Bank Marketing Project

Yufei Tang

Lei Wang

Hua Lu

Agenda

01

Project Introduction

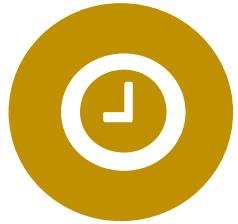
02

Data Information

03

Data Preprocessing

04

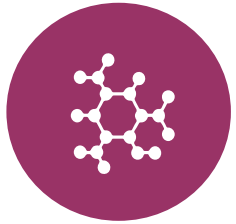


Data Model 1 – Logistical Regression

05

Data Model 2 – Decision Tree

06



Data Model 3 – Neural Network

07

Other Analyses

08

Summary

Project Introduction

Project Background

The whole national economy in the country is stagnant for a long time

Increasing competition among banks in the banking market

The Bank continues to lose its current customers and hardly to attract new customers

Customers have various investment opinions in the current market

The bank has challenges to identify and predict customer's consumption behaviours

Project Introduction

Project Goal & Benefits

Project Goal

- To distinguish potential targeted customers for our bank
- To analyse influences of each attribute on customer's final decision

Project Benefits

To predict whether a customer will subscribe a term deposit in our bank or not

To contact with most potential consumers effectively

To control and reduce the marketing cost

Data Information

The data is related with direct marketing campaigns of a Portuguese bank

The marketing campaigns were based on phone calls

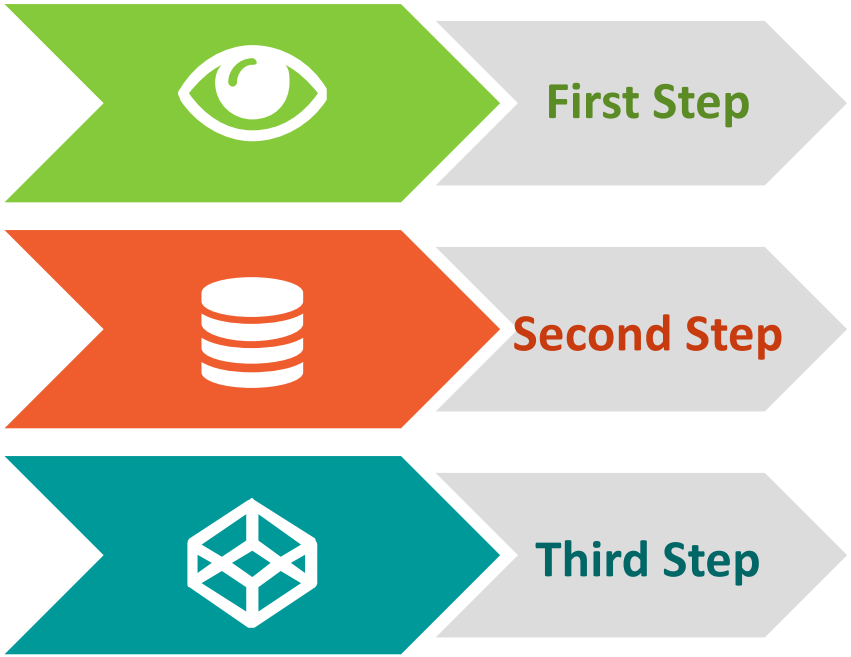
45211 pieces of customer information in total

16 input variables (Client information) ： such as age, education, housing, jobs, etc.

One output variable (Desired target) : Has the client subscribed a term deposit in our bank yet?

Data Preprocessing

Preprocessing Steps



- Data Elimination/ Attribute Deduction
- Variable Transformation / Normalization
- Binning

Data Preprocessing

Data Elimination/ Attribute Deduction

Missing Data Elimination:

Why 1) Large scale(over 50%) data missing in “contact” and “poutcom” columns
2) Occasional data(less than 5%) missing in “education” column

How Removed “contact” and “poutcom” columns
Deleted rows with missing date in “ education” column

Attribute Reduction:

Why 1) Information gained after result is obtained--- “duration” column

How Removed this column

Data Preprocessing

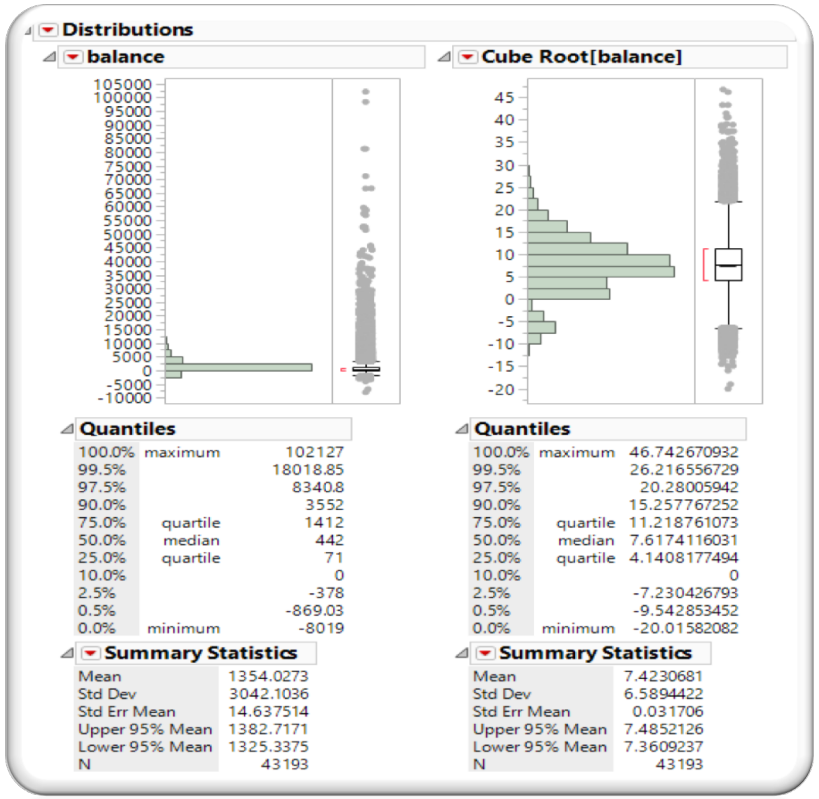
Variable Transformation / Normalization

The distribution of " balance" is left skewed and the value range is too large

Why

Using cube root function to transform this variable into a normal distribution

How



Data Preprocessing

Binning

1)Making the dataset easier to handle and interpret
2) Avoid over emphasize on specific continuous variables

Why

Binning categories under " job" and " month", based on the Row percentage result of each category as a function of Yes/No

How

job

Count	Old Values (11)	New Values (5)
9216	management	ATMS
7355	technician	
5000	admin.	
1540	self-employed	
9278	blue-collar	HEBSer
4004	services	
1411	entrepreneur	
1195	housemaid	
2145	retired	retired
775	student	student
1274	unemployed	unemployed

month

Count	Old Values (12)	New Values (5)
2820	apr	feb,apr
2533	feb	
6037	aug	jan,jun,aug,nov
4980	jun	
3842	nov	
1318	jan	
448	mar	mar
13192	may	may,jul
6601	jul	
690	oct	sep,oct,dec
532	sep	
200	dec	

Data Model 1 – Logistic Regression

Function 1

Make a Prediction about whether a customer will subscribe a term deposit

Step 1

Using Fit model to obtain a predictive formula

Function 2

Identify most important variables that will influence customers' depositing behaviour

Step 2

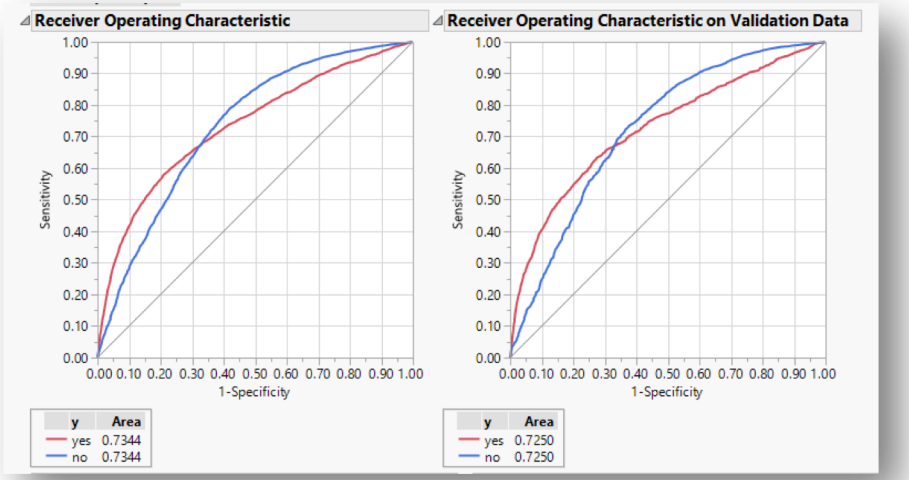
Using Stepwise regression to retain the most important ones

