Collision data from the city Seattle

Background

Traffic accident data recorded by Traffic Records and Seattle police department

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

Total

- 194’673 records
- 37 features
- Key parameter severity of injuries

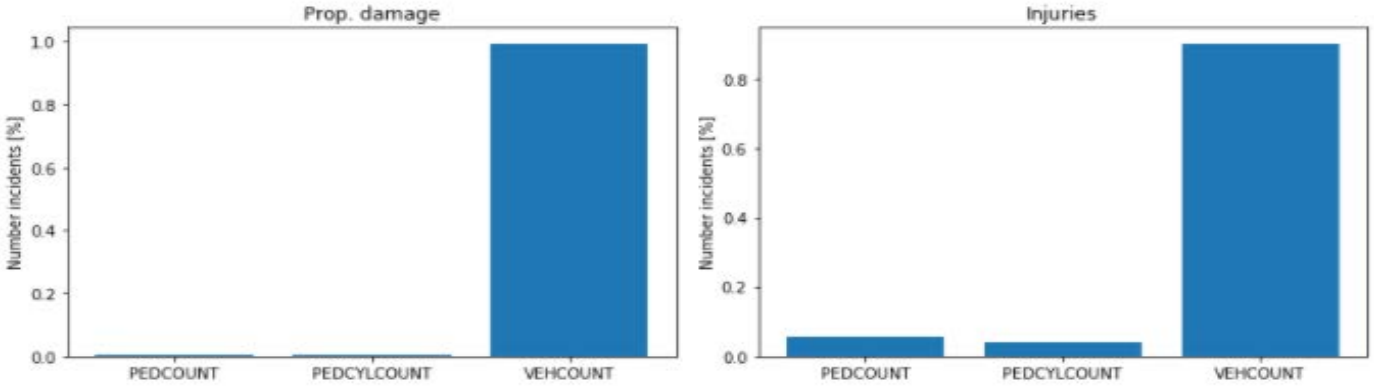
Severity levels (1&2 exist)
3=Fatality
2b=Serious injury
2=Injury
1=Proper damage
0=Unknown

Approach

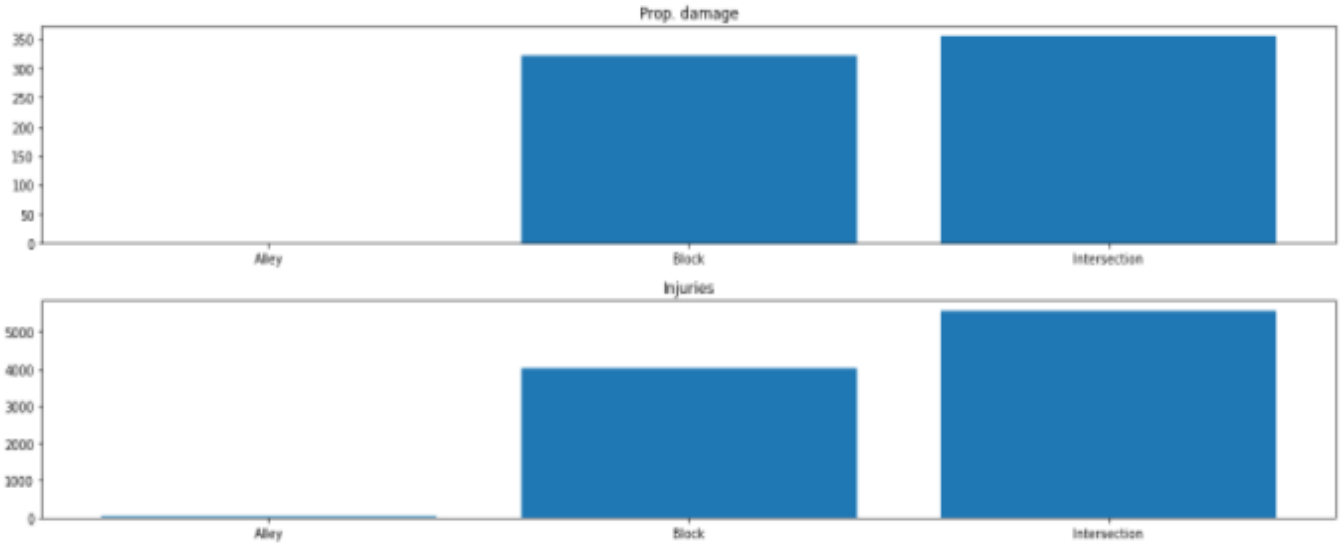
The analysis structured in the following

- 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

Data exploration



Most recorded accidents with vehicles. More investigation required to compute relative accident numbers.

