Actuaries Institute.

Whole person. Whole system.

IDSS 2023

12 – 14 November
Hobart

Predicting Health Conditions Using Census Data

Daniel Stone

What we'll cover today

- Hypothesis: Can we use census data to predict health conditions?
- Census data source, data & target
- Data wrangling census data
- Data transformations & variable treatment
- EDA
- Modelling: Baseline, Supervised and AutoML
- Future improvements: Other data enrichment, model refinement, alternate target data

Hypothesis: Can we use census data to predict health conditions?

2021 is the first time Census has collected information on diagnosed long-term health conditions.

Potential Uses:

- Public policy
- Life & Health Insurance
- Disability Insurance

28 Has the person been told by a doctor or nurse that they have any of these long-term health conditions?

- Include health conditions that have lasted or are expected to last for six months or more.
- Include health conditions that:
  - may recur from time to time, or
  - are controlled by medication, or
  - are in remission.
- Mark all that apply, like this: ▬

ⓘ Go to **www.census.abs.gov.au/questions** for more information.

- Arthritis
- Asthma
- Cancer (including remission)
- Dementia (including Alzheimer's)
- Diabetes (excluding gestational diabetes)
- Heart disease (including heart attack or angina)
- Kidney disease
- Lung condition (including COPD or emphysema)
- Mental health condition (including depression or anxiety)
- Stroke
- Any other long-term health condition(s)
- **No long-term health condition**

Hypothesis: Can we use census data to predict health conditions?

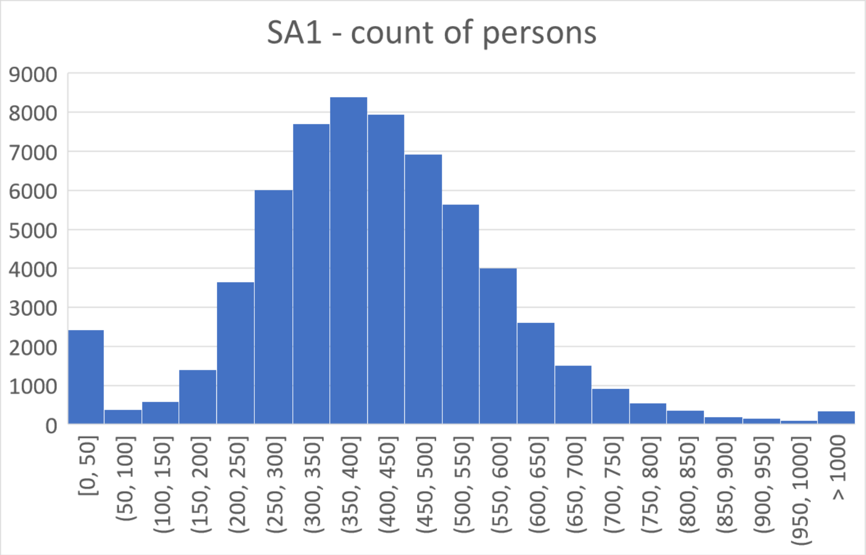
- 62 separate tables of data
- Data format not immediately usable = significant data wrangling needed to create model dataset
- Impact of perturbation of census data to protect anonymity
- Model dataset and explainability / usability of model output?

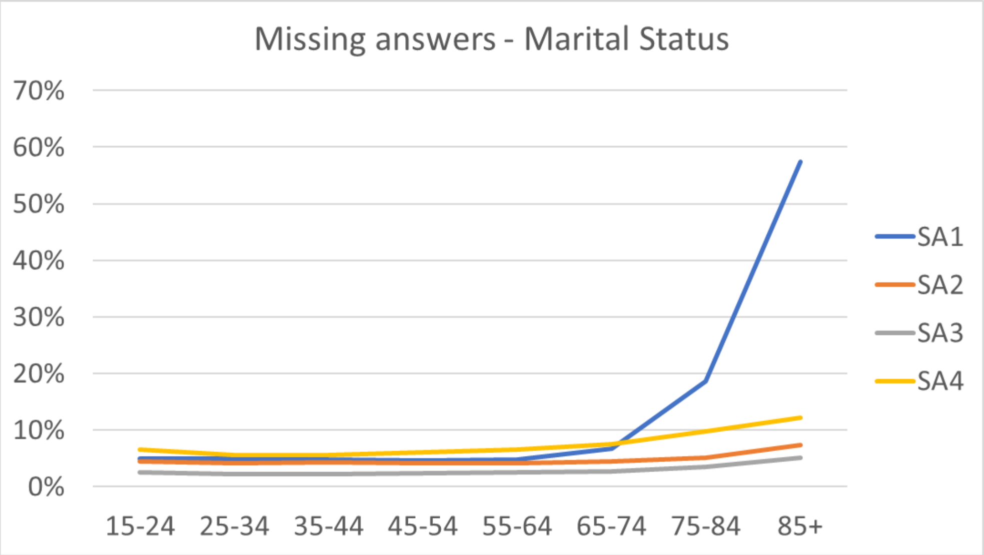
Census data source, data & target

IDSS 2023

Actuaries Institute.

