Team TOO HARD: 101C Final Report

Junpeng Jiang;Wenxin Zhou;Rosy Zhou

Data Cleaning

- After careful examination, we did not see any missing values in both training and testing datasets. However, we found some repetitive columns in the data when we went through the preliminary analysis of data.
- For example, we found values of "VT.TS.fga" are exactly the same as values of "HT.OTS.fga". "VT.TS.fga" represents total field goals scored by visiting team and "HT.OTS.fga" represents total field goals scored Home team's opposing team, which is just VT.
- The same logic applies to VT.TA.xxx == HT.OTA.xxx, VT.S1.xxx == HT.OS1.xxx. Therefore, we cleaned up all columns that has ".O" in it, this removes all repetitive columns.
- We also removed id, game id, data, in training and test dataset because these variables are unique in each row.
- After data cleaning, dim(train) = 9520,113 / dim(test) = 1648,112

New Variables Description(4 new variables)

Variable pht

- pht represents the prior winning rate of a team plays as a home team across all training data. For example, if team 1 played 100 games as home team in training data, and it won 50 games, all rows having team 1 as HT will have the hr with value 0.5.

Variable pvt

- Similar to pvt represents the prior winning rate of a team plays as visiting team across all training data. For example, if team 1 played 100 games as visiting team in training data, and it won 50 games, all rows having team 1 as VT will have the vr with value 0.5.

Variable winthisvt

- winthisvt is a home team's winning rate when it is against a specific visiting team, it is calculated using all training data.
- For example, if a row has HT=Team 1, VT=Team 2, winthisvt = 0.9, then it means of all games Team 1 plays as HT against Team 2 in training data, Team 1 wins Team 2 90% of the time.

Variable p

- Variable p represents the total prior winning probability for a team. It is the winning rate of this team as both the visiting team and home team. So p represents the overall winning rate of a team based on the data in train dataset.
- For example, if a team plays 100 games; 40 of them are home team and and 60 as visiting team. Among these 100 games this team wins 60 times. So the p for this team is 0.6.

Adding New variable Reasoning

- Our final goal is to predict whether a home team wins. We believe adding variables related to winning rate will assist in prediction.

- Our first step is to add each team's home team winning rate(hr) and visiting team winning rate(vr).This is a reasonable step to take, because these winning rate shows historically how probable a team win as either HT or VT. This could potentially forecast a team's win/loss in future games.
- To explore winning rate even further, we decided to look at the specific rivalry history. We believe historically how well team 1(HT) played against team 2 could be a very strong win/loss indicator of team 1's future game as HT against team 2.
- To start with, we first check that all teams exists in the home team and visiting team lists in traning and testing datasets. Then calculate the winning rates based on the following metrics and add them to the test and train dataset.

Description of our models.

- We applied a variaty of classification algorithms, including logistic regression, ridge regression, lasso regression, gradient boosting, random forest and partial least squares regression, and ensembling. While testing out the models, we tried to explore whether our added 4 variables significantly improves prediction. We basically keeps all 113 columns of cleaned data (so we don't remove any potentially useful information), add different combinations of (pvt,pht,p, winthisvt) and calculate misclassification rate.

Comparisions of the models and explaination

```r
Misclassification_Rate <- c(0.3206349,0.3174603,0.2853175)
MODEL <- c("Ridge Original","Ridge + pht/pvt/p","Ridge + all four")
result <- cbind(MODEL,Misclassification_Rate)
print(as.data.frame(result))
```

```
##               MODEL Misclassification_Rate
## 1    Ridge Original              0.3206349
## 2 Ridge + pht/pvt/p              0.3174603
## 3  Ridge + all four              0.2853175
```

- Among all the models we have tried, ridge performs best. It is resonable because there are too many variables and ridge and lasso are very effect methods to give penalty on too many variables. They are more restrictive and less flexible than the least quares.
- Our finding was that (winthisvt,hr,vr,p) all improves prediction, and among all models ridge regression performs the best. Another finding was that, by adding winthisvt, we once had reached the score of 0.71 in our own testing(by spliting training data). However, adding winthisvt made our trial submission scores very low(0.638). We reached a conclusion that this variable winthisvt is so strong that it overfits. We finally decided to remove winthisvt from our submission model.
- Our final model is ridge regression:HTWins ~ 113 remaining variables + pht + pvt + p