Part II Project Progress Presentation

"Predicting the Outcomes of English Premier League Matches"

Joseph Marchant

Introduction

- English Premier League (EPL) has 20 teams

- Season is 38 matches

- Data covers 26 seasons

- Goal is to predict the outcome of matches from 3 classes: Home, Draw, Away.

- Can only use data that is known before game begins.

- Goal is 40% accuracy.

- Focus is on finding unique/ effective features.

Data Collection - 11v11.com

- A website for international football statistics

- Web-scraped over 10000 pages

- Stored into csv

Data Link	Referee	Stadium
nal-v-norwich-city-15-august	Alan Gunn	Highbury
ea-v-oldham-athletic-15-augu	Jim Borrett	Stamford Bridge
y-city-v-middlesbrough-15-au	Howard King	Highfield Road

HomePlayers	AwayPlayers	HomeScorers	AwayScorers
4377(90)', '7800(90)', '37097(90	, '98(58)', '38680(90)', '846(90	['196', '484']	'511', '2266', '2266', '38680']
90)', '38546(90)', '31113(76)', '2	0)', '3620(90)', '1427(90)', '488	['37518']	['1427']
4796(90)', '34414(75)', '24329(9	0)', '356(54)', '33173(90)', '280	['1062', '5386']	['11296']

Data Collection - Other data sources

- Web-scraping

    - premierleague.com

    - Wikipedia.org

    - transfermarkt.co.uk

- Downloads

    - Football-data.co.uk

- Manual Collection

    - Physioroom.com

Data Cleaning/ Preparing for Models

- Missing/ incorrect values

- League table generator

  - Points Per Game, Goals Per Game, Clean Sheets Per Game, Position, Goals Conceded Per Game

- Prepare for models

  - Remove string columns

  - Add target column: {H = Home Win, D = Draw, A = Away Win}

	A	B	C	D	E	F	G	H	I	J	K
	hPos	hGSPG	hGCPG	hCSPG	hPtsPG	aPos	aGSPG	aGCPG	aCSPG	aPtsPG	target
	6	2	1.5	0.5	2	17	0.5	1.5	0	0.5	3
	22	0.5	2.5	0	0	8	1	0.5	0.5	2	1
	9	1.5	1	0.5	1.5	7	1.5	1	0	2	3
	1	3	1.5	0	3	3	2	0.5	0.5	2	1

Binary Models

- 80-20 train-test splits used

- Binary Classification

    - 1 = Home Team Win, 0 otherwise

- 5 "difference" features instead of 10

- Logistic Regression

    - 62% Accuracy

    - 51% Precision

    - 61% Recall

- Tensorflow NN

    - 61% accuracy best

3-way Classification Models

- 3-class Classification

    - Points: 3 = Home Team Win, 1 = Draw, 0 = Home Team Loss

- Tensorflow NN

- 30 hidden layers

- Learning rate = 0.1

- 100 epochs

- Tf.nn.sparse_softmax_cross_entropy_with_logits

- 49% accuracy on test set

Next Steps

- Keras NN

- More features

- Other ML Techniques: kNN, SVM, RandomForest

- Pipeline for predicting new matches

- In-depth evaluation

Thank you for listening.

Joseph Marchant