- 2021 Census of Population and Housing - General Community Profile Tables https://www.abs.gov.au/census/find-census-data/community-profiles/2021/AUS
- ~62,000 SA1 locations with Median of 400 people
- SA2, SA3, SA4 less granular, but less sparsity of data





Census data source, data & target

Australian Bureau of Statistics

2021 Census of Population and Housing
General Community Profile Tables

Table Number	Table Name	Table population			status
G01	Selected Person Characteristics by Sex	Persons			Selected
G02	Selected Medians and Averages				Considered
G03	Place of Usual Residence by Place of Enumeration on Census Night by Age	Persons (excludes overseas visitors)			
G04	Age by Sex	Persons			Considered
G05	Registered Marital Status by Age by Sex	Persons aged 15 years and over			Selected
G06	Social Marital Status by Age by Sex	Persons aged 15 years and over			Considered
G07	Indigenous Status by Age by Sex	Persons			
G08	Ancestry by Country of Birth of Parents	Responses and persons			
G09	Country of Birth of Person by Age by Sex	Persons			Selected
G10	Country of Birth of Person by Year of Arrival in Australia	Persons born overseas			
G11	Proficiency in Spoken English by Year of Arrival in Australia by Age	Persons born overseas			
G12	Proficiency in Spoken English of Parents by Age of Dependent Children	Dependent children in couple families			
G13	Language Used at Home by Proficiency in Spoken English by Sex	Persons			
G14	Religious Affiliation by Sex	Persons			
G15	Type of Educational Institution Attending (Full-time/Part-Time Student Status by Age) by Sex	Persons attending an educational institution			
G16	Highest Year of School Completed by Age by Sex	Persons aged 15 years and over who are no longer attending primary or secc			Selected
G17	Total Personal Income (Weekly) by Age by Sex	Persons aged 15 years and over			Selected

Table Number, Name, Population | Cell Descriptors Information

Ready | Accessibility: Investigate

Census data source, data & target

Selected Tables of information, (by SA1, Sex and AgeBand)

- AgeBand by Sex
- Registered Marital Status
- Number of Children Ever Born
- Labour Force Status
- Industry of Employment
- Occupation
- Total Personal Income (Weekly)
- Country of Birth of Person
- Highest Year of School Completed
- Highest Non-School Qualification: Level of Education

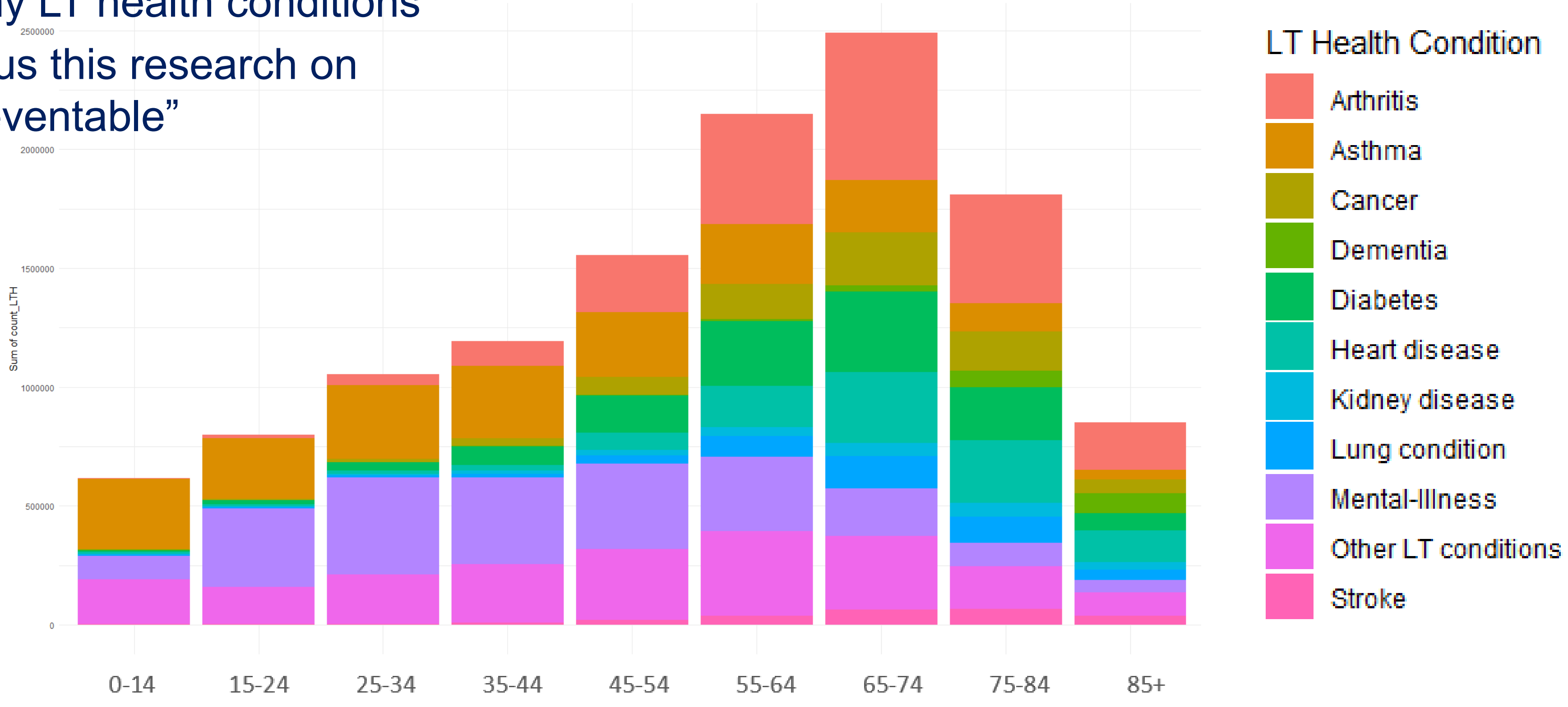
Target: (by SA1, Sex and AgeBand)

