PREDICTING SOCCER MATCH RESULTS TO IMPROVE CHANCES OF WINNING BETS

EESHPAUL | JOSEPH | SANTHOSH | RAJ

Introduction

Presentation Lineup

Eeshpaul
(Fall '17 BA)

Raj
(Fall '18 BA)

Santhosh
(Fall '17 BA)

Joseph
(Fall '18 BA)

Data Collection

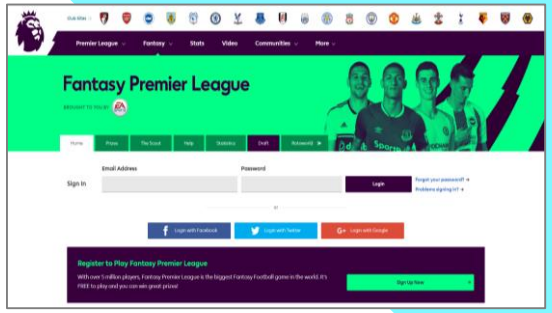
Webscraping

Understat.com

- Team Data (Goals, Shots, Expected Goals, Build Up etc.)
- Team Metrics (Formation, Goal difference, Timing, Attack Speed etc.)
- Player Data (Position, Goals, Time, Expected goals, Key passes etc.)

Football-uk.com

- Team Data (Goals, Shots, Fouls, Cards etc.)



Fantasy.premierleague.com

- Player Data (Clean sheets, Big chance missed, Own Goal etc.)

Data Collection

Team Data

CHELSEA
16/17

Data Collection

Team Data

CHELSEA
16/17

CHELSEA
17/18

CHELSEA
18/19

Data Collection

Player Data

6

Paul POGBA

16/17

GOAL	SHOT	CARD
5	92	7

EXPECTED GOAL	KEY PASS	ASSIST
7	57	4

Data Collection

Player Data

6
Paul
POGBA

16/17

6
Paul
POGBA

17/18

6
Paul
POGBA

18/19

500+ PLAYERS

16-19

10
Kun
AGUERO

16-19

10
Harry
KANE

Data Collection

Team Metrics

SITUATION	FORMATION	SHOT ZONE
OPEN PLAY FROM CORNER SET PIECE DIRECT FREEKICK PENALTY	4–3–3 4–1–4–1 3–5–2 3–4–3 …	OWN GOAL OUT OF BOX PENALTY AREA SIX–YARD BOX

GAME STATE	TIMING (MIN)	RESULT
GOAL DIFFERENCE 0 GOAL DIFFERENCE +1 GOAL DIFFERENCE > +1 GOAL DIFFERENCE −1 GOAL DIFFERENCE < −1	16–30 31–45 46–60 61–75 76+	MISSED SHOT BLOCKED SHOT GOAL SAVED SHOT SHOT ON POST