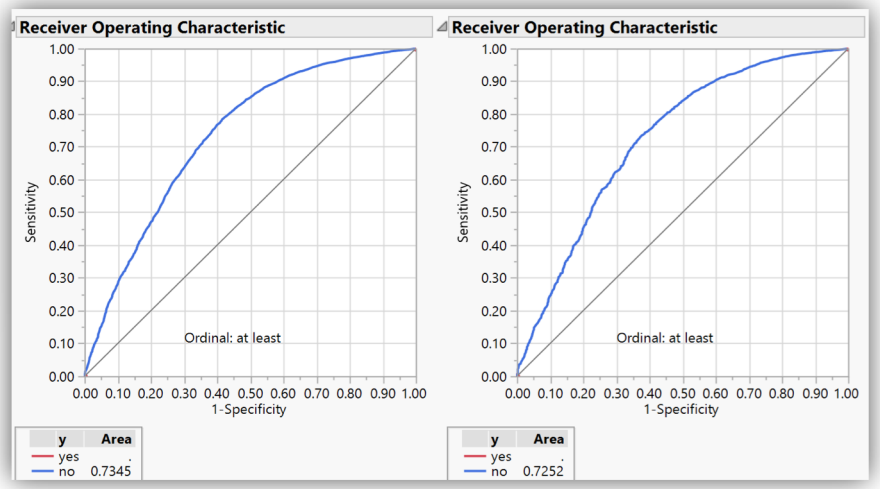
Data Model 1 – Logistic Regression

Model Evaluation

Roc Curves

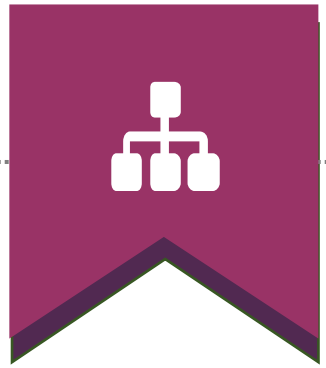
According to the ROC curves of Nominal Logistic and Stepwise Regression, the model has stable and predictive ability.





Data Model 1 – Logistic Regression

Misclassification Rate



Misclassification Rate

Nominal Logistic Regression: 0.1158 in validation dataset
Stepwise Regression: 0.1158 in validation dataset
No significant performance difference in making predictions

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.1138	0.1106	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1531	0.1493	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3179	0.3214	∑ -Log(ρ[j])/n
RMSE	0.3028	0.3041	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1834	0.1841	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1154	0.1158	∑ (ρ[j]≠ρMax)/n
N	32377	10816	n

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.1139	0.1107	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1533	0.1494	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3178	0.3214	∑ -Log(ρ[j])/n
RMSE	0.3028	0.3041	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1833	0.1841	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1154	0.1158	∑ (ρ[j]≠ρMax)/n
N	32377	10816	n

Data Model 1 – Logistic Regression

Cutoffs

Cutoffs

Opportunity costs: High opportunity cost which is the profit we could earn, if we predict a False Negative.

