Analysis of the White-House Road Fatalities Data (2015)

Dwipam Katariya (ddkatari), Neelam Tikone (ntikone)

***Abstract*— Analyzing the Road Fatalities all over the USA in the year 2015. Analyzing the effects of unsupervised clustering on the categorical data to analyze different types of road fatalities. Utilizing Supervised approaches to predict and study various factors correlated and causal to Drunken Driving. Study how clustering can play a major role in modeling.**

Machine Learning Algorithm — K-means, K-modes, Decision Tree, Gradient Boosting Machine, SVM.
Programming Languages — R, Python
API—Numpy, Pandas, Matplotlib, H2O, sklearn, graphlab, Google Embedding Projector
Tools —ARCGIS

I. INTRODUCTION

As there has been an increase in the average road fatalities each year more than the previous, the government has released the data for the road Accidents for the year 2015 so as to get some insights from the data which can help combat the amount of fatalities. Currently, on an Average, 92 people get killed due to road accidents in the USA each day.

The some of the various variables which are initially considered for cluster analysis are (All categorical variables):

1. Number of people travelling
2. Route
3. Harmful Environmental condition
4. Type of Road Intersection
5. Weather condition
6. State
7. Time of Day
8. Day of Week
9. Drowsy Driver
10. Pedestrian Involved.

II. SECTION 1 : PRE- MODELLINGIN R

1. Loading the train set data and the US county data into the data frames: This step involved loading the data into the R data frame.
2. Merging the data: Merging the Accident dataframe with the US county dataframe with respect to the state names.

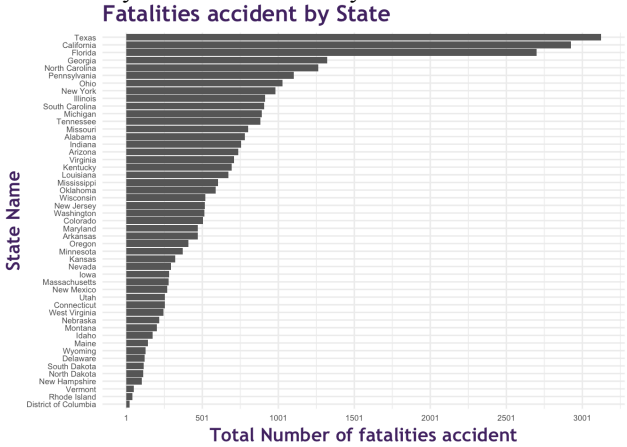
III. SECTION 2: FEATURE ENGINEERING

1. Identifying NA values: Before applying feature transformation we, checked the presence of NA values and we removed the rows containing those NA values from the data frame.
2. Taking in account the most important features and removing the unwanted features with the help of domain expert and the list of specified important variable list.
3. Removing outliers:
   Following ways to remove outliers: Removed fatalities which were greater than 4 as they were not significant.
   Removed the unknown and unrecorded data which had less frequency on variables: Relationship_to_road, Route, Special_Jurisdiction, Junction, Highways, Work_Zone, Harm_Event.
   We looked at Multi variate data with respect to driver age, drunken_driver and removed the samples which seems to be outliers because we found that there were few type of fatalities in state Washington with harmful event and no drunken driving. There were two rows with drunken driving. Hence, were moved such rows. We looked at univariate boxplots.

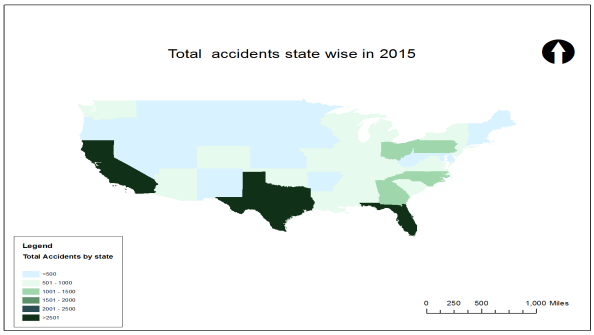
IV. SECTION 3: EXPLORATORY DATA ANALYSIS IN R

Used the library 'ggplot2' to plot data.

1. **Accidents per state:**
   The plot of the number of accidents in each state for the year 2015 was created. It is seen in the below figure 4.1.1 that the state 'Texas' is the highest for the year 2015 followed by California.

