Seattle Traffic Incidents: From Insurance Perspective.

Figure 1: Seattle skyline, picture copied from freeimages.com (website claims to be no copyright)

Introduction:

The City of Seattle

Seattle is the largest city in the Washington State with a population of around 610,000. It has an area of 4,903,675 sq. Miles, making it the 14 largest cities in the US. Recently (2016). The total number of cars hit a record high of 444,000. With the increase in number of cars, insurance demands are rising. It is necessary for insurance companies to identify customers with low and high risks. A quote based on this information may make potential customer to purchase policy on one hand and reduce insurance company's liability on the other hand.

The Insurance:

An Insurance is necessary for every motor vehicle in the state of Washington. This means that whenever a person in Washington buys a car, he must purchase an insurance with the car. There are 2 types of insurances that the insurance companies provide,

i. Full coverage: Insures everyone and everything involved in an accident

ii. Liability: Insures the person who if not at fault if the holder of the insurance at fault. It is also known as third party insurance.

Insurers vs. Insured:

While the insurance is mandatory by law, often conflict between insurers (insurance company) and Insured (insurance bearer) arise due to non-payment or underpayment of benefit by the company. The company evaluates each case independently and assigns benefits accordingly. Whereas the bearers believe that they are under-paid or not paid the benefits. It would be false statement that there is a general dislike about the insurance companies.

Change:

A change is what this report is all about. A change done by the insurance companies to provide better monthly payment rates to its customers. And another change, that we the people as a society can do to reduce the incidents and improve the road conditions. If the insurance companies can effectively improve the identification of Risks and provide the better rate to customers, customers win. If the society (made of these customers) reduce the incidents the insurance companies win.

Although an ideal, this report is intended to create a win-win situation for Insurance companies and its customers. The report

In near future, we might have robotically driven cars (self-drive). Imagine the risk an insurance companies will have to take. If there is a glitch in software, we will have major incidents all over the world. And if the software (and all other systems) is perfect then we might not need insurance at all. What future holds, only time can tell. Or may be a data scientist can….

Data:

Source:

For this analysis, data is available online on different platform. The data used here is from https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions website. The data is 221738 x 40, that is 221738 rows and 40 columns. These 40 columns give information about Location (Longitudes and Latitudes), Driver's faults (Inattention, speeding under-influence), How many Pedestrians, Cyclist where hit, severity of accident, number of vehicles involved, driving conditions (weather, road condition, light condition), Address type etc.

A full detail of information is given on the website. The page also gives information about the data filled in each case.

Application of Data:

The report is based on analysis of 4 important factors.

1. Drivers Fault
2. Accident Time

3. Distance from city center
4. Analysis of Parked Car in Incident

The first factor, driver's fault, takes into consideration if the driver was under-influence, speeding or inattentive (distracted). This is not the put the blame on the driver (or Insurance bearer) but to make the readers aware of the consequences of the actions done by them. This is the part where we take into consideration of our social responsibility.

The second factor is Accident Time. The report analysis the time of the day when the accidents happened. On one hand, it will help the insurance company identify the risk based on the time at which the customer will travel within the city of Seattle, on the other hand, it gives an insight of peak hours of incident, can these times be avoided? Can we change work hour to be safe? Can we not be just a number?

The third and fourth part (Hit Parked Vehicle and Distance from the city center) is mostly intended to be used by insurance companies to provide a better insurance quote to the customer.

Data Set:

The data used in this report (analysis) is derived from the incidents reported on the website above. The data is provided in comma separated values (csv) format. This data is stored in a dataframe simply named df.

From the df dataframe, the following columns were taken, rest all were dropped.

- X - Longitude
- Y - Latitude
- SPEEDING – whether the driver was speeding
- INATTENTIONIND - whether the driver was distracted
- UNDERINFL - whether the driver was under influence
- VEHCOUNT - number of vehicles hit in the accident
- INJURIES - number of (minor) injuries in the accident
- SERIOUSINJURIES - number of (major) injuries in the accident
- FATALITIES - - number of fatalities in the accident
- INCDTTM – Date and Time of the incident
- WEATHER – Weather condition at the time of incident
- LIGHTCOND – light condition at the time of incident
- ROADCOND - Road condition at the time of incident
- HITPARKEDCAR – weather a parked car was hit in the incident.

