Data Mining for Bank Telemarketing

Mengjiao Li, Jie Wang

Umass Amherst

August 18, 2018

Motivation

- Positive response rate to mass campaigns are typically very low. In order to save costs and time, it is important to filter the contacts but keep a certain success rate.
- GOAL: To build a classifier to predict whether or not a client will subscribe a term deposit. Besides, we plan to find out which factors are influential to customers decision, so that a more efficient and precise campaign strategy can be designed to help to reduce the costs and improve the profits.

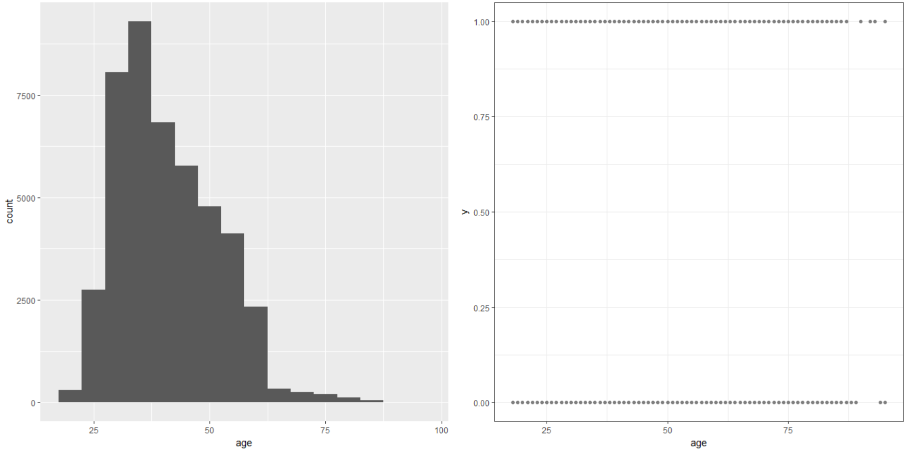
Data Source

Our data were collected from a Portuguese marketing campaign related with bank deposit subscription for 45211 clients and 16 features, and the response is whether the client has subscribed a term deposit. Our data set is downloaded from
http://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

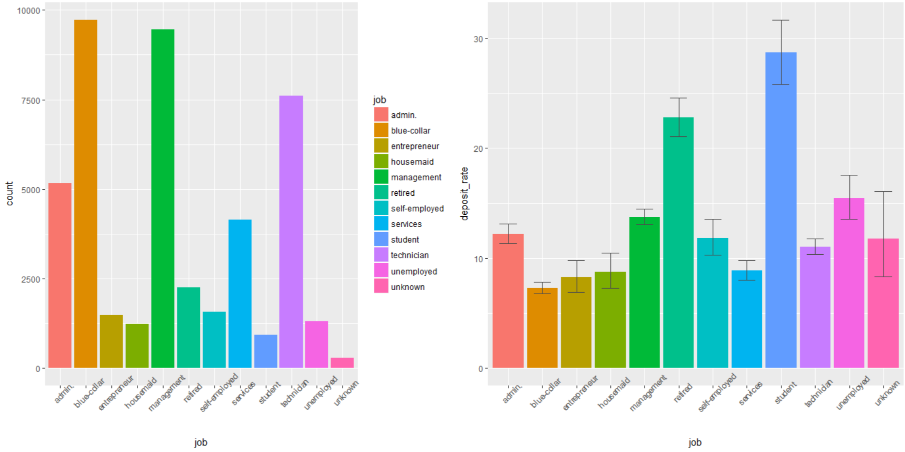
Two kinds of features:

- customer feature:
  - age, yearly balance (continuous var.)
  - education, default, job, marital status, housing loan, personal loan (categorical var.)
- Phone call feature:
  - duration (continuous var.)
  - contact communication type, day, month, campaign, pdays, previous, poutcome(categorical var.)

