UNIVERSITY OF CONNECTICUT

OPIM 5671- DATA MINING AND BUSINESS INTELLIGENCE

Sudip Bhattacharjee

Text Mining Group Project Report

By

TEAM 3

Nikhila

Irfan Shaik

Raheem Abdul

Amazon Fine Food Reviews

Table of Contents

EXECUTIVE SUMMARY

Embarking on the "Amazon Fine Food Reviews" project has been a journey of unraveling customer sentiments in the vast landscape of online reviews. Beyond the technical intricacies, this project holds immense significance for businesses seeking to truly understand the voices of their customers. By diving into the intricacies of user profiles, we've moved beyond just sentiment analysis; we've laid the groundwork for personalized recommendations and insights into individual preferences through features like Profile Name and User Id. The inclusion of time as a dynamic factor adds a real-world touch, allowing us to grasp how sentiments ebb and flow over time, helping businesses stay ahead of evolving trends. Our commitment to ethical considerations ensures that this analysis respects user privacy, establishing a foundation of trust between consumers and businesses.

As we chart the implications of our findings, the scalability of our models becomes apparent—these aren't just tools for e-commerce but adaptable frameworks with potential applications in hospitality, healthcare, and finance. The project isn't merely about analyzing reviews; it's about shaping how businesses engage with their customers, fostering a proactive approach to meet changing needs.

Looking forward, the project isn't just a static analysis; it's a steppingstone for future exploration. We're not just content with the current state of technology; we're laying the groundwork for what's next. Deepening our understanding of customer feedback through emerging technologies like advanced neural networks and deep learning is on the horizon. In essence, this project isn't just about numbers and algorithms; it's about harnessing technology to better connect businesses with the people they serve, providing not just insights but a pathway for continual innovation in the dynamic digital landscape.

1. Introduction

In the rapidly evolving landscape of e-commerce, customer feedback has become a cornerstone for businesses aiming to stay attuned to consumer sentiments. The "Amazon Fine Food Reviews" project endeavors to navigate this vast sea of opinions by employing sophisticated Text Mining techniques. By delving into the nuances of user reviews, the project seeks to uncover not only the overall sentiment but also the underlying patterns that drive consumer satisfaction or dissatisfaction. Leveraging advanced algorithms, such as Neural Networks and Decision Trees, this project aims to provide businesses with actionable insights, enabling them to enhance customer experiences and refine product offerings.

1.1 Problem Statement

The surge in online reviews has created a wealth of unstructured data, making it challenging for businesses to distill meaningful information from the vast array of customer feedback. The "Amazon Fine Food Reviews" project addresses the need for a comprehensive sentiment analysis solution. The problem at hand involves deciphering the sentiments expressed in customer reviews accurately. Additionally, the project tackles the inherent challenges of user-generated content, including varied writing styles, potential biases, and the ever-changing nature of language. The goal is to not only categorize reviews as positive or negative but to delve deeper, extracting valuable insights that can inform strategic business decisions.

1.2 Data Description

The dataset consists of 568454 rows which comprise reviews of users who ordered and returned the products on Amazon portal.

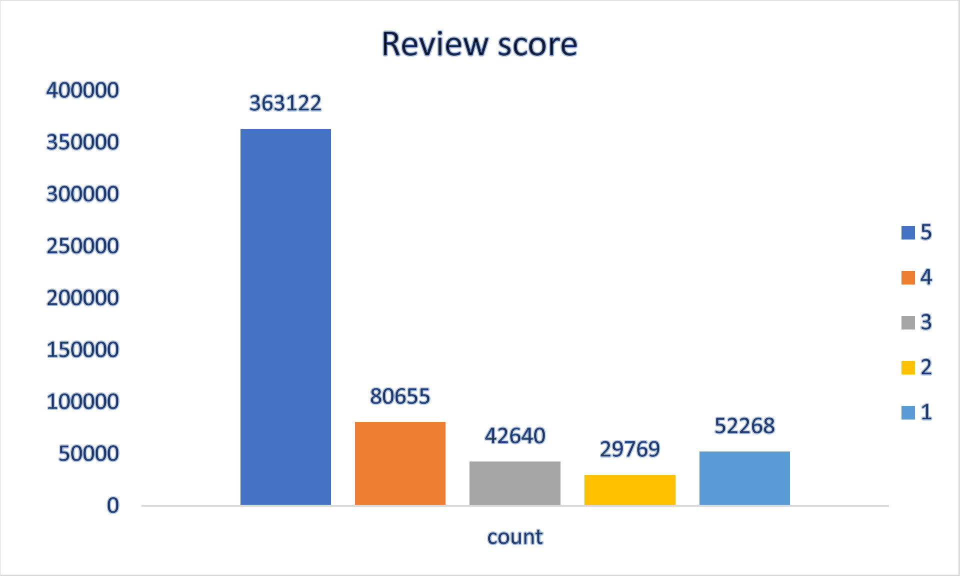
Data Set:

Review attribute	Description	Variable type
Product ID	Unique identifier for the product	Categorical
User ID	Unique identifier for the user	Categorical
Profile Name	Profile of the user	Text
Helpfulness Numerator	Number of users who found the review helpful	Numerical
Helpfulness Denominator	Number of users who voted whether the review was helpful or not	Numerical
Score	Rating between 1 and 5	Ordinal
Time	Timestamp of the review	Numerical
Summary	Brief summary of the review	Text
Text	Text of the review	Text

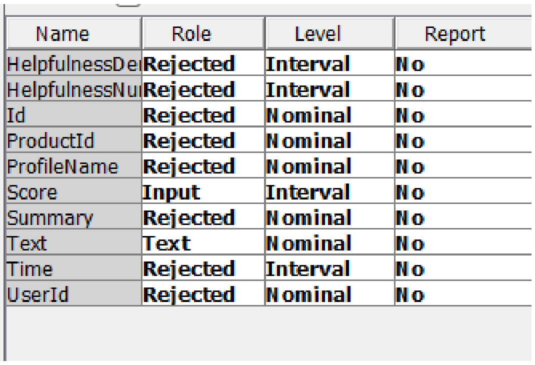
Source of Data set:

https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

Data Visualizations:



As it can be seen from the Figure 4-3, the distribution of the 'Score' attribute is highly skewed towards the score '5' which means the majority of the reviews have 'extremely high' rating.



Name	Role	Level	Report
HelpfulnessDe	Rejected	Interval	No
HelpfulnessNu	Rejected	Interval	No
Id	Rejected	Nominal	No
ProductId	Rejected	Nominal	No
ProfileName	Rejected	Nominal	No
Score	Input	Interval	No
Summary	Rejected	Nominal	No
Text	Text	Nominal	No
Time	Rejected	Interval	No
UserId	Rejected	Nominal	No

1.3 Software Used

We have used SAS Enterprise Miner Workstation 15.1 to create this project. SAS Enterprise Miner Workstation offers a comprehensive environment for text mining projects. It provides a range of features and functionalities to import, preprocess, analyze, model, and evaluate text data efficiently and effectively.

1.4 Data Exploration

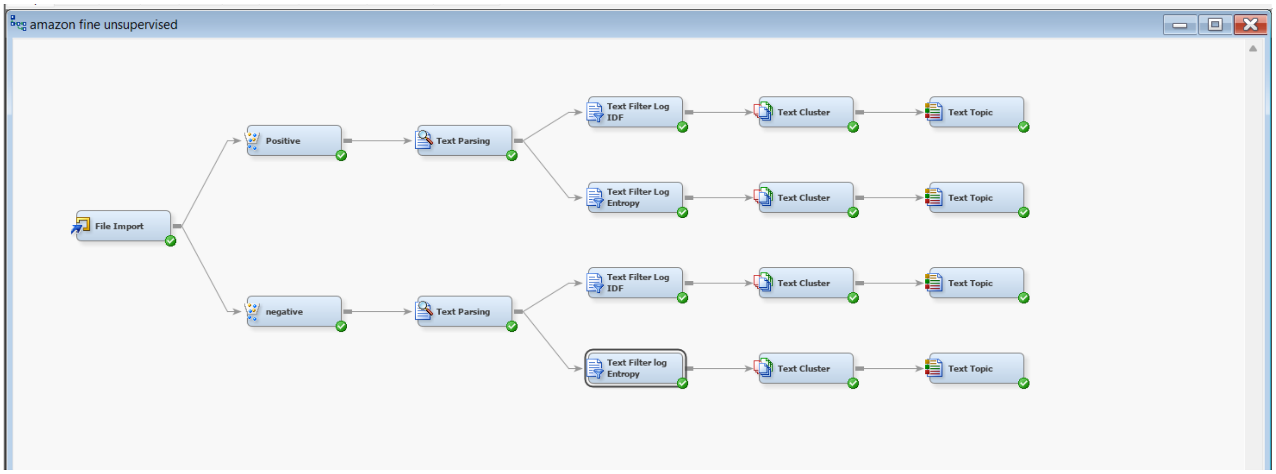
- This dataset with Reviews is of from Oct 1999 - Oct 2012
- There are total of 568,454 reviews
- The total unique users who given the reviews to products are 256,059 users
- There are total 74,258 products in the provided dataset
- 260 users had given > 50 reviews

2. Model Learning

2.1 Unsupervised Model

We created an Unsupervised model first for the Cluster analysis.

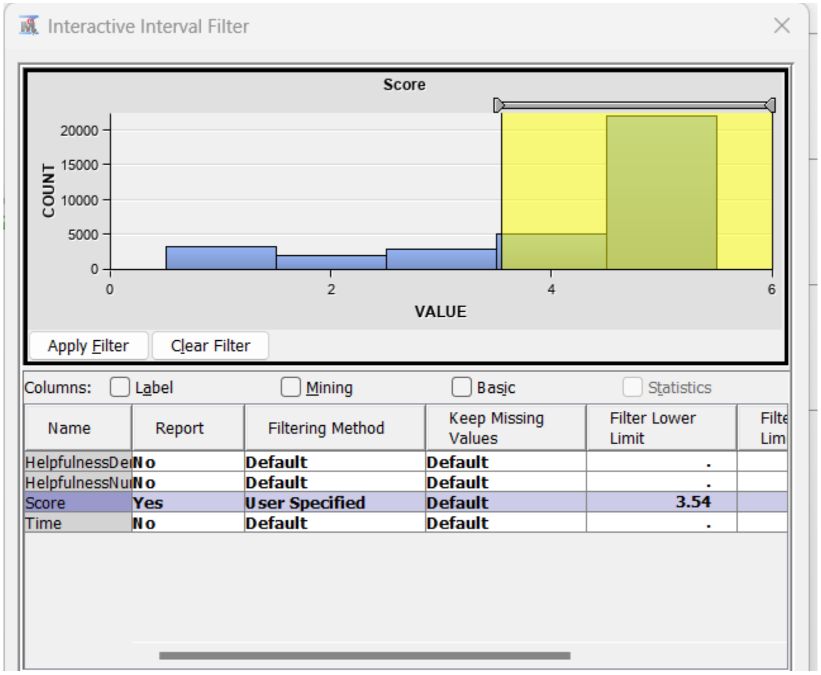
The diagram below depicts the model.



