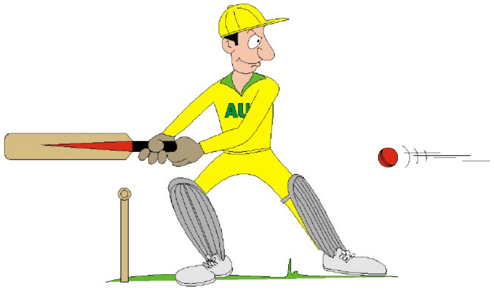
Can we predict a player's [Strike-Rate] in Cricket?

Harish Vasudevasarma
Data Scientist @ Metis

What's Cricket?

- 17 countries play this sports internationally.
- 2nd most watched sports after Soccer.
- **A One-Day International (ODI) game:**
  - **50 overs = 300 balls delivered to a team, consisting of 11 batsmen.**
  - **Game duration: ~6hrs / day.**

Motivation



- Good batsmen are like quarterbacks.
- A batsman's skill in ODI is judged by his/her,
  - Strike Rate: (Runs)/(100 balls) served.
- Predicting *Strike Rate per player* is key in *game outcome prediction*.

Key issues that could affect prediction

- Complex rules governing the game
  - Batting order, runs, bowling errors, etc

- Number of external parameters:
  - Weather related, Pitch

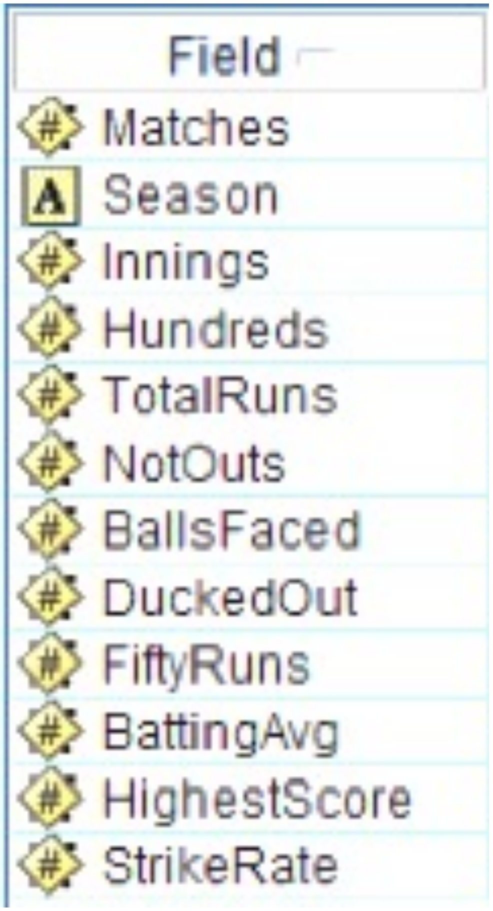
- Different formats of games:
  - ODI, Test, T20