Predicted Yes VS. Yes				
Cutoff	0.5	0.4	0.3	0.2
PredictY/Y	8.99%	9.04%	22.71%	36.99%

Data Model 1 – Logistic Regression

Conclusion

1. It is difficult to make an accurate prediction using logistic regression
2. We can compare the effect of variables in Nominal Logistic and Stepwise Regression to identify and decide which variables to retain.
3. The performances of Nominal Logistic and Stepwise Regression are alike in terms of making predictions

Data Model 2 – Decision Tress

Leaf Report

Leaf Report

Leaf report of customer characteristics and subsequent consequences

Days since last contact between 176 and 193, no housing=>
P(Y=Yes) = 58% (255 Yes, 184 No)

Younger than 61, less than 10 days since last contact, in months of September, October and December, call day is after 21st of the month =>
P(Y=Yes) =57.57% (77 Yes, 56 No)

Days since last contact between 81 and 101 =>
P(Y=Yes) = 52.94%(318 Yes, 282 No)

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.1618	0.1438	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2141	0.1918	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3007	0.3094	∑ -Log(ρ[j])/n
RMSE	0.2929	0.2984	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1716	0.1754	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1116	0.1163	∑ (ρ[j]≠ρMax)/n
N	32377	10816	n

Confusion Matrix

Training

Actual	Predicted Count	
y	no	yes
no	27927	697
yes	2917	836

Validation

Actual	Predicted Count	
y	no	yes
no	9281	267
yes	991	277

Data Model 2 – Decision Tress

Conclusion

1. Easier to interpret the results, target groups could be prioritized for decision makers
2. Depicted specific characteristics of certain groups
3. Dispersed groups and large number of leaves make it difficult for the bank to initiate marketing campaigns

Data Model 3 – Neural Networks

Neural Networks

Combining different amount of nodes in layers and different activation Functions to generalize or recognize unrealized patterns

Finding 1

Comparison of Activation functions(TanH, Linear, Gaussian)

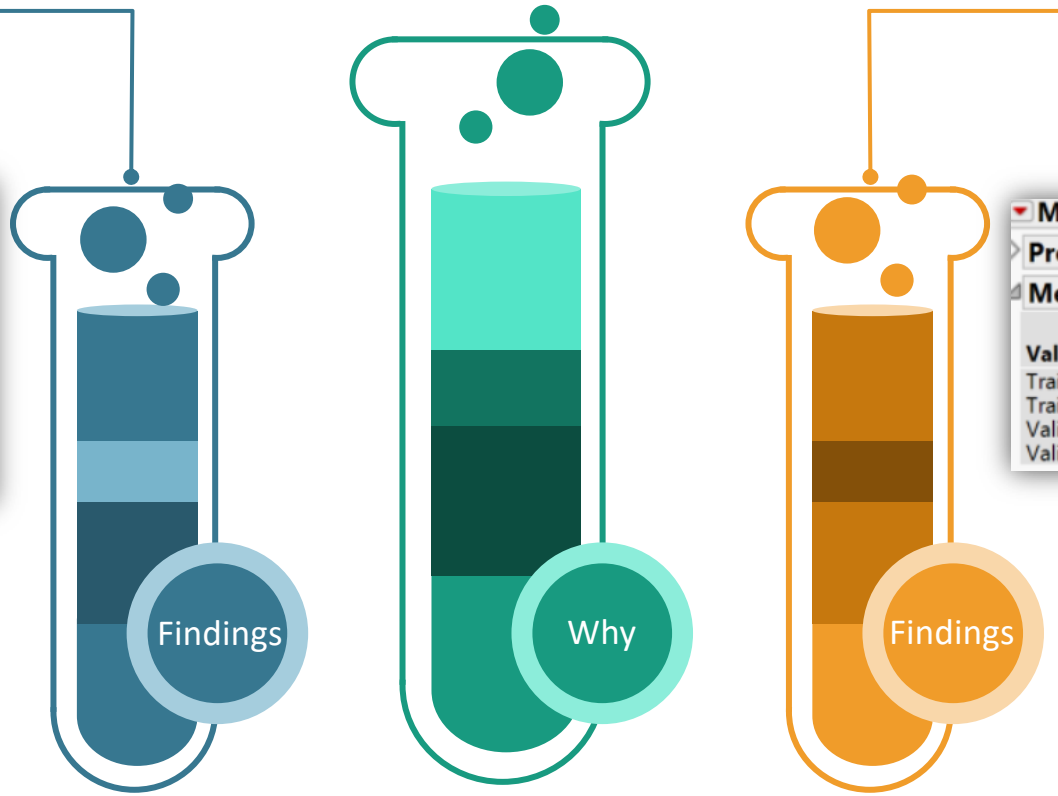
Model Comparison

Predictors

Measures of Fit for y

Validation	Creator	.2 .4 .6 .8	RMSE	Mean Abs Dev	Misclassification Rate
Training	Neural		0.2956	0.1747	0.1122
Training	Neural		0.3027	0.1834	0.1153
Training	Neural		0.2977	0.1774	0.1126
Validation	Neural		0.2996	0.1768	0.1153
Validation	Neural		0.3042	0.1842	0.1155
Validation	Neural		0.3010	0.1792	0.1156

