

Car Accident Severity in Seattle

IBM Data Science
Capstone

Introduction

We have been approached by the Seattle local government who want to understand the features that impact the severity of the car accidents in their city.

They want us to develop a model to predict the severity of a car accident to improve the efficacy of their first responders.

The business problem is that there are limited resources available and so the Seattle local government want a way to optimize their response to an accident, sending more personal to more severe accidents.

The stakeholders in our research are the Seattle local government and the people of Seattle.

Data

Data about car accidents in Seattle can be found here:

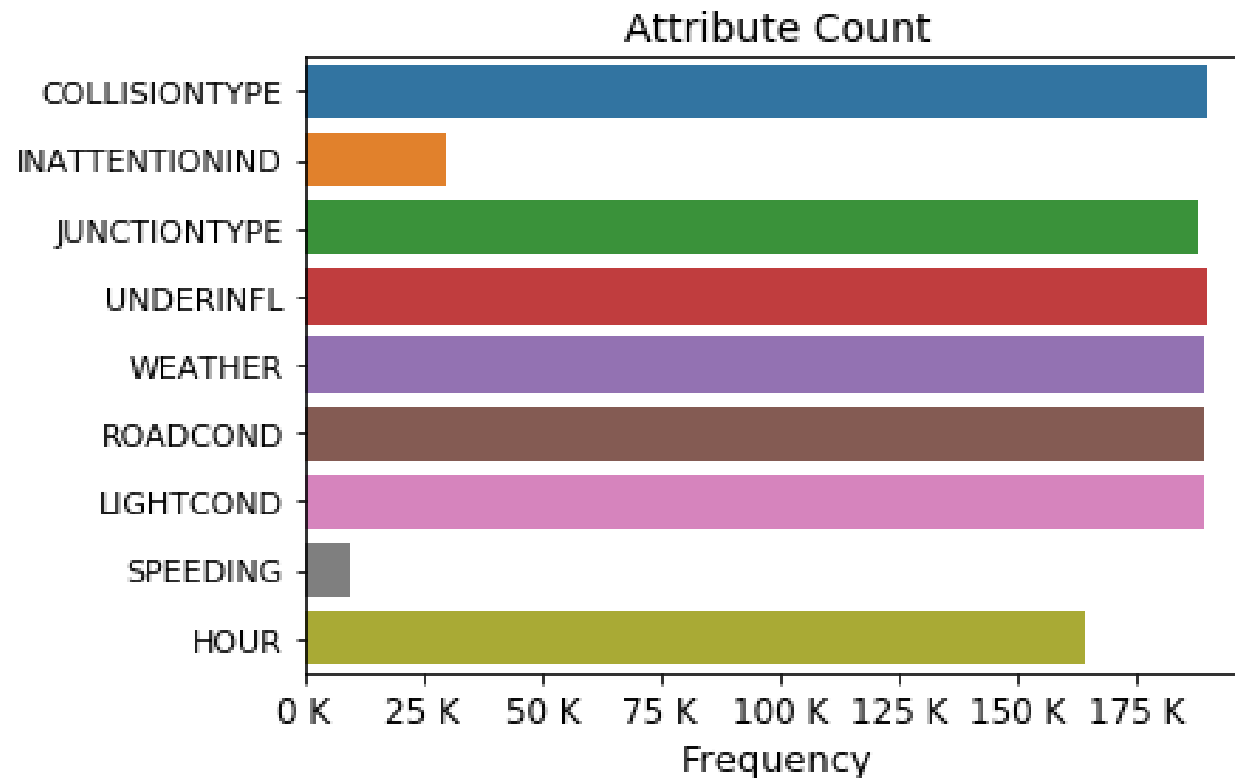
<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>

Metadata about the database can be found here:

<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>

Data Cleaning

There are some significant differences in the number of entries between attributes with large absences for INATTENTIONIND and SPEEDING.



Data Cleaning

For the following attribute some instances are labelled as unknown:

JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND (incl Dark - Unknown Lighting)

This label adds no information and so will be replaced by nan and dropped from the dataframe.

For the INATTENTIONIND and SPEEDING attributes the only label is Y for the affirmative. We will encode Y as 1 and nan as 0, under the assumption that the absence of an affirmative implies the negative.

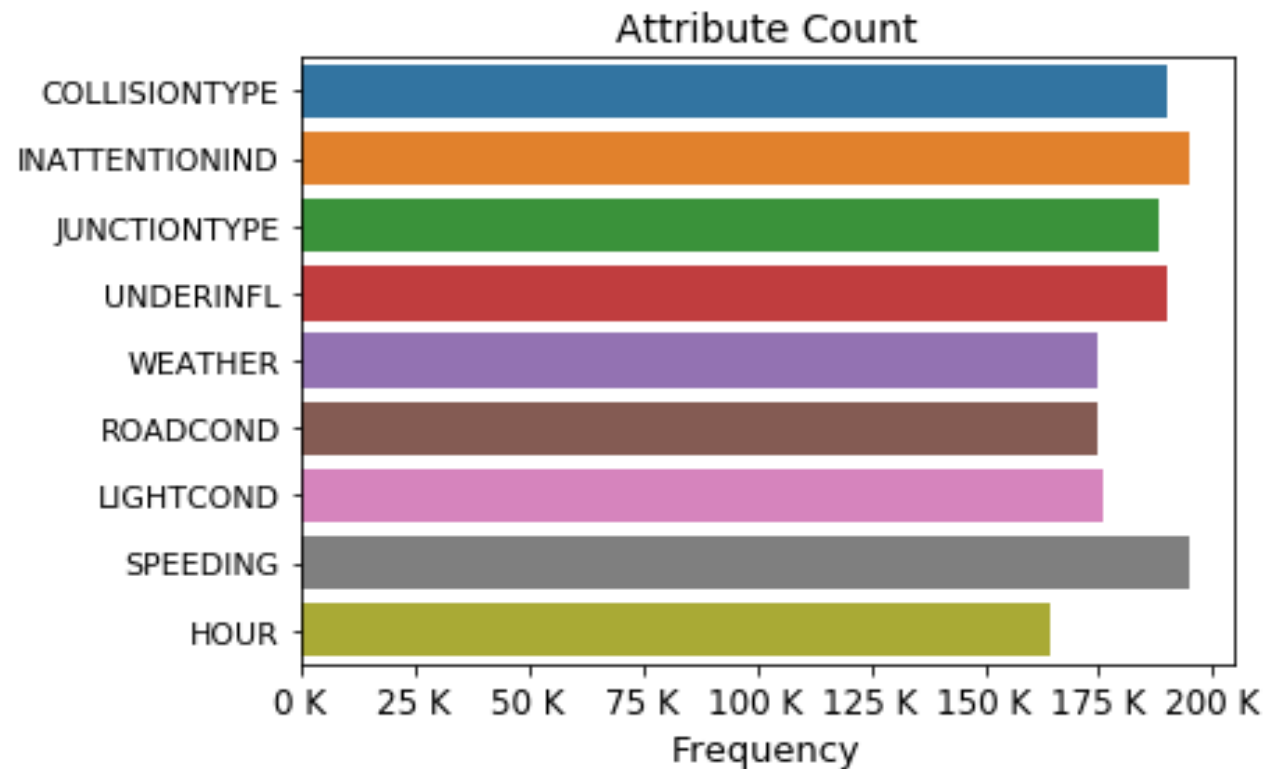
For the UNDERINFL attribute enteries are labelled as either "Y", "1", "N", "0".

