COMPLAINALYZER

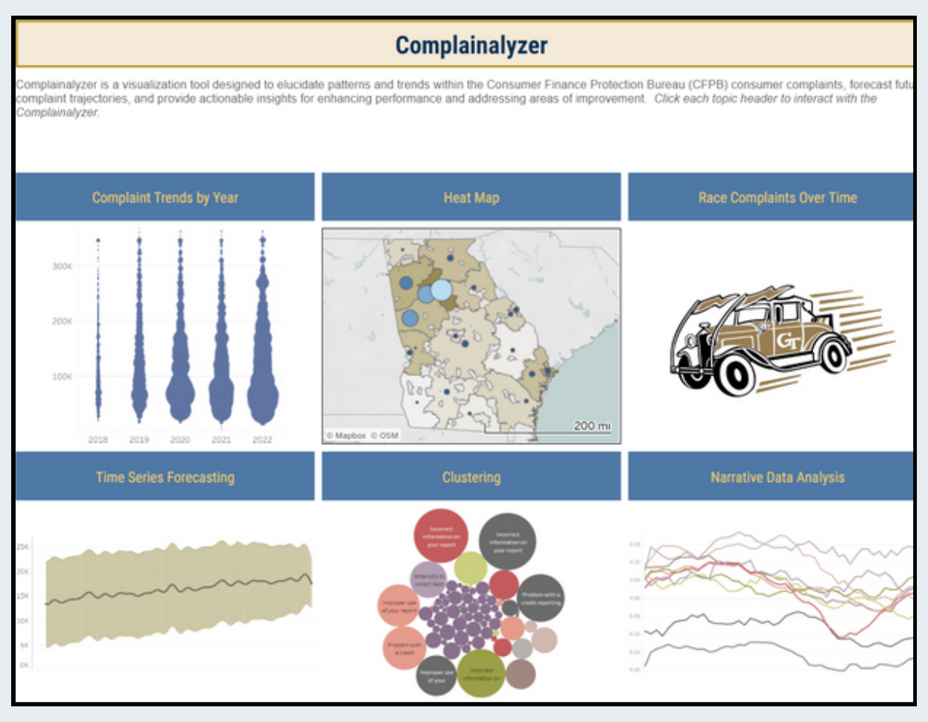
Team 95: Nā Mele Menehune
Ethan Maluhia Roberts, Divya Chandrasekaran, Hillary Latham,
James McGarr, Suganya Natarajan, Chaitanya Evani

Georgia Tech

INTRODUCTION AND MOTIVATION

What is the problem? The Consumer Financial Protection Bureau (CFPB) collects consumer complaint data about financial issues, including banks and credit companies. The CFPB website shares categorical and numerical trends for complaints submitted over time, but lacks forecasts and trends in complaint narratives. It has limited demographic and socioeconomic data about each complaint (for obvious privacy reasons).

Why is it important and why should we care? Understanding trends aids the CFPB's commitment to protect consumers. Publishing forecasting insights could bring attention to areas where financial institutions are failing to comply with regulations before they become widespread. Analysis of complaint narratives can provide deeper insights into specific problems, leading to better services and proactive solutions. Predictive trends can inform policymakers about changes needed in regulatory policy creation. We care because the effectiveness of the CFPB affects the fairness and quality of financial services that we, as consumers, receive and ensures financial institutions operate with integrity and transparency and so consumers have a voice that is heard.



public.tableau.com/views/Complainalyzer/Complainalyzer

OUR APPROACH

Rationale Blending quantitative forecasting with qualitative insights offers a holistic view of consumer sentiment and behavior. Merging complaint and census data unlocked deeper insights into the *why* behind complaint patterns.

What's new We tapped into advanced natural language processing to reveal prevalent themes in customer narratives and measure sentiment, offering a richer picture of consumer experiences. Innovative clustering and sentence structure analysis techniques identified and categorized common complaints, sharpening focus on critical issues.

1. Load and clean consumer finance complaint data. Merge complaints data with census data by zip code.
2. Train Time Series models to forecast future complaints, accounting for seasonal variations and trends.
3. Sentence structure analysis, looking common sentence structures for common complaints.
4. Clean, tokenize and lemmatize complaint narratives for a multi-pass Latent Dirichlet Allocation (LDA) model. Using the model, determine an optimal number of topics with a coherency score and grade each narrative to identify the keywords.
5. Utilize K-means clustering to group demographic attributes and complaint data (race population, issue, state, zip code, complaint type, and complaint count)
6. Create an interactive visualization in Tableau to share the story of our modeling results and share results publicly.