TanH has the lowest Misclassification Rate



Finding 2

Comparison of Number of Nodes/ Layers (3/3, 3/0)

Model Comparison

Predictors

Measures of Fit for y

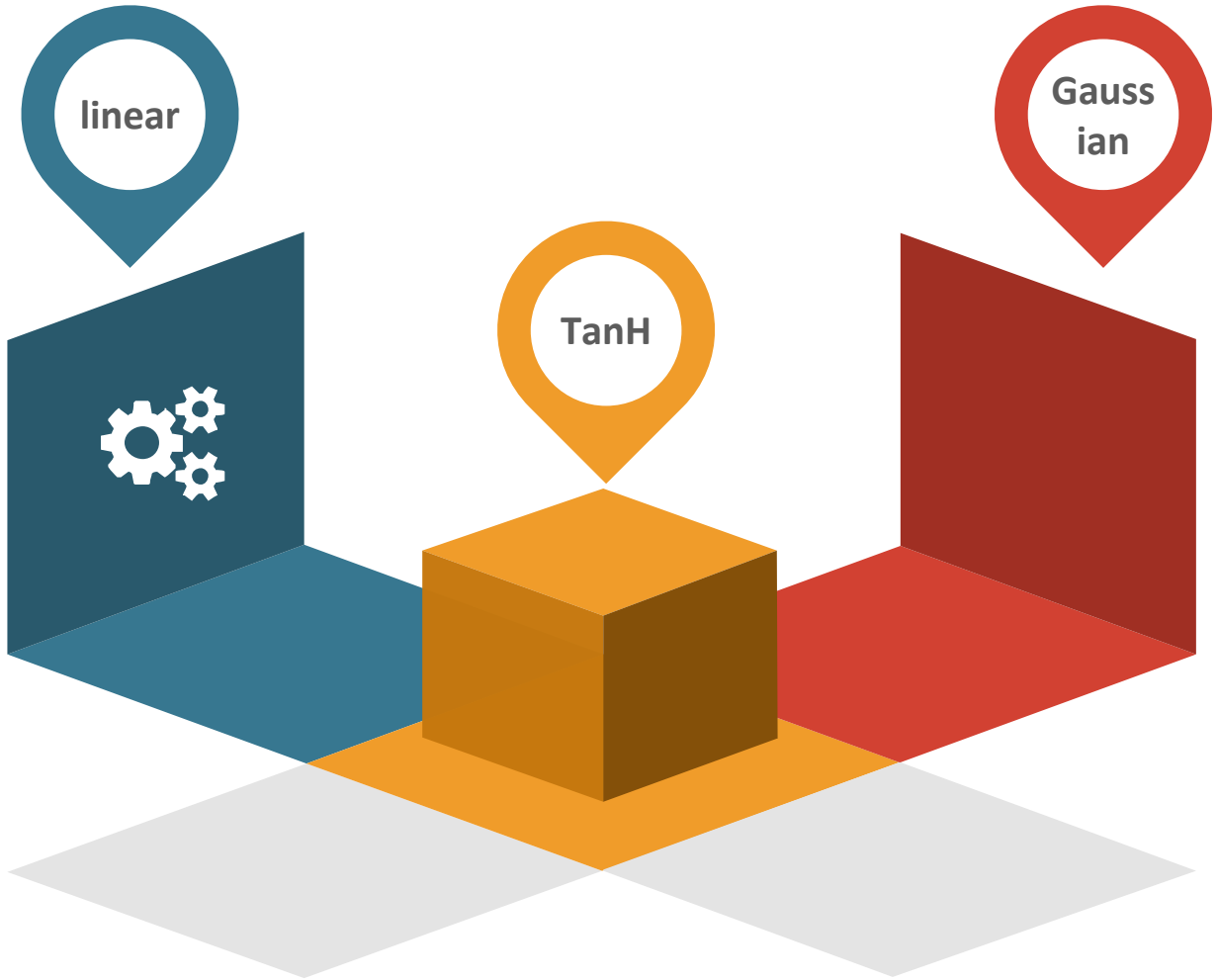
Validation	Creator	.2 .4 .6 .8	RMSE	Mean Abs Dev	Misclassification Rate
Training	Neural		0.2956	0.1747	0.1122
Training	Neural		0.2972	0.1774	0.1152
Validation	Neural		0.2996	0.1768	0.1153
Validation	Neural		0.3015	0.1800	0.1168

