“A Study on traffic accident fatalities in Michigan using Clustering & Classification techniques”

Suresh Ooty Krishnaswamy

Mansur Blackman

Abstract: Given the tremendous loss of life associated with traffic accidents any ability to predict what circumstances lead to fatalities will be very valuable and beneficial. This project explored causality behind the fatal accidents, using entire Michigan fatal incidents dataset [source: FARS, NHTSA] and all traffic incidents in Southeast Michigan that includes both fatal & non-fatal incidents [source: SEMCOG]. On the FARS dataset, a study was made using k-means clustering to establish a hypothetical spatial correlation between clusters of accidents caused by Rain, Snow & Blowing snow vs, accidents occurred in Normal weather conditions. Using SEMCOG data, an exploratory study was done using Naive Bayes Classification Method in predicting whether a set of circumstances surrounding an accident will result in a fatality or not. The results from 1. FARS dataset indicate a more precise weather data at the GPS level would be required to build models to predict fatal incidents based on weather condition. 2. SEMCOG dataset indicate that fatal accidents based on associated events could be predicted up to 50% using Naïve Bayes classifier.

Introduction:

According to National Highway Traffic and Safety Administration (NHTSA) report[6] 35,092 lives have been lost nationwide in 2015 involving motor vehicle crashes leaving 2.44 million people injured, which is a 7.2% increase from 2014. A report from ‘Fatality Analysis Reporting System’ (FARS) of NHTSA on fatal crashes by Weather Condition & Light Condition about 9% of fatal crashes occurred in Rain & Snow/Sleet conditions across USA and the same is 12.6% in Michigan The number of fatalities show an increasing trend in 2016, when the first half data is compared with 2015[7]. A Transportation research project [4] on climate change modeling & weather related road accidents (done in Canada) suggested that there are fewer traffic incidents reported in snowy weather conditions. But a similar study carried out in US[3] suggest that driving under unfavorable driving conditions, such as dark, adverse weather, or poor road surface might result in more severe injuries. During adverse weather conditions such as rain, snow, fog or storm drivers take different maneuvers to avoid crashes. Newly registered drivers, young drivers (15 - 34 years age) and Aged drivers (65 years and above) are vulnerable in such conditions [4]. This study also suggests that a sudden rain after a long gap leading to oil wash on road could make the road slippery & prone to accidents. Traffic accidents are considered one of the most important and dangerous problems [1] causing fatalities to the society. The study explores factors surrounding the fatal accidents occurred in Michigan.

Approach:

Two sets of datasets from two different data sources were explored for identification of a pattern, or causality. FARS dataset show that there are about 2000 fatal accidents each year and these occurred at different circumstances of weather, road conditions, road segments, time of the day, etc., Due to lack of GPS attributes before 2009, data from 2010 to 2015 were taken in to consideration. And as the scope of the study was on the weather factors, the data was cleansed by dropping the incidents related to alcohol influence.

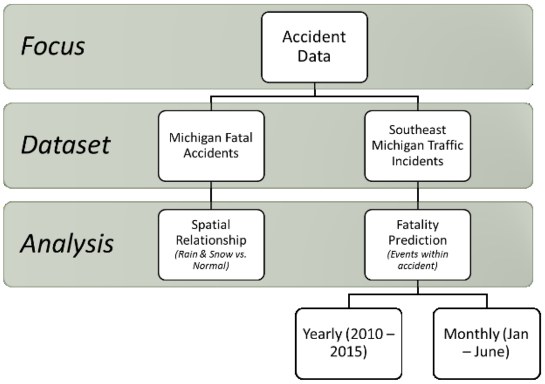


Figure 1 - Approach

Southeast Michigan Council of Governments (SEMCOG) dataset show that, there are about 300 fatal accidents in Southeast Michigan against 100000+ non-fatal traffic incidents each year from 2010 to 2015. An increasing trend of accidents in Southeast Michigan could be noted from the graph. With more than 100000 records for each year, SEMCOG dataset was comprehensive, but unlike FARS dataset, it did not have GPS attributes.

