

Moneyball: Minimizing Luck in Sports Betting

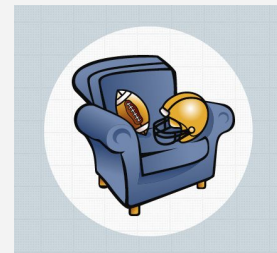
Kevin Cole, Metis NYC Summer '16



The Question, The Data

- Can we predict the if the home team will beat the Vegas spread?

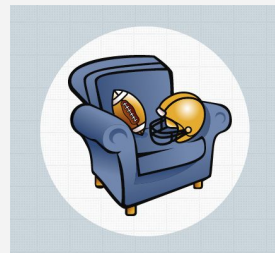
The Question, The Data



- Can we predict the if the home team will beat the Vegas spread?
- Data obtained from private provider **Armchair Analysis**
 - From 2000-2015 (over 3,000 games)



The Question, The Data



- Can we predict the if the home team will beat the Vegas spread?
- Data obtained from private provider Armchair Analysis
 - From 2000-2015 (over 3,000 games)
- Loaded into AWS & PostGreSQL, joined game & team data



PostgreSQL

The Question, The Data



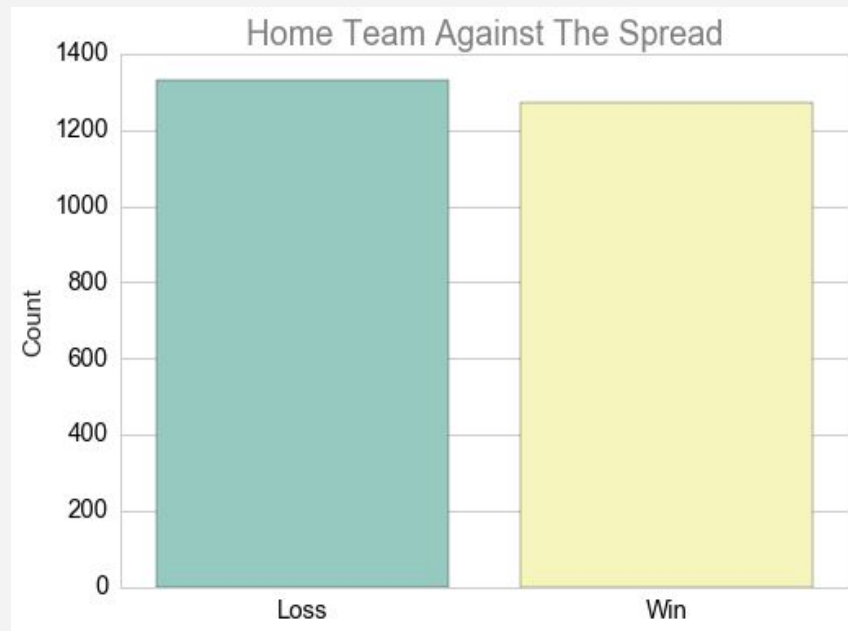
- Can we predict the if the home team will beat the Vegas spread?
- Data obtained from private provider Armchair Analysis
 - From 2000-2015 (over 3,000 games)
- Loaded into AWS & PostGreSQL, joined game & team data

gid	seas	wk	day	v	h	stad	temp	humd	wspd	wdir	cond	surf	ou	sprv	ptsv	ptsh	ptdiffh	beatsprh
1	2000	1	SUN	SF	ATL	Georgia Dome	79	None	None	None	Dome	AstroTurf	42.5	7.0	28	36	8	1.0

