Paper 3917-2019

# From Words to Actions: Using Text Analytics to Drive Business Decisions

Reid Baughman, Zencos; Chris St. Jeor, Zencos

## ABSTRACT

Companies from a variety of industries collect free-form text data that can be used to identify new patterns and relationships. Because unstructured text data does not fit the standard row and column format, it can be more difficult to analyze and utilize. This growing array of data has the potential of yielding new insights for companies seeking to better understand customers and gain an edge relative to competitors. By transforming free-form text data into a structure that can be analyzed and visualized, analysts can use supervised and unsupervised data mining techniques to shed new light on old business problems or develop fresh insights. Call center transcripts, medical records, practitioner's notes, survey responses, or any free-form response fields can all hold valuable insights for any organization. This paper will outline an example of how to manage and transform free-form text data, apply advanced analytical methods to extract useful patterns, and develop actionable insights.

## INTRODUCTION

In the wake of the 2007 financial crisis, the Consumer Protection Financial Bureau (CFPB) was created to "promote fairness and transparency for mortgages, credit cards, and other consumer financial products and services". As part of that mission, they established a system for consumers to log complaints regarding financial products and made the database of complaints available to the public. There are two columns of data that are of interest for this analysis: "Consumer complaint narrative" and "Company response to customer". The first column provides the opportunity to use unsupervised modeling to listen to the voice of the customer and understand common reasons for dissatisfaction. Combining those results with the second variable – Company response to customer – allows you to see which compliant topics are potentially avoidable by employing a supervised approach with relief as the target. The assumption here is that if the company offered relief, then they accept wrongdoing and that the complaint should have been avoidable. To demonstrate these two types of modeling, we'll use SAS® Visual Text Analytics due to its robust text mining capabilities.

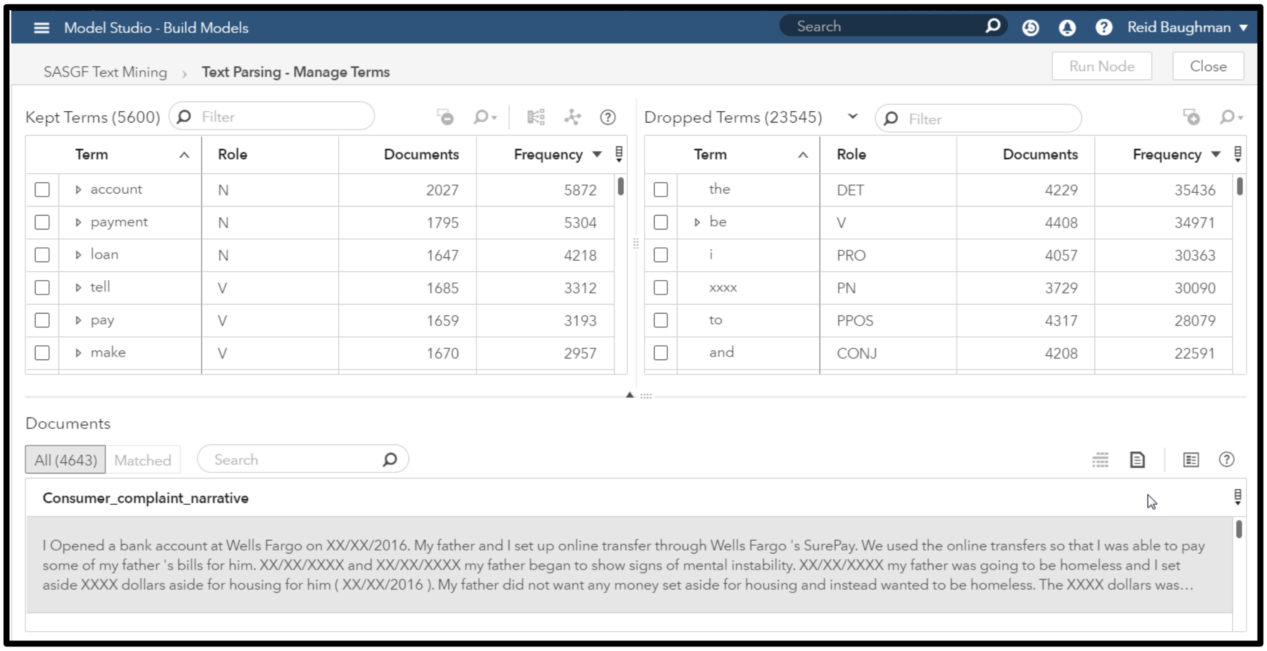
For the analysis, I chose to model data from Wells Fargo since they ran afoul of the CFPB and were required to pay a multi-million dollar fine in large part due to illegal business practices uncovered from the complaint database. In the period measured (March 2015 – May 2017) there were 4,643 complaints that included a consumer narrative and of those 756 resulted in the bank providing some sort of relief. With SAS® Visual Text Analytics, it is easy to get from raw data to insights within minutes.

