Motor Vehicle Accident Analysis

Applied Data Science Capstone

Charles Gagalac

Introduction

For the year 2018, the National Safety Council estimated 36,000 motor vehicle crashes occurred in the United States that involved 54,100 vehicles. 39,000 people lost their lives while 4.5 million were injured. If drivers had information regarding potential dangers on the roads, there would be less traffic incidents that would result in injury or fatality. Therefore, an analysis on whether road conditions or weather is related to motor vehicle accidents may be helpful. Machine learning will be used to classify motor vehicle accident severity relative to weather and road conditions.

Data and Methodology

The collisions data from Seattle GeoData (https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions) is used for the analysis. The dataset contains 221,266 records of 40 attributes ranging from 2004 to 2018. Of the 40 attributes, only 4 are relevant to the analysis: severity code, weather, road conditions, and light conditions. The severity code is the target, while the weather, road conditions and light conditions attributes are the features.

The classification machine learning models utilized to determine the relationship between the severity code target variable and the aforementioned feature attributes are K Nearest Neighbors, Decision Tree, Support Vector Machine, and Logistic Regression.

Data Preparation

Prior to building the classification machine models, the data needs to be pre processed. First, all rows with missing values from the severity code, weather, road conditions, light conditions attributes are eliminated. Rows that contain ‘Unknown’ or ‘Other’ values from the weather, road conditions, and light conditions labels are also dropped. Missing and ambiguous values may needlessly convolute the analysis.

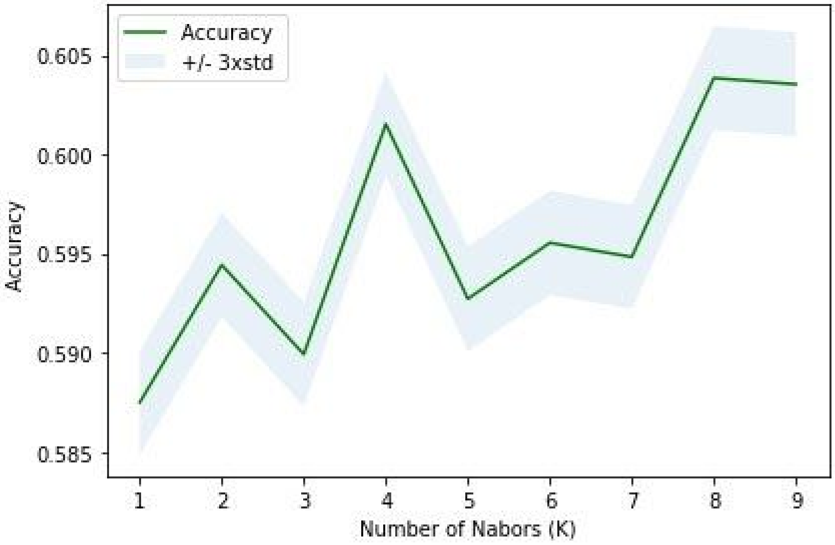
Next, the personal injury codes are consolidated into one personal injury code. The severity codes are broken down into ‘1’ for property damage, ‘2’ for injury, ‘2b’ for serious injury, and ‘3’ for fatality. After removing the missing, ‘Unknown’, and ‘Other’ values, property damage accounts for 115,548 incidences, while the injury, serious injury, and fatality designations have 56,378, 2985, and 333 values respectively. Merging the codes involving personal injury makes personal injury more meaningful relative to property damage.

Lastly, one hot encoding is used on the weather, road conditions, and light conditions features to balance out the data. Labeling, where values are numerically remapped, would increase the likelihood of bias towards the higher numerical values.

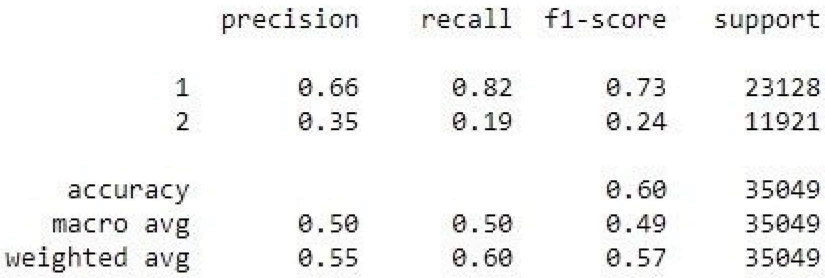
Modeling Results

k Nearest Neighbor

The k Nearest Neighbor algorithm achieved a best accuracy of 60% with a k of 8.



