Freddie Mac Case study

Table of Contents

Context Overview:

- The Federal Home Loan Mortgage Corporation (FHLMC), commonly known as Freddie Mac, is a publicly traded, government-sponsored enterprise, headquartered in Virginia, USA.
- The FHLMC was created in 1970 to expand the secondary market for mortgages in the US.
- Along with the Federal National Mortgage Association (Fannie Mae), Freddie Mac buys mortgages, pools them, and sells them as a mortgage-backed security to private investors on the open market.
- This secondary mortgage market increases the supply of money available for mortgage lending and increases the money available for new home purchases.
- So, essentially, Freddie Mac does not lend directly to the home buyers, instead they buy the mortgages from the lenders so that the lenders can continue to lend to home buyers.
- The concerned data corresponds to the mortgages that have opened home loan accounts during Mar 2018 to Dec x2020 and how regularly they have made their payments all the way till Mar'22 – starting from account opening date.
- The objective of exercises detailed in subsequent sections requires the data scientist to conduct exploratory data analysis on the datasets provided – based on guidelines provided, and then build certain models that may help Freddie Mac make a better decision on which home loans are potentially more risky, where should recommendation of certain collection efforts be focussed towards and what should be fair estimates of loss for onboarded applicants.
- Details of the data are provided in the next section.

Data & Metadata Details:

- The datasets are broadly the following:
    - Originations Data: this data corresponds to the details of the applicant(s) who have been accepted by some bank for a home-loan product. It contains details regarding the home loan account and the details about the applicant – information like credit score, Loan-to-value, Debt-to-Income Ratio, whether it is the first home loan availed by the applicant, etc. – all captured at the point the applicant had applied for the home loan. This file also contains information about the home loan, the principal amount (UPB), interest rate, first payment date, property type, etc. Note that this file contains information only about accepted home loan applicants – there is no information corresponding to applicants whose home loan applications were rejected.
    - Behaviour Data: this data contains how the applicant performed once their home loan started. This file essentially contains the month on month delinquency bucket

information. If an account has no pending dues as of a certain month, then their delinquency bucket is 0. If they have missed 1 payment, then delinquency bucket is 1 and so on. Delinquency bucket is 99 for cases for whom delinquency information is not available – records having such information can be excluded from further analysis. The behaviour data also contains any recorded loss from a certain loan. This information is also captured at an account level – only for accounts that have gone 6 cycles delinquent or more ever during their tenure.

- The Originations and Behaviour data can be merged using the Loan Sequence Number. Both of these files should ideally also be unique based on the UID Loan Sequency Number.
- The DataLayout.xslx file provided separately – contains details of the fields contained in the Originations and Behavioural datasets.

Case Study 1 [Exploratory Data Analysis]:

- Originations EDA:
  - Report the following for different kinds of variables – in excel:
    - Total number of records
    - Missing Counts and % - for each variable
    - UID variables – number of duplicates, number of unique values
    - For discrete variables, report frequencies of all discrete values (maximum 100). If more than 100, then club smaller frequencies together under Others so that no more than frequency of 100 elements are reported. Avoid variables like Postal Code – which are anyways expected to have discrete but lots of values
    - For continuous numeric variables – like amount variables or ratio or percentage variables, report minimum, the 1, 5, 10, 25, 50, 75, 90, 95, 99 percentiles, maximum and mean values
    - Plot, within excel, the monthly volumes of applications (use First Payment Date for getting the monthly volumes) – from Mar'18 to Dec'20. Alongside, plot the mean of the Credit Score for applications corresponding to each of those months – on a secondary vertical axis.
    - Share your observations based on the above plot.
- Merge the Originations file with the Behaviour File – using inner join. Loan Sequence Number is the UID based on which these 2 files should be merged. Are there any records in the Originations file which do not have a match in the Behaviour File, and vice versa?
- Using the monthly delinquency buckets describing payment regularity of the applicants, create a variable called Maximum Delinquency – which captures the maximum delinquency buckets across all the months that each applicant has ever touched. Records having delinquency bucket value of 99 in any of the months should be ignored from this analysis
- Vintage Analysis:
  - Vintage Analysis is conducted to understand how long do accounts typically need to go bad, if they have to turn bad i.e. delinquent.
  - Here, let us do a vintage analysis of 3 cycles delinquency, since 3 cycles is a default definition used quite often in the banking industry.
  - The aim is to look at the accounts originated every month (First Payment Date belongs to a certain month) and see what percentage of accounts hit 3 cycles

delinquency after 1 months, 2 months, etc. for the entire range of monthly data available.

- So, accounts originated in Mar'18 shall have nearly 48 months of data till Mar'22 whilst accounts originated in Dec'20 shall have only around 15 months of data till Mar'22.
- The objective is to create a 2-dimensional matrix – where there is Month of Origination on one-side and Months to Become Bad (i.e. 3+ cycles delinquent) on the other. Of course, a greater chunk of the population would have never touched 3 cycles or more – which can be recorded separately.
- Now, plot the above matrix in a chart with Months to Become Bad as the horizontal axis and different legends for each Month of Origination (from Mar'18 to Dec'20).
- Analyse the vintage curves. Share your insights on what do you think is typically time taken for accounts to go bad.
- Think how a vintage curve can help provide early indication about the quality of a particular month of bookings.