.92	0.34	1.00	0.83	0.37	0.62	0.58	0.00	1.00	0.04	0.68
Delinquent Accounts (Postcode)	0.01	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.09	0.01	0.18	0.09	0.06	0.10
Healthy Accounts (Postcode)	0.01	0.07	0.24	0.02	0.08	0.07	0.01	0.00	0.16	0.04	0.58	0.00	0.23	0.37

Panel C: Factor Rankings (4-5 year Maturities)

	Pre-Covid										Covid			
Factor	Dec-17	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
Total Debt	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.94	0.40
Secured Debt-to-Income	0.00	0.09	0.07	0.11	0.25	0.42	0.06	0.13	0.15	0.10	0.05	0.63	0.42	1.00
# Soft Credit Checks	0.10	0.09	0.20	0.38	0.53	0.56	0.23	0.26	0.16	0.09	0.28	0.18	0.36	0.33
# Hard Credit Checks	0.02	0.01	0.15	0.31	0.18	0.19	0.07	0.08	0.04	0.10	0.08	0.14	0.16	0.00
ex.Mortgage Balance to Limits	0.17	0.12	0.17	0.17	0.05	0.13	0.06	0.09	0.13	0.15	0.13	0.04	0.01	0.04
Revolving Balance to Limits	0.03	0.06	0.06	0.06	0.18	0.08	0.10	0.08	0.09	0.00	0.00	0.04	0.00	0.12
Term	0.07	0.07	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.02	0.05	0.09
Oldest Account Age	0.36	0.36	0.50	0.60	0.96	0.87	0.60	0.52	0.68	0.39	0.69	1.00	1.00	0.45
Delinquent Accounts (Postcode)	0.03	0.02	0.00	0.00	0.01	0.02	0.01	0.04	0.00	0.01	0.03	0.02	0.04	0.06
Healthy Accounts (Postcode)	0.07	0.00	0.05	0.17	0.15	0.15	0.08	0.07	0.02	0.02	0.03	0.00	0.06	0.15

Table 9: **Stability over Train/Validation/Test Window Permutations**

The below panels demonstrate the consistency of both out-of-sample model ROC-AUC performance and default factor rankings over time for a multitude of train, validation and test windows. As with our prior strategy, training + validation windows are used for hyperparameter optimisation while test data is used for out-of-sample evaluation of our calibrated 10-factor model

Panel A: XGBoost Out-of-Sample Performance [6m train, 3m validation, 3m test]

	Pre-Covid									Covid			
Maturity	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
All Maturities	0.803	0.803	0.785	0.796	0.767	0.763	0.757	0.780	0.777	0.705	0.722	0.713	0.694
1-2 Years	0.776	0.726	0.751	0.787	0.745	0.702	0.726	0.702	0.700	0.599	0.632	0.671	0.592
4-5 Years	0.815	0.836	0.792	0.793	0.795	0.793	0.78	0.821	0.806	0.782	0.755	0.728	0.771

Panel B: Default Factor Rankings [6m train, 3m validation, 3m test]

	Pre-Covid									Covid			
Factor	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
Total Debt	1.00	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.84
Secured Debt-to-Income	0.03	0.07	0.09	0.12	0.55	0.02	0.16	0.12	0.05	0.03	0.19	0.08	1.00
# Soft Credit Checks	0.16	0.30	0.41	0.67	0.67	0.49	0.38	0.16	0.19	0.24	0.16	0.14	0.21
# Hard Credit Checks	0.08	0.14	0.11	0.17	0.28	0.12	0.20	0.07	0.14	0.05	0.02	0.01	0.00
ex.Mortgage Balance to Limits	0.14	0.19	0.12	0.16	0.23	0.08	0.13	0.12	0.12	0.05	0.02	0.03	0.09
Revolving Balance to Limits	0.05	0.10	0.06	0.14	0.27	0.18	0.15	0.08	0.07	0.04	0.01	0.00	0.05
Term	0.05	0.10	0.13	0.37	0.20	0.09	0.20	0.01	0.03	0.07	0.22	0.11	0.12
Oldest Account Age	0.20	0.43	0.49	0.64	1.00	0.65	0.78	0.43	0.54	0.51	0.90	0.35	0.42
Delinquent Accounts (Postcode)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10
Healthy Accounts (Postcode)	0.01	0.01	0.09	0.11	0.11	0.05	0.05	0.04	0.02	0.01	0.06	0.02	0.10

Panel C: XGBoost Out-of-Sample Performance [6m train, 6m validation, 6m test]

	Pre-Covid				Covid	
Maturity	Sep-18	Mar-19	Sep-19	Mar-20	Sep-20	Mar-21
All Maturities	0.791	0.781	0.757	0.778	0.722	0.709
1-2 Years	0.747	0.763	0.716	0.708	0.644	0.634
4-5 Years	0.812	0.794	0.779	0.810	0.762	0.752

Panel D: Default Factor Rankings [6m train, 6m validation, 6m test]

	Pre-Covid				Covid	
Factor	Sep-18	Mar-19	Sep-19	Mar-20	Sep-20	Mar-21
Total Debt	1.00	1.00	1.00	1.00	1.00	1.00
Secured Debt-to-Income	0.04	0.03	0.07	0.10	0.07	0.11
# Soft Credit Checks	0.27	0.53	0.42	0.22	0.23	0.18
# Hard Credit Checks	0.10	0.09	0.16	0.13	0.09	0.08
ex.Mortgage Balance to Limits	0.12	0.10	0.10	0.13	0.00	0.04
Revolving Balance to Limits	0.03	0.11	0.09	0.06	0.03	0.00
Term	0.04	0.17	0.08	0.08	0.09	0.07
Oldest Account Age	0.31	0.49	0.46	0.58	0.52	0.48
Delinquent Accounts (Postcode)	0.00	0.00	0.00	0.00	0.03	0.01
Healthy Accounts (Postcode)	0.01	0.05	0.02	0.05	0.05	0.05

Table 10: **Payment Holiday Piecewise Logit Model**

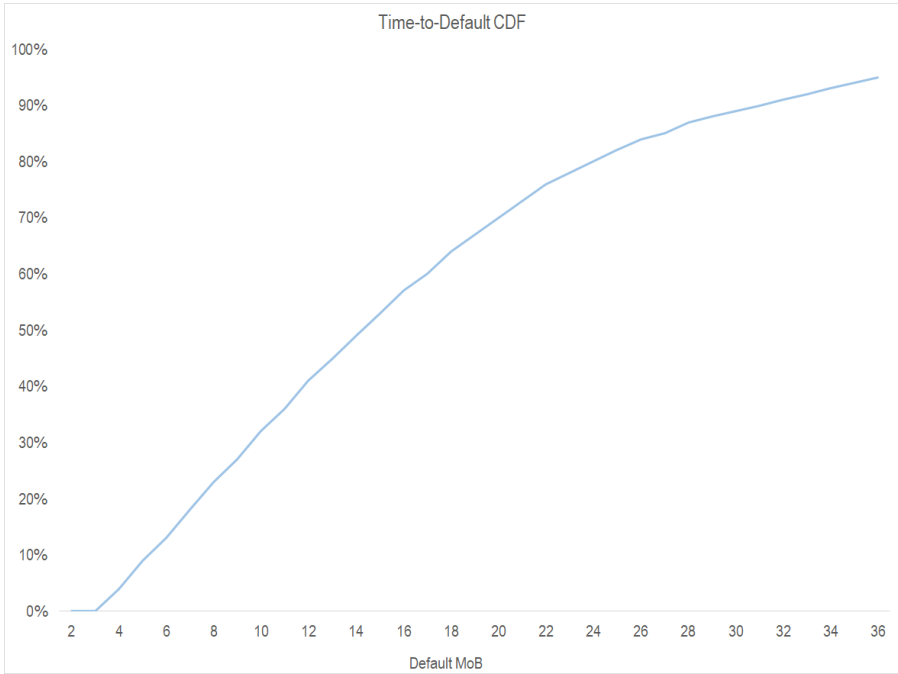
This table reports the results of our payment holiday piecewise-logit models. Two dummy variables are introduced to examine a potential structural break - *mid_dummy* corresponds to all borrowers with an XGBoost model-implied 9-month default probability (IDP) of 40-60% at loan origination, while *high_dummy* corresponds to borrowers with a model-implied 9-month default probability of >=60% at loan origination. We report coefficients and t-statistics for a variety of explanatory variable × IDP dummy interactions.

