Paper 3508-2015

Using Text from Repair Tickets of a Truck Manufacturing Company to Predict Factors that Contribute to Truck Downtime

Ayush Priyadarshi and Dr. Goutam Chakraborty, Oklahoma State University

ABSTRACT

In this era of big data, the use of text analytics to discover insights is rapidly gaining popularity in businesses. On average, more than 80 percent of the data in enterprises may be unstructured. Text analytics can help discover key insights and extract useful topics and terms from the unstructured data. The objective of this paper is to build a model using textual data that predicts the factors that contribute to downtime of a truck. This research analyzes the data of over 200,000 repair tickets of a leading truck manufacturing company. After the terms were grouped into ten key topics using text topic node of SAS® Text Miner, a regression model was built using these topics to predict truck downtime, the target variable. Data was split into training and validation for developing the predictive models. Knowledge of the factors contributing to downtime and their associations helped the organization to streamline their repair process and improve customer satisfaction.

INTRODUCTION

Trucks can break down on the road due to number of reasons. Whenever a truck breaks down, it is usually brought back to the dealer for repair. Downtime of a truck is the difference between the repair start date and the date when the truck is ready to run on road. In simple terms, it is just the number of days truck is not available for service which equals the time the dealer takes to repair the truck.

Downtime of a truck = Truck repair end date - Truck repair start date

Delay in repairs can lead to customer dissatisfaction which in turn may negatively impact the reputation of the company. Therefore, it is important to understand the types of repairs and determine how those influence downtime of a truck. Here, we illustrate the importance of analyzing textual data and using it for predictive modeling. In this paper, we will analyze the text content in the repair tickets data and identify the factors that impact truck downtime. Text in repair ticket is analyzed using various nodes in the text mining tab of SAS® Enterprise Miner 12.3.

DATA PREPARATION

The data for this paper was collected from a leading truck manufacturing company who wishes to remain anonymous. The data set consists of records of one year from Jan 2013 to Dec 2013 with about 200,000+obervations.

The table below illustrates the variables available in the repair ticket dataset. The average truck downtime is10.23 days:

Variable Name	Level	Description
Unique ID	ID	This field represents the ticket number
Operation	Text	This variable describes the problem and the solution that was performed for the particular ticket.
RO_START_DATE	Date	This represents date the ticket was created.

RO_END_DATE	Date	This field provides the date the ticket was closed.
Truck_Downtime	Interval	It is the difference between ticket start and end date.
RO_REPAIR_AT_SEL_DLR	Binary	This field indicates whether ticket was repaired at selling dealer or not. 1 = Yes, 0 = No.
RO_IN_WARRANTY	Binary	This field indicates whether the truck was repaired in warranty or not. 1 = Yes, 0 = No.

Table 1. Data dictionary for dealer ticket data

Summary Statistics

Results

The MEANS Procedure

Analysis Variable : TRUCK_DOWNTIME				
Mean	Std Dev	Minimum	Maximum	N
10.2382213	10.2946223	0	302.0000000	212562

Fig 1. Summary statistics of downtime

Data had many outliers with downtime as high as 302 days. Upon investigation of some of the outliers, we found following types of texts:

“Parts dept didn’t had one in stock”

“Had to take one from a new truck.”

“Truck did not return”

“Customer will repair at their shop”

Clearly, some of these long repair times have been because some particular parts of the truck used in repair was back ordered or the customer took the truck back and didn’t come back. For the purpose of this paper we will restrict our analysis to repair tickets data only and ignore situations that did not involve any actual repair. In addition, our analysis would not control for vehicle characteristics such as age, other dealer factors, parts back order and geographical factors. There is a separate study going on in the company taking in consideration all these factors.

