INDIVIDUAL ASSIGNMENT

QUESTION 1: Data-description

(A) Interpret findings (based on the chart outputs of TVA)

- **LTV_time by Time**

The loan to value ratio over time is low during the time 0 to 30. After the Global Financial Crisis in 2008 (period 30), the loan to value ratio increases significantly and is higher than the previous periods as the rates are above 90%. In the period 35, the loan to value ratio is highest which is above 100% is very high risk. A higher level of lending is because the money is being lent to someone who has less capital, therefore, the lenders normally requires higher interest rate for these lendings as they are very high risk. During GFC, the economy is in recession period, and people lost their jobs, homes and wealth, so the LTV ratio is very high since the GFC starts in period 30.

- **LTV_time Vintage**

Vintage is defined as the orgination time. The loan to value ratio of the loans originated before GFC (period 30) increases drammatically which the ratio reaches 100%, and it decreases after the GFC. There is a huge plummet in period 37, and the loan to value ratio is down to below 30% after the economy recession period. It indicates the GFC might have been improved during this period, and the bank is doing better in making decision. With the low loan to value ratio, there is higher chance to get loan approved.

- **LTV_time Age**

Loan age is to reflect the time-varying contractual features such as loan amortization schedule. With age, the loan to value rate decreases, and the shape is right skewed from period 0 to 50. The loan to value is high at the age of 20, and the rate decreases after that age.

- **Cep_time by Time**

Cumulative excess payments indicate the liquidity ability of the borrower. The CEP over time is calculated by the difference between balance scheduled time and observed balance over the property price. The property value at origination is floored at $100,000. Before period 30 which is the GFC period, the CEP increases and is high at 0.005. After the GFC, the CEP decreases and is down to -0.030 in the period 50. It indicates that borrowers have higher liquidity ability before GFC.

- Note: There is missing values (using isnull function to find out there is missing values from annuity scheduled payments) and some values of the data are infinitive that creates crack when computing the shape of cep_time. The function of (pd.options.mode.use_inf_as_na = True) is to define all the infinitive value are true, and the function of interpolate is to estimate the missing values.

- **Cep_time by Vintage**

CEP for loan originated has a hump shape, and it has an increasing trend. After period 30 (GFC), the CEP has a plummet , and then it startes to increase from period 35 to high rate of 0.06 in the period 40. It indicates the recovery after the Global Financial Crisis which liquidity of borrowers increase.

- **Cep_time by Age**

The loan age is calculated by the different of time and the origination of loan. The loan age indicates the amortisation schedule for loan. The CEP is low at the age of 20 to 40, and it increases after the age of 40. It shows that the liquidity of loan repayment of borrower is high at the beginning of the loan age and at the later age of the loan.

- **FICO_orig_time by Time**

FICO score at the origination time is low before the time of GFC which is between 600 to 660. After the GFC, the FICO score is above 660 which is good credit score.

- **FICO_orig_time by Vintage**

FICO score of loan originated has an increasing trend. After GFC period, it increases and then has a huge plummet in the period 37. This plummet point is which the FICO score is below 600, so it indicates there is subprime lending offered to individuals by the bank in this period with the exceptional higher interest rate than prime borrowers.

- **FICO_orig_time by Age**

With age, the FICO score decreases, and the shape is right skewed from period 0 to 50. The FICO score is high at the age of 20 which is above 680, and the score decreases after that age.

(B) Interpret findings (based on the chart outputs)

- **Visualize LTV_time bucket and default rates**

Forming 15 classes for the loan to value category is by self-defined boundaries. There is a moderate relationship between default rate and LTV class 1 to 6. The relationship between LTV class 6 to 12 is a strong positively. From LTV class 12 to 15, there is a moderate negatively. The LTV is high at above 80%, which means the default rate is higher compared to the LTV below 80%. Nevertheless, if the LTV exceeds 120%, it indicates that there is a saturation effect which translates to the negative relationship between LTV and default rate.

- **Visualize cep_time bucket and default rates**

Forming 14 classes for Cummulative Excess Payment over time category is by self-defined boundaries. There are between CEP and default rate: CEP class 1 to 5 (moderate positive), CEP class 5 to 7 (strong negative) and CEP class 7 to 11 (moderate positive). The low rate of CEP indicates higher default rate, therefore, when the CEP is below -0.02 indicating high default rate of 0.03. The higher CEP means lower default rate. As such, the CEP is above 0.02 indicating the very low rate of default rate which is below 0.010 as CEP represents the liquidity of the borrower.

