Capstone Project - Car Accident Severity

1. Introduction

1.1 Background

Car accident causes damages to both properties and injuries, even death. Approximately 1.35 million people die in road crashes each year, on average 3,700 people lose their lives every day on the roads. An additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities[1].

Everyone wants to have a safe road journey, however, various of factors could lead to an accident in a second and human beings cannot take 100% perfect measures to prevent the accident from happening.

1.2 Business Problem

The problem that we are going to solve is to predict the severity of having a car accident by taking into consideration multiple factors that could lead to a car accident.

This prediction can help to warn the driver in advance and can therefore, hopefully decrease the real accident rate.

1.3 Interest

The government could be potentially interested in the project as it could help to decrease local the injury or fatality rate.

Besides, insurance companies could be a potential user as they can prevent the car accident and therefore, decrease the insurance policy claim cost.

2. Data acquisition and cleaning

2.1 Data source

I got the data from coursera suggested list: you can also find the data in

Applied Data Science Capstone – week1 – Downloading Example Data Set.

The Data has 38 types of car accident relevant factors and include 1 946 734 rows of records.

The data set looks very comprehensive and should have enough sample to train the ML model and get a decent prediction result.

2.2 Feature selection

After analyzing all the attributes, I decide to use the following 7 attributes as input:

- INCDTTM: The date and time of the incident.
- INATTENTIONIND: Whether or not collision was due to inattention. (Y/N)
- UNDERINFL: Whether or not a driver involved was under the influence of drugs or alcohol
- WEATHER: A description of the weather conditions during the time of the collision
- ROADCOND: The condition of the road during the collision
- LIGHTCOND: The light conditions during the collision
- SPEEDING: Whether or not speeding was a factor in the collision

[1] Data from Association for Safe International Road Traveling, weblink

One common point is that almost all the data is categorial and not continuous.

2.3 Data cleaning

There are several problems with the dataset.

Firstly, there are many missing values for different attributes. I mainly used the following three ways to handle with it:

- For attribute with Yes and No value, as there is no NO answer, therefore, I assume all the missing value means No
- For attribute with few missing values (less than 5%) and hard to predict the value, I dropped the line directly
- For attribute with few missing values (less than 5%) but easy to predict the value, like the time per day, I use the mean of the time series to fill the empty value.

Secondly, most attributes are in text format. Therefore, I choose to define the text with numbers and convert them into integer for further modeling.

