Accident analysis on Allegheny County Intersections

Group members:
Marianne Liu, Weixuan Sun, Arushi Vyas, Alexander Waldron

Table of Contents

1.Business Objective

Car Accidents occur everyday with the majority of them happening in an intersections. The Federal Highway Administration reports 2.5 million accidents happen intersections and 20% of those accidents are fatal.[1]

Unfortunately, it is a morbid business but somebody must carry the costs for the damages done as the average person may not have the capital to cover the expenses of an accident, so they buy auto insurance.

That raises the question, how does the insurance company construct its policies so that they do are not losing money on claims, especially large ones? Additionally, is it possible to know what factors are the most commonly associated with the occurrence of a car accident; for example was there drunk driver involved? Statical models are able to help us answer these types of questions.

We plan to examine the local data from the Allegheny County Crash records to identify the key factors that are associated with fatal car accidents and those where at least one person was injured. With this information we are able to figure out how auto insurance companies create policies along developing ideas for our own policies that are different than existing policies already provided by local companies.

2. Data Exploration

2.1 Merging data

As we have two data sources:
1. Allegheny County Crash Data from 2004 - 2017
2. Municipality Codes: County Mapping with sub county name and code
We used Python to join the two sources on the sub county code.

```
merged = pd.merge(df, municipality, on='MUNICIPALITY')
```

After merging the data, we got 170,365 rows and 190 rows.

2.2 Feature Selection

To select features out of 190 available columns, we referred to Allegheny County Crash Data Dictionary explaining the metadata information.

We selected 35 columns which we believe would make the most sense out of all variables which contribute to the injury or fatal incidents in the allegheny county.

After selection, we filter the accident rows which only happens in the intersection:

```
columns = ['AGGRESSIVE_DRIVING','ALCOHOL_RELATED','AUTOMOBILE_COUNT',
           'CELL_PHONE','COLLISION_TYPE','CRASH_CRN','CRASH_MONTH',
          'CRASH_YEAR','DAY_OF_WEEK','DISTRACTED','DEER_RELATED','INJURY_OR_FATAL',
          'MAJOR_INJURY', 'MODERATE_INJURY','MINOR_INJURY','DRIVER_16YR','MUNICIPALITY','Municipality_Name',
          'DRIVER_17YR', 'DRIVER_65_74YR','DRIVER_75PLUS','MUNICIPALITY','HIT_DEER',
          'HOUR_OF_DAY','ILLEGAL_DRUG_RELATED','ILLUMINATION_DARK','INTERSECT_TYPE','PEDESTRIAN',
          'INTERSECTION','LOCATION_TYPE','ROAD_CONDITION','SPEEDING_RELATED','HIT_TREE_SHRUB',
          'URBAN_RURAL','WEATHER']
merged= merged[columns]
```

```
merged_1 = merged[(merged['INTERSECTION']== 1)]
merged_1.shape
```

```
(74273, 35)
```

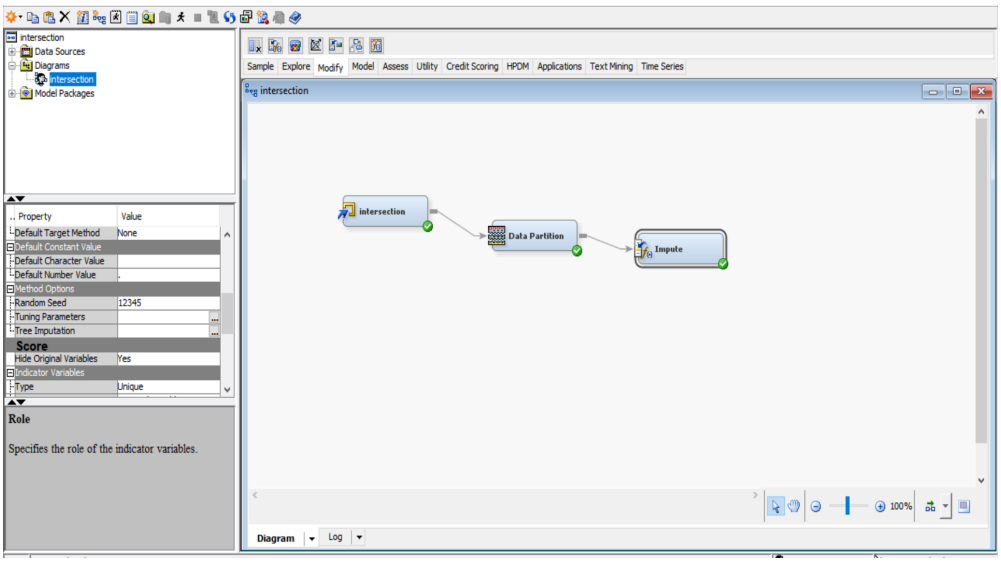
2.3 Column Encoding

According to the Allegheny County Crash Data Dictionary, columns such as “Weather_Type”,“Road_Condition”, “Intersect_Type” contains categorical values like 0,1,2,3,4, each of which represents a certain textual type.

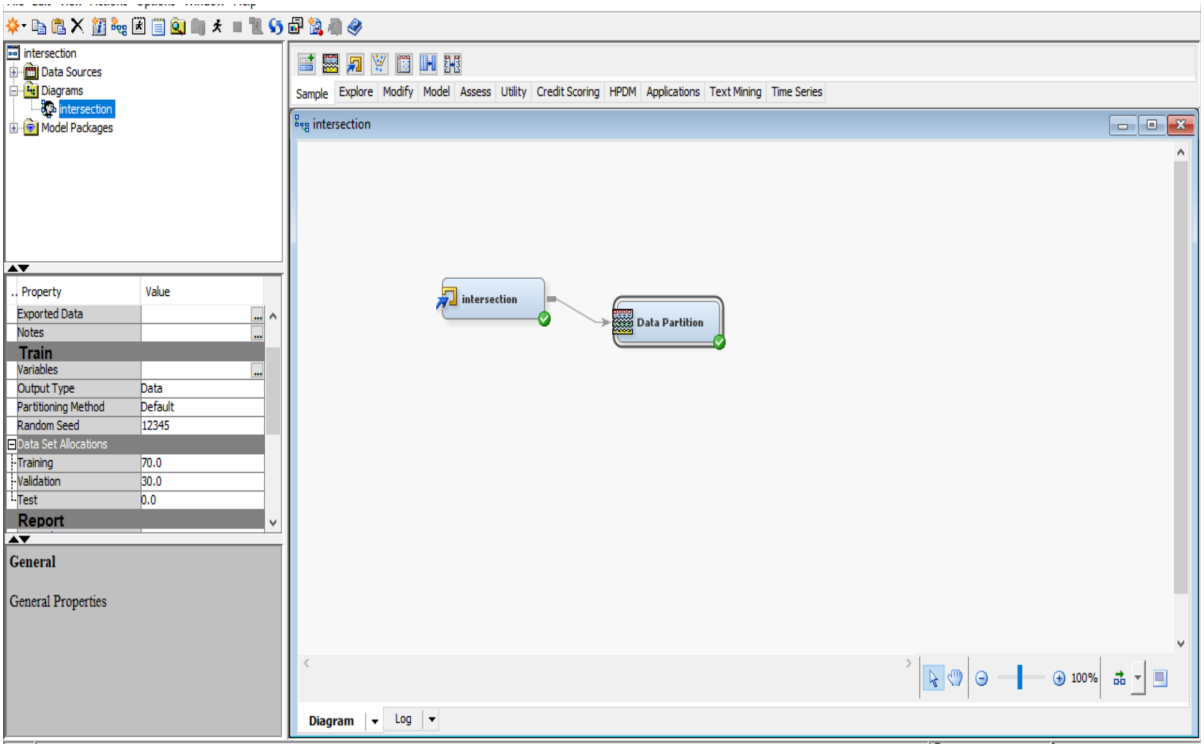
While conducting association rule learning, clustering, decision tree and logistic regression, it might be hard to interpret the result using those categorical values. Hence, we created “Weather_Type_Name”,“Road Condition_Name”, “Intersect_Type_Name” which contains the real textual values.

