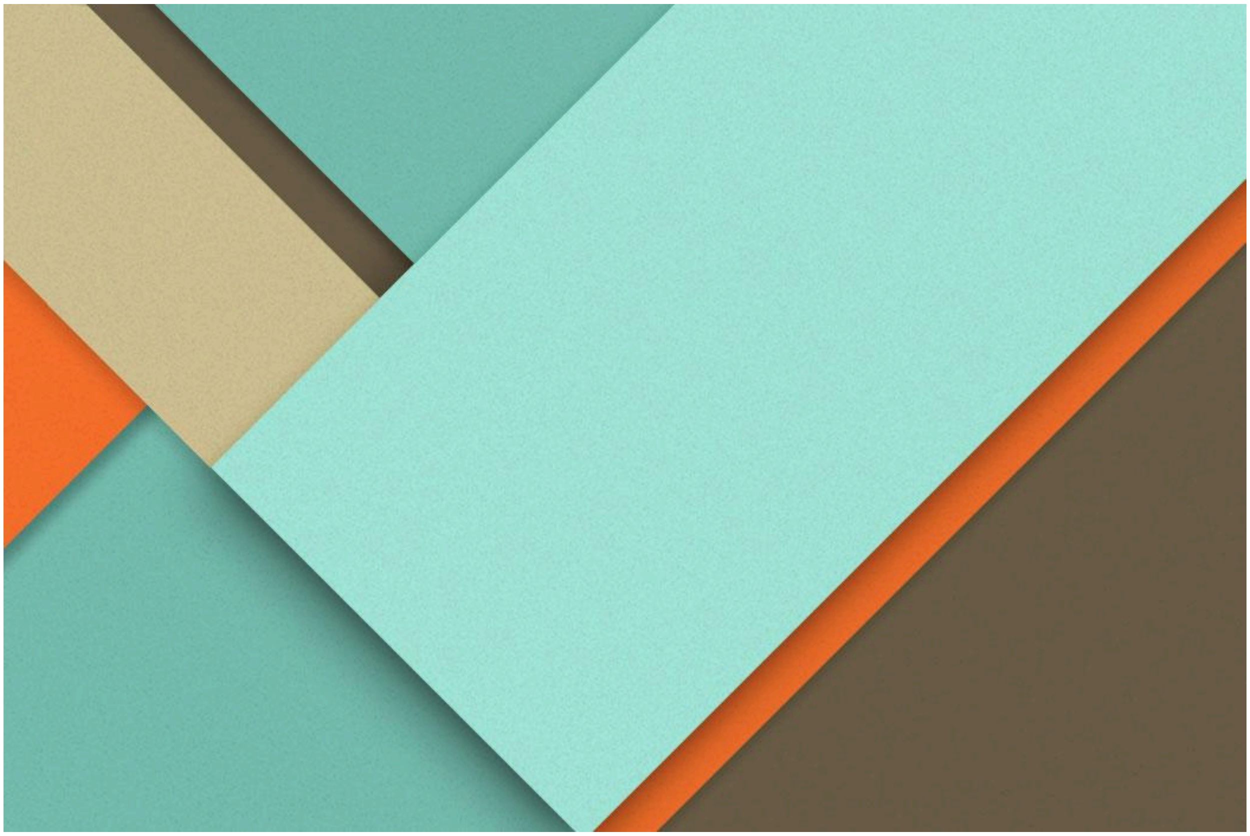


Capstone 2 - Mortgage default prediction

04.01.2024

Rohan Mani

Why?

Homeownership remains a foundational aspiration for many Americans, but one that has become more challenging to achieve in the 2020s. Home prices soared in the wake of the COVID-19 pandemic, as many Americans looked to upgrade their housing situation. Then the Federal Reserve (Fed) began raising interest rates (from near 0% to 5.50%) in 2022 and 2023 to offset the impact of inflation. The housing market was particularly hard hit by resulting higher mortgage rates. This project will help us understand the factors that impact mortgage loan defaults. It will help both mortgage lenders and consumers get a better understanding of the criteria that would influence the outcome of loan applications.

Banks are facing a challenge in lending due to mortgage rates. Loan default prediction is crucial for financial institutions to assess the risk associated with lending money to individuals and businesses. This would help banks to make better lending decisions such as adjusting loan terms/conditions, reserve additional funds to cover potential losses, or even deny/refuse loans to high risk borrowers.

The source data is a single family loan level dataset from Freddie Mac and has credit performance data on all mortgages that the company purchased or guaranteed from 1999 to 2023.

The dataset covers approximately 1.2 million mortgages originated between January 1, 1999 and March 31, 2023.

Audience

Banks are facing a challenge in lending due to mortgage rates. Loan default prediction is crucial for financial institutions to assess the risk associated with lending money to individuals and businesses. This would help banks to make better lending decisions such as adjusting loan terms/conditions, reserve additional funds to cover potential losses, or even deny/refuse loans to high risk borrowers.

