Car Collisions in Seattle: Analysis & Prediction

Abraham Mauleon-Amieva

(Dated: October, 2020)

Car collisions are one of the major problems across different societies worldwide. It is estimated that 1.35 million deaths occur every year due to a car accident, and around 3,700 fatal incidents occur daily. Over 5,419,000 crashes, 30,296 of them being fatal, and 2,239,000 causing an injury occurred in the United States, just in 2010 [1]. Here, we analyse collisions occurred in Seattle, WA. from 2004 to present. We investigate the regions with high rate, and build several predicting models based on the conditions provided. Among the models tested, we find best results when using logistic regression.

I. INTRODUCTION

According to the Washington State Department of Transportation (WSDOT), car accidents happen in Seattle, WA. with a high frequency. It is estimated that an incident occurs every 4 minutes and fatal accidents that leave victims dead occur every 20 hours. In this sense, fatal accidents went from 508 in 2016 to 525 in 2017, leaving a total of 555 fatalities. Typical conditions that contribute to such fatal events include alcohol use, motorcycle use, and pedestrians. Here we evaluate the different types of incidents that occurred in Seattle and consider the different conditions that led to them. For this, the data record of incidents, locations and important conditions since 2004 is used. Based on the data provided, correlations between different areas and parameters, such as alcohol use, vehicle type, weather and street conditions, etc., are of use to detect the most common type of accident and its causes.

For this report, we use data provided by the SPD [2] and Traffic Records. The data set consist of *194,673* incidents reported in Seattle from 2004 to the present. There are primary and secondary keys associated with every accident, including the specific location, date and time. All kinds of collisions have been recorded, including cars, pedestrians and cyclists. In total, more than 15 different conditions have been considered, e.g. weather, types of location (alley, mid-block, intersections), alcohol or drug abuse, etc. In addition, the severity of the accident is denoted by the severity code. We start with looking in the areas with the most fatal incidents reported. Then, a classification using common conditions is used to assess the severity of the accident. Based on our findings, we report specific conditions and areas with high collision risk in Seattle. Prediction models are build and trained using the key conditions identified. We then evaluate the several models tested using metric tools.

II. METHODS

We use data provided by the SPD [2] for our analysis and model prediction. The original data set includes a total of 41 parameters, of which we select the ones we consider of most relevance. First, we evaluate the different collisions. These are classified by either by the severity or the type of collision itself. Figure 1 **a** and **b** show the severity classification and the type of collision respectively. It is noted that ≈ 70% of the collisions result in property damage only, while incidents involving parked cars and collisions coming from different angles are among the most common.

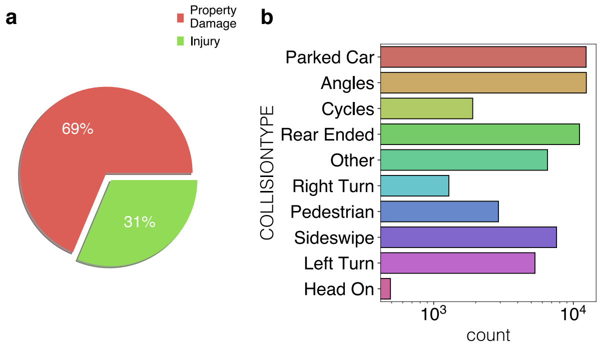

FIG. 1: **Collision Classification** by **a.** Severity: (i) Property Damage or, (ii) Injury. **b.** Different types of collision, i.e. parked cars, rear end, pedestrian involved, etc.

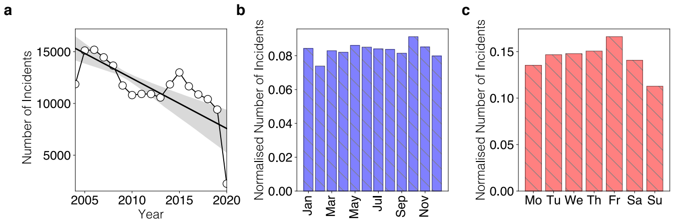

FIG. 2: **a.** Number of collisions per year. Open symbols are incidents recorded by the SDP, and the solid line is the best linear fit. Normalised counts per **b.** month and **c.** weekday.

