UNDERWRITING SCORECARD BASED ON FINANCIAL MODELING

JIAQI.CHEN

09/04/2015

RESEARCH PURPOSE AND BACKGROUND

- Data source: Freddy Mac & Fannie Mae
- Constructing a financial model by given data
- Predicting default rate via various features
- Create underwriting scorecard by actural default rate and forecasted default rate
- Data duration: 1999~2015
- Data volume: more than 20,000,000 loan level data

DATA DESCRIPTION

- FREDDY MAC & FANNIE MAE

- ORIGINATION DATA FILE & PERFORMANCE DATA FILE

Dataset	File Name Format	Contents	File Type	Delimiter
Full	historical_data1_QnYYYY.zip	historical_data1_QnYYYY.txt	Origination Data	Pipe ("|")
		historical_data1_time_QnYYYY.txt	Monthly Performance Data	

- -separated in to 4 quarters, such as: Q1,Q2,Q3,Q4
- -origination data: containing loan level origination information for all loans originated in that quarter.
- -performance data: containing monthly loan-level credit score performance and actual loss for each loan

FEATURES SELECTED

- Origination data
- -continuous features: 1.debit to income ratio (DTI) from 0~65%
  2.loan to value ratio (LTV) from 0~100%
  3.credit score (FICO) from 301~850
  4. original unpaid principal balance (UPB) from 0~1,000,000
  -leveled features： 1. first time homebuyer flag ‘Y’=yes, ‘N’=no
  2. loan purpose ‘P’=purchase, ‘C’=cash-out refinance, ‘N’=no cash-out refinance

- Performance data
- -features: 1.delinquency status
  2. ages

DEFAULT IDENTIFICATION

- - PERFORMANCE DATA
- 1.Filter data by age: 0~48 2.Identify loan's delinquency status>=3 as default 3. Record the default ID.
- -ORIGINATION DATA
- -Merge the default flag('0', '1') by ID with the origination dataset.

Formal Name	Loan Sequence Number	Monthly Reporting Period	Current Actual UPB	Current Loan Delinquency Status	Loan Age
Monthly Performance Data	F108Q4000374	201004	79930.72	0	16
	F108Q4000374	201005	79844.71	0	17
	F108Q4000374	201006	79844.71	1	18
	F108Q4000374	201007	79844.71	2	19
	F108Q4000374	201008	79844.71	3	20
	F108Q4000374	201009	79844.71	4	21
	F108Q4000374	201010	79844.71	5	22
	F108Q4000374	201011	79844.71	6	23
	F108Q4000374	201012	0	7	24

ID	DTI	LTV	FICO	LOAN PURPOSE	FHBF	Default
F108Q4000374	80	65	630	P	Y	1

SAMPLING AND WEIGHTING

- Purpose: Increase the efficiency of data processing and prevent incident that the financial model neglects the correct rate of default sample's prediction.
- -Sampling 50% of the default data from the origination data set.
- -Sampling 5% of the non-default data from the origination data set.
- Weight on samples:
- Default loan : 2
- Non-default loan: 20

ID	DTI	LTV	FICO	Loan purpose	First time homebuyer flag	Default	weight
F108Q4000374	57	80	630	N	Y	1	2
F108Q4000676	32	54	800	C	N	0	20
F108Q4000325	43	34	760	P	N	0	20
F108Q4000879	20	40	750	C	Y	0	20

LOGISTIC REGRESSION ANALYSIS

- Model selection: Logistic regression;
- Dependent variable: Default
- Independent variable : DTI, LTV, FICO, original UPB, Loan Purpose, First time homebuyer flag.
- Purpose: forecasting the default rate of loans.
- Model coefficients:
- Fannie Mae

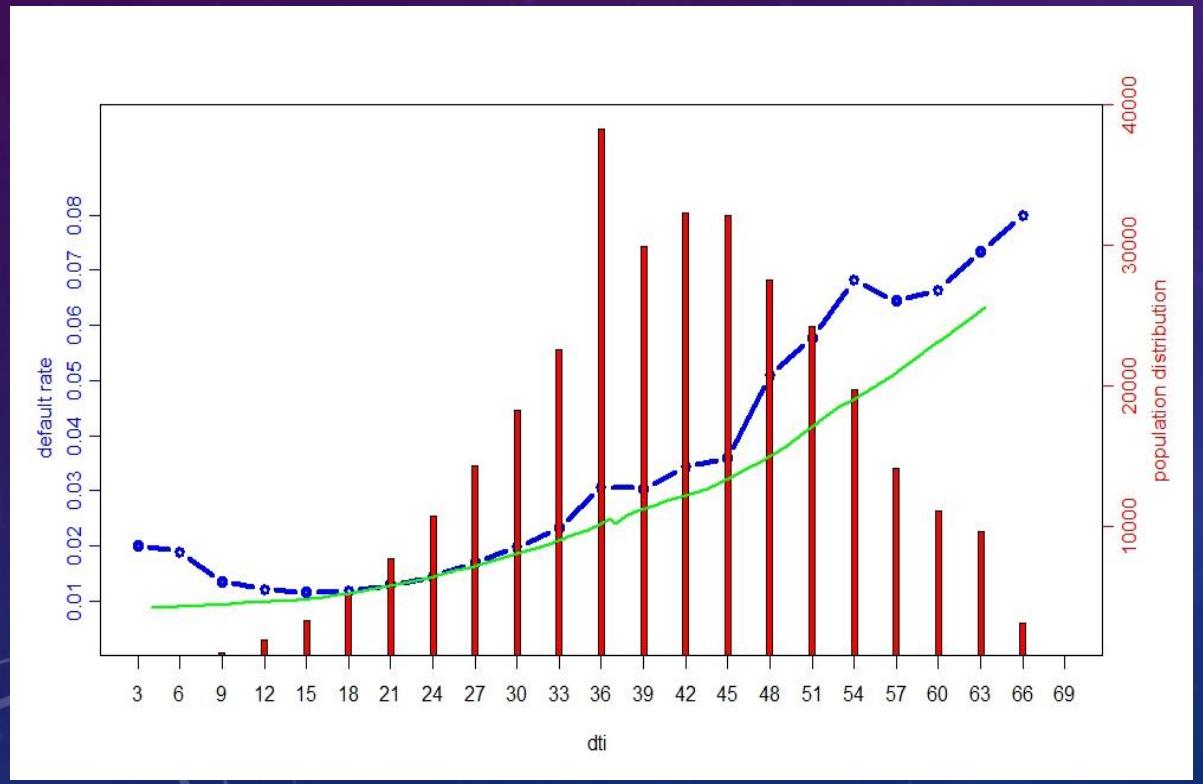
```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          3.762e+00  1.787e-02  210.60   <2e-16 ***
dataset1$fico       -1.499e-02  2.215e-05 -676.75   <2e-16 ***
dataset1$homefirstY  8.182e-02  4.698e-03   17.41   <2e-16 ***
dataset1$dti         3.471e-02  1.149e-04  302.15   <2e-16 ***
dataset1$upb         9.123e-07  1.377e-08   66.25   <2e-16 ***
dataset1$ltv         3.107e-02  1.147e-04  270.98   <2e-16 ***
dataset1$loanpurposeP -7.508e-01  3.663e-03 -204.99   <2e-16 ***
dataset1$loanpurposeR -5.003e-01  3.371e-03 -148.41   <2e-16 ***
dataset1$loanpurposeU -1.314e+00  7.505e-02  -17.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Freddy Mac

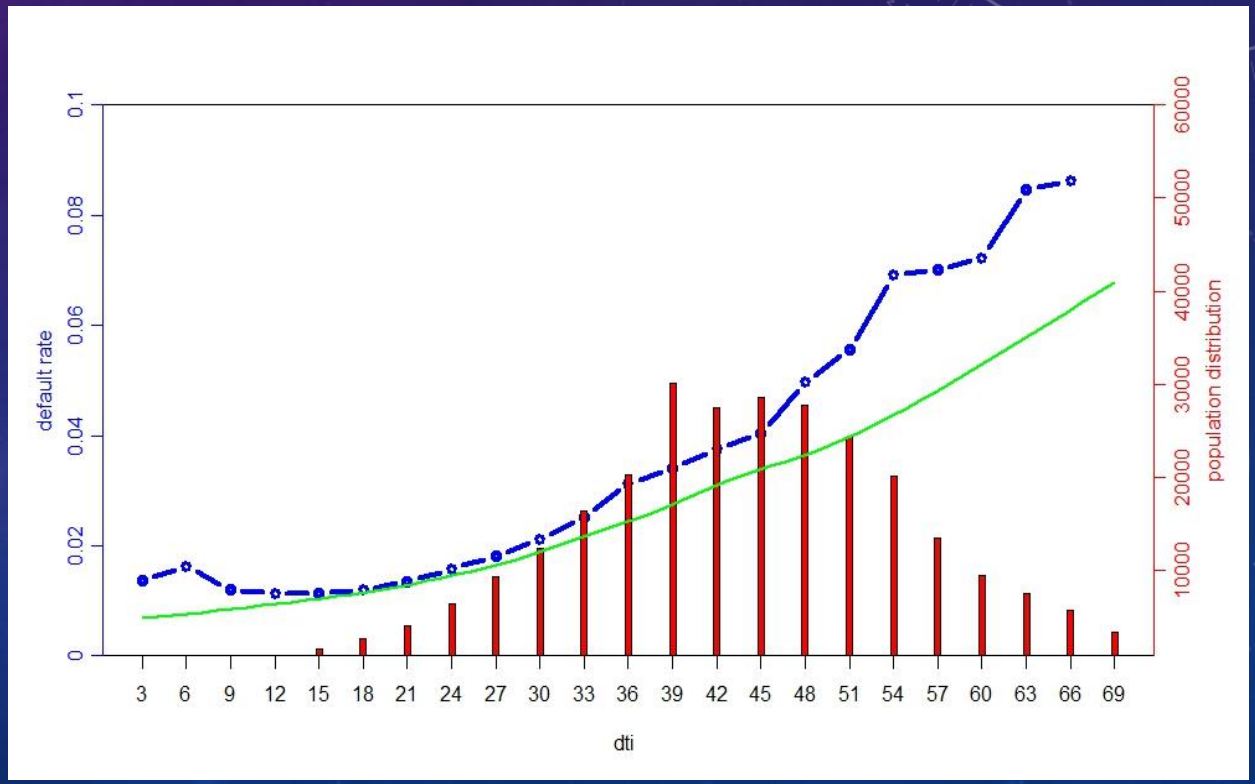
```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          2.818e+00  2.131e-02  132.21   <2e-16 ***
dataset1$fico       -1.427e-02  2.523e-05 -565.69   <2e-16 ***
dataset1$homefirstY  5.979e-02  5.061e-03   11.81   <2e-16 ***
dataset1$dti         3.622e-02  1.330e-04  272.32   <2e-16 ***
dataset1$upb         1.302e-06  1.563e-08   83.32   <2e-16 ***
dataset1$ltv         3.540e-02  1.310e-04  270.15   <2e-16 ***
dataset1$loanpurposeN -5.422e-01  3.711e-03 -146.11   <2e-16 ***
dataset1$loanpurposeP -7.870e-01  3.912e-03 -201.20   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

