March Data Crunch Madness

Hosted and Sponsored by: Center for Digital Transformation, FBAS, Deloitte.

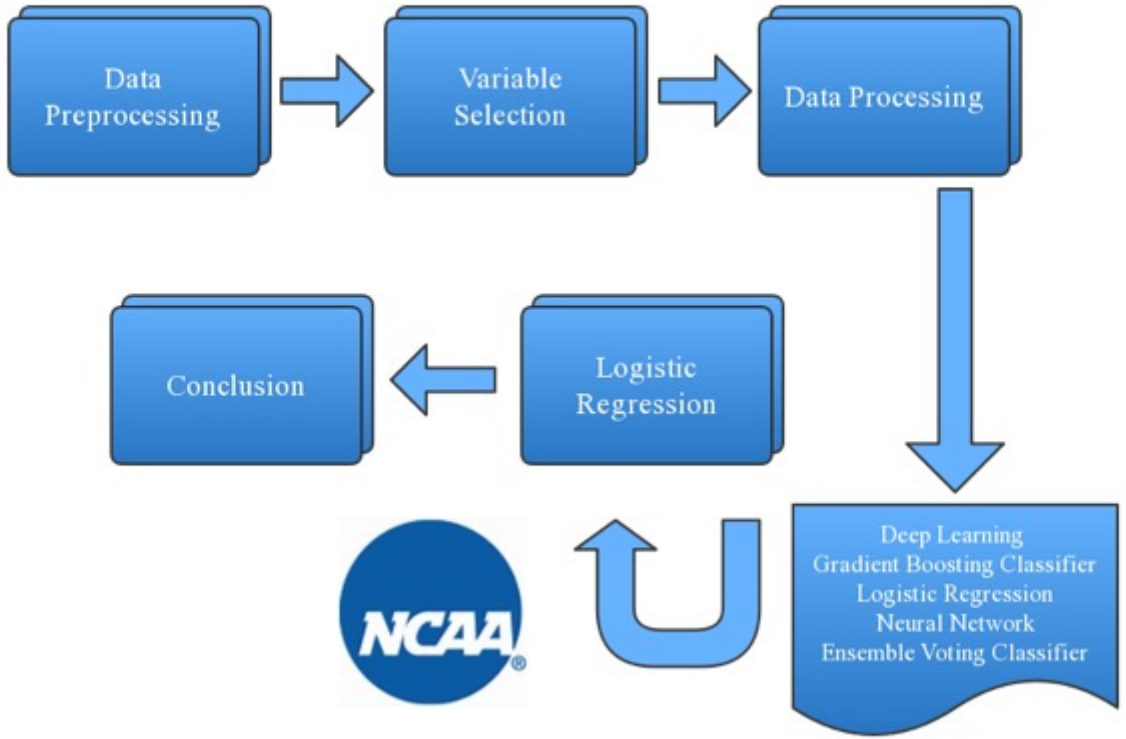
March Data Crunch Madness

Team Orange

Yichen Pan Ziyue Zhong Haojunzhi Yu Teena George

March 21, 2018

Project Procedures



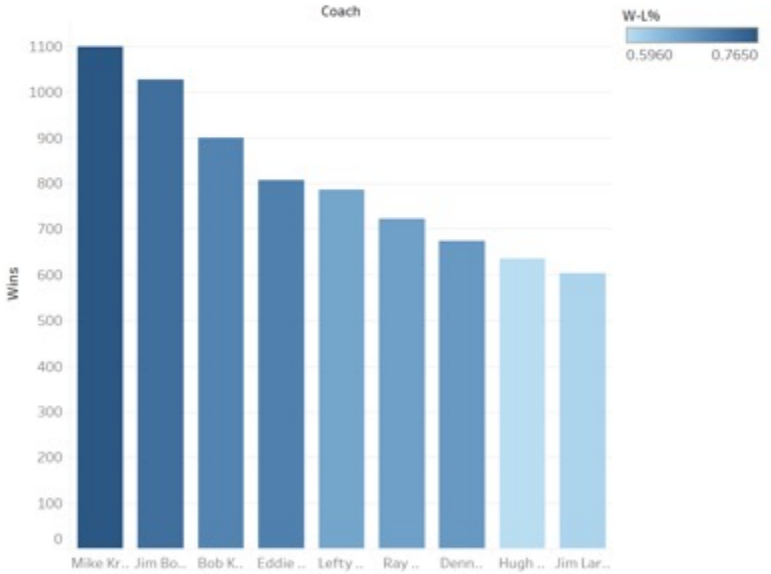
Data Description

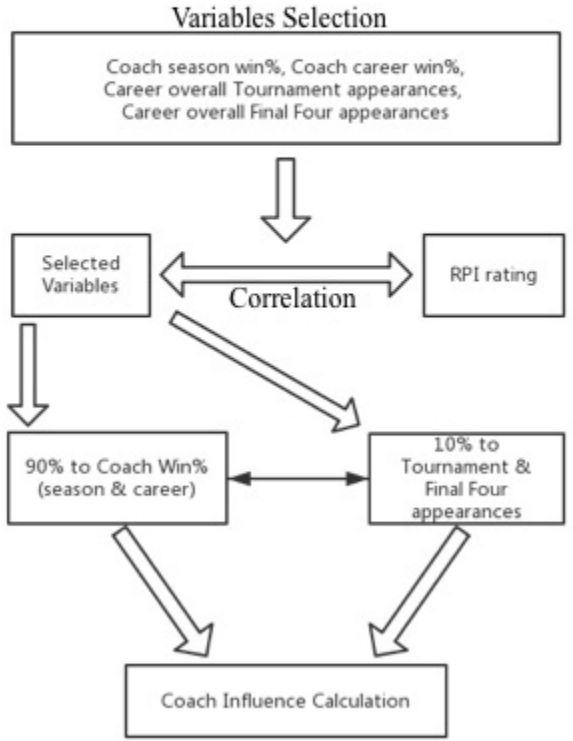
Variables	Calculation
E(W%)	How many games a baseball team "should" have won based on the number of runs they scored and allowed.
Log 5	The probability that team 1 will win a game, based on the winning percentage of Team 1 and Team 2
EFG-OEFG	How much a team exceed its opponent on effective field goal percentage
TPP-DTPP	Turnover committed per possession minus defensive rebounding caused per possession
SRS	A team rating that takes into account average point differential and strength of schedule. The rating is denominated in points above/below zero
SOS	A rating of strength of schedule. The rating is denominated in points above/below zero.
Adj_win	(0.8*Home Win + 1.2*Away Win + Host Win)/Total Games
Host_win	Host wining percentage
Coach Influence	Combination of Coach season win%, Coach career win%, Career overall Tournament appearances, Career overall Final Four appearances
Star Player	Players whose win share in the top 20 at regular season assume as star player

Data Pre-processing

- Process Missing value by using average RPI rating from all other variables
- Add variables from other sources (https://www.sports-reference.com/cbb/)
- Normalize data by using normalize function from sklearn.preprocessing package in python

