Predicting product subscription for Direct Marketing Campaigns

Annu Sharma

Problem Description

- Bank Marketing Data Set: UCI Repository
- Data contains details related to marketing campaign customer calls
- Data has – client information and previous campaign related data
- Goal – To predict if the customer will subscribe to a term deposit!

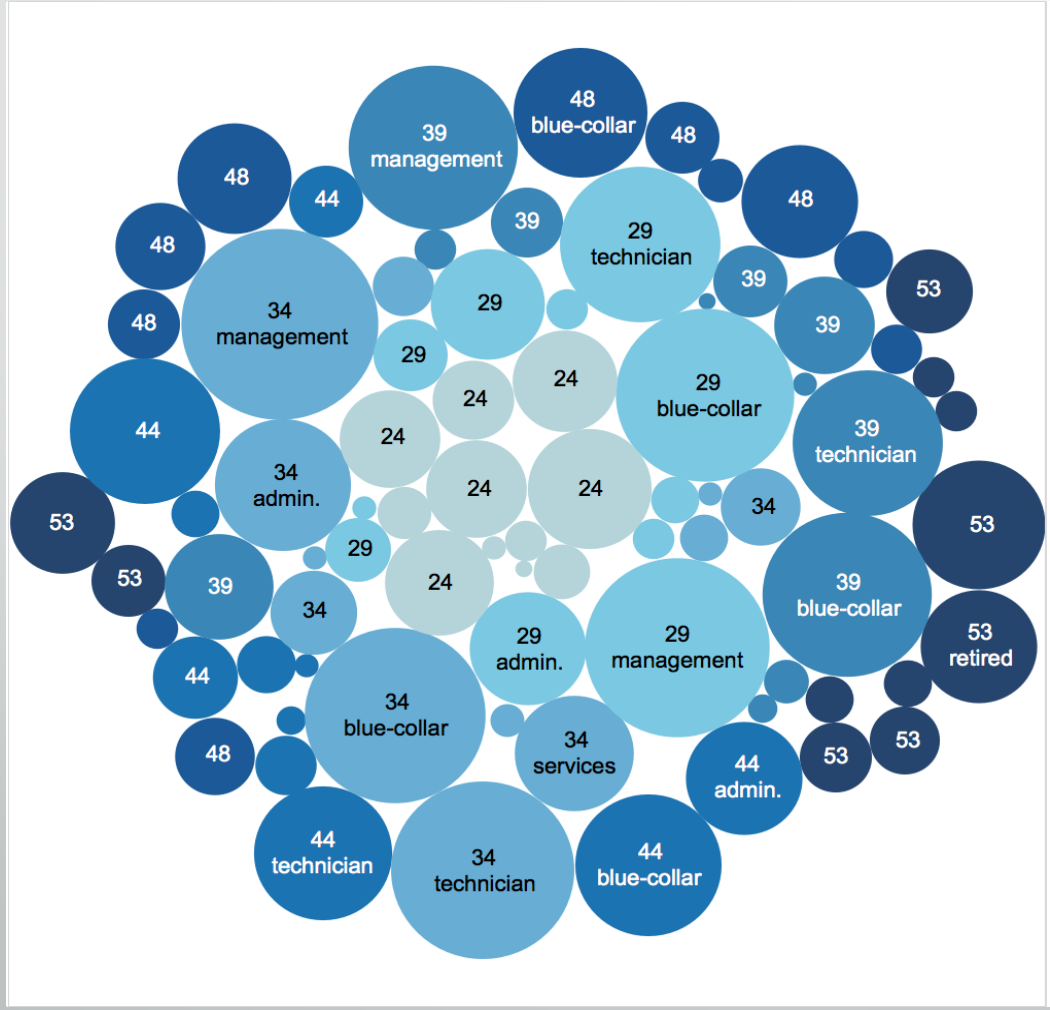
Exploratory Insights

Age (bin)	
15	2
19	43
24	254
29	465
34	554
39	407
44	347
48	275
53	254
58	83
63	22
68	11
73	15
77	8
82	3

