

# 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



# F1 Background

1st	25 pts
2nd	18 pts
3rd	15 pts
4th	12 pts
5th	10 pts
6th	8 pts
7th	6 pts
8th	4 pts



# 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)
- On average, placing on the podium nets +7.33 points, relative to placing 4th

# Executive Summary

Our best-performing classification model accurately identified at least one podium finisher correctly 76.2% of the time. It outperformed purely random prediction by ~10 win equivalents (108.2 championship points).

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).

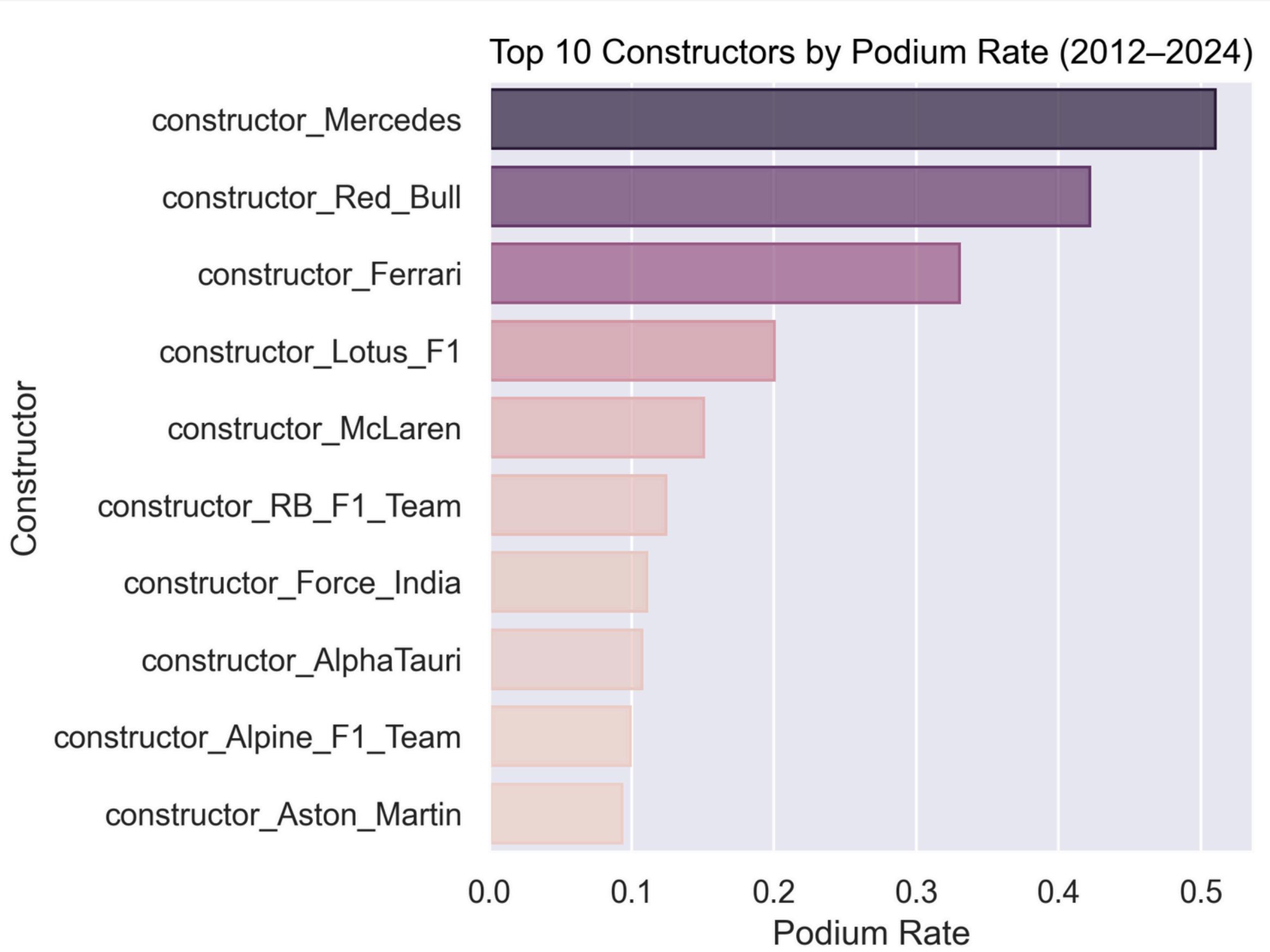
Each observation represents a single driver and race outcome.

