INTRODUCTION

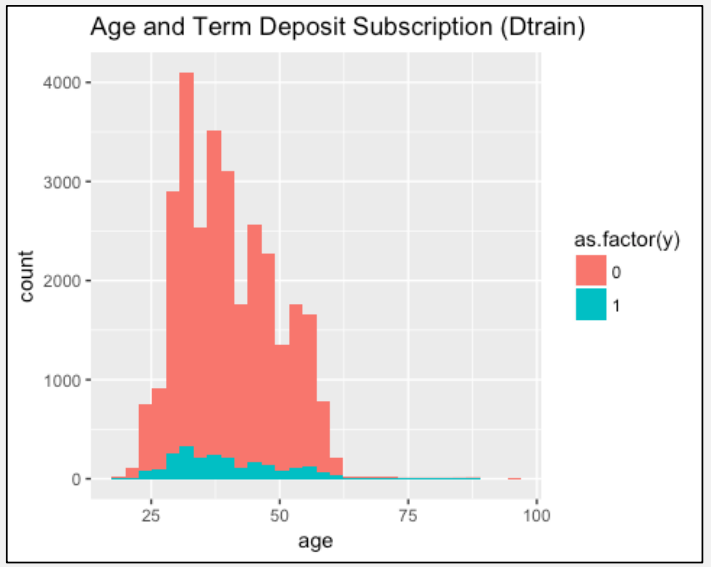
- Goal: predict the probability that a client would subscribe to a term deposit based on a variety of predictors
- Data:
  - Ten categorical and five integer variables
  - collected during a direct marketing campaign
  - 40000+ observations.

METHODOLOGY

- Subsetting: Data was divided into a training set and a validation set.
- Log-loss function used with the validation set to evaluate models.
- Approaches attempted:
- XGB
- LDA
- Random forest
- QDA
- Clustering
- Bruteforce to minimise log-loss
- **Linear model (logit LM chosen for final model)**

METHODOLOGY (CONT.)

- Data cleaning
- 'Age' and 'log(age)' were included to account for both non-linear and linear trends in the age data.
- Trial and Error
- Treatment of categorical variables:



Categorical variable → Dummy variables → Significance of variables test → Exclude insignificant dummies

Group:The Outliers

RESULTS

- $$\log\left(\frac{\hat{y}}{1-\hat{y}}\right) = \underset{(1.55)}{7.72} + \underset{(0.014)}{0.083age} - \underset{(0.576)}{3.45\log(age)} + \underset{(0.057)}{0.14default_{no}} - \underset{(0.098)}{0.25contact\_telephone} - \underset{(0.120)}{1.28month\_may} - \underset{(0.125)}{1.04month\_jun} - \underset{(0.070)}{0.99month\_jul} - \underset{(0.078)}{1.20month\_aug} + \underset{(0.275)}{2.24month\_oct} - \underset{(0.084)}{1.09month\_nov} + \underset{(0.134)}{1.137month\_mar} - \underset{(0.064)}{0.18day\_of\_week\_mon} + \underset{(0.061)}{0.17day\_of\_week\_wed} + \underset{(0.058)}{0.14day\_of\_week\_thu}$$

- All variables statistically significant
- AIC:15563

Group:The Outliers

RESULTS

EVALUATION METRIC	SCORE
Log-Loss$_{(Validation)}$	0.255
False Negative Rate$_{(Valid.)}$	0.48
False Positive Rate$_{(Valid.)}$	0.077
Log-Loss$_{(Public\ Kaggle)}$	0.556
Log-Loss$_{(Private\ Kaggle)}$	0.580
Null Deviance	16511
Deviance	15533

