Text Mining on Donor's Conversation with Solicitors

Abstract

Solicitors pursue alumni all the time into giving back to the school. Solicitors contact alumni via all the means and note down every conversation they have with them, which are summarized conversations with personal notes added. The objective of this paper is to analyze the text to uncover insights and in the end do predictive modelling. Using text mining nodes in SAS Enterprise Miner 14.1, we will perform text analytics on a database obtained from a university foundation that contained 38,000 observations.

In text analytics, text clustering was conducted and meaningful clusters were obtained and utilized for donor segmentation. We also used text topic node. The rule builder node was used to find key words that were associated with a donation.

Predictive modelling was conducted on text data and text/ numerical combinations and various models were compared. The numeric variables were three internal ratings, gender, degree, school, marital status and state. We would be seeing if textual data when combined with numeric data outperforms numeric data alone or textual data alone.

This paper is going to benefit any fundraising organization and widen the scope of their methods and the way they reach out to constituents.

Data preparation

Data has been obtained from XYZ University foundation database. There were seven tables. Our main table consisted of Constituent ID and Conversation (Text) column. Since the main table had multiple conversations with the constituents and each conversation recorded as unique observation, text was concatenated whenever ID was the same. The other six tables were joined to this table using left joint in SAS and keeping Constituent ID as primary key (ID is foreign key in all the tables). The final dataset had 38000 unique observations and 10 variables (9 numeric variables and 1 text variable). Our target variable is binary variable. 1 – if person made a donation. 0 – if they didn't.

Later for predictive modelling, the data was partitioned into training (80%) and validation (20%) data. Numeric variables were transformed to adjust skewness and kurtosis. Tree based imputation was used to impute missing numeric values. The variable Ethnicity was rejected for modelling because it had 55 percent missing value and for the remaining values, 90% of the values had ethnicity of a single race.

Name	Role	Level
ConstituentID	ID	Nominal
Degree	Input	Nominal
Ethnicity	Rejected	Nominal
Gender	Input	Binary
MaritalStatus	Input	Nominal
SchoolType	Input	Nominal
Target	Target	Binary
TextConversatic	Text	Nominal
bwfRating	Input	Ordinal
evertrueRating	Input	Ordinal
grenRating	Input	Ordinal
state	Input	Nominal

Figure 1: Data dictionary

Methodology and Results

We imported 38,000 observations in SAS Enterprise Miner 14.1. Our first part was finding text clusters, text topics and text rules.

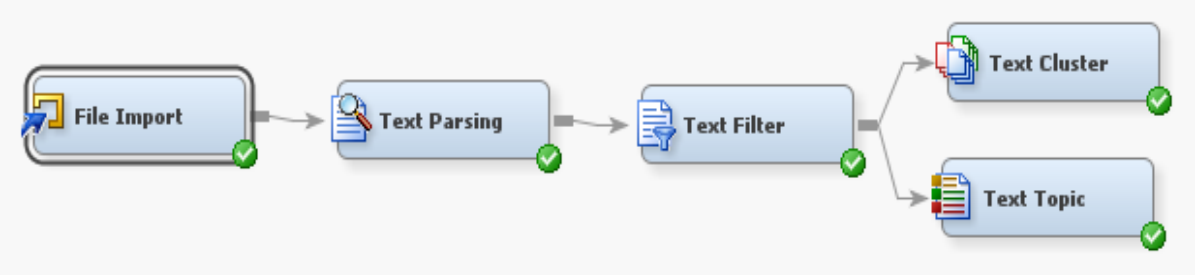
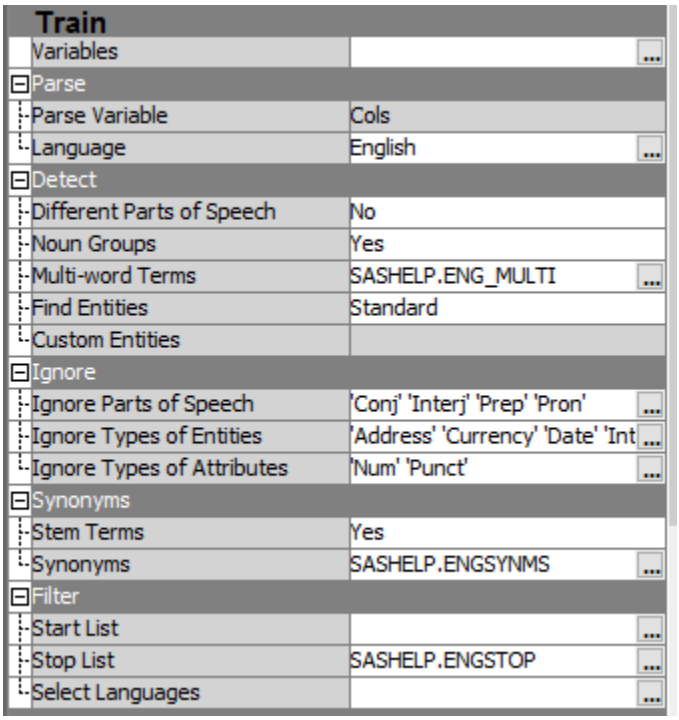