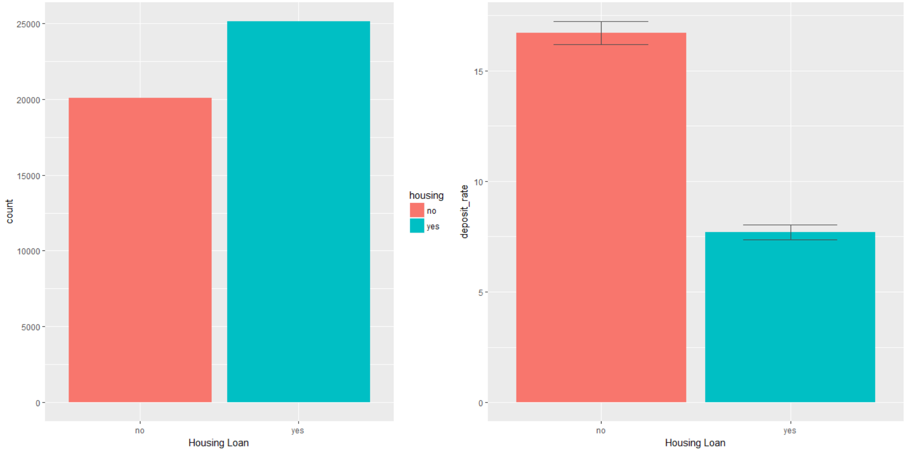
Data Visualization



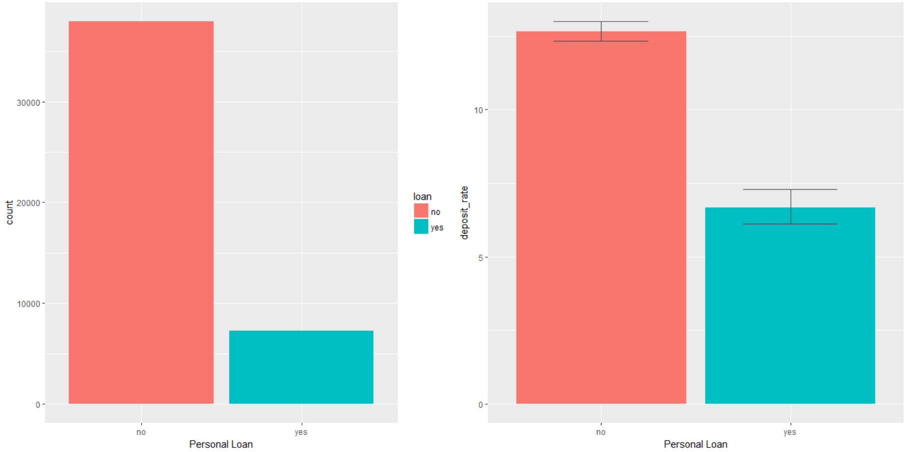
Data Visualization (Cont.)



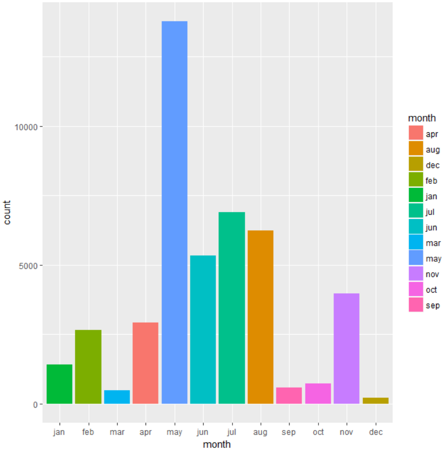
Data Visualization (Cont.)

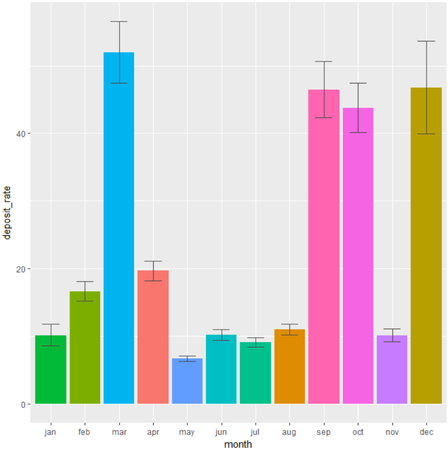


Data Visualization (Cont.)

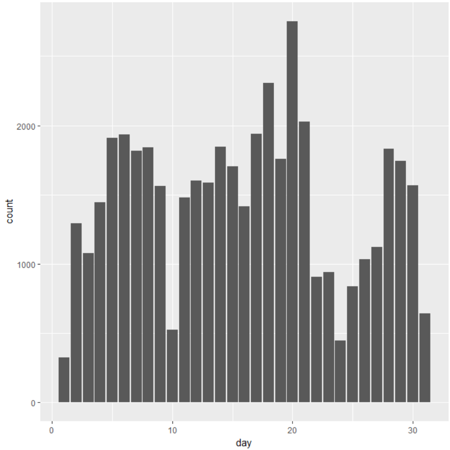


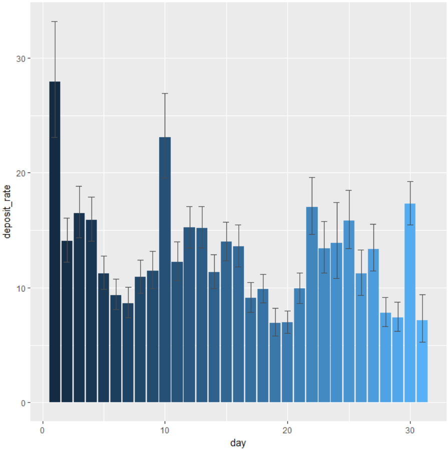
Data Visualization (Cont.)