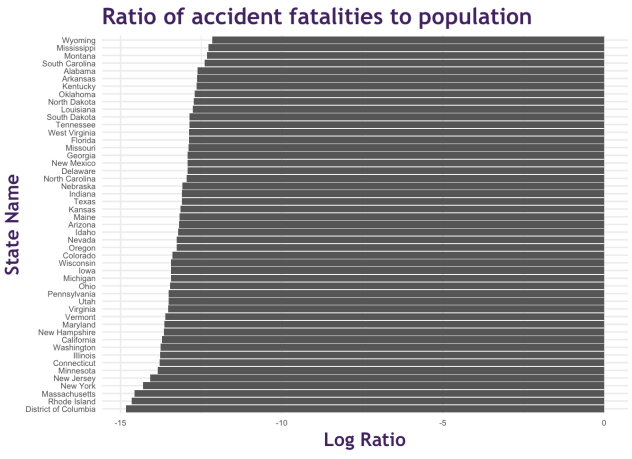


(figure 4.1.1)

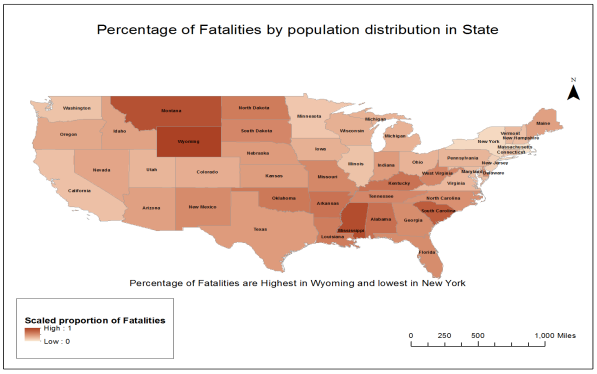


(figure 4.1.2)

As the above states showing the highest number of fatalities are also highly populated, we normalized it using the population count of each state. The *figure 4.1.3* shows us the graph normalized by population.



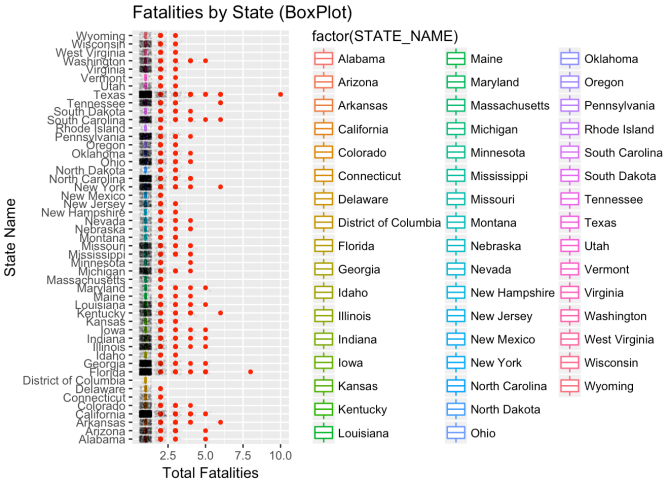
(figure 4.1.3)



(figure 4.1.4)

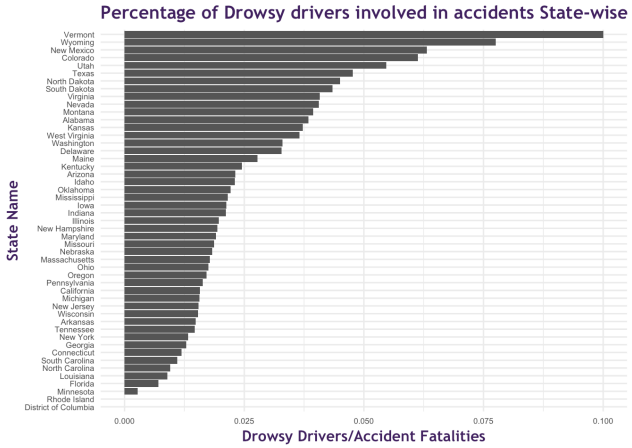
After normalizing according to the population of each state, it has been observed that Wyoming and Mississippi has the highest number of accident fatalities.

The figure 4.1.5 shows us the box plot for fatalities of each state. The boxplot shows us that the fatalities are not much affected by the state.

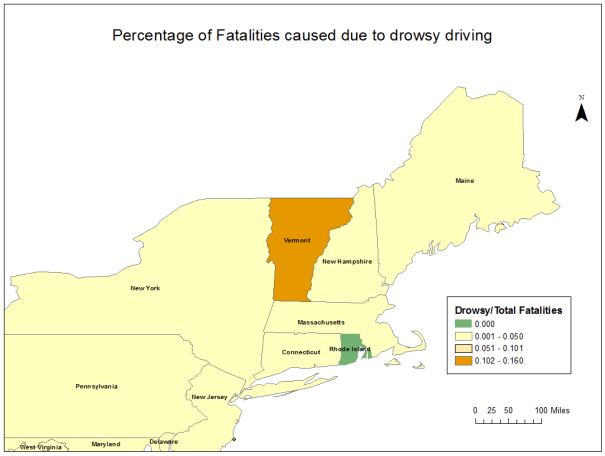


(figure 4.1.5)

