OPIM 5671: Text Mining

Categorization of Resumes for Enhanced Job Matching

Team 5

Lahari Maddula

Pradeepti Dokka

Sai Deepika Bandari

Sanchita Godse

Shuang Ma

Table of Contents

Executive Summary

The project aims to transform the recruitment process by implementing an automated text mining technology that efficiently categorizes resumes into specific job segments. This innovative program intends to enhance applicant-job matching, significantly reduce the time and resources currently allocated to manual resume screening, and elevate the overall candidate job search experience. Additionally, it promises to yield valuable labor market insights.

The research progressed through distinct stages, beginning with the setup and collection of data from livecareer.com, encompassing data preprocessing, labeling, model creation, and subsequent evaluation. The initial phase involved configuring the environment and procuring data, which included extracting text from PDFs, followed by rigorous cleaning, preparation, and manual data labeling to prepare it for model training. Subsequently, the focus shifted to constructing a robust text mining model, involving the exploration of various techniques and the training of the model with optimized hyperparameters. Finally, the evaluation process encompassed an in-depth review of the model's performance, utilizing SAS Enterprise Miner for error analysis.

The **Neural Network model incorporating text topic** analysis exhibited exceptional performance, demonstrating a strong fit with a ROC Index of 0.99 and Misclassification values at 0.28. This translates to a noteworthy 72% accuracy rate for the model. The **interpretation model** identified key resume keywords for distinct categories; for instance, **HR category resumes** should contain terms like **compensation, resource, HR experience, and recruitment**, while **Designer category resumes** should emphasize keywords like **graphic, design, and adobe**.

The successful culmination of this project aims to deliver a robust text mining solution, streamlining the hiring process and facilitating a data-driven approach to human resource management.

1. Introduction:

Background

In today's competitive recruitment landscape, companies are constantly seeking ways to streamline their processes and identify the best candidates more efficiently. Traditional resume screening methods, which often involve a manual review of each resume, can be time-consuming and labor-intensive. Additionally, these methods can be biased, as recruiters may unconsciously favor certain candidates over others based on gender, race, or age.

Text Mining as a Solution

Text mining offers a promising solution to these challenges by automating the process of resume categorization. By using natural language processing (NLP) techniques, text mining algorithms can extract and analyze key information from resumes, such as skills, experience, and education. This information can then be used to categorize resumes into predefined job categories automatically.

We used the SAS Enterprise Miner Workstation 15.1 to create this project because it's a powerful tool for text mining projects. It has a lot of features that make it easy to import, clean, analyze, model, and evaluate text data.

Data Description

Raw Data:

Variable	Variable Name	Description	Type
Variable 1	Resume_str	The resume content in plain text format.	Object
Variable 2	ID	A unique identifier for each resume, also serving as the	Numerical

		filename for the corresponding PDF.	
Variable 3	Resume_html	The resume content in HTML format as obtained from web scraping.	Object
Variable 4	Category	The job category for which the resume was submitted, with present categories including HR, IT, Education, Legal, and more.	Object

A snapshot of how the data looks like,

A	B	C	D	E	F	G
ID	Resume_str	Resume_html	Category			
16852973	HR ADMINISTRATOR/MARKETING ASSOCIATE	<div class="fontsize fontface vmargins	HR			
22323967	HR SPECIALIST, US HR OPERATIONS	<div class="fontsize fontface vmargins	HR			
33176873	HR DIRECTOR Summary Over 20 years	<div class="fontsize fontface vmargins	HR			
27018550	HR SPECIALIST Summary Dedicated,	<div class="fontsize fontface vmargins	HR			
17812897	HR MANAGER Skill Highlights HR	<div class="fontsize fontface vmargins	HR			
11592605	HR GENERALIST Summary Dedicated	<div class="fontsize fontface vmargins	HR			
25824789	HR MANAGER Summary HUMAN	<div class="fontsize fontface vmargins	HR			
15375009	HR MANAGER Professional Summary	<div class="fontsize fontface vmargins	HR			
11847784	HR SPECIALIST Summary Possess 15+ yea	<div class="fontsize fontface vmargins hmargi	HR			
32896934	HR CLERK Summary Translates business	<div class="fontsize fontface vmargins	HR			
29149998	HR ASSISTANT Summary Highly motivatec	<div class="fontsize fontface vmargins hmargi	HR			
11480899	HR MANAGER Summary Human Resou	<div class="fontsize fontface vmargins hmargi	HR			
23155093	HR MANAGER Summary To obtain a	<div class="fontsize fontface vmargins	HR			
11763983	HR GENERALIST Summary A people-orier	<div class="fontsize fontface vmargins hmargi	HR			
27490876	HR COORDINATOR Summary Applicant S	<div class="fontsize fontface vmargins hmargi	HR			
32977530	HR CLERK Summary I am an ethical, team	<div class="fontsize fontface vmargins hmargi	HR			
93002334	HR ANALYST Summary Experienced pro	<div class="fontsize fontface vmargins hmargi	HR			
24184357	HR DIRECTOR Summary Human	<div class="fontsize fontface vmargins	HR			
73077810	HR GENERALIST/RECRUITER Summary	<div class="fontsize fontface vmargins	HR			

39970711	HR & SAFETY MANAGER Summary	<div class="fontsize fontface vmargins	HR			
20806155	HR SPECIALIST (INFORMATION SYSTEMS)	<div class="fontsize fontface vmargins				
r Process Improvement	Human Resources	Process Improvement	Process Ir	Proposals	Solutions	Training
28640735	DIRECTOR OF HR Executive Profile	<div class="fontsize fontface vmargins	HR			
15575117	HR SENIOR SPECIALIST Career Overview	<div class="fontsize fontface vmargins	HR			
27496514	HR CUSTOMER SERVICE REPRESENTATIVE	<div class="fontsize fontface vmargins	HR			
14256329	HR SERVICES REPRESENTATIVE Summary	<div class="fontsize fontface vmargins	HR			
19336728	HR ASSISTANT INTERN Summary New gra	<div class="fontsize fontface vmargins hmargi	HR			
10694288	HR BENEFITS/LEAVE COORDINATOR	<div class="fontsize fontface vmargins	HR			
28175164	REGIONAL HR BUSINESS PARTNER Human F	<div class="fontsize fontface vmargins hmargi	HR			
10399912	HR PERSONNEL ASSISTANT Summary I ai	<div class="fontsize fontface vmargins	HR			
20417897	EXECUTIVE ASSISTANT HR Summary Skillf	<div class="fontsize fontface vmargins	HR			

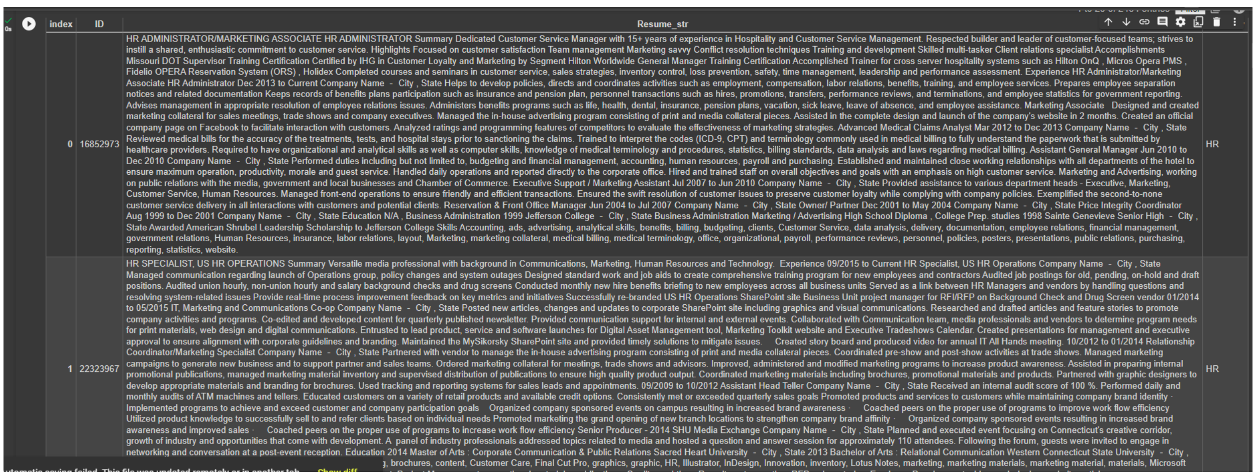
Our data originally has 16 different categories in the category variable such as HR, ADVOCATE, BUSINESS-DEVELOPMENT, CONSULTANT, DESIGNER, DIGITAL-MEDIA,

FITNESS, HEALTHCARE, INFORMATION TECHNOLOGY, SALES, TEACHER, BUDGETING, CHEF, BPO, AGRICULTURE, AUTOMOBILE.

Cleaning Dataset:

We had to clean our dataset as we found that we no longer require the column ‘Resume_html’ as it is a duplicate version of the already existing column but written using the html language. As this is something we won’t use in our dataset we removed the column using python along with cleaning the dataset where the ID column had unnecessary text values which are unexpected.

This is how our cleaned dataset looks like,



index	ID	Resume_str	
0	16852973	HR ADMINISTRATOR/MARKETING ASSOCIATE HR ADMINISTRATOR Summary Dedicated Customer Service Manager with 15+ years of experience in Hospitality and Customer Service Management. Respected builder and leader of customer-focused teams; strives to instill a shared, enthusiastic commitment to customer service. Highlights Focused on customer satisfaction Team management Marketing savvy Conflict resolution techniques Training and development Skilled multi-tasker Client relations specialist Accomplishments Missouri DOT Supervisor Training Certification Certified by IHG in Customer Loyalty and Marketing by Segment Hilton Worldwide General Manager Training Certification Accomplished Trainer for cross server hospitality systems such as Hilton OnQ , Micros Opera PMS , Fidelio OPERA Reservation System (ORS) , Holidex Completed courses and seminars in customer service, sales strategies, inventory control, loss prevention, safety, time management, leadership and performance assessment. Experience HR Administrator/Marketing Associate HR Administrator Dec 2013 to Current Company Name - City , State Helps to develop policies, directs and coordinates activities such as employment, compensation, labor relations, benefits, training, and employee services. Prepares employee separation notices and related documentation Keeps records of benefits plans participation such as insurance and pension plan, personnel transactions such as hires, promotions, transfers, performance reviews, and terminations, and employee statistics for government reporting. Advises management in appropriate resolution of employee relations issues. Administers benefits programs such as life, health, dental, insurance, pension plans, vacation, sick leave, leave of absence, and employee assistance. Marketing Associate Designed and created marketing collateral for sales meetings, trade shows and company executives. Managed the in-house advertising program consisting of print and media collateral pieces. Assisted in the complete design and launch of the company's website in 2 months. Created an official company page on Facebook to facilitate interaction with customers. Analyzed ratings and programming features of competitors to evaluate the effectiveness of marketing strategies. Advanced Medical Claims Analyst Mar 2012 to Dec 2013 Company Name - City , State Reviewed medical bills for the accuracy of the treatments, tests, and hospital stays prior to sanctioning the claims. Trained to interpret the codes (ICD-9, CPT) and terminology commonly used in medical billing to fully understand the paperwork that is submitted by healthcare providers. Required to have organizational and analytical skills as well as computer skills, knowledge of medical terminology and procedures, statistics, billing standards, data analysis and laws regarding medical billing. Assistant General Manager Jun 2010 to Dec 2010 Company Name - City , State Performed duties including but not limited to, budgeting and financial management, accounting, human resources, payroll and purchasing. Established and maintained close working relationships with all departments of the hotel to ensure maximum operation, productivity, morale and guest service. Handled daily operations and reported directly to the corporate office. Hired and trained staff on overall objectives and goals with an emphasis on high customer service. Marketing and Advertising, working on public relations with the media, government and local businesses and Chamber of Commerce. Executive Support / Marketing Assistant Jul 2007 to Jun 2010 Company Name - City , State Provided assistance to various department heads - Executive, Marketing, Customer Service, Human Resources. Managed front-end operations to ensure friendly and efficient transactions. Ensured the swift resolution of customer issues to preserve customer loyalty while complying with company policies. Exemplified the second-to-none customer service delivery in all interactions with customers and potential clients. Reservation & Front Office Manager Jun 2004 to Jul 2007 Company Name - City , State Owner/ Partner Dec 2001 to May 2004 Company Name - City , State Price Integrity Coordinator Aug 1999 to Dec 2001 Company Name - City , State Education N/A , Business Administration 1999 Jefferson College - City , State Business Administration Marketing / Advertising High School Diploma , College Prep. studies 1998 Sainte Genevieve Senior High - City , State Awarded American Shrubel Leadership Scholarship to Jefferson College Skills Accounting, ads, advertising, analytical skills, benefits, billing, budgeting, clients, Customer Service, data analysis, delivery, documentation, employee relations, financial management, government relations, Human Resources, insurance, labor relations, layout, Marketing, marketing collateral, medical billing, medical terminology, office, organizational, payroll, performance reviews, personnel, policies, posters, presentations, public relations, purchasing, reporting, statistics, website.	HR
1	22323967	HR SPECIALIST, US HR OPERATIONS Summary Versatile media professional with background in Communications, Marketing, Human Resources and Technology. Experience 09/2015 to Current HR Specialist, US HR Operations Company Name - City , State Managed communication regarding launch of Operations group, policy changes and system outages Designed standard work and job aids to create comprehensive training program for new employees and contractors Audited job postings for old, pending, on-hold and draft positions. Audited union hourly, non-union hourly and salary background checks and drug screens Conducted monthly new hire benefits briefing to new employees across all business units Served as a link between HR Managers and vendors by handling questions and resolving system-related issues Provide real-time process improvement feedback on key metrics and initiatives Successfully re-branded US HR Operations SharePoint site Business Unit project manager for RFI/RFP on Background Check and Drug Screen vendor 01/2014 to 05/2015 IT, Marketing and Communications Co-op Company Name - City , State Posted new articles, changes and updates to corporate SharePoint site including graphics and visual communications. Researched and drafted articles and feature stories to promote company activities and programs. Co-edited and developed content for quarterly published newsletter. Provided communication support for internal and external events. Collaborated with Communication team, media professionals and vendors to determine program needs for print materials, web design and digital communications. Entrusted to lead product, service and software launches for Digital Asset Management tool, Marketing Toolkit website and Executive Tradeshows Calendar. Created presentations for management and executive approval to ensure alignment with corporate guidelines and branding. Maintained the MySikorsky SharePoint site and provided timely solutions to mitigate issues. Created story board and produced video for annual IT All Hands meeting. 10/2012 to 01/2014 Relationship Coordinator/Marketing Specialist Company Name - City , State Partnered with vendor to manage the in-house advertising program consisting of print and media collateral pieces. Coordinated pre-show and post-show activities at trade shows. Managed marketing campaigns to generate new business and to support partner and sales teams. Ordered marketing collateral for meetings, trade shows and advisors. Improved, administered and modified marketing programs to increase product awareness. Assisted in preparing internal promotional publications, managed marketing material inventory and supervised distribution of publications to ensure high quality product output. Coordinated marketing materials including brochures, promotional materials and products. Partnered with graphic designers to develop appropriate materials and branding for brochures. Used tracking and reporting systems for sales leads and appointments. 09/2009 to 10/2012 Assistant Head Teller Company Name - City , State Received an internal audit score of 100 %. Performed daily and monthly audits of ATM machines and tellers. Educated customers on a variety of retail products and available credit options. Consistently met or exceeded quarterly sales goals Promoted products and services to customers while maintaining company brand identity · Implemented programs to achieve and exceed customer and company participation goals Organized company sponsored events on campus resulting in increased brand awareness · Coached peers on the proper use of programs to improve work flow efficiency Utilized product knowledge to successfully sell to and refer clients based on individual needs Promoted marketing the grand opening of new branch locations to strengthen company brand affinity · Organized company sponsored events resulting in increased brand awareness and improved sales · Coached peers on the proper use of programs to increase work flow efficiency Senior Producer - 2014 SHU Media Exchange Company Name - City , State Planned and executed event focusing on Connecticut's creative corridor, growth of industry and opportunities that come with development. A panel of industry professionals addressed topics related to media and hosted a question and answer session for approximately 110 attendees. Following the forum, guests were invited to engage in networking and conversation at a post-event reception. Education 2014 Master of Arts : Corporate Communication & Public Relations Sacred Heart University - City , State 2013 Bachelor of Arts : Relational Communication Western Connecticut State University - City , [illegible], brochures, content, Customer Care, Final Cut Pro, graphics, graphic, HR, Illustrator, InDesign, Innovation, inventory, Lotus Notes, marketing, marketing materials, marketing material, materials, Microsoft	HR

