**Student Name**

* **Pranav Vinod**

### *Notes:*

### *First, please re-save this document on your computer, RENAMING the file to contain your last name.*

### *Point values of each part are shown below; 10 points will be allocated for the quality of your business writing (organization, clarity, grammar, etc.).*

### *Type or paste your responses into the boxes below. The boxes will expand to fit your answers.*

### *Deliverables- Upload following 2 files on the course website:*

### *This completed file.*

### *R file used.*

### Classification and Prediction Models

1. Read the betting.csv data file into RStudio. Run set.seed(XXX) by using last three digits of your student ID in place of XXX followed by partitioning of the dataset into training (50%) and testing (50%). Report on how many cases of Win, Draw, and loss exist in the training and testing data. (10 pts).

A picture containing graphical user interface

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

There are 216, 215 and 349 cases of Draw, Loss and Win respectively in the Train dataset.

There are 188, 177 and 375 cases of Draw, Loss and Win respectively in the Test dataset.

1. Develop a multinomial logistic regression model for predicting match outcome based on the training dataset and report the final model, prediction equations, confusion matrix for training and test datasets. What conclusions can you derive regarding betting about match outcome at game half-time? (20 points):

Graphical user interface, text, application, email

Description automatically generated

The prediction equations are as follows:

Confusion Matrix for Train

A screenshot of a computer

Description automatically generated with medium confidence

Confusion Matrix for test data

A screenshot of a computer

Description automatically generated with medium confidence

From the confusion matrices, we can see that the overall accuracy is 66% for the training set and 64% for the testing set. Among the three different classes, the algorithm is best at correctly predicting “Loss” for both the train and test set with sensitivity of 0.83 and 0.72 respectively. It is the worst at correctly predicting “Draw” for both the sets.

1. Develop a decision tree for predicting match outcome using training dataset and report the final tree, related code, confusion matrix for both training and test datasets. What conclusions can you derive (20 points):

Graphical user interface, text

Description automatically generated

Timeline

Description automatically generated

From the final tree, we can see that FGS0, the first goal scored by the away team is the most important factor in deciding how the game ends. Depending on whether the first goal was scored by the away team or not, the next most important attribute is the half time goal difference, HTGD.

If the first goal was not scored by the away team, then HTGD is the only other factor which decides the outcome of the match. If it is less than 0 then the match likely ends in a draw and if it is greater than 1, the home team usually wins the game.

If the first goal was scored by the away team, then along with HTGD, TOTALHP is also a factor in predicting the outcome.

Confusion Matrix for Train set

A screenshot of a computer

Description automatically generated with medium confidence

Confusion Matrix for Test set

A screenshot of a computer

Description automatically generated with medium confidence

From the confusion matrices, we can see that the overall accuracy is 67% for the training set and 60% for the testing set. Among the three different classes, the algorithm is best at correctly predicting “Loss” for both the train and test set with sensitivity of 0.79 and 0.74 respectively. It is the worst at correctly predicting “Draw” for both the sets.

Chart

Description automatically generated

The model has an area under the curve of 72.2% with the threshold point being (0.686, 0.746).

1. Develop a random forest model for predicting match outcome using training dataset and provide confusion matrix for both training and test datasets. What conclusions can you derive (20 points):

Chart

Description automatically generated

From the Variable Importance plot we can see for Win and Draw, HTGD is the most important variable followed by FGS1 for Win and POINTSH for Draw. For Loss, FSG0 is the most important variable followed by HTGD.

Training Confusion Matrix

A screenshot of a computer

Description automatically generated with medium confidence

Testing Confusion Matrix

A screenshot of a computer

Description automatically generated with low confidence

Further, from the confusion matrices we can see that the overall accuracy is 88% for the training set and 63% for the testing set. Among the three different classes, the algorithm is best at correctly predicting “Loss” for both the test set and “Win” for the train set with sensitivity of 0.81 and 0.93 respectively. It is the worst at correctly predicting “Draw” for both the sets.

1. Develop an extreme gradient boosting model for predicting match outcome using training dataset and provide confusion matrix for both training and test datasets. What conclusions can you derive (20 points):

Chart

Description automatically generated

From the variable importance plot, HTGD is found to be the most important variable in predicting match outcome followed by FGS0 and TOTALHP.

Confusion Matrix for Training

A screenshot of a computer

Description automatically generated with medium confidence

Confusion Matrix for Testing

A screenshot of a computer

Description automatically generated with medium confidence

The parameters have been set to maximize testing accuracy with the lowest possible loss in accuracy and overfitting with the training set.

So, from the confusion matrices we can see that the overall accuracy is 77% for the training set and 64% for the testing set. Among the three different classes, the algorithm is best at correctly predicting “Loss” for both the test set and the train set with sensitivity of 0.78 and 0.87 respectively. It is the worst at correctly predicting “Draw” for both the sets.

1. Provide a summary of key results from the three models used above and compare results. Which classification and prediction method do you find to be the best for betting on the match outcome and why? (10 points):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Train DT | Test DT | Train RF | Test RF | Train XGB | Test XGB |
| Accuracy (%) | 67 | 60 | 88 | 63.1 | 77 | 64 |
| Win Sensitivity | 0.69 | 0.59 | 0.93 | 0.67 | 0.85 | 0.70 |
| Loss Sensitivity | 0.79 | 0.75 | 0.93 | 0.81 | 0.87 | 0.78 |
| Draw Sensitivity | 0.51 | 0.46 | 0.75 | 0.37 | 0.54 | 0.37 |

Considering the overall testing accuracy of each model, the Extreme Gradient Boosting method has the best performance. Among the three, the decision tree performs the worst.

At the same time, it must be noted that the random forest implementation is the best at correctly predicting if the match outcome will be “Loss” with the highest test sensitivity. So, if someone wants to bet that the home team would lose the match, they could consider using the random forest because of its high sensitivity.