We then identify the key conditions that promote the type of incidents described above. Using the data in Ref. [2], the collision annual rate is investigated. Our first observation is that the number of collisions per year shows a non-monotonic decrease, as indicated in Fig. 2**a**. In addition, we look for seasons and weekdays with a high incident frequency. As shown in Fig. 2**b**, the historic number of collisions is distributed evenly across the year. On the other hand, Fridays show the highest rate, but

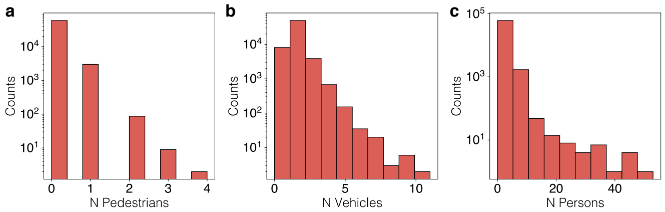

FIG. 3: Number of **a.** pedestrians, **b.** vehicles, and **c.** persons involved in collisions occurred in neighbourhoods with high rate. Note that the counts are plotted on a log scale for clarity.

the difference with other weekdays is not significant (see Fig. 2**c**).

It is also noted that the data contains latitude and longitude coordinates that are of use to identify areas of high collision rate. We use such coordinates, along the unique key for each incident to classify the different neighborhoods in Seattle. In Fig. 7**a** at the Appendix, we identify the different neighbourhoods in Seattle. Note that each region is coloured according the local number of historic collisions. Among the neighbourhoods with highest rate we find First Hill and Central Business District (see Fig. 7**b** and **c**). Given that the collision density varies across the different neighbourhoods, we select only the regions where the number of cases exceeds 4,000. We take this number as our threshold to classify low and high risk regions. Only high risk neighbourhoods are investigated for convenience.

After the neighbourhood selection, we analyse the different conditions that lead to the different types of collisions. Predicting models are build based on this analysis. Here we use k Nearest Neighbours (kNN), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR) machine learning algorithms available in Python to build our models.

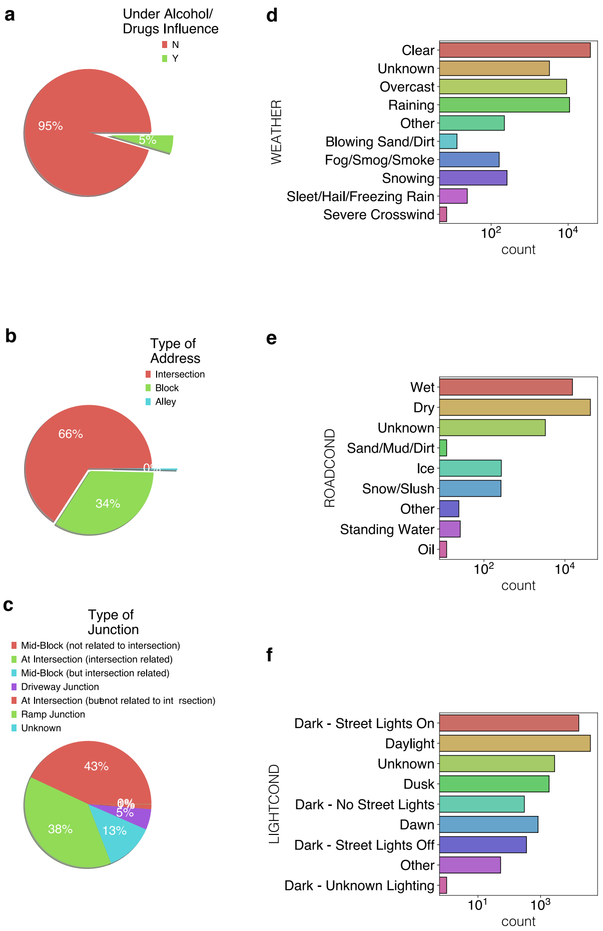

FIG. 4: **Incident Conditions a.** Influence of alcohol or drugs. **b.** Type of address: (i) Intersection, (ii) Block or, (iii) Alley. **c.** Type of junction, including: (i) Mid-Block, (ii) Road Intersection, (iii) Driveway Junction, etc. **d.** Weather, **e.** Road, and **f.** Light conditions. Note the log scales in plots **d-f**.

