Detecting Latent Defects in Consumer Finance Complaint Data

Jacob Perricone (jacobp2@stanford.edu), Blake Jennings (bmj@stanford.edu), Liu Jiang, and Ben Etringer

Stanford University

Overview

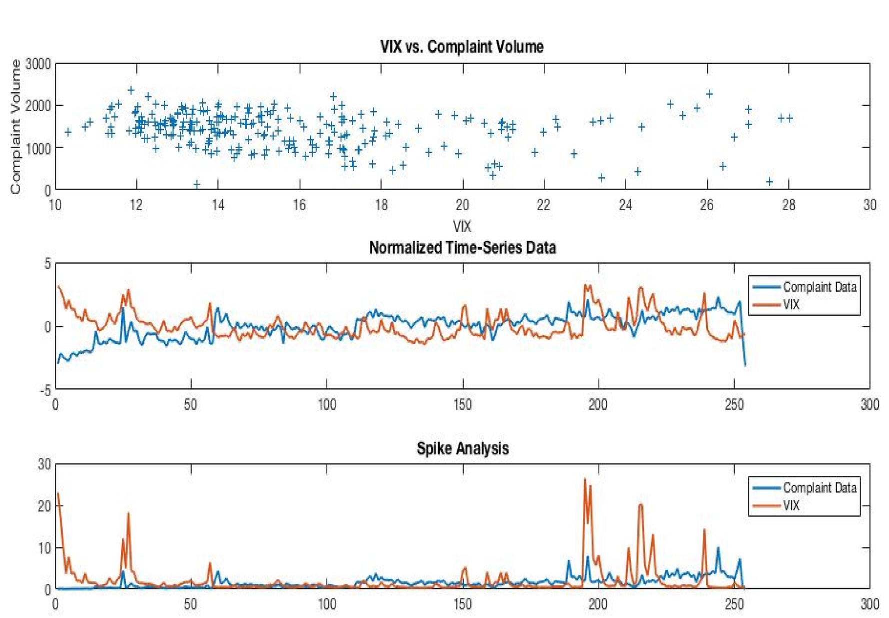
We analyze the Consumer Financial Protection Bureau's (CFPB) consumer complaint database, leveraging both the data's structured fields and unstructured complaint narratives. Employing a probabilistic defect model to synthesize the complaint narratives, we are able to compare financial products and institutions across defects. Encapsulated in a web application, consumers could utilize our model to evaluate a particular company/product before purchase/use.

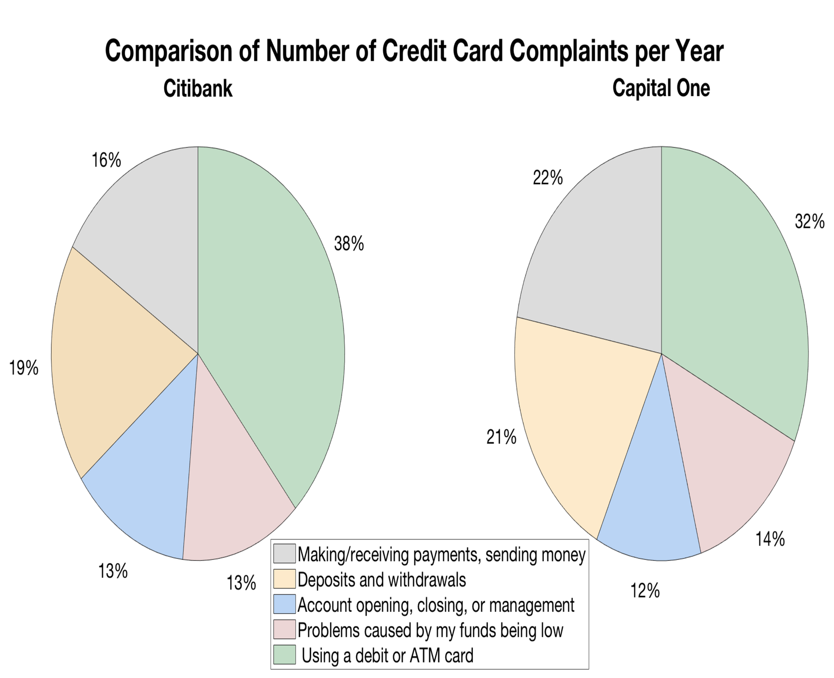
Data and Data Processing

Our main data source for analysis is the CFPB (Consumer Financial Protection Bureau) Consumer Complaint Database, which has logged nearly 700,000 consumer complaints since its launch in 2012. We examined both the database's structured fields as well as its complaint narratives.

In terms of other data sets, we investigated the relationship between the CFPB time series data and the macro-market environment. In processing the CFPB data, we aggregated CFPB data points by week to smooth out intra-week cyclical variability, normalized CFPB data points by maximum number of complaints received by an institution, and removed stop words and redundant information from the CFPB complaints.

