Introduction/Business Problem:

The Seattle administration has been reviewing accident cases as a means to deploy preventive measures to avoid damage to individuals and property. An ability to convert historic accident reports and use that to help state police and authorities to create targeted awareness measures to prevent accidents will be helpful. Further, insurance companies would also benefit by preventive and pro-active counter-measures that will reduce third-party property damage.

Using existing fields from records such as environmental factors, the state can use this model to preempt accidents and plan its response accordingly. The target audience for this model are the state police department, emergency services, town planners and insurance companies that can use this model to plan better and reduce accidents.

Data

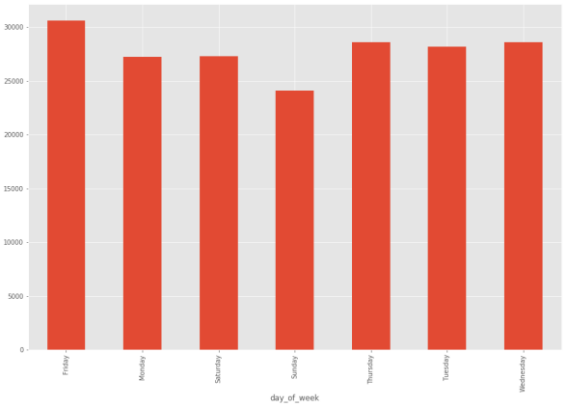
There are 136,485 severity 1 incidents that indicate property damage and 58,188 injury related incidents represented as Severity 2.

Pre-processing and analysing data

From an assessment of the data, the following features are considered to evaluate the model for severity prediction.

- Junction Type - Midblock, intersections have higher incidents.
- Weather - Rain/Overcast do indicate a higher incident rate compared to other non-Clear conditions.
- Road Cond - Wet conditions indicate higher rates. This is related to the weather. No other road conditions (other than dry) seem to drive the incident rates.
- Light Conditions - Dark(with Streetlights) has a higher rate of incidents.

- Day of the week: Derived column based on incident date - shows that incidents are higher on a Friday and lowest on Sundays. Under influence and Speeding were not considered as completeness and accuracy of data was not established. *(image alongside)*



Missing values and Balancing the Dataset

In the above graphs, there are "Unknown" values. I have replaced the "Unknown" values with Nan and dropped them. The number of incidents drops from 194,673 to 168,923. 113,271 of Severity 1 and only 55,652 of severity 2. In order to prevent any bias, Severity 1 counts have been reduced to 55,652.

Normalizing, Training and Test Data¶

Normalizing data has not been done as all of the features selected are categorical. Training and Testing split has been done in a 70:30 ratio.

Methodology - Machine Learning Algorithms

The following machine learning algorithms were used to predict severity of the incidents - whether it was property damage (severity 1) or resulted in Injury (severity 2):

- Logistic Regression
- K Nearest Neighbour
- Decision Tree
- Support Vector Machine

Model	F1 score	Jaccard Score
Logistic Regression	59%	59.3%
K-nearest Neighbour	52.3%	55.2%
Decision Tree	61.3%	59.1%
Support Vector Machine	58.5%	58.9%

Results

The results from the 4 machine learning algorithms used provided an accuracy closer to 60%. The summary is provided below:

1. Logistic Regression: 59.3%
2. KNN Classifier: 55.2%
3. Decision Tree: 59.1%
4. SVM: 58.9%

Discussion

From the data, the prediction accuracy wasn't very high. However, there were patterns when accidents were more likely to occur - such as Fridays. Authorities could create more awareness on Fridays to make people more cautious. Sundays saw relatively lower incidents.

Similarly, Wet days, intersections or Dark light conditions proved to be riskier and more accident prone.

Conclusion

At 60%, the accuracy is not great. It is higher than 50% indicating that the algorithms provide a better chance at predicting an incident regarding property damage or injury than a random guess (which is a probability of 50%). This is largely due to the insufficient size of the data that is causing a variance. There were many features such as speeding and driver under the influence that were not complete resulting in them being excluded from the model. With more data, we will be able to get better prediction abilities.