# Door-to-Door Charity Collection

Richard Warwick - 30th May 2016

# What is the problem

- The charity marketplace is extremely competitive in AU
- Accentuated by diminishing government and corporate funding
- Large-scale fundraising campaigns essential to maintain viability
- Although largely executed by volunteers, there are significant costs involved in the administration and promotion of door-to-door fundraising
- To efficiently target resources (including marketing) and maximise ROI, need to:
  - Understand profile of collectors
    - Are there any defining characteristics or distinct groups?
  - Identify key geographic areas for targeted messaging
  - Understand key factors that predict whether a collector is ‘profitable’

# What data do I have?

- 2015 collection data from *anonymous* national charity
  - Postcode, age, number of streets covered, new or existing volunteer, total collected, total donated (by the individual collector), total received (collected + donated), profitability (binary whether or not return exceeds average cost)
  - Private
- 2013/14 ATO Postcode Data
  - No. of individuals, salary, income, tax deductible donations
  - Public
- 2011 ABS Census Community Profile - Postcode level
  - Comprehensive demographic information to build out postcode level data
  - Public

# What data do I have?

Collectors profile skews towards older people, women, and new collectors

Inherent limitations/bias - bleeding between postcodes, houses per street etc.

| | postcode | num_streets | collection_amount | donation_amount | total_received | age | gender | weekends | profitable | Acq |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 26523 | 26523.000000 | 26523.000000 | 26523.000000 | 26523.000000 | 26523.000000 | 26523.000000 | 26523.000000 | 26523.00000( | 26523.000000 |
| unique | 1728 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| top | 4350 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| freq | 230 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| mean | NaN | 2.883610 | 77.820176 | 3.724618 | 81.544794 | 62.458621 | 0.653998 | 3.352298 | 0.413377 | 0.614372 |
| std | NaN | 2.548894 | 70.843846 | 18.707591 | 70.658179 | 14.072699 | 0.475703 | 0.871464 | 0.492449 | 0.486752 |
| min | NaN | 0.000000 | -414.950000 | -150.000000 | -500.000000 | 0.000000 | 0.000000 | 2.000000 | 0.000000 | 0.000000 |
| 25% | NaN | 1.000000 | 37.000000 | 0.000000 | 40.000000 | 54.000000 | 0.000000 | 2.000000 | 0.000000 | 0.000000 |
| 50% | NaN | 2.000000 | 61.000000 | 0.000000 | 64.000000 | 65.000000 | 1.000000 | 4.000000 | 0.000000 | 1.000000 |
| 75% | NaN | 3.000000 | 100.000000 | 0.000000 | 101.000000 | 72.000000 | 1.000000 | 4.000000 | 1.000000 | 1.000000 |
| max | NaN | 40.000000 | 2123.000000 | 1000.000000 | 2123.000000 | 103.000000 | 1.000000 | 4.000000 | 1.000000 | 1.000000 |

# What did I do?

- Collated and prepared the data for testing/analysis
    - Needed to create a business relevant metric not present in the data - ‘profitability’
- Tried the following models:
    - **K-means clustering**
    - PCA
    - Linear Regression
    - Logistic Regression
    - Decision Trees
    - **Random Forest**
- Transformed clustering and predictions into features for further analysis

# What Were the Results?

- Clustered collectors into 5 distinct groups
  - Gave them descriptive labels
  - Mapped them
  - Ready to send to outbound call centre/media agency
- Model that predicts profitability at an individual level
  - 5 key features:
    - Age, Number of streets, Postcode, New or returning collector, Gender
  - .97 accuracy on training data (tick), generalises at .57 (cross)
- Model that predicts profitability at a postcode level
  - Evenly distributed feature importance
    - Income and correlated features dominant
  - 1.0 accuracy training, generalises at .68 (better)