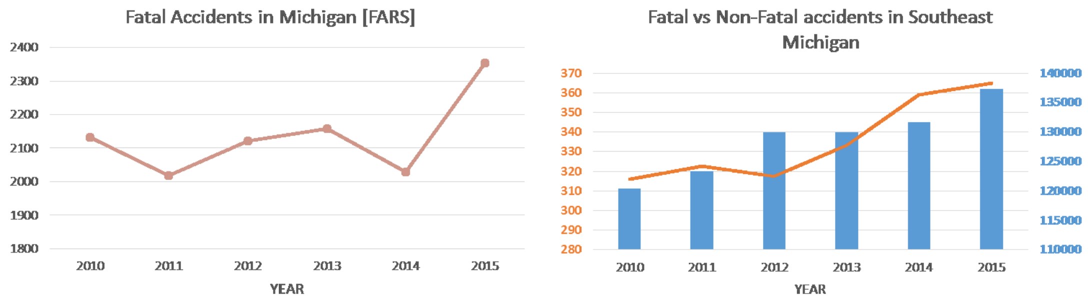


Figure 2 – Fatal Accidents in Michigan Region

Using FARS dataset, a study was done using k-means cluster if the locations/zones of weather related accidents could essentially predict that the area would be prone to a preponderance of non-weather related accidents. The dataset was further reduced to extract the unique set of records for each incidents, out of which 1552 records were related to rain, snow & blowing snow conditions, against 8685 unique records occurred in normal weather conditions. To study the spatial clusters, K-means cluster with 'Euclidean' was an obvious choice, as the scope of clusters were using GPS longitude & latitude. It was a challenging pursuit to identify areas that were prone to weather related accidents were also prone to non-weather related accidents. Number of iterations were run to construct a reasonable set of clusters across the map of Michigan, by scrutinizing the 'within sum of square errors' of clusters, in a spatial perspective. Additionally the clusters produced often spanned very large areas of 100+ square miles and it didn't seem realistic to claim that any common set of factors affected an area this size. Further, errors on data capture (esp. on weather) by a police officer reporting to the incident spot after few minutes could not be ruled out. To avoid such discrepancies of weather records for reported incidents, an approach of associating a third dataset that consists "Local Climatological Data" (LCD) from National Oceanic and Atmospheric Administration (NOAA) [8]. Spatial techniques such as identifying nearest neighbors using *Alteryx* [8] was applied to identify the nearest weather stations that contain precipitation data for every 15 minutes. The precipitation data could be extracted within ± intervals, but there were considerable amount of data points that were 100 miles away from the nearest weather station. This lead to a challenge of accurately associating the precipitation levels for some of the data points. A combination of non-availability of non-fatal incidents, accurate weather details and size of clusters lead to a challenge of establishing meaningful clusters to compare against the accidents occurred in normal weather conditions. Also a real time precipitation levels captured by on-vehicle sensors might have helped this study, but such data is not available to public today. A superimposed image of accidents occurred in normal weather on the clusters show that there are many outliers leading to questioning the approach adopted. However during this study, some interesting facts were observed such as

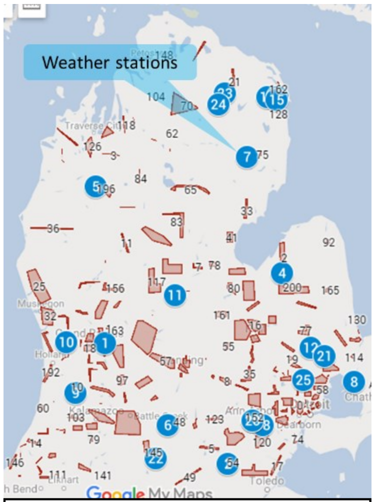


Figure 3 - Weather Stations

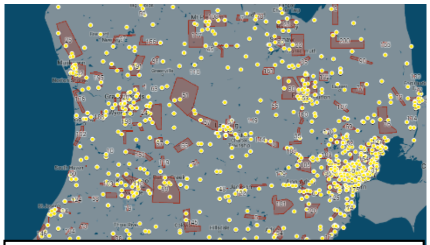


Figure 4 - Normal accidents superimposed on clusters

1) Southeast Michigan has high density of fatal accidents than any other part of Michigan (This observation lead us to explore Southeast Michigan traffic incidents from SEMCOG)

2) US24 (Telegraph road) seem to be “No Fatal Accident” zone in any weather condition in this period.

