Bank Marketing Analysis Project

Zhihao Guo, Sherry Nie, Jiahui Ji, Yanmeng Song

Data Description

- From UCI Machine Learning Repository
- A Portuguese banking institution's client data from a campaign
- Number of observations: 41188
- Number of variables: 20
- Response: if the client will subscribe a bank term deposit (YES/NO)

# bank client data		
1 - age		numeric
2 - job		categorical
3 - marital		categorical
4 - education		categorical
5 - default	Credit in default (YES/NO)	categorical
6 - housing	Housing loan (YES/NO)	categorical
7 - loan	Personal loan (YES/NO)	categorical
# related with the last contact of the current campaign		
8 - contact	Communication type	categorical
9 - month	Last contact month of the year	categorical
10 - day_of_week	Last contact day of the week	categorical
11 - duration	Last contact duration	numeric
# other attributes		
12 - campaign	Number of contacts performed during this campaign	numeric
13 - pdays	Number of days passed since a previous campaign	categorical
14 - previous	Number of contacts performed before this campaign	numeric
15 - poutcome	Outcome of the previous campaign	categorical
# social and economic context attributes		
16 - emp.var.rate	Employment variation rate (quarterly)	numeric
17 - cons.price.idx	Consumer price index (monthly)	numeric
18 - cons.conf.idx	Consumer confidence index (monthly)	numeric
19 - euribor3m	Euribor 3 month rate (daily)	numeric
20 - nr.employed	Number of employees (quarterly)	numeric

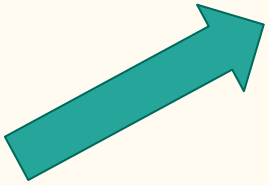
Goals

- Investigate which variables are significant in predicting whether a client will subscribe a term deposit or not
- Predict the subscription result applying multiple classification methods
- Compare prediction results before and after adding 5 social and economic attributes

