ON THE RIGHT TRACK

WHO WINS?

PREDICTING F1 RACE POINT FINISHERS USING AN ENSEMBLE TREE MODEL REGRESSOR

PROJECT

ESSENTIALS & DETAILS

FAST FACTS

TEAMS AND DRIVERS

10 TEAMS ON THE GRID

2 DRIVERS PER TEAM

RACES PER SEASON

>20 RACES IN A SEASON

RACE WEEKEND

3 PRACTICE SESSIONS

1 QUALIFYING SESSION

1 RACE SESSION

POINTS SYSTEM

TOP 10 WILL SCORE POINTS AT THE END OF THE RACE

PROJECT OBJECTIVE

PREDICT POINTS FINISHERS

How can we **develop a regression model that predicts Formula 1 drivers likely to finish in the top 10**, considering the criticality of every point towards the championship?

BUSINESS VALUE

TEAM STRATEGY

Establish **expectations and strategies** while **identifying areas for improvement.**

SPORTS BETTING

Allow **bettors to make informed decisions on who to bet on** during the race

F1 FANTASY

Aid in **building F1 fantasy teams and planning for transfers** on race weekends

$75M

Up to 75 Million Dollars in sponsorship per year

>$1B

Estimated market size of the Formula 1 betting market

$87B

Fantasy sports projected to grow by year 2031

MODEL DETAILS

HOW DID WE CHOOSE THE BEST MODEL?

01

SELECT BEST MODEL

Build baseline model: walk-forward
Time series split for assessing robustness

02

IMPLEMENT MODEL

Produce race result predictions
for the final few races of the season

03

EXPLAINABILITY

Use interpretation techniques and
make actionable suggestions

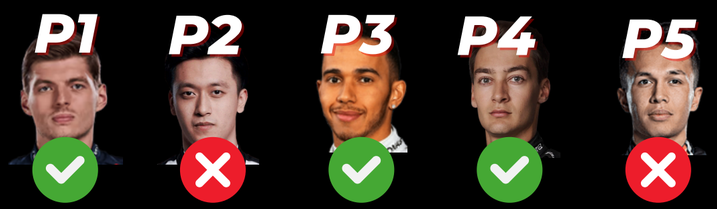
METHODOLOGY

RANKING EVALUATION METRIC

AVERAGE PRECISION @ K

- measures the **percentage of relevant results** among top k results
- Takes into consideration the **order of relevant results**