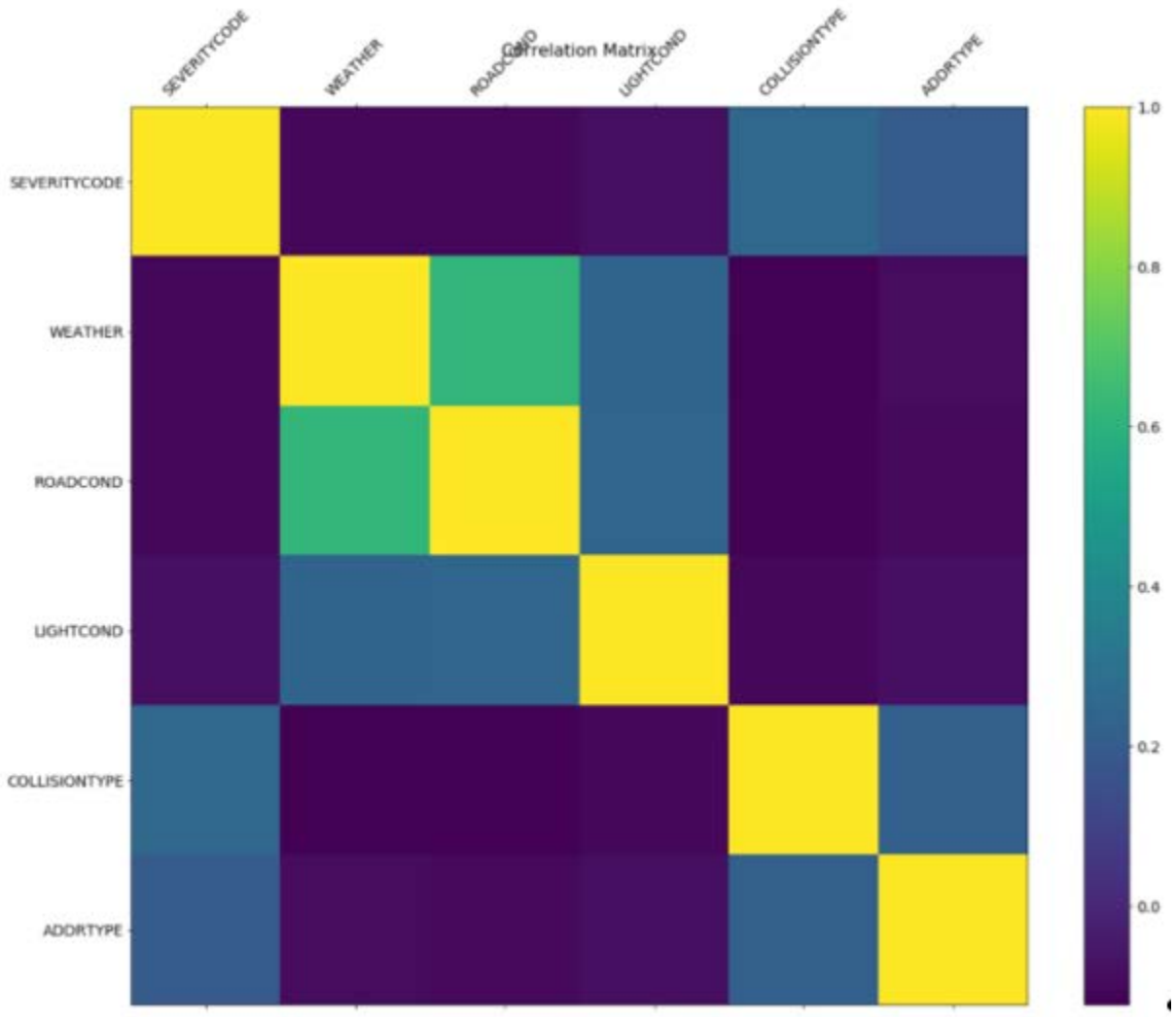


Accident numbers for cyclists for the severity level property damage(top) and injuries (bottom) split by where the accidents occured.

Intersections are hotspots for cycle accidents

For pedestrians, most accidents occur at intersections, mostly with vehicles

Selected features to build model



Weather, road conditions, and light conditions appear to be linked. This may mean that they not necessarily reflect to be chosen as features all together.

Data pre-processing

1. Label encoding
2. Drop or select features
3. Remove data rows with no value entries
4. Balance dataset to have equal data points within the two severity key parameter range. Data point numbers before
    - 130634 for severity code 1
    - 56870 for severity code 2
5. Normalize variables

Define goal

Build model to be able to predict severity levels depending on available information.

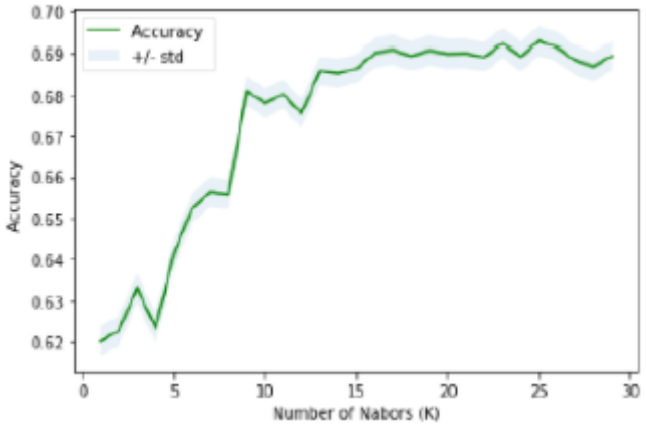
Approach: classification problem

Selected algorithms

- KNN
- Decision trees
- Logistic regression
- SVM

Tuning the model parameter

- KNN cluster number => 15



Tuning and evaluation results

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
```

Conclusion

- Most accidents for walkers and cyclists occur at intersections
- Most cyclists and pedestrian accident occur in combination with a vehicle
- Overcast and rainy weather conditions appear to important although most accidents occur on clear weather.

The built model can now be used to do predictions of severity levels depending on the available information interesting for

- Police
- Ambulance
- City planners
- Insurance companies