Image by Patrick Rodriguez on Wikimedia Commons

Seattle Accident Severity

1. Introduction

With traffic accidents claiming more than 40,000 lives per year in the U.S., I would like to determine if there are actions we can take to avoid accidents or reduce their severity by analyzing data from the Seattle Department of Transportation. The principal stakeholders would be the city of Seattle, as well as anyone who believes injuries are preventable and values their safety.

Seattle was founded in 1851, and as primarily a logging town in the early days, it has grown and evolved into one of the largest cities in North America. With a hilly land footprint of only 84 square miles, the estimated population in 2019 is over 750,000 and a metropolitan population of 4,000,000. This makes Seattle the 15th largest in the United States and the northernmost U.S. city with at least 500,000 people.

After the economic recession between 2007 - 2009, Seattle gained an average of nearly 15,000 residents per year for the next 5 years. When Amazon.com relocated its headquarters to South Lake Union, a construction boom commenced, which resulted in the completion of nearly 10,000 apartments in Seattle in 2017 (more than any previous year and double that of 2016). Although unemployment dropped from 9% to 3.6%, the city found itself bursting at the seams and had the 6th worst rush hour traffic.

Seattle is considered to have a temperate climate. Although it is the cloudiest area of the U.S. (294 days per year) and one of the five rainiest major cities in the U.S. (at least 0.01 inches, 150 days per year), it has significantly less total precipitation than many other U.S. cities because most rainy days are a slight drizzle. However, from November to January, Seattle experiences its heaviest rainfall (half of total precipitation for the year). In November alone, Seattle averages the most rainfall than any other U.S. city of more than 250,000 people. Heavy snow is rare - there have only been 17 days since 1948 where it has snowed 6 inches or more in a single day.

In 2000, Washington State adopted "Target Zero", a plan aimed to end traffic deaths and serious injuries by 2030. The push for safer streets expanded to Seattle's "Vision Zero" in 2015, after collaborating with the state and achieving a 28 percent reduction in fatalities and serious

injuries on Aurora Avenue North. Originally implemented in Sweden in the 1990's, Vision Zero has proved successful across Europe, and many cities across the U.S. have signed on to the idea that even just one death caused by these accidents is unacceptable and preventable. This study will hopefully reveal what, if any, measures we can take as individuals and municipalities to make travel in Seattle safer.

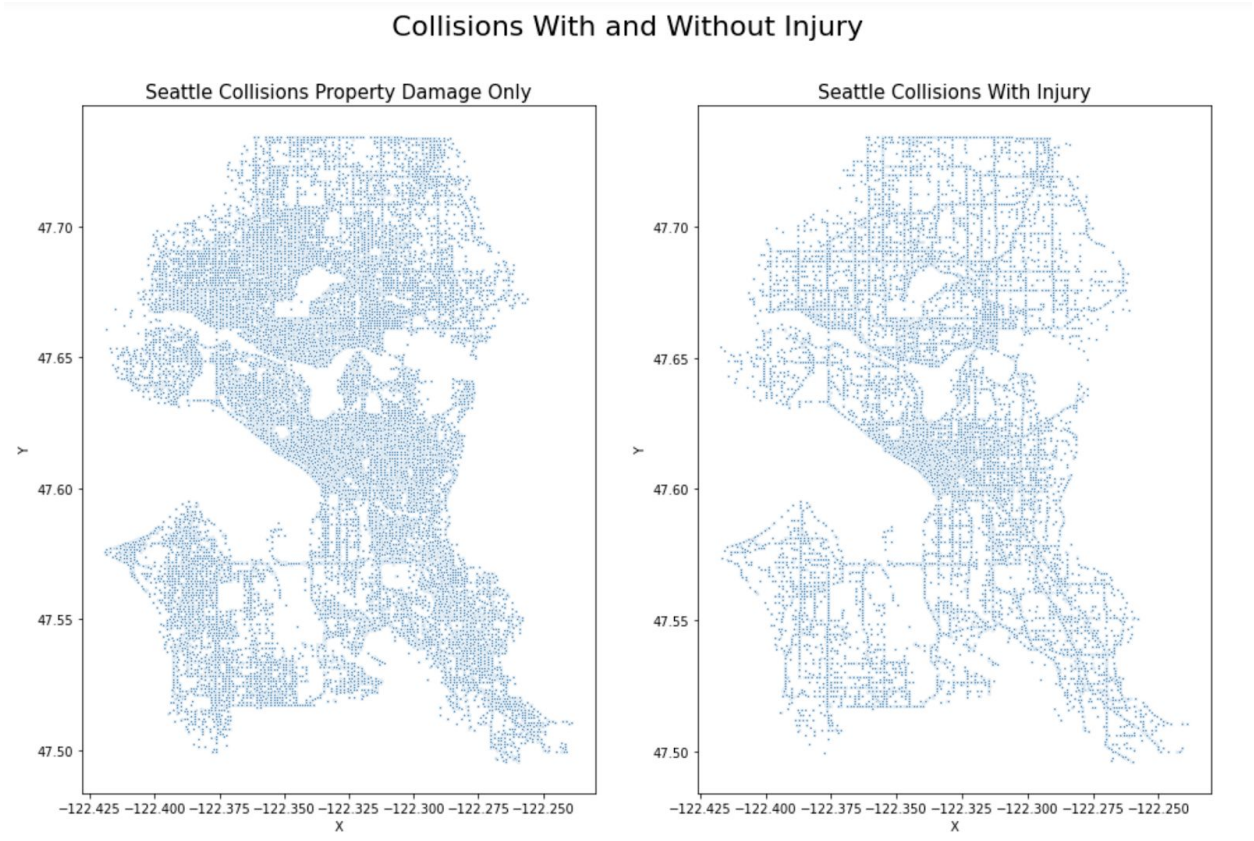
2. Data

The data used for this study was obtained from Seattle's Department of Transportation.