EXAMPLE



= 0.81

= 0.53

COMPARE MODELS

RANDOM FOREST

n_estimators: **300**
max_depth: **1**
max_features: **0.3**

MAP@K SCORE:
89%

XGBOOST REGRESSOR

n_estimators: **300**
max_depth: **8**
learning_rate: **0.05**

MAP@K SCORE:
83%

GRADIENT BOOSTING

n_estimators: **300**
max_depth: **10**
learning_rate: **0.1**

MAP@K SCORE:
78%

LINEAR REGRESSION

MAP@K SCORE:
64%

Ensemble Models were mainly **used to capture the complexity of the data set**

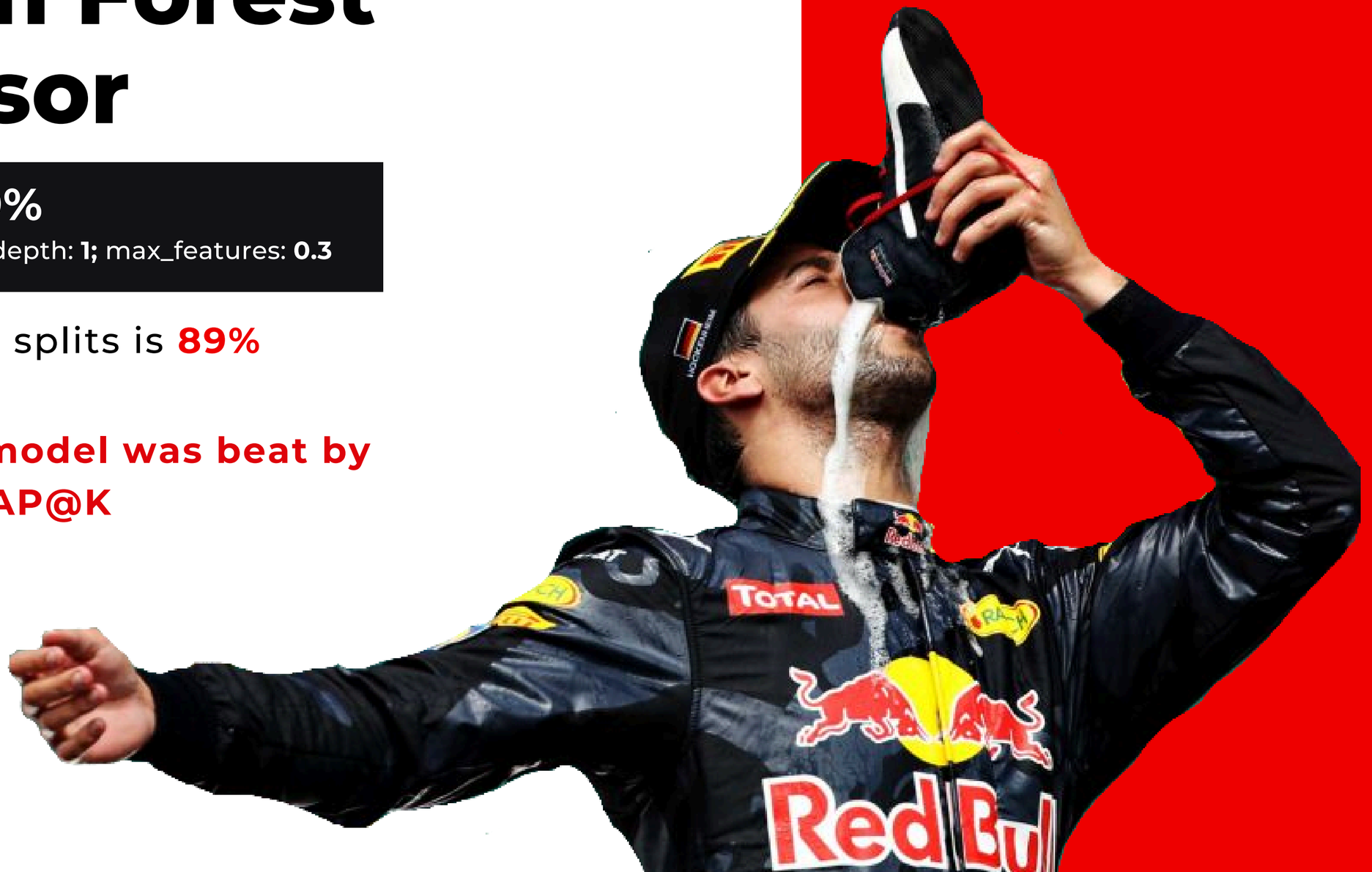
BEST MODEL PREDICTION

Random Forest Regressor

MAP@K SCORE: **89%**

n_estimators: **300;** max_depth: **1;** max_features: **0.3**

- Average for all splits is **89% MAP@K**
- The **baseline model was beat by almost 10% MAP@K**



PREDICTIONS

LAST FOUR RACES OF THE 2023 SEASON

FEATURES USED FOR PREDICTIONS:

CURRENT RACE	QUALIFYING
PREVIOUS RACE	DRIVER/TEAM PTS DRIVER/TEAM WINS DRIVER POSITION TEAM POSITION QUALIFYING POS. FASTEST LAP RANK AVE. LAP SPEED
2 - 5 RACES AGO	END RACE POSITION QUALIFYING POS.

2022 AND 2023 F1 SEASONS

DATA SET

573 ROWS

20 FEATS.

Data was extracted from the **Ergast Development website,** an **API that provides Formula 1 data** since the 1950s World Championship, and **Wikipedia** for additional information.