Most targeted age groups are the bins from ages **24-39** and gradual drop till 53

* We will focus on these age groups to get more detailed insights

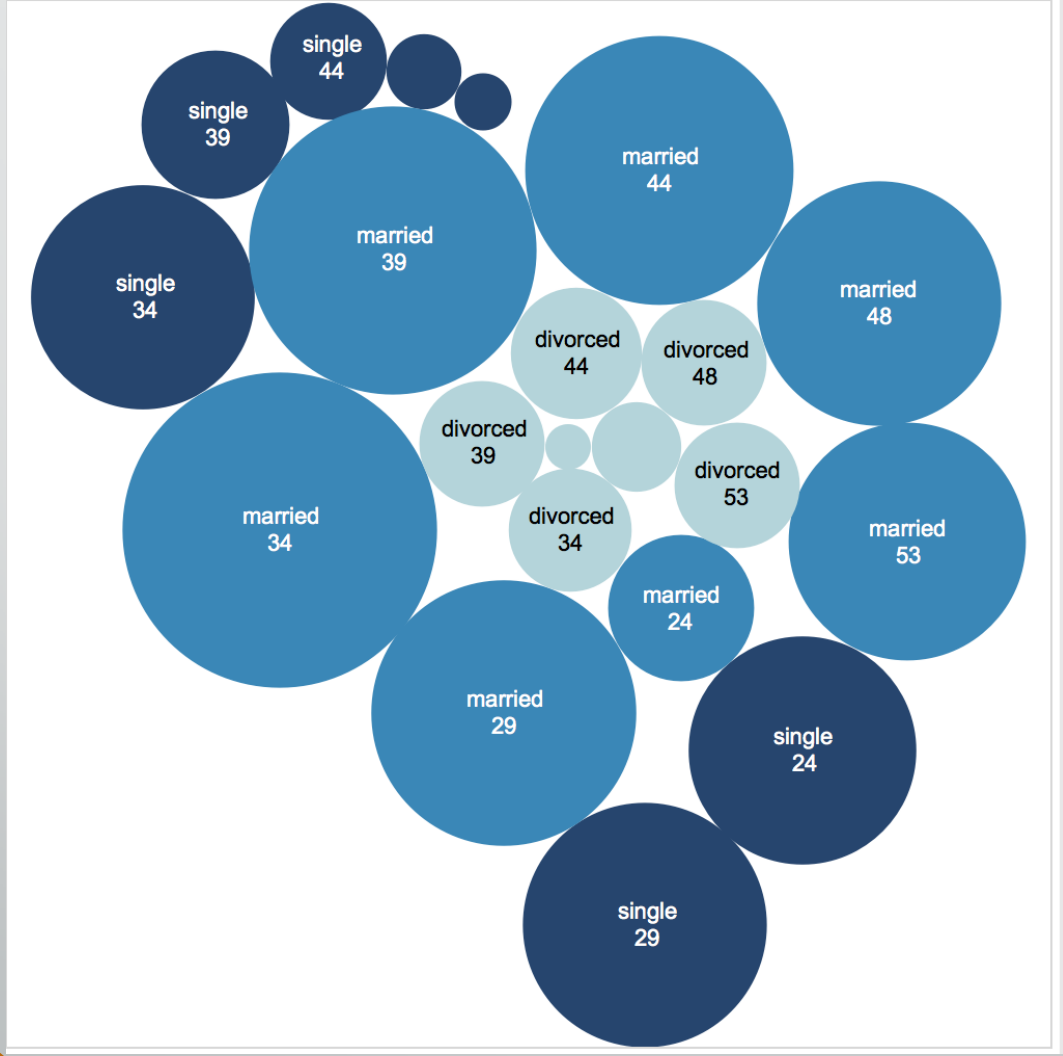
