

**Capstone Project - Car Accident Prediction**

Yoshie Toriya

October 25, 2020

## Introduction

**Background**

Deaths and injuries by road crash are highly concerning issues, with 1.35 million and 50 million people affected around the world, respectively.

In 2018, the United States, being one of the busiest countries with nearly 280 million vehicles in operation and more than 227.5 million people holding a valid driving license, counted more than 38,000 deaths and 12 million vehicles involved in road crashes.

Road crashes are the leading cause of deaths in the U.S. for people aged 1-54. An additional 4.4 million are injured seriously enough to require medical attention. According to OECD’s 2019 U.S. annual report, the economic costs of traffic crashes represents 1.6% of the GDP. When quality of life valuations are considered, the total value of societal harm from motor vehicle crashes is nearly 6% of GDP.

**Problem**

In 2018, in Seattle alone, the dataset provided by Seattle Department of Transportation (SDOT) counts over 3,500 injuries and 7,000 property-damage-only collisions in the same year. Thus, it is absolutely reasonable that various stakeholders would wish to minimize such numbers. This project addresses the request to choose the suitable predictive model(s) of the possibility and severity of road crashes.

Potential stakeholders include:

* Drivers, pedestrians, bicyclists, and others using transportation services
* Transportation service providers
* Hospitals
* Government bodies, e.g. EMS, SPD, School Board, DoTransportation, DoHealth
* Other local and federal government bodies (e.g., policy-making side, social welfare-side, etc.)
* Insurers
* Corporations and employees
* Nonprofit organizations

This assignment poses a question of how to predict road crashes when an individual sets off on a drive. To put it in a broader and socially beneficial context, I am setting **the goal of this project on finding**, if any, the determining variables that cause crashes and **the most suitable predictive modeling** for the responsible parties such as SDOT or SPD to be able to alert and educate drivers sufficiently to reduce the number of crashes.

I also considered predicting the severity of crashes but have decided not to pursue it at this point as people are not likely to decide whether or not to drive based on the possible severity of a crash. Thus I have expected the output will be supervised, binary-labeled classifications.

### Data Description

The following is a list of datasets I have looked into and decided how or whether to use them for the defined goal:

|  |  |  |
| --- | --- | --- |
| **1.** | **Collision Records (SDOT)** | provided by Seattle Department of Transportation (SDOT) found at [Kaggle](https://www.kaggle.com/jonleon/seattle-sdot-collisions-data), of more than 221,000 records over the period of 2004 - 2019 |
| **2.** | **Traffic Flow Counts** | [ArcGis](https://www.esri.com/en-us/home) datasets found at [Seattle GeoData](https://data-seattlecitygis.opendata.arcgis.com/). |
| **3,** | **Collision records (WSP)** | provided by [Washington State Patrol](https://www.wsp.wa.gov/driver/collision-records/) which offers per-involved-party (driver/pedestrian/pedcyclist) detail records while 1. provides per-crash-a-record summary |
| **4.** | **Weather data** | [NOAA](https://search.usa.gov/search?utf8=%E2%9C%93&affiliate=ncdc&query=seattle) datasets |

(5. Negative sample dataset - this will be explained in the data preparation section.)

In the following section, I describe each dataset, beginning with 1. SDOT collision records.

#### Data Understanding

#### Dataset: SDOT Collision Records

#### 1. SDOT Collision Records is a set of a-record-per-incident data which contains geospatial details, severity of the collision, conditions of the road, lighting, and weather at the time of the accident, and the feature of the junction.

#### There are 221,266 rows and 40 columns.

|  |  |
| --- | --- |
| SEVERITY breakdown  **Screen Shot 2020-10-25 at 1.47.55 PM.png** | ADDTYPE breakdown  **Screen Shot 2020-10-25 at 1.49.01 PM.png** |
| JUNCTION breakdown  **Screen Shot 2020-10-25 at 1.48.12 PM.png** |

#### We can see, most prominently, the majority of which are property-damage-only cases, which have been declining since 2016.

