Enhancing Financial Literacy Prevention: An Investigation of Socioeconomic and Geographic Influences

Mikayla Edwards

Table of Contents

Introduction

Syndicated data is external data sources businesses and researchers can use to enrich their internal customer data. This helps them to build more of a profile on their customer segments and adjust their business model (locations, marketing, products, etc.) as necessary. Examples of popular syndicated data sources include data from open sources like the Census, social media, and surveys conducted by the company itself or buying survey data from other companies (Lin, 2015). Example vendors that sell datasets include Experian Marketing Services and TowerData where companies can buy data at a granular level to do things from learn more about their ideal customer profile to expand their email marketing server. Combining company data with other data sources, syndicating data, can help companies better tailor their offerings and learn more about their customer base. However, the disadvantage is that the average user's data may be sold to companies without their explicit approval. This often happens by burying such practices deep within lengthy terms and conditions that are difficult for the average person to understand.

This research will focus on the intersection of syndicated data and financial literacy. Financial literacy involves an individual's understanding of the ways in which they can both manage and optimize their money (McGurran, 2022). Common examples of financial literate principles include budgeting, investing, paying off debt, and maintaining an emergency fund. This research will focus on the United States and examine financial complaints and socioeconomic indicators such as employment status, education, and income by age groups. The purpose of this research is to determine if there are any correlations between financial complaints and financial literacy to see where assistance programs can shift their efforts to create preventative programming and better assists during financial difficulties. The project will be laid out in the following order: 1) Introduction with background on the topic; 2) Objectives; 3) Methodology; 4) Data; 5) Variables, descriptive stats, and data visualizations; 6) Expected

results; 7) Literature review; 8) References; The study will include financial complaint data from the past year, specifically from July 9, 2022, to July 9, 2023. This data will be sourced from Consumer Financial Protection Bureau and includes information on 728 complaints. The data that it will be syndicated with will be from the 2020 Census and pulled into R, where analysis will be done. The plan on the study is as followed:

July 9th, 2023: Proposal completion

July 10th, 2023- July 15th, 2023: Data collection and analysis in R

July 16th, 2023: Data analysis completion

July 17th, 2023 – Create visualizations including charts, graphs, and tables

July 23rd, 2023: Mid-Project update report

July 24th, 2023 – July 29th, 2023: Create, train, and validate a random forest

July 30th, 2023: Create discussion of model estimation and results

August 1st, 2023 – August 5th, 2023: PowerPoint creation to showcase research insights

August 6th, 2023: PowerPoint project presentation

August 10th, 2023: Final project submission

This objective of this research is to explore whether there's a link between how well people understand financial matters and where they live. It will do this by analyzing whether the types of money-related complaints people make have any relation to their geographic location.

Literature Review

Studies on the connection between financial literacy and socioeconomic status have been done by numerous researchers. Outside of organizations that collect data for the general public such as the Bureau of Labor Statistics and the Federal Reserve (collecting economic data as FRED), many individual researchers and universities have conducted their own studies utilizing

publicly available data syndicated with data they created from their own focus groups or online surveys.

One example of this was done by researchers from the University of Iceland where they explored gender differences in relation to financial literacy (Gudjonsson et all, 2022). Their study was based on survey data with 803 participants. They collected data from both genders including information like age, education level, and the results of a financial literacy survey they conducted. They concluded that men and women value different thing which drives their spending habits and knowledge/care of financial literacy. Women value relationships and people while men value things. Although the percentages were close, men had a higher accuracy in answering the true/ false questions compared to women. Although the results were on par with their hypothesis, the researchers did express worry that their study, done in Iceland, a forefront for gender equality, had poor outcomes for women. However, this only made them even more interested in further research and created wonder as to how a similar study could be replicated in other countries and what their results would be.

In another light, researchers have also looked at how financial literacy, socioeconomic status, and politics are connected. A term donned ‘financial inclusion’ is used to measure how often and who utilizes financial services. Financial inclusion is then used in partnership with poverty levels to determine where there is a lack of resources available. This then gives policy makers and other stakeholders insights to drive local and national resource and policy changes. One example of this can be seen in a study conducted by Falak Khan, Dr. Siddiqui, and Dr. Imtiaz, Pakistan based researchers, in which they determined with a sample of over 850,000 individuals from over 10,000 studies from the past 45 years that research on financial literacy

and financial inclusion is few and skewed. Research is mostly in the fields of finance and economics, which makes sense with population and financial based metrics being analyzed.

However, this research is mostly done in or by the United States (22%) with India and Uganda following behind in a close second and third with 17% and 14% respectively. This study also found that researchers are mostly using the same citations and studying the same measures. They concluded that this just makes new distinctive research even more needed as the constant almost identical replication of studies will only increase the chance for overlap and not contribute new knowledge for the general public or policy makers ("Role of Financial Literacy," n.d.).

In a study conducted by researchers from the University of Utah and the University at Buffalo, the focus was on using Consumer Financial Protection Bureau (CFPB) data to examine the types of complaints filed against banks. The researchers explored how social norms and the level of trust in different areas can impact the frequency of financial complaints. The findings revealed that banks with a strong emphasis on customer service, community involvement, and a positive overall culture tend to have fewer financial complaints.

