University of Connecticut

MS in Business Analytics and Project Management

OPIM-5671- Data Mining and Business Intelligence

Professor Cruz

Text Mining Project

Consumer Complaints Dataset

Group 4:
Abhinav Dubey
Jiaxuan Wang
Marina Suberlyak
Shaista Usman

Table of Contents

Problem Statement

Our dataset compiles consumer complaints received by the Bureau of Consumer Financial Protection about financial products and services for the years 2021-2016. Complaints are processed by the institutions and each complaint has to identify information about the issue, the resolution or latest status, and whether the consumer filed a dispute to the response provided by the institution. The intention is to predict whether the resolution will be disputed by the consumer from the variables available; as some variables are text, text mining techniques will be applied to derive input values for modeling. We presume the business application here is twofold: inform which products, services, and resolution responses offered by the institution require a review to improve clarity, usability, and value to the consumer; and understand when a consumer is likely to file a dispute to guide the response and training for the institution's employees processing the claims.

Methodology

Sample

The data contains a variety of fields, including company name, issue within the complaint, the products and sub-products for the complaint, company response and timing of the response, and identifying data such as dates, state, and zip code. There are 670,598 rows and 15 columns in the dataset, providing sufficient predictor information to build our models. All columns are listed below:

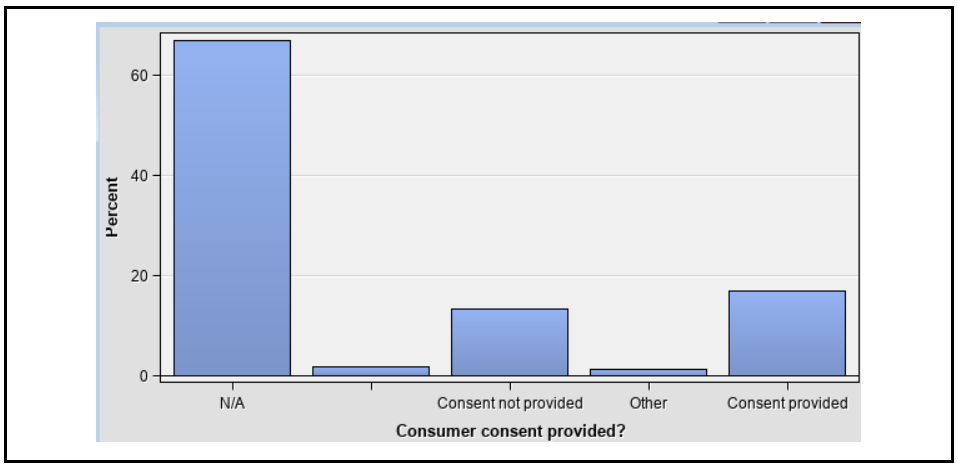
S.No.	NAME	DESCRIPTION	Data type
1	Company	Company names like Wells Fargo & Company, Citibank, Bank of America, etc	Nominal
2	Company_response_to_consumer	Company response to the customer	Nominal
3	Complaint_ID	Complaint ID	ID
4	Consumer_consent_provided_	Consumer consent provided	Nominal
5	Consumer_disputed_	If a consumer disputed or not	Binary
6	Date_received	The date on which the consumer files the complaint	Interval
7	Date_sent_to_company	The date on which complaint is sent to the company	Interval
8	Issue	Issues for which consumer has filed a complaint	Text
9	Product	Product for which complaint is filed	Nominal
10	State	Geographic identifier to know in which state complaint is filed like CT, VA, MA, etc	Nominal
11	Sub_product	Specific sub-product of the product to which the consumer files the complaint.	Text
12	Submitted_via	The mode used by the consumer to file the complaint.	Nominal

13	Tags	Gives extra information about the consumer	Nominal
14	Timely_response_	Tell if the response to the complaint was on time or not	Binary
15	ZIP_code	Geographic identifier to know where the complaint was registered.	Numeric

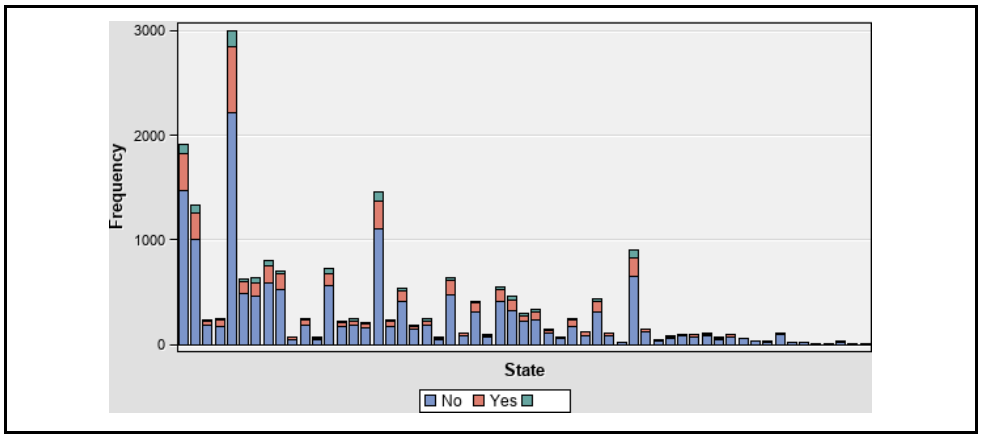
Explore

During the review of the data, the following decisions were made to reduce the dataset to its most useful form.

1. Company**:** This variable indicates a unique name of the company receiving the complaint and in itself holds no information for prediction. However, the company name may be useful for tabulation once we complete the analysis.
2. Consumer_consent_provided**:** This variable has "n/a" for over 60% of the data. Without clarity on why "n/a" might be an acceptable value, we choose to reject the variable.



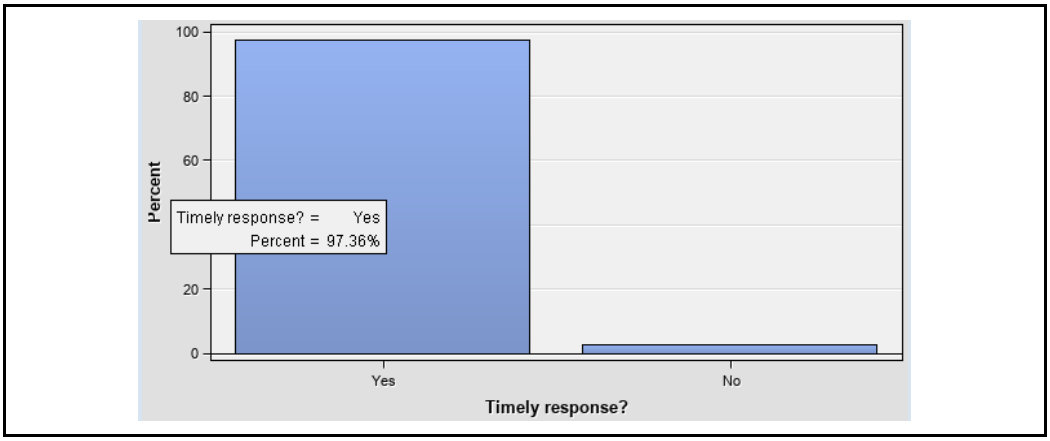
3. Date_received and Date_sent_to_company**:** These variables contain the time id of the complaint. We hypothesize that time separation between the consumer submitting the complaint and company receiving the complaint has no relationship to the type of complaint and subsequent resolution steps and is therefore not relevant in predicting consumer likelihood to dispute the resolution.
4. State**:** This variable is a geographic identifier and on visual inspection, the consumer dispute variable is equally represented across all states. Due to the complexity of incorporating geo data into modeling, we choose to reject this analysis.



5. Tags**:** This variable is dominated by blank values and is therefore not valuable as a predictor.

Variable Name	Percent Missing ▼
Tags	85.94

6. Timely_response: This variable is dominated by a single value, “Yes”, and is therefore not useful in this analysis.

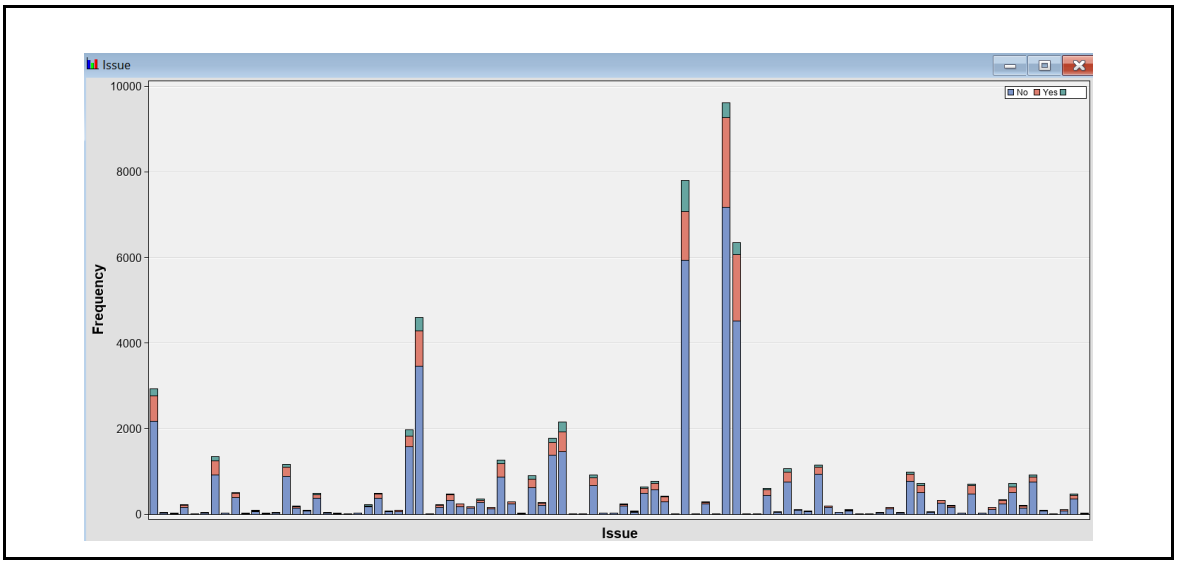


7. Zipcode**:** This variable is another geographic identifier and requires additional context in order to become useful. We hypothesize there is not a meaningful relationship between consumer likelihood to dispute and geography and therefore reject zipcode.