We will encode (Y, 1) = 1 and (N, 0) = 0.

Data Cleaning

The number of valid entries for each attribute after cleaning:

We can now remove instances where attributes are undefined (nan):

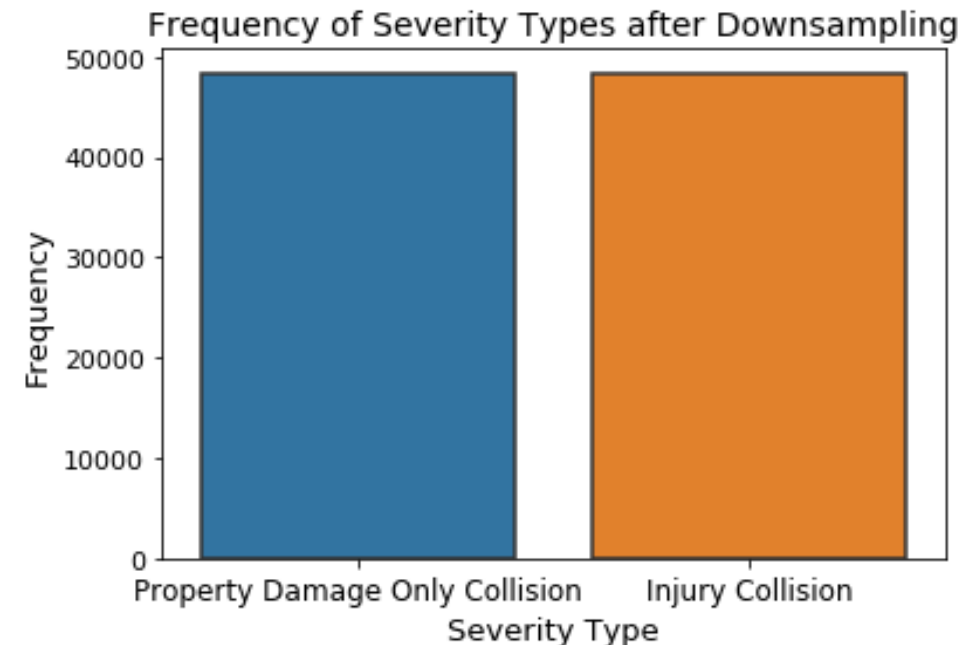
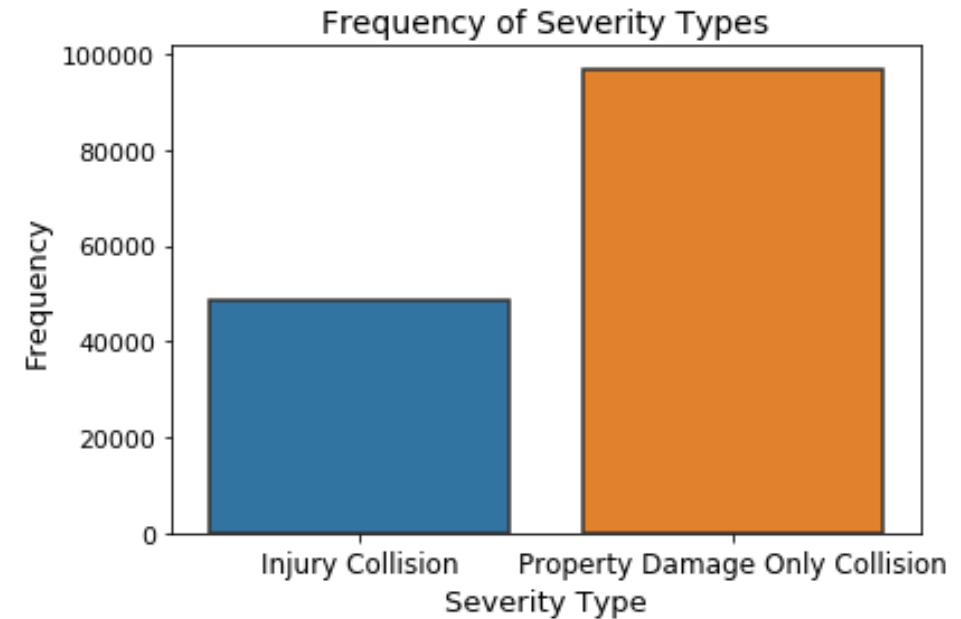


Balancing

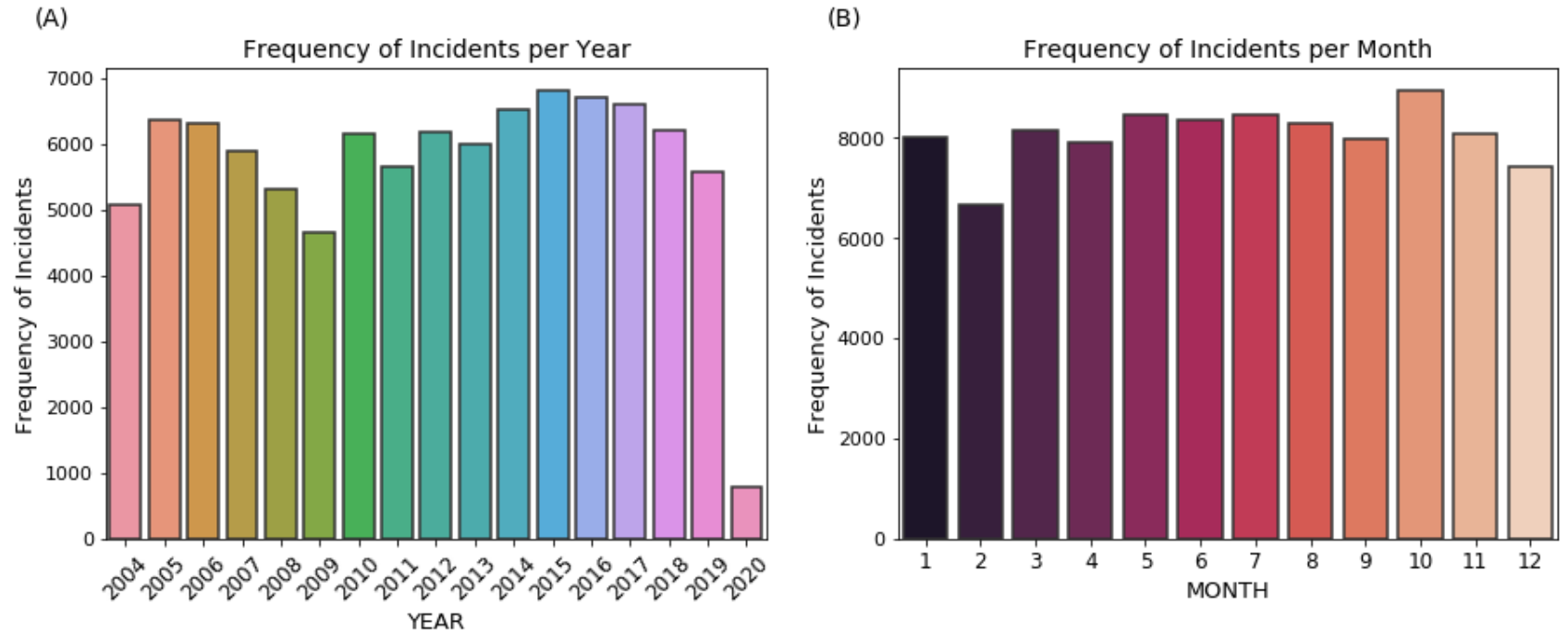
There are [48394, 96951] instances of Severity type ['Injury Collision', 'Property Damage Only Collision'] respectively.