Preliminary Data Exploration



Preliminary Data Exploration

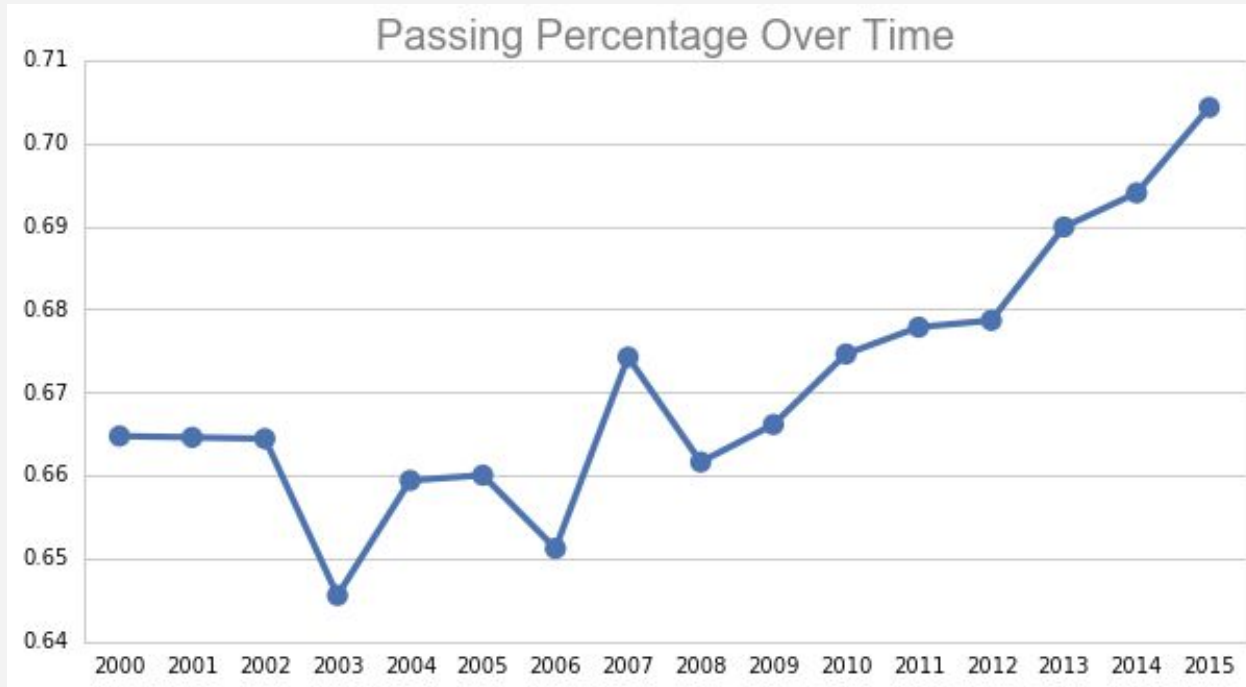


Creating Additional Features

- 25 historical game-level statistics for each teams, plus opponents
- 10 derived efficiency statistics
- Calculated rolling averages, periods of 1, 2, 3, 4, 5 & 15 games
- Total of 758 features