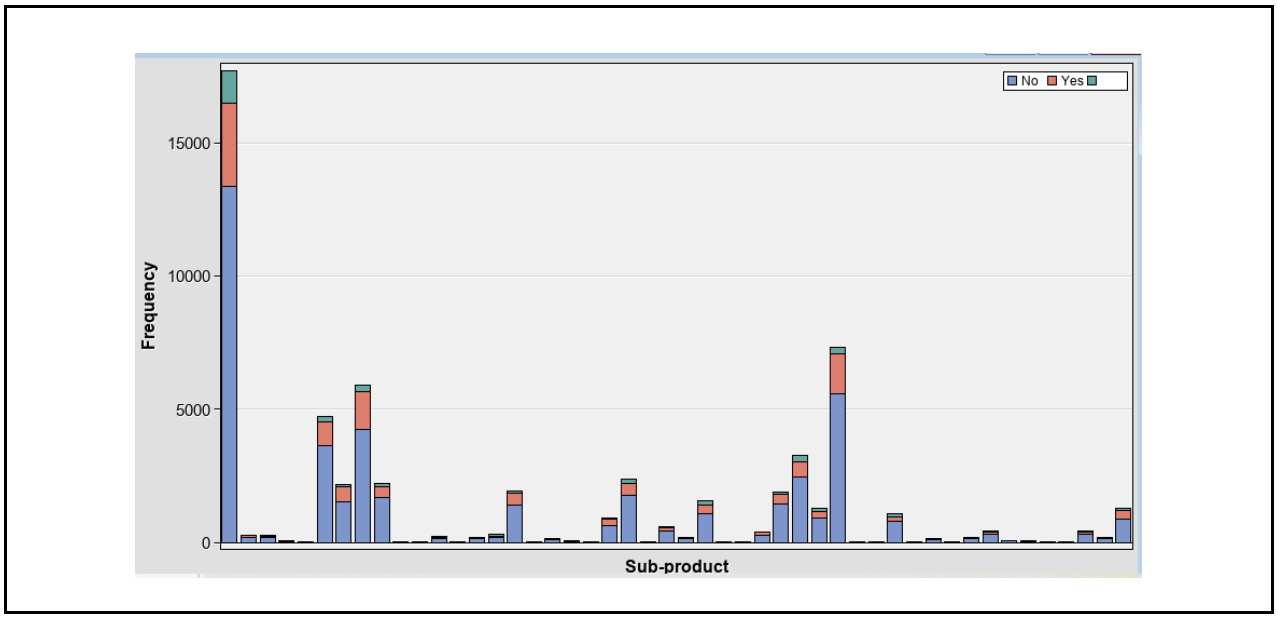
Please find the screenshot below of the metadata after rejecting the above variables.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Company	Rejected	Nominal	No		No	.	.
Company_response_to_consumer	Input	Nominal	No		No	.	.
Complaint_ID	ID	Nominal	No		No	.	.
Consumer_consent_provided_	Rejected	Nominal	No		No	.	.
Consumer_disputed_	Target	Binary	No		No	.	.
Date_received	Rejected	Interval	No		No	.	.
Date_sent_to_company	Rejected	Interval	No		No	.	.
Issue	Text	Nominal	No		No	.	.
Product	Label	Nominal	No		No	.	.
State	Rejected	Nominal	No		No	.	.
Sub_product	Label	Nominal	No		No	.	.
Submitted_via	Input	Nominal	No		No	.	.
Tags	Rejected	Nominal	No		No	.	.
Timely_response_	Rejected	Binary	No		No	.	.
ZIP_code	Rejected	Nominal	No		No	.	.

We now take a look at two text variables we identified - issue and sub-product - to understand how diverse and complex the data is within. In all, variable issue has 92 levels and a handful of categories contain a lot of the frequency; at the same time, target variables appear distributed across all levels and do not immediately suggest a pattern. Given that out of 15K observations, we discover 92 unique categories, we hypothesize that the issue contains some level of discrete answers, which were predetermined by the institution and made available for the consumer to select during complaint filing.



Variable sub-product has 48 levels and also has a handful of categories with a lot of frequency.



Modify

Filter Node

We identify that the sub-product variable has ~30% values missing and we cannot perform imputation without industry knowledge or additional data. Instead, with the dataset of 670K

observation, we can afford to remove the missing rows without compromising the overall information. We apply a Filter node to achieve this.

Variable Name	Percent Missing ▼	Number of Levels
Tags	85.944	
Sub_product	29.58547	

Sample and Partition Node

The reduced dataset includes 472K observations. To make it more manageable in modeling, we extract a random stratified sample of 15,000 observations, using the below setup:

Sample method: stratify
Type: number of observations
Observations: 15,000
Stratified criterion: proportional to preserve the proportions from original data

Sampling Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Filter_TRAIN	472396
SAMPLE	EMWS1.Smpl_DATA	15001

Lastly, we partition the sample into train/validate/test to allow model assessment and comparison; we will allocate 60% of the data to training, and split remainder into validation and test where we will compare model performance to confirm no overfitting and no significant deterioration in key indicators. Our setup and confirmation are below:

Training: 60%
Validation: 20%
Test: 20%

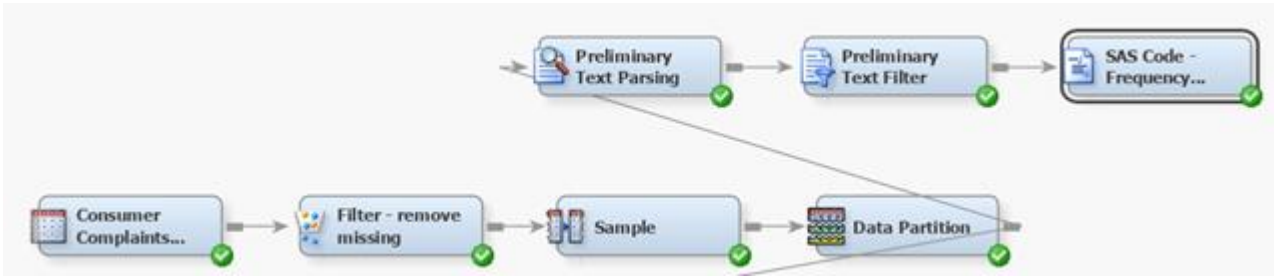
Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Smpl_DATA	15001
TRAIN	EMWS1.Part_TRAIN	8998
VALIDATE	EMWS1.Part_VALIDATE	2999
TEST	EMWS1.Part_TEST	3004

SAS Code Node

Without a start or stop list provided directly, we chose to perform a preliminary analysis where we combined SAS Code node with Text Parsing and Text Filter nodes to generate start and stop lists against the source variable issue. We relied on frequency filtering to derive the lists to allow us to minimize noise in the terms frequently repeated but without material information.

Creating a Start/Stop List



SAS Code Node – Code Editor

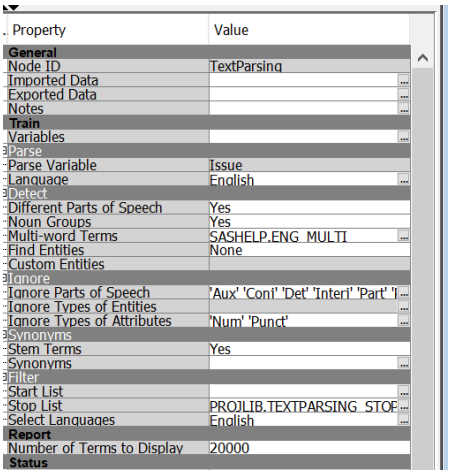
```
Training Code
/*----  SCN_CreateStartStopList_AW.sas  ----------------*/
/*----  Create Start List using frequency filtering  ----*/

%global LastParsing LastFilter TermData FTermData
        StartList MaxDocs MinDocs;

/*!!!!  Edit the following 5 lines  !!!!*/
%let StartList=ProjLib.CCstart;
%let StopList=ProjLib.CCstop;
%let MaxDocs=15001;
%let MinDocs=10;
%let AllowNumbers=N;

%let LastParsing= ;
%let LastFilter= ;

proc print data=&EM_IMPORT_DATA_EMINFO;
run;

proc sql noprint;
  select data into :LastFilter
  from &EM_IMPORT_DATA_EMINFO
  where key="LastTextFilter";
  select data into :LastParsing
  from &EM_IMPORT_DATA_EMINFO
  where key="LastTextParsing";
quit;

%put NOTE: Last SAS Text Parsing Node: &LastParsing;
%put NOTE: Last Text Filter Node: &LastFilter;

%let TermData=%sysfunc(strip(&EM_LIB)).%sysfunc(strip(&LastParsing))_terms;
%let FTermData=%sysfunc(strip(&EM_LIB)).%sysfunc(strip(&LastFilter))_terms_data;
```

Text Parsing Node

We have previously identified 2 text variables: issue and sub-product. However, variable issue contains twice as many unique categories (92 vs. 48 levels), and we choose this variable to derive additional information.