METHODOLOGY

TEXT PARSING

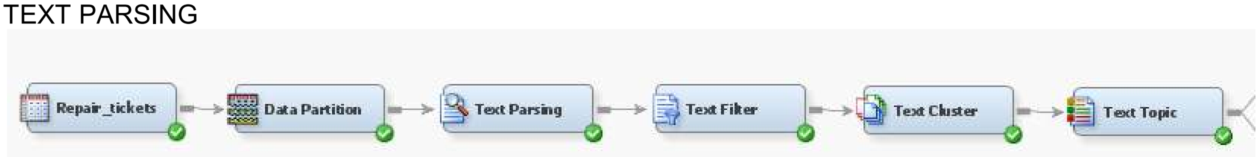

Fig 2 Enterprise Miner process flow

Once the data is imported, we use the data partition node in SAS Enterprise Miner to split the data into training and validation in the ratio of 60:40 respectively. Next, SAS® Text Parsing node is used to break down complete text into small tokens. Some of the properties that have been changed in the properties panel of Text Parsing node are:

- Detect Different Parts of Speech is turned off.
- “Find entities” is turned to Standard so to detect names of products.
- In the Ignore Types of Entities we have checked everything except Product and Prop_Misc as we need the names of product but can ignore other entities such as organization, currency, phone, etc.
- In Ignore parts of Speech we have added abbreviations, prop and number.

Terms

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
c ...	Noun	Alpha	194788	73816	N		30	1
+ replace ...	Verb	Alpha	119791	47989	Y	+	70	2
+ engine ...	Noun	Alpha	102832	39116	Y	+	11	3
+ be ...	Verb	Alpha	107755	38167	N	+	629	4
+ check ...	Verb	Alpha	77388	35167	Y	+	26	5
w ...	Miscellane...	Entity	87711	35109	Y		5	6
+ find ...	Verb	Alpha	63019	32113	Y	+	1364	7
+ install ...	Verb	Alpha	80882	31350	Y	+	1021	8
found ...	Adj	Alpha	56043	30894	Y		16	9
+ perform ...	Verb	Alpha	56925	29741	Y	+	585	10
+ repair ...	Noun	Alpha	57270	29260	Y	+	25	11
no ...	Adv	Alpha	52777	29232	N		248	12
+ code ...	Noun	Alpha	78662	29148	Y	+	1175	13
+ new ...	Adj	Alpha	63624	28459	N	+	670	14
+ truck ...	Noun	Alpha	62427	28099	Y	+	443	15
+ have ...	Verb	Alpha	54885	27447	N	+	583	16
+ not ...	Adv	Alpha	48907	26165	N	+	116	17
+ customer ...	Noun	Alpha	50325	25794	Y	+	543	18
check ...	Adj	Alpha	52432	25421	Y		28	19
+ test ...	Noun	Alpha	45494	25301	Y	+	363	20

Table 2. Text parsing results

We can see from the above table that some of the terms with highest frequency are “replace”, “engine”, “check”, “repair”, “repair”, “truck”, etc. which are obvious because it is analyzing repair ticket data.

TEXT FILTER

EMWS1.TextFilter2_spellDS

	Parent # Docs	Term	# Docs	Parent	Role	Parent Role	Min Distance	Dictionary
83	9191.0	vihicle	1.0	vehicle	PROP_MISC		14.0	
84	9191.0	v ehicle	2.0	vehicle	PROP_MISC		14.0	
85	9191.0	ve hicle	1.0	vehicle	PROP_MISC		14.0	
86	9191.0	vechicle	1.0	vehicle			7.0	
87	9191.0	vehcile	5.0	vehicle	PROP_MISC		7.0	
88	9191.0	vehcle	1.0	vehicle			8.0	
89	9191.0	vehcle	2.0	vehicle	PROP_MISC		8.0	
90	9191.0	vehi cle	1.0	vehicle	PROP_MISC		14.0	
91	9191.0	vehic le	1.0	vehicle	PROP_MISC		14.0	
92	9191.0	vehice	1.0	vehicle	PROP_MISC		8.0	
93	9191.0	vehicl	1.0	vehicle	PROP_MISC		5.0	
94	9191.0	vehicl e	1.0	vehicle	PROP_MISC		14.0	
95	9191.0	vehicle	3.0	vehicle	PROP_MISC		0.0	
96	9191.0	vehicvle	1.0	vehicle	PROP_MISC		7.0	
97	9191.0	vehilce	3.0	vehicle	PROP_MISC		7.0	
98	9191.0	vehocle	1.0	vehicle	PROP_MISC		14.0	
99	9191.0	vevicle	1.0	vehicle	PROP_MISC		14.0	
100	9191.0	vhicle	1.0	vehicle	PROP_MISC		8.0	

Table 3. Spell check results of term "vehicle"

Text filter node is used to simplify token used in the analysis. We used the default options of this node. In addition, we enabled the "Check Spelling" option in the properties panel under Text Filter node. An example of correction of the wrong spellings of the term "vehicle" is shown in Table 3. We also created some custom synonyms to treat groups of terms that have similar meanings.

