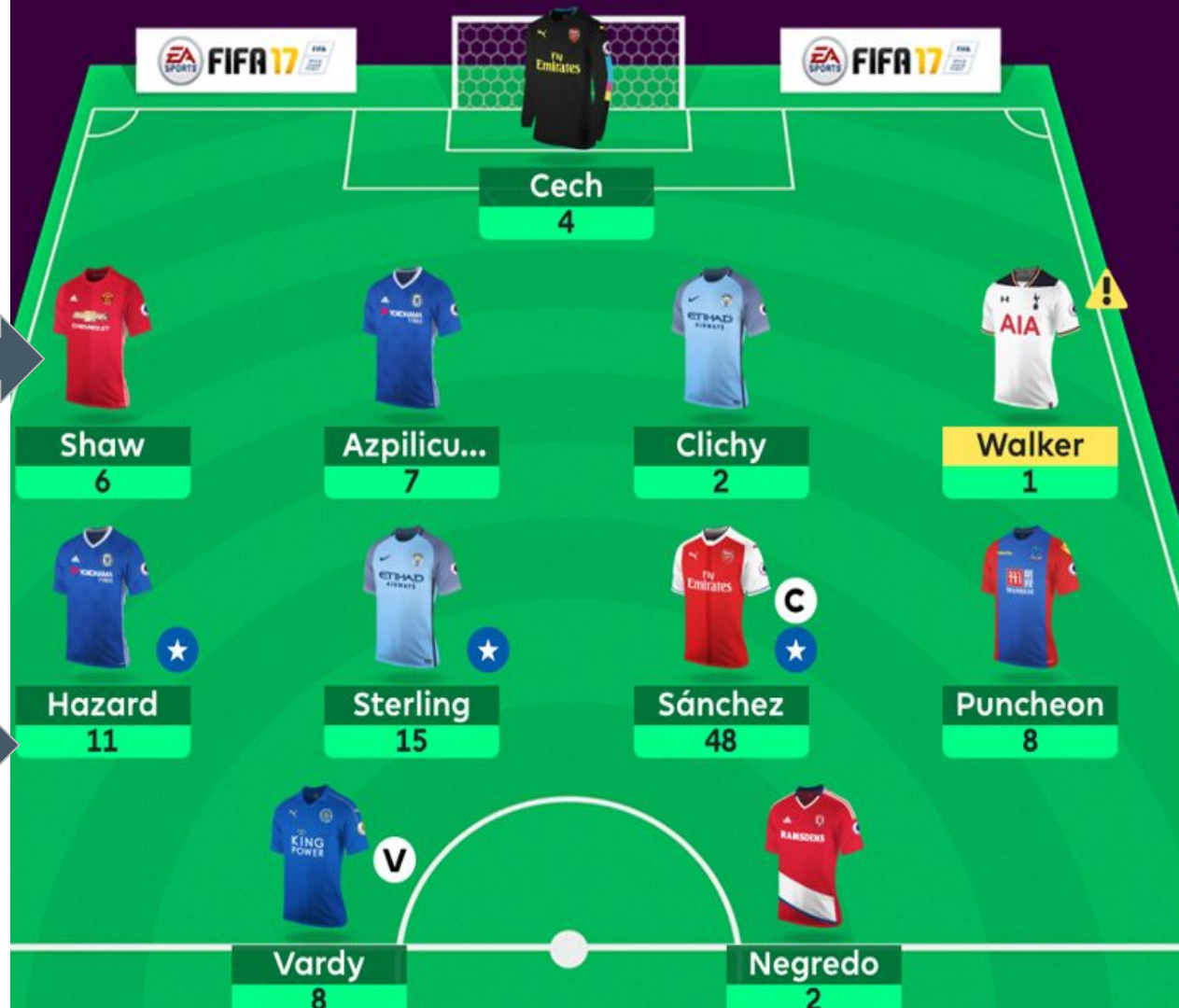r/English Premier League

Vs   r/Fantasy Premier League

Marketing Insights from Reddit

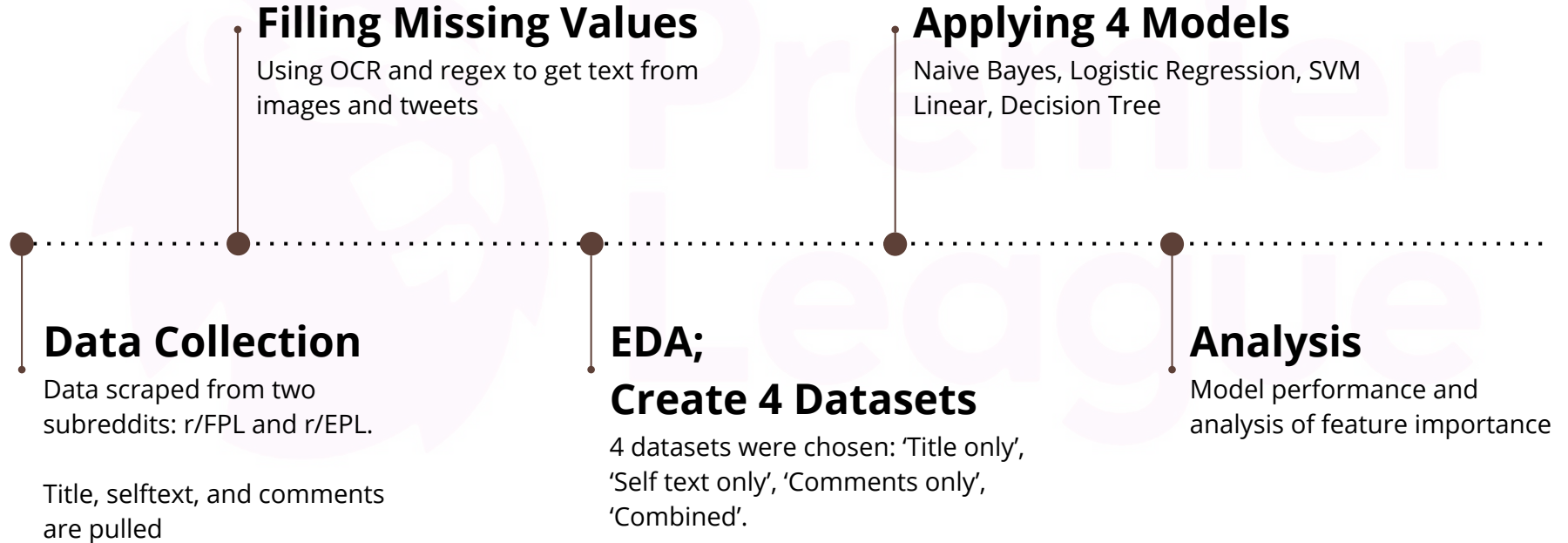# What is FPL?

Pick 14 Players

Score based on real
player's performance

Project Overview

Filling Missing Values

Using OCR and regex to get text from images and tweets

Applying 4 Models

Naive Bayes, Logistic Regression, SVM Linear, Decision Tree

Data Collection

Data scraped from two subreddits: r/FPL and r/EPL.

Title, selftext, and comments are pulled

**EDA;
Create 4 Datasets**

4 datasets were chosen: 'Title only', 'Self text only', 'Comments only', 'Combined'.

Analysis

Model performance and analysis of feature importance

Data Cleaning

Code applies an OCR to images and returns a text string

```python
#Converts image to text string
def ocr_core(filename):
    #Try/except handles 404 and other similar errors
    try:
        response = requests.get(filename)
        img = Image.open(BytesIO(response.content))