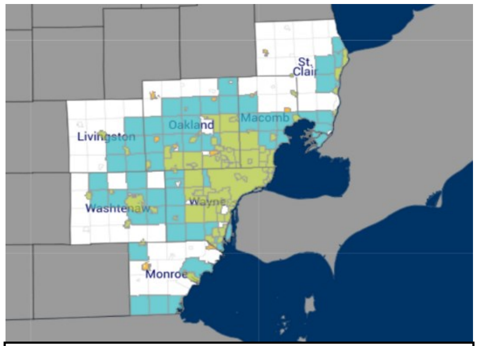


Figure 5 - Map of Southeast Michigan

SEMCOG dataset enabled study on a focused area, with limited surface area compared to FARS data. But it did not have the GPS coordinates found in the FARS dataset. In analyzing traffic accident data there are many different factors that can impact whether or not a fatality occurs. The dataset provided the police reports from thousands of traffic accidents in the southeast Michigan area (seen in figure 5). To be as close as possible to the current day patterns and trends, only the data from the six year period of 2010-2015 was looked at. All of the instance attributes available in the dataset used were attributes present in police reports from traffic accidents, so any incidents where no police report was filed or where the report had incomplete information was not available. Ten attributes were used to predict the fatalities using Naïve Bayes classifier. An application of Naïve Bayes classifier, as the chosen attributes were independent to each other.

Methodology:

Data Sampling:

FARS: A total of 1552 records were identified as influenced by weather conditions, by filtering on the attribute atmcond1 with 2, 4 & 11 (factors used for Rain, Snow & Blowing Snow). These records were further used to construct clusters using k-means.(see figure 3.)

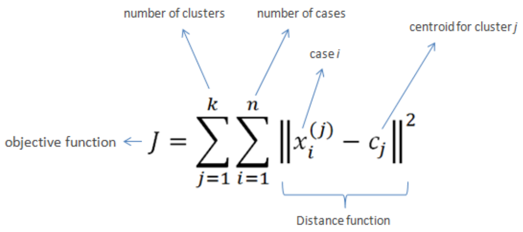
SEMCOG: The yearly analysis datasets were broken into training and testing files for each year from 2010-2015, making for 12 files in total. For the monthly analysis one file was made for months from Jan-Jun for 2010-2015. In the end, the monthly analysis generated one training and one test CSV also making for a total of 12. The CSV files and their corresponding ARFF files from *Weka* can be found in the project appendix. It is observed that the traffic data is unbalanced with fatal accidents and non-fatal accidents, with a mere 0.3% of the annual total accidents have fatalities. The testing and training files were separated using Alteryx so that one percent of the training set would be fatal ones. The training set generated had about 20000 non-fatal accidents and about 200 fatal accidents. Such approach gave a reasonably balanced training dataset. 1% was judged to allow a sufficient sampling while staying close enough to the 0.3% trend. The response variable ‘fatal’ is codified as ‘0’ for non-fatal and ‘1’ for fatal.

WEKA models:

K-means clustering on FARS data: Using *Weka*, few iterations were run to build a reasonably sized clusters that have spatial polygons. The within sum of squared errors was used as an indicator, a 0.79 of WSS had clusters with polygons of perimeter 57.3 miles. The data points that had atmcond1=2 or 4 or 11 were taken in to building the clusters. The scheme used was as follows:

Scheme: weka.clusterers.SimpleKMeans -init 1 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -V -M -N 201 -A "weka.core.EuclideanDistance -R first-last" -I 500 -O -num-slots 1 -S 100 -output-debug-info

