Using machine learning to predict default risk of credit card

Yiying Zhang (524900)
Jingmin Zhang(578516)

Wirtschaftswissenschaftliche Fakultät
Humboldt–Universität zu Berlin

Outline

1. Introduction ✓
2. Data pre-processing
3. Prediction models
4. Model performance
5. Conclusion

Dataset Information

- Motivation: From the perspective of risk management, the result of estimated probability of default will be very valuable.
- The dataset is related with customer default payment of credit card in Taiwan, which includes 23 variables as explanatory variables, and 1 binary variable as the response variable.
- 30000 observations.
- *UCI Machine Learning Repository*

Dataset Information(2)

- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).

Dataset Information(3)

- From April to September, 2015
- X6 - X11: History of past payment.
  - -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- X12-X17: Amount of bill statement (NT dollar).
- X18-X23: Amount of previous payment (NT dollar).
- Y: Default Payment(yes=1,no=0)

Outline

1. Introduction
2. Data pre-processing ✓
3. Prediction models
4. Model performance
5. Conclusion

Handling outliers

- Category variables: Education and Marriage
  - Method: Use mode to replace abnormal value
- Continuous variables: Amount of *billstatement* and *payment* in previous six months.
  - Method: Clustering use euclidean distance.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

- delete observations whose distance is larger than 98 quantile, in total are 981 enries.

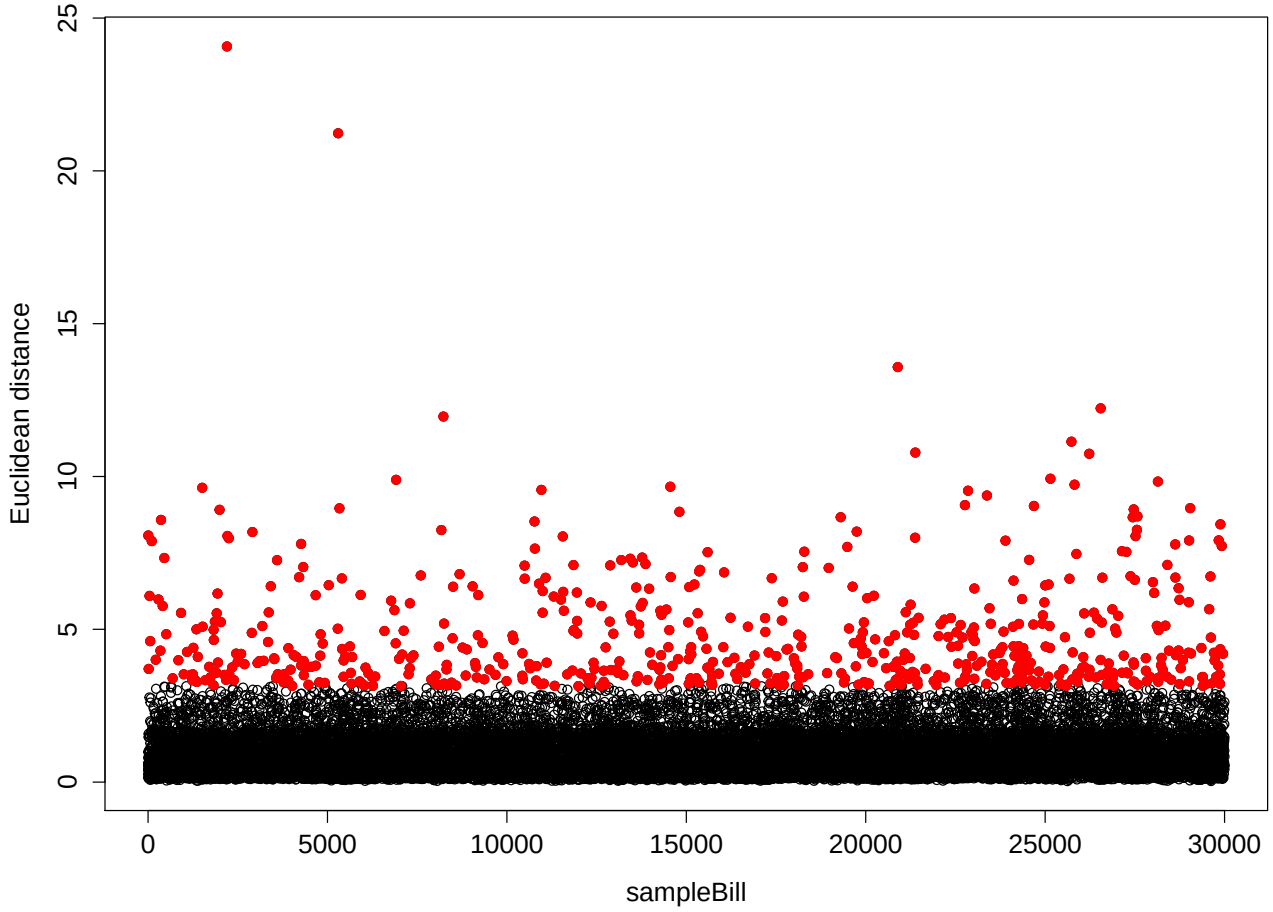
Code

```
DataBill=scale(data.frame(df$BILL_AMT1,df$BILL_AMT2,
  df$BILL_AMT3,df$BILL_AMT4,df$BILL_AMT5,df$BILL_
  AMT6))
set.seed(123)
km=kmeans(DataBill,center=3)
km$centers
c1=matrix(km$centers[1,],nrow=30000, ncol =6 , byrow
   = T)
dist1=sqrt(rowSums((DataBill-c1)^2))
c2=matrix(km$centers[2,], nrow=30000, ncol =6 ,
  byrow = T)
dist2=sqrt(rowSums((DataBill-c2)^2))
c3=matrix(km$centers[3,], nrow=30000, ncol =6 ,
  byrow = T)
dist3=sqrt(rowSums((DataBill-c3)^2))
dist.Bill=data.frame(dist1,dist2,dist3)
```

