FORMULA ONE RACE PREDICTION MODEL

SPENCER STAUB

THE GEORGE WASHINGTON UNIVERSITY

FALL 2022

Formula One Background

- Season Format
  - 20+ events per season organized into Grand Prix weekends
  - 20 drivers per year
  - 10 teams per year with 2 drivers per team
- Grand Prix Format
  - 3 Practice sessions total on Friday and Saturday
  - 1 Qualifying session
  - 1 Race
- Practice Format
  - Open track
  - Drivers record 10-40 laps per session
- Qualifying Format
  - Used to determine the starting order for the race
  - Quickest driver starts at the front
- Race Format
  - 50-75 laps
  - 2 hours long
  - Top 10 finishers score points

Why Formula One?

- A unique challenge with nearly unlimited scope
  - Challenges:
    - **Millions of data points**
      - Race Data
      - Qualifying Data
      - Practice Data
      - Telemetry Data
      - Weather Data
    - **Target Variable?**
      - Categorical?
      - Numerical?
      - Binary?
  - Unlimited Scope:
    - **Feature Engineering Methods**
      - Different aggregation methods
      - Varying historical window
      - Relative Performance of Car, Driver, Teammates, ect.
    - **Use of Modeling for Feature Engineering**
      - Tracks
        - Percentage of tight corners
        - Length of straights
        - Percentage of on-throttle
      - Cars
        - Top Speed
        - Cornering Performance
      - Driver Style

Opportunities to make money through Sports Betting

As a Formula One fan this topic is particularly interesting

Data

- Retrieved via the Ergast API through the FastF1 Python package
- Years Available
  - Race results & weather 1951-2022
  - Qualifying results1994-2022
  - Telemetry 2018-2022
- Available Data includes:
  - Session Results – Qualifying and Race finishing order
  - Lap Results – Lap Time, Sector Time, Tire Compound, Speed Trap
  - Car Telemetry Data – Track Position, Throttle, Brake, Distance of Driver Ahead
  - Weather Data
- 2+ Million observations of telemetry data per Grand Prix Weekend
  - 40+ Million per Season
  - 200+ Million total
- Relatively Clean Data
  - Low NA /Duplicate Rate

VODAFONE McLAREN MERCEDES

VODAFONE McLAREN MERCEDES

DRIVER

JENSON BUTTON

3

POSITION

2

- 1.075

CURRENT LAP
1:23.052

BEST LAP
1:22.565

TOP LAP SPEED

287 km/h

RACE
321km/h

AUTODROMO DI MONZA

LAP
30/53

WEATHER

Cloudy

25°c

WIND
3.5 m/s

TRACK TEMP.
45°C

DRS ZONE

POSITION

1

+1.075

CURRENT LAP
1:24.052

BEST LAP
1:23.781

TOP LAP SPEED

291 km/h

RACE
322km/h

DRIVER

LEWIS HAMILTON

4

LATITUDE
45.6301

LONGITUDE
9.29103

ON SAFETY CAR
GO TO PIT

ENGINE

EMPTY ON
LAP 51

DRS ON

In the Zone

BRAKES

THROTTLE

ENGINE

BRAKES

THROTTLE

LATITUDE
45.61628

LOGITUDE
9.28078

ON SAFETY CAR
DO NOT PIT

EMPTY ON
END

DRS

Not In the Zone

TYRES AND LAPS RACED

8 LAPS

TYRE PRESSURE AND LIFE

1.1bar 14 17 1.0bar

1.3bar 20 15 1.4bar

NEXT PIT WINDOW LAPS

36 – 39

PIT STOP TIMES
3.024 sec

ON LAP
21

RAW TELEMETRY DATA

vCar	=259.3	=258.4	kph
NGear	=6	=N	
rThrottlePedal	=100.0	=0.0	%
pBrakeR	=8.5	=0.0	°

