Big Data Project

Car Accident Severety^{Prediction} 😎

Made by: Ayhem Bouabid & Sinii Viacheslav

Year: **2023**

Data Characteristics

The dataset does not contain any NaN values:

	accident_index	loc_east	loc_north	lng	lat	accident_severity	n_veh	n
0	0	0	0	0	0	0	0	

Data Preprocessing

Accidents

Going through the explanations of the fields, it is possible to reduce the number of features as some of them do not introduce any additional value for our predictive task. The following will be dropped because these values represent the location of the accident with respect to a local geospatial system. This information duplicates another two columns which are longitude and latitude.

1. Location Easting OSGR (Null if not known)
2. Location Northing OSGR (Null if not known)

Other columns:

3. The police attendance - it happens after the accident happens
4. Longitude and Latitude might be dropped as the data is already clustered into different districts
5. Accident Severity: this value is practically equivalent to the target: the severity of casualities
6. Police: The police's intervention takes place generally after the accident. Such intervention could not possible affect the accident's severity and the casualities' seriousness

Additional remarks:

- The most seemingly important features are:
  - 1st /2nd Road Class / if it reflects quality
  - weather conditions / Light Conditions
  - Pedestrian Crossing Human control: we don't expect many accidents in conjuctions controlled by police officer: HOWEVER IT MIGHT HAVE SOME OVERLAPPING WITH CONJUNCTION CONTROL
  - Urban / Rural area: Rural area are more likely to have more fatal accidents: more serious casualities
  - SPEED LIMIT

Vehicles

- The data provides a detailed description of the vehicle
- Each unique vehicle is defined by the (accident_index, vehicle_reference) tuple. 'vehicle_reference' column indicates vehicle w.r.t. the accident.
- The fields most likely should be combined into a fewer but more general representations
- Certain fields might be dropped:

- Vehicle Location: can be deduced to a certain extent by the type of the road / location the accident took place
- Vehicle Maneouver is to be dropped
- There are two Hit Object features that can be merged into one
- The IMD level as well as the home area of the driver do not seem to have direct relation with the seriousness of the casuality any information about the driver can be found in the casuality table, so it should be dropped from the vehicle table

Results of Data Preparation

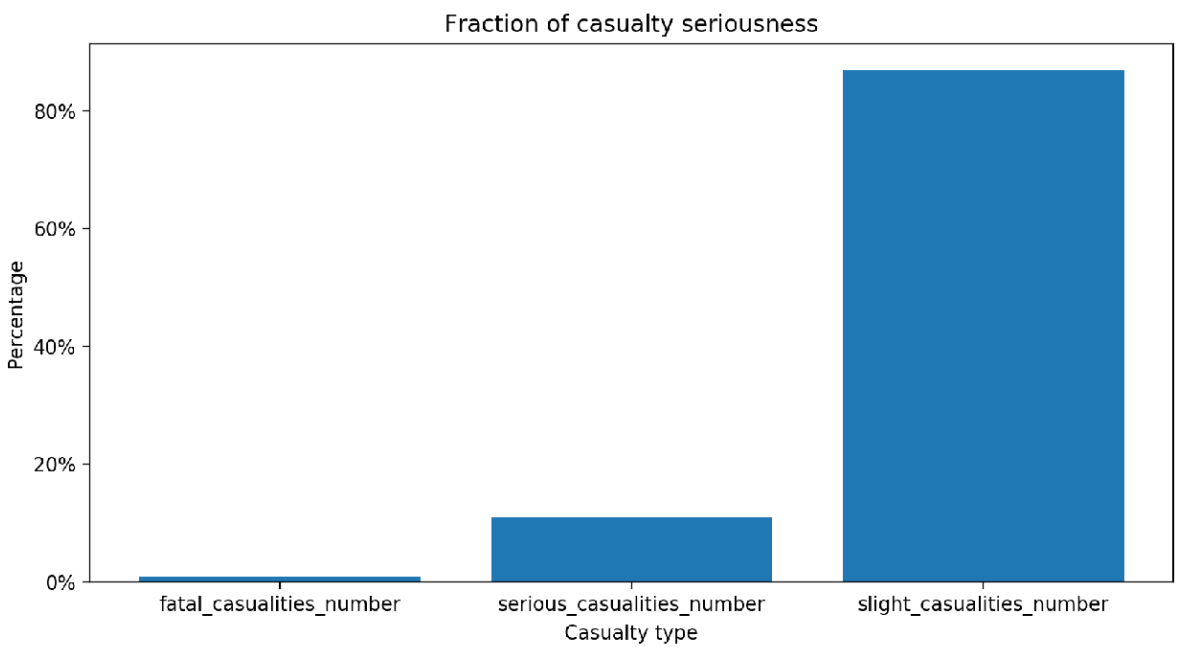
Before the cleaning there were 66 columns. We performed extensive cleaning and were left with 36 columns.

