CS 6923 Machine Learning
Spring 2019
Final Project Report

Name: Zhenghan He

NetID: zh1158

Name: Weihao Bi

NetID: wb832

PART I: Preprocessing

1. **How does your program handle missing value? And why?**
   - We noticed that there are four features have missing value. So, the best way to handle it is to define a function and use the function to automatically impute missing values.

   - For imputed part, we tried three machine learning models to imputed data. After comparation, we found out the KNN algorithm is the best one to do so.

   - The next step is choosing the order of imputed, we decided to sort the number of missing values of each feature and choose the lowest one as the first feature to imputed, then the second lowest and so on. This process makes sure that the training set is adequate and improve the accuracy.

2. **If your program converts numeric features to categorical features, or categorical features to numeric features. Describe how it does it.**
   - We first selected all feature with only two values, which means those features can be seen as a binary feature (e.g. 0 and 1).

   - There are so many features have more than two values and all of them are character value. To make the classifier work better, we convert them into numeric by use pandas dummy to achieve this. Compare to one-hot, dummy make the value of feature to k-1 new feature rather than k feature.

   - Once we get new k feature with 1 and 0 value. For one row, it presents the old category by set 1 at that column. And the row with all 0 denote the rest category of original category.

3. **Describe any feature selection, combination or creation, and any feature values combination performed by your program and the reasons for doing so.**
   - First of all, we calculate every column's score by using 'Selectkbest', sklearn's API. This function return k columns and scores of them. The function we used is 'f_classif'. We chose those columns whose score is more than 1. This step deducts several columns to make the model more efficient and effective.

   - To use SelectKbset function, we need to do label encode for all columns with character value. So, we use LabelEncoder to achieve this.

   - Then we eliminate several columns which only one value has, since those features will not reflect any useful information and not impact the prediction.

   - For some model we do poly transform for all features to make the performance better.

4. **Describe other preprocessing used in your program (e.g. centralizing, normalization)**
   - We used centralizing as well to prevent specific column become a decisional column.

   - To make sure the training dataset is balanced, we duplicate those sample of which the label is YES after splitting the training data to training set and validation set. And in the following training part, we use 'class_weight = 'balanced'' on those models that support the parameter to make each category is balanced.

PART II: Classification

Model One:

1. **Supervised learning method used in this model is**
   Logistics regression.

2. **Why you choose this supervised learning method?**

- If the data to be classified has many meaningful features, each of which has more or less impact on the final classification results, then the simplest and most effective way is to linearly weight these features and put them into the decision-making process. That is what the logistic regression does. The data we have meet the requirement.

- The amount of feature we used is large. Using logistics regression will have higher calculate speed.

3. **Describe the method you used to evaluate this method.**

- Firstly, we divided the training data into 2 parts: 70% for training, 30% for validation.

- Secondly, because the amount of negative sample is much larger than the amount of positive one. To prevent unbalanced in the training data. we duplicate the positive sample in training set 3 times to make the amount of negative sample and positive sample almost the same.

- Thirdly, we do 5-cross validation on the training sample and use the one that has the highest AUC score to test the validation set. The reason why we choose AUC (area under curve (ROC)) to evaluate the model is that it can prevent the effect of unbalanced data and give a fair evaluation. The accuracy may not indicate the performance of the model correctly when data is unbalanced. However, in our test, the accuracy and AUC are positive correlated.

- We used the confusion matrix, the accuracy and the AOC score to evaluate the model we get from the cross validation. Those metrics are also used to compare the model with other supervised method.

4. **Describe process of experimenting different parameter settings or associated techniques.**
   Parameter name: Regulation (penalty)
   - Parameter values: L2 or L1
   - Performance of different values:

- L2: accuracy: 0.6498180615282831 AUC: 0.6110832347991102
- L1: accuracy: 0.6504465762487595 AUC: 0.6111777223420931

- Analysis:

- The aim of the L2 norm is to prevent overfitting. The aim of the L1 norm is to reduce the number of features. In our case, L1 norm has a little bit better performance because the input matrix is sparse and the feature that we input into the model might have redundancy.

