|  |
| --- |
| SINGLE FAMILY LOAN PERFORMANCE MODEL  MOMODE |
| hone: [Your Phone] |

# PROBLEM STATEMENT

Financial institutions conduct stress tests to gauge the resilience of their balance sheets to substantial macroeconomic shocks. One way to measure the performance of a financial institution is by assessing the institution’s loan portfolio loss under stressed scenarios. The first step in assessing loan loss is to estimate the probability of default (PD).

 Understanding PD is necessary for the purpose of stress testing and risk management. Financial institutions may also find it beneficial as insights from default modeling can be incorporated to guide improvements on good underwriting practice and competitive mortgage pricing.

The proposed Model attempts to predict default risk by estimating relationships between default risks and default determinants.

The Model will help financial institutes to run their portfolio and get understanding of risks

# TARGET AUDIENCE

  Though we are using dataset provided by Freddie Mae, which is one of the biggest mortgage buyers from various mortgage banks. The use of this data analysis and model can also be extended to banks that hold big mortgage portfolios.

The clients have incentive to identify the loans that are risk of default and provide help to high-risk loans to minimize loss.

DATA DESCRIPTION USED FOR MODELING

This data is made available from Freddie Mae.

<http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html>

For every year the loan contains zipped file for each quarter example.

Each zipped file has two dataset in tab separated txt file

1. Single Family Loan-Level Dataset.
2. Monthly Performance Dataset

**Single Family Loan-Level Dataset has data at the time of loan was originated**

* CREDIT SCORE
* FIRST TIME HOMEBUYER FLAG
* MORTGAGE INSURANCE PERCENTAGE (MI %) -
* ORIGINAL COMBINED LOAN-TO-VALUE (CLTV)
* OCCUPANCY STATUS
* ORIGINAL UPB
* ORIGINAL LOAN-TO-VALUE (LTV)
* ORIGINAL DEBT-TO-INCOME (DTI) RATIO
* LOAN SEQUENCE NUMBER

**Monthly Performance Dataset has monthly performance of data from date of origination to end of 2015.**

* LOAN SEQUENCE NUMBER
* MONTHLY REPORTING PERIOD (YYYYMM)
* CURRENT UPB
* CURRENT LOAN DELINQUENCY STATUS
* LOAN AGE
* REMAINING MONTHS OF MATURITY
* REPURCHASE FLAG
* MODIFICATION FLAG
* ZERO BALANCE CODE
* ZERO BALANCE EFFECTIVE DATE
* CURRENT INTEREST RATE
* CURRENT DEFERRED UPB

APPROACH TO DEFINE A MODEL

Most of the banks or financial institutes publish their report or analyses the risk quarterly we take do the risk analysis for each quarter.

In mortgage industry if a loan payment is past the due date it gets marked as a Non-Performing loan.

Our aim is to find a model that will predict if a given loan will miss the payment and hence become a Non-Performing loan in the portfolio. Since expected outcome is a binary result where loan can fall in either of the two categories.

We make use of Logistic Regression model to measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a [logistic function](https://en.wikipedia.org/wiki/Logistic_function), which is the cumulative logistic distribution

STEPS TO DEFINE THE MODEL

1. DATA WRANGLING:

We get the data from published dataset by Freddie Mae. This data is not clean and needs to be transformed and cleaned. The missing value by understanding the understanding the user guide provided with the dataset.

1. FEATURE CORRELATION ANALYSIS:

The dataset defines and captures many different fields for a given loan. These columns are captured at the time of loan origination and also to define monthly loan performance. We study the correlation of various features with the loan delinquency status, which tells us how many days a given loan is behind its payment date.

We also study the correlation between each feature to avoid including highly correlated features in our modeling.

1. TRAIN TEST THE MODEL:

The dataset is divided into train and test data. The training data is used to define best-fitted model. The test data is then used to compare predicted output of the model with the expected output.

1. EVALUATE THE MODEL:

After we get our model, we run the standard evaluation and scoring function to test the quality of the mode.

DELIVERABLES

* Code creating data set modeling and evaluation
* Slide Deck telling a compelling data story.