Ass2

Wangzhihui Mei, HongYi Huang, ChangXu, Zijia He

November 2020

1 Task1

1.1 data describe

Cardiovascular diseases (CVDs) is the leading cause of mortality in India. Ischemic heart disease and stroke are the predominant causes and are responsible for nearly 80% of CVD deaths.

```
dt <- read.csv("heart1.csv",na.strings = "?")
dt <- na.omit(dt) # handle NA
head(dt)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
6	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1

This database contains 14 attributes. The ”target” field refers to the presence of CVD in the patient. It is integer valued 0 (no presence) or 1 (presence). Next we look in detail at the data characteristics of each attribute.
Age: Age in years
Sex: (1 = male; 0 = female)
CP:Chest pain type(1-typical angina, 2-atypical angina, 3-non-anginal pain, 4-asymptomatic)
trestbps:Resting blood pressure (in mm Hg on admission to the hospital)
Chol:Serum cholestoral in mg/dl
Fbs:Indicator of whether fasting blood sugar¿120 mg/dl (1-true; 0-false)
restecg:Resting electrocardiographic results
exang:Exercise induced angina (1-yes; 0-no)
oldpeak:ST depression induced by exercise relative to rest
slope:Slope of the peak exercise ST segment (1-upsloping, 2-flat, 3-downsloping)
ca:Number of major vessels (0-3) colored by flourosopy
thal:Summary of heart condition(3 = normal, 6 = fixed defect, 7=reversable defect)

target:the “The Disease Diagnosis” field refers to the presence of heart disease in the patient(0-No presence,1-Presence)

```
summary(dt)
```

```
     age             sex      cp         trestbps          chol         fbs      restecg     thalach        exang       oldpeak    
 Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258   0:147   Min.   : 71.0   0:204   Min.   :0.00  
 1st Qu.:47.50   1:207   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45   1:152   1st Qu.:133.5   1: 99   1st Qu.:0.00  
 Median :55.00           2: 87   Median :130.0   Median :240.0           2:  4   Median :153.0           Median :0.80  
 Mean   :54.37           3: 23   Mean   :131.6   Mean   :246.3                   Mean   :149.6           Mean   :1.04  
 3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:274.5                   3rd Qu.:166.0           3rd Qu.:1.60  
 Max.   :77.00                   Max.   :200.0   Max.   :564.0                   Max.   :202.0           Max.   :6.20  
 slope         ca          thal        target      
 0: 21   Min.   :0.0000   0:  2   Min.   :0.0000  
 1:140   1st Qu.:0.0000   1: 18   1st Qu.:0.0000  
 2:142   Median :0.0000   2:166   Median :1.0000  
         Mean   :0.7294   3:117   Mean   :0.5446  
         3rd Qu.:1.0000           3rd Qu.:1.0000  
         Max.   :4.0000           Max.   :1.0000  
```

1.2 logistic regression

By looking at the details of each data item in the dataset, it was found that the values of age as well as maximum heart rate were quite different and needed to be processed for both data items.

```
summary(dt$age)
dt$age<-cut(as.numeric(dt$age),breaks=3,labels=c("low1","normal1","high1"))
levels(dt$age)
table(dt$age)
```

```
> summary(dt$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  29.00   47.50   55.00   54.37   61.00   77.00 
> dt$age<-cut(as.numeric(dt$age),breaks = 3,labels=c("low1","normal1","high1"))
> levels(dt$age)
[1] "low1"    "normal1" "high1"  
> table(dt$age)

   low1 normal1   high1 
     64     168      71 
> 
> summary(dt$chol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  126.0   211.0   240.0   246.3   274.5   564.0 
> dt$chol<-cut(as.numeric(dt$chol),breaks =3,labels=c("low","normal","high"))
> levels(dt$chol)
[1] "low"    "normal" "high"  
> table(dt$chol)

   low normal   high 
   222     80      1 
```

Dividing the data set into training and test sets according to a 7 to 3 ratio