Furthermore, the study investigated the relationship between social norms and how effective consumer protection laws. They discovered that higher levels of trust in a particular location are associated with a lower number of complaints in that area. These results suggest that a government agency that handles consumer complaints can influence how banks treat their customers. The study sheds light on the interaction between informal culture and formal institutions, as well as the impact of stakeholders like customers and government agencies in shaping corporate policies (Hayes et al., 2021).

In addition to macroeconomics, researchers are delving into financial literacy at the microeconomics level, focusing on households and individuals. Researchers in Indiana, looked into household financial literacy and how families implement their knowledge. Despite households having a decent level of financial knowledge, achieving a mean score of 75% in their survey, their ability to put this knowledge into practice, as indicated by their low financial planning skills mean score of 59.8%, remains limited. Researchers determined the best way to fix this gap would be by introducing financial literacy as early as possible, with a preference to when individuals are college aged (Alhenawi et al., 2013).

However, in another investigation Susan J. Crain at Missouri State University explored how colleges are impacting the financial literacy skills of students (Crain, 2013). Crain's study involved looking at the undergraduate curriculum of over 400 universities to see if some sort of financial literacy course was a requirement for students to graduate and fulfill a general education credit. Out of all 435 universities, only 37 schools have some sort of financial literacy course as an elective students can opt to take. When talking with administrations and professors, Crain discovered that it's pretty split on how people think students should be taught to be financially literate with half thinking the classroom is the place and the other half thinking that's something students should learn on their own, from their families, or in extra-curriculars. From this, Crain concluded that if administration is not on board with full-fledged financial literacy courses, then the addition of financial literacy topics to courses such as freshmen seminars, survey courses, or other major related courses is the best course of action.

The conclusions and additional research stemming from these studies present conflicting viewpoints. While one study advocates for colleges to impart financial education, the other asserts that colleges are neither doing so nor inclined to do so. From this we learn that for

college-aged individuals to become finically literate it is up to them to pursue resources. Ohio based researchers at Youngstown State University ran a study consisting of 924 college students to learn about their personal financial literacy. Although results showed slightly higher for business majors, on average participants scored correctly on around 53% of questions. Therefore, college students tend to not be very knowledgeable, and their incorrect opinions informed their improper decision-making skills on these questions (Chen et al., 1998). Another study by Bryce Jorgensen followed with expanding knowledge by conducting what Jorgensen called 'The College Student Financial Literacy Survey' (CSFLS) in which Jorgensen found that as students matriculated through the education system from their freshmen year to master's degrees, their financial literacy knowledge significantly increased. Additionally, Jorgensen investigated how one's family can impact these skills and noted that parents with strong financial knowledge were a positive influence on their students own financial understanding, attitude towards finance, and scoring on the survey (Jorgensen, 2007).

Apart from teaching financial literacy early, there are still questions about what exactly needs to be taught. In today's economy, there's a growing need for people to possess this knowledge from a younger age. A study by Jing Jian Xiao, Sun Young Ahn, Joyce Serido, and Soyeon Shim delved into the difference between subjective and objective financial knowledge. Subjective knowledge was tested by asking students to rate their own knowledge and objective knowledge was tested by testing what students know. Using this information, they tracked how students' financial behaviors changed over time and divided them into two categories: risky spending and risky barrowing.

Their findings found that student's subjective knowledge had a stronger impact on predicting their risky spending and borrowing behaviors than their objective knowledge. Thus,

this research pointed out the importance of myth busting as students' beliefs drove a lot of their risky behaviors. The research indicated that individuals who engaged in less risky financial behavior tended to have more confidence in their subjective knowledge. Lastly, this study concluded that men tended to engage in more risky behaviors than females and students with higher Grade Point Averages (GPAs) tended to have more subjective and objective knowledge which led them to engage in fewer risky financial behaviors ((Xiao et al., 2014).

Methodology

In this study a sample of the data from the Consumer Complaint Database from the past year, designated as CCDB, will be ran through a random forest model (IBM, n.d.). This model allows for the exploration of relationships involving multiple independent variables and a single dependent variable. The model also uses decision trees to understand complex relationships, is able to handle large amounts of data (the CCDB dataset has over 1 million observations, 50,000 of which are used as a sample for model creation) and can accommodate missing values and outliers. The analysis also returns a confusion matrix, showing the models decision-making process and outcomes.

The predicting variable will be Product. Product is a multi-categorical field. The independent variables will be state, zip code, and issue. The goal of this model is to examine if State has a relationship with Product. As mentioned in the variable description, the Product variable is what the consumer complained about. The goal is to see if there is any sort of statistical significance between the Product and State. Then alternating the model, State will be used as the predicting variable with the goal of seeing whether we can use State to determine Product or Product to determine State.

Both models are run in R. The aim of these models is to investigate whether location is related to the type of product the consumer complained about.

