Car Accidents Severity in Seattle

Nastaran Dashti

November 16, 2020

1 Introduction

Road traffic injuries are within the top ten cause of death worldwide[1]. Every year, over 1.35 million people lose their life on roads which 150 thousands of them are from the United States. Over half of all road traffic deaths are among pedestrians, cyclists and motorcyclists.

Seattle, one of the fast-growing city in the West Coast of the US with a population of 3.95 million [2], recorded the highest number of car accidents in the entire country [3] in 2015. To reduce the number and severity of accidents in the future, we need to study the car accident data over previous years. Here, an analysis and prediction of the severity and number of accidents using various factors such as weather, road condition, location and sobriety of driver, and many other parameters that I will elaborate later in data section, is reported.

This report will guide first, the Seattle government on how to manage and reduce severity of car accidents, and second, the drivers on how to prevent getting into a car accident by changing either their travel plan or driving behavior.

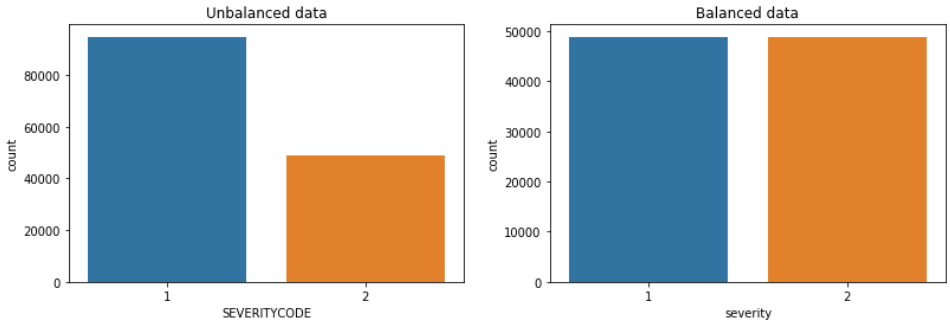
2 Data

The data is provided by SDOT Traffic Management Division [4] from 2014 to present. Here, for each accident around 37 features are considered. Some of them are listed below:

- LOCATION: Latitude (X) and longitude (Y) of accident.
- ADDRTYPE: Collision address type included alley, block, intersection.
- JUNCTIONTYPE: Category of junction at which collision took place.
- COLLISIONTYPE: Collision type such as parked Car, Angles and etc.
- WEATHER: A description of the weather conditions during the time of the collision.
- ROADCOND: The condition of the road during the collision.
- LIGHTCOND: The light conditions during the collision.
- INCDTTM: The date and time of the incident.

To build a good model, the dataset should be rich and contains many observations. Unfortunately, EXCEPTRSNCODE, EXCEPTRSNDESC, PEDROWNOTGRNT, SPEEDING, INATTENTIONIND and INTKEY have a high number of missing data, therefore I decided not to include them in the analysis.

The features SEVERITYCODE/SEVERITYDESC, representing different levels of severity caused by the accident, are used as a target. As you can see in the Fig 1(a), the majority of the accidents are related to the property damage and there is no data on the fatality. This can be interpreted in two ways; there is either a lack of information on fatality or no serious accident occurred during those years.



(a) (b)

Figure 1: The severity of the collision with (a) unbalanced, and (b) balanced data. Here (1) is related to property damage and (2) is related to injury collision

In order to improve the accuracy of the predictive machine learning models, the data needs to be balanced between the two categories. For this purpose, the resample library has been used to reduce the number of property damage, as it is shown in Fig 1(b).

Here first I analyze the feature that can predict future sever accident, then I use a supervised learning model, different classifier algorithms such as K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM) to predict the severity of an accident based on the relevant features.

3 ANALYSIS OF THE DATA

First we look at the location of accidents in Seattle map. As it is shown in Fig 2(a) the frequancy of acccidents is located at the center. By zooming in the center area, Fig 2(b), we

see most of the accident accrued in Capitol Hill, Belltown, Seattle center. We can conclude that mentioned above area are hot zones where is most probable to have accident.

