Informing a Formula 1 Decision

A comprehensive overview of data-driven racing strategies

Ben Gonzalez, Ryan Mudhole, Matt Peters, Jack Wentworth

F1 Background

- Race durations are usually 50-70 laps
- Teams constantly monitor and analyze efficiency
- 3 qualifying rounds before race
- Pit stops average 2.5-3 seconds

Business Problem

Can we accurately predict which drivers will stand on an F1 Grand Prix podium, before the race starts?

- F1 is a $15 billion industry
- One podium can be worth millions in sponsorships, prize money, and exposure
- Teams already know who qualifies (starting grid)

Executive Summary

Our best-performing classification model accurately identified at least one podium finisher correctly 76.2% of the time. (Note for Marc: The last thing we are testing is a predictive fit for our “profit function” in terms of points gained/lost in F1 scoring terms)

Grid improvement translates far more reliably than measurables like driver age, nationality, team, pit strategy, circuit characteristics, or round. Holding all other variables constant, a one-place gain on the grid is worth far more in podium probability than any other measurable we tested.

Using the F1 Data

5,000+ unique race outcome records from 2012-2024 from one of the leading sources in historical F1 data (Ergast).

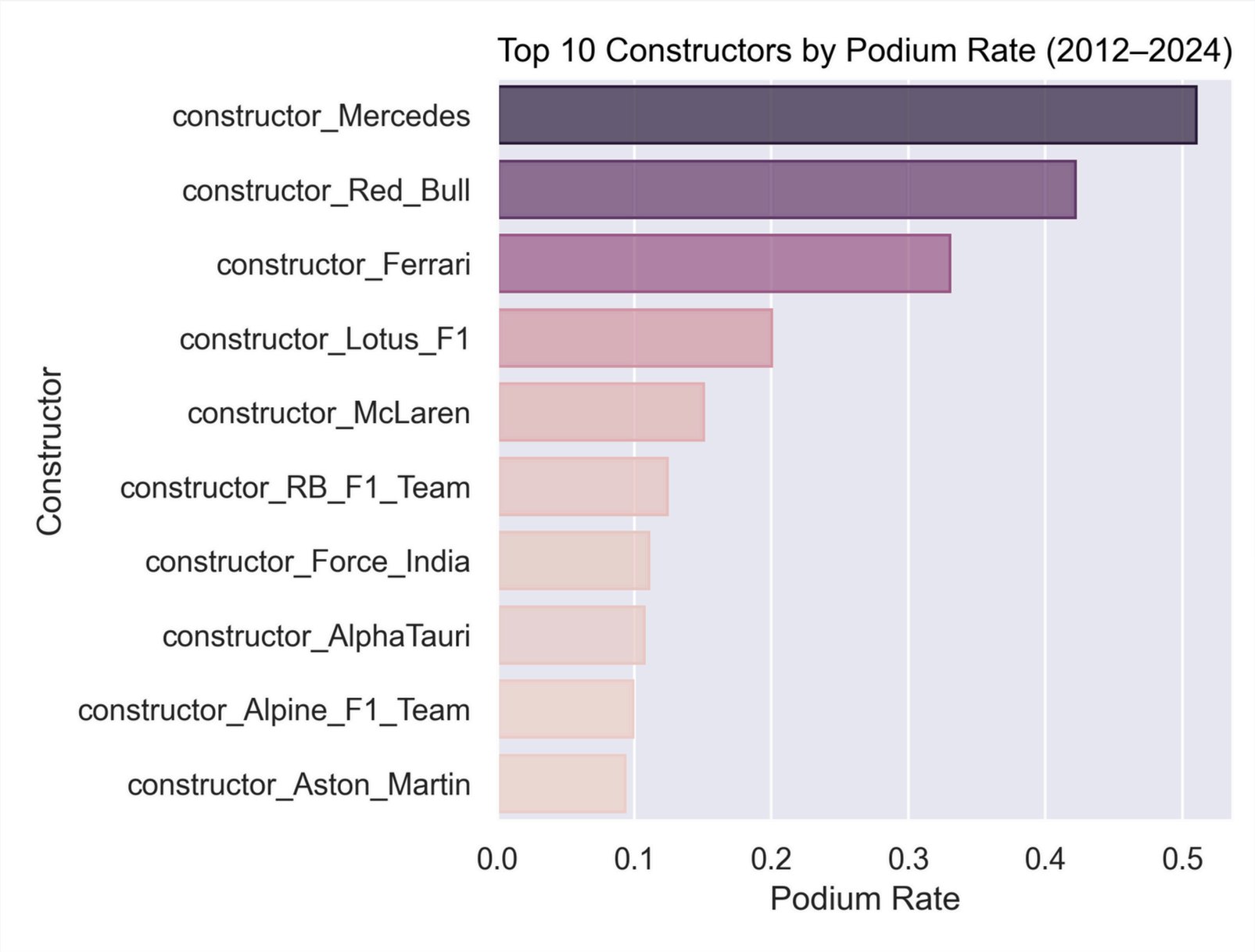
Each observation represents a single driver and race outcome.

- **podium**: Whether the race outcome was first, second, or third place (0 = no podium finish)
- **grid**: Starting position (lower = further up)
- **round**: Round number in the season
- **driver_age**: Driver age at time of race

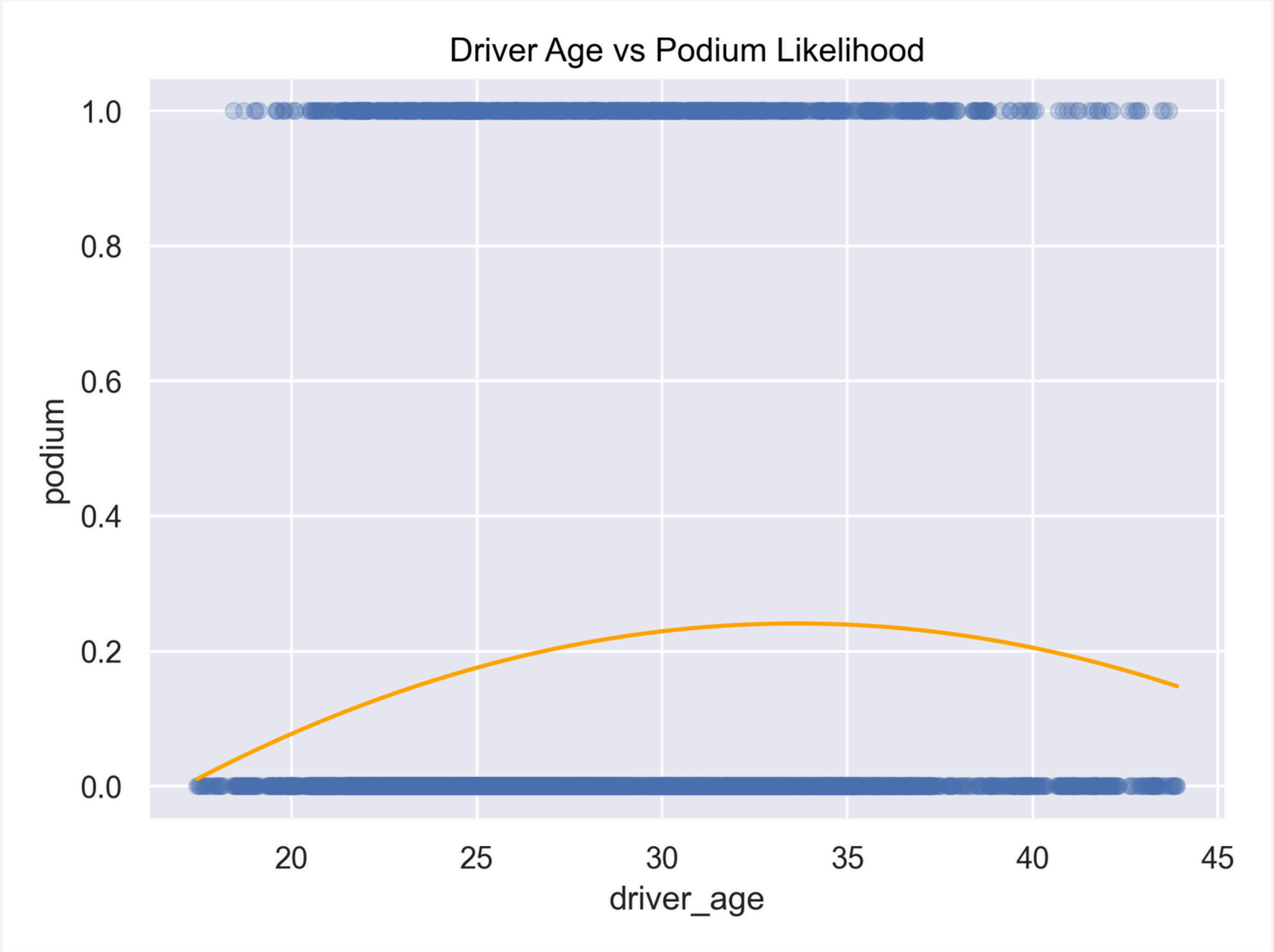
Using the F1 Data

- **constructor_***: The constructor/organizer of the racing team
- **lat**: Latitude of the circuit
- **lng**: Longitude of the circuit
- **alt**: Altitude of the circuit (meters)
- **laps_per_pit**: Average number of laps between pit stops for the driver in that race
- **no_pit**: Whether the driver did not pit in that race (1 = no pit)
- **nationality_***: Driver nationality