Data

I utilized data form the Consumer Complaint Database to learn more about financial issues consumers are facing in the United States. Analyzing this data will help to developing better effective preventative financial literacy efforts and enable us to focus on specific topics in particular locations. This data is known as ‘CCDB’ in my research.

I also incorporated 2020 Census data to access key socioeconomic indicators, including education level completed, median income, and employment status. This data was obtained from individual tables and merged into a single dataset named 'CensusData,' organized by state.

By analyzing the CensusData with the CCDB data, I aim to identify any geographical financial literacy issues or socioeconomic factors that may contribute to financial difficulties. This insight will aid in efforts to enhance financial literacy prevention and support individuals in managing their finances more effectively.

The variables for CCDB include:

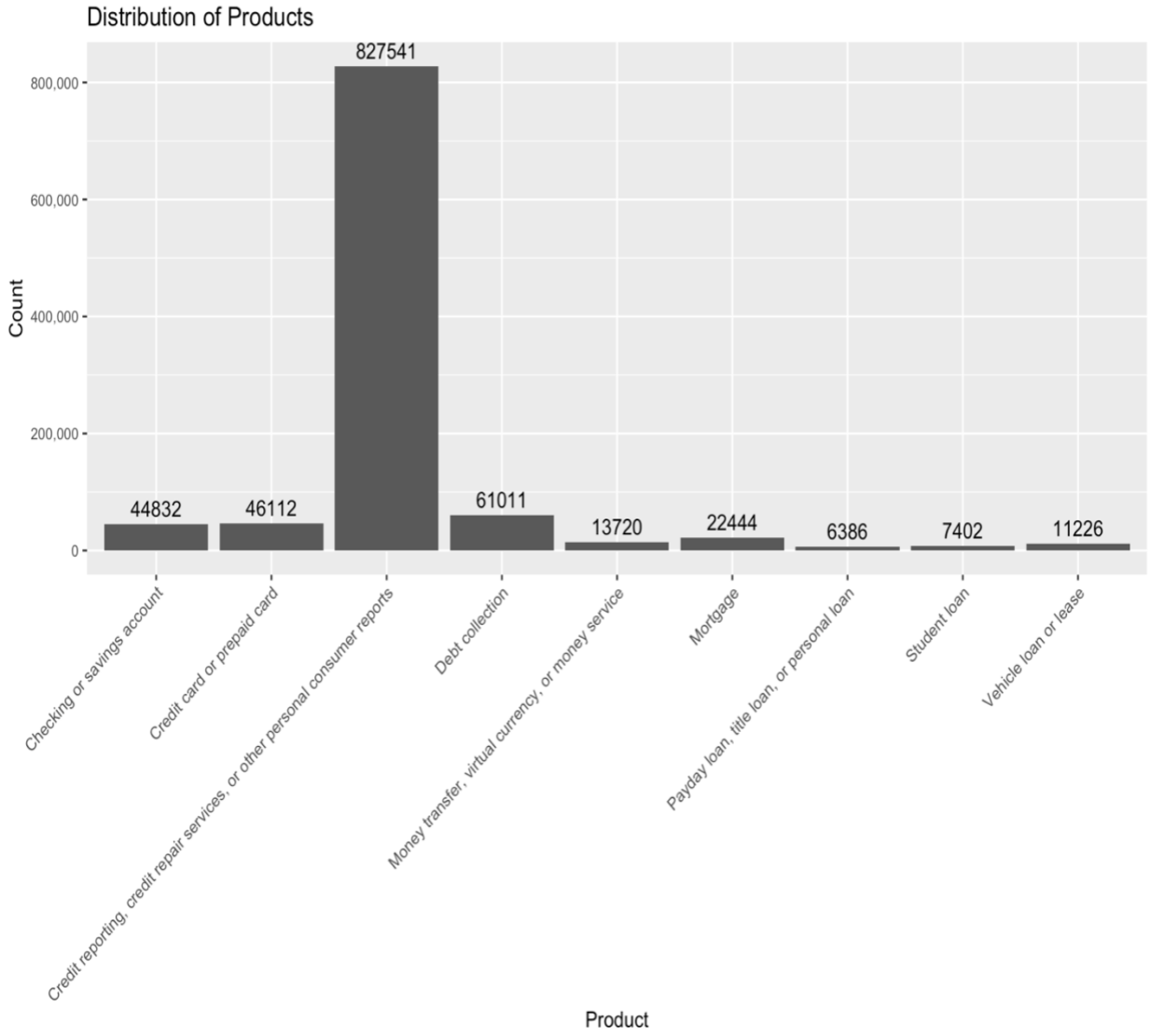
- Data received: The date the complaint was received (Date and time)
- Product: The type of product the consumer complained about (text, categorical)
- Sub-product: The type sub-product the consumer identified (text, categorical)
- Issue: The issue the consumer complained about (Text, categorical)
- Sub-issue: The sub-issue the consumer identified (text, categorical)
- Consumer complaint narrative: Consumer description of the issue. Consumers must opt-in or out for this to be publicly shared (text)