DATA RATE

Tranferring =3.3 MB/sec

NEXT PIT WINDOW LAPS

35 – 38

PIT STOP TIMES
3.174 sec

ON LAP
19

TYRE PRESSURE AND LIFE

1.0bar 10 13 1.1bar

1.3bar 15 10 1.2bar

TYRES AND LAPS RACED

11 LAPS

PREDICTIVE TIMELINE

SAP

Powered by HANA

PLANNED PIT 1

3.024 sec

PIT WINDOW

PLANNED PIT 2

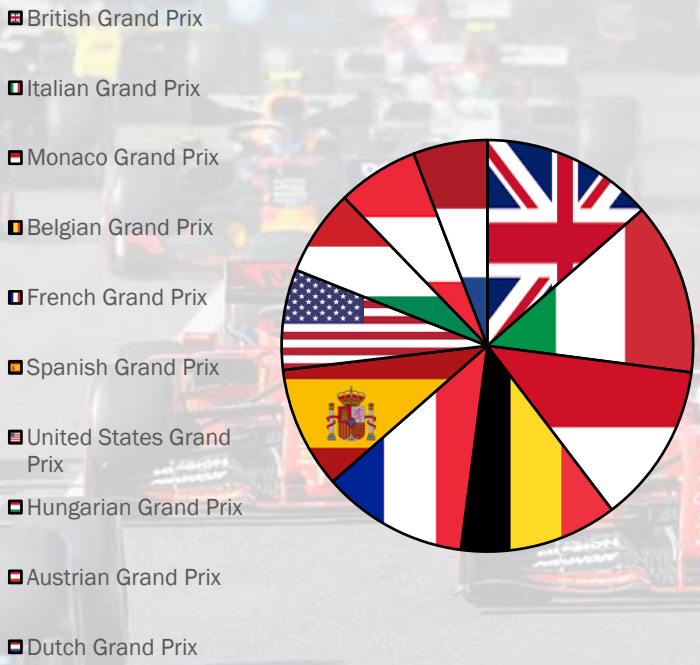
TIRE FAILURE

EMPTY

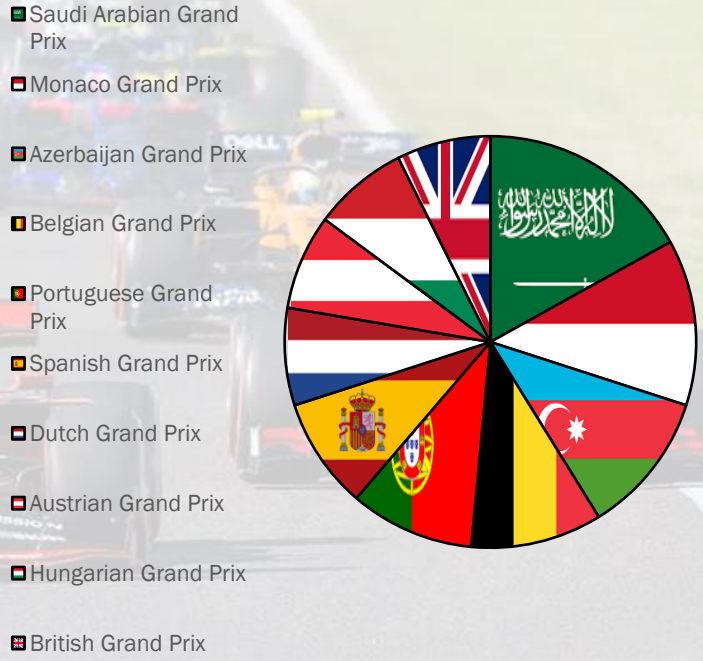
Track EDA

- 56 Different Tracks Raced (1951-2022)
- Top Tracks:
  - British Grand Prix
  - Italian Grand Prix
  - Monaco Grand Prix
- Most Dangerous Tracks
  - Saudi Arabian Grand Prix
  - Monaco Grand Prix
  - Azerbaijan Grand Prix
- Drivers with Most Track Experience
  - Hamilton (15 brit, 15 Ital, 14 Mon)
  - Verstappen (7 brit, 7 Ital, 6 Mon)
  - Leclerc (4 brit, 4 Ital, 3 Mon)
- Teams with Most Track Experience
  - Ferrari (151 brit, 191 Ital, 145 Mon)
  - Mercedes (30 brit, 31 Ital, 26 Mon)
  - Red Bull (26 brit, 26 Ital, 26 Mon)

