A Work Project, presented as part of the requirements for the Award of a Master Degree in Finance from Nova School of Business and Economics

An Approach to Securitisation in Europe
NPLs – Machine Learning Model
Field Lab Project Nova SBE | Moody's Analytics

Achille Cornelis Touchais (nr. 4173)
Filipe José Charneca Barreto (nr. 3922)
Henrique Barata Gameiro (nr. 3926)
José Eduardo de Sousa Pedro dos Reis (nr.3916)
Roberta Trento (nr. 4069)

A project carried out on the Master in Finance Program, under the Supervision of:

Moody's Analytics Advisors:
Carlos Castro, Director, Economics and Structured Analytics
Mike Mueller, Senior Director – Software Engineering, Structured Solutions
Lavinia Roma, Financial Engineer, Structured Analytics & Valuations

Faculty Advisor:
Professor João Pedro Pereira

January 4th 2019

Table of Contents

Abstract – ROBERTA 4

Executive Summary – ALL GROUP 5

1. Approaches to Credit Risk Modelling – FILIPE 8

1.1 Corporate Credit Risk – FILIPE, HENRIQUE 8

1.2 Mortgage Risk Approaches - FILIPE 9

1.2.1 Linear Regression 11

1.2.2 Logistic Model 12

1.2.3 Survival Analysis 13

1.2.4 Optimization model 15

1.3 Deep Learning Models and credit and liquidity risk - ACHILLE 17

1.3.1 Corporate credit risk evaluation 18

1.3.2 Private Loans credit risk evaluation 19

2. Transition states – FILIPE, JOSÉ, ROBERTA 21

2.1 Default and Prepayment risk 21

3. Literature and new Variables 24

3.1 Overall analysis – FILIPE, JOSÉ 24

3.2 Created variables 26

3.2.1 Ability to cover the loan with the property value – ACHILLE E FILIPE 26

3.2.2 Time elapsed since evaluation – ACHILLE E FILIPE 32

3.2.3 Number of valuations per Loan 35

3.2.4 Loan Age and related variables – FILIPE E HENRIQUE 35

3.2.5 Loan Balance related variables - FILIPE 39

3.2.6 Age of borrower – HENRIQUE, ROBERTA 42

3.2.7 Income related variables – HENRIQUE, ROBERTA 44

3.2.8 Balance in arrears in proportion to loan's outstanding value - ACHILLE . 47

4. The Dataset – FILIPE, JOSÉ, ROBERTA 48

4.1. Handling the dataset - FILIPE, JOSÉ, ROBERTA 48

4.1.1. Gap Flags 48

4.1.2. Date transformation 49

4.1.3. Data harmonizing 49

4.1.4 Dealing with missing values 49

4.1.5. Loan Status: Removal of categories 49
4.1.6. Loan Status: Prepayment and Default States 49
4.1.7. Loan Status: 12-Months Lag 50
5. Model – ACHILLE, JOSÉ 52
5.1 Buckets - JOSÉ 52
5.2 Network Mechanics – ACHILLE, JOSÉ 53
5.3 Optimization – ACHILLE, JOSÉ 58
5.4 Hyperparameter Selection – JOSÉ 59
5.5 Imbalanced Classes - José 60
6. Results – ACHILLE, JOSÉ 62
6.1 Variable Significance - JOSÉ 63
6.2 Variable Impact - JOSÉ 65
6.3 ROC curve - ACHILLE 66
6.4 Graphical predicted default rate – JOSÉ 69
6.5 Predicted default rate – HENRIQUE, JOSÉ 72
7. Conclusion – ACHILLE, ROBERTA 73
8. Next Steps – HENRIQUE, JOSÉ, ROBERTA 73
9. Limitations - ROBERTA 75
Appendix 77
Appendix A – Structural Credit Risk Models 77
Appendix B –Analysis of the created variables 78
Appendix C – List of Variables 91
Appendix D – Loss Graphs 93
Appendix E – Variable Significance 94
Appendix F – Variable Impact 98
Appendix G – Metrics and ROC curve 106
Appendix H – Predicted Default Rate Analysis 108
Appendix I – Alternative Methods 111
Appendix J – Structured Finance Portal – ALL GROUP 113
1.1 - Key Differentiators Explanation 113
1.2 - Credit Migration Probabilities on the Tranche 114
1.3 - Filters on Market Performances 115
1.4 - Functions 116

1.5 - Transition Matrix 117
References 118

Abstract – ROBERTA

As a consulting project, we were proposed to develop a neural network (NN) to predict mortgage states in one year, based on the paper 'Deep Learning for Mortgage Risk' by Justin A. Sirignano, Apaar Sadhwani, Kay Giesecke (2018). We developed a neural network model with the aim of being able to capture the relationships between the different variables, with respect to each other and to the response variable (the loan status in 12 months), better than traditional classification methods, such as logistic regressions, which constitute the benchmark set. Data was provided by Moody's, relating borrower, property and loan/financing characteristics for several mortgages over several periods in time (over 350 thousand mortgages). The purpose of our model is to predict the probabilities to transition to different states at a certain point in time. The best results were obtained with a 10 layer, 500 nodes per layer network. The model can identify a large portion of defaults. At the cost, however, of a general overestimation of the default rate over the years. The capability of identifying loans that will be in arrears is also acceptable, with, again, an overestimation of the verified rate. Variables relating to borrower characteristics and history as well as financing are found to be the most significant.

Executive Summary – ALL GROUP

This project investigates whether a deep learning model (a type of neural network) can be useful to predict the performance of a pool of mortgage loans. In traditional econometric models, the relationship between explanatory and dependent variables is constrained by the functional form of the model itself and by the limited data transformations that can be incorporated (squaring variables, logs, etc.). Those models may thus be insufficient to fully capture the complexity of the relationships between dependent and independent variables. In contrast, neural networks are well suited to capture complex non-linear relationships in the data.

Our work is mainly inspired by the paper "Deep Learning for Mortgage Risk", by Justin A. Sirignano, Apaar Sadhwani, Kay Giesecke (2018). One of their main findings was precisely the existence of non-linear relationships in the data. They found that the most important component of this non-linear relationship was the interaction between the variables, meaning that the sensitiveness of the loan performance to one variable depends on other variables.

In this report, we aim at developing a non-linear deep learning model and replicate the promising results in Sirignano et al. (2018). Our dataset was provided by Moody's and, due to hardware constraints, a small sample of 20 000 loans (out of 350 thousand) was randomly extracted. Several other filters were applied to the sample, such us guaranteeing all loans to have 24 consecutive observations, as well as to the construction of some variables. These are detailed in section 4.

From more traditional literature (section 3.1), we found that variables are usually categorized as borrower, financing (loan) and property, with some papers also trying to gauge how macroeconomic conditions can affect borrower's behavior. Sirignano et al. (2018) find evidence for strong macro effects, namely the unemployment rate, which is the most significant variable for explaining states transition.

Other papers and Moody's recommendation, due to good performance logistic regression, considered Loan age and Current LTV to have strong ability to differentiate the different states: default, prepayment, delinquency and performing. As well as these, we created other interaction variables we found relevant, namely *Completion*

(percentage of the contractual length of the loan already completed). The created variables are detailed in section 3.2.

The best results were obtained with 10 layer-network, trained over 800 epochs, using mini-batch stochastic gradient descent (SGD). Since there is a severe imbalance between classes in our dataset (with “performing” being the dominant class), an artificial weight (twice the inverse proportion of each class) was applied during optimization. The full model mechanics and specification are detailed in section5.

Our results reveal that borrower related risk factors, namely income related factors, such as the proportion of delinquent balance compared to the borrower’s income and installments as a proportion of income, as well as stability (employment status), are important at predicting the future state of the loan. Past information about the borrower, namely court related events and previous defaults, also prove to be significant (section 6.1 and 6.2 details the results, in section 6.1 we found the most significant variables – worse performance when omitted)

There is a general overestimation of the default probability by our model. Still when analyzing the receiver operating characteristics (ROC) curve (Section 6.3) we can see that a decision rule can be made to identify transitions to default and delinquency well, from originally performing loans. The overestimation can also be seen in section 6.5 where the predicted default rate is plotted against the actual one (across time).

Since our model was best at predicting defaults, analyzing the transition from performing to default state became the focus in section 6.4. Different variables were assed. Interesting patterns can be observed in the predictions. For instance, the higher estimated probability of default in situations where principal payment is delayed, indicating that agents with low home equity have higher propensity to default. The contrast between geographic regions, that is, mortgages from specific areas are more prone, to enter in default, according to our model, indicating that the property related risk factors (in this case location) can also be a differentiating factor (although not found as crucial in section 6.1 proved to have close relation with the probability of transitioning to default – section 6.2). More variables are discussed in section 6.4.

Given that the process to obtain an optimal network specification is an iterative one, we were constrained by the time necessary to train the model, so only a few iterations were possible. Hardware constraints also made it difficult to use a larger sample, which would likely reveal more general patterns (more in section 9).

M1 (our final model) is still a very imperfect model and further steps may be taken in order to improve on it. These relate to: the way the sample is obtained (pooled cross-section); the way the variables are encoded before being used to train the model (normalization would be preferred to the current method); the necessary iterations through possible network specifications (parameters that need to be manually chosen), with preference for larger networks (large amount of nodes and layers) which we found to work better during our testing. Research on the methodology to assess the significance of variables would also be beneficial, since it is too costly to simply retrain models without them. Finally, an ensemble model, which would combine the strengths of different networks (more details in section 8).

There are also possible alternatives to our framework, namely the use of decision trees, which allow for an easier identification of the importance of each variable, and the use of a recurrent network which would work with panel data (tracking observations belonging to the same loan) and not a pooled cross-section. Appendix I touches on these 2 methods.

1. Approaches to Credit Risk Modelling – FILIPE

In this section, we will start by covering briefly the conceptual corporate credit risk model approaches, namely the two big classes: Structural and Reduced-Form models. Afterwards, we will discuss, with more detail, the residential mortgage risk, particularly the difficulties to model it and the conceptual loan-level models applied to measure and manage this type of loans. Finally, we will cover the new techniques, namely the machine learning models and how can they be used to evaluate residential mortgage risk.

1.1 Corporate Credit Risk – FILIPE, HENRIQUE

The concerns towards credit risk exposure are relatively recent, although the banking institutions are present in the global economy for a long time. The first class of credit risk models was born, in the late 60s. In fact, the Altman's Z-Score (1968) and other credit scoring models assigned a score to companies according to their bankruptcy risk, based exclusively in their financial information. The fact they were not considering market information was the biggest flaw of this type of models, moreover they did not estimated the probabilities of default: "it may indicate that the mortgage is likely to default, but it does not tell how likely it is to default (i.e., whether there is a 90 per cent or 60 per cent probability of default)" (Li, M.; 2014).

The importance of the market information started to be noticed, and the first generation of structural market-based credit risk models appeared in the seventies, especially after 1974, when it was founded the BCBS, the Basel Committee on Banking Supervision. These models are known as Structural models since they relate the credit risk management to fundamental variables, such as the firm assets' value: if the assets become lower than the liabilities, the firm will default. They attempt to price the credit risk, i.e., the price of the exposure and they are based in the option pricing models: Black Scholes's and Merton's models. Although they give some good insights to compute probability of default, their assumptions are too strictly and hardly hold in the real world, considering, for instance, non-stochastic interest rates; too simple capital structures; and default occurring only at maturity.

The second generation of structural models was born from changes to Merton's model to accommodate less strict assumptions, for example, they started allowing for default before maturity and using stochastic processes to define interest rates. However, these models, have a low number of inputs failing to capture some information and generating poor performances, in certain scenarios. Despite the limitations, structural models are broadly used in the corporate credit risk industry, such as Moody's-KMV Portfolio Manager, Credit Metrics, or Credit Portfolio View. (a brief information regarding these models can be seen in Appendix A.

Their application in mortgage risk estimation is not so frequent, even though, Cunningham & Hendershott (1986) applied "the Black and Scholes (1973) option pricing model as modified by Brennan and Schwartz (1977)" in order to analyze the default risk of different types of mortgages and loan programs and, in the last instance, to compute the optimal default premia Federal Housing Administration (FHA) should charge to different borrowers.

The other class of credit risk models is the Reduced-form, which contrasts deeply with the structural approach. Instead of having the probabilities of default derived from the assets' value, the default event is considered exogenous, this way the probability is derived from a random variable that follows a Poison distribution. The default will happen if this "exogenous random variable jumps instantaneously from one to another at random times", (Zhang, X.; 2017). In fact, nowadays, we are able to use reduced-form models that make predictions over several periods, that focus on time-varying covariates, instead of static covariates. This type of models uses this statistic based stochastic process, instead of a typical theoretical model, being, then, less dependent on assumpions. One example of a reduced-form model is the *CreditRisk+*, used by Credit Suisse. This model is pretty easy to implement but has the disadvantage of only consider default state besides performing; then it only computes the default probability, ignoring all the other rating levels.

1.2 Mortgage Risk Approaches - FILIPE

As we have discussed, the focus of our project is mainly related to private mortgage loans, this is, loans borrowed by householders to finance real state acquisition. Residential loans' risk is not so easy to measure as corporate loans' risk, since individuals' information is not accessible in the market. This information unavailability is a big constraint when building a residential mortgage risk model, however there are other characteristics that make quite hard to analyze, measure and manage residential mortgages lenders' exposure.

Firstly, it is necessary a lot of scenarios to capture all the possible borrower behaviors (loan status) in different economies. The loan-level behaviors are not homogeneous, in other words, in different economies, there are strong evidence of different performances and correlations, for the same loan. The mortgages loans performance is much more dependent on the economic state than commercial loans and the volatility is relatively higher as well. Summarizing: one loan can have different behaviors according to the economic scenario and different loans, in the same scenario, can present very different performances.

Secondly, and like most of the loans, the mortgages are path dependent instruments. This means that loan history (historical information and performance) is relevant in future performance, and the past and current behavior will have a tremendous impact in future behavior. Thus, mortgages analysis requires a multi-period model.

Lastly, mortgages may have also call and put options: option to prepay and option to run away, respectively.

Nevertheless, there are some models broadly used to study mortgage risk, they are divided in Loan-level models and portfolio level models, according to the scope of the analysis, the implementation, and the data used.

A Loan-level model, as the name indicates, would be suitable to study the performance of individual loans. The input, usually, relies on borrower and mortgage individual information (instead of macroeconomic data) and the output is the loan's default probability, that eventually could be aggregated in order to estimate the loss of a given portfolio.

On the other hand, a portfolio-level model is applied when we want to study the default rate of a mortgage loans portfolio, as a whole. The interaction between the loans (i.e. correlation) is quite significant, this way, the data is mainly composed by macroeconomic explanatory variables, and less by borrower or mortgage individual information. In general, the inputs of the portfolio models are aggregated, this is, each input of the portfolio model is the weighted average of the inputs of each individual loan that composed the portfolio. For instance, the LTV ratio of the portfolio model will be the weighted average of the individual loans' LTV ratios. A similar situation occurs with the output: it is obtained the portfolio's probability of default, instead of the loan by loan probabilities.

Bottom line, we can infer that portfolio models are more restrictive than loan-level models: the losses predicted by a loan-level model can be aggregated in order to obtain portfolio loss, whereas the portfolio output cannot be insulated.

Thus, in these sub-sections, we will focus on explaining and analyzing the advantages and disadvantages of four widely used loan-level mortgage risk models:

1.2.1 Linear Regression

The first model developed to study mortgage risk was a basic linear regression, in which the default risk of certain loan, known as Loan Status, - the dependent variable that takes the value of 0, if it defaults, or 1, if not - is determined by k independent variables. The model follows the regression:

$$Default\ risk = \propto + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Where $\propto$ is the constant, x_i are the variables that try to explain the default risk, β_i the linear coefficients, that measure the sensitivity of default risk to a change of certain magnitude in the i^{th} explanatory variable, and ε the error, this is, what is not explained by the model.

This model is quite simple and easy to implement: from a sample analysis, it estimates the value of the coefficients that, posteriorly, are used to predict the default risk of a specific loan. Moreover, both panel and cross-sectional data can be used, and

the coefficients and output are interpreted in a straightforward way. For instance, the significance tests are pretty much easy to perform.

Still, the main problem of this model is the linearity assumption; according to it, even if the default determinants are transformed before joining the regression (for instance, through a log transformation), the relationship between the dependent and the explanatory variables is assumed to be linear. However, there are evidences of non-linear relationships between borrower behavior and the variables, which this linear model fails to capture, as it will be proved in section 3.2. The main reason we are building our machine learning model is to overcome this problem.

Other issue is related to the dependent variable, the default risk. This is not the default probability but a proxy, that can be seen as the predicted Loan Status. Instead of giving us a number between zero and one, the model outputs the 0, if it predicts default, or 1, otherwise. Thus, the model assumes only two scenarios: Default or No default, ignoring, in a certain way, how much close (or far) the mortgage seems to be from defaulting.

1.2.2 Logistic Model

The logistic model is an improvement to the linear regression, solving some of its problems. The logit model, as it is broadly known, applies a positive monotonic transformation to the linear regression, through the following logit formula, transforming its output in a default probability.

$$Probability\ (Loan\ Status = 1) = \frac{1}{1 + e^{-(\propto + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k)}}$$

Where the dependent variable is the probability of default (status equal to 1).

This formula is applied when we have got a binary dependent variable. For instance, we were considering only two loan status, the mortgage could default (loan status=1) or do not default (loan status=0).

However, in real world, when accessing the credit risk, a mortgage may assume different categories within the non-default class: performing, delinquent and prepaid. Thus, the dependent variable is not binary anymore, and to incorporate more than two

states we can use a multinomial logistic regression, which will not be developed in this paper, once it is a complex procedure and not common to implement. So, logit model does not necessarily limit us to a simple binary framework.

$$Probability\ (Loan\ Statu = j) = \frac{e^{(\propto+\beta_1 x_1+\beta_2 x_2+\cdots\beta_k x_k)}}{1+e^{(\propto+\beta_1 x_1+\beta_2 x_2+\cdots\beta_k x_k)}}$$

Campbell & Dietrich (1983) applied a multinomial logit regression to access residential mortgage risk.

Most of the mortgage papers applies logistic regressions. As discussed above, logit model overcomes some challenges of the linear regression, not only in extending the status the mortgage can be considered in, but also regarding the outcome, in the sense that this model, contrary to the linear regression, give us the prediction of how close a loan is from default, measured by the probability of default. In testing the significance of the explanatory variables, both models give similar outcomes.

Nevertheless, the positive monotonic transformation of the linear regression does not relax the linearity assumption. This way the flaw persists: under the logistic model, the logarithmic function of the odds is a linear function of the explanatory variables. This may have implications in fitting specific datasets, for example, an explanatory variable, having a significant non-linear relationship with default probability may be wrongly excluded from the model, due to lack of significance.

Failing to capture the non-linearity relationship sometimes can be mitigated by including in the regression the squared-term of the variable and cross-terms with other explanatory variable, or even by using categorical and dummy variables; however, these are not fool proof methods, and do not guarantee the overall non-linearity effect capture.

1.2.3 Survival Analysis

Survival analysis is an alternative method for the previous models. This analysis gives a special emphasis to the life course of the mortgage.

As we discussed before, a mortgage can be marked as performing, delayed, prepaid or default. These loan statuses are mutually exclusive, meaning that, at each

point of time, the mortgage loan status can be one and only one of these states. However, over, the time, the status can change across categories, in accordance with the borrower behaviour. It is frequent a loan, that starts by being performing, enters in delinquency, in default or gets prepaid.

Survival analysis in precisely studying this relationship between loan status and time passage, more specifically, how long a mortgage survives, i.e., the time it takes to migrate from performing to default or delinquent. Note that delinquency risk is not considered under the assumptions of this model and default and prepaid are assumed absorbing states, after reaching one of these states the mortgage will remain with that state, thenceforth.

The conditional probability of survival until t, also known as hazard rate, is defined by the following expression:

$$h(t) = h_0(t)e^{(\beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k)}$$

where, t is the loan age, h_0 is the empirical baseline hazard, which captures the shape of the hazard function, according to t. Like the previous models, x_i are the explanatory variables and β_i the coefficients, that studies the impact of each variable in the conditional probability of survival. This formulation assumes a discrete approach to time, the continuous would have $h(t)\Delta t$, as dependent variable.

Green & Shoven (1986), for instance, looked over mortgage and borrower variables and try to predict which of them may contribute to a default, in a defined time period.

Elul et al (2010) estimated a dynamic logit model (explained in the last section), using a hazard function varying nonparametrically, applied typically in these survival models. In fact, “survival analysis models (namely the Cox regression) and logistic regression models sometimes include quadratic or other nonlinear transformations of certain variables” (Sirignano et al.; 2018). The Cox proportional hazards model, referred before, is one category of survival analysis that “allows to analyze the effect of several risk factors on survival”, and may include a nonparametric baseline hazard function in order to capture non-linear effects, (Sirignano et al.; 2018).

One advantage of the survival analysis is the fact that it can be structured to capture each one of two types of mortgage risk: default risk and/or prepayment risk. As it will be explained further in section 2, both termination risks, even being mutually exclusive, should be considered by the lender, since a default or prepayment event will impact his cashflows.

Deng et al. (2000) included alongside both termination options estimations: for prepayment and for default and estimated the unobserved heterogeneity ("risk preferences and other idiosyncratic differences across borrowers") of the borrowers, that seems to be meaningful.

This matching between the loan age and the termination event is quite important and extends the typical frameworks, since it is not only being studied the impact of the conceptual determinants in the probability of the event of interest, but also the impact of the life course of the mortgage, i.e., the model is formulated to generate the probabilities as a function of loan age and the other explanatory factors (also assumed by logistic and linear regressions). The time horizon, trough survival is flexible from loan to loan, contrary to the logistic regression, in which each estimation process generates the default probability for a specific (fixed) time period.

Other advantage of the survival analysis is the incorporation and adjustment of 'censored data' in the estimation procedure, while the logistic model would abandon the observations for unavailable information. The assumptions of survival model are also quite flexible, when estimating semi-parameters such as the hazard baseline.

The main problem of this analysis in related to the data handling: firstly, the data be difficult to treat, secondly, the output analysis could require some developed programming techniques. These are the main reasons this model is not broadly used, in practice, although the good predictions it generates.

1.2.4 Optimization model

The previous three models are statistical models, the estimation procedures applied are related to the fundamentals of statistics, such as linear regressions and logit transformations. The following model is an economic model, in the sense that it is structured to capture the economic decision behind the termination event. Even though

this is not statistically related, we are going to study this model, once it can help us to understand the economic reasoning of certain determinants that may impact the borrower behaviour, and even which explanatory variables we should incorporate in the machine learning model.

The optimization model assumes that a borrower takes default or prepayment decisions in a way to maximize his wellbeing, measured by his wealth and utility functions, what is equivalent to say that the borrower minimizes the costs related to the house. The borrower, in each period, will choose one of three alternatives, according to lower cost. The alternatives are: to pay the instalment, to prepay the mortgage (refinance) or to default the current mortgage. Basically, he will choose the option that generates a larger benefit to himself, i.e., that optimizes his wellbeing, minimizing the cost function.

There are several different functions to define the borrower's choice, we opted for the following one, according to Capozza, Kazarian, and Thomson (1996), which is the most consensual:

$$P_t(H_t, r_t) = \min[{P_t}^d(H_t, r_t), {P_t}^p(H_t, r_t), {P_t}^c(H_t, r_t)]$$

We will not do an exhaustive analysis to this function, but rather understand the importance and the limitations of this type of models. Summarizing the function, the cost of the house for the borrower, in each month, given by $P_t(H_t, r_t)$, will be the lowest cost among the previous three alternatives: defaulting, P^d, prepaying the current mortgage and refinancing, P^p, or performing the current mortgage, P^c.

The cost, in each alternative, depends on the current house value and on the interest rate payed, in each month, defined by H_t and r_t, respectively. These explanatory factors are estimated, in each month, t, in several scenarios, through stochastic processes, that, afterwards generate the decisions distribution, from where is computed the probability of each state.

The model follows a simple framework that can eventually include other determinants of default, for example, the monthly income or the loan age. There are more complete models that include, as well, "trigger events", such as divorce or

unemployment. However, these variables selection process needs to be done carefully since the estimations are quite sensitive to inputs, then, bad assumptions may distort the outcomes of the model.

Optimization models link the borrower behaviour to economic forces, depending less on loan historical data, what can be positive, if we have poor information on loan.

On the other hand, since they are not statistically-heavy, economic models require developed programming to make predictions, and this is what makes the implementation harder when compared to statistical models. Not considering the delinquency scenario, in the decision process, can be seen as a caveat; moreover, the assumptions made in order to compute the cost under each decision, P^i, is relatively subjective.

We feel important to reinforce that the model (i.e. equation) studied is one of the several optimization models developed. There is literature that uses the utility maximization approach, instead of cost minimization, in which they "define household utility as a function of non-durable consumptions over time, housing consumptions over time and/or terminal wealth (financial wealth and housing wealth)", Li, M. (2014).

1.3 Deep Learning Models and credit and liquidity risk - ACHILLE

Due to the nature of their business, banks have plenty of data on their customers. However, many of them struggle with the stochastic behavior of their customers and misevaluate their credit risk, which can lead to inaccurate estimations of their liquidity buffer's need. Yet a correct management of liquidity risk is key since banks are required to maintain a healthy balance between investing to maximize their shareholders' profit and maintain high liquidity levels to respect their obligations to depositors (Tavana, et al., 2018). Moreover, as seen in previous examples, mismanaging liquidity and misevaluating credit risk can lead to the bank's failure. Too little liquidity and you risk insolvency, too much and you risk inefficiency (Matz, 2007).

The models from last section fail to capture the non-linear relationships between the borrowers' behaviour and the explanatory variables. Moreover, the correlation between variables can be harmful to prediction, resulting on bad performances and

misevaluation of credit risk. A machine learning model would overcome this conceptual models' limitations. In fact, artificial intelligence (AI) and machine learning models are being developed to model credit risk, not only for corporation loans but also for residential mortgages.

The subject of Deep Learning's use for credit risk assessment and management has been widely studied in the academic literature (Barboza, Kimura, Altman; 2017) (Huang, Liu, Ren; 2018) (Angelini, Tollo, Roli; 2008). Tavana, Abtahi, Di Caprio and Poortarigh have researched the potential positive impact deep learning's techniques could have on banks' liquidity risk measurement and management (2018).

Tavana, Abtahi, Di Caprio and Poortarigh found that neural network's technology can detect better liquidity risks' occurrences using data available in any banks' balance sheet. In addition, as neural network can deal with very noisy data and missing values (Angelini, Tollo, Roli; 2008), their use does not require extensive preprocessing of the data, facilitating the job of banks' managers (Tavana, Abtahi, Di Caprio, Poortarigh; 2018).

1.3.1 Corporate credit risk evaluation

There is also an extensive literature on how deep learning models can improve corporate credit risk's evaluation (Zhao, Xu, Kang, Kabir, Liu; 2015) (Khashman; 2010). Among the most praised assets of deep learning models are their ability to predict credit events with high accuracy, despite restrictive conditions, such as variables' endogeneity, an important number of outliers and missing values. The results found in the literature continuously showed these models achieved better results than traditional ones such as (Tavana, et al., 2018) logistic regressions (Angelini, Tollo, Roli; 2008)), and have also proven successful in credit scoring using only CDS data (Luo, Wu, Wu; 2016).

Furthermore, SMEs are more likely to fail due to short-term difficulties rather than long-term characteristics (Carton & Hoffer; 2006). In the context of their credit assessment, considering the constantly changing nature of many explanatory variables is therefore key, and deep learning models have proven successful in incorporating information on short-term evolutions of variables (Barboza, Kimura, Altman; 2017). The suggested improvement in credit assessment's accuracy despite restrictive conditions brought by deep learning techniques can have significant implications. For banks, small

improvements in predictions accuracy can largely improve their profitability while reducing their balance sheet's portion of Non-Performing Loans. As for SMEs facing financing problems due to their complexity or lack of data to hand-in, deep learning models could facilitate their credit profile evaluation, which would in turn give them access to funding (Huang, Liu, Ren; 2018).

1.3.2 Private Loans credit risk evaluation

Sirignano et al. (2018) have expressed the need to use this technology to evaluate mortgage loans by showing the highly non-linear relationships existing between borrowers' behaviors and risk factors, and proving the interactions existing between the explanatory variables. They found that prepayment events in particular, are significantly affected, looking at their relationship with the difference between initial mortgage rate and market rate. As for the interaction between explanatory variables, they illustrated this phenomenon with the impact of a borrower's FICO score [1] on the explanatory power of unemployment rate. Following these results, they questioned the use of more traditional models based on linear interpretations and suggested the use of deep learning techniques to predict mortgage loans' behavior. The neural network they developed on a very large US data base have proven very successful.

Additional research on private loan credit risk

Training a neural network on consumers' credit card data have also proven to give excellent credit risk predictions. Focusing on these transactions and excluding other data usually considered in credit scoring (i.e. socioeconomic data, loan balance and payment history, or credit bureau data), Kvamme, Sellereite, Aas and Sjursen managed to build a neural network predicting accurately mortgage defaults in Norway (2018). Eventually, to counter the problem of delicate and rare situations in credit scoring, normally assessed by human experts, an emotional neural network has been developed and once again has proven successful. Its two "emotional" responses, anxiety and confidence, change during the learning phase. As the loss function decreases meaningfully, anxiety decreases, and confidence increases (Khashman; 2011). This aspect of emotional neural networks allows for interpretation of data inputs with a degree of confidence. All the

[1] Fair, Isaac and Company. A data analytics company focused on credit scoring founded in 1956 and based in San Jose, California: https://www.fico.com/en/about-us (Last assessed in December 2018).

results justify the need to investigate further the use of deep learning technology to improve credit and liquidity risk evaluation and management.

2. Transition states – FILIPE, JOSÉ, ROBERTA

There are four possible states a mortgage can be in:

- Performing: all payments occur as predicted by the contract.
- Prepayment: A mortgage is considered prepaid when the borrower decides to partly or entirely (the one we will focus on the paper) pay in advance the principal of the loan.
- Delinquent: A loan is considered delinquent (or in arrears) whenever the payment is not made within one month from when the instalment was due.
- Default: A mortgage is considered in default when the cumulative amount in arrears is higher than three monthly installments (i.e. +90 days cumulative delay).

Section 4.1 will further explain how we applied this to our data. This is, how these classes were built and how observations were classified.

2.1 Default and Prepayment risk

An important role is given to the default and prepayment classes. These two statuses are extremely relevant to the lender. Whenever the mortgage is behind payment or is prepaid the lender experiences a disruption in cashflows from the missed payments.

Default occurs when the counterparty fails to meet its contractual obligations. The likelihood of this happening is called rate of default and it is one of the most important parameters to define the credit exposure of the lender. In this paper, the rate of default is calculated monthly, over twelve-month horizon (Sections 4.1.5 and 4.1.6). Thus, the monthly rate of default describes the likelihood of a mortgage to default within the following twelve months. The time horizon was set to twelve months as nowadays it is widely used in the financial industry for the calculation of credit risk and related

capital requirements. Moreover, the IFRS 9 requires impairment of financial assets to be measured as the expected credit losses over a twelve-month horizon[2].

Therefore, during the process to calculate the monthly rate of default, a new column with the twelve-month lagged mortgage status was added in order to have for each month the number of loans that defaulted within the next twelve months.

By the same token, whenever a mortgage is prepaid, the lender would lose either partly or entirely his future interest cashflows. With a decrease in current rates in the market, mortgages loans are paid off earlier in order to incur in lower interest rates by refinancing the loan, and the lender would have to deal with reinvestment risk. Prepayment is perceived as a financial risk as the investor would not be able to reinvest the cashflow at the same rate of return as the one locked in the mortgage and would have to use the current market interest rate. The mortgage can be partially prepaid in case the borrower wants to pay less in interest rates and prefer to pre-pay part of the principal amount.

A mortgage consists of a straight bond and an option that gives the borrower the right to prepay and refinance the loan at any time. The decision of prepay can be considered as a call option exercisable on the mortgage by the counterparty, giving the right to the borrow to redeem the mortgage before the maturity date[3]. The call option would be exercised whenever the value of the future instalments exceeds the value of the balance and the cost of refinancing the loan, both explicit costs, such as fees, and implicit costs, such as costs incurred when asking for another mortgage. However, as already seen in other studies conducted on prepayment risk, the behavior is unpredictable since it can be caused by other factors linked to the single borrower.

Understanding the behavior of prepayment would be profitable to the mortgagee, to decrease his exposure to prepayment, reinvestment[4] and liquidity risks. Liquidity risk is especially important for banks, which have to correctly estimate their

[2] "The rate of default under IFRS 9: multi-period estimation and macroeconomic forecast", Tomáš Vaněk, David Hampel, 2017
[3] "Modelling Prepayment Risk", J.P.A.M. Jacobs, R.H. Koning, E. Sterken, 2005
[4] "Modelling Prepayment Risk", J.P.A.M. Jacobs, R.H. Koning, E. Sterken, 2005

liquidity profile, strongly influenced by the maturity of their assets and liabilities[5], consult section 1.3 for more information on this subject.

[5] "Mortgage Prepayment Rate Estimation with Machine Learning", Taiyo Saito, 2018

3. Literature and new Variables

3.1 Overall analysis – FILIPE, JOSÉ

Our dataset comprises 56 different Variables, some with dynamic and other with static features. Some of these variables are extensively discussed and studied, in the mortgage risk related literature, due to their explanatory power regarding the borrower behavior.

In addition to the provided variables, Moody's recommended the creation of new interaction variables, namely Loan Age and Current LTV that were proven to have high explanatory power in the logistic regression estimation, tackling the same issue.

The variables try to capture the different types of risk a lender may be exposed to. Von Furstenberg (1969) and Gau (1978), group the risk factors in three areas: borrower and property developed by, who considered three major determiants: loan financing, borrower, and property. Financing risk factors try to gauge potential disruption to payments originating from the loan contract, such as the loan amount, balances, term, loan-to-value (included in our modeL: *CurrentLTV, PaymentFrequency, Completion, LoanTermInMonths* among others – see appendix C for full list of variables)

Borrower risk variables try to capture the risk related to borrower's information, such as his age, the occupation or the income. In fact, income is established as an important influence factor, translating the level of wealth, one of the most important characteristics when assessing borrower risk (ability of the borrower to meet the commitments agreed) (Gau, 1978).

Finally, property risk variables try to capture the influence that the underlying property can have on the performance of the mortgage, for instance, how the borrower behaves as his house gets more deteriorated. *Property type* or *Valuation Volatility* are examples of variables linked to property risk we are using in our model.

Vandell, K. (1978) and Webb, Bruce G. (1982) extend on the variables used, giving also importance to the relationship between instalments and income (included in our model as well – *Installpropincome)* income sources as risk of delinquency.

The project was mainly based on Sirignano et al (2018). On it, data related to borrower, property and loan financing characteristics is used as well as local and national economic variables such as unemployment and lagged default rate. Some of these variables change during the life of the loan, others remain constant. Our dataset also covers borrower, property and loan financing characteristics with variables such as *AgeOfBorrower, PropertyType* and *CurrentInterestRate* for each category respectively. As in Sirignano et al. (2018), some refer to the origination of the loan and remain constant through the life of the loan others are updated with every observation. Economic variables such as unemployment are not included, with the date (as distance form year 0) being the only proxy to mirror the economic reality of the particular year and month of an observation. It is unclear if the inclusion would be of much value. One of the contributing factors for these macroeconomic variables to be so significant in Sirignano et al. (2018) is tied to the large time period the observation range in and the sample, on which we worked on, ended up ranging only from 2013 to 2017.