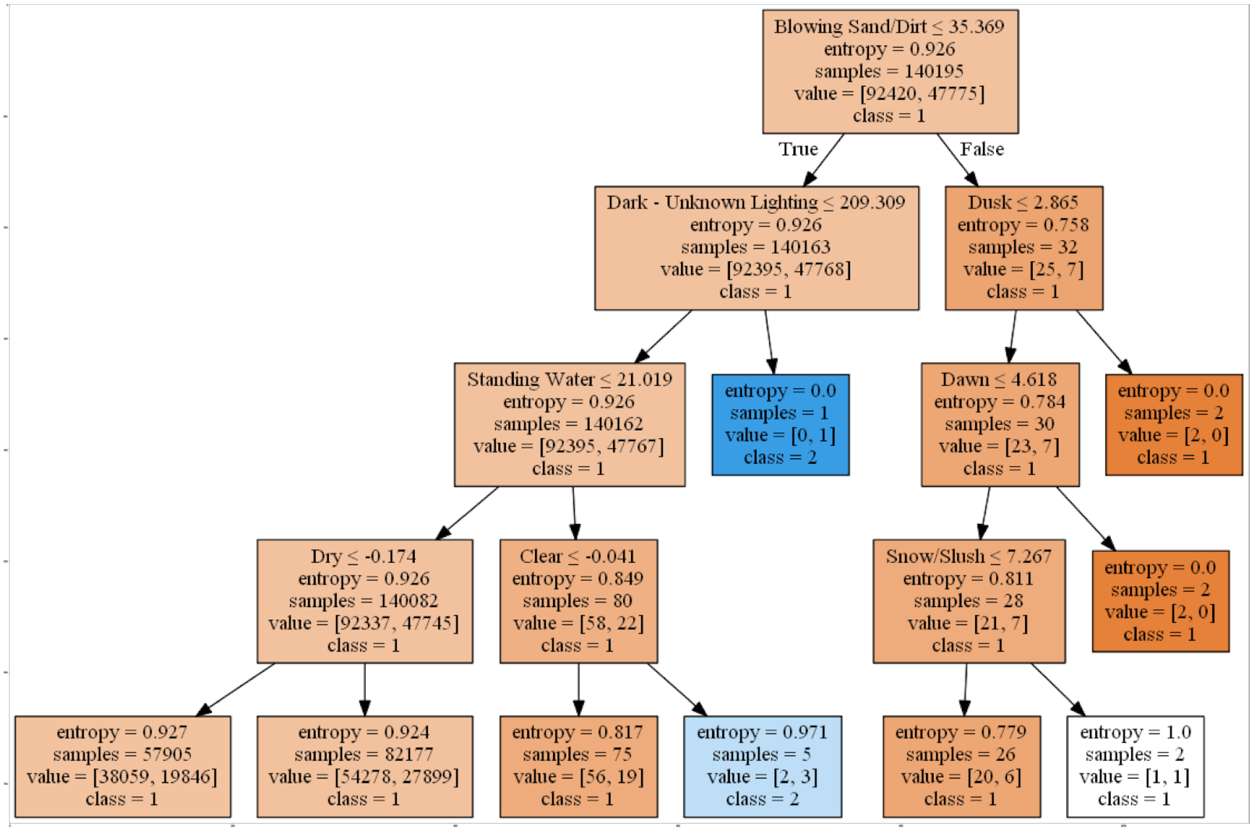
k-NN Classification Report



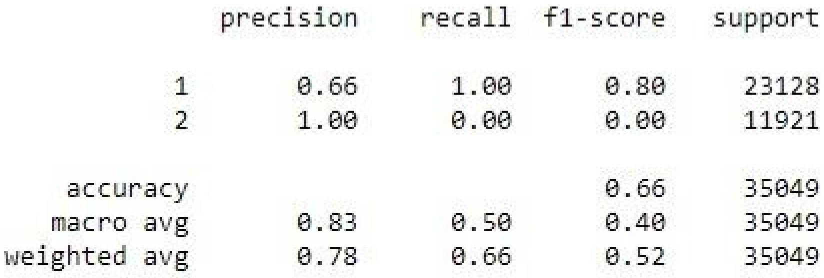
```
              precision    recall  f1-score   support

           1       0.66      0.82      0.73     23128
           2       0.35      0.19      0.24     11921

    accuracy                           0.60     35049
   macro avg       0.50      0.50      0.49     35049
weighted avg       0.55      0.60      0.57     35049
```