## UNSUPERVISED MODELING: TOPICS

### DATA PREP MADE EASY

The first step in any text mining project is to parse and filter all the free form text data so that we can cut through the noise and try to preserve the most important words. In SAS® Visual Text Analytics, this is accomplished simply with the Text Parsing node. After importing data, the dataset can be fed into the default text mining pipeline (collection of nodes) to run the data through the various text mining routines. The first of these is the

Parse node which parses the complaints text into individual words, stems (combines all variations of the same word), drops unimportant filler words and allows the selection of individual "Kept" or "Dropped" terms. By checking the boxes below in the "Kept Terms" or "Dropped Terms" panes, individual words can be included or excluded in the topic creation.


Model Studio - Build Models
Search
Reid Baughman
SASGF Text Mining › Text Parsing - Manage Terms
Run Node
Close

Kept Terms (5600)

| | Term | Role | Documents | Frequency |
|---|---|---|---|---|
| □ | account | N | 2027 | 5872 |
| □ | payment | N | 1795 | 5304 |
| □ | loan | N | 1647 | 4218 |
| □ | tell | V | 1685 | 3312 |
| □ | pay | V | 1659 | 3193 |
| □ | make | V | 1670 | 2957 |

Dropped Terms (23545)

| | Term | Role | Documents | Frequency |
|---|---|---|---|---|
| □ | the | DET | 4229 | 35436 |
| □ | be | V | 4408 | 34971 |
| □ | i | PRO | 4057 | 30363 |
| □ | xxxx | PN | 3729 | 30090 |
| □ | to | PPOS | 4317 | 28079 |
| □ | and | CONJ | 4208 | 22591 |