        text = pytesseract.image_to_string(img)
        return text
    except:
        return np.nan
```


FACEBOOK
SCREENSHOT

Data Cleaning
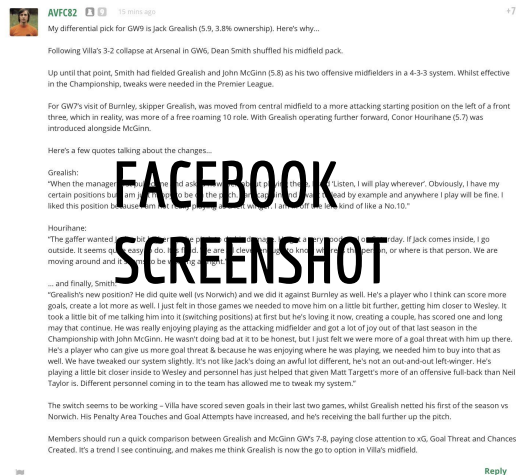
Regex extracts text from embedded tweet in reddit scrape

```python
rep = {'"ltr"': "", '&gt': "", '&lt' : "", '&amp;#39' : "", 'br&gt' : " ", "p&gt" :
    "&amp" : "", "amp;": "", "mdash" : "", ';' : ""}
rep = dict((re.escape(k), v) for k, v in rep.items())
pattern = re.compile("|".join(rep.keys())) #Creates pattern based on rep dictionary

#Extracts tweet content from media column in dataframe (which is in a bastardized h
def extract_tweet(tweet):
    #Try/except in case of formatting problems in the raw data.
    try:
        twit_string = pattern.sub(lambda x: rep[re.escape(x.group(0))], tweet) #get
        twit_string = re.search('dir=(.*)blockquote', twit_string) #finds main cont
        twit_string = re.sub(r'(?s)(a href)(.*?)(\/a)', " ", twit_string.group(1))
        return(twit_string)

    except:
        return np.nan
```