2.4 Manage Missing Values

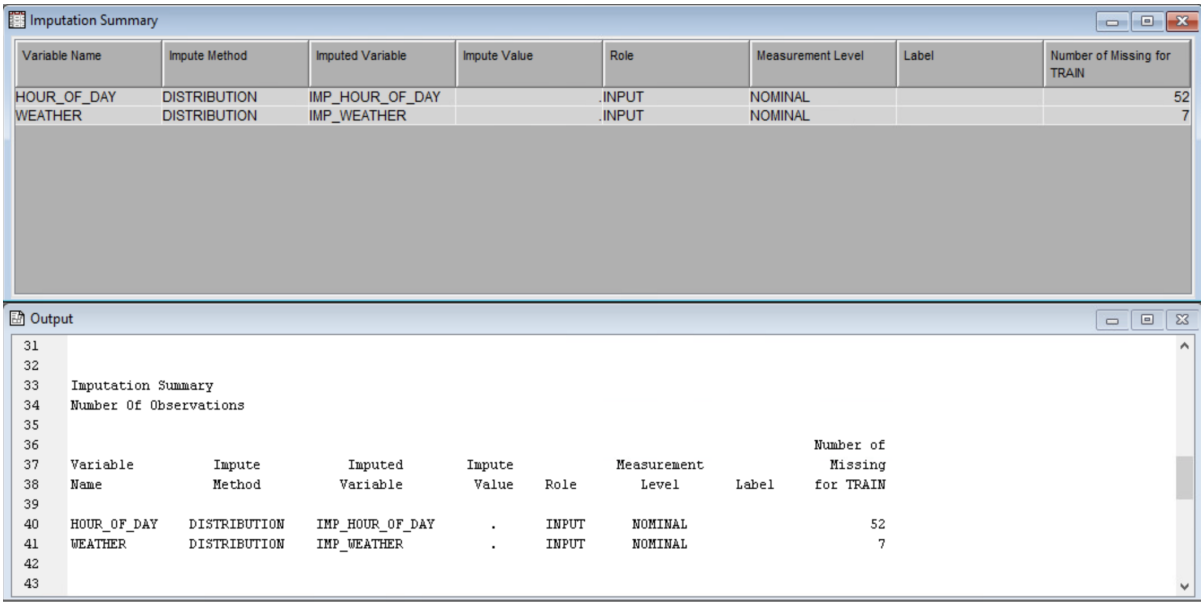
(1) Setting Municipality Name as ID and Crash type as Target, we got the variables result as follow:



(2) Before managing missing values, we add data partition to the diagram workplace. First, we change training to 70, validation to 30, test to 0 to see the result.



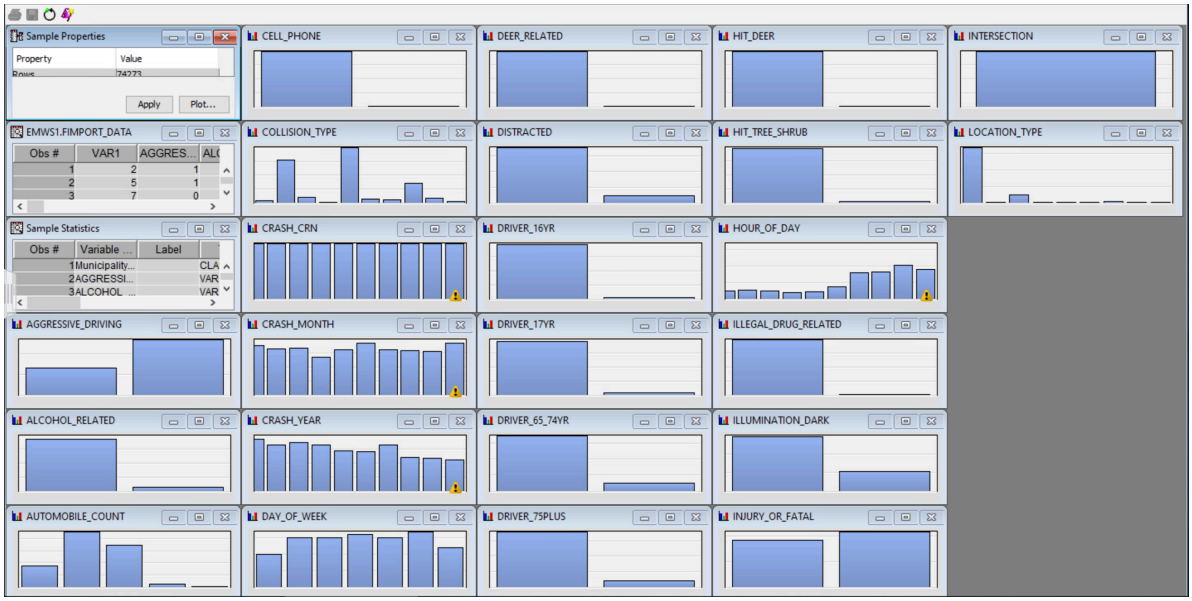
(3) By using impute method of Max and changing type-unique, Source-imputed variables, Role-input, we can see from the imputed variable and indicator variable that there are HOUR_OF_DAY and WEATHER. The numbers of missing for train are 52, 7. Compared to our large dataset, which is not big enough to affect our results.

