TRAFFIC ACCIDENT SEVERITY PREDICTION

Predicting the traffic accident severity using Machine Learning Algorithms

Applied Data Science Capstone by IBM/Coursera

Predicting traffic accident severity

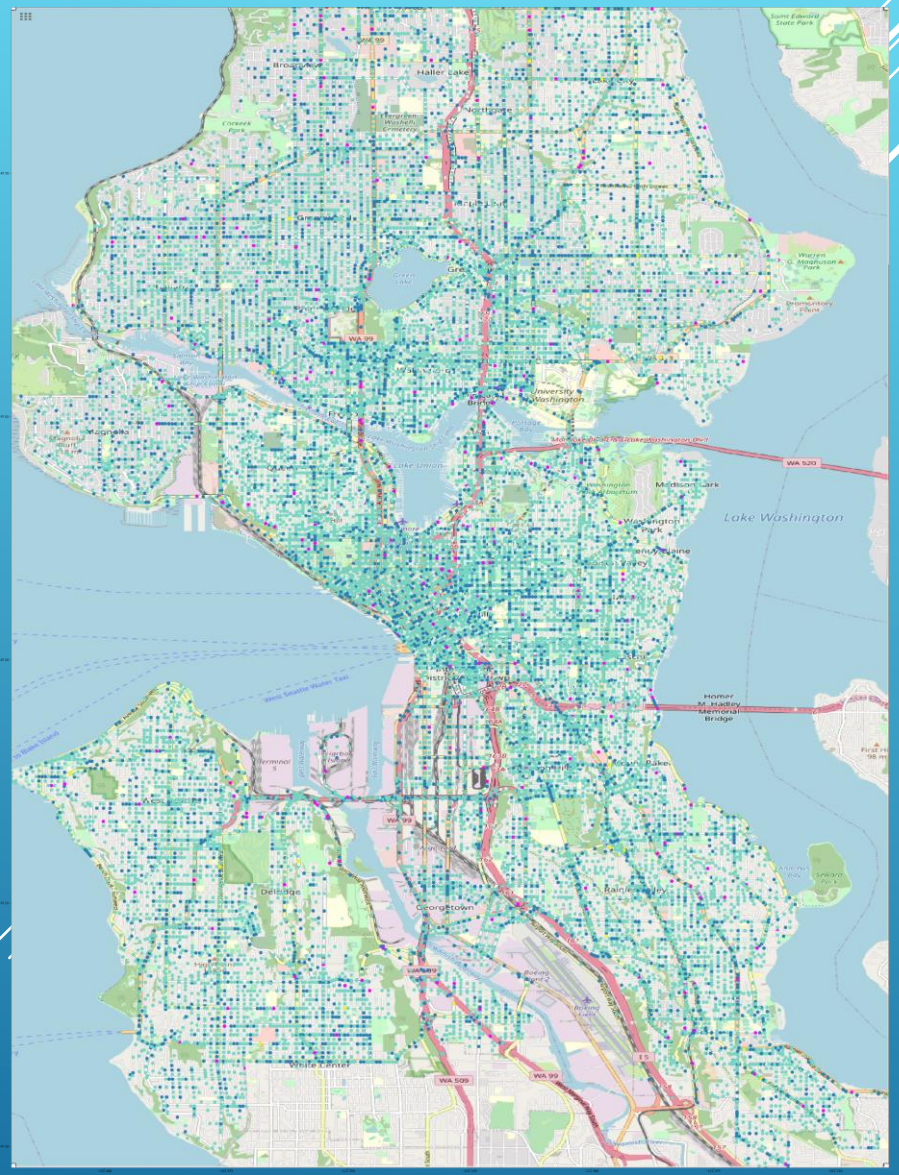
- More than 3700 people lose their lives everyday in road accidents
- Weather conditions, Road conditions and Light conditions are enough to predict an accident severity by their own? Which variable affect the algorithm the most?
- Predicting the accident severity could help each driver, telling them when to pay more attention or slow down in case of adverse conditions