We used two filters in the model which are Positive and Negative for the ratings to be grouped.

2.1.1 Positive Categories

In the score we considered 4& 5 as positive ratings.

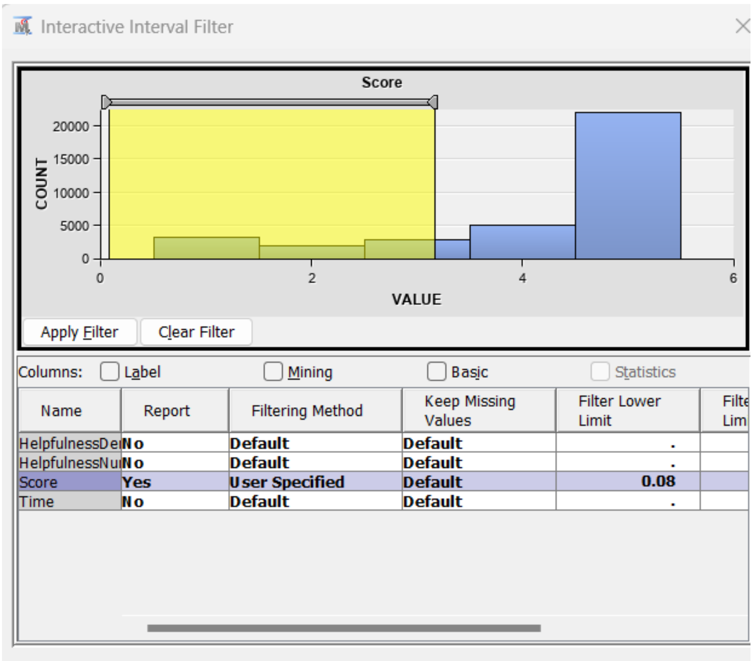


Columns: □ Label □ Mining □ Basic □ Statistics

Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filte Lim
HelpfulnessDe	No	Default	Default	.	
HelpfulnessNu	No	Default	Default	.	
Score	Yes	User Specified	Default	3.54	
Time	No	Default	Default	.	

2.1.2 Negative Categories

In the score we considered 1-3 as negative ratings.

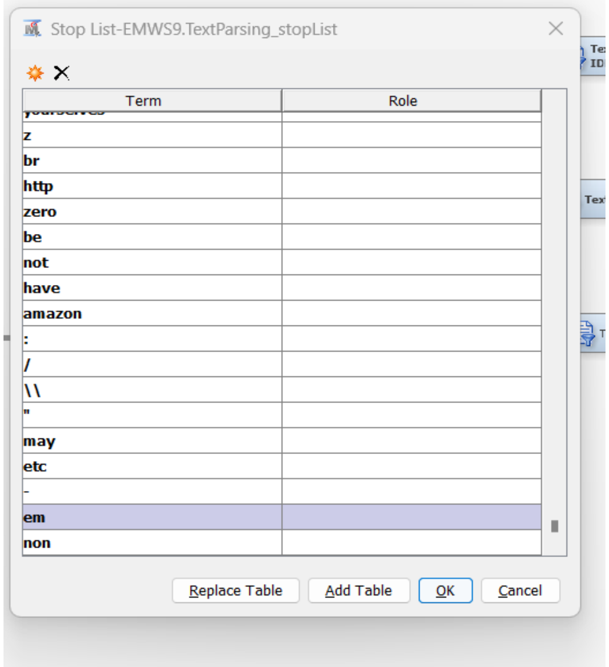


Columns: □ Label □ Mining □ Basic □ Statistics

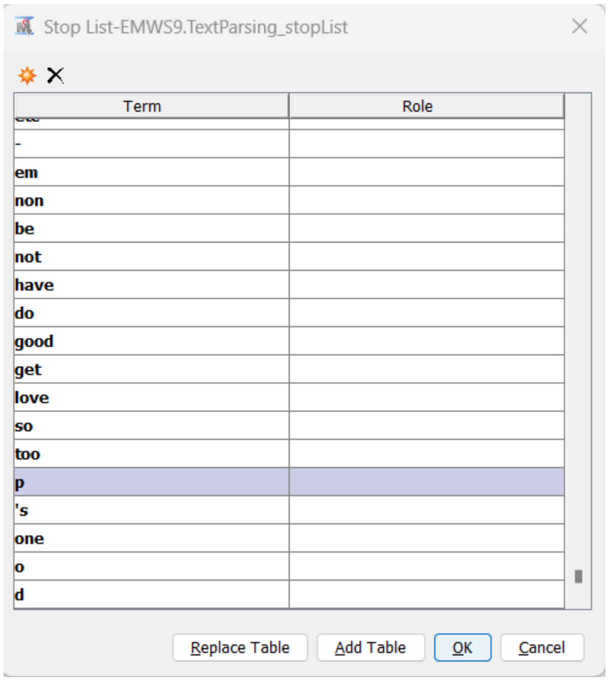
Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filte Lim
HelpfulnessDe	No	Default	Default	.	
HelpfulnessNu	No	Default	Default	.	
Score	Yes	User Specified	Default	0.08	
Time	No	Default	Default	.	

2.1.3 Text Parsing

To refine our analysis and focus on meaningful terms, we implemented a stop list—a carefully curated collection of words like so, have, be, too, not, punctuation marks which are deemed irrelevant to sentiment analysis. This stop list aimed to filter out common words that might not contribute significantly to discerning positive or negative sentiments, enhancing the accuracy and efficiency of our unsupervised learning models.



Term	Role
z	
br	
http	
zero	
be	
not	
have	
amazon	
:	
/	
\\	
"	
may	
etc	
-	
em	
non	



Term	Role
-	
em	
non	
be	
not	
have	
do	
good	
get	
love	
so	
too	
p	
's	
one	
o	
d	

2.1.4 Text Filter

our text filtering methodology, incorporating Log IDF and Entropy measures, represents a meticulous effort to curate a dataset that encapsulates the richness and diversity of language used in customer reviews. This approach sets the stage for more nuanced and accurate sentiment analyses, empowering us to extract deeper insights from the wealth of textual data at our disposal.

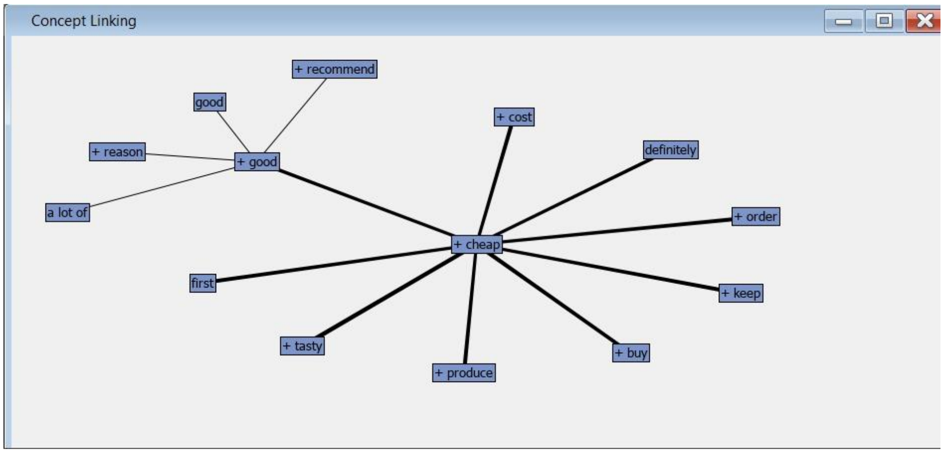
2.1.5 Cluster Analysis

The objective is to categorize and cluster text documents or data points that express favorable sentiments, positive opinions, or related themes.

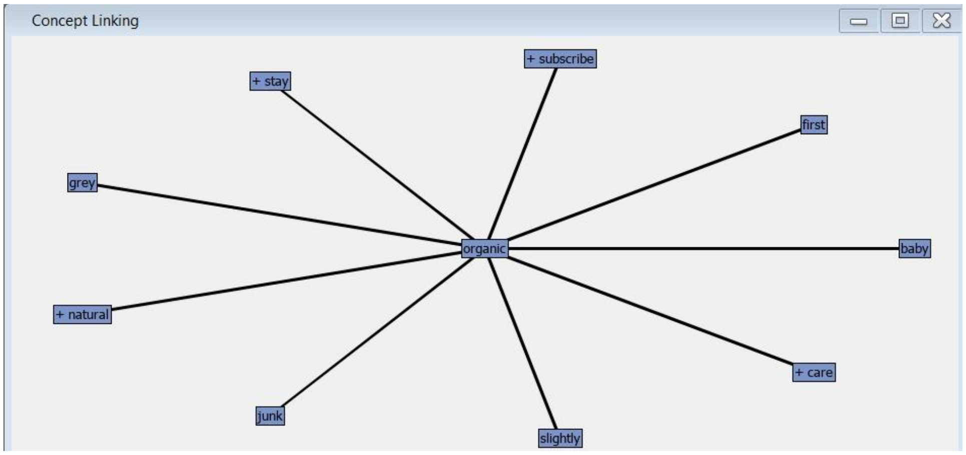
2.1.5.1 Positive Cluster Analysis

Positive Cluster	Cluster	Cluster Descriptive terms
	+Cheap	+Affordable, +Budget-friendly, +Cost-effective, +Inexpensive, +Economical, +Low-cost, +Wallet-friendly, +Thrifty, +Reasonable, +Bargain, +Discounted, +Frugal, + Value-for-money, +Discount
	+Organic	+Natural, +Sustainable, +Chemical-free, +Eco-friendly, +Pure, +Non-GMO, +Locally sourced, +Environmentally, +conscious, +Farm-to-table, +Fresh, +Healthy, +Whole, +Unprocessed, +Biorange, +Earth-friendly, +Clean, +Organic-certified

- Within the positive sentiment cluster, a prominent conceptual grouping revolves around the term "cheap." This suggests a thematic connection where customers express satisfaction with affordable pricing, portraying a positive perception of the cost-effectiveness of the products.