- **Visualize FICO_orig_time bucket and default rates**

Forming 9 classes for the FICO score category is by self-defined boundaries. There is relationship between FICO score and default rate: FICO class 1 to 3 (moderate positive) and FICO score class 4 to 9 (strong negative). The default rate is low when there is high FICO score above 550 which is good credit score. There is high default rate when the FICO score is below 500, as such the default is above 0.040 for these subprime borrowers.

<u>QUESTION 2</u>: PD modelling

(A) Estimate a credit risk model for mortgage default probabilities (PD)

Applying a logistic regression model to predict mortgage default probabilities (PD) which model is as below $PD = \beta 0 + \beta 1 * \text{feature1} + \beta 2 * \text{feature2} + \beta 3 * \text{feature3} + \epsilon$. The intercept which is $\beta 0$= -2.5379. The three features that are estimated for the PD model are LTV time, CEP time and FICO score. The LTV time coefficient is 0.0257, and the CEP time coefficient is 2.5453, and the FICO score coefficient is -0.0052. The mean of the PD is fitted with the default rate which is 0.024. The computation of the estimated PD for all mortgage loans and periods is shown in the data 2 as above under the column PD_logistic_model.

Plot and compare the PD with default rate over time and with each feature is shown in the above charts.

In the chart of average PD which is computed from the logistic regression model versus the default rate: it shows a straight line as the computed PD is fitted with the default rate

In the chart of average PD versus CEP time: there is a linear line from scatter points for the CEP time below -4, and above -1. The CEP time coefficient is 2.5453 showing the positive relationship between PD and CEP time.

In the chart of average PD versus FICO score: there is a strongly negative relationship that makes a linear line as shown in the chart. The FICO score coefficient is -0.0052 which indicates that is the negative relationship. The increase in FICO score means the low PD.

In the chart of average PD versus LTV time: there is a strong positive relationship between PD and LTV time as the LTV time coefficient is 0.0257. From the chart, it shows the higher rate of LTV creating very high probabilities of default.

(B) Estimate a credit risk model for mortgage default probabilities (PD) by including a non-linear transformation of features set

In this PD model, a non-linear transformation of the feature set (LTV time, CEP time and FICO score) is used to get a better model accuracy. Non-linear transformations applies polynomial terms for the feature set with the power of one, two and three. The out-of-sample may observe a better prediction for probabilities of default. The PD model is as following: PD=β0 + β1*feature1^1 + β2*feature1^2 + β3*feature1^3 +...

The intercept which is β0= 18.9856. The LTV time p1 coefficient is -0.0612, the LTV time p2 coefficient 0.0012, the LTV time p3 coefficient is -5.082e-06. The FICO score of p1, p2, p3 is as following -0.1084, 0.0002, -9.902e-08. The CEP time coefficient of p1, p2, p3 is as following -1.3270, -10.9449, 13.8780. The coefficient of these out-of-sample is larger than the model in the question 2a. The mean of the PD is fitted with the default rate which is 0.024. The computation of the estimated PD for all mortgage loans and periods is shown in the data 3 as above under the column PD2_nonlinear.

Plot and compare the PD with default rate over time and with each feature is shown in the above charts.

In the chart of computed average PD versus the default rate: it shows a straight line as the computed PD is fitted with the default rate

In the chart of average PD versus CEP time: there is a nonlinear line from scatter points for the CEP time above -1. There are high probabilities for the CEP time above -1

In the chart of average PD versus FICO score: there is a strongly negative relationship that makes a linear line as shown in the chart as in the one in question 2a. The increase in FICO score means the low PD, and the higher FICO score which is above 600 creates lower probabilities of default.

In the chart of average PD versus LTV time: there is nonlinear line PD and LTV time. From the chart, it shows the observed samples which are the LTV rates above 100 creating very high probabilities of default.

(C) Compare the accuracy of two models from sub-questions 2A and 2B

Creating test and train data for the two models in 2A and 2B is to compare the accuracy of two models. The better accurate prediction model is evaluated based on the fit of the models regarding to adjusted R^2, mean outcome and mean fit.

From the validation of the 2A model, the fitting sample in the training data is favourable as the data is used to train the data, whereas in the testing data does not fit well for the time below 45 as in the time-series fit, and it is fitted in the large deviation of the PD model.

