U.S. Mortgage Analysis

U.S. residential mortgage-backed securities securitization portfolios

Zhengyu Li
Yingying Li
Hongyu Ma
Rongxin Zhang
Tianwei Zhang

Table of Contents

1. Summary

1.1. Description of the Data Set

In order to build models for dataset Mortgage, we need to understand the dataset. The data set mortgage is in panel form and reports origination and performance observations for 50,000 residential U.S. mortgage borrowers over 60 periods. The periods have been deidentified. As in the real world, loans may originate before the start of the observation period. The loan observations may censored as the loans mature or borrowers refinance. The data set is a randomized selection of mortgage-loan-level data collected from the portfolios underlying U.S. residential mortgage-backed securities (RMBS) securitization portfolios and provided by International Financial Research.
Key variables include:
• id: Borrower ID
• time: Time stamp of observation
• orig_time: Time stamp for origination
• first_time: Time stamp for first observation
• mat_time: Time stamp for maturity
• balance_time: Outstanding balance at observation time
• LTV_time: Loan-to-value ratio at observation time, in %
• interest_rate_time: Interest rate at observation time, in %
• hpi_time: House price index at observation time, base year = 100
• gdp_time: Gross domestic product (GDP) growth at observation time, in %
• uer_time: Unemployment rate at observation time, in %
• REtype_CO_orig_time: Real estate type condominium = 1, otherwise = 0
• REtype_PU_orig_time: Real estate type planned urban development = 1, otherwise = 0
• REtype_SF_orig_time: Single-family home = 1, otherwise = 0
• investor_orig_time: Investor borrower = 1, otherwise = 0
• balance_orig_time: Outstanding balance at origination time
• FICO_orig_time: FICO score at origination time, in %
• LTV_orig_time: Loan-to-value ratio at origination time, in %
• Interest_Rate_orig_time: Interest rate at origination time, in %
• hpi_orig_time: House price index at origination time, base year = 100
• default_time: Default observation at observation time
• payoff_time: Payoff observation at observation time
• status_time: Default (1), payoff (2), and nondefault/nonpayoff (0) observation at observation time

1.2. Data Processing

We also need data processing before to build models. We processed our dataset mortgage before to make it easy for our data processing. In this step, we create variables as many as we can to fit models. However, for the simple calculations of appreciation interest rate between orig_time and time data, we rank variables to improve the prediction of models. Then, we use the cluster technique to label out the relationship between the FICO Ranks and Passed due payment numbers which is $0 payment. (the ones have both high FICO ranks, and high passed due payment numbers will be put in the same cluster. People in this cluster could have high risk which are different information from only FICO scores.)

```
/*Create appreciations to indicates the change of the loan*/
data mortgage_variable;
    set data.mortgage;
    by id;
    hprate = ifn(hpi_orig_time=0,0,(hpi_time-hpi_orig_time)/hpi_orig_time);
    interest_rate_appreciation=ifn(Interest_Rate_orig_time=0,0,(interest_rate_time-Interest_Rate_orig_time)/Interest_Rate_orig_time);
    balance_appreciation=ifn(balance_orig_time=0,0,(balance_time-balance_orig_time)/balance_orig_time);
    LTV_appreciation=ifn(LTV_orig_time=0,0,(LTV_time-LTV_orig_time)/LTV_orig_time);
    time_to_mature=mat_time-time;
    PMT=lag(balance_time)-balance_time;
    PMT_change=ifn(PMT=0,-1,lag(PMT)/PMT - 1);
    if lag(id) NE id then
        do PMT=0;
           PMT_change=0;
    end;
    if PMT =. then PMT=0;
    if PMT_change=. then PMT_change=0;
    if interest_rate_appreciation =. then interest_rate_appreciation=0;
    PMT_to_balance=PMT/balance_time;
    if first.id then nopmt = 0;
    if not first.id and PMT = 0 then do;
    nopmt + 1;
    end;
run;
```

```
proc sort data=mortgage_variable;
    by default_time;
run;
```

```
proc rank data=mortgage_variable out=mortgage_variable groups=10;
    var FICO_orig_time PMT_to_balance balance_orig_time LTV_appreciation balance_appreciation hprate;
    ranks FICO_Ranks PTB_Ranks balance_orig_Ranks LTV_appreciation_Ranks balance_appreciation_Ranks hprate_Ranks;
    by default_time;
run;
```

```
proc fastclus data=mortgage_variable out=mortgage_clusters maxclusters=10;
    var FICO_Ranks nopmt;
run;
```

After data processing, we could fit our models for dataset Mortgage. In this report, we plan to fit three types of model, including Probabilities of Default (PD) model, Loss Given Default (LGD) model and Exposure at Default (EAD) model.

2. Probabilities of Default (PD)

2.1. Description of Probabilities of Default

The Probabilities of Default is the most scrutinized parameter in credit risk analytics and subject to minimum standards imposed by prudential regulators. A PD model describes the likelihood of a default event. Banks observe whether borrowers' default, and generally indicate this with a default indicator:

$$D_{it} = \begin{cases} 1 & borrower\ i\ defaults\ at\ time\ t \\ 0 & otherwise \end{cases} \quad \text{With i} = 1, \dots, \text{I and t} = 1, \dots, \text{T}.$$

We assume that the default event is random and use "D" as the random variable and a "d" as its realization. A default event may be defined by any of the following events:
Payment delinquency of a number of days or more; popular thresholds are 30, 60, and 90 days; Bankruptcy of the borrower; Collateral owned by a bank (e.g., real estate owned after an unsuccessful sale at a foreclosure auction); Foreclosure of loan; Short sale of loan; Loss/write-down amount; Involuntary liquidation; Debt modification as a positive interest, expense, or principal forgiveness. Probabilities of default can be modeled with different methods, in this report, we focus on non-linear models. In the following, we show three non-linear models, including probit model, logit model and cloglog model.

2.2. Probabilities of Default Models

2.2.1. Probit model

2.2.1.1. Model Construction

The probit model equation is:

$$\mathrm{P}(D_{it} = 1 | X_{it-1}) = \Phi(\beta' x_{it-1})$$

With $\Phi(\cdot)$ the cumulative density function of the standard normal distribution, β a vector of sensitivity parameters, and x_{it-1} a vector of time-lagged (with regard to the observable default event) covariates. We can estimate probit model by using PROC LOGISTIC or PROC PROBIT. We choose to use PROC LOGISTIC because it has the same link functions and the link function can easily be changed for robustness checks. The PROC LOGISTIC statement involves the procedure that estimates the probit, logit, and some other models. SAS sorts the dependent variable from low to high and models the first category. The option DESCENDING reverses the default order and specifies that the probability of $y_{it} = 1$ is modeled. The output statement is the combination of the PREDICTED statement that the default probabilities are calculated for the estimation sample and stored in a separate variable in the input data set and also saved in the output. This file can

be used for estimation and prediction of default probabilities. The contents of the item store can be processed with the PROC PLM procedure.

2.2.1.2. Implementation

```
/*Probit Model*/
ods graphics on;
ods rtf file='c:\Users\lizhe\Downloads\probit_model_1.rtf';
proc logistic data=mortgage_clusters descending;
model default_time = balance_time-numeric-hpi_orig_time hprate-numeric-CLUSTER/ link=probit
selection=stepwise slentry=0.05 slstay=0.01
unequalslopes=(CLUSTER hprate_Ranks FICO_Ranks REtype_CO_orig_time-numeric-
REtype_SF_orig_time PTB_Ranks balance_orig_Ranks LTV_appreciation_Ranks balance_appreciation_Ranks investor_orig_time);
output out=probabilities predicted=PD_time;
run;
ods graphics off;
ods rtf close;
```

2.2.1.3. Output

The LOGISTIC Procedure

Model Information	
Data Set	WORK.MORTGAGE_CLUSTERS
Response Variable	default_time
Number of Response Levels	2
Model	binary probit
	partial unequal slopes
Optimization Technique	Newton-Raphson

Step 28. Effect investor_orig_time entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Number of Observations Read	622489
Number of Observations Used	621896

Response Profile		
Ordered Value	default_time	Total Frequency
1	1	15153
2	0	606743

Probability modeled is default_time='1'.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	142509.64	48790.721
SC	142520.98	49119.596
-2 Log L	142507.64	48732.721

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	93774.9185	28	<.0001
Score	72369.1059	28	<.0001
Wald	42013.3267	28	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
1.7398	3	0.6281

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	95.7	Somers' D	0.935
Percent Discordant	2.2	Gamma	0.955
Percent Tied	2.1	Tau-a	0.044
Pairs	9193976679	c	0.967

Summary of Stepwise Selection								
Step	Effect Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	LTV_time		1	1	4763.4893		<.0001	
2	LTV_appreciation_Ranks		1	2	6431.7625		<.0001	Rank for Variable LTV_appreciation
3	hprate		1	3	4889.0283		<.0001	
4	hprate_Ranks		1	4	4107.9834		<.0001	Rank for Variable hprate
5	gdp_time		1	5	3299.3527		<.0001	
6	balance_appreciation		1	6	2297.0050		<.0001	
7	LTV_appreciation		1	7	22352.0577		<.0001	
8	time_to_mature		1	8	5830.1781		<.0001	
9	interest_rate_time		1	9	2489.1500		<.0001	
10	nopmt		1	10	2156.9391		<.0001	
11	LTV_orig_time		1	11	2029.8455		<.0001	
12	FICO_orig_time		1	12	1687.3399		<.0001	
13	FICO_Ranks		1	13	15657.5543		<.0001	Rank for Variable FICO_orig_time
14	PTB_Ranks		1	14	881.8863		<.0001	Rank for Variable PMT_to_balance
15	hpi_orig_time		1	15	871.3861		<.0001	
16	hpi_time		1	16	11393.0058		<.0001	
17	uer_time		1	17	1328.7484		<.0001	
18	CLUSTER		1	18	560.3152		<.0001	Cluster
19	Interest_Rate_orig_time		1	19	235.3794		<.0001	
20	PMT		1	20	111.3359		<.0001	
21	balance_time		1	21	15.7206		<.0001	
22	balance_orig_time		1	22	548.6927		<.0001	
23	balance_orig_Ranks		1	23	143.0349		<.0001	Rank for Variable balance_orig_time
24	balance_appreciation_Ranks		1	24	162.1880		<.0001	Rank for Variable balance_appreciation
25	interest_rate_appreciation		1	25	54.1188		<.0001	
26	REtype_PU_orig_time		1	26	11.9059		0.0006	
27	PMT_to_balance		1	27	7.5360		0.0060	
28	investor_orig_time		1	28	7.4094		0.0065	

Analysis of Maximum Likelihood Estimates						
Parameter	default_time	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-0.3742	0.2146	3.0401	0.0812
balance_time		1	0.000017	4.172E-7	1724.2535	<.0001
LTV_time		1	0.0586	0.00161	1333.9847	<.0001
interest_rate_time		1	0.1348	0.00305	1954.2163	<.0001
hpi_time		1	-0.1071	0.00120	7920.0336	<.0001
gdp_time		1	-0.0800	0.00299	715.6226	<.0001
uer_time		1	0.1662	0.00533	971.6320	<.0001
REtype_PU_orig_time	1	1	-0.0625	0.0192	10.6008	0.0011
investor_orig_time	1	1	-0.0521	0.0192	7.4074	0.0065
balance_orig_time		1	-0.00002	4.449E-7	1614.9404	<.0001
FICO_orig_time		1	-0.0296	0.000285	10812.2204	<.0001
LTV_orig_time		1	-0.0534	0.00175	926.2271	<.0001
Interest_Rate_orig_t		1	-0.0290	0.00199	212.6012	<.0001
hpi_orig_time		1	0.1262	0.00131	9287.2241	<.0001
hprate		1	9.6040	0.0951	10208.9853	<.0001
interest_rate_apprec		1	-0.0607	0.00856	50.1835	<.0001
balance_appreciation		1	-10.0877	0.1592	4013.4316	<.0001
LTV_appreciation		1	3.5688	0.1579	510.6306	<.0001
time_to_mature		1	0.0279	0.000451	3827.4396	<.0001
PMT		1	0.000027	8.027E-7	1166.0606	<.0001
PMT_to_balance		1	-12.3851	0.3278	1427.3408	<.0001
nopmt		1	0.0424	0.00154	763.5771	<.0001
FICO_Ranks	1	1	0.8072	0.00834	9360.5480	<.0001
PTB_Ranks	1	1	0.0885	0.00318	772.8873	<.0001
balance_orig_Ranks	1	1	0.0625	0.00432	209.4461	<.0001
LTV_appreciation_Ran	1	1	-0.4913	0.00723	4622.2823	<.0001
balance_appreciation	1	1	-0.0385	0.00396	94.5282	<.0001
hprate_Ranks	1	1	1.2011	0.0106	12822.7244	<.0001
CLUSTER	1	1	-0.1161	0.00487	568.0682	<.0001