Data acquisition and cleaning

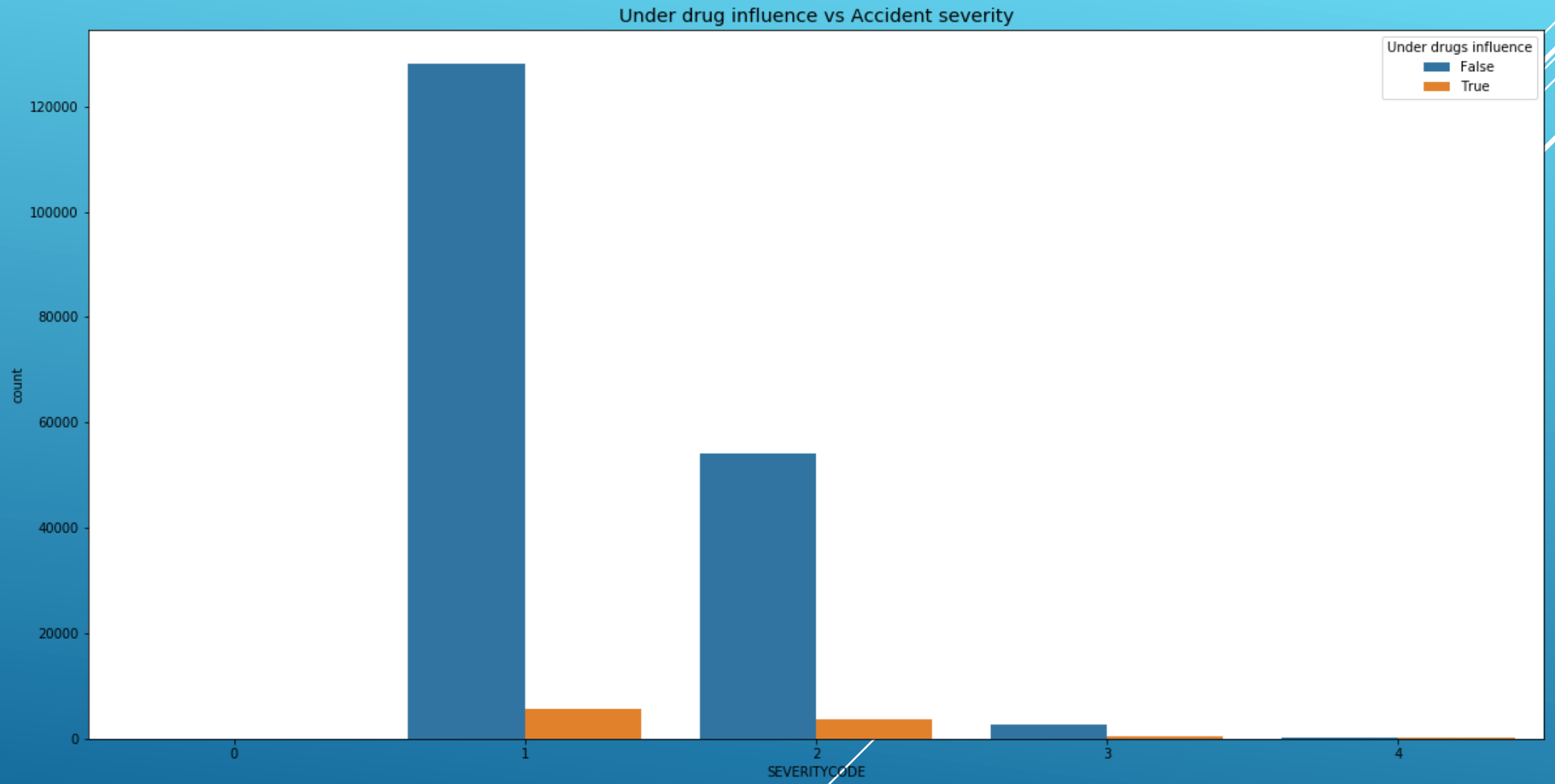
- The Seattle Collisions Dataset (2003-2020), available here
- The Seattle Collisions Dataset Metadata, available here
- Png Image of Seattle City that I especially made for this project, available here
- 221738 row in total and 40 columns
- Missing data will be dropped/replaced
- After cleaning and transformed categorical columns into numerical columns: 194827 rows and 41 columns (1st analysis); 194827 rows and 29 columns (2nd analysis)

Even if not in the scope of this project, the accidents map shows a concentration in the city-center and in the north of Seattle

Further analysis are required to establish a certain correlation with locations and density of population/business locations with the accidents

