Are there unidentified frequent issues among student loan providers?

Dan Ludwig *Faculty Advisor: Zsolt Ugray*

Abstract

There are millions of students, many of whom cannot afford to pay for their schooling without a loan. Student loans are an important business for financial institutions. It's common for students to run into issues at some point when dealing with their loan providers and servicers. I'm using Natural Language Processing (NLP) Topic Modeling algorithms to analyze customer complaint narratives to identify important emerging problems beyond the ones defined prior to complaint collection.

Background

The Consumer Finance Protection Bureau (CFPB), started by the United States government in 2011, collects and publishes data on consumer complaints against financial companies, including complaints related to the handling of student loans. These complaints are grouped by and analyzed by predefined issues. Some important student loan handling issues may not be listed in the dataset, which should be added to provide a more accurate understanding of frequent problems with the student loan providers. The CFPB dataset organizes consumer complaints into predefined common issues to see who are having what issues and how many. I performed a large- scale text analysis of consumer's complaint narratives to find if there are any existing prominent issues related to student loan servicing that had not been defined or identified.

Methods

Data Collection and Exploration:

- Used the CFPB's public dataset for complaints against student loan providers. (59,155 rows of data; 26,854 having narratives).
- Used python's 'pandas' and 'matplotlib' packages to initially perform exploratory analysis to break down data into digestible amounts of information.
- Found 11 distinct categories of issues that each complaint was assigned to which I can use to compare to the topics I find.

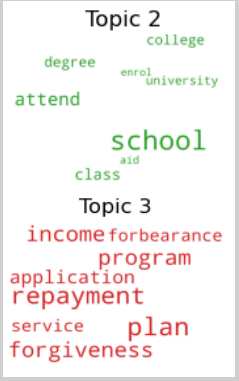
Data Preprocessing:

- Since text data is messy, I coded a function to clean the data by breaking down each word to its root meaning and keeping only the relevant words for each narrative.

Choosing and Tuning a Model:

- Used Latent Dirichlet Allocation (LDA) to read and group each narrative.
- Used a coherence score to measure the model's accuracy (the higher the score the better the model is fit to the data).
- To maximize the score, I built functions to test different numbers of topics and tune the model's features and give me the parameters for the highest score.

Existing Issues	count
Can't repay my loan	3057
Credit monitoring or identity theft protection services	39
Dealing with my lender or servicer	7909
Dealing with your lender or servicer	9348
Getting a loan	467
Improper use of your report	21
Incorrect information on your report	539
Problem with a credit reporting company's investigation into an existing problem	238
Problem with fraud alerts or security freezes	8
Struggling to repay your loan	3562
Unable to get your credit report or credit score	7



Topic #	Issues	count
0	Agreement Disputes	2093
1	Ameritech Issues	1
2	University Attended Closed Early	262
3	Dept Forgiveness	2267
4	Credit Issues	1023
5	Need New Payment Plan	10169
6	Issues Processing Payment	2727
7	Customer Service Issues	8312

Results

The model with the highest coherence score gave 8 distinct topics, two of which aren't listed under the original issues. These topics are *i) debt forgiveness programs not following through*, and *ii) the university attended had closed early*. These are two topics that should be recognized by the CFPB so they can help make sure the financial institutions are handling these problems in an appropriate manner and the consumers aren't being taken advantage of.