Introduction

PROBLEM DESCRIPTION

The main business problem is about collisions in Seattle and conditions causing them. Intention is to build predictive model which will provide reasonable prediction if under certain conditions is higher probability of collision with certain severity (certain places and on exact days) and which conditions are the mostly involved in collisions.

Audience consists of people who lives in Seattle or travel through it and goal is to provide them information about current situation on a roads and possible dangers.

Audience should be interested in this problem because knowing the relationships between conditions and likelihood of collision can save their money and life.

DATA DESCRIPTION

I've used data set provided in Week 1 about collisions in Seattle (Data-Collisions.csv). Data set has 194673 rows and 38 columns. After checking the data there is a need to clean them and select main features used in the further analysis. Cleaning of data mainly consists of:
- unification of values in columns (some used Y,N,1,0) - UNDERINFL
- binary codification of Y/N values into 1/0
- convert values in column to same data type (int)
- convert INCDATE from str into Date type
- drop NaN and empty rows in the selected features which we are going to use
- removed "Unknown" values from dataset

I dropped following columns because they are duplicated or irrelevant for my analysis:
- SEVERITYDESC (long text)
- SEVERITYCODE.1 (duplicate)
- SDOTCOLNUM (A number given to the collision by SDOT)
- EXCEPTRSNCODE
- ST_COLDESC (A detailed description of the severity of the collision)
- STATUS
- SPEEDING (only one value Category)
- INATTENTIONIND (only one value Category)
- PEDROWNOTGRNT (only one value Category)
- INCKEY
- COLDETKEY
- OBJECTID
- INCDTTM (similar to INCDATE but with concrete hours - missing hour attribute in 25% of rows)

I've selected following features (independent values) for prediction of dependent value SEVERITYCODE:
- ADDRTYPE: Collision address type
- LIGHTCOND: The light conditions during the collision.
- WEATHER: A description of the weather conditions during the time of the collision.
- ROADCOND: The condition of the road during the collision.

Numerical data (counts):
- PERSONCOUNT
- VEHCOUNT
- PEDCYLCOUNT

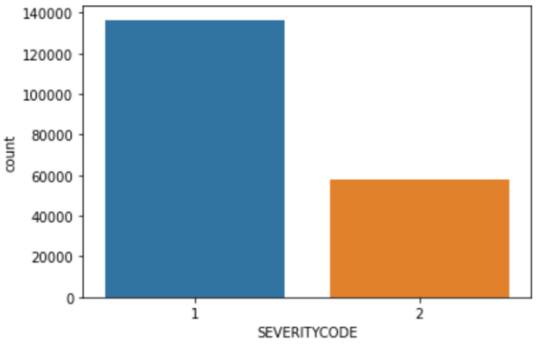
Numerical data (keys):
- SEGLANEKEY
- CROSSWALKKEY
- INTKEY

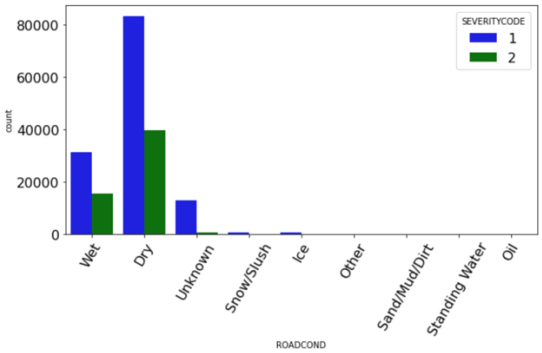
Space data (coordinates):
- X
- Y

By utilising independent values in machine learning models I determined under which conditions (WEATHER, ROADCOND, LIGHTCOND, ADDRTYPE, ADDRESS) it's most likely to predict certain SEVERITYCODE in concrete places.

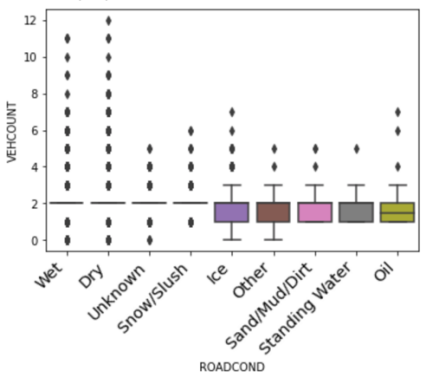
METHODOLOGY

I've done several Exploratory analyses, mainly consisting of:
- check data types with df.dtypes
- check value counts of each column in order to find NaN/Unknown values/corrupted values and accordingly adjusted them (e.g. filtered and used only LOCATIONs where number of Collisions
are higher than 10)
- use of plots such as bar charts for visualising counts of data distribution





