**Final report of SAS project**

# Abstract

We use data from FreddieMac to analysis the Expected Loss based on mortgage shown as liquidated from the year 2000 Sample datasets with 50,000 mortgage loans. We use the Expected Loss (EL) model to calculate the Expected loss on each loan. The Exposure At Default (EAD) and Loss Given Default (LGD) in this model are calculated by the formula we have known, the Probability of Default (PD) is calculated by using the result of logistic regression. In this project, we select appropriate variables, clean the data, run regressions and analyze the results.

# Introduction

In this project, we need to use data from FreddieMac to create an Expected Loss (EL) model based on mortgage shown as liquidated from the year 2000 Sample datasets with 50,000 mortgage loans. The Expected Loss is calculated as a product of Exposure At Default (EAD), Loss Given Default (LGD) and Probability of Default (PD):

The values of the first two variables (LGD and EAD) are easy to obtain and have been calculated in past assignments. We know that the Loss Given Default is calculated by

and the Exposure At Default is calculated by

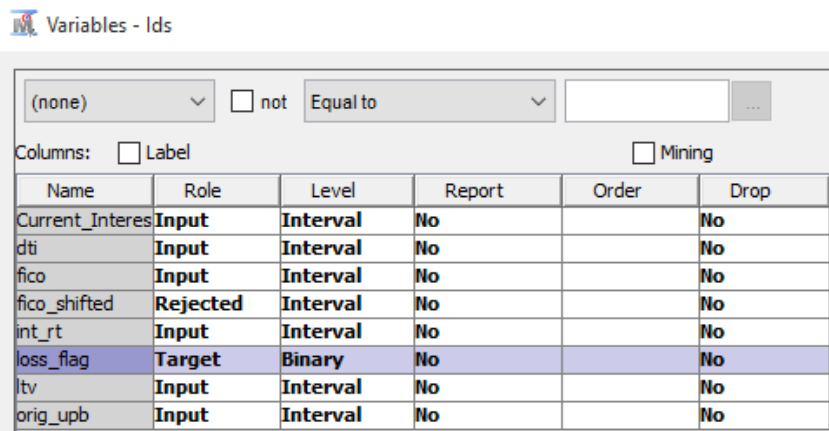
The Probability of Default cannot be calculated easily.

In this project, we are using logistic regression approach to calculating the Probability of Default.

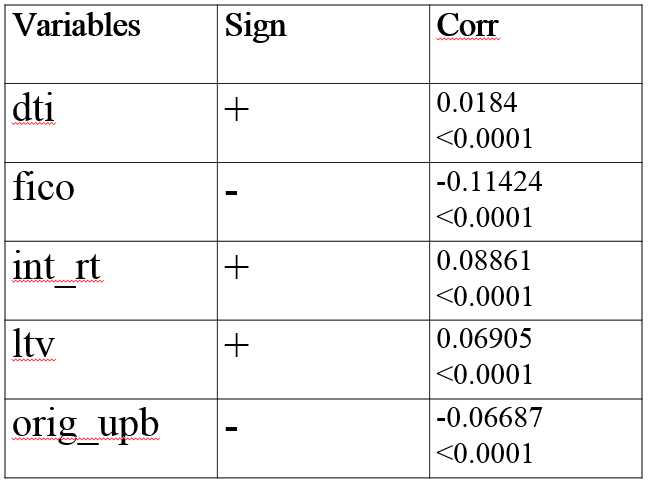
This base report is organized as follows: Section 2 gives an explanation and statement of the variables, section 3 describes and explains the result of logistic regression for Probability of Default, Section 4 reports and discusses the total results obtained of EL model and finally Section 5 concludes the report.

# Inputs and methodology

**Variables**

To get an appropriate PD model, first we need to select efficient variables as inputs to get the model. The variable selection satisfies three principles. The first is the effect on dependent variable can be explained logically. So we will give logical reason on why the variables can be efficient. The second is the variable should have significant correlation with dependent variable with correct sign, we will do correlation test with proc corr step in SAS. And the third is the variables can be proved to be significant in regression. If we get the regression result and some variables are not significant, then we need another regression with different set of variables. 

The variables we chose are original debt-to-income ratio (dti), credit score (fico), original interest rate (int\_rt), original loan-to-value (ltv), and original unpaid balance (orig\_upb). Their effect on dependent variable can be explained logically. A higher dti means it’s harder to pay debt with income, so it should have positive effect on PD. A high fico means better credit status, which means a lower PD. A higher interest rate means higher cost of paying debt which can reduce the motivation of paying debt. A high ltv can make people think it’s worthless to pay the debt, so it also has a positive effect on PD. For orig\_upb, although it means a larger amount of unpaid debt, it can also means a better credit status in the past because only those who always pay the debt on time can be given larger amount of debt to make the unpaid balance higher. For those who have worse credit status, their amount of debt will be limited to smaller amount, so they cannot accumulate a high unpaid balance. So orig\_upb has negative effect on PD.



