

Deep Learning for Mortgage Risk

Kay Giesecke

Center for Financial and Risk Analytics
Department of Management Science and Engineering
Stanford University

`people.stanford.edu/giesecke/`

Joint work with Justin Sirignano and Apaar Sadhwani

- We analyze mortgage risk using data for over **120 million loans** originated across the US between 1995 and 2014
- We develop, estimate, and test **dynamic machine learning** models for the transitions of a mortgage between states (current; 30, 60, 90+ days late; foreclosure; REO; paid off)
 - Basic building block is a deep neural network
- State transitions are allowed to depend upon both **static** and **time-varying variables**, including:
 - Loan-level features at origination
 - Loan-level performance variables
 - Local, regional, and national economic variables
- We develop an efficient **GPU parallel computing** approach to model fitting, testing, and prediction

Some takeaways

- The relationships between transitions rates and explanatory factors are often highly **non-linear**
- **Local risk factors** have a statistically and economically significant influence on transition rates
 - County-level unemployment rates
 - Zip-code level housing prices
 - Lagged foreclosure and prepayment rates in zip-code
- The **out-of-sample predictive performance** of our deep learning model is a significant improvement over that of other available models, such as logistic regression

- Data for 120 million prime and subprime mortgages originated across the US between 1995 and 2014
 - Source: CoreLogic
 - Extensive loan-level features at origination
 - Monthly performance update
- Data for local and national economic factors
 - Sources: Zillow, FHA, BLS, Freddie Mac, Powerlytics, CoreLogic
- ~ **3.5 billion monthly observations**, each described by roughly **300 feature variables**

Why don't we take a sample?

- Taking a truly random sample is difficult
- Some state transitions are moderately rare, and the wealth of training data improves model accuracy
- Sufficient geographic coverage is required to accurately measure the influence of local risk factors
- Larger data sets allow the fitting of richer models that capture the variety of risk and cashflow characteristics found across the entire range of mortgage products

Mortgage products in the data set

Product type	Total Data Set	Subprime	Prime
Fixed Rate	80.6 %	48 %	86.3 %
ARM	11.7 %	29 %	8.7 %
Hybrid 2/1	1 %	6.7 %	0 %
Hybrid 3/1	.63 %	2.2 %	.35 %
Hybrid 5/1	1.9 %	.22 %	2.2 %
Hybrid 7/1	.5 %	.005 %	.64 %
Hybrid 10/1	.24 %	.02 %	.28 %
Hybrid Other	.02 %	.02 %	.02 %
Balloon 5	.03 %	0 %	.03 %
Balloon 7	.03 %	.004 %	.04 %
Balloon 10	.004 %	.006 %	.004 %
Balloon 15/30	.2 %	1.07 %	.005 %
ARM Balloon	.2 %	1.3 %	.01 %
Balloon Other	.7 %	3.3 %	.26 %
Two Step 10/20	.003 %	0 %	.003 %
GPARM	.002 %	0 %	.002 %
Other	.7 %	4.3 %	.01 %

Summary statistics for some features

Feature	Mean	Median	25%	75%
FICO	720	730	679	772
LTV	74	79	63	90
Interest rate	5.8	5.8	4.9	6.6
Balance	190,614	151,353	98,679	238,000

Table: Prime mortgages

Feature	Mean	Median	25%	75%
FICO	634	630	580	680
LTV	74	80	68	90
Interest rate	7.8	7.8	6.3	9.6
Balance	160,197	124,000	68,850	210,000

Table: Subprime mortgages

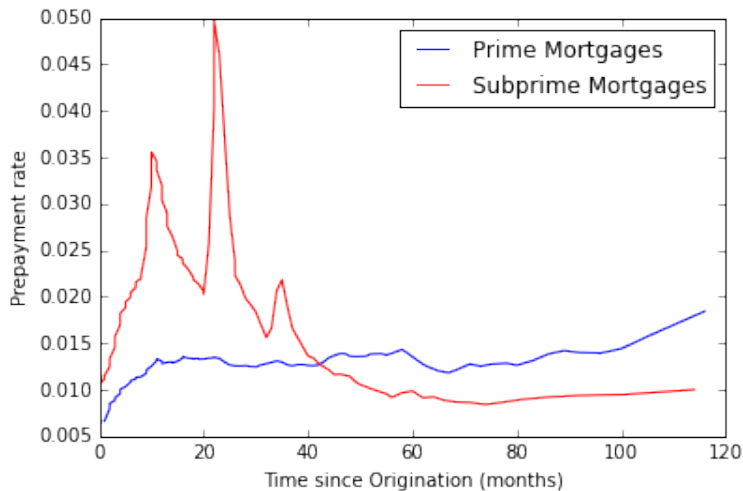
Monthly transition matrix for prime loans (95 million)