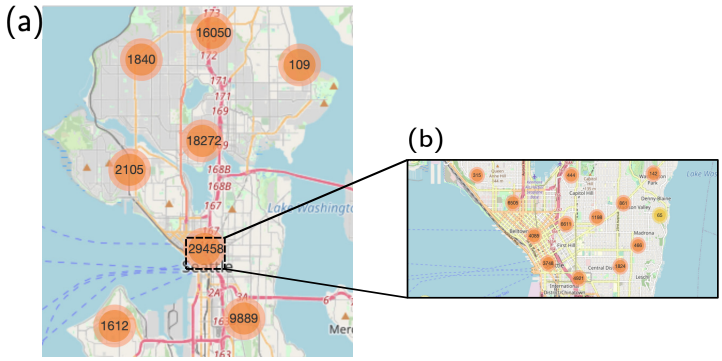


Figure 2: The number of accidents in (a) Seattle and (b) the downtown area of Seattle..

The second feature to be looked is collision type for different severity, (1) property damage and (2) injury collision . As it is expected, parked care collisions is the most type of the accident that lead to property damage, see Fig 3. However Rear end and angles collisions are ones of the most common types of car accidents that lead to injury.

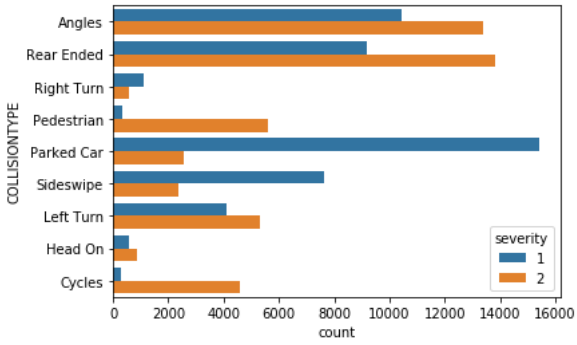
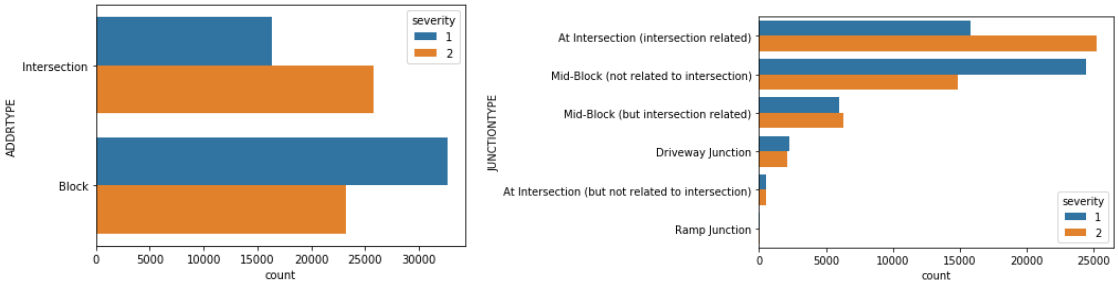


Figure 3: Number of accident based on collision type for different severity, (1)property damage (2) injury collision.

We can get the same information by looking at address type and junction type, Fig 4. While most of the collision with property damage are accrued at block, most of the collision with injury are accrued in intersection.