Overview of Variables:

Variable	Variable Name	Description	Type
Target Variable	Category	The job category for which the resume was submitted, with present categories including HR, IT, Education, Legal, and more.	Object

Input Variable	ID	A unique identifier for each resume, also serving as the filename for the corresponding PDF.	Numerical
Input Variable	Resume_str	The resume content in plain text format.	Object

2. Text Mining - Essential components:

The essential components of the text mining are defined below.

Tokenization:

The process of breaking down text into individual units called tokens is known as tokenization. These tokens can be words, sentences, or even smaller units like characters or n-grams. Tokenization serves as a fundamental step in text mining and natural language processing (NLP) tasks.

Stop Words:

Stop words are common words that are frequently removed from text during preprocessing due to their lack of significant meaning. Examples of stop words include "the," "is," "and," and "in." Removing stop words aids in noise reduction and enhances the efficiency of text mining algorithms.

Term Frequency-Inverse Document Frequency (TF-IDF):

TF-IDF is a numerical representation of a term's significance within a document or a corpus. It considers both the frequency of a term within a document (TF) and its rarity across the entire corpus (IDF). TF-IDF finds widespread application in text classification, information retrieval, and keyword extraction.

Term Frequency-Entropy (TF-Entropy):

TF-Entropy assesses the relevance of a term by combining its frequency in a document (TF) with its distribution across a corpus (Entropy). It accentuates terms that are common in one document but uncommon in others, providing a new viewpoint on term relevance.

Term Frequency-Mutual Information (TF-MI):

TF-MI is a text analysis metric that combines a term's frequency in a document (TF) with its unique association with that document (Mutual Information, MI). It discovers phrases that are both frequent and highly suggestive of the document's specific subject.

Term-Document Matrix (TDM):

A term-document matrix (TDM) represents a corpus of documents where each row corresponds to a unique term (word) in the corpus, and each column corresponds to a document. The matrix entries represent the frequency or presence of the term in each document.

3. Modeling and Forecasting : Key Evaluation Metrics

Test ROC Index:

The Test ROC Index, also known as the receiver operating characteristic curve (ROC AUC), is a performance measure commonly used in binary classification tasks. It assesses the ability of a classification model to distinguish between positive and negative instances by plotting the true positive rate (TPR) against the false positive rate (FPR). A higher ROC AUC indicates better model performance, as it represents a greater ability to correctly identify positive instances while minimizing the misclassification of negative instances.

Test Misclassification Rate:

Test misclassification, in the context of text mining, refers to the proportion of instances that a classification model incorrectly classifies. It serves as a measure of the model's accuracy in predicting the correct class labels for text data. A lower misclassification rate indicates a more accurate model.

Prediction Errors:

Prediction errors in text mining represent the discrepancies between the actual values or labels of text instances and the predicted values or labels assigned by a text mining model. These errors quantify the differences between the model's predictions and the true values. Analyzing prediction errors can help identify areas where the model may require improvement.

Mean Absolute Percentage Error (MAPE):

MAPE is a widely used error metric for evaluating the performance of prediction models. It calculates the average absolute difference between the actual values or labels of text instances and the predicted values or labels, expressed as a percentage. MAPE provides a straightforward interpretation of prediction accuracy, as a lower MAPE value indicates closer predictions to the actual values.

Akaike Information Criterion (AIC):

AIC serves as a model selection criterion in text mining, aiming to assess the goodness-of-fit of models without overfitting. It considers both the likelihood of the observed data given the model and the complexity of the model. A lower AIC value indicates a preferred model, as it suggests a better balance between fit and parsimony.

Bayesian Information Criterion (BIC):

BIC, also known as the Schwarz Information Criterion (SIC) or Schwarz Bayesian Criterion (SBC), is another model selection criterion used in text mining. Similar to AIC, it considers the likelihood function and penalizes complex models. Lower BIC values indicate preferred models.

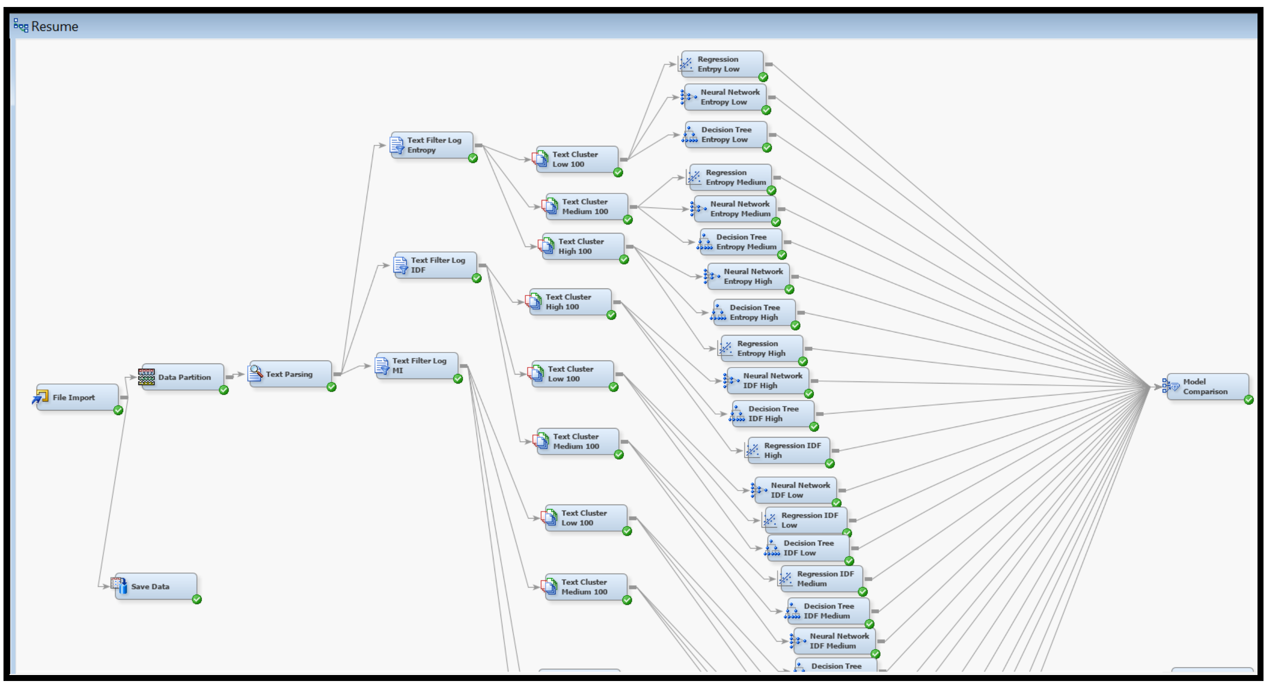
Root Mean Squared Error (RMSE):

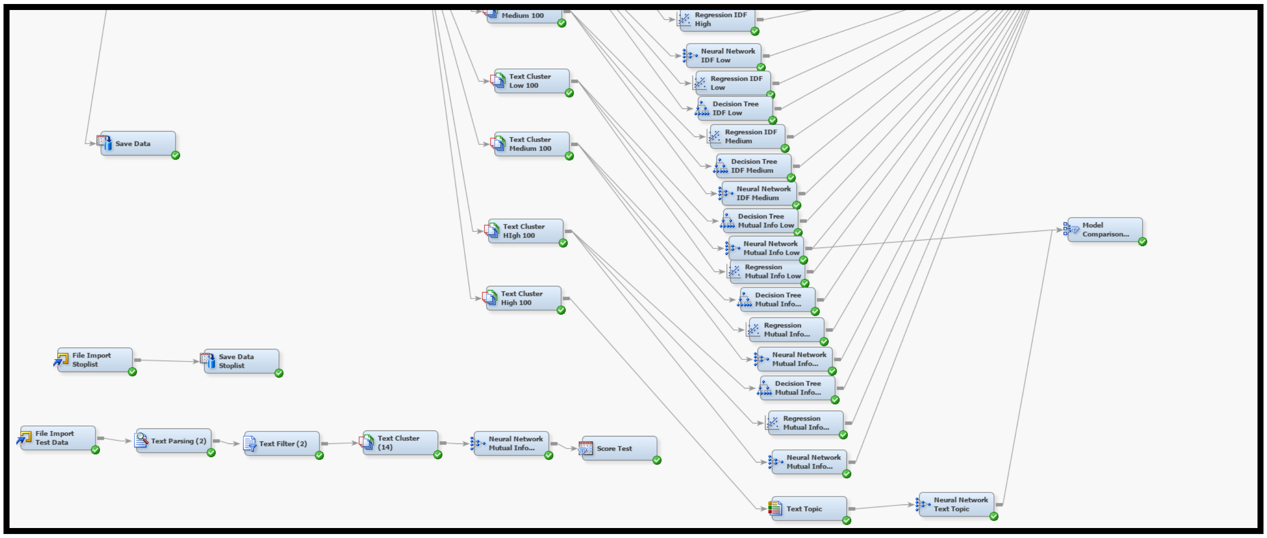
RMSE is a frequently employed metric for evaluating the accuracy of text mining models. It calculates the square root of the mean of the squared differences between the actual values or labels and the predicted values or labels. RMSE provides an indication of the average magnitude of the prediction errors. A lower RMSE value indicates better model performance, as it suggests smaller discrepancies between predictions and true values.

These key evaluation metrics provide valuable insights into the performance of text mining models, enabling data scientists and analysts to assess the effectiveness of different modeling approaches and select the most appropriate models for specific applications.

4. Full Model Diagram:

This is the full model diagram with all the models we tried with different input settings.





5. Model Description:

Now going through each of the nodes shown in the above diagram in detail.

5.1 File Import Node:

SAS Enterprise Miner Workstation streamlines the import of text data from diverse sources, including plain text files, Microsoft Word documents, PDFs, and web pages. The built-in data import capabilities seamlessly integrate your text data into the tool for comprehensive analysis.

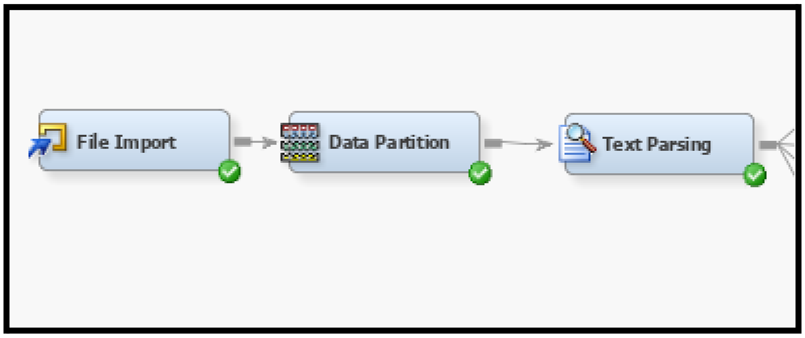
5.2 Data Partition Node:

This node effectively manages your text data with SAS Enterprise Miner Workstation's data partitioning tools. Divide your dataset into training and testing sets to ensure unbiased performance evaluation. Train your models using one set and test their accuracy on the other, ensuring their generalizability to unseen data.