Code

```
y=apply(dist.Bill, 1, min)
upper.limit = quantile(y,.98)
plot(1:30000,y,xlim=c(0,30000),xlab="sampleBill",
  ylab="Euclidean distance")
points(which(y>upper.limit),y[which(y>upper.limit)],
  pch=19,col="red")
df[which(y > upper.limit),13:18] = NA
```

Outliers in Billstatement



Code

```
1 library(corrplot)
2 loans_numeric<-sapply(clean.df,is.numeric)##only to
    numeric variables!
3 correlation<-cor(clean.df[,loans_numeric])
4 corrplot(correlation,type = "upper"
5           ,tl.pos = "d",tl.col = "black",tl.cex =
              0.6)
6 corrplot(correlation,add = TRUE,type = "lower",
7           method = "number",diag=FALSE
8      ,tl.pos = "n",
9      cl.pos = "n",cl.ratio= 0.2,addCoefasPercent =
        TRUE)
```

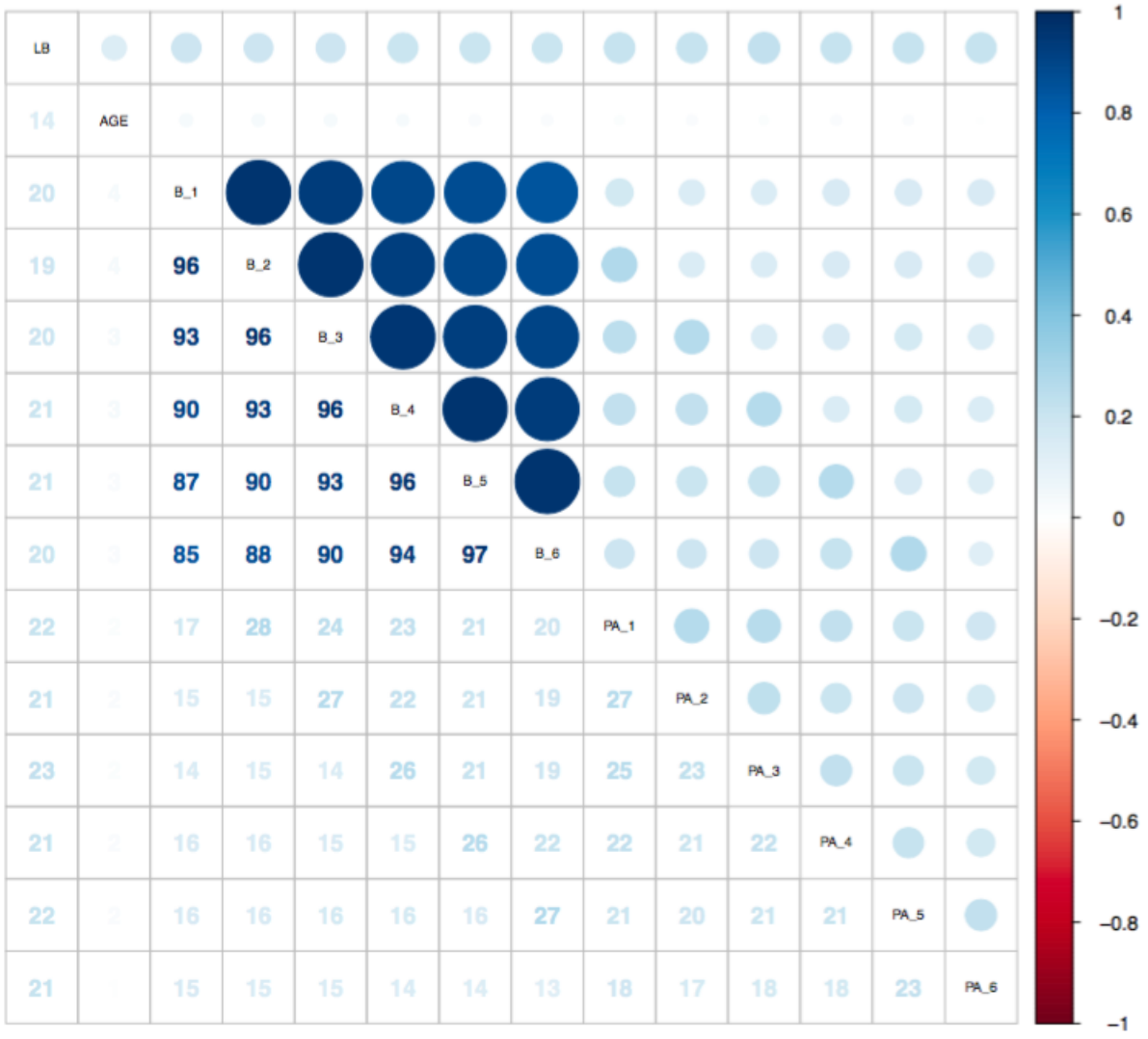


Figure 1: Correlation Table

Outline

1. Introduction
2. Data pre-processing
3. Prediction models ✓
4. Model performance
5. Conclusion

Logit Regression Model

- The Linear Regression Model is a basic model in our research, because it is fast to implement and more easily interpretable than other models.
- With the Linear Regression Model, we can get the output between 0 to 1, which can be interpreted as a probability.

Code

```
train_lr[,(13:17)]<-NULL
test_lr[,(13:17)]<-NULL
lr<-glm(default~.,data=train_lr,family = binomial(
  link="logit"))
pred.lr <- predict(lr, newdata = test_lr,type = "
  response")
summary(lr)
```

- y=-1.175-0.14sex.female-1.05edu.others
  -0.18mar.single+0.45sep.pay.duly
  +0.79sep.delay.one+2.01sep.delay.two
  +1.97sep.delay.three+1.93sep.delay.four
  -0.21Aug.pay.duly-0.68Aug.delay.four+
  1.18sep.delay.five+0.4July.delay.two
  +0.51July.delay.three-0.22May.pay.duly +0.65April.delay.three

Naive Bayes Model

- Naive Bayes is a simple classifier based on Bayes theorem.
- It works on the assumption that the values of all variables are independent.

Code

```
library("e1071")
df_nb<-naiveBayes(train$default~.,data=train)
pred.df<-predict(df_nb,newdata = test,type = "raw")
test$pred.df<-pred.df[,2]
```

Neural Network

Neural Network mimic the functioning of the brain!