```
> samp<-sample(2,nrow(dt),replace = T,prob = c(0.7,0.3))
> training<-dt[samp==1,]
> test<-dt[samp==2,]
> head(training)
      age sex cp trestbps   chol fbs restecg thalach exang oldpeak slope ca thal target
1   high1   1  3      145    low   1       0     150     0     2.3     0  0    1      1
3    low1   0  1      130    low   0       0     172     0     1.4     2  0    2      1
4 normal1   1  1      120    low   0       1     178     0     0.8     2  0    2      1
5 normal1   0  0      120 normal   0       1     163     1     0.6     2  0    2      1
6 normal1   1  0      140    low   0       1     148     0     0.4     1  0    1      1
7 normal1   0  1      140 normal   0       0     153     0     1.3     1  0    2      1
> head(test)
       age sex cp trestbps   chol fbs restecg thalach exang oldpeak slope ca thal target
2     low1   1  2      130    low   0       1     187     0     3.5     0  0    2      1
18   high1   0  3      150    low   0       1     114     0     2.6     0  0    2      1
25    low1   1  3      140    low   0       1     178     1     1.4     2  0    3      1
29   high1   0  2      140 normal   1       0     157     0     0.8     2  1    2      1
36 normal1   0  2      142    low   0       0     160     1     1.4     0  0    2      1
38 normal1   1  2      150    low   0       0     165     0     1.6     2  0    3      1
> colSums(is.na(test))
     age      sex       cp trestbps     chol      fbs  restecg  thalach    exang  oldpeak    slope       ca     thal   target 
       0        0        0        0        0        0        0        0        0        0        0        0        0        0 
>
```

Next, a logistic regression prediction model is built using the training set data

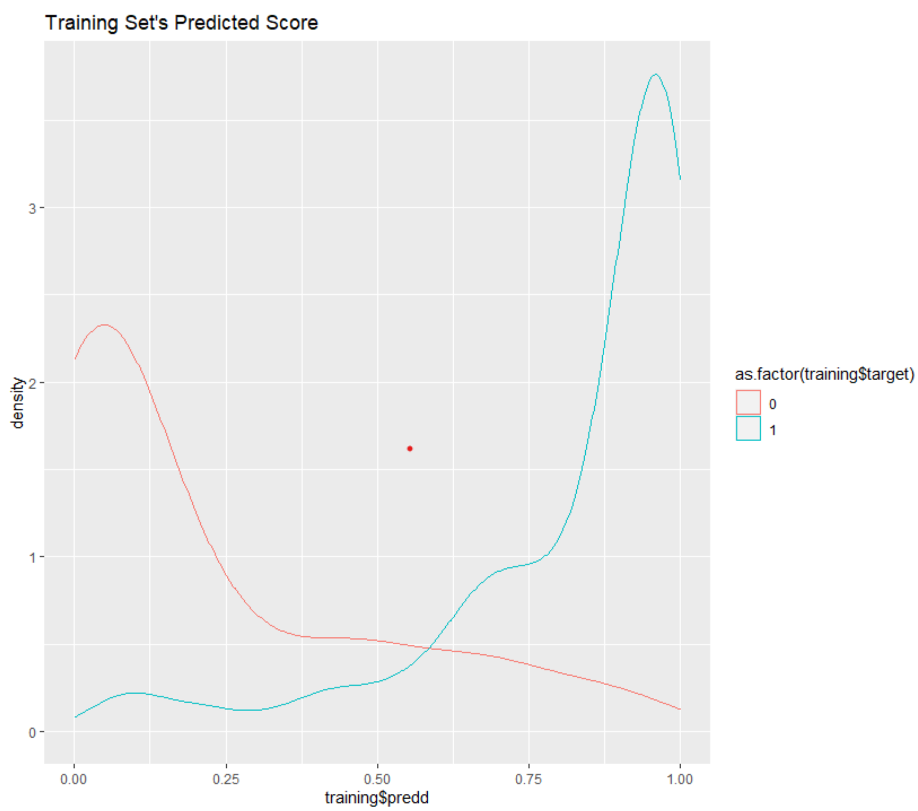

