Prediction Model of Potential Road Crashes Bases on the Spatial and Temporal Factors

Mohana Bansal
University of Maryland, College Park
mohanab@umd.edu

Kushal Thakkar
University of Maryland, College Park
kthakkar@umd.edu

Gaurav Shahane
University of Maryland, College Park
gshahane@umd.edu

Abstract

According to the National Safety Council report, approximately 38,300 people were killed and about 4.4 million injured in the United States in 2015. There are a variety of reasons that contribute to accidents some are internal to the driver but many are external. For example, adverse weather conditions like fog, rainfall or snowfall cause partial visibility and it may become difficult as well as risky to drive on such roads. At the same time, clear weather condition but poor road pavement condition might also lead to accident-prone situations. Predicting any likelihood of crashes as the effect of these features would be a major step towards achieving better public safety. This paper, attempts to create models that account for spatial and temporal features to determine likely road crash conditions. It is expected that the findings from this paper would help civic authorities to take proactive actions on likely crash prone weather and traffic conditions. These models can also be distributed as APIs and can form as an additional safety layer on existing navigational applications used for commutes.

Introduction

The preliminary research suggests that there are numerous data points like road weather condition which comprises of road surface condition (dry, wet, chemically wet, and more), wind speed, wind direction, precipitation, road geometry, speed limit and more that are available through means of open data initiatives, Department of Transportation archives and open and commercial APIs that can provide interesting features to model predictability of vehicular crashes.

The paper focuses on the IOWA Road Network which has over 80 weather and traffic sensors that archive information on road weather conditions, average traffic speed, density of different sized vehicles and more, 24 hours a day 7 days a week. This data is combined with the State of IOWA's open road crash dataset.

The aim of this paper is to predict the occurrences of crashes in around the vicinity of Road Weather Information Systems (RWIS) sensors . The paper attempts to find the relationship between the accidents, road surface and weather conditions. It describes the data and the relationship between the parameters that results in crash prone situations. The paper will then use this data to build models that will predict the occurrences of such crashes happening at a given geographical area. It is expected that these models can find their use in existing navigation systems and will help in rolling out alerts to drivers or road authorities of probable crashes conditions.

Research Question

Predicting the occurrences of vehicular crashes on State of Iowa roadways based on spatial and temporal features of weather and traffic.

State of the Art

The research papers referred were on the lines of accident prediction attempts using different approaches. Each paper has unique way data collection approach, different focus of predictions (predicting crash frequencies on a yearly level, predicting safer routes in relation to a route and so on), they deploy different statistical models. One of the papers suggests that the speed difference on the adjacent roads is a major factor (Abdel-Aty, M. A., Pemmanaboina, R, 2006) and the other discusses how the weather conditions are a major contributor (Yu,

R., Abdel-Aty, M. A., Ahmed, M. M., et.al 2014). The paper introduces various dimensions that may be considered when building statistical models for this paper. For instance, it's not just the weather condition but also traffic conditions that in combination might result in a better predictive models.

Data Collection and Aggregation

The data is extracted from diverse sources to accommodate factors like weather, road conditions and crash details. The datasets consists of the following: i. State of Iowa - Crash Data Set (https://catalog.data.gov/dataset/crash-data). This data contains logged details about vehicle crashes ranging from 2006 to 2016. It includes information about the date, time, location, weather and the road details for the crashes. The weather dataset and the traffic dataset are from the State of Iowa- Road Weather Information System's archives.

The Road Weather Information System's Department has over 80 weather, traffic sensing units located on the highway system, mostly in rural areas and areas with recurrent weather conditions. This dataset was extracted from Iowa Environmental Mesonet (https://mesonet.agron.iastate.edu/request/rwis/traffic.phtml). The data contains information about the average wind speed, maximum wind speed, humidity, visibility, wind direction, precipitation rate, and etc.

The third dataset is also based on the State of Iowa's Historic Traffic Data and is extracted from Iowa Environmental Mesonet (IEM) (https://mesonet.agron.iastate.edu/request/rwis/traffic.phtml). This data was captured in real time and made available by IEM. Thus, the archived information about the exact traffic conditions near these stations in regular interval was used for the analysis. This is a highly granular data and it provides information about parameters such as average speed, average headway, occupancy, and other traffic related features.

