ETC3250 TENNIS PROJECT

Federer's Friends – Nicolas Alexiou, Yijia Pan, Wen Zheng Tan, Gao Kassia Yue

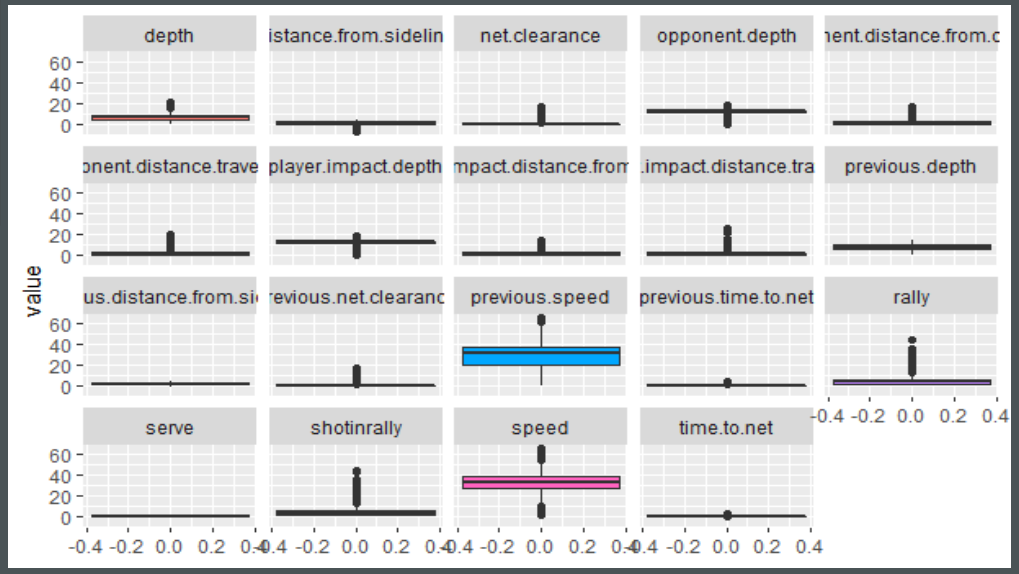
Introduction

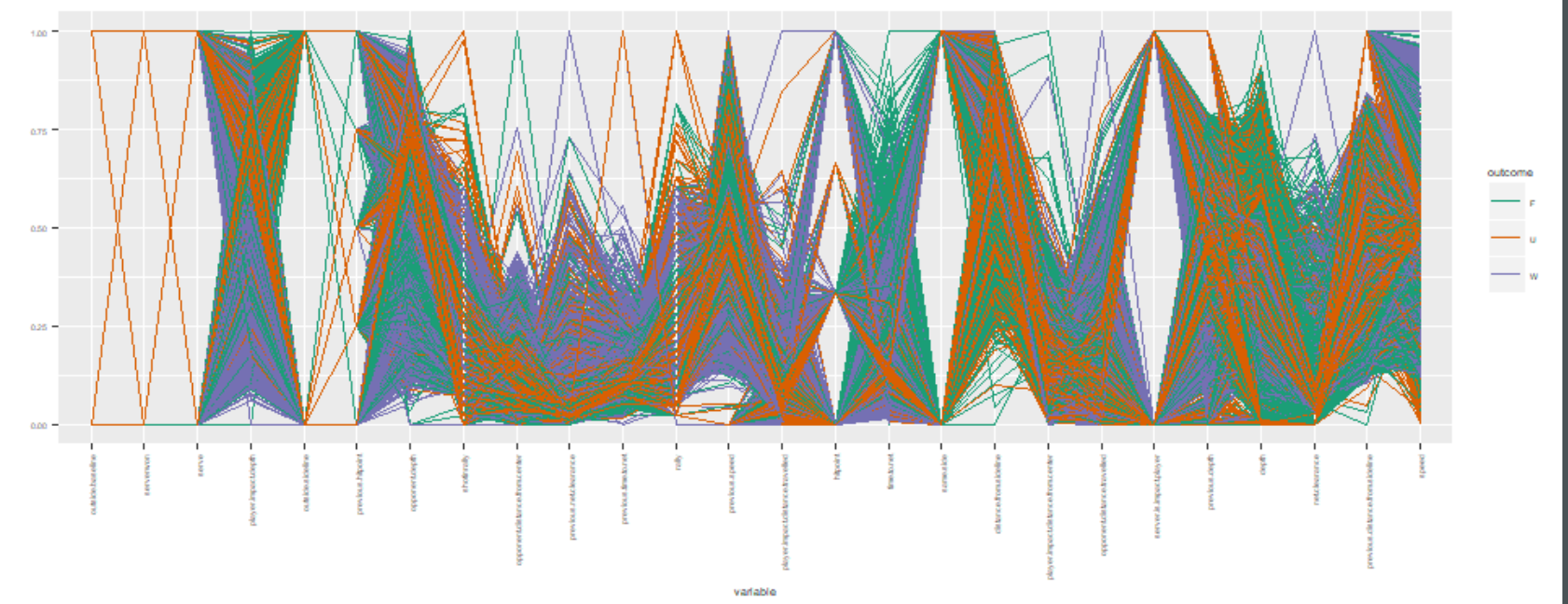
Aim

To produce a model that can reliably predict the outcome of a tennis point based on automatically observed features of the point, to move to a future of automatic point coding

Data Overview and Visualisation

- Data : the 2017 and 2018 Australian Opens – 49,914 observations of 31 variables
  - Four identify the point, One is the outcome
  - The other 25 are characteristics of the point that can be used to predict future points





- Little variance in most variables, apart from the speed of the point-ending shot
- Some shots required players to travel significantly more than in most points
