PUBLIC RECORD DATA ANALYSIS

Trunal Wadibhasme, MT2135

Internal Guide: Mihir Arjunwadkar External Guide: Naik, Vinayak

Introduction

- In "PUBLIC RECORDS DATA ANALYSIS," we have several datasets.
- Dataset includes mortgage, assignment, deed, release, NOD, and listing data from various sources like tax office, county office, and listing services.
- The dataset is used to create forecasting models for predicting real estate transactions that may occur in a given month and location (state or county).
- Two modules are used for data analysis:
- Data Quality Assessment: Determines data completeness by column and location.
- Identify The Gap And Its Patterns: Identifies gaps in county recorder data based on recording dates.
- Gaps refer to periods without transactions, occurring daily, monthly, or yearly.

Problem Statement

1. Gap Analysis in County Recorder Data: Programmatically detect and catalog gap patterns for the States and counties in the sample data.
   - Over a period of time, there is no transaction that happened. (1st type of Gap)
   - Might be happened but didn't send through the files. (2nd type of Gap)
2. Predict when the next update should be expected(New Records).

Average Percentage Of Data Populated By States

Figure 1 Shows the average percentage of data populated by states.the high percentage of data populated in "Nevada" State with **50.15** percent data populated.

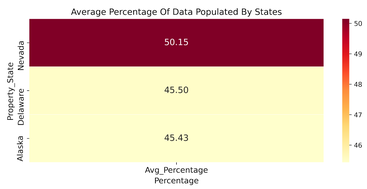

Figure 1. Data Populated By States

Monthly Average Interest Rate By States

Figure 2 displays the average interest rate by month over multiple years, categorized by states. It provides insights into the monthly trends of interest rates across different states. The data reveals that interest rates fluctuate within each state from month to month.

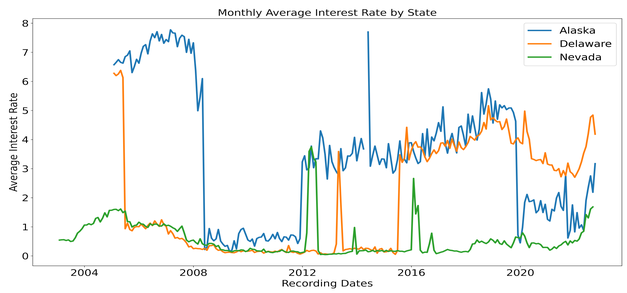

Figure 2. Monthly Average Interest Rate In each State

Gap Analysis Using Time Series By County

Figure 3 Show the number of transactions that happened in each month.However, there are some months where transactions didn't occur or might be happened but were not sent through the files. The first transaction was recorded on "2002-10-30" and the last transaction was recorded on "2022-09-30".In between, there are gaps in several counties.

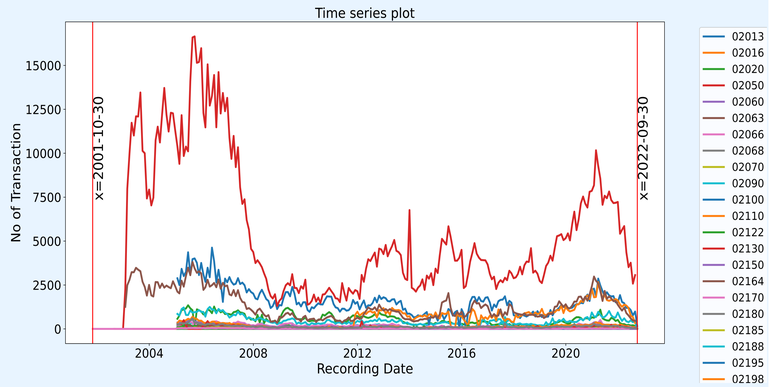

Figure 3. Gap Analysis By Time Series For Mortgage Data By County

Gap Pattern In Most Gap County:02066 In Mortgage Data

Figure 4 Represent the gap pattern in most gap county 02066. We can observe there no transactions has recorded in February from the year 2006 to 2021.Only one transaction has been recorded in February 2022.

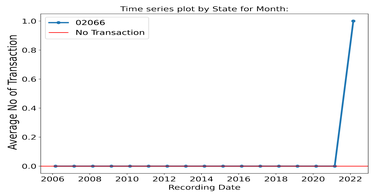

Figure 4. Gap pattern In February Month

Percentage Of Gap By County

Figure 5 show the percentage of the gap by county in "Mortgage Data".The high percentage of gap occurs in Plymouth County(02066) with **98.04** percentage gap.

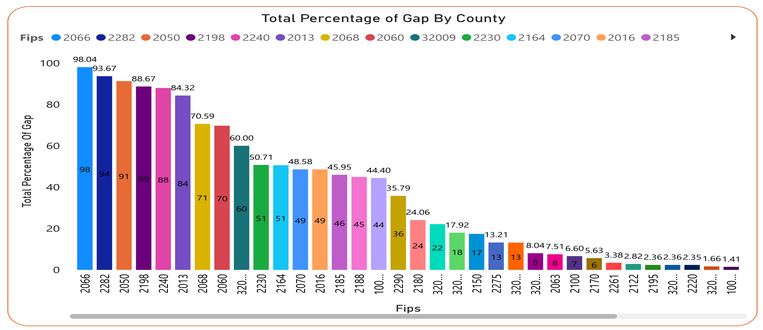

Figure 5. Percentage Of Gap For Each County

Predict When the next update should be expected by county Using ML Models.

Bayes'Theorem : Finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{1}$$

where A and B are events and P(B) ≠ 0.
Now, with regard to our dataset, we can apply Bayes' theorem in the following way:

$$P(y|X) = \frac{P(X|y) \cdot P(y)}{P(X)} \tag{2}$$

where y is the class variable and X is a dependent feature vector (of size n) where: X = (x1,x2,x3,.....,xn). Just to clarify, an example of a feature vector and corresponding class variable
X = (month, day,is-holiday, day of year, week number,state-name,city-name)
y = (0,1), Here "0" means didn't get update and "1" means "got update".

Figure 6 shows the performance of machine learning models. here "Naive Bayes" perform better as compared to Decision trees and random forests and logistic regression.The F1 Score for Naive Bayes is **0.77** .

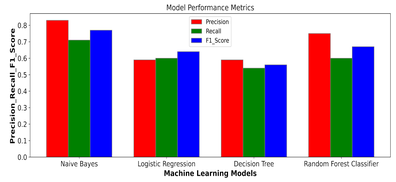

Figure 6. Performance Of ML Models

Summary and Conclusions

The below table shows the summary of the gap in "Mortgage Data" across the three "states".

Content	Alaska	Delaware	Nevada
Total Counties	29	3	16
Total Counties Have Gap	29	3	10
Total Counties Have No Gap	0	0	6
Most Percentage Of Gap By County	98.04	44.40	60
Percentage Of Gap By State	0	8.2	5.2

Table 1. Summary Of Gap In Mortgage Data

Conclusions:

1. The most average percentage of data populated in "Nevada" State in "Mortgage" Data with **50.15** percent average data.
2. There are **Six** county in "Nevada" state which don't have any gap.These are the clean counties.
3. The most percentage of gap occurred in 02066(Plymouth County) with **98.04** percent gap across different County.
4. The next update will occur on **"August" "Friday" on fourth week in "2023"** for county "32005"(Douglas County) and "32023" (Nye County).

Future Scope

- Identify the type of gaps in a gap month.
- How often the data is backfield?

Department Of Scientific Computing, Modeling and Simulation, Savitribai Phule Pune University.

trunalwadibhasme12@gmail.com