- **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

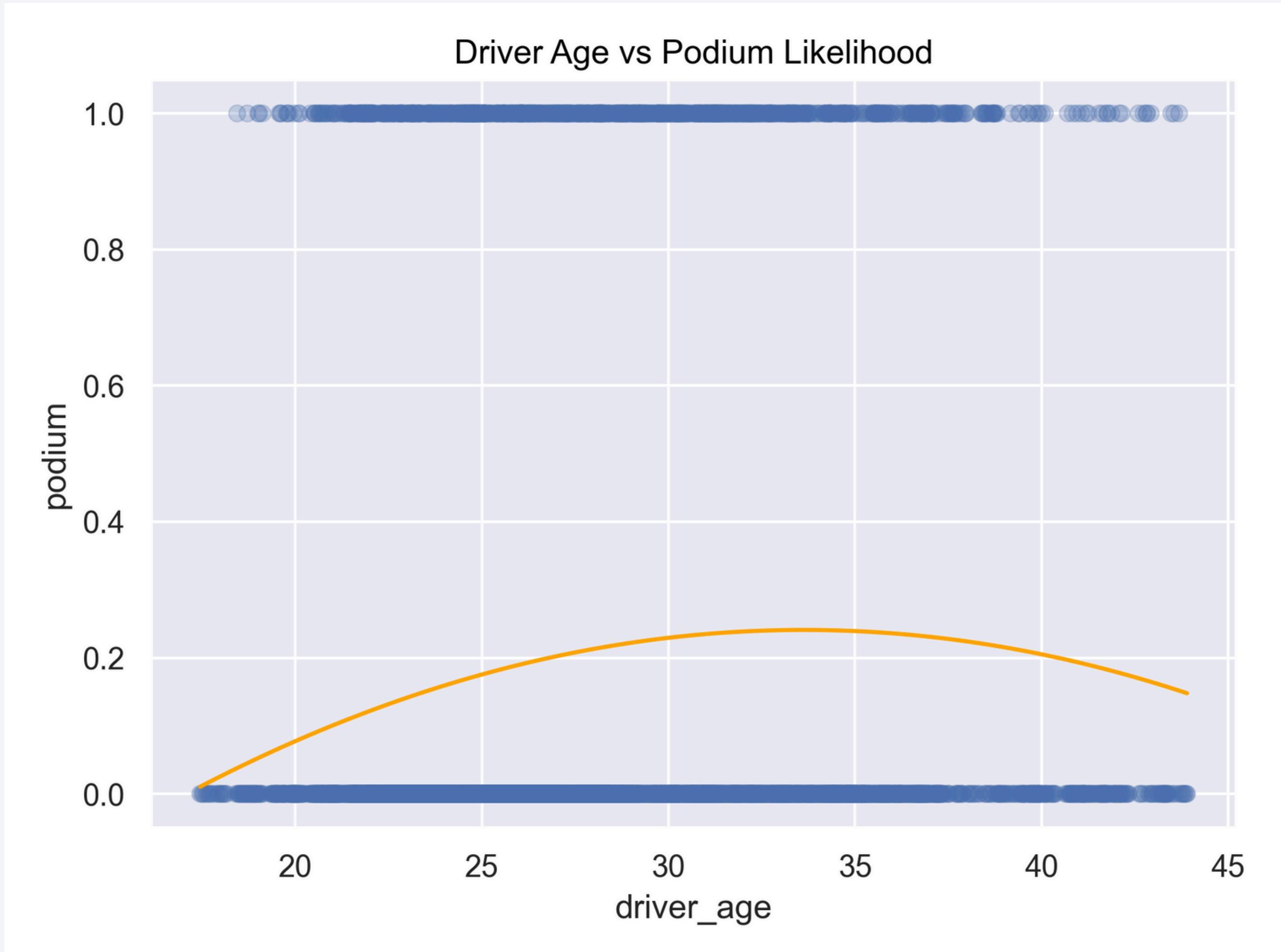
# Using the F1 Data

- **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

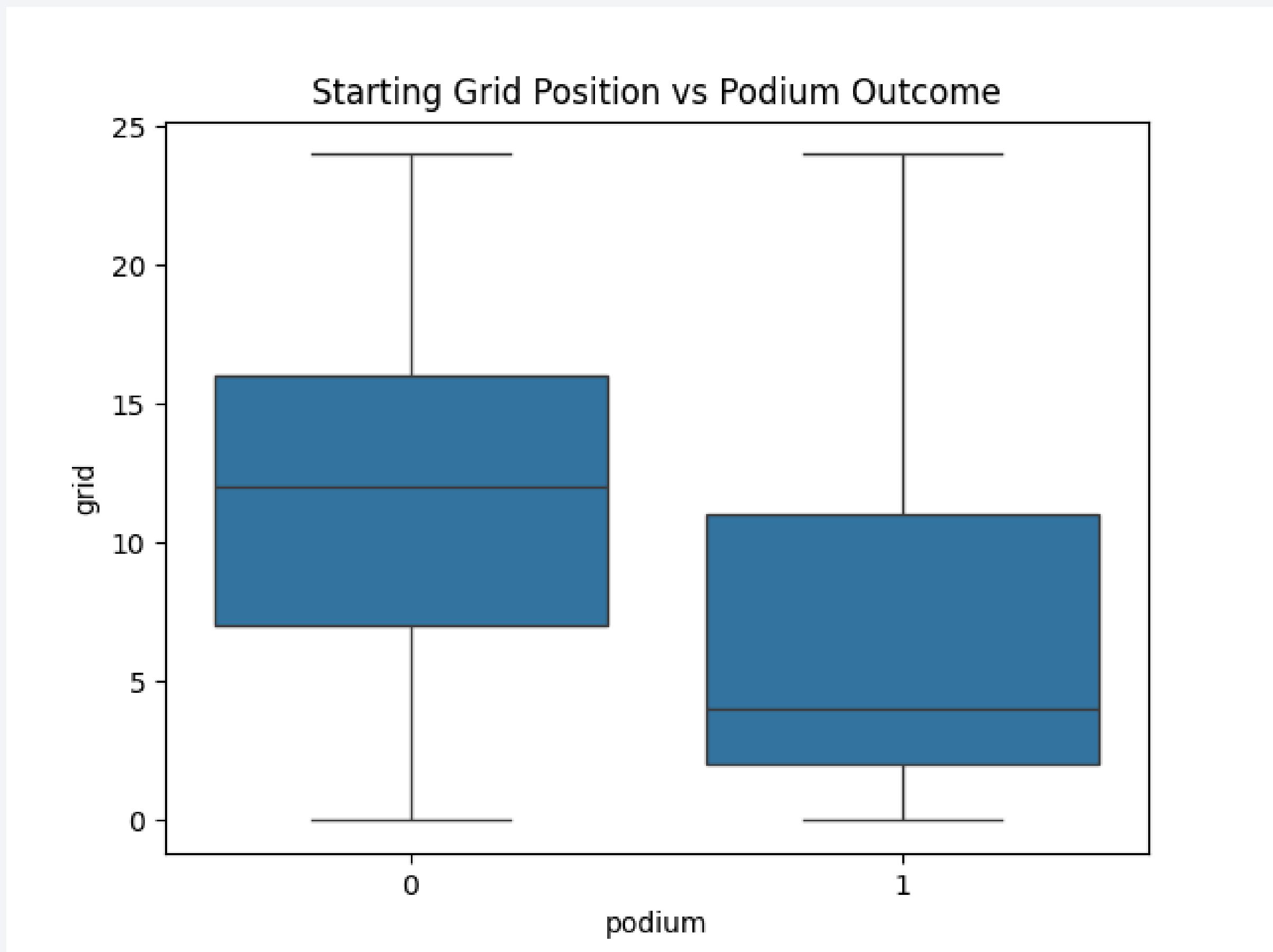
# Constructor success rates vary



# Driver age and outcomes



# Grid and podium relationship

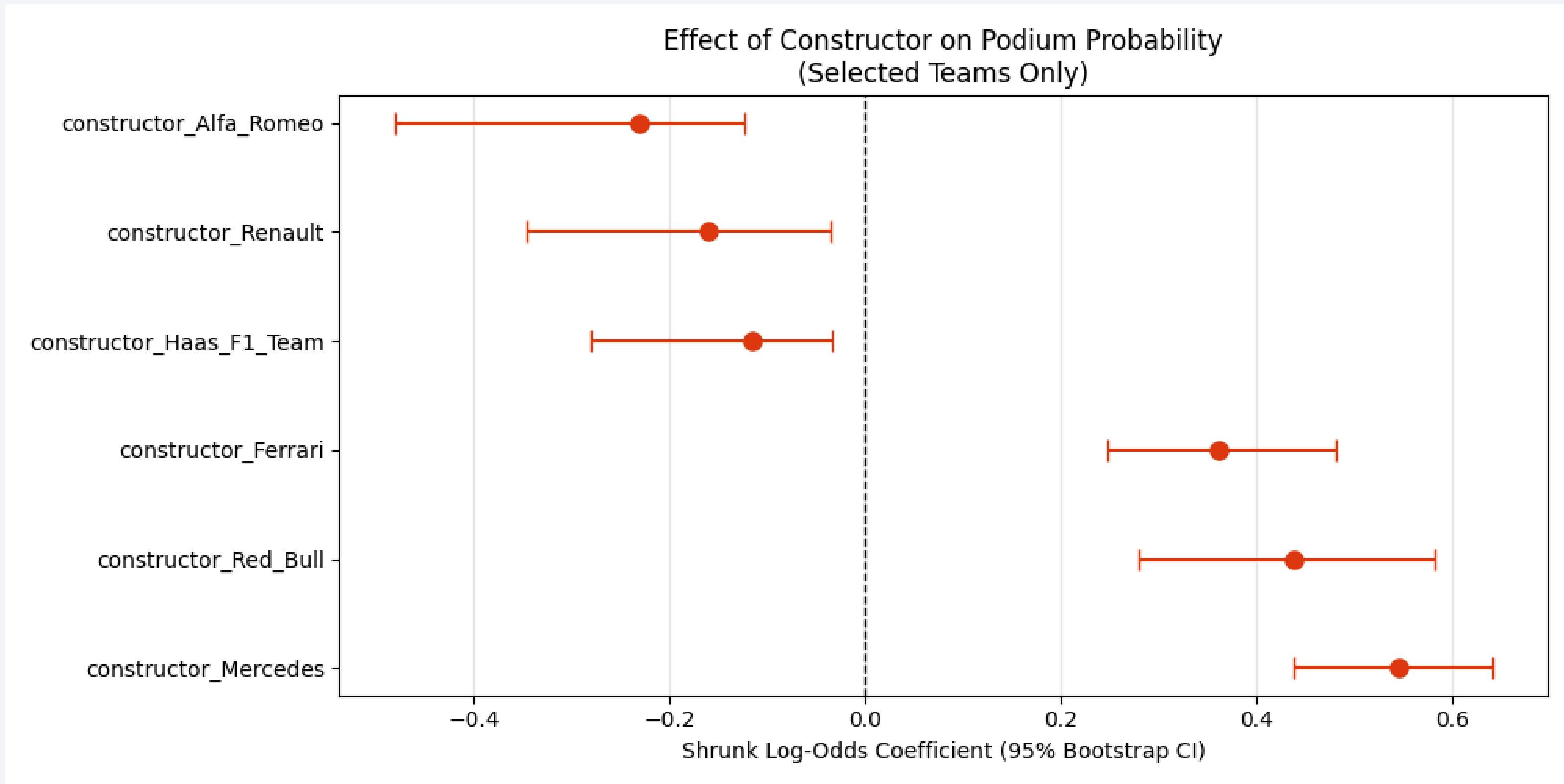


# Modeling podium outcomes

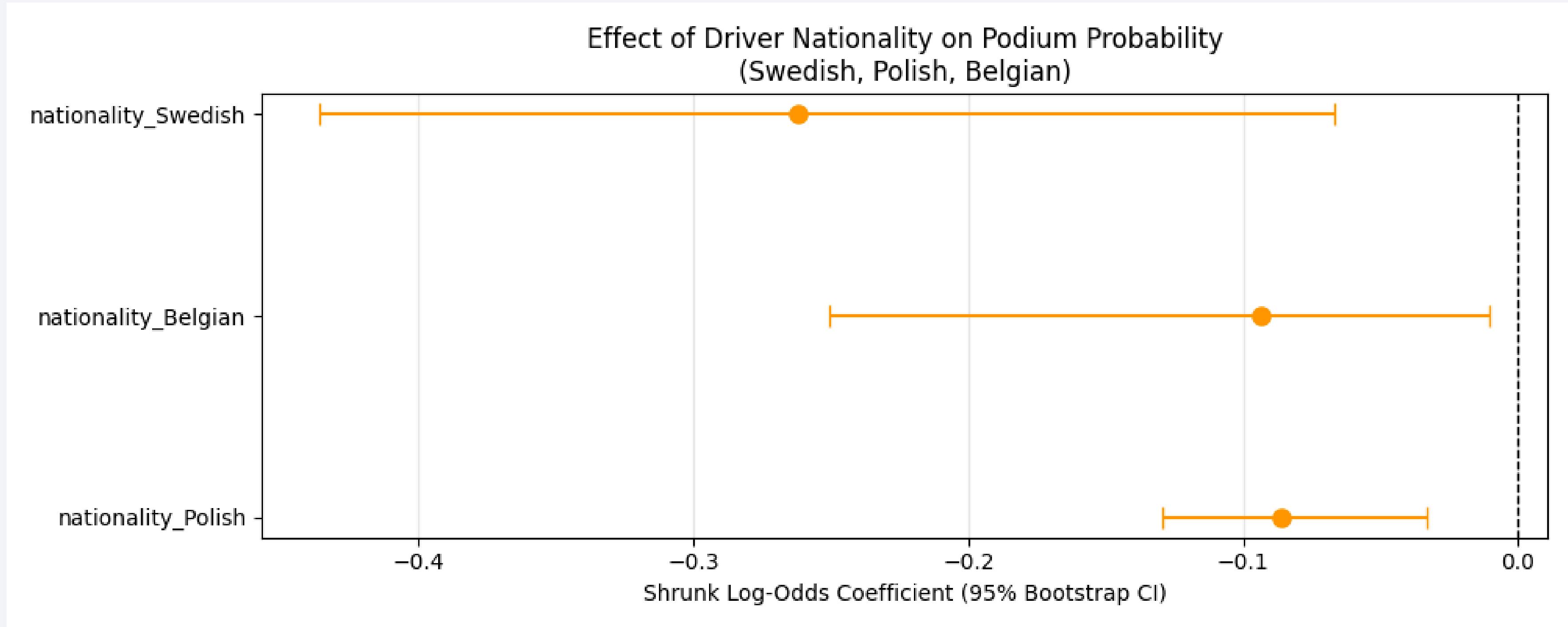
We fit a penalized **Elastic Net regression model** to the race data:

- 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 satisfied

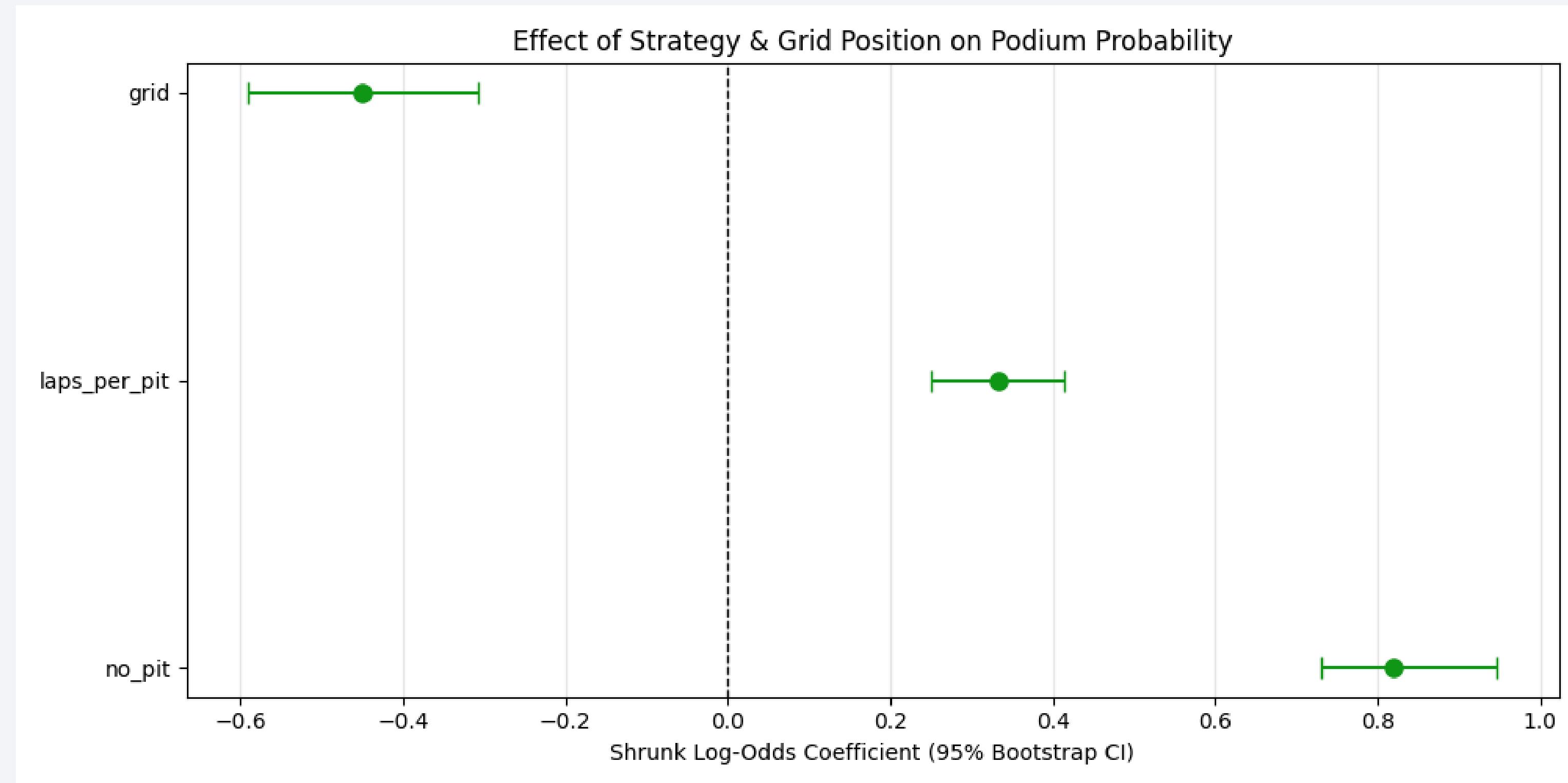
