Mortgage loan default prediction

Rohan Mani
04.02.2024

Problem Statement Worksheet (Hypothesis Formation)

Predict risk of mortgage default for any homeowner/borrower.

H

1 Context

Since mortgage rates have gone up recently, we would like to predict the default risk of current homeowners/borrowers based on data from 1999 to 2023.

2 Criteria for success

Model performance metrics.

3 Scope of solution space

We are using mortgages lent from 1999 to 2023.

4 Constraints within solution space

Not considering all the assets of the borrower

5 Stakeholders to provide key insight

Bank/lending company and the homeowner/borrower

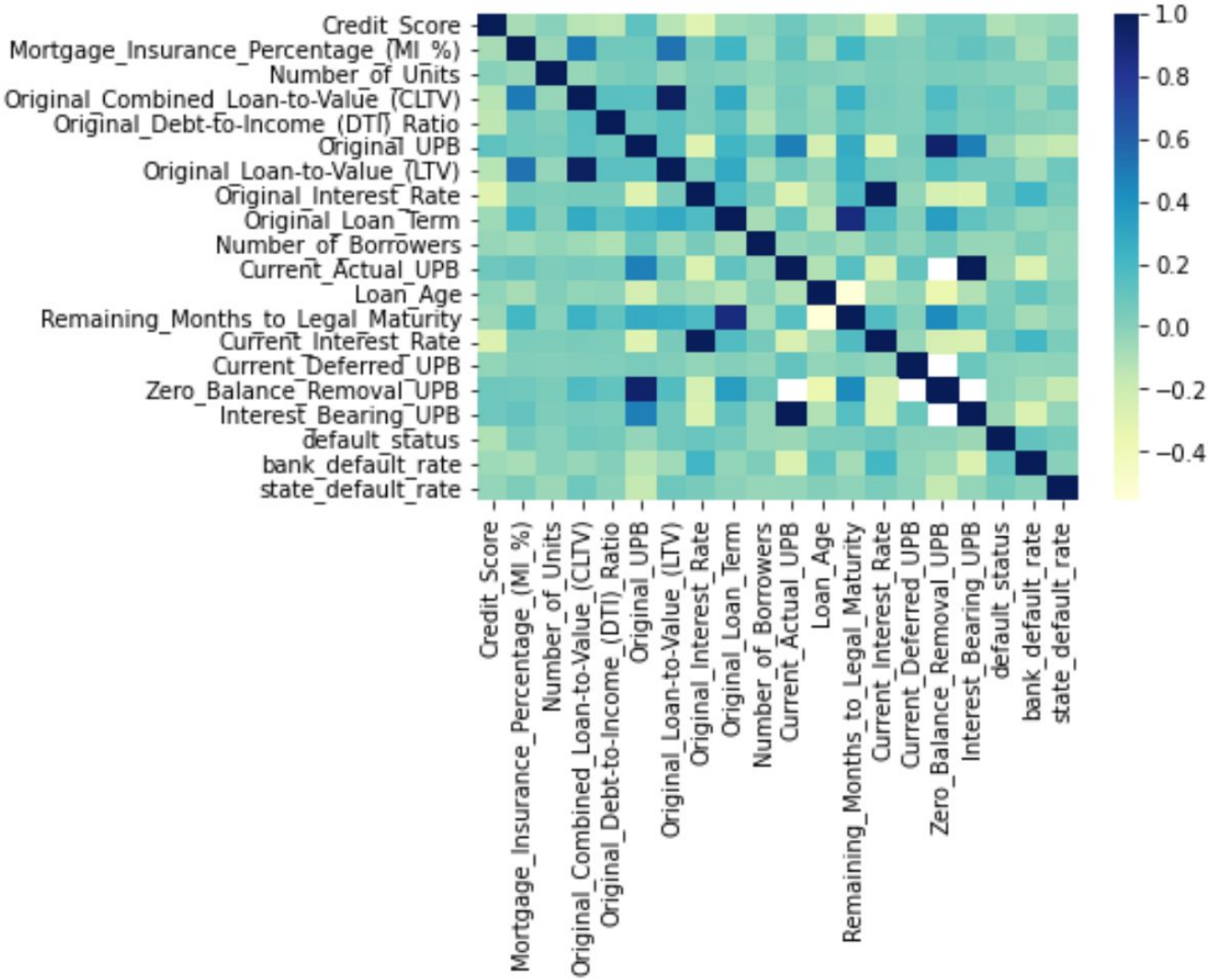
6 Key data sources

Data-https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset

Data wrangling

- Freddie Mac loan data from years 1999 to 2023.
  - 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 1,225,000 loans with 63 loan attributes.
- 29 attributes were removed since 90% of the values were missing.
- For the remaining attributes, missing values were replaced with the mean, median or mode.
- 1.3% loans (16,429/1,225,000) flagged as default in the dataset indicating the dataset is imbalanced.