Like Sirignano et al. (2018), we also have a data's static-dynamic division, following what was done in Moody's dataset. We have some variables that were evaluated at mortgage's orgination and others that are change on a monthly basis. For instance, *Interest rate type* or *Geographic region* (of the property) are considered static variables while *Current LTV* or *Distance to Maturity.*In fact, we also transformed the static variables in dynamic by interacting different types of variables.

However, contrary to Sirignano, we do not study the macroeconomic factors, in our analysis. As referred before, the only economic factors that may have implications in our neural network are: the *YrM*, the observation month, that may be influecned by macro factors, and the United Kingdom House Price Index, used to build the *Current LTV*. (section 3.2.1 – Current LTV).

In fact, as explained in section 1.2, a loan-level model relies more in borrower-level variables than in macroeconeomic determinants of default. Although, the inclusion of macroeconomic explanatory variables in the model increases the model's overall fitness and performance. Moroever, there are evidences of some macroeconomic variables having the highest explanatory power in borrowers behaviour, like the state

unemployment rate, or the interest rate margin, the difference between borrower's and market's interest rates. (Sirignano et al.; 2018)

3.2 Created variables

3.2.1 Ability to cover the loan with the property value – ACHILLE E FILIPE

The group wanted to capture the impact of the value of the house on the loan states' probabilities. We believed considering this amount in proportion to the remaining loan value to be paid – ending pool balance – should allow the capture of how much a borrower is covered by the value of her house. As a result, we analyzed the current Loan-To-Value as well as a similar variable considering the last official valuation. We thought of creating these variable after having found in the literature that, as the value of the house increases, the rate of being performing should increase (Bian, Lin, Liu; 2018). In addition, high changes in the property value largely affect the default probability (Kelly, McCarthy, McQuinn; 2014).

Current Loan-to-Value Ratio

Original LTV, defined by the loan value divided by the house price, both at origination, was one of the static variables initially considered. However, this ratio is not considering the macroeconomic factors over time, which have an implication in the house price, neither the fact that loan value is changing over time, according to what has been repaid. In several papers, it has been emphasized the impact of house prices and home equity accumulation in the default event.

For these reasons we created the Current Loan to Value (*Current LTV*), which transforms original LTV in a dynamic variable, capturing the loan value and the house price, in each month.

$$Current\ LTV_t = \frac{Ending\ Pool\ Balance_t}{Current\ Valuation_t}$$

The *Current LTV*, in each month, takes the current loan valuation, defined by the loan's ending pool balance and divides it by the Current House Valuation, which was computed considering the UK Government House Pricing Index (HPI), considering the following formula:

$$Current\ Valuation_t = Valuation_{t-n} * \frac{HPI_{t-n}}{HPI_t}$$

We feel important to refer that it was considered the overall UK HPI, regardless houses' specific locations.

The Current LTV, according to several studies, is one of the most important factors in explaining the borrower behavior. This variable tries to capture the effect of the house finance strategy, in each month - the percentages of debt and equity financing the house - in the loan states' probabilities. The LTV has a positive impact in probability of default and a negative impact in prepayment probability. (Campbell, Tim S., and J. Kimball Dietrich; 1983)

In Von Furstenberg, George M. (1969), although not developed, it is suggested to include a cross-term variable relating the income level and the LTV ratio, being expected a higher probability of default for borrowers with low income values and higher LTV ratios. This would be interesting to do.

However, the *Original LTV* also has been considered as a significant variable to explain borrower behavior, in particular default. Recently, Campbell, John Y., and Joao F. Cocco (2014), incorporated this original ratio in their household's utility-maximization model, which tries to predict the default decision. According to them, "a higher (initial) LTV ratio increases the probability of negative home equity and mortgage default".

Theoretically, the Loan to Value ratio compares the mortgage amount with the appraised value of the respective property. A higher mortgage relatively to the house price will make the borrower more dependent on debt to finance his house, and consequently more susceptible to do not comply with his obligations. Therefore, following the related literature, the lenders usually consider loan with higher Loan to Value ratio riskier than loans with lower LTV and, in order to protect themselves against

the exposure, they will increase the borrowing cost, meaning they will set higher interest rates on mortgages with high LTV ratios.

Considering a small sample of loans, we computed the median of Current LTV, given us approximately 0.46, meaning that, on average, within this sample, each mortgage, on each month, supports around 46% of the house's value. We chose to use the median instead of the average because it is better excluding outliers. Then, we divided the observations in two sets: High LTV and Low LTV, considering if they have a Current LTV above or under the median, respectively, and we computed the observable rate of each state, for both LTV levels, in each month.

On the one hand, we can observe that Current LTV seems to follow a pattern for borrower behavior, namely for performing and defaulting.

Mortgages with a lower loan to value ratio had, on average, higher performing rate, i.e. keep paying on time, and lower default rates, compared to higher LTV ratios mortgages, as it can be seen:

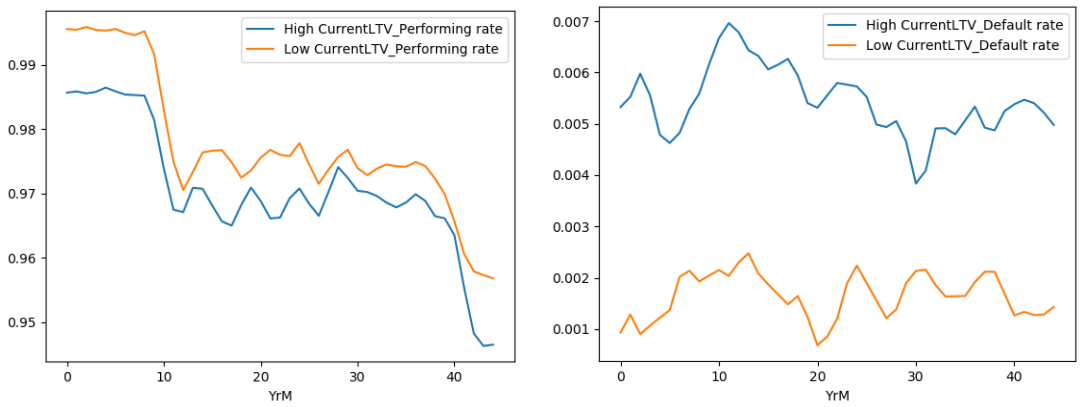

Figure 1

This was theoretically expected: a borrower less dependent on debt to finance his house will have a higher probability of complying with the payments, and lower probability of default the loan.

In fact, for default the aforementioned difference looks more prominent, in other words, the gap between high LTV and low LTV functions is broader than for any other state, which emphasizes the explanatory power of default probability by Current LTV, i.e., Default levels are more sensitive to changes in LTV (Von Furstenberg, George M.,

1969). Important to refer that a higher LTV does not mean that the borrower is more indebted, in absolute terms, instead he is more dependent of debt to pay his house.

On the other hand, for prepayment and delinquency rates the distinction between LTV levels does not look so clear. Nevertheless, when looking to the scatter plot, which plots the Current LTV against the prepayment rate, we can observe a downward trend: borrowers with low values of Loan to Value ratio paid their commitments sooner than expected, in other words, loans with low LTV ratios were prepaid more frequently than loans with high values of LTV ratio. From delinquency scatter plot we cannot make a big inference, maybe a slightly upward trend for LTV ratios lower than 1, pointing a more frequent delaying in payments, for higher LTV values.

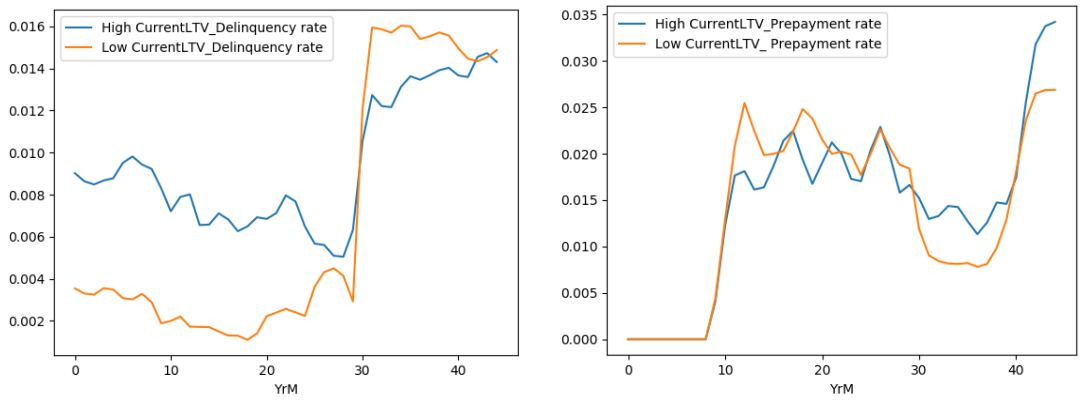

Figure 2

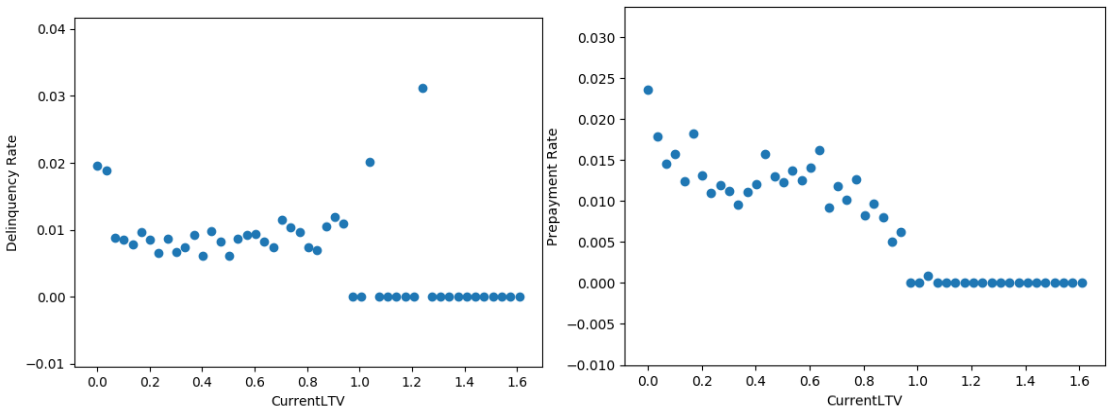

Figure 3

A studied conducted by PI Analytics, on a 50 thousand loans sample originated during 1999-2013 from Freddie Mac Loan Level Dataset, related the mark-to-market LTV (our *Current LTV*) with the default and prepayment probabilities. For LTVs under 1 the prepayment rate is flat near 70%. As the LTV increases, the prepayment rate starts to decrease. For LTVs above 2.25, the prepayment rate is virtually 0%. The default rate has an analogous behavior: for LTV ratios under the 1 threshold, default rate exhibits a constant flat behavior, near 0%. For LTV above 1 the default event starts to be more frequent, "as the amount of loan increases relative to the value of the house, the willingness of the homeowners to default on their mortgages increases", reaching a 100% default rate, for rates above 2.25.

Besides this research, there are other papers that empathize the fact that LTV effect on states' probabilities kicks in after a certain LTV threshold (Li, M. 2014).

Testing Current LTV Sample 2 we can observe the same patterns found in-sample analysis, supporting the literature, especially for default and performing scenarios. The graphs can be seen in Appendix B – Figure 21.

The *Current LTV* is one of the main determinants of loan states' probabilies, therefore it will be considered in our machine learning model

Ability to cover the loan with property value – LTV with last official valuation

This variable differs from current LTV as the value of the property taken for current LTV takes the original valuation of the house and changes it according to the house price index, whereas here we take into consideration the last official valuation amount of the house:

$$Ability\ to\ pay\ the\ loan\ back\ with\ the\ property = \frac{Ending\ pool\ balance}{Current\ property\ valuation}$$

- called IncentiveToSell on the python notebook.
- A high ratio = a low ability, a low ratio = a high ability.

We called it ability to sell the property, from the fact that lower ending pool balance to house value would make the borrower able to sell the property to pay the loan back.

We divided the data into two groups to assess this relationship: the ones with higher ability to sell their house, and the ones with lower ability. The threshold used to divide them was the ratio's median value.

Overall, this variable analysis gave us expected results for the performing state and default state: a much higher default rate in the low ability category and a strong positive correlation between the two categories in the performing state. The high ability category outperforming the low ones. The graph below illustrates this.

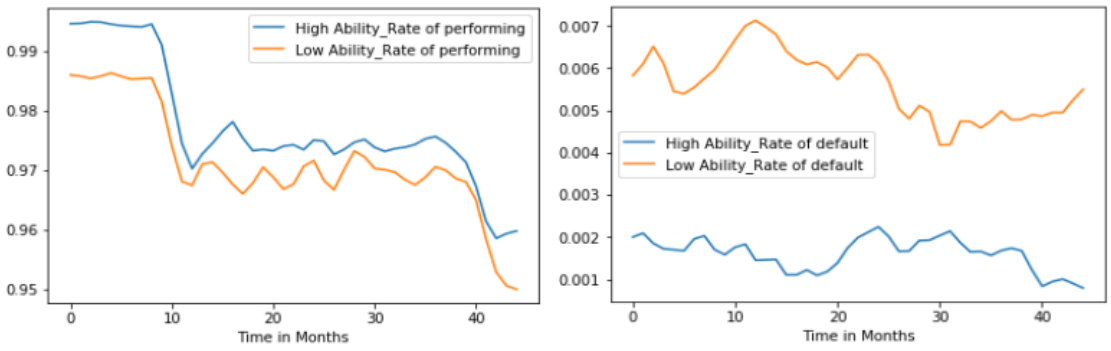

Figure 4

These relationships justify the need to consider the examined variable in our neural network for its explanatory power. These observations have been confirmed when tested In Sample 2 (Appendix B – Figure 23). However, we noticed a surprising effect of this variable's categories on the delinquency rate. As for its relationship with prepayment, the overlapping of both categories weakens the explanatory power observed before.

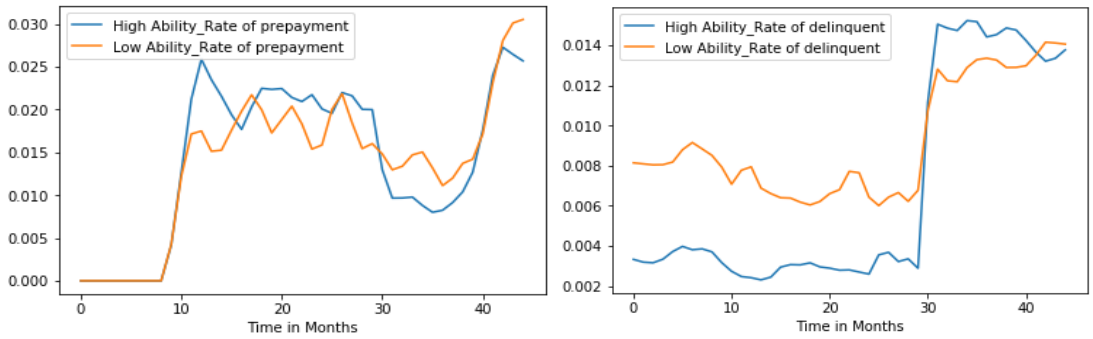

Figure 5

Looking at the relationship with delinquency in particular, one can see an expected effect for the first two thirds of the time. Yet, reaching the 30th month, when an important jump in delinquency occurs – probably caused by a significant drop in UK

households savings (OECD; 2018)(BBC; 2017), the high ability category's delinquency rate gets higher than the low one. As a result, this variable's ability to explain delinquency is relatively strong, but is very sensitive to chocs. As for prepayment, the graph above suggests a weak ability to explain this state. Nevertheless, despite these weaknesses, this variable use in our neural network is justified by its explanatory power for the performing and default rates.

3.2.2 Time elapsed since evaluation – ACHILLE E FILIPE

Additionally, the group decided the effect of time should be considered further. Indeed, most of mortgage and other personal credit score providers insist on the necessity to continuously update credit holders' score (Equifax, 2018), (Experian, 2018), (TransUnion, 2018). As lenders usually report monthly data on their borrowers, the borrowers' credit score is therefore adjusted, and important fluctuations can happen (NerdWallet, 2018). Having in our data inputs many dates regarding both the borrower and the property's valuation, we decided to create "distance" variables capturing the effect of time that passed from a certain valuation event until today. We considered the following variables:

$$\text{Time elapsed since last property valuation} = \text{Today's date} - \text{last property valuation}$$

- called *DistanceFromValuation* in the python notebook

$$\text{Distance since original property valuation} = \text{Today's date} - \text{Original vlaluation date}$$

- called *DistanceFromOriginalValuation* in the python notebook

$$\text{Distance since credit evaluation} = \text{Today's date} - \text{Bureau score year}$$

- called *DistanceFromEvaluation* in the python notebook

$$\text{Distance since last loan status} = \text{Today's date} - \text{Loan status' date}$$

- called TimeSinceStatus in the python notebook

Looking at these variables' relationships with the different mortgage loan's states (Appendix B), together with the sole effect of time, we noticed recurring trends that we expected, such as declines in performing loans' rates and (in most cases) increases in the default rates. However, what is more important in the context of our analysis with a multilayer perceptron is that the effects of the variables' as well as the time effect are not perfectly correlated with each other and have very distinct intensity and volatility, with relationships sometimes linear or non-linear (Appendix B). Therefore, each event's time interval's effect having their own specificities, their consideration represents relevant inputs for our multilayer perceptron. The next paragraphs explain the variables' specific effects noticed.

Looking, first of all, at the distance since property valuation, one can notice a very intense effect of time in first 5 months. The two graphs below illustrate this effect on the performing rate and the delinquent rate:

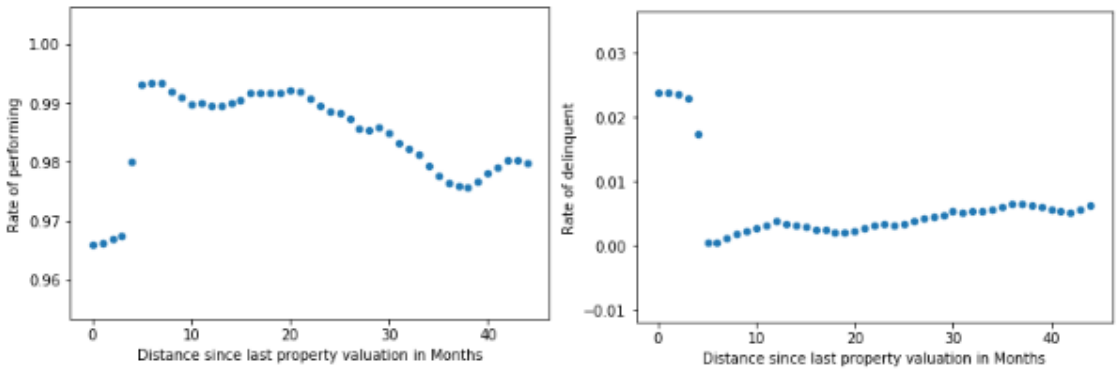

Figure 6

Here a significant shock can be seen, which intensity and direction were not expected (one would expect performing rates to decline and delinquency rate to increase over time). After which the recurring trends return (decline with time for performing rate and increase in time for delinquency rate). This effect has shown consistent when tested Sample 2 (Appendix B – figure 25). Moreover, this shock can also be distinguished by its singularity. It was not found in either the relationship between the distance since original property valuation and the loan states' rate, nor in the relationship between the loan states and time. On the contrary, much different effects are visible for these two variables, as illustrated by the graph below:

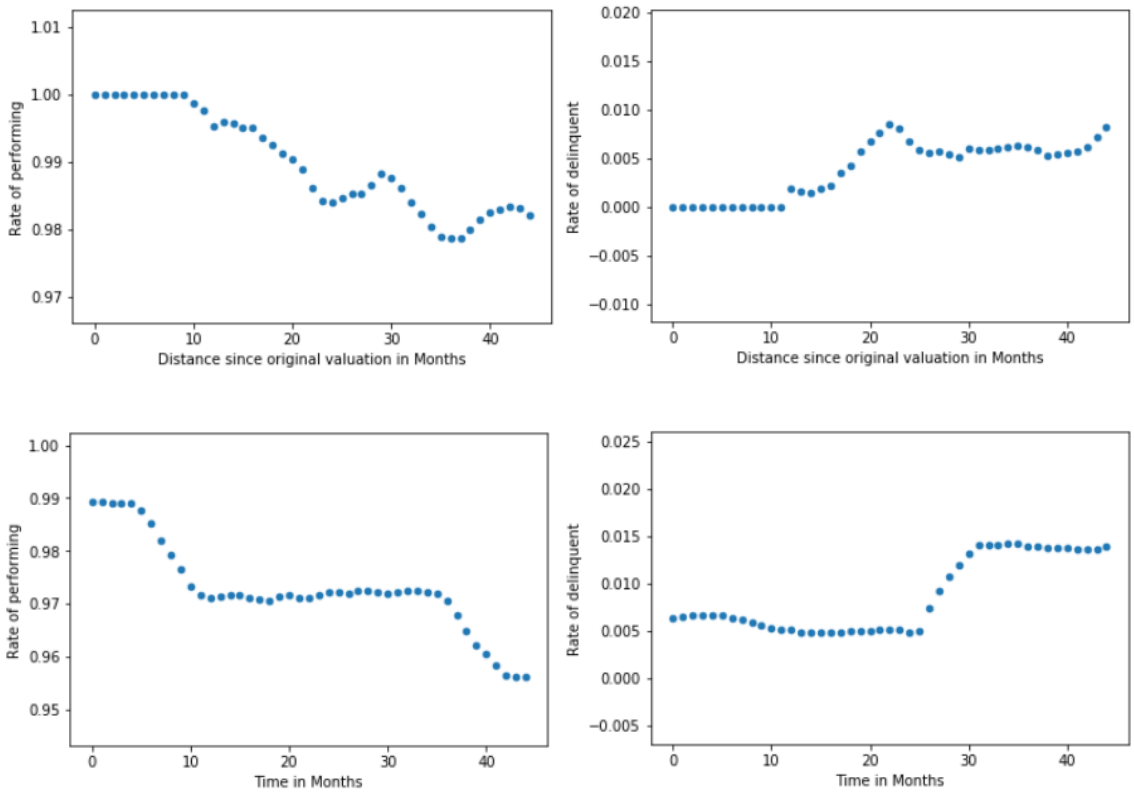

Figure 7

Looking at the relationship between time and performing rate, one can even observe an opposite effect at the same time with less intensity. This effect has also proven to be consistent when tested in Sample 2 (Appendix B – Figure 33). As for the relationships between the distance since original valuation and the loan states, together with the relationship between time and delinquency rate, one can see an effect happening much later with a smaller intensity. Consequently, the consistency of this intense effect, together with its singularity when compared to other related time interval factors suggest that it may not simply be noise in our data sample. It therefore carries relevant information for our neural network's input.

Holistically, we noticed these time intervals between different valuation and evaluation events have their own relationship with the loan states, either linear or non-linear, which appeared to be consistent in most cases when tested in Sample 2. We therefore concluded that they should be included in our model as the information they bring is relevant.

3.2.3 Number of valuations per Loan

The group also created a variable counting the number of times the loan has been valued. We wanted to assess whether trends could be observed between loans valued several times and a particular state. The combination done to create this variable was the following:

$$Number\ of\ valuation\ per\ loan = count\ of valuation\ date\ per\ loan$$

- called ValuationVolatility in the python notebook

Loans were valued from one to five times, the vast majority of them were valued once (see Appendix B – figure 34). No major trends were found, other than loans valued several times, if not classified performing, tended to be delinquent (Appendix B – figure 36).

3.2.4 Loan Age and related variables – FILIPE E HENRIQUE

The mortgage time path is quite important to understand the probability of a default or a prepayment. Three variables were created, in order to capture this effect of the mortgage track history in the probability of each state.

For each observation, we obtained the *Loan Age* which is the age of the mortgage, in months, the *Distance to maturity*, which gives the number of months until maturity and the *Percentage of Loan Completion*, which measures the loan's age as a percentage of its length.

$$LoanAge\ =\ Today's\ date(YrM) -\ LoanOriginationDate$$

$$DistanceToMaturity\ =\ DateOfLoanMaturity -\ Today's\ date$$

$$\%\ of\ Loan\ Completion\ =\ \frac{LoanAge}{(DateOfLoanMaturity -\ LoanOriginationDate)}$$

Loan age, in particular, is one of the commonly studied determinants. For instance, the survival analysis, explained in section 1.2.3, bases its entire framework on the age of

the mortgage when there is a change in its state, computing a time conditional probability.

Some studies documented a positive effect of mortgage age in the probabilities of default, delinquency and prepayment; and a negative effect of mortgage's age squared in the three referred probabilities (Campbell, Tim S., and J. Kimball Dietrich; 1983). This paper, even though it is quite old, also shows that excluding the age-related variables from the regressions generates poorer model performance (less significance), than when we include them, what proofs the importance of the loan age in explaining the states' probabilities.

Most of the studies documented a non-linear relationship between mortgage's age and probabilities of default and prepayment. (Von Furstenberg, 1969), but specially for default rates. In fact, "defaults display a rise-then-fall pattern as mortgage age", in the first years of the mortgage, it is common to have low default rates; as the time passes, the default frequency increase; however, it decreases again when the mortgage gets closer to its maturity.

Analyzing our sample, we can see a behavior between the both patterns, evidenced in the aforementioned literature.

In respect to default rate, we can see an approximation to the non-linear quadratic pattern described in literature, especially in the right tail, i.e., for older loans. Mortgages younger than 50 months and older than 250 months have a default rate close to zero; while ages between 50 and 250 months have default rates around 0.5%. For prepayment it is not so obvious, but the rate seems to be lower for loans older than 200 months (right tail), as expected, form literature; however, there is not a lower prepayment rate for younger loans. Loans younger than 200 months seem to have prepayment rates between 1 and 2%.

For performing and delinquent rates, the outcome is expected, although not extensively developed in related literature. Performing and delinquency rates follow the previous non-linear behavior, with a more prominent effect in older loans, like prepayment rate pattern. Delinquency seems to have a diminishing in older mortgages, while Performing has an upward trend.

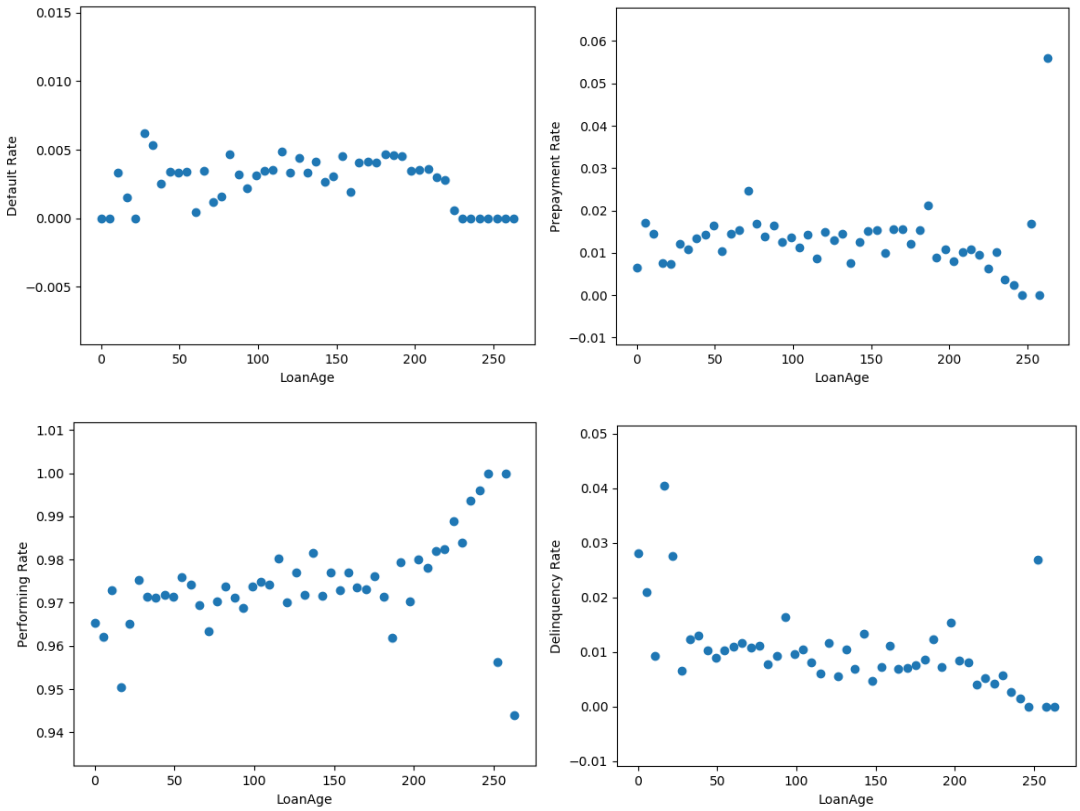

Figure 8

The distance to maturity is the opposite of loan age. Young mortgages have high distance to maturity, while old mortgages are closer to maturity (low distance). We seek to analyze how the proximity to each loan's maturity dates influence the probability of them defaulting or getting pre-paid. Therefore, this variable tries to capture how the borrower's decision is influenced by his mortgage's proximity to termination.

The expected behavior should be similar of what happens in Loan Age: Mortgages too far or too close from maturity will have lower probabilities of being prepaid or defaulted, than mortgages with intermediate distances to maturity. At the beginning of the loan there are less incentives to prepay or default. Over time, there is a higher probability of a change in financial situation (positive or negative) that leads the borrower to delay, miss payments or to pay installments sooner, increasing the probability of a default or prepayment. Closer to maturity, the incentives to repay the mortgage sooner or stop paying it are lower, again.

Overall, from our in-sample behaviors for the four states the patterns are not as evident as Loan Age outcomes, however it can be observed the right tailor evidence as

before, i.e., older loans exhibit low delinquency, default and prepayment rates and high performing rates. The default scatter plot seems to have the clearest pattern, similar to what has been described in the aforementioned literature, what emphasizes the explanatory of this variable to this particular probability.

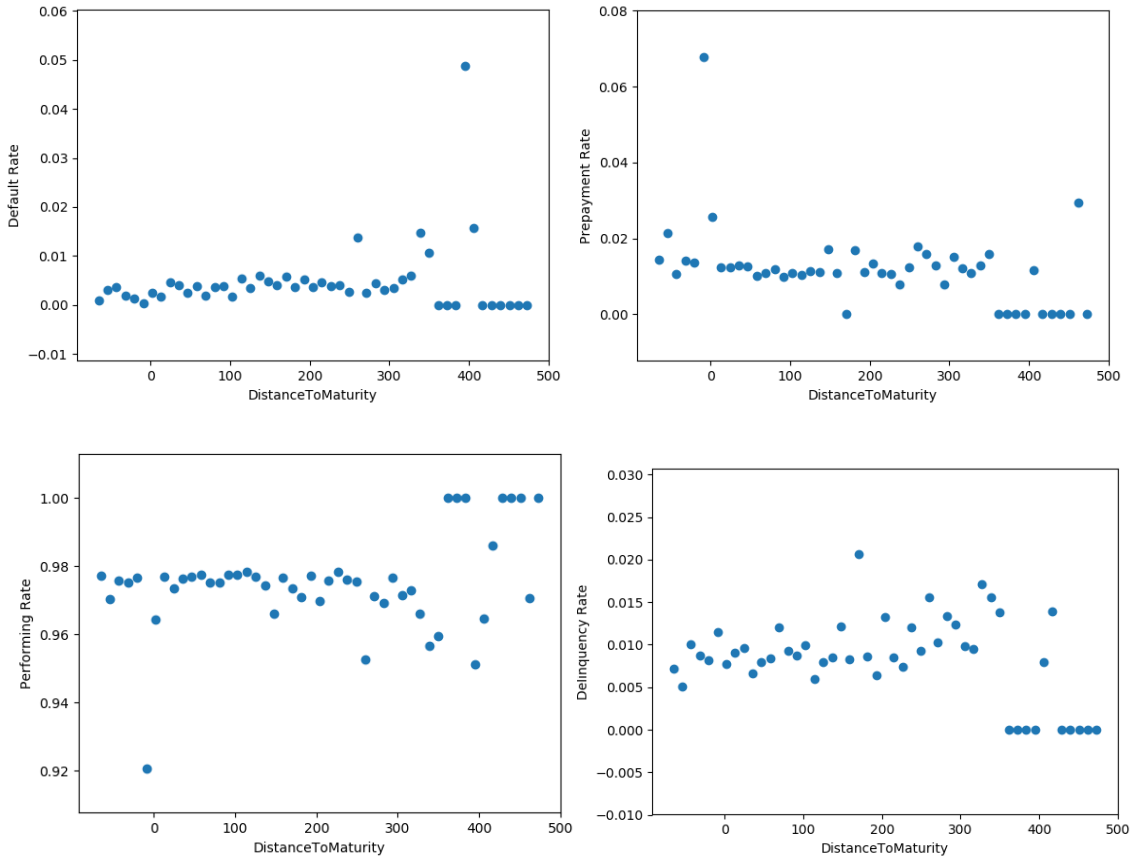

Figure 9

The negative values that we've found for this variable are most likely justified by a mismatch between the contractual and real maturity, i.e., some loans don't get paid in full at the maturity date, and the loan term requires to be extended.

In turn, the percentage of completion is the loan age in percentage of its term (maturity). There is not an extensive literature on this, but it is expected to follow the same pattern as the loan age: a loan with a higher (lower) completion percentage can be seen as an older (younger) loan.

The in-sample behavior for loans under 100% completion appear to describe a parabolic format, described in literature, although some jumps in the series, especially

when completion is around 0%. Again, the default rate seems to have the clearest pattern. The graphs for the four states can be observed in Appendix B – figure 39.

It was performed the same analysis, for the three variables, in a different sample (Appendix B - figures 37, 38 & 40), that generated similar results, especially for *Distance to Maturity* and *Percentage of Loan Completion.* Regarding *Loan's Age,* even with a big dispersion of the observations, the patterns can be deducted.

The three variables, specially the *Loan age*, seem to be significant in explaining borrower behavior, namely prepayment and default and, therefore, important to include in our neural network.

3.2.5 Loan Balance related variables - FILIPE

This section will comprise two variables we have created, related to the Loan balances. One is the *Distance to maximum balance*, how far from limit balance is the current loan, and the other is *Percentage of loan paid*, which measures how much of the loan has been paid, in percentage of initial amount.

$$Distance\ to\ Maximum\ Balance = Maximum\ Balance - Ending\ Pool\ Balance$$

$$Percentage\ of\ the\ loan\ paid = \frac{(OriginalBalance - EndingPoolBalance)}{OriginalBalance}$$

These variables are not extensively developed, in literature, for one reason: the high correlation between them and time, measured by *Loan age*. A low distance to maximum balance and a low percentage of loan paid are correlated with younger loans (lower loan age), while older loans are more susceptible to be far from the maximum balance and to have a higher percentage paid. Through conceptual approaches to mortgage risk, described in section 1.2, it is difficult to separate the effect of the loan time path from the real effect of the variable in the states' probability. For instance, the impact of the percentage of loan paid in default probability can be due to the loan age (passage of time), which is provoking the decreasing in the amount of loan to pay, and not due to what is missing to be paid. Nevertheless, using the neural network, this issue should not be problematic.