# Some constructors consistently perform



# 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. This strategy outperformed our model by 216.6 points or ~13 F1 wins.

# 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 top 3 grid positions 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

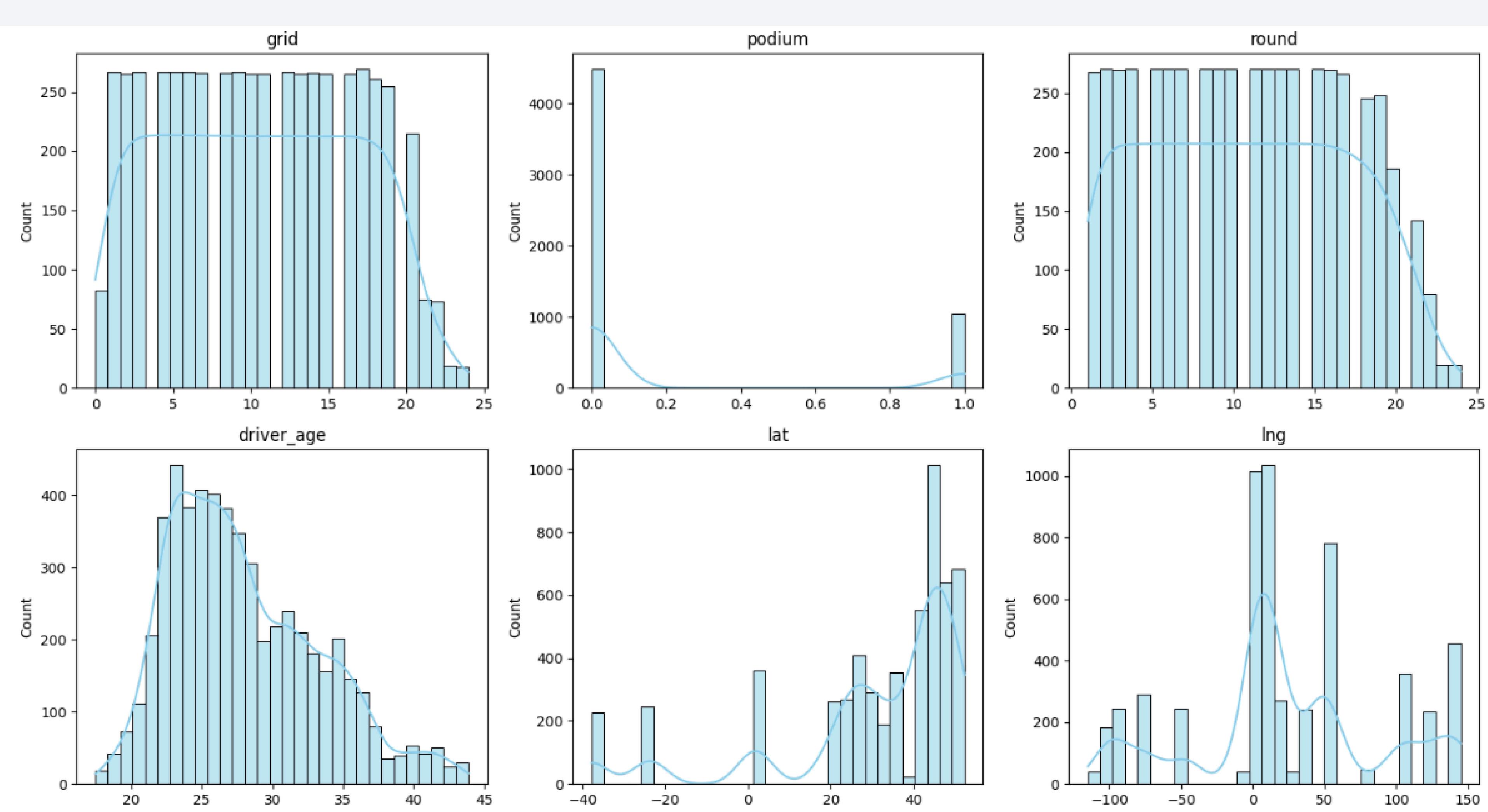
- **qualifyId**: Unique identifier
- **raceId**: Reference to races.csv
- **driverId**: Reference to drivers.csv
- **constructorId**: Reference to constructors.csv
- **number**: Racing number
- **position**: Qualifying position
- **q1**: Q1 lap time
- **q2**: Q2 lap time
- **q3**: Q3 lap time

