2017

THE POWER OF SAS ENTERPRISE MINER FOR TEXTUAL INTERNAL AUDITING

Team: Boiler Bank Busters (BBB)

Jeffrery (Kewa) Wei, Lukia (Xueyu) Chen, Shiva Biradar

Faculty Sponsor: Matthew A. Lanham (lanhamm@purdue.edu), Clinical Assistant Professor, Purdue University, Department of Management, 403 W. State St., Krannert Bldg 466, West Lafayette, IN 47907

Introduction

In September 2016, Wells Fargo reached a $190 million settlement with federal regulators and prosecutors for opening more than two million deposit and credit card accounts without customer authorization. Of the settlement, $100 million is going to the Consumer Financial Protection Bureau (CFPB), the largest amount the CFPB has ever levied against banks.

Interestingly enough, the CFPB, widely commented as the "major winner" by the public and the press for the disclosure of Wells Fargo scandal, maintains the largest consumer complaints database about 13 major types of consumer financial products and services, including bank accounts or services, credit card, and consumer loans, etc. Since July 2011, the CFPB has collected 647,795 complaints on a range of consumer financial products and services from the public, and sent to nearly 3,000 financial services companies notices for response. The CFPB has used the database to support its regulatory responsibilities including market supervision and enforcement activities, but it is not so clear as to what extent the CFPB has utilized the consumer complaints data. In the monthly compliant reports published by the CFPB in 2016, most of the analysis is based on statistical summary of the consumer complaints data.

The objective of this study is to explore the power of using SAS analytical tools, such as SAS Enterprise Miner, to efficiently generate business insights for organizations. Specifically, we are interested in exploring how financial institutions could utilize text mining for their internal audit function. Our study makes two contributions. First, the models we've developed can help financial services companies to predict business risks and potential issues based on the text data (customer complaints in this case). Secondly, we show that the SAS text mining analytics capabilities of identifying key terms is very powerful at automatically identifying terms compared to manual human reading and tagging of complaints. We posit here that those considering performing text analytics might be skeptical of the results obtained using this tool. Thus, our study helps remove these doubts and should provide confidence for others performing textual modeling building to support decision-making.

Data

The data set investigated in this study are customer complaints against various financial institutions from March 19, 2015 thru September 22, 2016. This data is publically available from The Consumer Financial Protection Bureau (CFPB) (https://data.consumerfinance.gov/dataset/Consumer-Complaints/s6ew-h6mp), and was filtered specifically to "Bank account and services" which corresponds to the focus of our primary research question. It contains 9055 complaints for all the banks. This file contains the following information:

- Customer contact information, partially censored,
- Complained banks' name,
- Details of customer complaints,
- A brief of response from both banks and customers.

Problem

As stated previously the goal of this project is to transform the unstructured complaints into useful information financial institutions and regulators could use for decision support. For this reason, we only keep the column of complaints and discard all other columns. The 9055 complaints were randomly partitioned into two datasets. One set we used for model training and testing purposes, while the other served as a mock scoring set. The data used for training and testing is tagged (i.e. manually categorized) by human individuals.

We developed a structured tagging protocol that provided examples of how to tag consistently among all tagging volunteers (Appendix 3). Next we distributed this protocol to the volunteers, had them read the instructions and begin classifying each complaint they received based on four categories. There were 141 individual taggers (including the authors), there were randomly provided 150 complaints among the 4532 complaints in the training/test set. Based on our design all complaints were read and tagged by multiple individuals. On average each complaint was tagged by 4 individuals.

The final target variable for each of the four unique topics was based on majority voting. This means that if two or more people voted "Yes" on a topic, that complaint topic was assigned as a "Yes," otherwise it was assigned as a "No".