The probit model has 28 steps in total, as in the report, we only put the result of the last step. From the result, we can see the AIC value, SC value and -2LogL value of this model.
The model fit statistics are measures for model fit based on −2 times the log-likelihood (−2LogL). Both the Akaike information criterion (AIC) and the Schwartz criterion (SC) are based on −2 times the log-likelihood (−2LogL) and include a penalty for the number of estimated parameters. A lower AIC, SC, or −2LogL indicates a better fit. These measures are absolute measures, which depend on the sample size. In other words, these measures cannot be used to compare models based on different sample sizes, which may be a result of the availability of dependent and independent variables. We also generalized R-squared measure for the model by using the RSQUARE after the model statement. The measure is a relative performance measure, as it includes a comparison with a noninformative model that assigns all default observations the same average default rate and is defined between zero and one.
We can see from the output that the final model consists of the variables LTV_time, interest_rate_time, FICO_orig_time, FICO_Ranks, hpi_orig_time, time_to_mature, balance_appreciation, gdp_time, time_from_first, hpa, uer_time, LTV_orig_time, interest_rate_appreciation, interest_Rate_orig_time, investor_orig_time, Retype_PU_orig_time, balance_time,etc. Figures above show that Somers' D, the accuracy ratio (AR), is 0.935 and c, the area under the ROC curve (AUROC), is 0.967 which is a very good performance.

2.2.2. Logit model

2.2.2.1. Model Construction

The SELECTION indicates our stepwise logistic regression. SAS will report the output for each of the intermediate variable selection steps.

2.2.2.2. Implementation

```
/*Logit Model*/
ods rtf file='c:\Users\lizhe\Downloads\logic_model_1.rtf';
ods graphics on;
proc logistic data=mortgage_clusters descending;
model default_time = balance_time-numeric-hpi_orig_time hprate-numeric-CLUSTER /
selection=stepwise slentry=0.05 slstay=0.01
unequalslopes=(CLUSTER hprate_Ranks FICO_Ranks REtype_CO_orig_time-numeric-REtype_SF_orig_time
PTB_Ranks balance_orig_Ranks LTV_appreciation_Ranks balance_appreciation_Ranks investor_orig_time);
run;
ods graphics off;
ods rtf close;
```

2.2.2.3. Output

The LOGISTIC Procedure

Model Information	
Data Set	WORK.MORTGAGE_CLUSTERS
Response Variable	default_time
Number of Response Levels	2
Model	binary logit
	partial proportional odds
Optimization Technique	Newton-Raphson

Number of Observations Read	622489
Number of Observations Used	621896

Response Profile		
Ordered Value	default_time	Total Frequency
1	1	15153
2	0	606743

Probability modeled is default_time='1'.

Note: 593 observations were deleted due to missing values for the response or explanatory variables.

Summary of Stepwise Selection							
	Effect			Number	Score	Wald	
Step	Entered	Removed	DF	In	Chi-Square	Chi-Square	Pr > ChiSq
1	LTV_time		1	1	4786.4012		<.0001
2	LTV_appreciation_Ranks		1	2	12192.7548		<.0001
3	LTV_appreciation		1	3	42061.7803		<.0001
4	balance_appreciation		1	4	5283.4891		<.0001
5	hprate_Ranks		1	5	17662.9587		<.0001
6	interest_rate_time		1	6	2433.7734		<.0001
7	time_to_mature		1	7	1654.8192		<.0001
8	nopmt		1	8	1065.3258		<.0001
9	gdp_time		1	9	755.8666		<.0001
10	FICO_orig_time		1	10	690.8419		<.0001
11	FICO_Ranks		1	11	10256.9061		<.0001
12	PTB_Ranks		1	12	593.3230		<.0001
13	CLUSTER		1	13	455.6189		<.0001
14	hprate		1	14	296.6149		<.0001
15	hpi_orig_time		1	15	373.2768		<.0001
16	hpi_time		1	16	3903.4773		<.0001
17	uer_time		1	17	662.4124		<.0001
18	LTV_orig_time		1	18	344.7682		<.0001
19	Interest_Rate_orig_time		1	19	93.8462		<.0001
20	PMT		1	20	50.1436		<.0001
21	balance_appreciation_Ranks		1	21	23.7418		<.0001
22	REtype_PU_orig_time		1	22	11.1680		0.0008
23	PMT_to_balance		1	23	5.6420		0.0175
24		balance_appreciation_Ranks	1	22		3.4254	0.0642
25	balance_orig_time		1	23	4.8095		0.0283
26		balance_orig_time	1	22		4.8219	0.0281

Analysis of Maximum Likelihood Estimates						
Parameter	default_time	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	0.4806	0.5027	0.9143	0.3390
LTV_time		1	0.1006	0.00413	594.1719	<.0001
interest_rate_time		1	0.2848	0.00727	1532.9029	<.0001
hpi_time		1	-0.1926	0.00329	3427.5888	<.0001
gdp_time		1	-0.1665	0.00664	629.1950	<.0001
uer_time		1	0.3073	0.0121	640.6429	<.0001
REtype_PU_orig_time	1	1	-0.1470	0.0450	10.6968	0.0011
FICO_orig_time		1	-0.0609	0.000664	8435.9847	<.0001
LTV_orig_time		1	-0.0931	0.00466	399.3387	<.0001
Interest_Rate_orig_t		1	-0.0463	0.00461	100.7239	<.0001
hpi_orig_time		1	0.2263	0.00353	4120.4609	<.0001
hprate		1	19.5451	0.2509	6069.6768	<.0001
balance_appreciation		1	-26.1960	0.3991	4307.8352	<.0001
LTV_appreciation		1	22.1626	0.5227	1797.9315	<.0001
time_to_mature		1	0.0491	0.000977	2519.3620	<.0001
PMT		1	0.000021	2.133E-6	96.2017	<.0001
PMT_to_balance		1	-20.8320	0.9745	456.9446	<.0001
nopmt		1	0.0755	0.00357	447.9327	<.0001
FICO_Ranks	1	1	1.6674	0.0195	7276.3844	<.0001
PTB_Ranks	1	1	0.2031	0.00674	908.8767	<.0001
LTV_appreciation_Ran	1	1	-1.3189	0.0192	4729.8094	<.0001
hprate_Ranks	1	1	2.7629	0.0274	10200.7701	<.0001
CLUSTER	1	1	-0.2424	0.0113	456.2217	<.0001

Odds Ratio Estimates				
Effect	default_time	Point Estimate	95% Wald Confidence Limits	
REtype_PU_orig_time		0.863	0.790	0.943
FICO_orig_time		0.941	0.940	0.942
LTV_orig_time		0.911	0.903	0.919
Interest_Rate_orig_time		0.955	0.946	0.963
hpi_orig_time		1.254	1.245	1.263
hprate		>999.999	>999.999	>999.999
balance_appreciation		<0.001	<0.001	<0.001
LTV_appreciation		>999.999	>999.999	>999.999
time_to_mature		1.050	1.048	1.052
PMT		1.000	1.000	1.000
PMT_to_balance		<0.001	<0.001	<0.001
nopmt		1.078	1.071	1.086
FICO_Ranks		5.298	5.099	5.505
PTB_Ranks		1.225	1.209	1.241
LTV_appreciation_Ranks		0.267	0.258	0.278
hprate_Ranks		15.846	15.019	16.719
CLUSTER		0.785	0.767	0.802

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	96.3	Somers' D	0.944
Percent Discordant	1.9	Gamma	0.961
Percent Tied	1.8	Tau-a	0.045
Pairs	9193976679	c	0.972

From this output figures, we can get the following infomation:
The final model consists of the variables time_to_mature, time_from_first, balance_time, LTV_time, interest_rate_time, hpi_time, gdp_time, uer_time, Retype_PU_orig_time, investor_orig_time, banlance_orig_time, FICO_orig_time, LTV_orig_time, interest_Rate_orig_time, hpi_orig_time, interest_rate_appreciation, LTV_appreciation, FICO_Ranks, etc. The coefficient corresponds to the area under the receiver operating characteristic (ROC) curve. For the moment, it suffices to say that the number 0.969 indicates a very good performance.

2.2.3. Cloglog model

2.2.3.1. Model Construction

We need to fit a Cloglog model in the same condition as in Probit model and Logit model. In the SAS code, we are going to use LINK = CLOGLOG.

2.2.3.2. Implementation

```
/*Cloglog Model*/
ods rtf file='c:\Users\lizhe\Downloads\cloglog_model_1.rtf';
ods graphics on;
proc logistic data=mortgage_clusters descending;
model default_time = balance_time-numeric-hpi_orig_time hprate-numeric-CLUSTER /link=cloglog
selection=stepwise slentry=0.05 slstay=0.01
unequalslopes=(CLUSTER hprate_Ranks FICO_Ranks REtype_CO_orig_time-numeric
-REtype_SF_orig_time PTB_Ranks balance_orig_Ranks LTV_appreciation_Ranks balance_appreciation_Ranks investor_orig_time);
run;
ods graphics off;
ods rtf close;
```

2.2.3.3. Output

The LOGISTIC Procedure

Model Information	
Data Set	WORK.MORTGAGE_CLUSTERS
Response Variable	default_time
Number of Response Levels	2
Model	binary cloglog
	partial unequal slopes
Optimization Technique	Newton-Raphson

Number of Observations Read	622489
Number of Observations Used	621896

Response Profile		
Ordered Value	default_time	Total Frequency
1	1	15153
2	0	606743

Probability modeled is default_time='1'.

Summary of Stepwise Selection								
	Effect			Number	Score	Wald		Variable
Step	Entered	Removed	DF	In	Chi-Square	Chi-Square	Pr > ChiSq	Label
1	LTV_time		1	1	4791.2325		<.0001	
2	interest_rate_time		1	2	2815.0261		<.0001	
3	gdp_time		1	3	1467.1641		<.0001	
4	LTV_appreciation_Ranks		1	4	1868.0530		<.0001	Rank for Variable LTV_appreciation
5	hpi_orig_time		1	5	6698.9005		<.0001	
6		LTV_time	1	4		0.0000	1.0000	
7	LTV_appreciation		1	5	35920.4763		<.0001	
8	hpi_time		1	6	6457.2803		<.0001	
9	hprate_Ranks		1	7	69991.6478		<.0001	Rank for Variable hprate
10	hprate		1	8	10481.6698		<.0001	
11	uer_time		1	9	2905.2857		<.0001	
12	time_to_mature		1	10	1948.1389		<.0001	
13	balance_appreciation		1	11	1318.1412		<.0001	
14	nopmt		1	12	725.5903		<.0001	
15	FICO_orig_time		1	13	834.1155		<.0001	
16	FICO_Ranks		1	14	8916.4517		<.0001	Rank for Variable FICO_orig_time
17	LTV_time		1	15	447.6542		<.0001	
18		LTV_time	1	14		0.0000	1.0000	

Analysis of Maximum Likelihood Estimates						
Parameter	default_time	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-11.8133	0.3197	1365.8062	<.0001
interest_rate_time		1	0.1723	0.00361	2284.6618	<.0001
hpi_time		1	-0.2613	0.00236	12214.6047	<.0001
gdp_time		1	-0.1569	0.00529	880.0590	<.0001
uer_time		1	0.3950	0.00862	2102.2696	<.0001
FICO_orig_time		1	-0.0435	0.000422	10643.3559	<.0001
hpi_orig_time		1	0.2979	0.00257	13476.9860	<.0001
hprate		1	18.2474	0.2047	7943.3810	<.0001
balance_appreciation		1	-10.2705	0.2180	2219.0226	<.0001
LTV_appreciation		1	7.8236	0.1487	2769.2565	<.0001
time_to_mature		1	0.0382	0.000688	3078.2632	<.0001
nopmt		1	0.0531	0.00217	601.8755	<.0001
FICO_Ranks	1	1	1.1167	0.0124	8118.8874	<.0001
LTV_appreciation_Ran	1	1	-0.3729	0.00720	2685.3744	<.0001
hprate_Ranks	1	1	2.8972	0.0209	19274.2613	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	93.6	Somers' D	0.894
Percent Discordant	4.2	Gamma	0.914
Percent Tied	2.2	Tau-a	0.043
Pairs	9193976679	c	0.947

From this output, we can see that the final model consists of the variables interest_rate_time, FICO_orig_time, balance_appreciation, FICO_Ranks and LTV_time have no impact on the default

risk, which are different from Logit model and Probit model. The number of effective independent variables in Cloglog model are less than that in Logit model and Probit model.
From the output result, we can tell the c of Cloglog model is 0.947, the Logit model is 0.972 and that of Probit model is 0.967. Therefore, the performance of the Cloglog model is only a little lower than Logit model and Probit model.