5.3 Text Parsing Node:

It unlocks the intricacies of text data with SAS Enterprise Miner Workstation's advanced text parsing functionalities and transforms unstructured text into structured information by tokenizing words, identifying sentences, tagging parts of speech, and conducting syntactic parsing. These techniques pave the way for meaningful analysis.

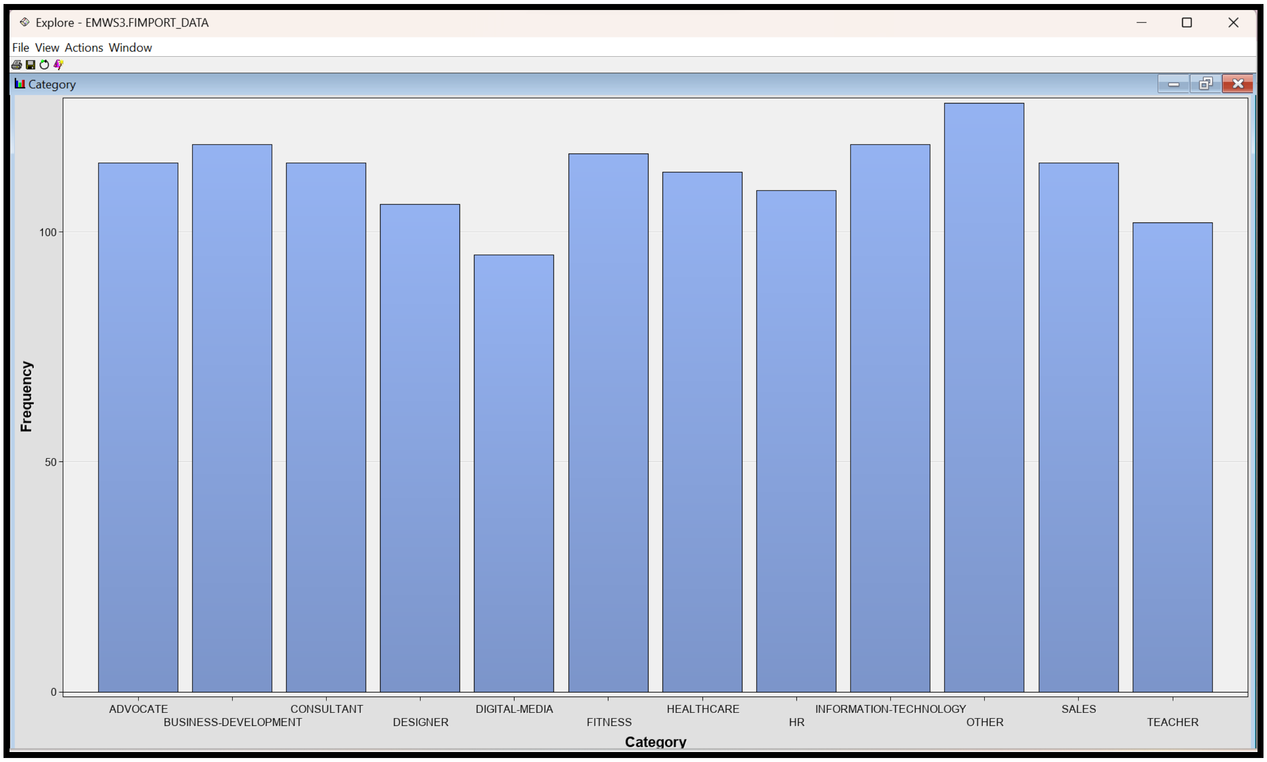
Going through each of the node settings,



In the file import node, we have provided the path of the files and edited the variables as per our needs. We have made the Category as the Target variable and the ID and the Resume_str and other properties as default.

Name	Partition Role	Role	Level
Category	**Default**	Target	Nominal
ID	**Default**	ID	Nominal
Resume_str	**Default**	Text	Nominal

Under the variables section in file import node, frequency of each category is shown. Earlier from the 16 categories, BUDGETING, CHEF, BPO, AGRICULTURE, AUTOMOBILE have the least frequency i.e., less than 50 and hence they are merged together and formed a new category called OTHER to balance the data in the categories to give more accurate results.



Exploring the File import exported data :

Here we are exploring the train dataset where we can see four columns with the observations in the first column followed by the ID, Resume_str and the Category.

EMWS13.FIMPORT_train

Obs #	ID	Resume_str	Category
1	16852973	HR ADMINISTRATOR/MARKETING ASSOCIATEHR ADMINISTRATOR Summar...	HR
2	27018550	HR SPECIALIST Summary Dedicated, Driven, and Dynamic with over 20 years ...	HR
3	17812897	HR MANAGER Skill Highlights HR SKILLS HR Department Startup Three...	HR
4	11592605	HR GENERALIST Summary Dedicated and focused Administrative Assistant wh...	HR
5	25824789	HR MANAGER Summary HUMAN RESOURCES MANAGER Extensive backgro...	HR
6	15375009	HR MANAGER Professional Summary Senior HR professional with a continuou...	HR
7	11847784	HR SPECIALIST Summary Possess 15+ years of experience as an HR Classific...	HR
8	29149998	HR ASSISTANT Summary Highly motivated, and a dynamic Human Resources ...	HR
9	11480899	HR MANAGER Summary Human Resources Manager with practical understa...	HR
10	23155093	HR MANAGER Summary To obtain a position that offers many opportunities for...	HR
11	11763983	HR GENERALIST Summary A people-oriented, results-driven professional with ...	HR
12	27490876	HR COORDINATOR Summary Applicant Screening, Background Checks, Ben...	HR
13	93002334	HR ANALYST Summary Experienced professional with background in Human ...	HR
14	24184357	HR DIRECTOR Summary Human Resource ProfessionalConfident, Resourcef...	HR
15	73077810	HR GENERALIST/RECRUITER Summary Human Resource Generalist who i...	HR
16	13879043	HR CONSULTING Summary 7+ years of Experience as a HR Partner with exper...	HR
17	30163002	HR GENERALIST Summary Young, dedicated and focused office administrati...	HR
18	18827609	HR ASSOCIATE Professional Summary Enthusiastic and goal-oriented HR Pro...	HR
19	25676643	HR SPECIALIST Summary An Human Resources Specialist with over 9 years i...	HR
20	87968870	HR GENERALIST Summary Energetic, Bilingual Human Resources Profession...	HR
21	46258701	HR COORDINATOR Professional Summary Highly efficient Hr Coordinator well ...	HR
22	14225422	HR MANAGER/GENERALIST Summary Background of progressively responsibl...	HR
23	29297973	HR REPRESENTATIVE Summary Experienced human resources professional w...	HR
24	19717385	HR INTERN Summary An enthusiastic student highly motivated and committed t...	HR

Data Set Allocation:

To ensure effective model training and evaluation, we employed data partitioning within the Data

Partition node. This process divided the dataset into three distinct subsets: training, validation, and testing. We opted for a partition ratio of 50-30-20, allocating 50% of the data for training, 30% for validation, and the remaining 20% for testing.

This partitioning strategy provided us with clearly defined subsets of data for specific purposes. The training set served as the foundation for developing and training our models. The validation set played a crucial role in fine-tuning and selecting the most suitable model. Finally, the testing set acted as an independent dataset to rigorously evaluate the performance of the chosen model. This approach ensured that we objectively assessed the model's generalizability to unseen data.

Property	Value
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Metho	Default
Random Seed	12345
Data Set Allocatio	
Training	50.0
Validation	20.0
Test	30.0
Report	
Interval Targets	Yes
Class Targets	Yes

Data partition node results:

Data=TEST

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Category	.	ADVOCATE	23	8.39416	Category
Category	.	BUSINESS-DEVELOPMENT	23	8.39416	Category
Category	.	CONSULTANT	23	8.39416	Category
Category	.	DESIGNER	22	8.02920	Category
Category	.	DIGITAL-MEDIA	19	6.93431	Category
Category	.	FITNESS	24	8.75912	Category
Category	.	HEALTHCARE	23	8.39416	Category
Category	.	HR	22	8.02920	Category
Category	.	INFORMATION-TECHNOLOGY	24	8.75912	Category
Category	.	OTHER	27	9.85401	Category
Category	.	SALES	23	8.39416	Category
Category	.	TEACHER	21	7.66423	Category

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Category	.	ADVOCATE	57	8.45697	Category
Category	.	BUSINESS-DEVELOPMENT	60	8.90208	Category
Category	.	CONSULTANT	57	8.45697	Category
Category	.	DESIGNER	53	7.86350	Category
Category	.	DIGITAL-MEDIA	47	6.97329	Category
Category	.	FITNESS	58	8.60534	Category
Category	.	HEALTHCARE	57	8.45697	Category
Category	.	HR	55	8.16024	Category
Category	.	INFORMATION-TECHNOLOGY	59	8.75371	Category
Category	.	OTHER	63	9.34718	Category
Category	.	SALES	58	8.60534	Category
Category	.	TEACHER	50	7.41840	Category

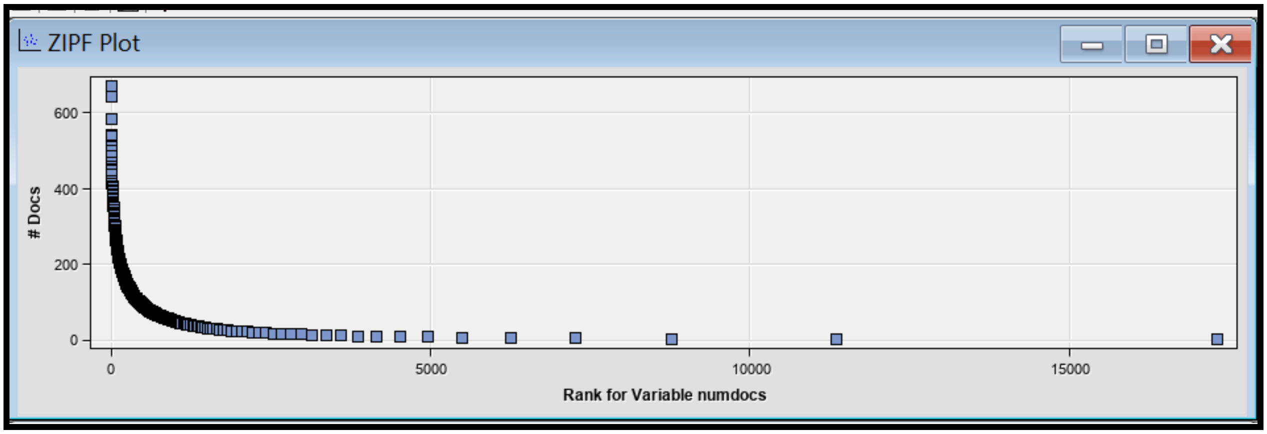
Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Category	.	ADVOCATE	35	8.64198	Category
Category	.	BUSINESS-DEVELOPMENT	36	8.88889	Category
Category	.	CONSULTANT	35	8.64198	Category
Category	.	DESIGNER	31	7.65432	Category
Category	.	DIGITAL-MEDIA	29	7.16049	Category
Category	.	FITNESS	35	8.64198	Category
Category	.	HEALTHCARE	33	8.14815	Category
Category	.	HR	32	7.90123	Category
Category	.	INFORMATION-TECHNOLOGY	36	8.88889	Category
Category	.	OTHER	38	9.38272	Category
Category	.	SALES	34	8.39506	Category
Category	.	TEACHER	31	7.65432	Category

The screenshot depicts the classification of resumes into job positions such as ADVOCATE, BUSINESS-DEVELOPMENT, and CONSULTANT, among others, across VALIDATE, TEST, and TRAIN databases. Each entry provides the job category, the number of resumes in that category, and the proportion of resumes that fall into that category, ensuring that our text mining model is trained, tested, and validated on a well-distributed sample. This balanced distribution is critical for the accuracy and fairness of your automated recruiting tool, which promises to speed the resume screening process, improve the job search experience, and provide labor market insights by automatically sorting resumes into relevant job categories.

Zipf’s Plot

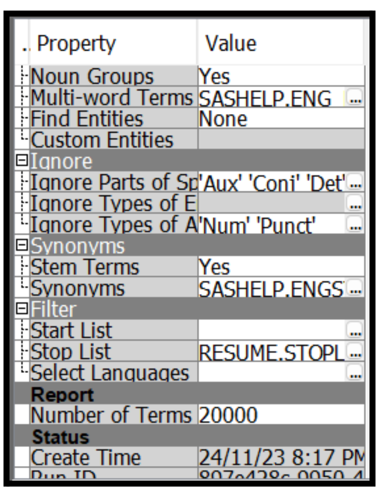
Zipf’s Law suggests that terms appearing with low frequency and terms appearing with high frequency are irrelevant.



Seeing the above plot, we can say that the frequency counts of terms are very long tailed. That is, there is a small number of very common terms that are used over and over again in most of the documents.

Text parsing configuration:

In the text parsing node, we've used all the default properties except for one change which is a modification of the stop list. We have added words that SAS has to ignore such as 'new', 'state', 'city' etc.



Property	Value
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Sp	'Aux' 'Conj' 'Det'
Ignore Types of E	
Ignore Types of A	'Num' 'Punct'
Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGS
Filter	
Start List	
Stop List	RESUME.STOPL
Select Languages	
Report	
Number of Terms	20000
Status	
Create Time	24/11/23 8:17 PM

Snap Shot of Stoplist:

These are a few works we added to our Stop list as shown below. The criteria we used are term weight to be less than 0.1, words not repeating in more than five documents and some domain knowledge.