Problem Formulation

Given the historical ‘match’ (=game) data and assumptions,

Learn a model for predicting the likely Strike Rate per player

1. Scraped Data from ESPN Cricinfo



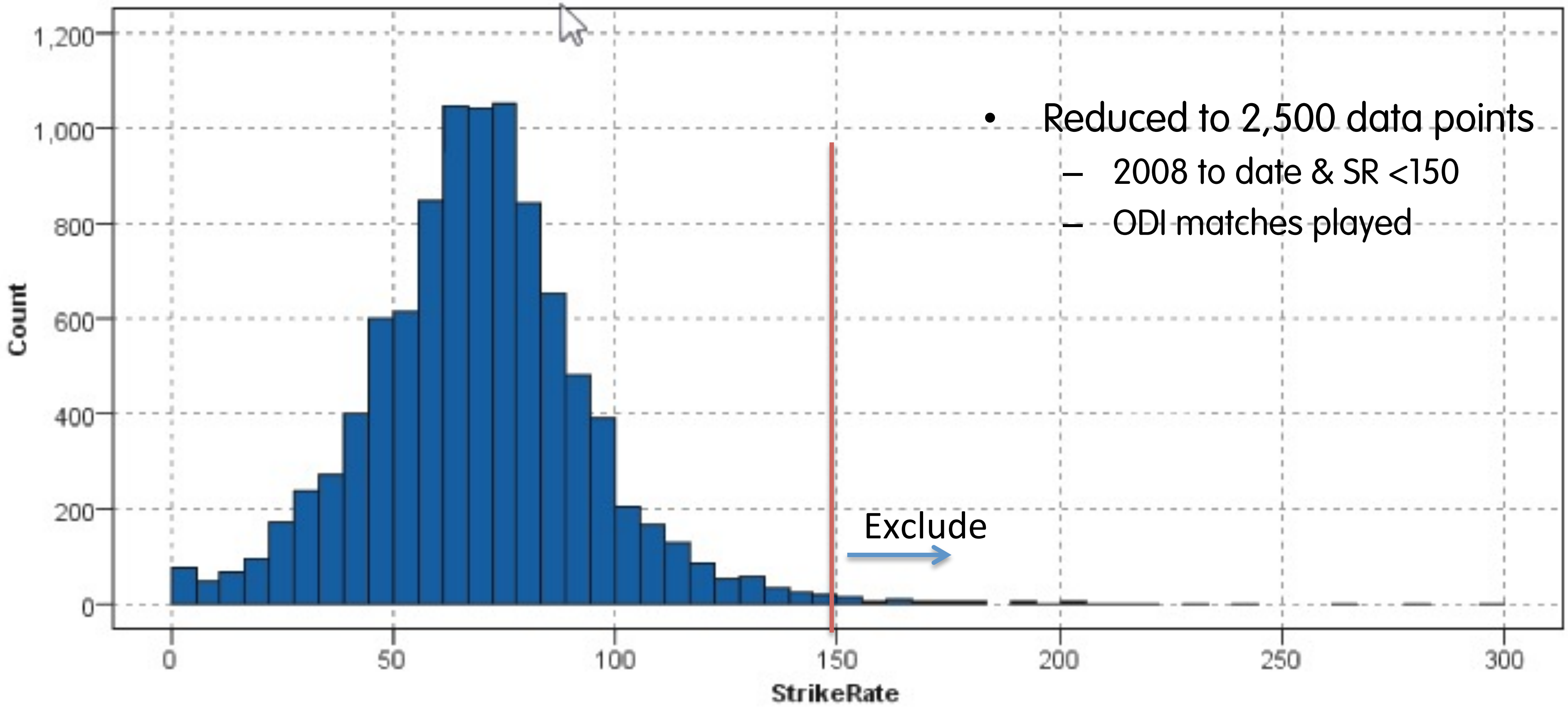
Field
Matches
A Season
Innings
Hundreds
TotalRuns
NotOuts
BallsFaced
DuckedOut
FiftyRuns
BattingAvg
HighestScore
StrikeRate

10 Potential
Predictors (X's)

Response (Y)