COMPARISON BETWEEN ACTURAL DEFAULT RATE AND FOARCASTED DEFAULT RATE: DTI

- Fannie Mae

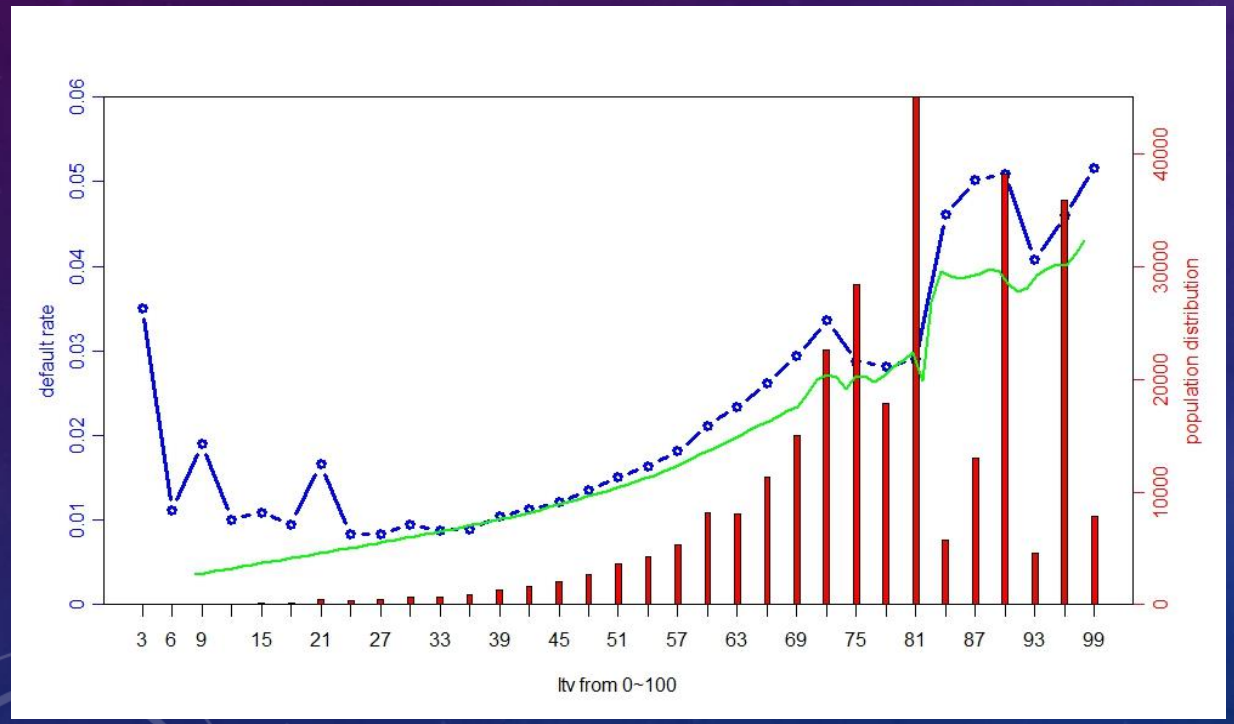


Freddy mac

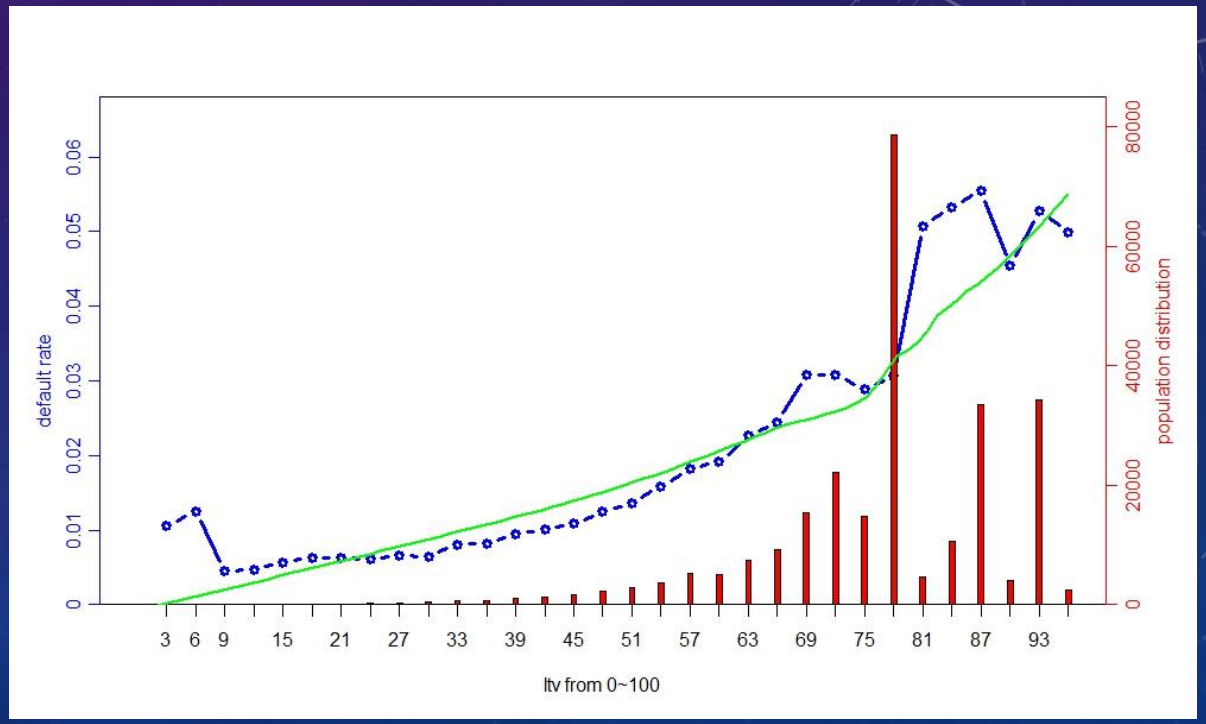


LTV

- Fannie Mae

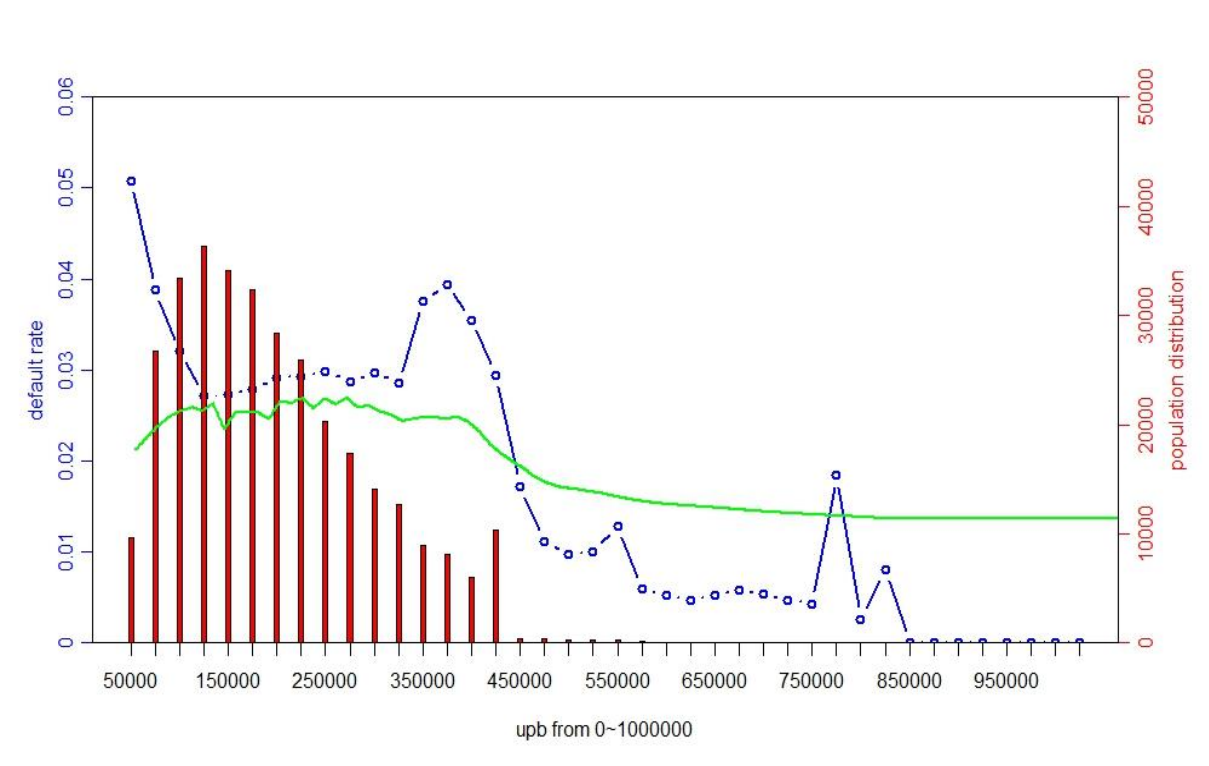


Freddy mac

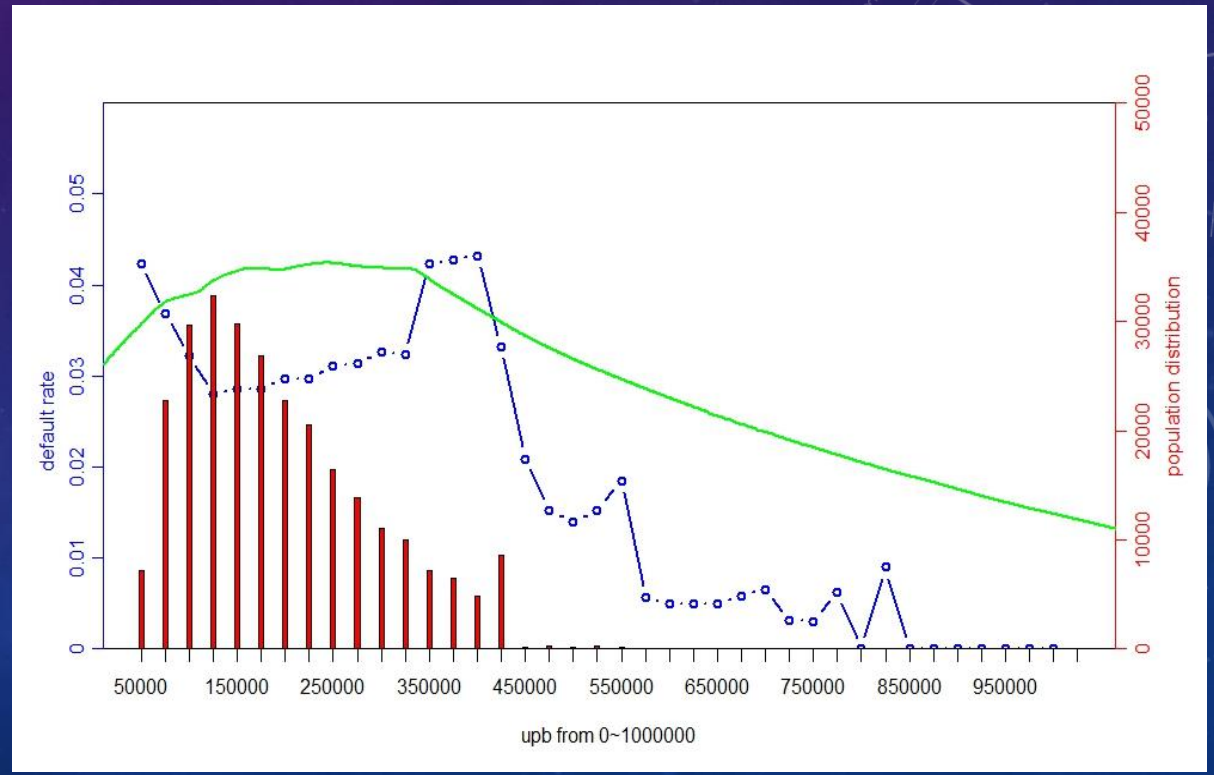


ORIGINAL UPB

Fannie Mae

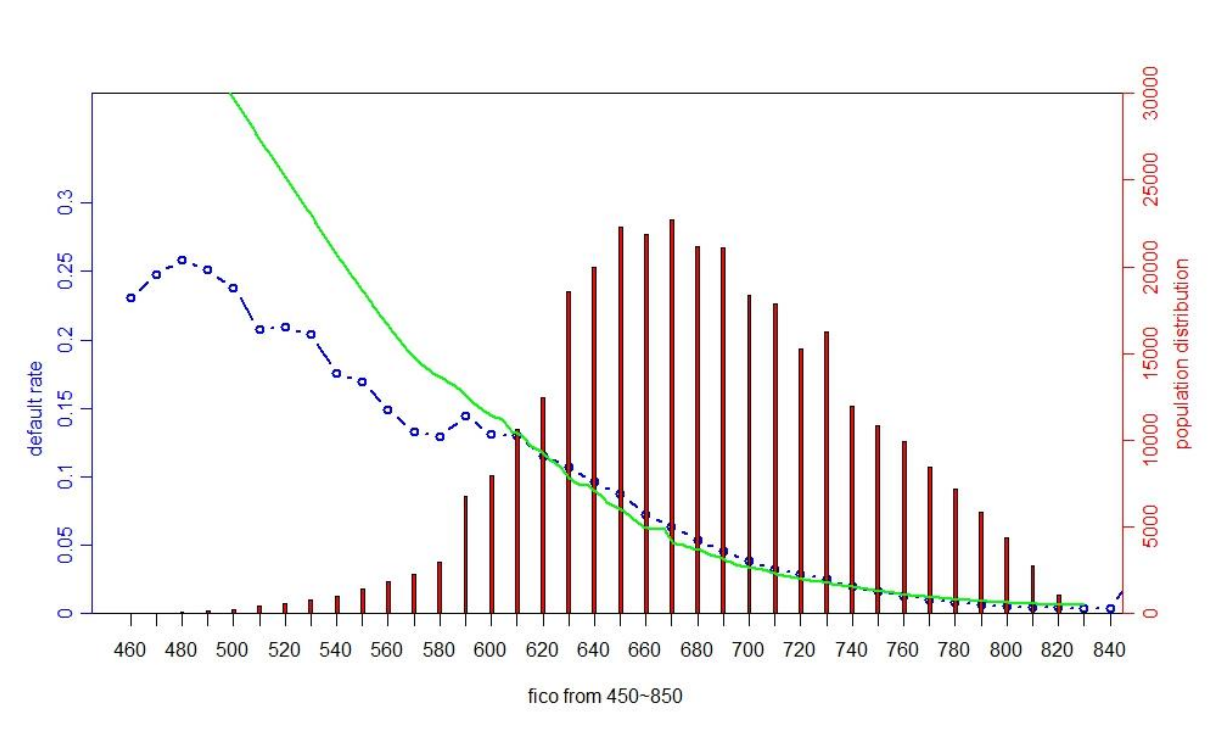


Freddy mac

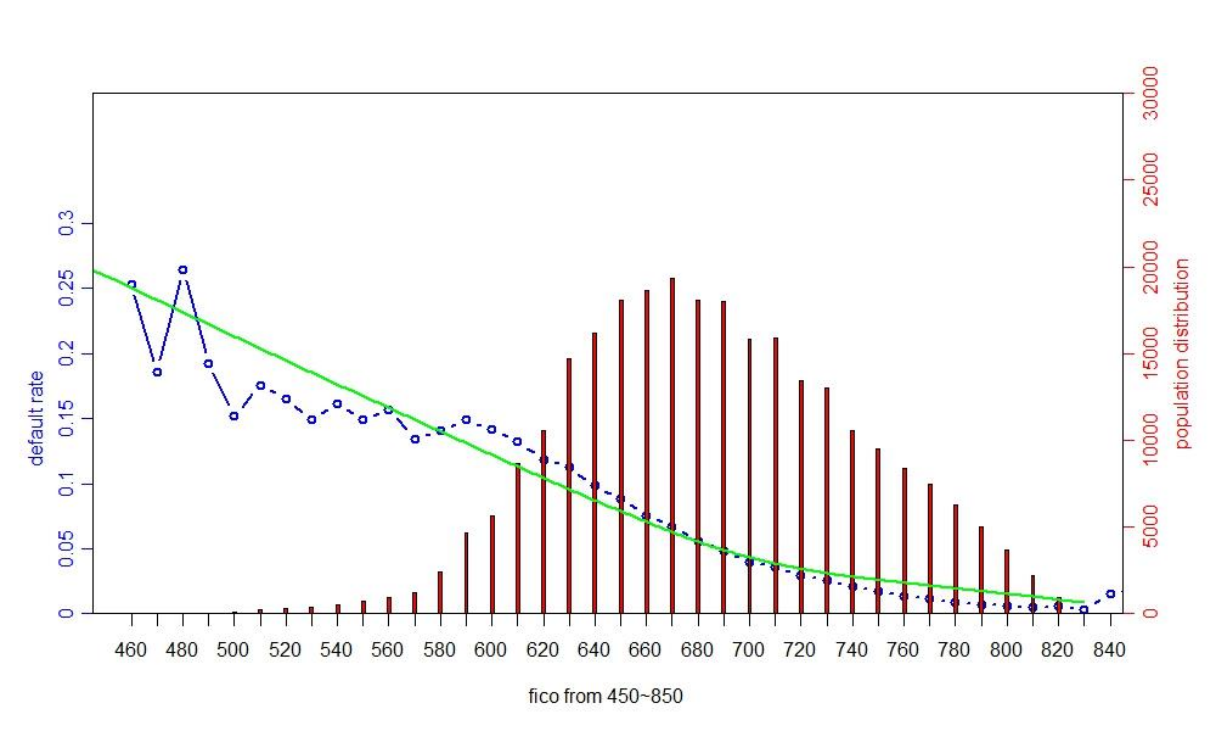


FICO SCORE

Fannie Mae

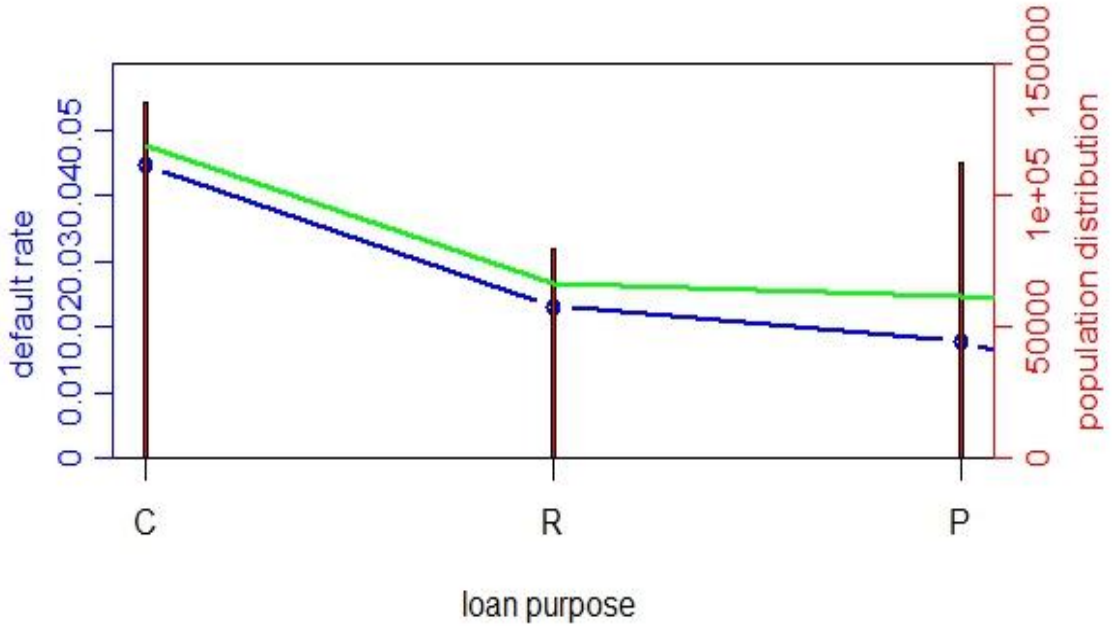


Freddy mac

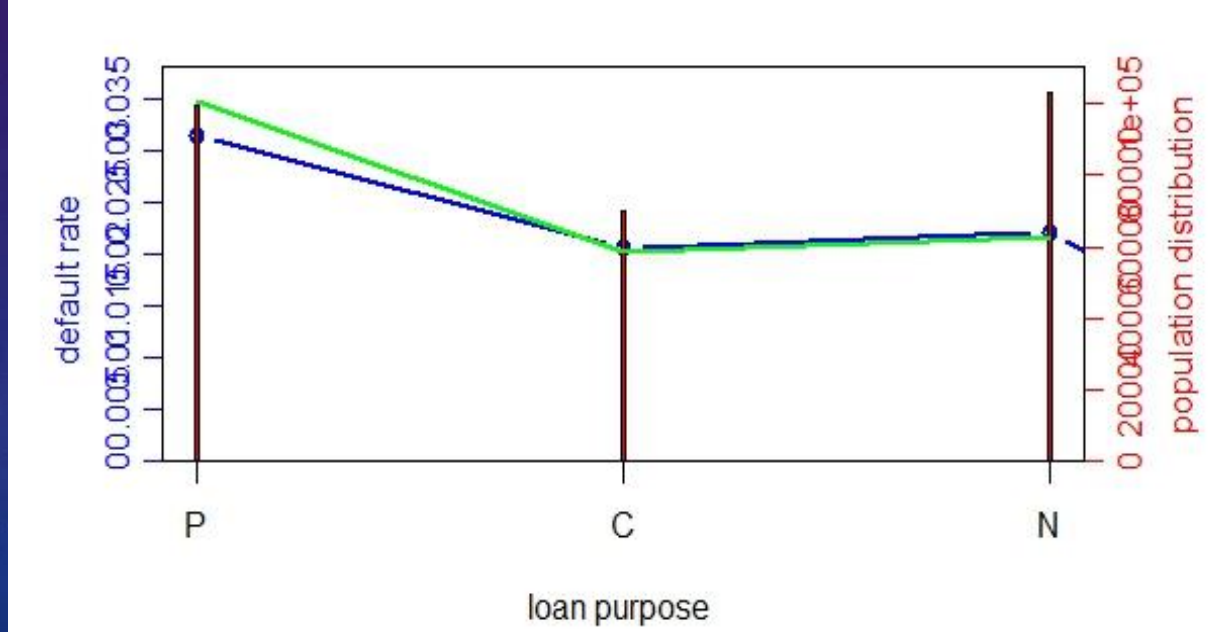


LOAN PURPOSE

Fannie Mae

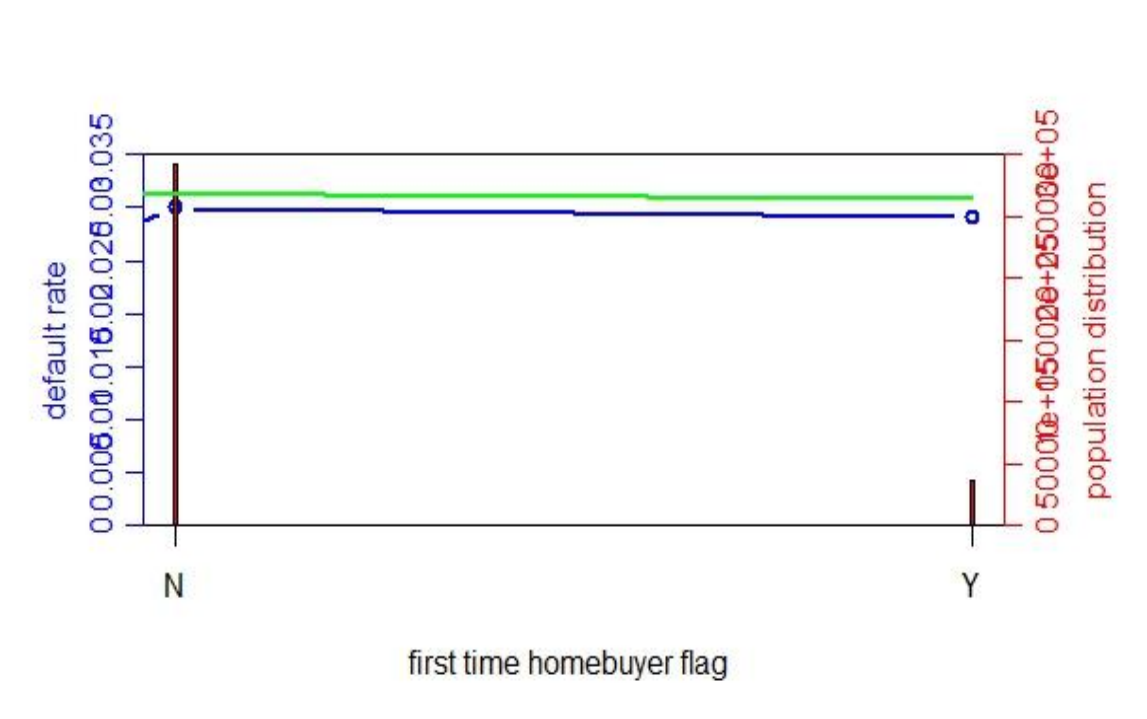


Freddy mac