Data Conditioning:

For SPEEDING, INATTENTIONIND and UNDERINFL the data is collected only if it was true ('Y' only of driver is speeding in SPEEDING column and so on). It would be fair to assume that for all the other rows (except where value is inserted) it can be not true. And hence, all the NAN values are replaced by 'N' in these columns. After this, all the remaining rows with NAN are dropped using the dropna command.

For the time of the day and distance data, it was necessary to scrap data from the dataframe df and create additional rows. 3 columns that were added are DIST for distance, HOUR for the time of the day and DAYIME for time complete time of the day.

Data Pre-Processing:

While some data is given, some of the data must be derived. In the driver's fault dataframe an extra column is created to find if the driver has any of the speeding, inattention, or influence, if any one of them is true then the driver's fault column shows true. For the time if the accident- the time given in the 'Collisions' csv file (column name: INCDTTM) in the format Date – Time (MM/DD/YYYY HH: MM: SS AM/PM). This was cleaned and another column with time is created.

Also, the 'X' and 'Y', that is, longitudes and latitudes are rounded to 2 digit as we will use this data in mapping the high frequency incidents on the Seattle map.

Methodology:

1. Driver's Fault Analysis

   For this analysis, a new dataframe was created and named as df_drvFlt and counts (or frequencies) was taken using groupby function. To identify if there is any drivers fault like Influence, Distraction or Speeding a new column was created. This column is named as FAULT and using 'or' condition (|) in python found if there was any of the driver's fault in the incident. (True if any one, any two or all three are True). This was then sorted by the frequency of the incidents.

   For analysis, and SNS plot with hue showing Fault values (True or False) is done with variables as Vehicle count, Injuries, Serious Injuries, Fatalities.

   Although a KNN, Logistic Regression or Decision Tree Analysis could have been done, there is no point in it. This analysis is to show the general audience the consequences of impaired, distracted driving or speeding.

   The result is discussed in the Result section of this report

2. Time of Accident Analysis

   For time of accident analysis, Polynomial regression is used. Time is rounded off, a value count for number of incidents happening in an hour are taken. So, for example, if the time is 12, it means that number of accidents happening from 12 pm to 1 am.

   Time vs. Frequency plot was created, and it showed a trend. Further analysis, using Linear Regression confirms that there is a trend in incidents with respect to time.

Also, going further, a 3D chart with Hour, Distance from City Center and Frequency was created and showed a possibility of trend. However, this will be a Multiple Polynomial Regression, this will make the analysis complex and lead to possible confusion.

Which leads us to finding the trend in the number of incidents with respect to distance from the center of the city.

3. Distance from City Center
   Like the time analysis, distance (from city center) vs. Frequency of accident was plotted and showed some trend. For this a value count for distance was taken. A polynomial regression confirmed this, however for an accurate result (one that is acceptable) the degree of polynomials was 6, making this a complex analysis. This was simplified by creating a dataframe df_hotSpot and using groupby function for longitudes and latitudes, incidents were sorted by their geographic location. Using a simple folium map, and creating bin for accidents, the frequency of incidents was marked with different colors on the map of Seattle. Green color was used for frequency between 1 to 1500, blue for 1500 to 3000, yellow for 3000 to 4500 and red for 4500 to 6000 incidents.

4. Analysis of Parked Car in Incidents

   A new dataframe was create to find the frequency (number of) incidents based on Weather condition, Road Condition, Light Condition, Hit Parked Vehicle and Address type data from the original dataframe df. Initially the results were inconclusively (some work done using value count). However, using a decision tree analysis the incidents showed some trend and were used in analysis. A decision tree with maximum depth of 4 was used.

   A simple value count for Address Type was done.

All the results of all the four analysis are shown in the Result section below and a conclusion from the results are discussed in the sections.