Figure 2: Process Flow for text clusters and text topic

Property	Value
General	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Import File	C:\Users\hgupta\Desktop\sas
Maximum rows to import	1000000
Maximum columns to import	10000
Delimiter	,
Name Row	Yes
Number of rows to skip	0
Guessing Rows	500
File Location	Local
File Type	csv
Advanced Advisor	No
Rerun	No

Figure 3: File Import Settings

After importing data, text parsing node was connected to file import node. Some settings were changed.



Train	
Variables	
Parse	
Parse Variable	Cols
Language	English
Detect	
Different Parts of Speech	No
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI
Find Entities	Standard
Custom Entities	
Ignore	
Ignore Parts of Speech	'Conj' 'Interj' 'Prep' 'Pron'
Ignore Types of Entities	'Address' 'Currency' 'Date' 'Int
Ignore Types of Attributes	'Num' 'Punct'
Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	
Stop List	SASHELP.ENGSTOP
Select Languages	

Figure 4: Text Parsing Settings

- "Find Entities" was changed from 'None" to "Standard"
- In "Ignore Parts of Speech", Conjunction, Interjection, Preposition, and Pronoun was ignored.
- In "Ignore Types of Entities", Address, currency, date, internet, person, phone, product, ssn and time were ignored.
- Number and punctuation were ignored in "Ignore types of attributes".

Then we attached text filter node to text parsing node. All settings were default except "Check Spelling" was marked 'yes'.

TEXT CLUSTERING

After changing names to hide identity of foundation and removing some terms in interactive view filter property of Text filter node, text cluster node was used.

Number of clusters was set to 10 and Descriptive terms to be 15. Clustering algorithm was set to default Expectation – Maximization.

The following clusters were obtained –

Terms in Cluster	Percentage	Possible summary
Work +interest+ year +back +school +talk +major +gift+ plan +time+ want +discuss+ major gift prospect	28	This group has shown interest in giving back to school in the form of major gift (any gift above $25,000)
Visit +email +meeting +schedule +request +attempt +area +meet +message + time +trip +leave +true left	32	In this group, a meeting is being requested by solicitor to constituents.
Mail +dr +letter recent +fund +contribution +donor +scholarship +member +support + student +information +note	29	Members are discussing about funding student's tuition as scholarship money and possibly asking for student information for whom they are going to pay tuition
Stadium +host +attend +game +Sports Z +event +letter +invite amp +dean +plan next +discuss +update	11	Constituents are invited for a popular game (in university) and dean would discuss and plan next opportunities with constituents in giving back to campus.

Figure 5: Text clusters and their possible explanation

TEXT TOPIC

Text topic node was attached to text filter node. Again names were changed to hide identity of foundation and some terms dropped terms in interactive view filter property of Text filter node.

In text topic node, settings were default.

Topic terms	Explanation
+Veterinary School X, +acknowledgment, +recent contribution	About Veterinary School and acknowledging recent contribution there.
+Meet, +request, +time	Solicitor is requesting donor time for meetup
+City X, +donor visit, +true	Solicitors visiting donor in City X

+Unmanaged prospects, +prospects, +county, + County X	About prospects who are not followed up in county X
+Regional, +regional event, +donor visit	Some regional event visited by donor
+Raise money, +value state educator, +inaugural copy	Cant decipher
+disconnect, +phone number, +donor	Donor disconnected phone
+member, +donation, +life member, +donor, +email	Donors who have been donating their life
+upcoming alumnus, student event, State Y, upcoming, City Y	Upcoming alumnus event at City Y, State Y.

