Predicting Soccer Match Results With Bet Data

Zhepei Wang, Zhenghan Zhang

What is gambling bet?

Premier League

Date | League

Today - 5th Of Dec		Home	Draw	Away	
20:00	Middlesbrough v Hull	4/5	12/5	4/1	+127 More Bets
Saturday - 10th Of Dec		**Home**	**Draw**	**Away**	
12:30	Watford v Everton	2/1	9/4	7/5	+59 More Bets
15:00	Arsenal v Stoke	4/11	4/1	15/2	+60 More Bets
15:00	Burnley v Bournemouth	11/5	9/4	13/10	+59 More Bets
15:00	Hull v Crystal Palace	15/8	23/10	6/4	+59 More Bets
15:00	Swansea v Sunderland	11/10	11/5	11/4	+59 More Bets
17:30	Leicester v Man City	7/2	11/4	3/4	+59 More Bets

Source: http://www.paddypower.com/football/football-matches/premier-league

Data Source

- http://football-data.co.uk/data.php
- 10 betting companies
- 20 seasons
- 16 leagues across Europe
- ~350 matches per league
- 90% training, 10% validation

A Glimpse of Data

- Features (Gambling bet)
  - $x_i = (h_i, d_i, a_i)$
- Scores
  - $y_i = (HG_i, AG_i)$

Algorithms

- Naive Guessing
- Polynomial Regression
- Multi-class classification with Integrated Feature
- Support Vector Machine (SVM)

Data Processing - Score

- Calculate goal difference

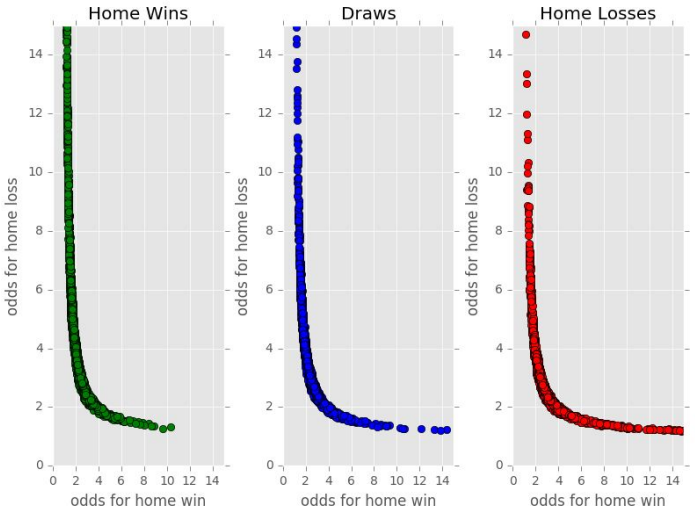
$$yd_i = HG_i - AG_i$$

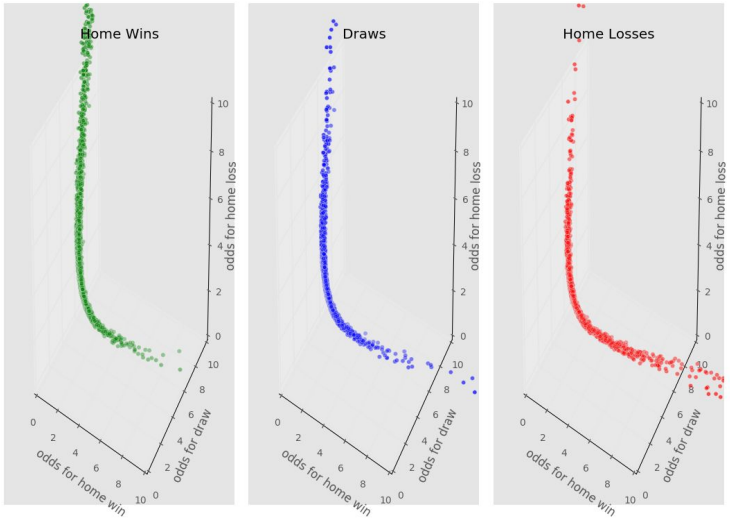
- Convert goal difference to match results

  - If $yd_i > 0$, we define $y_i = 1$, indicating a home win.
  - If $yd_i = 0$, we define $y_i = 0$, indicating a draw.
  - If $yd_i < 0$, we define $y_i = -1$, indicating a home loss.

