Vintage Model Notes

Contents

The final model output is lifetime expected credit losses for loss reserving under current expected credit loss (CECL). This paper will include …

1 CECL

1.1 What is CECL?

> "The Financial Accounting Standards Board (FASB) voted 5–2 on April 27th to move forward with the proposed Accounting Standards Update, Subtopic 326-20, more commonly known as the current expected credit loss (CECL) model. The final standard, expected to be passed in June 2016, will require institutions to reserve against losses on loans when they originate or acquire them and to re-estimate losses on an ongoing basis. The move to a CECL model is a departure from current GAAP, which requires institutions to defer the recognition of a credit loss until the loss is "probable and estimable," or has been incurred. The change is in direct response to the most recent global financial crisis.
>
> In the draft standard and subsequent Transition Resource Group (TRG) meetings, FASB members were intentionally non-prescriptive regarding acceptable methodologies available to institutions performing their quantitative loss calculations under the CECL standard. Board member Lawrence W. Smith reiterated the non-prescriptive nature, saying, "We are not prescribing specific methods of doing the allowance at all." Instead, the FASB listed examples of CECL-compliant calculations in their draft, including vintage analysis and the historical loss rate approach. Institutions will be free to use their judgment when developing estimation techniques as long as they are consistently applied over time and aim to faithfully estimate an actual life of loan loss."

Source: https://www.sageworks.com/banking/resources/CECL-Historical-Loss-Misconceptions/

1.2 Comments:

1. How institutions adopt CECL depends on what types of data they collect currently and plan to in the future
2. The policy is non-prescriptive -> more flexibility in what models can be used
3. Allowed models: vintage analysis, historical loss rate approach
4. Given model is compliant with CECL, each institution has different modeling requirements and standards

2 Data

This paper will use Fannie Mae's Single-Family Loan Performance credit dataset to test model hypotheses. Economic data will be obtained from FRED.

> "Fannie Mae provides loan performance data on a portion of its single-family mortgage loans to promote better understanding of the credit performance of Fannie Mae mortgage loans…
>
> The Single Family Fixed Rate Mortgage (primary) dataset contains a subset of Fannie Mae's 30-year and less, fully amortizing, full documentation, single-family, conventional fixed-rate mortgages. This dataset does not include data on adjustable-rate mortgage loans, balloon mortgage loans, interest-only mortgage loans, mortgage loans with prepayment penalties, government-insured mortgage loans, Home Affordable Refinance Program (HARP) mortgage loans, Refi Plus™ mortgage loans, or non-standard mortgage loans. Certain types of mortgage loans (e.g., mortgage loans with LTVs greater than 97 percent, Alt-A, other mortgage loans with reduced documentation and/or streamlined processing, and programs or variances that are ineligible today) have been excluded in order to make the dataset more reflective of current underwriting guidelines. Also excluded are mortgage loans originated prior to 1999, mortgage loans subject to long-term standby commitments, sold with lender recourse or subject to certain other third-party risk-sharing arrangements, or that were acquired by Fannie Mae on a negotiated bulk basis.
>
> The initial population of mortgage loans in the primary dataset included Fannie Mae acquisitions between January 1, 2000 and March 31, 2012 with corresponding monthly performance data as of December 31,

2012. Every quarter following the initial release, Fannie Mae updates the acquisition data to include a new quarter of acquired mortgage loans as of the prior year in addition to providing updated performance data as of the previous quarter. Fannie Mae releases updated information on or after the 20th of the month following the end of the quarter."

Source: http://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html

2.1 Description

Fannie Mae's loan data come in 2 files:

1. "Acquisition" file: static mortgage loan data at the time of the mortgage loan's origination and delivery to Fannie Mae (1 row per loan)
    a. Loan identifier
    b. Origination date
    c. Original interest rate
    d. Original Loan-to-Value ratio
    e. Credit score
    f. Origination channel
    g. And more…
2. "Performance" file: monthly performance data for each loan, from acquisition up until its current status as of the previous quarter (1 row per month of loan activity per loan)
    a. Loan identifier
    b. Calendar date
    c. Loan age
    d. Current unpaid principal balance (UPB)
    e. And more…

3 Model

Our analysis will aggregate loans by vintage. We are free to choose the time interval of the vintage and will use monthly periods. Loans that originated in the same month of the same year are grouped together as one vintage.

Final output of the model is the percentage of a vintage's original balance expected to be lost (ie. % of original balance that defaults plus any proceeds from resale/modification). To illustrate the methodology, below we discuss a simplified example. The columns are:

- ORIG_DTE: Year and month in which loans were originated (ie. vintage)
- ORIG_AMT: Original amount loaned for each vintage
- Net loss ratio: % of ORIG_AMT that was lost in each subsequent month (Month 0 is the month the loans were originated). Negative value if there was a gain due to default.

ORIG_DTE	ORIG_AMT	Net loss ratio				
	Loan age ->	Month 0	Month 1	Month 2	Month 3	Month 4
2018Q2	500000					
2018Q1	1000000	0.10%				
2017Q4	800000	0.00%	0.23%			
2017Q3	700000	0.05%	0.00%	1.00%		
2017Q2	1100000	0.00%	0.00%	0.04%	0.00%	
2017Q1	900000	0.20%	0.35%	0.40%	0.15%	0.04%

Data in the gray cells indicate the current period and we do not yet know how much was lost. Our goal is to project losses for loans originated in 2018/2. We make the assumption that we can estimate month 'X's' net loss ratio by taking some average of month 'X's' net loss ratios of previous loans. For instance, we estimate the net loss ratio of 2018/2 loans in Month 1 by using the net loss ratio of 2017/12 and 2017/11 loans in Month 1 (0.23% and 0% respectively). Of course, as discussed, we have more data than just historical net loss ratios of past loans. Our full model will include these other variables.

3.1 Dependent Variable

The dependent variable is defined as the percentages in the table. It is the net loss divided by the original balance for each vintage:

$$\text{Net loss ratio}_t = \text{Net loss of defaulted loans}_t \Big/ \text{Total ORIG\_AMT of vintage}_0$$

3.2 Independent Variables

The model has 3 groups of independent variables:

1. Age
   - Average loan default rates grow from a low level to a peak, say, six months after origination, before tapering off as only the best risks remain on the books
2. Loan quality
   - These include the origination characteristics of the loan, such as initial rate, loan-to-value ratio and credit score. With vintage analysis, we take a weighted average of these variables.
3. Selection of economic and risk drivers
   - The economic environment has a significant impact on default. Variables such as GDP growth rate and inflation will be included.
   - three broad classes of variables:

   1. overall level of economic activity (e.g., the unemployment rate, the rate of inflation, and the NBER recession indicator)
   2. the direction in which the economy is moving (e.g., the growth rates of GDP and industrial production)
   3. conditions in the financial markets (e.g., interest rates and stock market returns) (Figlewski)

4 Monthly time interval

Given greater computational capacity, we reduce the time interval from quarterly to monthly, enabling us to conduct a finer level of analysis.

Data: Quarter 4 data from years 2000 to 2005, inclusive.

A much wider range of data are now available. We use the vintage months between 2000 and 2005 August for training and 2005 September to December as the test set. We fit a tobit model using best forward subset selection based on AIC as above. The model is:

```
Call:
vglm(formula = "dflt_pct ~ bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5]) +
PROP_TYP_CO_wm+PROP_TYP_CP_wm+PROP_TYP_MH_wm+PROP_TYP_PU_wm+
bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5])*(PURPOSE_C_wm+PURPOSE_P_wm)+
bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5])*sqrt(I_count)+
bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5])*ORIG_RT_wm+
NUM_UNIT_wv*monthly_rGDP",
```

family = tobit(Lower = 0), data = train, trace = T)

Pearson residuals:

	Min	1Q	Median	3Q	Max
mu	-2.152	-0.5238	-0.2167	0.2443	13.08
loge(sd)	-5.186	-1.0167	0.0368	0.1900	140.99

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-8.692e-04	2.235e-03	-0.389	0.6974
(Intercept):2	-8.042e+00	1.564e-02	-514.139	< 2e-16 ***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1	4.012e-05	3.131e-03	0.013	0.9898
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2	-4.945e-04	2.004e-03	-0.247	0.8051
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3	-3.734e-04	2.479e-03	-0.151	0.8803
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4	2.563e-03	2.219e-03	1.155	0.2482
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5	-1.196e-03	2.597e-03	-0.461	0.6450
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6	4.068e-03	3.134e-03	1.298	0.1942
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7	-1.207e-02	9.862e-03	-1.224	0.2210
PROP_TYP_CO_wm	2.531e-04	8.393e-04	0.302	0.7629
PROP_TYP_CP_wm	-9.807e-03	1.858e-03	-5.279	
PROP_TYP_MH_wm	-1.132e-02	5.117e-03	-2.212	
PROP_TYP_PU_wm	6.538e-04	3.142e-04	2.081	
PURPOSE_C_wm	-1.521e-03	2.483e-03	-0.613	
PURPOSE_P_wm	-2.463e-04	7.750e-04	-0.318	
sqrt(I_count)	1.621e-06	1.298e-06	1.249	
ORIG_RT_wm	3.849e-05	2.577e-04	0.149	
NUM_UNIT_wv	-4.688e-04	2.595e-04	-1.806	
monthly_rGDP	-1.100e-02	2.663e-03	-4.130	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:PURPOSE_C_wm	2.383e-03	3.449e-03	0.691	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:PURPOSE_C_wm	2.167e-03	2.227e-03	0.973	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:PURPOSE_C_wm	6.986e-03	2.741e-03	2.548	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:PURPOSE_C_wm	9.102e-04	2.460e-03	0.370	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:PURPOSE_C_wm	3.966e-03	2.873e-03	1.380	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:PURPOSE_C_wm	-1.231e-03	3.667e-03	-0.336	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:PURPOSE_C_wm	1.123e-02	8.401e-03	1.337	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:PURPOSE_P_wm	2.497e-04	1.096e-03	0.228	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:PURPOSE_P_wm	6.283e-04	6.969e-04	0.902	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:PURPOSE_P_wm	2.704e-03	8.709e-04	3.105	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:PURPOSE_P_wm	7.224e-04	7.732e-04	0.934	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:PURPOSE_P_wm	7.982e-04	9.334e-04	0.855	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:PURPOSE_P_wm	8.983e-04	1.284e-03	0.700	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:PURPOSE_P_wm	-3.223e-04	5.394e-03	-0.060	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:sqrt(I_count)	1.160e-07	1.881e-06	0.062	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:sqrt(I_count)	-8.925e-07	1.172e-06	-0.762	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:sqrt(I_count)	2.931e-07	1.495e-06	0.196	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:sqrt(I_count)	-4.555e-07	1.310e-06	-0.348	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:sqrt(I_count)	1.113e-06	1.545e-06	0.721	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:sqrt(I_count)	1.521e-06	1.640e-06	0.927	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:sqrt(I_count)	-2.225e-06	1.902e-06	-1.170	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:ORIG_RT_wm	5.167e-05	3.636e-04	0.142	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:ORIG_RT_wm	9.584e-05	2.309e-04	0.415	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:ORIG_RT_wm	-2.678e-04	2.882e-04	-0.929	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:ORIG_RT_wm	-3.202e-04	2.567e-04	-1.247	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:ORIG_RT_wm	6.094e-05	3.028e-04	0.201	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:ORIG_RT_wm	-5.537e-04	3.606e-04	-1.536	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:ORIG_RT_wm	1.435e-03	1.481e-03	0.969	
NUM_UNIT_wv:monthly_rGDP	1.437e-01	2.694e-02	5.333	

PROP_TYP_CP_wm 1.30e-07 ***
PROP_TYP_MH_wm 0.0270 *
PROP_TYP_PU_wm 0.0375 *
PURPOSE_C_wm 0.5401
PURPOSE_P_wm 0.7506
sqrt(I_count) 0.2116
ORIG_RT_wm 0.8813
NUM_UNIT_wv 0.0709 .
monthly_rGDP 3.62e-05 ***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:PURPOSE_C_wm 0.4897
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:PURPOSE_C_wm 0.3303
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:PURPOSE_C_wm 0.0108 *
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:PURPOSE_C_wm 0.7114
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:PURPOSE_C_wm 0.1674
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:PURPOSE_C_wm 0.7372
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:PURPOSE_C_wm 0.1813
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:PURPOSE_P_wm 0.8198
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:PURPOSE_P_wm 0.3673
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:PURPOSE_P_wm 0.0019 **
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:PURPOSE_P_wm 0.3502
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:PURPOSE_P_wm 0.3925
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:PURPOSE_P_wm 0.4841
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:PURPOSE_P_wm 0.9524
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:sqrt(I_count) 0.9508
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:sqrt(I_count) 0.4463
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:sqrt(I_count) 0.8445
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:sqrt(I_count) 0.7281
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:sqrt(I_count) 0.4711
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:sqrt(I_count) 0.3539
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:sqrt(I_count) 0.2420
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:ORIG_RT_wm 0.8870
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:ORIG_RT_wm 0.6781
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:ORIG_RT_wm 0.3528
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:ORIG_RT_wm 0.2122
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:ORIG_RT_wm 0.8405
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:ORIG_RT_wm 0.1246
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:ORIG_RT_wm 0.3326
NUM_UNIT_wv:monthly_rGDP 9.66e-08 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

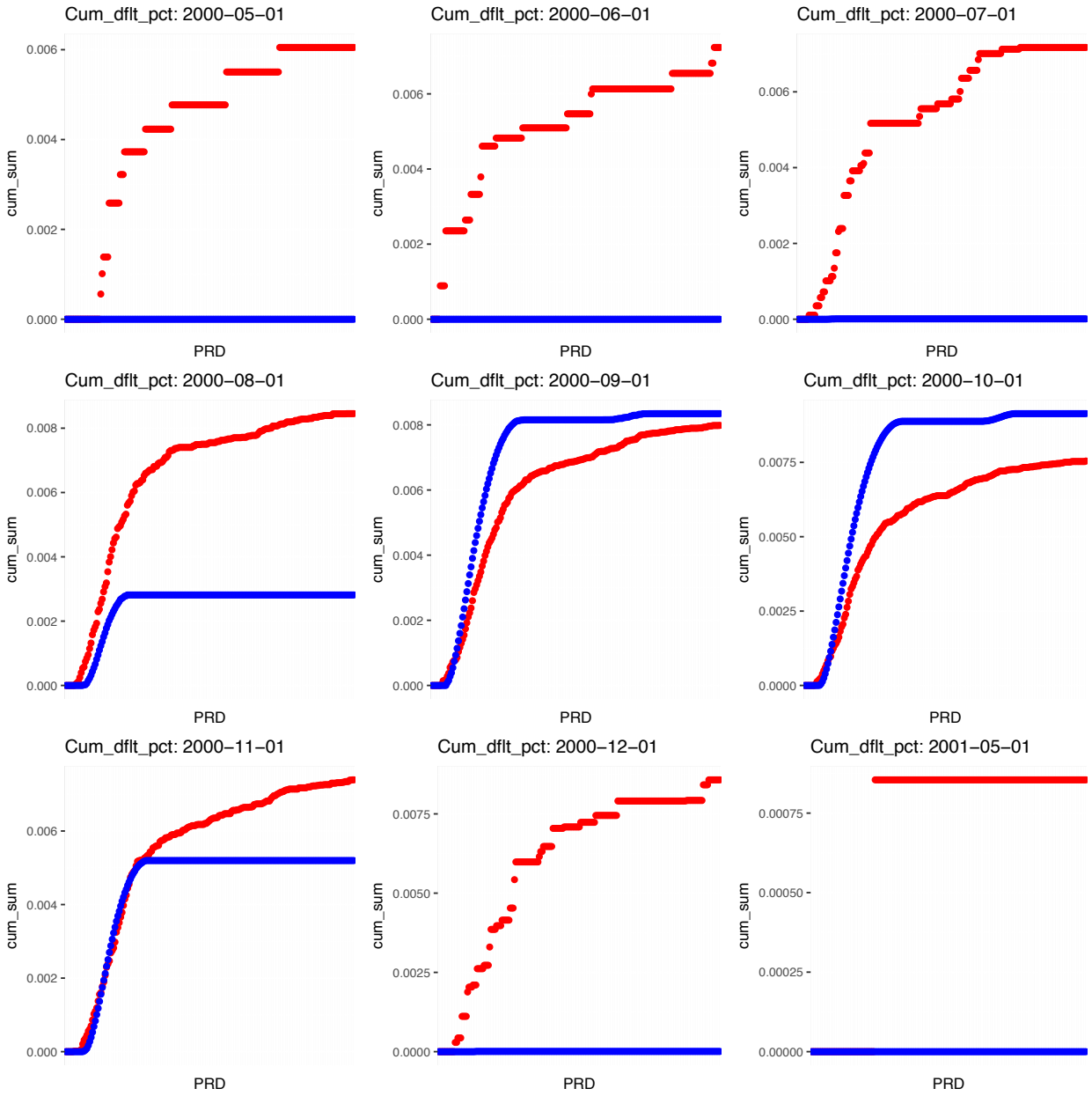
Number of linear predictors: 2

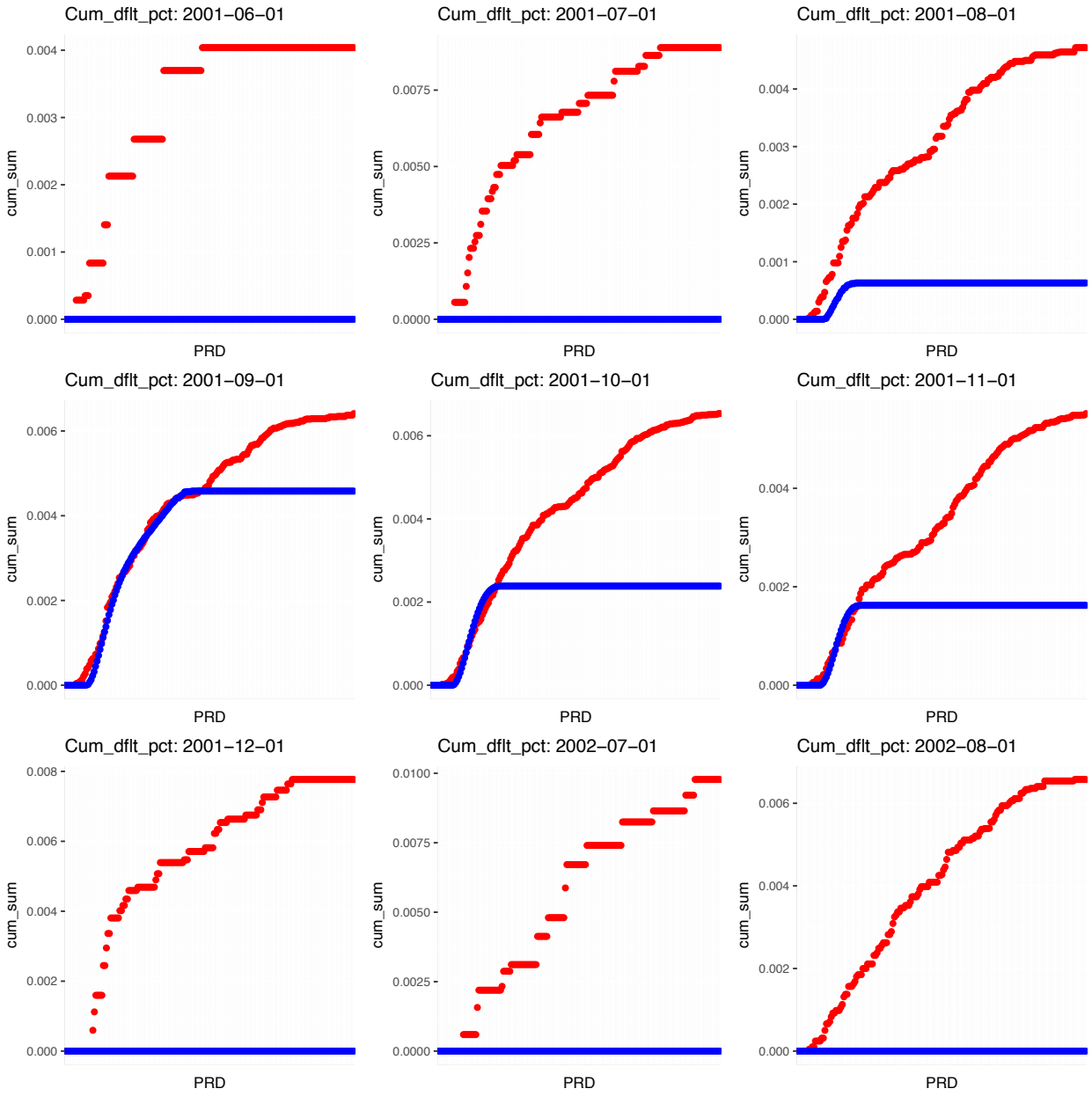
Names of linear predictors: mu, loge(sd)

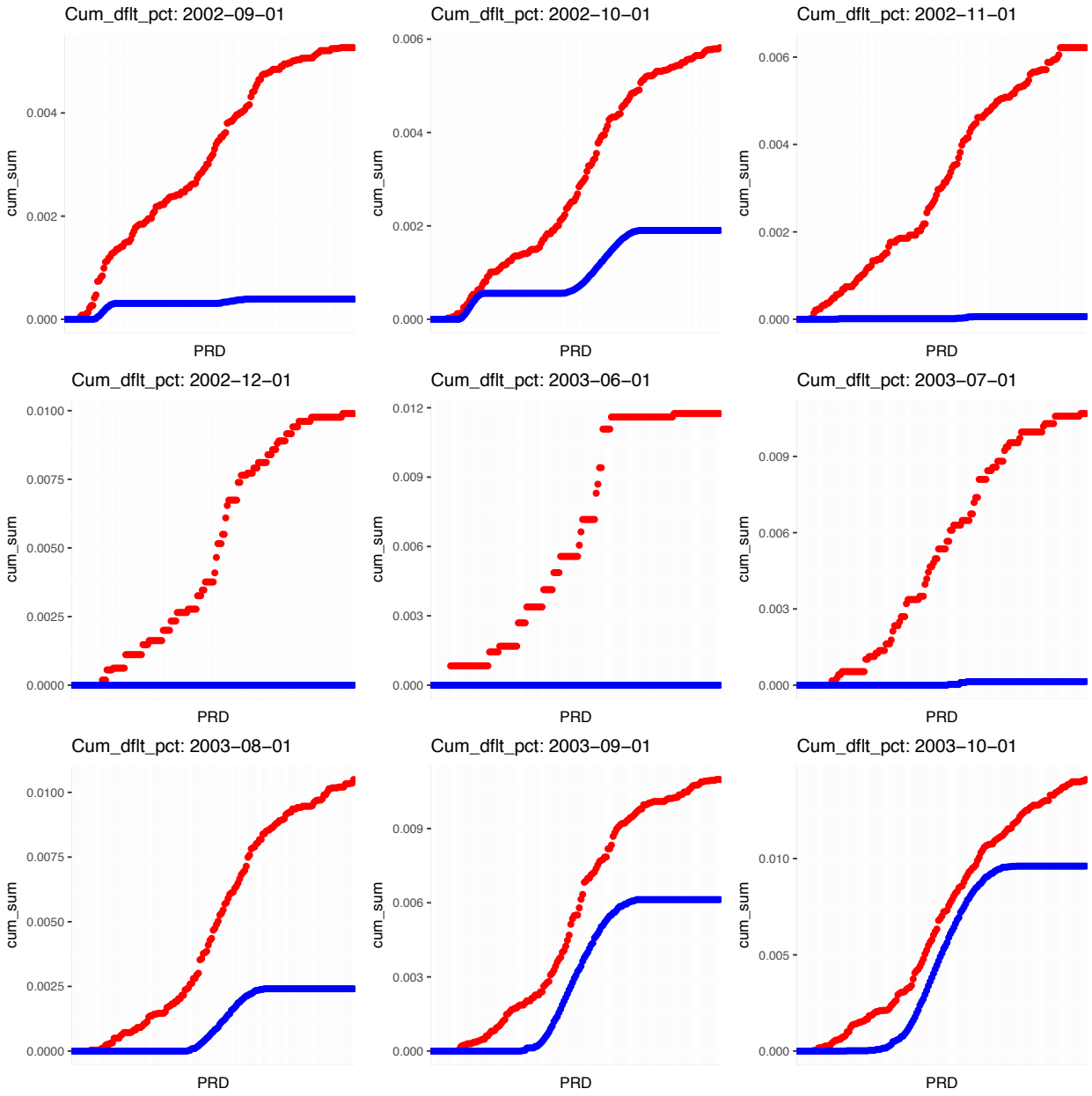
Log-likelihood: 18867.93 on 14052 degrees of freedom

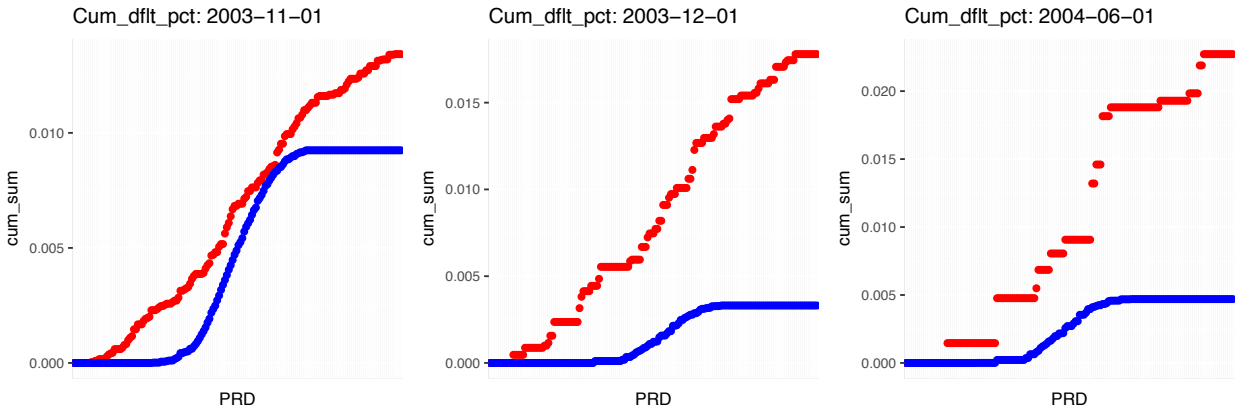
Number of iterations: 13

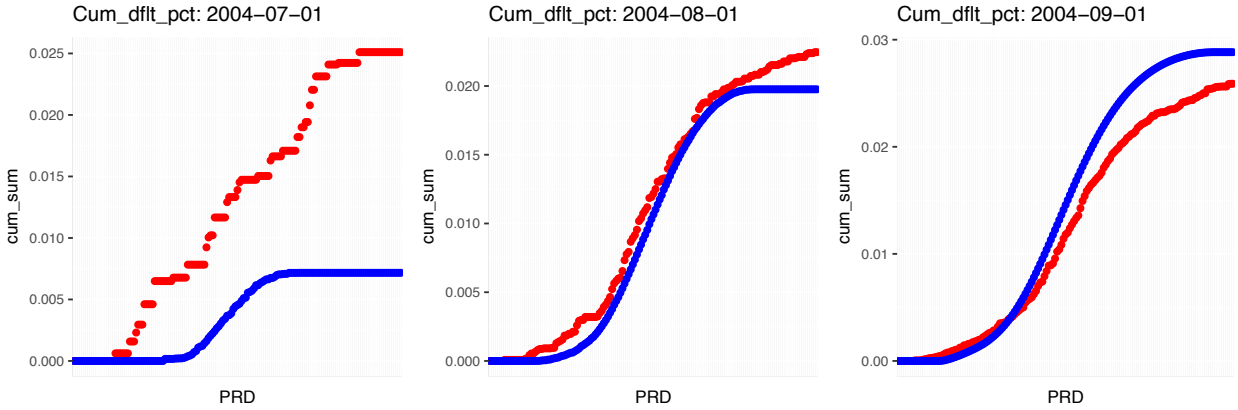
4.1 Fit to training set

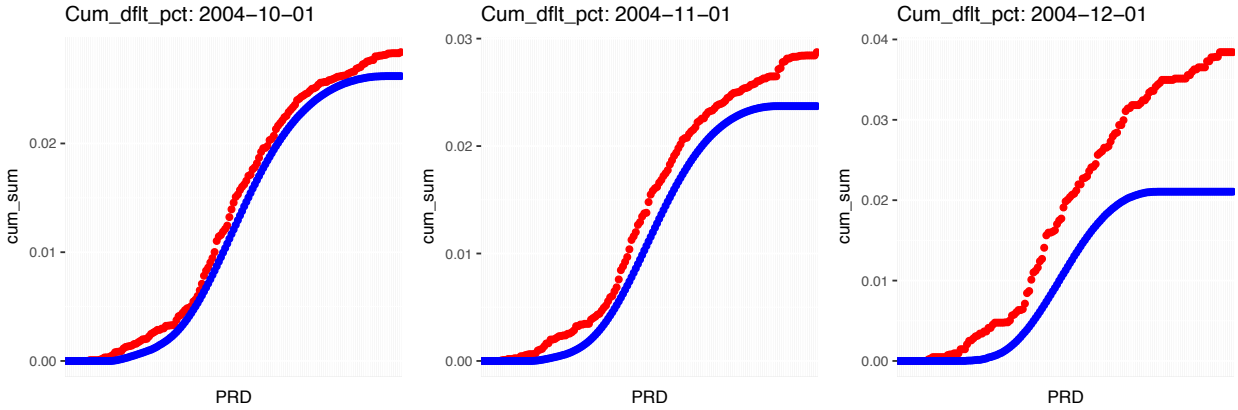


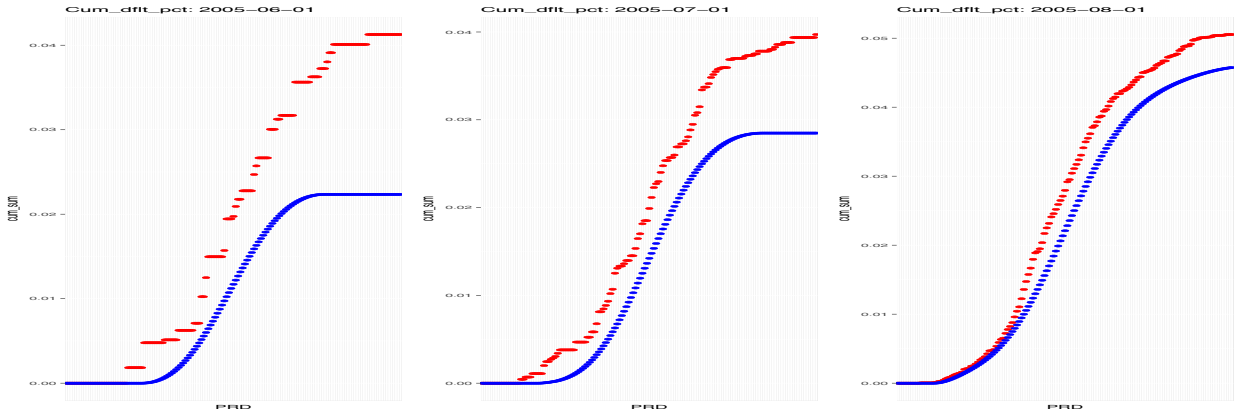

Cum_dflt_pct: 2001−06−01
Cum_dflt_pct: 2001−07−01
Cum_dflt_pct: 2001−08−01
Cum_dflt_pct: 2001−09−01
Cum_dflt_pct: 2001−10−01
Cum_dflt_pct: 2001−11−01
Cum_dflt_pct: 2001−12−01
Cum_dflt_pct: 2002−07−01
Cum_dflt_pct: 2002−08−01
cum_sum
PRD