2.3. Application

2.3.1. Description of Through-the-Cycle (TTC) and Point-in-Time (PIT)

Through-the-cycle (TTC) and Point-in-time (PIT) are different modeling methodologies. Through-the-cycle (TTC) models generally abstract from the state of the overall economy by excluding macroeconomic risk drivers. Point-in-time (PIT) models explicitly control for the state of the economy.
We are going to estimate two models: a TTC model and a PIT model for comparison.
The TTC model is based on application data (information that is observable at loan origination). The model is a logit model based on the FICO_orig and LTV_orig. While the PIT model is based on application data and time-varying information. It is a logit model based on the FICO_orig and LTV_orig, as well as the following macroeconomic indicators: GDP growth rate, unemployment rate, and house price index.

2.3.2. Implementation

```
/*TTC Model*/
ods graphics on;
proc logistic data=mortgage_clusters descending;
model default_time = balance_time-numeric-hpi_orig_time hprate-numeric-CLUSTER /link=probit
selection=stepwise slentry=0.05 slstay=0.01
unequalslopes=(CLUSTER hprate_Ranks FICO_Ranks REtype_CO_orig_time-numeric-
REtype_SF_orig_time PTB_Ranks balance_orig_Ranks LTV_appreciation_Ranks balance_appreciation_Ranks investor_orig_time);
output out=probabilities_TTC predicted=PD_TTC_time;
run;
ods graphics off;
/*PIT Model*/
ods graphics on;
proc logistic data=mortgage_clusters descending;
model default_time = balance_time-numeric-hpi_orig_time hprate-numeric-CLUSTER /link=probit
selection=stepwise slentry=0.05 slstay=0.01
unequalslopes=(CLUSTER hprate_Ranks FICO_Ranks REtype_CO_orig_time-numeric-
REtype_SF_orig_time PTB_Ranks balance_orig_Ranks LTV_appreciation_Ranks balance_appreciation_Ranks investor_orig_time);
output out=probabilities_PIT predicted=PD_PIT_time;
run;
ods graphics off;
```

2.3.3. Output

Output for Through-the-cycle (TTC)

The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.MORTGAGE_CLUSTERS
Response Variable	default_time
Number of Response Levels	2
Model	binary probit
Optimization Technique	Newton-Raphson ridge

Number of Observations Read	622489
Number of Observations Used	621896

Response Profile		
Ordered Value	default_time	Total Frequency
1	1	15153
2	0	606743

Probability modeled is default_time='1'.

Summary of Stepwise Selection								
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	LTV_time		1	1	4763.4893		<.0001	
2	LTV_appreciation_Ranks		1	2	6431.7625		<.0001	Rank for Variable LTV_appreciation
3	hprate		1	3	4889.0283		<.0001	
4	hprate_Ranks		1	4	4107.9834		<.0001	Rank for Variable hprate
5	gdp_time		1	5	3299.3527		<.0001	
6	balance_appreciation		1	6	2297.0050		<.0001	
7	LTV_appreciation		1	7	22352.0577		<.0001	
8	time_to_mature		1	8	5830.1781		<.0001	
9	interest_rate_time		1	9	2489.1500		<.0001	
10	nopmt		1	10	2156.9391		<.0001	
11	LTV_orig_time		1	11	2029.8455		<.0001	
12	FICO_orig_time		1	12	1687.3399		<.0001	
13	FICO_Ranks		1	13	15657.5543		<.0001	Rank for Variable FICO_orig_time
14	PTB_Ranks		1	14	881.8863		<.0001	Rank for Variable PMT_to_balance
15	hpi_orig_time		1	15	871.3861		<.0001	
16	hpi_time		1	16	11393.0058		<.0001	
17	uer_time		1	17	1328.7484		<.0001	
18	CLUSTER		1	18	560.3152		<.0001	Cluster
19	Interest_Rate_orig_time		1	19	235.3794		<.0001	
20	PMT		1	20	111.3359		<.0001	
21	balance_time		1	21	15.7206		<.0001	
22	balance_orig_time		1	22	548.6927		<.0001	

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.3742	0.2146	3.0401	0.0812
balance_time	1	0.000017	4.172E-7	1724.2535	<.0001
LTV_time	1	0.0586	0.00161	1333.9847	<.0001
interest_rate_time	1	0.1348	0.00305	1954.2163	<.0001
hpi_time	1	-0.1071	0.00120	7920.0336	<.0001
gdp_time	1	-0.0800	0.00299	715.6226	<.0001
uer_time	1	0.1662	0.00533	971.6320	<.0001
REtype_PU_orig_time	1	-0.0625	0.0192	10.6008	0.0011
investor_orig_time	1	-0.0521	0.0192	7.4074	0.0065
balance_orig_time	1	-0.00002	4.449E-7	1614.9404	<.0001
FICO_orig_time	1	-0.0296	0.000285	10812.2204	<.0001
LTV_orig_time	1	-0.0534	0.00175	926.2271	<.0001
Interest_Rate_orig_t	1	-0.0290	0.00199	212.6012	<.0001
hpi_orig_time	1	0.1262	0.00131	9287.2241	<.0001
hprate	1	9.6040	0.0951	10208.9853	<.0001
interest_rate_apprec	1	-0.0607	0.00856	50.1835	<.0001
balance_appreciation	1	-10.0877	0.1592	4013.4316	<.0001
LTV_appreciation	1	3.5688	0.1579	510.6306	<.0001
time_to_mature	1	0.0279	0.000451	3827.4396	<.0001
PMT	1	0.000027	8.027E-7	1166.0606	<.0001
PMT_to_balance	1	-12.3851	0.3278	1427.3408	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	96.6	Somers' D	0.932
Percent Discordant	3.4	Gamma	0.932
Percent Tied	0.0	Tau-a	0.044
Pairs	9193976679	c	0.966

Output for Point-in-time (PIT)

The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.MORTGAGE_CLUSTERS
Response Variable	default_time
Number of Response Levels	2
Model	binary probit
Optimization Technique	Newton-Raphson ridge

Number of Observations Read	622489
Number of Observations Used	621896

Response Profile		
Ordered Value	default_time	Total Frequency
1	1	15153
2	0	606743

Probability modeled is default_time='1'.

Summary of Stepwise Selection								
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	LTV_time		1	1	4763.4893		<.0001	
2	LTV_appreciation_Ranks		1	2	6431.7625		<.0001	Rank for Variable LTV_appreciation
3	hprate		1	3	4889.0283		<.0001	
4	hprate_Ranks		1	4	4107.9834		<.0001	Rank for Variable hprate
5	gdp_time		1	5	3299.3527		<.0001	
6	balance_appreciation		1	6	2297.0050		<.0001	
7	LTV_appreciation		1	7	22352.0577		<.0001	
8	time_to_mature		1	8	5830.1781		<.0001	
9	interest_rate_time		1	9	2489.1500		<.0001	
10	nopmt		1	10	2156.9391		<.0001	
11	LTV_orig_time		1	11	2029.8455		<.0001	
12	FICO_orig_time		1	12	1687.3399		<.0001	
13	FICO_Ranks		1	13	15657.5543		<.0001	Rank for Variable FICO_orig_time
14	PTB_Ranks		1	14	881.8863		<.0001	Rank for Variable PMT_to_balance
15	hpi_orig_time		1	15	871.3861		<.0001	
16	hpi_time		1	16	11393.0058		<.0001	
17	uer_time		1	17	1328.7484		<.0001	
18	CLUSTER		1	18	560.3152		<.0001	Cluster
19	Interest_Rate_orig_time		1	19	235.3794		<.0001	
20	PMT		1	20	111.3359		<.0001	
21	balance_time		1	21	15.7206		<.0001	
22	balance_orig_time		1	22	548.6927		<.0001	

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.3742	0.2146	3.0401	0.0812
balance_time	1	0.000017	4.172E-7	1724.2535	<.0001
LTV_time	1	0.0586	0.00161	1333.9847	<.0001
interest_rate_time	1	0.1348	0.00305	1954.2163	<.0001
hpi_time	1	-0.1071	0.00120	7920.0336	<.0001
gdp_time	1	-0.0800	0.00299	715.6226	<.0001
uer_time	1	0.1662	0.00533	971.6320	<.0001
REtype_PU_orig_time	1	-0.0625	0.0192	10.6008	0.0011
investor_orig_time	1	-0.0521	0.0192	7.4074	0.0065
balance_orig_time	1	-0.00002	4.449E-7	1614.9404	<.0001
FICO_orig_time	1	-0.0296	0.000285	10812.2204	<.0001
LTV_orig_time	1	-0.0534	0.00175	926.2271	<.0001
Interest_Rate_orig_t	1	-0.0290	0.00199	212.6012	<.0001
hpi_orig_time	1	0.1262	0.00131	9287.2241	<.0001
hprate	1	9.6040	0.0951	10208.9853	<.0001
interest_rate_apprec	1	-0.0607	0.00856	50.1835	<.0001
balance_appreciation	1	-10.0877	0.1592	4013.4316	<.0001
LTV_appreciation	1	3.5688	0.1579	510.6306	<.0001
time_to_mature	1	0.0279	0.000451	3827.4396	<.0001
PMT	1	0.000027	8.027E-7	1166.0606	<.0001
PMT_to_balance	1	-12.3851	0.3278	1427.3408	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	96.6	Somers' D	0.932
Percent Discordant	3.4	Gamma	0.932
Percent Tied	0.0	Tau-a	0.044
Pairs	9193976679	c	0.966

From the comparison of TTC and PIT, we can tell that both PIT and TTC model have the same c value which is 0.966. therefore, we can conclude that PIT and TTC has the same performance.

2.4. Conclusion

Different functions have important applications with a little difference of link functions. For example, the logit link function has useful properties if the resulting mean PDs have to be calibrated to a different level. The probit link function is particularly useful for estimating parameters that are in line with the internal ratings-based models under the Basel regulations, as these assume the probit link.
In our mortgage analysis, we fit three different kinds of non-linear regression model. The area under the ROC curve of these models are shown in the following table.

model	Probit Regression	Logit Regression	Cloglog Regression
AUROC	0.967	0.972	0.947

We can see from the table that these three regressions all have good performance. But Logit Regression has a little better performance than others. We also try TTC and PIT modeling and we know from the result that PIT modeling methodology has same performance with TTC in our situation.

3. Loss Given Default (LGD)

3.1. Description of Loss Given Default

Other than modeling probability of defaults and exposures at defaults, we also want to model severity of the loss given that default events have happened.