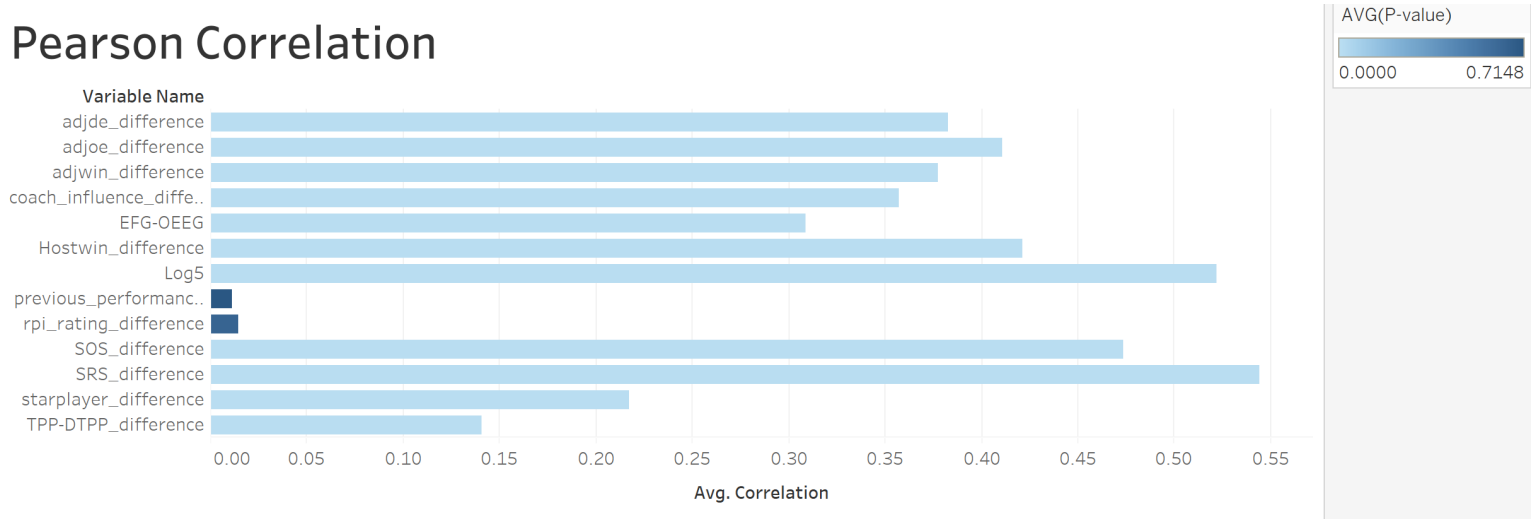
Coach Influence (New created variable)





Variable selection

- Use decision tree from SPSS modeler to generate first 15 important variables
- Select data based on Pearson Correlation between each variable and ‘result’

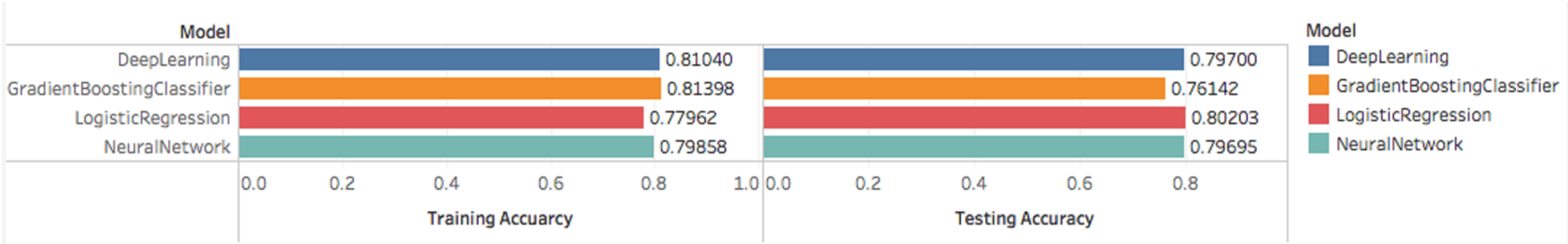


Applied Models

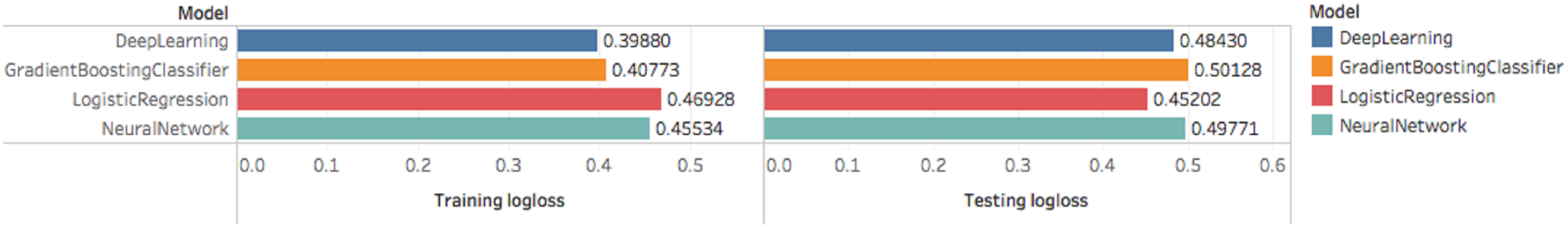
- Deep Learning
- Logistic Regression
- Neural Network
- Gradient Boosting Classifier

