A Survey of Advanced Modeling Techniques for Forecasting College Football Game Outcomes

Charles South[1], PhD and Edward Egros[2], MS

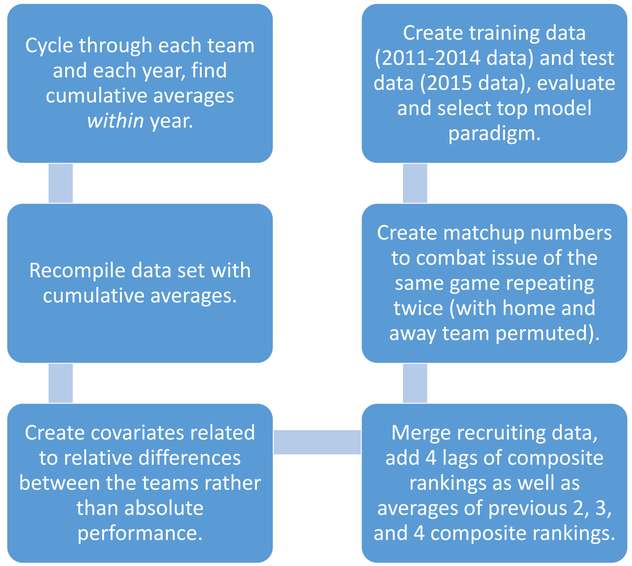
Purpose:

Can we use box score, point spread, and recruiting data to accurately forecast outcomes of college football games using modern machine learning and Bayesian modelling approaches?

Data:

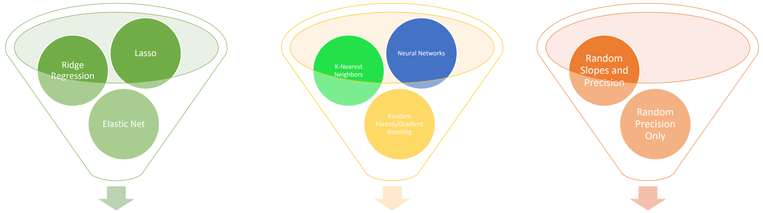
The data used for this college football forecast consisted of 4,339 games between Football Bowl Subdivision (FBS) teams between the 2011 and 2016 seasons[3]. The data set was augmented with composite team ratings[4] for each recruiting class dating back to 2008 to account for seniors in the 2011 season.

Data Prep/Manipulation Process:



[3]http://www.seldomusedreserve.com/?page_id=8805
[4]http://247sports.com/Article/247Rating-Explanation-81574

Modelling Paradigms:



Penalized Regression | **Non-Parametric** | **Bayesian**

Modelling Details:

- Dependent Variable: point differential
- R packages utilized: *glmnet*, *randomForest*, *FNN*, *nnet*, *xgboost*.
- Where applicable, the *caret* package was used to tune the models.
- For the Bayesian models, a the model was compressed to only include variables selected by the lasso.
- Bayesian framework:

predicted point difference for team i on game j

$$y_{ij} \sim Normal(\mu_{ij}, \tau_i),$$
$$\mu_{ij} = \beta_0 + \beta_1 X_{1ij} + \cdots + \beta_p X_{pij},$$
$$\beta_p \sim Normal(0.001, 0.001),$$
$$\tau_i \sim Gamma(1, 1)$$

random precision for team i

Lasso Retained Variables		
Yards Per Pass Attempt (YPPA)	Yards Per Rush Attempt (YPRA)	Rush Attempts
Total Yards	Yards Per Play (YPP)	Turnovers (TO)
Opponent Points Scored	Opponent YPRA	Opponent Total Yards
Opponent Turnovers	Opponent Penalty Yards	Average Point Differential
Opponent Offense Passing Yards	Opponent Offense YPRA	Opponent Offense Total Yards
Opponent Offense YPP	Opponent Defense Total Rush Yds	Opponent Defense YPRA
Opponent Defense Total Yards	Opponent Def YPP	Opponent Defense TO
Opponent Average Points Differential	Difference in Win Percentage	Home Field Advantage
Composite Ranking, Lag 2 (CR)	Average CR (Last 2 Years)	Average CR (Last 3 Years)

Final Predictions:

All generated models made predictions for the same game twice. Our decision rule was designed to be as intuitive as possible – if both predictions resulted in the same team being favored, the favored team was deemed the predicted winner. If the predictions diverged, the team with the larger predicted point differential was deemed the predicted winner.

Summary of Results:

Model	Overall Prediction	PPV	NPV
Lasso	**75.0%**	**77.3%**	**72.4%**
Random Forest	74.1%	78.9%	69.0%
Gradient Boosting	72.9%	76.9%	68.5%
K-Nearest Neighbors	69.0%	71.3%	66.3%
Neural Network	71.6%	75.8%	66.4%
Bayesian	74.1%	76.5%	71.5%

Conclusions/Future Work:

- The lasso was the top performer, slightly edging out the Bayesian model and random forests.
- In terms of variable importance, composite recruit rankings are important, though it appears it takes at least 2 years for players to develop.
- Volume and efficiency are important – as is home field advantage.
- Of the 27 predictors identified by the lasso, 15 were directly related to opponent strength. Schedule matters.
- An ensemble method may improve predictive ability.

[1]South Statistical Consulting and Analytics, LLC
[2]Fox 4 News (Dallas), InsideSportsAnalytics.com