**Determining the Severity of an Accident**

James Drake

**Introduction: Business Problem**

In this project we will try to determine if we are likely to get in to an accident on a car journey depending on certain weather and road conditions. The report is targeted at insurance companies. With this model an insurance company may be able to handle claims more efficiently as they can predict whether a claim is likely to be to do with a personal injury claim or a property damage claim. If when claiming a person enters the circumstances of their accident then the model should allow them to be directed to the correct claims department.

**Data**

Based on the business problem we will use the Coursera provided data set (link shown later) and extract particular attributes of interest that can help predict whether an accident results in property damage or injury. Some key factors are:

* Road Conditions
* Light Conditions
* Weather Conditions

From this raw data set it is necessary to remove unwanted attributes and then go through various data cleaning stages to ensure the data is in the right format to model.

Any missing data is removed for the attributes of interest and any attributes that are 'unknown' or 'other' are also removed as these will not be useful in helping the predictiveness of our model.

Categorical data is encoded into binary representation by one-hot encoding in python.

The target variable is referred to as 'Severity code'. For our data set we had two outcomes for this, injury or accident.

Once the data has been cleaned we are left with a data set of shape of 167,857 samples and 26 features.

**Methodology**

As we have a categorical target variables the most effective machine learning technique here will be classification techniques. Used in this paper will be K nearest neighbours, decision tree and support vector machines

**Results**

|  |  |  |
| --- | --- | --- |
| **Model Type** | **Jaccard Index Score** | **F1 Score** |
| K-Nearest Neighbours | 0.622789 | 0.767554 |
| Decision Tree | 0.665257 | 0.798984 |
| Support Vector Machines | 0.665048 | 0.798834 |

As can be seen from the results above, the most accurate model we have is using the decision tree. This would be the model I would take forward to use on future data. With additional points added later it should be possible to further refine our model as the data set gets larger.

**Discussion**

To enhance this analysis it would be useful to get this data from many different regions both nationally and internationally. This could then try to explain any regional differences. Some demographic data for the driver may also be useful to see if there are any correlations between driver age, sex etc that may be more effective in prediction

**Conclusion**

Our chosen decision tree model has a fairly low Jaccard and F1 score and therefore further data attributes may be worth exploring to understand if there are other data points that would increase the accuracy of our model.