Analysis

Data cleaning

Field	% Agreement	N Agreements	N Disagreement	N Case	Gohen's Kappa
Fraud Complaint	0.688	846	383	1229	0.377
Misleading information or Policy	0.648	797	42	1229	0.297
Unauthorized Transactions	0.741	911	31188	1229	0.183
Penalty Fees due to Insufficient funds and late payments	0.826	1015	214	1229	0.652

Table 1. Kappa score of four issues

Cohen's Kappa statistics were calculated to measure inter-rater reliability of taggers versus the author's tagging. We assume the study author's tagging was the standard which has been done in similar studies. Figure 2 shows that unauthorized transactions and penalty fees due to insufficient funds and late payments have sufficiently good quality as their values are exceed the 0.40 threshold.

Exploratory data analysis

Looking at the text conceptual link, we found that the term "fee" is connected to term late fee, overdraw and its synonyms.

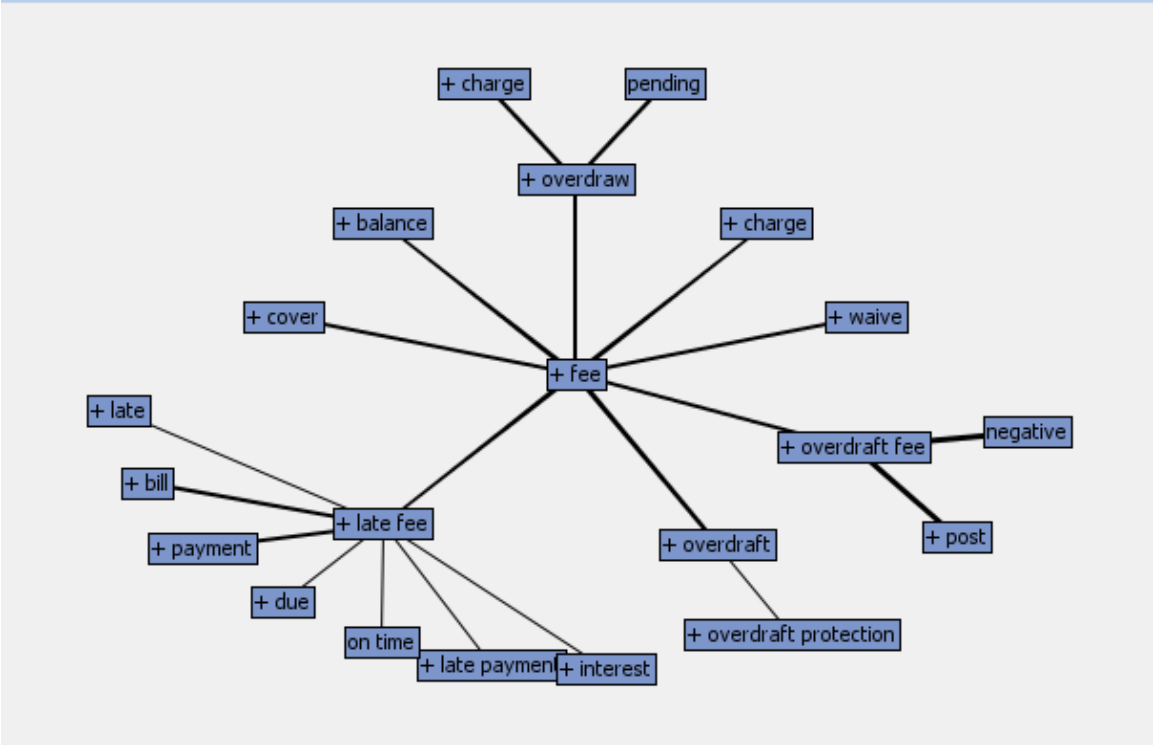


Figure 1. Conceptual link of word "fee".