III. RESULTS

Having identify the neighbourhoods of interest, we proceed to analyse the different conditions that promote such a high number of collisions. Following the classification shown in Fig. 1**b**, we investigate the number of pedestrians, vehicles and persons involved. The histograms in Fig. 3**a-c** show that for most of the collisions the number of pedestrians is minimum, equally to the number of vehicles and persons. Nonetheless, few collisions involving 10 cars and more that 40 persons have been also recorded.

In addition, we investigate the influence of other considerable conditions, i.e. use of drugs and alcohol, and road conditions (see Fig. 4). Regarding the use of alcohol or drugs, this represents only 5% of the collisions, as shown in Fig. 4**a**. However, 42% of these collisions result in injury. Interestingly, 30% of the collisions result also in injury without alcohol and drug consumption. Intersections, blocks and alleys are considered types of addresses (Fig. 4**b**). Note that incidents in alleys have not been recorded in the observed neighbourhoods. Accidents at intersections are the most common, causing a large number of crashes coming at different angles (Fig. 1**b**). On the other hand, the most common type of collision in blocks involves parked cars.

Looking more in detail the collisions occurred at intersections, Fig. 4**c** shows the different types of junction. Most of the incidents correspond to mid-block crashes not relating an intersection (with 43%), followed by crashes at intersections (38%). We now look at the weather, road and light conditions. Interestingly, we find that a large number of accidents happened during optimal conditions. While a large amount of events remain unknown, most of them have occurred in clear conditions, followed by overcasts and raining days, as shown in Fig. 4**d**. Equally for the road conditions, the majority of the collisions oc-