Figure 6: Text Topics

Text Rules based model

We are now going to look at words or combination of words that lead to donation and also those words which do not.

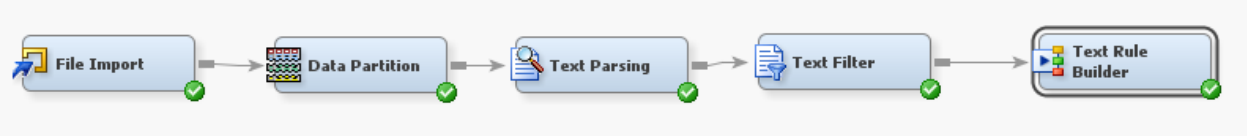


Figure 7: Modelling diagram for generating text rules.

We tried running text rule builder node under various modes and properties. The one with the lowest misclassification rate was chosen. The winner text rule builder model had 'Generalization Error', 'Purity of rules' and 'Exhaustiveness' as 'low'. The misclassification rate of validation dataset was 14.9%

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Target		_ASE_	Average Squared Error	0.044781	0.044978
Target		_DIV_	Divisor for ASE	5038	1262
Target		_MAX_	Maximum Absolute Error	0.583058	0.697646
Target		_NOBS_	Sum of Frequencies	2519	631
Target		_RASE_	Root Average Squared Error	0.211614	0.212079
Target		_SSE_	Sum of Squared Errors	225.6042	56.76176
Target		_DISF_	Frequency of Classified Cases	2519	631
Target		_MISC_	Misclassification Rate	0.132989	0.14897
Target		_WRONG_	Number of Wrong Classifications	335	94

Figure 8: Fit Statistics of winner text rule builder.

Here 1 means donation made and 0 means donation not made. Many words had to be dropped from 'interactive filter viewer' property in text filter node to hide identity of foundation and donors.

Rules obtained for Target =1

Rules Obtained

Target Value	Rule #	Rule	Precision	True Positive/Total	Valid True Positive/Total	Valid Precision
1	1	gift & invitation	100.0%	115/115	23/26	88.46%
1	2	recent contribution	100.0%	89/89	26/26	94.00%
1	3	support & ~sample bequest language & ~cal & schedule	100.0%	158/158	44/51	90.48%
1	4	game & host	100.0%	121/121	30/32	91.18%
1	5	set up & request	100.0%	72/72	20/20	92.11%
1	6	email & ~cal & alumni association	100.0%	63/63	12/15	91.13%
1	7	office & ~gift planning & ~cal & alum	100.0%	72/72	14/16	91.54%
1	8	1k	100.0%	41/41	5/7	90.98%
1	9	miss	100.0%	48/48	9/10	90.65%
1	10	follow & ~cal & love	100.0%	108/108	24/27	89.73%
1	11	e-mail & ~cal & interest	100.0%	103/103	14/19	89.86%
1	12	athletics	100.0%	44/44	12/12	90.07%

Figure 9: Text rules obtained

Rules obtained for Target =0

Rules Obtained

Target Value ▲	Rule #	Rule	Precision	True Positive/Total	Valid True Positive/Total	Valid Precision
0	29	corpus	80.00%	4/5	0/3	0.00%
0	30	coe magazine	76.92%	6/10	2/6	25.00%
0	31	socal	72.73%	6/10	3/7	38.46%
0	32	gift planning	74.07%	5/8	0/4	35.71%
0	33	area & ~year	72.22%	6/9	1/2	37.50%
0	34	employment & ~back	66.04%	9/19	0/2	33.33%

Target Value

Figure 10: Text rules obtained for Target =0

Predictive Modelling (Text and numeric data)

We apply predictive modelling and use various approaches and modelling techniques to determine the winner model. We are here comparing models with only numeric variable, only text variable and a combination of numeric and text variables. In the combination model, text clusters is the input text variable. Here the target variable is binary. 1-donation made. 0-no donation made.

The Text Parsing and Text Filter have same properties as laid out previously.

In Data Partition, 80% dataset is training and 20% is validation.

Text Rule Builder has Generalization Error', 'Purity of rules' and 'Exhaustiveness' as 'low'.

Tree Method has been used in Impute node.