# Collector Clusters

#0 = heart and soul - above average incomes, okay returns

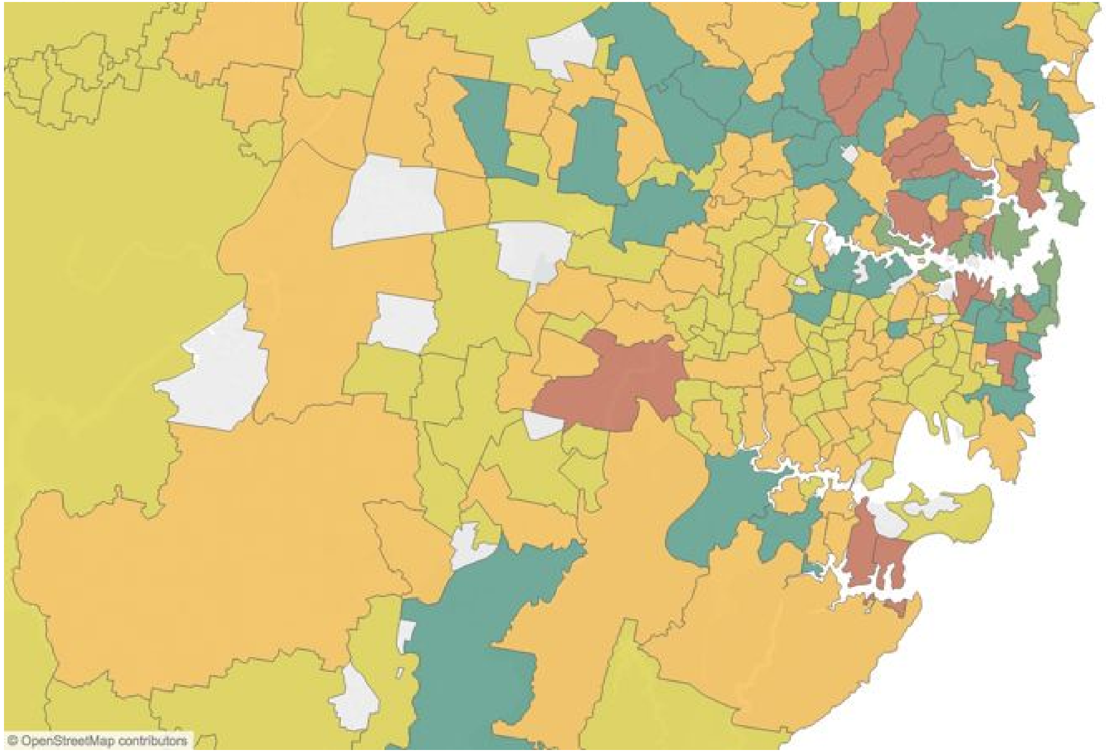
#1 = old money - high avg incomes, avg charity donations $5.5k each year, older & newer collecters, fewer streets

#2 = scraping by - collectors in low affluence neighbourhoods often having to chip in themselves to be profitable

#3 = blue collar collectors - average volunteers, average returns

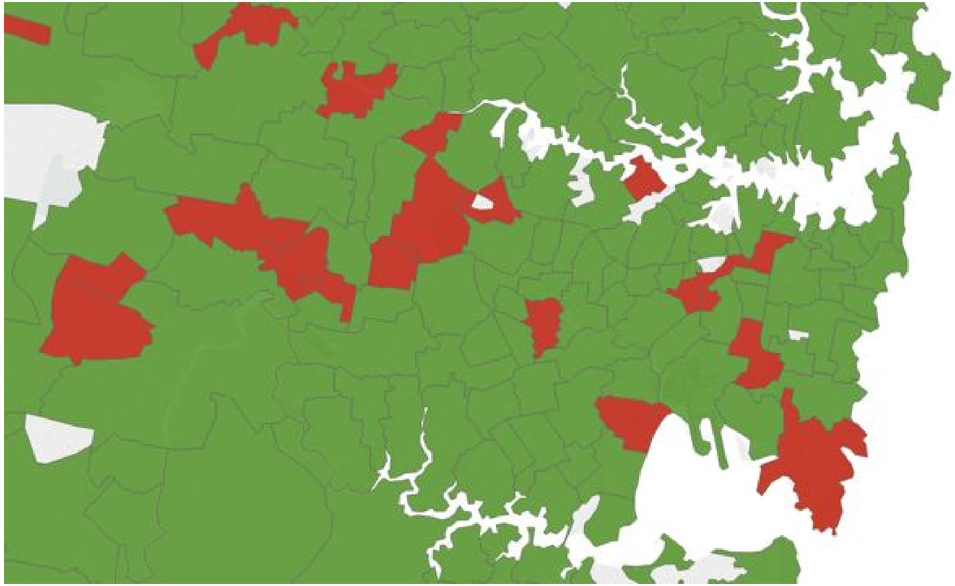
#4 = underperformers - well populated, high incomes, modest returns