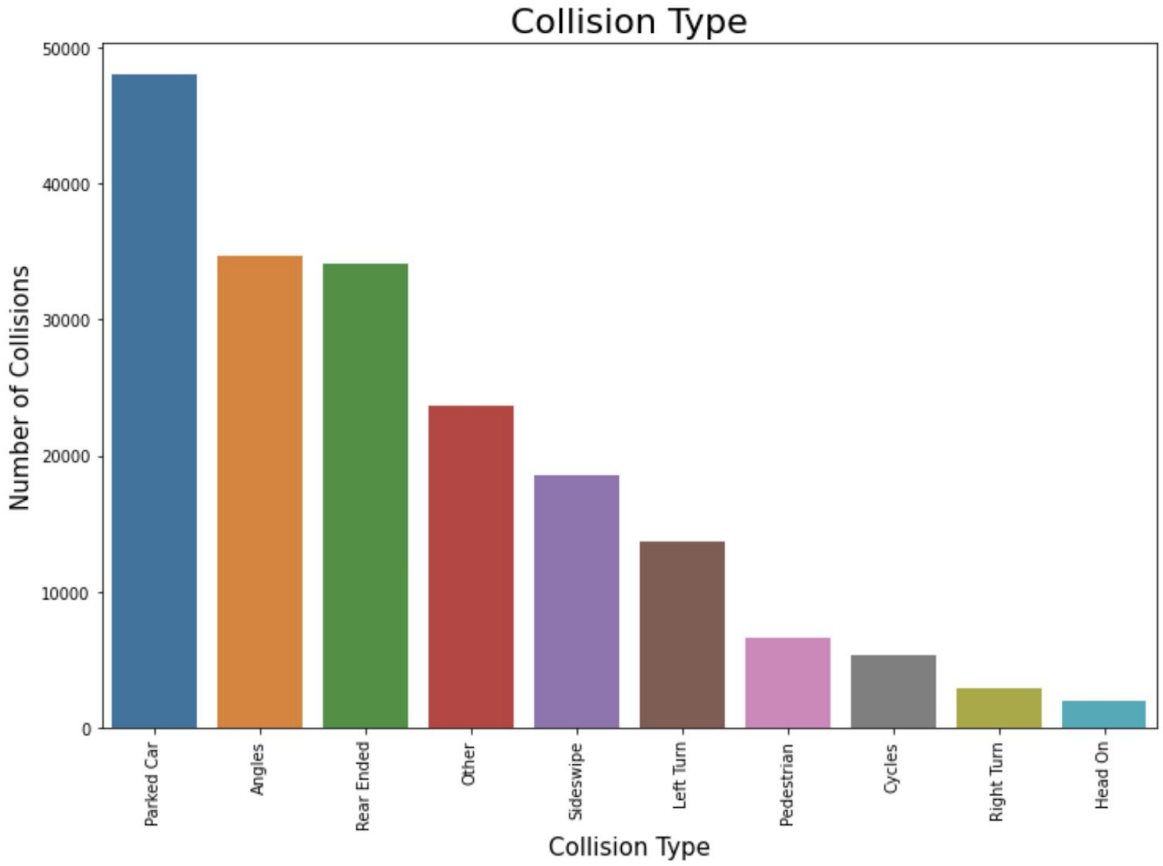
The data used is from Seattle's Collision GIS (Geographic Information System), a computer system used for capturing, storing and displaying data related to positions on Earth's surface. The data is from 2004 to the present and contains various features such as location, the severity of the collision, number of vehicles/cyclists/pedestrians involved, date/time of incident, weather, road conditions and more. There are almost 200,000 collisions in the dataset and 38 features. While some of the features won't be useful or have many missing values, the ones that I will explore in more detail will be:

- Severity Code - this will be the target that we'll compare the features' impact on.
- Severity Description - description of the severity codes.
- X and Y values (coordinates) - are there areas where collisions are more concentrated?
- Address Type - alley, block or intersection of collision.
- Collision Type - 10 types of collisions such as parked car, angles, rear end, pedestrian, etc.
- Person Count - # of people involved in collision.
- Pedestrian Count - # of pedestrians involved in collision.
- Cyclist Count - # of cyclists involved in collision.
- Vehicle count - # of vehicles involved in collision.
- Date/Time - are number of or severity of collisions more likely to occur on certain days or times?
- Junction Type - 7 types describing collision at intersection, mid-block, driveway and whether collision is related to intersection.
- Seattle Collision Code - Seattle codes to describe each collision.
- Seattle Collision Description - description of Seattle collision codes.
- Under Influence - was alcohol or drugs involved?
- Weather - do more collisions occur because of adverse weather?
- Road Conditions - do more collisions occur because of adverse road conditions?
- Light Conditions - do more collisions occur because of adverse light conditions?
- State Collision Code - 84 codes the state uses to describe each collision.
- State Collision Description - description of state collision codes.
- Hit Parked Car - was a parked car involved in the collision?

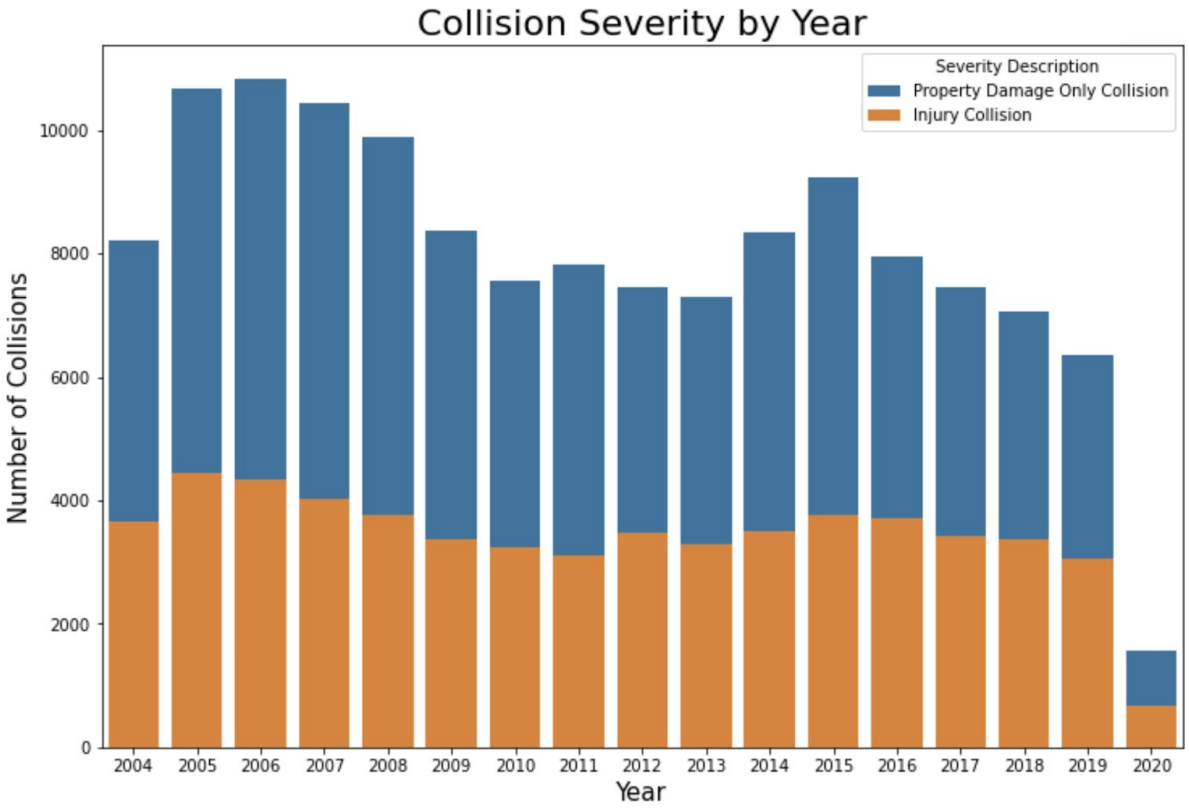
In the following section I will use graphs to gain insight into the data, and decide which features will be useful when predicting injuries when these collisions occur.



