A description of the problem and a discussion of the background. (**15 marks**)

Introduction

**Background**

Road accidents happen everyday and there may be certain common factors that contribute to an increased likelihood of getting into an accident that is severe. By being able to predict the severity of an accident and the likelihood of getting into one, one is then able to gauge if one should change his/her route of travel or avoid traveling if conditions are poor. This can help to reduce the number of accidents, injuries and fatalities.

**Problem**

To predict the severity of any car accident given a variety of dependent variables including weather conditions, light conditions, road conditions etc.

**Interested parties**

Clearly, the police force/government agencies would be very keen on these predictions so that they can educate the public and advise caution to travel (or not) and while travelling especially during risky conditions. The public themselves will also be keen to know when they should avoid driving so as to reduce risks of getting into accidents that then lead to injuries and fatalities.

A description of the data and how it will be used to solve the problem. (**15 marks)**

Data acquisition and cleaning

**Data sources**

Data utilised is what is provided in the Example dataset. Data given include 37 columns of which one column is the target/dependent (i.e. SEVERITYCODE).

**Data cleansing**

* The dataset is slightly imbalanced, with 136,485 data points with Severity Code 1 and only 58,188 data points with Severity Code 2. Therefore, I use the methodology of upsampling to remove data imbalance to end up with an equal number of data points with Severity Codes 1 and 2 (i.e 136,485 data points each).
* Many of the variables provided are categorical – for example, weather conditions, road conditions, inattention etc. I therefore utitlised label encoders to convert these features into numerical representations.
* There is a repeat column in that of SEVERITYCODE.1 which I dropped from my dataset. Also, in my opinion, certain columns (such as coordinates, object ID) were probably not very useful in predicting the severity of an accident – I opine that knowing the nature of the location of an accident (intersection, block) for example is more useful in determining the severity of an accident rather that the specific X,Y coordinates. I therefore dropped those columns as well.
* I added two new columns – Day of Week and Time of Day – as I think that during certain day of the week, traffic is probably more busy and therefore severity of an accident may increase. Similarly time of day may be important too as visibility at night is lower than in the day and people may be more tired as well. Therefore these two new columns are added using information from the Incident dates and incident time columns.

**Final feature selection**

|  |  |  |
| --- | --- | --- |
| Kept features | Dropped features | Added features |
| ADDRTYPE | X | DAYOFWEEK |
| COLLISIONTYPE | Y | TIMEOFDAY |
| PERSONCOUNT | INCKEY |  |
| PEDCOUNT | COLDETKEY |  |
| PEDCYLCOUNT | REPORTNO |  |
| VEHCOUNT | STATUS |  |
| JUNCTIONTYPE | INTKEY |  |
| SDOT\_COLCODE | LOCATION |  |
| INATTENTIONIND | EXCEPTRSNCODE |  |
| UNDERINFL | EXCEPTRSNDESC |  |
| WEATHER | SEVERITYCODE.1 |  |
| ROADCOND | INCDATE |  |
| LIGHTCOND | INCDTTM |  |
| PEDROWNOTGRNT | SDOTCOLNUM |  |
| SPEEDING | SEGLANEKEY |  |
| HITPARKEDCAR | CROSSWALKKEY |  |