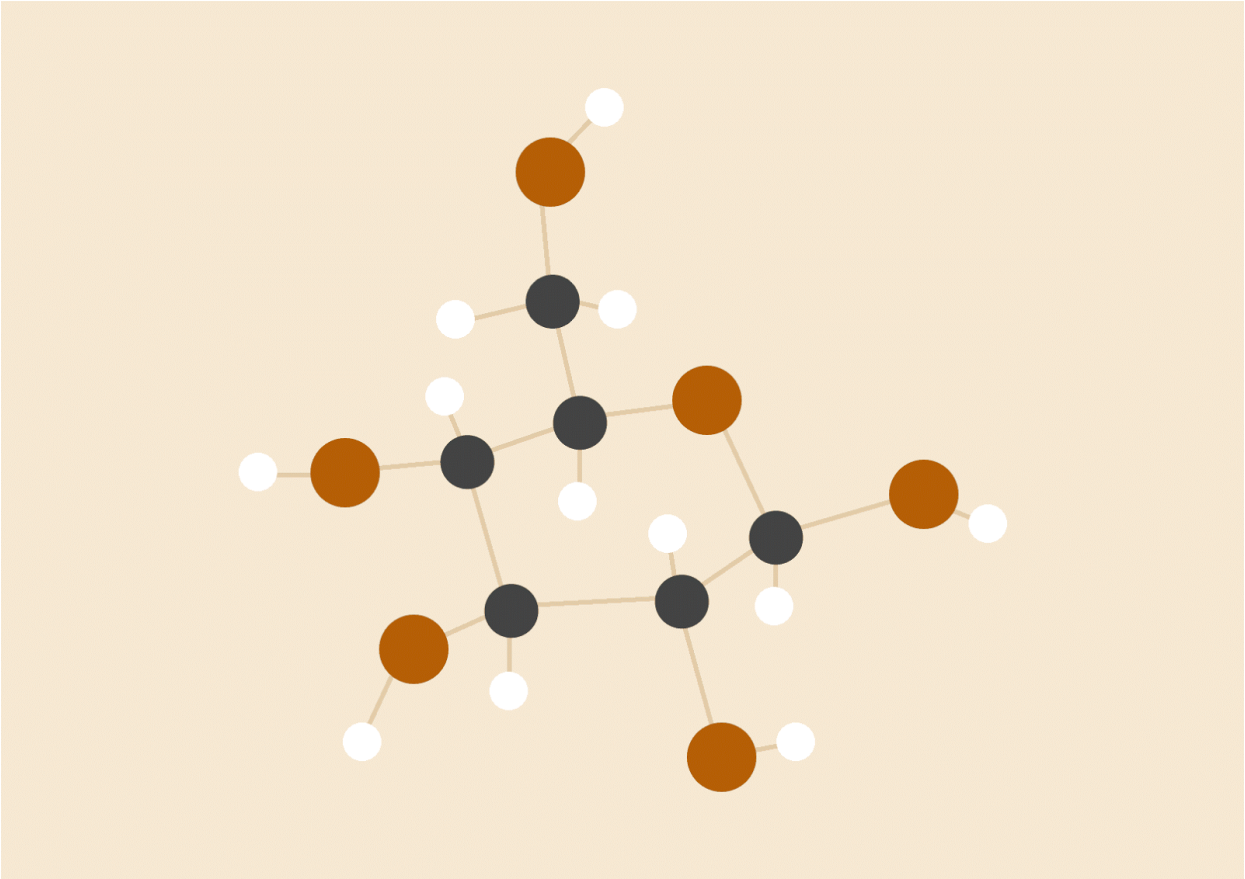
Final Progress Report



Shringa Bais
Swapnil Sopan Gaikwad
Joel Ponder
Lawrence Amadi

06.05.2017

Introduction

Analysing Consumer Financial Protection Bureau (CFPB) database which records complaints about financial products and services to companies for response. Every complaint provides insight into problems that people are experiencing, helping CFPB identify inappropriate practices and allowing us to stop them before they become major issues. and creating an **automated customer response system.**

1- Primary Goal

We are majorly trying to create a model and able to categorize data from the given input and predict their respective categories. There are primarily 3 categories on which we are focusing:

1) **General Explanation Case** - where complaints not regarding to money.
2) **Non-Monetary Relief Case** - where complaints regarding to money but because of some reason customer not get any monetary relief.
3) **Monetary Relief Case** - where complaints regarding to money and customer get monetary relief after complaint and data provided checked.