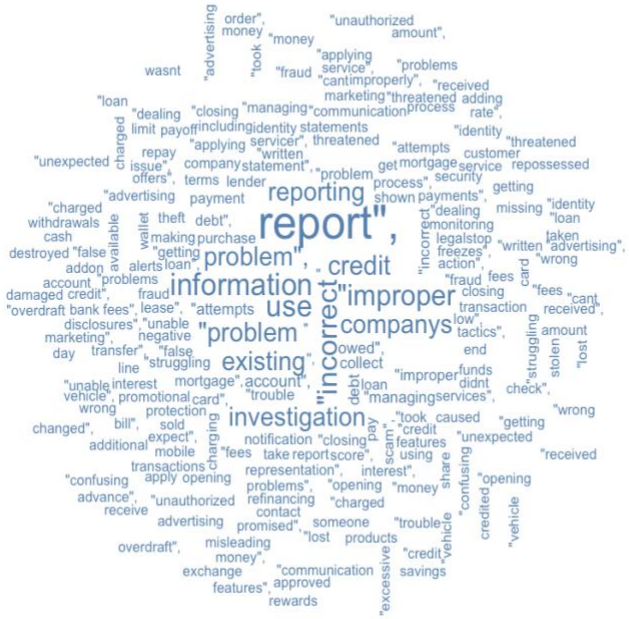
- Company public response: The company's optional public response to the complaint (text)
- Company: The name of the company the complaint is about (text, categorical)
- State: The state of the mailing address of the consumer (text, categorical)
- ZIP code: The zip code of the mailing address of the consumer. Limited to 5 digits. (text)
- Tags: Tags to support search ease. Ex: 'Older American' tag for those 62 years old or older and 'Servicemember' tag for those in or previously were in the military. (text)
- Consumer consent provided?: Whether or not the consumer opted for their complaint narrative to be published. (text, categorical)
- Submitted via: How was the complaint submitted (text, categorical)
- Date sent to company: The date the complaint was sent to the company by the Consumer Financial Protection Bureau. (date and time)
- Company response to consumer: How the company responded to the consumer. (text, categorical)
- Timely response?: Whether the company responded in a timely manner or not. (text, binary, categorical)
- Consumer disputed?: Whether the consumer disputed about the company responded. (text, binary, categorical)
- Complaint ID: The unique identifier for the complaint. (number)

A summary of the data's descriptive stats can be seen below:

```
 Date.received        Product          Sub.product          Issue           Sub.issue
 Length:1040674     Length:1040674     Length:1040674     Length:1040674     Length:1040674
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


 Consumer.complaint.narrative Company.public.response   Company            State             ZIP.code
 Length:1040674               Length:1040674          Length:1040674     Length:1040674    Length:1040674
 Class :character             Class :character        Class :character   Class :character  Class :character
 Mode  :character             Mode  :character        Mode  :character   Mode  :character  Mode  :character


    Tags           Consumer.consent.provided. Submitted.via      Date.sent.to.company
 Length:1040674    Length:1040674             Length:1040674     Length:1040674
 Class :character  Class :character           Class :character   Class :character
 Mode  :character  Mode  :character           Mode  :character   Mode  :character


 Company.response.to.consumer Timely.response.   Consumer.disputed.  Complaint.ID
 Length:1040674               Length:1040674     Length:1040674     Min.   :5750679
 Class :character             Class :character   Class :character   1st Qu.:6150987
 Mode  :character             Mode  :character   Mode  :character   Median :6519622
                                                                    Mean   :6508456
                                                                    3rd Qu.:6870948
                                                                    Max.   :7236351
```

