Project 2 Description

Data Mining - Virtual

Dr. Farahani

Mohammad Javad Aghaie

Contents

** Please upload "kaggle.jason" file to your google drive

Task 1:

1.

Function 'Forward Selection' in the code use forward stepwise selection method to select best features by AUC metric.

2.

First, we use forward selection function to gain important features. 15 of 20 features have been selected. Then, we use a logistic regression for evaluating output of previous section. The results have come in the following:

```
Train set Evaluation is:
F1-score of Model for selected features is:  0.528
Precision_score of Model for selected features is:  0.527
Recall-score of Model for selected features is:  0.542
Test set Evaluation is:
F1-score of Model for selected features is:  0.496
Precision_score of Model for selected features is:  0.492
Recall-score of Model for selected features is:  0.52
```

3.

We create a PCA with 15 component (as it can be seen in previous question) and fit it with our training data. Then, we transpose both train and test sets.

4.

We applied regression to our current data. Here are the results:

```
Train set Evaluation is:
F1-score of Model for selected features is:  0.873
Precision_score of Model for selected features is:  0.877
Recall-score of Model for selected features is:  0.875
Test set Evaluation is:
F1-score of Model for selected features is:  0.87
Precision_score of Model for selected features is:  0.874
Recall-score of Model for selected features is:  0.87
```

As it can be seen, there is huge improvements with PCA.

6.

We have merged Q6 and Q7. The result of each part has come too.

6.1.

We use " KBinsDiscretizer" for binning the date. We choose 5, 10 and 15 bins. The following is result of each bins.

```
LogisticRegression Model Evaluation for 5 Bins

Train set Evaluation is:
F1-score of Model for selected features is:  0.882
Precision_score of Model for selected features is:  0.883
Recall-score of Model for selected features is:  0.882

Test set Evaluation is:
F1-score of Model for selected features is:  0.86
Precision_score of Model for selected features is:  0.873
Recall-score of Model for selected features is:  0.85
LogisticRegression Model Evaluation for 10 Bins
```

LogisticRegression Model Evaluation for 10 Bins

```
Train set Evaluation is:
F1-score of Model for selected features is:  0.936
Precision_score of Model for selected features is:  0.936
Recall-score of Model for selected features is:  0.936

Test set Evaluation is:
F1-score of Model for selected features is:  0.86
Precision_score of Model for selected features is:  0.876
Recall-score of Model for selected features is:  0.85
```

```
LogisticRegression Model Evaluation for 15 Bins

Train set Evaluation is:
F1-score of Model for selected features is:  0.962
Precision_score of Model for selected features is:  0.961
Recall-score of Model for selected features is:  0.963

Test set Evaluation is:
F1-score of Model for selected features is:  0.883
Precision_score of Model for selected features is:  0.899
Recall-score of Model for selected features is:  0.87
```

6.2.

Usually if we attribute a number to each class, by default, many model consider it as its class importance so it prefers some classes more than the others. To preventing this, we use encoding such as One hot encoder.

```
Train set Evaluation is:
F1-score of Model for selected features is:  0.53
Precision_score of Model for selected features is:  0.533
Recall-score of Model for selected features is:  0.549
Test set Evaluation is:
F1-score of Model for selected features is:  0.504
Precision_score of Model for selected features is:  0.515
Recall-score of Model for selected features is:  0.53
```

6.3.

Log transformation is a data transformation method in which it replaces each variable x with a log(x). When our original continuous data do not follow the bell curve, we can log transform this data to make it as "normal" as possible so that the statistical analysis results from this data become more valid. In other words, the log transformation reduces or removes the skewness of our original data. The important caveat here is that the original data has to follow or approximately follow a log-normal distribution. Otherwise, the log transformation won't work.

Due to tremendous samples, Shapiro test is not suitable so we use Kolmogorov-Smirnov test for checking whether data's features have normal distribution or not.

As a result, none of the features have normal distribution so it is appropriate to use log transform. We use Function Transformer in Sklearn and also Numpy log1p with calculate log(x+1). The result of regression is:

```
Train set Evaluation is:
F1-score of Model for selected features is:  0.823
Precision_score of Model for selected features is:  0.833
Recall-score of Model for selected features is:  0.824
Test set Evaluation is:
F1-score of Model for selected features is:  0.813
Precision_score of Model for selected features is:  0.831
Recall-score of Model for selected features is:  0.81
```

6.4.

We use the Create "surface" feature by multiplying screen height and width.

We use our function to create model and evaluate it.

```
Train set Evaluation is:
F1-score of Model for selected features is:  0.539
Precision_score of Model for selected features is:  0.539
Recall-score of Model for selected features is:  0.554
Test set Evaluation is:
F1-score of Model for selected features is:  0.494
Precision_score of Model for selected features is:  0.497
Recall-score of Model for selected features is:  0.52
```

7.

We first perform Log Transformation than 15 bins binning. After that, use one hot encoder.

```
Train set Evaluation is:
F1-score of Model for selected features is:  0.85
Precision_score of Model for selected features is:  0.87
Recall-score of Model for selected features is:  0.83
Test set Evaluation is:
F1-score of Model for selected features is:  0.74
Precision_score of Model for selected features is:  0.77
Recall-score of Model for selected features is:  0.72
```

8.