# Collector Clusters - Maps (Tableau)



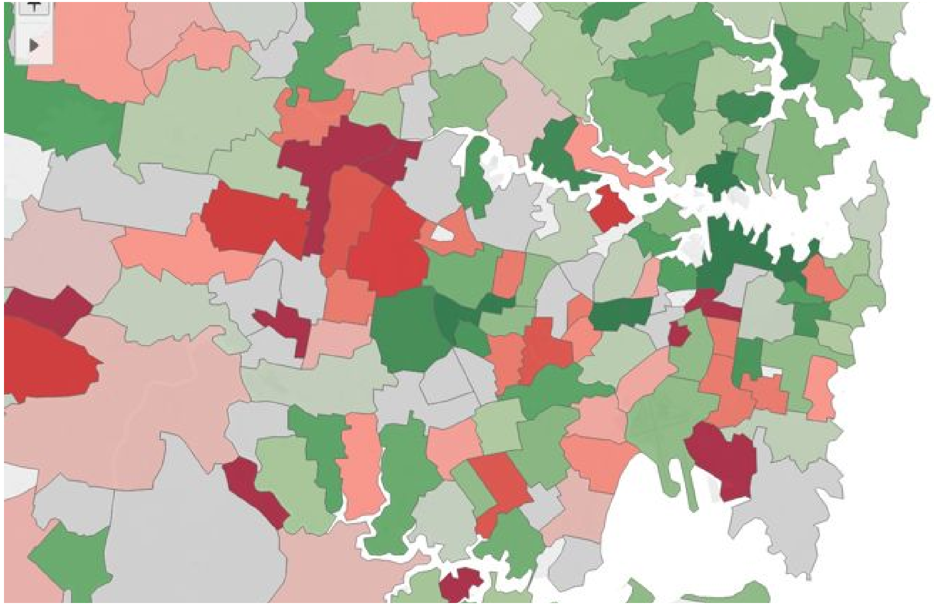
# Profitability Prediction - Postcode

- Averages profitability across all collectors in an area
- Useful for identifying problem/interesting areas
- Highly correlative features
  - Demographic data
  - Duplication between ATO and ABS data
- Even spread of feature importance within model
  - .08 highest value
- OOB score of 0.68



# Profitability Prediction - Individuals

- Same analysis at an individual level
- Feature importance helps identify key levers
- Individual features more important
    - Age, Gender, Num Streets, etc.
- OOB score of 0.57
- Potential for improvement



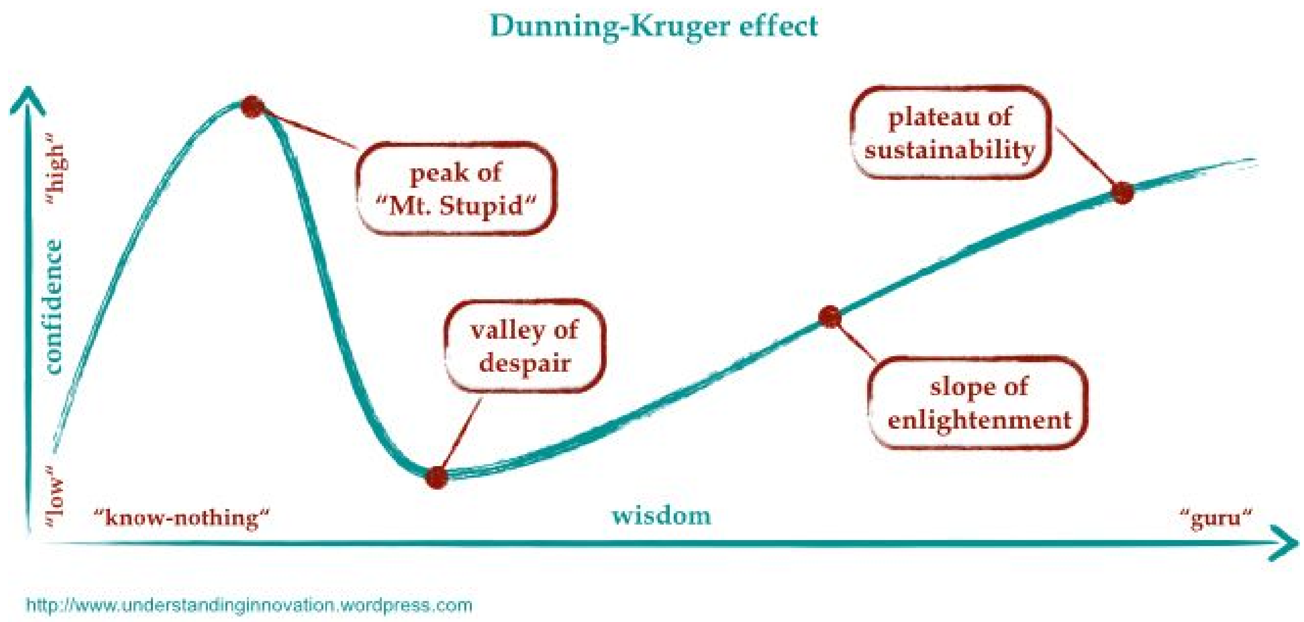
# Did I Achieve What I Set Out To Do?

- In a way...
  - Near minimum viable product
  - Clustering interesting and potentially useful
  - Successfully implemented some DS techniques
  - Established a potential use case
- Would have liked to produce a more accurate model
- Overall confident have added value to the dataset through unsupervised and supervised learning techniques

# What Did I Learn?

- Not to lose the wood for the (decision) trees
    - Desire to build a model that predicts nicely moved me away from actionable business insights
- Better understanding of characteristics of data and at what level DS techniques can be employed
- How to wrangle Python effectively (VCR last thing I've ever programmed)
- Balancing preparation, experimentation, and time (resources)

# I Took an All Too Familiar Path



# What I Will Do Next

- Go back to the data
- Refine the feature set further
- Look to improve prediction model
  - Try xgboost
- Try and enrich individual-level data (I've stripped all the PII out)
- Present results internally (workplace)
- *Maybe* implement some recommendations off the back
- Potentially create searchable map (dashboard)