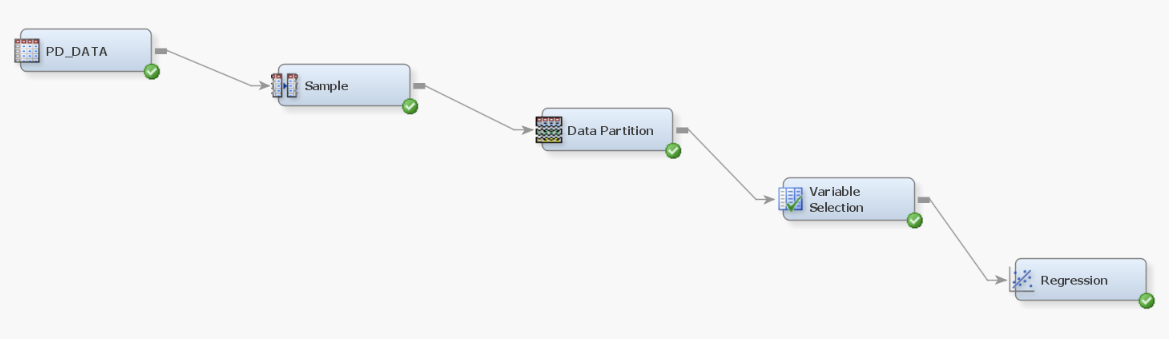
From this table we can find that our result of correlation test is the same as our expectation. So we use them to develop our model.

**Methodology**

We need to get PD which is a probability. So we choose logistic regression to develop PD model and set dependent variable as binary variable. In this way when we get the parameters of the model, we can get predicted values of PD between 0 and 1 which represents a probability of default.

Since our model is used for future predictions, we should use cross validation in regression process to test whether it has a good predicting ability. So we divided the original data set into a training set and a validation set. Here, we set training set as 60% of original set and validation set as 40%.

The instrument we used to get regression result is Enterprise Miner of SAS. And we use SAS to clean the original data.



# Summary of outputs and conclusion

1. **Evaluation of model’s logic**

The estimated logit model for predicting probability of default is expressed in this form:

Here the asterisks represent significance of coefficient, \* for significance at 90% confidence level, \*\* for 95% and \*\*\* for 99% accordingly.

Now it is necessary to evaluate the signs of coefficients and how they fall in line with our theoretical expectations.

* For *dti* a positive coefficient was expected since higher debt-to-income ratio leads to a higher probability of default on mortgage, i.e. a positive change of dti’s values must increase the probability of default. This coefficient is significant at 90% level and has right sign.
* For *fico* a negative coefficient was expected due to negative correlation between credit score and probability of default. In other words, the higher is the credit score, the smaller is chance that this borrower would experience a default, hence a negative coefficient makes sense. The resulting coefficient is highly significant and negative, this agrees with our assumption.
* The interest rate is positively correlated with probability of default since the higher interest rate will cause borrower to pay higher monthly payments which create a lot of additional pressure on his solvency. Hence, a positive sign is expected.
* The loan-to-value ratio is expected to be positively associated with probability of default because the higher is the ratio of loan to house’s value, the higher amount of payments must be received from borrower. The sign of *ltv* in model is positive, hence it is correct.
* The unpaid balance amount is expected to be positively related to probability of default since larger unpaid balance means larger payments and longer duration of mortgage. However, the obtained coefficient is negative. This result can be due to several problems with dataset. First, data should have been thoroughly investigated for presence of outliers, i.e. observations with extremely unusual values such as very high unpaid balance or very high interest rate in comparison with sample average, etc. The second reason may be very low number of defaults in database which includes only 751 observed defaults amongst 50,000 observations or 1.4% of whole database. Hence, the logit model has only 1.4% observations to train to recognize harbingers of default, while 98.6% of loans didn’t default so model might be biased due to lack of sufficient training observations.

The overall conclusion on estimation part is that almost all variables have theoretically expected signs and are significant at least at 90% confidence level. The problem with negative sign of unpaid balance need further investigation.

1. **Evaluation of model’s performance**

Since model is, in general, correct, it is necessary to evaluate its forecasting power. The SAS Enterprise Miner provides several key characteristics which allow us to judge the predicting ability of model.

We will use indicator known as hit-rate which can be constructed on the basis of classification table outputted by SAS. The hit-rate can be defined as percentage of observations that were correctly forecasted by logit model.

In other words, we find total number of observations for which outcome of unity was predicted, and then find its percentage in actual number of observations with ones. Hence, the higher is hit-rate for predicted ones, the better model forecasts the probability that event of interest will happen. For instance, hit-rate for ones in mortgage model is 77.17%, which means that we have 77.17% confidence that a loan for which we predicted default will actually experience default.

The hit-rate for zeroes is calculated in a similar fashion, we find what share of actual observations with zeroes was correctly assigned a value of zero by model. Therefore, the higher is hit-rate for zeroes, the more confident we are when predicting that event of interest will not happen to observation being considered. This hit-rate for logit model is 72.92%, hence we can be 72.92 percent confident when predicting that this borrower will not default.

If logit model is good, both hit-rates must be relatively high. For our model the consolidated table of hit-rates is:

|  |  |  |  |
| --- | --- | --- | --- |
| **Actual**  **outcome** | **Predicted outcome** | | **Total**  **Percentage** |
| 1 | 0 |
| 1 | 77.17 | 22.83 | 100% |
| 0 | 27.08 | 72.92 | 100% |
| **Total**  **percentage** | 100% | 100% |  |

The hit-rate for ones is in intersection of column “1” (predicted outcome) and row “1” (actual outcome). The hit-rate for zeroes is in intersection of column “0” and row “0”.

Each hit-rate is accompanied by miss-rate which shows the percentage of cases when we predicted zero or one, but an inverse outcome was observed. Our model has high hit-rates for both situations, i.e. when predicting default of a loan, and when predicting that loan will not default.