Report «Collision data from the city Seattle”

Introduction

The traffic management division of Seattle is recording traffic incidents since 2004. These include all types of collisions and environmental data as well. The business interest of such data is to understand in which conditions and locations accidents occur. This information may be useful for car manufacturers to adjust and extend their accident warning systems, insurances, police and rescue corps, or for the city counsel of Seattle to put new traffic laws into place at hotspot zones.

We want to understand what are the most common factors that led to injuries and hence what we can learn for the future about it. The severity code in the dataset comprises the following classifications

- 3=Fatality
- 2b=Serious injury
- 2=Injury
- 1=Proper damage
- 0=Unknown

For example, is riding a bike dangerous? Are the weather conditions contributing? Are they mostly happening in Alleys, Blocks, or Intersections?

We also want to build a model that could predict the potential severity of an accident given certain conditions (e.g., bad weather, bad light). For that we need to build a robust model. This model could be used in various ways, ambulance with no information about the accident etc.

Data

The data were recorded by Traffic Records, Seattle police department in collaboration with SDOT Traffic Management division in Seattle.

The data are weekly updated and includes details about the accident (e.g., collision description and code), the environment (e.g., weather, location) and the severity of injuries caused and the number of involved people. There are 37 variables describing the accident. There are currently 194’673 recorded accidents in the system.

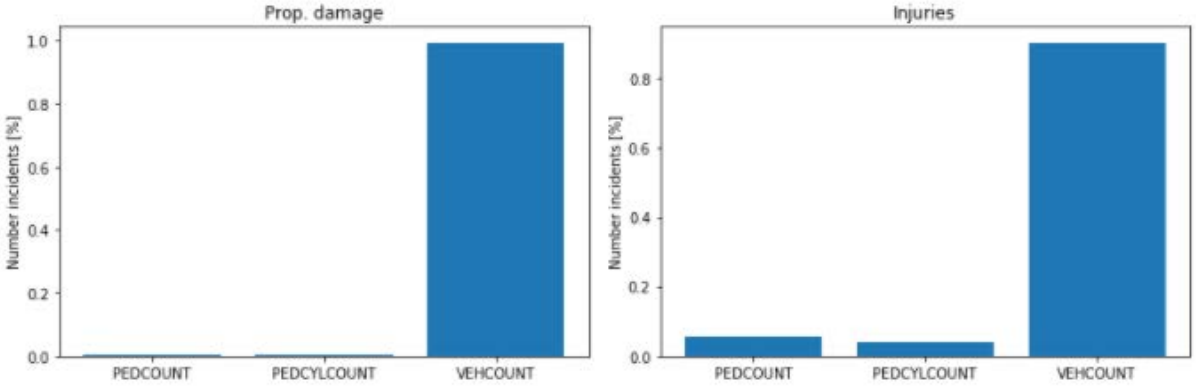
Data Set Basics	
Title	Collisions—All Years
Abstract	All collisions provided by SPD and recorded by Traffic Records.
Description	This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.
Supplemental Information	
Update Frequency	Weekly
Keyword(s)	SDOT, Seattle, Transportation, Accidents, Bicycle, Car, Collisions, Pedestrian, Traffic, Vehicle
Contact Information	
Contact Organization	SDOT Traffic Management Division, Traffic Records Group
Contact Person	SDOT GIS Analyst
Contact Email	DOT_IT_GIS@seattle.gov

