Addition Financial Competition 2024 - 2025

UCF Department of Statistics and Data Science

January 13, 2025

Group 10

Katelyn Choudhari and Jackson Windhorst

Overview

Addition Financial aims to understand the key factors influencing consumer decisions when selecting a financial institution for mortgage loans. By analyzing statewide data for Florida, from a variety of sources, aspects such consumer data, institutional data and consumer complaints were examined to determine the results laid out in this report. From this exploration, Addition Financial seeks to understand factors for consumer choice, insights that may drive strategic planning, enhance competitiveness in the mortgage market and promote sustainable growth while supporting the financial well-being in the communities it serves.

The decision-making process in mortgage lending is complex, influenced by a range of applicant, institution, loan and geographic factors. Understanding what drives these decisions is critical for improving transparency and fairness in the lending industry. This report analyzes the multiple recent datasets to identify the key determinants of consumer choice. By applying in depth analysis we plan to uncover some patterns in consumer choice.

Problem:

To efficiently target key information a research problem was created for starting analysis:
What causes consumers to choose one financial institution over another?

Before diving straight into analysis and research we decided that it would have been better if we approached the problem via a divide and conquer technique, simply put we chose to have two main models - loan details and consumer sentiment (specifically mortgages). We also included further analysis into geographic data that could be useful for future models.

Data Collection and Preprocessing

Outlined below are the datasets and Python files used and created in this analysis; as well as an overview of the preprocessing steps taken to ensure a usable dataset for each model.

Dataset and Model 1:

- modeling.csv
- loan_details.ipynb

This dataset was sourced from the HMDA database and was filtered for just Florida via the website API tool. For this research topic we are currently only interested in financial institutions from the state of Florida since the competition is hosted by a Florida financial institution.

For processing the dataset the csv was loaded into a pandas dataframe and some data cleaning was performed to remove irrelevant features. To start, all columns with 80% or more missing values (e.g. NAs, NaNs, NULLs) were removed along with any duplicate rows. Some columns were also removed but only based on the context they had to the overall dataset. State code was removed since the dataset was just for Florida, and activity year was removed since the dataset is only for 2023.

Multiple years were not used since we did not want to perform a time series analysis. Census tract was also removed but only because a county code was in the same dataset; and the census tract variable acts in the same format as the county code where this code shows data based on a certain region.

Some further processing was done such as encoding lei, or legal entity identifier, with the actual bank institution name and then assigning a number based on highest frequency. To elaborate, the bank with the highest frequency of customers was assigned the code 1 and so on.

Data processing was also done for columns that were objects (strings). Since the model will most likely require numerical values only NAs had to be filled with 0s and type casted as floats. There were also columns that contained different ranges such as debt to income ratio and age to name a few that needed to be encoded. All relevant codes are provided in the appendix (FIX THIS).

Dataset and Model 2:

- CFPB_Complaints_FL.csv
- CFPB_Complaints_FL_Model.ipynb

The complaints dataset was gathered via the Consumer Financial Protection Bureau's website, specifically the Consumer Complaint Database. The dataset was filtered to only include the product type of mortgages within the state of Florida.

Feature selection was done to select predictors that were only complaint or loan related. In addition to geographic and date data. The target variable set was the mortgage type, or

sub-product type. The most indicative predictors were: issue, sub-issue and consumer complaint narrative.
To preprocess the data typical methods of dealing with null or missing values were applied and zip codes were standardized. In addition, functions were created to clean and process the consumer complaint narrative column in the dataset. This column was cleaned using tokenization, lemmatization and stopword removal.

Lastly, a function was created to add two sentiment scores to the cleaned dataset. These scores were sentiment polarity and sentiment subjectivity; achieved via the utilization of TextBlob. An output file with the clean dataset and sentiment scores was also included for readability.

***Dataset 3 and Geographic Analysis**:*

- Institutions_FL.csv
- Institutions_FL_Model.ipynb

The purpose in establishing this dataset was mainly to visualize institutional data in Florida from a geographic or spatial viewpoint. Therefore, geographical and locational features were maintained only in the dataset. Preprocessing was similar to the first two datasets. Specifically the features latitude and longitude were cleaned to ensure no null values for spatial visualizations. Assessing this type of data can be useful in tandem with analysis from other datasets in order to create a full picture.

Methodology and Exploratory Data Analysis

Model 1:

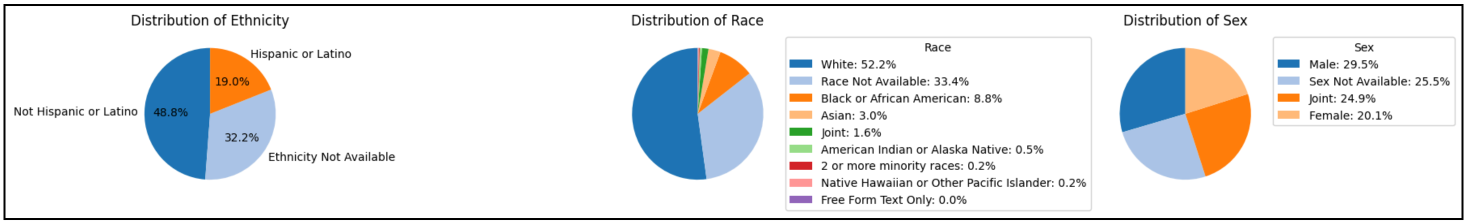
Loan Details: To get a more detailed analysis and understanding of the dataset certain variables were investigated. Details such as consumer demographics and loan characteristics are important for quick visualization to become more familiar with the topic of why a consumer would choose a certain financial institution.

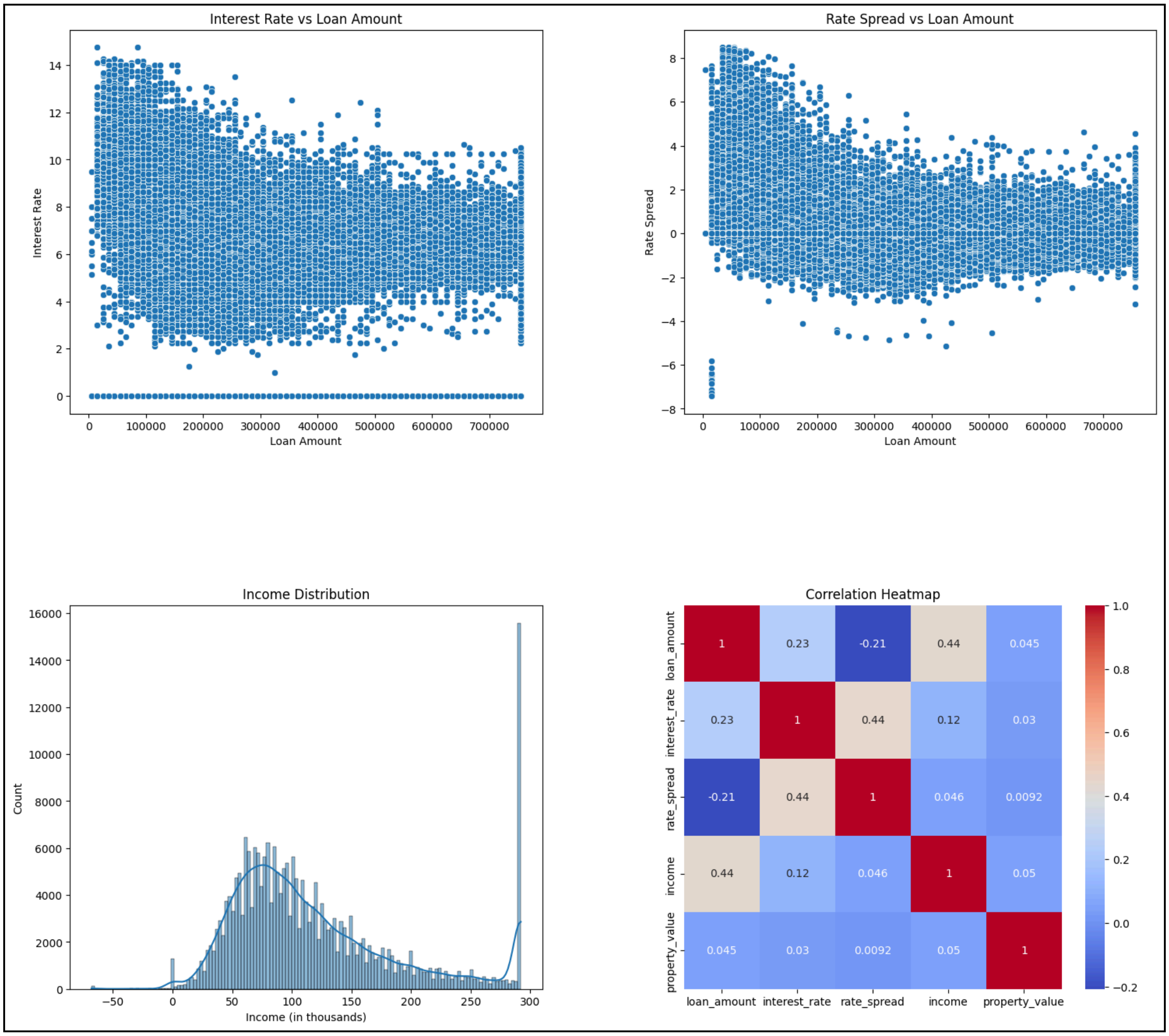
Before running an official model some feature selection needed to be done to choose the most relevant features for a model. For the feature utility metric we decided on the mutual information metric. This is an overall good base metric to choose features since it measures all kinds of relationships compared to something like correlation which only measures linear relationships. When running the feature selection metric a select best k was run with the mutual relationships to choose the top five features; *action_taken, loan_to_value_ratio', interest_rate*, and *submission_of_application.*

