Current Expected Credit Loss (CECL)
Project Report

Tianyu Cheng

Tharun Polamarasetty

Willies Mboko

Xu Yan

Contents

1. Objective

To test the effectiveness of Current Expected Credit Loss (CECL) framework and how the housing market of single family data provided by Fannie Mae would perform for the most volatile state California.

2. Introduction

Before we start about CECL, we should know about two terms: FASB and ALLL.

The Financial Accounting Standards Board (FASB) is a private, non-profit organization whose primary purpose is to provide a framework for establishing and improving generally Accepted Accounting Principles within the United States.

Before CECL, there used to be another framework called Allowance for Loan and Lease Losses (ALLL) which is backward looking that depends on incurred losses but not realized loss which means, unless it's know with certainty that future cashflow will not be collected it would not account for it. This resulted in a disaster during financial crisis as the economy's negative outlook was not considered for loss calculations which made it very difficult for adjusting reserves.

Therefore, FASB came up with newer framework called Current Expected Credit Losses (CECL) that focuses on estimation of expected losses over the life of the loans which is forward looking. There are several suggested methodologies to test this framework, few of which are: Vintage, PD x LGD method, Loss Rate, Discount Cash Flow, Roll rate. In order to test this new framework, we used Fannie Mae, single family data from 2007 to 2015 for vintage analysis to estimate future loss rate however, it doesn't seem like a good model.

So, we choose to test the dataset with PD x LGD model as the data set is imbalanced. We choose to test for the state of California as it's highly volatile from which we could access how the models performed during crisis and aftermath.

3. Probability of default model

Probability of Default (PD) model is used to estimate the probability of a borrower choose to default at a certain time. Usually in actual practice, a bank usually only care about if a borrower will default in this month because the bank will reevaluate the probability again next month. However, since CECL is accounting standard and ask to evaluate the potential loss for the whole life period, we not only need to calculate the loss at present but also in the future. This makes us to build a model that can be used for every period. Firstly, Vintage analysis and then transition matrix.

3.1 Vintage Analysis

Vintage Analysis is a quick estimating method using in banks. Usually Vintage Analysis use a baseline variable like some macroeconomics factors named Q factors. And an assumption for Vintage Analysis is that the loss rate will move the same direction and same length with the Q factor. So if we get the current vintage loss rate matrix and both current Q factor value as well as its forecast value, we can roughly estimate the future loss rate.

We separately build two loss rate estimate matrixes by vintage. One is based on 2005-2010 Fannie Mae dataset and the other one is based on 2011-2015 dataset. And we will take the former one to do analyze. From the difference matrix, we can see that if we use unemployment rate to predict, most of the predictions are over-estimated expect 3 cases. In the contract, it is just half to half if we use CPI to estimate. As for absolute error, CPI shows a smaller error than unemployment rate. From these results, we think that for 2005-2010, CPI is a better indicator than unemployment rate. The detailed results are in appendix 1.

3.2 Transition Matrix

So as for PD model, since the vintage is having high error rate, our group choose to use transaction matrix to estimate the probability to every status. By using this method, we get the potential default possibility of a borrower if we know his current delinquency status. And using transaction matrix method we can divide the borrowers into several groups and can better estimate the potential loss.

Transaction matrix is based on conditional Markov chain model. Conditional Markov chain model is a probability model that is used to predict the probability of an event with certain condition. However, there are several assumptions in Transaction Matrix model.

1) The probability of the transfer in this period only depend on last period. That means only the most recently information will influence the behavior of a borrower.

$$P(CLDS_{i+1} = C|CLDS_i CLDS_{i-1}, CLDS_{i-2} \dots) = P(CLDS_{i+1} = C|CLDS_i)\ for\ i = 1,2, \dots$$

2) The borrower will only move no more than one step. Since our data is monthly and CLDS is also measured by one month, so it is not realistic for a borrower to jump from 1 to 3 in month.

$$CLDS_i \in \{CLDS_{i-1} - 1, CLDS_{i-1}, CLDS_{i-1} + 1\}$$

With those assumptions, we build transaction matrix model based on all macroeconomics factors and a generated variable CLTV. Since CLTV is build on HPI, so in our modeling HPI will not be added. In our dataset, we have three labels named -1, 0, 1. -1 means prepaid in group 0 and go away from default in other groups. 0 means the borrower choose to remain in the same status. And 1 means the borrower choose to increase their CLDS level. And if a person's CLDS is bigger than 3, we regard him as a default borrower.

Besides, as for the dataset, since group 0 and group 3 is unbalanced due to there are too much 0 to 0 and 3 to default in the dataset, the indicators of the model for those groups may be useless. And we will not emphasize the indicators for those groups.