Methodology

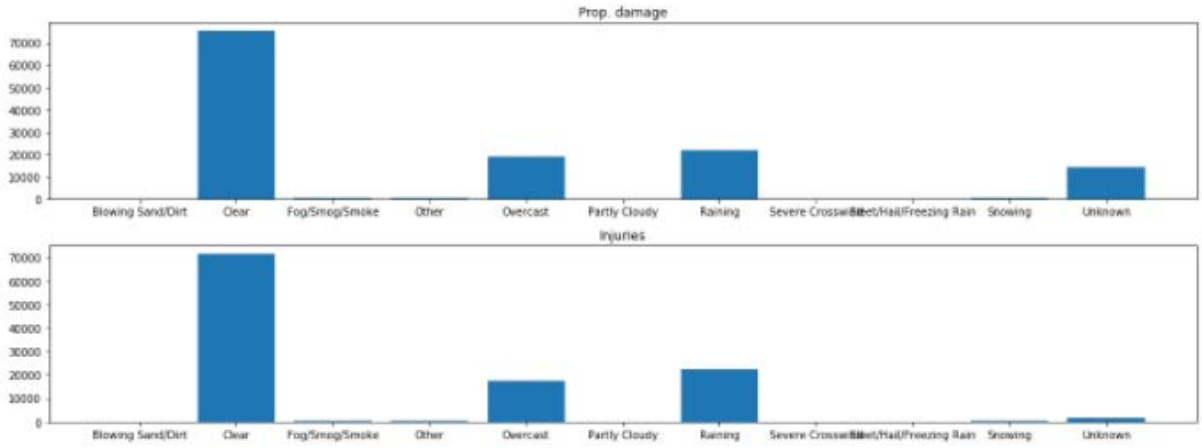
My approach to analyze the data was the following

- Load data
- Investigate dimensions and data statistic description
- Visualize the data and identify some key features that are interesting with regards to understand the severity of injuries in accidents.
- Check if some features are correlated
- Perform some data pre-processing
    - Label encoding
    - Label balancing
    - Select features (or create new ones if needed)
- Training the model
    - Normalize
    - Select train-test split
    - Select appropriate algorithms for the task (classification algorithms)
    - Find the right parameters to tune the algorithms
- Evaluate the model
- Ready to be pushed into production.

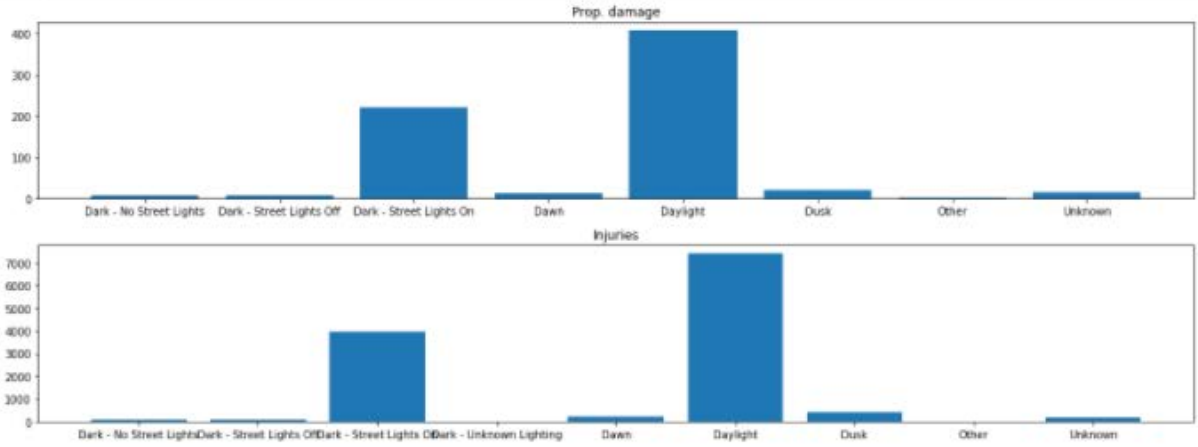
Exploratory data results



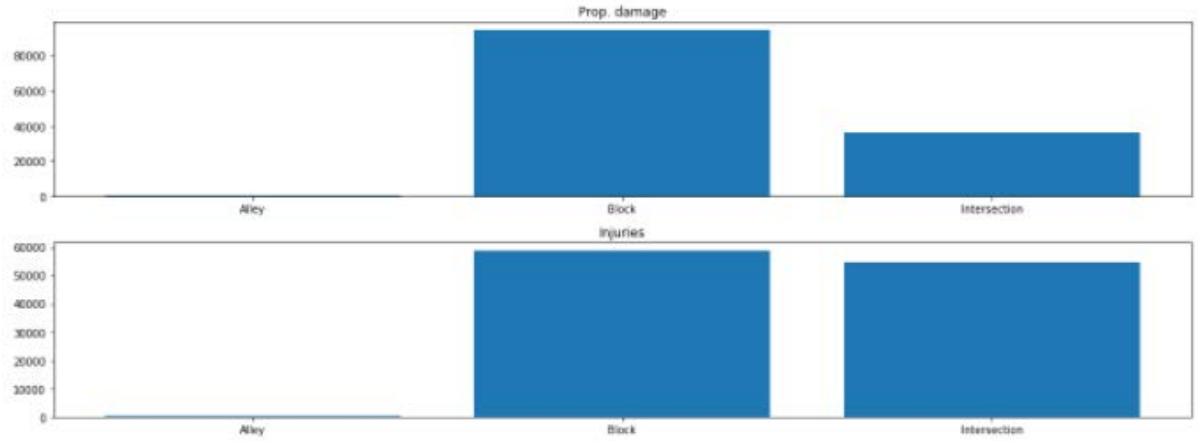
Most accidents occur and are recorded for vehicles. Accidents with pedestrians and cyclist show more often injuries compared to just property damage.