Bootstrapping is a statistical way to reduce uncertainty. In bootstrapping, we create B number of dataset (tables) from the main one. We randomly choose between the samples in main table and put it in the first table. Then we randomly choose another one and put it in another table. We fill all the tables with this procedure. As an average, we cover two over third of samples of main table (dataset). This method is used to reduce estimation error and have a reduction in variance (increase in bias) in estimation.

The difference between this method and CV is that in bootstrapping we choose between samples but in CV, we choose between features. On the other hand, CV tends to reduce bias (which drives to high variance) and bootstrapping tends to reduce variance (which drives to high variance).

9.

10.

There are 4 popular types of decision tree algorithms: ID3, CART (Classification and Regression Trees), Chi-Square and Reduction in Variance. We just describe ID3.

ID3 (Iterative Dichotomiser)

ID3 decision tree algorithm uses Information Gain to decide the splitting points. In order to measure how much information we gain, we can use entropy to calculate the homogeneity of a sample.

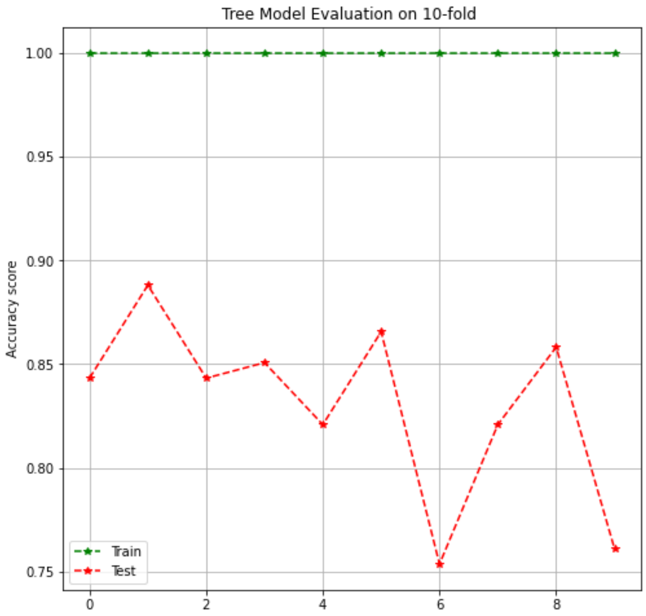
Recursion on a subset may stop in one of these cases:

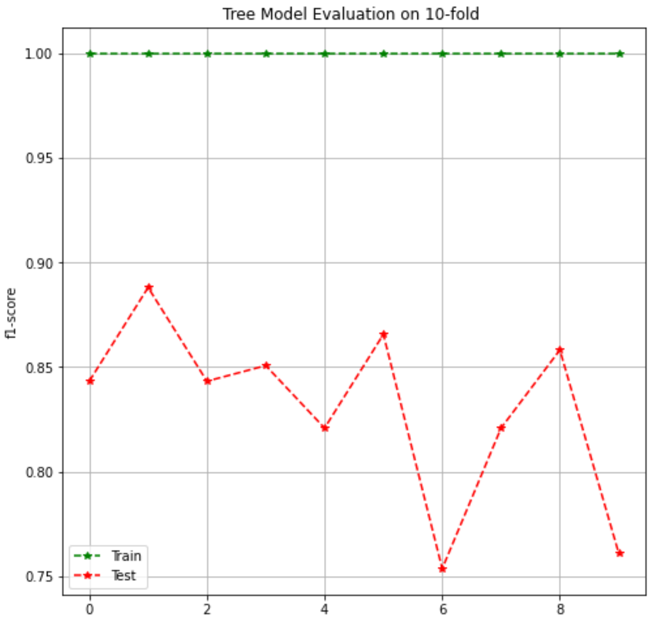
- every element in the subset belongs to the same class; in which case the node is turned into a leaf node and labelled with the class of the examples.
- there are no more attributes to be selected, but the examples still do not belong to the same class. In this case, the node is made a leaf node and labelled with the most common class of the examples in the subset.

- there are no examples in the subset, which happens when no example in the parent set was found to match a specific value of the selected attribute. An example could be the absence of a person among the population with age over 100 years. Then a leaf node is created and labelled with the most common class of the examples in the parent node's set.

11.

We've used Sklearn package for creating Decision Tree. We've also used 10-fold cross validation to overcome the high possibility of overfitting. The result of each fold are shown in the following figure.





For final evaluation, we use x_val with has resulted 83% f1-score.

12.

We changed the criterion and max depth so we could reach better result.

```
Accurracy is 0.84 and f1-score is 0.84
Accurracy is 0.85 and f1-score is 0.85
```

13.

There are two general ways to gain good model. One of them is iteratively divide feature space into different parts (Top-down approach) . Although it is a good method, it is very likely to overfit and leading to poor result in test dataset. For preventing this happens, we make a big tree and then we start to omit some leaves which doesn't affect on model or make it worse. This is called pruning a tree and it is common for making a tree.

14.

For some models it is impossible to draw bias and variance of a model. As the result, this method (elbow) isn't appropriate for finding model complexity. On the other hand, bias decreases and reaches to zero from 5 dimension and it would be impossible to be compared with variance in higher dimensions.

Nevertheless, in some linear regression and K-means models it is useful in finding best dimension and K, respectively.