Deep Learning for disentangling Liquidity-constrained and Strategic Default

Arka Prava Bandyopadhyay and Yildiray Yildirim *

Preliminary Draft (do not quote)

Abstract

We implement a Deep Neural Network (DNN) methodology for *disentangling* liquidity-constrained and strategic default using a proprietary Trepp data set of commercial mortgages and motivate a model-agnostic interpretation of variable importance that is robust (insensitive) to severe Financial Crisis (2008). We use 29 loan-specific, property-specific, tranche-specific and deal-specific co-variates, 4 macro-economic variables, 2 indices for approximately 200,000 loans over a period of 19 years from 1998 to 2016 on a monthly basis. The non-linear activation function in the DNN captures the highly non-linear interaction among co-variates, geographical cross-correlation and macro-economic factors at state level beyond linear unobserved time-fixed and loan-fixed effect. We are able to capture *strategic delinquency* from the Variable Importance table and Shapley values for Deep Learning, e.g., Net Operating Income, Appraisal Reduction Amount, Prepayment Penalty Clause, Balloon Payment, etc. co-contribute and interact in a highly non-linear way to impact the endogenous choice of delinquency class compared to other more statistically significant variables such as Loan-to-Value. Further, we show that our results on variable importance are robust to the Financial Crisis of 2008. There is a significant increase in accuracy of predictions for the classes beyond 90 days of delinquency when the Deep Learning Model is compared with Naive Bayes, Multinomial & Ordered Logistic Regression, Support Vector Machine, Distributed Random Forest, Gradient Boosting Machine, Deep Neural Network by gradually relaxing the specification structure, thereby increasing flexibility. These findings have significant implications for CMBS investors, Rating agencies, and Commercial Property Finance policymakers.

Key words: Strategic default, CMBS, Machine learning

*Bandyopadhyay, ArkaPrava.Bandyopadhyay@baruch.cuny.edu, PhD Student, William Newman Department of Real Estate, Baruch College, New York, NY 10010 USA; Yildirim, Yildiray.Yildirim@baruch.cuny.edu, William Newman Department of Real Estate, Baruch College, New York, NY 10010 USA. We thank Linda Allen, Liuren Wu and the seminar participants at Baruch College.

1 Introduction

The residential real estate bubble from 2004 (emergence) to 2008 (burst) has attracted lot of scholarly work and policy attention. Surprisingly, the commercial real estate price impact (potential bubble) has been ignored, in comparison. Since the inception of securitization as a means of financing commercial real estate (CRE) mortgages from 1998, the sophisticated B-piece investors have been outbid beyond sustainable long-run fundamentals, over time (from 2004) by investors who "originated to securitize", thereby, resulting in decline in underwriting standards in CRE (Levitin and Wachter (2012)).

The current outstanding balance of CMBS is almost a trillion dollars, but they are different from RMBS with respect to delinquency behavior in several aspects, e.g., lack of standardization of appraisals, underwriting criteria, legal documentation requirements, the mortgage instrument, the security instrument and even credit risk rating standards. These borrowers are not households, but savvy businessmen and hence their delinquency behavior is possibly much more strategic/ P&L - oriented based on mortgage contractual features (prepayment penalty clause, balloon payment indicator), macroeconomic conditions, supply and demand in the local geography and financial constraints, such as **Net Operating Income** (NOI), emanating from the unbalance in terms of the amount and time lag between cost of funding and income cash flows.

Despite some overlap in mutli-familiy property type, Commercial and Residential Real Estate (RRE) are markedly different markets and hence have attracted dissimilar government (e.g., GSE) intervention. Non-recourse Residential and Commercial Real Estate has an implicit put-option structure equivalent to repurchase of the loan with the value of the property as the strike, wherein the borrower can satisfy the debt obligation by surrendering the property to the lender. This is the primary reason why almost all of previous literature have found **Loan-to-Value** (LTV) as the primary driver of default behavior (e.g. Ambrose and Jr. (2012), Ambrose, Capone, and Deng (2001)). Since, RRE is both investible and consumable, tax-deduction acts an incentive and the foreclosure and recourse laws act as disincentives for strategic default for households. Although the individual Commercial Real Estate loans are much bigger in size compared to their Residential counterpart, partially amortizing structure, defeasance, yield maintainance clauses discourage refinancing and hence is exposed to **strategic** default, where borrowers choose to stay in 90-120 days delinquency bucket for a while.

Commercial mortgages are used to finance income-producing properties. Therefore, a borrower's default decision depends on not only the asset value (i.e., borrower equity) but also the property liquidity (i.e., property income). A rational borrower would not default when property net cash flow is positive and is enough to service the scheduled debt obligation even if the owner's equity position is negative. To properly reflect a rational borrower's default decision, a model for commercial mortgages needs to include both property value and property income as default triggers. Also, unlike residential mortgages that are typically fully amortizing, most commercial mortgages are partially amortizing; that is, a balloon payment is due when the mortgage matures. Typical commercial mortgages have a 7-12 year term and a 25-30 year amortization schedule. Borrowers usually fund the balloon payment by refinancing the current mortgage, which may be complicated at maturity due to higher interest rates or tighter under-

writing standards even for a borrower in good standing.

The complexity of CMBS modeling is due to the simultaneous inclusion of four significant risks: market, credit, prepayment (Christopoulos, Jarrow, and Yildirim (2008)) and liquidity (Ambrose and Sanders (2003)). The cash flows to the underlying CMBS loan pools, the cash flow allocation rules to the various bond tranches, the prepayment restrictions and the prepayment penalties differ across the different CMBS trusts. The estimation of the relevant parameters is itself a nontrivial problem, given the sparsity and the diversity of historical CMBS data. The empirical mortgage literature identified a number of variables to predict commercial mortgage credit and prepayment risk including creditworthiness and free cash flow of the entity, current leverage ratio, loan age, interest rates, and CMBS indices (e.g. von Furstenberg and George (1969), Curley and Guttentag (1974), Campbell and Dietrich (1983)). The commercial mortgage performance is typically specified in a linear form in terms of these factors. The commercial mortgage performance data, however, tell a different story. The presence of nonlinear effects in Figure 5 obviates the need for a more general form but it is difficult to identify all the factors and their mutual interactions. Instead of specifying a functional form for commercial mortgage performance, we include all possible factors and let the data dictate the model, which also allows for highly non-linear interaction terms between factors. Since, our data set is nationally representative, the pooled model computes an estimate of aggregate default risk in the commercial mortgages especially well for 2007-2009. Our estimation result provides a ranking of individual commercial mortgages in terms of their delinquency behavior and can be aggregated to a systemic measure of default risk in the commercial sector.

The sophisticated delinquency behavior of commercial mortgage borrowers is highly endogenous and hence cannot be captured by the standard loan-specific and macroeconomic variables in a linear specification. Rather than assume a single default trigger based on property value (measured by contemporaneous loan-to-value, LTV), our model incorporates a second trigger based on contemporaneous property income (NOI)[1]. We also explicitly consider balloon risk as a second source of credit risk in commercial mortgages. Our findings reveal that the effect of a property income along with prepayment penalty clause and balloon risk is significant to assess total credit risk adequately.

To test the robustness and stability of our DNN, we present the Variable Importance Plots of Predicted Default Rate from June 2006 to December 2008 with several features in Distributed Random Forest (DRF) in Figure 8a, Gradient Boosting Machine (GBM) in Figure 8b and DNN in Figure 8c, trained on data before June 2006 and motivate why we need a highly non-linear model and also why we allow for high-dimensional interaction among the borrower-specific, macroeconomic, spatial, vintage effects in the features. Time-to-Maturity, Geographical cross-correlation, NOI, Appraisal Reduction, Bankruptcy Flag, Property Type, Non-Recoverability, Appraised Value supercede Securitized LTV in the Variable Importance chart for DNN in Figure 8c. Morover, Balloon Payment supercedes LTV, corroborating the robustness of our DNN model. DRF captures non-lineary of the covariates but still ranks LTV much above NOI and

[1]This is in line with Foote, Gerardi, Goette, and Willen (2009) studying that when equity is negative but above a hreshold, default occurs with negative income shock. Our findings are consistent in the case of Commercial Real Estate without imposing any model specification.

other strategic variables in Figure 8a, even after tuning and bagging. GBM is a greedy algorithm and hence finds more occurances of local minima for LTV and hence ranks LTV higher than NOI in Figure 8b, even after boosting.[2]

We list the possible combinations of LTV and NOI that can *disentangle* Liquidity-constrained and Strategic Default behavior in Figure 9. We use DNN and show that case (1) is liquidity-constrained default in Figure 4b since there is no spread in terms of predictability of NOI. The effect is verified by ensuring high LTV values in the Bivariate Heatmap in Figure 5b. In fact, DNN algorithm can identify the threshold of NOI^* in Figure 4b Case (2) is more interesting since there is some spread in default predictability w.r.t NOI. We claim from Figure 4b, this is where strategic defaulters cannot be identified from non-strategic defaulters after $NOI > NOI^*$. But, most likely a strategic defaulter would not default in Case (1) to have the option to default in Case (2). This gives us a mechanism to identify the strategic defaulters from non-strategic ones once the threshold is identified. Cases (3), (4), described in Figure 5b, behave in a similar way since LTV is still very high. In fact, DNN shelps us identify LTV^*, NOI^{**} in Figure 5b. Financial friction acts as a liquidity-constraint for non-strategic defaulters. Strategic Defaulters are also in this cohort and NOI^{**} determines (in Figure 5b) the cutoff beyond which again the behavior of Strategic Defaulters from the Non-Strategic defaulters. The heterogeneity occurs because of constraint on time or limited attention for Non-Strategic Defaulters. Cases (5), (6) (in Figure 4a) are much more interesting and there is a whole host of factors that we conider in DNN to identify the strategic defaulters and their incentives. Again, DNN indicates a *possible* LTV^{***} & NOI^{***} in the lower portion of Figure 5b.

One possible determinant of strategic default is the moral hazard, that is less time-varying, but country-specific. Both consumer bankruptcy and commercial foreclosure (not necessarily arising from corporate bankrutcy) Laws are lenient in USA. On top of this, since, the probability of default is priced in at origination of commercial loans, delinquency can be viewed by corporate borrowers as insurance. Although, moral hazard is time-invariant, but the incentive of a borrower for moral hazard needs to be triggered. We use several key covariates, e.g., Net Operating Income, Appraisal Reduction Amount, Prepayment Penalty Clause, Balloon Payment at Maturity, Non-Recoverability, etc. to identify when moral hazard is triggered vis-a-vis higher order non-linear interactions during severe stress in the Financial Crisis in Figure 5.

[2]We take a deeper dive and investigate LTV in the following way:

$$LTV_t = \frac{AOB_{t-1} + DS_t + B_T}{MV_t - AR_t} \tag{1}$$

where AOB_t is the Outstanding Balance at time t-1 that is amortized , DS_t is the scheduled payment due for servicing the debt obligation at time t, B_T is the Balloon Payment due at maturity, MV_t is the Market value of the property/properties at time t (which varies significantly with respect to macroeconomic conditions and spatial/location context) for which the mortgage has been issued, AR_t is the Appraisal Reduction at time t.

AOB remains consistent, since, prepaymant penalty clauses discourage voluntary curtailment/full prepayment. SP obligations are not met both when the borrower is cash-constrained and also when the borrower chooses to strategically default. Proximity to balloon payment at maturity further complicates the endogenous behavior of the commercial borrowers towards maturity of the loan. The market value of a property is a function of the macro-economic factors like state GDP, Unemployment Rate, geographical location, 2 Year and 10 Year Treasury Rates. Until valuation is obtained, Appraisal Reduction Amount (ARA) may be calculated based on the scheduled principal balance or some other formula as defined in the servicing agreement.

Strategic defaults are de facto unobservable events. Although defaults are observed, one cannot observe whether a default is strategic as strategic defaulters disguise themselves among borrowers who cannot afford to pay. Bajari, Chu, and Park (2008) assess the likelihood of strategic default by estimating a structural model of default that includes both cash flow considerations and negative equity considerations. Guiso, Sapienza, and Zingales (2013) evaluate the likelihood of strategic default by resorting to a quarterly survey of a representative sample of U.S. households. By asking about a person's willingness to default at different levels of negative equity in a survey, one can measure the effect of the shortfall in equity while keeping all the other individual characteristics constant, including the level of wealth. Survey data provides an opportunity to separate contagion effects from sorting effects, which is difficult to do with field data. By asking questions about social and moral attitudes toward default, one can identify whether the high propensity to default in areas where foreclosures are more frequent is due to a clustering in those areas of individuals prone to default or to a contagion effect. Also, survey data allows asking about other attitudes and perceptions of the respondents that are not otherwise observable and that can be used to disentangle where certain effects, such as the correlation between knowing somebody who defaulted strategically and willingness to default strategically.

We motivate the highly strategic delinquency behavior of the savvy commercial borrowers/ business-owners from two different angles. We provide evidence from the Trepp data in Figure 1a that from 2012, the number of loans have remained flat but the outstanding balance of loans have steadily increased until 2016. This could have serious implications. There can only be two possibilities: if the same loans stay and there is no origination at all, and further if the outstanding balance is increasing, it means there is serious delinquency in the loans and the servicers are unable to secure the payment from the borrower and all these loans could potentially become limbo loans.

Figure 1b furthers the narrative. From mid-2014, the age of the loans is decreasing and the time-to-maturity is increasing. This could mean that from mid-2014, there are an equal number of originations to the number of maturing loans. But the fact that the Outstanding Balance is increasing in this entire period could only mean that the same loans are getting rolled over to new contracts, when balloon payments are missed during maturity.

Figure 1c clearly shows that LTV (widely used in previous literature and used by most banks/asset managers for credit risk calculations) is flat throughout the data horizon. The interest rate is decreasing almost monotonically in the data and there seems to be no sensitivity of LTV to interest rate. This means LTV is probably not the right way to think about credit risk. It could also be that the commercial borrowers **target** LTV. They strategically make payments towards their obligation so that the ratio of "Book Value of Loan" and the "Value of the Property" remains relatively stable over time. It would make sense for them to do this as banks/asset managers use LTV at origination as the primary determinant of creditworthiness of the borrowers. Further, the Contemporaneous LTV (CLTV) is used to calculate LGD (Loss given Default or 1-Recovery Rate). So, CLTV could also be targeted and there is no evidence of voluntary deleveraging from the borrower inspite of widely changing macro-economic conditions, e.g., interest rate.

Figure 1d corroborates that the NOI monotonically increases in the data and the occupancy is almost 100% in the entire data. So, there may be strategic saving of internal cash flow from income producing properties. Because of the strictly increasing NOI level, the strategic dominance of NOI over other factors can have disastrous aggregate macroeconomic consequences. To capture this, we try different methodologies like vanilla models (Naive Bayes, Multinomial and Ordered Logistic) and machine learning models (Distributed Random Forest, Gradient Boosting) and finally Deep Learning and find that Deep Neural Network (DNN) is best positioned to address the above issue and does capture NOI as the most significant strategic variable from the Variable Importance (VI) tables of the models. This difference does not stem from sample bias. This is the core reason for our choice of big data for training all the models. Also, Trepp is the largest provider of CMBS data and hence the sample is representative of the entire market and does not have any selection bias.