Data Pre-Processing

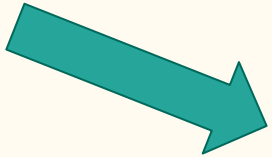
Deal with unbalanced data

no	36548
yes	4640



no	4000
yes	4000

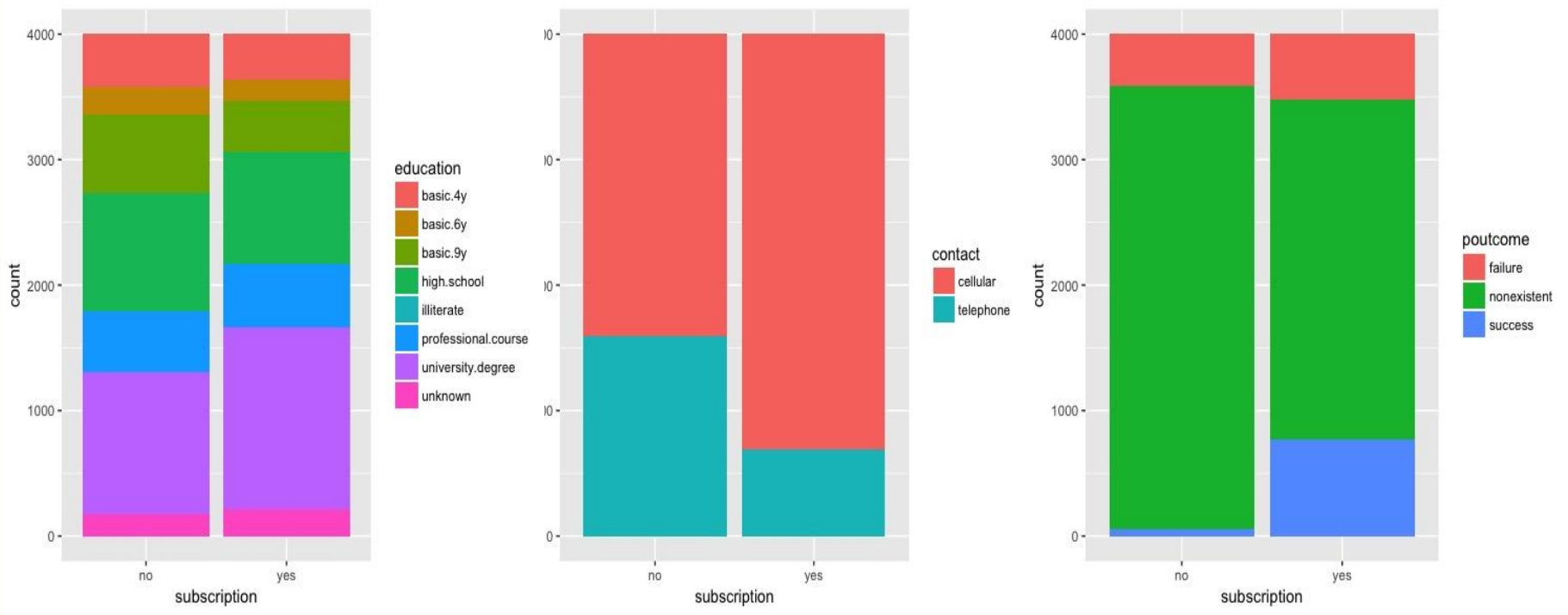
Training data



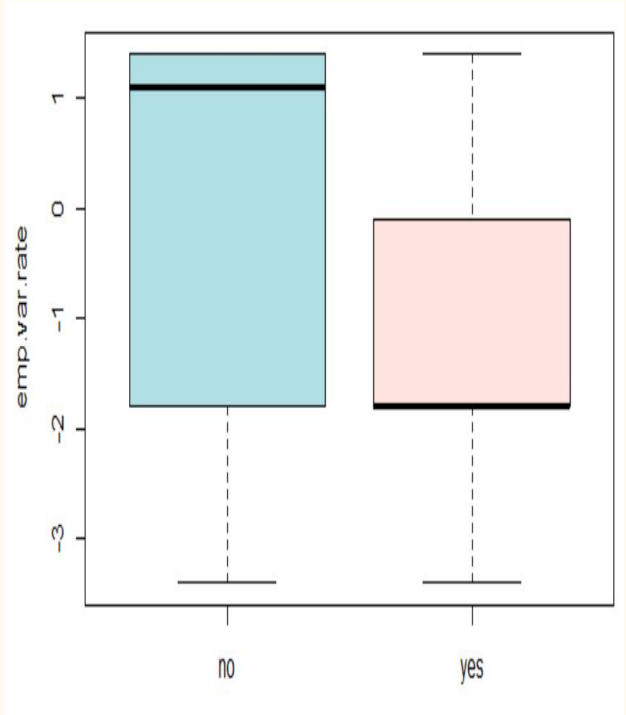
no	640
yes	640

Test data

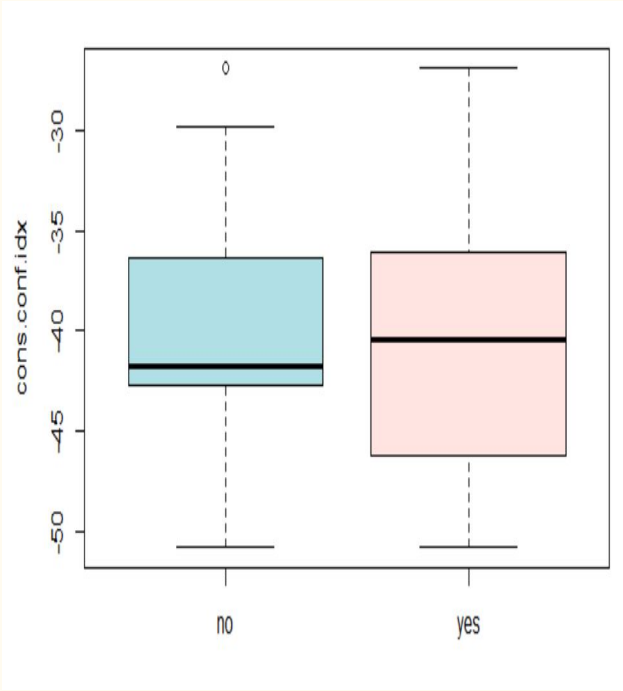
Visualization-- Categorical variables



