STATS 101C Final Project

Xunye Qian (504779593), Xingruo Zhang (304756494), Lucy Zhao (304917901)

1. Data Cleaning & Initial Variable Selection

We first removed all the duplicated columns under different names and stored the resulting dataset in a new .csv file called 'train_rm.csv'. The new dataset contains 156 variables. We then recoded the response variable *'HTWins*' into a 0/1 variable called *'HTwins_01'* ('Yes' as 1 and 'No' as 0) and calculated the correlations between all numeric variables and the response variable.

Among the variables that have the highest correlation with the response variable, we added up the variables for different starting player positions and created 3 new variables called '*VT.OS.plmin.total*', '*VT.S.plmin.total*', and '*HT.total.pts'*. We substituted the single variables with the totals and used the remaining 11 variables to fit a logistic regression.

2. Models Description & More Variable Selection

2.1. Logistic Regression with 11 Variables

With this first logistic regression model, we reached a training accuracy of 0.6523.

2.2. Logistic Regression with 9 Variables

From the summary of the logistic regression model above, we found that the coefficients of *'VT.TA.pts'* and *'VT.pmxU'* are not statistically significant. Hence, we removed *'VT.TA.pts'* and *'VT.pmxU'* from our selected variables. We then fitted another logistic regression using the resulting 9 variables. After removing the 2 variables, our training accuracy improved from 0.6523 to 0.6533.

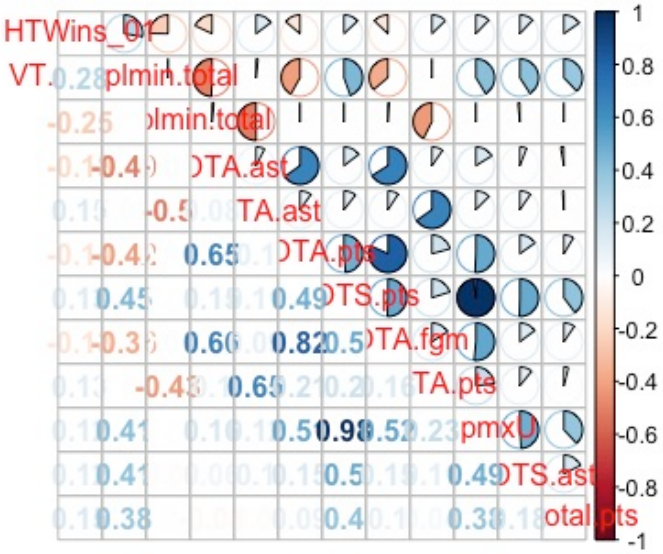

Figure 1. Correlation between the response variable and 9 selected predictors

2.3. Support Vector Classifier (SVC) with 9 Variables (Final Model)

We applied the same 9 variables to train a support vector classifier. We saw that the new model significantly improved the training accuracy to 0.6765.

		Reference	
		No	**Yes**
Prediction	**No**	1858	1067
	Yes	2013	4582
Accuracy		0.6765	

Table 1. Confusion matrix of support vector classifier with 9 variables

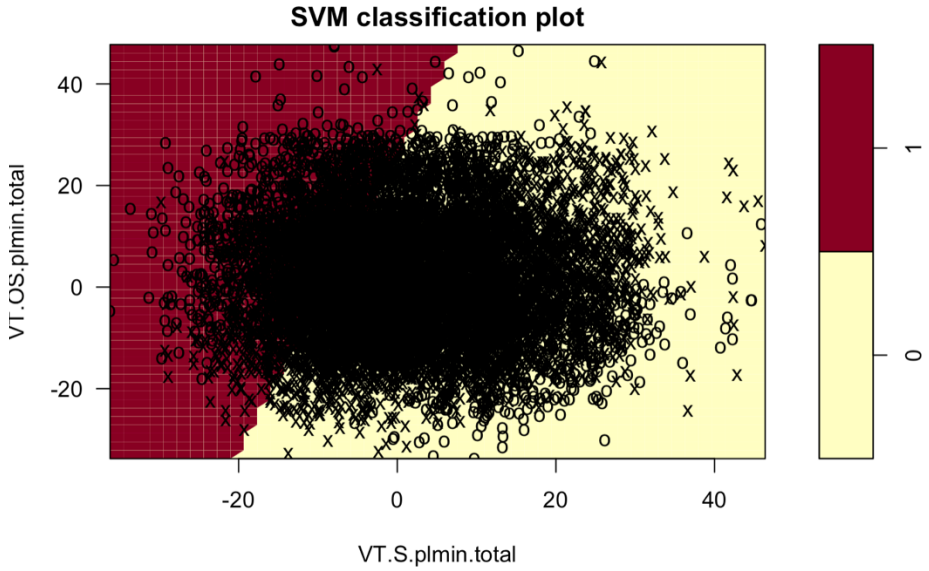

Figure 2. Example visualization of support vector classifier decision boundary

3. Final Model & Classification Rate

Our final model's public classification rate on Kaggle is 0.67475, and its private classification rate is 0.67597.

4. Analyses of Final Model Performance

The advantage of the Support Vector Classifier lies in its robust performance when training high dimensional data. The model has low sensitivity to observations far from the decision boundary. To decide our kernel, we adopted cross validation and concluded that linear kernel provided the highest test accuracy. We also selected a cost parameter of 10 after cross validating the model. The margin of the model is relatively narrow, so the numbers of support vectors and violations are small. Furthermore, we only used 9 predictors in the model to achieve computational efficiency and prevent overfitting. In conclusion, with a small set of predictors, the Support Vector Classifier provides high accuracy rates for the test data.