3. Exploratory Data Analysis

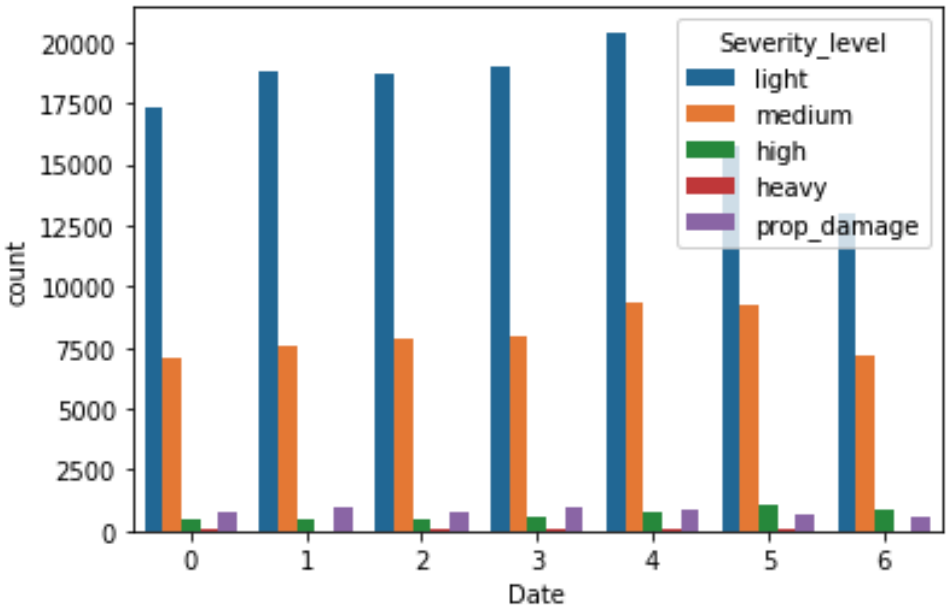
3.1 Calculation of target variable

The output is about PERSONCOUNT - The total number of people involved in the collision. I strongly believe that the number could be a good measurement for the severity of the car accident and I categorize the severity as below:

- Level 1 means only property damage (0 person involved)
- Level 2 means light car accident(1-2 persons involved)
- Level 3 means medium (3-5 persons involved)
- Level 4 means high (5-10 persons involved)
- Level 5 means heavy (more than 10 persons involved)

Here the output is also categorial instead of continuous.

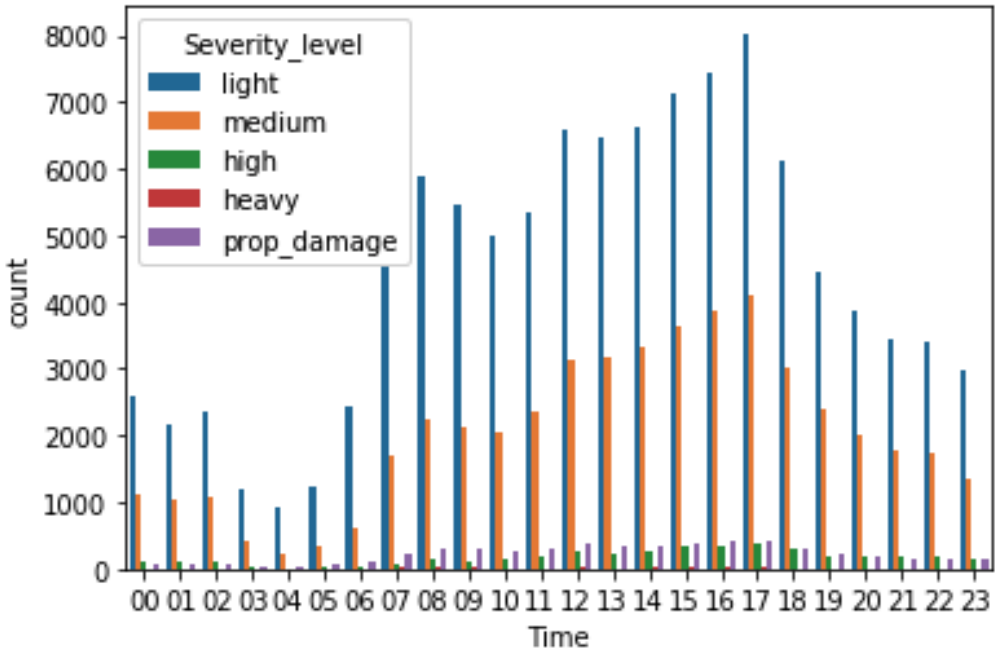
3.2 Relationship between severity and weekday



The frequency of light car accident is significantly lower during weekend (number 5 and 6). Besides, Friday is more likely to have car accident no matter of the severity of the accident.
Please note that based on the code rule of pandas, 0 stands for Monday, 5 stands for Saturday and 6 represents Sunday.