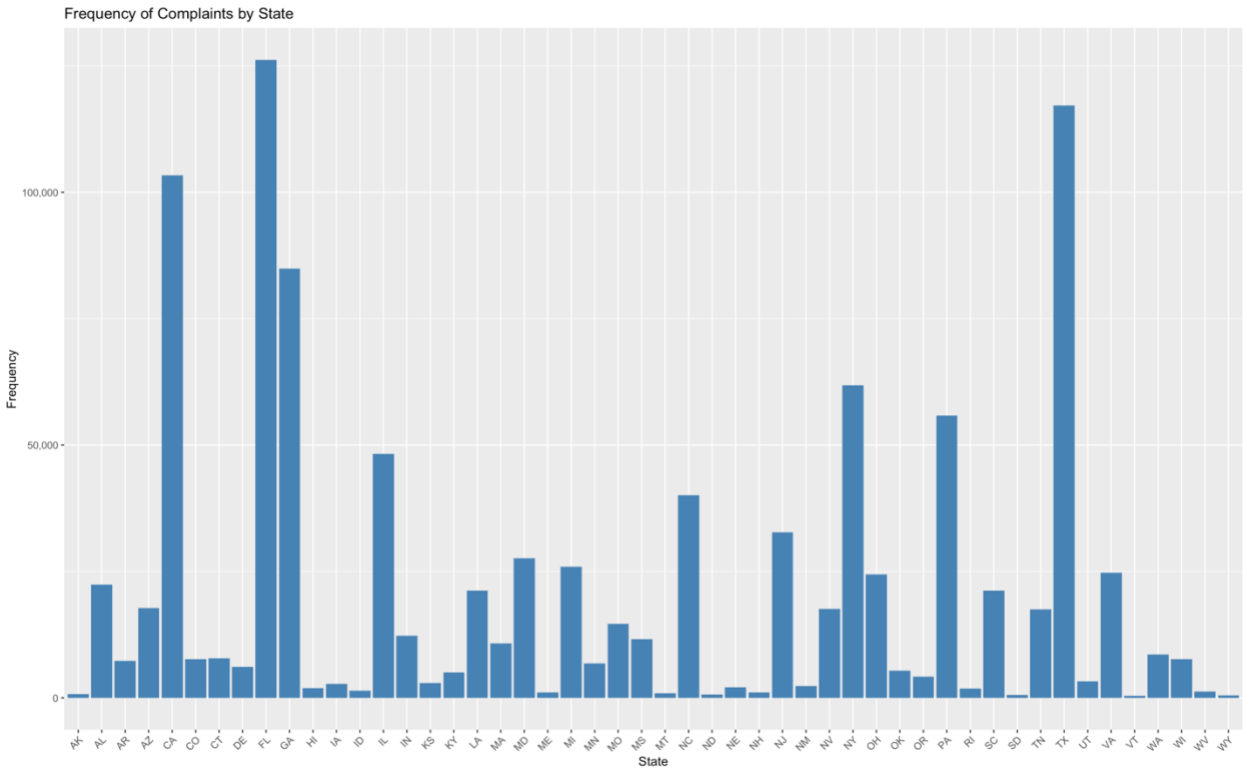
Visualizations for this data include:





Top Companies Complaints are about

Company	Count
TRANSUNION INTERMEDIATE HOLDINGS, INC.	264697
EQUIFAX, INC.	256611
Experian Information Solutions Inc.	242112
WELLS FARGO & COMPANY	19257
CAPITAL ONE FINANCIAL CORPORATION	15057
BANK OF AMERICA, NATIONAL ASSOCIATION	13519
JPMORGAN CHASE & CO.	12901
CITIBANK, N.A.	8816
SYNCHRONY FINANCIAL	6289
Bread Financial Holdings, Inc.	6166



The variables for the CensusData include:

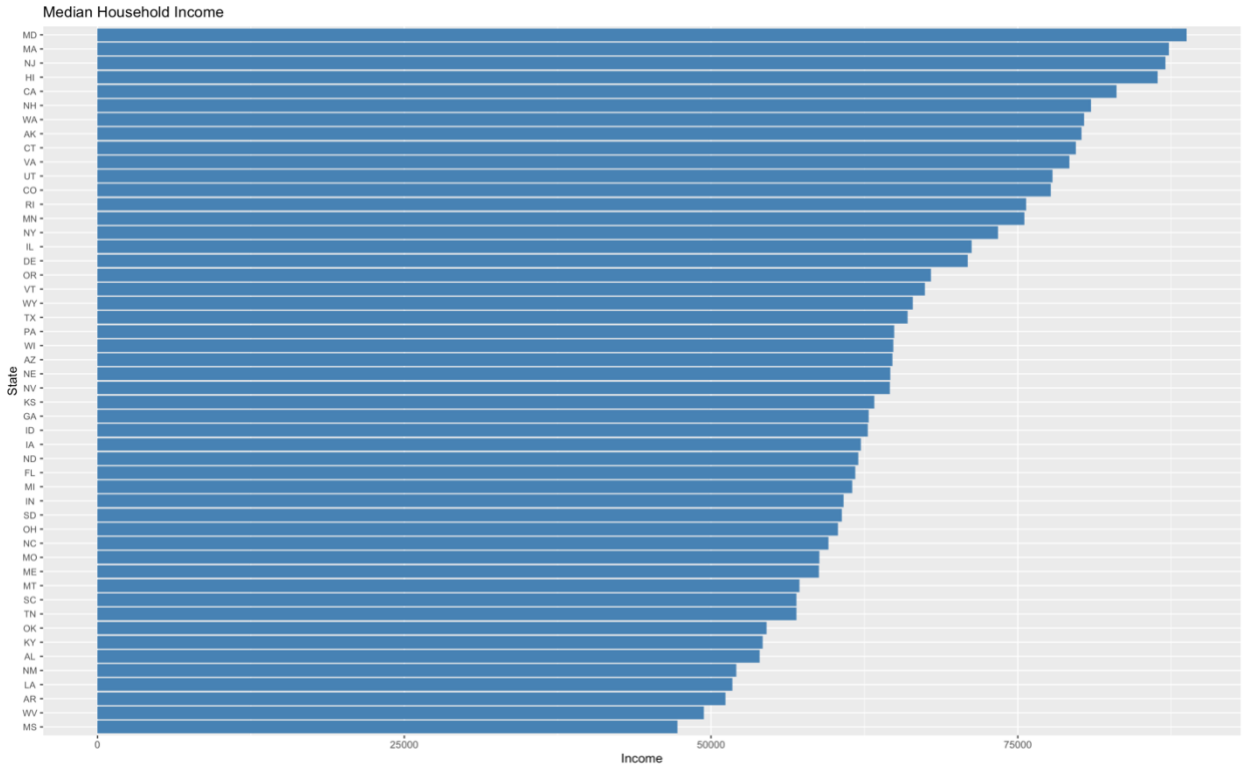
- State: The state the data is from. (text)
- Label: The estimate and margin of error for each state. (text, binary, categorical)
- Less than 9th grade: The number of people who have less than a 9th grade education. (numerical)
- 9th to 12th grade, no diploma: The number of people who have a 9th to 12th grade education but no diploma. (numerical)
- High school graduate (includes equivalency):The number of people who are high school graduates or equivalent. (numerical)
- Some college, no degree: The number of people who have some college but no degree. (numerical)
- Associate's degree: The number of people who have an Associate's degree. (numerical)