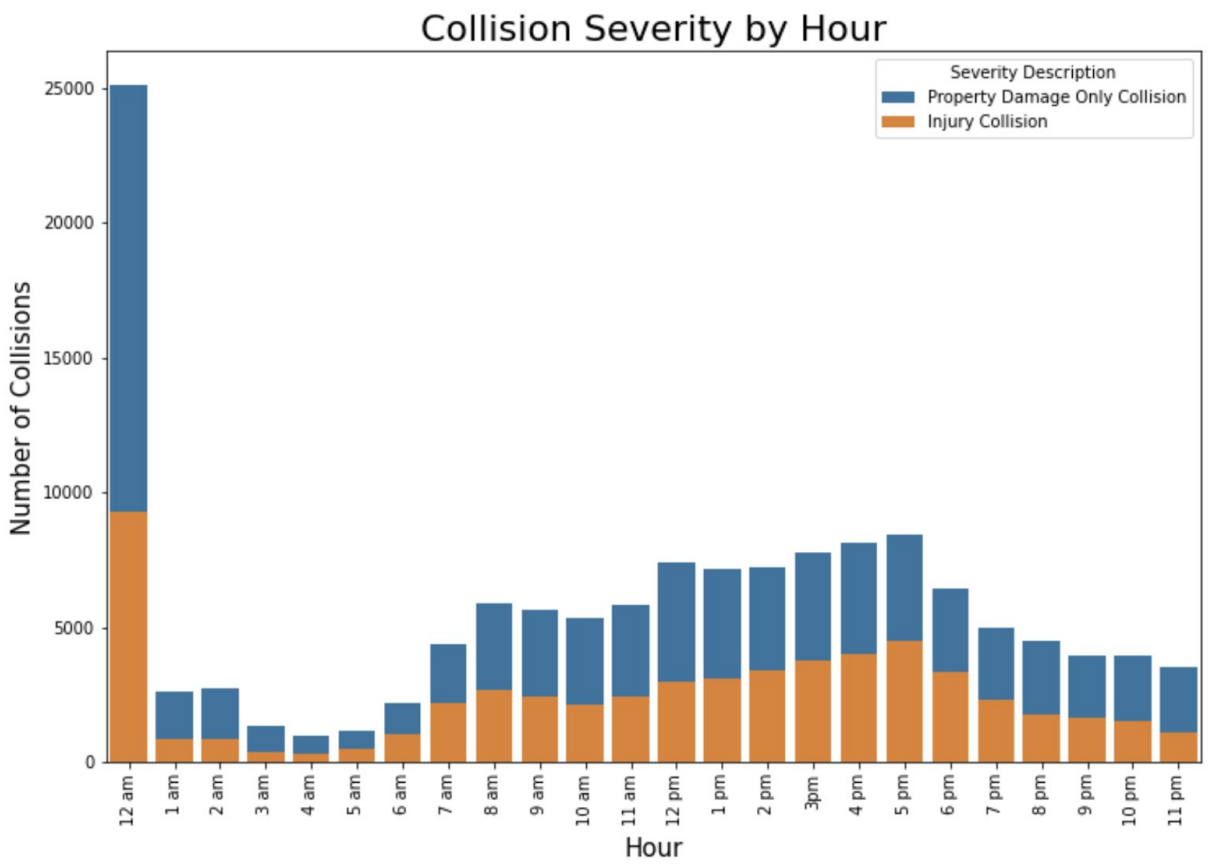
Maps of Seattle with points representing collisions - left map property damage only, right map injury. Injuries appear mostly in the downtown area and major roads with higher speed limits.



Many collisions occur with parked cars, rear ends and angles. Those could be from inattention or poor infrastructure. There are also many collisions with pedestrians.



The two downward yearly trends may be correlated with the recession in 2007-2009, and Seattle's adoption of Vision Zero in 2015. Collisions decreased even though there has been a huge population growth spurt in 2017.



It appears as though collisions go up throughout the day culminating around 5 pm. The other thing to note is the number of collisions (including injury) between midnight and 1 am. These high volumes of collisions may correlate to people getting off work and establishments' closing times.

After looking at the graphs, here are some observations about the data:

- The collisions involving injury tend to happen in and around the downtown area and major highways.
- Most collision types are with parked cars. Angles and rear end collisions are also common.
- It is rare for a collision to involve pedestrians or cyclists but collisions with pedestrians is still high.
- Although the number of collisions have fluctuated over the last 16 years, it appears to be trending down
- Accidents, including collisions that involve injury, spike between midnight and 1 am. 3-5 pm also has a higher chance for collisions.

Seattle does seem to be making some progress towards their Vision Zero plan, but have far to go. Some of the most relevant features to predict collision severity may be location, collision type and time of day.

3. Methodology

In this section I prepared the data for different machine learning classifiers. This includes missing data, restructuring data and normalizing the data. Then I fed the data into 5 different classifiers to get statistics on the results.