Structure Fields Analysis





Methods

In order to gather more insight from the raw complaints, we employ a probabilistic defect model to extract latent defects from the complaint data. The motivation for using a probabilistic latent defect model is that it is able to extract underlying defects that span across both the unstructured data, the complaint narrative, and the structured fields. We develop a multi-aspect topic model that incorporates both the structured and unstructured field and is able to synthesize a massive body into a small set of defects. We are able to extract a synthesized set of complaint characteristics and attribute them both cross-sectionally, across institutions and products, and through time.

EM Analysis of Complaint Narratives

We assume that each complaint record is generated from a latent distribution of defects, and we set out to model the joint distribution of complaints and defects. For each complaint record, $x^{(i)} : x \in \mathbf{C}$, we aim to model the joint distribution $p(x^{(i)}, d^{(i)}) = p(x^{(i)}|d^{(i)})p(d^{(i)})$ where $d^{(i)} \sim$ Multinomial(ϕ) (where $\phi_j > 0, \sum_{j=1}^{|\mathbf{D}|} \phi_j = 1$) with the parameter ϕ_j specifying $p(d^{(i)} = j)$. The complaint generation process is modeled as follows.

1: Complaint Generation

1. Choose a defect $d_j \sim$ Multinomial(ϕ_j)
2. Generate a complaint $x^{(i)} \sim p(x^{(i)}|d_j)$ where for each complaint-entity:
   1. $p(b^{(i)}, i^{(i)}, p^{(i)}, n^{(i)}, t^{(i)}|d_j) = p(b^{(i)}|d_j)p(i^{(i)}|d_j)p(p^{(i)}|d_j)p(n^{(i)}|d_j)p(t^{(i)}|d_j)$ (independence)
   2. bank $b^{(i)}|d_j \sim$ Multinomial(θ_b^j)
   3. issue $i^{(i)}|d_j \sim$ Multinomial(θ_i^j)
   4. product $p^{(i)}|d_j \sim$ Multinomial(θ_p^j)
   5. word $w^{(i)}|d_j \sim$ Multinomial(θ_w^j)
   6. product $t^{(i)}|d_j \sim N(\mu_j, \sigma_j)$

The parameters of our model, $\Phi = \{\theta_k^j | k \in (b,p,i,w,t), \mu_j, \sigma_j\}$, are estimated by the EM algorithm. The log-likelihood of our data is:

$$\mathcal{L}(X;\Phi,\phi) = \sum_{i=1}^{m} \log\Big[\sum_{j=1}^{k} p(x^{(i)}|d^{(i)};\Phi)p(d^{(i)};\phi)\Big]$$

The complaint posteriors and defect posteriors can be decomposed into:

$$p(x^{(i)}|d^{(i)} = j;\Phi) = \prod_{k \in (b,p,i,n,t)} p(k^{(i)}|d^{(i)} = j)$$

$$p(d^{(i)} = j|x^{(i)})_t = \frac{p(d^{(i)} = j)_t p(x^{(i)}|d^{(i)} = j)_t}{\sum_{l=1}^{k} p(d_t^{(i)} = j)_t p(x^{(i)}|d^{(i)} = l)_t}$$

To avoid underflow the defect posteriors are transformed into log space and scaled by $m^{(i)} = \max_i z_i^{(i)}$:

$$z_j^{(i)} = \log[p(d^{(i)} = j)] + \sum_{k \in (b,p,i,w,t)} \log[p(k^{(i)}|d^{(i)} = j)] + \sum_{w \in \mathbf{V}} C_w \log[p(w|d^{(i)} = j]$$

$$p(d^{(i)} = j|x^{(i)}) = \begin{cases} 0 \text{ if } z_j^{(i)} - m^{(i)} < -K \\ \dfrac{e^{z_j^{(i)} - m^{(i)}}}{\sum_{l: z_l^{(i)} - m^{(i)} \geq -K} e^{z_l^{(i)} - m^{(i)}}} \end{cases}$$

In the M-Step, we update the posterior densities, Φ as follows:

$$p(w|d^{(i)} = j)_{t+1} = \frac{1 + \sum_{i=1}^{M} p(d^{(i)} = j|x^{(i)})_{t+1} tf(i,w)}{|V| + \sum_{i=1}^{M} \big[p(d^{(i)} = j|x^{(i)})_{t+1} \cdot \sum_{s=1}^{V} tf(i,s)\big]}$$