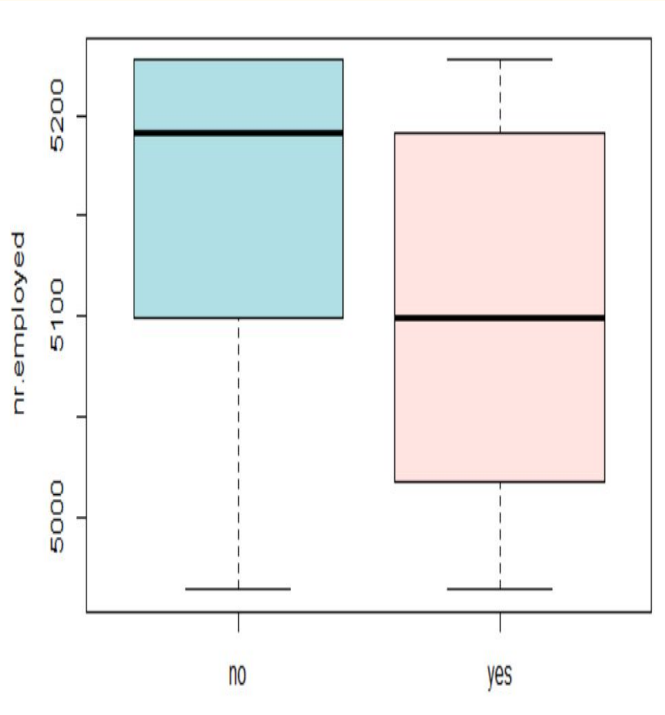
Visualization--Numeric variables



emp.var.rate



Consumer Confidence Index



Number of Employment

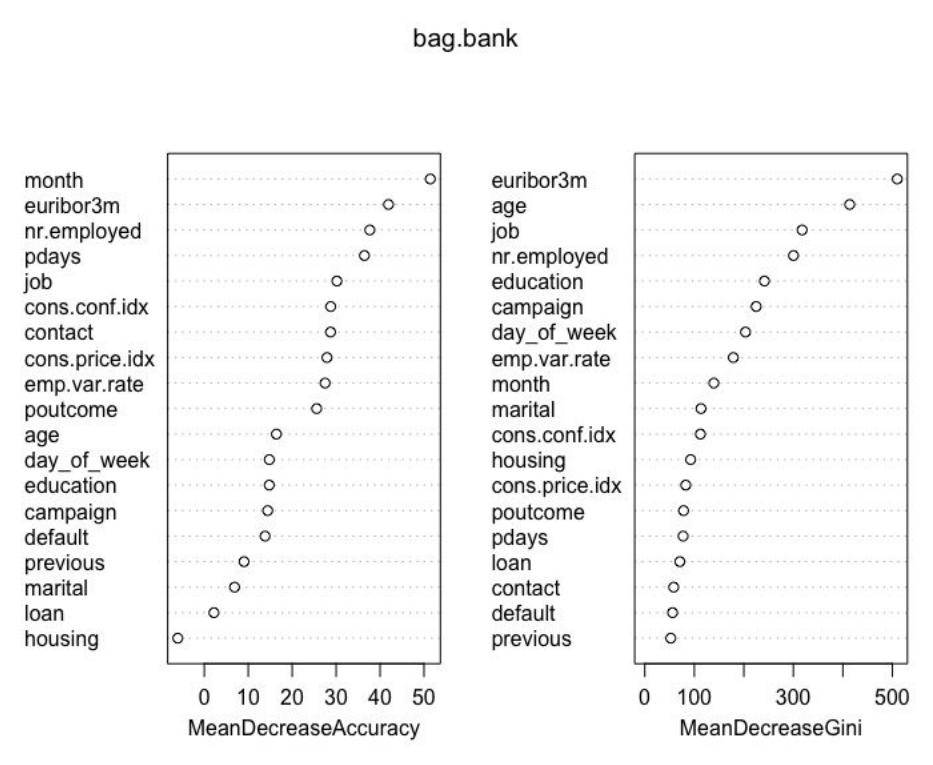
Variable Selection

- Random Forest
- Lasso
- Forward Selection
- Backward Selection

Choose variables that are selected by AT LEAST 3 of them!