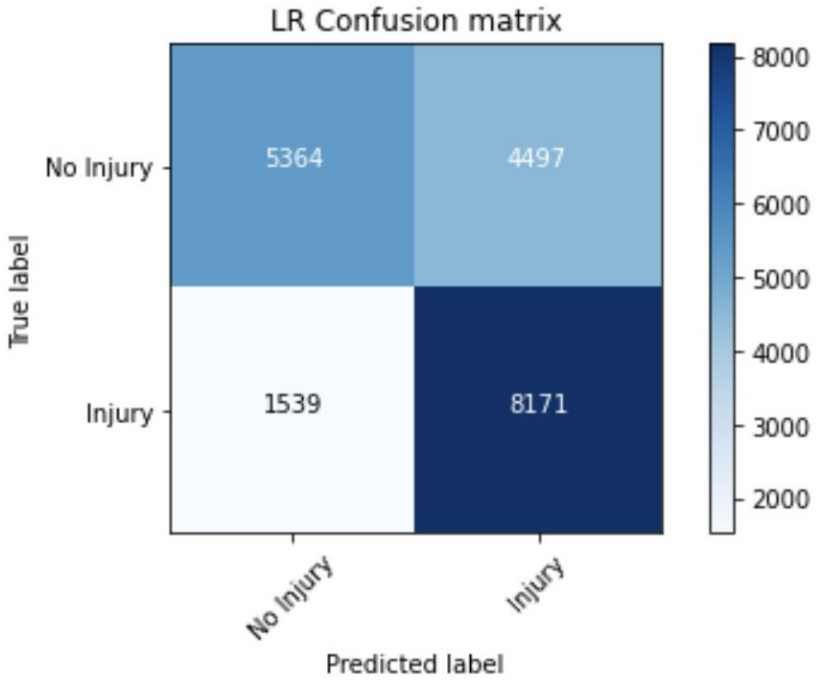
Since we have a fairly large dataset, and the number of incidents with missing data is fairly low, I dropped the rows with missing values. I then went through each feature to make sure it is of a type our algorithm will take, and make sure the values made sense. For some, they needed to be encoded as a '0/1' for each value. Once the data was fully prepared, I made a copy for use at the end of the project to feed into our final model.

The next step is to balance the data. We have far more collisions without injury than we do for injury, so I undersampled the majority class (class 1). I chose this because we still end up with close to 100,000 rows of clean, balanced data. Next is feature selection. Most of the data was compatible for the algorithm except for data such as coordinates or transportation department codes. More work could be done here to refine the selection of features or adding new ones, such as a clustering algorithm (DBSCAN) on the coordinates to produce a 'High Risk Area' feature.

I then split the data into training data (80%) and testing data (20%). In the final step, we standardize the data to feed into the classifiers. Standardizing data scales it to unit variance and is used so that all features' scales are equal. I chose the 5 most used classifiers to see what each model would produce.

3.1 Logistic Regression Classifier

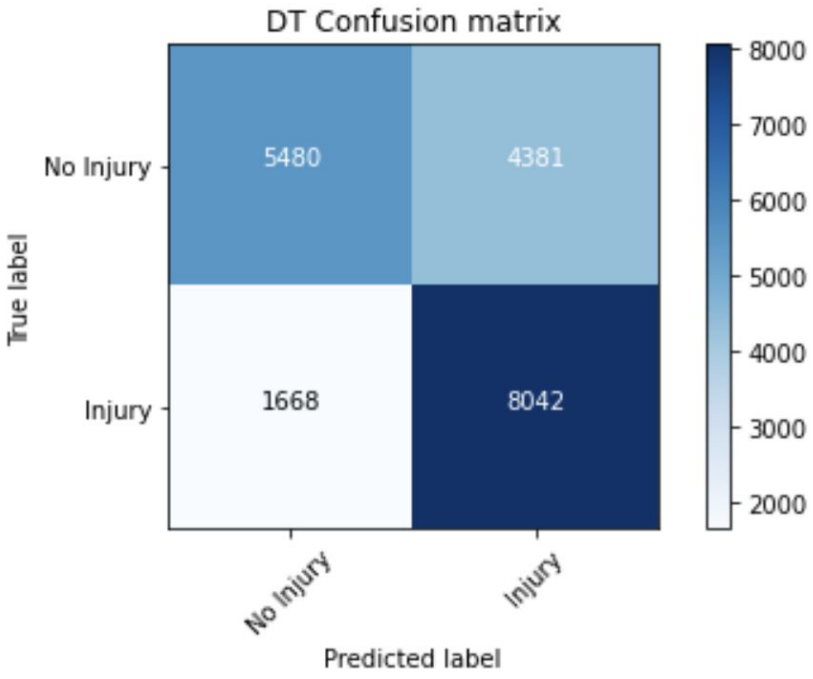
```
Confusion matrix, without normalization
```



The Logistic Regression model correctly identified over 8000 of the injury class. It didn't do so well identifying the no injury class, incorrectly predicting almost 4500 as injuries. It achieved an accuracy of 69.16%.

3.2 Decision Tree Classifier

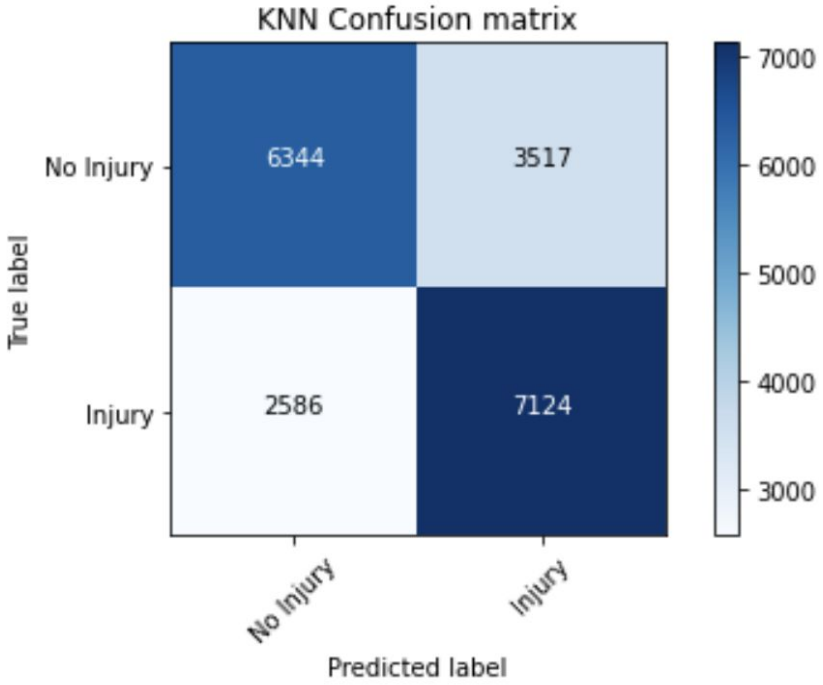
```
Confusion matrix, without normalization
```



The Decision Tree model has a similar accuracy of 69.09%. It is also more sensitive to the injury class while incorrectly predicting almost 4400 of the no injury class.