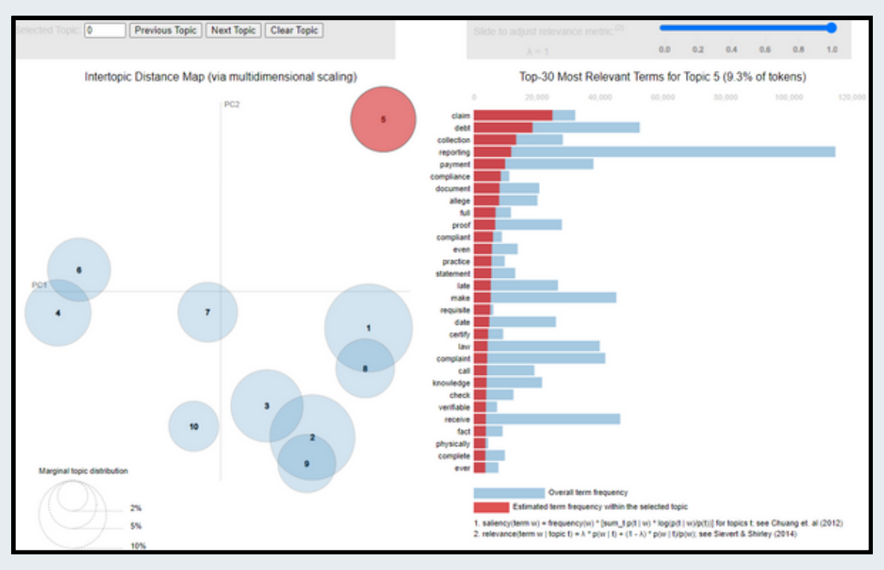
DATASETS

Consumer complaint data from the CFPB website was joined with US 2020 census data using official US zip codes.

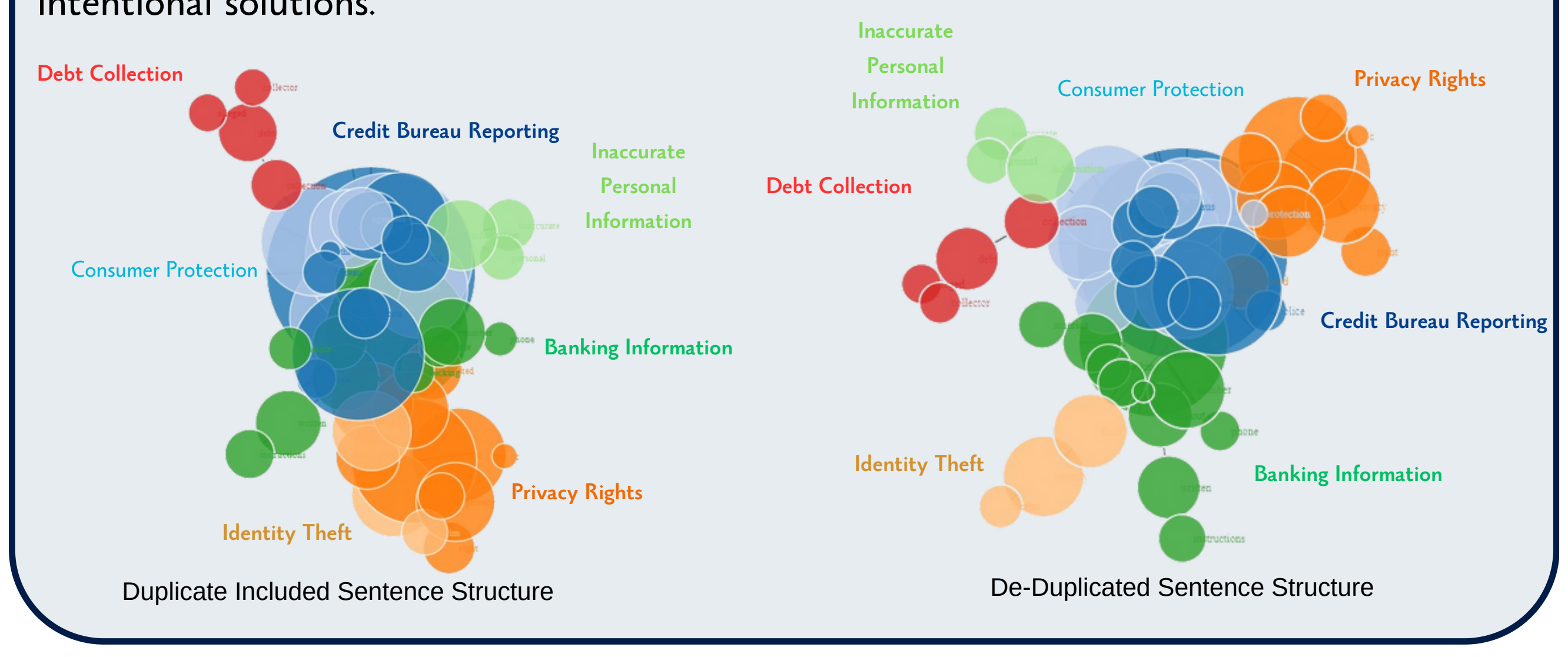
Dataset Characteristics The combined cleaned dataset is 1.5gb, with two million rows of complaint data reported from 2018 - 2022. There are 18 complaint attributes (date, categorical, narrative, geographic) and 20 census attributes (geographic, numeric).

