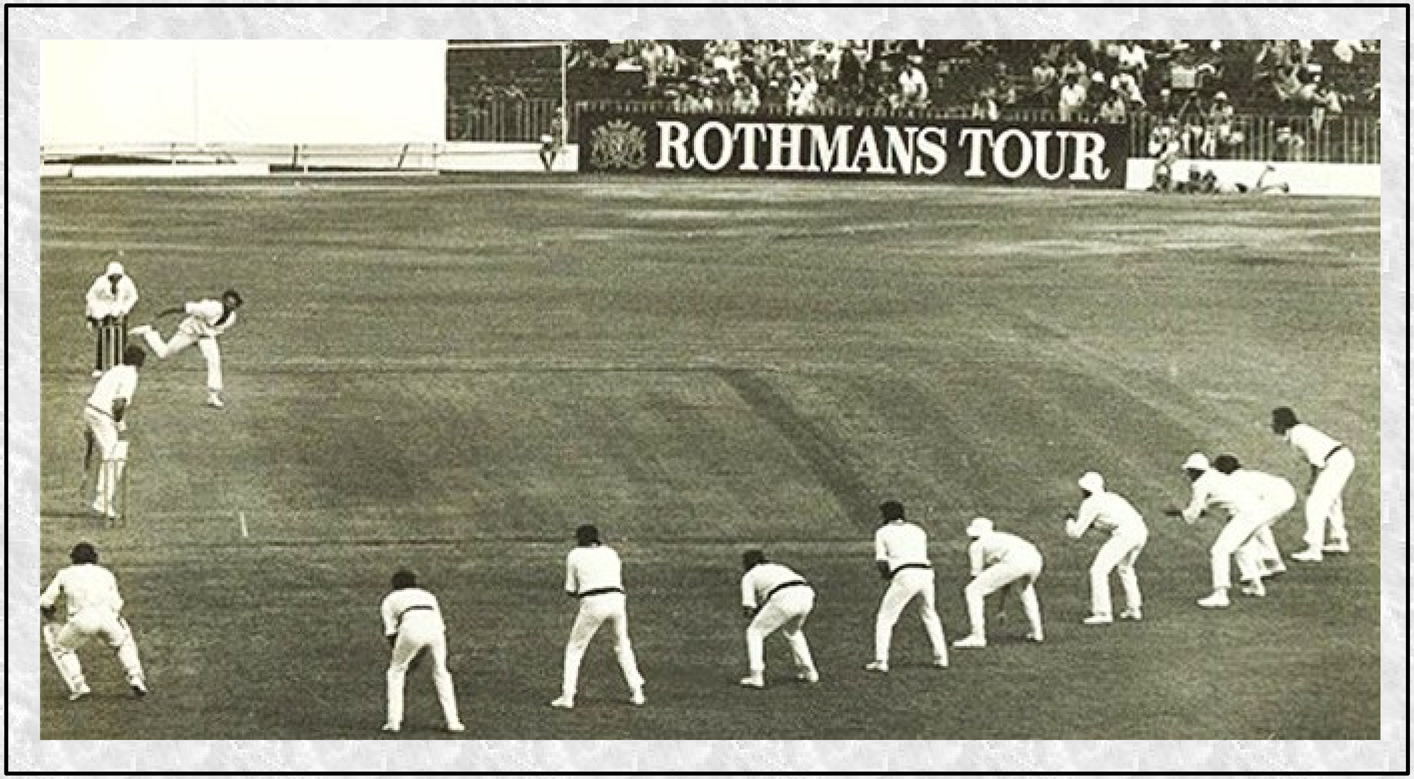
Bowling Performance Prediction in Test Cricket



Prantik Ghosh

LinkedIn: www.linkedin.com/in/prantik-ghosh-899701106/

Github: github.com/prantik-ghosh/bowling_performance_predictor_in_test_cricket

Motivation

“In no other game does the law of averages get to work so potently, so mysteriously" – Sir Neville Cardus.

- Outcome of a match is very hard to predict.
- An individual’s performance is even harder to gauge.
- In test cricket, a bowler’s contribution is absolutely paramount.
- Ability to predict it will help coaches/captains to pick the bowling squad more efficiently.