Results and Discussion:

The objective of the first part of the analysis was to show the consequences of impaired driving, speeding and distracted driving. The results are showing in the SNS plot below,

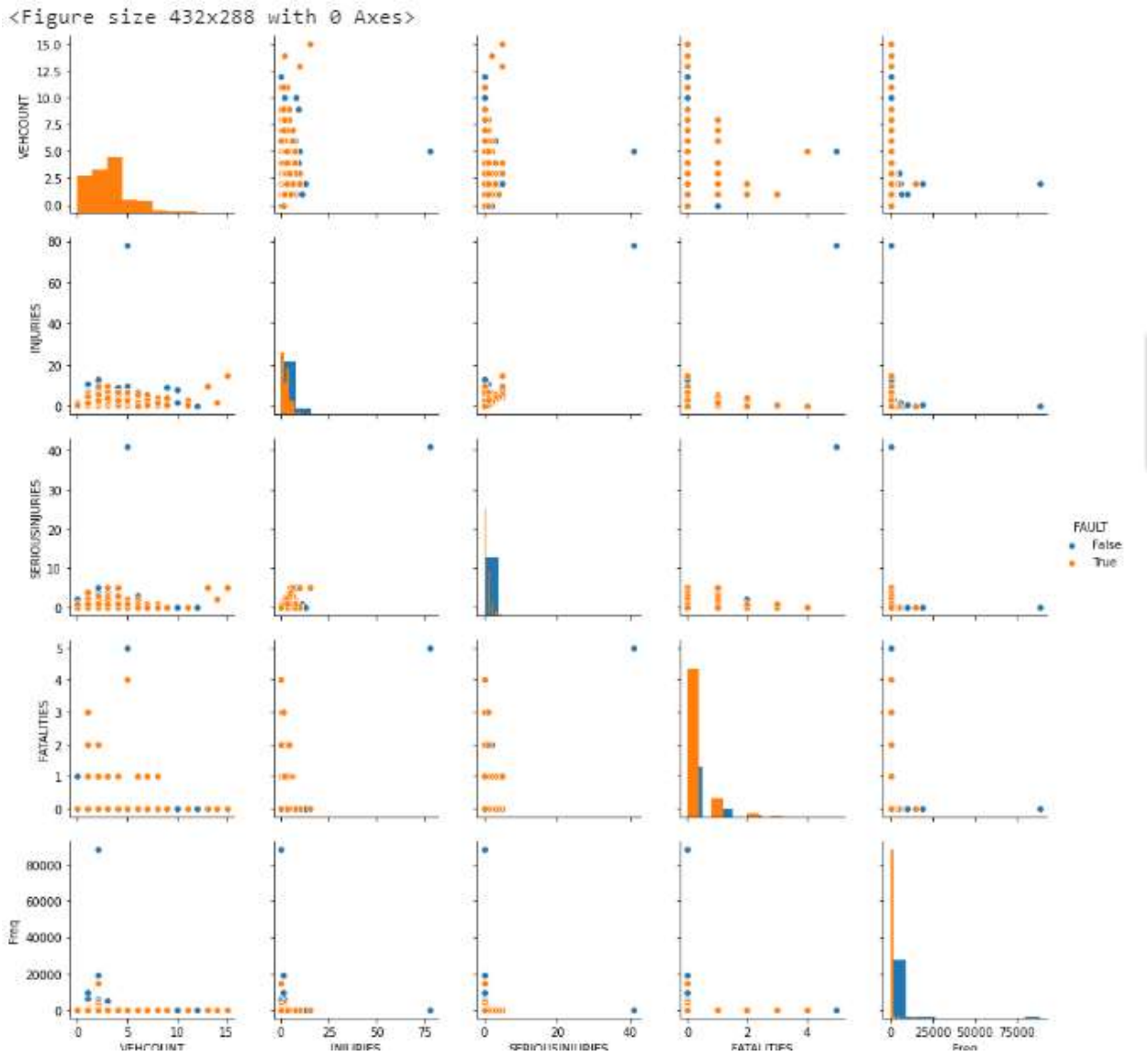

Figure 2: SNS plot showing impart of impair driving with injuries, fatalities and Vehicles Involved.

The above plot show that higher vehicle counts with respect to fatality, serious injury and number of injuries were done with drivers' fault. Something like decision tree or KNN would have given us better result in identifying if the driver were at fault based on the factors like fatality numbers, injury numbers or vehicles involved. However, to identify a mistake after committing it is like starting to dig a well after dying with thirst.

This analysis is to make the general audience aware of the consequences of impaired, distracted driving or speeding. The results give a message that is loud and clear.

