Machine Learning Capstone project

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Link | Features | Records |
| Heart Failure Prediction | https://www.kaggle.com/andrewmvd/heart-failure-clinical-data?select=heart\_failure\_clinical\_records\_dataset.csv | 12 | 500 |
| Car insurance cross sell | https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction?select=train.csv | 10 or 11 | 380k |
| UK car accidents | https://www.kaggle.com/benoit72/uk-accidents-10-years-history-with-many-variables?select=Road-Accident-Safety-Data-Guide.xls | 1. ish | >1m |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# 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?

## 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