IBM's APPLIED DATA SCIENCE CAPSTONE PROJECT

PREDICTING THE SEVERITY OF A CAR COLLISION IN SEATTLE

GIBRÁN MENDOZA MAGAÑA

Project to get the:

IBM Data Science Professional Certificate

Date: November 13, 2020

INTRODUCTION

The number of car crashes in the US have fallen since 2005; nonetheless, motor vehicle crashes are still the leading cause of death for Americans under 30 [1]. Furthermore, the US Department of Transportation estimated that in 2010 car collisions took 242 billion USD off the American economy [2] due to lost of productivity, medical costs, legal costs, emergency services, insurance, property damage, etc.

According to Rolison et al. [3], the main factors behind a car collision are: failure to look properly, loss of control, failure to judge another person's path or speed, traveling too fast for conditions, slippery roads, and carelessness. Knowing the causes behind a car collision could help to make sure they do not happen in the first place. Preventing vehicle collisions could not only help to avoid the loss of many young American lives (and the emotional burden that brings to their loved ones), but also save billions of dollars to the US economy. In other words, preventing vehicle collisions will improve the quality of life of millions of people.

The objective of this project is to use machine learning to develop a model that can warn the driver for the possibility of them getting into a car accident, as well as its severity, given the weather and road conditions. This could help the driver pay more attention when driving in areas where the probability of an accident is higher, or even change the route to a safer one. Additional to drivers, other stakeholders that might be interested in this project are: emergency services, city planners, insurance companies, etc.

Data acquisition

The City of Seattle has an "Open Data Program" in which it took the initiative to release all the data the city generates into its "Open Data Portal" [4]. The objective of this program is to increase the quality of life of its residents [5], and in this portal different files on transportation, public safety, finance, community, etc can be found.

The portal has a file called "Collisions" [6] with information containing all types of collisions that have taken place since 2004 up to September 15th, 2020. The data has 40 different attributes describing the weather, lighting, and road conditions at the time of the accident; the driver conditions like speeding, driving under influence, being inattentive, etc; and the location of the accident with X, Y coordinates. More information on these attributes can be found on the "Attribute Information" (Metadata) pdf file[7].

Furthermore, information on Seattle's streets was gathered as well in order to know how wide a street is, whether it is part of a main artery in a city or not, as well as the type of

neighborhood (downtown, industrial, residential) where the collisions took place. More information on this file can be found in the Seattle data portal [8].

Data cleaning

The Street and Collisions data downloaded from the Seattle Open Portal was combined into one table, then redundant variables were eliminated, i.e. two different codes where use to describe the same collision, one was used by the Seattle Department of Transportation (SDOT), and the other one was used by the State of Washington. Only the SDOT collision code was used in this project in order to avoid redundancy; furthermore, the SDOT code is more complete as it indicates the object colliding as well as the object being collided.

Timestamps were separated into new attributes that would describe separately the year, month, day of the week and hour of the collision in order to inquire about the seasonality of the data. Additionally, coordinates were also transformed from EPSG 2926 used by the State of Washington to the world wide standard UTM (EPSG 4326/WGS84) using the pyproj library in python.

Exploratory Data Analysis

Seasonality

An analysis was made on the behaviour of collisions throughout the year, the week, and the day as a first approach in finding patterns in the collisions data. The following figure was created:

Figure 1 presents a boxplot of the collisions registered every month, from 2004 to mid September 2020. The "Number of Collisions vs Month" box plot shows that, in average, October registers more collisions than any other month while February registers the least. It can also be observed that November, December and January have the highest spread in yearly reported monthly collisions.

The aforementioned behaviour might be due to weather patterns changing across the year. In subplot 3: "Inches of rain vs Month", it can be seen that there is a correlation between the amount of rainfall and the number of collisions per month, but only during certain months. The following table shows the Pearson correlation values between amount of rainfall and number of collisions per year period.

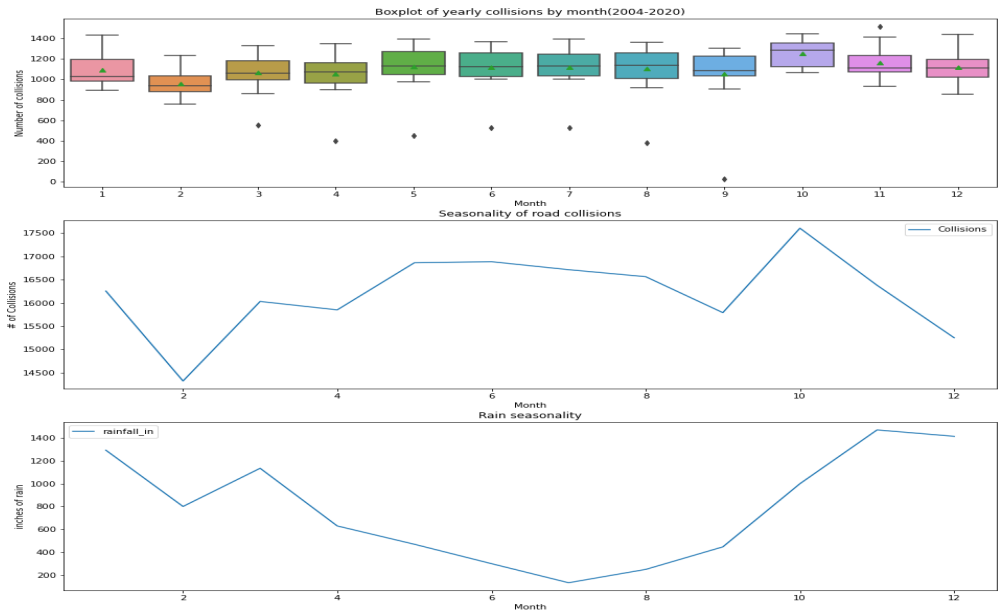

Figure1: Rain patterns vs collisions

Table 1: Correlation between rain and collisions	
Year period	Correlation
January - April	0.977
April - September	-0.709
September - December	0.286