Model Comparison

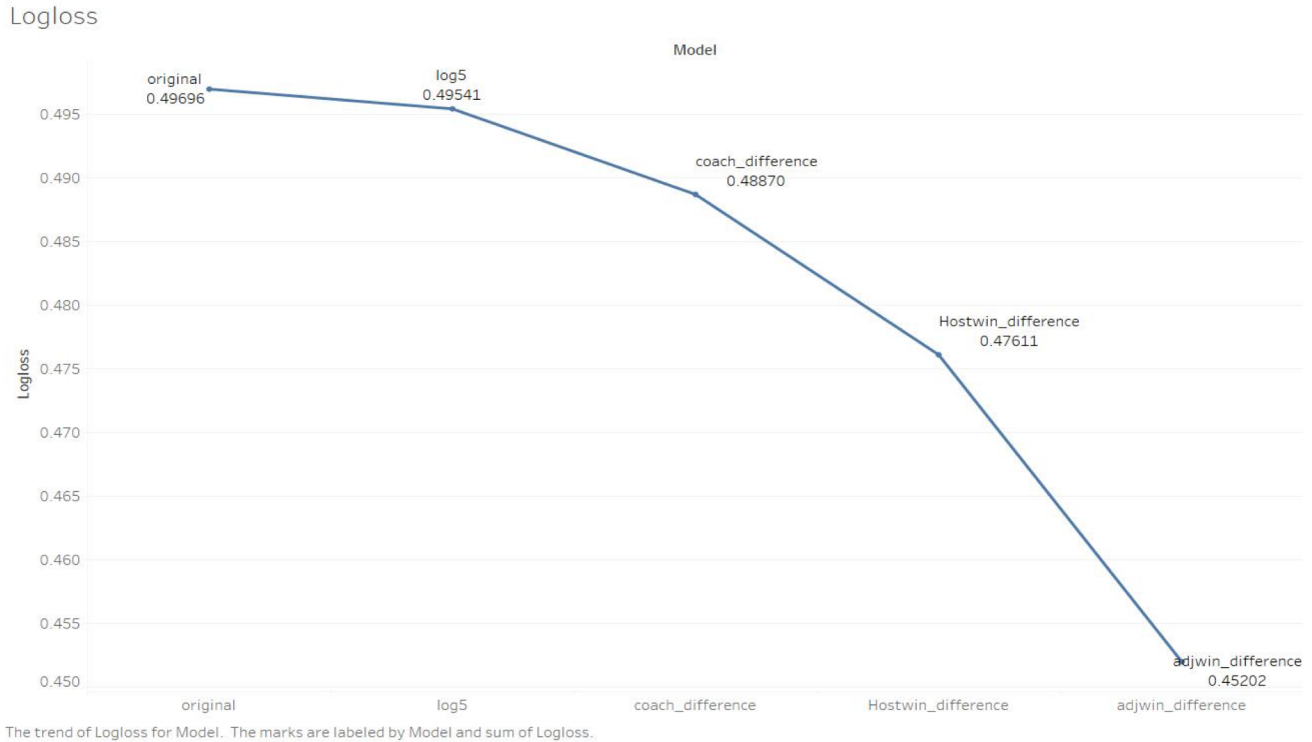
Accuracy for each model



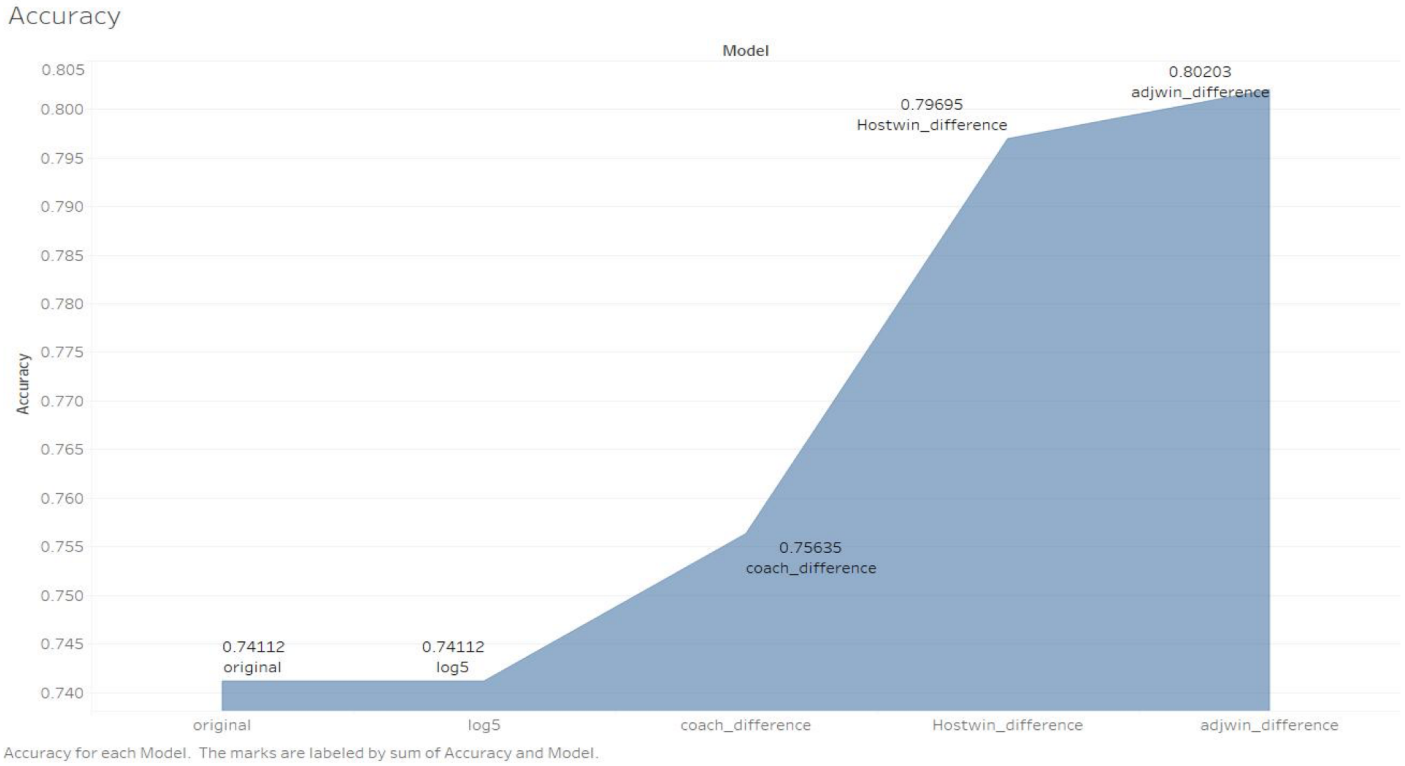
Logloss for each model



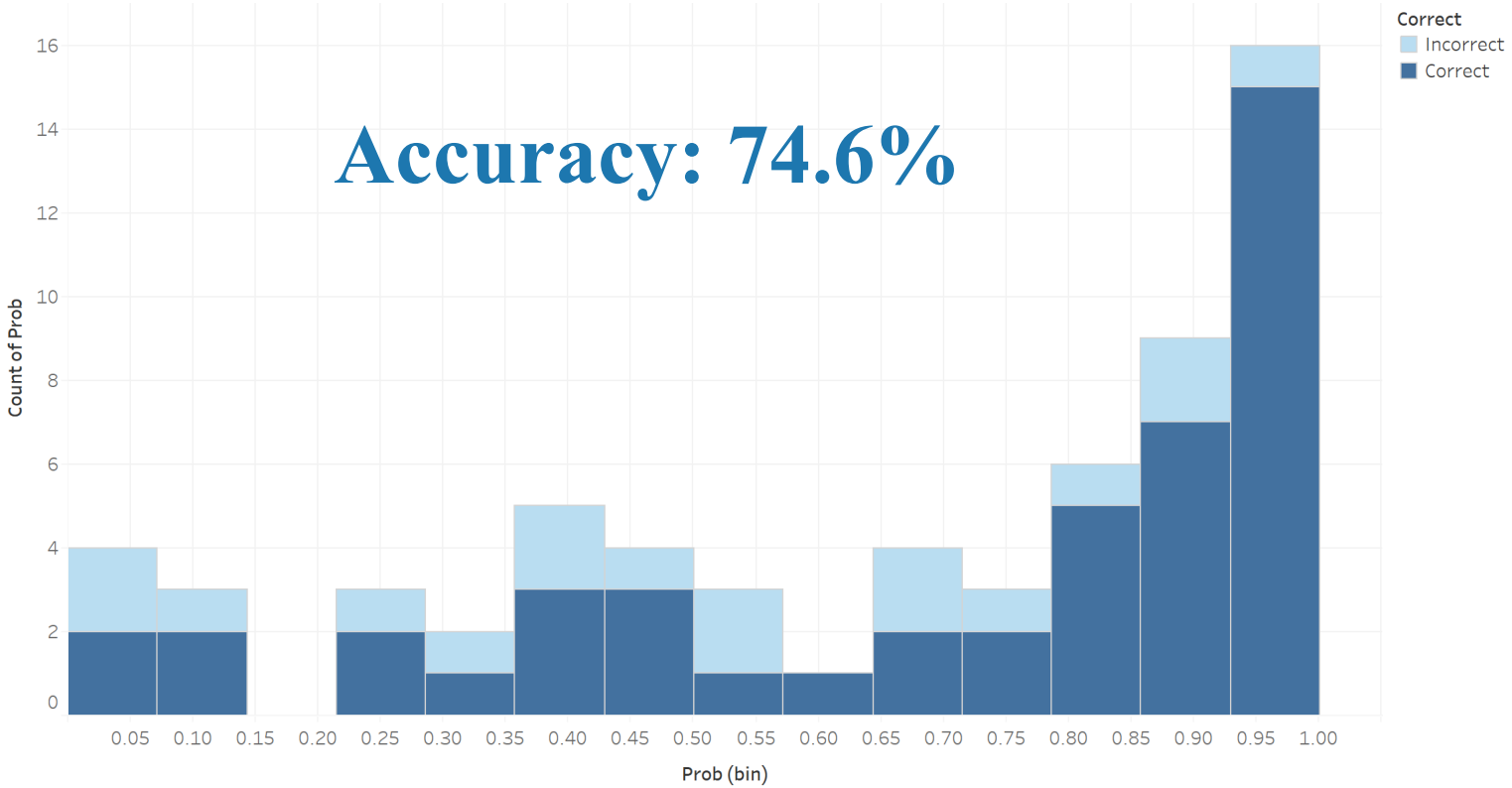
Log loss decreased with added variable



Model accuracy improved with added variable



Predicted probability distribution and correctness



Model Combination

- We find Ensemble Machine Learning Algorithms in scikit-learn by using python.
- We use voting function to build multiple models and simple statistics to combine our predictions from Neural Network , Gradient Boosting Classifier , Logistic Regression and Artificial Neural Network.
- However, the result is not better than our original result from Logistic Regression.
- Finally, we decide to use Logistic Regression as our model to generate probability and result of 2018 NCAA March Madness

Final Prediction

Final 4 prediction

- Kentucky University vs. Xavier University

- Villanova University vs. Kansas University


National Championship prediction :

- Kansas University