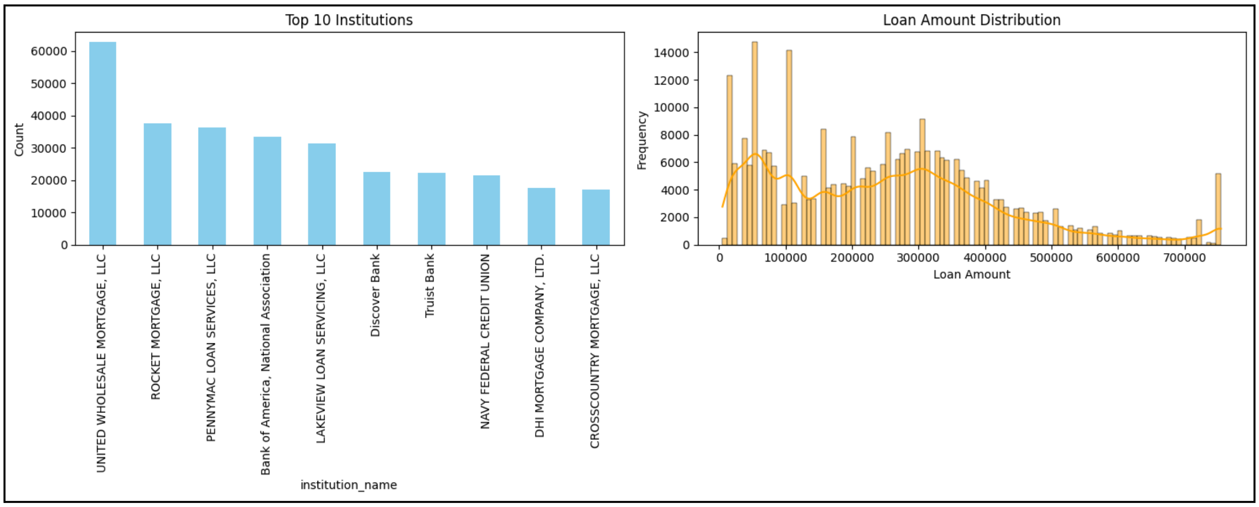
For the model a random forest classification model was chosen. Random forest classification models are great for handling large datasets and are robust to overfitting. This is due to the random forest classification model making multiple decision trees and averaging the predictions of all the trees (hence the forest term).

This ensemble model can also handle numerical and categorical features and show importance of the features. Feature importance is necessary when investigating what influences a consumer's choice when choosing a mortgage.

Although there are disadvantages to the random forest such as being computationally expensive, less interpretability than simpler models it falls within the perfect scope of our project. Simple models (linear regression) can sometimes not capture underlying relationships that more complex models can and extremely complex models (neural networks, transformers) can be hard to interpret.




Interest Rate vs Loan Amount
Interest Rate
0
2
4
6
8
10
12
14
Loan Amount
0
100000
200000
300000
400000
500000
600000
700000
Rate Spread vs Loan Amount
Rate Spread
−8
−6
−4
−2
0
2
4
6
8
Loan Amount
0
100000
200000
300000
400000
500000
600000
700000
Income Distribution
Count
0
2000
4000
6000
8000
10000
12000
14000
16000
Income (in thousands)
−50
0
50
100
150
200
250
300
Correlation Heatmap
loan_amount
interest_rate
rate_spread
income
property_value
1
0.23
-0.21
0.44
0.045
0.23
1
0.44
0.12
0.03
-0.21
0.44
1
0.046
0.0092
0.44
0.12
0.046
1
0.05
0.045
0.03
0.0092
0.05
1
loan_amount
interest_rate
rate_spread
income
property_value
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Top 10 Institutions
Count
0
10000
20000
30000
40000
50000
60000
UNITED WHOLESALE MORTGAGE, LLC
ROCKET MORTGAGE, LLC
PENNYMAC LOAN SERVICES, LLC
Bank of America, National Association
LAKEVIEW LOAN SERVICING, LLC
Discover Bank
Truist Bank
NAVY FEDERAL CREDIT UNION
DHI MORTGAGE COMPANY, LTD.
CROSSCOUNTRY MORTGAGE, LLC
institution_name
Loan Amount Distribution
Frequency
0
2000
4000
6000
8000
10000
12000
14000
Loan Amount
0
100000
200000
300000
400000
500000
600000
700000

Model 2:

Consumer Sentiment: Includes a Naives Bayes model with classification report, sentiment analysis and keyword and loan type analysis from complaints by consumers sourced from the Consumer Protection Financial Bureau. Various visualization and report formats were used to display the data, including, charts, a heatmap, a count report, word clouds and time series analysis of when complaints were made.

Given the nature of the dataset, a Naive Bayes model was first decided to be implemented. This method can be beneficial for tasks involving classification and text analysis, particularly because Naive Bayes excels in handling textual data. However, it should be noted that it is precisely that – naive.

Yet, this method was chosen as it lends itself to further sentiment analysis, topic classification and filtering duplicate complaints; as well as being simple and efficient for a task with a large amount of data. The model was traditionally split into training and testing sets and took advantage of the technique TF-IDF Vectorizer, which measures the originality of a word by comparing the frequency that a word occurs within a text. The model was evaluated with an accuracy percentage and classification report.

Further sentiment analysis was conducted to classify complaints as positive, neutral or negative. This can assist in prioritizing complaints and identifying improvement areas. Other visualization methods were used to represent sentiment trends, including a boxplot to show sentiment distribution across loan types and a heatmap that demonstrates the relationship between loan type and keywords.

Word clouds by loan type were created to help visualize top words or phrases. Basic bar and pie charts were incorporated to represent the proportion of complaints by loan type and top keywords across all complaints. Lastly, a top complaint issues bar graph and historical outlook of complaints overtime line chart was incorporated.

***Geographic Analysis**:*

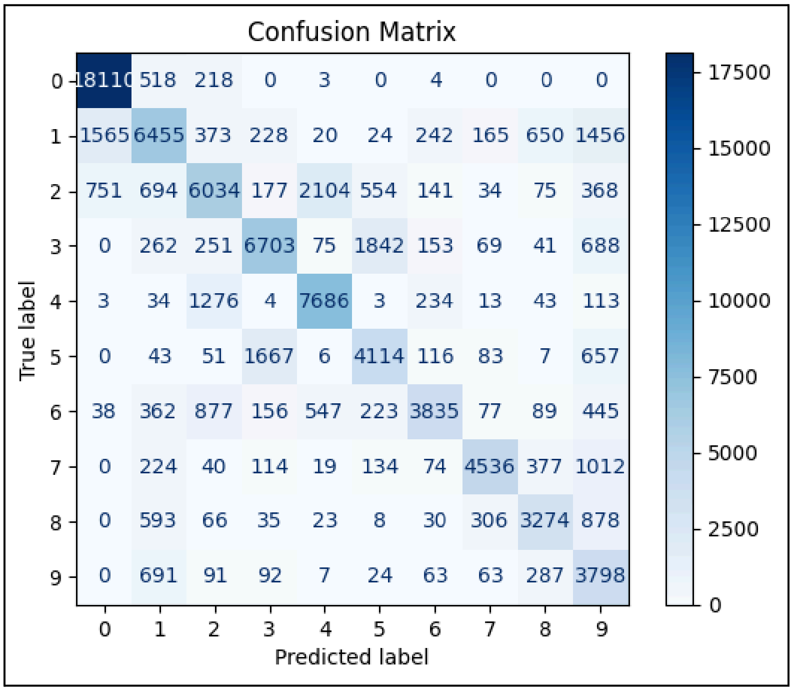
The model includes mainly an EDA via visualizations related to financial institutions located in Florida. An interactive marker map, heatmap of institutions locations, choropleth map; as well as more standard bar and line graphs were created to provide an understanding of the data.

Specifically, the folium library was used to create a map centered on Florida and markers for each institution with tooltips and pop-ups displaying key information were made. Juxtaposition to that is a heatmap that visualizes the density of institutions across Florida; darker areas indicating more institutions. With that, a choropleth map showing the number of institutions by county was added to show the spread of institutions across regions of Florida.