3.3 K-Nearest Neighbors (KNN) Classifier

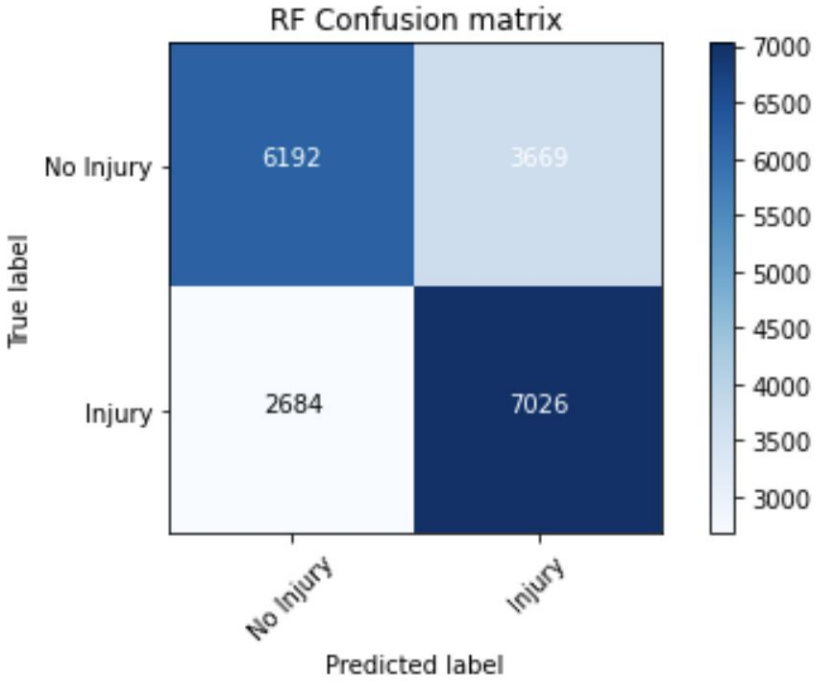
```
Confusion matrix, without normalization
```



Although the KNN classifier scored slightly lower in accuracy at 68.82%, it has classified many more of the no injury class while maintaining a high number of correct injury class predictions.

3.4 Random Forest Classifier

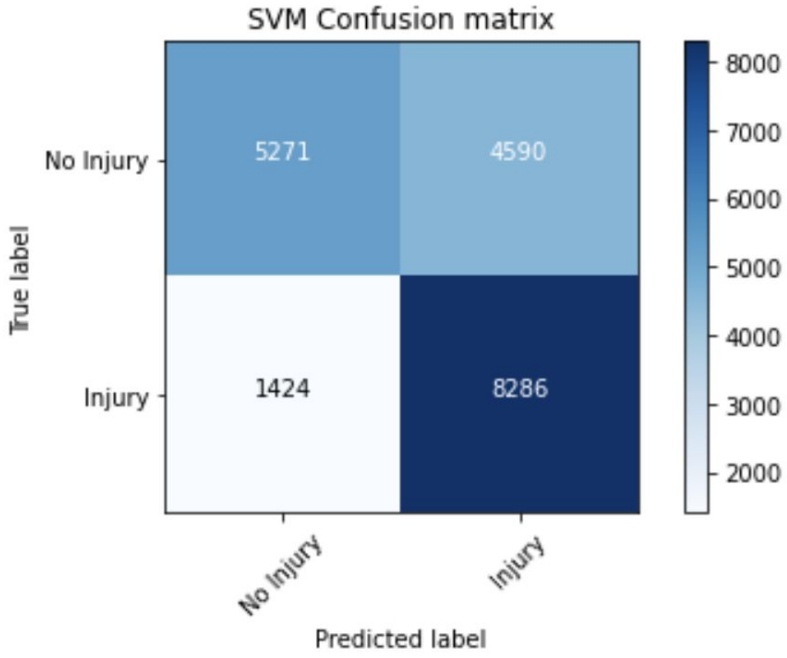
```
Confusion matrix, without normalization
```



The Random Forest classifier also is sensitive to both classes but has performed worse than KNN. It's accuracy is only 67.54%.

