**### Project Problem and Hypothesis:**

The Consumer Financial Protection Bureau (CFPB) provides a public record of consumer complaints against financial services firms of all kinds. Not all companies respond to consumer grievances without the CFPB getting involved. Can we identify types of cases that are going unaddressed (i.e., receiving untimely responses at a constant or increasing rate)? If so, we will gain insight into the drivers preventing consumers from receiving prompt consideration of their concerns.

The value to be predicted is binary (whether or not the accused responded within 15 days), and I think the top predictors of timeliness will be (1) product type, (2) issue type, (3) defendant (some companies may have developer more efficient processes to respond quickly, and (4) zip code of the plaintiff (claims from individuals in wealthier areas may be taken more seriously).

**### Datasets:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Source:** | <https://cfpb.github.io/api/ccdb/fields.html> |  |  |
|  |  |  |  |
| **Field Name** | **Description** | **Data Type** | **Notes** |
| **Date received** | The date the CFPB received the complaint | date & time |  |
| **Product** | The type of product the consumer identified in the complaint | plain text | This field is a categorical variable. |
| **Sub-product** | The type of sub-product the consumer identified in the complaint | plain text | This field is a categorical variable. Not all Products have Sub-products. |
| **Issue** | The issue the consumer identified in the complaint | plain text | This field is a categorical variable. Possible values are dependent on Product. On December 18, 2013 the issues for student loan complaints were revised. “Repaying your loan” and “Problems when you are unable to pay” were removed and “Can’t pay my loan” and “Dealing with my lender or servicer” were added. Complaints received beginning on that date reflect this change. Complaints received before that date remain unchanged. |
| **Sub-issue** | The sub-issue the consumer identified in the complaint | plain text | This field is a categorical variable.  Possible values are dependent on product and issue.  Not all Issues have corresponding Sub-issues.  On December 18, 2013, sub-issues were added for student loan complaints. Previously, sub-issues were not used for this type of complaint. Complaints received beginning on that date reflect this change. Complaints received before that date remain unchanged. |
| **Consumer complaint narrative** | Consumer complaint narrative is the consumer-submitted description of "what happened" from the complaint. Consumers must opt-in to share their narrative. We will not publish the narrative unless the consumer consents, and consumers can opt-out at any time. The CFPB takes reasonable steps to scrub personal information from each complaint that could be used to identify the consumer. | plain text | Consumers' descriptions of what happened are included if consumers consent to publishing the description and after we take steps to remove personal information. |
| **Company public response** | The company's optional, public-facing response to a consumer's complaint. Companies provide a public response to the CFPB, for posting on the public database, by selecting a response from a set list of options. | plain text | Companies' public-facing responses to complaints are included if companies choose to publish one. Companies may select a public response from a set list of options as soon as they respond to the complaint, but no later than 180 days after the complaint was sent to the company for response. |
| **Company** | The complaint is about this company | plain text | This field is a categorical variable. |
| **State** | The consumer’s reported mailing state for the complaint | plain text | This field is a categorical variable. |
| **ZIP code** | The consumer’s reported mailing ZIP code for the complaint | plain text | Mailing ZIP code provided by the consumer. This field may: i) include the first five digits of a ZIP code; ii) include the first three digits of a ZIP code (if the consumer consented to publication of their complaint narrative); or iii) be blank (if ZIP codes have been submitted with non-numeric values, if there are less than 20,000 people in a given ZIP code, or if the complaint has an address outside of the United States). |
| **Tags** | Data that supports easier searching and sorting of complaints submitted by or on behalf of consumers.  For example, complaints where the submitter reports the age of the consumer as 62 years or older are tagged, ‘Older American.’ Complaints submitted by or on behalf of a servicemember or the spouse or dependent of a servicemember are tagged, ‘Servicemember.’ Servicemember includes anyone who is active duty, National Guard, or Reservist, as well as anyone who previously served and is a Veteran or retiree. | plain text |  |
| **Consumer consent provided?** | Identifies whether the consumer opted in to publish their complaint narrative. We do not publish the narrative unless the consumer consents and consumers can opt-out at any time. | plain text | This field shows whether a consumer provided consent to publish their complaint narrative, as listed below:  Consent provided:Consumer opted in to share their complaint narrative.  Data populates in this field 60 days after the complaint was sent to the company for response or after the company provides an optional company public response – whichever comes first, and after steps have been taken to scrub personal information from the complaint narrative.  Consent not provided:Consumer did not opt-in to publish their complaint narrative. Data populates in this field 60 days after the complaint was sent to the company for response or after the company provides an optional company public response – whichever comes first.  Consent withdrawn:Consumer opted in to publish their complaint narrative and later withdrew their consent.  N/A: Consumers did not have the option to publish their consumer complaint narrative or the complaint was received before March 19, 2015. Data populates in this field immediately.  Other: Complaint does not meet criteria for narrative publication.  Blanks appear until at least 60 days after the complaint is sent to the company for response or until the company provides an optional company public response – whichever comes first. |
| **Submitted via** | How the complaint was submitted to CFPB | plain text | This field is a categorical variable. |
| **Date sent to company** | The date the CFPB sent the complaint to the company | date & time |  |
| **Company response to consumer** | This is how the company responded to the complaint | plain text | This field is a categorical variable. |
| **Timely response?** | Whether the company gave a timely response | plain text | yes/no |
| **Consumer disputed?** | Whether the consumer disputed the company’s response | plain text | yes/no |
| **Complaint ID** | The unique identification number for a complaint | number |  |

