How severe could a Car-Crash be in the UK ?

Data

Data Source -
The data come from the **Open Data** website of the UK government, where they have been published by the Department of Transport.

The dataset comprises of two csv files:

- Accident_Information.csv: every line in the file represents a unique traffic accident (identified by the Accident_Index column), featuring various properties related to the accident as columns. Date range: 2005-2017
- Vehicle_Information.csv: every line in the file represents the involvement of a unique vehicle in a unique traffic accident, featuring various vehicle and passenger properties as columns. Date range: 2004-2016

The two above-mentioned files/datasets can be linked through the unique traffic accident identifier (Accident_Index column).

Data PreProcessing -
This dataset covers a wider date range of events. Most of the coded data variables have been transformed to textual strings using relevant lookup tables, enabling more efficient and "human-readable" analysis. It features detailed information about the vehicles involved in the accidents. While the first dataset contains 34 columns/attributes, the second dataset contains 24 of them. But before merging the datasets a number of observations about missing data was made and much of necessary data was formatted.

Based on initial intuition and an educated guess, many columns in the accidents-dataset like ‘InScotland’, which do not seems to contribute to good prediction, were dropped. Then about 0.5% of the records which had missing valued were dropped too. It seems important that time-of-day seemed to play an import role and so the ‘time’ attribute for was divided to 5 time-periods and numerical values were assigned to each. After formatting the ‘date’ attribute , vehicle-dataset, with chosen attributes only, was merged with formatted dataset, with the Accident_Index as the guiding column, to give final dataset, where features to needed to be formatted – categorical to numercal.

Features selected -
The features for model building were selected carefully based on data exploration and analysis:

(From Accidents_Information dataset)Time_of_Day, Road_Surface_Condition, Urban_or_Rural_Area, Light_Condition, Speed_Limit, Weather_Conditions

(From Vehicle_Information dataset) Age_of_Vehicle, Age_Band_of_Driver, Vehicle_Manoeuvre.