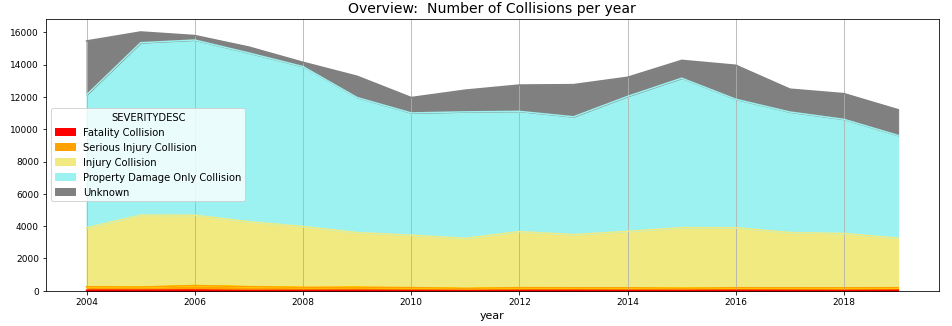


Figure 1. Number of collisions over the period of 2004 to 2019 with severity breakdown

Furthermore, I wanted to see if there is any seasonality.

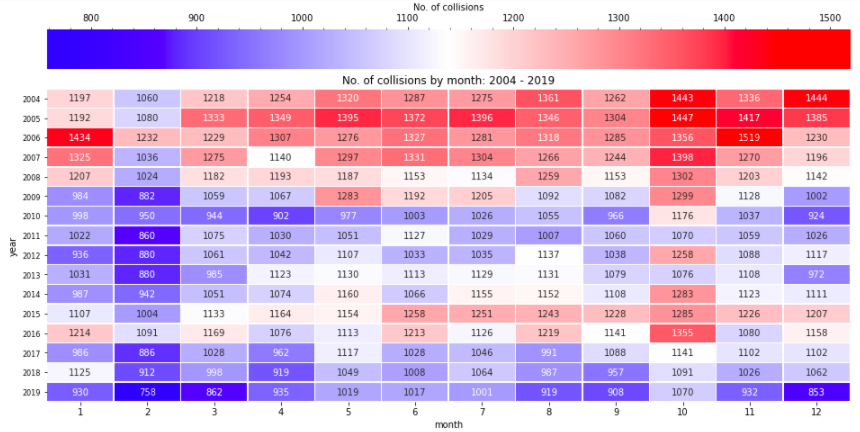


Figure 2. Number of collisions per month over the period of 2004 to 2019

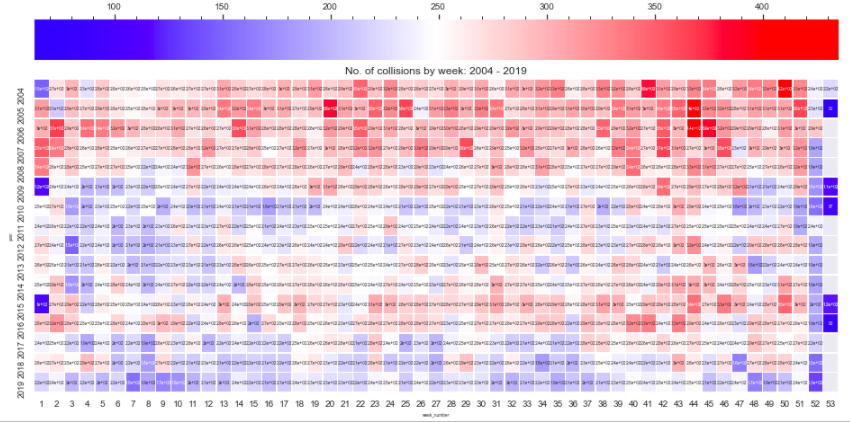


Figure 3. Number of collisions per week over the period of 2004 to 2019

There appears to be no apparent seasonality. I have also zoomed in on the recent six years.

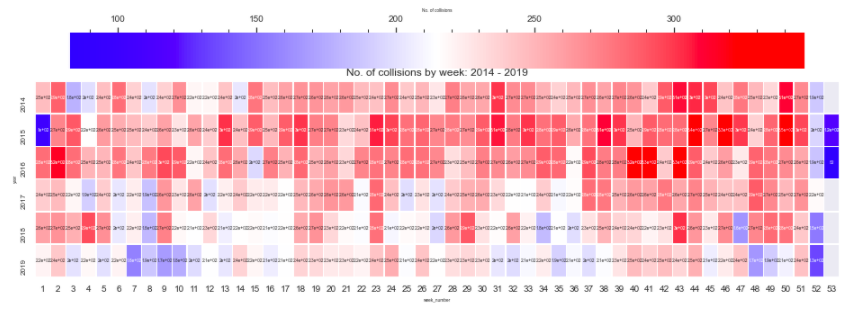


Figure 4. Number of collisions per week over the period of 2014 to 2019

Not every year is the same. There is no significant seasonality to be considered.