3.2. Loss Given Default Models

3.2.1. LGD Computing

In the dataset, we do not have LGD value ready to use, in this case we need to compute our own LGD values. To compute LGD values, the mortgage dataset provided enough information for use

to use. We are going to use the LGD formula as showed below:

$$LGD = 100 \times \frac{CUPB + ACRINT + FCLEXP + PROEXP - NETREC}{CUPB}$$

Based on the definition online, we use outstanding balance at default time as unpaid balance at default; for the accrued interest we used interest rate at default time to calculate the interest amount. Given the data, we assume the salvage value of mortgage is similar as the property value. The model assume that the bank would sell the houses three months after default, so we can believe that houses will not significantly be depreciated in these three months period of time. After we computed LGD values, we are going to apply linear transformation on LGD to get other values to prepare further models.

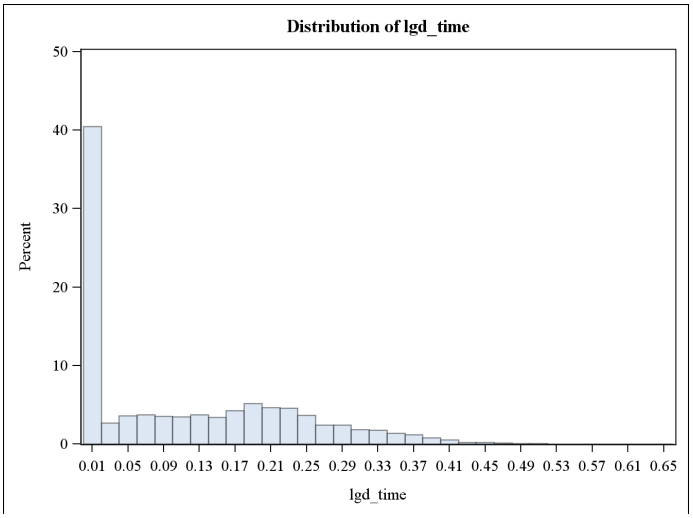
3.2.2. Data Sets Preparation

After we compute LGD values, we can tell there are some values are not good enough for the model to use. First, we have some negative LGD values. We can assume the reason of it is because all the loans are not only default in the first year. For all these negative LGD values, we decided to use 0.00001 to instead them. In this way, we do not lose any value and we also can use them as in positive values. Second, we have some large values that are bigger than 1. We cannot continuous out modeling with those values because they are going to affect our Beta distribution modeling. In this case, we decided to give all these values a new value with 0.9999.

Moments			
N	15153	Sum Weights	15153
Mean	0.11216123	Sum Observations	1699.57917
Std Deviation	0.11990944	Variance	0.01437827
Skewness	0.75668319	Kurtosis	-0.4165109
Uncorrected SS	408.486505	Corrected SS	217.859609
Coeff Variation	106.908098	Std Error Mean	0.0009741

Basic Statistical Measures			
Location		Variability	
Mean	0.112161	Std Deviation	0.11991
Median	0.077656	Variance	0.01438
Mode	0.000010	Range	0.65341
		Interquartile Range	0.20430

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	115.1433	Pr > |t|	<.0001
Sign	M	7576.5	Pr >= |M|	<.0001
Signed Rank	S	57407141	Pr >= |S|	<.0001



3.2.3. Model Results

	Variable #	R-square	AIC
linear	10	0.8862	
Logistic	8	0.7472	
Probit	10	0.8864	
Non-linear regression	4		-55047
Fractional Logit Regression	2		8537.1
Beta Regression	5		-113e3
Tobit Regression	2		-40942
Heckman Model			4096

In this section, we fit data to the model so that we can model the probabilities that the LGD events that have happened. We ran both linear and non-linear models. For linear models, we have linear regression on LGDs, Logistic, Probit. While for non-linear models, we have nonlinear regression, fractional logit regression, Beta regression, Tobit regression, and Heckman model. Among all the linear models we have, we conclude that Probit model fits best because it has the highest R-square as it shown at the table above. As for the non-linear regression, we can know that Beta Regression perform best because it has the lowest AIC value.

3.3. Conclusion

After we run the validation on the model we choose at the part of result, which is Probit model for linear model and Beta regression for non-linear model, we have the R-squared and AIC value shown below:

R-square	Train	Test
Probit	0.8880	0.8829

AIC	Train	Test
Beta Regression	-56126	-57372

We can see that the both for both models, the train set and the test set match with each other.

Thus, both models would work with our datasets. However, it would be more convenient to use linear regression models but the non-linear regression models might perform more stable. So when come to practical use, we should also take real factor into account to choose the best model. To further improve the models, we may work further to control the amount of the variables in the variable to improve the efficiency of the model.

4. Exposure at Default (EAD)

4.1. Description of Exposure at Default

Exposure at default (EAD) is defined as the nominal outstanding balance, net of specific provisions. In other words, EAD represents the net outstanding debt. It also has a linear impact on both the expected loss and the Basel capital. That's also the reason why accuracy is important in the model.

4.2. Exposure at Default (EAD) Models

4.2.1. Conversion Measures and Data Preparation

4.2.1.1. Description of Conversion Measures

We can model exposures at default directly by building a model for exposure amount. We also can relate the EAD to a scaling variable and derive conversion measures. As we know, models for conversion measures are generally more robust than EAD models.

The following table shows four common conversion measures:

Conversion Measure	Formula
Credit Conversion Factor (CCF)	EAD = Drawn + CCF * (Limit-Drawn)
Credit Equivalent (CEQ)	EAD = Drawn + CEQ * Limit
Limit Conversion Factor (LCF)/Loan Equivalent (LEQ)	EAD = LCF * Limit
Used Amount Conversion Factor (UACF)	EAD = UACF * Drawn

The credit conversion factor (CCF) is defined as the portion of the undrawn amount that will be converted into credit. Note that the undrawn amount is equal to the limit minus the drawn amount. The EAD thus becomes the drawn amount plus the CCF times the limit minus the drawn amount. The credit equivalent (CEQ), is defined as the portion of the limit likely to be converted into credit. The EAD is then defined as the drawn amount plus the CEQ times the limit. The limit

conversion factor (LCF), or loan equivalent (LEQ), is defined as a fraction of the limit representing the total exposure. The EAD is then defined as the LCF or LEQ times the limit. Finally, the used amount conversion factor (UACF) is defined using the drawn amount as the reference. Hence, the EAD is then computed as the UACF times the drawn amount.

4.2.1.2. Data Preparation

For defaulted exposures, the EAD at the moment of default can be determined. We need to understand how we can create the development sample for EAD. Considering a period δ_t before the time of default to determine the risk factors and drawn amount. The problem is the determination of the time lag between the observed exposure amount at default and the observed drawn amount and limit. We choose to use variable time horizon method. This approach is a variant of the fixed time horizon approach whereby several reference times within a chosen time horizon are used to determine the drawn or undrawn amounts and risk factors. This is the dataset of exposures with flexible payment schedules and that low-risk mortgage borrowers often make prepayments, while high-risk borrowers usually do not.

Data preparation steps:

First - Set up data by computing the lagged exposure amounts (balance_time) for mortgage borrowers as a measure for the drawn amount at the beginning of the reference period. Consistent with the variable time horizon method, we consider one to four lags under the assumption that one period equals one quarter (a lag of four periods is equal to one year). The ARRAY command in SAS allows us to set the first observations for which no lagged value is available to a missing value for multiple lags. The data are first sorted by the variable "id." We apply the calculations on all observations because later we will estimate regression models that control for the selecting default event.

Second - Generate the dependent variable based on the four concepts: CCF, CEQ, LCF, and UACF.

As the formula shown in the following table.

Measure	Formula	Lower Bound	Upper Bound	Transformation
CCF	(EAD-Drawn)/(Limit-Drawn)	$-\infty$	1	-ln(1-CCF)
CEQ	(EAD-Drawn)/Limit	-1	1	ln((1+CEQ)/(1-CEQ))
LCF	EAD/Limit	0	1	ln(LCF/(1-LCF))
UACF	EAD/Drawn	0	∞	ln(UACF)

We define the limit as the outstanding loan amount at origination (balance_orig_time). The drawn amount is the outstanding loan amount prior to the observation period. We focus on the four-period lag (lag4), which is equal to one year. The exposure amount is the outstanding loan amount in the observation period (balance_time). Then, we have the variables' exposure and drawn amount by the credit limit. As we are interested in EAD and mortgages are exposed to prepayments, we do not impose the regulatory floor of the exposure by the drawn amount. We

add one restriction: CCF is set to zero if the drawn amount is equal to the limit (IF drawn=limit THEN CCF=0;) as the data has observations where the limit is completely drawn.

Third - Compute the 1st and 99th percentiles of the conversion measures. Floor measures at 1st percentile and cap measures at 99th percentile. As result, for CCF and LCF, we substitute an observation of one with 0.9999999. We also transform the variables CCF, CEQ, LCF, and UACF by using transformation functions to match the range of normally distributed residuals in regressions.

Last - Generate moments of the four dependent variables for the complete data set using PROC MEANS for CCF, CEQ, LCF, and UACF, as well as their transforms for all observations and by the default indicator. We also plot histograms using PROC UNIVARIATE for CCF, CEQ, LCF, and UACF, as well as their transforms.

Implementation of Data Preparation

```
/*Create appreciations to indicates the change of the loan*/
/*ead*/
data data.mortgage_variable;
    set data.mortgage;
    hprate = (hpi_time-hpi_orig_time)/hpi_orig_time;
    interest_rate_appreciation=(interest_rate_time-Interest_Rate_orig_time)/Interest_Rate_orig_time;
    balance_appreciation=(balance_time-balance_orig_time)/balance_orig_time;
    LTV_appreciation=(LTV_time-LTV_orig_time)/LTV_orig_time;
    time_to_mature=mat_time-time;
run;
proc sort data=data.mortgage_variable;
    by default_time;
run;
proc rank data=data.mortgage_variable out=data.mortgage_variable groups=10;
    var FICO_orig_time;
    ranks FICO_Ranks;
    by default_time;
run;


proc sort data=data.mortgage_variable; by id; run;

data data.mortgage_variable;
    set data.mortgage_variable;
    hprate = (hpi_time-hpi_orig_time)/hpi_orig_time;
    interest_rate_appreciation=(interest_rate_time-Interest_Rate_orig_time)/Interest_Rate_orig_time;
    balance_appreciation=(balance_time-balance_orig_time)/balance_orig_time;
    LTV_appreciation=(LTV_time-LTV_orig_time)/LTV_orig_time;
    time_to_mature=mat_time-time;
run;


data mortgage(drop=i count);
    set data.mortgage_variable;
    by id;
```

```
    /* create lagged variables */
    array x(*) lag1-lag4;
    lag1=lag1(balance_time);
    lag2=lag2(balance_time);
    lag3=lag3(balance_time);
    lag4=lag4(balance_time);

    /* reset count at the start of each new by-group */
    if first.id then count=1;

    /* assign missing values to first observation of a by-group */
    do i=count to dim(x);
        x(i)=.;
    end;
    count+1;
run;
```