Cum_dflt_pct: 2002−09−01
Cum_dflt_pct: 2002−10−01
Cum_dflt_pct: 2002−11−01
Cum_dflt_pct: 2002−12−01
Cum_dflt_pct: 2003−06−01
Cum_dflt_pct: 2003−07−01
Cum_dflt_pct: 2003−08−01
Cum_dflt_pct: 2003−09−01
Cum_dflt_pct: 2003−10−01
cum_sum
PRD

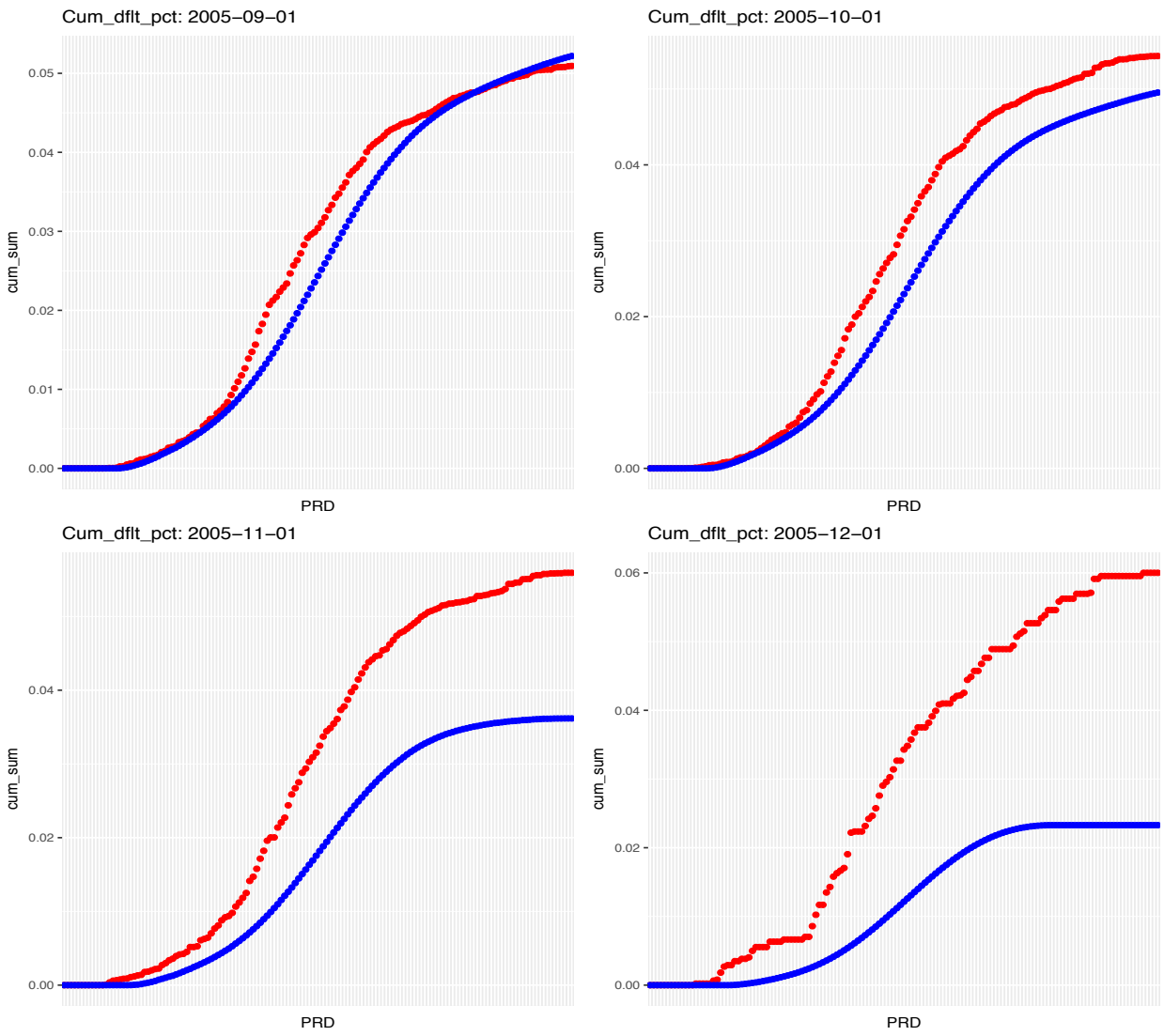

Cum_dflt_pct: 2003−11−01
Cum_dflt_pct: 2003−12−01
Cum_dflt_pct: 2004−06−01
cum_sum
PRD


Cum_dflt_pct: 2004−07−01
Cum_dflt_pct: 2004−08−01
Cum_dflt_pct: 2004−09−01
cum_sum
PRD


Cum_dflt_pct: 2004−10−01
Cum_dflt_pct: 2004−11−01
Cum_dflt_pct: 2004−12−01
cum_sum
PRD


Cum_dflt_pct: 2005−06−01
Cum_dflt_pct: 2005−07−01
Cum_dflt_pct: 2005−08−01
cum_sum
PRD

4.2 Fit to test set



4.3 Comments

- The fit to the training data are poor. In particular, many projections are flat at zero, indicating the model's lack of ability to capture structure in the data.
- The out-of sample fit is reasonable.

5 Two stage modeling

Data: Quarter 4 data from years 2000 to 2005, inclusive.

We fit a model that is done in 2 stages. The model is:

$$E(\log \mathrm{Y} \mid X) = P(Y > 0 \mid X) * E(\log Y \mid Y > 0, X)$$

We fit 2 separate models for each term on the right side of the equation. Below we will denote $P(Y > 0 \mid X)$ as stage 1 and $E(\log Y \mid Y > 0, X)$ as stage 2. Note on a minor detail: because we have logged Y, we must multiply our final prediction by a term that accounts for the exponential transformation.

5.1 Stage 1: Logistic regression

The model first identifies months when loans default using a binomial model with a logit link function. Using best forward subset selection based on AIC to select the variables, the model is:

```
Call:
vglm(formula = "did_dflt ~ bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5]) +
sqrt(I_count)*LIBOR1+
bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5])*(PURPOSE_C_wm+PURPOSE_P_wm)+
bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5])*ORIG_AMT_sum+
bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5])*ORIG_RT_wm+
OCLTV_wv*monthly_rGDP",
    family = binomialff(link = logit), data = train, trace = T)


Pearson residuals:
                Min      1Q  Median     3Q   Max
logit(prob) -4.791 -0.4319 -0.1179 0.4929 10.73

Coefficients:
                                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                   -1.237e+00  1.616e+01  -0.077  0.93895
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1   1.957e-01  2.270e+01   0.009  0.99312
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2  -1.146e+01  1.495e+01  -0.766  0.44339
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3   1.548e+01  1.789e+01   0.866  0.38662
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4   1.190e+01  1.598e+01   0.745  0.45652
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5   5.341e-01  1.821e+01   0.029  0.97660
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6   2.936e+01  2.079e+01   1.413  0.15779
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7  -5.756e+01  5.763e+01      NA       NA
sqrt(I_count)                                                  4.586e-02  3.472e-03  13.209  < 2e-16 ***
LIBOR1                                                         2.989e-01  1.740e-01   1.718  0.08587 .
PURPOSE_C_wm                                                  -5.595e+00  1.768e+01  -0.316  0.75166
PURPOSE_P_wm                                                  -9.620e-01  5.631e+00  -0.171  0.86435
ORIG_AMT_sum                                                  -1.401e-09  3.255e-10  -4.303 1.69e-05 ***
ORIG_RT_wm                                                    -2.374e-01  1.830e+00  -0.130  0.89678
OCLTV_wv                                                      -1.520e-02  2.951e-03  -5.152 2.58e-07 ***
monthly_rGDP                                                  -1.321e+01  5.593e+01  -0.236  0.81325
sqrt(I_count):LIBOR1                                          -1.538e-03  1.110e-03  -1.385  0.16597
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:PURPOSE_C_wm  8.940e+00  2.510e+01   0.356  0.72169
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:PURPOSE_C_wm  1.723e+01  1.579e+01   1.091  0.27535
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:PURPOSE_C_wm  1.586e+01  1.961e+01   0.809  0.41858
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:PURPOSE_C_wm  9.855e+00  1.745e+01   0.565  0.57215
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:PURPOSE_C_wm  1.191e+01  2.011e+01   0.592  0.55371
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:PURPOSE_C_wm -1.923e+01  2.377e+01  -0.809  0.41845
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:PURPOSE_C_wm  5.281e+01  4.971e+01   1.062  0.28814
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:PURPOSE_P_wm  1.112e+00  8.724e+00   0.127  0.89858
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:PURPOSE_P_wm  4.763e+00  4.906e+00   0.971  0.33158
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:PURPOSE_P_wm  1.348e+01  6.226e+00   2.166  0.03033 *
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:PURPOSE_P_wm  2.185e+00  5.688e+00   0.384  0.70082
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:PURPOSE_P_wm  5.158e+00  6.518e+00   0.791  0.42875
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:PURPOSE_P_wm  1.136e+00  8.577e+00   0.132  0.89467
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:PURPOSE_P_wm -2.950e+00  3.187e+01  -0.093  0.92624
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:ORIG_AMT_sum  1.077e-09  5.010e-10   2.150  0.03153 *
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:ORIG_AMT_sum  9.243e-10  2.856e-10   3.237  0.00121 **
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:ORIG_AMT_sum  7.391e-10  3.617e-10   2.043  0.04104 *
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:ORIG_AMT_sum  7.927e-10  3.294e-10   2.406  0.01611 *
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:ORIG_AMT_sum  5.394e-10  3.656e-10   1.475  0.14008
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:ORIG_AMT_sum  9.849e-10  3.774e-10   2.610  0.00907 **
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:ORIG_AMT_sum  1.898e-10  4.113e-10   0.461  0.64449
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:ORIG_RT_wm    5.365e-01  2.620e+00   0.205  0.83777
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:ORIG_RT_wm    1.476e+00  1.718e+00   0.859  0.39015
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:ORIG_RT_wm   -3.070e+00  2.034e+00  -1.509  0.13125
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:ORIG_RT_wm   -1.538e+00  1.851e+00  -0.831  0.40624
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:ORIG_RT_wm   -1.368e-01  2.080e+00  -0.066  0.94758
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:ORIG_RT_wm   -3.414e+00  2.437e+00  -1.401  0.16128
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:ORIG_RT_wm    7.053e+00  8.674e+00   0.813  0.41611
OCLTV_wv:monthly_rGDP                                          1.000e-01  2.053e-01   0.487  0.62608
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors: 1

Name of linear predictor: logit(prob)

Residual deviance: 5822.891 on 7004 degrees of freedom

Log-likelihood: -2911.445 on 7004 degrees of freedom

Number of iterations: 6

Warning: Hauck-Donner effect detected in the following estimate(s):
'bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7'2.

5.2 Test error

Test error:

	0	1
0	0.8253968	0.1746032
1	0.1045752	0.8954248

Area under the curve (AUC): 0.860

5.3 Stage 2: Logistic regression

Next, we fit a logistic regression to months that had at least one default (ie. Default ratio > 0). We make our projection by applying the two stages of models to the test set. Using best forward subset selection based on AIC to select the variables, the model is:

Call:
vglm(formula = "dflt_pct ~ bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5]) +
(PROP_TYP_CO_wm+PROP_TYP_CP_wm+PROP_TYP_MH_wm+PROP_TYP_PU_wm)*unemp+
sqrt(I_count)*LIBOR1+
bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5])*ORIG_AMT_sum+
bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5])*(PURPOSE_C_wm+PURPOSE_P_wm)+
(ORIG_CHN_B_wm+ORIG_CHN_C_wm)*unemp+
(OCC_STAT_I_wm+OCC_STAT_P_wm)*LIBOR1",
family = logistic(), data = train.positive, trace = T)

Pearson residuals:

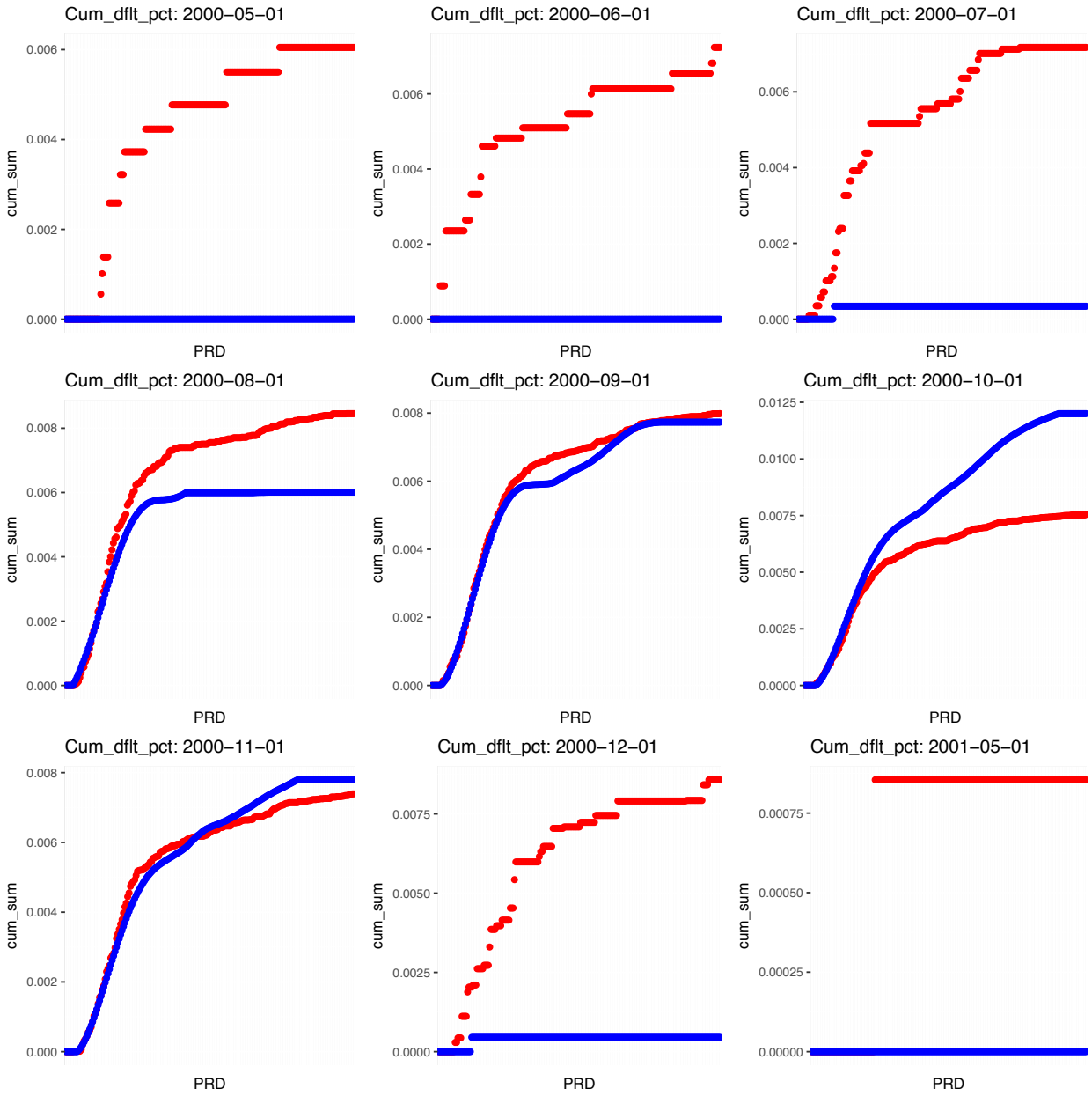
	Min	1Q	Median	3Q	Max
location	-1.7314	-0.4970	-0.06027	0.4692	1.733
loge(scale)	-0.8363	-0.8109	-0.70353	-0.1795	41.672

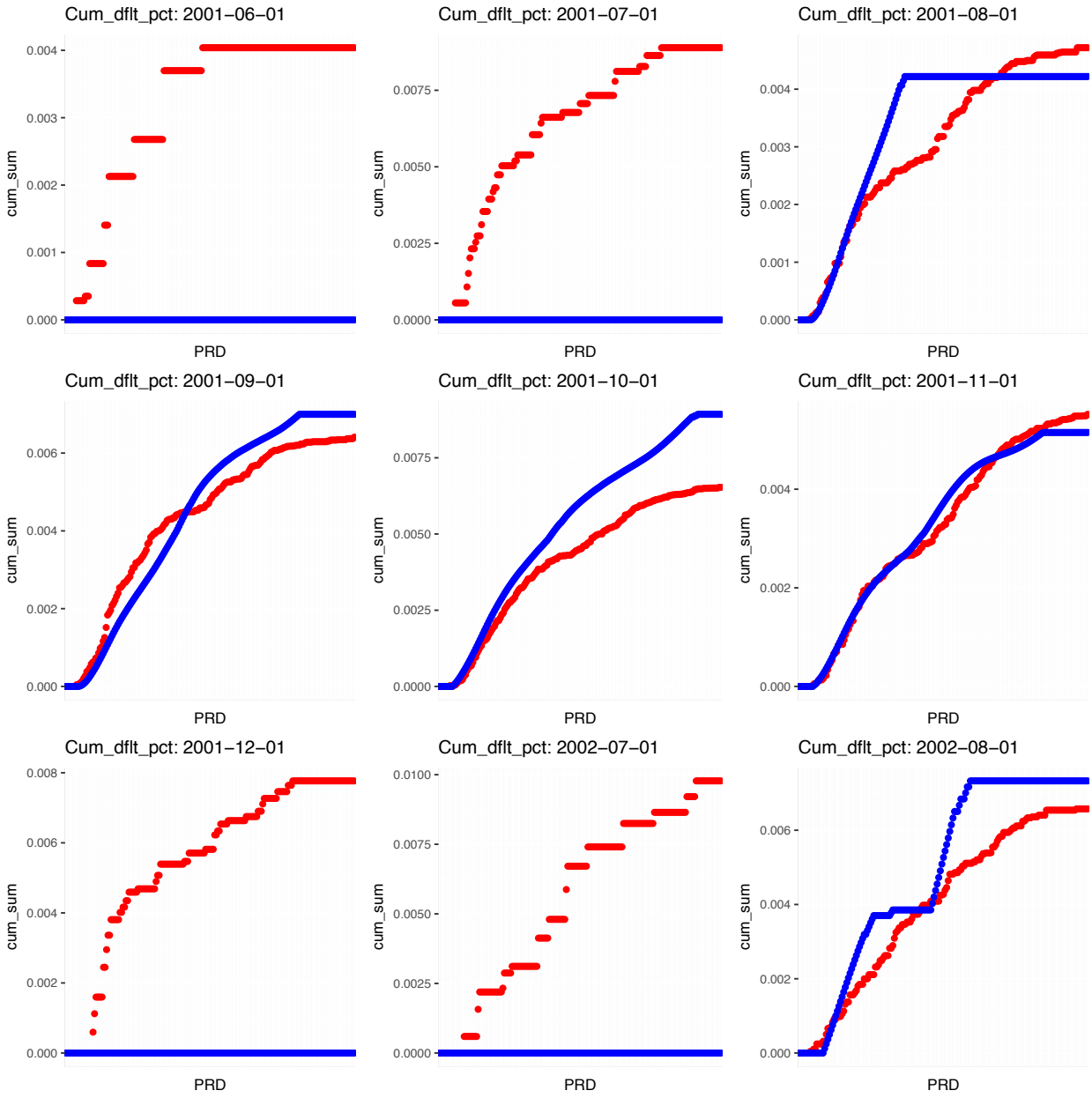
Coefficients:

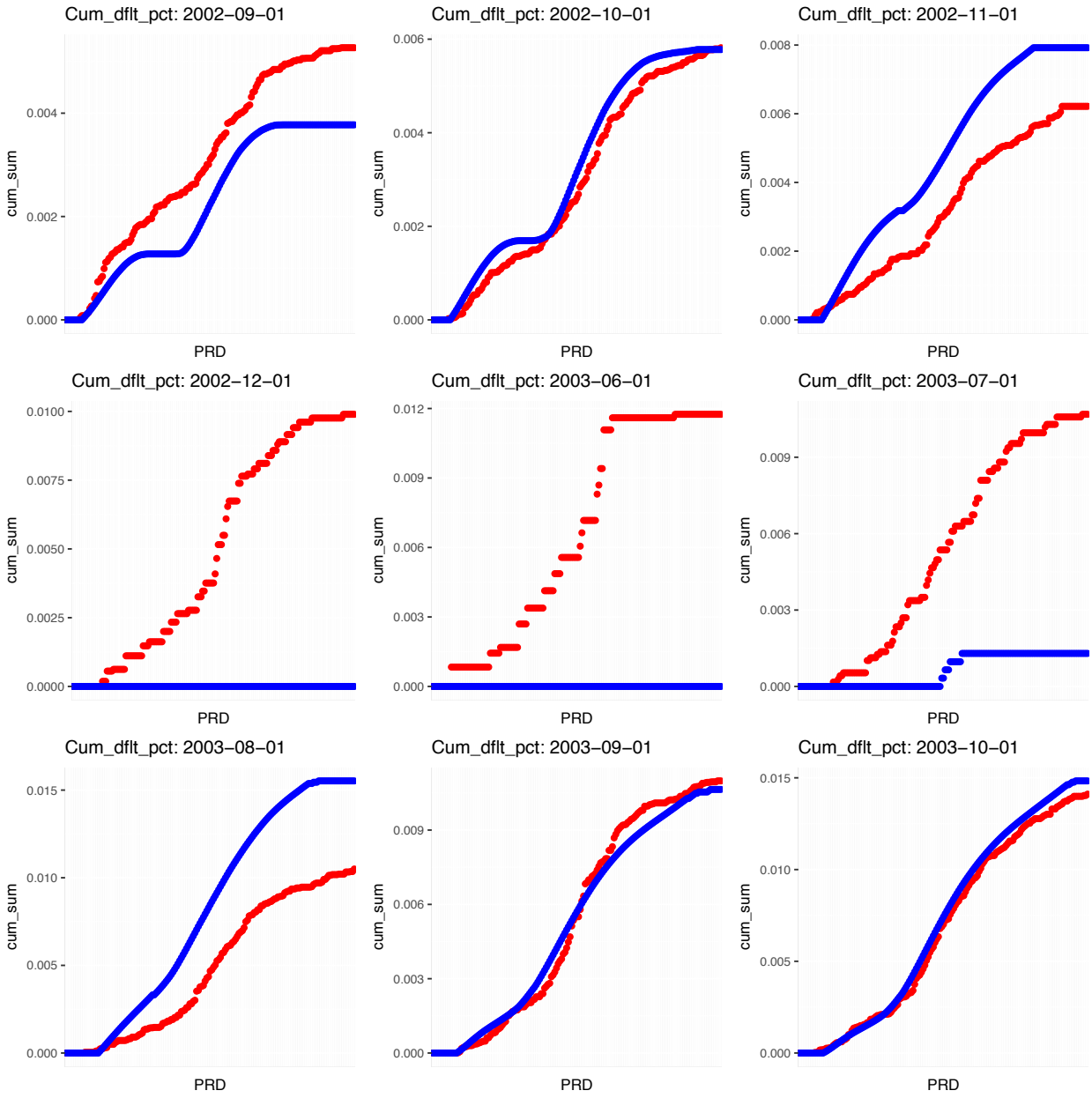
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept):1	-3.635e-03	9.373e-04	-3.878	0.000105	***
(Intercept):2	-9.572e+00	1.508e-02	-634.909	< 2e-16	***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1	-6.926e-04	4.693e-04	-1.476	0.139965	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2	1.080e-04	2.716e-04	0.398	0.690882	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3	-1.832e-03	3.386e-04	-5.412	6.25e-08	***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4	-6.509e-04	3.092e-04	-2.105	0.035271	*
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5	-7.895e-04	3.790e-04	-2.083	0.037235	*
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6	-5.015e-04	5.151e-04	-0.974	0.330256	
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7	-7.268e-04	1.292e-03	-0.563	0.573770	
PROP_TYP_CO_wm	2.313e-03	6.398e-04	3.615	0.000301	***
PROP_TYP_CP_wm	1.677e-02	1.539e-03	10.897	< 2e-16	***
PROP_TYP_MH_wm	1.304e-02	3.421e-03	3.812	0.000138	***

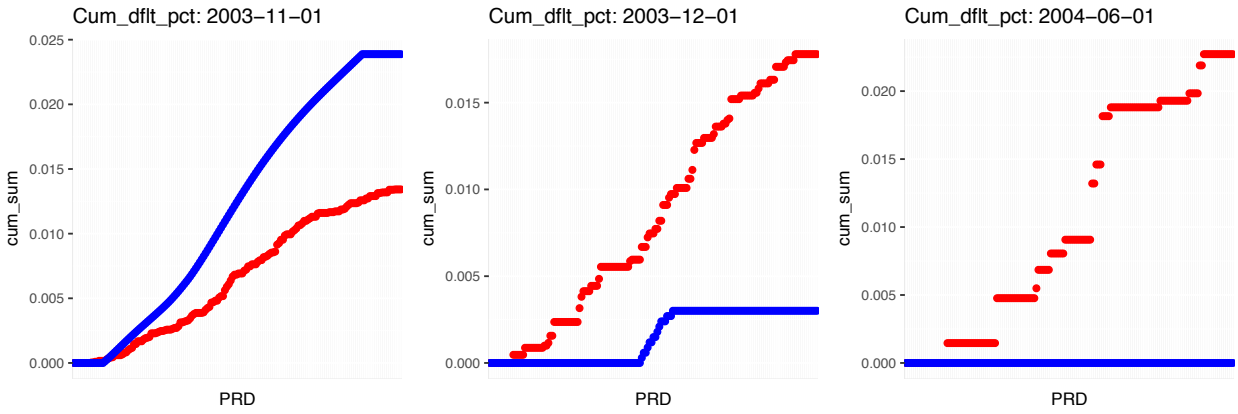
```
PROP_TYP_PU_wm                                          1.956e-03  2.423e-04    8.073 6.88e-16 ***
unemp                                                3.394e-05  3.761e-05    0.902 0.366792
sqrt(I_count)                                       -5.877e-06  2.893e-07  -20.312  < 2e-16 ***
LIBOR1                                               6.823e-04  6.013e-04    1.135 0.256473
ORIG_AMT_sum                                            1.197e-13  1.470e-14    8.143 3.87e-16 ***
PURPOSE_C_wm                                           -1.127e-03  6.145e-04   -1.833 0.066741 .
PURPOSE_P_wm                                           -2.806e-04  2.684e-04   -1.045 0.295795
ORIG_CHN_B_wm                                           -3.764e-04  1.160e-04   -3.245 0.001177 **
ORIG_CHN_C_wm                                           -4.044e-04  3.661e-05  -11.044  < 2e-16 ***
OCC_STAT_I_wm                                           2.924e-03  1.263e-03    2.315 0.020619 *
OCC_STAT_P_wm                                           4.532e-03  8.733e-04    5.189 2.11e-07 ***
PROP_TYP_CO_wm:unemp                                       -5.560e-04  4.797e-04   -1.159 0.246398
PROP_TYP_CP_wm:unemp                                        1.345e-03  1.292e-03    1.042 0.297560
PROP_TYP_MH_wm:unemp                                       -8.634e-03  3.194e-03   -2.703 0.006866 **
PROP_TYP_PU_wm:unemp                                        1.165e-03  1.631e-04    7.147 8.86e-13 ***
sqrt(I_count):LIBOR1                                   -2.085e-07  8.546e-08   -2.439 0.014715 *
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:ORIG_AMT_sum  4.650e-15  2.147e-14    0.217 0.828497
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:ORIG_AMT_sum  1.116e-15  1.249e-14    0.089 0.928792
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:ORIG_AMT_sum  1.543e-14  1.636e-14    0.943 0.345832
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:ORIG_AMT_sum -4.097e-15  1.446e-14   -0.283 0.776937
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:ORIG_AMT_sum  9.871e-15  1.782e-14    0.554 0.579575
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:ORIG_AMT_sum -1.708e-14  2.005e-14   -0.852 0.394346
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:ORIG_AMT_sum  2.054e-14  2.008e-14    1.023 0.306330
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:PURPOSE_C_wm  1.682e-03  9.689e-04    1.736 0.082596 .
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:PURPOSE_C_wm -2.325e-04  5.503e-04   -0.423 0.672638
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:PURPOSE_C_wm  4.365e-03  6.951e-04    6.279 3.41e-10 ***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:PURPOSE_C_wm  1.665e-03  6.273e-04    2.655 0.007942 **
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:PURPOSE_C_wm  1.589e-03  7.873e-04    2.018 0.043581 *
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:PURPOSE_C_wm  1.612e-03  1.089e-03    1.480 0.138788
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:PURPOSE_C_wm  1.335e-03  2.889e-03    0.462 0.644137
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1:PURPOSE_P_wm  7.828e-04  4.216e-04    1.857 0.063315 .
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2:PURPOSE_P_wm  1.038e-04  2.549e-04    0.407 0.683733
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3:PURPOSE_P_wm  1.457e-03  3.036e-04    4.798 1.60e-06 ***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4:PURPOSE_P_wm  6.566e-04  2.765e-04    2.375 0.017562 *
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5:PURPOSE_P_wm  7.819e-04  3.416e-04    2.289 0.022087 *
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6:PURPOSE_P_wm  4.058e-04  4.698e-04    0.864 0.387711
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7:PURPOSE_P_wm  5.480e-04  1.228e-03    0.446 0.655392
unemp:ORIG_CHN_B_wm                                        -3.850e-04  8.393e-05   -4.587 4.49e-06 ***
unemp:ORIG_CHN_C_wm                                         1.078e-05  2.998e-05    0.360 0.719196
LIBOR1:OCC_STAT_I_wm                                       -1.700e-03  8.364e-04   -2.033 0.042079 *
LIBOR1:OCC_STAT_P_wm                                        -6.312e-04  6.121e-04   -1.031 0.302445
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  2