Our Deep Neural Network (DNN) methodology extends the scope of "Frailty Model" [3]. DNN not only captures latent time-fixed macroeconomic effect but also loan specific idiosyncratic effects beyond what has been captured in prior literature in Commercial Mortgages. We include 31 variables from Trepp in our DNN model along with state-level macro variables like unempoyment, GDP growth and 2Yr & 10 Yr treasury rates and recently created indices. These 35 variables capture the loan-specific unobserved effects and the macroeconomic variables proxy for unobserved common latent variables. Morover, using the deep learning methodology, we can capture the highly non-linear interaction among the covariates. The unbiasedness of the factor loadings is not the objective here, the accuracy of prediction is. We conduct a horse racing among all the models we test and we determine the best metric/visualization scheme to do that.

Furthermore, our deep learning model also captures non-linear effect of the covariates and all possible variable interactions of any order existing in the data, which cannot be captured by the extensions of models with linear specification, thereby eliminating the bias due to linear model mis-specification in Figure **??**.

Our findings yield important new insights into the interplay of borrower behavior, various risk triggers and the macroeconomy. They significantly differ from the findings of Campbell and Dietrich (1983), Cunningham and Capone (1990), Deng (1999), Elul, Souleles, Chomsisengphet, Glennon, and Hunt (2010), Foote et al. (2009) and others. These prior studies highlight loanlevel variables such as loan-to-value ratio, loan age, etc. as major predictors of borrower behavior. The significance of loan-to-loan correlation due to the common exposure of borrowers to geography, vintage and property types emphasizes the need for CMBS investors as well as commercial mortgage lenders to diversify mortgage risk, beyond the conventional borrower characteristics highlighted in the literature.

We test whether by adding macroeconomic variables, we can delve into the realm of omitted variable bias found in all shedonic models. We extend the literature on hedonic models by

[3]Duffie (2009) created an MCMC methodology that updates the posterior distribution of unobserved risk factors based on Bayes' rule whenever defaults cluster at a given point in time. In the event forecasting literature, such a dynamic unobserved covariate's effect is termed "frailty". Yildirim (2008) also propose a mixture model to disentangle the probability of long term survivorship and the timing of the default event.

systematically different macro variables which are exogenous in the shedonic regression model beyond the characteristics (used as covariates) and can explain a lot of the unobserved effects Childs, Ott, and Riddiough (1996). Other than 2Yr Treasury Rate and State Unemployment Rate, the macroeconomic variables do not directly affect the strategic delinquency behavior and timing.

National interest rates, e.g., 2 Year and 10 Year Treasury rates impact occupancy of commercial properties directly, as well as through state-level GDP. The unemployment is also captured at the state level. The local State level Unemployment Rate in urban centers and occupancy of lessees in commercial properties are highly correlated, because there lessees are the job-creators locally. We claim that all of the above can be captured by NOI, since both occupancy and unemployment rates affect the NOI from the property.

We normalize Net Operating Income (NOI) as a percentage in the pooled data for loans and create histograms of relative frquency of the number of loans in different delinquency classes with respect to the different percentiles of NOI. We see a sheavy support for the relative frequency across all the delinquency classes at the NOI percentages 5%-7%. We call them **dominant** NOI buckets. We show the distribution of different delinquency classes with all the NOI buckets including the dominant ones (see figure 2). The significant heterogeneity across the delinquency classes and the highly non-linear effect of NOI towards the strategic choice of the borrower to be in a specific delinquency class is not bourne out of this diagram.

Beyond the above dominant buckets, we see highly non-linear **strategic** behavior for commercial mortgage borrowers to choose different delinquency classes for different buckets of NOI. To visualize this, we zoom in and remove the dominant buckets and form the **rescaled** (without dominant NOI bucket masses) relative frequency histogram across all delinquency classes. Figure 3 highlights the complex relationship that exists between the percentage of loans across the different delinquency classes "Within 30 Days"(**W0_30D**), "30 Days to 60 Days"(**W30_60D**), "60 Days to 90 Days"(**W60_90D**), "90 Days to 120 Days"(**W90_120D**), "Beyond 120 Days" (B120D) and the buckets of net operating income (NOI) excluding the dominant NOI buckets, which can be incentivized by the macro-economy. The sensitivity varies significantly in a highly non-linear way in both magnitude and sign.

There is a **U-shaped** choice between NOI buckets 37%-45%for the borrowers in the delinquency class W90_120D. This means that when a borrower is already beyond the default threshold of 90 days, but less than the cutoff of 120 days, they are incentivized to stay there for a while and time their future payments based on cash flow. Since these NOI bucktes are higher, the borrowers make some profit from the income generated from the property, but they still stay at the same delinquency classes and do not pay-off the earlier missed payments to come back to the Current State (less than 30 days of delinquency). Similarly, the borrowers in delinquency class B120D choose to be in lower NOI buckets in a non-linear way. This is because of the lack of net cash flow income for them to be able to pay off the earlier missed payments. They end up in a vicious cycle of making less money from the property and becoming worse off in terms of their creditworthiness. We call them "limbo" loans as these loans stay in this state for a while before they are resolved. The sensitivity estimates generated by vanilla models can misrepresent the influence of risk factors because of naive choice of linear specification. This can

make it difficult to make economic conclusions from the borrower behavior. In our approach, the relationship is entirely dictated by data, thereby minimizing model misspecification and bias of the variable estimates.

Our data indicates that transitions between the Current loans to at most 90 days of missed/late payment are in fact frequent, e.g., a meaningful number of loans enter foreclosure but eventually return to current. Similarly, many loans are consistently behind payment but do not ever enter foreclosure. We do not find any model that can correctly predict which bucket of delinquency classes a loan belongs within 90 days.

We also add more broadly to the literature on neural networks. Several authors have used shallow neural networks in other areas of financial economics. Bansal and Viswanathan (1993) approximate the pricing kernel using a neural network. Hutchinson, Lo, and Poggio (1994) pioneered the use of neural networks for nonparametric option pricing. Brown, Goetzmann, and Kumar (1998) use neural networks to predict stock markets. Swanson and White (1997) propose the use of neural networks for macroeconomic forecasting. Lee, White, and Granger (1993) construct tests for neglected nonlinearities in time series models using neural networks. Granger (1995) and Kuan and White (1994) study nonlinear or neural network modeling of financial time series. Khandani, Kim, and Lo (2010b) and Butaru, Chen, Clark, Das, Lo, and Siddique (2016) examine other machine learning models of financial default. Recent applications of deep learning in financial economics include Klabjan (2007) who model market movements. Heaton, Polson, and Witte (2017) use deep learning for portfolio selection.

The remainder of the paper proceeds as follows. We provide details on the big data we use in Section 3 and provide descriptive statistics and conduct some exploratory analysis to give an idea of the trends in data. In Section 4, we motivate all Naive Bayes, Multinomial Logit, Distributed Random Forest, Gradient Boosting Machine models and provide empirical results and list the deficiencies in each of them. In Section A, we describe the Deep Learning (DNN) Model and motivate how DNN can alleviate most of the issues in the earlier models. We also point out the key findings of the paper in this section and how they differ from the literature. In Section 7, we provide concluding remarks followed by references.

2 A Simple Theoretical Motivation

We provide a simple model framework to motivate that the "Optimists" (or Strategic Defaulters) would prefer to maintain a consistently higher LTV during the good portion of business and economic cycle. They will then have the option to strategically default in the future. Whereas, "Pessimists" (Non-Strategic Defaulters) would prefer to continually reduce LTV, in anticipation of different forces increasing LTV in the future and also to alleviate the consequences of default in the event they are liquidity-constrained. This differential behavior across the cohort of borrowers will price in their heterogenous beliefs (π) in the expectation of occupancy of the property.

In residential market, while negative equity (whenever the value of the mortgage exceeds

the value of the property), in nonrecourse states, is a necessary condition for strategic default Guiso, Sapienza, and Zingales (2009), it is not sufficient. Even in nonrecourse states, there are frictions that make defaulting less appealing. Consider a borrower who at time t owns a house worth A_t and faces a mortgage balloon payment equal to B_T. From a purely financial point of view the borrower will not default as long as $A_t > B_T$. In the decision whether to default strategically, however, there are considerations other than the financial gain or loss from defaulting. For example, by not defaulting, a borrower enjoys the benefit of defaulting in dire conditions in the future. The intertemporal substitution of default choice is co-determined by timing of Appraisal Reduction, Non-recoverability as to whether the Master Servicer/Special Servicer has ceased advancing (P&I and/or Servicing) for the related mortgage loan, etc. Also, by defaulting she faces higher cost of borrowing in the future due to differential credit-rationing by the lender, since lenders are generally NPV-neutral and default is a deadweight loss for them. Let us define K_t as the net benefit (opportunity cost of cash) of not defaulting at t. Then a rational borrower will not default if $A_t - B_T + K_t > 0$.

If the commercial borrower does not have a balloon payment due, then her decision of whether to default strategically is more complex, because she must trade off the decision to default today with postponing the decision and possibly defaulting tomorrow. In addition, the option to default tomorrow is conditional on the ability of the borrower to serve her mortgage debt, which is highly correlated with the probability of occupancy and positive cash flow from the lessee in the property. If the property is vacant or if the lessee does not pay up, the borrower is likely to default next period and thus loses the value of the option. Let $V_T = A_T - B_T + K_T$, where T is the day the balloon payment is due. Then the value Bajari et al. (2008) of not defaulting at T-1 is:

$$V_{T-1} = A_{T-1} - m_{T-1} - B_T + K_{T-1} + (1 - \pi_{T-1})E_{max}(V_T, 0) \tag{2}$$

where A is the monetary value of the cashflow and the serviceflow enjoyed between time T - 1 and T, m is the mortgage payment (scheduled and unscheduled) between T-1 and T, π_{T-1} is the probability of vacancy of the property (i.e., not having a lessee, and E is the expectation operator. The value of not defaulting at a generic date t is then:

$$V_t = A_t - m_t - B_T + K_t + (1 - \pi_t)E_{max}(V_{t+1}, 0) \tag{3}$$

From (1), the decision to default strategically at a generic time t can be described by the following relationship

$$StrategicDefault = F(A - B, A, m, \pi, K) \tag{4}$$

Analogous to Residential Real Estate borrowers, when the owner uses the property for her own business instead of leasing it, the property serves the purpose of "Consumption" instead of "Investment". So, the size of the property along with the owner's firm's productivity (equivalently, cashflow from lessee) is no longer scalable in terms of default decision and has an upper cap. This is further clouded by the fact that Recourse Laws are not strictly implemented in most states. Bankruptcy Laws need to be fairly strong in a state to reinforce recourse laws.

3 Data

We have monthly proprietary novel data set from January 1998 to September 2016 from Trepp, the leading provider of analytics, information, and technology to the global CMBS, commercial mortgage finance, and banking industries. Trepp is the largest commercially available database containing detailed information on over 1,800 deals and more than 100,000 loans, which support close to $800 billion in securities. Deal coverage includes North American, European, and Asian CMBS, as well as Commercial Real Estate backed CDOs.

For the initial review, we use data from 2007 to 2009 (36 months) for 91767 loans for model fitting purposes, since going beyond this size requires access to GPU/Clouds, which we have adapted now. Our final cleaned data has around 8.4 million observations of 35 covariates.

We include the variables used in previous CMBS literature, like An, Deng, and Gabriel (2009), Ambrose and Sanders (2003) and preclude the following key loan-specific variables: log(original balance), LTV, time of amortization, time to maturity, lockout, lockout expiration, corporate bond credit spread Titman, Tompaidis, and Tsyplakov (2005), yield curve, mortgage-treasury rate spread, region dummy, seasonal/quarter dummy, among others.

We finally decide to use loan-to-value (ltv), occupancy rate (occ), tranche loan-to-value, (securltv), tranche weighted average cost (securwac), annualized gross rate (actrate), outstanding scheduled principal balance at end of current period (obal), derived most recent net operating income (noi), outstanding legal remaining outstanding principal balance reflecting defeasance of the loan as of the determination date (balact), securitization balance of the loan predged to the trust (face), most recent appraised value else securitization appraised value (appvalue), total amount of principal and interest due (actpmt), regularly scheduled principal to be paid to the trust (curschedprin), principal prepayments and prepayments (full or partial), discounted payoffs, and/or other proceeds resulting from liquidation, condemnation, insurance settlements (curunschedprin), interest basis of an adjustable rate loan (pmtbas), net proceeds received on liquidation of loan (liqproceeds), expenses associated with the liquidation (liqexpense), difference between Net Proceeds (after Liquidation Expenses) and Current Beginning Scheduled Balance (realizedloss), amount received from a borrower as a pay off a loan prior to the maturity or anticipated repayment date (pppenalties) as the loan-specific variables. Age of the property is include as a control in addition to the age of the loan. We add age^2 as as a control variable too to capture the non-linear relationship of aging of the loan with the delinquency classes. We calculate "time to maturity" to extract any strategic default behavior closer to the realized maturity of the loans.

We use the loan vintage (to capture if origination and underwriting standards have an effect on the delinquency class of the loans), 51 states in USA (msa, county, zip have severe missing values, hence the identification comes at a state level), property type (we bucket thousands of property types into 8 unique types), fixed/floating as dummy variables. We use "Number of Properties" (numprop) in a deal as a deal-specific variable.

We control for refinance pipeline and/or balloon payment by assigning a dummy if a loan is within 3 months threshold to its original scheduled maturity date. We use MIT Commercial In-

dex, National Council of Real Estate Investment Fiduciaries (NCREIF) regional property value indices. Additionally, we include state-level quarterly GDP (converted to monthly), monthly historical unemployment data by state and historical interest rates of different maturities.

The summary statistics for the cleaned data containing 9,617,333 observations of continuous variables is provided in Table 1. "One hot encoding" technique converts categorical variables as binary vectors without any order.

4 Parametric Models and Empirical Results

Our empirical results are based on models harnessing the unprecedented size of our sample set and the heterogeneity in the incentives of default and beliefs we investigate. The models calculate the accuracy of prediction for 7 different delinquency states starting from Current/Performing classes **W0_30D** which includes "loans with payments not received but still in grace period or not yet due", Late/Non-Performing classes **W30_60D**, **W60_90D** which includes loans with "Late Payment beyond 30-days but less than 60 days, beyond 60-days but less than 90-days, Default state **W90_120D** ((within 90 to 120 days of delinquency), Liquidation Proceedings & Final Resolution state **B120D** (beyond 120 days of delinquency), combined together as "limbo" loans. We try further states in the **PrfMatBal** (Performing, Mature and Balloon Payment due) and **NPrfMatBal** (Non-Performing, Mature and Balloon Payment due) classes to capture the incentives delay in resolution for foreclosed loans to REO/prepaid. Our deep learning model for commercial mortgage state probabilities is a nonlinear extension of the multinomial/ordered logistic regression model in Figure **??**. It can be naively viewed as a recursive formulation of logistic regression of nonlinear transformations of the explanatory variables.