The analysis is based on the above three data sets. They provide different dimensions to cover various aspects related to crash and noncrash conditions. The data points are detailed and collected over a longer period of time and are geo-located. This makes the data very promising for spatial and temporal statistical analysis and modeling.

Data Description

The fields included in the final dataset are as following:

- Station: the field contains the name of the stations. Some station names examples include: "RAVI4", "RBIFI4" and more.
- Date: The field contains the date corresponding to a traffic and weather condition.
- Time: The field contains time corresponding to the traffic and weather conditions.
- Longitude: The field contains the longitude of the station.
- Latitude: The field contains the latitude of the station.
- tmpf: The fields holds the median air temperature of the station at a particular hour and date. The data type of the field is 'numeric'. Example values: "19.4", "22.9".
- dwpf: The field holds the median dew point temperature for a station at a particular hour and date. Example values: "18.4", "19.5".
- sknt: The field contains the median wind speed in knots for a station at a particular hour and date. Example values: "5.4", "4.32".
- drct: The field contains the median of the wind direction for a station at a particular hour and date. The data type of the field is 'numeric'. Some of the values are "258", "272".
- gust: The field contains the median of the wind gust in knots for a station at a particular hour and date. Some of the values are "8.64", "5.94".
- tfs: The median pavement temperature for a station at a particular hour and date. Some of the values are "16.3", "19.9".
- lane_id: Id of the lanes in the State of Iowa. The data type of the field is 'numeric'.

- avg_speed: The median of the average traffic speed for a station at a particular hour. The data type of the field is 'numeric'.
- normal_vol: The median traffic volume on a lane of Iowa at a particular hour.
- occupancy: The median occupancy on the lane of Iowa at a particular hour.

Data Cleaning

The data cleaning started with merging Weather Data, and Traffic Data for the State of Iowa. Following strategies were used for merging the data.

To create a single dataset for the monthly data files for 6 months from January 2016 to June 2016, R code was written to create a single CSV file for the purpose of analysis. Single files were generated for weather and traffic data respectively.

The time was rounded up to the hourly window it falls in. After this, remaining parameters were aggregated on hourly basis.

For aggregating the records in both the files, disparate reading from different sensors of the same RWIS station, occasional faulty sensors and so on were accounted. The records were aggregated on the median values for such parameters. This eliminated extreme values in the merged dataset due to faulty readings from RWIS sensors or equipment failures.

Another effort for Data Cleaning went into aggregating the nominal type records of Road Pavement Conditions. Each station in the Road Weather Station Data, there were records from four different sensors about the road conditions. The data also had "Error", "Other", and "No Reports" as the weather conditions. To train the model, there should be one specific condition for each station. Thus, the frequency of each weather condition category for a specific station at a particular time was calculated. The condition with the highest frequency was assigned as the weather condition for that station at that time. However, in the case when there were two conditions with the same frequency for a station at the same time, the less severe condition was chosen.

Finally, the Weather and Traffic data were merged together based on RWIS Station, Date and Hour features. The size of this dataset was much less than the size of the individual datasets of Weather and Traffic due to aggregation of the data, omission of records with missing values and matched records between the various datasets for merging. This dataset served as the base on which the Road Weather conditions and number of crashes happened per hour per date per station were added in the subsequent step.

For aggregating crash dataset with the road weather and the traffic data set, Haversine distance between the crash site and each of RWIS stations in state of IOWA was calculated. The formula used is as below:

Haversine formula:

$$a = \sin^2(\quad /2) + \cos \quad 1 \quad \cos \quad 2 \quad \sin^2(\quad /2)$$

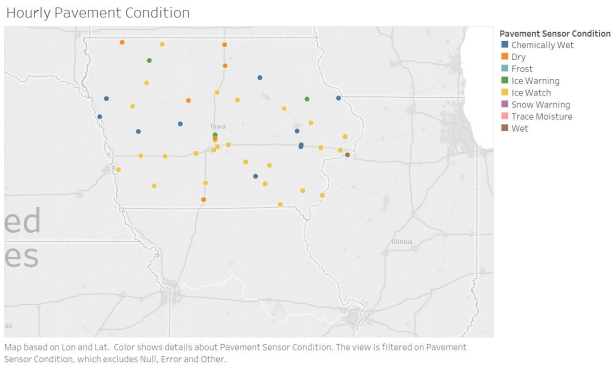
$$c = 2 \quad \text{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \quad c$$