	Current	30	60	90+	Foreclosure	REO	Paid Off
Current	97	1.4	0	0	.001	0	1.6
30 days	34.6	44.6	19	0	.004	.003	1.8
60 days	12	16.8	34.5	34	1.6	.009	1.1
90+ days	4.1	1.4	2.6	80.2	10	.3	1.3
Foreclosure	1.9	.3	.1	6.8	87	2.5	1.3
REO	0	0	0	0	0	100	0
Paid off	0	0	0	0	0	0	100

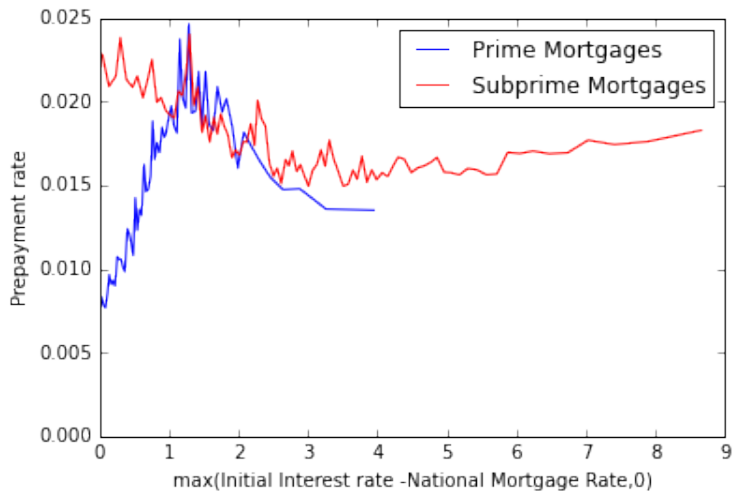
Prepayment Rate vs. Borrower FICO



Prepayment Rate vs. Loan Age



Prepayment Rate vs. Prepayment Incentive



Dynamic multi-state model framework

- Modeling the state transitions over time is a **dynamic supervised learning** problem (soft classification)
- The conditional probability that the n -th loan transitions from its state U_t^n at time t to state u at time $t + 1$ is

$$\mathbb{P}(U_{t+1}^n = u \mid \mathcal{F}_t) = h_\theta(u, X_t^n)$$

where X_t^n is a vector of explanatory variables including:

- The current state of the mortgage, U_t^n
 - The features of the n -th loan at t
 - Local, regional, and national economic factors at t
- Formulation captures loan-to-loan correlation due to geographic proximity and exposure to common risk factors

Baseline model: Logistic regression (LR)

- For g the softmax function $g(z) = \left(\frac{e^{z_1}}{\sum_{k=1}^K e^{z_k}}, \dots, \frac{e^{z_K}}{\sum_{k=1}^K e^{z_k}} \right)$ and $W \in \mathbb{R}^K \times \mathbb{R}^{d_x}$, $b \in \mathbb{R}^K$, take

$$h_{\theta}(u, x) = (g(Wx + b))_u$$

- To allow for nonlinear relationships, take basis functions $\phi : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_{\phi}}$, $W \in \mathbb{R}^K \times \mathbb{R}^{d_{\phi}}$, $b \in \mathbb{R}^K$, and set

$$h_{\theta}(u, x) = (g(W\phi(x) + b))_u$$

- This is a LR of the basis functions $\phi = (\phi_1, \dots, \phi_{d_{\phi}})$
 - Traditional examples: Polynomials, step functions, splines
- In a neural network (NN), the basis functions are chosen via learning a parameterized function ϕ_{θ} using the data