Data Retrieval

Match wise data downloaded from the web

Years and Data Volume
7000+ records from years 2000-2017

Bowling statistics
– Overs bowled
– Runs conceded
– Wickets taken

Bowler Information
– Name
– Country
– Bowler type (pace/spin)
– Bowling arm

Other Information
– Start date
– Opposition
– Ground/Stadium
– Home or away

Feature Engineering

Following features were extracted from the base data:

- Bowler’s performance in each of the last 5 years (Average #wickets captured per match)
- Bowler’s “popularity” in each of the last 5 years (#matches played)
- Bowler-opposition interaction
- Bowler-home/away interaction
- Bowling type-opposition interaction
- Bowling type-ground/stadium interaction

Target (Grouping by Player)

Player-1, match-1 prediction

Player-1, match-2 prediction → Player-1 aggregate

Player-1, match-3 prediction

Player-2, match-1 prediction

Player-2, match-4 prediction → Player-2 aggregate

Player-2, match-5 prediction

...

Setting up the Baseline

- Last year’s performance is generally a very good indicator of a player’s current year’s performance.
- Average number of wickets taken per match is a straightforward measure of performance.
- Hence, the baseline for each player is set as follows:

Baseline = (Avg #wkts/match last year) x (#matches this year)

Cross validation

Purpose	Training Data (year range)	Test/Validation (year)
Validation set 1	2005-2010	2011
Validation set 2	2006-2011	2012
Validation set 3	2007-2012	2013
Validation set 4	2008-2013	2014
Validation set 5	2009-2014	2015
Validation set 6	2010-2015	2016
Final Testing	2011-2016	2017

Model Fitting Strategy

Models to compare
– Linear Regression
– Random Forest
– Gradient Boosted Decision Trees

Metrics to use
– Explained Variance
– Mean Squared Error

Cross validation
Cross validate using the six training/validation set

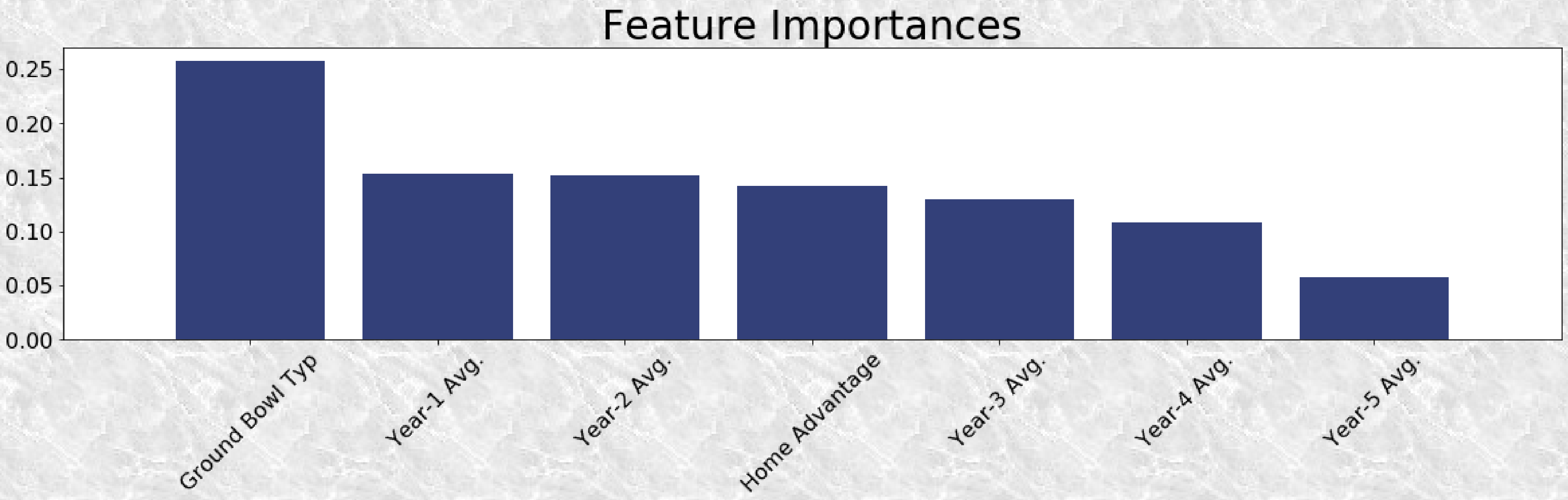
Picking the winner
Pick the winning model based on average test score

Results: Feature Importance

After testing with different models, only the following features proved to be significant:

- Bowler’s past performance
- Bowler-home/away interaction
- Bowling type-ground interaction

The final model run revealed the following feature importances:



Results: Model Selection

- All three models delivered best test score for some validation set or other.
- Gradient boosting did slightly better on average.
- Gradient boosting model’s score was always close to the best model when it was not the best.

