Predicting the use of the Sacrifice Bunt in Major League Baseball

Charlie Gallagher
Brian Gilbert
Neelay Mehta
Chao Rao

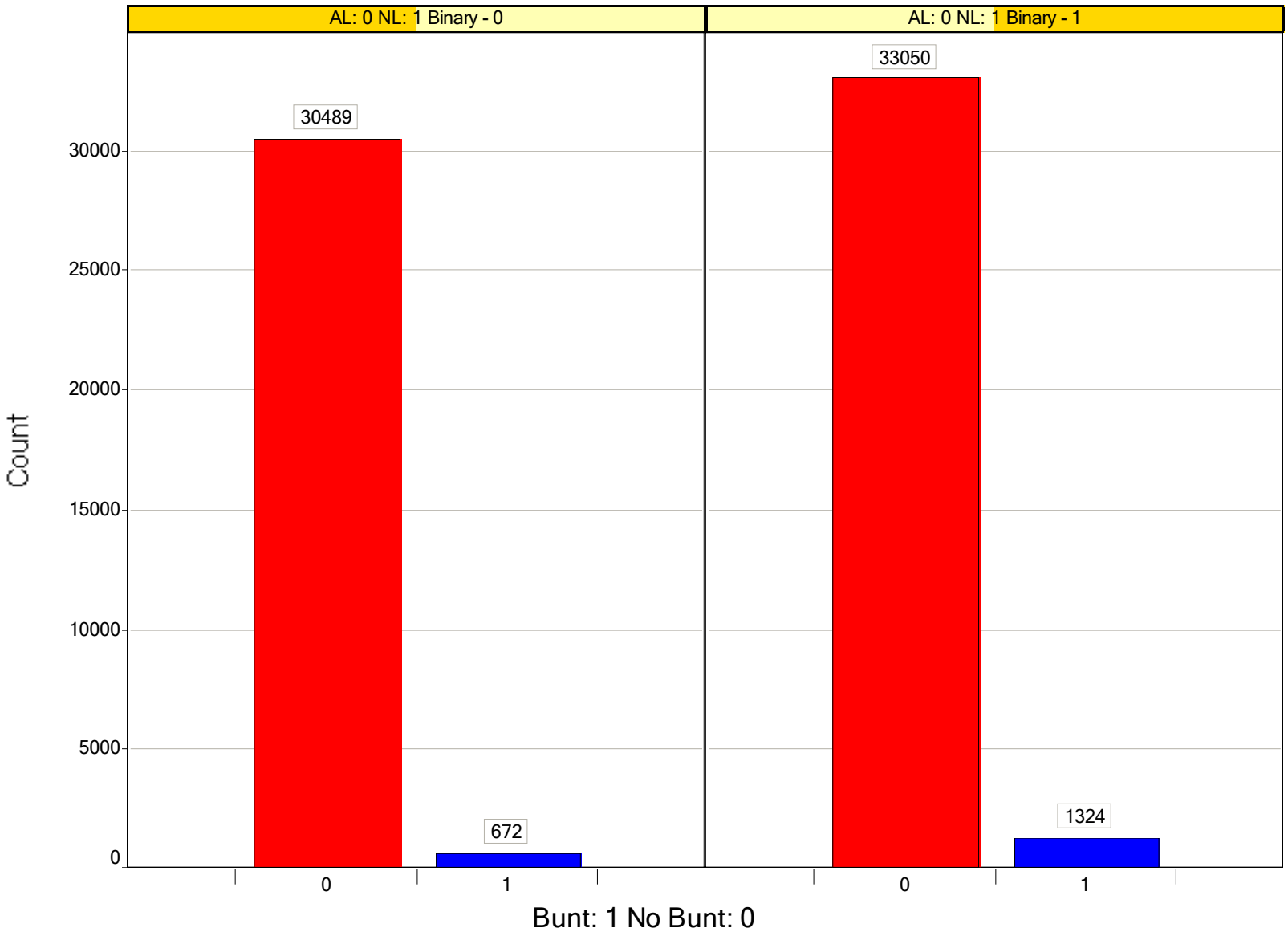
Understanding the Data

- Data from the St. Louis Cardinals
  - Sig Mejdal, Senior Quantitative Analyst
- Consists of plate appearances with men on base
  - The teams involved in the occurrence
  - Inning/Score/# of Outs/Batter
- Need to determine which variables matter the most in bunt prediction