Don't Drink and Drive

and

For the time of the accident, a polynomial regression was done to find the trend with number of accidents vs. Time of the day (24 hour clock). The results show that number of accidents increase during the office hours (starting 8 am) reach peak at 12:00 pm and reduce as 5:00 pm approaches. The R2 score of the analysis is 0.8, which can be considered good. From insurance perspective, people travelling in vehicles during lunch hours have more chances to get in an accident than regular commuters. The degree of polynomial was 4. The equation can be written as

$$Yt = Xt^4 - 2574.8 * Xt^3 + 744.598 * Xt^2 - 51.67 * Xt + 1.0556$$

Where,

Xt = Hour of the day.

Yt = Frequency of Accidents.

With intercept of 3843.84.

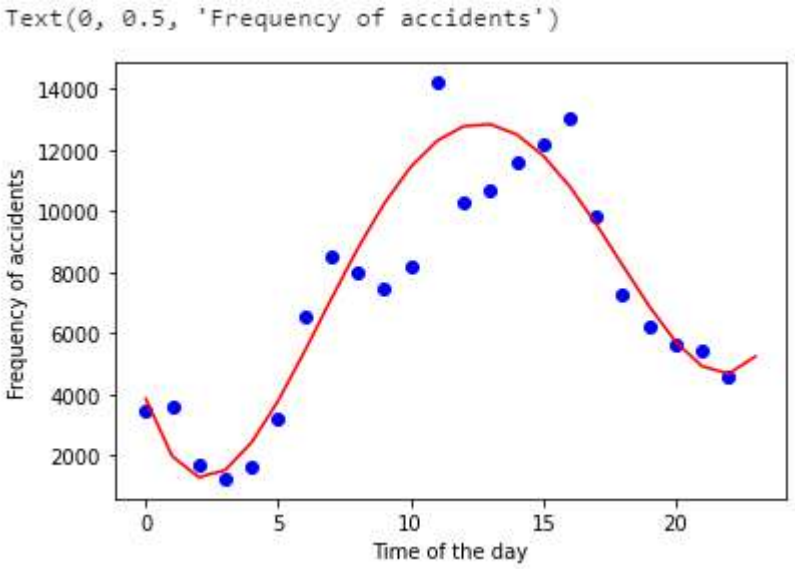

Figure 3: Scatter Plot and Trendline for Hour of Incident and Frequency of Incidents

When a 3D plot of Distance vs. Time vs. Frequency of the accident was plotted it showed a trend like this.

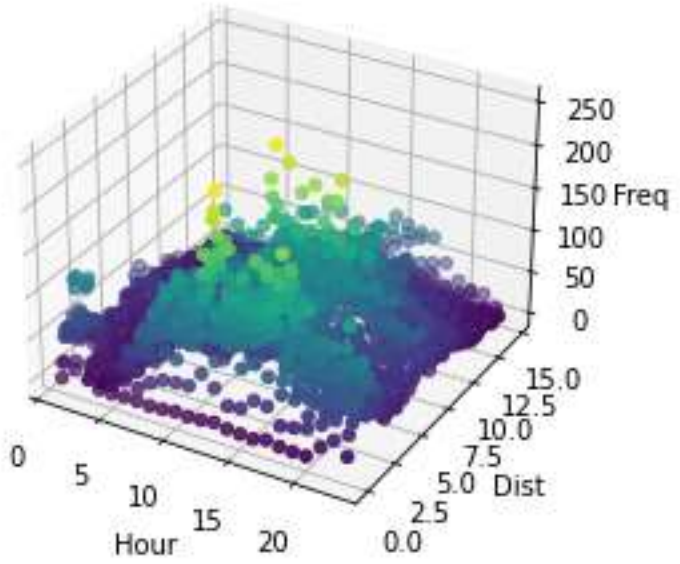

Figure 4: 3D Plot showing relation between Time, Distance and Frequency of Incidents

When a decision tress, KNN analysis and Logistic Regression (By Binning the results as this is not a classification data but continuous data) was done on this we were not able to get any accurate result (best accuracy of 0.056). A regression can be performed on this; however, this would require us to perform a Multiple Polynomial Regression. This is beyond the scope of this project.

Here we come to the next analysis, which is analysis of distance and accidents from the city center.