With the help of Figure 1 and Table 1, the following can be observed regarding the correlation between rainfall levels and collisions in Seattle:

1. In September and October the correlation is positive, but from October to December it is negative.

2. The correlation is very strong between January and April.

3. The correlation is strong and negative between between April and September.

It can be seen that the seasonality of precipitations do not completely explain the variation of collisions in Seattle throughout the year. This means that other factors contribute more to

the amount of collisions, based on the subplot "Yearly collisions by month". It seams that during winter, when there are less people on the streets, there are less accidents.

Figure 2 shows the number of collisions by day of the week and by hour of the day. It can be seen that the day with less collisions is Sunday, and the hour with the least collisions is 4 a.m. On the other side the day with most collisions is Friday, and the hour with the most collisions is 5 p.m. This is consistent with the theory that the more people outside, the more likely an accident will happen, regardless of the weather.

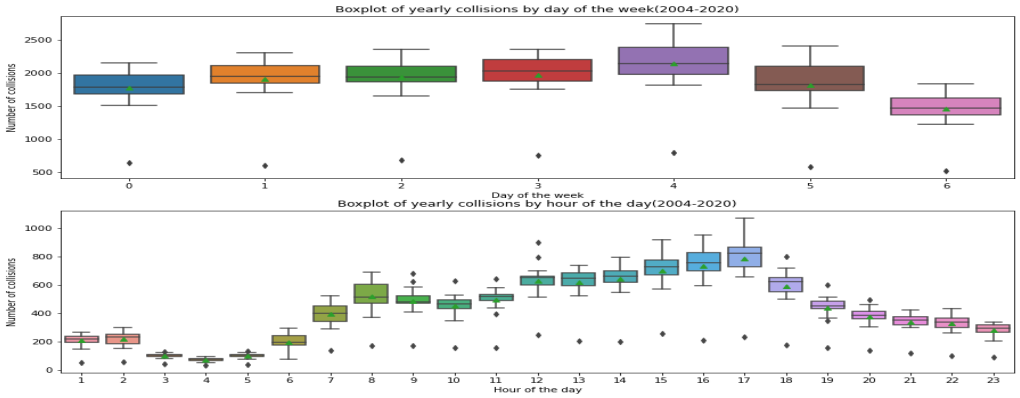

Figure2 : Seasonality of collisions by month, day of the week and hour

Weather

Weather conditions do not always influence the amount of collisions taking place, but do they influence on the severity of the collision taking place? Figure 3 shows the number of collisions by severity vs weather conditions at the time of the incident.

No matter how severe a collision is, most accidents occur when skies are clear. A minority of collisions (almost 1 in 4) occur with raining condition and a similar proportion of them with overcast skies. Weather is a contributing factor to collisions in Seattle, but does not seem to have a predominant role.

Light conditions

As discussed previously in the "seasonality" section, most accidents occur during daytime, but it is important to determine if there's an effect between the severity of an accident and the light conditions. Figure 4 summarizes this information, where it can be seen that dark

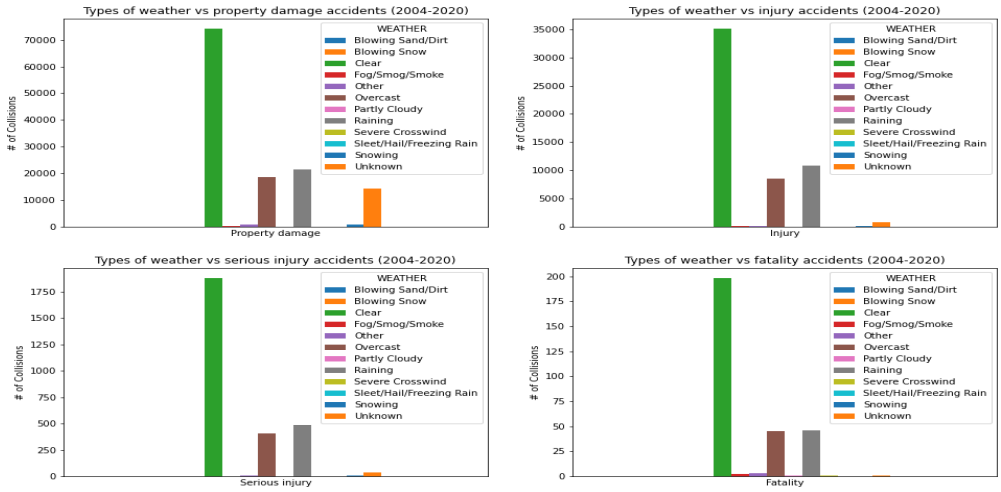

Figure3 : Weather vs severity code

conditions have a more prominent role in severe accidents. For instance, fatal accidents are almost 3/4 of day time accidents.

