Kaggle 案例分析

- Santander Customer Satisfaction

顏家佑、沈祐珍、黃侲艗

Kaggle 案例分析

Santander Customer Satisfaction

01. 案例簡介

02. Script介紹

03. Xgboost應用

案例簡介

Santander Customer Satisfaction

- Which customers are happy customers?

Santander

Santander Customer Satisfaction

2016由桑坦德銀行發起的Kaggle挑戰
找出對銀行服務很滿意與不滿意的顧客
第一名獎金：30,000美元

案例簡介

Santander Customer Satisfaction

- 資料簡介

	ID	var3	var15	imp_ent_var16_ult1	imp_op_var39_comer_ult1	saldo_medio_var44_ult3	var38	TARGET
1	1	2	23	0	0	0	39205.17	0
2	3	2	34	0	0	0	49278.03	0
3	4	2	23	0	0	0	67333.77	0
4	8	2	37	0	195	0	64007.97	0
5	10	2	39	0	0	0	117310.98	0
6	13	2	23	0	0	0	87975.75	0

The "TARGET" column is the variable to predict.
It equals 1 for unsatisfied customers and 0 for satisfied customers.

案例簡介

Santander Customer Satisfaction

- 資料簡介

Training data 訓練資料

Used to build the model

Testing data 驗證資料

Assess the quality of the model

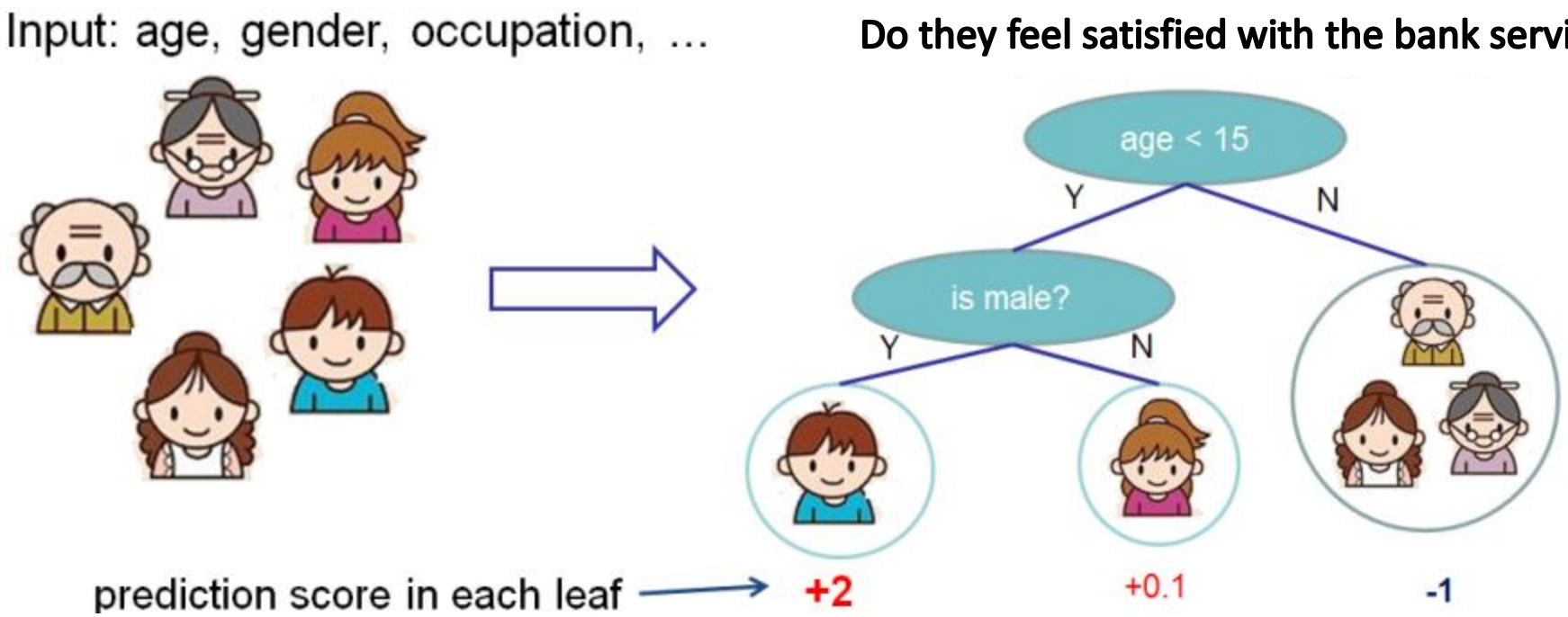
Predict the TARGET variable using the XGBOOST model

