TRANSITIONS IN FOCUS: DELINQUENCY TO RESOLUTION IN CALIFORNIA

FM_CA_30_Size_4

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

MEET THE TEAM

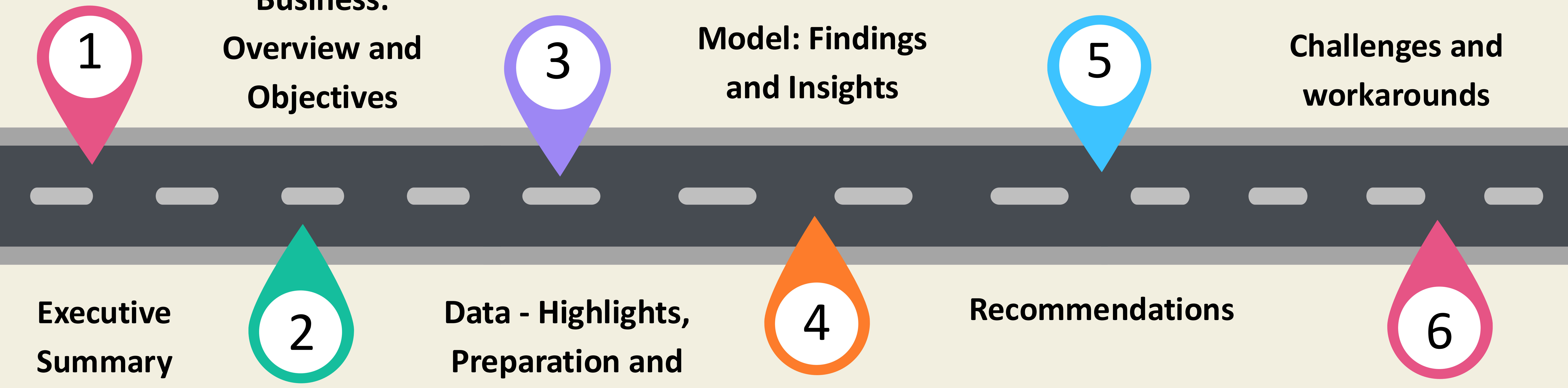
Rishi Rao

Devni Shah

Davya Vuyyuru

Chandan N A

ROADMAP



1 Executive Summary

2 Business: Overview and Objectives

3 Data - Highlights, Preparation and Feature Engineering

4 Model: Findings and Insights

5 Recommendations

6 Challenges and workarounds

Executive Summary

Executive Summary

BACKGROUND & OBJECTIVE

- This project aims to predict and assess the progression of **30-day delinquent** mortgage loans within the secondary market. The study is vital to Freddie Mac's mission of supporting liquidity and stability in the U.S. housing sector.

CORE STRATEGIES

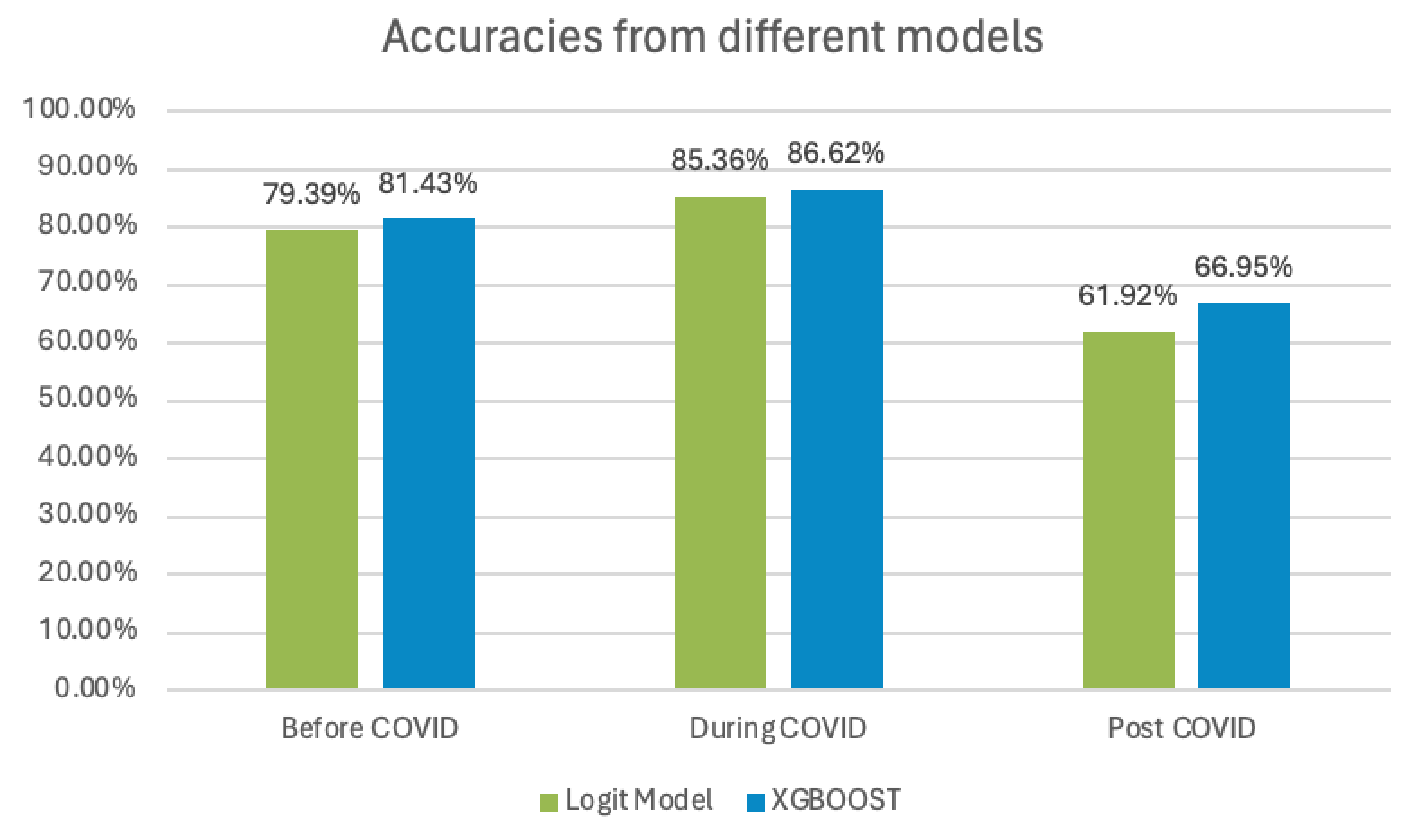
- Focus on analyzing delinquency transitions in the **California mortgage market**.
- Employ advanced machine learning models to predict loan performance and trends.
- Integrate external economic data to improve predictions.

METHODS

- Cleaned and standardized monthly performance data from Freddie Mac's loan-level dataset.
- Selected key predictive variables such as **Loan Age, Interest Rate, Credit Score, Delinquency Due to Disaster, and Current Actual UPB.**
- Combined internal loan data with external economic indicators for context.

TECHNIQUES & EVALUATION

- Multinomial Logistic Regression
- XGBoost
- Evaluate model on 3 time periods: Pre-COVID, COVID, and Post-COVID
- Accuracy + RMSE



RMSE - XGBoost	
Pre COVID	7.58%
During COVID	1.47%
Post COVID	4.74%

Business-Overview and Objectives

OVERVIEW

Freddie Mac helps make housing affordable by buying loans from banks, giving them more money to lend. It bundles these loans into investments sold to investors, keeping the housing market stable and accessible for buyers and renters.

BUSINESS IMPACT

- Prepayments reduce long-term interest revenue.
- Defaults lead to unrecoverable losses and high foreclosure costs.

REGIONAL MARKET CHALLENGES

- California’s home prices are unpredictable due to economic factors like employment fluctuations and interest rate changes, as well as natural risks such as wildfires and earthquakes.

ECONOMICS

- Borrower behavior may shift with economic changes, making accurate forecasting essential for managing risks

OBJECTIVES

Prediction of Mortgage Transitions

Develop predictive models to analyze how loans transition from 30 days past due to other statuses and evaluate the model's performance across three distinct time periods to assess its effectiveness.

California Market Focus

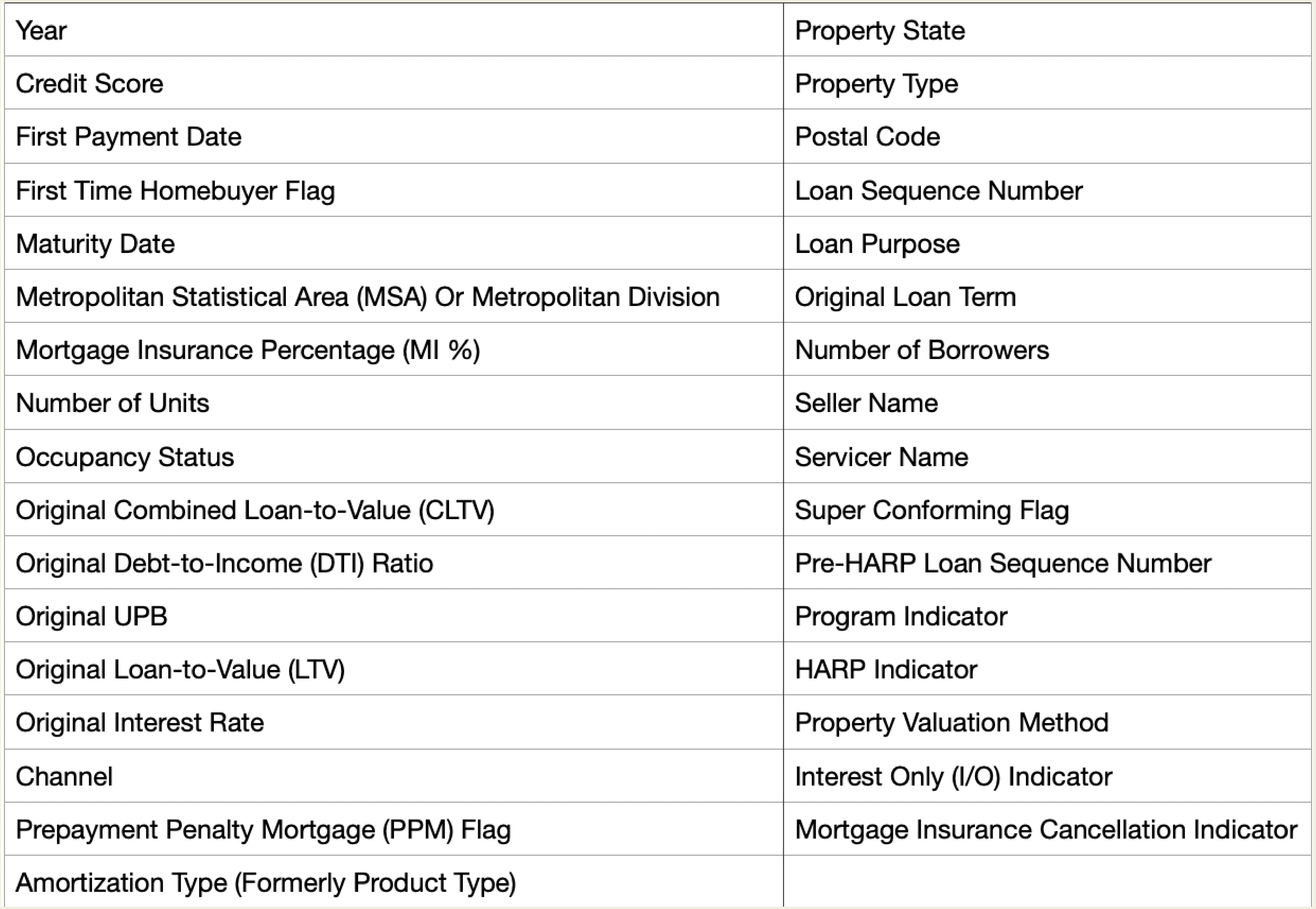
Compare trends and predictions across 3 distinct periods to help understand the disruption on borrower behavior.

Minimize Forecasting Errors

Reduce discrepancies between predicted and actual mortgage transition rates, measured through root mean square error, to enhance model precision.

