AN ADVANCED ANALYTICS APPROACH FOR RECALL IN AUTOMOBILE INDUSTRY

SAS Global Forum 2017

Prathap Maniyur
Prathap.maniyur@fractalanalytics.com

Prashant Nayak
Prashdnayak@gmail.com

Mansi Bhat
Bhat.r.mansi@gmail.com
Contents

Background Research: .......................................................... 3
General Motors Recall: ......................................................... 4
Analytics Workflow: ............................................................ 7
  a. Data Collection Source: .................................................. 8
  b. Variable selection: .......................................................... 11
  c. Text parsing/Text Filter: .................................................. 9
  d. Cluster Analysis: ............................................................ 11
Models Building: ............................................................... 12
ROC Analysis: ................................................................. 13
Conclusion: ................................................................. 14
References: ................................................................. 17
Background Research:

A recall typically happens after discovering safety issues or product defects that might endanger the customer and might put the seller at a legal risk and financial risk. Apart from this there is risk to endangering the corporate image of the company. Every country’s consumer protection laws have specific requirements with respect to product recall. Automobiles have one of the highest number of recalls and it is one of the largest purchases for the consumer after their house. A recall is issued in the automobile industry when the seller or NHTSA determines that the car creates additional safety risk or failure to adhere to standards. The federal government issued a large number of recalls in 2014; close to 900 separate recalls affected 51 million vehicles nationwide. The ones in our memories are Toyota’s problem with sudden acceleration, General Motors’ faulty ignition switches, and Volkswagen’s emissions tests cheating. This study helps automobile manufacturers understand customers who are talking about defects in their cars and to be proactive in recalling the product at the right time before the Government acts.

Source: National Highway Traffic Safety Association
Tesla lead the way in recall proactiveness, as the manufacturer was the best at identifying problems with its cars and initiating recalls within a three-year period. It scored at 100% in manufacturer-initiated recall campaigns, which exceeds the industry average of 46.1%. This means that Tesla has strong internal systems in place to identify issues even before the NHTSA does.

In contrast, Ford had the lowest manufacturer-initiated recall campaign rate of 29.6%. This was 16.5% below the industry average. NHTSA investigations accounted for over 70% of the recalls that Ford issued in 31 years.²

### General Motors Recall:

GM had one of the biggest recalls in 2014 which involved cars such as Chevrolet Malibu, Chevrolet Impala, Monte Carlo, Cadillac CTS and Cadillac SRX.³ In May 2014, GM had recalled more cars and trucks in 2014 than it sold in the last five years since it filed for bankruptcy. The last five years were good for “new GM” with record profits and gaining market share. The recall surge was the results of new standards at GM where the calls are recalled more quickly as the problems emerge. 30 million cars were recalled worldwide and paid compensation was done for 124 deaths. The recall was due to the ignition switch problem which causes the engine to shut down while driving. At that point of time the web and social media analysis would have helped GM to anticipate the recall problems better. Towards this end, the forums and social media were researched for the GM cars. The GM cars had a lot of issues discussion on social media.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Manufacturer</th>
<th># Cars Recalled</th>
<th># Cars Sold</th>
<th>Recall Rate per 1,000 Cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Porsche</td>
<td>599,563</td>
<td>739,812</td>
<td>511</td>
</tr>
<tr>
<td>2</td>
<td>Mercedes-Benz</td>
<td>3,664,182</td>
<td>5,874,888</td>
<td>624</td>
</tr>
<tr>
<td>3</td>
<td>Kia</td>
<td>5,169,239</td>
<td>5,557,319</td>
<td>788</td>
</tr>
<tr>
<td>4</td>
<td>Tesla</td>
<td>95,184</td>
<td>91,046</td>
<td>998</td>
</tr>
<tr>
<td>5</td>
<td>Mazda</td>
<td>8,783,819</td>
<td>9,201,683</td>
<td>955</td>
</tr>
<tr>
<td>6</td>
<td>General Motors</td>
<td>129,215,450</td>
<td>134,895,276</td>
<td>958</td>
</tr>
<tr>
<td>7</td>
<td>Subaru</td>
<td>5,648,076</td>
<td>6,749,209</td>
<td>985</td>
</tr>
<tr>
<td>8</td>
<td>Toyota</td>
<td>92,614,717</td>
<td>51,166,734</td>
<td>1,026</td>
</tr>
<tr>
<td>9</td>
<td>Nissan</td>
<td>28,771,128</td>
<td>27,705,771</td>
<td>1,038</td>
</tr>
<tr>
<td>10</td>
<td>Jaguar Land Rover</td>
<td>1,709,017</td>
<td>1,601,206</td>
<td>1,067</td>
</tr>
<tr>
<td>11</td>
<td>Mitsubishi</td>
<td>5,418,810</td>
<td>4,973,757</td>
<td>1,089</td>
</tr>
<tr>
<td></td>
<td><strong>Industry Average</strong></td>
<td><strong>527,406,265</strong></td>
<td><strong>472,971,556</strong></td>
<td><strong>1,115</strong></td>
</tr>
</tbody>
</table>

