Paper 3472-2015

Analyzing Marine Piracy from Structured & Unstructured data using SAS® Text Miner

Raghavender Reddy Byreddy, Globe Life and Accident Insurance Company; Anvesh Reddy Minukuri, Comcast Corporation; Tejeshwar Gurram, OSU; Nitish Byri, H&R Block; Goutam Chakraborty, OSU

ABSTRACT

Approximately 80% of world trade at present uses the seaways, with around 110,000 merchant vessels and 1.25 million marine farers transporting almost 6 billion tons of goods every year. Marine piracy stands as a serious challenge to sea trade. Understanding how the pirate attacks occur is crucial in effectively countering marine piracy. Predictive modeling using the combination of textual data with numeric data provides an effective methodology to derive insights from both structured and unstructured data. 2,266 text descriptions about pirate incidents that occurred over the past seven years, from 2008 to the second quarter of 2014 were collected from the International Maritime Bureau (IMB) website. Analysis of the textual data using SAS®Enterprise Miner™ 12.3, with the help of concept links, answered questions such as: what are the arms used by pirates for attacks ,how do pirates steal the ships, how do pirates escape after the attacks, what are the reasons for occasional unsuccessful attacks? Topics are extracted from the text descriptions using a text topic node, and the varying trends of these topics are analyzed with respect to time. Using the cluster node attack descriptions are classified into different categories based on attack style and pirate behavior described by a set of terms. A target variable attack type is derived from the clusters and is combined with other structured input variables such as Ship type, Status, Region, Part of day, Part of year. Predictive model is built with attack type as a target variable and other structured data variables as input predictors. The model predicts the possible type of attack given the details of ship and travel. Thus, the results of this paper could be helpful for the shipping industry to become more aware of possible attack type for different vessel types and to devise counter-strategies in reducing the effects of piracy to crews, vessels, and cargo.

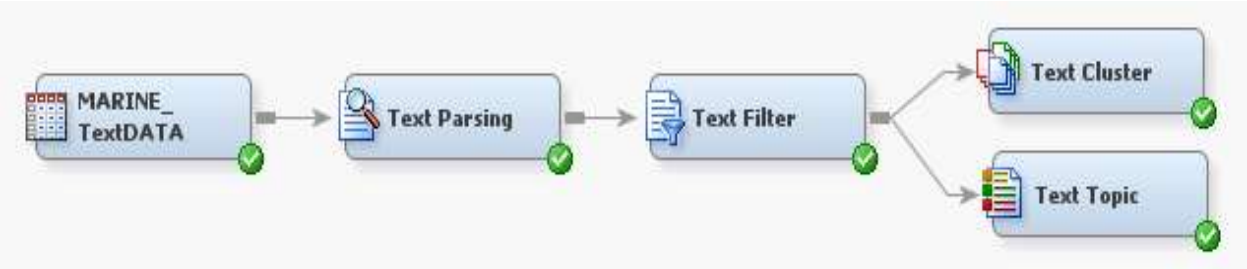
INTRODUCTION

This paper is broadly divided in to two parts. In the first part unstructured data (textual description of the pirate incidents) is analyzed using text mining techniques to derive valuable insights out of the data and also to analyze spot trends over time. In the second part structured data is combined with unstructured data and a predictive model is built to predict the type of attack. The target variable for this purpose is derived from the unstructured data using text cluster node and explained using rule based node. The relationship between the derived target variable and other structured input variables is established by building different predictive models such as Decision tree, Logistic Regression, Neural networks and the best model among them is selected based on the least validation misclassification rate.

PART I-INSIGHTS FROM UNSTRUCTURED DATA

Unstructured data (textual description of the pirate incidents) are analyzed using the text mining techniques to derive valuable insights out of the data.

SAS® Enterprise Miner 12.3 and SAS® Enterprise Guide and the following nodes of text mining were used for the analysis following the process suggested by Chakraborty, Pagolu and Garla(2014).



Error! Reference source not found.

RESULTS

Concept links

Concept link shows the association of a particular term with other terms in the document corpus document based on their co-occurrences.

The below concept links helps in answering the questions on pirate attacks.

Arms used by pirates?