2. Data exploring, and filtering

- 12,100 data points
  - Player level batting history in ODI
  - 1972 to date

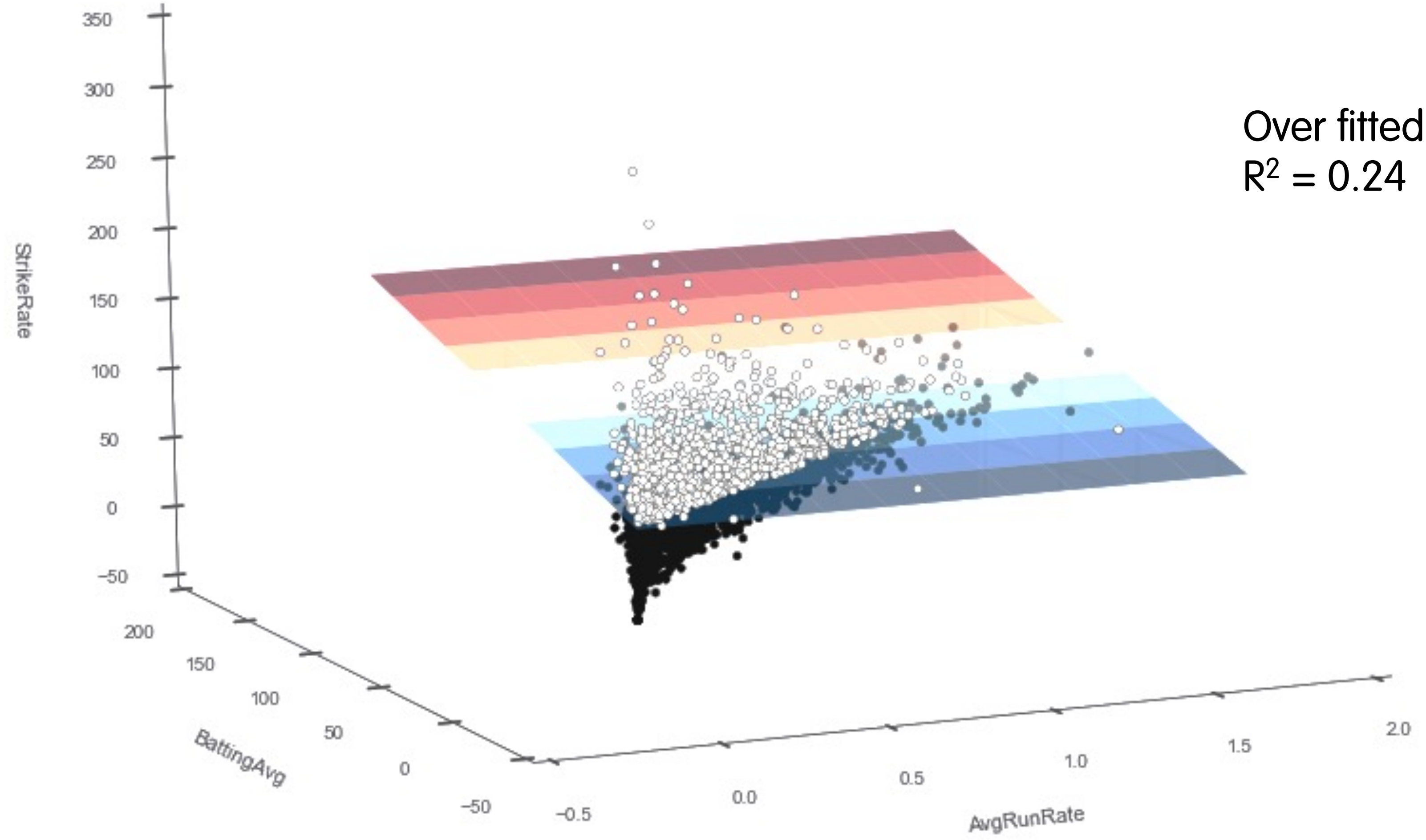


Exclude

- Reduced to 2,500 data points
  - 2008 to date & SR <150
  - ODI matches played