Model building

We used SAS Enterprise Miner (EM) for our project. The same diagram was used twice for the two variable. As shown in our diagram (Appendix 1a), we created four different paths for either analysis:

1. Build model right after filtering the word. In this method we heavily rely on the text mining capability of SAS EM.
2. Create 25 single-term topics with text topic node and build quantitative model. The topics were created according to the target variable.
3. Create text cluster with text cluster node, using default setting, and build quantitative model. The clusters were created according to the target variable.
4. Use the default setting of text topic node to create 25 single-term topics (Appendix 2). There was no target variable set up in this process (Appendix 1.b). Later we manually feed these topics into the user topic in text topic node and generate models using them as input (Appendix 1.c).

A model comparison node has been added at the end of the model building process.

Generalization

From the comparison of models (Appendix 4 and 6) derived from tagging (with human efforts) (models' description not ending with "m") and automatic generating topics (models' description ending with "m"), we can conclude that using SAS to generate text

topics without having a clear target variable can still help company to get insight of the operation. Models built from unsupervised training have similar, if not the same, performance comparing to the supervised trained models.

By utilizing human tagging, we can create a model with high precision in terms of identification of potential issues. As most of generated models had similar performances, we choose to select the regression model for the simplicity (appendix 5 and 7).

Suggestions for Future Studies

From this project we can see that utilizing SAS EM text mining tools could help the financial institutions increase internal audit efficiency, by identifying potential issues/ risks based on the customer complaints. We intended to perform sentiment analysis with Natural Language Process (NLP); but we found out that such function is only provided in SAS sentiment analysis studio, which is not available to us. Therefore, this project is finished with Bag of Words (BoW) analysis. For future studies, we recommend other groups to conduct sentiment analysis using SAS sentiment analysis studio.

Conclusion

Using human tagging procedure does help us to get a better performance for model building. We also noticed that by using SAS EM text mining tools alone can generate a fairly good list of text topics, which can be used as inputs to predict potential issues.

We recommend companies to use SAS EM text mining to automatically generate text topics, which could be used to predict the potential issues/ risks within the company. This could help firms to utilize the text data (consumer complaints in this case), as well as to increase efficiency at the internal audit department especially at stage of risk identification, so that firms can develop prompt strategies to address the risks. In business practice for financial institutions, the internal audit department can create periodical review reports on customer complaints to highlight business risks and potential problems within the organizations.

Reference:

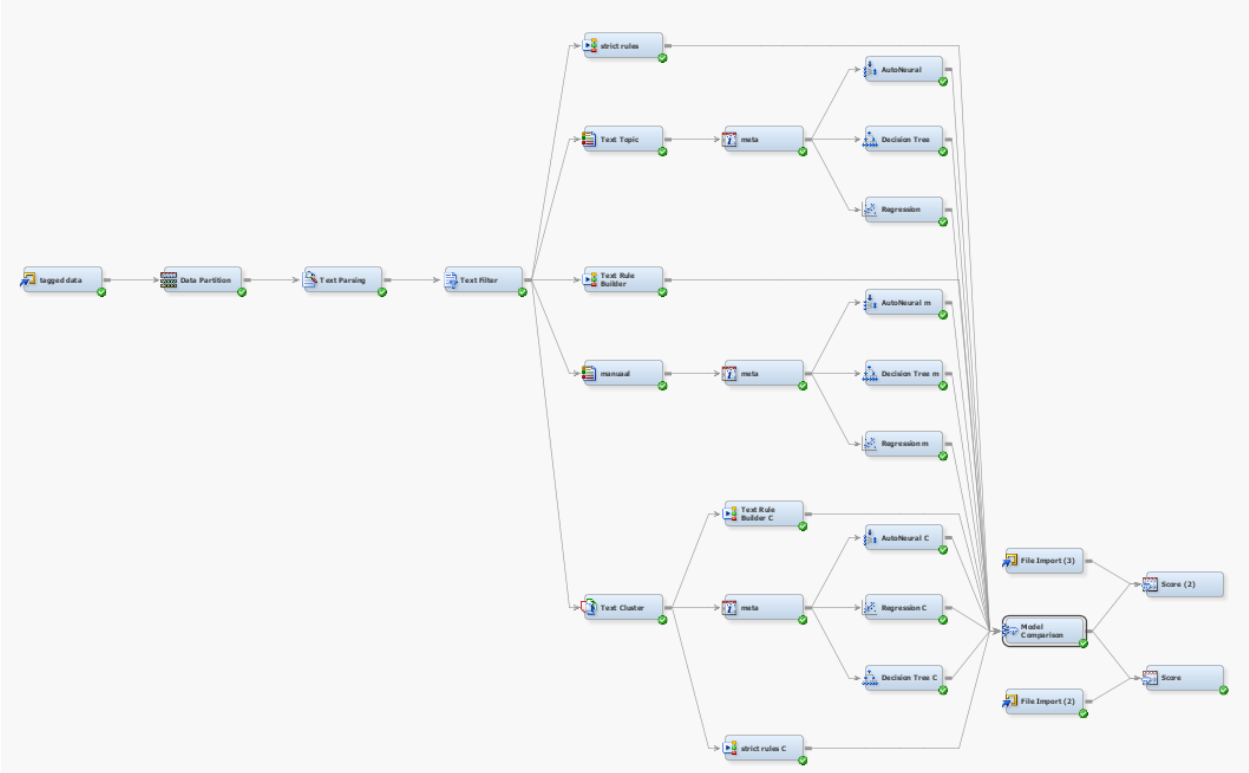
Chakraborty, G., Pagolu, M., & Garla, S. (2013). Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS. SAS Institute.