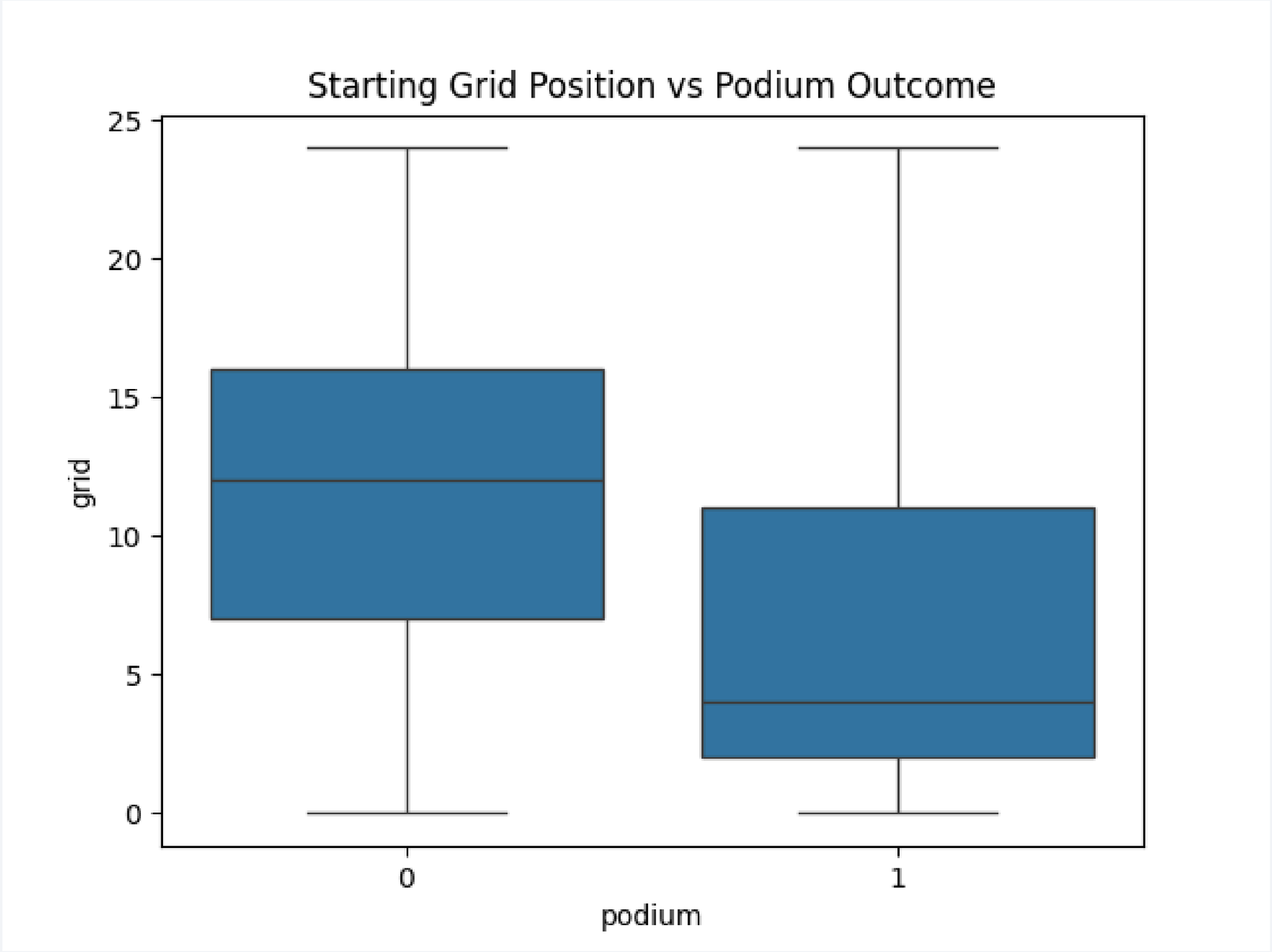
Constructor success rates vary



Driver age and outcomes



Grid and podium relationship

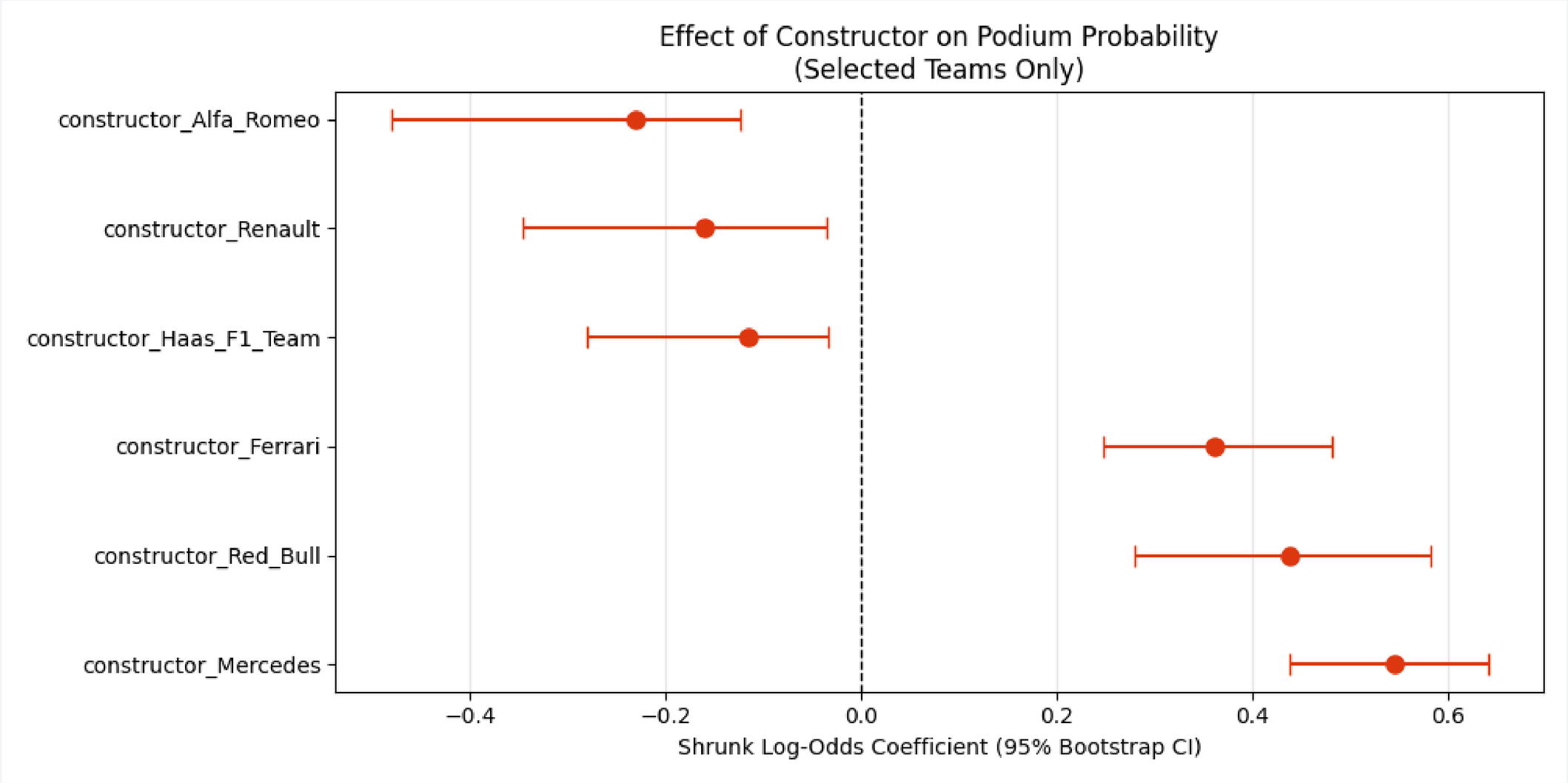


Modeling podium outcomes

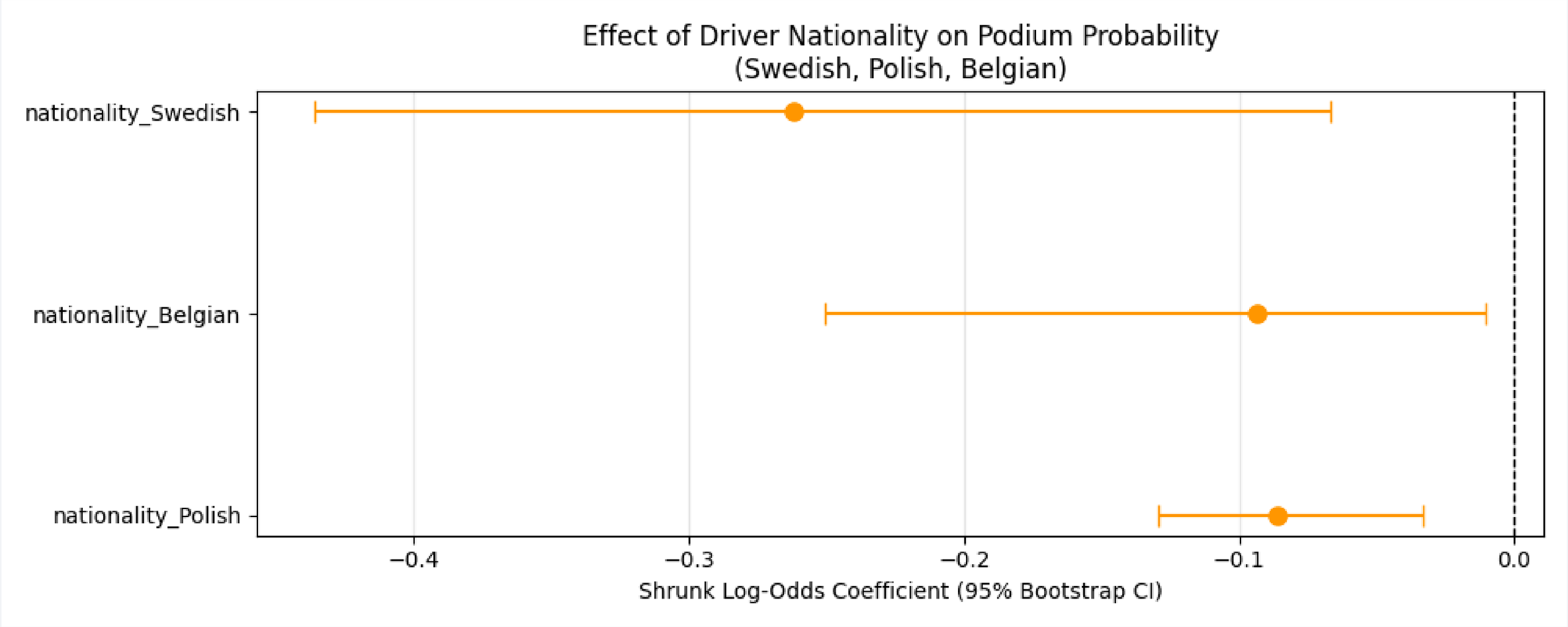
We fit a penalized **Elastic Net regression model** to the race data