Data - Highlights, Preparation and Feature Engineering

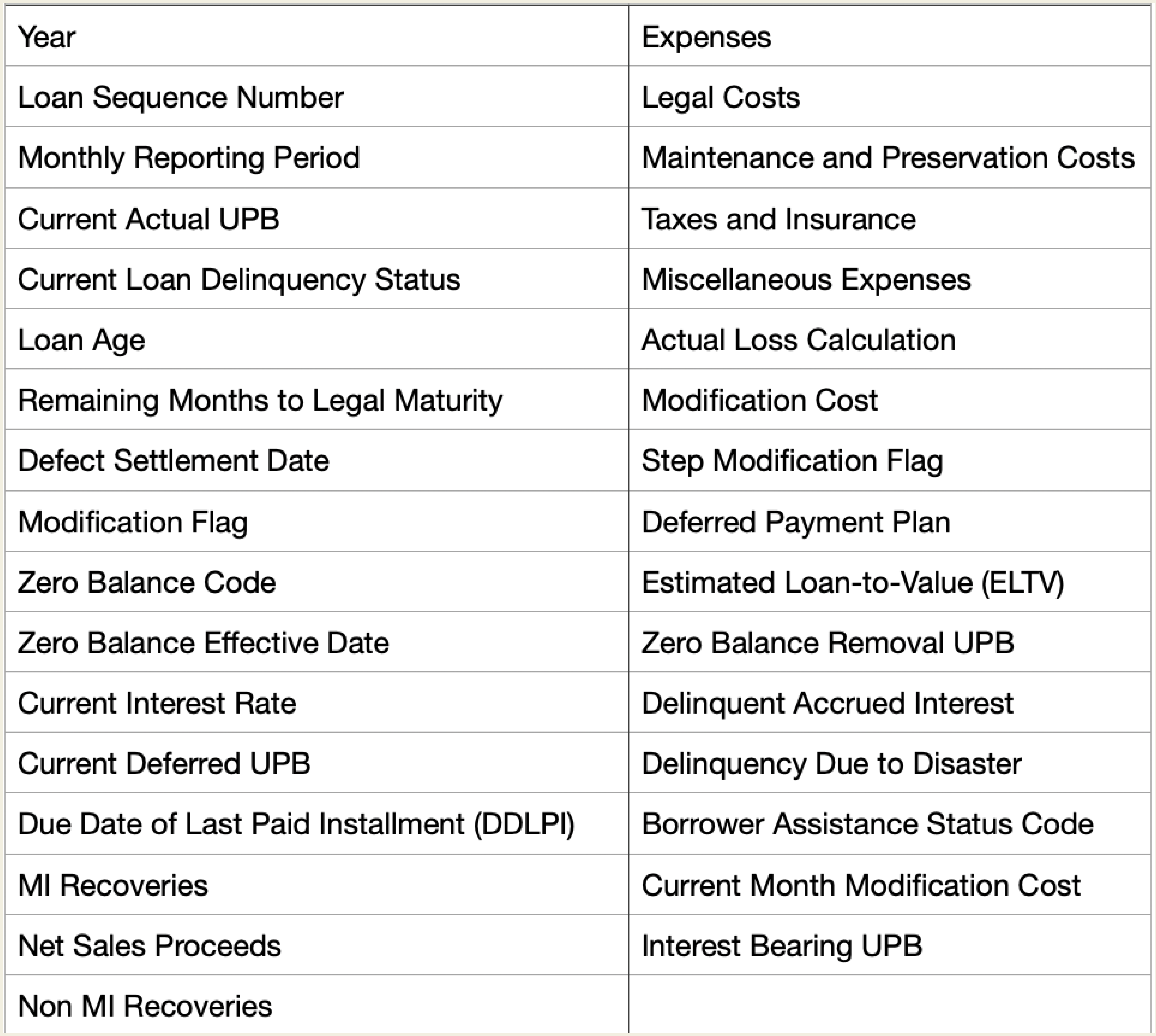
Origination Data File



Year	Property State
Credit Score	Property Type
First Payment Date	Postal Code
First Time Homebuyer Flag	Loan Sequence Number
Maturity Date	Loan Purpose
Metropolitan Statistical Area (MSA) Or Metropolitan Division	Original Loan Term
Mortgage Insurance Percentage (MI %)	Number of Borrowers
Number of Units	Seller Name
Occupancy Status	Servicer Name
Original Combined Loan-to-Value (CLTV)	Super Conforming Flag
Original Debt-to-Income (DTI) Ratio	Pre-HARP Loan Sequence Number
Original UPB	Program Indicator
Original Loan-to-Value (LTV)	HARP Indicator
Original Interest Rate	Property Valuation Method
Channel	Interest Only (I/O) Indicator
Prepayment Penalty Mortgage (PPM) Flag	Mortgage Insurance Cancellation Indicator
Amortization Type (Formerly Product Type)	

- Loan-level details captured at the time of a mortgage's origination.
- 1,215,024 loans across the U.S, 33 columns.
- **143,276 loans** originated in California.

Monthly Performance Data



- Year
- Loan Sequence Number
- Monthly Reporting Period
- Current Actual UPB
- Current Loan Delinquency Status
- Loan Age
- Remaining Months to Legal Maturity
- Defect Settlement Date
- Modification Flag
- Zero Balance Code
- Zero Balance Effective Date
- Current Interest Rate
- Current Deferred UPB
- Due Date of Last Paid Installment (DDLPI)
- MI Recoveries
- Net Sales Proceeds
- Non MI Recoveries
- Expenses
- Legal Costs
- Maintenance and Preservation Costs
- Taxes and Insurance
- Miscellaneous Expenses
- Actual Loss Calculation
- Modification Cost
- Step Modification Flag
- Deferred Payment Plan
- Estimated Loan-to-Value (ELTV)
- Zero Balance Removal UPB
- Delinquent Accrued Interest
- Delinquency Due to Disaster
- Borrower Assistance Status Code
- Current Month Modification Cost
- Interest Bearing UPB

- Tracks loan performance month-over-month from origination to termination.
- **7,436,931 rows** for California, 33 columns.
- Captures transitions across statuses.
- Identify high-risk loans based on payment behavior trends.

External Data Sources

US GDP

CALIFORNIA MONTHLY UNEMPLOYMENT RATE

HOUSE PRICE INDEX (HPI)

DATA CLEANING AND PREPARATION

Inner Join for Unified Dataset: Combined datasets using Loan Sequence Number to create a comprehensive, unified dataset.

Data Cleaning and Preprocessing: Created missing value flags, replaced invalid or missing codes (e.g., 999, NA) with NA, converted columns to appropriate data types (e.g., numeric, factor), and resolved inconsistencies like blank spaces.

Delinquency Categorization: Categorized loans into delinquency groups (0, 1, 2, 90+, etc.) based on the Current Loan Delinquency Status. Created lagged data by arranging and grouping data by Loan Sequence Number and reporting period to track transitions over time.

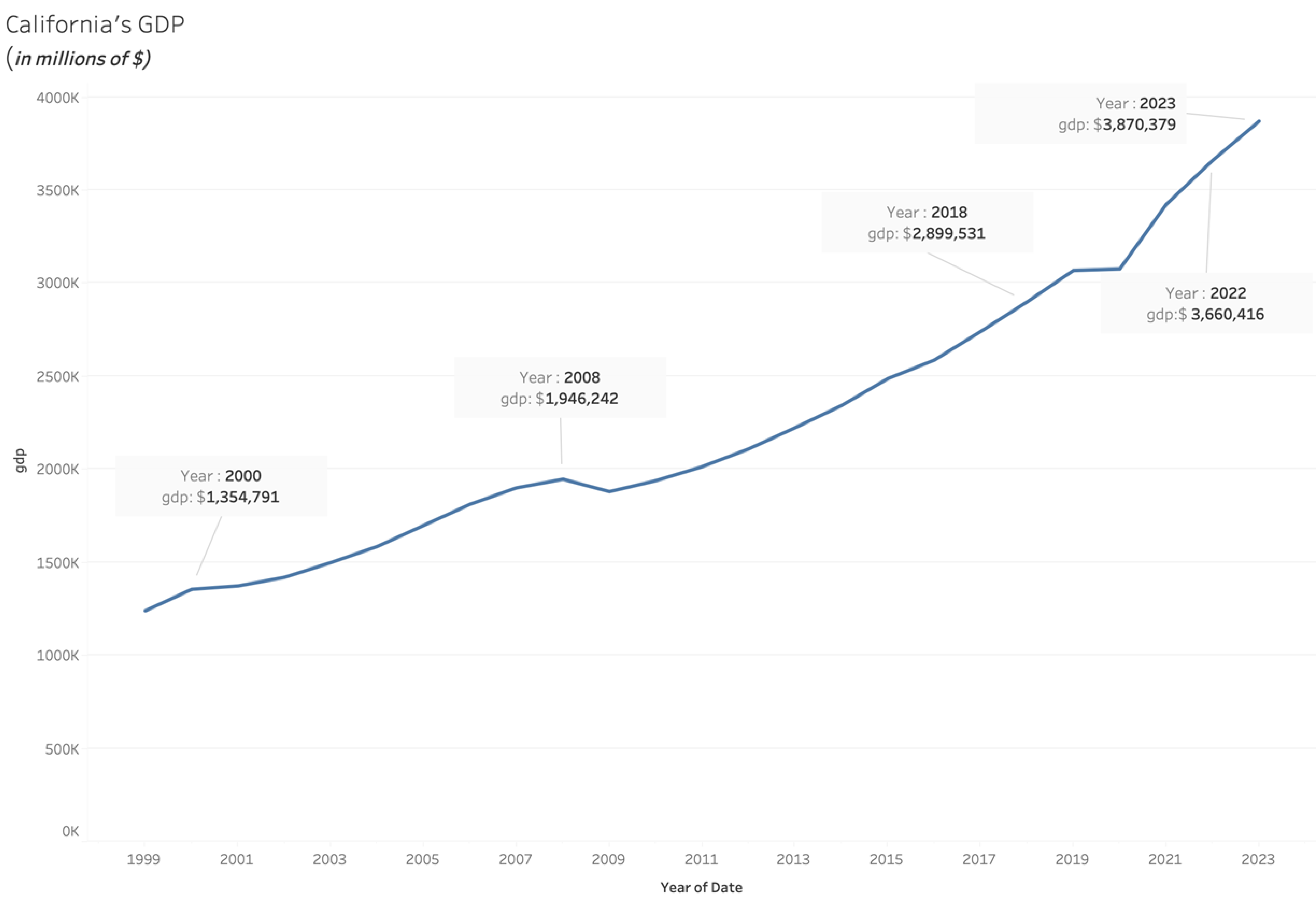
Focus on 30-Day Transitions: Analyzed transitions starting from "30 days past due," summarizing types of transitions (e.g., 1 → 0, 1 → 90+), and identifying monthly transition patterns to uncover trends.

Streamlining Dataset: Removed irrelevant columns with significant NA values to enhance processing efficiency and streamline the dataset for analysis.

External Data Integration: Incorporated external data sources, including GDP and unemployment rates, to analyze the impact of macroeconomic factors on loan performance.

Data Splitting for Analysis: Split the dataset into training and validation sets, further dividing validation data into pre-COVID, COVID, and post-COVID periods for detailed testing and analysis under varying economic conditions.