Consumer Financial Protection Bureau (CFPB). Press Release. "Consumer Financial Protection Bureau Fines Wells Fargo $100 Million for Widespread Illegal Practice of Secretly Opening Unauthorized Accounts", September 8, 2016. Retrieved from: http://www.consumerfinance.gov/about-us/newsroom/consumer-financial-protection-bureau-fines-wells-fargo-100-million-widespread-illegal-practice-secretly-opening-unauthorized-accounts/

CNBC News. "Wells Fargo case is a loud, serious warning to banks, CFPB says". September 12, 2016. Retrieved from: http://www.cnbc.com/2016/09/12/wells-fargo-case-is-a-loud-serious-warning-to-banks-cfpb-says.html

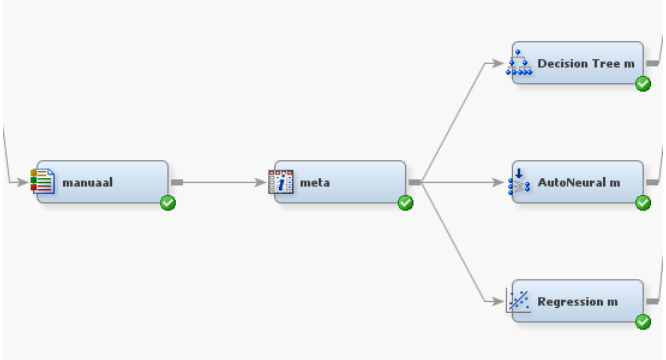
Appendix 1. SAS EM Diagram

Appendix 1.a.



Appendix 1.b.

Appendix 1.c.



Appendix 2. Results from Text Topic Node

Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Single	1	0.001	0.001	+card	1	1257
Single	2	0.001	0.001	+check	1	1141
Single	3	0.001	0.001	+payment	1	868
Single	4	0.001	0.001	+transaction	1	924
Single	5	0.001	0.001	+overdraft	1	782
Single	6	0.001	0.001	+charge	1	750
Single	7	0.001	0.001	+number	1	786
Single	8	0.001	0.001	+credit	1	1073
Single	9	0.001	0.001	+deposit	1	867
Single	10	0.001	0.001	+business	1	771
Single	11	0.001	0.001	+fund	1	1046
Single	12	0.001	0.001	+debit	1	740
Single	13	0.001	0.001	+balance	1	832
Single	14	0.001	0.001	+close	1	1008
Single	15	0.001	0.001	+company	1	495
Single	16	0.001	0.001	+information	1	869
Single	17	0.001	0.001	+letter	1	631
Single	18	0.001	0.001	+branch	1	799
Single	19	0.001	0.001	+bank	1	618
Single	20	0.001	0.001	+deposit	1	818
Single	21	0.001	0.001	+statement	1	521
Single	22	0.001	0.001	+complaint	1	695
Single	23	0.001	0.001	+fraud	1	541
Single	24	0.001	0.001	+claim	1	342
Single	25	0.001	0.001	+bill	1	537