Text Rule Builder has Generalization Error', 'Purity of rules' and 'Exhaustiveness' as 'low'.

Misclassification rate on Validation dataset is used to assess models.

Stepwise Selection was used in Regression node.

Default settings were used in Decision Tree, AutoNeural, Dmine Regression.

Variable Selection node is Decision tree node used to select variables for Neural Network.

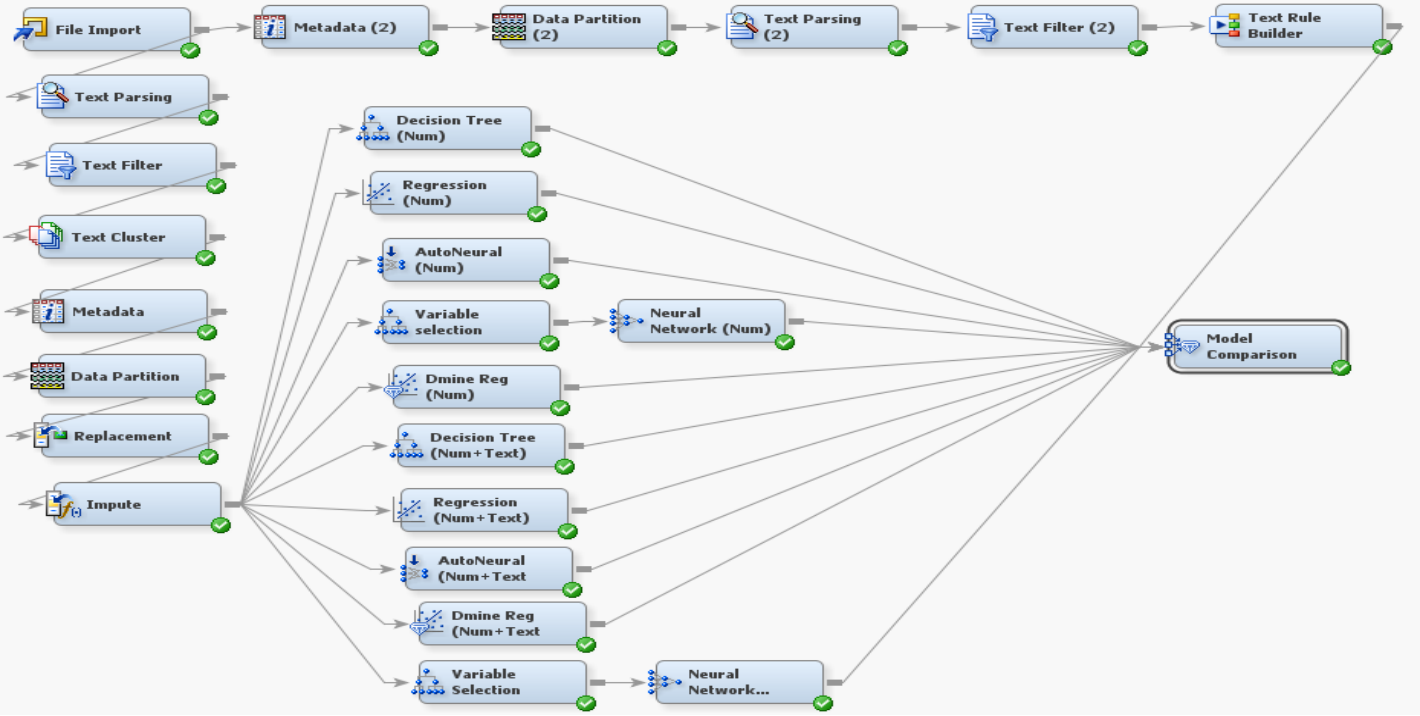


Figure 11: SAS Enterprise Miner 14.1 Screen

Selected Model	Model Description	Selection Criterion: Valid: Misclassification Rate
Y	Regression (Num+Text)	0.051876
	AutoNeural (Num+Text	0.06181
	Dmine Reg (Num+Text	0.064018
	Decision Tree (Num)	0.067329
	Regression (Num)	0.068433
	Decision Tree (Num+Text)	0.069536
	Neural Network (Num)	0.07064
	Neural Network (Num+Text)	0.07064
	AutoNeural (Num)	0.072848
	Dmine Reg (Num)	0.075055
	Text Rule Builder	0.130243