Data Processing - Odds

- Downward sloping distribution
- 2D vs. 3D





Data Processing - Odds

- Combined feature for single company:

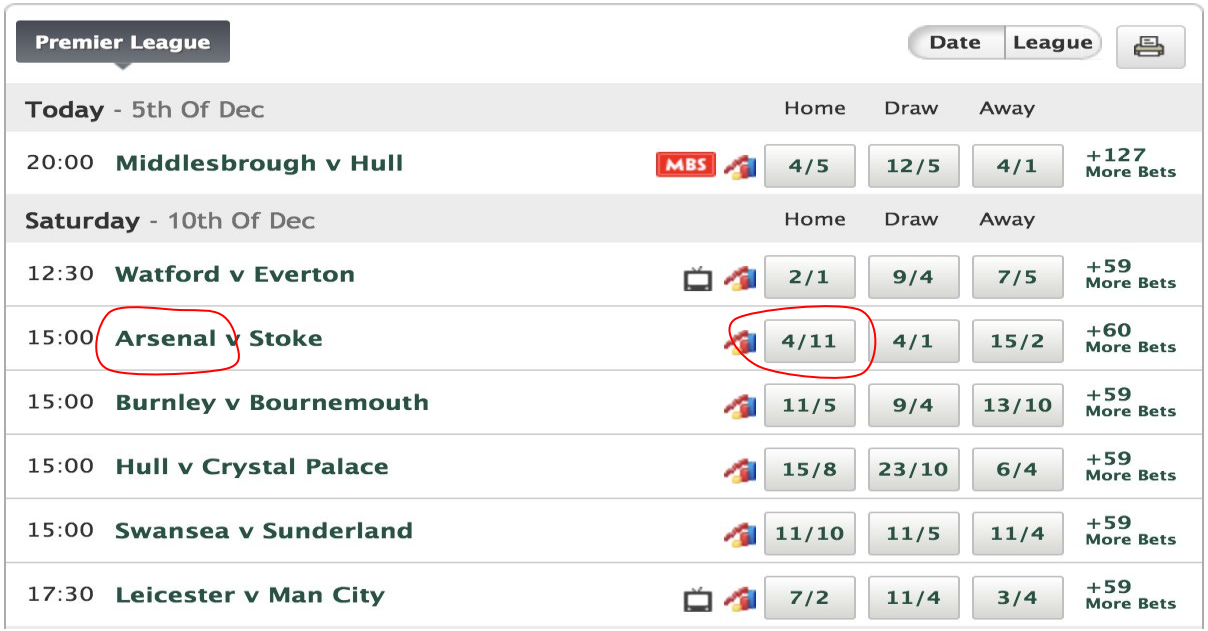
$$x_i = \log(\frac{1}{3}(\frac{h_i}{a_i} + \frac{h_i}{d_i} + \frac{d_i}{a_i}))$$

- Combined feature for multiple companies:

$$x_i = \log(\frac{1}{3}(\frac{E(h_i)}{E(a_i)} + \frac{E(h_i)}{E(d_i)} + \frac{E(d_i)}{E(a_i)}))$$