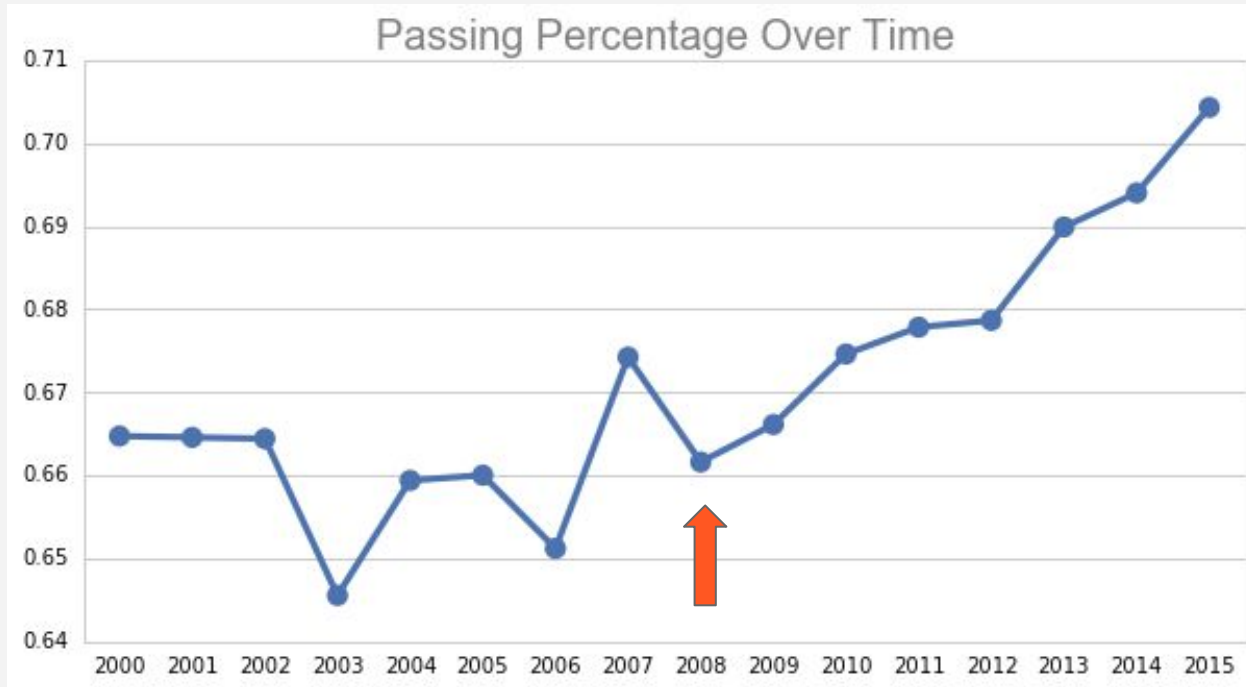
Adjusting the Data

- Narrowed data to last five years (2011-2015), 980 games



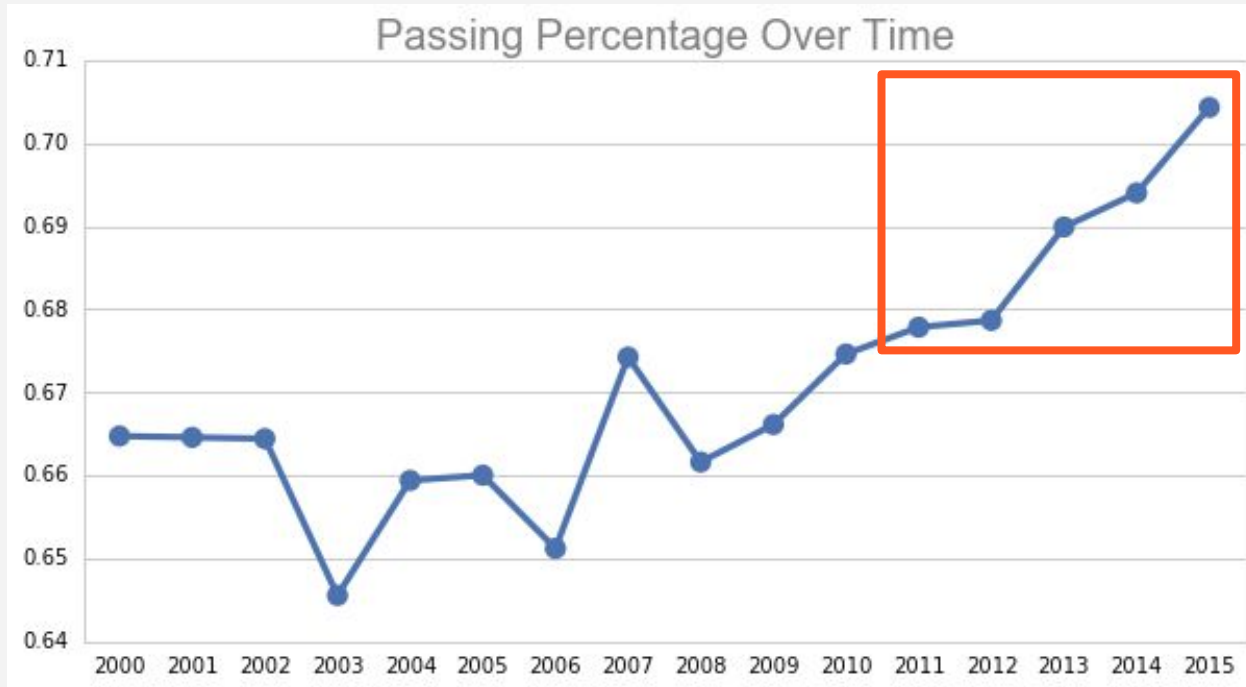
Adjusting the Data

- Narrowed data to last five years (2011-2015), 980 games



Adjusting the Data