California's GDP Growth: 1999-2023



- GDP grew from $1,355B (2000) to $3,870B (2023), nearly **tripling over 20 years.**

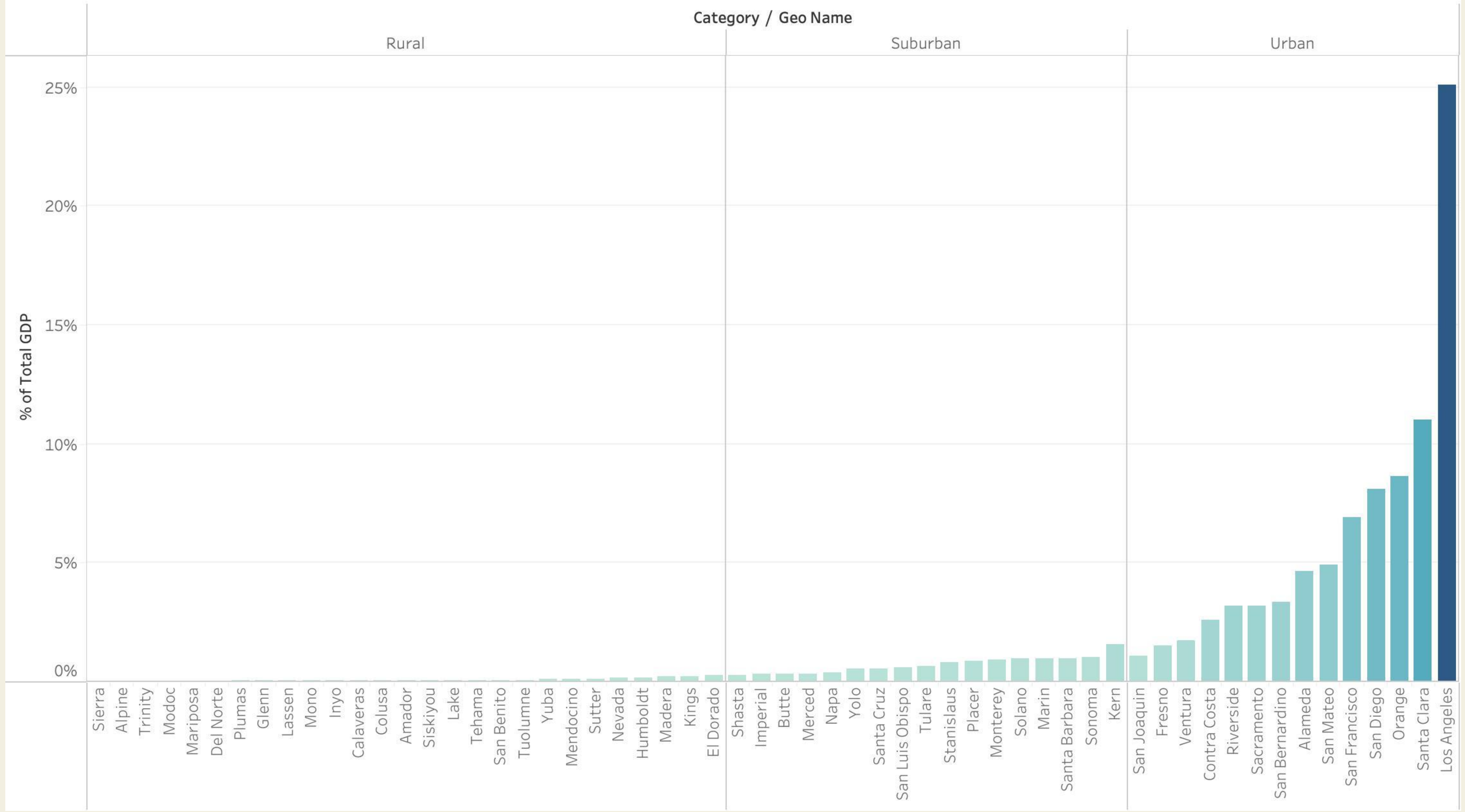
- **Slowed to $1,946B in 2008**, showing resilience during financial turmoil.

- A surge from $2,022B to $3,463B during 2010-2010, a **71% increase**.

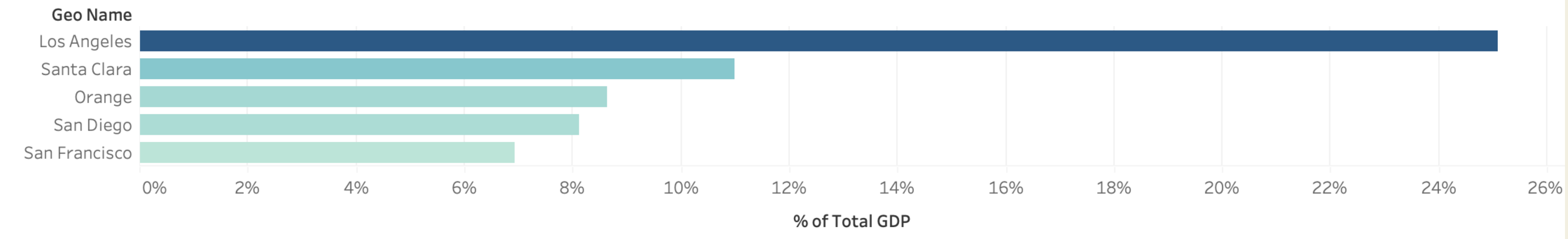
- COVID-19 Impact (2020): **Growth slowed**, but GDP rebounded strongly post-pandemic, reaching $3,870B by 2023.

Urban Centers Drive A Disproportionate Share of State GDP

GDP Contribution by Category

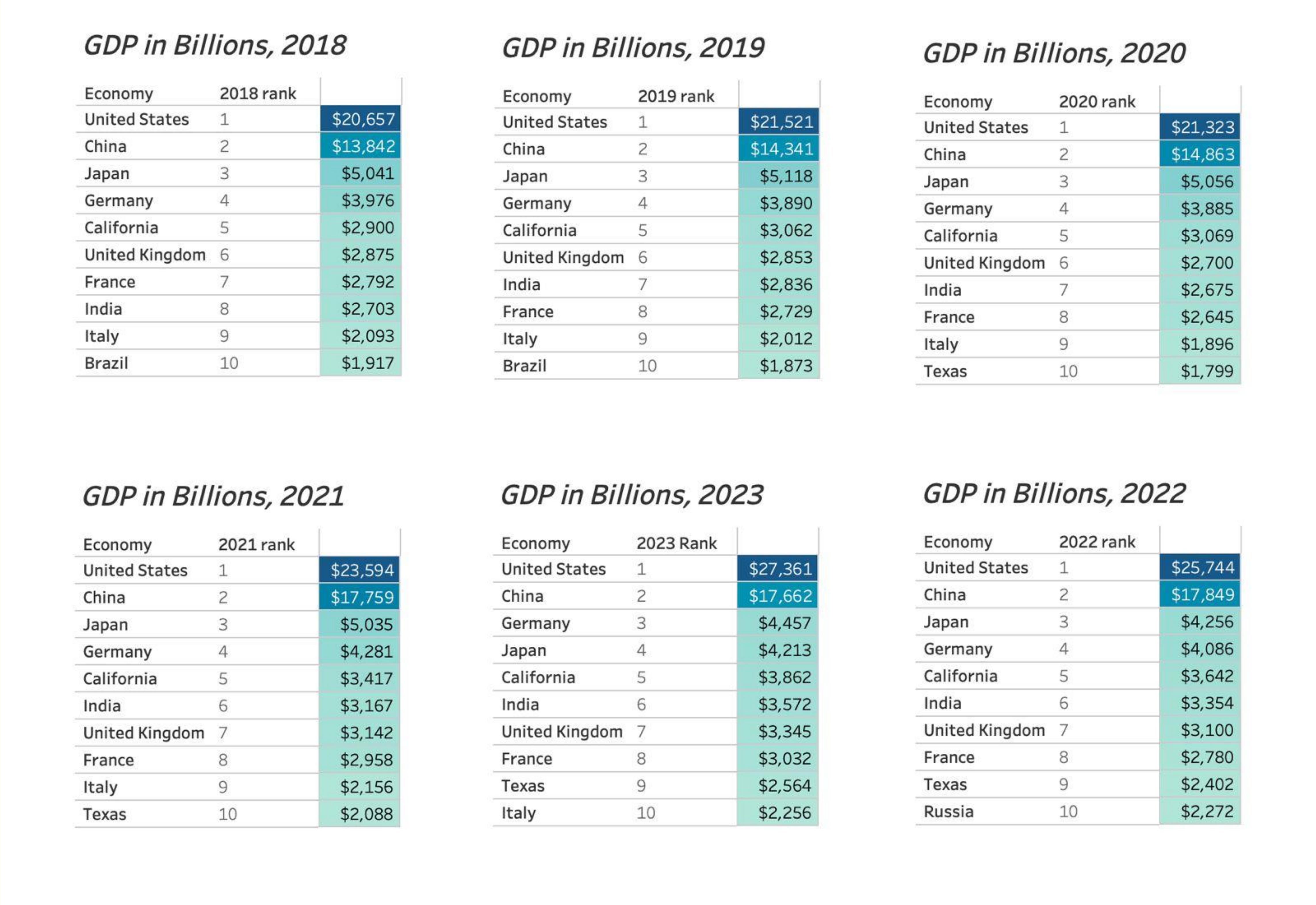


The Top 5 Counties



- **Los Angeles** contributes ~20% of California's GDP, the highest among all counties.
- **Urban counties** dominate GDP contributions, with ~70% of total GDP.
- **Suburban counties** contribute moderately, with notable contributors like **Placer and Napa.**
- **Rural counties** collectively contribute <5%.
- **Economic disparity is evident**, with rural counties lagging far behind urban centers.

California Vs The World



GDP in Billions, 2018

Economy	2018 rank	
United States	1	$20,657
China	2	$13,842
Japan	3	$5,041
Germany	4	$3,976
California	5	$2,900
United Kingdom	6	$2,875
France	7	$2,792
India	8	$2,703
Italy	9	$2,093
Brazil	10	$1,917

GDP in Billions, 2019

Economy	2019 rank	
United States	1	$21,521
China	2	$14,341
Japan	3	$5,118
Germany	4	$3,890
California	5	$3,062
United Kingdom	6	$2,853
India	7	$2,836
France	8	$2,729
Italy	9	$2,012
Brazil	10	$1,873

GDP in Billions, 2020

Economy	2020 rank	
United States	1	$21,323
China	2	$14,863
Japan	3	$5,056
Germany	4	$3,885
California	5	$3,069
United Kingdom	6	$2,700
India	7	$2,675
France	8	$2,645
Italy	9	$1,896
Texas	10	$1,799

GDP in Billions, 2021

Economy	2021 rank	
United States	1	$23,594
China	2	$17,759
Japan	3	$5,035
Germany	4	$4,281
California	5	$3,417
India	6	$3,167
United Kingdom	7	$3,142
France	8	$2,958
Italy	9	$2,156
Texas	10	$2,088

GDP in Billions, 2023

Economy	2023 Rank	
United States	1	$27,361
China	2	$17,662
Germany	3	$4,457
Japan	4	$4,213
California	5	$3,862
India	6	$3,572
United Kingdom	7	$3,345
France	8	$3,032
Texas	9	$2,564
Italy	10	$2,256

GDP in Billions, 2022

Economy	2022 rank	
United States	1	$25,744
China	2	$17,849
Japan	3	$4,256
Germany	4	$4,086
California	5	$3,642
India	6	$3,354
United Kingdom	7	$3,100
France	8	$2,780
Texas	9	$2,402
Russia	10	$2,272

- **Global Rank**: 5th largest economy from **2018–2023**.

- **GDP Growth**: Rose from **$2,900B (2018)** to **$3,862B (2023)**.