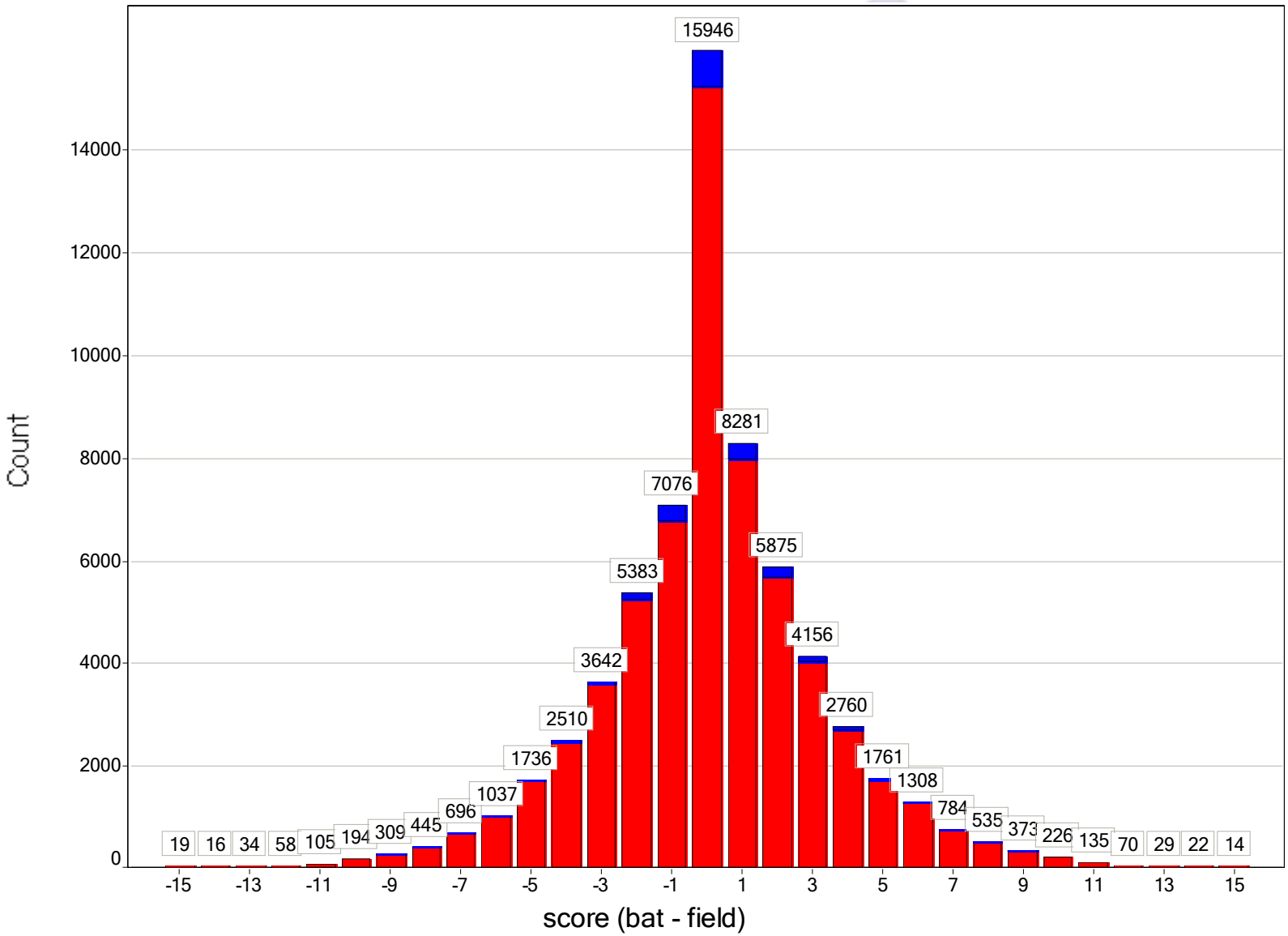
Data Pre-Processing Challenges

- Partitioning the data
- Numerical vs. Categorical
- Eliminating variables
- Creating dummy variables
- Binning variables
- Deriving new variables
- Using outside data sources to validate our assumptions
  - Baseball-reference.com