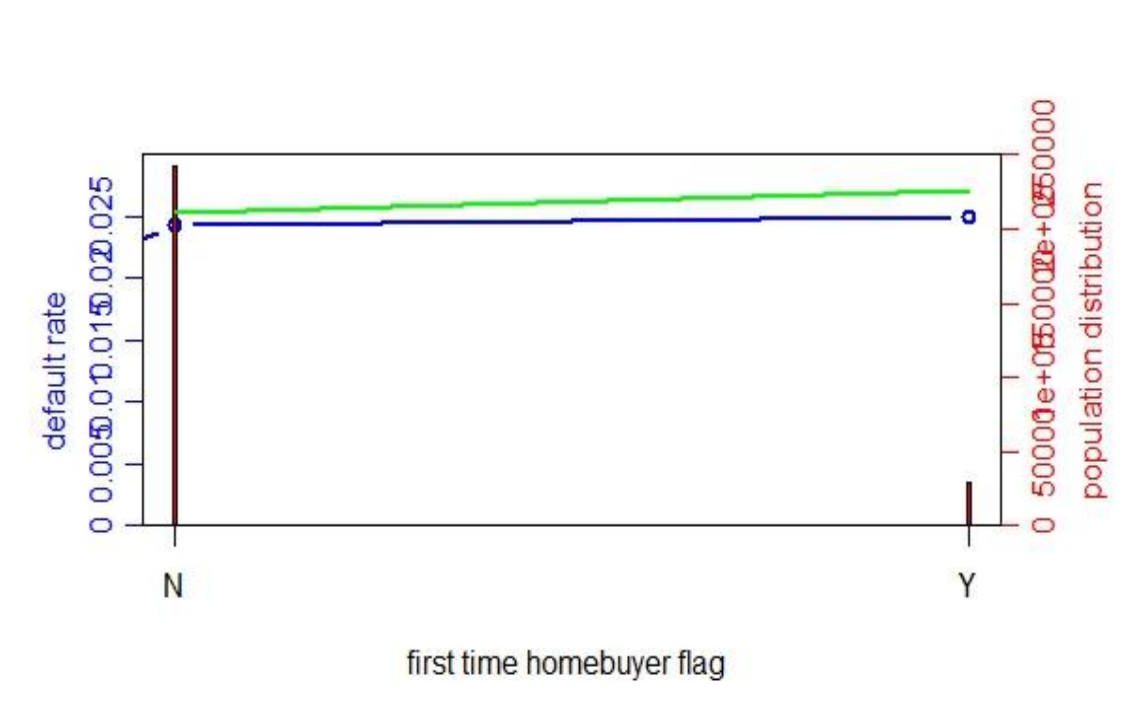


FIRST TIME HOMEBUYER FLAG

Fannie Mae



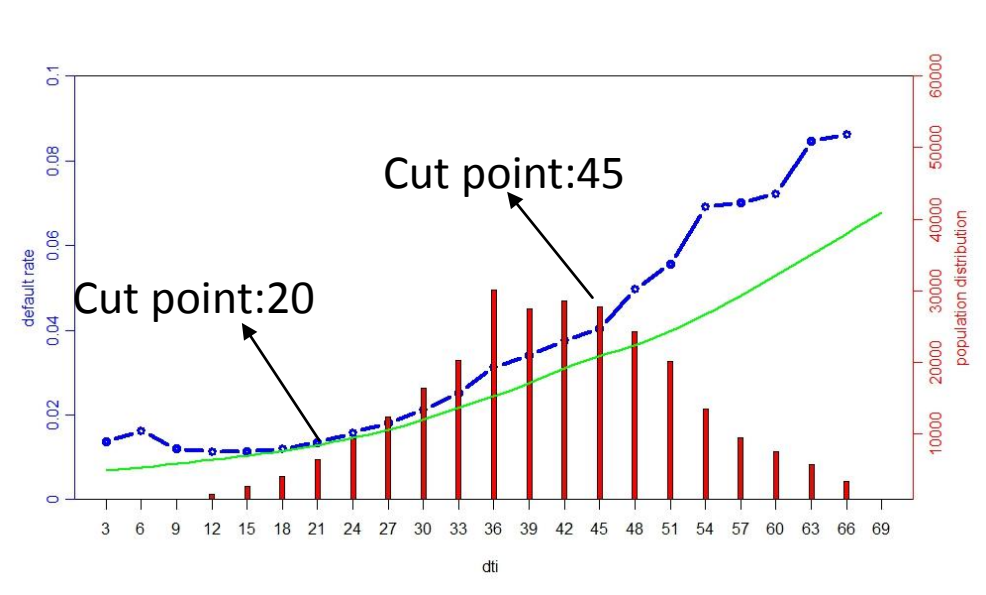
Freddy mac

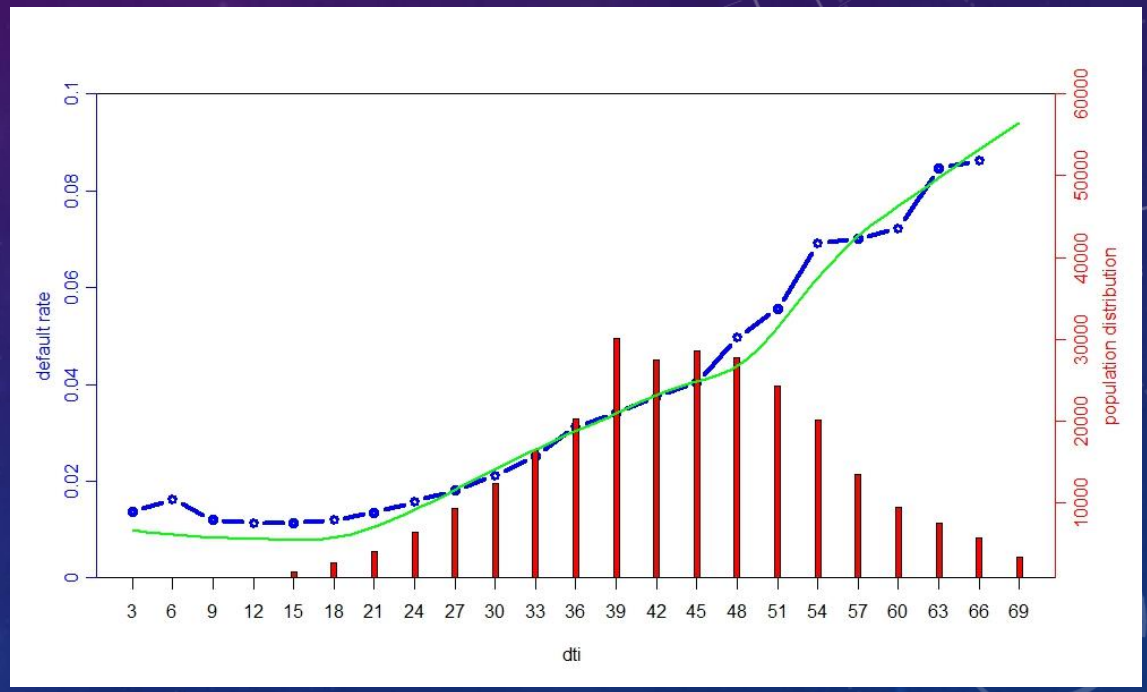


ADJUSTMENT ON FEATURES

- Problem: model fitness should be improved
- Method: make a spline on each continuous feature(DTI,LTV,FICO,UPB)
- 1. identify the cut point
- 2. create sub features according to the cut point
- 3. apply sub features on the logistic regression model
- 4. analyze the goodness of fitting again

DETAILS ABOUT SPLINE FUNCTION: DTI(FREDDY MAC)



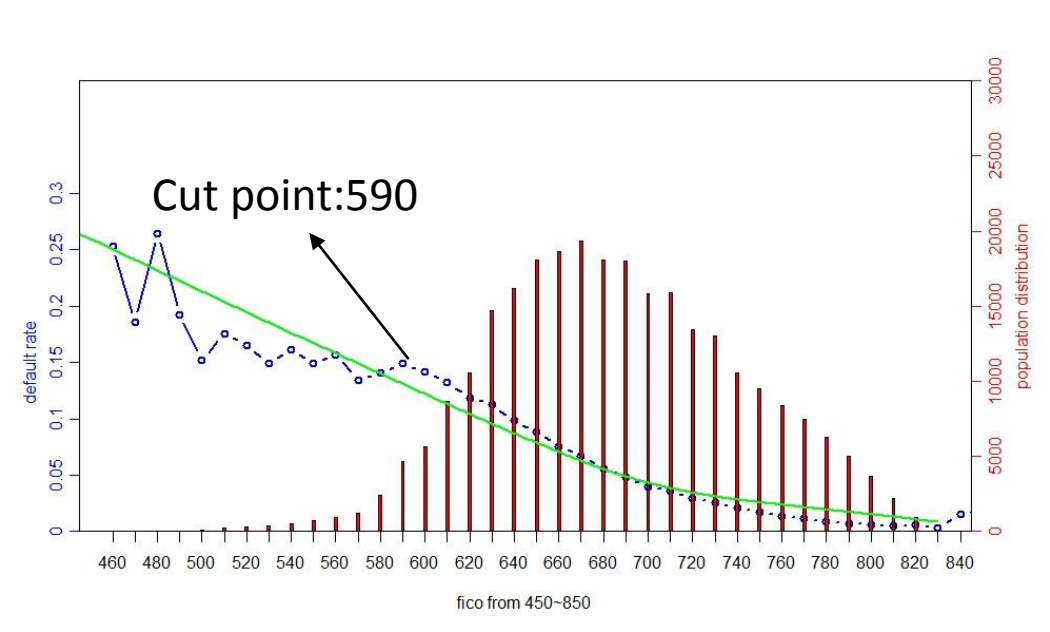


$$\text{dti1}=\begin{cases} dti & if\ dti \le 20 \\ 20 & if\ dti > 20 \end{cases}$$