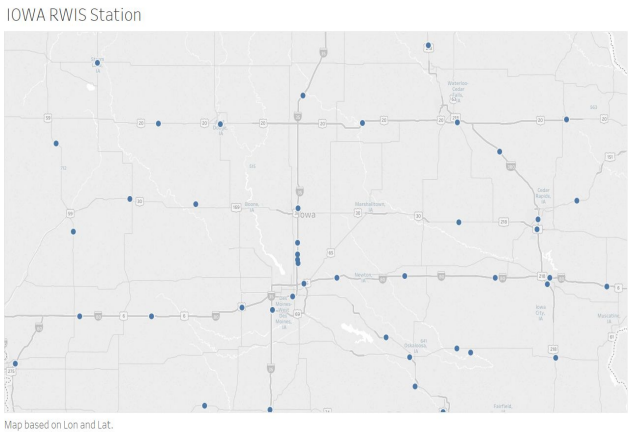
where φ *is latitude,* λ *is longitude, R is earth's mean radius = 6,371km*

After calculating distance of each crash site from the RWIS station, nearest weather station to the accident site and the corresponding Haversine distance of that weather station was assigned. This information was used as a connecting link between Crash Data, Road Weather Data and Traffic Data.

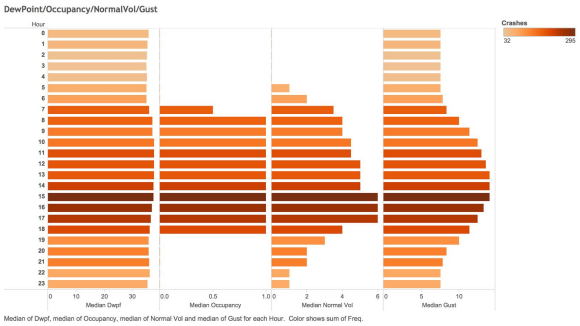
Exploratory Analysis



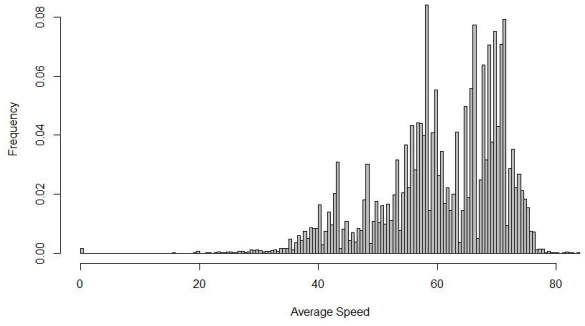
Road weather condition snapshot as on January 3, 12:00 PM.



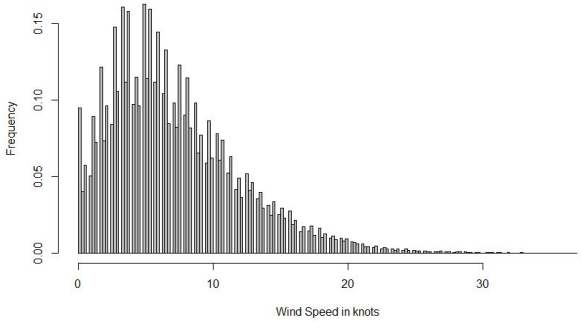
The above graph was plotted on the State Map of Iowa. The blue dots represent the RWIS weather station locations. The data covers comprehensive collection of the weather conditions for all parts of the State of Iowa.



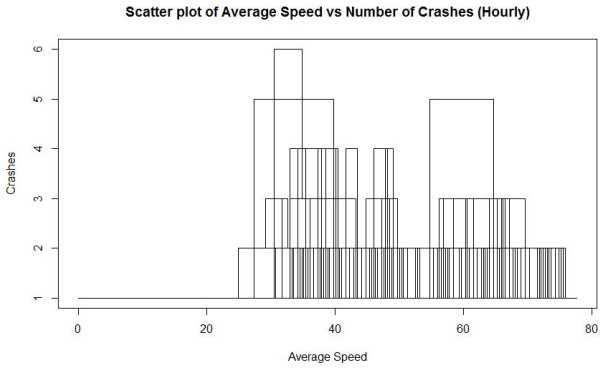
The above graph was color coded on the number of crashes and has dew point temperature, occupancy, normal volume and gust on the x-axis and hours on the y-axis. From the plot, the crash pattern can be observed for various conditions around the clock. It can be further observed that the darkest color is approximately between 14-16 hours. These are the hours with maximum crashes in all the graphs.