- Bachelor's degree: The number of people who have a Bachelor's degree. (numerical)
- Graduate or professional degree: The number of people who have a graduate ro professional degree. (numerical)
- Median household income in the past 12 months (in 2020 inflation-adjusted dollars): The median household income by state for the past 12 months. (numerical)
- In labor force: The number of people in the labor force. (numerical)
- Employed: The number of people employed. (numerical)
- Unemployed: The number of people employed. (numerical)
- Armed Forces: The number of people in the armed forces. (numerical)
- Not in labor force: The number of people NOT in the labor force. (numerical)

A summary of the data's descriptive stats can be seen below:

```
   State              Label           Total Education:   Less than 9th grade 9th to 12th grade, no diploma
Length:51          Length:51          Min.   :  396006   Min.   :   7059     Min.   :  17165
Class :character   Class :character   1st Qu.: 1242742   1st Qu.:  38151     1st Qu.:  62026
Mode  :character   Mode  :character   Median : 3055262   Median :  96663     Median : 179160
                                      Mean   : 4446870   Mean   : 204214     Mean   : 267547
                                      3rd Qu.: 5222064   3rd Qu.: 189760     3rd Qu.: 293677
                                      Max.   :26985739   Max.   :2339290     Max.   :1861951
High school graduate (includes equivalency) Some college, no degree Associate's degree Bachelor's degree
Min.   :  81740                             Min.   :  61199         Min.   :  11586    Min.   :  72511
1st Qu.: 320971                             1st Qu.: 260778         1st Qu.: 114303    1st Qu.: 239796
Median : 833658                             Median : 670587         Median : 269657    Median : 581334
Mean   :1152147                             Mean   : 873536         Mean   : 389699    Mean   : 946397
3rd Qu.:1432902                             3rd Qu.:1020682         3rd Qu.: 453990    3rd Qu.:1258298
Max.   :5319879                             Max.   :5354965         Max.   :2144060    Max.   :6117707
Graduate or professional degree
Min.   :  39329
1st Qu.: 149288
Median : 362236
Mean   : 613331
3rd Qu.: 864716
Max.   :3847887
Median household income in the past 12 months (in 2020 inflation-adjusted dollars) Total Employment Level:
Min.   :47247                                                                      Min.   :  463496
1st Qu.:59209                                                                      1st Qu.: 1446711
Median :64591                                                                      Median : 3589338
Mean   :67145                                                                      Mean   : 5198311
3rd Qu.:76678                                                                      3rd Qu.: 6088868
Max.   :96762                                                                      Max.   :31588105
In labor force:    Civilian labor force:    Employed          Unemployed       Armed Forces
Min.   :  298045   Min.   :  294229      Min.   :  280323   Min.   :  13906   Min.   :   800
1st Qu.:  841048   1st Qu.:  837167      1st Qu.:  786926   1st Qu.:  47285   1st Qu.:  4633
Median : 2141969   Median : 2119098      Median : 1968308   Median : 135943   Median : 14026
Mean   : 3245159   Mean   : 3220485      Mean   : 3001495   Mean   : 218990   Mean   : 24674
3rd Qu.: 3850628   3rd Qu.: 3816268      3rd Qu.: 3568396   3rd Qu.: 242920   3rd Qu.: 22010
Max.   :19857737   Max.   :19706307      Max.   :18090052   Max.   :1616255   Max.   :151430
Not in labor force state_short
Min.   :  165451   create_choropleth_map
1st Qu.:  506935   Class :character
Median : 1457046   Mode  :character
Mean   : 1953152
3rd Qu.: 2341804
Max.   :11730368
```

Visualizations for this data include:



Expected Results

Based on the research, I expect that 1) Income will have a positive correlation with the states that make the most complaints due to their size 2) Location will be a significant factor in which types of financial complaints are submitted. 3) Larger states have higher median household income and higher frequencies of complaints; 4) Banks will be in the top companies people complain about;

Models and Results

In this study the data from the Consumer Complaint Database, designated as CCDB, will be ran through a random forest model (Yiu, 2019). This type of model will allow me to analyze the relationships between multiple independent variables (predictor variables) and one dependent variable (the target variable). In one model, the dependent variable will be the State. State is a

field I created that utilizes the two-character abbreviation for each state. For example, 'North Carolina' is identified as 'NC'. This was done to ease visualization purposes as the Census data, designated as 'CensusData', already has the same abbreviations. The independent variables will be Issue and Product. This was done to decrease the noise in the dataset. In the other model, the dependent variable will be the product. Product is a multi-categorical text field. The independent variables will be State and Issue. The goal of this model is to examine if location has anything to do with the product. As mentioned in the variable description, the Product variable is what the consumer complained about. The goal is to see if there is any sort of relationship between the Product and State.