Applying Text Parsing node, our setup included selecting detection for parts of speech, removing default synonym list, adding the previously derived stop list and selecting English as the language.

Property Panel



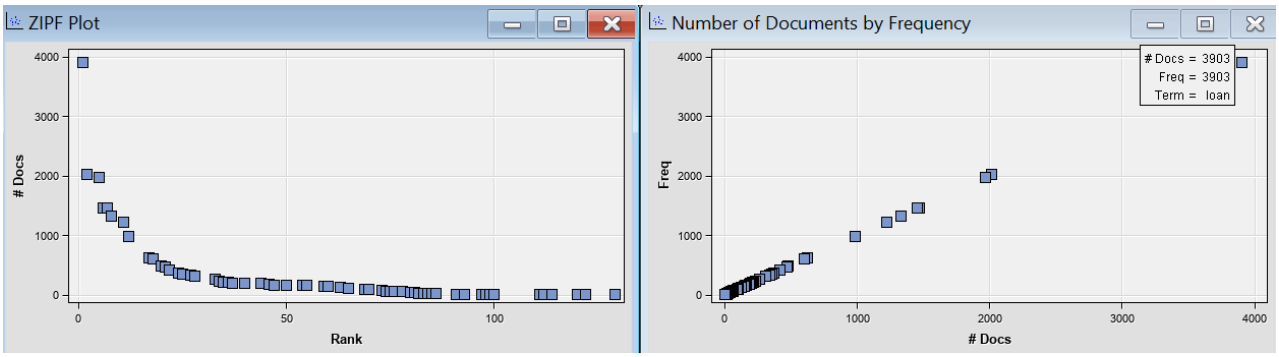
Completing this node, we were able to decompose text into tokens and identify terms, associate each term with parts of speech and perform stemming to equate terms with different tenses.

Text Filter Node

To further improve the quality of terms derived from variable issue, we applied a Text Filter node, which assists in correcting misspellings, establishing frequency and terms weights. Our setup included “Check Spelling” and default settings for frequency and term weighting.

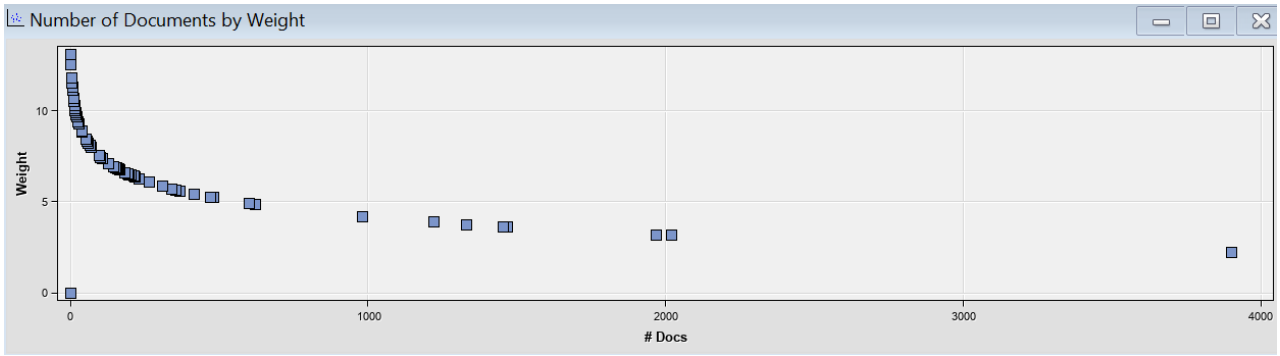
From the initial run, we discover that a single term, “loan”, appears in an extraordinary number of documents. We also observe that term frequency plotted against # of documents is a 45° line, so many terms appear approximately once in many documents.

Zipf Plot and Number of Documents by Frequency



We choose to select Default frequency weighting and Inverse Document Frequency term weighting to give more emphasis to infrequent terms in the document collection and produce variables for modeling. The resulting weight to doc # plot is below:

Number of Documents By Weight

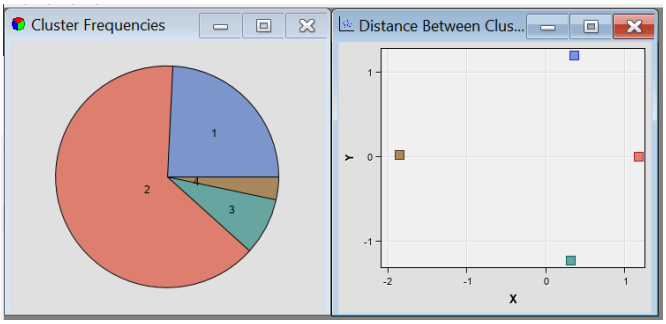


Text Cluster Node

In order to reduce the dimensionality of our dataset and to transform the weighted, term-document frequency matrix generated by the filter node into SVD values, we add a Text Cluster node. After trialing cluster sizes, we found 4 clusters to be optimal for our dataset as they provide adequate distance among them such that each cluster possesses a distinct theme or concept and is mutually exclusive, as evidenced by the distances between each of the four clusters. We do note that cluster size is uneven, with cluster #2 hosting majority of the observations, with terms around account servicing and payment issues. In a dataset around financial product complaints, it makes sense that most of the issues would be associated with consumer's accounts and their grievances against the financial products or services availed by them. We also note the 2nd largest cluster contains terms around debt collection; again, this makes sense as consumers are likely to have disagreements over such financially impactful judgement.

Property Panel, Cluster Frequencies and Distance Between Clusters

Train	
Variables	
Transform	
SVD Resolution	Low
Max SVD Dimensions	100
Cluster	
Exact or Maximum Number	Exact
Number of Clusters	4
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15



Text Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage
1	debt not 'collect debt' +attempt +owe collect cont +disclosure verification lender...	2183	24%
2	loan collection foreclosure modification account +payment 'escrow account' +ser...	5759	64%
3	+problem +be +cause +fund low +'false statement' +statement false representat...	745	8%
4	'mortgage broker' application broker mortgage originator ...	311	3%

Text Topic Node

We further added a text topic node in series with the cluster node to refine our term understanding and increase the number of variables derived from text to aid in modeling later. We used all the default values for the topic node. We settle on fifteen topics and observe that only a few terms define each topic. This again supports our hypothesis that the variable issue contains predetermined text that likely already separates issue topics into discrete categories.

We used these distinct topics as inputs to our models for accurate classifications.

Text Topics

Topic ID ▲	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
1	0.555	0.104	foreclosure,collection,modification,loan,management	4	2019
2	0.484	0.110	+service,escrow,escrow account,+payment,account	9	1330
3	0.435	0.110	+attempt,collect debt,cont,+owe,collect	9	983
4	0.332	0.106	management,opening,+close,account,+service	4	600
5	0.281	0.105	verification,+disclosure,debt,lease,+manage	3	473
6	0.237	0.104	mortgage broker,application,broker,mortgage,originator	5	311
7	0.241	0.099	communication,+tactic,other service,other transaction,other	2	418
8	0.222	0.099	+withdrawal,+deposit,other service,other transaction,other	2	354
9	0.205	0.107	+problem,+cause,+fund,low,+be	7	370
10	0.219	0.108	lease,+manage,loan,+repay,+shop	3	377
11	0.184	0.103	representation,false,+false statement,+statement,other service	4	198
12	0.175	0.104	lender,servicer,+deal,contact,+disclosure	3	213
13	0.170	0.106	improper contact,contact,+share,improper,info	5	162
14	0.175	0.106	+take,illegal action,illegal,action,+threaten	6	233
15	0.153	0.101	+cost,process,settlement,other service,other transaction	3	147

Text Profile Node

Given our observation that text within variable issue is pre-crafted to contain discrete responses, we take a look at the Text Profile node to uncover whether certain terms, and therefore topics, have a relationship with the target variable. As per SAS documentation, "For each level of a target variable, the node outputs a list of terms from the collection that characterize or describe that level."[1]

Profiled Variables

Name	Value	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
Corpus		relation/Nn	vehicle/Nn	sell/Vb	repossess/Vb	excessi...	excessive/Adj	unexpe...	other fee/...
Consumer_dispute...	No	communicat...	tactic/Nn	deposit/Nn	withdrawal/Nn	other/Nn	not/Adv	collect ...	owe/Vb
Consumer_dispute...	Yes	lease/Vb	disclosure/Nn	verification...	mortgage/Nn	broker/Nn	originator/Nn	mortga...	applicatio...

The output shows that target variable consumer_disputed=yes commonly has terms around the financial product itself or the application process; whereas consumer_disputed=no often contains terms around communication, account services and debt collection.

