The Voice of Customers

CFPB data analysis from 2005 by now

Danfeng Wang
Big Data Analytics
Letterkeney Institute of Technology
Donegal, Ireland
L00162010@student.lyit.ie

***Abstract*—This is a customer complaint dataset about the financial product consumption in the USA. We would like to hear the voice from these customers by analysing it. How many valid complaint data in this dataset? What is the most concerned problem? What is the strongest voice of these complaints? Have those financial companies improved their service from this? Can we predict the trends of the complaints? With these questions, we are going to have a exploration to this dataset. Please take a deep breath, hear the voices from the customers. Let the millions of complaints records tell us the truth, charts never lie.**

I. Introduction

In this project, we are going to analysis the consumers complaints about financial products in the USA. The dataset of this project obtained from the data.gov website which is collected by the Consumer Financial Protection Bureau, an agency of the United States government responsible for consumer protection in the financial sector [1]. It includes complaints dating from 2015 by now.

The volume of this dataset is roughly 2 million items. We apply big data analytic related techniques, such as machine learning, natural language process (NLP) and so force, to achieve the goal of the project:

(i) providing descriptive analysis combined with visualization, such as bar charts and word cloud charts, to give an insight and summarization of these data.
(ii) providing prediction model by machine learning approaches.

- What's the most complained from different dimensions, such as complained channels,companies, products. Have they changed over time?
- Pictures speak louder than text, what can we get from the narrative of those complaints.
- Can we predict the issue type? Can we predict the process result to guide the urgency of these complaints, so that the companies can give a higher priority when dealing with the urgent complaints.
- The quality of service - responding time, where, what and which providers got the most complaints. Any improvement later?

II. Code list

III. System preparation

We apply big data and machine learning related techniques to implement this project.

In the process of achieving the above goal, we have three aspects to consider:

Data engineering

- - data ingestion
- - data preparation
- - visualization

Machine learning

- - data loading
- - feature engineering
- - training and prediction
- - model tuning

Databrick is a cloud platform for big data analysis. In this project, we use Databrick as the data storage, analysis and modelling platform. Follow the next steps for system preparation.

- Sign up for a free Databrick trial.
- Create a cluster 6.4 (includes Apache Spark 2.4.5, Scala 2.11)
- Create notebook.
- Start cluster. The cluster will be terminated automatically after 60 minutes of inactivity.

At the data modelling stage, we use MLFlow on Databrick to implement the machine learning pipeline. MLFlow is an platform for managing the machine learning lifecycle.

IV. Dataset acquisition

- Download dataset file from the web site.
- Upload the file into DBFS.
- Unzip and load the data into table.
- Check the integrity.

During this process, we found the csv format cannot be correctly loaded by spark SQLContext, . By reading few line from the file, we found the fields are seperated by comma, one of the descriptive field *Consumer*

complaint narrative is enclosed by double quotes. By applying loading parameters such as *escape* and *quotion*, the problem remained. In order to avoid this problem, we tried loading json file.

A. Fields description

see Table1.

- *Company public response* - standard items. This is how the company attributed the complaint. Some are believed the company should not be responsible for the issue, it should be a third party, policy, consumer, misunderstanding and so forth. Some are resolved privately which we can not distinguish the attribute, and some are believed invalid complaints.
- *Company response to consumer* - The process status is close, in progress or untimely response in which the close status has several different way including close with explanation, monetary relief, non-monetary relief, relief or without relief. We can see percentage of issue resolved and in which way the issues have been resolved.

TABLE I
DATASET

Category[1]	Column name	Description[2]
identity	complaint_id	numeric string
company	company	string
	date_sent_to_company	yyyymmdd
	company_public_response	
	company_response	category
	consumer_consent_provided	category
	timely	yes/no
product	product	string
	sub_product	string
consumer	submitted_via	channel/category
	issue	category
	sub_issue	
	complaint_what_happened	long text
	tags	string
	consumer_disputed	yes/no
	state	category / geo
	zip_code	category / geo

[1] The business purpose of the column.
[2] The column type and other description.

B. Business process

The complaint data will go through the following steps when it is submitted to the CFPB, see Figure 1.

C. More about the data

The complaint facts will not be verified by CFPB, but it gives the companies opportunity to confirm the consumer has commercial relationship with them.