Decision Tree

The Decision Tree model produced a 66% accuracy.



Decision Tree Classification Report

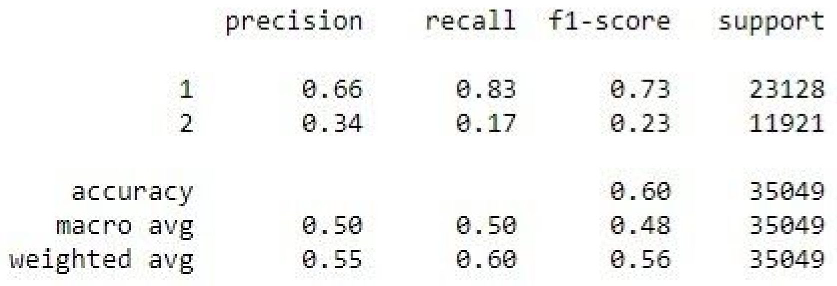


```
              precision    recall  f1-score   support

           1       0.66      1.00      0.80     23128
           2       1.00      0.00      0.00     11921

    accuracy                           0.66     35049
   macro avg       0.83      0.50      0.40     35049
weighted avg       0.78      0.66      0.52     35049
```

Support Vector Machine

The Support Vector Machine model produced a 56% accuracy.

Support Vector Machine Classification Report

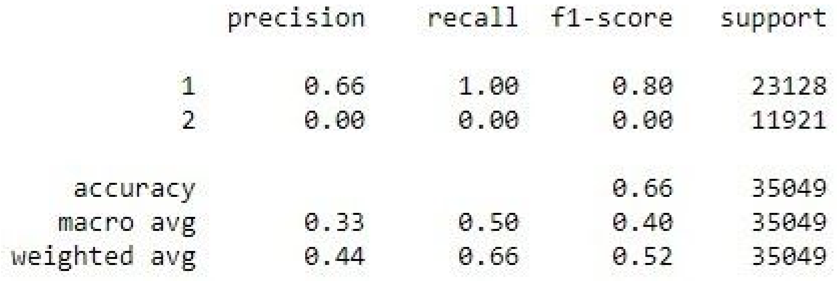


```
              precision    recall  f1-score   support

           1       0.66      0.83      0.73     23128
           2       0.34      0.17      0.23     11921

    accuracy                           0.60     35049
   macro avg       0.50      0.50      0.48     35049
weighted avg       0.55      0.60      0.56     35049
```

Logistic Regression

With a C set at .01 and a liblinear solve, Logistic Regression obtained a 66% accuracy.

Logistic Regression Classification Report



```
              precision    recall  f1-score   support

           1       0.66      1.00      0.80     23128
           2       0.00      0.00      0.00     11921

    accuracy                           0.66     35049
   macro avg       0.33      0.50      0.40     35049
weighted avg       0.44      0.66      0.52     35049
```

Discussion

Although k-Nearest Neighbor and Support Vector Machine did reasonably well, decision tree and logistic regression performed the best. Both had identical and highest f1 score, precision, recall, and accuracy scores.

However, these results really only apply to classifying incidents involving property damage. All models were simply not able to reliably classify accidents that resulted in personal injury. The analysis seems to have suffered from an imbalanced dataset where property damage outnumbers personal injury 2 to 1.

Conclusion

The analysis shows that road conditions and weather are linked to motor vehicle accidents, accidents resulting in property damage. When it comes to personal injury outcomes, the modeling is deficient. A dataset where the personal injury and property damage outcomes are balanced or similar in number would result in better machine learning models. Nevertheless, drivers should take heed of their driving conditions prior to setting out on the roads.