IBM Data Science Capstone Car Accident Severity

Introduction

Car accidents occur on an hourly basis, but what if we could predict whether a car accident will be more severe depending on the weather conditions preceding the accident, and whether the implications of these weather conditions can help the area plan for possible life and death situations.

The question I will be trying to answer is “Does rainy weather increase the severity of accidents?”

Target Audience: The results of the data from the model can help the City of Seattle government improve road safety and local hospitals whether to have more staff on call for a specific period or weather event.

Data

The data is sourced from the Seattle Police Department from 2004 to present. Included in the data is the weather during the accident, whether speeding was a factor, and the severity of injuries from the collision. Using this data, I will run a supervised model to predict if rainy conditions have an effect on the severity of injuries that occur from an resulting accident given specific weather conditions at the time.

The data set contains 37 variables. “SEVERITYCODE” ranks the accident severity from 1 to 2, with 2 indicating serious injury. “WEATHER” indicates the weather conditions present at the time of the accident

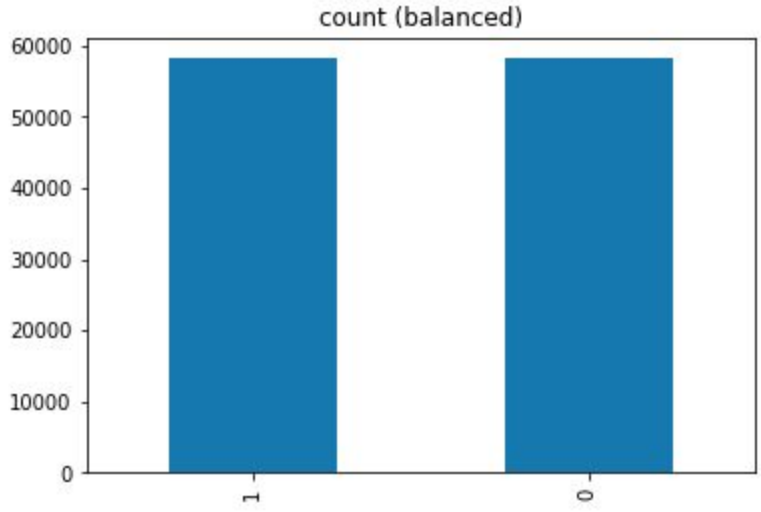
Methodology

Running through the data analysis, first SEVERITYCODE had to be converted from 1 as property damage to 0 and 2 for serious injury to 1 to index for the models.

For the weather condition, I transformed the inputs to simplify the categories, ranking the conditions from clear to adverse weather conditions: 0 for Unknown, 1 for clear, 2 for cloudy, 3 for wind and visibility impaired weather conditions, and 4 for rain and snow which affect visibility and road surface conditions, along with traction.

Multiple variables in the data set contained Null values, such and thus were dropped from the data set as they were insignificant due to the missing data. While these factors are probably important in real life in determining accident severity, the missing data makes the variables useless.

Another issue that arose during analysis is that most accidents are property damage only, thus I had to balance the data. Balancing was required as given the data set weighing heavily with property damage, this would cause a bias in machine learning algorithms to indicate property damage no matter the weather condition. Using the already converted SERVERITYCODE I randomly under-sampled property damage(0) data to even out with existing data containing injuries(1).



Finally, the following models were chosen to be used for machine learning: Decision Tree, Logistic Regression, and K Nearest Neighbors.

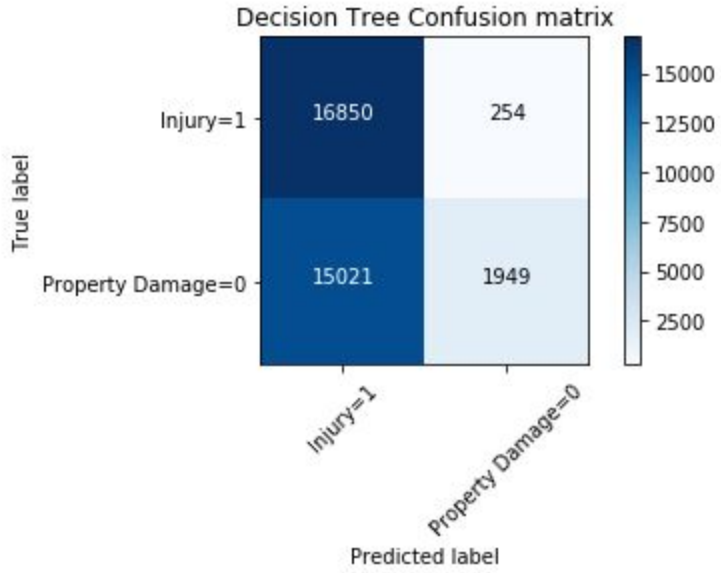
(Please continue to next page for results)

Results

Decision Tree

Max depth of 4 used in the Decision Tree model.