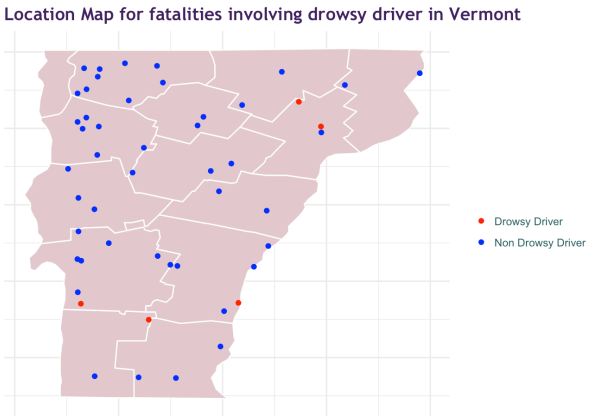
2. **Accidents due to drowsy drivers:**
   The number of accidents sue to drowsy drivers was plotted (figure 4.2.1) and we found that Vermont has the highest number of accidents due to drowsy drivers.



(figure 4.2.1)



(figure 4.2.2)



(figure 4.2.3)

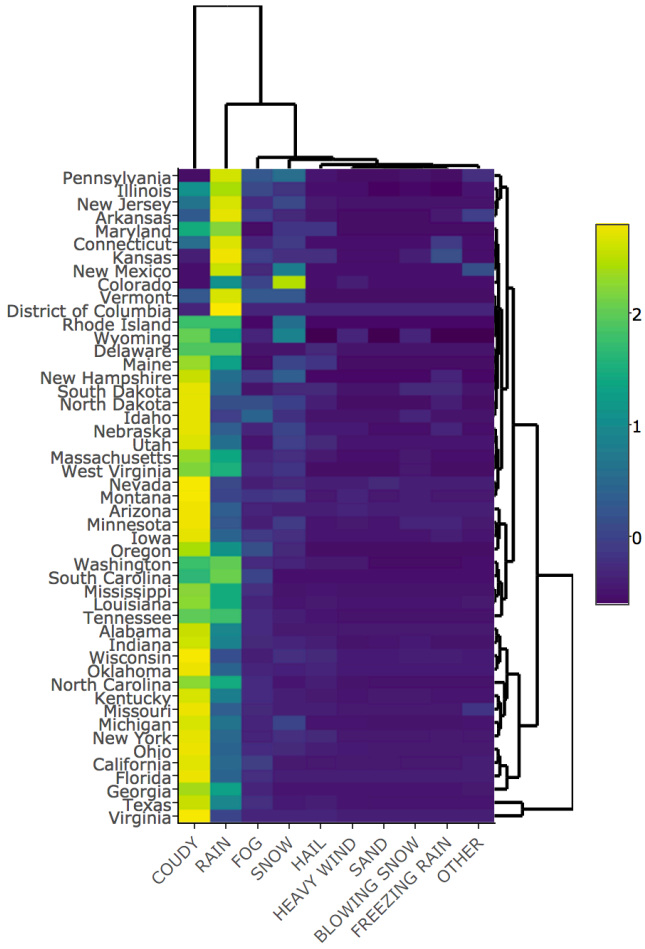
Rhode Island has no Fatalities Involving Drowsy driving. Vermont has 16% of fatalities Involving drowsy driving. According to USA today:
www.usatoday.com/story/money/careers/2015/07/05/10-worst-states-to-make-a-living-2015/29626223
Vermont is the 5th worst place to work because of the bad work culture and highest number of night shifts which can be one of the reasons for the drivers being drowsy.

3. **Drunken driver Dependency on age group:**
   Chi-squared test was performed to check the dependency of the drunken driver fatalities on the age group. It was observed that the P-value is less than the significance value (figure 4.3.1). Hence, we can say that the drunken driver fatalities are not independent of the age group.

```
##
##  Pearson's Chi-squared test
##
## data:  ageDrunkFat
## X-squared = 645.78, df = 2, p-value < 2.2e-16
```

(figure 4.3.1)

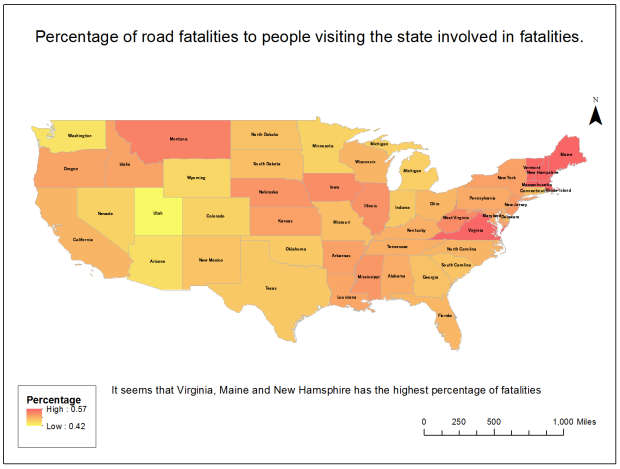
4. **Effect of weather on road fatalities:**
   The following weather conditions were taken into account to find out if any of them affect the road fatalities.