We implement Naive Bayes, Multinomial (with Lasso and Ridge regularization) & Ordered Logit, parallelize Random Forest and implement adaptive gradient boosting after bagging and finally implement a Deep Recurrent Neural network and compare the prediction on different mortgage states on the holdout sample.

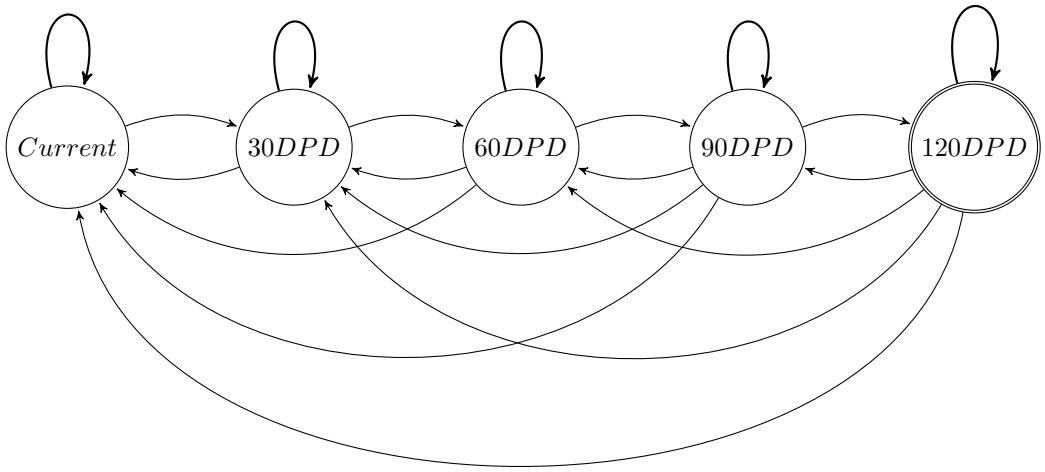
A **Naive Bayes** classifier estimates the conditional a-posterior probabilities of a categorical variable given independent covariates using the Bayes rule. The assumption of **independence** of the covariates is key to the success of the Naive Bayes classifier. We see that **W0_30D**, **W30_60D** & **W60_90D** classes have less mis-classification in Table 4 error than other models, since the assumption of independence among the co-variates holds until a loan is in these classes. But this analysis is still kept in the paper to motivate why we eventually need deep learning as a means of avoiding this strong assumption of independence among the covariates. [4]

[4]For classification, ROC curve analysis is conducted on each predictor. For two class problems, a series of cutoffs is applied to the predictor data to predict the class. The sensitivity and specificity are computed for each cutoff and the ROC curve is computed. The trapezoidal rule is used to compute the area under the ROC curve. This area is used as the measure of variable importance. For multi-class outcomes, the problem is decomposed into all pair-wise problems. For a specific class, the maximum area under the curve across the relevant pair-wise AUC's is used as the variable importance measure.

Ordered Logit exploits the natural order of delinquency classes and computes transition probabilities in that order. Ordered Logit does not allow all the back transitions shown in the above Finite State Automaton[5] of the loan delinquency classes. **Multinomial Logit** assumes Independence of Irrelevant Alternatives (IIA), which is not true in this situation as we will see in the next section. Suppose, hypothetically, there are two choices given to a borrower to be either **within 30 days of delinquency** or **between 90 days and 120 days of delinquency**. Clearly, the borrower would like to stick with the first choice, as the second choice classifies him/her in the default category and is detrimental for her creditworthiness from a lender's perspective. Now suppose, one more choice for being in **30 days to 60 days of delinquency** is given to the borrower, s/he may choose to rather be in this new state instead of less than 30 days of delinquency and may **strategically** miss one payment if there is a great investment opportunity for him/her in that one month horizon. In fact, none of the models (except Naive Bayes) can distinguish these three classes (**W0_30D**, **W30_60D** & **W60_90D**) and considers all of them as **Current Loans** in Table 4.

The borrower can undertake this decision as she/she is already some days in delinquency and she/she wouldn't mind going to the next bucket until she/she falls in the bucket for **90 days to 120 days of delinquency**. In this situation, the borrower's creditworthiness doesn't change that much from a lender's perspective. hence, the odds for being in the **"less than 30 days delinquency"** to being in the classes of **90 days to 120 days of delinquency** will change drastically in the presence of this new choice of being in **30 days to 60 days of delinquency**. hence the IIA assumption is clearly violated. The following Finite State Automaton details all possible transitions so that the above arguments can be visualized.[6]

The different delinquency classes will seriously affect the default behavior of the loans, e.g., simply having a cutoff for default would imply that we are assuming that loans "Within 30



[5]

[6]In a Multinomial Logit Model, log-odds of each delinquency state with respect to the "Current" state assumes a linear specification. The odds that a loan has a delinquency classes j as opposed to the baseline, depending only on individual loan-specific covariates is defined as:

$$\frac{Pr(Y_i = j|Z_i = z)}{Pr(Y_i = 0|Z_i = z)} = exp(Z^{'}\gamma_j) \tag{5}$$

days delinquency", "Between 30 days and 60 days delinquency" and "Between 60 days and 90 days delinquency" have the same default risk. We motivate why this assumption is not true. If this were to be true, that a borrower would only pay off just before 90 days delinquency in order to avoid default and facing derogatory consequences. The fact that the above three buckets represent different defauly risk categories implies that the borrower's default behavior will change when she/she is between 30 days and 60 days of delinquency compared to the situation when all the above three categories are bucketed together as "Non-Default".

In this Table 4 the row labels are the predicted classes and the column labels are the actual classes. As is evident from the Sensivity and Error , the Multinomial Logistic Model can correctly classify the Current or "**W0_30D**" really well, but the Specificity is really low, i.e., the model cannot classify the loans that are **not** in "**W0_30D**" correctly vis-a-vis the "**W0_30D**" class. Also the error rates for the classes "**W30_60D**", "**W60_90D**" are 100% which means the model cannot identify any those classes correctly. Similarly, the classes "**W90_120D**" and "**B120D**" are also identified very poorly the Multinomial Logistic Model. In fact, some of the risks (Current Note Rate, LTV, Unemployment Rate, etc.) are misrepresented in Multinomial Logit, e.g., if local Unemployment increases, the *Current* Commercial Loan Default should increase (Table 2). **Lasso** and **Ridge** does not improve the performance of Multinomial Logit in Table 4.[7]

$$Pr(Y_i = j|Z_i = z) = \frac{exp(Z^{'}\gamma_j)}{1 + \sum_{l=1}^{J} exp(Z^{'}\gamma_l)} \tag{6}$$

$$Pr(Y_i = j|Z_i = z) = \frac{1}{1 + \sum_{l=1}^{J} exp(Z^{'}\gamma_l)} \tag{7}$$

the choice Y_i takes on non-negative, un-ordered integer values $Y_i \in \{0, 1, ..., J\}$

Multinomial logistic regression does not assume normality, linearity, or homoskedasticity; it has a well-behaved likelihood function, a special case of conditional logit.

A more powerful alternative to multinomial logistic regression is discriminant function analysis which requires these assumptions are met. Multinomial logistic regression also assumes non-perfect separation.

The Independence of Irrelevant Alternatives (IIA) assumption inherent in Multinomial Logit Model implies that adding or deleting alternative outcome categories does not affect the odds among the remaining outcomes.

$$Pr(Y_i = j|Y_i \in \{j, l\}) = \frac{Pr(Y_i = j)}{Pr(Y_i = j) + Pr(Y_i = l)} = \frac{exp(X_{ij}^{'}\gamma)}{exp(X_{ij}^{'}\gamma) + exp(X_{il}^{'}\gamma)} \tag{8}$$

This can be tested by the Hausman-McFadden test. There are alternative modeling methods, such as alternative-specific multinomial probit model, or nested logit model to relax the IIA assumption.

[7]The random effects approach (BLP methodology) of multiple unobserved choice characteristics to avoid IIA is attractive, substantively and computationally, compared to the nested logit or unrestricted multinomial probit models.

The more accurate our model, the more we can trust the importance measures and other interpretations. Measuring linear model goodness-of-fit is typically a matter of residual analysis. (A residual is the difference between predicted and expected outcomes). The problem is that residual analysis does not always tell us when the model is biased. Also the marginal effect of features towards classifying the response set does not have a clear interpretation in terms of sensitivity and directionality. We list the co-efficients of Multinomial Logit for the sake of completeness. (Table 2)

5 Vanilla Machine Learning Models & Empirical Results

5.1 Distributed Random Forest

The confusion matrices of the delinquency classes for in-sample/training set are calculated for the entire data in Table 3 and also subsample in Table 3 until the December, 2006 for stress testing the robustness for Out-of Sample Prediction during the Financial crisis in Figure 8 As is evident from the Error in Table 4, the **Distributed Random Forest** Model can correctly classify the Current or "**W0_30D**" **completely** in Figure **??**. Also the error rates for the classes "**W30_60D**", "**W60_90D**" are 98% which means the model cannot identify any those classes correctly but better than Multinomial Logit Model. Similarly, the classes "**W90_120D**" and "**B120D**" are also identified very poorly but better than the Multinomial Logistic Model.[8]

As is evident from the **Out-of-Sample** Errors in Table 4, the Distributed Random Forest Model can correctly classify the Current or "**W0_30D**" **completely**. here the column labels

[8]Recursive partitioning, a critical data mining tool, shelps in exploring the stucture of a data set, while developing easy to visualize decision rules for predicting a categorical (classification tree) or continuous (regression tree) outcome. This section briefly describes CART modeling, conditional inference trees, and random forests.

Random Forests are developed by aggregating decision trees and can be used for both classification and regression. Each tree is a weak learner created from bootstrapping from subset of rows and columns. More trees will reduce the variance.It alleviates the issue of overfitting, can handle a large number of features. It shelps with feature selection based on importance. It is user-friendly with two parameters: number of trees (default 500) and variables randomly selected as candidates at each split, $\sqrt{ntree}$ for classification and $ntree/3$ for regression.

1. Draw ntree bootstrap samples.
2. For each bootstrapped sample, grow un-pruned tree by choosing the best split based on a random sample of mtry predictors at each node.
3. Predict new data using majority votes for classification and average for regression based on ntree trees.

For each bootstrap iteration and related tree, prediction error using data not in bootstrap sample, namely, "Out Of Bag Error" is estimated.

R's randomForest splits based on the Gini criterion and H2O trees are split based on reduction in Squared Error (even for classification). H2O also uses histograms for splitting and can handle splitting on categorical variables without dummy (or one-hot) encoding. Also, R's randomForest builds really deep trees, resulting in pure leaf nodes, leading to constant increments in prediction and ties and hence relatively lower AUC.The trees in H2O's random forest aren't quite as deep and therefore aren't as pure, allowing for predictions that have some more granularity to them and that can be better sorted for a better AUC score.

CART models an outcome y_i for an instance i as:

$$y_i = f(x_i) = \sum_{m=1}^{M} c_m I x_i \in R_m \tag{9}$$

where each observation x_i belongs to exactly one subset R_m, c_m is the mean of all training observations in R_m.

The estimation procedure takes a feature and computes the cut-off point that minimizes the Gini index of the class distribution of $\mathbf{y}$, making the consequent subsets as different as possible. The algorithm is repeated until a stopping criterion is reached. Tree-based methods are invariant to monotonic feature transformations and can handle both continuous and categorical variables. A tree of depth L can capture L-1 interactions, making interpretations straightforward, providing counterfactuals. However, they are not so good at handling linear relationships, as the use of the step function in splitting is inherently non-linear.

are the predicted classes and the row labels are the actual classes.Also the error rates for the classes "**W30_60D**", "**W60_90D**" are 100% which means the model cannot identify any those classes any better than Multinomial Logit Model. Similarly, the classes "**W90_120D**" and "**B120D**" are also identified very poorly but better than the Multinomial Logistic Model **Out-of-Sample**. The Out-of-sample predictions worsen during the Financial Crisis. [9]

5.2 Gradient Boosting Machine (GBM)

As is evident from the **In-Sample** Errors in Table 4, the **Gradient Boosting Machine** can correctly classify the Current or "**W0_30D**" **completely**. Also the error rates for the classes "**W30_60D**", "**W60_90D**" are almost 100% which means the model cannot identify those classes any better than Multinomial Logit Model. Similarly, the classes "**W90_120D**" and "**B120D**" are also identified very poorly but better than the Multinomial Logistic Model **In-Sample** in Figure 3. We also attach the Variable Importance for GBM during the using data before Financial Crisis in Figure 8

Here the column labels are the predicted classes and the row labels are the actual classes. The out-of-sample predictions for GBM perform as good as Deep Learning in our preliminary analysis. This method uses the same approach as a single tree, but sums the importances over each boosting iteration (see the gbm package vignette).[10]

[9]Along with training a model that classifies accurately in a hold-out sample, one needs to be able to interpret the model results. Feature importance is the most useful interpretation tool (such as the coefficients of linear models), to identify important features. Most random Forest (RF) implementations also provide measures of feature importance via permutation importance. Permutation importance is obtained by observing the effect on model accuracy of randomly shuffling each predictor variable. This technique is broadly-applicable because it doesn't rely on internal model parameters even while using Lasso or Ridge regularization in the presence of highly correlated features.

For each tree, the prediction accuracy on the out-of-bag portion of the data is recorded. Then the same is done after permuting each predictor variable. The difference between the two accuracies are then averaged over all trees, and normalized by the standard error. If the standard error is equal to 0 for a variable, the division is not done. here is the Variable Importance table **??** for the Random Forest Model Khandani, Kim, and Lo (2010a). The Variable Importance for Out-of-Sample predictions during the Financial Crisis in Figure 8 give similar results.

[10]GBM **?** forms an ensemble Kuncheva (2003) of weak prediction models in a stage-wise fashion and is generalized by allowing an arbitrary differentiable loss function. Boosting trees increases their accuracy, but decreases speed and user interpretability. The gradient boosting method generalizes tree boosting to minimize these drawbacks.

At each step m, $1 \leq m \leq M$ of gradient boosting, an estiamtor h_m is computed from the residuals of the previous model predictions. Friedman (2001) proposed regularization by shrinkage:

$$F_m(x) = F_{m-1}(x) + \nu\gamma_m h_m(x) \tag{10}$$

where $h_m(x)$ represents a weak learner of fixed depth, γ_m is the step length and ν is the learning rate or the shrinkage factor. XGBoost Chen and Guestrin (2016) is a faster and more accurate implementation of the Gradient Boosting algorithm Chen, Lundberg, and Lee (2018).