From these variables we try to capture how the borrower's exposure to debt financing influences his ability to pay the loan, in other words, we will try to understand how the borrower will behave in accordance with his dependence from debt, when compared with his eventual maximum debt exposure (through *Distance to maximum balance* analysis) and when compared with his initial debt exposure (through *Percentage of loan prepaid* analysis).

The *Distance to maximum balance* (DMB) can be positive or negative. A positive value means that what is left to pay is lower the maximum loan amount the borrower could eventually get, and therefore less risky to the lender, since the borrower is under his debt threshold. In the opposite way, a negative distance to maximum balance represents a high risk to the lender since the amount borrower has left to pay is higher than the maximum amount of debt the borrower supposed could contract. Literature related to the debt levels, states that borrowers above their debt thresholds have, on average, a lower probability of performing and prepay, and a higher probability of default and being delinquent, than borrowers under their debt limits.

Overall, the findings from the data analysis sustain this relation, especially for default and delinquent, where we can see that negative DMB series is strictly above the positive DMB, meaning that over indebted borrowers are more likely to be in arrears or default the payments than "under indebted" borrowers. For prepayment and performing rates, we cannot observe a strict dominance of one series, although as expected positive maximum balance, this is, borrowers under the debt limit performed and prepaid their mortgages more frequently than borrowers above their debt maximum.

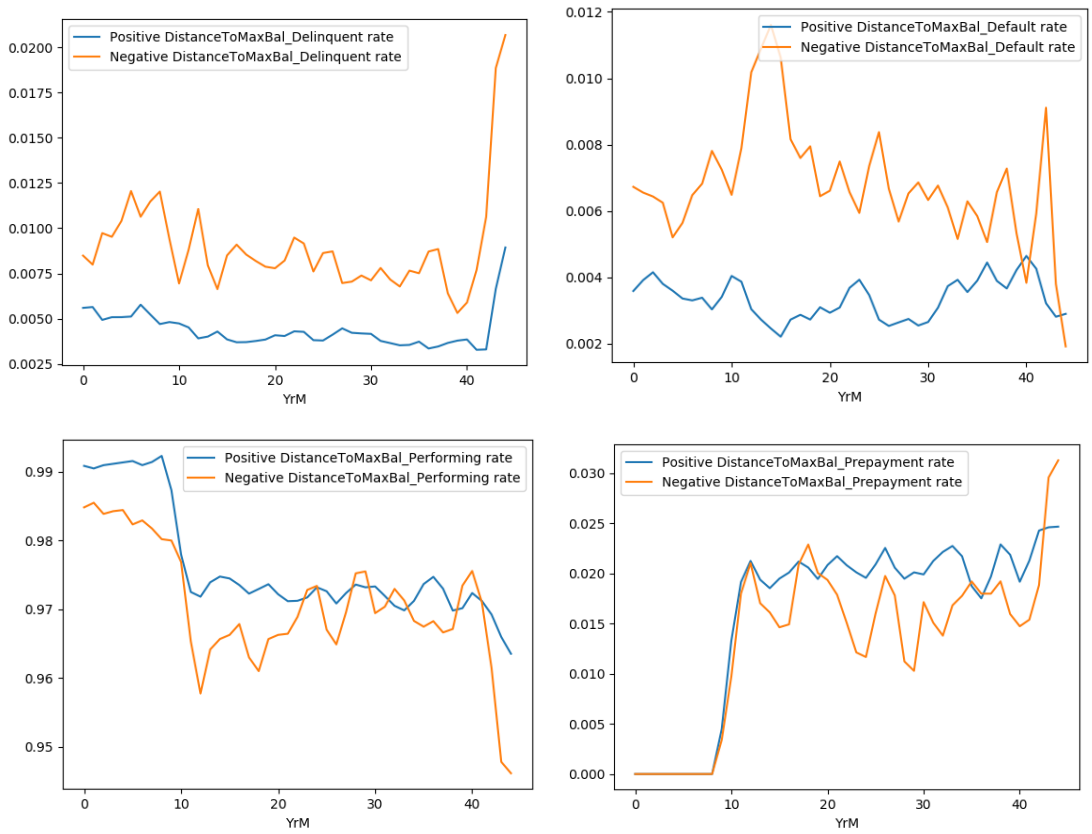

Figure 10

In turn, the Percentage of the loan that has already been paid is expected to be a determinant of probability states. As the borrower pays more and more of his loan, the amount of debt in relation to the initial amount is becoming lower, what incentives the borrower to keep paying the loan. A higher percentage of the loan paid should explain lower default and delinquency rates, and higher performing rates. The rate of prepayment is expected to be higher for mortgages that have still a high amount outstanding, as the borrower has more incentive to prepay, avoiding paying more interest, in the future.

Similarly, to DMB variable, our findings followed the theoretically expected, for default and delinquent, but also for performing. Mortgages paid above the median (22%) are more likely to perform and less likely to default and fall in delinquency. Regrading prepayment, we cannot see clear difference between the two percentage paid levels, they seem to have similar prepayment rates, over time. The borrowers who have repaid a higher percentage of the mortgage have similar prepayment availability as borrowers that repaid a lower percentage of their mortgages.

This can be observed in the following graphs:

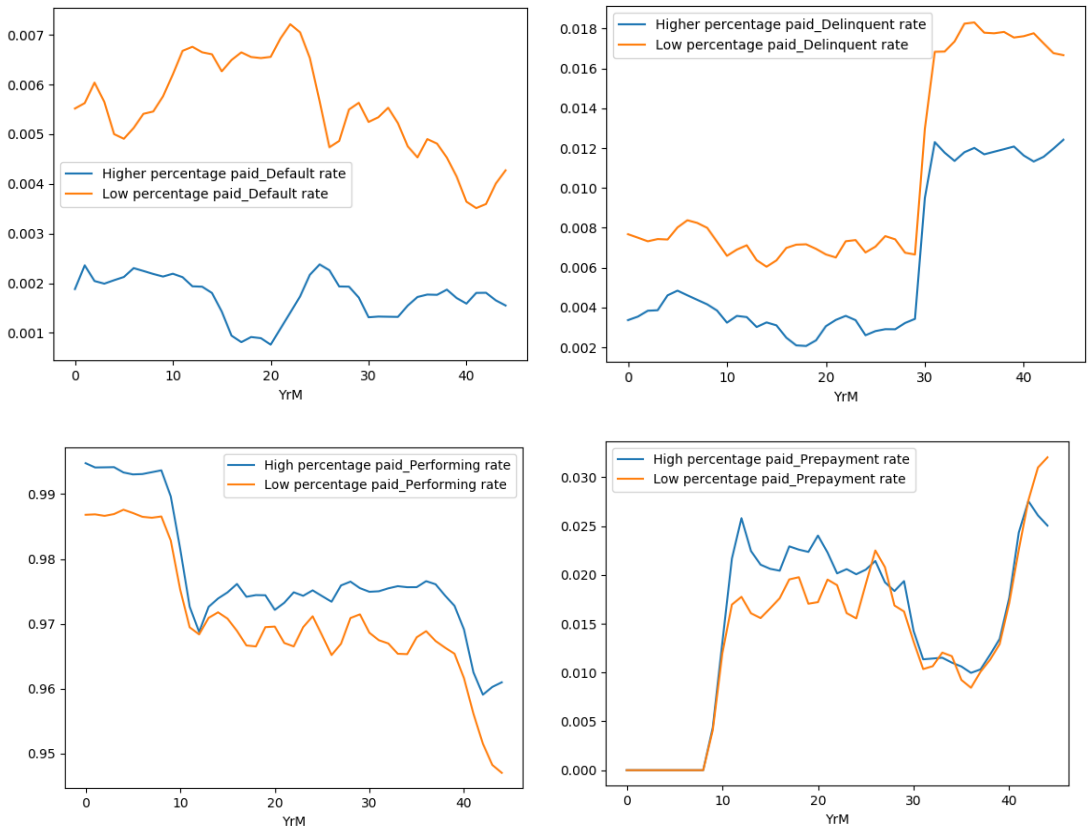

Figure 11

From the sample 2 testing, for both variables, the same results have emerged, as it can be seen in Appendix B, figures 41 & 42. *Distance to Maximum Balance* seems to have a strong explanatory power for default and delinquency, whilst *Percentage of Loan Paid* seems to be significant when explaining default, delinquency but also performing.

As referred above, the neural network must mitigate/solve the issue relative the correlation among explanatory variables. Therefore, they must be included in our machine learning model.

3.2.6 Age of borrower – HENRIQUE, ROBERTA

AgeOfBorrower variable was created to investigate the effect of the age of the borrowers on the loan status probabilities. The age of the borrower was calculated by calculating the difference between the *YrM* variable and the *DateOfBirth* variable.

In recent years, mortgages with longer terms have become more attractive to borrowers, shifting the repayment of the loan to later periods. For example, in 2017, an

article in Financial Times referred to the fact that more than 1/3 of the mortgages originated in that year would not be repaid before those borrowers turn 65, while some lenders in the UK are now setting the maximum term length of mortgages at 40 years (the standard term length is 25 years)[6]. These extra-long mortgages are more affordable as the monthly instalments can be reduced to an affordable level. However, extending the maturity leads to higher interest rate fees and an increase in the likelihood of their ability to repay being disrupted by some unexpected events[7]. Overall, these effects can have a major impact on the loan status' probabilities.

This trend might be explained by several factors: people are now working for more years, low interest rates might make it easier for borrowers to comply with their mortgage repayments given their retirement incomes, high house prices in the UK make longer term mortgages with smaller payments more affordable, and people are marrying and having children at a later stage in their lives, which then increases the age at which households buy a house and get a mortgage. Nevertheless, some lenders and regulators still have doubts whether elderly borrowers will be able to make their payments with their retirement income.

In C.A. Ajayi (1992), there has been found a small, significant, positive relationship between defaults and the age of borrowers, while Jones (1993) has found a negative correlation between the age of borrowers and default.

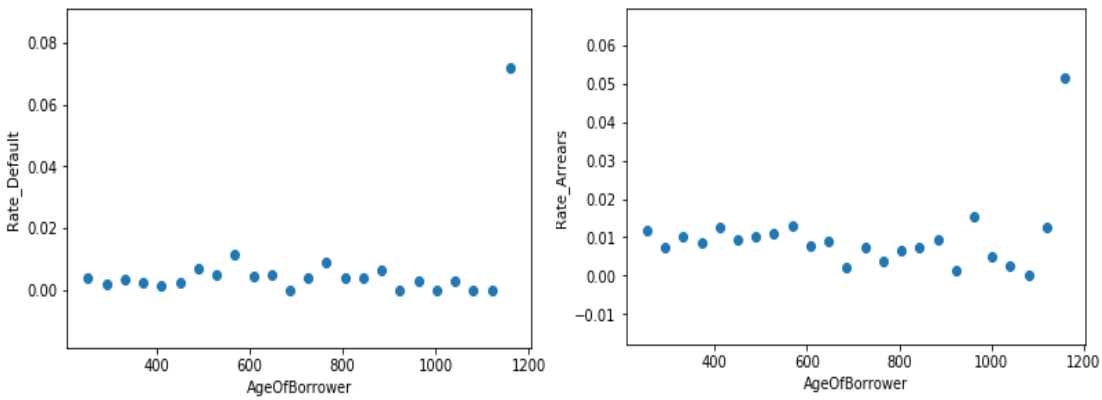

Figure 12

[6]**Financial Times; (2017); "Extra-long mortgages push up the age of borrowers"**; available at: https://www.ft.com/content/7711f8c8-7205-11e7-93ff-99f383b09ff9 (last assessed in December 2018).
[7] **Financial Times; (2017); "Extra-long mortgages push up the age of borrowers"**; available at: https://www.ft.com/content/7711f8c8-7205-11e7-93ff-99f383b09ff9 (last assessed in December 2018).

In both graphs, a non-linear relationship can be observed. Default rate is higher for people between 40 and 50 years old and between 70 and 80 years old. In the first case, the rate is highly influenced by other borrower specific variables like employment status, primary income or loan to value ratio. The Arrears rate is higher for younger people around 30 years old and older people around 80 years old. Higher arrears balance is expected for younger people as usually, being at the beginning of their career, they don't have very stable jobs and have lower primary income. By the same token, people that retire might have difficulties in being able to keep paying the instalments with just their pension. These non-linear relationships are caused by the more complex relation between the response variable and *Ageofborrower.*

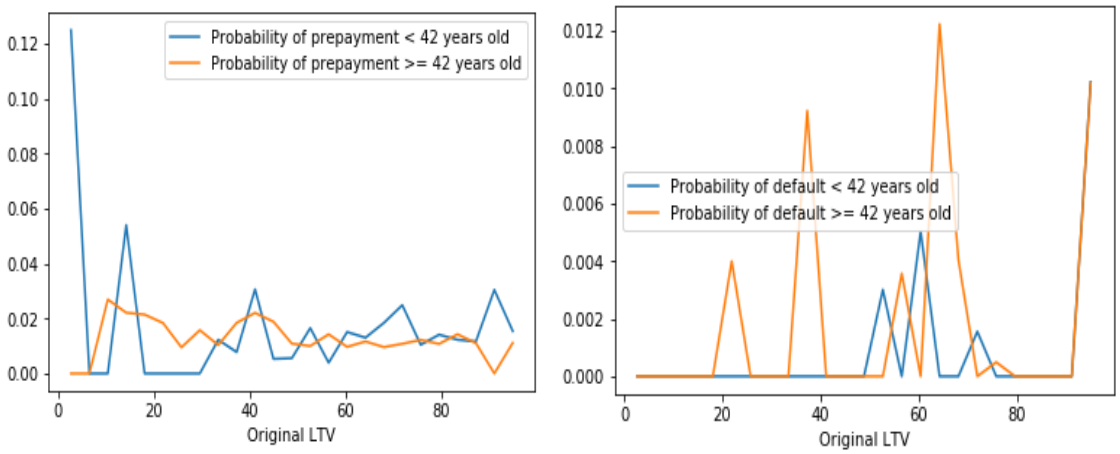

Figure 13

When relating *AgeOfBorrower* with *Original LTV*, by splitting the variable according to the median, we can see that older borrowers are generally less likely to prepay their loans than younger borrowers, with the prepayment rate decreasing as the Original LTV ratio increases, while the rate for younger borrowers is more volatile. Then, we can verify that older borrowers are generally more prone to default on their loans, being the magnitude of the Original LTV ratio a factor that seems to induce significantly older borrowers into default.

3.2.7 Income related variables – HENRIQUE, ROBERTA

The extent to which a borrower is able to meet their payments is another important factor that affects the rate of default and delinquency. To capture this effect, we created a variable called "Instalment as a Proportion of Income" by dividing the

amount of the monthly instalment by the monthly primary income, as described in the formula below. The monthly instalment was calculated by subtracting the Ending pool balance from the Beginning pool balance. Hence:

$$Installment\ as\ a\ Proportion\ of\ Income = \frac{Beggining\ Pool\ Balance - Ending\ Pool\ Balance}{\left(\frac{Primary\ Income}{12}\right)}$$

Looking at the proportion of the income that is represented by the monthly instalment can give a good representation of how likely the mortgage is to be in arrears or to be in default. If the instalment represents a high percentage of the income, the borrower has a higher likelihood of being delinquent or defaulting. Therefore, we would expect a high rate of being performing for mortgages with a low proportion of the income being absorbed by the instalment. In fact, prior academic work suggests that this ratio is generally expected to have a positive correlation with the default rate, as suggested by LaCour-Little and Malpezzi (2003), Kelly (2008) or Archer and Smith (2013).

Similarly, it is also important to define the relationship between the income and the cumulative amount of the instalments not paid and to capture their impact on the loan states' probabilities. As the amount of arrears balance increases, it would be more difficult for the mortgagor to be able to repay the loan in full, given his primary income. So, it would be expected that when the arrears balance increases the rate of being delinquent would increase, up to a point at which the arrears balance would be too high for the mortgagor to be able to continue paying and therefore defaulting on the mortgage. We thus created a new variable called "Arrears to Income", given by:

$$Arrears\ to\ Income = \frac{Arrears\ Balance}{Primary\ Income}$$

To analyze the first variable, "Instalment as a proportion of Income", the dataset was divided into two smaller samples (using 50% as the threshold ratio). Loans with high percentages of the income absorbed by the mortgage instalments are more volatile, and overall, they have higher rate of default and delinquency. We found interesting result especially when relating the proportion of the mortgage payments with the Current LTV of these loan to show how this variable significantly contributes for the loans' statuses.

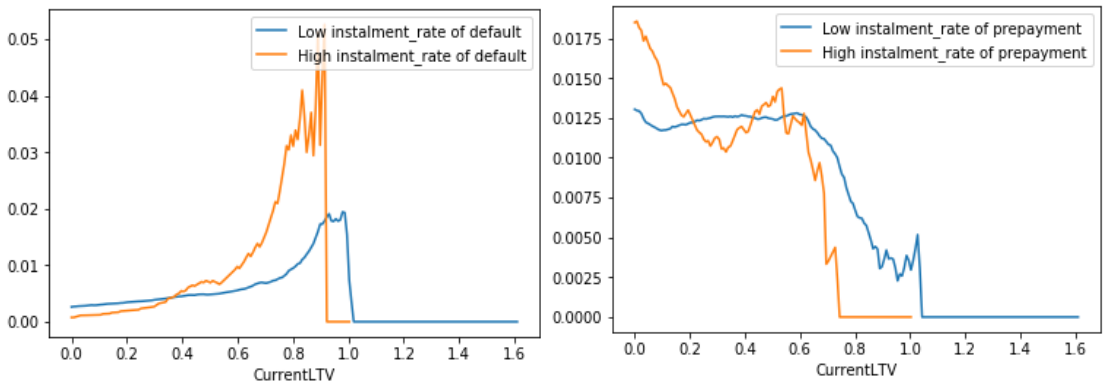

Figure 14

For both groups, as the Current LTV increases, so does the default rate in our sample. Nevertheless, the rise in the occurrence of defaults seems to be magnified in loans whose instalments are high relative to the primary income of borrowers. On the other hand, we can see on the second graph that, as the Current LTV increases, loans with high instalments relative to primary income seem to be much less likely to be prepaid for LTV ratio greater than 80%, a reason that further justifies the inclusion of this variable in our model.

The analysis conducted on the data set confirms this relationship, with a high rate of default for high values of "ArrearsIncome". Loans with a high ratio display a high

rate of delinquency and default and, when compared to loans with a low ratio, the gap in probabilities is very high. By following the reasoning presented above, we would expect the graph for the arrears rate to be the mirror image of the one presented above. This interesting effect on delinquency rate might be explained by the influence of other explanatory variables, such as the LTV ratio or employment status, which cause this non-linear relationship. Because of this interesting relationship, it is worth to include the variable in our neural network model.

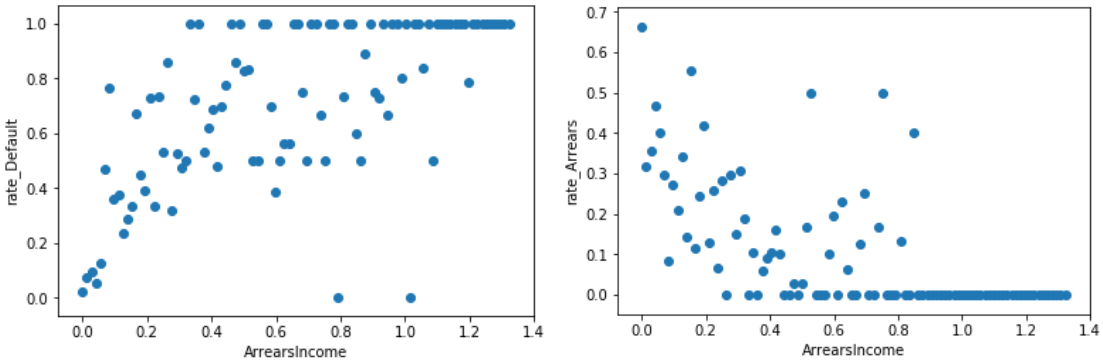

Figure 15

3.2.8 Balance in arrears in proportion to loan's outstanding value - ACHILLE

We thought of an additional way to capture delinquency's severity by considering the balance in arrears in proportion to the loan's outstanding value:

$$Balance\ in\ arrears\ to\ value\ outstanding = \frac{Balance\ in\ arrears}{Ending\ pool\ balance}$$

- Called ArrearsEndBalance in the python notebook.

The idea is to assess whether the remaining loan value to be paid can explain loan behavior when late payments occur. In our data sample, we noticed a non-linear relationship between this variable and the future loan states' rates.

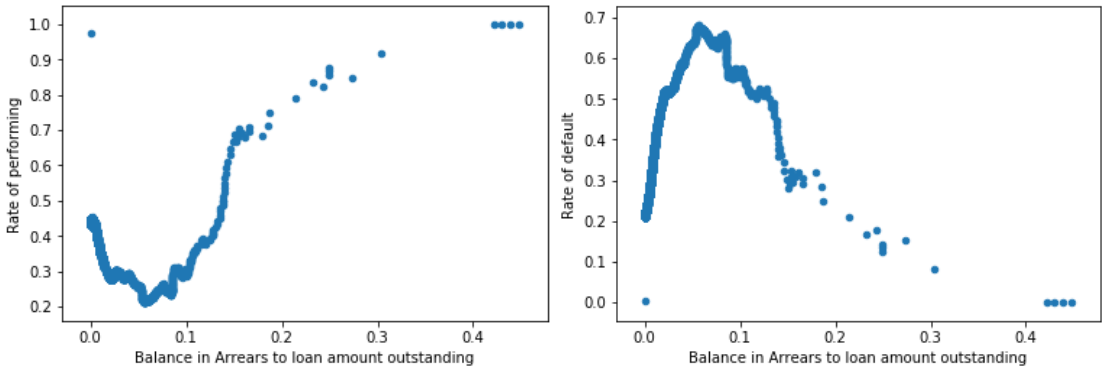

Figure 16

As seen in the graphs above, no linear trends can be deduced. However, this variable was judged to have the second most powerful predictive power by our neural network (see section 6.1). Therefore, only a neural network could capture the non-linear effect this variable has on the future loan states.

4. The Dataset – FILIPE, JOSÉ, ROBERTA

The dataset provided by Moody's Analytics is composed by over 2 million United Kingdom mortgages and around 44 million monthly observations. The dataset includes different kinds of mortgages, such as fixed rate, floating rate, capped rate and fixed with periodic resets.

The data set includes variables detailing characteristics of the loans. These variables can present static or dynamic information. The static variables are features registered at origination and they do not change over time. The dynamic variables show features that change over time and are 35 in total, including such variables as number of months in arrears, arrears balance, current interest rate, current interest rate index, current valuation amount and prepayment date. The variables can be further divided in continuous and categorical (also including dummy variables).

As stated in Section 2, the model's purpose is to predict the probabilities of each observation transitioning to four different states: Prepay, Performing, Delinquent and Default. Information that is given by the variable *LoanStatus.*

Due to the large size of the file, a small sample had to be taken in order to make easier the data preparation and model's estimation, due to our hardware constraints, in particular RAM amount and capacity to use GPU for estimation. For this purpose, 20 thousand random loans were selected, and their observations (about 760 thousand) extracted from the full dataset (to see shortcomings of this approach see section 9).

In the following sections, we will describe the steps taken to prepare the data.

4.1. Handling the dataset - FILIPE, JOSÉ, ROBERTA

4.1.1. Gap Flags

The first step of the data treatment process consisted in removing loans with missing observations. This decision was taken based on the idea that an incomplete time series ('series breaks') would not be suitable for panel data type model, at the time it was still not clear what model specification we would use and if it would require a time continuous stream of observations (not necessary since we ended up using a pooled

cross-section). Originally, these loans were easy to identify (already identified in the dataset) and remove.

4.1.2. Date transformation

Some of the variables in the dataset are dates. The date format was expressed with the year directly followed by the month i.e. 201801. This was an issue when applying mathematical operations to the dates. Therefore, the creation of a date function to transform each date (YrM) in a number was required. Each date was simply transformed in the number of months. For instance, February 2015 is represented by the number 24,182, which is 2015*12+2.

4.1.3. Data harmonizing

We take a sample from the full dataset, based on unique loan keys and since we are taking a 12-month lag, in order to guarantee that enough individual observations are extracted, we required a minimum of 24 consecutive observations per loan (12 observations once the Lag is created, explained in section 4.1.7).

4.1.4 Dealing with missing values

Some of the variables presented missing values or no information. Variables with more than 40% of missing values were removed from the dataset. We did not want to include variables with many missing observations, with fear of corrupting the results, but we tried to exclude the least number of variables possible. 40% seemed a good compromise.

4.1.5. Loan Status: Removal of categories

In the data set, the dynamic categorical variable called 'Loan Status' represents the account status of the loan, in each moment in time. This variable comprises six categories: 1-Performing; 2-Arrears; 3-Default or Foreclosure; 4–Redeemed; 5–Repurchased by seller and 6-Other. Loans with value 5-'Repurchased by seller' and 6-'Other' were immediately removed from the dataset, as they are not interesting for our analysis.

4.1.6. Loan Status: Prepayment and Default States

After removing the 'Repurchase by seller' and 'Other' categories, the remaining status need to be updated. Both the prepayment and default categories need to be

revised in order to make the loan status respect the criteria used to define the prepayment and default class, considering the criteria in section 2.

a) Prepayment Status creation

A mortgage is considered prepaid when the borrower decides to partly or entirely pay in advance (i.e. before contractual maturity) the principal of the loan. In our analysis we only study full prepayments, that is the is paid off in its entirety before the maturity date.

To accomplish this the following condition is imposed:7

LoanStatus = 4 (Redeemed) and *YrM* < *DateOfLoanMaturity*, the full payment (realized on *YrM*) happened before the contractual maturity (*DateOfLoanMaturity*).

b) Default Status Revision

A mortgage is considered in default whenever the borrower is three months delinquent, meaning that he falls three months behind with his payments. After defaulting, a mortgage cannot return to performing or delinquent states (default is an absorbing state). In the dataset, the variable *NumberOfMonthsInArrears* shows the cumulative number of months the borrower failed to pay, and it was used to identify defaults (*NumberOfMonthsInArrears* >= 3).

c) Other states

An observation is considered delinquent (or in arrears) whenever the months in arrears variable is between zero and three, and they are assigned to the delinquent status. Similarly, observations already assigned to the performing class and that did not suffer any change in the meantime are classified with the performing status.

After these changes, the new Loan Status is composed by four updated states: 0 – Prepayment, 1 – Performing, 2 – In Arrears or delinquent, 3 - Default.

4.1.7. Loan Status: 12-Months Lag

We're interested in predicting the state of a certain observation in 12 months, therefore *LoanStatus,* in 12 months' time, for a particular observation, is required to estimate the 'forward lag'. It can be easily obtained since we have a unique key for each

loan that allow us to keep track of the loan in several points in time. The process is similar to match each loan status at time t with the explanatory variables 12 months ago (trackable through the unique key):

$$y_{i,t} = f(X_{i,t-12})$$

Where $y_{i,t}$ is the probability of transition to a certain state, for each observation, at time t, and $X_{,i,t-12}$ the set of explanatory variables, at time t-12. The rates are calculated monthly, over a twelve-month horizon. When doing this, the observation amount will decrease, since there are observations that won't have a match, that is there are no observations with the same key 12 months after that observation. Therefore, during the process, the amount of observations will decrease to about 525 thousand (previously 760 thousand).

The full list of the variables used in our model can be found in Appendix C.

Once the loan state 1 year ahead had been calculated, the observations decreased to about 525 thousand, from here 20% of the data, roughly 105 thousand observations, were left out for testing purposes and from the remaining 80%, 70% was used for training and the remaining for validation, that is assessing the progression of the loss function and accuracy measure with every epoch on a set other than the one the model is using for training, in an effort to avoid overfitting – memorizing particular patterns of the training set, therefore not generalizing well in other datasets (poor out of sample performance).

5. Model – ACHILLE, JOSÉ

The following section specifies the model and its procedures to obtain the desired output (probabilities for the different states).

Within the aforementioned sample, 20% of the data, roughly 105 thousand observations, were left out for testing purposes and from the remaining 80%, 70% was used for training and the remaining for validation, that is assessing training performance on a set other than the one the model is using for training, in an effort to avoid overfitting – memorizing particular patterns of the training set, therefore not generalizing well in other datasets (poor out of sample performance).

5.1 Buckets - JOSÉ

Several types of explanatory variables are considered in the model, after the filters in the previous section, ranging from binary variables (6), categorical (22) and continuous (28). For continuous variables, we decided to group the possible values they can take into intervals (buckets) and then represented as several binary variables (one for each bucket, plus another for missing values)

Although some variables are labelled as continuous and theoretically could have an infinite amount of numbers, in our sample dataset they take only few values. For example, *CCJNumberSatisfied* (Number of County Court judgements against the primary borrower that had been solved at the origination time of the loan), could take any positive integer, however in our dataset it only takes on four values (0, 1, 3 or missing).

The fact that many of the variables used had some amount of missing values (no more than 40% of observations though), also contributed to the decision of bucketing the variables, since no consensus on how to replace certain variables' missing values was reached, due to the potential biases that would be imposed in the data. With this approach, they simply become a new category, an extra bucket.

Still, for variables like *CurrentInterestRate* and *LoanAge*, where there are many different values and few observations have missing values (which could be replaced by the mean for instance), simply normalizing them would be preferred, however we were still having difficulties getting the model to learn (loss decrease with each subsequent epoch). Simple normalization would be preferred because the biggest advantage of

using a neural network as classifier, is the fact that it can consider highly nonlinear relationships and patterns when estimating the probabilities of each state. Relationships that in some way are distorted and simplified when bucketing the variables due to the loss of information of each individual observation. As described in, George Cybenko (1989) and Hornik et al. (1989), neural networks can, with enough hidden layers and nodes, describe/ approximate any function to the desired level of accuracy, a limitation to other classifiers such as logistic regression.

The buckets are determined in the following way:

If the number of unique values in a particular variable is larger than 25, the max and minimum values for that dataset are calculated and 25 equally spaced intervals are created (division done only as a starting experiment, more and smaller intervals would mitigate the information loss effect described above), and each observation assigned to the matching interval. If less than 25, the different values the variable can take are identified and fewer buckets created, one potential change for these variables in future iterations would be to simply see each value as a category and encode them as if they were categorical variables.

After this process the input variables' dimension increases from 56 (the number of variables) to 659. Not all buckets created contain observations though (due to their equal length).

5.2 Network Mechanics – ACHILLE, JOSÉ

Our model's purpose is to predict the probabilities of a certain loan at a certain point in time, transitioning to different states, during the following 12 months. It takes as inputs several characteristics of the loan and borrower, and as output the probability of the loan status, 12 months into the future, being in one of four categories (Y), prepaid, performing, arrears or default ($Y = 0, 1, 2, 3$ respectively).

In Sirignano, et al. (2015) each loan could transition through several states multiple times during its lifetime. We took a simpler approach where different observations related to one loan are considered independent. We therefore deal with the dataset as if it were cross-sectional, and the problem becomes a basic classification one, without the need for a recurrent neural network.

A time variable and a 1-year lag of the loan status are included, so that potential time trends can be gauged, and the probabilities interpreted as transition probabilities.

Since we're interested in a 12-month prediction, all the information included in the regressors is lagged 12 months compared to the response variable. We are estimating $P(Y_{i,t} = y|X_{i,(t-12)})$, for simplicity, from now on, time subscripts won't be used.

With:

$I = number\ of\ observations$

$Q = number\ of\ variables$

$N = input\ dimension$

$L = number\ of\ layers$

$K_l = number\ of\ nodes\ in\ layer\ l$

$a^l_{k_l} = activation\ of\ node\ k_l\ in\ layer\ l$

$b^l_{k_l} = bias\ of\ node\ k_l\ in\ layer\ l$

$z^l_{k_l} = intermediate\ calculation\ for\ node\ k_l\ in\ layer\ l$

$w^l_{(k_l,k_{(l-1)})}$
$= weight\ for\ activation\ a^l_{k_l}\ applied\ to\ previous\ layer\ activation\ a^{l-1}_{k_{(l-1)}}$

$\sigma(a) = \max(\,0\,, a\,) = rectified\ linear\ unit\ activation$ (ReLU)

$s(z_1, \dots, z_{K_L}) = \left(\frac{e^{z_1}}{\sum_{k=1}^{K_L} e^{z_k}}, \dots, \frac{e^{z_{K_L}}}{\sum_{k=1}^{K_L} e^{z_k}}\right)$ = softmax activation function

The network works in the following manner:

An initial layer comprised by the input values is fed into the next layer (the first hidden layer) as follows:

A linear function is applied

$$\boldsymbol{z}^1 = \boldsymbol{W}^1 x + \boldsymbol{b}^1$$

$$\begin{bmatrix} z_0^1 \\ \vdots \\ z_{K_1}^1 \end{bmatrix} = \begin{bmatrix} w_{(0,0)}^1 & \cdots & w_{(0,N)}^1 \\ \vdots & \ddots & \vdots \\ w_{(K_1,0)}^1 & \cdots & w_{(K_1,N)}^1 \end{bmatrix} . \begin{bmatrix} x_0 \\ \vdots \\ x_N \end{bmatrix} + \begin{bmatrix} b_0^1 \\ \vdots \\ b_{K_1}^1 \end{bmatrix}$$

That is, each input is multiplied by a weight, and to the total sum a bias is added, this is done for every node in the first hidden layer (K_1 nodes in the first layer).

A non-linearity is applied

$$\boldsymbol{a}^1 = \sigma(\boldsymbol{z}^1)$$

$$\begin{bmatrix} a_0^1 \\ \vdots \\ a_{K_1}^1 \end{bmatrix} = \begin{bmatrix} \sigma(z_0^1) \\ \vdots \\ \sigma(z_{K_1}^1) \end{bmatrix}$$

An activation function is applied to each operation to incorporate non-linearities making the network able to represent highly nonlinear functions of its inputs.

This process repeats itself with every layer transition, with the previous layer's activations becoming inputs for the next layer's activations, in general:

$$\boldsymbol{a}^l = \sigma(\boldsymbol{W}^l \boldsymbol{a}^{l-1} + \boldsymbol{b}^l)$$

The function used to introduce nonlinearities across the network was ReLU. Today it is generally accepted to perform very well across different cases (a conclusion also arrived at in Sirignano, et al. (2015)) due to training performance (faster and easier conversion, loss decreases quicker), no other function was used due to time and computational constraints, though finding the right activation function for the problem is an iterative process and different combinations should be attempted, other common choices include tanh, sigmoid functions and exponential linear unit (ELU).

Regarding the last layer, because we wish to obtain probabilities for each state, the activation used was the softmax, which turns its inputs into probabilities, making sure that all the activations in the last layer are positive and sum up to 1. Guaranteeing that they are positive first by calculation the exponential of each of the intermediate calculations for the last layer (and making all sum to one by dividing each by the sum of all exponentials).

