

# Capstone Project – Traffic Accident Severity Seattle

## Introduction

This report looks into traffic accidents in the Seattle Area in order to help inform drivers about how different factors can contribute to server traffic accidents that can result in fatalities.

Producing a model which can predict the severity of accidents could provide useful information to several stakeholders.

Drivers can use this information to better understand how factors can increase or decrease their chances of being involved in a serious accident. This could help improve driver's performance and awareness of other on the road.

This could also help traffic officials ensure they are prepared for severe accidents as the model may be able to predict when such accidents are more likely to occur based on the conditions. As a result, they can adjust staff hours to accommodate for when sever accidents are more likely.

Hospital and fire departments would be able to plan to ensure enough resources available for when accidents happen.

## Data

The data consists of accidents reports from 2004, this contains the severity of the accidents as well as many features which could be used as predictors of the severity of an accident. The dataset contains almost 200,000 rows. Although some rows are incomplete therefore the data needs to be cleansed before models can be build using the dataset.

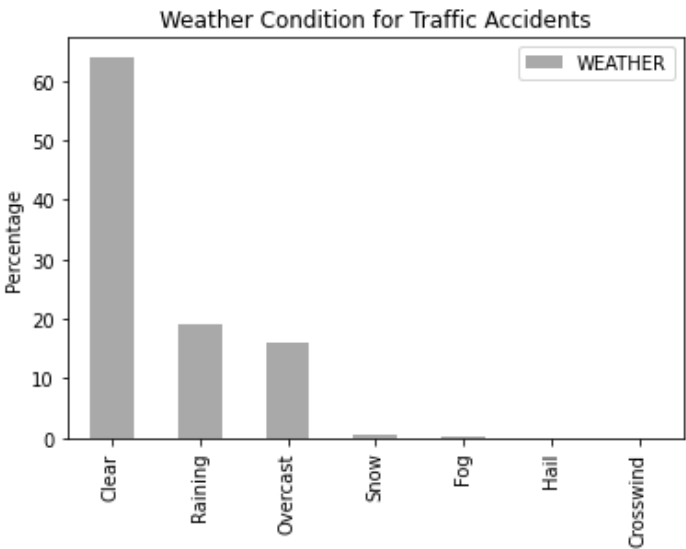
The features selected were weather, light condition, speeding, collision type and vehicle count. These were selected as I believe that these features would have the most influence on the severity of an accident.

The data were split into training and testing, with 20% being for testing. The features used as predictors have been normalised to prepare the data for the machine learning models.

## Methodology

Once the dataset had been cleansed, an exploratory analysis of the weather conditions for the accidents in the dataset was conducted.

This was to see which weather conditions were most common as this could have an influence on the severity of the accident. For example, snow conditions tend to make vehicles more challenging to control which could have an influence on the severity of an accident. As the graph shows over 60% of accidents occurred in clear conditions.



Features were grouped with the severity code in order to visualise how the severity of the accidents were distributed across the variables in the given feature.

WEATHER		SEVERITYCODE	
Clear	1	73671	
	2	35662	
Fog/Smog/Smoke	1	376	
	2	184	
Overcast	1	18589	
	2	8678	
Raining	1	21596	
	2	11085	
Severe Crosswind	1	18	
	2	7	
Sleet/Hail/Freezing Rain	1	82	
	2	28	
Snowing	1	662	
	2	166	

This table shows the grouping of the severity codes with the different weather conditions. The most common was severity code 1 (property damage) in clear conditions. As most days have clear conditions this being the most common condition makes sense.

This table shows severity of an accident grouped with light conditions. The level of visibility can impact a driver's ability to identify potential causes of accidents. Therefore, using this as a predictor is logical as light conditions could have a significant impact on the severity of an accident.

LIGHTCOND		SEVERITYCODE	
Dark - No Street Lights	1	1097	
	2	323	
Dark - Street Lights Off	1	807	
	2	310	
Dark - Street Lights On	1	32660	
	2	14299	
Dawn	1	1618	
	2	807	
Daylight	1	75035	
	2	38163	
Dusk	1	3777	
	2	1908	

The next stage is to create machine learning models which could help predict future accidents severity using the data from the dataset.

In order to determine which model would be the most more accurate the data were split into training and testing sections.

K-nearest neighbour, decision tree, support vector machine and logistic regression models were built and then each model was evaluated using Jaccard index and f1-score in order to determine which should be used for the prediction.

The next model built was a decision tree. Testing various models will give an indication as to which will be the most appropriate to use to predict the severity of an accident in the future.

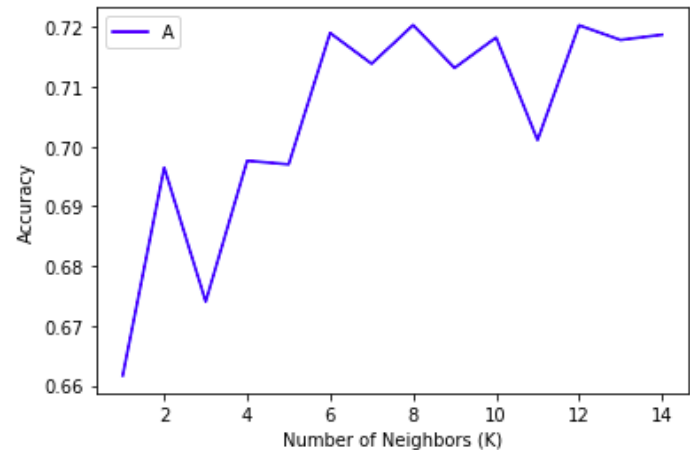
Support Vector machine creates a boundary a hyperplane between the two categories. This separates the two values of the target. As with every model the SVM was evaluated in order to determine which model would be most useful.