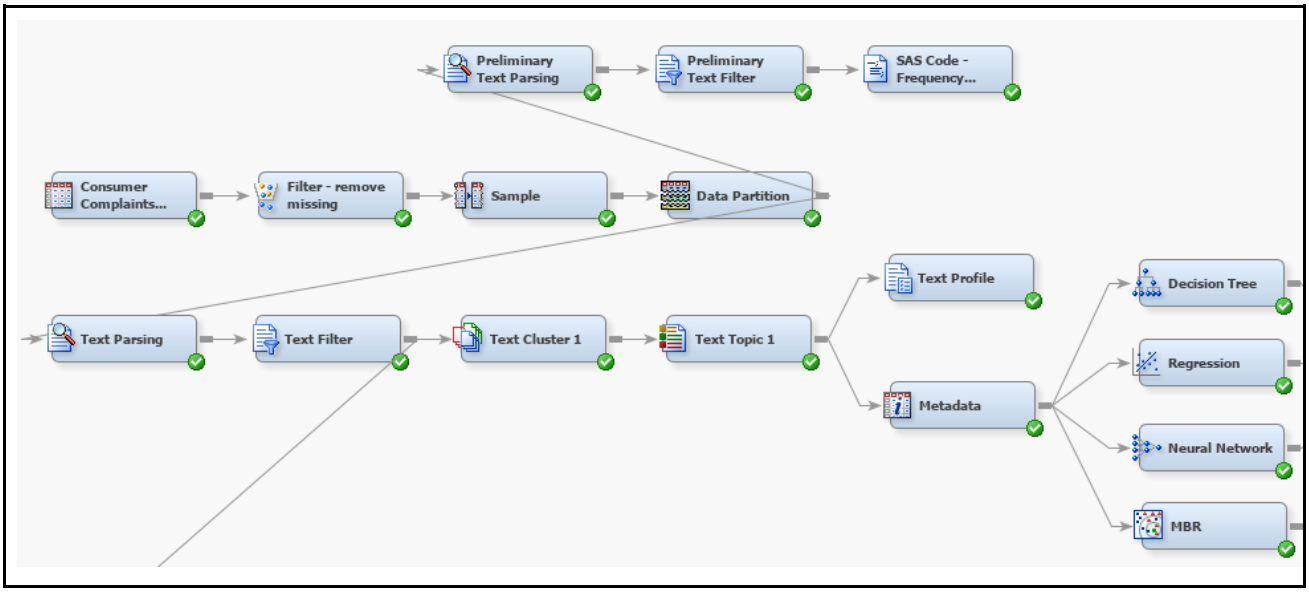
Metadata Node

We apply Metadata node to assign roles to the new variables we derived during text analysis steps. Specifically, we will set the first 10 topic variables as inputs, and keep all other variables at default.

Text Import node is not applicable since our dataset is directly provided. Text Rule-Builder is also not applicable since we are relying on predictive models to provide results instead.

Model

Predictive modeling helps us to find good rules(Models) for guessing the values of one or more variables in a data set from the values of the other variables in the data set. Once a good rule has been found, it can be applied to the new datasets to predict future scenarios. As in the given consumer complaints dataset our target variable “Consumer_disputed_” is categorical, which gives us information about the customer if he/she is disputing or not. To predict future values we have done four models: decision tree, regression, MBR and Neural Network.



Decision Tree Node

To do decision tree modeling attach a decision tree node to the Metadata node. Change the Assessment Measure property in the subtree section to Average Square Error. Run the Decision Tree node.

Property Panel - Decision Tree Node

Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25

Once the model executed successfully after that check the results as our primary goal is to run different models and select one from it, but it might be informative to examine how the decision tree chose to partition the data.

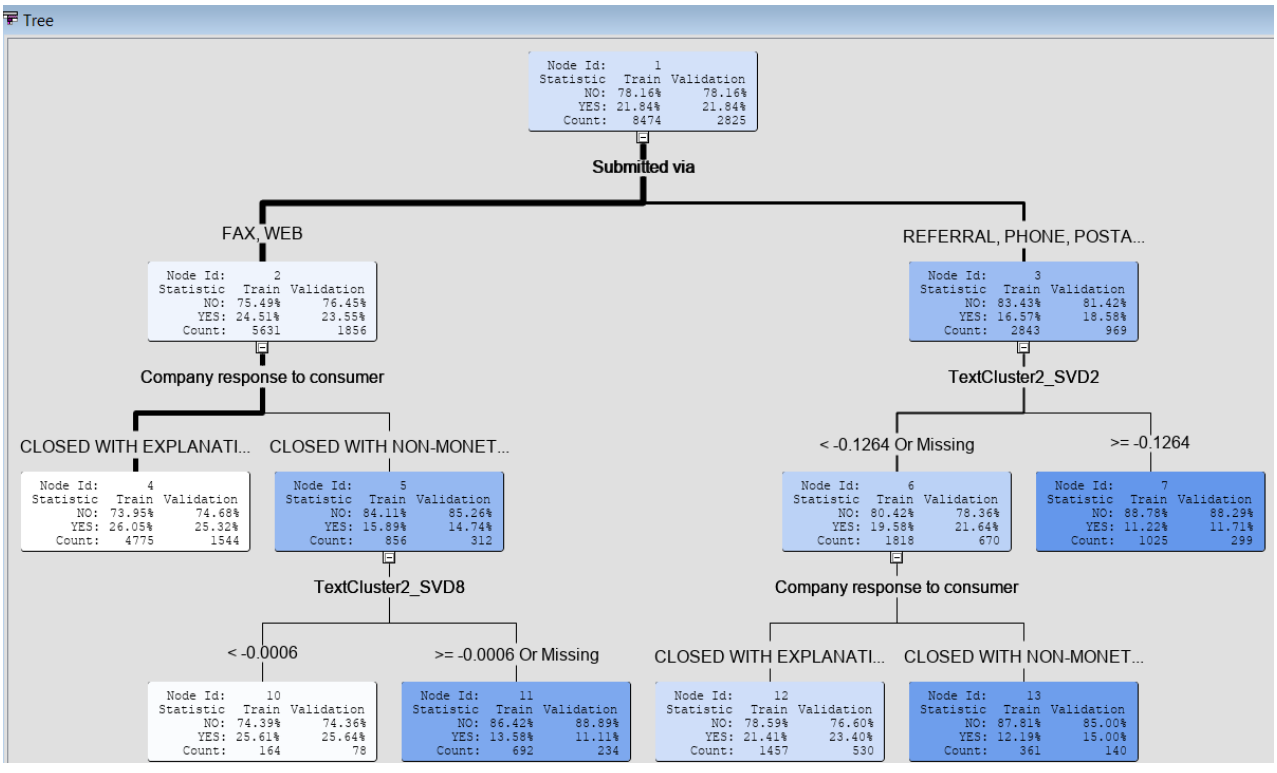
Please find the screenshots of the results which we got from the running the decision tree node.

Variable Importance

Variable Importance

Variable Name	Label	Importance	Validation Importance	Ratio of Validation to Training Importance	Number of Splitting Rules
Submitted via	Submitted via	1.0000	0.5223	0.5223	1
Company response t...	Company response to ...	0.9145	1.0000	1.0934	2
TextCluster2 SVD2		0.6203	0.7606	1.2261	1
TextCluster2 SVD8		0.4012	0.5533	1.3790	1
TextCluster2 SVD1		0.0000	0.0000	.	0
TextCluster2 SVD3		0.0000	0.0000	.	0
TextCluster2 SVD4		0.0000	0.0000	.	0
TextCluster2 SVD10		0.0000	0.0000	.	0
TextCluster2 SVD11		0.0000	0.0000	.	0
TextCluster2 SVD7		0.0000	0.0000	.	0
TextTopic2 raw3	+attempt,collect debt,c...	0.0000	0.0000	.	0
TextCluster2 SVD9		0.0000	0.0000	.	0