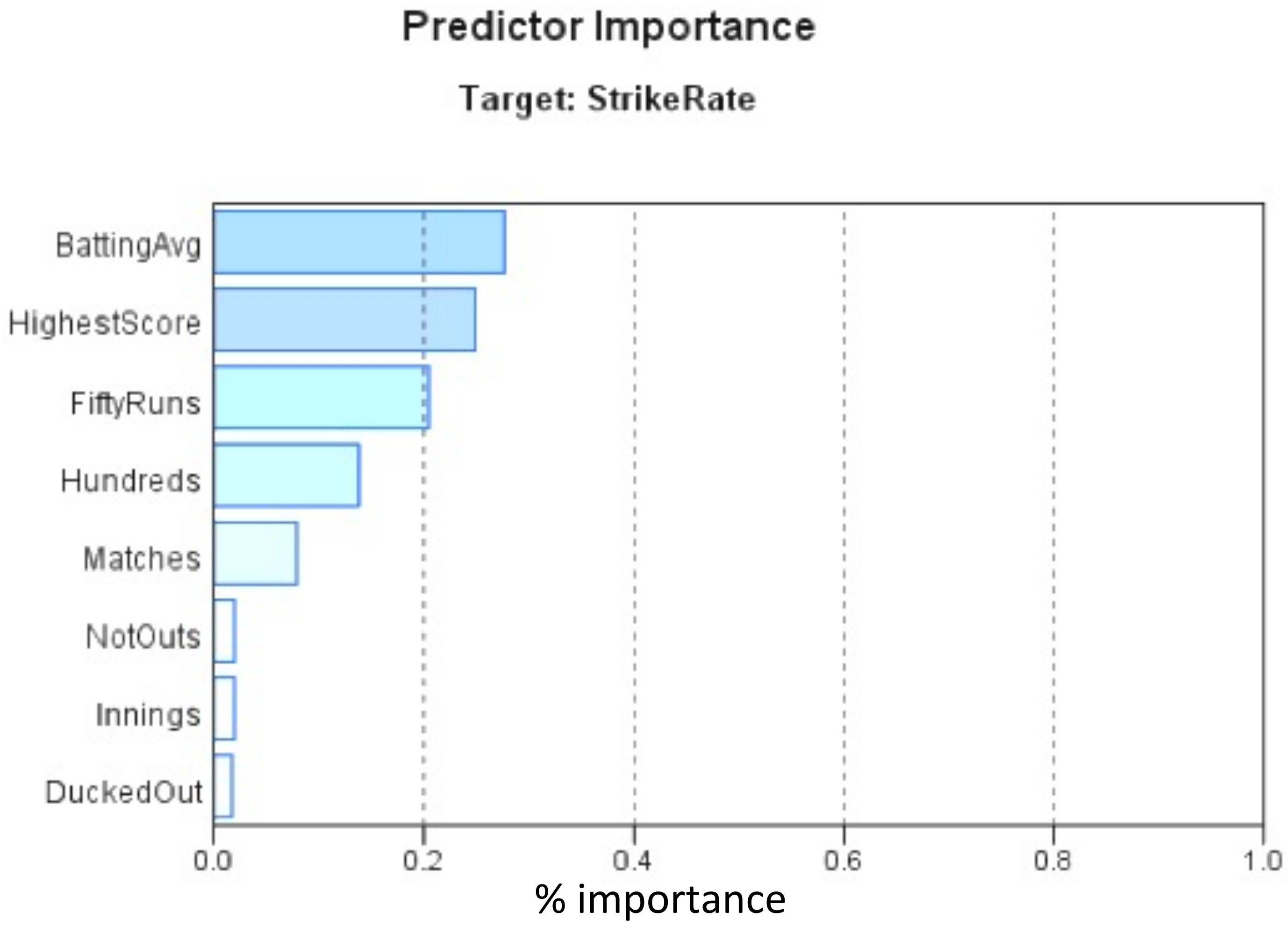
3. Run a linear regression with all of it

Explore how each player level stats affect Strike Rate



Over fitted with
$R^2 = 0.24$

4,5: Which ones are most influential features (X's) to Strike Rate (Y)