The misclassification rate of TanH with 3/0 of nodes is slightly lower

Data Model 3 – Neural Networks

Conclusion

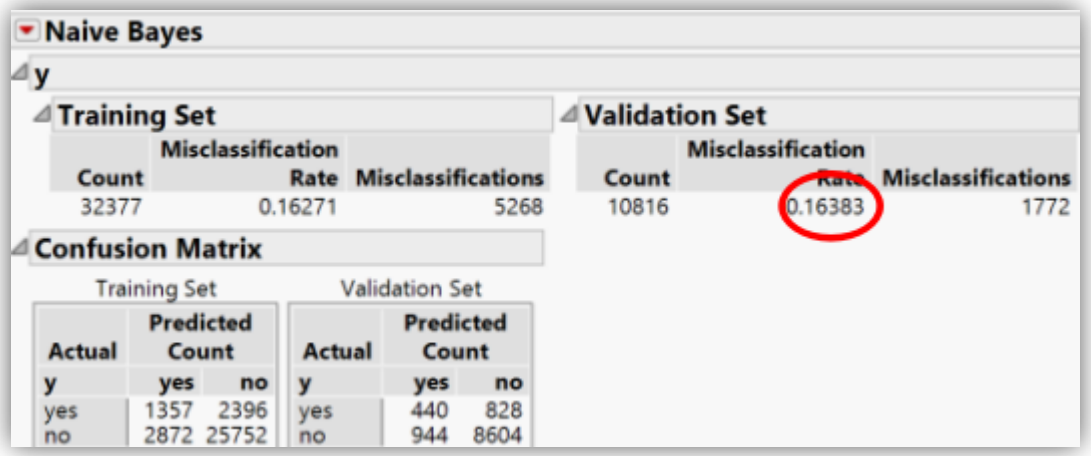
By using TanH function with 3/0 nodes which has the highest rate of predicting “Yes”, the Bank could make predictions based on Neural Network method



Other Analyses– Naïve Bayes

It is time effective to obtain a predictive model

Comparison basis to find best model to solve the problem



Naive Bayes

y

Training Set

Count	Misclassification Rate	Misclassifications
32377	0.16271	5268

Validation Set

Count	Misclassification Rate	Misclassifications
10816	0.16383	1772

Confusion Matrix

Training Set

Actual	Predicted Count	
y	yes	no
yes	1357	2396
no	2872	25752

Validation Set

Actual	Predicted Count	
y	yes	no
yes	440	828
no	944	8604

Conclusion

01 Number One

With best performance in predicting probability of people who would accept the offer

