Machine Learning Capstone project

# <https://www.kaggle.com/benoit72/uk-accidents-10-years-history-with-many-variables>

# Casualty severity

Serious injury: An injury for which a person is detained in hospital as an “in-patient”, or any of the following injuries whether or not they are detained in hospital: fractures, concussion, internal injuries, crushings, burns (excluding friction burns), severe cuts, severe general shock requiring medical treatment and injuries causing death 30 or more days after the accident. An injured casualty is recorded as seriously or slightly injured by the police on the basis of information available within a short time of the accident. This generally will not reflect the results of a medical examination, but may be influenced according to whether the casualty is hospitalised or not. Hospitalisation procedures will vary regionally.

# Car accidents

* This is not what is safer, we aren’t predicting if an accident will occur. This is if an accident occurs, is it likely to be serious
  + Could say motivation is understanding for emergency services? Prioritisation?
* Could restrict records to certain vehicle types – just cars and motorbikes?
* Could strip out some features before assessing them and justify, i.e. not interested in location, police force etc
* Could make it binary – fatal or not. Is that easier than 3 levels?
* How to deal with multiple casualties (records) for single accident?
  + Could classify an accident as any fatalities?
* How to deal with multiple vehicles in a single accident?
* Data is (strongly?) unbalanced, less than 2% are fatal – how to deal with that
* Deal with missing data

## Accidents

* Accident\_Index
* Date
* Time
* 1st\_Road\_Class
* Road\_Type
* Speed\_limit
* Junction\_Detail
* Light\_Conditions
* Weather\_Conditions
* Road\_Surface\_Conditions

## Vehicles

* Accident\_Index
* Vehicle\_Reference

All these for both vehicles where present:

* Vehicle\_Type
* Vehicle\_Manoeuvre
* Skidding\_and\_Overturning
* Vehicle\_Leaving\_Carriageway
* Hit\_Object\_off\_Carriageway
* 1st\_Point\_of\_Impact
* Sex\_of\_Driver
* Age\_of\_Driver

## Casualties

* Accident\_Index
* Vehicle\_Reference
* Casualty\_Reference
* Casualty\_Class
* Sex\_of\_Casualty
* Age\_of\_Casualty
* Casualty\_Severity (Target factor)
* Car\_Passenger

# High level plan

## Introduction

* What is the purpose
* Data source and pre-processing
  + Details in an appendix?

## EDA

* Distributions for key variables
* Missing values
* Initial observations
* Correlations
* Outliers

## Initial data prep

* Encoding
  + Avoid dummy variable trap, perfect collinearity
  + Check how NAs are handled

## Split into train / validation / test sets

## Data wrangling

* Missing data
* Additional fields
  + Dates, age group, vehicle group….
* Tweak for no second vehicle
* Tweak for pedestrians –> second vehicle = first vehicle, first vehicle = foot

## Factor selection

* Identify best factors and justify

## Model selection

* How many methods to use?
* Select optimal parameters
  + Which method?
  + Cross validation?
* Score against the validation set
* What is the metric of success?
  + Confusion matrix
  + AUC

## Choose best model

* Explain why it is the best model

## Test against the test set

* Final assessment of the model

## Final conclusions, next steps etc

* How could it be improved further etc

## Other factors – things that are important but we have no data

* Speed at impact (could be higher on slower road if not paying attention)
* Using mobile phone
* Drink or drugs
* Car unroadworthy
* Wearing seatbelt
* Missing data
  + Age of casualty and drivers (key factors)

Other next steps

* How could this be extended to accidents involving more than 2 vehicles
* Repeat for other years – same results?

## Models

* (Guessing)
* Random Forest (First so we can select features?)
  + Tune: mtry
* Logistic regression
* CART
  + Tune: cp
* KNN
  + Tune: K
* QDA
* One or two more?
  + Naïve Bayes?
* Ensemble

# To do list

* ~~Ensure random forest final model uses tuned mtry~~
* ~~Investigate the “Check this out” comment~~
* ~~Check out what method(s) to use for the imputation~~
* Why is the CART tree so basic?
  + Does it align with the key features from the random forest?
* Why does QDA give vast majority probability of 0 or 1?
* How to explain what factors are most important
  + What’s the relationship to casualty severity?
* ~~Show all the fields, grouped by category and with brief explanation (check Kaggle?)~~
* Calculate example accuracy to achieve a negative prediction rate of say 50%?
* Change KNN to tune only when we have scaled the data
* Adjust driver age under 16 to 16
* Change the positive class to be 1 in the confusion matrix