MLB Game Predictor

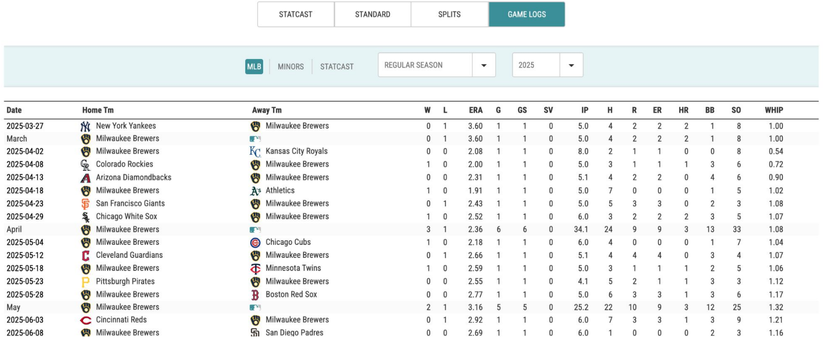
By Tymon Vu and Ian Wong

Project Goals - “Why Should You Care?”

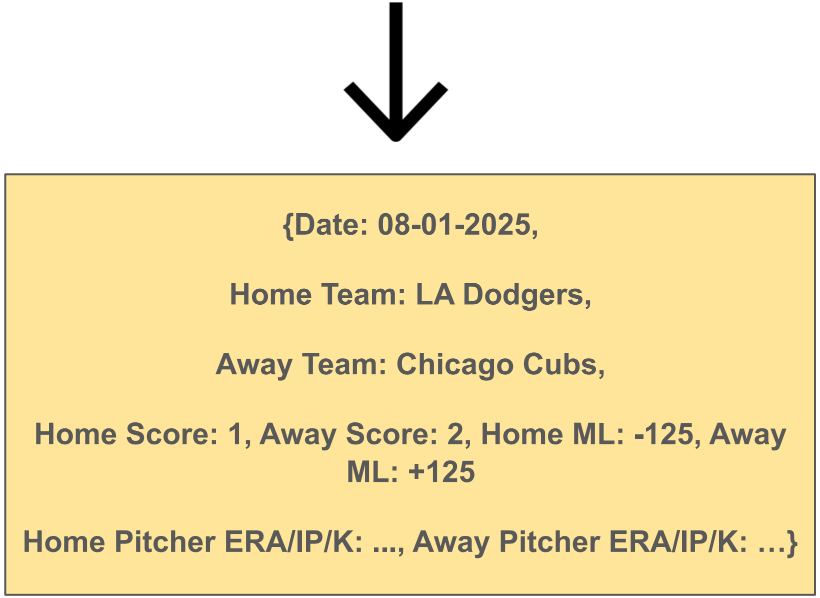
- Test how effective a simple ML model can be at predicting MLB game outcomes (Conquer the sports betting world potentially??)
  - Compare how similar odds are to Vegas
  - Predict Winners and Losers correctly
- Identify how pitching stats and team trends influence game outcomes
  - Pitchers controls the rhythm and pace of game, making them the most important player on the field
  - Widely available data on pitching matchups and statistics
- Evaluate and compare multiple models (XGBoost, LR, RF) for specific use case

How we got our data

- MLB Schedule/Final Scores + Vegas odds sourced from “Shane McDonald’s Online Repo”
- Scraped pitcher data from Baseball Savant
- Calculated rolling pre-game stats so each game only uses information that existed **up to that date**