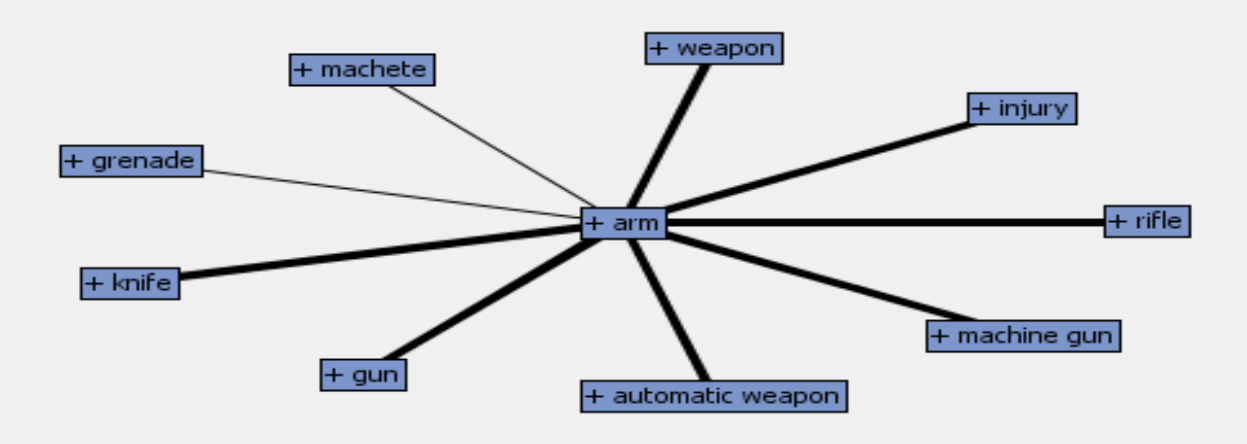


Figure 2. Concept link - “arm”

The above concept link shows the most frequent words that are associated with the term ‘’arm’’ and strength of association between them. Those words include ‘’weapon’’, ‘’gun’’, ‘’knife’’ ‘’automatic-weapon’’, ‘’machinegun’’, ‘’rifle’’, ‘’grenade’’, and ‘’machete’’.

The conclusion is that pirates are not only using the regular rifles, guns, knives and machetes for attacks, but also deadlier weapons such as machine guns, grenades and automatic weapons.

How pirates steal ships?

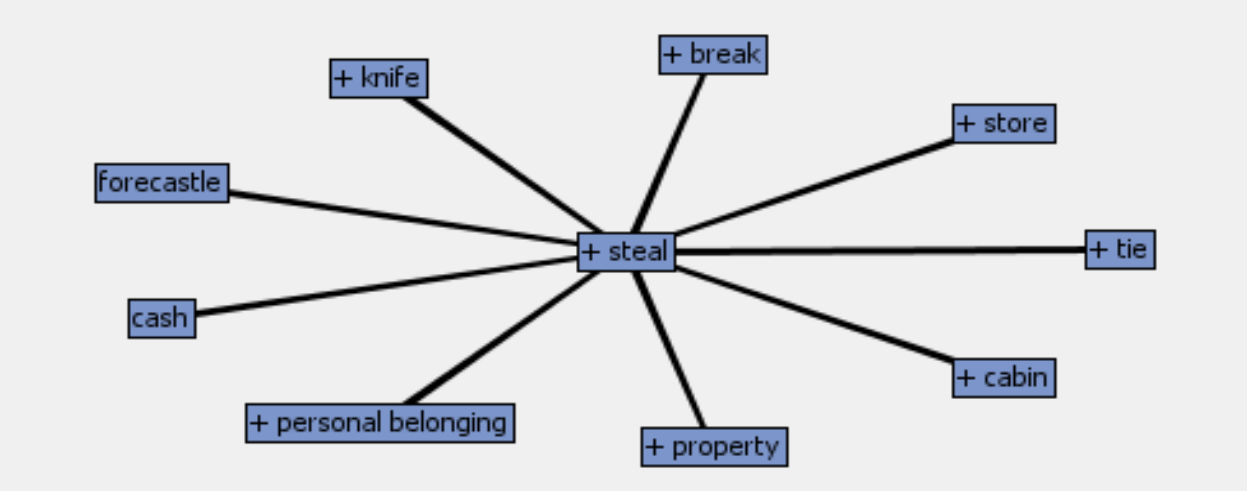


Figure 3. Concept link - “steal”

The above concept link shows the association of the word ‘’steal’’ with the terms ‘’break’’, ‘’store’’, ‘’forecastle’’, ‘’cabin’’, ‘’cash’’, ‘’tie’’, ‘’personal belonging’’, ‘’knife’’, ‘’property’’.

Pirates are mainly stealing ships by breaking in to the stores and cabins of ships by moving through the forecastle. Pirates tie down the crew during steal. In many of the incidents pirates also stole cash, personal belongings of the crew along with the ship property.

How pirates escape after attacks?