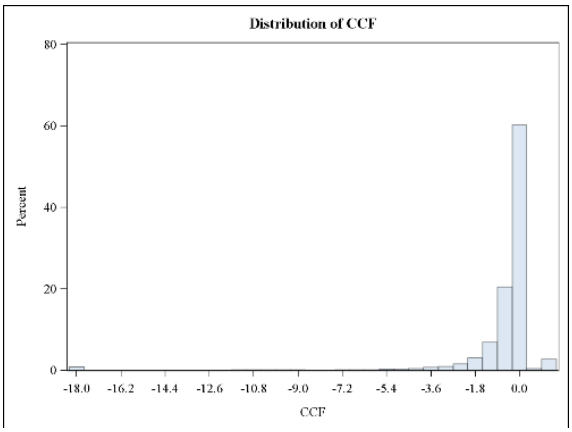
```
data mortgage1(where=(drawn ne . and limit ne . and exposure ne . and exposure ne 0));
    set mortgage;

    /*Definitions*/
    drawn=lag4;
    limit=balance_orig_time;
    exposure=balance_time;

    /*Caps for exposure and draw*/
    if exposure>limit then exposure=limit;
    if drawn>limit then drawn=limit;

    /*Conversion measures*/
    if drawn=limit then CCF=0;
    else CCF=(exposure-drawn)/(limit-drawn);

    if limit=0 then CEQ=0;
    else CEQ=(exposure-drawn)/limit;

    if limit=0 then LCF=0;
    else LCF=exposure/limit;
```

```
    if drawn=0 then UACF=0;
    else UACF=exposure/drawn;
run;

proc means  p1 p99;
    var CCF CEQ LCF UACF;
run;

data mortgage2;
    set mortgage1;

    /* Floors*/
    if CCF<=-18.0502849 then CCF=-18.0502849;
    if CEQ<=-0.1297378 then CEQ=-0.1297378;
    if LCF<=0.3724166 then LCF=0.3724166;
    if UACF<=0.7492250 then UACF=0.7492250;

    /*Caps*/
    if CCF>=0.9999999 then CCF=0.9999999;
    if CEQ>=0.0102912 then CEQ=0.0102912;
    if LCF>=0.9999999 then LCF=0.9999999;
    if UACF>=1.0105358 then UACF=1.0105358;

    /*Transformations*/
    CCF_t=-log(1-CCF);
    CEQ_t=log((1+CEQ)/(1-CEQ));
    LCF_t=log(LCF/(1-LCF));
    UACF_t=log(UACF);

run;

ods rtf file='c:\Users\lizhe\Downloads\ead_1.rtf';
proc means data=mortgage2(where=(default_time=1));
    var CCF CEQ LCF UACF CCF_t CEQ_t LCF_t UACF_t;
run;

ods graphics on;
proc univariate data=mortgage2(where=(default_time=1));
    var CCF CEQ LCF UACF CCF_t CEQ_t LCF_t UACF_t;
    histogram;
run;
```

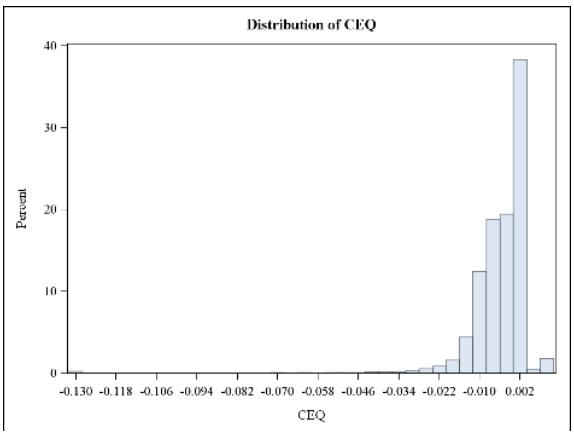
Output of Data Preparation

Variable	N	Mean	Std Dev	Minimum	Maximum
CCF	11673	-0.6358660	2.0406573	-18.0502849	0.9999999
CEQ	11673	-0.0044892	0.0095413	-0.1297378	0.0102912
LCF	11673	0.9781223	0.0505083	0.3724166	0.9999999
UACF	11673	0.9949136	0.0146982	0.7492250	1.0105358
CCF_t	11673	0.1218104	2.6773378	-2.9470821	16.1180957
CEQ_t	11673	-0.0089835	0.0191379	-0.2609463	0.0205831
LCF_t	11673	8.1114383	5.5361960	-0.5218635	16.1180956
UACF_t	11673	-0.0052234	0.0162799	-0.2887159	0.0104807

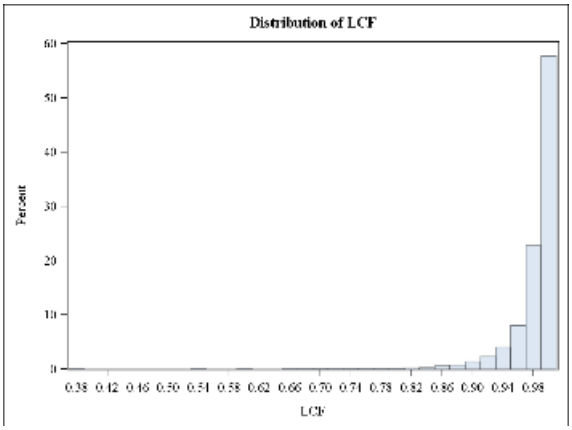
Histograms for CCF, CEQ, LCF, and UACF, as well as their transforms



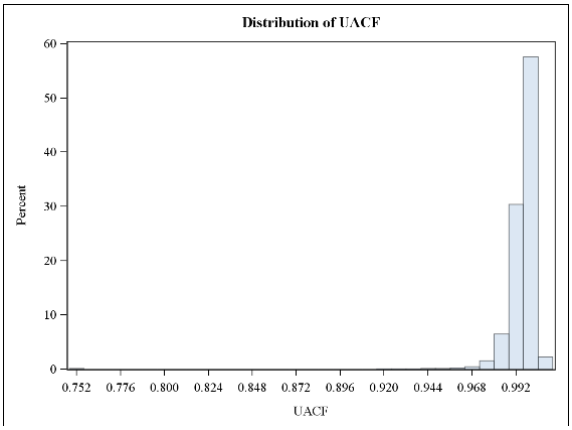
Histogram Credit Conversion Factor (CCF)



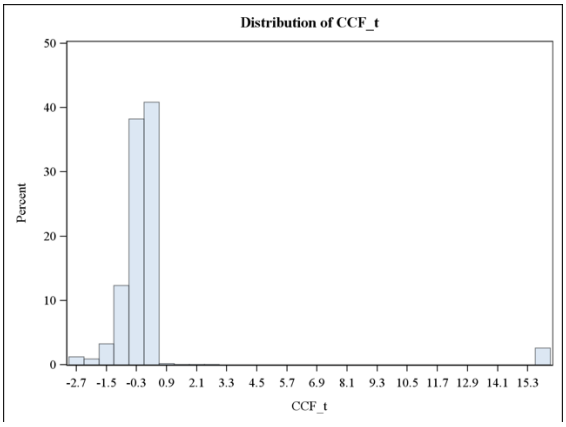
Histogram Credit Equivalent (CEQ)



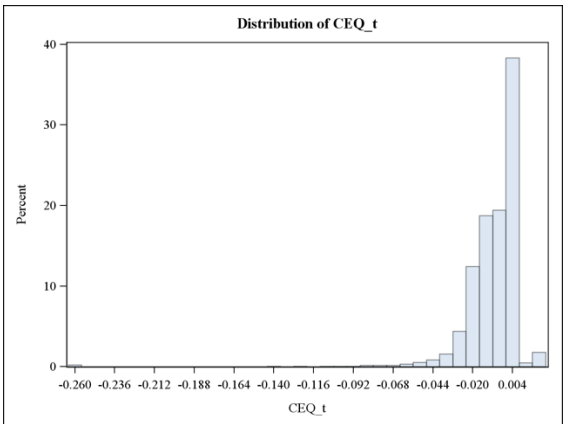
Histogram Limit Conversion Factor (LCF)



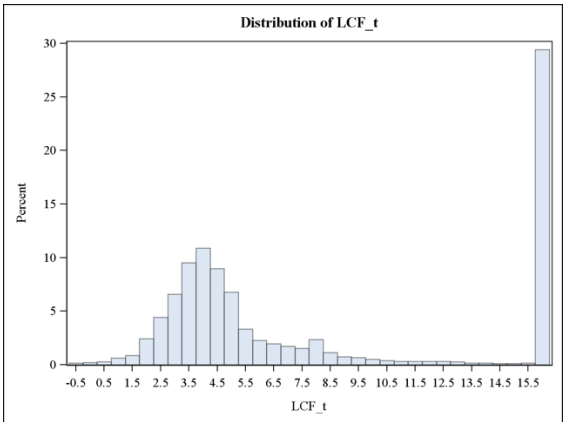
Histogram Used Amount Conversion Factor (UACF)



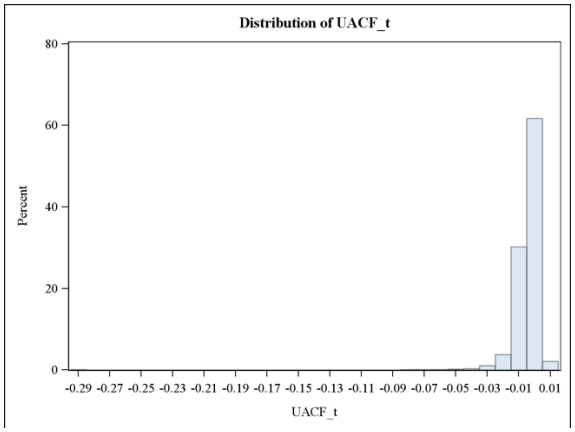
Histogram Transformed Credit Conversion Factor (CCF_t)



Histogram Transformed Credit Equivalent (CEQ_t)



Histogram Transformed Limit Conversion Factor (LCF_t)



Histogram Transformed Used Amount Conversion Factor (UACF_t)

4.2.2. Linear Regression

4.2.2.1. Model Construction

We estimate a linear regression model for each transform of CCF, CEQ, LCF, and UACF. And then use stepwise process to select our covariates from the following variables: LTV_time time_since_orig interest_rate_time uer_time LTV_appreciation time_to_mature hpa interest_rate_appreciation gdp_time REtype_CO_orig_time REtype_PU_orig_time REtype_SF_orig_time investor_orig_time.

4.2.2.2. Implementation

```
/***optimal linear regression***/
proc reg data=mortgage2(where=(default_time=1))plots(maxpoints=20000 stats=all)=diagnostics;
model CCF= uer_time interest_rate_appreciation time_to_mature LTV_appreciation hprate interest_rate_time;
run;

proc reg data=mortgage2(where=(default_time=1))plots(maxpoints=20000 stats=all)=diagnostics;
model CEQ=time_to_mature LTV_appreciation hprate uer_time gdp_time interest_rate_appreciation
interest_rate_time investor_orig_time REtype_SF_orig_time;
run;

proc reg data=mortgage2(where=(default_time=1))plots(maxpoints=20000 stats=all)=diagnostics;
model LCF=time_to_mature LTV_appreciation hprate time_since_orig uer_time gdp_time interest_rate_appreciation REtype_SF_orig_time;
run;

proc reg data=mortgage2(where=(default_time=1))plots(maxpoints=20000 stats=all)=diagnostics;
model UACF= time_to_mature LTV_appreciation hprate uer_time gdp_time interest_rate_appreciation
interest_rate_time Retype_SF_orig_time investor_orig_time;
run;
```

4.2.2.3. Output

The REG Procedure
Model: MODEL1
Dependent Variable: CCF

Number of Observations Read	11673
Number of Observations Used	9787
Number of Observations with Missing Values	1886

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1481.03600	246.83933	60.77	<.0001
Error	9780	39726	4.06199		
Corrected Total	9786	41207			

Root MSE	2.01544	R-Square	0.0359
Dependent Mean	-0.62063	Adj R-Sq	0.0353
Coeff Var	-324.74184		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.84047	0.23982	-3.50	0.0005
uer_time	1	0.13244	0.01480	8.95	<.0001
interest_rate_appreciation	1	0.17341	0.02474	7.01	<.0001
time_to_mature	1	-0.00882	0.00163	-5.41	<.0001
LTV_appreciation	1	1.62119	0.21483	7.55	<.0001
hprate	1	1.65461	0.20383	8.12	<.0001
interest_rate_time	1	-0.00095256	0.01073	-0.09	0.9293

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: CEQ

Number of Observations Read	11673
Number of Observations Used	9787
Number of Observations with Missing Values	1886

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	0.25690	0.02854	411.46	<.0001
Error	9777	0.67827	0.00006937		
Corrected Total	9786	0.93517			