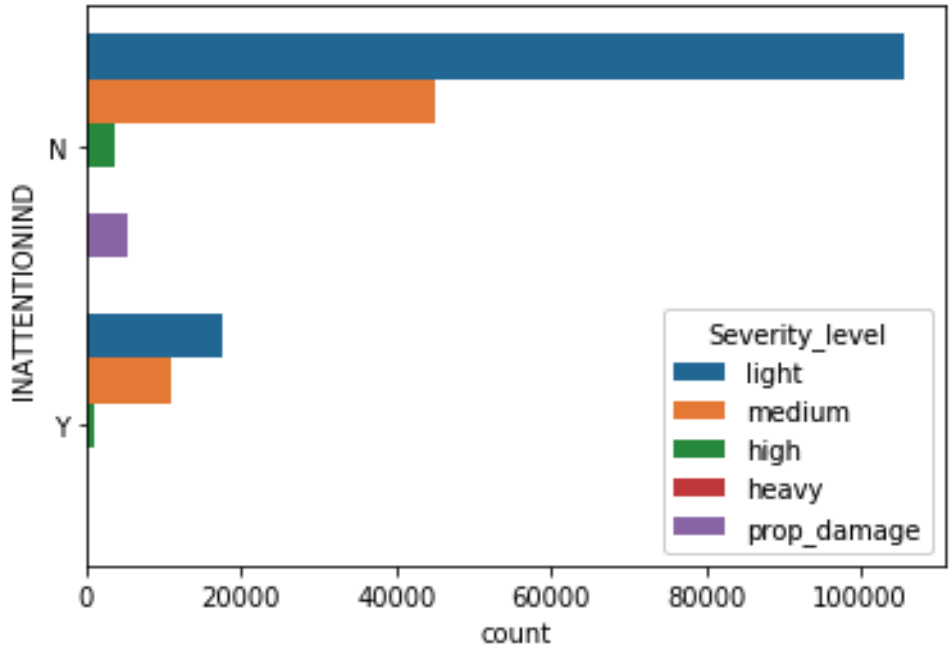
3.3 Relationship between severity and day time

The frequency of car accident starts to increase after noon and reaches a peak at around 17h no matter of the car accident severity level. It could be that in the afternoon, drivers tend to be fatigue and is less focused.



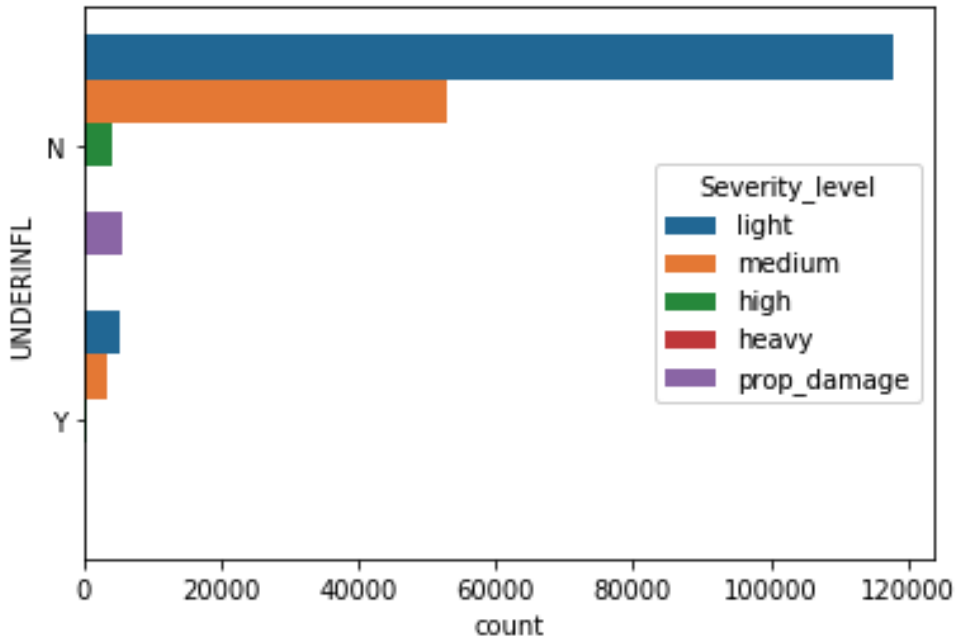
3.4 Relationship between severity and inattentioning

A certain number of car accidents happens due to lacking of attentions. More accidents happen for others reasons no matter of the severity level of the car accident.