```
> samp<-sample(2,nrow(dt),replace = T,prob = c(0.7,0.3))
> training<-dt[samp==1,]
> test<-dt[samp==2,]
> head(training)
      age sex cp trestbps   chol fbs restecg thalach exang oldpeak slope ca thal target
1   high1   1  3      145    low   1       0     150     0     2.3     0  0    1      1
3    low1   0  1      130    low   0       0     172     0     1.4     2  0    2      1
4 normal1   1  1      120    low   0       1     178     0     0.8     2  0    2      1
5 normal1   0  0      120 normal   0       1     163     1     0.6     2  0    2      1
6 normal1   1  0      140    low   0       1     148     0     0.4     1  0    1      1
7 normal1   0  1      140 normal   0       0     153     0     1.3     1  0    2      1
> head(test)
       age sex cp trestbps   chol fbs restecg thalach exang oldpeak slope ca thal target
2     low1   1  2      130    low   0       1     187     0     3.5     0  0    2      1
18   high1   0  3      150    low   0       1     114     0     2.6     0  0    2      1
25    low1   1  3      140    low   0       1     178     1     1.4     2  0    3      1
29   high1   0  2      140 normal   1       0     157     0     0.8     2  1    2      1
36 normal1   0  2      142    low   0       0     160     1     1.4     0  0    2      1
38 normal1   1  2      150    low   0       0     165     0     1.6     2  0    3      1
> colSums(is.na(test))
     age      sex       cp trestbps     chol      fbs  restecg  thalach    exang  oldpeak    slope       ca     thal   target 
       0        0        0        0        0        0        0        0        0        0        0        0        0        0 
>
```


```
> mod<-glm(target~.,data = training,family = binomial('logit'))
> summary(mod)

Call:
glm(formula = target ~ ., family = binomial("logit"), data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6397  -0.3260   0.1472   0.4614   2.4521

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.43597    3.14892   0.456  0.64838
agenormal1    -1.08551    0.66165  -1.641  0.10088
agehigh1       0.19073    0.82340   0.232  0.81682
sex1          -1.27014    0.63533  -1.999  0.04559 *
cp1            1.07789    0.64981   1.659  0.09716 .
cp2            1.76202    0.59697   2.952  0.00316 **
cp3            2.33462    0.87380   2.672  0.00754 **
```

```
trestbps        -0.02011     0.01359  -1.480  0.13899
cholnormal      -0.87024     0.56281  -1.546  0.12204
cholhigh        13.46342  1455.39800   0.009  0.99262
fbs1             1.40058     0.78331   1.788  0.07377 .
restecg1         0.75994     0.47349   1.605  0.10850
restecg2         0.18667     3.24270   0.058  0.95409
thalach          0.01516     0.01288   1.177  0.23910
exang1          -1.06170     0.53172  -1.997  0.04585 *
oldpeak         -0.52696     0.26619  -1.980  0.04774 *
slope1          -0.74014     1.04764  -0.706  0.47989
slope2           0.01442     1.13336   0.013  0.98985
ca              -0.57276     0.23372  -2.451  0.01426 *
thal1            1.48527     2.16992   0.684  0.49367
thal2            1.70341     1.83567   0.928  0.35343
thal3            0.18380     1.87081   0.098  0.92174
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 292.99  on 213  degrees of freedom
Residual deviance: 134.65  on 192  degrees of freedom
AIC: 178.65

Number of Fisher Scoring iterations: 14
```

After getting the training model, we need to evaluate the suitability of the model for the scenario by the following metrics