The unbalanced dataset needs to be addressed before applying the machine learning algorithms in order to properly address the predictive capabilities of the machine learning models.

We have a large amount of data and so we undersample the majority severity type (Property Damage Only Collision). Due to the large minority data set the trends present in the full data set should be preserved.



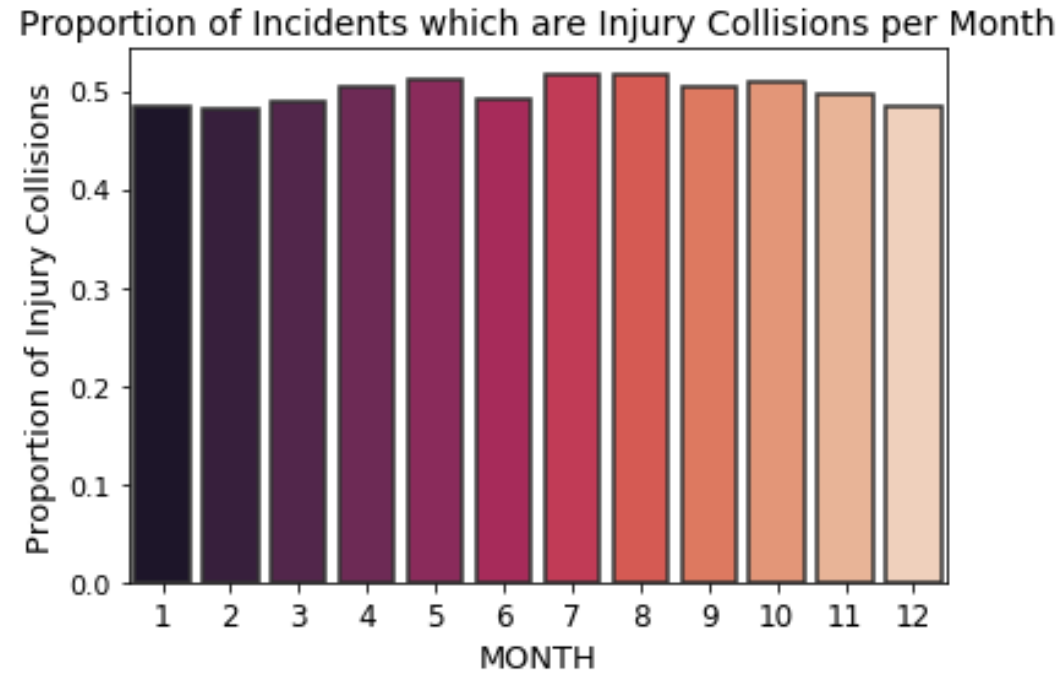
Time Attributes



(A) There doesn't appear to be any clear correlation with the number of instances in each year. We should note that clearly the data is incomplete for 2020.

(B) There doesn't appear to be any seasonal dependence in the frequency of incidents. The absence of any seasonal dependence is somewhat surprising given that variable weather conditions, daylight hours etc.

Time Attributes

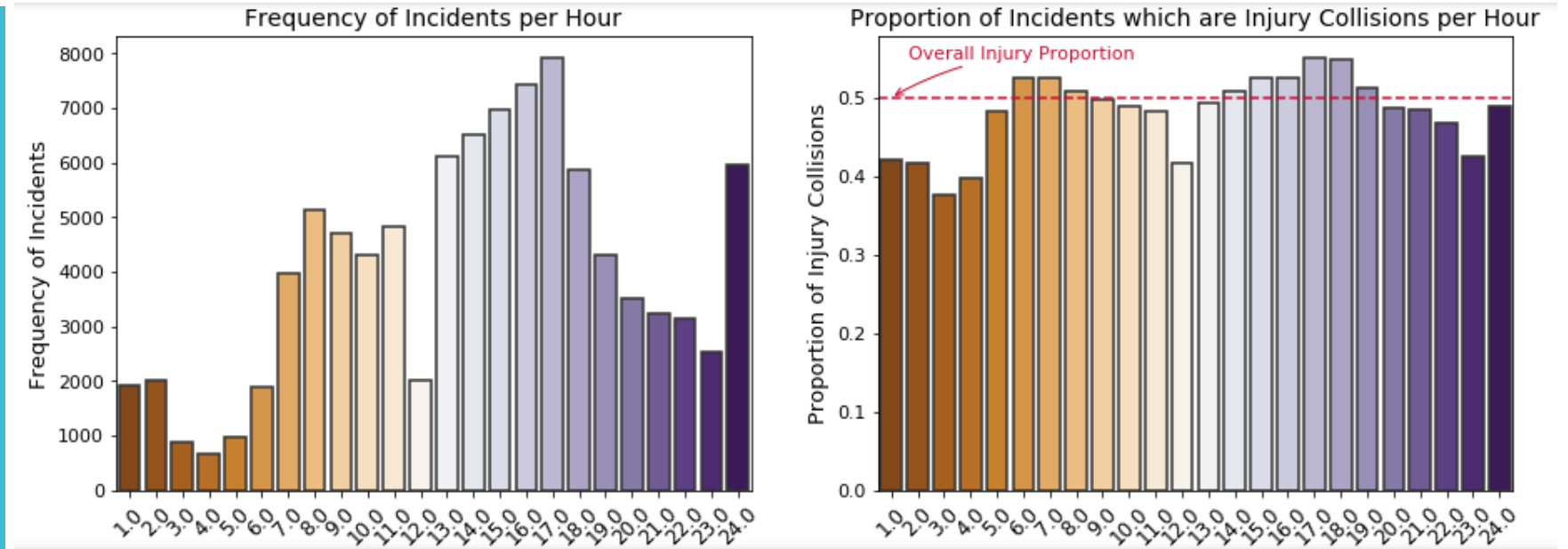


There doesn't appear to be any seasonal dependence in the proportion of incidents classified as an injury collision.