From the validation of the 2A model, the training data is fitted well, whereas the testing data is fitted favourably from the time above 50 shown in the time-series fit.

In summary, the 2A training model adjusted R^2 is 0.0084 which means 0.84% of the data fits the regression model, whereas the 2B training model adjusted R^2 is 0.0094 meaning the data is more fitted the regression model compared to the 2A model of 0.94%. In the perfect model, the mean fit and the mean outcome will be matched, however, the testing model of the two models does not fit favourably.

In conclusion, the model 2B creates higher variables which are nonlinearities to forecast the PD more accuracy in comparison the two models' adjusted R^2. Moreover, there are many variables related to the economic variables which might be good choices to choose for better accuracy prediction model such as GDP of the country or unemployment rate as they are related to the financial ratio of the bank to forecast the probabilities of default for the bank.

QUESTION 3: LGD modelling

(A) Estimate a linear regression model to predict LGD

Firstly, calculating the resolution period is to find out the LGD The computation of the LGD based on the formula: $LGDit=EADit-\sum T\tau=1(CFt+\tau/(1+rt+\tau)\ t+\tau)\ EADit$

Using the function resolution bias is to correct observed LGD for resolution bias. The prediction of LGD is estimated by a linear regression model, and the independent variables are in the feature set (LTV time, CEP time and FICO score). The intercept which is $\beta 0$= 0.5219. The LTV time coefficient is 0.0042 (positive relationship), and the CEP time coefficient is 0.5091 (positive relationship), and the FICO score coefficient is -0.0004 (negative relationship). The mean of the implied LGD is nearly fitted with the observed LGD which is 0.670. The computation of the estimated implied LGD for all mortgage loans and periods is shown in the data_default3 as above under the column LGD_linearmodel.

Plot and compare the implied LGD with observed LGD over time and with each feature is shown in the above charts.

In the chart of average LGD implied which is computed from the linear regression model versus the observed LGD: it shows that they are not correlated with each other.

In the chart of average LGD versus CEP time: there is no correlation for the CEP time and the implied LGD. The CEP time coefficient is 0.5091 showing the positive relationship between LGD and CEP time. There is high LGD when the CEP time is between -0.2 and 0 as CEP shows the liquidity ability of the borrowers, so the lower CEP indicating high loss rates of given default.

In the chart of average LGD versus LTV time: there is a strong positive relationship between LGD and LTV time (linear line) as the LTV time coefficient is 0.0042. From the chart, it shows the higher rate of LTV creating very high LGD, which LTV is above 100% indicating a very high risk and high loss rate given default

In the chart of average LGD versus FICO score: there is a strongly negative relationship that makes a linear line as shown in the chart. The FICO score coefficient is -0.0004 which indicates that is the negative relationship. The high FICO score means the lower LGD and vice versa.

(B) Estimate a linear regression model to predict LGD by including a non-linear transformation of features set

In this LGD model, a non-linear transformation of the feature set (LTV time, CEP time and FICO score) is included to get a better model accuracy. Non-linear transformations apply polynomial terms for the feature set with the power of one, two and three. The out-of-sample may observe a better prediction for probabilities of default. This LGD model is estimated by a linear regression model to predict LGD.

The intercept which is $\beta 0$= -3.6749. The LTV time p1 coefficient is -0.0160, the LTV time p2 coefficient 0.0003, the LTV time p3 coefficient is -1.041e-06. The FICO score of p1, p2, p3 is as following 0.0215, -3.371e-05, 1.715e-08. The CEP time coefficient of p1, p2, p3 is as following 0.5365, 1.1710, -1.8673. The coefficient of these out-of-sample is larger than the model in the question 3A. The mean of the LGD is nearly fitted with the observed LGD which is 0.671. The computation of the estimated LGD for all mortgage loans and periods is shown in the data_default4 as above under the column LGD2_linearmodel.

Plot and compare the implied LGD with observed LGD over time and with each feature are shown in the above charts.

In the chart of average LGD implied which is computed from the linear regression model versus the observed LGD: it shows that they are not correlated with each other.

In the chart of average LGD versus CEP time: there is no correlation for the CEP time and the implied LGD. There is high LGD when the CEP time is between -0.2 and 0 as CEP shows the liquidity ability of the borrowers, so the lower CEP indicating high loss rates of given default. There is out-of-sample observation which are the value of CEP time above 0.6 and below -0.4 indicating high LGD.

