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:

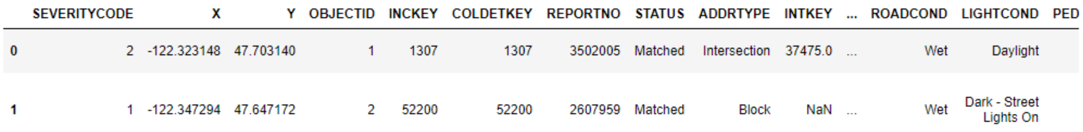
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

Method

1. **Acquire Incident Data**
   The first part of the project will involve uploading data on car incidents in Seattle from the above sources.
2. **Data Cleaning**
   Using data wrangling techniques this data will be processed and the dataset balanced between the different target categories.
3. **Feature Selection**
   We will then explore the predictive capability of each feature and select those which would be most useful to include within our machine learning algorithm.
4. **Machine Learning Training**
   We will use four different machine learning algorithms; Decision Tree, Logistic Regression, KNN and SVM to categorise the type of incident base on the associated features.
5. **Machine Learning Evaluation**
   We will evaluate the performance of the machine learning algorithms based on different metrics including accuracy, f1-score and the confusion matrix.
6. **Discussion**
7. **Conclusions**

Results

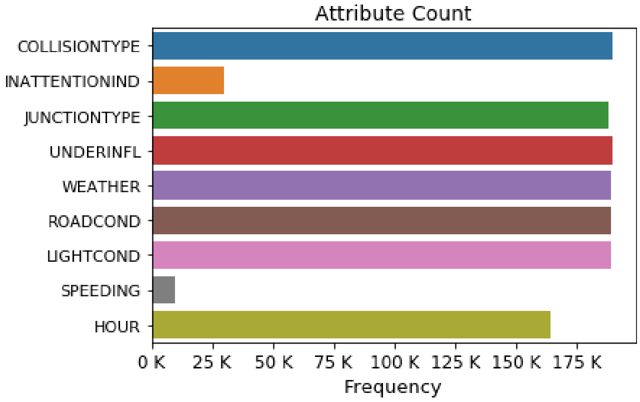
After downloading the Seattle incident data we get the following dataframe:



	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PED
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	

Data Cleaning

From the dataframe we extract the following features/attributes:



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

For the following attribute some instances are labelled as unknown:

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

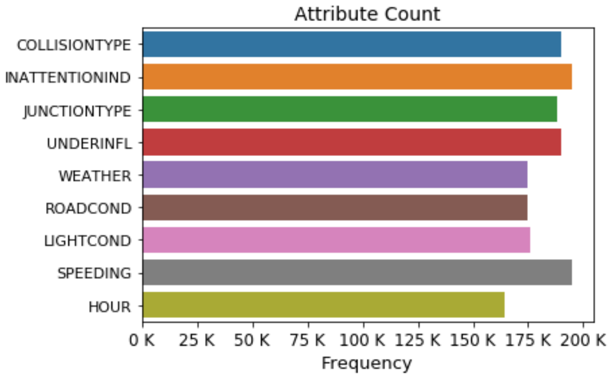
This label adds no information and so will be replaced by nan and dropped from the dataframe.

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.

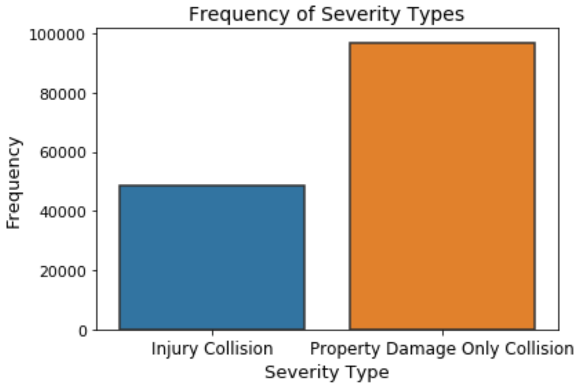
For the UNDERINFL attribute enteries are labelled as either "Y", "1", "N", "0".