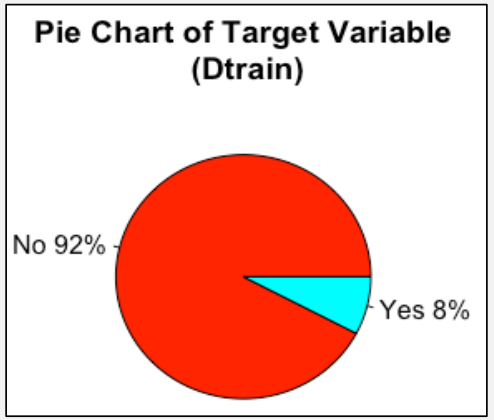
Group:The Outliers

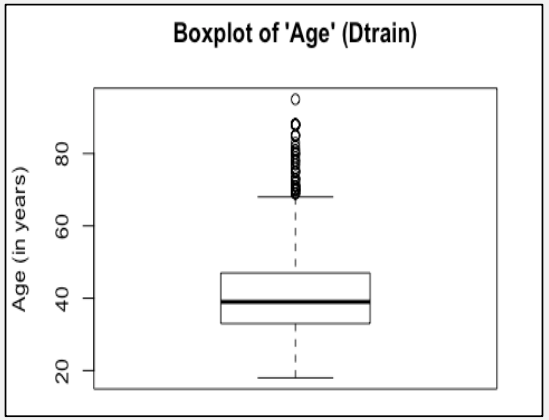
RESULTS - DISCUSSION

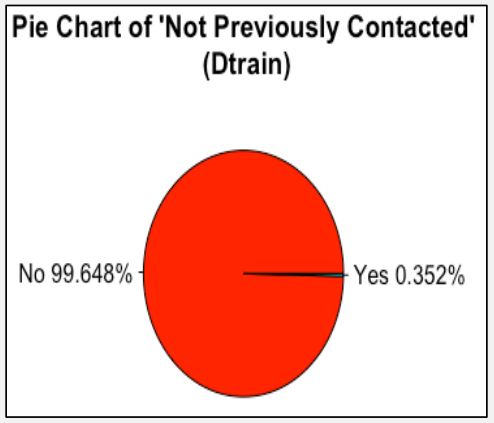
- Deviance < Null Deviance points towards model explaining the relationship within data
- False Negative Rate significantly greater than False Positive Rate
- Noting the months that appear to result in an increase or decrease in the probability of a client subscribing to a term deposit, is there seasonality?
- Clients contacted via cellular appear to have a higher probability of subscribing then those via telephone keeping all else constant
- Clients contacted in the middle of the week appear to have a higher probability of subscribing than those contacted earlier and/or later

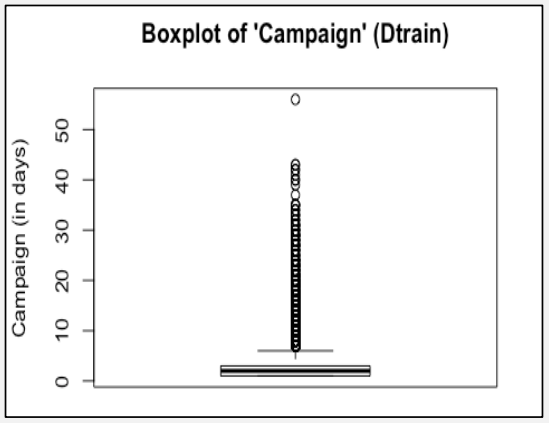
RESULTS – IN RETROSPECT...

- Possibility of overfitting?
- Class imbalance in variable 'y'
  - Solutions?
- Significant number of outliers in 'age' and 'campaign'
  - Solutions?
- In variable 'pdays', a value of 999 corresponded to someone not previously contacted.
  - This comprised the overwhelming majority of observations.
  - Solutions?









Group:The Outliers

CONCLUSION

- A logit model comprising of significant predictors was produced
- This model could be used by businesses in order to identify the certain individuals that should be targeted through direct marketing
- Furthermore, this model could be used to identify certain strategy that could be implemented when direct marketing occurs
- Through analysis of the plots of various predictors, a variety of alternative methodology that perhaps should have been used were recognised for future models to consider