The R code used to create the random forest to predict State can be seen here:

#Predicting State based on Product

```{r pressure, echo=FALSE}
# install.packages("randomForest")
# install.packages("dplyr")

library(randomForest)
library(dplyr)

set.seed(42)

# Sample 50,000 rows from the CCDB data and select the columns 'Product', 'Issue', and 'State'
sampled_data <- CCDB %>% sample_n(50000, replace = FALSE, select = c("Product", "Issue", "State",
"ZIP.code","Consumer.complaint.narrative"))

# Convert 'Issue' to factor for the random forest function
sampled_data$Issue <- factor(sampled_data$Issue)

# Convert 'State' to uppercase two-letter abbreviations using built-in 'state.abb'
sampled_data$State <- toupper(sampled_data$State)

# Remove rows where 'State' is not a valid two-letter abbreviation (NA rows)
sampled_data <- sampled_data %>% filter(!is.na(State) & State %in% state.abb)

# Convert 'Product' and 'State' to factors
sampled_data$Product <- factor(sampled_data$Product)
sampled_data$State <- factor(sampled_data$State)

# Create the random forest model with n = 2000
rf_model_state <- randomForest(State ~ Product, data = sampled_data, ntree = 2000)

# Print the model summary
print(rf_model_state)


```

Here are the model's results:

```
Call:
 randomForest(formula = State ~ Product, data = sampled_data,      ntree = 2000)
               Type of random forest: classification
                     Number of trees: 2000
No. of variables tried at each split: 1

        OOB estimate of  error rate: 87.26%
```

This model, which used Product to predict State included 2000 trees and an OOB estimate of error rate of 87.26%, did not perform well in accurately predicting the State based on the Product variable. The error rate suggests that at only around 10% of the time the model was correct in predicting State. However, the confusion matrix shows the model was mostly incorrect in attempting to predict the state for almost all places except for California, Florida, and Texas. For California, the model had around an 83% error rate, while Florida and Texas have 14% and 8% respectively. California had some correct predictions, but its error rate is closer the overall model rate and is not very effective. On the other hand, Florida and Texas have error rates lower than 15% meaning that almost 85% of the time the model is able to correctly predict the State based on the Product. When taking a look at the rankings of the states in 'CensusData', the median income of California is ranked much higher than Texas with Florida following up both states even lower results. Since these states do not appear to be closely affiliated with one another in the CensusData I followed up by further exploring this in the CCDB data. A count code was ran to see how often these states appear in the CCDB by focusing on the top 10 states in the data. The result are as follows:

```
    FL     TX     CA     GA     NY     PA     IL     NC     NJ     MD
126200 117142 103375  84899  61779  55806  48287  40055  32724  27614
```

This analysis confirmed the suspicion that Florida, Texas, and California are in the top states with complaints in the CCDB. This makes sense as the more data a model has to learn from the better the model will be at making predictions. However, even though California is the third highest complaining state, it still had an individual error rate of 83%. This indicates a need for further analysis to understand why this high error rate is occurring.

The R code used to create the random forest to predict Product can be seen here:

```
# Predicting Product based on State
```{r pressure, echo=FALSE}
# install.packages("randomForest")
# install.packages("dplyr")

# library(randomForest)
# library(dplyr)

set.seed(42)

# Sample 50,000 rows from the CCDB data and select the columns 'Product', 'Issue', 'State', and 'ZIP.code'
sampled_data <- CCDB %>% sample_n(50000, replace = FALSE, select = c("Product", "Issue", "State","ZIP.code",
"Consumer.complaint.narrative"))

# Convert 'Issue' to factor for the random forest function
sampled_data$Issue <- factor(sampled_data$Issue)

# Convert 'State' to uppercase two-letter abbreviations using built-in 'state.abb'
sampled_data$State <- toupper(sampled_data$State)

# Remove rows where 'State' is not a valid two-letter abbreviation (NA rows)
sampled_data <- sampled_data %>% filter(!is.na(State) & State %in% state.abb)

# Convert 'Product' and 'State' to factors
sampled_data$Product <- factor(sampled_data$Product)
sampled_data$State <- factor(sampled_data$State)

# Create the random forest model with n = 2000
rf_model_product <- randomForest(Product ~ State, data = sampled_data, ntree = 2000)

# Print the model summary
print(rf_model_product)


```
```

Here are the model's results:

```
Call:
 randomForest(formula = Product ~ State, data = sampled_data,      ntree = 2000)
               Type of random forest: classification
                     Number of trees: 2000
No. of variables tried at each split: 1

        OOB estimate of  error rate: 20.43%
```