Root MSE	0.00833	R-Square	0.2747
Dependent Mean	-0.00463	Adj R-Sq	0.2740
Coeff Var	-179.73273		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.01781	0.00101	-17.61	<.0001
time_to_mature	1	0.00013059	0.00000674	19.37	<.0001
LTV_appreciation	1	0.03564	0.00090171	39.52	<.0001
hprate	1	0.02124	0.00084316	25.19	<.0001
uer_time	1	-0.00088923	0.00006125	-14.52	<.0001
gdp_time	1	0.00030501	0.00003819	7.99	<.0001
interest_rate_appreciation	1	-0.00060511	0.00010228	-5.92	<.0001
interest_rate_time	1	0.00016631	0.00004561	3.65	0.0003
investor_orig_time	1	-0.00083461	0.00024195	-3.45	0.0006
REtype_SF_orig_time	1	-0.00059094	0.00017324	-3.41	0.0006

The REG Procedure
Model: MODEL1
Dependent Variable: LCF

Number of Observations Read	11673
Number of Observations Used	11673

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	11.37028	2.84257	1801.97	<.0001
Error	11668	18.40604	0.00158		
Corrected Total	11672	29.77632			

Root MSE	0.03972	R-Square	0.3819
Dependent Mean	0.97812	Adj R-Sq	0.3816
Coeff Var	4.06059		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.75851	0.00385	196.91	<.0001
LTV_time	1	0.00087266	0.00002170	40.22	<.0001
time_to_mature	1	0.00143	0.00002875	49.72	<.0001
interest_rate_time	1	0.00126	0.00018929	6.64	<.0001
uer_time	1	-0.00382	0.00024420	-15.64	<.0001

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: UACF

Number of Observations Read	11673
Number of Observations Used	9787
Number of Observations with Missing Values	1886

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	0.33162	0.03907	345.88	<.0001
Error	9777	1.66970	0.00017078		
Corrected Total	9786	2.00132			

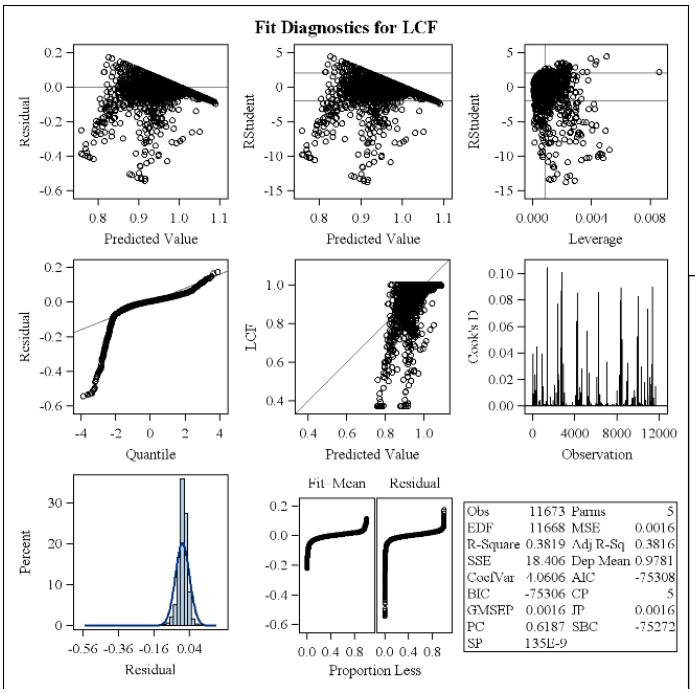
Root MSE	0.01307	R-Square	0.2415
Dependent Mean	0.99472	Adj R-Sq	0.2408
Coeff Var	1.31375		

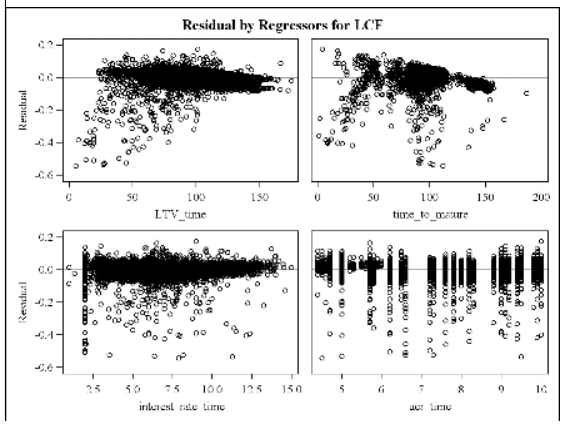
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.97660	0.00159	615.29	<.0001
time_to_mature	1	0.00017731	0.00001058	16.76	<.0001
LTV_appreciation	1	0.03358	0.00141	37.87	<.0001
hprate	1	0.03236	0.00132	24.46	<.0001
uer_time	1	-0.00139	0.00009610	-14.44	<.0001
gdp_time	1	0.00052851	0.00005992	8.82	<.0001
interest_rate_appreciation	1	-0.00119	0.00016048	-7.43	<.0001
interest_rate_time	1	0.00035237	0.00007157	4.92	<.0001
REtype_SF_orig_time	1	-0.00080361	0.00027181	-2.96	0.0031
investor_orig_time	1	-0.00102	0.00037962	-2.69	0.0072

conversion measures	CCF	CEQ	LCF	UACF
R-squared	0.0359	0.2747	0.3819	0.2415

The R-squared is highest for LCF and then we give the Linear Regression Fit Diagnostics and Residuals for LCF in the following figure.





4.2.3. Transformed Linear Regression

4.2.3.1. Model Construction

We use the similar method to estimate a transformed linear regression model for each transform of CCF, CEQ, LCF, and UACF.

4.2.3.2. Implementation

```
proc reg data=mortgage2(where=(default_time=1))plots(maxpoints=20000 stats=all)=diagnostics;
   model CCF_t=LTV_time time_to_mature interest_rate_time uer_time;
run;
```

```
proc reg data=mortgage2(where=(default_time=1))plots(maxpoints=20000 stats=all)=diagnostics;
   model CEQ_t=LTV_time time_to_mature interest_rate_time uer_time;
run;
```

```
proc reg data=mortgage2(where=(default_time=1))plots(maxpoints=20000 stats=all)=diagnostics;
   model LCF_t=LTV_time time_to_mature interest_rate_time uer_time;
run;
```

```
proc reg data=mortgage2(where=(default_time=1))plots(maxpoints=20000 stats=all)=diagnostics;
   model UACF_t=LTV_time time_to_mature interest_rate_time uer_time;
run;
```

4.3.2.3. Output

The REG Procedure
Model: MODEL1
Dependent Variable: CCF_t

Number of Observations Read	11673
Number of Observations Used	11673

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2277.68434	569.42108	81.63	<.0001
Error	11668	81389	6.97539		
Corrected Total	11672	83667			

Root MSE	2.64110	R-Square	0.0272
Dependent Mean	0.12181	Adj R-Sq	0.0269
Coeff Var	2168.20177		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.53850	0.25615	-2.10	0.0355
LTV_time	1	0.00997	0.00144	6.91	<.0001
time_to_mature	1	-0.00061195	0.00191	-0.32	0.7489
interest_rate_time	1	-0.11393	0.01259	-9.05	<.0001
uer_time	1	0.07454	0.01624	4.59	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: CEQ_t

Number of Observations Read	11673
Number of Observations Used	11673

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.75002	0.18750	620.66	<.0001
Error	11668	3.52497	0.00030211		
Corrected Total	11672	4.27498			

Root MSE	0.01738	R-Square	0.1754
Dependent Mean	-0.00898	Adj R-Sq	0.1752
Coeff Var	-193.48001		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.06897	0.00169	-40.91	<.0001
LTV_time	1	0.00022337	0.00000949	23.53	<.0001
time_to_mature	1	0.00036785	0.00001258	29.23	<.0001
interest_rate_time	1	0.00015942	0.00008284	1.92	0.0543
uer_time	1	-0.00033576	0.00010686	-3.14	0.0017

The REG Procedure
Model: MODEL1
Dependent Variable: LCF_t

Number of Observations Read	11673
Number of Observations Used	11673

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	83024	20756	881.56	<.0001
Error	11668	274717	23.54447		
Corrected Total	11672	357741			

Root MSE	4.85226	R-Square	0.2321
Dependent Mean	8.11144	Adj R-Sq	0.2318
Coeff Var	59.82002		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.08206	0.47060	0.17	0.8616
LTV_time	1	0.07273	0.00265	27.44	<.0001
time_to_mature	1	0.08353	0.00351	23.78	<.0001
interest_rate_time	1	-0.74674	0.02313	-32.29	<.0001
uer_time	1	-0.35763	0.02983	-11.99	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: UACF_t

Number of Observations Read	11673
Number of Observations Used	11673

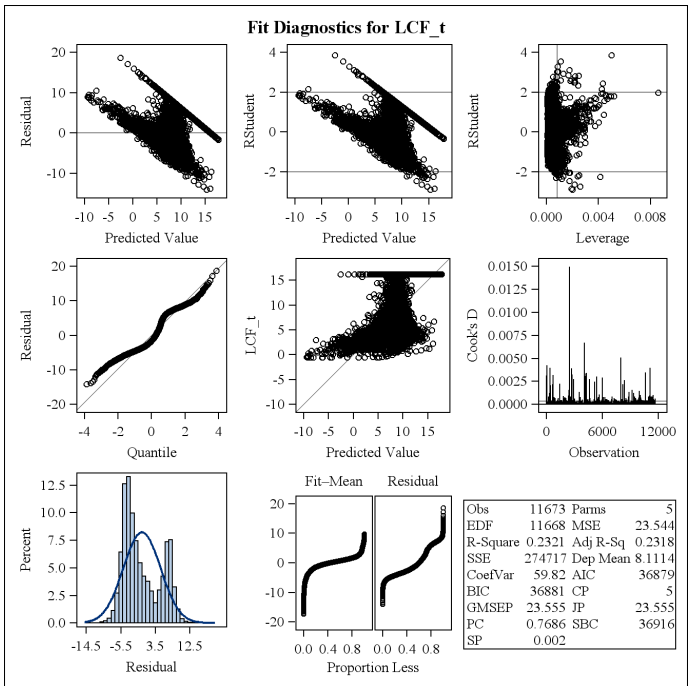
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.40084	0.10021	434.24	<.0001
Error	11668	2.69263	0.00023077		
Corrected Total	11672	3.09347			

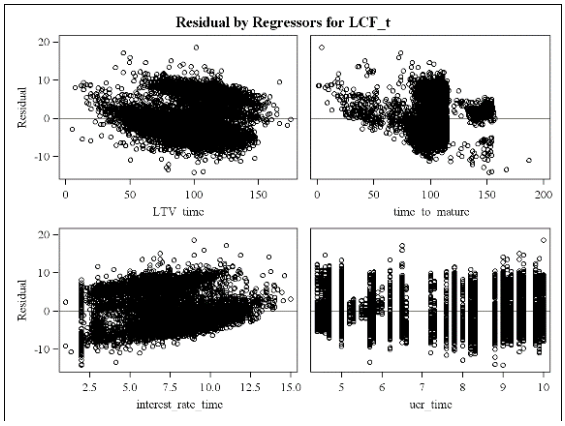
Root MSE	0.01519	R-Square	0.1296
Dependent Mean	-0.00522	Adj R-Sq	0.1293
Coeff Var	-290.82821		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.04937	0.00147	-33.51	<.0001
LTV_time	1	0.00016951	0.00000830	20.43	<.0001
time_to_mature	1	0.00026167	0.00001100	23.79	<.0001
interest_rate_time	1	0.00027746	0.00007240	3.83	0.0001
uer_time	1	-0.00034668	0.00009340	-3.71	0.0002

conversion measures	CCF_t	CEQ_t	LCF_t	UACF_t
R-squared	0.0272	0.1754	0.2318	0.1296

LCF_t has the highest R-squared value. But the R-squared for transformed linear regression are lower than linear regression overall. And then we give the Linear Regression Fit Diagnostics and Residuals for LCF_t in the following figure.