When it comes to predict the probability of default for one next delinquency under the condition of a particular status, we use logistic regression method to estimate. Logistic regression is a statistic model used to predict a label of an independent variable. In our model, we will use those factors to predict the next delinquency. And since most macroeconomics factors have strong multicollinearity, in order to make our model effective, we use ridge regression to estimate the coefficients of the model.

As mentioned above, as we want to estimate the probability of default for all time, we will mix the dataset and will not regress based on a particular year. The baseline of the model is the mean of every variables and we will use the mean to calculate the matrix. This choice is a common use in the industry. After calculating the coefficients, we have the transaction matrix as below.

Table 1 Transaction Matrix Result

	-1	0	1	2	3	D
0	0.016308	0.980186	0.003506			
1		0.361465	0.450551	0.187984		
2			0.132569	0.416776	0.450655	
3				0.063522	0.247910	0.688568

From this matrix, we can clearly find that most of the loans will remain the ‘0’ delinquency status, while if a borrower has already reached to a delinquency higher than 1, there is great possibility of him to go to default. And this result is similar with some of the other researches.

We also picked two variables’ coefficients to make an analysis. The first variable is business climate from Bloomberg and its coefficient is in the appendix. Business climate is an indicator that is used to measure whether the economy is booming or not. If business climate index increase, which means people feels positive to the future, they will go to decrease their delinquency status next month or even prepaid. And this matches our coefficient.

Another variable is CLTV, which measures the ratio of remaining loan value to current collateral value. If this indicator increases, since for a particular month his remaining loan value is the same, this means his collateral’s value decreases and will make the borrower feel not worth to pay his loan. So he or she will probability have a higher possibility to go to default. And the coefficient of this variable also matches our analysis.

4. Loss Given Default Model

Loss Given Default (LGD) is an important element in modeling credit risk in the PD-LGD model. For our current data we estimate LGD as:

$$\mathrm{LGD} = \frac{\text{Loan Balance add Expenses less proceeds}}{\text{Loan Balance}}$$

We also set a restriction that LGD will lie between 0 and 1.

Mathematical representation of LGD model:

$$let\ Y = LGD$$

Then

$$\mu_i = E[Y_i \mid x_i] = \beta_0 + \boldsymbol{\beta_i x_i}$$

where $\boldsymbol{x_i}$ are the explanatory variables i.e GDP, CLTV and FICO and $\boldsymbol{\beta_i}$ is a vector of coefficients.

$$\eta = \beta_0 + \boldsymbol{\beta_i x_i}$$

We now express η in the form a logit link function, $log\frac{\mu_i}{1-\mu_i}$ in order to ensure that μ_i lies between 0 and 1.

Notice how

$$log\frac{\mu_i}{1-\mu_i} = \eta = \beta_0 + \boldsymbol{\beta_i x_i}$$

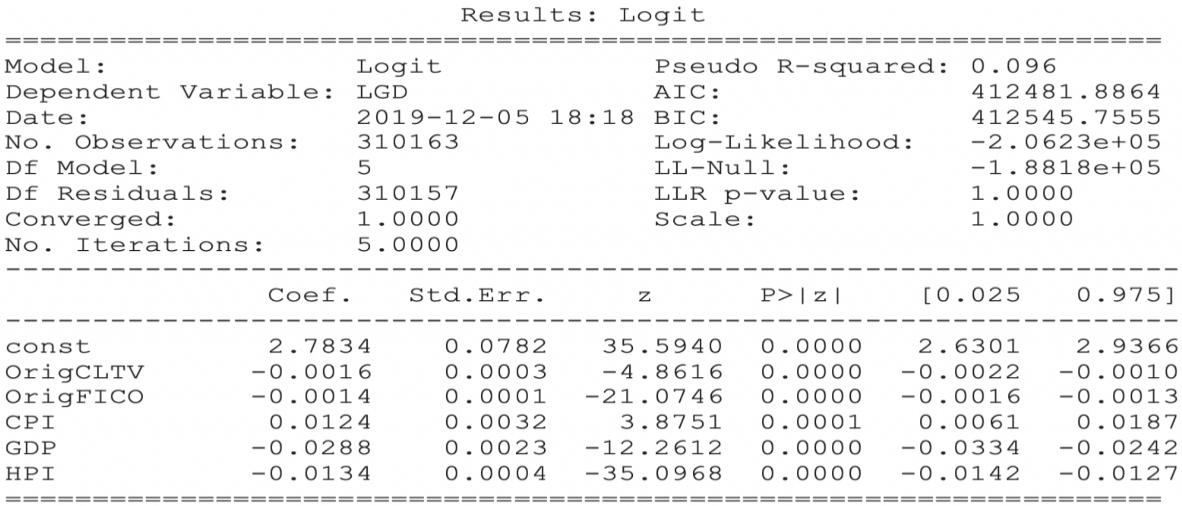
$$\Rightarrow \mu_i = \frac{1}{1+e^{\eta}}$$

Notice also how we have not described the distribution of Y yet.