案例簡介

XGBoost

- eXtreme Gradient Boosting (極限梯度提升)

利用分類和回歸樹(Classification and regression tree, CART)。

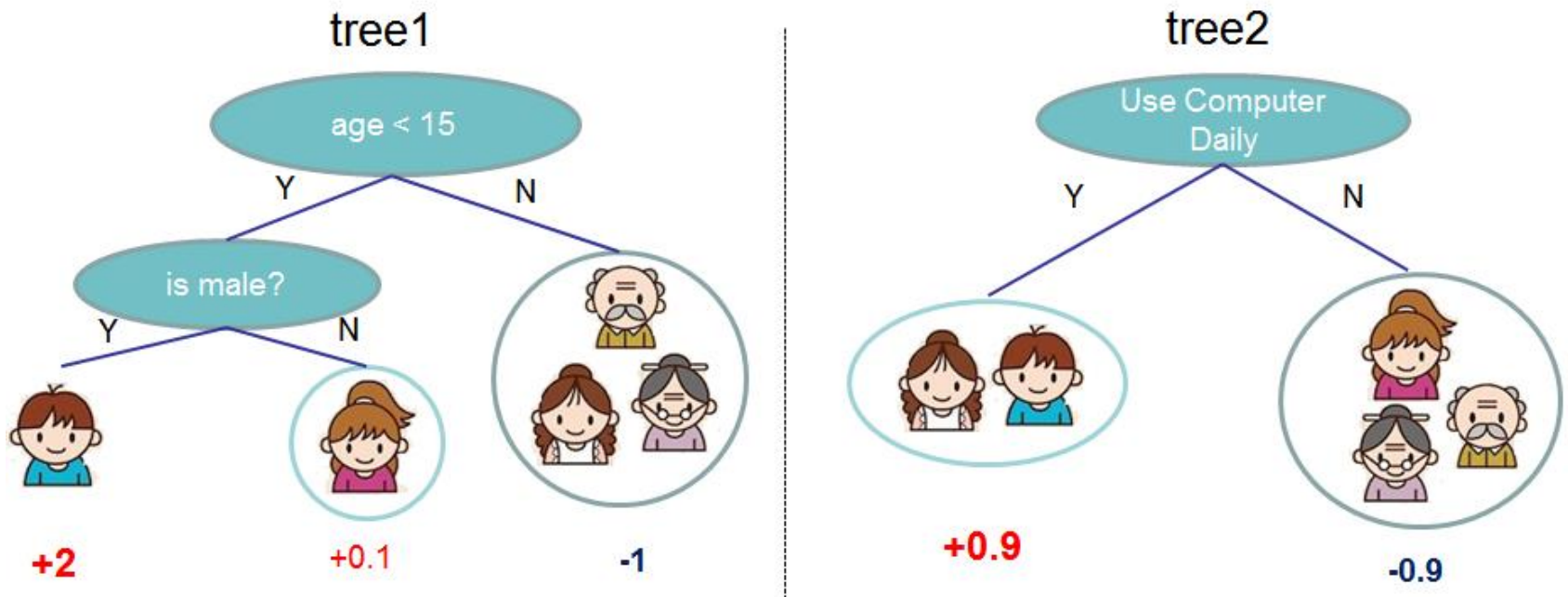


案例簡介

XGBoost

- eXtreme Gradient Boosting (極限梯度提升)

利用分類和回歸樹(Classification and regression tree, CART)。



f( ) = 2 + 0.9= 2.9 f( )= -1 - 0.9 = -1.9

案例簡介

XGBoost

- eXtreme Gradient Boosting (極限梯度提升)

利用分類和回歸樹(Classification and regression tree, CART)。




01
Train model

02
Perform prediction

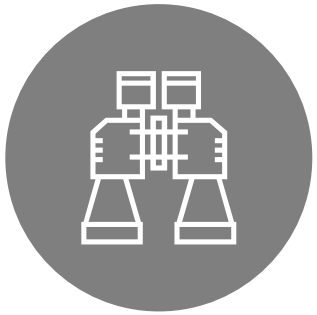
03
Transform regression result into binary classification

