Fannie Mae Loans Servity Analysis

Deriving features from crude data

Prithwish Maiti

12/09/2024

Abstract

Fannie Mae single-family mortgage loans data contains a lot of information on the loans they have acquired. It includes a robust comprehension of the performance of the loans and any delinquency procedure if conducted. However much of the data is in crude form, that is reflectes only the procedure in which it was collected and compiled. There is a need to refine the data and extract useful features . Here I present the procedure , that I have formulated to do the same and them form a predictive model to form a LGD prediction table and subsequently a servity model.

Content

1 Introduction to the data 2
1.1 Other observation 2
2 Feature extraction 2
2.1 Grouping columns and idea of feature extraction 3
2.2 Sequential Formation of features 6
3 Merging Macroeconomic data 7

1 Introduction to the data

After navigating to Fannie Mae's single family loans performance dashboard, we find that the data is arranged in quarter-wise manner. Each file's name is acquisition and performance which states that, historically, the acquisition data and performance data were present in separate files, which now have been merged together. The loans are entered into the particular file(year and quarter) when it was acquired by Fannie Mae, which may be unequal to their origination date.

This also means that the last activity of that loan is contained in the same report.

Loan ID	Acquisition Columns	Delinquency Status	Dispsition ?	Costs
1542683	....	0	No	N/A
	......(May have changes)	..	No	N/A
	....	10 (Deed-in-Lieu)	Yes	56000 4500 3200
1545679(Next)	....			N/A

Table 1 - A visualisation shown how the data is arranged. Note that any losses appear in the very last row and not before that.

If the loan is cleared without any default then the status is generally 01 – Which stands for prepaid or cleared. The challenge of the data set is that any losses appear in the very last row and not before that. So we cannot build continuous evaluation process, and resort to a prediction model. Currently I'm flattening the data over the time by attracting features out of this time series. We can train a time series model, but that would require higher level of engineering as the time period of the loans are not constistent.

1.1 Other observation

There might be some mistakes in entering the data. So we always rely on the later rows as we assume that the mistakes might be corrected later.

Some of the features were added after 2020. Eg:- ADR_TYPE, ADR_COUNT, Total Defferal Amount, Some after 2023- Payment Deferral Modification Event Indicator, So they are of no use to us as our primary focus is on the period before 2020.

HIGH_LOAN_TO_VALUE_HLTV_REFINANCE_OPTION_INDICATOR, DEAL_NAME have all null values in most of the files(check).

2 Feature extraction

The performance columns are left joined to the loans' acquition data. The acquisition data consists of fields which are static throughout the loans existances until closure. They are not suppose to change but we relay on the last column as it represents the most recent information. The are several fields that the performance columns measure like delinquency status and remaining upb which are shown over time and forclosure related costs and proceeds(capital seized/obtained to subdue the losses) which are presented once at the final row of the loan. Hence, we can finally visualise the data as a combination of static fields combined with time series data and some other fields which may not be present.

So far, we can only highlight the one useful feature of the data, that is all the rows (denoting a time period) of a loan, are included within the same file, making the aggregation of all the data per loan easy. We use this advantage to process a particular loan altogether by processing an acquisition and performance file of a quarter(file).

2.1 Grouping columns and idea of feature extraction

We demonstrate the sequential formation of features by forming tables at different stages which highlight the extraction of some particular features The first one being the acquisition data itself It's worth noting that we have to add the acquisition date, which is not mentioned in data.

Here is a description of the acquition data-

Loan_id, act_period	Primary ID of Loan	Acquisition Columns (static and not subjected to change unless there is some data discrepancy)
Channel, seller,	Lender Details	
Orig_rate, orig_upb, orig_term, orig_date, first_pay,	Loan agreement details	
Oltv, ocltv, num_bo, dti, cscore_b, cscore_c,	Credit Information variables	
First time buyer?, purpose, property type, no_units, occ_stat, mi_pct, FRM or ARM, mi_type	Purpose of loan , and profile description of buyer	
State, zip,	Geographical address	

There are some indicators that denote certain properties of the loan that was used/ effected at the origination.

1. Relocation_mortgage_indicator
2. Homeready_program_indicator
3. Property_inspection_waiver_indicator
4. High_balance_loan_indicator
5. High_loan_to_value_hltv_refinance_option_indicator
6. Deal_name

Next the description of Performance data-

LOAN_ID, SERVICER,	primary and servicer(tracked to see if changed)
CURR_RATE	The current rate at the point
CURRENT_UPB, LOAN_AGE, REM_MONTHS, ADJ_REM_MONTHS, DLQ_STATUS, MOD_FLAG, Zero_Bal_Code	Core performance Variables
ZB_DTE, LAST_PAID_INSTALLMENT_DATE, FORECLOSURE_DATE, DISPOSITION_DATE,	Related dates

Costs and proceeds to delinquent loans proceeding to foreclosure-

FORECLOSURE_COSTS, PROPERTY_PRESERVATION_AND_REPAIR_COSTS, ASSET_RECOVERY_COSTS, MISCELLANEOUS_EXPENSES_AND_CREDITS, ASSOCIATED_TAXES_FOR_HOLDING_PROPERTY, NET_SALES_PROCEEDS, CREDIT_ENHANCEMENT_PROCEEDS, REPURCHASES_MAKE_WHOLE_PROCEEDS, OTHER_FORECLOSURE_PROCEEDS,	Costs adding to losses
NON_INTEREST_BEARING_UPB, PRINCIPAL_FORGIVENESS_AMOUNT, repch_flag, LAST_UPB	Some relaxation on Losses
NET_SALES_PROCEEDS, CREDIT_ENHANCEMENT_PROCEEDS, REPURCHASES_MAKE_WHOLE_PROCEEDS, OTHER_FORECLOSURE_PROCEEDS,	Proceeds from Forclosure