Data source

The source data is a single family loan level dataset from Freddie Mac and has credit performance data on all mortgages that the company purchased or guaranteed from 1999 to 2023.

The dataset covers approximately 1.2 million mortgages originated between January 1, 1999 and March 31, 2023.

Data

The Freddie Mac site has the following files.

• One Origination Data file containing loan-level origination information such as Credit Score, Original Interest Rate, Loan Amount, Original Loan Period, Property Type, location, etc. for all the loans originated during the quarter.

• One Performance Data file for all of the respective loans originated during the quarter. All performance periods associated with a loan will be contained within the same Performance Data file. This file will have Loan Age, Loan Balance, Current Interest Rate, Loan Delinquency Status, etc.

For each year there is a sample origination data file and a performance data file. The sum total of the loans from the sample data file for years from 2020 to 2023 is 1.2 million loans. This will serve as the raw data for this project.

Data wrangling

Loan sample files for the period 1999 to 2023 (~1M loans) was downloaded from the Freddie Mac site. The original loan file and the loan performance history files were merged to obtain a single file with a unique entry for each loan along with all its attributes. The resulting raw loan data file has 1225000 loans with 63 loan attributes.

Columns with missing values

All these columns have more than 90% missing values, so we will drop them. Defect Settlement Date, Current Month Modification Cost, Due Date of Last Paid Installment, MI Recoveries, Non MI Recoveries, Expenses, Legal Costs, Maintenance and Preservation Costs, Taxes and Insurance, Miscellaneous Expenses, Actual Loss Calculation, Modification Cost, Zero Balance Removal UPB, Delinquent Accrued Interest, Current Month Modification Cost, Defect Settlement Date, Super Conforming Flag, Pre-HARP Loan Sequence Number, HARP Indicator, Property Valuation Method, Estimated Loan to Value, Modification Flag, Step Modification Flag, Deferred Payment Plan, Delinquency due to disaster, and borrower assistance status code.

- Columns with missing values were replaced with the median or mean or overwhelmingly majority value.
- The distribution for credit score is left skewed so the mean would be affected by the low credit scores. Hence, the missing values were replaced with the median which was 751.

- For Mortgage Insurance the missing values were replaced with the mean, 4.87.
- First Time Homebuyer Flag had 730 missing values which were replaced with 'N', because majority of the values are 'N'. The number of units column had 87 missing values. This which were replaced with the median value, 1.
- The Combined loan-to-value had 50 missing values which were replaced with the median which was 77.
- The Debt-to-Income column had 95,000 missing values which were replaced with the rate for HARP loans which is 45% according to Central Coast lending and the rest were replaced with 70% (as per freddie mac user guide the values missing in the data are either rates for HARP loans or exceed 65% for non HARP loans).
- The Loan-to-Value had 44 missing values which were replaced with the median which was 75. The channel column had 61 missing values which were replaced with 'TPO'.
- The Property Type variable had 46 missing values which were replaced with the 'Single Family' category because it was in the overwhelming majority. The Loan Purpose column had 1 missing observation which was replaced with the 'P'.
- The Number of Borrowers had 243 missing values which were replaced with the median which was 2.
- The Metropolitan/MSA column had missing values indicating it is either unknown or not MSA/Metropolitan. Replaced this with 0.

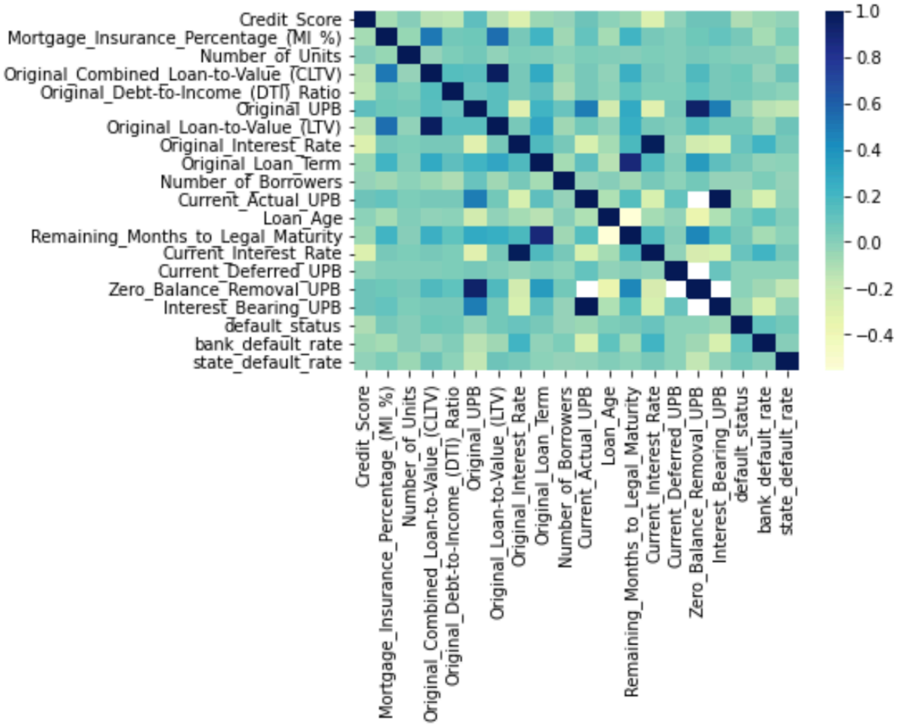
Identified loans that had an REO disposition code or an REO acquisition code and flagged them as default. This resulted in 1.3% loans (16429/1225000) flagged as default in the dataset indicating the dataset is imbalanced.

