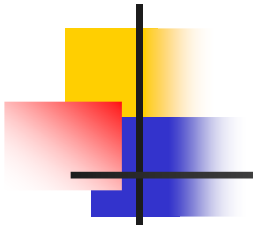


Data Science Course Capstone Project
Car Accident Severity Prediction - Introduction

- **The goal:** to build a predictor capable of predicting the severity of a road accident given traffic, weather and other environmental conditions. The severity is to be predicted in terms of a property damage only or with some type of bodily injury event in case of accident.
- **The purpose:** to help travelers to judge if the conditions they are currently encountering during their trip are a known factor relevant for serious consequences in case of accident or not.
- **The data:** accident dataset maintained by the Transportation Department of the City of Seattle, WA (https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions)
- **The methodology:**

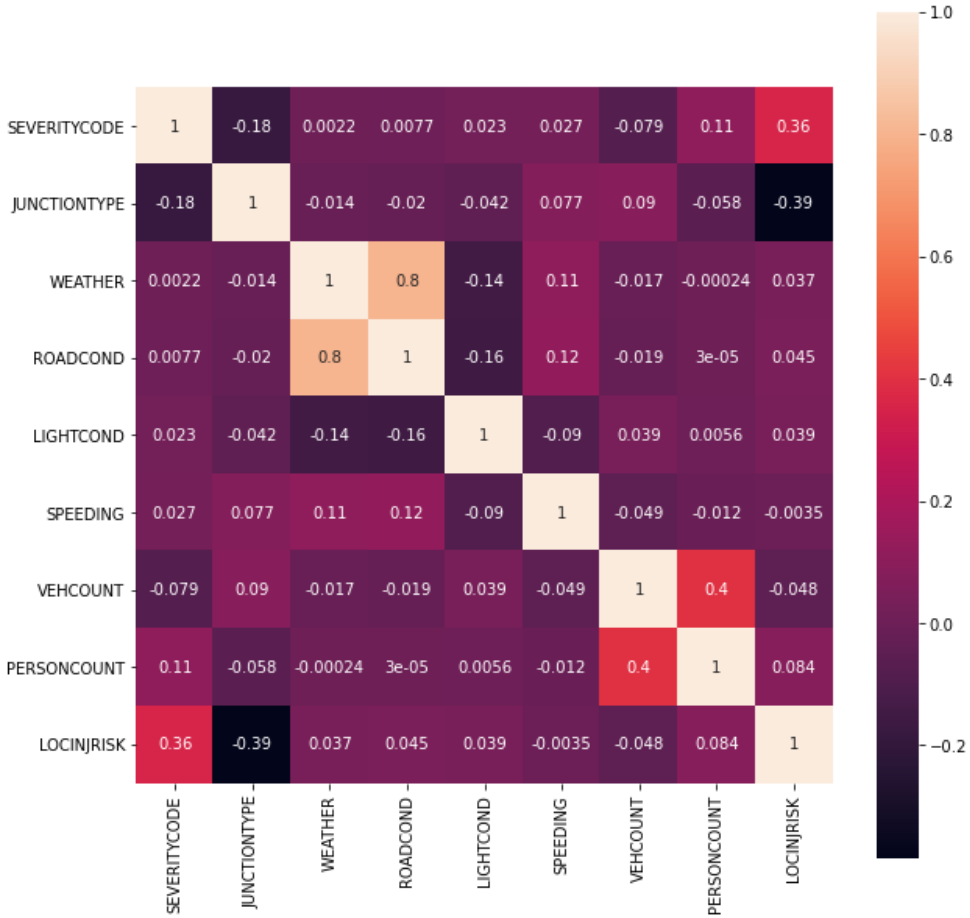
    1. Selection of the data features
    2. Preparation of the data and split into training and test sets
    3. Training of multiple predictors built with different techniques
    4. Testing of the built predictors using the test data set
    5. Evaluation of performances and selection of best predictor