This is again surprising given the variable conditions throughout the year.

- From this cursory analysis the month is probably not useful in predicting the type of collision.

Time Attributes



(A) There is evidence that there is some variation throughout the day in the frequency of incidents per hour. There are relatively few incidents in the early hours of the morning, with local peaks around commuting hours (08:00 - 09:00 and 17:00 - 18:00) and a further peak after midnight (24:00 - 01:00). We could speculate that the origin of this late-night peak could be associated with the closing of establishments.

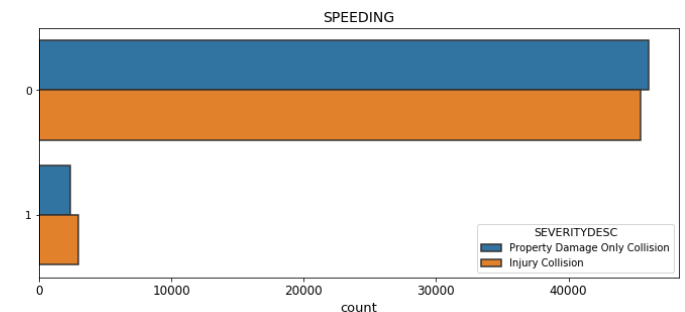
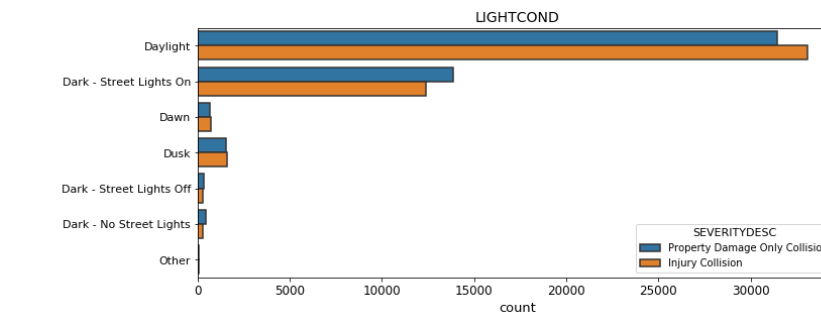
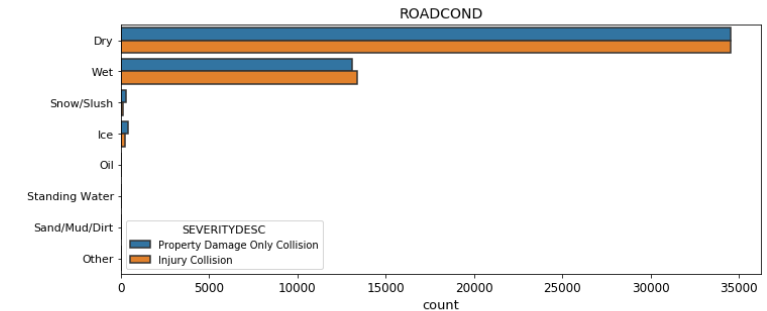
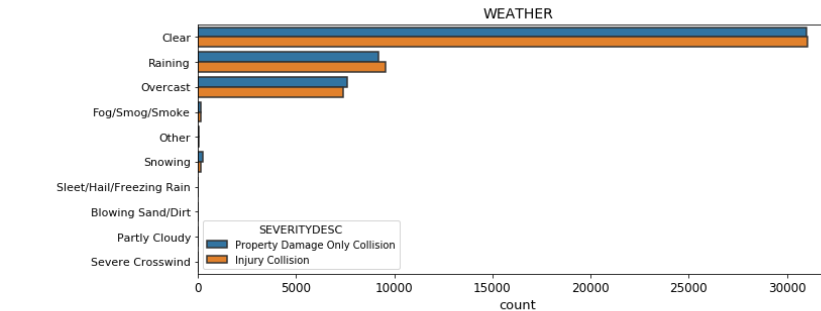
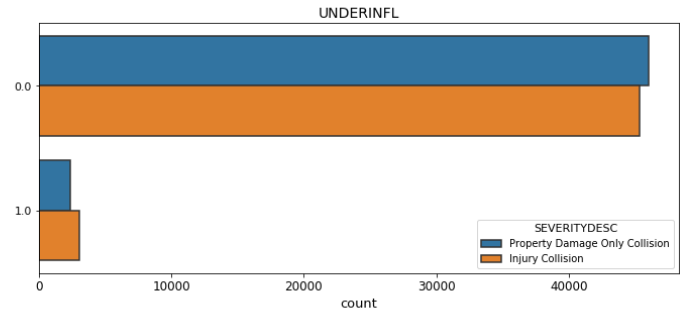
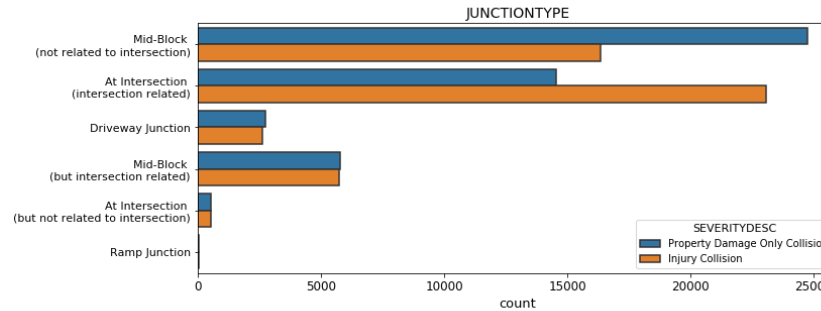
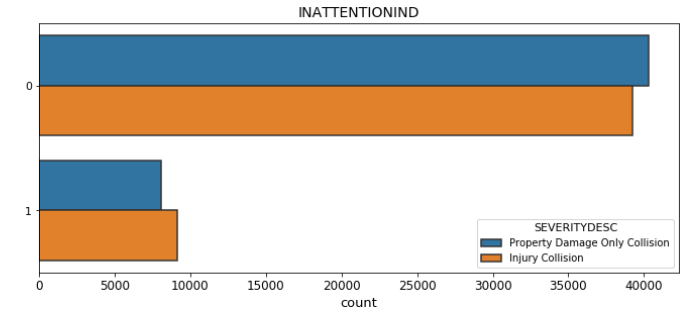
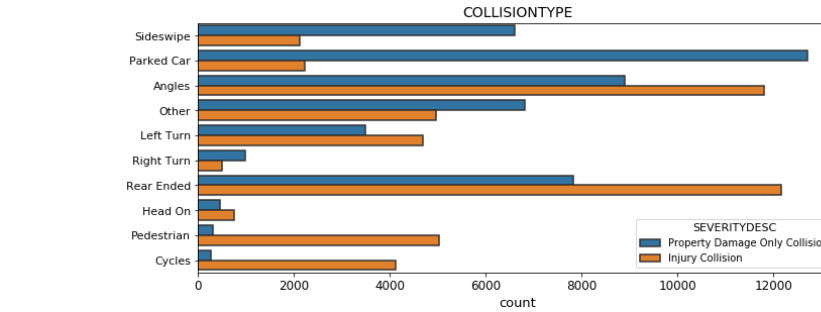
(B) There does appear to be some meaningful variation in the proportion of incidents which are injury collisions. The proportion of injury collisions exceeded the average daily proportion between 06:00 and 09:00 and then again between 13:00 and 19:00