```
predicted<-predict(mod, training ,type ="response")
training$predd<-round(predicted ,3)
View(training)
ggplot( training , aes( training$predd , color = as.factor(training$target) ) ) +
  geom_density( ) +
  ggtitle( "Training Set's Predicted Score" )
training$predw<-ifelse(training$predd >0.46,1,0)
```

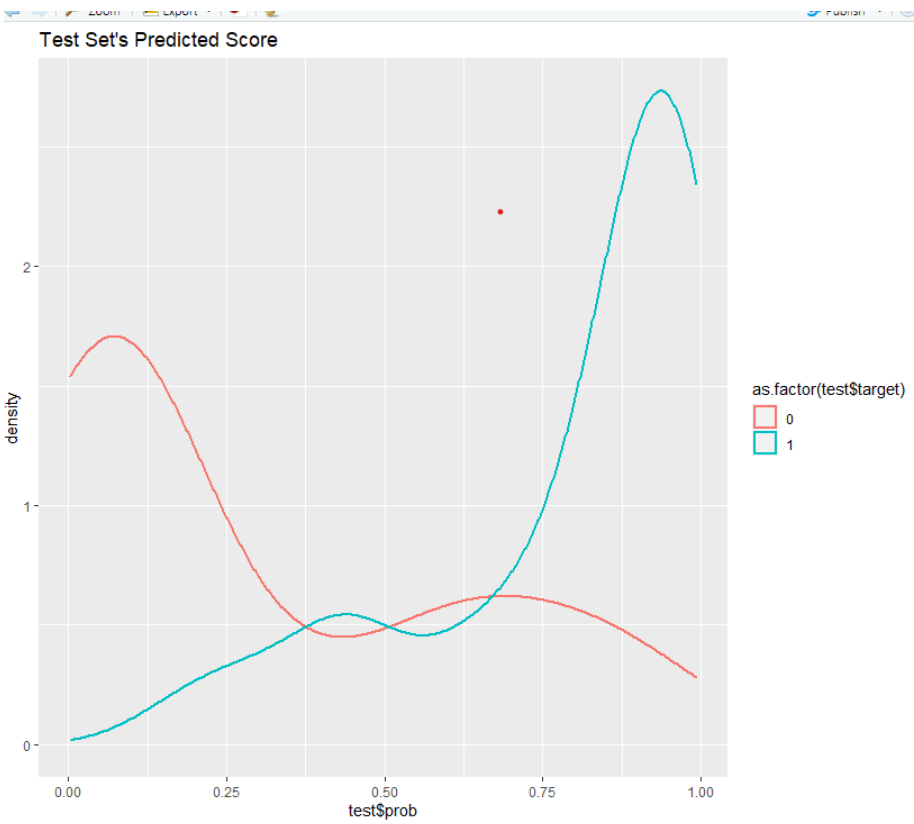



```
> training$predw<-ifelse(training$predd >0.46,1,0)
> confusn<-confusionMatrix(training$target ,
  training$predw ,threshold = 0.46)
> confusn
    0   1
0  74  11
1  19 110
> confusn<-as.matrix(confusn)
> AccuracyRate <- sum(diag(confusn))/sum(confusn)
> AccuracyRate
 [1] 0.8598131
```

Similarly, we need to evaluate this model in the test set

```
> test$prob <- predict(mod, test ,type = "response")
> View(test)
> ggplot( test , aes( test$prob , color = as.factor(
  test$target) ) ) + geom_density( size = 1 ) +
  ggtitle( "Test Set's Predicted Score" )
```

```
> test$predc<-ifelse(test$prob>0.475,1,0)
> confusn<-confusionMatrix(test$target,test$predc,
  threshold = 0.475)
> confusn
   0  1
0 31  7
1 14 37
> class(confusn)
[1] "data.frame"
> confusn<-as.matrix(confusn)
> AccuracyRate <- sum(diag(confusn))/sum(confusn)
> AccuracyRate
[1] 0.7640449
```



The evaluation summary of logistic regression can be seen below