Date	Home Tm	Away Tm	W	L	ERA	G	GS	SV	IP	H	R	ER	HR	BB	SO	WHIP
2025-03-27	New York Yankees	Milwaukee Brewers	0	1	3.60	1	1	0	5.0	4	2	2	2	1	8	1.00
March	Milwaukee Brewers		0	1	3.60	1	1	0	5.0	4	2	2	2	1	8	1.00
2025-04-02	Milwaukee Brewers	Kansas City Royals	0	0	2.08	1	1	0	8.0	2	1	1	0	0	8	0.54
2025-04-08	Colorado Rockies	Milwaukee Brewers	1	0	2.00	1	1	0	5.0	3	1	1	1	3	6	0.72
2025-04-13	Arizona Diamondbacks	Milwaukee Brewers	0	0	2.31	1	1	0	5.1	4	2	2	0	4	6	0.90
2025-04-18	Milwaukee Brewers	Athletics	1	0	1.91	1	1	0	5.0	7	0	0	0	1	5	1.02
2025-04-23	San Francisco Giants	Milwaukee Brewers	0	1	2.43	1	1	0	5.0	5	3	3	0	2	3	1.08
2025-04-29	Chicago White Sox	Milwaukee Brewers	1	0	2.52	1	1	0	6.0	3	2	2	2	3	5	1.07
April	Milwaukee Brewers		3	1	2.36	6	6	0	34.1	24	9	9	3	13	33	1.08
2025-05-04	Milwaukee Brewers	Chicago Cubs	1	0	2.18	1	1	0	6.0	4	0	0	0	1	7	1.04
2025-05-12	Cleveland Guardians	Milwaukee Brewers	0	1	2.66	1	1	0	5.1	4	4	4	0	3	4	1.07
2025-05-18	Milwaukee Brewers	Minnesota Twins	1	0	2.59	1	1	0	5.0	3	1	1	1	2	5	1.06
2025-05-23	Pittsburgh Pirates	Milwaukee Brewers	0	0	2.55	1	1	0	4.1	5	2	1	1	3	3	1.12
2025-05-28	Milwaukee Brewers	Boston Red Sox	0	0	2.77	1	1	0	5.0	6	3	3	1	3	6	1.17
May	Milwaukee Brewers		2	1	3.16	5	5	0	25.2	22	10	9	3	12	25	1.32
2025-06-03	Cincinnati Reds	Milwaukee Brewers	0	1	2.92	1	1	0	6.0	7	3	3	1	3	9	1.21
2025-06-08	Milwaukee Brewers	San Diego Padres	0	0	2.69	1	1	0	6.0	1	0	0	0	2	3	1.16



{Date: 08-01-2025,

Home Team: LA Dodgers,

Away Team: Chicago Cubs,

Home Score: 1, Away Score: 2, Home ML: -125, Away ML: +125

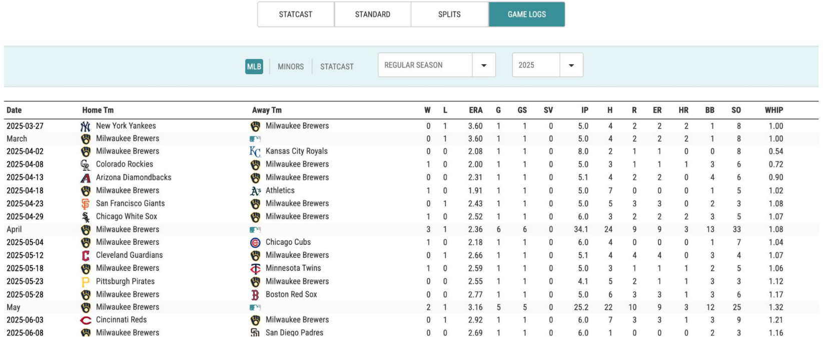
Home Pitcher ERA/IP/K: ..., Away Pitcher ERA/IP/K: …}

How we got our data

- **MLB Schedule/Final Scores** + **Vegas odds** from *Shane McDonald's Online Repo*