The hour in which an incidence occurs could therefore be a useful predictor of the type of incident

Other Attributes

Using the balanced dataset we would expect attributes where there is a discrepancy between the frequency of different incident types to be predictive.

Conversely an approximately even distribution between categories of an attribute suggests that the attribute has little or no predictive capability to distinguish between different types of incidents.



Other Attributes

For the attributes:

COLLISIONTYPE, JUNCTIONTYPE

There are some large discrepancies within categories for the different incident types. We would therefore expect this attributes to be useful predictors of the incident type.

For the attribute:

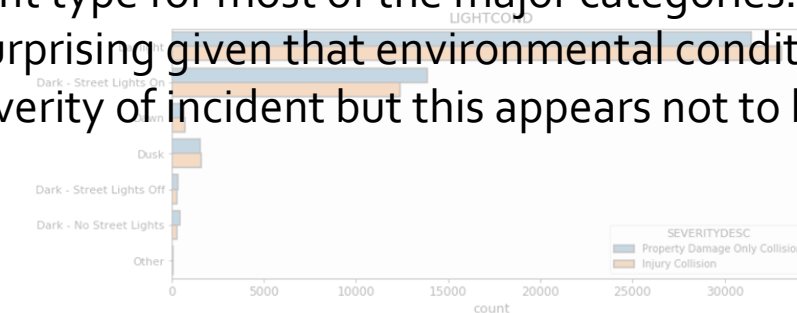
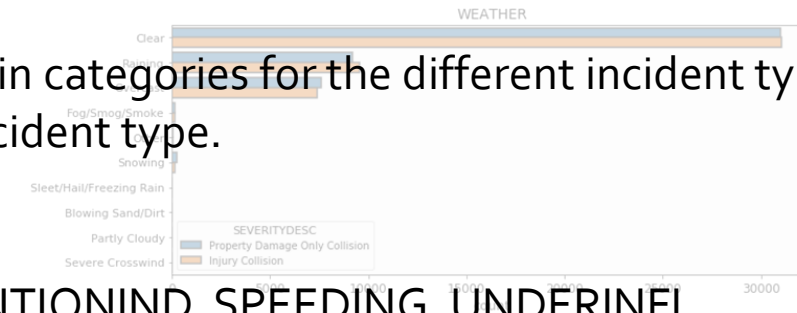
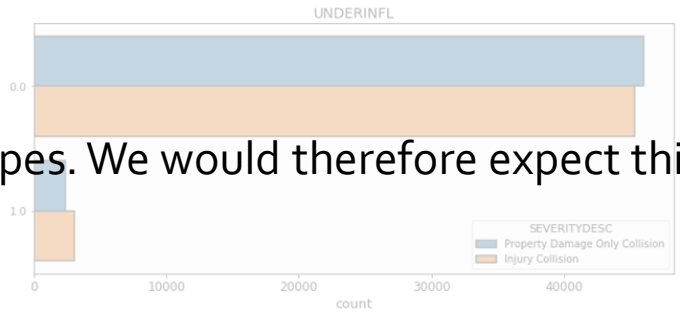
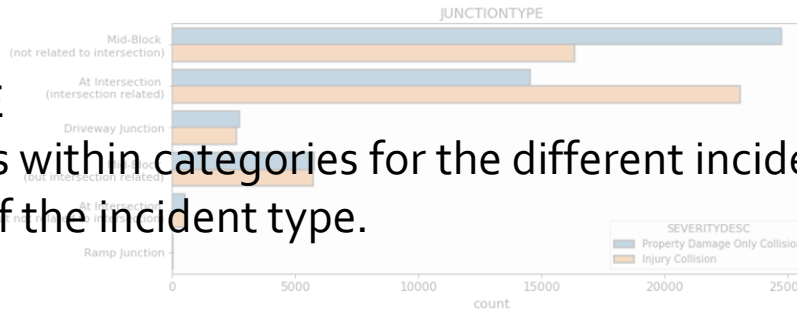
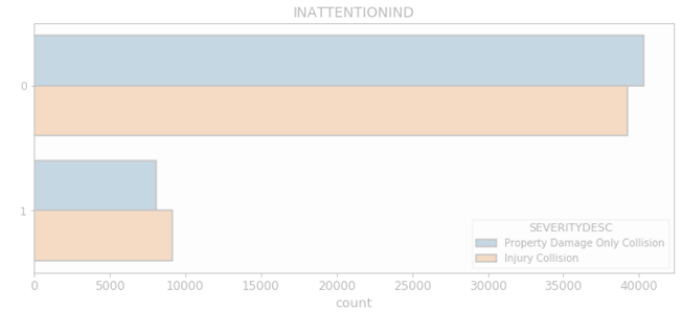
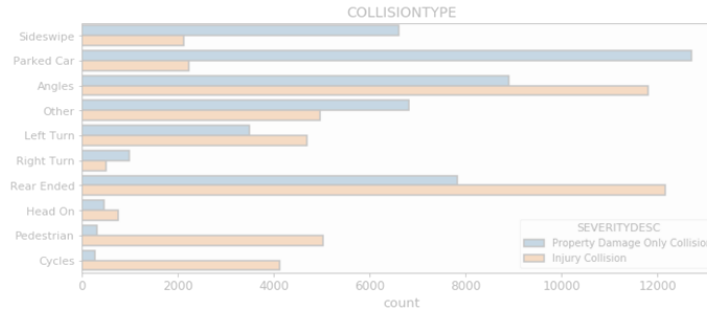
LIGHTCOND

There are some discrepancies within categories for the different incident types. We would therefore expect this attribute may be a useful predictor of the incident type.

For the attributes:

WEATHER, ROADCOND, INATTENTIONIND, SPEEDING, UNDERINFL

There similar counts of each incident type for most of the major categories. These appear to provide less information on the type of incident. This is fairly surprising given that environmental conditions, speeding and being under the influence might be expected to effect the severity of incident but this appears not to be the case.