- Some rallies were significantly above average in terms of length – typical of Grand Slam matches
- Points where players were further from the centre ended in forced errors
- The longest rallies ended in unforced errors, the shortest ended in winners
- Points where the penultimate shot had a very high net-clearance ended in winners – likely smashes

RF, XGBOOST, GBM

Chosen as these are three models commonly used in winning submissions on Kaggle

Methodology

1. Random Forest

Reasons to use:
- Simplicity
- Flexibility

Process:
1. Use bootstrapped training samples to form decision trees
2. Randomly chooses a subset of predictors to use as split candidates each time
3. Decision trees are combined

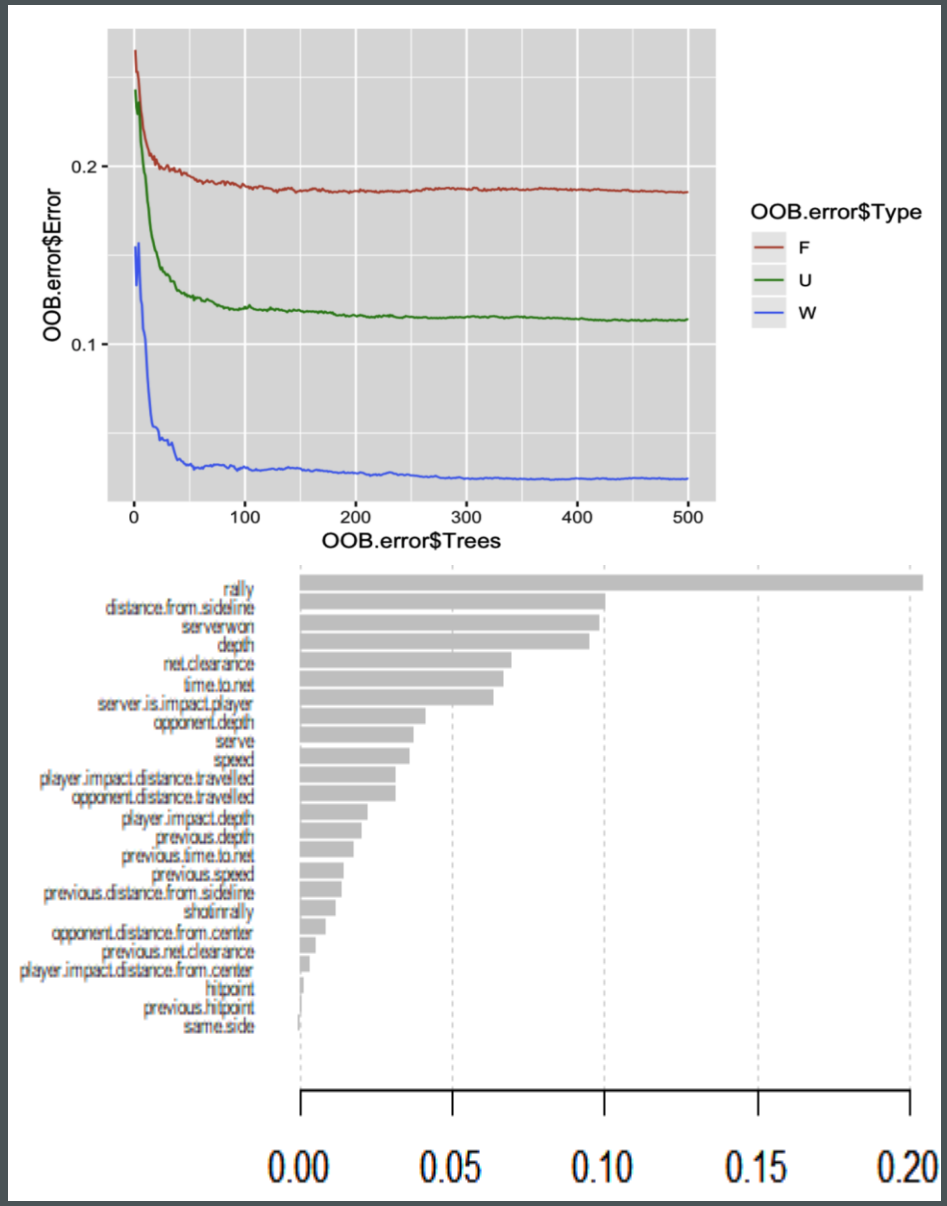
Result:
Public Kaggle score:
Without gender:
0.90901
With gender:
0.91050