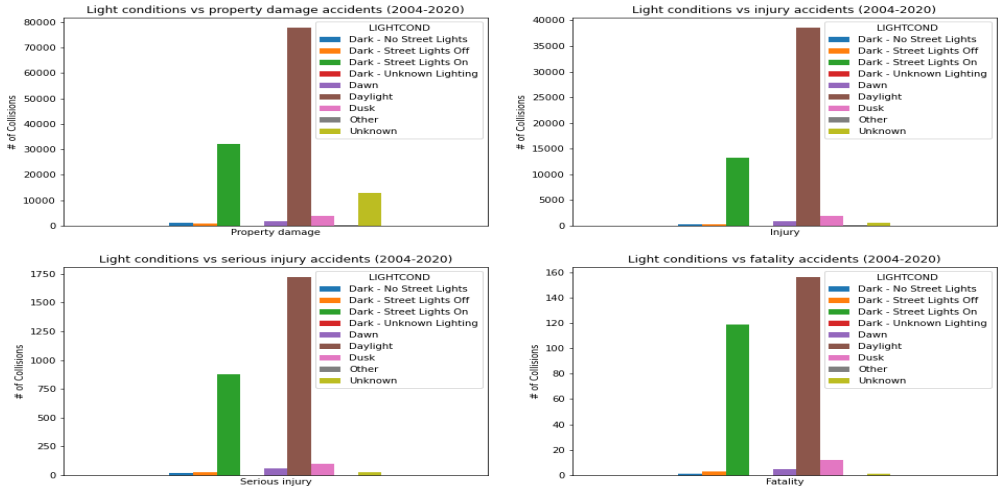

Figure4 : Lighting conditions vs severity code

Road conditions

As mentioned in the introduction, slippery roads are one of the leading factors behind a road accident[3]. The effect of rain on collisions has already been analyzed for this project, but other factors causing slippery roads like oil, ice, or sand have not been analyzed yet. Figure 5 shows the number of collisions by severity for each road condition. Notice how no significant information can be found regarding slippery conditions other than rain.

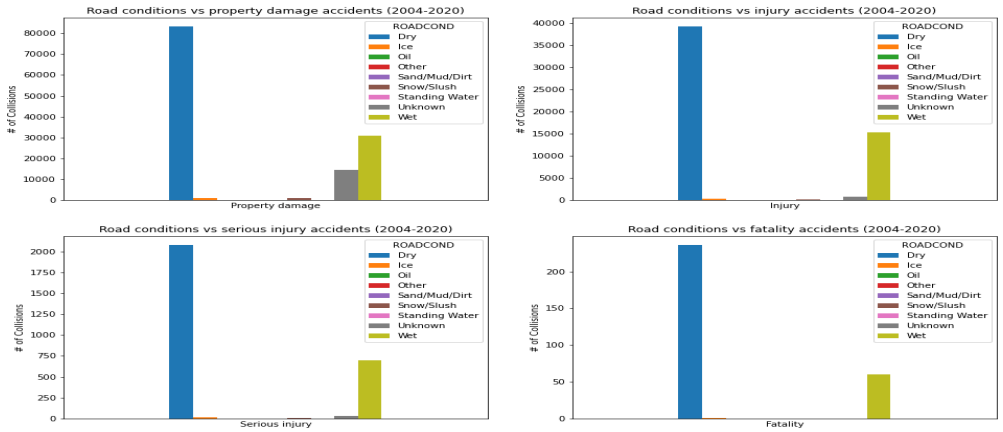

Figure5 : Road conditions vs severity code

Carelessness

Another factor contributing to road collisions according to Rolison et al. is carelessness[3]. Inattention to the road and speeding are some of the factors behind car collisions in Seattle; however, they represent a very low number of collisions compared to attentive drivers. Figure 6 shows the types of collisions generated by careless drivers vs the types of collisions generated by drivers paying attention to the road.

Inattentive drivers tend to hit the rear of the car in front, a parked car, or another car in angles. Speeding drivers fall mainly into the "other" category, but besides that they behave similar as inattentive drivers. It can also be observed that the amount of drivers being careless is very low compared to other type of drivers. The imbalance on this data does not make it a good variable for a model, so it will not be taken into account for this project.