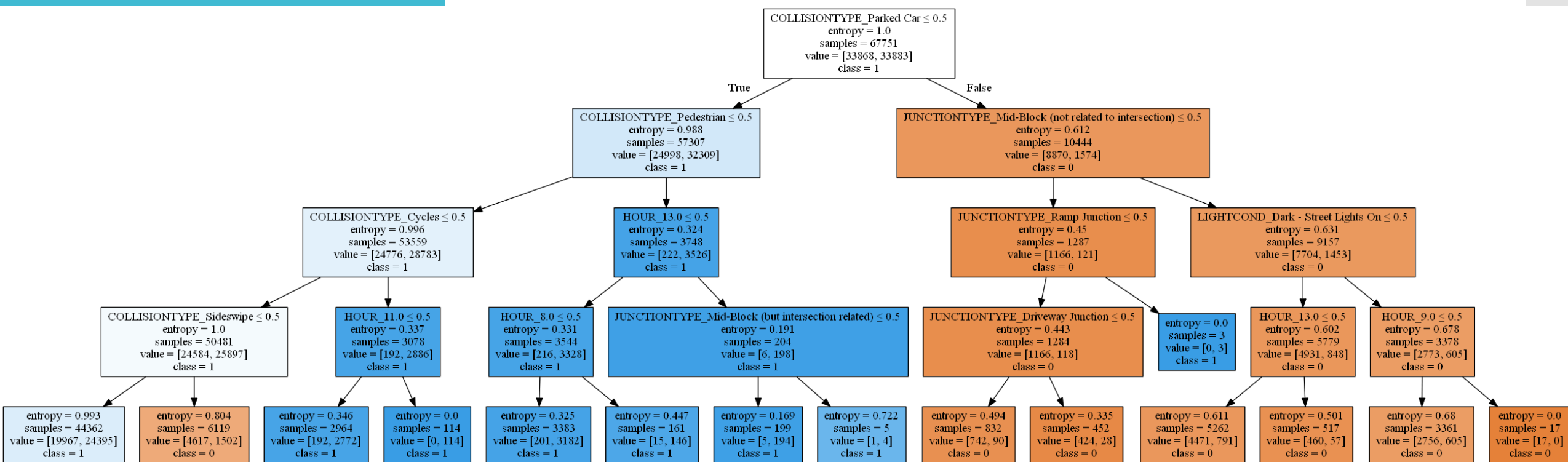
Machine Learning

Based on the earlier analysis of the different features we will limit our analysis to those features which are more predictive. This should increase the efficiency of our ML training algorithms and remove obscuring features.

We convert the categorical attribute features to numerical values using one-hot encoding.

	COLLISIONTYPE_Angles	COLLISIONTYPE_Cycles	COLLISIONTYPE_Head On	COLLISIONTYPE_Left Turn
0	0	0	0	0
1	0	0	0	0
2	1	0	0	0
3	0	0	0	0
4	0	0	0	1