6 DNN for disentangling Delinquency Incentives

Deep Learning is a form of machine learning with multiple layers that learns multiple levels of representations for different levels of abstraction Sirignano, Sadhwani, and Giesecke (2016). It captures associations and discovers regularities within sets of patterns; it is suited for high volume, high dimensional data. It performs well when the relationships are dynamic or non-linear in Figure 5, when the standard regression models perform very poorly. No assumptions on normality, linearity, variable independence are needed.

We use a multi-layer feedforward DNN, trained with stochastic gradient descent using back-propogation. Each compute node trains a copy of the global model parameters on its local data with asynchronous multi-threading and contributes periodically to the global model via model averaging across the DNN. We tune both and Optimizer and Model-specific Hyperparameters in Section A.2. We use SMOTE technique to reduce class imbalance. In Section A.3, we use Variable Importance to compare the most significant marginal contributions of the features in Section A.4 and plan to test Shapley values in Section A.7.

6.1 Model Results

As is clear from the similar counts of the loans of different categories in the in-sample confusion matrix in Table 3, we have **undersampled** the W0_30D class/Current Loans to alleviate the class imbalance problem. The Out-of-Sample predictions across different delinquency classes are as good as GBM in Table 4.

The accuracy of predictions change dramatically, if NOI is taken out. We also conduct a robustness check by leaving out each of the *strategic* variables from the DNN model. When year and month fixed effects are taken out in Figure 7, NOI loses its importance significantly! This clearly indicates that NOI is not a statistically significant variable by itself. It is used strategically by borrowers when clustering of macro-economic events happen and when NOI is taken out, the constraint variable like prepayment penalty clause and voluntary prepayment variable like current unscheduled pricinpal payment show up higher in the variable importance in Table 4 than LTV. Similarly, when Prepayment Penalties are taken out of the list of variables. When Balloon Payment constraints are taken out of the list of variables.

For neural networks, two popular methods for constructing Variable Importance (VI) scores are the Garson algorithm, later modified by Goh (1995), and the Olden algorithm Olden, Joy, and Death (2004). For both algorithms, the basis of these importance scores is the network's connection weights. The Garson algorithm determines VI by identifying all weighted connections between the nodes of interest. Olden's algorithm, on the other hand, uses the product of the raw connection weights between each input and output neuron and sums the product across all hidden neurons. This has been shown to outperform the Garson method in various simulations. For DNNs, a similar method due to Gedeon (1997) considers the weights connecting the input features to the first two hidden layers (for simplicity and speed); but this method can be slow for large networks. For Deep Learning, there is no impact of scaling, because the numbers were already scaled. hence, the relative importance is the same as the absolute importance in Figure 7.

Model-agnostic interpredibility separates interpretation from the model. Compared to model-specific approaches, model-agnostic VI methods are more flexible (since they can be applied to any supervised learning algorithm). We intend to further investigate model-agnostic methods for quantifying global feature importance using three different approaches: 1) PDPs, 2) ICE curves, and 3) permutation Greenwell, McCarthy, Boehmke, and Liu (2018).

As is clear from the preliminary analysis, the Net Operating Income (NOI), the Prepayment Penalty clause and the Balloon Payment triger are significantly high in the variable importance table 4. NOI is even higher than LTV as found in the VI tables for other previous models. This provides evidence on how these three less statistically significant features contribute much more towards the classification, via highly non-trivial and non-linear interactions with more statistically significant variables.

Net Operating Income (NOI), a key indicator for an investment property's financial standing, is the income generated by an investment property after subtracting the operating expenses and vacancy losses but before principal and interest payments, capital expenditures, depreciation, and amortization. NOI calculation involves the following key variables. Potential Rental Income assumes zero vacancy or could be based on a rental market analysis. Vacancy losses represent the loss of income due to tenants vacating the property and/or tenants defaulting on their lease payments. Total Operating Expenses on an Investment Property could include Property Taxes, Rental Property Insurance, Property Management Fees, Maintenance and Repairs, Miscellaneous Expenses, etc. Debt service, depreciation, leasing commissions, tenant improvements, repairs to wear and tear, income taxes, and mortgage interest expenses are not included in the calculation of net operating income. This is because NOI is unique to the property itself and does not include other expenses that are specific to the investor/borrower in Table 4.

Some of the calculations that rely on NOI include Cap Rate (property's potential rate of return), ROI, Debt Coverage Ratio, Cash Return on Investment. The use of NOI provides an overview of a property's ongoing operating revenue. NOI analysis can be manipulated since a property owner can choose to accelerate or defer certain expenses. The NOI of a property is not always constant, it can change depending on how the property is managed. Because other expenses are not considered in NOI (e.g., interest expense, debt service, income taxes, capital expenditures), the actual cash flow that a property can generate may differ after all these other expenses are paid. If projected rents are used to calculate NOI, it can throw off the net operating income formula if these rents differ from market rents.

Our Deep Learning Variable Importance table in Figure 7 shows that NOI is the key endogenous feature for understanding strategic delinquency behavior of the commercial mortgage borrowers. We intend to further investigate how prepayment penalty clause and indicator for balloon payments co-determine the strategic delinquency behavior along with the NOI using Shapley values by capturing the marginal contributions.

6.2 Future Work: Robust Causality via Irreversibility

We focus now on the feedforward part of our DNN. For the activation function of "ReLU With Dropout", the aggregation is done in the following way:

$$X^{(i+1)} = a(WX^{(i)} + b_i) \tag{11}$$

The Loss Function for such a recursive mechanism of applying Logistic Classification can be described as a Cross-Entropy between neighboring layers. Since the same transition function is used between the layers, maximizing the cross entropy between layers amounts to minimizing loss function. This is the essence through which parameter sharing is exceuted in DNN. We prove that this parameter sharing actually increases cross-entropy thereby increasing the accuracy of the predictions in an "Irreversible Way".

Since the support of the probability distributions, or in other words, the number of neurons with activation values varies across the different layers, the definition of Kullback-Leibler Divergence or Relative Entropy needs to be generalized. In order to calculate relative entropy between distributions with different support, a new formulation which is independent of the cardinality of the support set is required. We intend to develop that on a separate paper.

But for motivation, we calculate the absolute Shannon Entropy for each layer of the Deep Neural Network and show that is almost constant across the layers. This confirms that no information has been lost in the process, but a better abstraction has been obtained from the coarse representation from the input variables. We relate the above phenomenon to "reducing the number of counterfactuals" between neighboring layers in the DNN. This irreversible transformation defines a new and more robust form of "Causality" by reductio ad absurdum.

7 Conclusion

Using DNN, non-linearities of dependence of the response and interactions among features can be captured, without specifying the relationships apriori. The Variable Importance chart/diagram is different for different algorithms, e.g., Distributed Random Forest, Gradient Boosting Machine, Deep Learning. Net Operating Income, Prepayment Penalty Clause, Appraisal Reduction, Non-Recoverability, Bankruptcy Flag, Liquidity Proceeds, Liquidity Expense and Balloon Payment Indicator co-determine the strategic delinquency behavior of a commercial mortgage borrower. Surprisingly, Loan-to-Value is unable to capture this Strategic behavior and is not even statistically significant during Financial Crisis. Hyperparameter Tuning during the implementation of DNN is still an art and not a science. We intend to point out some heuristics/empirical relationships among the hyperparameters, both for optimization and the model itself. The classification of critical delinquency states of systems when the agent decisions are endogenous while the data is highly unbalanced across states can only be captured through deep learning. We intend to explore the heterogeneity of these Liquidity-constrained and Strategic Defaulters in a more theoretical way in future work, which is line with Geanakoplos (2010), that Optimists are price setters. This could have pricing implications too. Eventually, our goal is to create a Measure of Causality for Deep Learning Models in the near future.

References

Albanesi, S. and D. Vamossy (2019): "Predicting Consumer Default: A Deep Learning Approach," 70.

Albertini, F. and E. D. Sontag (1993): "For neural networks, function determines form," *Neural Networks*, 6, 975 – 990.

Ambrose, B. W., J. C. A. Capone, and Y. Deng (2001): "Optimal Put Exercise: An Empirical Examination of Conditions for Mortgage Foreclosure," *The Journal of Real Estate Finance and Economics*, 23, 213–234.

Ambrose, B. W. and R. J. B. Jr. (2012): "The Adjustable Balance Mortgage: Reducing the Value of the Put," *Real Estate Economics*, 40, 536–565.

Ambrose, B. W. and A. B. Sanders (2003): "Commercial Mortgage-Backed Securities: Prepayment and Default," *The Journal of Real Estate Finance and Economics*, 26, 179–196.

An, X., Y. Deng, and S. A. Gabriel (2009): "Value Creation through Securitization: Evidence from the CMBS Market," *Journal of Real Estate Finance and Economics*, 38, 302–326.

Bajari, P., C. S. Chu, and M. Park (2008): "An Empirical Model of Subprime Mortgage Default From 2000 to 2007," NBER Working Papers 14625, National Bureau of Economic Research, Inc.

Bansal and Viswanathan (1993): "No Arbitrage and Arbitrage Pricing: A New Approach," *Journal of Finance*, 48, 1231–1262.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984): *Classification and Regression Trees*, Monterey, CA: Wadsworth and Brooks.

Brown, S., W. Goetzmann, and A. Kumar (1998): "The Dow Theory: William Peter Hamilton's Track Record Reconsidered," *Journal of Finance*, 53, 1311–1333.

Butaru, F., Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique (2016): "Risk and risk management in the credit card industry," *Journal of Banking & Finance*, 72, 218–239.

Campbell, T. S. and J. K. Dietrich (1983): "The Determinants of Default on Insured Conventional Residential Mortgage Loans," *Journal of Finance*, 38, 1569–81.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002): "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, 16, 321–357.

Chen, H., S. Lundberg, and S. Lee (2018): "Hybrid Gradient Boosting Trees and Neural Networks for Forecasting Operating Room Data," *CoRR*, abs/1801.07384.

Chen, T. and C. Guestrin (2016): "XGBoost: A Scalable Tree Boosting System," *CoRR*, abs/1603.02754.

Childs, P., S. H. Ott, and T. J. Riddiough (1996): "The Pricing of Multiclass Commercial Mortgage-Backed Securities," *The Journal of Financial and Quantitative Analysis*, 31, 581–603.

Christopoulos, A. D., R. A. Jarrow, and Y. Yildirim (2008): "Commercial Mortgage-Backed Securities (CMBS) and Market Efficiency with Respect to Costly Information," *Real Estate Economics*, 36, 441–498.

Cleveland, W. S. (1979): "Robust Locally Weighted Regression and Smoothing Scatterplots," .

Cunningham, D. and C. Capone (1990): "The Relative Termination Experience of Adjustable to Fixed-Rate Mortgages," *Journal of Finance*, 45, 1687–1703.

Curley, A. J. and J. M. Guttentag (1974): "The Yield on Insured Residential Mortgages," in *Explorations in Economic Research, Volume 1, Number 1*, National Bureau of Economic Research, Inc, 114–161.

Daelemans, W., B. Goethals, and K. Morik, eds. (2008): *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML/PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part I*, vol. 5211 of *Lecture Notes in Computer Science*, Springer.

Deng, Y. (1999): "Network Power: Japan and Asia. Edited by Katzenstein Peter and Shiraishi Takashi. Ithaca, NY: Cornell University Press, 1997. 399p. $55.00 cloth, $22.50 paper," *American Political Science Review*, 93, 226–227.

Duffie, D. (2009): "Frailty Correlated Default," *Journal of Finance*, LXIV, 34–52.

Elul, R., N. Souleles, S. Chomsisengphet, D. Glennon, and R. Hunt (2010): "What "Triggers" Mortgage Default?" *American Economic Review*, 100, 490–94.

Foote, C., K. Gerardi, L. Goette, and P. Willen (2009): "Reducing Foreclosures: No Easy Answers," NBER Working Papers 15063, National Bureau of Economic Research, Inc.

Friedman, J. H. (1991): "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19, 1–67.

Garson, G. D. (1991): "Interpreting Neural-network Connection Weights," *AI Expert*, 6, 46–51.

Geanakoplos, J. (2010): "The Leverage Cycle," *NBER Macroeconomics Annual*, 24, 1–66.

Gedeon, T. D. (1997): "Data Mining of Inputs: Analysing Magnitude and Functional Measures," *Int. J. Neural Syst.*, 8, 209–218.

Goh, A. T. C. (1995): "Back-propagation neural networks for modeling complex systems," *AI in Engineering*, 9, 143–151.

Goldstein, E. B. and G. Coco (2015): "Machine learning components in deterministic models: hybrid synergy in the age of data," *Frontiers in Environmental Science*, 3, 33.

Goodfellow, I., Y. Bengio, and A. Courville (2016): *Deep Learning*, New York: The MIT Press.

Granger, C. (1995): "Modelling Nonlinear Relationships between Extended-Memory Variables," *Econometrica*, 63, 265–79.

Greenwell, B. M., A. J. McCarthy, B. C. Boehmke, and D. Liu (2018): "Residuals and Diagnostics for Binary and Ordinal Regression Models: An Introduction to the sure Package," *The R Journal*, 10, 381–394.

Guiso, L., P. Sapienza, and L. Zingales (2009): "Moral and Social Constraints to Strategic Default on Mortgages," Working Paper 15145, National Bureau of Economic Research.

——— (2013): "The Determinants of Attitudes toward Strategic Default on Mortgages," *Journal of Finance*, LXVIII.

Harrison, D. and D. L. Rubinfeld (1978): "Hedonic housing prices and the demand for clean air," *Journal of Environmental Economics and Management*, 5, 81–102.

Heaton, J. B., N. G. Polson, and J. H. Witte (2017): "Deep learning for finance: deep portfolios," *Applied Stochastic Models in Business and Industry*, 33, 3–12.

Hornik, K. (1991): "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, 4, 251 – 257.

Hornik, K., M. Stinchcombe, and H. White (1989): "Multilayer feedforward networks are universal approximators," *Neural Networks*, 2, 359 – 366.

Hutchinson, J. M., A. Lo, and T. Poggio (1994): "A Nonparametric Approach to Pricing and Hedging Derivative Securities via Learning Networks," *Journal of Finance*, 49, 851–89.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): *An Introduction to Statistical Learning: with Applications in R*, Springer.

Khandani, A., A. Kim, and A. Lo (2010a): "Consumer Credit Risk Models Via Machine-Learning Algorithms," *SSRN Electronic Journal*.

Khandani, A. E., A. J. Kim, and A. Lo (2010b): "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking and Finance*, 34, 2767–2787.

King, G. and L. Zeng (2001): "Logistic Regression in Rare Events Data," *Political Analysis*, 9, 137–163.

Klabjan, D. (2007): "Subadditive approaches in integer programming," *European Journal of Operational Research*, 183, 525–545.

Kuan, C.-M. and H. White (1994): "Adaptive Learning with Nonlinear Dynamics Driven by Dependent Processes," *Econometrica*, 62, 1087–1114.