K-means Clustering Method: K-Means clustering intends to partition n objects into *k* clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly *k* different clusters of greatest possible distinction. The best number of clusters *k* leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of *k*-Means clustering is to minimize total intra-cluster variance, or, the squared error function:



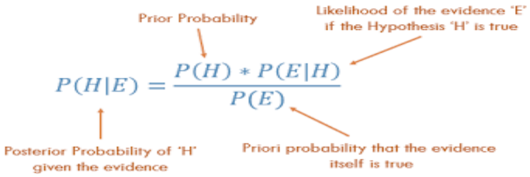
Naïve Bayes Classifier on SEMCOG data: Within the *Weka* tool, several attributes judged to be irrelevant or seen to violate the assumption of independence between attributes were removed. Since the SEVERITY attribute directly identified fatalities in addition to the A-C level injuries and property damage, it could not be used in the training set. In specific the CRASH_ID, ROADNAME, MILE, POLICE REPORT NUMBER, MONTH, DATE, YEAR, and SEVERITY attributes were removed. This left the DAY, TIME, TYPE, WEATHER, LIGHTING, ROADCONDITION, OFFSETDISTANCE, FACTOR and FATALITY attributes with FATALITY being used as the class identifier. Since three or more of the factors sometimes occurred in the same incident this attribute was broken into 16 separate attributes with values of zero or one if that factor was present in the accident. Some common factors include the involvement of a young driver, large truck, or a motorcycle. The TYPE of the accident described the specific manner of collision (i.e. head-on, rear-end, single vehicle, etc.). The offset distance is the distance between where an accident occurred and where it was at the time of report.

Bayes Classification Method:

The Naive Bayes classification method was used to predict the fatal accidents on the test data sets. The Naive Bayes method begins with the assumption that all attributes used in prediction are independent of each other. The independence assumption is not 100% valid for the car accident dataset but through deleting some attributes that had clear dependencies, a set of very nearly independent attributes were developed. The idea for applying the Naive Bayes method came from Mujallli[1], in that work the technique was applied in a similar situation in Jordan for analyzing traffic accidents. We were interested to see how effective this same technique would be on the Michigan. Additionally this is one of the built in functions in the *Weka* tool and allowed very comprehensive reports to be generated for each run.

The Bayes classification method relies on the Bayes theorem which is:

The classification method builds on the above theorem. Through this method we are able to evaluate the likelihood of future events given the prior history of events and the attributes of those instances around them.


Prior Probability
Likelihood of the evidence 'E' if the Hypothesis 'H' is true
$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$
Posterior Probability of 'H' given the evidence
Priori probability that the evidence itself is true

Tools Used:

For the computing work done in the project there were two primary tools used. The first was *Alteryx*, which allowed for a very smooth preprocessing of data sets. The second was *Weka*, and this tool performed the Naive Bayes classification and generated very comprehensive result summaries for all of the test runs.

Alteryx is a tool that the team was familiar with due to data analysis work previously done using it. Through *Alteryx*, the team was able to take the CSV files downloaded from FARS & SEMCOG website and split them up, join them together, reformat, and identify some issues with the data that needed to be addressed or cleaned. A great deal of time was saved being able to manipulate several CSV files in *Alteryx* rather than trying to perform operations on several files individually.

Weka has the Naive Bayes function built in and was used to perform the final preprocessing in picking the attributes to be used and then perform the computational heavy lifting of the NB classifications. Although the team realized there were more things that the tool is capable of, including some of the steps done in *Alteryx*, the level of comfort both team members had with the tools made the aforementioned division of labor between tools the most effective.

Results:

Naive Bayes Classification on Yearly Data: In reviewing the output of the models generated most years were able to achieve nearly a 50% true positive rate in accurately predicting fatalities, while all years had less than a 3% false positive rate. This output is owed in large part to the unbalanced nature of the dataset with traffic accidents causing fatalities very infrequently and accounting for less than one percent of the total accidents. Since so few accidents result in fatalities the model tends to predict too many fatalities since the training set had Additionally an ROC area of 0.9 was reached for most years.