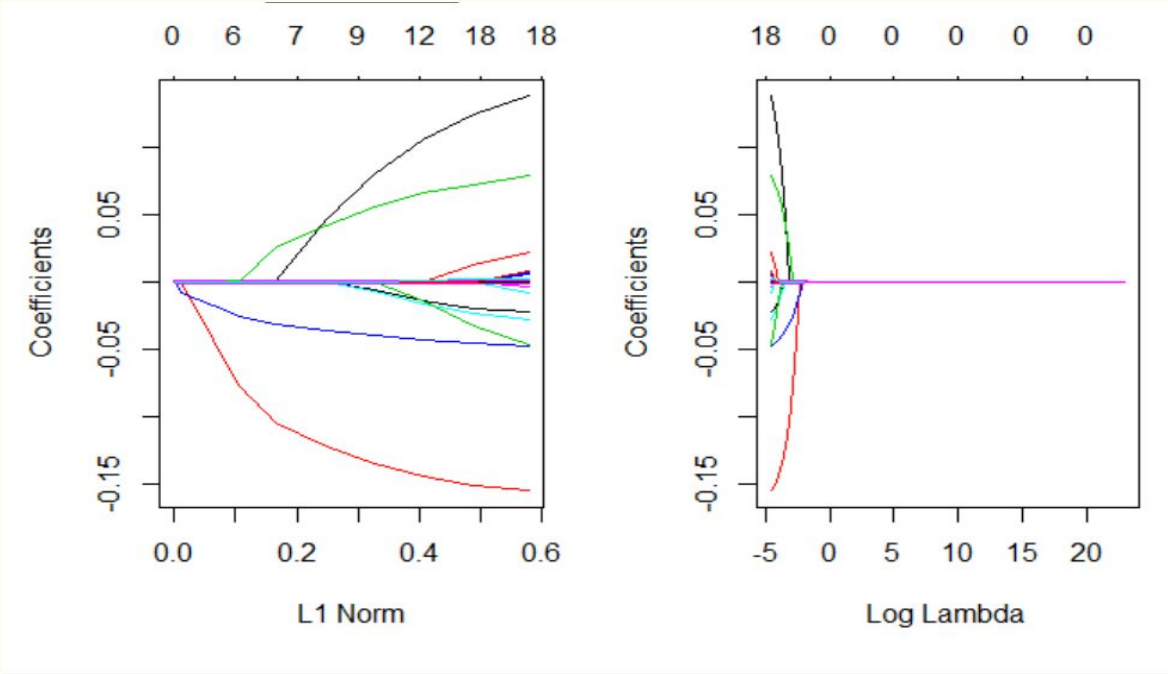
Variable Selection - RandomForest

- “MeanDecreaseAccuracy” criterion: I pick ten variables from “months” to “poutcome”.
- “MeanDecreaseGini” criterion: I pick eight variables from “euribor3m” to “emp.var.rate”.