Decision Tree



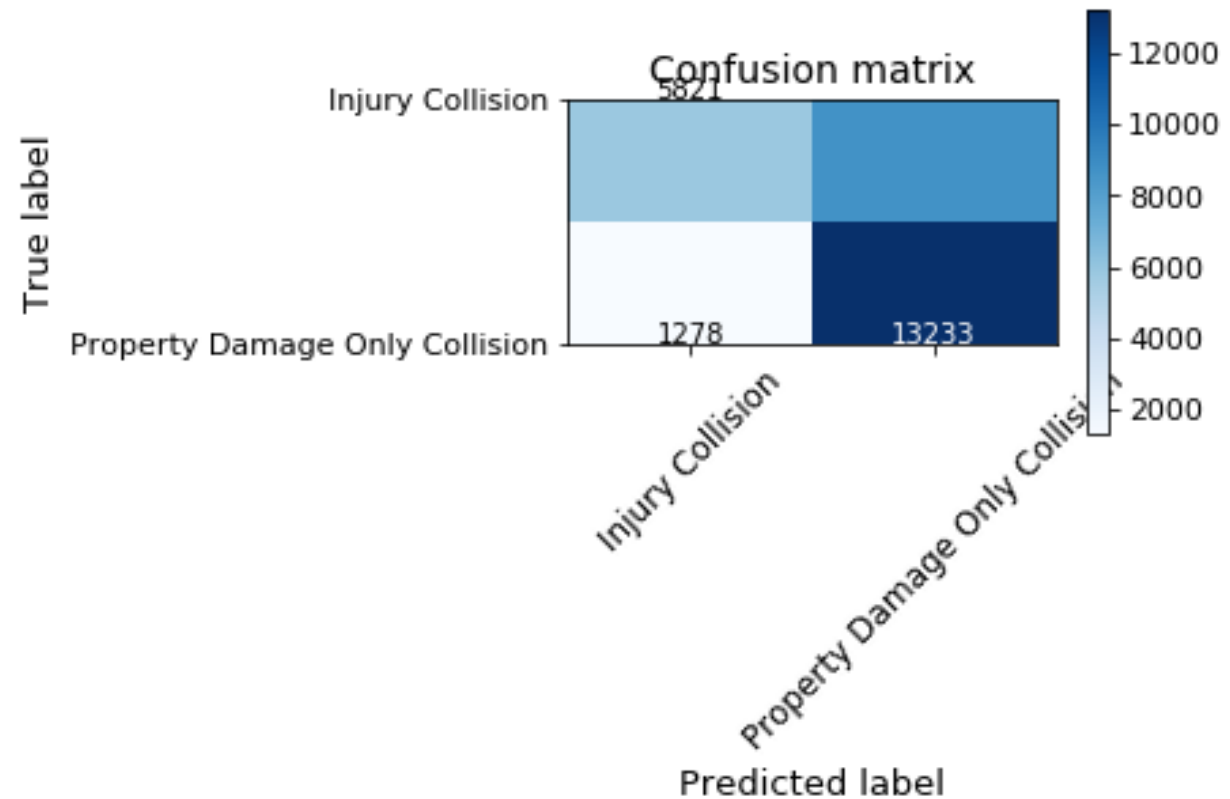
Decision Tree

Accuracy score: 0.66

f1-score: 0.63

Confusion matrix, without normalization

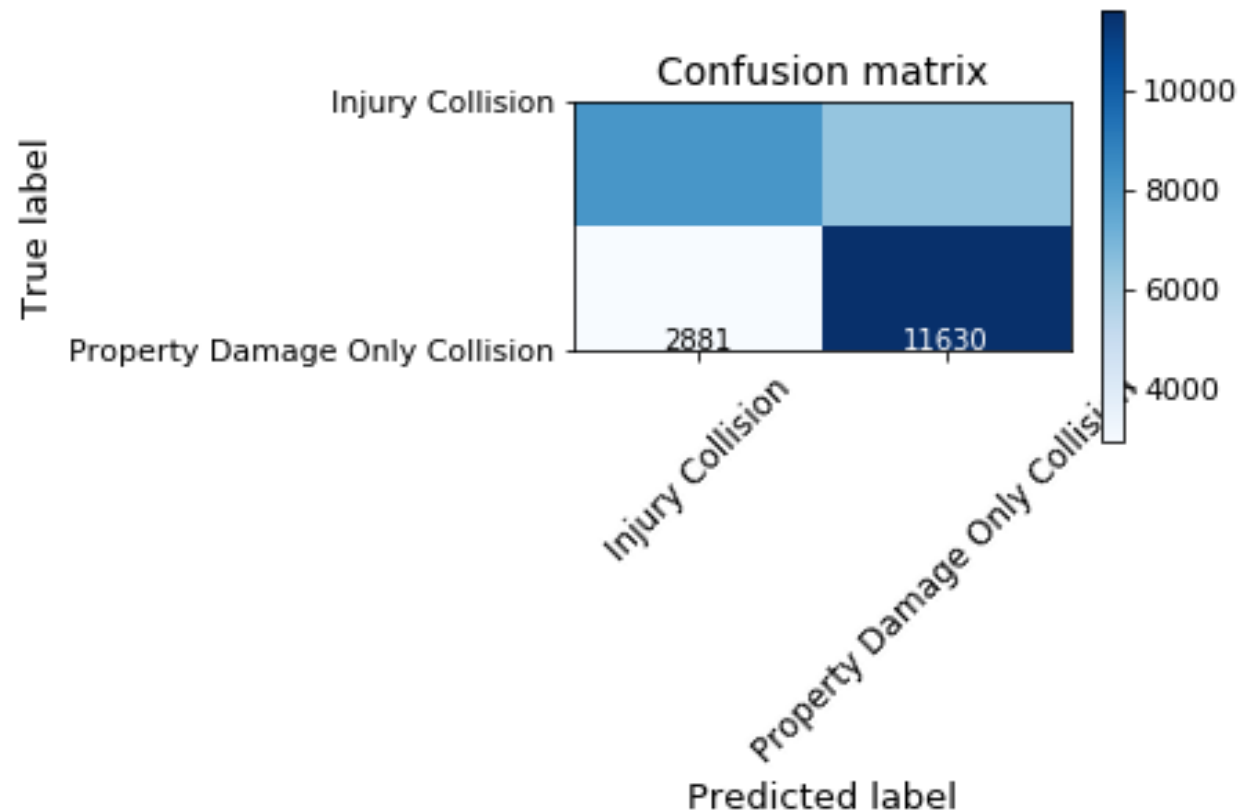
```
[[ 5821  8705]  
 [ 1278 13233]]
```



Logistic Regression

Accuracy score: 0.68
f1-score: 0.68

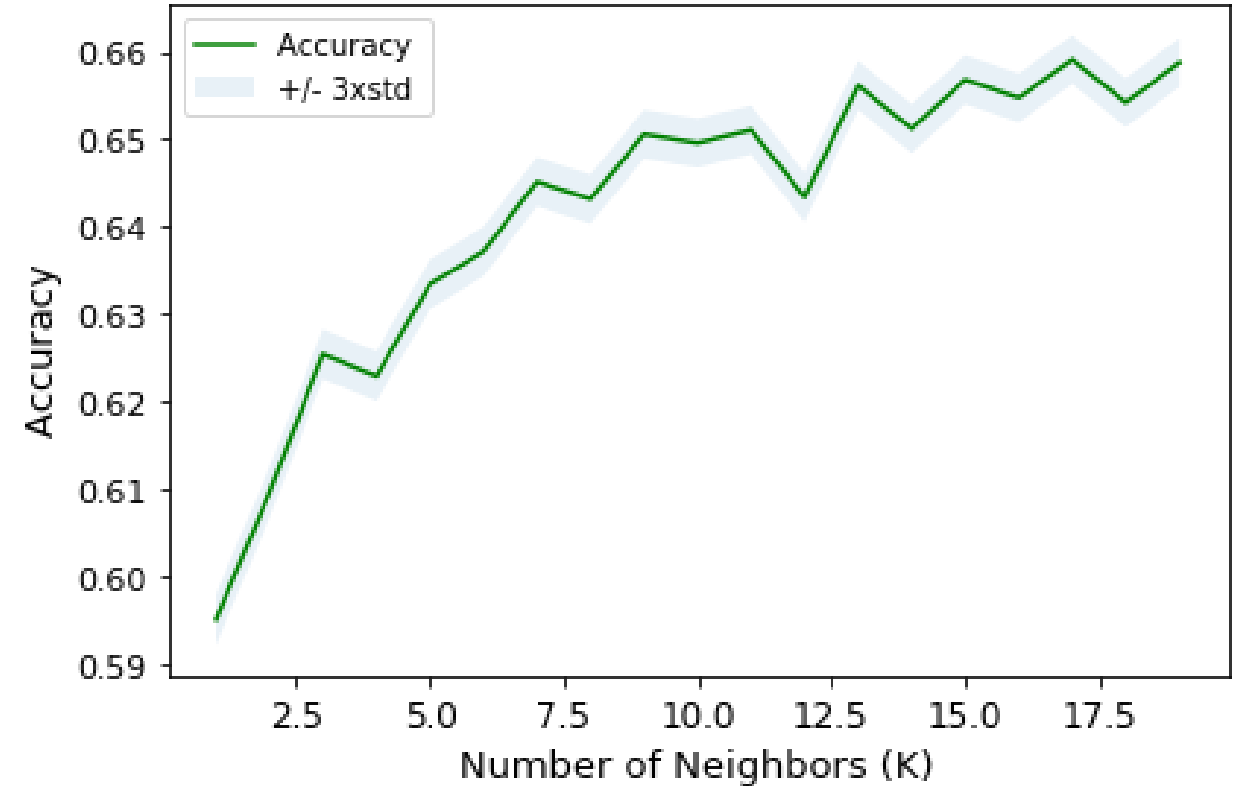
Confusion matrix, without normalization
[[8167 6359]
 [2881 11630]]



KNN

Accuracy of the KNN model appears to plateau after around $k = 12$.

The best value found in the parameter sweep was $k = 17$. This value is used to train the model.



KNN

Accuracy of the KNN model appears to plateau after around $k = 12$.

The best value found in the parameter sweep was $k = 17$. This value is used to train the model.

