PREDICTING THE SEVERITY OF A CAR CRASH

By Andrew Lin

INFORMATION ABOUT THE SEVERITY OF A CAR CRASH IS USEFUL

- People can make decisions on where, when, or how to travel if they knew how severe a crash could be going a certain route
- This could also be beneficial to first responders if they were notified about how severe a crash could be

DATA

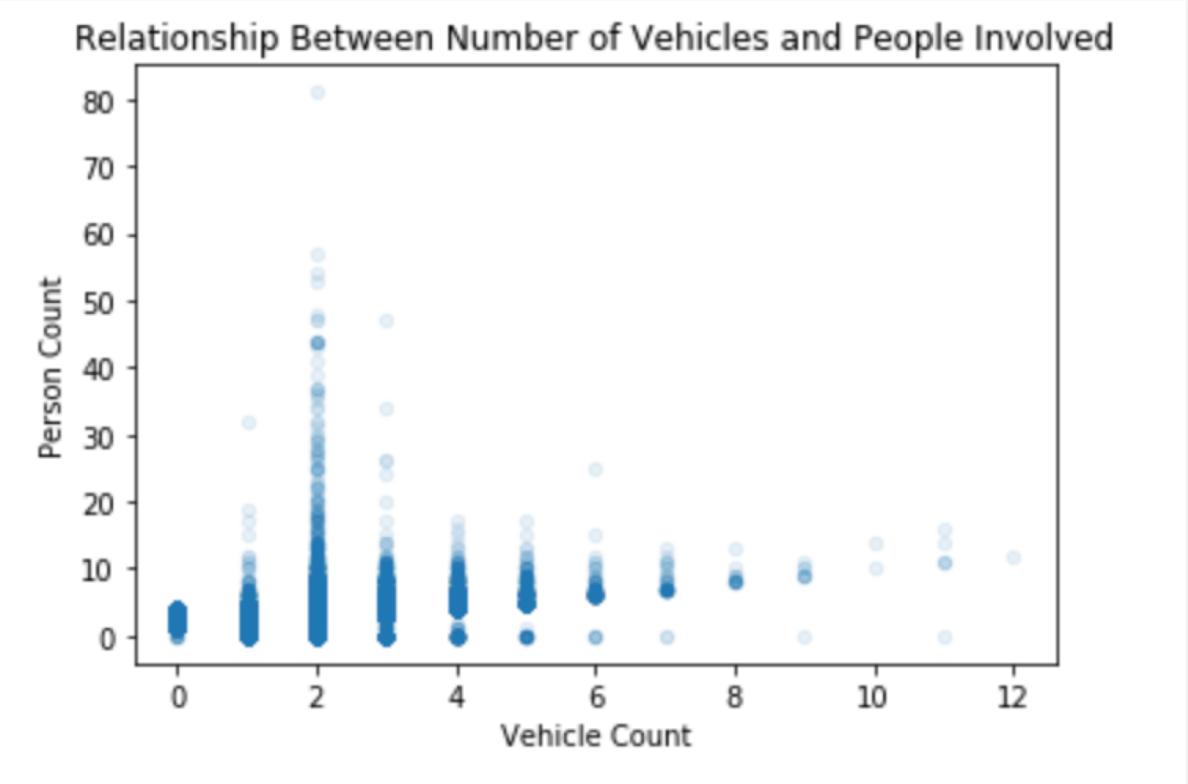
- The data I am using is from the Seattle Department of Transportation
- It has data from 2004 – Present
- 194673 observations and 37 features in the raw dataset

DATA EXPLORATION

- Insights
  - There are some observations where there were 0 people affected or 0 vehicles involved
  - It doesn't take a lot of cars to affect a lot of people
    - 81 people affected, 2 vehicles involved

PERSONCOUNT VS. VEHCOUNT

(3.1 IN REPORT)



DATA CLEANING

- I got rid of all observations that had 0 as a value in PERSONCOUNT or VEHCOUNT because those are not very informative
- I calculated what 2 standard deviations above the mean for PERSONCOUNT and VEHCOUNT would be and dropped those observations to get rid of the outliers and reduce noise
- I decided to fill the null values in my chosen features by imputing the mode for each respective column