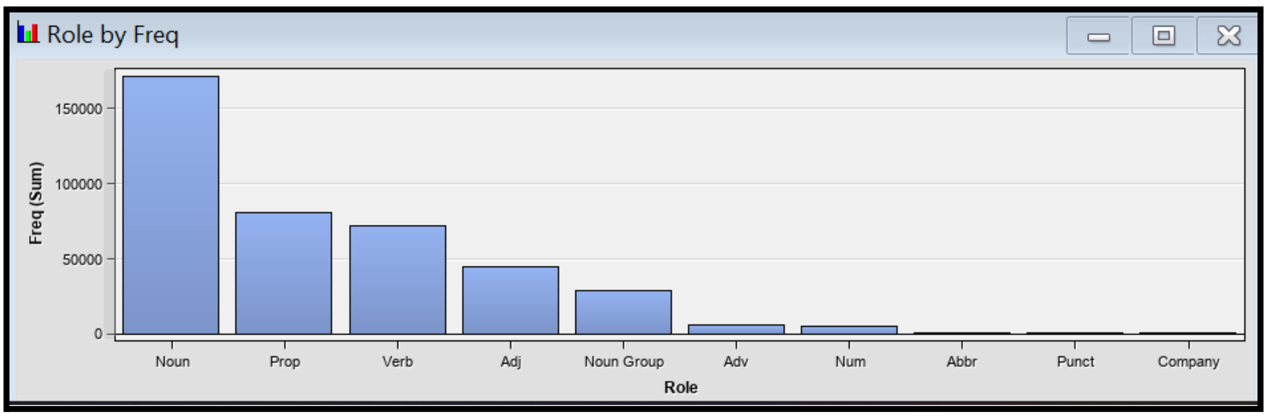
	A	B
1	TERM	ROLE
2	name	FALSE
3	company	FALSE
4	state	FALSE
5	education	FALSE
6	skill	FALSE
7	management	FALSE
8	new	FALSE
9	skills	FALSE
10	city	FALSE
11	experience	FALSE
12	maintain	FALSE
13	develop	FALSE
16	manage	FALSE
17	all	FALSE
18	experience	FALSE
28	be	FALSE
30	year	FALSE
43	use	FALSE
45	other	FALSE
54	ï	FALSE
58	knowledge	FALSE
64	need	FALSE
68	company name ï	FALSE
76	professional	FALSE
77	include	FALSE
78	need	FALSE
93	make	FALSE
110	â	FALSE
135	may	FALSE
193	have	FALSE
226	accomplishments	FALSE

Text parsing results:

SAS Enterprise Miner Workstation groups the terms and plots them based on their part of speech. This visualization organizes the terms into different categories corresponding to their respective parts of speech.

This visualization provides insights into the linguistic composition of the text corpus. For instance, a high proportion of nouns might indicate that the text is primarily factual, while a high proportion of adjectives might suggest that the text is more descriptive or emotional. The visualization can also be used to identify potential errors in the text parsing process, such as misclassified parts of speech.

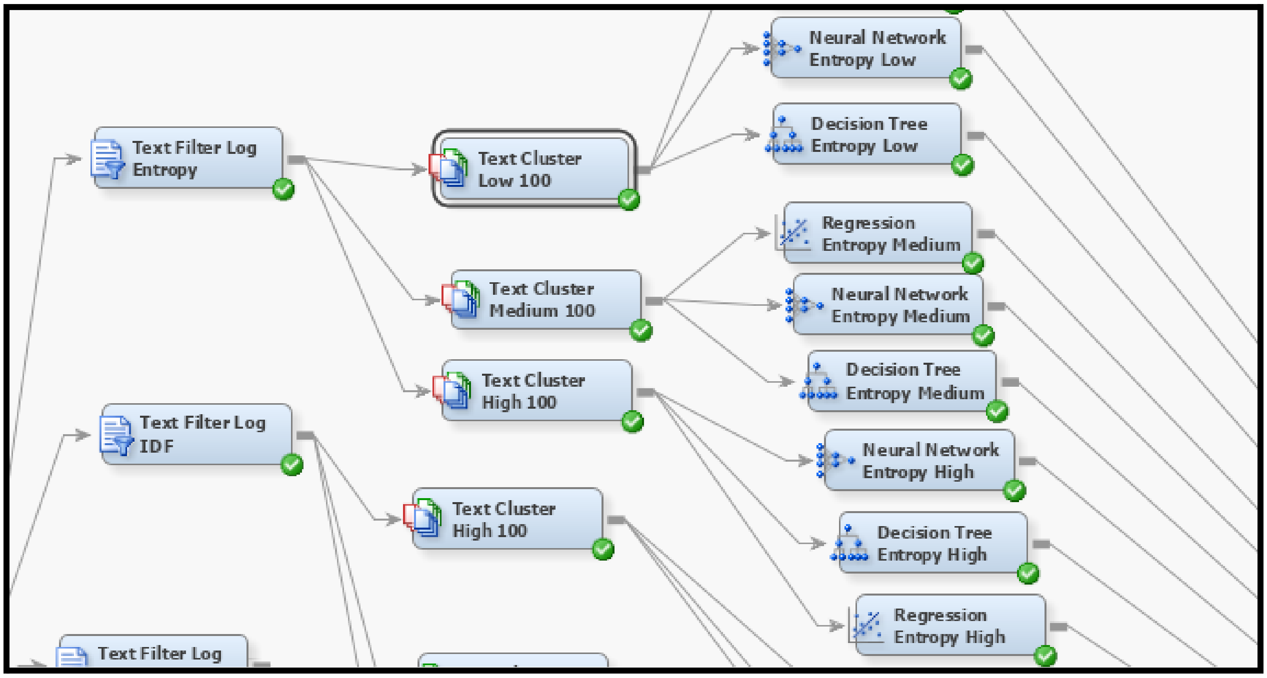
Overall, the part-of-speech visualization is a valuable tool for understanding the structure and content of text data. It can be used to inform a variety of text mining tasks, such as topic modeling, sentiment analysis, and information extraction.



5.4 Text Filtering Node:

Helps to cleanse and prepare your text data for analysis with SAS Enterprise Miner Workstation's text filtering capabilities. It also applies filters to eliminate stop words, punctuation, special characters, or other unwanted elements, ensuring your data is free from noise and ready for further exploration.

5.4.1 Logarithmic Frequency Weight and Entropy-based Term Weight



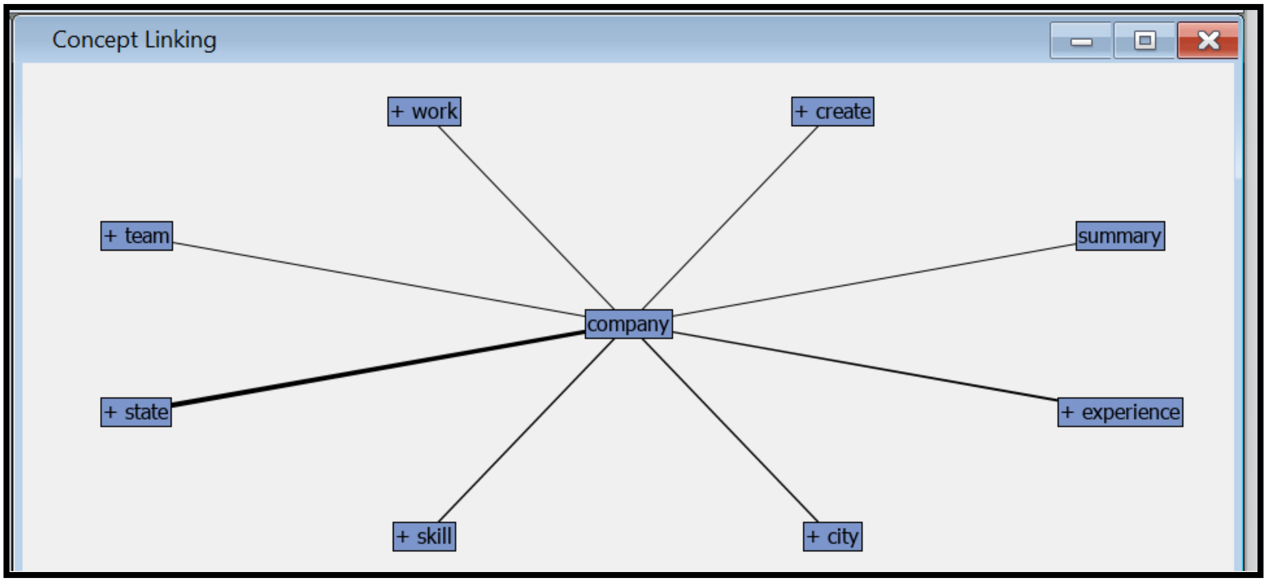
Text filter configuration:

Here we can see that the weightings section settings where the frequency weight is set as Log and the term weight as the Entropy.

Weightings	
Frequency Weight	Log
Term Weight	Entropy

Concept linking:

We can see the concept linking for one of the terms which is Company is more associated with. This can be also interpreted as follows, the child term state is contained in 642 documents, and 642 of these documents contain the parent term company..



Text filter results:

Terms

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Parent/Child Status	Parent ID
name ...	Prop	Alpha	Drop	0.000	3178	3178	673	673	1		73529
compa...	Prop	Alpha	Keep	0.021	2954	2954	672	672	2		52513
+ state...	Noun	Alpha	Keep	0.022	4405	4405	642	642	3	+	50035
+ educ...	Noun	Alpha	Keep	0.050	869	869	584	584	4	+	40062
+ skill ...	Noun	Alpha	Keep	0.068	1523	1523	541	541	5	+	38037
+ man...	Noun	Alpha	Keep	0.078	2042	2042	538	538	6	+	18399
+ new ...	Adj	Alpha	Drop	0.000	1757	1757	515	515	7	+	73703
skills ...	Prop	Alpha	Keep	0.062	659	659	511	511	8		21148
+ city ...	Noun	Alpha	Keep	0.067	2766	2766	501	501	9	+	8808
+ expe...	Noun	Alpha	Keep	0.085	963	963	485	485	10	+	51521
+ main...	Verb	Alpha	Keep	0.093	1439	1439	469	469	11	+	47924
+ devel...	Verb	Alpha	Keep	0.102	1560	1560	458	458	12	+	49849
+ servi...	Noun	Alpha	Keep	0.110	1687	1687	450	450	13	+	38042
+ custo...	Noun	Alpha	Keep	0.132	3162	3162	449	449	14	+	7623
+ man...	Verb	Alpha	Keep	0.103	1431	1431	447	447	15	+	45848
all ...	Adj	Alpha	Drop	0.000	1290	1290	436	436	16		73672
experie...	Prop	Alpha	Keep	0.078	460	460	419	419	17		22401
+ busin...	Noun	Alpha	Keep	0.124	1736	1736	413	413	18	+	20216
+ com...	Noun	Alpha	Keep	0.112	1237	1237	413	413	18	+	26139
+ devel...	Noun	Alpha	Keep	0.119	1291	1291	412	412	20	+	37630
+ plan ...	Verb	Alpha	Keep	0.120	1143	1143	405	405	21	+	31513
+ client...	Noun	Alpha	Keep	0.128	1936	1936	404	404	22	+	549
current...	Prop	Alpha	Keep	0.092	493	493	404	404	22		52350
+ provi...	Verb	Alpha	Drop	0.000	908	908	403	403	24	+	73562
summa...	Prop	Alpha	Keep	0.083	401	401	396	396	25		20322
inform...	Noun	Alpha	Keep	0.124	952	952	394	394	26		19634
+ be ...	Verb	Alpha	Drop	0.000	1233	1233	394	394	26	+	73664

The screenshot above shows the result of a text filter, which details the extraction of key terms from a corpus based on their structural role, frequency, and relevance. Terms are given a status to indicate whether they will be kept ('Keep') or dropped ('Drop') in further analysis, with weights indicating their importance. The frequency counts for both the imported and current datasets are displayed, and terms may be ordered based on these parameters. Parent/Child status and IDs indicate a hierarchical structuring of phrases, which could be part of a classification used for efficiently categorizing resumes, assisting in the automation of the recruiting process by recognizing significant terms that correlate with job categories.

Interactive Filter Viewer:

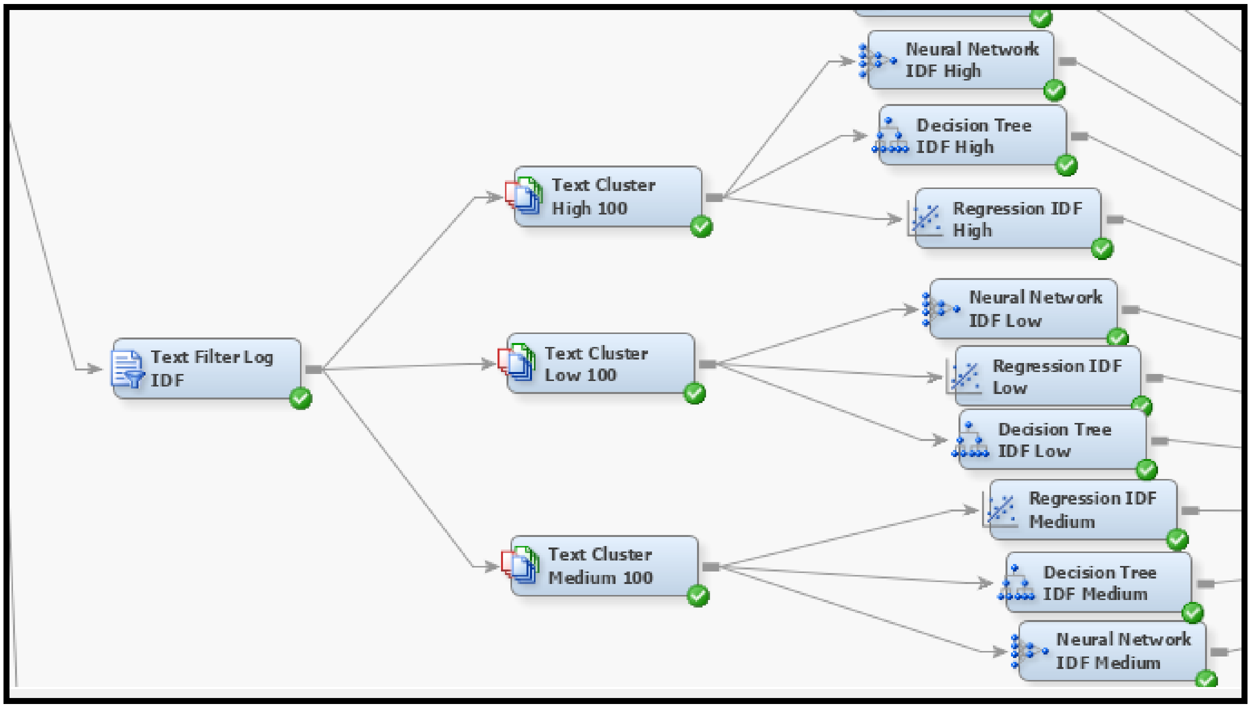
The provided image displays the frequency weights of terms. A stop list can be formulated by retaining words that have weights surpassing the 0.1 threshold and excluding those below it.