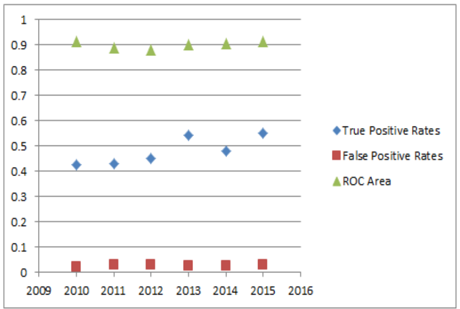


Figure 6 – TPR, FPR & ROC on Yearly data

Table 1 – YEARLY model results

YEAR	Class	TP Rate	FP Rate	Precision	Specificity	Recall	F-Measure	MCC	ROC Area	PRC Area	Confusion Matrix
2010	0	0.977	0.573	0.999	0.427	0.977	0.988	0.093	0.914	1	a b <-- classified as
	1	0.427	0.023	0.022	0.977	0.427	0.042	0.093	0.914	0.031	99629 2369 | a = 0
	wt.avg	0.976	0.572	0.998	0.976	0.976	0.987	0.093	0.914	0.999	71 53 | b = 1
2011	0	0.971	0.57	0.999	0.43	0.971	0.985	0.083	0.89	1	a b <-- classified as
	1	0.43	0.029	0.018	0.971	0.43	0.034	0.083	0.89	0.032	101174 3033 | a = 0
	wt.avg	0.97	0.57	0.998	0.97	0.97	0.984	0.083	0.89	0.999	73 55 | b = 1
2012	0	0.972	0.551	0.999	0.449	0.972	0.985	0.091	0.882	1	a b <-- classified as
	1	0.449	0.028	0.021	0.972	0.449	0.039	0.091	0.882	0.029	99580 2912 | a = 0
	wt.avg	0.971	0.551	0.998	0.971	0.971	0.984	0.091	0.882	0.999	75 61 | b = 1
2013	0	0.976	0.456	0.999	0.544	0.976	0.987	0.118	0.901	1	a b <-- classified as
	1	0.544	0.024	0.028	0.976	0.544	0.052	0.118	0.901	0.043	105132 2616 | a = 0
	wt.avg	0.975	0.455	0.998	0.975	0.975	0.986	0.118	0.901	0.999	62 74 | b = 1
2014	0	0.974	0.522	0.999	0.478	0.974	0.986	0.096	0.904	1	a b <-- classified as
	1	0.478	0.026	0.021	0.974	0.478	0.041	0.096	0.904	0.047	113240 3051 | a = 0
	wt.avg	0.973	0.521	0.998	0.973	0.973	0.985	0.096	0.904	0.999	72 66 | b = 1
2015	0	0.972	0.448	0.999	0.552	0.972	0.986	0.11	0.914	1	a b <-- classified as
	1	0.552	0.028	0.024	0.972	0.552	0.045	0.11	0.914	0.054	115052 3296 | a = 0
	wt.avg	0.972	0.448	0.998	0.972	0.972	0.984	0.11	0.914	0.999	65 80 | b = 1

Naive Bayes Classification on Monthly Data:

Data Elements:

The exploratory study on the data, revealed some interesting coincidences such as higher number of fatal accidents occurred during the lighting conditions “Dark” and “Lights” are correlated to the ones that occurred between 6pm to 6am. There was a peak of fatal incidents noted against the data attribute “Time” for these values. To study the behavior of the model, a new factor attribute “*Dark_r_Lights*” (here after referred as DRL) was introduced with 0s and 1s to represent the existence of either

Fatal incidents in Light conditions	
Dark	**548**
Dawn	35
Lights	**556**
Daylight	825
Dusk	44
Uncoded	1
Unknown	9

“Dark” or “Lights” condition. The idea was to eliminate the variables TIME and LIGHTING with DRL and to study on how the Naive Bayes model when a dependent variable is introduced. This approach was taken only on the monthly slice of the dataset. It could be noted that the total month wise fatal accidents is low for June (200) when compared with rest.