- A multi-layer NN repeatedly passes linear combinations of learned ϕ_θ through simple nonlinear link functions to produce a highly nonlinear function
- Formally, the output $h_{\theta,l} : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_l}$ of the l -th layer is:

$$h_{\theta,l}(x) = g_l(W_l h_{\theta,l-1}(x) + b_l),$$

where $W_l \in \mathbb{R}^{d_l} \times \mathbb{R}^{d_{l-1}}$, $b_l \in \mathbb{R}^{d_l}$, $h_{\theta,0}(x) = x$, and

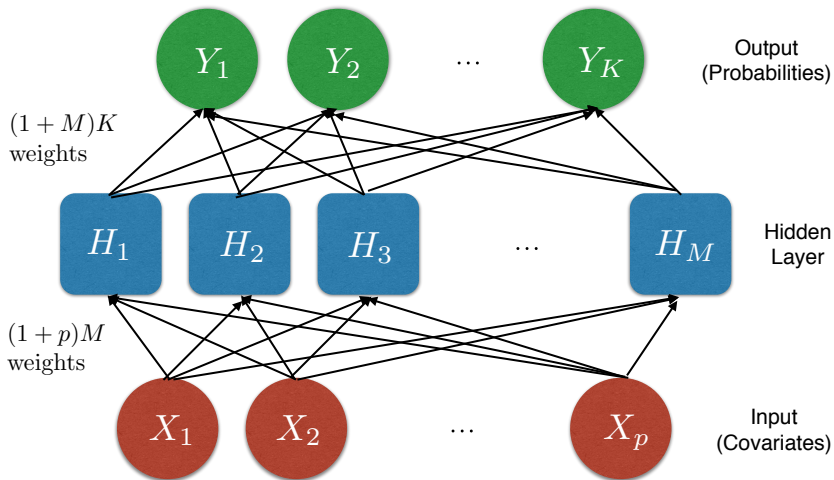
$$g_l(z) = (\sigma(z_1), \dots, \sigma(z_{d_l})), \quad z = (z_1, \dots, z_{d_l}) \in \mathbb{R}^{d_l}$$

$$g_L(z) = g(z) = \text{Softmax}$$

- The final output of the NN is given by:

$$h_\theta(u, x) = (h_{\theta,L}(x))_u$$

Neural network with single layer



- Number of hidden layers (“depth”)
 - Build up multiple layers of abstraction; each layer extracts features of the input for classification
- Number of hidden units M
 - The hidden units capture the nonlinearities in the data
- Activation function $\sigma(x)$
 - Sigmoid $1/(1 + e^{-x})$
 - Rectified linear unit (ReLU) $\max(x, 0)$
- Selection via cross-validation: 5 layers, 200-140 ReLU units

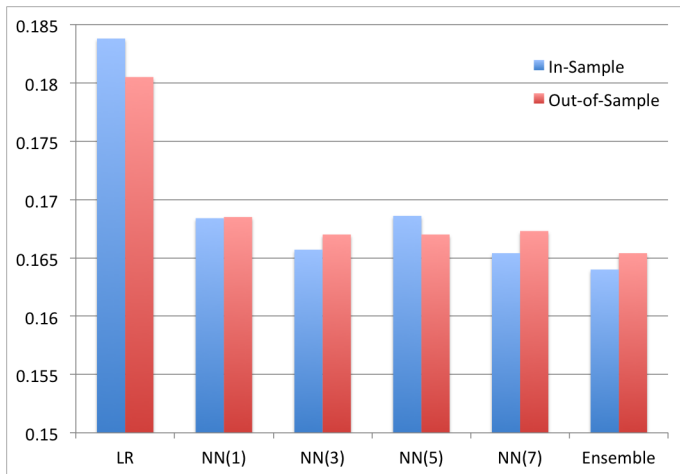
- We observe the variables $(X_t^1, \dots, X_t^N)_{t=0,1,\dots,T}$ for N loans
- Assuming the states U_t^1, \dots, U_t^N are independent given \mathcal{F}_{t-1} , the conditional log-likelihood of the states given the exogenous covariate data takes the form