### circuits.csv

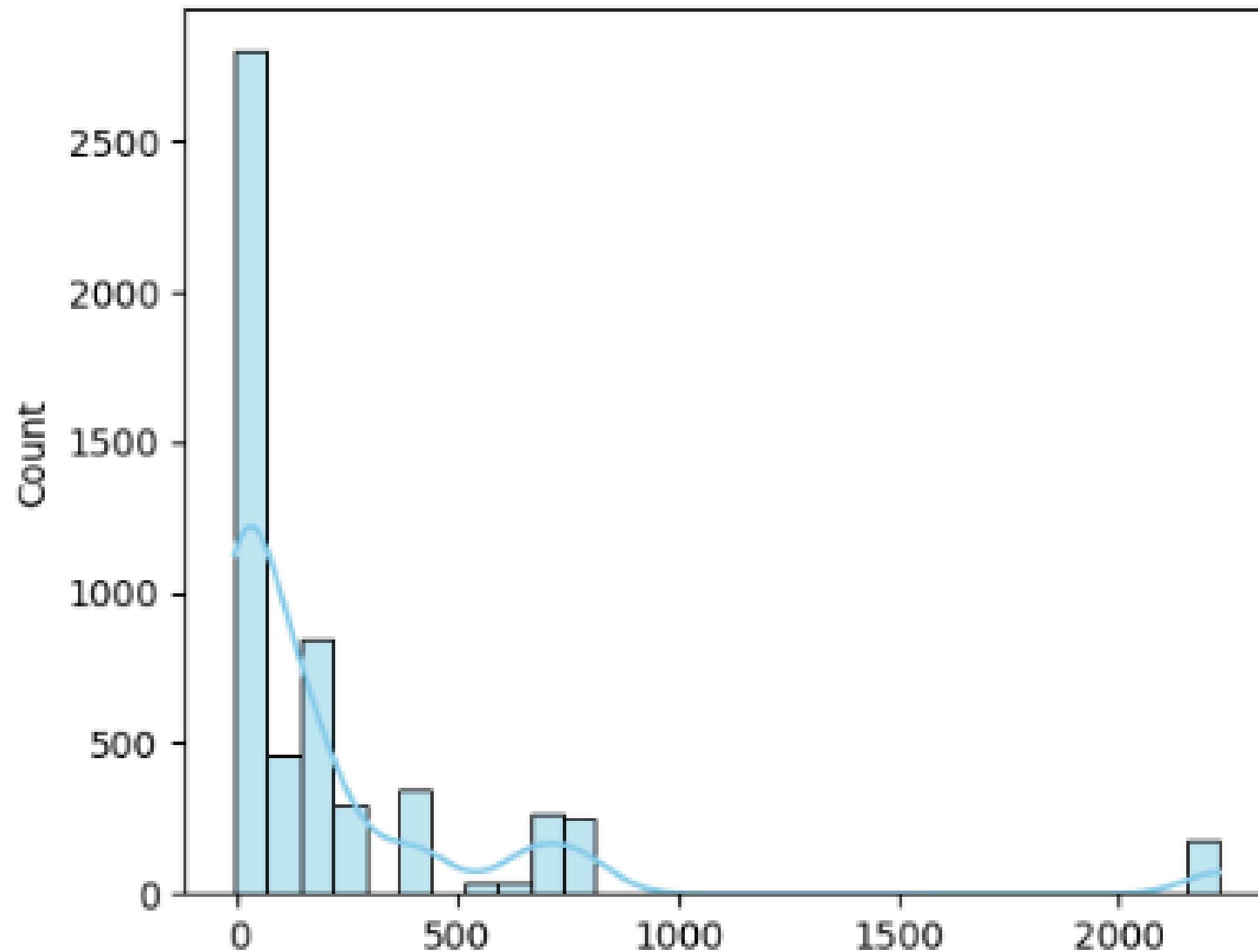
- **circuitId**: Unique identifier
- **circuitRef**: Reference name
- **name**: Circuit name
- **location**: City or location
- **country**: Country
- **lat**: Latitude
- **long**: Longitude
- **alt**: Altitude
- **url**: Link to circuit information

### pit\_stops.csv

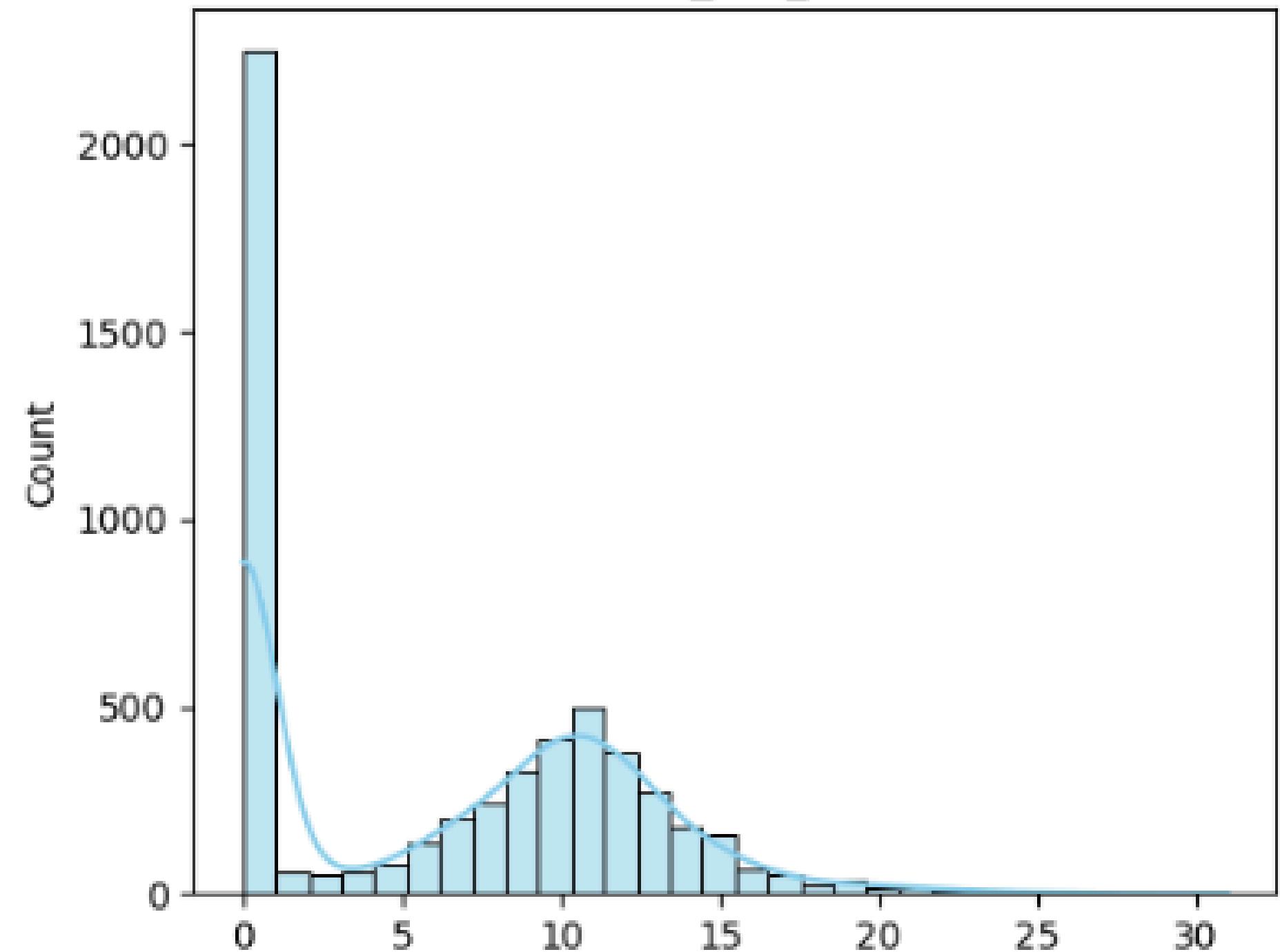
- **raceId**: Reference to races.csv
- **driverId**: Reference to drivers.csv
- **stop**: Pit stop number for the driver
- **lap**: Lap when the pit stop occurred
- **time**: Pit stop time string
- **duration**: Duration in seconds
- **milliseconds**: Duration in milliseconds