(figure 4.4.1)

Heatmap was drawn to check the relationship. The lighter the color, the higher are the accidents observed in that particular weather. Heatmaps have cells arranged as per the fatalities in weather with scaled data. It's obvious that Clear weather would be having Larger amount of accidents, however it is interesting to see that after Clear weather, Cloudy has the highest accidents and then Rainy. We can also see that California and Texas are clustered together and its well-reasoned that both states share approximately similar weather conditions, fatalities and population.

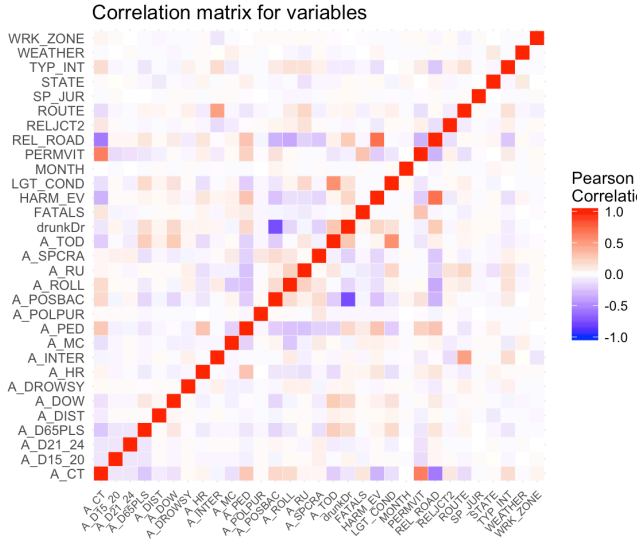
5. **People visiting the state involved in fatalities.**
   The below map shows the total number of the fatalities compared to the total number of people visiting in each state. Maine has 57% of 155 people involved in accidents. Whereas, Virginia has 54% of 711 people involved in fatal accidents.



(figure 4.5.1)

6. Correlation between variables:

The figure 4.6.1 gives us the correlation matrix which shows how much each variable is correlated with the other variable.

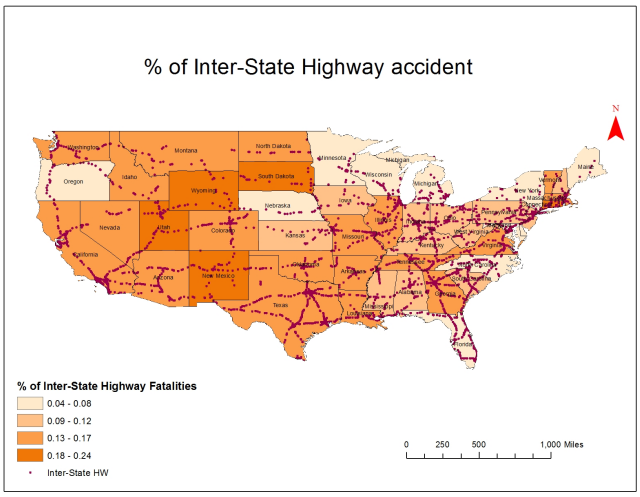


(figure 4.6.1)

It seems that Types of crashes such as one car crash, two car crash are correlated to Number of people fatalities, and types of event. Also Positive check of BAC and drunken driver are correlated, hence we can think to remove some correlated variables, as correlation won't reduce the power of prediction, as correlated variables does not mean they are dependent.

7. Interstate highway fatalities:

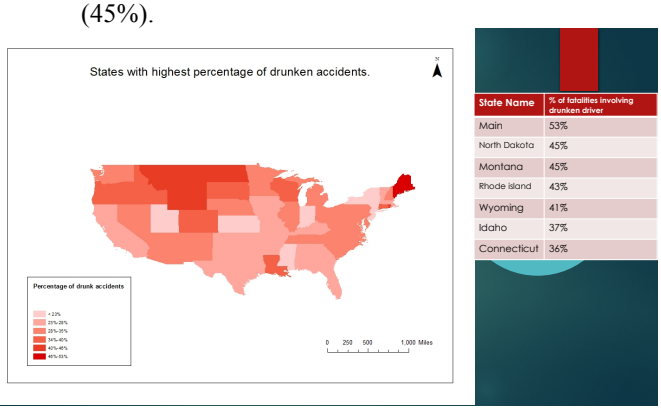
It is observed that New Mexico has the highest percentage of Interstate accident (24%) followed by South Dakota (23%). The below map shows us the fatalities on the interstate highway throughout the USA.