- Narrowed data to last five years (2011-2015), 980 games



Parameter Tuning & Feature Elimination

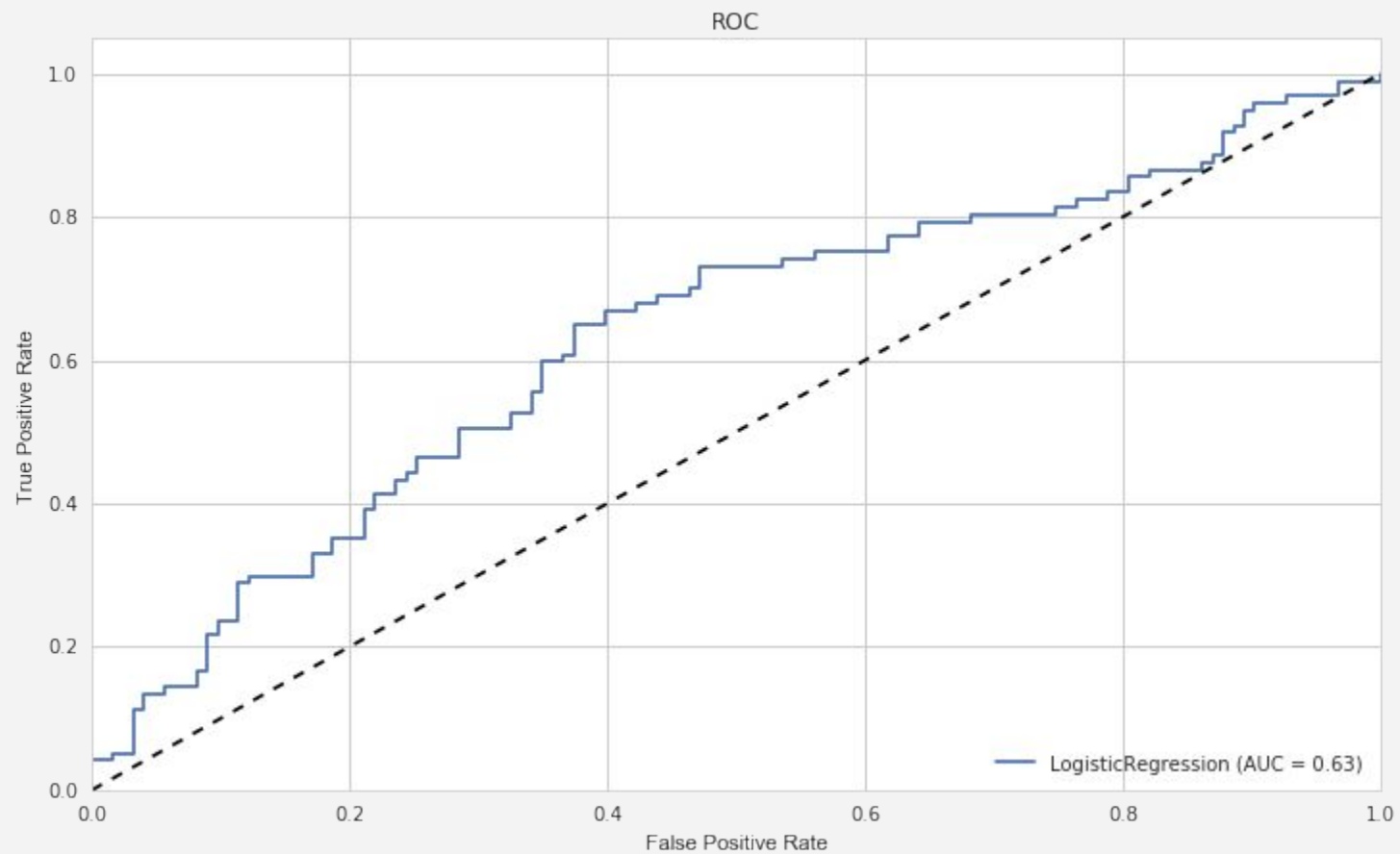
	Random Forest	Logistic Regression	SVC
Mean Cross-Validation	0.512		0.508
Accuracy Score	0.505		0.523