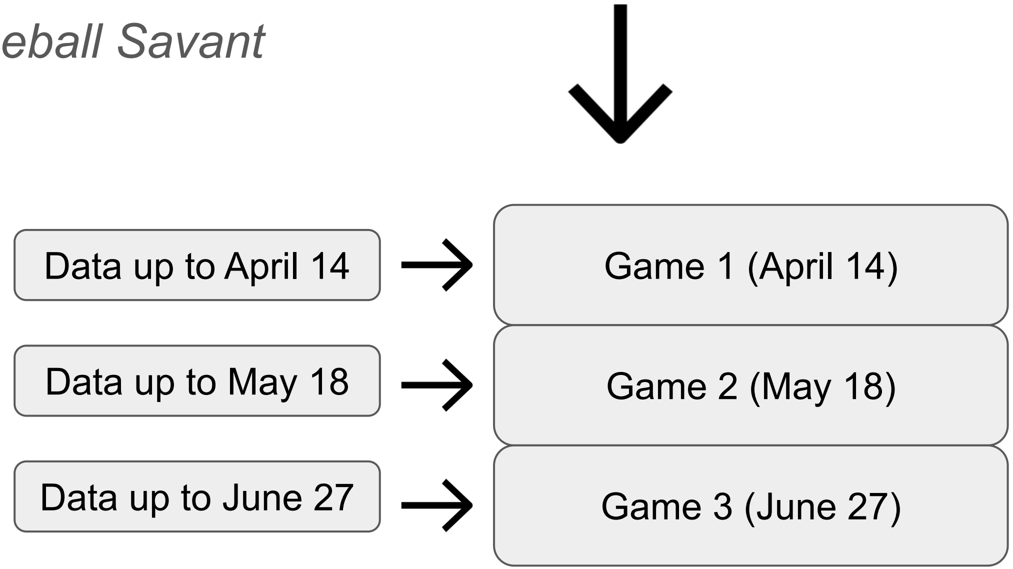
- Scraped **pitcher data** from *Baseball Savant*

- Calculated rolling pre-game stats so each game only uses information that existed **up to that date**



Date	Home Tm	Away Tm	W	L	ERA	G	GS	SV	IP	H	R	ER	HR	BB	SO	WHIP
2025-03-27	New York Yankees	Milwaukee Brewers	0	1	3.60	1	1	0	5.0	4	2	2	2	1	8	1.00
March	Milwaukee Brewers		0	1	3.60	1	1	0	5.0	4	2	2	2	1	8	1.00
2025-04-02	Milwaukee Brewers	Kansas City Royals	0	0	2.08	1	1	0	8.0	2	1	1	0	0	8	0.54
2025-04-08	Colorado Rockies	Milwaukee Brewers	1	0	2.00	1	1	0	5.0	3	1	1	1	3	6	0.72
2025-04-13	Arizona Diamondbacks	Milwaukee Brewers	0	0	2.31	1	1	0	5.1	4	2	2	0	4	6	0.90
2025-04-18	Milwaukee Brewers	Athletics	1	0	1.91	1	1	0	5.0	7	0	0	0	1	5	1.02
2025-04-23	San Francisco Giants	Milwaukee Brewers	0	1	2.43	1	1	0	5.0	5	3	3	0	2	3	1.08
2025-04-29	Chicago White Sox	Milwaukee Brewers	1	0	2.52	1	1	0	6.0	3	2	2	2	3	5	1.07
April	Milwaukee Brewers		3	1	2.36	6	6	0	34.1	24	9	9	3	13	33	1.08
2025-05-04	Milwaukee Brewers	Chicago Cubs	1	0	2.18	1	1	0	6.0	4	0	0	0	1	7	1.04
2025-05-12	Cleveland Guardians	Milwaukee Brewers	0	1	2.66	1	1	0	5.1	4	4	4	0	3	4	1.07
2025-05-18	Milwaukee Brewers	Minnesota Twins	1	0	2.59	1	1	0	5.0	3	1	1	1	2	5	1.06
2025-05-23	Pittsburgh Pirates	Milwaukee Brewers	0	0	2.55	1	1	0	4.1	5	2	1	1	3	3	1.12
2025-05-28	Milwaukee Brewers	Boston Red Sox	0	0	2.77	1	1	0	5.0	6	3	3	1	3	6	1.17
May	Milwaukee Brewers		2	1	3.16	5	5	0	25.2	22	10	9	3	12	25	1.32
2025-06-03	Cincinnati Reds	Milwaukee Brewers	0	1	2.92	1	1	0	6.0	7	3	3	1	3	9	1.21
2025-06-08	Milwaukee Brewers	San Diego Padres	0	0	2.69	1	1	0	6.0	1	0	0	0	2	3	1.16

online data



Data up to April 14 → Game 1 (April 14)