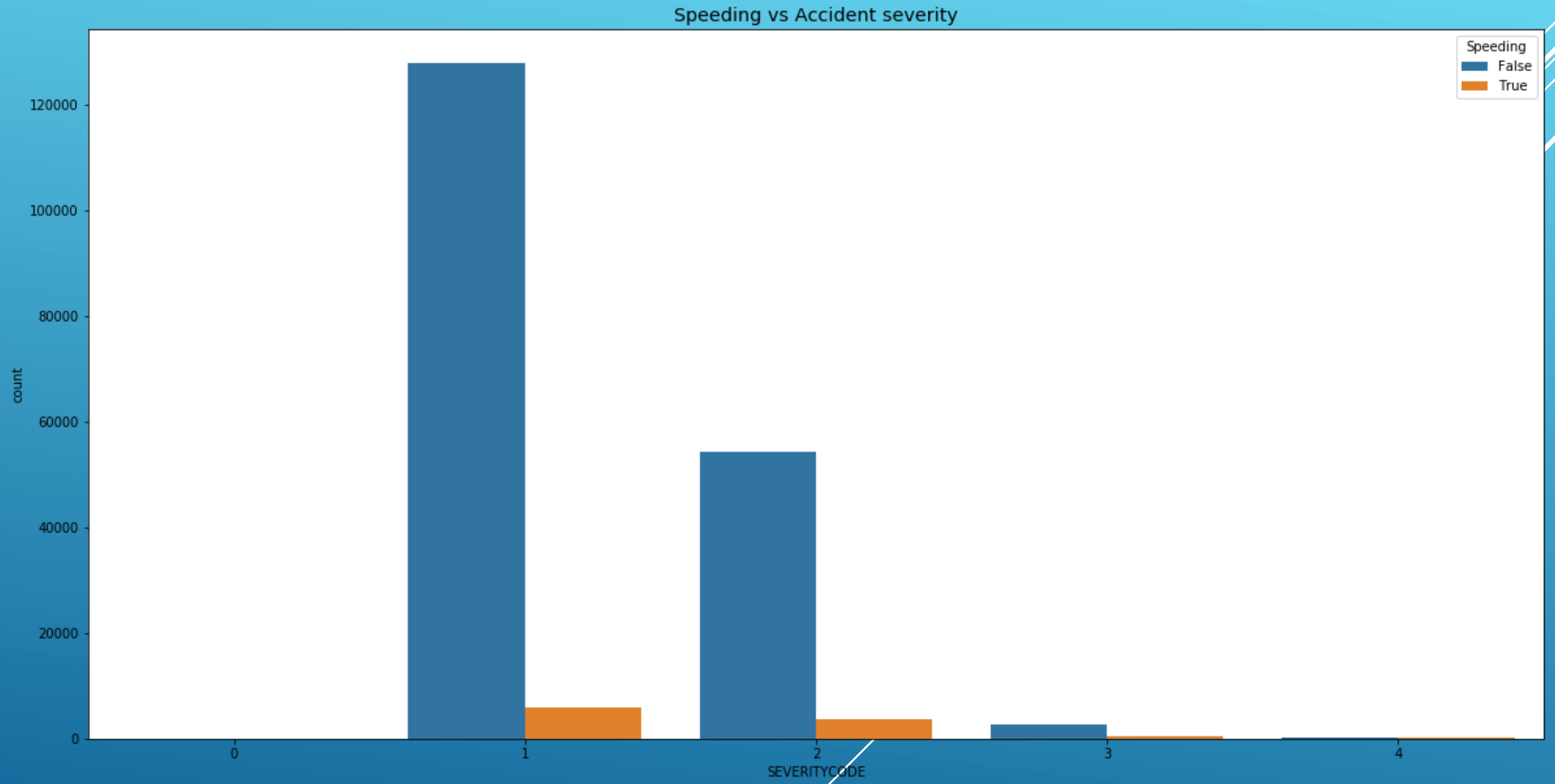


Under drug influence



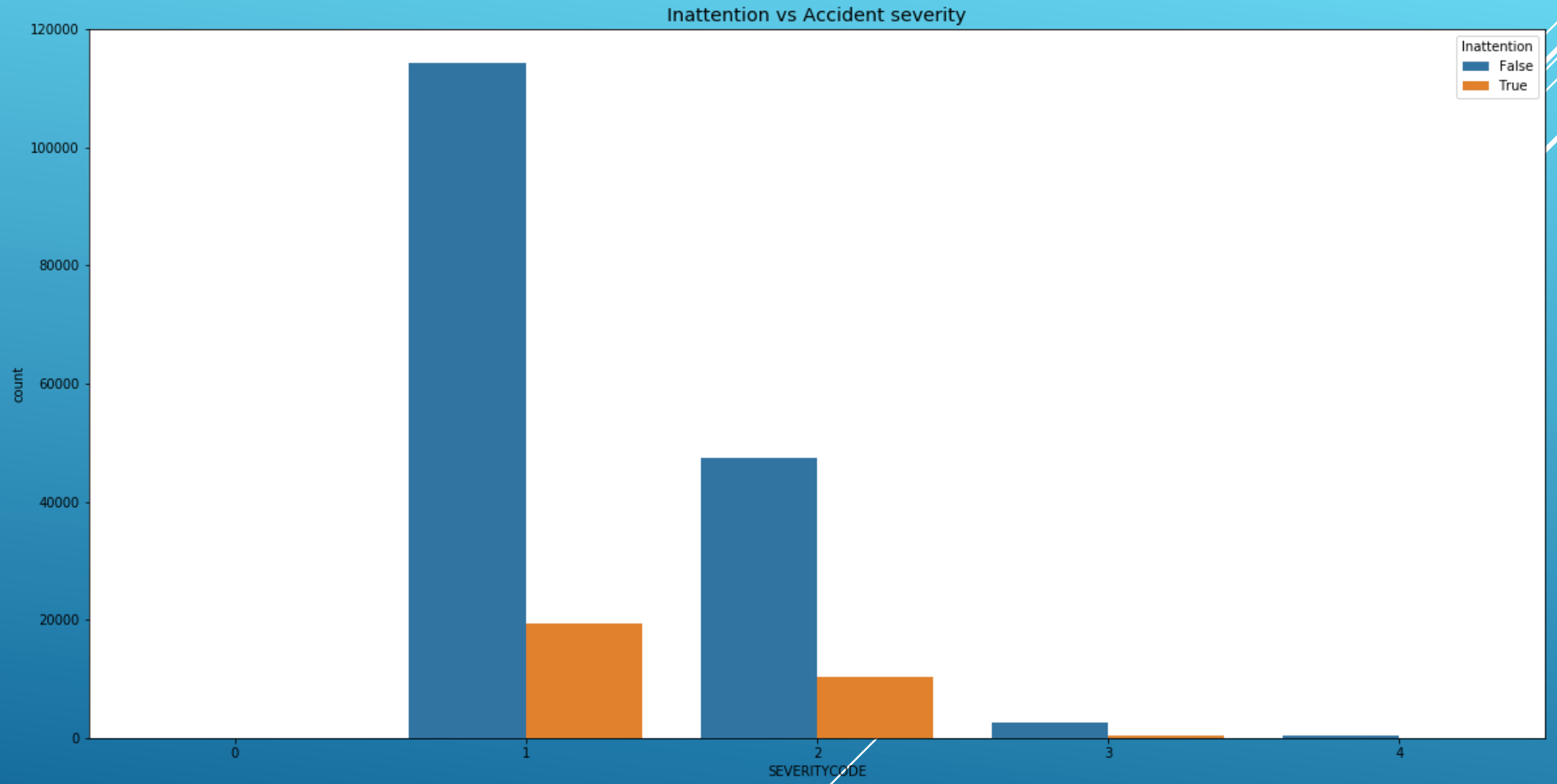
Under the use of drugs, the probability to incur in a serious accident (severity = 4) is ~ 1% against 0.1% of not being under drugs

Speeding



Similar to the drug use, the probability to incur in a serious accident while speeding are almost 1%, significally higher in comparison to normal conditions (0.13%)