- Type of Long-Term Health Condition

Target – LTH condition

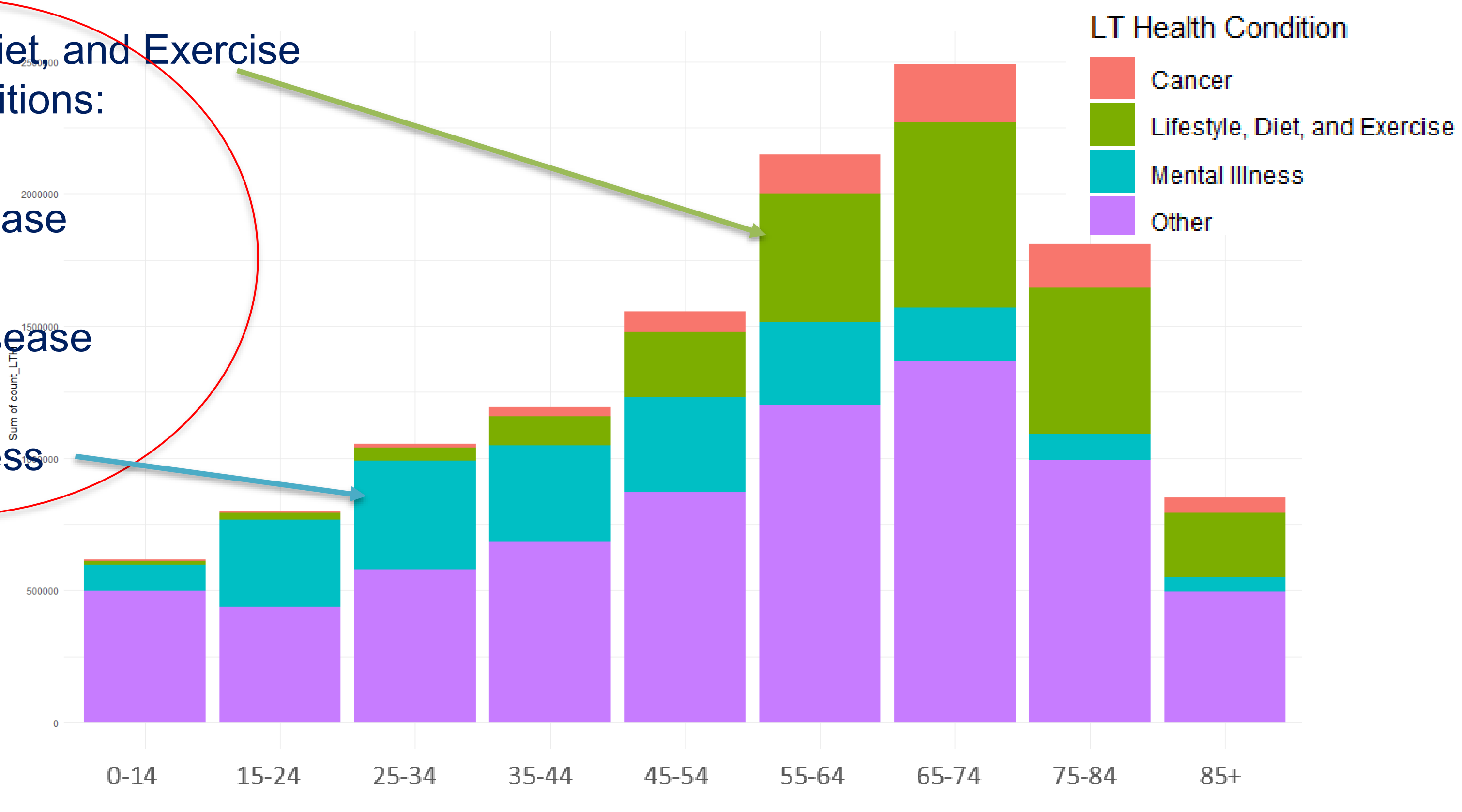
Many LT health conditions

Focus this research on “Preventable”