The above graph is plotted to observe the average speed across lanes in the State of Iowa. It can be observed that the average speed for maximum traffic is between 65-75.



The above histogram was plotted to observe the pattern of the wind speed in knots in the State of Iowa. The maximum wind speed in knots varies between 3-7.



The above plot is between median the average speed across the lanes of Iowa against the number of crashes. As it can be observed that the number of crashes increase when the average speed is between 35-50 miles/hour and 55-70 miles/hour, we can say that there is a relationship between the two variables.

Experimental Model

Multivariate Linear Regression

The number of accidents occurring in an hour was modelled as a result of all the other variables. It was assumed that the number of accidents may have a linear relationship with the

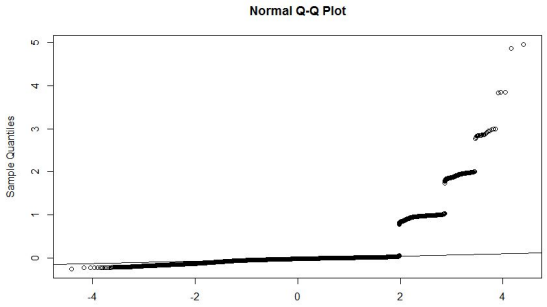
other variables. Training dataset was divided into training and testing dataset in a 75% and 25% proportion respectively. The result was as below.

```
Call:
lm(formula = Freq ~ ., data = dataLm)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3440 -0.1486 -0.0814 -0.0242  4.7486

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.2700558  0.0620200  20.478  < 2e-16 ***
hour         0.0016878  0.0012672   1.332 0.182988
tmpf        -0.0038650  0.0018863  -2.049 0.040550 *
dwpf         0.0027470  0.0010296   2.668 0.007670 **
sknt        -0.0174114  0.0050906  -3.420 0.000634 ***
drct        -0.0000450  0.0000690  -0.652 0.514329
gust         0.0139828  0.0035320   3.959 7.70e-05 ***
tfs          0.0003880  0.0010928   0.355 0.722604
lane_id      0.0076856  0.0111770   0.688 0.491741
avg_speed   -0.0040257  0.0007474  -5.386 7.74e-08 ***
avg_headway         NA         NA      NA       NA
normal_vol   0.0058192  0.0014985   3.883 0.000105 ***
long_vol    -0.0018703  0.0020873  -0.896 0.370305
occupancy   -0.0029733  0.0012729  -2.336 0.019562 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3656 on 3019 degrees of freedom
Multiple R-squared:  0.0417,    Adjusted R-squared:  0.03789
F-statistic: 10.95 on 12 and 3019 DF,  p-value: < 2.2e-16
```

It was observed that the variables: pavement temperature (tfs), Dew Point Temperature (dwpf), Wind Speed (sknt), Wind Gust (gust), Average Speed (avg_speed), and Traffic Volume (normal_vol) are predictive of the Number of Crashes per hour. However, the adjusted R Squared value is significantly less (0.03789) which explains only roughly 3.7% of the variations in the Number of Accidents. The correlation between predicted values and testing data for number of accidents came only 0.2128064. Further the regression was cross validated with regularization. The result was as below.

```
(Intercept)  0.0297120845
hour         .
tmpf         .
dwpf         .
sknt         .
drct         .
gust         .
tfs          .
lane_id      0.0174244313
avg_speed   -0.0004638787
avg_headway  .
normal_vol   0.0004661011
long_vol     .
occupancy    .
```

The predictive features of the model were observed as the Lane ID (lane_id), Average Speed (avg_speed), and Traffic Volume (normal_vol). The correlation with predicted values and the test data was found to be only 0.1924064. This correlation was observed as less than the one for Multivariate Linear Regression without Regularization. We can conclude that the regularization did not improve the model.



The Residuals were not normal as per the QQ plot above. It was concluded that the Linear models were not suitable for the data being used for vehicular crash prediction.

