Predicting Accident Severity Based on Collisions Data in Seattle

Capstone Project for completing IBM Data Scientist Professional Course

Author : Andre K.

September 2020

Introduction

Prevent or minimize traffic accidents injury

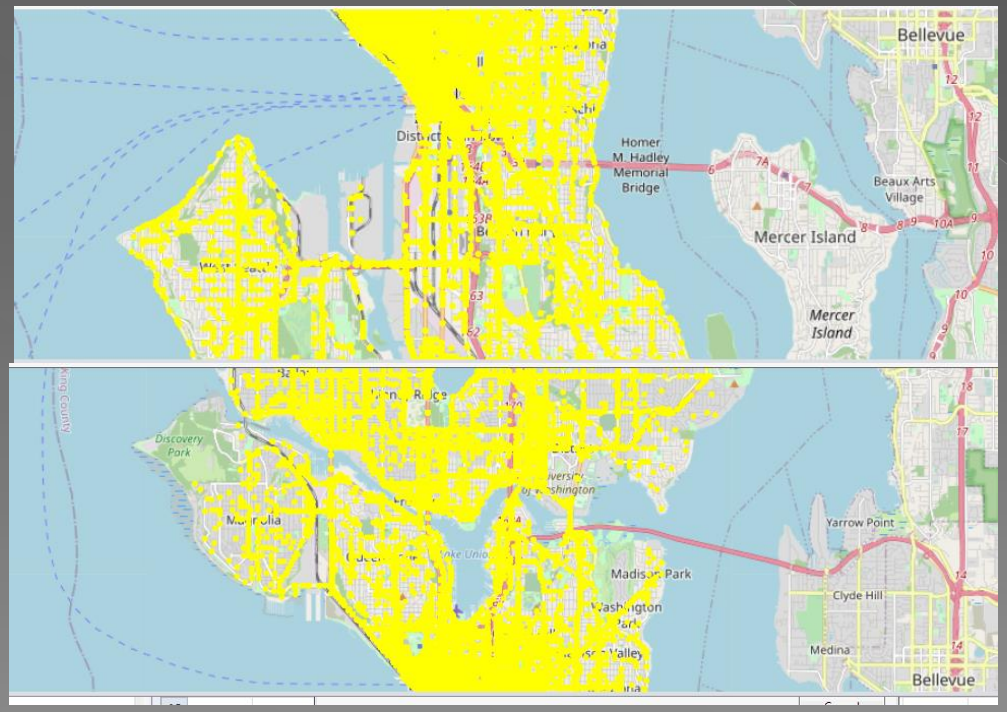
- Apply data science on collisions data in Seattle.
- Build model to predict severity level of accidents under certain traffic circumstances/parameters.
- Traffic parameters are categorized into: location type, collision type, number of persons involved, day and time, violations and environment/conditions. Target parameter is whether the collision causing injury or not.
- Model outcome shows 70% accuracy, in which traffic authorities can utilize to predict accident severity and take necessary measures to prevent or minimize traffic accidents related injury,

Data Acquisition & Cleaning

- Utilize car collisions data made available by city of Seattle, covering 15+ years with nearly 200,000 of car collisions data.
- 37 attributes and 1 target data (severity level code 1 or 2).
- Code 1 : collision causing property damage only
  Code 2 : collision causing injury
- Some columns have missing data, which will be cleaned in our process

Features : Location

Location attributes contain information of coordinate of location (X: longitude, Y: latitude), Address Type, Location/Address, Junction Type, lane segment key and crosswalk key. While the location information illustrate how collisions data are scattered, our objective is to model collisions of Seattle in general, not location specific.



"Address Type" is selected as feature as it can be applied as location feature in general. We can see that most collision occurs at Block and Intersection.

ADDRTYPE	SEVERITYCODE	
Alley	1	669
	2	82
Block	1	96830
	2	30096
Intersection	1	37251
	2	27819

Features : Collision Type

There are several data to explain type of accidents: Collision Type, SDOT code, SDOT description, ST code, ST description and Hit Parked Car. Using all of these column will make the feature redundant.

"Collision Type" is chosen as it has more complete data and intuitive classification.