4.2.4. Non-linear Regression

4.2.4.1. Model Construction

We estimate a non-linear model for LCF, as this might provide a better fit of the residuals. In this non-linear model, we choose LTV_appreciation hpa time_since_orig uer_time gdp_time interest_rate_appreciation REtype_SF_orig_time as covariates.

4.2.4.2. Implementation

```
/**** Nonlinear Regression ****/
ODS GRAPHICS ON;
PROC NLMIXED DATA = mortgage2 TECH = TRUREG;
PARMS b0=0 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0 sigma=1;
xb = b0 + b1 * time_to_mature + b2*LTV_appreciation + b3*hprate +b4*uer_time+
b5*gdp_time+b6*interest_rate_appreciation+b7*REtype_SF_orig_time;
/*
xb = b0 + b1 * LTV_time + b2*time_since_orig + b3*interest_rate_time + b4*uer_time;
*/
mu = 1 / (1 + EXP(- xb));
lh = PDF('NORMAL', LCF, mu, sigma);
ll = LOG(lh);
model LCF~general(ll);
predict mu out=out_mu_nl;
RUN;
ODS GRAPHICS OFF;
```

4.2.4.3. Output

The NLMIXED Procedure

Specifications	
Data Set	WORK.MORTGAGE2
Dependent Variable	LCF
Distribution for Dependent Variable	General
Optimization Technique	Trust Region
Integration Method	None

Dimensions	
Observations Used	391027
Observations Not Used	54096
Total Observations	445123
Parameters	10

Parameters										
b0	b1	b2	b3	b4	b5	b6	b7	b8	sigma	NegLogLike
0	0	0	0	0	0	0	0	0	1	398223.08

Iteration History						
Iter		**Calls**	**NegLogLike**	**Diff**	**MaxGrad**	**Radius**
1	*	25	320021.734	78201.35	429730.1	-86639.8
2	*	39	212596.541	107425.2	565041.7	-105264
3	*	53	-164077.73	376674.3	1407875	-320653
4	*	66	-292377.8	128300.1	1798366	-127289
5	*	79	-322026.76	29648.96	1864987	-29710.7
6	*	93	-385656.24	63629.48	1823493	-65045.7
7	*	106	-428693.4	43037.16	1094601	-46829.1
8	*	118	-440415.32	11721.93	1307083	-21686.5
9	*	130	-450737.9	10322.58	222286.8	-9821.89
10	*	144	-517102.5	66364.6	41137636	-166708
11	*	156	-610534.79	93432.29	7654150	-119016
12	*	168	-639449.04	28914.25	2250836	-23935.8
13	*	180	-644220.31	4771.268	408225.4	-4241.34
14	*	192	-644436.76	216.4509	21574.11	-209.035
15	*	204	-644437.41	0.649581	68.42249	-0.64822
16	*	216	-644437.41	6.56E-6	0.005984	-6.54E-6

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	-129E4
AIC (smaller is better)	-129E4
AICC (smaller is better)	-129E4
BIC (smaller is better)	-129E4

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b0	2.5894	0.006644	39E4	389.75	<.0001	0.05	2.5764	2.6024	3.204E-6
b1	0.009307	0.000039	39E4	237.17	<.0001	0.05	0.009230	0.009384	0.005984
b2	4.5123	0.005912	39E4	763.26	<.0001	0.05	4.5007	4.5239	2.641E-7
b3	2.0491	0.004515	39E4	453.79	<.0001	0.05	2.0402	2.0579	4.905E-7
b4	-0.1053	0.000589	39E4	-178.95	<.0001	0.05	-0.1065	-0.1042	0.00006
b5	-0.00684	0.000560	39E4	-12.21	<.0001	0.05	-0.00794	-0.00574	5.051E-6
b6	0.2634	0.005277	39E4	49.91	<.0001	0.05	0.2530	0.2737	1.05E-8
b7	-0.00715	0.002047	39E4	-3.49	0.0005	0.05	-0.01116	-0.00314	6.546E-7
b8	0	0	39E4	.	.	0.05	.	.	0
sigma	0.04656	0.000053	39E4	884.33	<.0001	0.05	0.04646	0.04666	0.001102

From the result, we can tell in this non-linear regression, the AIC is -129E4, BIC is -129E4. And the estimate of b0 is 2.5894, b2 is 4.5123, b3 is 2.0491, b4 is -0.1053, b5 is -0.00684, b6 is 0.2634, b7 is -0.00715, b8 is 0 and sigma is 0.04656, b0,b2,b3 and b6 are positive, which means higher b0,b2,b3 and b6 or sigma go with higher LCF, while b1, b4, b5 and b7 are negative, which means higher b1, b4, b5 and b7 go with lower LCF.

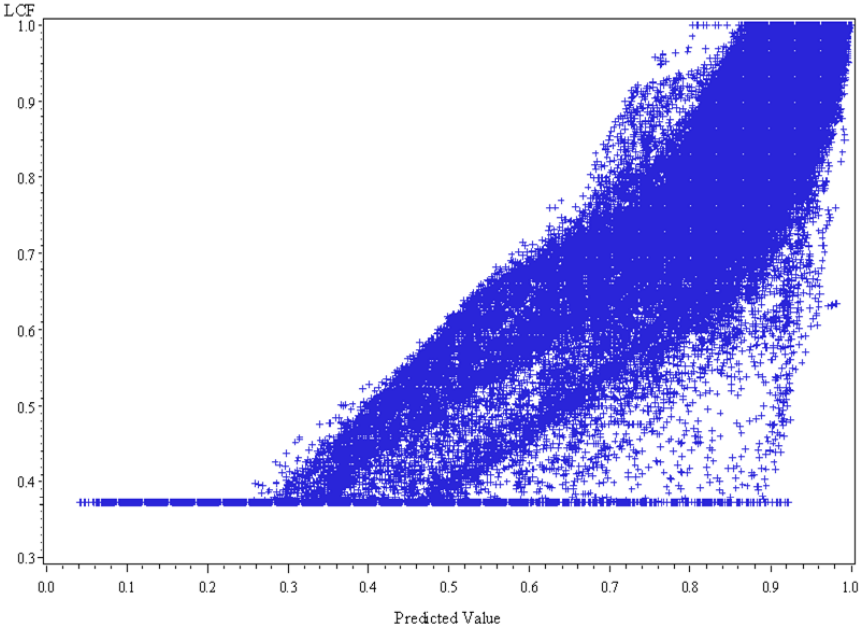
The REG Procedure
Model: MODEL1
Dependent Variable: LCF

Number of Observations Read	445123
Number of Observations Used	391027
Number of Observations with Missing Values	54096

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	4119.74728	4119.74728	1925270	<.0001
Error	391025	836.72654	0.00214		
Corrected Total	391026	4956.47382			

Root MSE	0.04626	**R-Square**	0.8312
Dependent Mean	0.93157	**Adj R-Sq**	0.8312
Coeff Var	4.96562		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	0.02207	0.00065964	33.46	<.0001
Pred	Predicted Value	1	0.98150	0.00070737	1387.54	<.0001



Variable	N	Mean	Std Dev	Minimum	Maximum
exposure	11673	253399.02	196298.34	9816.33	6500000.00
EAD_CCF	9787	260402.09	202784.37	14849.24	6500000.00

The R-squared of nonlinear regression is 0.8312, which is significantly higher than R-squared of linear regression. The estimated EAD with CCF has the mean of 260402.09 and standard deviation of 202784.37.

4.3. Controlling for Adverse Selection in PD Models

4.3.1. Interaction of PD and EAD

Low-risk borrowers are more likely to reduce the loan balance below the expected balance, while high-risk borrowers are more likely to increase the loan balance above the expected balance. In other words, exposures generally increase in the case of default. Borrowers typically have a prepayment option, which they generally exercise if they are low risk and have excess cash to repay the mortgage loan. A prepayment has a limited impact and can be controlled by a separate risk factor. However, after the complete mortgage has been paid off, low-risk borrowers leave the observed population, while high-risk mortgage borrowers remain. It is common in these situations to model the competing states default, payoff, and non-default by the following:

$$S_{it} = \begin{cases} 1 & \text{borrower i defaults at time t} \\ 2 & \text{borrower i pays loan off at time t} \\ 0 & \text{otherwise} \end{cases}$$

We consider a discrete-time hazard model such as the multinomial logit or probit model, or a continuous-time hazard model.

4.3.2. Discrete Time Hazard Model

4.3.2.1. Multinomial Logit Model

Description

Multinomial logit model is a representative of the class of discrete-time hazard models as follows,

$$\mathrm{P}(S_{it} = s \mid x_{it-1}) = \frac{exp(\beta_s' x_{it-1})}{1 + \sum_{s=1}^{2} exp(\beta_s' x_{it-1})}$$

We can estimate a multinomial logit model in SAS using PROC LOGISTIC.

Implementation

```
proc logistic data=data.mortgage;
    class status_time (ref='0');
    model status_time=FICO_orig_time LTV_time gdp_time interest_rate_time uer_time Retype_PU_orig_time/link=glogit rsquare;
    output out=probabilities predicted=p;
run;
```

Output

Model Information	
Data Set	DATA.MORTGAGE
Response Variable	status_time
Number of Response Levels	3
Model	generalized logit
Optimization Technique	Newton-Raphson

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Number of Observations Read	622489
Number of Observations Used	622219

Response Profile		
Ordered Value	**status_time**	**Total Frequency**
1	0	580484
2	1	15153
3	2	26582

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	360828.66	341904.05
SC	360851.35	342062.83
-2 Log L	360824.66	341876.05

R-Square	0.0300	**Max-rescaled R-Square**	0.0682

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
FICO_orig_time	2	835.8137	<.0001
LTV_time	2	3851.0248	<.0001
gdp_time	2	1074.6705	<.0001
interest_rate_time	2	1804.5392	<.0001
uer_time	2	1974.0504	<.0001
REtype_PU_orig_time	2	16.0727	0.0003

Analysis of Maximum Likelihood Estimates						
Parameter	**status_time**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**1**	1	-3.6624	0.1112	1083.9209	<.0001
Intercept	**2**	1	-1.9635	0.0907	468.5575	<.0001
FICO_orig_time	**1**	1	-0.00369	0.000128	825.5240	<.0001
FICO_orig_time	**2**	1	0.000190	0.000097	3.8390	0.0501
LTV_time	**1**	1	0.0198	0.000419	2247.0533	<.0001
LTV_time	**2**	1	-0.0117	0.000301	1512.4431	<.0001
gdp_time	**1**	1	-0.0949	0.00396	574.5102	<.0001
gdp_time	**2**	1	0.1069	0.00492	472.2119	<.0001
interest_rate_time	**1**	1	0.1447	0.00427	1146.4102	<.0001
interest_rate_time	**2**	1	0.0979	0.00338	839.1543	<.0001
uer_time	**1**	1	-0.0385	0.00521	54.6254	<.0001
uer_time	**2**	1	-0.2012	0.00457	1935.0847	<.0001
REtype_PU_orig_time	**1**	1	0.0858	0.0257	11.1260	0.0009
REtype_PU_orig_time	**2**	1	0.0466	0.0202	5.3457	0.0208

Odds Ratio Estimates				
Effect	**status_time**	**Point Estimate**	**95% Wald Confidence Limits**	
FICO_orig_time	**1**	0.996	0.996	0.997
FICO_orig_time	**2**	1.000	1.000	1.000
LTV_time	**1**	1.020	1.019	1.021
LTV_time	**2**	0.988	0.988	0.989
gdp_time	**1**	0.910	0.902	0.917
gdp_time	**2**	1.113	1.102	1.124
interest_rate_time	**1**	1.156	1.146	1.165
interest_rate_time	**2**	1.103	1.096	1.110
uer_time	**1**	0.962	0.952	0.972
uer_time	**2**	0.818	0.810	0.825
REtype_PU_orig_time	**1**	1.090	1.036	1.146
REtype_PU_orig_time	**2**	1.048	1.007	1.090

From the results above, we know that we obtain two parameter estimates for every covariate. All the variables have positive effects for both probabilities of default and payoff, which mean that, for example, increasing the value for the uer_time will increase both the probabilities of default and payoff.