(figure 4.7.1)

8. Drunken Fatalities in each State:

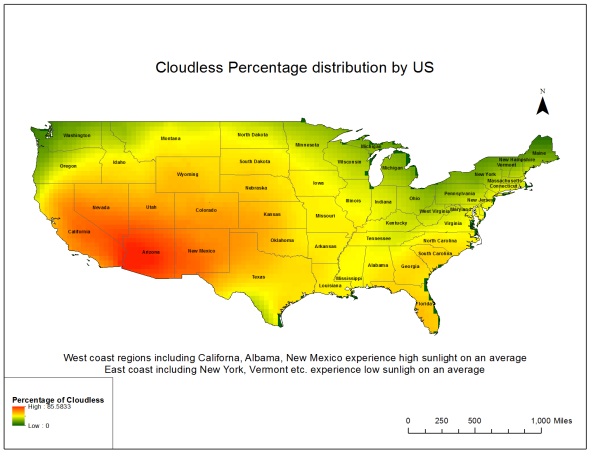
The below map shows us the drunken accidents which took place in each state. Its observed that Maine has the highest amount of accidents due to drunken drivers (53%) followed by North Dakota (45%).



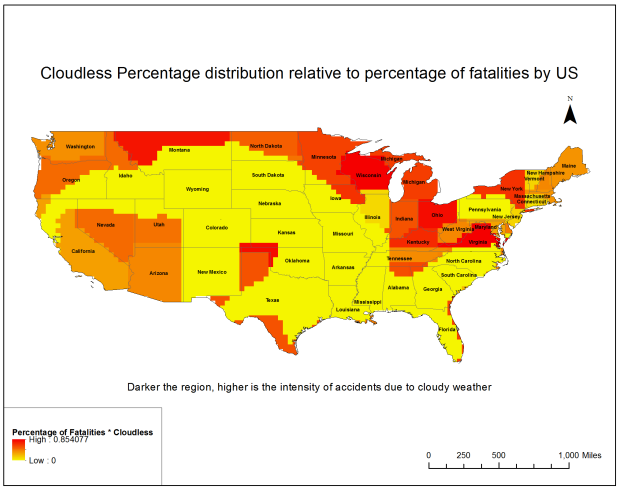
State Name	% of fatalities involving drunken driver
Main	53%
North Dakota	45%
Montana	45%
Rhode island	43%
Wyoming	41%
Idaho	37%
Connecticut	36%

(figure 4.8.1)

9. Effects of Cloudy weather:



T



Scaled the cloudlessness with the percentage of accidents per population. States like California , Arizona who experience low cloudy weather are scaled to lower scale while states like Wyoming, Washington are relatively scaled to higher scaled.
This means higher the effect of cloudiness, darker the region in the above map. This was converted from Vector data to raster data for the purpose.

V. K-MEANS CLUSTERING

To see how many similar types of Fatalities are involved we can perform K-means clustering. As we have categorical data we will evaluate the results based upon K-means and K-modes.

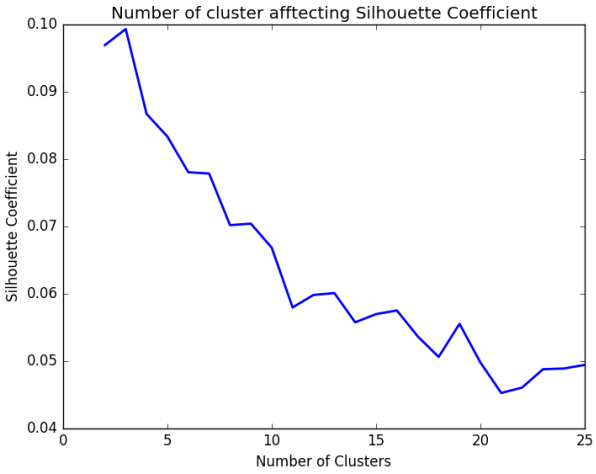
1. **One- Hot Encoding**:
   The data consisted of categorical variables. Hence to perform K-means clustering using Euclidian distance as a distance metric, we performed One-Hot Encoding on the data.

