**Framework for the CECL Project**

1. **Aim:**
   1. To develop a linear model that can fulfill Current Expected Credit Losses (CECL) requirements.
   2. To communicate the results of our model to our Financial mathematics colleagues.
2. **Data Used:**
   1. We use Fannie Mae single-family loan data from 2011 to 2017.
   2. The loan data consists of 2 data sets, an acquisition file, and a performance file.
   3. The acquisition file includes loan details and parameters picked at loan origination. This data set never changes over time.
   4. The performance file contains a historical record of how the servicing of the loan. It will indicate instances of prepayments, delinquencies and default. This file is updated every quarter.
   5. The acquisition and performance data sets both have a reporting lag of one quarter.
   6. We also use macroeconomic data sourced from online pages such as Bloomberg.
   7. We will focus on the state of California for this report.
   8. We will use a random sample of 30% if the data.
   9. We will further split the data into a train data set and a test data set.
3. **Data Structure:**
   1. This section describes the data that we will use for the exercise:
   2. We will combine data from the acquisition and performance data sets.
   3. Our final data set will have the following fields: ID, Original date, Payment date, Current Delinquency Status, Original UPB and Current UPB.
   4. We will then combine the data set above with select macroeconomic variables. We intend to use the following macroeconomic variables: CPI, HPI, Unemployment rate, 10-year Treasury bill interest rate(Yield), S&P 500 index, GDP, currency strength, building permits, housing starts, house sales, lumber, household income, fixed-rate mortgage average, business climate, housing inventory estimate for sale and Rental Vacancy Rate.
   5. The macroeconomic variables are often given in quarterly or annual periods. We will need to adjust these variables into a monthly estimate using appropriate assumptions. For instance, we may opt to uses the change in values over the period to achieve the required stationarity. Inputs such as unemployment can be input as is since we expect it to be naturally stationary.
4. **Methods:**
   1. Our goal is to predict the future credit losses of the loans as per the data provided.
   2. The most popular framework for credit losses is the PD \* LGD \* EAD
   3. The PD stands for Probability of Default. For that part, we will use logistic regression.
   4. The PD section is the most intensive section of this framework and we will spend the most time and effort on it.
   5. **Logistic Regression:**
      1. To get the default probability of each loan, we need to build a model based on the sampled data.
      2. We will then validate our model by predicting a train data set.
      3. We will need to be careful when interpreting the effects of macro-economic variables on default. The factors may be important if there is significant change over periods. However, it may not be that predictive if the variable is stable.
      4. While in this case we elected to use a linear modeling framework, we would still recommend looking into a general Markov transition modeling framework.
      5. This is because of the greater flexibility and ease of use and implementation as a result of the switch.
      6. The above model can be used as a natural check for this model.
   6. After obtaining probabilities from the logistic regression above we will plug the figures into a probability transition matrix with six states; Prepayment, Stage1, Stage2, Stage3 and Default
   7. LGD stands for Loss Given Default. For this part, we will need to obtain recovery rates of the loans and model their likelihood based on similar parameters as the PD model.
   8. EAD stands for Exposure at Default. EAD is the size of the portfolio exposed to the risk of loss.