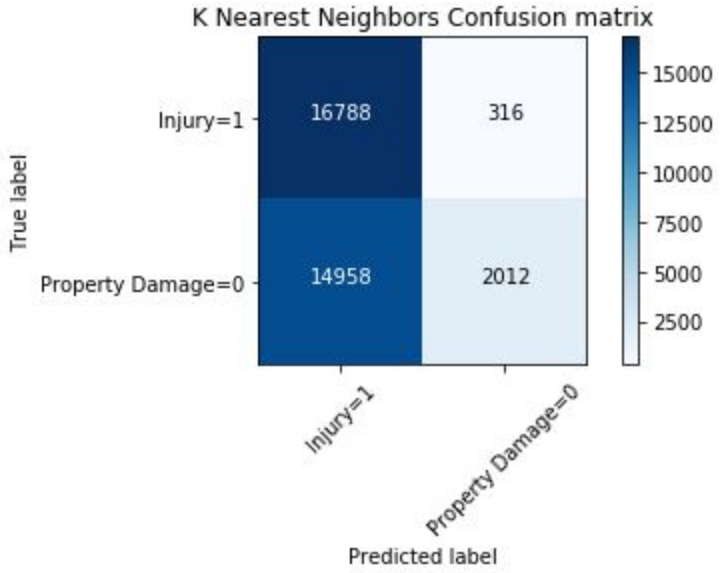
```
DecisionTrees's Accuracy:  0.5517109819803956
```



	precision	recall	f1-score	support
0	0.88	0.11	0.20	16970
1	0.53	0.99	0.69	17104
micro avg	0.55	0.55	0.55	34074
macro avg	0.71	0.55	0.45	34074
weighted avg	0.71	0.55	0.45	34074

K nearest neighbors

```
The best accuracy was with 0.5517403298702823 with k= 3
```



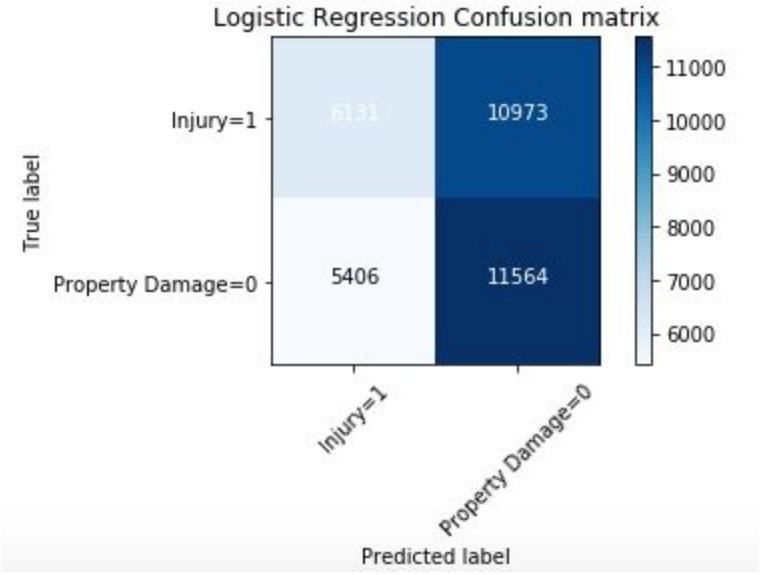
	precision	recall	f1-score	support
0	0.86	0.12	0.21	16970
1	0.53	0.98	0.69	17104
micro avg	0.55	0.55	0.55	34074
macro avg	0.70	0.55	0.45	34074
weighted avg	0.70	0.55	0.45	34074

Logistic Regression

C of .01 used in the Logistic Regression model.

Logistic Regression Accuracy

```
0.5193109115454598
```



	precision	recall	f1-score	support
0	0.51	0.68	0.59	16970
1	0.53	0.36	0.43	17104
micro avg	0.52	0.52	0.52	34074
macro avg	0.52	0.52	0.51	34074
weighted avg	0.52	0.52	0.51	34074

Discussion

While both the decision tree and K nearest neighbors model had a precision of 70%, the logistic regression model was the outlier at 52% precision. Both decision tree models and K nearest neighbors had high recall too when Weather condition was the only variable.

Only a slight correlation between weather and an injury occurring in the F1 scores of the decision tree and K nearest neighbors was shown with both at 69%. Surprisingly, the weather alone showed only about a 20% correlation when only property damage occurred.

On a side note but important note, most accidents in the overall dataset from the Seattle Police Department cause property damage. This needs to be taken into account by the decision makers.

Conclusion

After analysis, the data showed the likelihood of an accident increasing in severity is only slightly correlated to the weather condition present at the time alone. Additionally, the weather condition at the time the accident occurred did not have an impact on the severity no matter if it was day or night, despite the different lighting conditions. It was noted by the models that when only property damage occurred, little to none correlation existed with what the weather was at the time of the accident. We have to also look at the bigger picture of the data. While a correlation is only slightly shown in some of the models, most accidents in the dataset result in property damage only, with fatalities being the rarest category to occur. This ties back to our target audience.

The question has been answered of does weather affect accident severity, but does the result warrant our target audience to make a change in staffing or speed limits during a weather event, given that most accidents involve property damage more often than fatalities? That decision will be left to the decision makers.