3.5 Support Vector Machine (SVM) Classifier

```
Confusion matrix, without normalization
```

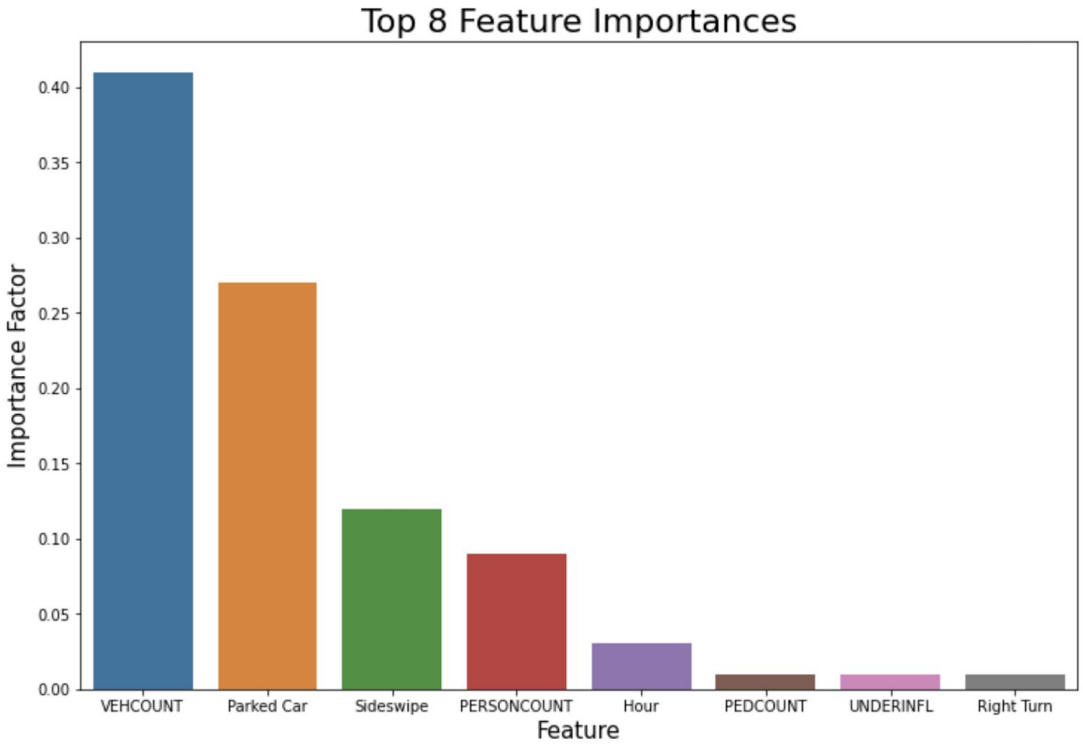


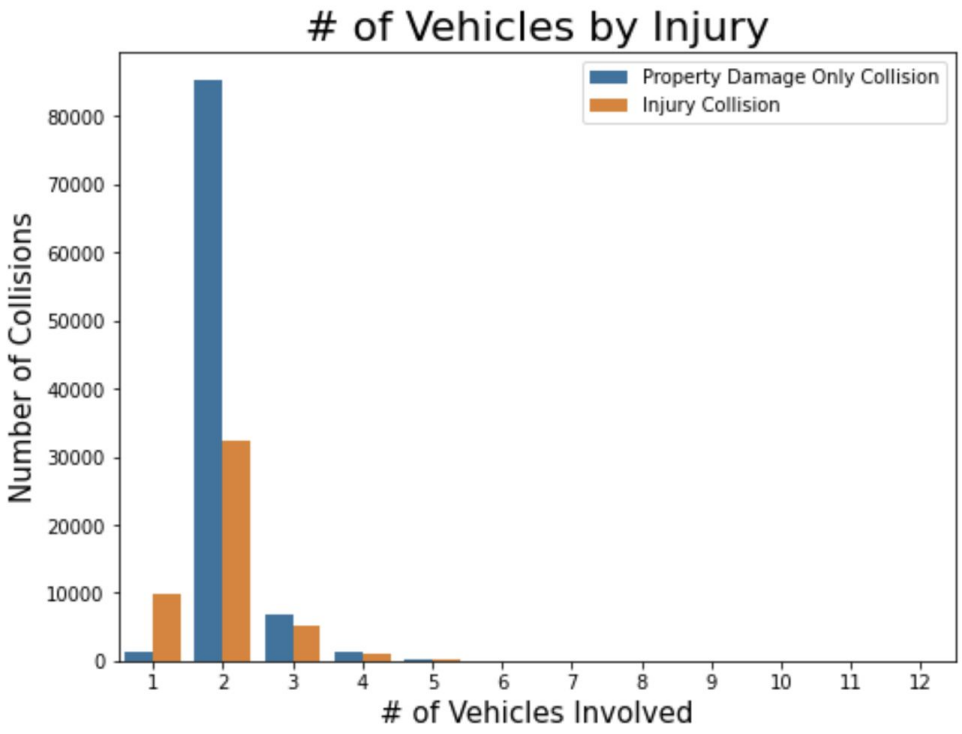
The SVM classifier was the most sensitive to the injury class by classifying 85% of those correctly. The overall accuracy is 69.27%, the highest of all of them, but it is not sensitive to the no injury class.

