Extracting Demographic Clusters From Consumer Financial Protection Bureau Complaint Data

Jacob Metzger

May 13, 2016

1 Introduction

The Consumer Financial Protection Bureau (CFPB) was founded in 2011 in response to the 2008 housing crisis in order to monitor and evaluate the financial marketplace and address consumer concerns regarding the financial sector and its products. [2] A part of the CFPB's goal regarding transparency and response to consumer concerns, the Bureau provides open access to regularly updated data for the complaints it receives. The task set forth in this project is to explore the CFPB consumer complaint database and other publicly available data sources to discover demographic groups that may be of interest for financial sector companies seeking to take a proactive approach to addressing perceived problems in products or practices. In particular, the goal is to identify subpopulations of consumers that have made complaints to the CFPB on their behalf. This identification can serve as the basis of a targeted action plan by those in the financial sector seeking to proactively address consumer concerns in products, processes, or practices for their own consumer bases. These cluster groups may also be of interest to those seeking to understand how different consumers perceive harms by financial actors.

2 Scope

For the purpose of this project, I have elected to concentrate analysis on the complaints received for the continental United States, leaving out (where applicable) considerations for Alaska, Hawaii, and US territories such as Puerto Rico and the Virgin Islands. This project will focus primarily on categorical and numerical data, since these data are most easily amenable to clustering analysis. As such, the free-form consumer narrative text provided by the CFPB is not within the scope of this analysis.

3 Datasets

3.1 Primary Sources

Three primary sources were used to provide data for analysis: The Consumer Finance Bureau Consumer Complaint Database [3], IRS Individual Income Tax Statistics [4], and demographic data provided by the US Census Bureau [5]. Each of these datasets is available publicly from the web sites of the respective government agencies. In addition, supplemental data from the HUD USPS Zip Code Crosswalk Files [6] was used to aid in merging these datasets.

3.2 Transformation and Reconciliation

Data transformations were performed on the CFPB dataset in order to facilitate alignment with the IRS income statistics and the Census Bureau demographics data. For example, the State column was transformed from state name ("Alabama") to state abbreviation ("AL"). Additionally, because the Census bureau data was aggregated on a county level and the CFPB data was aggregated by zip code, I used HUD USPS Zip

Code Crosswalk data [6] in order to translate zip codes into counties. Because counties do not always circumscribe zip codes, in the case that a zip code for an observation corresponded to multiple counties, one of these counties was chosen uniformly at random. This introduces some noise into the dataset, but this seemed to be a reasonable solution given the lack of resolution between the data sets. This strategy is not expected to significantly impact the aggregated results of this project. Additionally, the Census Bureau's population counts were transformed into percentages by dividing by the total county population to allow for smaller and larger counties with similar demographic profiles to be compared meaningfully.

3.3 Shape and Features

After preprocessing, a total of 462,939 observations are used. The merged dataset itself contains 93 attributes: 17 attributes are directly taken from the CFPB complaint dataset, 2 are attributes added to help with identifying counties along with the county name, 1 is the county-level adjusted gross income taken from the IRS dataset, and the 73 columns thereafter are county-level population demographics taken from the Census dataset. An enumeration and brief description of the columns mentioned in the results of the project is provided in Appendix I. A merged list of the attributes selected for consideration from these sets is shown Appendix II. Complete descriptions for each of the columns made available by the CFPB, IRS, and Census Bureau datasets are available with those datasets.

4 Cluster Modeling

A k-means algorithm was used in order to extract clustering information from this dataset. In order to input columns containing categorical data into k-means, one-hot encoding was used to transform these columns. While this significantly increases the dimensionality of the intermediate data, clustering algorithms like k-means make use of distance metrics that categorical data does not provide, so this transformation was necessary as a preliminary step. After that, the dimensionality of the data was reduced by using principal component analysis (PCA). This linearly projects the data from its sparse, high-dimensional representation into a lower-dimensional space. This reduction of sparsity is meant to improve the quality of the clusters returned by k-means. For this project, the total feature-space was reduced to 5 dimensions before applying the clustering algorithm. Additionally, because k-means is sensitive to the parameter k, the number of clusters to search for, an estimate of an optimal k was obtained by graphing the distortion of the hypothetical clusters. [1] k was chosen from this graph by selecting the value that reduces cluster distortion while minimizing the total number of clusters introduced into the model.