This model, which included 2000 trees and an OOB estimate of error rate of 20.43%, did very well at accurately predicting the Product based on the State variable. The error rate suggests that at around 79% of the time the model was correct in predicting Product. The confusion matrix's class.error showcases the models success even further by showing that it accurately predicted the Product 100% of the time in the following categories: "Credit reporting, credit repair services, or other personal consumer reports" and "Checking or savings account.". However, the model did experience some misclassification in the following categories: "Student loan," "Payday loan, title loan, or personal loan," "Vehicle loan or lease," "Debt collection," "Money transfer, virtual currency, or money service," and "Mortgage". This prompted me to want to see a count of the various Product categories both in my sample data used by the model and in the CCDB as a whole. The results are as follows for the sampled data:

```
Credit reporting, credit repair services, or other personal consumer reports
                                                                      39458
                                                        Debt collection
                                                                       2823
                                                Credit card or prepaid card
                                                                       2223
                                                Checking or savings account
                                                                       2145
                                                                   Mortgage
                                                                       1079
                                Money transfer, virtual currency, or money service
                                                                        647
                                                     Vehicle loan or lease
                                                                        554
                                                               Student loan
                                                                        352
                                          Payday loan, title loan, or personal loan
                                                                        310
```

As stated earlier, the 100% accuracy versus misclassification results make sense since the model had less of those Product category types to learn from. However, it makes it even more impressive that it still was relatively accurate most of the time. Further research could include looking into the wording used by consumers in the complaints of the misclassified data as there could be common wording overlap that the model is confused by.

Conclusion

While predicting the specific state might not always be reliable, the second model showed that Product can be a successful predictor for State. Determining the financial literacy needs on a state-by-state basis presents challenges due to the CCDB data composition. The analysis draws from a sample dataset featuring over 39,000 complaints on ‘credit reporting, credit repair services, and personal consumer reports’ which also constitutes 78% of the full CCDB data set. Therefore, the model is able to accurately predict credit-related concerns only because it makes up majority of both the sample data and the CCDB set used for this analysis.

The idea of applying these models to a more expansive Consumer Complaint Database spanning multiple years to include a broader spectrum of product types would be good further research and would allow room to engage in deeper state-based factors such as median income and employment level and how they affect Product complaints reported.

Nonetheless, focusing efforts on the US credit system would be a good place to start as it appears that most issues are stemming from it. Such initiatives can contribute to enhancing financial literacy and simultaneously reducing formal consumer complaints across the United States.

References

Alhenawi, Yasser and Elkhal, khaled, Financial Literacy of U.S. Households: Knowledge vs. Long-Term Financial Planning (2013). Financial Services Review, Vol. 22, 2013, Available at SSRN: https://ssrn.com/abstract=2532068

Chen, H., & Volpe, R. P. (1998). An analysis of personal financial literacy among college students. Financial Services Review, 7(2), 107-128. https://doi.org/10.1016/S1057-0810(99)80006-7.

Consumer complaint database. Consumer Financial Protection Bureau. (n.d.). https://www.consumerfinance.gov/data-research/consumer-complaints/

Crain, S. J. (2013). Are Universities Improving Student Financial Literacy? A Study of General Education Curriculum. Journal of Financial Education, 39(1/2), 1–18. http://www.jstor.org/stable/41948694

Full article: Role of Financial Literacy in achieving financial ... (n.d.). https://www.tandfonline.com/doi/full/10.1080/23311975.2022.2034236

Gudjonsson, S., Minelgaite, I., Kristinsson, K., & Pálsdóttir, S. (2022, November 28). *Financial Literacy and gender differences: Women choose people while men choose things?*. MDPI. https://www.mdpi.com/2076-3387/12/4/179#:~:text=Financial%20literacy%20%3D%200.364–0.067%20gender,the%20worse%20the%20financial%20literacy.

HAYES, R. M., JIANG, F., & PAN, Y. (2021). Voice of the customers: Local Trust Culture and consumer complaints to the CFPB. *Journal of Accounting Research*, *59*(3), 1077–1121. https://doi.org/10.1111/1475-679x.12364

IBM. (n.d.). Random Forest. https://www.ibm.com/topics/random-forest

Jorgensen, B. L. (2007, October 2). Financial Literacy of College Students: Parental and Peer Influences. [Master's Thesis, Virginia Tech]. Virginia Tech Works. https://vtechworks.lib.vt.edu/handle/10919/35407

Multinomial Logistic Regression | Stata Data Analysis Examples. (n.d.) UCLA: Statistical Consulting Group. from https://stats.oarc.ucla.edu/stata/dae/multinomiallogistic-r regression/

Xiao, J. J., Ahn, S. Y., Serido, J., & Shim, S. (2014). Earlier financial literacy and later financial behaviour of college students. International Journal of Consumer Studies, 38(6), 593-601. https://doi.org/10.1111/ijcs.12122

Yiu, T. (2019, June 12). Understanding Random Forest. Towards Data Science. Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2