² **Image source:** iSeeCars.com

² In contrast, Ford had the lowest manufacturer-initiated recall campaign rate of 29.6%. This was 16.5% below the industry average. NHTSA investigations accounted for over 70% of the recalls that Ford issued in 31 years.

³ **General Motors Recall:**

GM had one of the biggest recalls in 2014 which involved cars such as Chevrolet Malibu, Chevrolet Impala, Monte Carlo, Cadillac CTS and Cadillac SRX.³ In May 2014, GM had recalled more cars and trucks in 2014 than it sold in the last five years since it filed for bankruptcy. The last five years were good for “new GM” with record profits and gaining market share. The recall surge was the results of new standards at GM where the calls are recalled more quickly as the problems emerge. 30 million cars were recalled worldwide and paid compensation was done for 124 deaths. The recall was due to the ignition switch problem which causes the engine to shut down while driving. At that point of time the web and social media analysis would have helped GM to anticipate the recall problems better. Towards this end, the forums and social media were researched for the GM cars. The GM cars had a lot of issues discussion on social media.
Over the last few years hundreds of GM car owners have submitted their problems to CarComplaints.com, a completely free website dedicated to exposing car trouble spots based on the complaints of real car owners. There's no better way to accurately assess car problems then to get unfiltered feedback directly from those who have driven the cars and experienced the problems first-hand. GM had problems with its ignition. The ignition switch would move out of the “run” or “on” position, causing a partial loss of electrical power and the engine turning off, according to GM. The risk increases if a driver’s key ring is carrying added weight or if the vehicle encounters rough road conditions. When the ignition switch is not in the run position, the airbags may not deploy if the vehicle is involved in a crash. Post the ignition switch recall there was 1-2 percentage point decline in the average used car price.
The google trend showed that there was an increase in car recall search term in year 2013.

To understand the problem better, a word cloud was created to see if it could explain the GM recall problem better but nothing concrete came up. The cause for recall was still unclear and it warranted further deep dive.

Since the companies might spend huge amount of time and resources, wanted to understand if it is possible to build a predictive approach for recall. Hence an
analytic approach was designed to check if the prediction can happen through finding the right patterns in complaints to predict car problems and avoid damage.

**Analytics Workflow:**
The web was researched to find important websites and forums where people complained about car problems. The websites and NHTSA complaints were scrapped to collect the complaints. The unstructured data such as customer comments or feedback can enhance the power of existing predictive models. The numeric data such as mileage of the cars where the issue is reported, severity issue rating and the cost incurred for the repair is captured in the numeric and the comments are captured as text.

The Text Parsing node not only enables to gather statistical data about the terms in a document collection but also enables to modify the output set of parsed terms by dropping terms that are a certain part of speech, type of entity, or attribute. SAS® Text Miner can generate SVD (Singular value decomposition) units from text documents which is a vectorial representation of terms in documents. These SVDs when used as additional inputs along with the existing structured input variables often prove to capture the response better. Different models with and without complaints were designed. Few models were given with just numeric and few were analysed with unstructured data. Finally the models were compared to find the best model which could have predicted and helped us analyse car problems and recall. We demonstrate below how SVDs or the text cluster components along with numeric attributes help us predict recall in automobile industry.

Below is the social listening and predictive analytics workflow which is designed to check if we can predict recall by collecting complaints and attributes of people and car at the time of complaint.

To conduct the exercise for the data analysis three steps were considered:
1. Data collection
2. Text analysis
3. Variable selection

Post this model building and comparison for prediction was done. For this exercise purpose, we have assumed that the cars of consumers who have posted stalling/stoping complaints were eventually recalled.

a. Data Collection Source: Research was done to select forums, website social media groups where car users complain about their cars. Below forums and websites were chosen because of the high popularity amongst car users.

4. https://twitter.com/carcomplaints: Used to understand the current sentiments of the people.