Intermediate calculation of last layer:

$$\boldsymbol{z}^L = \boldsymbol{W}^L \boldsymbol{a}^{L-1} + \boldsymbol{b}^L$$

Use the vector $\boldsymbol{z}^L$ as input to the softmax functions and obtain the probabilities for each of the possible states.

In our case, as in Sirignano, et al. (2015), the probability of default, for example, for an observation would be given by (Probability under a particular model's architecture and related set of parameters - M):

$$P_M(Y_i = 3|X_i) = \frac{e^{z_3^{(L-1)}}}{\sum_{k_L=0}^{3} e^{z_{k_L}^{(L-1)}}}$$

So, the network works as a function $\mathbb{R}^N \rightarrow \mathbb{R}^{dY}$, dY = 4 (all possible loan states in one-year time).

Again, neural networks are flexible because, we can alter the weights and biases (the set of parameters M) to make complex transformations to the inputs, with the help of implemented nonlinearities, and more importantly these parameters won't be predetermined, as we would, for instance, square or take the log of a variable to transform it in a normal regression, but will be dictated by the data.

This choice of parameters will be done, in such a way that, the model's outputs match reality as closely as possible. A loss function, a function whose output translates the error between the model's predictions, and the actual occurrences will be chosen and minimized by changing these parameters.

Let p_i denote the true distribution of Y_i and p_i^M the fitted distribution by the network for Y_i under the set of parameters M.

Because we are estimating probabilities for the different outcomes the natural choices for loss functions (that can be derived through maximum likelihood ((Goodfellow, et al., 2015), (Sirignano, et al., 2015)) are the categorical cross entropy (H) loss and Kullback Leibler divergence (D_{KL}) The two are related as: $D_{KL}\,(p_i||p_i^M) = H(p_i, p_i^M) - H(p_i)$ and both measure in a sense the distance between the true and the predicted probability distributions for the loan state for each particular observation. But

because we're dealing with a supervised learning model, that is we know with certainty the true state of the sample loan in 12 months, the entropy of the actual distribution is 0 ($H(p_i) = -\sum_y p_{i,y}\log(p_{i,y}) = 0$), making the two measures equivalent.

We then want to minimize H with respect to M, which is defined as:

$$H(p_i, p_i^M) = -\frac{1}{I}\sum_{i=1}^{I}\sum_y p_{i,y} log(p_{i,y}^M)$$

(For the entire dataset)

For example, an observation showing a loan in arrears in 1 year has the following predicted distribution:

	Prepaid	Performing	Arrears	Default
Actual	0	0	1	0
Predicted	0.02	0.5	0.3	0.18

In this example observation, we would compute the cross entropy as:

$$-(0 * \log(0.02) + 0 * \log(0.5) + 1 * \log(0.3) + 0 * \log(0.18))$$

As we can see the only contribution for the loss comes from the negative log of the predicted probability for the correct state, in a sense 'how far' the probability of being in arrears is from 1: $-\log(0.3) \cong 1.2$, it is easy to see that as $p_{i,2}^M \to 1$ the contribution to the loss will decrease and in the limit be zero.

One other loss function was experimented with, Categorical Hinge. With Y_i^M representing the class with highest score, the function is defined as: $L_i = -\frac{1}{I}\sum_{i=1}^{I}\sum_{y \neq Y_i^M} \max(0, p_{i,y}^M - p_{i,Y_i^M}^M + 1)$, in the example above the loss for the individual observation would be: $\max(0\ , 0.02 - 0.3 + 1) + \max(0\ , 0.5 - 0.3 + 1) + \max(0\ , 0.18 - 0.3 + 1) = 2.8$.

The results of training using this loss function were interesting and a similar decrease in loss to the model with categorical cross entropy was obtained. However, even though the final outputs of the model are positive, and sum to 1, they can no longer be interpreted as probabilities of each state. This loss function penalises more missing a prediction (having a lower score for the correct class compared to other classes), driving

the scores either very close to 0 or 1, there is however a method to obtain the desired probability distribution in practice (Platt, 1999) but not explored during the project.

5.3 Optimization – ACHILLE, JOSÉ

The loss function could theoretically be minimized by finding, analytically, the combination of weights and biases that would yield the smallest possible value of the function, however these problems generally have many parameters, making this process impossible to follow in a practical, real world situation.

An alternative, iterative process, Gradient Descent (GD), is used instead in these minimization problems. In this process we calculate the loss function's average gradient vector $\nabla H(p_i, p_i^M)$ over all the dataset and measure the impact of every parameter.

This gradient is obtained using backpropagation, because neural networks work as a composition of functions (as many as the number of layers and nodes), to get the impact on the loss function due a weight in layer l, $\frac{\partial H}{\partial w^l_{k_l,k_{l-1}}}$, we would need to take into account all the layer outputs (activations), this weight will affect later in the network. This derivative would be equal to the sum of several chain derivatives representing the paths that were affected by this weight.

We then want to 'move' from the current point M in a way that is the most efficient, that is in the direction where the new combination of parameters M' we suspect will have the lowest loss, given a certain step size (γ), we achieve this by subtracting the gradient vector to the current point, so $M' = M - \gamma \frac{1}{I} \sum_{i=1}^{I} \nabla H(p_i, p_i^M)$. The balancing of the step size or learning rate (lr) is important because we are compromising efficient training time by choosing a very small lr and the risk of not converging if we pick too big of a lr, with the possibility of observing an increase in loss in the new point M'. Even if a quick conversion to a minimum is achieved, we're not sure if we are at the true or satisfactory minimum of the function, therefore the random initialization (random assignment of values to weights and biases when beginning training) can lead to different result in different training runs.

This approach, although the most accurate, can be very inefficient for large datasets, instead, Stochastic Gradient Descent (SGD) can be used, where the gradient is calculated

and each step is decided by evaluating only one observation (in no particular order) at a time, or Mini-Batch SGD, where each step is decided based on an average gradient, but only for a small (random) sample of training data at a time. These 2 methods allow for several steps to be taken with every epoch (with epoch corresponding to one full passage over the data, that is all observations have been used to assess the gradient) compared to the single one that would be taken with GD. However because we're looking at individual subsets in SGD, the 'path' that will be constructed by the several gradients won't necessarily represent the shortest or most direct path to the minimum as would the gradient in the GD method, Mini-Batch SGD is preferred because this variance in the direction of each gradient compared to GD is somewhat mitigated when we take the average even of a small sample, so the variations won't be as pronounced, and faster convergence to a minimum is expected.

5.4 Hyperparameter Selection – JOSÉ

Although the minimization of the loss function will help us find the improvements on weights and biases, many other parameters (hyperparameters) of the network need to be specified manually, such as the number of hidden layers, nodes in each layer, activation functions, loss functions, learning rate, decay, number of epochs and optimizers (variants of the GD and SGD methods). The choice of the loss and activations functions was already discussed and can be narrowed down depending on the different kinds of problems being tackled, however for the remaining a grid search should be performed to assess the different combinations of hyperparameters.

The best performing model, **M1**, consist of a network, trained over 800 epochs with 10 hidden layers and 500 nodes per layer. Using regular SGD as optimizer and the Keras' default batch size (32 observations), a learning rate $lr = 10^{-5}$ and a decay (reduction of lr with every epoch) $d = \frac{lr}{2+\frac{Epochs}{800}}$ so by the last epoch the lr was one third of the original one.

For building and estimating the model the Keras package with Tensorflow as backend is used. Here we have several choices of hyperparameters that can be pieced together easily the model built in a very user-friendly by simply picking the desired hyperparameters from the available pre-sets.

The choice of optimizer was the trickiest one, since there are many optimizers available. The most common optimizers used to train neural networks are the traditional SGD Adam (P.Kingma & Ba, 2015) and Root Mean Square Propagation, RMSProp (Geoff Hinton in Lecture 6e of his Coursera Class), the last two combining traditional gradient descent with a momentum concept and other techniques. For example, dealing with decay without necessity for external input, to mitigate the variance/divergence from the correct path that can happen in SGD without loss of efficiency (still training in mini-batches and computing the same number of derivatives per step).

5.5 Imbalanced Classes - José

Due to the fact that the amount of loans in default (1.0266%), arrears (0.6982%) and in prepayment (1.2829%) state is much less than the amount performing (96.9922%), some measures need to be taken to guarantee that the model gives enough importance to these classes when training. Otherwise the impact on the loss function of missing these would be neglectable once the average loss of the dataset was calculated. Two approaches were tried:

Weight the loss function according to the class being tested. Penalising missing these underrepresented classes would counter balance their small natural impact on the loss. The weight was calculated as $j_y = \frac{1}{frequency\ of\ class\ y}$ so we want the inverse proportion, all classes end up with the same weight. To take a more conservative approach since the minority classes are the ones of interest and may dictate performance of a pool of loans, an even higher weight can be considered fort the minority classes, in our case, in model **M1**, for all but the performing class, the weight was multiplied by 2 (although results for no weights – **M1_0** and no doubling of the minority classes' weights - **M1_1** are discussed to justify this step), turning $j_y = \frac{2}{frequency\ of\ class\ y}$, for $y \neq 1$ (performing state, the dominant class). The loss function becomes:

$$H(p_i, p_i^M) = -\frac{1}{I}\sum_{i=1}^{I}\sum_{y} \boldsymbol{j}_y p_{i,y} log(p_{i,y}^M)$$

One second approach, commonly used in problems such as disease diagnosis (Mazurowski, et al., 2008), where the rare events are of great importance, would be to

artificially increase the minority classes (over-sampling) or decrease the majority class (under-sampling) or a combination of both. Models were estimated using the methods in the imbalanced-learn package (Lemaitre, et al., 2018). One of the methods analysed was the SMOTENC (Synthetic Minority Oversampling Technique extended to categorical variables (Chawla, et al.; 2002), the results however, as in other attempts (of similar methods) were not satisfactory (no training loss decrease and large swings in validation loss, both sharp increases and decreases). In any case, further investigation into these methods is required because the model specification used when testing these was not the current one (**M1**), in particular the usage of RMSprop optimizer and much less nodes per layer (never above 200), so further investigation into these processes is needed.

6. Results – ACHILLE, JOSÉ

Model performance with every epoch:

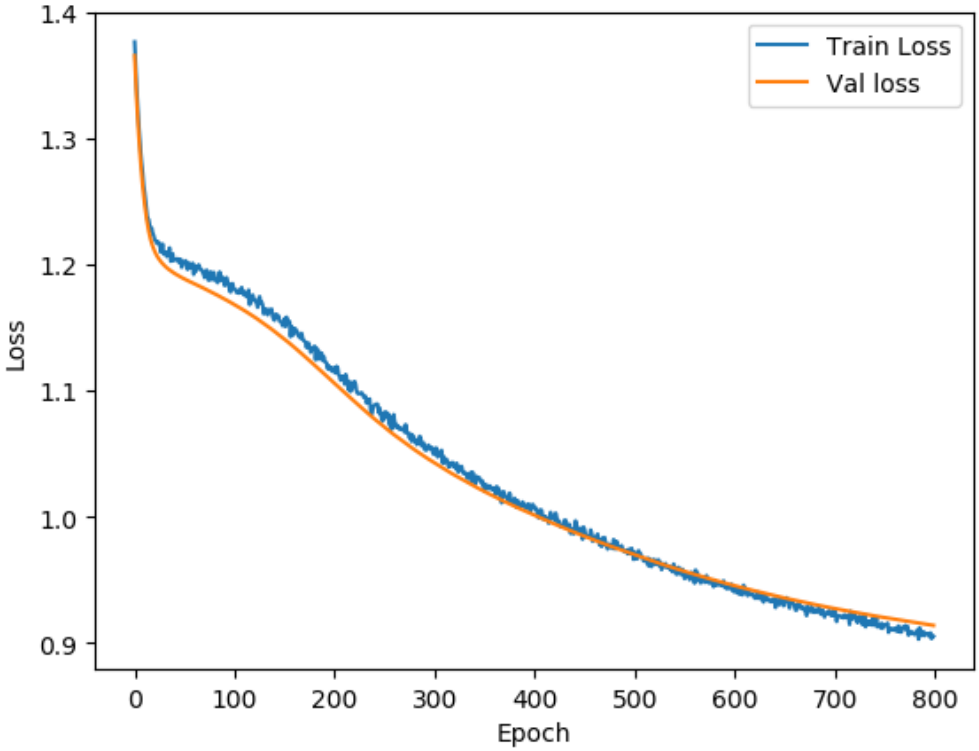

Figure 17: Training and Validation loss evolution with every epoch for M1

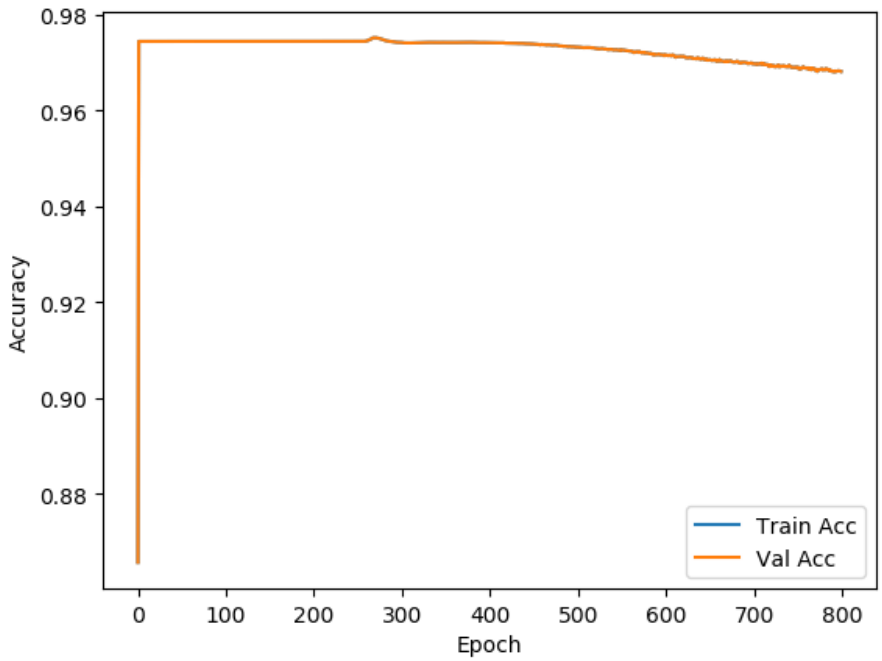

Figure 18: Accuracy evolution with every epoch for M1

Looking at figure 17, the classifier's performance during training (that is if and how the loss is decreasing with every epoch) can be observed, in both the training and validation set. The decreasing loss signalling that predictions are closer to the true distribution of probabilities for each state. After a certain point the loss becomes concave, signalling a small learning rate, perhaps needing adjustment for these epochs. There is also absence of overfitting since the validation loss closely follows the trend of the training loss. Both still decrease by the later epochs, so, it is likely that more iterations would improve the model's predictions.

Turning to accuracy (Figure 18), that is the fraction of correct predictions out of all predictions, the results are somewhat puzzling. We verify a sharp increase in accuracy in the beginning (matching the sharp decrease in loss) but after, at the later epochs, accuracy starts decreasing while loss remains decreasing. The fact that the loss function is being weighted according to the possible states (and more importance given to the minority classes), is likely making the optimizer tune the parameters not to miss those underrepresented classes, even if it means wrongfully predicting a distribution that understates the probability of the loan being performing. Which is the desired result. A conservative approach is preferred to be missing defaults or prepayments that can have dire consequences for example when analysing a pool of loans to create credit derivatives. It is also important to mention that accuracy is an incomplete measure due to the imbalance of classes in the dataset (high accuracy can mask the failure to correctly predict transitions to one of the rare states). The ROC curve analysis (precision and recall) will provide extra information on the model's observation level performance.

In appendix D it is available the training history for a network trained without weighting classes – **M1_0** (figure 45) and with simple proportion as weight, that is without doubling the minority class weights – **M1_1** (figure 46).

6.1 Variable Significance - JOSÉ

In order to measure the importance of each variable, a leave one out analysis was performed, that is the model was evaluated and the variable of interest (x_q) omitted. It would be, however, too costly to rerun the model every time we wished to test the significance of a given variable by leaving it out. We cannot, however, evaluate the model on a test set missing certain attributes, the input dimension needs to be the same

as the observations used for training. In alternative since there is a column for every possible value x_q can take, all these columns were set to zero for all observations of the test set (something that was never 'seen' by the model during training). An example for *LoanStatus* is given in appendix E.

After evaluating the model with all categories of x_q set to zero, the loss is calculated and compared to the original loss ('Difference' computes the difference between the loss with the variable left out and the original loss - 0.534165). Table 1 (appendix E) summarizes the results for the top variables.

LoanStatus is, without surprise, the most significant variable. Partially because some of the observations that are currently in default (*LoanStatus* = 3) will, by design, be in default in one year, therefore increasing the importance of the current loan state (*LoanStatus*). Regardless, when considering only observations initially performing (*LoanStatus* = 1) it remains an important variable (table 2 in appendix E).

It is visible the impact of the agent's income, with the variables *ArrearsEndBalance, ArrearsIncome* and *Installmentpropincome* (defined in section 3.2) showing relative high importance, giving most importance to borrower's risk factors. Same holds true when looking only at initially performing loans (table 3). It is therefore in agreement with Vandell, K. (1978) and Webb, Bruce G. (1982) and Von Furstenberg (1969) and Gau (1978).

And in general, we can see that borrower related variables dominate the top of the table with *YrM* being the first variable to deviate. *YrM* in this case translating the macroeconomic reality the particular period of the observation and potential trends.

In table 2(appendix E) we can see the variables that, when omitted improved the loss in the test set. This should not happen. If a variable is poor at predicting, the optimization process should set parameters in such a way that it won't affect the output estimation. Having said that, this method has its limitations and **M1**, is very much an imperfect model. Its specification may not be the most suited to the problem and currently may not be learning the correct and generalizable patterns in between the variables. One other issue that may be distorting the results is the weighting of the loss function, since this evaluation (both with the original and the altered attributes) does

not take into consideration the weight of the different classes given when training the model.

Some results are expected, variables capturing Loan-To-Value (*CurrentLTV and Incentivesell)*, consider the most important in Von Furstenberg (1969) and Gau (1978) for instance, may display some collinearity and therefore loss is not affected when one is removed, this may happen between other variables where the effect of some are mirrored by other or a combination of others. Variables which were initially thought to have great explanatory power (*LoanAge, DistanceFromValuation for example).*

6.2 Variable Impact - JOSÉ

The variable impact is measured by analysing the magnitude of the change in each of the probabilities in the distribution when the variable changes from one value (in our case, class, since even the continuous variables are separated in binary variables representing each bucket) to the other. It is considered not in a point, but across the dataset so an expected distribution is used (as simple average of the individual ones).

$$p^M = \frac{1}{I}\sum_{i=1}^{I} p_i^M$$

Here the ideal scenario would be to evaluate the model on a randomly generated dataset that would have every other possible combination of attributes. The test set is, however, considered to be representative of the full dataset and distribution of attributes, therefore the analysis is done over the test set and not over a randomly generated sample.

Because we want just one value associated with each variable x_n representing impact on each of the probabilities, the expected impact magnitude ($E\left(\left|\frac{\partial p^M}{\partial x_n}\right|\right)$) is calculated as the simple arithmetic average. An example for *CurrentInterestRate* is given in appendix F. In the same appendix the full list of variables and their impact is reported as well as the top variables affecting each category's probability.

In appendix F (tables 16 - 19) are presented the most impactful variables for each of the different state's probabilities. Again, the values presented are the average size of the changes in the probabilities due to changes in each of the variables.

We can see that the current state of the loan is decisive in all future states' probabilities. Looking then to originally performing loans (tables 20 - 23).

This analysis is not without is drawbacks though, as, categorical and binary variables, will have less possible states than continuous ones. Possible improvements are referred in section 8).

6.3 ROC curve - ACHILLE

Looking at the individual performance of the model (loan level), several measures can be considered when measuring the performance of a classification model. Very common ones are Accuracy, Sensitivity and Specificity (Appendix G). Accuracy can be very useful as it provides a global overview of how well the model did perform by taking its correct classifications in proportion to the total number of observations. However, it tends to always be high when classifications' cut-offs are high, or generally when the data set is significantly imbalanced with one category largely dominating the others (Notesbyanerd; 2014)(Ritchie NG; 2018). It is a significant issue for a model like ours, trained on a data set where the vast majority of the mortgage loans observed are classified performing twelve months forwards. As a result, Accuracy should not be the preferred measure of this multi-class credit model. Instead, the use of Receiver Operating Characteristic (ROC)[8] curves, assessing a model's sensitivity and specificity is a better option. A ROC curve illustrates the ability of a binary classifier to correctly categorize its observations by showing how its false positive classifications increase (1-specificity) in relation to improvements in its true positive classifications (sensitivity). As a result, the problem encountered with very high accuracy measures for imbalanced data set is neutralized since the effect of changing correct and incorrect classifications' rates is visible. Since ROC curves assess binary categorizers, a multi class model like ours needs to have a ROC curve plotted for each category. When analyzing ROC curves, the researcher must know the closer the curve is to the top left corner, the better. Such curve would suggest that the assessed model is able to improve its correct positive classifications without increasing its proportion of incorrect positive classifications.

[8] See: https://www.youtube.com/watch?v=MUCo7NvB9SI for a quick explanation on ROC curves for credit risks models.

Inversely, should the ROC curve be close to a 45º line crossing the graphs origin, it would mean that the model's results are effectively random, as increasing the proportion of correct positive classifications could not be done without equally increasing the proportion of incorrect positive classifications. The Area Under the Curve (AUC) for each category has also been calculated to quantify this analysis. The bigger the area, the closer the curve is to the top left corner, and the better are the results. Below, are plotted the ROC curves[9] corresponding to model predictions over the test set (not used for training) of our model's predictions for out of sample.

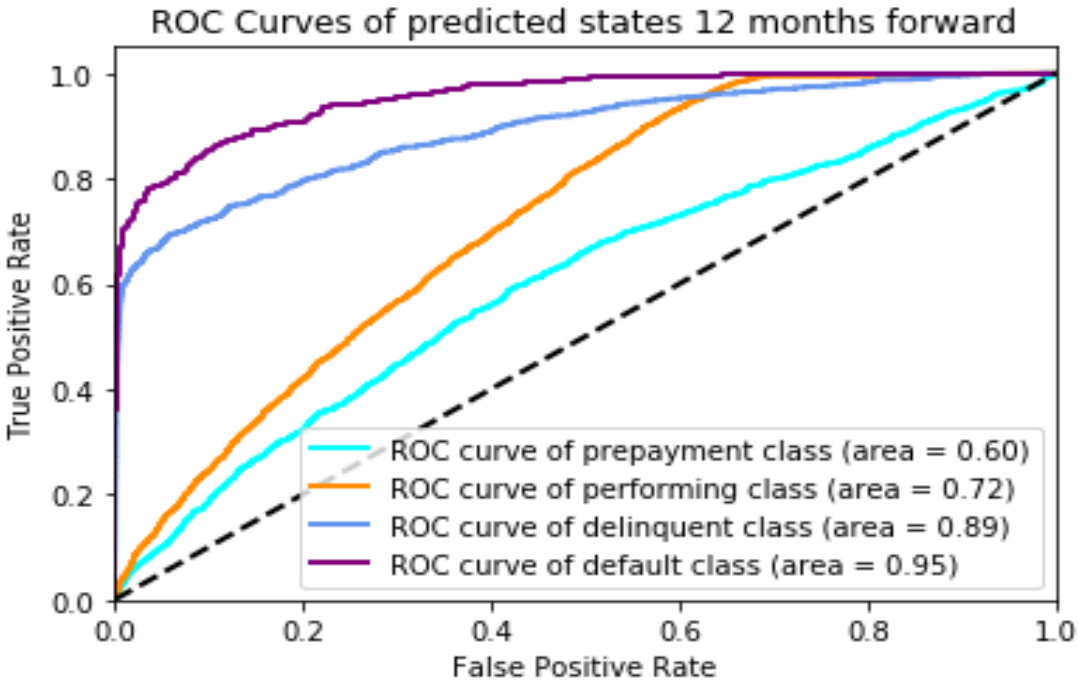

Figure 19

- **Time horizon of predictions**: 12 months.
- **AUC strength index**[1]: 0.90 - 1 = excellent / 0.80 - 0.90 = good / 0.70 - 0.80 = fair / 0.60 - 0.70 = poor / 0.50 - 0 .60 = fail

Looking at this graph[10], we can see that the model's potential to predict default at a time horizon of twelve months is excellent. The model can highly increase its

[9] See: https://www.dlology.com/blog/simple-guide-on-how-to-generate-roc-plot-for-keras-classifier/ for a guide on codding ROC curves for Keras classifiers.

[10] AUC strength index from: Thomas G. Tape; "Interpreting Diagnostic Tests"; *University of Nebraska*; available at: http://gim.unmc.edu/dxtests/roc3.htm (last assessed in December 2018).

proportion of correct classification of this state (true positive) without increasing too much its proportion of incorrect classifications (false positive). As a result, feeding the neural network with more data, easily doable with computers more powerful than our personal laptops, should lead to excellent default predictions. The same can be said for the delinquent state. However, our model's potential to predict prepayment and performing 12 months forward is poor and fair respectively. Regarding the performing predictions, such result probably comes from the way we parameterized the loss function, correcting itself with more severity when missing alternative states rather than performing state (see 5.5 Imbalanced Classes) For prepayment however, these disappointing result probably comes from the fact we trained the model on a largely imbalanced sample with few prepaid transmitting less information on this particular state to help our model recognize it.

We evaluated our model's precision in predicting future states of initially performing mortgage. We did so to analyzes the model's performance on predicting the transition from performing to other states as this ability is the most sought after, because transitioning from a performing to default is synonymous of a cash flow deterioration which is the concerning effect we wish to try to predict. A transition from performing to performing is something we would be less interested in. The ROC curves for originally performing loans can be found below:

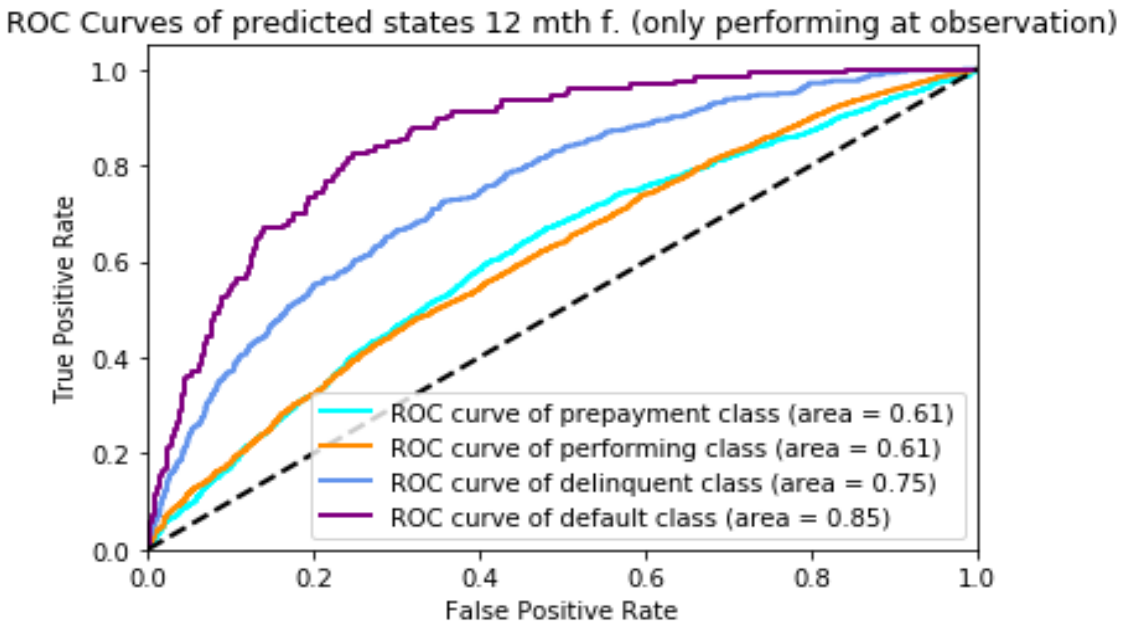

Figure 20

- **Time horizon of predictions**: 12 months.
- **AUC strength index**: 0.90 - 1 = excellent / 0.80 - 0.90 = good / 0.70 - 0.80 = fair / 0.60 - 0.70 = poor / 0.50 - 0 .60 = fail

The model decreases in precision overall when we consider only initially performing loans. This was somewhat expected as removing default and prepaid as a possibility for the loan status at the time of the observation removes certainty on future prepayment and future default states. In case of currently defaulted loans as these (by design), will remain in default in 12 months. In addition, delinquent loans are more likely to either default or go back in arrears twelve months forward than currently performing loans. Holistically, the results of this analysis are coherent with the findings of the ROC curves' analysis on the whole data sample, that is, high potential for predicting default and delinquent states, and difficulties in classifying correctly future performing and prepaid loans.

6.4 Graphical predicted default rate – JOSÉ

As referred in the previous section, the model is most proficient at analysing the transition from performing to default. Estimated average probabilities (for each class an average across all observations is taken) in relation to some variables are presented

in appendix H (calculated over the whole sample) and will be discussed in this section, there are however many more whose behaviour can be analysed, we, therefore focus first on some of the identified variables in table 23, that cause the most significant shifts (magnitude wise) of the predicted probability of default, CurrentLTV and Completion (which also reflects effects tied to *LoanAge*) will also be analysed as they were important variables in the logistic regression case. Actual values won't be of much significance, since there is a general overestimation of the probability of default by our model, likely due to the extra weight given to the minority classes during training, instead, changes with a variable's values/categories will be analysed.

Looking at *GeographicRegion* (figure 49), properties located in the regions 'YOHU' and 'SCTL' present the highest estimated default probabilities with the ones located in 'EAST' and 'SOEA' regions the lowest predicted rate, showing that house characteristics (location in particular), as well as possible economic inequalities between regions are a significant factor. It is observed through figure 53 that regulated loans (under the Consumer Credit Act in UK) have a higher estimated probability of default, a result somewhat surprising at first due to the fact that this legislation demands a certain degree of information to be collected before conceding credit, which should allow lenders to make better decisions over weather or not to concede credit to an individual, though further investigation into on the topic is needed. In the same figure we can see also the influence of the variable *BankruptcyOrIVAFlag,* which indicates if the borrower has been bankrupt in the past (or an Individual Voluntary Agreement or equivalent). An interesting result on the borrower's history, and logical, indicating that some of the reasons that led to the previous financial difficulties may still be present.

Looking then at *PaymentType*, agreed at the mortgage's origination, payment via increasing instalments shows to be more propitious to defaults, displaying the highest estimated probability, with borrowers likely to be able to meet the initial amount of payments but later incurring in more difficulties. It also may leave borrowers more exposed to variations in the property's value (further investigation also required on how the variable would to *CurrentLTV* for instance) since a decrease in property value, for instance, would mean a sharper decrease in the agent's home equity compared, for example to an annuity type of payment, where larger principal payments are made at

the beginning. And in fact, annuity type payments are the ones that lead to a lower estimated probability of default. There is, however, contrasting evidence. Bullet loans, where the full principal payment would be made at the end of the loan term, display a smaller probability of defaulting (though still higher than annuity type payments).

Looking now at *Completion* (figure 51), we can see that as the share of time passed since origination increases, so does the probability of default, we can also see that, only passed the contractual time frame for the loan (100%), does the estimated probability increase significantly when compared to early values a result somewhat expected, since the farther away from the initial contract the more variables can change compared to the initial reality, weather it is related to the borrower (changes in income, type of employment among others), property (value) and financing (more adverse rates compared to origination). Specially if the reason why the loan has lasted more than the originally agreed maturity is tied delinquency, the capacity of the borrower to meet the payments is questionable and the default rate should reflect it. The effect is however not monotonic, with the model predicting loans around 210 – 218% *Completion* to be less likely to default when compared to loans with *Completion* around 150%,

Finally Looking at Current loan-to-value (both *CurrentLTV* and *Incentivesell* variables), we observe an increase in the estimated default rate, until about a LTV of 100%, after which higher LTVs will correspond to increasingly lower estimated probabilities (with more variance in the case of *CurrentLTV*) hitting a minimum around 150% and sharply increasing afterwards, to highest estimated probability.

6.5 Predicted default rate – HENRIQUE, JOSÉ

Plotted in figure 3 are the expected default rates taken as the simple average of probabilities of default for every observation in the test set in the different periods.

Matching these with the observed default rate (in the sample), there is a clear overestimation of the probability of default. Again, the extra weight in the loss function to the minority classes (to which default belongs) may be making the model overestimate these probabilities. When looking at the predictions, without it, they are much closer, albeit still not matching, displaying opposite shifts.

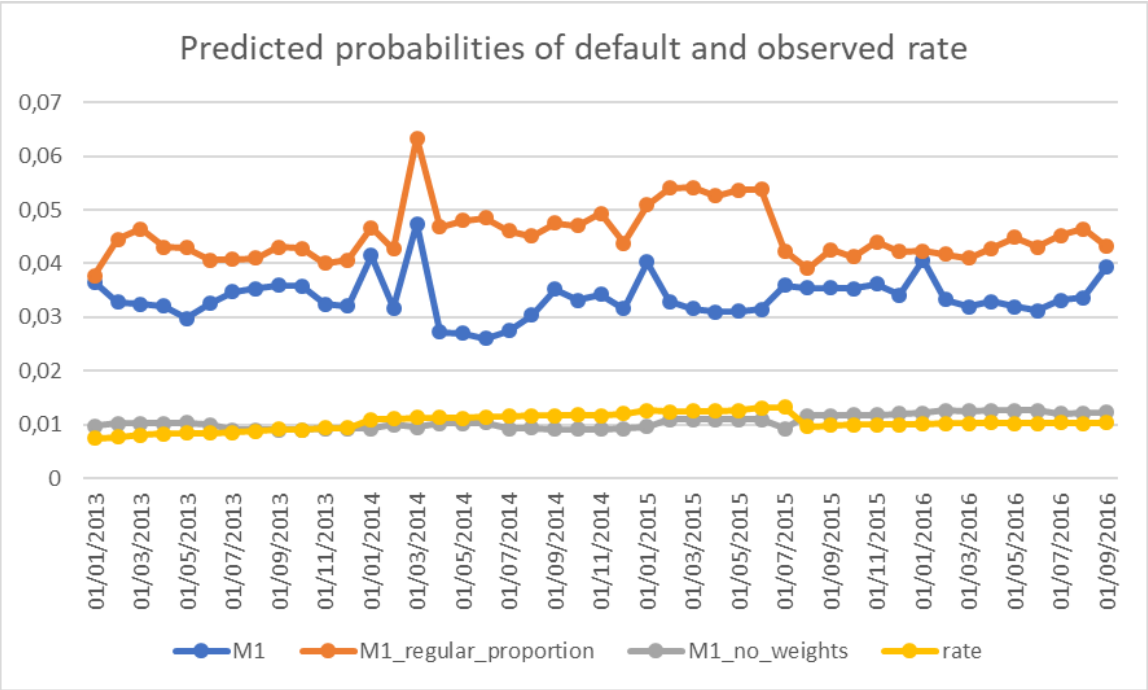

Figure 49: Default predictions and observed rate. M1_no_weights refer to a model equal to M1 but trained without weighting the loss function and M1_regular_proportion refers to a model equal to M1 but trained using simply the inverse proportion

7. Conclusion – ACHILLE, ROBERTA

The main task asked by Moody's Analytics was to assess the relevance of using deep learning technology to evaluate credit risk of mortgage loans, based on Sirignano, Sadhwani and Giesecke paper (2018). Analyzing data from UK mortgage owners, we built a multilayer perceptron of 10 hidden layers with 500 activations each using Keras layers, trained on 800 EPOCHS, using SGD optimization. It predicts mortgage loan behavior on a twelve months horizon. Due to computing power limitations, we were only able to perform this task on a very small sample (20 thousand loans) of the full data set. We managed to achieve high accuracy with this network although masking the imbalance existence in classes in the dataset. The ROC curve results suggest our model is very good at distinguishing future default and delinquent loans and assessing the transition from originally performing loans to both classes. Despite struggling with future prepayment classifications, the encouraging results we achieved with defaults and delinquency with all the limitation and constraints of our model, strongly lead us to believe the use of neural network should be investigated further and is likely to outperform traditional methods based on linear analysis. Regarding variables' impact, we found a strong dependence on borrower characteristics, in particular income related risk factors and past history of the borrower. Looking at the predictions of our model from the full dataset, we can also verify then estimation of nonlinearities (Current LTV and Completion mentioned in section 6.5).

