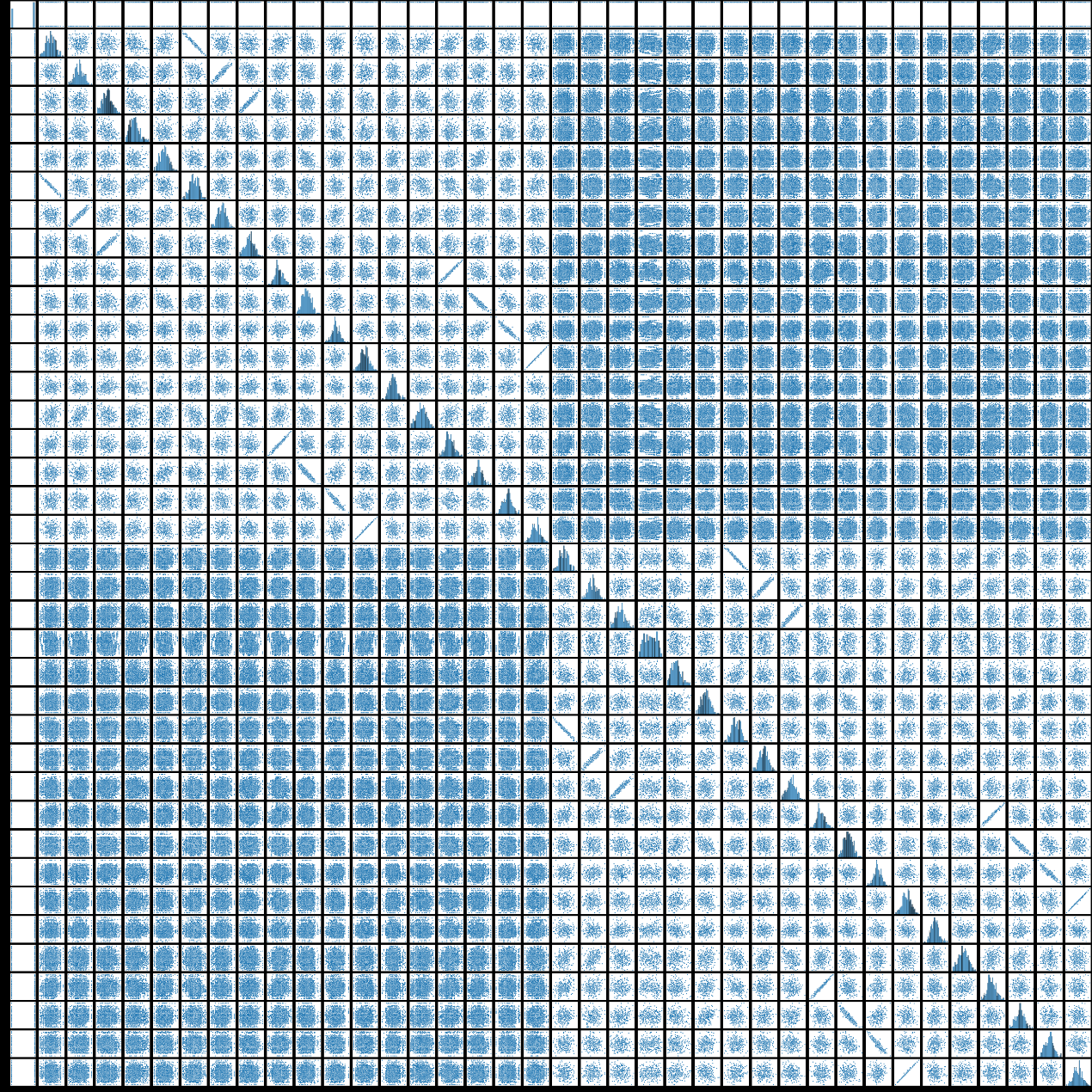
**character encoding**

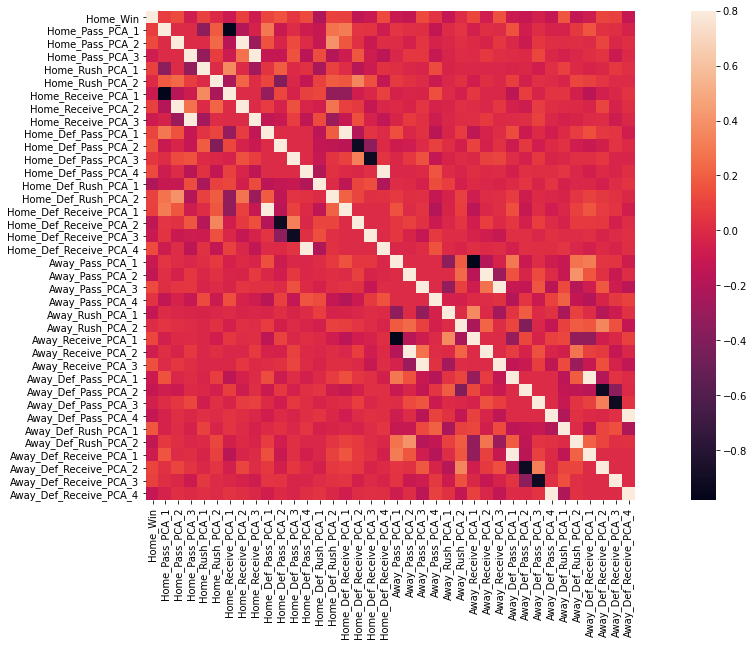
We can do character encoding in some columns in the RAW\_DATA.csv. We choose the column of Home\_win in the Raw\_Data as the target variable. It shows that the column of data is of type bool. We try to convert Boolean values to numbers, changing ‘True’ to 1，changing ‘False’ to 0.

**data visualization**

After combine Home\_win data with all PCA data, we got the complete data. Then we call Seaborn, which is a Python visualization library based on Matplotlib. We first use pairplot, which mainly shows the relationship between variables. You can see that the diagonal line is the histogram of each attribute, and the non-diagonal line is the correlation graph between two different attributes. As can be seen from the histogram on the diagonal, there are orders of magnitude differences between different features. Many statistical methods are very sensitive to the distribution of data. Statistical models tend to focus on data with high concentrations or large variations in multiples, whereas it is the changes in low concentrations that may make a difference. We can normalize the data in the later step using preprocessing.scale from sklearn.



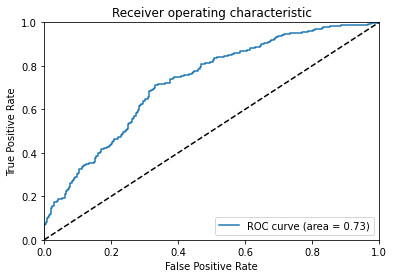
