Capstone Project

PREDICTING ACCIDENT SEVERITY

Introduction

- Despite the fact that the US population has increased threefold since the beginning of the 20th century and the total number of cars cracking the 280 million mark in 2019 (source), leading to a whopping 3000 billion miles travelled p.a., fatality rates in traffice continously decline. This decrease of deaths in car accidents is related to measures, enforced by the law (e.g. seat belt law, 1968), advanced safety features (mandatory air bags (1998)), but also improved road safety (e.g. signs, traffic lights etc.).
- Moreover, thanks to technical advances in computer technology in the last decades, efficient measures can be taken even after an accident happened, e.g. by minimizing the response time of emergency teams and police through the smart analysis of accident records.
- This coding exercise demonstrates how collision records can be analysed and provide insight into predicting car accident severity using the example of Seattle, USA.

Data Overview

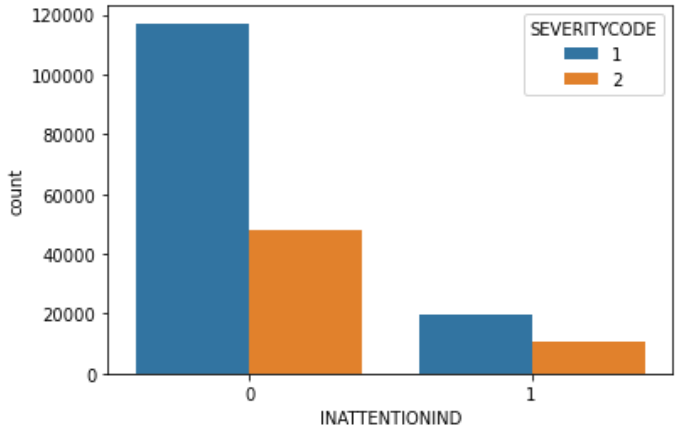
- The Data which is used for development of the model is the Collisions data that is provided by the Seattle department of transportation and Seattle police department and recorded in the traffic records. The data consists of all the Accidents taking place from year 2004 to Current year and is updated weekly. The data includes information about various attributes including weather conditions, road conditions, visibility, Accident severity. Since data is updated continuously the model can be trained and tested for continuous improvement and increasing the accuracy. For example:- In case of condtions such as when visiblity is poor or the road condition is not good the model will give a warning to the travellers.

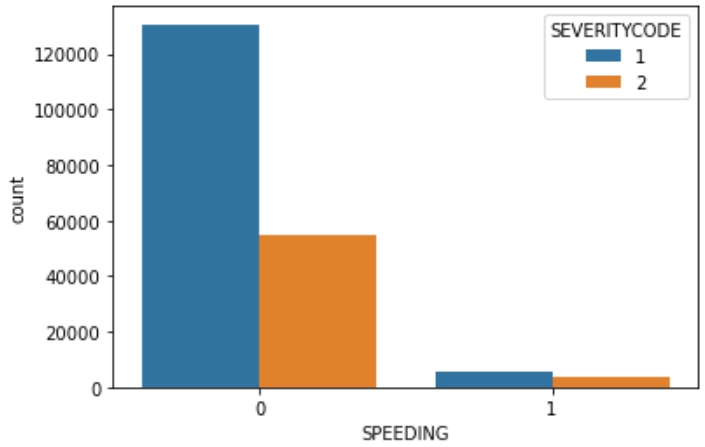
Methodology

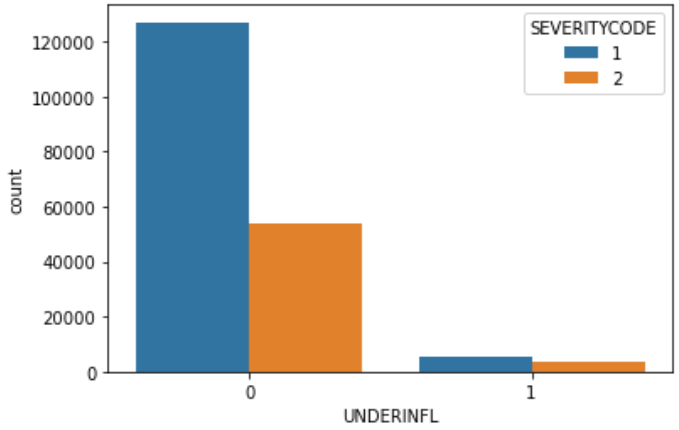
The workflow for the data analysis is as follows: -