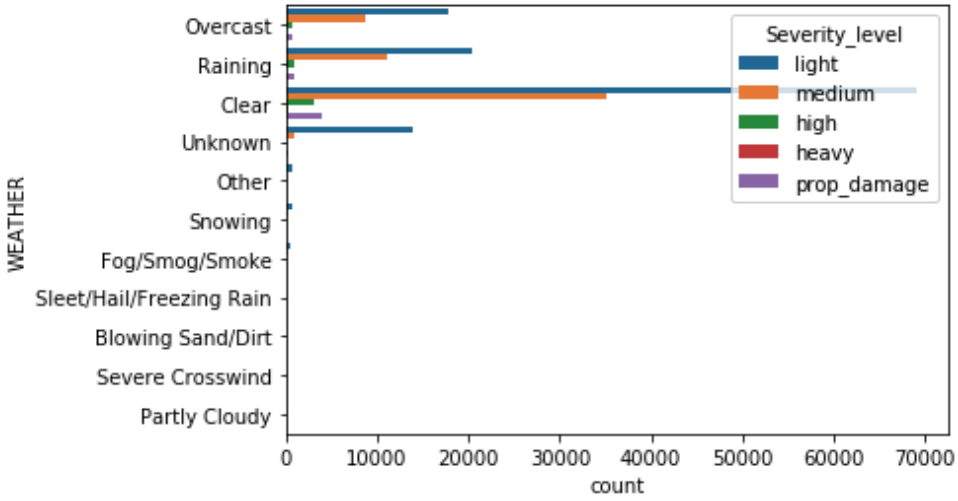
3.5 Relationship between severity and UNDERINFL (impact of alcohol)

A certain number of car accidents happens due to alcohol impact. More accidents happen for others reasons no matter of the severity level of the car accident.



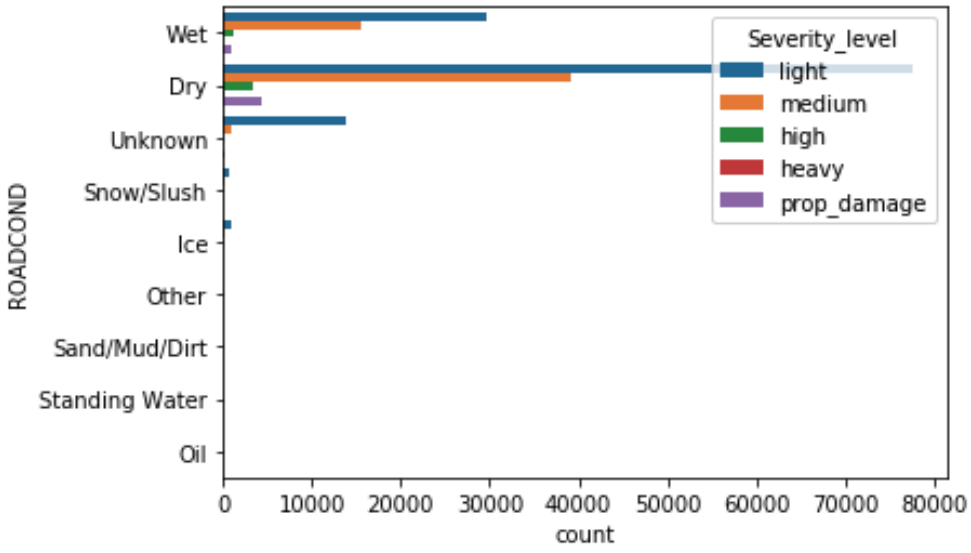
3.6 Relationship between severity and weather

It's more likely to have car accident when the weather is clear no matter of the severity of the accident. Based on the generated graph below, Clear, Raining and Overcast are top three weathers that have the most cases of car accident.