CONCEPT LINKS

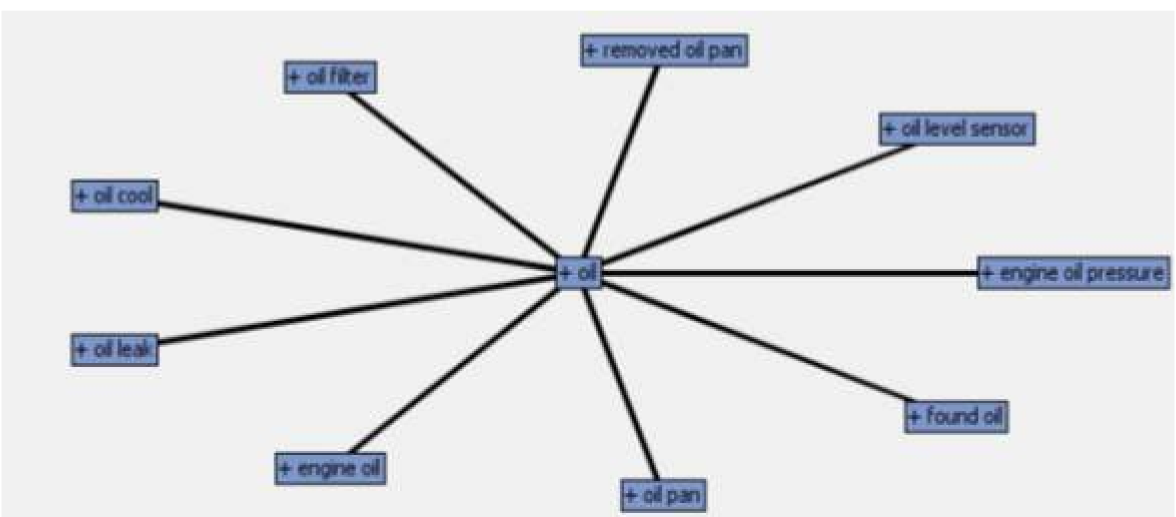

Fig 3.1 Concept link of the term "oil"

Concept links are great tools to understand association among terms in a corpus of documents. In high number of cases the system generated multiple codes for low oil level or due to oil leakage. As shown via the concept links, the term Oil is strongly associated with other terms such as “oil leak”, “found oil” and “oil level sensors”. Problem of oil leaking is one of the major problems associated with oil pressure and oil level sensors.

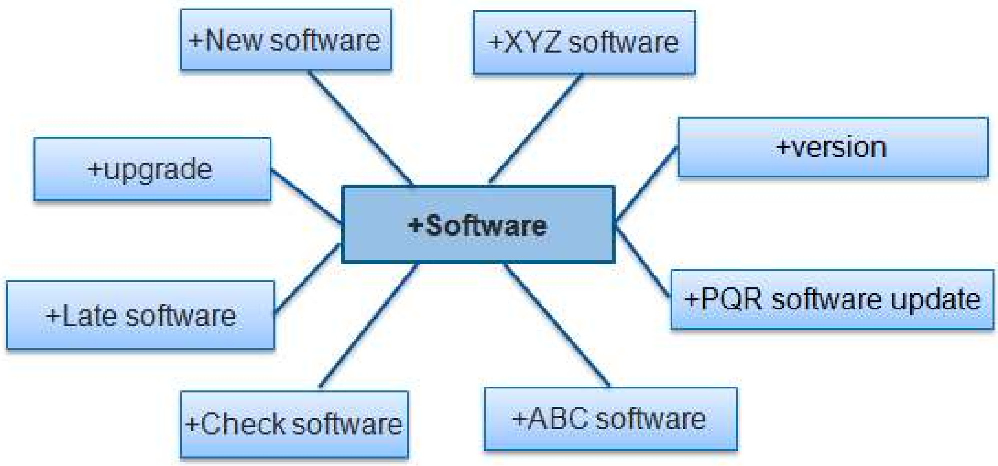

Fig 3.2 Concept link of the term “software”