The graphs below display the circumstantial conditions. We see most cases happen on dry roads and in daylight.

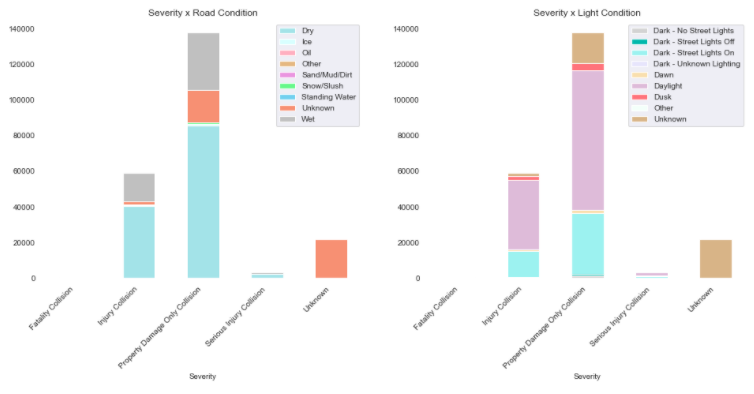


Figure 5. Breakdown of Road Conditions per Severity of Collision (of all cases)

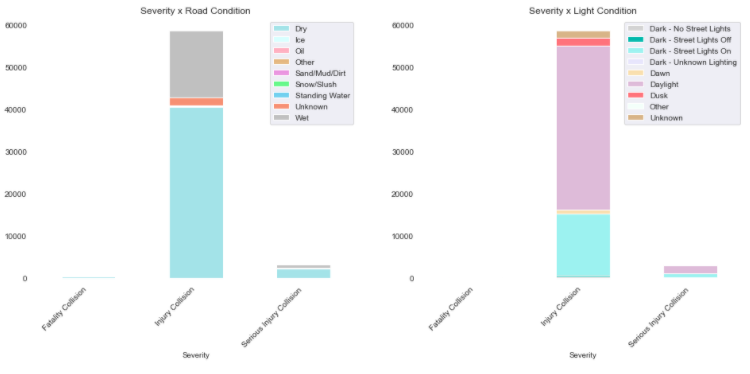


Figure 6. Breakdown of Road Conditions per Severity of Collision (excluding Property-damage-only and Unknown)

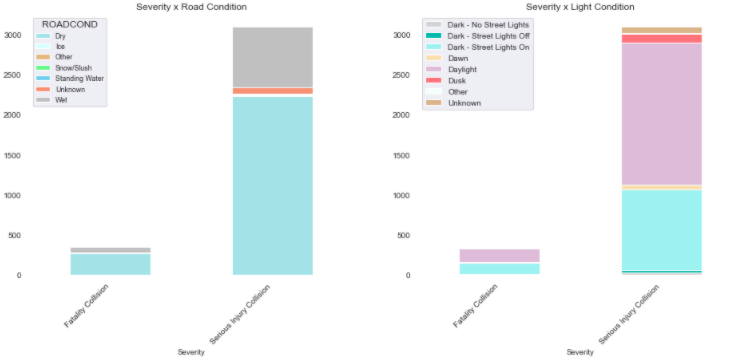


Figure 7. Breakdown of Road Conditions per Severity of Collision (Fatal and Serious Injury only)

Figure7 and 8 illustrate whether there is any correlation between variables that may provide a hint to where else I could probe further.

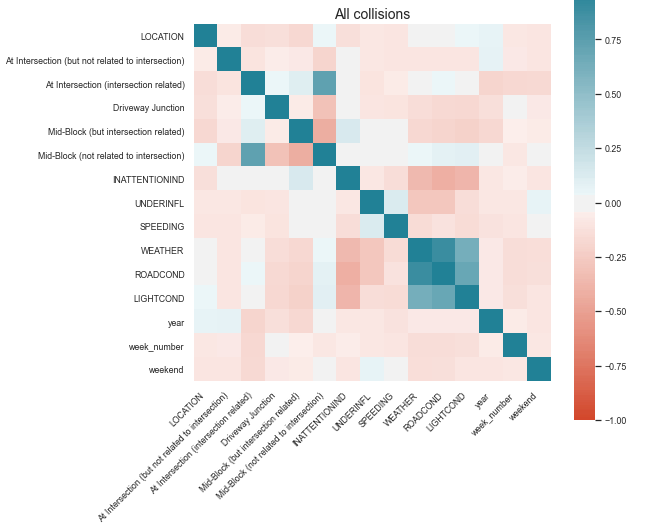


Figure 8. Correlation between variables (of all cases)