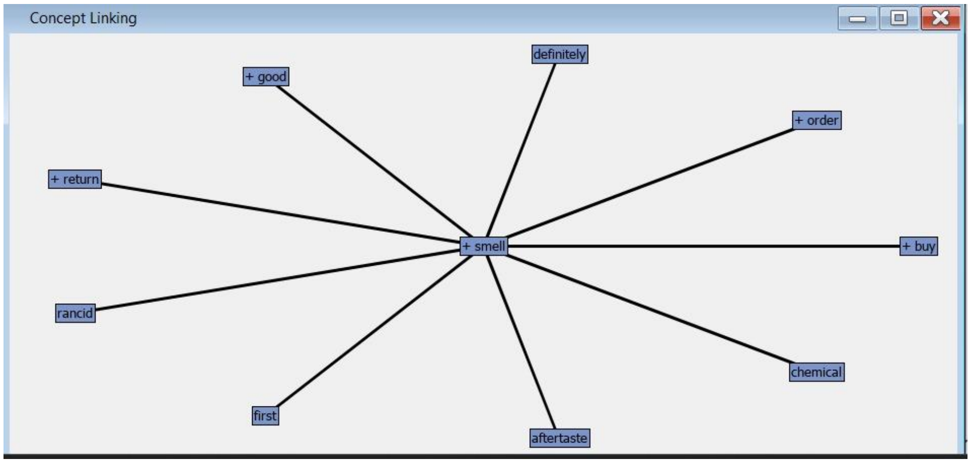
- These additional positive analysis points provide deeper insights into the specific facets of positivity associated with the "organic" concept. The identified terms unveil customer sentiments related to health, sustainability, and lifestyle choices, contributing to a nuanced understanding of positive sentiments within the dataset.



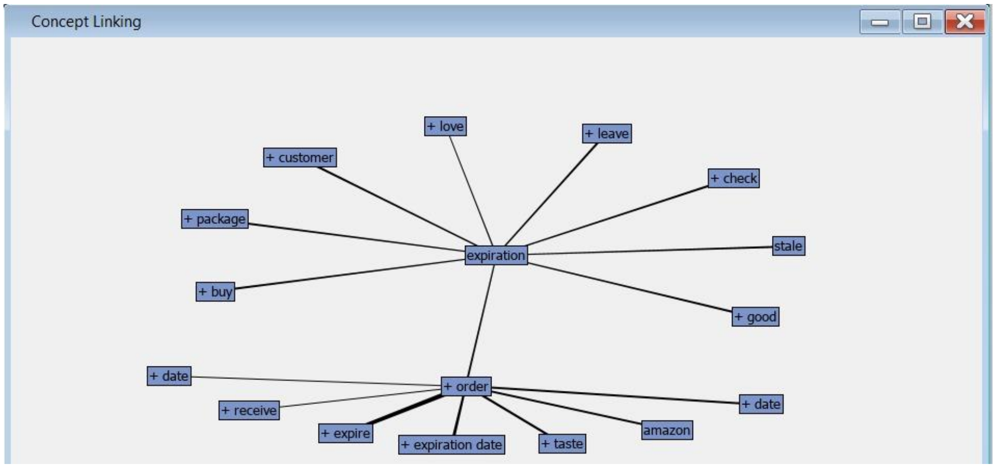
2.1.5.2 Negative cluster analysis

	Cluster	Cluster Descriptive terms
Negative Cluster	+Smell	+Pungent, +Repugnant, +Stale, +Rancid, +Offensive, +Putrid, +Displeasing, +Foul +Stinky, +Tasteful +Noxious, +Fetid, +Unpleasant, +Unappetizing, +Musty, +Moldy, +Damp, +Subtle +Delicate
	+Expiration	+Short shelf life, +Limited Freshness, +Imminent expiration, +Close expiration date, +Near-end expiration, +Restricted storage time, +time-sensitive Freshness, +Brief shelf viability, +Expiring soon, +Limited storage lifespan, +Close use-by date, +Imminent spoilage, +Expiry proximity

- These identified negative analysis points shed light on specific aspects of customer dissatisfaction within the dataset, such as unpleasant smells, aftertaste concerns, chemical perceptions, and challenges related to return policies. Understanding these negative sentiments is crucial for businesses to address customer concerns and enhance overall product satisfaction.



- The terms "extraction date," "receive," "expire," and "taste" collectively highlight negative sentiments related to date-related concerns. Customers express dissatisfaction with the extraction date, the taste of products nearing expiration, and potential issues with the expiration date itself.



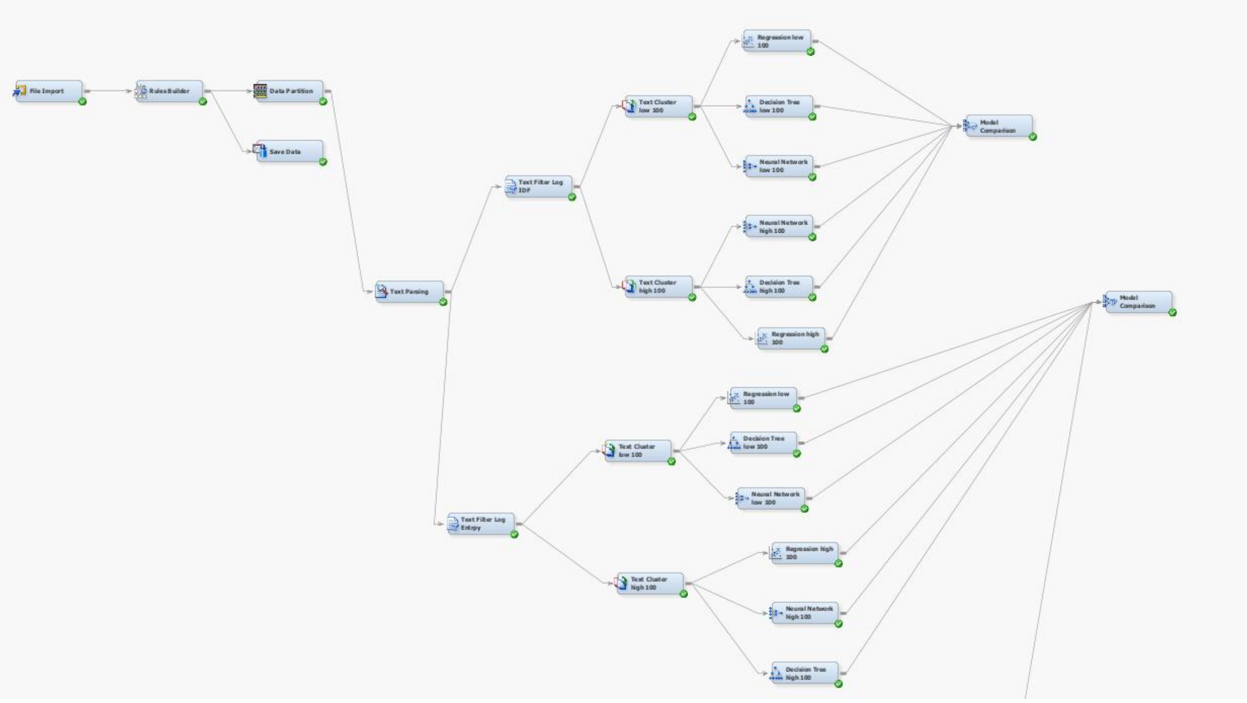
2.1.6 Cluster Analysis Model Findings

- In the **Cheap** Cluster, some favor products are +Affordable, +Budget-friendly, +Cost-effective, +Inexpensive, and +Economical, meeting the standards of such cost-effectiveness can pose a challenge. Striking a balance between being +Low-cost and +Wallet-friendly is important, and being +Thrifty in resource allocation is valued. Seeking +Reasonable deals and +Bargains is common, and opportunities for +Discounted offerings are appreciated. Managing to be +Frugal and delivering +Value-for-money is key in this cluster.
- In the For the **Organic** Cluster, users are preferring for +Natural and +Sustainable choices, preferences lean towards +Chemical-free, +Eco-friendly, and +Pure options. A strong emphasis is placed on +Non-GMO and +Locally sourced products, with an awareness of being +Environmentally conscious. The appeal lies in a commitment to +Farm-to-table practices, ensuring offerings are +Fresh, +Healthy, and +Whole. Prioritizing +Unprocessed and +Biorange choices.
- while some users ordered fine foods with expectations of a +Pungent, +Repugnant, and +Stale-free experience, unfortunately, the reality revealed an +Offensive, +Putrid, and +Displeasing aroma. The food turned out to be +Foul and +Stinky, leading to an overall +Unpleasant and +Unappetizing dining experience.
- Despite the initial promise of fine foods, the experience was tainted by a +Short shelf life, leaving limited room for +Freshness. The proximity to an +Imminent expiration and a +Close expiration date was evident, leading to concerns about a near-end expiration. The +Restricted storage time and the need for time-sensitive +Freshness added complexity to the situation in the **Expiration** cluster.

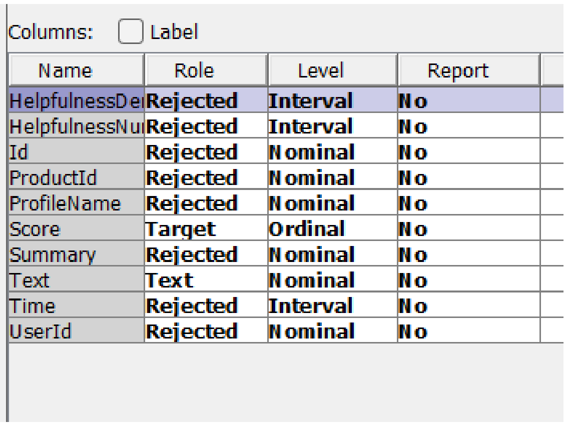
2.2 Supervised Learning Model

- Development of prediction or classification model requires a target variable.

2.2.1 Basic Model in Sentiment Analysis



- In the next phase of our analysis, we transitioned into supervised learning, employing a rule-based classifier to discern sentiment labels based on predefined conditions. The model construction process unfolded with the importation of data using the import node, setting the stage for a structured approach to sentiment classification.
- Variables included to forecast the positive & negative feedback of the product are: Score as a Target variable and Text as Text Variable.

