Predicting Winners of Basketball Games

Tuyen Duong, Dara Hashemi, Ki Hyun Park

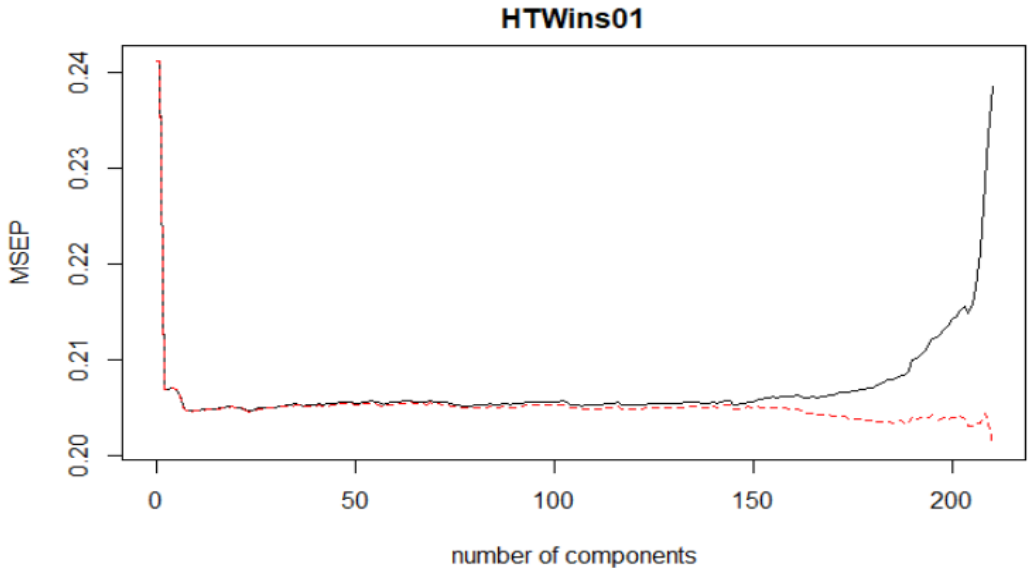
Organizing Data

- Removed variables which did not play a big role in our data (id, gameID, HT, VT, HT league, VT league)
- Created a binary response variable
- Removed perfectly collinear (redundant) variables
- Tested methods: Ridge Regression, Lasso, PCR, LDA, Logistic Regression, KNN-cv, Boosting

Top 5 Methods

Best Model-PCR

Our best model utilized Principal Component Regression accompanied by cross validation



Why PCR?

- Resolves Correlation Issues
- PCR reduces complexity while capturing the variance explained by variables
  - ex) PCR resolved multicollinearity between the variables “fgm” and “fga”

- Works very well with large data sets

- One of the higher accuracy rates among all the methods

Summary

Accuracy Rate: about 70%

What we tried: Different generalized models and classification methods.

What worked: Ridge Regression and Principal Component Regression

What didn’t work: Predicting using only more recent data