Predicting the car accidents severity

1 Introduction

1.1 Background

Accidents has little to high impact on other people travelling to/through that particular location. So if we can predict accident severity based on past data about weather/location/time, people can accordingly plan their travel carefully. And lot of traffic jam can also be avoided.

1.2 Problem

Based on past data which has Weather/Road/location/time conditions we can predict accident severity So that commuter can change travel plans accordingly. And hence travel safe. Infact accidents can also be prevented if one knows about such information report from past in some kind of app or from news.

1.3 Interest

There will be many drivers of all kind who might be interested in knowing such accident predictions which might help them to plan their travel accordingly. Also many urgent task based on time constraints can use such type of information as needed. Definitely this is useful information to drivers.

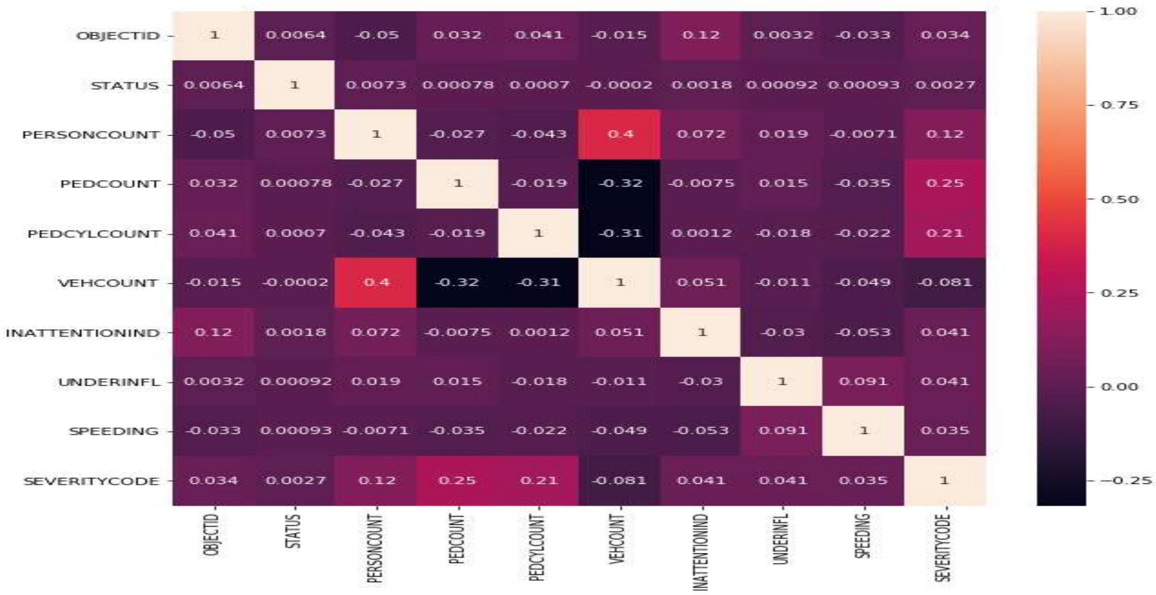
2. Data acquisition and cleaning

2.1 Data sources

Here collision data from Seattle police department will be used. The target or label columns are accident " severity" in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact. Collision Data will be loaded to IBM cloud project notebook to work on it. Data parameters like location and weather can be used to predict where more number of accidents happened in the past and with what type of severity. And to get more specific like Alley, Intersection or Block we can use Address Type field. Junction type like "Mid-Block (not related to intersection)" will give more such information. Similarly Intersection key INTKEY can be used which corresponds to the intersection

associated with a collision. Collision code can be used to get accidents with such same collision descriptions like "MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END". Further we can use Weather, Road condition & Light conditions parameters (like Overcast, Dry & Daylight) to see come up with such type of predictions. ST_COLCODE can be used to get more details about description by state's code designation for example "One parked--one moving". We need to balance the target data and eliminate the variables with many NaN value. For the missing data, we can either impute the missing value or eliminate it directly.

- Data rows: 194673
- Attributes: 37
- Target: SEVERITYCODE



The categorical data need to be converted into numerical data for the model to understand and predict. Below fields were encoded.

- 'LIGHTCOND'
- 'ROADCOND'
- 'WEATHER'
- 'COLLISIONTYPE'

Library used : IMBLearn

Before Balancing	
1	84748
2	37980

After Balancing	
1	75960
2	75960

3. Methodology

We can build a logistics regression model to predict the probability of a high severity accident.

- Multinominal NaïveBayes
  HyperParameter – Alpha =1.0
- Gaussian NB
- Random Forest Classifier
  HyperParameter – max_depth=15, random_state=0,n_estimators=100,max_features=20

4. Results and Discussion section

Confusion matrix and area under ROC curve can be used to evaluate the model performance. There are two types of models, regression and classification, that can be used to predict accident severity. Regression models can provide additional information on the severity of accident, while classification models focus on the probabilities an accident might happen. The underlying algorithms are similar between regression and classification models, but different audience might prefer one over the other. For example, an traffic mangement team executive might be more interested if accident happens (regression models), but a general driver might find the results of classification models more interpretable. Therefore, in this study, I carried out both regression and classification modeling.

Algorithm	Confusion Matrix		Precission	Recall	F1 Score	Accuracy
GaussianNB	34954	6919	66	68	68	67
	12228	6448				
MultinomialNB	24961	16812	75	66	67	66
	3685	14991				
RandomForestClassifier	21999	15261	73	71	71	71
	6110	31461				

5. Conclusion section

A report is prepared for the deployment. Other users can treat it as reference to know the project details. In this study, I analyzed the relationship between accidents probability and their severity data. I identified location, weather, time, accidents from past years among the most important features that affect an accident. I built both regression models and classification models to predict whether and how much an accident would be severe. These models can be very useful in helping traffic team management in a number of ways. For example, it could help identify locations/time/days to acquire, estimate the size of the effort needed to clear the traffic, plan for alternate routes if needed. So there are lot of possibilities.