$$L_N(\theta) = \sum_{t=1}^T \sum_{n=1}^N \log h_{\theta}(U_t^n, X_{t-1}^n)$$

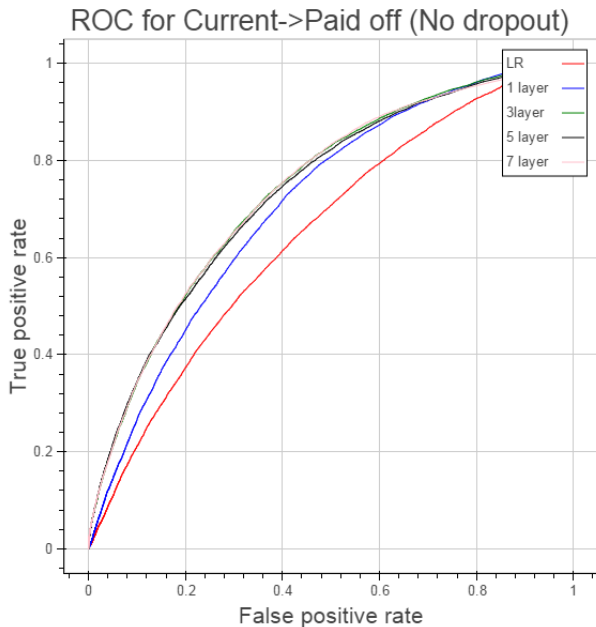
- Under mild conditions, the MLE $\arg \max_{\theta} L_N(\theta)$ is consistent and asymptotically normal as $N \rightarrow \infty$
- We use ℓ_2 -regularization, dropout, and ensembles to address overfitting

- We have 3.5 billion samples, each with 294 features
- We develop a **GPU parallel computing** environment running on a cluster of Amazon Web Services nodes
- We optimize $L_N(\theta)$ using **minibatch gradient descent** on a sequence of blocks of data
 - Gradient is available in closed form
 - Random starting values for θ
 - Batch size chosen by cross-validation
 - Adaptive learning rate (momentum based)
- We use the Torch scientific computing language for the optimization and the Python language for data processing

In- and out-of-sample errors vs. network depth



Out-of-sample ROC curves for month-ahead prediction



Out-of-sample AUCs for month-ahead prediction

Model	Current	30	60	90+	Forecl.	REO	Paid off
LR	.92719	.93206	.99069	.99670	.99781	.98980	.63521
NN (1)	.94142	.94081	.99155	.99690	.99798	.99113	.73764
NN (3)	.94211	.94117	.99168	.99691	.99799	.99187	.74250
NN (5)	.94254	.94140	.99170	.99691	.99799	.99205	.74679
NN (7)	.94052	.94109	.99169	.9969	.99798	.99187	.73336
Ensemble	.94423	.94200	.99181	.99696	.99802	.99251	.75814

Table: We report the AUC for the two-way classification of whether u or another event $u' \neq u$ occurs.

Out-of-sample AUCs for month-ahead prediction using ensemble

	Current	30	60	90+	Forecl.	REO	Paid off
Current	.762	.888	NA	NA	.556	.500	.754
30	.705	.694	.679	NA	.736	.564	.826
60	.668	.639	.701	.701	.807	.911	.734
90+	.719	.815	.915	.683	.690	.913	.792
Foreclosure	.836	.904	.928	.687	.661	.768	.739

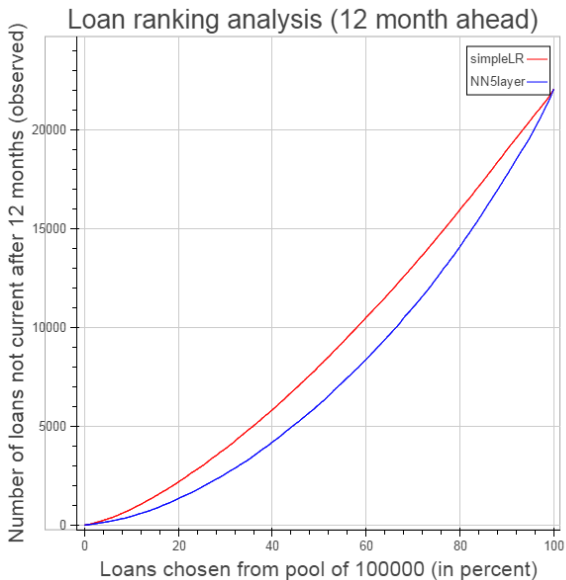
Table: The AUC for event $u \rightarrow u'$ is the AUC for the two-way classification of whether the transition $u \rightarrow u'$ or another transition $u \rightarrow u'' \neq u'$ occurs.