Documents

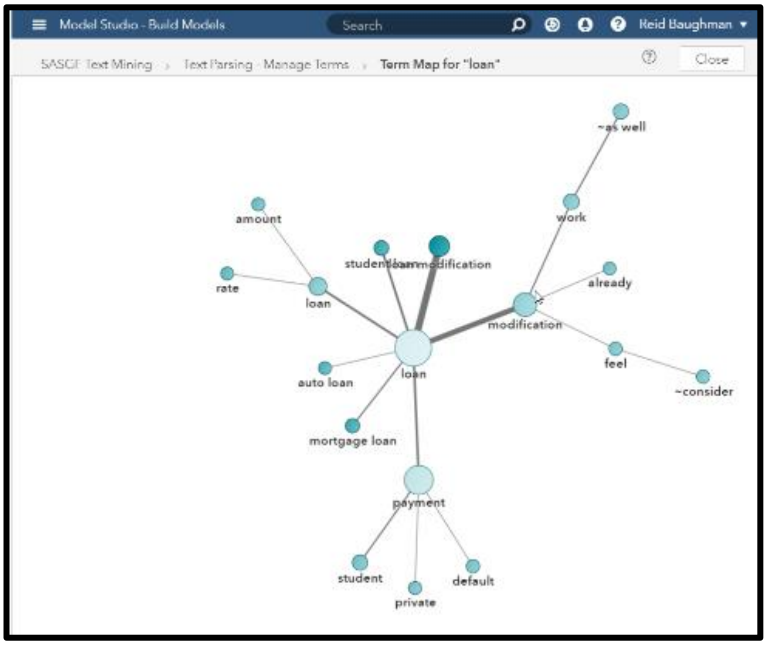
All (4643) Matched Search

Consumer_complaint_narrative

I Opened a bank account at Wells Fargo on XX/XX/2016. My father and I set up online transfer through Wells Fargo 's SurePay. We used the online transfers so that I was able to pay some of my father 's bills for him. XX/XX/XXXX and XX/XX/XXXX my father began to show signs of mental instability. XX/XX/XXXX my father was going to be homeless and I set aside XXXX dollars aside for housing for him ( XX/XX/2016 ). My father did not want any money set aside for housing and instead wanted to be homeless. The XXXX dollars was…

**Figure 1. Parsing Node**

## Term Map



An additional useful feature of the Parsing node is the term mapping option. After checking the box of one the terms in the "Kept Terms" pane, a term map can be created for any term. This map shows the relationship between that particular term and any others it is most related to. In Figure 2. Term Map, the term "loan" has been selected and some of the terms it is most closely related to include "modification", "payment", "mortgage loan" and "auto loan". Thus, the Parsing node can be a useful tool for beginning to explore the connections between terms. To dive more deeply into this aspect of text mining, the "Topics" node is next.