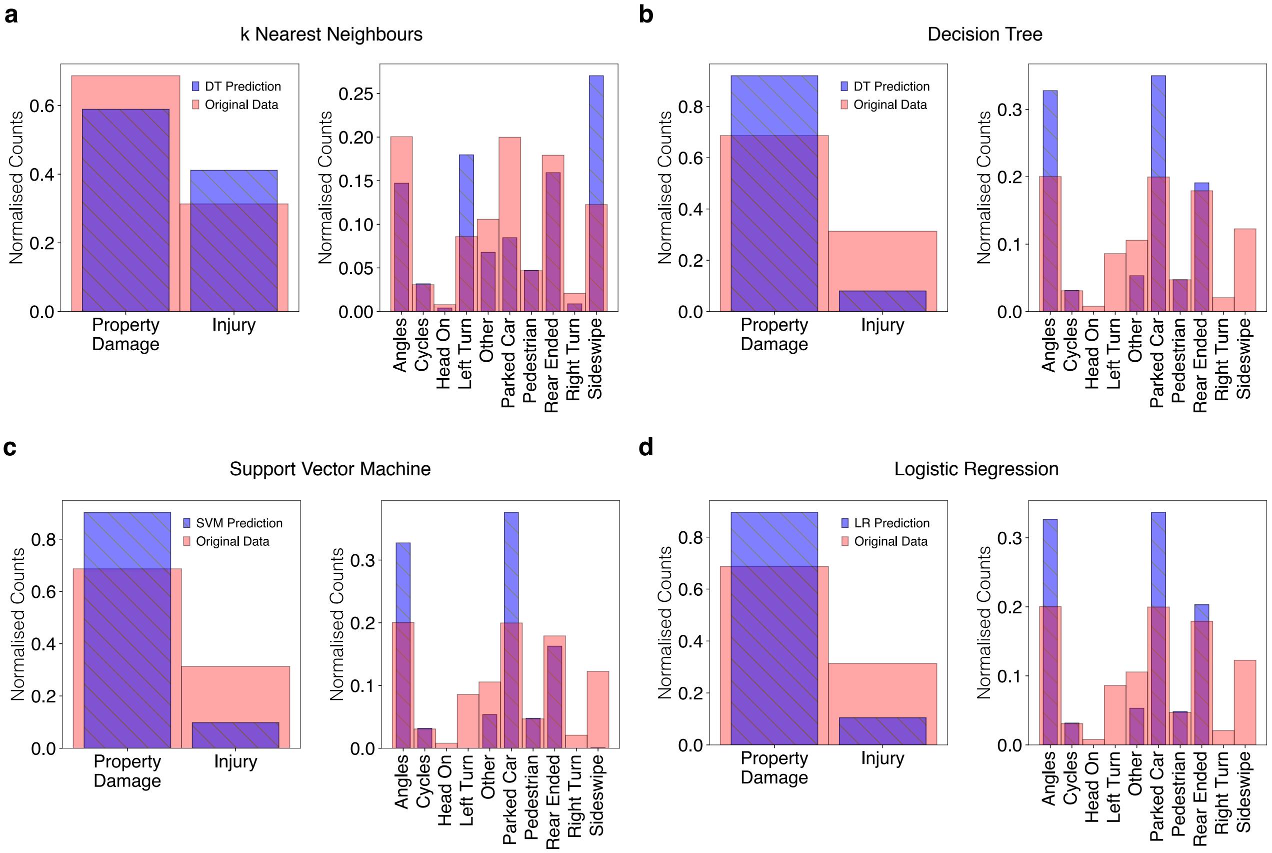

FIG. 5: **Model predictions** for the severity and collision type using **a.** k Nearest Neighbours, **b.** Decision Tree, **c.** Support Vector Machine, and **d.** Logistic Regression models. The original testing data is represented by red bars, while the predictions are plotted in blue.

curred in dry conditions, followed by wet roads, snow and ice (see Fig. 4**e**). Regarding light conditions, collision during daylight are more frequent than during dark, even with lights on. Nonetheless, the order of magnitude between these two conditions is comparable, as indicated in Fig. 4**c**.

Moving on the predicting models, we employ four different methods: (i) k Nearest Neighbours (kNN), (ii) Decision Tree (DT), (iii), Support Vector Machine (SVM), and (iv) Logistic Regression (LR). To train the models we use the data described in Fig. 3 and 4: number of pedestrians, vehicles, and persons, address type, influence of alcohol/drugs, weather, light and road conditions. For this, more than 60,000 incidents are used. Each model is used to predict the severity (property damage or injury), and the type of accident (Fig. 1**b**). First, the data set is split between training and testing data. We use 80% of the available data for training purposes. Our results are summarised in Fig. 5.

`k Nearest Neighbours` — We use the k Nearest Classifier, which relies on a given number of neighbours. For each prediction we identify the optimal number of neighbours $\in [1, 15]$. Results from predictions are shown in Fig. 5**a**. We test the accuracy of the model, which yields 61% for the severity, and 38% for the type of accident.

`Decision Tree` — Decision tree models are used as second classifier method. Both models for severity and type of accident use *entropy* as a function to measure the quality of a split. The maximum depth of the tree is chosen to be 4. A schematic representation of the tree model to predict the type of collision is included in Fig. 8 at the Appendix. For the severity of the collisions, an accuracy of 75% is obtained, while the the prediction for the collision type yields an accuracy of 53%.

`Support Vector Machine` — A different classification method is the C-Support Vector Classification (SVC). Training data for the severity and type of accident is fitted, and predictions for the testing data are generated. The accuracy obtained for the severity is 75%, and 52% for the collision type.

`Logistic Regression` — Despite its name, logistic regression corresponds to a linear model for classification rather than regression. For this model we chose the inverse of regularisation strength $C = 0.01$, which means strong regularisation. The prediction results are shown in Fig. 5**d**. For the severity we obtain an accuracy of 75%, and for the collision type the accuracy is 54%.

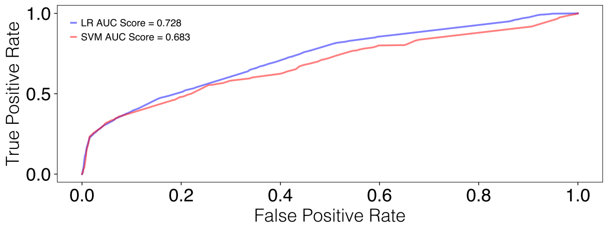

FIG. 6: **ROC curves** for LR and SVM collision type prediction models. The respective area under the curve (AUC) for each ROC line is also specified.