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

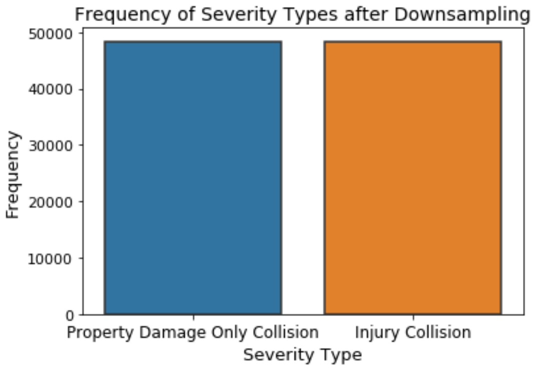
The number of valid entries for each attribute after cleaning:



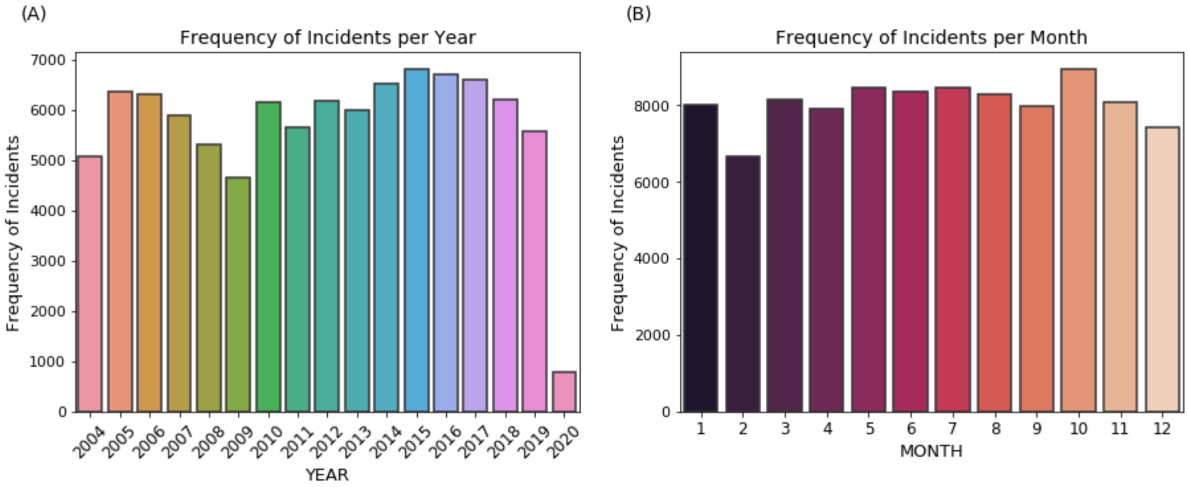
There are [48394, 96951] instances of Severity type ['Injury Collision’, 'Property Damage Only Collision'] respectively. The unbalanced dataset needs to be addressed before applying the machine learning algorithms in order to properly address the predictive capabilities of the machine learning models.



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.

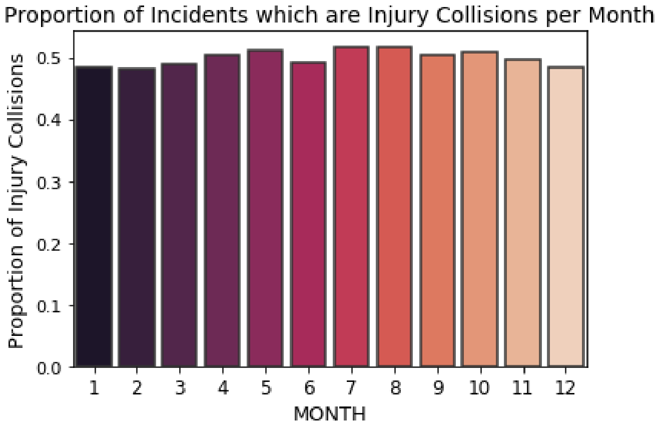


Feature Selection



(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.

