A Data-driven Approach to Bank Telemarketing

Summary

Dataset

Source: A direct marketing campaign launched by the Portuguese Banking Institute.

Features: Previous campaign contact info and success + demographical data

Outcome/Dependent Variable: Campaign Success/Whether or not the contact signed up for term deposit

Size: **43600** entries and **17** variables

Analytics

Statistics: correlations, chi-squares,

Visualizations: histogram, bubble charts, tree diagrams

Machine Learning: Classification Tree, ROC curve, Random Forest Partial Dependence Plot

Tools: Rstudio

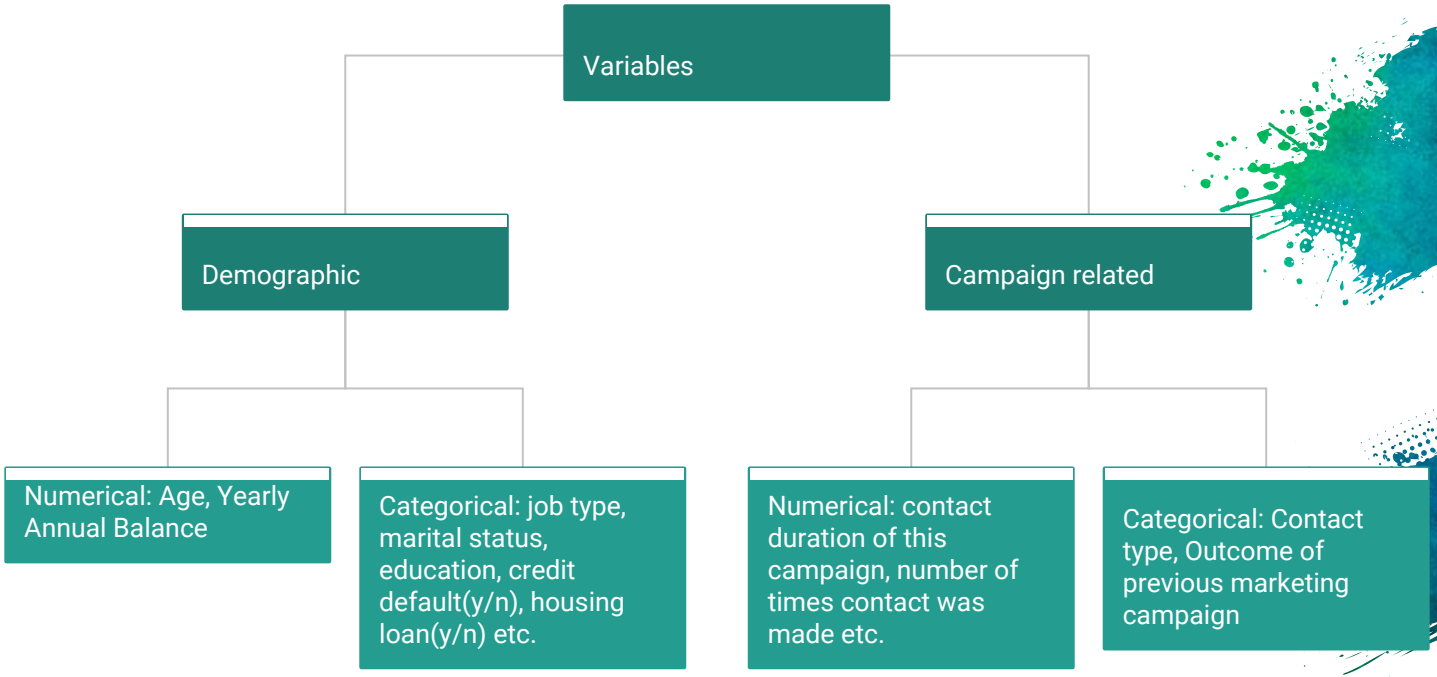
So what factors decide the outcome of campaign success in telemarketing?