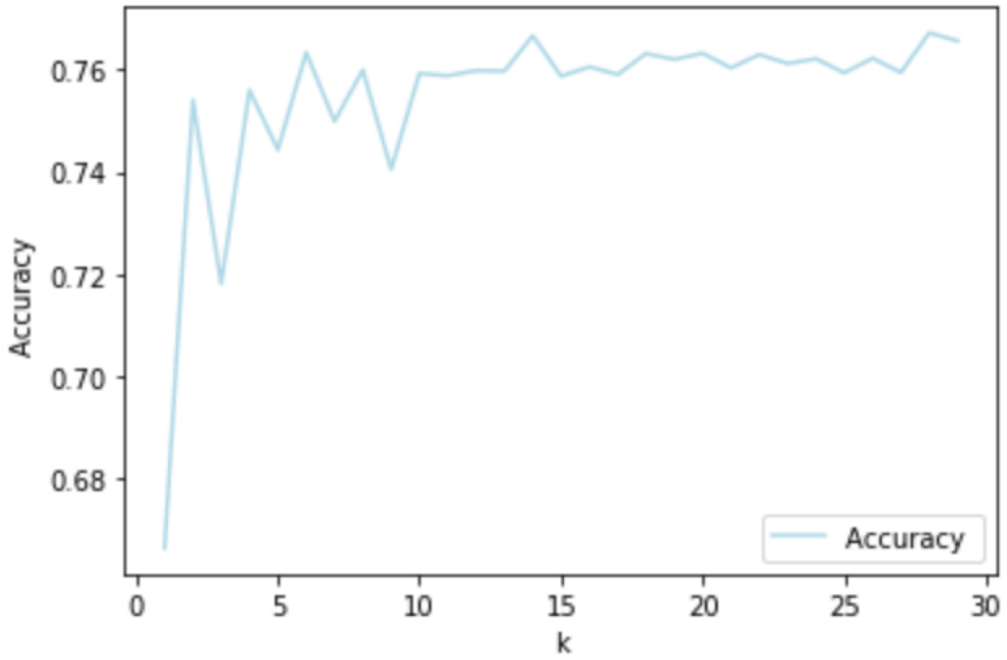
THE FEATURES

- I decided to make a vehicle to person ratio variable because I thought it would be more informative and it did show an accuracy improvement in the models
- The features I used in the model were the encoded versions of ADDRTYPE, COLLISIONTYPE, ROADCOND, LIGHTCOND, and veh_per_pep
- The total features in the final model is 32

THE MODELS

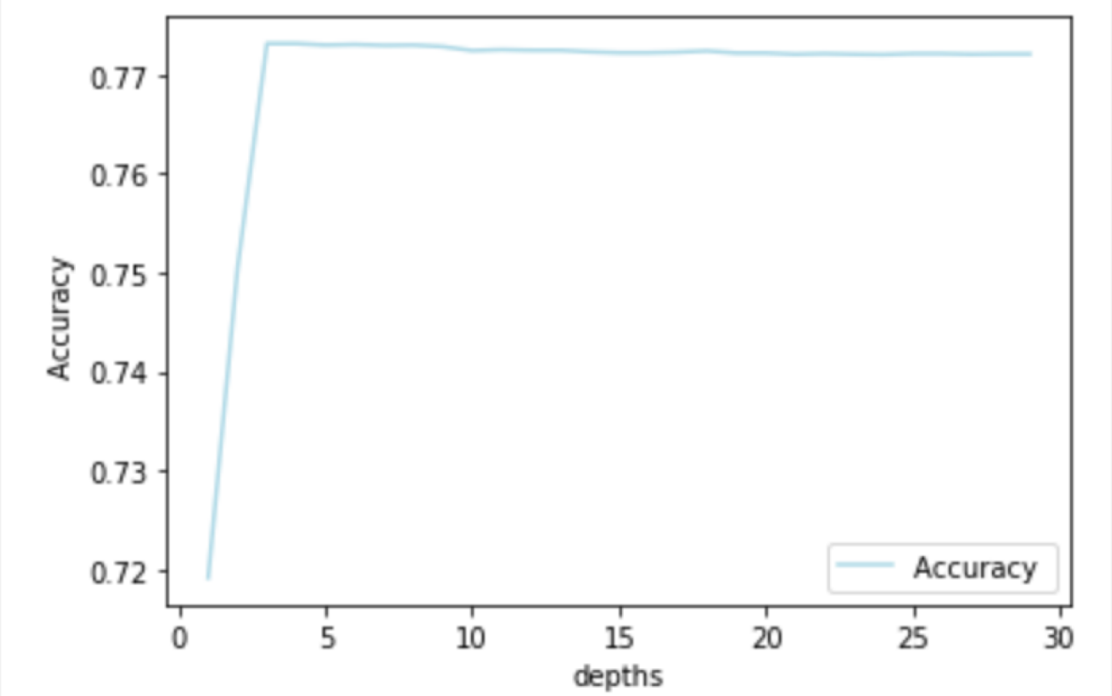
- I tried out 3 different models to predict severity of a car crash
- The 3 I tried out are KNN, Decision Tree, and Logistic Regression

