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.

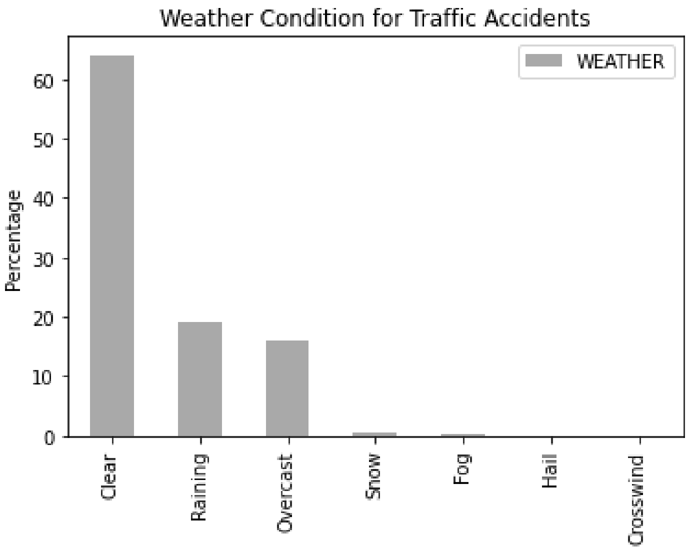
Drivers can use this information to better understand how factors can increase or decrease their chances of being involved in a serious accident. This could help improve driver's performance and awareness of other on the road.

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.

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.

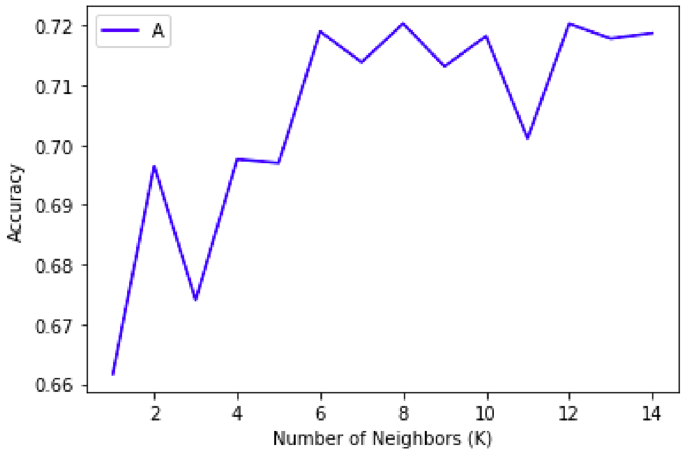
Support Vector machine creates a boundary a hyperplane between the two categories. This separates the two values of the target. As with every model the SVM was evaluated in order to determine which model would be most useful.

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

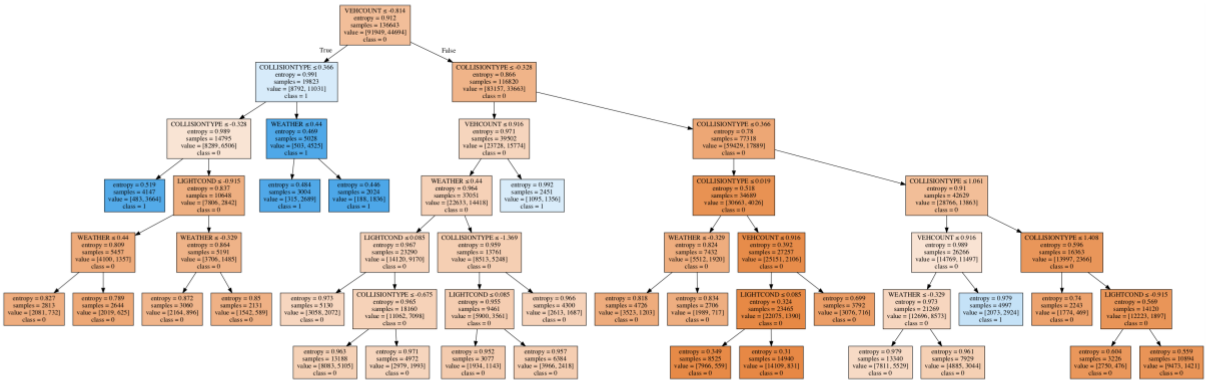
Before the KNN model could be evaluated the best value of K had to be found. The graph shows the values of K which were tested in a for loop in order to determine the most accurate value and consequently the value which the model would be evaluated on. Although the chart shows several values being very 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.