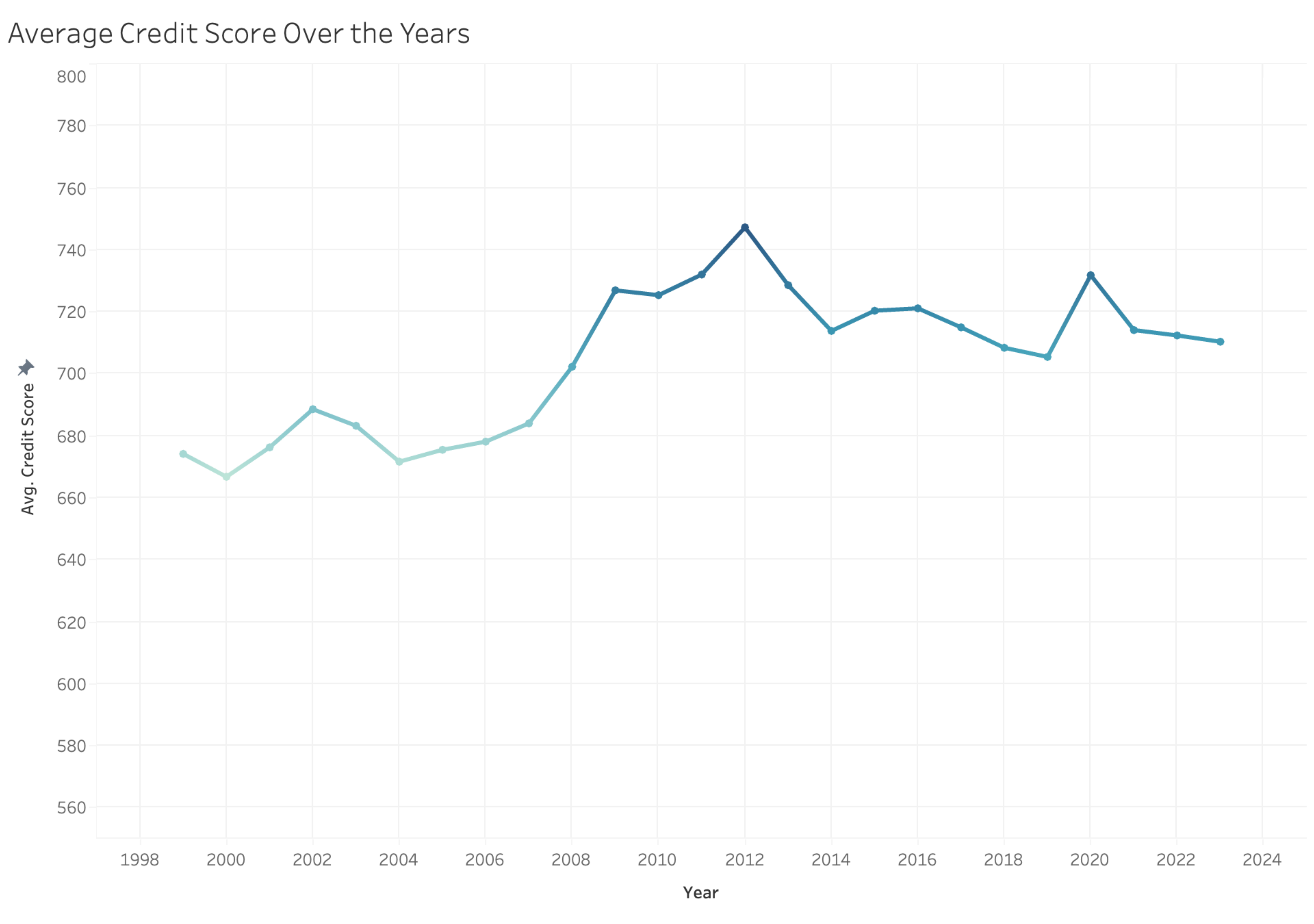
- Consistently ahead of **UK, India, France**, and **Italy**.

- **2023 Growth**: Increased by **6.1%** to reach **$3,862B**, nearing **Germany** and **Japan**

- Critical to the U.S.’s global **#1 ranking**.

- Leads in **technology, entertainment**, and **renewables**.

Credit scores are influenced by major economic events like the 2008 crisis and covid-19



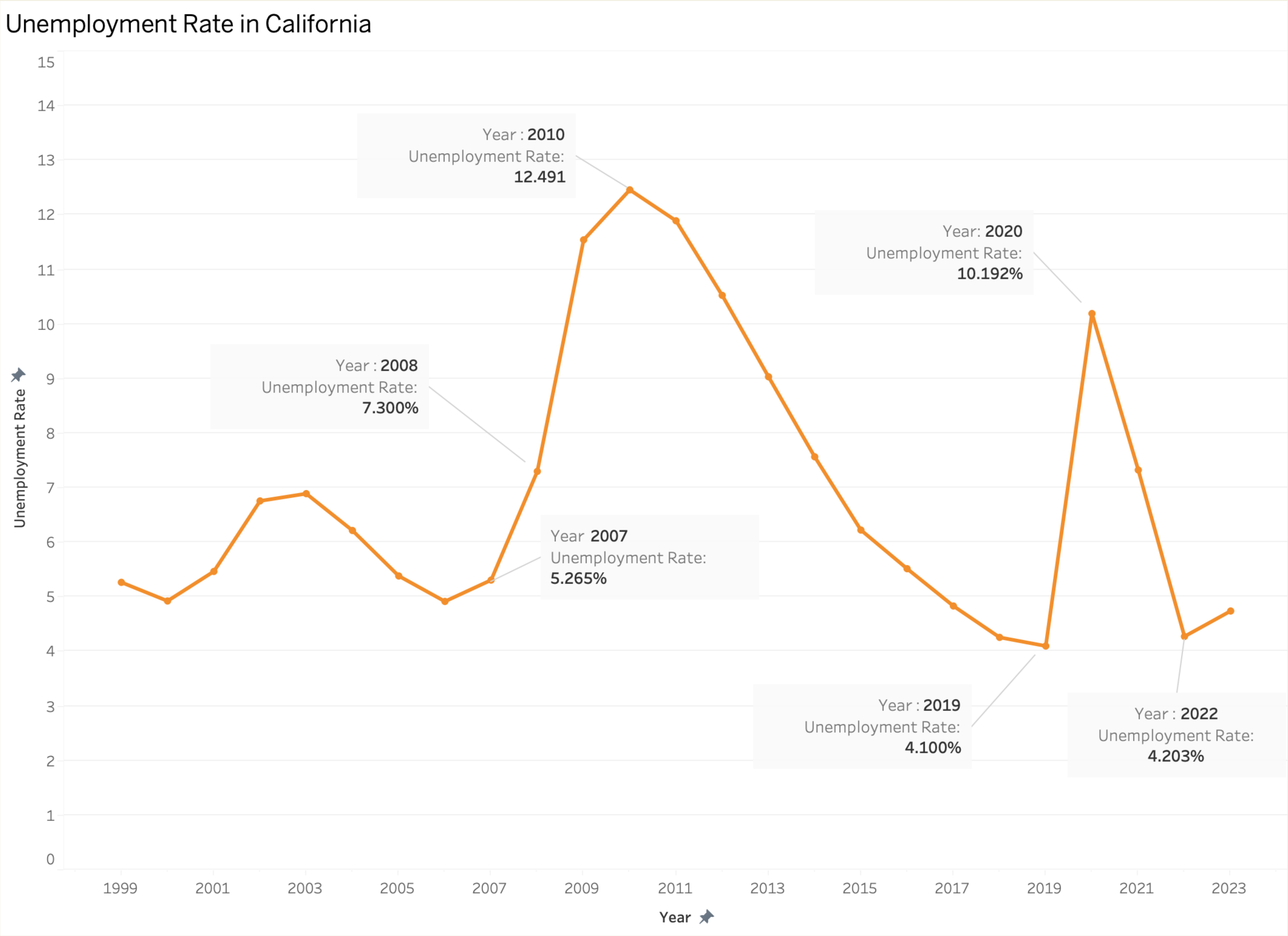
- **Steady Growth Before and After 2008**, suggesting other factors in play rather than direct crisis impact.

- A **small drop** is seen in 2008, aligning with the immediate financial crisis effects, but the recovery is rapid.

- Shaky in the 2010's decade.

- Increase during the financial crisis and increase in 2020 suggest that the two major events of the decade did not have an adverse effect.

Unemployment spikes during crises and recovers in stability


Unemployment Rate in California

- Unemployment rose sharply from 5.27% in 2007 to 12.49% in the aftermath of the 2008 financial crisis.
- A **small drop** is seen in 2008, aligning with the immediate financial crisis effects, but the recovery is rapid.
- Periods of economic stability, like 2007 and 2019, show low unemployment rates of around 4–5%.
- Strong link between economic crises and unemployment surges.

Model Description

Models Used for Loan Delinquency Prediction

Multinomial LOGISTIC MODEL

XGBOOST

MODEL 1 : MULTINOMIAL LOGISTIC MODEL

Multinomial Logistic Regression:

- Used for multiclass classification of loan delinquency status.
- Easy to implement and understand

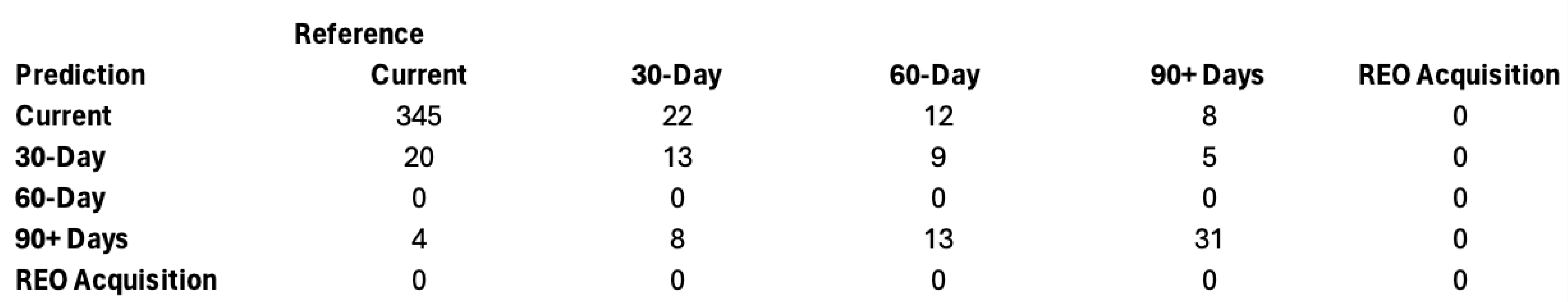
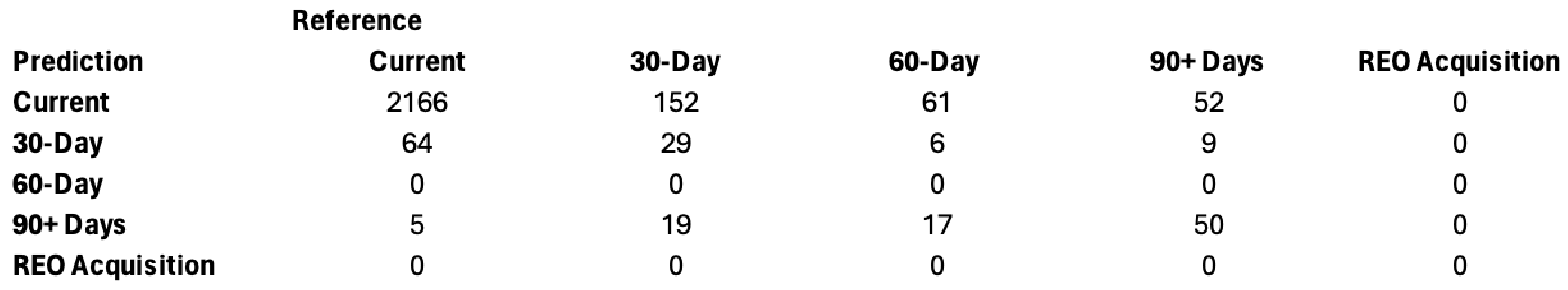
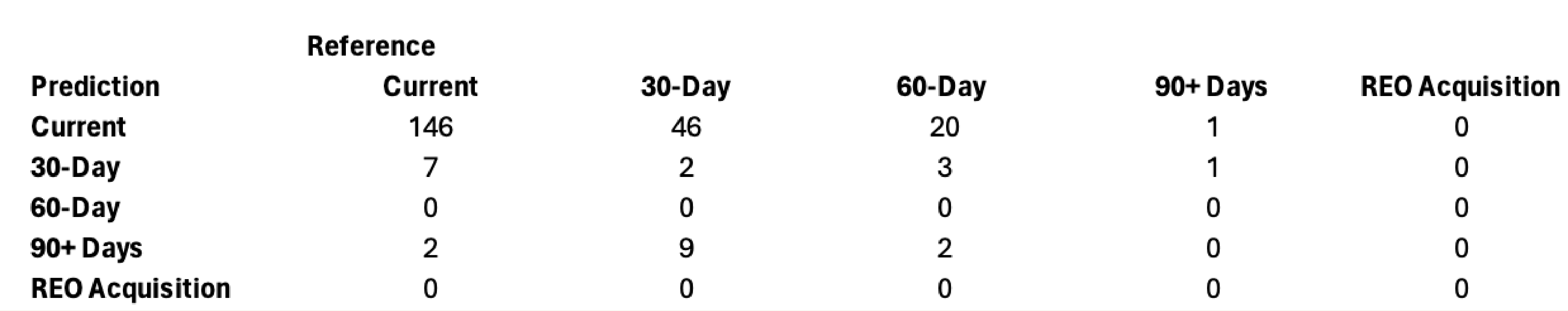
Input Features:

- 25 Key variables used, like Seller Name, Credit Score, Loan Age etc.

Performance Evaluation:

- Accuracy metrics were evaluated across Pre-COVID, COVID, and Post-COVID periods.