	Logit Model Results			
Variable	(1)	(2)	(3)	(4)
Intercept	-2.474**	-2.472**	-2.475**	-2.476**
	(-21.94)	(-21.94)	(-21.94)	(-21.94)
# Soft Credit Checks	0.079**	0.050	0.051	0.078**
	(9.40)	(1.90)	(1.86)	(9.37)
Term	0.007**	0.007**	0.007**	0.007**
	(9.46)	(9.51)	(9.53)	(9.49)
# Hard Credit Checks	0.105**	0.105**	0.105**	0.105**
	(9.48)	(9.53)	(9.52)	(9.46)
ex.Mortgage Balance to Limits	0.000**	0.000**	0.000	0.000
	(4.18)	(4.10)	(1.37)	(1.26)
Revolving Balance to Limits	0.008**	0.008**	0.008**	0.008**
	(17.59)	(17.41)	(17.39)	(17.56)
Oldest Account Age	-0.002**	-0.002**	-0.002**	-0.002**
	(-14.78)	(-14.73)	(-14.64)	(-14.69)
Delinquent Accounts (Postcode)	0.001**	0.001**	0.001**	0.001**
	(4.42)	(4.40)	(4.40)	(4.42)
Healthy Accounts (Postcode)	-0.003**	-0.003**	-0.003**	-0.003**
	(-2.84)	(-2.83)	(-2.83)	(-2.85)
Secured Debt-to-Income	-0.007**	-0.007**	-0.007**	-0.007**
	(-10.39)	(-9.16)	(-8.82)	(-9.93)
ex.Mortgage Balance to Limits ×*mid_dummy*			0.000	0.000
			(1.14)	(1.38)
ex.Mortgage Balance to Limits ×*high_dummy*			0.000	0.000
			(0.64)	(0.44)
# Soft Credit Checks ×*mid_dummy*		0.041	0.039	
		(1.48)	(1.37)	
# Soft Credit Checks ×*high_dummy*		0.013	0.012	
		(0.44)	(0.39)	
Secured Debt-to-Income ×*mid_dummy*	0.004**	0.003**	0.003**	0.003**
	(4.83)	(3.43)	(3.13)	(4.32)
Secured Debt-to-Income ×*high_dummy*	-0.003	-0.001	-0.001	-0.003
	(-0.97)	(-0.25)	(-0.30)	(-0.89)
n	94,865	94,865	94,865	94,865

Figure 1: **Time-to-Default CDF Curves**

Panel A captures the proportion of all defaults occurring on or prior to a given post-origination month on the x-axis. 98% of all defaults occur withing 36-months of loan origination, whilst 27% of all defaults occur within our chosen 9-month default window. Panel B shows default breakdown my loan maturity, and demonstrates that a shorter default window does not bias the analysis against long-maturity loans

Panel A: All Maturities



Panel B: Individual Maturities

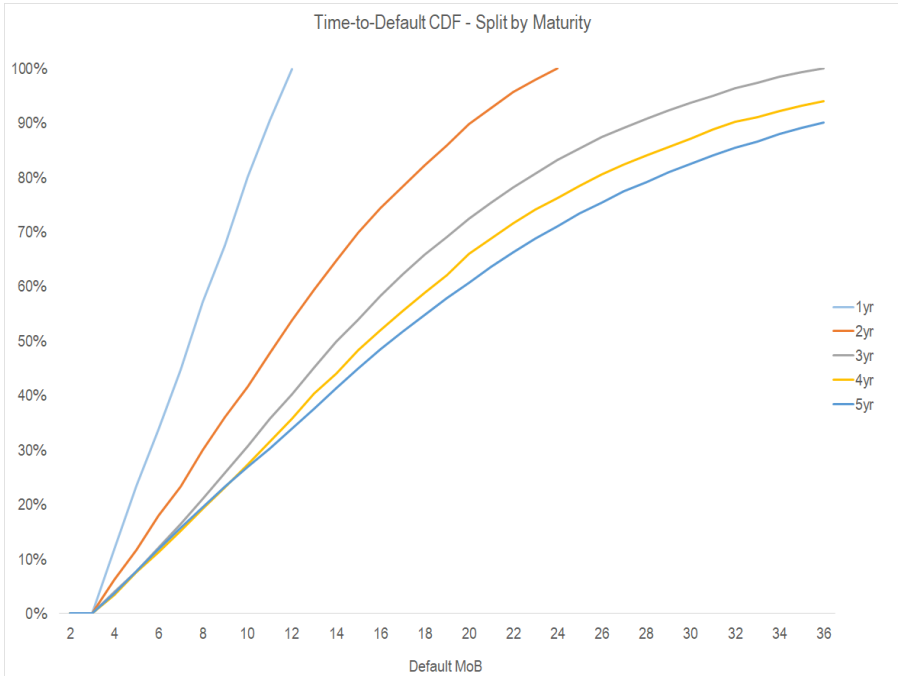


Figure 2: **Default Curves by Maturity**

This figure shows 9-month default rates by loan vintage. "All Maturies" covers all loan maturities in our sample, i.e. 1-5 years. Long-term loans represent 4-5 year maturities, whilst short-term loans represent 1-2 year maturities. "Intermediate" represent 3-year maturity loans. Default refers to a borrower missing 3 or more consecutive monthly repayments within 9 months of loan origination

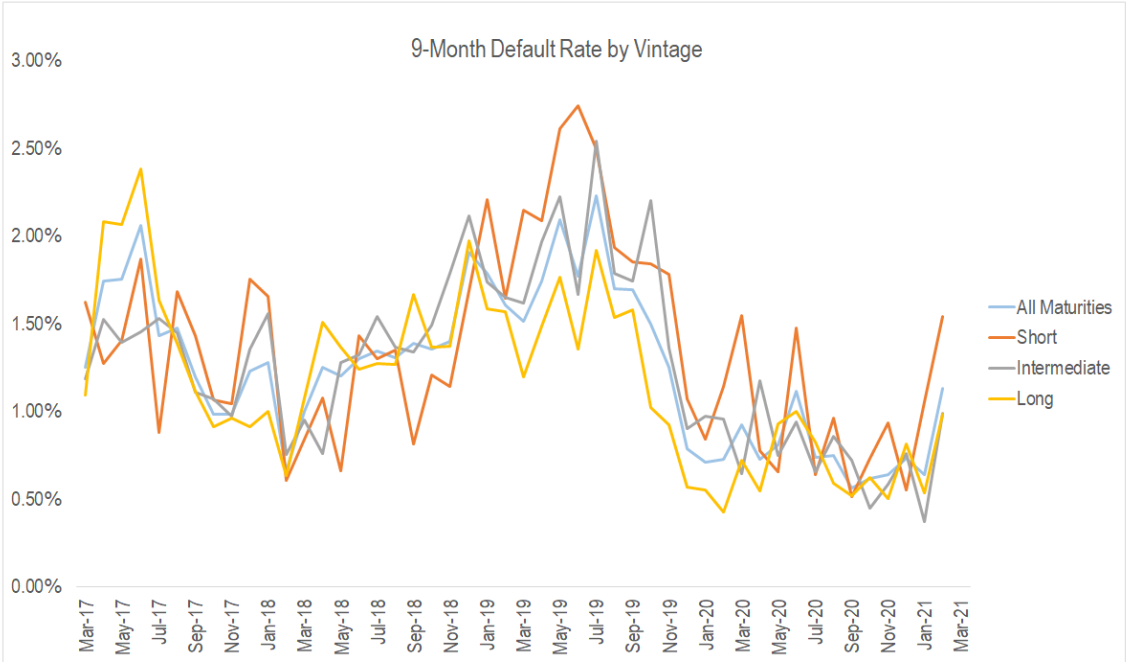


Figure 3: **Default Factor Histograms**

This figure highlights histograms for each of the features in our final 10-factor model.