1. Data Loading
    1. Involves downloading the data from the repository and storing it as a dataframe using pandas
2. Data Overview
    1. Exploring the data in terms of dimensions, format, type of data and volume
3. Data Cleanup
    1. Cleaning up the data of non value adding information and making it fit for the model development
4. Exploratory Data Analysis
    1. Getting Insights from the data about frequency, distribution and pre-selection of features for Model Building
5. Model Building
    1. Normalizing the data
    2. Benchmarking Differnet Models

Exploratory Analysis

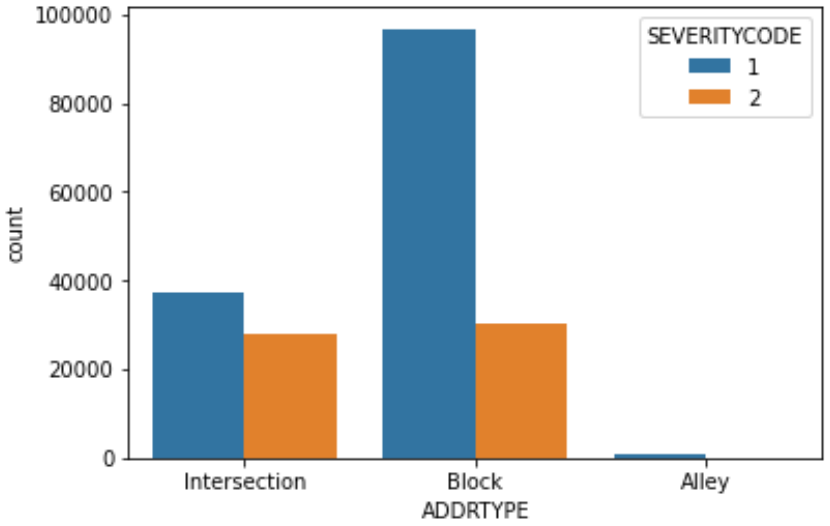






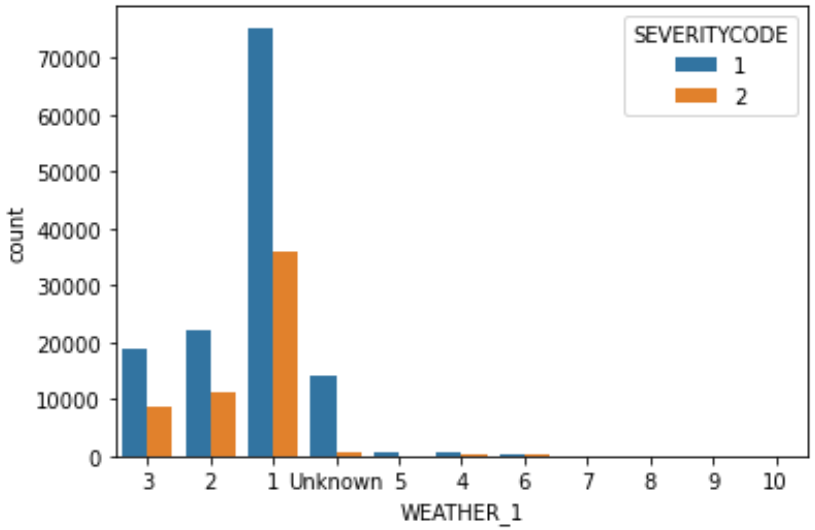
- Since INATTENTION, DRIVING UNDER INFLUENCES and SPEEDING are criminal offences we can say that on Analysis of CRIMINAL OFFENCES v/s SEVERITY CODE it is found that criminal offences increases the chances of accidents resulting in injuries(Severity Code=2) by approximately 8-9%

Exploratory Analysis



- On Analysis of data feature ADDRTYPE, describing the location of accidents shows that majority of accidents that result in injury(SEVERITYCODE=2) are intersection related as compared to any other location.

