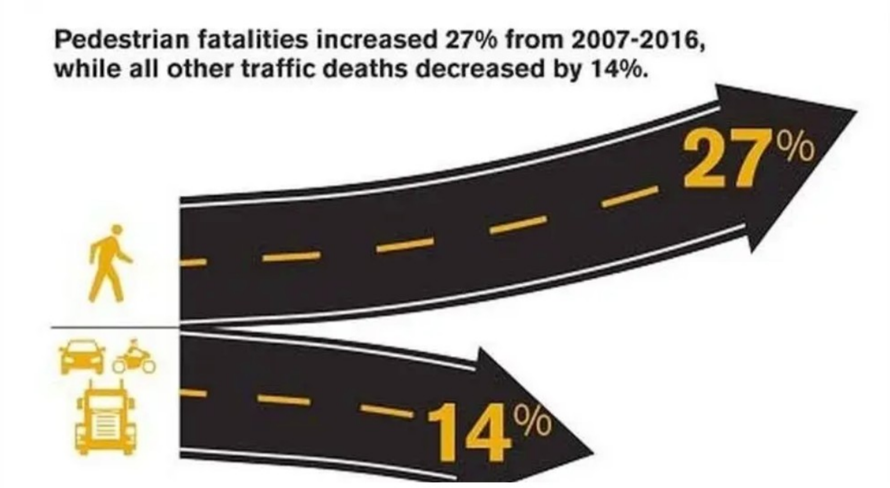
IBM Data Science Capstone

Guidance for pedestrians in Seattle

Jan Stroppel - 10/2020

Current status

- National Highway Traffic Safety Administraion (NHTSA) sees an increase in the number of accidents
- Accidents with pedestrians often leads to severe injuries

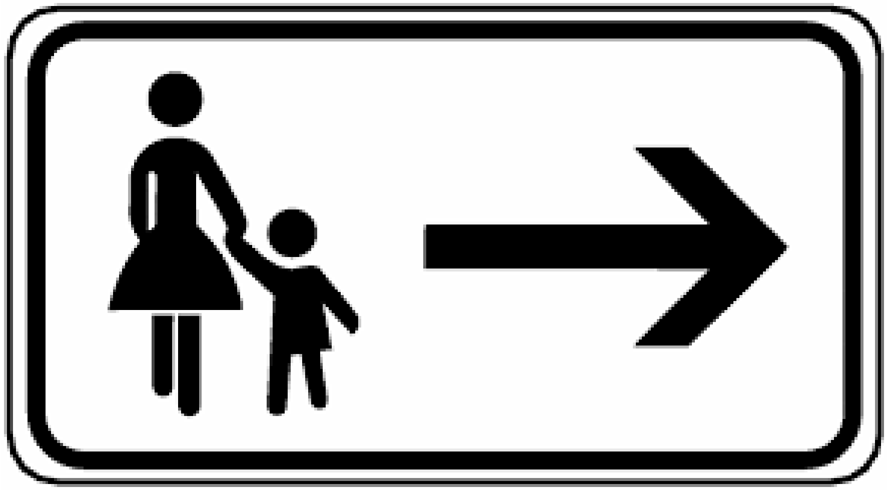


Pedestrian fatalities increased 27% from 2007-2016, while all other traffic deaths decreased by 14%.

