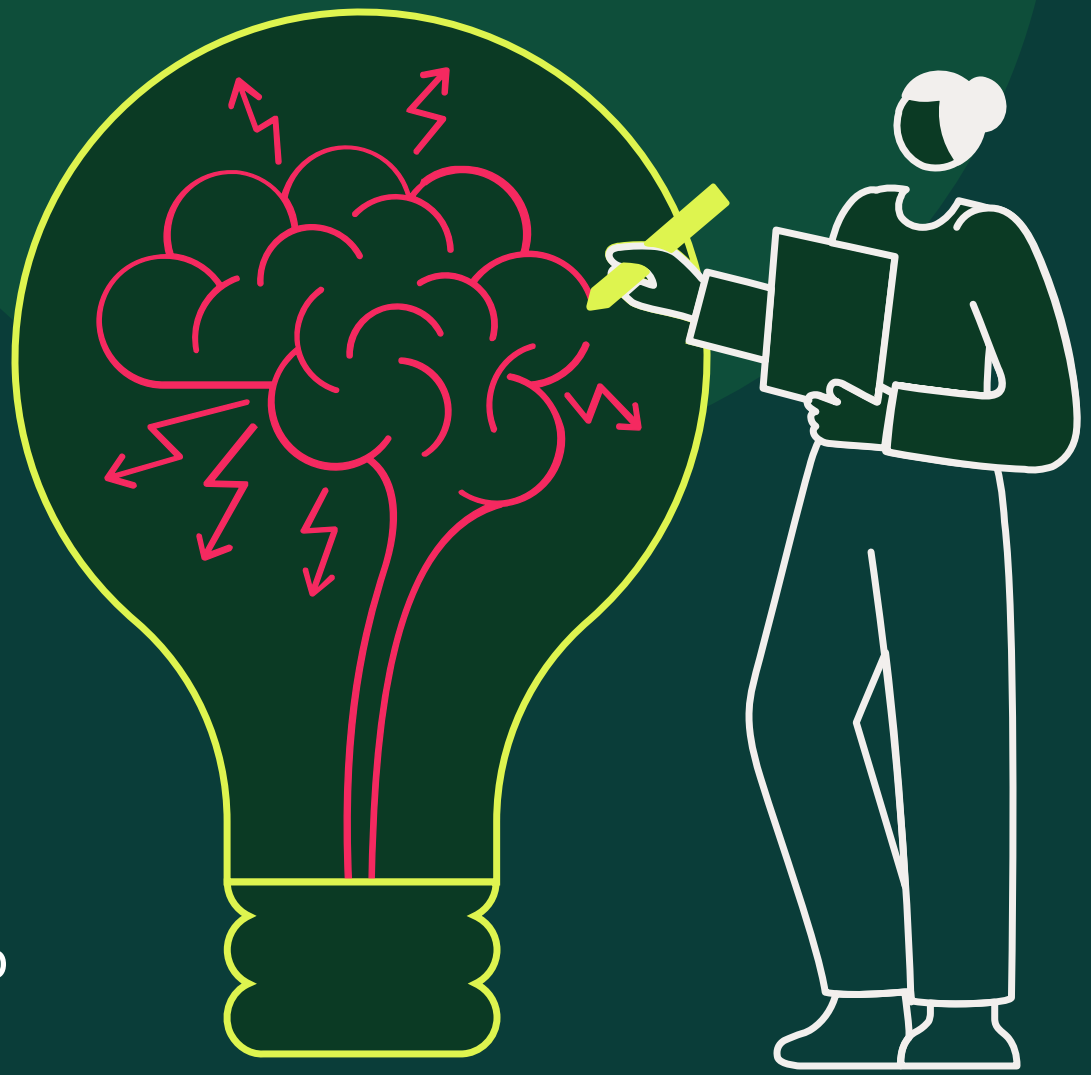
Freddie Mac
We make home possible®

Data Assessment Document



Davya Vuyyuru, Devni Shah, Chandan N A, Rishi Rao

FM_CA_30_size5

Agenda

Business – Problem and Objectives

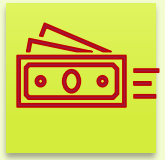
Data Source

Data - Preparation and Cleaning

Model Considerations – EDA and Feature Engineering

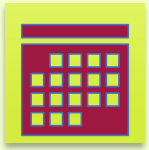
Project Success Assessment

Business Problem And Objectives



Prepayment vs. Default Risks: Freddie Mac must balance the risks of early payoffs and defaults for California-based mortgages 30 days past due, as these directly impact financial stability and require strategic monitoring to mitigate losses.