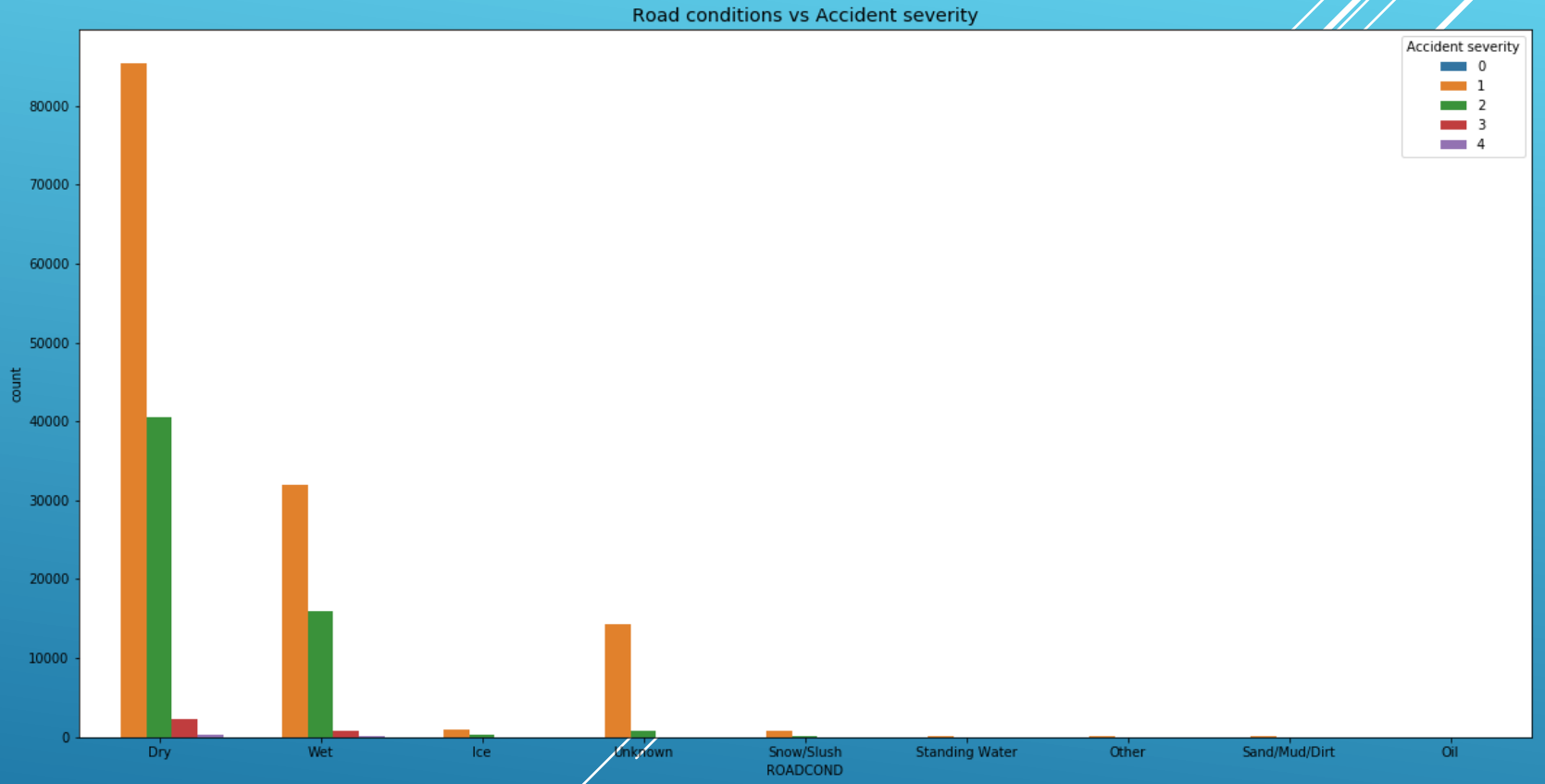
Inattention



The inattention is less impactful, since the probably to incur in a serious accidents are ~ 0.05%. Anyway, always pay attention at driving!!