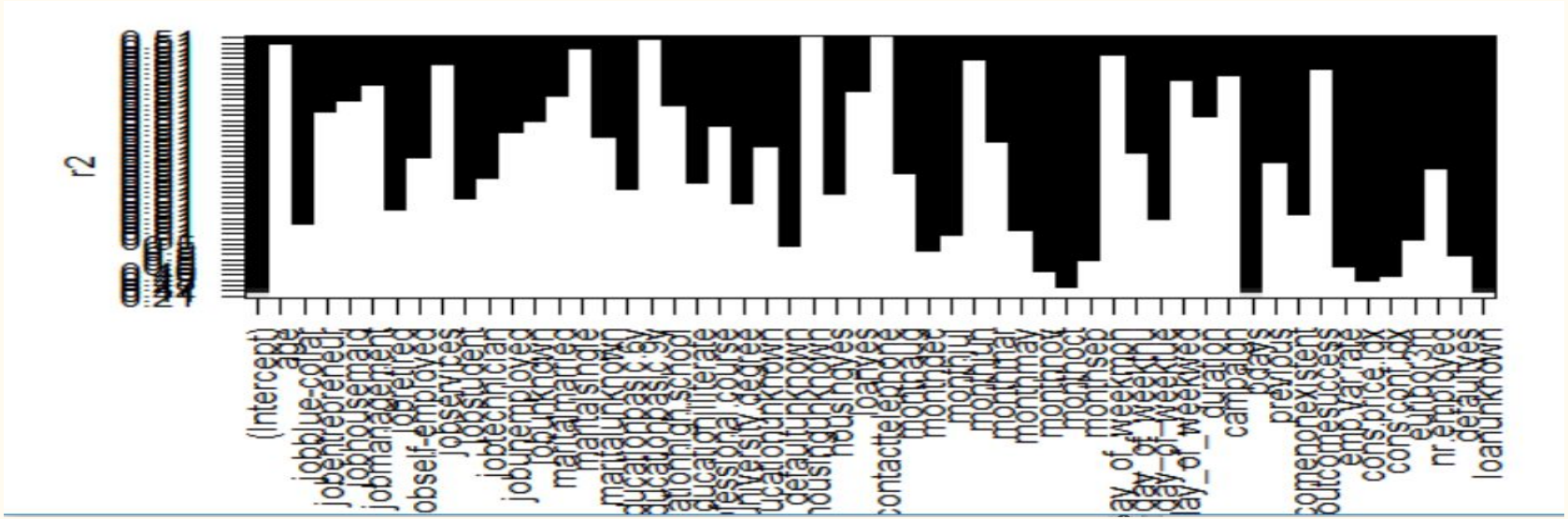
Variable Selection - Lasso

Variables:
Job, Education, Contact,
Default, Day_of_week,
Month, poutcome,
Emp.var.rate, Nr.employed.



Variable Selection - Forward Method

- Forward method and Backward method gives the same result
- Variables: Job, Education, Default, Month, Campaign, Day_of_week, Previous, Poutcome, Emp.var.rate, Cons.price.idx, Nr.employed, Housing, loan, Euribor3m.