Data Engineering

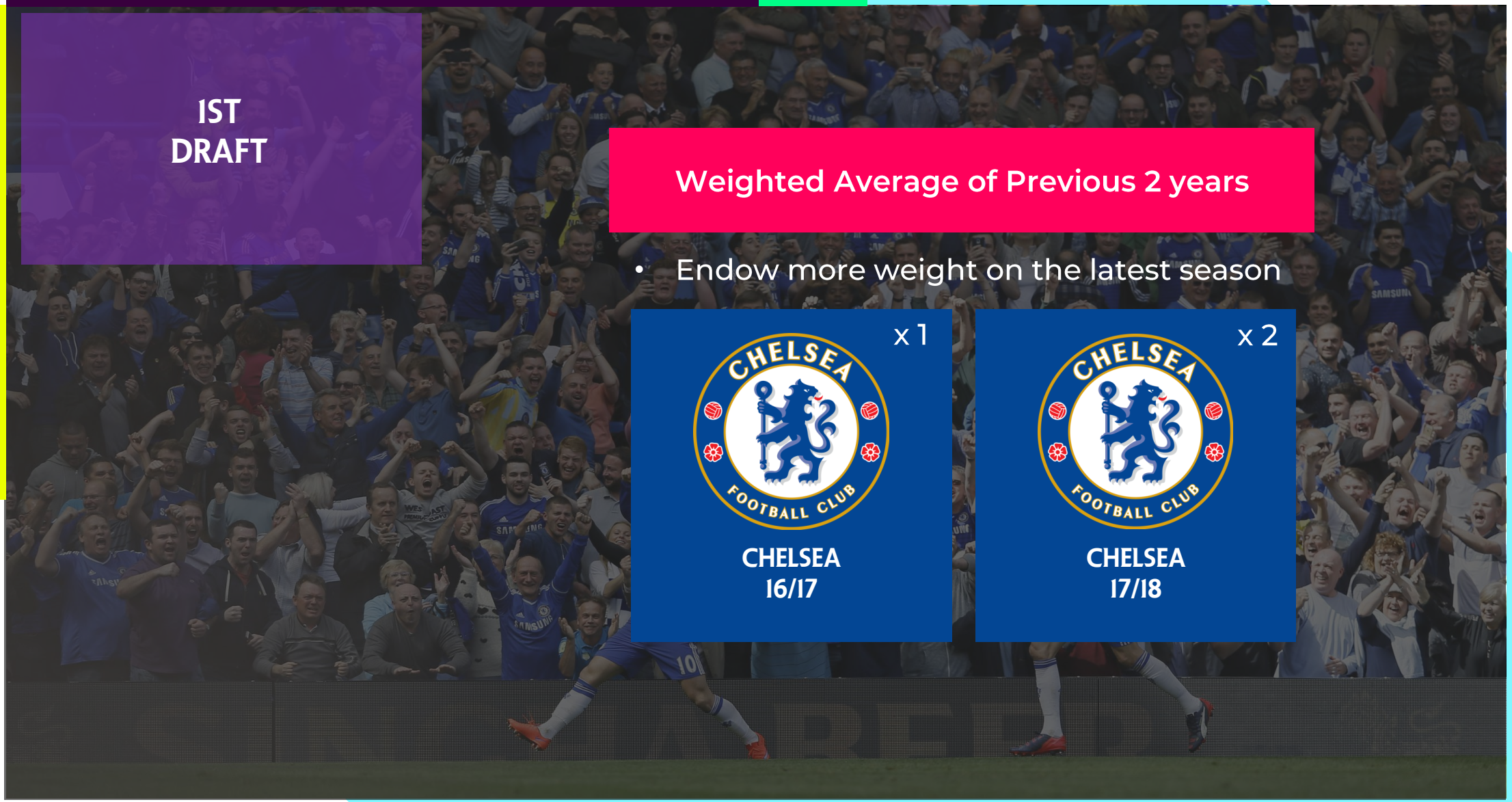
Dataset

			HOME TEAM						AWAY TEAM						TARGET VARIABLE
HOME	AWAY	SEASON	GOAL	SHOT	XG	76-90	...	XGA	GOAL	SHOT	XG	76-90	...	XGA	RESULT
		16/17	21	109	19	10	...	40	21	109	19	10	...	40	A
		17/18	39	171	42	17	...	32	39	171	42	17	...	32	H
		16/17	56	255	53	13	...	11	56	255	53	13	...	11	A
		17/18	19	142	13	9	...	59	19	142	13	9	...	59	H
		17/18	24	126	21	6	...	66	24	126	21	6	...	66	D
		16/17	17	96	11	5	...	63	17	96	11	5	...	63	A
		16/17	62	276	58	16	...	24	62	276	58	16	...	24	H
		17/18	49	199	49	12	...	23	49	199	49	12	...	23	A

Data Engineering

Final Dataset

1ST
DRAFT



Weighted Average of Previous 2 years

- Endow more weight on the latest season

x 1

CHELSEA
16/17

x 2

CHELSEA
17/18

Data Engineering

Final Dataset

1ST DRAFT

2ND DRAFT

Last Year & Last Last Year Column

- In game match stats of previous years

16/17	CHELSEA	MANCHESTER UTD.	GOAL GOAL AGAINST EXPECTED GOAL AGAINST SHOT SHOT AGAINST EXPECTED SHOT AGAINST 76 – 90 31– 45 ...
17/18	CHELSEA	MANCHESTER UTD.	GOAL GOAL AGAINST EXPECTED GOAL AGAINST SHOT SHOT AGAINST EXPECTED SHOT AGAINST 76 – 90 31– 45 ...