cfpb
United States Census 2020
United States Zip Codes.org

COMPLAINT CLUSTERING


Intertopic Distance Map (via multidimensional scaling)
Top-30 Most Relevant Terms for Topic 5 (9.3% of tokens)
Previous Topic
Next Topic
Clear Topic
Marginal topic distribution
Overall term frequency
Estimated term frequency within the selected topic

We compared the Louvain algorithm output (below) with the LDA algorithm (right) visualization in order to determine the best way to cluster the complaint narratives. Using the coherency score, LDA recommended 10 clusters for complaints. The Louvain clustering displayed decided 7 topics was appropriate. The CFPB has 9 products that users can complain about. This is where the artform of analytics comes in, determining which algorithm best suits the needs of the target audience. If the CFPB's budget allows them to target three topics, one way to address the complaints would be to use 3 clusters to cover the most complaints, or it may be best to maintain the separation of focused clusters and target the 3 with the most complaints for less coverage, but with more intentional solutions.

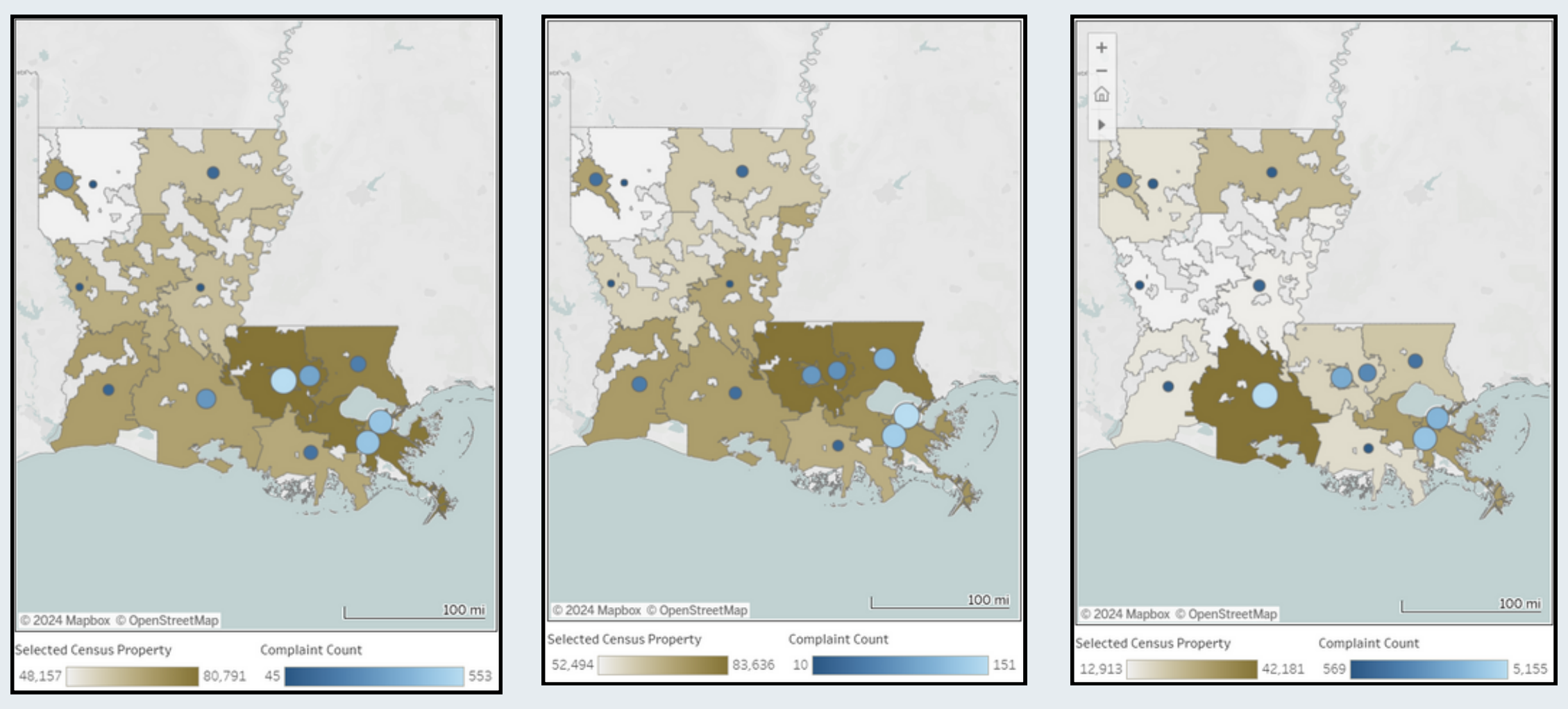


Duplicate Included Sentence Structure

De-Duplicated Sentence Structure

GEOGRAPHIC EXPLORATION

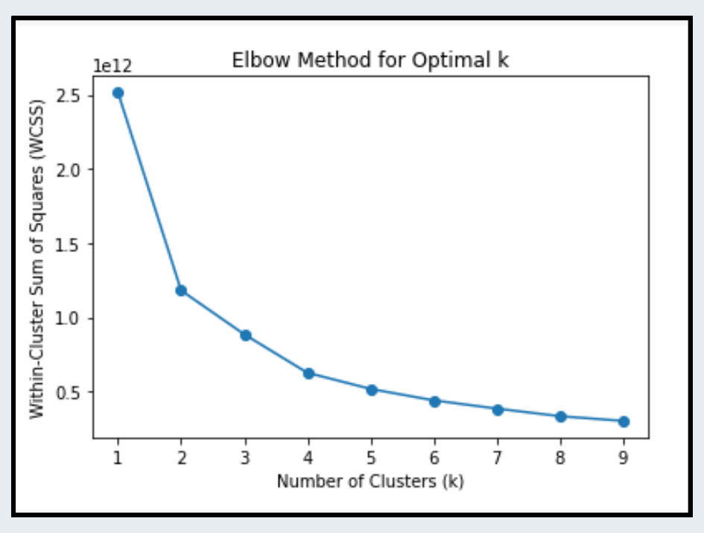
In Tableau, a complaint heat map by zip code region is overlaid with a census heat map. In Louisiana, for example, the below images show varying complaint volume in zip regions with the same mean income (1,2) but align with expectations when highlighting public assistance income in areas with complaints for credit reporting (3).