Economic Sensitivity: California's 30-days past due borrowers are highly sensitive to economic shifts, necessitating adaptable forecasting models to reflect changing regional conditions and borrower responses.



Monthly Transition Rate Tracking: Monitor and predict monthly status changes of California-based mortgages that are 30 days past due, aiming to identify trends and transitions in delinquency status over time.

Error Minimization: Minimize forecast errors by comparing predicted versus actual 30-day delinquency transition rates using root mean square error (RMSE) as the primary metric.

Period Analysis: Assess the predictive model's performance by evaluating transition rate trends across pre-COVID, COVID, and post-COVID periods, specifically focusing on 30-days past due mortgages in California, to capture the impact of economic shifts during these phases.

California_30 - Housing Market Dynamics

High Volatility in Prices

California's housing prices are highly volatile, especially in cities like Los Angeles and San Francisco. Fluctuations can impact borrowers' ability to pay, increasing default risk if property values fall below the loan balance.

Foreclosure Protections

The state has strong foreclosure protection laws that may delay the transition to default. California was also one of the states with the highest levels of mortgage forbearance requests during the peak of the pandemic.

30 days - Early Indicator

Due to high living costs, many borrowers may begin to experience financial distress earlier in the delinquency process. A "30 days past due" status can thus be a critical early warning sign.

Data Source - Freddie Mac's Loan-Level Dataset

ORIGINATION DATA FILE	
Index	**Column Name**
1	Year
2	Credit Score
3	First Payment Date
4	First Time Homebuyer Flag
5	Maturity Date
6	Metropolitan Statistical Area (MSA) Or Metropolitan Division
7	Mortgage Insurance Percentage (MI %)
8	Number of Units
9	Occupancy Status
10	Original Combined Loan-to-Value (CLTV)
11	Original Debt-to-Income (DTI) Ratio
12	Original UPB
13	Original Loan-to-Value (LTV)
14	Original Interest Rate
15	Channel
16	Prepayment Penalty Mortgage (PPM) Flag
17	Amortization Type (Formerly Product Type)
18	Property State
19	Property Type
20	Postal Code
21	Loan Sequence Number
22	Loan Purpose
23	Original Loan Term
24	Number of Borrowers
25	Seller Name
26	Servicer Name
27	Super Conforming Flag
28	Pre-HARP Loan Sequence Number
29	Program Indicator
30	HARP Indicator
31	Property Valuation Method
32	Interest Only (I/O) Indicator
33	Mortgage Insurance Cancellation Indicator

Origination Data File:

- 1 Million+ rows and 33 Columns.
- Geographic Data: Grouped by fields such as "Metropolitan Statistical Area (MSA)" and "Postal Code
- Borrower Characteristics: Includes "Credit Score," "Debt-to-Income (DTI) Ratio," and "First Time Homebuyer Flag"
- Loan Characteristics: "Loan Purpose," "Original Combined Loan-to-Value (CLTV)," "Interest Only (I/O) Indicator"

Data Source - Freddie Mac's Loan-Level Dataset

MONTHLY PERFORMANCE DATA FILE	
Index	**Column Name**
1	Year
2	Loan Sequence Number
3	Monthly Reporting Period
4	Current Actual UPB
5	Current Loan Delinquency Status
6	Loan Age
7	Remaining Months to Legal Maturity
8	Defect Settlement Date
9	Modification Flag
10	Zero Balance Code
11	Zero Balance Effective Date
12	Current Interest Rate
13	Current Deferred UPB
14	Due Date of Last Paid Installment (DDLPI)
15	MI Recoveries
16	Net Sales Proceeds
17	Non MI Recoveries
18	Expenses
19	Legal Costs
20	Maintenance and Preservation Costs
21	Taxes and Insurance
22	Miscellaneous Expenses
23	Actual Loss Calculation
24	Modification Cost
25	Step Modification Flag
26	Deferred Payment Plan
27	Estimated Loan-to-Value (ELTV)
28	Zero Balance Removal UPB
29	Delinquent Accrued Interest
30	Delinquency Due to Disaster
31	Borrower Assistance Status Code
32	Current Month Modification Cost
33	Interest Bearing UPB

Monthly Performance Data:

- 7 million+ rows and 33 columns.
- Tracks detailed loan performance metrics over time for Freddie Mac's mortgage portfolio, focusing on delinquency, prepayment, and default statuses.

Additional Data Sources:

Freddie Mac House Price index data

FRED - Federal Reserve Economic Data

Google Trends

Data Preparation

1. Combined origination data from each year (1999-2018) into a single file with 1 million rows and 33 columns.

2. Filtered this to keep only California (CA) data, reducing it down to around 120,000 rows.

3. Merged monthly data files from 1999-2024, creating a huge dataset with millions of rows and added a Year column.

4. Matched the Loan Sequence Number in R to filter the monthly dataset for CA, ending up with over 7 million rows and 33 columns.