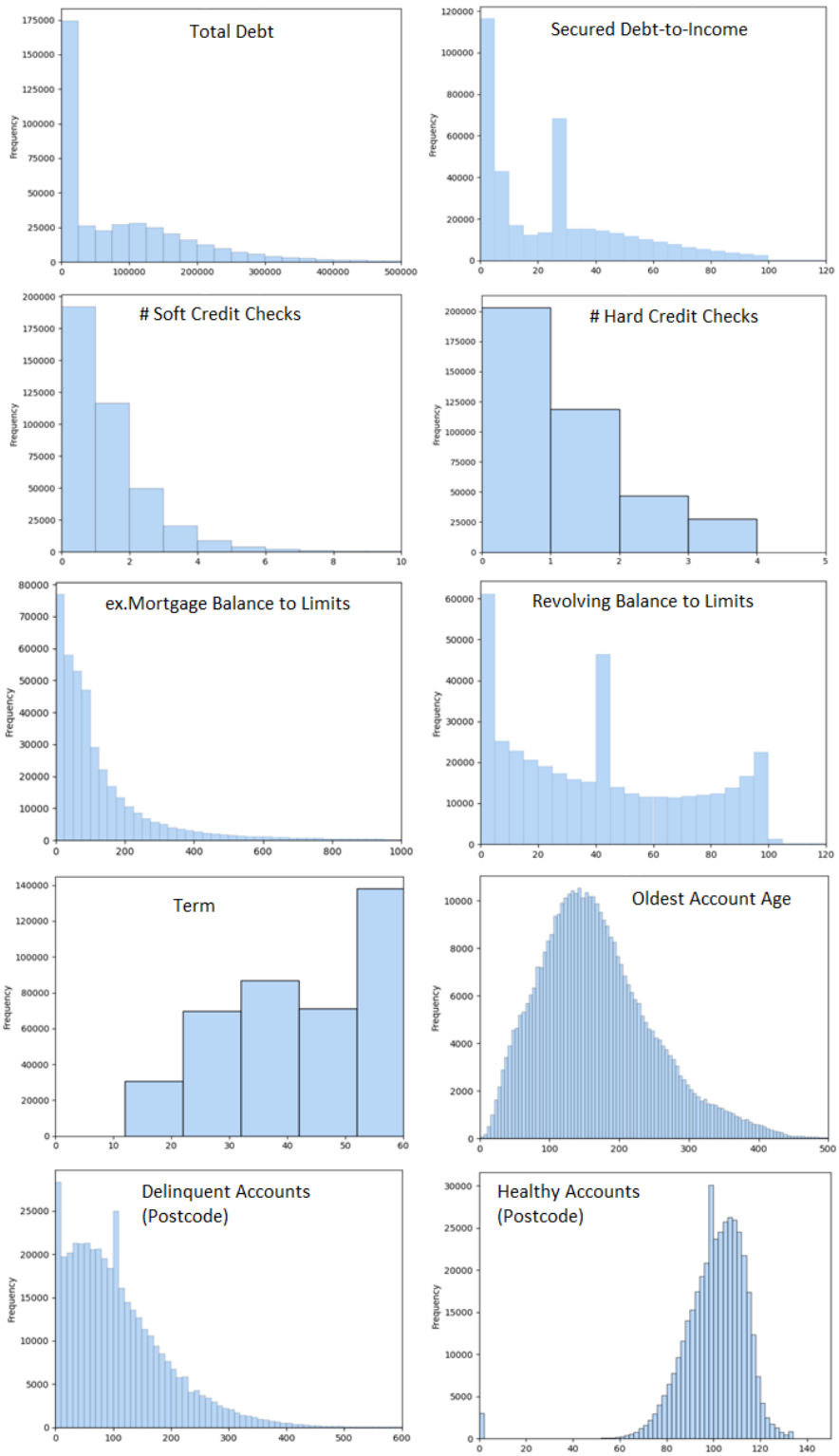


Figure 4: **Default Factor Box Plots**

The figure below highlights box plots for each of the features in our final 10-factor model. Quarter-by-quarter plots allow us to capture time-series variation in default factor value spreads at loan origination