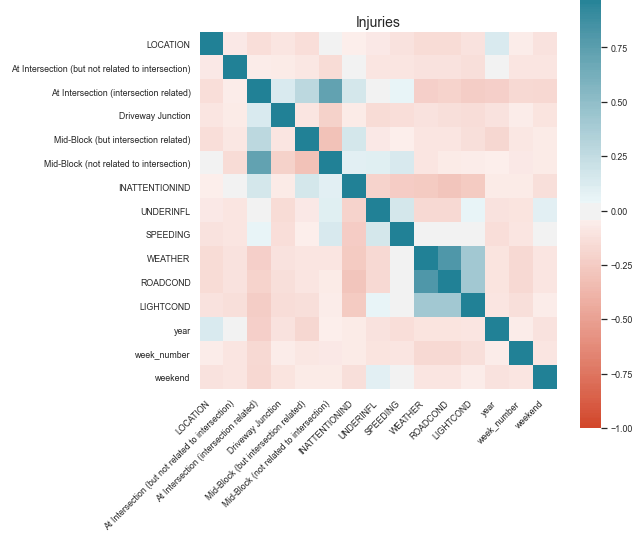


Figure 9. Correlation between variables (of injury cases)

##### Aside from the obvious correlation between weather and road and light conditions (road and light conditions are dependent on weather), no other correlations can be specifically pointed out.

Thus far we do not see particular patterns such as seasonality or variables that are so distinctive as the causes of crashes. One more area we could explore is personal attributes, which has led to finding the WSP collision records that contains age, gender, vehicle details and more.

#### Dataset: WSP Collision records

I have separated Understanding WSP Dataset section in another notebook as it involves a much larger dataset. Upon evaluating the correlation among variables, and given the fact that these records do not have geospatial data or mappable IDs to SDOT data, I decided not to use it at this point in this project. It is included in the [Appendix](https://github.com/Yoshie-T/Coursera_Capstone/blob/master/Appendix_Understanding%20other%20datasets.ipynb).

#### Datasets: Weather data (NOAA) and Traffic Flow Counts datasets

#### Upon observing in the above two datasets, I decided to not go further with the ****weather data****. As for the ****Traffic Flow Counts data****, while it should provide the perfect negative crash dataset, the number of negative records could hypothetically be as many as 13.5 billion, which would make the positive data and negative data significantly unbalanced. The prediction to positive results would be cosmically too small and would defy the purpose of alerting and educating about the risks. The data also appears critically lacking to contain latitude and longitude points and would thus require massive text processing work on the street names in both this and SDOT datasets to merge them, as they seem to be the only thing close to a mappable ID. Dataset: Negative records

For the **negative data set**, therefore, I will proceed with building it up by using the SDOT positive dataset and randomizing it, which is a commonly used method [a senior Data Scientist at Esri](https://medium.com/geoai/using-machine-learning-to-predict-car-accident-risk-4d92c91a7d57) introduces. Alternately, large research institutes may be able to develop an algorithm to train data without negative data ([reference](https://www.riken.jp/en/news_pubs/research_news/pr/2018/20181126_2/index.html)). Hopefully such an algorithm will be available in a Python library soon if it has not already been.

Lastly as part of data understanding, I have created **HeatMapWithTime geo maps** with SDOT data that displays hourly collision occurrences that are also in the [Appendix](https://github.com/Yoshie-T/Coursera_Capstone/blob/master/Appendix_Hourly%20HeatMap.ipynb).

#### Data Preparation

### The following describes how I have prepared the datasets:

### 1. Made negative records three times in volume of SDOT (positive) records by copying SDOT records three times. Randomized them by adding/subtracting random small random figures to datetime values and geospatial values.

### 2. Added the target column ‘Crash’ to SDOT with value =1. Append it with the negative dataset. Applied value =0 to ‘Crash’ of negative records, and wherever the value is null in the table.

### 3. Separate it in two sets; (1) for Train & Test with records up to year 2019, (2) with records of 2020. The latter is to be used in the later Test phase.

### 4. Feature, Test\_Train\_Split, and Normalized the datasets. Training the models in the later phase required (1) to be smaller size, thus records of 2004 – 2017 have been dropped. (2) had to be equally downsized, so random records have been selected.