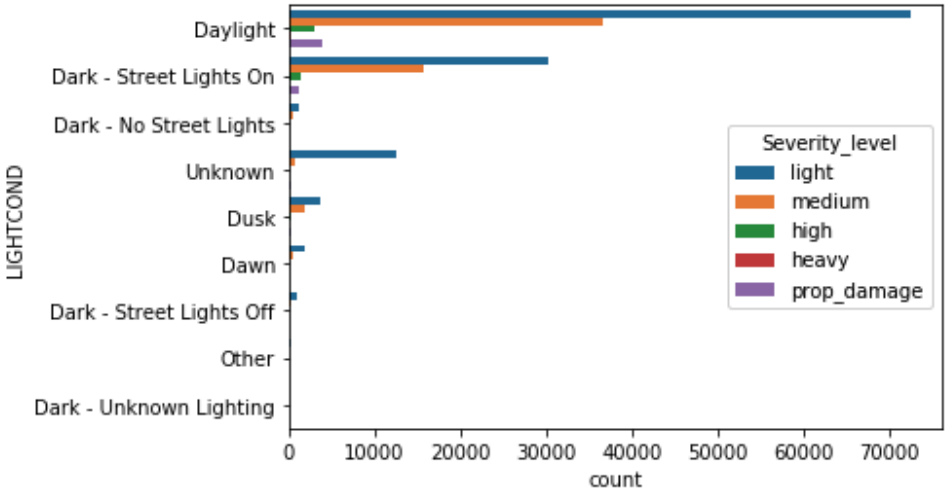
3.7 Relationship between severity and road condition

Most car accidents happen when the road condition is dry, the next high one is wet.



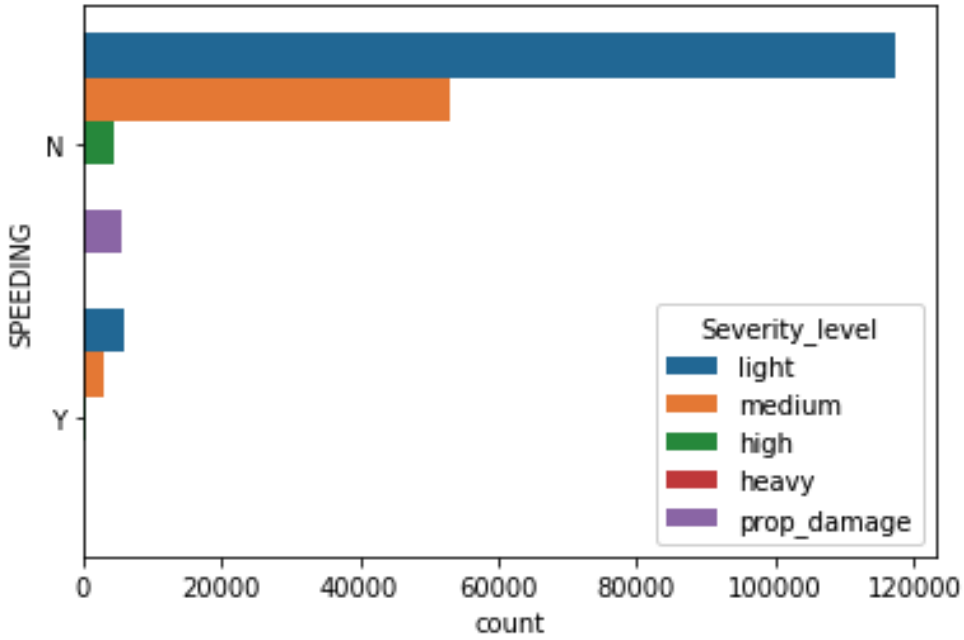
3.8 Relationship between severity and light condition

Most car accidents happen when there is daylight, the next high one is when it is dark but the street lights on. It seems that the more normal the light condition, the more likely that driver get diverted and lead to car accident.



3.8 Relationship between severity and speeding

A certain number of car accidents happens due to over speed. More accidents happen for others reasons no matter of the severity level of the car accident.