Software update was identified as a critical problem in the analysis. Due to NDA restrictions, various software levels in the above concept link have been re-coded as ABC, XYZ and PQR. Most of the technicians faced issues because the trucks didn’t have the updated version of ABC, PQR or XYZ software levels for the control unit. We observe that the term “software” is strongly associated with terms such as “new software”, “check software”, “upgrade”, “version”, “ABC software”, “XYZ software” and “PQR software update”.

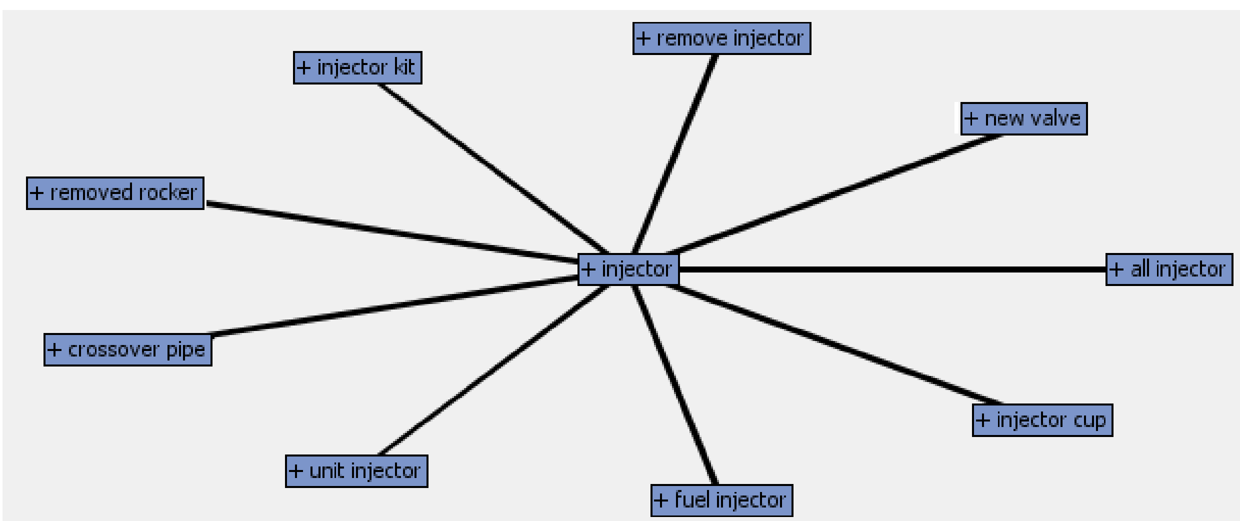

Fig 3.3 Concept link of the term “injector”

The above concept link of the term “injector” provides us critical information that whenever there is an “injector” problem then “Compressor crossover pipe”, “valve”, “rocker” and “injector cup” are also replaced. It also illustrates that injector problem is mostly associated with “fuel injector” and “unit injector”.

TEXT CLUSTERS AND TEXT TOPICS

After data filtering, we grouped the documents into clusters using the Text Cluster node. We used the Expectation Maximization algorithm for clustering. Using default options of Text Cluster node weinitially generateda 15 cluster solution. Looking at the terms that describe these clusters, we found many overlapping terms. To improve the clarity of the cluster solutions, we forced SAS EM to arrive at a 6 cluster solution. In the 6cluster solution, distance between clusters indicates that clusters are well differentiated with each other.