$$\text{dti2}=\begin{cases} 0 & if\ dti \le 20 \\ dit-20 & if\ 20 < dti \le 45 \\ 25 & if\ dti > 45 \end{cases}$$

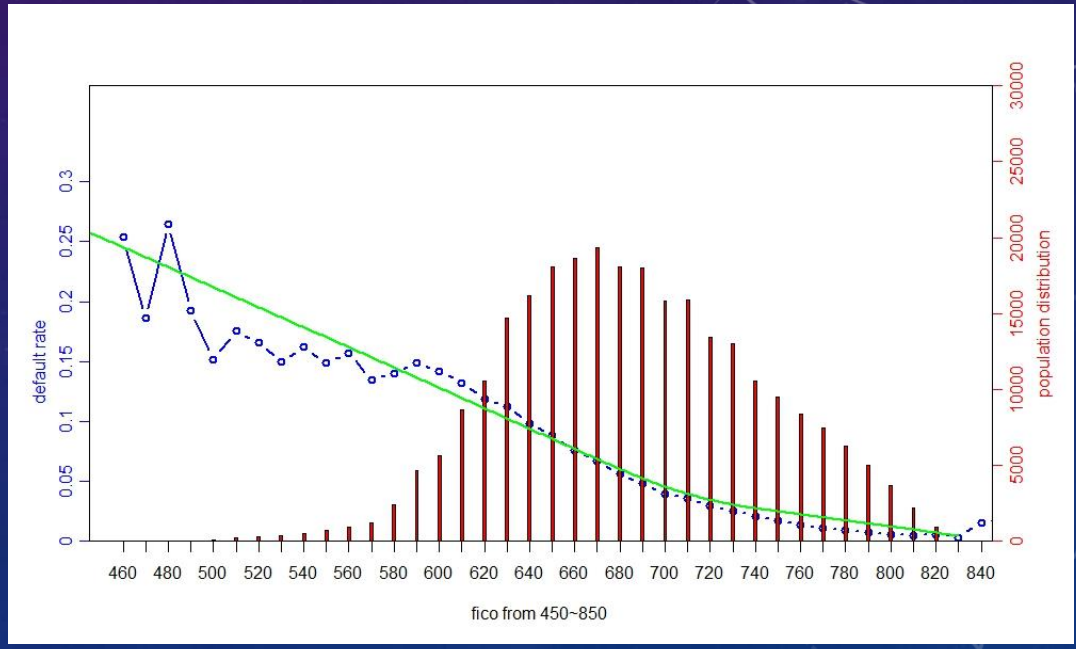
$$\text{dti3}=\begin{cases} 0 & if\ dti \le 45 \\ dit-45 & if\ dti > 45 \end{cases}$$

DETAILS ABOUT SPLINE FUNCTION: FICO(FREDDY MAC)

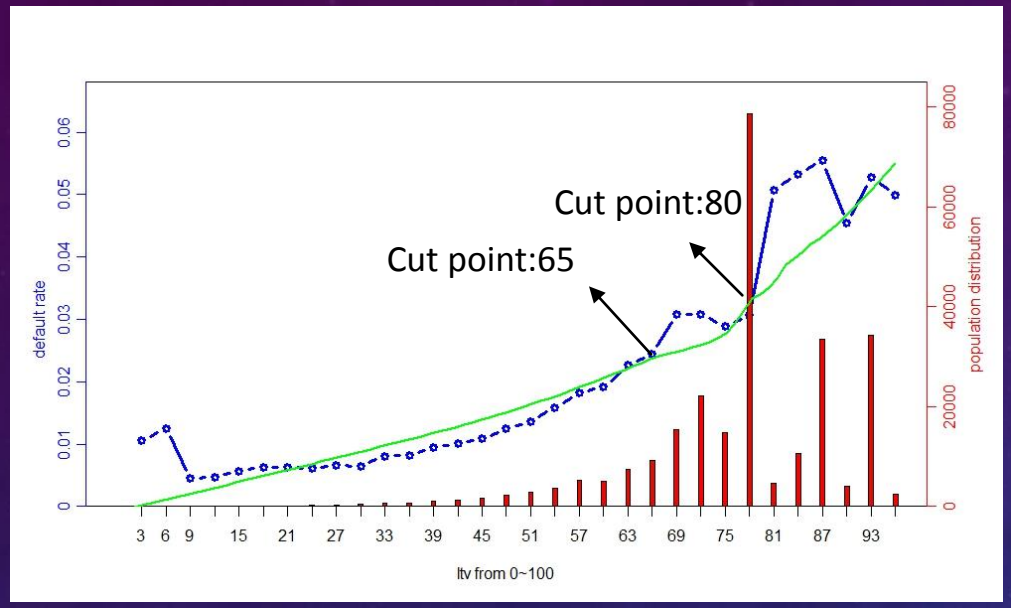


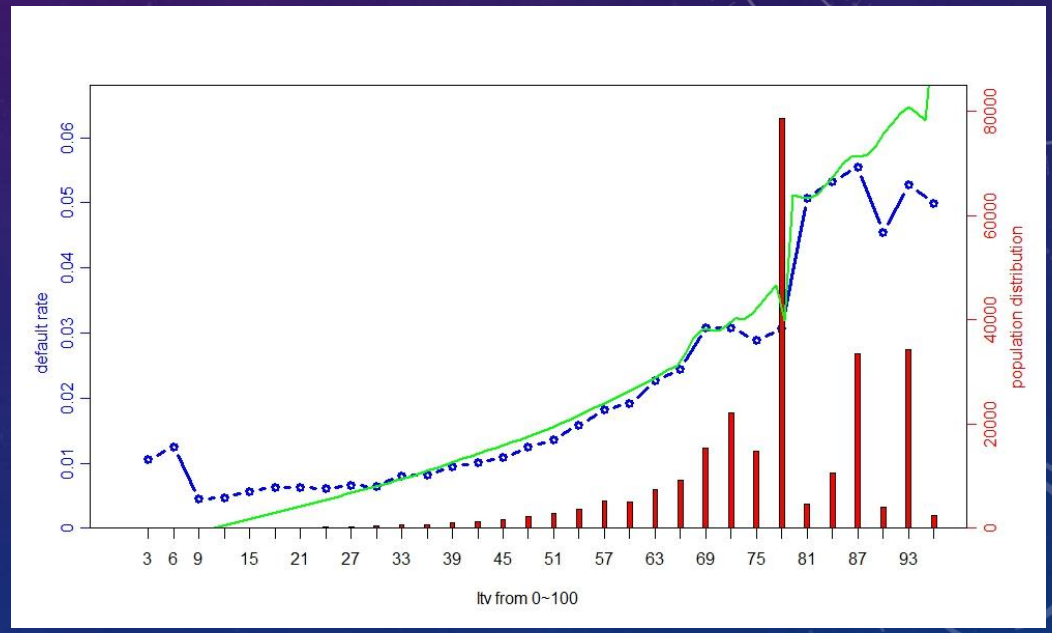
$$\text{fico1}=\begin{cases} fico & if\ dti \le 590 \\ 590 & if\ fico > 590 \end{cases}$$

$$\text{fico2}=\begin{cases} 0 & if\ dti \le 590 \\ fico - 590 & if\ fico > 590 \end{cases}$$



DETAILS ABOUT SPLINE FUNCTION: LTV(FREDDY MAC)