Exploratory Data Analysis (Correlation heatmap)



Exploratory Data Analysis Findings

- There is correlation between variables like Original UPB, Interest Bearing UPB, Zero Balance Removal, and Current Actual UPB.
- We also have high cross correlation between CLTV, LTV, and Mortgage Insurance percentage.
- State default rates vary from 2.5% (Nevada) to below 1% (California).
- Bank/lender default rates vary from over 1.2% to below 0.8%.

Feature engineering

- Created default rates by state and bank/servicer as new feature columns to inform the model.
- 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.
- Encoded categorical columns such as First Time Homebuyer, Loan Purpose, Property Type, Channel to numeric.
- 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

Cox survival model metrics

- The loan default prediction problem is akin to a survival analysis problem, therefore the Cox survival analysis was used.
- The Concordance Index (C-index) quantifies the rank correlation between actual survival times and a model's predictions.

Model	Concordance
Cox PH	0.853

Indicates high correlation

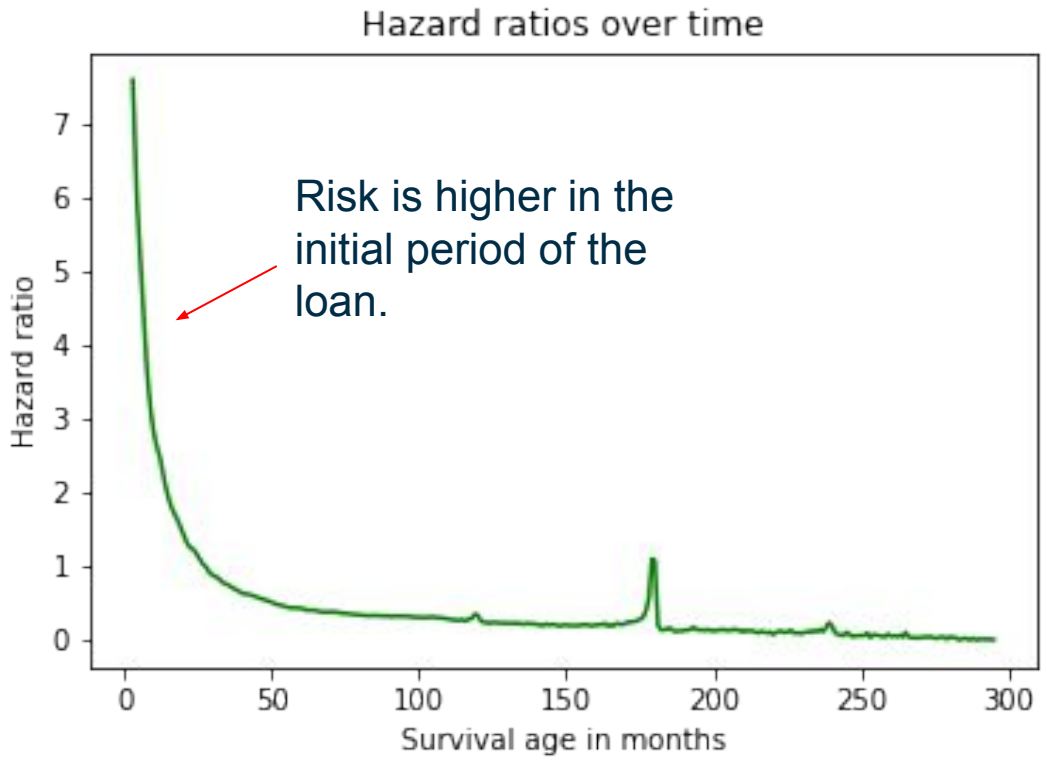
Some important and interesting Feature Coefficients

Variable/Feature	Coefficient/Weight	
state_default_rate	0.97	Highly correlated
bank_default_rate	0.34	
Channel_B (Broker)	0.14	
Current_Interest_Rate	0.13	
Channel_C (Correspondent)	0.1	
Occupancy_Status_P (Primary Home)	0.05	
MSA (Metropolitan Area)	-0.11	
HARP_Indicator (Home Affordability Refinance Program)	-0.16	Negatively correlated