Kuhn, M. and K. Johnson (2013): "Applied predictive modeling," .

Kuncheva, Ludmila I.and Whitaker, C. J. (2003): "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Machine Learning*, 51, 181–207.

Kvamme, H., N. Sellereite, K. Aas, and S. Sjursen (2018): "Predicting mortgage default using convolutional neural networks," *Expert Syst. Appl.*, 102, 207–217.

Lee, T. H., H. White, and C. Granger (1993): "Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests," *Journal of Econometrics*, 56, 269–290.

Levitin, A. and S. Wachter (2012): "Explaining the Housing Bubble," 36.

Lundberg, S. M. and S.-I. Lee (2017): "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Curran Associates, Inc., 4765–4774.

Montufar, G. F., R. Pascanu, K. Cho, and Y. Bengio (2014): "On the Number of Linear Regions of Deep Neural Networks," in *Advances in Neural Information Processing Systems 27*, ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Curran Associates, Inc., 2924–2932.

Olden, J. D., M. K. Joy, and R. G. Death (2004): "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data," .

Sirignano, J., A. Sadhwani, and K. Giesecke (2016): "Deep Learning for Mortgage Risk," *SSRN Electronic Journal*.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014): "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, 15, 1929–1958.

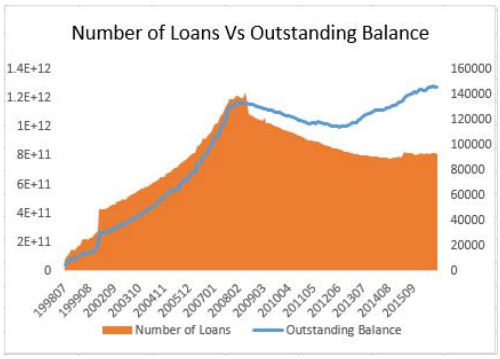
Sussmann, H. J. (1992): "Uniqueness of the weights for minimal feedforward nets with a given input-output map," *Neural Networks*, 5, 589 – 593.

Swanson, N. and H. White (1997): "A Model Selection Approach To Real-Time Macroeconomic Forecasting Using Linear Models And Artificial Neural Networks," *The Review of Economics and Statistics*, 79, 540–550.

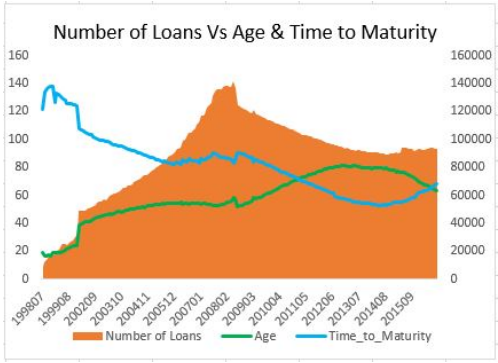
Titman, S., S. Tompaidis, and S. Tsyplakov (2005): "Determinants of Credit Spreads in Commercial Mortgages," *Real Estate Economics*, 33, 711–738.

von Furstenberg and George (1969): "Default Risk on FHA-Insured Home Mortgages as a Function of the Terms of Financing: A Quantitative Analysis," *Journal of Finance*, 24, 459–77.

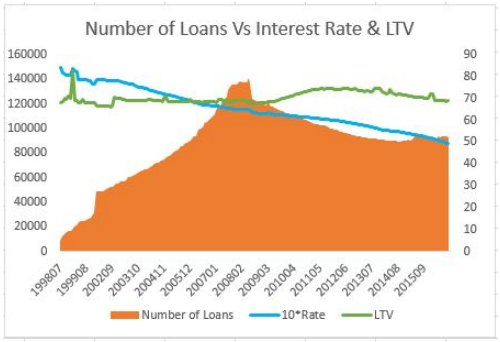
Yildirim, Y. (2008): "Estimating Default Probabilities of CMBS Loans with Clustering and Heavy Censoring," *The Journal of Real Estate Finance and Economics*, 37, 93–111.



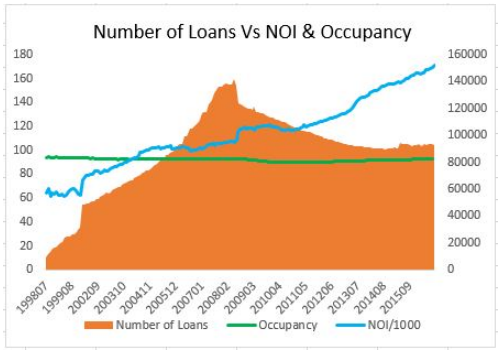
(a) Number of Loans vs. Outstanding Loan Balance



(b) Number of Loans vs. Age and Time to Maturity



(c) Number of Loans vs. Interest Rate and LTV



(d) Number of Loans vs. NOI and Occupancy

Figure 1: Motivating Diagrams of NOI, LTV

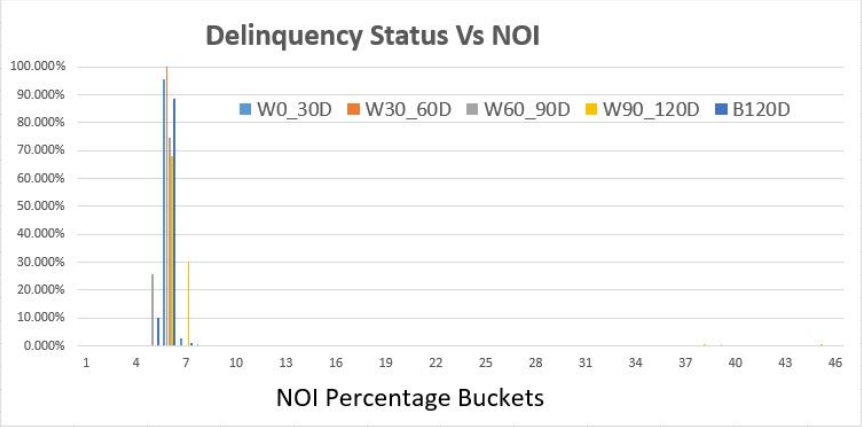


Figure 2: Distribution of Delinquency classes Vs NOI including dominant buckets

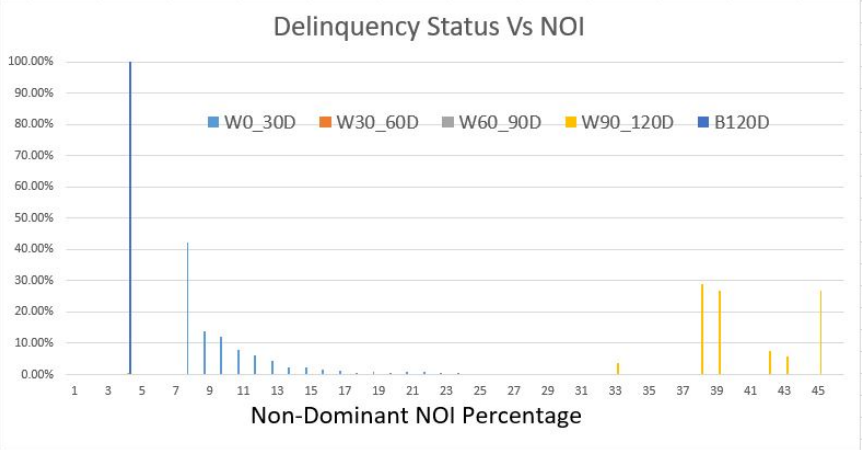
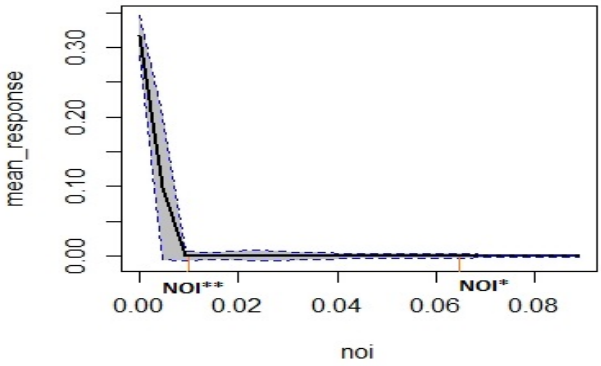
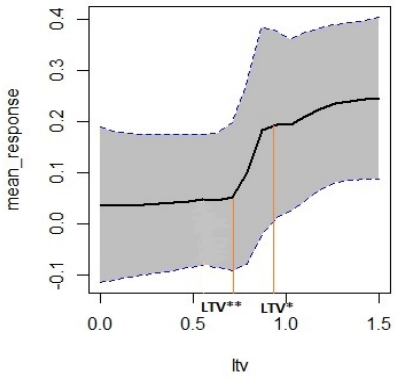


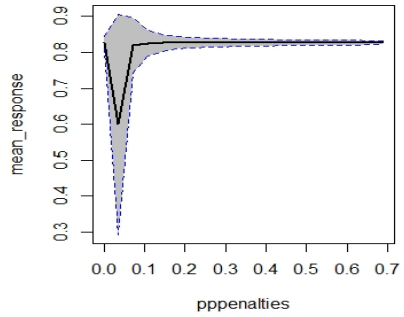
Figure 3: Distribution of Delinquency Classes Vs NOI excluding dominant buckets



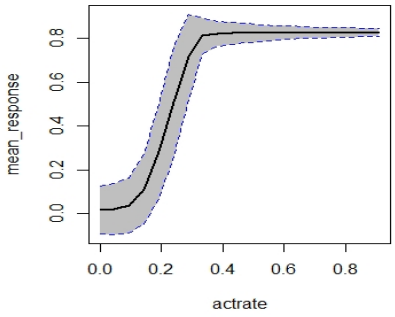
(a) Default Rate Vs Net Operating Income



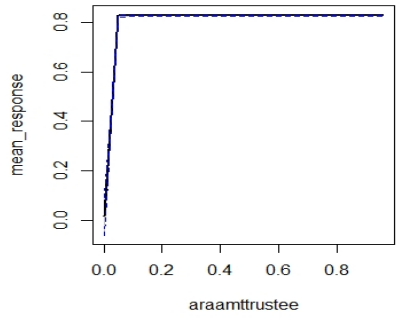
(b) Default Rate Vs Loan-to-Value



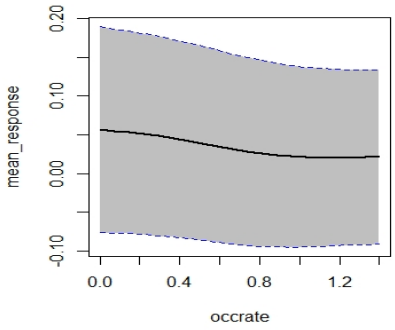
(c) Default Rate Vs Prepayment Penalties



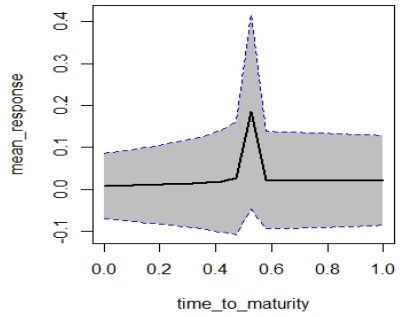
(d) Default Rate Vs Current Note Rate



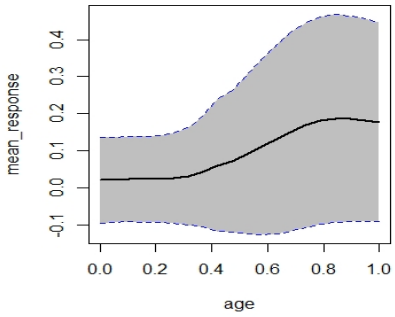
(e) Default Rate Vs Appraisal Reduction Amount



(f) Default Rate Vs Occupancy Rate

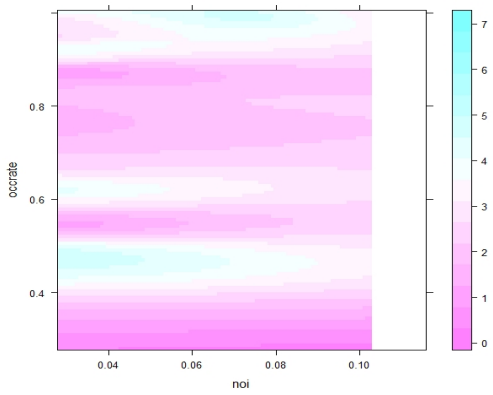


(g) Default Rate Vs Time to Maturity

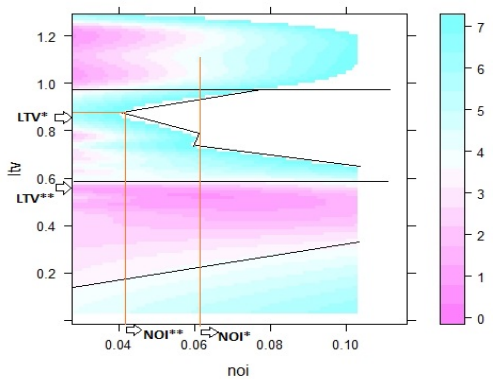


(h) Default Rate Vs Age of Loan

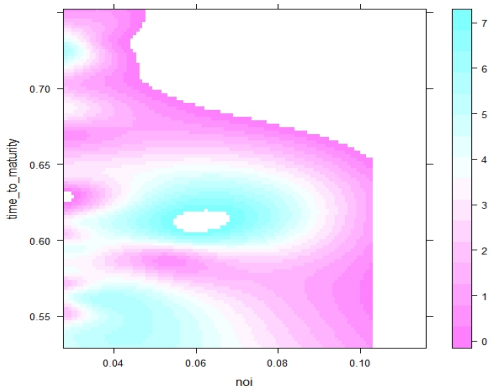
Figure 4: Partial Dependence Plots for Predicted Default Rate



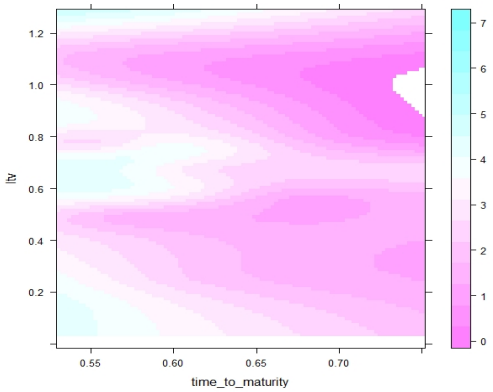
(a) Default Rate Vs NOI & Occupancy Rate