8. Next Steps – HENRIQUE, JOSÉ, ROBERTA

The sample was obtained, by extracting observations based on randomly selected loan keys (a unique identification for every loan), however, afterwards individual observations are considered as different loans. We assume independence between them however this is a very weak assumption, and something could be addressed with a different method for sampling. The alternative, and more conventional way, would be to select random observations at each point in time, so instead of a unique key, select a unique combination of that key and a date (for example out of all observation in March 2015, take 10 thousand at random). This is not without its computational hurdles for us, since we're also obtaining our response variable from the data (besides picking a random key in a particular point in time, we would need to

guarantee it had an observation 12 months after, March 2016, and extract the *LoanStatus* at that time). A computational simple approach could be taken, but for a sample size similar to what we ended up (525 thousand observations) it would take a long time to extract and for time reason, as well as doubt about the final type of model (some panel data alternatives were analyzed) no further research was done on other techniques to do it.

We would also use standard normalization when preparing variables for the model. Buckets provide an excellent solution to dealing with missing values (simply adding a new category/bucket), however when using it we forfeit some of NN strengths, that is the capacity of the network to mirror the most complex relationships. With buckets these are captured but at a much shallower degree, since we lose some information, when limiting the possible values/categories that a variable can take to 25. Therefore, in next iteration, replacing the missing values for continuous variables would be done with either with the mode or the mean, or even the lowest possible value, depending on the scenario, and implications it may imply to the data, and a binary variable created to identify if the value of the variable in that observation was originally missing. This was the original approach (and the one used in Sirignano et al. (2018)), and the results were not satisfactory (no loss decrease), however almost all other network specifications, in particular, the weighting of the loss functions and the size of the network (number of nodes and layers) were not the same as M1, with much less nodes, and no weighting to the loss, which as we can see if not given, results using buckets and the same architecture as M1 gives unsatisfactory results (Figure 47).

Research on more and different methodology to assess the significance of variables would also be beneficial, since it is too costly to simply retrain models without them. And the approaches taken in section 6.1 and 6.2, although sound, are still imperfect because we're not fully omitting the variable and retraining the model (6.1) nor are we seeing the different magnitudes for different values of the continuous variables, obtaining only an average. The graphical analysis used in section 6.4 for a few variables being the most detailed view to see the patterns estimated by the model.

There are also possible deviations from our framework, namely the use of decision trees, which, more clearly, highlights what variables the role of each variable,

and the use of a recurrent network which would work with panel data (tracking observations belonging to the same loan) and not a pooled cross-section. Appendix I touches on these 2 methods.

9. Limitations - ROBERTA

The first one regarding data inputs. The academic paper we inspired this project from, based its analysis on a much larger data set and considered a much longer time frame and a larger amount of loans. The provided dataset was also quite large and redundant however we could not take full advantage of it due to hardware restrictions (more specifically the RAM amount and the ability to use GPUs while training the models, which proved quite unstable with many crashes in Windows)

The fact that we took a small sample from the dataset, the variety of data also suffered, with observations ranging only from 2013 to 2017 (very little when compared to the much larger sample considered in Sirignano et al. (2018) which range from 1995 and 2014) when data on the Moody's set was available from 2008. Variety also suffered as our sample had a few variables that were constant or missing (therefore not adding any information). The variables, and respective classes are listed below:

BorrowerType – All are individuals;

ClassOfborrower - Prime borrowers;

CreditQuality: Pass type B;

IsUnderLitigation – Very few borrowers under litigation, none present in the testing set;
Lien – 1st Lien (Lender) – First to be paid when borrower default (seniority)

OccupancyType: Owner occupied

PastToCurrent – expected since Prior Balances are all 0.

PaymentFrequency – All monthly payments (1)

Our last limitation was our lack of proficiency in python. With none of the member of the group having programmed in it. Python has several very efficient libraries and we did not have the level of proficiency to fully leverage these, making some processes more time consuming then what they could have been otherwise, worsening

the time situation. things as none of us had any experience in any programming language.

Appendix

Appendix A – Structural Credit Risk Models

1. Moody's-KMV Portfolio Manager:

It is an improvement of Black-Scholes-Merton's model which estimates the Expected Default Frequency (EDF). This model comes up as an attempt to solve the outdated credit ratings problem. Indeed, the ratings provided by credit rating agencies, usually, are not updated with the regularity desired. The slow adjustment would make the ratings outdated and misestimate the risk. The KMV model, since it is using data from the stock market with more regularity, permits their rating to be adjusted faster and continuously, allowing the investor to have a perception of risk closer to the real one. Therefore, EDF frequently anticipates credit migrations, when comparing to the ratings. One problem of this model is the fact that it is only applicable to listed companies, in which the information is publicly available to the market.

2. Credit Metrics:

Credit Metrics is also an extension of Merton's model. It was proposed by JP Morgan and it is used to evaluate and manage the risk exposure of a portfolio composed by several loans. It connects the portfolio's value changes with the credit ratings migration (up or down), computing the portfolio's credit value at risk. Credit Metrics problems are related to portfolios with large number of obligors - in which is necessary to use a factor's model - and to the constant transition matrix, insensitive to business cycles.

3. Credit Portfolio View:

This structural model, commercialized by McKinsey, is an economic state dependent model, meaning that takes into consideration the economic state, when computing the default probabilities. In fact, these probabilities depend on the macroeconomic fluctuations, measured by indicators such as the GDP growth rate, interest rates, exchange rates, unemployment rate, etc. This procedure solves the previous issue of *Credit Metrics* static transition probabilities matrix; however, it faces calibration problems, because it requires a high number of in-sample defaults.

Appendix B –Analysis of the created variables

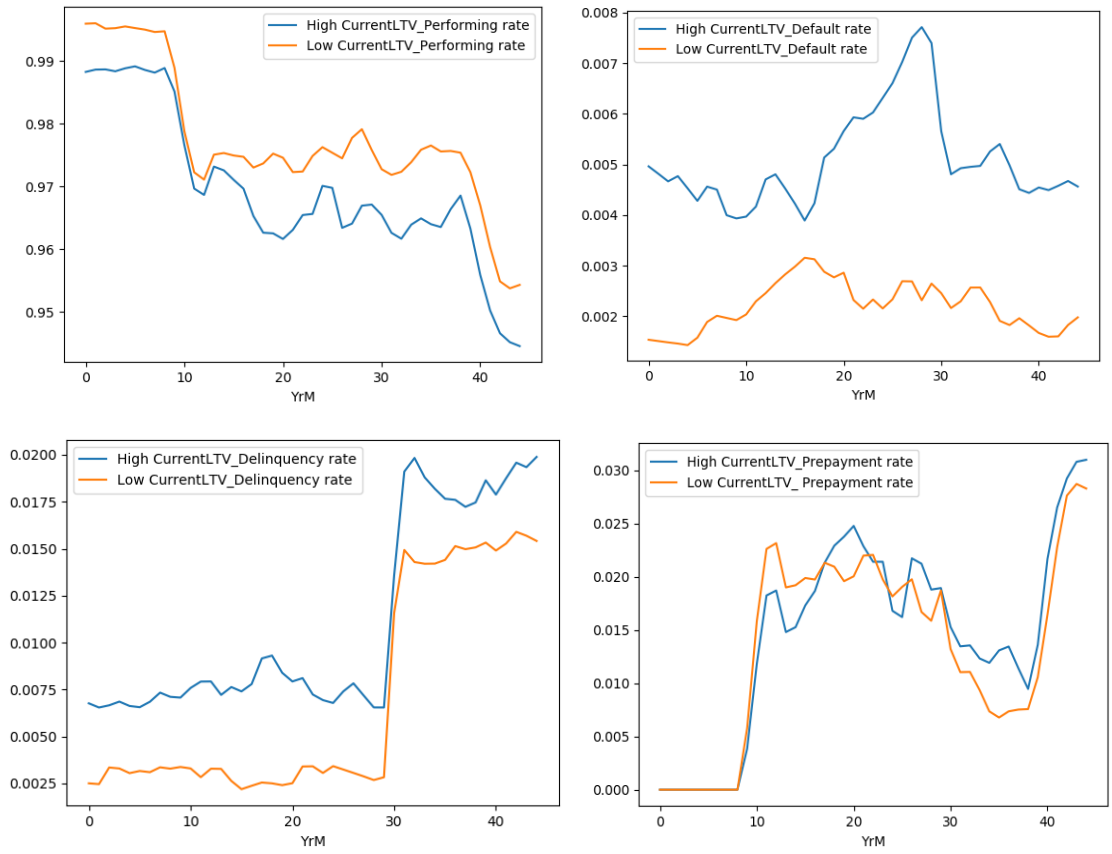

Figure 21 – Current Loan-to-Value for Sample 2

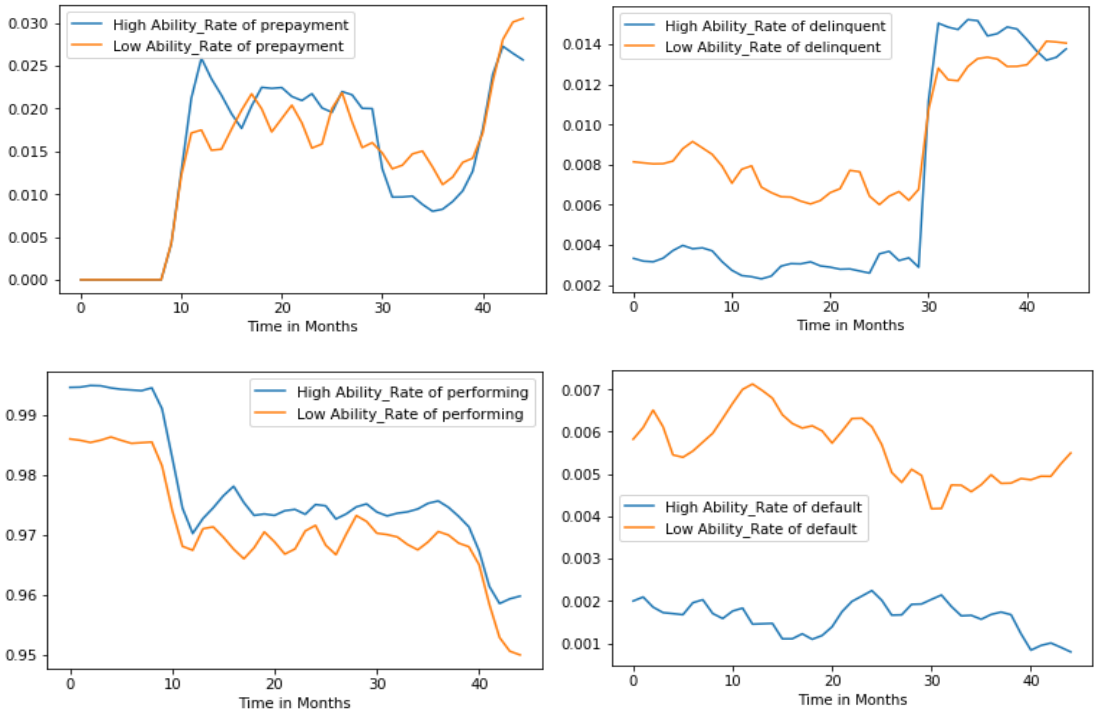

Figure 22 - Ability to cover the loan with property value – LTV with last official valuation for sample 1

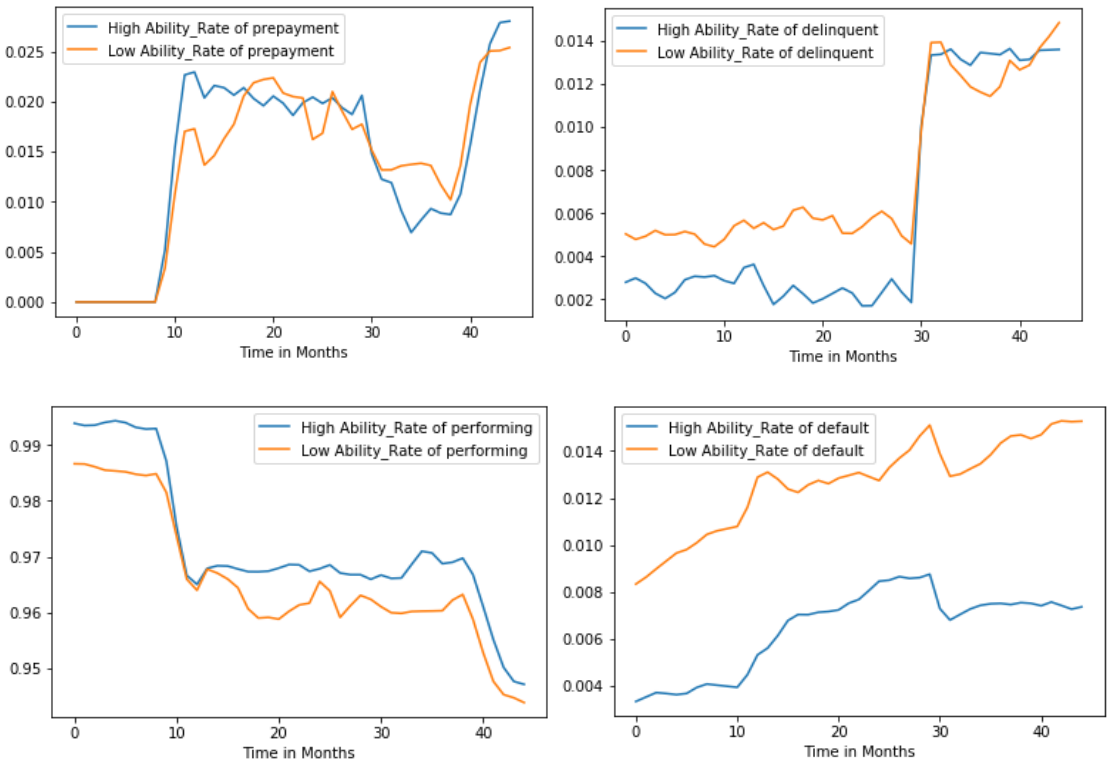

Figure 23 - Ability to cover the loan with property value – LTV with last official valuation for sample 2