Data Science Course Capstone Project
Car Accident Severity Prediction – Feature Selection

- **Target feature**: the "Severity Code" ("1" damage, "2" injury)
- **Selected features:** following an analysis of the 37 available data columns, the following features are selected:

Location	Road condition
Weather condition	Junction junction
Car Speeding	number of people involved
Light conditions	number of vehicles involved in

- **Heatmap**: despite appearing to be the most significant among the available ones, only few features appears to be significantly correlated among themselves and with the target feature



Data Science Course Capstone Project
Car Accident Severity Prediction – Data Preparation

- **Location**: the column location was transformed into a new feature "LocInjRisk" representing the probability of an injury in case of accident at that location, calculated based on the counts of damage and injury events at that location

	SEVERITYCODE	LOCATION	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND	SPEEDING	VEHCOUNT	PERSONCOUNT	LOCINJRISK
0	2	5TH AVE NE AND NE 103RD ST	At Intersection (intersection related)	Overcast	Wet	Daylight	0	2	2	0.483871
1	1	5TH AVE NE AND NE 103RD ST	At Intersection (intersection related)	Overcast	Dry	Dark - Street Lights On	0	3	4	0.483871
2	2	5TH AVE NE AND NE 103RD ST	At Intersection (intersection related)	Clear	Dry	Daylight	0	3	5	0.483871
3	2	5TH AVE NE AND NE 103RD ST	At Intersection (intersection related)	Overcast	Wet	Daylight	0	2	2	0.483871
4	1	5TH AVE NE AND NE 103RD ST	At Intersection (intersection related)	Overcast	Wet	Daylight	0	2	2	0.483871
...	...	...	...	...	...	...	...	...	...	...
169282	2	5TH AVE W BETWEEN W FULTON ST AND W BARRETT S ST	Mid-Block (not related to intersection)	Clear	Dry	Dusk	0	2	2	0.000000
169283	1	SW OTHELLO ST BETWEEN 29TH AVE SW AND 30TH AVE SW	Mid-Block (not related to intersection)	Raining	Wet	Dark - Street Lights On	0	2	2	0.000000
169284	2	5TH AVE S BETWEEN S LUCILE ST AND S FINDLAY ST	Driveway Junction	Clear	Dry	Daylight	0	2	2	0.000000
169285	1	NE PARK RD AND NE RAVENNA WB BV	At Intersection (intersection related)	Clear	Dry	Daylight	0	2	2	0.000000
169286	1	PUGET BLVD SW BETWEEN SW HUDSON ST AND DEAD END 1	Mid-Block (not related to intersection)	Raining	Wet	Dark - Street Lights On	0	1	1	0.000000

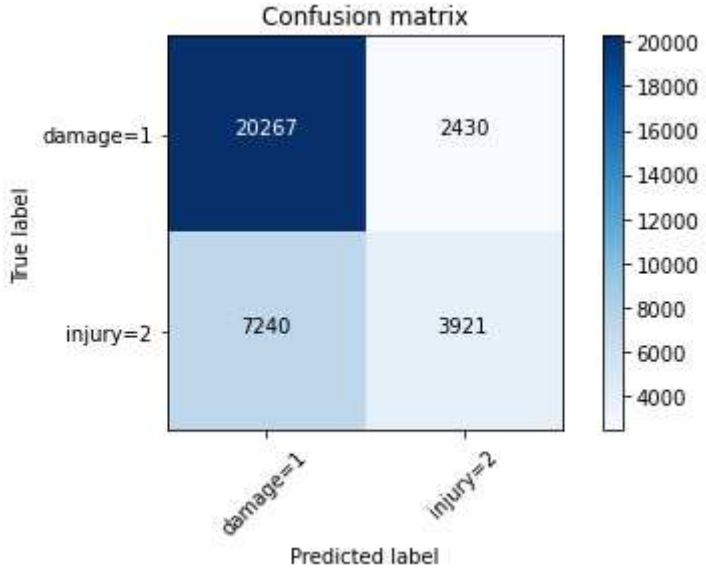
169287 rows × 10 columns

- **Additional Preparations**: transformation of the the categorical features "JunctionType", "Weather", "RoadCond" and "LightCond" into numerical values corresponding to their various categories possible
- **Training vs. Test Data Sets:** randomly separated 135429 records for training and 33858 records for testing purposes