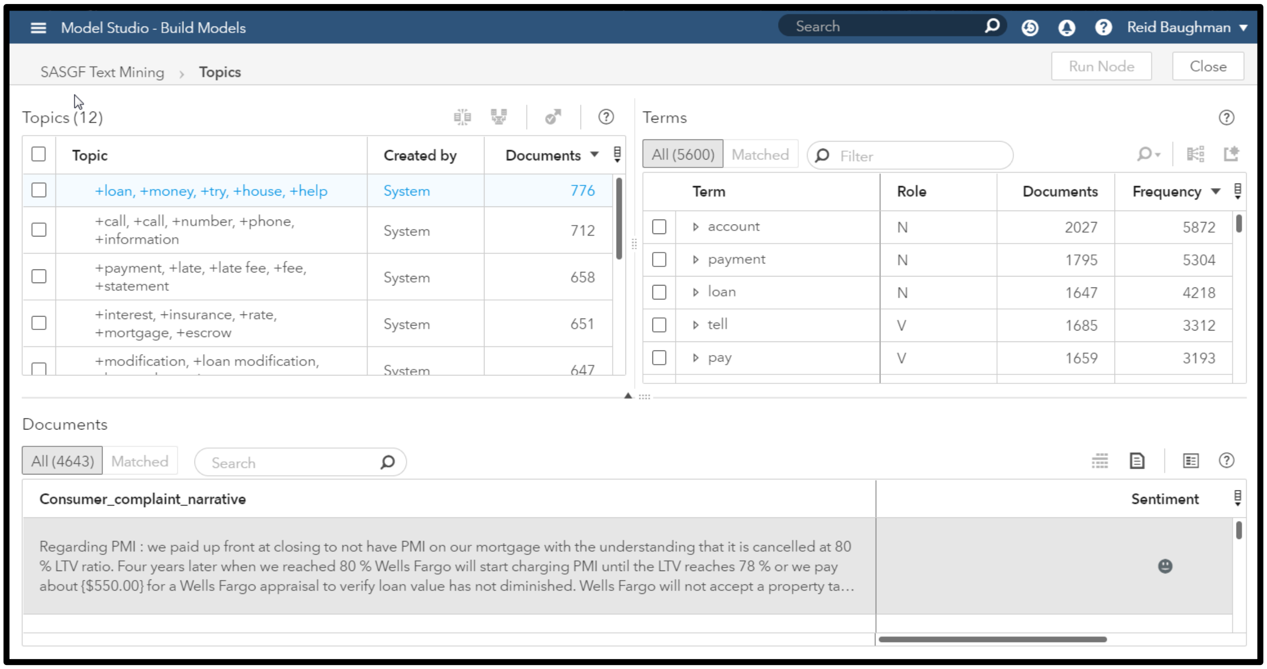
**Figure 2. Term Map**

## TOPIC CREATION

Once the data has been parsed, it's ready for analysis. The Topics node performs a form of unsupervised learning which derives topics that naturally arise from the text data. This type of text analytics is considered unsupervised because it lacks a target it's trying to predict. Terms cluster together if they appear together frequently. For example, if several complaints contained the words "late", "fee" and "unfair" then those words would likely comprise a topic. The Topics node does this topic clustering for us automatically and generates useful visuals.

### Topics Node

Figure 3. Topics node below shows the topics generated (left pane) and terms contained (right pane) in each topic. In the Topics pane, each row represents a topic and the most frequent terms in each topic. Additionally, the "Documents" column shows the number documents that topic is found in. An additional feature in the "Topics" pane is the ability to split or combine topics. If a default topic contains too many terms (or too few), it can be split or combined with others. At the bottom of the window, raw complaint data and sentiment analysis is shown for each topic.



| ☐ | Topic | Created by | Documents |
|---|---|---|---|
| ☐ | +loan, +money, +try, +house, +help | System | 776 |
| ☐ | +call, +call, +number, +phone, +information | System | 712 |
| ☐ | +payment, +late, +late fee, +fee, +statement | System | 658 |
| ☐ | +interest, +insurance, +rate, +mortgage, +escrow | System | 651 |
| ☐ | +modification, +loan modification, | System | 647 |

| ☐ | Term | Role | Documents | Frequency |
|---|---|---|---|---|
| ☐ | account | N | 2027 | 5872 |
| ☐ | payment | N | 1795 | 5304 |
| ☐ | loan | N | 1647 | 4218 |
| ☐ | tell | V | 1685 | 3312 |
| ☐ | pay | V | 1659 | 3193 |

| Consumer_complaint_narrative | Sentiment |
|---|---|
| Regarding PMI : we paid up front at closing to not have PMI on our mortgage with the understanding that it is cancelled at 80 % LTV ratio. Four years later when we reached 80 % Wells Fargo will start charging PMI until the LTV reaches 78 % or we pay about {$550.00} for a Wells Fargo appraisal to verify loan value has not diminished. Wells Fargo will not accept a property ta… | ☹ |

**Figure 3. Topics node**

### Key Business Points

From this unsupervised method, we now have a better understanding as to which complaint topics are most dominant. From Figure 3 we can see that the three most dominant topics are:

1. +loan, +money, +try, +house, +help
2. +call, +call, +number, +phone, +information
3. +payment, +late, +late fee, +fee, +statement

This information could be useful in various parts of the business. For decision makers it could serve as a starting point to discuss areas of focus for improving customer satisfaction.

For customer service reps, it could help focus their training to respond to the most common types of complaints. To get to the next level of insight, we can employ supervised methods to see not just *what* people are complaining about but *how* to potentially avoid some of the complaints in the first place.