```
  > Confusion
         Actual
```

```
Predicted   0   1
        0  31   7
        1  14  37
> AccuracyRate
[1] 0.7640449
> plot(rocCurve)
> rocCurve$sensitivities
[1] 1.0000000 0.6888889 0.0000000
> rocCurve$specificities
[1] 0.0000000 0.8409091 1.0000000
> rocCurve$au
Area under the curve: 0.7649
```

1.3 method2

1.4 method3

1.5 discuss

2 Task3

2.1 Data mining and cleaning

This data comes from Twitter sentiment analysis in kaggle, From a data set of nearly one million, 13,700 comments about Alan bryd were selected. Among these data, 9,700 positive sentiment data. 4000 negative emotion data. In order to balance the data set, 4000 positive sentiment data and 4000 negative sentiment data were extracted.

id	comment
0	gone to work miss me k leave me love er some shit
0	gone to work... mondays- wednesdays always suck! hope it goes by fast so I can come home to my
0	gonig to do some homework not fun. not fun at all.
0	gonig to work in like 20 mins. it upsets me
0	Gonna attempt to leave my phone in living room on charger and go to sleep alone. wish me luck!! N
0	Have to take my 10 yearold German Shepherd to vet tomorrow
0	Have to take the dog a walk now. And it's raning boo.
0	Have to throw some of her stuffs. Luggages are too full!
0	Have to trash all the carpet that was put in 1.5 years ago
0	Have to wait 3 weeks to find out if I am pregnant or not..
0	Have to wait another day to get the album
0	have to wait for hot water until monday
0	Have to wait for the maintenance guy today. Pipes in building getting a makeover and bathroom sinl
0	Have to wait til Wednesday to get internet back, they sent us the wrong modem
0	Have to wait 'till tomorrow to get my iPhone 3G S They're getting a shipment in around 12 tomorrov
0	Have to wait two weeks for my cord thankfully I had a great experience with Apple support. Thanks
0	have to wait until june 4 to see mtv award cause here in panama didnt transmited tonightt
0	Have to wear the penalty sombrero for the rest of the afternoon after losing in foosball
0	have to work 2night
0	have to work a double today
0	have to work at budabing's at 8 but i feel miserable
0	Have to work now... it will be a hard and long day and then preparing all for the painter on Monday.
0	have to work this arvo
0	have to work todayy oh well ill twitter in the back room ilec ilec ilec
0	Have to work tomorrow.. yay me
0	Have to work until 3am at the bar! Ughh why Im I doing this
0	Have to write a college application. Then i have to go to spanish.
0	Have to write up my case study today and it has to be in tomorrow god damn twitter made me forge
0	have too changes in my life
0	have too many things going on in my life and my mind that I can't seem to get a good nights rest.
0	have too much on my mind & cant sleep
0	Have torn the ligaments in my ankle and now walk like a pleb! xXxc
0	have totally failed one of my chemistry exams
0	have tried half an onion, warm oil, pain meds and nothing is helping my little girls sore ear any hoo
0	Have tried to be nice about it, but that term and "retard" really really hurt sometimes
0	Have u ever been so high your voice sounds 1000x's louder than it really is...?
0	Have u ever been spooked so bad u didnt wanna move at all.... that me right now!
0	Have u ever done somethin' you thought at the time was right, but afterwards realised u wer dead w
0	Have u ever realized that something u never imagined u would need is the one thing u want the mos
0	have u every created an elaborate query from the master database, forget to set the limiter and exe
0	Gonna be a great message at NorthPoint today but I couldn't find my journal this morning.
0	Gonna be a long day here @ home - Everton lost to Chelsea.
0	Gonna be a long day. Working until 7.
0	gonna be all alone tonite I want my girl to come by
0	Gonna be another hot day! We will have water shortage if it doesn't rain soon
0	Gonna be hard but Rpattz and I have to quit smoking. http://bit.ly/90Dql =1

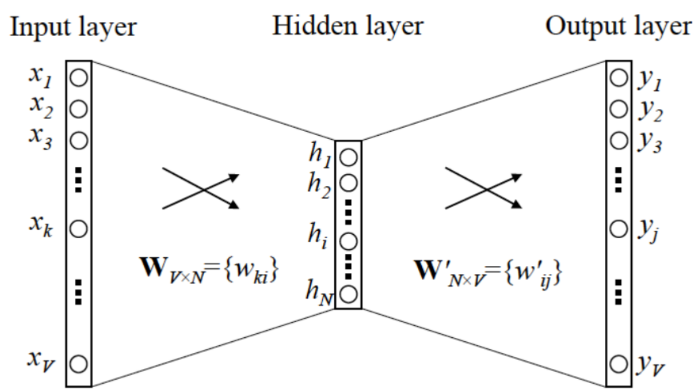
There are most characters in the comments, such as emoji, @ other users and some garbled characters are inconsistent in capitalization

1	Finished my paper!!!!!!!!!! But............ smh
1	first days are always good..wish my dentist gave me more hours though
1	fone off....cant talk 2 my love....imu marcus!!!!!
1	for the fact that *I* didn't get any work done, not that you all did.
1	forgot XBL was off today, was about to check to see if a game was on XBLA that I wanted to buy
1	found a mosquito bite. (Those tend to get really swollen and red for me) Distracting myself with
1	Found a way to make this Private woot woot
1	Four more fake people added me. Is this why people don't like Twitter?
1	frank iero should be the sexiest vegetarian 2009.
1	french lost, #fb
1	fuck man .i hate this. =O. work suckss :'(
1	fuck you, :]
1	- fuck, my money is running out & i have no jobs...
1	fuck. I am an ugly person.
1	fucking rur mom
1	- Fwd: Good Morn,happy birthday ! regardless of what ppl said yesterday,they don't realize what
1	gah. so much less ok than i was trying to tell myself i was.
1	GCSE's clearly suck.
1	- Geez Chelsea already scored against Everton 1-0
1	gerald lost a friend............idk what to say........this is like day 2 of crazy events.......
1	getting "goodbye" e-mails from #Iran #iranelection
1	-- getting a Mani + Pedi with the husband!
1	getting a webcam today.
1	- Getting ready to leave for Spring City, TN - @EzraJane I'm going to miss your show tonight...

So a series of data cleaning operations are used to clean the data.