Hello!
My name is Alan Luo
I love data and sharing ideas
You can find me on linkedin

1. The Portuguese Banking Institute Telemarketing Campaign Dataset

Reviewing data structures, removing outliers and visualizing data

Data structure



Outliers

- Clients range from age 20 to 80
- Annual balance between -8,019 and 102,127
- Age frequencies show that customers aged before 25 and after 60 cover less than 1% of the data

Sample Age Chart

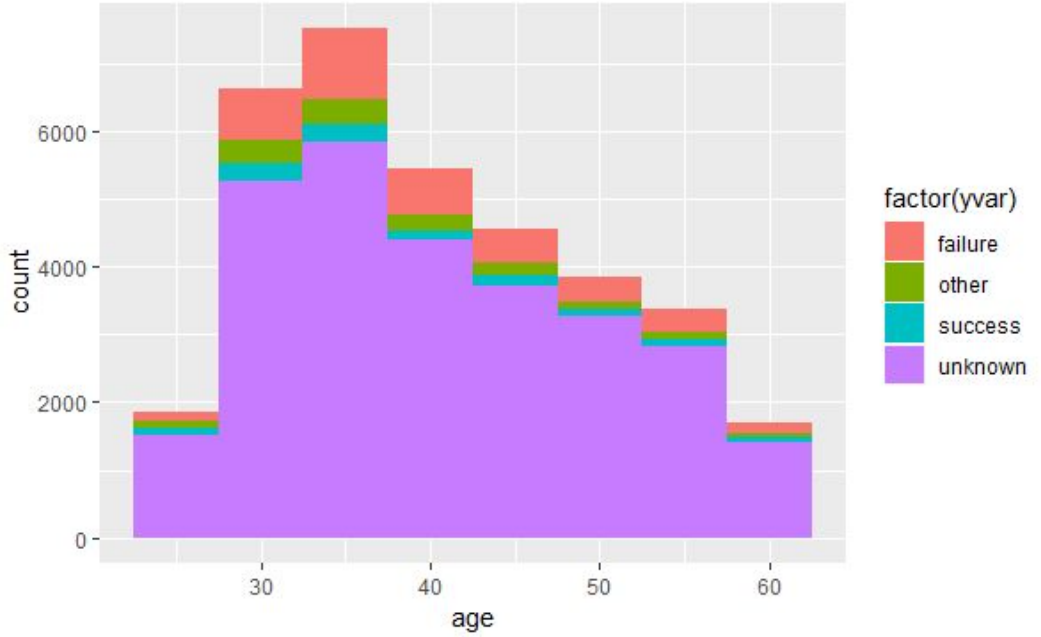
age	freq	percent
20	50	0.11
25	510	1.16
33	1922	4.40