2- Complaint Status

Closed (all types)	In Progress	Untimely response
665717	613	3448

Here we can see that 3448 complaints have 'Untimely response' and 613 are still 'In Progress'. If we do good automation for current system, we can also minimise these counts greatly.

3- Data

Data has been categorized into below shown categories:

Response Category	**Count**
Closed	15746
Closed with explanation	496628
Closed with monetary relief	45248
Closed with non monetary relief	85242
Closed with relief	5252
Closed without relief	17601
In Progress	613

Untimely response	3448

Here we can see that most of the complaints (count = 496628) are closed with explanation. We can take help of history data for such explanation as it is not involving any money. It just contain explanation. So we can give response within very less time for such cases.
For monetary relief we can see that 45248 cases are there.
Fo non monetary relief we can see that 85242 cases are there, which is double in count of monetary relief cases.

Experiments

1- Manual analysis

As per our first discussion with TA and Irina it was very important for us to look into data and find out relation between the given columns which will further help in analysing.
Here we observed and manipulated(using basic excel filtering) the .csv file to know more about important columns for data analysis and classifications. Look after each category of data required in experiments and find out its count.Checked for similar columns which could be added as additional features in model.

1.1 Compliant data in each category.

Complaint Category	Closed with monetary relief	Closed with non monetary relief	Closed with explanation	**Total Count**
Credit Card	267	120	733	1120
Debt Collection	44	386	1329	1759
Mortgage	90	201	3241	3532
Money Transfer	7	1	60	68
Student Loan	4	17	99	120

1.2 Actual Data vs Required data in database

Company_response_to_consumer	**Normal Count**	**Count with Consumer complaint narrative description present**
Closed with explanation	21749	8055
Closed with monetary relief	2173	290
Closed with non-monetary relief	3048	1766

2- TF-IDF

Me and Swapnil both opted for this approach as in earlier work of text classification this approach gets to be very helpful. However, when we started working on vectorization and applying cosine similarity I have also cross checked our output by using inbuilt functions as **TfidfVectorizer** by stemming and also applying **SVD** as preprocessing and dimensional reduction steps and another approach by creating function is explained below (Issues faced in both the approaches is explained in analysis part):

- We have taken only those rows of Column - 'Consumer complaint narrative' in which complaint description is present.
- Data preprocessing

1. Tokenise
2. Remove numbers
3. Stop word removal
4. Too low and too high frequency words removal

Preprocessing Vocabulary

Preprocessing Steps	Vocabulary Size	Time for calculations
Without preprocessing	66080	50 min
Only Characters, No Numbers	64201	48 min
Removal of stop words	62016	31 min
Removal of small length (<4) words and high length words (>15)	51989	20 min

- Build term-document matrix

1. Build a vocabulary
2. Added row tokens for each complaint
3. Calculate TF-IDF value for each token in a row
4. Build a CSR matrix

- Split data into test and train set
- Build classifier on train dataset i.e. history of complaints and their resolutions
- Check accuracy using column name 'Company Response to Customer'.
- Built a classifier test data i.e. treating it as a future complaints.

Approach points

- We took 10000 history data for training set. We divided training and test set on the basis of row number as well as on the basis of time. Above table shows data division.