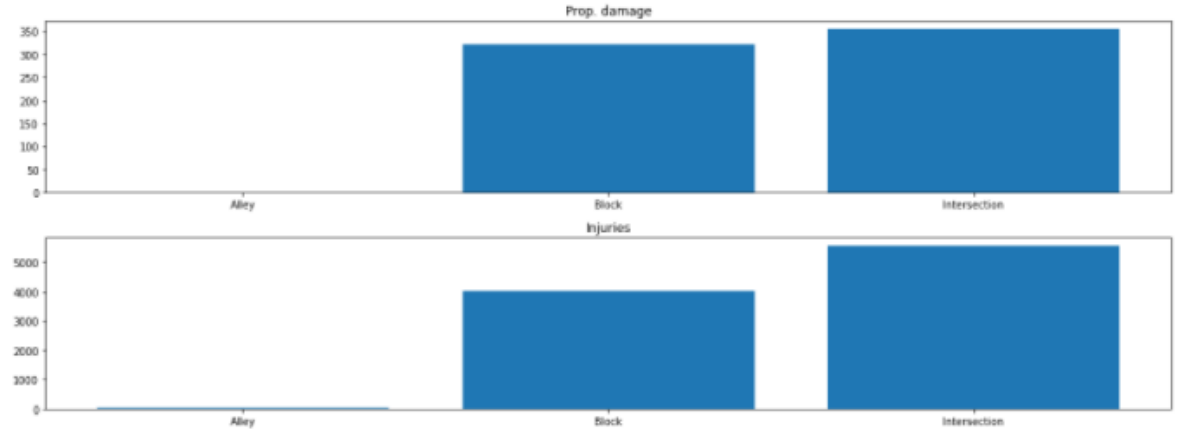
Most accidents occur during clear weather and only the second highest count during rainy weather. As there are also less rainy days, there might be no correlation with the weather conditions.



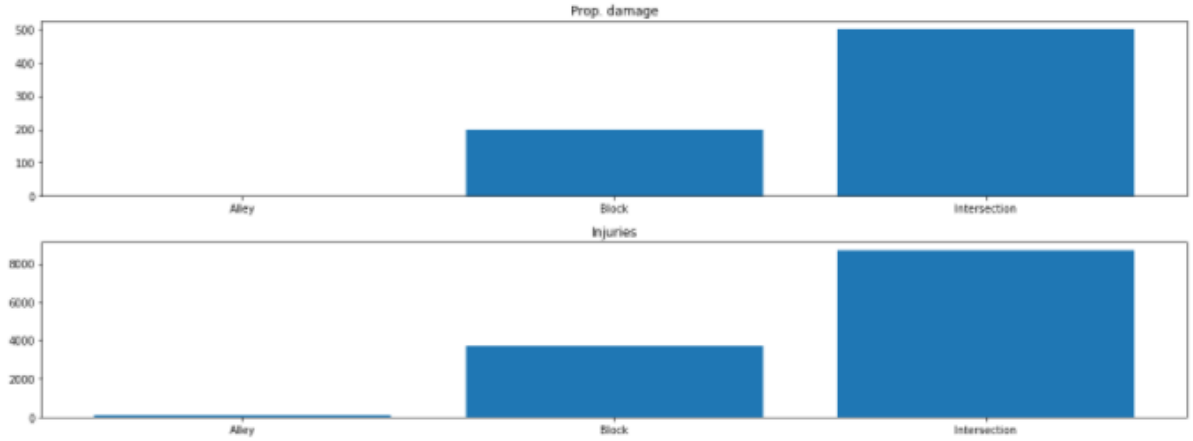
Most accidents occur during daylight suggesting whereas the second highest count occurs during dark hours where street light are on.