Terms

TERM	FREQ	# DOCS	KEEP ▾	WEIGHT	ROLE	ATTRIBUTE
company	2954	672	✓	0.001	Prop	Alpha
state	4405	642	✓	0.02	Noun	Alpha
education	869	584	✓	0.044	Noun	Alpha
skill	1523	541	✓	0.055	Noun	Alpha
managem...	2042	538	✓	0.057	Noun	Alpha
skills	659	511	✓	0.06	Prop	Alpha
city	2766	501	✓	0.047	Noun	Alpha
experience	963	485	✓	0.051	Noun	Alpha
maintain	1439	469	✓	0.089	Verb	Alpha
develop	1560	458	✓	0.086	Verb	Alpha
service	1687	450	✓	0.11	Noun	Alpha
customer	3162	449	✓	0.145	Noun	Alpha
manage	1431	447	✓	0.085	Verb	Alpha
experience	460	419	✓	0.063	Prop	Alpha
business	1736	413	✓	0.212	Noun	Alpha
company	1237	413	✓	0.131	Noun	Alpha
developm...	1291	412	✓	0.213	Noun	Alpha
plan	1143	405	✓	0.124	Verb	Alpha
client	1936	404	✓	0.114	Noun	Alpha
current	493	404	✓	0.068	Prop	Alpha
summary	401	396	✓	0.123	Prop	Alpha
information	952	394	✓	0.233	Noun	Alpha
create	946	391	✓	0.174	Verb	Alpha
year	838	387	✓	0.049	Noun	Alpha
office	1177	382	✓	0.108	Noun	Alpha
work	867	381	✓	0.071	Verb	Alpha
communic...	770	376	✓	0.14	Noun	Alpha
work	738	373	✓	0.1	Noun	Alpha
team	864	371	✓	0.131	Noun	Alpha
system	1475	370	✓	0.229	Noun	Alpha
training	887	364	✓	0.16	Noun	Alpha
managem...	1154	364	✓	0.177	Prop	Alpha
member	864	364	✓	0.137	Noun	Alpha
ensure	878	361	✓	0.132	Verb	Alpha
sale	2173	359	✓	0.273	Noun	Alpha
university	603	355	✓	0.097	Prop	Alpha
project	1150	342	✓	0.176	Noun	Alpha
product	1134	337	✓	0.19	Noun	Alpha
implement	808	336	✓	0.171	Verb	Alpha
train	800	334	✓	0.166	Verb	Alpha
report	814	328	✓	0.163	Verb	Alpha
staff	768	328	✓	0.128	Noun	Alpha
data	1098	323	✓	0.181	Noun	Alpha

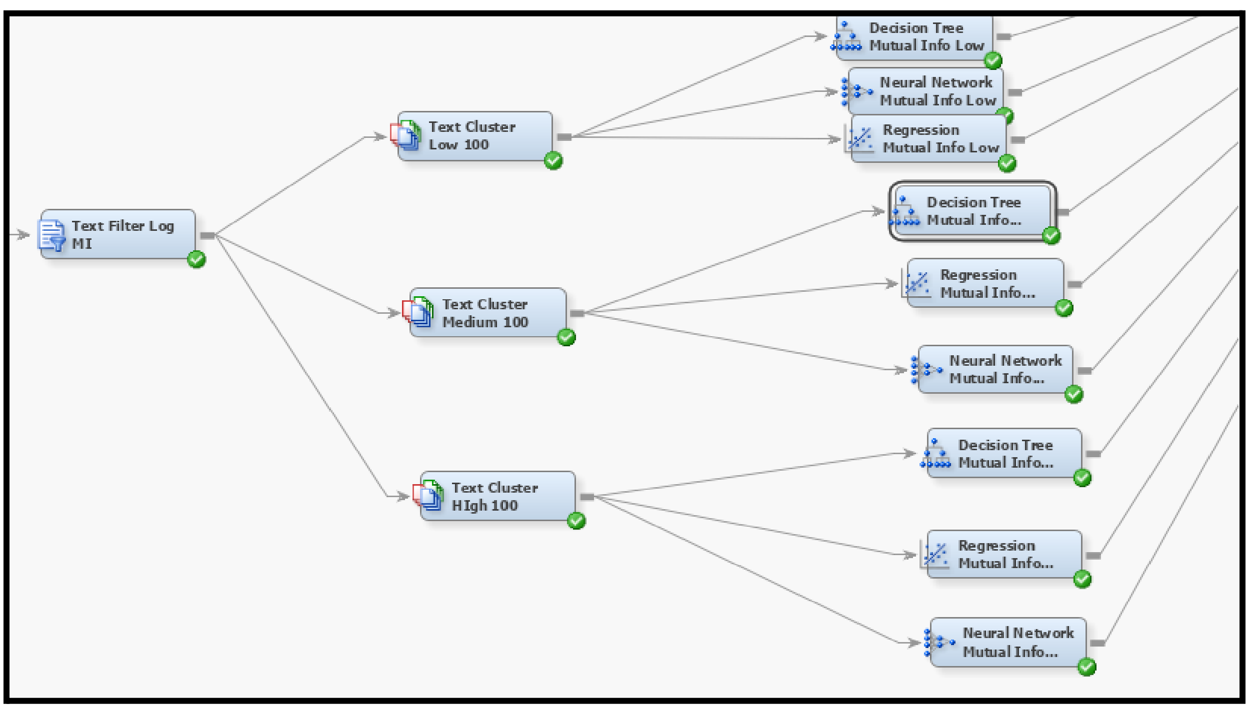
5.4.2 Logarithmic Frequency Weight and IDF-based Term Weight

A closer look at the diagram when the frequency weight is set to log and the term weight as IDF, with no change in configurations from the previous model.



5.4.3 Logarithmic Frequency Weight and Mutual Information-based Term Weight

We tried changing the term weight as mutual information as it is recommended as the default for documents that are associated with a categorical target variable which is category.



5.5 Text Clustering Node:

This node uncovers patterns and themes in your text data using SAS Enterprise Miner Workstation's clustering algorithms. Also groups similar documents together based on their content, revealing hidden relationships and thematic structures within your text corpus.

Text Cluster Configuration:

For the Text Cluster node, we configured the clustering settings as follows:

Train	
Variables	...
⊟Transform	
SVD Resolution	Medium
Max SVD Dimensi	100
⊟Cluster	
Exact or Maximum	Maximum
Number of Cluster	12
Cluster Algorithm	Expectation-Maxin
Descriptive Terms	15
Status	
Create Time	24/11/23 8:17 PM
Run ID	63927f09-154b-4
Last Error	

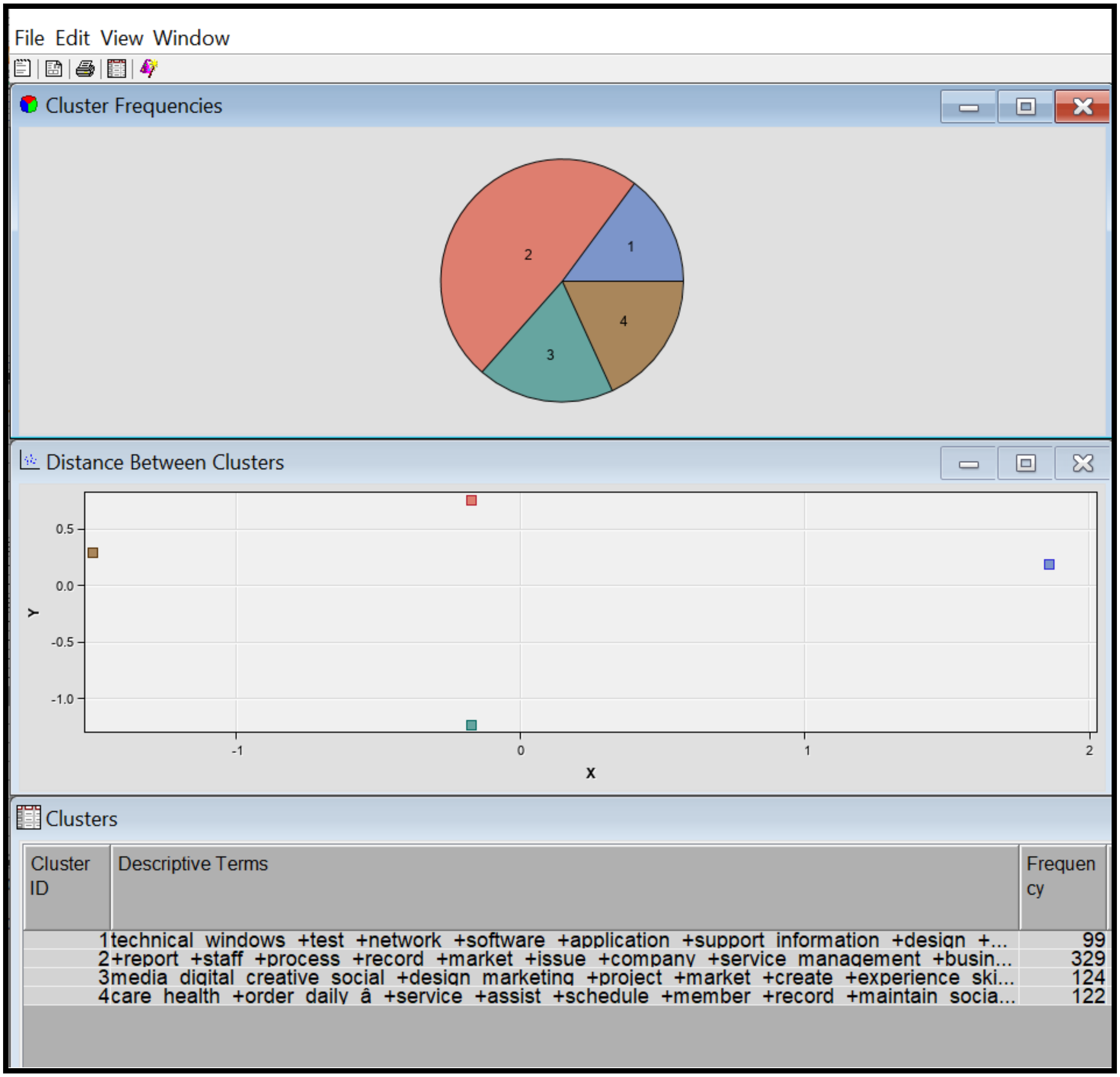
• To align with our dataset's 12 categories for the target variable, we designated the cluster number as precisely 12. This decision aimed to categorize text documents into distinct clusters of 12 based on their content and sentiment.

• With a dataset comprising 2400 resumes, we capped the maximum SVD (Singular Value Decomposition) dimensions at 100. This choice aimed to reduce data dimensionality while retaining a significant amount of information. The SVD dimensions played a critical role in extracting pertinent features from the text data, facilitating effective document clustering.

• We experimented with various SVD resolutions—low, medium, and high—to ascertain the optimal setting for our model's performance.

Text cluster results for low 100 and frequency weight as entropy:

Using a low SVD resolution resulted in the creation of 44 dimensions, while the mid-resolution generated a total of 69 dimensions. Opting for the high resolution led to the creation of all 100 dimensions, ultimately yielding the most favorable outcomes in our analysis.



Cluster Frequencies (Pie Chart): The pie chart displays the relative sizes of the dataset's clusters. Each segment is labeled with a cluster number, and the size of the segment represents the frequency of the cluster or the number of documents/terms it contains. Here we can see than most of the terms are present in category 2.

Distance Between Clusters (Scatter Plot): The scatter plot beneath the pie chart illustrates the distances between the clusters, with the x and y axes potentially reflecting multiple dimensions or principal components. The geographical layout of the squares (each representing a cluster) demonstrates how unique each cluster is in terms of distinguishing qualities from the others. Here, we can see that each cluster distance is large which indicates that there are few very terms that overlap with each other which is a good indication.

Clusters (Table with Descriptive Terms): At the bottom, the table lists the descriptive terms that are most distinctive of each cluster and enumerates the clusters by ID. These phrases reveal the thematic or subject substance of each cluster. The 'Frequency' column counts the number of times the descriptive terms appear, creating a link between each cluster ID and the exact terms that characterize it. The frequency count of cluster 2 is highest which has a count of 324 for one of the models we tried which contain terms like report, staff etc. This thorough split helps in interpreting each cluster's thematic focus, which is necessary for effectively categorizing text data, such as sorting resumes into job-related groups.

For all the other text cluster nodes, we have used the same configuration that we discussed above. These are connected to regression and decision tree models with the same configuration as mentioned above. All these models are connected to a model comparison node.

5.6 Model Nodes:

Regression Node - Logistic Regression:

Logistic regression attempts to predict the probability that a binary or ordinal target will acquire the event of interest as a function of one or more independent inputs. We have selected the Stepwise as the model selection so that all the inputs are used to fit the model and we have used the Validation error as the selection criterion.

Property	Value
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactio	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes

The fit statistics results for the regression model is as follows where we can see the various statistics for the train and validation values.