- Inbuilt functions of TF- IDF Vectorization needs vocabulary size to be fixed and passed a as a parameter in function which is causing to create proper matrix that will calculate proper accuracy of data.
- Each word has considered as a feature for tf-idf calculation.Example - We can select features by their frequencies.
- We built a matrix for each row of size (1 * Size of vocabulary), most of the features are having zero values. So for storage efficiency we used CSR matrix.
- We are using columns named **'Product'** and **'Issue'** for finding similarity between complaints.

Predicting Accuracy

Train data size	Test data size	Misclassified complaints	% Error	% Accuracy
80	20	3	22 %	78 %
1000	200	43	40 %	60 %
4500	500	178	37%	63%
9000	1000	414	41 %	59 %
60000	5000	975	32 %	62 %

Time based train and test set accuracy change

Train Data Size / Test Data Size	Closed with explanation	Closed with non-monetary relief	Closed with monetary relief	Closed without relief	Closed with relief
1000/100	848/87	107/9	23/4	5/0	17/0
2000/200	1737/173	201/18	27/6	12/3	23/0
3000/300	2601/233	296/37	53/23	15/5	35/2
4000/400	3455/367	387/46	83/27	34/9	41/4

Analysis

- Here we found out that as the vocab contain most usual words, numbers, noise etc. , the calculation time also high for such vocabulary.

- As we preprocess data, the vocabulary size get reduced and calculation time also get reduced.
- As we increase the train and test dataset, we are getting equal or more than 60 % accuracy.
- After preprocessing of data, we get good amount of reduction in processing time.
- The error percentage is around 40% for increased size of train data-set. Here we can say that as higher the number of train data, the good model can be built as include more different types of examples.
- Misclassified complaints are still more as we can see there is more weightage of 'Closed with Explanation' category and hence for other categories we can see very few accuracy percentage.
- Other major features which we have considered while predicting accuracy are **Product, Issue** and **Complaint ID** added in code while enhancing our model to calculate more accurate output.
- **Added SVD part for dimensional reduction ,however while creating a CSR matrix we have used each complaint highest frequency words as its features, so by reducing and increasing dimension won't affect much on our model. Below table shows the SSE error calculated by using SVD for different dimensions.**

The sum of squared errors table

Dimensions of matrices (k)	The sum of squared errors
50	31102.1
60	15814.93
70	4153.263
75	3.253855e-23

Class-wise Error %

- Here we can see that, we haven't all types of test data. Response type Closed without relief, Closed with relief and In progress has no examples in test data-set.
- We can prepare same table for all train and test combinations. We can see from above table that Closed with non-monetary relief, Closed with monetary relief and Closed type of responses have greater error percentage.
- We can ignore error % for Closed type of responses as those complaints mostly does not involve money. Simple explanation from history response to similar complaints will help.
- But we have to focus of Closed with non-monetary relief and Closed with monetary relief type of responses as money involved in it. Those complaints have to check at-least once manually.

Consider below train and test set.

Train data size	Test data size	Misclassified complaints	% Error	% Accuracy

5000	500	152	35 %	65%

Class Wise Error %

Response type or class	Right prediction count	Wrong prediction count	% of wrong predictions
Closed with explanation	239	70	22.653721682847898
Closed with non-monetary relief	15	39	72.22222222222221
Closed with monetary relief	7	12	63.1578947368421

Parameter Tuning

As we have opted for TF- IDF vectorization with cosine similarity to predict the accuracy of our model for the given dataset there is no particular requirement of parameter tuning. Major parameter required in our model are feature and its dimension and we did not fluctuate dimension values as mentioned above in SVD also. Used complete highly processed vocabulary data as features for each row of complaints.

Feature Selection

We have applied Infogain algorithm on the given columns and find out best attributes in the given data set. On the basis of its gini index we have ranked the best attributes : **Company**,**Product**,**Issue**,**Sub-Product**,**Sub- Issue** and **Consumer complaint Narrative**
Used Weka to find out best attributes by using Infogain attribute selector and Ranker as a search method.