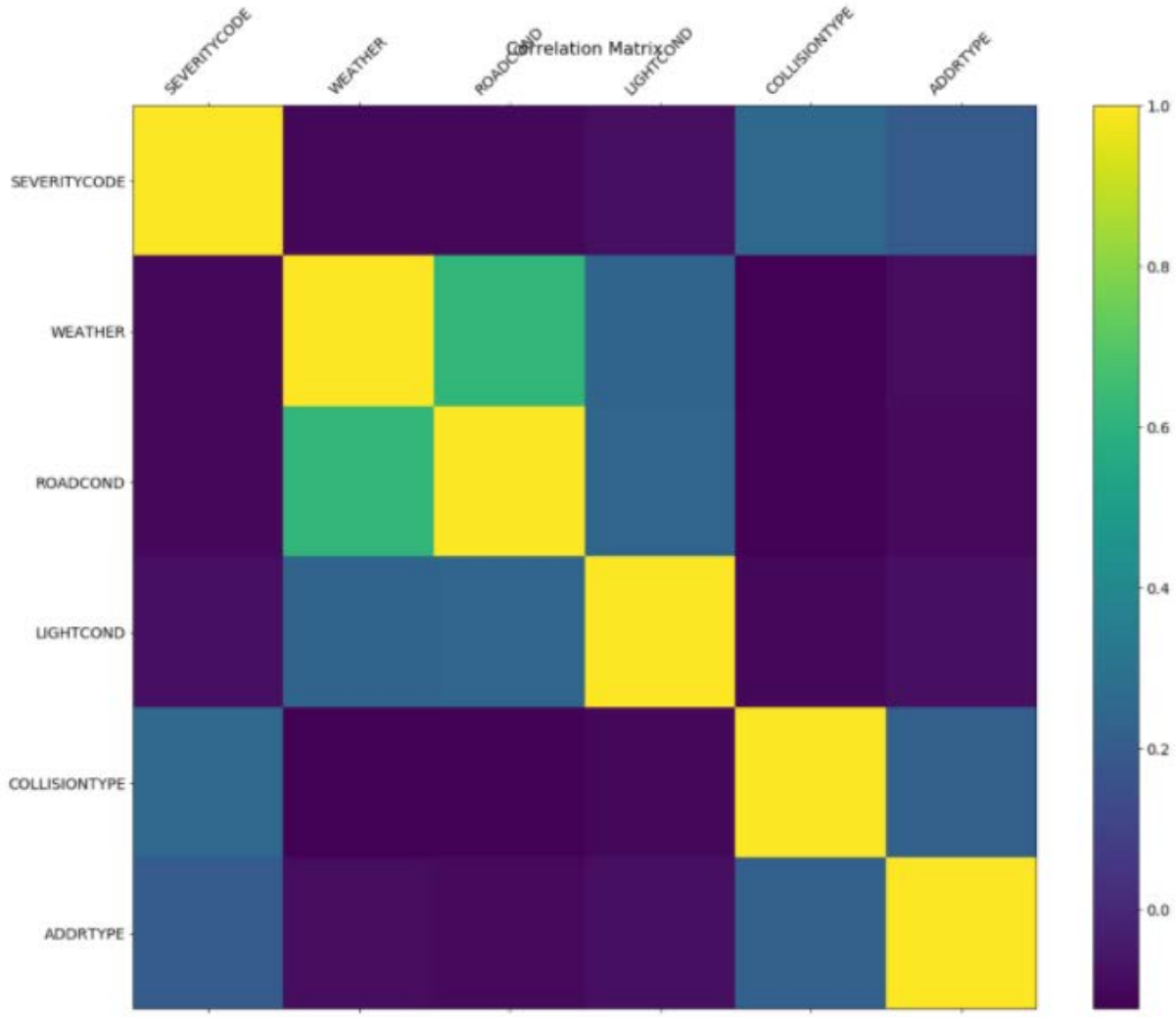
This histogram show that most accidents with vehicles occur on the block, while the accidents with injuries occur more often at intersections.



Cyclist accidents for property damage and injuries occur most frequently at intersections.