COMPLT_FLG = is disposition is done or not when loan is in default state. The defaulted state include:

- 09 = Deed-in-Lieu; REO Disposition

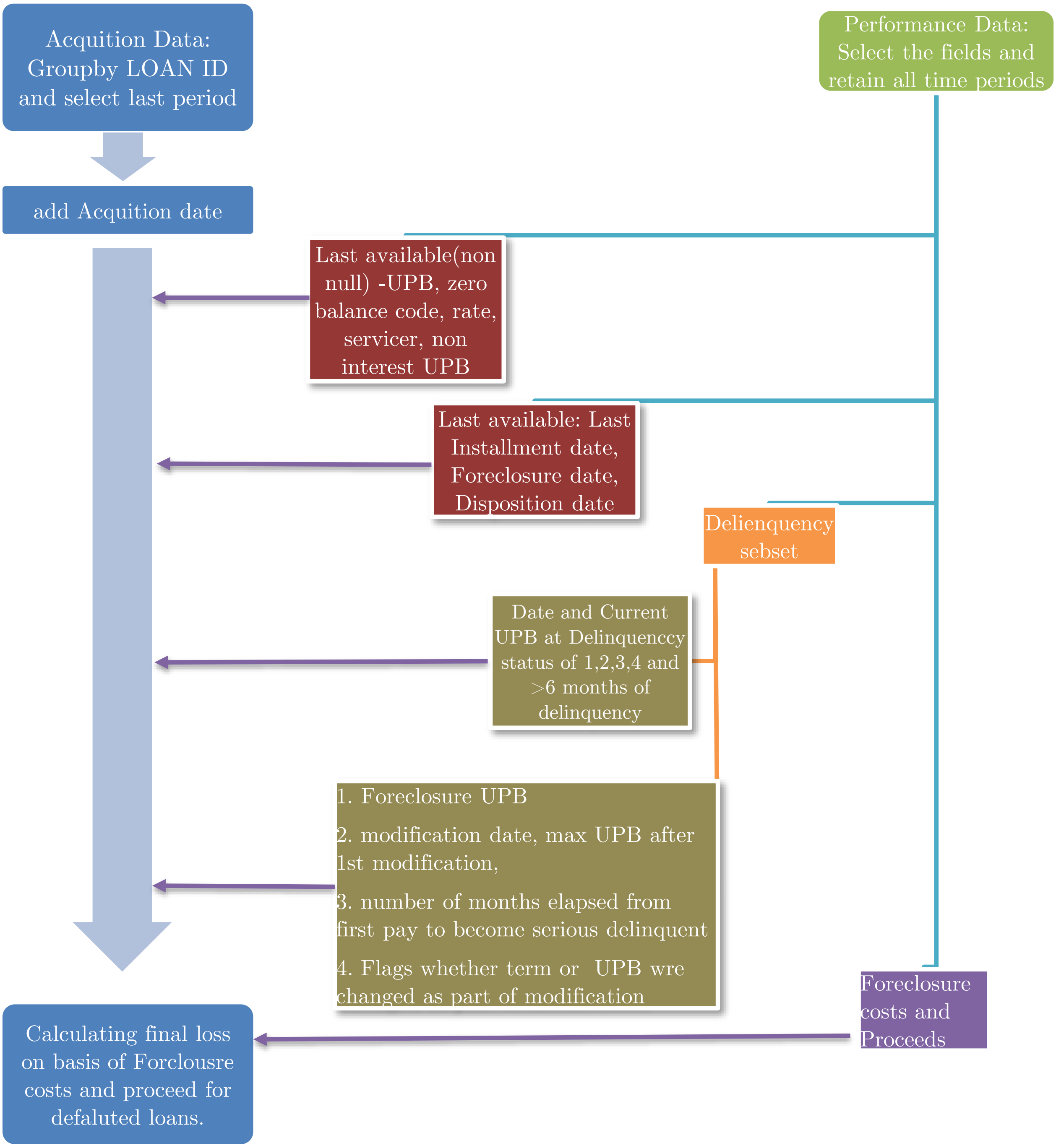
- 02 = Third Party Sale
- 03 = Short Sale
- 15 = Non-Performing Note Sale

INT_COST = the interest cost based on time between LPI date and LAST date

Modifying NA values in these columns to 0 or if COMPLT_FLG is 1

NET_LOSS = LAST_UPB + Sum of all costs - Sum of all proceeds

2.2 Sequential Formation of features


Acquition Data: Groupby LOAN ID and select last period
Performance Data: Select the fields and retain all time periods
add Acquition date
Last available(non null) -UPB, zero balance code, rate, servicer, non interest UPB
Last available: Last Installment date, Foreclosure date, Disposition date
Delienquency sebset
Date and Current UPB at Delingquenccy status of 1,2,3,4 and >6 months of delinquency
1. Foreclosure UPB
2. modification date, max UPB after 1st modification,
3. number of months elapsed from first pay to become serious delinquent
4. Flags whether term or UPB wre changed as part of modification
Foreclosure costs and Proceeds
Calculating final loss on basis of Forclousre costs and proceed for defaluted loans.

Problems faced – majority of the loans don't have a Current UPB data in delienquency status of 6 or more months(or the current UPB is 0). In that case, we use the origination amount to maintain consistency.

3 Merging Macroeconomic data