Data Science Course Capstone Project
Car Accident Severity Prediction – Multiple Predictors 1/4

- **Logistic Regression**

```
LogisticRegression(C=0.01, solver='liblinear')
```

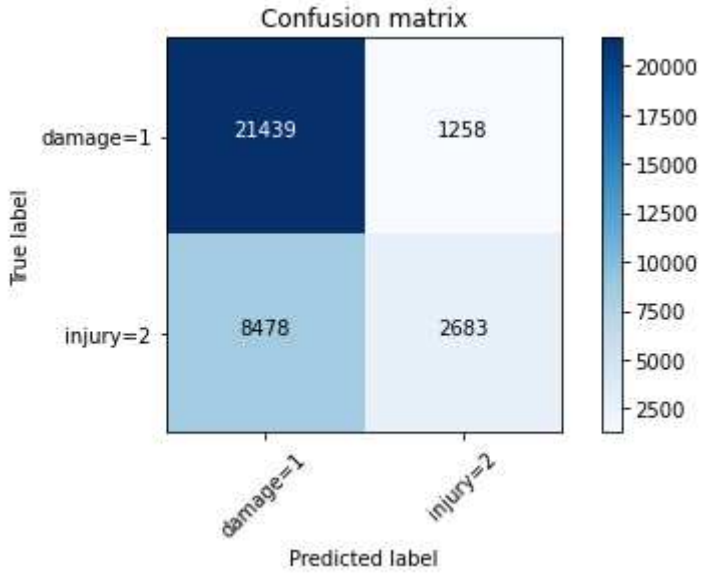


	precision	recall	f1-score	support
1	0.74	0.89	0.81	22697
2	0.62	0.35	0.45	11161
accuracy			0.71	33858
macro avg	0.68	0.62	0.63	33858
weighted avg	0.70	0.71	0.69	33858

Data Science Course Capstone Project
Car Accident Severity Prediction – Multiple Predictors 2/4