Parameter name: Inverse of Regularization Strength(C)

- Parameter values: 1,2,3
- Performance of different values:

- C = 1($\lambda = 1$) accuracy: 0.6498180615282831 AUC: 0.6110832347991102
- C = 2($\lambda = 0.5$) accuracy: 0.6500496195831955 AUC: 0.611343173512632
- C = 3($\lambda = 0.3$) accuracy: 0.6502811776381079 AUC: 0.6113438744833638

- Analysis:

- As we decrease the regularization strength, the accuracy increases as we expected. Since L2 regularization will prevent overfitting and, in some cases, the accuracy will increase when the regularization strength decrease.

Parameter name: degree of input training data.

- Parameter values: degree = 1 degree = 2 (poly transform)
- Performance of different values:

- Degree = 1 accuracy: 0.6498180615282831 AUC: 0.6110832347991102
- Degree = 2 accuracy: 0.6543830631822692 AUC: 0.6503345380058108

- Analysis:

- The accuracy increases when we transform the input's degree to 2. It indicates that the data has a better fitting with square model.

5. Accuracy and Confusion matrix with most suitable parameters

Confusion Matrix		Predict	
		Yes	No
Correct	Yes	2176	1197
	No	9251	17606

Accuracy: 0.6543830631822692

Parameter: C = 1, degree = 2, penalty='l2'

Model Two:

1. **Supervised learning method used in this model is**
   Neural Network.

2. **Why you choose this supervised learning method?**

- Neural networks have the ability to learn and construct models of nonlinear complex relationships.

- For the data we have, it is hard to determine the relationship between each feature and the classification result. Using neural network will help us find the underlying relationship between them.

3. **Describe the method you used to evaluate this method.**

- Firstly, we divided the training data into 2 parts: 70% for training, 30% for validation.

- Secondly, because the amount of negative sample is much larger than the amount of positive one. To prevent unbalanced in the training data. we duplicate the positive sample in training set 3 times to make the amount of negative sample and positive sample almost the same.

- Thirdly, we do 5-cross validation on the training sample and use the one that has the highest AUC score to test the validation set. The reason why we choose AUC (area under curve (ROC)) to evaluate the model is that it can prevent the effect of unbalanced data and give a fair evaluation. The accuracy may not indicate the performance of the model correctly when data is unbalanced. However, in our test, the accuracy and AUC are positive correlated.

- We used the confusion matrix, the accuracy and the AOC score to evaluate the model we get from the cross validation. Those metrics are also used to compare the model with other supervised method.

4. **Describe process of experimenting different parameter settings or associated techniques.**
   Parameter name: activation function
   - Parameter values: tanh, logistic(sigmoid)
   - Performance of different values:
     - Logistic(sigmoid):
       accuracy: 0.6300363876943433
       AUC: 0.699497490088743

- Tanh
  accuracy: 0.8641085014885875
  AUC: 0.8825613064310555

- Analysis:
- Tanh function has a much better performance. Because the result of tanh can be positive or negative. But logistic (sigmoid function) can only be positive.

Parameter name: hidden layer

- Parameter values: tanh, logistic(sigmoid)
- Performance of different values: (50, 50,50,30), (100,100,100,100,100,50), (100, 100, 100, 100, 100, 100, 50)

- 4 hidden layers
  accuracy: 0.8641085014885875
  AUC: 0.8825613064310555
- 6 hidden layers
  accuracy: 0.9514720476347999
  AUC: 0.9650411711019771
- 7 hidden layers
  accuracy: 0.9478332782004631
  AUC: 0.95080911395743

- Analysis:
- As we increase the amount of the hidden layers and the number of nodes in each layer, the accuracy increases first because more hidden layer and nodes means a better depiction of the complexity relationship between features and label. However, as there are too much hidden layers, overfitting happens. The accuracy decreases.