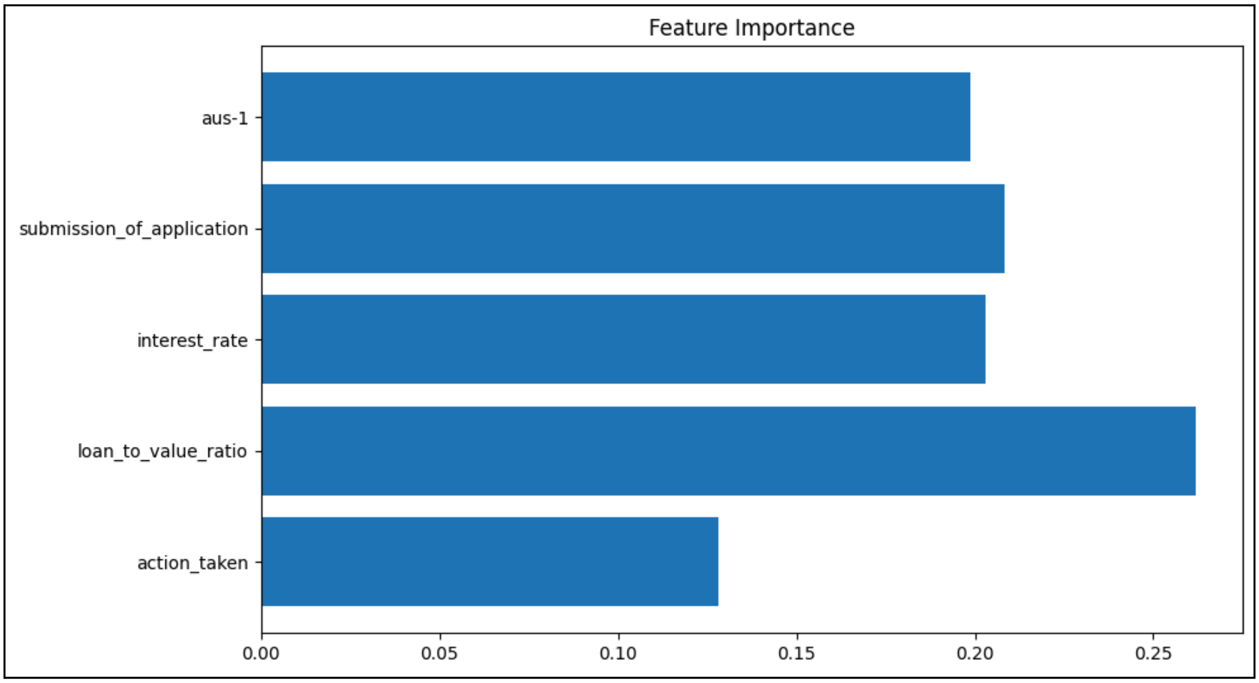
Results

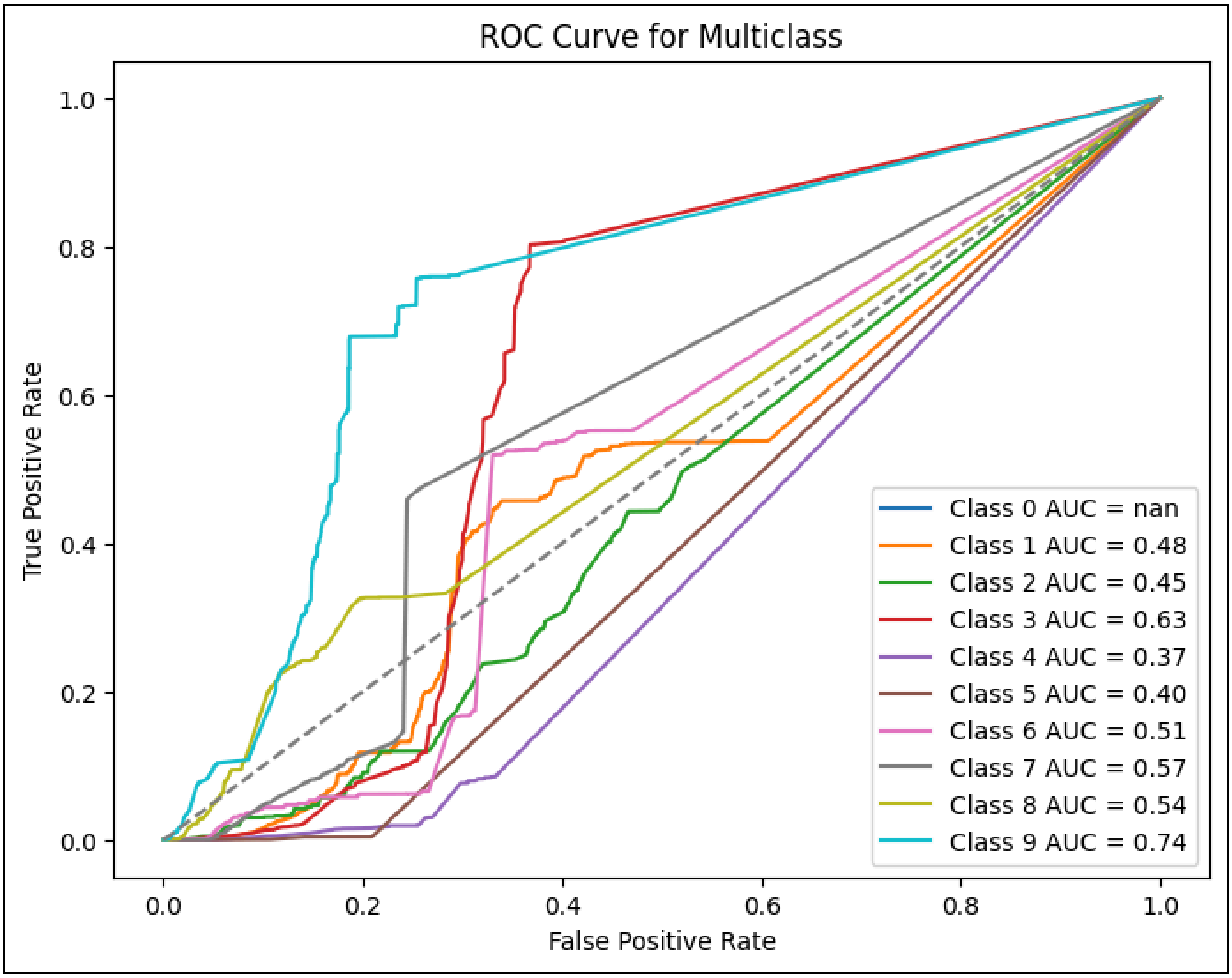
Model 1:

The Random Forest Classification model achieved an overall accuracy of 71.16%, and performed well on the classes with the highest frequency. This is to be expected since the top class is considered the most 'popular bank' among consumers. The macro average had a precision of 0.70, recall of 0.68 and F1-score of 0.68, suggesting a moderate performance on average across all classes. For the weighted average precision was 0.72, recall: 0.71 and F1-score: 0.71 indicating a relatively good performance. Keep in mind that the weighted average gives weight to classes with more occurrences. In the context of the research question the main features; loan to value ratio, action taken, submission of application, interest rate, automated underwriting system seem to contribute significantly to why consumers choose certain financial institutions over others.



	0	1	2	3	4	5	6	7	8	9
0	18110	518	218	0	3	0	4	0	0	0
1	1565	6455	373	228	20	24	242	165	650	1456
2	751	694	6034	177	2104	554	141	34	75	368
3	0	262	251	6703	75	1842	153	69	41	688
4	3	34	1276	4	7686	3	234	13	43	113
5	0	43	51	1667	6	4114	116	83	7	657
6	38	362	877	156	547	223	3835	77	89	445
7	0	224	40	114	19	134	74	4536	377	1012
8	0	593	66	35	23	8	30	306	3274	878
9	0	691	91	92	7	24	63	63	287	3798

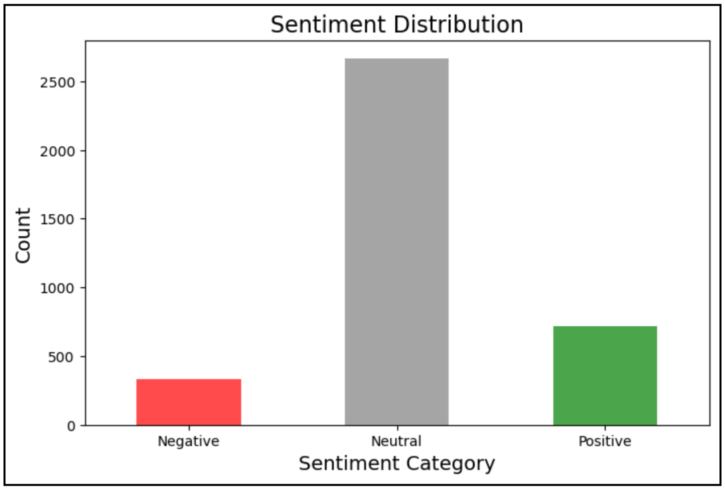

Feature Importance
aus-1
submission_of_application
interest_rate
loan_to_value_ratio
action_taken
0.00
0.05
0.10
0.15
0.20
0.25

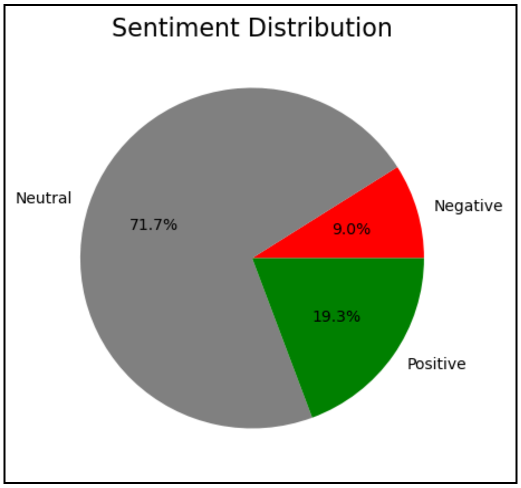

ROC Curve for Multiclass
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 AUC = nan
Class 1 AUC = 0.48
Class 2 AUC = 0.45
Class 3 AUC = 0.63
Class 4 AUC = 0.37
Class 5 AUC = 0.40
Class 6 AUC = 0.51
Class 7 AUC = 0.57
Class 8 AUC = 0.54
Class 9 AUC = 0.74

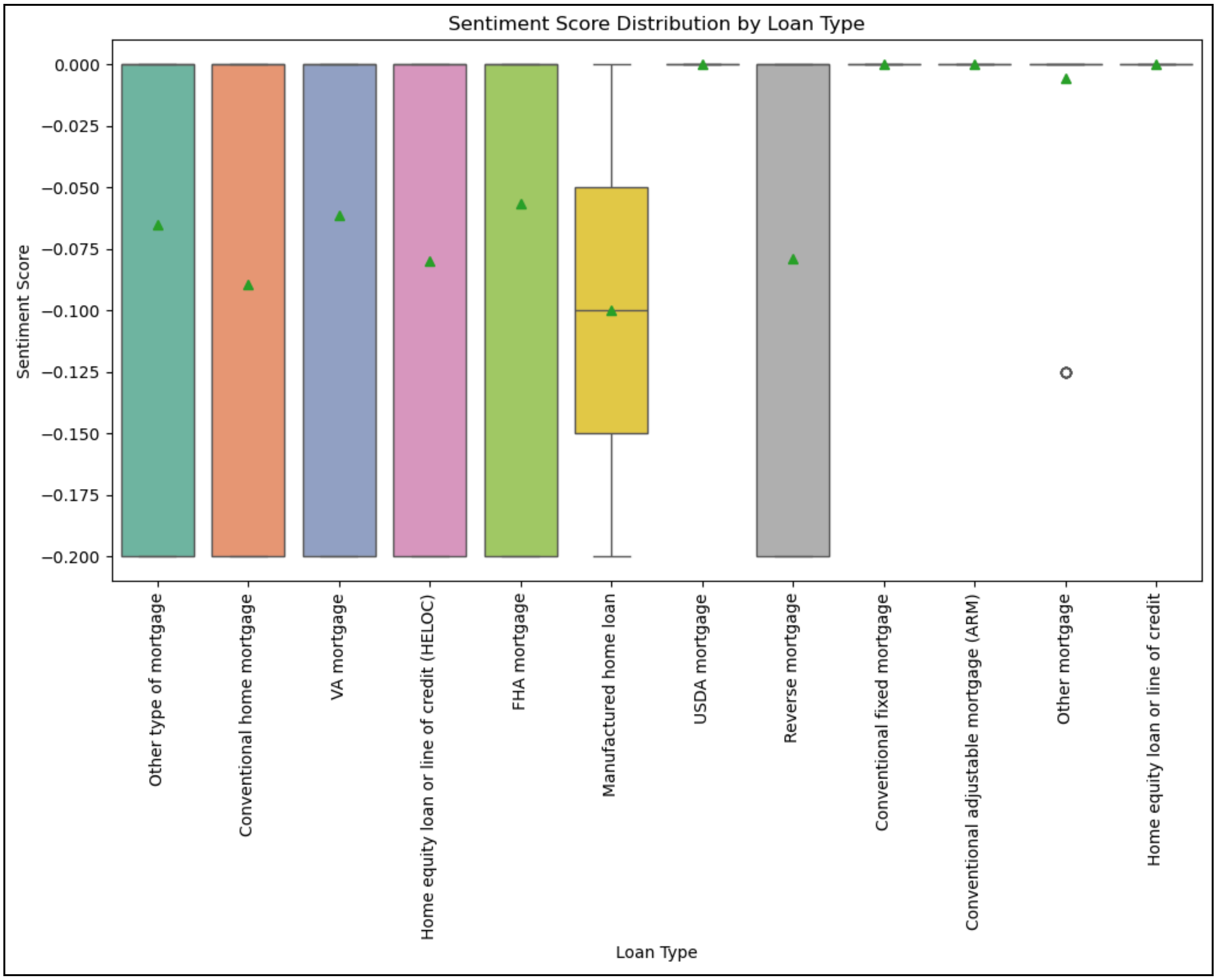
Model 2:

The Naive Bayes model achieved an overall accuracy of 47.04%, performing well for larger categories like conventional home mortgages (F1-score: 0.69) due to their distinct features and higher representation. However, it struggled with smaller, underrepresented classes (reverse mortgage and USDA mortgage), resulting in zero precision, recall and F1-scores for many categories. The model's reliance on feature independence and imbalance data led to frequent misclassifications, highlighting the need for improved data balance and more advanced modeling techniques. Nonetheless, creating this model was a good starting point for analysis of text.

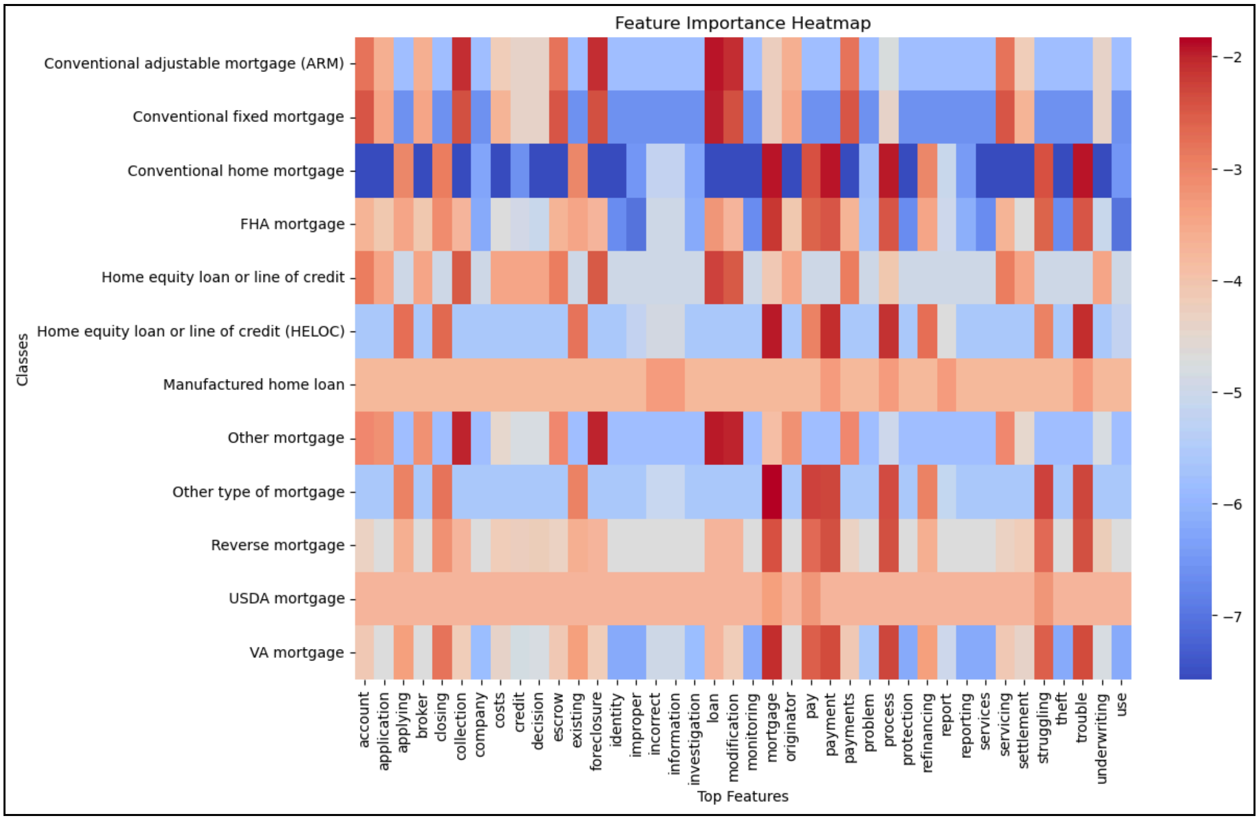
For sentiment analysis, some observations were that conventional fixed mortgages showed neutral sentiments, whereas FHA and VA mortgages exhibited more negative polarity, indicating customer dissatisfaction. Products like FHA and VA mortgages may require operational or policy reviews to address recurring complaints.