{'oembed': {'provider_url': 'https://twitter.com', 'url': 'https://twitter.com/FFScout/status/1177564416927252481', 'html':
'<blockquote class="twitter-video"><p lang="en" dir="ltr">👤 Pep Guardiola when asked if Manchester City
have any new fitness concerns: &quot;I have training this afternoon&quot;.<a
href="https://twitter.com/hashtag/FFScout?FPL=hash&amp;ref_src=twsrc%5Etfw">#FFScout <a
href="https://twitter.com/hashtag/FPL?src=hash&amp;ref_src=twsrc%5Etfw">#FPL <a
href="https://twitter.com/hashtag/GW7?src=hash&amp;ref_src=twsrc%5Etfw">#GW7 <a
href="https://twitter.com/hashtag/MCFC?src=hash&amp;ref_src=twsrc%5Etfw">#MCFC <a
href="https://twitter.com/hashtag/FantasyPL?
src=hash&amp;ref_src=twsrc%5Etfw">#FantasyPL</p>&mdash; Fantasy Football Scout
(@FFScout) <a href="https://twitter.com/FFScout/status/1177564416927252481?
ref_src=twsrc%5Etfw">September 27, 2019</blockquote>\n<script async
src="https://platform.twitter.com/widgets.js" charset="utf-8"></script>\n', 'author_name': 'Fantasy Football
Scout', 'height': 186, 'width': 350, 'version': '1.0', 'author_url': 'https://twitter.com/FFScout', 'provider_name': 'Twitter',
'cache_age': 3153600000, 'type': 'rich'}, 'type': 'twitter.com'}

👤 Pep Guardiola when asked if Manchester City have any new fitness concerns: I have training this afternoon. / Fantasy Football Scout (@FFScout) /

Number of Posts By Date
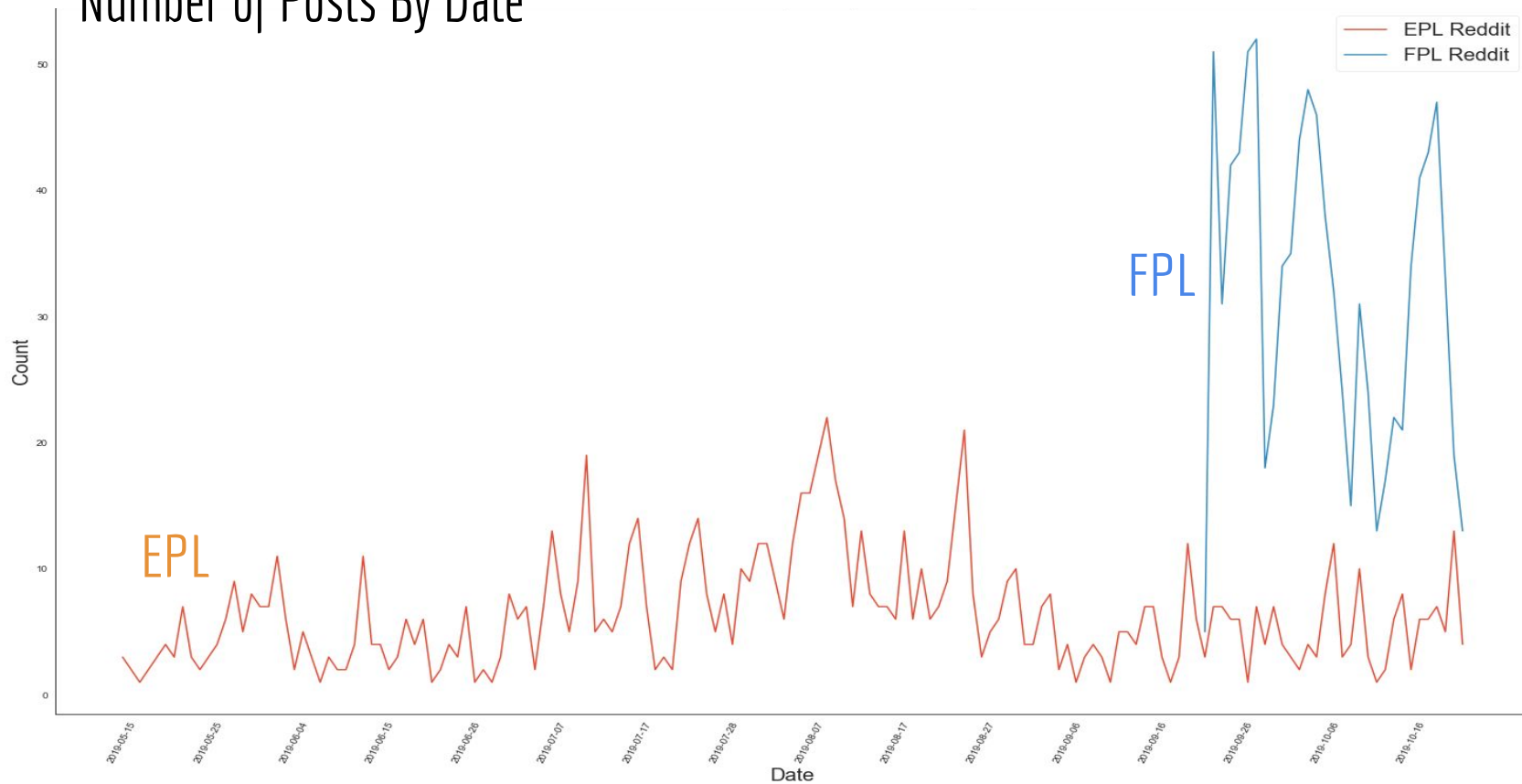
Naive Bayes (Multinomial)