Exploratory Data Analysis

Here are the key findings from EDA:

- The highest number of loans are taken from homeowners in California with 146450 loans, followed by Texas with 75269 loans.
- The maximum default rate appears to be above 0.025 for Nevada. The other states are below this rate with California in the low area of the spectrum with under 0.010 default rate (even though we previously observed that California has the maximum loans). It may be good to include this feature as well to inform the model.

- The maximum default rate appears to be above 0.12 for Ocwen loan servicing. The other servicers are below 0.08. It may be good to include the servicer default rate statistics as a feature to inform the model.
- There is some high cross correlation between variables like Original UPB, Interest Bearing UPB, Zero Balance Removal, and Current Actual UPB. We also have a high cross correlation between CLTV, LTV, and Mortgage Insurance percentage. The exclusion of highly correlated variables will be looked into further during model preprocessing.



Preprocessing and training development

Here are the preprocessing steps that were performed:

- Preprocessed the data and computed relevant features.
- Summarized the information in Metropolitan Area/Division and encoded MSA as a new feature column with 0 indicating non-metropolitan areas and 1 indicating metropolitan.
- Created default rates by state and bank/servicer as new feature columns to inform the model.

- Dropped unneeded columns. The Payment date, monthly reporting period, and maturity date columns are not necessary since the required information is captured in loan age and remaining months to maturity. The information provided by delinquency status and zero balance code columns has been captured in default status. The default rates by servicer name has been included as a feature hence the servicer name column is no longer needed. The information in the seller name has 30% missing data and therefore being dropped. The default rate by property state is also included as a feature hence property state is not needed.
- Encoded categorical columns such as First Time Homebuyer, Loan Purpose, Property Type, Channel to numeric since most machine learning models accept only numeric data.
- Channel gets encoded to following features
  - Channel_B : Broker
  - Channel_C : Correspondent
  - Channel_R : Retail
  - Channel_T : TPO
- Occupancy Status gets encoded to following features
  - Occupancy Status_P : Primary
  - Occupancy Status_I : Investment
  - Occupancy Status_S : Secondary

Algorithms & Machine Learning/Modeling

The following models were implemented:

1. Random Forest (Standard Classification Model)
2. Cox Survival model (Survival Analysis Model)

The table below shows the model performance metrics.

Random Forest (Standard Classification Model)

Model	Recall	Precision	Accuracy	F1

Random Forest	0.98	0.99	0.987	0.98
Random Forest using SMOTE to balance the data	0.99	0.99	0.989	0.99

Survival Analysis Model (response variable is default status/event and loan age/duration)

Model	Concordance
Cox PH	0.853

I implemented the Random Forest model to see what the results look like using the standard classification approach. The performance metrics for the Random Forest Model using SMOTE suggested overfitting.

I used the Cox survival analysis guide and scikit-learn modules to implement the Cox survival model. The loan default prediction problem is akin to a survival analysis problem. Since the loan default event of interest is a time-based event and the loans in the dataset have varying origination dates over the years 1999 to 2023, the Cox survival model is the preferred model to move forward with. This model also shows good performance based on the concordance ratio metric above.

The Concordance Index (C-index) is one of the most used metrics as it encompasses both observed events and censored (ongoing) cases. In doing so, it quantifies the rank correlation between actual survival times and a model's predictions.

Model Predictions

Some of the interesting and significant features and their importance (coefficient/weight) as predicted by the model is given below.