- Excluding country but including lat/lng/alt as proxies
- Adding pit strategy data like laps_per_pit without data leakage through total laps
- Including an additional variable for the relationship between latitude and drivers with British nationalities
- Confirming model assumptions were mostly satisifed

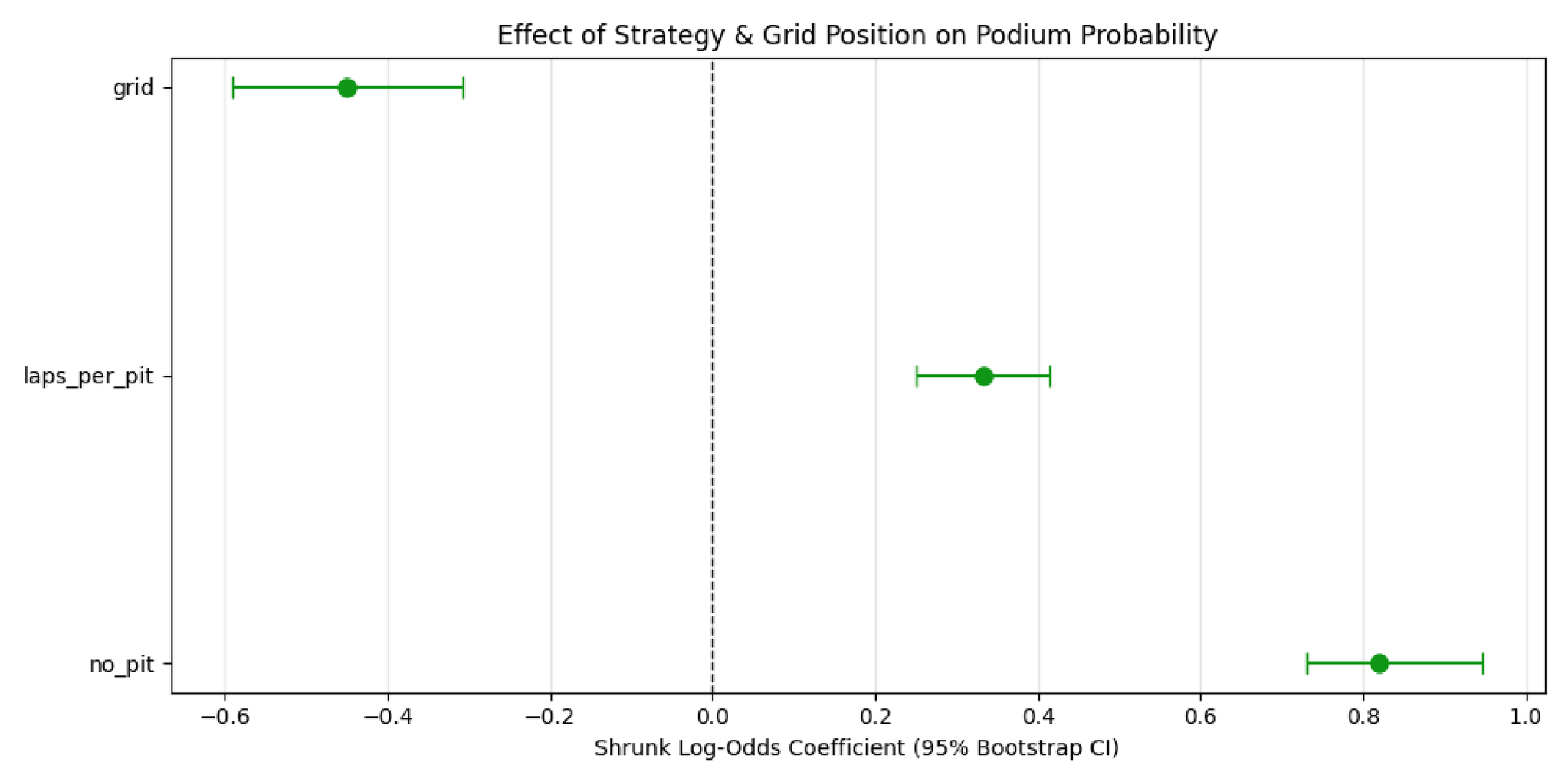
Some constructors consistently perform



Drivers that underperform



Pit strategy and grid really matter



Modeling 2025 Grand Prix outcomes

Our best model perfectly predicted all three podium finishers for one of the races (4.8% of the time).

- The model predicted at least one podium finisher correctly **76.2%** of the time, and at least two 19% of the time
- Simply predicting that grid positions 1-3 will podium is perfectly correct 28.6% of the time, and at least one of them did finish on the podium **100%** of the time

Recommendation

F1 constructors should invest the majority of their performance budget in single-lap qualifying pace and maximizing grid position gains.

- Pit-strategy and circuit characteristics have near-zero effects once grid is accounted for
- Baseline of grid 1-3 outpaces our model’s predictive fit

Next Steps

With richer data (like real-time tire degradation, detailed weather, drag reduction) we might find small additional edges. We would also love to explore more ways to introduce bias toward grid placement in our model.

But, the public data we have today says grid is still what matters most.


Questions?

Thank you!

Appendix

About the Dataset

The Ergast API data describes Formula 1 World Championship races from 1950-2024. It contains data on racers, drivers, constructors, qualifying, circuits, lap times, pit stops, and championships. Unfortunately, a lot of the useful data like lap times or laps completed would have resulted in data leakage.

Data: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020

`races.csv`

- **raceId**: Unique identifier for each race
- **year**: Year of the race
- **round**: Round number in the season
- **circuitId**: Reference to `circuits.csv`
- **name**: Race name
- **date**: Race date
- **time**: Race start time (nullable)
- **url**: Link to race information

`drivers.csv`

- **driverId**: Unique identifier
- **driverRef**: Reference name
- **number**: Racing number (nullable)
- **code**: 3-letter code
- **name**: Full driver name
- **dob**: Date of birth
- **nationality**: Driver nationality
- **url**: Link to driver information

`constructors.csv`

- **constructorId**: Unique identifier
- **constructorRef**: Reference name
- **name**: Full constructor name
- **nationality**: Constructor nationality
- **url**: Link to constructor information

`results.csv`

- **resultId**: Unique identifier for this result
- **raceId**: Reference to `races.csv`
- **driverId**: Reference to `drivers.csv`
- **constructorId**: Reference to `constructors.csv`
- **number**: Racing number (nullable)
- **grid**: Starting position
- **position**: Finishing position (numeric)
- **positionText**: String version of position (e.g., 'R', 'DSQ')
- **positionOrder**: Numeric finishing position
- **points**: Points earned in race
- **laps**: Number of laps completed
- **time**: Race time (nullable)
- **milliseconds**: Race time in milliseconds
- **fastestLap**: Lap number of fastest lap
- **fastestLapTime**: Fastest lap time string
- **fastestLapSpeed**: Speed on fastest lap
- **statusId**: Result status (1 = finished, other values = retired/DNF)

`qualifying.csv`

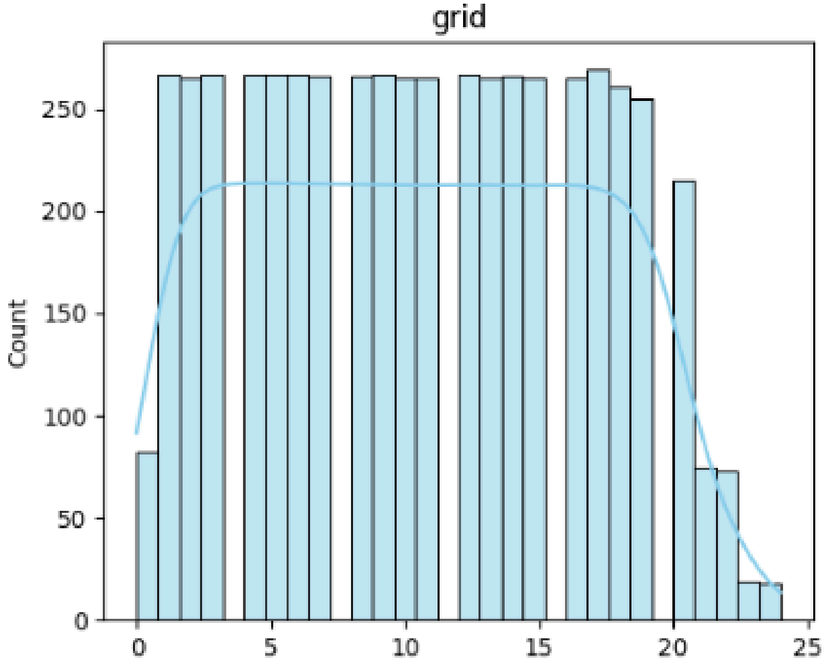
- **qualifyId**: Unique identifier
- **raceId**: Reference to `races.csv`
- **driverId**: Reference to `drivers.csv`
- **constructorId**: Reference to `constructors.csv`
- **number**: Racing number
- **position**: Qualifying position
- **q1**: Q1 lap time
- **q2**: Q2 lap time
- **q3**: Q3 lap time

