**Principal Component Analysis (PCA)**:

After we obtained the original dataset, there are 844 features for each game, which are placed in 12 files separately. Because the number of features for each game is large, we utilized the PCA to do the dimension reduction firstly.

Before we implemented the PCA, we analyzed the meaning of 12 datasets. We found that each of these 12 data files represent a main factor that could affect the game result. For example, in the file “Away\_Def\_Pass.csv”, there are 88 features that are related to the Away Team’s ability to pass ball when they are in defense. Hence, we decided to do dimension reduction of 12 datasets respectively before we concatenate them. When implementing PCA on each dataset, we utilized the Cumulative Proportion of Variance Explained to determine how many components to preserve. When the Cumulative Proportion of Variance Explained just exceeded 99%, we selected the corresponding number of components as the number of dimensions we want to preserve after the dimension reduction.

For example, the graph of Cumulative Proportion of Variance Explained curve for “Away\_Def\_Pass.csv” is as follows,

Chart

Description automatically generated

(Graph 1)

In this example, when the number of components equal to 4, the Cumulative Proportion of Variance Explained just exceeded 99%. Then we kept 4 components from the PCA. Since all of the remaining 11 datasets exhibited the similar pattern as the Graph 1, this can also validate our inference that the features in each dataset have high correlation.

**Support Vector Machine (SVM)**:

When developing the SVM model, we constructed a pipeline to assemble the range of tuning parameter and kernel type selection. And then we applied the Cross Validation to optimize the Pipeline and obtained the corresponding optimized value for the parameters in Pipeline.

In the Pipeline, we set the tuning parameter can be 0.01,0.1,1,10, or 100 and the kernel can be linear or radial basis function. The optimized pipeline provided that and the kernel is linear. And then we run this model on the datasets and obtained the corresponding precision, recall, and f1-score as in the following Table1,

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | f1-score |
| True | 0.70 | 0.76 | 0.73 |
| False | 0.64 | 0.56 | 0.60 |

(Table 1: SVM)

Let , , , and represent the number of True Positive, False Positive, True Negative, and False Negative respectively. And , . Then, can imply 70% of Home Team Win result that the model predicted is correct. means that the model can correctly identify 76% of games which have Home Team Win result in reality. Because and , we can also know that 64% of Away Team Win result the model predicted is correct and the model can identify 56% of games which have Away Team Win. In addition, , which combines the information from both of Precision and Recall and provides an evaluation score for the model. From Table 1, we can find that the SVM model can predict the Home Team Win Result more accurately than the Away Team Win Result.

**Voting Classifier:**

We have trained the Logistic Regression Model, SVM, and Random Forest model respectively. In order to take all of the three models we have trained into account, we applied the Bagging method to combine these three models and constructed the Soft Voting Classifier model.

In the Soft Voting Classifier model, we set the weight of each model equal to . After running the Soft Voting Classifier model on the dataset, we obtain the Precision, Recall, and f-1 score which are listed in the following Table 2.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | f1-score |
| True | 0.70 | 0.78 | 0.74 |
| False | 0.66 | 0.55 | 0.60 |

(Table 2: Soft Voting Classifier)

The evaluation of the result from the Soft Voting Classifier model is almost consistent with result from the Logistic Regression Model, SVM, and Random Forest model. Because the Soft Voting Classifier model takes all of the three models into consideration, the Table 2 can imply that our trained models can perform better when predicting and identifying the Home Team Win result compared with predicting and identifying Away Team Win result.