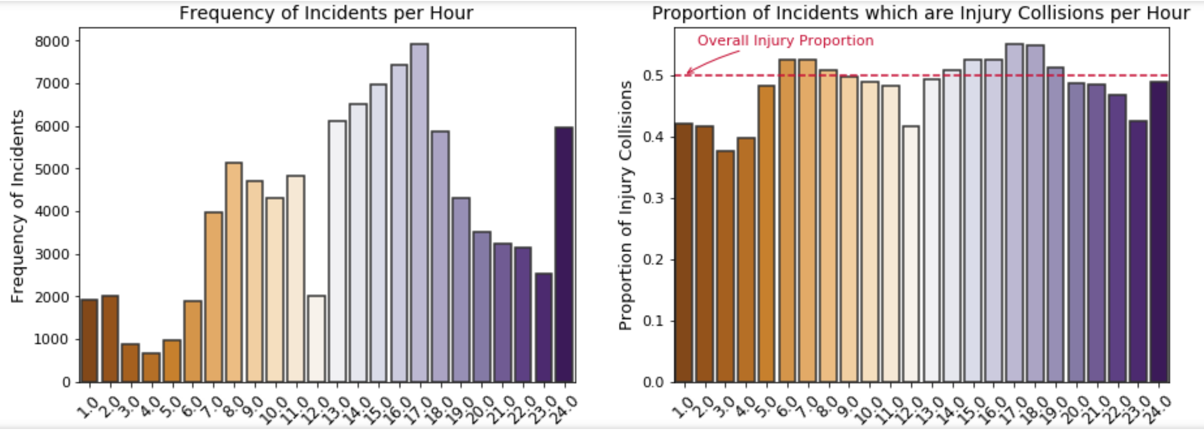
(B) There doesn't appear to be any seasonal dependence in the frequency of incidents. The absence of any seasonal dependence is somewhat surprising given that variable weather conditions, daylight hours etc.



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.



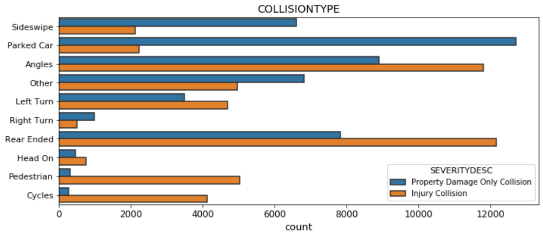
(A) There is evidence that there is some variation throughout the day in the frequency of incidents per hour. There are relatively few incidents in the early hours of the morning, with local peaks around commuting hours (08:00 - 09:00 and 17:00 - 18:00) and a further peak after midnight (24:00 - 01:00).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

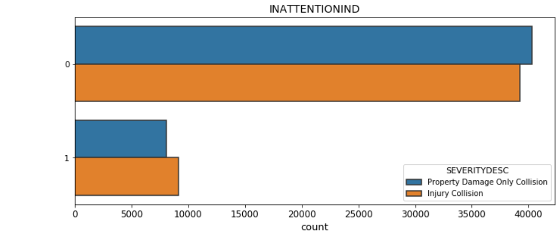
(B) There does appear to be some meaningful variation in the proportion of incidents which are injury collisions. The proportion of injury collisions exceed the average daily proportion between 06:00 and 09:00 and then again between 13:00 and 19:00

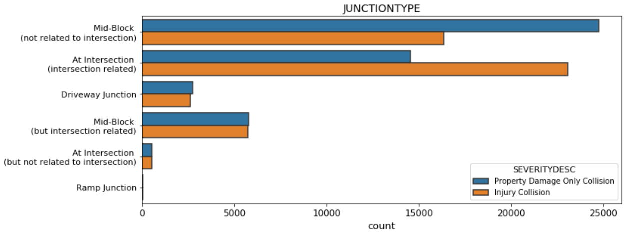
The hour in which an incidence occurs could therefore be a useful predictor of the type of incident

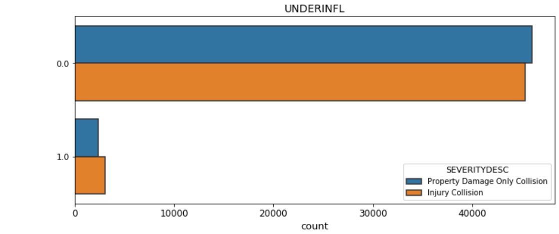
Using the balanced dataset we would expect attributes where there is a discrepancy between the frequency of different incident types to be predictive.

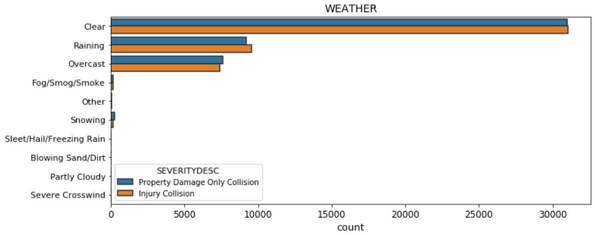
Conversely an approximately even distribution between categories of an attribute suggests that the attribute has little or no predictive capability to distinguish between different types of incidents.