Names of linear predictors: location, loge(scale)

Log-likelihood: 22962.82 on 6101 degrees of freedom

Number of iterations: 18

No Hauck-Donner effect found in any of the estimates
```

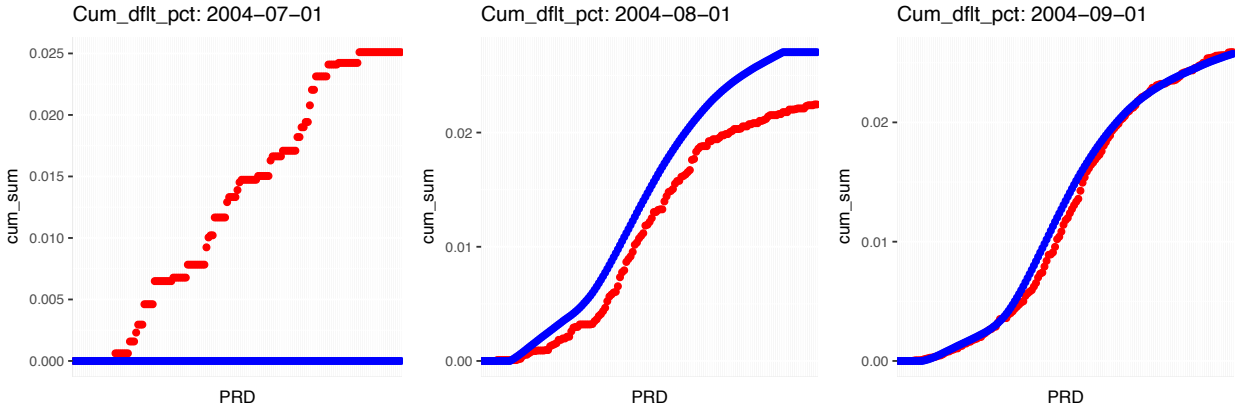
5.4 Fit to training set

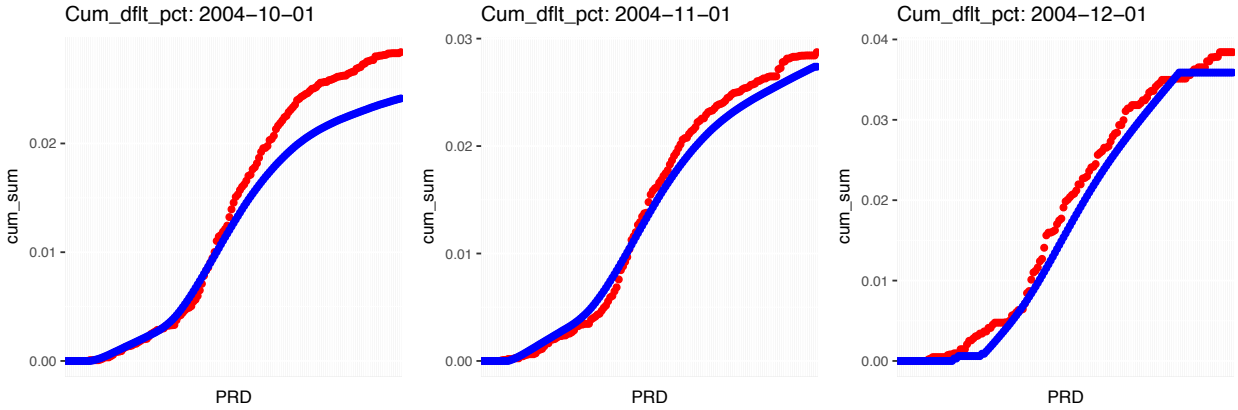


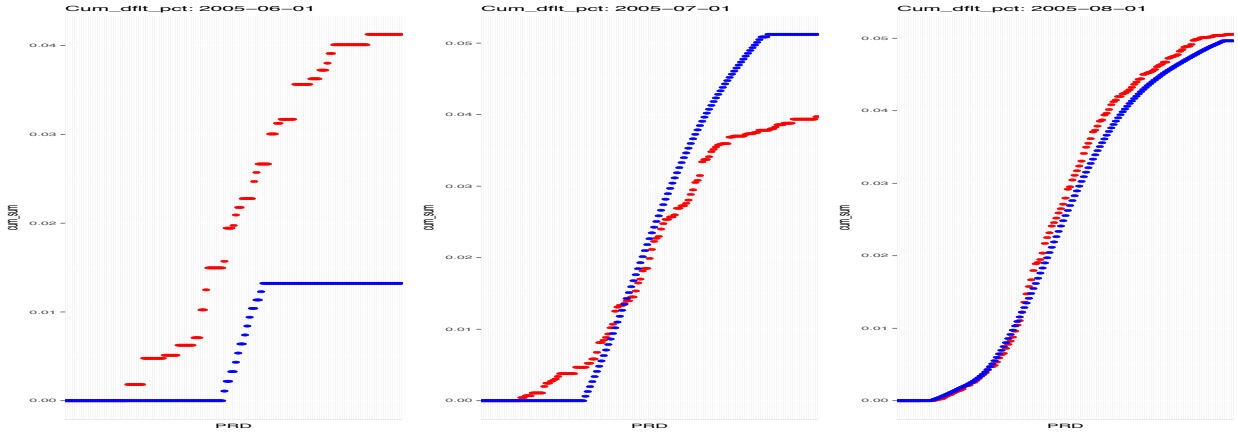

Cum_dflt_pct: 2001−06−01
Cum_dflt_pct: 2001−07−01
Cum_dflt_pct: 2001−08−01
Cum_dflt_pct: 2001−09−01
Cum_dflt_pct: 2001−10−01
Cum_dflt_pct: 2001−11−01
Cum_dflt_pct: 2001−12−01
Cum_dflt_pct: 2002−07−01
Cum_dflt_pct: 2002−08−01
cum_sum
PRD


Cum_dflt_pct: 2002−09−01
Cum_dflt_pct: 2002−10−01
Cum_dflt_pct: 2002−11−01
Cum_dflt_pct: 2002−12−01
Cum_dflt_pct: 2003−06−01
Cum_dflt_pct: 2003−07−01
Cum_dflt_pct: 2003−08−01
Cum_dflt_pct: 2003−09−01
Cum_dflt_pct: 2003−10−01
cum_sum
PRD

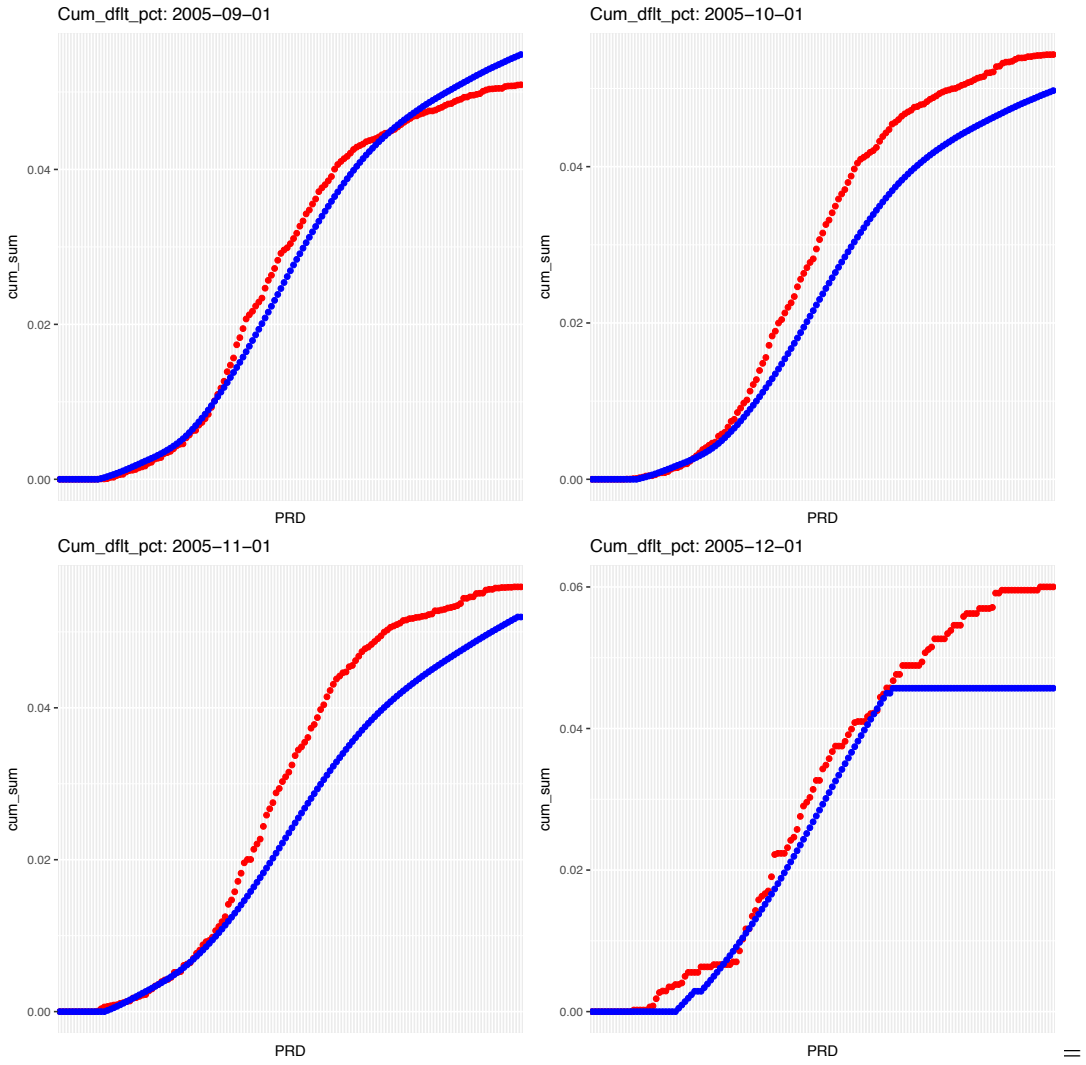

Cum_dflt_pct: 2003−11−01
Cum_dflt_pct: 2003−12−01
Cum_dflt_pct: 2004−06−01
cum_sum
PRD


Cum_dflt_pct: 2004−07−01
Cum_dflt_pct: 2004−08−01
Cum_dflt_pct: 2004−09−01
cum_sum
PRD


Cum_dflt_pct: 2004−10−01
Cum_dflt_pct: 2004−11−01
Cum_dflt_pct: 2004−12−01
cum_sum
PRD


Cum_dflt_pct: 2005−06−01
Cum_dflt_pct: 2005−07−01
Cum_dflt_pct: 2005−08−01
cum_sum
PRD

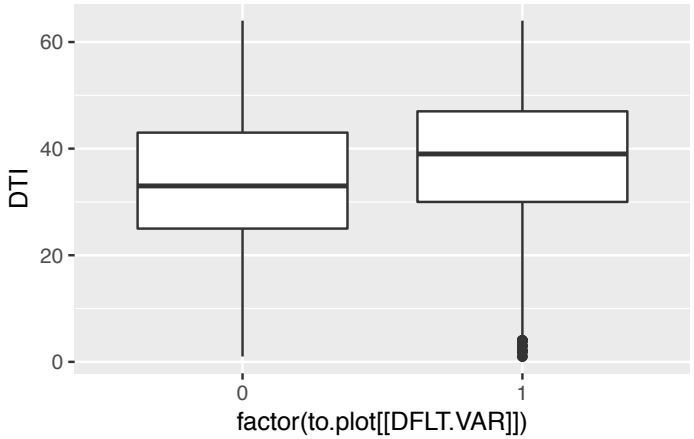
5.5 Fit to test set

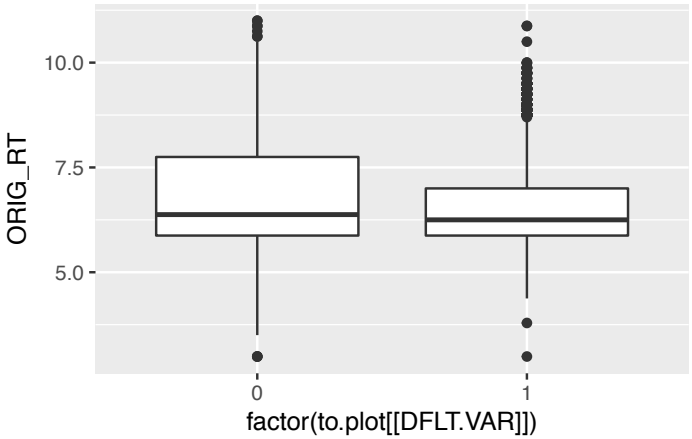


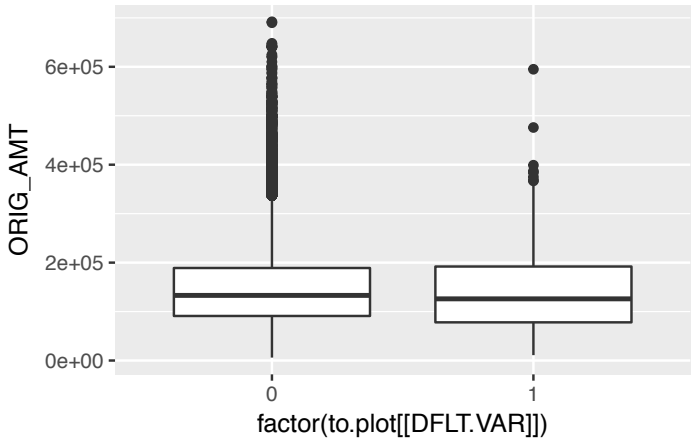
6 Specifying economically meaningful variables

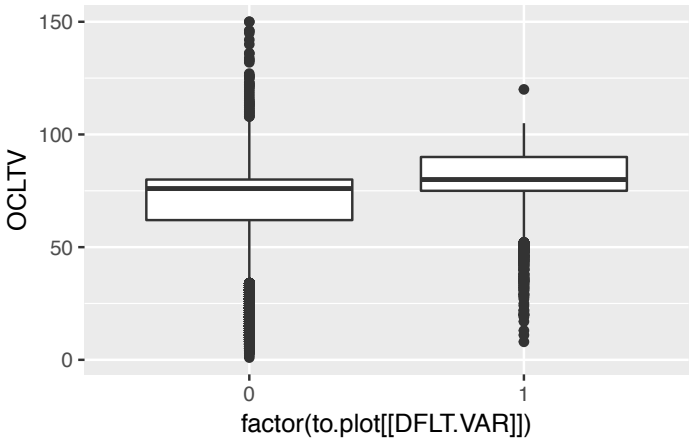
Previously, we chose those variables that gave the greatest reduction in AIC. In this section, we attempt to choose variables based on economic rationale. As a first step, we look at the relationships between the available covariates and whether a loan defaults. The covariates are either continuous or categorical, which we analyze separately. In all results shown below, the latter is a binary variable that is 1 if a loan defaults within 9 years since its origination and 0 otherwise. The definition of each covariate is available on the fannie mae website (http://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html; Glossary and File Layout under 'Related Links' on the right)

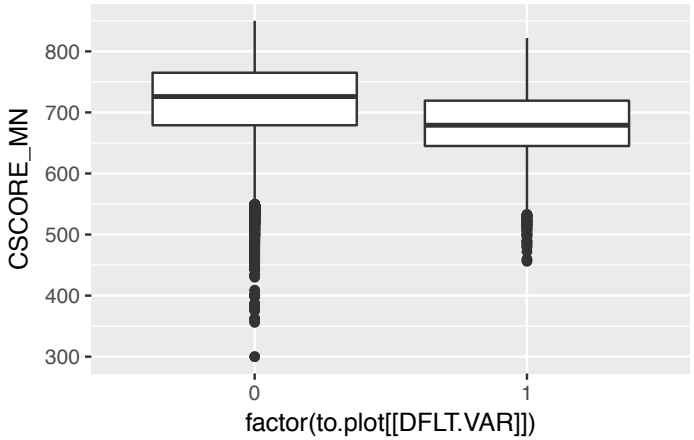
6.1 Continuous

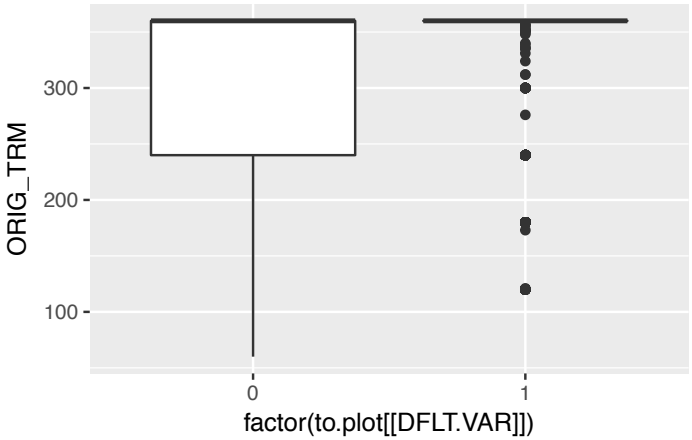


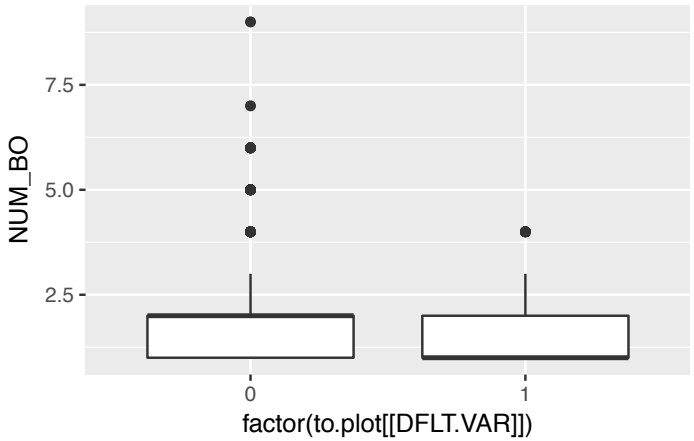


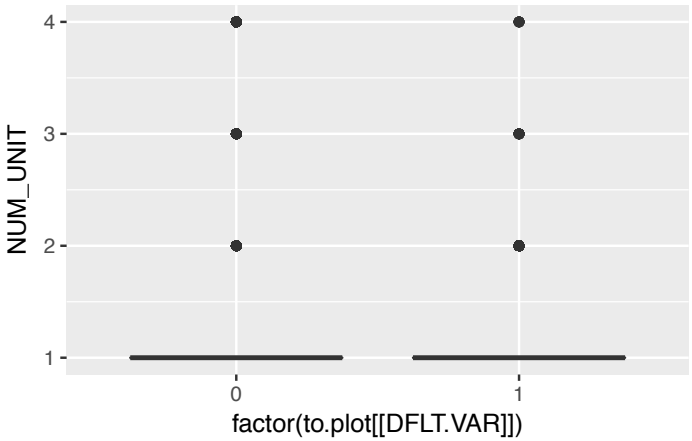












Regression on Y:
Df Sum Sq Mean Sq F value Pr(>F)
DTI 1 16 16.284 1092 <2e-16 ***
Residuals 564604 8416 0.015 --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 27391 observations deleted due to missingness

Df Sum Sq Mean Sq F value Pr(>F)
ORIG_RT 1 2 1.780 119 <2e-16 ***
Residuals 591995 8857 0.015 --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Df Sum Sq Mean Sq F value Pr(>F)
ORIG_AMT 1 0 0.29016 19.39 1.06e-05 ***
Residuals 591995 8858 0.01496 --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Df Sum Sq Mean Sq F value Pr(>F)
OCLTV 1 36 36.29 2435 <2e-16 ***
Residuals 591994 8822 0.01 --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness

Df Sum Sq Mean Sq F value Pr(>F)
CSCORE_MN 1 55 54.54 3684 <2e-16 ***
Residuals 587629 8699 0.01 --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
4366 observations deleted due to missingness

Df Sum Sq Mean Sq F value Pr(>F)
ORIG_TRM 1 21 20.716 1388 <2e-16 ***
Residuals 591995 8838 0.015 --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Df Sum Sq Mean Sq F value Pr(>F)
NUM_BO 1 22 22.338 1496 <2e-16 ***
Residuals 590662 8822 0.015 --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1333 observations deleted due to missingness

Df Sum Sq Mean Sq F value Pr(>F)
NUM_UNIT 1 0 0.03151 2.106 0.147
Residuals 591995 8858 0.01496

6.2 Categorical

Categorical tables and Chi-squared tests:

[1] "ORIG_CHN"
[1] "Table"

	B	C	R
0	112857	196601	273544
1	2341	3355	3299

	B	C	R
0	0.1935791	0.3372218	0.4691991
1	0.2602557	0.3729850	0.3667593

[1] "Statistic and P-value"

```
   X-squared
4.347071e+02 4.023014e-95
 X-squared
0.02709809

[1] "FTHB_FLG"
[1] "Table"

       N     U     Y
 0 527664  1652 53686
 1   8042    36   917

            N           U           Y
 0 0.905080943 0.002833609 0.092085447
 1 0.894052251 0.004002223 0.101945525
[1] "Statistic and P-value"
   X-squared
1.477053e+01 6.203254e-04
  X-squared
0.004995033

[1] "PURPOSE"
[1] "Table"

       C      P     R     U
 0 177573 239396 165744   289
 1   3427   3496   2068     4

             C            P            R            U
 0 0.3045838608 0.4106263786 0.2842940504 0.0004957101
 1 0.3809894386 0.3886603669 0.2299055031 0.0004446915
[1] "Statistic and P-value"
   X-squared
2.720433e+02 1.115306e-58
 X-squared
0.02143677

[1] "PROP_TYP"
[1] "Table"

      CO     CP     MH     PU     SF
 0  40784   2661   4390  53881 481286
 1    649     14    203    975   7154

            CO           CP           MH           PU           SF
 0 0.069955163 0.004564307 0.007529991 0.092419923 0.825530616
 1 0.072151195 0.001556420 0.022568093 0.108393552 0.795330739
[1] "Statistic and P-value"
   X-squared
3.107341e+02 5.237208e-66
 X-squared
0.02291051

[1] "OCC_STAT"
[1] "Table"

       I      P      S
 0  30959 531470  20573
 1   1004   7672    319

            I          P          S
 0 0.05310273 0.91160922 0.03528804
```

```
  1 0.11161757 0.85291829 0.03546415
[1] "Statistic and P-value"
    X-squared
 5.952790e+02 5.455137e-130
  X-squared
0.03171031

[1] "RELOCATION_FLG"
[1] "Table"

        N      Y
  0 579836   3166
  1   8984     11

            N          Y
  0 0.994569487 0.005430513
  1 0.998777098 0.001222902
[1] "Statistic and P-value"
   X-squared
2.859772e+01 8.908690e-08
  X-squared
0.006950339
```

6.3 Variables

Based on the above analysis, the reasonable choices are:
For continuous:

```
cont.vars <- c(
   "CSCORE_MN_var", "CSCORE_MN_wm", "CSCORE_MN_wv", "CSCORE_MN_vm",
   "DTI_var", "DTI_wm", "DTI_wv",
   "sqrt(I_count)",
   "OCLTV_var", "OCLTV_wm", "OCLTV_wv")
```

For categorical:

```
cat.vars <- c(
  "ORIG_CHN_B_wm+ORIG_CHN_C_wm",
  "OCC_STAT_I_wm+OCC_STAT_P_wm",
  "PROP_TYP_CO_wm+PROP_TYP_CP_wm+PROP_TYP_MH_wm+PROP_TYP_PU_wm",
  "PURPOSE_C_wm+PURPOSE_P_wm"
)
```

Thus, our initial choice of covariates and interaction terms for the linear model is:

```
"did_dflt ~ bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5]) +
  CSCORE_MN_wm + DTI_wm + DTI_wv + sqrt(I_count) +
  LIBOR1*ORIG_RT_wm + OCLTV_wv*unemp +
  (PROP_TYP_CO_wm+PROP_TYP_CP_wm+PROP_TYP_MH_wm+PROP_TYP_PU_wm) +
  (OCC_STAT_I_wm+OCC_STAT_P_wm)*unemp"
```

- OCC_STAT proxy for wealth. Poor is hardest hit
- In low interest rate environment, borrower loses value

6.4 Ideas for other potential variables not included

- Bucketized CSCORE, ORIG_AMT etc
- Mortgage Premium = (Current Interest Rate - Market Mortgage Rate)/Current Interest Rate
- Loan Size = Original un-payment balance (UPB) / Average UPB in the same state and same date
- Market Loan to Value (MLTV) = (Current UPB/HPI Factor) * Original Housing Value
- HPI Factor = Current HPI / Original HPI (House price index)

6.5 Results

vglm(formula = "did_dflt ~ bs(AGE, knots=quantile(AGE, probs=seq(0, 1, 0.2))[2:5]) + \n CSCORE_MN_wm +
OCLTV_wv*unemp + DTI_wm + DTI_wv +\n + sqrt(I_count) + LIBOR1*ORIG_RT_wm +\n
(PROP_TYP_CO_wm+PROP_TYP_CP_wm+PROP_TYP_MH_wm+PROP_TYP_PU_wm) +\n
(OCC_STAT_I_wm+OCC_STAT_P_wm)*unemp",
    family = binomialff(link = logit), data = train, trace = T)

Pearson residuals:

	Min	1Q	Median	3Q	Max
logit(prob)	-4.577	-0.4805	-0.1182	0.5236	16.23

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	5.432e+01	1.517e+01	3.581	0.000342	***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])1	9.324e+00	7.749e-01	12.032	< 2e-16	***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])2	8.282e+00	4.954e-01	16.719	< 2e-16	***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])3	7.020e+00	6.007e-01	11.686	< 2e-16	***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])4	7.518e+00	5.421e-01	13.868	< 2e-16	***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])5	6.300e+00	6.140e-01	10.260	< 2e-16	***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])6	4.266e+00	6.682e-01	6.385	1.72e-10	***
bs(AGE, knots = quantile(AGE, probs = seq(0, 1, 0.2))[2:5])7	4.754e+00	8.010e-01	5.936	2.92e-09	***
CSCORE_MN_wm	-4.770e-03	1.579e-02	-0.302	0.762632	
OCLTV_wv	-2.530e-02	3.791e-03	-6.674	2.49e-11	***
unemp	8.170e+00	4.370e+00	1.870	0.061531	.
DTI_wm	5.837e-02	4.904e-02	1.190	0.234016	
DTI_wv	4.877e-03	6.228e-03	0.783	0.433550	
sqrt(I_count)	2.130e-02	9.877e-04	21.567	< 2e-16	***
LIBOR1	-2.197e+00	4.148e-01	-5.296	1.18e-07	***
ORIG_RT_wm	-1.163e+00	1.272e-01	-9.141	< 2e-16	***
PROP_TYP_CO_wm	1.100e+01	7.446e+00	1.478	0.139471	
PROP_TYP_CP_wm	-2.438e+01	1.752e+01	-1.391	0.164084	
PROP_TYP_MH_wm	4.732e+01	4.369e+01	1.083	0.278784	
PROP_TYP_PU_wm	-7.305e+00	2.878e+00			
OCC_STAT_I_wm	-7.021e+01	1.218e+01			
OCC_STAT_P_wm	-4.965e+01	9.060e+00			
OCLTV_wv:unemp	5.344e-03	1.606e-03			
LIBOR1:ORIG_RT_wm	2.748e-01	6.635e-02			
unemp:OCC_STAT_I_wm	-1.239e+01	6.328e+00			
unemp:OCC_STAT_P_wm	-9.803e+00	4.444e+00			

```
PROP_TYP_PU_wm                          -2.538 0.011136 *
OCC_STAT_I_wm                          -5.765 8.17e-09 ***
OCC_STAT_P_wm                           -5.480 4.26e-08 ***
OCLTV_wv:unemp                         3.327 0.000878 ***
LIBOR1:ORIG_RT_wm                          4.141 3.45e-05 ***
unemp:OCC_STAT_I_wm                        -1.958 0.050274 .
unemp:OCC_STAT_P_wm                         -2.206 0.027407 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  1

Name of linear predictor: logit(prob)

Residual deviance: 6067.598 on 7024 degrees of freedom

Log-likelihood: -3033.799 on 7024 degrees of freedom

Number of iterations: 6

Warning: Hauck-Donner effect detected in the following estimate(s):
'OCC_STAT_P_wm'
```

6.6 Fit to test data