Decision Tree

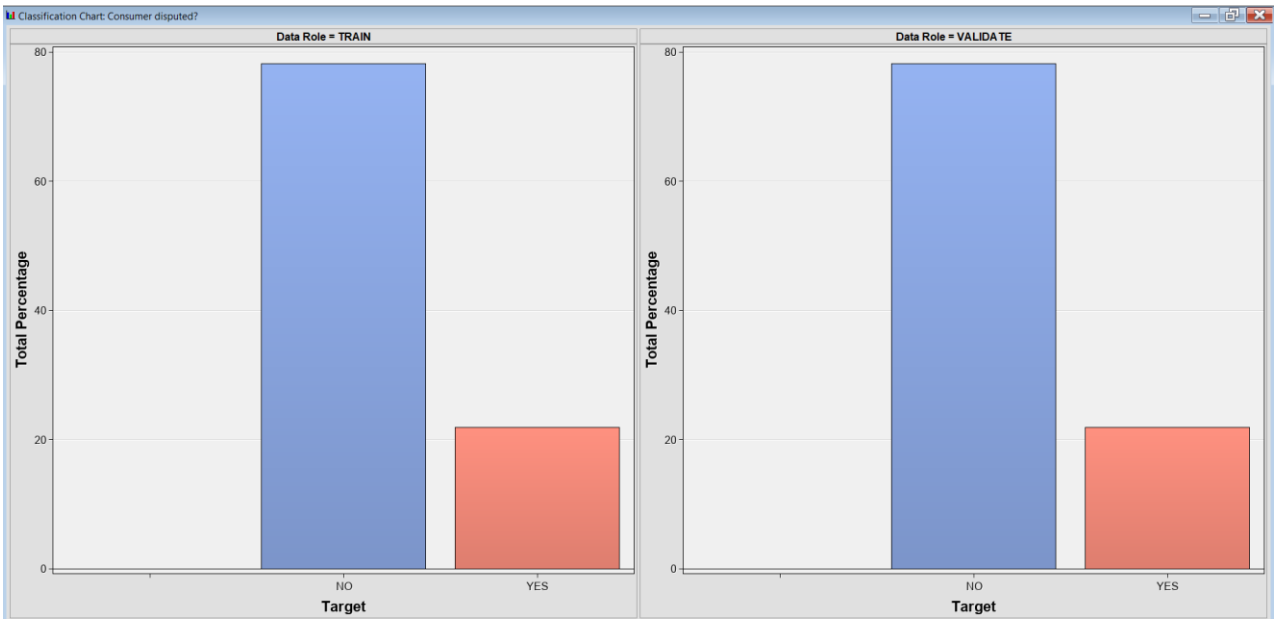


Fit Statistics

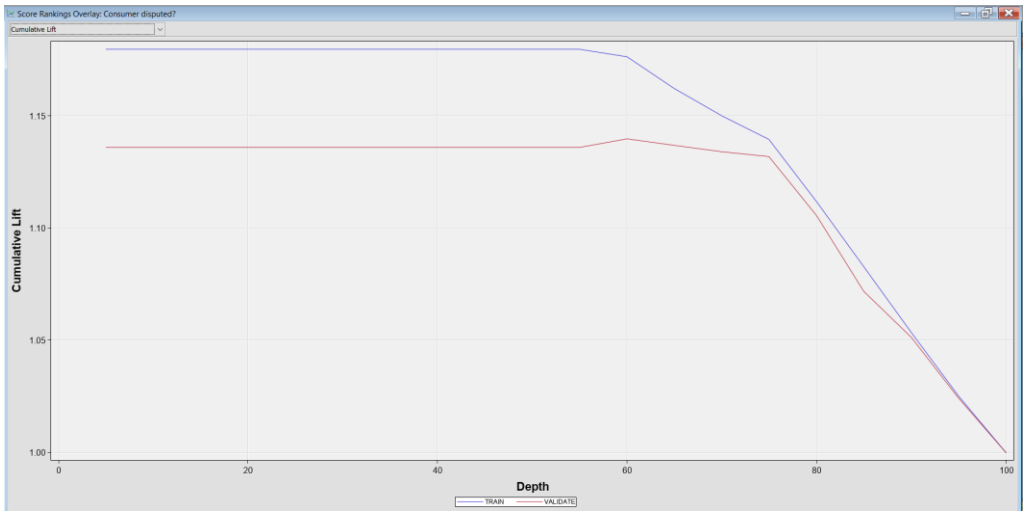
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Consumer disput...	Consumer dispute...	NOBS	Sum of Frequencies	8474	2825	2828
Consumer disput...	Consumer dispute...	MISC	Misclassification R...	0.218433	0.218407	0.218883
Consumer disput...	Consumer dispute...	MAX	Maximum Absolut...	0.887805	0.887805	0.887805
Consumer disput...	Consumer dispute...	SSE	Sum of Squared E...	2836.614	948.5189	949.1142
Consumer disput...	Consumer dispute...	ASE	Average Squared ...	0.167372	0.167879	0.167807
Consumer disput...	Consumer dispute...	RASE	Root Average Squ...	0.409111	0.409731	0.409642
Consumer disput...	Consumer dispute...	DIV	Divisor for ASE	16948	5650	5656
Consumer disput...	Consumer dispute...	DFT	Total Degrees of F...	8474	.	.

Classification Chart



Cumulative Lift



From the above screenshots we can see in the fit statistics results the model is not overfitting because the difference between the training misclassification rate and validation misclassification rate is not much. And same for the training misclassification rate and test misclassification rate.

In this model, we have identified the most important predictor variable to predict our target. The most important variable is “Sumitted_via”

Regression Node

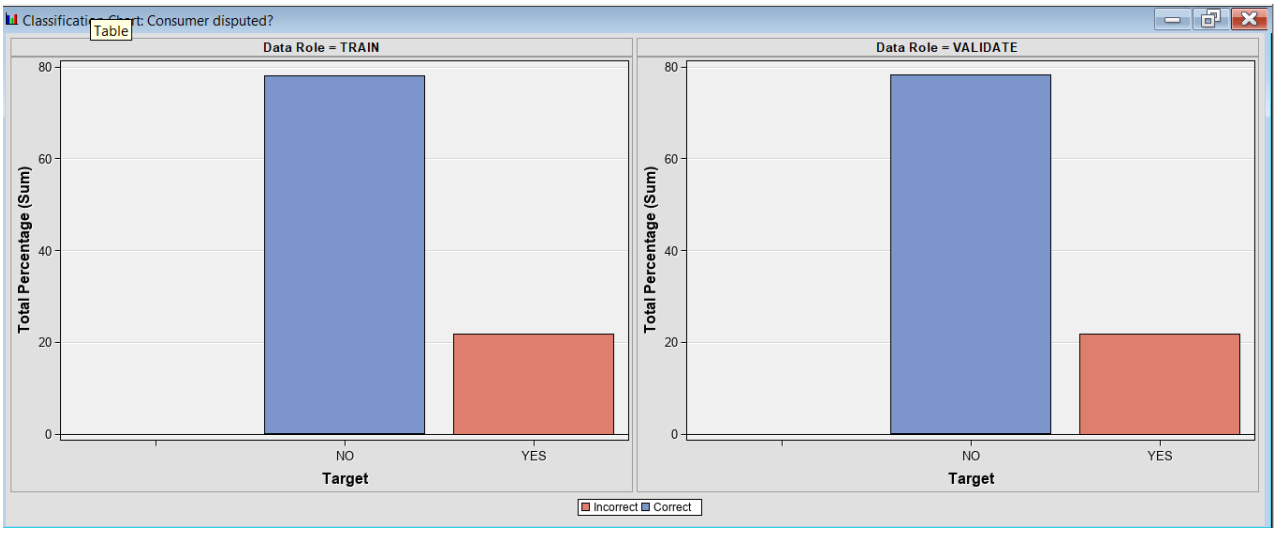
To do the regression modeling attach a Regression node to the Metadata node. After attaching the node, we have not changed any property panel properties. Run the Regression node.

Fit Statistics

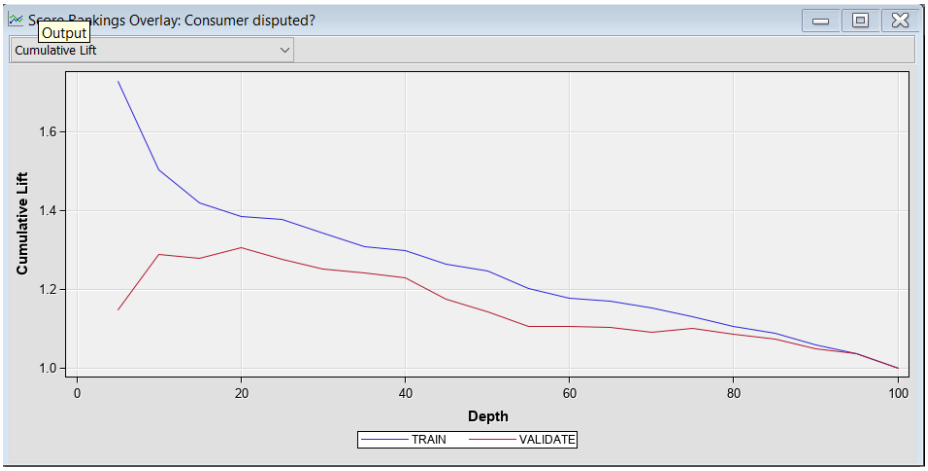
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Consumer_disputed	Consumer disputed?	AIC	Akaike's Information Criteri...	8744.805	.	.
Consumer_disputed	Consumer disputed?	ASE	Average Squared Error	0.166299	0.169321	0.168467
Consumer_disputed	Consumer disputed?	AVERR	Average Error Function	0.511022	0.522353	0.522139
Consumer_disputed	Consumer disputed?	DFE	Degrees of Freedom for Er...	8432	.	.
Consumer_disputed	Consumer disputed?	DFM	Model Degrees of Freedom	42	.	.
Consumer_disputed	Consumer disputed?	DFT	Total Degrees of Freedom	8474	.	.
Consumer_disputed	Consumer disputed?	DIV	Divisor for ASE	16948	5650	5656
Consumer_disputed	Consumer disputed?	ERR	Error Function	8660.805	2951.294	2953.219
Consumer_disputed	Consumer disputed?	FPE	Final Prediction Error	0.167955	.	.
Consumer_disputed	Consumer disputed?	MAX	Maximum Absolute Error	0.942151	0.998331	0.998548
Consumer_disputed	Consumer disputed?	MSE	Mean Square Error	0.167127	0.169321	0.168467
Consumer_disputed	Consumer disputed?	NOBS	Sum of Frequencies	8474	2825	2828
Consumer_disputed	Consumer disputed?	NW	Number of Estimate Weights	42	.	.
Consumer_disputed	Consumer disputed?	RASE	Root Average Sum of Squ...	0.407797	0.411486	0.410447
Consumer_disputed	Consumer disputed?	RFPE	Root Final Prediction Error	0.409824	.	.
Consumer_disputed	Consumer disputed?	RMSE	Root Mean Squared Error	0.408812	0.411486	0.410447
Consumer_disputed	Consumer disputed?	SBC	Schwarz's Bayesian Criteri...	9040.685	.	.
Consumer_disputed	Consumer disputed?	SSE	Sum of Squared Errors	2818.429	956.6618	952.848
Consumer_disputed	Consumer disputed?	SUMW	Sum of Case Weights Tim...	16948	5650	5656
Consumer_disputed	Consumer disputed?	MISC	Misclassification Rate	0.218197	0.219115	0.218529

Classification Chart



Cumulative Lift



From the above screenshots we can see in the fit statistics results the regression model is not overfitting because the difference between the training misclassification rate and validation misclassification rate is not much. And same for the training misclassification rate and test misclassification rate.