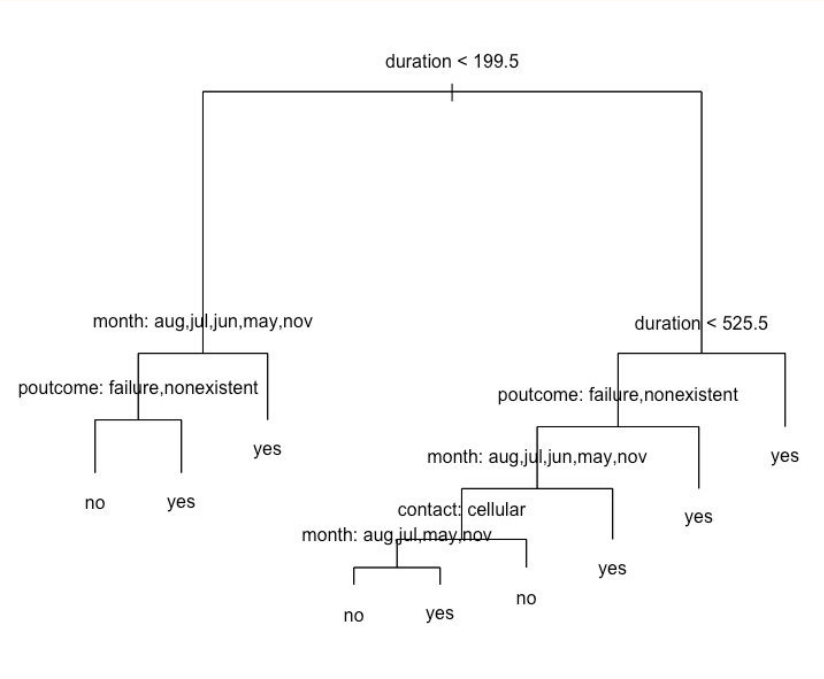


Classification

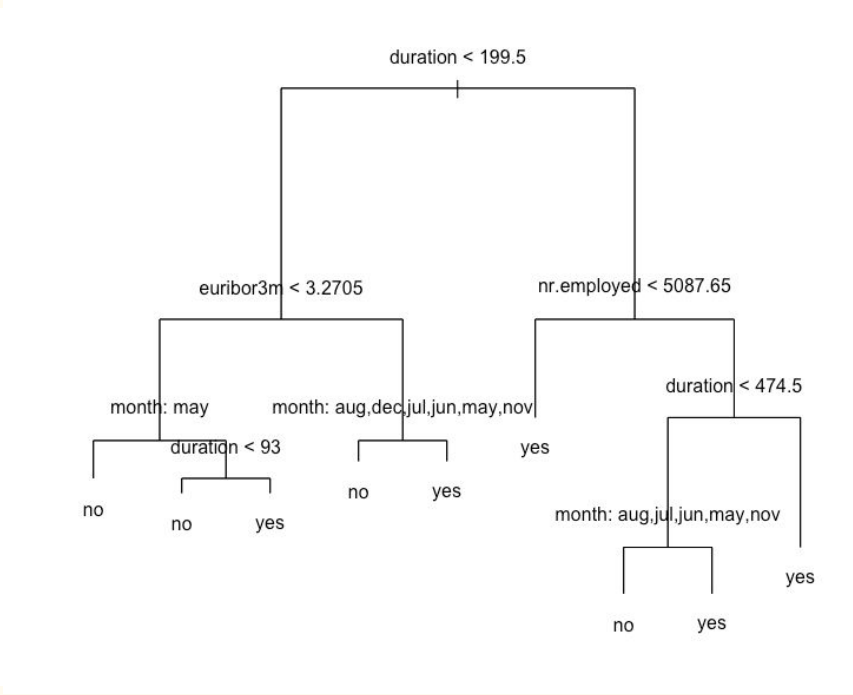
- Final variables: job, education, contact, month, day_of_week, poutcome, duration, campaign, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed
- Methods: Tree, RandomForest, Logistics Regression, SVM, KNN

Classification - Tree

Single Tree



Random Forest



Classification - KNN (Highlight!)

★ Self-defined Distance Matrix Calculation (many categorical!!!)

Distance between two observations:

- For numeric variables: square of Euclidean distance
- For categorical variables: If the two values are the same, set distance = 0;