5. Now, both origination and monthly datasets are CA-specific, ready for analysis.

Data Cleaning

Performed

1. Standardized data for consistency by replacing out-of-range values and placeholder codes with NA or designated placeholders across relevant columns in Origination File.

2. Reformatted Monthly Reporting Period to a readable date format and mapped Zero Balance Code to descriptive labels for improved clarity in analysis and visualization.

3. Isolated 30-day delinquent loans to examine key metrics, converted necessary columns to numeric, and identified missing values in Credit Score and DTI Ratio.

4. Merge external economic indicators like unemployment rate, GDP growth, or house price index data (especially for California) to the dataset.

5. Create a flag for loans with high Debt-to-Income, low Credit Score, or high LTV to help the model focus on loans more likely to default.

• • •

Planned

Model Considerations and Feature Engineering

The model must forecast mortgage portfolio transition rates, focusing on how many loans move from "current" to "30 days past due" status, with predictions compared to actual transitions. The goal is to minimize the RMSE between predicted and actual transition rates over time.

Logistic Regression for Binary Classification

- *Why*: Simple, interpretable model for predicting 30-day past-due loans.
- *Consideration*: Estimates probabilities, useful for error forecasting

Random Forest for Non-linear Relationships

- *Why*: Captures non-linear relationships and interactions between features
- *Consideration*: Handles mixed data types and offers feature importance.

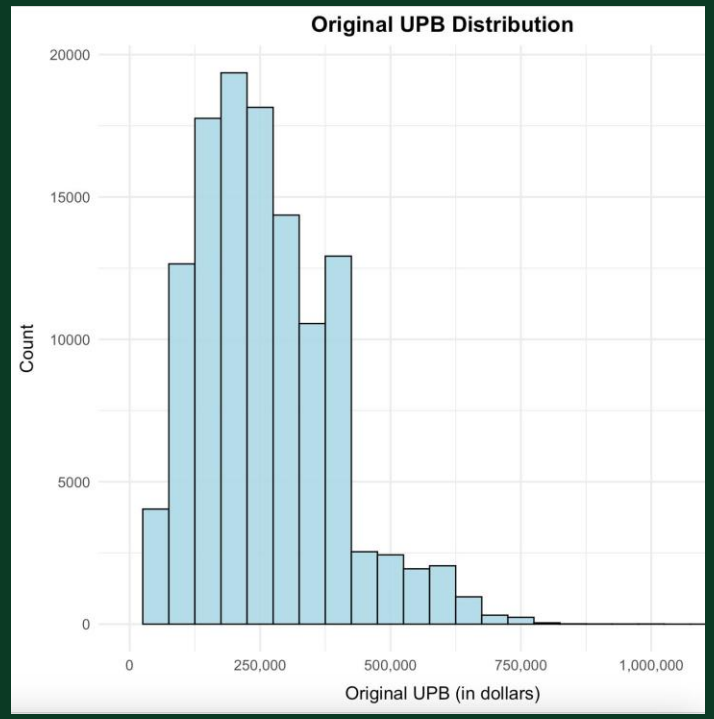
Gradient Boosting (XGBoost) for High Accuracy

- *Why*: Strong predictive performance, handles missing data.
- *Consideration*: More complex but highly accurate.

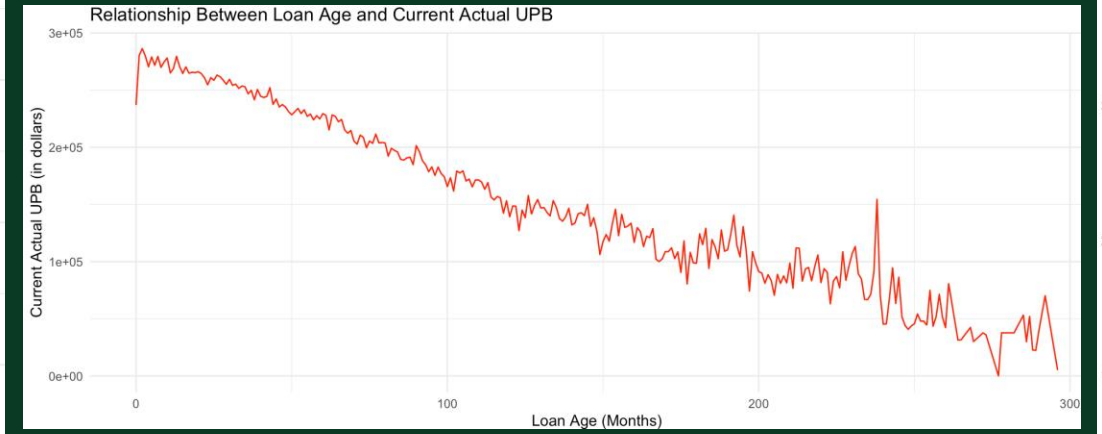
Cross-Validation for Evaluation

- *Why*: Improves model generalizability and stability.
- *Consideration*: Useful for all models, minimizes overfitting.