5. Accuracy and Confusion matrix with most suitable parameters

Confusion Matrix		Predict	
		Yes	No
Correct	Yes	3314	59
	No	1408	25449

Accuracy: 0.9514720476347999
Parameter: activation function: tanh. hidden layer: (100,100,100,100,100,50)

Model Three:

1. **Supervised learning method used in this model is**
   Random Forest

2. **Why you choose this supervised learning method?**

- Because randomness in the random forest, it is hard to overfitting in the training data.

- Also, for the reason of randomness, random forest can avoid the effect of noise data and have a good performance when there is noise in training data.

- Random forest run much faster than other algorithms like neural network or SVM.

- Random forests can process very high-dimensional data without feature selection

3. **Describe the method you used to evaluate this method.**

- Firstly, we divided the training data into 2 parts: 70% for training, 30% for validation.
- Secondly, because the amount of negative sample is much larger than the amount of positive one. To prevent unbalanced in the training data. we duplicate the positive sample in training set 3 times to make the amount of negative sample and positive sample almost the same.
- Thirdly, we do 5-cross validation on the training sample and use the one that has the highest AUC score to test the validation set. The reason why we choose AUC (area under curve (ROC)) to evaluate the model is that it can prevent the effect of unbalanced data and give a fair evaluation. The accuracy may not indicate the performance of the model correctly when data is unbalanced. However, in our test, the accuracy and AUC are positive correlated.
- We used the confusion matrix, the accuracy and the AOC score to evaluate the model we get from the cross validation. Those metrics are also used to compare the model with other supervised method.

4. **Describe process of experimenting different parameter settings or associated techniques.**

   Parameter name: max_depth (The maximum depth of the tree)

   - Parameter values: 20, 50
   - Performance of different values:
     - max_depth = 20

     accuracy: 0.8496857426397618

     AUC: 0.8761292873067194

     - max_depth = 50

accuracy: 0.9950380416804498
AUC: 0.9972074319544253

- Analysis:
  - When the maximum depth is small, the model may be underfitting. After we increase the max_depth, the model fits well to the data.

Parameter name: n_estimators (The number of trees in the forest.)

- Parameter values: 10, 30
- Performance of different values:
  - n_estimators = 10

accuracy: 0.9950380416804498
AUC: 0.9972074319544253

  - n_estimators = 30

accuracy: 0.9980482963943103
AUC: 0.998901589902074

- Analysis:
  - When we increase the number of tress in the forest, the model performs better because as the increase of the trees, the complexity relationship between features and label can be depicted more precisely.

5. Accuracy and Confusion matrix with most suitable parameters

Confusion Matrix		Predict	
		Yes	No
Correct	Yes	3373	0
	No	59	26798

Accuracy: 0.9980482963943103
Parameter: criterion='Gini', n_estimators=30, max_depth = 50

PART III: Best Hypothesis

1. **Which model do you choose as final method?**
   Model number: Model Three
   Supervised learning method used in this model: Random Forest

2. **Reasons for choosing this model.**
   - The average accuracy and AUC of the model in cross validation are the highest among the three model we choose.

   - The accuracy and AUC of the model on validation set are the highest among the three model we choose.

3. **What are the reasons do you think that make it has the best performance?**
   - Because of the randomness in random forest, it is hard to overfitting on the training data. However, when we use logistic regression and neural network, overfitting is an important problem for us to consider. In fact, overfitting actual happened when we use neural network and add too many levels in the hidden layers.

   - Neural network and logistic regression are more sensitive to the noise data in training data. They may adjust the model to fit the noise in the training set and make the prediction less precisely. However, random forest can avoid the noise data affect the model for the reason of randomness.

   - When using other supervised method, we need to do feature selections. In the process of feature selection, the importance of some feature may be over emphasis or ignore. And that will affect the performance of the model we get. However, random forest does not require feature selection. The relationship between the label and the feature can be precisely depicted by random forest.