Data Engineering

Final Dataset

1ST DRAFT

2ND DRAFT

3RD DRAFT

Specially Engineered Variables

- Finish Flag of previous year
  - Champion's League, Europa, FA Cup etc.
- Key Players
  - 30 Top players per seasons based on skills stats

Data Engineering

Final Dataset

TRAINING DATA

16/17 – 17/18

TEST DATA

18/19 (100 MATCHES)

NUMBER OF FEATURES

489

SELECTED FEATURES

38

Feature Selection

Important Variables

TIME
(76-90, 31-45)

"Time is important to Win"

- The end time of the half and the game is important
  ex) Shots, Shots Against, Expected Goals, Expected Goals Against, Goals

SHOT > GOAL

"Shot is better than Goal"

- Shots represents the performance of a team better
  ex) 76-90 | Shots Against, Blocked Shots | Shots, Shot in Six-yard Box | Shots Against

EXPECTED GOAL > GOAL

"Expected Goal is better than Goal"

- Expected Goal also represents the performance of a team better
  ex) Goal Difference 1 | Expected Goals Against, 76-90 | Expected Goals Against

Model

Ensemble

Logistic Regression	Random Forest	Support Vector Machine
Feature Selection: Grid Search K-fold CV : 4-fold CV	# of trees : 1,000 # of Total Variables: 38 Mtry(split at each tree): 12	Feature Selection: PCA (16 components explaining 99% variance) Model Type: Linear SVM K-fold CV: 10-fold CV
Accuracy **56.98%**	Accuracy **59.76%**	Accuracy **57.32%**