Zero-Inflated Model

Zero Inflated models are used in scenarios where majority of outcome variables are zeros. For modelling the accident occurrence, the traffic volume (normal_vol) was used as the logit part of the model for explaining the higher number of non-crash situations. The rest of the variables were used to model the accident occurrence. The result and the confusion matrix was found to be as below.

```
Pearson residuals:
     Min       1Q   Median       3Q      Max
-1.04648 -0.15965 -0.11468 -0.08177 24.33535

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.3901041  0.1328326   2.937 0.003316 **
hour         0.0150260  0.0040066   3.750 0.000177 ***
tmpf         0.0258727  0.0055309   4.678 2.90e-06 ***
dwpf        -0.0117290  0.0031258  -3.752 0.000175 ***
sknt        -0.2288127  0.0152308 -15.023  < 2e-16 ***
drct        -0.0024946  0.0002048 -12.181  < 2e-16 ***
tfs         -0.0192291  0.0032347  -5.945 2.77e-09 ***
avg_speed   -0.0530864  0.0018707 -28.377  < 2e-16 ***
gust         0.1533793  0.0104694  14.650  < 2e-16 ***
long_vol     0.0335939  0.0039087   8.595  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.79259    0.08331   21.52   <2e-16 ***
normal_vol  -0.48772    0.03085  -15.81   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 40
Log-likelihood: -9584 on 12 Df
```

	Predicted		
Actual	0	1	Row Total
0	30599	3	30602
	0.977	0.000	
1	731	0	731
	0.023	0.000	
Column Total	31330	3	31333

We can observe that all the features are significant predictors of crash occurrences. The logit part (Traffic Volume) is significant. It says with every unit traffic increase, the odds of generating

zeros goes down by 0.48772. However, as per the confusion matrix, it can be observed that the model accurately predicts nonoccurrence of the crashes, but it fails to predict occurrences of the crashes.

In case of regression using the number of accidents per hour as the outcome variable, following results was observed.

```
Pearson residuals:
     Min       1Q   Median       3Q      Max
-1.15591 -0.16073 -0.11392 -0.08214 40.65090

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.809157   0.125163   6.465 1.01e-10 ***
hour         0.018410   0.003937   4.677 2.92e-06 ***
tmpf         0.019296   0.005383   3.584 0.000338 ***
dwpf        -0.007507   0.003029  -2.479 0.013190 *
sknt        -0.231428   0.014704 -15.740  < 2e-16 ***
drct        -0.002612   0.000196 -13.328  < 2e-16 ***
tfs         -0.017882   0.003136  -5.701 1.19e-08 ***
avg_speed   -0.054965   0.001778 -30.910  < 2e-16 ***
gust         0.159368   0.010074  15.821  < 2e-16 ***
long_vol     0.025042   0.004108   6.095 1.09e-09 ***

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.02096    0.07358   27.46   <2e-16 ***
normal_vol  -0.37680    0.02711  -13.90   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 44
Log-likelihood: -1.024e+04 on 12 Df


Total Observations in Table:  31333
```

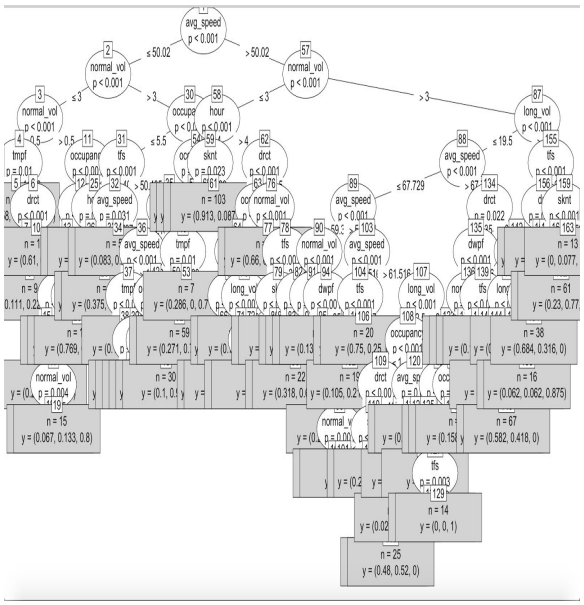
	Predicted				
Actual	0	1	5	149	Row Total
0	30592	8	1	1	30602
	0.976	0.000	0.000	0.000	
1	663	5	0	0	668
	0.021	0.000	0.000	0.000	
2	52	0	0	0	52
	0.002	0.000	0.000	0.000	
3	9	0	0	0	9
	0.000	0.000	0.000	0.000	
4	2	0	0	0	2
	0.000	0.000	0.000	0.000	
Column Total	31318	13	1	1	31333

For regression, all the variables are significant and the logit part (traffic volume) is also significant. As per the confusion matrix, the model predicted no crash situations very well, however, the number of crashes were not predicted correctly.