Appendix 3. Tagging Protocol

1. Complaints on Fraud

Description: Customer complains that the bank has conducted fraudulent/ deceptive activities to his/her bank account.

Example: "...I have attached a bank statement that shows just how egregious the fraud.... Over {$800000.00} of unauthorized ACH transfers occurred..."

2. Complaints on Misleading information or Policy

Description: Customer mentions receiving misleading information from bank representatives, bank newsletters, etc.); customer complaints about confusing bank policy.

Example: "I opened up CapitalOne 360 Savings account after seeing advertisement on-line. Bank promised {$500.00} bonus after keeping XXXX in the account for 90 days. ... Now, Bank is claiming that I am not eligible for bonus ..."

3. Complaints on Unauthorized Transactions

Description: Customer claims that he/ she did not authorize or have knowledge about some transactions associated with his/ her bank account. Example could be unknown accounts creation/ closure, unauthorized money transfer or card use.

Example: *"BOA refuses to refund my account for an unauthorized transfer of {$1300.00} that occurred on … despite BOA having actual notice …."*

4. Complaints on Penalty Fees due to insufficient funds and late payments

Description: Customer mentions he/she has been charged extra fees involuntary, due to overdraft (withdrawal of money from account with insufficient balance), or late payments of credit card.

Example: *"… On the next billing cycle of my credit card, the statement says I never paid and it charged me late fees and interest. ....*

Appendix 4. Models Created on Predicting Unauthorized Transaction

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Selection Criterion: Valid: Misclassification Rate	Target Variable
Y	AutoNeural	AutoNeural	AutoNeural C	0.190301	Unauthorized_Judge
	Reg2	Reg2	Regression	0.197649	Unauthorized_Judge
	Reg	Reg	Regression C	0.199853	Unauthorized_Judge
	Tree2	Tree2	Decision Tree	0.204996	Unauthorized_Judge
	Tree	Tree	Decision Tree C	0.206466	Unauthorized_Judge
	TextRule	TextRule	Text Rule Builder	0.22263	Unauthorized_Judge
	TextRule3	TextRule3	Text Rule Builder C	0.22263	Unauthorized_Judge
	Tree3	Tree3	Decision Tree m	0.229978	Unauthorized_Judge
	TextRule2	TextRule2	strict rules	0.233652	Unauthorized_Judge
	TextRule4	TextRule4	strict rules C	0.233652	Unauthorized_Judge
	Reg3	Reg3	Regression m	0.237325	Unauthorized_Judge
	AutoNeural2	AutoNeural2	AutoNeural	0.237325	Unauthorized_Judge
	AutoNeural3	AutoNeural3	AutoNeural m	0.237325	Unauthorized_Judge