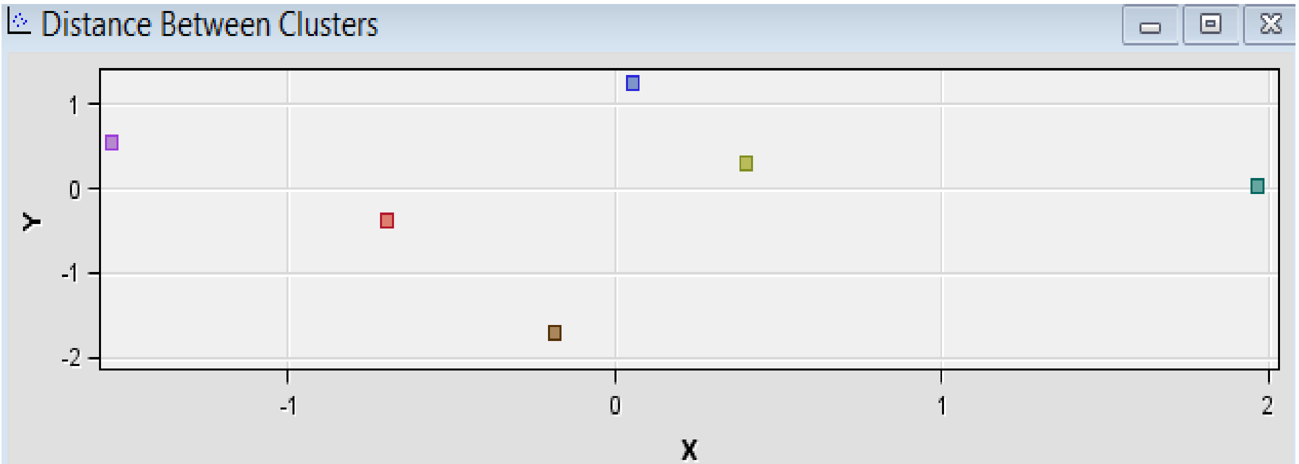

Fig 4 Distance between clusters

Cluster ID	Descriptive Terms	Count (Percent)	Meaningful themes
1	check +clutch+ transmission +remove +fluid +cylinder + replace +fault + cool	4,317(14%)	Deals with transmission related issues and involves clutch repair
2	repair +cup +injector +code +coolant +cover +valve +filter +fuel +hook +pipe +replace +pull	5,243(17%)	Describes issues relating to Injectors and the parts associated with its repair such as valve, cup, etc
3	replace +sensor +diagnostic +charge +engine +fault +filter +oil +perform +pressure +update	4,626(15%)	Describes truck breakdown due to sensor issues and the fault codes generated for the sensor failure.
4	clean +coolant +back +cool +clear +code +oil +diagnostic +leak +fault +filter +find	7,402(24%)	Diagnosing the problem, coolant leak and filter problems. Involves cleaning the engine because of the coolant and oil leaks.
5	update +software +program +check +ABC software +XYZ software +perform +diagnostic	3,392(11%)	It deals with problems in diagnostics due to outdated software level.
6	adjust +air +axle +brake +front +inspection +install +light +rear +repair +right +steer +wheel	6,477(21)%	Deals with problems such as front brake repair, light and steering wheel issues

Table 5 Clusters and terms that describe them

Text Topic node is used to extract key topics from the document. A topic is a collection of terms that represent a common theme.We have changed the “number of multi term topics” from the default option of 25 to 10 to attain better clarity in topics.

Topic ID	Topic	Meaningful themes
1	clutch, +transmission, +cylinder, +oil, +gear	Illustrates transmission related issues and involves clutch repair
2	sensor, +code, +fault, +engine, +pressure	Describes truck breakdown due to sensor issues and the fault codes generated for sensor failure.
3	injector, +cup, +fuel, +valve, +rocker	It includes issues relating to Injectors and the parts associated with its repair such as valve, cup, etc
4	coolant, +leak, +cool, +pipe, +turbo	Describes coolant leak and problems associated with turbochargers
5	wheel, +axle, +rear, +front, +brake, +steer	Deals with problems such as front brake repair, axle and steering wheel issues
6	def, +regen, +pump, +derate	It involves regeneration related repairs in trucks.
7	update, +software, +perform, +ABC, +XYZ software	It deals with problems in diagnostic due to an outdated software level of control unit.
8	wiper, +motor, +gear, +windshield, +recall	Illustrates windshield and wiper motor related repairs.
9	fuel, +filter, +oil, +inspection, +leak	Problems associated with inspection, fuel filter and oil leak in the engine.
10	wire, +light, +truck, +check, +fuse, +harness	Describes problems dealing with light, fuse wire and harness.