Target

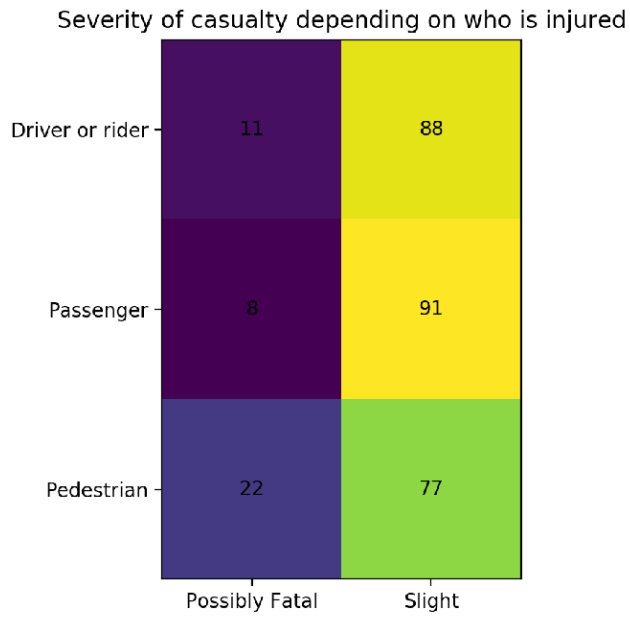
Predict casualty severity

EDA

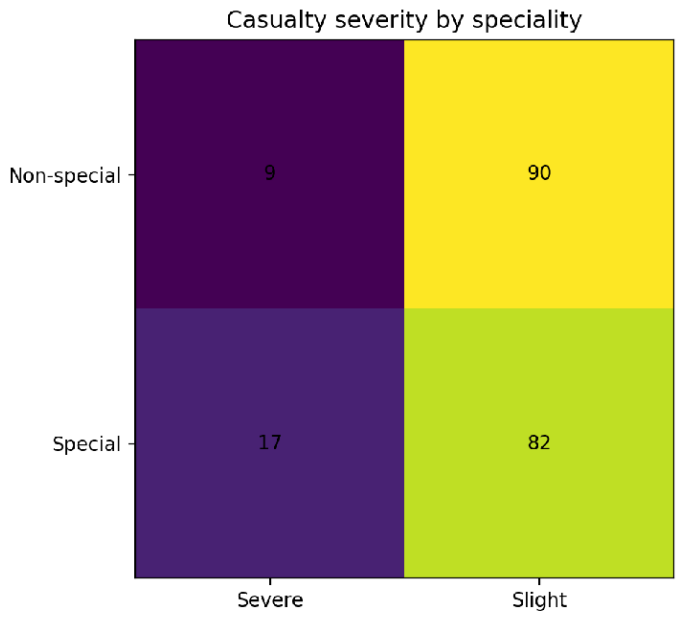
First let's see how many casualties of each types are present in the dataset. There is a serious class imbalance in our problem with slight casualties dominating the data.
We decided to merge 'Fatal' and 'Serious' casualties into a single category and to transform our task into binary classification. 0 will stand for Slight casualties and 1 for non-Slight.



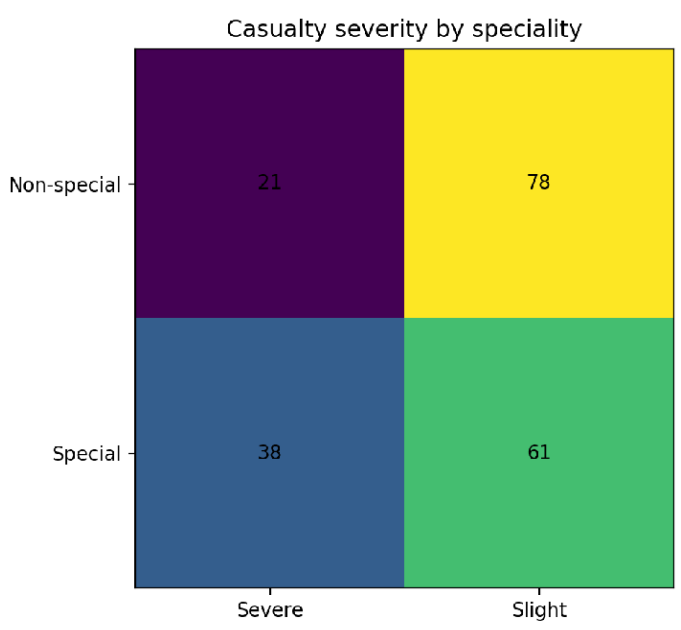
Pedestrians have a much higher risk of a serious injury compared to people located in the car.