For pedestrians, a large part of accidents occur at intersections.



The correlation matrix of the different features show that the weather, road conditions and light conditions are somehow related. This makes sense as for example during rainy days, the sight condition are rather dark, light is needed and the road conditions are often also worse.

Data processing

The data processing included to select features that were populated and add information not covered by others. Here, one could spend more time calculating the effect of in-/excluding certain features on the result improvement (e.g., F1 score). But for now we just choose here some features.

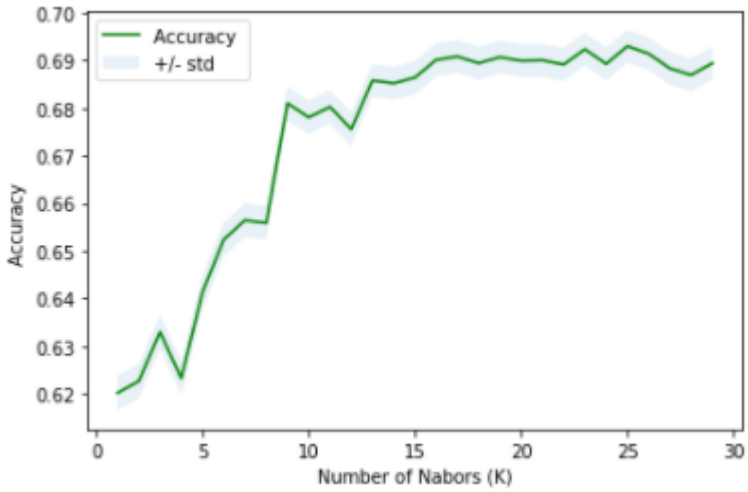
Next step is label encoding. This can be tricky as some algorithms may be sensitive to the chosen encoding.

Next step is to remove the NaN values as we have enough data samples for training. The last two steps before model building to balance the target variables in the data set to not have one heavily over-represented and finally normalize the feature range.

Selecting the algorithms and tuning their parameters

This is a classification problem and some of the most popular algorithms are KNN, Decision Tree, Logistic regression, and SVM to find the patterns to predict the accident severity.

For the Decision Tree and KNN some parameters must be tuned. For the KNN algorithm I found that 15 clusters give the best accuracy.



For the decision tree model, I found that a maximum depth of 7 gives a good accuracy vs complexity trade-off.

Evaluating the models

KNN

```
KNN's evaluation F1-score:        0.67
KNN's evaluation Jaccard-score:        0.67
```

Decision Tree

```
DecisionTrees's F1-score:        0.70
DecisionTrees's  evaluation Jaccard-score:        0.70
```

Logistic regression

```
Logistic Regression evaluation F1-score:        0.59
Logistic Regression evaluation Jaccard-score:        0.60
Logistic Regression evaluation Logloss-score:        0.65
```

SVM

```
SVM evaluation F1-score:        0.68
SVM evaluation Jaccard-score:        0.68
```

Result

From the evaluated models, the decision tree presents the highest score to predict future severity levels before SVM and KNN.

Discussion

Already the exploratory step in the project showed that most accidents for walkers and cyclists occur at intersections. Most cyclists and pedestrian accident occur in combination with a vehicle as we can read in the data table columns. For vehicles, the number of injuries is as high for intersection and block accidents.

Overcast and rainy weather conditions appear to important although most accidents occur on clear weather. Most recorded accidents involve a vehicle. To get a normalized view whether they have an especially high percentage of generating accidents compared to cyclists, the total number of cyclists, pedestrian and vehicles should be considered.

The logistic regression did not perform as expected, but perhaps the predictive power is limited due to the limited parameter tuning I performed. The SVM and KNN are very similar in terms of the evaluation score. Further fitting tools could be tested for the SVM instead of the used ‘rbf’.

To potentially enhance and make the models more robust, ensemble averaging could be considered using KNN, SVM and decision tree.

Conclusion

The limited data study performed suggest that cyclists and pedestrians are more likely to experience severe injury if it occurs at intersections. Many of these occur together with a vehicle. Other factors for numbers and severity levels are rainy and dark weather conditions although the majority of accidents happen during the day in clear weather.

Considering the built model, a first responder call could estimate what the severity level of an accidents is if there are no further information.