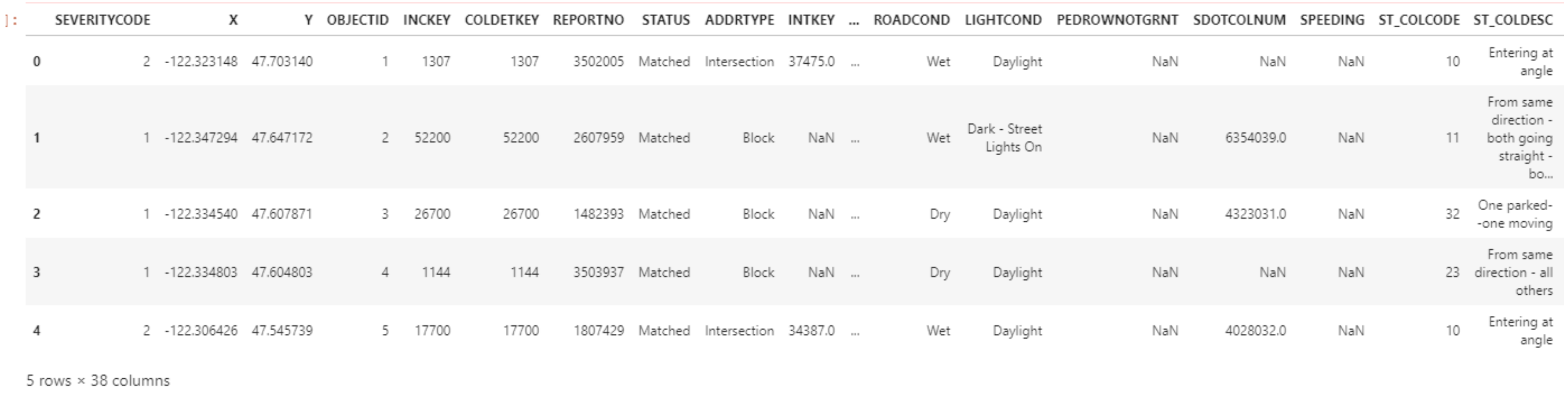
Bild: NHTSA

Business Problem

- Create a model to predict the probability of pedestrians involed in an accident
- Find parameters which indicates a high risk for pedestrians
- Enable authorities in Seattle to create a guidance for pedestrians



Data



]:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCOLNUM	SPEEDING	ST_COLCODE	ST_COLDESC
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	NaN	NaN	NaN	10	Entering at angle
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	NaN	6354039.0	NaN	11	From same direction - both going straight - bo...
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	NaN	4323031.0	NaN	32	One parked--one moving
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	NaN	NaN	NaN	23	From same direction - all others
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	NaN	4028032.0	NaN	10	Entering at angle

5 rows × 38 columns

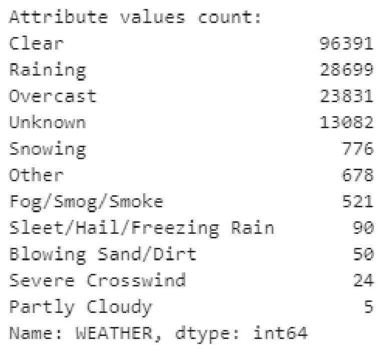
The dataset „Collisions – All Years“ for Seattle was used for creating, training and testing the model

Selected Columns

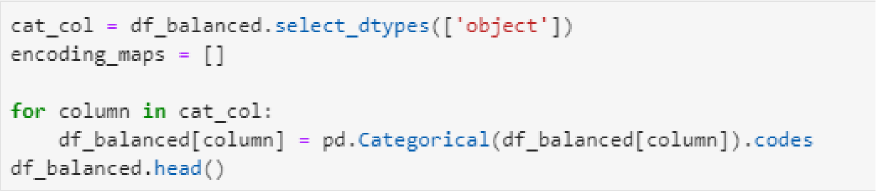
- Speeding: Was the vehicle too fast?
- Weather: Was it raining?
- Road Condition: Was the road slippy?
- Light condition: Was the pedestrian visible?
- Date and Time: Are there differences for week days and the time?
- Inattention: Was the driver distracted?
- Drugs: Was the driver under influence of alkohol or drugs?
- Dependent variable was PEDESTRIANINVOLVED, which is a boolean derived from the collision type

Data Preparation

- Unify entries for booleans (N,Y instead of 0,1)
- Setting meaningful default values (Unknown for missing data)
- Transforming categorical values into numerical values (0,1 instead of Dry, Wet)
- Balancing data set (same number of rows for involved/not involved pedestrians)