04
Evaluate model performance

案例簡介

XGBoost

- eXtreme Gradient Boosting (極限梯度提升)



training

Basic : xgboost
Advanced: xgb.train

data

Sparse matrix
Dense matrix
Xgb.DMatrix

booster

nonlinear link: “gbtree”
linear link: “gblinear”

01 Train model

02 Perform prediction

03 Transform regression result into binary classification

04 Evaluate model performance

案例簡介

XGBoost

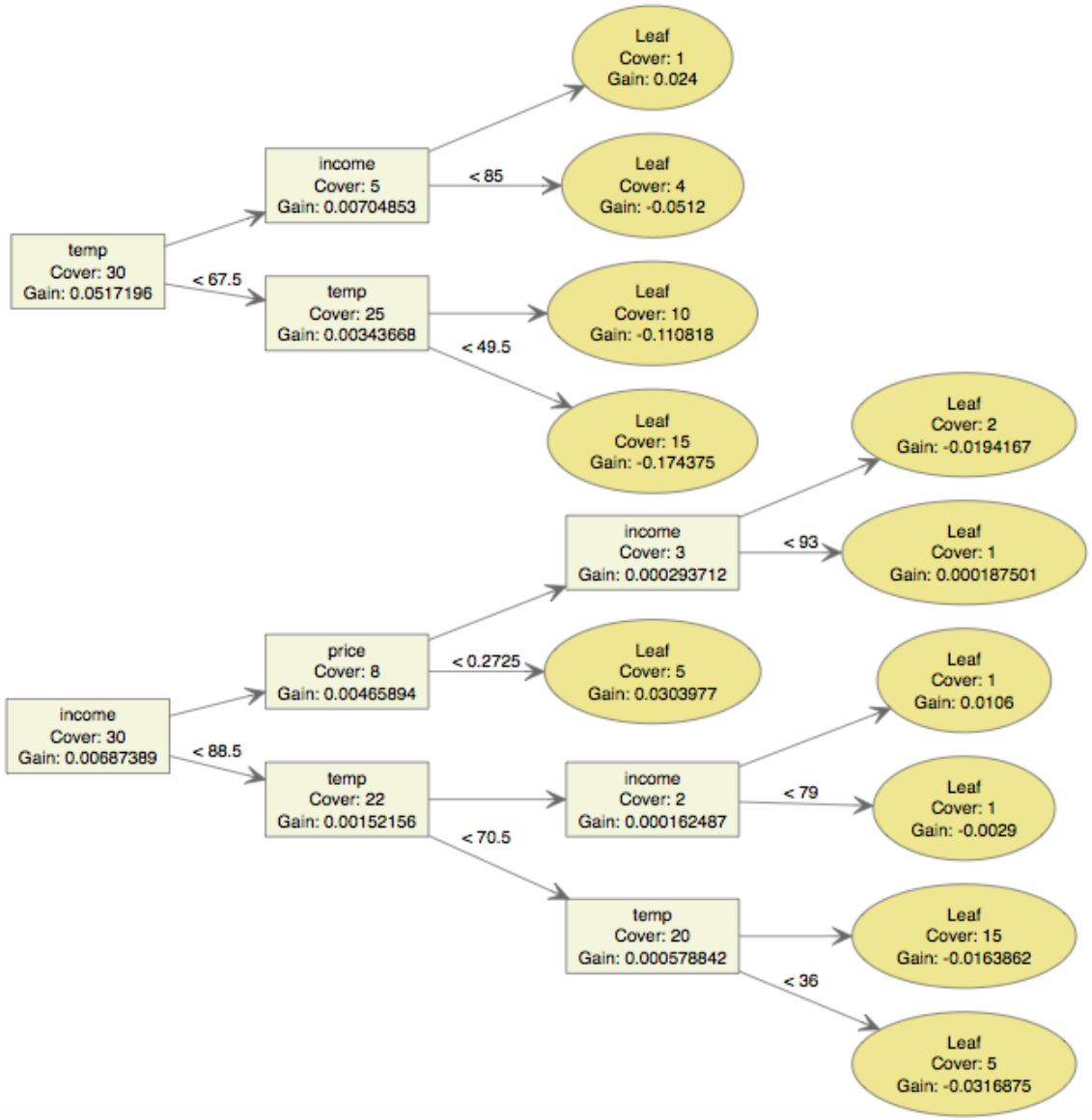
- eXtreme Gradient Boosting (極限梯度提

objective	regression/classific binary:logis
nrounds	the max number o (the number of dec
max_depth	maximum depth o
eta	step size of each b
nthread	number of cpu thre