PREDICTION STEPS

01 Predict **one race at a time** per driver for all races.

02 Each driver has **different set of historical data**

03 Lags: **most recent completed race** by a driver.

04 **Exclude drivers affected by unpredictable** race events*

* crashes or engine failures

2023 MEXICO GRAND PRIX

VER	PIA
LEC	NOR
SAI	OCO
HAM	GAS
RUS	RIC

AP@K: 89%

2023 LAS VEGAS GRAND PRIX

VER	RUS
LEC	ALO
PER	OCO
HAM	PIA
SAI	STR

AP@K: 100%

2023 SAO PAULO GRAND PRIX

VER	NOR
PER	OCO
HAM	PIA
SAI	STR
ALO	GAS

AP@K: 88%

2023 ABU DHABI GRAND PRIX

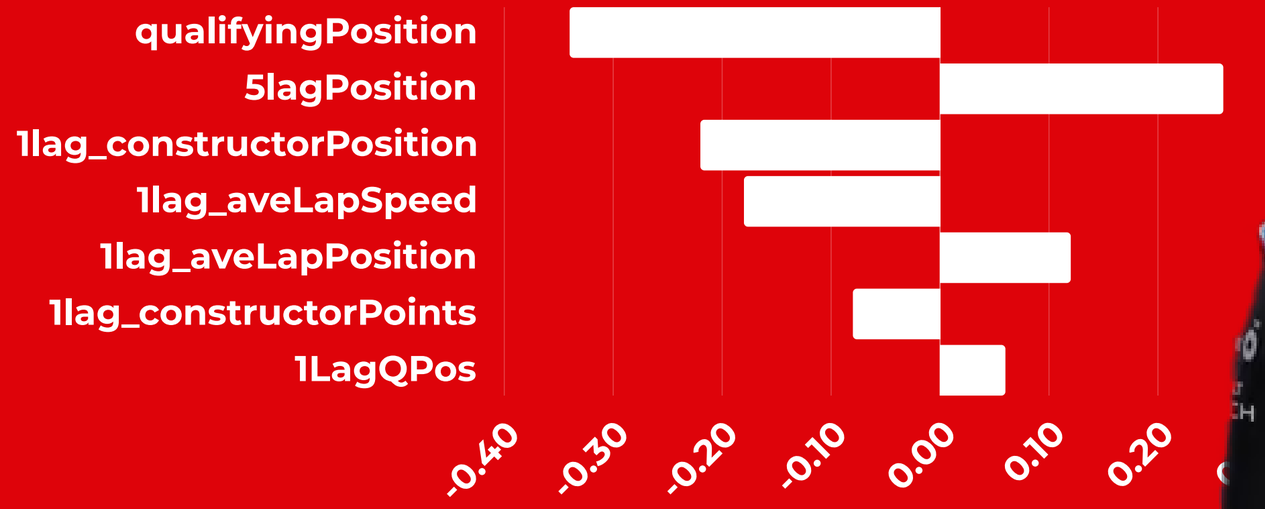
VER	RUS
PER	NOR
LEC	ALO
HAM	OCO
PIA	STR

AP@K: 89%