Fit Statistics

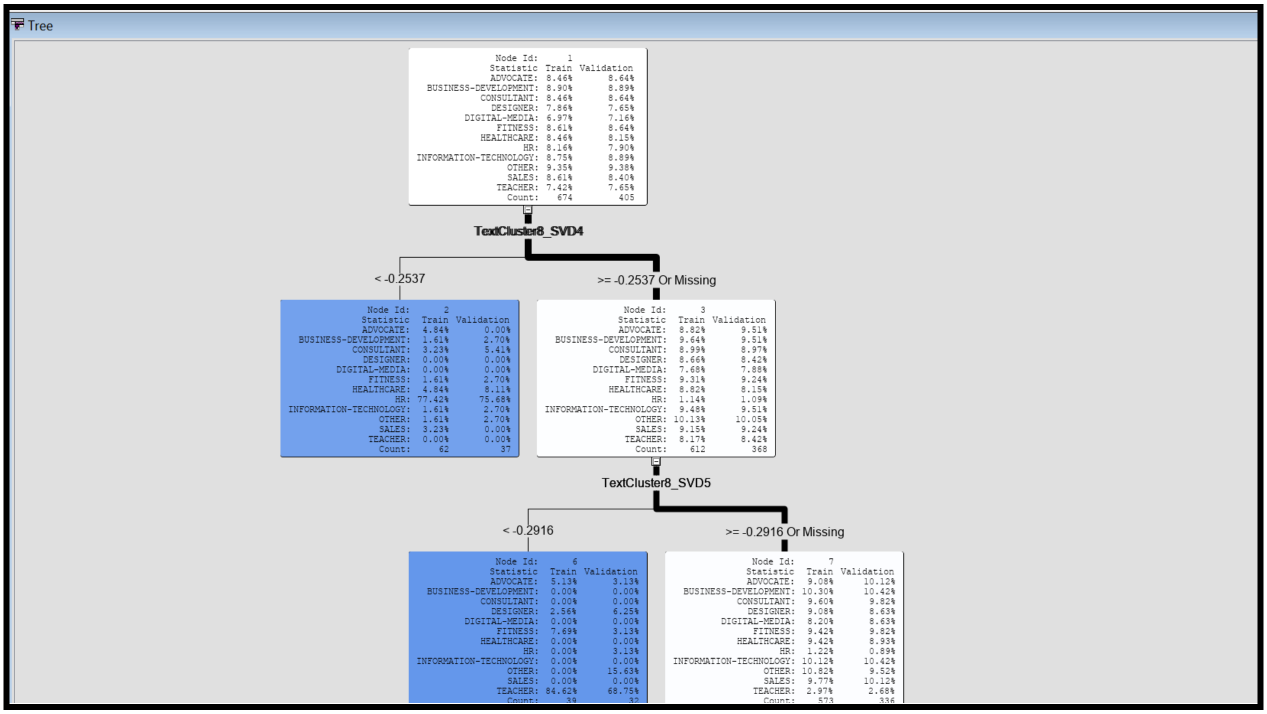
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Category	Category	AIC	Akaike's Infor...	1746.095	.	.
Category	Category	ASE	Average Squa...	0.042053	0.0473	0.044034
Category	Category	AVERR	Average Error...	0.188686	0.230945	0.224761
Category	Category	DFE	Degrees of Fr...	7304	.	.
Category	Category	DFM	Model Degree...	110	.	.
Category	Category	DFT	Total Degrees...	7414	.	.
Category	Category	DIV	Divisor for ASE	8088	4860	3288
Category	Category	ERR	Error Function	1526.095	1122.393	739.0148
Category	Category	FPE	Final Predictio...	0.043319	.	.
Category	Category	MAX	Maximum Abs...	0.997945	0.999995	0.99999
Category	Category	MSE	Mean Square ...	0.042686	0.0473	0.044034
Category	Category	NOBS	Sum of Frequ...	674	405	274
Category	Category	NW	Number of Est...	110	.	.
Category	Category	RASE	Root Average ...	0.205068	0.217485	0.209844
Category	Category	RFPE	Root Final Pre...	0.208133	.	.
Category	Category	RMSE	Root Mean Sq...	0.206606	0.217485	0.209844
Category	Category	SBC	Schwarz's Ba...	2506.319	.	.
Category	Category	SSE	Sum of Squar...	340.1223	229.877	144.785
Category	Category	SUMW	Sum of Case ...	8088	4860	3288
Category	Category	MISC	Misclassificati...	0.373887	0.42716	0.405109

Decision Tree Node:

It is used as a decision-making framework that partitions data into progressively smaller subsets based on specific characteristics. It's constructed by applying a sequence of simple rules, each of which assigns an observation to a subset based on the value of a single input variable. This process of successive rule applications creates a hierarchical structure, resembling a tree, where each subset is represented by a node. The initial, all-encompassing subset is called the root node, and the ultimate, non-divisible subsets are termed leaves. For each leaf node, a decision is made and applied to all observations belonging to that leaf. The nature of the decision depends on the specific context. Most of the settings were left as is.

Use Multiple Targets	No
Splitting Rule	
Interval Target Criteri	ProbF
Nominal Target Criter	ProbChisq
Ordinal Target Criteri	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate	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	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validat	No
Number of Subsets	10
Number of Repeats	1
Seed	12345

The following is the first two depths of the tree given in the output results. Here we can see the train and the validation split percentages along with the best path of the tree along with the clusters chosen.

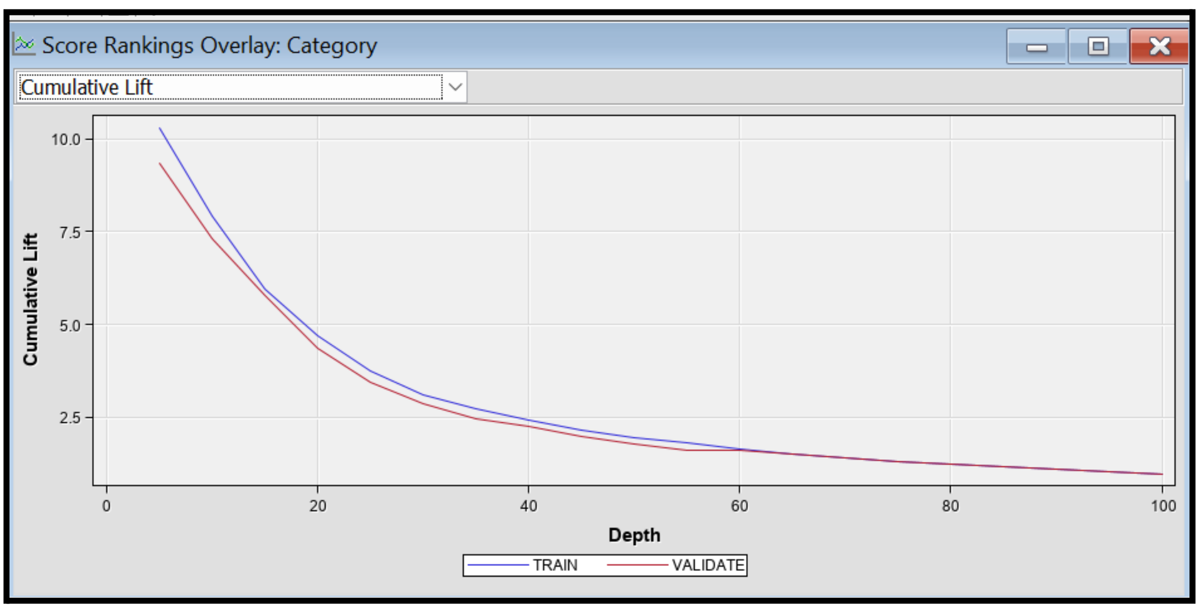


Neural Network Node:

Neural networks are powerful for capturing complex, nonlinear relationships within data. Text data often contains intricate patterns and dependencies that may not be well-suited for linear models like regression or decision trees. Neural networks excel at learning hierarchical representations and abstract features from raw data. All values given in the properties are default.

Property	Value
General	
Node ID	Neural7
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Continue Trainir	No
Network	...
Optimization	...
Initialization See	12345
Model Selection	Profit/Loss
Suppress Outpu	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No
Status	
Create Time	11/19/23 7:24 F
Run ID	7b57e7cd-1a9b-
Last Error	
Last Status	Complete
Last Run Time	11/23/23 10:13
Run Duration	0 Hr. 0 Min. 9.0
Grid Host	
User-Added Noc	No

The cumulative lift for this model is as follows.

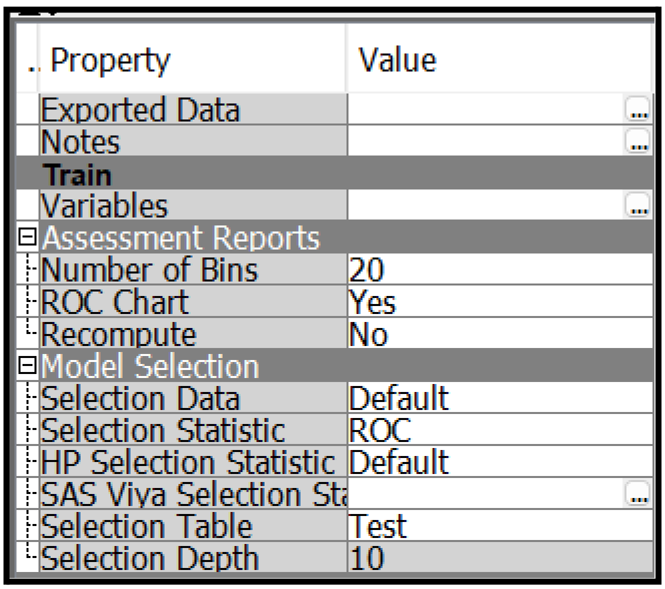


At a depth of 20%, the cumulative lift for the training data is approximately 4.03, which means that the model is able to identify 4.03 times as many targets in the top 20% of the population as would be expected by random chance.

5.7 Model Comparison Node:

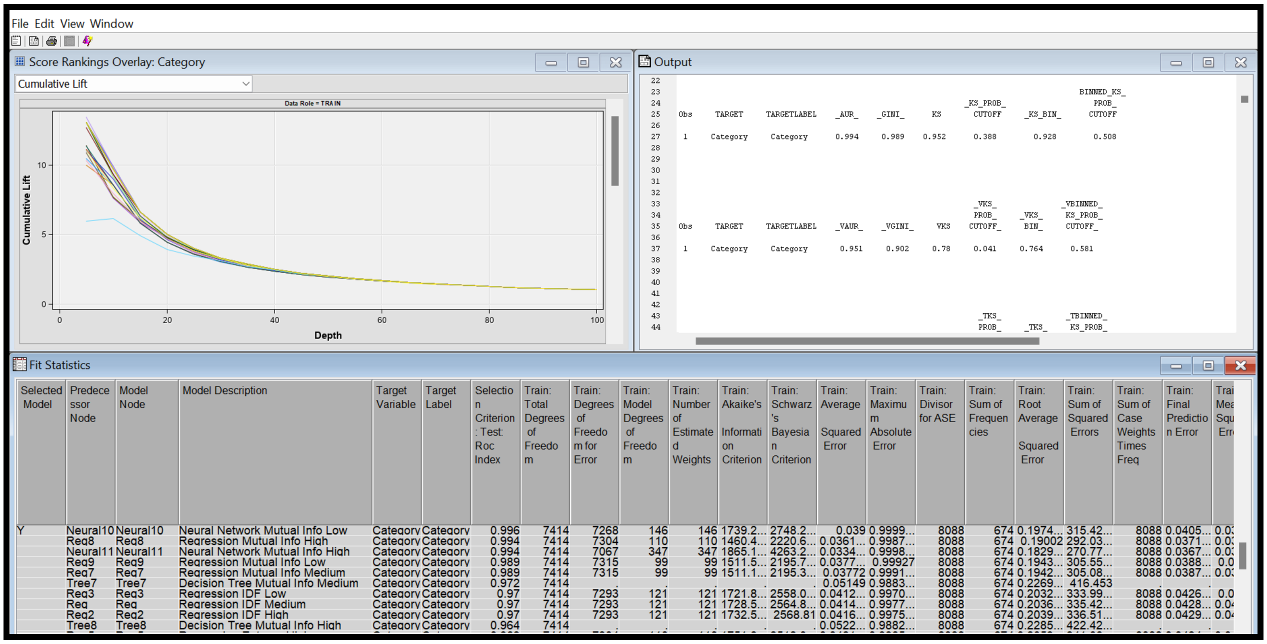
Make informed decisions with SAS Enterprise Miner Workstation's comprehensive model comparison tools. Evaluate and compare the performance of various text mining models using metrics like Test ROC Score, Misclassification rate, RMSE, accuracy, precision, recall, F1-score, and other domain-specific measures. Select the most suitable model for our specific text mining task based on rigorous performance assessment.

We have selected the Selection data as default and the selection statistic as ROC.