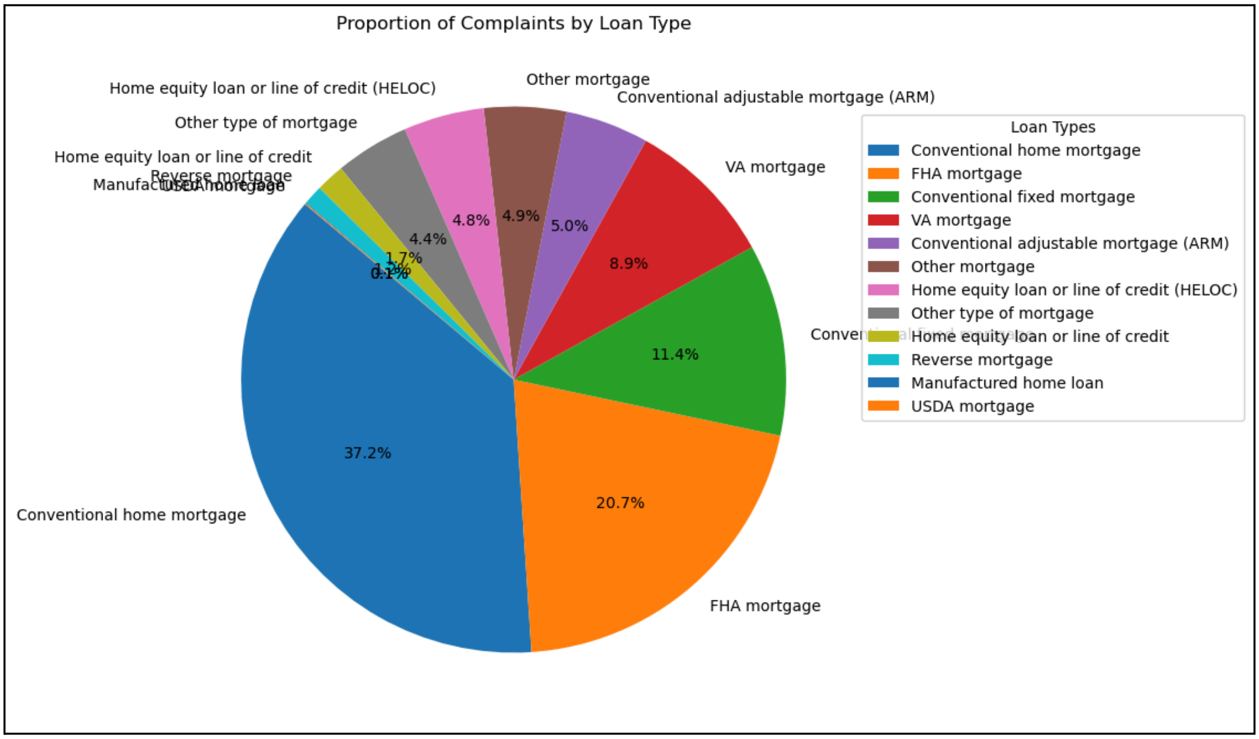


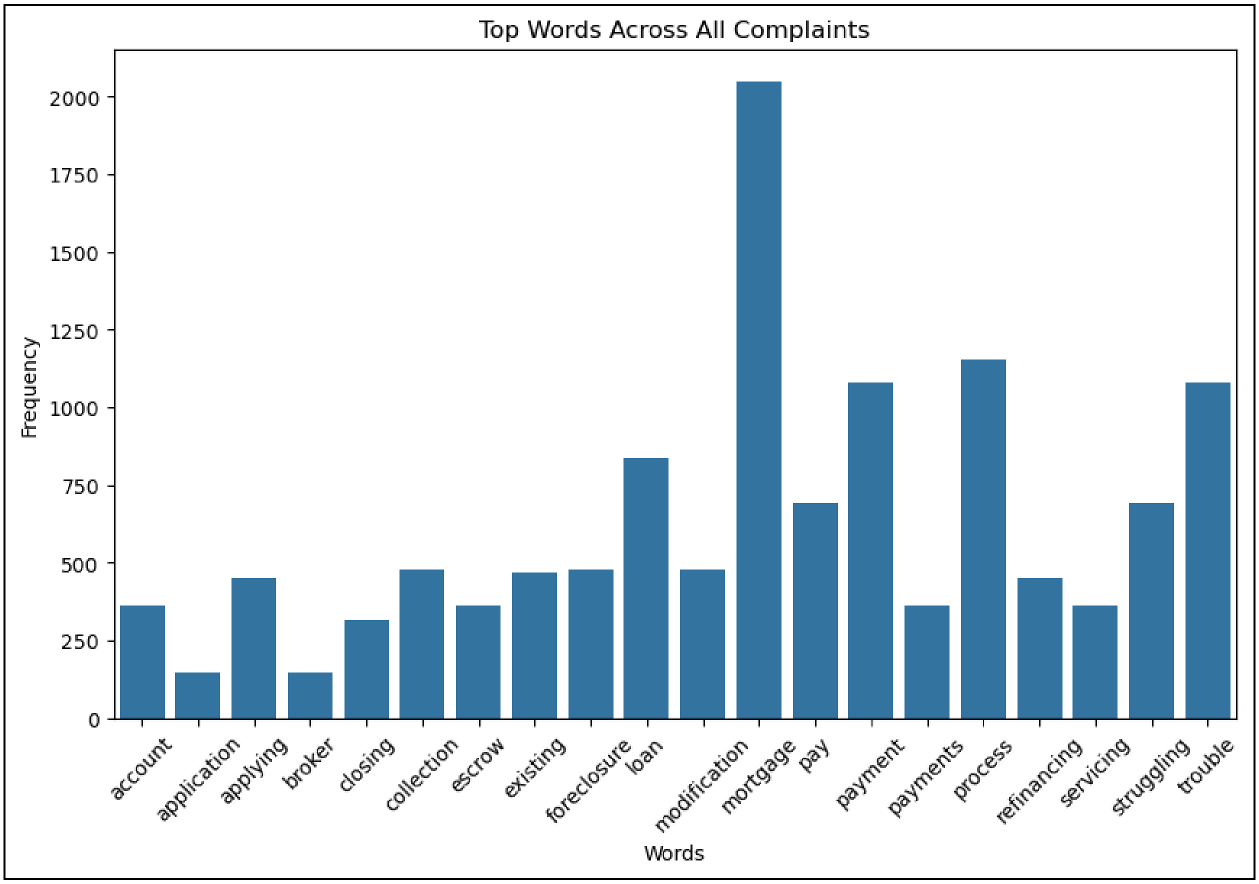


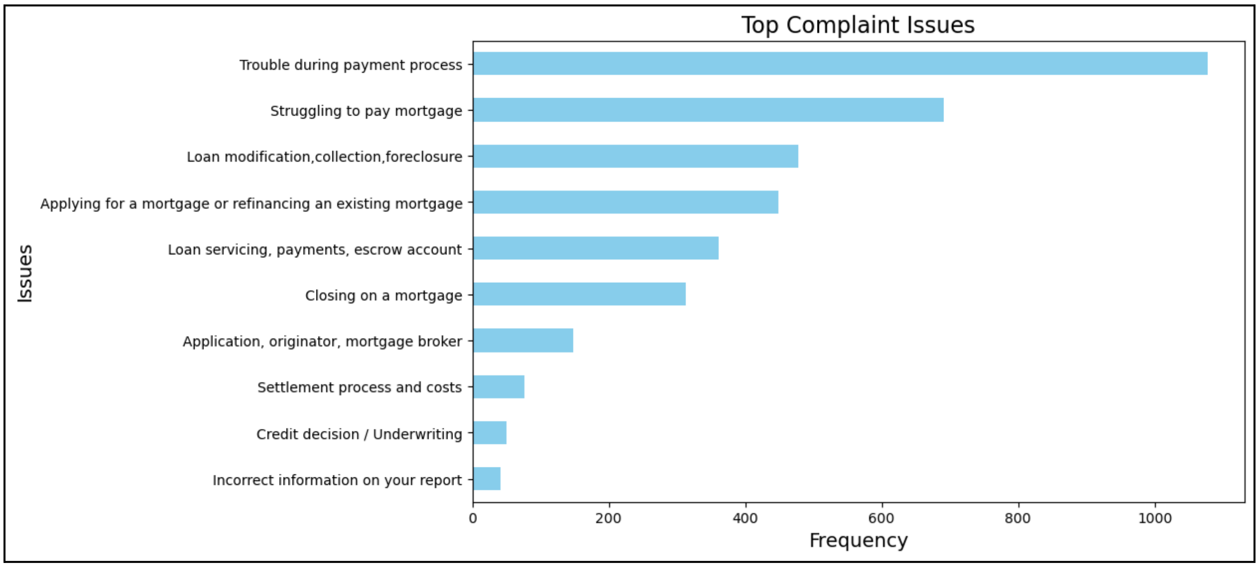


The heatmap visualized the importance of individual words, or features, across different complaint categories, sub-product or loan type, as determined by the log probabilities in the Naive Bayes model. Darker colors in the heatmap indicate higher importance of a word for a specific class. For example, words that frequently appeared in the complaints were: interest, loan, credit, fixed, adjustable. It should be noted that due to prior misclassifications from the Naive Bayes model there is feature overlap in the heatmap which may cause overlap in word importance across categories.

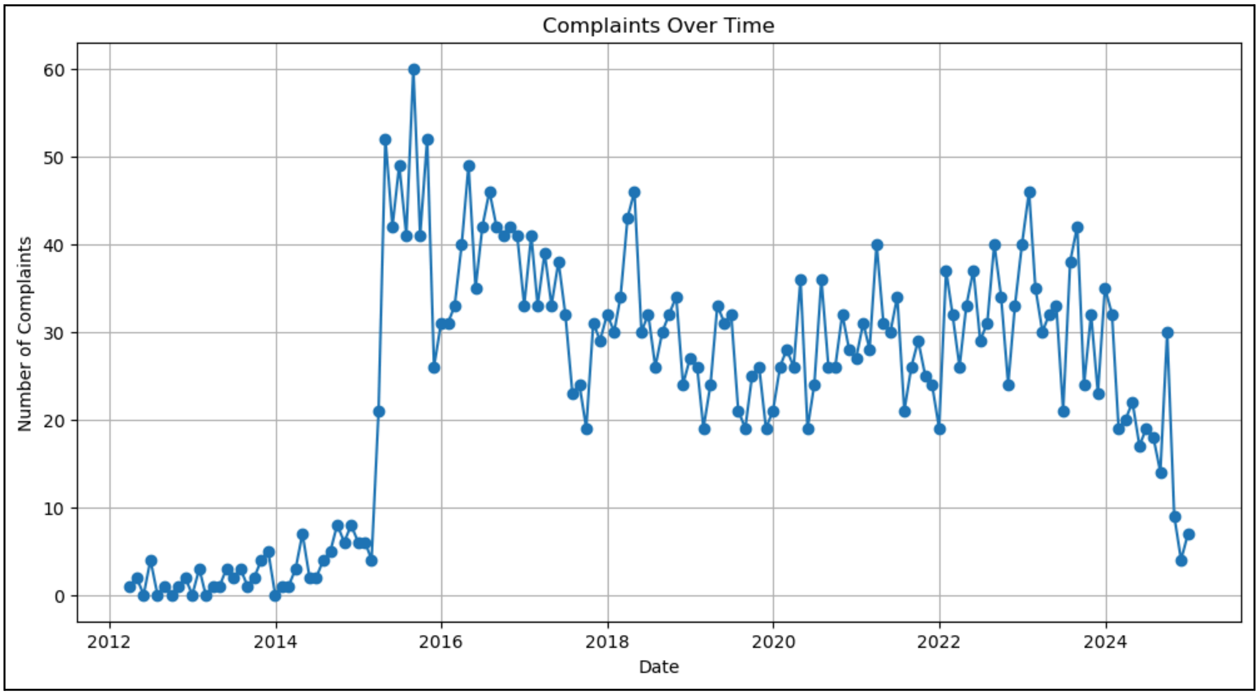

Feature Importance Heatmap
Classes
Conventional adjustable mortgage (ARM)
Conventional fixed mortgage
Conventional home mortgage
FHA mortgage
Home equity loan or line of credit
Home equity loan or line of credit (HELOC)
Manufactured home loan
Other mortgage
Other type of mortgage
Reverse mortgage
USDA mortgage
VA mortgage
account
application
applying
broker
closing
collection
company
costs
credit
decision
escrow
existing
foreclosure
identity
improper
incorrect
information
investigation
loan
modification
monitoring
mortgage
originator
pay
payment
payments
problem
process
protection
refinancing
report
reporting
services
servicing
settlement
struggling
theft
trouble
underwriting
use
Top Features
−2
−3
−4
−5
−6
−7

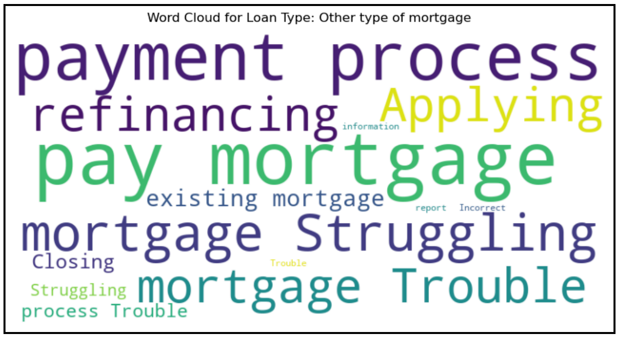

Proportion of Complaints by Loan Type
Home equity loan or line of credit (HELOC)
Other mortgage
Conventional adjustable mortgage (ARM)
Other type of mortgage
VA mortgage
Home equity loan or line of credit
Reverse mortgage
Manufactured home loan
USDA mortgage
4.8%
4.9%
5.0%
4.4%
1.7%
1.2%
0.1%
0.1%
8.9%
11.4%
37.2%
20.7%
Conventional fixed mortgage
Conventional home mortgage
FHA mortgage
Loan Types
Conventional home mortgage
FHA mortgage
Conventional fixed mortgage
VA mortgage
Conventional adjustable mortgage (ARM)
Other mortgage
Home equity loan or line of credit (HELOC)
Other type of mortgage
Home equity loan or line of credit
Reverse mortgage
Manufactured home loan
USDA mortgage

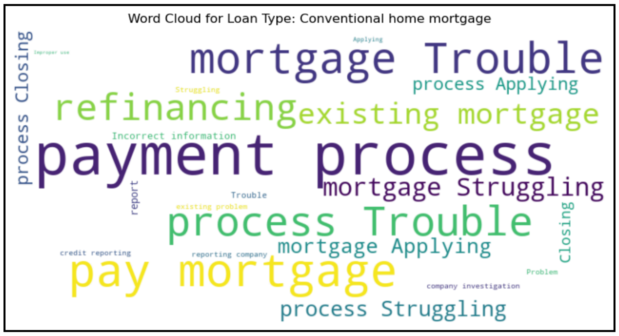


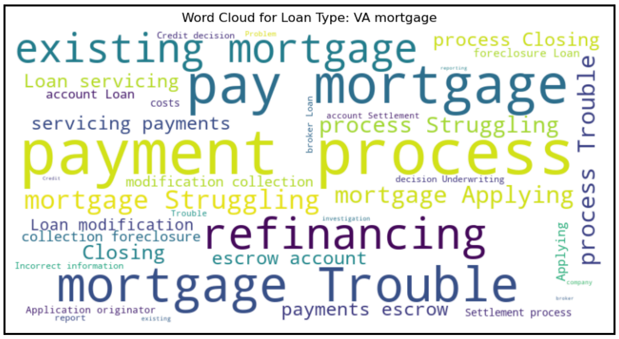


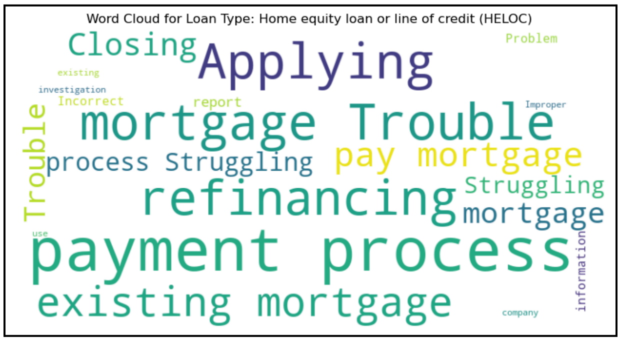
Similar to the heatmap, the word clouds for each class show more prevalent words as larger. Although this is not a detailed analysis, it is a quick indication into the more frequent words customers mention in their complaints. Likewise, the Complaints Over Time line chart gives a high level indication as to which years had more frequent complaints. This could be due to a variety of factors, economic or business specific; however it may be telling for certain financial institutions based on their own dealings.

