Week 3.2
Using Python Tools to Generate a Custom Dataset, Add new features and Experiment with Selection Tools.

Produced By: Abdullah Abdul Majid*(linkedin)*

Completion Date: July 19, 2024

Email: abdullahpisk@gmail.com

Jupyter Notebook Links:

GitHub

Table of Contents:

Problem Statement:

Create a random custom dataset and add more features to it. Then experiment with feature selection techniques to improve the model efficiency.

Purpose:

The purpose of this assignment is to create a very large custom dataset using the given command. Then use the tools from the Sklearn library to add more features to it. And then use multiple feature selection tools to reduce the number of features for the model training and efficiency.

Process:

Dataset Creation and Initial Inspection

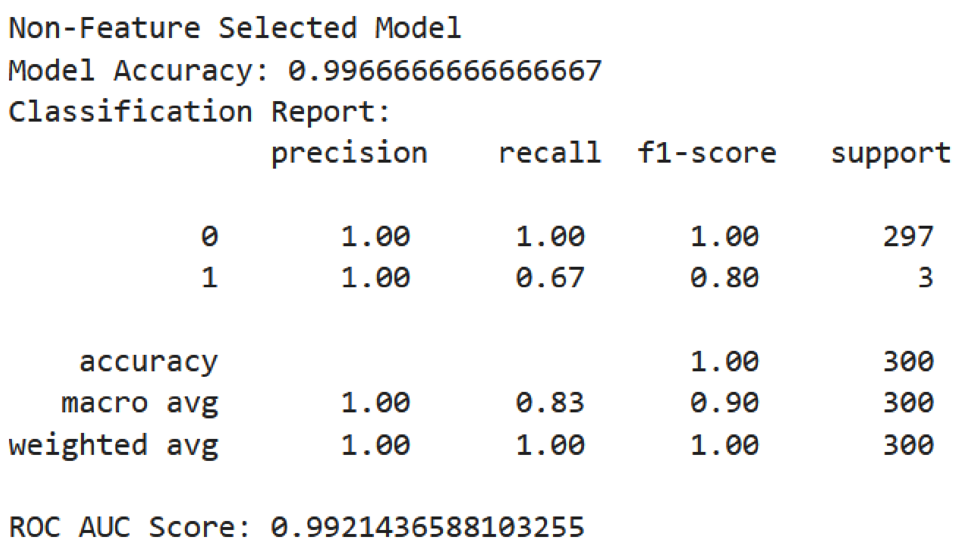
- Imported all the necessary libraries, i.e pandas, seaborn, matplotlib and sklearn.
- Created the dataset using the following command:
  X, y = make_classification(n_samples=1000, n_features=20, n_informative=2, n_redundant=10, n_clusters_per_class=1, weights=[0.99], flip_y=0, random_state=1)
- Inspected the dataset shape.
- Printed the first five samples.
- Computed and plotted the correlation matrix.

Adding More Features to the Dataset

- Used the Polynomial Features function to add polynomial features (Exponential of numerical features) and interaction terms (Interaction/products between different pictures).
- Printed the result.
- Split the new dataset into test and train sets for model training.

Training a Basic Logit Regression Model

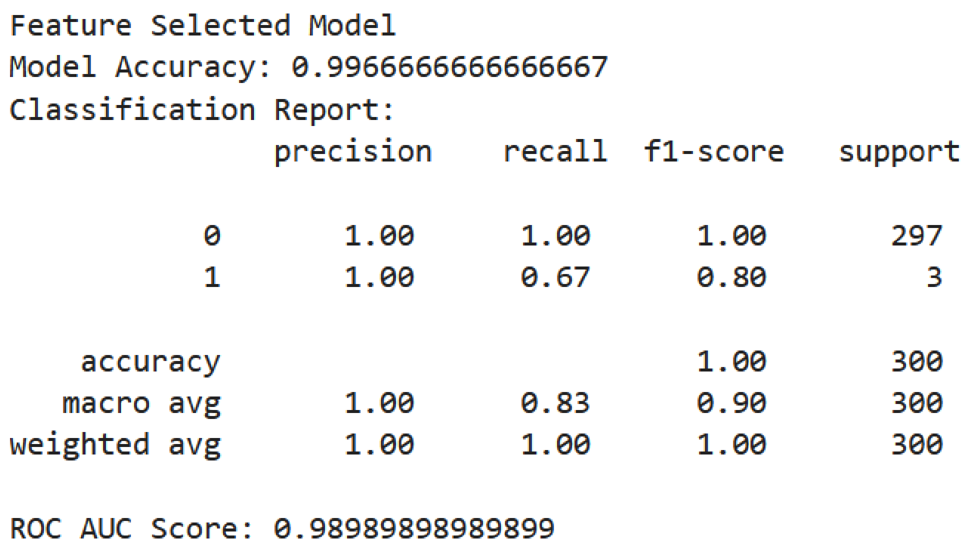
- Train a logit regression model on the polynomial features.
- Evaluate the model's performance using the classification_report, roc_auc and accuracy.



```
Non-Feature Selected Model
Model Accuracy: 0.9966666666666667
Classification Report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00       297
           1       1.00      0.67      0.80         3

    accuracy                           1.00       300
   macro avg       1.00      0.83      0.90       300
weighted avg       1.00      1.00      1.00       300

ROC AUC Score: 0.9921436588103255
```

Using Recursive Feature Elimination to select the Most Impactful Features

- Used Sklearn’s RFE tool to reduce the number of features to the top five most impactful features.
- View the selected features.
- Train the Logit Regression model again.
- Evaluate the model performance again using the same techniques used above.



```
Feature Selected Model
Model Accuracy: 0.9966666666666667
Classification Report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00       297
           1       1.00      0.67      0.80         3

    accuracy                           1.00       300
   macro avg       1.00      0.83      0.90       300
weighted avg       1.00      1.00      1.00       300

ROC AUC Score: 0.98989898989899
```

Conclusion

The model performance was pretty much the same after using RFE, which I used because it’s a wrapper-based feature reduction model. Such models also consider interaction terms for feature selection, hence are better in our case. However, our model wasn’t impacted.