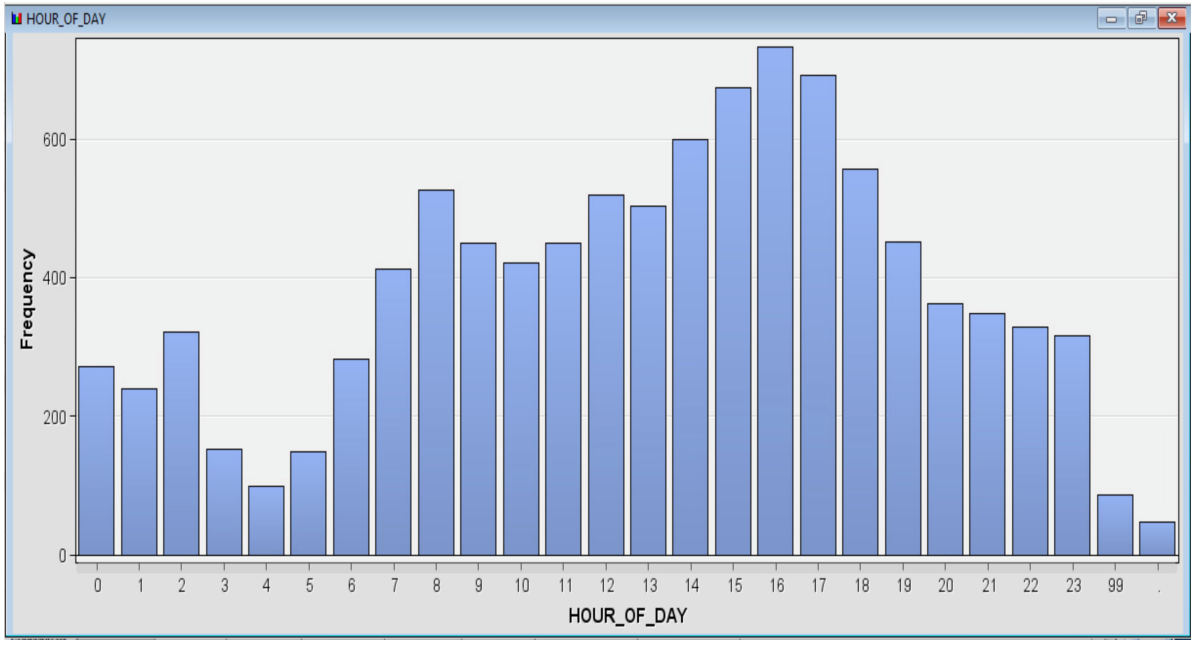


2.5 Distribution

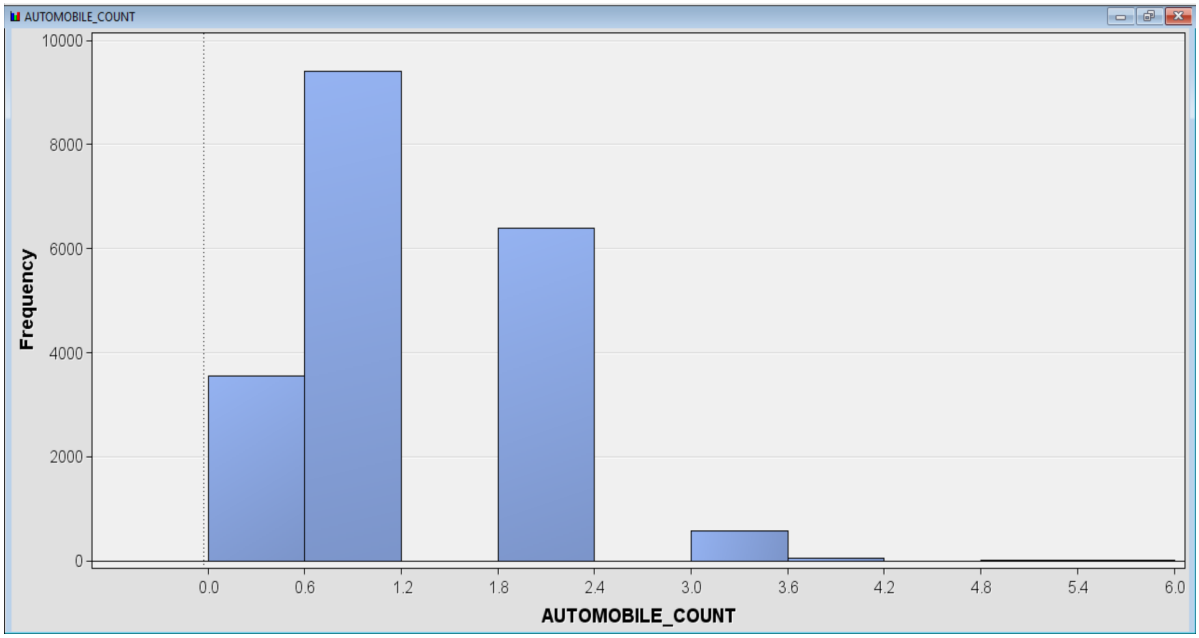
(1) Because we want to do logistic regression analysis, we set most of our variables to binary level.

(2) By exploring variables of the intersection data, we can see from the histogram that the majority of data distributions are not skewed, except for HOUR_OF_DAY and AUTOMOBIL_COUNT.





(3) We can see that in HOUR_OF_DAY, the overall distribution is a bit skewed towards right side, especially during 14:00 pm to 19:00 pm, which makes sense that crashes happen frequently in later time of the day due to poor illumination and traffic jam when people drive from downtown offices to residential areas.



(4) AUTOMOBILE_COUNT refers to generally how many automobiles crash together in an intersection. Intersection roads are typically four-direction or three-direction. That is to say, there are at least two cars from two directions crashing together. In the distribution, the average automobile counts are 0.6-0.2 or 1.8-2.4, which makes sense that there are at least one or two cars in an intersection accident.

(5) From the skewness and distribution, we find that it is unnecessary to transform variables.

3.Unsupervised Learning

3.1 Association Rule Learning

(1) Setting **Municipality Name as ID** and **Crash type as Target**, we got the result as follow:

https://www.autoaccident.com/statistics-on-intersection-accidents.html

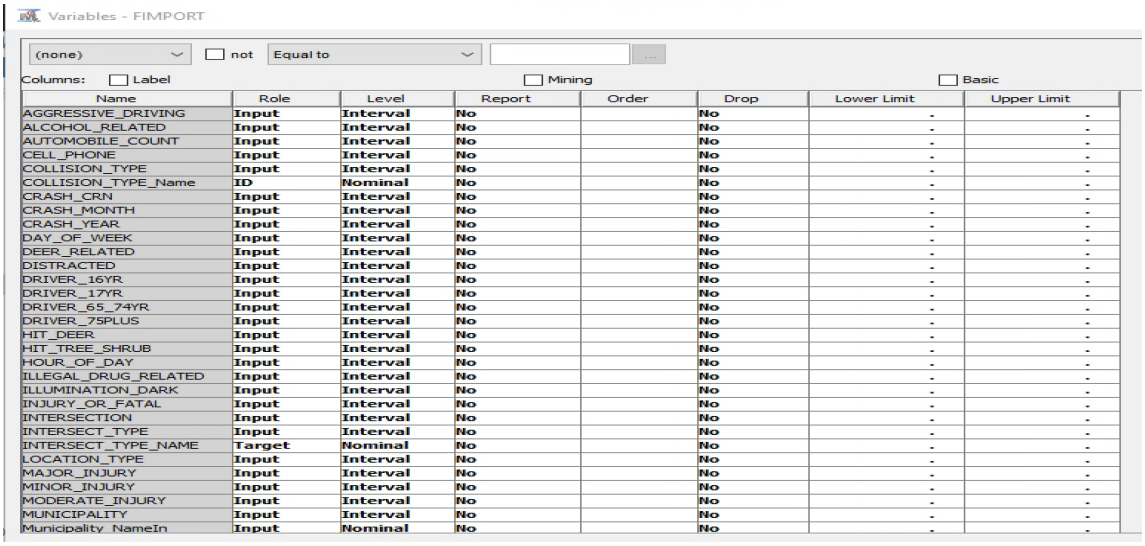
Variables - FIMPORT

(none) not Equal to

Columns: Label Mining Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
AGGRESSIVE_DRIVING	Input	Interval	No		No	.	.
ALCOHOL_RELATED	Input	Interval	No		No	.	.
AUTOMOBILE_COUNT	Input	Interval	No		No	.	.
CELL_PHONE	Input	Interval	No		No	.	.
COLLISION_TYPE	Input	Interval	No		No	.	.
COLLISION_TYPE_Name	Target	Nominal	No		No	.	.
CRASH_CRN	Input	Interval	No		No	.	.
CRASH_MONTH	Input	Interval	No		No	.	.
CRASH_YEAR	Input	Interval	No		No	.	.
DAY_OF_WEEK	Input	Interval	No		No	.	.
DEER_RELATED	Input	Interval	No		No	.	.
DISTRACTED	Input	Interval	No		No	.	.
DRIVER_16YR	Input	Interval	No		No	.	.
DRIVER_17YR	Input	Interval	No		No	.	.
DRIVER_65_74YR	Input	Interval	No		No	.	.
DRIVER_75PLUS	Input	Interval	No		No	.	.
HIT_DEER	Input	Interval	No		No	.	.
HIT_TREE_SHRUB	Input	Interval	No		No	.	.
HOUR_OF_DAY	Input	Interval	No		No	.	.
ILLEGAL_DRUG_RELATED	Input	Interval	No		No	.	.
ILLUMINATION_DARK	Input	Interval	No		No	.	.
INJURY_OR_FATAL	Input	Interval	No		No	.	.
INTERSECTION	Input	Interval	No		No	.	.
INTERSECT_TYPE	Input	Interval	No		No	.	.
INTERSECT_TYPE_NAME	Input	Nominal	No		No	.	.
LOCATION_TYPE	Input	Interval	No		No	.	.
MAJOR_INJURY	Input	Interval	No		No	.	.
MINOR_INJURY	Input	Interval	No		No	.	.
MODERATE_INJURY	Input	Interval	No		No	.	.
MUNICIPALITY	Input	Interval	No		No	.	.
Municipality_NameIn	ID	Nominal	No		No	.	.

(2) Given the Collision type as ID and Intersect type as target value, we got the result as follow:



Variables - FIMPORT

(none) not Equal to

Columns: Label Mining Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
AGGRESSIVE_DRIVING	Input	Interval	No		No	.	.
ALCOHOL_RELATED	Input	Interval	No		No	.	.
AUTOMOBILE_COUNT	Input	Interval	No		No	.	.
CELL_PHONE	Input	Interval	No		No	.	.
COLLISION_TYPE	Input	Interval	No		No	.	.
COLLISION_TYPE_Name	ID	Nominal	No		No	.	.
CRASH_CRN	Input	Interval	No		No	.	.
CRASH_MONTH	Input	Interval	No		No	.	.
CRASH_YEAR	Input	Interval	No		No	.	.
DAY_OF_WEEK	Input	Interval	No		No	.	.
DEER_RELATED	Input	Interval	No		No	.	.
DISTRACTED	Input	Interval	No		No	.	.
DRIVER_16YR	Input	Interval	No		No	.	.
DRIVER_17YR	Input	Interval	No		No	.	.
DRIVER_65_74YR	Input	Interval	No		No	.	.
DRIVER_75PLUS	Input	Interval	No		No	.	.
HIT_DEER	Input	Interval	No		No	.	.
HIT_TREE_SHRUB	Input	Interval	No		No	.	.
HOUR_OF_DAY	Input	Interval	No		No	.	.
ILLEGAL_DRUG_RELATED	Input	Interval	No		No	.	.
ILLUMINATION_DARK	Input	Interval	No		No	.	.
INJURY_OR_FATAL	Input	Interval	No		No	.	.
INTERSECTION	Input	Interval	No		No	.	.
INTERSECT_TYPE	Input	Interval	No		No	.	.
INTERSECT_TYPE_NAME	Target	Nominal	No		No	.	.
LOCATION_TYPE	Input	Interval	No		No	.	.
MAJOR_INJURY	Input	Interval	No		No	.	.
MINOR_INJURY	Input	Interval	No		No	.	.
MODERATE_INJURY	Input	Interval	No		No	.	.
MUNICIPALITY	Input	Interval	No		No	.	.
Municipality_NameIn	Input	Nominal	No		No	.	.