Differences in AUCs matter

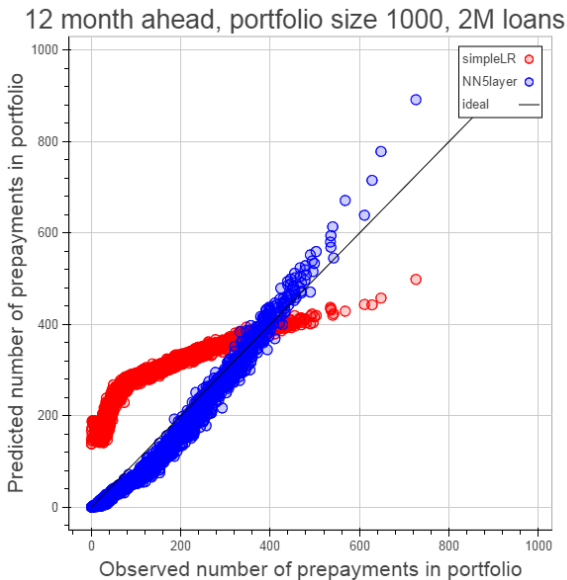
State	NN (5)	LR
Paid off	4.06	8.14
Current	93.28	89.09
30 days delinquent	1.60	1.54
60 days delinquent	0.36	0.36
90+ days delinquent	0.49	0.55
Foreclosure	0.19	0.30
REO	0.02	0.03

Table: Select best 20,000 out of 100,000 loans according to predicted probability of being current in 12 months. Performance of portfolio after (out-of-sample) 12 months recorded via percent of portfolio in each state.

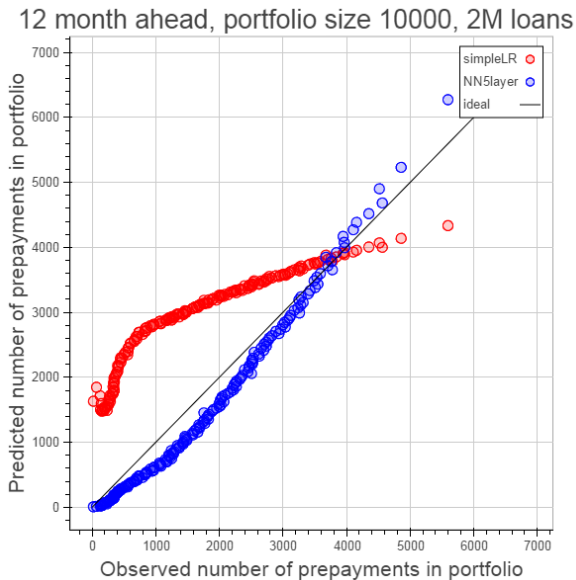
Loan ranking analysis



Out-of-sample prediction of pool-level prepayment



Out-of-sample prediction of pool-level prepayment




Global variable ranking by “leave-one-out”


Variable	Test Loss
State unemployment rate	1.160
Current outstanding balance	.303
Original interest rate	.233
FICO score	.204
Number of times 60dd in last 12 months	.179
Number of times current in last 12 months	.175
Original loan balance	.175
Total days delinquent ≥ 160	.171
Lien type = first lien	.171
Original interest rate - national mortgage rate	.170
LTV ratio	.169
Time since origination	.168
Debt-to-income ratio	.168
\vdots	\vdots

