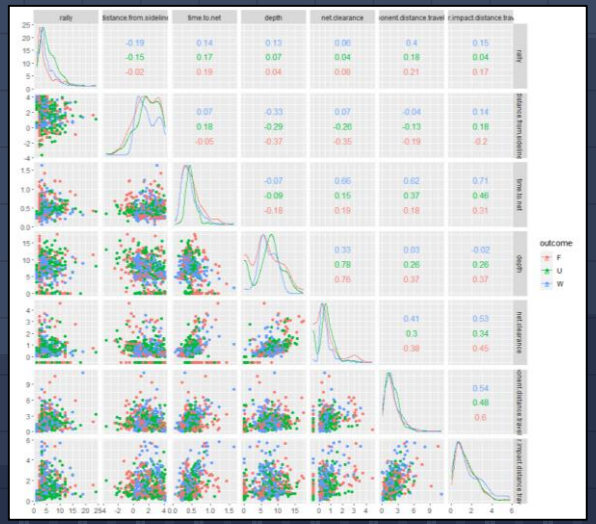
Using Data Science to Revolutionize the Sport of Tennis

Team Monba

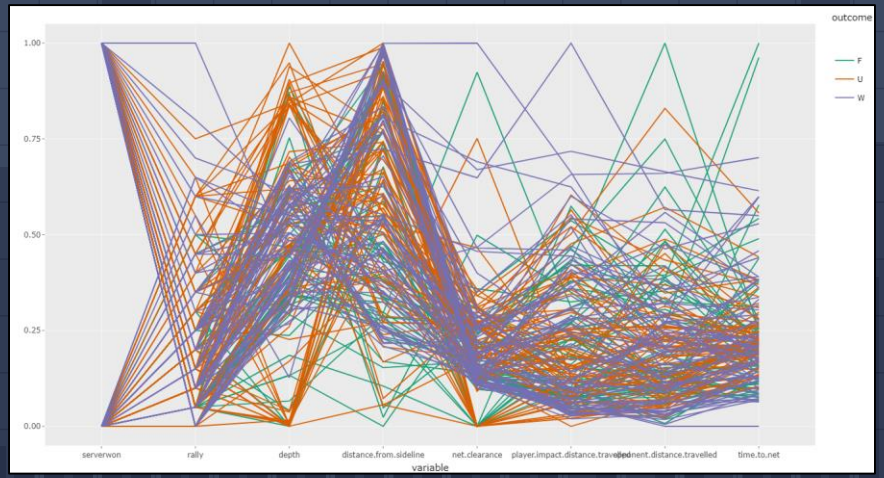
ETC3250 Project

Methodology

- **1. Partitioning:** 80/20 split of training data.
- **2. Sample Code:** Used to establish a baseline.
- **3. Visualisation:** Investigating most important variables from random forest model.



Correlation Plot



Parallel Coordinates Plot

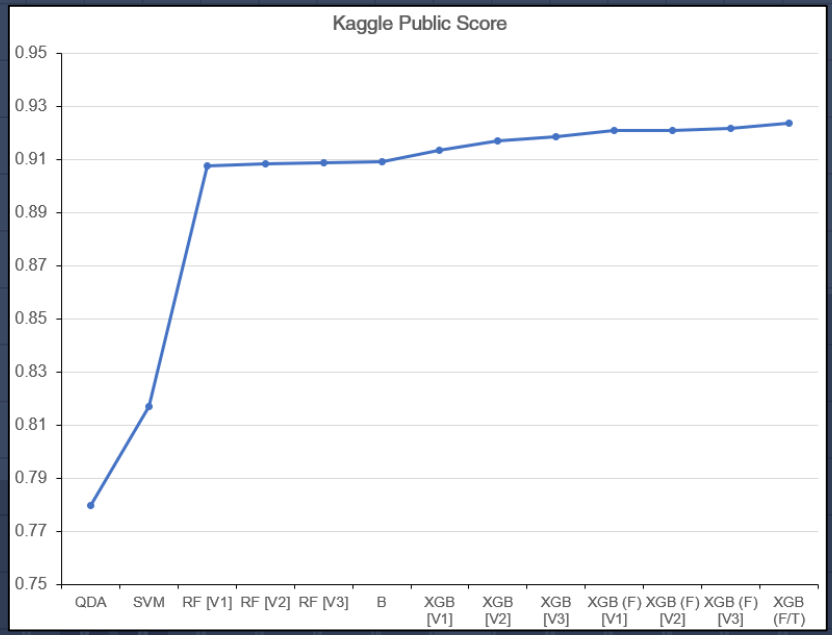
Methodology

- **4. Model Selection:** XGBoost used for final model.
- **5. Feature Engineering:**
  - $rally.calc = rally - shotinrally$
  - $rally.prop = \frac{rally - shotinrally}{rally}$
  - $speed.calc = speed - previous.speed$
- **6. Tuning:** Final parameters:

max_depth	min_child_weight	gamma	subsample	colsample_bytree	eta (learning rate)
6.000	1.000	0.000	0.975	1.000	0.200

Results

Model	Score
QDA	0.77963
SVM	0.81706
Random Forest	0.90864
Boost	0.90916
XGBoost	0.91867
XGBoost (with features)	0.92104
XGBoost (with features and tuning)	0.92349



Kaggle Public Scores for Different Models

Improvements

- **Data preparation:** More time spent cleaning and filtering dataset.
- **Alternative models and software:** Neural networks, TensorFlow, other boosting algorithms (AdaBoost, Gradient Boosting).
- **Feature engineering is key:**
  - Further investigation into additional, highly significant variables.
  - Exploration into automatic feature engineering using deep feature synthesis and other automation methods.