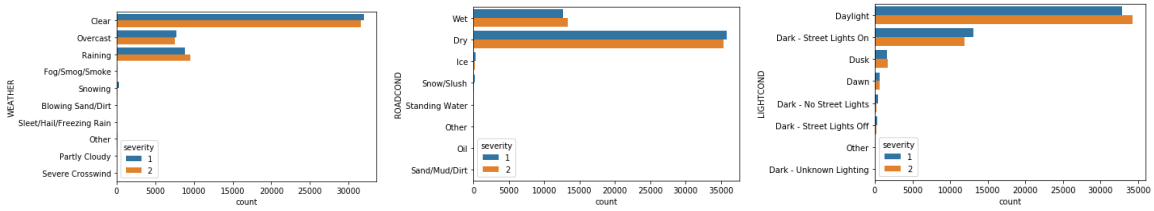


(a) (b)

Figure 4: Number of accident based on (a) address type and (b) junction type, for different severity, (1) injury collision (2) property damage.

The features of weather condition and road condition are related to each other. Supportingly, most of the accidents are accrued in clear weather and dry road, as it is shown in

Fig 5(a) and (b). By looking at light condition, Fig 5(b), we see that despite of our expectation, most of the accidents more likely to happen on daylight.



(a) (b) (c)

Figure 5: Number of accident based on (a) weather condition, (b) road condition and (c) light condition, for different severity, (1) injury collision (2) property damage.

Other feature to look at is the number of accidents in each year, Fig 6. We see that number of accident increase till 2008 and start declining between 2010–2019. Note the apparent drop-off in 2020 is likely due to a delay in reporting.

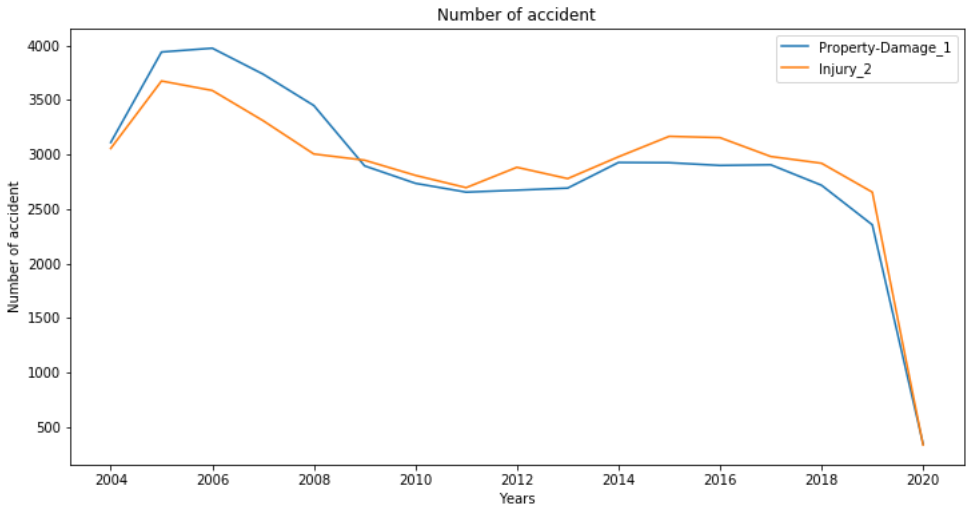


Figure 6: Number of accident per year for different severity, (1)property damage (2) injury collision.

4 Machine Learning Models

I select the most important features to predict the severity of accidents in Seattle. The predictors are weather, road and light condition. I use the following classifier algorithms: K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression(LR) and Random Forest (RF) to predict future severity.

Note that machine Learning models require numerical data, and cannot handle alphanumeric strings. Therefore using the "One-Hot Encoding" technique, the input textual column I can be replaced with a series of binary columns (containing 1 or 0) relating to each possible value in the input column. We also convert the accident severity to a binary variable 0 for property Damage and 1 for injuries.