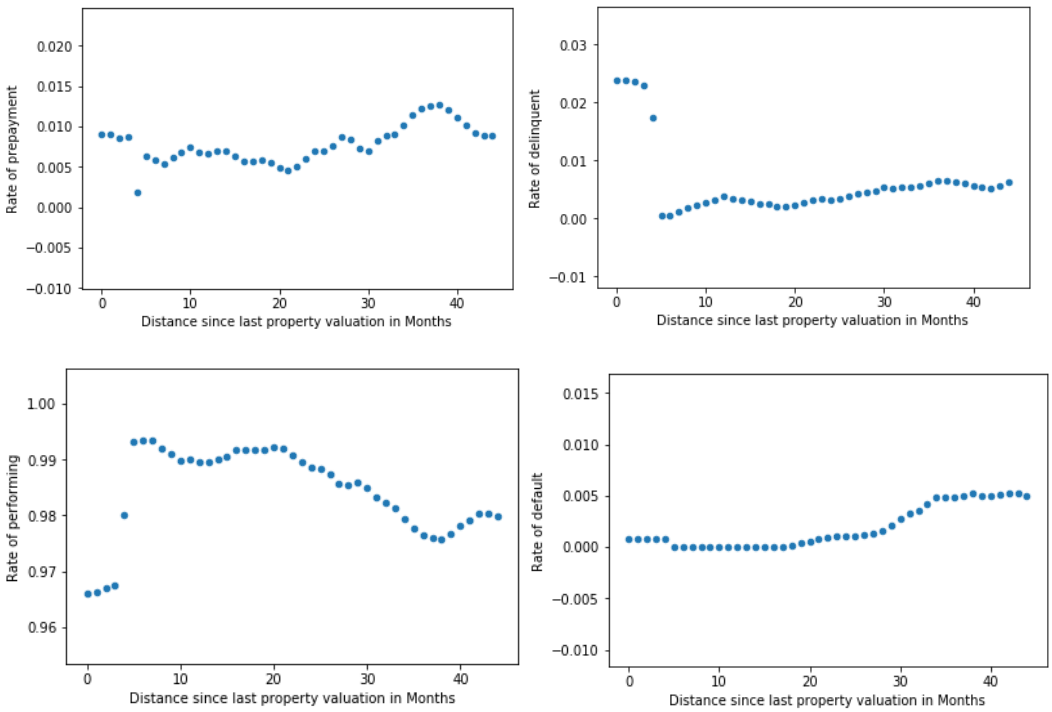

Figure 24 - Distances since last property valuation for sample 1

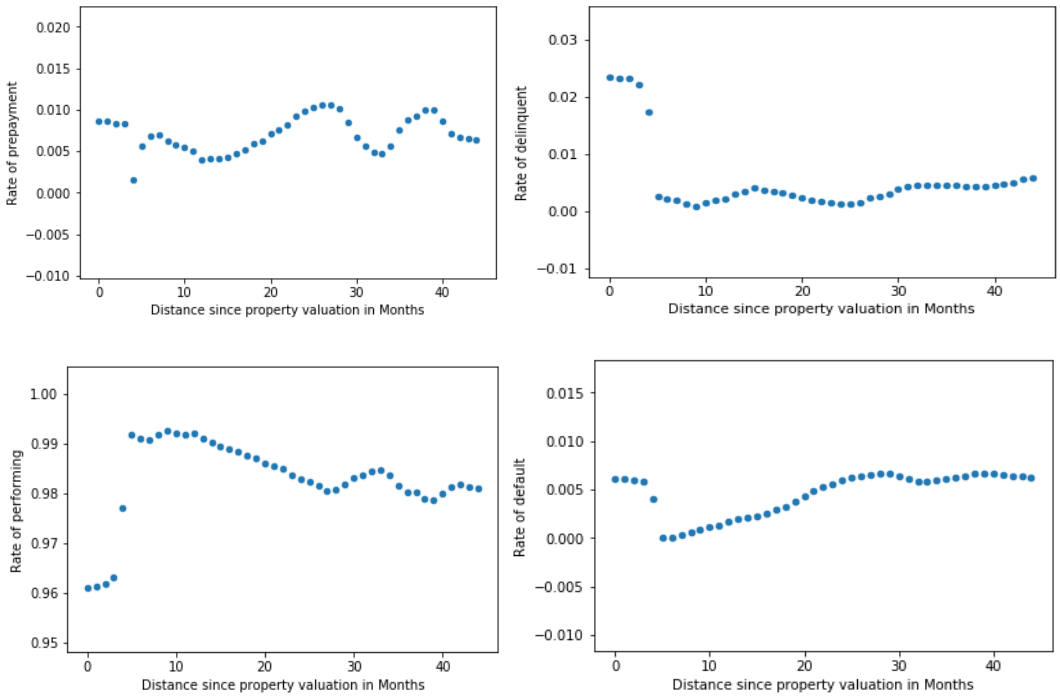

Figure 25 - Distances since last property valuation for sample 2

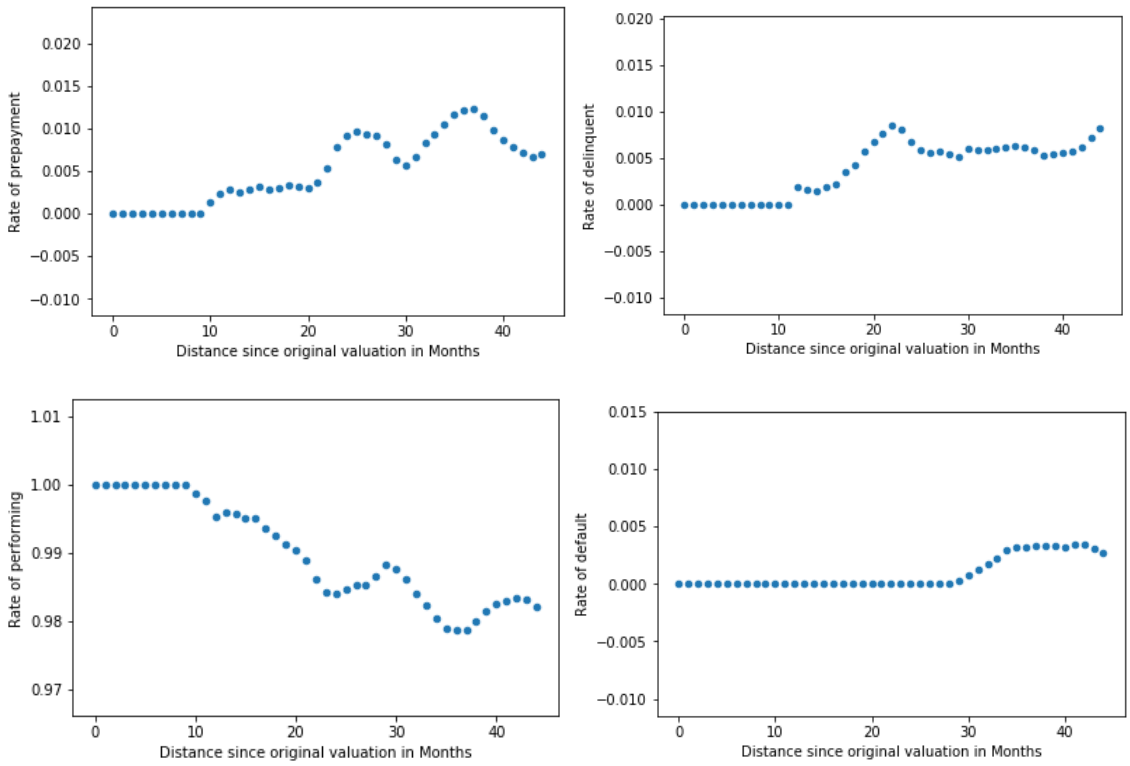

Figure 26 - Distances since original property valuation for sample 1

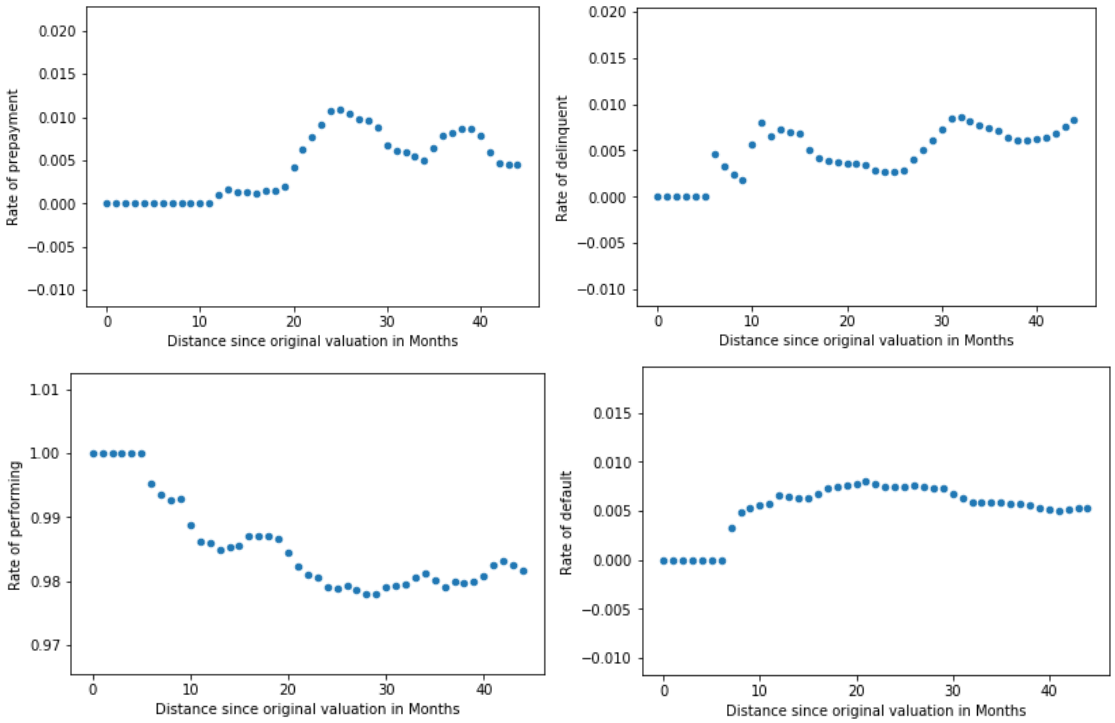

Figure 27 - Distances since original property valuation for sample 2

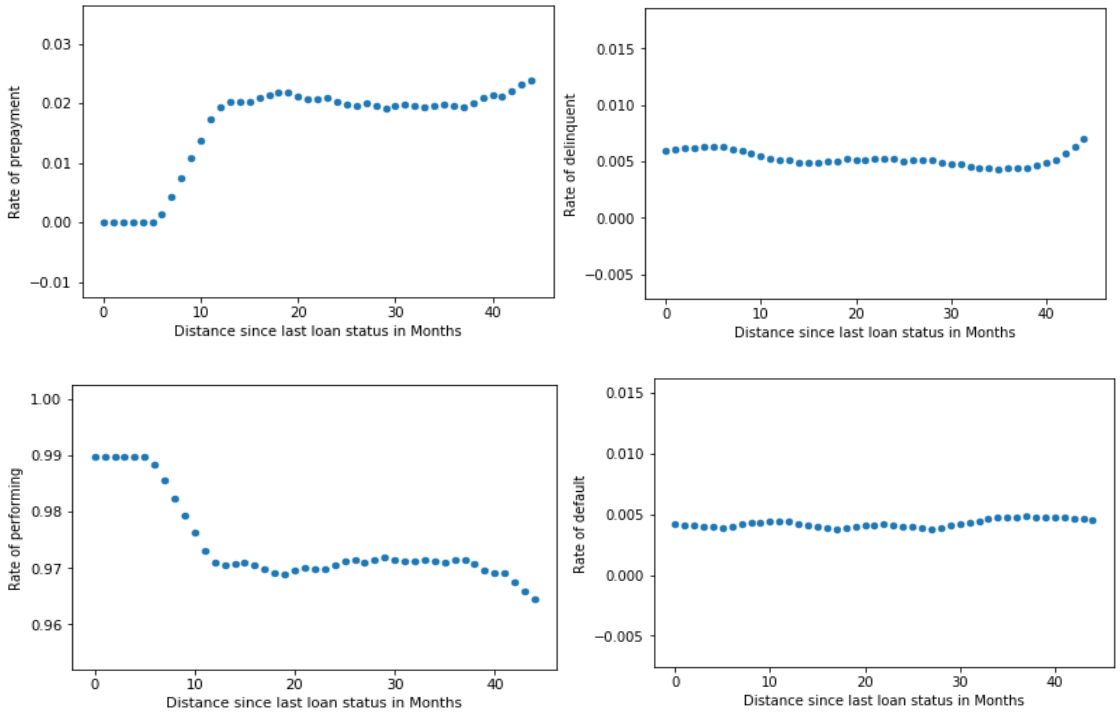

Figure 28 - Distances since last loan status for sample 1

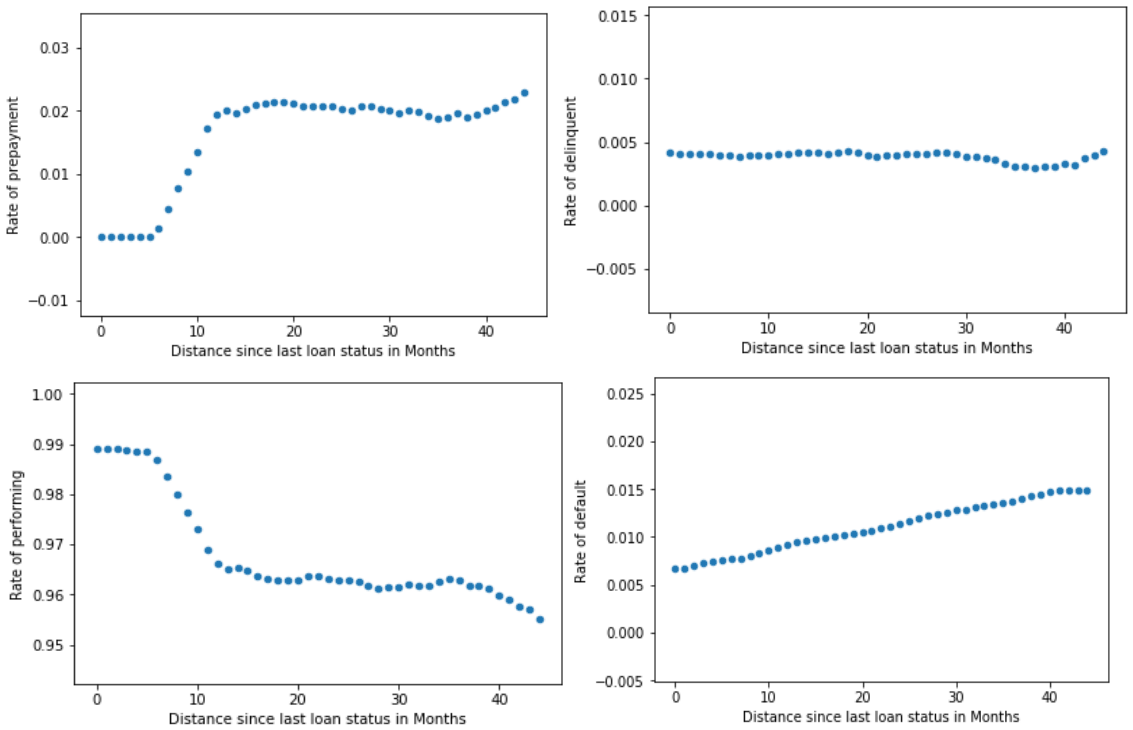

Figure 29 - Distances since last loan status for sample 2

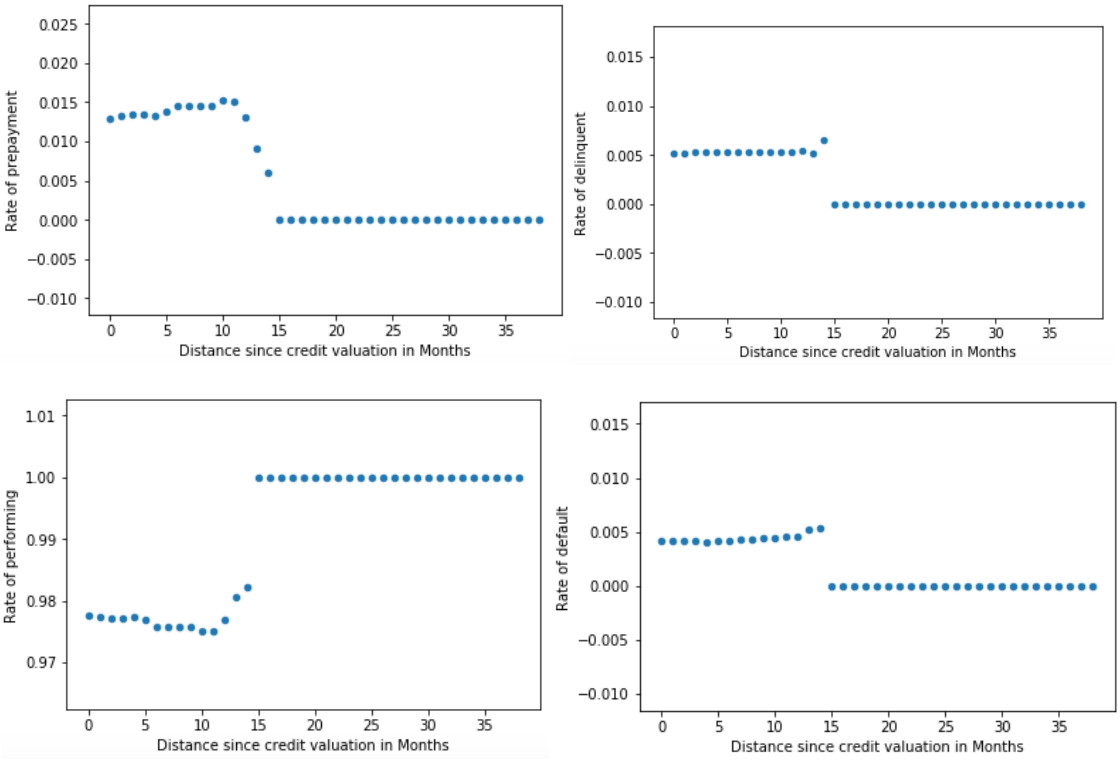

Figure 30 - Distances since original credit evaluation for sample 1

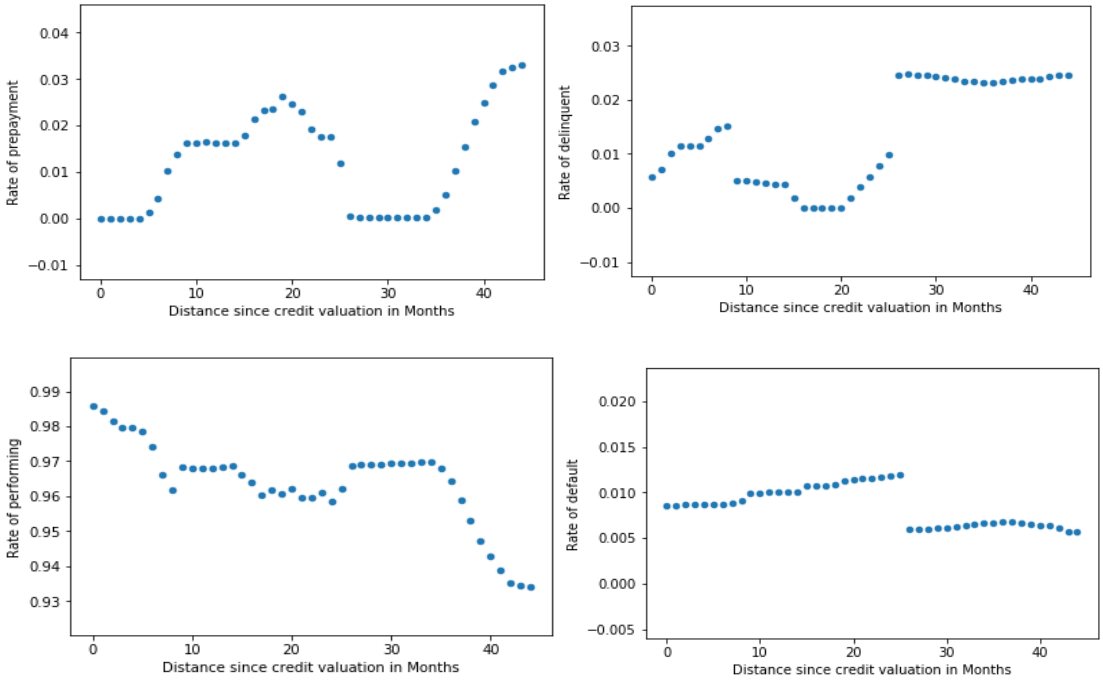

Figure 31 - Distances since original credit evaluation for sample 2

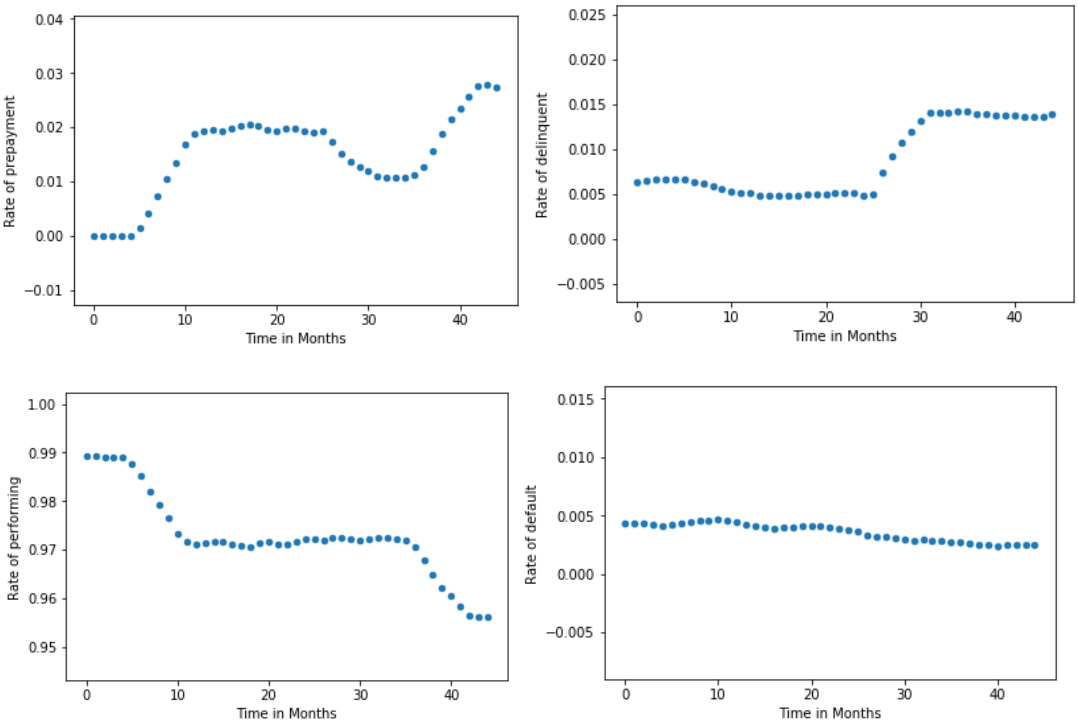

Figure 32 – Effect of Time for sample 1

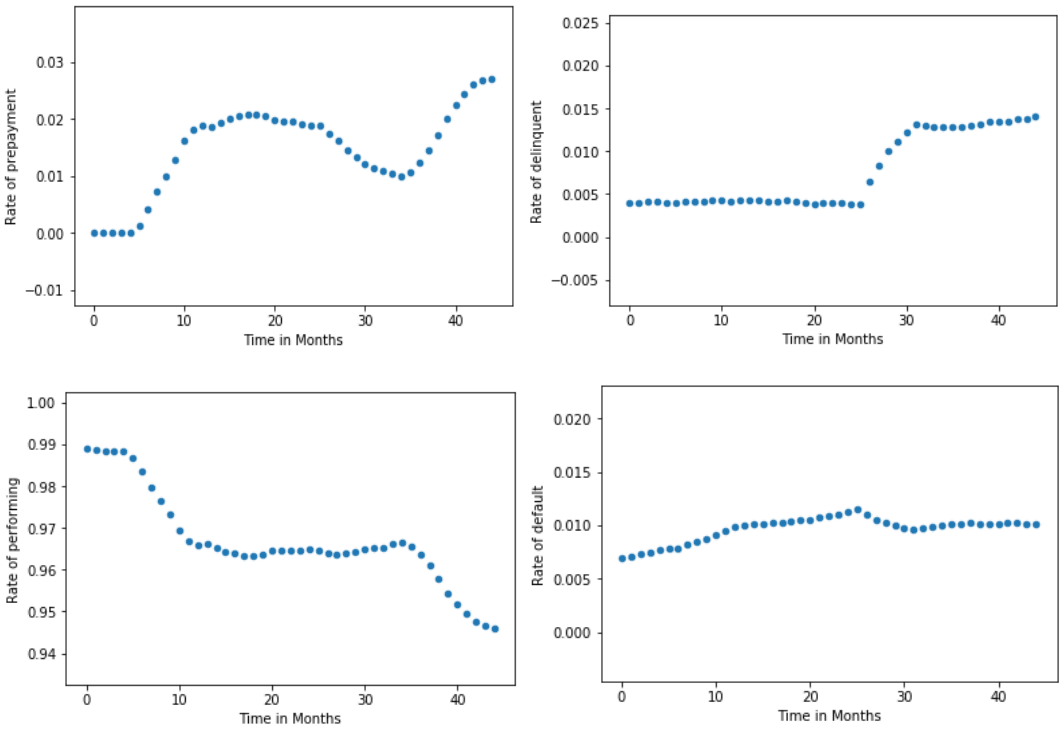

Figure 33 - Effect of Time for sample 2

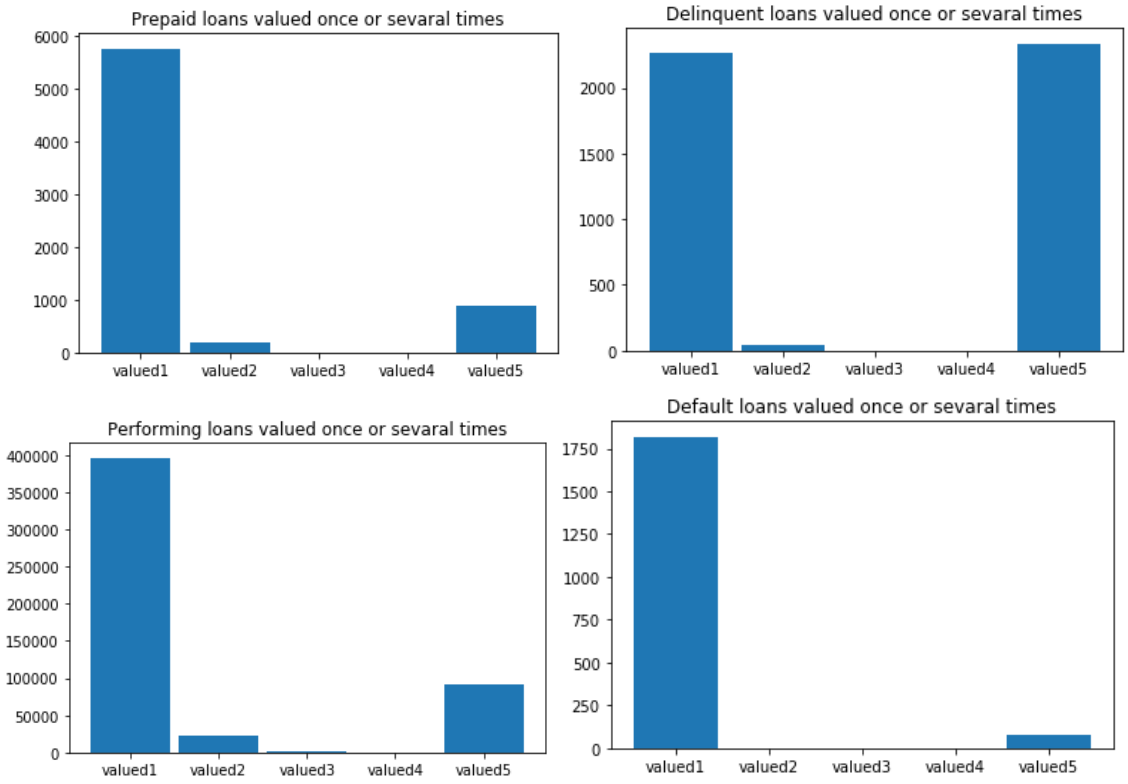

Figure 34 - Number of valuation per loan for sample 1

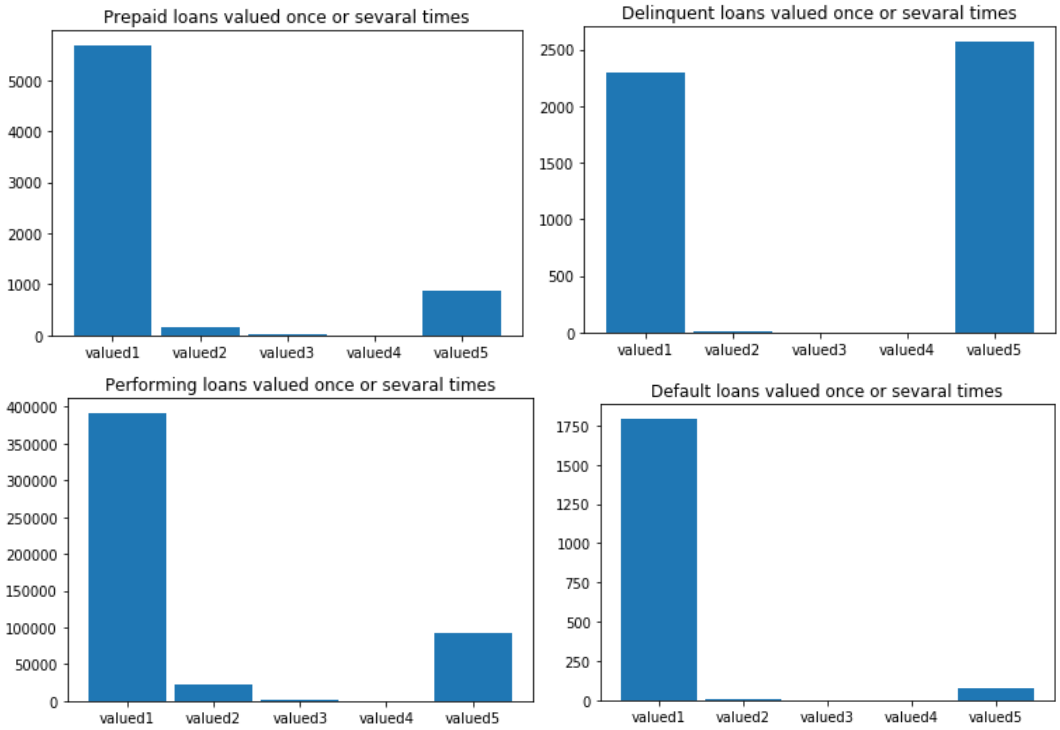

Figure 35 - Number of valuation per loan for sample 2

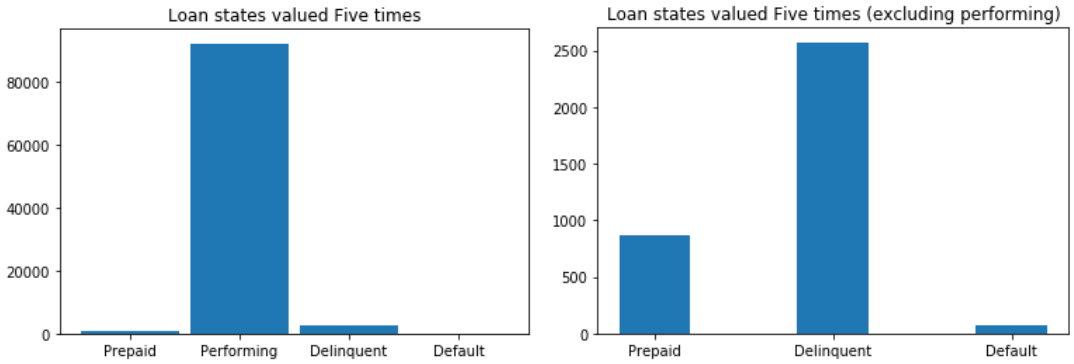

Figure 36 – Observations whose valuations changed five times, for sample 1

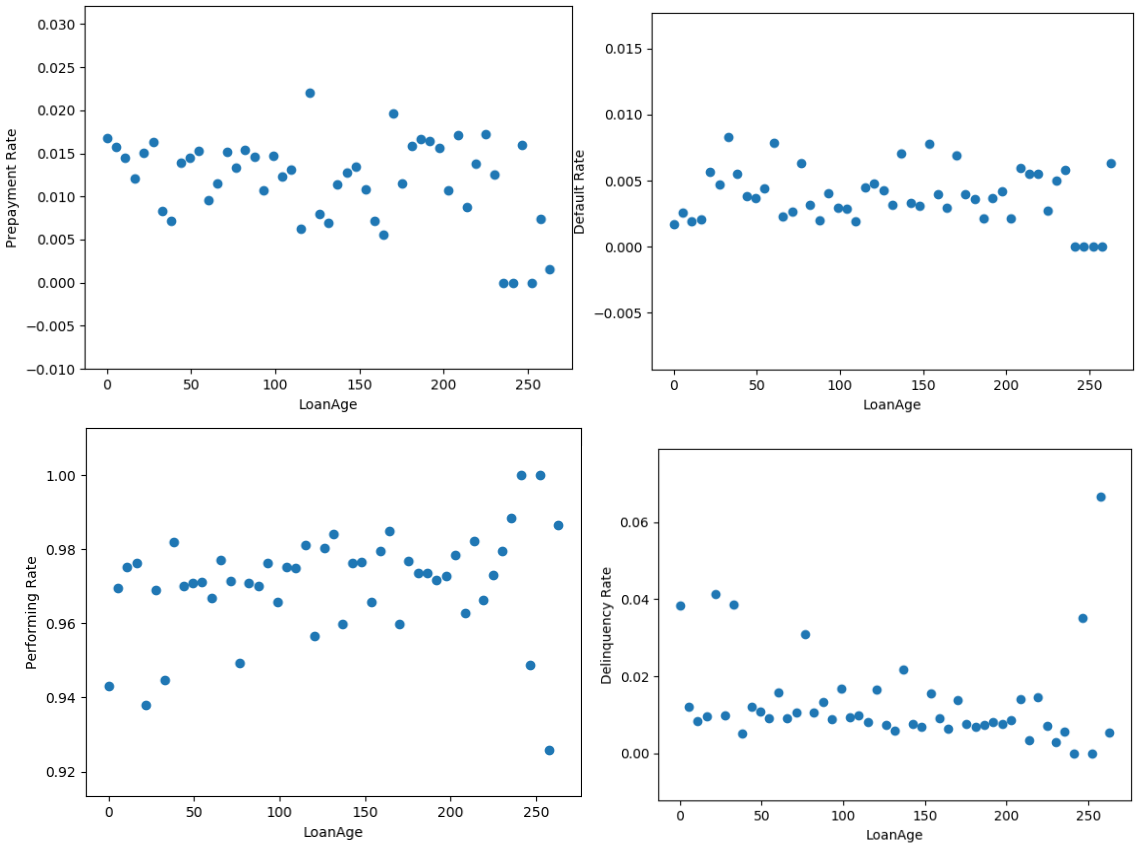

Figure 37 – Loan Age for sample 2

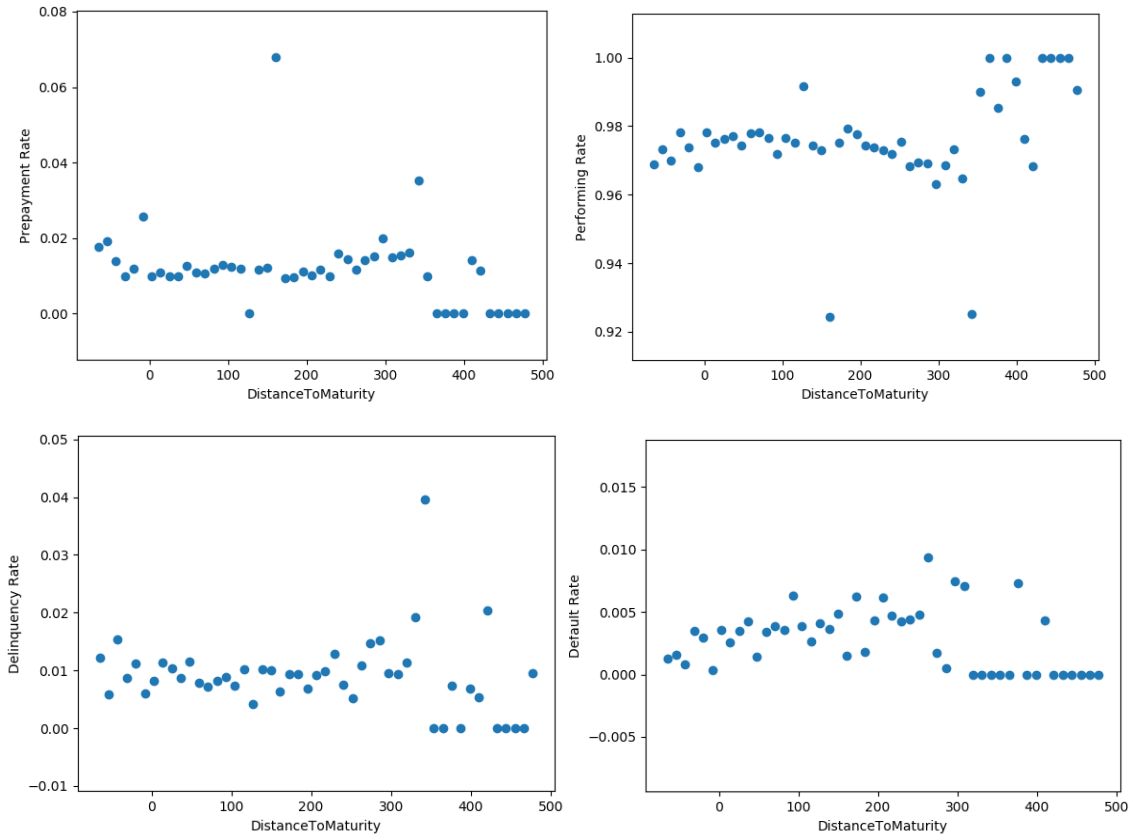

Figure 38 – Distance to Maturity for sample 2

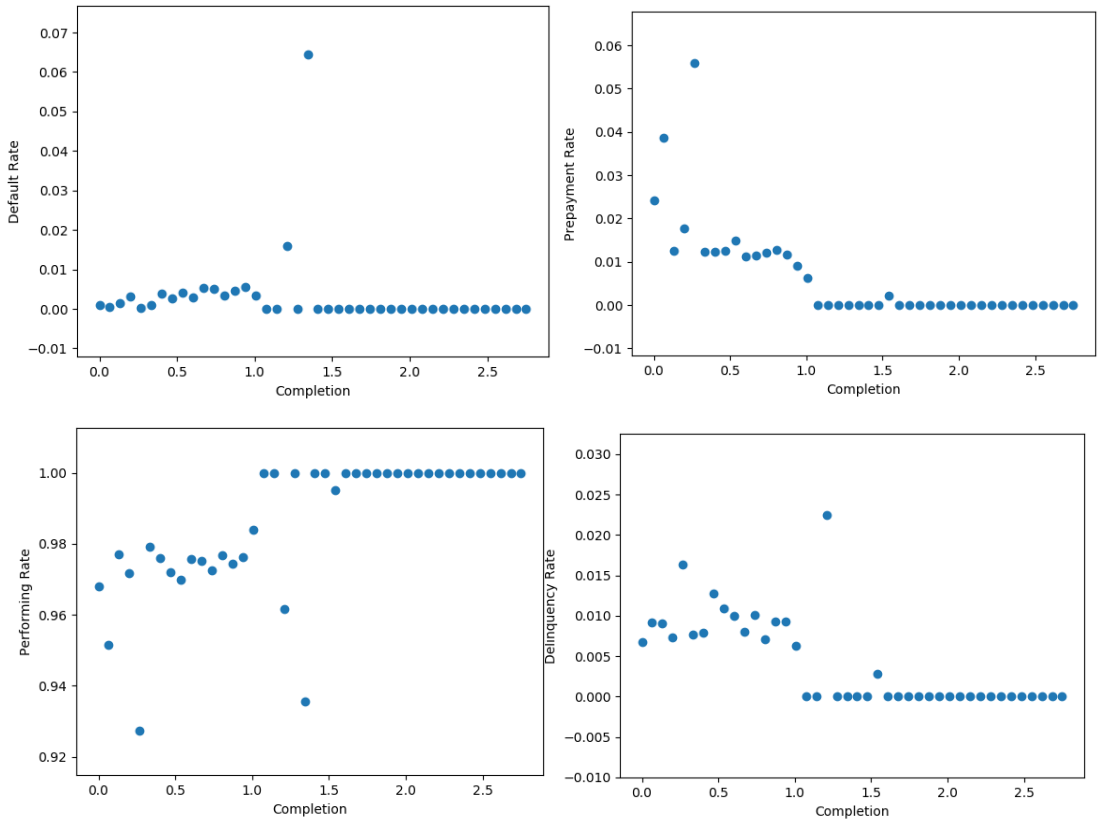

Figure 39 – Percentage of Loan Completion for sample 1

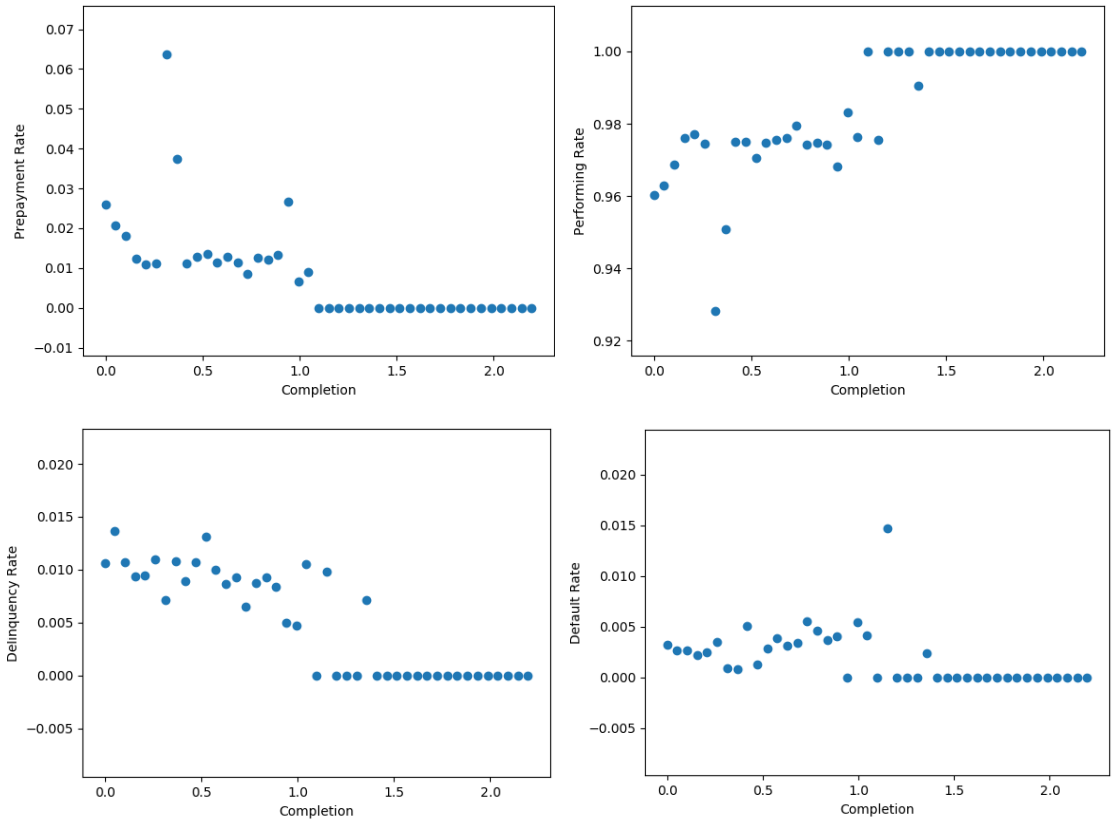

Figure 40 – Percentage of Loan Completion for sample 2

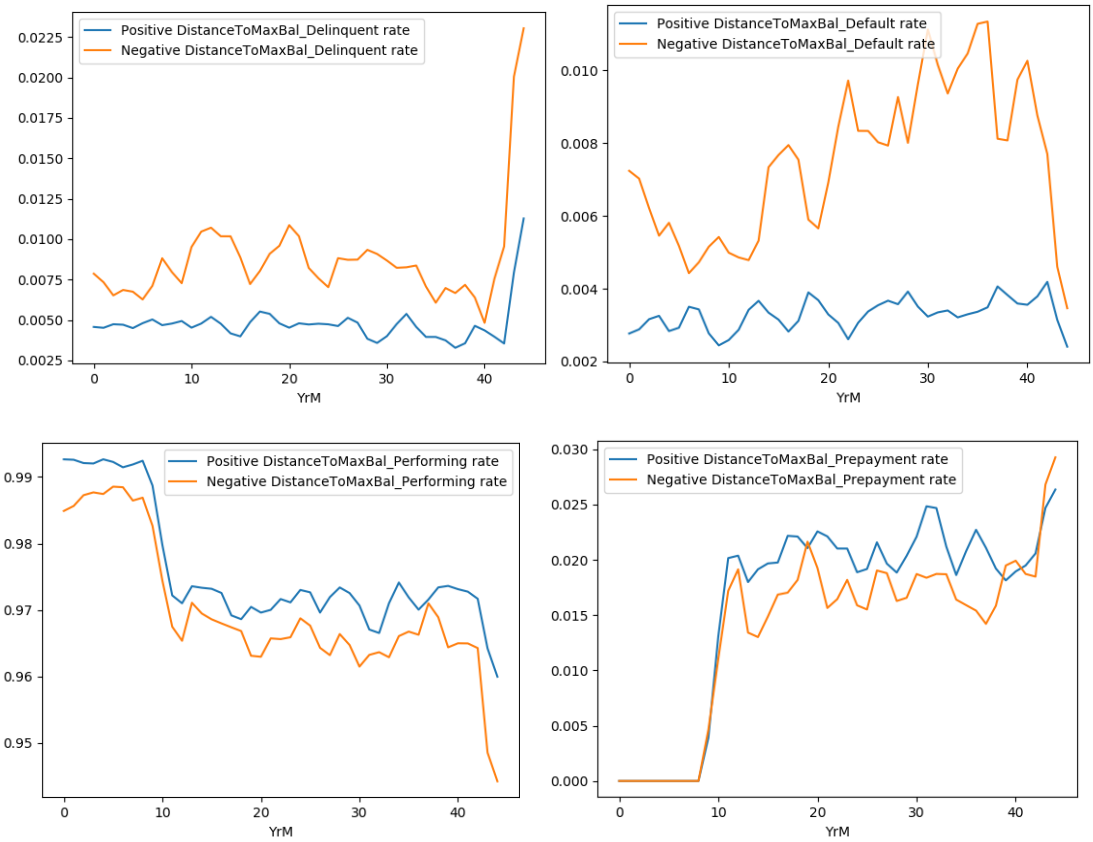

Figure 41 – Distance to Maximum Balance for sample 2

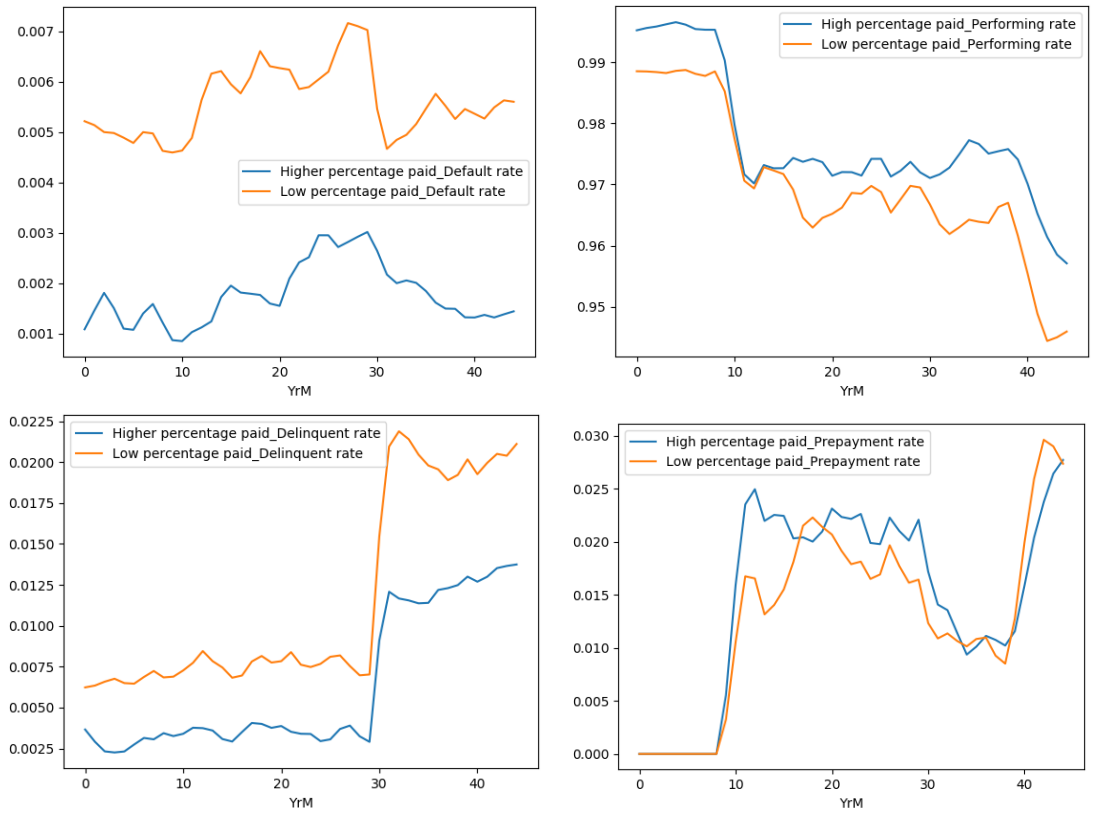

Figure 42 – Percentage of Loan Paid for sample 2

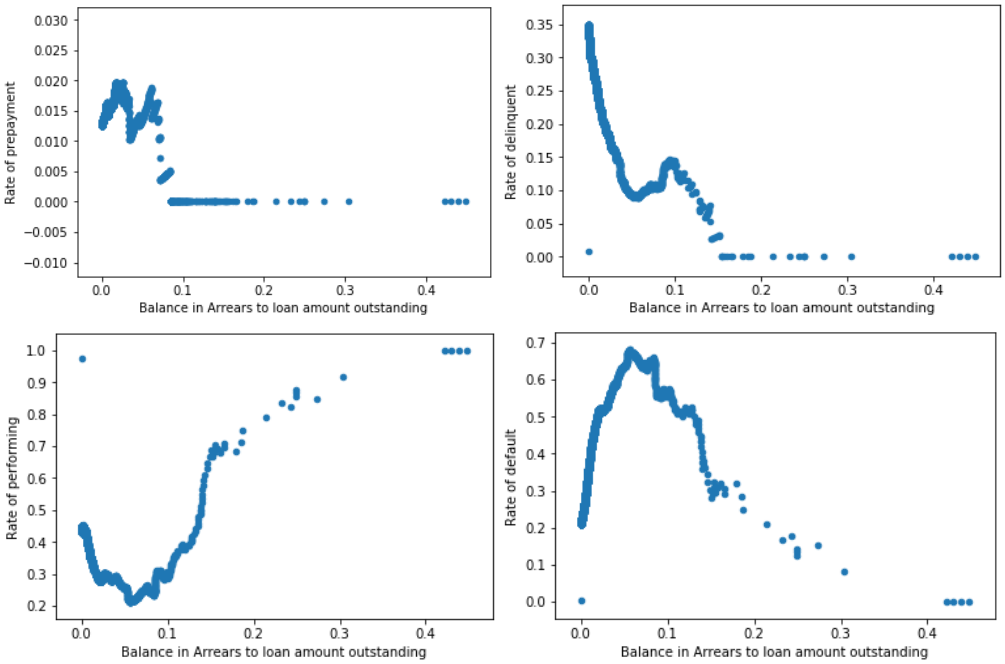

Figure 43 - Balance in arrears in proportion to loan's outstanding value for sample 1

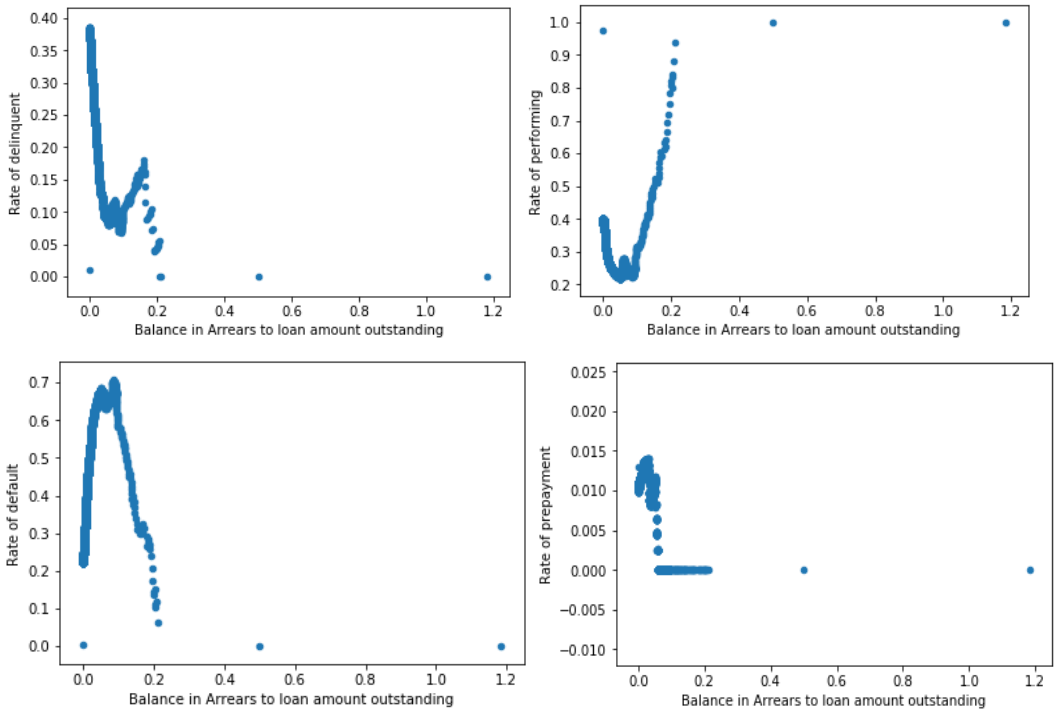

Figure 44 - Balance in arrears in proportion to loan's outstanding value for sample 2