Naive Guessing

- Choose the smallest bet of the three
- No training
- Accuracy: **23.78%**



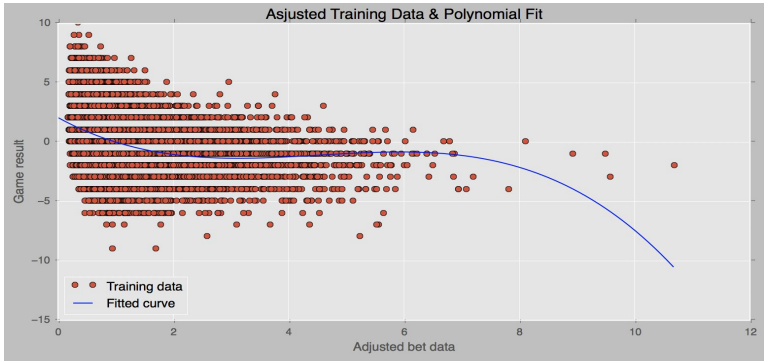
Premier League

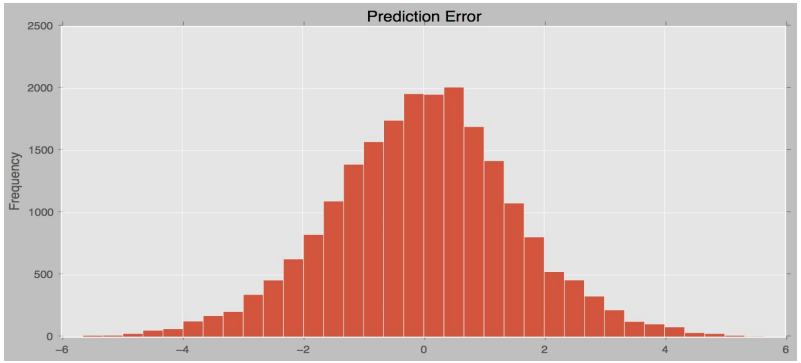
Date | League

Today - 5th Of Dec		Home	Draw	Away	
20:00	Middlesbrough v Hull	4/5	12/5	4/1	+127 More Bets
Saturday - 10th Of Dec		Home	Draw	Away	
12:30	Watford v Everton	2/1	9/4	7/5	+59 More Bets
15:00	Arsenal v Stoke	4/11	4/1	15/2	+60 More Bets
15:00	Burnley v Bournemouth	11/5	9/4	13/10	+59 More Bets
15:00	Hull v Crystal Palace	15/8	23/10	6/4	+59 More Bets
15:00	Swansea v Sunderland	11/10	11/5	11/4	+59 More Bets
17:30	Leicester v Man City	7/2	11/4	3/4	+59 More Bets

Polynomial Regression

- Use score difference instead of match result as y_i
- $y = -0.048x^3 + 0.653x^2 - 2.668x + 2.081$
- Convert back to match result w/ threshold of 0.5
- Accuracy: **44.43%**





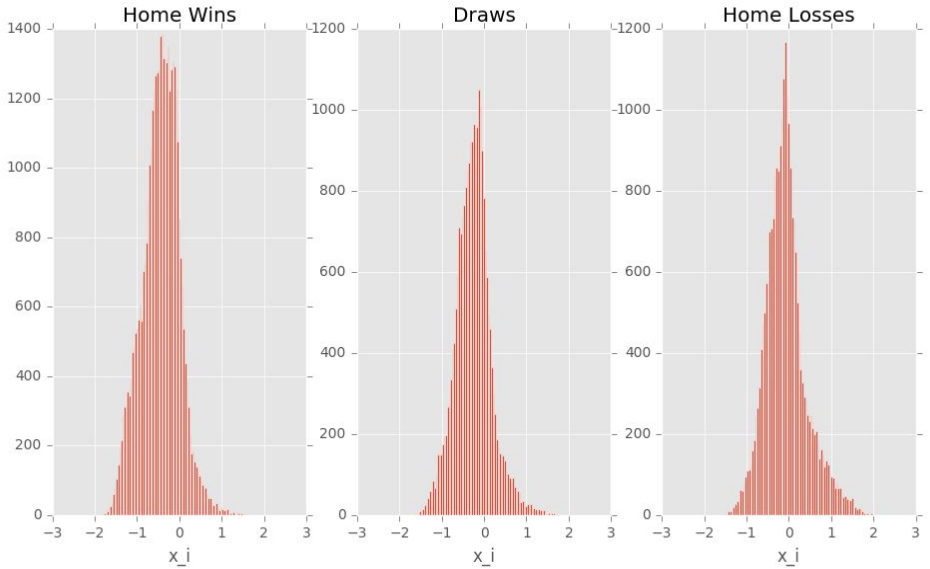
Multi-class Classification

- Recall the combined feature for multiple companies:

$$x_i = \log(\frac{1}{3}(\frac{E(h_i)}{E(a_i)} + \frac{E(h_i)}{E(d_i)} + \frac{E(d_i)}{E(a_i)}))$$

Multi-class Classification

- Generate $\mathcal{N}(\mathbb{E}_k[x], \mathbb{V}_k[x])$
- Calculate pdf for each distribution
- Accuracy: **48.14%**



Support Vector Machine

- RBF Kernel (Radial Basis Kernel)
- Three features
  - $x_1 = h_i^2$
  - $x_2 = \sqrt{d_i}$
  - $x_3 = a_i$
- Accuracy: **49.17%**

Discussion

Algorithms	Accuracy
Polynomial Regression	44.43%
Classification with Integrated Feature (Gaussian)	48.14%
Classification with Integrated Feature (Laplace)	47.54%
Support Vector Machine	49.17%

- Classification algorithms are more applicable
- Support vector machine algorithm has the best prediction accuracy
- Classification with a single integrated feature has the best performance

Next Steps

- Expand on existing algorithms
- Extend features
  - Country
  - Competitiveness
- Parallel data processing