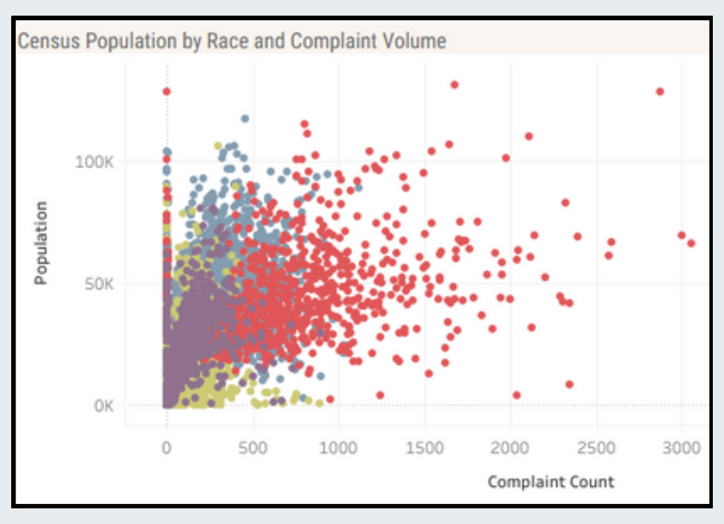


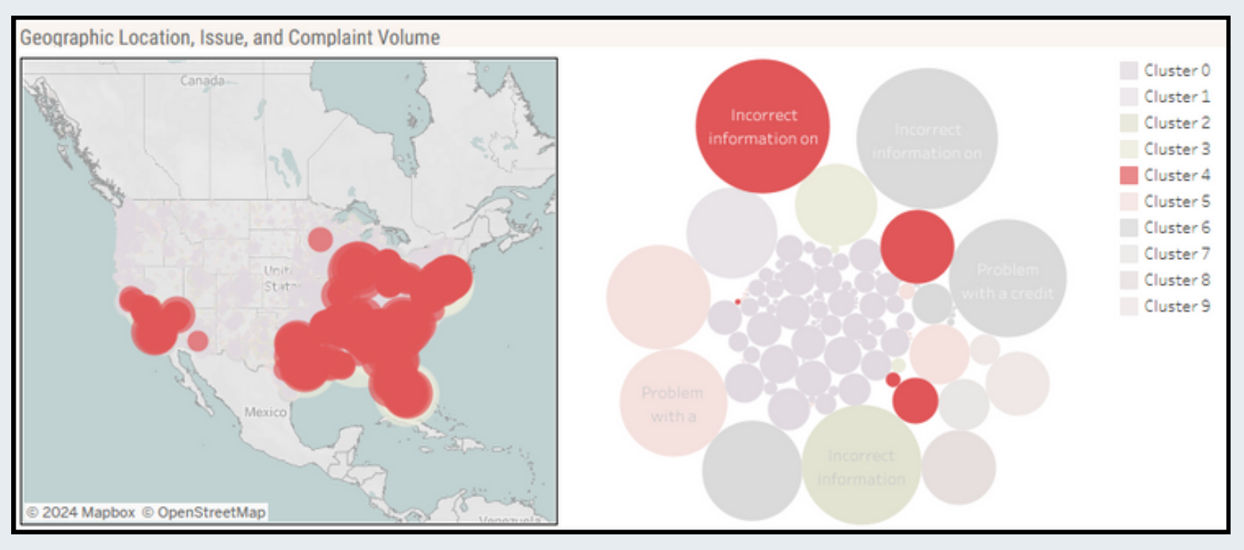
From left to right, (1) Mean Household Income with Debt Collection complaints, (2) Mean Household Income with Mortgage complaints, and (3) Households with Public Assistance Income with Credit Reporting complaints.

CLUSTERING

Modelling We performed K-Means clustering to group zip codes and races based on complaint types. We utilized a preprocessing pipeline to standardize and encode features and determined the optimal number of clusters using the elbow method, resulting in k=9 for Issue and Zip clustering and k=4 for race population clustering. We then established a clustering pipeline using Pipeline to calculate various statistics to identify common products, zip codes, and states within each cluster.