MBR Node

To do MBR modeling attach an MBR (Memory Based Reasoning) node to the Metadata node. Change the Number of Neighbors property in the Train section to 8. Select the Variables property, and change the Use status of all input variables to Rejected. Then change the Use status of all TextCluster_SVDn variables to
Yes. These variables are orthogonal, and hence can be used as inputs to the MBR node. Otherwise, you would need to use a method such as principal components to convert inputs to orthogonal inputs. Run the MBR node.

MBR Property Panel

.. Property	Value
General	
Node ID	MBR
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Method	RD-Tree
Number of Neighbors	16
Epsilon	0.0
Number of Buckets	8
Weighted	Yes
Create Nodes	No
Create Neighbor Variables	Yes
Status	
Create Time	6/3/21 5:48 PM
Run ID	03b785ba-baee-4584-8f03-996763abeb16
Last Error	
Last Status	Complete
Last Run Time	6/3/21 5:52 PM
Run Duration	0 Hr. 0 Min. 11.27 Sec.
Grid Host	
User-Added Node	No

Variable Property

Variables - MBR

(none) not Equal to

Columns: Label Mining Basic Statist

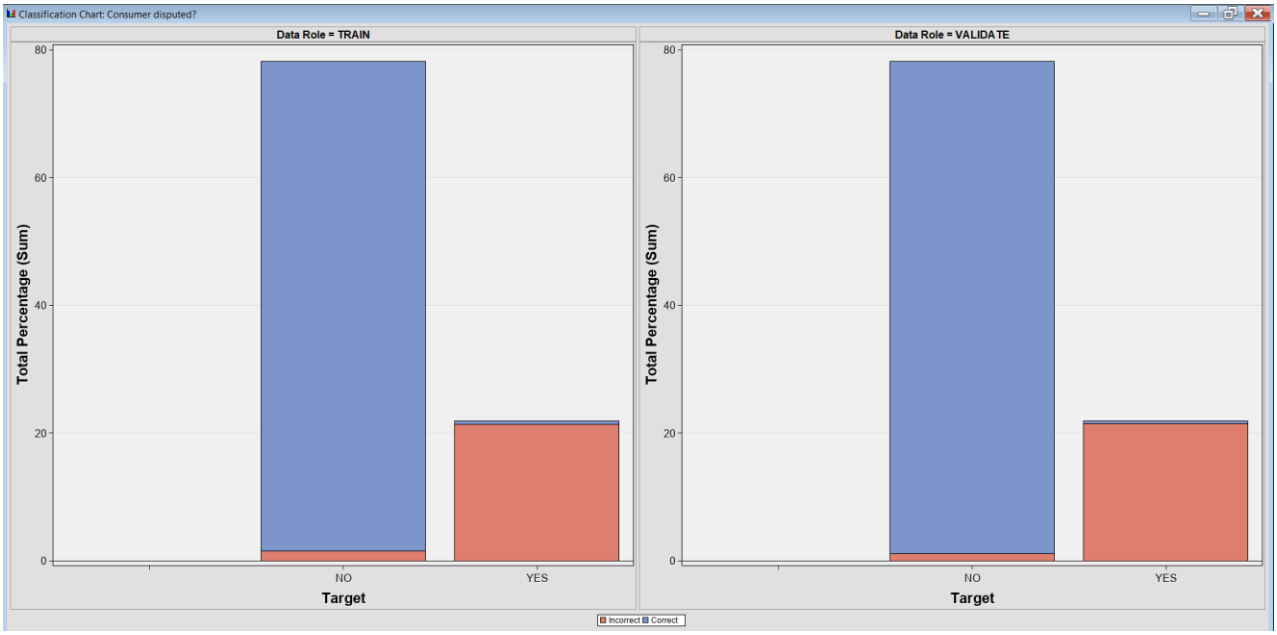
Name	Use	Report	Role	Level
Complaint_ID	Yes	No	ID	Nominal
Consumer_disputed_	Yes	No	Target	Binary
Date_received	Default	No	Rejected	Interval
Date_sent_to_company	Default	No	Rejected	Interval
TextCluster2_SVD1	Yes	No	Input	Interval
TextCluster2_SVD10	Yes	No	Input	Interval
TextCluster2_SVD11	Yes	No	Input	Interval
TextCluster2_SVD2	Yes	No	Input	Interval
TextCluster2_SVD3	Yes	No	Input	Interval
TextCluster2_SVD4	Yes	No	Input	Interval
TextCluster2_SVD5	Yes	No	Input	Interval
TextCluster2_SVD6	Yes	No	Input	Interval
TextCluster2_SVD7	Yes	No	Input	Interval
TextCluster2_SVD8	Yes	No	Input	Interval
TextCluster2_SVD9	Yes	No	Input	Interval
TextCluster2_prob1	Yes	No	Rejected	Interval
TextCluster2_prob2	Yes	No	Rejected	Interval
TextCluster2_prob3	Yes	No	Rejected	Interval
TextCluster2_prob4	Yes	No	Rejected	Interval
TextTopic2_raw1	No	No	Input	Interval
TextTopic2_raw10	No	No	Input	Interval
TextTopic2_raw11	No	No	Input	Interval
TextTopic2_raw12	No	No	Input	Interval
TextTopic2_raw13	No	No	Input	Interval
TextTopic2_raw14	No	No	Input	Interval
TextTopic2_raw15	No	No	Input	Interval
TextTopic2_raw2	No	No	Input	Interval
TextTopic2_raw3	No	No	Input	Interval
TextTopic2_raw4	No	No	Input	Interval
TextTopic2_raw5	No	No	Input	Interval
TextTopic2_raw6	No	No	Input	Interval
TextTopic2_raw7	No	No	Input	Interval
TextTopic2_raw8	No	No	Input	Interval
TextTopic2_raw9	No	No	Input	Interval
DOCUMENT	No	No	ID	Nominal
dataobs	No	No	ID	Interval

Fit Statistics

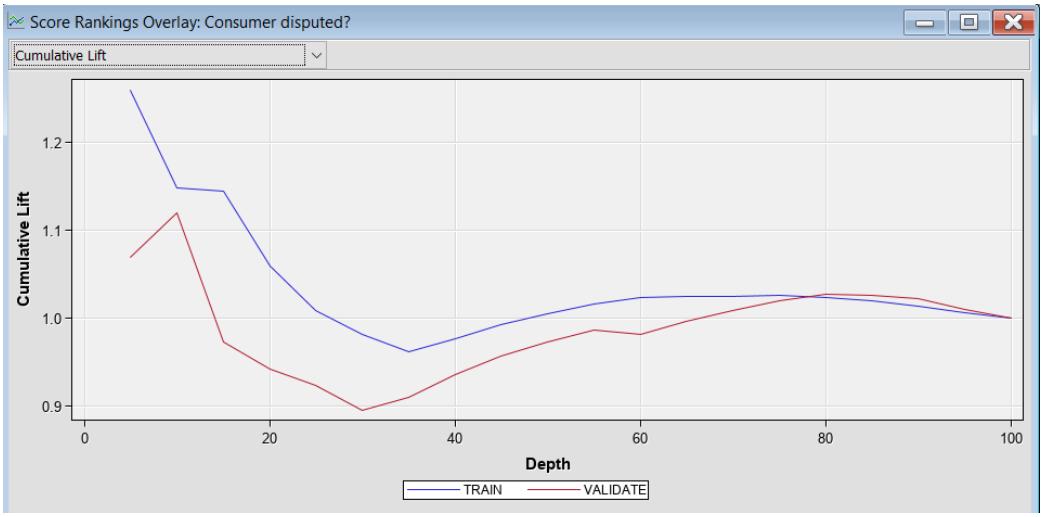
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Consumer disputed	Consumer disputed?	NW	Number of Estimated Weig...	0	.	.
Consumer disputed	Consumer disputed?	NOBS	Sum of Frequencies	193	2999	3004
Consumer disputed	Consumer disputed?	SUMW	Sum of Case Weights Tim...	579	8997	9012
Consumer disputed	Consumer disputed?	DFT	Total Degrees of Freedom	386	.	.
Consumer disputed	Consumer disputed?	DFM	Model Degrees of Freedom	0	.	.
Consumer disputed	Consumer disputed?	DFE	Degrees of Freedom for Er...	386	.	.
Consumer disputed	Consumer disputed?	ASE	Average Squared Error	0.116244	0.169771	0.174033
Consumer disputed	Consumer disputed?	RASE	Root Average Squared Error	0.340946	0.412033	0.417173
Consumer disputed	Consumer disputed?	DIV	Divisor for ASE	579	8997	9012
Consumer disputed	Consumer disputed?	SSE	Sum of Squared Errors	67.30555	1527.429	1568.385
Consumer disputed	Consumer disputed?	MSE	Mean Squared Error	0.116244	0.169771	0.174033
Consumer disputed	Consumer disputed?	RMSE	Root Mean Squared Error	0.340946	0.412033	0.417173
Consumer disputed	Consumer disputed?	AVERR	Average Error Function	0.45899	0.909429	0.915815
Consumer disputed	Consumer disputed?	ERR	Error Function	265.7554	8182.129	8253.323
Consumer disputed	Consumer disputed?	MAX	Maximum Absolute Error	0.998436	1	1
Consumer disputed	Consumer disputed?	FPE	Final Prediction Error	0.116244	.	.
Consumer disputed	Consumer disputed?	RFPE	Root Final Prediction Error	0.340946	.	.
Consumer disputed	Consumer disputed?	AIC	Akaike's Information Criteri...	265.7554	.	.
Consumer disputed	Consumer disputed?	SBC	Schwarz's Bayesian Criteri...	265.7554	.	.
Consumer disputed	Consumer disputed?	MISC	Misclassification Rate	0.196891	0.27109	0.276298
Consumer disputed	Consumer disputed?	WRONG	Number of Wrong Classific...	38	813	830