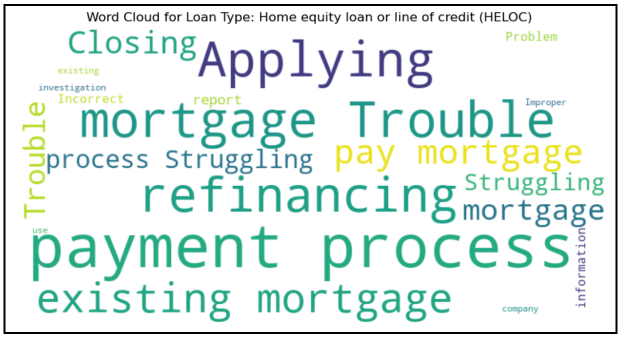

Complaints Over Time
Number of Complaints
0
10
20
30
40
50
60
Date
2012
2014
2016
2018
2020
2022
2024

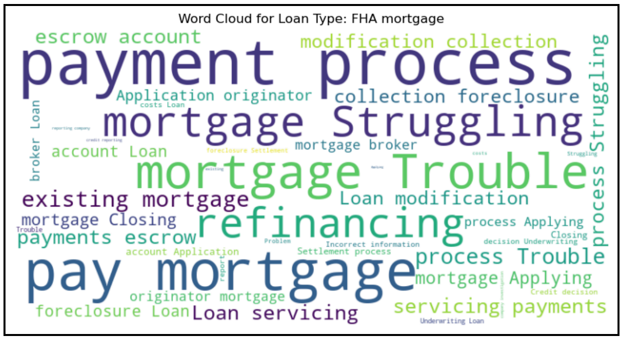

Word Cloud for Loan Type: Other type of mortgage
payment process
refinancing
Applying
information
pay mortgage
existing mortgage
report
Incorrect
mortgage Struggling
Closing
Trouble
Struggling
mortgage Trouble
process Trouble

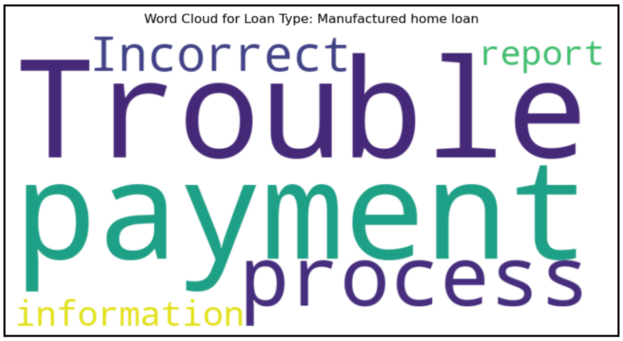

Word Cloud for Loan Type: Conventional home mortgage
process Closing
Applying
mortgage Trouble
process Applying
Struggling
refinancing
existing mortgage
Incorrect information
payment process
Trouble
report
mortgage Struggling
existing problem
process Trouble
Closing
credit reporting
reporting company
mortgage Applying
pay mortgage
Problem
company investigation
process Struggling

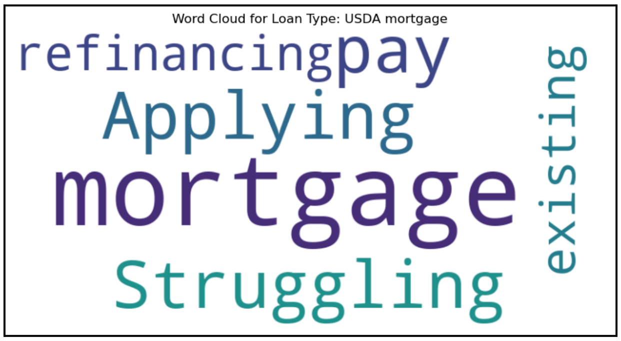

Word Cloud for Loan Type: VA mortgage
existing mortgage
Credit decision
Problem
process Closing
foreclosure Loan
reporting
Loan servicing
account Loan
pay mortgage
costs
broker Loan
account Settlement
servicing payments
process Struggling
process Trouble
payment process
Credit
modification collection
decision Underwriting
mortgage Struggling
mortgage Applying
Trouble
investigation
Loan modification
collection foreclosure
refinancing
Applying
Closing
escrow account
company
Incorrect information
broker
mortgage Trouble
Application originator
report
existing
payments escrow
Settlement process

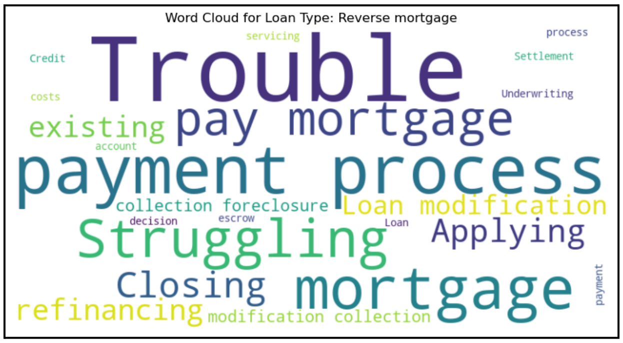

Word Cloud for Loan Type: Home equity loan or line of credit (HELOC)
Closing
Applying
Problem
existing
investigation
Incorrect
report
Improper
Trouble
mortgage Trouble
process Struggling
pay mortgage
Struggling
refinancing
mortgage
use
payment process
information
existing mortgage
company

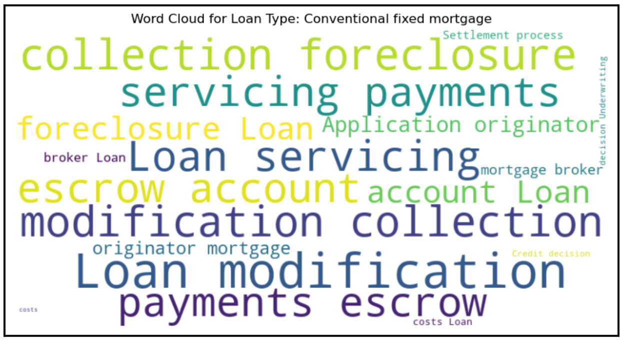

Word Cloud for Loan Type: Home equity loan or line of credit (HELOC)