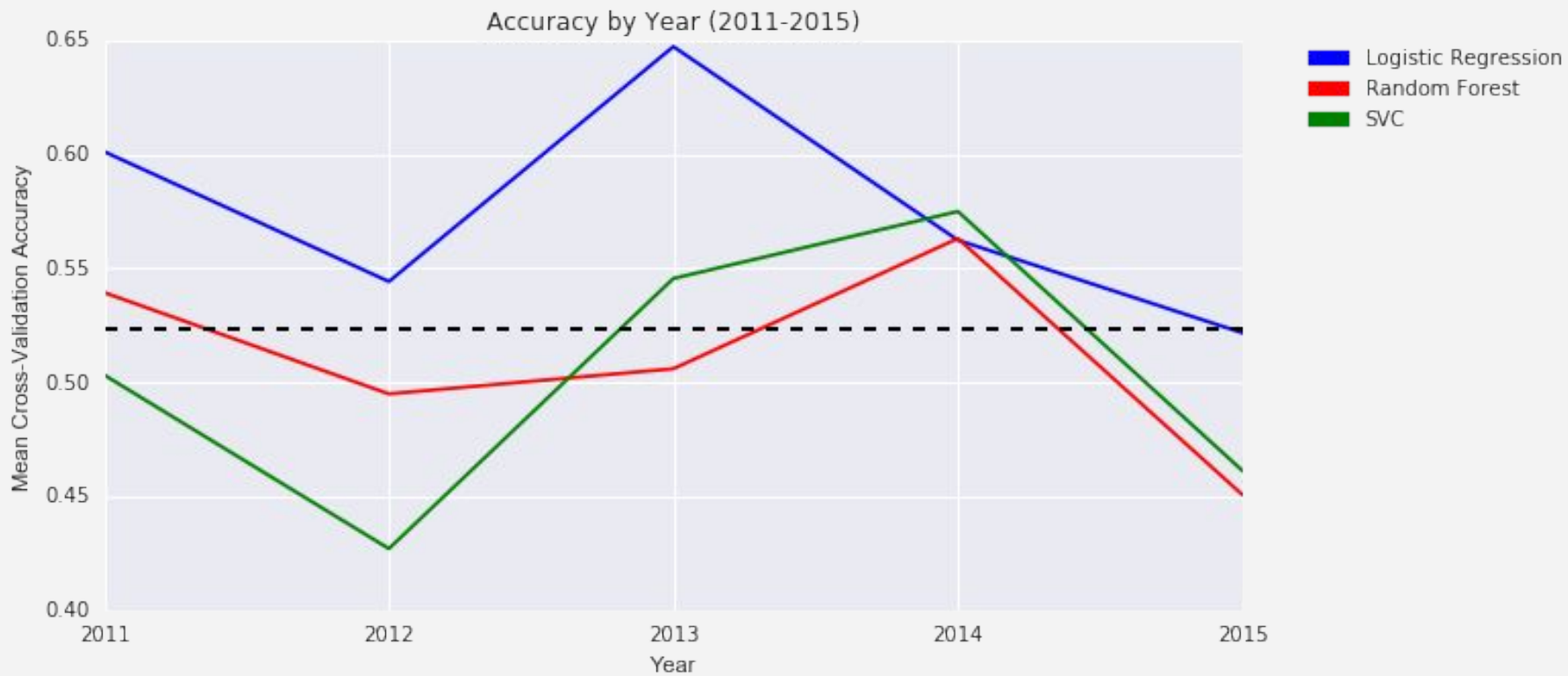
Parameter Tuning & Feature Elimination

	Random Forest	<i>Logistic Regression</i>	SVC
Mean Cross-Validation	0.512	0.565	0.508
Accuracy Score	0.505	0.573	0.523