Classification Chart



Cumulative Lift



From the above screenshots we can see in the fit statistics results the MBR model may be overfitting because the difference between the training misclassification rate and validation misclassification rate is significant. And same for the training misclassification rate and test misclassification rate.

Neural Network Node

To do neural network modeling Attach a Neural Network node to the Metadata node. We have not changed any property panel settings. We executed the node to see the results.

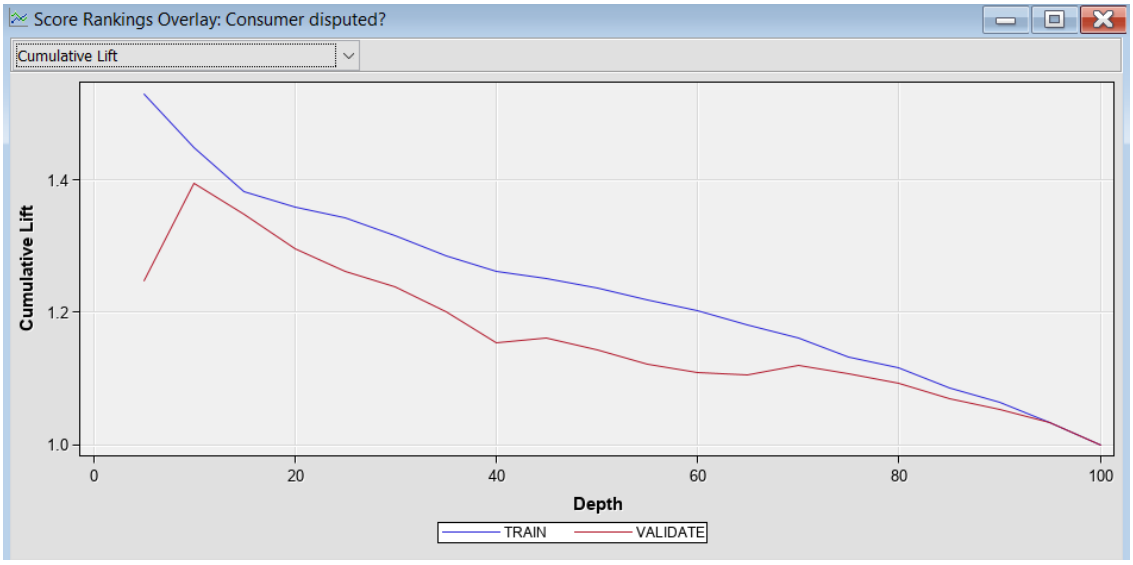
Fit Statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Consumer disputed	Consumer disputed?	DFT	Total Degrees of Freedom	8474	.	.
Consumer disputed	Consumer disputed?	DFE	Degrees of Freedom for E...	8323	.	.
Consumer disputed	Consumer disputed?	DFM	Model Degrees of Freedom	151	.	.
Consumer disputed	Consumer disputed?	NW	Number of Estimated Wei...	151	.	.
Consumer disputed	Consumer disputed?	AIC	Akaike's Information Crite...	8992.232	.	.
Consumer disputed	Consumer disputed?	SBC	Schwarz's Bayesian Crite...	10055.99	.	.
Consumer disputed	Consumer disputed?	ASE	Average Squared Error	0.166841	0.168586	0.168346
Consumer disputed	Consumer disputed?	MAX	Maximum Absolute Error	0.939589	0.930596	0.945659
Consumer disputed	Consumer disputed?	DIV	Divisor for ASE	16948	5650	5656
Consumer disputed	Consumer disputed?	NOBS	Sum of Frequencies	8474	2825	2828
Consumer disputed	Consumer disputed?	RASE	Root Average Squared Er...	0.408462	0.410593	0.4103
Consumer disputed	Consumer disputed?	SSE	Sum of Squared Errors	2827.62	952.5121	952.1652
Consumer disputed	Consumer disputed?	SUMW	Sum of Case Weights Tim...	16948	5650	5656
Consumer disputed	Consumer disputed?	FPE	Final Prediction Error	0.172895	.	.
Consumer disputed	Consumer disputed?	MSE	Mean Squared Error	0.169868	0.168586	0.168346
Consumer disputed	Consumer disputed?	RFPE	Root Final Prediction Error	0.415806	.	.
Consumer disputed	Consumer disputed?	RMSE	Root Mean Squared Error	0.41215	0.410593	0.4103
Consumer disputed	Consumer disputed?	AVERR	Average Error Function	0.512759	0.518115	0.517323
Consumer disputed	Consumer disputed?	ERR	Error Function	8690.232	2927.348	2925.979
Consumer disputed	Consumer disputed?	MISC	Misclassification Rate	0.218433	0.218407	0.218883
Consumer disputed	Consumer disputed?	WRONG	Number of Wrong Classifi...	1851	617	619

Classification Chart

Cumulative Lift



From the above screenshots we can see in the fit statistics results the neural model is not overfitting because the difference between the training misclassification rate and validation misclassification rate is not much. And same for the training misclassification rate and test misclassification rate.

Assess

Model Comparison Node

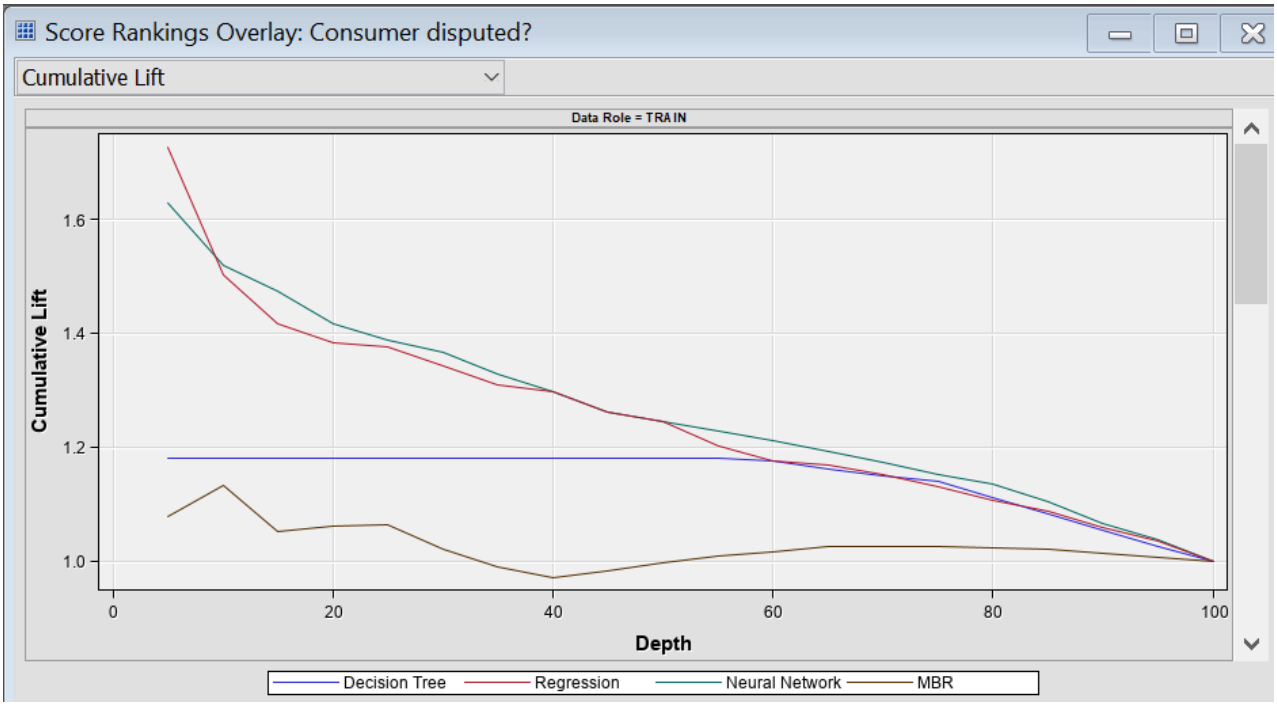
After creating and running our models, we used the Model Comparison node to assess the corresponding results and to determine our best model. We decided on measuring each model's performance based on fit statistics like misclassification rate, RMSE, and MSE and the ROC and lift charts.

