COMP9313 (20T2) Project 2

Yu Zhang (z5238743)
05.08.2020

1. Evaluation of your stacking model on the test data.

Answer:

F1-score:

```
evaluator = MulticlassClassificationEvaluator(predictionCol="prediction",metricName='f1')
```

Accuracy:

```
evaluator=MulticlassClassificationEvaluator(predictionCol="prediction",metricName='accuracy')
```

Precision:

```
evaluator = MulticlassClassificationEvaluator(predictionCol="prediction",metricName='weightedPrecision')
```

Recall:

```
evaluator = MulticlassClassificationEvaluator(predictionCol="prediction",metricName='weightedRecall')
```

	F1-score	Accuracy	Precision	Recall
result	0.7483312619309965	0.75375	0.7467152515850681	0.7537499999999999

2. How would you improve the performance (e.g., F1) of the stacking model.

For task 2.2, you may try from the following directions:

- **the base feature generation**
- **the meta feature generation**
- **the hyper-parameters of base and meta models**

Answer:

First, we should focus on preprocessing. Raw datasets have many punctuations and some words we are meaningless. So we can use *stopwords* function and *PorterStemmer* function form *nltk* to remove meaningless word, and remove punctuations. Then we can use *base_features_gen_pipeline* function to pipeline the data.

Second, we can generate more meta feature, From lab 3, we also can add logical regression to train and test data, Besides, we can use other method to train data, such as Decision Tree, random forest, Gradient Boosting Decision Tree

Third, we can tune hyperparameters, we can decrease underfitting and overfitting.