### Thus, I have (1) of 93,608 and (2) of 6,688 records.

#### Method

I have run the following steps with Classifiers: **Decision Tree**, **Random Forest**, **K-Nearest Neighbors**, **Support Vectors Machine**, and **Logistic Regression**:

* Run Train, Fit, while checking the accuracy score of each
* Evaluate each with Precision, Recall, F1 score
* Evaluate each with Confusion Matrix (Logistic regression will be dealt with log loss evaluation.)

The accuracy score suggests the depth of Decision Tree 1 and 2 offer the highest accuracy followed by that of 3 through 6.

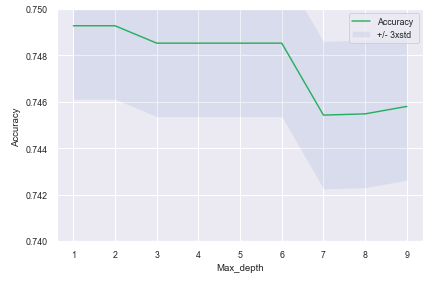


Figure 10. Accuracy of Decision Tree per depth

For visualization and training the model I am applying the depth 6.

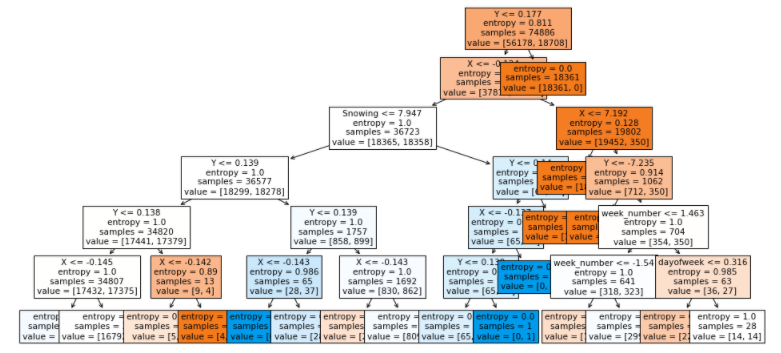


Figure 11. Decision Tree with the depth 6

As for K-Nearest Neighbors, 2 for the k marks the highest accuracy.

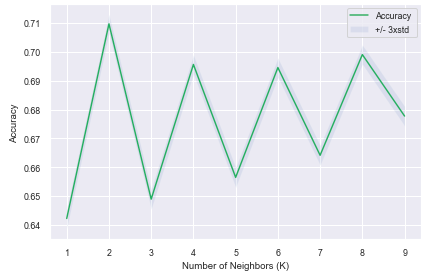


Figure 12. Accuracy of KNN per K

For Support Vector Machine, Polynomial, Linear, and RBF kernels report the same accuracy rate.

#### Screen Shot 2020-10-26 at 5.15.37 PM.png

Figure 13. Accuracy of SVM per kernel

With Random Forest and Logistic Regression, below is the summary of the accuracy. As seen through Figure 11 through 14 below, most models report similar scores, while logistic regression reports much lower score. Therefore logistic regression will not be considered as an option to proceed with.

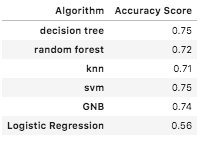


Figure 14. Accuracy Score Summary

Evaluation continues with Classifier Report and Confusion Matrix. Because we know the positive dataset is genuine and the negative dataset is not, I weigh more on predicting the positive rather than predicting the negative. In other words, Recall > Precision, and Accuracy > Precision.

Therefore at this point **Decision Tree**, **KNN**, and **SVM with polynomial kernel** appear to be the viable options to adopt. Confusion Matrix shows difference among the rest of the models.

|  |  |
| --- | --- |
| Screen Shot 2020-10-25 at 6.04.21 PM.png | Screen Shot 2020-10-25 at 6.41.36 PM.png |
| Screen Shot 2020-10-25 at 6.04.01 PM.png | Screen Shot 2020-10-25 at 6.41.46 PM.png |
| Screen Shot 2020-10-25 at 6.04.09 PM.png | Screen Shot 2020-10-25 at 6.41.55 PM.png |

SVM with Polynomial, Sigmoid, Linear and RBF kernel:

|  |  |
| --- | --- |
| Screen Shot 2020-10-26 at 5.19.38 PM.png | Screen Shot 2020-10-25 at 6.45.59 PM.png |
| Screen Shot 2020-10-26 at 5.19.46 PM.png | Screen Shot 2020-10-25 at 6.46.10 PM.png |
| Screen Shot 2020-10-26 at 5.19.56 PM.png | (Confusion Matrix turned out insufficiently) |
| Screen Shot 2020-10-26 at 5.20.04 PM.png | (Confusion Matrix turned out insufficiently) |