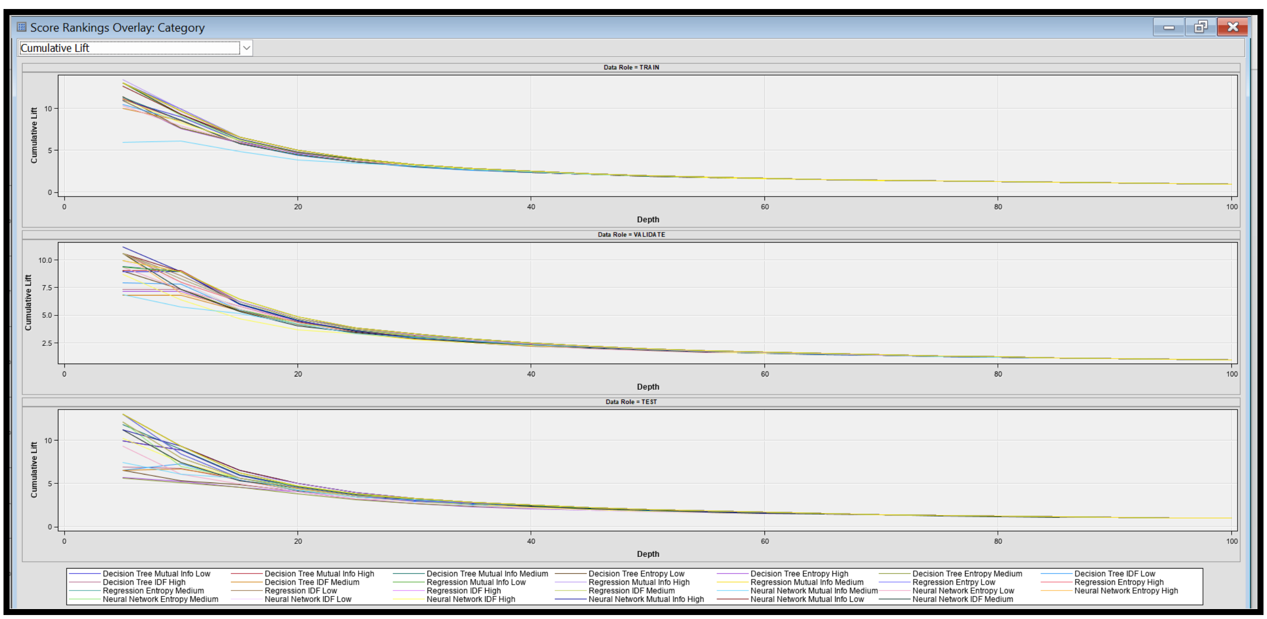


Property	Value
Exported Data	...
Notes	...
Train	
Variables	...
Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
Model Selection	
Selection Data	Default
Selection Statistic	ROC
HP Selection Statistic	Default
SAS Viya Selection Sta	...
Selection Table	Test
Selection Depth	10

The results of the model comparison node is as follows,



Looking at the cumulative fit closely,



The cumulative fit of the model in the image shows that it is able to learn the relationship between the input and output variables with a high degree of accuracy. The cumulative fit is a measure of the proportion of the variance in the output variable that can be explained by the model. In this case, the cumulative fit is 0.95, which means that the model can explain 95% of the variance in the output variable.

This suggests that the model is a good fit for the data and can be used to make accurate predictions about the output variable for new input data points.

Now looking at the Fit Statistics window and breaking the results in a table according to their term weight and frequency weight.

5.7.1 Model Comparison Results for Log frequency weight and Entropy term weight:

Now going through the model comparison results, we see that the following values

SVD Values	Model type	ROC	Misclassification Rate
Low 100	Regression	0.962	0.3738
	Neural network	0.913	0.4332
	Decision Tree	0.908	0.5385
Medium 100	Regression	0.962	0.3738
	Neural network	0.937	0.3902
	Decision Tree	0.87	0.5326
High 100	Regression	0.963	0.3783
	Neural network	0.937	0.3364
	Decision Tree	0.861	0.5890

We can see that we have obtained the highest ROC for the Regression model of 0.963 when the SVD settings were high and 100.

5.7.2 Model Comparison Results for Log frequency weight and IDF term weight:

SVD Values	Model type	ROC	Misclassification Rate
Low 100	Regression	0.97	0.3635
	Neural network	0.904	0.473
	Decision Tree	0.909	0.5934
Medium 100	Regression	0.97	0.3664
	Neural network	0.931	0.4228
	Decision Tree	0.926	0.58
High 100	Regression	0.97	0.3664
	Neural network	0.914	0.5712
	Decision Tree	0.92	0.5771

We can see that we have obtained the highest ROC for the Regression model of 0.97 when the SVD settings are either high or low. The value is near to 1 which indicates that the terms in the documents are not so overlapping.

5.7.3 Model Comparison Results Log frequency weight and MI term weight:

SVD Values	Model type	ROC	Misclassification Node
Low 100	Regression	0.989	0.3293
	Neural network	0.996	0.3308
	Decision Tree	0.957	0.5341
Medium 100	Regression	0.989	0.3293
	Neural network	0.929	0.3916

	Decision Tree	0.972	0.5281
High 100	Regression	0.994	0.3278
	Neural network	0.994	0.2818
	Decision Tree	0.964	0.5326

We can see that Mutual information did increase the ROC and the Neural network model turned out to be the best model for our data.

Comparison between the 5.7.1, 5.7.2, 5.7.3 models:

From the above models obtained, we can say that the Neural network model of Mutual information with SVD resolution of high is the best model for this dataset when both misclassification rate and ROC are taken into account. The mutual information makes sense as we have a categorical target variable making it a better term weight choice when compared to the Entropy and IDF.

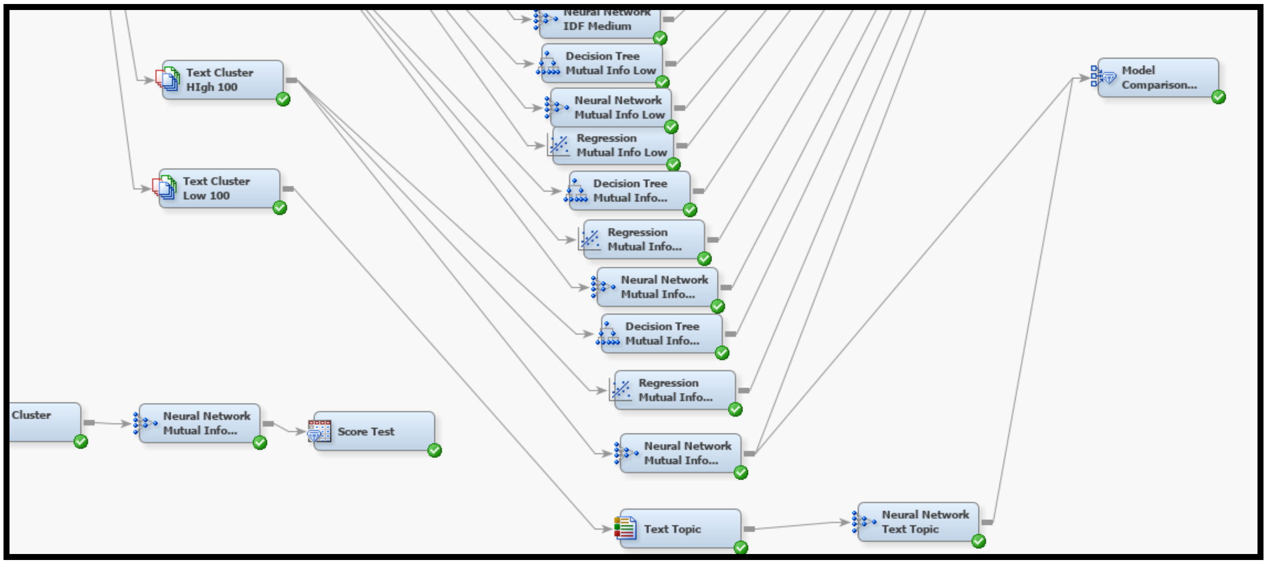
The misclassification rate of the best model along with the ROC is as follows.

Term weight	**Test misclassification rate**	**ROC**
Mutual Information	0.28190	0.994

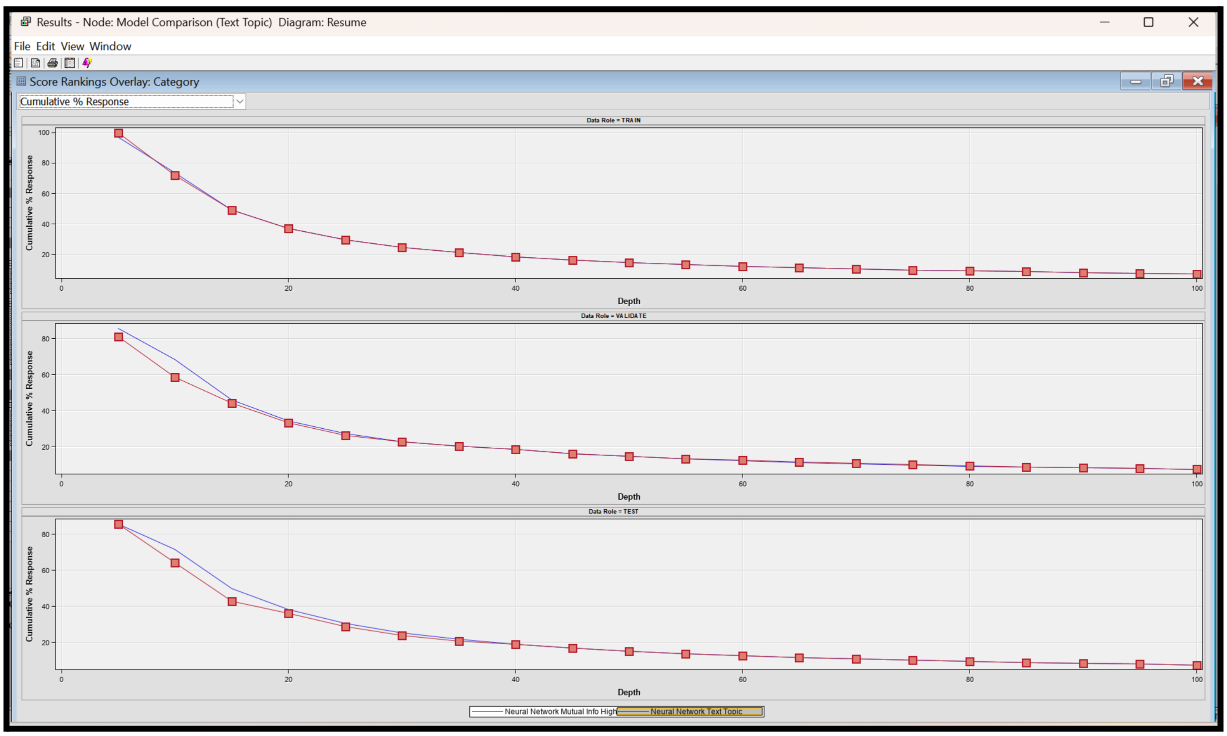
5.8 Text Topic Node:

Helps to delve into the underlying topics of your text data with SAS Enterprise Miner Workstation's topic modeling techniques. Employs algorithms like Latent Dirichlet Allocation (LDA) to identify the key themes and subjects that permeate your text corpus, gaining valuable insights into the core discussions within your data.

Comparison between the Text topic node results with the best model:



We included Text Topic node in our best model path and compared the results with text topic and without text topic. For our data, the best results are achieved with the text topic node.



From the above screenshot, we can infer that the cumulative percent response is high for the Neural Network with text topics for the test data with 85%.

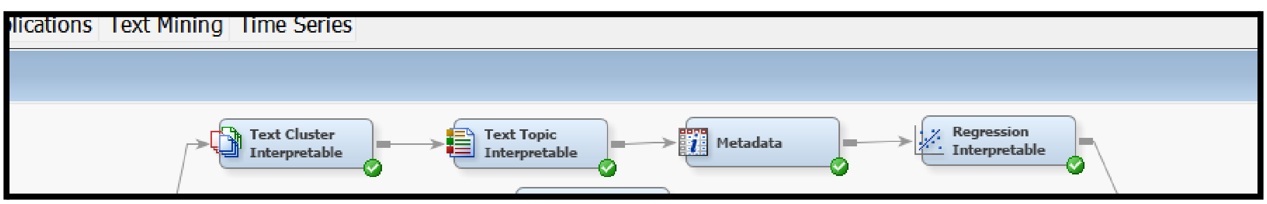
	Misclassification Rate	**ROC**

With Text Topic	0.25	0.998
Without Text Topic	0.28	0.998

6. Best Model:

Observing the test misclassification rates in the table, it is clear that using the **mutual information high (MI High) setting with text cluster high and SVD 100 appended with text topic with neural network** gives the best results and is considered as the best model. Neural networks are powerful for capturing complex, nonlinear relationships within data. Text data often contains intricate patterns and dependencies that may not be well-suited for linear models like regression or decision trees. Neural networks excel at learning hierarchical representations and abstract features from raw data. In the case of text data, neural networks can automatically extract meaningful features from words or phrases, capturing semantic relationships that may be challenging for traditional models. Text data often results in high-dimensional feature spaces due to the presence of a large vocabulary. Neural networks, especially deep learning models, can handle high-dimensional data more effectively than traditional models.

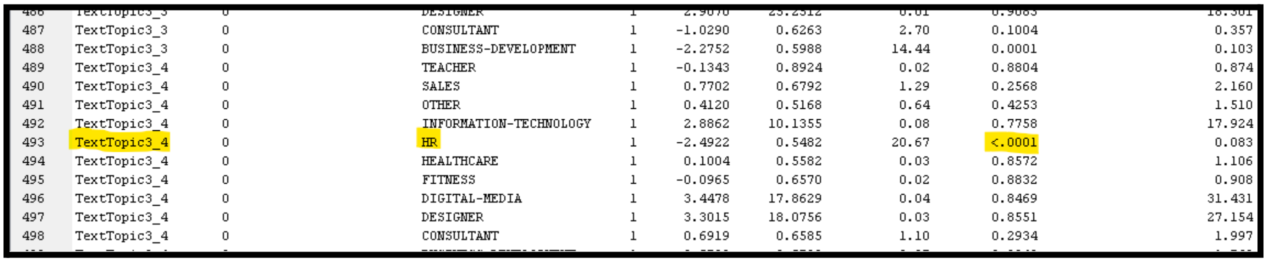
7. Interpretable Model:



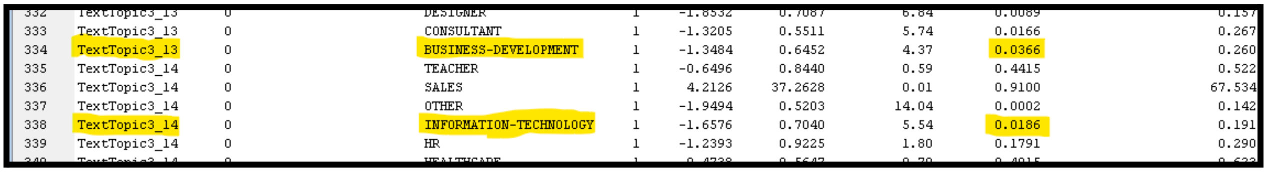
For interpretable model, we added a text cluster with Maximum SVD dimensions of 50 and text topic and metadata to the regression model. Here we modified the input variables in the metadata node. For example, we rejected the ‘Text cluster_SVD’ related and ‘text_topic_raw’ related variables and passed ‘Text_cluster_prob’ and ‘Text_Topic’ related variables as input to the regression model. After running the model, the misclassification rate is 44% which is higher than our best model which is the Neural network. And hence interpretable is not considered as the best model for this dataset.