2. **Choosing appropriate K:**
   As we used max 300 iterations for converging the K-Means algorithm. We tried K from 2 to 25 and the appropriate K was chosen by comparing the Silhouette Coefficient and the Calinski-Harabaz Index.
   a) Silhouette Coefficient:
   Figure 5.2.1 shows the coefficient mapped against number of clusters. Silhouette number of clusters. Which is given by the formula:

$$s = \frac{b - a}{max(a, b)}$$

a: The mean distance between a sample and all other points in the same class.

b: The mean distance between a sample and all other points in the next nearest cluster.
The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters.
The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.

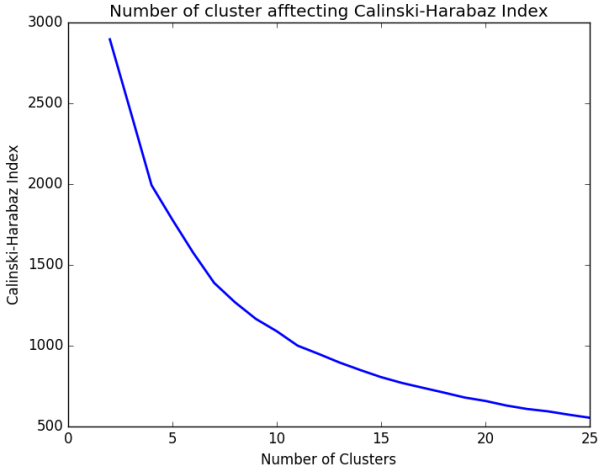


(figure 5.2.1)

b) Calinski-Harabaz Index:
Figure 5.2.2 shows the Calinski-Harabaz Index mapped against number of clusters.
For k clusters, the Calinski-Harabaz score s is given as the ratio of the between-clusters dispersion mean and the within-cluster dispersion.

$$s(k) = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N - k}{k - 1}$$

The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.



(figure 5.2.2)

For a K of 3 we observed the highest Silhouette coefficient and the highest Calinski-Harabaz Index. Hence we chose our K to be 3.

1. **Distance Metric:**
We used Euclidean distance to calculate the distance between two points.

$$dist((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

2. **Cluster Analysis:**
As we have performed One-hot encoding i.e. converting to dummy variable, each column for each centroid specifies the percentage of true labels associated to particular centroid.

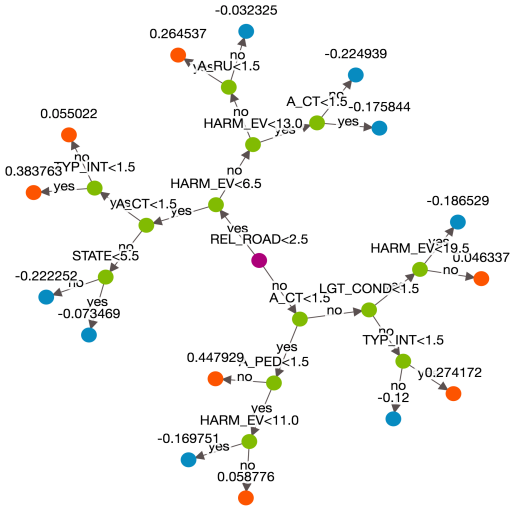
This is because while computing new centroid, we add up all the points i.e. adding just one and then dividing by total points, that is percentage of points having the truth. Looking at the three centroids we have following preliminary analysis:

Clusters	***Percentage of Data associated***
1	40%
2	42%
3	18%

Clusters	1	2	3
Pedestrian Involved			
1	62	113	4275
2	2	10478	11196

cluster	*1*	*2*	*3*
Relationship to Road			
On Roadway	930	11020	4516
On shoulder	150	84	108
On Median	847	72	30
On Roadside	8159	131	106
Outside Traffic way	413	1	23
Off roadway	41	1	3

3. **Decision Trees for K-Means Clustering:**
To observe what differentiates the clusters formed above, we used decision tree to find the most influential variables.

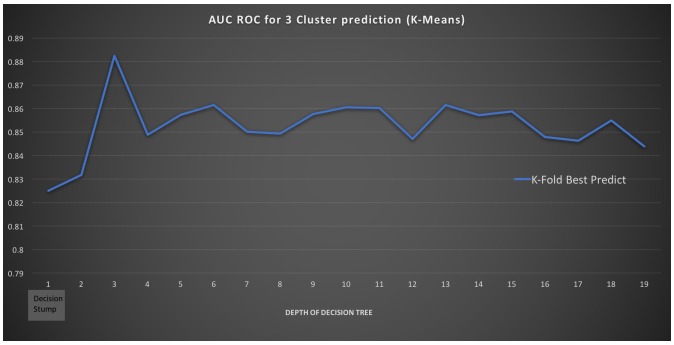


(figure 5.4.1)

Cluster Confusion Matrix:

Target Cluster	Predicted Cluster	Frequency
3	2	8
2	3	16
2	1	22
3	3	910
1	1	2048
1	3	28
1	2	33
3	1	39
2	2	2251

AUC-ROC = 88%
Accuracy = 98.4%
Depth = 3
Test data = 20% sampled
Most influential variables as per priority: Relation to Road, Harmful Event, Crash Type, Pedestrian involved, Light Condition.