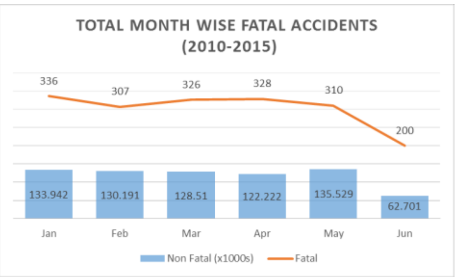


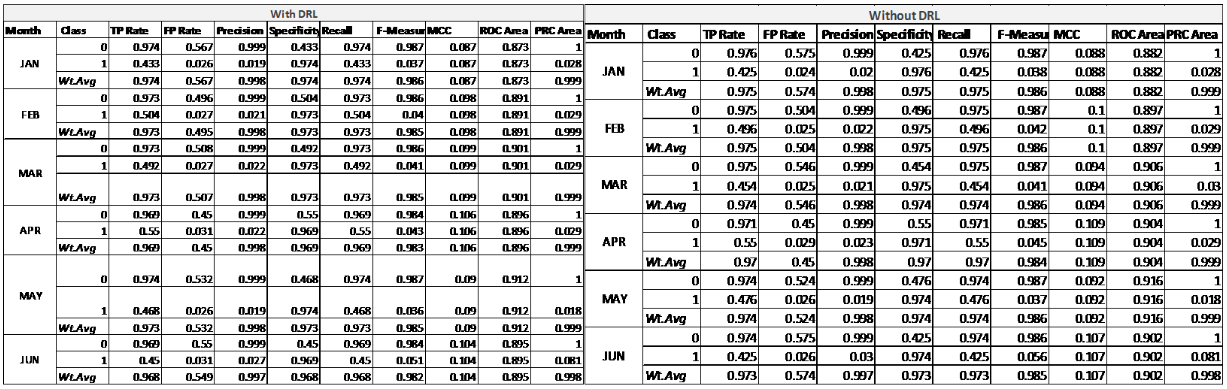
Figure 7 – Total Month wise fatal accidents (2010-2015)

Analysis: Two models using Weka’s Naïve Bayes classifier,

1.With all three variables TIME, LIGHTING and DRL in the model and

2. Without DRL, were applied on six datasets. Both the models contained DATE, YEAR, DAY, TIME, TYPE,WEATHER, ROADCONDITION, OFFSETDISTANCE, UNITS, FACTORS (separated as A, B, C, D, E, F, G, I, L, M, P, R, S, T, W, Y variable) and FATAL. Both the models contained DATE, YEAR, DAY, TIME, TYPE,WEATHER, ROADCONDITION, OFFSETDISTANCE, UNITS, FACTORS (separated as A, B, C, D, E, F, G, I, L, M, P, R, S, T, W, Y variable) and FATAL.

Table 2 – JAN-JUN model results



With DRL										
Month	Class	TP Rate	FP Rate	Precision	Specificity	Recall	F-Measur	MCC	ROC Area	PRC Area
JAN	0	0.974	0.567	0.999	0.433	0.974	0.987	0.087	0.873	1
	1	0.433	0.026	0.019	0.974	0.433	0.037	0.087	0.873	0.028
	Wt.Avg	0.974	0.567	0.998	0.974	0.974	0.986	0.087	0.873	0.999
FEB	0	0.973	0.496	0.999	0.504	0.973	0.986	0.098	0.891	1
	1	0.504	0.027	0.021	0.973	0.504	0.04	0.098	0.891	0.029
	Wt.Avg	0.973	0.495	0.998	0.973	0.973	0.985	0.098	0.891	0.999
MAR	0	0.973	0.508	0.999	0.492	0.973	0.986	0.099	0.901	1
	1	0.492	0.027	0.022	0.973	0.492	0.041	0.099	0.901	0.029
	Wt.Avg	0.973	0.507	0.998	0.973	0.973	0.985	0.099	0.901	0.999
APR	0	0.969	0.45	0.999	0.55	0.969	0.984	0.106	0.896	1
	1	0.55	0.031	0.022	0.969	0.55	0.043	0.106	0.896	0.029
	Wt.Avg	0.969	0.45	0.998	0.969	0.969	0.983	0.106	0.896	0.999
MAY	0	0.974	0.532	0.999	0.468	0.974	0.987	0.09	0.912	1
	1	0.468	0.026	0.019	0.974	0.468	0.036	0.09	0.912	0.018
	Wt.Avg	0.973	0.532	0.998	0.973	0.973	0.985	0.09	0.912	0.999
JUN	0	0.969	0.55	0.999	0.45	0.969	0.984	0.104	0.895	1
	1	0.45	0.031	0.027	0.969	0.45	0.051	0.104	0.895	0.081
	Wt.Avg	0.968	0.549	0.997	0.968	0.968	0.982	0.104	0.895	0.998