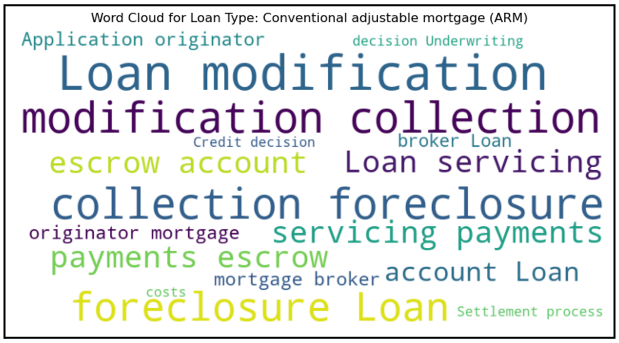

Word Cloud for Loan Type: FHA mortgage

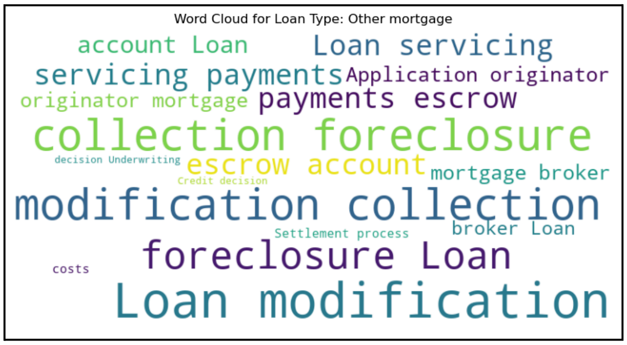

Word Cloud for Loan Type: Manufactured home loan

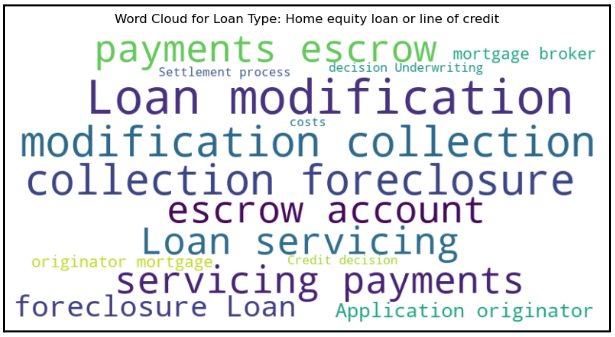

Word Cloud for Loan Type: USDA mortgage


Word Cloud for Loan Type: Reverse mortgage


Word Cloud for Loan Type: Conventional fixed mortgage

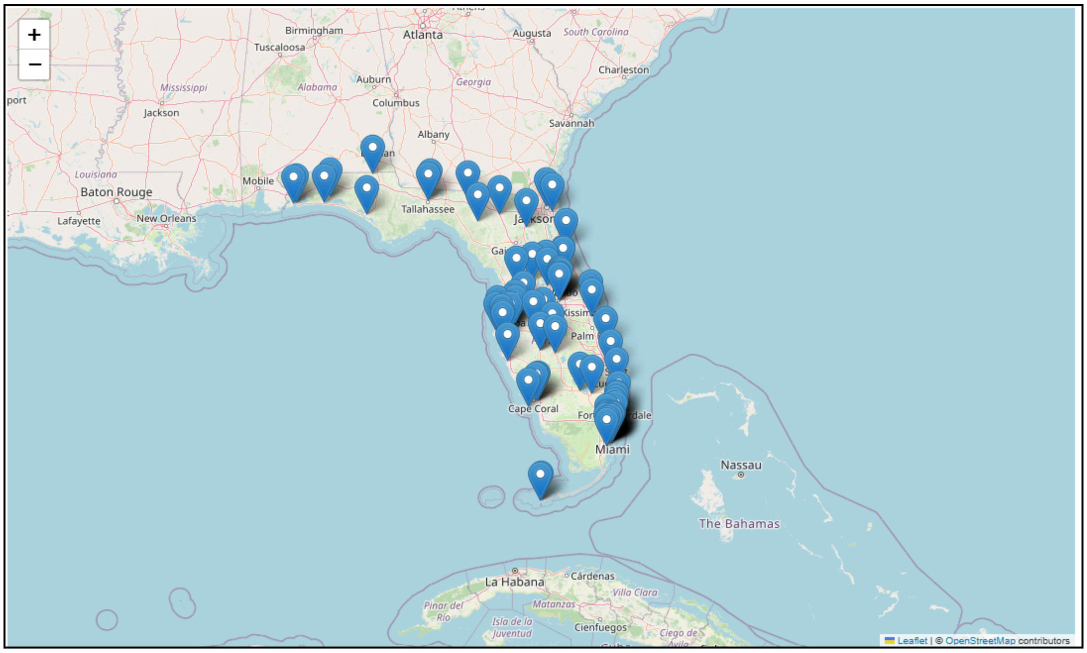

Word Cloud for Loan Type: Conventional adjustable mortgage (ARM)

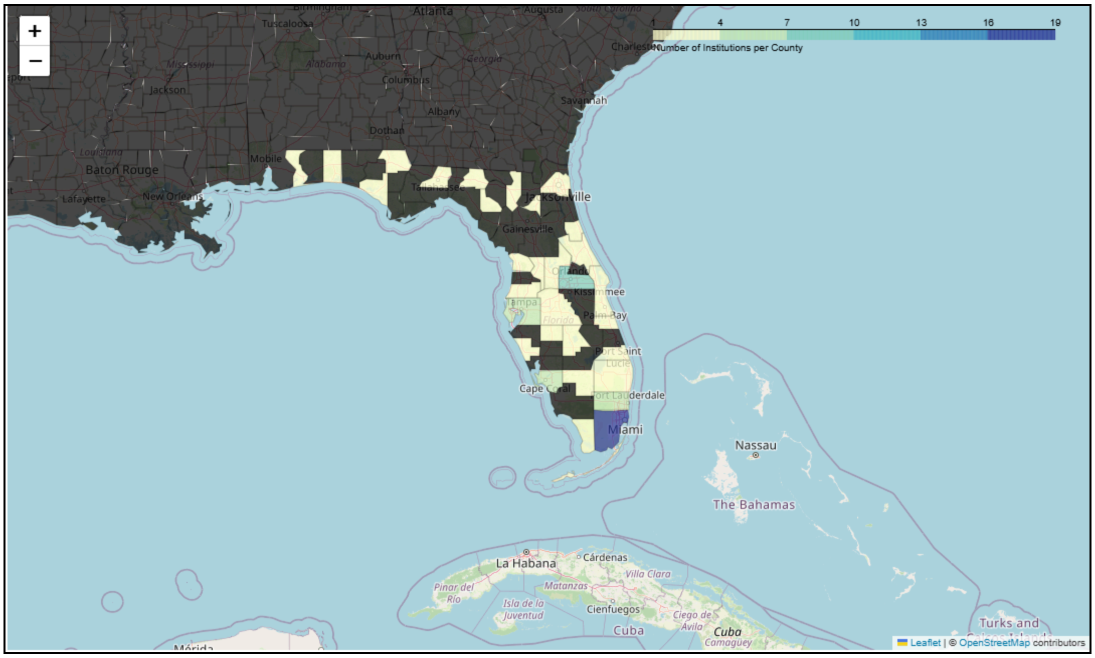

Word Cloud for Loan Type: Other mortgage


Word Cloud for Loan Type: Home equity loan or line of credit

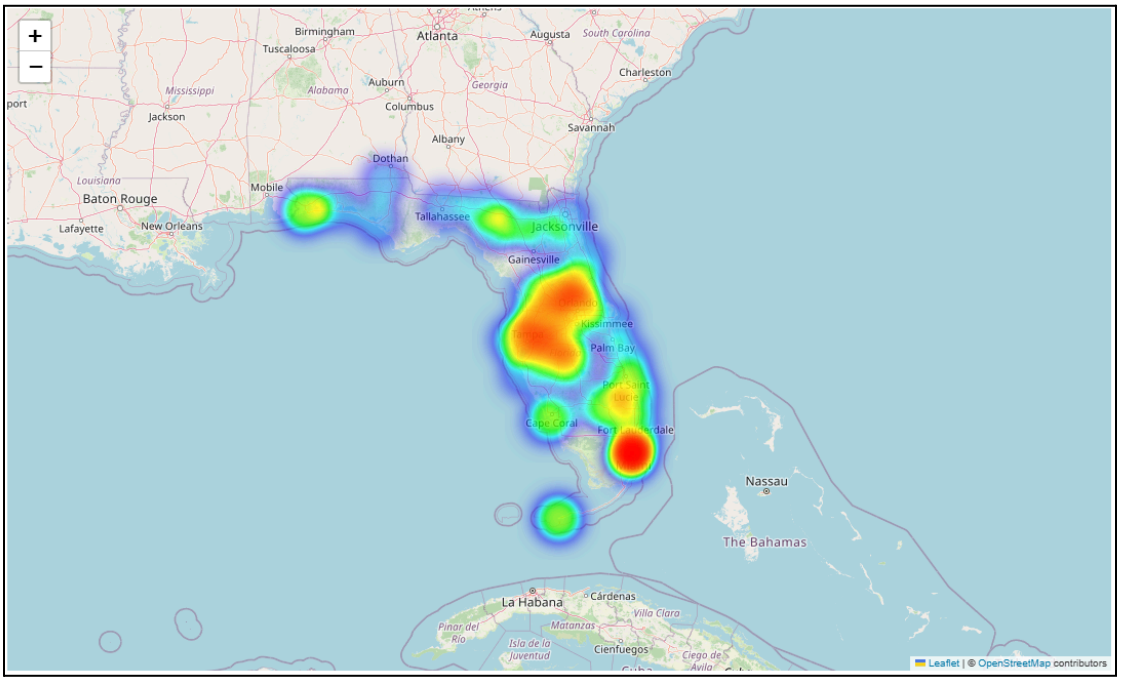
Geographic Analysis

Regarding the interactive marker map, major clusters of institutions are seen in urbanized regions such as Miami-Dade, Broward and Orange counties. Sparse markers in rural countries like Liberty and Lafayette indicate fewer financial institutions, suggesting limited access in these areas.