> Title only
> Selftext only
> Comments only
> Combined

Logistic Regression

> Title only
> Selftext only
> Comments only
> Combined

SVM Linear

> Title only
> Selftext only
> Comments only
> Combined

Decision Tree

> Title only
> Selftext only
> Comments only
> Combined

Modelling Strategy

Four models selected; each model then run on 4 datasets for a total of 16 results.

These models were selected because they facilitate feature analysis

——

Modelling Results (Accuracy and AUC scores)

Accuracy

	title	selftext	comments	combined
Naive Bayes	0.834171	0.782245	0.869347	0.917923
Log Reg	0.740369	0.793970	0.857621	0.926298
SVM Linear	0.844221	0.783920	0.855946	0.926298
Decision Tree	0.800670	0.738693	0.743719	0.825796

AUC Score

	title	selftext	comments	combined
Naive Bayes	0.919	0.868	0.925	0.971
Log Reg	0.818	0.856	0.910	0.975
SVM Linear	0.928	0.815	0.907	0.976
Decision Tree	0.804	0.750	0.750	0.816

- LogReg and SVM were best performers

- Title alone is a decent classifier

- Self text alone performed poorly; similar content is shared

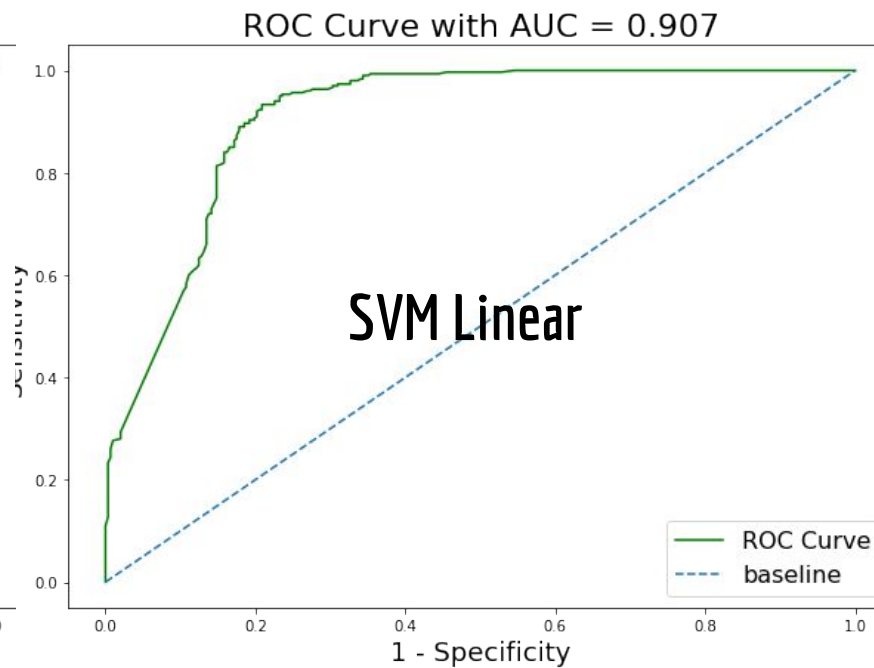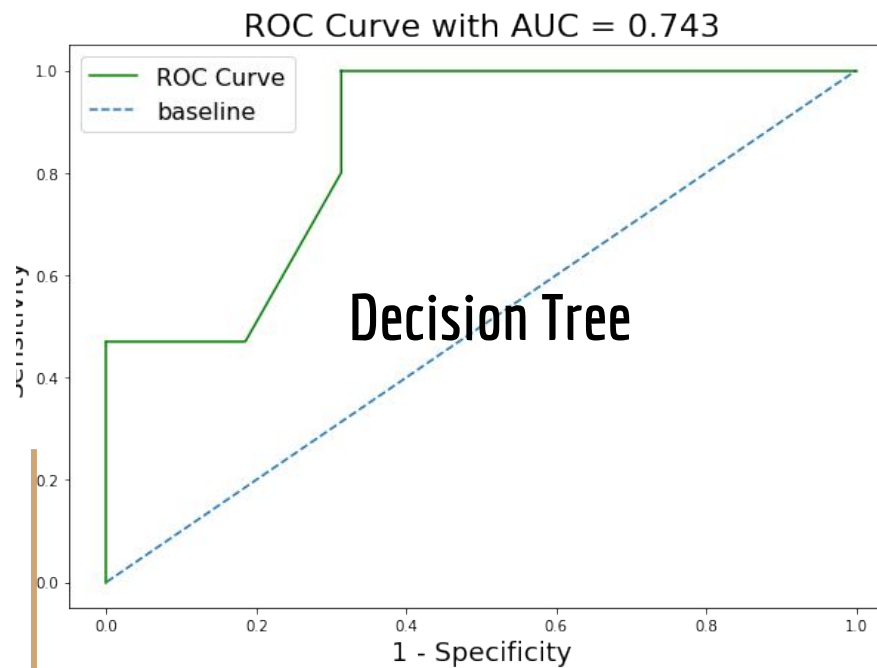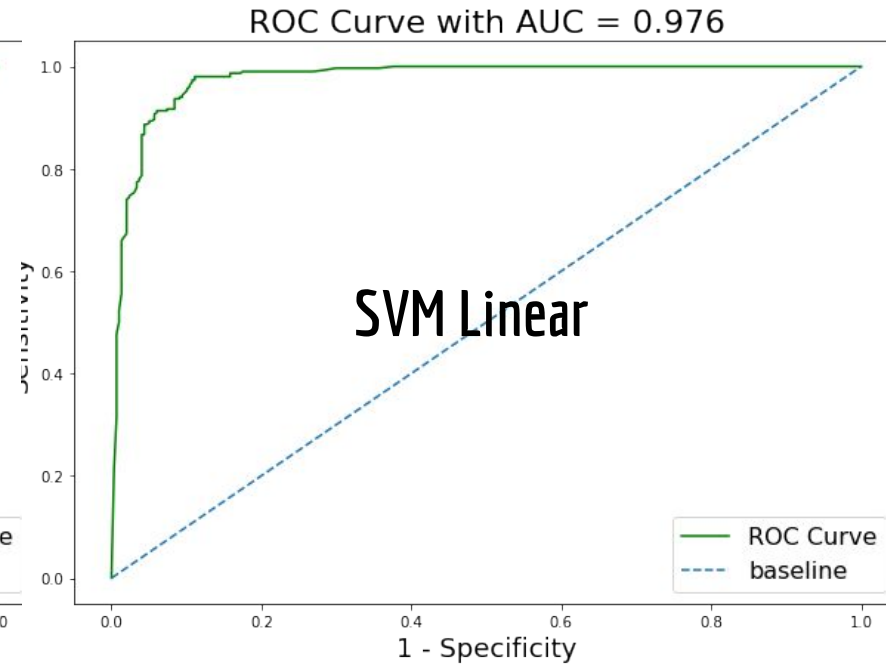- The more features you put in, the better the accuracy