Exploratory Analysis – Jobs targeted



Most targeted customers in –

- Management
- Blue Collar
- Technician

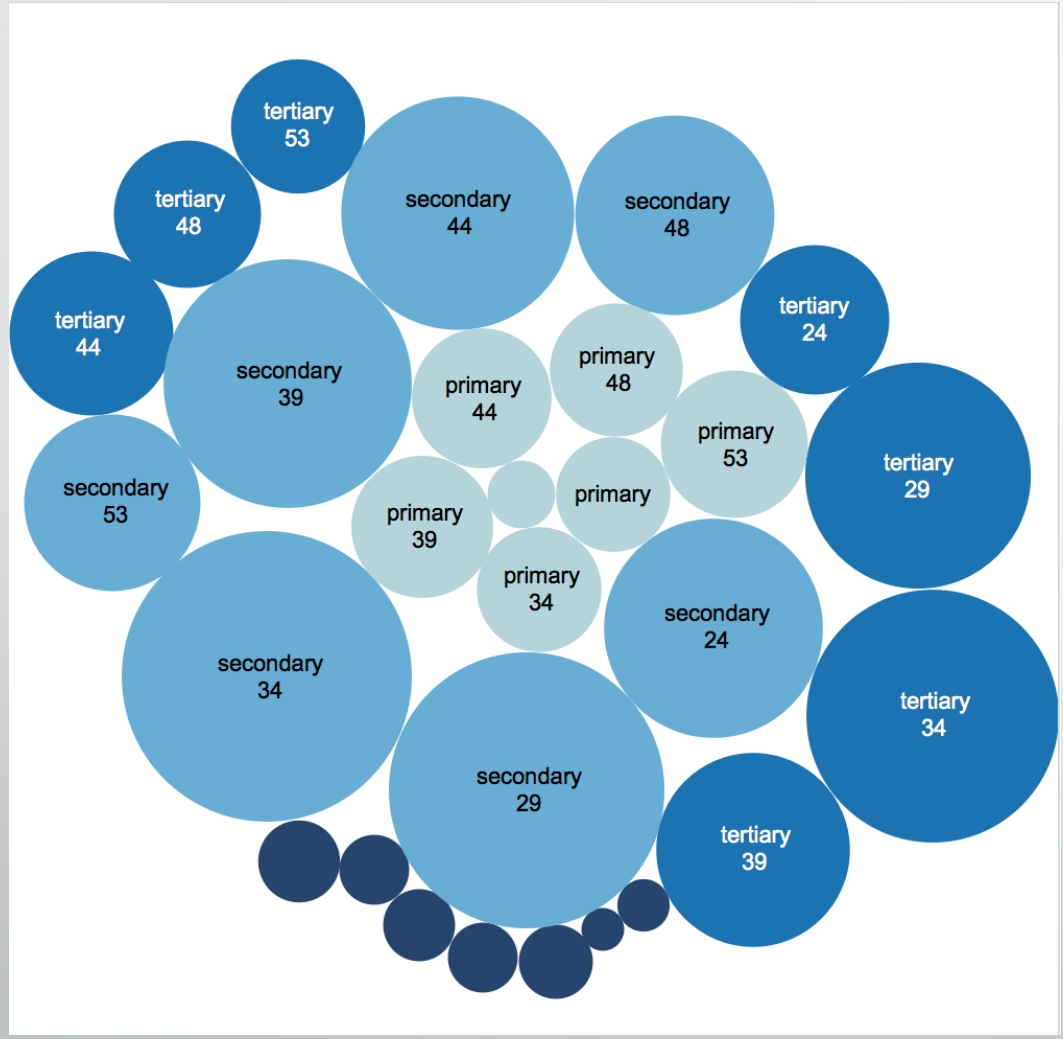
Exploratory Analysis – Marital Status



Targeted :