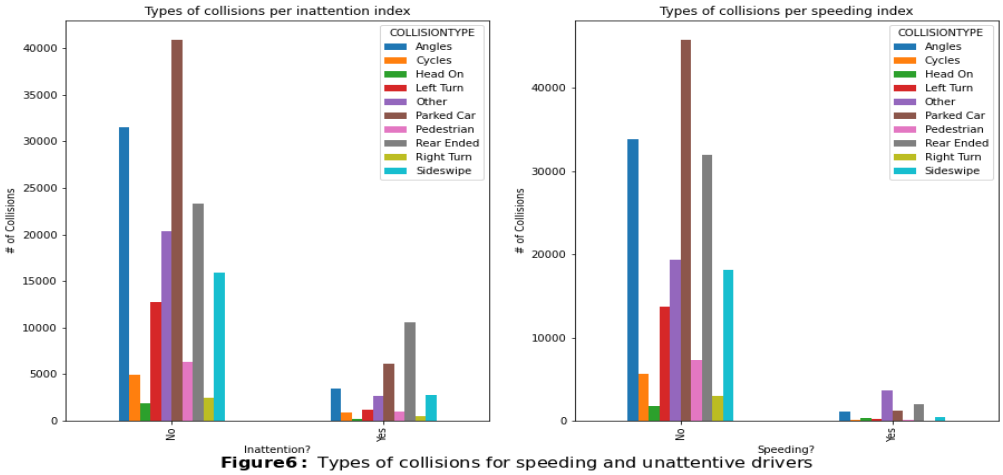

Figure6 : Types of collisions for speeding and unattentive drivers

Actions and Actors involved

Failure to judge another parson's path or speed is another reason behind car collisions [3]. Among the most common types of collision listed by the Seattle Department of Transportation are hitting a parked car, colliding in angles, hitting the rear end of a car, turning left, hitting a cyclist, hitting a pedestrian, etc. Figure 7 shows the most popular types of collisions per severity of incident. The results are consistent with the findings of Rolison et al. [3] where "failure to look properly" and "failure to judge another person's path or speed" can be translated into rear ended collisions, angle collisions and even head-on collisions.

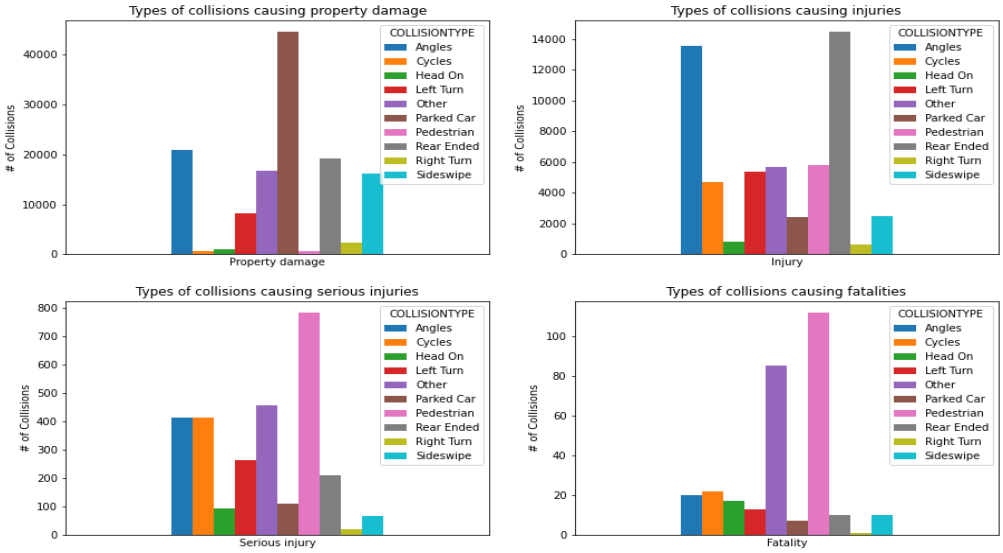

Figure7 : Types of collisions per severity code

Most collisions causing property manage or injuries are related with hitting a parked car, colliding in angles, hitting a car's rear end, turning left, etc. On the other hand, serious injuries and fatalities are caused by head on collisions, cars hitting pedestrians, and cars hitting cyclist, but also during left turning and colliding at angles. Notice how right turning produces very few collisions.

Location of collisions

As mentioned on the section before, some actions are more likely to end in a collision than others, and some actions are more likely to produce a more serious collision than others. For instance, head on collisions tend to produce more serious damage than rear ended collisions. Head on collisions are unlikely in freeways since there is a division between lanes, hitting a pedestrian there is also less likely. The following image shows the distribution of the different types of collisions across Downtown Seattle from 2017 to mid September.

The smaller the circle on the map, the less severe a collision is. It can be seen on figure 8 that collisions in downtown Seattle vary from severity 1: Property damage up to severity 3: Serious injury. It can also be seen that around main arteries/avenues collisions are more present (and in all types of severity) than in other types of streets. This can be further observed in figure 9.

