ETC3250 Project

The Knitting Club

James Nguyen - 27839397
William J Dunne - 27833534
Patrick Graham - 25942476
Chew Kian Wei - 28985362
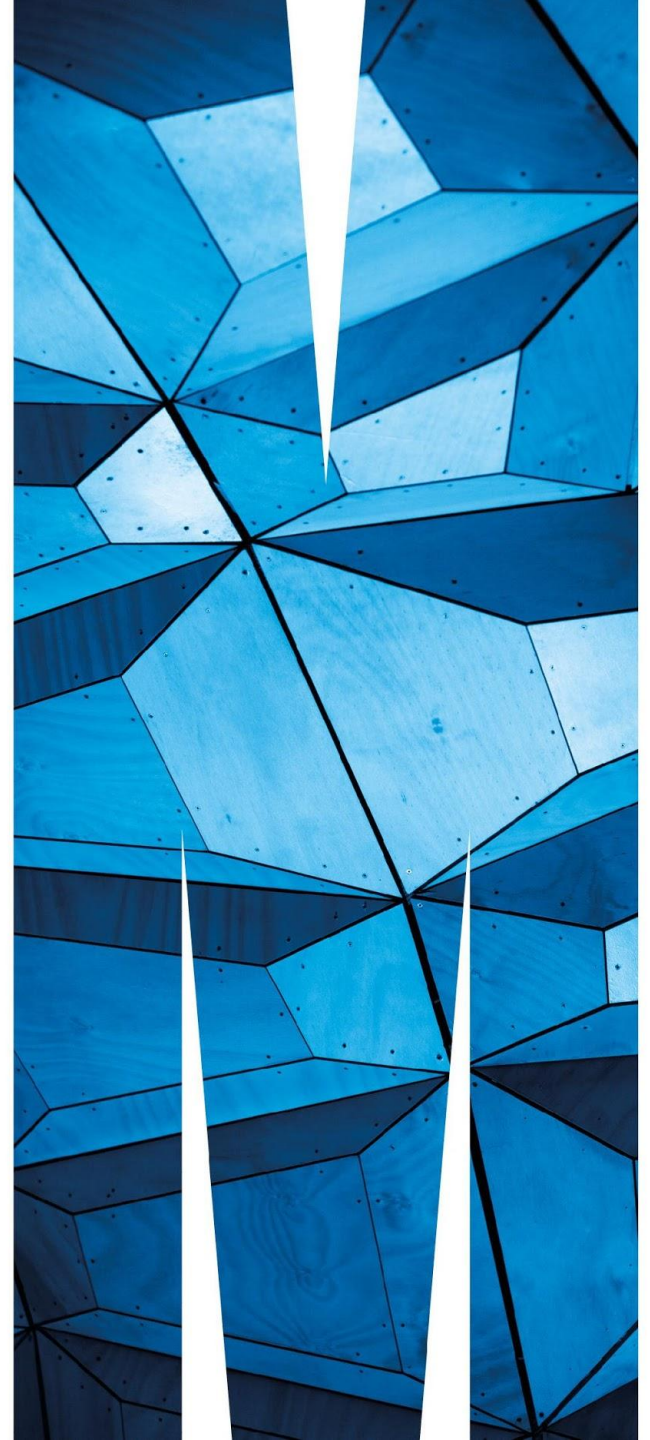
The Data

- Some winners marked outside baseline or sideline
- Most winners hit close to the net
- Forced areas most commonly occur when the ball is close to the sideline
- Unforced errors usually have high ball speed

Source: https://www.gristforthemills.com/posts/2018-03-21-ao-to-ai/

Our Approach

- Different Models
  - RF
  - XGBoosted trees
  - Stacked Ensembles
- Feature Engineering

3
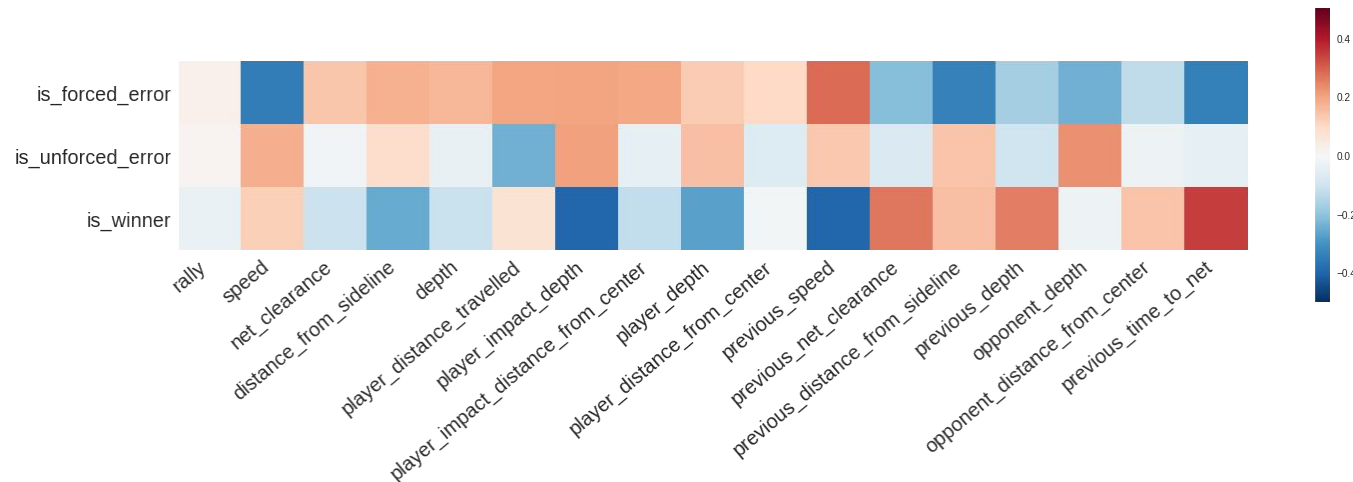
Kaggle Results

Model	Feature Engineering	Extended Features	Prediction (%)
Random Forest	N	N	90.76
XGBoost	N	N	91.09
XGBoost	Y	N	91.71
h2o.ai Ensemble	Y	N	92.05
h2o.ai Ensemble	Y	Y	92.62

Discussion

- The random forest model performed surprisingly well
- Speed difference, net clearance difference and difference in distance between players were most import features engineered
- XGBoost performed well, but ensembles were better
- 69 different models used, with 20% of the data used for cross-validation

Questions?