```python
data=pd.read_csv("AlanBryd.csv").astype(str)
'''
clean the dataset
'''


def remove_pattern(input_txt, pattern):
    r = re.findall(pattern, input_txt)
    for i in r:
        input_txt = re.sub(i, '', input_txt)

    return input_txt
data['comment'] = np.vectorize(remove_pattern)(data['comment'], "@[\w]*")
data['comment']= data['comment'].str.replace("[^a-zA-Z#]", " ")
data['comment'] = data['comment'].apply(lambda x: ' '.join([w for w in x.split() if len(w)>3]))




text = data['comment']
sentences = []
```

2.2 Use word2vec to build word vectors



Use word2vec's word vector for word embedding as input to the model

```
text = data['comment']
sentences = []
for item in text:
    sentence = str(item).split(' ')
    sentences.append(sentence)
#训练
model = word2vec.Word2Vec(sentences, size = 200)

def buildWordVector(imdb_w2v, text, size):
    vec = np.zeros(size).reshape((1, size))
    count = 0.
    #print text
    for word in text.split():
        #print word
        try:
            vec += imdb_w2v[word].reshape((1, size))
            count += 1.
        except KeyError:
            print (word)
            continue
    if count != 0:
        vec /= count
    return vec
result = buildWordVector(model, data.loc[1]['comment'] , 200)
for i in range(1, len(data)):
    result = np.concatenate((result, buildWordVector(model, data.loc[i]['comment'] , 200)), axis = 0)
#读取出数据
#把series 转换成dataframe格式，并且将200维的特征都赋值