Ensemble Model

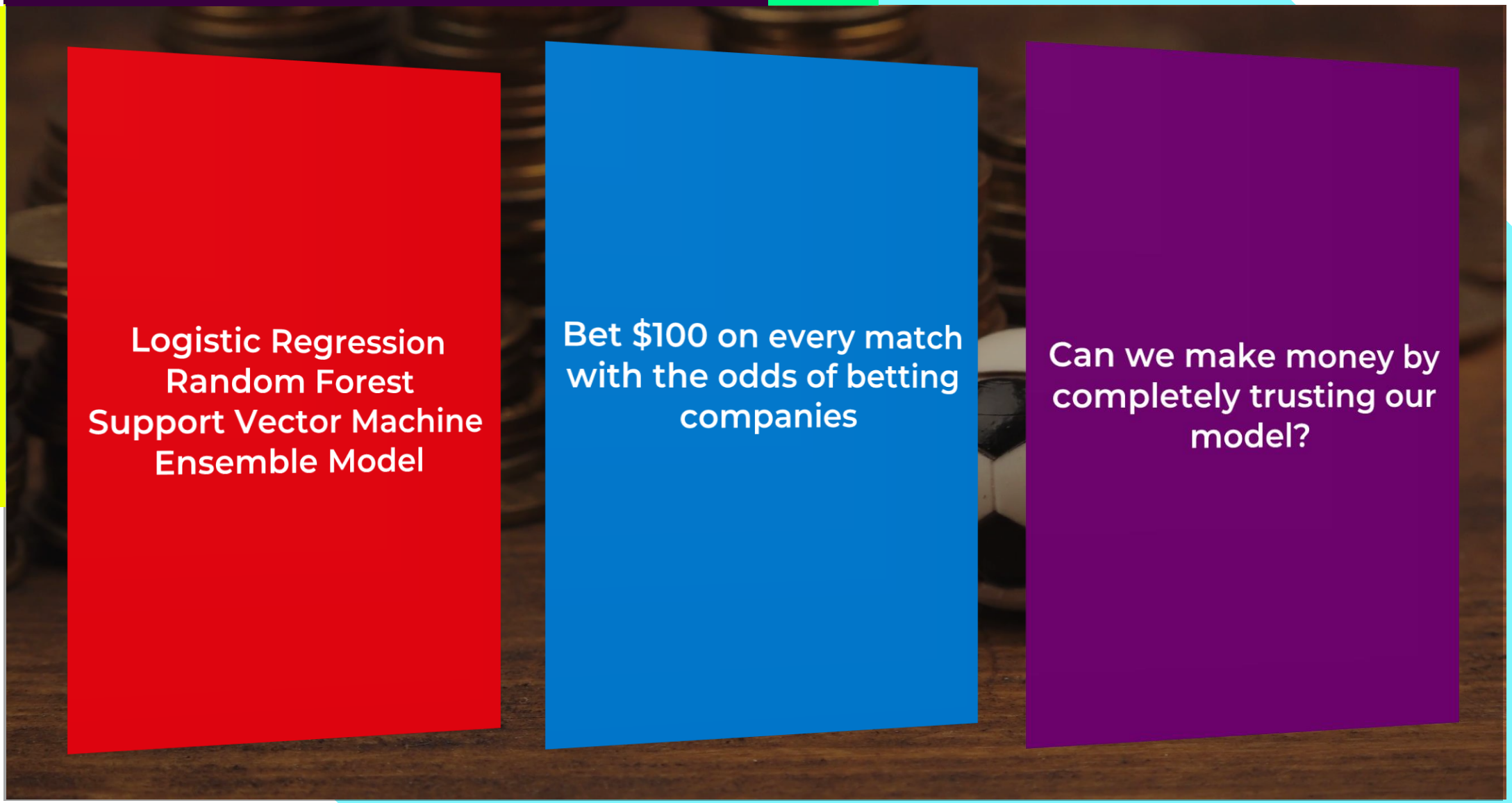
Choose the model with the best accuracy

Accuracy **57.32%**

Row Labels	A	D	H	Grand Total	
A	19	10	2	31	61.29%
D	4	7	4	15	46.67%
H	5	10	21	36	58.33%
Grand Total	28	27	27	82	57.32%

Model

Application



Logistic Regression
Random Forest
Support Vector Machine
Ensemble Model

Bet $100 on every match with the odds of betting companies

Can we make money by completely trusting our model?

Profit Comparison

Betting Companies	bet365	bwin.com	William Hill	**TOTAL 82 Games** **Earn $8,200** In 100% accuracy
	?	?	?	
Friend Groups	**Smart Fan(1)**	**Regular Fan(3)**	**Bad Fan(1)**	
	?	?	?	
Our Model	**Random Forest**	**Logistic Regression**	**SVM**	**Ensemble**
	?	?	?	?

Profit Comparison

Betting Companies	bet365	bwin	William Hill	TOTAL 82 Games
	$5,311*	$5,203	$5,257	Earn $8,200 In 100% accuracy
Friend Groups	Smart Fan(1)	Regular Fan(3)	Bad Fan(1)	
	?	?	?	
Our Model	Random Forest	Logistic Regression	SVM	Ensemble
	?	?	?	?

Profit Comparison

Betting Companies	bet365	bwin	William Hill	TOTAL 82 Games
	$5,311*	$5,203	$5,257	Earn $8,200 In 100% accuracy

Friend Groups	Smart Fan(1)	Regular Fan(3)	Bad Fan(1)
	$7,247*	$5,311	$2,194

Our Model	Random Forest	Logistic Regression	SVM	Ensemble
	?	?	?	?

Profit Comparison

Betting Companies	bet365	bwin	William Hill	TOTAL 82 Games
	$5,311*	$5,203	$5,257	Earn $8,200 In 100% accuracy

Friend Groups	Smart Fan(1)	Regular Fan(3)	Bad Fan(1)
	$7,247*	$5,311	$2,194

Our Model	Random Forest	Logistic Regression	SVM	Ensemble
	$6,267*	$5,812	$5,347	$5,407

Future Scope

- Work on further improving our predictions by exploring new models and engineering better explanatory features
- Operationalizing this model to update every week
- Build a GUI for an easy visualization of actual results, predicted results and their comparisons with betting websites
- MAKE SOME MONEY $$$

Q&A