Variable/Feature	Coefficient/Weight
state_default_rate	0.97

bank_default_rate	0.34
Current_Interest_Rate	0.13
Occupancy_Status (Primary Home)	0.05
MSA (Metropolitan Area)	-0.11
HARP_Indicator (Home Affordability Refinance Program)	-0.16

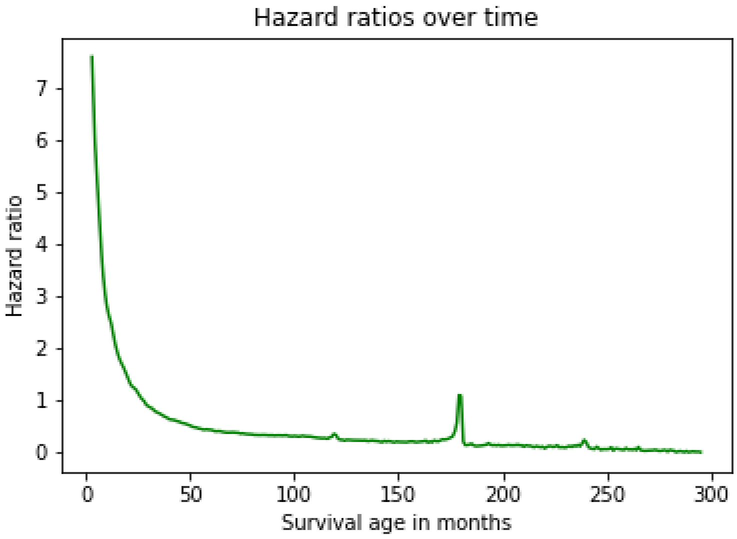
A key quantity in survival analysis is the so-called survival function, which relates time to the probability of surviving beyond a given time point. Let T denote a continuous non-negative random variable corresponding to a loan's survival time (time to default). The survival function S(t) returns the probability of survival beyond time and is defined as S(t)=P(T>t)

The Cox proportional hazards regression model can be written as follows:

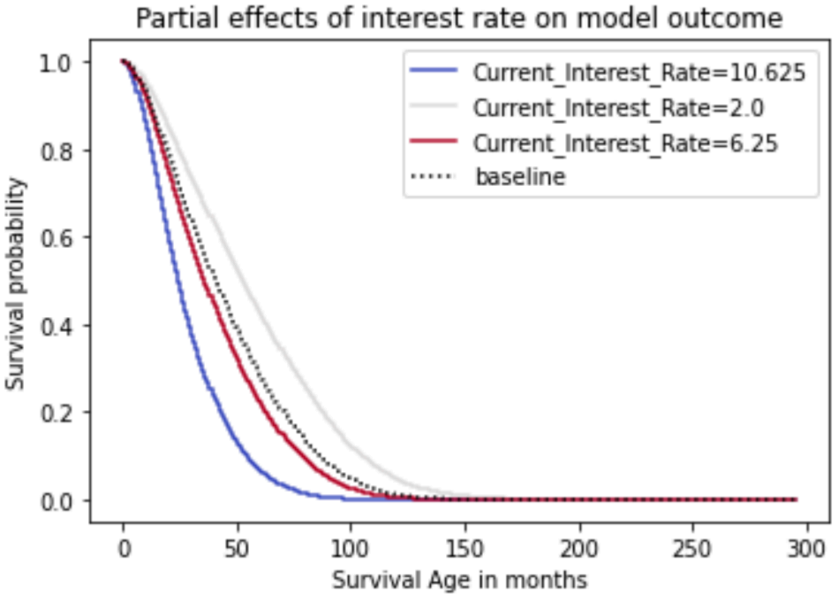
$$h(t) = h_0(t)\,\mathbf{exp}(b_1X_1 + b_2X_2 + \ldots + b_pX_p)$$

where h(t) is the expected hazard at time t, h0(t) is the baseline hazard and represents the hazard when all of the predictors (or independent variables) X1, X2 , Xp are equal to zero. Notice that the predicted hazard (i.e., h(t)), or the rate of suffering the event of interest in the next instant, is the product of the baseline hazard (h0(t)) and the exponential function of the linear combination of the predictors. Thus, the predictors have a multiplicative or proportional effect on the predicted hazard.

The survival curve graph below shows the risk (hazard ratio) as predicted by the model for the various survival ages (months). As shown, the risk is higher in the initial period of the loan.



The partial effects on the outcome graph below plots the survival probabilities over time represented by the survival age (months). The model training generated a positive coefficient for the current interest rate feature. As expected, the above 'partial effects on outcome' graph shows that having a higher interest rate leads to a lower probability of survival.



How a bank or lender will use this model?

A bank or lender can use this model to help decide whether a loan will default or not within a specific timeframe. A bank can use this model to get the survival time for a new borrower. Then they can use the survival time to predict the risk of default for a loan. The bank will have to enter 25 variables such as Original loan amount, Current Interest rate, Credit Score, type of loan, number of borrowers, type of property (single/multi family), etc. After the model is run on this loan data, running the predicted survival times function will output the survival time for this loan. This survival time is a predicted measure of the time until a default. The bank can use this predicted time to default as an additional criteria to determine if the borrower qualifies for the loan.