Visualizing the data

- Histograms of relation between age and contact type

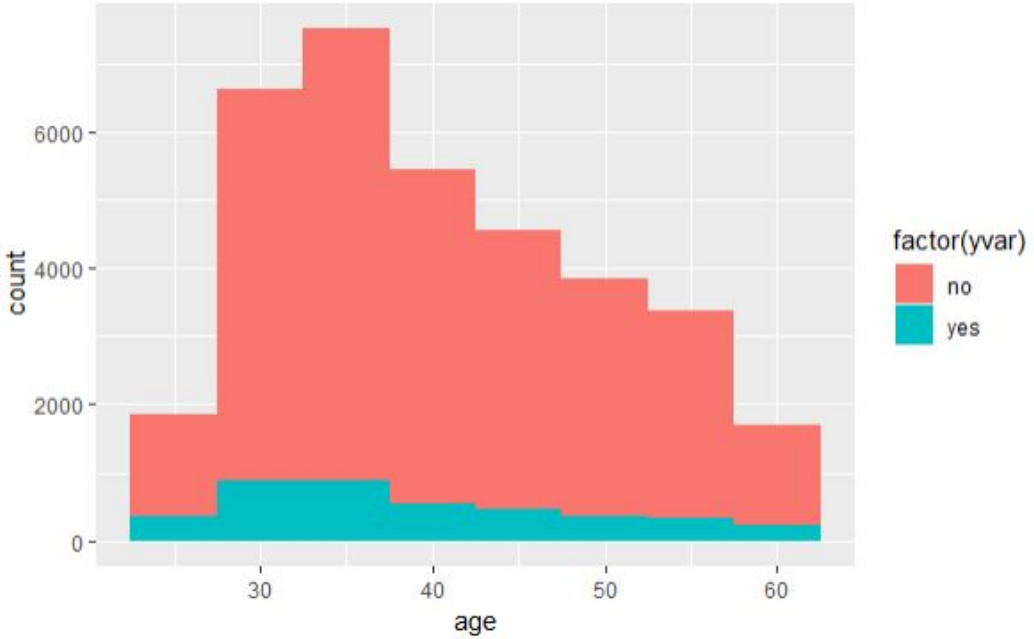
Visualizing the data

- Histograms of relation between age and success of previous campaign



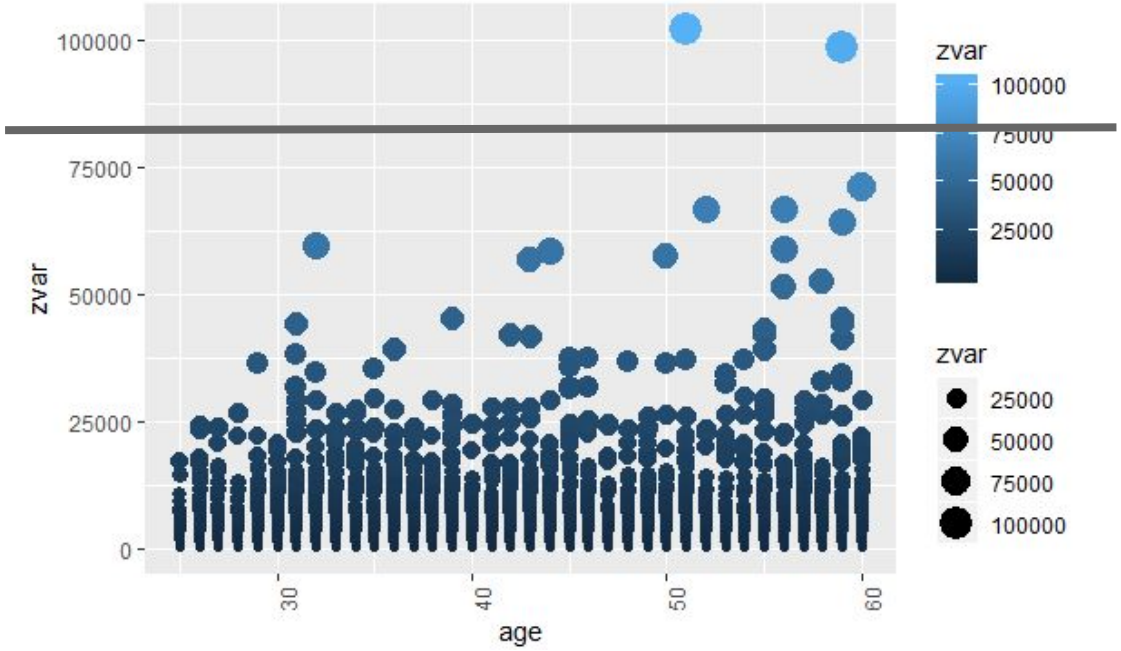
Visualizing the data (outcome)