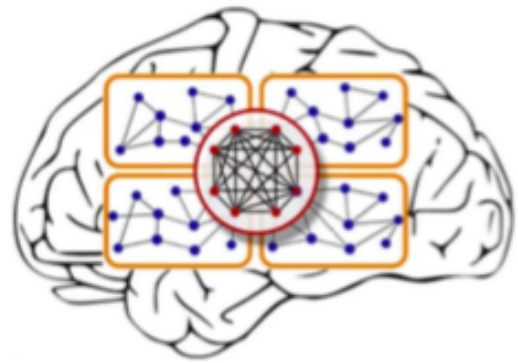


Figure 2: The human brain as a network of many cells called neurons

Neural Network

- pocess summarised as following:
  - Receives input signals(variable values)
  - Aggregates input signals
  - Non-linear transformation
  - sends output signal(results)

Code

```
1 model.control<- trainControl(method = "cv",  number
    = 5, classProbs = TRUE,summaryFunction =
    twoClassSummary, allowParallel = TRUE, returnData
    = FALSE)
2 nn.parms <- expand.grid(decay = c(0, 10^seq(-3, 0,
    1)), size = seq(3,15,2))
3 nn <- caret::train(default~., data = train.nn,
4                    method = "nnet", maxit = 200,
                       trace = FALSE,
5                    tuneGrid = nn.parms,
6                    metric = "ROC", trControl = model
                       .control)
```

Decision Trees

Decision trees model is one of the most basic machine learning algorithms used for classification and regression

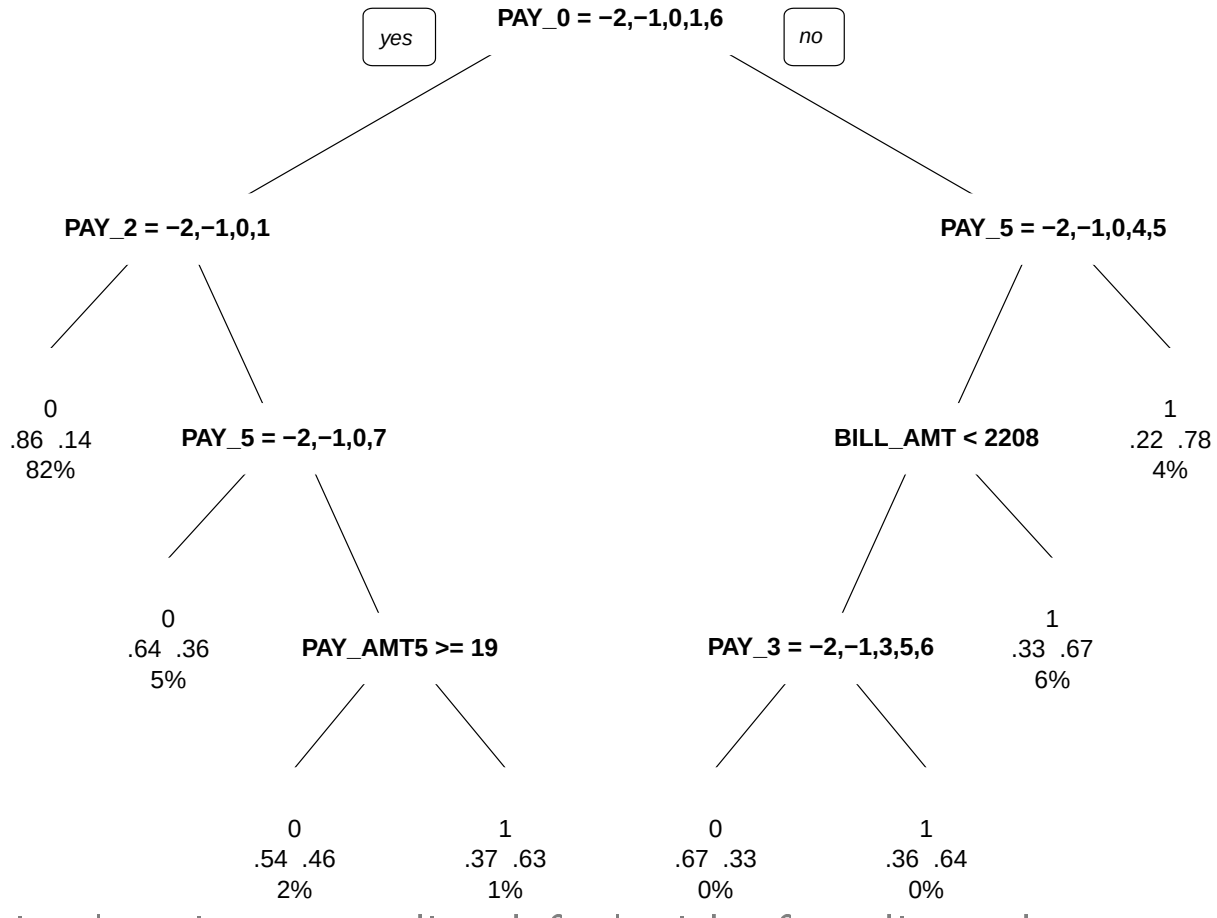
- Classification Trees
  - Trees work by spilting the data into smaller, more homogeneous groups,i.e minimize impurity.
  - Each node of trees denotes a test on an attribute, and look for the variable which most increases homogeneity in the resulting partitions