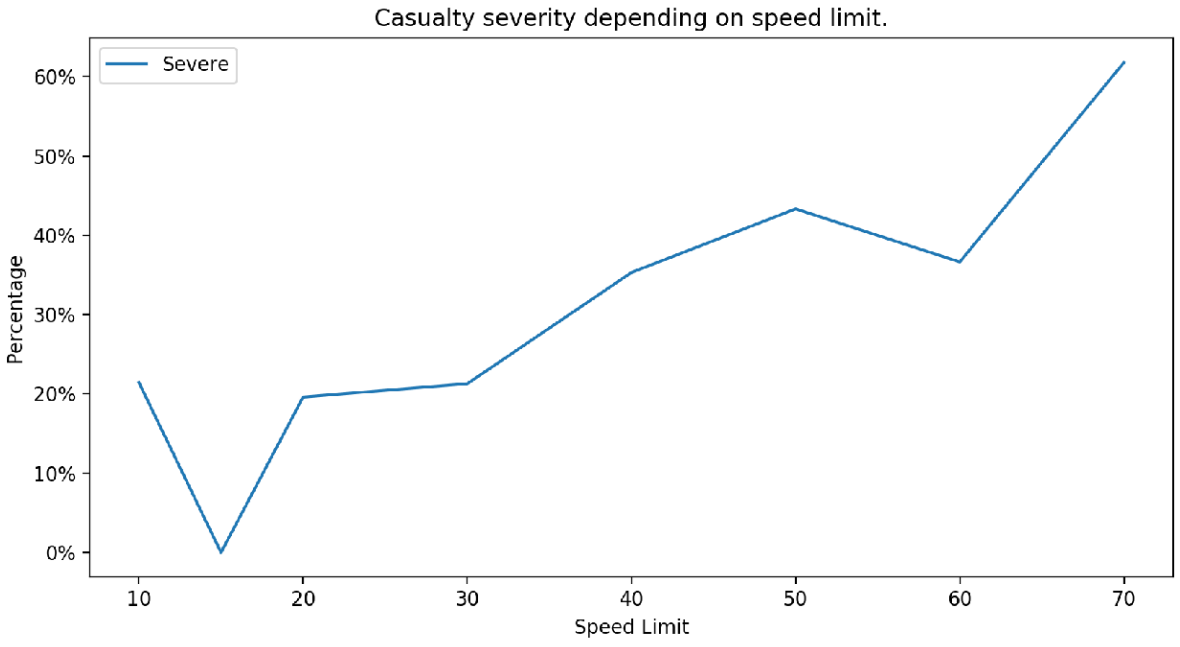
We can see that accidents with special accidentscircumstances such as the vehicle leaving the road, having the car not in its natural position / direction are 2 likely to have severe casualties. The first visualization corresponds to the entire dataset.