- Histograms of relation between age and success of **current** campaign



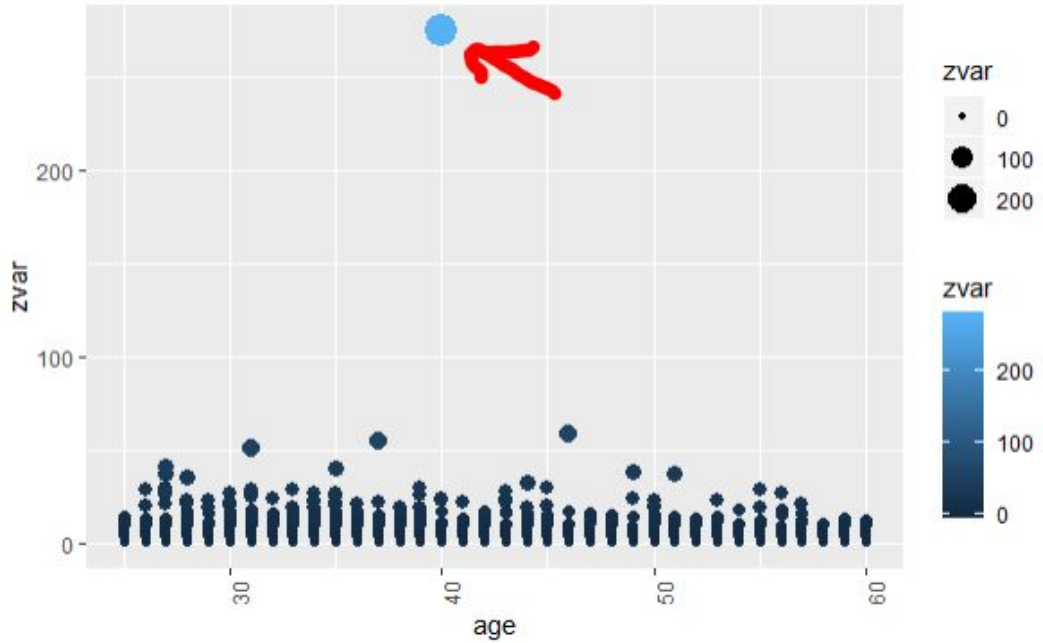
Visualizing the data

- Bubble Chart of relation between age and annual balance



Visualizing the data

- Histograms of relation between age and previous contact times



2. Statistical Analysis

Researching variable relations using correlations and chi-squares

Correlation Table

age annual_balance	duration	contact_times	previous_contact_times
age	1.00000000	0.083356651	-0.030597602
annual_balance	0.08335665	1.000000000	0.015113375
duration	-0.03059760	0.015113375	1.000000000
contact_times	0.05036392	-0.006504731	-0.048924942
previous_contact_times	-0.02750818	0.009739325	-0.008585207

Chi-Square Matrix

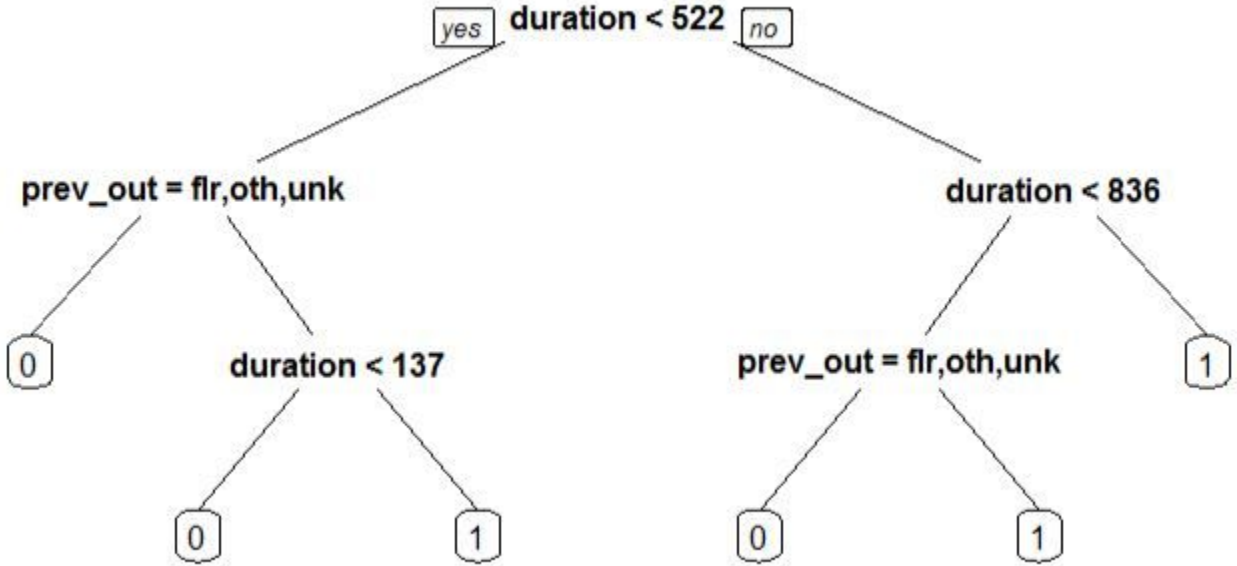
Chi-square Matrix	job	marital	housing_loan	personal_loan	contact_type	prev_outcome	outcome_term_deposit
job	NA	0.00	0.00	0.00	0.00	0.00	0.00
marital	NA	NA	0.00	0.00	0.00	0.00	0.00
education	NA	NA	0.00	0.00	0.00	0.00	0.00
housing_loan	NA	NA	NA	0.00	0.00	0.00	0.00
personal_loan	NA	NA	NA	NA	0.14	0.00	0.00
contact_type	NA	NA	NA	NA	NA	0.00	0.00
prev_outcome	NA	NA	NA	NA	NA	NA	0.00