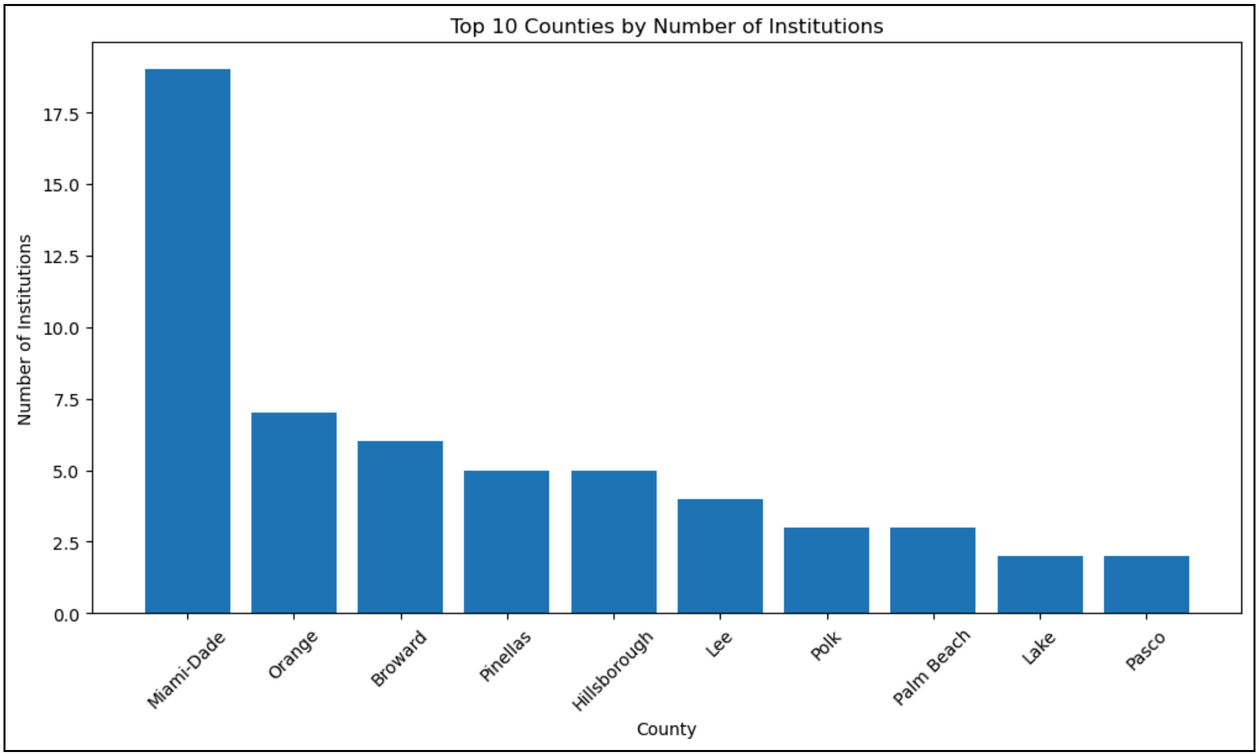


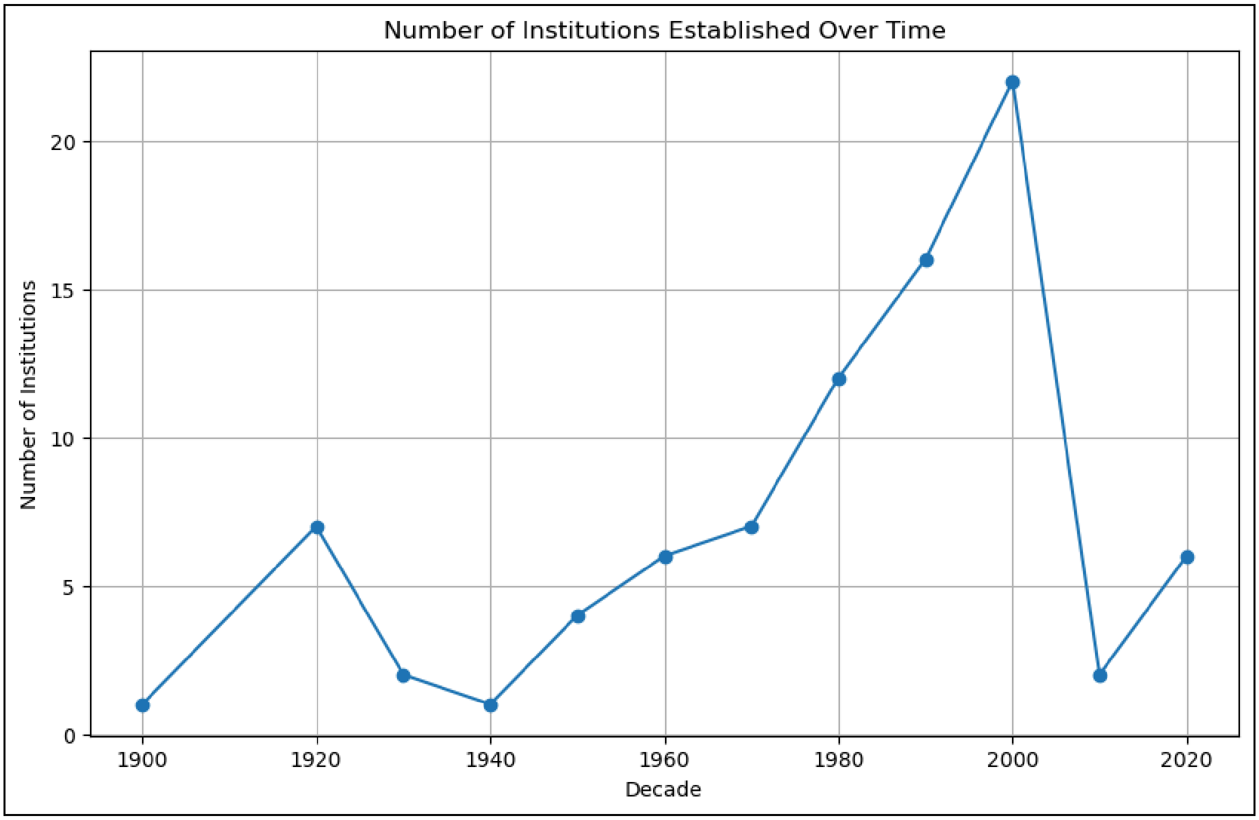
For the heatmap of institutions, the highest density regions overlap with metropolitan areas, particularly South and Central Florida - not surprisingly. This may indicate competitive markers, suggesting a focus on differentiating serves rather than increasing physical appearance. Unlike more northern and rural regions which could be opportunities for growth through placement of new branches or outreach.



Combining sentiment analysis with this spatial data, demonstrates that countries with higher institutional density show more neutral to positive sentiments due to better resource availability. Transposed, rural countries exhibit more negative sentiments, reflecting dissatisfaction with accessibility or service quality.

Overall, these visualizations can be clues as to what competition exists in which regions of Florida for financial institutions. It also can also be telling of the population and economic activity of the area; which could help institutions determine how to market to demographics by county. Or, it highlights potential regions for expansion or increased digital banking services in underserved rural areas. Assumingly, consumers tend to make decisions based on proximity and recognizability – if they know you and have seen you, they are more likely to consider your services. However, this can only be surmised on a surface level based on the data thus far. Other financial factors as discussed in Model 1 could have larger impacts in consumer decisions.


Top 10 Counties by Number of Institutions
Number of Institutions
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Miami-Dade
Orange
Broward
Pinellas
Hillsborough
Lee
Polk
Palm Beach
Lake
Pasco
County


Number of Institutions Established Over Time
Number of Institutions
0
5
10
15
20
1900
1920
1940
1960
1980
2000
2020
Decade

Limitations

Given the high dimensionality of the datasets, the computational time was extreme for all models. In order to address this, more simplistic methods and approaches were applied across the three models. As previously stated, features needed to be preprocessed and transformed to hone in on the most important predictors. Due to this limitation, the overall conclusions that could be drawn from each model to answer the initial problem, is only indicative to a degree of what could computationally be achieved.

During this project, unforeseen challenges also arose related to task distribution and execution. While initial responsibilities were delegated among team members, certain contributions were not completed as planned. As a result, a significant portion of the work was managed by two team members, which constrained the time and resources available for exploring advanced modeling techniques or optimizing performance.

These limitations affected the depth of analysis and the scope of model evaluation, but regardless the project was prioritized in order to deliver a functional and coherent solution. This showcases the importance of adaptability and perseverance in overcoming obstacles outside of the data.

Conclusion

Based on the information from all datasets, several factors likely influence a consumers decision when choosing a financial institution for mortgages.

However, the top five factors, based off this data analysis, that consumers are likely to choose a financial institution are:

1. ***Accessibility and convenience*** - being near branches or availability of online services remains a crucial factor. Having a clear approval process and providing clear communication on how approvals or denials are made.
2. ***Significance of Financial Terms***- consumers often look to favorable interest rates and lower loan to value ratios since a lower interest rate could significantly reduce long term mortgage costs.
3. ***Reputation and trust*** - institutions with better resolution rates and fewer complaints are more likely to attract customers
4. ***Competitive and transparent rates and fees*** - transparency and affordability of rates remain essential for consumer trust. Consumers may prefer lenders that offer quicker and more transparent decisions through automation.
5. ***Regional economic factors*** - Florida specific trends, like regional housing activity and property types, play a significant role in mortgage decisions

Therefore, it is recommended Addition Financial focus on the following:

1. ***Data-driven marketing*** - use insights from interest rates and loan costs across the mortgage market to influence consumer choices.
2. ***Service optimization*** - improve resolution times and address frequent complaints to build consumer trust
3. ***Regional expansion*** - focus on underserved areas while tailoring product offerings to align with local market needs
4. ***Technological investment*** - leverage online platforms to simplify applications and enhance consumer experiences

References

Aggarwal, S. (2018, August 8). *Latent Dirichlet Allocation (LDA)*. *Towards Data Science*. Retrieved January 3, 2025, from https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

Anandhu. (n.d.). *Word cloud in Python for beginners*. *Kaggle*. Retrieved December 20, 2024, from https://www.kaggle.com/code/anandhuh/word-cloud-in-python-for-beginners

Consumer Complaint Database. Consumer Financial Protection Bureau. (n.d.-a). https://www.consumerfinance.gov/data-research/consumer-complaints/search/?dateRange=All&date_received_max=2025-01-10&date_received_min=2011-12-01&page=1&searchField=all&size=25&sort=created_date_desc&tab=List

GeeksforGeeks. (n.d.). *Python | Part of speech tagging using TextBlob*. Retrieved December 18, 2024, from https://www.geeksforgeeks.org/python-part-of-speech-tagging-using-textblob/

HMDA Dataset Filtering. HMDA - Home Mortgage Disclosure act. (n.d.). https://ffiec.cfpb.gov/data-browser/data/2023?category=states&items=FL

Home Mortgage Disclosure Act (HMDA) Data. Consumer Financial Protection Bureau. (n.d.). https://www.consumerfinance.gov/data-research/hmda/

Python Visualization Authority. (n.d.). *Getting started with folium*. Retrieved January 3, 2025, from https://python-visualization.github.io/folium/latest/getting_started.html

Real Python. (n.d.). *Natural language processing with NLTK in Python*. Retrieved December 18, 2024, from https://realpython.com/nltk-nlp-python/

Real Python. (n.d.). *Natural language processing with spaCy in Python*. Retrieved December 11, 2024, from https://realpython.com/natural-language-processing-spacy-python/

Real Python. (n.d.). *Using collections.Counter in Python to count objects*. Retrieved December 11, 2024, from https://realpython.com/python-counter/

Scikit-learn Developers. (n.d.). *CountVectorizer: Convert a collection of text documents to a matrix of token counts*. Retrieved December 20, 2024, from https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

Scikit-learn Developers. (n.d.). *LabelEncoder: Encode target labels with value between 0 and n_classes-1*. Retrieved December 20, 2024, from https://scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.LabelEncoder.html

Scikit-learn Developers. (n.d.). *LatentDirichletAllocation: Latent Dirichlet Allocation for topic modeling*. December 6, 2024, from https://scikit-learn.org/1.5/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

Scikit-learn Developers. (n.d.). *MultinomialNB: Multinomial Naive Bayes implementation*. Retrieved December 18, 2024, from https://scikit-learn.org/1.5/modules/generated/sklearn.naive_bayes.MultinomialNB.html

Scikit-learn Developers. (n.d.). *TfidfVectorizer: Convert a collection of raw documents to a matrix of TF-IDF features*. Retrieved December 18, 2024, from https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

University of Central Florida, Department of Statistics. (2024). *Addition FI 2024 mortgage loan choices data science competition: Description.* Retrieved from https://sciences.ucf.edu/statistics/wp-content/uploads/sites/2/2024/10/Addition-FI-2024-Mortgage-Loan-Choices-Data-Science-Competition_Description.pdf

W3Schools. (n.d.). *Python requests module*. Retrieved December 18, 2024, from https://www.w3schools.com/python/module_requests.asp