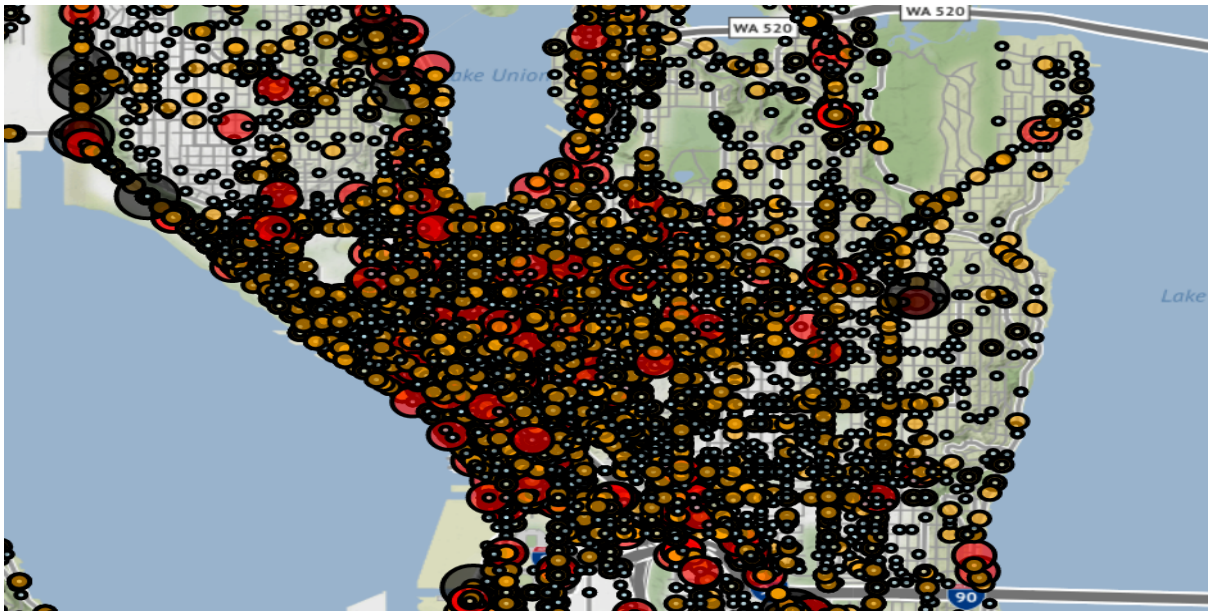

Figure 8: Collisions in Downtown Seattle from 2017 to mid-September 2020.

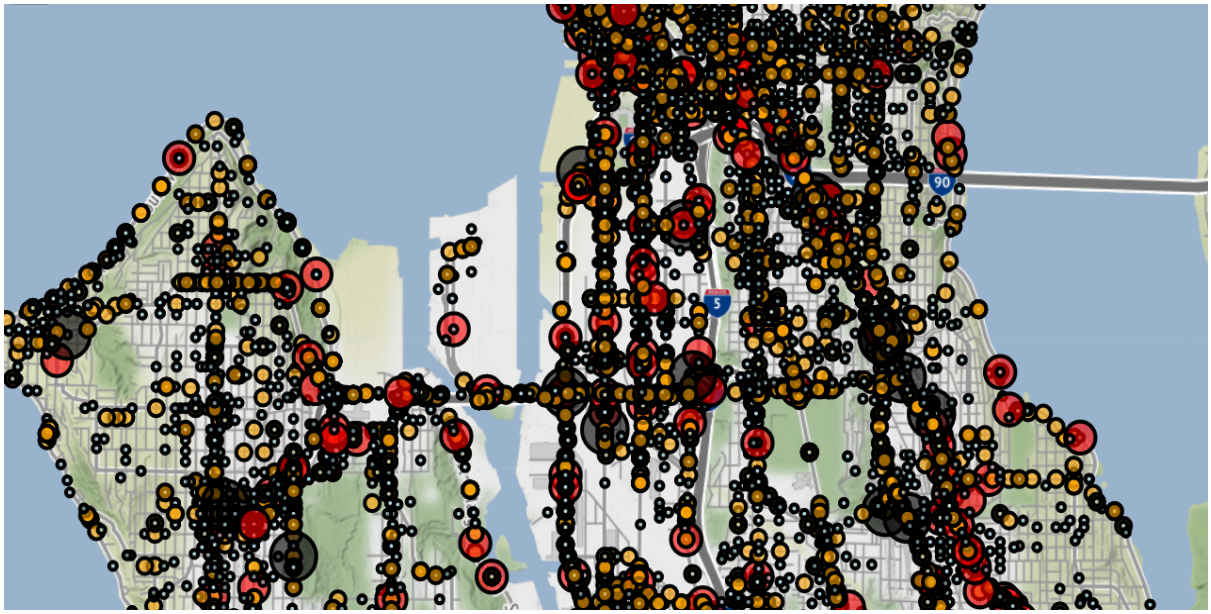

Figure 9: Collisions in Southern Seattle from 2017 to mid-September 2020.

Some neighborhoods seem to have more of a certain severity of collisions than others, also the arteries of a city have significantly higher amounts of collisions than non-primary

streets. According to the 'Seattle Right-of-Way Improvements Manual' [9], the city classifies its streets based on movement (principal arterial, minor/collector arterial, and non-arterial) and place (industrial, urban, downtown, and urban village). Figure 10 shows the severity of collisions per type of street.