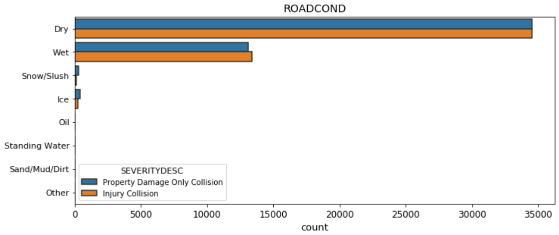


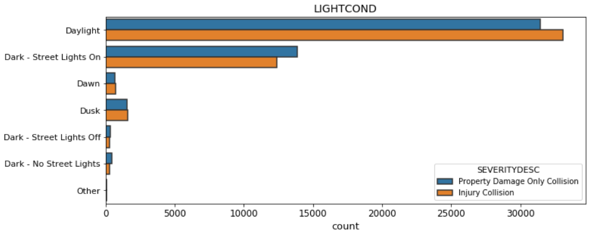


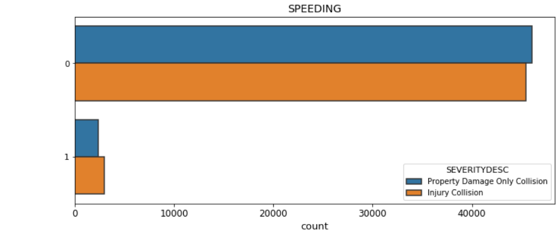












For the attributes:

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:

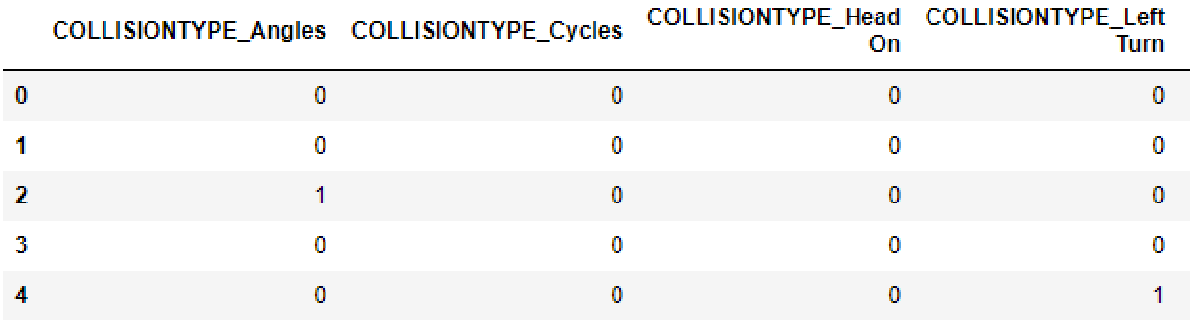
WEATHER, ROADCOND, INATTENTIONIND, SPEEDING, UNDERINFL

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 affect the severity of incident but this appears not to be the case.

Machine Learning Training

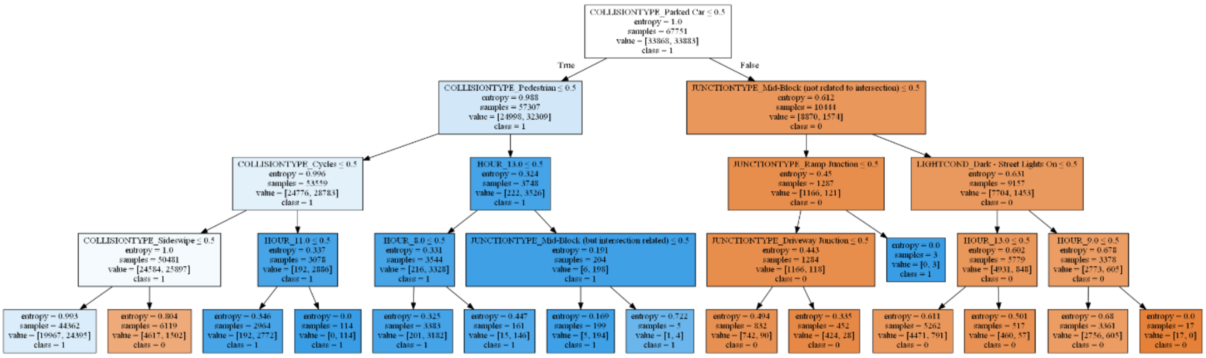
Based on the earlier analysis of the different features we will limit our analysis to those features which are more predictive. This is should increase the efficiency of our ML training algorithms and remove obscuring features.