V. DATA PREPARATION

In order to make it easier to manipulate in the next stages, the dataset will be prepared as follows:

	0	1
date_received	2019-09-24	2019-09-19
product	Debt collection	Credit reporting, credit repair services, or o...
sub_product	I do not know	Credit reporting
issue	Attempts to collect debt not owed	Incorrect information on your report
sub_issue	Debt is not yours	Information belongs to someone else
complaint_what_happened	transworld systems inc. \nis trying to collect...	
company_public_response		Company has responded to the consumer and the ...
company	TRANSWORLD SYSTEMS INC	Experian Information Solutions Inc.
state	FL	PA
zip_code	335XX	15206
tags		
consumer_consent_provided	Consent provided	Consent not provided
submitted_via	Web	Web
date_sent_to_company	2019-09-24	2019-09-20
company_response	Closed with explanation	Closed with non-monetary relief
timely	Yes	Yes
consumer_disputed	N/A	N/A
complaint_id	3384392	3379500

Fig. 1. sample data

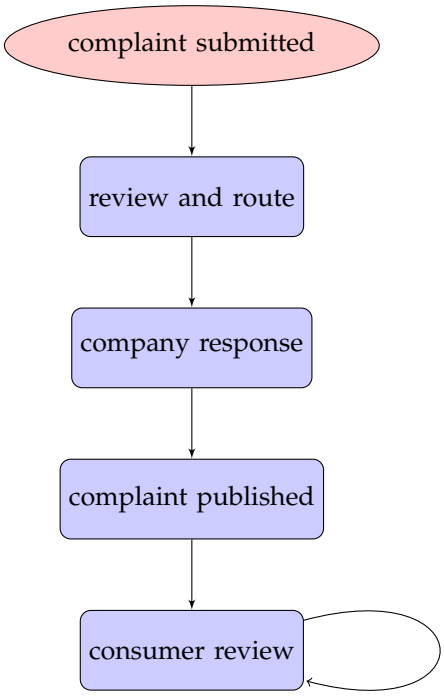

Fig. 2. data flow

A. Data integration

The purpose of this stage is to make sure the dataset is analysable.

By looking at the data,

- check null value to target fields.
- validate the date fields.
- Check the value of state fields.Standardizing?

item type conversion.
item eliminating meaningless words such as stop words, symbols, and convert the text into lower case.

B. Data transformation

- date: generate corresponding fields such as year, month, week and so forth.
- location: generate geographical columns, such as longitude, latitude and so forth.
- standardized text: convert text to code. Split the text into words, extract nouns.

VI. Descriptive Analysis

In this process, we visualized the data from different dimensions to see how many complaints happened, as well as the trend over time.

For example, the figure below shows the top 6 complained products by year. During the first 4 years, most complaints mainly focus on the mortgage (green bar), and the amount of complaints declined gradually after 2013 and remains static from 2018. While in 2017, the *credit reporting* product (yellow bar) became the most-complained-about financial product, and it has been growing rapidly since then.

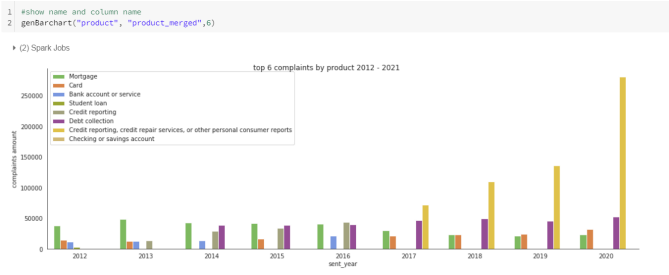

Fig. 3. top 6 complained product by year

In this way, the complaints change over years can be shown by other dimensions, such as issue type, channel, companies, and so on and so forth. So functions have be defined here for further use.

VII. NLP

As we listed in the *table1*, the *complaint_what_happened* column is the consumer's narrative of the complaint. We apply NLP analysis to dig into this column for a deep insight.

The pipeline of dealing with NLP usually consists of several steps:

1) normalization - convert the text content into lower case and remove the non words such as punctuation and other non alphanumeric characters. The private information of this column has been replaced by x with the corresponding length. In order to remove them, they were added into the stop words array.

```
def genBarchart(name, group_field_name, top_n=4):
  df = getRankByYear(group_field_name)
  years = df["sent_year"].unique()
  #begin year
  year1 = years[0];
  #end year
  year2 = years[-1]
  with sns.axes_style('white'):
    #show complaints amount by group_field_name by year
    g = sns.catplot(x="sent_year", y="count", data=df, aspect=3.0, kind='bar',
                    palette=palette_colors,ci=None,hue=group_field_name,
                    order=range(year1,year2),sharex=False,legend_out=False)
    g.set_ylabels('complaints amount')
    plt.legend(loc='best')
    #title - top n complaints by xxx begin_year - end_year
    g.fig.suptitle('top {} complaints by {} {} - {}'.format(top_n,name,year1,year2))

# group count and get top n of each group
def getRankByYear(field_name, top_n=4):
  #top n complaints amount by year by field_name
  df_topn = sqlContext.sql("SELECT year(date_received) as sent_year," + \
                field_name +\
                " ,count(1) AS count " +\
                " FROM t_complaints " +\
                " GROUP BY year(date_received)," + field_name)
  w = Window.partitionBy("sent_year").orderBy(F.col("count").desc())
  df_topn_pd = df_topn.select("*",F.row_number().over(w).alias('rank')).\
    filter(F.col('rank') <= top_n).orderBy(F.col("sent_year")).toPandas()
  return df_topn_pd
```

Fig. 4. chart function

2) tokenization - split the sentences into words by space.
3) stop words - remove meaningless words.
4) tagging - only keep specific word types, such as noun, adjective.
5) stemming - reducing the words to it's word stem so that the similar words lie under a same stem.