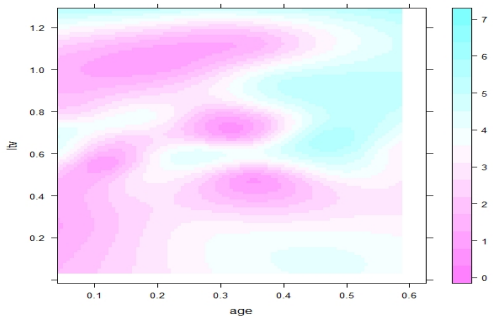
(b) Default Rate Vs NOI & LTV



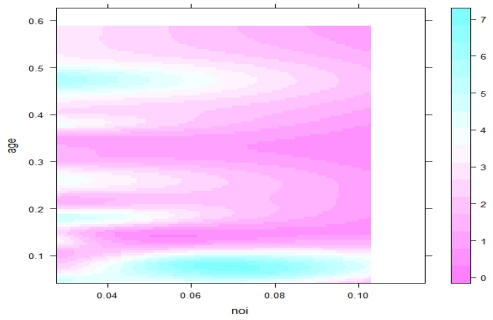
(c) Default Rate Vs NOI & Time to Maturity



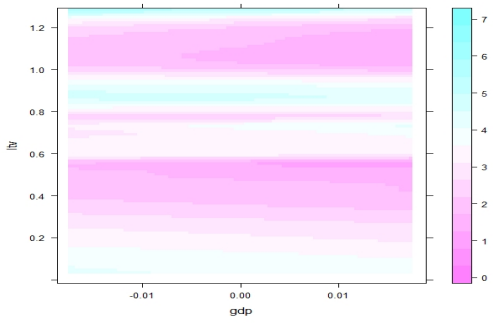
(d) Default Rate Vs Time to Maturity & LTV



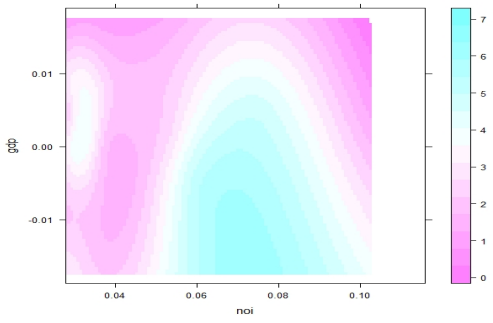
(e) Default Rate Vs Age & LTV



(f) Default Rate Vs NOI & Age

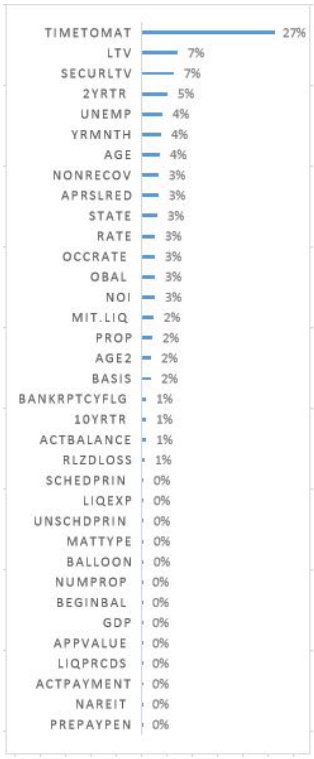


(g) Default Rate Vs GDP & LTV

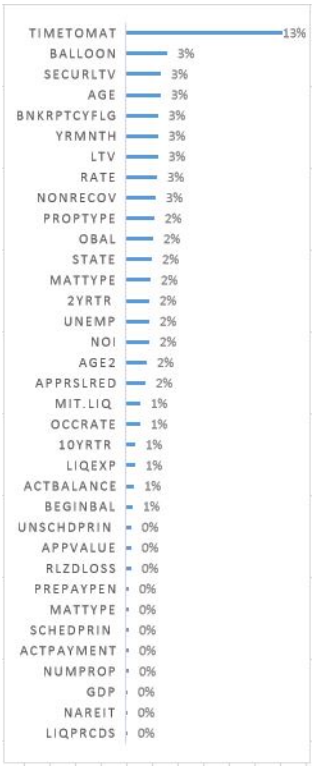


(h) Default Rate Vs NOI & GDP

Figure 5: Bivariate Heatmaps

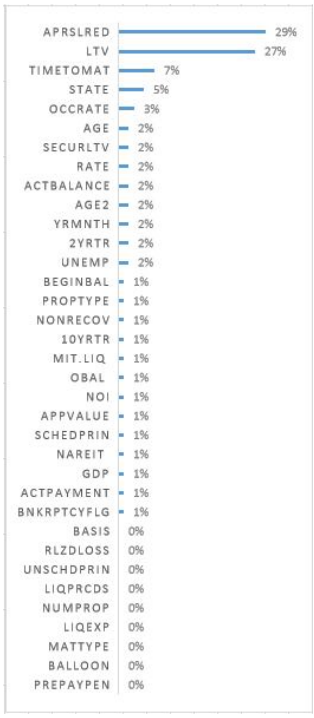


(a) Variable Importance: Lasso

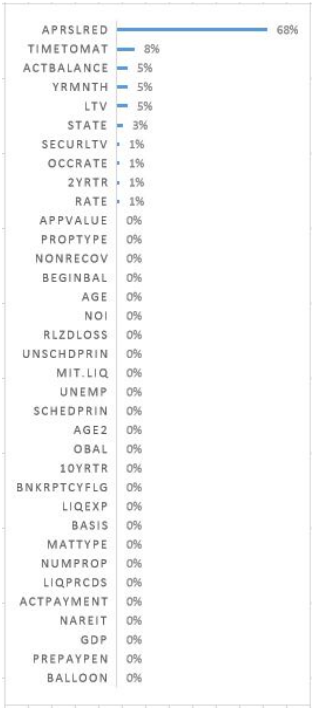


(b) Variable Importance: Ridge

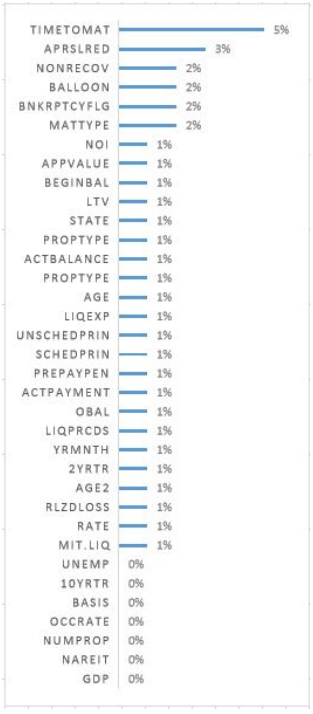
(c) Variable Importance: Ordinal



(d) Variable Importance: DRF

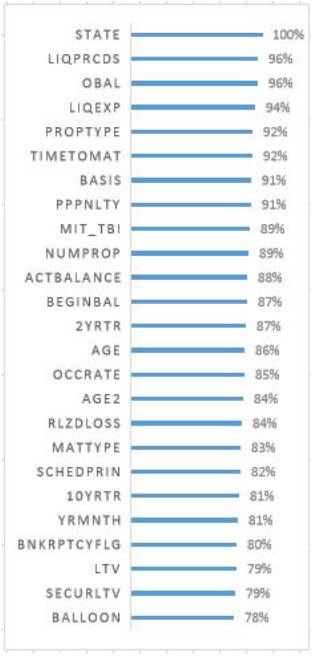


(e) Variable Importance: GBM

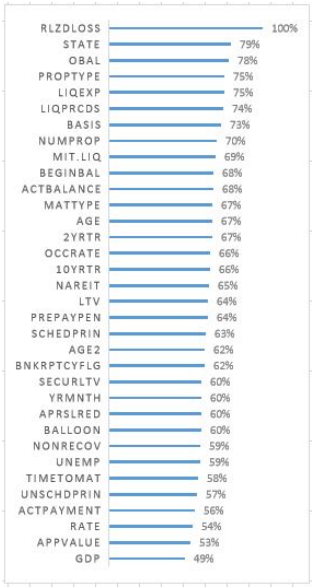


(f) Variable Importance: DNN

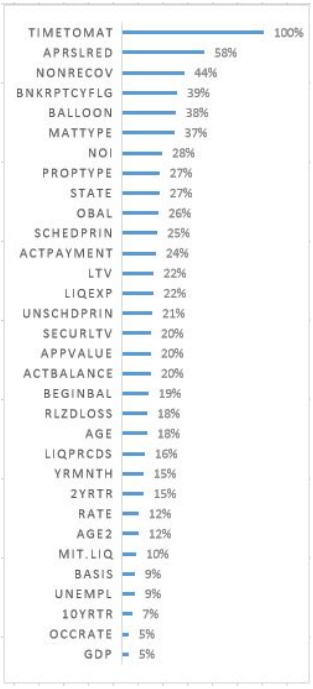
Figure 6: Variable Importance



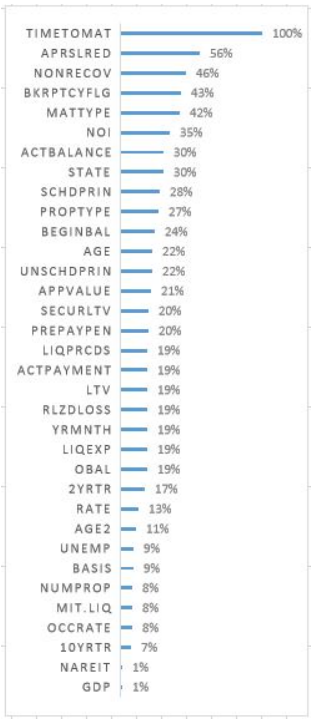
(a) VI without NOI



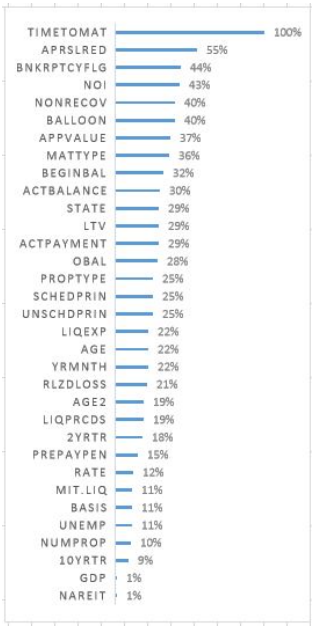
(b) VI without Year_Month



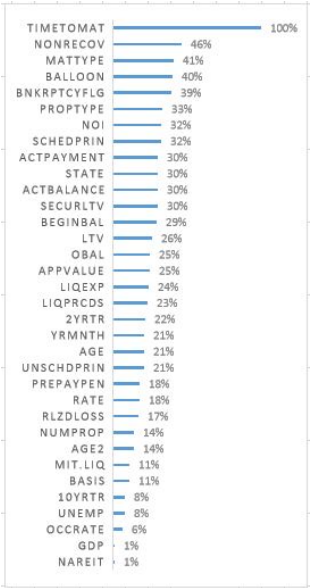
(c) VI without PrePayPen



(d) VI without Balloon Payment

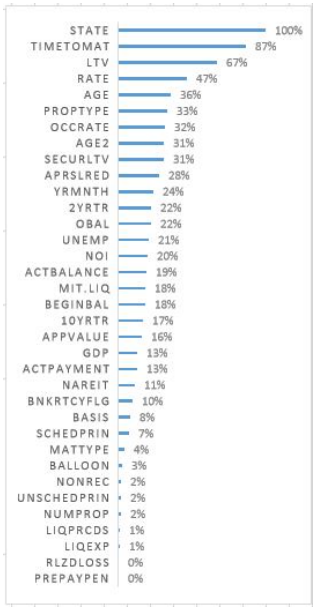


(e) VI without Occupancy

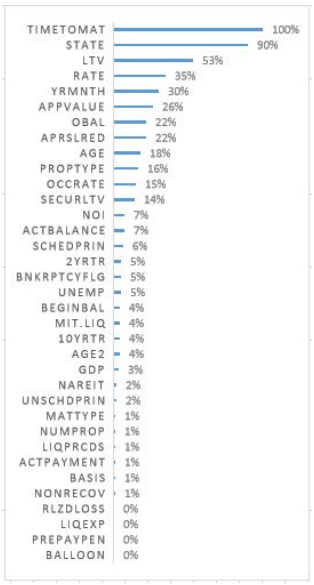


(f) VI without Appraisal Reduc

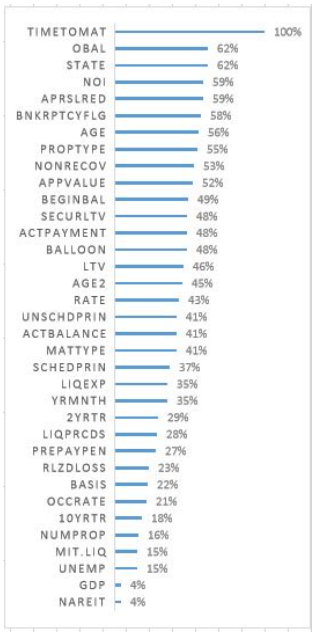
Figure 7: Variable Importance Leaving One Out



(a) VI in 2008: DRF



(b) VI in 2008: GBM



(c) VI in 2008: DNN

Figure 8: Variable Importance during Financial Crisis

A Deep Learning Model for CMBS

The purpose for a deep learning model is bourne out of the need to have transparency and accountability Albanesi and Vamossy (2019). By the very nature of the deep learning model, we do not have to add interaction terms in the specification of the model, especially in the case of high dimensional data. The sequential layers embody highly non-linear and non-trivial interaction among the variables and capture several latent fundamental features in the process. The causal interpretation of the covariates both in default Kvamme, Sellereite, Aas, and Sjursen (2018) and prepayment calculations have not been explored in details. The broader impact could be traced out by improved allocation of credit and aid in policy design (macroprudential, bankruptcy, foreclosure, etc.).

With the provision of enough hidden units, a neural network can mimic continuous functions on closed and bounded sets really well Hornik (1991), vis-a-vis the product and division of relevant features and their interactions. The advantage of more layers (as opposed to simply adding more units to existing layers) is that the later layers learn features of greater complexity by utilizing features of the lower layers as their inputs. Deep neural networks 10, with three or more hidden layers, need exponentially fewer units than shallow networks or logistic regressions with basis functions; see Montufar, Pascanu, Cho, and Bengio (2014) and Goodfellow, Bengio, and Courville (2016).

