Logistic regression is a technique often used to model the probability of events with an outcome falling into one of two categories (e.g. yes/no, true/false, red/black) given a data set with a defined set variables. The input data can be continuous, discrete ordinal (for example numbers on a scale) or discrete nominal (categories, for example colours).

Logistic regression has a vast range of applications, for example to predict the probability of : voter choice at elections (candidate a/candidate b), students course completion, consumers taking up a specific marketing offer (responder/non-responder), homeowners defaulting on mortgages, and many many more.. Whilst the simplest form of logistic regression is binary, multinomial and ordinal versions exist, whereby input data can be used to predict for example, the most popular choice of transport in the year 2030, (car, bus, tram, bike) or one of an ordered set of outcomes (for example teeshirt sizes in a certain demographic (S/M/L/XL).

Logistic regression itself dates back to the 19th Century and has been used widely in medicine for many years, for example to predict patient mortality given trauma or injury, the likelihood of disease from patient information. Predictions are made using a model of the relationship between input variables (**x0 … xn**) and the log of the odds (*l*) of the given outcome, and the probability (*p*) of the event occurring, for example for binary logistic regression as shown below, where the model parameters *b*, and b0..n have been calculated from research data by modelling…



Then by reversing the log/exponentiation, the actual odds of the event are calculated.

Text

Description automatically generated

Logistic regression algorithms are used in Supervised ML and work by guessing the initial parameters to calculate a predicted answer and adjusting these iteratively until they provide the best fit to the training data. They thus incorporate the ability to learn and refine the model further as more and more data sets are added over time, possibly even changing the model substantially if the input data, for example, becomes widened in the demographic of its sampling or the event the data reflect change in themselves. The bias of human training data in ML has notably been raised in recent years with the observation that historically data has often been restricted to predominantly western/white male subjects and doesn’t represent fully represent the demographic the ML is being applied to, leading to potential discrimination when ML informs automated decision making or evidence which can radically affect peoples prospects, as discussed by the author Caroline Criado Perez in ‘Invisible Women: Data Bias in a World Designed for Men’