```
          0         1
  0 0.8095238 0.1904762
  1 0.1612200 0.8387800
Area under the curve (AUC): 0.824
```

7 Gradient boosted model for Stage 1

Along with other model-averaging (ensemble) methods, gradient-boosted regression trees or gradient-boosted models (GBMs) differs fundamentally from conventional regression based techniques such as generalised additive models (GAM – Hastie & Tibshirani 1990). Whereas the latter seek to fit the single most parsimonious model that best describes the relationship between a response variable and some set of predictors, ensemble methods fit a large number of relatively simple models whose predictions are then combined to give more robust estimates of the response. In boosted regression trees (BRT) each of the individual models consists of a simple classification or regression tree, i.e. a rule-based classifier that partitions observations into groups having similar values for the response variable, based on a series of binary rules (splits) constructed from the predictor variables (Hastie et al. 2001). The boosting algorithm uses an iterative method for developing a final model in a forward stage-wise fashion, progressively adding trees to the model, while re-weighting the data to emphasize cases poorly predicted by the previous trees. A BRT model can therefore be seen as a regression model in which each of the individual model terms is a simple regression tree (Friedman et al. 2000).

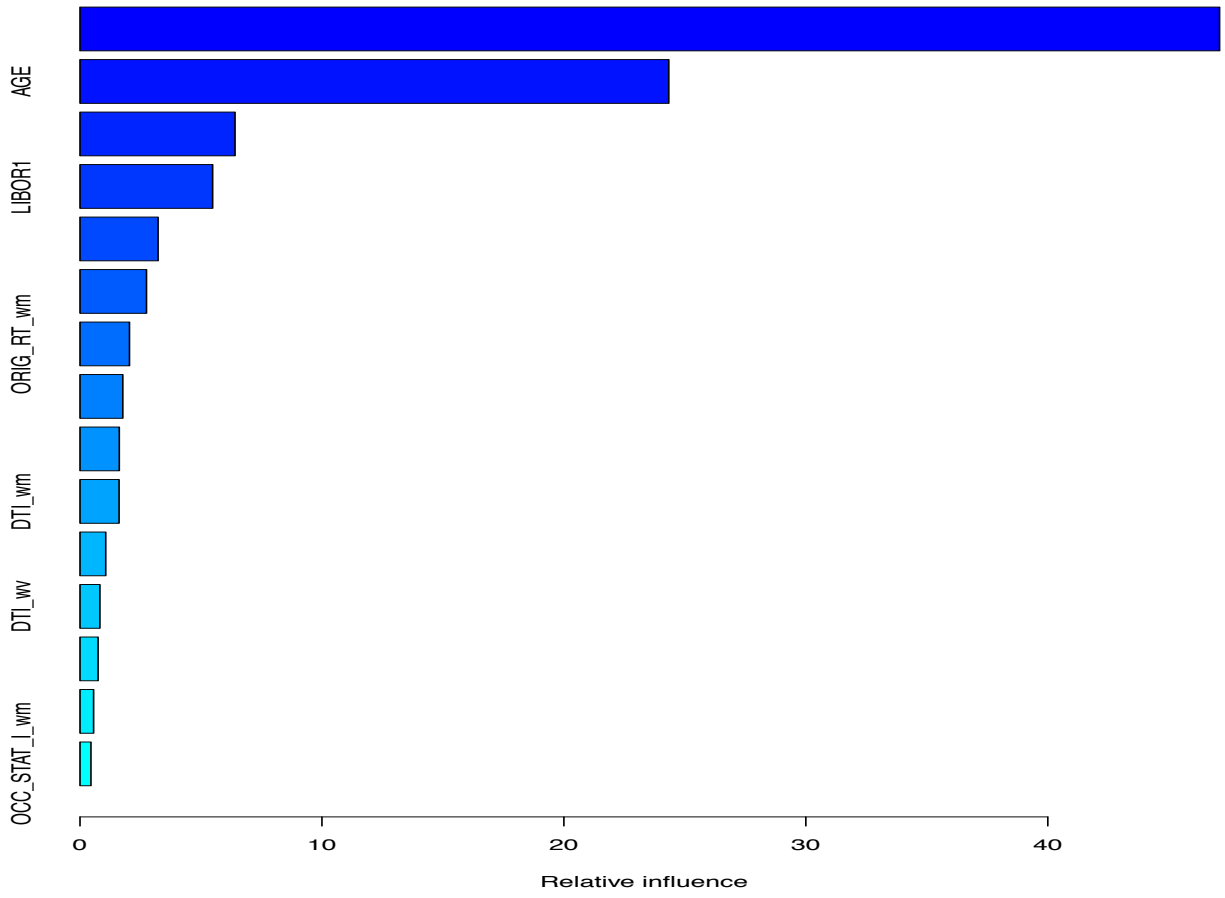
Advantages offered by a BRT model include its ability to accommodate both different types of predictor variables and missing values, its immunity to the effects of extreme outliers and the inclusion of irrelevant predictors, and its facility for fitting interactions between predictors (Friedman & Meulman 2003). Fitting of interaction effects is controlled by varying the size of the individual regression trees. Where the individual tree terms consist of a single rule constructed using just 1 predictor variable, no interaction effects are fitted, and the final model is likely to approximate closely one fitted using any conventional regression technique that allows the fitting of nonlinear responses, e.g. a GAM. However, where the individual trees consist of 2 or more rules, the function fitted for any one predictor may vary depending on the value taken by another predictor, with the potential complexity of these interaction effects increasing as the size of the individual tree terms increases.

(Leathwick, 2006)

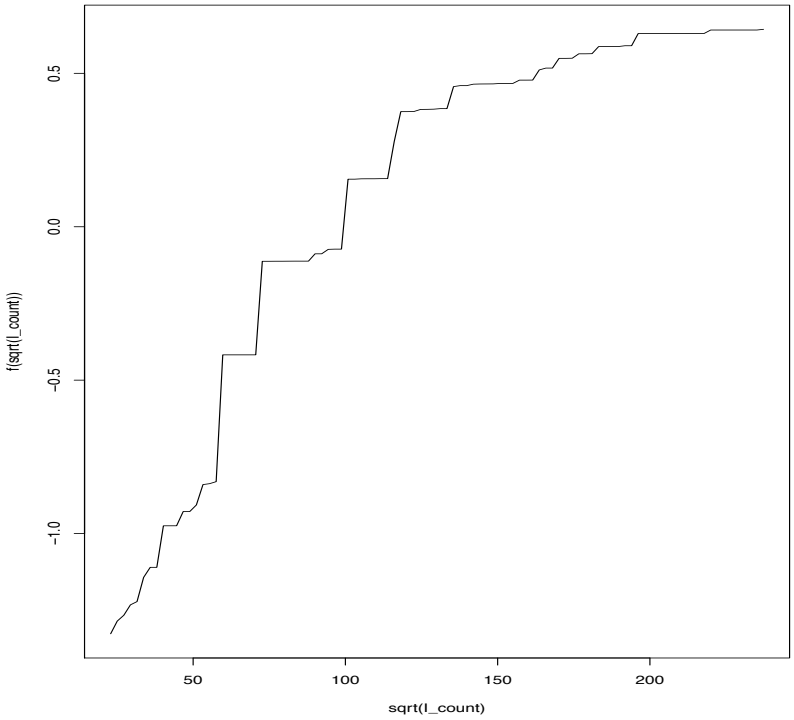
Pros:

- No age splines, no need to consider interactions, easy interpretability, scalable
- The model allows us to rank variables by importance as shown below:

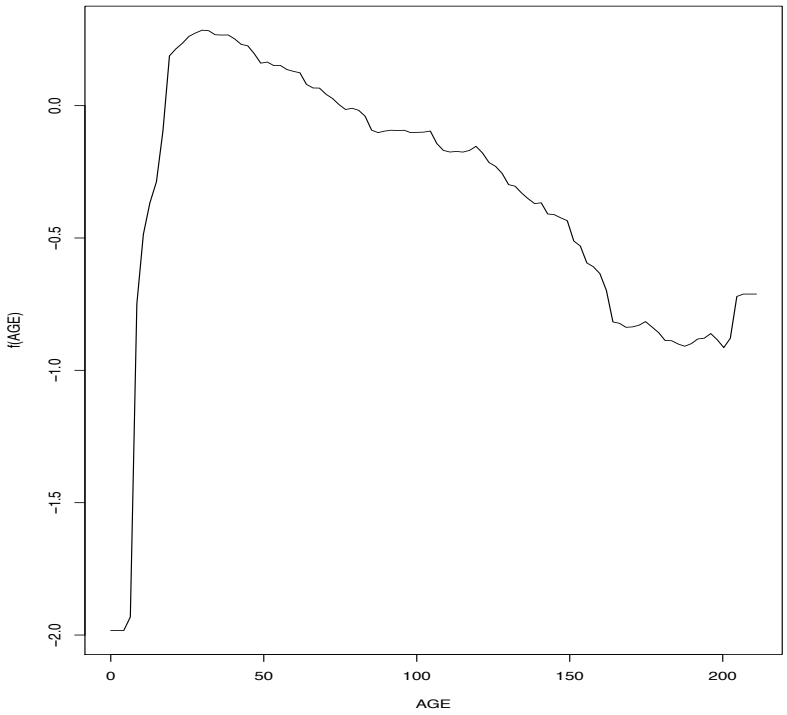
	var	rel.inf
sqrt(I_count)	sqrt(I_count)	47.1001199
AGE	AGE	24.3448080
unemp	unemp	6.3898901
LIBOR1	LIBOR1	5.4361702
OCC_STAT_P_wm	OCC_STAT_P_wm	3.2952521
PROP_TYP_PU_wm	PROP_TYP_PU_wm	2.7467181
ORIG_RT_wm	ORIG_RT_wm	1.9595021
CSCORE_MN_wm	CSCORE_MN_wm	1.7469519
OCLTV_wv	OCLTV_wv	1.6579038
DTI_wm	DTI_wm	1.5800054
PROP_TYP_CP_wm	PROP_TYP_CP_wm	1.0887741
DTI_wv	DTI_wv	0.8060106
PROP_TYP_CO_wm	PROP_TYP_CO_wm	0.7478713
PROP_TYP_MH_wm	PROP_TYP_MH_wm	0.6188146
OCC_STAT_I_wm	OCC_STAT_I_wm	0.4812076

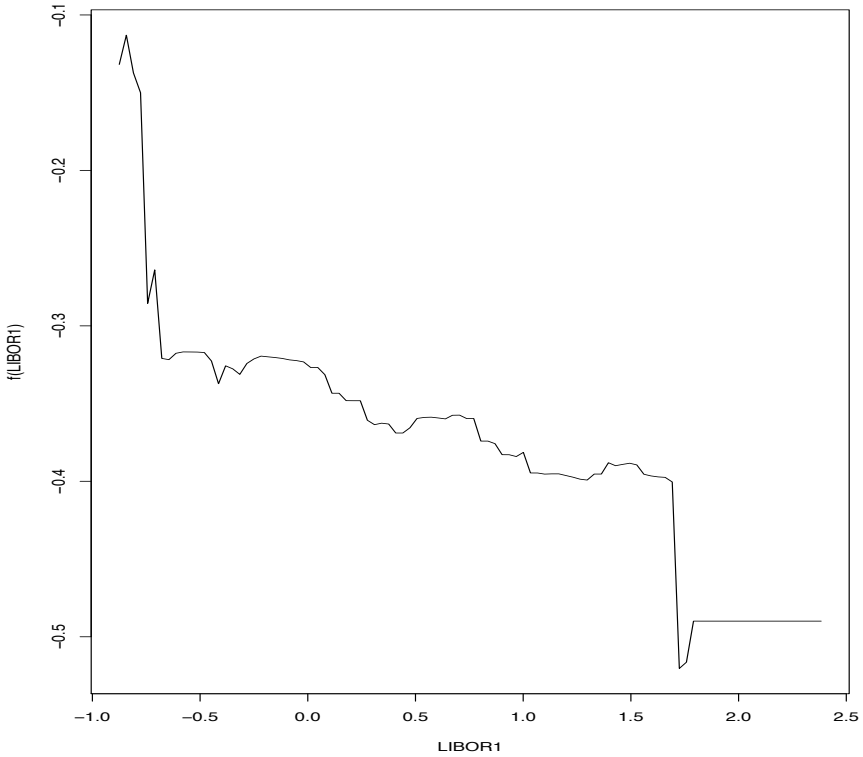


- We are also able to see the relationship of each covariate with the dependent variable after the model is fit. A few examples are shown below:



-





7.1 Fit to test data

	var	rel.inf
ORIG_AMT_sum	ORIG_AMT_sum	38.69498197
AGE	AGE	22.30303537
I_count	I_count	6.96493703
unemp	unemp	4.56987197
monthly_rGDP	monthly_rGDP	3.69409400
LIBOR1	LIBOR1	3.55930055
cpi	cpi	2.78220249
OCC_STAT_S_wm	OCC_STAT_S_wm	2.07522615
ORIG_RT_var	ORIG_RT_var	1.04837571
PURPOSE_C_wm	PURPOSE_C_wm	1.02574224
ORIG_CHN_R_wm	ORIG_CHN_R_wm	0.95874057
OCC_STAT_P_wm	OCC_STAT_P_wm	0.89954551
PROP_TYP_PU_wm	PROP_TYP_PU_wm	0.81005355
OCLTV_wv	OCLTV_wv	0.64884770
PROP_TYP_CP_wm	PROP_TYP_CP_wm	0.63869934
CSCORE_MN_var	CSCORE_MN_var	0.50291653
ORIG_RT_wm	ORIG_RT_wm	0.50174784
OCLTV_var	OCLTV_var	0.47246314