4. Predictive Modeling

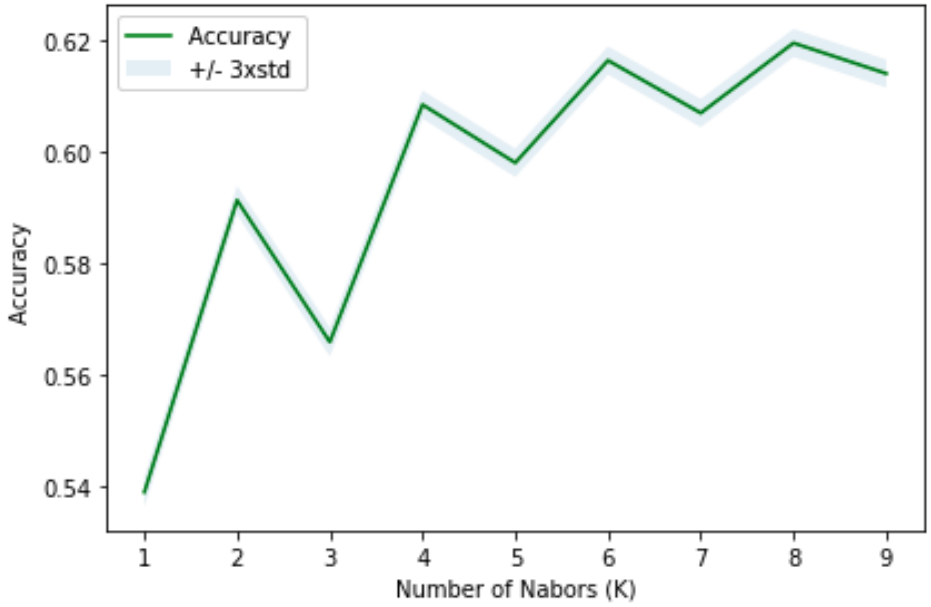
In general, there are two types of modeling in supervised learning: regression and classification. I choose to use different models in classification instead of regression for the following reasons:

- Most of the attributes are categorial instead of continuous
- The prediction target is to predict the severity of the car accident, and it is also categorial result.

I split the example data set into train (80%) and test set part (20%).

4.1 K-nearest Neighbors

The best accuracy was with 0.62 with k= 8:



4.2 Decision Tree

The decisionTrees's Accuracy is 0.64.

4.3 Support Vector Machine (SVM)

I use the train set to fit the model and test set to generate yhat.

4.4 Regression

I use the train set to fit the model and test set to generate yhat probability.

In summary, the decision tree model performs the best with 0.64 accuracy.

5. Performance of classification models

I choose to use F-1 score and Jaccard to evaluate all the models and plus log loss evaluation for regression.

	F1-score	Jaccard	Log loss
K-nearest Neighbors	0.55	0.61	na
Decision Tree	0.51	0.65	na
Support Vector Machine	0.51	0.65	na
Regression	0.51	0.65	0.82

6. Conclusions

In this study, I analyzed the relationship between car accident severity (by number of persons involved) and selected 7 features of datetime, inattentioning, impact of alcohol, weather, road condition, light condition and speeding. I used 4 different classification model: K-nearest Neighbors,

Decision Tree, SVM and regression. These models can be very helpful to predict the severity of the car accident, warn drivers in advance and then to decrease the car accident potentially.

7. Future directions

I was able to achieve ~64% accuracy in the classification problem. However, there was still significant variance that could not be predicted by the models in this study:

- The example data set has some missing values: I made some assumptions to fill the missing values. If we could get more qualified data or data from other data source, the prediction could be more accurate.
- The type of attributes is not diversified. As most of the available attributes are categorial, it could be hard to train other types of model such as non-linear regression.
- The intercorrelation of attribute is not clear and could potentially impact the accuracy of the prediction. For example, the weather could impact the road condition and the impact of alcohol may have correlation with inattentioning.