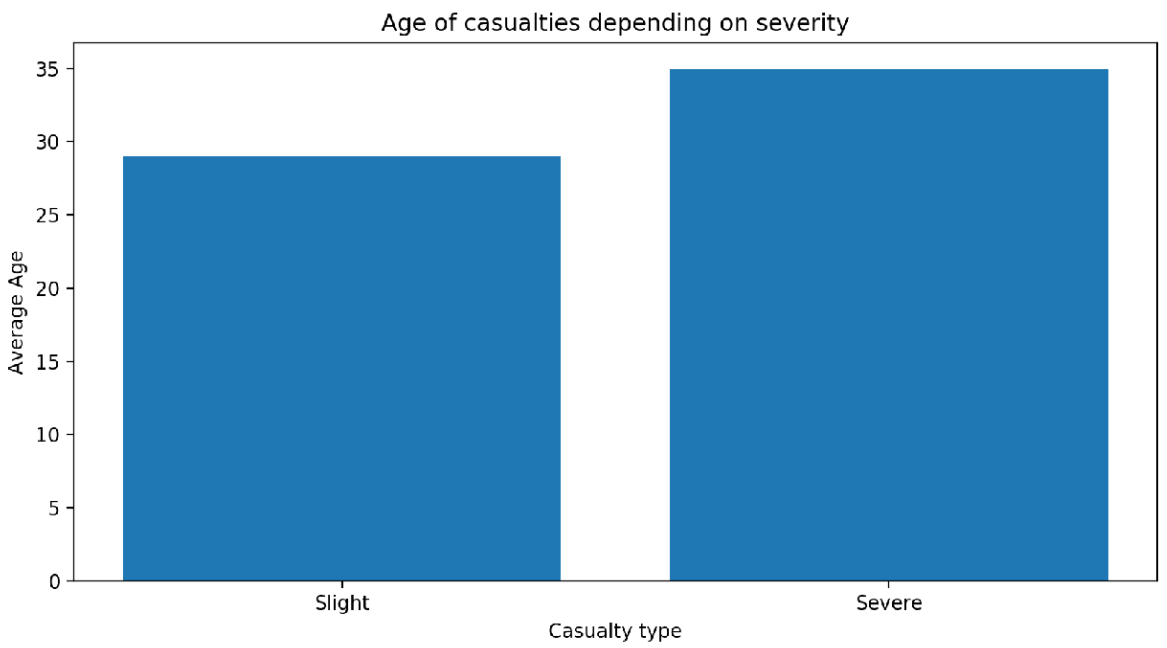
We can see that this observation is supported further in the 2nd visualization as it demonstrates the influence of such special circumstances on pedestrians' casualties



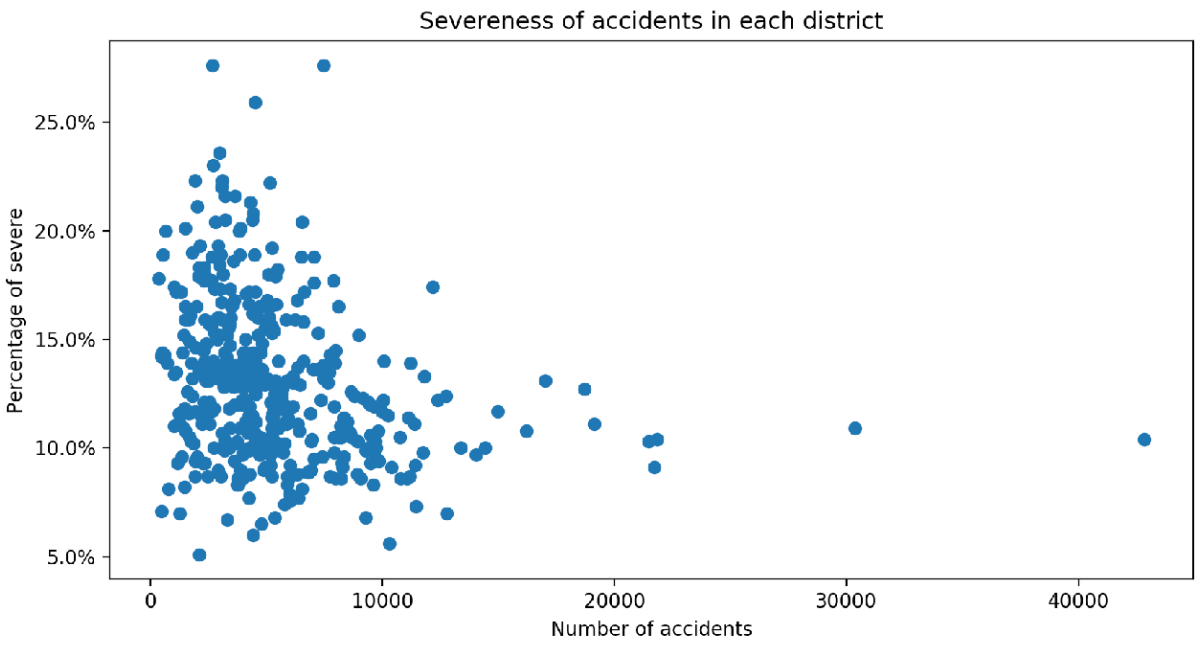
Here we study the severity of the casualty depending on the speed limit of the road where the accident happened. As the speed limit increases, the vehicles become deadlier leading to more serious injuries. Local regulations should be more cautious about setting the speed limit on some roads.
Thus speed limit math be a very informative feature to predict the casualty severity.