Code

```
#creat the train and test set with 80% and 20%
idx.train <- createDataPartition(y = clean.df$
  default.payment.next.month, p = 0.8, list = FALSE)
clean.df.train <- clean.df[idx.train, ]
clean.df.test <-  clean.df[-idx.train, ]
```

```
#pre-pruning the tree and build model#
rpart.control=rpart.control(minsplit = 6, minbucket
  = 6,cp=0.001, xval = 5, maxdepth = 6)
dt<-rpart(default.payment.next.month ~ ., data =
  clean.df.train, method="class", control = rpart.
  control)
##predict the test set
pred.dt <- predict(dt, newdata = clean.df.test, type
   = "prob")[, 2]
```

Visualization of Decision Trees model



Random Forest

The random forest is a tree-based classifier and is very popular in many applications

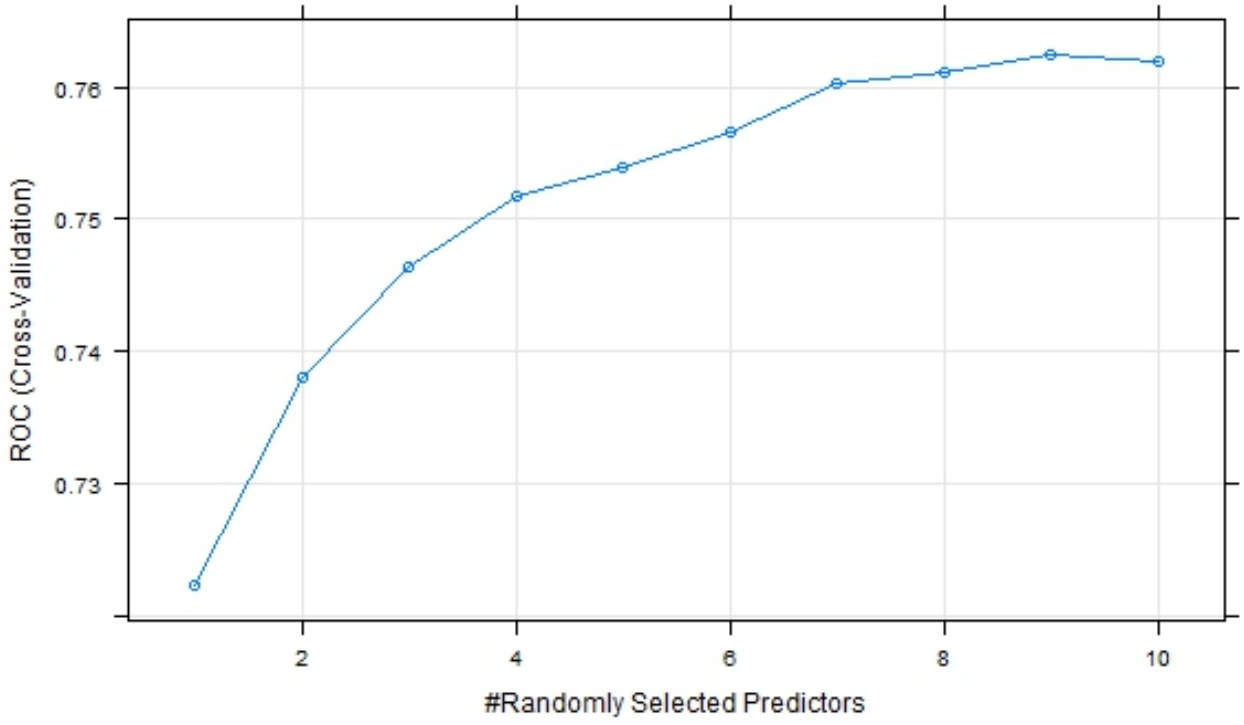
- Instead of one big tree, Random forest has many small trees.
- With less correlation between each tree, the model will perform better.
- Use package"caret" to develop the model, the evaluation criteria is Gini index.