4.1 K-Nearest Neighbor

K-Nearest Neighbors is an algorithm for supervised learning, where the data is 'trained' with data points corresponding to their classification. Once a point is to be predicted, it takes into account the 'K' nearest points to it to determine it's classification. First I find the optimized k parameters to train the model. From Fig 7 we can see, the best k to be chosen is 6. In Fig 8 we can see at classification report and confusion matrix

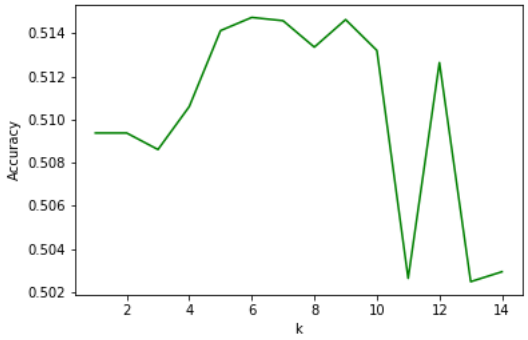
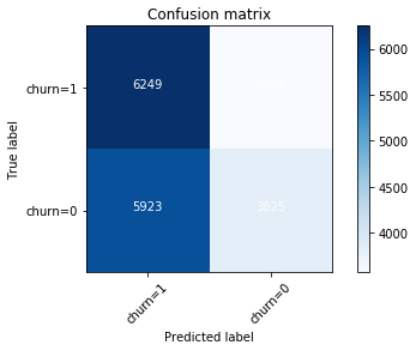


Figure 7: Accuracy of the model in terms of k.



(a)

	precision	recall	f1-score	support
0	0.52	0.39	0.45	9748
1	0.51	0.64	0.57	9823
accuracy			0.51	19571
macro avg	0.52	0.51	0.51	19571
weighted avg	0.52	0.51	0.51	19571

(b)

Figure 8: (a) Knn confusion matrix and (b)classification report.

4.2 Decision Tree

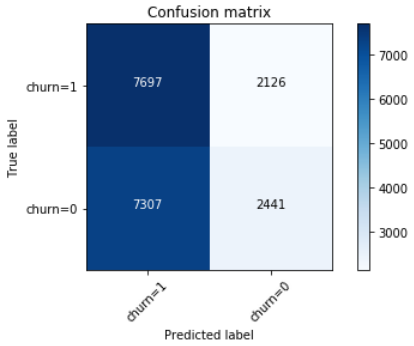
In Decision Tree algorithm build a model from historical data of accidents. Then I use the trained decision tree to predict the class of a unknown severity.

In Fig 9, we can see at classification report and confusion matrix.

4.3 Support Vector Machine

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data is transformed in such a way that the separator could be drawn as a hyperplane.

In Fig 10 we can see at classification report and confusion matrix.

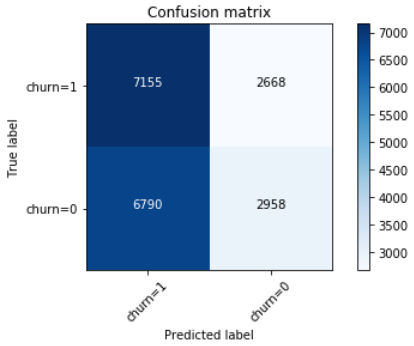


(a)

	precision	recall	f1-score	support
0	0.53	0.25	0.34	9748
1	0.51	0.78	0.62	9823
accuracy			0.52	19571
macro avg	0.52	0.52	0.48	19571
weighted avg	0.52	0.52	0.48	19571

(b)

Figure 9: (a) DT confusion matrix and (b)classification report.



(a)

	precision	recall	f1-score	support
0	0.53	0.30	0.38	9748
1	0.51	0.73	0.60	9823
accuracy			0.52	19571
macro avg	0.52	0.52	0.49	19571
weighted avg	0.52	0.52	0.49	19571

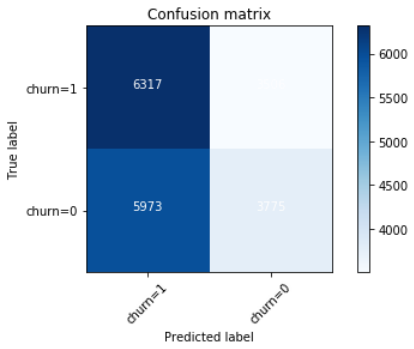
(b)

Figure 10: (a) SVM confusion matrix and (b) classification report.