The longitude and latitude data was rounded to 2 digits (after zero) and using a value_counts function, the frequency (number of accidents) were found. By plotting, number of accidents vs. distance from city center it showed some trend.

A polynomial regression was done, and the R2 Score was 0.6 for the 6th degree of polynomial, if we lower the degree of polynomial the score reduces further. The equation can be written as

$$Yd = Xd^6 + 3129.44 * Xd^5 - 2065.35 * Xd^4 + 527.85 * Xd^3 - 63.12 * Xd^2 + 3.544 * X - 0.0756$$

Where,

Xd = Distance from Center of City (in Kms).

Yd = Frequency of Accidents.

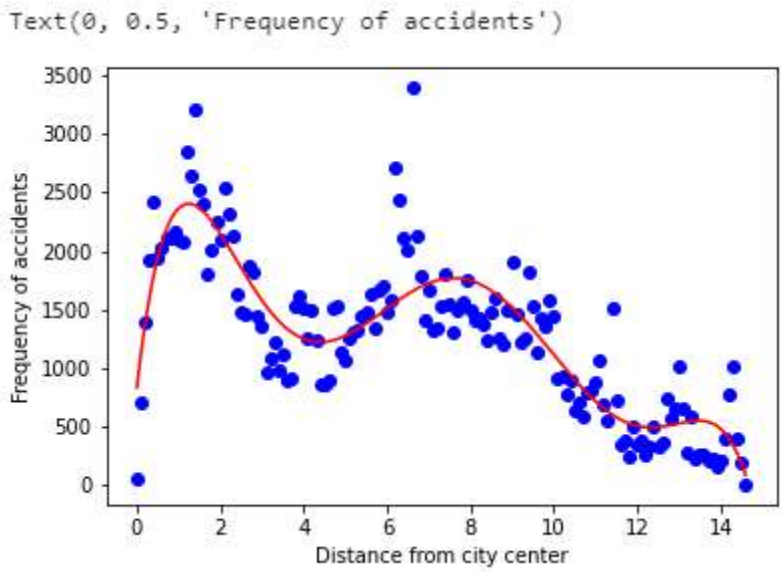

Figure 5: Scatter Plot and Trendline for Distance from Center of City and Frequency of Incidents

Considering the degree of the equation, this is too complicated and confusing. Another thing is it doesn't give any idea about the region where it happened. For example, 2 kilometers from the city can be in any direction. While giving an initial quote the computer / agent may use this formula but would not be able to accurately predict the frequency of accidents in that region. Hence, another analysis using folium map was the done. The results show the following map,