Confusion Matrix

Obtained Confusion matrix from our analysis.Training Data size: 30000

Company Response To Consumer	**Closed with Explanation**	**Closed With Monetary Relief**	**Closed With Non-Monetary Relief**	**Closed**
Prediction				
Closed with Explanation	**22152**	**239**	**344**	**2**
Closed With Monetary Relief	**1777**	**258**	**5**	**0**

Closed With Non-Monetary Relief	**4046**	**16**	**385**	**3**
Closed	**688**	**22**	**5**	**0**

By including above features mentioned in Feature selection Confusion matrix has also achieved some of its accuracy level. Training Data size :4000

Company Response To Consumer	**Closed with Explanation**	**Closed With Monetary Relief**	**Closed With Non-Monetary Relief**	**Closed**
Prediction				
Closed with Explanation	**1160**	**17**	**75**	**5**
Closed With Monetary Relief	**100**	**23**	**10**	**2**
Closed With Non-Monetary Relief	**165**	**4**	**88**	**0**
Closed	**30**	**1**	**1**	**16**

Cost Matrix

As per comments provided by Professor on progress report I tried to calculate Cost matrix for the mentioned approach.Find how much our approaches cost if our results shows false values also

Let set up cost matrix for monetary values as the highest:

Cost matrix	Closed with Explanation	Closed With Monetary Relief	Closed With Non-Monetary Relief	Closed
Closed with Explanation	0.0	2.0	1.0	1.0
Closed With Monetary Relief	1.0	0.0	1.0	1.0
Closed With Non-Monetary Relief	1.0	2.0	0.0	1.0
Closed	1.0	2.0	1.0	0.0

For 1st confusion matrix- Total cost =2.0* 239+344+2+1777+5+0+4046+2.0*16+3+688+2.0* 22+5

= 478+344+2+1777+5+0+4046+32+3+688+44+5

= 7,424

Average cost :0.2456

For 2nd confusion matrix- Total cost = 2.0*17+75+5+100+10+2+165+2.0*4+0+30+2.0*1+1
= 34+75+5+100+10+2+165+8+0+30+2+1
= 432

Average cost :0.231

As for both of the Confusion matrices we have taken different training datasets size.therefore Cost matrix values also differed.So, lesser the average cost matrix value more the approach is better.

Accuracy of different Classifiers:

Used Weka tool to apply given data with best selected attributes and applied multiple classifiers trees to check the percentage of correctly classified instances in Confusion matrix.

Classifiers	Random Forest	J48	Decision Tree	Random Tree
Correctly classified instances	75%	73.9%	72%	76%
Incorrectly classified Instances	24.2%	26%	28%	22.2%

Learning:

- Learned to apply Data mining concepts on real world data and how it should be handled
- We could improve our accuracy of the approach by selecting correct set of data.
- We would like to have had more examples from “Closed with Monetary Relief” and “Closed with Non-Monetary Relief”.
- Response to the Complaint may be a result of information from the Company's own data set that we do not have access to. It may not be possible to consistently predict these results.
- Multiple approaches have been opted by our team to analyse the data and we cannot say any of them is not useful, from each analysis we have got information about the data and categorization.
- It is very important to understand your data properly before start applying any algorithms over it.
- My analysis has not taken any classifier into account and has opted to use approach of checking historical values while predicting for the future data.Each approach has its own advantages and disadvantages.

Conclusions:

- Data need to be cleaned properly before processing and building a model.

- Large number of history data affecting performance.whenever analysis is done it's important to have more and more amount of data
- SVD won’t help us more in terms of categorising responses by using TF IDF approach.VD reduces dimensions but in our case more dimension gives more result and dimensionality reduction wouldn’t help much.
- We can’t focus only on TD-IDF values for prediction. While applying different classifier output in weka ,we got to know some other analysis of our data.
- How model results costs get affected by using cost matrix.
- For better result, we can find out the company or region from where maximum complaints received to get quicker response to similar complaints.
- As given data gets filed online by Customers and they provide data from the provided filters, it's not necessary that selected filter option depicts the actual issue. Correct filter data should be provided.