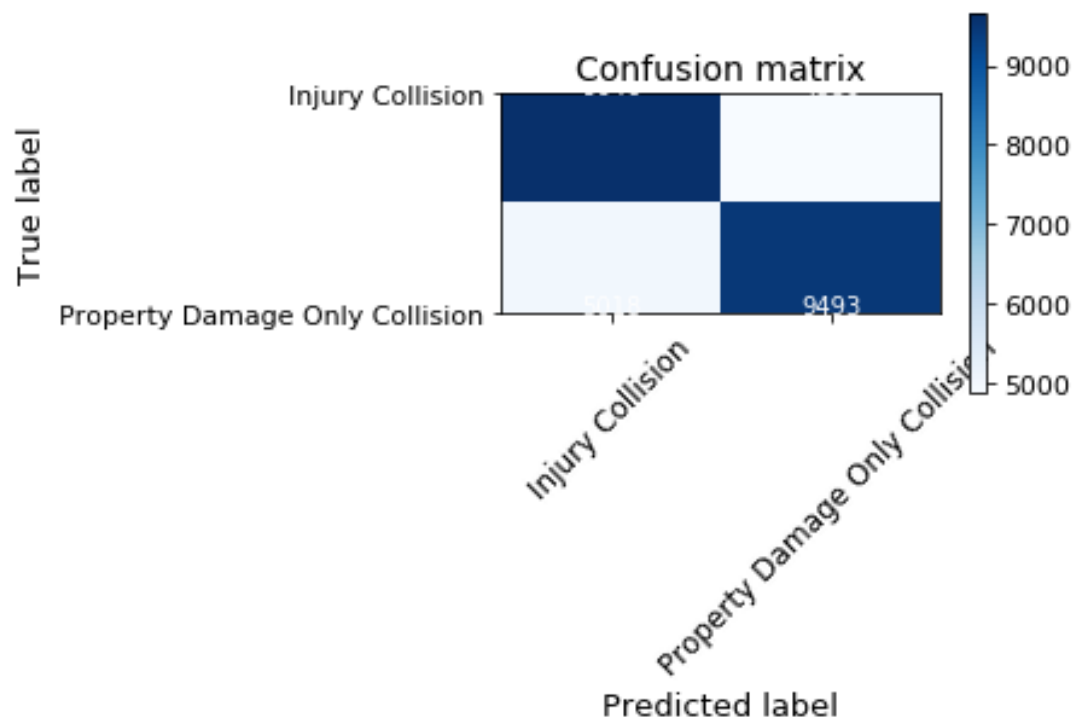
Accuracy score: 0.66

f1-score: 0.66

Confusion matrix, without normalization

```
[[9646 4880]
```

```
[5018 9493]]
```



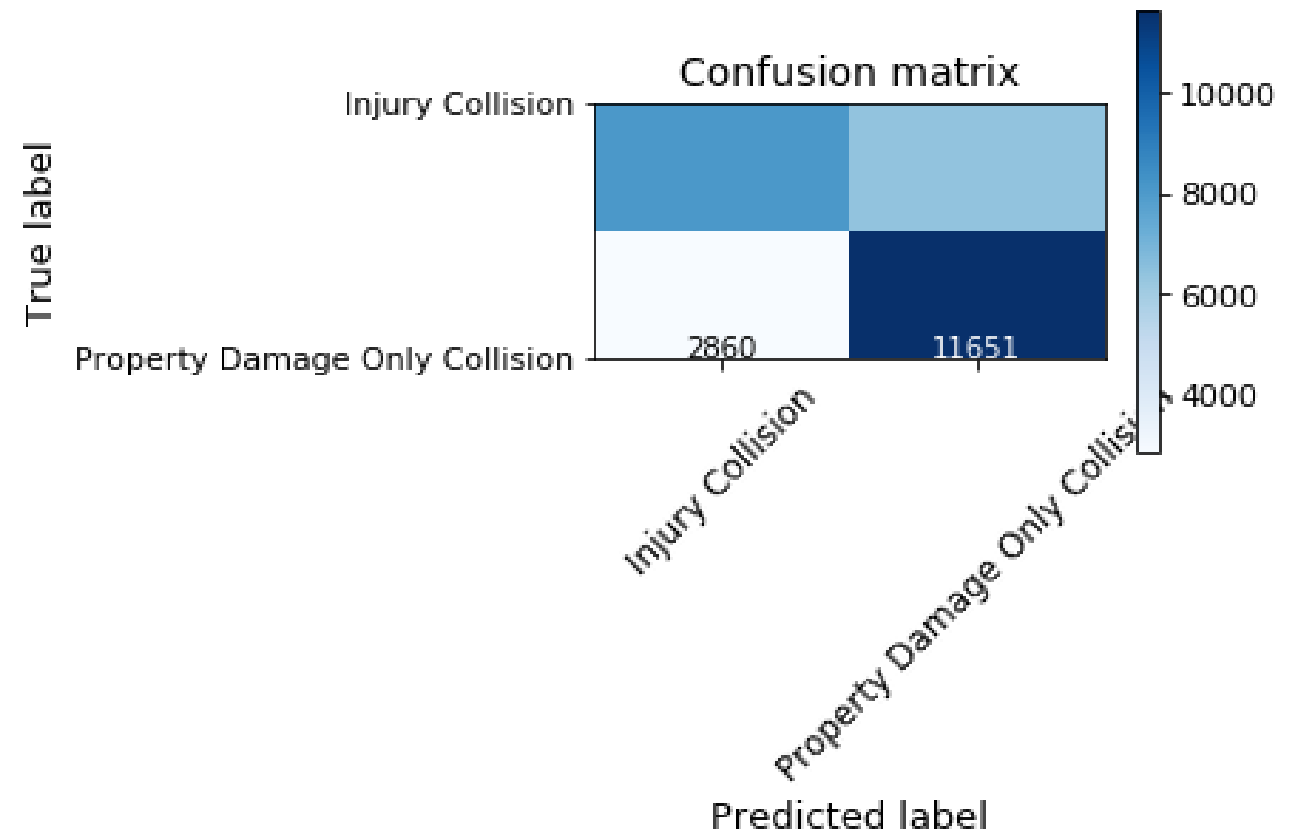
SVM

Accuracy score: 0.68

f1-score: 0.68

Confusion matrix, without normalization

```
[[ 8107  6419]
 [ 2860 11651]]
```



Summary

Decision Tree

Accuracy score: 0.66
f1-score: 0.63

Confusion matrix, without normalization
[[5821 8705]
 [1278 13233]]

The best k = 17
Accuracy score: 0.66
f1-score: 0.66

Confusion matrix, without normalization
[[9646 4880]
 [5018 9493]]

KNN

Logistic Regression

Accuracy score: 0.68
f1-score: 0.68

Confusion matrix, without normalization
[[8167 6359]
 [2881 11630]]

Accuracy score: 0.68
f1-score: 0.68

Confusion matrix, without normalization
[[8107 6419]
 [2860 11651]]

SVM

Conclusions

We have trained several machine learning techniques to classify an incident based on attributes such as the collision type, junction type and time.

The accuracy achieved with each algorithm ≥ 0.66 , compared to a base line 0.5.

The best algorithms in terms of performance metrics were logistic regression and SVM. However the algorithm most successful at correctly identifying the more serious class of incidents (incidents with injuries) was KNN.

We will need to engage with discussions to understand what the most important performance metrics are before deploying the chosen algorithm.