Guidance for Pedestrians in Seattle

Jan Stroppel

09.10.2020

1. Introduction

1.1 Background

The National Highway Traffic Safety Administration (NHTSA) has found out, that from 2007 to 2016 the traffic accidents in the United States went down by 14 %. Unfortunately in the same time the pedestrian fatality went up by 27 %. Modern traffic planning should therefore invest more money in pedestrian friendly concepts. Also, the local authorities should identify situations in which pedestrians have a high risk on being involved in accidents to create a guidance for pedestrians, warning them about potential risks.

1.2 Problem

With accurate data we could identify situations where pedestrians have a higher risk, like for example in the night, when they are not seen well by a driver.

1.3 Interest

Local authorities would be interested in knowing which circumstances lead to a higher risk for pedestrians. They could create a guidance for pedestrians as well as consider it in future projects.

2. Data acquisition and cleaning

2.1 Data sources

The data set "Data Collision – All Years" was used for this project. It includes the statistics of accidents in Seattle for the last years. It has information about if a pedestrian was involved in the accident as well as data about the conditions that lead to the accident, like weather or light conditions.

2.2 Data cleaning

First the description of every property was read and it was decided which values were used for the prediction. As a second step, the values for the accidents for this properties were evaluated which lead to the following cleaning activities:

INCDTTM: This property contains the datetime of the accident. In our modell we are only interested in the day of the week, when an accident occurs, and the time of the day. Therefore this column was split into two columns – **ACCWEEKDAY** and **ACCHOUR** – to be able to use them in the model. The column **ACCHOUR** also had a significant too high amount of the time 00:00, which lead to the assumption that this was a default value if the time was unknown. For this reason accidents with this time were removed from the dataset.

WEATHER, ROADCOND, LIGHTCOND: This columns had null values for some accidents. This null values were exchanged with the value "Unknown".

INATTENTIONIND, SPEEDING: Only the value Y was set in this columns, in case of a negation it was kept empty. Therefore the null values were exchanged with a N.

UNDERINFL: In this column the values were mixed up, one part was filled with a Y or N to as boolean parameter, another part was filled wit 0 and 1. It was obvious that 0 was used for N and 1 for Y because the relation of the given values 0 to 1 and N to Y were nearly identical, making this assignment easy. Every 0 was exchanged with an N and every 1 with a Y.

PEDESTRIANINVOLVED: This variable was created as a boolean variable. In case the COLLISIONTYPE variable was Pedestrian it was set to true, otherwise to false. The dataset contained 158256 accidents with no pedestrian involved and just 5891 accidents with pedestrians involved. To balance the dataset 5891 accidents without pedestrians involved were selected so that the complete dataset consists of 11782 accidents.

Some columns consists of categorical values, like the weather conditions. To better handle this data in a statistical model they were transformed into numerical values.

There was one property which has a direct connection to involved pedestrians, PEDROWNOTGRNT. It is a boolean value that states if the right for the pedestrian to walk was granted. As this property is only set for accidents with pedestrians it has no value for our prediction and was therefore ignored.

2.3 Feature selection

Only a few columns were considered to have an influence on the fact if a pedestrian is involved in an accident. The following columns were selected:

:

	PEDESTRIANINVOLVED	SPEEDING	WEATHER	ROADCOND	LIGHTCOND	INATTENTIONIND	UNDERINFL	HITPARKEDCAR	ACCWEEKDAY	ACCHOUR
10649	False	0	1	0	5	0	0	0	4	14
127198	False	1	4	8	2	0	0	0	2	18
136695	False	0	9	7	2	0	0	0	2	17
114776	False	0	1	0	5	0	0	0	3	13
187836	False	0	1	0	5	0	1	0	1	14

Speeding: It was suggested that the speed of the vehicle might have influence on the question if a pedestrian is involved in the accident. Speeding is a boolean value.

Weather: The suggestion is that under some weather conditions the risks for pedestrians might be higher, for example, if it is raining, pedestrians might run more often over the street to get into a dry place. Also, if it is raining, drivers might see pedestrians later and the way to stop the car might be longer.

Road conditions: Also the road conditions could have influence on the risk for pedestrians. The risk might be higher if the road is slippy and cars need a longer way to stop.

Light conditions: There are also some light conditions that might result in a higher risk, for example, if it is dark and the pedestrian is wearing black clothes it might be hard to see a pedestrian.

Under influence of drugs/alcohol: Drugs could also lead to a higher risk for pedestrian. A drunken driver might ignore a red traffic light and hit a pedestrian.

Inattention of the driver: Same might happen if the driver is inattentional, maybe using his smart phone and don't see the red traffic light.

Hit parked car: If a parked car is hit in an accident could also increase the risk for pedesitrians, as cars often park near the sidewalk.

Day of week: The question was if for example the risk for pedestrians is higher on weekends, because people like to make a walk in their spare time.

Hour of the day: Is the risk higher at night when it is dark and the driver might be tired?

3. Predictive Modeling

There are two types of models, regression and classification, that can be used to predict the risk for pedestrians. For our case a classification model was chosen, the reason for it id that the dependent variable, PEDESTRIANINVOLVE, is a categorical binary value. For the prediction of a categorical value classification models fit perfectly. The underlying algorithms are similar between regression and classification models, but different audience might prefer one over the other.

3.1 Classification models

I decided using a decision tree to predict the risk for a pedestrian for the following reasons:

Good Visualization: The algorithm is simple to understand and the visualization is easy readable by human beings.