In the chart of average LGD versus LTV time: there is nonlinear line LGD and LTV time.

In the chart of average LGD versus FICO score: there is outliers with the FICO score below 500 and the score above 800. The higher FICO score also indicates high LGD as observed from the chart.

From the two model 3A and 3B, the adjusted R^2 of linear regression model 3A is 0.074 and the adjusted R^2 of linear regression model 3B is 0.084. It indicates the model 3B is more fitted to the regression model than the 3A model.

(C) Suggest two additional features that can explain for mortgage LGD

The two additional features are suggested for predicting LGD mortgage, which are interest rate time and GDP over time as these two variables indicate the economic variables affecting the loss rate given default. Logically when there is low interest rate which indicates the government is implementing monetary policy expansionary to increase consumption and investment. The higher GDP indicates the economy is in good situation which might be in the boom period. These two variable features are reasonable to add in the linear regression model to predict LGD.

The validation of the training and testing model with two additional features shows that the model is quite perfect as the mean outcome and the mean fit are matched. The training model does fit well with the large deviation, and it does not fit well for the periods below 25 as in the time-series chart. The testing model is fit well.

The intercept which is $\beta 0$= -0.2561. The LTV time coefficient is 0.0031 (positive relationship), and the CEP time coefficient is 0.0894 (positive relationship), and the FICO score coefficient is 0.0003, and the interest rate coefficient is 0.0218, and the GDP coefficient is 0,0081. However, the adjusted R^2 for this model which is 0.034 is smaller than the model 3A and 3B.

From the two charts, it shows the lower GDP implies higher LGD, and the higher interest rate indicates high LGD.

<u>QUESTION 4</u>: Bank capital allocation

In the setting I, the PDs are inferred from the model 2A and LGDs are inferred from the model 3A. In the setting II, the PDs are inferred from the model 2B and LGDs are inferred from the model 3B. The Basel capital ratio measures the Credit value at risk with the formula:

$$CVaR=(WCDR-PD)*DLGD*DEAD$$

The Downturn LGD is assumed as the following function DLGD: 0.08+0.092* LGD which US regulators apply for bank. The correlation is assumed at 15%. The Basel capital for mortgages is the result of multiplying then annual WCPD with the Downturn LGD and the Downturn EAD.

The chart in the setting I and setting II shows the annual WCDR, relative capital ratio and annual PD of the default rate and mean PD throughout the time series. The annual WCDR indicates the worst case of default rate (green line). The capital adequacy is calculated by the difference between the annual WCDR and the annual PD with the Downturn LGD. The relative capital in this setting is at the stable default rate of 0.2%, and it is higher from the period 175. In the setting II, the default rate is more fluctuate from period 175, whereas the default rate is quite smooth in the setting I.

QUESTION 5: Loan pricing

Assuming those following features are LTV_time =80, cep_time=0, FICO_orig_time=700 to compute the appropriate interest rate for a borrower with these assuming features. The model prediction is using linear regression model to forecast interest rate, and the independent variables are LTV time, CEP time and FICO score. From the model, the appropriate interest rate is 6.699111%

QUESTION 6: Bonus question – Create a dashboard application using package JupyterDash

(A) Draft a dashboard application for loan pricing

Using regression model to compute the interest rate estimate with choices of features. The choices of variable are GDP rate, Unemployment rate and Time as they are relative to economic variables which might be reasonable and related to forecast interest rate. The mean of interest rate is 6.702 from the model. Plot the interest rate prediction with each variable to draft a dashboard application. The input for the x-axis is the three choices of features (gdp_time, uer_time and time) and the y-axis is the interest rate.

(B) Create a dashboard application using package JupyterDash

Creating a dashboard application using package JupyterDash from the draft application in question 6A and uses the framework of the Plotly library for Python which is in the reference list.

REFERENCE

Giordani, P., Jacobson, T., Schedvin, E.V. & Villani, M. 2014, 'The Journal of Financial and Quantitative Analysis', *Cambridge University Press on behalf of the University of Washington School of Business Administration*, vol. 49, no. 4, pp. 1071-1099

Plotly 2021, *Plotly express in Python*, viewed 25 October 2021, <https://plotly.com/python/plotly-express/>

Plotly 2021, *Dash Python user guide,* viewed 25 October 2021, < https://dash.plotly.com/>

Rosch, D. & Scheule, H. 2016, *Deep credit risk - Machine learning with Python*