Figure 12: Validation Misclassification rate of all models.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
IMP_REP_SchoolType	10	300.9561	<.0001
IMP_bwfRating	4	5030.4485	<.0001
IMP_evertrueRating	5	1000.2063	<.0001
MaritalStatus	4	145.9427	<.0001
TextCluster_cluster_	3	148.2913	<.0001

Figure 13: Important variables of winner Regression Model.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	3.2976	3.9237	0.71	0.4007	27.047
IMP_REP_SchoolType	Agriculture	1	0.3170	0.0981	10.44	0.0012	1.373
IMP_REP_SchoolType	Arts	1	-0.2775	0.0745	13.87	0.0002	0.758
IMP_REP_SchoolType	Business	1	-0.0335	0.0802	0.17	0.6761	0.967
IMP_REP_SchoolType	Education	1	-0.0965	0.0969	0.99	0.3193	0.908
IMP_REP_SchoolType	Engineering	1	-0.1595	0.0810	3.87	0.0491	0.853
IMP_REP_SchoolType	Graduate	1	-0.3444	0.4793	0.52	0.4725	0.709
IMP_REP_SchoolType	Health Science	1	0.4757	0.1538	9.57	0.0020	1.609
IMP_REP_SchoolType	Human Science	1	-0.2568	0.1142	5.05	0.0246	0.774
IMP_REP_SchoolType	Other	1	-1.2499	0.1324	89.17	<.0001	0.287
IMP_REP_SchoolType	Unknown	1	1.4605	0.1221	143.10	<.0001	4.308
IMP_bwfRating	1	1	-5.3843	3.9237	1.88	0.1700	0.005
IMP_bwfRating	2	1	-3.6827	3.9233	0.88	0.3479	0.025
IMP_bwfRating	3	1	-0.5217	3.9234	0.02	0.8942	0.594
IMP_bwfRating	4	1	0.3698	3.9235	0.01	0.9249	1.448
IMP_evertrueRating	0	1	-0.5300	0.1090	23.66	<.0001	0.589
IMP_evertrueRating	1	1	-1.1096	0.0711	243.40	<.0001	0.330
IMP_evertrueRating	2	1	-0.3255	0.0613	28.17	<.0001	0.722
IMP_evertrueRating	3	1	0.0683	0.0567	1.45	0.2282	1.071
IMP_evertrueRating	4	1	-0.1867	0.0546	11.69	0.0006	0.830
MaritalStatus	Divorced	1	0.4630	0.1081	18.35	<.0001	1.589
MaritalStatus	Married	1	-0.0823	0.0518	2.53	0.1118	0.921
MaritalStatus	Single	1	-0.2631	0.0642	16.81	<.0001	0.769
MaritalStatus	Unknown	1	-0.8146	0.0718	128.86	<.0001	0.443
TextCluster_cluster_	1	1	0.0728	0.0455	2.56	0.1095	1.075
TextCluster_cluster_	2	1	-0.4627	0.0463	99.86	<.0001	0.630
TextCluster_cluster_	3	1	0.3981	0.0413	92.88	<.0001	1.489

Figure 14: Maximum Likelihood Estimates

Following variables increases chances of donation –

School type – Agriculture, Health Science and Unknkwon

Bwf or wealth rating =4

Evertrue score = 3

Marital Status = Divorced

Text Cluster 1 and Text Cluster 3

Odds Ratio Estimates

Effect		Point Estimate
IMP_REP_SchoolType	Agriculture vs Veterinary	1.164
IMP_REP_SchoolType	Arts vs Veterinary	0.643
IMP_REP_SchoolType	Business vs Veterinary	0.820
IMP_REP_SchoolType	Education vs Veterinary	0.770
IMP_REP_SchoolType	Engineering vs Veterinary	0.723
IMP_REP_SchoolType	Graduate vs Veterinary	0.601
IMP_REP_SchoolType	Health Science vs Veterinary	1.365
IMP_REP_SchoolType	Human Science vs Veterinary	0.656
IMP_REP_SchoolType	Other vs Veterinary	0.243
IMP_REP_SchoolType	Unknown vs Veterinary	3.653
IMP_bwfRating	1 vs 5	<0.001
IMP_bwfRating	2 vs 5	<0.001
IMP_bwfRating	3 vs 5	<0.001
IMP_bwfRating	4 vs 5	<0.001
IMP_evertrueRating	0 vs 5	0.073
IMP_evertrueRating	1 vs 5	0.041
IMP_evertrueRating	2 vs 5	0.090
IMP_evertrueRating	3 vs 5	0.133
IMP_evertrueRating	4 vs 5	0.103
MaritalStatus	Divorced vs Widowed	0.791
MaritalStatus	Married vs Widowed	0.459
MaritalStatus	Single vs Widowed	0.383
MaritalStatus	Unknown vs Widowed	0.221
TextCluster_cluster_	1 vs 4	1.084
TextCluster_cluster_	2 vs 4	0.635
TextCluster_cluster_	3 vs 4	1.501