Target – LTH condition Grouping

1. Lifestyle, Diet, and Exercise Related Conditions:
   - Diabetes
   - Heart Disease
   - Stroke
   - Kidney Disease
2. Mental Illness
3. Cancer
4. Other



Data wrangling census data

IDSS 2023 Actuaries Institute.

WIDE format M_15_19_yr_Divorced

	SA1_CODE_2021	M_15_19_yr_Married	M_15_19_yr_Separated	M_15_19_yr_Divorced	M_15_19_yr_Widowed	M_15_19_yr_Never_married	M_15_19_yr_Tot	M_20_24_yr_Married	M_20_24_yr_Separated	M_20_24_yr_Divorced
1	10102100701	0	0	0	0	16	16	0	0	0
2	10102100702	0	0	0	0	3	3	0	0	0

source

melt

	SA1_CODE_2021	variable	value
1	10102100701	M_15_19_yr_Married	0
2	10102100701	M_15_19_yr_Separated	0
3	10102100701	M_15_19_yr_Divorced	0
4	10102100701	M_15_19_yr_Widowed	0
5	10102100701	M_15_19_yr_Never_married	16
6	10102100701	M_15_19_yr_Tot	16
7	10102100701	M_20_24_yr_Married	0
8	10102100701	M_20_24_yr_Separated	0
9	10102100701	M_20_24_yr_Divorced	0
10	10102100701	M_20_24_yr_Widowed	0
11	10102100701	M_20_24_yr_Never_married	0
12	10102100701	M_20_24_yr_Tot	0
13	10102100701	M_25_34_yr_Married	6
14	10102100701	M_25_34_yr_Separated	0
15	10102100701	M_25_34_yr_Divorced	0
16	10102100701	M_25_34_yr_Widowed	0
17	10102100701	M_25_34_yr_Never_married	11
18	10102100701	M_25_34_yr_Tot	21
19	10102100701	M_35_44_yr_Married	5
20	10102100701	M_35_44_yr_Separated	0

LONG format

Parse

Str, %like

	SA1_CODE_2021	variable	value	gender	age_band	marital_status
1	10102100701	M_15_19_yr_Married	0	Male	15-19	Married
2	10102100701	M_15_19_yr_Separated	0	Male	15-19	Separated
3	10102100701	M_15_19_yr_Divorced	0	Male	15-19	Divorced
4	10102100701	M_15_19_yr_Widowed	0	Male	15-19	Widowed
5	10102100701	M_15_19_yr_Never_married	16	Male	15-19	Never Married
6	10102100701	M_20_24_yr_Married	0	Male	20-24	Married
7	10102100701	M_20_24_yr_Separated	0	Male	20-24	Separated
8	10102100701	M_20_24_yr_Divorced	0	Male	20-24	Divorced
9	10102100701	M_20_24_yr_Widowed	0	Male	20-24	Widowed
10	10102100701	M_20_24_yr_Never_married	0	Male	20-24	Never Married

Male, 15-19, Divorced

Pivot wide

	SA1_CODE_2021	gender	age_band	Divorced	Married	Never Married	Separated	Widowed	age_band2
			15-19	0	0	12	0	0	15-24
			20-24	0	0	3	0	0	15-24
			25-34	0	6	9	0	0	25-34
			35-44	0	7	3	0	0	35-44
			45-54	3	13	4	0	0	45-54
			55-64	6	24	0	0	0	55-64
			65-74	0	10	0	0	4	65-74
			75-84	3	7	0	0	0	75-84
9	10102100701	Female	85+	0	0	0	0	4	85+
10	10102100701	Male	15-19	0	0	16	0	0	15-24
11	10102100701	Male	20-24	0	0	0	0	0	15-24
12	10102100701	Male	25-34	0	6	11	0	0	25-34
13	10102100701	Male	35-44	0	5	12	0	0	35-44
14	10102100701	Male	45-54	3	19	9	0	0	45-54
15	10102100701	Male	55-64	0	27	4	0	0	55-64
16	10102100701	Male	65-74	11	11	3	0	4	65-74
17	10102100701	Male	75-84	0	13	0	0	0	75-84
18	10102100701	Male	85+	0	0	0	0	0	85+