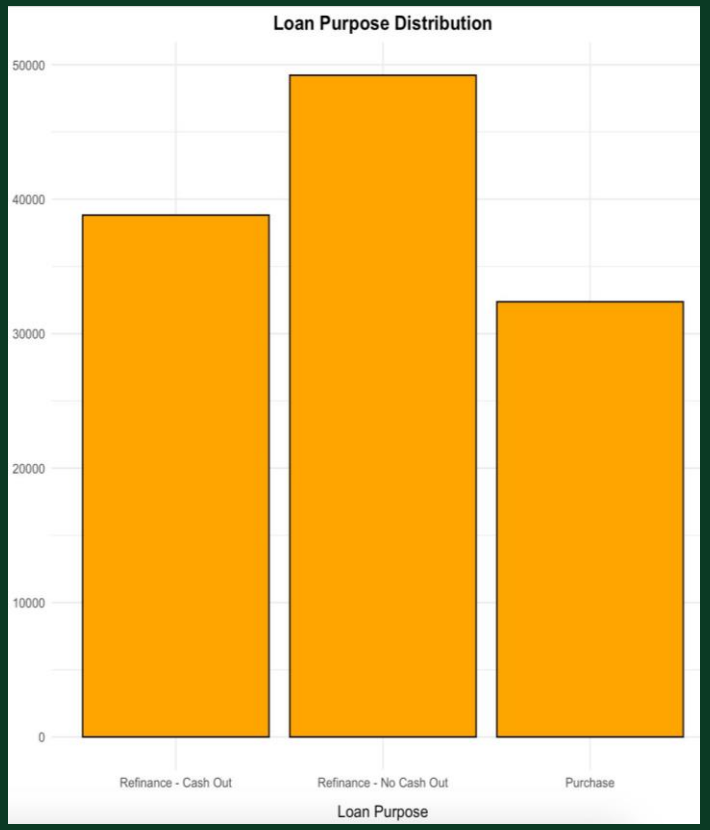
Exploratory Data Analysis



Most loans have an Original UPB between $100,000 and $400,000, with fewer higher-balance loans.

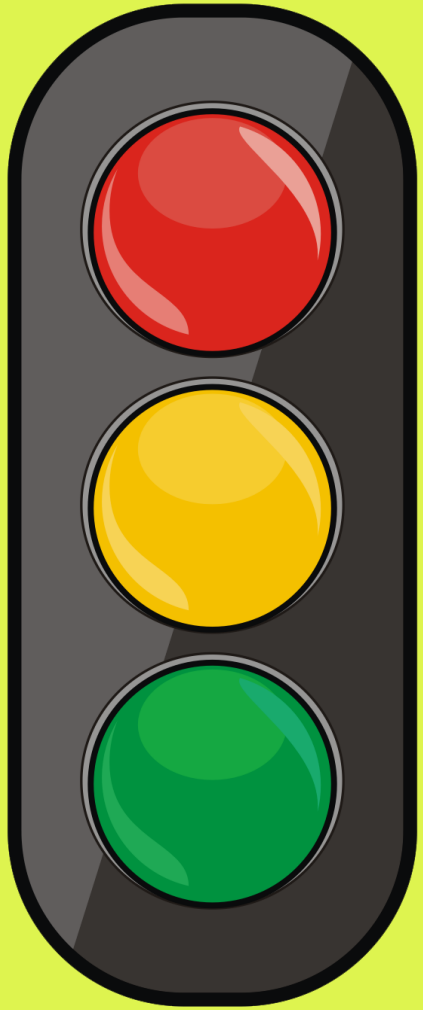


For 30-days past due loans, as loan age increases, the unpaid balance generally decreases, reflecting gradual repayment despite delinquencies.



Most borrowers prioritize refinancing without cashing out, indicating a focus on debt restructuring over equity extraction

Project Success Assessment



- Managing and integrating complex, vast data sources is challenging.

- Model considerations
- Feature engineering
- EDA

- Business problem and objectives
- Data cleansing and preparation

Thank You