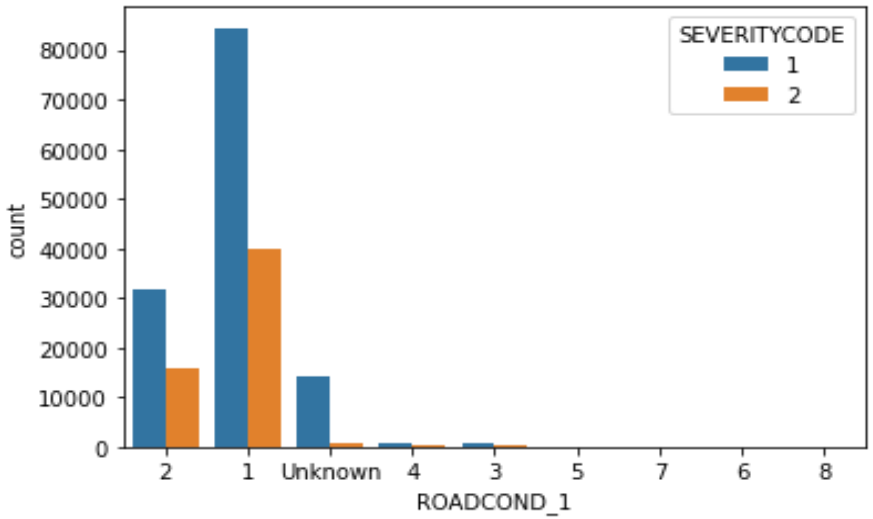
Exploratory Analysis



Weather	Value
Clear	1
Raining	2
Overcast	3
Snowing	4
Other	5
Fog/Smog/Smoke	6
Sleet/Hail/Freezing Rain	7
Blowing Sand/Dirt	8
Severe Crosswind	9
Partly Cloudy	10

- Maximum accidents that result in injury takes place mostly in partly cloudy weather condition that is around 60% but this result cannot be considered statistically stable as only 5 counts of such accidents are recorded. The other weather conditions that have high rate of injury related to accidents are rainy(33.72%), fog/smog/smoke(32.86%), overcast(31.55%) and clear(32.25%). It seems logical that due to lower visiblity conditions and wet roads the chances of accident leading to injuries increases.
- With Clear being the one proving to give a counter intutive idea at first glance it can be explained further in the report.

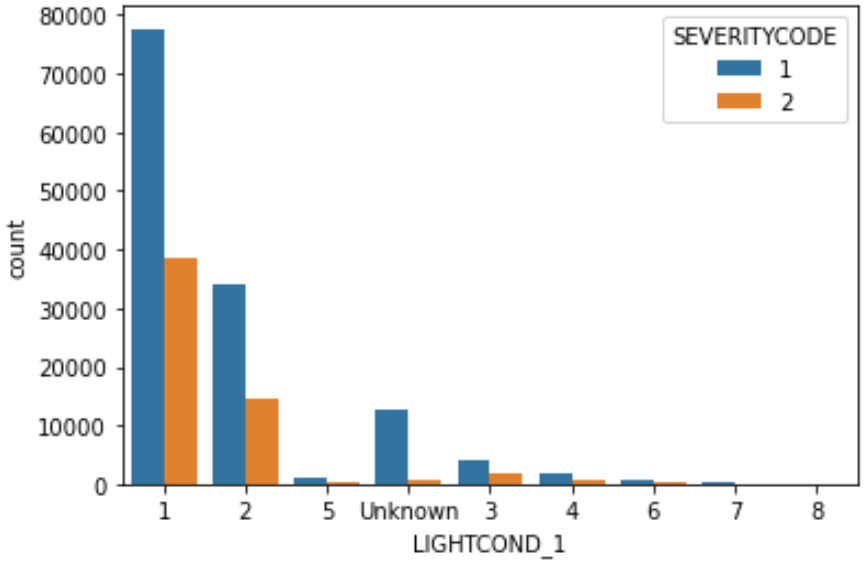
Exploratory Analysis



Road Condition	Value
Dry	1
Wet	2
Ice	3
Snow/Slush	4
Other	5
Standing Water	6
Sand/Mud/Dirt	7
Oil	8