	0	1
0	0.83333333	0.16666667
1	0.09803922	0.90196078

Area under the curve (AUC): 0.868

According to AUC, GBM gives a more accurate prediction than a linear model.

8 GBM-GAM: 2 part model

Data: Quarter 4 data from years 2000 to 2005, inclusive.

Our final model is fit in 2 stages. We reiterate the model specification, which is:

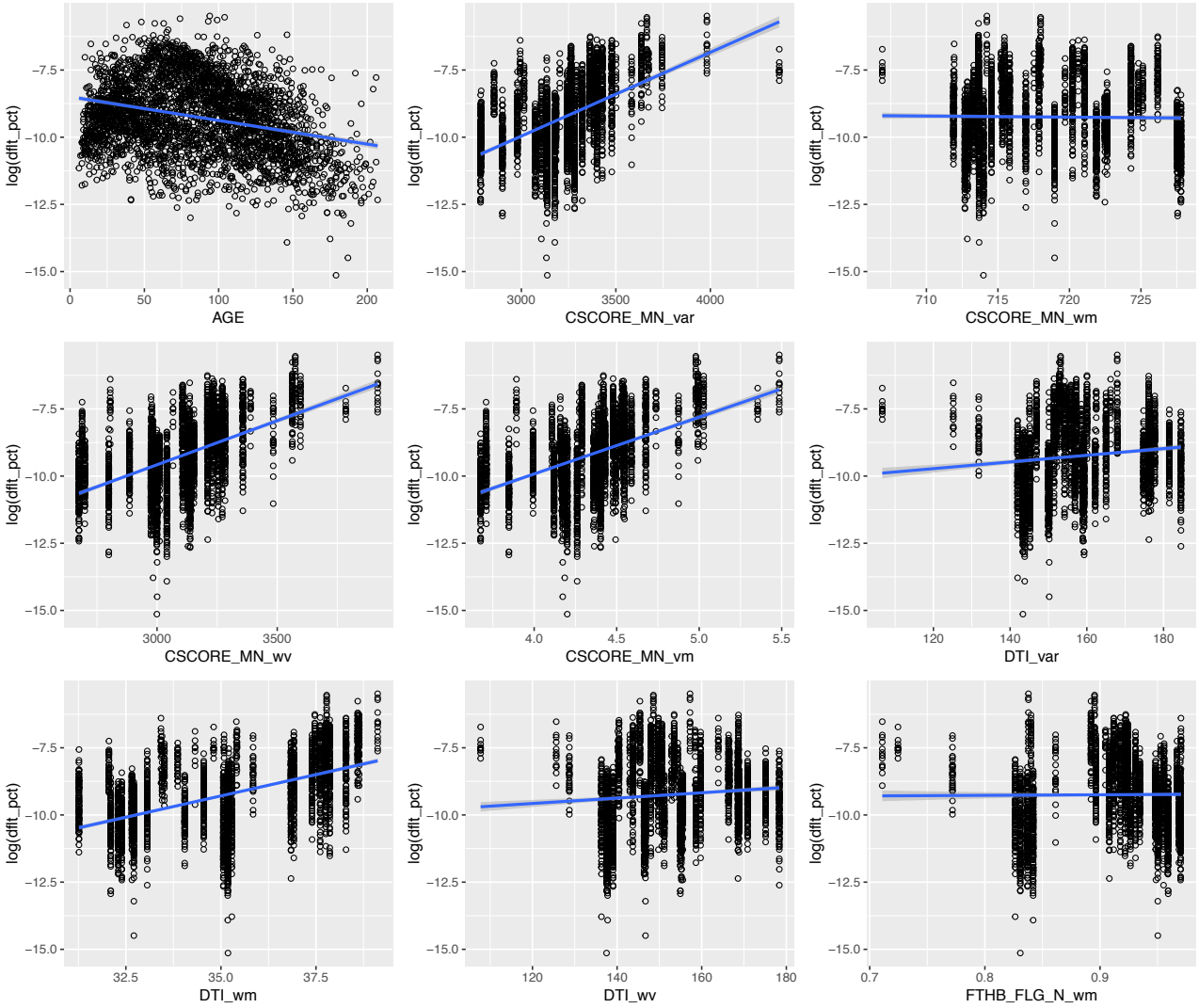
$$E(\log \mathrm{Y} \mid X) = P(Y > 0 \mid X) * E(\log Y \mid Y > 0, X)$$

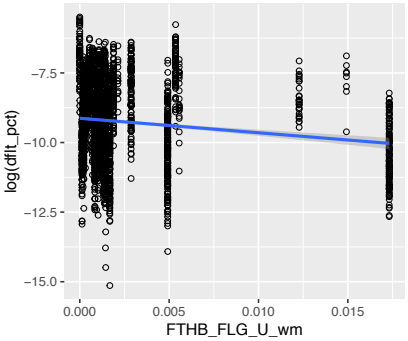
Where Y is the default ratio (ie. % amount defaulted within a vintage on a particular month)

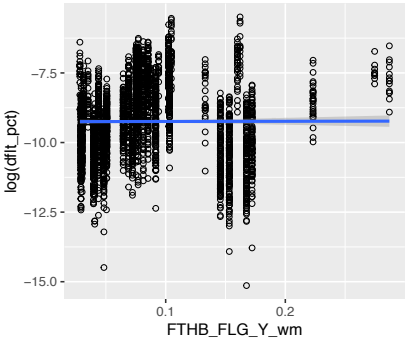
We fit 2 separate models for each term on the right side of the equation. $P(Y > 0 \mid X)$ is stage 1 and $E(\log Y \mid Y > 0, X)$ is stage 2. Our final model will be a gradient-boosted model (GBM) for the first stage and a generalized additive model (GAM) for the 2nd stage. Since GAM, like logistic regression, requires us to select the covariates, we investigate the relationship between each covariate with the dependent variable in order to select the most meaningful variables. Note we use a smoothing spline with 5 degrees of freedom for age.

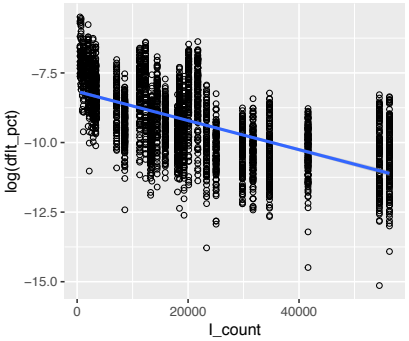
8.1 Variable selection for Stage 2 GAM

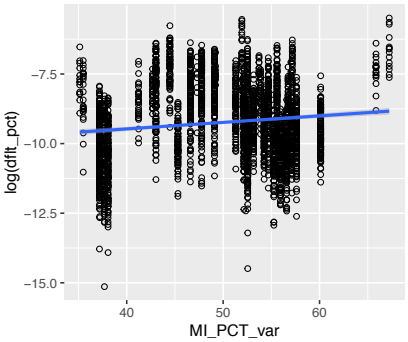
Plots below show a linear fit of each covariate to the dependent variable, which is the log of the default ratio.

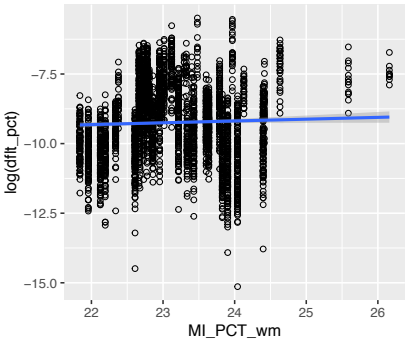


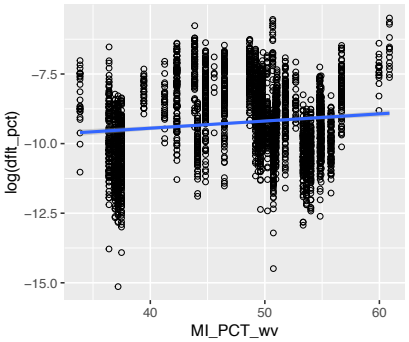

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0.000
0.005
0.010
0.015
FTHB_FLG_U_wm

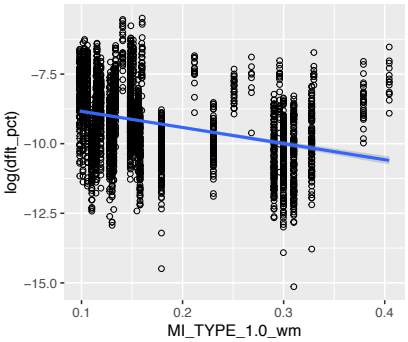

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0.1
0.2
FTHB_FLG_Y_wm

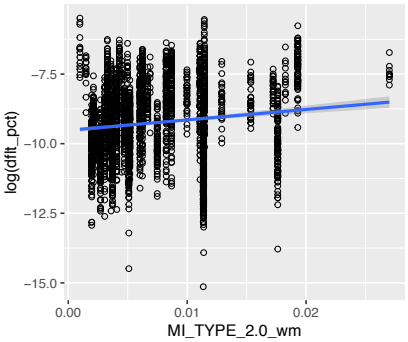

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0
20000
40000
l_count

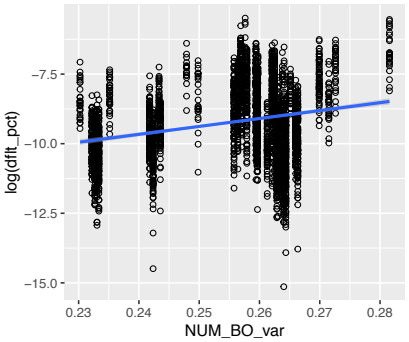

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
40
50
60
MI_PCT_var

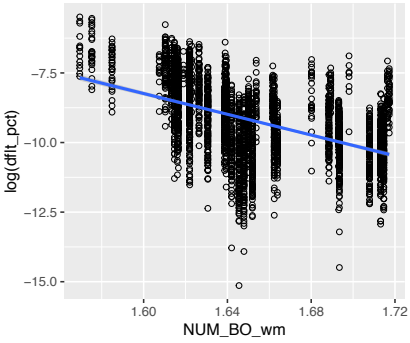

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
22
23
24
25
26
MI_PCT_wm

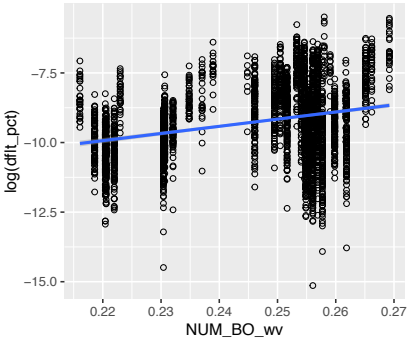

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
40
50
60
MI_PCT_wv

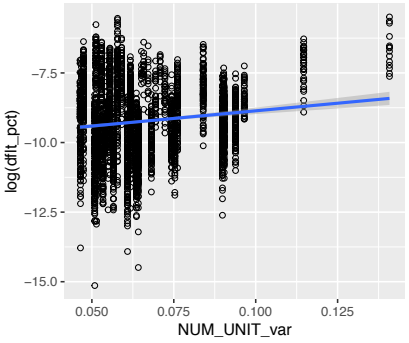

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0.1
0.2
0.3
0.4
MI_TYPE_1.0_wm

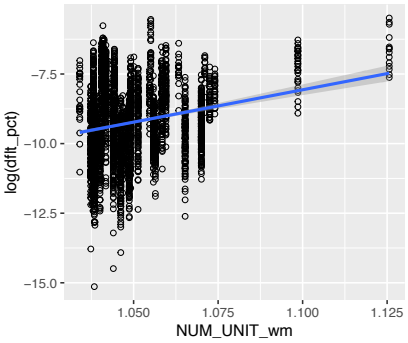

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0.00
0.01
0.02
MI_TYPE_2.0_wm

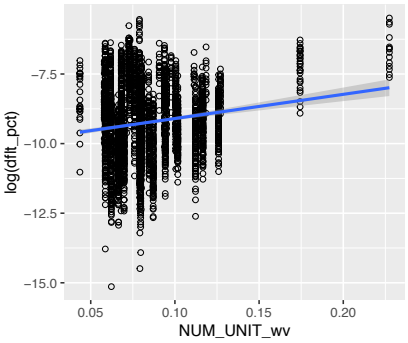

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0.23
0.24
0.25
0.26
0.27
0.28
NUM_BO_var

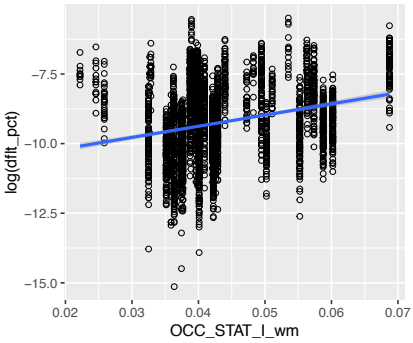

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
1.60
1.64
1.68
1.72
NUM_BO_wm

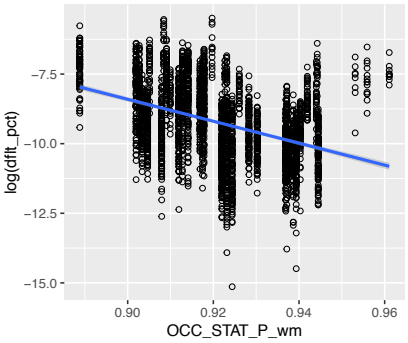

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0.22
0.23
0.24
0.25
0.26
0.27
NUM_BO_wv

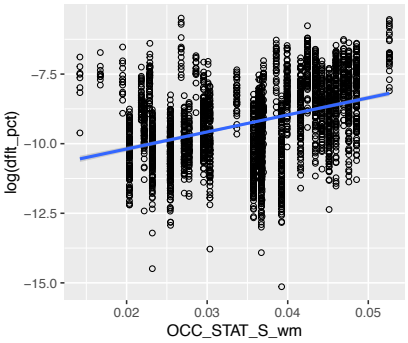

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0.050
0.075
0.100
0.125
NUM_UNIT_var

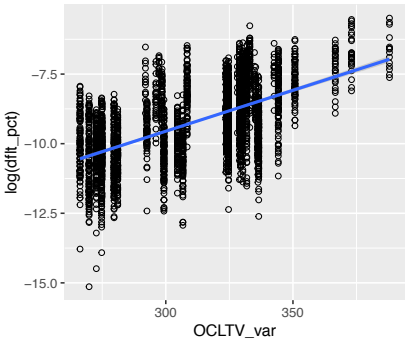

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
1.050
1.075
1.100
1.125
NUM_UNIT_wm

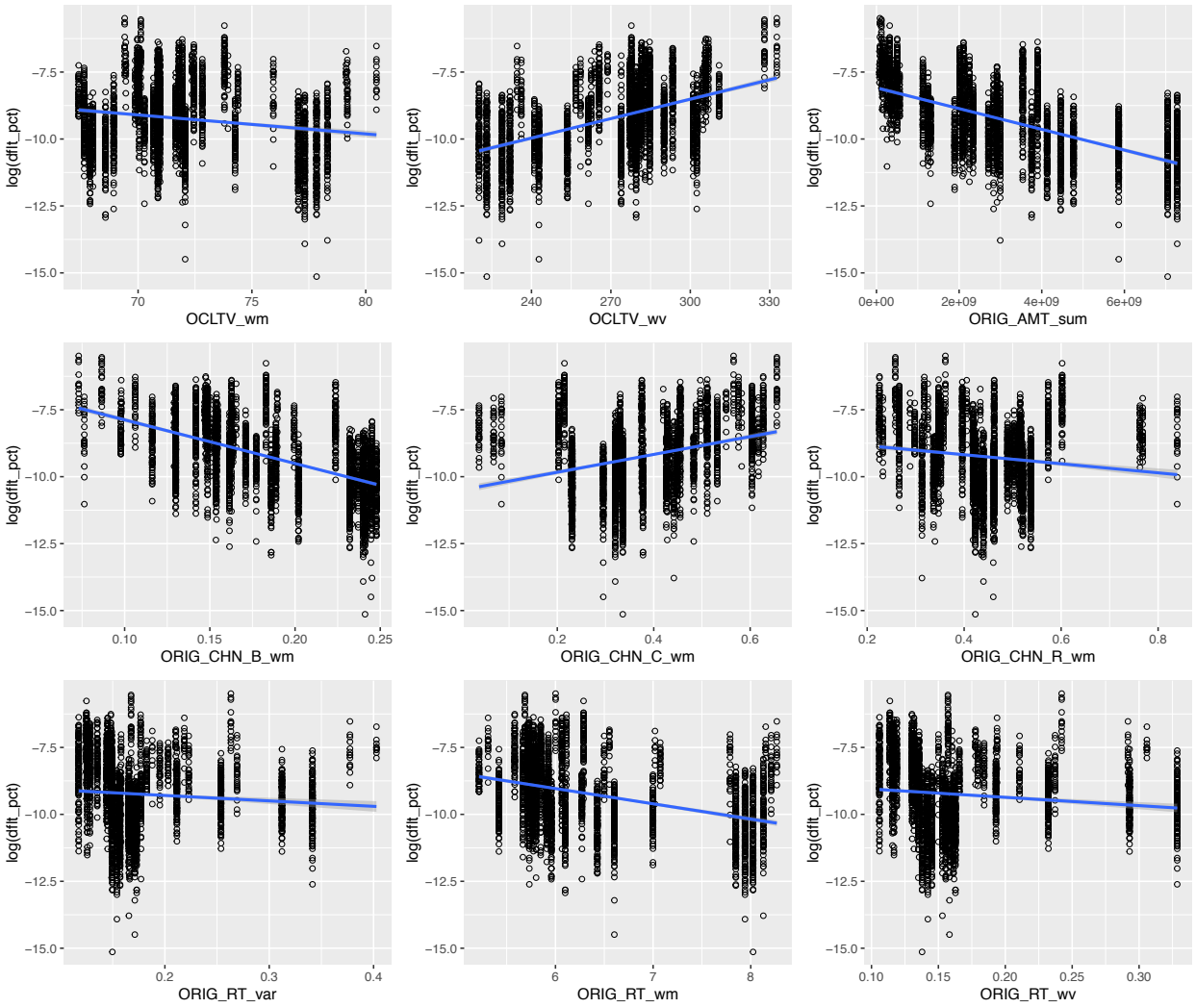

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0.05
0.10
0.15
0.20
NUM_UNIT_wv

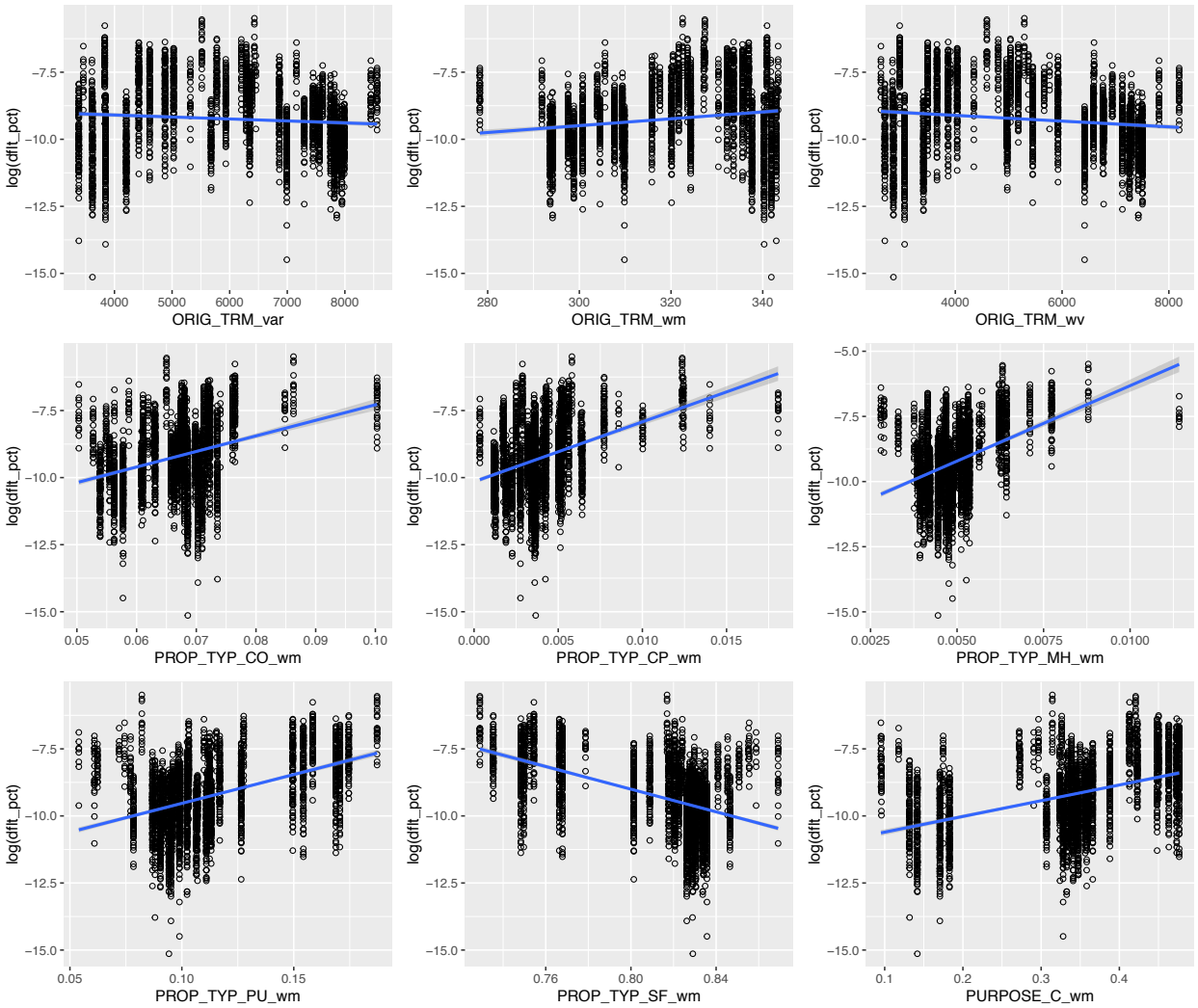

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0.02
0.03
0.04
0.05
0.06
0.07
OCC_STAT_I_wm

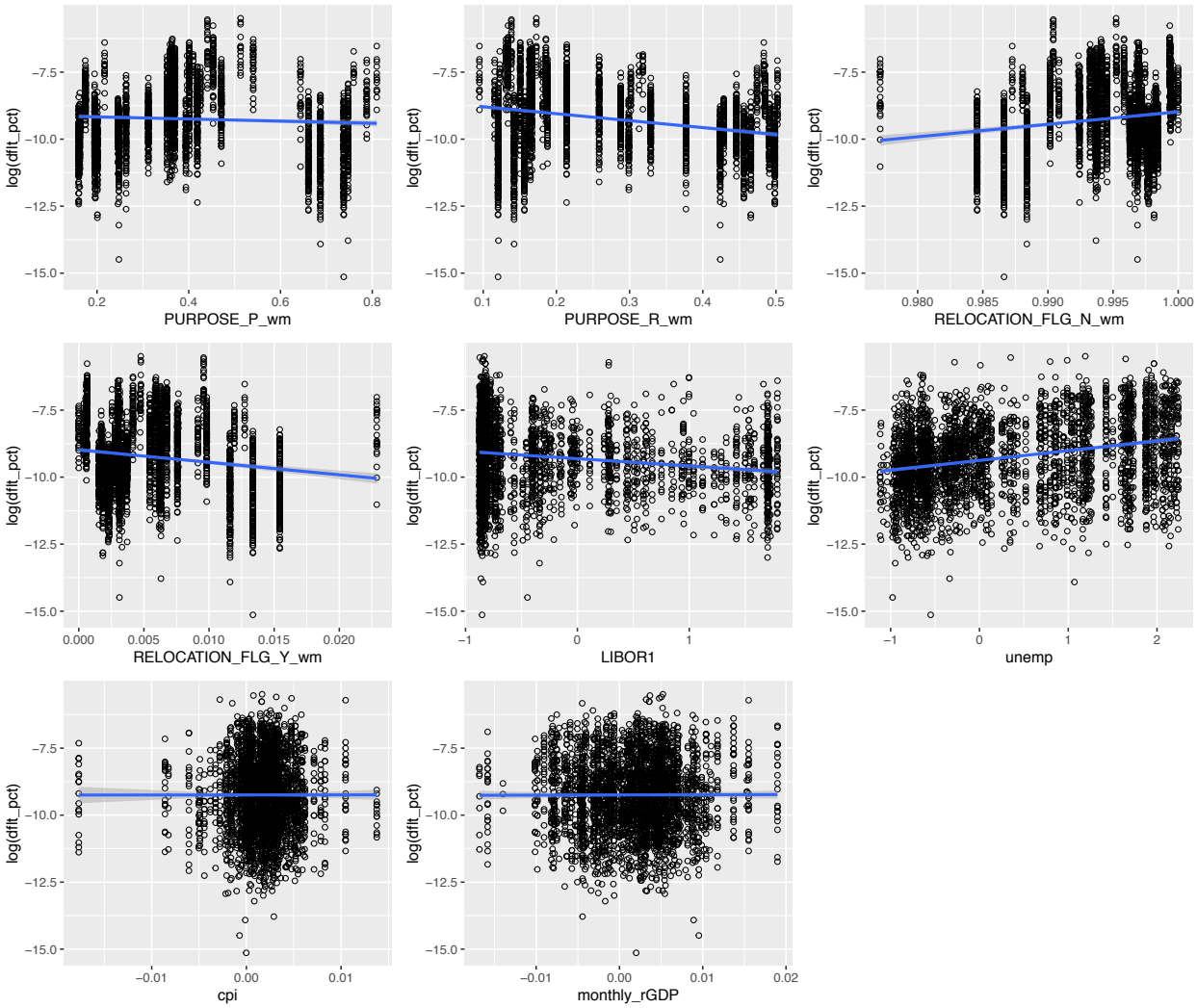

log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0.90
0.92
0.94
0.96
OCC_STAT_P_wm


log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
0.02
0.03
0.04
0.05
OCC_STAT_S_wm


log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
300
350
OCLTV_var


log(dflt_pct)
−7.5
−10.0
−12.5
−15.0
70
75
80
OCLTV_wm
240
270
300
330
OCLTV_wv
0e+00
2e+09
4e+09
6e+09
ORIG_AMT_sum
0.10
0.15
0.20
0.25
ORIG_CHN_B_wm
0.2
0.4
0.6
ORIG_CHN_C_wm
0.2
0.4
0.6
0.8
ORIG_CHN_R_wm
0.2
0.3
0.4
ORIG_RT_var
6
7
8
ORIG_RT_wm
0.10
0.15
0.20
0.25
0.30
ORIG_RT_wv


log(dflt_pct)
ORIG_TRM_var
ORIG_TRM_wm
ORIG_TRM_wv
PROP_TYP_CO_wm
PROP_TYP_CP_wm
PROP_TYP_MH_wm
PROP_TYP_PU_wm