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



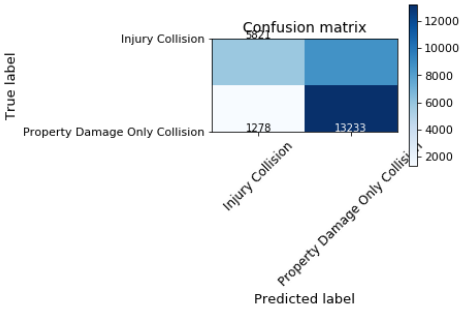
	COLLISIONTYPE_Angles	COLLISIONTYPE_Cycles	COLLISIONTYPE_Head On	COLLISIONTYPE_Left Turn
0	0	0	0	0
1	0	0	0	0
2	1	0	0	0
3	0	0	0	0
4	0	0	0	1

The trained decision tree:

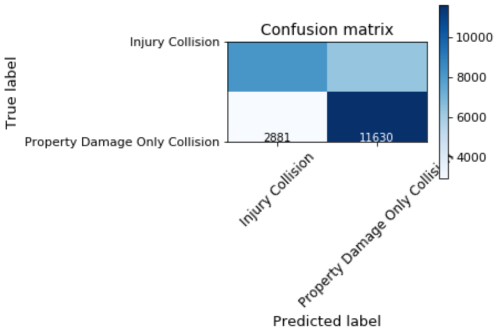

COLLISIONTYPE_Parked Car ≤ 0.5 entropy = 1.0 samples = 67751 value = [33868, 33883] class = 1
True
False
COLLISIONTYPE_Pedestrian ≤ 0.5 entropy = 0.988 samples = 57307 value = [24998, 32309] class = 1
JUNCTIONTYPE_Mid-Block (not related to intersection) ≤ 0.5 entropy = 0.612 samples = 10444 value = [8870, 1574] class = 0
COLLISIONTYPE_Cycles ≤ 0.5 entropy = 0.996 samples = 53559 value = [24776, 28783] class = 1
HOUR_13.0 ≤ 0.5 entropy = 0.324 samples = 3748 value = [222, 3526] class = 1
JUNCTIONTYPE_Ramp Junction ≤ 0.5 entropy = 0.45 samples = 1287 value = [1166, 121] class = 0
LIGHTCOND_Dark - Street Lights On ≤ 0.5 entropy = 0.631 samples = 9157 value = [7704, 1453] class = 0
COLLISIONTYPE_Sideswipe ≤ 0.5 entropy = 1.0 samples = 50481 value = [24584, 25897] class = 1
HOUR_11.0 ≤ 0.5 entropy = 0.337 samples = 3078 value = [192, 2886] class = 1
HOUR_8.0 ≤ 0.5 entropy = 0.331 samples = 3544 value = [216, 3328] class = 1
JUNCTIONTYPE_Mid-Block (but intersection related) ≤ 0.5 entropy = 0.191 samples = 204 value = [6, 198] class = 1
JUNCTIONTYPE_Driveway Junction ≤ 0.5 entropy = 0.443 samples = 1284 value = [1166, 118] class = 0
entropy = 0.0 samples = 3 value = [0, 3] class = 1
HOUR_13.0 ≤ 0.5 entropy = 0.602 samples = 5779 value = [4931, 848] class = 0
HOUR_9.0 ≤ 0.5 entropy = 0.678 samples = 3378 value = [2773, 605] class = 0
entropy = 0.993 samples = 44362 value = [19967, 24395] class = 1
entropy = 0.804 samples = 6119 value = [4617, 1502] class = 0
entropy = 0.346 samples = 2964 value = [192, 2772] class = 1
entropy = 0.0 samples = 114 value = [0, 114] class = 1
entropy = 0.325 samples = 3383 value = [201, 3182] class = 1
entropy = 0.447 samples = 161 value = [15, 146] class = 1
entropy = 0.169 samples = 199 value = [5, 194] class = 1
entropy = 0.722 samples = 5 value = [1, 4] class = 1
entropy = 0.494 samples = 832 value = [742, 90] class = 0
entropy = 0.335 samples = 452 value = [424, 28] class = 0
entropy = 0.611 samples = 5262 value = [4471, 791] class = 0
entropy = 0.501 samples = 517 value = [460, 57] class = 0
entropy = 0.68 samples = 3361 value = [2756, 605] class = 0
entropy = 0.0 samples = 17 value = [17, 0] class = 0