Top 10 Tracks by Frequency

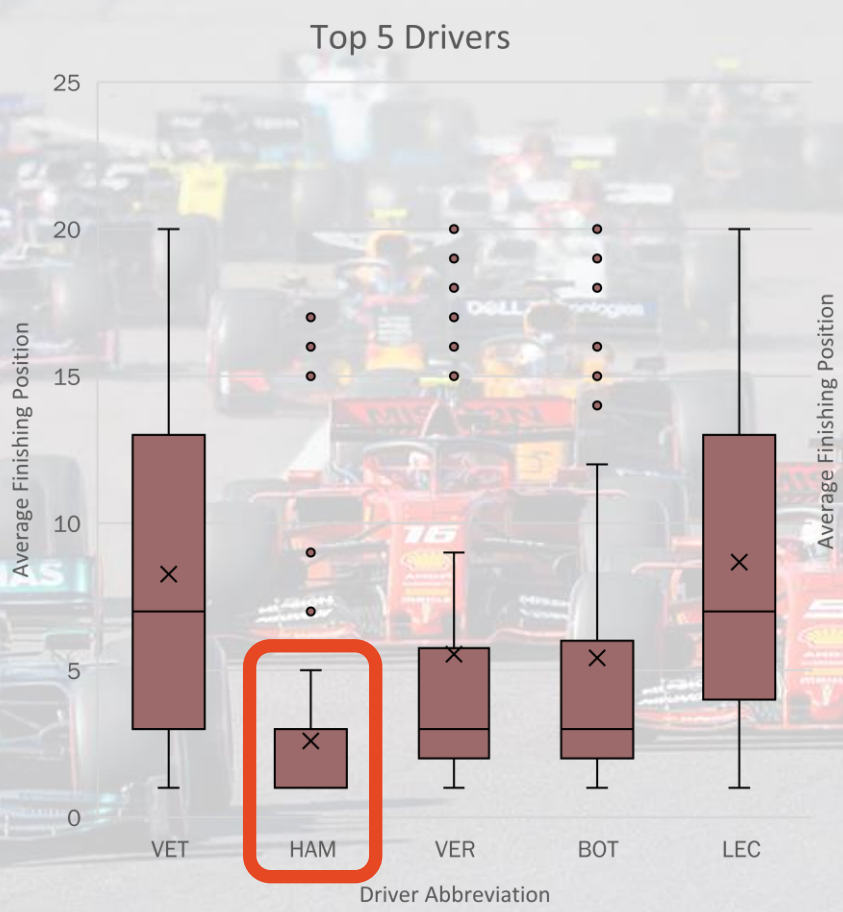


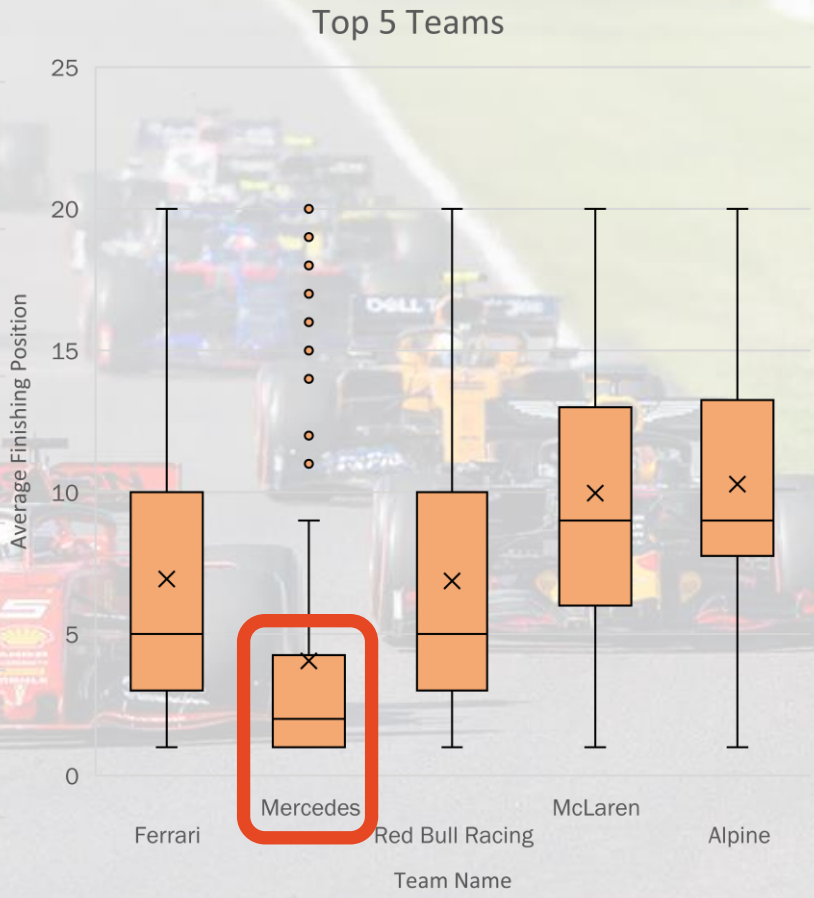
Most Dangerous Tracks By Average Accident Frequency