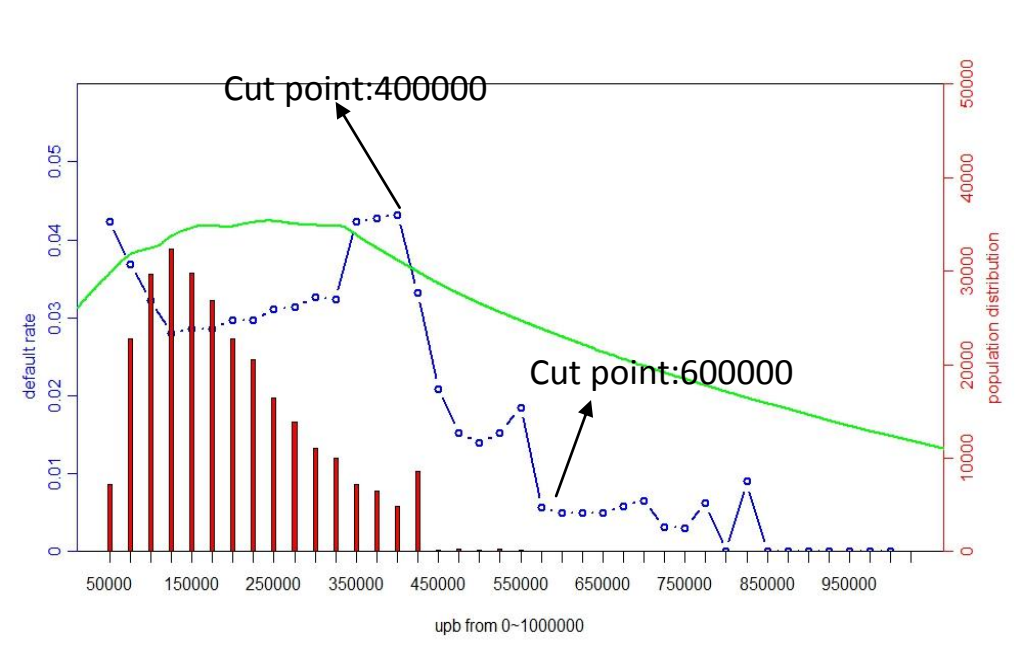


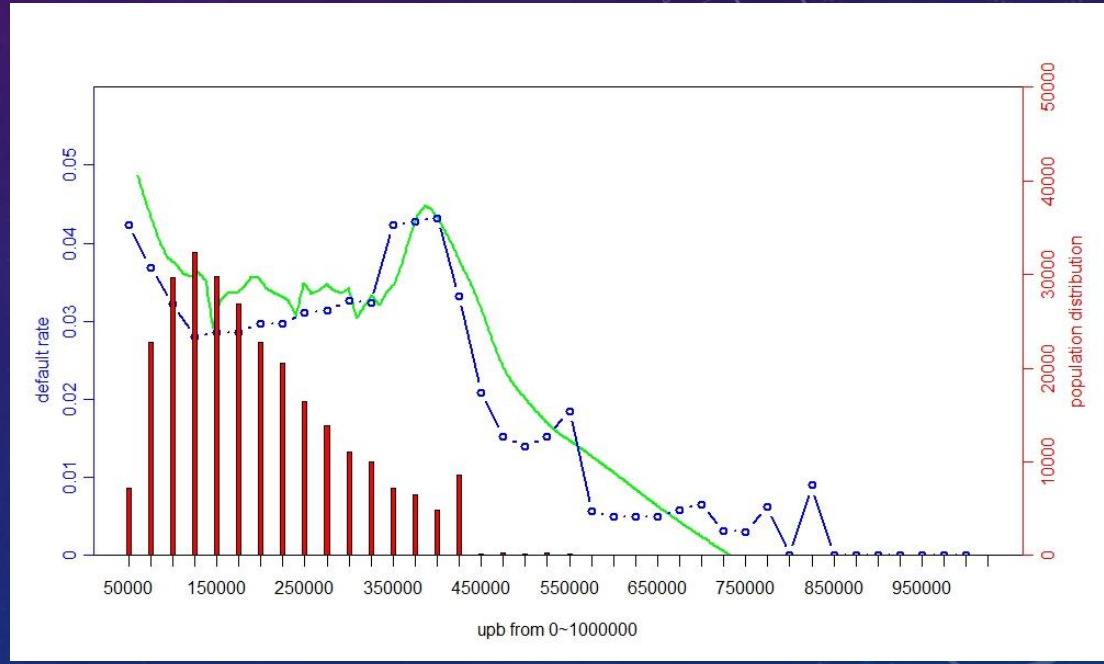
$$ltv1=\begin{cases} ltv & if\ ltv \le 65 \\ 65 & if\ ltv > 65 \end{cases}$$

$$ltv2=\begin{cases} 0 & if\ ltv \le 65 \\ ltv-65 & if\ 65 < ltv \le 80 \\ 15 & if\ ltv > 80 \end{cases}$$

$$ltv3=\begin{cases} 0 & if\ ltv \le 80 \\ ltv-80 & if\ ltv > 80 \end{cases}$$

DETAILS ABOUT SPLINE FUNCTION: UPB (FREDDY MAC)





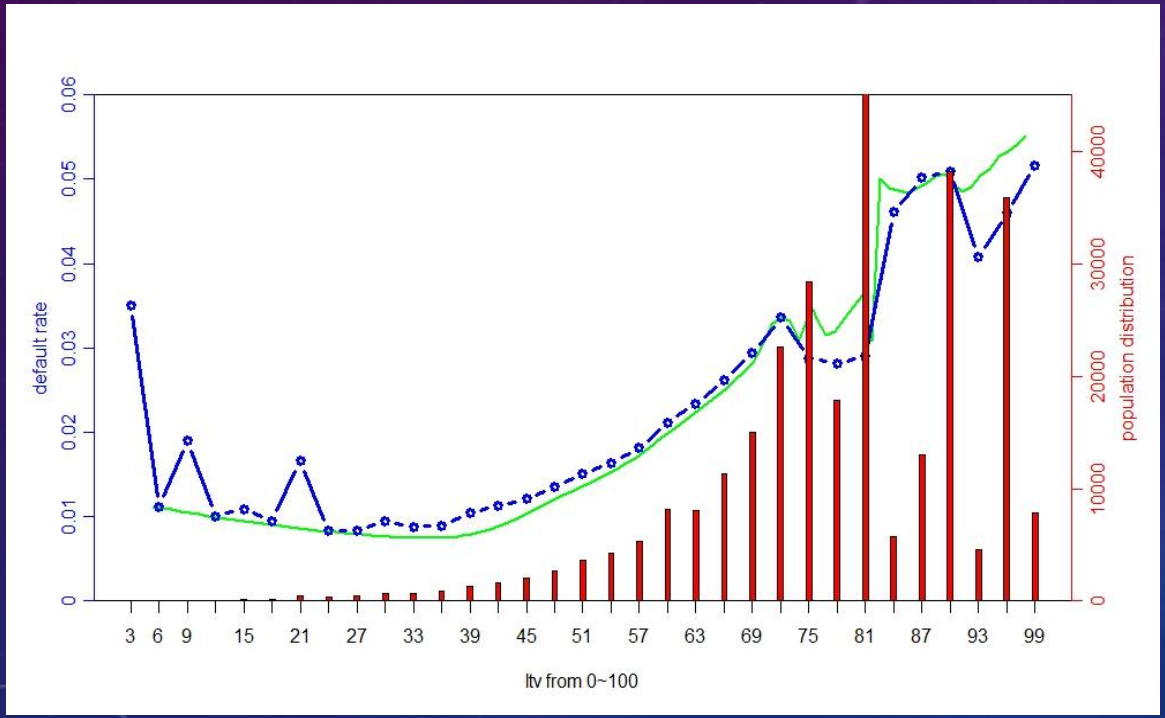
$$\text{upb1}=\begin{cases} upb & if\ upb \le 400000 \\ 400000 & if\ upb > 400000 \end{cases}$$

$$\text{upb2}=\begin{cases} 0 & if\ upb \le 400000 \\ upb - 400000 & if\ 400000 < upb \le 600000 \\ 200000 & if\ upb > 600000 \end{cases}$$

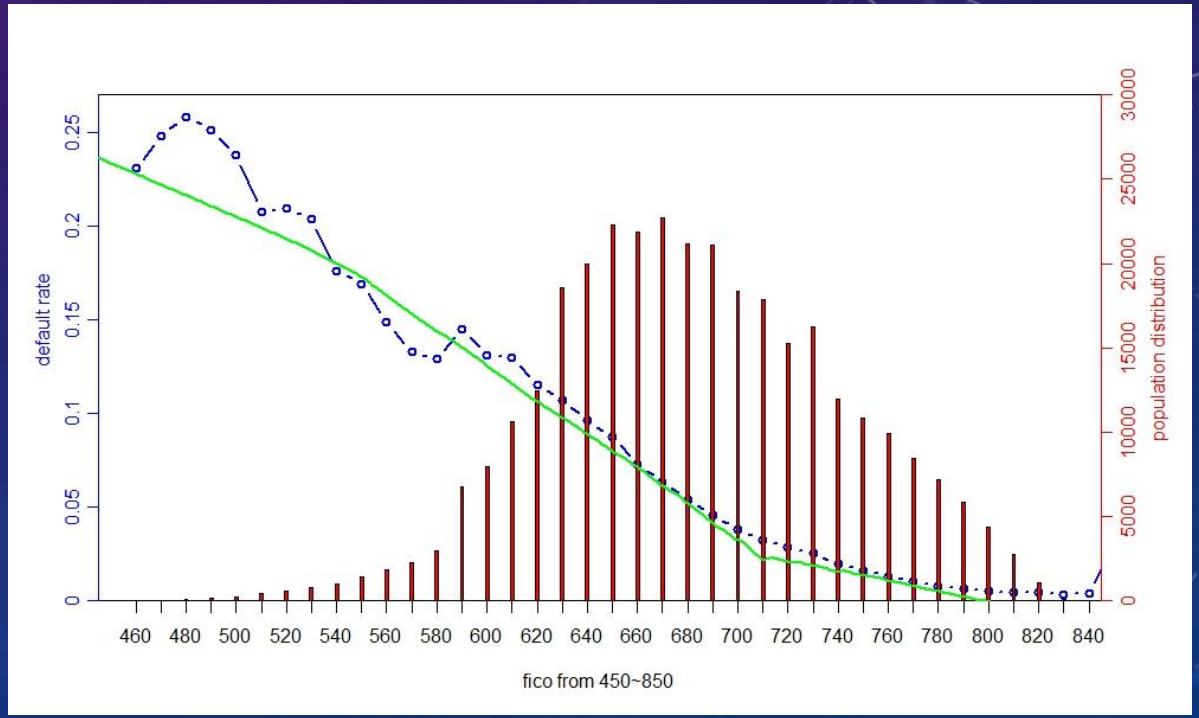
$$\text{upb3}=\begin{cases} 0 & if\ ltv \le 600000 \\ upb - 600000 & if\ ltv > 600000 \end{cases}$$