Logistic Regression Best Features

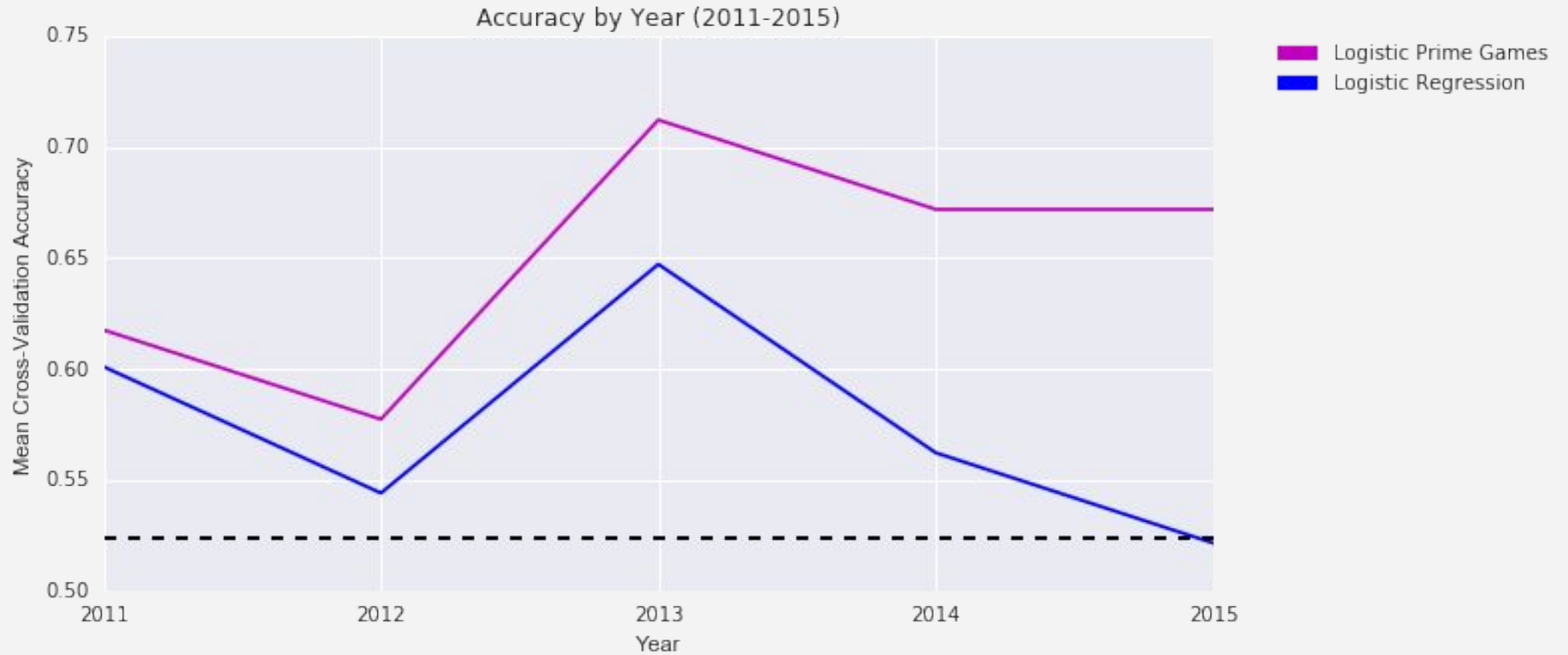
Statistic	Team	Own or Opponent	Time Period
Points Over Spread	Home	Own	15 Games
Kick Return Yards	Home	Own	15 Games
4th Quarter Points	Home	Own	4 Games
QB Rushing Atts	Home	Opponent	5 Games
Wins vs Spread	Vistor	Own	15 Games
Pass Attempts	Vistor	Own	2 Games
Pass Yards	Vistor	Own	5 Games
Rushing Yards	Vistor	Opponent	15 Games





Finding Prime Betting Matchups

- **Practically you can't bet on every game, nor would you want to**
 - Filtered testing data to highest (>0.60) & lowest (<0.40) probabilities estimates
 - 71 of 220 games or roughly 1/3
 - 256 games per season, 15 per week, 5 prime matchups per week
- **Accuracy score of 0.64**



Thoughts, Potential Next Steps

- **Game is rapidly changing**
 - Trends could lower accuracy of model when applied to future games
- **Volatility of best features**
 - Could vary by week of season, home/away, favorite/underdog
- **Monte Carlo**
 - Expand the data set through simulation, focusing on recent trends