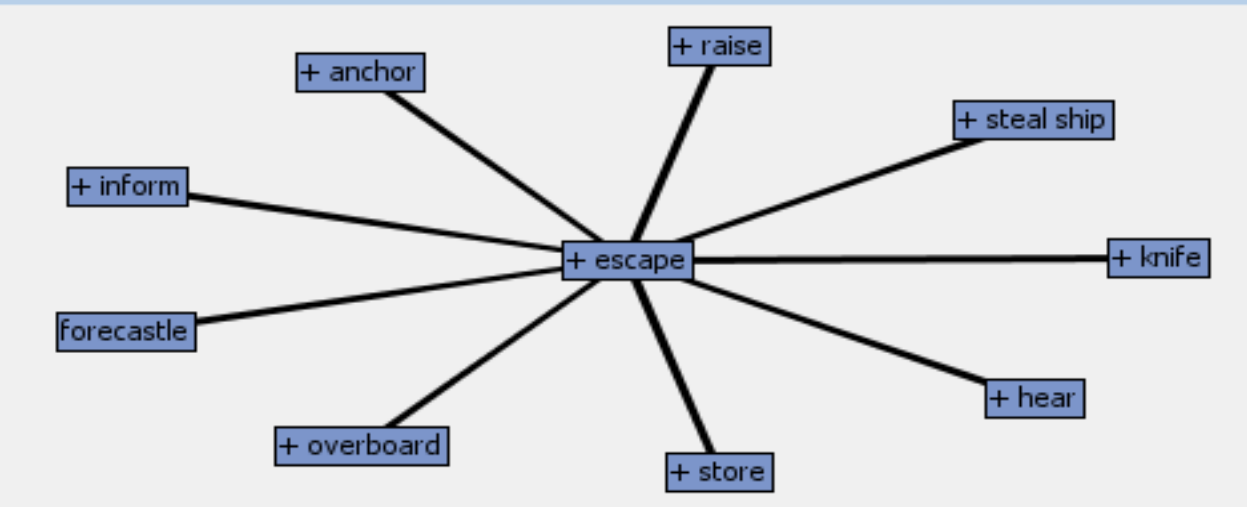


Figure 4. Concept link - “escape”

The above concept link shows the association of the word ‘’escape’’ with the terms ‘’raise’’, ‘’steal ship’’, ‘’over board’’, ‘’anchor’’, ‘’knife’’, ‘’inform’’, ‘’store’’, ‘’hear’’.

Following insights can be derived from the concept link such as pirates tend to escape when they hear the alarm sound. Pirates involved in stealing the anchored ships are usually not heavily armed, they just carry knifes and tend to escape immediately after the theft. In majority of these incidents they stole from the ship by breaking in to the stores secretly and escaped jumping overboard through forecastle. So, shipmasters have to be extremely cautious when they anchored their ships at the ports.

Reasons for unsuccessful attacks:

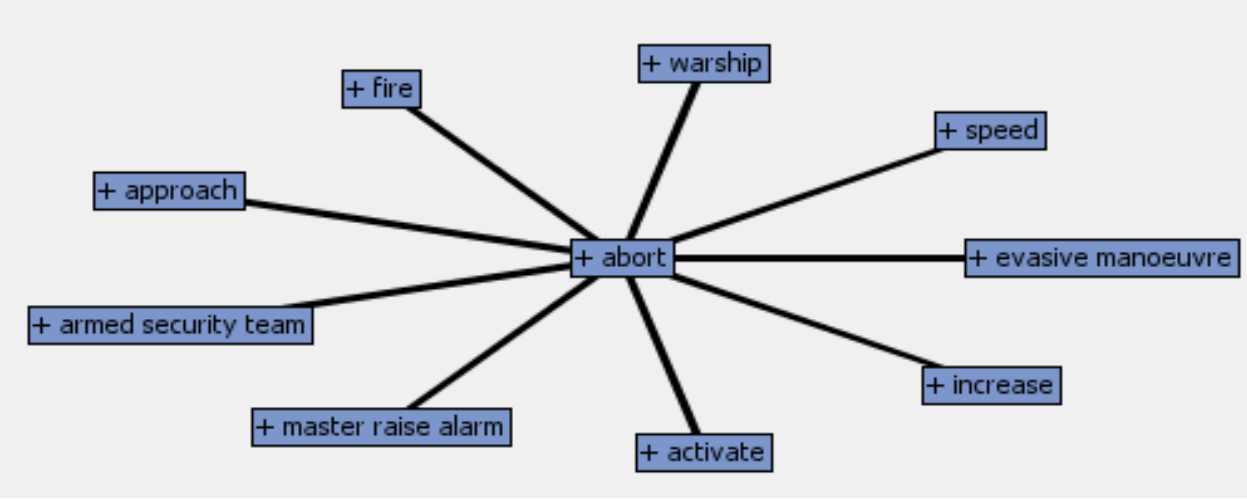


Figure 5. Concept link - “abort”

The above concept link shows the association of the word ‘’abort’’ with terms ‘’warship’’, ‘’speed’’, ‘’evasive-maneuver’’, ‘’increase’’, ‘’speed’’, ‘’activate’’, ‘’master raise alarm’’, ‘’armed security team’’, ‘’ fire’’.