PROP_TYP_SF_wm
PURPOSE_C_wm



As another test for variable selection, we fit the model 8 times on a randomly chosen test set and count the number of times each interaction term appears as statistically significant in our model (as measured by AIC). The following are the counts of variables/interaction terms:

5(PURPOSE_C_wm+PURPOSE_P_wm)*log(ORIG_AMT_sum)+
3OCLTV_wm*LIBOR1+
OCLTV_wm*unemp
4OCLTV_wv*LIBOR1+
4OCLTV_wv*unemp+
4(PURPOSE_C_wm+PURPOSE_P_wm)*unemp+
5s(AGE, df=5)*OCLTV_wv+
2s(AGE, df=5)*OCLTV_wm+
5sqrt(I_count)*LIBOR1+
8s(AGE, df=5)*sqrt(I_count)+
2(PURPOSE_C_wm+PURPOSE_P_wm)*DTI_wm+
2(PURPOSE_C_wm+PURPOSE_P_wm)*LIBOR1
4(PURPOSE_C_wm+PURPOSE_P_wm)*unemp
(PURPOSE_C_wm+PURPOSE_P_wm)*CSCORE_MN_wv
(OCC_STAT_I_wm+OCC_STAT_P_wm)*sqrt(I_count)+
(OCC_STAT_I_wm+OCC_STAT_P_wm)*unemp
(OCC_STAT_I_wm+OCC_STAT_P_wm)*LIBOR1+
(OCC_STAT_I_wm+OCC_STAT_P_wm)*log(ORIG_AMT_sum) +
s(AGE, df=5)*(PURPOSE_C_wm+PURPOSE_P_wm)+
3log(ORIG_AMT_sum)*LIBOR1+
4log(ORIG_AMT_sum)*unemp+
2DTI_wm+

CSCORE_MN_wm
3DTI_wm*LIBOR1
3(PROP_TYP_CO_wm+PROP_TYP_CP_wm+PROP_TYP_MH_wm+PROP_TYP_PU_wm)*log(ORIG_AMT_sum)
PROP_TYP_CO_wm+PROP_TYP_CP_wm+PROP_TYP_MH_wm+PROP_TYP_PU_wm

Based on the above analyses, the reasonable choices of variables are:
cont.vars: s(AGE, df=5) + CSCORE_MN_vm + DTI_wm + sqrt(I_count) + NUM_BO_wm + OCLTV_wv + log(ORIG_AMT_sum)
cat.vars: (OCC_STAT_I_wm+OCC_STAT_P_wm) + (ORIG_CHN_B_wm+ORIG_CHN_C_wm+ORIG_CHN_R_wm) +
(PROP_TYP_CO_wm+PROP_TYP_CP_wm+PROP_TYP_MH_wm+PROP_TYP_PU_wm) +
(PURPOSE_C_wm+PURPOSE_P_wm)

Our initial model specification is:
log(dflt_pct)~ s(AGE, df=5) +
CSCORE_MN_vm + DTI_wm + sqrt(I_count) + NUM_BO_wm + OCLTV_wv + log(ORIG_AMT_sum) +
(OCC_STAT_I_wm+OCC_STAT_P_wm) + (ORIG_CHN_B_wm+ORIG_CHN_C_wm+ORIG_CHN_R_wm) +
(PROP_TYP_CO_wm+PROP_TYP_CP_wm+PROP_TYP_MH_wm+PROP_TYP_PU_wm) +
(PURPOSE_C_wm+PURPOSE_P_wm) +unemp
s(AGE, df=5)*sqrt(I_count)+
(PURPOSE_C_wm+PURPOSE_P_wm)*log(ORIG_AMT_sum)+
OCLTV_wv*LIBOR1+
s(AGE, df=5)*OCLTV_wm

9 3 Stage Vintage model

3 stage model:

$$E(\mathrm{Y} \mid X) = P(Y > 0 \mid X) * \exp\big(E(log\, Y \mid Y > 0, X)\big) * \phi$$

Y = % of ORIG_AMT_sum exposed at default (= 0 if didn't default)

Phi = smearing estimator (addresses the exponential transformation)

$$E(L \mid X) = E(\mathrm{Y} \mid X) * E(\mathrm{Z} \mid X)$$

Z = % of exposed amount that is lost

L = % of ORIG_AMT_sum lost at default

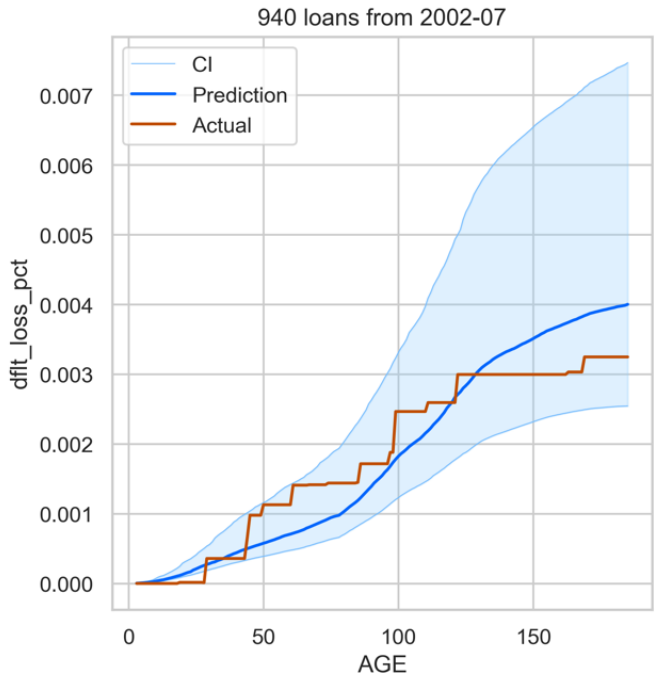
9.1 Model specs

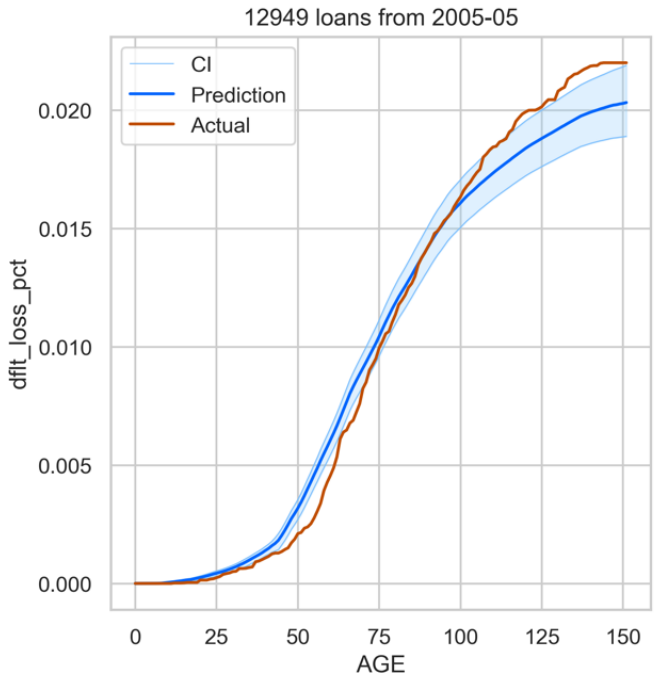
1. selected = ['ORIG_AMT_sum', 'UNEMP', 'ORIG_RT_wm', 'MR', 'HPI', 'DTI_wm', 'CPI', 'LOAN_ID_count',
'LIBOR', 'rGDP', 'AGE', 'DTI_wv', 'ORIG_CHN_R_wv', 'ORIG_RT_wv', 'ORIG_RT_cv', 'DTI_cv',
'CSCORE_MN_wm']
gbm_formula1 = 'did_dflt ~ -1 + {0}'.format(' + '.join(selected))

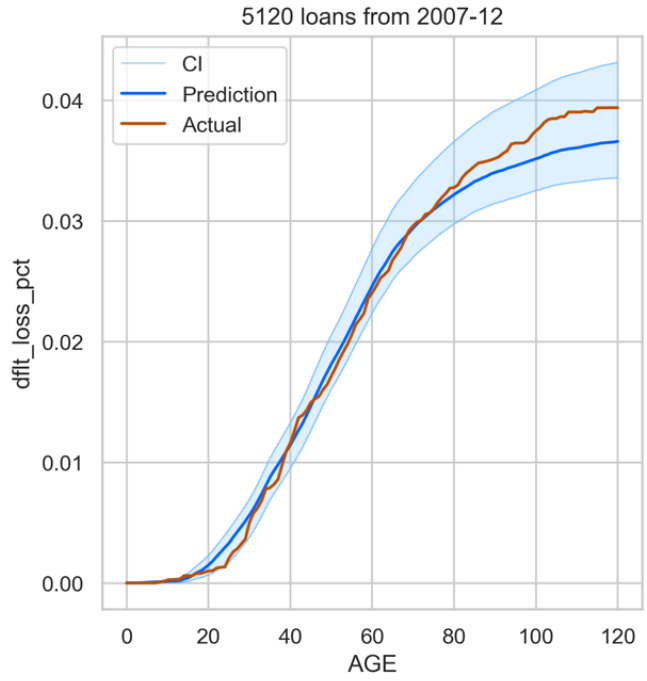
2. selected = ['(PURPOSE_P_wm + PURPOSE_R_wm)*(np.log(ORIG_AMT_sum) + DTI_wm)',
'DTI_wm*cr(AGE, df=5)',
'ORIG_RT_wm*cr(AGE, df=5)', 'np.log(ORIG_AMT_sum)*UNEMP',
'CSCORE_MN_wv*MR', 'np.sqrt(LOAN_ID_count)*UNEMP', 'DTI_wm*UNEMP',
'np.sqrt(LOAN_ID_count)*cr(AGE, df=5)', 'LIBOR']
gam_formula2 = ('np.log(dflt_pct) ~ {0}'.format(' + '.join(selected)))

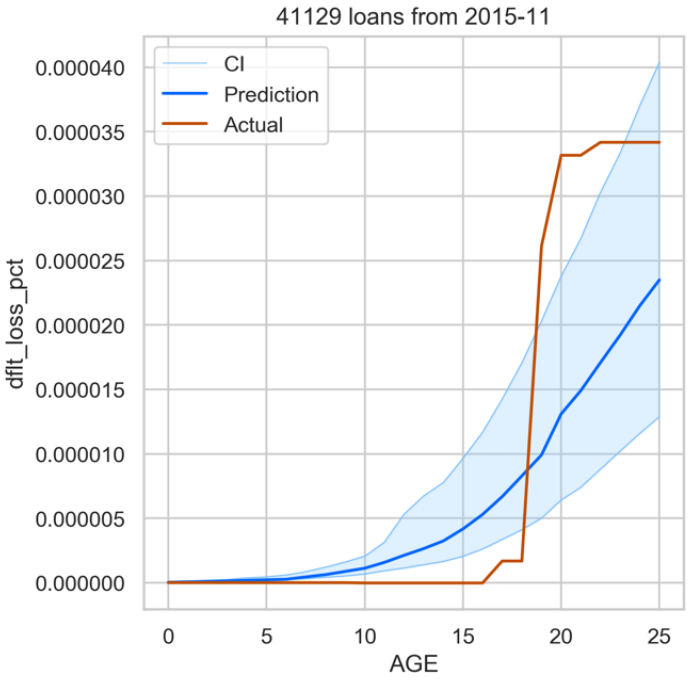
3. selected = ['AGE*UNEMP', 'DTI_wm', 'HPI', 'OCC_STAT_S_wm', 'PURPOSE_P_wm*AGE',
'NUM_BO_wm*UNEMP']
gam_formula3 = ('net_loss_pct ~ {0}'.format(' + '.join(selected)))

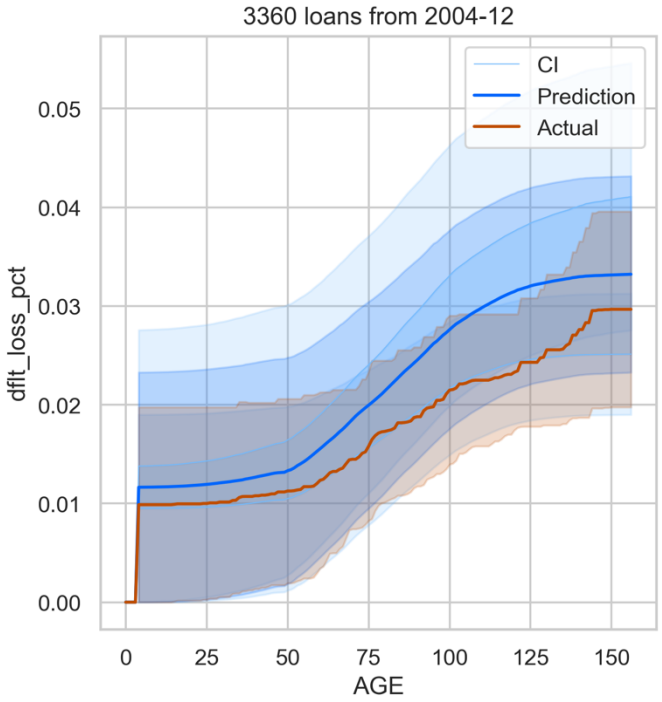
9.2 Vintage level results

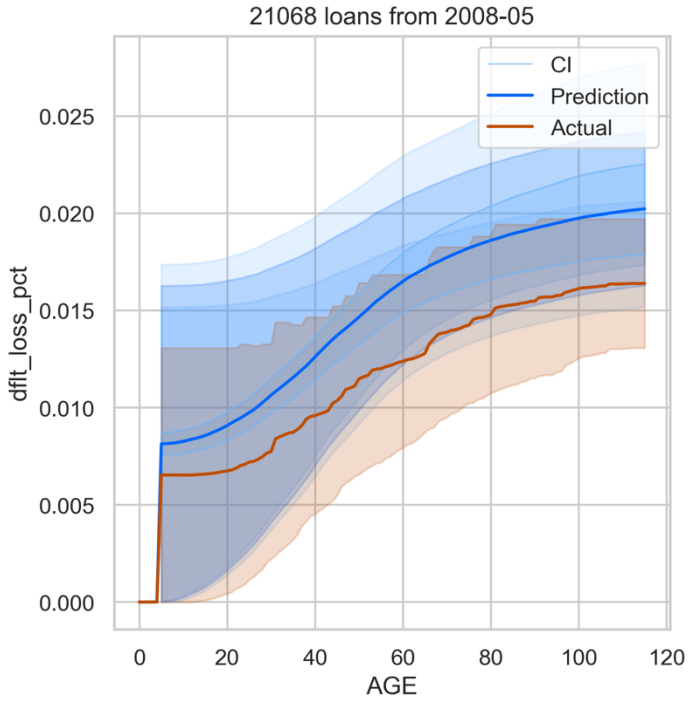


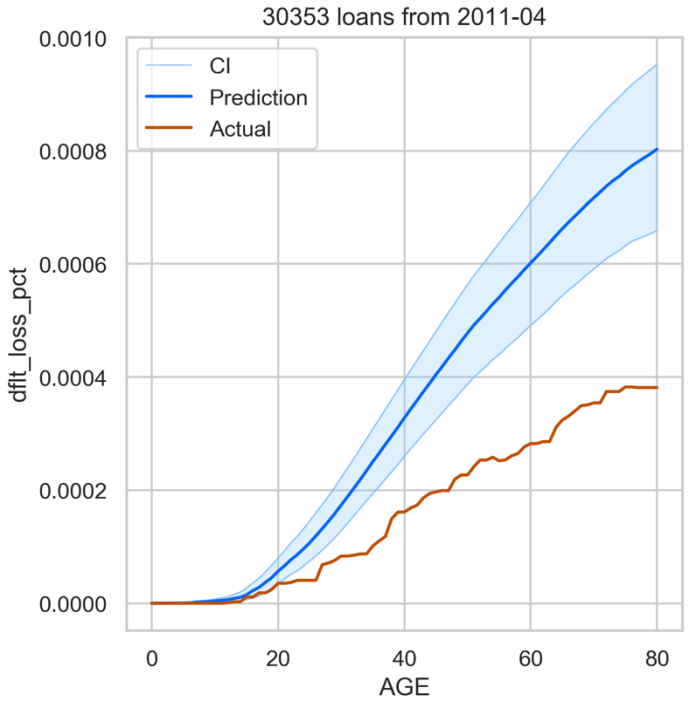


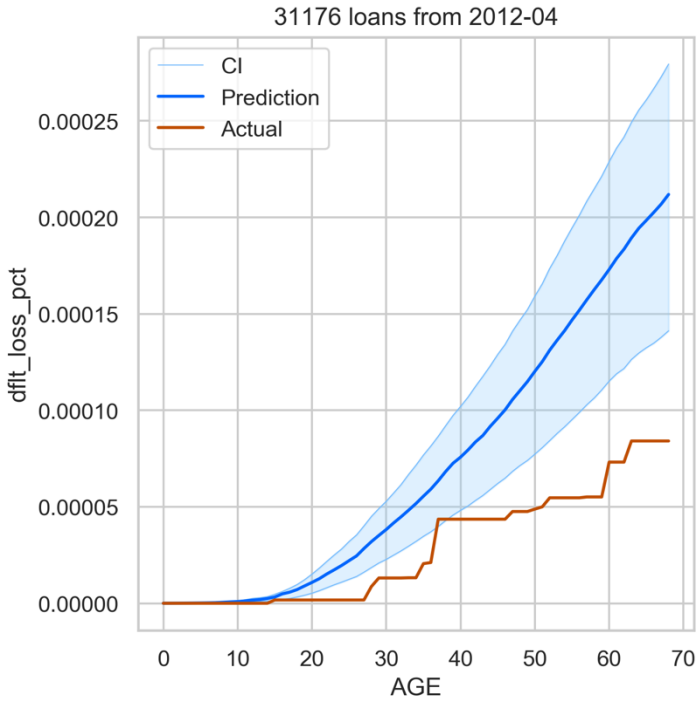


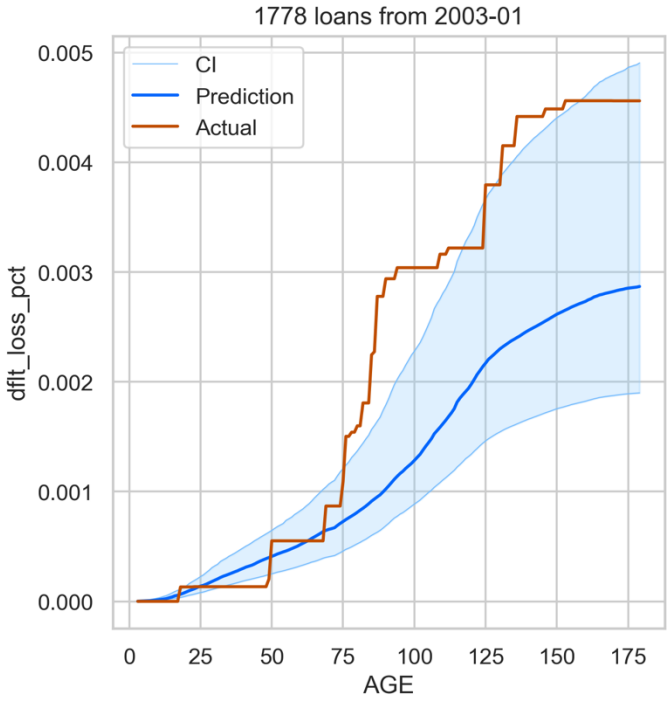


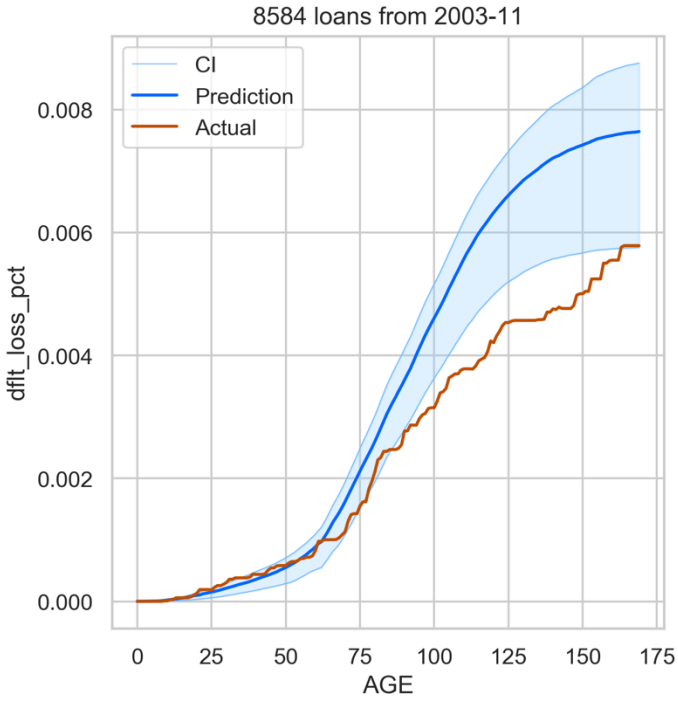

3360 loans from 2004-12
CI
Prediction
Actual
dflt_loss_pct
0.00
0.01
0.02
0.03
0.04
0.05
0
25
50
75
100
125
150
AGE


21068 loans from 2008-05
CI
Prediction
Actual
dflt_loss_pct
0.000
0.005
0.010
0.015
0.020
0.025
0
20
40
60
80
100
120
AGE

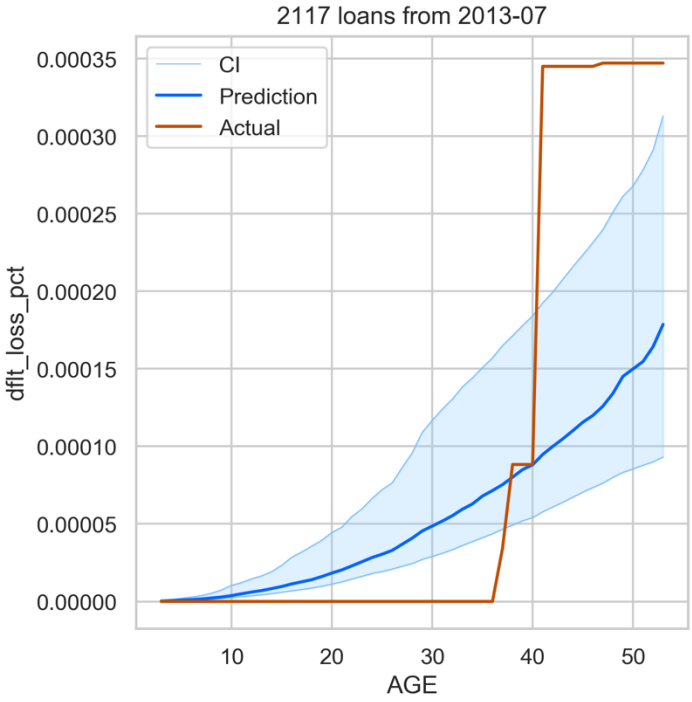

30353 loans from 2011-04
CI
Prediction
Actual
dflt_loss_pct
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0
20
40
60
80
AGE

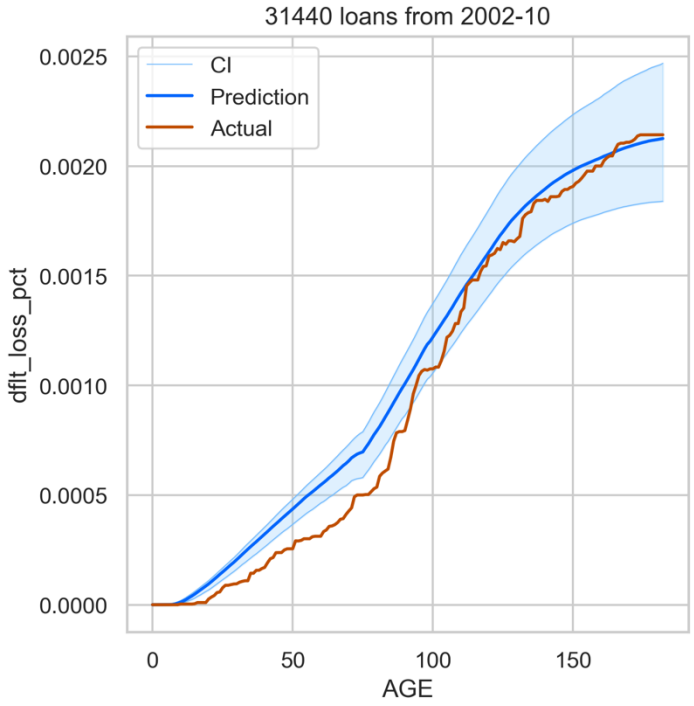