Figure 15: Odds Ratio Estimates

Some of the key odds ratio –

The Unknown School Type has the highest odds of making donation.

Divorced has odds of 0.8 vs odds of widowed. Widowed make more donation.

TextCluster 3 (Mail +dr +letter recent +fund +contribution +donor +scholarship +member +support + student +information +note) has 1.5 times odds of making donation as compared to text cluster 4 (Stadium +host +attend +game +Sports Z +event +letter +invite amp +dean +plan next +discuss +update)

Conclusion

With an accuracy of 95%, we can predict who is going to donate based on our best model.

In fundraising world, we have often ignored textual conversations. Many non profits don't keep transcripts of conversations; those who do, don't know how to use them for better decision making. Of course textual conversations, give solicitor an idea of constituent's plans, things discussed etc., but this is not the full picture. As we have seen in the results, applying analytics to textual conversations not just gives us clusters, text topic terms and word rules but also improves our predictive modelling when used with numeric variables. The improvement seems small but difference becomes significant when we are talking about millions of alumni and billions of money raised every year. A donor makes multiple donations and missing out a few could make a difference of millions easily.

Future Scope

We can further do many other things with this research –

- We could predict how much somebody is going to donate. This could well tell us the financial ability of constituent.
- We can form clusters of constituents in terms of homogeneity.
- We could divide donations into small, medium and large donation. So now we would have nominal target rather than binary target.
- Solicitors usually in text column, write positive, neutral and negative sentiments together. Its very difficult to classify them into sentiment, since the text column has all kinds of word. If a column were to be included in foundation software that gave option to express sentiment in few words, the results would be further accurate and constituents could be easily classified into kind of sentiment.

References

http://support.sas.com/documentation/onlinedoc/txtminer/14.1/tmref.pdf

https://support.sas.com/resources/papers/proceedings14/1288-2014.pdf

http://support.sas.com/resources/papers/proceedings11/223-2011.pdf

https://www.sas.com/content/dam/SAS/en_us/doc/event/analytics-experience-2016/analyzing-sentiments-tweets-tesla-model3.pdf

Acknowledgment

I am really thankful to Dr. Goutam Chakraborty and Dr. Miriam McGaugh for guidance and support throughout the paper.

I am also thankful to University Foundation, who wishes to remain anonymous, for providing me access to database

Contact Information

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

Harsh Gupta

Oklahoma State University

Email: harsh.gupta@okstate.edu

Harsh Gupta is a graduate student currently pursuing Masters in Business Analytics at the Spears School of Business, Oklahoma State University. He has worked as a Business Analyst for 2 years in IT. He also recently completed his 13-week summer internship with University Foundation as Information Strategy Intern. He also works for School of Entrepreneurship as Graduate Research Assistant, helping startups with their marketing plans and web analytics.

Dr. Goutam Chakraborty

Oklahoma State University

Email: goutam.chakraborty@okstate.edu

Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS and OSU data mining certificate and SAS and OSU marketing analytics certificate at Oklahoma State University. He has published many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has over 25 Years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.

Dr. Miriam McGaugh

Oklahoma State University

Email: miriam.mcgaugh@okstate.edu

Dr. Miriam McGaugh is Clinical Professor in Business Analytics program at Oklahoma State University. She was a Community Health Epidemiologist at Oklahoma State Department of Health for almost 15 years. During that period she used to be a Lecturer at Spears School of Business, Oklahoma State University and played a big role in teaching and encouraging students to achieve honors like SAS® Certified Base Programmer and SAS® Certified Advance Programmer.