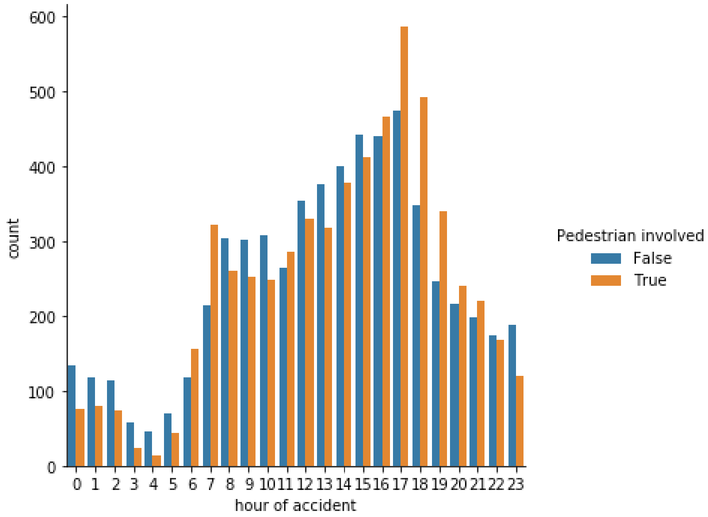


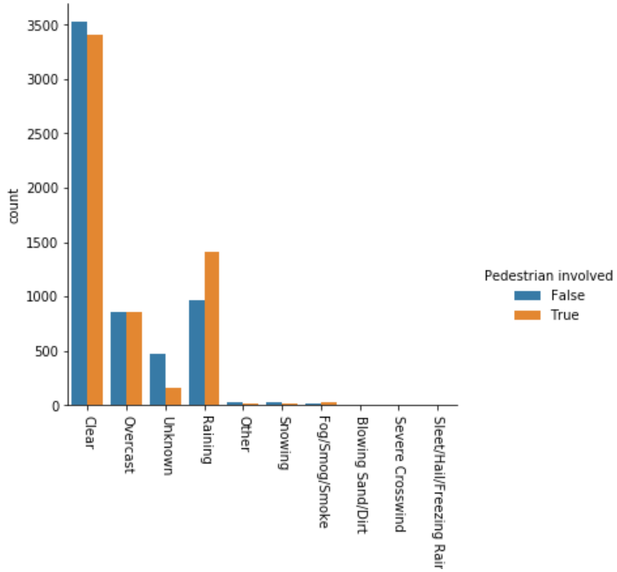
```
Attribute values count:
Clear                      96391
Raining                    28699
Overcast                   23831
Unknown                    13082
Snowing                      776
Other                        678
Fog/Smog/Smoke               521
Sleet/Hail/Freezing Rain      90
Blowing Sand/Dirt             50
Severe Crosswind              24
Partly Cloudy                  5
Name: WEATHER, dtype: int64
```

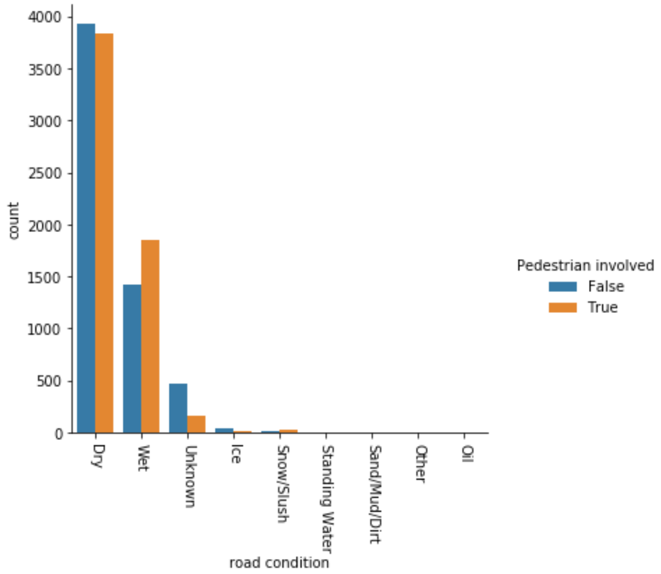


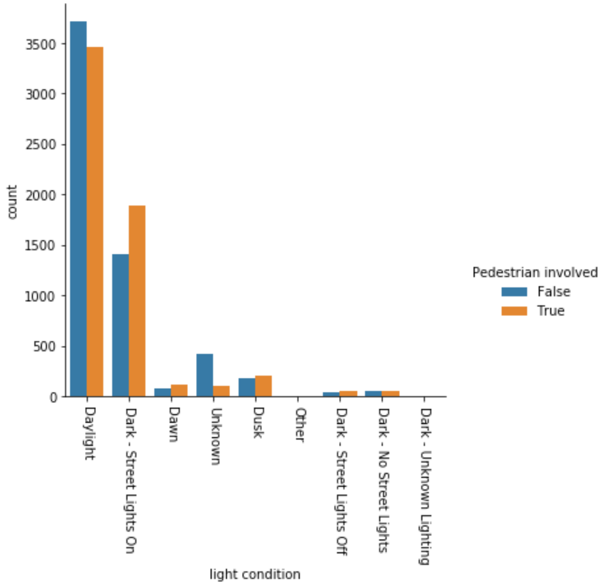
```
cat_col = df_balanced.select_dtypes(['object'])
encoding_maps = []

for column in cat_col:
    df_balanced[column] = pd.Categorical(df_balanced[column]).codes
df_balanced.head()
```