(figure 5.4.2)

VI. K-MODES CLUSTERING

K-modes centroids are effective if all the values in the dataframe are categorical. K-Modes centroids are the

categorical values that are majority for the given cluster (Modes)

1. **Distance Metrics:**
   Distance measure is the Huang distance which is given by:

$$d(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j)$$

where,

$$\delta(x_j,y_j)=\begin{cases}0, & x_j=y_j\\ 1, & x_j\neq y_j.\end{cases}$$

Hence higher the distance measure, better the clusters.

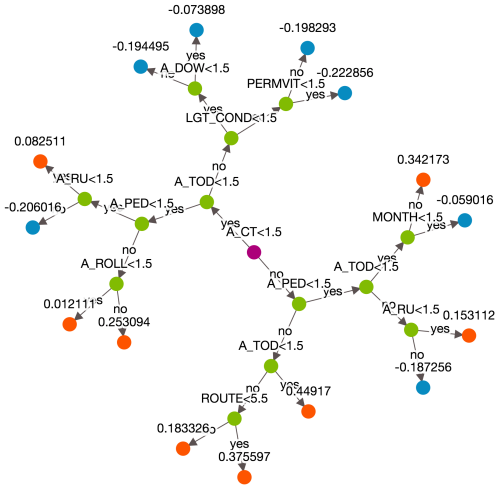
2. **Cluster Analysis:**

Clusters	Percentage of Data associated
1	51%
2	27%
3	22%

Clusters	1	2	3
Pedestrian Involved			
1	264	86	4100
2	13548	6884	1743

cluster	1	2	3
Relationship to Road			
On Roadway	10979	725	4762
On shoulder	158	76	108
On Median	316	463	170
On Roadside	2115	5557	724
Outside Trafficway	140	225	72
Off roadway	14	24	7

3. **Decision Tree for K-Modes:**
   To observe what differentiates the clusters formed above, we used decision tree to find the most influential variables

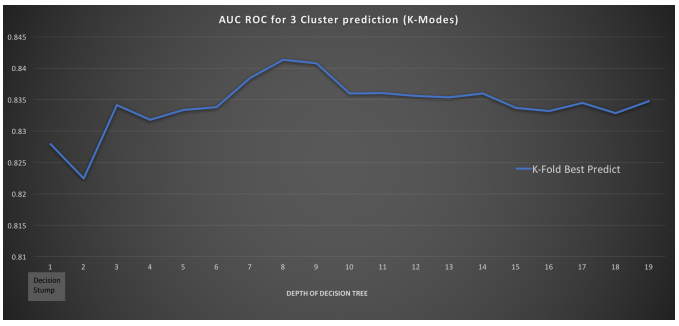


Confusion Matrix:

target_cluster	predicted_cluster	Frequency
3	2	40
2	2	1074
1	3	181
1	2	69
3	1	76
2	1	106
2	3	138
3	3	1213
1	1	2564

AUC-ROC = 84%
Accuracy = 98.4%
Test data = 20% sampled
Most influential variables as per priority: Crash type, Pedestrian Involved, Time of Day, Light Condition and Persons involved

The figure 6.2.1 shows us the AUC ROC for 3 cluster prediction in K-Modes.

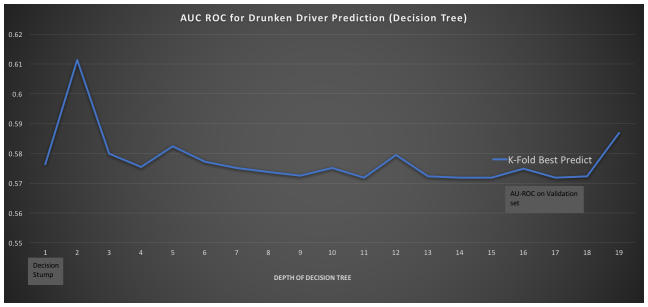


(figure 6.2.1)

VII. PREDICTING DRUNKEN DRIVERS

1. **Balancing the data:**
   As the data was skewed, as the examples having the fatalities including the drunken drivers were very few, we had to balance the data to remove the skewedness.
   We tuned the parameters for different k, where k is number of nearest neighbors to construct synthetic samples and number of nearest neighbors to determine if minority sample is a danger. We tune kind of smote with regular borderline and SVM.

