HW1 Technical Report

Bryce Grover

Analysis

Pre-processing

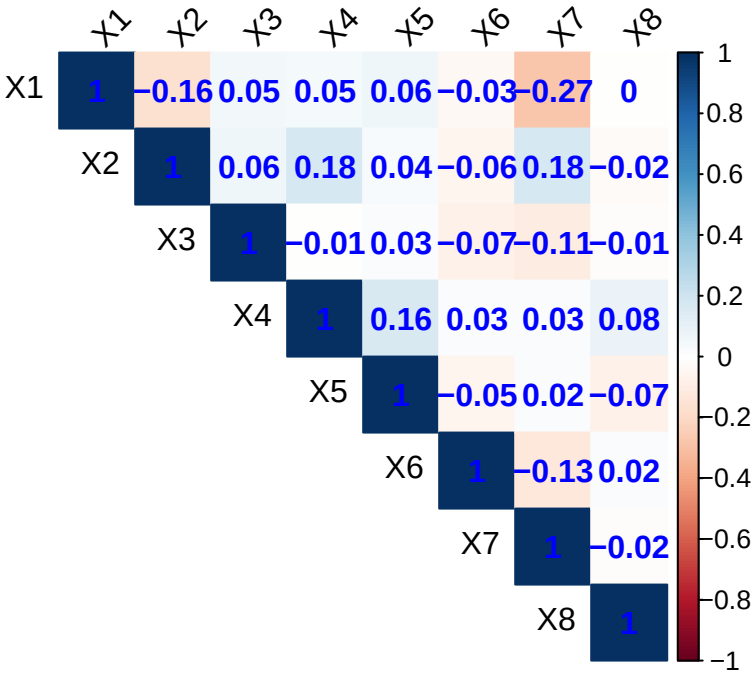
The dataset is comprised of eight continuous numeric variables (X1 through X8) used to predict a binary outcome ('Group'). Intitial data cleaning involved removing Null values and z-scoring features so they are on the same scale. Such scaling is essential for model and data comparison. Data was split using a random state, '123', to ensure consistent model results across multiple model executions.

Below are the summary statistics.

```
       X1                X2                X3                X4
 Min.   :-2.8704   Min.   :-3.2248   Min.   :0.008168   Min.   :0.002185
 1st Qu.:-0.9882   1st Qu.:-0.7538   1st Qu.:0.241284   1st Qu.:0.241352
 Median : 0.3285   Median : 0.6484   Median :0.495069   Median :0.510937
 Mean   : 0.2548   Mean   : 0.5363   Mean   :0.494275   Mean   :0.498409
 3rd Qu.: 1.4669   3rd Qu.: 1.8753   3rd Qu.:0.737394   3rd Qu.:0.757189
 Max.   : 3.2896   Max.   : 4.1234   Max.   :0.992868   Max.   :0.988585
       X5                 X6                X7                 X8
 Min.   :0.0006907   Min.   :0.00548   Min.   :0.0003876   Min.   :0.002317
 1st Qu.:0.2244917   1st Qu.:0.19112   1st Qu.:0.3140703   1st Qu.:0.176609
 Median :0.5798780   Median :0.49945   Median :0.5243598   Median :0.479805
 Mean   :0.5157830   Mean   :0.47572   Mean   :0.5133883   Mean   :0.478011
 3rd Qu.:0.7710559   3rd Qu.:0.69750   3rd Qu.:0.7196874   3rd Qu.:0.733842
 Max.   :0.9988967   Max.   :0.96149   Max.   :0.9859866   Max.   :0.999016
```

Correlation Matrix

Additionally, the data features seem largely uncorrelated.



Methods

Three models were developed:

1. **SVM:** Implemented with a range of kernels, C values, and gamma parameters to find the optimal boundary for class separation. The best kernel was RBF, the best c value was 5, and the best gamma was 0.1. Model ROC/AUC and accuracy was measured to evaluate model performance.
2. **Logistic Regression:** In addition to SVM, a logistic regression model was trained on the same train-test-split. Model ROC/AUC and accuracy was measured to evaluate model performance.
3. **KNN:** Finally, a KNN model was fit on the same train-test-split. A GridSearchCV was used to find the optimal number of K's. The optimal number of K's is 5 for this particular train-test-split. Model ROC/AUC and accuracy was measured to evaluate model performance.

Results

The models' performances were assessed using ROC/AUC and accuracy metrics. The selection of the best model for production will depend on its performance according to these standards, which include its ability to accurately make true positive predictions.

SVM

1. *The SMV model training accuracy was 0.85 and the testing accuracy was 0.74*
2. *The SMV model ROC/AUC was 0.85*

Logistic Regression

1. *The LR model training accuracy was 0.76 and the testing accuracy was 0.78*
2. *The LR model ROC/AUC was 0.86*

KNN

1. *The KNN model training accuracy was 0.76 and the testing accuracy was 0.78*
2. *The KNN model ROC/AUC was 0.78*

Model Metric Conclusion

The best performing model based on both testing accuracy and ROC/AUC scores was the Logistic Regression model. An ROC/AUC score demonstrates a model's ability to differentiate between true positives and false positives. A score close to 1 is better. The 78% accuracy means that out of every 100 predictions the model makes, 78 are likely to be correct. With an ROC/AUC score of .78, the Logistic Regression model has a good ability to differentiate between true positives and false positives. Given that the Logistic Regression model outperforms both the SVM and KNN models, it is the best choice to implement in production.

Reflection

This was my first time using sklearn pipelines in combination with GridSearchCV create a more optimized model. When I took 392, pipelines weren't used in the curriculum, although we did learn about GridSearchCV when trying to find the optimal hyperparameters of other models. Using both in this homework assignment helped me practice using a more srteamlined work flow.