Data visualization





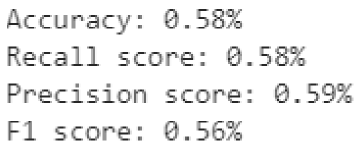




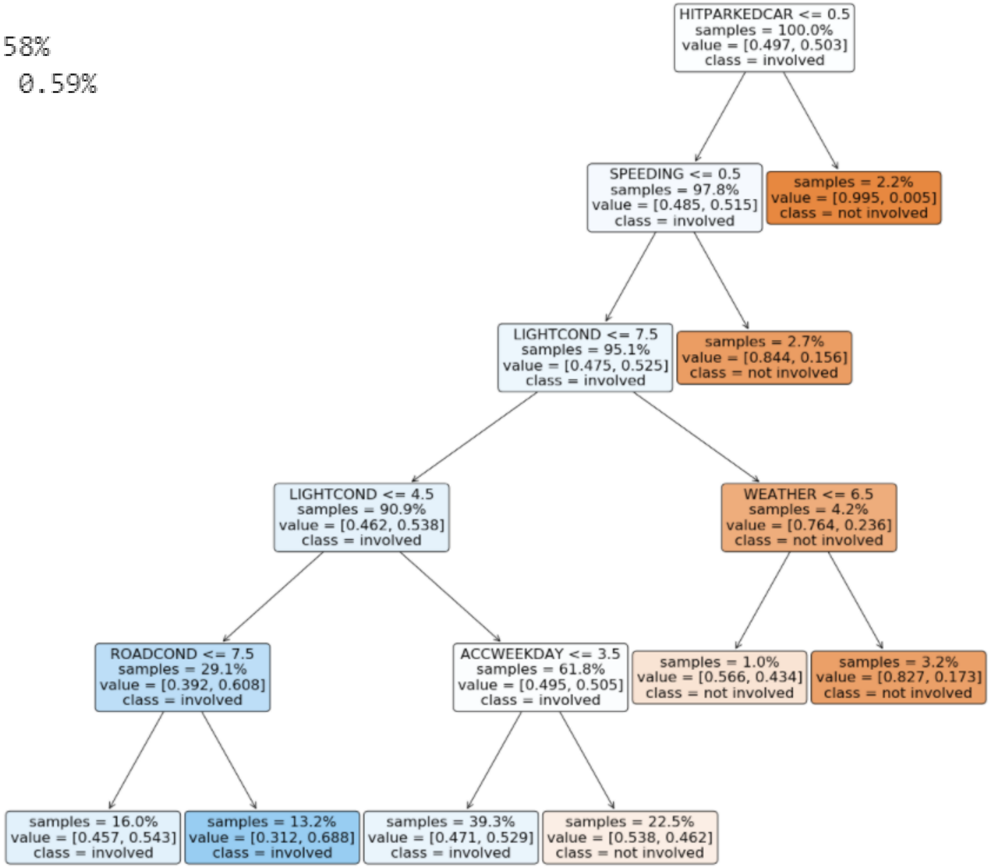
Methodology

- As most variables are categorical, a classifier model was chosen
- It was decided to create a decision tree
- The dependent variable is binary, so ideal for a decision tree
- The dataset was split 30/70 in training/test set

Decision Tree



```
Accuracy: 0.58%
Recall score: 0.58%
Precision score: 0.59%
F1 score: 0.56%
```



Evaluation

- Model performance low (50 to 60%)
- Reason could be not enough data to train and test the model (only 5891 entries for each value)
- Another reason could be that business problem can not be handled by the data set