KNN




```
The best accuracy was 0.76725 at k = 28
```

DECISION TREE




```
The best accuracy was 0.77312 at depth =  3
```

LOGISTIC REGRESSION

- For Logistic Regression I used a Grid Search to find the best parameters
- The best parameters were
  - C = 0.001
  - Penalty = l1
  - Solver = liblinear

The accuracy score was 0.77315, the highest out of the three models

EVALUATION METRICS OF EACH MODEL

	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.77	0.73	NA
1	Decision Tree	0.77	0.72	NA
2	LogisticRegression	0.77	0.72	0.46

THE CHOSEN MODEL

- The model I decided to use is the Logistic Regression model
- The accuracy was the best and the evaluation metrics were relatively great
- This model also gives probabilities for each severity code, so that could be useful for a first responder to receive, so they can better prepare

```
LogisticRegression(C = .001, solver = 'liblinear', penalty = 'l1')
```

CONCLUSION AND RECOMMENDATIONS

- Built a model that can help predict the severity of the crash

Looking back and forward

- I would like to have imputed the data in a better way that would have led to some sort of improvement rather than just imputing the mode
- If there is any way to get data on the type of vehicle or age of the people involved in the accident then that would probably help out a lot
- I would also have liked to seen more than just severity codes 1 and 2. In the metadata, it referenced severity codes 2b and 3.

THE PHOTOS MISSING IN THE REPORT

(5.1 IN REPORT)

```
[[22867   263]
 [ 7044  1988]]
              precision    recall  f1-score   support

           1       0.76      0.99      0.86     23130
           2       0.88      0.22      0.35      9032

    accuracy                           0.77     32162
   macro avg       0.82      0.60      0.61     32162
weighted avg       0.80      0.77      0.72     32162
```

Decision Tree

Confusion Matrice

Classification Report

```
[[22591   539]
 [ 6796  2236]]
              precision    recall  f1-score   support

           1       0.77      0.98      0.86     23130
           2       0.81      0.25      0.38      9032

    accuracy                           0.77     32162
   macro avg       0.79      0.61      0.62     32162
weighted avg       0.78      0.77      0.73     32162
```

KNN

```
[[22764   366]
 [ 6931  2101]]
              precision    recall  f1-score   support

           1       0.77      0.98      0.86     23130
           2       0.85      0.23      0.37      9032

    accuracy                           0.77     32162
   macro avg       0.81      0.61      0.61     32162
weighted avg       0.79      0.77      0.72     32162
```

Logistic Regression

CORRELATION MATRIX

(3.1 IN REPORT)

	SEVERITYCODE	PERSONCOUNT	VEHCOUNT
SEVERITYCODE	1.000000	0.128745	-0.249046
PERSONCOUNT	0.128745	1.000000	0.367250
VEHCOUNT	-0.249046	0.367250	1.000000