VEHCOUNT ≤ -0.814 entropy = 0.912 samples = 136643 value = [91949, 44694] class = 0
True
False
COLLISIONTYPE ≤ 0.366 entropy = 0.991 samples = 19823 value = [8792, 11031] class = 1
COLLISIONTYPE ≤ -0.328 entropy = 0.866 samples = 116820 value = [83157, 33663] class = 0
COLLISIONTYPE ≤ -0.328 entropy = 0.989 samples = 14795 value = [8289, 6506] class = 0
WEATHER ≤ 0.44 entropy = 0.469 samples = 5028 value = [503, 4525] class = 1
VEHCOUNT ≤ 0.916 entropy = 0.971 samples = 39502 value = [23728, 15774] class = 0
COLLISIONTYPE ≤ 0.366 entropy = 0.78 samples = 77318 value = [59429, 17889] class = 0
entropy = 0.519 samples = 4147 value = [483, 3664] class = 1
LIGHTCOND ≤ -0.915 entropy = 0.837 samples = 10648 value = [7806, 2842] class = 0
entropy = 0.484 samples = 3004 value = [315, 2689] class = 1
entropy = 0.446 samples = 2024 value = [188, 1836] class = 1
WEATHER ≤ 0.44 entropy = 0.964 samples = 37051 value = [22633, 14418] class = 0
entropy = 0.992 samples = 2451 value = [1095, 1356] class = 1
COLLISIONTYPE ≤ 0.019 entropy = 0.518 samples = 34689 value = [30663, 4026] class = 0
COLLISIONTYPE ≤ 1.061 entropy = 0.91 samples = 42629 value = [28766, 13863] class = 0
WEATHER ≤ 0.44 entropy = 0.809 samples = 5457 value = [4100, 1357] class = 0
WEATHER ≤ -0.329 entropy = 0.864 samples = 5191 value = [3706, 1485] class = 0
LIGHTCOND ≤ 0.085 entropy = 0.967 samples = 23290 value = [14120, 9170] class = 0
COLLISIONTYPE ≤ -1.369 entropy = 0.959 samples = 13761 value = [8513, 5248] class = 0
WEATHER ≤ -0.329 entropy = 0.824 samples = 7432 value = [5512, 1920] class = 0
VEHCOUNT ≤ 0.916 entropy = 0.392 samples = 27257 value = [25151, 2106] class = 0
VEHCOUNT ≤ 0.916 entropy = 0.989 samples = 26266 value = [14769, 11497] class = 0
COLLISIONTYPE ≤ 1.408 entropy = 0.596 samples = 16363 value = [13997, 2366] class = 0
entropy = 0.827 samples = 2813 value = [2081, 732] class = 0
entropy = 0.789 samples = 2644 value = [2019, 625] class = 0
entropy = 0.872 samples = 3060 value = [2164, 896] class = 0
entropy = 0.85 samples = 2131 value = [1542, 589] class = 0
entropy = 0.973 samples = 5130 value = [3058, 2072] class = 0
COLLISIONTYPE ≤ -0.675 entropy = 0.965 samples = 18160 value = [11062, 7098] class = 0
LIGHTCOND ≤ 0.085 entropy = 0.955 samples = 9461 value = [5900, 3561] class = 0
entropy = 0.966 samples = 4300 value = [2613, 1687] class = 0
entropy = 0.818 samples = 4726 value = [3523, 1203] class = 0
entropy = 0.834 samples = 2706 value = [1989, 717] class = 0
LIGHTCOND ≤ 0.085 entropy = 0.324 samples = 23465 value = [22075, 1390] class = 0
entropy = 0.699 samples = 3792 value = [3076, 716] class = 0
WEATHER ≤ -0.329 entropy = 0.973 samples = 21269 value = [12696, 8573] class = 0
entropy = 0.979 samples = 4997 value = [2073, 2924] class = 1
entropy = 0.74 samples = 2243 value = [1774, 469] class = 0
LIGHTCOND ≤ -0.915 entropy = 0.569 samples = 14120 value = [12223, 1897] class = 0
entropy = 0.963 samples = 13188 value = [8083, 5105] class = 0
entropy = 0.971 samples = 4972 value = [2979, 1993] class = 0
entropy = 0.952 samples = 3077 value = [1934, 1143] class = 0
entropy = 0.957 samples = 6384 value = [3966, 2418] class = 0
entropy = 0.349 samples = 8525 value = [7966, 559] class = 0
entropy = 0.31 samples = 14940 value = [14109, 831] class = 0
entropy = 0.979 samples = 13340 value = [7811, 5529] class = 0
entropy = 0.961 samples = 7929 value = [4885, 3044] class = 0
entropy = 0.604 samples = 3226 value = [2750, 476] class = 0
entropy = 0.559 samples = 10894 value = [9473, 1421] class = 0

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.