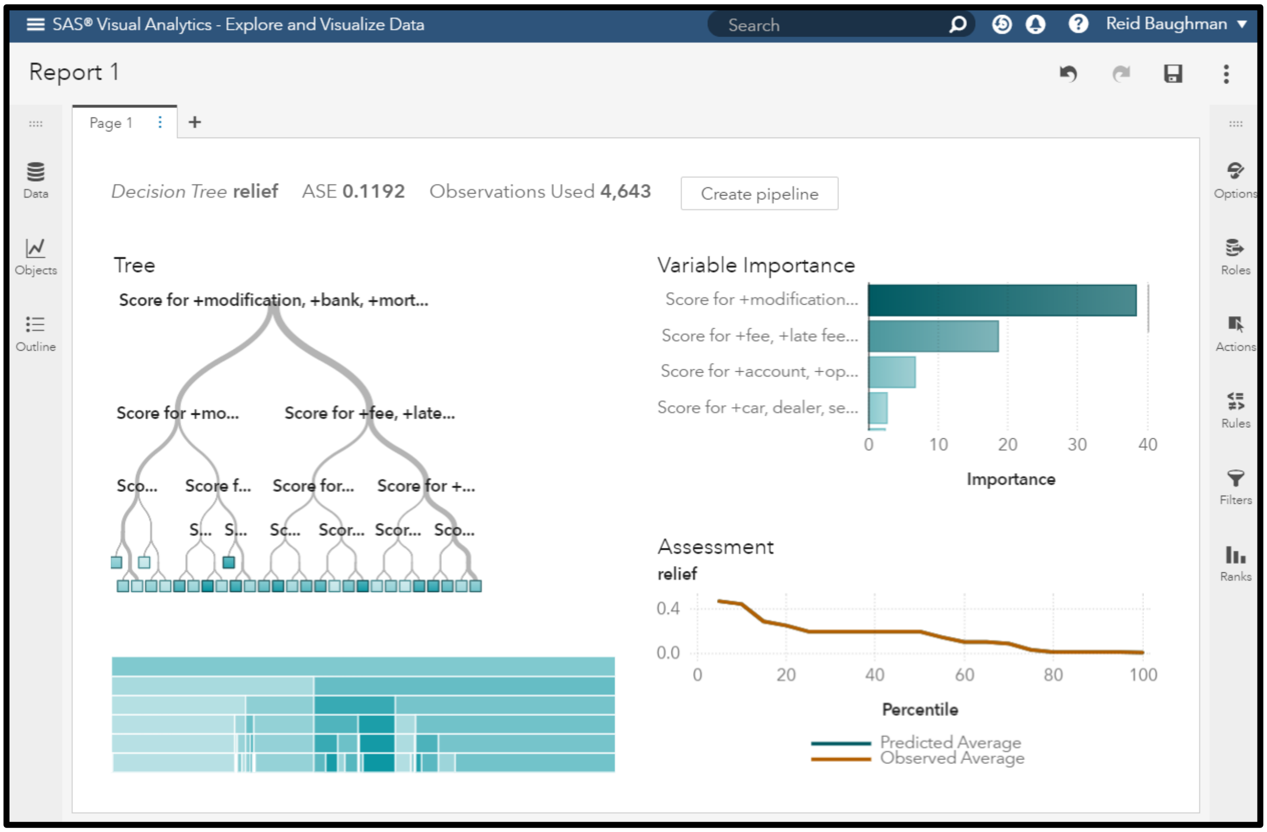
## SUPERVISED MODELING: PREDICTING AVOIDABLE COMPLAINTS

Now that we have a better understanding of some of the dominant complaint topics, let's look now to see which topics are potentially avoidable. Using the "relief" indicator explained earlier, we'll attempt to model which topics are associated with relief to uncover areas where the company can improve and hopefully avoid preventable complaints.

SAS® Visual Text Analytics contains an array of methods for supervised learning. After running the data through the Text Analytics pipeline, the newly created topics can be leveraged as features in training a model to predict which complaints receive relief. The output data from the pipeline can be easily loaded into SAS® Visual Analytics to both model and visualize the data. The two supervised methods we'll discuss here are the decision tree and random forest methods.

### DECISION TREE

The decision tree is a powerful and flexible model that provides a good combination of predictive power and interpretability. By feeding in the topics as input variables and setting the "relief" variable as the target, we can analyze the text to determine which complaint topics received monetary relief. Figure 4. Decision Tree shows the results of training a decision tree on the data.