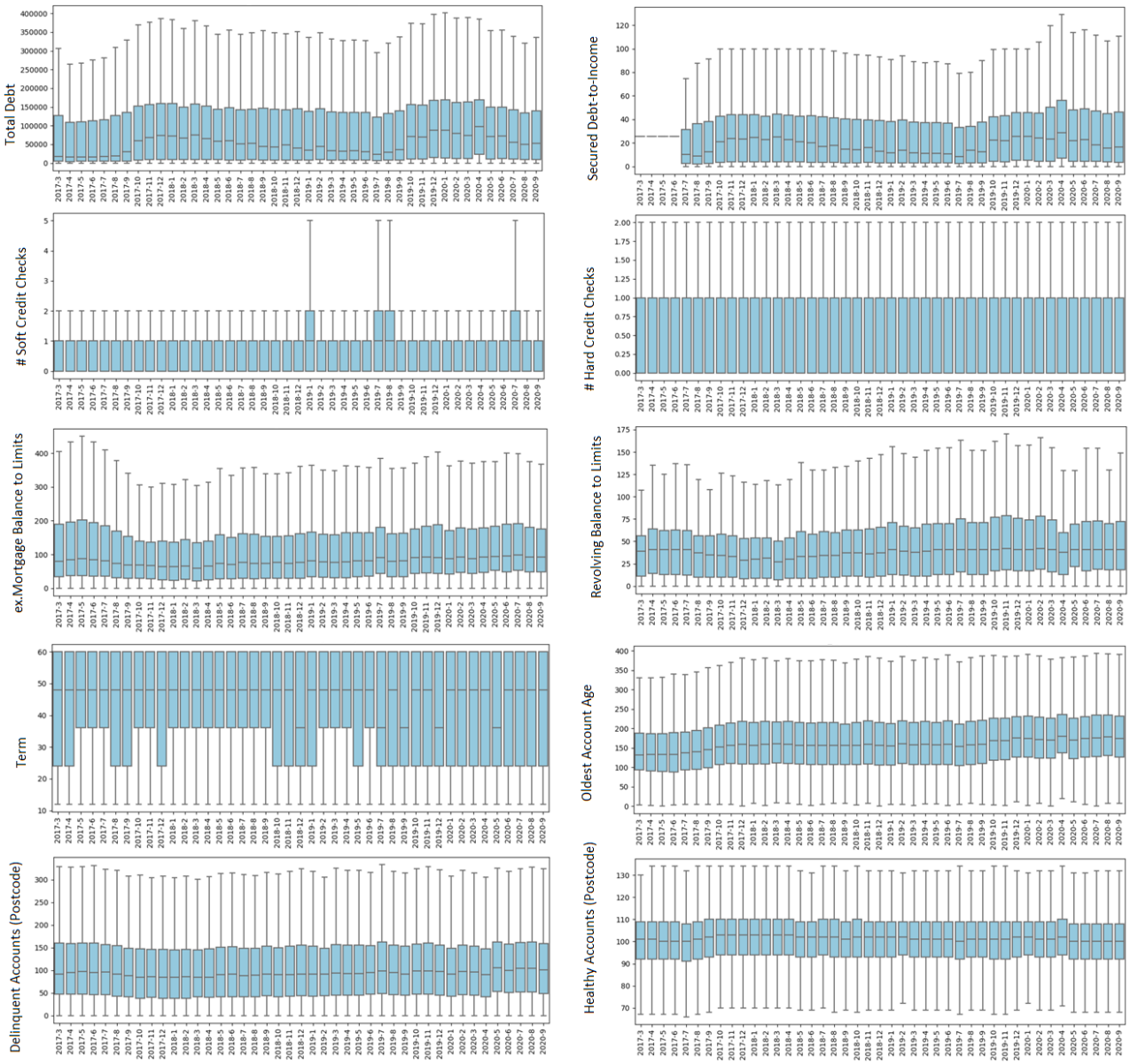


Figure 5: **Explanatory Variable Correlations**

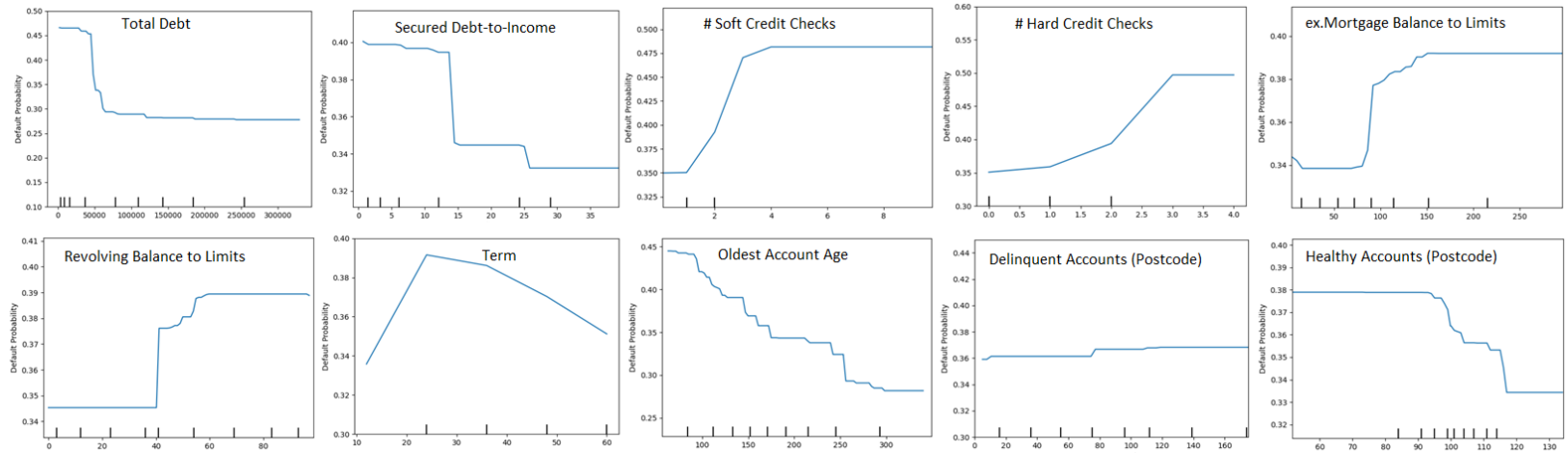
This figure provides a view on the cross-correlations of all explanatory variables in our final 10-factor default model. As highlighted in Section 4.1.1, low correlation is imperative when determining the individual importance rank of each standalone feature

Variable	Total Debt	Secured Debt-to-Income	# Soft Credit Checks	# Hard Credit Checks	ex.Mortgage Balance to Limits	Revolving Balance to Limits	Term	Oldest Account Age	Healthy Accounts (Postcode)	Delinquent Accounts (Postcode)
Total Debt	1.00	0.47	0.01	0.01	0.00	-0.07	0.16	0.15	-0.19	0.16
Secured Debt-to-Income	0.47	1.00	-0.02	-0.02	0.01	-0.07	0.13	0.10	-0.10	0.08
# Soft Credit Checks	0.01	-0.02	1.00	0.22	0.07	0.09	-0.08	-0.09	0.01	-0.02
# Hard Credit Checks	0.01	-0.02	0.22	1.00	0.04	0.08	-0.06	-0.14	0.02	-0.02
ex.Mortgage Balance to Limits	0.00	0.01	0.07	0.04	1.00	0.10	-0.03	-0.06	0.02	-0.02
Revolving Balance to Limits	-0.07	-0.07	0.09	0.08	0.10	1.00	-0.11	-0.10	0.04	-0.05
Term	0.16	0.13	-0.08	-0.06	-0.03	-0.11	1.00	0.11	-0.07	0.09
Oldest Account Age	0.15	0.10	-0.09	-0.14	-0.06	-0.10	0.11	1.00	-0.10	0.10
Healthy Accounts (Postcode)	-0.19	-0.10	0.01	0.02	0.02	0.04	-0.07	-0.10	1.00	-0.57
Delinquent Accounts (Postcode)	0.16	0.08	-0.02	-0.02	-0.02	-0.05	0.09	0.10	-0.57	1.00

Figure 6: **Partial Dependency Plots**

The panels below present partial dependency plots (PDPs) for the in-sample 6-month pre-Covid and 6-month Covid periods. PDPs aim to capture the ceterus-paribus effect of a given explanatory variable on the target variable, and allow us to infer visually the extent to which the causal relationship between an input and target variable is linear/quadratic in nature.

Panel A: Pre-Covid



Panel B: Covid period

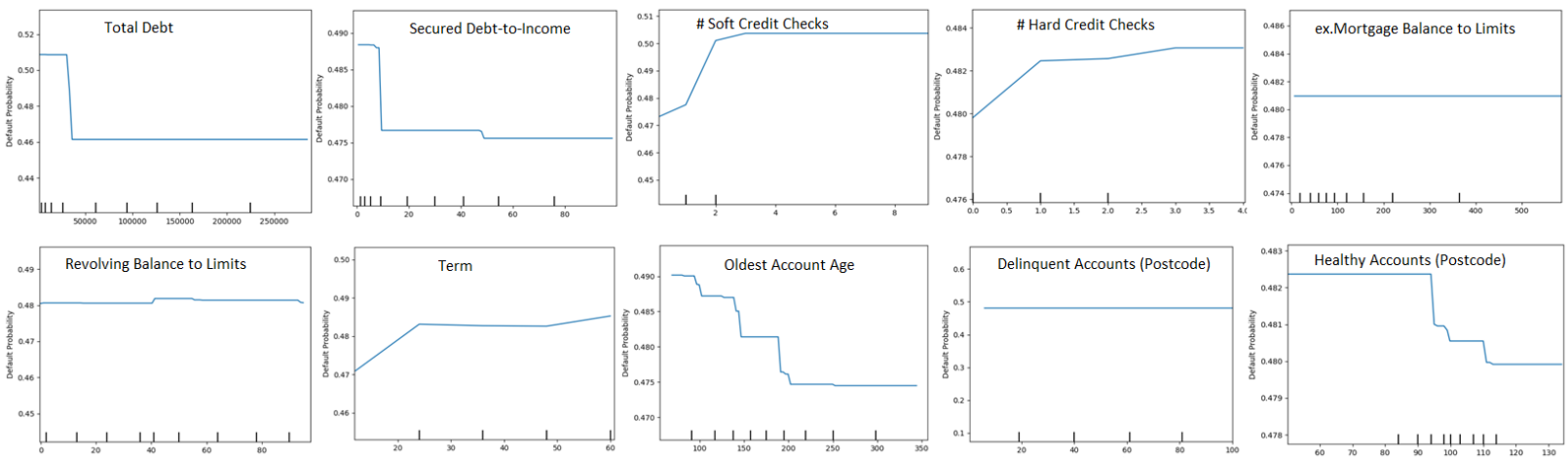


Figure 7: **Payment Holiday Adoption Rates by Maturity**

The figure below shows 9-month payment holiday adoption rates by loan vintage. We include pre-Covid vintages to explicitly highlight the dates during which payment holidays are introduced and utilised. Payment holiday "adoption rates" refer to the proportion of borrowers in a given vintage month who were granted a payment holiday within 9 months of loan origination. Given our default prediction window is 9 months, we choose a 9-month payment holiday window to allow a fair, "apples-to-apples" comparison between implied default probabilities and payment holiday adoption rates

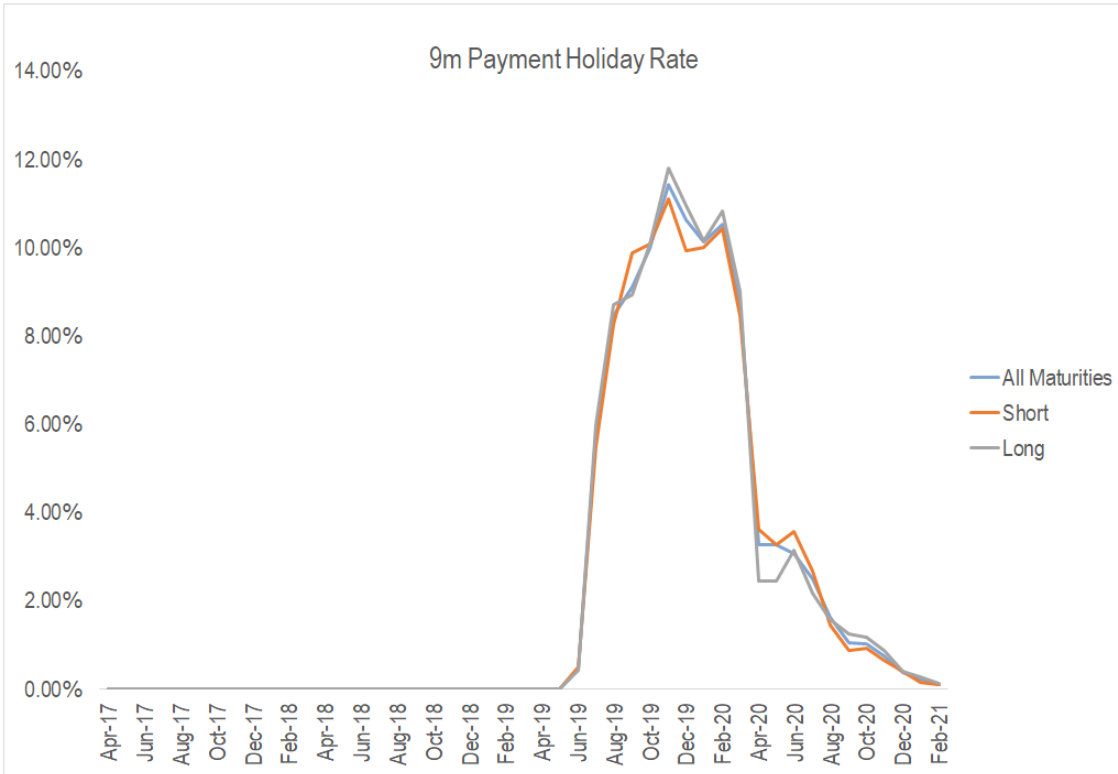


Figure 8: **Payment Holiday Adoption Rates vs Implied Default Probability**

The figure below captures payment holiday adoption rates for varying levels of borrower implied default probability (IDP). Given the variation and granularity in IDPs, we choose to bucket customers by a variety of IDP increments from 3% to 5% and calculate payment holiday adoption rates for borrowers in each IDP bucket