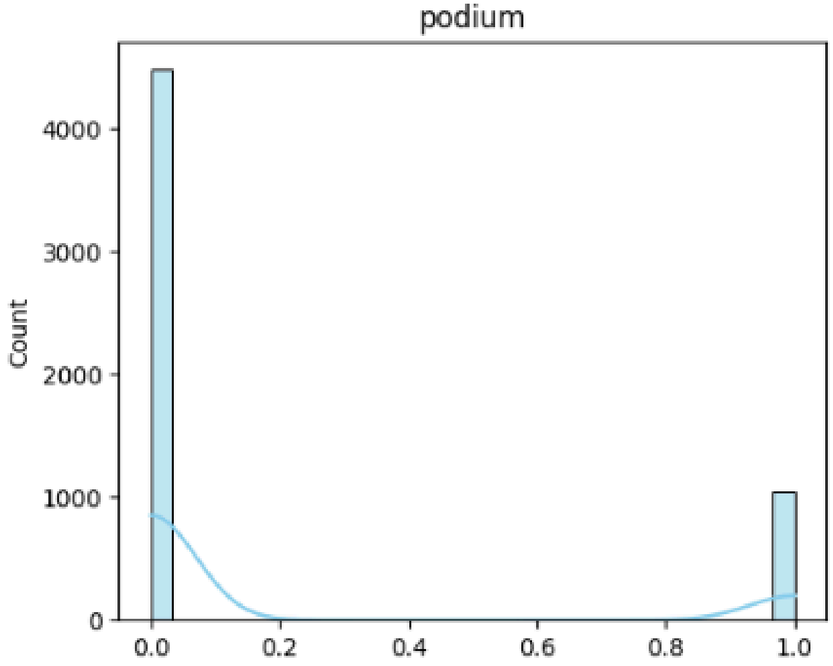
`circuits.csv`

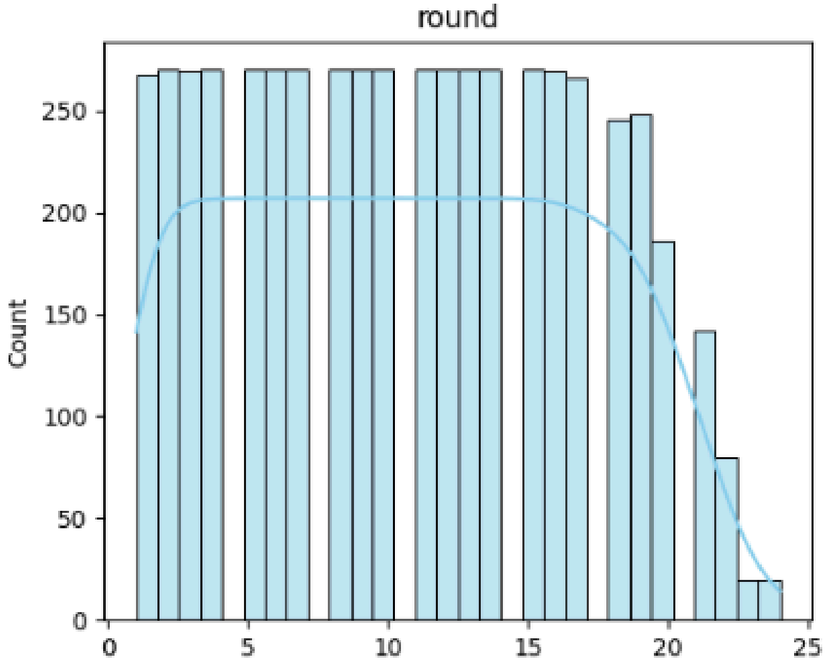
- **circuitId**: Unique identifier
- **circuitRef**: Reference name
- **name**: Circuit name
- **location**: City or location
- **country**: Country
- **lat**: Latitude
- **lng**: Longitude
- **alt**: Altitude
- **url**: Link to circuit information

`pit_stops.csv`

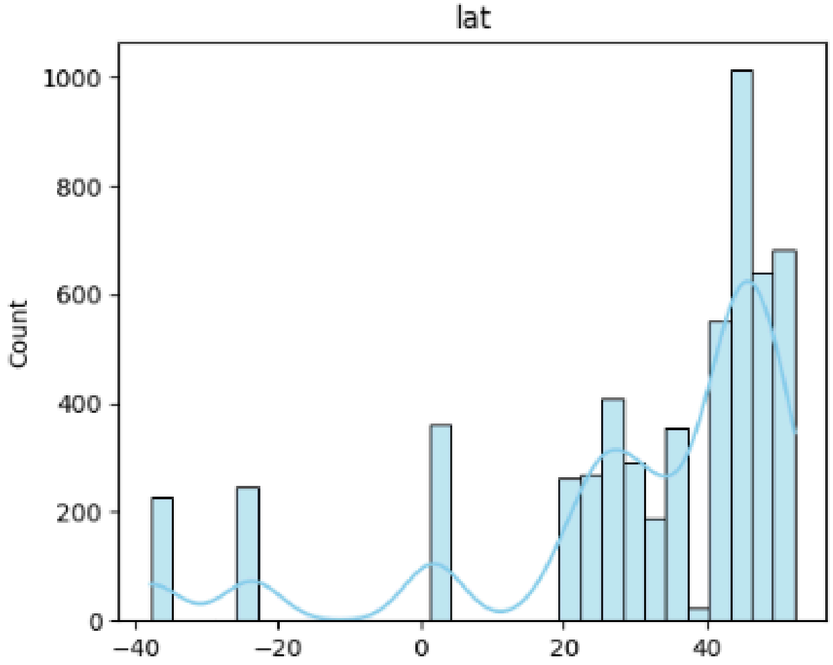
- **raceId**: Reference to `races.csv`
- **driverId**: Reference to `drivers.csv`
- **stop**: Pit stop number for the driver
- **lap**: Lap when the pit stop occurred
- **time**: Pit stop time string
- **duration**: Duration in seconds
- **milliseconds**: Duration in milliseconds