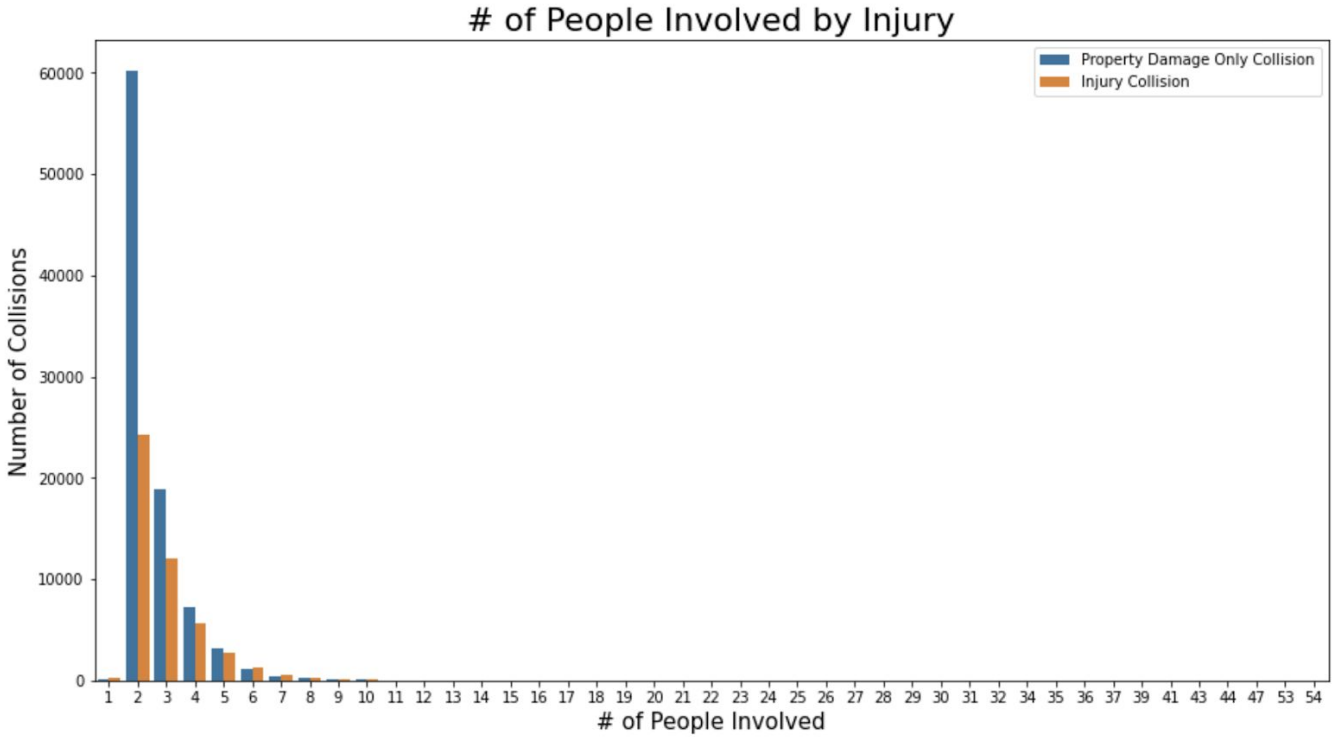
4. Results

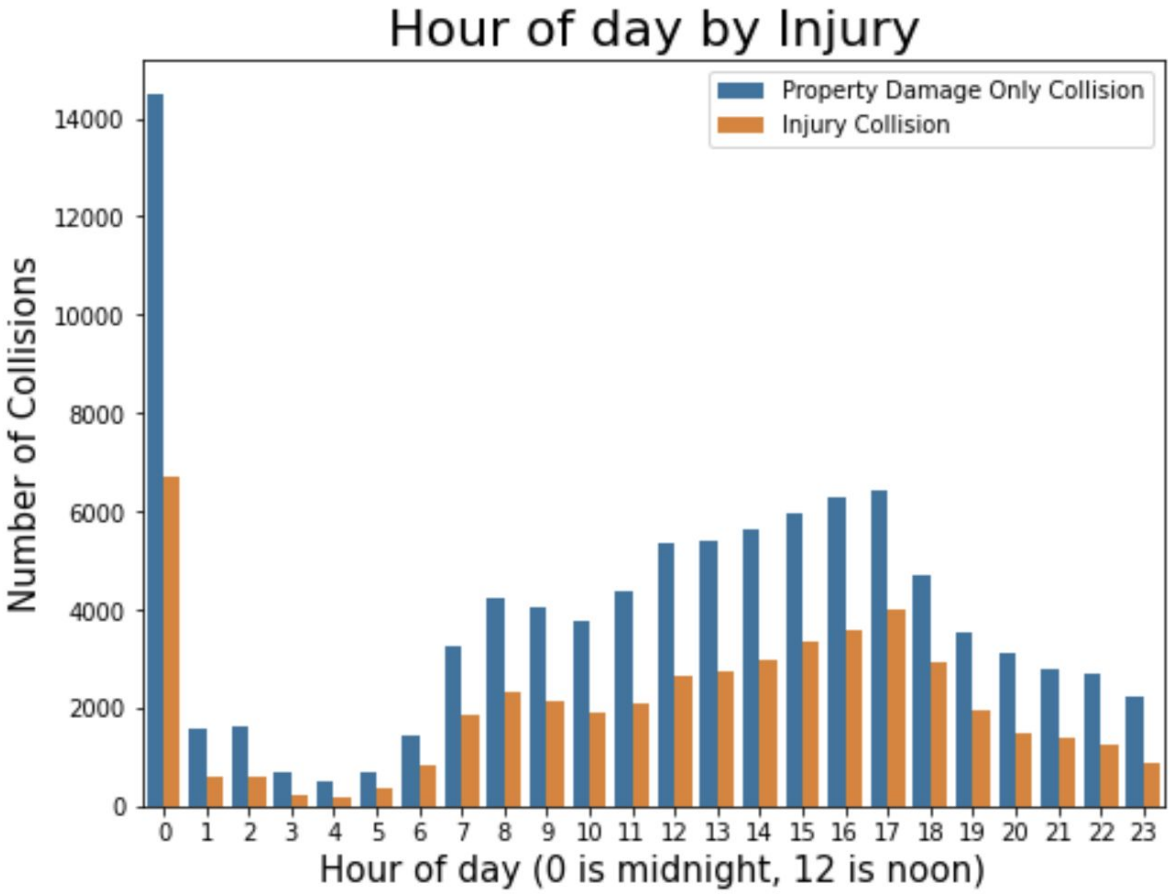
	Decision Tree	K-Nearest Neighbors	Logistic Regression	Random Forest	Support Vector Machine
Jaccard Score	0.6909	0.6882	0.6916	0.6754	0.6927
F1 Score	0.6852	0.6876	0.6848	0.6747	0.6849
Injury Class F1 Score	0.7267	0.7001	0.7303	0.6887	0.7337
Injury Class Recall Score	0.8282	0.7337	0.8415	0.7236	0.8533

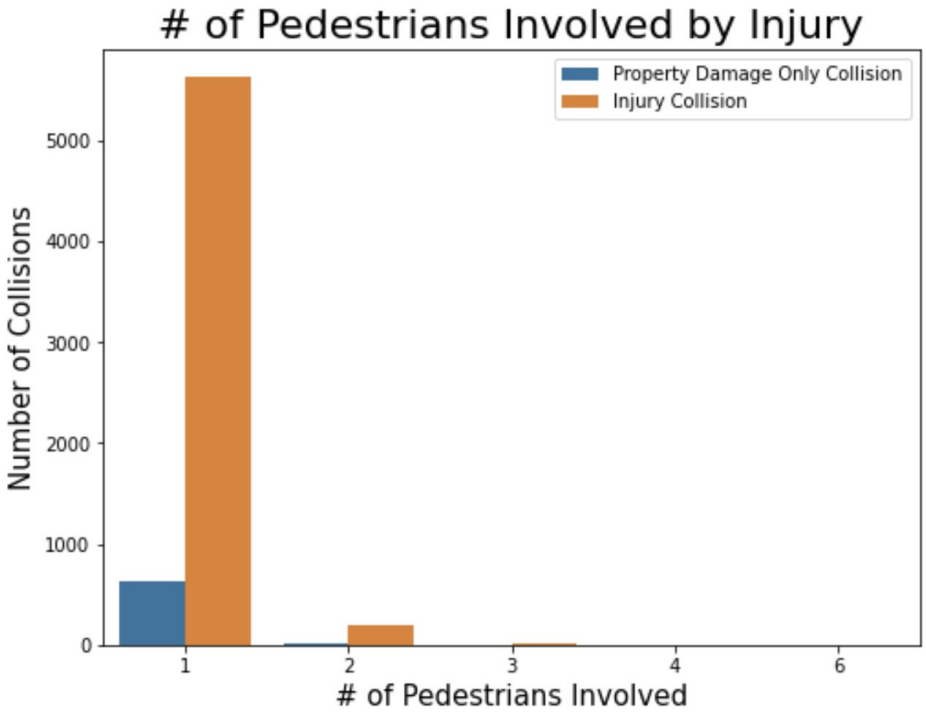
For accuracy (jaccard score) of the classifiers, Support Vector Machine has the best score. The F1 score, or the weighted average of the precision and recall, is highest with the K-Nearest Neighbors classifier. I also included the F1 Score and the recall score for just the injury class. These are also important scores if it's important to predict more of the injuries correctly. For a higher recall score in the injury class, the classifier will also mis-classify more of the 'property damage only' categories as 'injuries'. This may be desired so that a larger portion of the injuries will get a quicker response from an ambulance, however, we wouldn't want to waste ambulance resources to false alarms, when a real injury is happening elsewhere. I will go with K-Nearest Neighbors since it scored the highest F1 score and it is sensitive to both the 'property damage only' class and the 'injury' class. However, in order to get insight into feature importance, I will also run the data through the Decision Tree classifier. Then we can see which features are the biggest factor in determining whether or not injury occurs when there's a collision. Additionally, I will graph some features with respect to injury for detail.