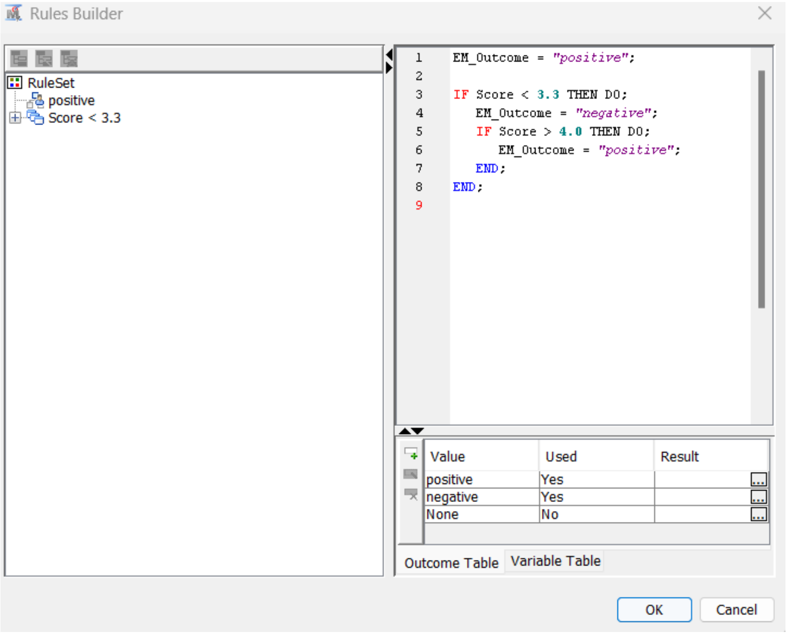


Columns: □ Label

Name	Role	Level	Report
HelpfulnessDe	Rejected	Interval	No
HelpfulnessNu	Rejected	Interval	No
Id	Rejected	Nominal	No
ProductId	Rejected	Nominal	No
ProfileName	Rejected	Nominal	No
Score	Target	Ordinal	No
Summary	Rejected	Nominal	No
Text	Text	Nominal	No
Time	Rejected	Interval	No
UserId	Rejected	Nominal	No

2.2.2 Rule Builder Classifier

This rule-based approach categorizes reviews into 'Negative' if the score is less than 3. 'Positive' if the score is greater than 4.0. The defined rules create a clear-cut framework for sentiment assignment, offering a straightforward yet effective method for classifying reviews based on their numerical scores. This would create a new column with name “EM_Outcome”.

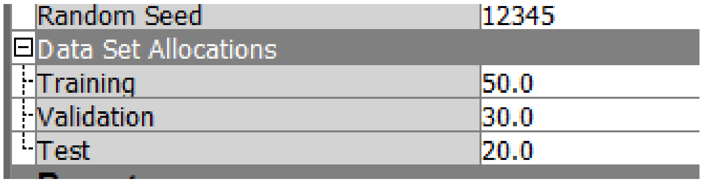


```
EM_Outcome = "positive";

IF Score < 3.3 THEN DO;
   EM_Outcome = "negative";
   IF Score > 4.0 THEN DO;
      EM_Outcome = "positive";
   END;
END;
```

Value	Used	Result
positive	Yes	
negative	Yes	
None	No	

2.2.3 Data Recoding and Partitioning

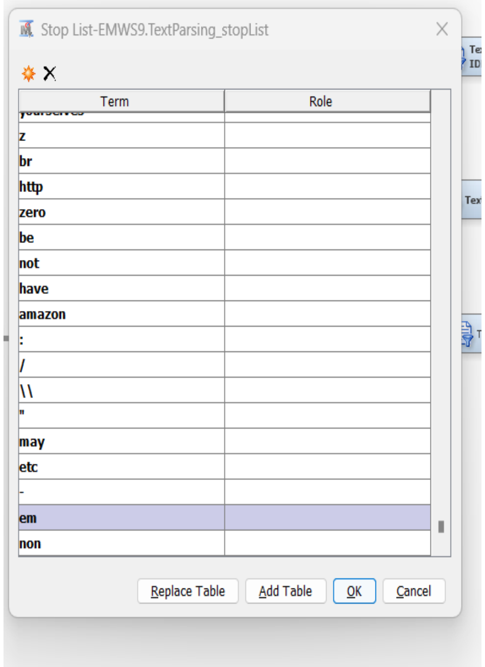
The recoded data, as performed in the above method, was subsequently saved into a new CSV file, setting the stage for the upcoming steps in our model development. Continuing from this point, the dataset was strategically partitioned into Training (50%), Validation (30%), and Test (20%) sets, forming a robust framework for the subsequent processes of training, fine-tuning, and evaluating our sentiment analysis model.



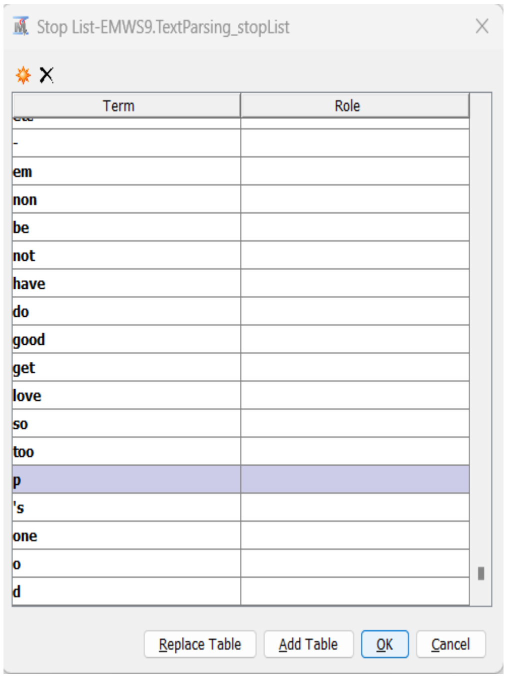
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	30.0
Test	20.0

2.2.4 Text Parsing

In the next phase of our analysis, the introduction of the Text Parsing node marked a significant refinement step. Configured to exclude specific parts of speech, including 'Auxiliary Verbs,' 'Conjunctions,' 'Determiners,' and 'Interjections,' the Text Parsing node enhanced the relevance of the textual data for sentiment analysis. By focusing on content-rich words, this step ensures that subsequent sentiment predictions are based on meaningful linguistic elements, optimizing the accuracy and depth of our sentiment analysis model.

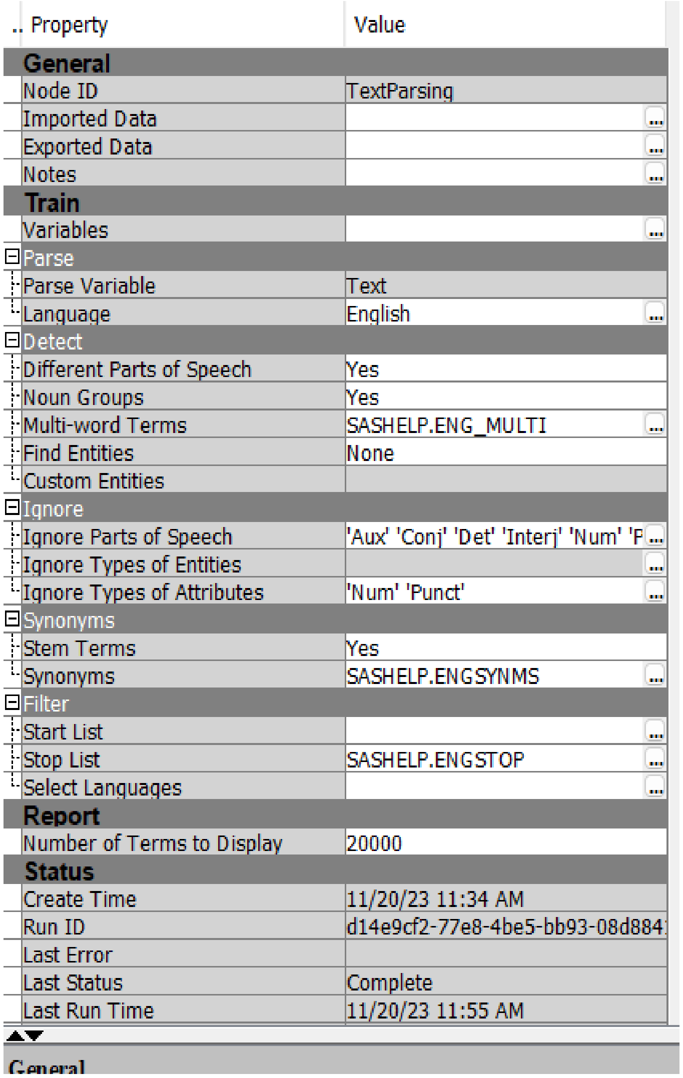


Term	Role
yourselves	
z	
br	
http	
zero	
be	
not	
have	
amazon	
:	
/	
\\	
"	
may	
etc	
-	
em	
non	



Term	Role
etc	
-	
em	
non	
be	
not	
have	
do	
good	
get	
love	
so	
too	
p	
's	
one	
o	
d	

The above are some of the words which we added in existing stop in the text parsing node.

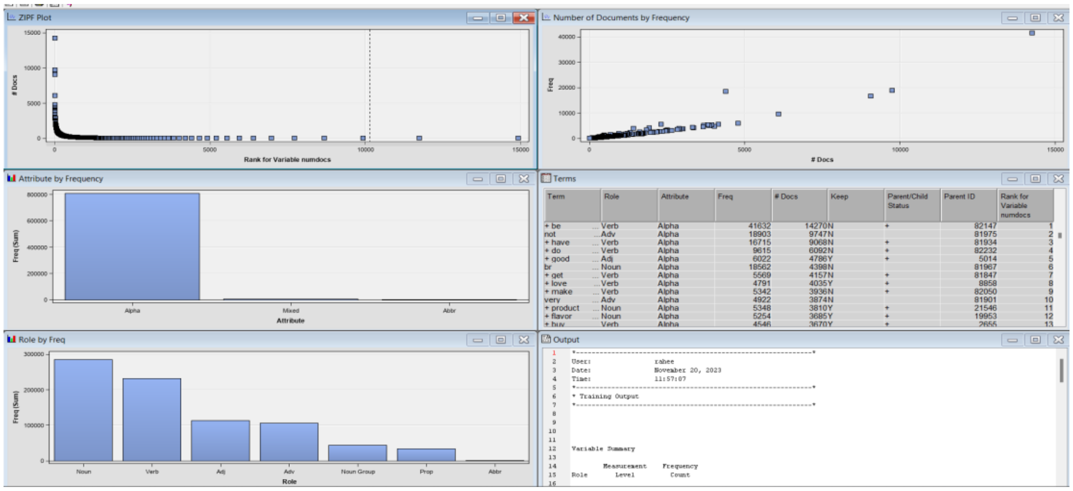


.. Property	Value
General	
Node ID	TextParsing
Imported Data	
Exported Data	
Notes	
Train	
Variables	
⊟ Parse	
Parse Variable	Text
Language	English
⊟ Detect	
Different Parts of Speech	Yes
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI
Find Entities	None
Custom Entities	
⊟ Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Interj' 'Num' 'P
Ignore Types of Entities	
Ignore Types of Attributes	'Num' 'Punct'
⊟ Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS
⊟ Filter	
Start List	
Stop List	SASHELP.ENGSTOP
Select Languages	
Report	
Number of Terms to Display	20000
Status	
Create Time	11/20/23 11:34 AM
Run ID	d14e9cf2-77e8-4be5-bb93-08d884
Last Error	
Last Status	Complete
Last Run Time	11/20/23 11:55 AM

General

We have utilized every default property in the text processing node with the exception of one, which is the stop list modification. We have manually added some stop list words to the existing table. We have included terms like be, not, have, do, good, get, using, br, http, :, /, \\, put, may, ok ,taste, ", amazon. The Zipf diagram helped me remember these terms.