Appendix 5. Logistic Regression on Predicting Unauthorized Transaction

Analysis of Maximum Likelihood Estimates

Parameter	Unauthorized_ Judge	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-1.7515	0.0601	848.65	<.0001
TextTopic_raw1	1	1	3.4051	0.3703	84.55	<.0001
TextTopic_raw12	1	1	2.5905	0.5431	22.75	<.0001
TextTopic_raw13	1	1	1.1937	0.5692	4.40	0.0360
TextTopic_raw16	1	1	1.0968	0.5427	4.08	0.0433
TextTopic_raw19	1	1	2.0665	0.6726	9.44	0.0021
TextTopic_raw2	1	1	1.2879	0.2372	29.49	<.0001
TextTopic_raw22	1	1	2.2896	0.6368	12.93	0.0003
TextTopic_raw24	1	1	2.7271	0.7520	13.15	0.0003
TextTopic_raw3	1	1	2.5776	0.3393	57.72	<.0001
TextTopic_raw4	1	1	1.8001	0.3878	21.55	<.0001
TextTopic_raw6	1	1	0.9650	0.4414	4.78	0.0288
TextTopic_raw7	1	1	1.9527	0.5093	14.70	0.0001
TextTopic_raw9	1	1	1.8481	0.5372	11.83	0.0006

Topic ID	Topic
1	+unauthorized
2	+transaction
3	+authorize
4	+withdraw
5	+fraud
6	+withdrawal
7	+authorization
8	+charge
9	+police
10	+card
11	+claim
12	+permission
13	+dispute
14	+fraudulent
15	+unauthorized transaction
16	+investigation
17	+police report
18	+company
19	pin
20	+report
21	theft
22	+consent
23	+unauthorized charge
24	+fraudulent charge
25	+file

Appendix 6. Models created on predicting Hidden Fee

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Misclassification Rate
Y	AutoNeural	AutoNeural	AutoNeural C	Fees_Judge	0.119677
	Reg	Reg	Regression C	Fees_Judge	0.128488
	TextRule	TextRule	Text Rule Builder	Fees_Judge	0.132159
	TextRule3	TextRule3	Text Rule Builder C	Fees_Judge	0.132159
	Reg2	Reg2	Regression	Fees_Judge	0.133627
	Tree	Tree	Decision Tree C	Fees_Judge	0.138767
	TextRule2	TextRule2	strict rules	Fees_Judge	0.141703
	TextRule4	TextRule4	strict rules C	Fees_Judge	0.141703
	Tree2	Tree2	Decision Tree	Fees_Judge	0.145374
	Tree3	Tree3	Decision Tree m	Fees_Judge	0.157856
	Reg3	Reg3	Regression m	Fees_Judge	0.159325
	AutoNeural3	AutoNeural3	AutoNeural m	Fees_Judge	0.265051
	AutoNeural2	AutoNeural2	AutoNeural	Fees_Judge	0.284875

Appendix 7. Logistic Regression on Predicting Hidden Fee

Analysis of Maximum Likelihood Estimates

Parameter	Fees_ Judge	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-3.3393	0.1318	642.12	<.0001
TextTopic_raw1	1	1	3.3719	0.2700	156.00	<.0001
TextTopic_raw12	1	1	2.0404	0.8743	5.45	0.0196
TextTopic_raw13	1	1	4.5738	0.9105	25.23	<.0001
TextTopic_raw14	1	1	5.6609	1.2430	20.74	<.0001
TextTopic_raw2	1	1	4.5287	0.3954	131.18	<.0001
TextTopic_raw25	1	1	4.9919	0.9125	29.92	<.0001
TextTopic_raw3	1	1	2.5490	0.3319	58.98	<.0001
TextTopic_raw4	1	1	2.2057	0.7206	9.37	0.0022
TextTopic_raw5	1	1	2.2745	0.3963	32.94	<.0001
TextTopic_raw6	1	1	2.0562	0.4755	18.70	<.0001
TextTopic_raw7	1	1	2.8818	0.6691	18.55	<.0001
TextTopic_raw8	1	1	3.0390	0.6024	25.45	<.0001

Topic ID	Topic
1	+fee
2	+overdraft
3	+charge
4	+overdraft fee
5	+charge
6	+balance
7	+cover
8	+overdraw
9	negative
10	+protection
11	+transaction
12	+cause
13	+insufficient
14	draft
15	+negative
16	+overdraft protection
17	+item
18	+negative balance
19	+post
20	+draft fee
21	+xxxx overdraft fee
22	+pay
23	pending
24	+refund
25	+late fee