Code

```
rf.parms <- expand.grid(mtry = 1:10)
rf.caret <- train(default.payment.next.month~., data
   = clean.df.train, importance = T,
                     method = "rf", ntree = 500,
                       tuneGrid = rf.parms,
                     metric = "ROC", trControl = model.
                       control)
```

```
###outcome of random forest###
            Reference
Prediction good default
   good     4281     805
   default   225     492
   Accuracy : 0.8225       AUC : 0.7701
```

ROC with different "mtry"



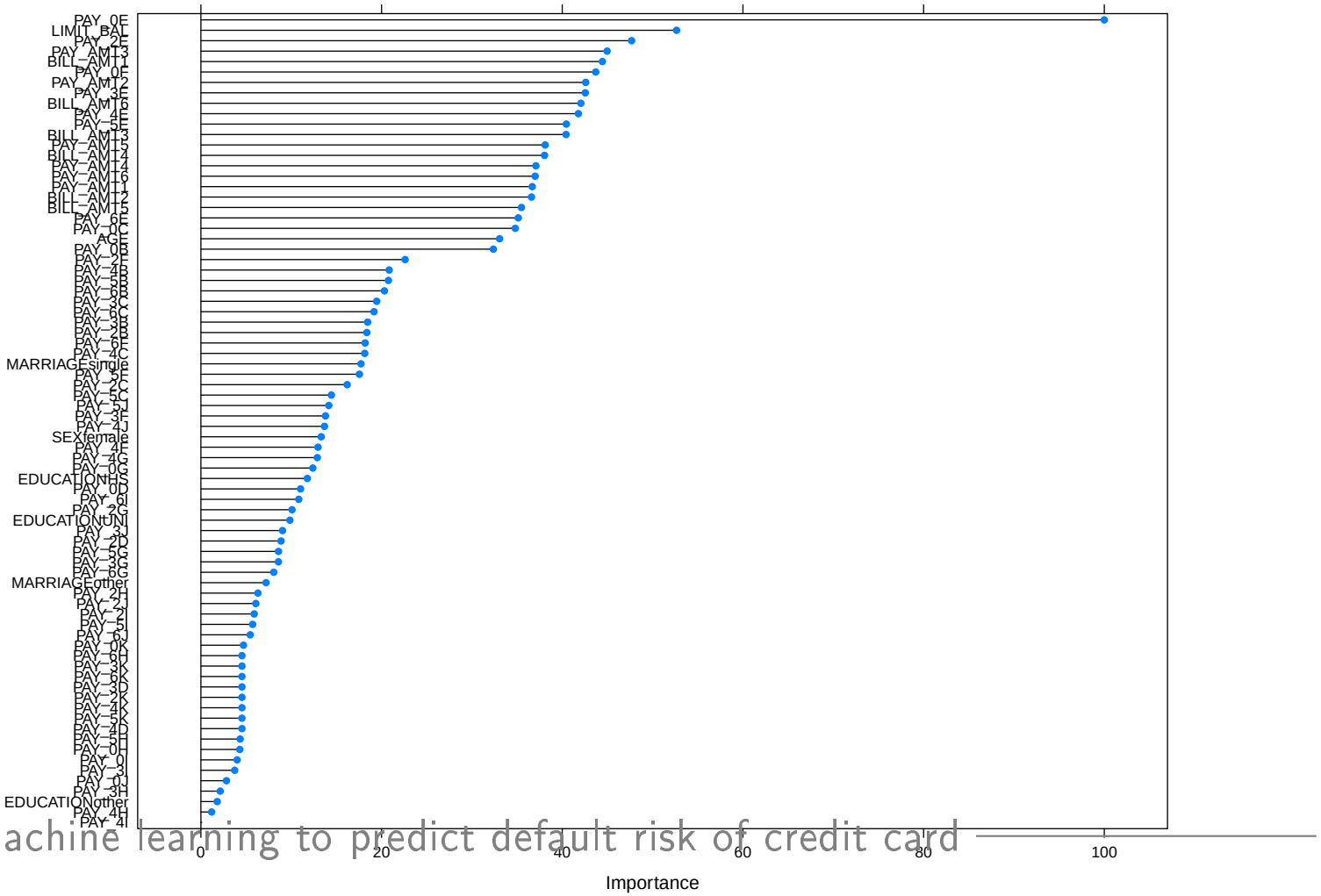
Variable importance in Random forest

- Out-of-bag importance: The decrease in accuracy when randomly change the values of each variable in turn for each tree, averaged over all trees.