Results from the model:

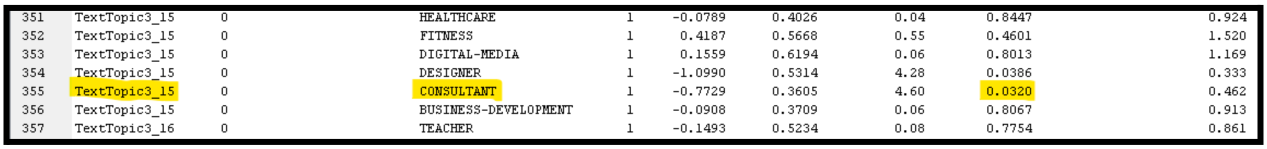
Though it is not considered as best model, we can draw a few interpretations and keywords that can be used in our resumes for a particular category.



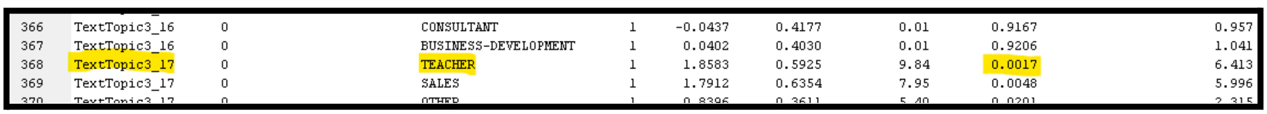
487	TextTopic3_3	0	CONSULTANT	1	-1.0290	0.6263	2.70	0.1004	0.357
488	TextTopic3_3	0	BUSINESS-DEVELOPMENT	1	-2.2752	0.5988	14.44	0.0001	0.103
489	TextTopic3_4	0	TEACHER	1	-0.1343	0.8924	0.02	0.8804	0.874
490	TextTopic3_4	0	SALES	1	0.7702	0.6792	1.29	0.2568	2.160
491	TextTopic3_4	0	OTHER	1	0.4120	0.5168	0.64	0.4253	1.510
492	TextTopic3_4	0	INFORMATION-TECHNOLOGY	1	2.8862	10.1355	0.08	0.7758	17.924
493	TextTopic3_4	0	HR	1	-2.4922	0.5482	20.67	<.0001	0.083
494	TextTopic3_4	0	HEALTHCARE	1	0.1004	0.5582	0.03	0.8572	1.106
495	TextTopic3_4	0	FITNESS	1	-0.0965	0.6570	0.02	0.8832	0.908
496	TextTopic3_4	0	DIGITAL-MEDIA	1	3.4478	17.8629	0.04	0.8469	31.431
497	TextTopic3_4	0	DESIGNER	1	3.3015	18.0756	0.03	0.8551	27.154
498	TextTopic3_4	0	CONSULTANT	1	0.6919	0.6585	1.10	0.2934	1.997



333	TextTopic3_13	0	CONSULTANT	1	-1.3205	0.5511	5.74	0.0166	0.267
334	TextTopic3_13	0	BUSINESS-DEVELOPMENT	1	-1.3484	0.6452	4.37	0.0366	0.260
335	TextTopic3_14	0	TEACHER	1	-0.6496	0.8440	0.59	0.4415	0.522
336	TextTopic3_14	0	SALES	1	4.2126	37.2628	0.01	0.9100	67.534
337	TextTopic3_14	0	OTHER	1	-1.9494	0.5203	14.04	0.0002	0.142
338	TextTopic3_14	0	INFORMATION-TECHNOLOGY	1	-1.6576	0.7040	5.54	0.0186	0.191
339	TextTopic3_14	0	HR	1	-1.2393	0.9225	1.80	0.1791	0.290



351	TextTopic3_15	0	HEALTHCARE	1	-0.0789	0.4026	0.04	0.8447	0.924
352	TextTopic3_15	0	FITNESS	1	0.4187	0.5668	0.55	0.4601	1.520
353	TextTopic3_15	0	DIGITAL-MEDIA	1	0.1559	0.6194	0.06	0.8013	1.169
354	TextTopic3_15	0	DESIGNER	1	-1.0990	0.5314	4.28	0.0386	0.333
355	TextTopic3_15	0	CONSULTANT	1	-0.7729	0.3605	4.60	0.0320	0.462
356	TextTopic3_15	0	BUSINESS-DEVELOPMENT	1	-0.0908	0.3709	0.06	0.8067	0.913
357	TextTopic3_16	0	TEACHER	1	-0.1493	0.5234	0.08	0.7754	0.861



366	TextTopic3_16	0	CONSULTANT	1	-0.0437	0.4177	0.01	0.9167	0.957
367	TextTopic3_16	0	BUSINESS-DEVELOPMENT	1	0.0402	0.4030	0.01	0.9206	1.041
368	TextTopic3_17	0	TEACHER	1	1.8583	0.5925	9.84	0.0017	6.413
369	TextTopic3_17	0	SALES	1	1.7912	0.6354	7.95	0.0048	5.996

From the above screenshot we can see that yellow highlighted content represents the topics which contain the keywords that are necessary for that particular category. All these are significant as their $p<0.05$. To be more specific, TextTopic3_4 with HR contains the important words used for resume related to HR and similarly with TextTopic3_13, TextTopic3_15, TextTopic3_17 for BUSINESS DEVELOPMENT, CONSULTANT and TEACHER respectively.

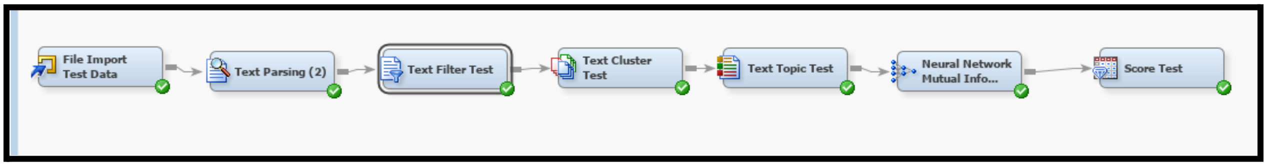
Parameter	**Category**	**Some Keywords**
TextTopic3_4	HR	hr,+compensation,+employee, human,human+recruitment
TextTopic3_13	Business-Development	food,+clean,+item,+guest
TextTopic3_15	Consultant	_+tax,financial,accounting,+statement,+loan
TextTopic3_17	Teacher	child,+teacher,+reinforcement

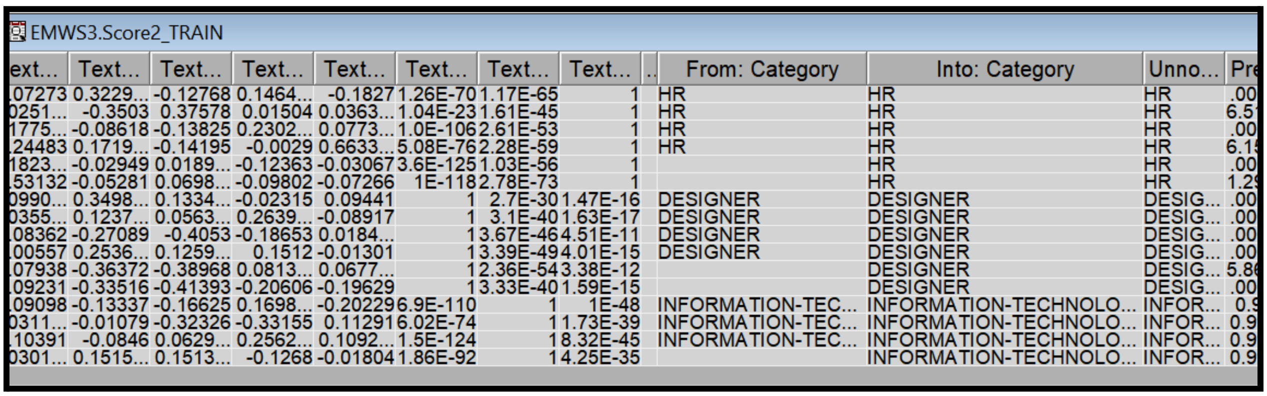
		method,,+play

Few words might not be appropriate for the category, because this is not the best model.

8. Testing the tuned model:

To assess the effectiveness of the best model, we input resumes with both blank and specified categories. The input resumes marked as blank categories are correctly classified into the appropriate categories, as indicated by the output category in the "Into" field in the provided screenshot.





EMWS3.Score2_TRAIN

ext...	Text...	Text...	Text...	Text...	Text...	Text...	Text...	..	From: Category	Into: Category	Unno...	Pre
.07273	0.3229...	-0.12768	0.1464...	-0.1827	1.26E-70	1.17E-65	1		HR	HR	HR	.00
0251...	-0.3503	0.37578	0.01504	0.0363...	1.04E-23	1.61E-45	1		HR	HR	HR	6.5
1775...	-0.08618	-0.13825	0.2302...	0.0773...	1.0E-106	2.61E-53	1		HR	HR	HR	.00
.24483	0.1719...	-0.14195	-0.0029	0.6633...	5.08E-76	2.28E-59	1		HR	HR	HR	6.1
1823...	-0.02949	0.0189...	-0.12363	-0.03067	3.6E-125	1.03E-56	1			HR	HR	.00
.53132	-0.05281	0.0698...	-0.09802	-0.07266	1E-118	2.78E-73	1			HR	HR	1.2
0990...	0.3498...	0.1334...	-0.02315	0.09441	1	2.7E-30	1.47E-16		DESIGNER	DESIGNER	DESIG...	.00
0355...	0.1237...	0.0563...	0.2639...	-0.08917	1	3.1E-40	1.63E-17		DESIGNER	DESIGNER	DESIG...	.00
.08362	-0.27089	-0.4053	-0.18653	0.0184...	1	3.67E-46	4.51E-11		DESIGNER	DESIGNER	DESIG...	.00
.00557	0.2536...	0.1259...	0.1512	-0.01301	1	3.39E-49	4.01E-15		DESIGNER	DESIGNER	DESIG...	.00
.07938	-0.36372	-0.38968	0.0813...	0.0677...	1	2.36E-54	3.38E-12			DESIGNER	DESIG...	5.8
.09231	-0.33516	-0.41393	-0.20606	-0.19629	1	3.33E-40	1.59E-15			DESIGNER	DESIG...	.00
.09098	-0.13337	-0.16625	0.1698...	-0.20229	6.9E-110	1	1E-48		INFORMATION-TEC...	INFORMATION-TECHNOLO...	INFOR...	0.9
0311...	-0.01079	-0.32326	-0.33155	0.11291	6.02E-74	1	1.73E-39		INFORMATION-TEC...	INFORMATION-TECHNOLO...	INFOR...	0.9
.10391	-0.0846	0.0629...	0.2562...	0.1092...	1.5E-124	1	8.32E-45		INFORMATION-TEC...	INFORMATION-TECHNOLO...	INFOR...	0.9
0301...	0.1515...	0.1513...	-0.1268	-0.01804	1.86E-92	1	4.25E-35			INFORMATION-TECHNOLO...	INFOR...	0.9

9. Conclusion:

To conclude, developing a text mining solution for resume categorization offers a transformative approach to streamlining the recruitment process and enhancing the overall hiring experience. By automating resume screening and leveraging the power of natural language processing, businesses can significantly reduce manual review efforts, improve the accuracy of candidate-job matching, and gain deeper insights into labor market trends. This innovative solution holds the potential to revolutionize recruitment practices, enabling job seekers and businesses to identify and attract top talent with greater efficiency and effectiveness.

10. Business Insights & Future Recommendations:

From the perspective of students:

- Identify and incorporate effective keywords given by model according to the category in resumes and cover letters.
- Showcase alignment with the preferred culture and values of target companies.
- Enhance employability by aligning skills with industry needs.

Students can improve their career prospects by connecting their talents and resumes with text mining insights in recruitment. Focusing on in-demand abilities, designing resumes with successful keywords, and recognizing changing market trends are all examples of this. Students can also highlight their unique origins if they are aware of diversity-focused hiring methods.

From the perspective of businesses and HR:

- Gain insights into candidate interests through resume content analysis.
- Stay informed about dynamic job market changes to guide recruitment strategies.
- Maintain a balance between automation and human insight in recruitment.

Our project can be used by businesses to discover skill gaps in their personnel pool, informing recruitment and training plans. Analyzing resume and job posting patterns aids in the optimization of job descriptions in order to attract qualified candidates and the development of more inclusive hiring processes. Future developments may include more intuitive job-matching technology, which will help to refine the recruitment process even further.

Our project can help HR teams detect skill gaps and labor trends, allowing them to optimize recruiting and link training with market demands. Resume analysis aids in the customization of corporate branding and the improvement of recruiting diversity. The future focus will be on harnessing analytics and automation to improve the efficiency and strategicity of HR procedures.

11. References

https://www.kaggle.com/code/mubtasimahasan/resume-classification-machine-learning-sklearn

Text Analytics using SAS Miner Course notes