Data up to May 18 → Game 2 (May 18)

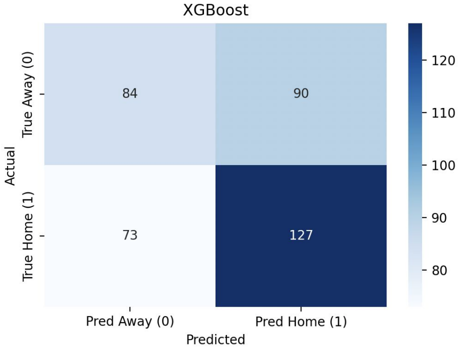
Data up to June 27 → Game 3 (June 27)

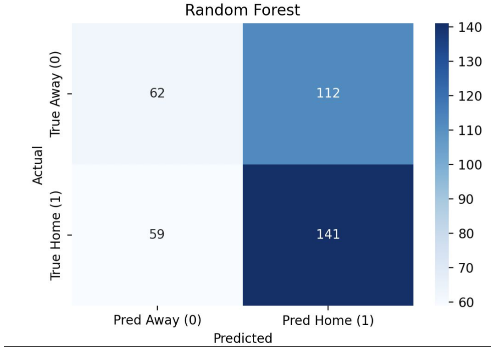
Feature Engineering

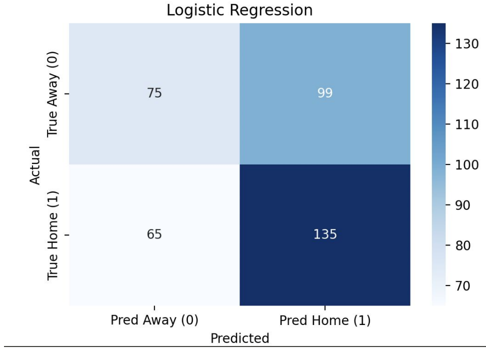
- Models fed in:
    - Away + Home Pitcher ERA - How many runs a pitcher gives up
    - Away + Home Pitcher K9 - How many strikeouts pitcher gets
    - Away + Home Pitcher BB9 - How many walks pitcher gets
    - WHIP + IP - Baseball pitches + innings pitched
- Each stat has also rolling average of last three games before
    - Account if a baseball pitcher is “hot” or whether they are in a “slump”

Models Compared

Model	Accuracy
Vegas Heuristic	0.543
Logistic Regression	0.561
Random Forest	0.542
XGBoost	0.564
Random	0.495





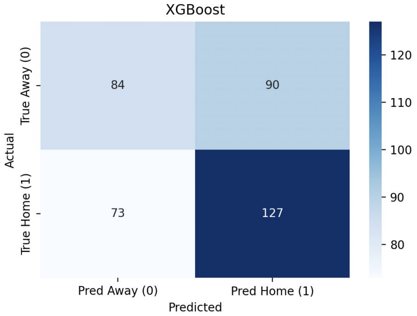


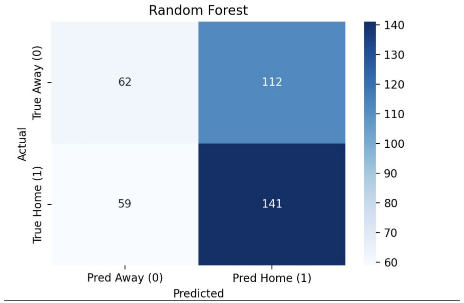
- Train/Test Split - Around 10% Test (Games after September 1st) -> Sequential Data