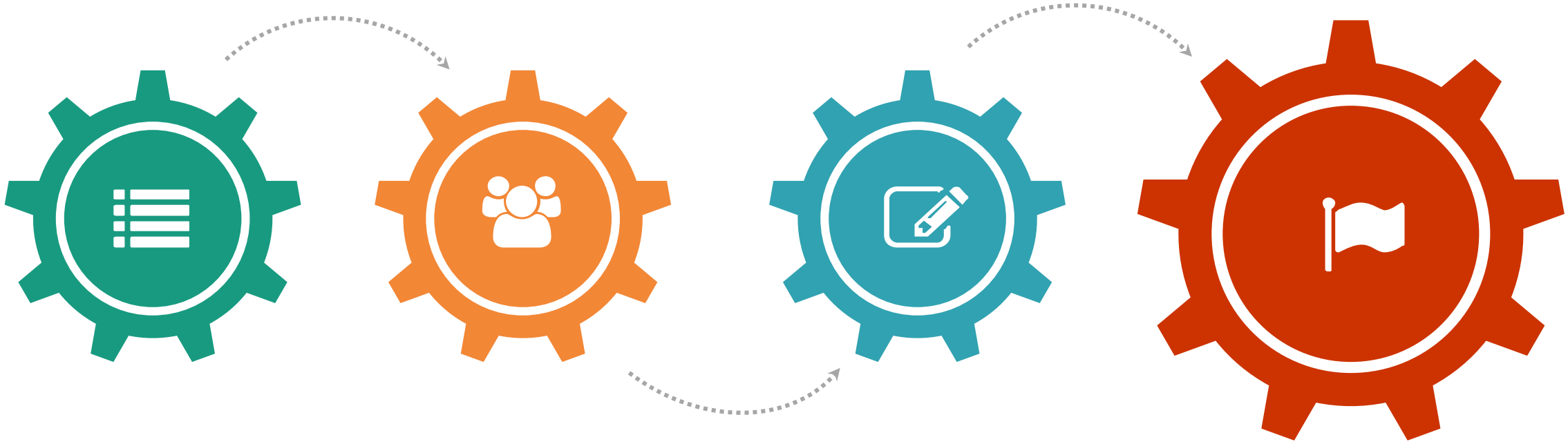
02 Number Two

With noticeably higher Misclassification in exchange

03 Number Three

The bank should use this model with caution

Other Analyses– Hierarchical Clustering



Reason 1
To segment customers and find factors that may affect their decisions

Reason 2
To find target groups of customers that have higher probability to accept the offer

Reason 3
To help the bank to design marketing strategies toward target customers

How
Start from a larger number of clusters, and gradually approach the best number where the objective of the clustering is met

Other Analyses– Hierarchical Clustering

Conclusion

- Out Eight customer clusters we proposed to study, we identified 3 of them as target groups
- The three target groups have symbolic characteristics to label and to distinguish from the other
- The counts of the three groups are not large, efforts could be made to enlarge target groups by formulating a larger customer basis
- Marketing campaigns can be devised according to the target clusters to expand cluster size with high value or to improve converting rates or potential customers