SAS® Visual Analytics - Explore and Visualize Data

Report 1

*Decision Tree* **relief** ASE **0.1192** Observations Used **4,643**

**Figure 4. Decision Tree**

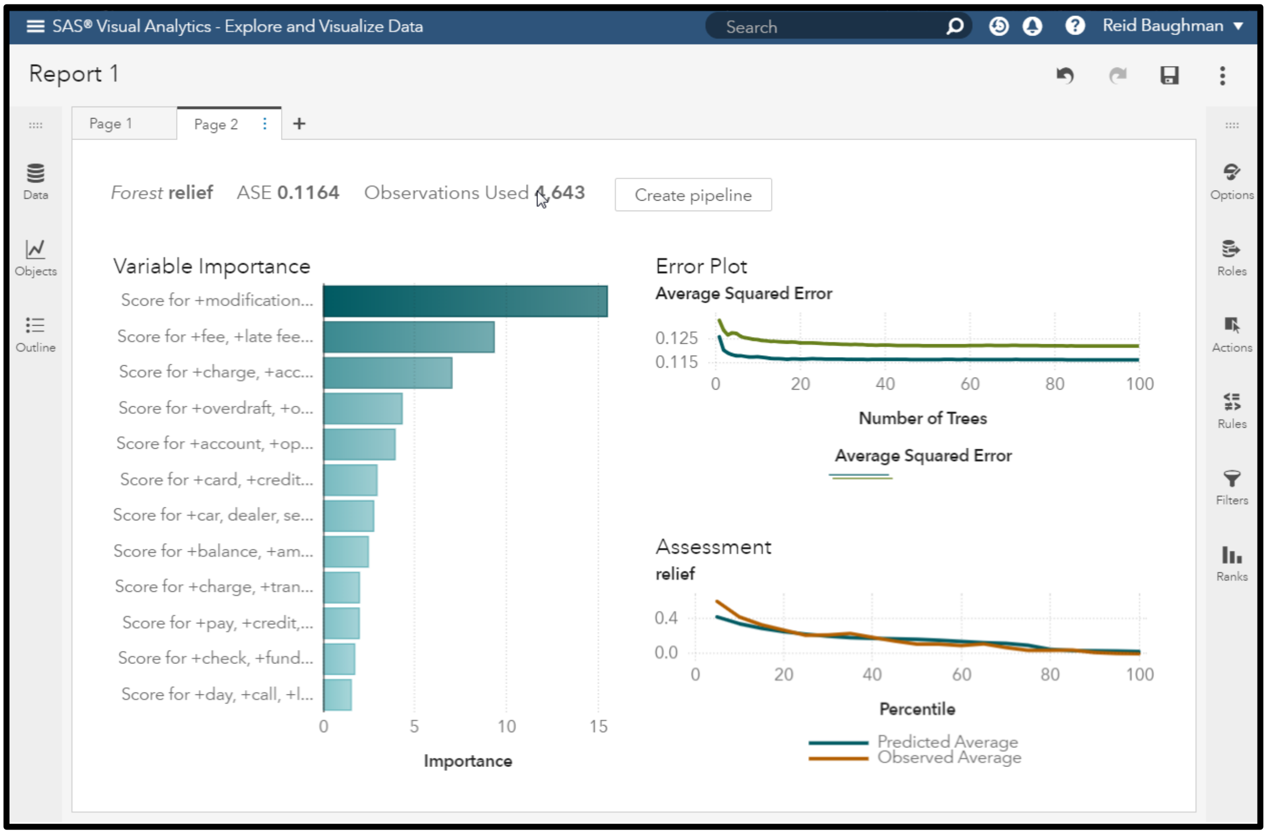
There is a lot going on in Figure 4 so let's focus on two key areas: the "Tree" pane on the left side and the "Variable Importance" pane on the right. The Tree pane shows us our decision tree and how different topics are used as branches to route the complaints in different directions. The goal here is to see if the topics we've derived from the text data can be used to correctly predict whether a complaint will receive relief or not. Some topics do a better job at this than others (i.e. they are more predictive). In the tree diagram, these will appear towards the top of the tree. The thickness of the lines is also important because it shows how many complaints are routed each way at each branch – the thicker the line the more observations are travelling along that branch.