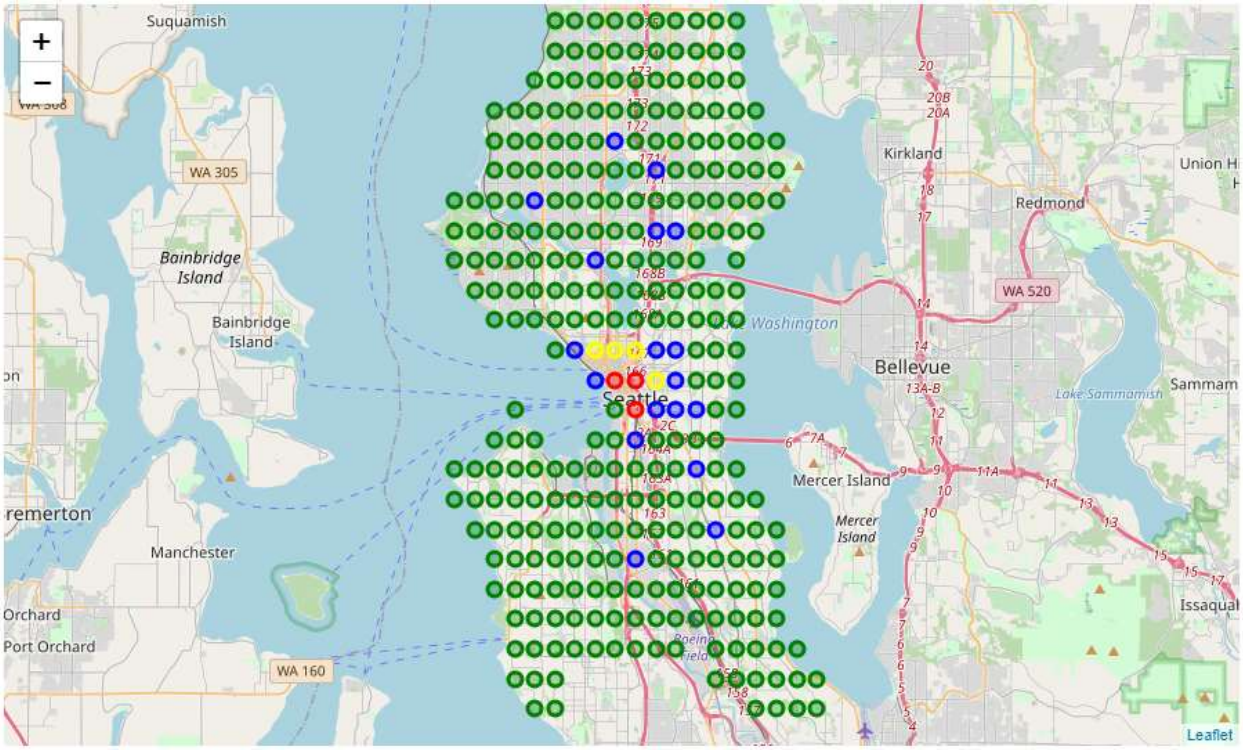

Figure 6: Map of Seattle and Frequency of Incidents

The region with red markers have between 4500 to 6000 incidents, the region with yellow markers have 3000 to 4500 incidents, the blue markers have 1500 to 3000 incidents and the green have 1 to 1500 incidents.

The results show that, the incients are high near the regions like Pioneer square, Westlake and First Hill have very high frequency of accidents, followed by areas like University Street, Seattle Center, Yesler Terrace. Mostly the heart of the city is where the frequency of the accidents is high. From insurance perspective this is the very high risk zone.

Lastly, analysis on incidents involving parked cars was done. This was to analyse the conditions like weather, road, lighting, and parking location. A decision tree analysis was done to come to conclusive results.

From the decision tree, parked vehicle are hit when

- Good Road Conditions, Weather and Light Condition. (Entropy = 0.997, pure node)
- Road Conditions are Oily or snow/slush with good lighting conditions. (Entropy = 0.811, pure node)
- Road Conditions are unknown, Weather is snow, sleet or unknown, light condition is known and it is not daylight. (Entropy = 0.996, pure node)
- and in other 2 cases, but since the entropy is 0.0, not worth to be investigating

In short, the road condition and weather (snow or sleet) make a big impact on incidents that involve parked vehicles. This is important as the parked vehicle in incident may not be paid today but becomes a liability for tomorrow. In cases like ‘hit and run’ or ‘uninsured motorist’ the insurance company has to pay its customer even if it is not the customer’s fault.

Also, it is important to see that 167 parked cars were hit near intersection. Now, in US, usually we have no-parking zone close to any intersection. These cars were therefore (possibly) parked in that no parking zone. Also, from all the parked cars only 2 cars were in an incident out of a total 11,782 hit parked cars, that is less than 0.2%.

A person who parks in alley can get a better quote than one parking on the block.