Title Only

ROC Curve with AUC = 0.821

Decision Tree

ROC Curve with AUC = 0.928

SVM Linear

Selftext Only

ROC Curve with AUC = 0.786

ROC Curve with AUC = 0.815

Decision Tree

SVM Linear

1 - Specificity

1 - Specificity

Comments Only

ROC Curve with AUC = 0.743

Decision Tree

ROC Curve with AUC = 0.907

SVM Linear

Combined

ROC Curve with AUC = 0.826

ROC Curve with AUC = 0.976

Decision Tree

SVM Linear

Top 25 Features
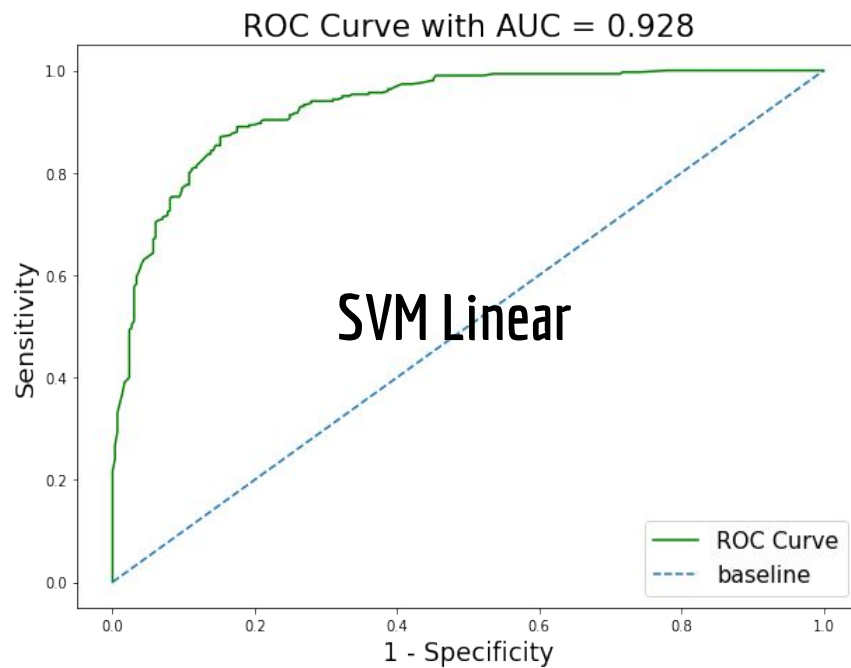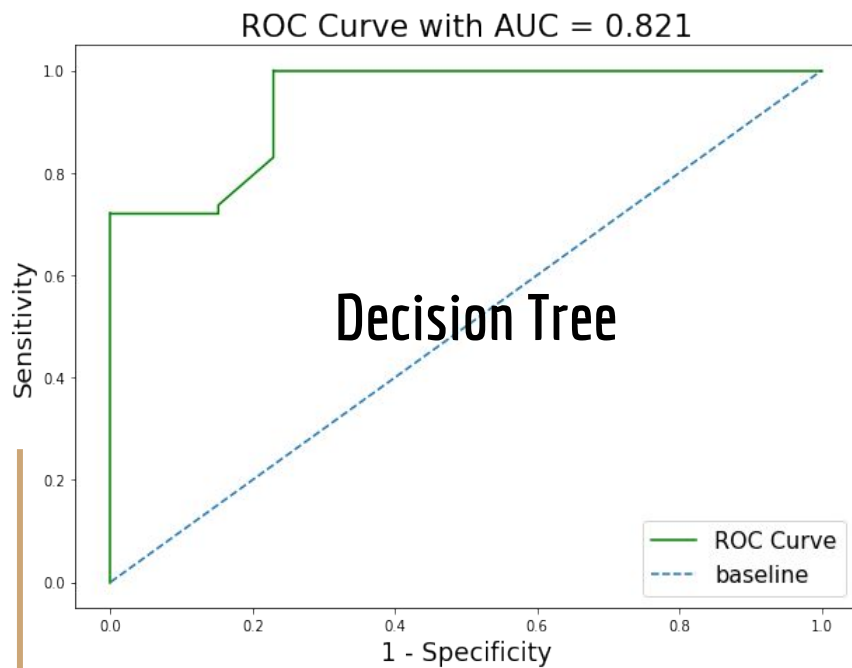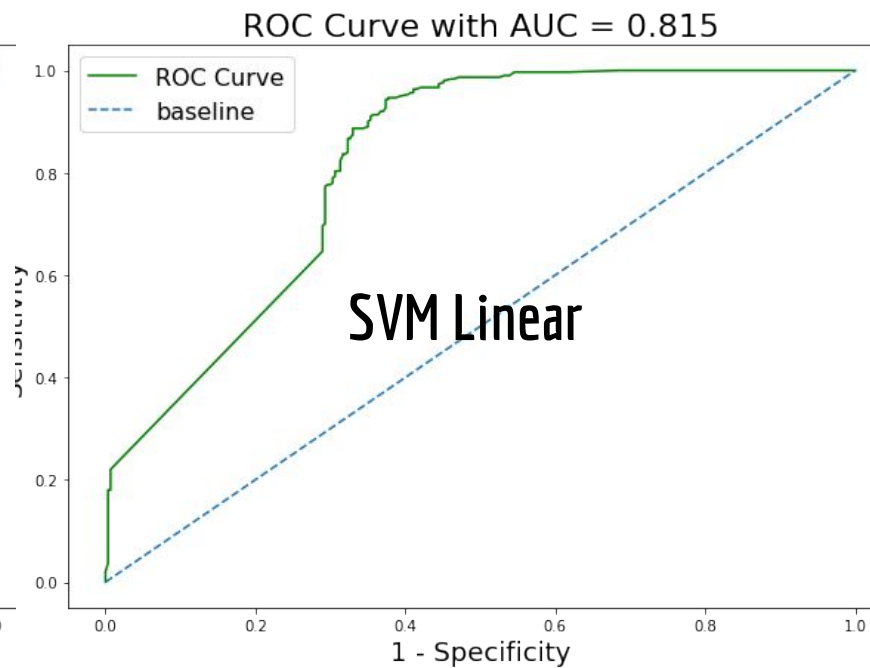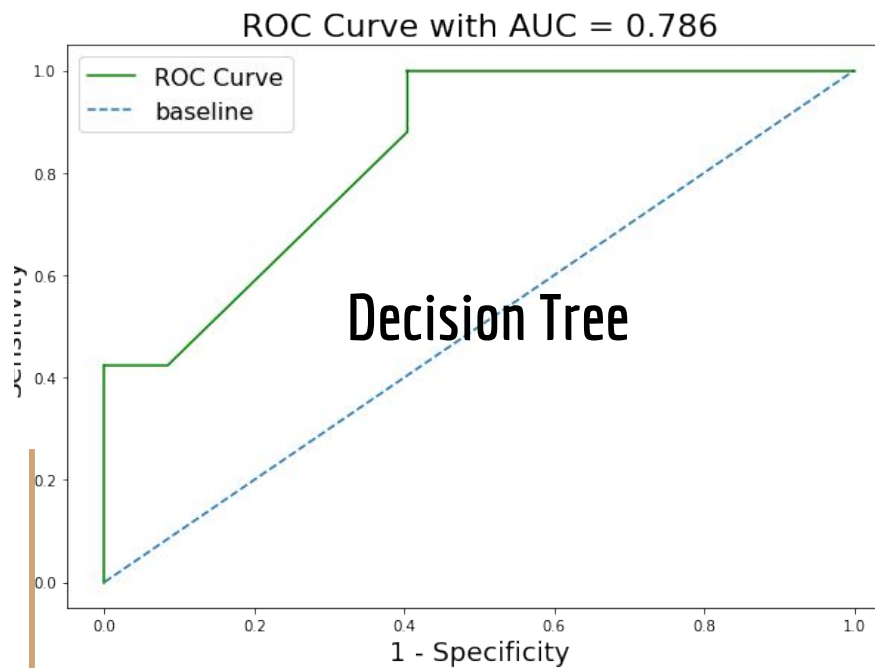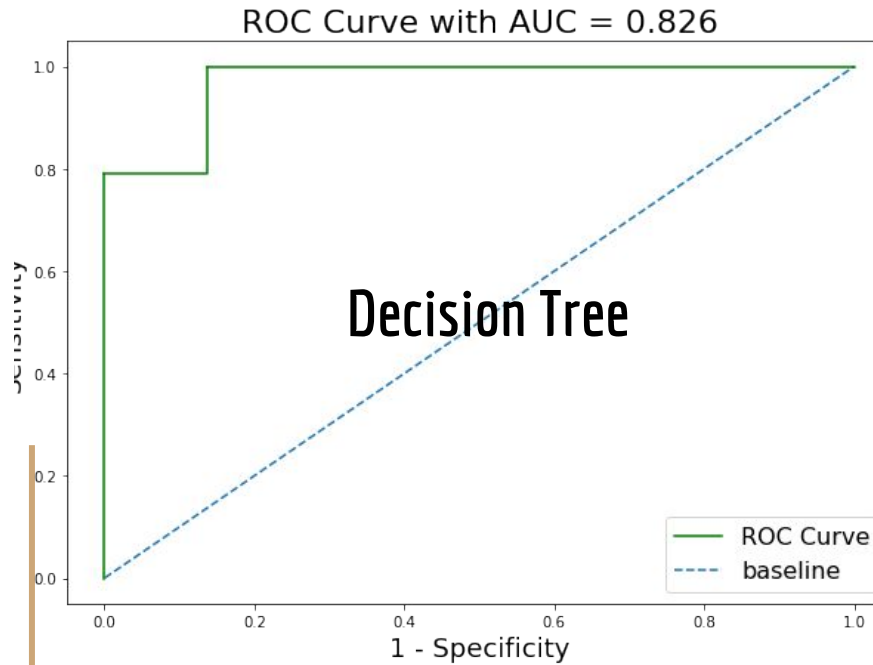
FPL