  If not, set distance = 1

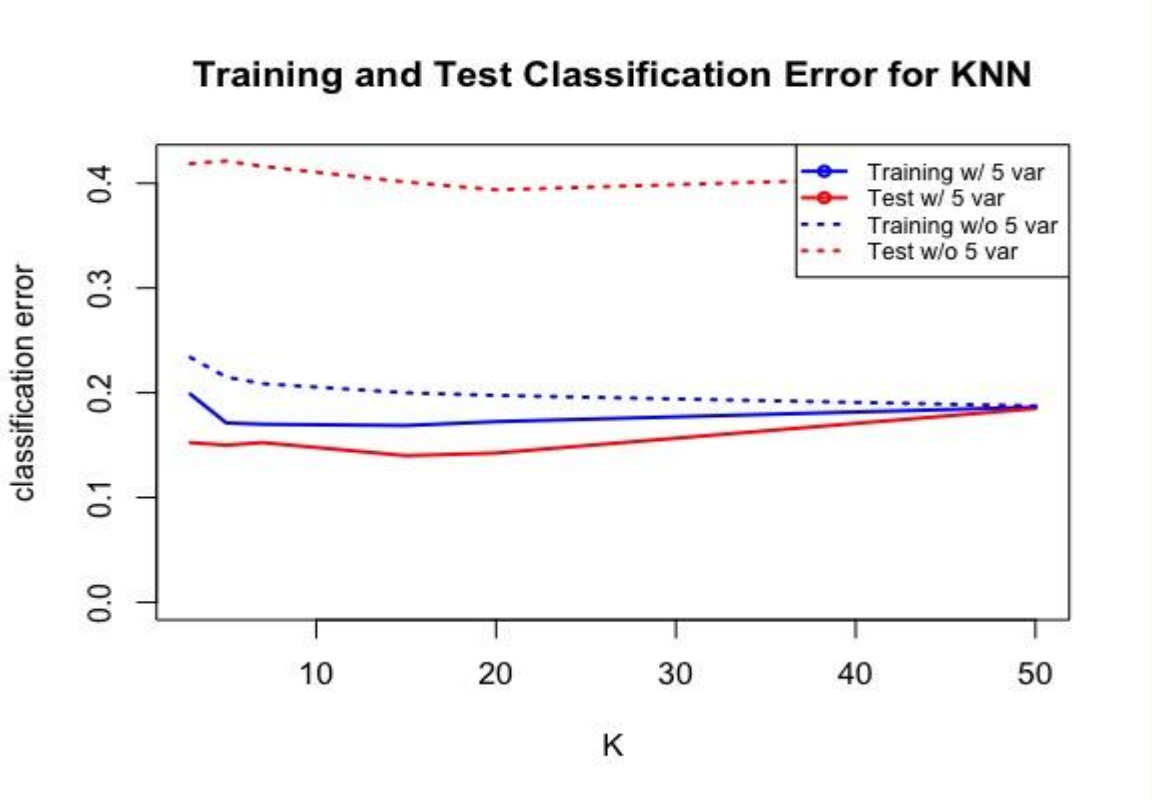
- Take square root of sum (as in Euclidean distance)

Compute pairwise distance matrix (symmetric, with diagonal 0)

Classification - KNN

- Sampled 800 training data and 400 test data
- Write our own KNN function based on the distance matrix (majority vote of the closest k observations)
- Choose k = (3, 5, 7, 15, 20, 50)
- Run KNN twice: with and without the 5 social/economic variables

KNN Errors


Training and Test Classification Error for KNN
Training w/ 5 var
Test w/ 5 var
Training w/o 5 var
Test w/o 5 var
classification error
0.0
0.1
0.2
0.3
0.4
10
20
30
40
50
K

Clustering

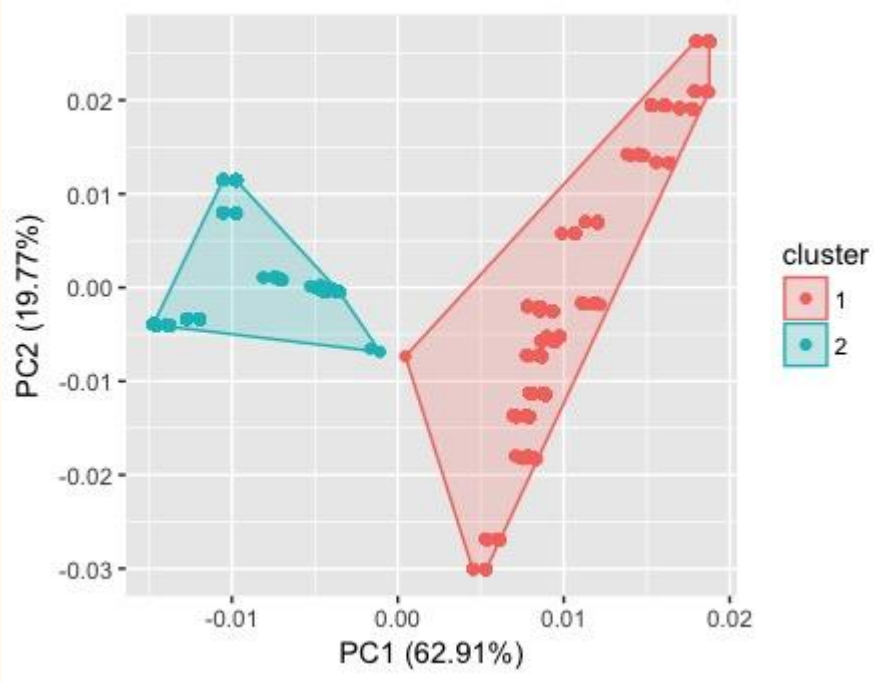
Variables: 5 social and economic attributes