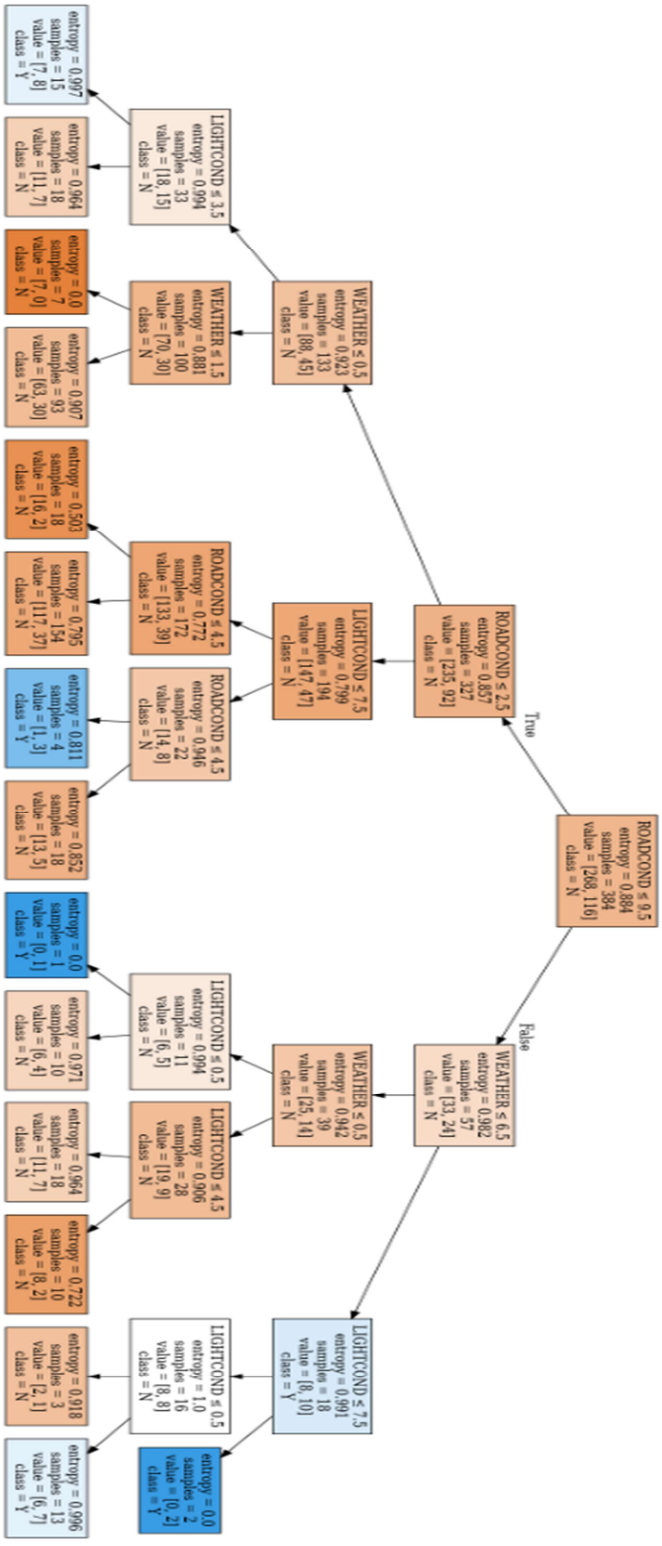

Figure 7: Decision Tree Analysis of Parked Vehicles involved in an incident.

Conclusion:

Insurance, although mandated by law, is not a compulsion but convenience. It gives motorist a peace of mind that they are taken care off in case they get in accident. With the changes in law in recent time, general view about insurance is that it is a burden not an ease. Often a person involved in an incident talks about going to a lawyer to file a suit against an insurance company.

This report was to make change the complete outlook of general populace about the idea of insurance and to improve their insurance experience on one hand and help the insurance companies give a better quote and assist them to help their customers in the worst time.

The objective can be reached by participation of the customers and insurance company working hand in hand with each other. Can the customer (and everyone in society) not drink and drive? Can they not be speeding? Can they be attentive while driving? These may be simple questions, but the data analysis shows otherwise.

Also, can everyone of us (Seattle specifically) avoid driving in the peak hour (11:00 am to 1 pm)? We can have the companies change the work timings that will distribute the commutation time. We can ask the commuters, the pedestrians, cyclist to be careful during these peak hours. There may be company vehicles that might be involved in incidents. Can the companies ask their drivers to avoid the rush hours for service? Distributing the traffic and spreading our travel hours may help to avoid incident rates.

Now if everyone makes the above changes, we can always have insurance companies change their policies and reduce insurance rate and pay off claims appropriately.

Insurance companies too have responsibility towards the society. If people depend on Insurance Companies in times of need, the insurance company also depends on the people to run their business. By getting complete information about the driver's travelling information, his commute routes and time, parking information a lower quote and insurance premium can be given. As the numbers decrease, it will be the moral responsibility of the company to pay complete remuneration to its customer (involved in incident) as promised when giving quote.

Lastly, even if no one takes the suggestions in this report seriously. Incidents continue as they are, people do not follow the precautions and insurance companies follow the same business practices, is it worth to be just a number in this report?