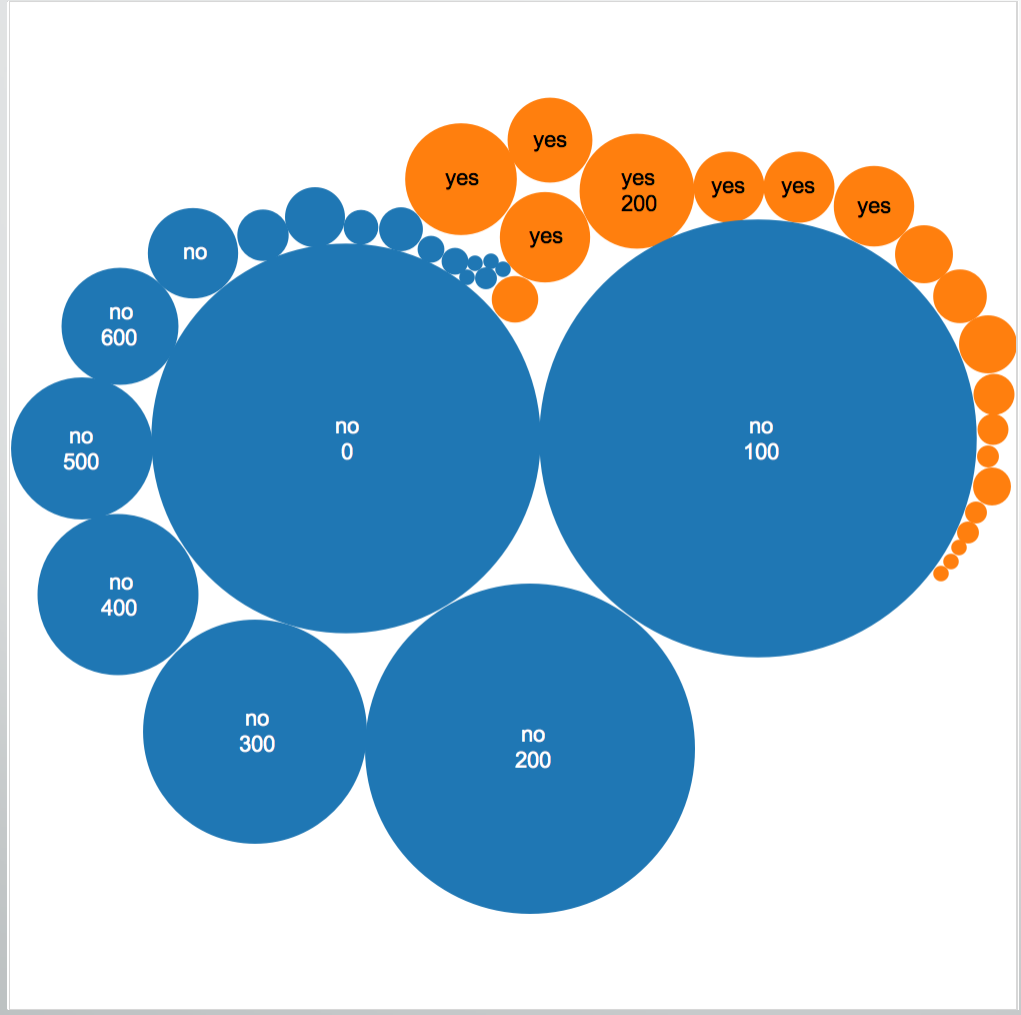
- Primarily, married customers in age bins 24 -53
- Single customers in middle age bins of 29-34