From the above concept link it can be interpreted that in majority of the unsuccessful attacks the reasons that led the pirates to abort the attack are the warships employed to counter marine piracy, evasive measures such as evasive maneuver, increase speed of the ships to escape attack, activate Ship Security Alert System, fire hoses, raise alarm and armed security teams firing on the pirate skiffs approaching for attack.

Text Topic Analysis

Using text topic node 7 user topics described by a group of carefully selected terms are defined as shown below.

Topic Identifier	Text Topic	Descriptive terms
T1	Crew Alertness	'alarm' '+raise' '+alertness' '+muster'
T2	External support	'+contact' '+investigation' '+guard' 'assistance' '+helicopter' 'coastal'
T3	Hijack release	'+hijack' '+safe' '+hostage' '+release' 'ransom' '+pay'
T4	Physical attack	'+damage' '+tie' '+threaten' '+injure' '+assault' '+kill' '+beat' '+die' '+kick'
T5	Robbery	'+steal' '+store' '+property' 'cash' '+belonging' '+engine' '+padlock'
T6	Unsuccessful attack	'+abort' 'empty'
T7	evasive measures	'+chase' '+speed' '+manoeuvre' 'evasive' '+increase' '+measure' '+warship' '+activate' 'coalition' '+warning' '+anti' 'distress'

Error! Reference source not found.

Trend Analysis

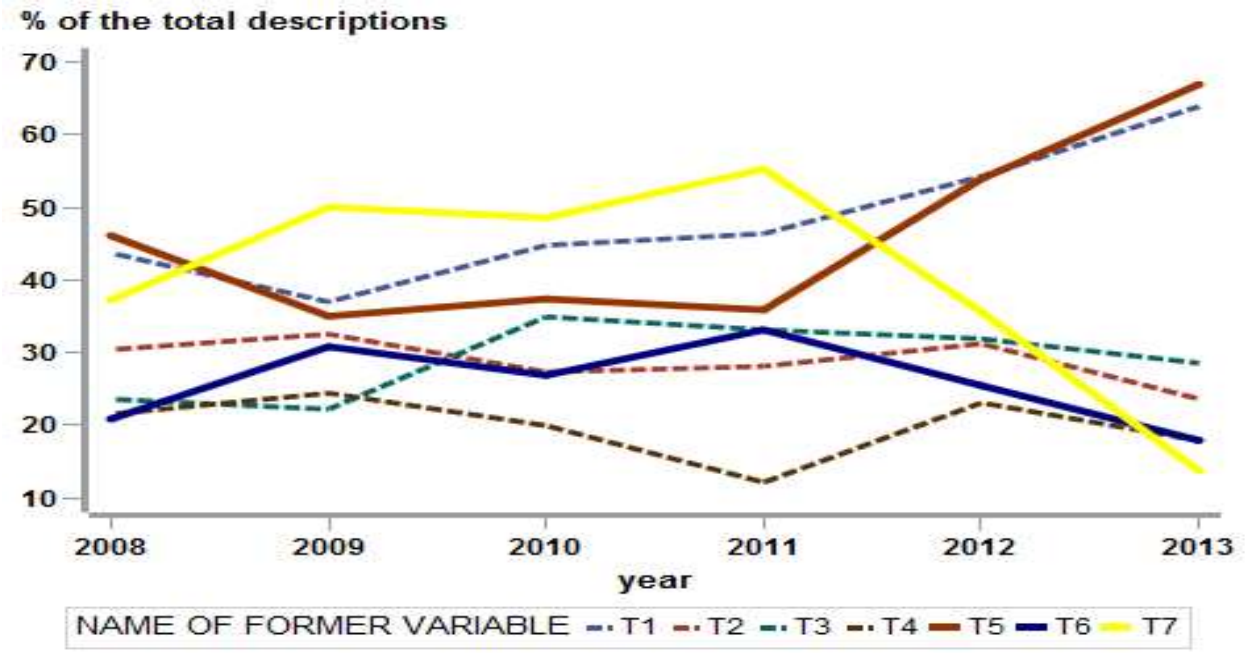
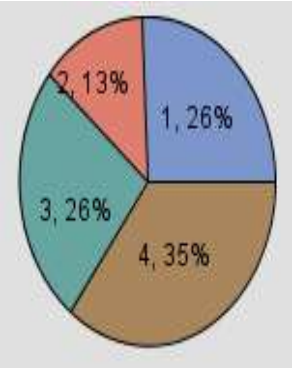


Figure 6- Line plot showing topic trends

From the above line plot it can be observed that after the year 2011, topics T7 (evasive measures) and, topicT6 (unsuccessful attacks) are trending downwards while the topic T5 (Robbery) is trending upwards indicating that the performance of evasive measures is sliding down and the success rate of the pirates has increased and the activity of robbery has increased.