Although there are no significant correlations between numerical variables, most categorical data display significant relations between each other.

3. Machine Learning

Building models to predict campaign outcome and results

Tree Based Model - Classification Tree



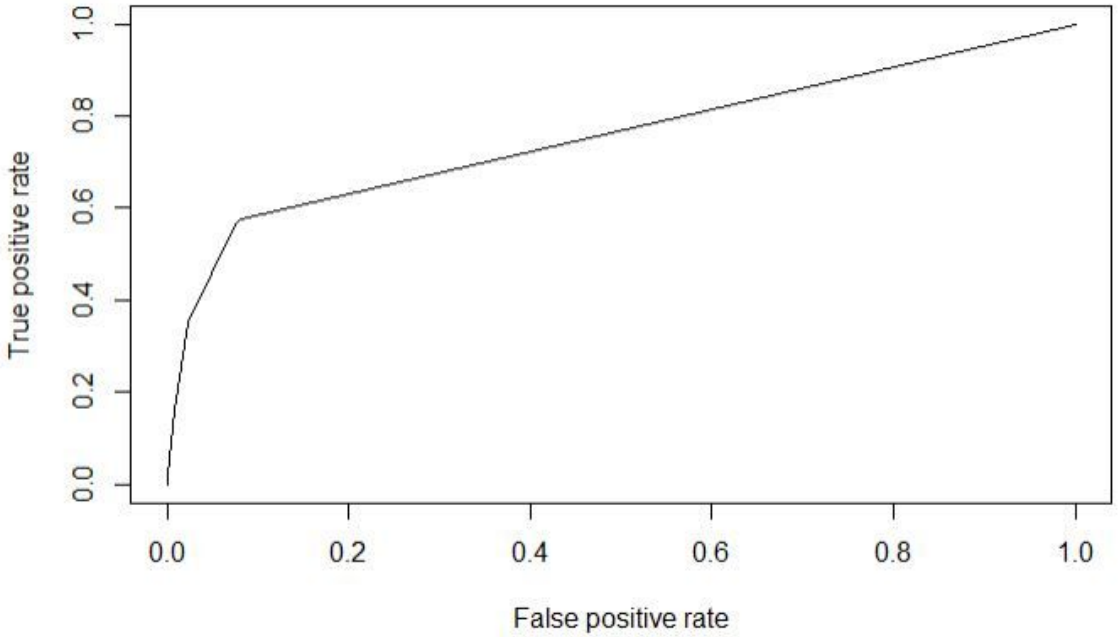
Classification Tree - Conclusions

We can draw a few conclusions based on the tree that R generated above:

1. The tree is solely based on duration and previous contact outcomes, which is an indication that other features did not stand out or show significance compared to these two features.
2. When the contact duration exceeds **14 minutes**, the likelihood for success increases.
3. If contact duration is below that, then the results would be based on the outcome of previous campaign, if the previous campaign was successful then the consumer would be more open to a term deposit if not then the consumer would less likely open up a term deposit.

ROC Curve

Our model is able to generate an accuracy rate of 90.41% and a left leaning ROC curve which indicates a reasonable accuracy of the prediction.