Web Crawlers - Information Retrieval Studio:

SAS Information Retrieval Studio is a framework and graphical administration interface for crawling, normalizing, analyzing, indexing, and searching text documents. It serves as the default user interface to SAS Web Crawler and SAS Search and Indexing, and provides integration with other products including SAS Content Categorization and SAS Sentiment Analysis.

Information retrieval studio was used to crawl the websites carcomplaints.com and kbb.com.
Nodes used in analysis:

b. Text parsing/Text Filter:
First use the Text Parsing and Text Filter nodes as seen in the diagram flow. The Text Parsing Node takes the raw text from the data source and has the ability to parse different languages and different parts of speech. The Text Filter Node needs to immediately follow the Text Parsing Node.

The Text Filter Node applies filters to the text data and can define the own dictionary, term weighting, frequency weighting, and term filters, which can use the defaults provided in Enterprise Miner. The Text Filter Node creates a new Transaction Data Set that details which observations contain which words. Words like camshaft, ignition, fuel injection which were important for our analysis were added to our dictionary. It was also found that there were some words like “conatct” which was a typing error for “contact”. More of such errors were identified and filtered in the Text Filter.

<table>
<thead>
<tr>
<th>Terms</th>
<th>FREQ</th>
<th># DOCS</th>
<th>KEEP</th>
<th>WEIGHT</th>
<th>ROLE</th>
<th>ATTRIBUTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>bunch</td>
<td>3</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>camshaft</td>
<td>5</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Noun Group</td>
<td>Alpha</td>
</tr>
<tr>
<td>everytime</td>
<td>3</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>list</td>
<td>4</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>tow</td>
<td>4</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>order</td>
<td>3</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Verb</td>
<td>Alpha</td>
</tr>
<tr>
<td>tty</td>
<td>3</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>regular</td>
<td>3</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>occasion</td>
<td>3</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Adj</td>
<td>Alpha</td>
</tr>
<tr>
<td>ignition</td>
<td>3</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Noun Group</td>
<td>Alpha</td>
</tr>
<tr>
<td>module</td>
<td>5</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>daycare</td>
<td>5</td>
<td>3</td>
<td></td>
<td>0.0</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
</tbody>
</table>

The Text Topic Node follows the Text Filter Node in the flow. The Text Topic Node uses the transaction data created by the Text Filter Node and creates “topics,” which are groups of words that are automatically determined to be related. In the results of the Text Topic Node there is a Topics table that contains a summary of information on each topic. Each topic is characterized by several key words, though the total number of terms in a topic is indicated in the Topics table. In the picture below you can see that the top three topics characterized by the keywords “stall, stop, restart, die, power off, tow, shut” indicating there are complaints talking about car stall.
It is evident that the word ‘stall’ is associated with ‘move’ and ‘jerk’ and it is a common problem in traffic. Thus when we keep deep diving into the nodes the problem becomes increasingly clear that the car stalls when it is supposed to move.

The above table is shown in the below matrix as distribution.
c. Variable selection:
After cleaning the data, stat explorer in SAS E Miner was used to do the selection of important variables. Variable selection node was used to find the variable worth. Variables like miles, year and region were kept for analysis. All other variables like, injuries, fire, weather, etc proved to be unimportant.

d. Text Cluster Analysis:
Cluster analysis node is used to analyse clusters generated. Based on scree plot 3 clusters were selected. The highest number of cluster has “stall”, “drive”, “start” which accounts for 67% which is a good number for problem analysis.
The biggest cluster showed frequency of 151 and terms like stall, car, start chevy were evident, Chevy was recalled for stalling problem. This gives enough doubt to dig more and analyse comments with these words to find for patterns like where they all have same problems. Did customers run the car for same miles before stalling? Did they complaint had same pattern before stalling. Are they from same region? Do they come from same manufacture or distribution centre? Should the text mining be done? Are the comments posted useful for predicting recall?

**Models Building**: From the text topic and text cluster node stalling is a major problem for the GM cars. The next step is to understand if the text available in social media can help in predicting the recall better.

The parsed data is then sent to four regression models to predict the recall better. Since recall takes the value one or zero logistic regression is done.

1. Regression model which includes just the numeric values such as miles drove and year the car was purchased.
2. Regression + Text- which includes attributes like miles, year and region which is turned into nominal value and it includes text.
3. Regression +SVD- Which includes numeric values and text which is composed using single value decomposition method which is used for analysis.
4. Neural Network: Formula is used to adjust the weights to calculate the neuron’s output and see the difference between the neuron’s output and the desired output.