The multinomial and gaussian parameters $y \in (p,b,i,t)$ are updated as follows:

$$p(y^{(i)} = l|d^{(i)} = j)_{t+1} = \frac{1 + \sum_{i=1}^{M} p(d^{(i)} = j|x^{(i)})_{t+1} \mathbf{1}_{c^{(i)} = l}}{|Y| + \sum_{i=1}^{M} p(d^{(i)} = j|x^{(i)})_{t+1}}$$

$$\mu_j^{t+1} = \frac{\sum_{i=1}^{M} p(d^{(i)} = j|x^{(i)})_{t+1} t^{(i)}}{\sum_{i=1}^{M} p(d^{(i)} = j|x^{(i)})_{t+1}}$$

$$(\sigma_j^{t+1})^2 = \frac{\sum_{i=1}^{M} p(d^{(i)} = j|x^{(i)})_{t+1} (t^{(i)} - \mu_j^{t+1})^2}{\sum_{i=1}^{M} p(d^{(i)} = j|x^{(i)})_{t+1}}$$

$$p(d_j)_{t+1} = \frac{p(d^{(i)} = j|x^{(i)})_{t+1}}{M}$$

Results

Now, once the parameters of the model have been estimated for the complaints by EM, we can use the model to find the words, products, banks, and complaints that are most likely given a defect. The results below show the top companies, issues, products, and words associated with each defect. The EM algorithm returns the posterior densities of the entities conditioned on each defect. Selecting the complaints with the largest posterior densities, we are able to synthesize the latent defects into a vector of representative attributes.

Table: Representative Entities Per Defect

	Defect 1	Defect 2	Defect 3
Company	CoreLogic Experian USAA Savings	PNC Bank N.A. Barclays PLC Seterus, Inc.	OneMain Financial Holdings, LLC,Toyota Motor Credit Corporation, Amex
Issue	Communication tactics Closing/Cancelling account Managing	Loan servicing, payments, escrow account	Money was not available when promised, Taking out the loan or lease, Improper use of my credit report
Product	Mortgage, Credit reporting	Debt collection, Bank account or service	Consumer Loan, Credit card
Word	credit, information, accounts,verify disputed, rude	loan, mortgage, modification, debt collector	zone, delay experience,unauthorized, expired

The results of our topic model analysis proved interesting. The defects appeared unique and provided a means of summarizing issues across banks and institutions. For a global run, excluding the date entity, on a limited data-set $\approx 20\%$ of the complaints, we extracted the following entities.

Table: Representative Entities Per Defect

	Defect 1	Defect 2	Defect 3
Company	CoreLogic Experian USAA Savings	Ocwen Barclays PLC Seterus, Inc.	OneMain Financial Holdings, LLC,Toyota Motor Credit Corporation, Amex
Issue	Communication tactics Closing/Cancelling account Managing	Loan servicing, payments, escrow account	Money was not available when promised, Taking out the loan or lease, Improper use of my credit report
Product	Mortgage, Credit reporting	Debt collection, Bank account or service	Consumer Loan, Credit card
Word	credit, information, accounts,verify disputed, rude	loan, mortgage, modification, debt collector	credit, delay experience,unauthorized, information expired

On a per product scale, we found similarly interesting results but noticed that by restricting the product universe, the defects found were more similar. Nevertheless, it is helpful in distinguishing which companies most commonly have defects associated with particular product. An example below shows the results using the debt collection complaint universe:

Table: Debt Collection Representative Entities

	Company	Issue	Prior	Word
Defect 1	Encore Capital Group Capital One JPMorgan Chase & Co.	Disclosure verification of debt, False statements or representation, Taking/threatening an illegal action	0.170263252	debt, threaten,letter
Defect 2	Citibank Encore Capital Group Navient Solutions, Inc.	Communication tactics, Cont'd attempts collect debt not owed, Improper contact or sharing of info	0.321562717	debt,number,phone
Defect 3	Encore Capital Group Bank of America Citibank	Cont'd attempts collect debt not owed, Disclosure verification of debt, False statements or representation	0.508174031	debt,credit,account

Another useful result of our model is that it allows us to find the most representative complaint overall given a defect. For example, for Credit Cards, the defect found with the highest prior had the top issue "Account Opening, closing or management," with the representative complaint narrative:

> *I cancelled my XXXX phone service XX/XX/XXXX. XXXX had provided me signal booster as I was not getting enough signal at my home. When I cancelled my XXXX service I requested for them to send me Signal booster return kit in order for me to return the signal booster to XXXX. The return kit simply contains the XXXX return label. But even after 1 month I never received the return Kit. So, I called XXXX again and again for the issue. Every time their response was that they forgot to send the return kit and will send me as soon as possible. At that time, I informed them that the delay is from their end ...]*

Conclusion

The results of the algorithm are exciting and warrant future research and discovery. Given limitation in computation power and time, we were unable to process all of the complaints in the CFPB in a single trial for this paper. Further work will include the whole dataset. Secondly, a more complicated topic model could be implemented that includes background noise as a generator for the complaint entities. Thirdly, we hope to eliminate our smoothing procedure for the prior, fixing an ϵ, and rather throw out a defect cluster if the prior is too small. Lastly, we aim to expand the dataset, incorporating real-time Twitter data, in order to cross-validate the information quality relayed through our defects.

Acknowledgments

We thank our advisors at the CFPB, Professors Kay Giesecke and KK Jain, and teaching assistants Apaar Sadhwani and Enguerrand Horel for their guidance and support.