31176 loans from 2012-04
CI
Prediction
Actual
dflt_loss_pct
0.00000
0.00005
0.00010
0.00015
0.00020
0.00025
0
10
20
30
40
50
60
70
AGE

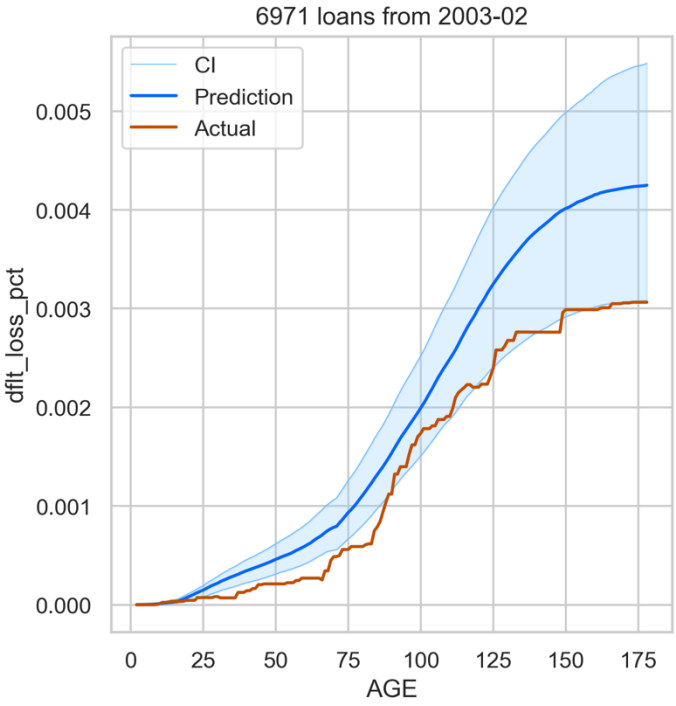

1778 loans from 2003-01
CI
Prediction
Actual
dflt_loss_pct
0.005
0.004
0.003
0.002
0.001
0.000
0
25
50
75
100
125
150
175
AGE

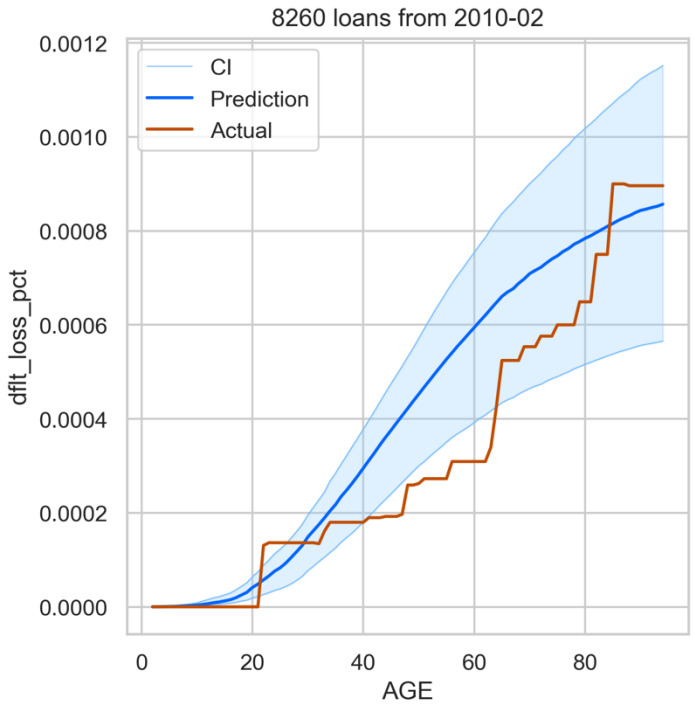

8584 loans from 2003-11
CI
Prediction
Actual
dflt_loss_pct
0.008
0.006
0.004
0.002
0.000
0
25
50
75
100
125
150
175
AGE

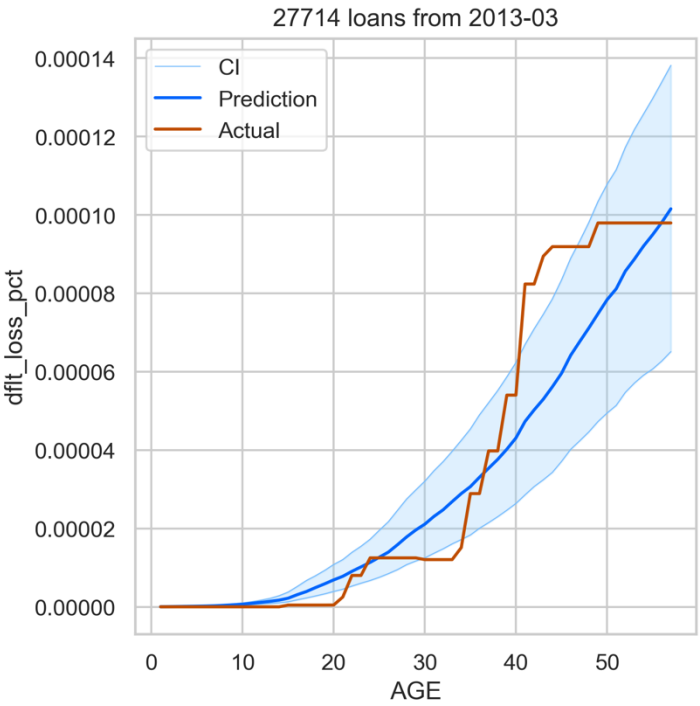

25850 loans from 2013-05
CI
Prediction
Actual
dflt_loss_pct
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
10
20
30
40
50
AGE

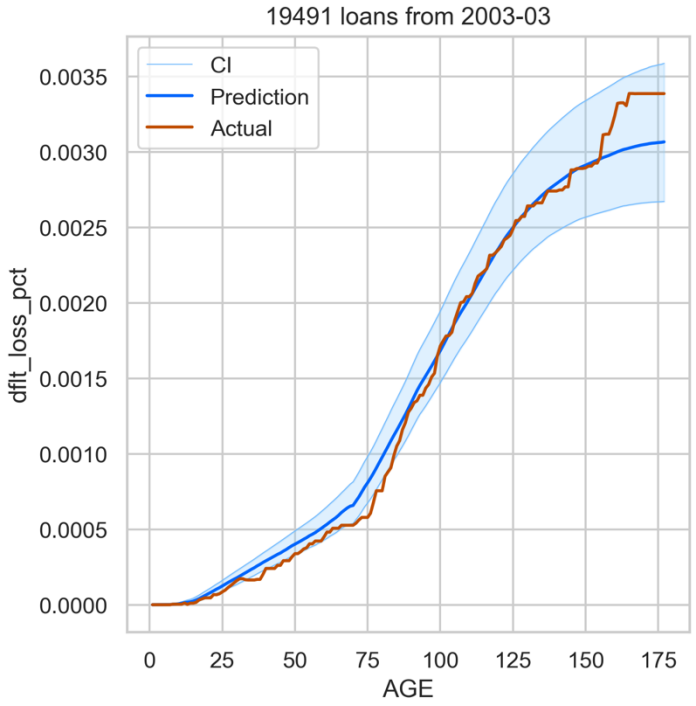

2117 loans from 2013-07
CI
Prediction
Actual
dflt_loss_pct
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
10
20
30
40
50
AGE

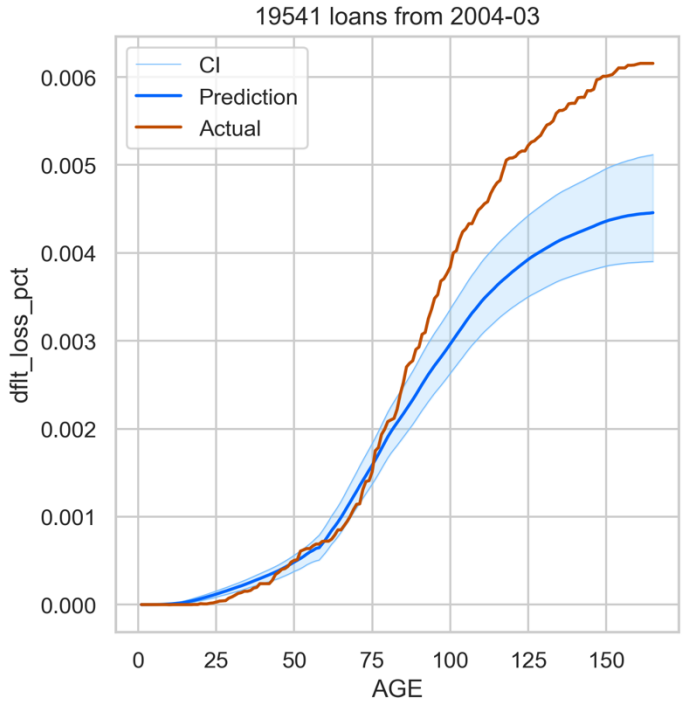

31440 loans from 2002-10
CI
Prediction
Actual
dflt_loss_pct
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
50
100
150
AGE

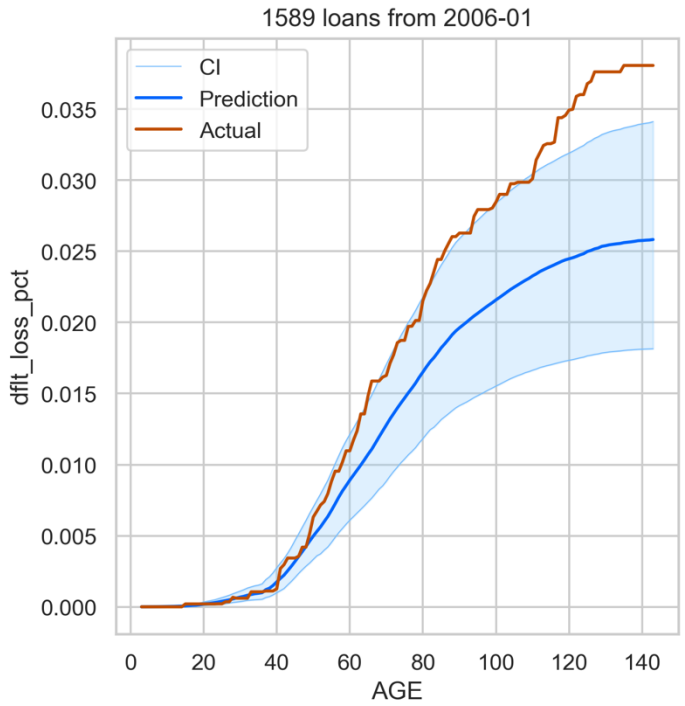

6971 loans from 2003-02
CI
Prediction
Actual
dflt_loss_pct
0.005
0.004
0.003
0.002
0.001
0.000
0
25
50
75
100
125
150
175
AGE

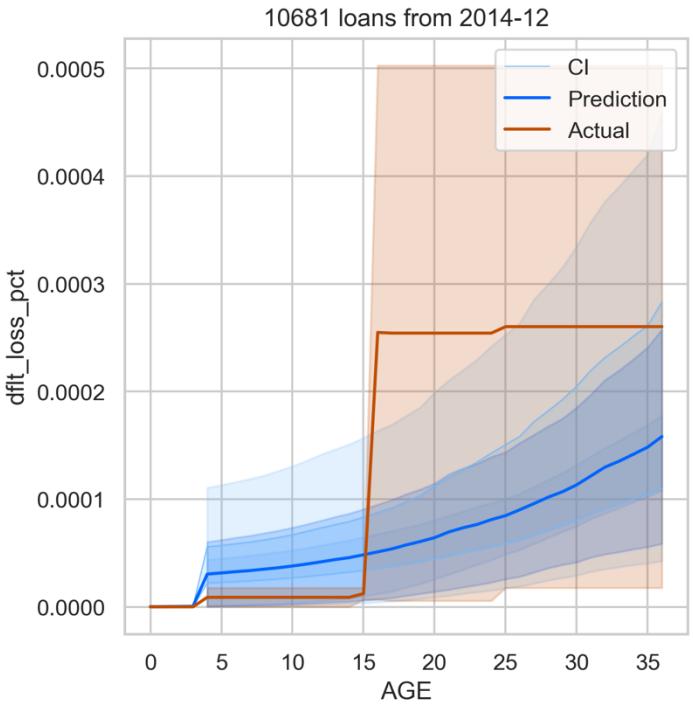

8260 loans from 2010-02
CI
Prediction
Actual
dflt_loss_pct
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
20
40
60
80
AGE

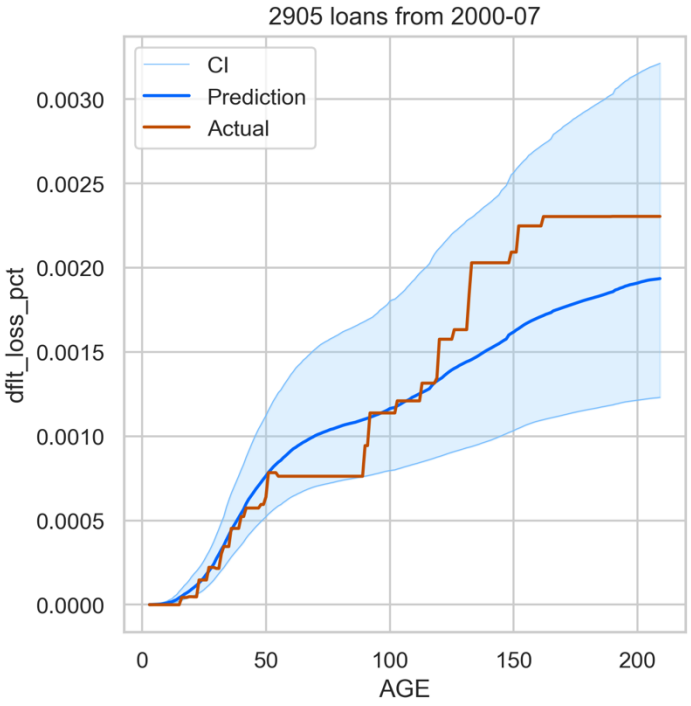

27714 loans from 2013-03
CI
Prediction
Actual
dflt_loss_pct
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
10
20
30
40
50
AGE

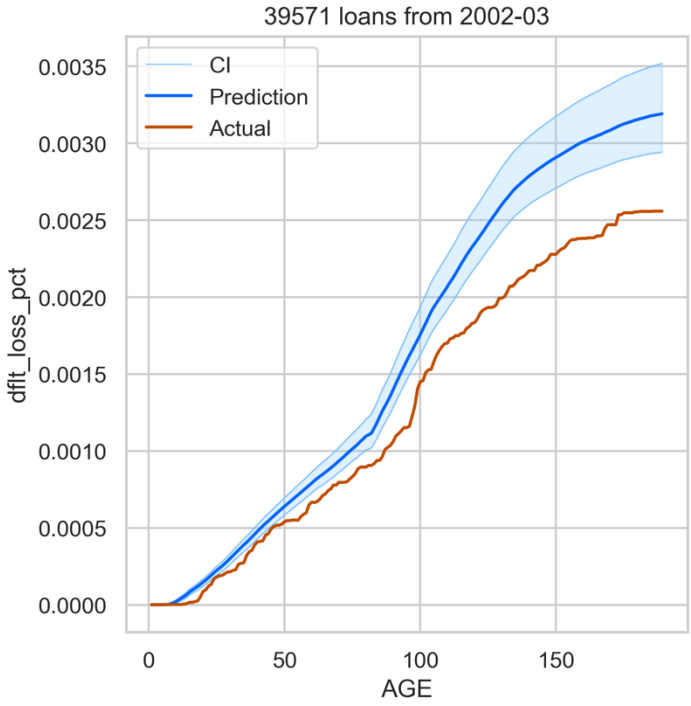

19491 loans from 2003-03
CI
Prediction
Actual
dflt_loss_pct
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
25
50
75
100
125
150
175
AGE

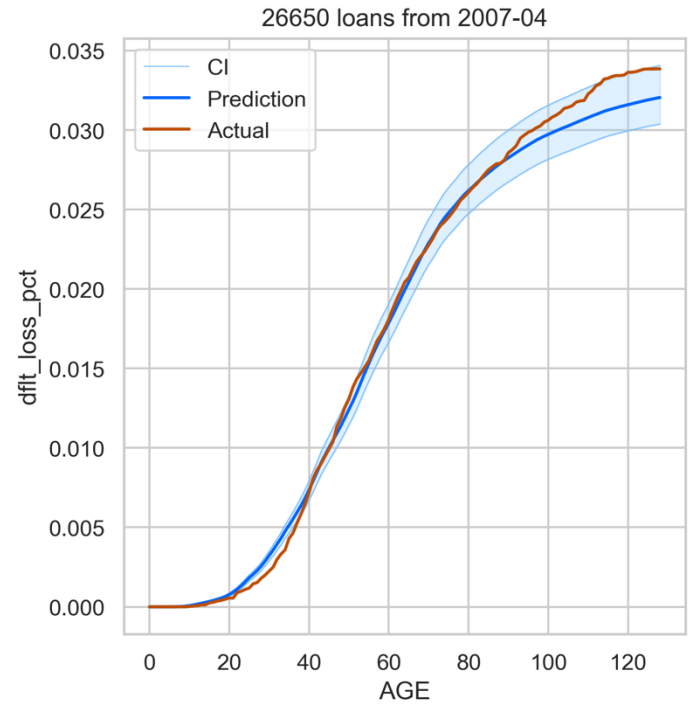

19541 loans from 2004-03
CI
Prediction
Actual
dflt_loss_pct
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
25
50
75
100
125
150
AGE

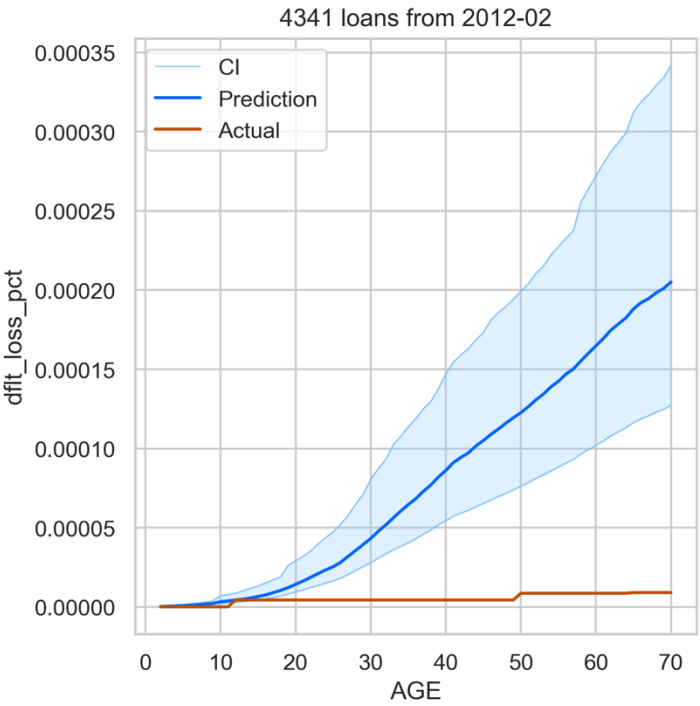

1589 loans from 2006-01
CI
Prediction
Actual
dflt_loss_pct
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
140
AGE

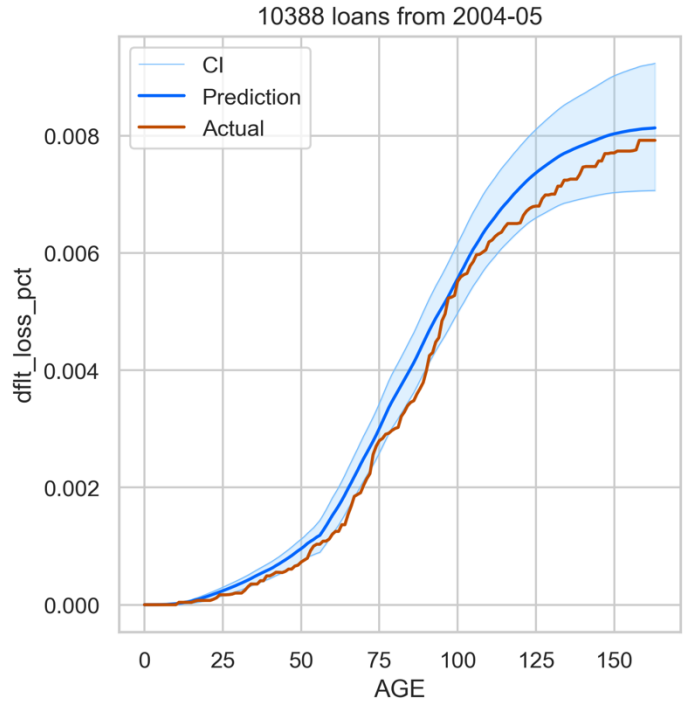

10681 loans from 2014-12
CI
Prediction
Actual
dflt_loss_pct
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
5
10
15
20
25
30
35
AGE

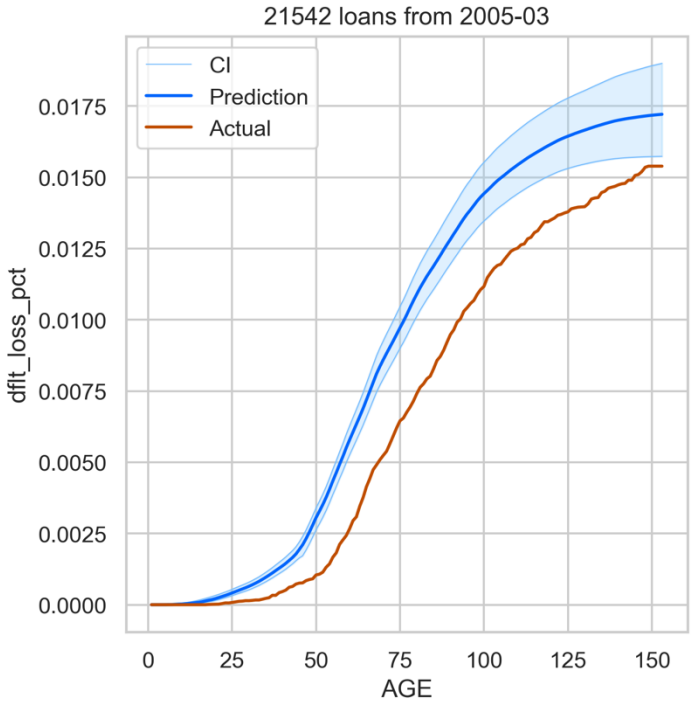

2905 loans from 2000-07
CI
Prediction
Actual
dflt_loss_pct
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
50
100
150
200
AGE

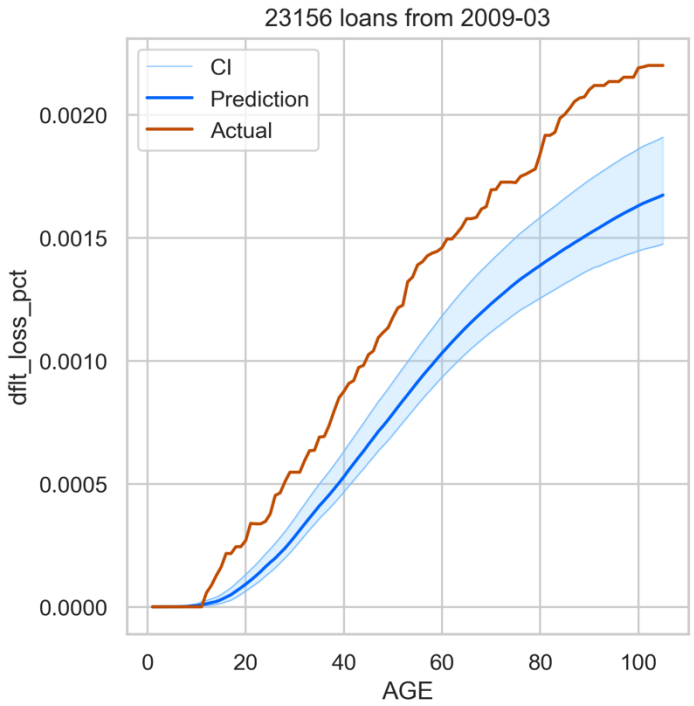

39571 loans from 2002-03
CI
Prediction
Actual
dflt_loss_pct
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
50
100
150
AGE

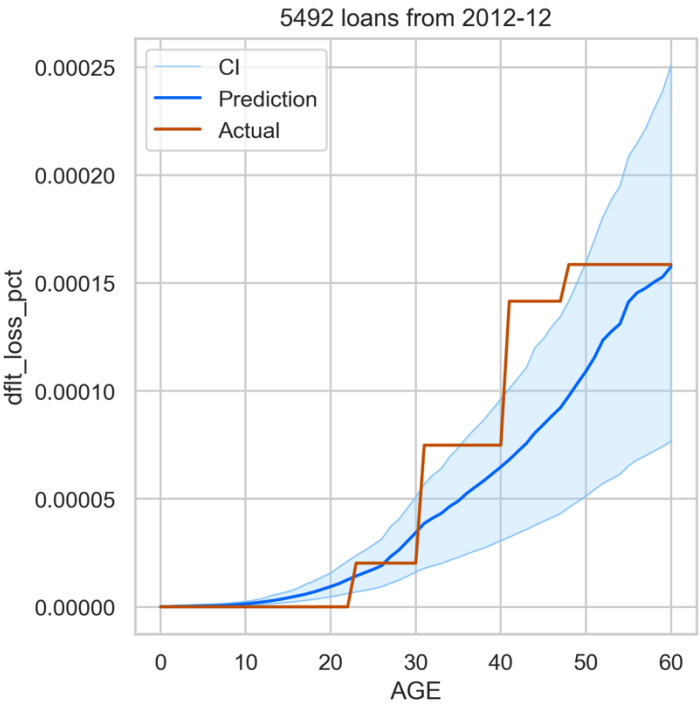

26650 loans from 2007-04
CI
Prediction
Actual
dflt_loss_pct
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
AGE

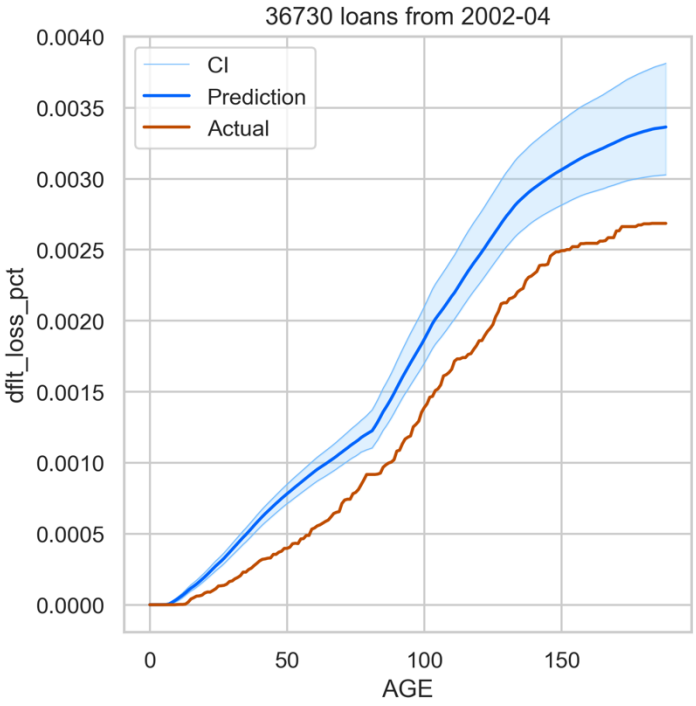

4341 loans from 2012-02
CI
Prediction
Actual
dflt_loss_pct
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
10
20
30
40
50
60
70
AGE

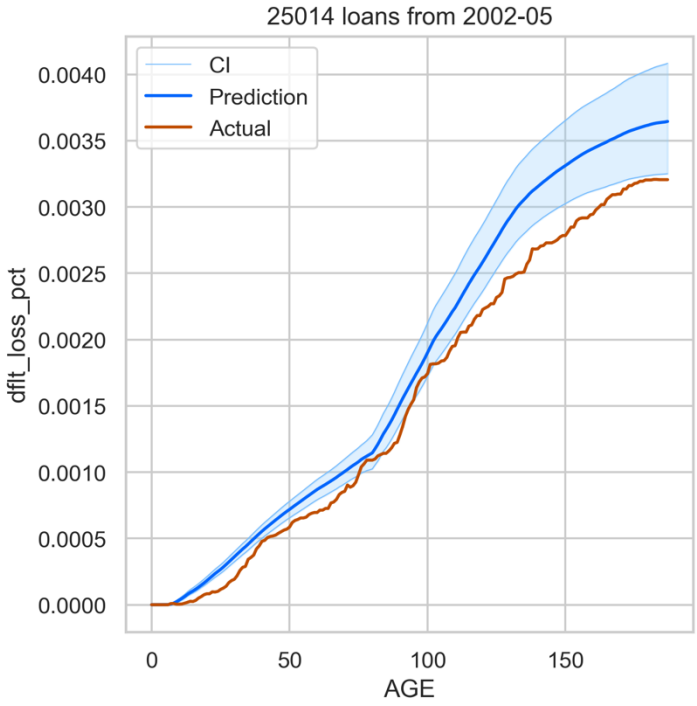

10388 loans from 2004-05
CI
Prediction
Actual
dflt_loss_pct
0.000
0.002
0.004
0.006
0.008
0
25
50
75
100
125
150
AGE

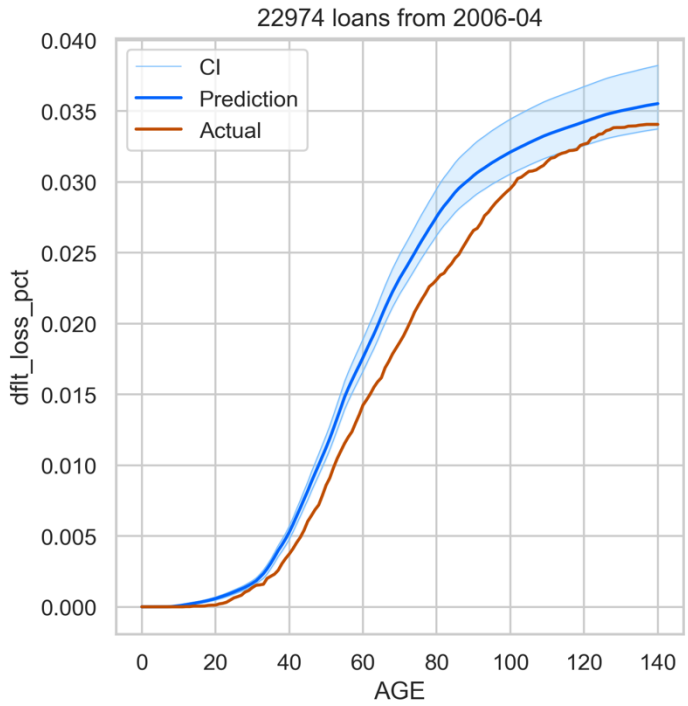

21542 loans from 2005-03
CI
Prediction
Actual
dflt_loss_pct
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0
25
50
75
100
125
150
AGE

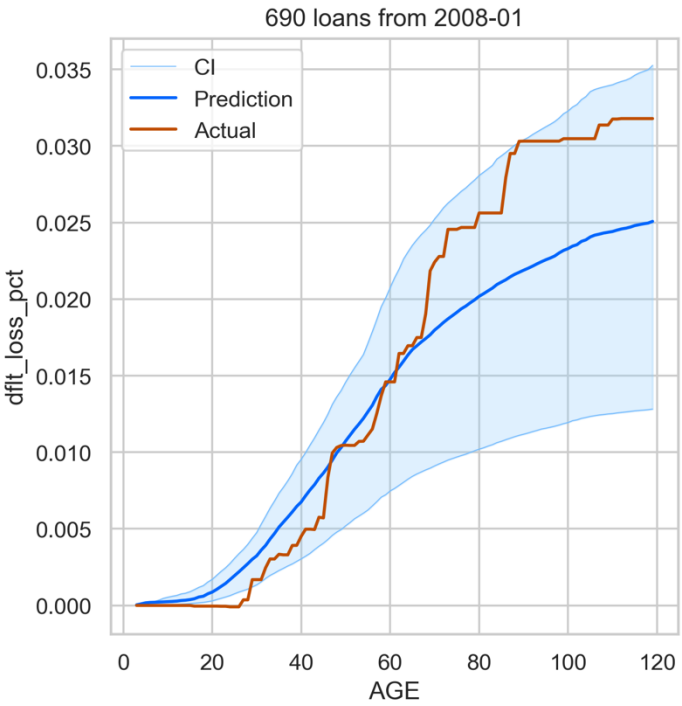

23156 loans from 2009-03
CI
Prediction
Actual
dflt_loss_pct
0.0000
0.0005
0.0010
0.0015
0.0020
0
20
40
60
80
100
AGE


5492 loans from 2012-12
CI
Prediction
Actual
dflt_loss_pct
0.00000
0.00005
0.00010
0.00015
0.00020
0.00025
0
10
20
30
40
50
60
AGE


36730 loans from 2002-04
CI
Prediction
Actual
dflt_loss_pct
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
50
100
150
AGE

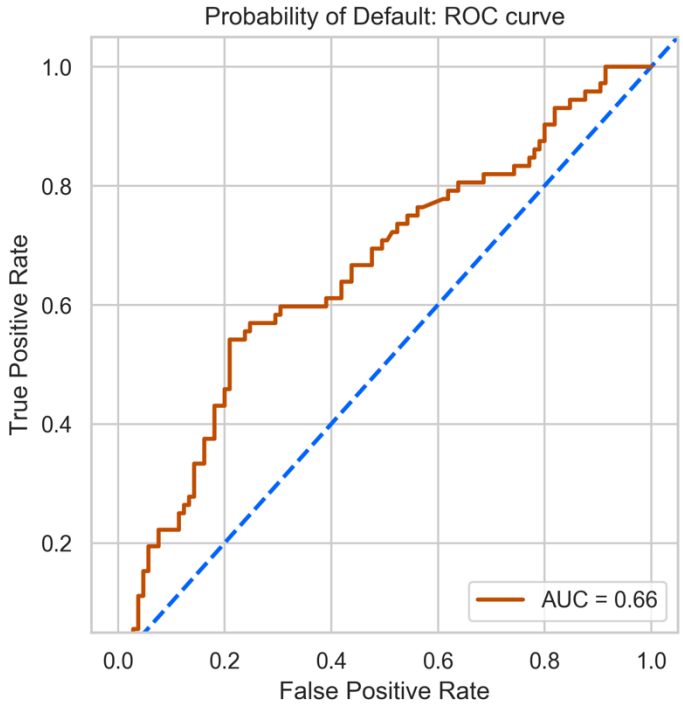

25014 loans from 2002-05
CI
Prediction
Actual
dflt_loss_pct
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
50
100
150
AGE

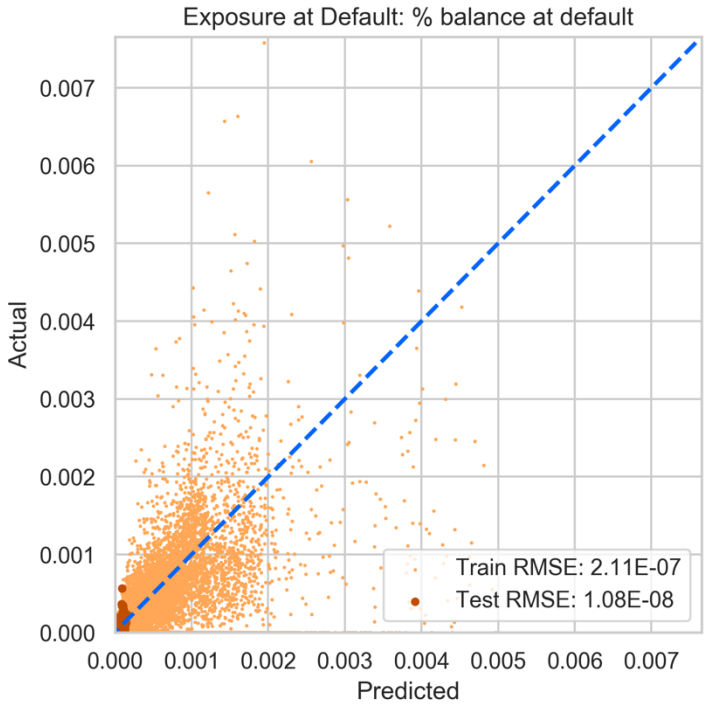

22974 loans from 2006-04
CI
Prediction
Actual
dflt_loss_pct
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
140
AGE

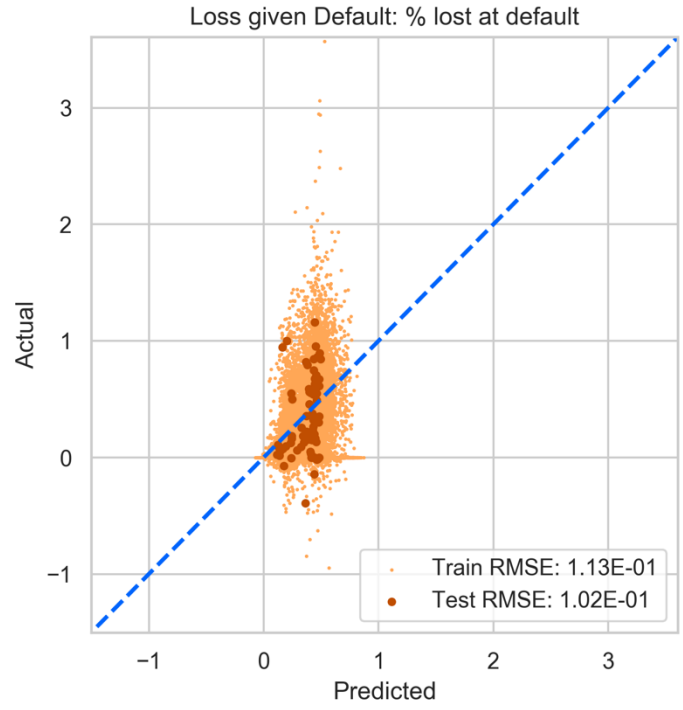

690 loans from 2008-01