Output

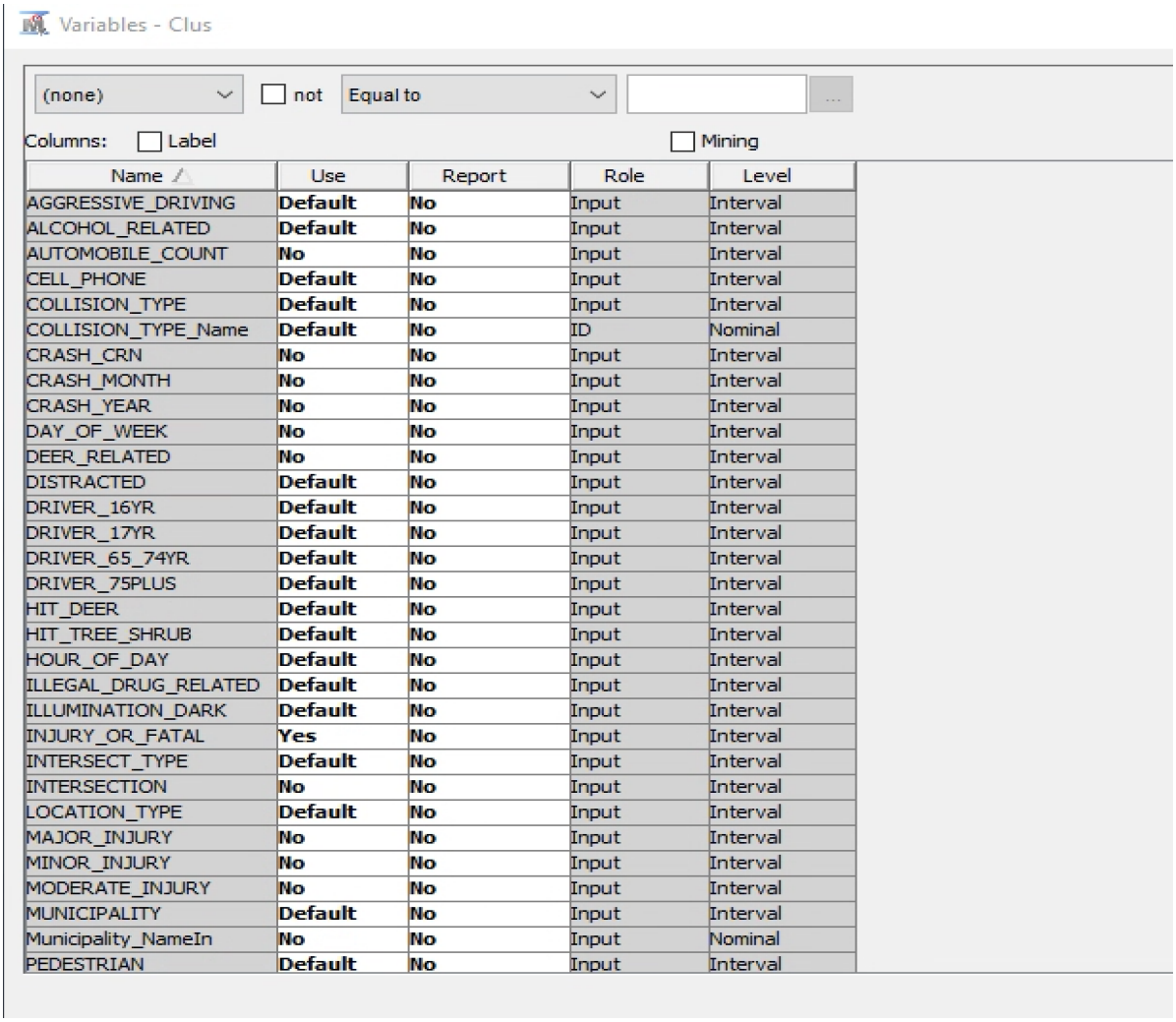
	Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule
25							
26							
27							
28							
29	2	60.00	75.00	60.00	1.25	6.00	Railroad crossing ==> Crossover
30	2	80.00	100.00	60.00	1.25	6.00	Crossover ==> Railroad crossing
31	2	100.00	100.00	100.00	1.00	10.00	Traffic circle or Round About ==> "T" intersection
32	2	100.00	100.00	100.00	1.00	10.00	On ramp ==> "T" intersection
33	2	100.00	100.00	100.00	1.00	10.00	Off ramp ==> "T" intersection
34	2	100.00	100.00	100.00	1.00	10.00	Multi-leg intersection ==> "T" intersection
35	2	100.00	100.00	100.00	1.00	10.00	Four way intersectionMid-block ==> "T" intersection
36	2	100.00	100.00	100.00	1.00	10.00	Four way intersection ==> "T" intersection
37	2	100.00	100.00	100.00	1.00	10.00	"Y" intersection ==> "T" intersection
38	2	100.00	100.00	80.00	1.00	8.00	Railroad crossing ==> "T" intersection
39	2	100.00	100.00	60.00	1.00	6.00	Crossover ==> "T" intersection
40	2	100.00	100.00	100.00	1.00	10.00	Traffic circle or Round About ==> "Y" intersection
41	2	100.00	100.00	100.00	1.00	10.00	On ramp ==> "Y" intersection
42	2	100.00	100.00	100.00	1.00	10.00	Off ramp ==> "Y" intersection
43	2	100.00	100.00	100.00	1.00	10.00	Multi-leg intersection ==> "Y" intersection
44	2	100.00	100.00	100.00	1.00	10.00	Four way intersectionMid-block ==> "Y" intersection
45	2	100.00	100.00	100.00	1.00	10.00	Four way intersection ==> "Y" intersection
46	2	100.00	100.00	100.00	1.00	10.00	"T" intersection ==> "Y" intersection
47	2	100.00	100.00	80.00	1.00	8.00	Railroad crossing ==> "Y" intersection
48	2	100.00	100.00	60.00	1.00	6.00	Crossover ==> "Y" intersection
49	2	60.00	60.00	60.00	1.00	6.00	Traffic circle or Round About ==> Crossover
50	2	60.00	60.00	60.00	1.00	6.00	On ramp ==> Crossover
51	2	60.00	60.00	60.00	1.00	6.00	Off ramp ==> Crossover
52	2	60.00	60.00	60.00	1.00	6.00	Multi-leg intersection ==> Crossover
53	2	60.00	60.00	60.00	1.00	6.00	Four way intersectionMid-block ==> Crossover
54							

Based on the association rule learning, we have identified the following top rules:
Top Rules:
Railroad Crossing => Crossover (Lift: 1.25, Support: 60%, Confidence: 75%)
Traffic Circle/Roundabout => ‘T’ Intersection (Lift: 1.00, Support: 100%, Confidence: 100%)
Off Ramp => ‘T’ Intersection (Lift: 1.00, Support: 100%, Confidence: 100%)
4-way Intersection => ‘T’ Intersection (Lift: 1.00, Support: 100%, Confidence: 100%)
Multi-leg Intersection => ‘T’ Intersection (Lift: 1.00, Support: 100%, Confidence: 100%)

3.2 Clustering Analysis

We want to understand the factors lead to fatal or injury intersection accidents. To do so, we categorize factors to subjective and objective factors. For example, putting the weather condition, road type, illumination condition into the objective category (environment condition) and putting whether the driver has taken alcohol and drug, is the driver licensed and is this person an underage (<17) or old driver (>75) into the subjective category