Road Conditions

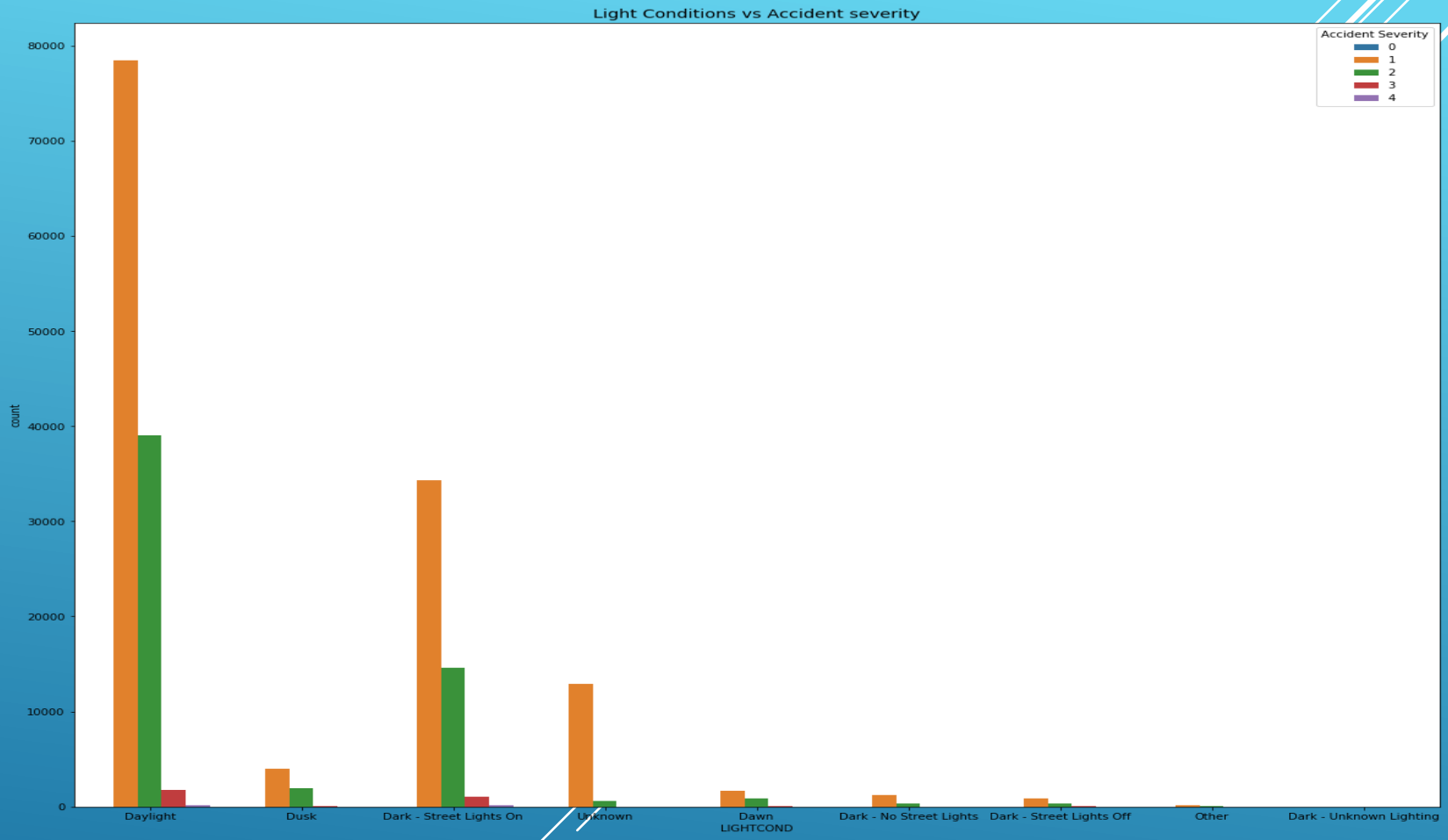
SEVERITYCODE	0	1	2	3	4
ROADCOND					
Dry	0	85349	40580	2232	268
Ice	0	935	275	18	1
Oil	0	40	24	0	0
Other	0	89	43	3	0
Sand/Mud/Dirt	0	53	23	0	0
Snow/Slush	0	835	166	8	0
Standing Water	0	86	30	3	0
Unknown	0	14307	753	30	1
Wet	2	31954	15892	758	69



The majority of the accidents were registered in a Dry condition. Anyway, the probability of a serious accident are 0.21% (dry) against 0.14% (wet)

Light Conditions

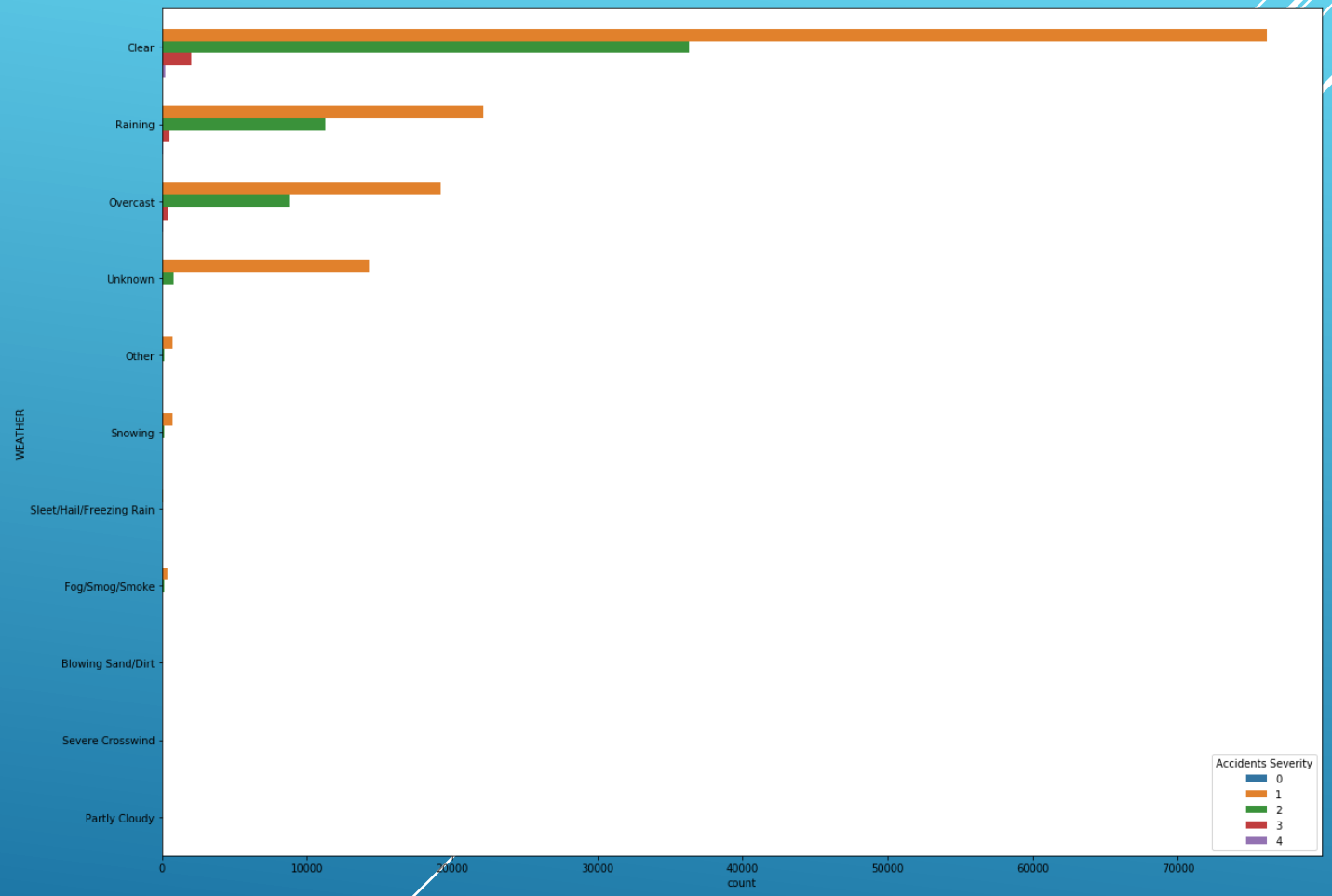
SEVERITYCODE LIGHTCOND	0	1	2	3	4
Dark - No Street Lights	0	1212	340	25	1
Dark - Street Lights Off	0	883	317	28	4
Dark - Street Lights On	0	34300	14602	1015	146
Dark - Unknown Lighting	0	16	8	0	0
Dawn	0	1706	837	60	5
Daylight	2	78475	39061	1785	168
Dusk	0	3990	1958	110	14
Other	0	185	54	4	0
Unknown	0	12881	609	25	1