CI
Prediction
Actual
dflt_loss_pct
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
AGE

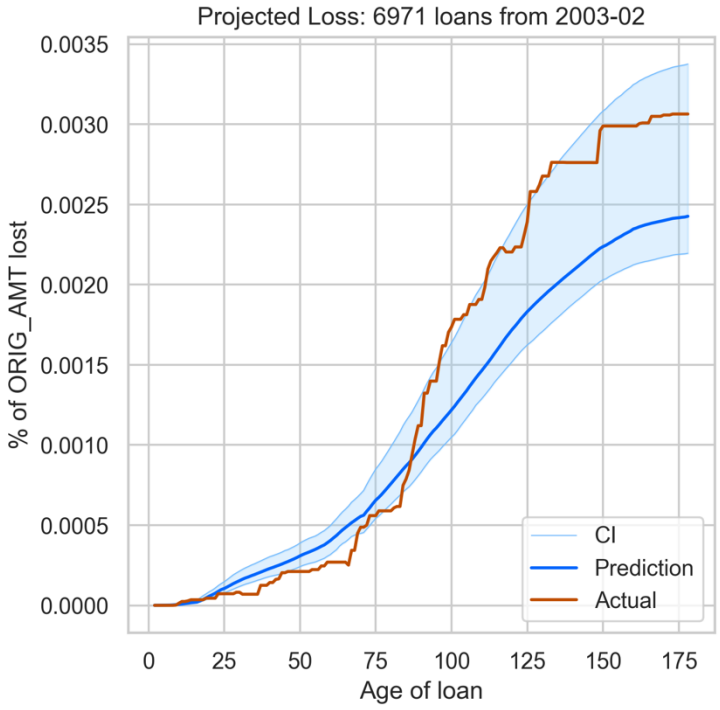
9.3 Model Diagnosis (GBC-OLS-OLS)

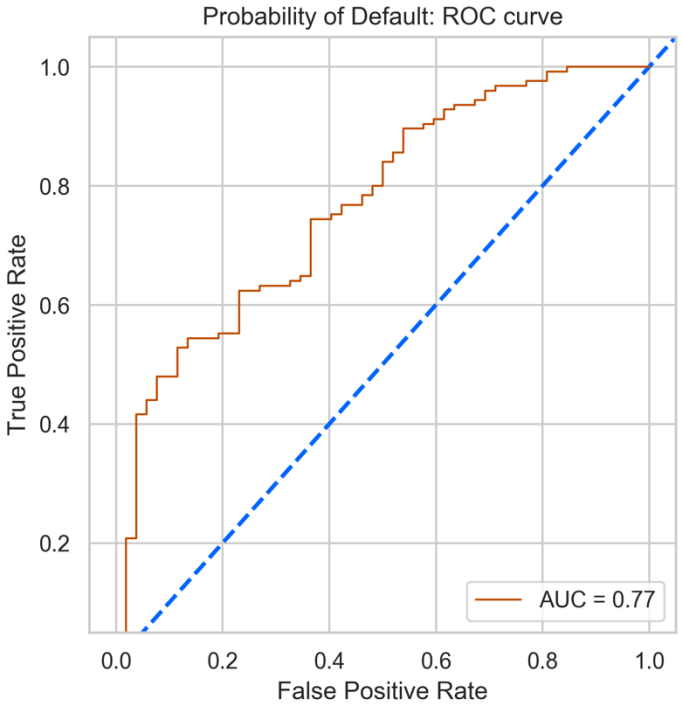
Dfdf

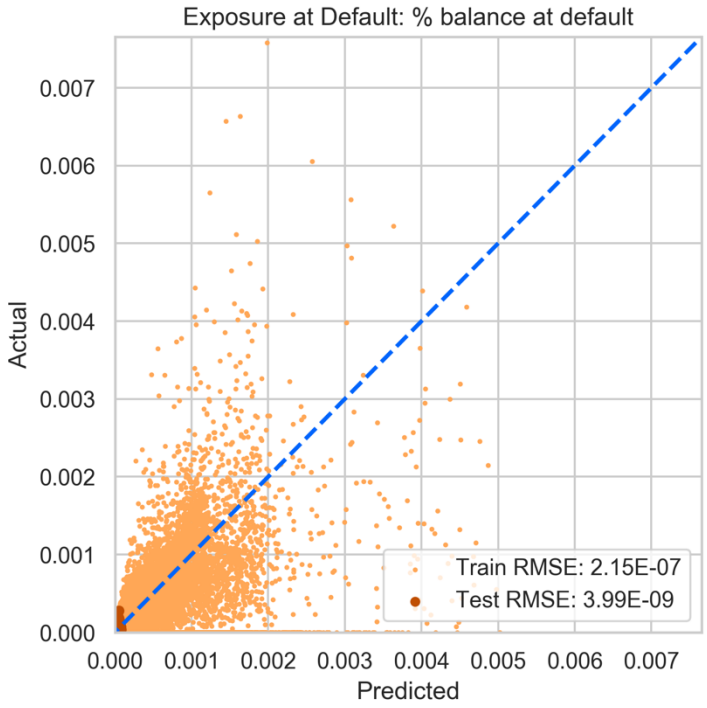


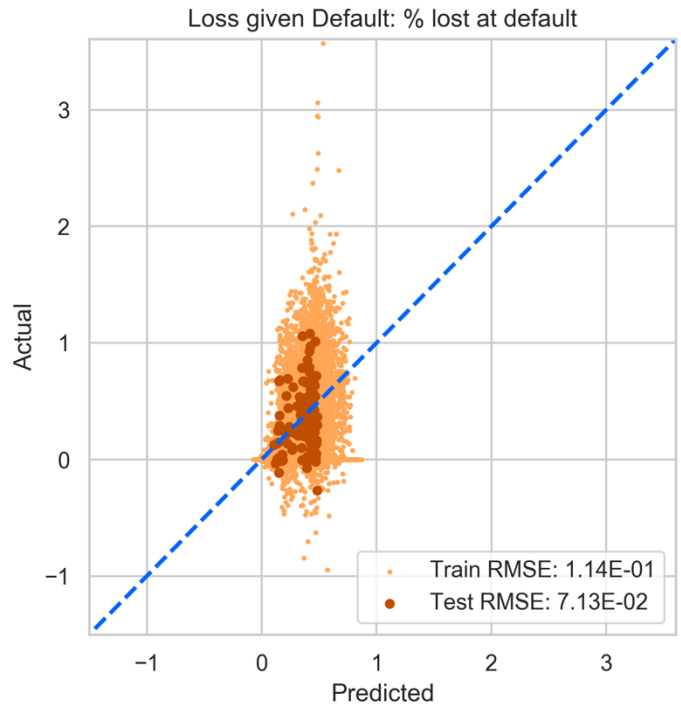


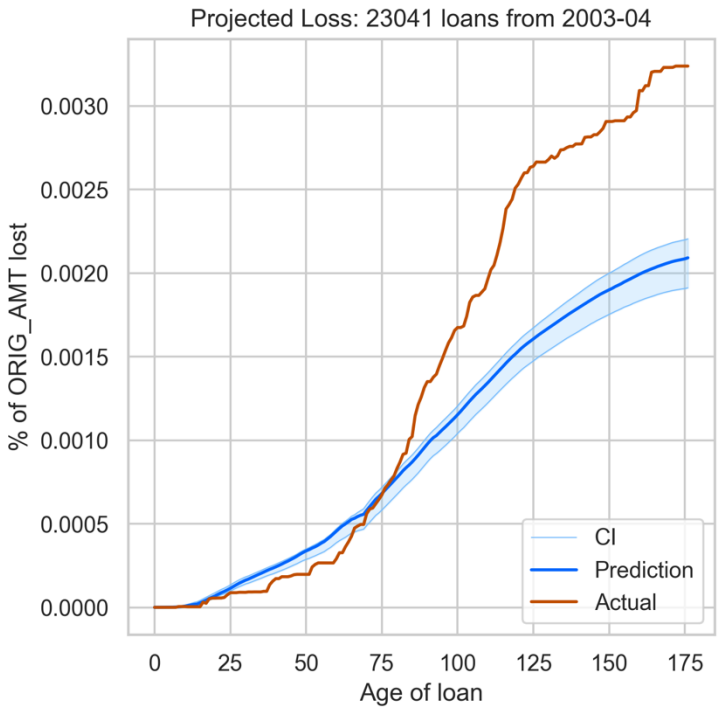





Probability of Default: ROC curve
True Positive Rate
False Positive Rate
AUC = 0.77
0.0
0.2
0.4
0.6
0.8
1.0

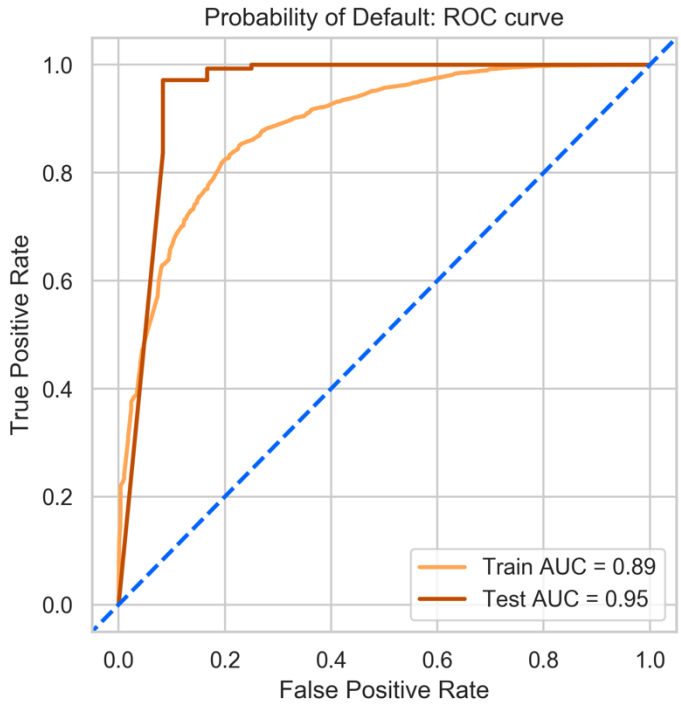

Exposure at Default: % balance at default
Actual
Predicted
Train RMSE: 2.15E-07
Test RMSE: 3.99E-09
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007

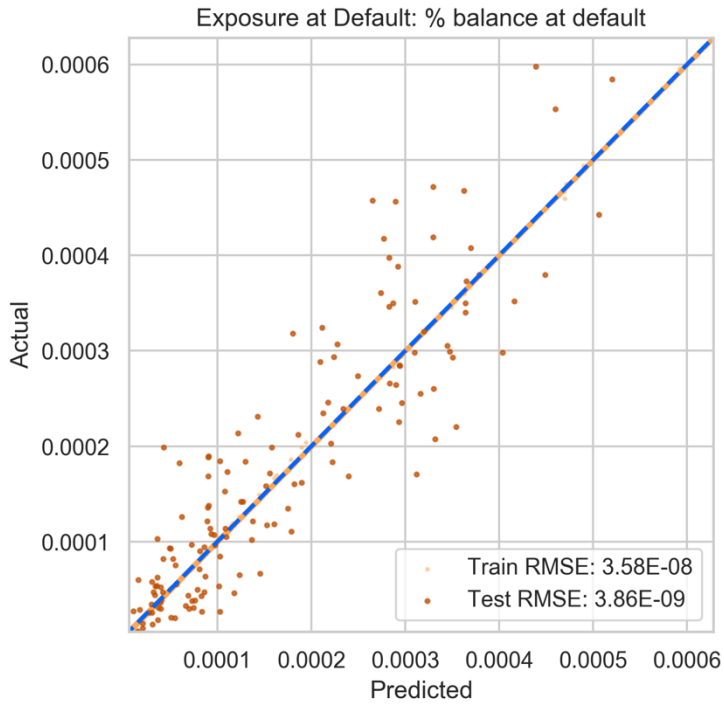

Loss given Default: % lost at default
Actual
Predicted
Train RMSE: 1.14E-01
Test RMSE: 7.13E-02
−1
0
1
2
3

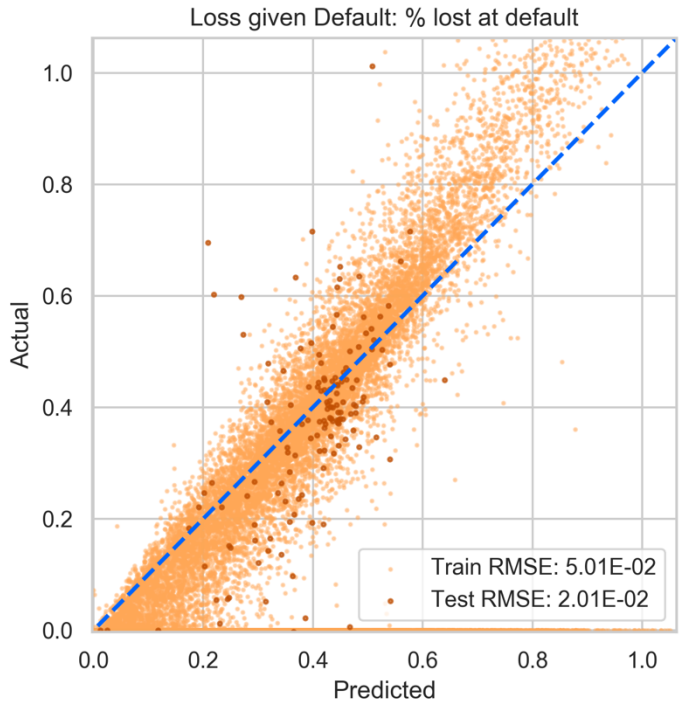

Projected Loss: 23041 loans from 2003-04
% of ORIG_AMT lost
Age of loan
CI
Prediction
Actual
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0
25
50
75
100
125
150
175

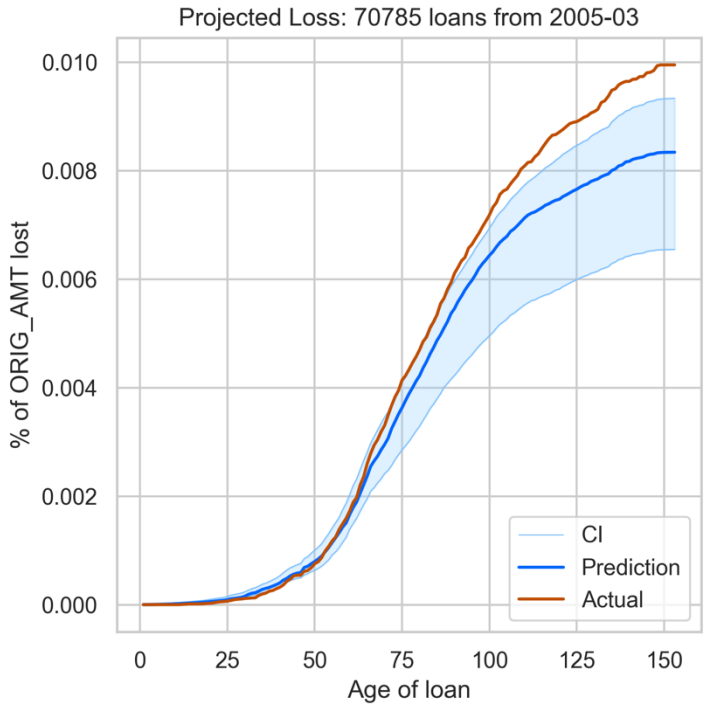
9.4 Model Diagnosis (GBC-RFR-RFR)

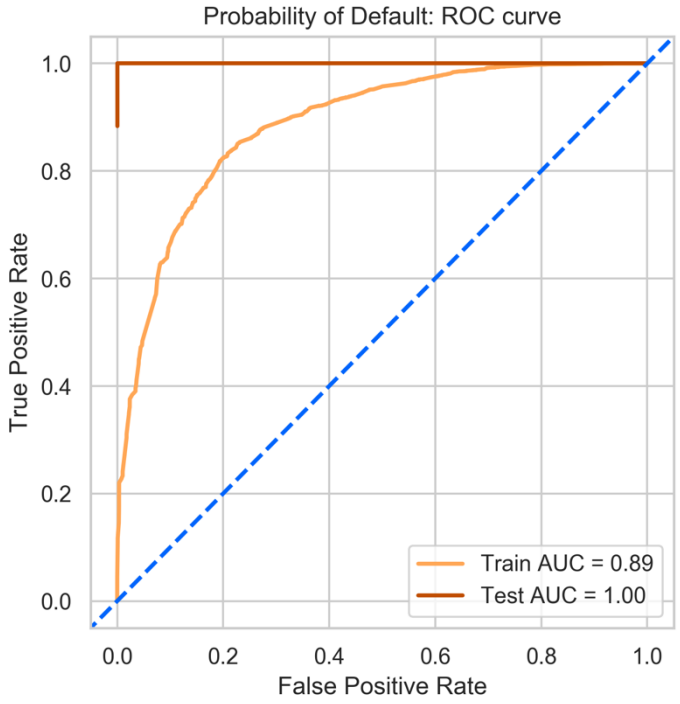
dfds

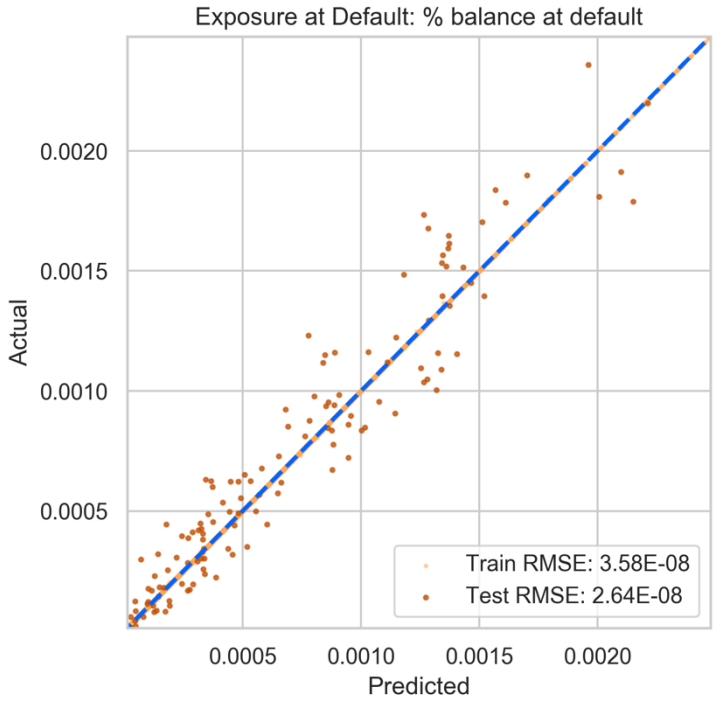


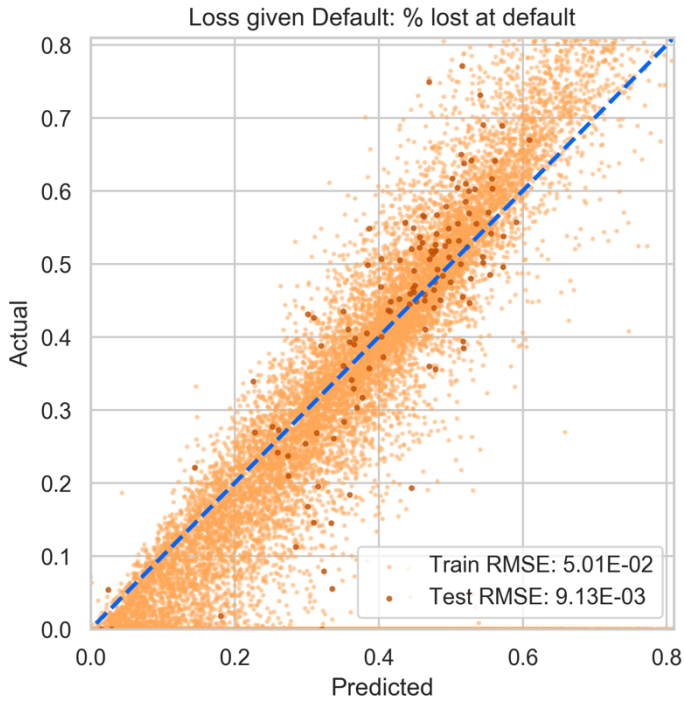


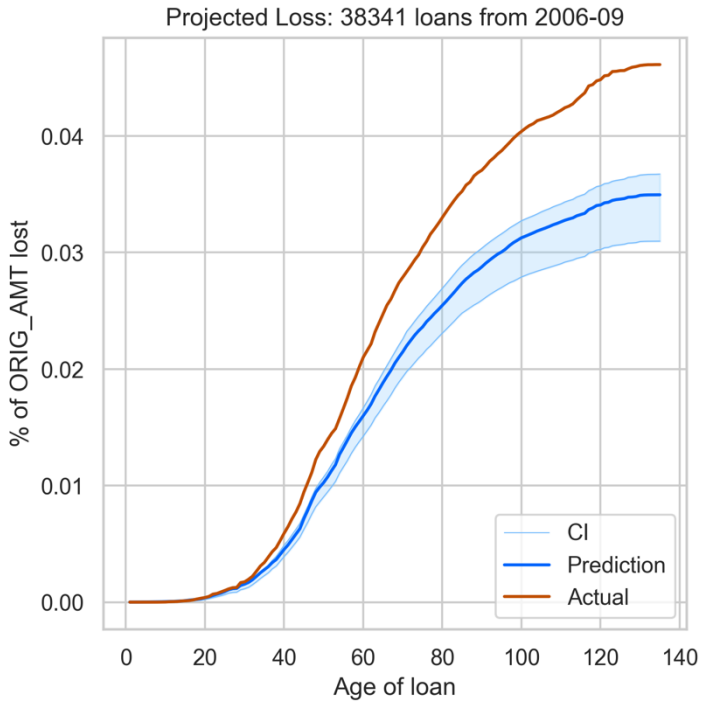


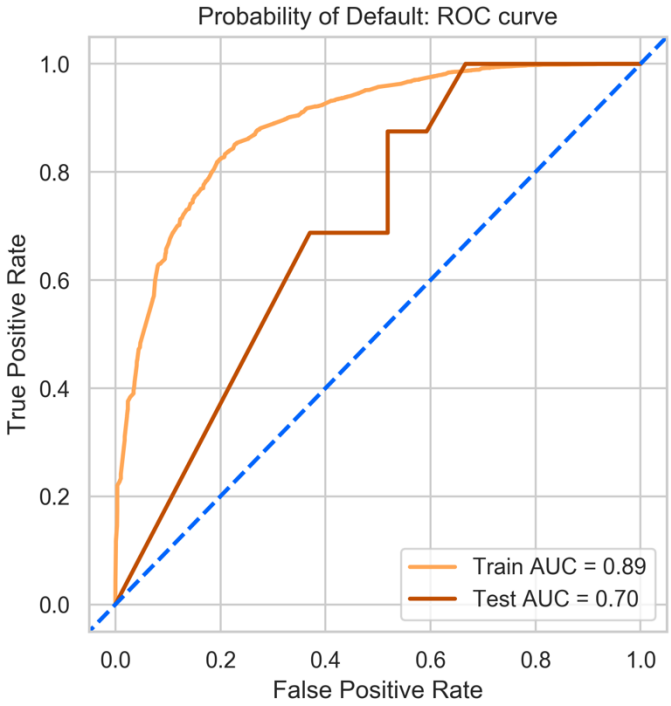


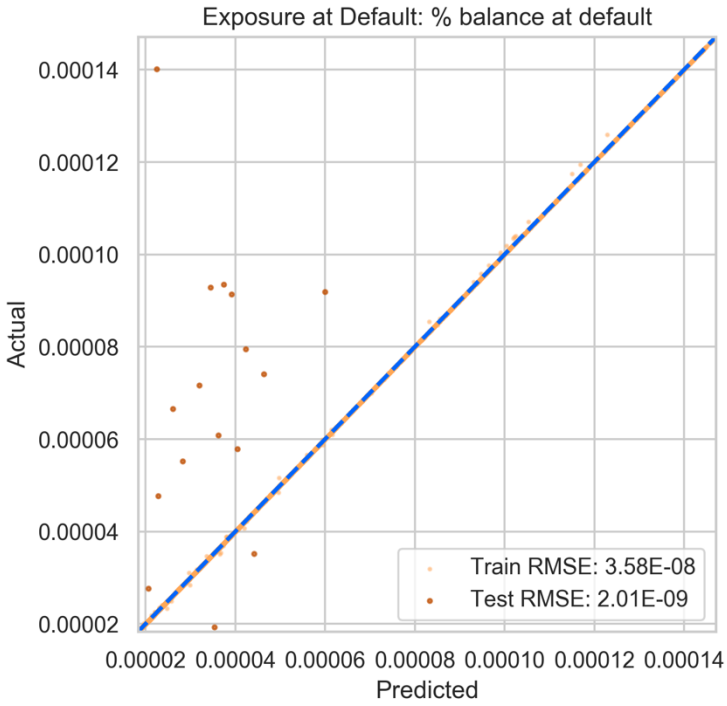

Probability of Default: ROC curve
True Positive Rate
False Positive Rate
Train AUC = 0.89
Test AUC = 1.00

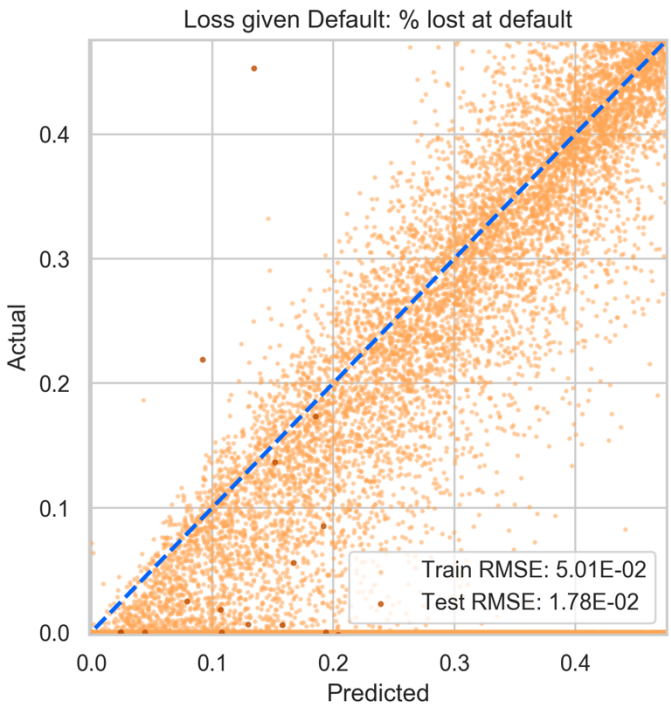

Exposure at Default: % balance at default
Actual
Predicted
Train RMSE: 3.58E-08
Test RMSE: 2.64E-08

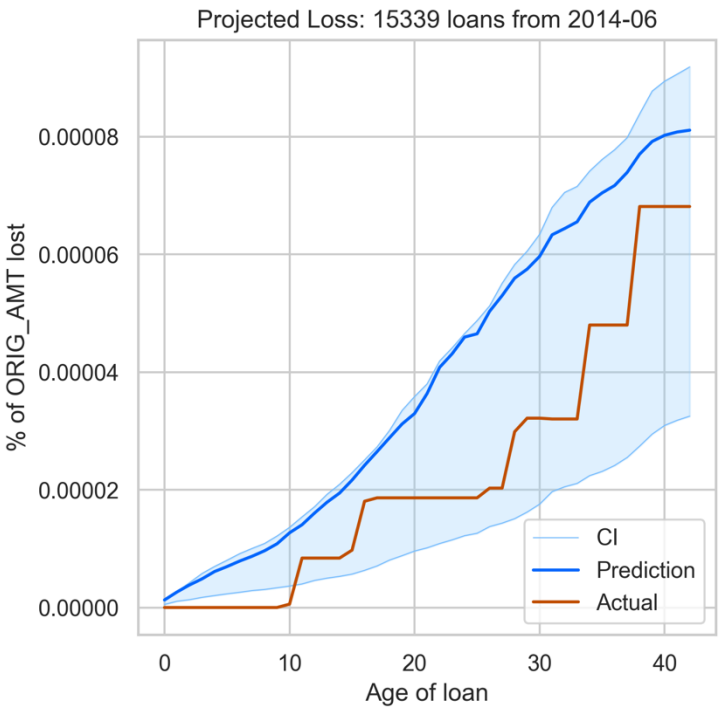

Loss given Default: % lost at default
Actual
Predicted
Train RMSE: 5.01E-02
Test RMSE: 9.13E-03

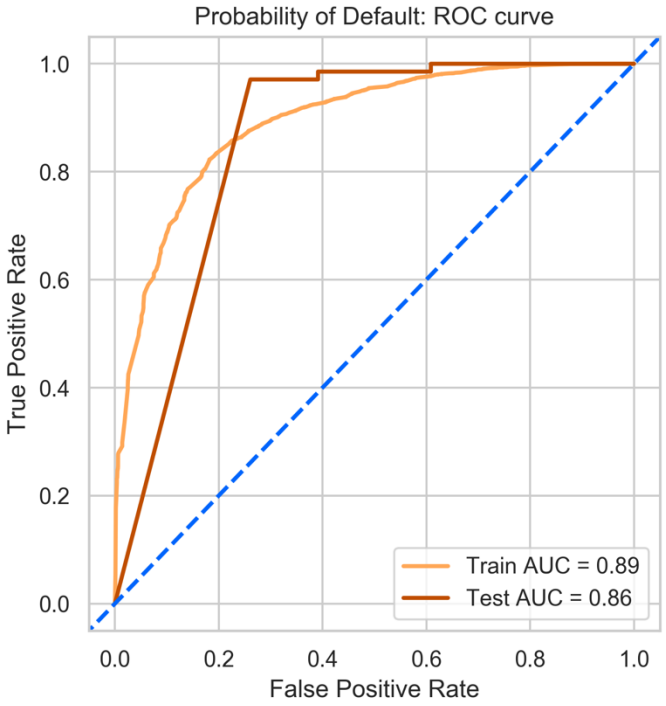

Projected Loss: 38341 loans from 2006-09
% of ORIG_AMT lost
Age of loan
CI
Prediction
Actual

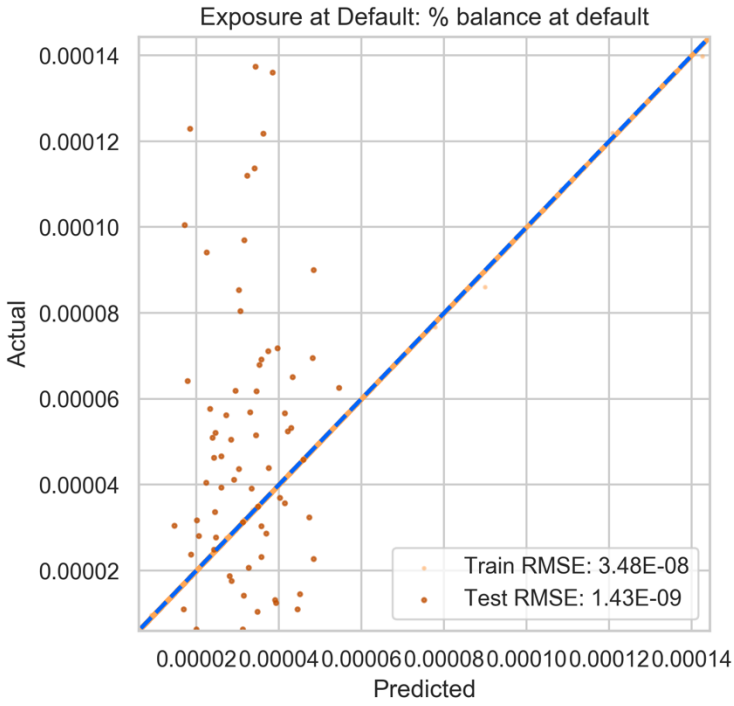

Probability of Default: ROC curve
True Positive Rate
False Positive Rate
Train AUC = 0.89
Test AUC = 0.70

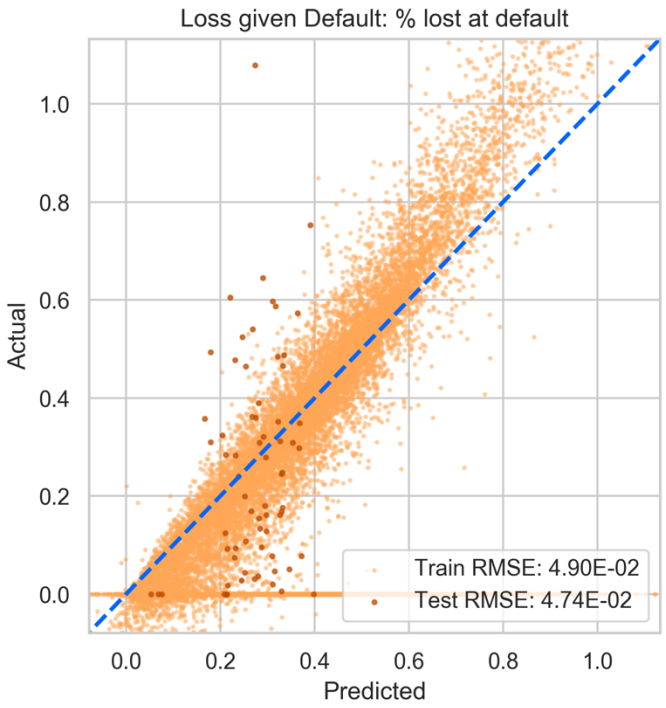

Exposure at Default: % balance at default
Actual
Predicted
Train RMSE: 3.58E-08
Test RMSE: 2.01E-09

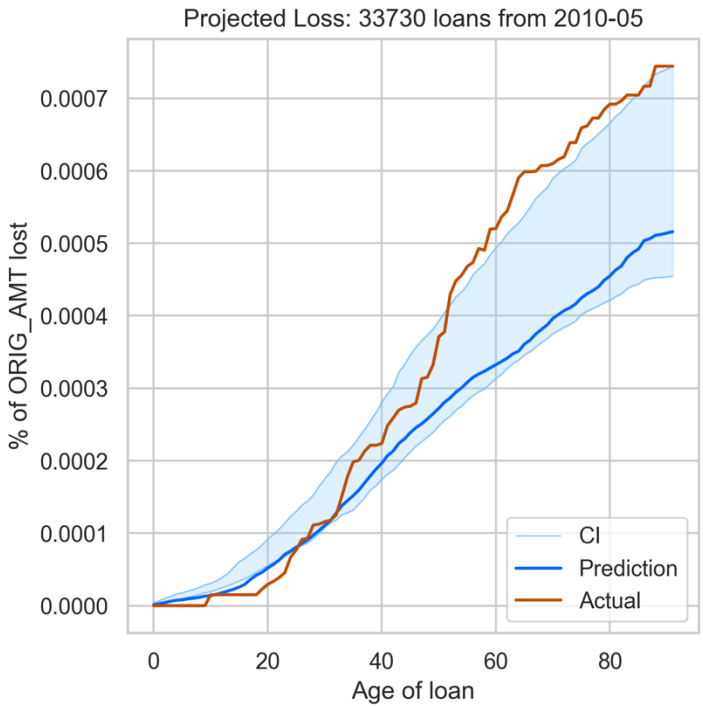

Loss given Default: % lost at default
Actual
Predicted
Train RMSE: 5.01E-02
Test RMSE: 1.78E-02

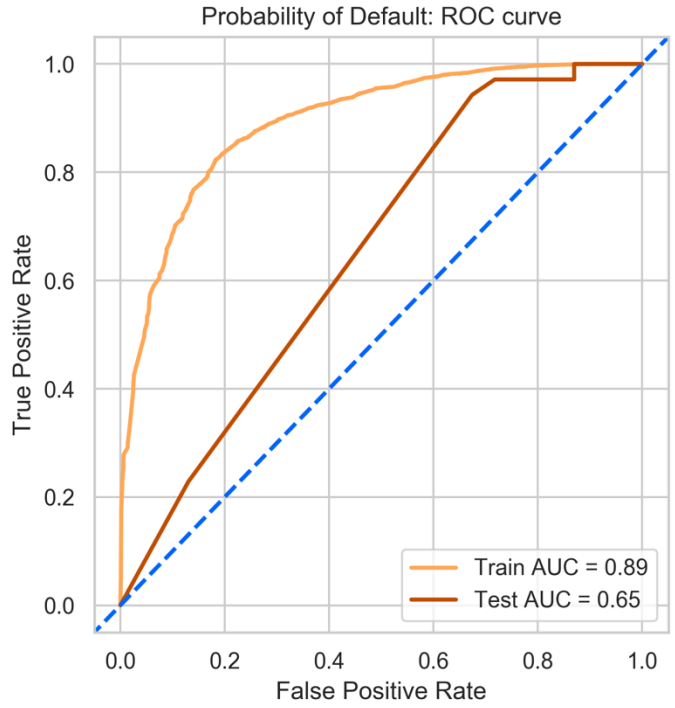

Projected Loss: 15339 loans from 2014-06
% of ORIG_AMT lost
Age of loan
CI
Prediction
Actual

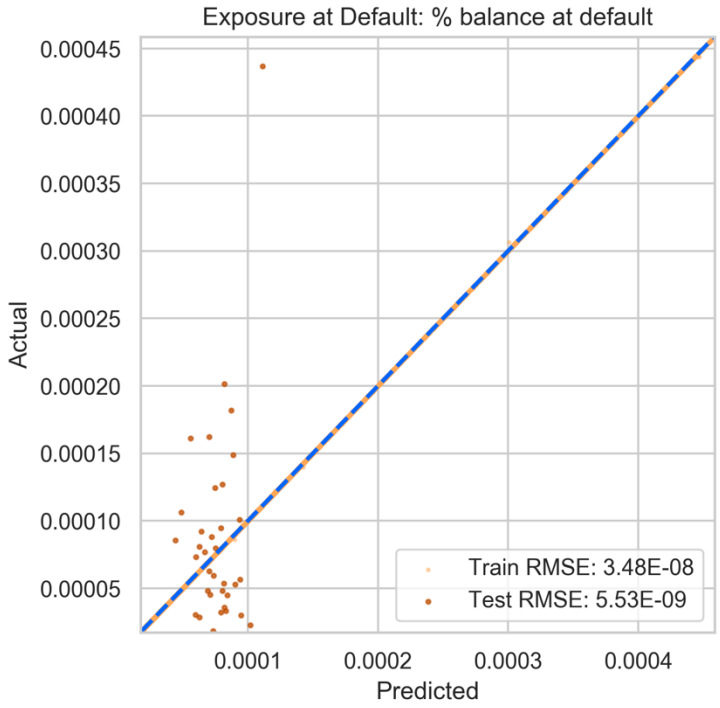

Probability of Default: ROC curve
True Positive Rate
False Positive Rate
Train AUC = 0.89
Test AUC = 0.86

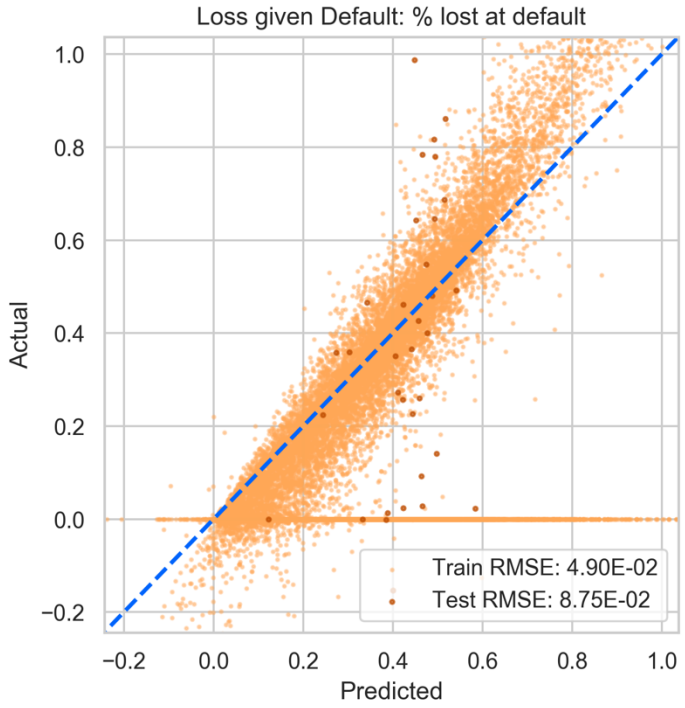

Exposure at Default: % balance at default
Actual
Predicted
Train RMSE: 3.48E-08
Test RMSE: 1.43E-09

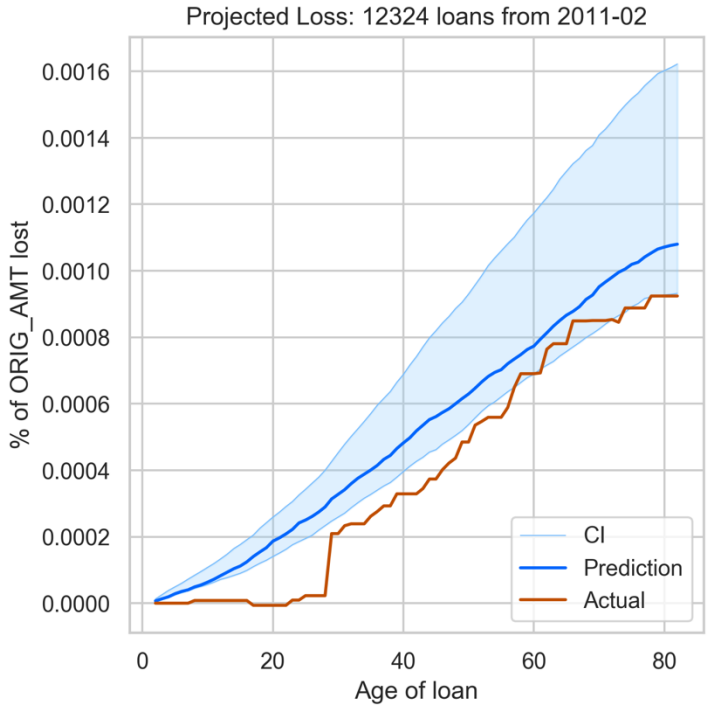

Loss given Default: % lost at default
Actual
Predicted
Train RMSE: 4.90E-02
Test RMSE: 4.74E-02

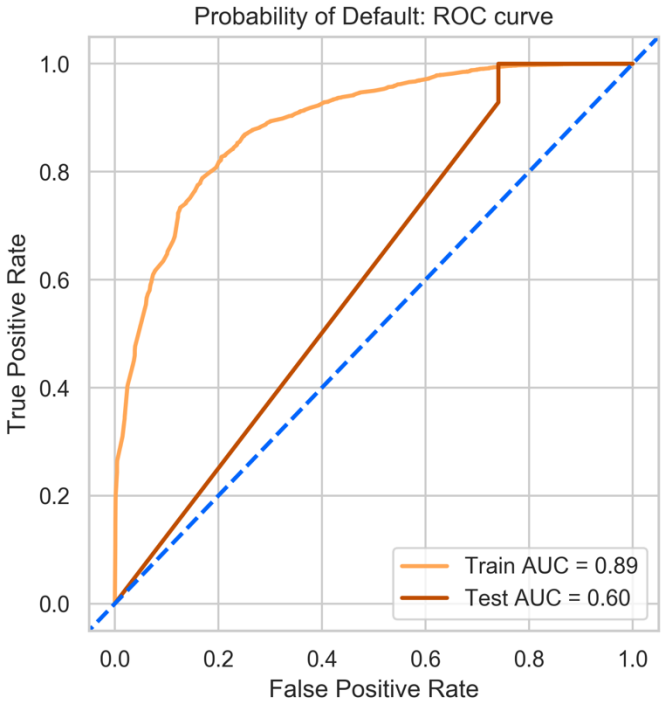

Projected Loss: 33730 loans from 2010-05
% of ORIG_AMT lost
Age of loan
CI
Prediction
Actual

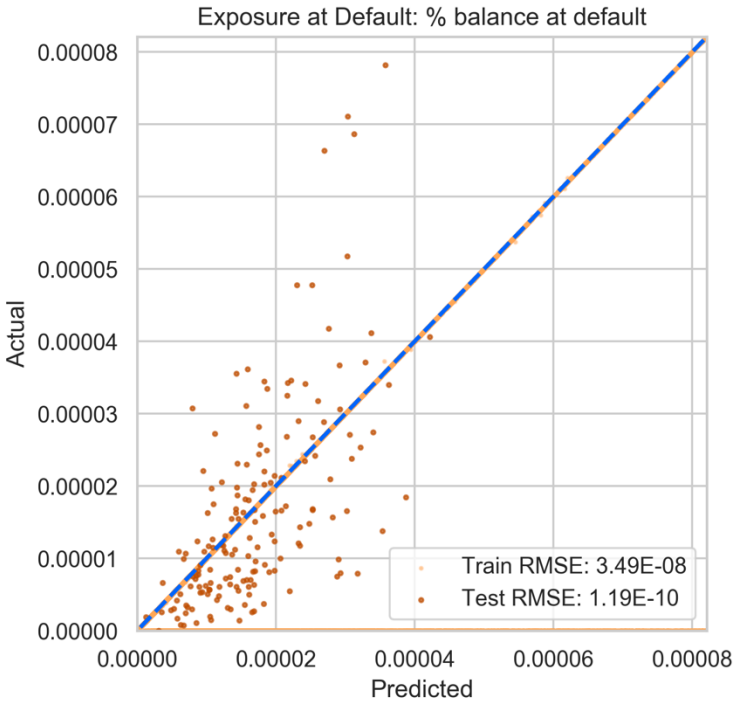

Probability of Default: ROC curve
True Positive Rate
False Positive Rate
Train AUC = 0.89
Test AUC = 0.65

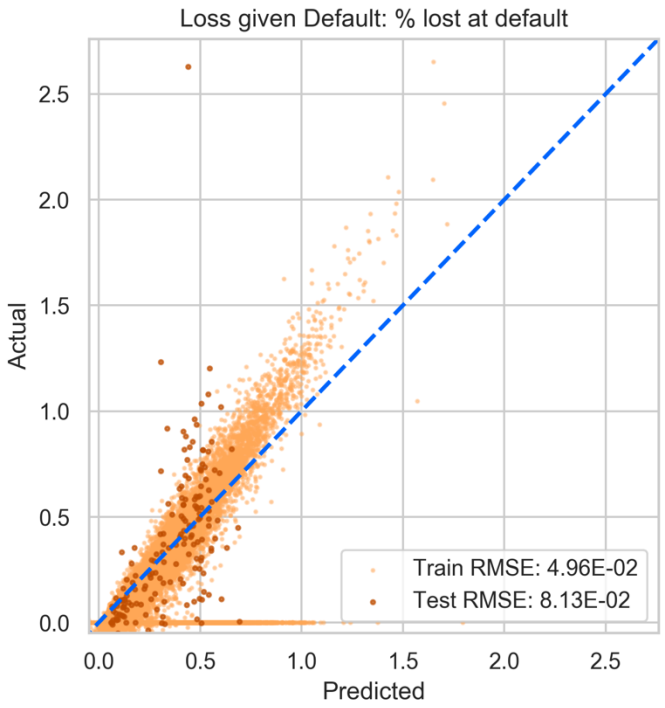

Exposure at Default: % balance at default
Actual
Predicted
Train RMSE: 3.48E-08
Test RMSE: 5.53E-09

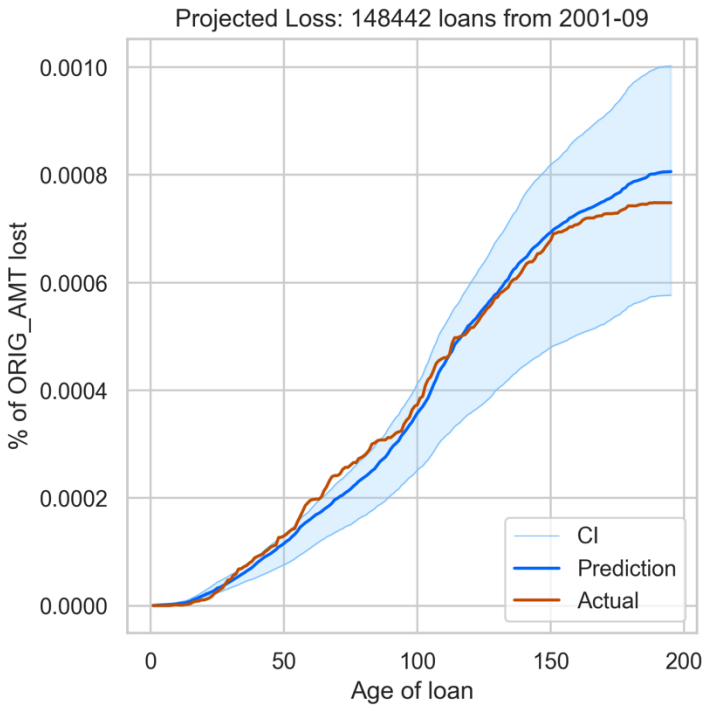

Loss given Default: % lost at default
Actual
Predicted
Train RMSE: 4.90E-02
Test RMSE: 8.75E-02


Projected Loss: 12324 loans from 2011-02
% of ORIG_AMT lost
Age of loan
CI
Prediction
Actual









10 Next steps

	Probability of Default_t P(loan defaults in month t)	Exposure at Default_t UPB after dflt in month t	Loss given Default_t Net loss after disposal in t
Vintage	GBM	LM	LM
Individual	Cox?		

References

Leathwick, J. R., Elith, J., Francis, M. P., Hastie, T., & Taylor, P. (2006). Variation in demersal fish species richness in the oceans surrounding new zealand: An analysis using boosted regression trees. *Mar Ecol Prog Ser, 321*, 267-281. Retrieved from https://www.int-res.com/abstracts/meps/v321/p267-281/