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?

# Caravan Insurance policy

* Binary classification problem
  + Will individual buy a policy or not?
* Data inputs
  + 85 input attributes
  + 1 – 43: Socio-demographic data
    - Derived from zip codes. All customers living in areas with the same zip code have the same socio- demographic attributes
  + 44 – 85: Product ownership
  + Some variables are one hot encoded
* Target variable
  + CARAVAN: Number of mobile home policies

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | # | Name | Description | Labels |
|  | 1 | MOSTYPE | Customer Subtype | L0 |
|  | 2 | MAANTHUI | Number of houses 1 - 10 |  |
|  | 3 | MGEMOMV | Avg size household 1 - 6 |  |
|  | 4 | MGEMLEEF | Avg age | L1 |
|  | 5 | MOSHOOFD | Customer main type | L2 |
| Religion | 6 | MGODRK | Roman catholic | L3 |
| 7 | MGODPR | Protestant ... |  |
| 8 | MGODOV | Other religion |  |
| 9 | MGODGE | No religion |  |
| Marital status | 10 | MRELGE | Married |  |
| 11 | MRELSA | Living together |  |
| 12 | MRELOV | Other relation |  |
| Children | 13 | MFALLEEN | Singles |  |
| 14 | MFGEKIND | Household without children |  |
| 15 | MFWEKIND | Household with children |  |
| Education | 16 | MOPLHOOG | High level education |  |
| 17 | MOPLMIDD | Medium level education |  |
| 18 | MOPLLAAG | Lower level education |  |
| Employment type | 19 | MBERHOOG | High status |  |
| 20 | MBERZELF | Entrepreneur |  |
| 21 | MBERBOER | Farmer |  |
| 22 | MBERMIDD | Middle management |  |
| 23 | MBERARBG | Skilled labourers |  |
| 24 | MBERARBO | Unskilled labourers |  |
| Social class | 25 | MSKA | Social class A |  |
| 26 | MSKB1 | Social class B1 |  |
| 27 | MSKB2 | Social class B2 |  |
| 28 | MSKC | Social class C |  |
| 29 | MSKD | Social class D |  |
| Home ownership | 30 | MHHUUR | Rented house |  |
| 31 | MHKOOP | Home owners |  |
| Car ownership | 32 | MAUT1 | 1 car |  |
| 33 | MAUT2 | 2 cars |  |
| 34 | MAUT0 | No car |  |
| Health insurance | 35 | MZFONDS | National Health Service |  |
| 36 | MZPART | Private health insurance |  |
| Income | 37 | MINKM30 | Income < 30.000 |  |
| 38 | MINK3045 | Income 30-45.000 |  |
| 39 | MINK4575 | Income 45-75.000 |  |
| 40 | MINK7512 | Income 75-122.000 |  |
| 41 | MINK123M | Income >123.000 |  |
| 42 | MINKGEM | Average income |  |
|  | 43 | MKOOPKLA | Purchasing power class |  |

# Labels

|  |  |
| --- | --- |
| # | L0 |
| 1 | High Income, expensive child |
| 2 | Very Important Provincials |
| 3 | High status seniors |
| 4 | Affluent senior apartments |
| 5 | Mixed seniors |
| 6 | Career and childcare |
| 7 | Dinki's (double income no kids) |
| 8 | Middle class families |
| 9 | Modern, complete families |
| 10 | Stable family |
| 11 | Family starters |
| 12 | Affluent young families |
| 13 | Young all American family |
| 14 | Junior cosmopolitan |
| 15 | Senior cosmopolitans |
| 16 | Students in apartments |
| 17 | Fresh masters in the city |
| 18 | Single youth |
| 19 | Suburban youth |
| 20 | Ethnically diverse |
| 21 | Young urban have-nots |
| 22 | Mixed apartment dwellers |
| 23 | Young and rising |
| 24 | Young, low educated |
| 25 | Young seniors in the city |
| 26 | Own home elderly |
| 27 | Seniors in apartments |
| 28 | Residential elderly |
| 29 | Porchless seniors: no front yard |
| 30 | Religious elderly singles |
| 31 | Low income Catholics |
| 32 | Mixed seniors |
| 33 | Lower class large families |
| 34 | Large family, employed child |
| 35 | Village families |
| 36 | Couples with teens 'Married with children' |
| 37 | Mixed small town dwellers |
| 38 | Traditional families |
| 39 | Large religious families |
| 40 | Large family farms |
| 41 | Mixed rurals |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | L1 | L2 | L3 | L4 |
| 0 |  |  | 0% | $0 |
| 1 | 20-30 years | Successful hedonists | 1 - 10% | $1 - 49 |
| 2 | 30-40 years | Driven Growers | 11 - 23% | $50 -99 |
| 3 | 40-50 years | Average Family | 24 - 36% | $100 - 199 |
| 4 | 50-60 years | Career Loners | 37 - 49% | $200 - 499 |
| 5 | 60-70 years | Living well | 50 - 62% | $500 - 999 |
| 6 | 70-80 years | Cruising Seniors | 63 - 75% | $1000 - 4999 |
| 7 |  | Retired and Religious | 76 - 88% | $5000 - 9999 |
| 8 |  | Family with grown ups | 89 - 99% | $10000 - 19999 |
| 9 |  | Conservative families | 100% | $20000 + |
| 10 |  | Farmers |  |  |