- Oily Roads have one of the highest risk causing accidents that lead to injuries at 37.5% but with only 24 counts of such actual accidents being recorded shows that this statement is not stable statistically. The other conditions that impose a high risk are wet and dry which can be directly correlated with weather conditions.

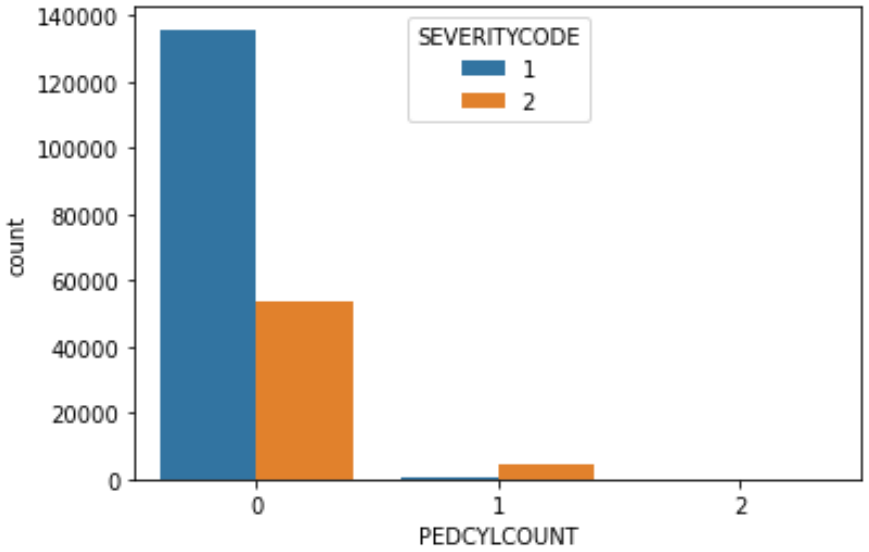
Exploratory Analysis

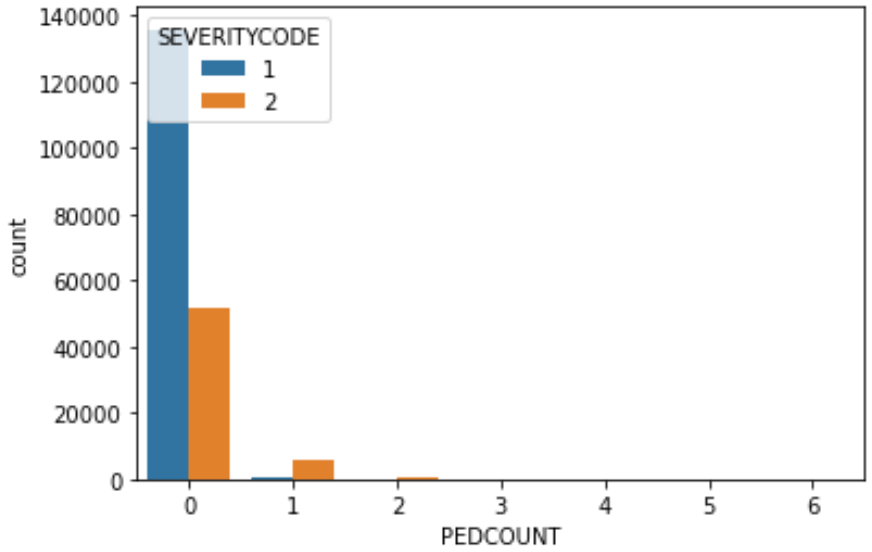


Light Condition	Value
Daylight	1
Dark-Street Lights On	2
Dusk	3
Dawn	4
Dark- No Street Lights	5
Dark- Street Lights Off	6
Other	7
Dark- Unknown Lighting	8

- On anlysing the data feature of LIGHT CONDITION it can be said that almost majority of injury related accidents take place at time of day, dawn and dusk with 33.19%, 32.94% and 32.94% respectively.