DETAILS ABOUT SPLINE FUNCTION: FANNIE MAE

LTV

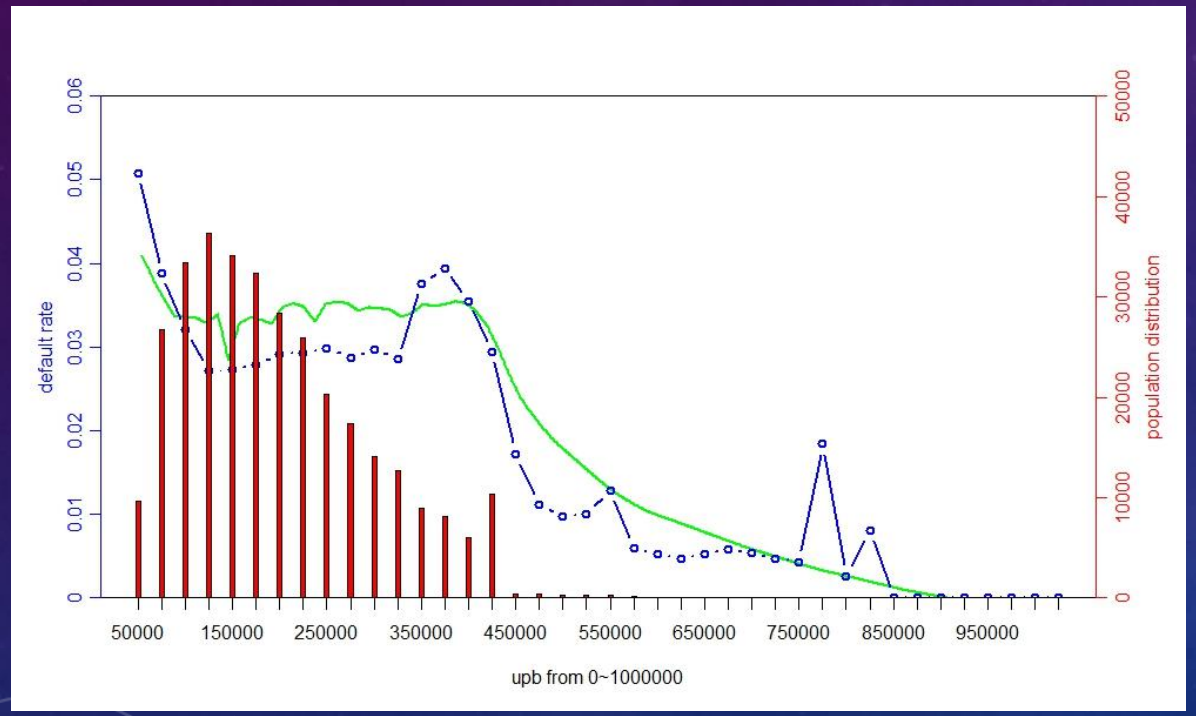


FICO

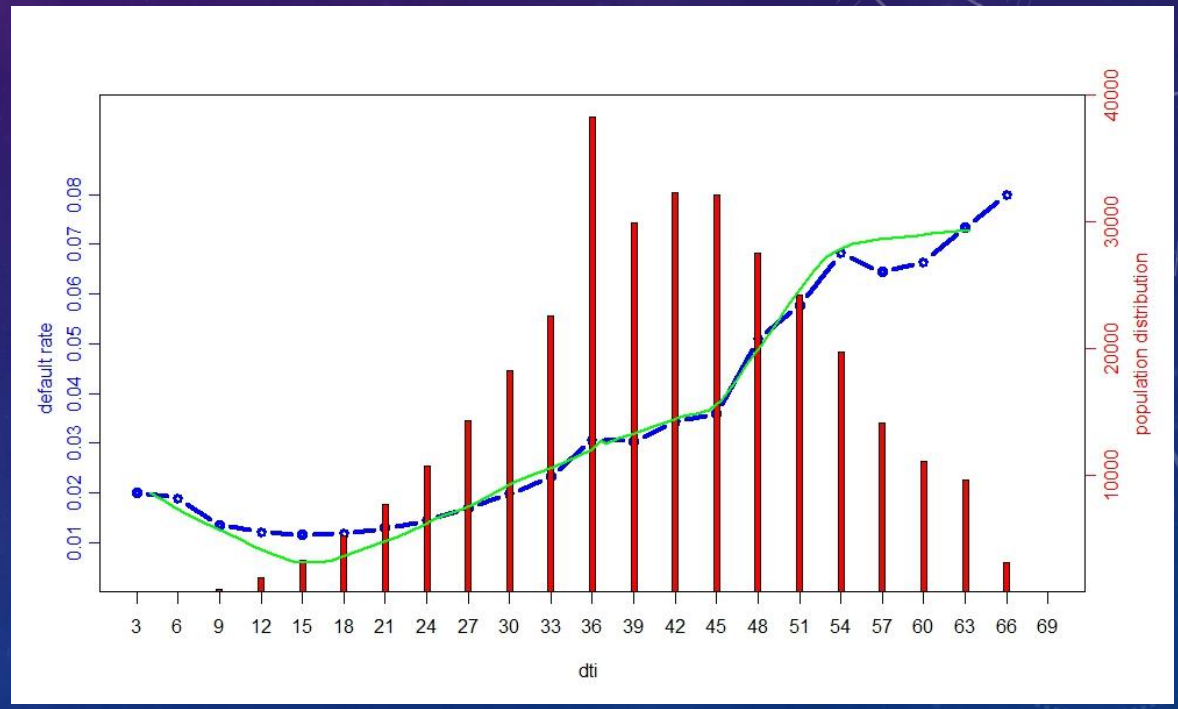


DETAILS ABOUT SPLINE FUNCTION: FANNIE MAE

UPB



DTI



LOGESTIC REGESSION MODEL WITH SPLINED FEATURES

FANNIE MAE

```
Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)              4.910e-01  6.176e-02    7.949 1.88e-15 ***
dti1                    -2.232e-03  2.607e-04   -8.562  < 2e-16 ***
dti2                     5.632e-04  1.756e-05   32.073  < 2e-16 ***
dti3                     3.274e-03  8.334e-05   39.284  < 2e-16 ***

ltv1                     2.652e-04  2.671e-05    9.932  < 2e-16 ***
ltv2                     8.735e-04  3.159e-05   27.652  < 2e-16 ***
ltv3                     1.264e-03  4.095e-05   30.862  < 2e-16 ***
fico1                   -5.348e-04  1.155e-04   -4.629 3.67e-06 ***
fico2                   -8.700e-04  6.780e-06 -128.310  < 2e-16 ***

fico3                   -1.160e-04  5.637e-06  -20.571  < 2e-16 ***
upb1                     3.646e-08  1.626e-09   22.431  < 2e-16 ***
upb2                    -4.371e-08  7.875e-09   -5.551 2.84e-08 ***
upb3                     1.897e-08  2.534e-08    0.748    0.454
dataset1$homefirstY      4.886e-04  4.987e-04    0.980    0.327
dataset1$loanpurposeP   -2.077e-02  4.068e-04  -51.064  < 2e-16 ***
dataset1$loanpurposeR   -1.334e-02  3.770e-04  -35.382  < 2e-16 ***
dataset1$loanpurposeU   -3.791e-02  6.793e-03   -5.581 2.40e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.4367739)
```

FREDDY MAC

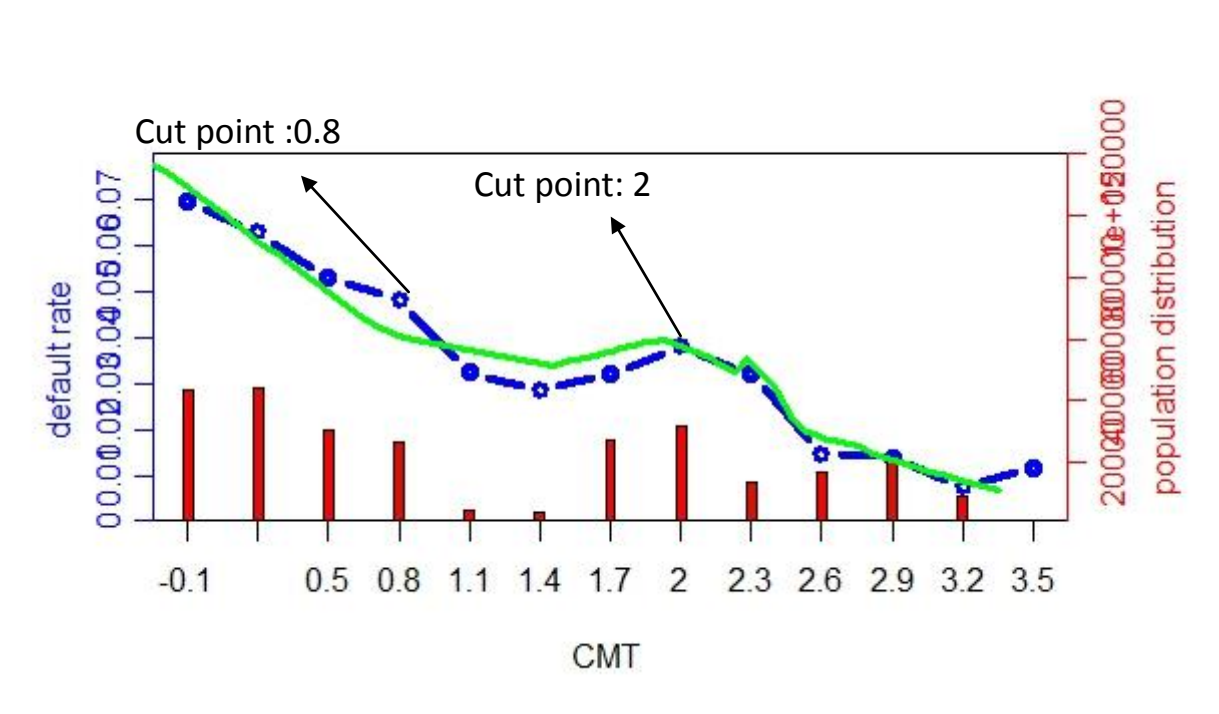
```
Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)              5.260e-02  8.248e-02    0.638    0.524
dti1                    -1.881e-03  2.635e-04   -7.138 9.45e-13 ***
dti2                     5.223e-04  2.004e-05   26.065  < 2e-16 ***
dti3                     3.601e-03  9.967e-05   36.127  < 2e-16 ***

ltv1                     1.916e-04  2.347e-05    8.163 3.27e-16 ***
ltv2                     9.220e-04  3.445e-05   26.763  < 2e-16 ***
ltv3                     1.857e-03  4.534e-05   40.953  < 2e-16 ***
fico1                    2.093e-04  1.543e-04    1.356    0.175
fico2                   -8.359e-04  7.425e-06 -112.581  < 2e-16 ***

fico3                   -8.059e-05  6.664e-06  -12.094  < 2e-16 ***
upb1                     4.608e-08  1.881e-09   24.500  < 2e-16 ***
upb2                    -5.455e-08  1.149e-08   -4.745 2.08e-06 ***
upb3                    -6.887e-09  4.025e-08   -0.171    0.864
dataset1$homefirstY     -4.082e-05  5.735e-04   -0.071    0.943
dataset1$loanpurposeN   -1.489e-02  4.327e-04  -34.399  < 2e-16 ***
dataset1$loanpurposeP   -2.234e-02  4.554e-04  -49.045  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.4510141)
```

APPLY MACRO VARS FEATURES

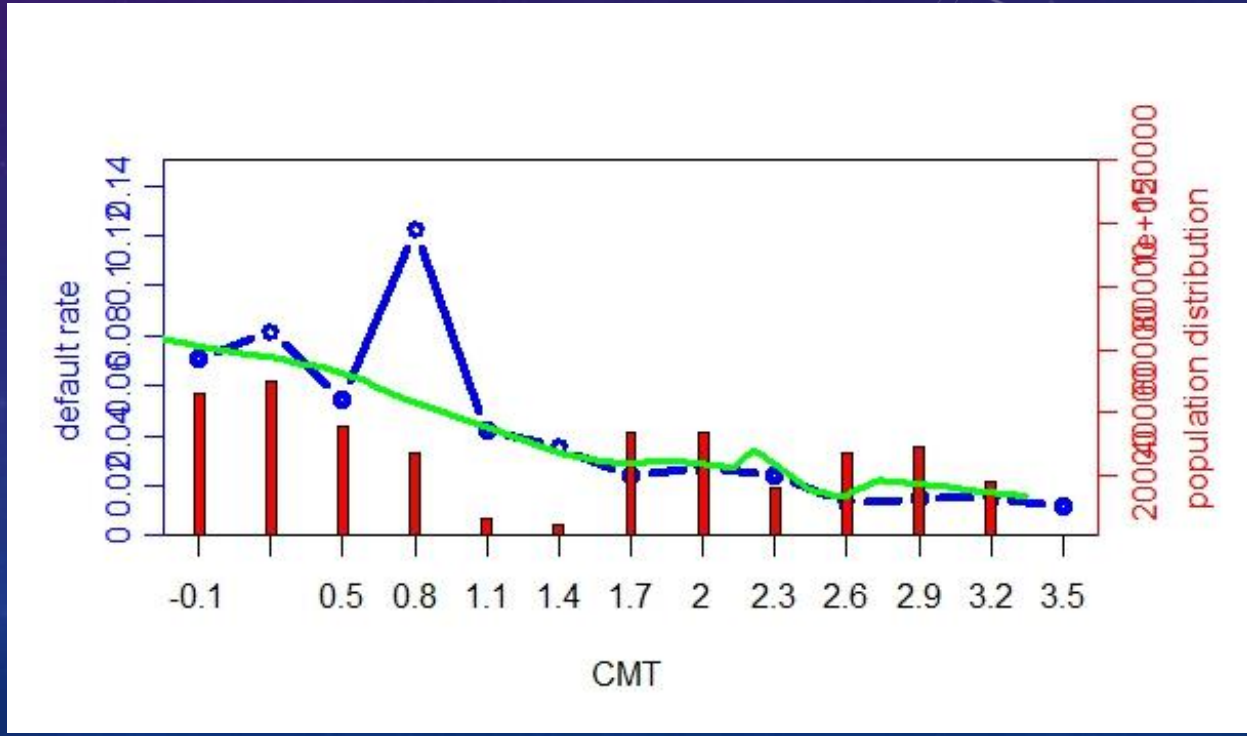
- 1. relative employment rate (RUE)
- - matching method: by Loan's origination year, quarter , location
- 2 . Two year Home price appreciation (HPA)
- - HPA=[HPI2year(housing price index 2 years)/HPI (housing price index original year)]-1
- - matching method of HPI 2 year : by Loan's origination year +2, quarter, location
- - matching method of HPI : by Loan's origination year +2, quarter, location
- 3. mortgage spread rate
- - spread rate = original loan interest rate- cmt10 (10 years' constant maturity treasure rate)
- - matching method of cmt10 : by origination year ,quarter
- 4.cmt difference
- -CMT difference=cmt10-cmt01(1 year's constant maturity treasure rate)
- - matching method of cmt01 : by origination year ,quarter