The final model that was tested was logistic regression. The goal of this model is to reduce the cost, the different between the predicted values and the real values of  $y$ . This model differentiates itself from the others as a predicted value probability is also calculated. This gives insight into the model's confidence level for the values it has outputted. This model was also evaluated using log loss which provides another method of evaluating the model.

## Results

### KNN

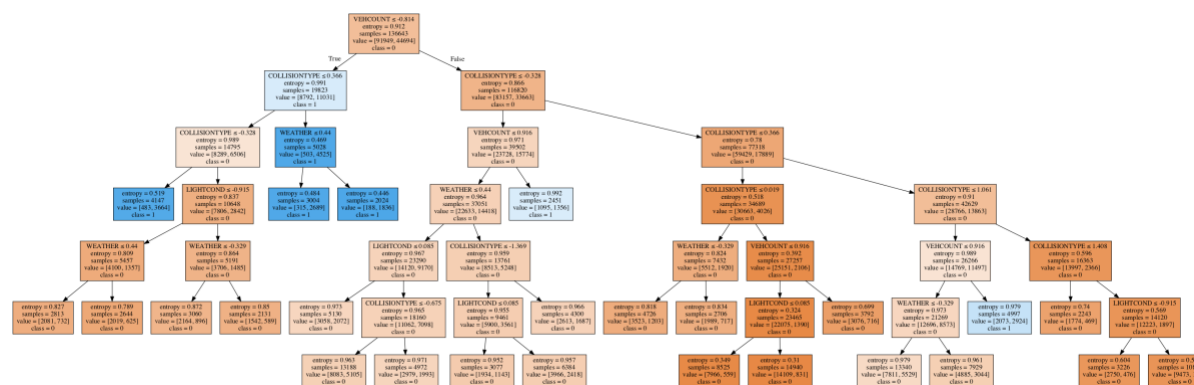
Before the KNN model could be evaluated the best value of K had to be found. The graph shows the values of K which were tested in a for loop in order to determine the most accurate value and consequently the value which the model would be evaluated on. Although the chart shows several values being very similar. The most accurate value was found to be 8. This was through using the `max()` function on the `mean_acc` list that saved the results for each value of K.



Once the K value was determined an F1-Score and Jaccard score were calculated which is what each model will be compared on. The value of the Jaccard index was 0.68. The F1-score was 0.69, using a weighted average. Each model's jaccard and F1 score can be found at the end of the results section in a table to make comparison of the models easier.

### Decision Tree

The decision tree was tested using different values for the max depth and min sample leaf, the most accurate model was found using a max depth of 6 and a min leaf sample of 2000. The decision tree scored higher values for the Jaccard and f1 score than the KNN with 0.71 and 0.78 respectively. The graph shows a visualisation of the decision tree the model produced.



## Support Vector Machine

The support vector machine had some difficulties fitting the data. This was not loading in my notebook. Although attempts were made to resolve this issue in the end a smaller dataset had to be used in order to fit the model and be able to evaluate the SVM's accuracy. The model was amongst the most accurate with a Jaccard score of 0.71 and an f1-score of 0.78. This made determining which model to select more difficult as the SVM was more accurate than others but had used a smaller dataset. Having to train the model with the whole dataset in order for it to be used to create future prediction would not be possible with the computer being used.

## Logistic Regression

The logistic regression model also calculated a log loss score using the probability of y values. The Jaccard score was in line with the other models being used at 0.68. The log loss score was 0.62 while the f1-score was far lower than any other model at 0.55.

Using the accuracy scores to compare the models, the model that is suggested to be used to create future accident severity predictions is the Decision Tree model as this model preformed the best during the training.

Below is a table showing the evaluation metrics of each model.

Model	Jaccard Score	F1-Score	Log-Loss
KNN	0.67	0.69	N/A
Decision Tree	0.71	0.78	N/A
SVM	0.71*	0.78*	N/A
Logistic Regression	0.68	0.55	0.62

\*Reduced dataset was used by to computing limit.

## Discussion

The recommendation I would make based on the results would be to use a decision tree model to predict future accident severity. An observation that came from initial analysis of the data was that over 60% of accidents occur during clear conditions. Although this figure was expected to be high as most days have clear weather conditions. This highlights to drivers that accidents can occur during all conditions and paying full attention to the road is important even on good condition days.

The dataset used only contained property damage and injury as the classes in the severity column. Using a dataset that contains data with more classifications of severity could help

to better predict the type of accident rather than a generalised injury which using this dataset produces.

The issue of the SVM not loading when attempting to fit the full training data may have impacted the final decision as to which model to use. Finding a way to speed up processing of fitting the data to the model would give a better result of the accuracy score as it would have used the same amount of data as the other models therefore making comparison fairer.

## Conclusion

The project objectives were to produce a model a model using the Seattle traffic accident data in order to predict the severity of future accidents. This was completed through testing several machine learning models in order to determine which would be the most suitable to make future predictions.

Collecting information of individuals involved in accidents could also be an area for future exploration. Do people who are younger tend to get in more severe accidents? Do people who wear glasses more likely to be involved in an accident when light conditions are poor. The locations of accidents could also be explored. Identifying specific locations that tend to yield more severe accidents could contribute to alterations being made to the road to attempt to reduce this issue.

This could also aid first responses as they are aware of areas that tend to need their help more.