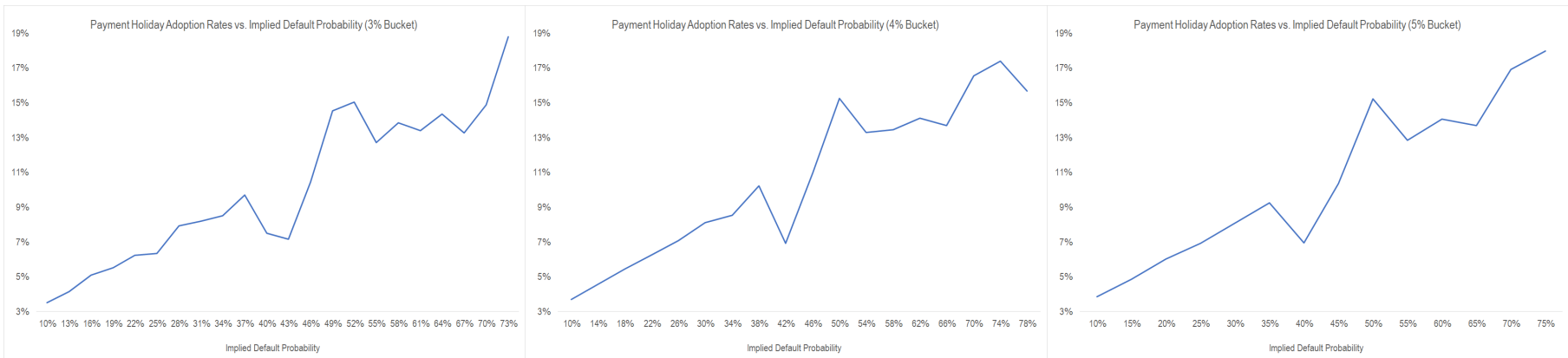
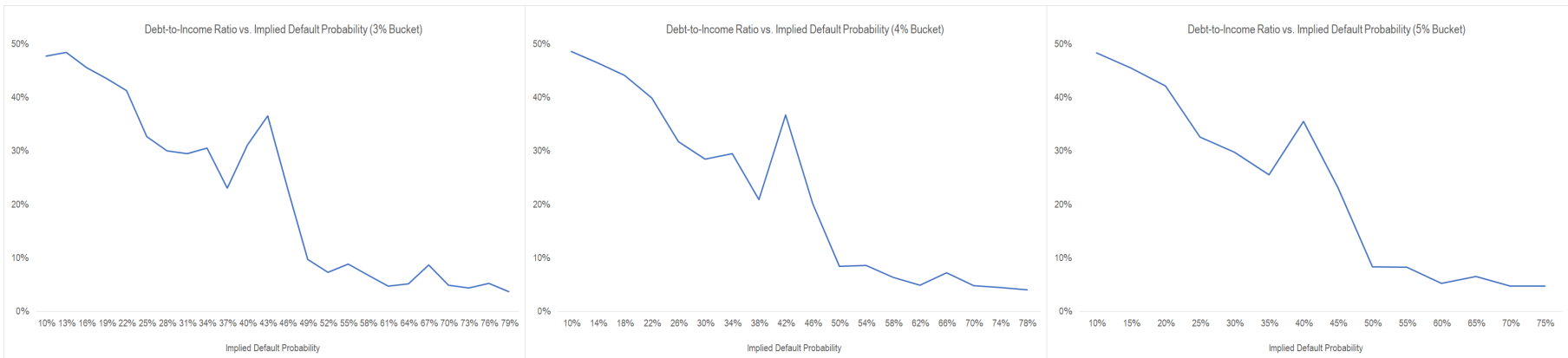


Figure 9: **Debt-to-Income Ratios vs. Implied Default Probability**

The figure below captures average borrower debt-to-income levels for varying levels of borrower implied default probabilities (IDP). Given the variation and granularity in IDPs, we bucket customers by a variety of IDP increments from 3% to 5% and calculate average debt-to-income rates at loan origination for borrowers in each IDP bucket



A Additional Tables

Table A.1: **Default Factor Rankings (Alternative Default Windows)**

The panels below indicate default factor rankings over time for each variable in our 10-factor model, with a focus on 6-month and 12-month default window sizes. Our goal is to demonstrate that factor stability over time is not affected by our choice of default window. All rankings were initially determined via permutation feature importance, before being normalised to lie in the interval [0,1]. As with Table 8, we observe *TotalDebt* to be the most significant default factor for the majority of out-of-sample quarters. *OldestAccountAge* remains the second-most important factor across both pre-Covid and Covid periods. Finally, both postcode variables remain the least important variables. As with our main 9-month default window analysis, we observe a decrease in the importance of unsecured borrowing factors *ex.MortgageBalancetoLimits* for both 12-month and 6-month default windows as we enter the Covid period.

Panel A: Factor Rankings - All Loan Maturities (12m Default Window)

	Pre-Covid										Covid			
Factor	Dec-17	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
Total Debt	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.76	0.41
Secured Debt-to-Income	0.00	0.13	0.13	0.02	0.09	0.46	0.12	0.10	0.22	0.05	0.07	0.18	0.18	1.00
# Soft Credit Checks	0.22	0.24	0.36	0.42	0.58	0.52	0.41	0.33	0.26	0.24	0.20	0.17	0.37	0.34
# Hard Credit Checks	0.07	0.04	0.17	0.20	0.13	0.26	0.20	0.13	0.11	0.09	0.06	0.02	0.14	0.10
ex.Mortgage Balance to Limits	0.24	0.45	0.21	0.33	0.30	0.24	0.19	0.18	0.14	0.19	0.14	0.00	0.05	0.07
Revolving Balance to Limits	0.11	0.20	0.22	0.20	0.23	0.49	0.30	0.11	0.18	0.16	0.59	0.11	0.24	0.13
Term	0.06	0.11	0.11	0.27	0.30	0.17	0.05	0.05	0.00	0.09	0.19	0.03	0.15	0.00
Oldest Account Age	0.28	0.52	0.46	0.42	0.50	0.72	0.61	0.46	0.64	0.47	1.00	0.62	1.00	0.45
Delinquent Accounts (Postcode)	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.05
Healthy Accounts (Postcode)	0.03	0.00	0.06	0.04	0.02	0.08	0.05	0.04	0.06	0.06	0.02	0.00	0.05	0.14

Panel B: Factor Rankings - All Loan Maturities (6m Default Window)

	Pre-Covid										Covid			
Factor	Dec-17	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
Total Debt	1.00	1.00	1.00	1.00	0.46	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Secured Debt-to-Income	0.02	0.21	0.14	0.13	0.92	0.78	0.10	0.39	0.14	0.05	0.19	0.07	0.26	0.20
# Soft Credit Checks	0.19	0.15	0.17	0.37	0.66	0.49	0.34	0.41	0.42	0.13	0.25	0.20	0.37	0.33
# Hard Credit Checks	0.13	0.10	0.15	0.15	0.12	0.14	0.11	0.12	0.06	0.09	0.34	0.02	0.14	0.16
ex.Mortgage Balance to Limits	0.04	0.04	0.03	0.07	0.16	0.03	0.01	0.04	0.12	0.08	0.00	0.03	0.00	0.10
Revolving Balance to Limits	0.01	0.04	0.06	0.00	0.03	0.05	0.00	0.03	0.04	0.01	0.13	0.04	0.08	0.14
Term	0.13	0.12	0.13	0.18	0.57	0.34	0.15	0.13	0.04	0.10	0.53	0.09	0.05	0.00
Oldest Account Age	0.63	0.55	0.38	0.53	1.00	1.00	0.66	0.93	0.81	0.47	0.79	0.34	0.78	0.78
Delinquent Accounts (Postcode)	0.00	0.03	0.00	0.01	0.00	0.00	0.05	0.00	0.00	0.00	0.18	0.00	0.07	0.09
Healthy Accounts (Postcode)	0.08	0.00	0.01	0.13	0.15	0.10	0.05	0.02	0.04	0.03	0.27	0.01	0.04	0.10