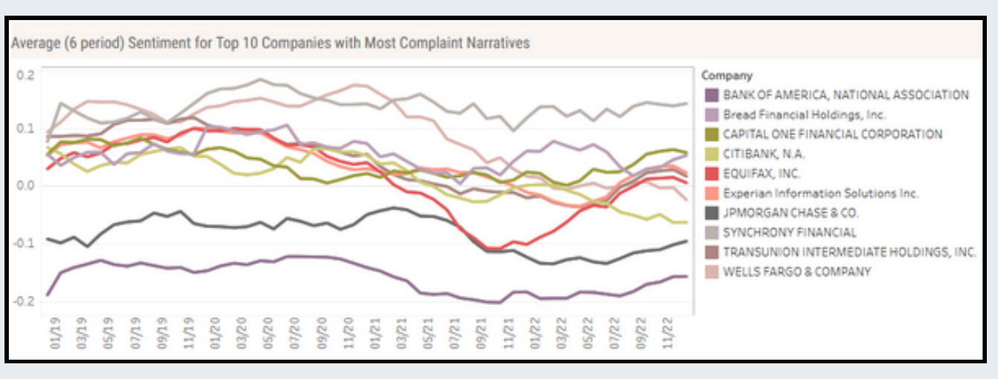


Important Inferences

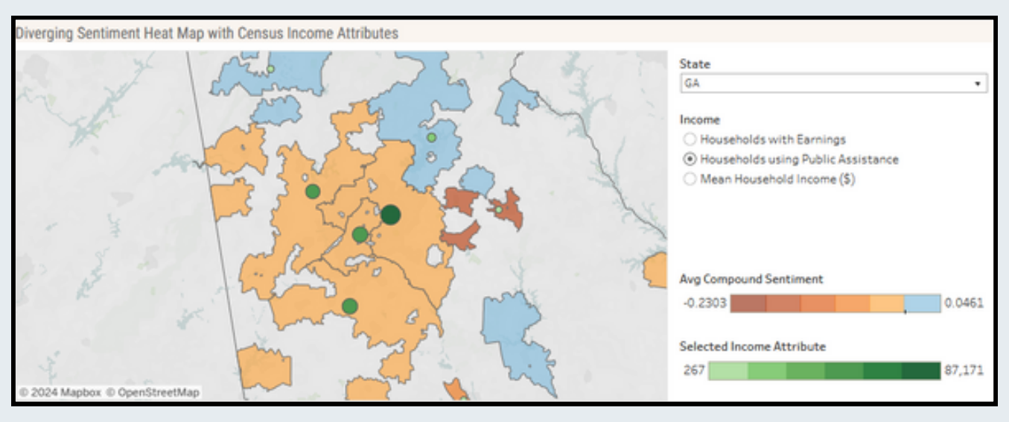
There are widespread instances of incorrect credit report information across multiple regions and a predominance of mortgage-related complaints in south Los Angeles, highlighting a potential need for assistance. The majority of clusters in CA, NY and GA are in areas of high African American populations and in CA in suburban and urban communities in areas with a high proportion of white residents. "Credit reporting, credit repair services" is the highest product category seen in all clusters with the major issue being "Incorrect information on your report."