Another way to look at which clusters are most predictive is the "Variable Importance" pane on the right which ranks the topics by predictive power. This is often an easier way to determine which topics are most predictive.

When performing supervised learning tasks, it is often advisable to fit several different models to any dataset to see which fits the data best. Every dataset has its own peculiarities and some models are better suited to certain kinds of data than others. With that in mind, we'll look at using a variant of the decision tree called a random forest.

## RANDOM FOREST

A random forest is like a decision tree in that it is comprised of many, many decision trees (hence the term "forest"). It is especially helpful when you have issues with highly dimensional data, or data with too many variables. Some text analytics use cases can result in a very high number of topics so a random forest can be useful in efficiently modeling highly dimensional data. For random forests, SAS® Visual Analytics provides a similar output to the decision tree. The results are below in Figure 5.



*Forest* relief ASE **0.1164** Observations Used **4,643**

SAS® Visual Analytics - Explore and Visualize Data
Search
Reid Baughman
Report 1
Page 1
Page 2
+
Data
Objects
Outline
Create pipeline
Options
Roles
Actions
Rules
Filters
Ranks
Variable Importance
Score for +modification...
Score for +fee, +late fee...
Score for +charge, +acc...
Score for +overdraft, +o...
Score for +account, +op...
Score for +card, +credit...
Score for +car, dealer, se...
Score for +balance, +am...
Score for +charge, +tran...
Score for +pay, +credit,...
Score for +check, +fund...
Score for +day, +call, +l...
0
5
10
15
Importance
Error Plot
Average Squared Error
0.125
0.115
0
20
40
60
80
100
Number of Trees
Average Squared Error
Assessment
relief
0.4
0.0
0
20
40
60
80
100
Percentile
Predicted Average
Observed Average

**Figure 5. Random Forest**

Since the random forest is a composite of many smaller decision trees, there isn’t a tree chart like we saw in Figure 4. We do however see a “Variable Importance” pane (on the left) as well as two charts on the right showing the Error Plot and the Assessment. The “Error Plot” shows us how the accuracy of the random forest increases as it adds more trees to the forest. You’ll notice that the accuracy increases for a short time and then flatlines. There isn’t any incremental benefit to adding more trees so it stops after 100 iterations. The Assessment chart shows the predicted vs observed average from the top decile (10th percentile) all the way to the 100th percentile.

### Key Business Points

The most important takeaway from these two graphs is the variable importance. From a business perspective, we now understand which complaint topics are associated with some form of relief being provided to the customer.  This suggests that some of these may have been avoidable. The key drivers here appear to be:

1. +modification, +bank, +mortgage, +loan modification, +home
2. +fee, +late fee, +late, +bank, +waive
3. +charge, +account, +service, +refund, +call

The interpretation here is that customers who complained about the above topics were more likely to receive relief than those who complained about other topics. The thought here is that if work can be done to understand why these are more likely to receive relief than others, maybe insights there could lead to adjustments in how the mortgage department does their business or how customer service reps handle complaints about late fees.

## CONCLUSION

The text mining techniques demonstrated here uncover otherwise unreachable insights trapped in potentially thousands of customer responses. Sifting through the deluge of text, these techniques allow us to quickly filter down to the most meaningful words that describe the customer’s experience. Perhaps more importantly, these techniques generalize to practically any body of text from tweets, to transcripts, to even books to name a few. Just think, what insights could be hiding in your stores of unstructured text?

## REFERENCES

Chakraborty, Goutam, et al. 2013. *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2018. SAS® Visual Text Analytics 8.3: User’s Guide. Cary, NC: SAS Institute Inc.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Reid Baughman
Zencos Consulting LLC
rbaughman@zencos.com

Chris St. Jeor
Zencos Consulting LLC
cst.jeor@zencos.com