PART II- PREDICTIVE MODELING COMBINING STRUCTRED & UNSTRUCTURED DATA

Text cluster node discovered four groups from the document corpus and each document (attack description) is assigned to one cluster described by the descriptive terms as shown below.



Cluster ID	Descriptive Terms
1	+'master raise alarm' +anchor +boat +duty +escape +inform +jump +knife +muster +notice +robber +steal ...
2	'safe release' +'armed pirate' +'crew member' +barge +hijack +hostage +member +pay +pirate +release +sail +tow ...
3	+'steal ship' +authority +board +escape +knife +property +robber +steal +store cash stole +anchor ...
4	+'evasive manoeuvre' +abort +approach +attack +attempt +chase +fire +gun +increase +master +measure +move ...

Figure 7- Cluster frequencies and descriptive terms

Based on the descriptive terms, the story line of each cluster may be described as shown below.

#	Descriptive terms	Story Line
1	+'master raise alarm' +anchor +boat +duty +escape +inform +jump +knife +muster +notice +robber +steal	Master raise alarm noticing robbers and muster the crew. Robbers jump in to boat and escape
2	'safe release' +'armed pirate' +'crew member' +barge +hijack +hostage +member +pay +pirate +release +sail +tow	Armed pirates hijack the ship and take crew members as hostages and release them after the payment
3	+'steal ship' +authority +board +escape +knife +property +robber +steal +store cash stole +anchor	Robbers steal the property and cash in the anchored ships and escape. Incident reported to authority
4	+'evasive maneuver' +abort +approach +attack +attempt +chase +fire +gun +increase +master +measure +move	Pirates attempt to attack the ship by chasing and firing with guns, ship takes evasive maneuvers increasing the speed to abort the attack.

Table 2. Story Line of clusters

Creating a target variable from textual descriptions

A variable called “Attack type” is derived from the above clusters which broadly classifies the type of attacks in to three levels as shown below.

Cluster no	Name
Cluster 1 & Cluster 3	Hit and Run
Cluster 2	Hijack
Cluster 4	Chase and Fire

Table3. Deriving Attack type variable

Clusters 1 and Cluster 3 are combined to form a single level as both the clusters are similar in terms of attack style. The derived variable will be used as a target variable and is combined with the other structured input data variables to build a predictive model.

Using Rule based node with Attack type as target variable and narration (textual description of the attack) as text variable the following content categorization codes are generated. These codes can be further improvised and used with the content categorization studio in categorizing the future textual descriptions automatically.

Content categorization code

F_Attacktype =Hijack::

(OR

, "ransom"

, (AND, (NOT, (OR, "masters" , "master")), (OR, "hostages" , "hostage"), (OR, "hijack" , "hijacked"), (NOT, (OR, "chase" , "chasing" , "chased")))

, (AND, (NOT, (OR, "ships" , "ship")), (OR, "tow" , "towing" , "towed"))

, (AND, (OR, "crew members" , "crew member"), (OR, "kidnapped" , "kidnap"))

, (AND, (NOT, (OR, "ships" , "ship")), (OR, "vessel" , "vessels"), (NOT, (OR, "attempt" , "attempted" , "attempts" , "attempting")), (NOT, (OR, "skiff" , "skiffs")), (OR, "fishing" , "fish"), (NOT, (OR, "raised" , "raise")))

, (AND, (NOT, (OR, "firing" , "fire" , "fires" , "fired")), (OR, "pirated" , "pirate" , "pirates"), (OR, "crew member" , "crew members"))

, (AND, (NOT, (OR, "fire" , "fired" , "fires" , "firing")), (OR, "pirates" , "pirate" , "pirated"), (NOT, (OR, "manoeuvred" , "manoeuvre" , "manoeuvres" , "manoeuvring")), (OR, "negotiation" , "negotiations"))

, (AND, (NOT, (OR, "fire" , "fired" , "fires" , "firing")), (OR, "pirates" , "pirate" , "pirated"), (NOT, (OR, "contacted" , "contact" , "contacting" , "contacts")), (NOT, (OR, "damages" , "damaged" , "damage" , "damaging")), (NOT, (OR, "manoeuvred" , "manoeuvre" , "manoeuvres" , "manoeuvring")), (NOT, (OR, "chase" , "chased" , "chasing")), (NOT, (OR, "aborting" , "aborted" , "abort")), (OR, "wear" , "wearing" , "wore"))

, (AND, (OR, "vessels" , "vessel"), "cash" , (OR, "tie" , "ties" , "tied"))

, (AND, "singapore" , (OR, "towing" , "tow" , "towed")))

F_Attacktype =Chase and Fire::

(OR

, "evasive"

, (AND, (NOT, (OR, "stole" , "steal" , "stolen" , "stealing")), "underway" , (OR, "team" , "teams"))