A downside of any unsupervised clustering model is that the resultant clusters can be difficult to interpret since they may not correspond to obvious relationships or existing categories. After producing the k-means model, I utilize the clustering results to train a Random Forest model in order to extract key factors explaining the clustering result.

5 Clustering Results

5.1 Determining k

Figure 1 displays the graph of cluster distortion based on different values of k, the total number of clusters. Based on this graph, k=3 is an optimal value, though 4 and 5 may also be acceptable in other cases. k=3 rests in a graphical "elbow" of the graph, providing an ideal point where the increased clusters increase the complexity of the model without proportional benefit.

5.2 Visualizing the k-means clusters

In Figure 2, the three clusters identified by k-means are displayed. The visualization was created by using a second, separate PCA projection into a two-dimensional space. The three clusters consist of the blue and the green to the left part of the graph and the cluster in the upper right in red. These are clusters 0, 1, and 2, respectively. We can see that cluster 2 is much smaller, much less cohesive, and much further removed

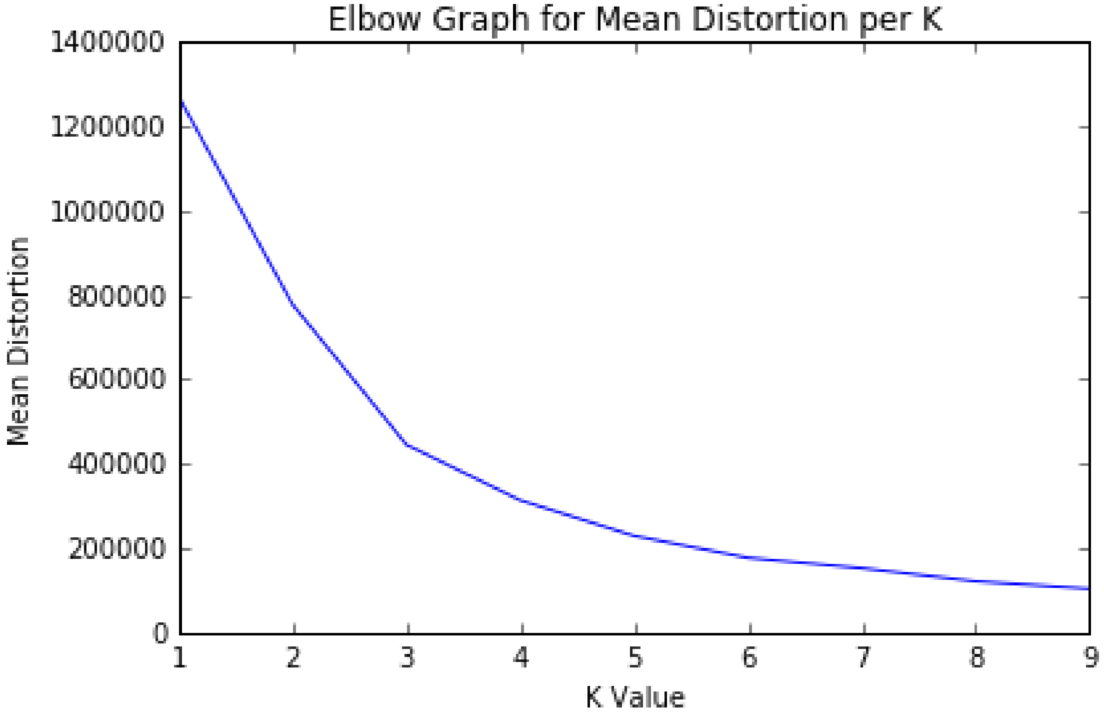

Figure 1: k-value Distortion Graph

from the bulk of the data. Figure 3 shows the support (population counts) for each cluster. Based on these observations, it is reasonable to consider cluster 2 as consisting of outliers or anomalous entries.