We can think of any distributions that take on values between 0 and 1 and fit it in the distribution. Fits that come to mind are

- Bernoulli, Binomial
- Beta distribution

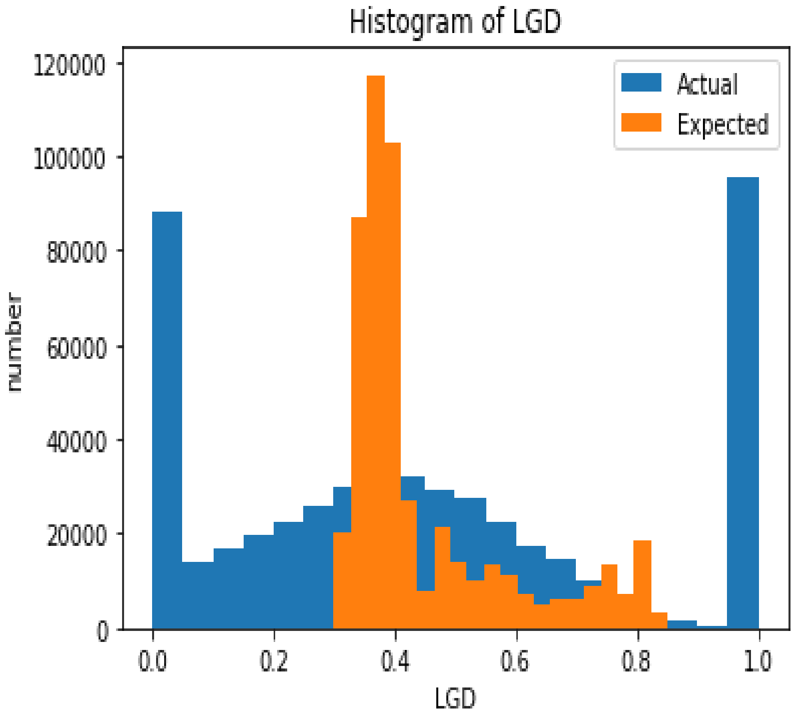
We fit a Binomial regression and the result was a below:



```
                        Results: Logit
==================================================================
Model:               Logit             Pseudo R-squared: 0.096
Dependent Variable:  LGD               AIC:              412481.8864
Date:                2019-12-05 18:18  BIC:              412545.7555
No. Observations:    310163            Log-Likelihood:   -2.0623e+05
Df Model:            5                 LL-Null:          -1.8818e+05
Df Residuals:        310157            LLR p-value:      1.0000
Converged:           1.0000            Scale:            1.0000
No. Iterations:      5.0000
-------------------------------------------------------------------
                Coef.   Std.Err.     z      P>|z|    [0.025   0.975]
-------------------------------------------------------------------
const          2.7834    0.0782   35.5940  0.0000   2.6301   2.9366
OrigCLTV      -0.0016    0.0003   -4.8616  0.0000  -0.0022  -0.0010
OrigFICO      -0.0014    0.0001  -21.0746  0.0000  -0.0016  -0.0013
CPI            0.0124    0.0032    3.8751  0.0001   0.0061   0.0187
GDP           -0.0288    0.0023  -12.2612  0.0000  -0.0334  -0.0242
HPI           -0.0134    0.0004  -35.0968  0.0000  -0.0142  -0.0127
==================================================================
```

The fit was had a low R-squared implying that the fit only explained about 10% of the overall variation.

A diagnostic plot of Actual v/s Expected LGD highlights the problem. The distribution of LGD is bimodal and fitting any distribution would need to account for it for us to get a better fit.



5. Potential Improvements

Although we go through PD-LGD Model and derive our own results of project on Current Expected Credit Losses, combining with process and regression results we have, there's still some potential improvement we could modify to fulfill our project.

Firstly, there is still unbalance exists when we arrange the datasets of PD model, especially between group 0 and group 3. And it is because of it exists, some indicators such as Recall, Precision, etc., cannot have significant value when we use to correct and prove our model. Thus, if we could oversample those datasets and make them balance in transaction matrix, our model can have better confidence level to show the results.

Secondly, as we could tell from our coefficient forms and also the assumptions of both models, the delinquency status can only be 0 and 1, but considering the fact that someone may default more than two, or even consecutive times, multi-value of delinquency status are needed to be set. But we all know after this setting, the simple model would become complex and more difficult to calculate and estimate. So, if time allows, a developed model will derive under multi-value delinquency status allowed.