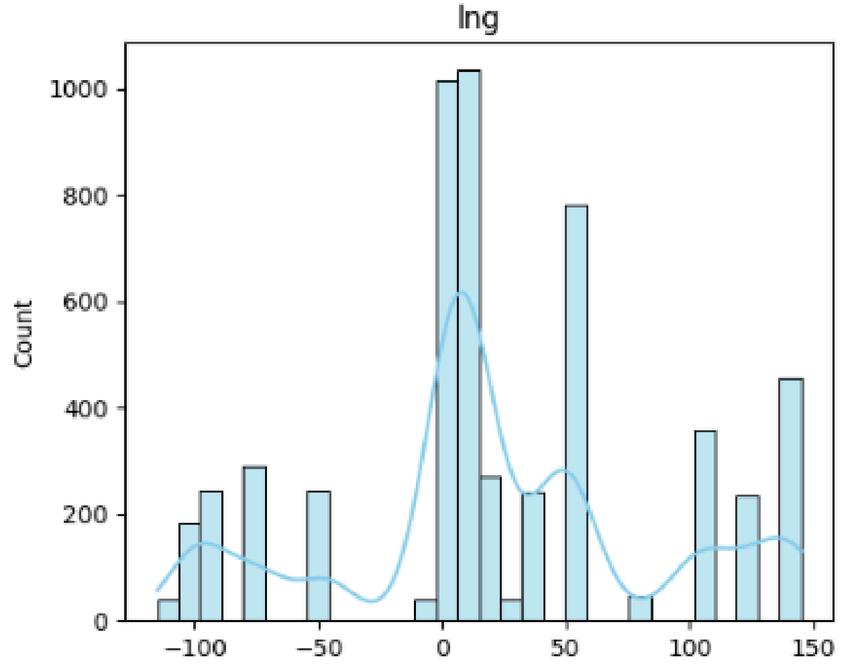

grid
Count
0
50
100
150
200
250
0
5
10
15
20
25

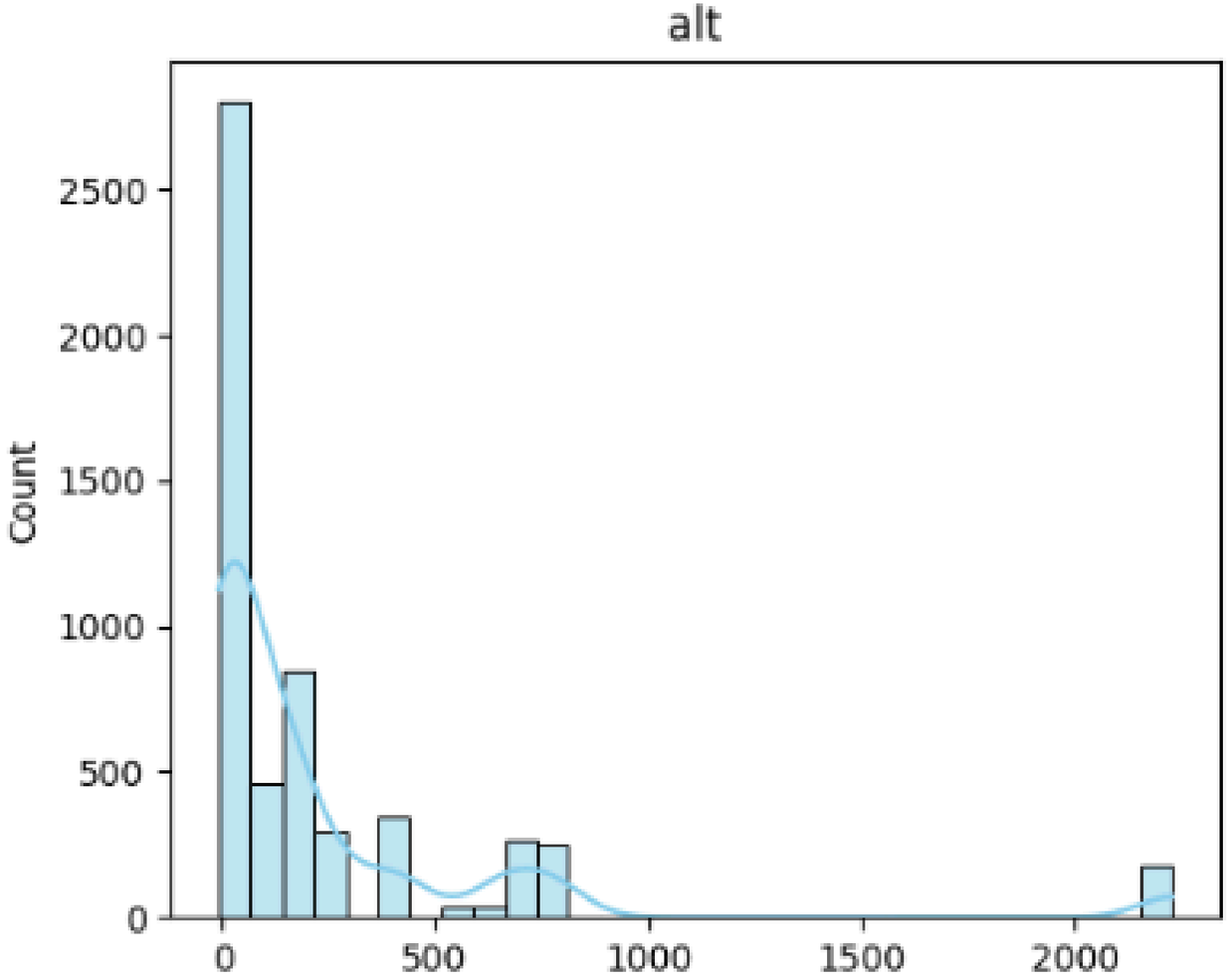

podium
Count
0
1000
2000
3000
4000
0.0
0.2
0.4
0.6
0.8
1.0

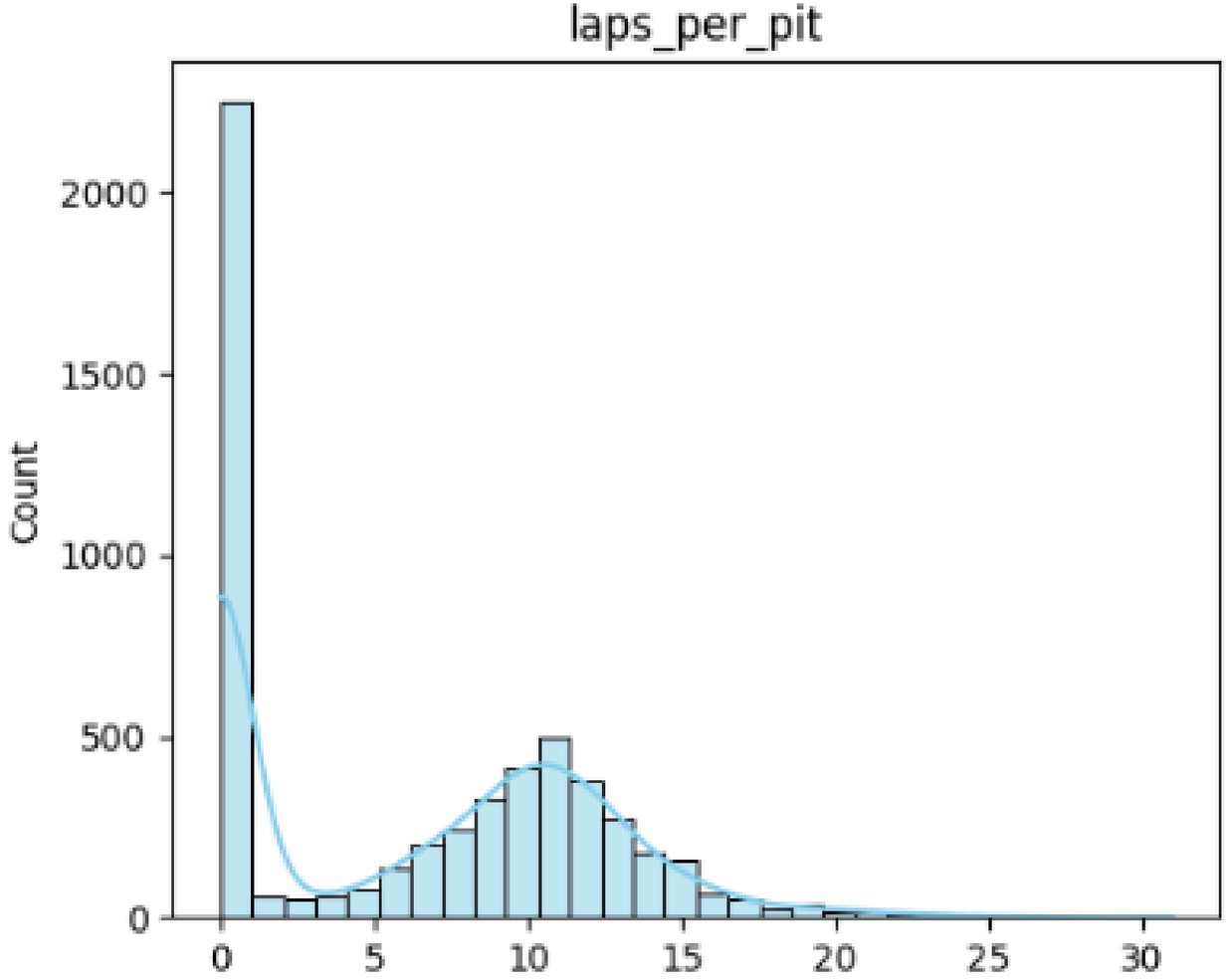

round
Count
0
50
100
150
200
250
0
5
10
15
20
25

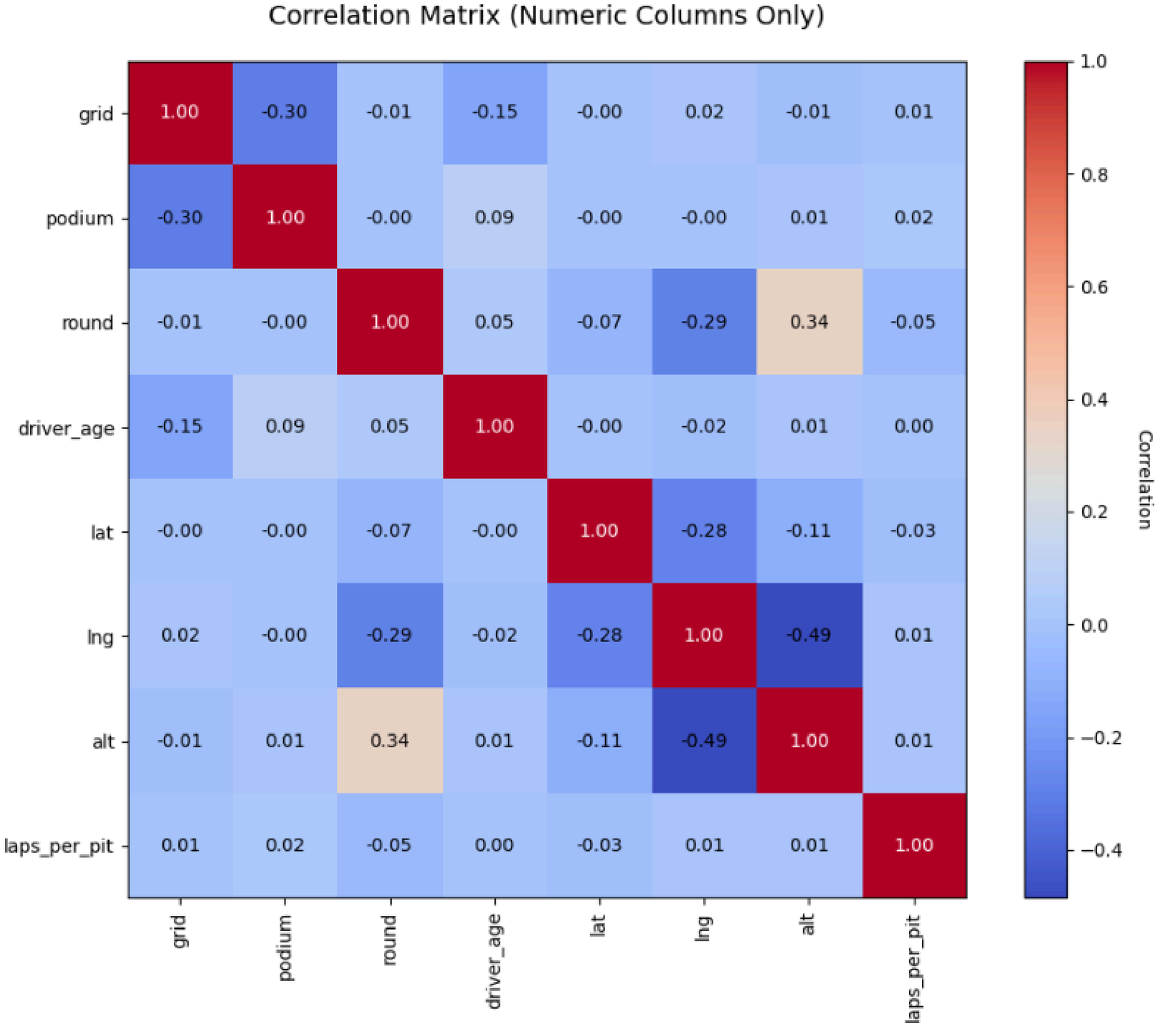

driver_age
Count
0
100
200
300
400
20
25
30
35
40
45

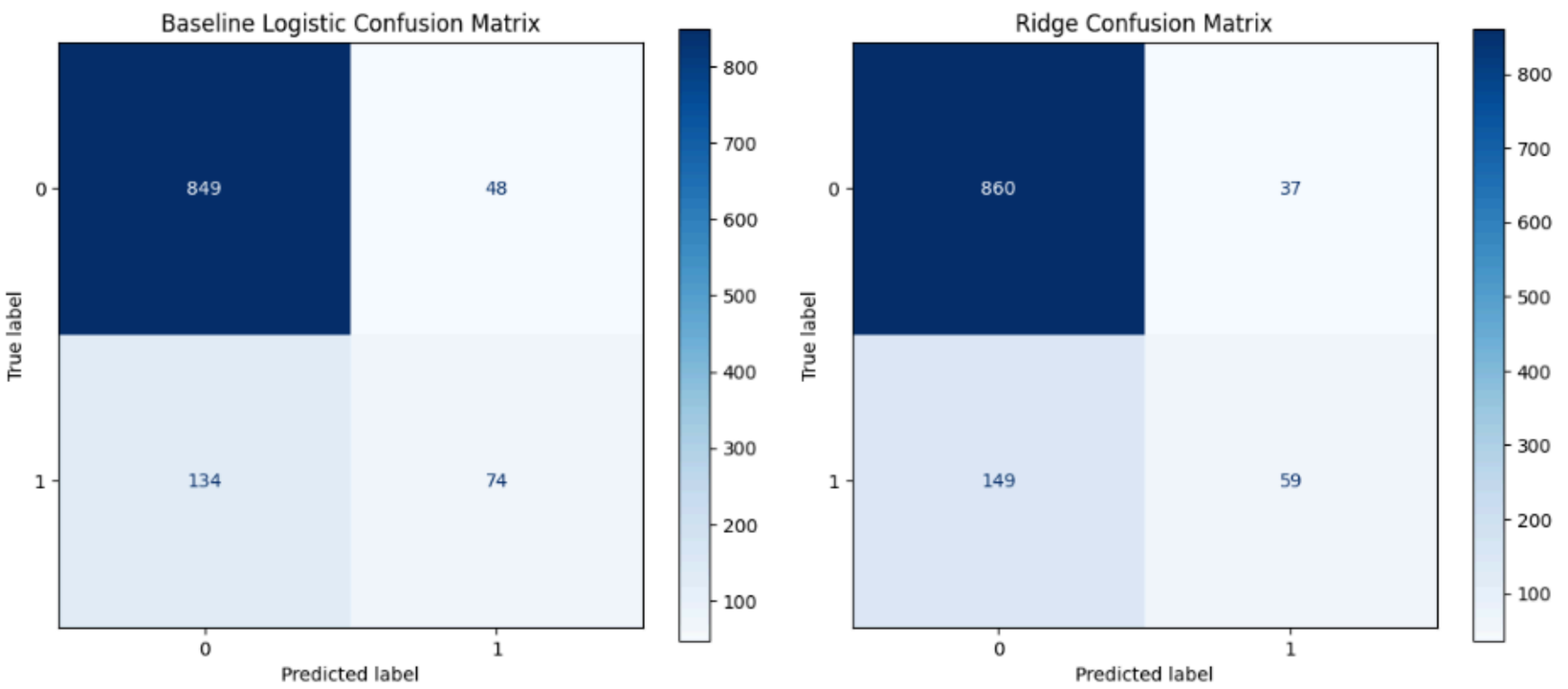

lat
Count
0
200
400
600
800
1000
−40
−20
0
20
40


lng
Count
0
200
400
600
800
1000
−100
−50
0
50
100
150

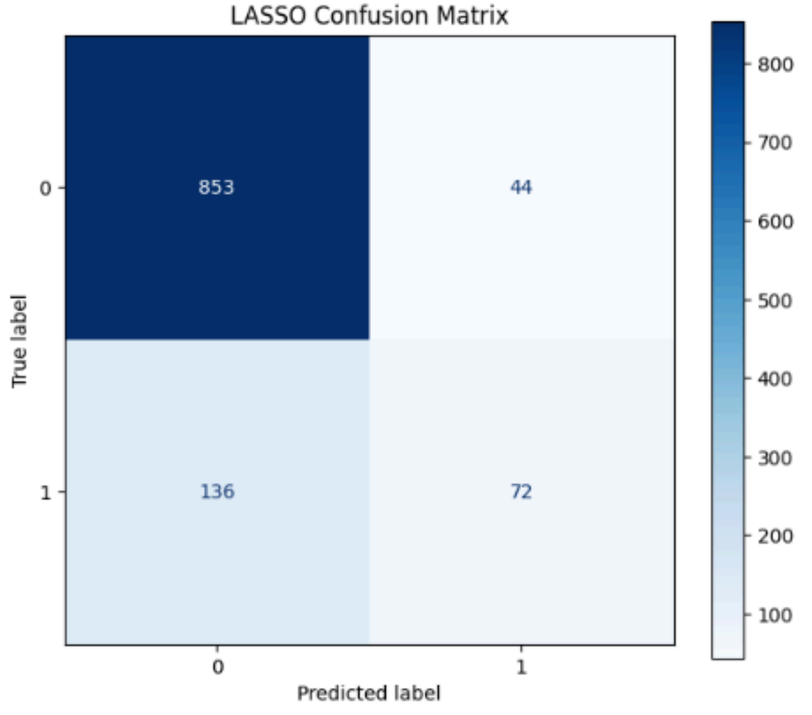

alt
Count
2500
2000
1500
1000
500
0
0
500
1000
1500
2000