LOGISTIC REGRESSION MODEL ANALYZE： CMT

FANNIE MAE

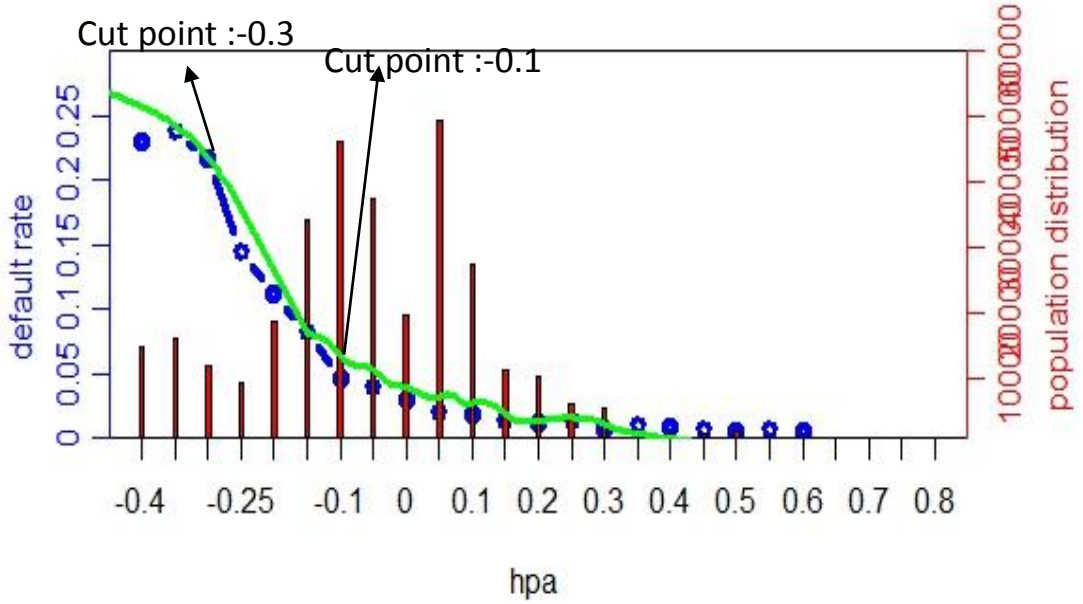


FREDDY MAC

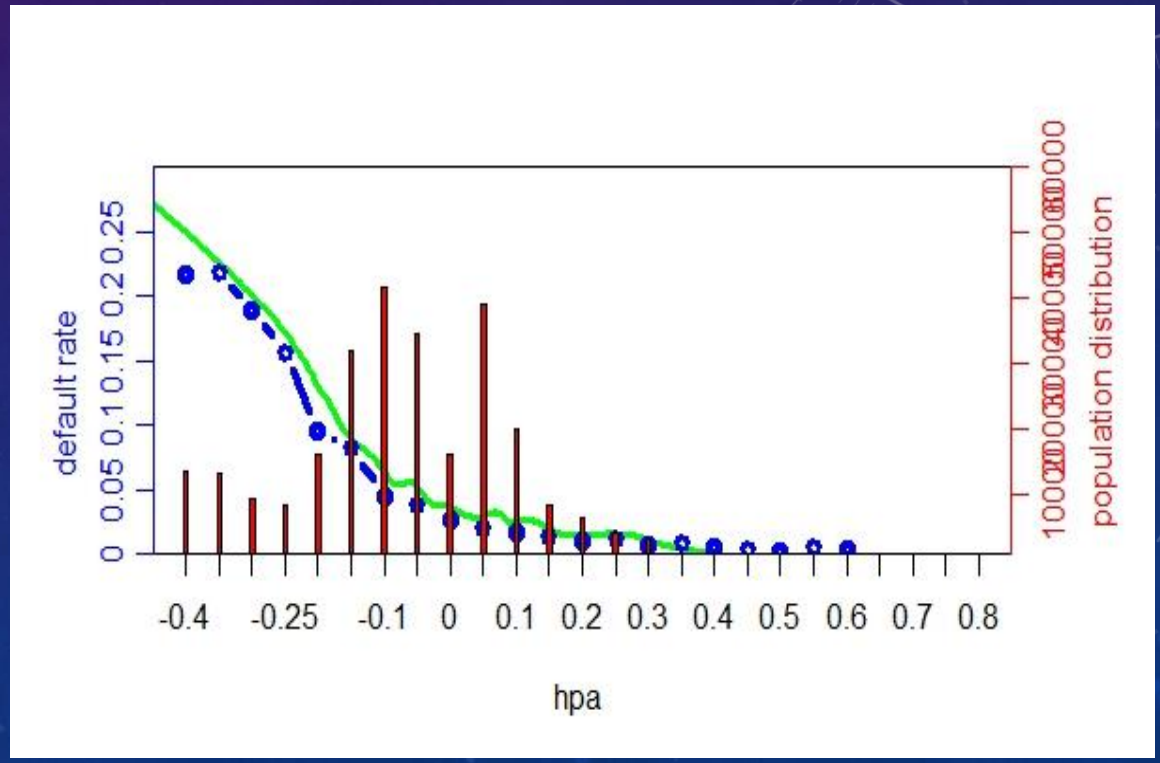


LOGISTIC REGRESSION MODEL ANALYZE： HPA

FANNIE MAE

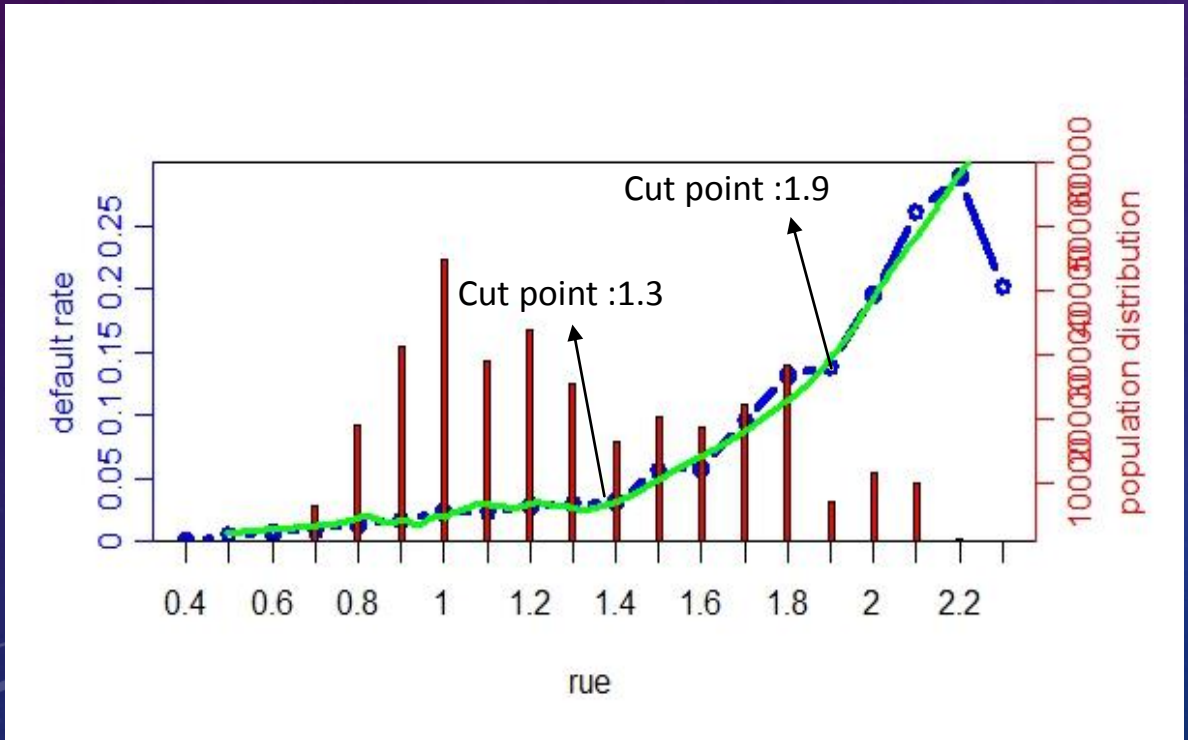


FREDDY MAC

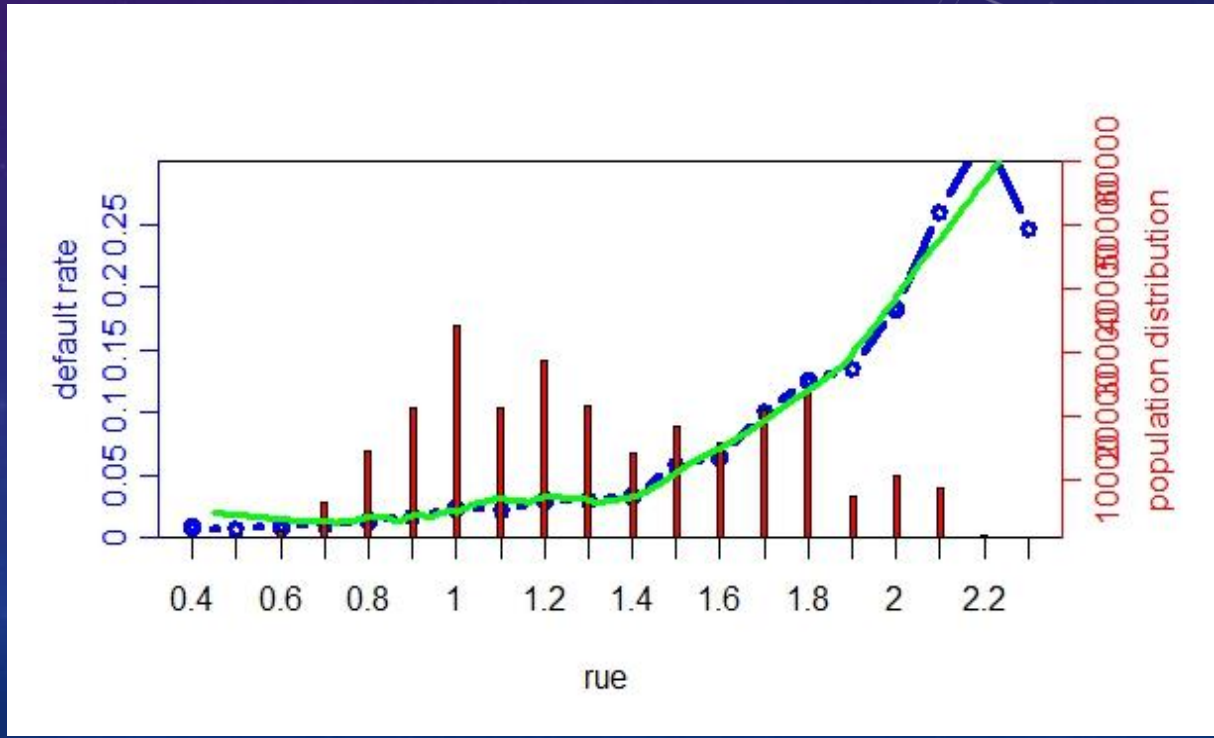


LOGISTIC REGRESSION MODEL ANALYZE： RUE

FANNIE MAE

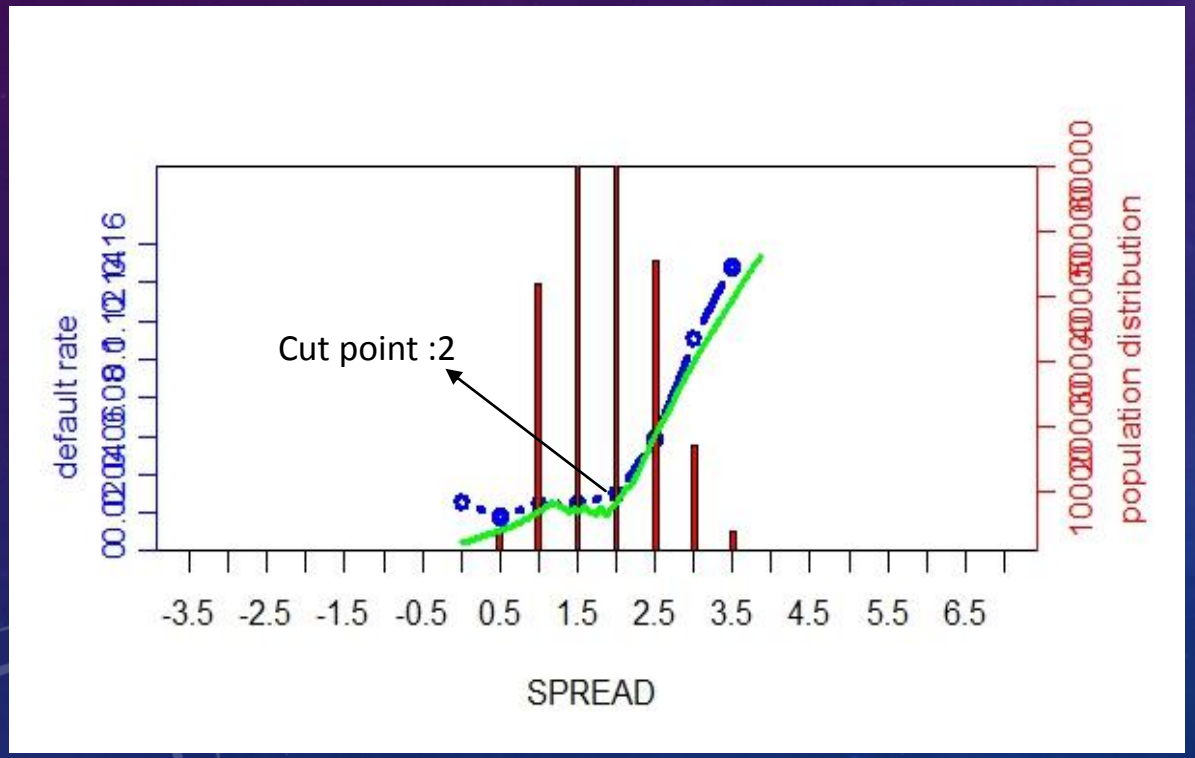


FREDDY MAC

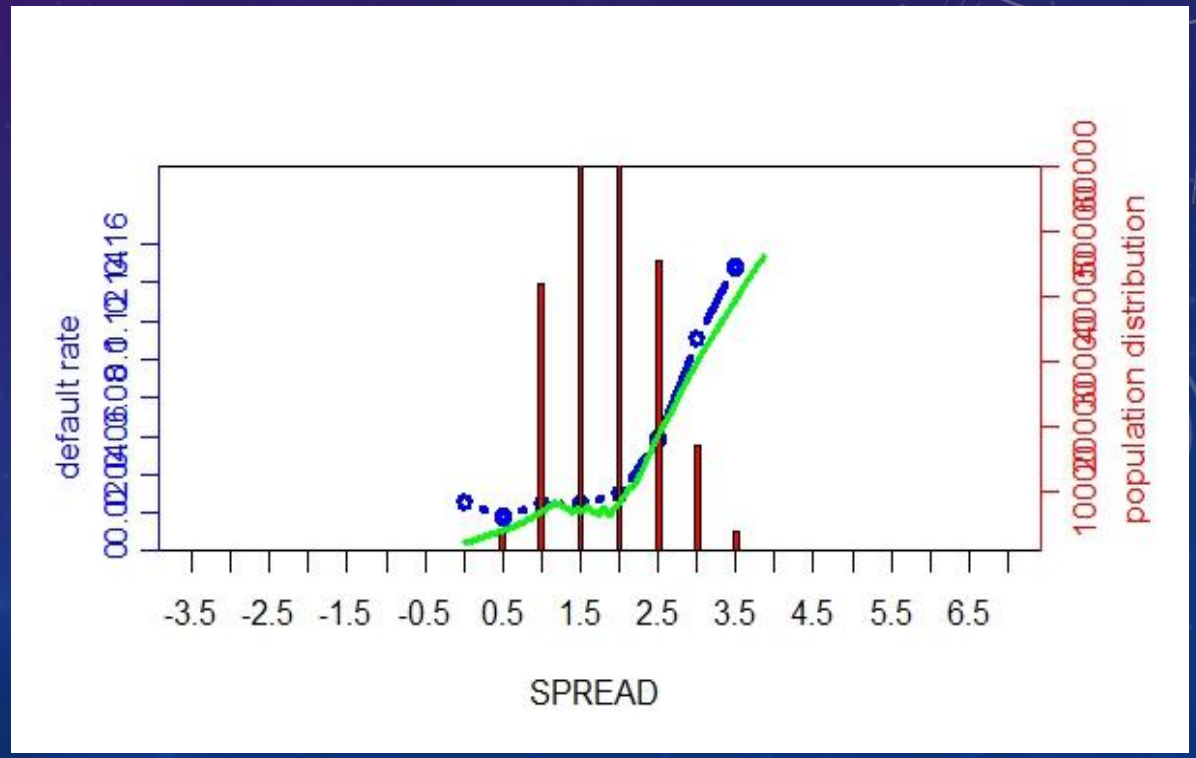


LOGISTIC REGRESSION MODEL ANALYZE： SPREAD

FANNIE MAE



FREDDY MAC



LOGISTIC MODEL COEFFICENT

FANNIE MAE

```
Coefficients: (2 not defined because of singularities)
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)               5.947e-01  6.071e-02    9.796  < 2e-16 ***
dti1                     -1.808e-03  2.548e-04   -7.098 1.27e-12 ***
dti2                      3.650e-04  1.719e-05   21.238  < 2e-16 ***
dti3                      1.989e-03  8.171e-05   24.338  < 2e-16 ***

cmt1                     -4.619e-03  7.536e-04   -6.129 8.85e-10 ***
cmt2                     -2.113e-02  8.000e-04  -26.407  < 2e-16 ***
cmt3                      3.095e-03  4.840e-04    6.394 1.62e-10 ***
spreada1                 -1.984e-03  5.712e-04   -3.474 0.000513 ***
spreada2                  2.965e-02  6.321e-04   46.900  < 2e-16 ***
hpa1                     -1.898e-01  2.146e-02   -8.843  < 2e-16 ***
hpa2                     -1.011e+00  1.376e-02  -73.455  < 2e-16 ***
hpa3                     -6.172e-02  1.003e-03  -61.548  < 2e-16 ***
rue1                      1.154e-01  1.602e-03   72.023  < 2e-16 ***
rue2                             NA         NA       NA       NA
rue3                             NA         NA       NA       NA
ltv1                      3.593e-04  2.314e-05   15.527  < 2e-16 ***
ltv2                      7.139e-04  3.058e-05   23.346  < 2e-16 ***
ltv3                      1.378e-03  4.009e-05   34.364  < 2e-16 ***
fico1                    -5.885e-04  1.129e-04   -5.214 1.85e-07 ***
fico2                    -7.923e-04  6.649e-06 -119.156  < 2e-16 ***

fico3                    -1.304e-04  5.563e-06  -23.437  < 2e-16 ***
upb1                      1.170e-08  1.636e-09    7.154 8.40e-13 ***
upb2                     -6.855e-09  7.720e-09   -0.888 0.374594
upb3                     -4.154e-08  2.478e-08   -1.676 0.093678 .
datasetset1$homefirstY   -2.458e-05  4.880e-04   -0.050 0.959834
datasetset1$loanpurposeP -1.773e-02  3.990e-04  -44.442  < 2e-16 ***
datasetset1$loanpurposeR -4.415e-03  3.704e-04  -11.919  < 2e-16 ***
datasetset1$loanpurposeU -1.389e-02  6.637e-03   -2.093 0.036378 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FREDDY MAC

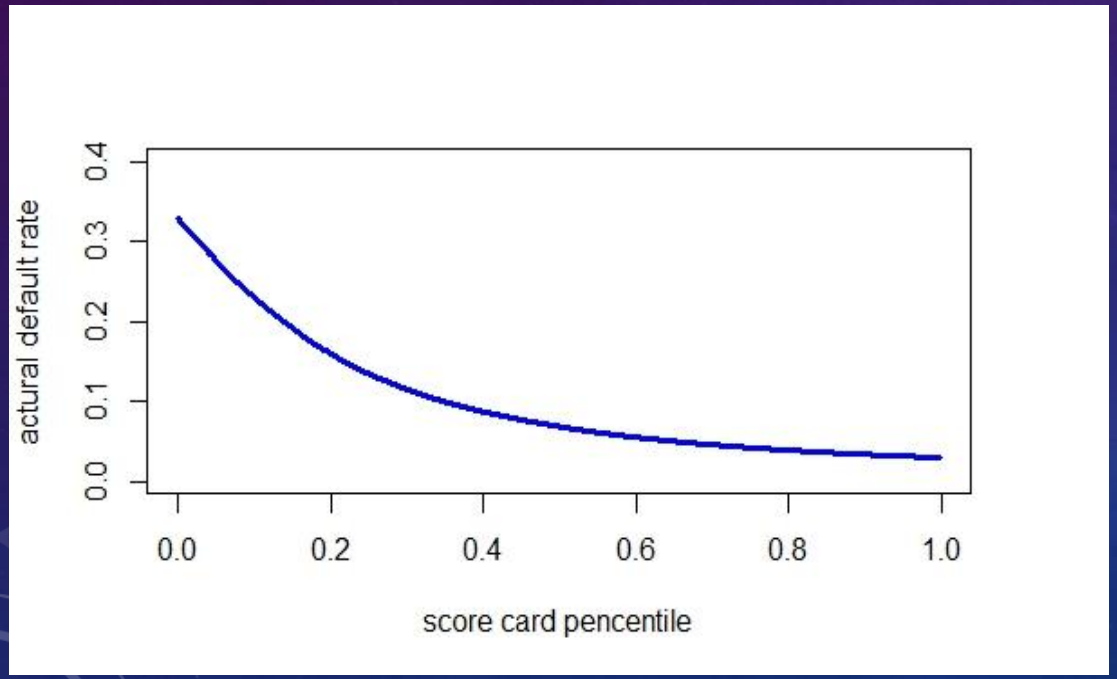
```
Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)               2.267e-01  8.086e-02    2.804  0.00505 **
dti1                     -1.495e-03  2.573e-04   -5.811 6.22e-09 ***
dti2                      3.798e-04  1.958e-05   19.394  < 2e-16 ***
dti3                      2.347e-03  9.754e-05   24.062  < 2e-16 ***