2. XGboost

- Reason to use: improves on random forest and gbm
  - Provision for regularisation, to avoid overfitting
  - Prunes trees backwards, stopping once improvement is insignificant
  - Internal cross-validation and handling of missing values

For loop used to search for the best parameters
Result: higher Kaggle Public score of 0.91904

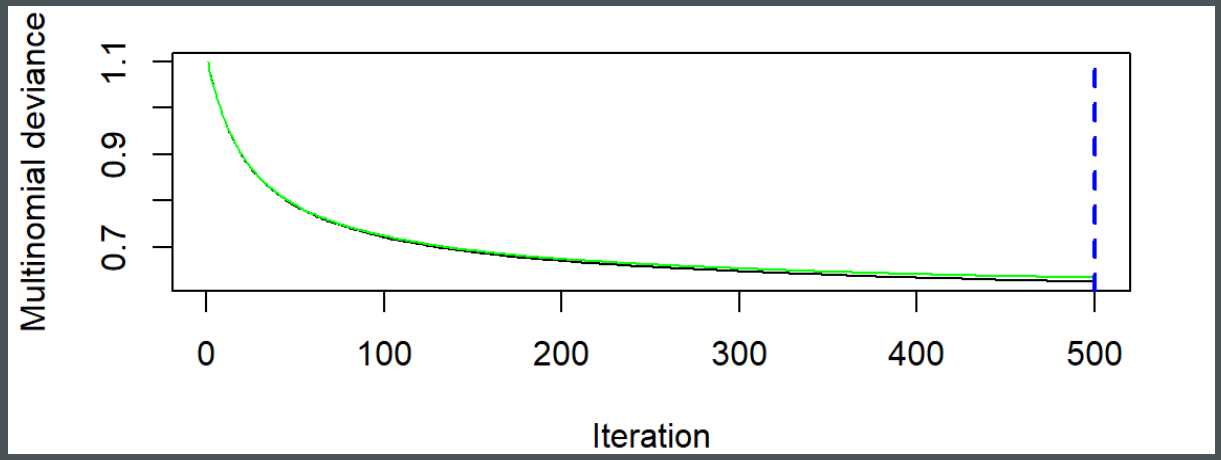
Plot showing the importance of each variable:
The higher the score, the greater the relative contribution of the corresponding variable to the XGboost model – rally most important by far