Random Forest - 200 Trees

```
#Deploy Random Forest - R code
if (!require(randomForest)) install.packages('randomForest')
library(randomForest)
bankForest <- randomForest(dummy_outcome_term_deposit ~ age + job + marital +
education + annual_balance + contact_type + duration + contact_times +
previous_contact_times + prev_outcome + dummy_default_credit +
dummy_housing_loan + dummy_personal_loan, data = train, nodesize = 25, ntree =
200)
predictForest <- predict(bankForest, newdata = test)
table(test$dummy_outcome_term_deposit,predictForest)
accuracy = (9067+386)/(9067+386+211+830)
print(accuracy)
```

The accuracy we received from the Random Forest Model is 90.08% which is a bit lower than the classification model, it is recommended that we use the classification model to make future predictions.

4. Further Research

Finding relations between demographic data and campaign outcome

Rerunning Classification Tree and Random Forest

We wanted to see if age/job/marital status/education could help us determine potential clients in the future.

So we removed duration and prev_out and only include variables highlighting personal demographic information.

And reran classification tree and random forest on our dataset

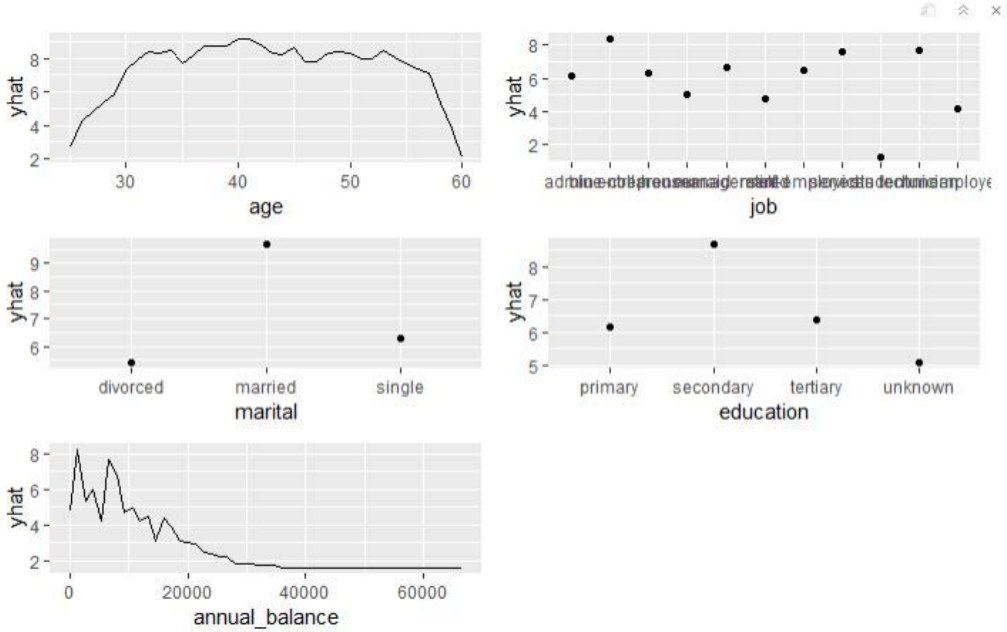
Random Forest - Conclusions

```
bankForest <- randomForest(dummy_outcome_term_deposit ~ age + job + marital +
education + annual_balance + dummy_default_credit + dummy_housing_loan +
dummy_personal_loan, data = train, nodesize = 25, ntree = 200)
table(test$dummy_outcome_term_deposit,predictForest)
accuracy = (8827+146)/(1070+8827+451+146)
print(accuracy)
```

The accuracy we received from the Random Forest Model is <u>85.5%</u> ,

Partial Dependence Plot

We deployed a partial dependence plot to most features in order to understand what affected the model the most.



Conclusion

1. We can conclude that the success of this particular campaign was determined by the contact duration and the success of the previous campaign.
2. From a marketing standpoint, it makes sense that consumers that are willing to listen to the seller for long periods of time and is already a customer from the beginning is more likely to accept similar services from the same organization.
3. This data and the classification tree we've built shows just that. While we are also able to generate a high accuracy machine learning model without data containing the two important features mentioned above, we are unable to determine what actually matters in terms of consumer demographics in relation to campaign success.
4. However, this model can still be used to predict future campaign success if marketers want to target specific groups of consumers. The only issue would be the lack of transparency and understanding of how the model works internally.


Thanks!
Any questions?