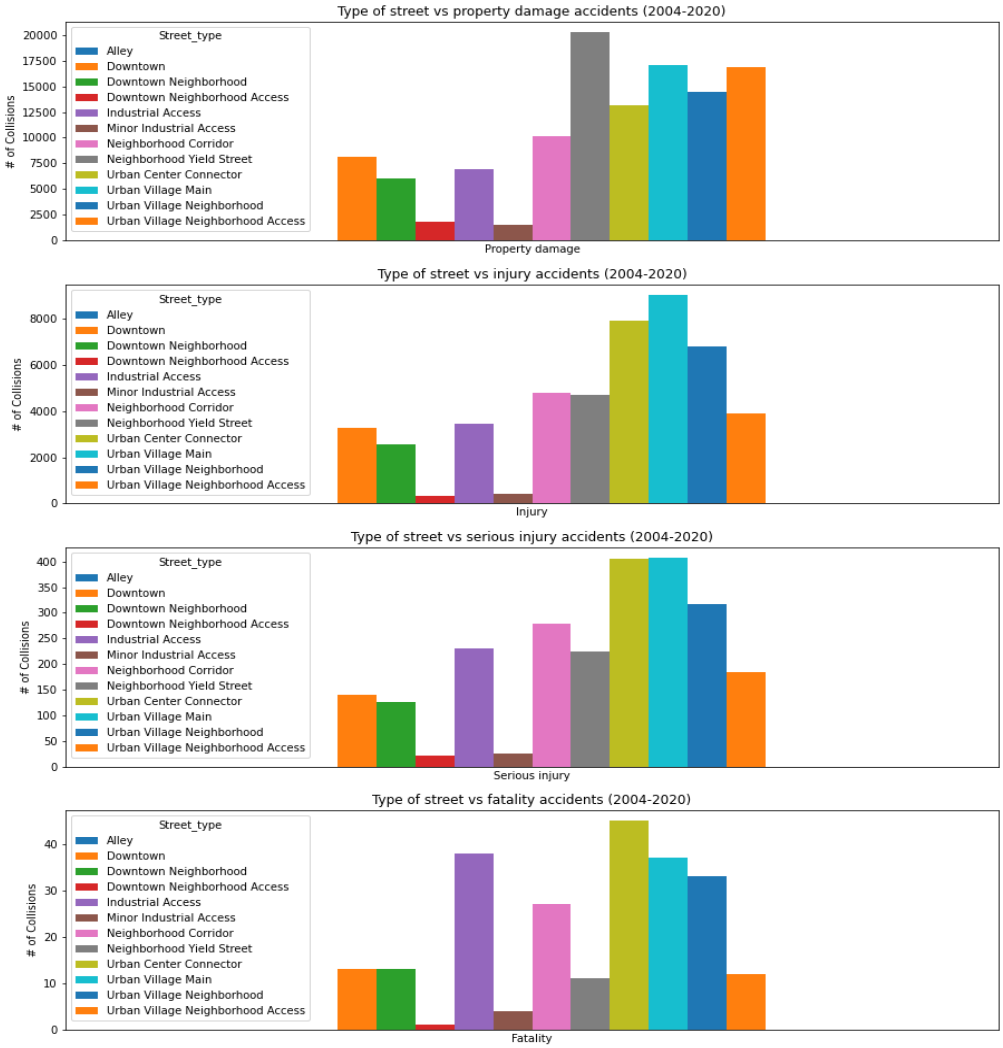

Figure 10: Type of street vs severity code

It can be seen that 'Neighborhood yield streets', which have speed limits of 15-20 mph have high property damage accidents and low fatality accidents. The same is true for 'Urban

Village Neighborhood' access, which has a similar speed limit. An opposite behaviour can be found for 'Industrial Access' streets, which are more predominant in more severe collisions. The same happens with 'Urban Center Connectors', which is a principal artery just like Industrial Access. Both Industrial Access streets and Urban center connectors have speed limits of 25-35 mph. Minor Arterial roads like 'Downtown Neighborhoods' are equally present in all types of accidents. For more information on the types of streets, please refer to Figure 11.

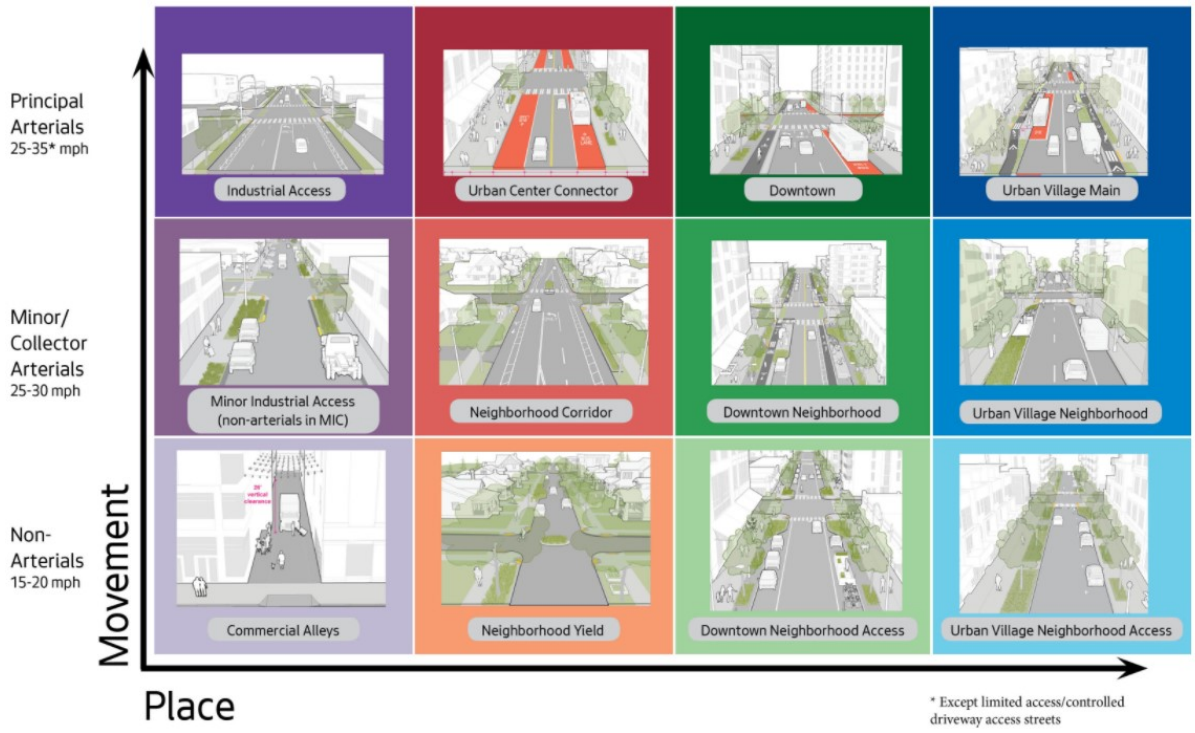

Figure 11: Type of streets in Seattle [9]

Predictive Modeling

One of the challenges for predicting the severity of a future collision based on this data set is that there is a heavily unbalanced number of labels. For every collision resulting in a fatality on this data set, almost 500 collisions causing property damage-only take place. Several approaches can be taken in order to deal with imbalance data. According to Truică et al. [10], decision trees and random forests are good for dealing with imbalance data. Additionally,

oversampling and under-sampling the labels of a model can further improve its accuracy.