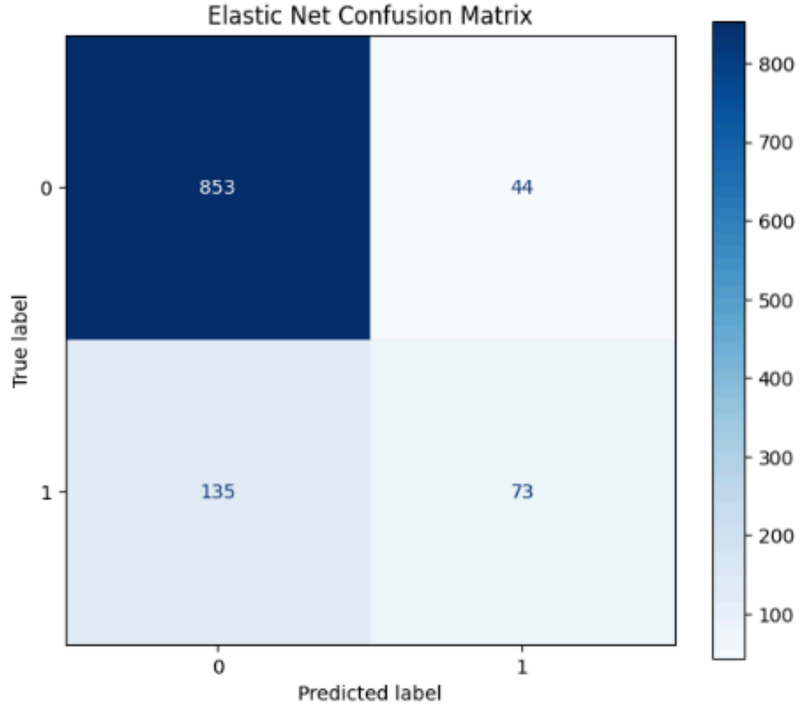

laps_per_pit
Count
2000
1500
1000
500
0
0
5
10
15
20
25
30

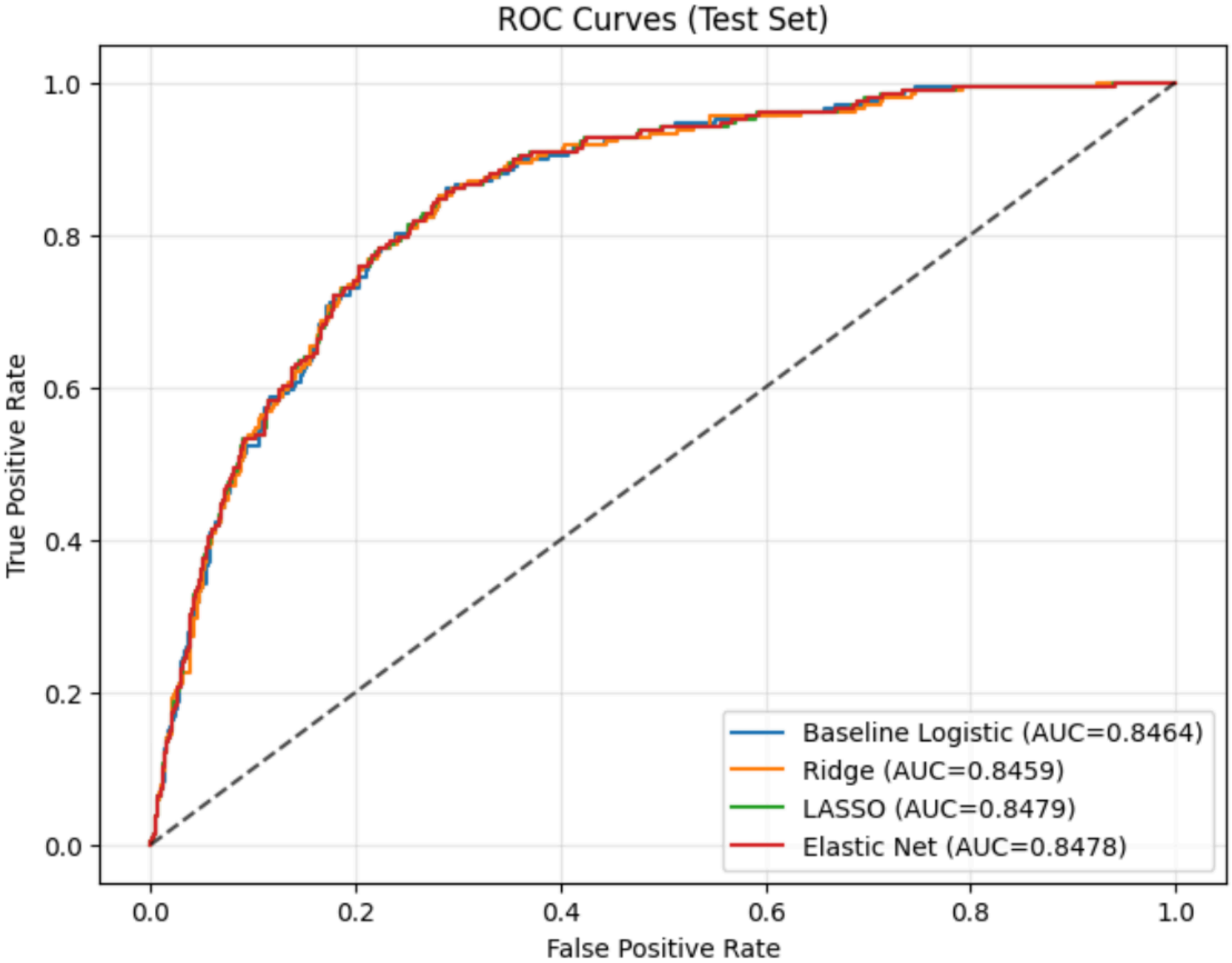

Correlation Matrix (Numeric Columns Only)
grid
podium
round
driver_age
lat
lng
alt
laps_per_pit
1.00 -0.30 -0.01 -0.15 -0.00 0.02 -0.01 0.01
-0.30 1.00 -0.00 0.09 -0.00 -0.00 0.01 0.02
-0.01 -0.00 1.00 0.05 -0.07 -0.29 0.34 -0.05
-0.15 0.09 0.05 1.00 -0.00 -0.02 0.01 0.00
-0.00 -0.00 -0.07 -0.00 1.00 -0.28 -0.11 -0.03
0.02 -0.00 -0.29 -0.02 -0.28 1.00 -0.49 0.01
-0.01 0.01 0.34 0.01 -0.11 -0.49 1.00 0.01
0.01 0.02 -0.05 0.00 -0.03 0.01 0.01 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Correlation