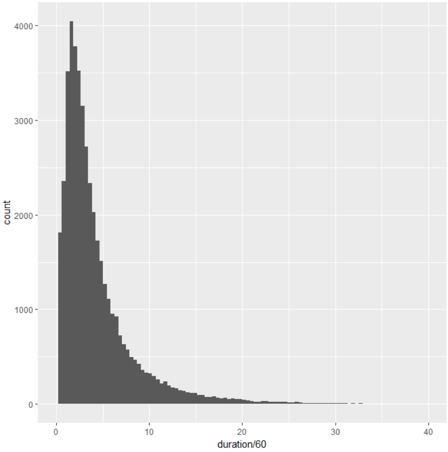


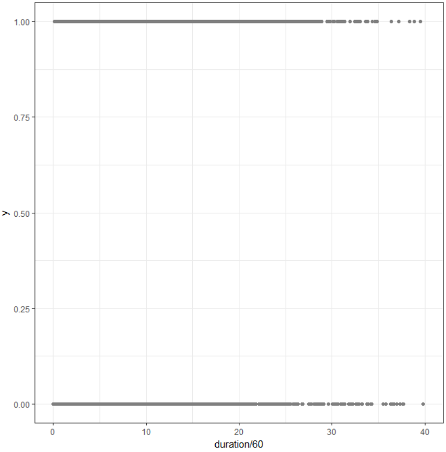
Data Visualization (Cont.)



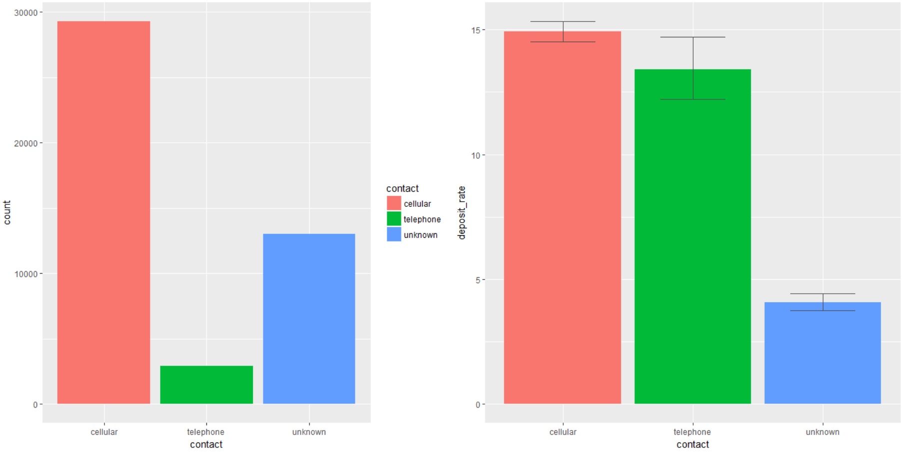


Data Visualization (Cont.)





Data Visualization (Cont.)



Feature Selection

select the most influential features from the original feature set and remove redundant features from your dataset, two ways:

- rank features by some criteria and select the ones that are above a defined threshold: chi-squared, information gain, linear correlation...
- Search for optimum feature subsets from a space of feature subsets: best-first search, back-ward search, forward search, hill climbing search...

R package "Fselector"

- install and load the package, FSelector:
  install.packages("FSelector")
  library(FSelector)
- calculate weights for each attribute using some function
  SOMEFUNCTION(class~., train),
  chi.squared, information.gain, random.forest.importance...

	attr_importance
age	0.0119225202
job	0.0080246244
marital	0.0016934420
education	0.0031712869
default	0.0002202107
balance	0.0054601413
housing	0.0092010547
loan	0.0027664915
contact	0.0145914521
day	0.0032258538
month	0.0253083723
duration	0.0703595860
campaign	0.0039872287
pdays	0.0245426595
previous	0.0120962089
poutcome	0.0293272594

- use the cutoff function to obtain the attributes of the top five weights:

```
subset = cutoff.k(weights, 5)
```