A.1 Dynamic Deep Neural Network

Recurrent Neural Networks (henceforth referred to as RNN) are neural networks that are preferred to Convolutional Neural Networks (CNN) for the purpose of processing sequential data $x^{(1)}, ..., x^{(\tau)}$ of variable length. Parameter sharing enables the extension of multilayer percep-

Figure 9: Interaction of LTV and NOI

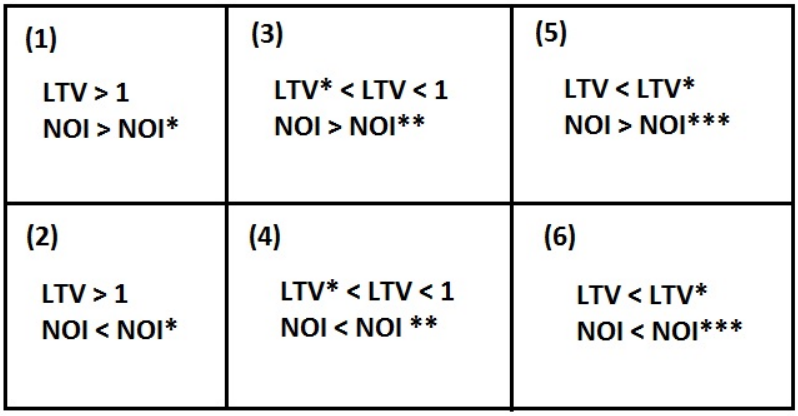


Table 1: Summary Statistics

Statistic	N	Min	Pctl(25)	Median	Mean	Pctl(75)	Max
beginbal	9,617,333	0	1,838,146	4,036,697	7,882,727	9,100,000	99,990,043
orig_bal	9,617,333	0	1.92e+06	4.10e+06	9.31e+06	9.14e+06	1.68e+09
rate	9,617,333	0	6	6	6	7	9
Scheduled_principal	9,617,333	0	1,060	4,235	15,665	9,838	430,000,000
Unscheduled_principal	9,617,333	0	0	0	5.03e+04	0	1.50e+09
balance_actual	9,617,333	0	1774511	3980981	7797502	9009824	99999000
payment	9,617,333	0	1.50e+04	2.86e+04	6.59e+04	5.78e+04	6.75e+08
pppenalties	9,617,333	0	0	0	420	0	29,477,125
liqproceeds	9,617,333	0	0	0	1.79e+04	0	2.56e+09
realizedloss	9,617,333	0	0	0	5.26e+03	0	2.04e+08
liqexpense	9,617,333	0	0	0	3.20e+03	0	1.06e+09
numprop	9,617,333	1	1	1	1	1	225
Appraisal_Reduc	9,617,333	0	0	0	1.32e+05	0	3.91e+08
SecurLTV	9,617,333	0	63	71	67	76	150
Face	9,617,333	0	6	6	6	7	9
NOI	9,617,333	0	2.68e+05	5.27e+05	1.22e+06	1.10e+06	1.09e+09
LTV	9,617,333	0	63	71	69	77	150
AppValue	9,617,333	1.62e+03	3.46e+06	6.80e+06	1.74e+07	1.45e+07	4.81e+10
OccRate	9,617,333	0.0	0.89	0.96	0.92	1.0	1.4
Basis	9,617,333	0.0	2.0	2.0	1.7	2.0	4.0
Unemp	9,617,333	0.019	0.049	0.061	0.066	0.080	0.154
GDP	9,617,333	0	0.0008	0.0016	0.0013	0.0048	0.0174
2YrTr	9,617,333	0.002	0.006	0.010	0.019	0.031	0.061
10YrTr	9,617,333	0.015	0.024	0.036	0.034	0.043	0.063
NAREIT	9,617,333	-0.45	-0.05	0.01	-0.01	0.04	0.38
MIT.Liq	9,617,333	0.017	0.022	0.036	0.025	0.105	0.184
time_to_maturity	9,617,333	0	35	66	80	100	765
age	9,617,333	0	31	60	63	90	409
age2	9,617,333	0	961	3600	5393	8100	167281

Table 2: Co-efficients of Multinomial Logistic Regression

	names	W0_30D	W30_60D	W60_90D	W90_120D	PrfMatBal	NPrfMatBal	B120D
1	Intercept	-24.83	-230.36	-223.45	-319.34	591.88	-170.20	272.02
2	State	-0	0.69	0.30	0.99	-2.12	-0.76	-0.08
52	PropType	-0.39	0.65	-0.12	0.61	0.55	0.62	0.49
60	MatType	0.23	-4.82	-2.87	-3.79	7.32	-57.90	-6.62
61	Balloon	0.33	-5.39	-3.48	-4.05	5.79	-53.02	-5.83
62	NonRecov	-1.59	-0.94	-0.75	0.69	-3.13	0.53	0.43
63	BnkrptcyFlg	-0.15	0.02	0.19	1.10	-0.78	-1.88	-0.09
64	BeginBal	-3.39	-2.13	-3.27	1.56	-0.67	-8.03	5.99
65	Obal	-1.18	21.62	19.06	-41.82	18.39	-7.16	-115.35
66	**Rate**	**-17.31**	29.77	29.74	34	-27.34	8.05	32.65
67	SchedPrin	134.71	153.33	54.63	89.44	156.85	173.07	41.50
68	Payment	44.44	-361.30	-6.64	-37.43	24.03	-1.95	57.25
69	UnschdPrin	45.46	-133.66	70.57	-47.70	57.08	-71.93	-172.29
70	PrePayPen	11.65	-253.05	-103.59	-61.27	-44.45	-370.52	16.91
71	ActBalance	2.25	0.94	2.73	-2.03	-0.28	8.07	-4.28
72	Liqprcds	-75.90	76.39	35.06	131.46	8.39	179.17	165.38
73	Liqexp	448.71	-1139.31	-2653.88	-518.47	-124.37	-1035.34	72.08
74	RlzdLoss	69.41	26.56	-48.44	-42.75	39.97	-43.45	-58.66
75	NumProp	-2.03	3.72	4.96	16.78	5.49	4.55	-23.82
76	AprslRed	-101.36	8.64	11.74	66.68	-20.31	49.31	87.47
77	SecurLTV	5.65	5.77	5.39	3.49	3.20	3.68	-4.19
78	NOI	159.73	-7.38	-39.41	27.03	124.60	-77.96	-149.62
79	**LTV**	**-6.55**	-3.10	-2.39	0.03	-3.56	-2.21	0.77
80	AppValue	-2.27	3.02	4.01	7.02	-201.11	11.14	9.66
81	OccRate	0.42	-3.10	-3.55	-3.51	-1.27	-2.28	-2.84
82	Basis	-7.14	-2.13	-1.89	0.28	19.61	48.04	5.77
84	Age	-6.55	-5.95	-4.29	-4.77	-2.70	6.02	8.03
85	Age2	16.16	0.62	-2.87	-1.52	12.32	-10.49	-8.52
86	**Unemp**	**-1.93**	17.51	21.77	18.87	25.90	21.55	**-12.23**
87	GDP	-0.84	0.75	0.27	3.30	8.21	2.35	-6.66
88	2YrTr	4.89	-7.05	-9.52	-34.94	-5.31	-50.57	-3.23
89	10YrTr	10.94	18.14	18.82	37.29	-20.60	7.96	15.26
90	NAREIT	0.10	0.13	0.32	0.20	0.02	0.41	0.17
91	MIT.Liq	0.29	-0.12	-0.35	-0.31	0.27	-0.34	3.71
92	TimeToMat	4.69	4.63	4.43	4.57	-101.20	-140.81	2.91

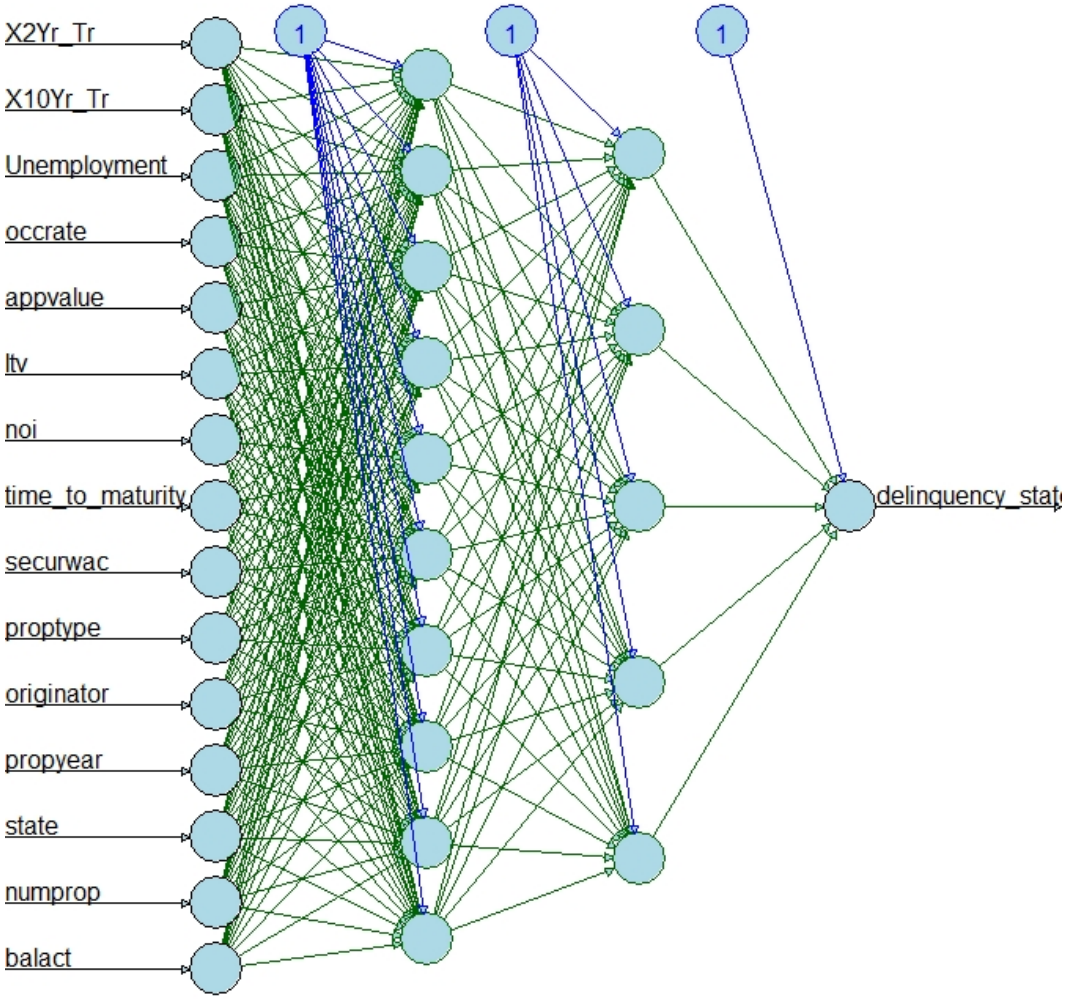


Figure 10: Deep Neural Network

Table 3: Cross-Validation Training Errors of Models across Delinquency Classes

	Ord	Mult	Lasso	Ridge	DRF	GBM	DL
W0_30D	0.01	0.01	0.01	0.01	0.01	0.01	0.0
W30_60D	**0.61**	1	1	1	**0.98**	1	1
W60_90D	**0.63**	1	1	1	**0.98**	1	1
W90_120D	0.36	0.48	0.48	0.47	**0.21**	**0.27**	**0.21**
PrfMatBal	1	1	1	1	**0.76**	**0.84**	**0.78**
NPrfMatBal	1	0.67	0.7	0.74	**0.16**	**0.12**	**0.12**
B120D	1	0.8	0.87	0.83	**0.22**	**0.21**	**0.2**
Totals	0.03	0.02	0.02	0.02	0.01	0.01	0.01

Table 4: Out-of-Sample Test Errors of Models across Delinquency Classes

	Ord	Multi	Lasso	Ridge	DRF	GBM	DL
W0_30D	0.01	0.01	0.01	0.01	0.01	0.01	0.01
W30_60D	**0.89**	1	1	1	1	1	1
W60_90D	**0.59**	1	1	1	1	1	1
W90_120D	0.2	0.37	0.37	0.37	**0.14**	**0.15**	**0.17**
PrfMatBal	0.98	1	1	1	**0.97**	**0.96**	**0.95**
NPrfMatBal	1	0.42	0.42	0.47	**0.19**	**0.13**	**0.13**
B120D	1	1	1	1	**0.07**	**0.08**	**0.07**
Totals	0.98	0.03	0.03	0.03	0.03	0.03	0.03

trons (MLP) to RNNs. Using a time dependent parameter, sequence lengths not seen during training cannot be generalized. Nor can we translate the statistical robustness across different sequence lengths and also across time. Each member of the output is computed uniformly by the same update methodology applied to previous outputs, which results in parameter sharing vis-a-vis a very deep computational graph.

A computational Graph is a visualization for comprehending the unfolding of a recursive or recurrent computation, corresponding to a sequence of events, allowing for parameter sharing.

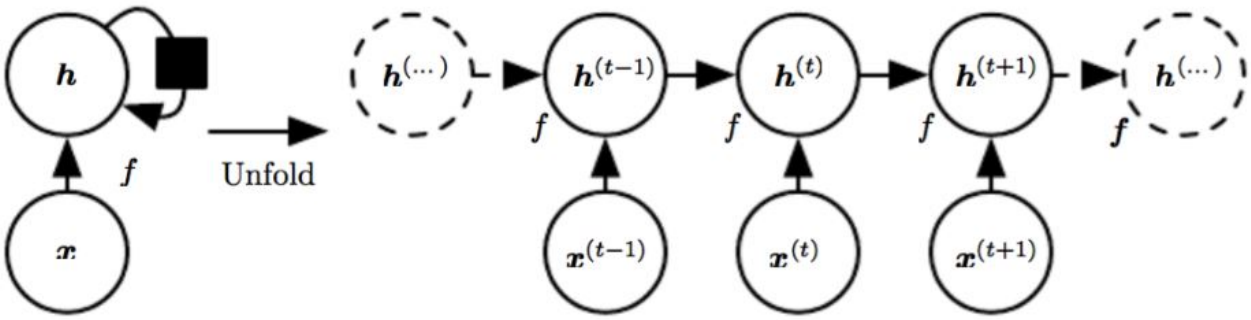


Figure 11: An recurrent network with no inputs

We consider a specification as in Goodfellow et al. (2016):

$$h^{(t)} = g^{(t)}(x^{(t)}, x^{(t-1)}, x^{(t-2)}, ..., x^{(2)}, x^{(1)}) \tag{12}$$

Here, $h(.)$ represents the hidden layers, which are lossy summaries. Depending on the training criterion, the summary will keep certain aspects of the past more accurately than others. The extreme case is when one can recover the input sequence from $h^{(t)}$, e.g., in autoencoder frameworks. $g^{(t)}$ needs the entire history $(x^{(t)}, x^{(t-1)}, x^{(t-2)}, ..., x^{(2)}, x^{(1)})$ as input to produce the current state.

The recursive specification described below has two distinct advantages:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}; \theta) \tag{13}$$

1. Whatever be the sequence length, the learned model has the same input size, because it is specified in terms of transition between consecutive states, rather than in terms of variable-length history of states.

2. The same transition function $f(.)$ can be used with the same parameters at every time step.

Learning a single shared parameter model is the key to generalizing to sequence lengths not in the training set. This facilitates the estimation with way less training examples than 0ther methods not using parameter sharing.

Under certain regularity conditions, the estimators are consistent and also asymptotically normal (see Hornik, Stinchcombe, and White (1989), Sussmann (1992) and Albertini and Sontag (1993) where they study the identifiability). One can regularize DNN using optimal hyperparameter tuning via cross-validation and "drop out" some sample values to reduce overfitting Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014).