Appendix C – List of Variables

Variable	Explanation	Type
LoanStatus	Current Loan Status	Categorical
ArrearsEndBalance	Section 4.1.8 - Balance in arrears to loan value outstanding	Continuous
ArrearsIncome	Section 4.1.7 - Cummulative amount in arrears to annual income	Continuous
CCJNumberSatisfied	Number of Satisfied County Court Judgements or equivalent	Continuous
OccupancyType	Type of property occupancy	Categorical
CCJValueSatisfied	Total Value of Satisfied County Court Judgements or equivalent	Continuous
PaymentType	Principal payment type	Categorical
Instalmentpropincome	Section 4.1.7 - Monthly installment to income ratio (monthly)	Continuous
IsUnderLitigation	Flag to indicate litigation proceedings underway	Dummy
IsFirstTimeBuyer	First time buyer flag	Dummy
EmploymentStatus	Employment status of the primary applicant	Categorical
YrM	Date of the observation	Continuous
Lien	Seniority on liquidation of property	Categorical
BureauScoreType	Type of scorecard provided	Categorical
BureauScoreProvider	Name of who has provided the score	Categorical
BorrowerType	The type of borrower	Categorical
Originator	Lender that advanced the original loan	Categorical
ClassOfBorrower	Class of borrower based on credit scoring or other classification	Categorical
CurrentInterestRateIndex	Reference rate off which the mortgage interest rate is set	Categorical
AreFurtherAdvancesPossible	Possibility to have further advances i.e. advances above the original loan balance.	Dummy
CreditQuality	Originators own definition of borrower credit quality	Categorical
BureauScoreValue	Borrower's score	Continuous
PctOfPrepaymentsAllowedPerYear	Percentage amount of pre-payments allowed under the product per year	Continuous
PropertyType	Property type/usage	Categorical
CurrentInterestRate	Current interest rate (%)	Continuous
BankruptcyOrIVAFlag	Mortgage's Bankruptcy or Individual Voluntary Arrangement Flag	Dummy
OriginationChannel	Origination channel, arranging bank or division for the loan	Categorical
PrimaryIncomeVerification	Income Verification for Primary Income	Categorical
LoanTermInMonths	Contractual length of the loan	Continuous

Completion	Section 4.1.4 - Loan Age in percentage of contractual term	Continuous
PastToCurrentLoan	Relation between past loans and current loans	Continuous
DistanceToMaturity	Section 4.1.4 - Number of months to contractual termination	Continuous
PaymentDue	Dynamic contractual payment due	Continuous
GeographicRegion	The region description of where the property is located	Categorical
Purpose	Purpose of the loan	Categorical
HasRightToBuy	Loan's right to buy flag	Dummy
OriginationValuationType	Valuation type at origination	Categorical
CurrentValuationType	Valuation type of last evaluation	Categorical
percentagepaid	Section 4.1.5 - Percentage of the Loan paid	Continuous
DistanceFromOriginalValuation	Section 4.1.2 - Distance since original property valuation	Continuous
DsitanceToMaxBal	Section 4.1.5 - How far from limit debt is the current loan	Continuous
CCJNumberUnsatisfied	Number of Unsatisfied County Court Judgements or equivalent	Continuous
TimeSinceStatus	Section 4.1.2 - Distance since last loan status	Continuous
PaymentFrequency	Frequency of payments due, i.e. number of months between payments	Categorical
InterestRateType	Interest rate type	Categorical
CurrentLTV	Section 4.1.1.1 - Updated Loan-to-Value Ratio	Continuous
NumberOfDebtors	Number of borrowers to the loan	Continuous
IsRegulatedLoan	Indication if the loan is regulated (Y) or not	Dummy
Incentivesell	Section 4.1.1.2 - Ability to cover the loan with property value LTV with last official valuation	Continuous
LoanAge	Section 4.1.4 - Age of the mortgage	Continuous
AgeOfBorrower	Section 4.1.6 - Age of the borrower, dynamic	Continuous
RepaymentMethod	Type of principal repayment	Categorical
DistanceFromValuation	Section 4.1.2 - Distance since last property valuation	Continuous
DistanceFromEvaluation	Section 4.1.2 - Distance since credit evaluation	Continuous
InterestRateResetIntervalInMonths	The interval in months at which the interest rate is adjusted (for floating loans)	Continuous
ValuationVolatility	Section 4.1.3 - Number of Valuation per loan	Continuous

Appendix D – Loss Graphs

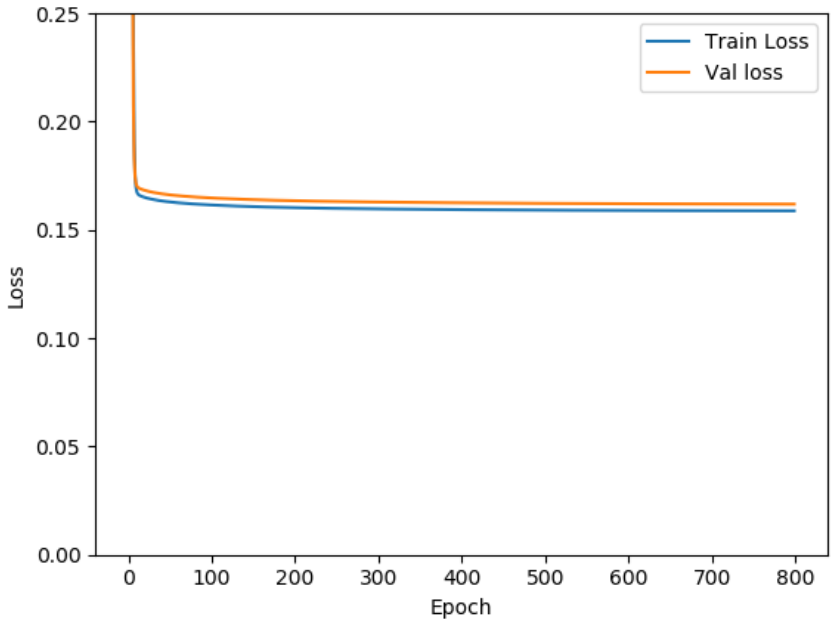

Figure 45: Training loss for network trained without weighting the loss function. There is a sharp initial decrease in loss but there is no more learning henceforth, the network mostly predicts the observations will be in state performing in 1-year time (as observed in figure 47, appendix G, no conclusion or rule can be drawn from the ROC curve analysis).

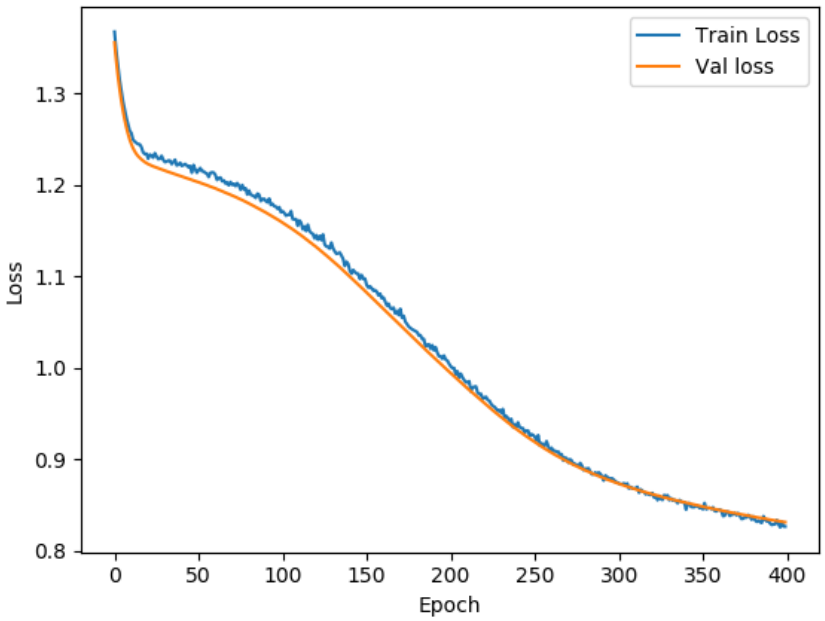

Figure 46:Figure 5:Training loss for network trained simply by using inverse proportion (M1_1 This network was trained with only 400 epochs for time reasons but as we can

see a very similar evolution to M1, the two models showing somewhat similar in general.

Appendix E – Variable Significance

Variable	Loss	Accuracy	Difference
LoanStatus	0.985447	0.687760	0.451282
ArrearsEndBalance	0.902535	0.775095	0.368371
ArrearsIncome	0.766091	0.869249	0.231926
CCJNumberSatisfied	0.610604	0.953899	0.076439
CCJValueSatisfied	0.596355	0.955061	0.062191
PaymentType	0.583634	0.958404	0.049470
Instalmentpropincome	0.579593	0.958432	0.045429
IsUnderLitigation	0.578550	0.962575	0.044385
IsFirstTimeBuyer	0.574679	0.964479	0.040514
EmploymentStatus	0.572628	0.962375	0.038463
YrM	0.558910	0.961727	0.024745
BureauScoreType	0.556429	0.964460	0.022264
BureauScoreProvider	0.555852	0.965641	0.021687
Originator	0.547932	0.965993	0.013768
CurrentInterestRateIndex	0.545102	0.967850	0.010937
AreFurtherAdvancesPossible	0.544534	0.966698	0.010369
BureauScoreValue	0.542219	0.971126	0.008054
PctOfPrepaymentsAllowedPerYear	0.541930	0.965546	0.007765
PropertyType	0.540369	0.970145	0.006204
CurrentInterestRate	0.537916	0.969603	0.003751
BankruptcyOrIVAFlag	0.535398	0.969669	0.001233
OriginationChannel	0.534939	0.969603	0.000774

Table 1: Significance, top variables, ('Difference' computes the difference between the loss with the variable left out and the original loss - 0.534165)

Variable	Loss	Accuracy	Difference
PrimaryIncomeVerification	0.534771	0.968165	0.000606
LoanTermInMonths	0.534005	0.969517	-0.000160
Completion	0.531504	0.968060	-0.002661
PaymentDue	0.530519	0.969660	-0.003646
DistanceToMaturity	0.530501	0.969155	-0.003663
GeographicRegion	0.529501	0.968793	-0.004663
Purpose	0.529308	0.969469	-0.004856
HasRightToBuy	0.528227	0.968860	-0.005938
OriginationValuationType	0.527371	0.969345	-0.006794
CurrentValuationType	0.526936	0.969250	-0.007228
percentagepaid	0.526637	0.970117	-0.007528
DistanceFromOriginalValuation	0.526194	0.968936	-0.007971
DsitanceToMaxBal	0.522988	0.968965	-0.011176
CCJNumberUnsatisfied	0.522830	0.970355	-0.011335
TimeSinceStatus	0.522030	0.969698	-0.012135
PaymentFrequency	0.519864	0.970641	-0.014301
InterestRateType	0.518847	0.970736	-0.015318
CurrentLTV	0.518812	0.970831	-0.015353
NumberOfDebtors	0.518601	0.971307	-0.015564
IsRegulatedLoan	0.518226	0.971450	-0.015938
Incentivesell	0.516722	0.971507	-0.017442
LoanAge	0.516486	0.970603	-0.017678
AgeOfBorrower	0.511484	0.970793	-0.022681
RepaymentMethod	0.506188	0.971555	-0.027977
DistanceFromValuation	0.503963	0.972174	-0.030201
DistanceFromEvaluation	0.502471	0.971850	-0.031694
InterestRateResetIntervalInMonths	0.495225	0.971802	-0.038940
ValuationVolatility	0.493733	0.970107	-0.040432

Table 2:Significance, low/negative impact variables

Variable	Loss	Accuracy	Difference
LoanStatus	0.9769	0.691857	0.458865
ArrearsEndBalance	0.893095	0.78079	0.375061
ArrearsIncome	0.754848	0.877529	0.236814
CCJNumberSatisfied	0.597191	0.96172	0.079157
CCJValueSatisfied	0.58278	0.963094	0.064746
PaymentType	0.569989	0.966141	0.051955
Instalmentpropincome	0.565772	0.966316	0.047737
IsUnderLitigation	0.565134	0.970514	0.047099
IsFirstTimeBuyer	0.560629	0.972565	0.042594
EmploymentStatus	0.559115	0.97032	0.041081
YrM	0.544627	0.969808	0.026593
BureauScoreType	0.542292	0.972652	0.024257
BureauScoreProvider	0.541794	0.973455	0.02376
Originator	0.533653	0.973997	0.015619
CurrentInterestRateIndex	0.530946	0.975835	0.012912
AreFurtherAdvancesPossible	0.530384	0.974751	0.012349
BureauScoreValue	0.528335	0.978688	0.0103
PctOfPrepaymentsAllowedPerYear	0.527486	0.973465	0.009451
PropertyType	0.52603	0.977789	0.007996
CurrentInterestRate	0.523709	0.977402	0.005675
BankruptcyOrIVAFlag	0.52119	0.97744	0.003155
OriginationChannel	0.520839	0.977634	0.002804
PrimaryIncomeVerification	0.520579	0.976164	0.002545
MismatchMat	0.519783	0.97745	0.001749

Table 3: Variable significance for initially performing loans (for the most significant variables). Initial loss was 0.518034. Borrower characteristics still dominate the top positions

If we were interested in measuring the impact of the variable *LoanStatus*, that is the loan state 12 months prior, we would set all binary variables corresponding to each of the possible classes to zero.

We would go from the setting in table 7 to table 8.

Obs	…	LoanStatus=0	LoanStatus=1	LoanStatus=2	LoanStatus=3	…
1	…	1	0	0	0	…
…	…	…	…	…	…	…
i-1	…	0	0	1	0	…
i-1	…	0	1	0	0	…
i+1	…	0	1	0	0	…
…	…	…	…	…	…	…
I-1	…	0	0	1	0	…
I	…	0	0	0	1	…

Table 4:Original attributes

Obs	…	LoanStatus=0	LoanStatus=1	LoanStatus=2	LoanStatus=3	…
1	…	0	0	0	0	…
…	…	…	…	…	…	…
i-1	…	0	0	0	0	…
i-1	…	0	0	0	0	…
i+1	…	0	0	0	0	…
…	…	…	…	…	…	…
I-1	…	0	0	0	0	…
I	…	0	0	0	0	…

Table 5:New attributes for testing

Same would happen for continuous variables since there are several binary variables representing the possible intervals the continuous variables can take.

Table 6:Variable significance for initially performing loans (for the most significant variables). Initial loss was 0.518034. Borrower characteristics still dominate the top positions.

Appendix F – Variable Impact

The following tables exemplify the steps to obtain variable impact discussed in Section 6.2. All values are exemplary except for the intervals for *CurrentInterestRate.*

		CurrentInterestRate (%)			
Obs	**…**	**]0,0.408]**	**]0.408,0.816]**	**…**	**]9.382,9.79]**	**…**
1	…	1	0	…	0	…
…	…	…		…	…	…
i-1	…	0	0	…	0	…
i-1	…	0	0	…	0	…
i+1	…	0	1	…	0	…
…	…	…		…	…	…
I-1	…	0	0	…	0	…
I	…	0	0	…	1	…

Table 7:Test set (CurrentInterestRate columns highlighted)

		CurrentInterestRate (%)			
Obs	**…**	**]0,0.408]**	**]0.408,0.816]**	**…**	**]9.382,9.79]**	**…**
1	…	1	0	…	0	…
…	…	…		…	…	…
i-1	…	1	0	…	0	…
i-1	…	1	0	…	0	…
i+1	…	1	0	…	0	…
…	…	…		…	…	…
I-1	…	1	0	…	0	…
I	…	1	0	…	0	…

Table 8:Make all observation belong to the first interval

	Probability Distribution for]0,0.408]			
Obs	**Prepayment - 0**	**Performing - 1**	**Arrears - 2**	**Default - 3**
1	0.29504	0.479749	0.155397	0.069814
…	**…**	**…**	**…**	**…**
i-1	0.246259	0.634933	0.092731	0.026078
i-1	0.233278	0.695198	0.057117	0.014407
i+1	0.296935	0.469743	0.164897	0.068425
…	**…**	**…**	**…**	**…**
I-1	0.283988	0.679874	0.025893	0.010245
I	0.224886	0.712858	0.049115	0.01314

Table 9:Predictions from the transformed test set in table 8

The process in tables 8 and 9 is repeated for every interval in the variable (tables 10 and 11 exemplify the next interval).

		CurrentInterestRate (%)				
Obs	**…**	**]0,0.408]**	**]0.408,0.816]**	**…**	**]9.382,9.79]**	**…**
1	…	0	1	…	0	…
…	…	…	…	…	…	…
i-1	…	0	1	…	0	…
i-1	…	0	1	…	0	…
i+1	…	0	1	…	0	…
…	…	…	…	…	…	…
I-1	…	0	1	…	0	…
I	…	0	1	…	0	…

Table 10:Same process and in Table 12 for the next interval

	Probability Distribution]0.408,0.816]			
Obs	**Prepayment - 0**	**Performing - 1**	**Arrears - 2**	**Default - 3**
1	0.287695	0.658554	0.039198	0.014553
…	**…**	**…**	**…**	**…**
i-1	0.27846	0.485536	0.16842	0.067584
i-1	0.188195	0.238544	0.388725	0.184536
i+1	0.269032	0.465684	0.185318	0.079966
…	**…**	**…**	**…**	**…**
I-1	0.299601	0.56062	0.100252	0.039527
I	0.288267	0.683982	0.019939	0.007812

Table 11:Predictions from the transformed test set in table 13

Once the predictions are obtained for all observations, a sample wise distribution is calculated as by averaging each class across all observations, obtaining the following:

Average distribution for]0,0.408]

Prepayment - 0	**Performing - 1**	**Arrears - 2**	**Default - 3**
0.287913	0.661227	0.035746	0.015114

Average distribution for]0.408,0.816]

Prepayment - 0	**Performing - 1**	**Arrears - 2**	**Default - 3**
0.281243	0.686612	0.023347	0.008798

And so on for every interval.

Afterwards the absolute change in each probability caused from moving from one interval to the other is calculated

	Absolute change between]0,0.408] and]0.408,0.816]
Prepayment - 0	|0.281243 - 0.287913| = 0.00667
Performing - 1	|0.686612 - 0.661227| = 0.025385
Arrears - 2	|0.23347 - 0.035746| = 0.012399
Default - 3	|0.008798 - 0.015114| = 0.006316

This is done for every pair of sequential intervals (]0,0.408] and]0.408,0.816],]0.408,0.816] and]0.816,1.224],]0.816,1.224] and]1.224,1.632], …). In this case, there

will be 24 values (25 intervals) for each state. The expected absolute change for each probability due to change in the variable in question (in this case *CurrentInterestRate*) is calculated as arithmetic average across the values of the 24 combinations of sequential intervals.

The results are presented in tables 12 and 13 for the impact on general probability and tables 14 and 15 for initially performing loans (values presented in percentage points).

	Prepayment	Performing	Arrears	Default
LoanStatus	1.846065	11.012204	9.391025	3.467245
BureauScoreValue	0.839868	2.889824	2.592207	1.123105
ArrearsEndBalance	1.144978	2.601697	2.656215	1.097659
DistanceFromValuation	0.795220	2.588846	2.444024	0.885494
Completion	0.638874	2.265495	2.145877	0.746648
CCJNumberSatisfied	0.251472	2.251598	1.640678	0.641636
ArrearsIncome	1.003235	2.215891	2.173412	1.065788
Incentivesell	0.555463	2.211835	2.063457	0.696191
EmploymentStatus	0.352837	2.084448	1.842334	0.589121
Instalmentpropincome	0.486972	2.034207	1.816547	0.698208
IsFirstTimeBuyer	0.265059	2.029741	1.441861	0.520655
TimeSinceStatus	0.649388	2.008971	1.776499	0.818509
CurrentLTV	0.579513	1.958099	1.844078	0.693535
LoanAge	0.488520	1.889044	1.699989	0.653174
Originator	0.233846	1.881147	1.559456	0.555538
NumberOfDebtors	0.207429	1.876289	1.276809	0.506644
ValuationVolatility	0.338154	1.868719	1.118677	0.411894
CurrentValuationType	0.159420	1.865614	1.443668	0.498386
RepaymentMethod	0.118721	1.829141	1.310538	0.466452
DistanceFromEvaluation	0.335587	1.774115	1.465801	0.574627
OriginationValuationType	0.159626	1.766135	1.332332	0.490383
GeographicRegion	0.438246	1.747629	1.585698	0.535560
PaymentType	0.153643	1.741057	1.250373	0.488905
AgeOfBorrower	0.493419	1.677540	1.520246	0.577034
IsRegulatedLoan	0.143194	1.642737	1.151276	0.405021

Table 12:Variable Impact, first set of variables

	Prepayment	Performing	Arrears	Default
MismatchMat	0.235616	1.589867	1.288499	0.457640
Purpose	0.168416	1.552940	1.216253	0.427922
HasRightToBuy	0.188008	1.540673	1.094948	0.390797
DsitanceToMaxBal	0.178044	1.503080	1.113689	0.454942
PropertyType	0.150590	1.501467	1.001424	0.379959
PaymentDue	0.126841	1.410124	1.133486	0.386820
CurrentInterestRateIndex	0.160334	1.407725	1.190458	0.377599
BureauScoreProvider	0.236461	1.393339	1.058805	0.380547
CurrentInterestRate	0.126686	1.313257	1.006729	0.353156
percentagepaid	0.201565	1.201920	0.917296	0.363623
BankruptcyOrIVAFlag	0.113043	1.161340	0.962611	0.311773
PctOfPrepaymentsAllowedPerYear	0.176655	0.977350	0.742402	0.268969
PrimaryIncomeVerification	0.217539	0.960892	0.616754	0.258679
InterestRateType	0.181004	0.947945	0.658720	0.221341
CCJValueSatisfied	0.148953	0.940509	0.690017	0.292997
DistanceToMaturity	0.211341	0.937570	0.712560	0.259920
DistanceFromOriginalValuation	0.172180	0.888320	0.647363	0.245763
BureauScoreType	0.148156	0.858033	0.732172	0.274016
OriginationChannel	0.151969	0.812142	0.639277	0.202221
InterestRateResetIntervalInMonths	0.036250	0.767362	0.614813	0.188796
CCJNumberUnsatisfied	0.110419	0.736576	0.617250	0.229744
AreFurtherAdvancesPossible	0.257580	0.371099	0.140563	0.074633

Table 13:Variable Impact, second set of variables

	Prepaid	Performing	Arrears	Default
CurrentValuationType	1.818469	11.074799	9.434116	3.459150
BankruptcyOrIVAFlag	1.085524	3.060122	3.031307	1.114341
BureauScoreProvider	0.931770	2.220437	2.342466	0.809744
BureauScoreType	0.921142	2.107678	2.214019	0.814804
AreFurtherAdvancesPossible	0.682766	1.230228	1.406682	0.506313
EmploymentStatus	0.393435	1.363112	0.947629	0.762459
GeographicRegion	1.376015	2.804182	3.146393	1.034009
HasRightToBuy	0.789579	1.636198	1.758030	0.667747
InterestRateType	0.538993	1.116675	1.200746	0.454922
IsFirstTimeBuyer	0.507253	0.936426	1.079728	0.401403
IsRegulatedLoan	1.108826	2.361511	2.403404	0.933044
OriginationChannel	0.376801	0.989002	0.999226	0.366575
OriginationValuationType	0.803005	1.449256	1.771700	0.567497
Originator	1.018140	2.003620	2.184744	0.837016
PaymentType	0.009011	0.016222	0.016854	0.008379
PrimaryIncomeVerification	0.901147	1.622164	1.685423	0.837887
PropertyType	0.380691	0.670154	0.740985	0.323655
Purpose	0.766325	1.366727	1.326885	0.806163
RepaymentMethod	0.715870	1.177879	1.347383	0.706385
CurrentInterestRateIndex	0.451281	0.684999	0.767132	0.442961
CurrentInterestRate	0.460832	0.586551	0.887752	0.180726
PaymentDue	0.533851	0.778762	0.769716	0.591005
CurrentLTV	0.544856	0.733312	0.731402	0.569840
LoanAge	0.277814	0.324474	0.303631	0.362504

Table 14:Variable Impact, first set of variables. Originally performing loans.

	Prepaid	Performing	Arrears	Default
DistanceToMaturity	0.116248	0.120936	0.224341	0.281281
Completion	0.092014	0.090493	0.208385	0.179070
Incentivesell	0.108528	0.094820	0.231759	0.235395
percentagepaid	0.049369	0.048126	0.151433	0.119944
Instalmentpropincome	0.037832	0.040789	0.126478	0.110986
BureauScoreValue	0.041644	0.039215	0.194331	0.154045
InterestRateResetIntervalInMonths	0.023588	0.022019	0.149206	0.127511
PctOfPrepaymentsAllowedPerYear	0.041851	0.045582	0.112095	0.078918
CCJNumberSatisfied	0.034977	0.033796	0.163665	0.133429
CCJNumberUnsatisfied	0.036057	0.016388	0.341988	0.289547
CCJValueSatisfied	0.039110	0.033341	0.105283	0.048989
NumberOfDebtors	0.044508	0.042301	0.273380	0.238483
AgeOfBorrower	0.058391	0.061615	0.018916	0.138922
ArrearsIncome	0.067343	0.052387	0.179636	0.125378
ArrearsEndBalance	0.029090	0.025309	0.134868	0.116340
DistanceFromEvaluation	0.027050	0.022693	0.172135	0.142843
DistanceFromValuation	0.024999	0.021682	0.138843	0.121235
DistanceFromOriginalValuation	0.024217	0.018616	0.152328	0.133933
TimeSinceStatus	0.023226	0.018103	0.169547	0.134941
DsitanceToMaxBal	0.024192	0.019902	0.139713	0.116115
MismatchMat	0.017454	0.015674	0.115024	0.100147
ValuationVolatility	0.025298	0.021361	0.153642	0.134774

Table 15: Variable Impact, second set of variables. Originally performing loans.

	Prepayment	Performing	Arrears	Default
LoanStatus	1.846065	11.012204	9.391025	3.467245
ArrearsEndBalance	1.144978	2.601697	2.656215	1.097659
ArrearsIncome	1.003235	2.215891	2.173412	1.065788
BureauScoreValue	0.839868	2.889824	2.592207	1.123105
DistanceFromValuation	0.795220	2.588846	2.444024	0.885494

Table 16:Top variables affecting prepayment probability

	Prepayment	Performing	Arrears	Default
LoanStatus	1.846065	11.012204	9.391025	3.467245
BureauScoreValue	0.839868	2.889824	2.592207	1.123105
ArrearsEndBalance	1.144978	2.601697	2.656215	1.097659
DistanceFromValuation	0.795220	2.588846	2.444024	0.885494
Completion	0.638874	2.265495	2.145877	0.746648

Table 17:Top variables affecting performing probability

	Prepayment	Performing	Arrears	Default
LoanStatus	1.846065	11.012204	9.391025	3.467245
ArrearsEndBalance	1.144978	2.601697	2.656215	1.097659
BureauScoreValue	0.839868	2.889824	2.592207	1.123105
DistanceFromValuation	0.795220	2.588846	2.444024	0.885494
ArrearsIncome	1.003235	2.215891	2.173412	1.065788

Table 18:Top variables affecting arrears probability

	Prepayment	Performing	Arrears	Default
LoanStatus	1.846065	11.012204	9.391025	3.467245
BureauScoreValue	0.839868	2.889824	2.592207	1.123105
ArrearsEndBalance	1.144978	2.601697	2.656215	1.097659
ArrearsIncome	1.003235	2.215891	2.173412	1.065788
DistanceFromValuation	0.795220	2.588846	2.444024	0.885494

Table 19: Top variables affecting default probability

	Prepaid	Performing	Arrears	Default
CurrentValuationType	1.818469	11.074799	9.434116	3.459150
PaymentType	1.397750	2.623916	2.650842	1.370816
GeographicRegion	1.376015	2.804182	3.146393	1.034009
IsRegulatedLoan	1.108826	2.361511	2.403404	0.933044
BankruptcyOrIVAFlag	1.085524	3.060122	3.031307	1.114341

Table 20:Top variables affecting the probability of transitioning from performing to prepaid.

	Prepaid	Performing	Arrears	Default
CurrentValuationType	1.818469	11.074799	9.434116	3.459150
BankruptcyOrIVAFlag	1.085524	3.060122	3.031307	1.114341
GeographicRegion	1.376015	2.804182	3.146393	1.034009
PaymentType	1.397750	2.623916	2.650842	1.370816
IsRegulatedLoan	1.108826	2.361511	2.403404	0.933044

Table 21:Top variables affecting the probability of keeping performing

	Prepaid	Performing	Arrears	Default
CurrentValuationType	1.818469	11.074799	9.434116	3.459150
GeographicRegion	1.376015	2.804182	3.146393	1.034009
BankruptcyOrIVAFlag	1.085524	3.060122	3.031307	1.114341
PaymentType	1.397750	2.623916	2.650842	1.370816
IsRegulatedLoan	1.108826	2.361511	2.403404	0.933044

Table 22: Top variables affecting the probability of transitioning from performing to prepaid.

	Prepaid	Performing	Arrears	Default
CurrentValuationType	1.818469	11.074799	9.434116	3.459150
PaymentType	1.397750	2.623916	2.650842	1.370816
BankruptcyOrIVAFlag	1.085524	3.060122	3.031307	1.114341
GeographicRegion	1.376015	2.804182	3.146393	1.034009
IsRegulatedLoan	1.108826	2.361511	2.403404	0.933044

Table 23: Top variables affecting the probability of transitioning from performing to default.

Appendix G – Metrics and ROC curve

- ***Accuracy***: The number of correct classifications in proportion to the number of observations:

$$\frac{(True\ Positive + True\ Negative)(Prepaid + Performing + Delinquent + Default)}{(True + False)(Positive + Negative)(Prepaid + Performing + Delinquent + Default)}$$

This formula is used for ease of understanding, Prepaid, for example does not represent a value, but *True Positive Prepaid* does. Top expression would become: $True\ Positive\ Prepaid + True\ Positive\ Performing + True\ Positive\ Delinquet$ …

- **Sensitivity**: For each class, its proportion of correctly classified as true against the number of observations for this particular class:

$$\frac{True\ Positive}{True\ Positive + False\ Negative}$$

- **Specificity**: For each class, its proportion of correctly classified as false against the number of observations for this particular class:

$$\frac{True\ Negative}{True\ Negative + False\ Positive}$$

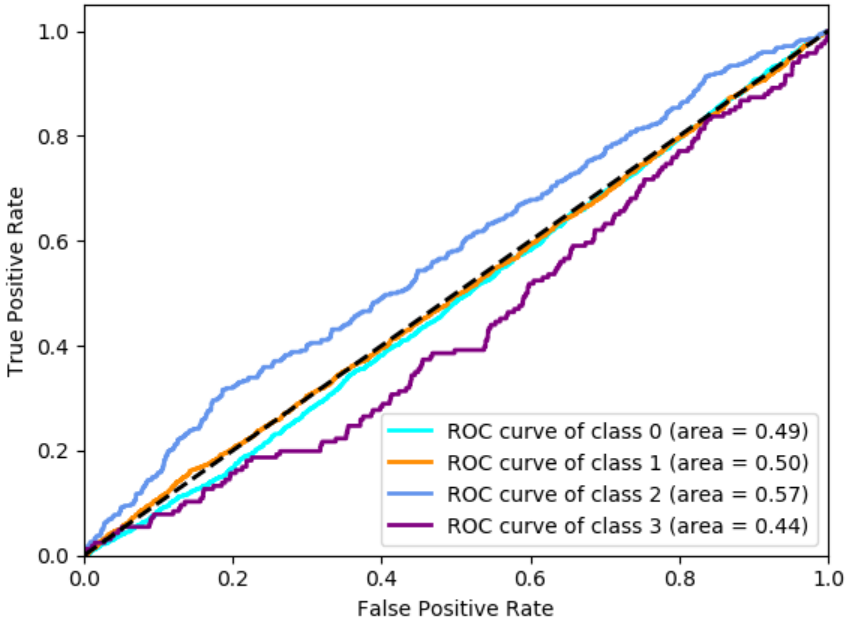

Figure 47: ROC for M1_0 - no weighting, transitioning from performing to all other states. We can see that if no weights are given to the loss function, no criteria can be set to accept if a performing loan will transition to a particular state based on the output probability distribution. It is effectively random.