Data wrangling census data

Calculate probabilities
Used in modelling

Create probabilities list
Used in creating sample population

	SA1_CODE_2021	gender	age_band2	Mar_Married_pct	Mar_Divorced_pct	Mar_Widowed_pct	Mar_Never Married_pct	Marital_Status_list_probs
1	10102100701	Female	15-24	0.0000000	0.00000000	0.0000000	1.0000000	c(Married = 0, Separated = 0, Divorced = 0, Widowe [...]
2	10102100701	Female	25-34	0.4000000	0.00000000	0.0000000	0.6000000	c(Married = 0.4, Separated = 0, Divorced = 0, Wido [...]
3	10102100701	Female	35-44	0.7000000	0.00000000	0.0000000	0.3000000	c(Married = 0.7, Separated = 0, Divorced = 0, Wido [...]
4	10102100701	Female	45-54	0.6500000	0.15000000	0.0000000	0.2000000	c(Married = 0.65, Separated = 0, Divorced = 0.15, [...]
5	10102100701	Female	55-64	0.8000000	0.20000000	0.0000000	0.0000000	c(Married = 0.8, Separated = 0, Divorced = 0.2, Wi [...]
6	10102100701	Female	65-74	0.7142857	0.00000000	0.2857143	0.0000000	c(Married = 0.714285714285714, Separated = 0, Divo [...]
7	10102100701	Female	75-84	0.7000000	0.30000000	0.0000000	0.0000000	c(Married = 0.7, Separated = 0, Divorced = 0.3, Wi [...]
8	10102100701	Female	85+	0.0000000	0.00000000	1.0000000	0.0000000	c(Married = 0, Separated = 0, Divorced = 0, Widowe [...]
9	10102100701	Male	15-24	0.0000000	0.00000000	0.0000000	1.0000000	c(Married = 0, Separated = 0, Divorced = 0, Widowe [...]
10	10102100701	Male	25-34	0.3529412	0.00000000	0.0000000	0.6470588	c(Married = 0.352941176470588, Separated = 0, Divo [...]
11	10102100701	Male	35-44	0.2941176	0.00000000	0.0000000	0.7058824	c(Married = 0.294117647058824, Separated = 0, Divo [...]
12	10102100701	Male	45-54	0.6129032	0.09677419	0.0000000	0.2903226	c(Married = 0.612903225806452, Separated = 0, Divo [...]
13	10102100701	Male	55-64	0.8709677	0.00000000	0.0000000	0.1290323	c(Married = 0.870967741935484, Separated = 0, Divo [...]
14	10102100701	Male	65-74	0.3793103	0.37931034	0.1379310	0.1034483	c(Married = 0.379310344827586, Separated = 0, Divo [...]
15	10102100701	Male	75-84	1.0000000	0.00000000	0.0000000	0.0000000	c(Married = 1, Separated = 0, Divorced = 0, Widowe [...]

Data transformations & variable treatment

- Age-band standardization
- Country of birth = Australia vs Overseas
- Avg # children estimate = 1xp(1) + 2xp(2) + … + 6xp(6+)
- Personal income estimate = Weighted estimate (midpoint of band)
- Missing data
  - Age 0-14 = (e.g. Never Married = 100%)
  - Other ages = (fill using SA2, SA3, SA4 probabliities)
    - Most NB for Ages 65+ ⇩

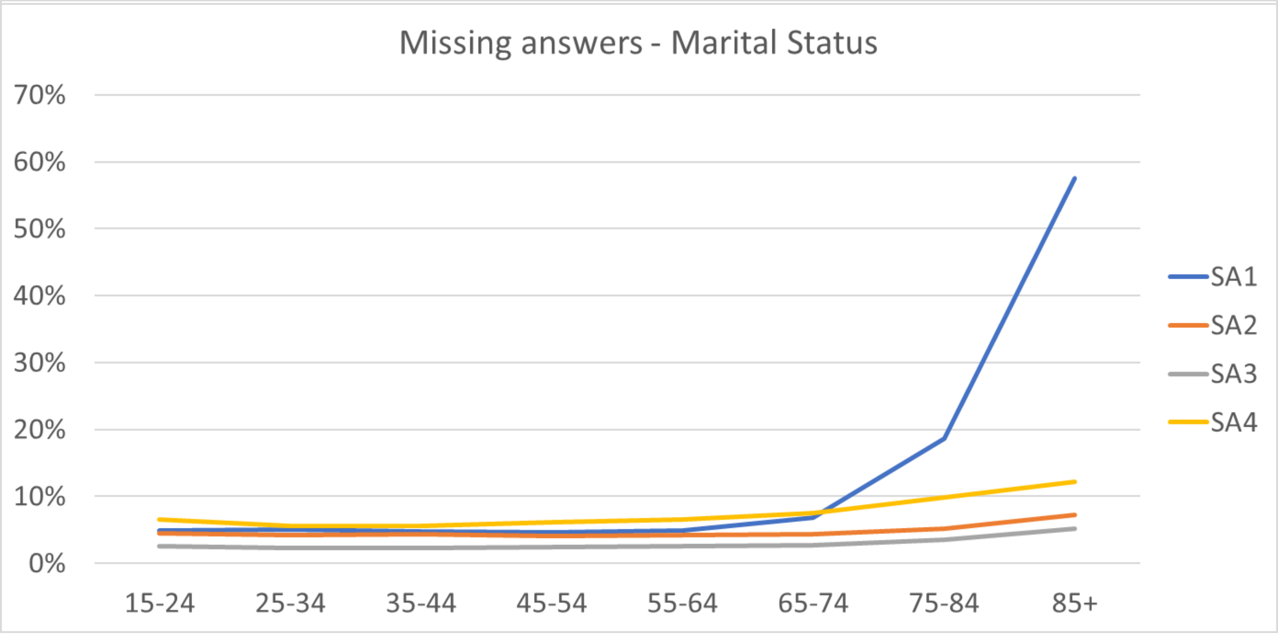
Data transformations & variable treatment