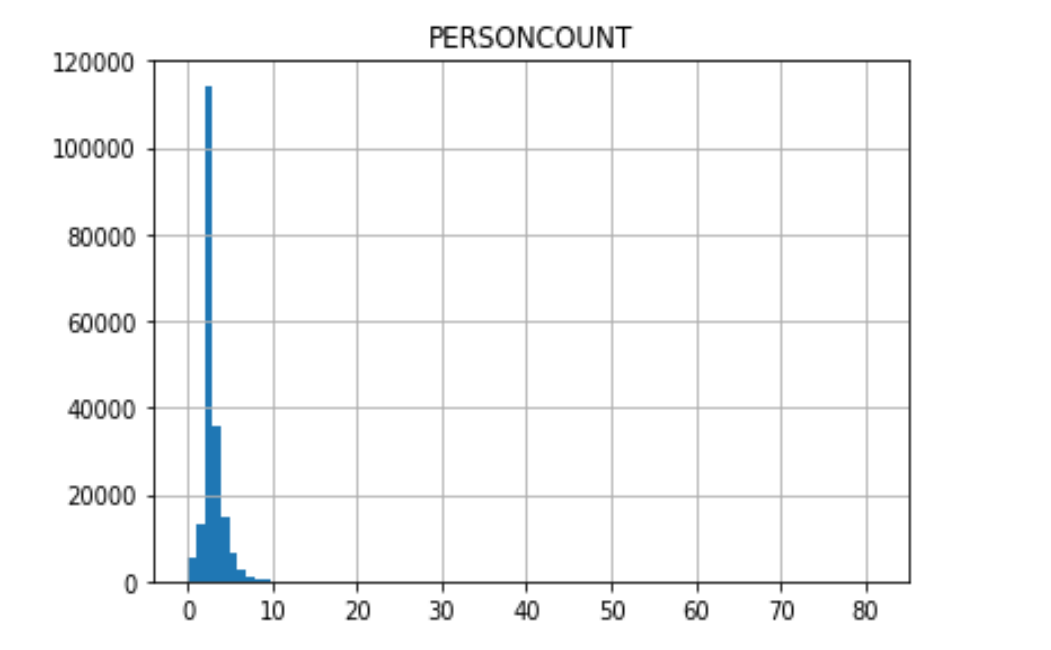
COLLISIONTYPE	SEVERITYCODE	
Angles	1	21050
	2	13624
Cycles	1	671
	2	4744
Head On	1	1152
	2	872
Left Turn	1	8292
	2	5411
Other	1	17591
	2	6112
Parked Car	1	45325
	2	2662
Pedestrian	1	672
	2	5936
Rear Ended	1	19419
	2	14671
Right Turn	1	2347
	2	609
Sideswipe	1	16103
	2	2506

Features : Number Involved

Data related to number involved in accident are the columns for person counts, pedestrian counts, cyclist counts and vehicle counts

"Person Counts" is selected as feature to make the model more general (not specific to pedestrian or cyclist).

Most of collision involves 3 persons.



Features : Date/Time

From Date and Time related data, we can categorize (bin) date into indication whether collision occurs on weekend or weekday. Meanwhile for time data, we can categorize (bin) into whether or not the collision occurs at rush hour.

Date/Time data is transformed into "weekend indicator" and "rush hour indicator", which both are used as features for the model.

weekend_indi	SEVERITYCODE	
0	1	101291
	2	44038
1	1	35194
	2	14150

rushhour_indi	SEVERITYCODE	
0	1	109659
	2	44832
1	1	26826
	2	13356

Features : Violations

Four columns can be attributed to violation of traffic rules, namely: Pedestrian rights are not granted, speeding, inattention driving and driving under influence. Each of this is used as features for the model.

PEDROWNOTGRNT	SEVERITYCODE	
Y	1	460
	2	4207

INATTENTIONIND	SEVERITYCODE	
Y	1	19408
	2	10397

SPEEDING	SEVERITYCODE	
Y	1	5802
	2	3531

UNDERINFL	SEVERITYCODE	
N	1	127071
	2	53597
Y	1	5559
	2	3562

Features : Conditions

Three attributes indicate environment/condition at the time collision happens : weather, road condition and light condition. Each of these is used as features for the model.

WEATHER	SEVERITYCODE	
Blowing Sand/Dirt	1	41
	2	15
Clear	1	75295
	2	35840
Fog/Smog/Smoke	1	382
	2	187
Other	1	716
	2	116
Overcast	1	18969
	2	8745
Partly Cloudy	1	2
	2	3
Raining	1	21969
	2	11176
Severe Crosswind	1	18
	2	7
Sleet/Hail/Freezing Rain	1	85
	2	28
Snowing	1	736
	2	171
Unknown	1	14275
	2	816

ROADCOND	SEVERITYCODE	
Dry	1	84446
	2	40064
Ice	1	936
	2	273
Oil	1	40
	2	24
Other	1	89
	2	43
Sand/Mud/Dirt	1	52
	2	23
Snow/Slush	1	837
	2	167
Standing Water	1	85
	2	30
Unknown	1	14329
	2	749
Wet	1	31719
	2	15755