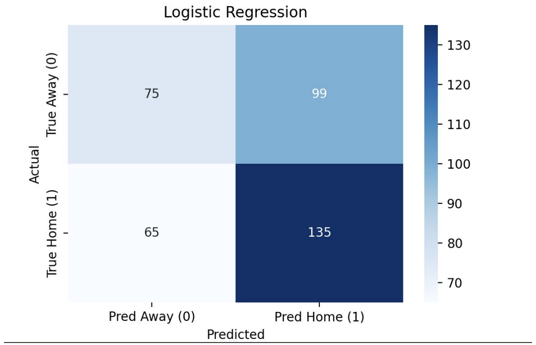
- All Models (including Vegas Heuristic) predict heavily that “Home Teams” win a lot more

Models Compared

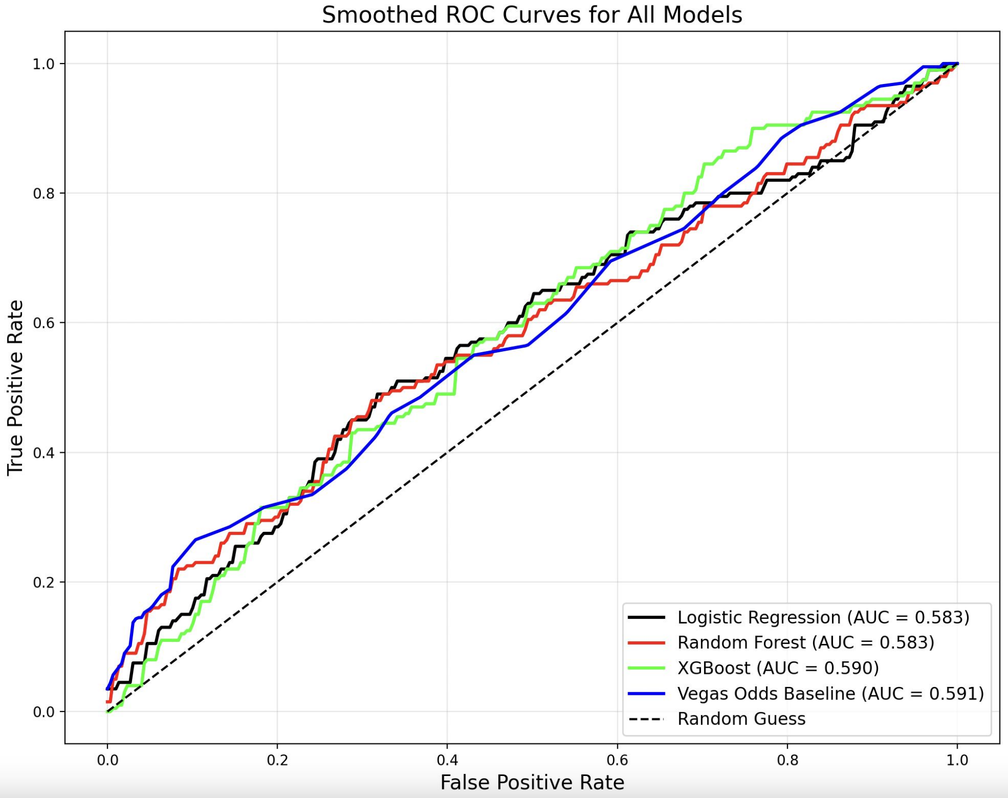
Model	Accuracy
Vegas Heuristic	0.543
Logistic Regression	0.561
Random Forest	0.542
XGBoost	0.564
Random	0.495