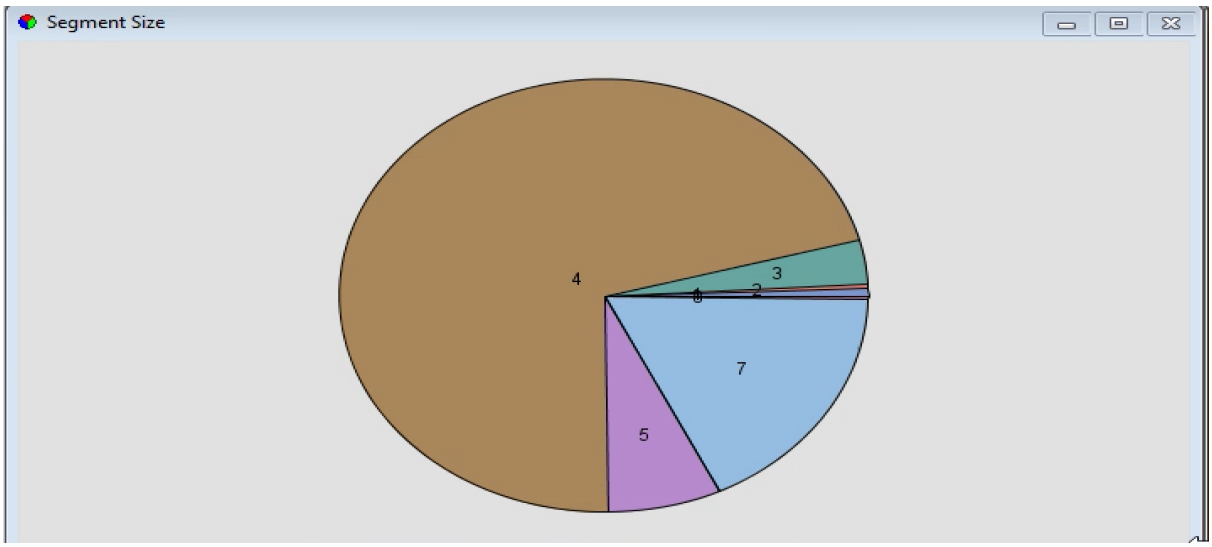
To understand how objective and subjective factors lead to the happening of intersection accidents (either level of injury or fatality), we set all objective and subjective factors to “Default”, make all irrelevant columns such as “CRASH_CRN” (which is the identifier of the crash incident) as “Rejected”, and make “INJURY_OR_FATAL” as yes as it is a binary factor in deciding whether there is an injury/ fatal accident.


Variables - Clus

(none) not Equal to ...

Columns: Label Mining

Name	Use	Report	Role	Level
AGGRESSIVE_DRIVING	Default	No	Input	Interval
ALCOHOL_RELATED	Default	No	Input	Interval
AUTOMOBILE_COUNT	No	No	Input	Interval
CELL_PHONE	Default	No	Input	Interval
COLLISION_TYPE	Default	No	Input	Interval
COLLISION_TYPE_Name	Default	No	ID	Nominal
CRASH_CRN	No	No	Input	Interval
CRASH_MONTH	No	No	Input	Interval
CRASH_YEAR	No	No	Input	Interval
DAY_OF_WEEK	No	No	Input	Interval
DEER_RELATED	No	No	Input	Interval
DISTRACTED	Default	No	Input	Interval
DRIVER_16YR	Default	No	Input	Interval
DRIVER_17YR	Default	No	Input	Interval
DRIVER_65_74YR	Default	No	Input	Interval
DRIVER_75PLUS	Default	No	Input	Interval
HIT_DEER	Default	No	Input	Interval
HIT_TREE_SHRUB	Default	No	Input	Interval
HOUR_OF_DAY	Default	No	Input	Interval
ILLEGAL_DRUG_RELATED	Default	No	Input	Interval
ILLUMINATION_DARK	Default	No	Input	Interval
INJURY_OR_FATAL	Yes	No	Input	Interval
INTERSECT_TYPE	Default	No	Input	Interval
INTERSECTION	No	No	Input	Interval
LOCATION_TYPE	Default	No	Input	Interval
MAJOR_INJURY	No	No	Input	Interval
MINOR_INJURY	No	No	Input	Interval
MODERATE_INJURY	No	No	Input	Interval
MUNICIPALITY	Default	No	Input	Interval
Municipality_NameIn	No	No	Input	Nominal
PEDESTRIAN	Default	No	Input	Interval



Mean Statistics

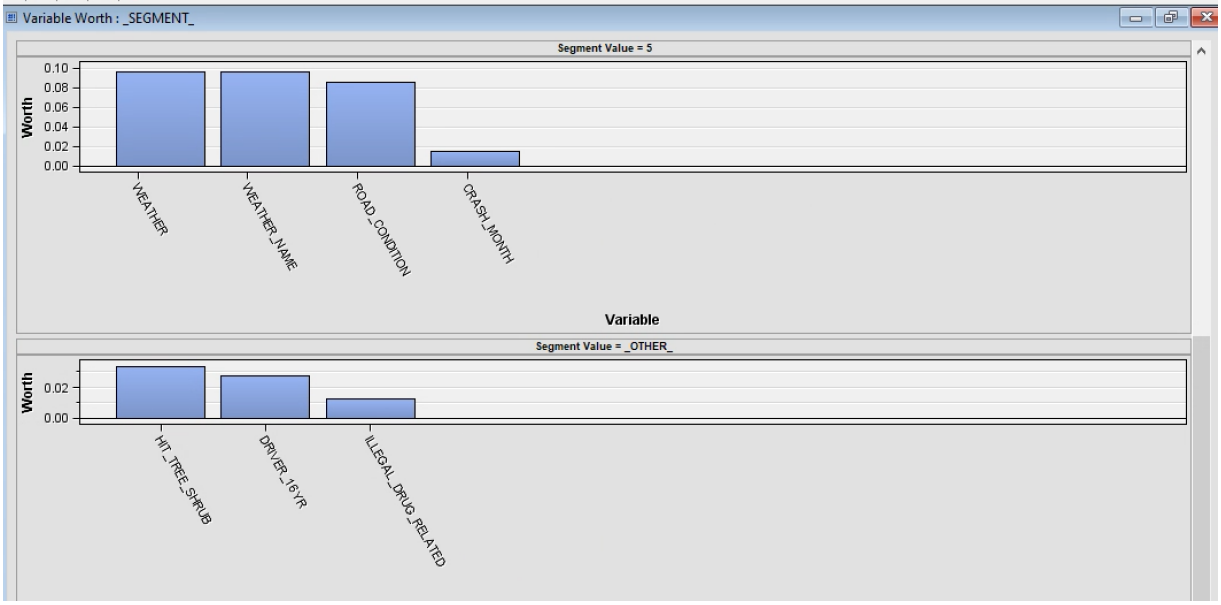
Clustering Criterion	Maximum Relative Change in Cluster Seeds	I m	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	AGGRESSIVE_DRIVING
0.854365	0.003991	.	1	486	0.923438	10.88913	4	12.43175	0.460905
0.854365	0.003991	.	2	172	0.760147	9.367314	4	21.02879	0.046512
0.854365	0.003991	.	3	2390	1.474406	13.09074	4	5.709828	0.570293
0.854365	0.003991	.	4	52762	0.742065	10.40656	7	2.694267	0.696297
0.854365	0.003991	.	5	5242	0.972819	11.98401	4	4.315335	0.498855
0.854365	0.003991	.	6	2	1.955651	6.631937	8	14.29209	1
0.854365	0.003991	.	7	12966	1.047022	13.73429	4	2.694267	0.208777
0.854365	0.003991	.	8	253	1.048454	13.30026	6	14.29209	0.498024

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
ROAD_CONDITION		1	1	1
WEATHER		1	0	0.926451
HIT_TREE_SHRUB		2	0	0.534782
ILLEGAL_DRUG_RELATED		1	0	0.322021
LOCATION_TYPE		1	0	0.23313
HIT_DEER		1	0	0.191649
DISTRACTED		1	0	0.116456
CELL_PHONE		0	1	0.114005