SENTIMENT ANALYSIS

Basic topic modeling using Latent Dirichlet Allocation (LDA). Narrative data was cleaned of identifying and redacted information. Cleaned sentences were tokenized and models built. Computing a coherency score helped evaluate the optimal number of topics which was identified to be 10 as seen on the inter-topic distance map. Compound sentiment analysis was performed using VADER sentiment analyzer on cleaned narratives. Sentiment scores range from -1 to 1 (most negative to most positive response).



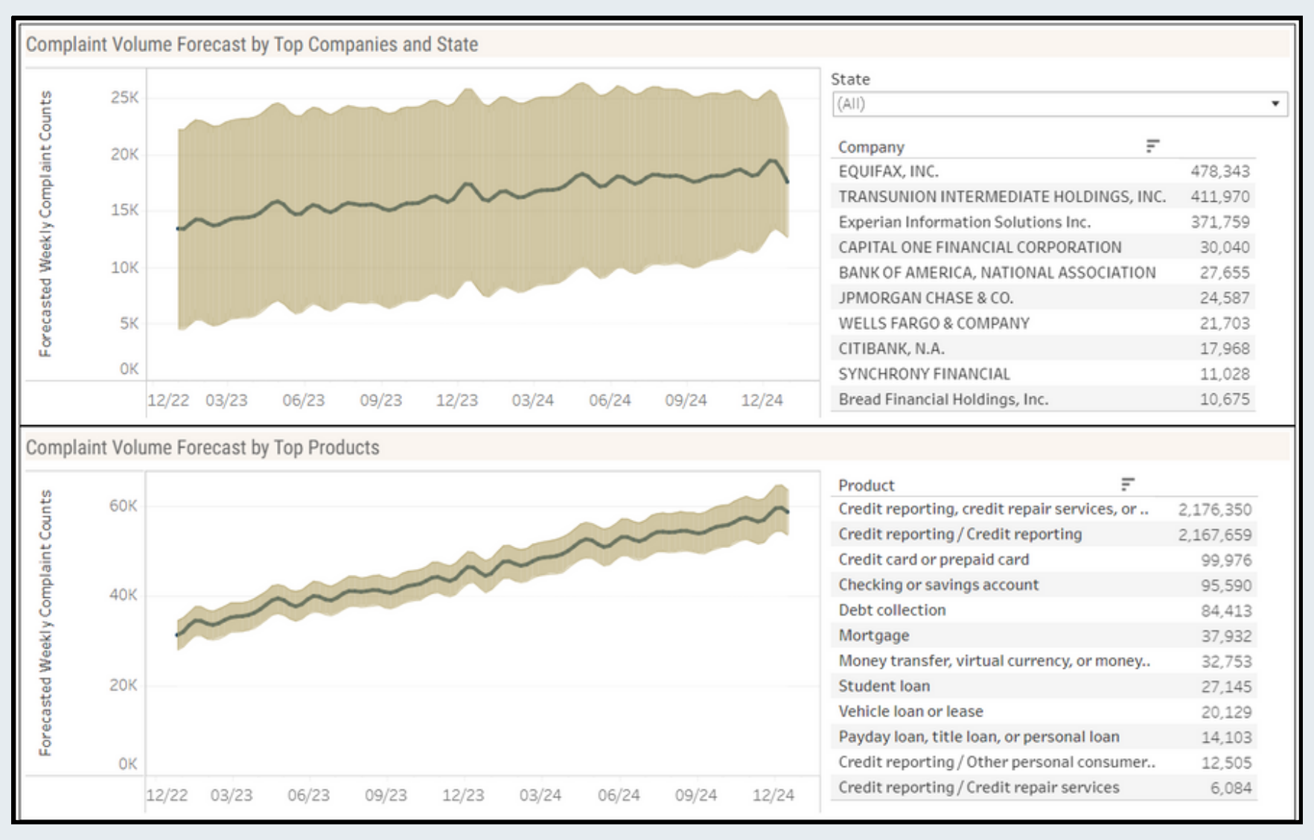
Observations: Combining sentiment scores with demographics and zip codes provides insights into how regional economic conditions could affect perception. A large portion of negative sentiments are from areas receiving public assistance, possibly indicating a disproportionate number lower income populations are affected by actions of financial and credit companies. Analyzing the trends of sentiment scores over time, it is important to note sudden dips for companies like Equifax where these could be attributed to major news events like data leaks or reported scams.