Machine Learning Evaluation

Decision Tree

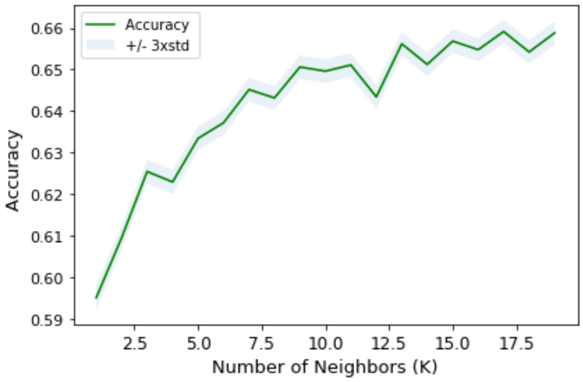


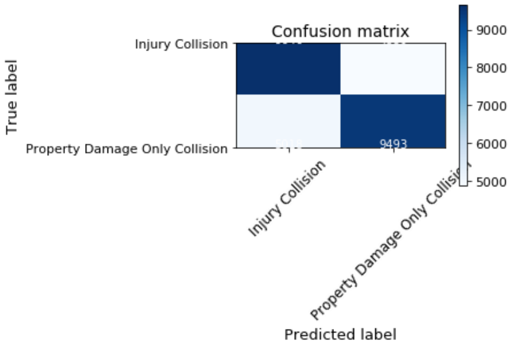
Logistic Regression



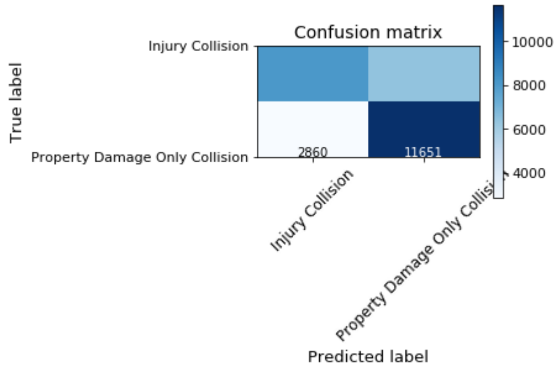
KNN

Optimisation: 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.

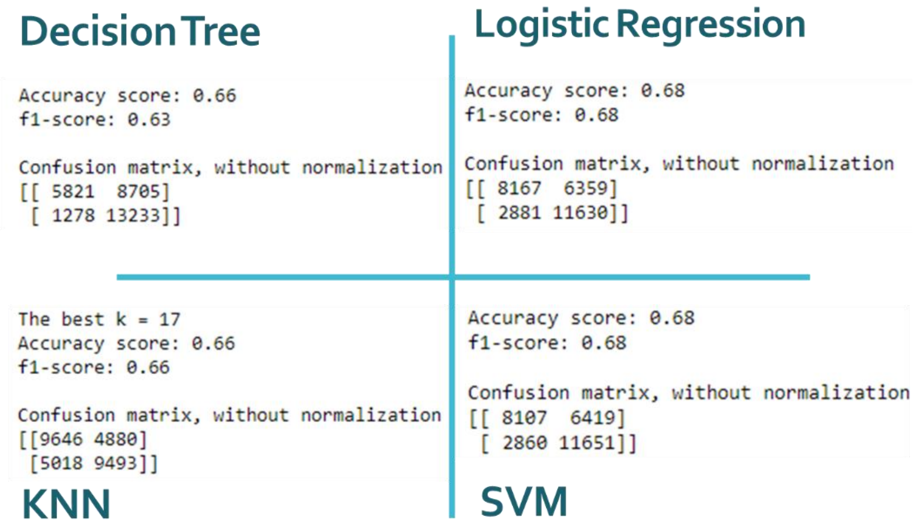




SVM



Discussion


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]]
```

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

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

KNN

```
Accuracy score: 0.68
f1-score: 0.68

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

SVM

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.

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.

Appendix

More information on car incident data is Seattle can be found in:

https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0