Correlation matrix can also be used to analyze the correlation between features. The correlation matrix can be obtained by corr(), and then the heatmap can be obtained by seaborn.heatmap. From the first line of the heat map, you know the dependency of Home\_win and the various PCA data. These dependencies are not very different, so we take all of the PCA data and used it to build the model.



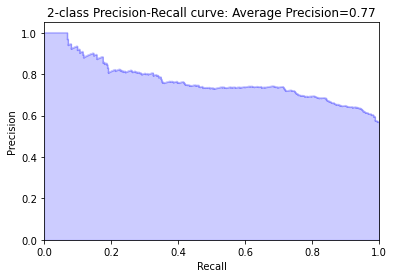
**Model 1:** **Logistic Regression**

Logistic Regression is actually a classification model and is often used for dichotomy. The essence of Logistic regression is to assume that the data obey this distribution, and then use maximum likelihood estimation to estimate the parameters. We use the LogisticRegression in sklear.linear\_model to build the logistic regression model. The optimal tunning parameter is found by GridSearchCV, and our best L1\_ratio is 0.5. The model can be evaluated by.best\_estimator\_.score, and our score is about 0.679.

After using the logistic regression model with the best parameter to predict the value corresponding to X\_test, we can use ROC Curve to evaluate the model. It is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. Our area of ROC curve is 0.73.



We also can evaluate the model by using recall curve. The average precision (AP) could be computed from prediction scores. This score corresponds to the area under the precision-recall curve. The average precision of logistic regression model is 0.76.



The optimal threshold could be found by maximizing the F1 score. When F1 Score is about 0.76, the threshold of logistic regression model is about 0.3.