Team and Driver EDA

- 831 Different Drivers (1951-2022)
- 206 Unique Teams (1951-2022)
- Top Current Drivers:
  - Lewis Hamilton
  - Max Verstappen
  - Charles Leclerc
- Top Current Teams:
  - Mercedes
  - Red Bull
  - Ferrari
- Accident Rates
  - Lewis Hamilton: 0.044983
  - Max Verstappen: 0.091549
  - Charles Leclerc: 0.109756
- Car Reliability
  - Mercedes: 0.075728
  - Red Bull: 0.103659
  - Ferrari: 0.207965





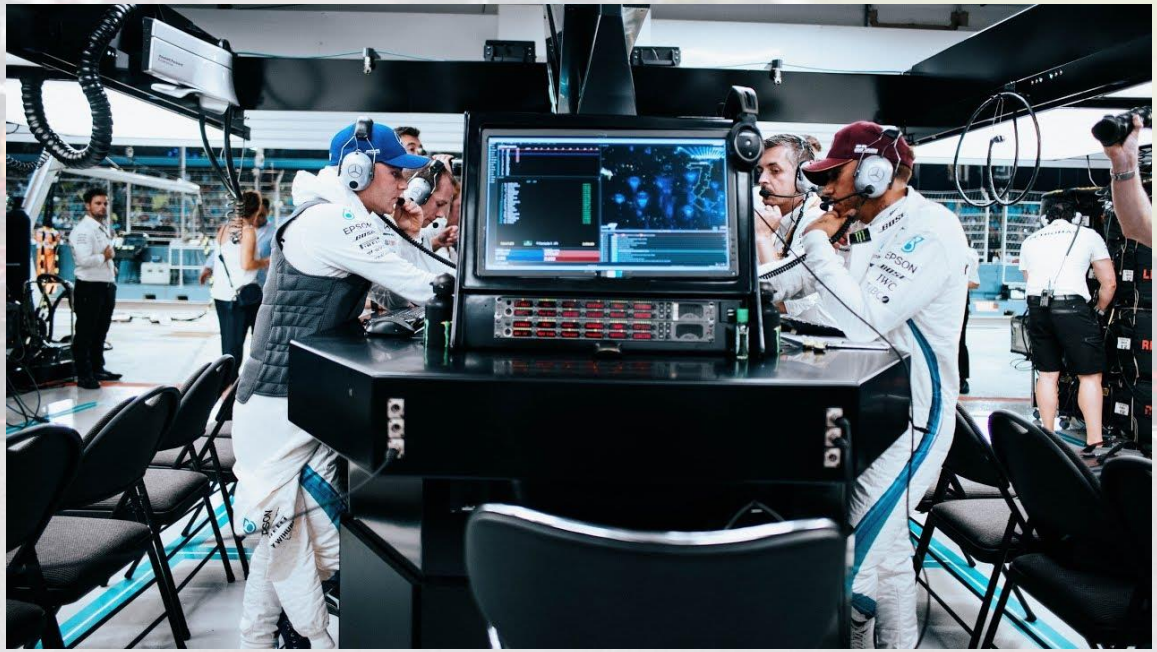
Qualifying Correlation

- Qualifying Results (where a driver starts) is highly correlated with finishing position
  - Starts in 1st: 42% Win Rate
  - Starts in 2nd : 23% Win Rate
  - Qualifying Feature Importance: 0.375
  - All other features are less than 0.02
- This CAN be a problem:
- When Qualifying is removed:
  - Max Feature Importance Value Decreases
  - Other Features Importance Values Increase
  - Model accuracy decreases, however, betting odds increase
    - The less data before a bet is made = higher betting odds + more opportunities to make money
    - Is it worth removing qualifying?

*** Monaco has a 51% First place win rate

Feature Engineering

- Aggregation Features (Grouped by Driver, Team and/or Race)
  - Historical Finishing Positions (Last, Past 10, Total History)
  - Historical Qualifying Positions (Last, Past 10, Total History)
  - Previous Years Performance (Points, Driver/Team Champion)
  - Historical Car Reliability