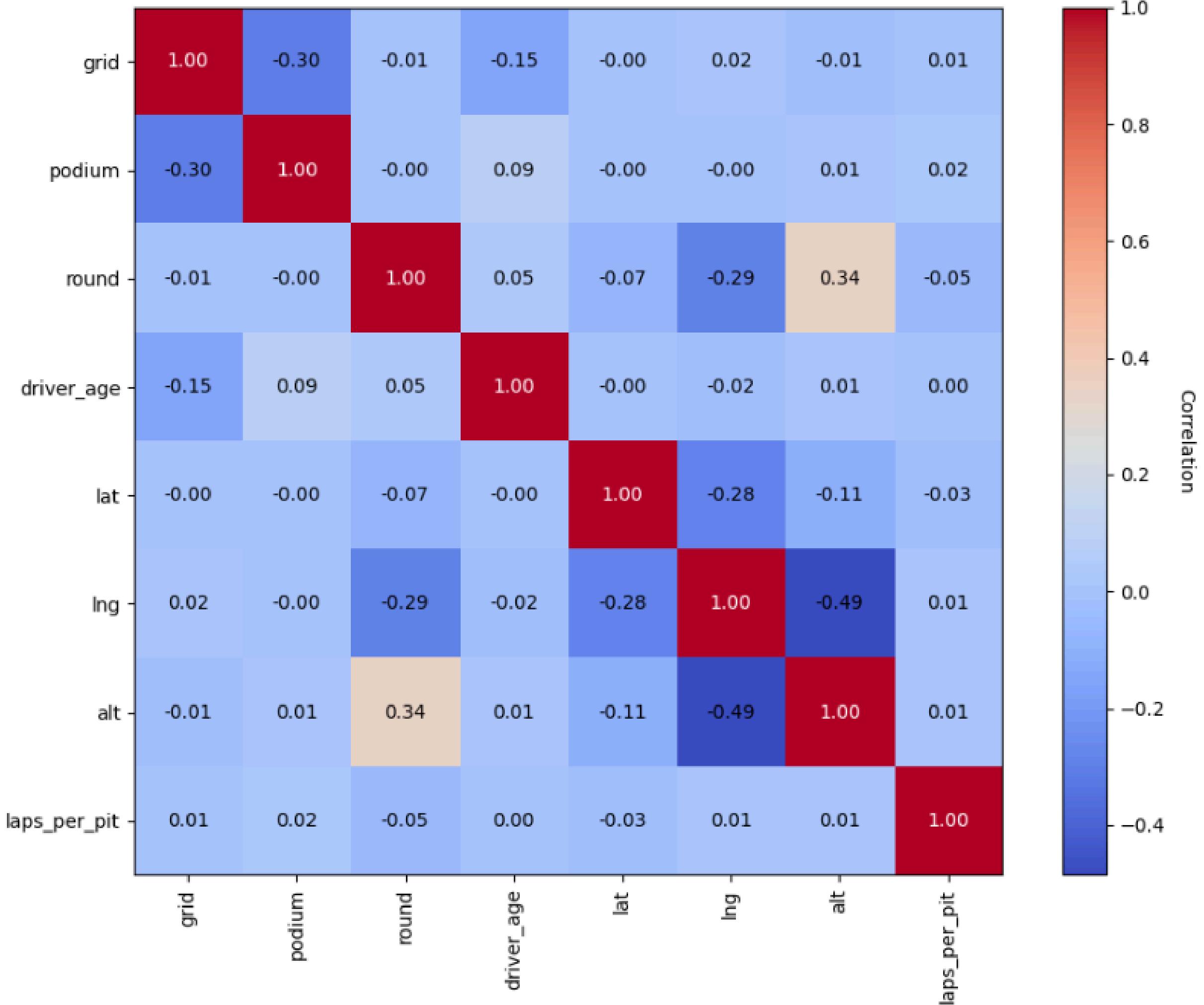
alt

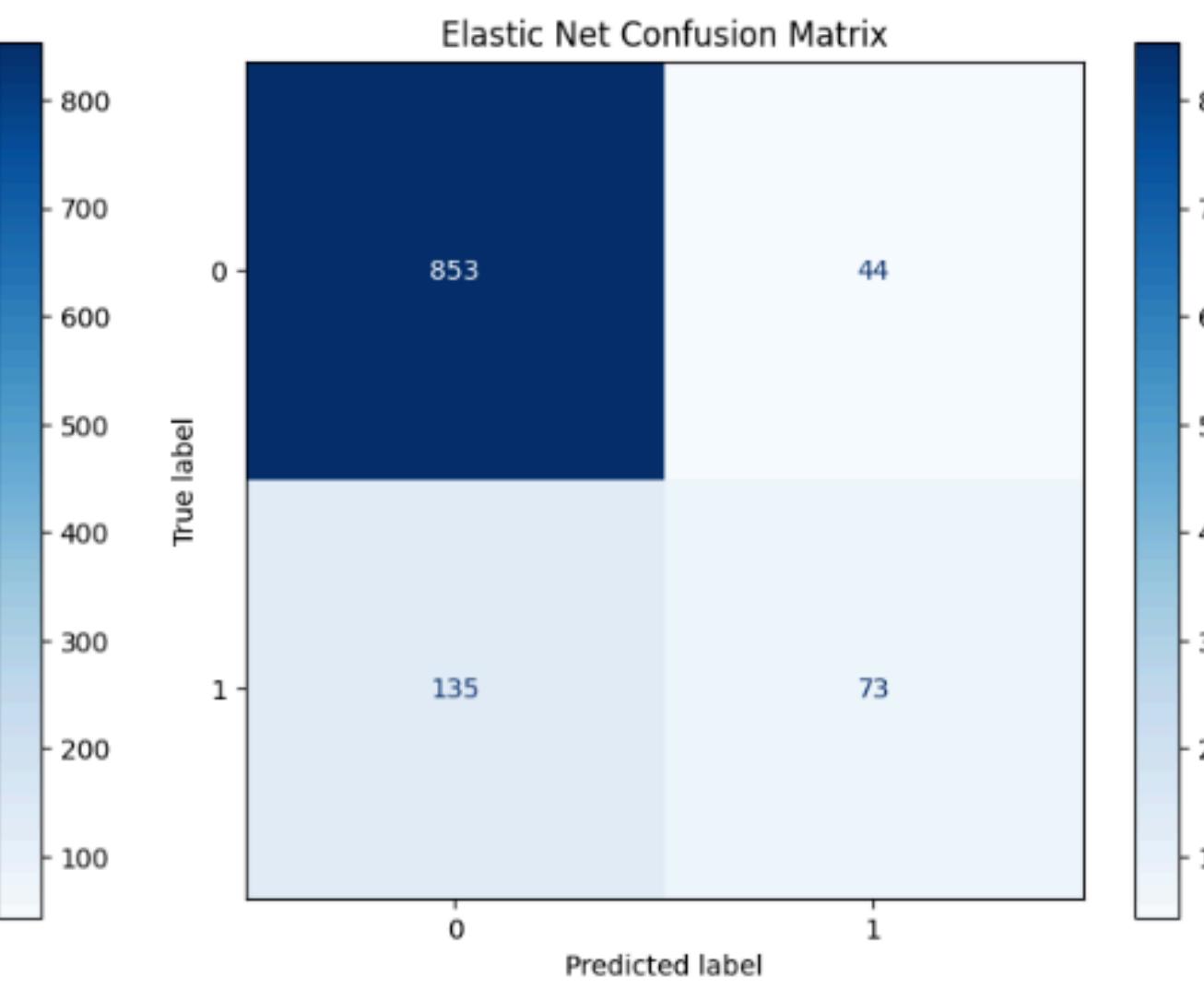