6 Cluster Analysis via Feature Importance in a Random Forest Model

Following the clustering result from k-means, a Random Forest model was trained using the same one-hot encoded data that was fed into k-means along with the clustering results. A Random Forest model was chosen for this due to its robust nature: it tends to not overfit data, has relatively few hyperparameters, and handles outliers reasonably. Further, the Random Forest model includes an intrinsic, robust measure of its performance, called the out-of-bag score, which removes the need for a separate training and test set with independent cross validation of the trained model. That score was used to evaluate the appropriateness of the model after training.

The Random Forest ensemble model was trained with 1000 estimators and yielded an out-of-bag score of approximately 0.9998, where a score of 1.0 is the score of a perfect model. This indicates that the Random Forest model accurately represents the k-means clustering result. After training, the Random Forest model yielded a list of the most important features used in its classification (see Figure 4). Since the importance of the individual features trails off quickly and the total number of features is very high, the top 10 features were given particular attention in the following analysis.

From the graph of the most important features used by the Random Forest model, we can see that each of the individual features used by the model have relatively low importance and comparatively high uncertainty. However, looking at these features, it's worthwhile to note a few points: First, all of these features are demographic features that were not part of the CFPB dataset but were introduced from the US Census Bureau dataset. Second, most of these features involve a self-identified hispanic, hispanic multiracial, or mixed hispanic county-level demographic, with only two of the top ten being non-hispanic demographics.

Table 1 shows the means of the county-level demographic percentages from cluster 0 and cluster 1 for these top 10 features. Note that for each of the hispanic subpopulations, their county-level means are

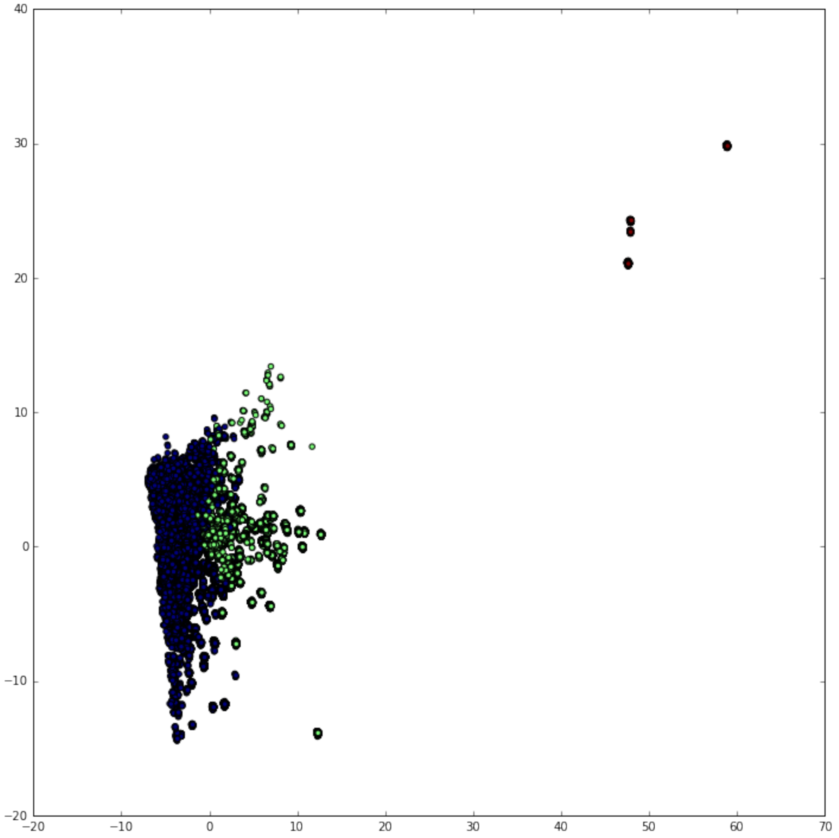

Figure 2: PCA Projection of Clustering Result
Blue = Cluster 0, Green = Cluster 1, Red = Cluster 2