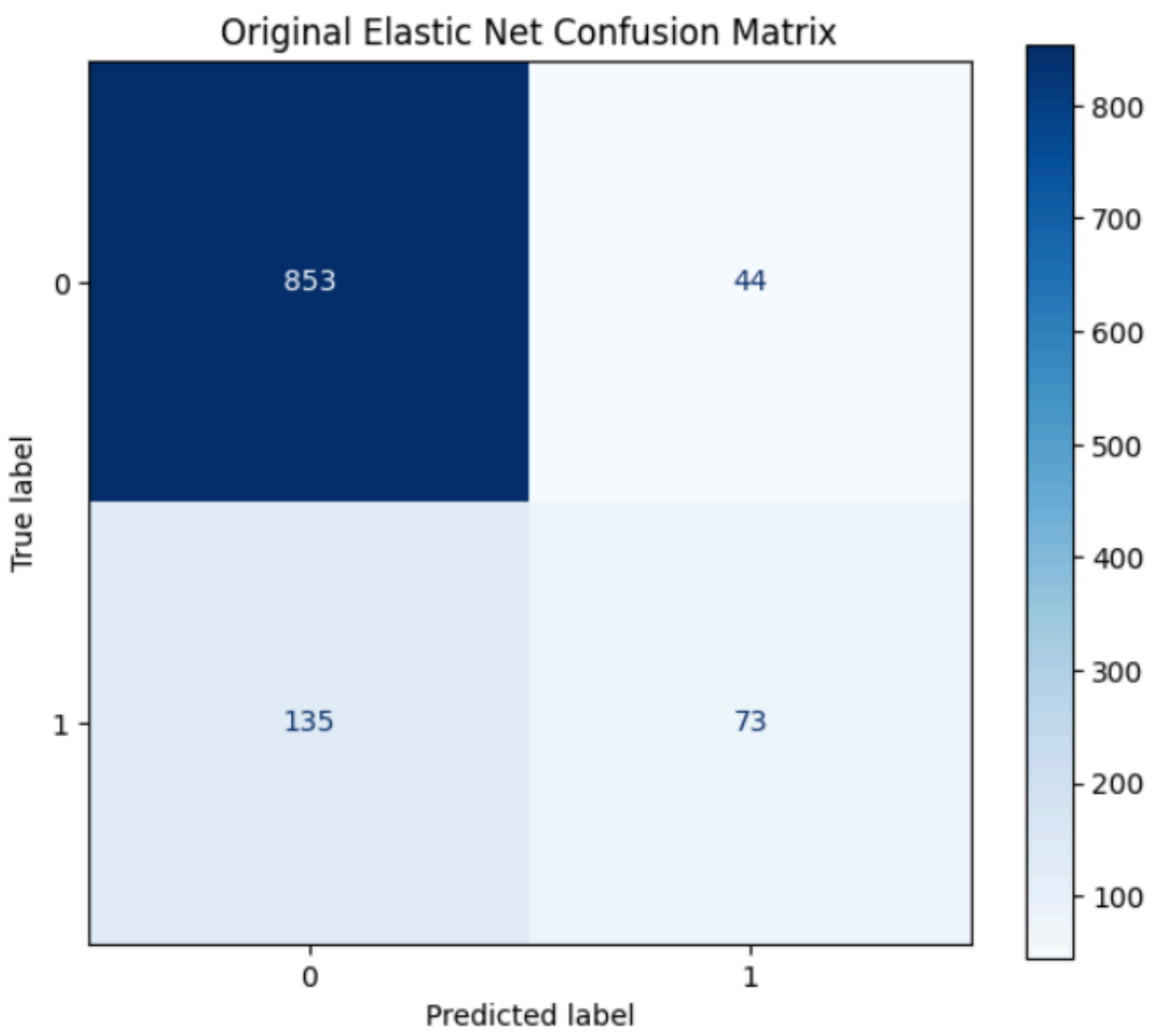

Baseline Logistic Confusion Matrix
849
48
134
74
True label
Predicted label
0
1
800
700
600
500
400
300
200
100

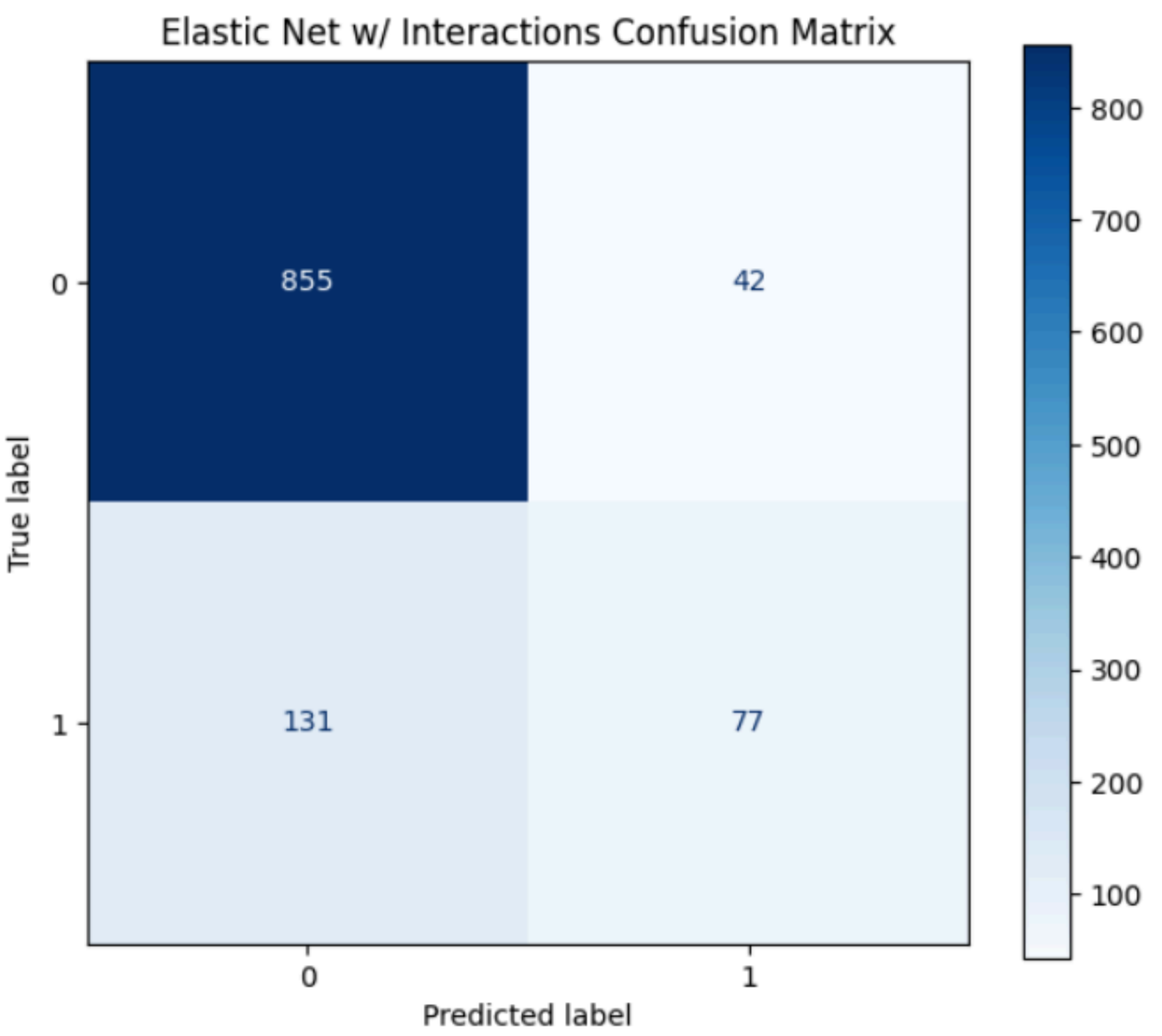
Ridge Confusion Matrix
860
37
149
59
True label
Predicted label
0
1
800
700
600
500
400
300
200
100