2.2.4.1 Text parsing Results



Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ be	... Verb	Alpha	41632	14270	N	+	82147	1
not	... Adv	Alpha	18903	9747	N		81975	2
+ have	... Verb	Alpha	16715	9068	N	+	81934	3
+ do	... Verb	Alpha	9615	6092	N	+	82232	4
+ good	... Adj	Alpha	6022	4786	Y	+	5014	5
br	... Noun	Alpha	18562	4398	N		81967	6
+ get	... Verb	Alpha	5569	4157	N	+	81847	7
+ love	... Verb	Alpha	4791	4035	Y	+	8858	8
+ make	... Verb	Alpha	5342	3936	N	+	82050	9
very	... Adv	Alpha	4922	3874	N		81901	10
+ product	... Noun	Alpha	5348	3810	Y	+	21546	11
+ flavor	... Noun	Alpha	5254	3685	Y	+	19953	12
+ buy	Verb	Alpha	4546	3670	Y	+	2655	13

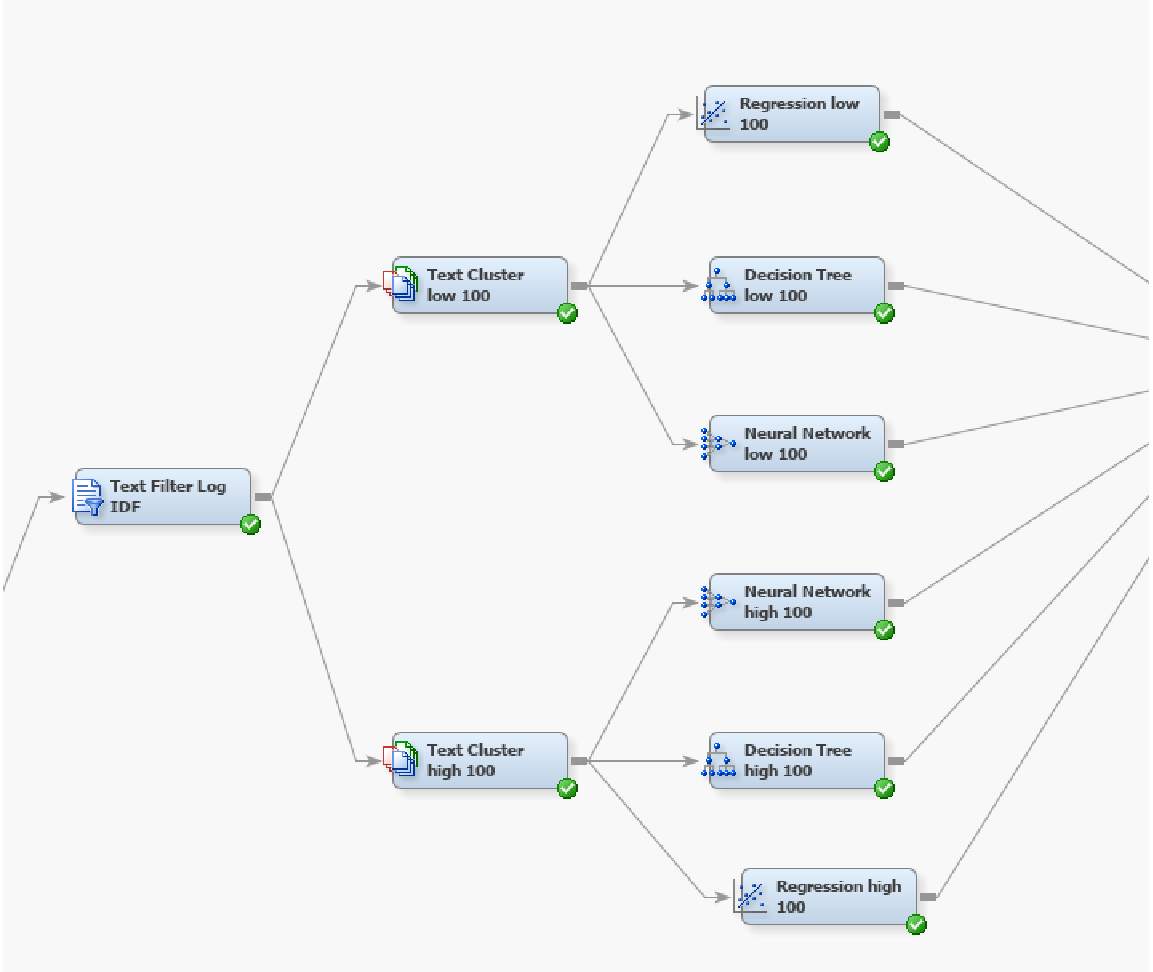
```
*------------------------------------------------------------*
User:                rahee
Date:                November 20, 2023
Time:                11:57:07
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

        Measurement    Frequency
Role       Level         Count

```

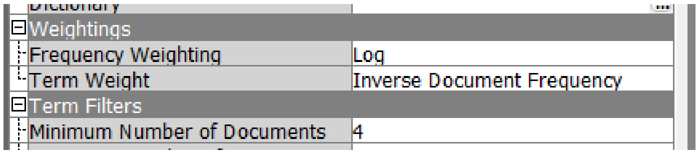
After conducting text parsing on all the documents, the SAS Enterprise Miner Workstation grouped the terms and plotted them based on their part of speech. This visualization organizes the terms into different categories corresponding to their respective parts of speech.

2.2.5 Text Filter

Upon evaluating the results of both the log and binary frequency weight settings, it was discovered that the Log frequency weight yielded the most favorable outcomes for the project. Consequently, a decision was made to uniformly apply the binary frequency weight setting across all models. This strategic choice significantly enhanced the performance and accuracy of the models, particularly in their capacity to effectively handle textual data.

2.2.5.1 Using frequency weight as Log & term weight as IDF

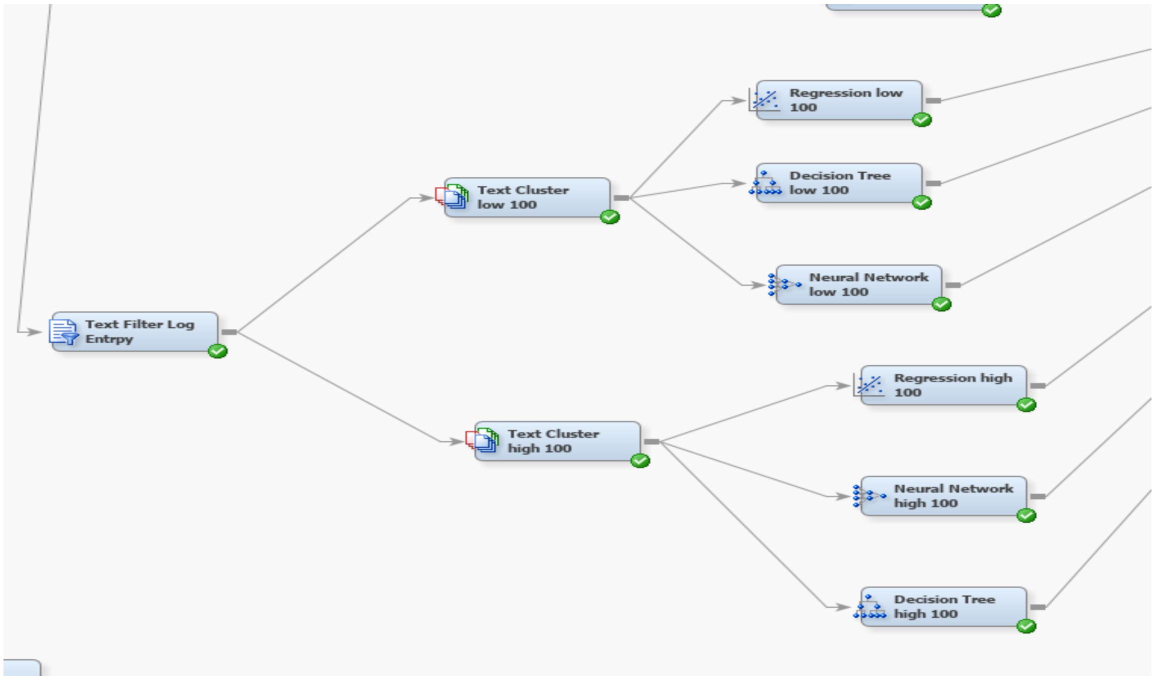


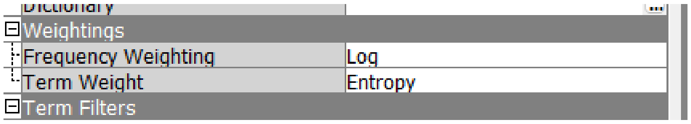


Dictionary	...
⊟Weightings	
Frequency Weighting	Log
Term Weight	Inverse Document Frequency
⊟Term Filters	
Minimum Number of Documents	4

2.2.5.2 Using frequency weight as Log & term weight as IDF & Entropy

We have maintained term weight as entropy in this iteration. Every other attribute has been maintained in its default state.