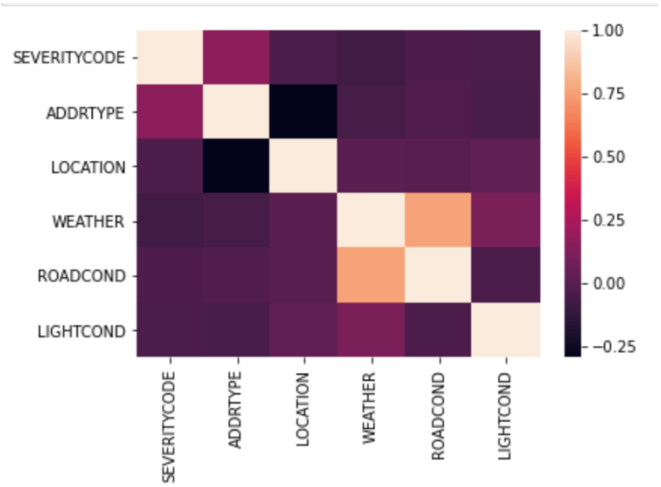
- use of boxplot for detecting outliers of ROAD conditions



- use of pd.crosstab(normalize=True) for finding ratios between WEATHER and SEVERITYCODE

SEVERITYCODE	1	2
WEATHER		
Blowing Sand/Dirt	0.000177	0.000077
Clear	0.429214	0.209722
Fog/Smog/Smoke	0.002151	0.001076
Other	0.001023	0.000461
Overcast	0.107367	0.050855
Partly Cloudy	0.000012	0.000018
Raining	0.126742	0.065431
Severe Crosswind	0.000106	0.000041
Sleet/Hail/Freezing Rain	0.000491	0.000160
Snowing	0.003907	0.000969

- use LabelEncoder for creation of correlation heatmap which helps to find which features are interrelated (obviously WEATHER conditions are influencing ROAD conditions)



- use of dummified features in order to make classification tree model

Machine learning algorithms used are for solving classification problems:
- DecisionTreeClassifier: Decision tree construction does not involve any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle multidimensional data. During tree construction, attribute selection measures are used to select the attribute that best partitions the tuples into distinct classes.
When decision trees are built, many of the branches may reflect noise or outliers in the training data. max_depth was set on 6.
Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data.
- Logistics regression: It can be used for Classification as well as for Regression problems, but mainly used for Classification problems.
In Logistic regression, it is not required to have the linear relationship between the dependent and independent variable. Logistic Regression uses Sigmoid function.

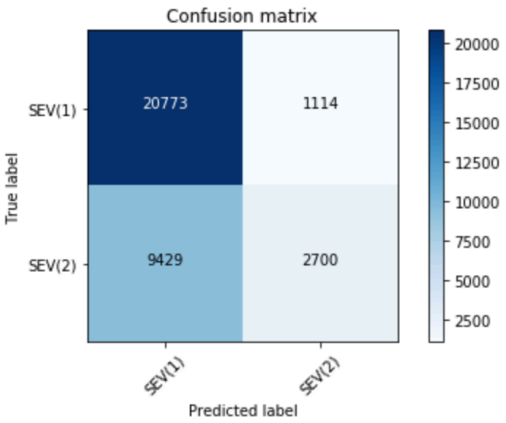
Data set was divided into training (0,7) and testing (0,3) set.

RESULTS

Decision tree accuracy is 0.69 and Confusion matrix is following:

	precision	recall	f1-score	support
1	0.69	0.95	0.80	21887
2	0.71	0.22	0.34	12129
accuracy			0.69	34016
macro avg	0.70	0.59	0.57	34016
weighted avg	0.69	0.69	0.63	34016

```
Confusion matrix, without normalization
[[20773  1114]
 [ 9429  2700]]
```



As we can see Precision for SEV 2 is higher than for SEV 1 but in the recall context we see the model is better for predicting SEV 1 than SEV 2. Weighted avg F1-score is 0.63

Logistics regression has Jaccard index: 0.64, F1-score: 0.60 and LogLoss: 0.65.

DISCUSSION

According to data provided we can predict with models what value will severity code have based on the model. It will be valuable to have complete dataset and minimise NaN values and also in some cases e.g. INATTENTIONIND, SPEEDING there is only one binary classifying value and it's impossible to deduct what the missing values are. These features will be useful for analysis of Severity of collision.

CONCLUSION

During the analysis we have observed several interesting relationships in data and this kind of classification problem is good to predict with a given accuracy because of minimal differences in Severity code 1 and Severity code 2 in reality. Generally, there is higher probability of SEVERITYCODE 1 collision, and they are mainly happening during daylight on a dry road, followed by dark - street lights on a wet road. The ratio between SEVERITYCODE 1 and SEVERITYCODE 2 is the lowest in the case of Dusk and Dawn light conditions.