To be noticed the number of accidents occurred in daily light and in dark with lights on, for severity = 3 and 4 the numbers are almost the same

Weather Conditions

SEVERITYCODE WEATHER	0	1	2	3	4
Blowing Sand/Dirt	0	40	15	0	0
Clear	1	76108	36317	2016	227
Fog/Smog/Smoke	0	383	188	3	3
Other	0	722	120	7	3
Overcast	0	19175	8847	446	53
Partly Cloudy	0	5	4	0	1
Raining	1	22151	11278	529	50
Severe Crosswind	0	18	7	0	1
Sleet/Hail/Freezing Rain	0	85	29	2	0
Snowing	0	734	169	10	0
Unknown	0	14227	812	39	1



Situation is slightly different for the weather conditions
Almost the same amount of accidents were registered for both severity 3 and 4 with overcast and raining weather conditions

1st Analysis – Accident severity prediction using all non-dropped features

- All the features were used to train the model (in exception of those already dropped)
- Random Forest Classifier and best Features for Random Forest algorithm
- Decision Tree Classifier and best Features for Decision Tree algorithm

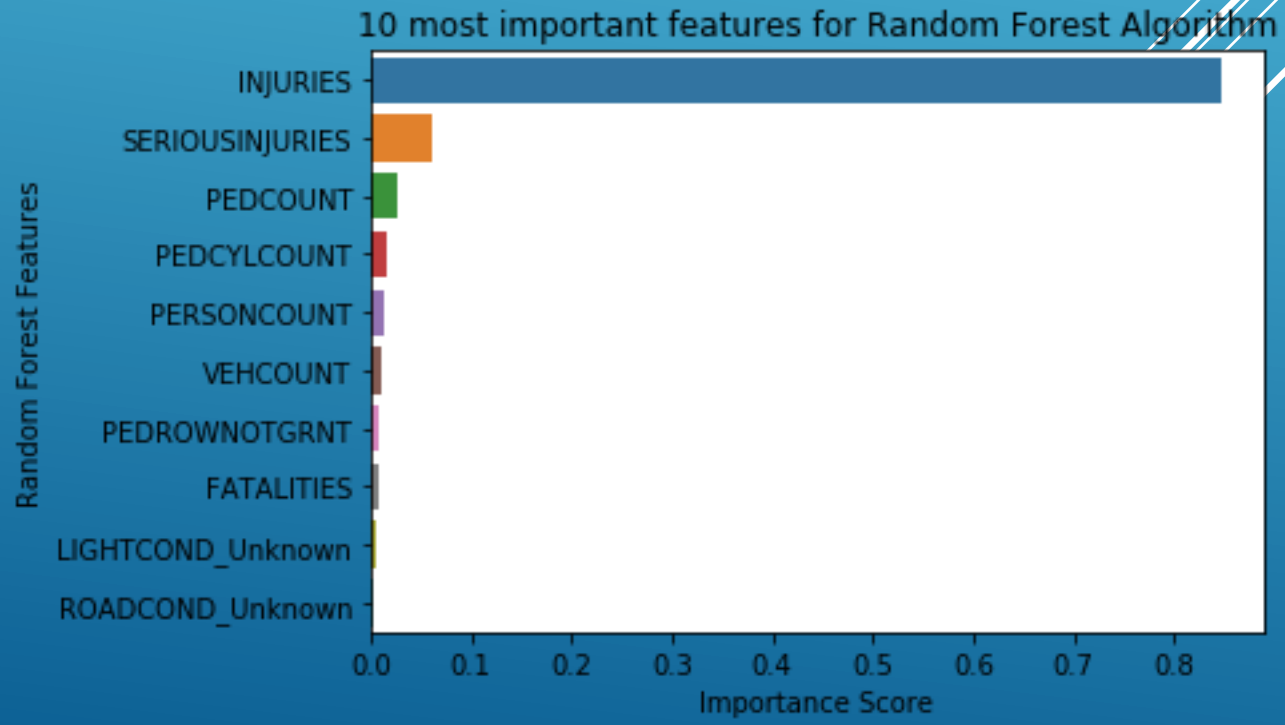
2nd Analysis – Accident severity prediction using weather, road and light conditions only

- Only the weather, road and light conditions were used
- Random Forest Classifier and Best Features for Random Forest algorithm
- Decision Tree Classifier and Best Features for Decision Tree algorithm

