What affects the chances of a player successfully stealing a base?

...

By Enikő Gergely

Base Stealing Model

Predict the probability of a successful stolen base given that the runner takes off for an empty second base with no other runners on the bases.

Input Data

	Original data	Calculated/Added
Catcher	Pop Time, Exchange time	Historical SB success rate against (RTM), Arm strength
Pitcher	Plate Time, Handedness	Historical SB success rate against (RTM), Pitch type percentage (arsenal usage)
Runner	Lead Distance, Secondary Lead Distance	Historical SB success rate (RTM), Sprint speed
Other	Game state (score, inning, outs), Count (balls, strikes), Batter handedness	Lead by the defensive team
Target variable	Advance	

Model evaluation metric

Accuracy:

- data is a unbalanced, accuracy is not a good metric for evaluation

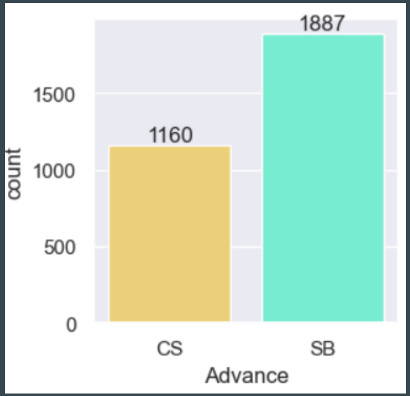
Precision or recall:

- As a goal we could maximize the number of stolen bases (high recall) but in this case we can end up with too many outs,
- or minimize caught steals (high precision) but we could miss too many steal opportunities and lose some additional runs.

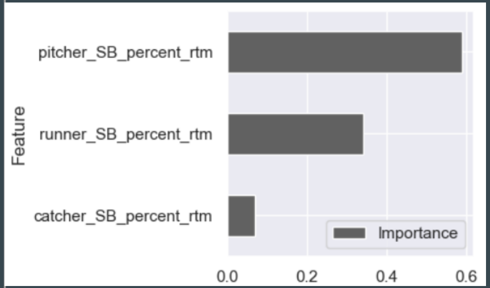
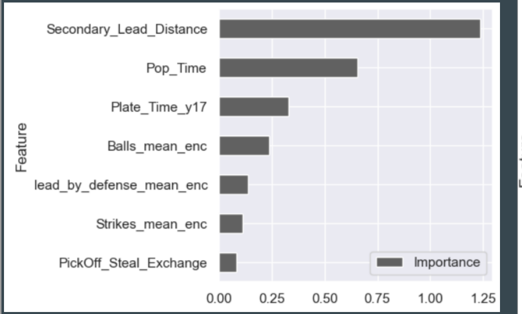
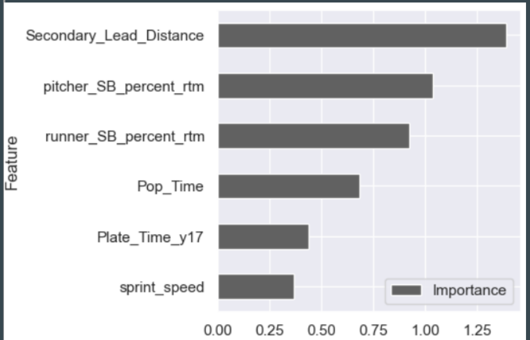
F1-score:

- balance between the two cases
- **Final Metric**

Number of CS and SB in data

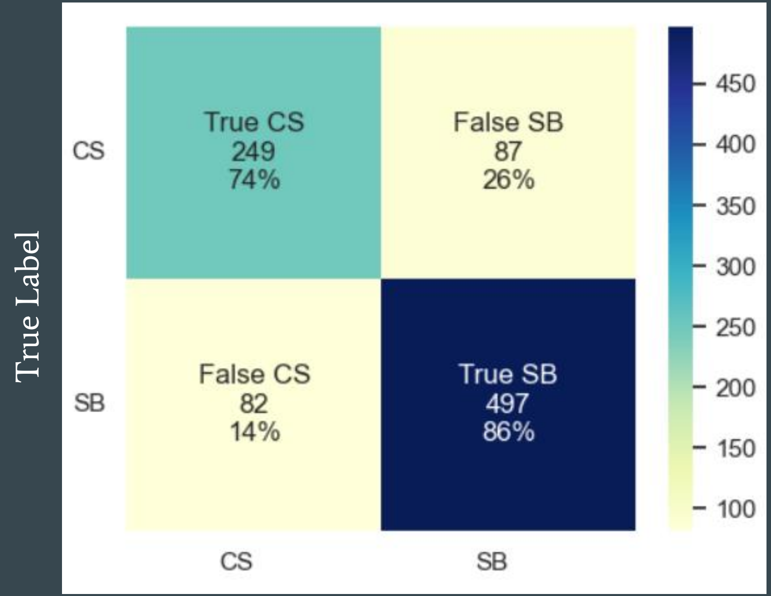


Model Results

<table>
<tr><th></th><th>Model 1: Historical SB success rates only</th><th>Model 2: original columns</th><th>Model 3: all data available</th></tr>
<tr><td>Model</td><td>Gradient Boosted Trees</td><td>Logistic Regression</td><td>Logistic Regression</td></tr>
<tr><td>F1-score</td><td>0.833</td><td>0.793</td><td>0.861</td></tr>
<tr><td>Feature Importance</td><td>
</td><td>
</td><td>
</td></tr>
</table>

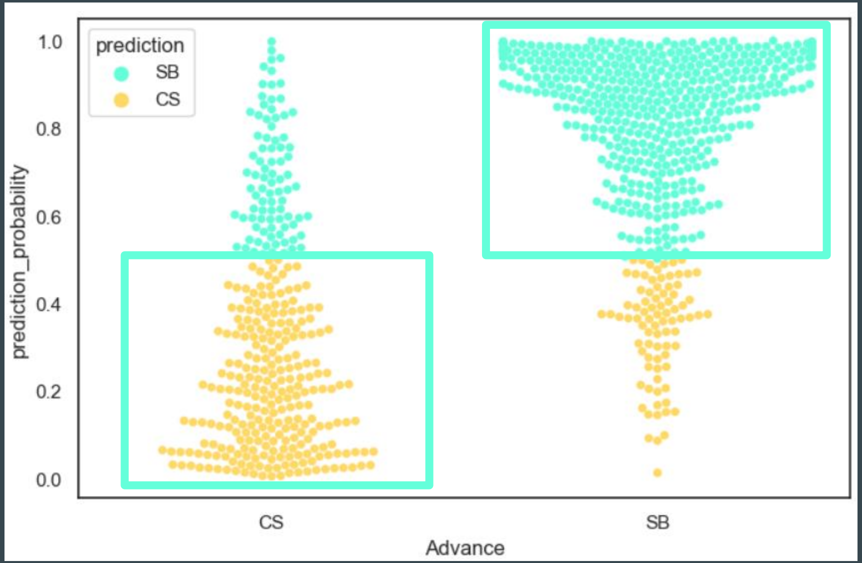
Model 3 Results: Confusion Matrix

Confusion matrix



Predicted Label

Confusion matrix with 0.5 cutoff value (probabilities larger than 0.5 will be categorized as SB and lower will be CS)



* correctly classified cases highlighted in green.

How to use it?

Using the run values we can calculate a break-even point where the steal attempt has 0 influence - zero expected change in the run expectancy.

If the model predicts higher probability → steal attempt can happen.

These break-even points can be different in different game situations:

- Average: 75% success
- 3-0 Count: 87% success , 15 attempt (alternative: BB can happen)
- 3-2 Count: 46% success , 388 attempt (alternative: K can happen)

Stolen Base Success rates by Count

Balls	Strikes	Advance	ThrownOut	total	SB_success
3	0	13	2	15	0.866667
2	0	110	26	136	0.808824
	2	283	71	354	0.799435
1	2	405	112	517	0.783366
0	1	489	136	625	0.782400
	0	836	238	1074	0.778399
	2	267	80	347	0.769452
3	1	26	8	34	0.764706
1	1	370	123	493	0.750507
	0	441	153	594	0.742424
2	1	190	71	261	0.727969
3	2	179	209	388	0.461340

Trends that might impact the model

The limited instances of stolen bases during the period of 2019-2021 suggest that teams likely opted for stealing bases when the odds were favorable – they expect a positive impact on their expected runs. Consequently, data on riskier stolen base scenarios is limited, leading to an increasing skew towards higher break-even points and the models can shift the game to less and less steals.

The significance of catcher input in the model is relatively minimal. Catchers had been more valued for their framing and less valued for the skill of avoiding stolen bases. This could indicate that catchers possess comparable skills in this aspect, leading to a reduced impact on the play's outcome.

How might the 2023 rule change influence this dynamic?

Thank you!

Questions?