1. Introduction

Accidents and mishaps happen every now and then, and causes damages to property and living beings. These uncalled events happen due to many reasons, few in our control and few beyond our control. Of all types of accidents ranging from accidents in factories, at home, in airplanes and road accidents etc, **Road accident** contribute significantly in higher frequent occurrences, damages and more wider demographics of people. We will be studying road accidents and will try to get some insights to reduce road accidents by the data being made available. This will help everyone whether one drives in car, walks on road or have property near road. City administration can make use of the insights of this analysis and take appropriate steps to reduce accidents and overall loss of society.

2. Data used for study

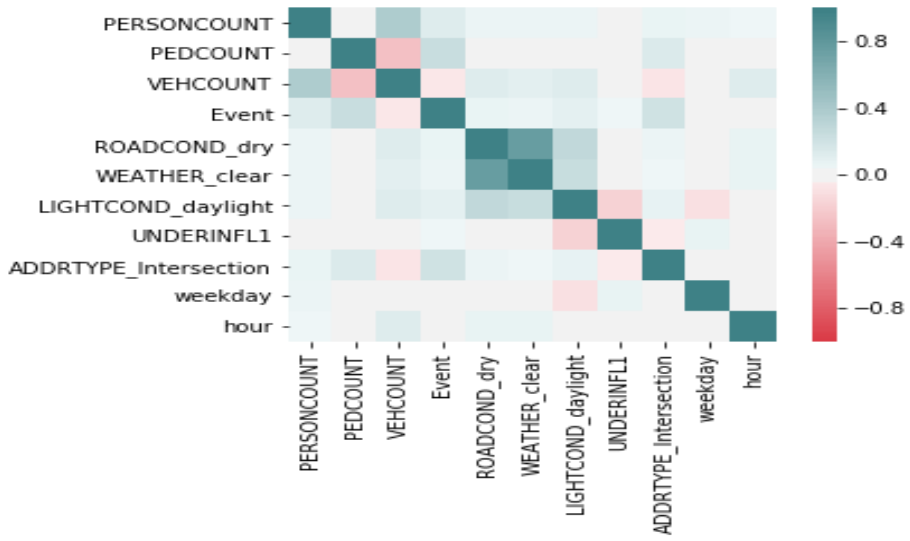
The data used is from Coursera's Applied Data Science Capstone project and owner of this data is "SDOT Traffic Management Division, Traffic Records Group". This data contains 194,673 cases of car accidents happened in Seattle, from 2004 to present. There are 37 attributes in data including road condition, light condition, weather etc.. during the accident.

3. Methodology

3.1. Data exploration and Data cleaning

We inspected all 37 attributes' data type, fill rate and description. We removed attributes that were not useful to solve our problem. Variables like case code, ObjectID and various keys for case type codes, junction types were excluded. We were left with 10 variables, there was no fatality information and the variable nearest to it was severitycode which distinguished between property damage and injury. We converted this variable into binary, with 1 being injury.

We clubbed similar and more frequent attributes of the variables together. Example in case of weather we clubbed all other than Clear weather together, Road condition dry and all rest together, lighting as daylight and all rest together. Also for date, we brought days of the week together and hours for time.



4. Results:

We ran logistic regression and got accuracy of 72%, with mean error of 0.27. From the matrix it is clear that we can to some extent predict if injury is likely to happen.

5. Discussion

We have predicted if accident happened then whether there will be injury or property damage. This limitation of our assumption is due to the fact that we have not analysed overall traffic of the region.

6. Conclusion

Since the overall set we analysed was only all accident happened and not all vehicles driven in the region, we can only use these numbers for "If accident happens, the probability of injury will be explained by model". But nevertheless, this itself is not to be ignored and whenever there is higher probability of injury suggested by model the traffic can be more thoroughly monitored, efforts for having the road dry, better lighting and traffic easing steps during peak hours should be incorporated.