Diverging Sentiment Heat Map with Census Income Attributes
State
GA
Income
Households with Earnings
Households using Public Assistance
Mean Household Income ($)
Avg Compound Sentiment
-0.2303
0.0462
Selected Income Attribute
267
87,171
© 2024 Mapbox © OpenStreetMap

Also notable is the fact that some companies like Bank of America have a consistent negative sentiment This data can used this along with topic data to address areas that require improvement by changing staffing / product policies.

TIME SERIES FORECAST

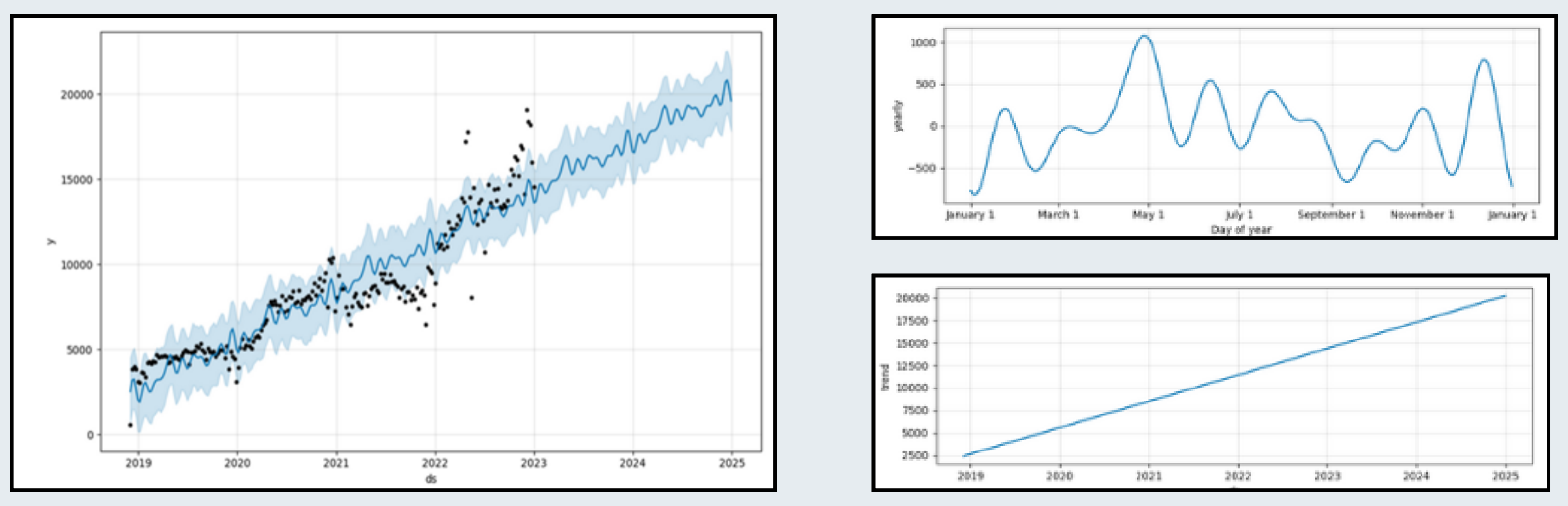
Forecast Weekly Average Complaint Count for Companies across States for Complaint Volume by Product 2023-2025: We achieved best results with Facebook Prophet, given the non-stationary data like consumer complaints.


Complaint Volume Forecast by Top Companies and State
Forecasted Weekly Complaint Counts
0K
5K
10K
15K
20K
25K
12/22
03/23
06/23
09/23
12/23
03/24
06/24
09/24
12/24
State
(All)
Complaint Volume Forecast by Top Products
0K
20K
40K
60K

Company	
EQUIFAX, INC.	478,343
TRANSUNION INTERMEDIATE HOLDINGS, INC.	411,970
Experian Information Solutions Inc.	371,759
CAPITAL ONE FINANCIAL CORPORATION	30,040
BANK OF AMERICA, NATIONAL ASSOCIATION	27,655
JPMORGAN CHASE & CO.	24,587
WELLS FARGO & COMPANY	21,703
CITIBANK, N.A.	17,968
SYNCHRONY FINANCIAL	11,028
Bread Financial Holdings, Inc.	10,675

Product	
Credit reporting, credit repair services, or ..	2,176,350
Credit reporting / Credit reporting	2,167,659
Credit card or prepaid card	99,976
Checking or savings account	95,590
Debt collection	84,413
Mortgage	37,932
Money transfer, virtual currency, or money..	32,753
Student loan	27,145
Vehicle loan or lease	20,129
Payday loan, title loan, or personal loan	14,103
Credit reporting / Other personal consumer..	12,505
Credit reporting / Credit repair services	6,084

Model Optimization and Evaluation: Models were tuned adjusting hyper parameters for number and flexibility. Cross validation withheld a prior year of data and compared predictions to actual values. Best performing model had the lowest RMSE, MAE and MSE.