PAY0.2	LIMIT.BAL	PAY2.2	PAY.AMT3	BILL.AMT1
40.46	20.39	18.29	17.13	16.92

Table 1: Importance of top 5 variables

Plot of Variable importance



Outline

1. Introduction
2. Data pre-processing
3. Prediction models
4. Model performance ✓
5. Conclusion

ROC curves of LR and NN

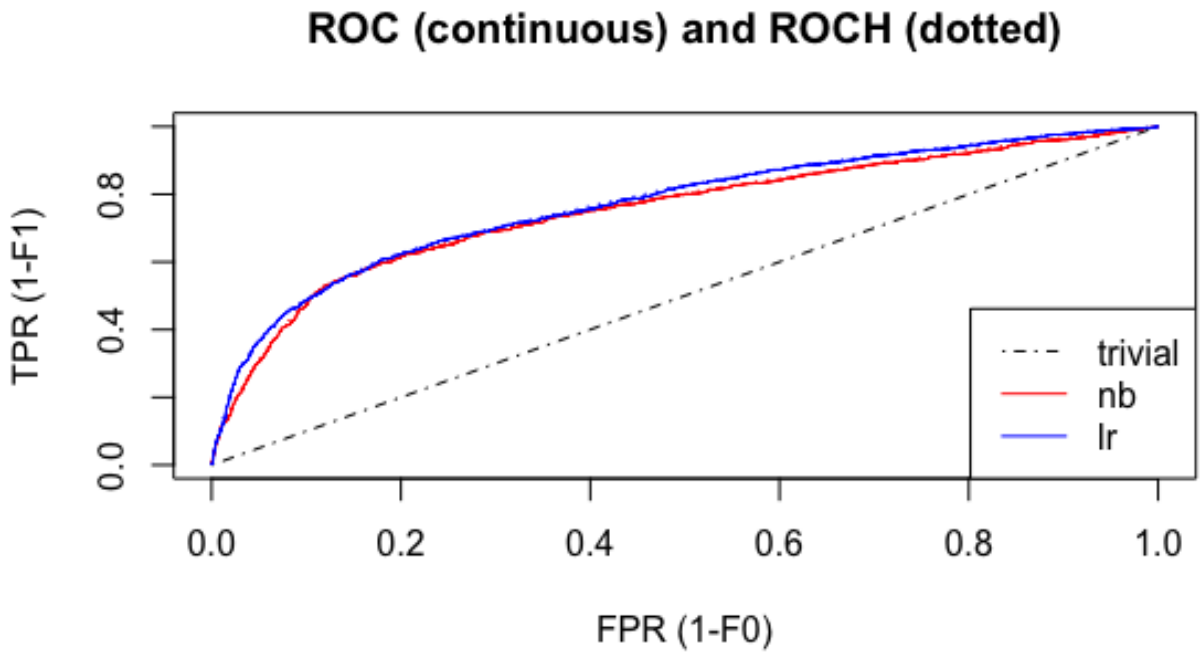


Figure 3: ROC of LR and NN

ROC curves of DT and RF

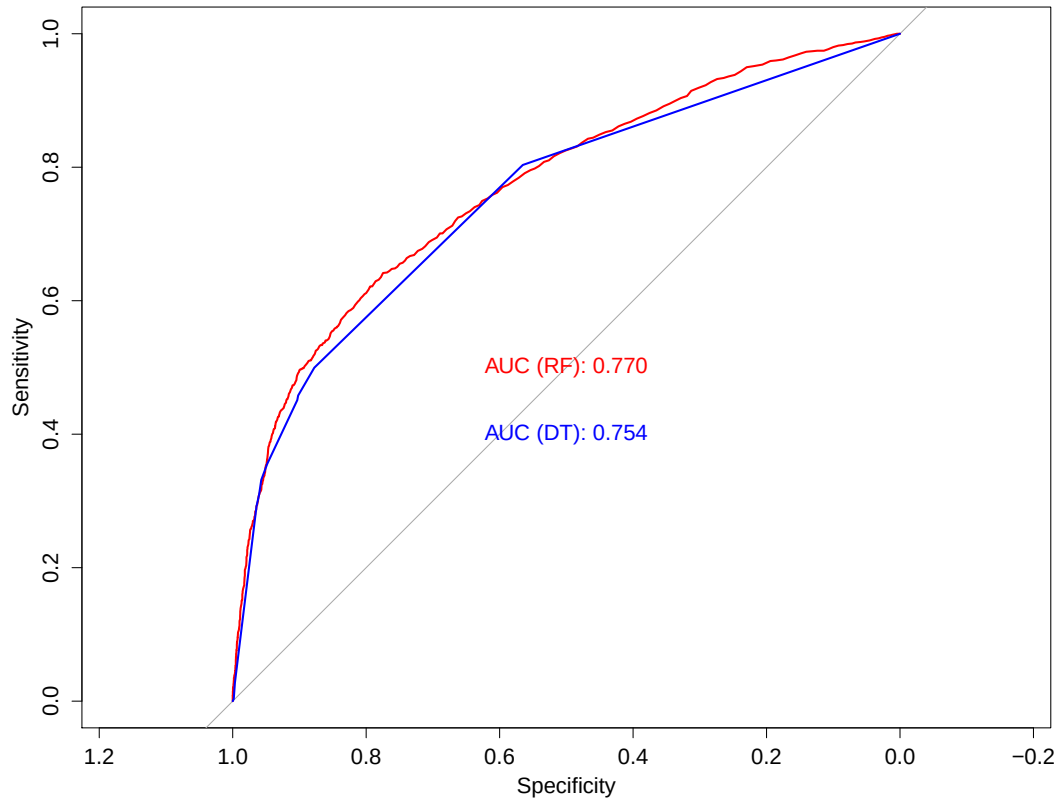


Figure 4: ROC of RF and DT

Results of all models

	LR	NB	NN	RF	DT
Acc	0.8199	0.6948	0.7139	0.8225	0.8179
AUC	0.7734	0.5732	0.7852	0.7701	0.7582

Table 2: AUC and Accuracy results

Outline

1. Introduction
2. Data pre-processing
3. Prediction models
4. Model performance
5. Conclusion ✓

Conclusion

- We conduct 5 different classification models of machine learning to predict the default risk of credit card and evaluate the performance using AUC and accuracy value.
- Apart from NB, the AUC of other 4 models is similar, all above 0.5 (random assign),Neural Network is the best model.
- Variables have different influence on different models.(e.g recent payment-situation is most important to random forest model, but is not for logit model)

Gihub Page:

SPL/R-programming(not the final version)
`https://github.com/Jingmin24/R-programming.git`