Without DRL										
Month	Class	TP Rate	FP Rate	Precision	Specificity	Recall	F-Measu	MCC	ROC Area	PRC Area
JAN	0	0.976	0.575	0.999	0.425	0.976	0.987	0.088	0.882	1
	1	0.425	0.024	0.02	0.976	0.425	0.038	0.088	0.882	0.028
	Wt.Avg	0.975	0.574	0.998	0.975	0.975	0.986	0.088	0.882	0.999
FEB	0	0.975	0.504	0.999	0.496	0.975	0.987	0.1	0.897	1
	1	0.496	0.025	0.022	0.975	0.496	0.042	0.1	0.897	0.029
	Wt.Avg	0.975	0.504	0.998	0.975	0.975	0.986	0.1	0.897	0.999
MAR	0	0.975	0.546	0.999	0.454	0.975	0.987	0.094	0.906	1
	1	0.454	0.025	0.021	0.975	0.454	0.041	0.094	0.906	0.03
	Wt.Avg	0.974	0.546	0.998	0.974	0.974	0.986	0.094	0.906	0.999
APR	0	0.971	0.45	0.999	0.55	0.971	0.985	0.109	0.904	1
	1	0.55	0.029	0.023	0.971	0.55	0.045	0.109	0.904	0.029
	Wt.Avg	0.97	0.45	0.998	0.97	0.97	0.984	0.109	0.904	0.999
MAY	0	0.974	0.524	0.999	0.476	0.974	0.987	0.092	0.916	1
	1	0.476	0.026	0.019	0.974	0.476	0.037	0.092	0.916	0.018
	Wt.Avg	0.974	0.524	0.998	0.974	0.974	0.986	0.092	0.916	0.999
JUN	0	0.974	0.575	0.999	0.425	0.974	0.986	0.107	0.902	1
	1	0.425	0.026	0.03	0.974	0.425	0.056	0.107	0.902	0.081
	Wt.Avg	0.973	0.574	0.997	0.973	0.973	0.985	0.107	0.902	0.998

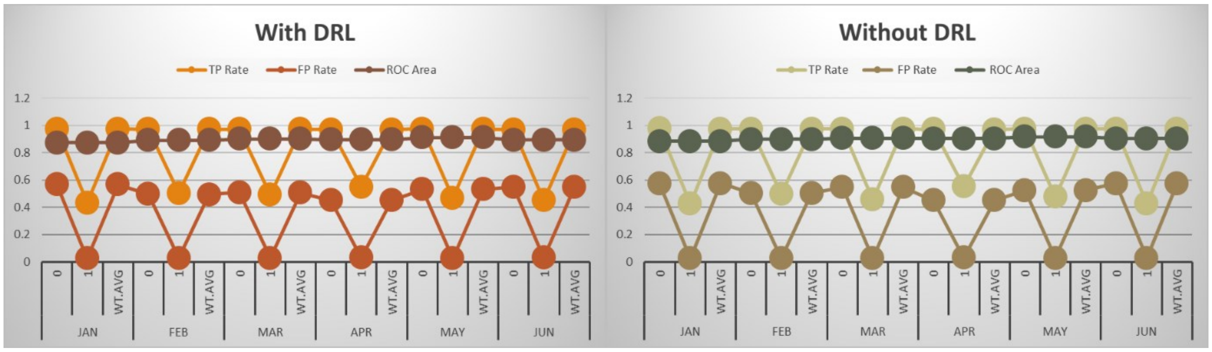


Figure 7 – TPR, FPR & ROC for monthly datasets

Performance Measures: The performance measures used in this study were accuracy, sensitivity, specificity and F-measure. Accuracy is the proportion of instances that were correctly classified among all instances. Accuracy only gives information on the classifier’s overall performance. In cases where there is a highly skewed data distribution, the overall accuracy is not