CONFUSION MATRIX : LOGISTIC MODEL

PERIOD	ACCURACY	RMSE	CONFUSION MATRIX
Before COVID	79.39%	7.18%	
During COVID	85.36%	2.24%	
Post COVID	61.92%	4.85%	

	Reference				
Prediction	Current	30-Day	60-Day	90+ Days	REO Acquisition
Current	345	22	12	8	0
30-Day	20	13	9	5	0
60-Day	0	0	0	0	0
90+ Days	4	8	13	31	0
REO Acquisition	0	0	0	0	0

	Reference				
Prediction	Current	30-Day	60-Day	90+ Days	REO Acquisition
Current	2166	152	61	52	0
30-Day	64	29	6	9	0
60-Day	0	0	0	0	0
90+ Days	5	19	17	50	0
REO Acquisition	0	0	0	0	0

	Reference				
Prediction	Current	30-Day	60-Day	90+ Days	REO Acquisition
Current	146	46	20	1	0
30-Day	7	2	3	1	0
60-Day	0	0	0	0	0
90+ Days	2	9	2	0	0
REO Acquisition	0	0	0	0	0

MODEL 2 : XGBOOST

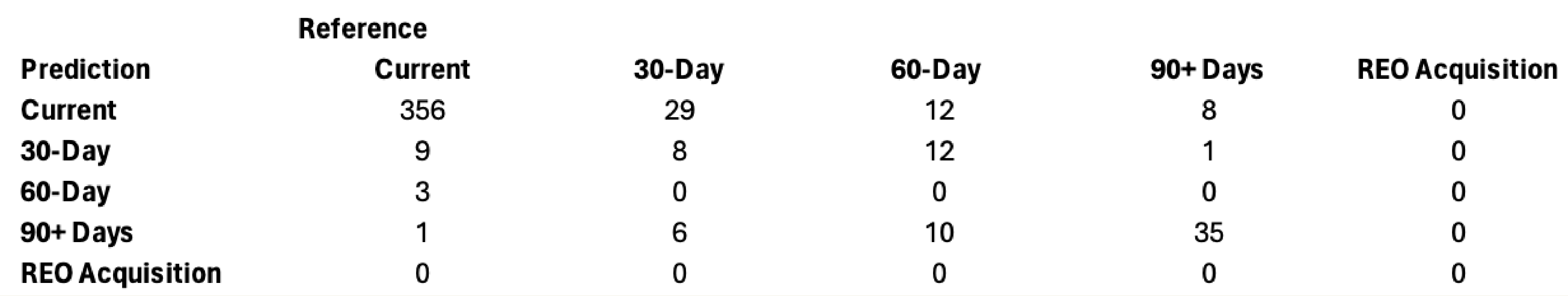
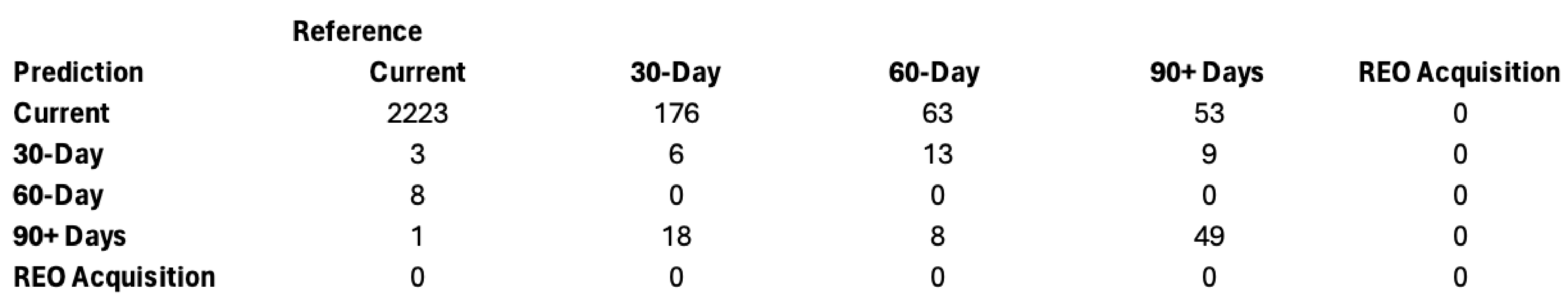
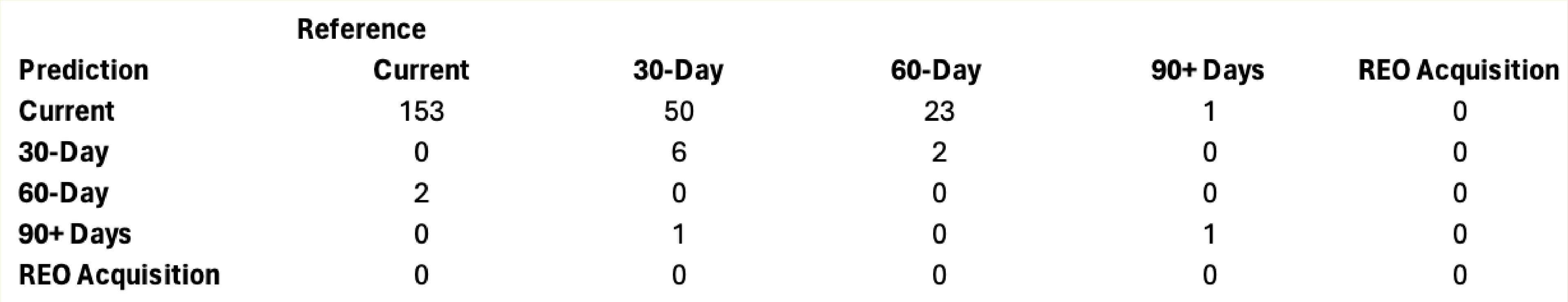
XGBoost:

- A high-performance, gradient boosting framework used for multiclass classification.
- Known for its scalability, efficiency, and ability to handle complex datasets with missing values and non-linear relationships.

- **Input Features**:
- Includes similar variables as in the logistic model.
- Advanced handling of categorical variables and interactions through automated feature importance ranking.

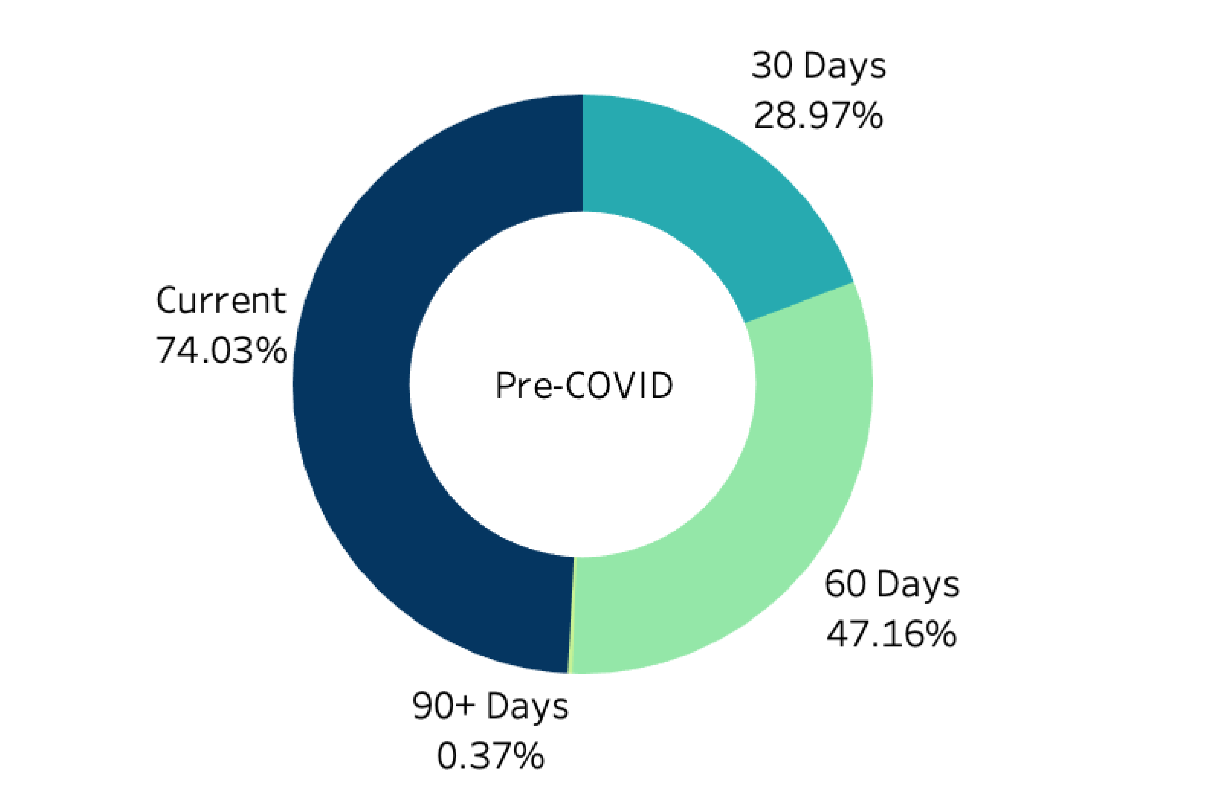
- **Performance Evaluation**:
- Evaluated across Pre-COVID, COVID, and Post-COVID periods.
- XGBoost performed better than the logistic model.