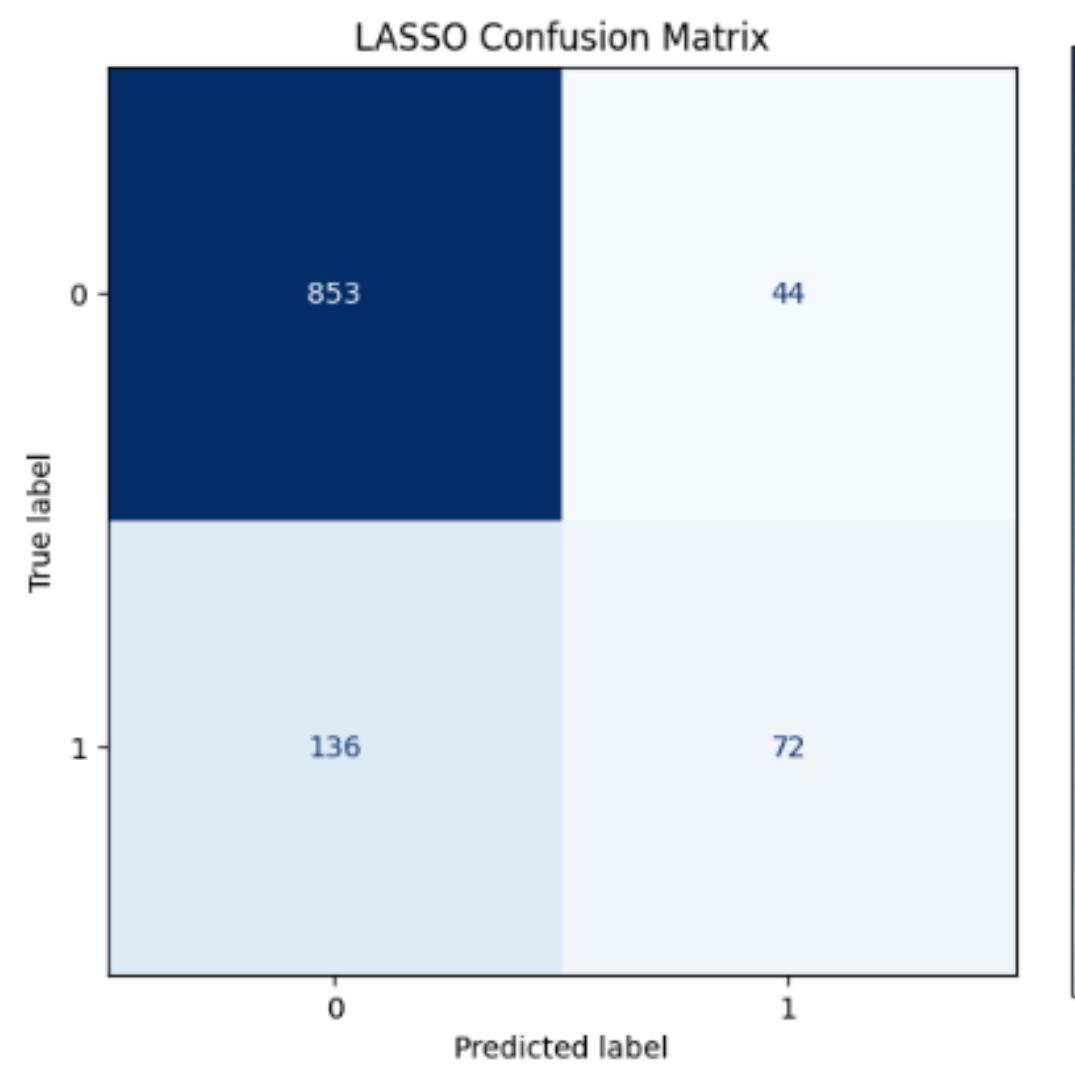
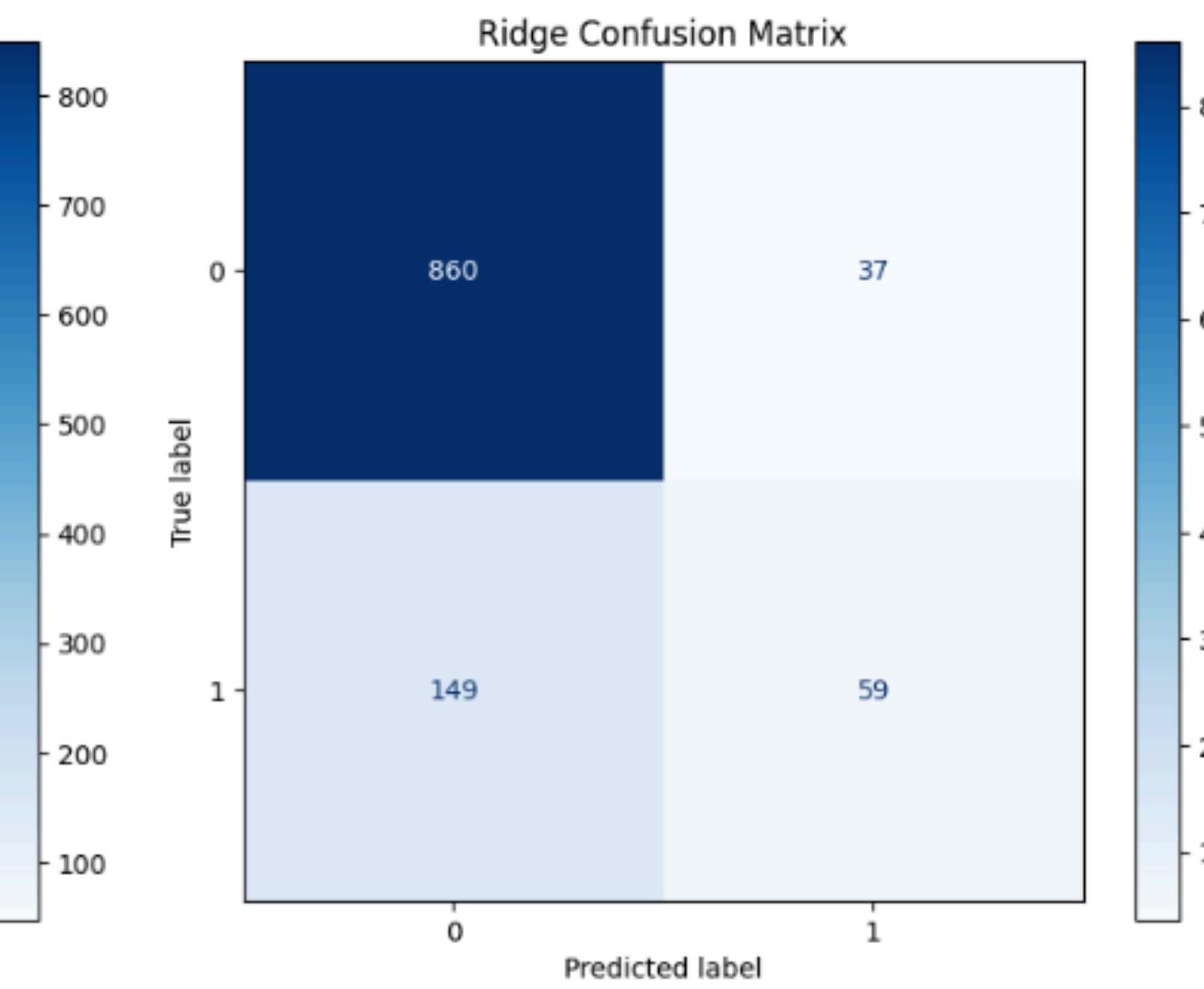
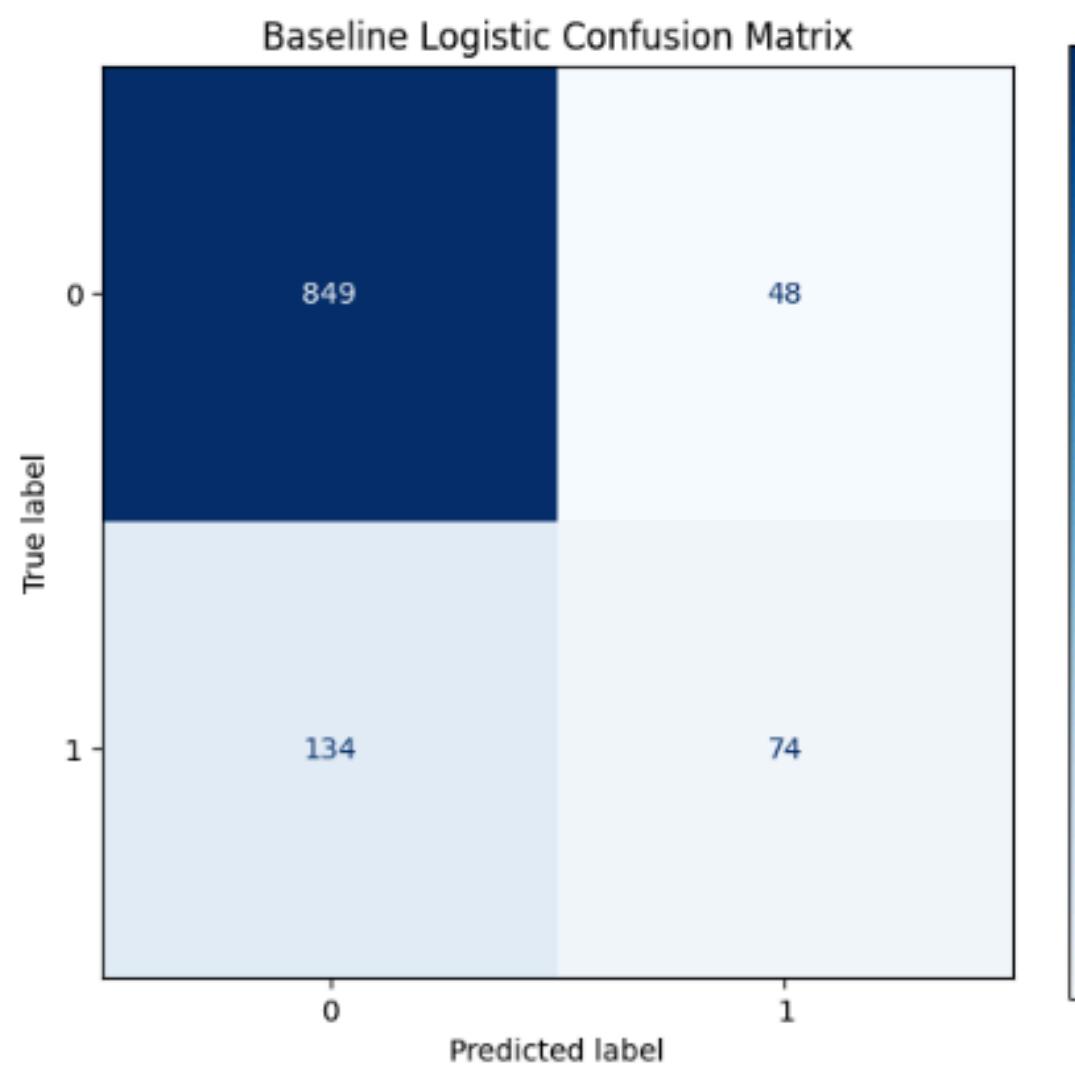