01
Train model

0
Perform

```
bst <- xgboost(data = train.data, label = Icecream$cons, max.depth = 3, eta = 1,
nthread = 2, nround = 2, objective = "reg:linear")
```



Kaggle 案例分析

Santander Customer Satisfaction

01. 案例簡介

02. Script介紹

https://rgmmmt4r.github.io/106-2_R_b04303117/project2/kaggle_group6_01copy.html

03. Xgboost應用

Xgboost應用

Xgboost
cForest
比賽誰預測鐵達尼號生存比較準

01. 資料來源

- https://www.kaggle.com/c/titanic/data
- Titanic: Machine Learning from Disaster

02. Xgboost

- Xgboost號稱kaggle競賽神器
- 那將Xgboost與當時練習鐵達尼號用的cForest來比較

Xgboost應用

Xgboost
cForest
比賽誰預測鐵達尼號生存比較準

03. 變數

- Pclass：艙等
- Sex：性別
- SibSp：兄弟姊妹、老婆丈夫數量
- Parch：父母小孩的數量
- Embarked：出發港口
- Title：利用NAME找出Miss、Mrs
- FsizeD：依據家庭大小分類(1,4,5)
- isAlone：是否只有一個人
- Fsize：家庭大小

04. Xgboost

- 將資料轉成sparse matrix才能跑

```
train = data[data$PassengerId %in% df_train$PassengerId,]
y_train <- train[!is.na(Survived),Survived]
train = train[,Survived:=NULL]
train = train[,PassengerId:=NULL]
train_sparse <- data.matrix(train)
```

Xgboost應用

Xgboost
cForest
比賽誰預測鐵達尼號生存比較準

05. 設定hyperparameters

```
param <- list(objective   = "binary:logistic",
              eval_metric = "error",
              max_depth   = 7,
              eta         = 0.1,
              gammma      = 1,
              colsample_bytree = 0.5,
              min_child_weight = 1)

set.seed(1234)

# Pass in our hyperparameteres and train the model
system.time(xgb <- xgboost(params  = param,
                           data    = dtrain,
                           label   = y_train,
                           nrounds = 500,
                           print_every_n = 100,
                           verbose = 1))
```

06. 預測模型，並利用原本的train資料測試準確率

- 準確率：0.863

```
pred <- predict(xgb, dtest)
pred.resp <- ifelse(pred >= 0.5, 1, 0)

pred2 <- predict(xgb, dtrain)
pred2.resp <- ifelse(pred2 >= 0.5, 1, 0)
success <- 0
for(i in c(1:891)){
  if(pred2.resp[i] == data$Survived[i]){
    success <- success + 1
  }
}
print(success / 1000)
```

Xgboost應用

Xgboost
cForest
比賽誰預測鐵達尼號生存比較準

07. 用同一份資料與變數換cForest預測

- 將資料轉成factor

```
data$Pclass <- factor(data$Pclass)
data$Sex <- factor(data$SibSp)
data$Parch <- factor(data$Parch)
data$Embarked <- factor(data$Embarked)
data$Title <- factor(data$Title)
data$FsizeD <- factor(data$FsizeD)
data$isAlone <- factor(data$isAlone)
data$Fsize <- factor(data$Fsize)
```

08. 建立cForest模型

```
train <- data[1:891,]
test <- data[892:1309,]
set.seed(102)
model <- cforest(factor(Survived) ~ Pclass + Sex + SibSp + Parch + Embarked + Title + FsizeD + isAlone + Fsize, data = trai
n)
```

Xgboost應用

Xgboost
cForest
比賽誰預測鐵達尼號生存比較準

09. 預測模型，並利用原本的train資料測試準確率

- 準確率：0.723

```
predict1 <- predict(model, test,  OOB=TRUE, type = "response")
predict2 <- predict(model, train,  OOB=TRUE, type = "response")
success <- 0
for(i in c(1:891)){
  if(predict2[i] == data$Survived[i]){
    success <- success + 1
  }
}
print(success / 1000)
```

```
## [1] 0.723
```

10. 結論

- Xgboost的準確率有0.863，而cForest只有0.723
- Xgboost效率高、準確率高

•Xgboost真的比較難！


Thank you