Table A.2: **XGBoost Model Performance by Maturity (Alternative Default Windows)**

In order to show the maturity effect identified in the main text isn't confined to a 9-month default window, we repeat the analysis for both a 12-month and 6-month default window. Analysing the below table, we see that for a 12-month default window, short-maturity loans underperform by an ROC-AUC of 4.9% during the pre-Covid window and 8.7% in the Covid window. For a 6-month default window, short-maturity loans underperform by an ROC-AUC of 7% during the pre-Covid period, rising to 16.2% during the Covid window. In terms of statistical significance, for a 12-month default window, short-maturity loans underperform long-maturity loans across the whole sample by a considerable margin (t-statistic of 24.87). Furthermore, the drop-off in short-maturity loan predictability (t-statistic of 20.34) exceeds the drop-off in long-maturity loan predictability (t-statistic of 17.94). Similar observations are had for a 6-month default window, with short-maturity loans under-performing long-maturity loans overall (t-statistic of 21.69), with the drop-off in short-maturity loan predictability (t-statistic of 11.65) exceeding the drop-off in long-maturity loan predictability (t-statistic of 6.11). We conclude that the maturity effect identified in the main text is observed across a range of default windows.

Panel A: XGBoost Out-of-Sample Performance by Maturity (12m Default Window)

	Pre-Covid										Covid			
Maturity	Dec-17	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
All Maturities	0.802	0.794	0.793	0.784	0.784	0.760	0.763	0.754	0.768	0.753	0.681	0.690	0.673	0.670
1-2 years	0.785	0.775	0.735	0.782	0.778	0.733	0.718	0.710	0.691	0.693	0.652	0.638	0.629	0.584
4-5 years	0.810	0.807	0.818	0.785	0.781	0.783	0.778	0.774	0.809	0.780	0.703	0.724	0.694	0.727

Panel B: XGBoost Out-of-Sample Performance by Maturity (6m Default Window)

	Pre-Covid										Covid			
Maturity	Dec-17	Mar-18	Jun-18	Sep-18	Dec-18	Mar-19	Jun-19	Sep-19	Dec-19	Mar-20	Jun-20	Sep-20	Dec-20	Mar-21
All Maturities	0.832	0.818	0.803	0.783	0.832	0.812	0.780	0.771	0.804	0.801	0.713	0.719	0.757	0.711
1-2 years	0.857	0.785	0.722	0.716	0.807	0.803	0.706	0.715	0.725	0.694	0.565	0.620	0.676	0.601
4-5 years	0.834	0.838	0.845	0.803	0.836	0.848	0.806	0.805	0.854	0.840	0.793	0.743	0.781	0.808

Table A.3: **Payment Holiday Piecewise Logit Model (12m Default Window)**

This table reports the results of our payment holiday piecewise-logit models. Two dummy variables are introduced to examine a potential structural break - *mid_dummy* corresponds to all borrowers with an XGBoost model-implied 9-month default probability (IDP) of 40-60% at loan origination, while *high_dummy* corresponds to borrowers with a model-implied 12-month default probability of >=60% at loan origination. We report coefficients and t-statistics for a variety of explanatory variable x IDP dummy interactions.

	Logit Model Results			
Variable	(1)	(2)	(3)	(4)
Intercept	-2.471**	-2.470**	-2.467**	-2.469**
	(-22.71)	(-22.69)	(-22.68)	(-22.71)
# Soft Credit Checks	0.067**	0.076**	0.076**	0.067**
	(3.53)	(9.12)	(9.19)	(3.53)
Term	0.009**	0.009**	0.009**	0.009**
	(11.82)	(11.81)	(11.81)	(11.82)
# Hard Credit Checks	0.101**	0.101**	0.102**	0.101**
	(9.56)	(9.62)	(9.66)	(9.58)
ex.Mortgage Balance to Limits	0.000**	0.000**	0.000**	0.000**
	(2.60)	(2.53)	(3.88)	(3.91)
Revolving Balance to Limits	0.009**	0.009**	0.009**	0.009**
	(18.97)	(19.17)	(19.18)	(18.97)
Oldest Account Age	-0.002**	-0.002**	-0.002**	-0.002**
	(-16.25)	(-16.43)	(-16.57)	(-16.35)
Delinquent Accounts (Postcode)	0.001**	0.001**	0.001**	0.001**
	(5.34)	(5.35)	(5.35)	(5.35)
Healthy Accounts (Postcode)	-0.002**	-0.002**	-0.002**	-0.002**
	(-1.98)	(-1.98)	(-1.99)	(-1.98)
Secured Debt-to-Income	-0.007**	-0.007**	-0.007**	-0.007**
	(-11.35)	(-12.11)	(-12.48)	(-11.57)
ex.Mortgage Balance to Limits $\times$*mid_dummy*	0.000	0.000		
	(0.27)	(0.29)		
ex.Mortgage Balance to Limits $\times$*high_dummy*	0.000	0.000		
	(0.97)	(1.14)		
# Soft Credit Checks $\times$*mid_dummy*	0.008			0.008
	(0.37)			(0.39)
# Soft Credit Checks $\times$*high_dummy*	0.015			0.017
	(0.71)			(0.81)
Secured Debt-to-Income $\times$*mid_dummy*	0.005**	0.005**	0.005**	0.005**
	(4.47)	(5.02)	(5.23)	(4.57)
Secured Debt-to-Income $\times$*high_dummy*	0.003	0.004	0.005	0.004
	(0.88)	(1.26)	(1.62)	(1.09)
n	94,865	94,865	94,865	94,865

B Algorithmic Details

B.1 Logistic Regression

The logistic regression is a supervised machine learning algorithm typically used for binary classification problems, first proposed by Cox (1958). The primary difference between a linear regression and logistic regression is the bounding of the logistic regression's output range between 0 and 1. In addition, as opposed to a linear regression, the logistic regression does not require a linear relationship between input and output variables; a linear combination of input features is fed into a nonlinear transformation via the sigmoid function. In the case of a logistic regression, the assumption is that decision boundaries are linear - that is, decision boundaries are hyperplanes in a high-dimensional feature space, where the dimension of the feature space is determined by the number of elements in the feature vector of a training example. Due to the simplistic assumption of linear decision boundaries, the logistic regression is often the first algorithm used for classification problems (Gudivada et al., 2016). As a result of linear, non-complex decision boundaries, the logistic regression is known to be less prone to overfitting.

In `scikit-learn`, logistic regression models are trained using a coordinate descent algorithm known as L-BFGS (Limited-memory Broyden Fletcher Goldfarb Shanno). Batch methods such as the L-BFGS algorithm, along with the presence of a line search method to automatically find the learning rate, are usually more stable and easier to check for convergence than stochastic gradient descent. L-BFGS uses an approximated second order gradient which provides faster convergence toward the minimum during parameter optimisation.

To begin with, we utilise the log-loss as our chosen cost function:

$$Cost(F_\theta(x), y) = -y.log(F_\theta(x) - (1-y).log(1 - F_\theta(x))$$

We then introduce an $\mathcal{L}2$ regularisation parameter λ to arrive at our overall loss function which we minimise via L-BFGS:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(F_\theta(x_i), y_i) + \frac{\lambda}{2m}\sum_{j=1}^{n} {\theta_j}^2$$

When tuning our logistic regression model, we grid-search a single parameter C, where $C = 1/\lambda$ and represents the inverse of regularisation strength λ. Very small values of C imply high regularization strength and promote parsimonious model structures that oftentimes under-fit the data. Large values of C imply lower regularization strength and tend to promote

over-fitting of the data. Simply put, optimisation of the C parameter in our logistic regression model serves as a regularisation procedure and allows us to prevent both over-fitting and under-fitting of the training data.