1st Analysis

Random Forest Algorithm

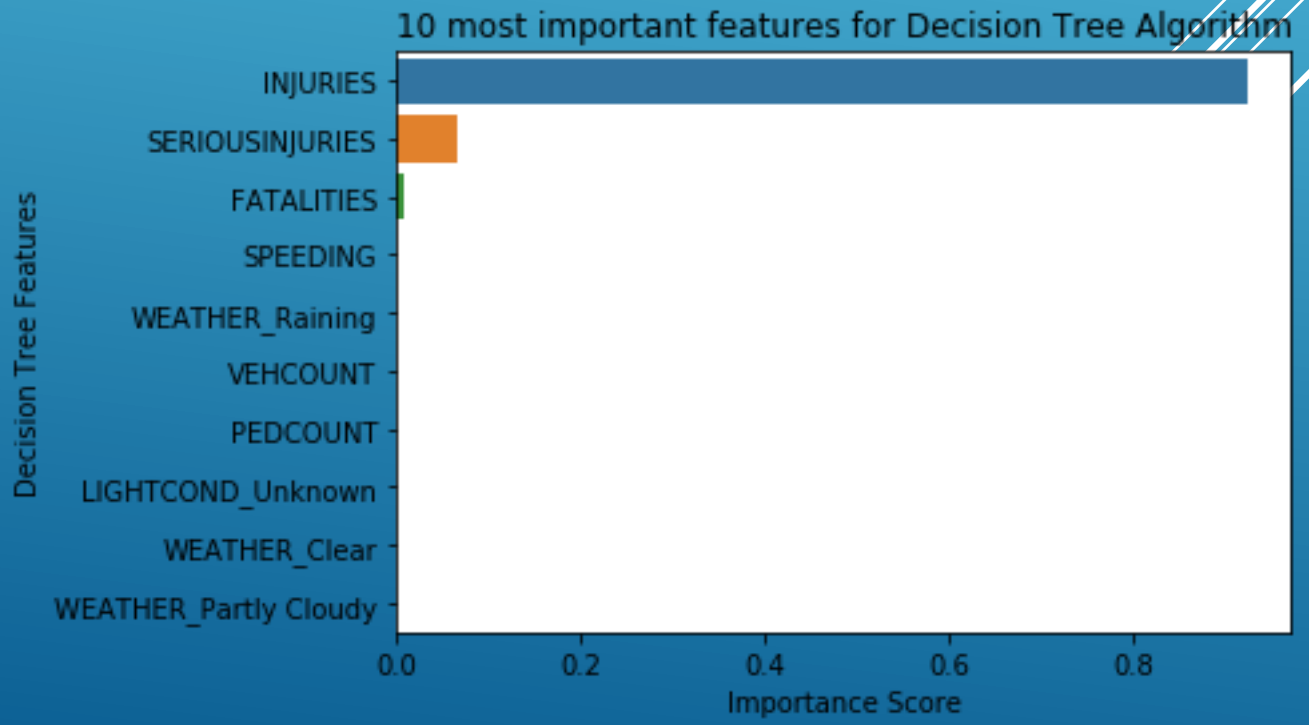
- Really accurate algorithm. Unfortunately it is slower than the Decision Tree algorithm.
- The accuracy scored was 0.9999828911
- The most important feature was Injuries one



1st Analysis

Decision Tree Algorithm

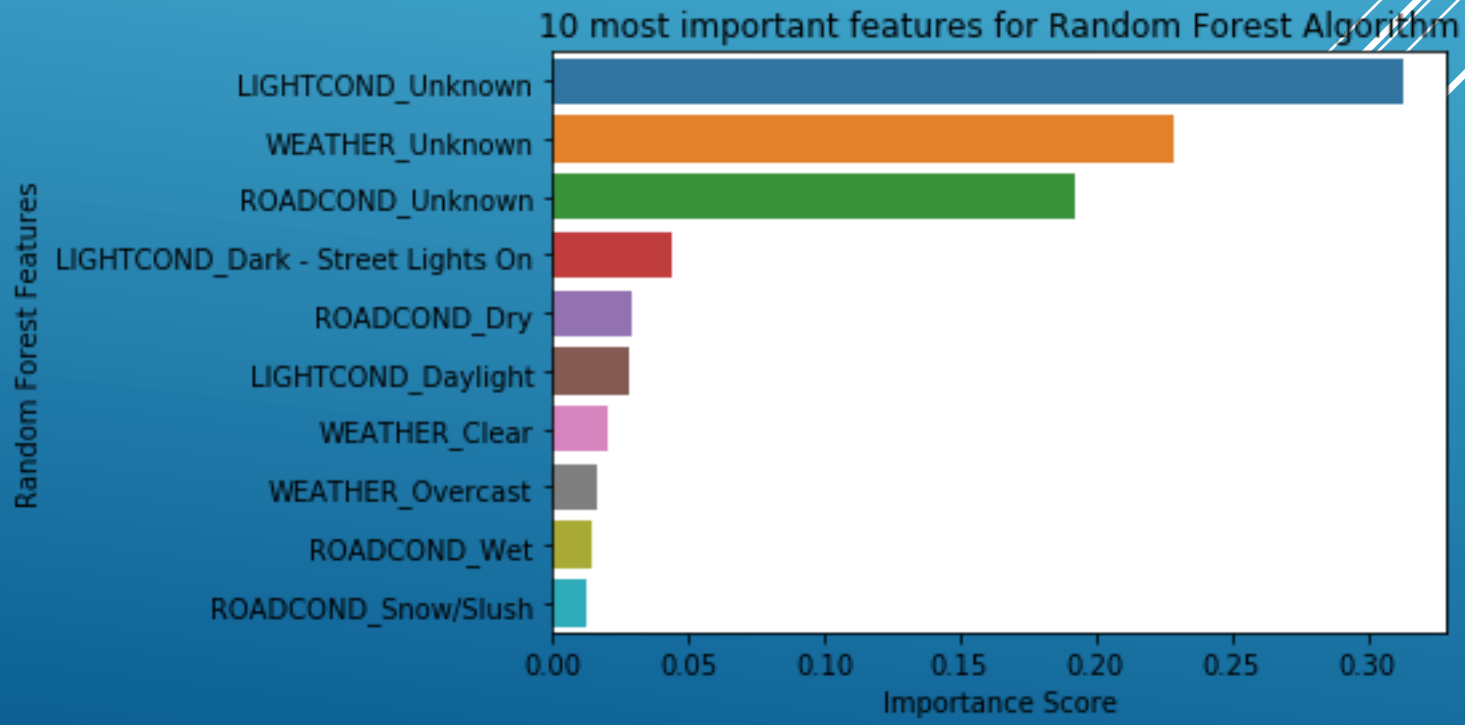
- Best accuracy scored: 1.0000000000
- Faster than Random Forest Algorithm
- Sensitive to overfitting, checked with cross_validation_score: 0.9999435494
- The most important feature was Injuries one