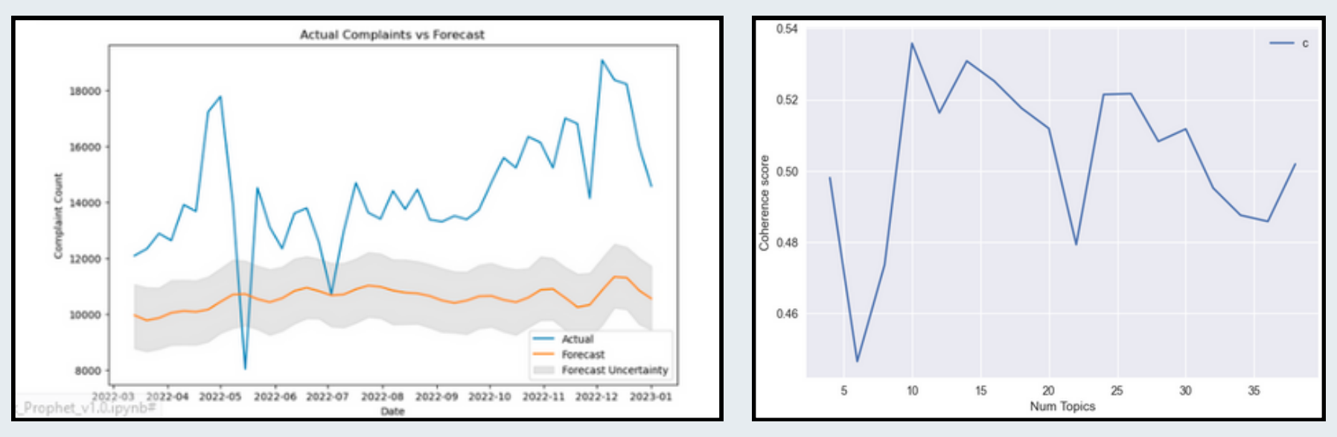


Usefulness of the Findings: The consistent upward slope without any plateaus or dips suggests that the trend in complaints is steadily increasing without any periods of leveling off or decline. Seasonality chart shows the number of complaints has regular patterns within a year, possibly lower complaints at the start and end of the year and peaks at specific times. This pattern suggests some regular annual events or behaviors that influence complaint numbers, potentially attributed to more service use or sales during certain seasons, leading to more complaints. Key learnings from the analysis can inform planning for financial companies, including: (1) staffing levels for financial companies, (2) and understanding of the products that need most attention, and (3) an attempt to proactively improve products and services that are most problematic for customers.

EXPERIMENTS AND RESULTS

Experiments and Evaluation Metrics: Time series models were evaluated using MAE, MSE, RMSE and cross validating forecasts to the actuals for 2022 (a), topic models by measuring coherency scores (b).

Results: Our Tableau comprehensive data visualization tool improves upon previous visualizations by adding sentiment analyses, bar chart races, and an interactive analysis of the structure of the complaint dataset. This allows forecasting of expected complaints by company and product type from 2023 -2024 showing (1) a clear upward trajectory for most categories of complaints, indicating that financial companies will need to staff-up to be prepared to properly deal with complaints. We also determined (2) highest complaint volume by company (Equifax, Transunion, and Experian), and (3) which types of complaints will be most and least common. Credit reporting complaints are expected to increase by 40% in 2024 as overall credit usage rises, mortgage and debt collection complaints will drop, and student loan complaints (which peaked in 2022 after President Biden's attempt at loan cancellation and post-COVID payment restarting) will stay below peak levels. We also discovered narratives are often duplicated, assumed in circumstances where an agency aided in the complaint filing, creating less distinct complaint topics which are more connected with each other, compared to unique complaints narratives which are less well connected.



(a) forecasting vs actual complaint volume (b) coherency scores

Comparative Advantage of Our Analytical Techniques: Our approach yields more granular forecasts and deeper insights into the volume of future complaints, including highest complaint topics, and sentiments driving consumer complaints, all which provide actionable intelligence for financial companies. Our methods demonstrate a superior capacity for capturing nuanced patterns in consumer data, outperforming standard methods by integrating advanced natural language processing and demographic analysis.