- print the results:

```
f=as.simple.formula(subset, "y")
print(f)
y ~ duration + poutcome + month + pdays + contact
```

Logistic Regression

- Linearity
- $$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$
  $$p = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

Logistic Regression

- Linearity
- $$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$
  $$p = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

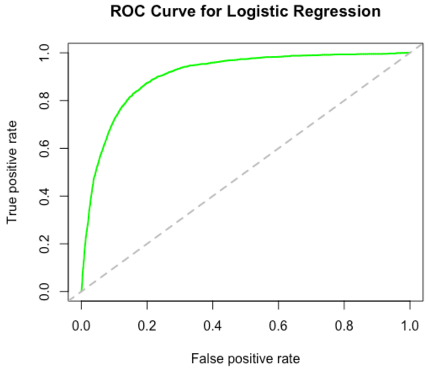
glm(f, family=binomial(link='logit'), data=train)
f: y $\sim$ duration + poutcome + month + pday +contact

Prediction

predict(model, newdata=test, type="response")

Prediction

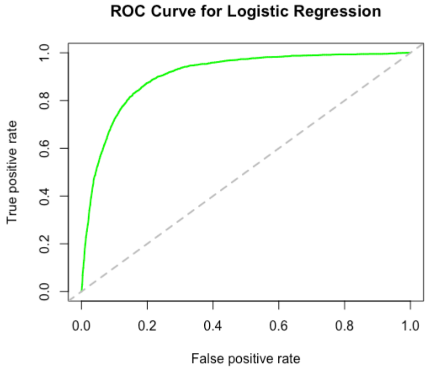
predict(model, newdata=test, type="response")



AUC=0.8976556

Prediction

predict(model, newdata=test, type="response")



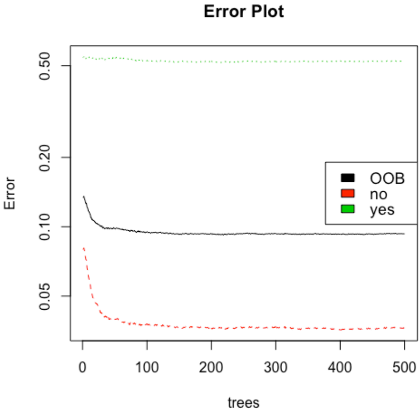
AUC=0.8976556 versus **AUC(full)=0.9086456**

Random Forest

- machine learning technique, nonlinear
- number of decision trees are built during the process
- reduce chances of over-fitting

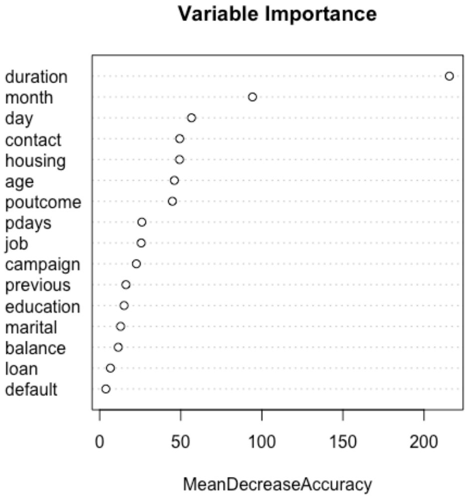
R package "randomForest"

model=randomForest(y ., data=train, importance=TRUE, ntree=500)
plot(model)

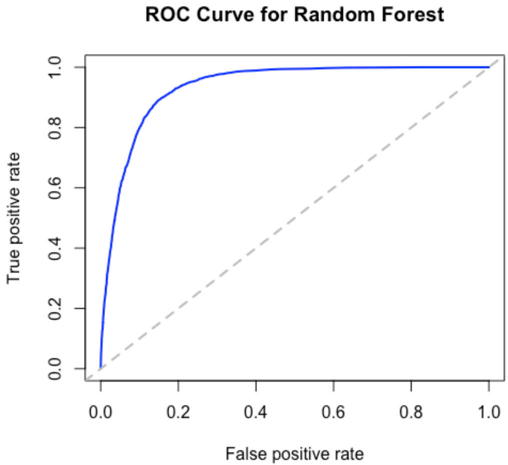


Variable Importance

varImpPlot(model, main="Variable Importance", type=1)



Prediction



AUC= 0.934956

Conclusion

- Used ggplot2 to visualize and analyze the dataset
- Introduced a R package "Fselector" for feature selection and select the most influential factors for customers decision
- Applied the classification methods logistic regression and randome forest to predict the success of bank telemarketing