University of Nebraska Lincoln

Practical Machine Learning
ENGR 491/891

Programming Assignment 2

Spring 2022

Linear Regression & K-Nearest Neighbors

ENGR 891: 100 points

Last Name 1:

First Name 1:

NUID 1:

Last Name 2:

First Name 2:

NUID 2:

Obtained Score:

1. **Part A:** degree of the polynomial of the optimal model and following results for both training and test data: Mean Squared Error & Coefficient of determination R^2 variance score.

Degree of the polynomial: 4

```
Train: Mean squared error: 4.43
Train: Coefficient of determination r^2 variance score [1 is perfect predi
ction]: 0.96

Test: Mean squared error: 4.40
Test: Coefficient of determination r^2 variance score [1 is perfect predic
tion]: 0.96
```

b) **Part A:** learning curve.

Please see Jupyter Notebook: Liew_assignment2_PartA

c) **Part A:** Answer to Q-1.

In general, the model is as good as it can be - as can be seen from the high R2 of 0.96 for both training and testing data. The MSEs for both train and test data are similar at about 4.4, which points to a well-performing model.

For the learning curve, the training data starts with relatively large negative RMSEs when few instances are used for training and with increasing number of training instances, first sharply, then gradually becoming closer to zero with the increasing number of instances and overcoming the noise, until eventually reaching a plateau of about -5 even when more data are added.

Likewise, the validation data shows a similar trend as the training data, albeit starting with smaller negative RMSEs, and reaches a plateau of about -5 as well like the training curve.

In short, the training and validation curves converge and reach a plateau at about -5 negative RMSE with increasing number of training instances. The model is still very slightly underfitting, therefore having a high bias, because the negative RMSEs for both training and testing data are not zero, but it is the best complex model for a SGD linear regression solution, for this particular data set.

d) **Part B:** observation from the EDA.

We observe that only 8.21% news articles are fake.

Long news articles (>480) are observed more frequently in real articles.

You will have to determine the optimal vectorization technique by comparing 3 techniques.

Among the three vectorization techniques, **binary features method** is chosen. The three methods have comparable test accuracy, precision and recall for the "Real" class and recall for the "Fake" class. Precision of the "Fake" class is the highest at 1 for the TF-IDF values method. However, ROC AUC is the highest for the binary features method at 0.75. Therefore, binary count matrix is used for Experiment 3 and 4.

Test accuracy, test confusion matrix and test classification report and Area under the ROC curve for the 3 techniques:

1. Feature binary counts

```
Test Accuracy:  0.9344262295081968

Test Confusion Matrix:
[[  5  19]
 [  1 280]]

Classification Report:
              precision    recall  f1-score   support

        Fake       0.83      0.21      0.33        24
        Real       0.94      1.00      0.97       281

    accuracy                           0.93       305
   macro avg       0.88      0.60      0.65       305
weighted avg       0.93      0.93      0.92       305


Area under the ROC curve:  0.7523724792408066
```

2. Feature frequency counts

```
Test Accuracy:  0.9311475409836065

Test Confusion Matrix:
[[  5  19]
 [  2 279]]

Classification Report:
```

```
              precision    recall  f1-score   support

        Fake       0.71      0.21      0.32        24
        Real       0.94      0.99      0.96       281

    accuracy                           0.93       305
   macro avg       0.83      0.60      0.64       305
weighted avg       0.92      0.93      0.91       305


Area under the ROC curve:  0.5954181494661922
```

3. TF-IDF values

```
Test Accuracy:  0.9377049180327869

Test Confusion Matrix:
[[  5  19]
 [  0 281]]

Classification Report:
              precision    recall  f1-score   support

        Fake       1.00      0.21      0.34        24
        Real       0.94      1.00      0.97       281

    accuracy                           0.94       305
   macro avg       0.97      0.60      0.66       305
weighted avg       0.94      0.94      0.92       305


Area under the ROC curve:  0.6041666666666667
```

e) **Part B:** for all experiments (2 to 4) report test accuracy, test confusion matrix and test classification report.

1. Experiment 2

```
Test Accuracy:  0.9344262295081968

Test Confusion Matrix:
[[  5  19]
 [  1 280]]

Classification Report:
              precision    recall  f1-score   support
```

```
              precision    recall  f1-score   support

        Fake       0.83      0.21      0.33        24
        Real       0.94      1.00      0.97       281

    accuracy                           0.93       305
   macro avg       0.88      0.60      0.65       305
weighted avg       0.93      0.93      0.92       305
```

2. Experiment 3

```
Performance Measures Using Optimal Threshold from Precision-Recall Curve:
Test Accuracy:  0.9114754098360656

Test Confusion Matrix:
[[ 11  13]
 [ 14 267]]

Classification Report:
              precision    recall  f1-score   support

        Fake       0.44      0.46      0.45        24
        Real       0.95      0.95      0.95       281

    accuracy                           0.91       305
   macro avg       0.70      0.70      0.70       305
weighted avg       0.91      0.91      0.91       305
```

3. Experiment 4

```
Test Accuracy:  0.8786885245901639

Test Confusion Matrix:
[[ 12  12]
 [ 25 256]]

Classification Report:
              precision    recall  f1-score   support

        Fake       0.32      0.50      0.39        24
        Real       0.96      0.91      0.93       281

    accuracy                           0.88       305
   macro avg       0.64      0.71      0.66       305
weighted avg       0.91      0.88      0.89       305
```

f) **Part B:** Answer to Q-2, Q-3 & Q-4.

- Q-2) Why is the performance (precision & recall) of the true/real class higher than the fake class? Explain.

The performance of the "Real" class is higher than the "Fake" class because the data set is overwhelmingly represented by the “Real” class (91.8%). Recall is especially affected because many “Fake” articles are predicted as “Real”.

- Q-3) Compare the results from experiment 2 and 3. What changes do you observe. Explain.

The recall for “Fake” class increases from 0.21 to 0.46 while precision decreases from 0.83 to 0.44. On the other hand, the recall for “Real” class decreases from 1 to 0.95 and precision increases a little from 0.94 to 0.95. Using the optimal threshold from the Precion-Recall curve, test accuracy also drops slightly from 0.93 to 0.91.

- Q-4) Compare the results from experiment 3 and 4. What changes do you observed. Explain.

For the misinformation class (“Fake”), test recall increases slightly from 0.46 to 0.50 but precision decreases from 0.44 to 0.32. The tradeoff is probably not worth it because the improvement of recall (+0.04) is a lot smaller than the decrease of precision (-0.12).