, (AND, (NOT, (OR, "stealing" , "stolen" , "steal" , "stole")), (OR, "pirate" , "pirates" , "pirated"), (OR, "measures" , "measure"))

, (AND, (OR, "increased" , "increasing" , "increase"))

, (AND, (OR, "fired" , "fires" , "firing" , "fire"), (OR, "skiff" , "skiffs"))

, (AND, (NOT, (OR, "stole" , "stolen" , "steal" , "stealing")), "underway" , (OR, "warships" , "warship"), (NOT, (OR, "escape" , "escaping" , "escaped")))

, (AND, (OR, "altered" , "alter" , "altering"))

, (AND, (NOT, (OR, "stole" , "stealing" , "stolen" , "steal")), (NOT, (OR, "robber" , "robbers")), "onboard" , (OR, "approach" , "approaching" , "approached" , "approaches"))

, (AND, (NOT, (OR, "steal" , "stole" , "stolen" , "stealing")), (OR, "pirate" , "pirated" , "pirates"), "coalition")

, (AND, (NOT, (OR, "crew" , "crews")), (NOT, (OR, "steal" , "stole" , "stolen" , "stealing")), (OR, "firing" , "fired" , "fires" , "fire"))

, (AND, (NOT, (OR, "board" , "boarded" , "boarding")), (NOT, (OR, "robbers" , "robber")), (OR, "abort" , "aborting" , "aborted")))

F_Attacktype =Hit and Run::

(OR

, (AND, (OR, "stores" , "store"), (NOT, "undertow"))

, (AND, (OR, "muster" , "mustered" , "mustering"))

, (AND, (NOT, "tug"), (OR, "anchor" , "anchored" , "anchoring"), (OR, "escaped" , "escape" , "escaping"))

, (AND, (NOT, "tug"), (OR, "stealing" , "stolen" , "steal" , "stole"), (NOT, "underway"), (NOT, "togo"))

, (AND, (OR, "port" , "ports"), (OR, "robbers" , "robber"))

, (AND, (OR, "spot" , "spotted"))

, "crew alertness"

, "local"

, (AND, (OR, "knife" , "knives" , "knifes"))

, (AND, (OR, "escaped" , "escape" , "escaping")))

From the above terms of the rule builder, we can see that the terms like ransom or the attacks including the hostages along with chase and fire are considered to belong to the category of attack type as Hijack ,similarly which stealing the boats and escaping in the last case can obviously lead to the category of Hit and Run. It might look quite interesting to note that the chase and fire has some terms that are common with the Hit and Run however we can see that the rules are different. The term “NOT board “is something that signifies the moving boat.

Combining Target variable with input variables

The Attack type variable (derived from unstructured data) is combined with the other structured input variables. The resultant data set is as shown below.

Role	Variable name	Variable Descriptions	Type	No of levels
Target	Attack type	Attack type derived from text descriptions	Categorical	3
Input	Narration	Textual description of the attack	Text	
Input	Shiptype	Type of the vessel	Categorical	9

Input	Status	Status of ship during attack	Categorical	2
Input	RMSDist	RMS value of the distances from the incident location to the three nearest sea ports. This value indicates at what distance from the security point the incident happened as the seaports are expected to have some surveillance and security.	Numeric	
Input	Region	Region in which the incident happened	Categorical	4
Input	Partofyear	Quarter of the year during which the attack has happened	Ordinal	4
Input	PartofDay	Part of day incident happened	Categorical	4

Table4. Data definition

Data Preparation

Data is explored using the STAT explore node and the missing values are imputed using mean imputation for interval variables and mode imputation for categorical variables. The results of the STAT explore node are as shown below