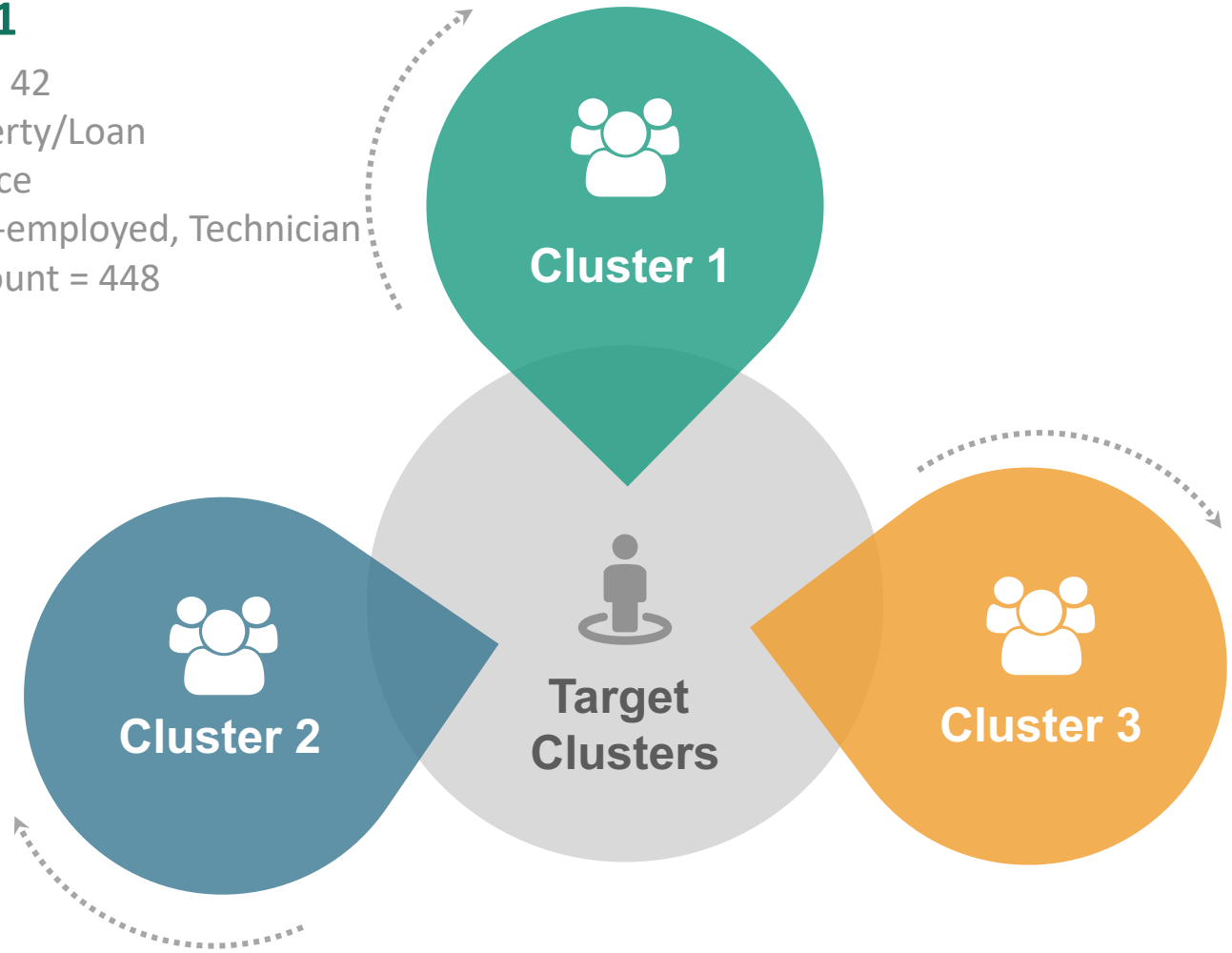
Other Analyses– Hierarchical Clustering

Cluster 1

Average age 42
No Default/Property/Loan
High Balance
Job: Admin, Management, Self-employed, Technician
P(Yes)= 53.79%, Count = 448

Cluster 2

Average age 54
No Default/Property/Loan
High Balance
Job: retired
P(Yes) = 29.81%, Count = 3096



Cluster 3

Average age 26
Single
No Default/Property/Loan
Average balance
Job: Student
P(Yes) = 28.80%, Count = 743

Summary

Models Comparison

Model	Misclassification	Predict "Yes"
Logistic Regression	0.1158	8.99%
Decision Tree	0.1163	21.85%
Neural Networks	0.1153	14.9%
Naïve Bayes	0.1638	34.7%

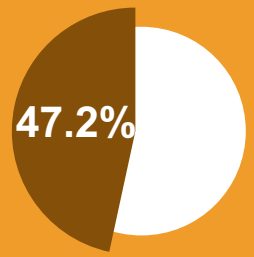
Predicted Yes VS. Yes (Naïve Bayes)				
Cut-off	0.5	0.4	0.3	0.2
P(Yes) VS. Yes	34.7%	38.26%	40.9%	44.4%
Predicted Yes VS. Yes (Logistic Regression)				
Cut-off	0.5	0.4	0.3	0.2
P(Yes) VS. Yes	8.99%	9.04%	22.71%	36.99%
Predicted Yes VS. Yes (Decision Tree)				
Cut-off	0.5	0.4	0.3	0.2
P(Yes) VS. Yes	21.85%	32.66%	43.23%	47.2%
Predicted Yes VS. Yes (Neural Network)				
Cut-off	0.5	0.4	0.3	0.2
P(Yes) VS. Yes	14.9%	29.47%	31.71%	41.76%

- Neural Networks has the lowest misclassification
- Naive Bayes has a significantly better accuracy in predicting "Yes"
- Decision Tree Model is recommended, taking into consideration of opportunity costs

Summary

Benefits

- Our project was able to provide a model that can predict response of potential customer with 47.2% accuracy
- The three customer groups we identified could help the bank to better interpret their most interested customers and to improve converting rates amongst potential customers



Increase Accuracy

Improvement

Further improvements could be made:

- by integrating results of Stepwise Regression and that of other predictive models
- by collect new data according to our finds to enhance stability, reliability and accuracy

Thank you !