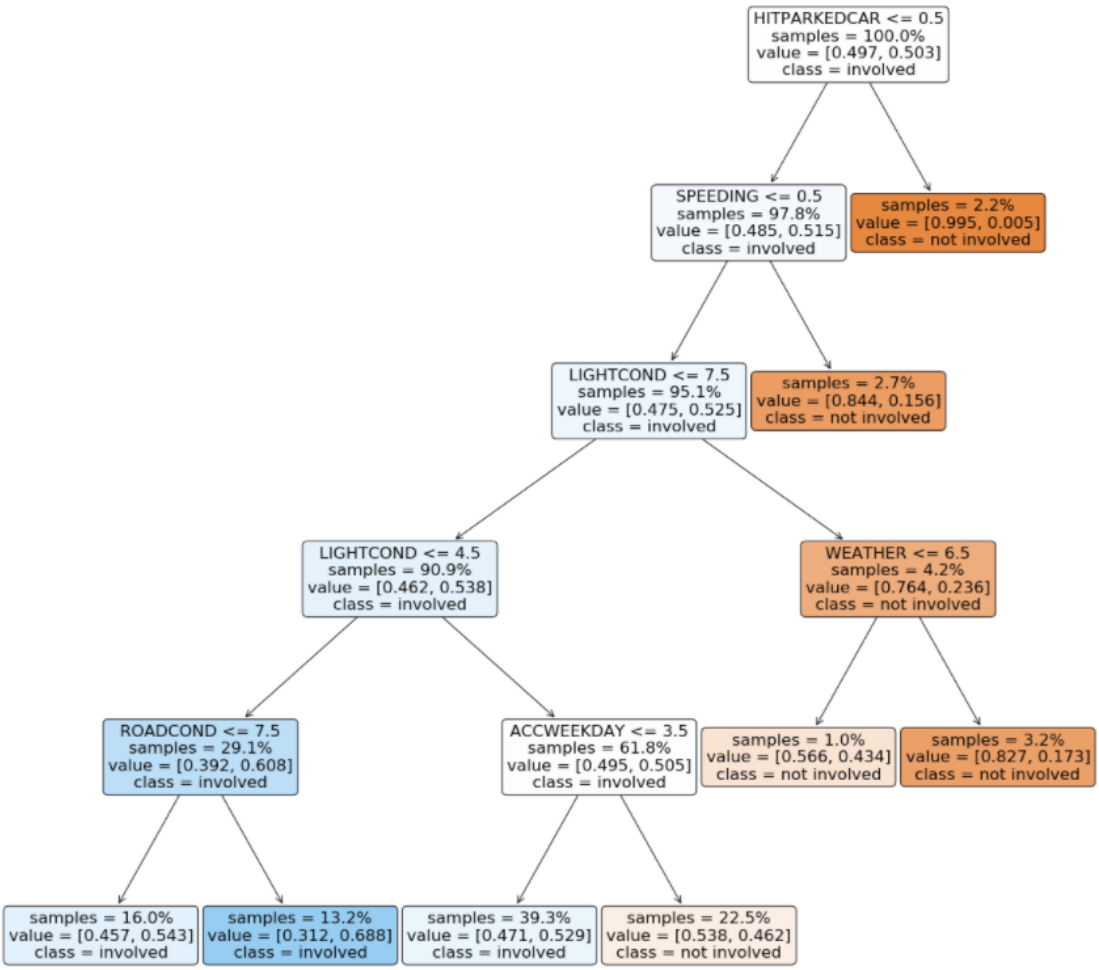
Categorical variables: Most of the variables are categorical, which makes a classification model and especially a decision tree fitting perfectly.

The dataset was divided into a training and a test set (70/30). With the training set the best parameters for the decision tree was calculated, the result was tested with the test set.

```
test_model = DecisionTreeClassifier(criterion=grid_search.best_params_.get('criterion'),
                                    max_depth=grid_search.best_params_.get('max_depth'),
                                    max_leaf_nodes=grid_search.best_params_.get('max_leaf_nodes'),
                                    min_samples_leaf=grid_search.best_params_.get('min_samples_leaf'))
test_model.fit(X_trainset,Y_trainset)

yh = test_model.predict(X_testset)

from sklearn import metrics

accuracy = metrics.accuracy_score(Y_testset, yh)
recall = metrics.recall_score(Y_testset, yh, average='weighted')
precision = metrics.precision_score(Y_testset, yh, average='weighted')
f1 = metrics.f1_score(Y_testset, yh, average='weighted')

print('Accuracy: ' + str(accuracy.round(2)) + '%')
print('Recall score: ' + str(recall.round(2)) + '%')
print('Precision score: ' + str(precision.round(2)) + '%')
print('F1 score: ' + str(f1.round(2)) + '%')
```

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

Unfortunately the result is not satisfactory. The statistical values all reside between 50 – 60 %, a good fit would be over 70 %. Let’s have a look on the resulting decision tree:



4. Conclusions

The resulting decision tree does not perform very well. The reason might be that the dataset (11782) was not big enough as most of the accidents do not include pedestrians. Another reaon might be that it is hard to predict if a pedestrian is involved in regards to the properties of thze dataset. Interesting is the fact that hitting a parked car and a too fast car seem to result no involved pedestrian. Also bad light conditions more often leads to pedestrians involv ed in an accident.

5. Future directions

In the future it might be evaluated if there are other factors that might lead to ahigher risk for pedestrians, and this should be included in the report that leads to the dataset.