IV. DISCUSSION

Observations plotted in Fig. 2**a** indicate that the number of collisions has decreased for the timescale considered here. In addition, no specific year seasons or weekdays have been identified as critical. Using the locations provided, we have classified the collision according neighbourhoods, and only regions with $\leq 4,000$ incidents have been considered. In total, we analysed the historic collision data for 13 neighbourhoods, of which First Hill and Central Business District show the highest rate (see Fig. 7 in the Appendix).

Model	Accuracy	Jaccard	F1-Score
kNN	0.745	0.2174	0.688
DT	0.615	0.308	0.625
SVM	0.749	0.245	0.699
LR	0.748	0.251	0.701

TABLE I: **Severity.** Model evaluation.

Model	Accuracy	F1-Score
kNN	0.534	0.465
DT	0.382	0.398
SVM	0.525	0.456
LR	0.537	0.468

TABLE II: **Collision Type.** Model evaluation.

We have looked into the number of persons, pedestrians and vehicles involved in the historic collisions, and for most cases these numbers are minimum. Nonetheless, few incidents involving 10 vehicles and more than 40 persons have been recorded previously. In addition, we have investigated the various conditions that promote accidents. It is noted the most of the accidents occurred without the influence of alcohol or drugs. However, 42% of collisions involving use of alcohol/drugs result in injury. Interestingly, the majority of accidents in high-rate neighborhoods occurred with non-adverse conditions, e.g. clear weather, dry roads and during daylight. Fortunately, most of the accidents result only in property damage rather than injury (Fig. 1**a**). Under these conditions, the most common types of incidents involve parked cars, and rear collisions.

Predicting models have been build using the previously analysed data. We have used four different algorithms: kNN, DT, SVM, and LR. Each model has been tested to predict both the severity and type of collision. Moreover, these models have been evaluated using metric tools, such as accuracy, Jaccard index, and F1 score. The evaluation results for each prediction are displayed in Tabs. I and II, respectively. Note that a multi variable approach has been used for predicting the collision type (Tab II, and thus, the lower scores with respect or results in Tab. I. While the accuracy is comparable for all the models when predicting the severity, the F1-score points to the Logistic Regression as the best. Similarly, for the collision type, both accuracy and F1-score show best results when using the same model. We evaluate the sensitivity of this model against the SVM model using receiver operating characteristic (ROC) curves (Fig. 6). Additionally, the area under the curve (AUC) shows a slight increase for sensitivity and specificity when using the LR model.

V. CONCLUSION

Car collisions in Seattle, MA. occur with a high rate, as recorded by the SDP from 2004 to present. Historically, specific regions of the city show higher rates than others, to mention, the First Hill neighbourhood. Here, we have analysed the different types of collisions and their severity only in neighbourhoods with high number of incidents. In addition, the different conditions that lead to these accidents have been investigated. It is noted that most of the collisions occur during optimal weather, road and light conditions. Fortunately, most of those result only in property damage.

Prediction models have been trained and tested using the data for the neighbourhoods of interest. Models for both severity and type of collision have been used. Overall, we find more accurate results when predicting the severity of the accident. On the other hand, the multi variable models for the collision type show less accurate predictions. Finding the critical conditions that promote collisions might be a way to improve the predictions for the type of collisions. The model evaluation indicates that the model with best results uses Logistic Regression for both severity and collision type.

It is worth noting that despite having a large number of collisions, Seattle is not listed among the top 25 cities with most accidents [4]. To mention, among the most car accident-prone cities in the United States, we find Columbus, OH., and Los Angeles, CA. Analysis and prediction of this kind might be beneficial for road planning, design of safe routes, car protection analysis, and most importantly, to reduce the number of accidents.