	Log Reg	SVM Linear	Decision Tree
0	gw	gw	gw
1	week	fantasy	kdb
2	point	gameweek	gameweek
3	gameweek	point	xi
4	fantasy	start	clean
5	meme	week	injury
6	start	xi	cantwell
7	kdb	greenwood	start
8	captain	option	posted
9	injury	captain	fantasy
10	clean	injury	mount
11	sterling	differential	update
12	fixture	chance	wa
13	sheet	xg	nailed
14	bench	clean	xg
15	xi	romero	ib
16	price	assist	ben
17	salah	asset	league
18	mount	meme	player
19	greenwood	fixture	tc
20	option	episode	chance

EPL

	Naive Bayes	Log Reg	SVM Linear
0	league	club	premier league
1	premier	premier	window
2	premier league	fan	manchester united
3	team	var	buy
4	united	football	ability
5	club	league	signing
6	season	sign	bbc
7	wa	signing	subbed minute
8	player	tottenham	predicted
9	fan	arsenal	van
10	city	deal	kevin bruyne
11	arsenal	ticket	time
12	football	manchester	manchester
13	game	year	interested
14	liverpool	united	championship
15	ha	english	sign
16	year	season	english
17	like	watch	football
18	good	icon	club
19	var	loan	tottenham
20	chelsea	contract	var

Technical FPL Terms

Teams

Players

Football News

First few Decision Tree splits are on Player Name

gw ≤ 0.002
gini = 0.5
samples = 1390
value = [693, 697]

True

False

kdb ≤ 0.008
gini = 0.478
samples = 1148
value = [454, 694]

deep ≤ 0.056
gini = 0.024
samples = 242
value = [239, 3]

gameweek ≤ 0.008
gini = 0.449
samples = 1047
value = [357, 690]

bruyne ≤ 0.15
gini = 0.076
samples = 101
value = [97, 4]

thrust ≤ 0.032
gini = 0.008
samples = 240
value = [239, 1]

gini = 0.0
samples = 2
value = [0, 2]

xi ≤ 0.036
gini = 0.429
samples = 1001
value = [312, 689]

addressed ≤ 0.079
gini = 0.043
samples = 46
value = [45, 1]

frightening ≤ 0.044
gini = 0.02
samples = 97
value = [96, 1]

doubt ≤ 0.043
gini = 0.375
samples = 4
value = [1, 3]

gini = 0.0
samples = 239
value = [239, 0]

gini = 0.0
samples = 1
value = [0, 1]

FPL

coef	feature
-1.650847	sterling
-1.314611	salah
-1.285031	mount
-1.238638	greenwood
-1.233254	son
-1.227731	aguero
-1.164465	lundstram
-1.126874	pukki
-1.101163	tomori
-1.081017	cantwell
-1.077149	chance
-1.050719	wilson
-1.020123	abraham
-0.911834	mahrez
-0.897539	mcginn

EPL

coef	feature
0.325824	crystal
0.340312	ashley
0.340989	david
0.351290	graham
0.358717	sane
0.386136	palace
0.389661	west
0.395960	chris
0.429823	kyle
0.432827	pochettino
0.442659	henry
0.456683	liverpool
0.491166	young
0.519032	sheffield

What players do r/FPL and r/EPL support?

Coefficients were passed through a dictionary of players + teams to see which players/teams were popular.

- r/FPL focuses on players
- r/EPL focuses on teams and managers (*managers highlighted in green*)

———

coef	feature
-0.692644	watford
-0.353725	leicester
-0.242555	pool
-0.199185	burnley
-0.161617	bournemouth
-0.096775	southampton
0.004072	norwich
0.139241	brighton
0.254606	everton
0.325824	crystal
0.519032	sheffield
0.527880	chelsea
0.605307	city
0.655225	newcastle
0.722146	villa
1.142986	united
1.191003	manchester
1.260671	arsenal
1.361363	tottenham

What teams do r/FPL and r/EPL support?

r/EPL discusses the big teams (e.g. Manchester United, Arsenal, Manchester City, Liverpool…)

r/FPL discusses the less popular teams (e.g. Watford, Leicester, Burnley…)

There may be opportunities to target fans of the smaller teams as their players are hot topics in FPL circles.

Conclusions

r/FPL is interested in players; r/EPL is interested in teams
Marketing campaign on r/EPL should be team focused

r/FPL is full of FPL jargon
How do we ease EPL redditors into FPL?

Most popular teams/players
Use the Big 6 football clubs, but also appeal to the supporters of the smaller clubs, particularly if their players are doing well in FPL