laps\_per\_pit

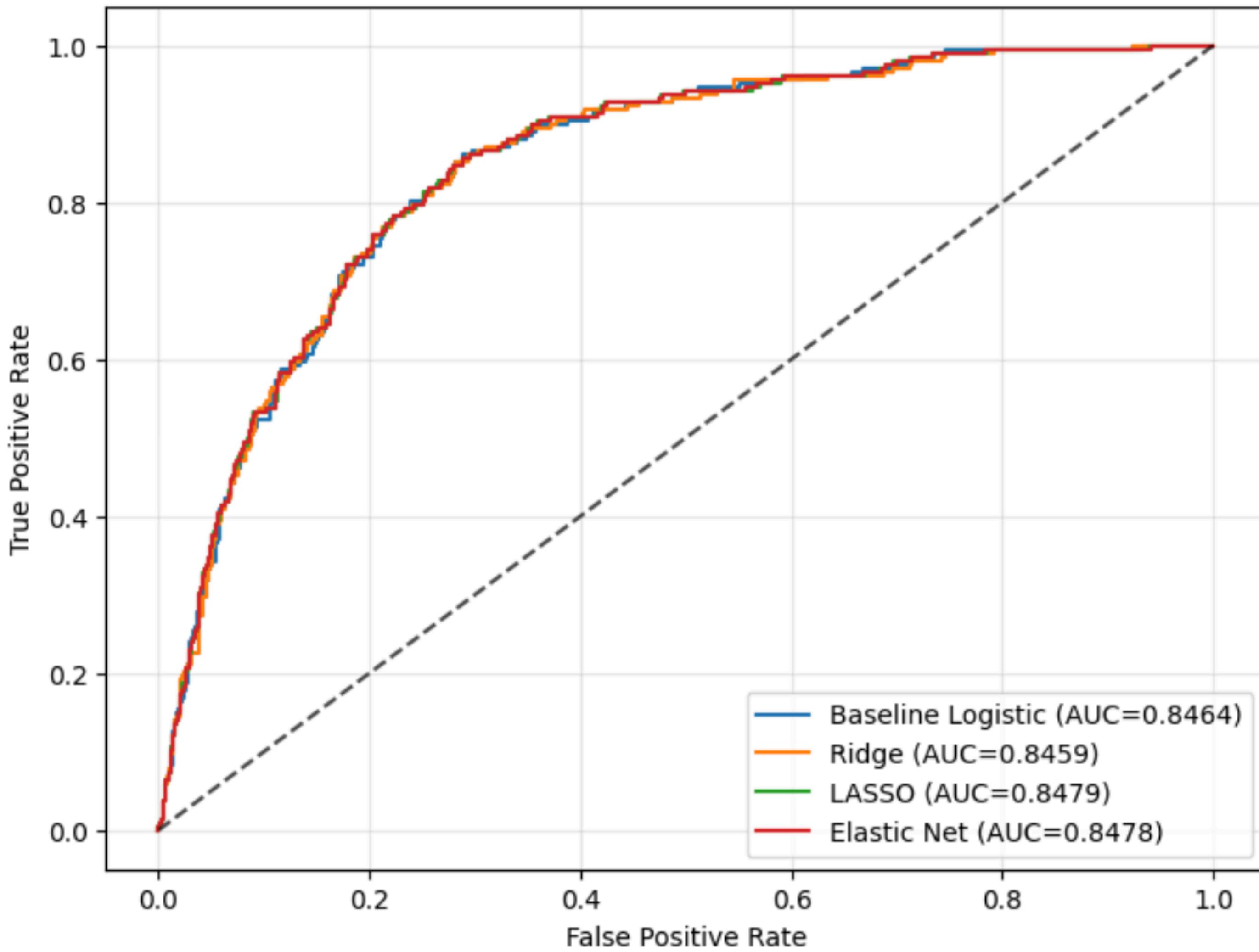


Correlation Matrix (Numeric Columns Only)