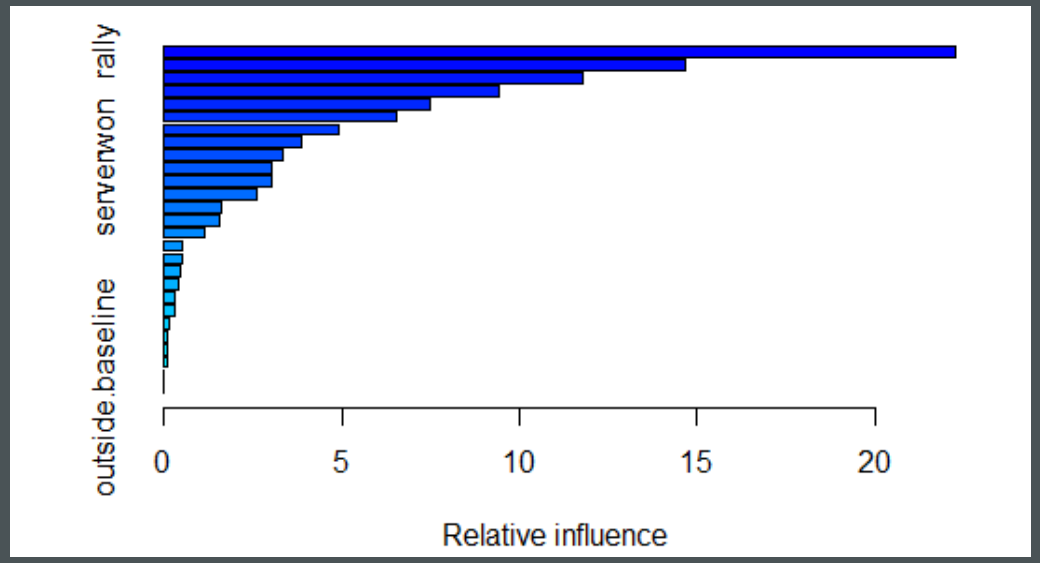


Methodology & Results

3. Gradient Boosting Machine (gbm)

- Process:
- Creates a decision tree (the observations are all weighted equally)
- Weights the data by giving more importance to hard-to-classify observations
- Creates a new model then consisting of both trees, using cross-validation
- 500 iterations were chosen to optimize the model & rally is the most important variable





- Result: Public Kaggle score of 0.75646
- Reasons:
  - Deficiencies compared to xgboost
  - Difficulty of tuning gbm
  - Sensitiveness to overfitting

Discussion

- **Results** – Confusion tables for the random forest, XGBoost and gbm models respectively
- Training error rates: 11.24% for random forest, 10.12% for xgboost, 26.26% for gbm

pred_random_tr

	F	U	W	Sum
F	4242	736	251	5229
U	611	5310	111	6032
W	49	96	5064	5209
Sum	4902	6142	5426	16470

[1] 0.1123862

xgbpred_xgb_tr

	F	U	W	Sum
F	4337	780	112	5229
U	606	5377	49	6032
W	51	69	5089	5209
Sum	4994	6226	5250	16470

[1] 0.1012143

gbmPredictions

	F	U	W	Sum
F	3511	1124	594	5229
U	712	4575	745	6032
W	309	811	4089	5209
Sum	4532	6510	5428	16470

[1] 0.2625987

- xgboost - the superior modelling technique (due to provision for regularisation, unique pruning method and handling of cross-validation and missing values)
- Random forest outperforms gbm (due to the relative easiness of tuning and relative unlikeliness to overfit)
- xgboost model with the data separated by gender had lower Public Kaggale score compared to the model on the initial data
  - Maybe the extra complexity added by splitting the data outweighed the extra information
- Rally the most important variable, followed by distance from sideline and server won
  - The longer a rally, the more tired and cautious players become and the more likely a point will end in an unforced error

Conclusion

1. xgboost is the superior method with regards to predicting the outcome of a point in tennis, compared to random forest and gbm

2. the length of the rally is the most important variable in determining its outcome.
3. With a success rate of around 90% (depending on the method of measurement), it is not quite ready just yet, but suggests we could one day be looking at a future free of manual point coders in tennis!