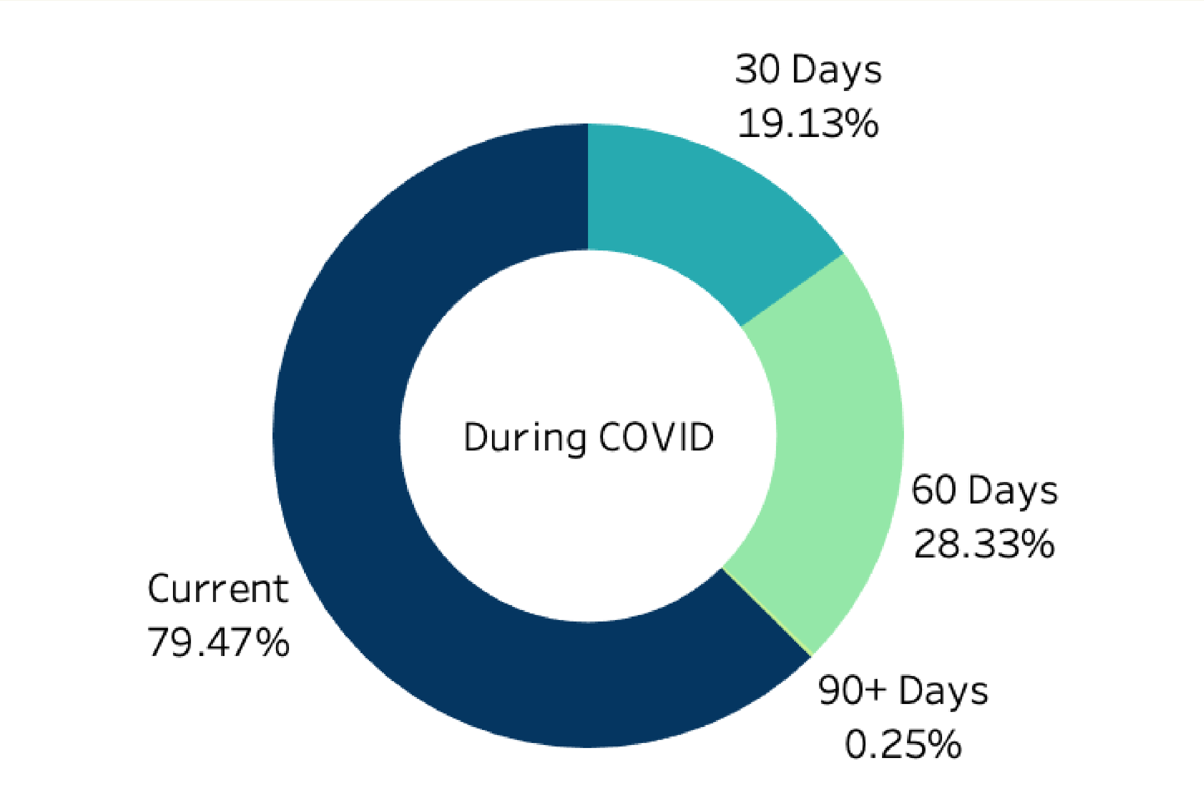
CONFUSION MATRIX : XGBOOST

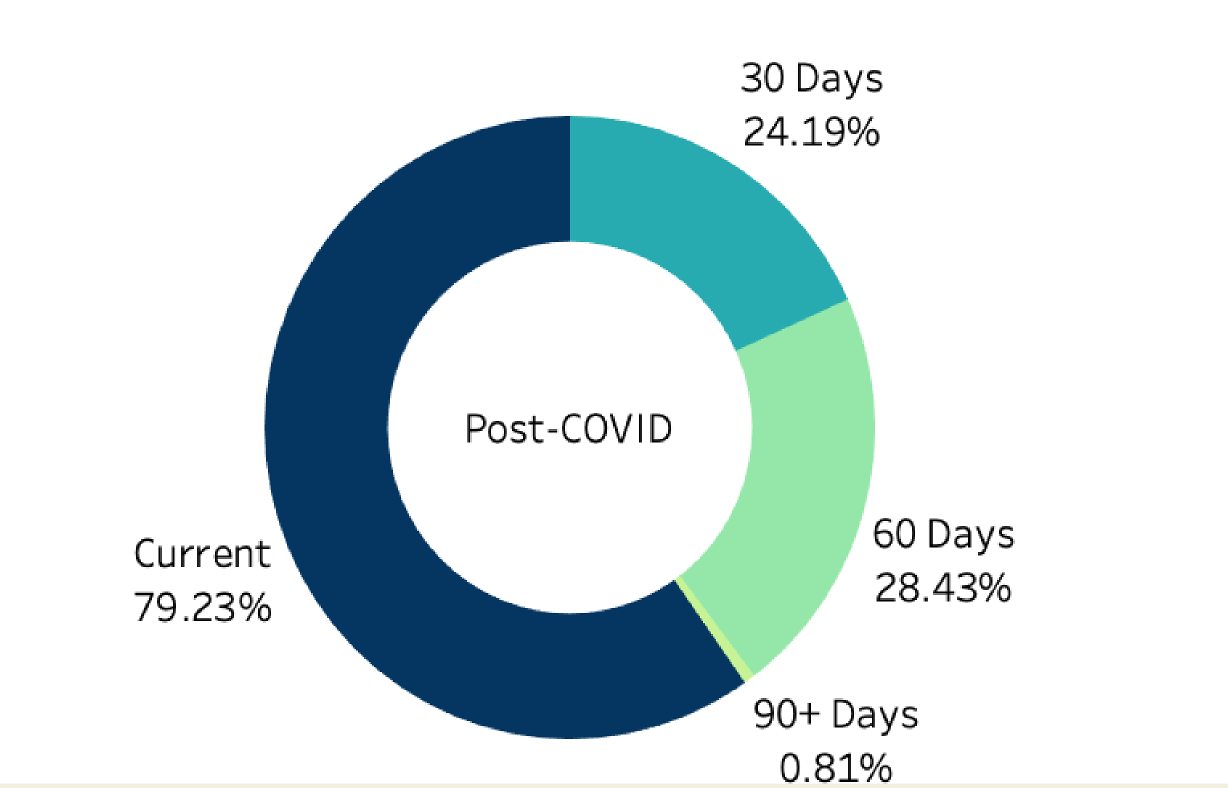
<table>
<tr><th>PERIOD</th><th>ACCURACY</th><th>RMSE</th><th>CONFUSION MATRIX</th></tr>
<tr><td>Before COVID</td><td>81.43%</td><td>7.58%</td><td>
<table>
<tr><th></th><th>Reference</th><th></th><th></th><th></th><th></th></tr>
<tr><th>Prediction</th><th>Current</th><th>30-Day</th><th>60-Day</th><th>90+ Days</th><th>REO Acquisition</th></tr>
<tr><td>Current</td><td>356</td><td>29</td><td>12</td><td>8</td><td>0</td></tr>
<tr><td>30-Day</td><td>9</td><td>8</td><td>12</td><td>1</td><td>0</td></tr>
<tr><td>60-Day</td><td>3</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>90+ Days</td><td>1</td><td>6</td><td>10</td><td>35</td><td>0</td></tr>
<tr><td>REO Acquisition</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
</table></td></tr>
<tr><td>During COVID</td><td>86.62%</td><td>1.47%</td><td>
<table>
<tr><th></th><th>Reference</th><th></th><th></th><th></th><th></th></tr>
<tr><th>Prediction</th><th>Current</th><th>30-Day</th><th>60-Day</th><th>90+ Days</th><th>REO Acquisition</th></tr>
<tr><td>Current</td><td>2223</td><td>176</td><td>63</td><td>53</td><td>0</td></tr>
<tr><td>30-Day</td><td>3</td><td>6</td><td>13</td><td>9</td><td>0</td></tr>
<tr><td>60-Day</td><td>8</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>90+ Days</td><td>1</td><td>18</td><td>8</td><td>49</td><td>0</td></tr>
<tr><td>REO Acquisition</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
</table></td></tr>
<tr><td>Post COVID</td><td>66.95%</td><td>4.74%</td><td>
<table>
<tr><th></th><th>Reference</th><th></th><th></th><th></th><th></th></tr>
<tr><th>Prediction</th><th>Current</th><th>30-Day</th><th>60-Day</th><th>90+ Days</th><th>REO Acquisition</th></tr>
<tr><td>Current</td><td>153</td><td>50</td><td>23</td><td>1</td><td>0</td></tr>
<tr><td>30-Day</td><td>0</td><td>6</td><td>2</td><td>0</td><td>0</td></tr>
<tr><td>60-Day</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>90+ Days</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>REO Acquisition</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
</table></td></tr>
</table>

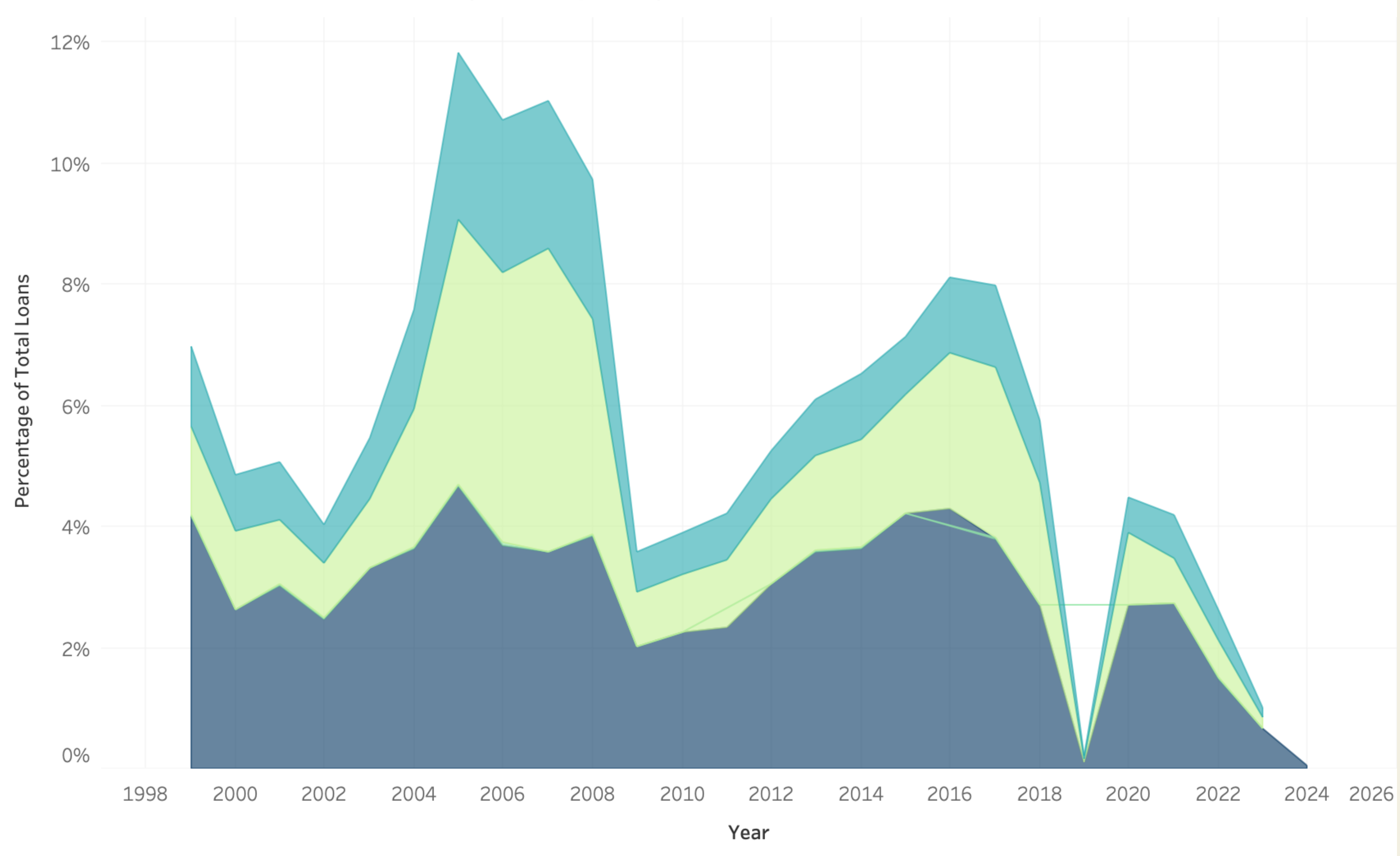
Findings and Insights

Shifting Trends in Loan Transitions: Pre-COVID, During, and Post-COVID Insights







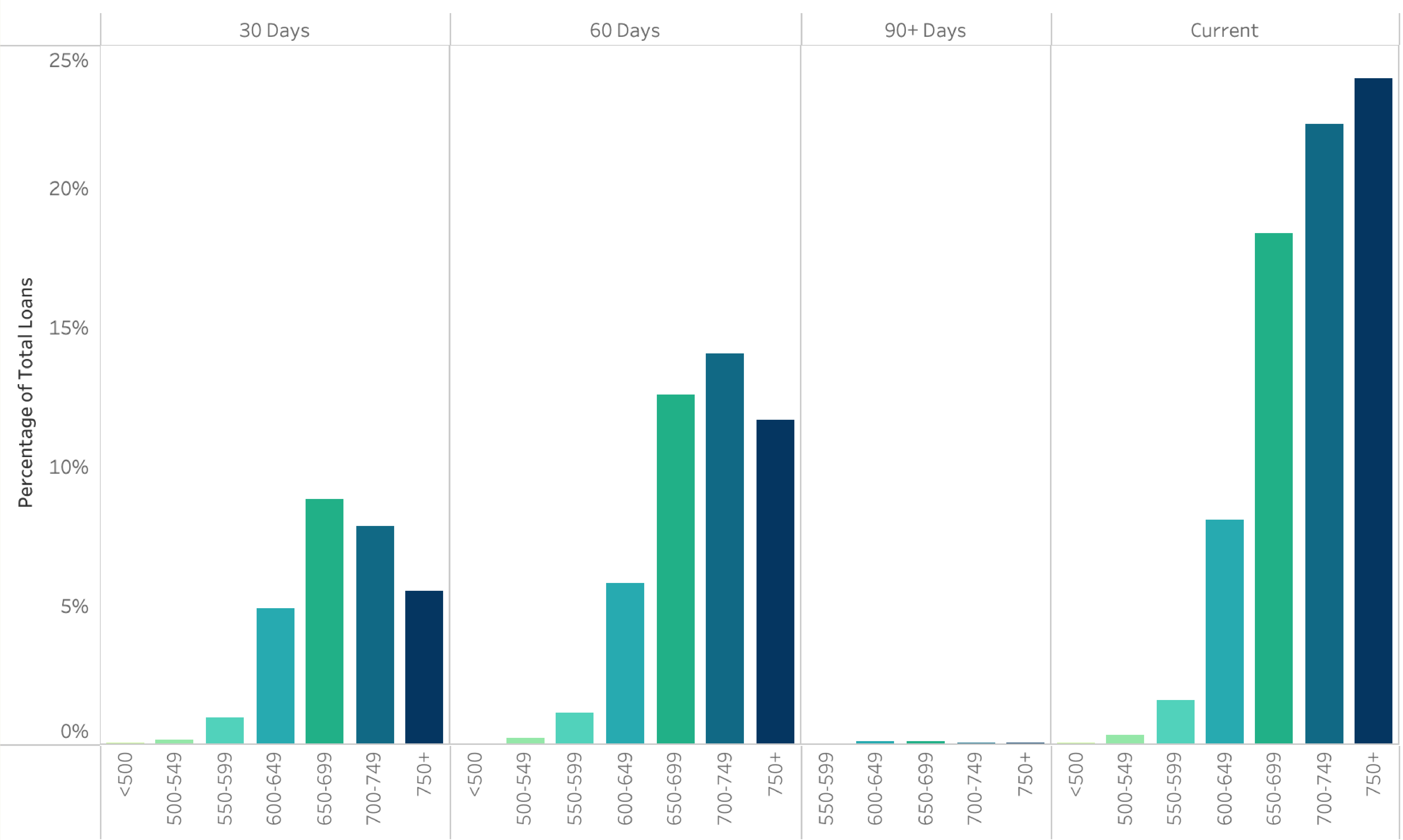


- Pre-COVID, 47.16% of loans transitioned to 60-day delinquency, highlighting **increased delinquency risks**.