4.4 Logistic Regression

Logistic Regression is a variation of Linear Regression, useful when the observed dependent variable, y, is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables.

In Fig 11 we can see at classification report and confusion matrix.



(a)

	precision	recall	f1-score	support
0	0.52	0.39	0.44	9748
1	0.51	0.64	0.57	9823
accuracy			0.52	19571
macro avg	0.52	0.52	0.51	19571
weighted avg	0.52	0.52	0.51	19571

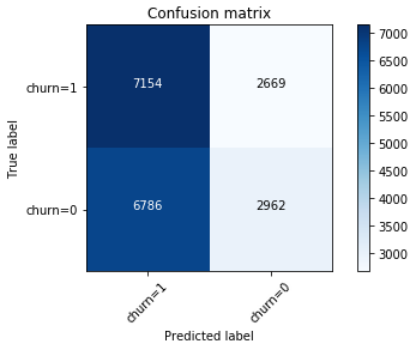
(b)

Figure 11: (a) LR confusion matrix and (b) classification report.

4.5 Random Forest

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forest has nearly the same hyperparameters as a decision tree

In Fig 12 we can see at classification report and confusion matrix.



(a)

	precision	recall	f1-score	support
0	0.53	0.30	0.39	9748
1	0.51	0.73	0.60	9823
accuracy			0.52	19571
macro avg	0.52	0.52	0.49	19571
weighted avg	0.52	0.52	0.49	19571

(b)

Figure 12: (a) DT confusion matrix and (b) classification report.

4.6 Results

We compare the actual values and predicted values to calculate the accuracy of the models. Evaluation metrics provide a key role in the development of a model, as it provides insight to areas that require improvement. There are different model evaluation metrics, I used Jaccard , F1 score and accuracy to compare theses models.

Jaccard is the size of the intersection divided by the size of the union of two label sets. If the entire set of predicted labels for a sample strictly match with the true set of labels, then the subset accuracy is 1.0; otherwise it is 0.0.

The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. To be conclude, all

Algorithm	Jaccard	F1-score	accuracy
KNN	0.40	0.51	0.51
Decision Tree	0.45	0.48	0.52
SVM	0.43	0.49	0.52
LogisticRegression	0.40	0.51	0.52
Random Forest	0.43	0.49	0.52

Figure 13: Performance of each machine learning classified algorithm.

algorithms have low scores, %50, that shows there is still significant variance that could not be predicted by the models in this study and need some more preprocessing.

One possibility to get low score is the similarity of the features for both type of accidents. In effect, the highest type 2 accident happen in the same light, road, weather condition of a type 1 severity accident. This similarity can cause the models to not be able to clearly define the strong attributes as most of them have the same tendency for both labels. This resemblance cause difficulty for the model to clearly define and classify a type of accident.

5 Conclusion

By studying some features, some important information about car accidents is revealed. Most of the accidents occur in the center of Seattle and not in the hayway which shows the problem is not the speed. Concerning injuries severity accident the focus has to be made in some important factors: intersections, rear end and angles collisions. The poor weather, road and light conditions do not produce significant car accident rate.

Finally, the results of the machine learning algorithms using predictors such as the weather, road and light conditions throws mediocre results. Other factors have to be considered to improve the prediction rate of the models being used.

References

[1] What do people die from? (2020). URL https://ourworldindata.org/what-does-the-world-die-from.

[2] Seattle - Wikipedia (2020). URL https://en.wikipedia.org/w/index.php?title=Seattle&oldid=984462802.

[3] Washington State Car Accident Statistics, journal = Colburn Law, year = 2020, month = Mar, url = https://www.colburnlaw.com/seattle-traffic-accidents.

[4] Traffic Management Division of Seattle, WA (2020). URL https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.