# Analytics in Banking

Will customers subscribe to a term deposit?

What should the bank do to increase subscription rates?

- Customer demographics
- Effectiveness of marketing campaigns
- Practical considerations for future marketing campaigns

Group Members: Michael Chan, Dennis Rigon, Ze Yu Zhong, Max Han, Xavier Villegas

# Exploration of Dataset

## 7 Demographic Variables:

- Age
- Marital Status
- Default Status
- Personal Loan
- Type of Job
- Education Level
- Housing Loan

## 7 Marketing Campaign Variables:

- Communication type
- Month
- Day of week
- no. of contacts made this campaign
- days since last contact
- no. of contacts made prior to campaign
- outcome of previous campaign

## Notable Issues:

- Large number of categorical independent variables (10/14)
- Redundant/collinear dummy variables
- Large number of “unknown” reponses
- Class imbalance (>90% did not subscribe)

# Feature Engineering

Dealing with a “meaningless” numerical value:

- pdays changed into a binary variable (0 = not previously contacted, 1 = previously contacted)

Dealing with unknowns:

- “unknowns” and “non-existents” kept as their own levels

Dealing with collinearity:

- Collinear dummy variables merged
  - previous = 0, poutcome = nonexistent
  - house = unknown, loan = unknown
- Month and day_of_week removed

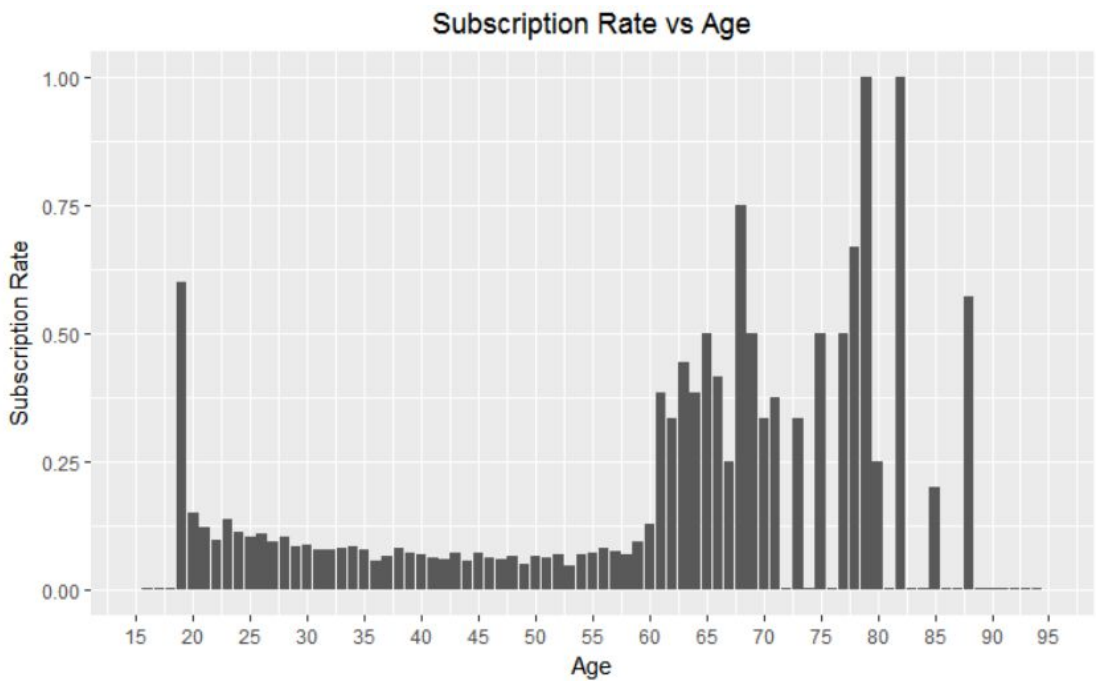
Increasing parsimony:

- default = yes merged with default = unknown

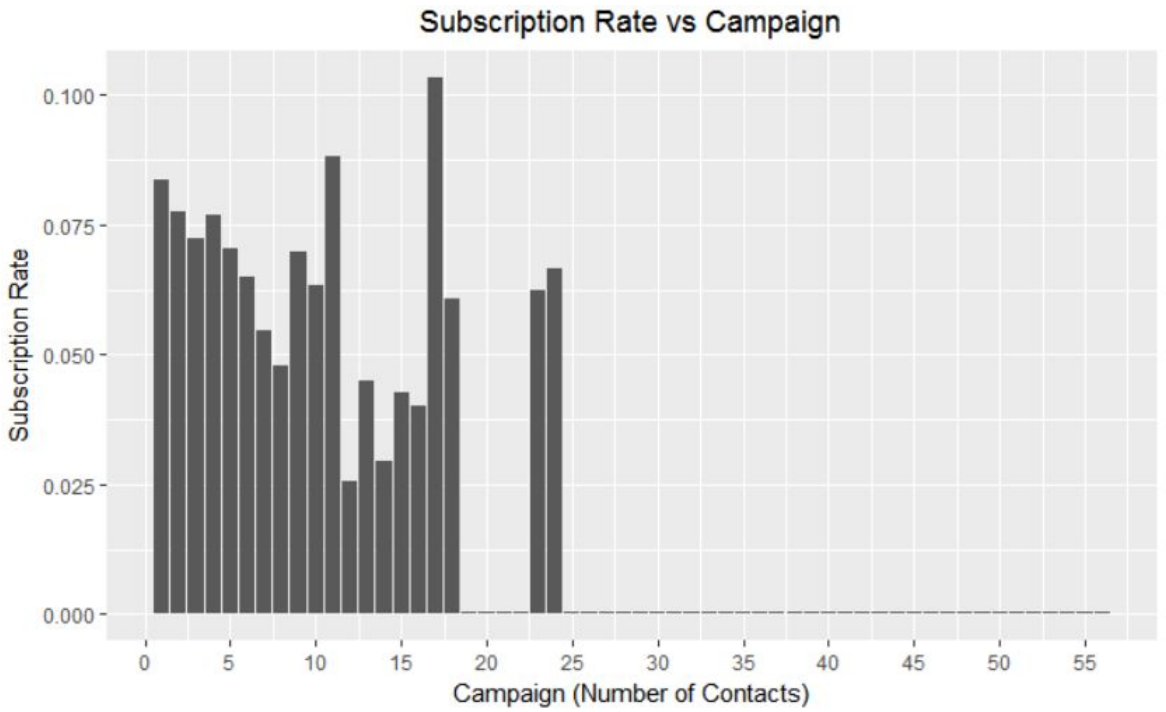
# Feature Engineering

## Transformed numerical variables to categorical

Subscription Rate vs Age




Subscription Rate vs Campaign Length



# Interaction Effects

Example of a dummy variable created:

- 1 if edu = 4 or 6 and age <= 30, 0 otherwise

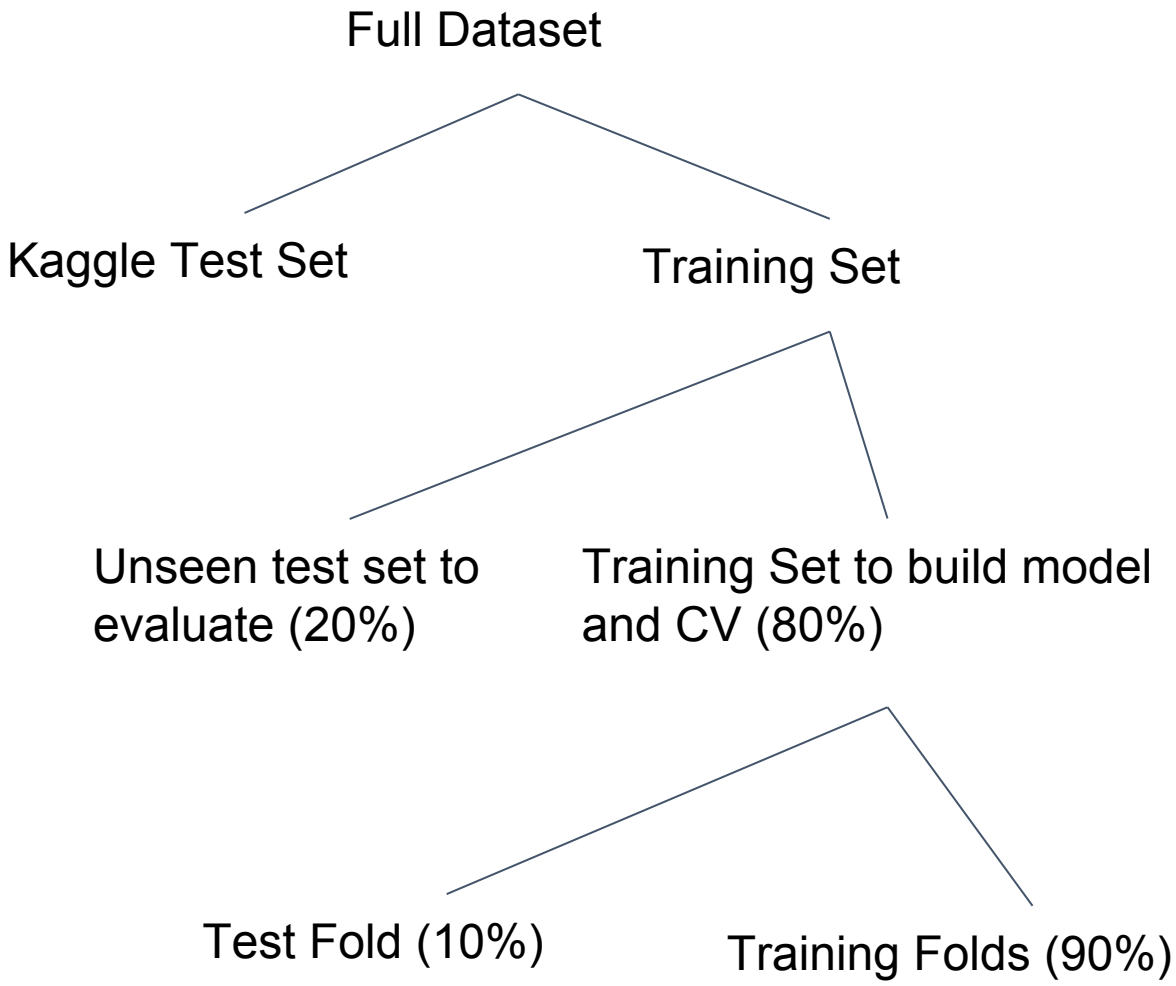
| | 30 or less <dbl> | 31 to 40 <dbl> | 41 to 50 <dbl> | 51 to 60 <dbl> | Overall for <60 <dbl> |
|---|---|---|---|---|---|
| Overall | 0.09923502 | 0.07606860 | 0.06330613 | 0.07059801 | 0.07544803 |
| 4 | 0.06010929 | 0.04285714 | 0.05152672 | 0.07662464 | 0.05959210 |
| 6 | 0.06285714 | 0.07910448 | 0.06525573 | 0.04477612 | 0.06769937 |
| 9 | 0.08816121 | 0.06952204 | 0.05917160 | 0.06687403 | 0.06897312 |
| high.school | 0.09109875 | 0.07134725 | 0.06206460 | 0.06483791 | 0.07332065 |
| professional.course | 0.09248555 | 0.07931666 | 0.05870237 | 0.06631763 | 0.07475459 |
| university.degree | 0.12173315 | 0.08577461 | 0.07111597 | 0.08353414 | 0.08804175 |
| unknown | 0.13286713 | 0.07246377 | 0.08373206 | 0.04444444 | 0.07383628 |

Similarly investigated were:

- 1 if marital = married and age <= 30, 0 otherwise
- 1 if edu = 6 and age = 51-60, 0 otherwise
- 1 if job = retired and age > 60, 0 otherwise

# Model Building and Selection

- Logistic Regression
- Stepwise Logistic Regression
- LASSO Logistic Regression
- Classification Trees
- Random Forests
- Boosted Trees (XGBoost)



# Results

| | CV log loss | Test log loss |
|---|---|---|
| **Logistical GLM based methods** | | |
| Logistical GLM on all transformed variables | 0.2631134 | 0.26382 |
| Stepwise Logistical GLM including all steps | N/A, took too long to compute | 0.2637668 |
| Stepwise Logistical GLM, first 20 steps | 0.2636601 | 0.2630271 |
| LASSO Logistical GLM | N/A, algorithm uses AUC | 0.2656045 |
| **Tree-based Methods** | | |
| Random Forest | N/A | 1.151856 |
| XGBoost | 0.244734 | 0.2500551 |

# Conclusions

The bank should focus on:

- Younger demographic, especially students
- Older demographic
- Customers who have been contacted in a previous campaign

The bank should avoid:

- Middle-aged demographic, especially those with low education levels
- Contacting clients too many times
- Contacting clients via landline, instead of mobile
- Spending too much time contacting clients who do not disclose credit status