Deep Neural Networks (DNN) 11 have an extensive set of current applications like: System identification and control (e.g., vehicle control, trajectory prediction, etc.), Game-playing and decision making (e.g., chess, poker, etc.), Pattern and sequence recognition (e.g., radar systems, face identification, signal classification, speech/image recognition, etc.), Medial diagnosis and finance (e.g., automated trading systems, cancer diagnosis, etc.).

A.2 Hyperparameter Tuning and Grid Search

Hyper-parameter tuning with Random Grid Search (RGS) tests different combinations of hyperparameters to find the optimal choice based on accuracy, without overfitting.

Hyperparameters can be divided into 2 categories:

- **Optimizer hyperparameters**
- **Model Specific hyperparameters**

There are several different **Optimizer hyperparameters**, e.g., a lower **Learning rate** will require a much longer time/epochs to reach the ideal state, whereas higher learning rate would overshoot the ideal state and the algorithm might not converge. The learning rate has to shepherd all of the parameters each with its own error curve. Error curves are not clean u-shapes. they have more complex shapes with local minima. **Mini-batch** size has an effect on the resource requirements of the training process (smaller leads to better fit, larger can speed up and generalize better). While the computational boost incentivizes us to increase the

minibatch size, this practical algorithmic benefit incentivizes us to actually make it smaller. **Number of Epochs** should be chosen based on the Validation Error. The manual way is to train the model as long as validation error keeps decreasing. There's a technique called Early Stopping to determine when to stop training the model.

Similarly, there are several different **Model hyperparameters**. Neural Networks are universal function approximator. **Number of hidden units** represents the required capacity to learn function. Another sheuristics involving the **first hidden layer** is that setting the number of hidden units larger than the number of inputs has been observed to give better results in number of tasks. **Non-Linear Activation Functions** could be Sigmoid, Hyperbolic Tangent, MaxOut, Leaky Recitified Liear Unit etc. **"WithDropout"** specification implies a random subset of the network is trained and the weights of all sub-networks are averaged. It works together with the parameter hidden_dropout_ratios, which controls the amount of layer neurons that are randomly dropped for each hidden layer. Hidden dropout ratios are useful for preventing overfitting on learned features. As to the **number of layers**, 3 layer Neural Net will generally outperform a 2 layer one. But going even deeper rarely shelps much more. **Nesterov accelerated gradient** includes an adaptive learning rate smoothing factor (to avoid divisions by zero and allow progress), adaptive learning rate time decay factor (similarity to prior updates), etc. **Hidden layers** are the most important hyper-parameter to set for deep neural networks, as they specify how many hidden layers and how many nodes per hidden layer the model should learn. **L1 penalty** lets only strong weights survive and **L2 penalty** prevents any single weight from getting too big. **Rho** is similar to prior weight updates) and **Epsilon** prevents getting stuck in local optima. **Early stopping metric** is a stopping criterion (tolerance and rounds).

A.3 Class Imbalance Problem

Most classifiers are unable to distinguish minor classes Kuncheva (2003) and are sheavily influenced by major classes, e.g., the conditional probability of minor classes are underestimated in a logistic regression King and Zeng (2001), Tree based classifiers, and KNN yield high recall but low sensitivity when the data set is extremely unbalanced Daelemans, Goethals, and Morik (2008). There are a plethora of techniques to balance the data, e.g., oversampling, undersampling and Synthetic Minority Oversampling Technique (SMOTE) proposed by Chawla, Bowyer, Hall, and Kegelmeyer (2002). Oversampling methods replicate the observations from the minority class to balance the data. However, adding the same observation to the original data causes overfitting, where the training accuracy is high but forecast accuracy over testing data is low. Conversely, the under-sampling methods remove the majority of classes to balance data. Obviously, removing observations causes the training data to lose useful information pertaining to the majority class. SMOTE finds random points within nearest neighbors of each minor observation and by boosting methods generates new minor observations. Since the new data are not the same as the existing data the overfitting problem won't be an issue anymore, and we won't loose the information as much as with the under-sampling methods. For these reasons, this study considers the SMOTE function to balance the data.

A.4 Variable Importance

Random forests and gradient boosted decision trees can naturally quantify the importance

or relative influence of each feature. Other algorithms like naive Bayes classifiers and support vector machines are not capable of doing so and model-free approaches are generally used to measure each predictor's importance.

It is often beneficial (from an accuracy standpoint) to train and tune multiple state-of-the art predictive models (e.g., multiple RFs, GBMs, and deep learning NNs (DNNs)) and then combine them into an ensemble called a super learner through a process called model stacking. Even if the base learners can provide there own measures of variable importance, there is no logical way to combine them to form an overall score for the super learner.

Decision trees probably offer the most natural model-based approach to quantifying the importance of each feature. In a binary decision tree, at each node t, a single predictor is used to partition the data into two homogeneous groups. The chosen predictor is the one that maximizes some measure of improvement bit . The relative importance of predictor x is the sum of the squared improvements over all internal nodes of the tree for which x was chosen as the partitioning variable; see Breiman, Friedman, Olshen, and Stone (1984) for details. This idea also extends to ensembles of decision trees, such as RFs and GBMs. In ensembles, the improvement score for each predictor is averaged across all the trees in the ensemble. Fortunately, due to the stabilizing effect of averaging, the improvement-based variable importance metric is often more reliable in large ensembles (see James, Witten, Hastie, and Tibshirani (2013), Pg. 368). RFs offer an additional method for computing variable importance scores. The idea is to use the leftover out-of-bag (OOB) data to construct validation-set errors for each tree. Then, each predictor is randomly shuffled in the OOB data and the error is computed again. The idea is that if variable x is important, then the validation error will go up when x is perturbed in the OOB data. The difference in the two errors is recorded for the OOB data then averaged across all trees in the forest.

In multiple linear regression, the absolute value of the t-statistic is commonly used as a measure of variable importance. The same idea also extends to generalized linear models (GLMs). Multivariate adaptive regression splines (MARS), which were introduced in Friedman (1991), is an automatic regression technique which can be seen as a generalization of multiple linear regression and generalized linear models. In the MARS algorithm, the contribution (or variable importance score) for each predictor is determined using a generalized cross-validation (GCV) statistic.

For NNs, two popular methods for constructing variable importance scores are the Garson algorithm (Garson (1991)), later modified by Goh (1995), and the Olden algorithm (Olden et al. (2004)). For both algorithms, the basis of these importance scores is the network's connection weights. The Garson algorithm determines variable importance by identifying all weighted connections between the nodes of interest. Olden's algorithm, on the other hand, uses the product of the raw connection weights between each input and output neuron and sums the product across all hidden neurons. This has been shown to outperform the Garson method in various simulations. For DNNs, a similar method due to Gedeon (1997) considers the weights connecting the input features to the first two hidden layers (for simplicity and speed); but this method can be slow for large networks.

A.5 Filter-based approaches to variable importance

Filter-based approaches, which are described in Kuhn and Johnson (2013), do not make use of the fitted model to measure variable importance. They also do not take into account the other

predictors in the model. For regression problems, a popular approach to measuring the variable importance of a numeric predictor x is to first fit a flexible nonparametric model between x and the target Y ; for example, the locally-weighted polynomial regression (LOWESS) method developed by Cleveland (1979). From this fit, a pseudo-R2 measure can be obtained from the resulting residuals and used as a measure of variable importance. For categorical predictors, a different method based on standard statistical tests (e.g., t-tests and ANOVAs) is employed; see Kuhn and Johnson (2013) for details.
For classification problems, an area under the ROC curve (AUC) statistic can be used to quantify predictor importance. The AUC statistic is computed by using the predictor x as input to the ROC curve. If x can reasonably separate the classes of Y , that is a clear indicator that x is an important predictor (in terms of class separation) and this is captured in the corresponding AUC statistic. For problems with more than two classes, extensions of the ROC curve or a one-vs-all approach can be used.

A.6 Partial dependence plots

Harrison and Rubinfeld (1978) analyzed a data set containing suburban Boston housing data from the 1970 census. They sought a housing value equation using an assortment of features; see Table IV of Harrison and Rubinfeld (1978) for a description of each variable. The usual regression assumptions, such as normality, linearity, and constant variance, were clearly violated, but through an exhausting series of transformations, significance testing, and grid searches, they were able to build a model which fit the data reasonably well (R2 = 0.81). Their prediction equation is given in Equation (1). This equation makes interpreting the model easier. For example, the average number of rooms per dwelling (RM) is included in the model as a quadratic term with a positive coefficient. This means that there is a monotonic increasing relationship between RM and the predicted median home value, but larger values of RM have a greater impact.

To help understand the estimated functional relationship between each predictor and the outcome of interest in a fitted model, we can construct PDPs. PDPs are particularly effective at shelping to explain the output from "black box" models, such as RFs and SVMs. Not only do PDPs visually convey the relationship between low cardinality subsets of the feature set (usually 1-3) and the response (while accounting for the average effect of the other predictors in the model), they can also be used to rank and score the predictors in terms of their relative influence on the predicted outcome, as will be demonstrated in this paper.

Let $x = \{x_1, x_2, ..., x_p\}$ represent the predictors in a model whose prediction function is $\hat{f}(x)$. If we partition x into an interest set, z_s, and its complement, $z_c = x \setminus z_s$, then the "partial dependence" of the response on z_s is defined as:

$$f_s(z_s) = E[\hat{f}(z_s, z_c)] = \int \hat{f}(z_s, z_c) p_c(z_c) dz_c \tag{14}$$

where $p_c(z_c)$ is the marginal probability density of z_c : $p_c(z_c) = \int p(x) dz_s$. The above equation can be estimated from a set of training data by:

$$\bar{f}_s(z_s) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(z_s, z_{i,c}) \tag{15}$$

where $z_{i,c}$ $(i = 1, 2, ..., n)$ are the values of z_c that occur in the training sample; that is, we average out the effects of all the other predictors in the model.
Constructing a PDP in practice is rather straightforward. To simplify, let $z_s = x_1$ be the predictor variable of interest with unique values $\{x_{11}, x_{12}, ..., x_{1k}\}$. The partial dependence of the response on x_1 can be constructed as follows:

Input: the unique predictor values $x_{11}, x_{12}, ..., x_{1k}$;
Output: the estimated partial dependence values $\bar{f}_1(x_{11}), \bar{f}_1(x_{12}), ..., \bar{f}_1(x_{1k})$.
for $i \in 1, 2, ..., k$ **do**
(1) copy the training data and replace the original values of x1 with the constant x_{1i};
(2) compute the vector of predicted values from the modified copy of the training data;
(3) compute the average prediction to obtain $\bar{f}_1(x_{1i})$.
end
The PDP for x1 is obtained by plotting the pairs $\{x_{1i}, \bar{f}_1(x_{1i})\}$ for $i = 1, 2, ..., k$.

Algorithm 1: A simple algorithm for constructing the partial dependence of the response on a single predictor x_1.

Algorithm 1 can be computationally expensive since it involves k passes over the training records. Fortunately, it is embarrassingly parallel and computing partial dependence functions for each predictor can be done rather quickly on a machine with a multi-core processor. For large data sets, it may be worthwhile to reduce the grid size by using specific quantiles for each predictor, rather than all the unique values. For example, the partial dependence function can be approximated very quickly by using the deciles of the unique predictor values. The exception is classification and regression trees based on single variable splits which can make use of the efficient weighted tree traversal method described in **?**.

While PDPs are an invaluable tool in understanding the relationships uncovered by complex nonparametric models, they can be misleading in the presence of substantial interaction effects Goldstein and Coco (2015). To overcome this issue, Goldstein et al. introduced the concept of individual conditional expectation (ICE) curves. ICE curves display the estimated relationship between the response and a predictor of interest for each observation; in other words, skipping step 1 (c) in Algorithm 1. Consequently, the PDP for a predictor of interest can be obtained by averaging the corresponding ICE curves across all observations. Although ICE curves provide a refinement over traditional PDPs in the presence of substantial interaction effects, in Section 3.2, we show how to use partial dependence functions to evaluate the strength of potential interaction effects.

A.7 Interpretation and Feature Relevance

Borrower or obligor behavior is a high-dimensional function of the loan, tranche, deal, index and macroeconomic co-variates over time. We want to investigate how different covariates, both macroeconomic and idiosyncratic, interact to cause a state transition. We measure second order interaction between variables, however, this cross-derivative can, of course, be generalized to higher order interactions. We implement this using LIME (Local Interpretable Model Agnostic Explanations).

Explainable Artificial Intelligence (XAI) is now a legal mandate in regulated sectors such

as banking, insurance, etc. XAI is critical in ensuring that there are no age, gender, race, disabilities or sexual orientation biases in decisions dictated by regulations such as the European General Data Protection Regulation (GDPR) and the Fair Credit Report Act (FCRA).

SHapley Additive exPlanations (SHAP) criterion is a game theory concept : it is a method for assigning payouts to players in a coalition/co-operative game proportional to their marginal contribution towards the total payout in a **fair** way based on the Llyod Shapley's axioms.

1. **Symmetry:** Agents i and j are **interchangable** relative to the payoff function if they always contribute the same amount to every coalition of the other agents.

2. **Dummy Players:** Agent i is a dummy player if she/she contributes nothing to any coalition. Dummy players should receive no payoff.

3. **Additivity:** For the same agent i, we can divide the payoff of two games separately as if both games are played simultaneously.

$Shapleyvalue_{ij}$ for feature j and instance i is how much the feature value x_{ij} contributed towards prediction for instance i compared to baseline prediction for dataset. The baseline prediction is the median model output over training dataset. So, Shapley value is the average marginal contribution of a feature value over all possible combinations of instances and it identifies main drivers of default risk for each individual borrower. The idea can be related to Variance Inflation Factor (VIF) used a in a regression context to rule out irrelevant features from a set of of regression on individual regressors.

We use a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP Lundberg and Lee (2017) assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

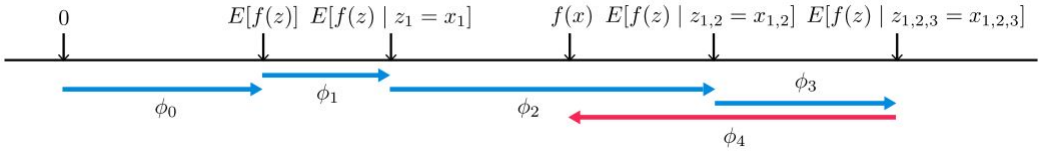


Figure 12: Shapley Value Interpretability

This method involves all feature subsets except a feature i, $S \subseteq F \setminus \{i\}$, where F is the set of all features. Shapley Value of a feature is the marginal impact of including that feature. Since, the effect of withholding a feature involves all possible subsets of other features, Shapley Value is the weighted average:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{\mid S \mid !(\mid F \mid - \mid S \mid -1)!}{\mid F \mid !} [f_{S \cup i}(x_{S \cup i}) - f_S(x_S)] \tag{16}$$