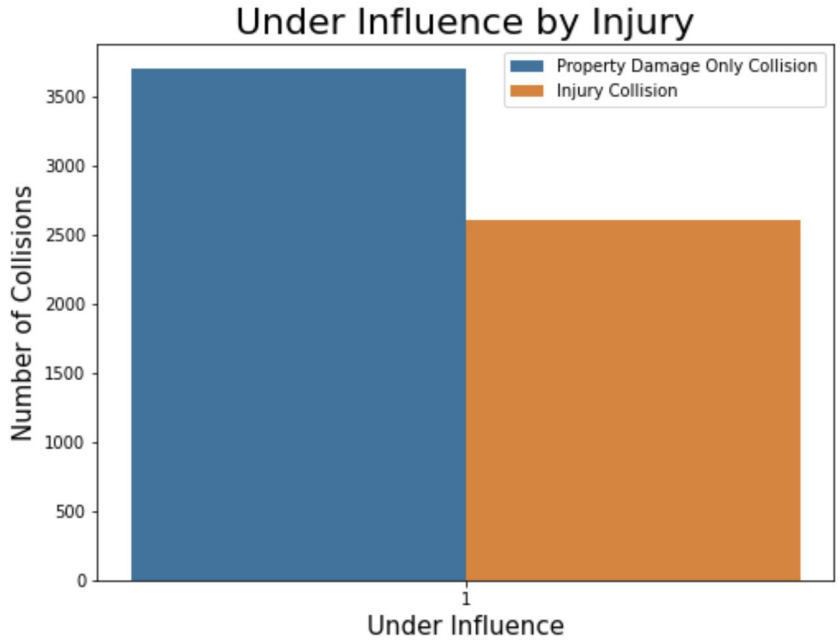


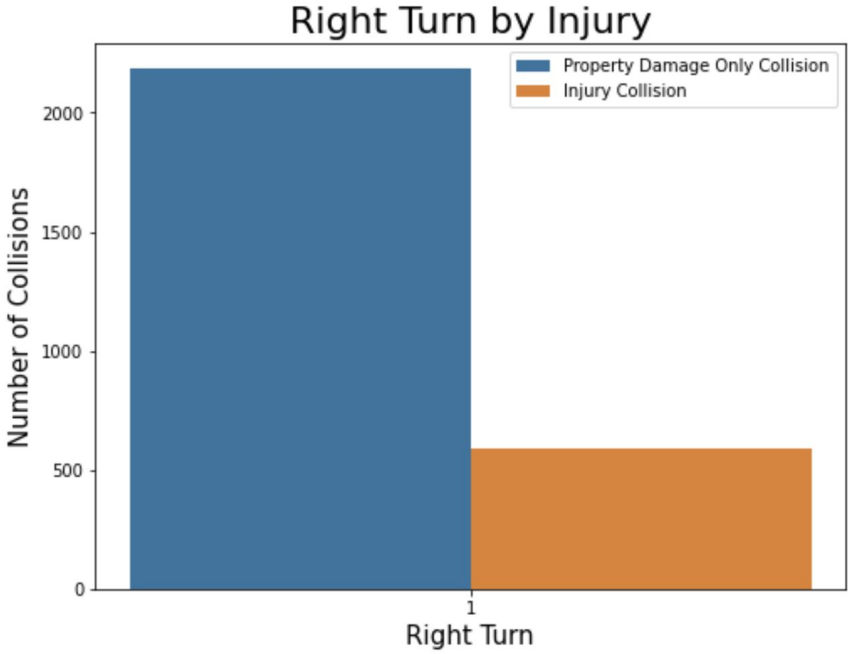

# of Vehicles by Injury
Property Damage Only Collision
Injury Collision
Number of Collisions
0
10000
20000
30000
40000
50000
60000
70000
80000
1
2
3
4
5
6
7
8
9
10
11
12
# of Vehicles Involved


# of People Involved by Injury
Property Damage Only Collision
Injury Collision
Number of Collisions
0
10000
20000
30000
40000
50000
60000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 34 35 36 37 39 41 43 44 47 53 54
# of People Involved


Hour of day by Injury
Property Damage Only Collision
Injury Collision
Number of Collisions
0
2000
4000
6000
8000
10000
12000
14000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of day (0 is midnight, 12 is noon)


# of Pedestrians Involved by Injury
Property Damage Only Collision
Injury Collision
Number of Collisions
0
1000
2000
3000
4000
5000
1 2 3 4 6
# of Pedestrians Involved


Under Influence by Injury
Property Damage Only Collision
Injury Collision
Number of Collisions
0
500
1000
1500
2000
2500
3000
3500
1
Under Influence


Right Turn by Injury
Property Damage Only Collision
Injury Collision
Number of Collisions
0
500
1000
1500
2000
1
Right Turn

5. Discussion

I listed the top 8 features by importance in correctly predicting injury from collision. These were the only ones to score above .001 and after looking at each feature, there are a few things to note. Whereas, 'parked car' and 'sideswipe' categories result in few injuries, 'number of vehicles involved' tells another story. When the number of vehicles is three or more, the chance of injury is nearly 50% and if the collision involves only 1 vehicle, the chance for injury far outweighs property damage only. Remember, this is the feature that is most predictive for the model. When you take this information and pair it with the feature, 'number of pedestrians involved', it's clear that any collision with a pedestrian is likely to cause injury. Chance of injury also goes up when the number of people involved goes up. Collisions between midnight and 1 am are frequent and result in many injuries. Collisions also tend to climb throughout the day culminating at 5 pm. When drugs or alcohol is involved, the chance for injury is nearly 50%. And 'right turn' also has an uncomfortably high number of injuries. There are definitely ways to reduce contact between automobiles and humans, such as giving pedestrians a head start before allowing vehicles to proceed. Lowering speed limits and no turn on red could be other ways. As individuals driving these automobiles, being sober and respecting pedestrians' right of way need to be top priority.

6. Conclusion

I believe there are things that can be done, from an individual and societal point of view, to reduce the number of collisions and their severity. As individuals, we can be more aware of ours and others driving habits in certain conditions, such as Friday at 5 pm, or at midnight when the bars close down. Downtown and highways are also areas where injury tends to occur. Driving only when sober and giving pedestrians right of way will reduce the risk for injury. As a city, implementing no turn on red and letting pedestrians go first may be ways to improve injury rates. Lowering speed limits in the areas where injuries occur most often may also be another way. Infrastructure that allows pedestrians to move more safely can not only save lives, but encourage more people to walk and use public transportation. Fewer vehicles may then have an impact on safety and the well-being of Seattle's population.