91.5% MAP@K

Predicting the last four races of the 2023 season

GASLY PREDICTED TO BE >10TH PLACE IN THE ABU DHABI GRAND PRIX

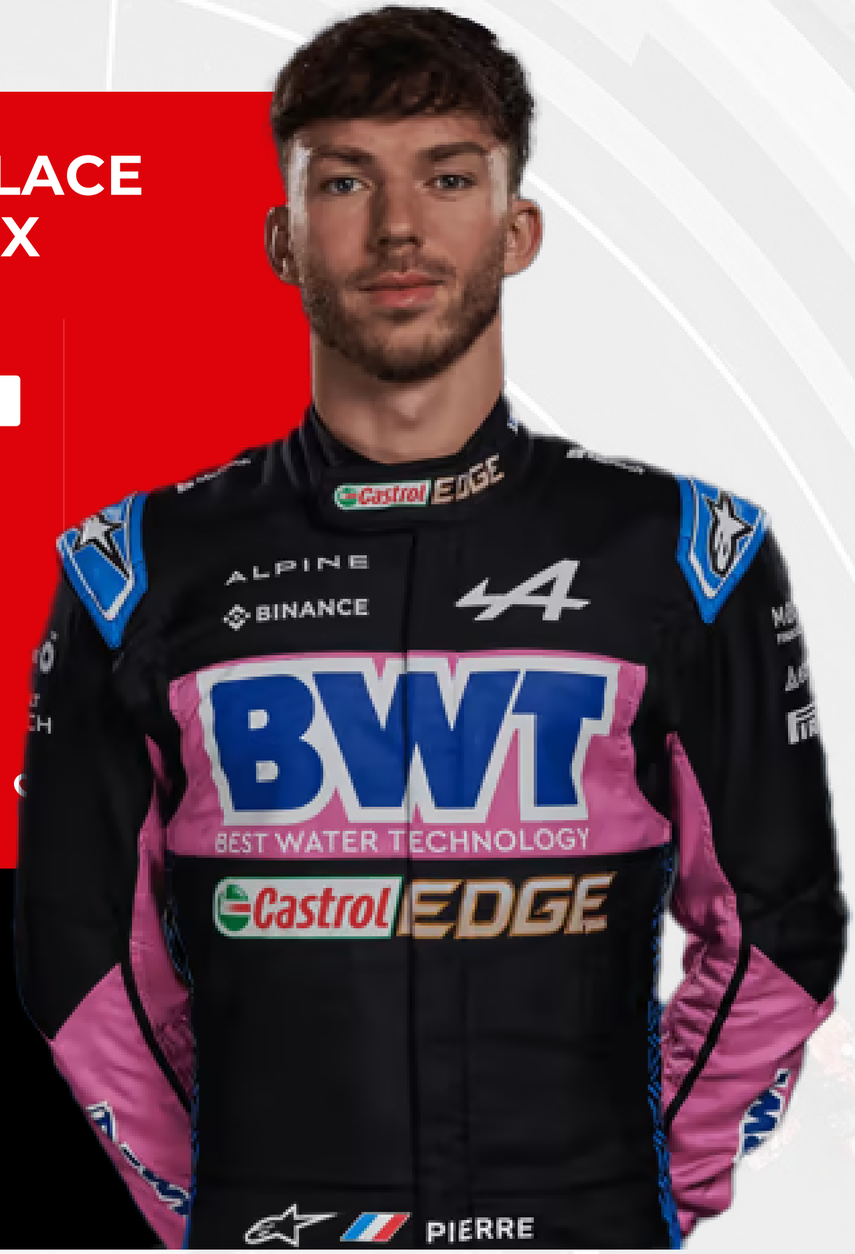


PIERRE GASLY

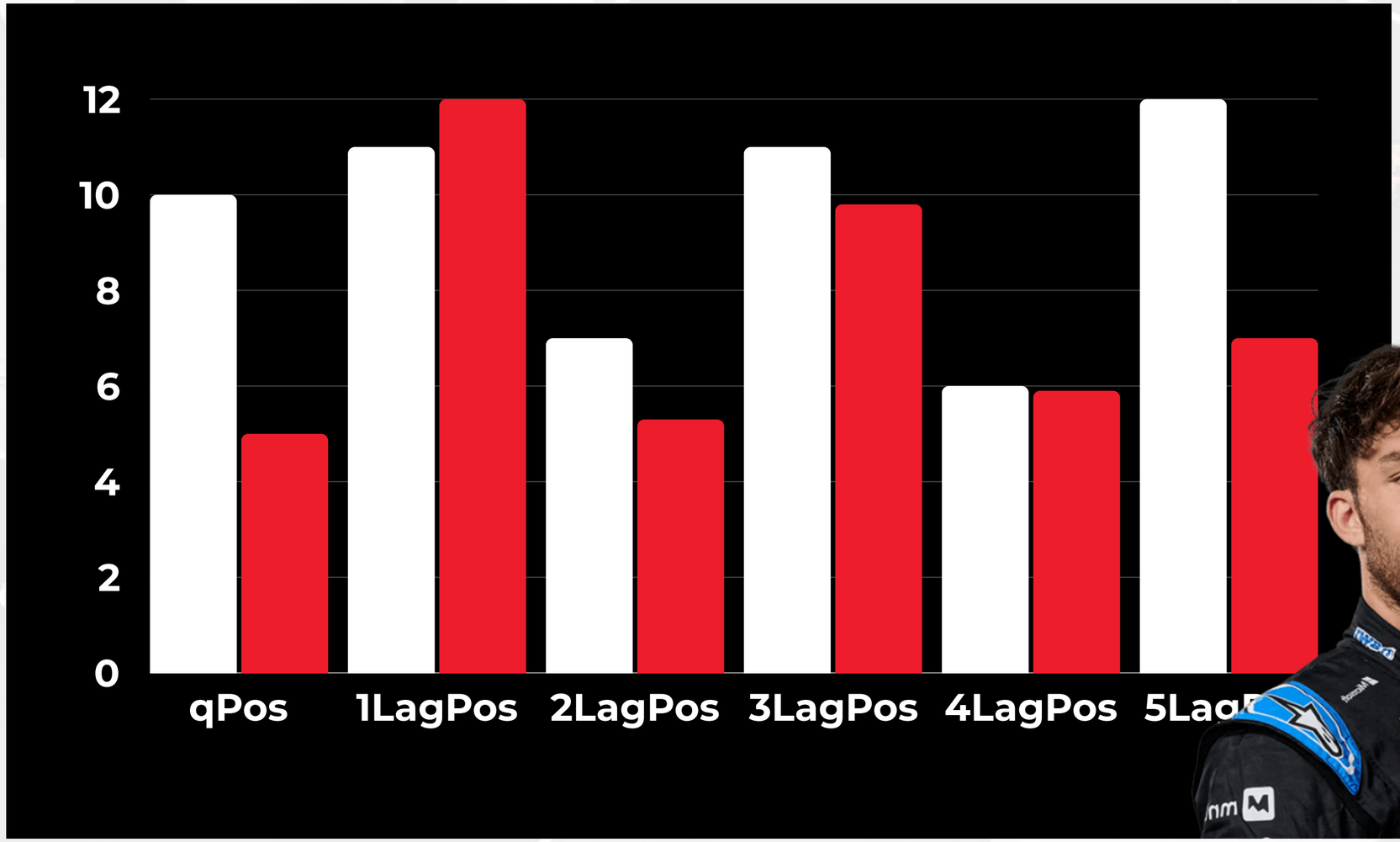
REASONS FOR PREDICTION

9th - 12th
Position in the current qualifying

11th - 13th
End race position 5 races ago



WHAT CAN GASLY DO TO IMPROVE?



10.0 to 5.0

improve qualifying position 5 races from now

9.4 to 8.0

average ending race position to target in the next 5 races

THANK YOU!

APPENDIX

BASE-LINE COMPARISON

Walk-forward

- Using **Finishing Position in the Previous Race** as the prediction
- Score to beat is **MAP@K score of 80%** which is the average scores of all test splits

TESTING MODEL ROBUSTNESS

Time Series Split

- **Maximum test size constrained** to the **driver with the lowest number of races**

COMPARE MODELS

RANDOM FOREST

n_estimators: **300**
max_depth: **1**
max_features: **0.3**

MAP@K SCORE:
89%

XGBOOST REGRESSOR

n_estimators: **300**
max_depth: **8**
learning_rate: **0.05**

MAP@K SCORE:
83%

GRADIENT BOOSTING

n_estimators: **300**
max_depth: **10**
learning_rate: **0.1**

MAP@K SCORE:
78%

LINEAR REGRESSION

MAP@K SCORE:
64%

Ensemble Models were mainly **used to capture the complexity of the data set**