**### Domain Knowledge:**

I don’t have direct prior experience working on this issues, but I have previously used Python to predict binary outcomes using logistic regression.

Because the database is relatively new (first established in 2011), little academic research exists. However, Yale Law School published a [2014 study](http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=5877&context=fss_papers) from Ian Ayres, Jeff Lingwall & Sonia Steinway of Yale Law School in which the authors used this database to illustrate disparities in complaint responsiveness and related variables. The authors demonstrate two broad findings that are directly relevant to my research. First, they show that the individual defendant matters in driving responsiveness. Specifically, they show that twenty-seven firms, including large banks such as Bank of America, Citibank, and PNC Bank, were significantly less likely to respond to consumers in a timely manner than the average accused party. Second, the authors argue that variables related to the individual plaintiff’s identity are also relevant in explaining variation in response times. Specifically, they show that defendants were significantly less likely to provide timely responses to plaintiffs filing from zip codes with higher percentages of black and hispanic residents, senior citizens, or college students.

The Ayres, Lingwall, and Steinway study provides the following benchmarks for my study:

1. Association between defendant-level variables and complaint response timeliness:
   1. 19 companies demonstrated untimely response rates at 5% significance
2. Association between plaintiff-level variables and complaint response timeliness:
   1. Procedure:
      1. Created dummy variable indicating whether the defendant’s response was timely or not
      2. Ran five logistic regressions, progressively adding additional additional IVs and controlling for:
         1. Product/issue occurrence frequency
         2. Fiscal quarter
         3. State
         4. Company
   2. Results:
      1. Demographic variables significant after controls:
         1. College enrollment (p < 5%)
         2. % high school graduates (p < 10%)
      2. Product categories significant at 5% level:
         1. Bank account or service
         2. Credit card

**### Project Concerns and Risks:**

Concerns:

1. I’m not sure how to control for variables as the authors did.
2. I don’t have yet have data for income (or other demographic information) as it relates to the zip codes provided in the dataset
3. I was surprised to find this (quite comprehensive) study, and I’m not sure how my work will add unique value beyond what has already been done in this study.

Caveats:

1. Consumers self-select into filing a complaint, so I’ll need to be careful about trying to infer from the sample of those who filed a complaint to all consumers with grievances (or, even more so, all consumers in the United States)
   1. Those with the ability and persistence to file a complaint may differ in significant ways, such as in having better internet access or more free time than the overall set of aggrieved consumers
   2. Complaint rates may not be consistent across industries, potentially skewing the data toward industries where aggrieved consumers are more likely to file
      1. This could be due to differences in (i) the importance of the service to a consumer’s financial circumstances, (ii) consumers’ awareness of whether a given firm is regulated by the CFPB, and/or (iii) the ease with which consumers can notice errors/inconsistencies in the service provided by the financial firm
2. The CFPB vets all filings to ensure that the complainant has a business relationship with the firm in question, but this verification process is not foolproof and may be more or less accurate for different product areas
3. The CPFB does not investigate the merit of claims before reporting them to defendants, so there may be effects from certain product areas being associated with higher (or lower) rates of unsubstantiated complaints

Risks:

* If the model is wrong, we will find spurious associations between delayed responses to consumers and zip code, defendant, or product/issue type. This could lead to making unwarranted inferences about income discrimination, company-specific incompetence, or certain product/issue types being more difficult than others to address in a timely manner.
* None of the data looks to be erroneous, but there appear to be a small number of null values scattered throughout various columns.

**### Outcomes:**

The output should be a logit regression with coefficients and p-values for each IV.

I don’t have a target audience per se, but the general public would expect a summary of which variables were significant (i.e., p < 5%).

If the project is a bust, we’ll be able to say that none of the IVs in the dataset significantly predict whether the consumer will receive an untimely response. This is not a problem per se; it just indicates that none of the available variables are particularly helpful in predicting this outcome.