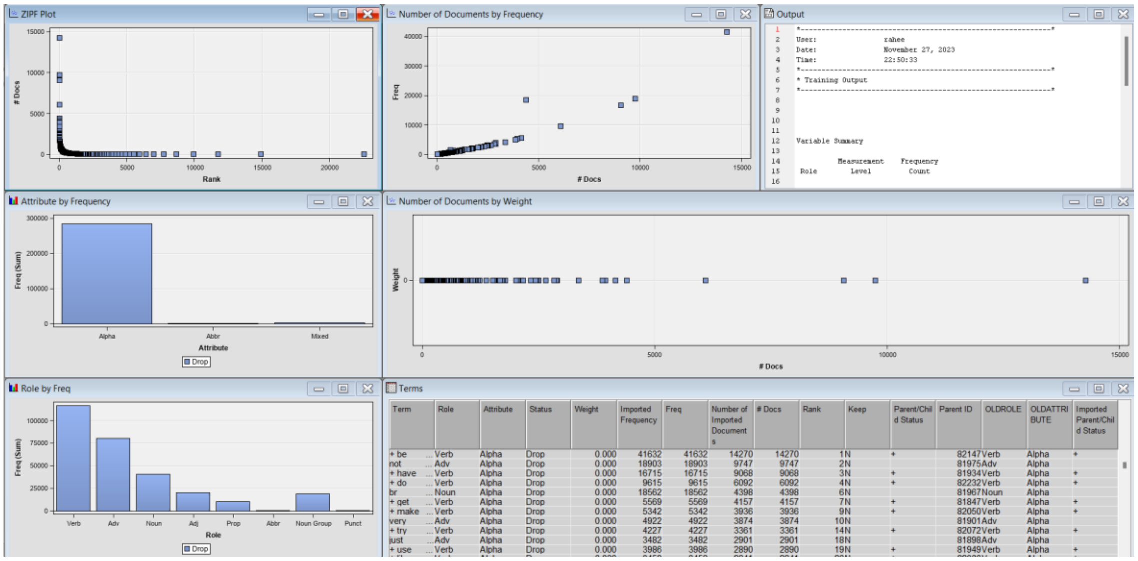




Dictionary	...
⊟ Weightings	
Frequency Weighting	Log
Term Weight	Entropy
⊟ Term Filters	

2.2.5.3 Text Filter Results

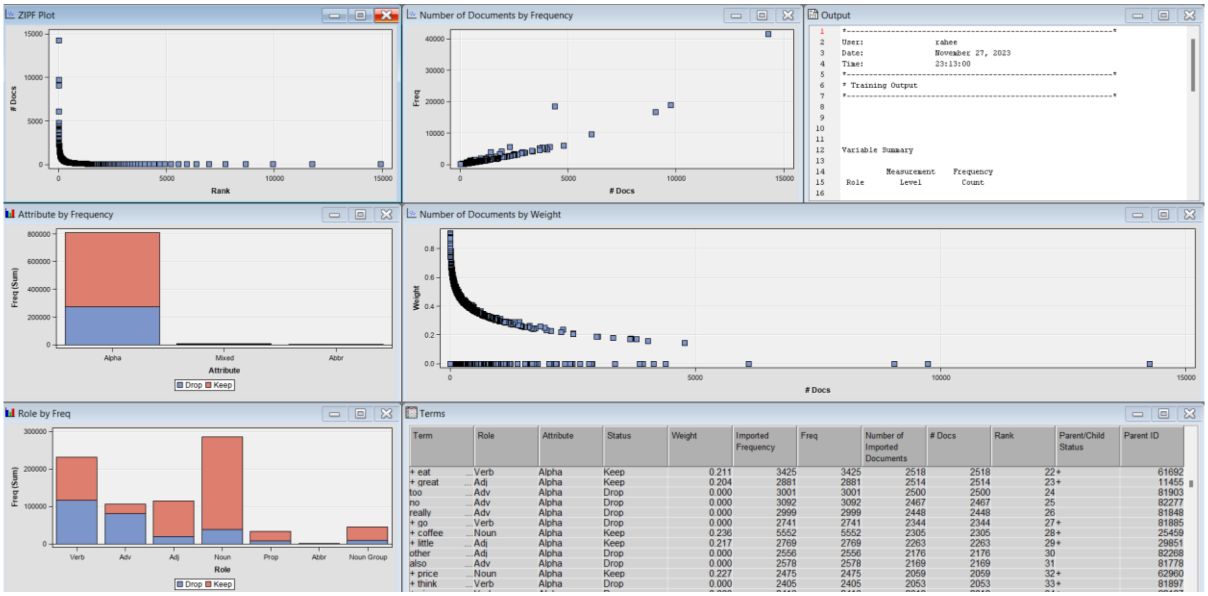
LOG-IDF



```
*------------------------------------------------------------*
User:                    rahee
Date:                    November 27, 2023
Time:                    22:50:33
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role         Level         Count

```

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Keep	Parent/Child Status	Parent ID	OLDROLE	OLDATTRIBUTE	Imported Parent/Child Status
+ be ...	Verb	Alpha	Drop	0.000	41632	41632	14270	14270	1	N	+	82147	Verb	Alpha	+
not ...	Adv	Alpha	Drop	0.000	18903	18903	9747	9747	2	N		81975	Adv	Alpha	
+ have ...	Verb	Alpha	Drop	0.000	16715	16715	9068	9068	3	N	+	81934	Verb	Alpha	+
+ do ...	Verb	Alpha	Drop	0.000	9615	9615	6092	6092	4	N	+	82232	Verb	Alpha	+
br ...	Noun	Alpha	Drop	0.000	18562	18562	4398	4398	6	N		81967	Noun	Alpha	
+ get ...	Verb	Alpha	Drop	0.000	5569	5569	4157	4157	7	N	+	81847	Verb	Alpha	+
+ make ...	Verb	Alpha	Drop	0.000	5342	5342	3936	3936	9	N	+	82050	Verb	Alpha	+
very ...	Adv	Alpha	Drop	0.000	4922	4922	3874	3874	10	N		81901	Adv	Alpha	
+ try ...	Verb	Alpha	Drop	0.000	4227	4227	3361	3361	14	N	+	82072	Verb	Alpha	+
just ...	Adv	Alpha	Drop	0.000	3482	3482	2901	2901	18	N		81898	Adv	Alpha	
+ use ...	Verb	Alpha	Drop	0.000	3986	3986	2890	2890	19	N	+	81949	Verb	Alpha	+

Log-Entropy



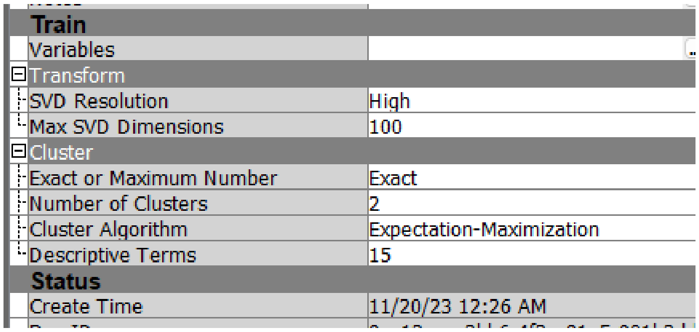
```
*------------------------------------------------------------*
User:                  rahee
Date:                  November 27, 2023
Time:                  23:13:00
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role        Level         Count

```

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Parent/Child Status	Parent ID
+ eat	...Verb	Alpha	Keep	0.211	3425	3425	2518	2518	22	+	61692
+ great	...Adj	Alpha	Keep	0.204	2881	2881	2514	2514	23	+	11455
too	...Adv	Alpha	Drop	0.000	3001	3001	2500	2500	24		81903
no	...Adv	Alpha	Drop	0.000	3092	3092	2467	2467	25		82277
really	...Adv	Alpha	Drop	0.000	2999	2999	2448	2448	26		81848
+ go	...Verb	Alpha	Drop	0.000	2741	2741	2344	2344	27	+	81885
+ coffee	...Noun	Alpha	Keep	0.236	5552	5552	2305	2305	28	+	25459
+ little	...Adj	Alpha	Keep	0.217	2769	2769	2263	2263	29	+	29851
other	...Adj	Alpha	Drop	0.000	2556	2556	2176	2176	30		82268
also	...Adv	Alpha	Drop	0.000	2578	2578	2169	2169	31		81778
+ price	...Noun	Alpha	Keep	0.227	2475	2475	2059	2059	32	+	62960
+ think	...Verb	Alpha	Drop	0.000	2405	2405	2053	2053	33	+	81897

- The notable change occurs in the 'Number of Documents by Weights' plot, where each setting exhibits varying weights depending on the respective range."
- The weighted frequencies and document counts contribute to a nuanced understanding of the relative significance of each term.

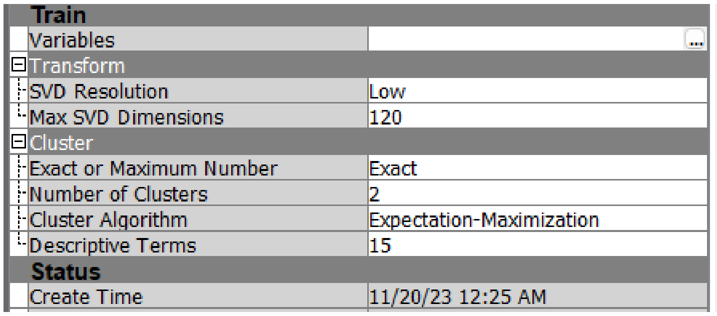
2.2.6 Cluster analysis

In the Text Cluster node, the resolution parameter defines the number of dimensions retained after applying SVD. In this case, the resolution was set with a low value of 100 and a high value of 100. This configuration influences the granularity of the latent structures identified through SVD, impacting the level of detail and complexity in the resulting clusters.



Train	
Variables	...
⊟ Transform	
SVD Resolution	High
Max SVD Dimensions	100
⊟ Cluster	
Exact or Maximum Number	Exact
Number of Clusters	2
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15
Status	
Create Time	11/20/23 12:26 AM

“We experimented with an SVD value of 120; however, this resulted in a degradation of model performance, particularly evident during comparisons with lower ROC values.”



Train	
Variables	...
⊟Transform	
SVD Resolution	Low
Max SVD Dimensions	120
⊟Cluster	
Exact or Maximum Number	Exact
Number of Clusters	2
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15
Status	
Create Time	11/20/23 12:25 AM

2.2.6.1 Text cluster Result

- In configuring our model, we elected to set the cluster number precisely to 2, aligning with the binary classification of the target variable into positive and negative categories. This decision aimed at effectively grouping the text documents into two distinct clusters based on their content and sentiment.
- Given the substantial size of our dataset, comprising 568,454 documents, we established the maximum Singular Value Decomposition (SVD) dimensions at 100. This strategic choice facilitated dimensionality reduction while retaining a meaningful amount of information. The SVD dimensions played a pivotal role in extracting pertinent features from the text data, thereby enhancing the efficacy of document clustering.
- Furthermore, our exploration of various SVD resolutions involved testing three settings: low, mid, and high. This comprehensive analysis sought to determine the optimal SVD resolution for our model, a critical step in refining the clustering process.
- In end, we have narrowed down our filter settings to just two options: low and high. This decision stems from the observation that the high and mid filters yield identical results.

Log – IDF		
SVD - 100	Model	ROC
Low	Decision	0.706
	Regression	0.833
	Neural network	0.84
High	Decision	0.696
	Regression	0.845
	Neural network	0.847

Log - Entropy		
SVD - 100	Model	ROC
Low	Decision	0.649
	Regression	0.834
	Neural network	0.841
High	Decision	0.695
	Regression	0.848
	Neural network	0.849

The Best Model from this model comparison comes out to be Neural network model with 84.9% test ROC index.