Results and Evaluation

##### KNN and SVM with polynomial kernel mark the same scores whereas Decision Tree scores lower.

|  |
| --- |
| Screen Shot 2020-10-25 at 7.26.56 PM.png |
| Screen Shot 2020-10-25 at 7.27.04 PM.png |
| Screen Shot 2020-10-25 at 7.27.15 PM.png |

##### Therefore I conclude, as per this modeling and testing procedure, **K-Nearest Neighbors** and **SVM** with polynomial kernel have been identified as the relevant models.

## Discussion

As seen above, most classifiers have marked close accuracy above 71% at the training phase, and K-Nearest Neighbors and SVM worked similarly through the post-train test phase with 100% recall and 75% accuracy. I believe the results are acceptable as the first step for the purpose of warning the risks to the drivers.

I also believe the results could be improved and output could offer more by:

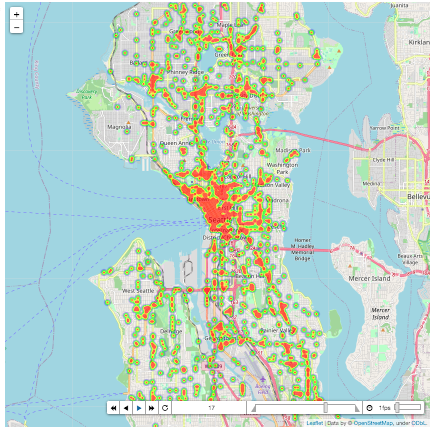
1. applying the real and accurate negative samples
2. mapping with other collision detail datasets, such as WSP records
3. incorporating dynamic geospatial and weather data

Currently, such as in the case of TSP data, 1. faces challenges in that the dataset lacks geospatial details as well as the significantly imbalanced volume compared to that of the positive dataset, which will require a way of optimizing. In this work, having to create negative data examples from scratch has become the source of no-full confidence in it.

Nonetheless, this work has been merely a starting point, and by continuing to probe into variables and to train and test, improvement is possible.

## Conclusion and Future Direction

We can continue to study further into the known datasets, explore possible other data sources, and continue to apply those to various algorithms until we are confident to conclude the accuracy has reached the highest within the feasible capacity. Once we have run sufficient training and testing, we can put this into practice, for example, in developing a prediction app. I have created HeatMapWithTime geo maps and present as a way to provide a visible and practical image of the adaptation that can be seen [here](https://github.com/Yoshie-T/Coursera_Capstone/blob/master/Appendix_Hourly%20HeatMap.ipynb). (Below is a screenshot)



I also personally find it highly interesting to incorporate more human and vehicle aspects into the prediction that WSP data could provide for example. That is To Be Continued.

Thank you for taking time to review this.

### References

[National Motor Vehicle Crash Causation Survey, US Department of Transportation](https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811059)  
[The Association for Safe International Road Travel (ASIRT)](https://www.asirt.org/safe-travel/road-safety-facts/)  
[The City of Seattle Transportation GIS Datasets](https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions)  
[Medium: Using Machine Learning to Predict Car Accident Risk](https://medium.com/geoai/using-machine-learning-to-predict-car-accident-risk-4d92c91a7d57)  
[Kaggle: Seattle SDOT Collision Data](https://www.kaggle.com/jonleon/seattle-sdot-collisions-data)  
[Riken: Smarter AI: machine learning without negative data](https://www.riken.jp/en/news_pubs/research_news/pr/2018/20181126_2/index.html)

### Acknowledgement

[Photography: Driving Test Org](https://driving-tests.org/beginner-drivers/top-10-traffic-jam-tips/)  
[Daniel Wilson: "Using Machine Learning to Predict Car Accident Risk"](https://medium.com/geoai/using-machine-learning-to-predict-car-accident-risk-4d92c91a7d57)

### Appendix

[Understanding other datasets](https://github.com/Yoshie-T/Coursera_Capstone/blob/master/Appendix_Understanding%20other%20datasets.ipynb)

[Seattle Crashes Hourly HeatMap](https://github.com/Yoshie-T/Coursera_Capstone/blob/master/Appendix_Hourly%20HeatMap.ipynb)