Missing answer means:
- Population count but no data row of variable
- SA1 highest missing due to low count

Approach:
- Merge SA1, SA2, SA3, SA4 tables: by SA code, Age-Band & gender
- Use SA1 probability where available then SA2 then SA3 then SA4

Result:
- Low missing probabilities
- Set As “Missing” and modelled

Note: Same approach for LTH target (fewer missing)



EDA

I did some…

- Mostly around missing data problem – Did SA1 still roll up closely to SA2>SA3>SA4
- A bit around how I might group the target
- I was looking to create some engineered groupings of some of the Occupations, Industries, High-School, …
- But, ultimately decided to let the model figure it out for me and passed all the data through

Modelling: Baseline, Supervised and AutoML

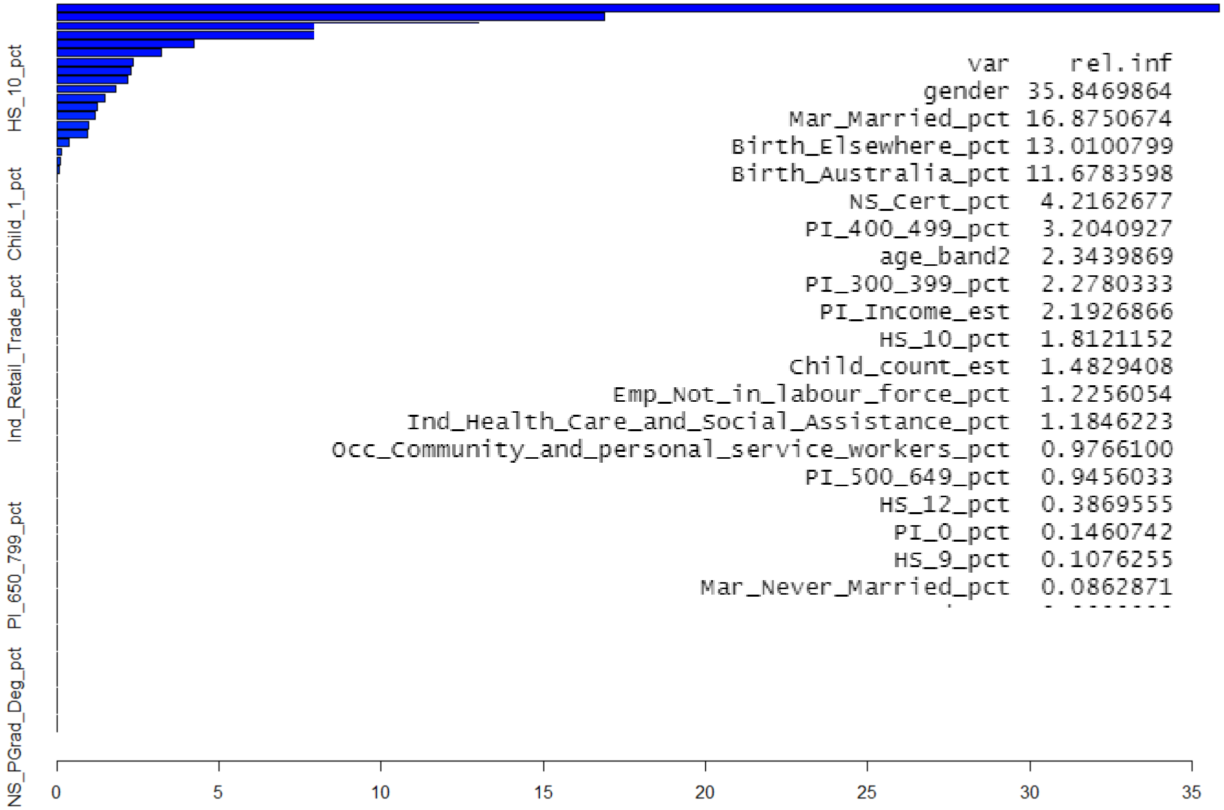
- Predictor inputs
  - 2 categorical factor variables age_band2, gender
  - Approx 85 continuous (0-1) predictor variables
  - 2 engineered features (child count, income estimate)
  - Weights – count_pop
- Baseline
  - GLM – log transformation of target
  - family = gaussian(link = "identity")
- Supervised – GBM
  - distribution = "gaussian“
- AutoML – H2O
  - DNF