Exploratory Analysis





- The involvement of pedestrians and cycle users increase the risk of accidents leading to injury(SEVERITY CODE=2) by approximately 88%. This can also explain the high rate of accidents with injuries during the clear weather, with dry roads as during such weather conditions the number of people coming out of their house increases

Classification Models Used

- Logistic Regression
- Support Vector Machine(SVM)
- Decision Tree
- K-Nearest Neighbors

Model Evaluation

Model	Without Resampling			With Resampling		
Algorithm	Jaccard Score	F-1 Score	Log Loss	Jaccard Score	F-1 Score	Log Loss
Logistic Regression	0.7385	0.6909	0.5335	0.5466	0.5115	0.7057
Support Vector Machine(SVM)	0.7382	0.6904	NA	0.5067	0.6360	
Decision Tree	0.7386	0.6910	NA	0.4983	0.6333	
KNN	0.7290	0.6905	NA	0.5276	0.5297	

Result

- There were two models constructed respectively: -

  1) In which the data was not resampled

  2) In which data was resampled

- Both the models were then trained on the same classification algorithms to make a fair comparison between the two. It was found that the models which were trained on the dataset that was not resampled had a higher accuracy for the predictions and lower uncertainty as compared to it's counterpart. Though the data without resampling might have a little bias due to a very high value of accidents having SEVERITYCODE 1 in respect to accidents having SEVERITYCODE 2. But due to resampling the accuracy and uncertainty increases and this means that some important information or trend goes missing. Thus, impacting the accuracy and predictive ability of the model.

- For Logistic Regression Model the value of log loss for data without resampling is almost 32.27% higher showing a very high level if uncertainty. Similar trend is visible for Jaccard score as well as F1 scores for all the other models. This model also gives a large value of false negatives for accidents of severity class 2 and fails to make accurate prediction for them.

Conclusion

- The aim of this project is to predict the severity of accidents and to reveal the factors affecting the severity. When examining the feature importance after the training step, it seems possible to say that there are several remarkable features. It seems quite effective that a pedestrian or cyclist was involved in the accident.

- Going back to the exploratory data analysis step, we were able to get many important following insights such as majority of accidents that take place in wet conditions of road are directly related with the rainy and overcast weather while majority of the accidents that take place in dry conditions involve majority of pedestrians and cyclists. The accidents that result in majority of injuries take place at dawn, daytime and dusk. This proves the logical conditions of lower visibility will result in higher injury accidents to be untrue on majority. The intersection as a place has a large amount of accidents with injuries and property damage as the collision of vehicles take place at an angle. The criminal offences such as inattention, driving under influences and speeding increases the risk of injury causing accidents and the timings of these offences explains the high accidents during dawn and dusk.

Suggestions

- Various points that can make suggestions based on all insights can be listed as follows,
- Drivers should be more careful where pedestrians and/or cyclists are concentrated. (*is_ped* and *is_bike*)
- Drivers should be more careful at intersections. (*addr_type_intersection*)
- Age as a factor should be considered by seattle transportation department which will be very useful in explaining criminal offences.
- Extra measures can be taken to prevent driving under drugs and alcohol. (*under_infl*)
- Apart from the features that distinguish the two classes from each other, suggestions for situations where accidents occur frequently can be listed as follows:
- At hours close to office hours drivers should be more careful.
- Special precautions can be taken by the relevant authorities, especially since there are angles type collisions at intersections.
- Authorities should look at districts with schools, universities and offices where the pedestrians and cyclist are concentrated.
- Authorities should also make rules, take precautions and make arrangements for the safety of pedestrains and cyclist.