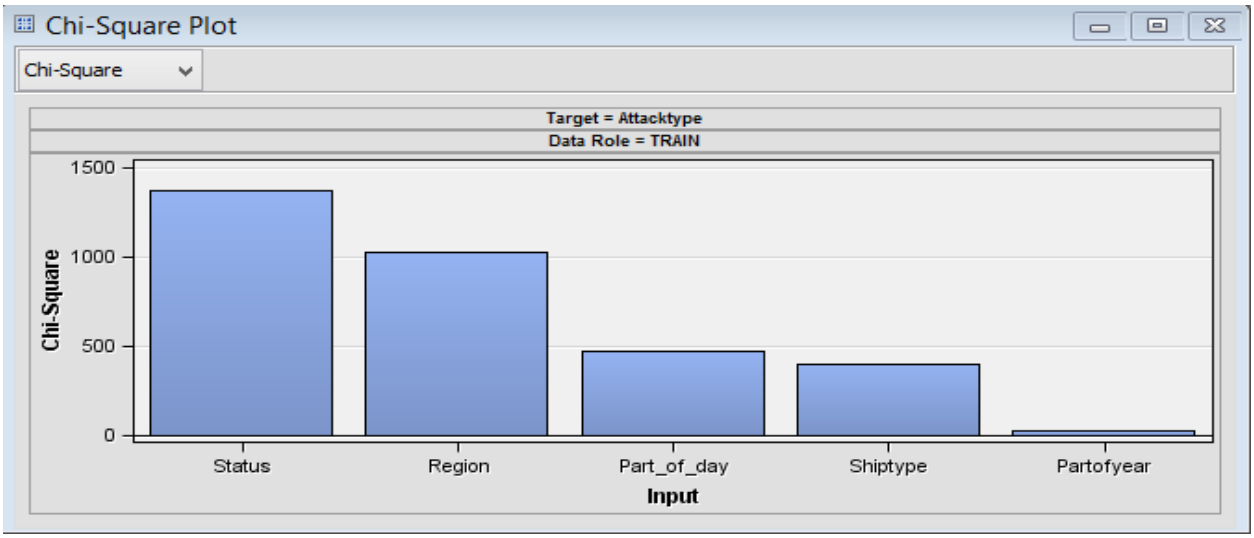


Figure 8- Chi-square Plot

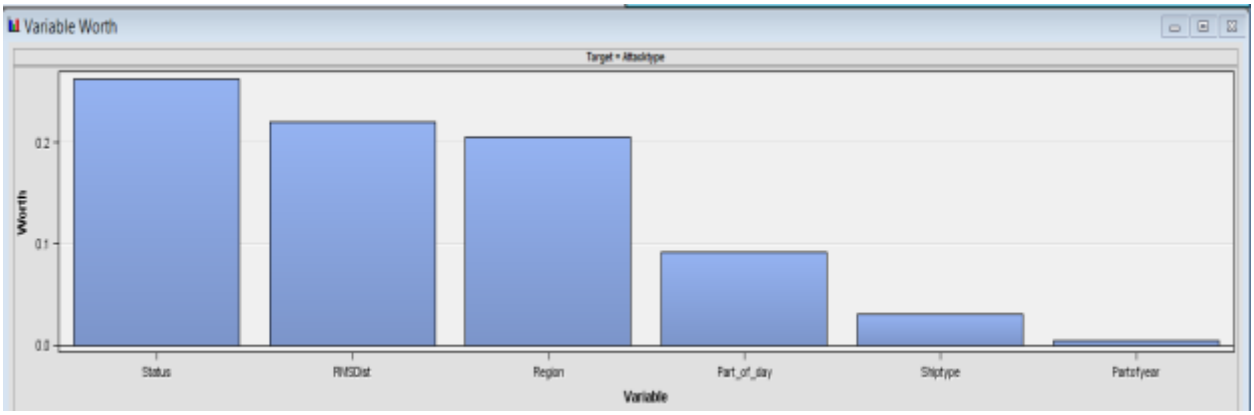


Figure 9- Variable worth

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	IMP_Part_of_day	INPUT	4	0	Midnight	40.33	Day	22.01
TRAIN	IMP_Status	INPUT	3	0	Steaming	56.91	Anchored	43.00
TRAIN	Partofyear	INPUT	4	0	2	29.04	1	27.07
TRAIN	Region	INPUT	4	0	Africa(Excluding somalia)	35.55	South East Asia	23.70
TRAIN	Shiptype	INPUT	9	0	Bulk Carrier	22.30	Product Tanker	20.42
TRAIN	Attacktype	TARGET	3	0	Hit and Run	51.24	Chase and Fire	37.10

Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	Attacktype	TARGET	Hit and Run	1094	51.2412
TRAIN	Attacktype	TARGET	Chase and Fire	792	37.0960
TRAIN	Attacktype	TARGET	Hijack	249	11.6628

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
RMSDist	INPUT	165.2812	231.2772	2135	0	0	47.06749	1223.404	1.898037	3.408225

Figure 10- Descriptive Statistics

Predictive modeling

Predictive model is built to predict the type of attack when a particular ship type is traveling through a particular route (defined by Region, RMS Dist variables) during particular time (defined by part of the year, part of the day) variables.

Predictive Models are built using Decision Trees, Logistic Regression, Neural networks with attack type as the target variable and other variables as input variables as shown below.