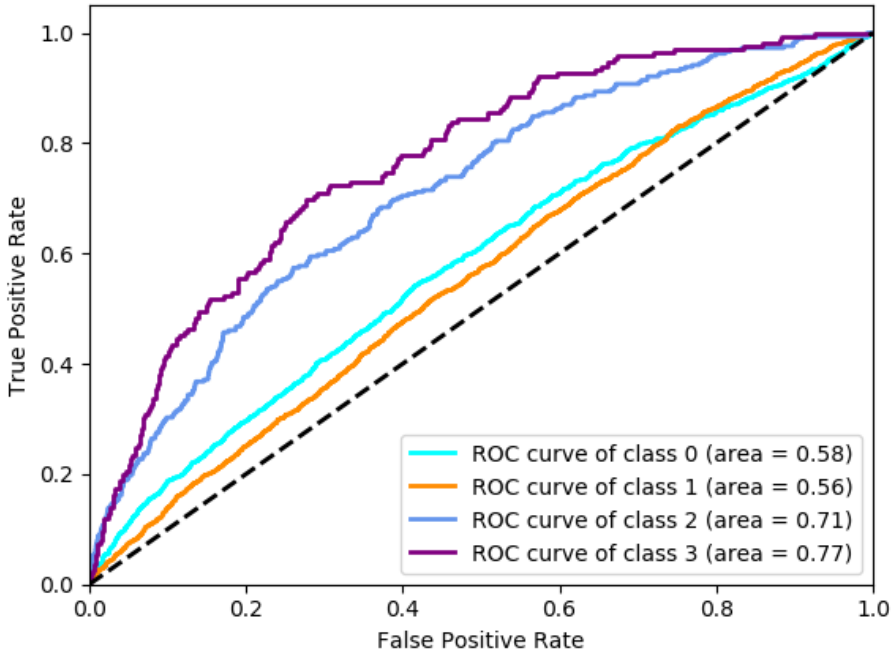

Figure 48: ROC curve for transition from

Appendix H – Predicted Default Rate Analysis

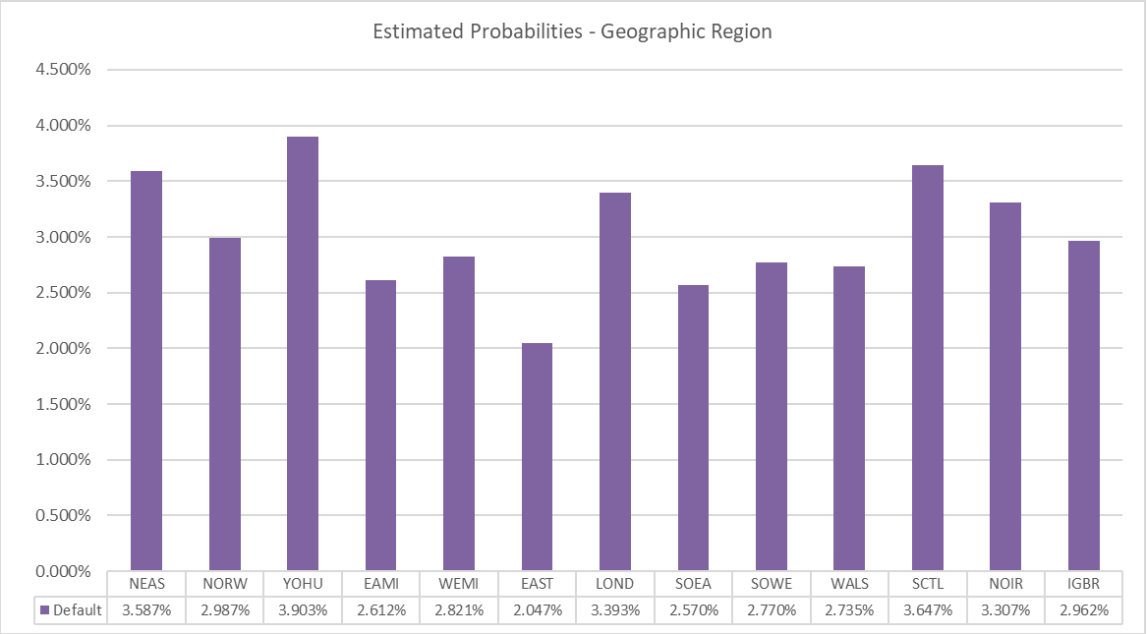

Figure 49

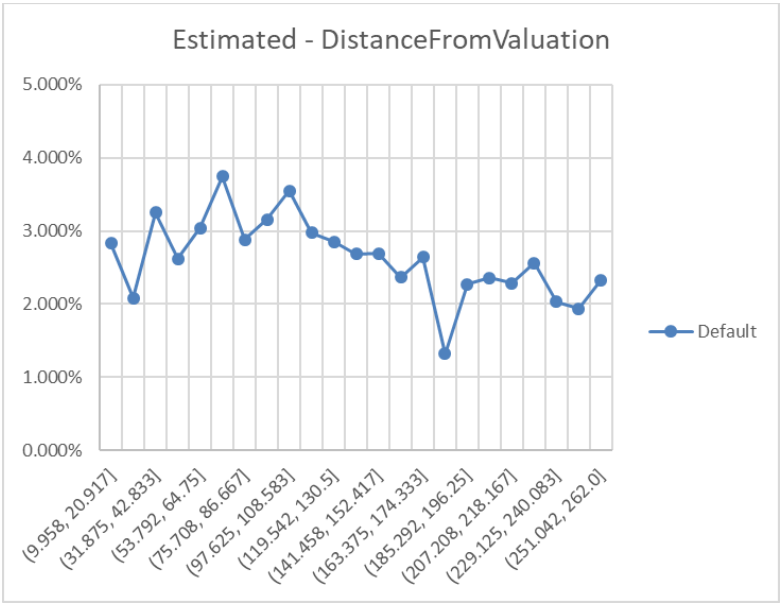

Figure 50

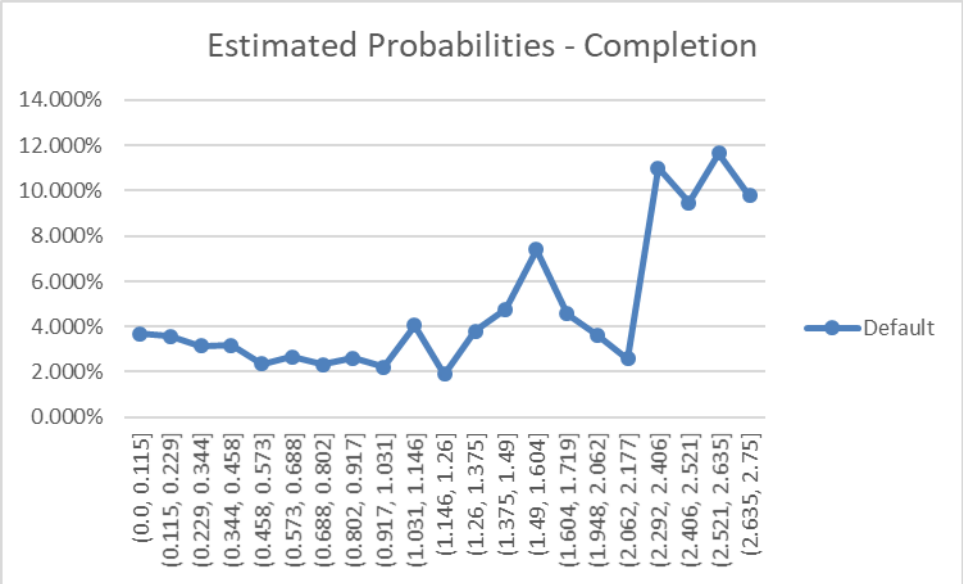

Figure 51

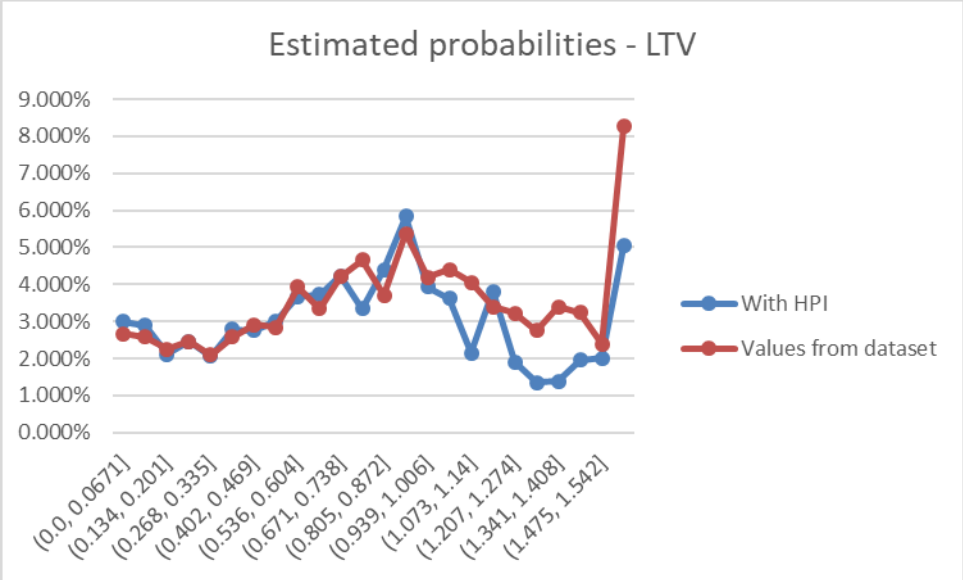

Figure 52

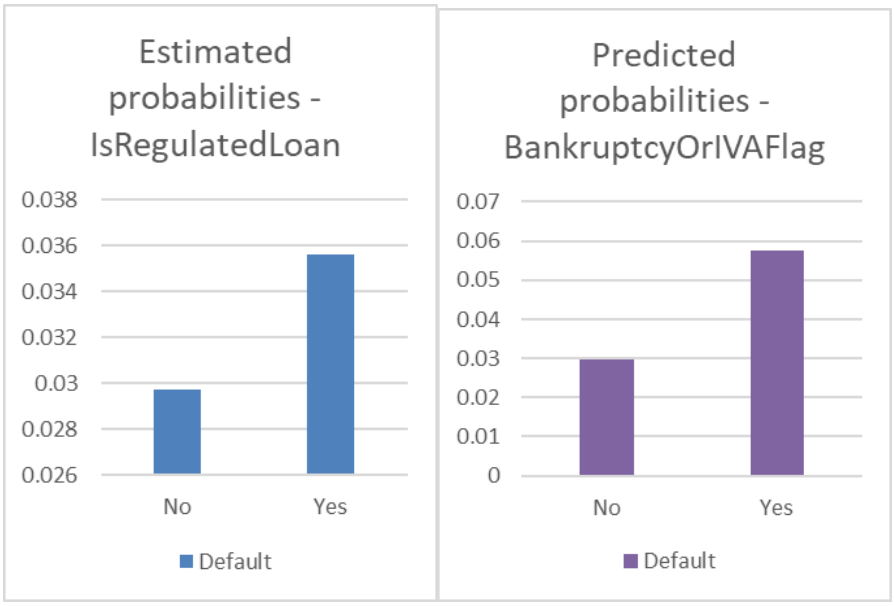

Figure 53

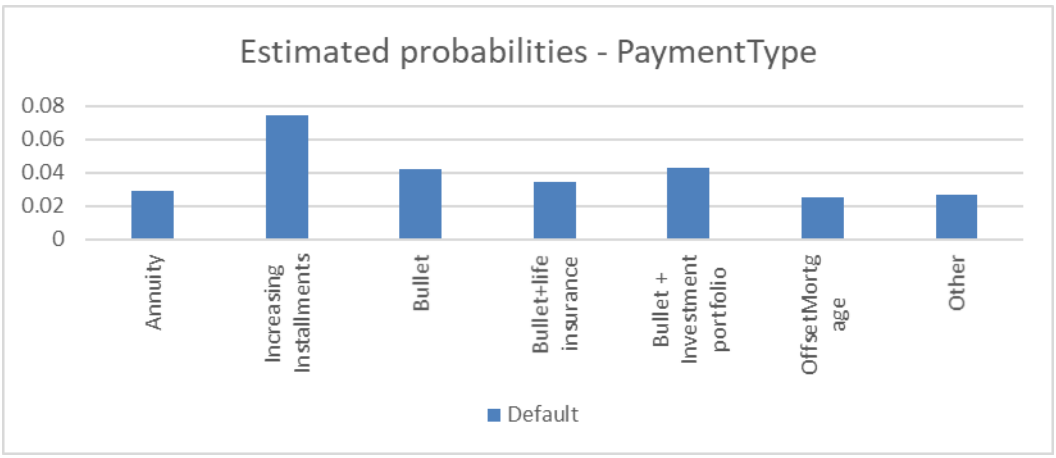

Figure 54

Appendix I – Alternative Methods

It is in our belief that a better predictive power could be achieved with a recurrent neural network[11]. The main difference between a feedforward neural network like the one we developed during this project and a recurrent neural network is the memory component. Recurrent neural networks save the output of a layer and feed this information back to the model as an input to assess the next observation, whereas a feedforward neural network simply classifies an observation and does not consider any information from this observation further once the final output is given (Brezak, Bacek, Majetic, Kasac, Novakovic; 2011) (Bengio, Mikolov, Pascanu; 2013).

The temporal aspect of our data and the information contained in previous observations makes the memory component of a recurrent neural network very interesting in our case of time-series forecasting (Brownlee; 2018). However, it adds some important complications in both the data handling part and the model training part (Bengio, Mikolov, Pascanu; 2013).

Looking now at alternatives to neural networks, an interesting method to predict mortgage loans behavior could be a decision tree learning model[12]. Such model will assess each attribute of the data set, dividing them into subsets depending on their variables' categories, until it reaches pure subsets allowing it to provide an interpretation of the data input. It will select the attribute it will use to build the most precise tree based on the attribute's information gain[13] (Kaur; 2017). There are three important advantages of this method that leads it to be an interesting alternative to consider in the next steps of this project. The first one being the confidence aspect of decision trees. As they keep count of the number of observations within the subset that led to the classification, the final output can be given with more or less certainty depending on how large this number is. Additionally, these models allow for visibility in the analysis that the neural network cannot provide, making it easier to assess the

[11]See https://www.youtube.com/watch?v=epS9UVRuoOE for useful explanation. Youtube channel: https://www.youtube.com/channel/UCs7alOMRnxhzfKAJ4JjZ7Wg (Last assessed in December 2018).
[12]See https://www.youtube.com/playlist?list=PLBLV0mgoy14rhd0DgjhRKp2TWn3rf-Lwn for useful explanation (Last assessed in December 2018).
[13] See https://www.python-course.eu/Decision_Trees.php for mathematical and programming basic overview (Last assessed in December 2018).

attribute with the most significant predictive power. On top of it, they require much less data curation that neural networks do, and can handle missing values and noise easily, as they'll focus on what will bring information gain (Gupta; 2017). However, decision tree learning models may have disappointing results out of sample since they tend to easily overfit, as they can only do axis-aligned splits of the data. Even though pruning can be done to alleviate this problem (Kaur; 2017) (Gupta; 2017). Additionally, it may be difficult to find the best tree as these models focus on the information gain step by step rather than x steps ahead or all the way to the end. Consequently, such model can deviate from what would be its ideal tree because at an intermediary step, a less powerful tree may have a greater information gain that the ideal tree. To counter these problems, a random decision forest method should be applied. A random forest consists of separating the data set into random sub-samples and build a decision tree for each sub-sample (Sekhar, Mina, Madhu; 2016) (Brownlee; 2016).

The final classification of the observation will be the average of the output of the independent decision trees (Bacham, Zhao; 2017) (Brownlee; 2016). As a result, by building trees on sub-samples different from each other, the method adds flexibility and avoid the over-fitting problem encountered with single decision trees. Such model should therefore be investigated further for our project of mortgage loan behavior's classification because of their ability to deal with missing value and noise and the visibility of their analysis – transmitting information on confidence of the classification and importance of data attributes.

Appendix J – Structured Finance Portal – ALL GROUP

The first part of our thesis consisted of an advisory project, in which we have analyzed the Moody's Analytics Structured Finance Portal and gave suggestions in order to improve it. This portal "is a premier web-based tool that offers data and analytics across all structured finance asset classes with advanced reporting and time-saving data normalization and aggregation. It provides structured finance professionals with cashflows, regulatory metrics, comparative analytics, and data aggregation in one integrated platform", (Moody's Analytics, Product List, Structured Finance Portal). Therefore, after studying the platform and all the features included in it, we were able to come out with the following improvement suggestions:

1.1 - Key Differentiators Explanation

Although the portal has already a glossary and its users are mostly financial experts, we think it would be convenient to include, in a straightforward way, a brief explanation of each indicator. One option could be, when hovering with the mouse on an indicator, an explanation would appear with a short description of the indicator followed by its formula, as you can see in *Figure 55.*

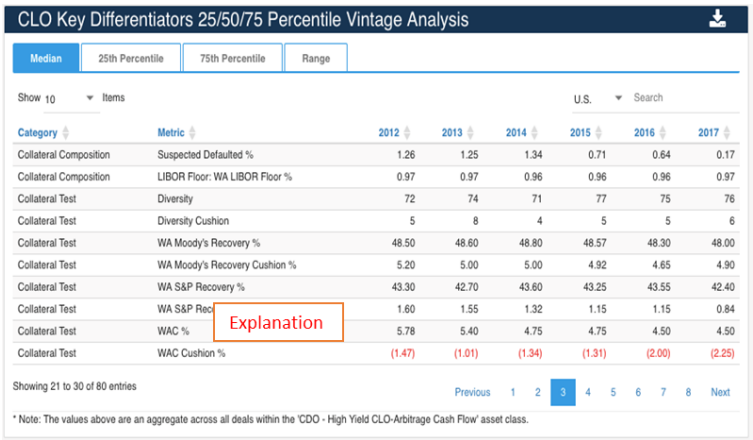

Figure 55

Another option could be to have a glossary available to download right next to the data download link – *Figure 56* . This option may be less quick than the first one, but it's much easier to implement and it is as useful. Either way, it would help clients not familiar with all aspects of structured finance to quickly grasp the meaning of the indicator, giving them a better view on their investment.

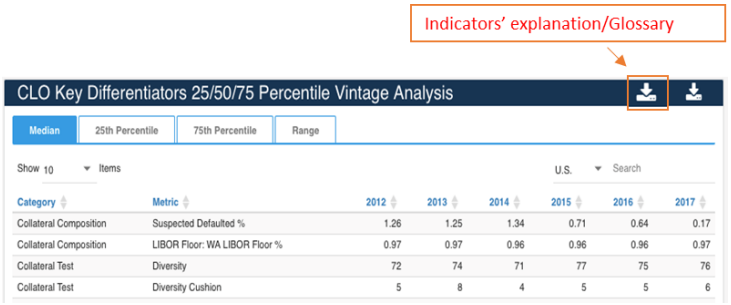

Figure 56

1.2 - Credit Migration Probabilities on the Tranche

Our second suggestion is related to the clients' perception of risk of their investments or the investments they are managing. This way, we consider convenient to display, in a straightforward way, the row of Moody's transition matrix corresponding to each tranche's rating.

In the portal, when a certain deal is 'open' we can observe several information including a C*haracteristics'* Board (Figure 58) in which is displayed, among other things, the *Moody's rating* for the specific tranche. In the deal we are using, we have got an 'Aaa grade' for the A tranche, what represents the highest degree of credit worthiness. From this issuance and/or current rating we know the original or current risk of investing in this tranche, however this is subject to change, until maturity.

Moody's Analytics' transition matrix gives some insights regarding this possibility of a credit rating being downgraded or upgraded, in fact, "it forecasts the probability of a credit migration, during the next year, for this each rating level", (Moody's Analytics, Credit Transition Model 2017 Update: Methodology and Performance Review). In this table (*Figure 57*), we have got the Moody's Analytics Historical Transition Matrix (1970-2017), which can be considered a good proxy for conditional transition matrix.

Initial	Rating at year end								
Rating	Aaa	Aa	A	Baa	Ba	B	Caa	Ca-C	Default
Aaa	91.37	7.59	0.85	0.17	0.02	0.00	0.00	0.00	0.00
Aa	1.29	90.84	6.85	0.73	0.19	0.04	0.00	0.00	0.07
A	0.09	3.10	90.23	5.62	0.74	0.11	0.02	0.01	0.08
Baa	0.05	0.34	4.94	87.79	5.54	0.84	0.17	0.02	0.32
Ba	0.01	0.09	0.54	6.62	82.76	7.80	0.63	0.06	1.49
B	0.01	0.06	0.20	0.73	7.10	81.24	5.64	0.57	4.45
Caa	0.00	0.03	0.04	0.24	1.04	9.59	71.50	3.97	13.58
Ca-C	0.00	0.00	0.14	0.00	0.55	3.76	8.41	64.19	22.96
Default	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00

Figure 57

Therefore, as we have said before, our suggestion would be to include this feature either directly displayed or through a button that would pop up the row of the transition matrix, *Figure 58*, presenting to the investor or portfolio manager the one-year probabilities of a credit migration. In the Silver Arrow example, we had an Aaa rated tranche, so it would display, directly or hidden, the first row of the transition matrix, showing the probabilities of an Aaa rated tranche to get downgraded to each one of the levels.

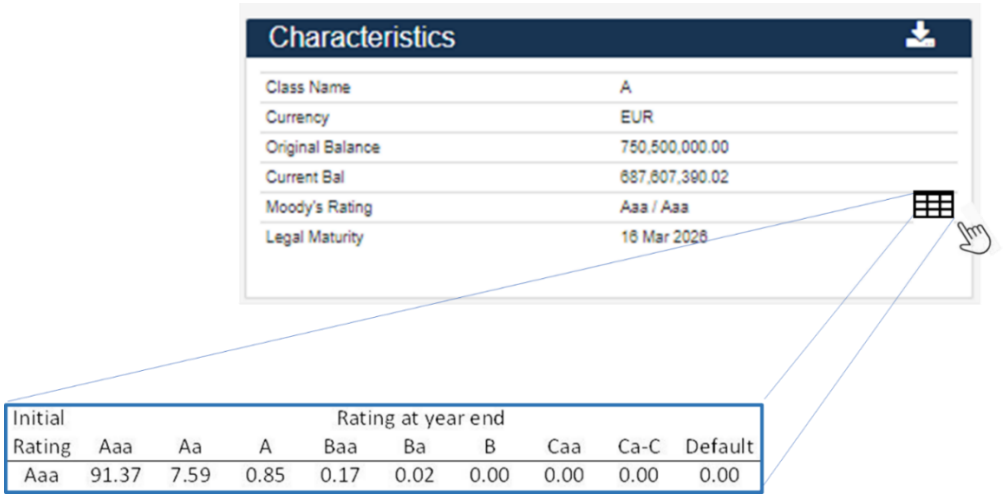

Figure 58

This feature would give us further information not only the default probability, but also on the probability of a credit rating migration, impacting directly the value of the asset. This way, the investors, CLO managers, and other counterparties that use the platform would have, almost immediately, a broader notion of the risk they are facing by investing in that particular security.

1.3 - Filters on Market Performances

The third suggestion to improve the portal would be adding filters on market performance, filtering, for example, by geography or deal manager (*Figure 59*)

By adding a filter for the asset managers, the client would be able to see how a certain asset class has performed with a specific manager. The client would also be able to compare the performance of a specific asset when under the management of different companies.

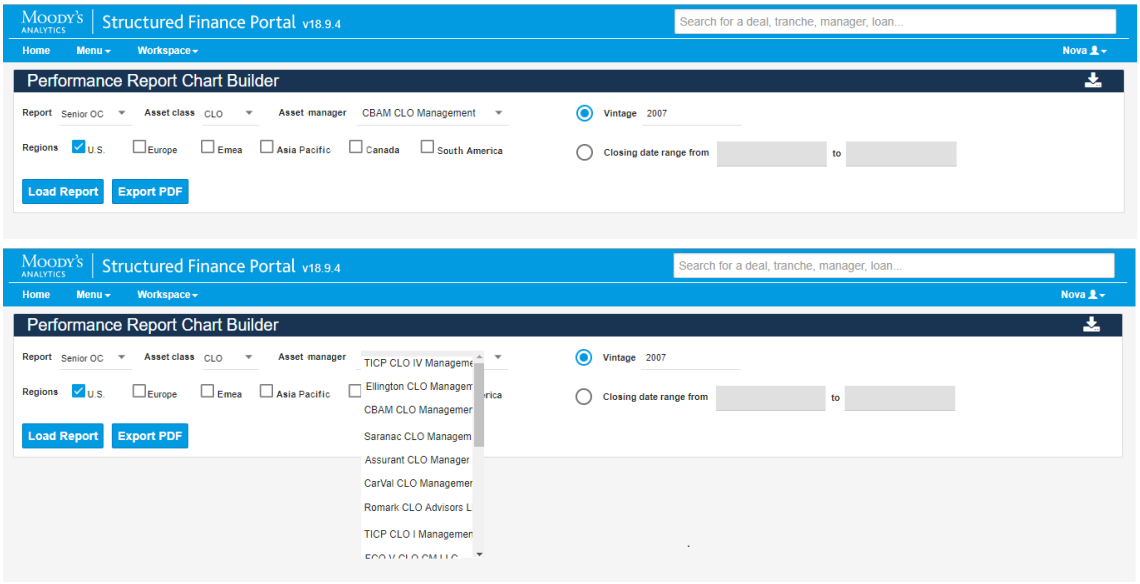

Figure 59

1.4 - Functions

Our fourth suggestion is related with some potential changes in the search engine, not with the purpose to replace the already good menu layout in the top left-hand corner of the screen, but to improve workflow and speed. In fact, we believe that a Bloomberg style shortcut function would be beneficial in improving the search engine. When opening the deal page, from the regular search menu, we would have another search bar (*Figure 60)* where we could use functions to directly give us the desired information. We could, for example, immediately write 'CF' to be directed to the Cash Flows page and after going to the same bar and write 'MA' for running the base case by Moody's.

Another example can be the performance indicators; we could type 'performance' into the touch bar and be directed to the usual screen, or write immediately the desired metric we wish to evaluate, for instance write 'WAM' to go directly do this metric (*Figure 61*), and perhaps focus more on it having besides the graph, the actual current value easily visible and perhaps the historical evolution (useful for instance in classes where prepayment risk is higher). The option to see other deals with asset pools with similar characteristics would also be interesting.

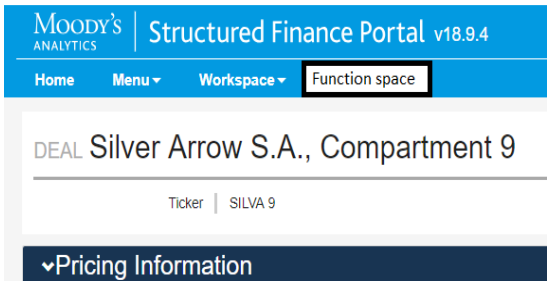

Figure 60

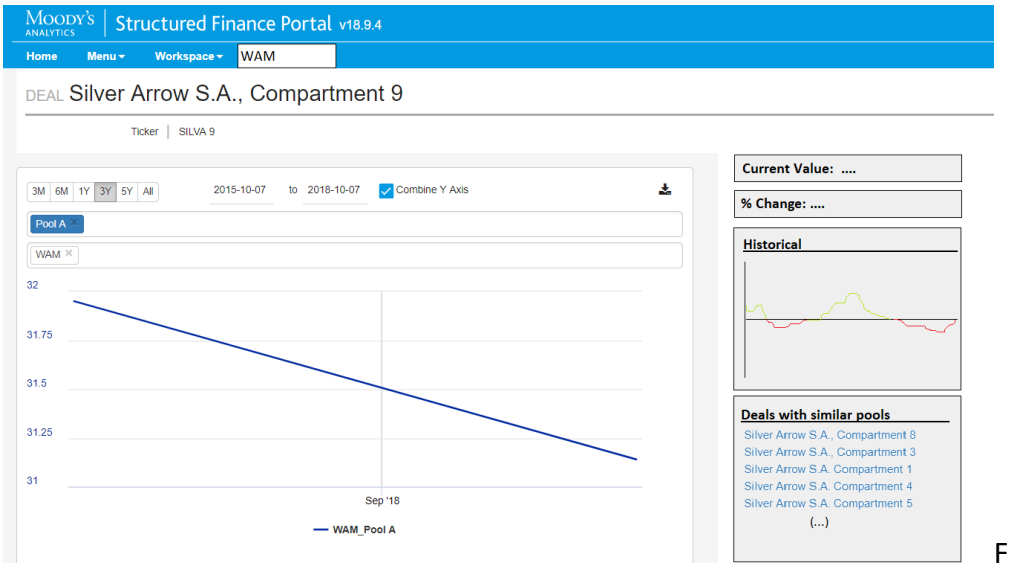

Figure 61

1.5 - Transition Matrix

This is not a brand-new suggestion, but perhaps something that can be improved. Looking to the portal transition matrixes, we noticed that there are a lot of bugs and inaccurate information. In fact, sometimes the matrixes state weird outcomes, such as 100% of a specific credit migration and no values appearing in some cells during crisis time spans. We believe it would be useful this framework to be corrected and developed a bit more, once it is an important risk assessment tool for the investors that use the portal.

References

The academic paper this study was based from:

Sirignano, J. A., Sadhwani, A. & Giesecke, K.; (2018) ; "Deep Learning for Mortgage Risk"; available at: https://arxiv.org/pdf/1607.02470.pdf (Last Accessed in December 2018).

All other references:

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. Md. J. & Liu, Y.; 2015; "Investigation and improvement of multi-layer perceptron neural networks for credit scoring"; *Expert Systems With Applications*.

Luo, C., Wu, D. & Wu, D.; 2016; "A deep learning approach for credit scoring using credit default swaps"; *Engineering Applications of Artificial Intelligence*.

Carton, R. B. & Hofer, C. W.; 2006; "Measuring organizational performance*"; Edward Elgar Publishing*.

Huang, X.; Liu, X. & Ren, Y.; 2018; "Enterprise credit risk evaluation based on neural network algorithm*"*; *Cognitive System Research*.

Barboza, F., Kimura, H. & Altman, E.; 2017; "Machine learning models and bankruptcy prediction"; *Expert Systems With Applications*.

Tavana, M., Abtahi, A. R., Di Caprio, D. & Poortarigh, M.; 2018; "An Artificial Neural Network and Bayesian Network model for liquidity risk assessment in banking"; *Neurocomputing*.

Angelini, E, Tollo, G. & Roli, A.; 2008; "A neural network approach for credit risk evaluation"; *The Quarterly Review of Economics and Finance*.

Chollet, F.; 2017; "Deep Learning with Python"; *Manning Publications*.

Shai, S.S. & Shai, B. D; 2014; "Understanding Machine Learning: From Theory to Algorithms"; *Cambridge University Press*; available at:

https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf (Last Accessed in December 2018).

Kashman, A.; 2011; "Credit risk evaluation using neural networks: Emotional versus conventional models"; *Applied Sof Computing*.

Kvamme, H., Sellereite, N., Aas, K. & Sjursen, S.; 2018; "Predicting mortgage default using convolutional neural networks"; *Expert Systems With Applications*.

Kashman, A.; 2010; "Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes"; *Expert Systems With Applications*.

Matz, L.; 2007; "Liquidity Risk Measurement and Management"; *John Wiley & Sons*.

Hu, J., Zhang, J., Zhang, C. & Wang, J.; 2016; "A new deep neural network based on a stack of single-hidden-layer feedforward neural networks with randomly fixed hidden neurons"; *Neurocomputing*.

Kelly, R., McCarthy, Y. & McQuinn, K.; 2012; "Impairment and negative equity in the Irish mortgage market"; *Journal of Housing Economics*.

Bian, X., Lin, Z. & Liu, Y.; 2018; "House price, loan-to-value ratio and credit risk"; *Journal of Banking & Finance*.

NerdWallet; 2018; "How Often Do Your Credit Scores Change?"; available at: https://www.nerdwallet.com/blog/finance/credit-scores-change/ (last accessed in December 2018).

Equifax; 2018; "Why do credit scores fluctuate?"; available at: https://www.equifax.com/personal/education/credit/score/why-do-credit-scores-fluctuate/ (last accessed in December 2018).

Experian; 2018; “How often does your credit score update?”; available at: **https://www.experian.com/blogs/ask-experian/how-often-does-your-credit-score-update/** (last accessed in December 2018).

TransUnion; 2018; “How Long Does it Take for a Credit Report to Update?”; available at: **https://www.transunion.com/blog/credit-advice/how-long-does-it-take-for-a-credit-report-to-update** (last accessed in December 2018).

OECD Data; 2018; “Household Savings”; available at:https://data.oecd.org/hha/household-savings.htm#indicator-chart (last accessed in December 2018).

BBC; 2017; “Household savings ratio falls to record low, says ONS”; available at: https://www.bbc.com/news/business-39453844 (last accessed in December 2018).

Cybenko, G., 1989. *Approximation by superpositions of a sigmoidal function,* s.l.: Mathematics of control, signals and systems.

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

Gau, G., 1978. *A taxonomic model for the risk-rating of residential mortgages,* s.l.: The Journal of Business.

Hornik, K., Stinchombre, M. & White, A., 1989. Multilayer feedforward networks are universal approximators. *Neural Networks,* 2(5), pp. 359-366.

Lemaitre, G., Nogueira, F., Oliveira, D. & Aridas, C., 2018. *imbalanced-learn.readthedocs.io.* [Online] Available at: https://imbalanced-learn.readthedocs.io/en/stable/index.html [Accessed 20 December 2018].

Mani, I. & Zang, I., 2003. *kNN approach to unbalanced data distributions: a case study involving information extraction.* s.l., s.n.

N.Vapnik, V., 1999. *The Nature of Statistical Learning Theory.* 2 ed. s.l.:Springer.

P.Kingma, D. & Ba, J., 2015. *Adam: A Method for Stochastic Optimization.* San Diego, s.n.

Platt, J., 1999. Probabilistic outputs for support vector machines to regularized likelihood methods. *Advances in large margin classifiers,* 10(3), pp. 31-74.

Ruder, S., 2016. An overview of gradient descent optimization algorithms.

Sutskever, I., Martens, J., Dahl, G. & Hinton, G., 2013. *On the importance of initialization and momentum in deep learning.* s.l., s.n.

von Furstenberg, G., 1969. *`Default risk on fha-insured home mortgages as a function of the terrms of financing: A quantitative analysis',* s.l.: Journal of Finance.

Weaver, C. E. a. W., 1964. *THE MATHEMATICAL THEORY OF COMMUNICATION.* s.l.:The University of Illinois Press.Urbana.

Brownlee, J.; 2018; "When to Use MLP, CNN, and RNN Neural Networks"; *Machine Learning Mastery*; available at:https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/ (Last accessed in December 2018).

Brownlee, J.; 2016; "Crash Course in Recurrent Neural Networks for Deep Learning"; *Machine Learning Mastery*; available at:https://machinelearningmastery.com/crash-course-recurrent-neural-networks-deep-learning/ (Last accessed in December 2018).

Brezak, D., Bacek, T., Majetic, D., Kasac, J. & Novakovic, B.; 2011; "A Comparison of Feed-forward and Recurrent Neural Networks in Time Series Forecasting"; *University of Zagreb*. Available at: https://bib.irb.hr/datoteka/575899.A_Comparison_of_Feed-forward_and_Recurrent_Neural_Networks_in_Time_Series_Forecasting.pdf (Last accessed in December 2018).

Bengio, Y., Mikolov, T. & Pascanu, R.; 2013; “On the difficulty of training recurrent neural networks”; *30th International Conference on Machine Learning, Atlanta, Georgia, USA*; available at: http://proceedings.mlr.press/v28/pascanu13.pdf (Last accessed in December 2018).

Kaur, R.; 2017; “An Introduction to Machine Learning With Decision Trees”; *AI Zone*; available at:https://dzone.com/articles/machine-learning-with-decision-trees (Last accessed in December 2018).

Gupta, P.; 2017; “Decision Trees in Machine Learning”; *Towards Data Science*; available at:https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 (Last accessed in December 2018).

Brownlee, J.; 2016; “Bagging and Random Forest Ensemble Algorithms for Machine Learning”; *Machine Learning Mastery*; available at: https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/ (Last accessed in December 2018).

Bacham, D. & Zhao, J.; 2017; “Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling”; *Moody’s Analytics*; available at: https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling?fbclid=IwAR3rX7pxZjQATnkJ2QHQ9BkL_PnQYOgn7Mj7wi27YSCdGp3Bu-rlWvY8RXo (Last accessed in December 2018).

Sekhar, R. & Minal, M. E.; 2016; “Mode Choice Analysis Using Random Forrest Decision Trees”; *Transportation Research Procedia*.

Chatterjee, S.; 2015; “Modelling credit risk”; Centre for Central Banking Studies; Bank of England

“Credit Risk Modelling: Current Practices and Applications”; 1999; *Basel Committee on Banking Supervision*

Roger, M. S., Ashish, D.; 2010; Residential Mortgage Portfolio Risk Analytics; *Moody's Research Lab*

Zhang, Xuan; 2017; Essays in credit risk management PhD thesis - University of Glasgow

Teaching notes, Credit risk, Nova SBE, João Pedro Pereira

Principles for the Management of Credit Risk, BIS

https://www.moodysanalytics.com/product-list/structured-finance-portal

https://www.moodys.com/sites/products/ProductAttachments/DRD/CTM_Methodology.pdf

Silva-Palacios, D., Ferri, C & Ramírez-Quintana, M.; 2017; "Improving Performance of Multiclass Classification by Inducing Class Hierarchies"; *International Conference on Computiational Science.*

Ritchie, N. G.; 2018; "Evaluating a Classification Model"; available at: https://www.ritchieng.com/machine-learning-evaluate-classification-model/ (last accessed in December 2018).

Notesbyanerd; 2014; "Multi-class Performance Measures"; available at: http://notesbyanerd.com/2014/12/17/multi-class-performance-measures/ (last accessed in December 2018).