Cluster: PCA and K-means (two cluster & three clusters)

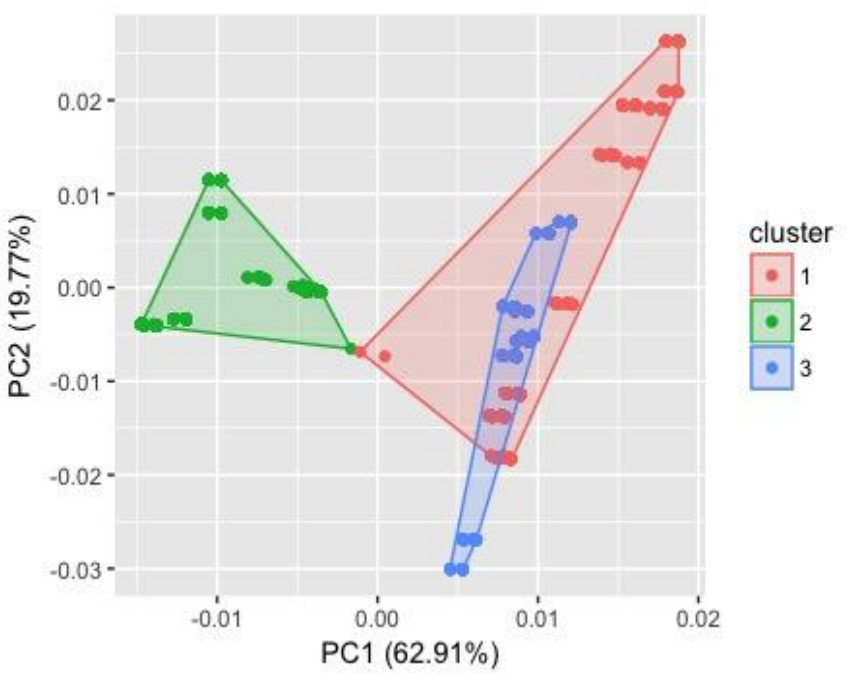
Accuracy: compare the actual cluster with the K-means cluster

Number of clusters	accuracy
2	0.711125
3	0.336

Clustering



Cluster = 2



Cluster = 3

Results

Train and Test Error without five social & economic variables			Train and Test Error with five social & economic variables	
Method	**Training error**	**Test error**	**Training error**	**Test error**
Single Tree	0.1833	0.1578	0.1314	0.1164
Random Forest	0.1564	0.1422	0.1156	0.0992
Logistic Regression	0.1645	0.1508	0.1341	0.1211
SVM (kernel = linear, cost = 10)	0.3048	0.2930	0.1908	0.1953
KNN	0.1975(K=20)	0.3938(K=20)	0.1688(K=15)	0.1400(K=15)

Limitations

Data limitation: almost 50% categorical variables

Interaction: not include in logistic regression model

Data contains a high portion of “unknown” levels, which may create difficulties for interpretation

Implication

- Social and economic attributes have significant impact on clients' decision of deposit subscription and bring improvement to our model performance. For all five predictors, the smaller value will result in clients being more likely to deposit.
- Suggestions for bank on future campaign:

  Make more contacts to clients, especially those with higher degrees

  Prefer cellular call communication type

  Prefer to choose March, April and October to make contact calls

THANKS

FOR WATCHING