- **Support Vector Machine**

```
SVC(kernel='linear', probability=True)
```

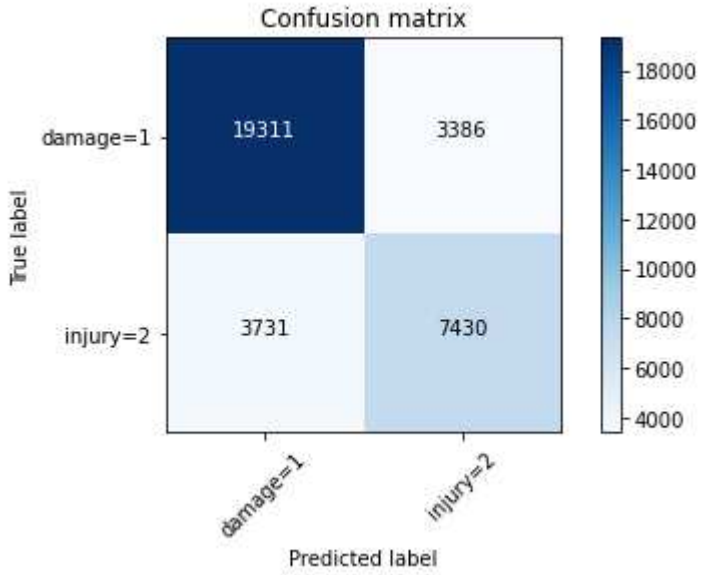


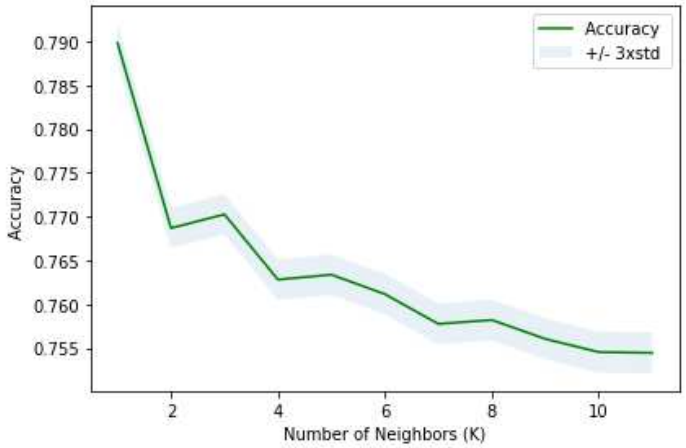
	precision	recall	f1-score	support
1	0.72	0.94	0.81	22697
2	0.68	0.24	0.36	11161
accuracy			0.71	33858
macro avg	0.70	0.59	0.59	33858
weighted avg	0.70	0.71	0.66	33858

Data Science Course Capstone Project
Car Accident Severity Prediction – Multiple Predictors 3/4

- **K-Nearest Neighbor**

```
KNeighborsClassifier(n_neighbors=1)
```





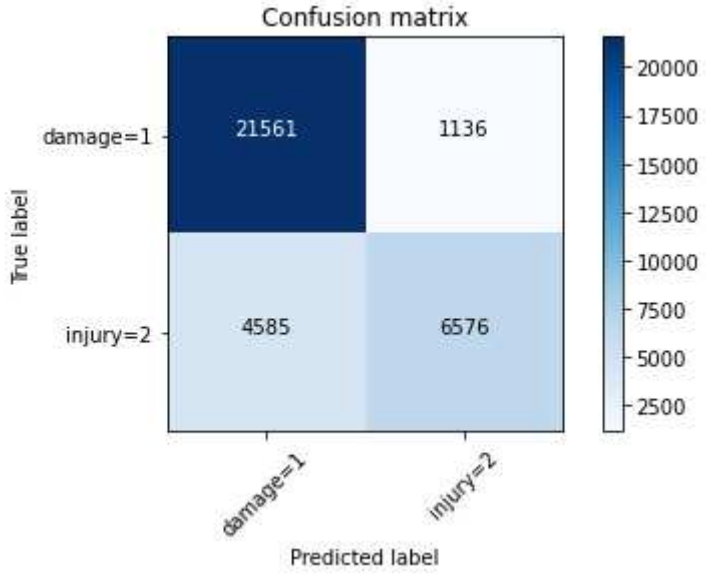

```
Best Accuracy is 0.7897985705003249 for k = 1
```

	precision	recall	f1-score	support
1	0.84	0.85	0.84	22697
2	0.69	0.67	0.68	11161
accuracy			0.79	33858
macro avg	0.76	0.76	0.76	33858
weighted avg	0.79	0.79	0.79	33858

Data Science Course Capstone Project
Car Accident Severity Prediction – Multiple Predictors 4/4

- **Decision Tree**

```
DecisionTreeClassifier(criterion='entropy', max_depth=30)
```



	precision	recall	f1-score	support
1	0.82	0.95	0.88	22697
2	0.85	0.59	0.70	11161
accuracy			0.83	33858
macro avg	0.84	0.77	0.79	33858
weighted avg	0.83	0.83	0.82	33858

Data Science Course Capstone Project
Car Accident Severity Prediction – Performance Summary

Predictor Type	Weighed Avg. Precision	Weighed Avg. Recall	Weighed Avg. F1 Score	Notes
Logistic Regression	0.70	0.71	0.69	
Support Vector Regression	0.70	0.71	0.66	
K-Nearest Neighbor	0.79	0.79	0.79	
Decision Tree	0.83	0.83	0.82	Best performer!