First approach

A first approach into creating a classification model consists of creating a training and testing set, using different oversampling and undersampling methods and tuning a decision tree classifier model.

Training and testing

A random selection of 70% data was chosen to train the model and 30% was randomly chosen to test the model using the 'train_test_split' function from sklearn.

Weighting

Weighting was applied to the data set in order to prevent the model from bias towards the dominant label.

Table 2: Weighting	
Severity	Weights
Property Damage	0.3568
Injury	0.8809
Serious Injury	17.66
Fatality	177.69

Oversampling

Smote was used to oversample the labels that were scarce. Smote uses a technique to create more points of an undersampled label by selecting the k-nearest neighbors of a label to calculate the mean of the points to create a new point. After SMOTE, all labels had the same number of points (81,526 each).

Second Approach

Random Forest Classifier was used on the second approach, data was trained and tested, oversampled and undersampled and finally the best parameters were determined using GridSearchCV.

Training and testing

A random selection of 70% data was chosen to train the model and 30% was randomly chosen to test the model using the 'train_test_split' function from sklearn.

Over and Undersampling

Data was oversample again by using SMOTE, then the remaining data set was undersampled by using the imblearn RandomUnderSampler.

Cross Validation

Model training was performed by using cross validation, a splitting method where the data set is divided in n groups. The model is then trained with n-1 groups and tested with the remaining group. The process is repeated but now another group is used for testing, the process is repeated n times until all groups have been used for both training and testing.

Performance

First approach

A confusion matrix, accuracy, and f1 macro average were calculated to test the performance of the first approach. The results are summarized in the table bellow, where it can be seen that even though the accuracy of the decision trees was around 60%, the model is only good for predicting collisions resulting in 'Property damage', it performs slightly better than randomness for predicting Injury yielding collisions, and does not predict serious injury nor fatality yielding collisions.

Table 3: Performance of first approach

Model	Acc.	Property Damage F1-score	Injury F1-score	Serious Injury F1-score	Fatality F1-score	Macro avg
Decision Tree (criterion='entropy', class_weights = 'weights)	0.62	0.73	0.35	0.02	0.01	0.28
SMOTE + Decision Tree (criterion='entropy)	0.60	0.72	0.36	0.03	0.02	0.28

Second approach

Since SMOTE was not enough to deal with the imbalance data, a second approach was tested with random forest classifiers, cross validation and a combination of over and undersampling. The following table summarizes the performance results.

Table 4: Performance of Second approach

Model	Train macro avg.	Property Damage F1-score	Injury F1-score	Serious Injury F1-score	Fatality F1-score	Testing macro avg
Random forest, cross validations = 5	0.27	0.80	0.28	0.01	0.00	0.27
SMOTE + Random forest, cross validations = 10	0.78	0.80	0.28	0.01	0.00	0.27
SMOTE + Random Under Sampler + Random forest, cross validations = 5	0.87	0.80	0.27	0.01	0.00	0.27

Hyper-parameter tuning

Pre-processing the data with a combination of over sampling and under sampling provided the best training performance for the model. After that, the best parameters for the model were determined using GridSearchCV with 5 cross validations.The following table summarizes some of the parameters use and the test scores obtained.

Table 5: Hyper parameters

Mean fit time [s]	Criterion	Max depth	Max features	# of estimators	Mean test score
216.57	'entropy'	10	5	700	0.484
400.38	'entropy'	15	5	1200	0.646
281.51	'gini'	15	8	700	0.683
480.21	'gini'	15	8	1200	0.684

Based on the mean fit time and the mean test score (jaccard index), the best parameters for fitting the data set are: Criterion = 'gini', Max depth = 15, Max features = 8, and number of estimators = 700. After testing the model, the f1-scores for Severity 1, 2, 3 and 4 collisions were of 0.76, 0.43, 0.03 and 0.02, respectively. The importances taken into consideration for the model are presented in the following figure.