Modelling: Results (Mental-Illness)

- GLM poor
  - Train, Validation & Test: $R^2 < 0.10$
  - but coefficients easier to interpret

- GBM: Not great but usable
  - Train, Validation & Test: $R^2 \sim 0.134$
  - Some expected features showing importance
  - Scope for

Variable importance (GBM)



Modelling: So What?

What does this mean?

1. Conclusion: We can model LTH conditions using Census data = Better? Yes
2. A person with a probability of being married, probability of children, probability of Occupation, …., etc has a LTH Mental-Illness probability of 30% (example) = How helpful is this?
3. Identify drivers of LTH condition = Age, Gender, Employment Status, Marital Status, Occupations, etc…= e.g. As probability of being married increases, probability of Mental-Illness decreases
4. Compare the predicted probability of LTH condition of one (or a group of) SA1 vs Census observed LTH condition to identify areas where experience is worse = Public policy?

Is this model actually meaningful? What about a different approach?

Create sample population

- Principle
  - Expand data to mimic individual-level population
  - Assign predictor labels based on probability
  - Target = still probability of "LTH condition category"
  - Model using 10 categorical variables (and any new engineered features)
- Result
  - Model impact of specific individual features: Age, Gender, Occupation, etc rather than based on a probability (which doesn't make sense for individuals)

Create sample population

Probabilities dataframe

	SA1_CODE_2021	gender	age_band2	age_category	count_pop	Marital_Status_list_probs	child_count_list_probs	emp_status_list_probs
3	10102100701	Female	25-34	Adult	12	c(Married = 0.4, Separated = 0, Divorced = 0, Wido [...]	c(`0` = 0.333333333333333, `1` = 0, `2` = 0.666666 [...]	c(Employed = 0.8, Unemployed = 0, `Not in labour f [...]
4	10102100701	Female	35-44	Adult	15	c(Married = 0.7, Separated = 0, Divorced = 0, Wido [...]	c(`0` = 0, `1` = 0, `2` = 1, `3` = 0, `4` = 0, `5` [...]	c(Employed = 0.5, Unemployed = 0, `Not in labour f [...]
5	10102100701	Female	45-54	Adult	16	c(Married = 0.65, Separated = 0, Divorced = 0.15, [...]	c(`0` = 0.428571428571429, `1` = 0.571428571428571 [...]	c(Employed = 0.666666666666667, Unemployed = 0, `N [...]
6	10102100701	Female	55-64	Adult	28	c(Married = 0.8, Separated = 0, Divorced = 0.2, Wi [...]	c(`0` = 0, `1` = 0.181818181818182, `2` = 0.242424 [...]	c(Employed = 0.709677419354839, Unemployed = 0, `N [...]

Expand to Aus population

NS_qual_list_probs	Marital_Status	Child_count	emp_status_count
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Married	1	Not in labour force
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Divorced	1	Employed
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Married	1	Employed
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Married	1	Employed
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Married	1	Employed
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Divorced	1	Employed
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Never Married	1	Not in labour force
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Married	0	Employed
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Married	1	Employed
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Never Married	0	Employed
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Divorced	1	Employed
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Married	1	Not in labour force
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Never Married	0	Employed
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Married	0	Not in labour force
c(`AdvDip and Dip` = 0.571428571428571, `Bach Deg` [...]	Married	1	Not in labour force

Assign labels using sample

```
sample_runif_combined <- function(combined_field) {
        # extract the input list and probabilities from the combined field
        input_list <- names(combined_field)
        probs <- unname(combined_field)

        # call the sample function using the extracted input list and probabilities
        sample(input_list, size = 1, prob = probs)
}

# call the sample_runif_combined function for each row of dt_tmp3
dt_tmp3[, Marital_Status := sapply(Marital_Status_list_probs,
sample_runif_combined)]
```

Future improvements

- Alternate data approach
  - Create sample population

- Model refinement
  - More variables
  - More sophisticated models (neural network) and fine tuning
  - Model at SA2 (or higher) rather than SA1?