Exploratory Analysis – Education Status



- Majority, secondary level educated
- Tertiary level educated

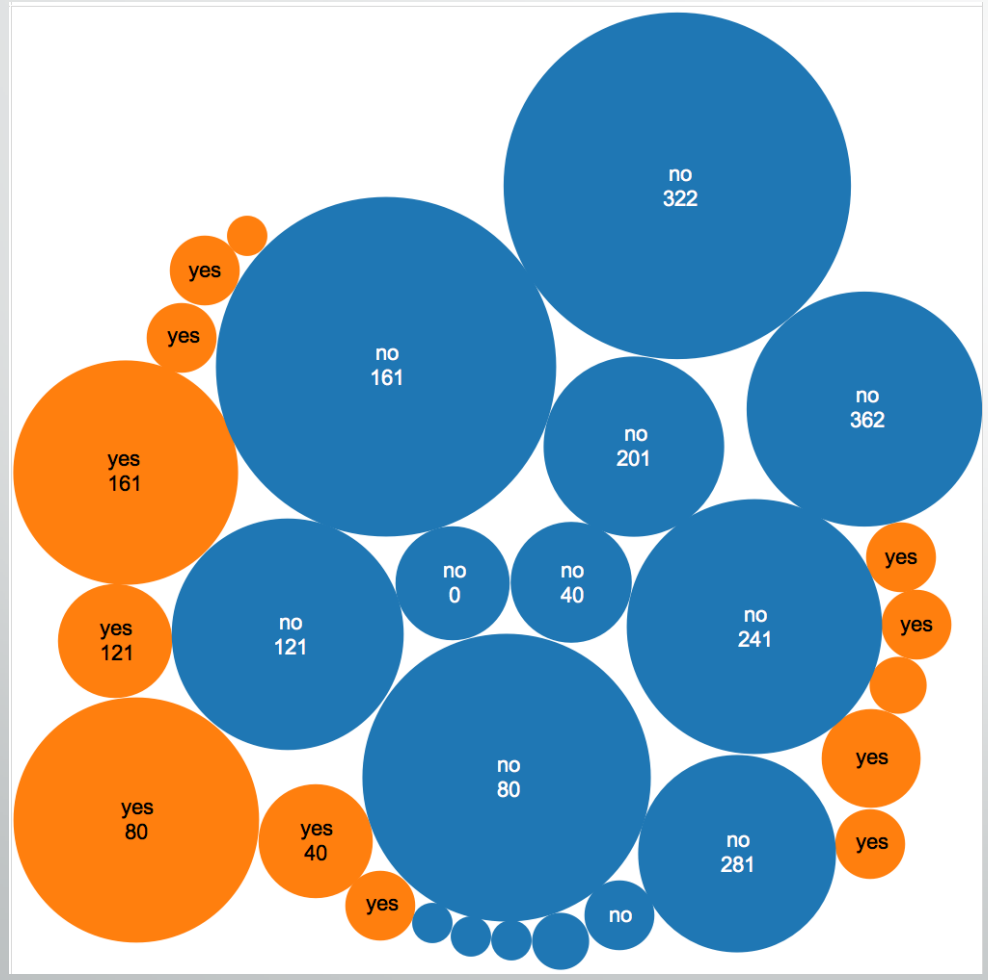
Exploratory Analysis
Relation between Call Duration and Conversion



- Majority of calls range from 0-300s
- Calls above 700s – almost always yes.
- Almost no conversion for up to 200s duration

* Need for more customer engagement in call to improve conversion rate?

Exploratory Analysis
Relation between recency of previous contact and conversion



- Most customers contacted >150 days prior to current contact
- Very less conversion for a lapse of 300 days and greater

* Continued customer contact to better chances?

Approach to Prediction

- 4522 data instances – 2714(Training) and 1808(Testing)
- Had both continuous and categorical data
- One of K encoding applied to categorical data for easier processing
- Algorithms tested
  - Decision Tree - Handles categorical attributes well
  - Random Forests - Simple and powerful ensemble classifier, must try!
  - Logistic Regression
  - Support Vector Machines

Prediction Results

- Decision Tree :
  - Default, pure leaf nodes error – 17.94% (Over-fitted!)
  - Max depth of 10 error – 13.55%
  - Best accuracy at max depth of 6 - 11.41%

- Random Forest :
  - Default(number of trees 10) error – 8.17%
  - Variation of number of tress/depth does not affect accuracy

Prediction Results

- Logistic Regression :
  - Default(Regularisation parameter, C = 1) error – 11.92%
  - Negligible change in accuracy on varying C
  - No better than Decision Trees!

- Support Vector Machine:
  - Default(Regularisation parameter, C = 1) error – 11.4%
  - Best performance at C = 1
  - No gain over Decision Trees/Random Forest Ensemble