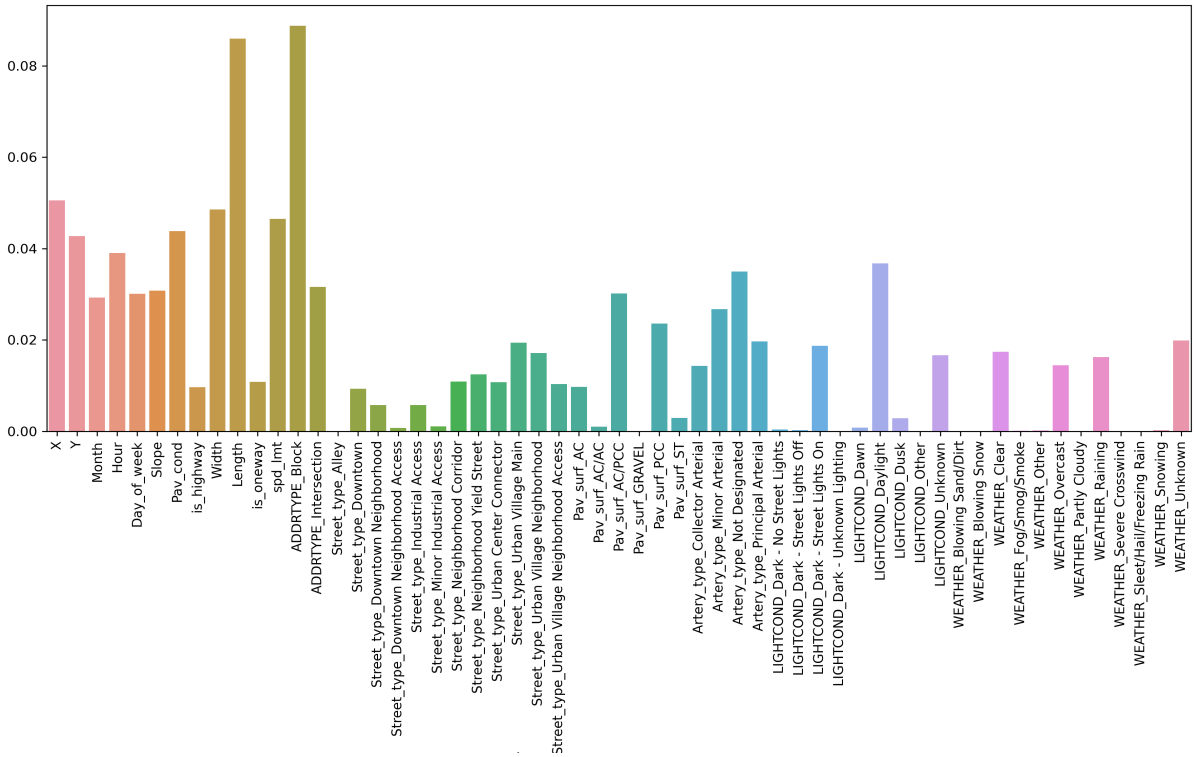

Figure 12: Random Forest Feature Importances

Results and Discussion

A combination of undersampling and oversampling the data set was implemented in order to improve the performance of a random forest classifier. After tuning the model, it performs better than randomness when trying to predict property damage incidents and incidents resulting in injuries. The most important parameters for determining the severity of a collision are: Weather it occurred along the block or at an intersection, the length of the street segment where the accident took place, the coordinates of the accident, the speed limit, the width of the street is, the weather, light conditions, pavement conditions, day of the week, and so on.

The results are consistent with the findings by Rolison et al. [3], but the model also helps determine which factors are more important than others for causing a collision. The feature importance results, along with the descriptive analytics of this project, have shed some light into the following facts:

1. Most accidents happen at daylight and when skies are clear.
2. Rain adds up to collisions but is not as important as the hour of the day and day of the week (amount of people on the streets).
3. Industrial and urban areas with higher speed limits tend to have more severe crashes than other types of streets.
4. Pedestrians and cyclist are the largest victims of severe injuries and fatalities.

Conclusion

The aim of the project was to warn about the possibility of a driver getting into an accident and how severe it could be, a model with 66% accuracy was successfully developed; however, the more severe accidents cannot be predicted by this model. Despite this, the data understanding section of this project shed some light on important factors behind car collisions in Seattle. Predicting the possibility of a driver getting into an accident yielding in a serious injury or fatality is not an easy task; in the meanwhile, the Seattle Department of Transportation can use the information analyzed in this project to make changes in the traffic regulations to make streets safer for everybody.

REFERENCES

1. Centers for Disease Control and Prevention. Motor Vehicle Injuries. Retrieved from: https://www.cdc.gov/winnablebattles/report/motor.html

2. "The Economic and Societal Impact Of Motor Vehicle Crashes, 2010" National Highway Traffic Safety Administration, page 5. Retrieved from: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013

3. Rolison, J; Regev, S. et al. "What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions. and road accident records." Accident Analysis & Prevention. Volume 115, June 2018, Pages 11-24. Retrived from: https://doi.org/10.1016/j.aap.2018.02.025

4. Seattle Open Portal. Retrieved from: https://data.seattle.gov/

5. Seattle Information Technology. About the Open Data Program. Retrieved on October 4th from: http://www.seattle.gov/tech/initiatives/open-data/about-the-open-data-program

6. Seattle Information Technology. Collisions. Retrieved on October 28th from: https://data.seattle.gov/dataset/Collisions/nuam-5pkc

7. Arcgis Metadata Form. Collisions - All Years. Retrieved on October 28th from: https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pd

8. Seattle Information Technology. Seattle Streets. Retrieved on October 28th from: https://data.seattle.gov/dataset/Seattle-Streets/b856-55i2

9. Seattle Right-of-Way Improvements Manual. 2 Street Type Standards. Retrieved on October 28th from: https://streetsillustrated.seattle.gov/https-streetsillustrated-seattle-gov-wp-content-uploads-2019-12-streettyperelationships-movementplacev4-jpg/

10. Truică, Ciprian-Octavian & Leordeanu, Catalin. (2017). Classication of an Imbalanced Data Set using Decision Tree Algorithms. University Politehnica of Bucharest Scientific Bulletin Series C - Electrical Engineering and Computer Science. 79. 69-.