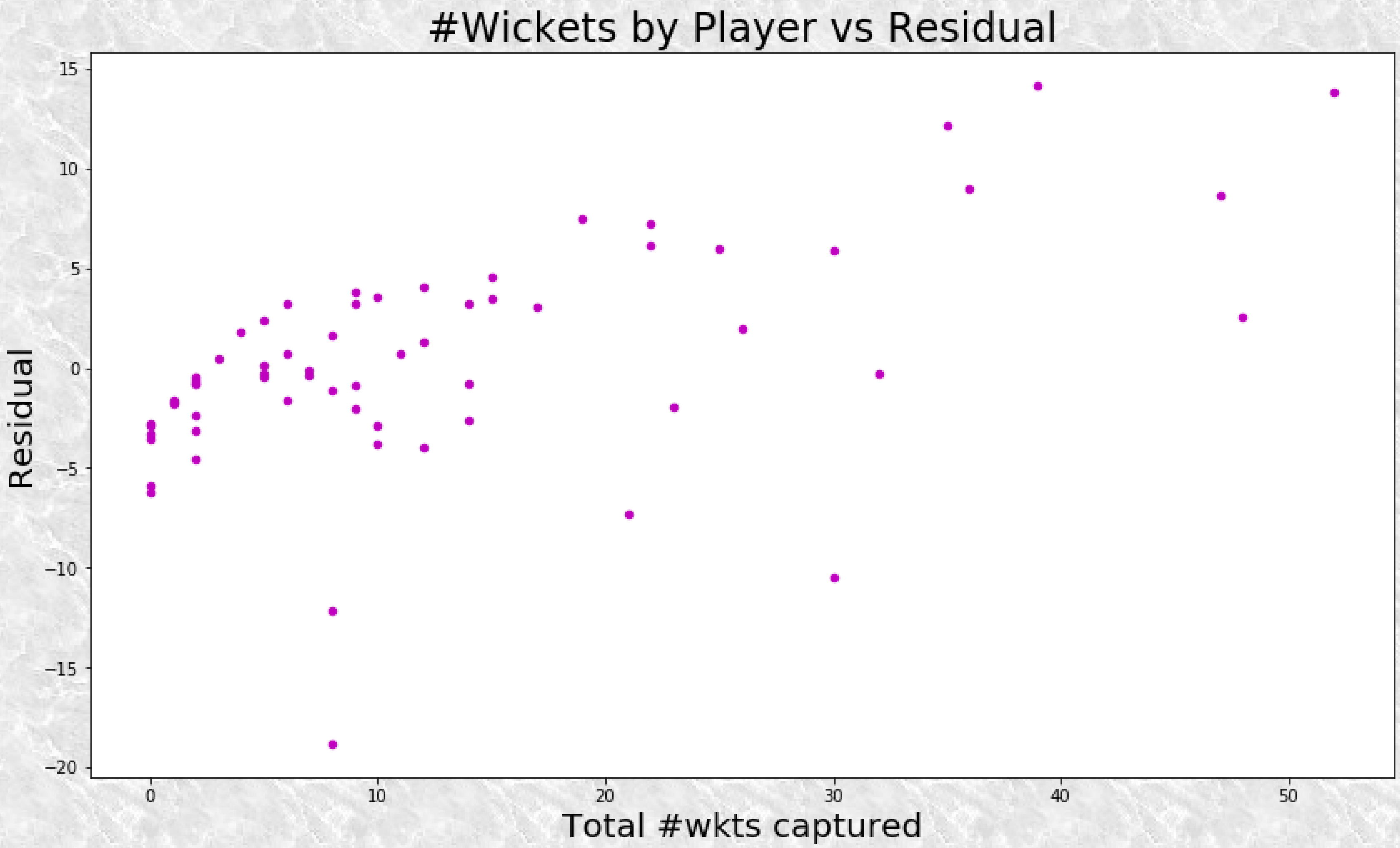
Gradient Boosting Decision Tree Regressor was chosen as the winning model!

Final Prediction Results

The optimized Gradient Boosting model, when trained on 2011-2016 data and tested on 2017 data returned the following scores:

- An Explained Variance of 81.4% against a Baseline Explained Variance of 65.9%
- A Mean Squared Error (MSE) of 30.2 against a Baseline MSE of 59.9
- Final results in line with results in the validation sets

Residual Plot Diagram



Next Steps

- Consider bowler sub-type (What kind of spinner? A leg-break, an off-break, a left-arm-orthodox or a Chinaman bowler?)
- Consider weather data and how it would interact with bowler type.
- Perhaps consider domestic performance for those bowlers who are new to test cricket.

Questions?

Prantik Ghosh

LinkedIn: www.linkedin.com/in/prantik-ghosh-899701106/

Github: github.com/prantik-ghosh/bowling_performance_predictor_in_test_cricket