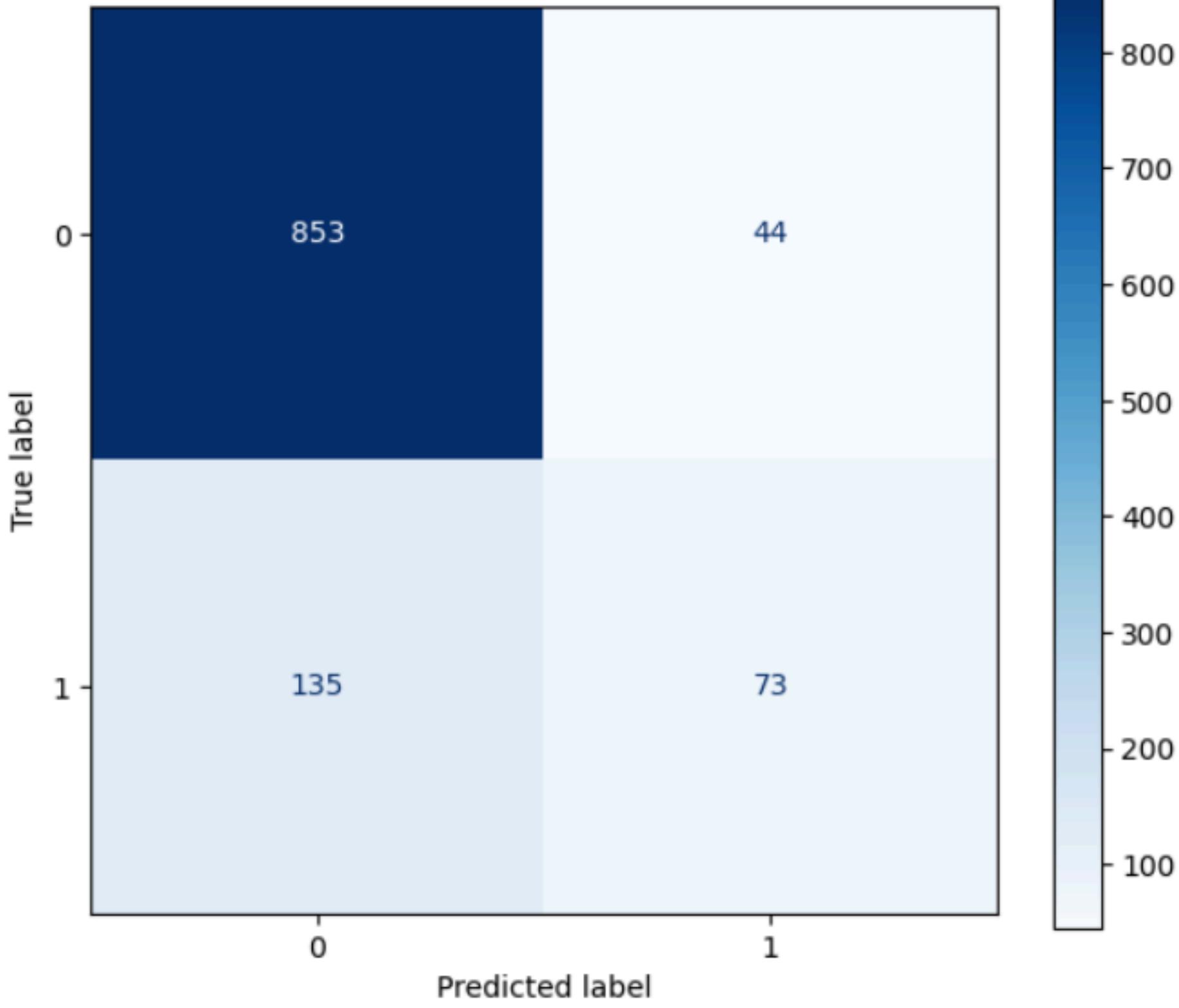




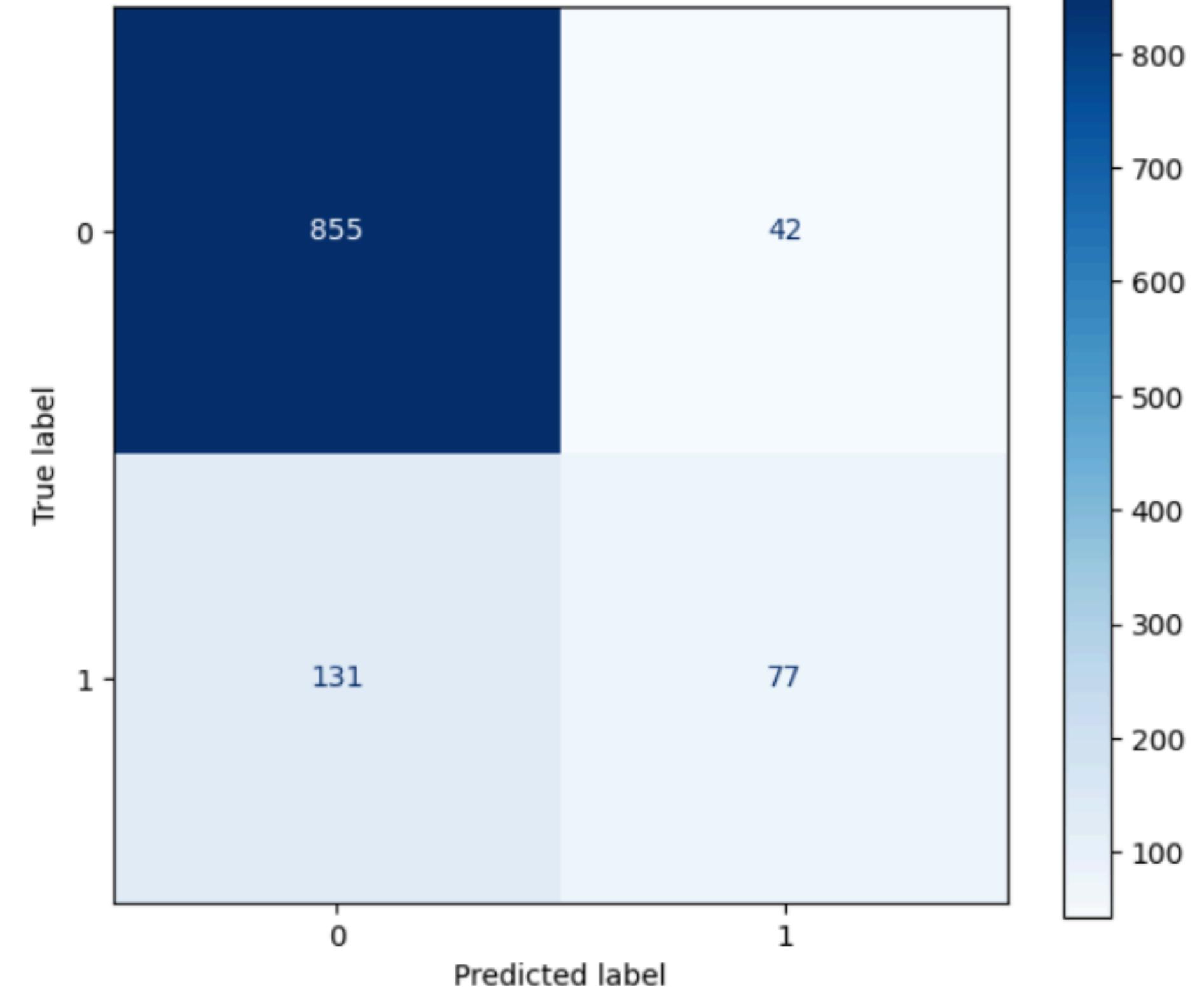
## ROC Curves (Test Set)



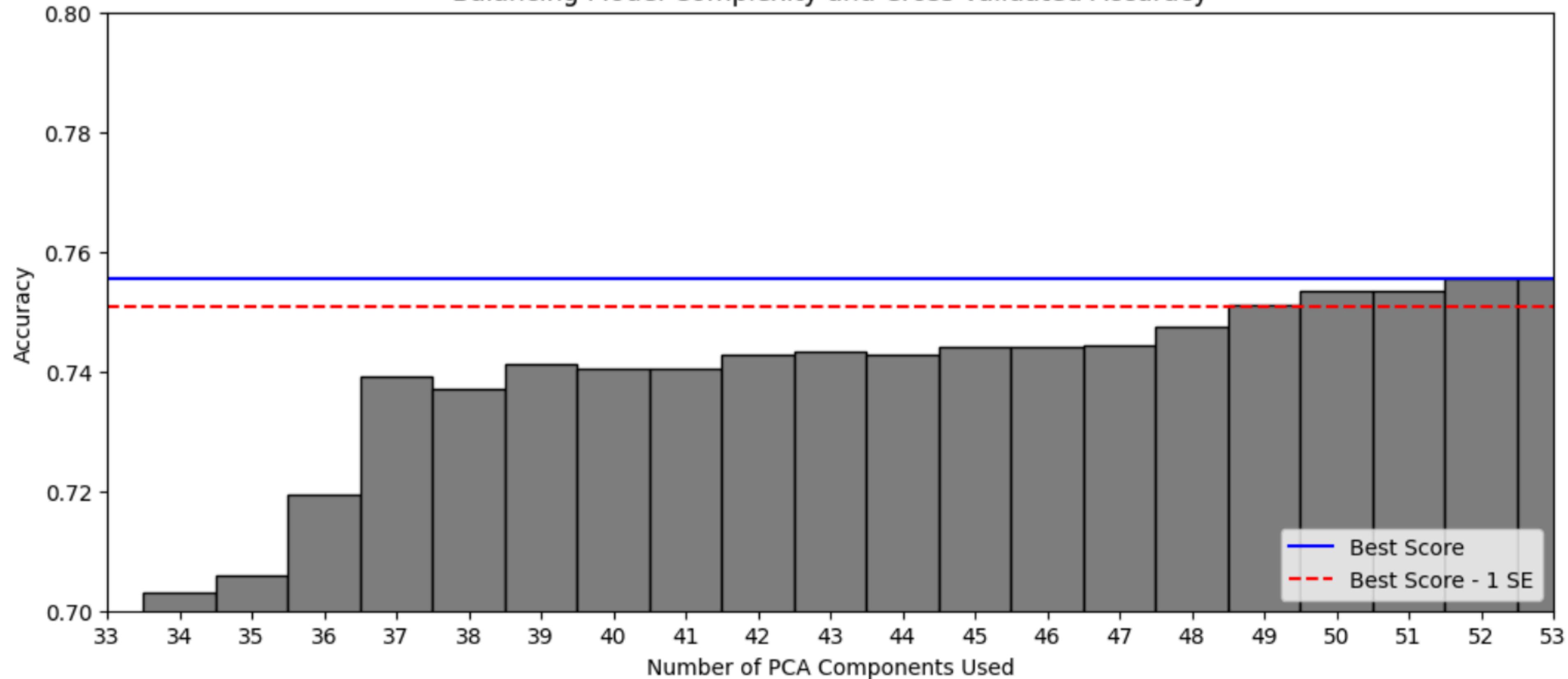
Original Elastic Net Confusion Matrix



Elastic Net w/ Interactions Confusion Matrix



## Balancing Model Complexity and Cross-Validated Accuracy



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	1.082036

Final Elastic Net Coefficients with Selected Interaction  
(95% Bootstrap CI not crossing 0)