vectors = pd.DataFrame(result)
```

2.3 model using and result

2.3.1 GBDT

theory of GBDT

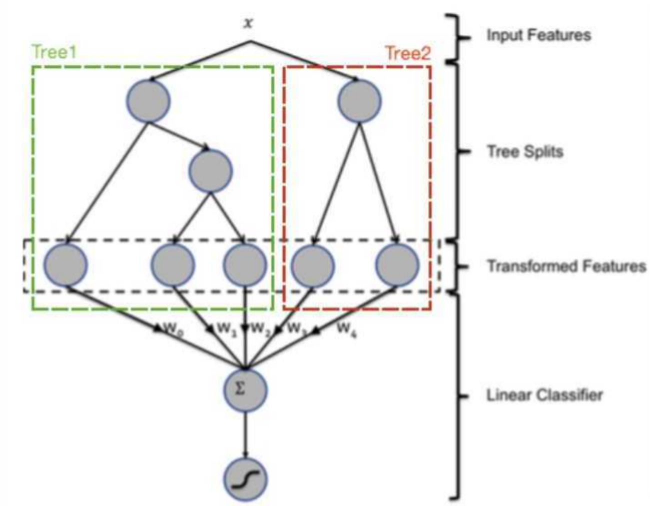
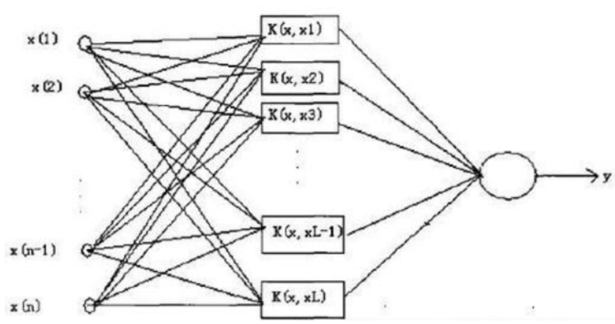


Figure 1: Hybrid model structure. Input features are transformed by means of boosted decision trees. The output of each individual tree is treated as a categorical input feature to a sparse linear classifier. Boosted decision trees prove to be very powerful feature transforms.

Parameters of gbdt:
$n_estimators = 1000$, $subsample = 0.8$, $loss =' deviance'$, $max_features =' sqrt'$,

2.3.2 SVM

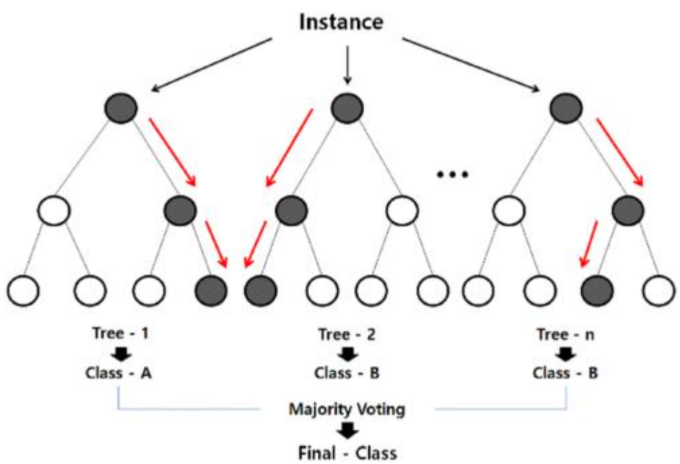
theory of SVM



Parameters of SVM: $kernel = `rbf'degree = 3$

2.3.3 RandomForest

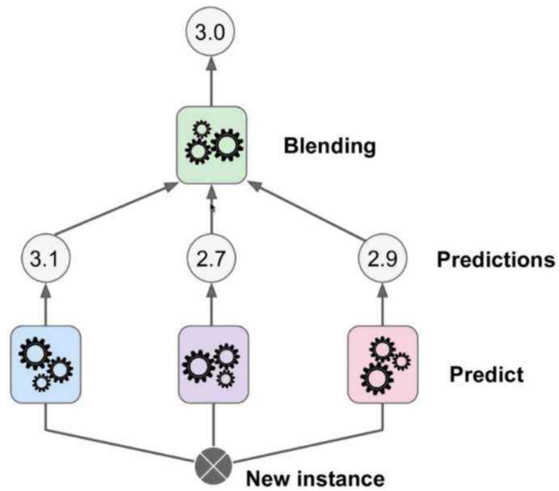
theory of RF



Parameters of RF:$oob_score = True, n_estimators = 400, max_features =' sqrt'$,

2.3.4 ExtraTrees

theory of ET

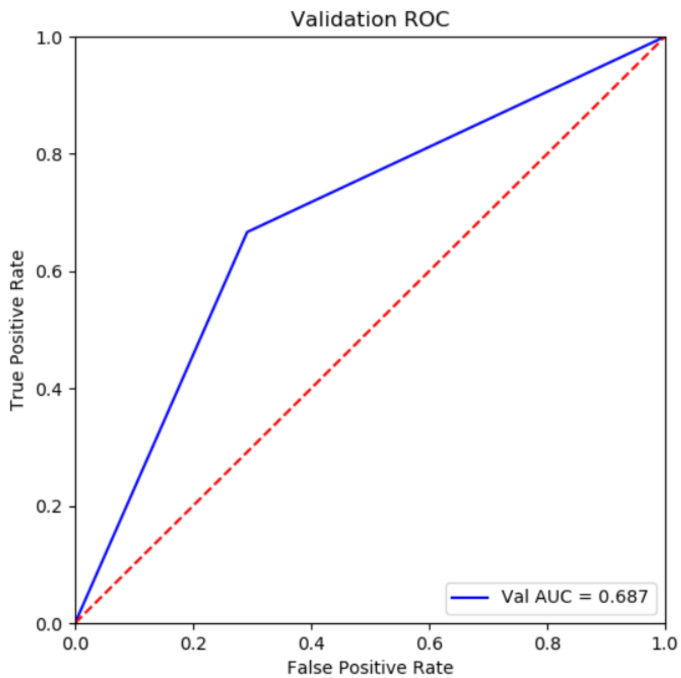


Parameter of ET: $criterion = gini, maxfeatures = log, maxdepth = 50$

2.3.5 Result

```
gbdt score:
model train err is 0.053545
auc is  0.6888644488777488
precision is  0.6891532910703071
recall is  0.6888644488777488
 test err is  0.310948905109489
svm score:
model train err is 0.446049
auc is  0.5516693095981039
precision is  0.5811703096539163
recall is  0.5516693095981039
 test err is  0.44525547445255476
RF score:
model train err is 0.015252
auc is  0.6916253175780518
precision is  0.6925907836786455
recall is  0.6916253175780518
 test err is  0.30802919708029197
ET score:
model train err is 0.015252
auc is  0.6756043219033654
precision is  0.6801474018098703
recall is  0.6756043219033654
 test err is  0.3236009732360097
staking(GBDT ET RF - LR) score:
model train err is 0.015252
auc is  0.6864247105563797
precision is  0.6970203449083667
recall is  0.6864247105563797
 test err is  0.3124087591240876
```

According to the picture above,The expressive power of gbdt is better than other models in all aspects, and the mathematical model in machine learning can fit the features well when the data set is not large. Compared with other tree models, GBDT has a stronger ability to fit data by calculating residuals. The auc figure of GBDT is :


Validation ROC
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Val AUC = 0.687