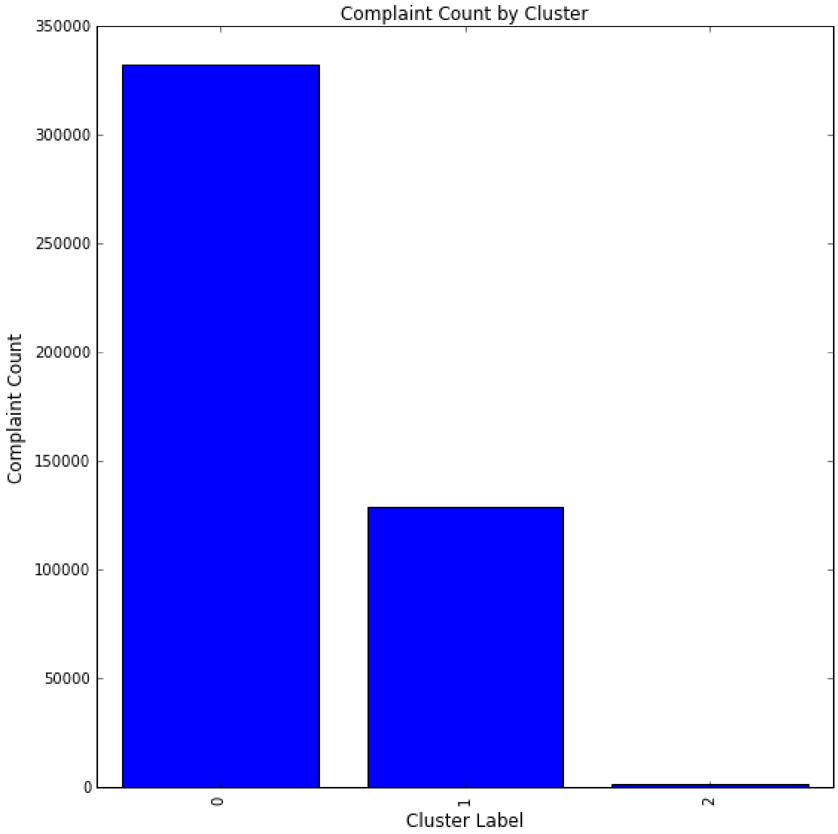

Figure 3: Cluster Support Counts

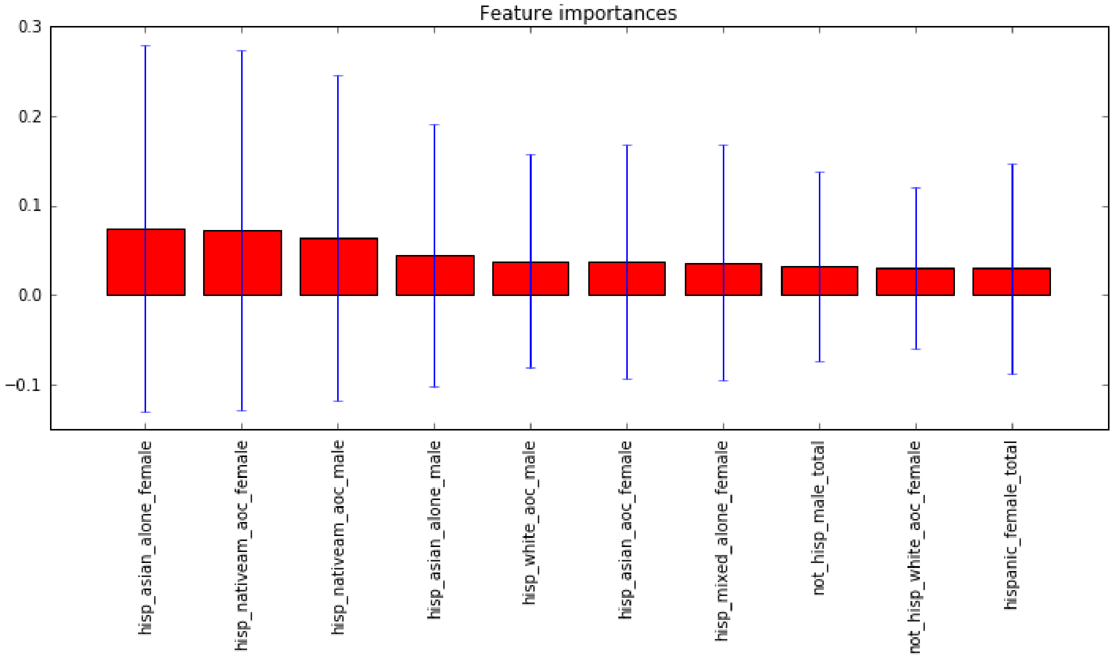

Figure 4: Top 10 Feature Importances from Random Forest

Feature Name	Cluster 0 Mean	Cluster 1 Mean
hisp asian alone female	0.000423595237060243	0.002193308261492692
hisp nativeam aoc female	0.0020776312208056168	0.007215892494361203
hisp nativeam aoc male	0.002312719910802029	0.007526149032218437
hisp asian alone male	0.0004325625329237277	0.002142435427032182
hisp white aoc male	0.04786420075582912	0.17415252155733882
hisp asian aoc female	0.0007691231017846226	0.0036182140314220497
hisp mixed alone female	0.0016473071853516253	0.004493868685872551
not hisp male total	0.43525209081457994	0.3031082418730597
not hisp white aoc female	0.3537952380157938	0.20179656984626024
hispanic female total	0.05072443869266556	0.18998654077791183

Table 1: Demographic Features and Means in Identified Clusters

higher for cluster 1, whereas for the white subpopulations the county-level means are higher for cluster 0. That is, for these attributes, hispanic populations tend to be assigned to cluster 1 while non-hispanic, white populations tend to be assigned to cluster 0. For contrast, note how cluster 0 is more than twice as populated as cluster 1, as shown in Figure 3. This suggests that counties with higher proportions of hispanic subpopulations are represented differently within the CFPB Consumer Complaint Database than counties with larger proportions of white, non-hispanic subpopulations.

7 Conclusion

The goal for this project was to explore the CFPB Consumer Complaint Database for interesting demographic clusters. Because the CFPB database does not directly contain demographic data, the dataset was supplemented by income statistics from the IRS and demographic statistics from the US Census Bureau. By use of the k-means clustering algorithm, non-trivial clusters within the CFPB data were located, and by the use of Random Forests, several explanatory factors were obtained. In analyzing the top factors used by the model, it was observed that hispanic subpopulations tend to be clustered differently from their non-hispanic, white counterparts within the CFPB dataset.

This project did not intend to explore causes for clusters within the CFPB dataset, and such speculation risks becoming very complex very quickly, requiring a level of nuance that may escape the relatively brute approach used here. Nonetheless, knowing that these clusters exist is valuable, actionable information. The existence of these groups indicates differences in how populations interact with the financial industry and provides the basis for a segmented strategy to assess and address their concerns in a targeted way.

References

[1] Bernico, Mike *K-Means Clustering Lecture* https://github.com/mbernico/CS570/blob/master/Kmeans_Clustering_Lecture.ipynb. Accessed March 28, 2016

[2] Consumer Financial Protection Bureau *The Bureau* http://www.consumerfinance.gov/about-us/the-bureau/. Accessed April 9, 2016.