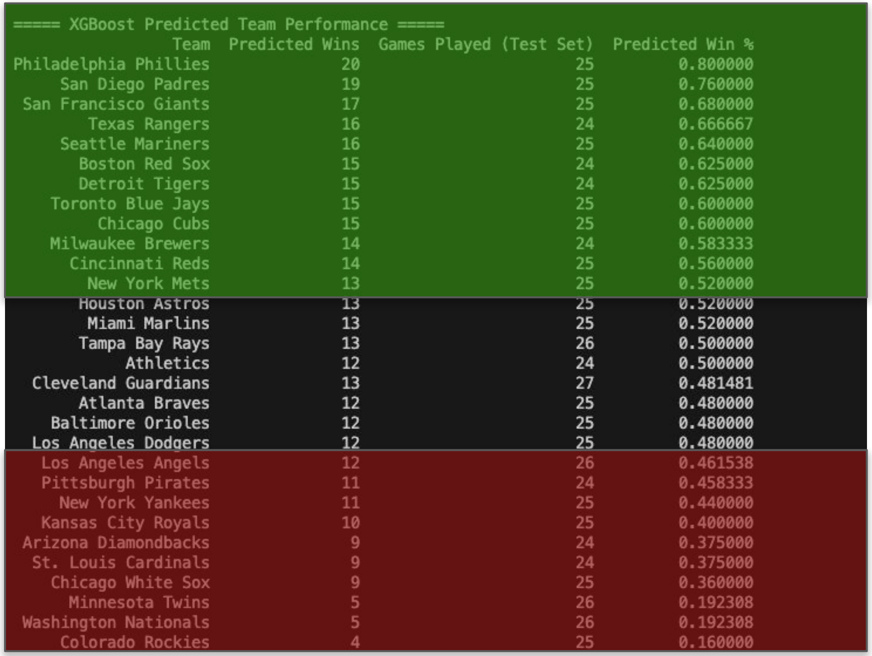


Models Compared



Model Conclusions

- XGBoost seems most balanced model for accuracy and ROC-AUC
- Correctly predicted 9/12 teams to make postseason
- Correctly predicted most of the terrible teams to have the worst performances (e.g., Rockies, Nationals, White Sox)



```
===== XGBoost Predicted Team Performance =====
                 Team  Predicted Wins  Games Played (Test Set)  Predicted Win %
Philadelphia Phillies              20                       25         0.800000
     San Diego Padres              19                       25         0.760000
 San Francisco Giants              17                       25         0.680000
        Texas Rangers              16                       24         0.666667
     Seattle Mariners              16                       25         0.640000
       Boston Red Sox              15                       24         0.625000
       Detroit Tigers              15                       24         0.625000
    Toronto Blue Jays              15                       25         0.600000
         Chicago Cubs              15                       25         0.600000
    Milwaukee Brewers              14                       24         0.583333
      Cincinnati Reds              14                       25         0.560000
        New York Mets              13                       25         0.520000
       Houston Astros              13                       25         0.520000
        Miami Marlins              13                       25         0.520000
       Tampa Bay Rays              13                       26         0.500000
            Athletics              12                       24         0.500000
  Cleveland Guardians              13                       27         0.481481
       Atlanta Braves              12                       25         0.480000
    Baltimore Orioles              12                       25         0.480000
  Los Angeles Dodgers              12                       25         0.480000
   Los Angeles Angels              12                       26         0.461538
   Pittsburgh Pirates              11                       24         0.458333
     New York Yankees              11                       25         0.440000
   Kansas City Royals              10                       25         0.400000
 Arizona Diamondbacks               9                       24         0.375000
  St. Louis Cardinals               9                       24         0.375000
    Chicago White Sox               9                       25         0.360000
      Minnesota Twins               5                       26         0.192308
 Washington Nationals               5                       26         0.192308
     Colorado Rockies               4                       25         0.160000
```

Closing Thoughts - “Why Should You Believe Us?”

- Very difficult to predict sports
  - Each game is an independent event
  - The MLB is very streaky and most forecasts are unreliable
- Even Vegas doesn’t get it right a lot of the time
  - Model outperforms Vegas using simple methods
- Our evaluations show the model is able to capture and use relevant signal
  - Produced reasonable accuracy and ROC-AUC given only pitcher stats
  - Model gives heavy bias towards home teams,
    which aligns with most other predictors

Thank You!