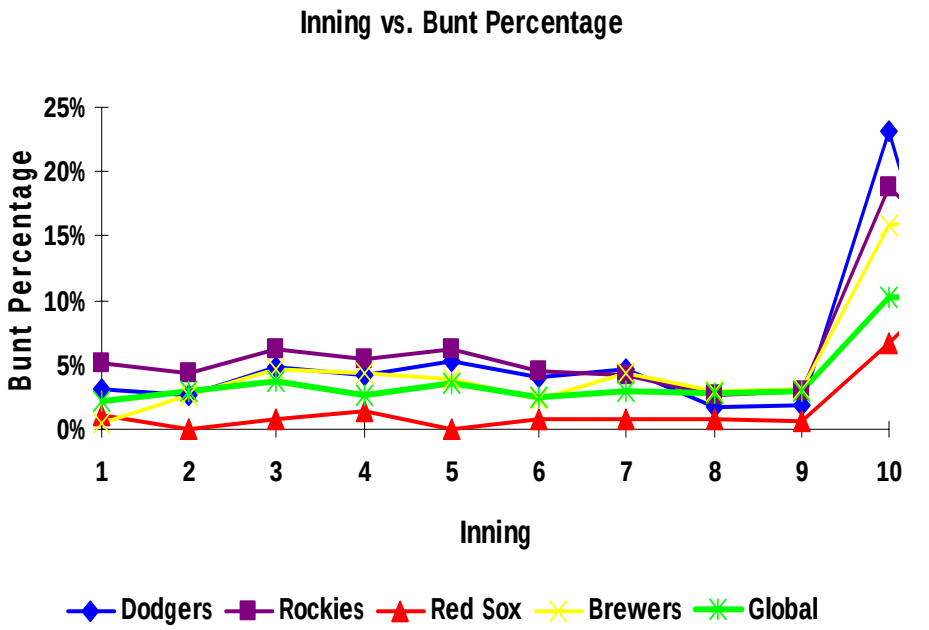
Naïve Rule and AL/NL Difference

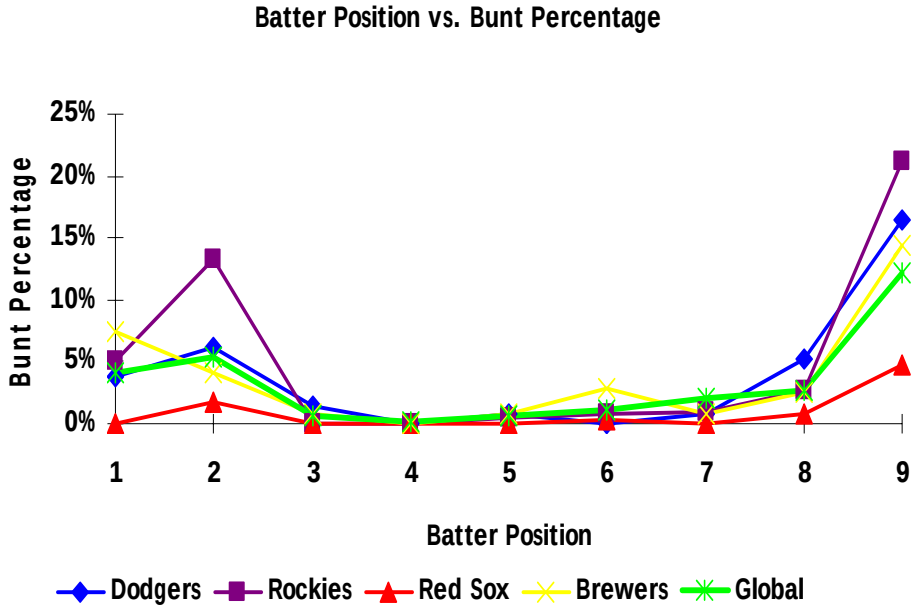


The Significance of Score Difference



Visualizing the Data

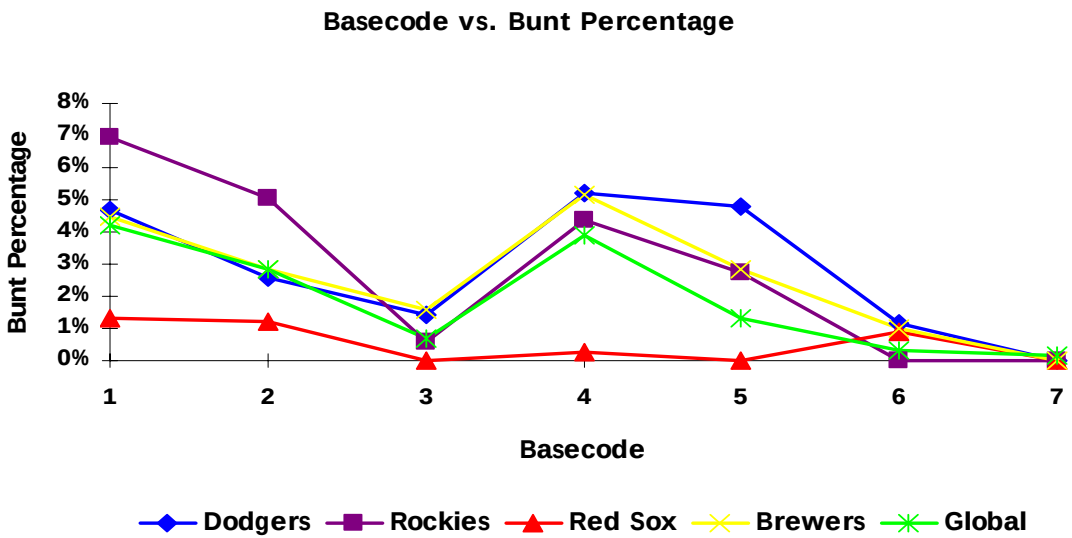




Basecode

Runners on:

1. 1st
2. 2nd
3. 3rd
4. 1st & 2nd
5. 1st & 3rd
6. 2nd & 3rd
7. 1st, 2nd & 3rd



Methodology

- Classification Trees
  - Desired by client
  - Low bunt frequency led to difficulties
  - Over-sampling helped, but still not accurate
- Logistic Regression: Global vs. Team-based
  - Initially, team-based models looked most ideal
  - Team-based models much more parsimonious and but not as accurate as global models

The Best Model

- Global Logistic Regression - Best Subset
  - Cutoff of 0.5
  - Validation Error Rate: 3.00%
  - NYY & CWS: 2.72% Error

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.91484118	0.19527394	0	*
Inning 1-9: 0 10+: 1	1.43896341	0.31070757	0.00000363	4.2163229
out.before	-2.06532502	0.12200621	0	0.12677707
basecode.before_3	-1.18709564	0.52609891	0.02404486	0.3051061
basecode.before_4	0.47677508	0.16713402	0.00433562	1.61087108
basecode.before_6	-2.63196373	1.02573335	0.01028984	0.07193705
basecode.before_7	-2.14123392	0.72531897	0.00315593	0.11750975
binned score diff_2	0.66862828	0.20002525	0.00082962	1.95155859
binned score diff_5+	-0.85927516	0.3422673	0.01205473	0.42346892
binned bpos_3-7	-1.65075326	0.1863011	0	0.1919053
binned bpos_9	1.85490274	0.15460882	0	6.39107656
Top Bunter	1.51956904	0.26843038	0.00000002	4.57025528
Team binned_2	0.90669906	0.20409924	0.00000889	2.47613549
Team binned_3	1.00761068	0.20899259	0.00000143	2.73904896
Team binned_4	1.47704506	0.21063162	0	4.3799839

Residual df	9985
Residual Dev.	1757.736206
% Success in training data	3.17
# Iterations used	9
Multiple R-squared	0.3749283

Performance Metrics

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	226	1317
0	184	48273

Error Report			
Class	# Cases	# Errors	% Error
1	1543	1317	85.35
0	48457	184	0.38
Overall	50000	1501	3.00

Elapsed Time

Overall (secs)	593.00

Caveats in the analysis

- Data Purged
  - 2004 Season
- Potential important data not provided
  - Opposing pitcher, weather…
- Naïve Rule is tough for a model to "beat"
  - Sacrifice bunts not a common occurrence
- Need more power!
  - More powerful software could have made the analysis more manageable

Final Thoughts

- Domain Knowledge extremely powerful
- Complicated Models: Marginal Improvement over Naïve Rule
- Cost-Benefit: How much is the model worth in wins compared to using the Naïve Rule?
  - Predict approximately 2-3 more bunts, prevent 1 run over the course of a season
  - Assume a competent manager's domain knowledge would be far more effective

Justification for binning teams

Red Sox	20	2375	0.84%
Blue Jays	26	2213	1.17%
Athletics	29	2296	1.26%
Rangers	28	2056	1.36%
Yankees	41	2193	1.87%
Devil Rays	41	2045	2.00%
Orioles	49	2288	2.14%
Indians	52	2328	2.23%
Royals	52	2012	2.58%
Mariners	58	2197	2.64%
Padres	62	2259	2.74%
Twins	58	2081	2.79%
Phillies	68	2221	3.06%
Angels	73	2235	3.27%
Tigers	72	2137	3.37%
Reds	71	2042	3.48%
Diamondba	70	1957	3.58%
Brewers	74	2037	3.63%
Cardinals	80	2133	3.75%
Giants	85	2204	3.86%
Braves	82	2100	3.90%
Dodgers	81	2051	3.95%
White Sox	79	1981	3.99%
Mets	81	2007	4.04%
Marlins	84	2056	4.09%
Pirates	86	2034	4.23%
Cubs	89	1935	4.60%
Astros	97	2076	4.67%
Rockies	106	2130	4.98%
Nationals	102	1860	5.48%