B.2 k-Nearest-Neighbours

The k-nearest neighbors algorithm (kNN) is a non-parametric classification method first developed by Fix and Hodges Jr (1952). Model input consists of the k closest training examples in a training dataset, where "close" in this case refers to Euclidean distance. In kNN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small - if $k = 1$, then the object is assigned to the class of the single nearest neighbor).The neighbors are taken from a set of objects for which the class is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

A drawback of the basic "majority voting" classification occurs when the class distribution is skewed, as is the case in our P2P default dataset. Examples of the more frequent class tend to dominate the prediction of a new example, as they tend to be common among the k nearest neighbors due to their large number (Coomans and Massart, 1982). One way to overcome this problem is to weight the classification, taking into account the distance from the test point to each of its k nearest neighbors. The class of each of the k nearest points is multiplied by a weight proportional to the inverse of the distance from each point in question to the test point.

Algorithm 1: k-Nearest-Neighbours

Input: Q, a set of query points and R, a set of reference points

Output: A list of k reference points for every query point $q \in Q$

for $q \in Q$ **do**

1. *Compute Euclidean distances between q and all $r \in R$.*
2. *Sort the computed Euclidean distances*
3. *Select k nearest reference points $r_{q1}, ..., r_{qk}$ corresponding to k smallest Euclidean distances*
4. $\hat{y}_q = \frac{1}{k} \sum_{j=1}^{k} r_{qj}$

end

For each training window, we grid-search the optimal number of neighbours k to use, with

this optimal figure ranging from 5-13 depending on the sample period

B.3 Naïve Bayes

The Naïve Bayes (NB) classifier is a probabilistic classifier based on Bayes' theorem and built on the assumption that each feature independently and equally contributes to the probability of a sample belonging to a specific class. The algorithm makes predictions about an instance belonging to a particular class by computing the class prior probability, the likelihood of belonging to a particular class, the posterior probability and the predictor prior probability as $P(C_j|x) = \frac{p(C_j)p(x|c_j)}{p(x)}$, where j is the number of classes and x is the feature vector. $P(C_j|x)$ is the posterior probability of class C_j given predictor x, $P(C_j)$ is the prior probability of class j, $p(x|c_j)$ is the likelihood of a predictor given a class and $p(x)$ is the prior probability of the predictor.

The NB classifier is simple to implement, computationally fast, performs well on large, high dimensionality datasets and is particularly suited to real-time applications.

B.4 Classification Trees

First proposed by Breiman et al. (1984), a classification tree is a hierarchically organized structure with each node splitting the data space into partitions based on value of a particular feature. This is equivalent to a partition of $\mathbb{R}^d$ into K disjoint feature sub-spaces $\{\mathcal{R}_1, ..., \mathcal{R}_k\}$, where each $\mathcal{R}_j \subset \mathbb{R}^d$. On each feature subspace $\mathcal{R}_j$ the same decision/prediction is made for all $x \in \mathcal{R}_j$.

Algorithm 2: Classification Tree

Initialise tree $T(D)$ where D denotes the depth; denote by $R_l(d)$ the covariates in branch l at depth d.

for d = *1,...,D* **do**

  for $\tilde{R}$ *in* $\{R_l(d), l = 1, ..., 2^{d-1}\}$ **do**

    Given splitting variable j and split point s, define regions

$$R_{left}(j,s) = \{X|X_j \leq s, X_j \cap \tilde{R}\} \text{ and } R_{right}(j,s) = \{X|X_j > s, X_j \cap \tilde{R}\}$$

    Find j, s that optimize

$$j,\, s = \underset{j,s}{\text{argmax}}\ \ Entropy(\tilde{R}) - Avg(Entropy(R_{left}(j,s), Entropy(R_{right}(j,s))$$

    Set the new partitions

$$R_{2l}(d) \leftarrow R_{right}(j,s) \text{ and } R_{2l-1}(d) \leftarrow R_{left}(j,s)$$

  end

end

Result: A fully grown classification tree T of depth D. The output is given by

$$f(x_i) = \sum_{k=1}^{2^D} \text{avg } (y_i|x_i \in R_k(D))\mathbb{1}_{x_i \in R_k(D)}$$

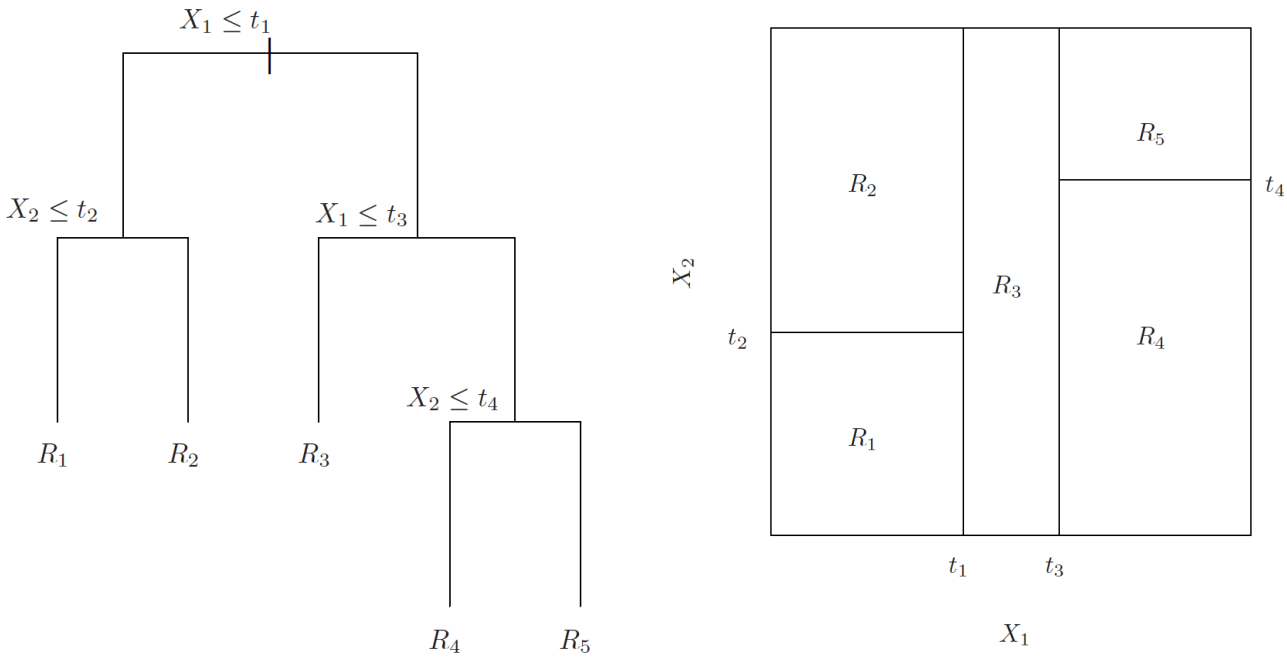
i.e. the average response in each region R_k at depth D.

Ideally, would like to find partition that achieves the lowest entropy for a classification problem. Given the number of potential partitions is too large to search exhaustively, greedy search heuristics must be used to determine the optimal partition - starting at the root node, we evaluate the loss for splitting on all combinations of features j and and split points s. The optimal pair (j, s) determines the members of each child node. Finally, we recurse on all child nodes iteratively until some stopping criterion is met.

B.5 Random Forest

While classification trees offer non-parametric, non-linear framework for modelling, they are often prone to overfitting training data - i.e. they record low bias and high variance (Mitchell et al., 1997). Random forests utilise an ensemble approach, combining the output of multiple

Figure B.1: **Classification tree**



(a) Recursive binary splitting

(b) Partition tree

decision trees in a bootstrap-aggregation format. This procedure relies on the notion that large numbers of weak learners perform better in aggregation relative to small numbers of more complex learners. While the hyperparameters for individual trees are similar in both classification tree and random forest model structures, random forests incorporate additional randomness at the tree-level; rather than searching through all possible features when evaluating split points, the algorithm searches for the best split point among a random subset of features. The resulting individual trees display lower correlation and offer more power when used in an ensemble format. As a result, the hyperparameters we tune in our random forest model are the number of features to randomly select when evaluating split points for individual trees, as well as the depth of each individual tree. For simplicity, we opt to fix the number of trees in our random forest at 100.

Algorithm 3: Random Forest

Determine forest size F

for $t = 1,\ldots,F$ **do**

Obtain bootstrap sample Z from original data.