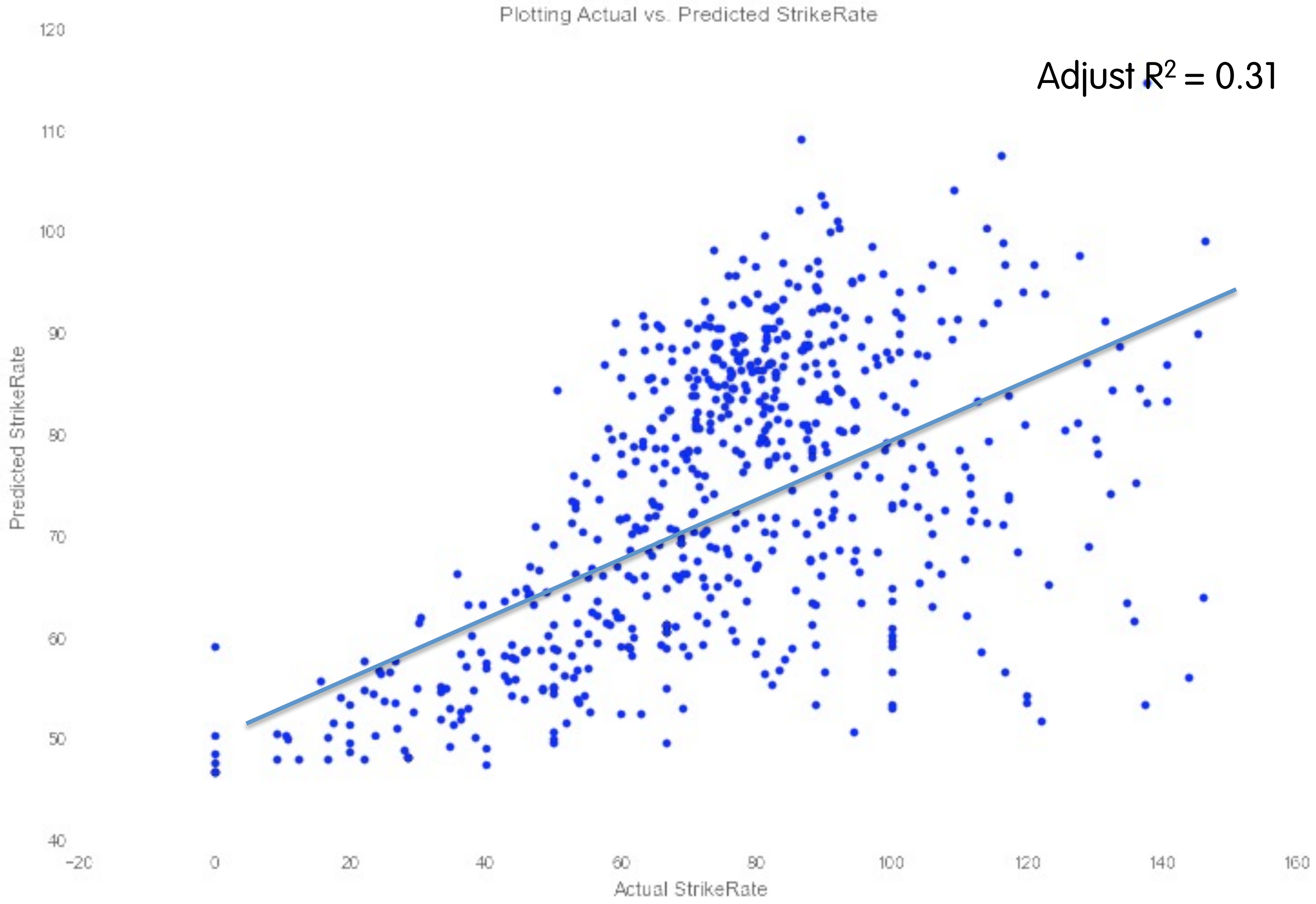


6. Revised model with relevant features only, after a few more iterations

Predicted_StrikeRate =
45.3+
0.9 * BattingAvg +
11.4 * AvgRunRate +
-0.61 * BattingAvg:AvgRunRate +
0.18 * HighestScore +
0.84 * Matches

Does this mean, a new
batsman with no
history hits 45.3 runs?

7. Predicted new Strike Rate



Reflections

Was it an ‘acceptable’ prediction? Yes, for the following reasons:

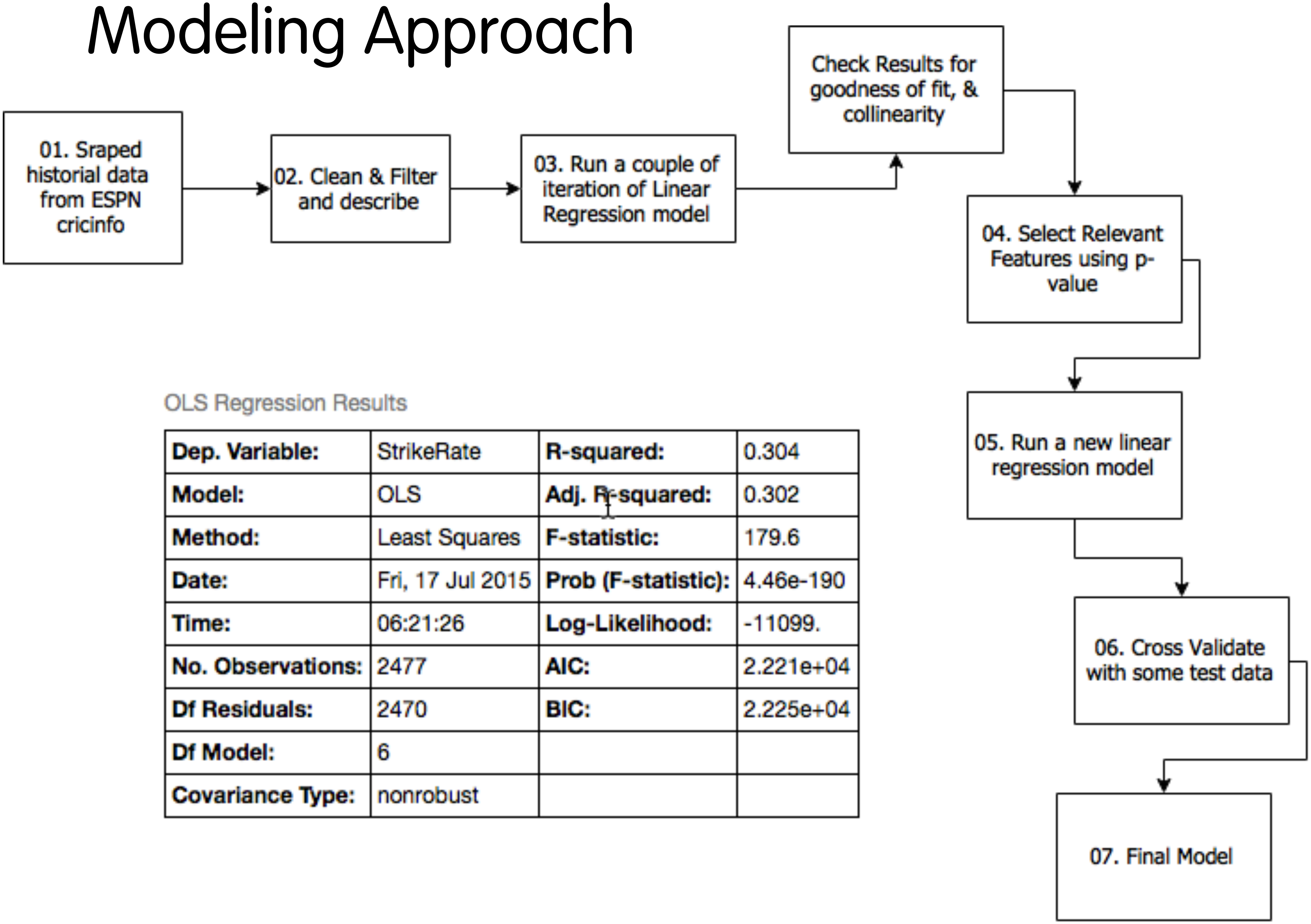
- Consistent results in 5-fold cross-validation
  (R^2 scores = 0.29, 0.31, 0.34, 0.28, 0.30)

- Reasonably sized sample (2,500 players) in making the prediction

- Improvement in the model from 24% to 31% explained variation, considering all the ‘unexplained’ noise in the data.

Appendix

Modeling Approach



01. Sraped historial data from ESPN cricinfo

02. Clean & Filter and describe

03. Run a couple of iteration of Linear Regression model

Check Results for goodness of fit, & collinearity

04. Select Relevant Features using p-value

05. Run a new linear regression model

06. Cross Validate with some test data

07. Final Model

OLS Regression Results

Dep. Variable:	StrikeRate	**R-squared:**	0.304
Model:	OLS	**Adj. R-squared:**	0.302
Method:	Least Squares	**F-statistic:**	179.6
Date:	Fri, 17 Jul 2015	**Prob (F-statistic):**	4.46e-190
Time:	06:21:26	**Log-Likelihood:**	-11099.
No. Observations:	2477	**AIC:**	2.221e+04
Df Residuals:	2470	**BIC:**	2.225e+04
Df Model:	6		
Covariance Type:	nonrobust		