As the research question focuses on identifying vehicular crashes, the large number of type 1 errors are fine if the type 2 errors are fewer. In these models, since the type 2 errors are large in number, it was concluded that this method was not suitable for predicting the vehicular crashes using the available data.

C-Tree

In the plotted c-tree as below, the variables at the top are the most deterministic predictors. This implies that the most significant variables in the c-tree are avg_speed, normal_vol, long_vol. In the tree "n" stands for the number of records that satisfies the condition of the branch. "y" has three values and its value varies between 0 and 1. The three values of "y" correspond to the class 0,1 and 2 respectively. It tells the fraction of nodes that belongs to each class. The sum of the values of y is always equal to 1.



Random Forest

The final aggregated data had over 125,000 records from 2016. The resultant dataset was a highly unbalanced dataset with 2.5% of records corresponding to logged crash occurrences. Random Forest model was trained against the following feature set: "hour", "tmpf", "dwpf", "sknt", "drct", "gust", "tfs", "avg_speed", "avg_headway", "normal_vol", "long_vol", "occupancy" and the dependent variable was the indicator of whether accident log was present for corresponding conditions.

Performance of ensemble machine learning technique of Random Forest was evaluated under following scenarios:

a. **Binary classification problem with upscaling of minority classes:**

Aggregated dataset consisted of numerical crash occurrences for each hour slice. The numerical occurrences were converted into 0 or 1 where 0 indicates no crash occurred and 1 indicated at least 1 crash instance was logged. The training dataset was balanced such that resulting data set had equal proportions of records corresponding to crash and noncrash instances. For balancing the datasets minority classes were artificially synthesized through upscaling techniques to match majority class of no crash occurrences. The trained model was fitted against the testing sample. Below was the result observed:

Total Observations in Table: 31333

	Predicted		
Actual	0	1	Row Total
0	30504	96	30600
	0.974	0.003	
1	700	33	733
	0.022	0.001	
Column Total	31204	129	31333

The ROC curve for the model was very poor in this setup and the accurate prediction are likely by chance.

b. Binary classification problem with downscaling of majority classes:

In this scenario, the balancing technique was changed from upscaling of minority classes to down scaling of majority classes. The models was trained and tested. It was observed with the change in balancing technique predictability of model improved significantly. On an average, the model was able to predict with predict crash occurrences with an accuracy of over 65%. It was observed that over 20% of the records were incorrectly predicted as crash occurrences (False positives). This however, was acceptable as it makes the model only more stringent.

	Predicted		
Actual	0	1	Row Total
0	23692	6869	30561
	0.756	0.219	
1	258	514	772
	0.008	0.016	
Column Total	23950	7383	31333

c. Classification problem with 3 classes: 0 - No crashes, 1 - One crash, >1 - For more than one crash logged in an hour. Upscaling of minority class of >1 to match with class 1 and downscaling of majority class of 0 to match with class 1:

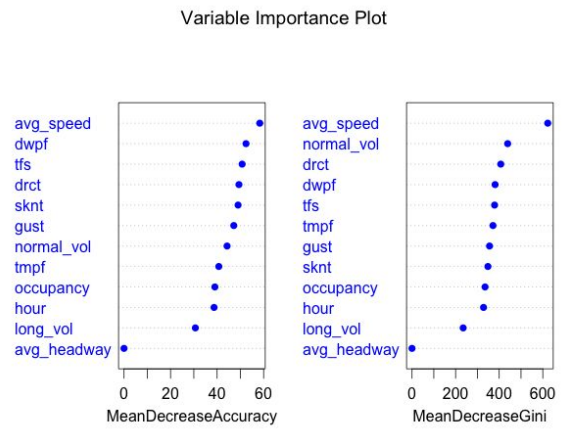
Extending the approach to three classes, we calculated our outcome variable of crash frequencies indicating scenarios were no crashes were reported as class 0, only one crash reported as 1 and more than one crash reported as class >1. For balancing the data, we upscaled the number of minority class records of >1 to match frequency of number of records with class 1, we also downscaled the majority class of 0 to match with the frequency of records with class 1.