Now we want to see whether there is a connection between casualty severity and the age of a person. The graph below indicates that on average older people have a higher risk of fatal or serious injury compared to youngsters. This information may be crucial when medical aid decide who has to have a higher priority for help.
The age feature may be useful for predictions.



For each district in the country we extracted two values: total number of accidents that took place there and the percentage of severe accidents. The graph below shows that districts with less accidents are more dangerous. Small number of accidents may be due to a low number of people living there, i.e. districts in the countryside or suburbs. Thus, they are probably being less developed and have worse conditions leading to more severe accidents taking place.



Predictions.

We trained two models - Logistic Regression and Random Forest. The table below shows the probabilities of class 1 (Non-Slight injury) in the opinion of each model along with real values of the samples. We took 10 random samples for illustrative purposes.

	Logistic Regression	Random Forest	Real
0	0.1230	0.3354	1
1	0.1230	0.1433	1
2	0.1230	0.0555	0
3	0.1230	0.1291	1
4	0.1230	0.0903	0
5	0.1230	0.0383	0
6	0.1230	0.0817	0
7	0.1230	0.2221	0
8	0.1230	0.0573	0
9	0.1230	0.2626	1

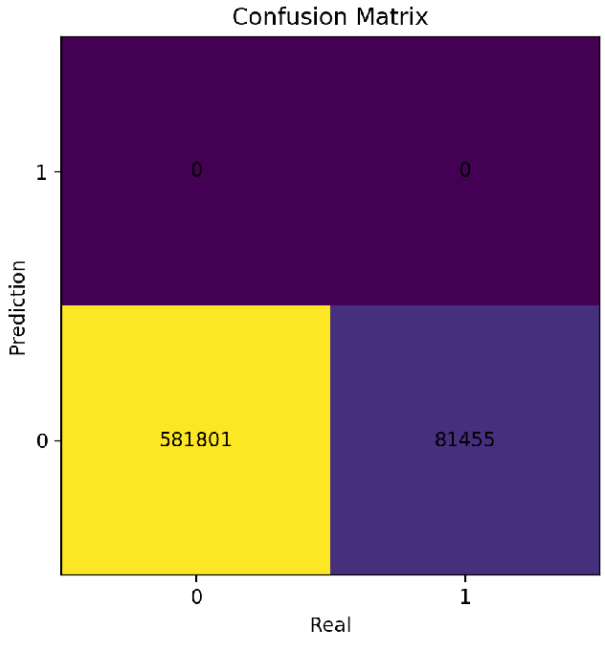
Evaluation.

Here are the Area Under ROC and Area Under PR metrics which we used to evaluate our models. As expected, a more complex and expressive model - Random Forest - have higher metrics values, i.e. performs better.

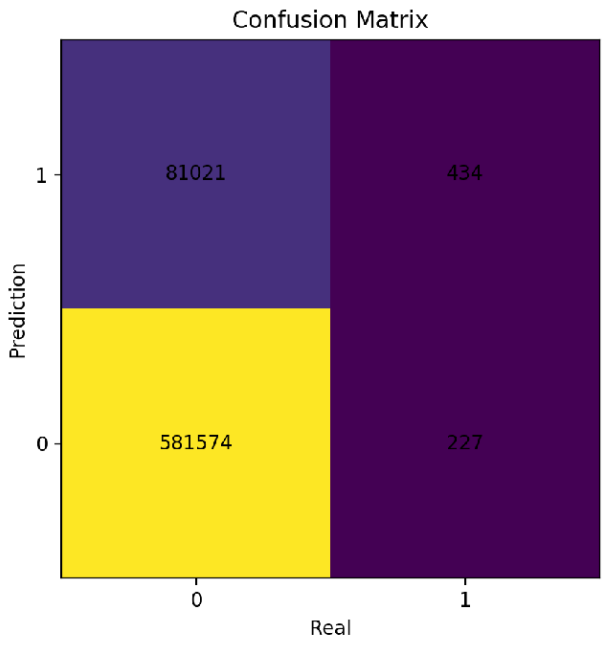
	model	area_under_curve	area_under_pr_curve
0	logistic_regression	0.5000	0.1228
1	random_forest	0.7453	0.2994

CONNECTING

Confusion Matrix. Logistic Regression



Confusion Matrix. Random Forest



Made with Streamlit