Grow full trees following Algorithm (2) with the following adjustments:

1. Select $\tilde{p}$ variables from the original set of p variables.
2. Choose the best combination (j, s) (c.f. Algorithm (2)) from $\tilde{p}$ variables
3. Create the two daughter nodes

Denote the obtained tree by T_t

end

Result: Ensemble of F many trees. The output is the average over the trees in the forest given as

$$f(x_i) = \frac{1}{F} \sum_{t=1}^{F} T_t(x_i)$$

B.6 XGBoost

For our XGBoost model, we tune three hyperparameters via grid-search: the learning rate, tree depth and percentage of column samples considered for split point evaluation. For simplicity, we fix the number of sequential trees in our model at 200.

The learning rate controls the impact sequential residual-correction models have on the overall model prediction at each iteration. Small learning rates are preferred to ensure convergence of the loss-function gradient. Tree depth controls the depth of each sequential tree fitted on the residuals from the previous model iteration. Given that boosting algorithms derive much of their benefit from the sequential addition of weak learners, shallow trees are usually preferred to deep trees.

The XGBoost algorithm has several in-built functions to control tree depth and complexity, namely alpha and lambda. For simplicity, we opt to leave these regularisation parameters at their default levels. XGBoost also has an in-built pruning functionality controlled via the gamma parameter (if the gain from an additional split point in a tree fitted on the model residuals from the previous iteration is lower than gamma, the node in question is pruned from the tree). For simplicity, we leave this at the default value also.

We also tune the percentage of feature variables available to the splitting algorithm when evaluating potential split points in a tree. As with the random forest model, limiting the num-

ber of potential variables to split on reduces the complexity of the model and helps to prevent over-fitting.

Algorithm 4: XGBoost

Input: Training set $(x_i, y_i)_{i=1}^N$, a differentiable loss function $L(y, F(x))$, a number of weak learners M and a learning rate α

Initialize model with a constant value:

$$\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta).$$

for *m = 1,...,M* **do**

Compute the gradients and hessians:

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)}$$
$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)}$$

Fit a base learner (or weak learner, e.g. tree) using the training set $\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\}_{i=1}^N$ by solving the optimization problem below:

$$\hat{\phi}_m = \arg\min_{\phi \in \mathbf{\Phi}} \sum_{i=1}^{N} \frac{1}{2}\hat{h}_m(x_i)\left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)\right]^2$$
$$\hat{f}_m(x) = \alpha\hat{\phi}_m(x)$$

Update the model:

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x)$$

end

Result: Final model output is given by:

$$\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x)$$

B.7 Neural Network

Neural networks are forecasting methods that are derived from simple mathematical models of the human brain, and allow complex nonlinear relationships between a response variable and its predictors. Neural networks began with the pioneering work of McCulloch and Pitts (1943)

- they outlined the first formal model of an elementary neural network and demonstrated its ability to represent common logical operators such as "AND" or "OR" functions. Later, they discovered combinations of neurons could be used to replicate the human brain's approach to pattern recognition and classification.

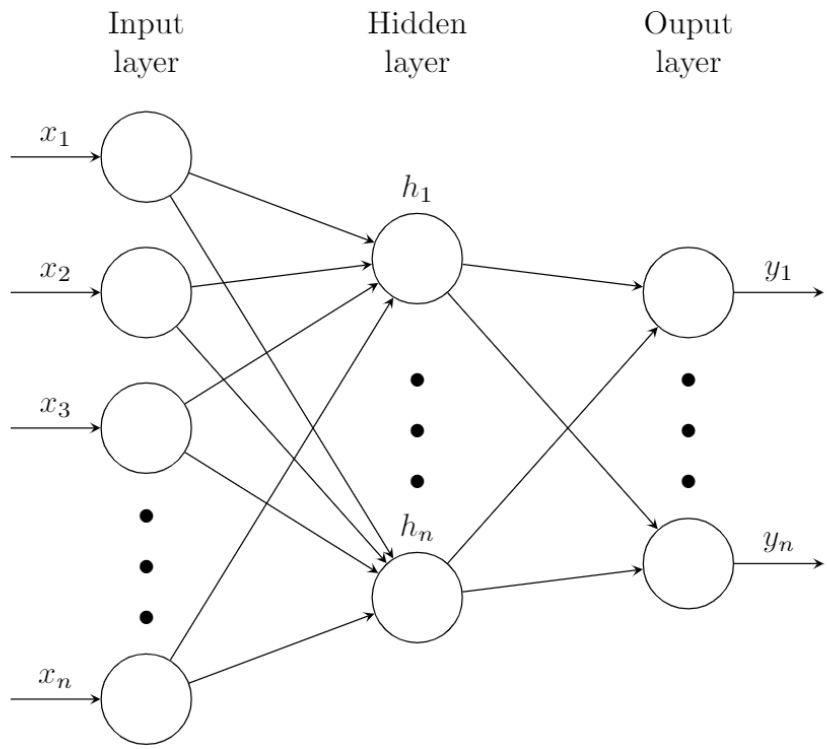
A neural network can be thought of as a network of neurons organised into layers. The predictors (or inputs) form the top layer, and the forecasts (or outputs) form the bottom layer. A simple network with no hidden layers replicates a traditional linear regression model. Once we add intermediate layers with hidden neurons, non-linearity is introduced. This approach is known as a multi-layer feed-forward network, where each layer of neurons receives inputs from the neurons in previous layers.

The outputs of the neurons in one layer are inputs to the next layer, where the inputs to each neuron are summed via a simple weighted linear combination. The weights are selected in the neural network framework using a learning algorithm that minimises a cost function, such as the MSE for regression-based problems or the log-loss for classification problems. This weighted linear combination is then modified using an activation function (the sigmoid function is often chosen for classification problems) to give the input for the next layer. The incorporation of an activation function reduces the effect of extreme input values, thus making the network more robust to outliers.

The weights take random values to begin with, and these are then updated using observed data via a backpropogation algorithm. Consequently, there is an element of randomness in the predictions produced by a neural network.

An typical example of a neural network architecture is shown below. For our architecture, we utilise 10 input nodes (one for each input variable), one hidden layer with 10 nodes (i.e. a fully-connected structure), and one output node for the output layer. The hidden layer makes use of the ReLu activation function, while a sigmoid activation function is chosen for the output layer. We use the binary cross-entropy loss function as is standard for classification problems, and opt for the efficient "Adam" optimiser (Kingma and Ba, 2014) to carry out the gradient procedure and subsequent node weight updates.

Figure B.2: **Neural Network Architecture**



Algorithm 5: Backpropagation

Initialise all weights **w** in the network and set learning rate η;

for *i = 1..max_epochs* **do**

  for *j = 1..n* **do**

    $\forall w \in \mathbf{w}$: $\Delta w = \frac{-\partial \bar{E}rr_j}{\partial w}$;

    where $\frac{-\partial \bar{E}rr_j}{\partial w} = Avg(\frac{-\partial Err_{j,1}}{\partial w}, ..., \frac{-\partial Err_{j,10}}{\partial w})$

    $\forall w \in \mathbf{w}$: $w_{new} \leftarrow w_{old} + \eta \Delta w$;

  end

end

Given the computational complexity associated with neural network training, we choose not to grid-search parameters. For our model, we opt for mini-batch gradient descent using a batch size of 10, i.e. after each batch of 10 training instances, the average gradient $\frac{-\partial \bar{E}rr_j}{\partial w}$ is calculated and backpropogated through the network. n represents the number of batches in our training dataset. Finally, we set the number of epochs to 50 to ensure convergence of the gradient function.

C Computational Details

Our machine learning libraries of choice are the popular `scikit-learn` and `keras` packages used within a Python 3 programming framework. Our XGBoost model implementation is derived from the `xgboost` library. We use `pandas` for data manipulation and `numpy` for mathematical operators. Our regression package of choice is the `statsmodels` API. For data preprocessing, we utilise the `PowerTransformer` class from the `scikit-learn` package, as well as making use of the `Pipeline` feature to prevent data leakage between test/train datasets.

C.1 Setup

While origination factors and monthly performance updates for 500,000+ loans appears to suggest intensive computational requirements, the small size of our rolling train/validation and test windows coupled with the relatively low complexity associated with training tree-based algorithms implies powerful hardware instances (such as the high-performance GPU computing capabilities offered via Amazon Web Services) are not required. All work was carried out on a single 2.60 GHz, 16GB RAM node with 6 cores.