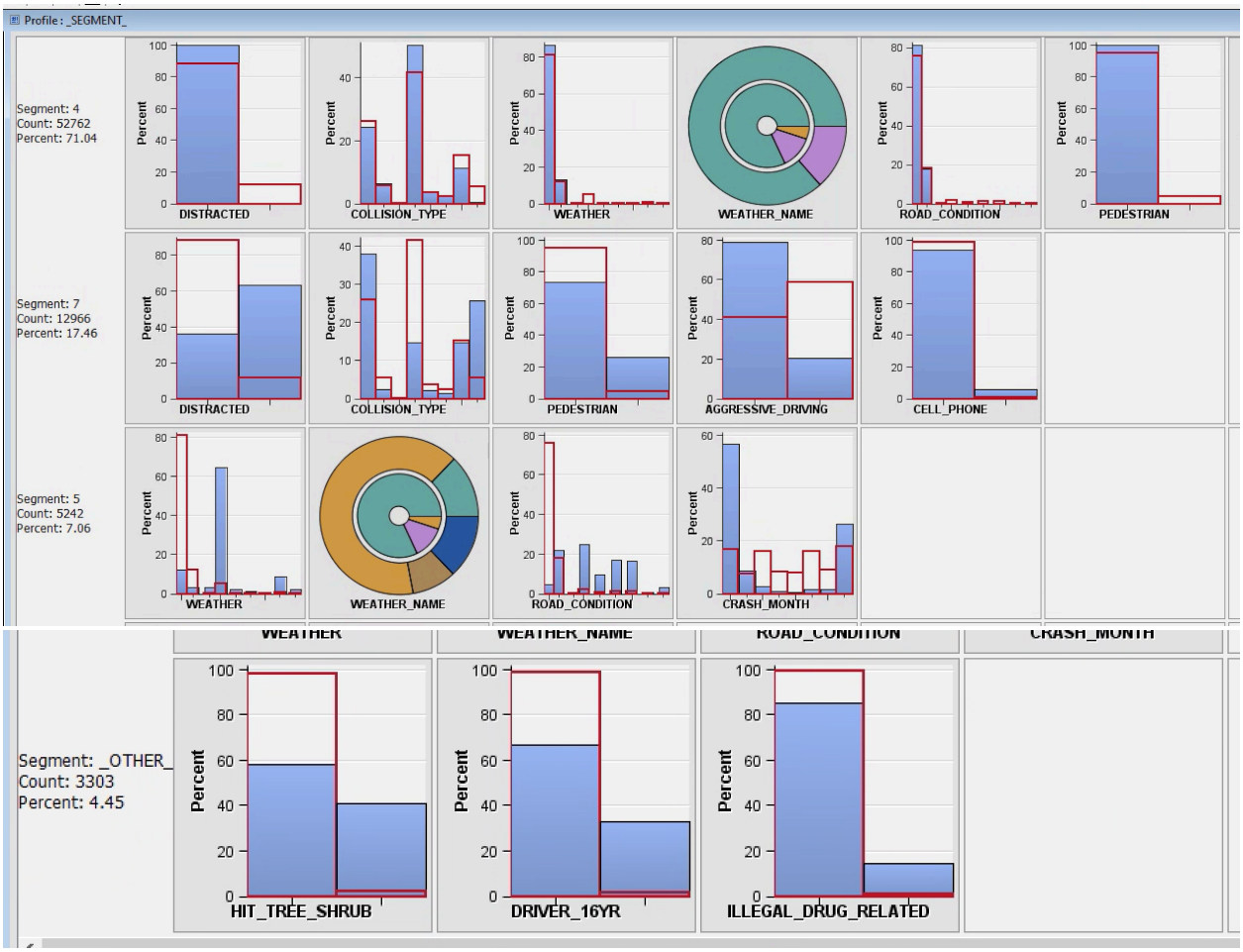
Based on the distribution of the bar chart, we found four primary clusters, clusters 4, 7, 5 and 3. Root mean square standard deviation of the clusters are ranging from 0.742 to 1.956. Regarding the important variables, we found that the **road condition, weather and whether it was hit by tree shrub, whether driver is related to illegal drug, whether the car was hit by deer** were the most important variables.

We used the “Segment Profile” node to evaluate in detail the traits of each cluster.



According to the screenshot above, we can see that for each cluster, the important variables varies. For the largest Cluster #4 and #7, they share the traits of “DISTRACTED” and “COLLISION TYPE” as the most important determining variables.

Across different clusters, there are some differences in priority of the importance of the variables and the weightage of each variable.



Examining the detailed distribution of attributes within the clusters, it can be analyzed that:

1.Cluster #4:
Cluster 4 has a low percentage of “DISTRACTION”, meaning that drivers of the incidents involved within this group were less likely to got distracted. And the most dominant “COLLISION TYPE” is angle (value of 4) collision. It showed that most of the accidents occurred on a angle with no impairments to the driver or signs of inclement weather. It was not initially intuitive that these results would be the most frequent. However, applying the cluster to everyday life, the cluster makes sense based on the fact inclement weather or drunk driving are not everyday occurrences unlike endless drivers making turns throughout the day.

2.Cluster #7:
Cluster 7 has a high percentage of “DISTRACTION”, 26.33% of pedestrian involved in the incidents and 20.87% has an indication of the aggressive driving.
It suggests drivers who are distracted are more likely to be in an accident which results in injury or fatality. The distractions in segment 7 involved a non collision (overturned vehicle, jackknifed trailer, or a vehicle going off the roadway) accident most often, and it does make sense a driver who is distracted will be more likely to either not turn properly or go off the road.

3.Cluster #5:
Cluster 5 has a high occurrence of snow as the “WEATHER TYPE”, angle as the most dominant “COLLISION TYPE”. The months of December-January. And the weather is snowy and the roads have slush or being wet. It was the most intuitive as the data shows that the months of December-January and when driving conditions are not optimal, are very frequent occurrence in the data set.

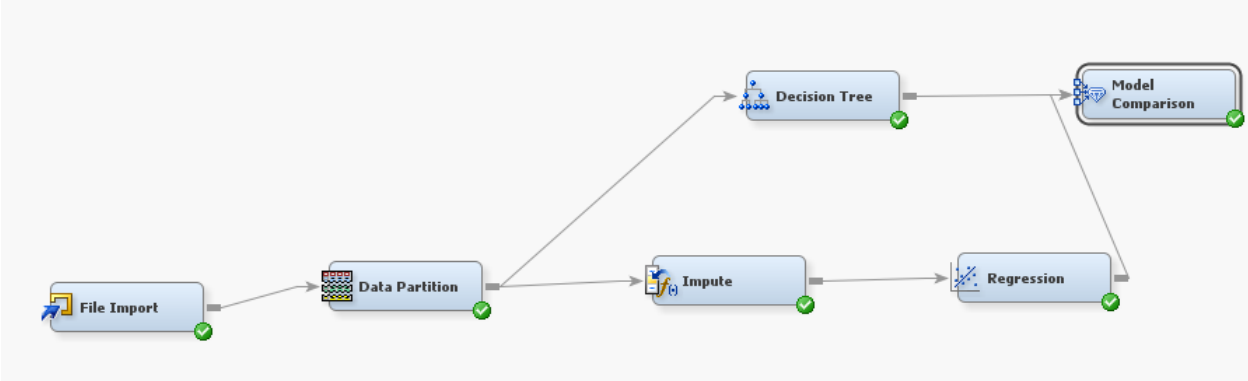
4. Supervised Learning

Based on the results of our unsupervised learning models, we decided to run 2 supervised learning models : decision tree and logistic regression to see which of the following clustered factors had the most significance in the occurrence of accidents at intersections.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
ROAD_CONDITION		1	1	1
WEATHER		1	0	0.926451
HIT_TREE_SHRUB		2	0	0.534782
ILLEGAL_DRUG_RELATED		1	0	0.322021
LOCATION_TYPE		1	0	0.23313
HIT_DEER		1	0	0.191649
DISTRACTED		1	0	0.116456
CELL_PHONE		0	1	0.114005

Our Diagram for supervised learning models in SAS is as shown below.