- During COVID, 79.47% of loans stayed "Current," likely due to **relief measures** supporting borrowers.

- Post-COVID, a **slight increase** of 0.81% in 90+ day delinquency reflects ongoing challenges as conditions normalized.

Credit Score Dynamics:
Loan Transition Patterns Across Statuses

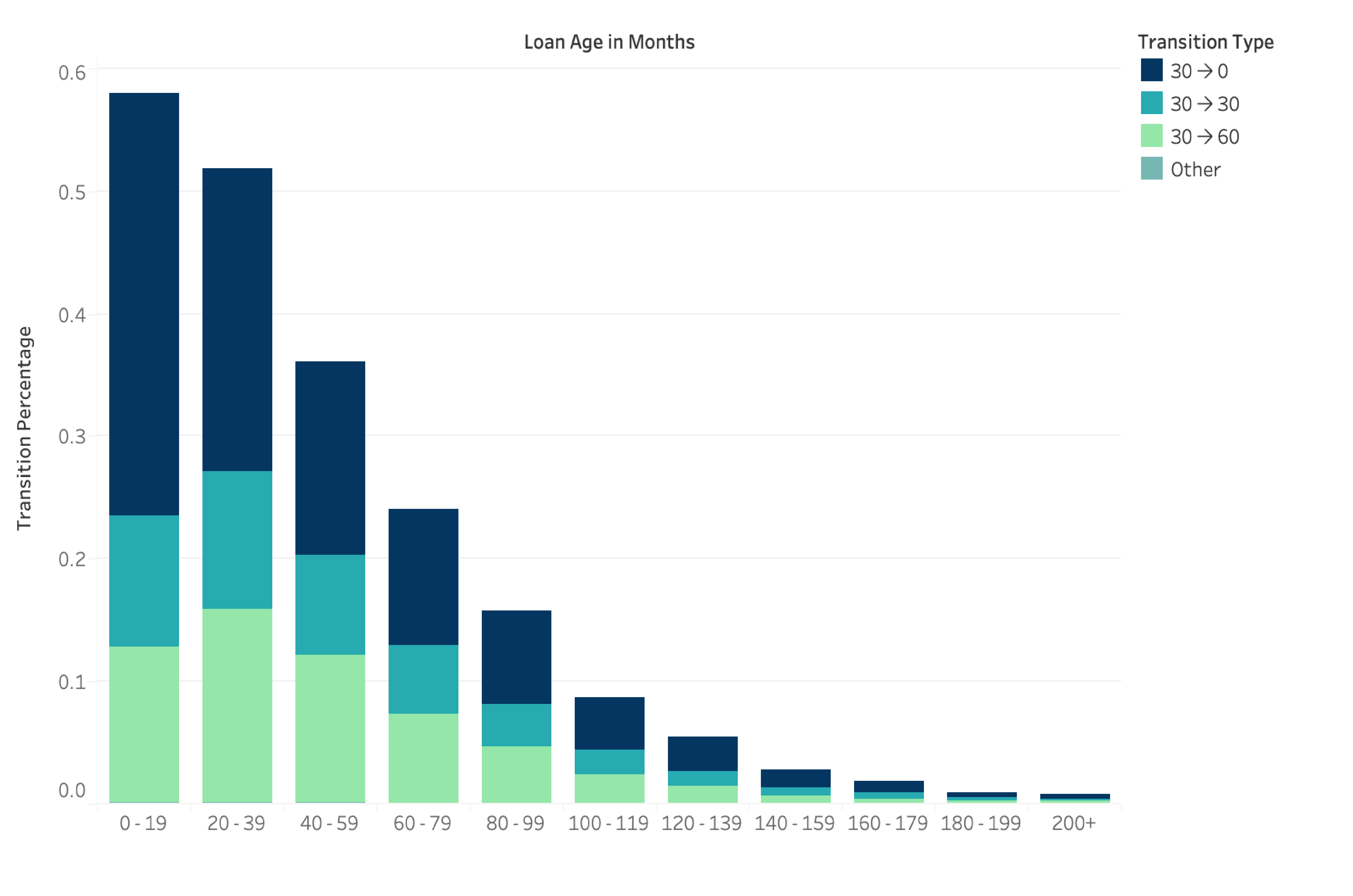


- Credit score distribution reveals clear differences in loan performance across statuses.

- **Higher credit scores (700+)** are strongly linked to "Current" status, indicating stable repayment.

- The chart provides insights into how c**redit risk varies** across score bands and loan statuses.

From Delinquency to Resolution or Escalation: Transition Patterns by Loan Age



- High transition activity in **newer loans (0–39 months**) reflects effective early interventions.
- **Older loans (60+ months)** face higher risks of persistent delinquency.
- **Declining** resolution rates with loan age stress the importance of timely action.

Recommendations

Recommendations

Credit Score Insights

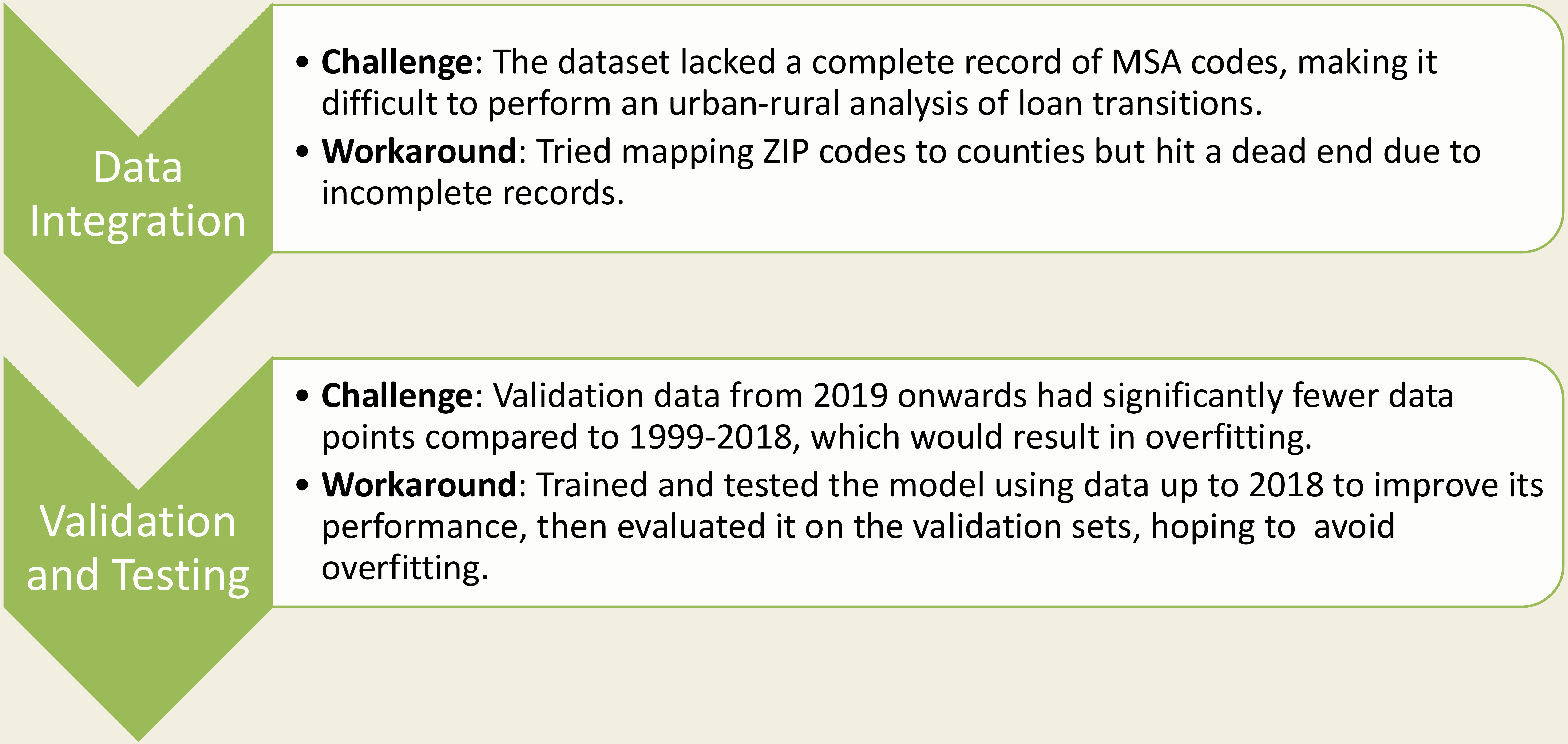
- Prioritize lending to borrowers with 700+ credit scores.
- Use stricter criteria for sub-700 scores.
- Offer refinancing to reward high-performing borrowers.

Loan Age Insights

- Engage borrowers early (0–39 months) to reduce delinquency risks.
- Provide tailored support for loans aged 60+ months to address persistent issues.
- Closely monitor older loans as resolution rates decline.

Challenges and Workarounds

Challenges and Workarounds



Data Integration

- **Challenge**: The dataset lacked a complete record of MSA codes, making it difficult to perform an urban-rural analysis of loan transitions.
- **Workaround**: Tried mapping ZIP codes to counties but hit a dead end due to incomplete records.

Validation and Testing

- **Challenge**: Validation data from 2019 onwards had significantly fewer data points compared to 1999-2018, which would result in overfitting.
- **Workaround**: Trained and tested the model using data up to 2018 to improve its performance, then evaluated it on the validation sets, hoping to avoid overfitting.