Cox hazard ratios over time

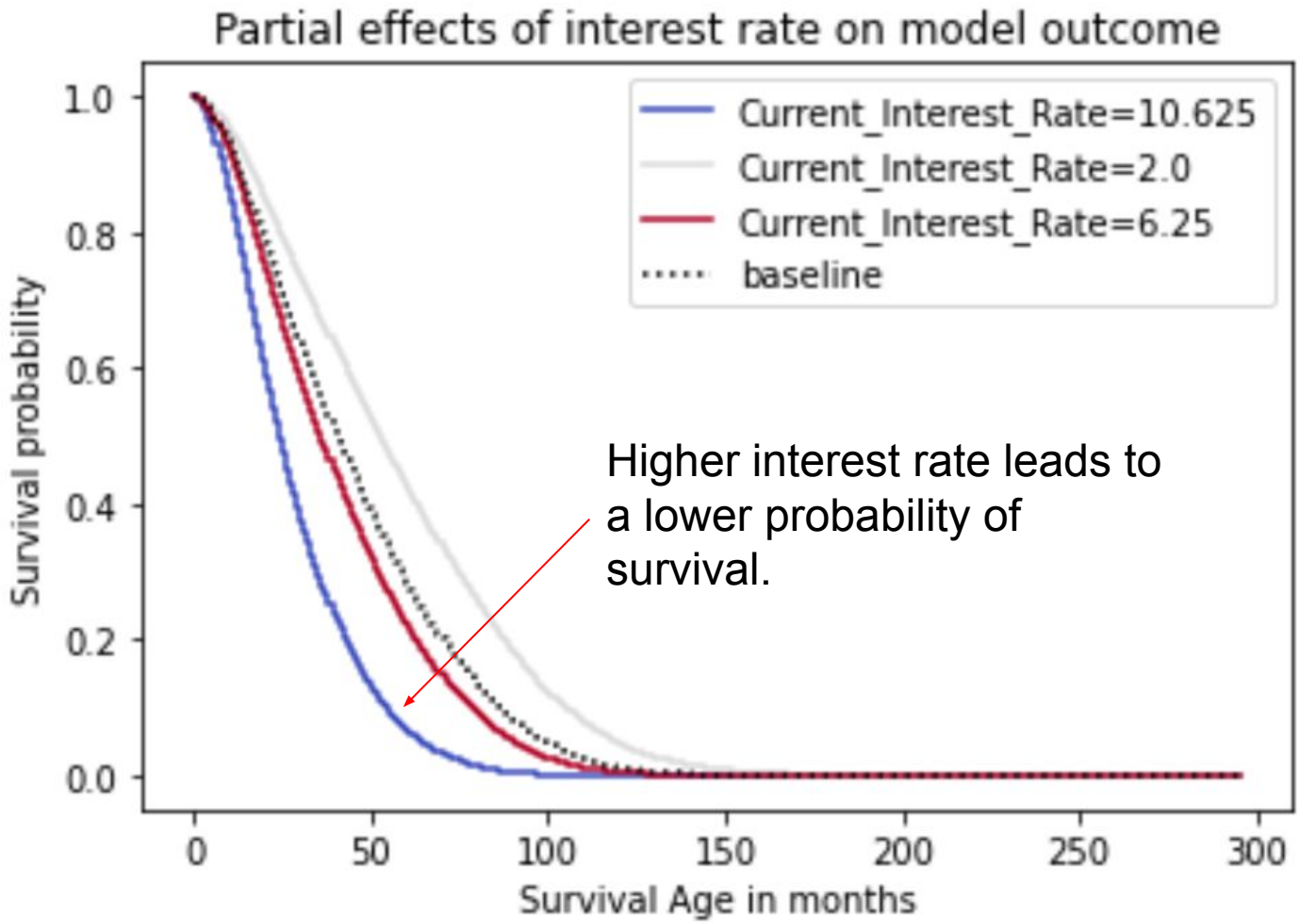
The Cox proportional hazards regression model can be written as follows

$$h(t) = h_0(t)\exp(b_1X_1 + b_2X_2 + \ldots + b_pX_p)$$



Risk (hazard ratio) as predicted by the model for the various survival ages.

Partial effects on outcome graph



How a lender will use this model?

- For each 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.