LIGHTCOND	SEVERITYCODE	
Dark - No Street Lights	1	1203
	2	334
Dark - Street Lights Off	1	883
	2	316
Dark - Street Lights On	1	34032
	2	14475
Dark - Unknown Lighting	1	7
	2	4
Dawn	1	1678
	2	824
Daylight	1	77593
	2	38544
Dusk	1	3958
	2	1944
Other	1	183
	2	52
Unknown	1	12868
	2	605

Modelling : Logistic Regression

Logistic Regression is chosen to model the problem, because:

- Target variable is binary : Severity level code 0 means collision does not cause injury, and 1 means collision does cause injury.
- Benefit from obtaining probabilistic result.
- The problem exhibits linear decision boundary. Features in the data show linear decision boundary that is suitable for logistic regression problem.
- We can explore and understand the impact of each feature to the severity level of accidents.

Modelling : Decision Tree

As second approach, Decision Tree model is applied to model the problem

- Decision tree predict the outcome of severity level of collision the model by creating branches based on each feature : address type, collision type, weekend indicator, rush hour indicator, violation, and environment condition.

- The model create decision treee hierarchy iteratively by maximizing the information gain (or minimizing entropy) produced by the decision tree.

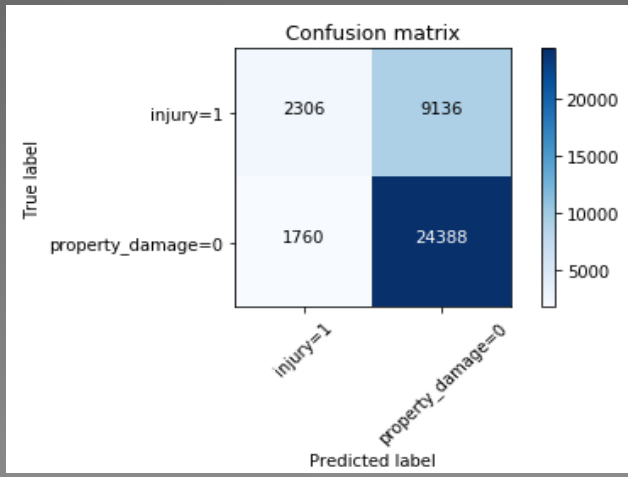
Model Evaluation (Logistic Regression)

- Jaccard Similarity Score : 0.71

```
from sklearn.metrics import jaccard_similarity_score
jaccard_similarity_score(y_test, yhat)
```

```
0.7101356743814844
```

- f1 Score : 0.66, obtained from confusion matrix




```
print (classification_report(y_test, yhat))
```

	precision	recall	f1-score	support
0	0.73	0.93	0.82	26148
1	0.57	0.20	0.30	11442
micro avg	0.71	0.71	0.71	37590
macro avg	0.65	0.57	0.56	37590
weighted avg	0.68	0.71	0.66	37590

Model Evaluation (Decision Tree)

Decision Tree's Accuracy Score : 0.75

```
predTree = CollisionTree.predict(X_test)
```

```
from sklearn import metrics
import matplotlib.pyplot as plt
print("DecisionTrees's Accuracy: ", metrics.accuracy_score(y_test, predTree))
```

```
DecisionTrees's Accuracy:  0.7502793296089385
```

Conclusions

- The model utilizes collisions data in Seattle over past 15 years, to predict whether a certain set of traffic conditions will likely cause injury or not

- Features of the model include : location type, collision type, number of persons involved, weekend/weekday /rush hour indicator, violation of traffic rules and condition such as weather, road and light condition.

- Two approaches were chosen (1) Logistic Regression model and (2) Decision Tree model. Evaluation of the model shows accuracy score around 70% for both models.

- Potential use by traffic authorities to prevent or minimize injury and potential fatalities, when facing various traffic conditions.

Future Developments

- Location specific features in the model

  To use location specific (latitude and longitude coordination) of collisions data, to spot specific location where high level severity collision happens and make appropriate safety enhancements for the locations.

- Location specific features in the model

  to utilize data of number of pedestrian and number of cyclist involved in the collisions. This project uses the overall number of persons involved in collision as feature, meanwhile it is reasonable to assume that pedestrian and cyclist are more prone to injury. Further use of these data can be helpful to tackle specific traffic policies toward pedestrian and cyclist.

Disclaimers

This project is conducted for the purpose of completing capstone project for IBM Data Science Professional certification. Although the project makes use of publicly available car accident data in Seattle, there is no relation of this project to the city of Seattle and result of this project, as a whole or part, should not be used outside of the purpose of completing the above-mentioned capstone project.