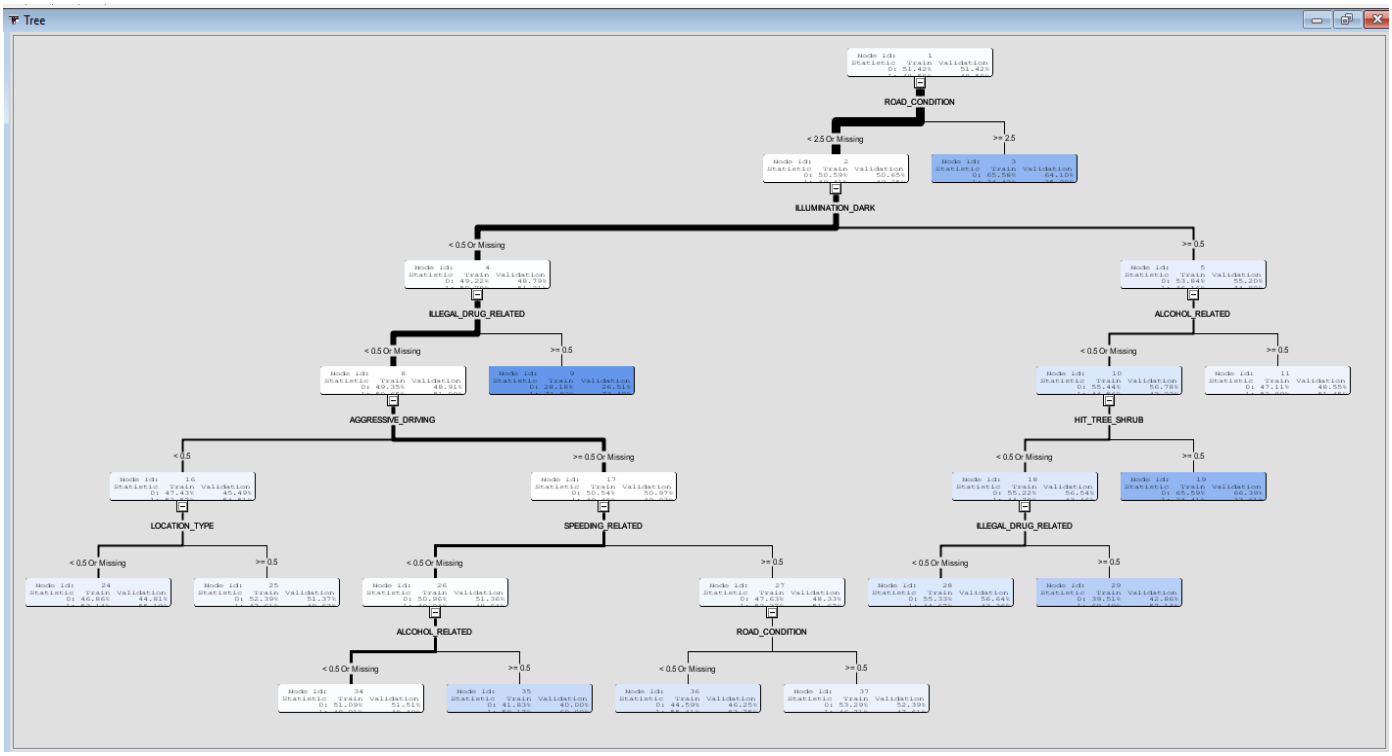
4.1 Decision tree Modelling

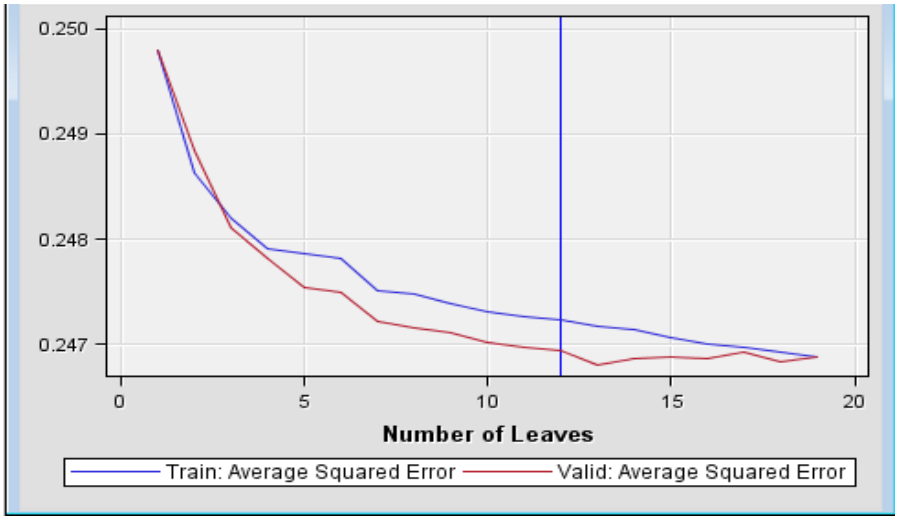
(1) We rejected the unnecessary variables and chose only the variables that were important for our analysis. The following figure shows the feature selection in SAS.

Name	Role	Level	Report
AGGRESSIVE_DR	Input	Interval	No
ALCOHOL_RELA	Input	Interval	No
AUTOMOBILE_C	Rejected	Interval	No
CELL_PHONE	Input	Interval	No
COLLISION_TYP	Rejected	Interval	No
COLLISION_TYP	Rejected	Nominal	No
CRASH_CRN	Rejected	Interval	No
CRASH_MONTH	Rejected	Interval	No
CRASH_YEAR	Rejected	Interval	No
DAY_OF_WEEK	Rejected	Interval	No
DEER_RELATED	Rejected	Interval	No
DISTRACTED	Input	Interval	No
DRIVER_16YR	Rejected	Interval	No
DRIVER_17YR	Rejected	Interval	No
DRIVER_65_74Y	Rejected	Interval	No
DRIVER_75PLUS	Rejected	Interval	No
HIT_DEER	Rejected	Interval	No
HIT_TREE_SHRU	Input	Interval	No
HOUR_OF_DAY	Rejected	Interval	No
ILLEGAL_DRUG_	Input	Interval	No
ILLUMINATION_	Input	Interval	No
INJURY_OR_FAT	Target	Binary	No
INTERSECTION	Rejected	Interval	No
INTERSECT_TYP	Rejected	Interval	No
INTERSECT_TYP	Rejected	Nominal	No
LOCATION_TYPE	Input	Interval	No
MAJOR_INJURY	Rejected	Interval	No
MINOR_INJURY	Rejected	Interval	No
MODERATE_INJ	Rejected	Interval	No
MUNICIPALITY	Rejected	Interval	No
Municipality_Nan	Rejected	Nominal	No
PEDESTRIAN	Rejected	Interval	No
ROAD_CONDITI	Input	Interval	No
SPEEDING_RELA	Input	Interval	No
URBAN_RURAL	Rejected	Interval	No
VAR1	Rejected	Interval	No
WEATHER	Rejected	Interval	No
WEATHER_NAM	Rejected	Nominal	No

(2) We split the data into 70% training set and 30% validation set. We then ran a decision tree in SAS.

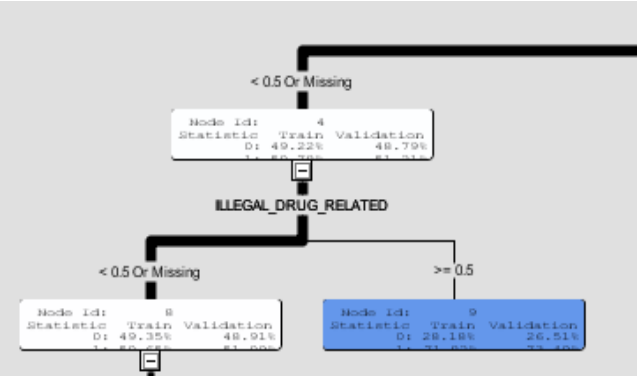
(3) After pruning the tree at the lowest validation error, we got an optimal decision tree with 12 leaf nodes. The images of the decision tree and the pruning of the tree are shown below.



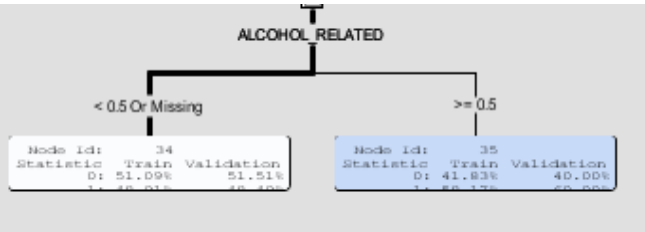


(4) Some of our important findings and takeaways from the decision tree are :

- If a person is driving on a bad and dark road, under the influence of drugs, it is highly probable that an accident would occur.