The variable weight for test svd is more. That’s says to predict recall we need to analysis text. Rank of the profiles is good is also better for text SVD.

Model comparison is done using ROC analysis where the true and false positive and negatives are used.
ROC Analysis:
Num + SVD was the best regression curve as it gave the best ROC curve. The next best was num + text. 20% of the data was taken for training and 80% of the data is used for validation. Thus through proper training the model can easily help to predict the recalls.

![ROC Chart](image)

The Regression num +svd model is ensuring that the true negative and true positive are higher. Although Neural Network is good at training data set it falls marginally short when it compares to the validation dataset and it is unclear on the process adopted in the neuron training. Hence finally regression num+svd model is chosen for the final modeling. Thus it is clear that numeric values and text SVDs help to predict the models better.

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>Target</th>
<th>False</th>
<th>True</th>
<th>False</th>
<th>True</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reg3</td>
<td>TRAIN</td>
<td>Recalled</td>
<td>7</td>
<td>67</td>
<td>12</td>
<td>74</td>
</tr>
<tr>
<td>Reg3</td>
<td>VALIDATE</td>
<td>Recalled</td>
<td>2</td>
<td>21</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>Reg3</td>
<td>TRAIN</td>
<td>Recalled</td>
<td>3</td>
<td>53</td>
<td>41</td>
<td>76</td>
</tr>
<tr>
<td>Reg3</td>
<td>VALIDATE</td>
<td>Recalled</td>
<td>1</td>
<td>14</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td>Reg3</td>
<td>TRAIN</td>
<td>Recalled</td>
<td>81</td>
<td>99</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Reg3</td>
<td>VALIDATE</td>
<td>Recalled</td>
<td>21</td>
<td>26</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Neural Neural Network</td>
<td>TRAIN</td>
<td>Recalled</td>
<td>5</td>
<td>91</td>
<td>9</td>
<td>76</td>
</tr>
<tr>
<td>Neural Neural Network</td>
<td>VALIDATE</td>
<td>Recalled</td>
<td>4</td>
<td>22</td>
<td>4</td>
<td>17</td>
</tr>
</tbody>
</table>

The Text Rule Builder node generates an ordered set of rules from small subsets of terms that together are useful in describing and predicting a target variable. Each rule in the set is associated with a specific target category. Each target category consists of a conjunction that indicates the presence or absence of one or a small
subset of terms (for example, “term1” AND “term2” AND (NOT “term3”)). A document matches this rule if and only if it contains at least one occurrence of term1 and of term2 but no occurrences of term3. This set of derived rules creates a model that is both descriptive and predictive. When categorizing a new document, the model will proceed through the ordered set and choose the target that is associated with the first rule that matches that document.

The SVDs are then sent through the text builder to understand the factors for recall. The top rules obtained through the text builder for recall is showcasing stall, push, daughter indicating the issues caused by stalling leading to recall where the target variable is set to recall =1. Thus stall is an important predictor for recall.

Conclusion: By using web and social media analysis one can find the key issue people are discussing more which might result in the eventual recall. Through the social media chatter and the sentiment analysis across Chevrolet and GMC, it could have identified the defect of ignition defect. And then the cluster analysis and SVDs nodes can be reassigned to the customer data and attributes and analysed to find the underlying pattern and would help us predict whether there is any impending recall. One of the recommended models is decision tree which could provide insights on the key parameters for the recall. For example, in the below decision tree for maximum recall prediction it is the text cluster node and SVD 4 and the miles involved. This decision tree would help in chalking out the major factors affecting recall.
The SVDs which are statistically significant are then sent to text builder to understand the key factors, decision tree is used to get the most important factors, text cluster and text topic which in turn can provide a final set of classification rules for both recall and no recall predictions in the total data for the GM cars. Thus GM can use the web and social media analytics for more effective predictive recall to ensure lesser brand equity hit.
References:

1) Carcomplaints.com
2) Kbb.com
3) http://money.cnn.com/2014/06/30/autos/general-motors-recall/
4) http://money.cnn.com/2014/05/21/news/companies/gm-recall-nightmare/
5) https://www.nhtsa.gov/
6) https://my.gm.com/recalls
8) http://fortune.com/2015/08/24/feinberg-gm-faulty-ignition-switch/
11) https://www.wsj.com/articles/gm-to-recall-7-6-million-more-vehicles-in-u-s-1404153705
12) http://support.sas.com/demosdownloads/001175/readme_all.pdf