Ranking by gradient for current → paid off

Variable	Abs. Gradient
Current outstanding balance	.1707
Original loan balance	.0731
Original interest rate	.0603
FICO score	.0589
Current interest rate - national mortgage rate	.0538
Time since origination	.0460
Lagged prime prepayment rate in same zip code	.0392
Scheduled interest and principal due	.0334
Current interest rate - original interest rate	.0320
Lagged prime default rate in same zip code	.0289
State unemployment rate	.0288
Zillow zip code housing price change since origination	.0241
Original interest rate - national mortgage rate	.0230
Original appraised value	.0185
Original term of the loan	.0169
LTV ratio	.0137
Zillow zip code median house price change since origination	.0135
⋮	⋮

- Analyzed unprecedented data set of 120 million mortgages
- Developed and tested dynamic deep learning models for the transitions of mortgages between various states
- Out-of-sample predictive performance is a significant improvement over that of other models
- Designed efficient GPU parallel computing approach to fitting, testing, and prediction
- Results highlight the importance of nonlinear relationships and local (i.e., zip-code level) risk factors
- Building block for portfolio risk analytics and optimization
 - Sirignano & Giesecke (2015)
 - Sirignano, Tsoukalas & Giesecke (2015)

 Prignano, J., G. Tsoukalas & K. Giesecke (2015), Large-scale loan portfolio selection. [Working Paper, Stanford University.](#)

 Prignano, J. & K. Giesecke (2015), Risk analysis for large pools of loans.

Working Paper, Stanford University.

Loan-level features at origination

Feature	Values
FICO score	Continuous
Original debt-to-income ratio	Continuous
Original loan-to-value ratio	Continuous
Original interest rate	Continuous
Original balance	Continuous
Original term of loan	Continuous
Original sale price	Continuous
Buydown flag	True, False
Negative amortization flag	True, False
Occupancy Type	Owner occupied, etc.
Prepayment penalty flag	True, False
Product type	Fixed-rate, etc.
Loan purpose	Purchase, etc.
Documentation	Full documentation, etc.
Lien type	1st Position, etc.
Channel	Retail Branch, etc.
Loan type	Conventional, etc.
Number of units	1,2,3,4,5

Loan-level features at origination (continued)

Feature	Values
Appraised value < sale price?	True, False
Initial Investor Code	Portfolio Held, etc.
Interest Only Flag	True, False
Original interest rate – natl rate	Continuous
Margin for ARM mortgages	Continuous
Periodic rate cap	Continuous
Periodic rate floor	Continuous
Periodic pay cap	Continuous
Periodic pay floor	Continuous
Lifetime rate cap	Continuous
Lifetime rate floor	Continuous
Rate reset frequency	1,2,3, ... (months)
First rate reset period	1,2,3, ... (months since origination)
Pool insurance flag	True, False
Alt-A flag	True, False
Prime flag	True, False
Geographic state	CA, FL, NY, MA, etc.
Vintage (origination year)	1995, 1996, ..., 2014

Loan-level performance features

Feature	Values
Current state	Current, etc.
Number of days delinquent	Continuous
Current interest rate	Continuous
Current interest rate – national mortgage rate	Continuous
Time since origination	Continuous
Current balance	Continuous
Scheduled principal payment	Continuous
Scheduled principal + interest payment	Continuous
# months the mortgage's interest been less than the national mortgage rate and the mortgage did not prepay	Continuous
# occurrences of current in past 12 months	0-12
# occurrences of 30 days delinquent in past 12 months	0-12
# occurrences of 60 days delinquent in past 12 months	0-12
# occurrences of 90+ days delinquent in past 12 months	0-12
# occurrences of Foreclosed in past 12 months	0-12

Local and national economic risk factors

Feature	Values
Monthly zip code median house price per square feet (Zillow)	Continuous
Monthly zip code average house price (Zillow)	Continuous
Monthly three-digit zip code average house price (FHA)	Continuous
Monthly state unemployment rate (BLS)	Continuous
Yearly county unemployment rate (BLS)	Continuous
National mortgage rate (Freddie Mac)	Continuous
Median income in same zip code (Powerlytics)	Continuous
Total number of prime mortgages in same zip code (CoreLogic)	Continuous
Lagged subprime default rate in same zip code (CoreLogic)	Continuous
Lagged prime default rate in same zip code (CoreLogic)	Continuous
Lagged prime paid off rate in same zip code (CoreLogic)	Continuous
Current year	1999-2014