- It also indicates that driving under the influence of alcohol may impair the control over the vehicle and the car may hit a tree/bush and cause and accident.



(5) The Accuracy for the above model was approximately 76%. Although it is not a very great accuracy, we get significant results based on out training validation split and there is no issue of overfitting.

(6) The results from SAS showed that the most significant factors were : ILLEGAL_DRUG_RELATED & ALCOHOL_RELATED.

4.2 Logistic Regression

In order to solidify our results from the decision tree, we also ran a logistic regression to verify the previous results.

(1) We first added an impute node in order to impute any missing values in the dataset.

(2) Keeping the data partition same, (Training- 70%, Validation- 30%), we ran a stepwise logistic regression model and got the following results.

(3) The most important factors in this model also turned out to be ILLEGAL_DRUG_RELATED & ALCOHOL_RELATED as can be seen in the figure below.

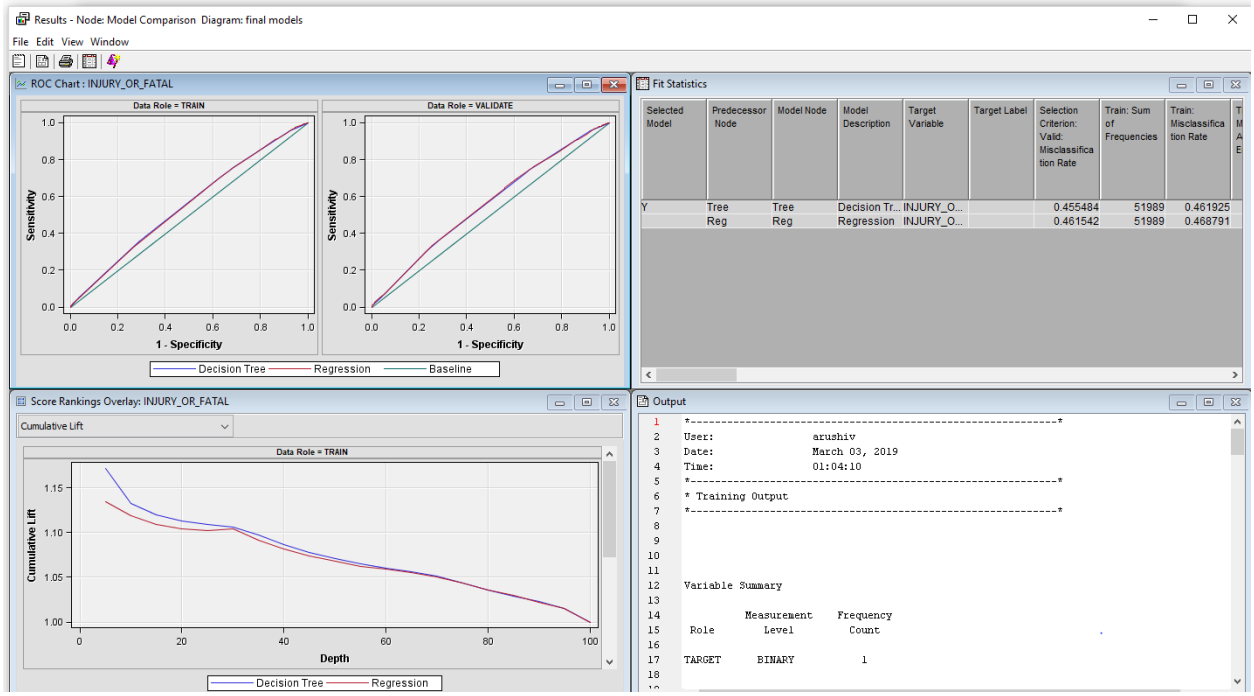
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	0.0710	0.0158	20.15	<.0001		1.074
AGGRESSIVE_DRIVING	1	-0.0616	0.0189	10.60	0.0011	-0.0167	0.940
ALCOHOL_RELATED	1	0.3253	0.0363	80.24	<.0001	0.0461	1.384
HIT_TREE_SHRUB	1	-0.1919	0.0667	8.28	0.0040	-0.0142	0.825
ILLEGAL_DRUG_RELATED	1	0.7631	0.1160	43.24	<.0001	0.0342	2.145
ILLUMINATION_DARK	1	-0.2475	0.0203	149.04	<.0001	-0.0627	0.781
ROAD_CONDITION	1	-0.1112	0.00840	175.23	<.0001	-0.0686	0.895
SPEEDING_RELATED	1	0.0603	0.0307	3.84	0.0500	0.0103	1.062

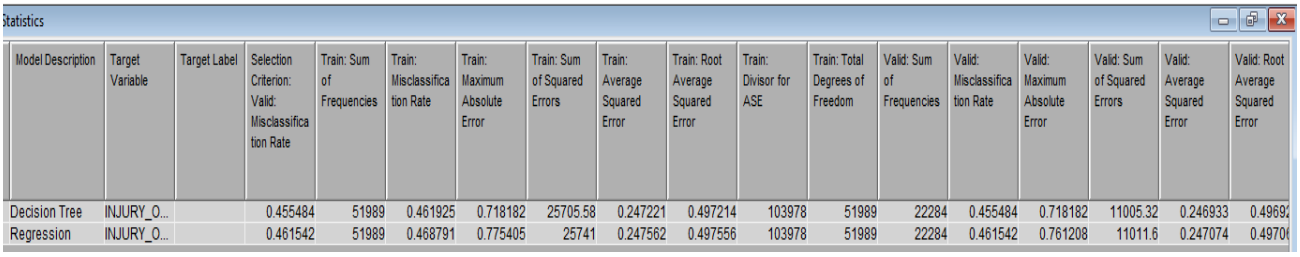
(4) The accuracy of this model also turned out to be approximately the same i.e. 76%.

4.3 Model Comparison:

(1) In order to compare the results of both the supervised learning models that we worked on, we added a “Model Comparison” node in our SAS diagram.
(2) On running the model comparison node, we achieved the following results.



(3) The Training and Validation Average Square Errors for both the models are also almost the same. See figure below.



5. Conclusion and business implications

We found through our data exploration key factors that correlate with auto accidents and with regression we were able to look at the probability of these factors being involved with an accident.

Clustering:

- Cluster 4 was the most frequent with accidents happening on an angle in dry conditions
- Cluster 7 involved a non collision (overturned vehicle, jackknifed trailer, or a vehicle going off the roadway) accident most often, and it does make sense a driver who is distracted will be more likely to either not turn properly or go off the road.
- Cluster 5 the data shows that the months of December-January and when driving conditions are not optimal, have very frequent occurrences in the data set.

Decision Tree:

- Bad conditions and darkness, while under the influence of drugs has the highest probability
- Driving under the influence of alcohol increases the likelihood a car may hit a tree/bush
- Most significant factors were : ILLEGAL_DRUG_RELATED & ALCOHOL_RELATED.

Stepwise Logistics Regression:

- The most significant factors contributing to the occurrence of fatal and injury incidents are ILLEGAL_DRUG_RELATED, ALCOHOL_RELATED, ILLUMINATION_DARK and ROAD_CONDITION

Business Implications and future improvements:

- Knowing the factors highly correlated to the occurrence of fatal/injury accidents happening in the intersection we can make policies that would not be detrimental to the agency financially.
- If the agency knows the area it servers or a specific a driver has a history with illegal substance use then perhaps offering a policy that with a higher premium or deductible is necessary to cover the costs, in the increased likelihood of an accident.
- Our data exploration did have its limitations and in the future improvements can be make more accurate predictions. More information is needed on drivers everywhere in the county to create the most efficient policy. Our data only focused on the intersections in Allegheny County rather than the county as a whole. Additionally, examining the other factors that are correlated to accidents like the age of the driver provide more detail in the data.

SAS and Python source code

https://drive.google.com/open?id=1a2cyglffdUp64W7Y6MfPC5Ac-s7yJDoq

Reference

[1] Statistics on Intersection Accidents (Updated for 2018). (2019). Retrieved from https://www.autoaccident.com/statistics-on-intersection-accidents.html