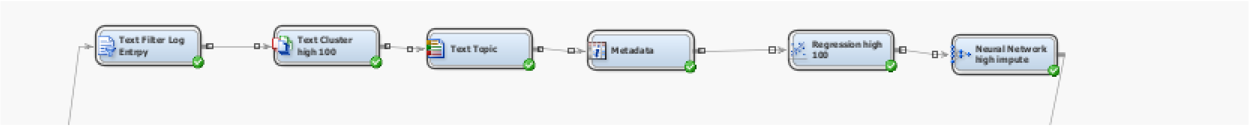
2.2.7 Test Misclassification rate

After thorough analysis, we have determined that the Regression model with a high SVD resolution is optimal for our specific requirements. In the selection of term weight methods, namely Entropy and IDF, a tie emerged. To break this tie and identify the most suitable term weight for our final model, we employed the test misclassification rate as the decisive criterion.

Term weight	Test Misclassification Rate
IDF	0.1696
Entropy	0.1659

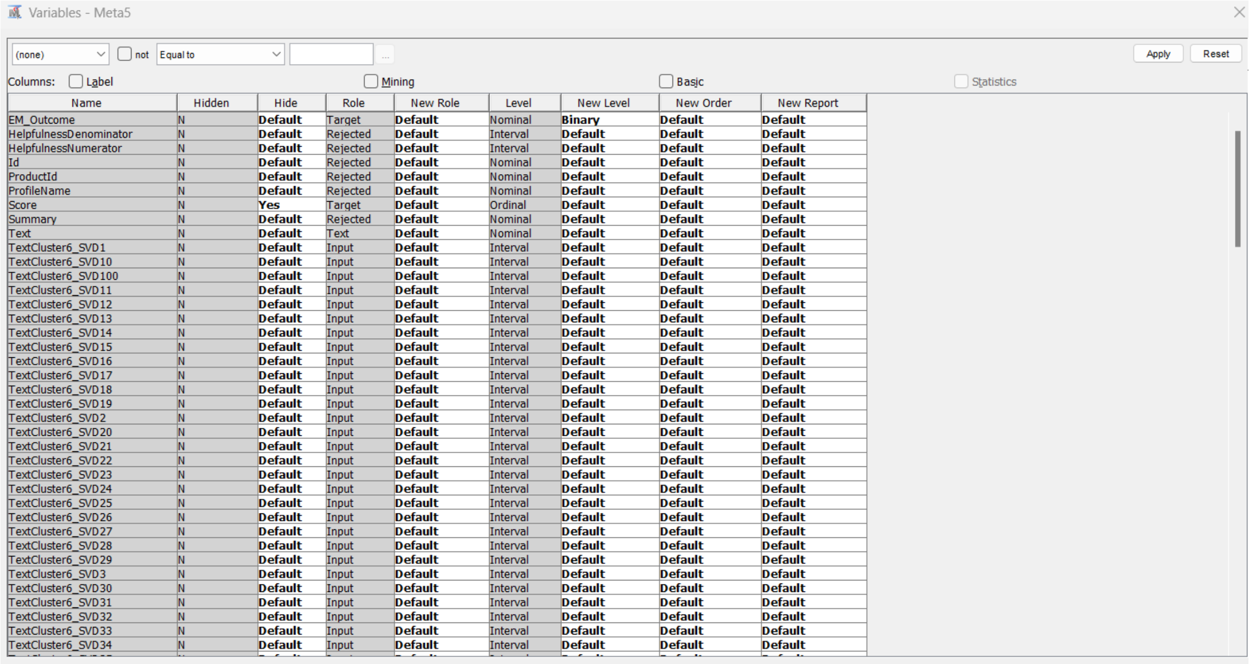
- Upon reviewing the test misclassification rates in the above table, it is evident that Entropy performs most effectively for our model, yielding a misclassification rate of 16.5% on the testing data.
- The ultimate configuration for our model includes Frequency weight set as Log, Term weight as Entropy, SVD dimensions at 100, and SVD resolution set to high. Subsequently, we will apply this identical configuration to both the Interpretable model, incorporating a neural network attached to the regression model. This approach aims to further explore and enhance the capabilities of our model.

2.2.8 Interpretable Model



- Utilizing the optimal configuration from our previous discussions, the text filter employs Log as the frequency weight and Entropy as the term weight. Meanwhile, the text cluster settings include a cluster number of 2, SVD dimension set to 100, and SVD resolution configured as High.
- For the text topic node, we maintain 20 multi-term topics with the same configuration as discussed in the interpretable model. Additionally, a neural model has been incorporated into the Regression model node to enhance and explore further capabilities of the overall model.

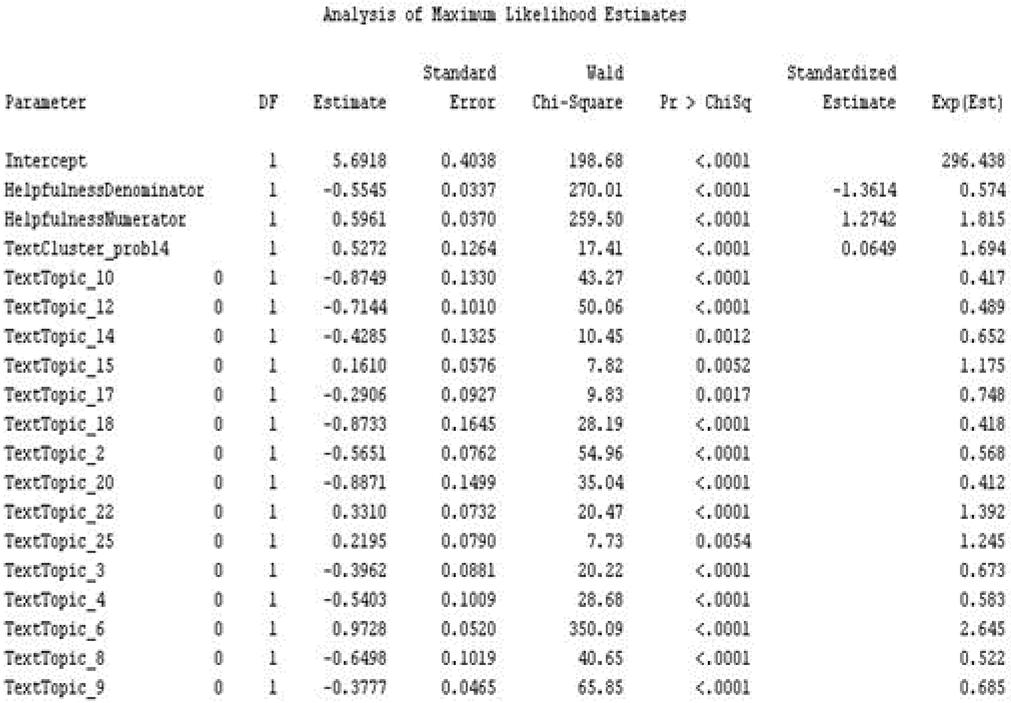
The provided screenshot displays the "Metadata" node, encompassing all variables acquired from the preceding node.



Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report
EM_Outcome	N	Default	Target	Default	Nominal	Binary	Default	Default
HelpfulnessDenominator	N	Default	Rejected	Default	Interval	Default	Default	Default
HelpfulnessNumerator	N	Default	Rejected	Default	Interval	Default	Default	Default
Id	N	Default	Rejected	Default	Nominal	Default	Default	Default
ProductId	N	Default	Rejected	Default	Nominal	Default	Default	Default
ProfileName	N	Default	Rejected	Default	Nominal	Default	Default	Default
Score	N	Yes	Target	Default	Ordinal	Default	Default	Default
Summary	N	Default	Rejected	Default	Nominal	Default	Default	Default
Text	N	Default	Text	Default	Nominal	Default	Default	Default
TextCluster6_SVD1	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD10	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD100	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD11	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD12	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD13	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD14	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD15	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD16	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD17	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD18	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD19	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD2	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD20	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD21	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD22	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD23	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD24	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD25	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD26	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD27	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD28	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD29	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD3	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD30	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD31	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD32	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD33	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster6_SVD34	N	Default	Input	Default	Interval	Default	Default	Default

We modified the variables of the "train" dataset in the metadata node by excluding TextCluster_SVD1 through TextCluster_SVD100 and TextTopic_Raw1 through TextTopic_Raw25. Instead, we opted for TextCluster6_cluster as the input variable. These adjustments were implemented to streamline our model's input variables, focusing on the most crucial ones. We believe these changes have contributed to an enhanced accuracy and efficacy in predicting the target variable.

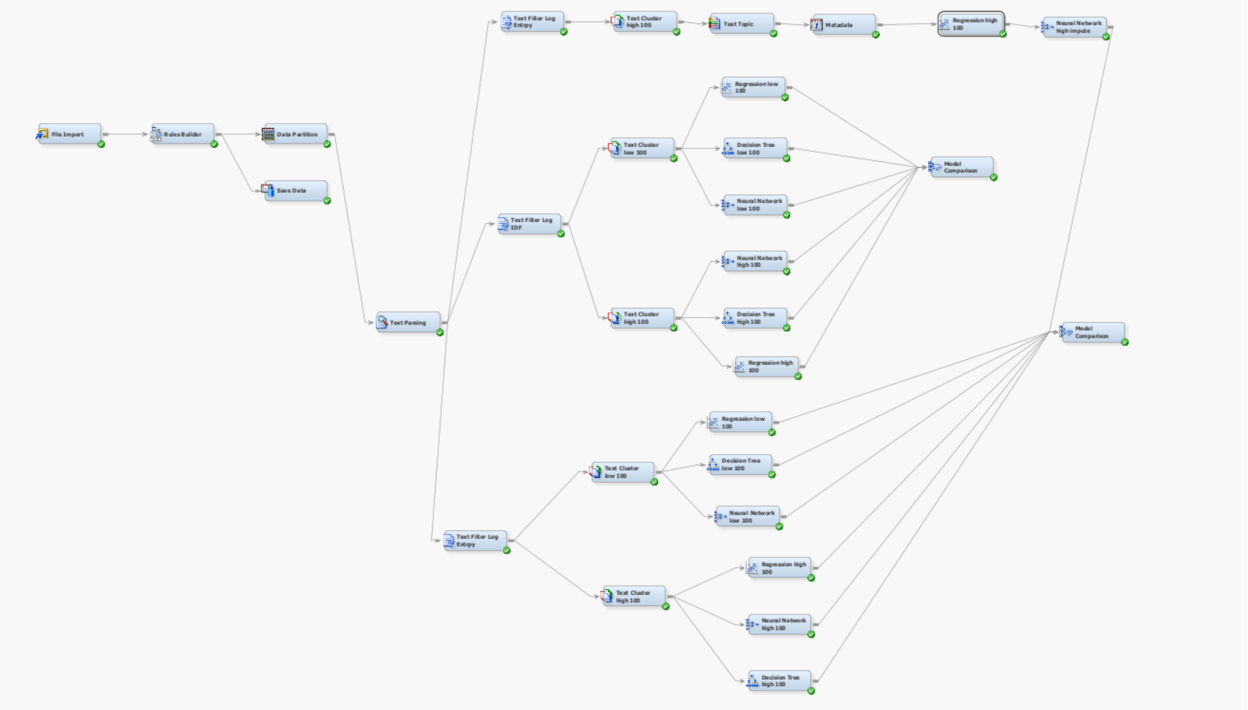
Maximum Likelihood Estimation (MLE) is a statistical technique employed to estimate model parameters by maximizing the likelihood function. Widely utilized in regression analysis, MLE is instrumental in determining the coefficients that optimally align with the provided data.
The Maximum likelihood estimates are obtained for the Regression model are shown below:



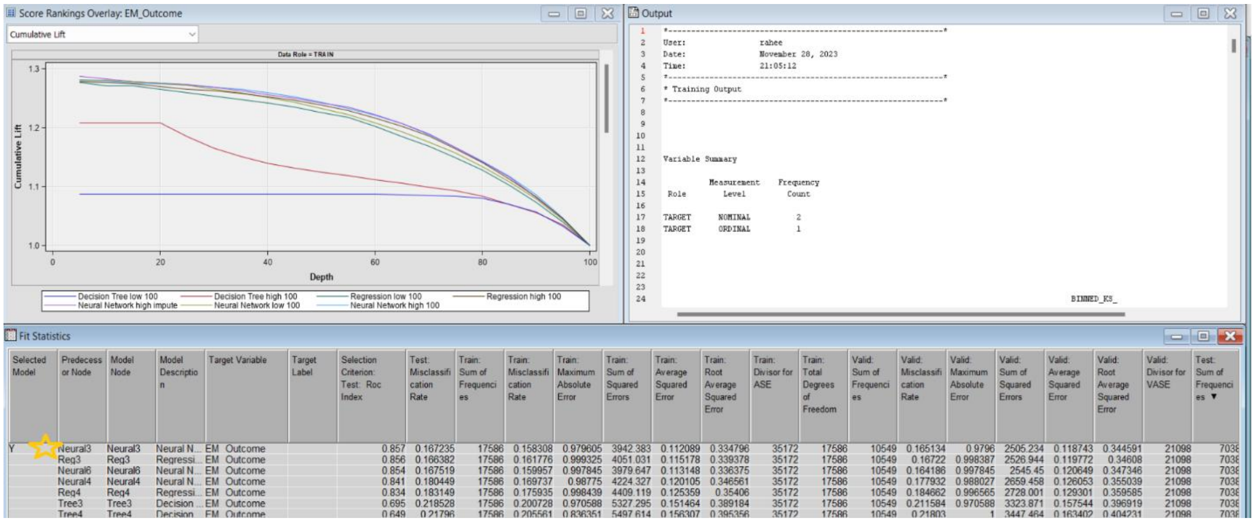
```
                                  Analysis of Maximum Likelihood Estimates

                                                Standard          Wald                       Standardized
Parameter                   DF     Estimate        Error    Chi-Square    Pr > ChiSq             Estimate     Exp(Est)

Intercept                    1       5.6918       0.4038        198.68        <.0001                           296.438
HelpfulnessDenominator       1      -0.5545       0.0337        270.01        <.0001              -1.3614        0.574
HelpfulnessNumerator         1       0.5961       0.0370        259.50        <.0001               1.2742        1.815
TextCluster_prob14           1       0.5272       0.1264         17.41        <.0001               0.0649        1.694
TextTopic_10            0    1      -0.8749       0.1330         43.27        <.0001                             0.417
TextTopic_12            0    1      -0.7144       0.1010         50.06        <.0001                             0.489
TextTopic_14            0    1      -0.4285       0.1325         10.45        0.0012                             0.652
TextTopic_15            0    1       0.1610       0.0576          7.82        0.0052                             1.175
TextTopic_17            0    1      -0.2906       0.0927          9.83        0.0017                             0.748
TextTopic_18            0    1      -0.8733       0.1645         28.19        <.0001                             0.418
TextTopic_2             0    1      -0.5651       0.0762         54.96        <.0001                             0.568
TextTopic_20            0    1      -0.8871       0.1499         35.04        <.0001                             0.412
TextTopic_22            0    1       0.3310       0.0732         20.47        <.0001                             1.392
TextTopic_25            0    1       0.2195       0.0790          7.73        0.0054                             1.245
TextTopic_3             0    1      -0.3962       0.0881         20.22        <.0001                             0.673
TextTopic_4             0    1      -0.5403       0.1009         28.68        <.0001                             0.583
TextTopic_6             0    1       0.9728       0.0520        350.09        <.0001                             2.645
TextTopic_8             0    1      -0.6498       0.1019         40.65        <.0001                             0.522
TextTopic_9             0    1      -0.3777       0.0465         65.85        <.0001                             0.685
```

2.3 Final Model

All the nodes connected have the same settings that previously ran for the Log – Entropy model. The diagram for sentiment analysis is given:



3. Model Comparison



```
*------------------------------------------------------------*
User:               rahee
Date:               November 28, 2023
Time:               21:05:12
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

         Measurement    Frequency
 Role       Level         Count

TARGET     NOMINAL          2
TARGET     ORDINAL          1






                                                          BINNED_KS_
```

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Test: Roc Index	Test: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE	Test: Sum of Frequencies ▼
Y	Neural3	Neural3	Neural N...	EM Outcome		0.857	0.167235	17586	0.158308	0.979605	3942.383	0.112089	0.334796	35172	17586	10549	0.165134	0.9796	2505.234	0.118743	0.344591	21098	7038
	Reg3	Reg3	Regressi...	EM Outcome		0.856	0.166382	17586	0.161776	0.999325	4051.031	0.115178	0.339378	35172	17586	10549	0.16722	0.998387	2526.944	0.119772	0.34608	21098	7038
	Neural6	Neural6	Neural N...	EM Outcome		0.854	0.167519	17586	0.159957	0.997845	3979.647	0.113148	0.336375	35172	17586	10549	0.164186	0.997845	2545.45	0.120649	0.347346	21098	7038
	Neural4	Neural4	Neural N...	EM Outcome		0.841	0.180449	17586	0.169737	0.98775	4224.327	0.120105	0.346561	35172	17586	10549	0.177932	0.988027	2659.458	0.126053	0.355039	21098	7038
	Reg4	Reg4	Regressi...	EM Outcome		0.834	0.183149	17586	0.175935	0.998439	4409.119	0.125359	0.35406	35172	17586	10549	0.184662	0.996565	2728.001	0.129301	0.359585	21098	7038
	Tree3	Tree3	Decision ...	EM Outcome		0.695	0.218528	17586	0.200728	0.970588	5327.295	0.151464	0.389184	35172	17586	10549	0.211584	0.970588	3323.871	0.157544	0.396919	21098	7038
	Tree4	Tree4	Decision	EM Outcome		0.649	0.21796	17586	0.205561	0.836351	5497.614	0.156307	0.395356	35172	17586	10549	0.21803	1	3447.464	0.163402	0.404231	21098	7038

From the final model discussed above, it is evident that the Test ROC stands at 0.857, surpassing the results of the various preceding models. However, the misclassification rate remains consistent at 0.16, equivalent to 16%, mirroring the outcome of the previous model. Despite this, the improved ROC accuracy in the interpreted model signifies progress in our model performance.

4. Findings and Business Insights

Through our comprehensive analysis of the Amazon Fine Food reviews dataset, several key findings and business insights have emerged. The sentiment analysis revealed that terms such as "good," "love," and "product" are frequently mentioned, with varying weights assigned to signify their importance in expressing sentiments. The identification of latent structures using Singular Value Decomposition (SVD) in the Text Cluster node provided a nuanced understanding of patterns and themes within the reviews.

The categorization of reviews into positive and negative sentiments, coupled with the analysis of specific terms and their weighted frequencies, unveils valuable insights into customer perceptions.

4.1 Business Insights

Leveraging Positive Sentiments:

Businesses can capitalize on the dominance of positive sentiments by incorporating terms like "good" and "love" in marketing materials. Emphasizing positive customer experiences can enhance brand perception.

Addressing Negative Sentiment Pain Points:

The identification of negative sentiment pain points related to smell and expiration provides a roadmap for quality improvement. Addressing these concerns through enhanced quality control and product freshness management is crucial.

Communication Strategy Enhancement:

Crafting communication strategies that align with positive sentiments can strengthen brand-customer relationships. Highlighting positive aspects of products, such as quality and love from customers, can positively influence perceptions.

Optimizing Product Lifecycle Management:

Insights into terms related to expiration underscore the importance of managing product lifecycles effectively. Clear communication about expiration dates and proactive management of freshness can build trust with customers.

Customer Engagement and Feedback Response:

Engaging with customers who express negative sentiments provides an opportunity to address concerns directly. Responding to feedback demonstrates a commitment to customer satisfaction and can turn negative experiences into positive ones.

4.2 Recommendations

- Invest in continuous quality improvement to address concerns related to product quality, including smells and expiration issues.
- Implement clear and proactive communication about product expiration dates to enhance customer transparency and trust.
- Develop a robust customer engagement strategy to actively respond to customer feedback, turning negative sentiments into positive experiences.
- Optimize marketing messaging by strategically incorporating positive terms like "good" and "love" based on their weighted frequencies to reinforce positive brand associations.
- Explore more sophisticated machine learning techniques and refine the sentiment analysis model to gain deeper insights into customer sentiments and preferences.

5. Conclusion

In summary, the analysis of Amazon Fine Food reviews reveals a predominance of positive sentiments, particularly associated with terms like "good" and "love," indicative of overall customer satisfaction. However, identified negative sentiments related to product quality, such as smells and expiration concerns, underscore areas for improvement. Leveraging a combination of rule-based sentiment classification, dual filtering with Log IDF and Entropy, and Singular Value Decomposition in text clustering has provided a nuanced understanding of customer feedback. Recommendations emphasize continuous quality enhancement, proactive communication, and a robust customer engagement strategy. These steps, along with optimizing marketing messaging and refining the sentiment analysis model, are key to addressing concerns, building positive brand associations, and fostering sustained growth in alignment with customer expectations.

6. References

- Data source:
  https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

- Text Analytics Using SAS Text Miner Course Notes