**Final Project Report**

**Team member: Daoyuan Chen, Haopeng HU**

Goal: Developing a classification model(s) to predict the potential of employees leaving the company (become terminated).

**First**: Make sure the models will be used to predict the potential of employees leaving the company and do some Pretreatment with Data(attrition\_data.csv).

Models: KNN, Naive Bayes, CART, C50, ANN

Pretreatment with Data:

1.Scan all data from attrition\_data.csv, and regard EMP\_ID as unless column.

2.Convert the data type of "JOBCODE" from "numeric" to "factor"

3.Replace all "" with a random ethnicity in "ETHNICITY"

4.Replace all "" with "Unknown” in "REFERRAL\_SOURCE” and remove "" level from "REFERRAL\_SOURCE”

5.Convert the data type of "REHIRE" from "logical" to "factor"

6.Use 30% test 70% training data

**Second**: Using models to Predict STATUS

1. Predict STATUS using KNN(three kind of models k=3,k=5, k=10), using kknn library, combine STATUS with other columns except EMP\_ID to build the prediction model with training data. Making a prediction compare with actual test data.

Prediction results:

K=3

手机屏幕截图

描述已自动生成

k=5

手机屏幕截图

描述已自动生成

K=10

手机屏幕截图

描述已自动生成

The error rate for k=3, k=5 and k =10 are closed, and the model of k = 10 is the most accurate one, so the number K has a little influence with the accuracy of KNN’s Prediction.

1. Predict STATUS using Naive Bayes

Using e1071 library, combine STATUS with other columns except EMP\_ID to build the prediction model with training data. Making a prediction compare with actual test data.

Prediction result:

手机屏幕截图

描述已自动生成

1. Predict STATUS using CART

Using rpart and rpart.plot library, combine STATUS with other columns except EMP\_ID to build the prediction model with training data. Making a prediction compare with actual test data.

Prediction result:

手机屏幕截图

描述已自动生成

1. Predict STATUS using C50

Using C50library, combine STATUS with other columns except EMP\_ID to build the prediction model with training data. Making a prediction compare with actual test data.

Prediction result:

**手机屏幕截图

描述已自动生成**

5.Predict STATUS using ANN

Because the ANN model needs data are numeric type, so we need more pretreatment with the data, and because the limit of R language, for this model, we decide use python to build this model and make a prediction., the pretreatment is as following

1.Drop "EMP\_ID" and "TERMINATION\_YEAR"

2.Perform min-max scaling each numerical feature column to the range [0,1]

3.Perform one-hot encoding on each categorical feature column

4.Perform one-hot encoding on label column

5.Concatenate normalized numerical features and one-hot encoded categorical features into X and denote label by y

Then we will use those data to build ANN model by using Keras to train a neural network with one 5-node hidden layer and a 1-node sigmoid output layer.

手机截图图社交软件的信息

描述已自动生成

Finally we Evaluate the accuracy of the prediction with ANN model.

Prediction result:

手机屏幕截图

描述已自动生成

So the error\_rete of ANN model is around 0.26074898243.

1. Predict STATUS using Random Forest

For the first, because of the same reason with ANN model, so we should do some same pretreatment with data, then we can build a Random forest model to make a prediction as following.

手机屏幕截图

描述已自动生成

And the Prediction result and the accuracy of the prediction with Random Forest model is

手机屏幕截图

描述已自动生成

So the error\_rete of Random Forest model is around 0.2444521498.

**Third:** Conclusion

The error rate of KNN models are 0.3866158(K=3), 0.3863691(k=5), 0.3692788(K=10)

For Naive Bayes model the error rate is 0.350208.

For CART model(Decision-Tree) the error rate is 0.2676838

For C50 model(Decision-Tree) the error rate is 0.2479196  
For Ann model, the error rate is around 0.26074898243.

For Random Forest model, the error rate is around 0.2444521498.

So the accurate order of each model is

Random Forest model, Ann model, C50 model and CART model’s accuracy are closed,

their accuracy are better than KNN models and Naive Bayes model .