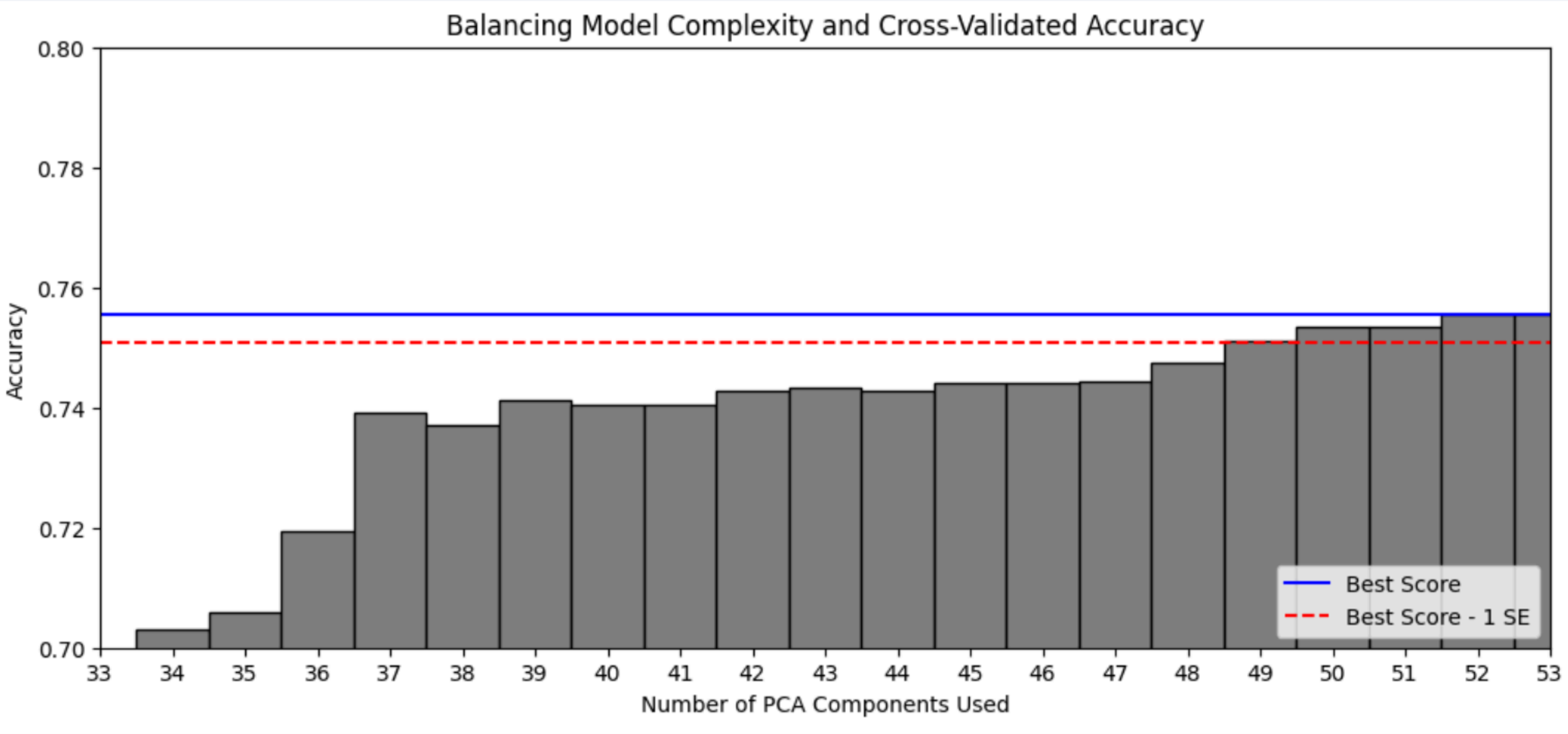

LASSO Confusion Matrix
853
44
136
72
True label
Predicted label
0
1
800
700
600
500
400
300
200
100


Elastic Net Confusion Matrix
853
44
135
73
True label
Predicted label
0
1
800
700
600
500
400
300
200
100

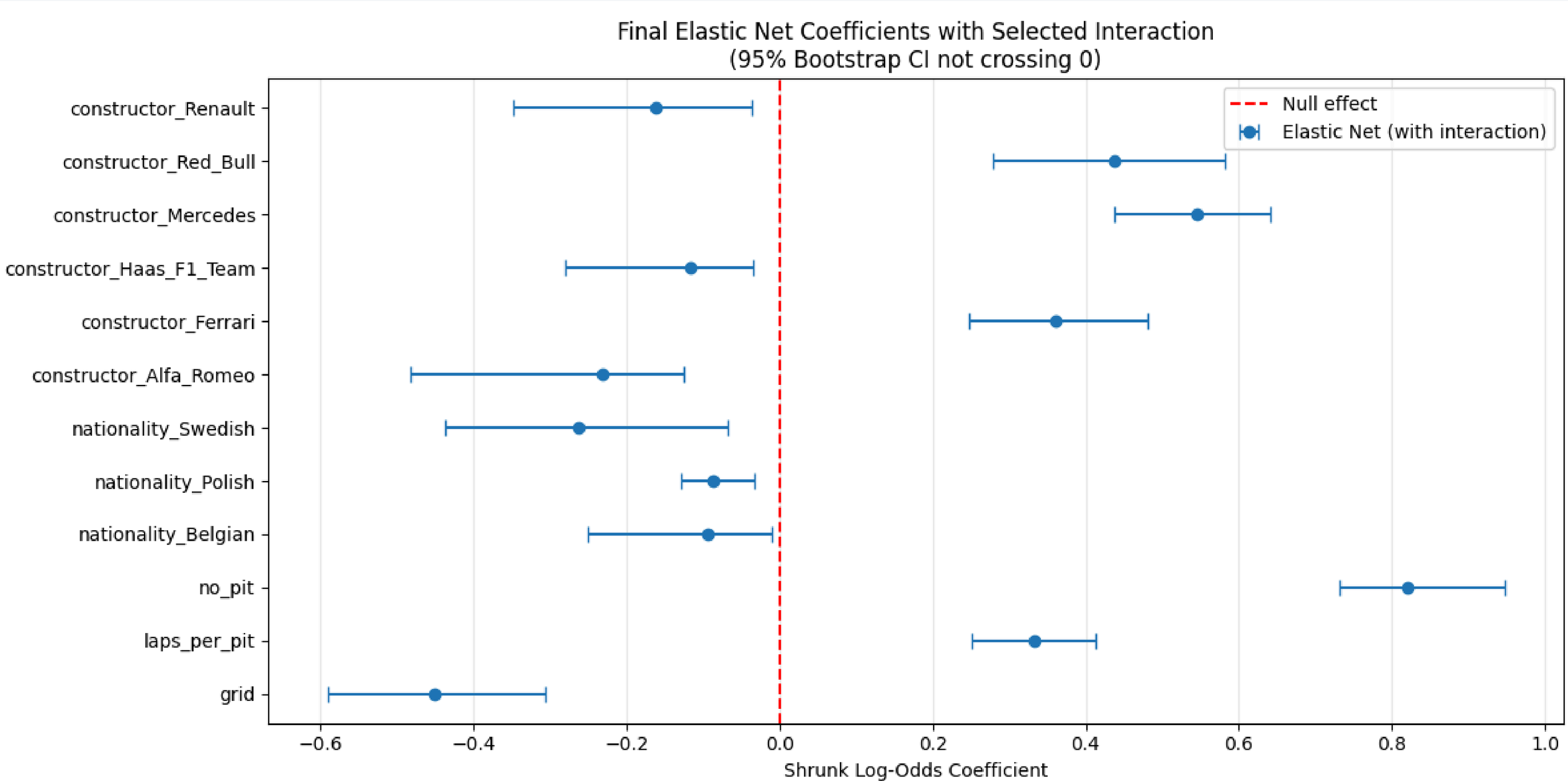

ROC Curves (Test Set)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Baseline Logistic (AUC=0.8464)
Ridge (AUC=0.8459)
LASSO (AUC=0.8479)
Elastic Net (AUC=0.8478)


Original Elastic Net Confusion Matrix
853
44
135
73
0
1
True label
0
1
Predicted label
800
700
600
500
400
300
200
100


Elastic Net w/ Interactions Confusion Matrix
855
42
131
77
0
1
True label
0
1
Predicted label
800
700
600
500
400
300
200
100


Balancing Model Complexity and Cross-Validated Accuracy
Accuracy
0.80
0.78
0.76
0.74
0.72
0.70
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
Number of PCA Components Used
Best Score
Best Score - 1 SE

predictors	vif
str	f64
grid	7.288288
round	4.948282
driver_age	44.337784
lat	3.367466
lng	1.776223
alt	1.918891
laps_per_pit	2.354971
no_pit	1.190781
nationality_Argentinian_	1.108827
nationality_Australian	6.367534
nationality_Belgian	1.701795
nationality_Brazilian	3.69868
nationality_British	12.148727
nationality_Canadian	2.971298
nationality_Chinese	2.114709
nationality_Danish	3.910597
nationality_Dutch	4.452569
nationality_Finnish	8.316766
nationality_French	9.851128
nationality_German	10.902765
nationality_Indian	2.291019
nationality_Indonesian	1.335948
nationality_Italian	2.286848
nationality_Japanese	3.163632
nationality_Mexican	5.960148
nationality_Monegasque	3.00206
nationality_New_Zealander	1.677814
nationality_Polish	1.404888
nationality_Russian	3.422855
nationality_Spanish	8.155351
nationality_Swedish	2.694039
nationality_Thai	2.359177
nationality_Venezuelan	2.189523
constructor_Alfa_Romeo	3.57553
constructor_AlphaTauri	2.631471
constructor_Alpine_F1_Team	2.461183
constructor_Caterham	2.026034
constructor_Ferrari	4.588594
constructor_Force_India	2.976788
constructor_HRT	2.40661
constructor_Haas_F1_Team	4.167377
constructor_Lotus_F1	2.264471
constructor_Manor_Marussia	1.728656
constructor_Marussia	1.962483
constructor_McLaren	4.807804
constructor_Mercedes	4.821305
constructor_RB_F1_Team	1.556553
constructor_Racing_Point	1.444232
constructor_Red_Bull	5.258543
constructor_Renault	2.373368
constructor_Sauber	3.693248
constructor_Toro_Rosso	3.550412
constructor_Williams	3.873961
residuals	1.032036


Final Elastic Net Coefficients with Selected Interaction
(95% Bootstrap CI not crossing 0)
Null effect
Elastic Net (with interaction)
constructor_Renault
constructor_Red_Bull
constructor_Mercedes
constructor_Haas_F1_Team
constructor_Ferrari
constructor_Alfa_Romeo
nationality_Swedish
nationality_Polish
nationality_Belgian
no_pit
laps_per_pit
grid
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Shrunk Log-Odds Coefficient