VI. APPENDIX

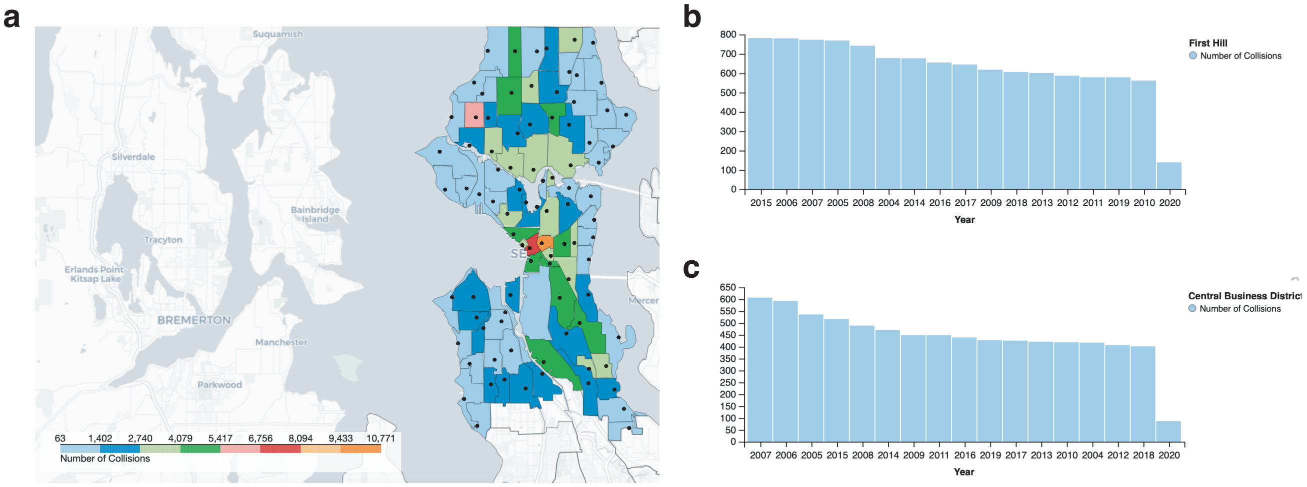

FIG. 7: **Collisions per Neighbourhood.** **a.** Coloured neighborhoods according their local number of historic collisions. Collision rates in **b.** First Hill and **c.** Central Business District. See more info in the Supplementary Material [3].

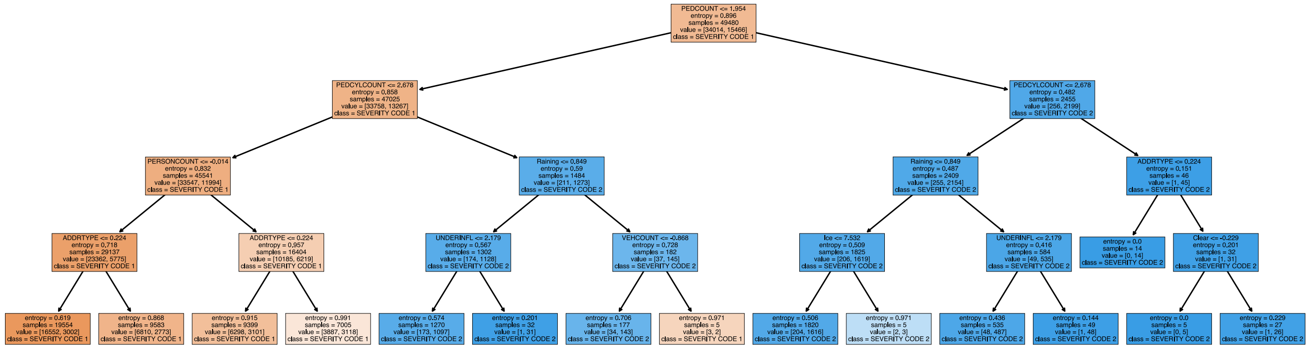

FIG. 8: Schematic diagram for the classification using a `Decision Tree` for the type of collision.

References

[1] Fatality Analysis Reporting System URL 2010

[2] Seattle Collisions — All Years URL 2020

[3] Supplementary Information for Collisions in Seattle. URL

[4] When Cars Collide: America's Most Car Accident-Prone Cities Owen, T. URL 2018