It was observed that the over 65% of crash occurrences with one crash reported were correctly predicted and about 4% of instances with more than one crash reported were correctly predicted. It is also observed that out of 63 instances of more than one crash instances for a given hour over 50 instances were predicted as class 1 indicating the model is able to identify crash instances from non crash instances. However, there is a spill in accuracy of records with class >1 into class 1, but they are still being predicted as crash instances nonetheless.

Total Observations in Table: 31333

	Predicted			
Actual	0	1	2	Row Total
0	23048	7505	49	30602
	0.736	0.240	0.002	
1	200	456	12	668
	0.006	0.015	0.000	
2	6	54	3	63
	0.000	0.002	0.000	
Column Total	23254	8015	64	31333

Below graph indicates the feature importance of variables in terms of accuracy and their ability to classify records in pure subsets of outcome variables. It is observed that average speed, dew point temperature and pavement temperature are most important features for the model's accuracy.



Conclusion

We compared various machine learning approaches that took spatial and temporal features associated with historic crashes occurrences and attempted to create a predictive model to determine crash occurrences given weather and traffic related informations from RWIS stations in IOWA. In our analysis we used Zero Inflated Negative Binomial Models, Multivariate Linear Regression, Random Forests with different levels of outcome variables and different balancing techniques. We found that Random Forest with down sampling for classification problem with two levels of crash occurrences (No crash occured vs crash occurred) predicted over 65% of crash instances correctly. For prediction with outcome variable as a three level class (No crash occurred, one crash occurred, and more than one crash occurred), and balancing the number of instances with majority class of no crashes and minority class of more than one crashes to match number of instances with one crash, we found the model was able to predict classes with one crash with an accuracy of over 65%.

Discussion

As a future scope to this research, we believe that more features which are related to the accidents such as Road geometry, Day of the week, and lighting condition can be used for explaining the remaining variance in our models. The accuracy of the models can also be increased by including the data for previous years 2015, 2015 and more in model building as well. As a potential use case, the developed models, can be used along with live Road Weather Information System data as APIs, to predict the vehicular crash occurrences in real time. This will help Highway Authorities in taking precautionary actions towards avoiding crashes, and also caution the drivers to be more careful, and thereby making the street more safe for travel.

References

Abdel-Aty, M. A., Pemmanaboina, R. (2006). Calibrating a Real- Time Traffic Crash-Prediction Model Using Archived Weather and ITS Traffic Data, vol. 7(2)

Bin Islam, M. and Kanitpong, K (2008). Identification of factors in road accidents through in-depth accident analysis. In IATSS Research Vol.32 No.2, 58-67

C. Oh, J. Oh, and S. Ritchie, “Real-time estimation of freeway accident likelihood,” presented at the 80th Annu. Meet. Transportation Research Board, Washington, DC, USA, 2001.

Chang, L. and Wang, H (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis & Prevention. Volume 38, Issue 5, 1019-1027

Chen, F., Chen, S., Ma, X. (2016). Crash Frequency Modeling Using Real Time Environmental and Traffic Data and Unbalanced Panel Data Models, International Journal of Environmental Research and Public Health

Chong, M., Abraham, A. and Paprzycki, M (2005). Traffic Accident Analysis Using Machine Learning Paradigms . Informatica 29 (2005), 89–98

Goodwin, L. C. (2002). Analysis of Weather Related Crashes on U.S. Highways

Lee, C., Hellinga, Bruce., Saccomanno, F. (2003). Real-Time Crash Prediction Model for the Application to Crash prevention in Freeway Traffic, pp. 03-2749

Lord, D., Manar, A., Vizioli, A. (2004). Modelling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. *Accident Analysis and Prevention 37 (2005) 185-199*

Lord, D. and Mannering, F (2010). The statistical analysis of crash frequency data: A review and assessment of methodological alternatives. Transportation Research Part A, 44, 291-305

Pisani, A. P., Goodwin, C, L., Rossetti, A.M. U.S. Highway Crashes in Adverse Road Weather Conditions

Srisuriyachai, S. (2007). Analysis of road traffic accidents in Nakhon Pathom province of Bangkok using data mining. Graduate Studies. Bangkok, Mahidol University

Yu, R., Abdel-Aty, M. A., Ahmed, M. M., Wang, X. (2014). Utilizing Microscopic Traffic and Weather Data to Analyze Real-Time Crash Patterns in the Context of Active Traffic Management, vol. 15(1)