- Misclassification Rate : The percentage of classifications that were incorrect, the values closer to zero are better.
- Mean Square Error : MSE measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. In general, a lower RMSE is better than a higher one.
- Root Mean Square Error : RMSE is the square root of the average of squared errors. In general, a lower RMSE is better than a higher one.
- ROC Chart : Shows the trade-off between sensitivity and specificity. Classifiers that give curves closer to the top-left corner indicate a better performance.

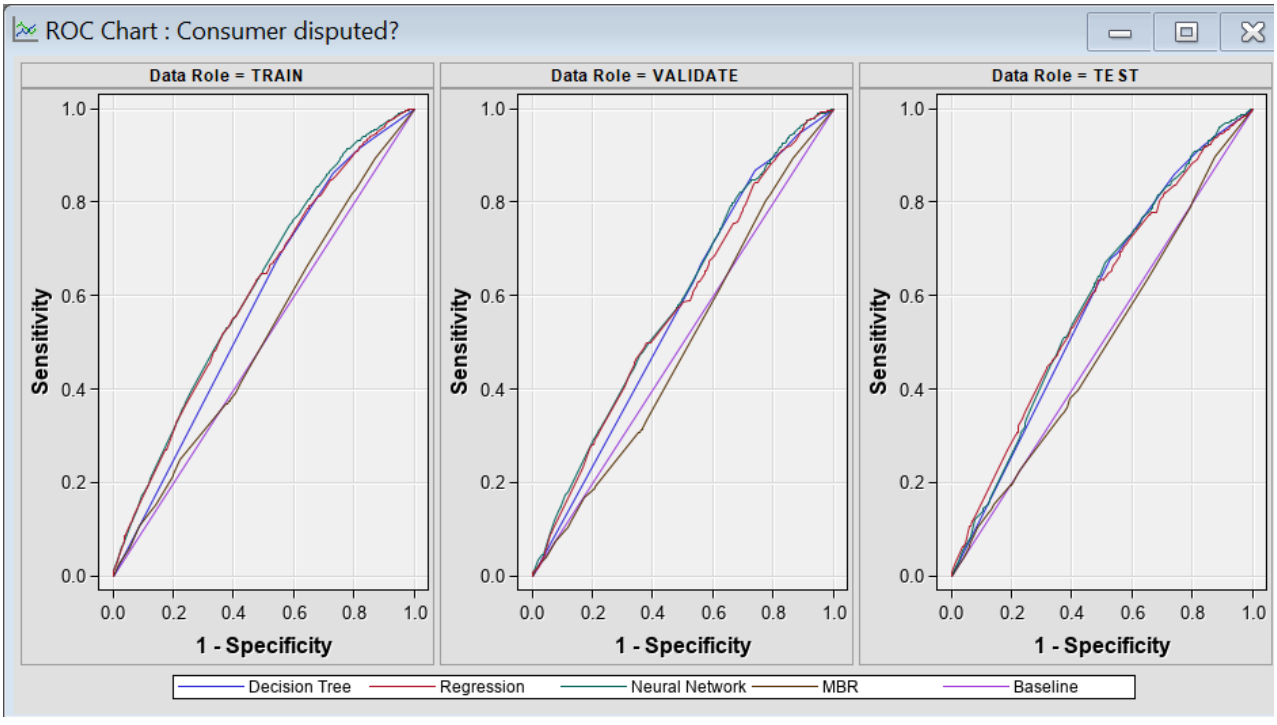
Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Train: Misclassification Rate	Train: Average Squared Error	Train: Root Average Squared Error	Valid: Misclassification Rate	Valid: Average Squared Error	Valid: Root Average Squared Error	Test: Misclassification Rate	Test: Average Squared Error	Test: Root Average Squared Error	Target Variable
Y	Tree	Tree	Decisio...	0.2184...	0.1673...	0.4091...	0.2184...	0.1678...	0.4097...	0.2188...	0.1678...	0.4096...	Consumer disputed
	Neural	Neural	Neural ...	0.2183...	0.1658...	0.4072...	0.2184...	0.1681...	0.4100...	0.2185...	0.1686...	0.41062	Consumer disputed
	Reg	Reg	Regres...	0.2181...	0.1662...	0.4077...	0.2191...	0.1693...	0.4114...	0.2185...	0.1684...	0.4104...	Consumer disputed
	MBR	MBR	MBR	0.1968...	0.1162...	0.3409...	0.27109	0.1697...	0.4120...	0.2762...	0.1740...	0.4171...	Consumer disputed

Cumulative Lift



ROC Chart

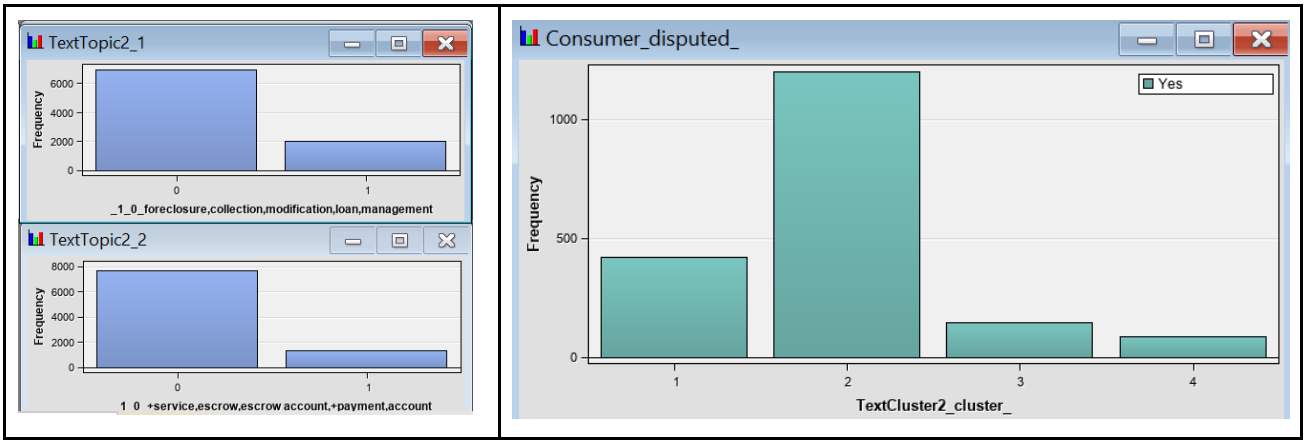


Results

Based on the results obtained from the ‘Model Comparison’ node, our best model overall was the Decision Tree Model. We made our decision based on the training, validation, and test misclassification rate and RMSE. The Decision Tree model has a training and validation misclassification of about 0.2184 and a validation RMSE of 0.2089 and MSE of 0.168 an overall accuracy of 71.86%. The tree model further has a test MSE of 0.1678 and a misclassification rate of 0.2188 and a RMSE of 0.4096. The model performs better on our test partition without overfitting the training or validation datasets. The tree model further has lower RMSE and misclassification rates when compared to our other models. The tree model provides a lift of 1.14 and 1.21 on validation and test partitions, respectively. The neural network model has a better lift but has similar misclassification rates and RMSE when compared with the tree mode. The tree model further has a ROC index of 0.59 on the test partition which is the same for most of our models and a sizable area under the curve which is indicative of better accuracy and of a stronger model. The tree model captures sufficient variance and inherent patterns from the training partition while remaining relatively less complex and performing well on the validation and test partitions. The tree model has a better rate of classification and a higher accuracy of predictions for both classes of the target variable.

Conclusions and Recommendations

From our text mining analysis, we can identify Cluster 2 and Topics 1-2 as groupings with the largest amount of absolute complaints. This provides immediate learnings for the business as they can examine the terms that typically accompany a dispute and begin to consider what actions they may wish to pursue.



For modeling, identifying potential disputes correctly in turn can lead to process efficiencies or training opportunities for the business. Therefore, we conclude misclassification rate is more important in evaluating the performance and any improvements to the model should aim to minimize this metric, particularly false negatives as those especially indicate missed opportunities for the business.

We see a few paths to improving our model from misclassification rate of 21.84%:

(1) **More precise stop and synonym lists**: For text parsing and filtering, we derived a stop list from the text variable itself. However, if the domain experts participate in the stop or start compilation it is likely more noise can be eliminated. Similarly, synonym list that accurately mimic industry language and commonly interchangeable terms can be helpful in deriving more value from the text variable.

(2) **Refine text variable**: As we noted previously, the variable issue appears to have predetermined statements that consumers selected from the list; we expect that these statements already have a certain level of information separation. We believe this impacts how terms should be weighted and consulting the domain experts may be helpful here to further refine this text variable for modeling.

References

[1] SAS Enterprise Miner 14.3 Reference Help
https://documentation.sas.com/?docsetId=tmref&docsetTarget=n0sebvirmou078n1sxso7lqkwyln.htm&docsetVersion=14.3&locale=en