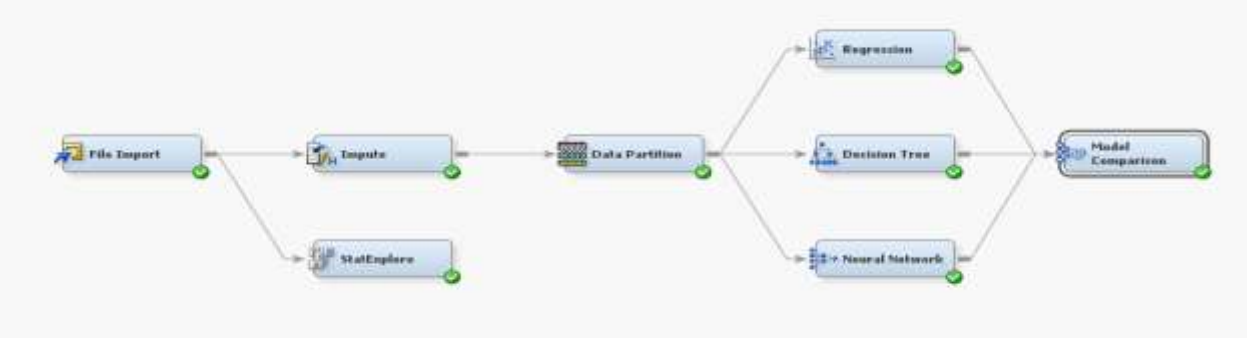


Figure11- EM Process flow diagram.

Neural Network is selected as the best model by the Model Comparison node because it has the least misclassification rate in the validation data among the three models.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate
Y	Neural	Neural	Neural Net...	Attacktype	Attacktype	0.148837	1705	0.151906
	Tree	Tree	Decision Tr...	Attacktype	Attacktype	0.160465	1705	0.157771
	Reg	Reg	Regression	Attacktype	Attacktype	0.169767	1705	0.156598

Figure 14- Model Fit Statistics

CONCLUSION

This paper illustrates an application of text analytics for generating insights into text comments and predictive modeling that combines both structured and unstructured data in an interesting context, marine piracy. The results of this paper could be helpful for the shipping industry to become more aware of the possible attack types and pirate behavior in advance. This in turn may help the shipmasters to take preventive measures and, devise counter strategies to reduce marine piracy's effects on crew, vessels and cargo.

REFERENCES

http://support.sas.com/documentation/cdl/en/emag/65762/PDF/default/emag.pdf

http://www.sas.com/en_us/software/analytics/text-miner.html

http://www.icc-ccs.org/piracy-reporting-centre

Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®

CONTACT INFORMATION

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

Raghavender Reddy Byreddy

Modeling Analyst

Globe Life and Accident Insurance Company

682-226-0480

Email: Raghavender.byreddy@okstate.edu

Raghavender Byreddy is a SAS Certified Predictive Modeler currently working as a Modeling analyst for Globe Life and Accident Insurance Company. He completed his masters in Management Information systems, SAS & OSU Data mining certificate program from Oklahoma State University in Dec 2014. His certifications include SAS® Base Programming, SAS® Certified Advance Programmer for SAS, Statistical Business Analyst using SAS9: Regression and Modeling, SAS® Certified Predictive Modeler Using SAS Enterprise Miner7, SAS® and OSU Data Mining Certificate, SAS® and OSU Predictive Modeling Certificate. He has also published posters in SCSUG conference and SAS Analytics Conference 2014.

Anvesh Reddy Minukuri

Comcast Corporation

405-780-5346

Anvesh_Reddy.minukuri@okstate.edu

Anvesh Reddy Minukuri is working as a Senior Analyst, Data Science at Comcast Corporation. Prior to this he pursued masters in management information systems at Oklahoma State University. He has completed OSU and SAS Data mining, Advanced Analytics Certificates. His certifications include SAS® Base Programming, SAS® Certified Advance Programmer for SAS, Statistical Business Analyst using SAS9: Regression and Modeling, SAS® Certified Predictive Modeler Using SAS Enterprise Miner7, SAS® and OSU Data Mining Certificate, SAS® and OSU Predictive Modeling Certificate. He has also published poster in SCSUG conference and SAS Analytics Conference 2014

Tejeshwar Gurram

Oklahoma State University

207-939-6286

Tejeshwar.gurram@okstate.edu

Tejeshwar Gurram is graduate student with strong background in SAS and solid work experience in Java/J2EE and is expecting to pursue the career in the field of data science and modeling. Besides having Various SAS and Java certifications, he also has working knowledge on various analytical and Big data Projects. He believe that he being Graduate teaching assistant for big data analytics has helped him gain profound knowledge on various Big Data technologies including Teradata Aster and IBM info sphere streams etc.

Nitish Byri

H&R Block

405-762-2971

Nitish.byri@okstate.edu

Nitish Byri is a SAS Certified Business Analyst working as a Pricing Data Analyst at H&R Block. He completed his Masters in Management Information Systems at Oklahoma State University and is very passionate about Analytics. He has a rich academic experience at OSU in data mining and also has an expertise in working with tools like SAS® Enterprise Guide, Enterprise Miner and SAS® 9.0. His list of certifications includes SAS and OSU Data Mining Certificate, SAS® Certified Advance Programmer and SAS® Certified Predictive Modeler.

Dr. Goutam Chakraborty

Oklahoma State University

Stillwater, OK, 74078

E-mail: 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. His research has been published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 Years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.