[3] Consumer Financial Protection Bureau *Consumer Complaint Database* http://www.consumerfinance.gov/data-research/consumer-complaints/. Accessed April 9, 2016.

[4] Internal Revenue Service *SOI Tax Stats - County Data - 2013* https://www.irs.gov/uac/SOI-Tax-Stats-County-Data-2013. Accessed April 9, 2016

[5] US Census Bureau *County Characteristics Datasets: Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2014* https://www.census.gov/popest/data/counties/asrh/2014/CC-EST2014-ALLDATA.html. Accessed April 9, 2016

[6] US Department of Housing and Urban Development *HUD USPS ZIP Code Crosswalk Files* https://www.huduser.gov/portal/datasets/usps_crosswalk.html. Accessed April 11, 2016

Appendix I: Specific Column Descriptions

Feature Name	Description (From Census Bureau)
hisp asian alone female	Hispanic-Asian (alone) Female
hisp nativeam aoc female	Hispanic-Native American (alone or in combination) Female
hisp nativeam aoc male	Hispanic-Native American (alone or in combination) Male
hisp asian alone male	Hispanic-Asian (alone) Male
hisp white aoc male	Hispanic-White (alone or in combination) Male
hisp asian aoc female	Hispanic-Asian (alone or in combination) Female
hisp mixed alone female	Hispanic-Mixed (alone) Female
not hisp male total	Non-Hispanic Male Total
not hisp white aoc female	Non-Hispanic White (alone or in combination) Female
hispanic female total	Hispanic Female Total

According to the Census Bureau survey, "Hispanic origin is considered an ethnicity, not a race. Hispanics may be of any race."[5] "Alone" and "in combination" refer to identifying with the specified demographic solely or together with additional groups. For example, an Asian American that does not identify with any other demographic may be identified as "Asian Alone" here, whereas a Native American that also identifies as Asian may be considered "Native American In Combination".

Appendix II: Column Names

date received

product

sub product

issue

sub issue

company response

company

state name

zip code

tags

consumer consent provided

submitted via

date sent

company response status

timely response

consumer disputed

complaint id

county code

county name

CountyAGI

total pop

total male pop

total female pop

white alone male pop

white female alone pop

black alone male

black alone female

nativeam male alone

nativeam female alone
asian male alone
asian female alone
pacific male alone
pacific female alone
mixed male total
mixed female total
white aoc male
white aoc female
black aoc male
black aoc female
nativeam aoc male
nativeam aoc female
asian aoc male
asian aoc female
pacific aoc male
pacific aoc female
not hisp male total
not hisp female total
not hisp white alone male
not hisp white alone female
not hisp black alone male
not hisp black alone female
not hisp nativeam alone male
not hisp nativeam alone female
not hisp asian alone male
not hisp asian alone female
not hisp pacific alone male
not hisp mixed alone male
not hisp mixed alone female
not hisp white aoc male
not hisp white aoc female
not hisp black aoc male
not hisp black aoc female
not hisp nativeam aoc male
not hisp nativeam aoc female
not hisp asian aoc male
not hisp asian aoc female
not hisp pacific aoc male
not hisp pacific aoc female
hispanic male total
hispanic female total
hisp white alone male
hisp white alone female
hisp black alone male
hisp black alone female
hisp nativeam alone male
hisp asian alone male
hisp asian alone female
hisp pacific alone male
hisp pacific alone female
hisp mixed alone male
hisp mixed alone female
hisp white aoc male
hisp white aoc female
hisp black aoc male
hisp black aoc female
hisp nativeam aoc male
hisp nativeam aoc female
hisp asian aoc male
hisp asian aoc female
hisp pacific aoc male
hisp pacific aoc female
not hisp pacific alone female
hisp nativeam alone female