4.3.2.2. Estimate the probabilities of default

Description

The probabilities of default can be estimated with the OUTPUT statement, which has evaluated the model equation and estimated parameters as follows:

$$\hat{P}(S_{it} = 1 \mid x_{it-1}) = \frac{exp(\hat{\beta}_1' x_{it-1})}{1 + exp(\hat{\beta}_1' x_{it-1}) + exp(\hat{\beta}_2' x_{it-1})}$$

With

$$\hat{\beta}_1' x_{it-1} = \hat{\beta}_{0,1} + \hat{\beta}_{1,1} * FICO\_orig\_time + \hat{\beta}_{2,1} * LTV\_time + \hat{\beta}_{3,1} * gdp\_time + \hat{\beta}_{4,1} * interest\_rate\_time + \hat{\beta}_{5,1} * uer\_time + \hat{\beta}_{6,1} * REtype\_PU\_orig\_time$$

$$\hat{\beta}_2' x_{it-1} = \hat{\beta}_{0,2} + \hat{\beta}_{1,2} * FICO\_orig\_time + \hat{\beta}_{2,2} * LTV\_time + \hat{\beta}_{3,2} * gdp\_time + \hat{\beta}_{4,2} * interest\_rate\_time + \hat{\beta}_{5,2} * uer\_time + \hat{\beta}_{6,2} * REtype\_PU\_orig\_time$$

With $\hat{\beta}_{0,1}$ to $\hat{\beta}_{6,2}$ being the estimated default parameters.
We use PROC MEANS to calculate the mean for the default indicators and the estimated PDs.

Implementation

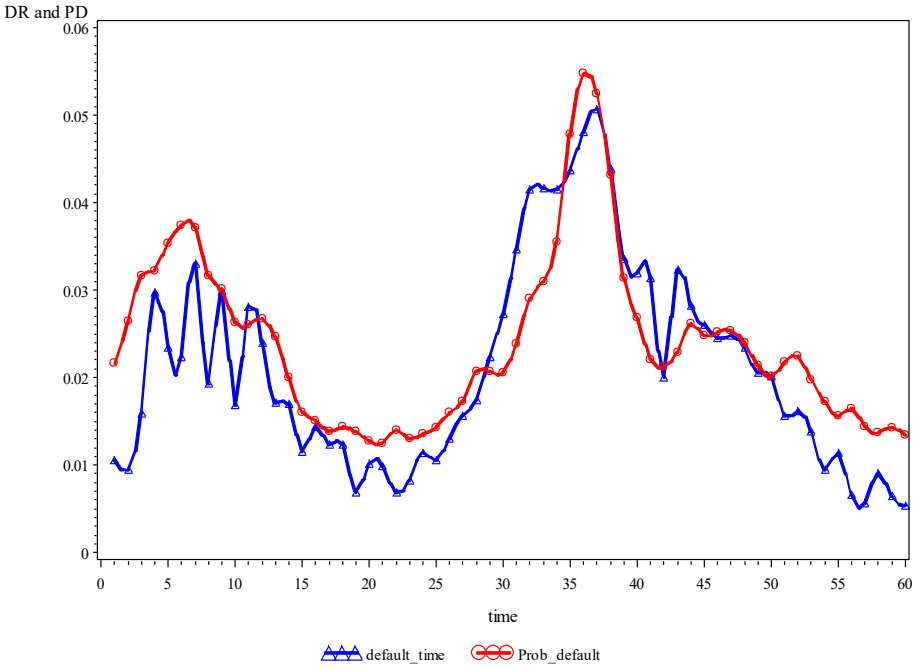
```
/**** Estimation of Default Probabilities *****/



/* calculate the mean for the default indicators */
proc means data=probabilities (where=(_level_=1)) mean nolabels;
var default_time p;
run;

/* compute default rates and avg estimated default probabilities by time */

proc sort data=probabilities;
by time;
run;

proc means data=probabilities (where=(_level_=1));
by time;
output out=means mean(default_time p)=default_time p;
run;

data means;
set means;
label p="Prob_default";
run;

ods graphics on;
axis1 order=(0 to 60 by 5) label=('time');
axis2 order=(0 to 0.06 by 0.01) label=('DR and PD');
symbol1 interpol=spline width=2 value=triangle c=blue;
symbol2 interpol=spline width=2 value=circle c=red;
legend1 label=none shape=symbol(4,2) position=(bottom outside);

proc gplot data=means;
plot (default_time p)*time/overlay haxis=axis1 vaxis=axis2 legend=legend1;
run;
ods graphics off;
```

Output

Calibration of Multinomial Logit Models: Comparison of Default Indicators and Estimated Default Probabilities. The calibration is clear, as the mean of the default event almost matches the mean of the estimated PDs.

Variable	Mean
default_time	0.0243506
p	0.0243532

Real-Fit Diagram for the Default Probabilities. In the figure below, we compare the observed default rate (DR) and the average of the estimated default probabilities (PD)



4.3.2.3. Estimate the probabilities of payoff

Description

Probabilities of payoff can also be estimated with the OUTPUT statement which has evaluated the model equation and estimated parameters as follows:

$$\hat{P}(S_{it} = 2 \mid x_{it-1}) = \frac{exp(\hat{\beta}_2' x_{it-1})}{1 + exp(\hat{\beta}_1' x_{it-1}) + exp(\hat{\beta}_2' x_{it-1})}$$

With

$$\hat{\beta}_1' x_{it-1} = \hat{\beta}_{0,1} + \hat{\beta}_{1,1} * FICO\_orig\_time + \hat{\beta}_{2,1} * LTV\_time + \hat{\beta}_{3,1} * gdp\_time + \hat{\beta}_{4,1} * interest\_rate\_time + \hat{\beta}_{5,1} * uer\_time + \hat{\beta}_{6,1} * REtype\_PU\_orig\_time$$

$$\hat{\beta}_2' x_{it-1} = \hat{\beta}_{0,2} + \hat{\beta}_{1,2} * FICO\_orig\_time + \hat{\beta}_{2,2} * LTV\_time + \hat{\beta}_{3,2} * gdp\_time + \hat{\beta}_{4,2} * interest\_rate\_time + \hat{\beta}_{5,2} * uer\_time + \hat{\beta}_{6,2} * REtype\_PU\_orig\_time$$

With $\hat{\beta}_{0,1}$ to $\hat{\beta}_{6,2}$ being the estimated default parameters.

We use PROC MEANS to calculate the mean for the payoff indicators and the estimated payoff probabilities.

Implementation

```
/**** Estimation of Pay-off Probabilities *****/

/* calculate the mean for the pay-off indicators */
proc means data=probabilities (where=(_level_=2)) mean nolabels;
var payoff_time p;
run;
```

```
/* compute payoff rates and avg estimated payoff probabilities by time */

proc sort data=probabilities;
by time;
run;
```

```
proc means data=probabilities (where=(_level_=2));
by time;
output out=means mean(payoff_time p)=payoff_time p;
run;
```

```
data means;
set means;
label p="Prob_payoff";
run;
```

```
ods graphics on;
axis1 order=(0 to 60 by 5) label=('time');
axis2 order=(0 to 0.14 by 0.02) label=('PR and PP');
symbol1 interpol=spline width=2 value=triangle c=blue;
symbol2 interpol=spline width=2 value=circle c=red;
legend1 label=none shape=symbol(4,2) position=(bottom outside);
```

```
proc gplot data=means;
plot (payoff_time p)*time/overlay haxis=axis1 vaxis=axis2 legend=legend1;
run;
ods graphics off;
ods rtf close;
```

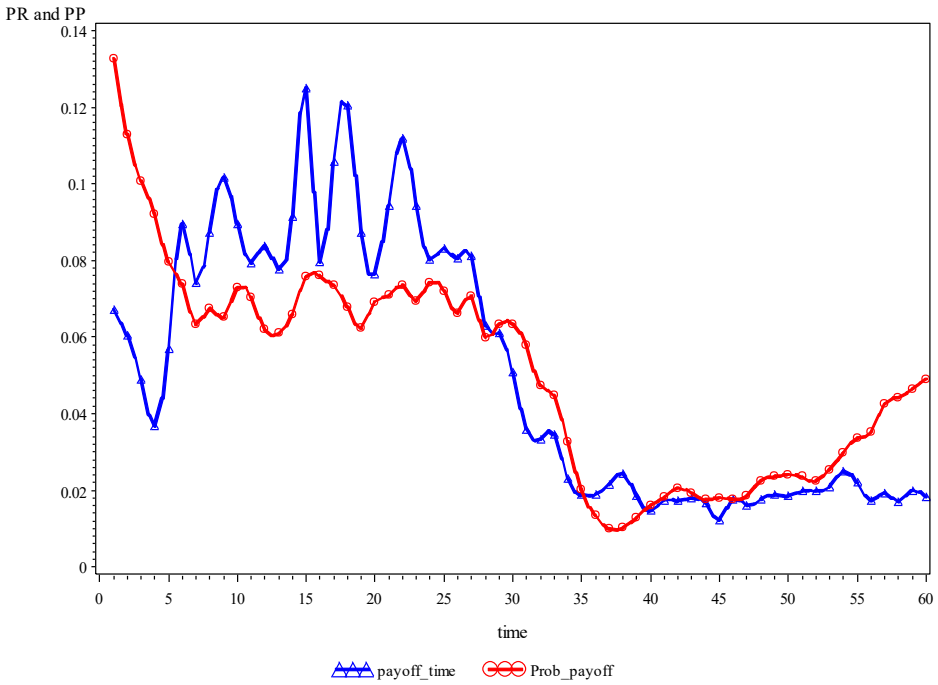
Output

Calibration of Multinomial Logit Models: Comparison of Payoff Indicators and Estimated Payoff Probabilities. The calibration is clear, as the mean of the payoff event almost matches the mean of the estimated payoff probabilities.

Variable	Mean
payoff_time	0.0427140
p	0.0427213

Real-Fit Diagram for the Payoff Probabilities. In the figure below, we compare the observed payoff

rate (PR) and the average of the estimated payoff probabilities (PP).



4.4. Conclusion

In this section, we fitted three EAD models in total, including linear regression, transformed linear regression and non-linear regression. The R-square of these models are shown in the following table.

model	Linear regression	Transformed linear regression	Non-linear regression
R-square of LCF	0.3819	0.2318	0.8312

R-square of nonlinear regression (0.8312) is significantly higher than R-square of both transformed linear regression (0.2318) and linear regression (0.3819), which indicates that nonlinear regression is the best fit according to our model results.

5. Expectation

During the whole project, we noticed that the estimations of PD, EAD, and LGD might have biases based on dataset has different needs before modeling. To improve this project and our models, we need to balance the possible margins and adjust estimates. For the sample of reference, they should have a longer period of time of information and should include real life economic affects with a relatively high default rate. For the dataset, all the data have different background of the locations, industries, and resource, those reasons should be considered into our model. For example, if one individual has international business from China, and all the investment money were from China, this individual should be considered as higher risk than people who in the same situation but with good FICO score. Because of all the risks that individual might face like currency change or legal policy difference.

By using all different types of estimation techniques by statistics methodology for three factors: PD, EAD and LGD we cannot have the perfect model for this dataset due to limit time and recourses. However, we have a good demonstration of some important methods. The purpose of PD, EAD, and LGD are to predict the possibility, exposure and loss of customer default event in the future time. As for now, we have an output for the question, but with the real-life decision we need to constantly update our determinants. Because the most well-known method of estimating PD, EAD and LGD is to calculate them according to the historical data. But it is still not perfect enough. We should explore model that leads to higher efficiency

In conclusion, through this project we were able to use various factors that impose significant impacts on forecasting future target values. We also learned that we need to keep balance between flexibility and stability during the model fitting. During the project, we were able to apply the textbook knowledge to the real-life scenario. We also had a deep understanding about banking risk management.