A pipeline is created for achieving the above steps.

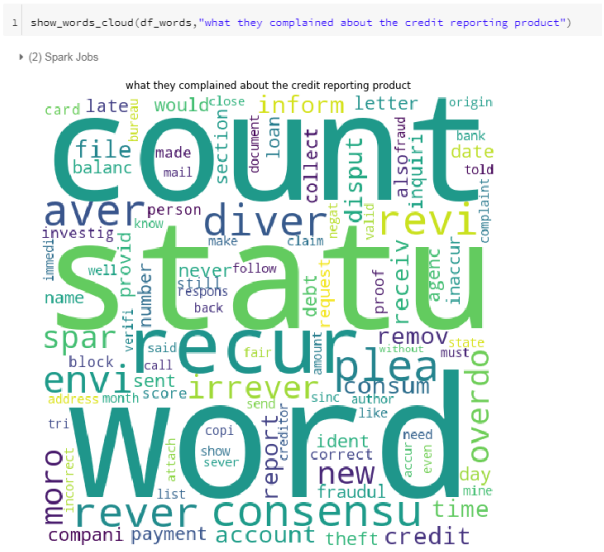

Fig. 5. the voice of credit reporting product

VIII. Predictive Analysis

When exploring this dataset, we found the part of within company response column, there are several ways for the company to deal with the issues. Some of the response involved with money or other non-monetary relief. It would be a valuable reference for the companies complained by the customers if we are able to predict the

settlement whether involves with any forms of relief or not?

In order to build predictive model to try to implement this task, we will follow the steps as below:

Data preparation - after the descriptive analysis stage, the dataset has been prepared and saved into database, we can use pyspark.sql api to retrieve data conveniently. For the narrative column, the sentences of that column has been processed by a nlp pipeline.

Feature engineering - From the dataset introduction section, we have known that most of the features are nominal categorical data, such as issue, sub-issue, product, company response, and so force. We will convert these columns into numeric by applying one hot encoding and create dummy variables of each categorical. By transforming these different values to columns, we can avoid the model regarding these numeric values as number instead of a specific category, because the number itself is meaningless [4].

Cross-validation - we prepare the dataset by using 5-folder cross-validation. This is a popular method which tends to provide less biased model than a simple $train_test_split$.

Algorithm selection - This is a classification prediction problem, we apply Logistic Regression and Random Forest model. With Random Forest, we don't have to think of the dimensionality deduction, while the disadvantage is the computational complexity. Logistic Regression as a regression model is simple and has good interpret ability.

model tuning - We will apply PCA(Principal Component Analysis) to do dimensionality reduction and apply Logistic Regression on the PCA dataset.

model evaluation - At this final stage, we will give an evaluation for the above models. We will explain the prediction capability of the model.

IX. Conclusion

From the above analysis, we have heard the voices from the customers in consuming financial products. From this process of manipulating this dataset, we have known the pipeline of dealing with big data. Big data is able to offer us insights into what happened in the financial products domain as well as the prediction of what will happen.

- feature selection
- one-hot encoding (Bag of words) - It represent the sentence by replacing each unique word into a specific number. With this Bag of Words model, the sequence of the words is ignored.
- relevant - plot the information. Since we have many categorical features in the dataset, we apply PCA to do dimensionality reduction.
- classification - Logistic Regression, which is simple and explainable.

X. Future work

For further analysis, analysis the data base on geographical features will be done. Moreover, the narrative of complaint is used to output word cloud chart. In order to make full use of the narrative content, the sentiment analysis can be applied further.

References

[1] https://catalog.data.gov/dataset/consumer-complaint-database. *The LaTeX Companion*. The Consumer Complaint Database, Metadata Updated: July 17, 2020.

[2] https://cfpb.github.io/api/ccdb/fields.html *The LaTeX Companion*. The Consumer Complaint Database fields description.

[3] ("Summary of product and sub-product changes," n.d.)

[4] Categorical Data. Strategies for working with discrete… | by Dipanjan (DJ) Sarkar | Towards Data Science [WWW Document], n.d. URL https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63 (accessed 3.3.21).