- Other data enrichment
  - SEIFA scores & deciles (e.g. IRSAD)
  - Other health data

- Alternate data source
  - Access to enhanced ABS Census data

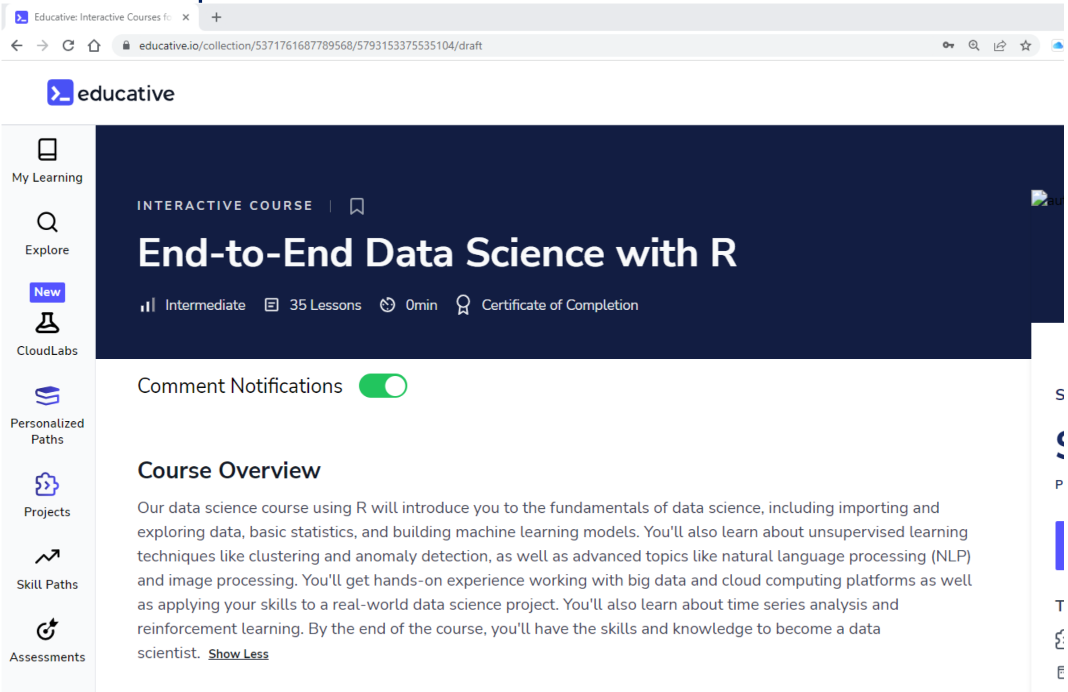
Course – End-to-End Data Science with R

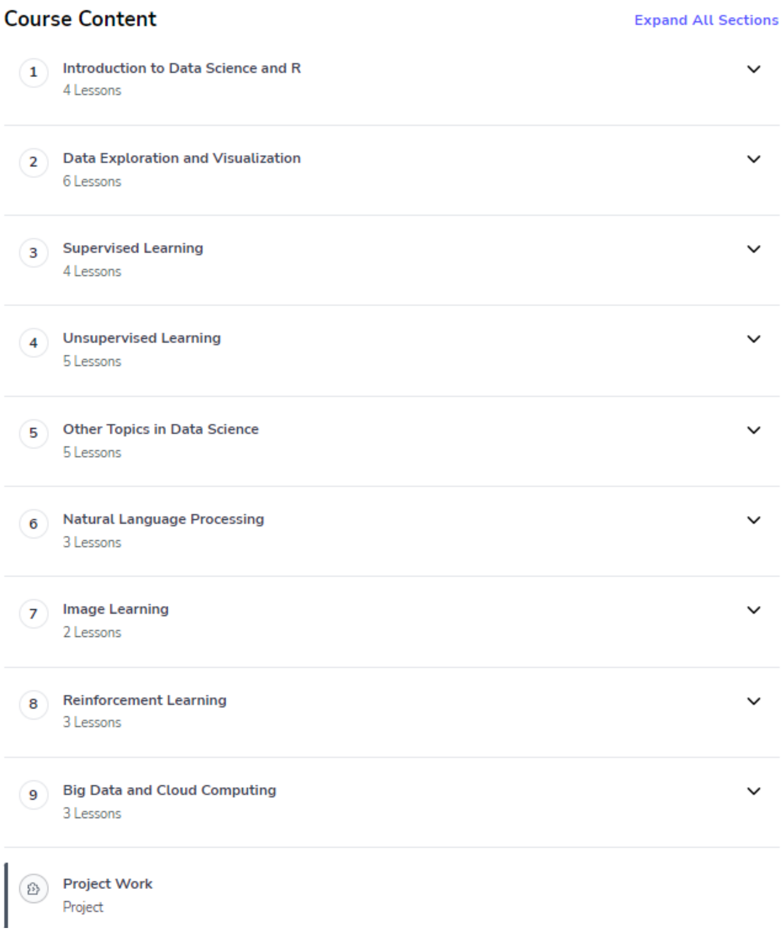
Course co-author with Rene Essomba

www.educative.io

Learning, labs and project

* Will post a link once course available.





Actuaries Institute.

Thank you

IDSS 2023

12 – 14 November
Hobart