SOURCES

Freddie Mac Loan Level Data

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

FRED Economic Data

https://fred.stlouisfed.org/

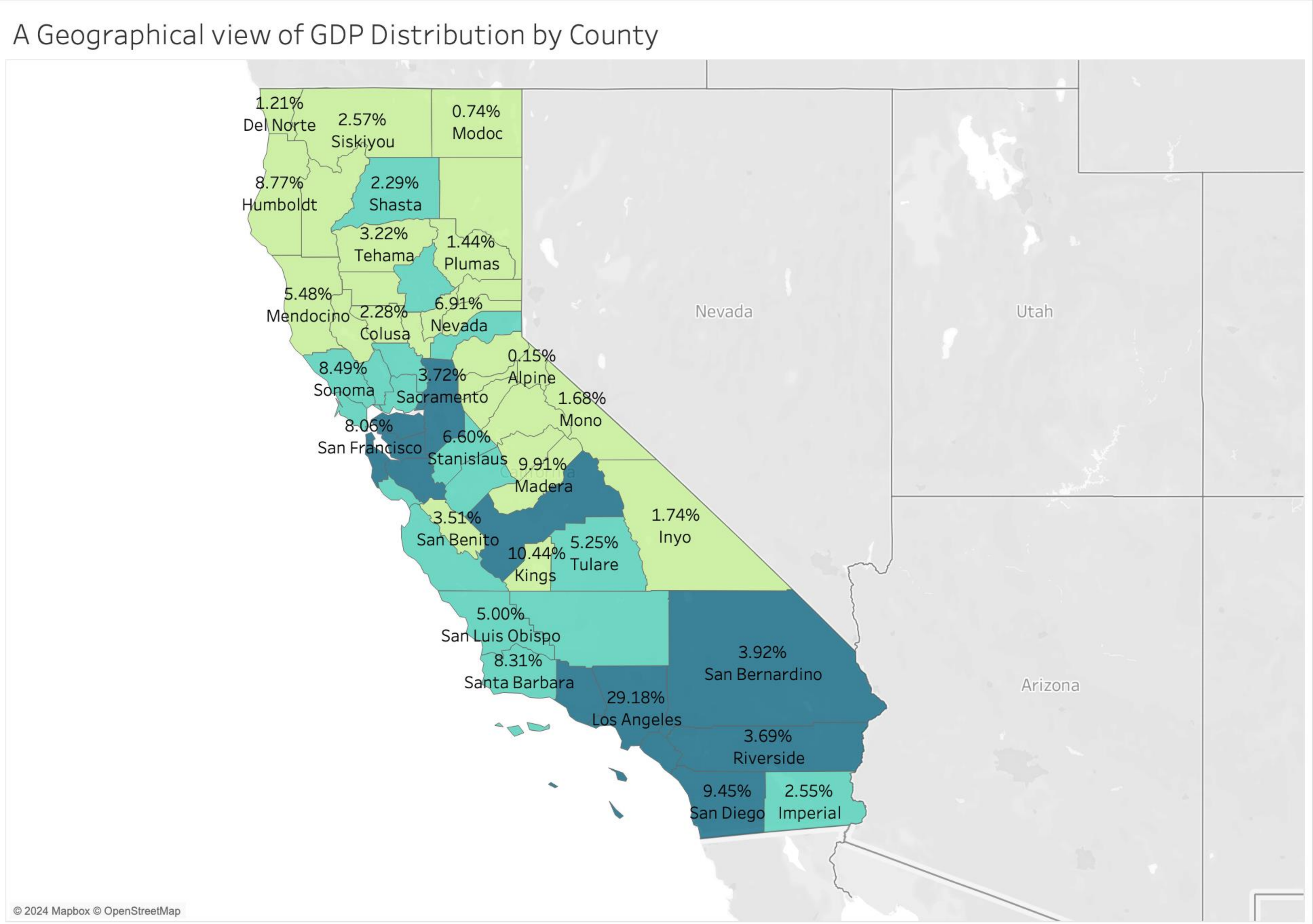
Bureau of Economic Analysis

https://www.bea.gov/

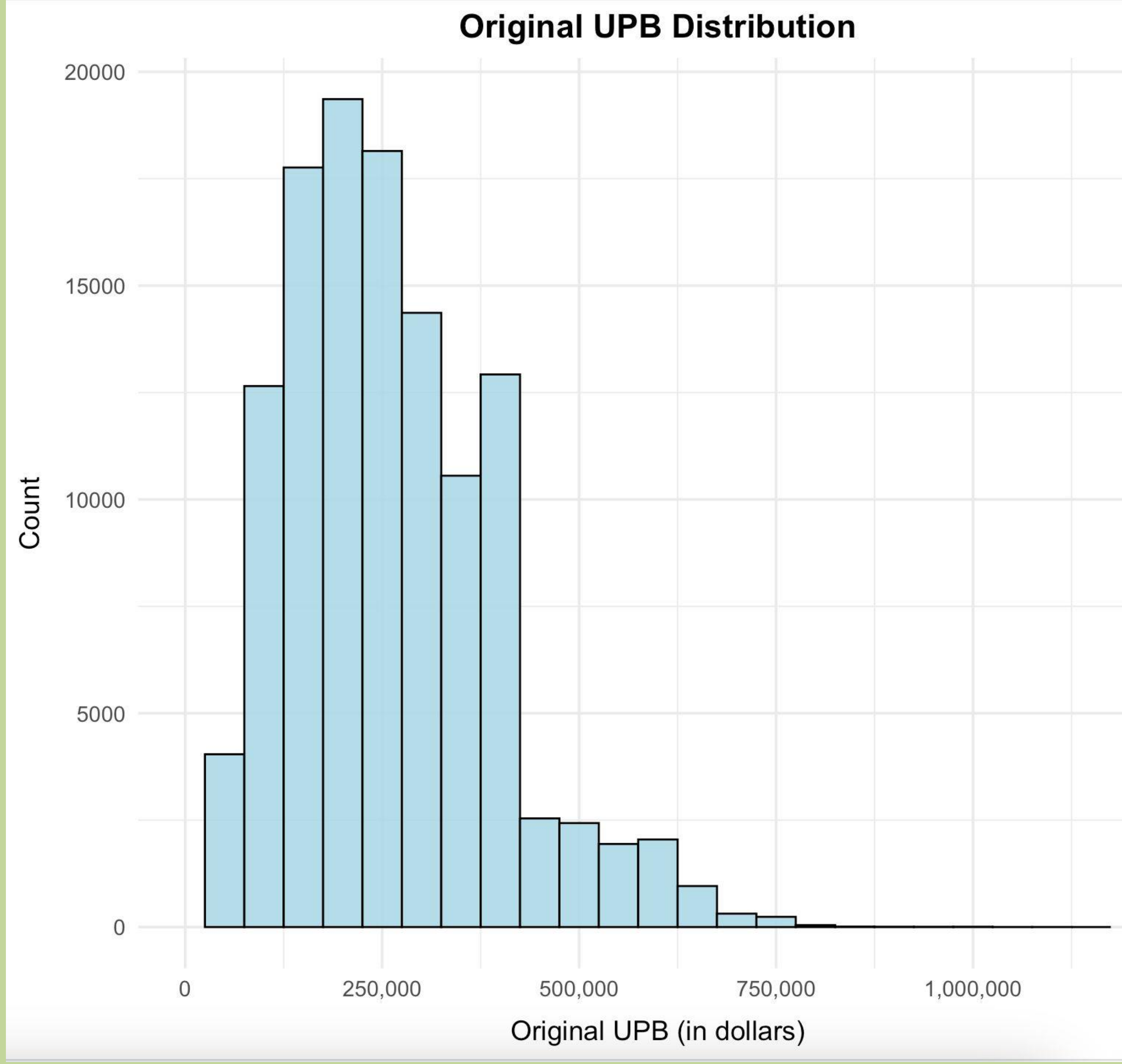
Thank You

Appendix

APPENDIX

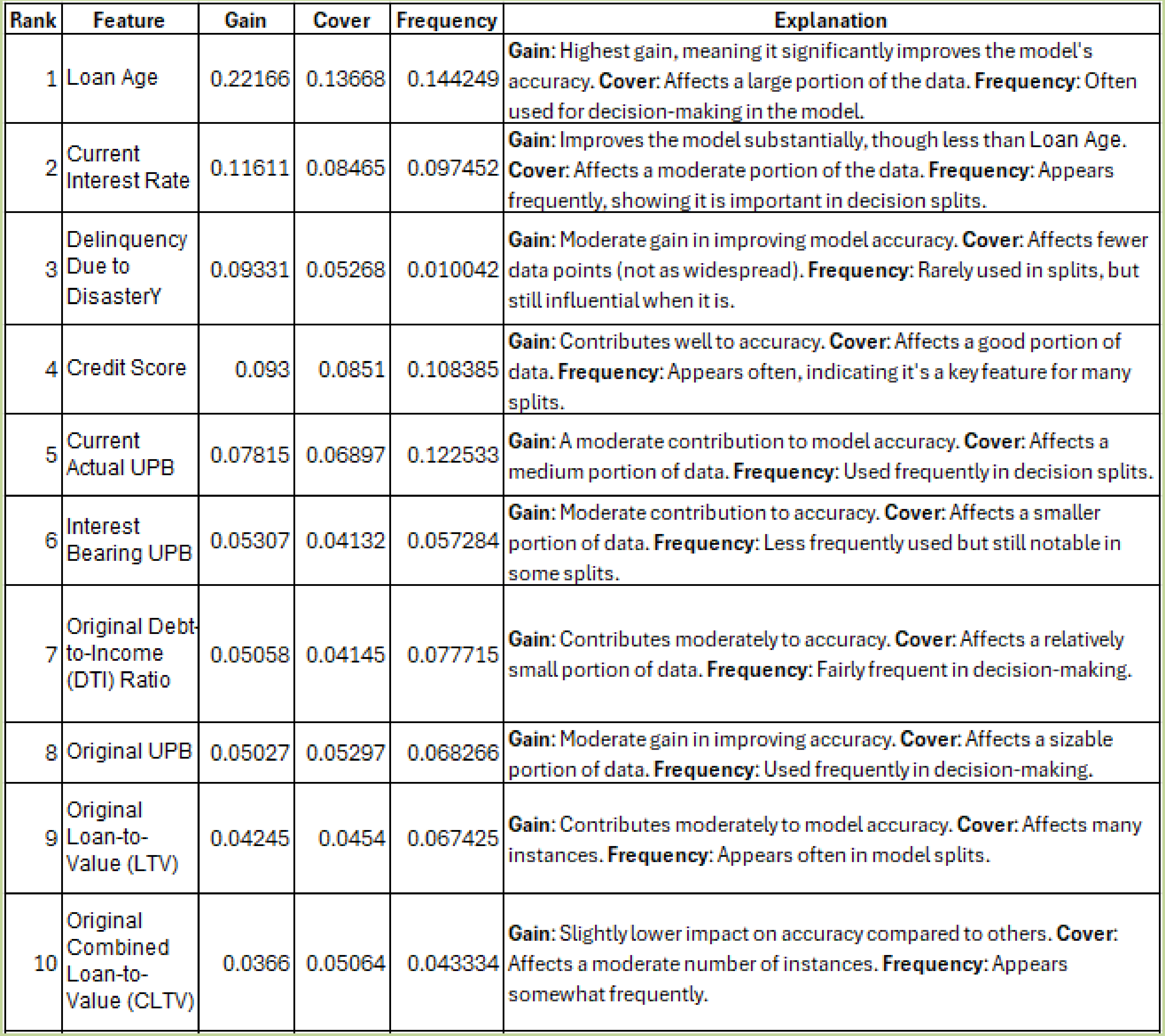


Rural, Urban and Suburban - The Divide



Distribution of Original Unpaid Principal Balance (UPB)

Important Features - XGBOOST



Rank	Feature	Gain	Cover	Frequency	Explanation
1	Loan Age	0.22166	0.13668	0.144249	**Gain**: Highest gain, meaning it significantly improves the model's accuracy. **Cover**: Affects a large portion of the data. **Frequency**: Often used for decision-making in the model.
2	Current Interest Rate	0.11611	0.08465	0.097452	**Gain**: Improves the model substantially, though less than Loan Age. **Cover**: Affects a moderate portion of the data. **Frequency**: Appears frequently, showing it is important in decision splits.
3	Delinquency Due to DisasterY	0.09331	0.05268	0.010042	**Gain**: Moderate gain in improving model accuracy. **Cover**: Affects fewer data points (not as widespread). **Frequency**: Rarely used in splits, but still influential when it is.
4	Credit Score	0.093	0.0851	0.108385	**Gain**: Contributes well to accuracy. **Cover**: Affects a good portion of data. **Frequency**: Appears often, indicating it's a key feature for many splits.
5	Current Actual UPB	0.07815	0.06897	0.122533	**Gain**: A moderate contribution to model accuracy. **Cover**: Affects a medium portion of data. **Frequency**: Used frequently in decision splits.
6	Interest Bearing UPB	0.05307	0.04132	0.057284	**Gain**: Moderate contribution to accuracy. **Cover**: Affects a smaller portion of data. **Frequency**: Less frequently used but still notable in some splits.
7	Original Debt-to-Income (DTI) Ratio	0.05058	0.04145	0.077715	**Gain**: Contributes moderately to accuracy. **Cover**: Affects a relatively small portion of data. **Frequency**: Fairly frequent in decision-making.
8	Original UPB	0.05027	0.05297	0.068266	**Gain**: Moderate gain in improving accuracy. **Cover**: Affects a sizable portion of data. **Frequency**: Used frequently in decision-making.
9	Original Loan-to-Value (LTV)	0.04245	0.0454	0.067425	**Gain**: Contributes moderately to model accuracy. **Cover**: Affects many instances. **Frequency**: Appears often in model splits.
10	Original Combined Loan-to-Value (CLTV)	0.0366	0.05064	0.043334	**Gain**: Slightly lower impact on accuracy compared to others. **Cover**: Affects a moderate number of instances. **Frequency**: Appears somewhat frequently.