cmt1                     -1.256e-02  7.601e-04  -16.528  < 2e-16 ***
cmt2                     -5.426e-03  8.366e-04   -6.485 8.87e-11 ***
cmt3                     -1.236e-03  5.757e-04   -2.147  0.03183 *
spreada1                 -6.648e-05  5.541e-04   -0.120  0.90450
spreada2                  3.240e-02  6.071e-04   53.359  < 2e-16 ***
hpa1                     -1.493e-01  2.269e-02   -6.582 4.65e-11 ***
hpa2                     -8.970e-01  1.495e-02  -60.001  < 2e-16 ***
hpa3                     -7.069e-02  1.445e-03  -48.908  < 2e-16 ***
rue1                     -4.769e-03  1.085e-03   -4.397 1.10e-05 ***
rue2                      1.194e-01  1.954e-03   61.077  < 2e-16 ***
rue3                      1.600e-01  8.504e-03   18.819  < 2e-16 ***
ltv1                      4.108e-04  2.295e-05   17.905  < 2e-16 ***
ltv2                      8.588e-04  3.367e-05   25.507  < 2e-16 ***
ltv3                      1.742e-03  4.439e-05   39.258  < 2e-16 ***
fico1                     9.911e-06  1.506e-04    0.066  0.94754
fico2                    -7.502e-04  7.277e-06 -103.095  < 2e-16 ***

fico3                    -1.109e-04  6.578e-06  -16.864  < 2e-16 ***
upb1                      1.757e-08  1.897e-09    9.262  < 2e-16 ***
upb2                      9.453e-09  1.124e-08    0.841  0.40050
upb3                     -3.840e-08  3.931e-08   -0.977  0.32867
datasetset1$homefirstY   -3.842e-04  5.605e-04   -0.686  0.49300
datasetset1$loanpurposeN -3.535e-03  4.262e-04   -8.295  < 2e-16 ***
datasetset1$loanpurposeP -1.818e-02  4.477e-04  -40.617  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.4296014)
```

CREATE UNDERWRITING SCORECARD

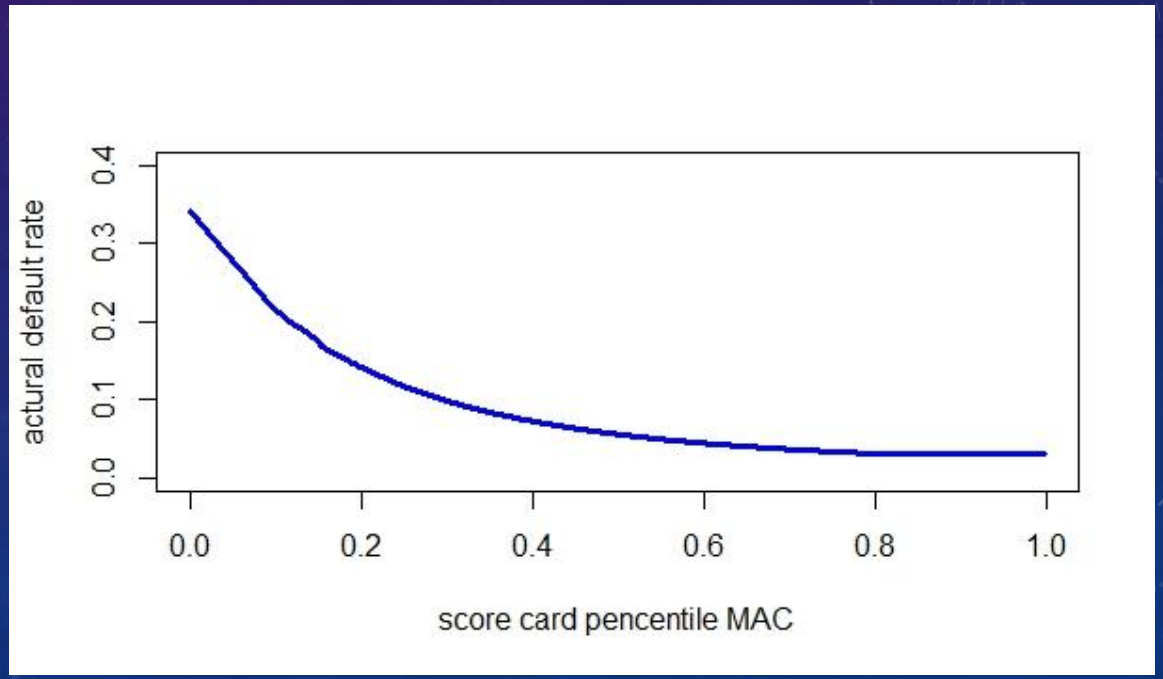
- 1.neutralize macro vars.
- - take the average of the variables(cmt , spread, hpa, rue)
- 2.apply the model coefficients to the neutralized data set, and get the predicted default rate
- 3. scorecard=(-1)*predicted default rate (the higher the default rate, the lower the score)
- 4.rank our scorecard rate from low to high.
- 5. make the scorecard graph by percentile.（1000 interval from 0 to 1）
- Example：0.1%：$\frac{2*(\#\ of\ actual\ defaults\ in\ 0\sim0.1\%\ interval)}{2*(\#\ of\ actual\ defaults\ in\ 0\sim0.1\%\ interval)+20*(\#\ of\ actual\ non-defaults\ in\ 0\sim0.1\%\ interval)}$
- 0.2%：$\frac{2*(\#\ of\ actual\ defaults\ in\ 0\sim0.2\%\ interval)}{2*(\#\ of\ actual\ defaults\ in\ 0\sim0.2\%\ interval)+20*(\#\ of\ actual\ non-defaults\ in\ 0\sim0.2\%\ interval)}$
- 0.3%：$\frac{2*(\#\ of\ actual\ defaults\ in\ 0\sim0.3\%\ interval)}{2*(\#\ of\ actual\ defaults\ in\ 0\sim0.3\%\ interval)+20*(\#\ of\ actual\ non-defaults\ in\ 0\sim0.3\%\ interval)}$

SCORE CARD GRAPH

FANNIE MAE



FREDDY MAC



CONCLUSION

- 1.sampling data by age and delinquency status
- 2. create logistic model by splined features
- 3. add macro Vars features to this model and revalue this model
- 4.neutralize Macro Vars and get the scorecard default rate
- 5. create the underwriting scorecard.

Thank you for your attention!