2. **Applying balanced data to Decision Tree:**
   After doing 10-fold cross validation and changing the depth indecision trees from 1 to 20, depth 2 had 62% accuracy on test data and 61% accuracy on validation set.



(*figure 7.2.1)*

Confusion Matrix:

target	*predicted*	*count*
0	0	2242
0	1	1671
1	1	1074
1	0	340

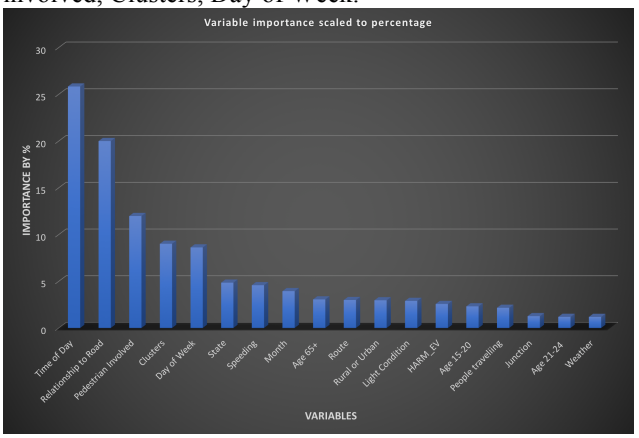
3. **Applying balanced data to Gradient Boosting Machine:**
   We tune the following parameters. As this was a binary classification, we kept the distribution as Bernoulli, number of trees, max_depth, learning rate, n-folds. We performed grid search to get the hyperparameters.
   We found the best AUC ROC on test data with the above parameters with precision 0.99, specificity 0.99, AUC 0.80.
   MSE: 0.153088003111
   RMSE: 0.391264620316
   LogLoss: 0.499809385663
   Mean Per-Class Error: 0.270416838515
   AUC: 0.799462929755
   Gini: 0.598925859509

Confusion Matrix:

	0	1	Error Rate
0	3137	776	0.1983 (776.0/3913.0)
1	485	929	0.343 (485.0/1414.0)
Total	3622	1705	0.2367 (1261.0/5327.0)

Important feature can be detected by:
(number of times a feature occurred/ total trees)

Following are the important variables which we found:
Time of day, Relationship_to_road, Pedestrian involved, Clusters, Day of Week.



4. **Applying SVM:**

RBF Kernel, Uniform Class weights:

	precision	recall	f1-score	support
0	0.79	0.92	0.85	3913
1	0.61	0.34	0.44	1414
avg / total	0.75	0.77	0.74	5327

RBF Kernel, Balanced Class weights:

	precision	recall	f1-score	support
0	0.87	0.74	0.80	3913
1	0.49	0.69	0.57	1414
avg / total	0.77	0.73	0.74	5327

Linear Kernel Uniform Class weights:

	precision	recall	f1-score	support
0	0.83	0.89	0.86	3913
1	0.61	0.49	0.54	1414
avg / total	0.77	0.78	0.77	5327

Linear Kernel Balanced Class weights:

	precision	recall	f1-score	support
0	0.87	0.73	0.79	3913
1	0.48	0.69	0.57	1414
avg / total	0.77	0.72	0.74	5327

precision recall f1-score support

It seems that RBF Kernal with uniform class weights perform better than other parameters.

VIII. CONCLUSIONS

The most important variable is Time_of_day, Relationship_to_Road, Crash_Type, Pedestrian_involved which plays a major role in predicting if the driver was drunk or not.
Decision trees performs bad on imbalance dataset due to the property of Entropy function and it overfits to different depths and hence it lacks good prediction on unseen data.
SVM computes solution globally using the dual formulation. Hence, applying linear and RBF Kernel was a good idea.
GBM outperforms all the approaches utilized in this paper.
Looking at the important variables, relative measures to combat road accident fatalities can be taken by FARS.
Cluster Analysis proved beneficial while exploring different types of fatalities and hence the clusters formed some importance in variable importance, though manually selecting features out of 250 features require domain knowledge and explicit definition of variables proved to be beneficial.
Removal of outliers played a major role in modelling.

IX. REFERENCES

[1] https://www.whitehouse.gov/blog/2016/08/29/2015-traffic-fatalities-data-has-just-been-released-call-action-download-and-analyze - DATA COLLECTION
ftp://ftp.nhtsa.dot.gov/fars/2015/

http://scikit-learn.org/stable/documentation.html - SKLEARN DOCUMENTATION
http://docs.h2o.ai/h2o/latest-stable/index.html - H2O DOCUMENTATION
http://www.irma-international.org/viewtitle/10828/ - Kmodes clustering
http://grid.cs.gsu.edu/~wkim/index_files/papers/kprototype.pdf -k-prototype