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

sufficient. In this case, accuracy might give a false indication that a classifier performance is high, where in fact the classifier is only predicting all samples as belonging to one class value, in which case it is biased in its results to majority class. Sensitivity and specificity are usually adopted to monitor classification performance on two classes separately. Sensitivity represents the proportion of correctly predicted as "fatal" among all the observed as "fatal". Specificity represents the proportion of correctly predicted as "Fatal" among all the observed "non-fatal". F-measure represents the harmonic mean of precision and sensitivity, and is frequently used in imbalanced datasets [1]. However, a trade-off exists between sensitivity and specificity. Therefore, the area under a Receiver Operating Characteristic (ROC) curve is also used as a target performance measure. ROC curve represents the true positive rate (sensitivity) vs. the false positive rate (1-specificity). ROC curves are more useful as descriptors of overall performance, reflected by the area under the curve, with maximum of one describing a perfect test and a ROC area of 0.50 describing a valueless test.

Conclusions:

Yearly Models:

1. ROC area, weighted average of Precision indicate that the overall model is satisfactory.
2. The "non-fatal" data points have a precision close to 1, indicate that the predictions for non-fatal incidents are accurate.
3. The TPR of "fatal" data points are around '0.5', which indicate that at least 50% of the "fatal" data classified accurately.
4. FPR for "fatal" classification are around '0.3'.

Monthly Models:

5. ROC area, Sensitivity, Specificity improved as the dependent attribute DRL was dropped. This also indicates that overall precision of the model improved, with slight improvement to the precision of "fatal" class.
6. ROC areas in the ranges above '0.8' indicates that the model is reasonably good.
7. False Positive Rate for condition b = 1 (fatal) The FPR rate for "fatal" class is below '0.3'.

Overall:

The performance measures of both YEARLY and MONTHLY models built using SEMCOG dataset indicate that they behave very similar in terms of precision, TPR, FPR and ROC area. This could be due to the reason that a similar approach was adopted to construct the training dataset from an imbalanced dataset. And on the FARS dataset, a more precise weather data at the GPS location & time of traffic incident are necessary to arrive at conclusions on weather influence on each individual fatal accidents. A more precise precipitation data obtained from on-vehicle sensors would enable such study.

References:

[1]Bayes classifiers for imbalanced traffic accidents datasets - Randa Oqab Mujallia,Griselda Lópezb, Laura Garachb

[2]Road Danger Estimation for Winter Road Management - Pavel Moiseets and Yuzuru Tanaka

[3]Influential Factors for Severe Traffic Crashes - Huanmei Wu, Sravani Malipeddi

[4]Climate change modeling and the weather-related road accidents in Canada - Md. Shohel Reza Amin a,Alireza Zareie b,1, Luis E. Amador-Jiménez

[5]Traffic indicators, accidents and rain: some relationships calibrated on a French urban motorway network - Maurice Arona, Romain Billotb, Nour-Eddin EL Faouzib, Régine Seidowsky

[6] NHTSA, US Department of Transportation, 2015 Motor Vehicle Crashes: Overview

[7] NHTSA, US Department of Transportation, Early Estimate of Motor Vehicle traffic fatalities for teh first half (Jan-Jun) of 2016

[8] National Oceanic and Atmospheric Administration [NOAA],(https://www.ncdc.noaa.gov/cdo-web/datatools/lcd *)*

[9] Alteryx tool (http://www.alteryx.com/ *)*

[10] http://www.saedsayad.com/clustering_kmeans.htm