Group segmentation is also an improvement being considered when processing data. During our model validation part, we go through the csv document and find out one of the best ways to sole significant problem is to split the data into two parts, say above 0.5 and below 0.5. Such group segmentation is easy to operate on Python, but the results will have better look after such group segmentation.

Least but not last, under real economic and financial world, we think beta distribution is a better option compared to a binomial distribution when doing regressions and model construction.

Beta distribution is useful for ratios while binomial distribution is the initial assumptions for the basic model.

6. References

[1] https://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html.

[2] Transition matrix models of consumer credit ratings by Madhur Malik and Lyn C. Thomas, Quantitative Financial Risk Management Centre, School of Management, University of Southampton, Southampton SO17 1BJ, UK

[3] Mortgage Transition Model Based on LoanPerformance Data by Shuyao Yang, Washington University in St. Louis

[4] https://www.sageworks.com/cecl-transition-content-license/

7. Appendix

7.1 Vintage result and difference

Table 2 Prediction using unemployment rate

Loss Rate (Estimate)	**Y1**	**Y2**	**Y3**	**Y4**	**Y5**	**Y6**
2005	0.1029%	0.2731%	0.5207%	1.2297%	2.0748%	1.7258%
2006	0.0861%	0.6558%	1.9857%	3.2418%	2.4826%	1.6243%
2007	0.1579%	2.0647%	4.0622%	3.0673%	2.2203%	1.4605%
2008	0.5167%	2.3029%	2.0319%	2.8294%	1.9963%	1.3429%
2009	0.0653%	0.2392%	2.7169%	2.5439%	1.8356%	1.1274%
2010	0.0150%	0.3318%	2.4428%	2.3392%	1.5410%	0.9529%
Average	0.1573%	1.1735%	2.2934%	2.5419%	2.0251%	1.3723%
Loss/Q factor	2.44%	16.51%	29.79%	31.02%	24.35%	17.81%

Table 3 Prediction using CPI

Loss Rate (Estimate)	**Y1**	**Y2**	**Y3**	**Y4**	**Y5**	**Y6**
2005	0.1029%	0.2731%	0.5207%	1.2297%	2.0748%	1.7258%
2006	0.0861%	0.6558%	1.9857%	3.2418%	2.4826%	0.5593%
2007	0.1579%	2.0647%	4.0622%	3.0673%	0.7644%	0.3687%
2008	0.5167%	2.3029%	2.0319%	0.9741%	0.5039%	0.2618%
2009	0.0653%	0.2392%	0.9354%	0.6422%	0.3579%	0.2885%
2010	0.0150%	0.3707%	0.6167%	0.4560%	0.3944%	0.0214%
Average	0.1573%	1.1813%	2.3704%	2.8958%	5.6849%	1.7644%
Loss/Q factor	6.47%	49.43%	107.91%	147.24%	356.05%	105.23%

Table 4 Difference between prediction using unemployment rate and actual values

Difference	Y1	Y2	Y3	Y4	Y5	Y6
2005	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%
2006	0.000%	0.000%	0.000%	0.000%	0.000%	-0.226%
2007	0.000%	0.000%	0.000%	0.000%	-0.105%	-0.084%
2008	0.000%	0.000%	0.000%	1.291%	1.005%	0.795%
2009	0.000%	0.000%	2.412%	2.271%	1.642%	0.973%
2010	0.000%	0.232%	2.301%	2.217%	1.427%	0.853%
Average	0.000%	0.039%	0.786%	0.963%	0.662%	0.385%

Table 5 Difference between prediction using CPI and actual values

Difference	Y1	Y2	Y3	Y4	Y5	Y6
2005	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%
2006	0.000%	0.000%	0.000%	0.000%	0.000%	-1.291%
2007	0.000%	0.000%	0.000%	0.000%	-1.561%	-1.176%
2008	0.000%	0.000%	0.000%	-0.564%	-0.487%	-0.286%
2009	0.000%	0.000%	0.631%	0.370%	0.165%	0.134%
2010	0.000%	0.271%	0.475%	0.333%	0.280%	-0.078%
Average	0.000%	0.045%	0.184%	0.023%	-0.267%	-0.450%

7.2 Coefficients in PD part

Table 5 Coefficients for Business Climate

	-1	0	1
0	0.0558	0.0180	0.0739
1	0.2734	0.1263	0.2607
2	0.0583	0.0187	0.0395
3	0.0131	0.0107	0.0242

Table 6 Coefficients for CLTV

	-1	0	1
0	-0.1552	-0.1524	0.3076
1	-0.3504	-0.09469	0.3598
2	-0.1620	-0.0278	0.1898
3	-0.1360	-0.0140	0.1500