Table 6. Topics identified from Text Topic node

PREDICTIVE MODELING

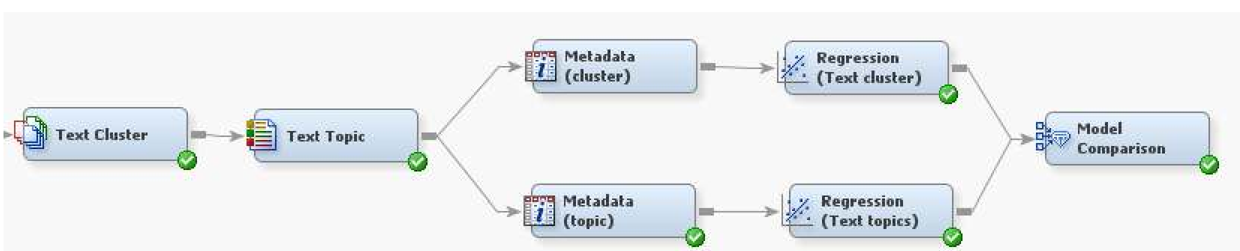

Fig 6. Enterprise Miner process flow

The cluster membership from the cluster node and/or the topics extracted using Text Topic node can also be used inputs in a predictive model. We used both of these types of variables as potential inputs in two different regression models to predict truck downtime. In the properties panel, we use “stepwise” as the Selection Model and “Validation error” as the selection criteria.

Model comparison node is used to compare the performance of the two regression models. The model selection node selects the Text Topics Regression model over the Text Cluster Regression model on the basis of lower average square error in the validation data.

Fit Statistics

Selected Model	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Average Squared Error
Y	Reg	Regression (Text topics)	Truck_Downtime	63.97175
	Reg2	Regression (Text cluster)	Truck_Downtime	66.54128

Table 7. Model Comparison

In the selected regression model, we find following factors increase truck downtime:

1) Transmission and clutch problems

2) Sensor failure

3) Injector repair

4) Issues with Turbo chargers and coolant leak

5) Software version outdated

Our model is limited by the constraints of non-availability of vehicle characteristics such as age, other dealer factors, parts back order and geographical factors. Efforts are currently under progress to merge those datasets to the repair ticket data. Addition of more numerical data such as these in our model would enhance the model accuracy significantly.

REFERENCES

1) Improving Customer Loyalty Program through Text Mining of Customers' Comments by MaheshwarNareddy and Goutam Chakraborty, Oklahoma State University

2) Replace Manual Coding of Customer Survey Comments with Text Mining: A Story of Discovery with Text as Data in the Public Sector by Jared Prins, Alberta Tourism

3) Classification of Customers' Textual Responses via Application of Topic Mining by Anil Kumar Pantangi, Goutam Chakraborty, Oklahoma State University

4) Opinion Mining and Geo-Positioning of Textual Feedback from Professional Drivers Mantosh Kumar Sarkar, Goutam Chakraborty, Oklahoma State University.

5) Getting Started with SAS® Text Miner 4.2, SAS Institute, Cary

ACKNOWLEDGMENTS

This paper utilized data from the leading truck manufacturing company. We are thankful to the company who preferred to stay anonymous and provided us the opportunity to perform analysis on the data for the purpose of the project.

CONTACT INFORMATION

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

Ayush Priyadarshi,
Oklahoma State University
Stillwater, OK, 74075
Work Phone: (405)780-5298
Email: ayushp@okstate.edu
Ayush Priyadarshi is a Graduate student enrolled in Management Information Systems at Spears School of Business, Oklahoma State University (OSU), Stillwater. He has over three years work experience in data analytics and Information Technology. He is a Base SAS® 9 certified professional, a certified SAS Statistical Business Analyst, JMP Software data exploration certified and holds the SAS and OSU Data Mining certificate. He has previously given three poster presentations at the SAS Analytics Conference 2014 and is also the author of another paper at SAS Global Forum 2015 named “Forecasting bike rental demand using SAS® Forecast Studio”.

Dr. Goutam Chakraborty
Oklahoma State University
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 in 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®.