2nd Analysis

Random Forest Algorithm

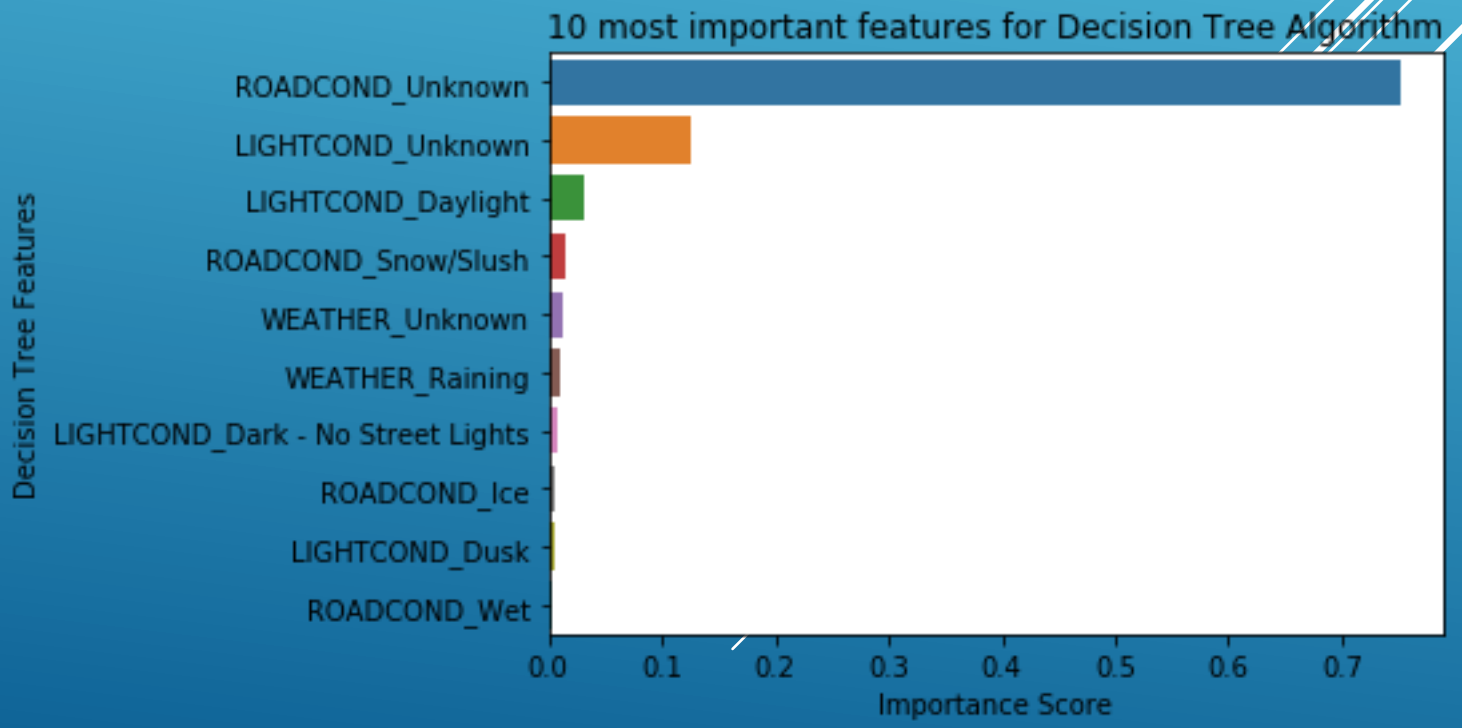
- Really accurate algorithm. Unfortunately it is slower than the Decision Tree algorithm.
- The accuracy scored in this case was: 0.6856062550
- The most important feature was the Lightcond_Unknown, followed by Weather_unknown and Lighcond_dark – Street Light On



2nd Analysis

Decision Tree Algorithm

- Registered again the best accuracy: 0.685857801
- Faster than Random Forest Algorithm
- The most important feature was Roadcon_unknown



Conclusions

- The classifications model I chose performed almost perfectly using all the variables, so, given the number of people injured, the model could predict with and extreme precision the severity of the accident
- Using only the weather, road and light conditions, the accuracy dropped significantly Despite this the model can be used anyway, giving acceptable results
- Many more data are required to have better results, such as:
  - Forecast Weather Data of the days in which the accidents occurred
  - Coordinates for each accident
  - Coordinate of each business location and their opening hours
  - Density of population