  - Historical Driver Reliability
  - Current Season Performance (Points)
  - Total Team Experience
  - Total Driver Experience
  - Relative Qualifying Time Deltas
- Raw Features
  - Grid Position
- Categorical Features
  - Track
  - Driver Name
  - Team Name



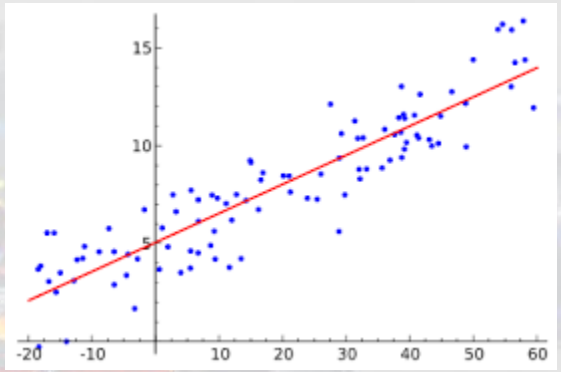
Target Engineering

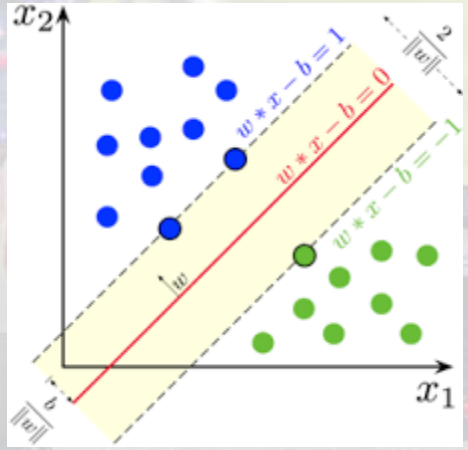
- Options:
  - Categorical Target
  - Numeric Target
  - Binary Target
- Binary Target
  - Winner
  - Top Two Finishers
  - Top Three Finishers (Podium)
- Target Imbalance
  - Binary targets will create imbalance
    - Winner 95% / 5%
    - Top Two 90% / 10%
    - Podium: 85% / 15%

Driver	Finishing Position	Binary Target: Winner
Max Verstappen	1	True
Lewis Hamilton	2	False
Lando Norris	3	False
Sergio Perez	4	False
Carlos Sainz	5	False
Valtteri Bottas	6	False
Charles Leclerc	7	False
Yuki Tsunoda	8	False
Esteban Ocon	9	False
Daniel Ricciardo	10	False
Fernando Alonso	11	False
Pierre Gasly	12	False
Lance Stroll	13	False
Antonio Giovinazzi	14	False
Sebastian Vettel	15	False
Nicholas Latifi	16	False
George Russell	17	False
Kimi RÃ¤ikkÃ¶nen	18	False

Model Development

- Models Used
  - Logistic Regression
  - SVM
  - XGboost
- Cross Validation
  - Standard Cross Validation
  - Historical Cross Validation
- Data Manipulation
  - Oversampling
  - Undersampling
  - Removal of Crash / Breakdown Data
- Scaling
  - Standard Scaling for Logistic Regression

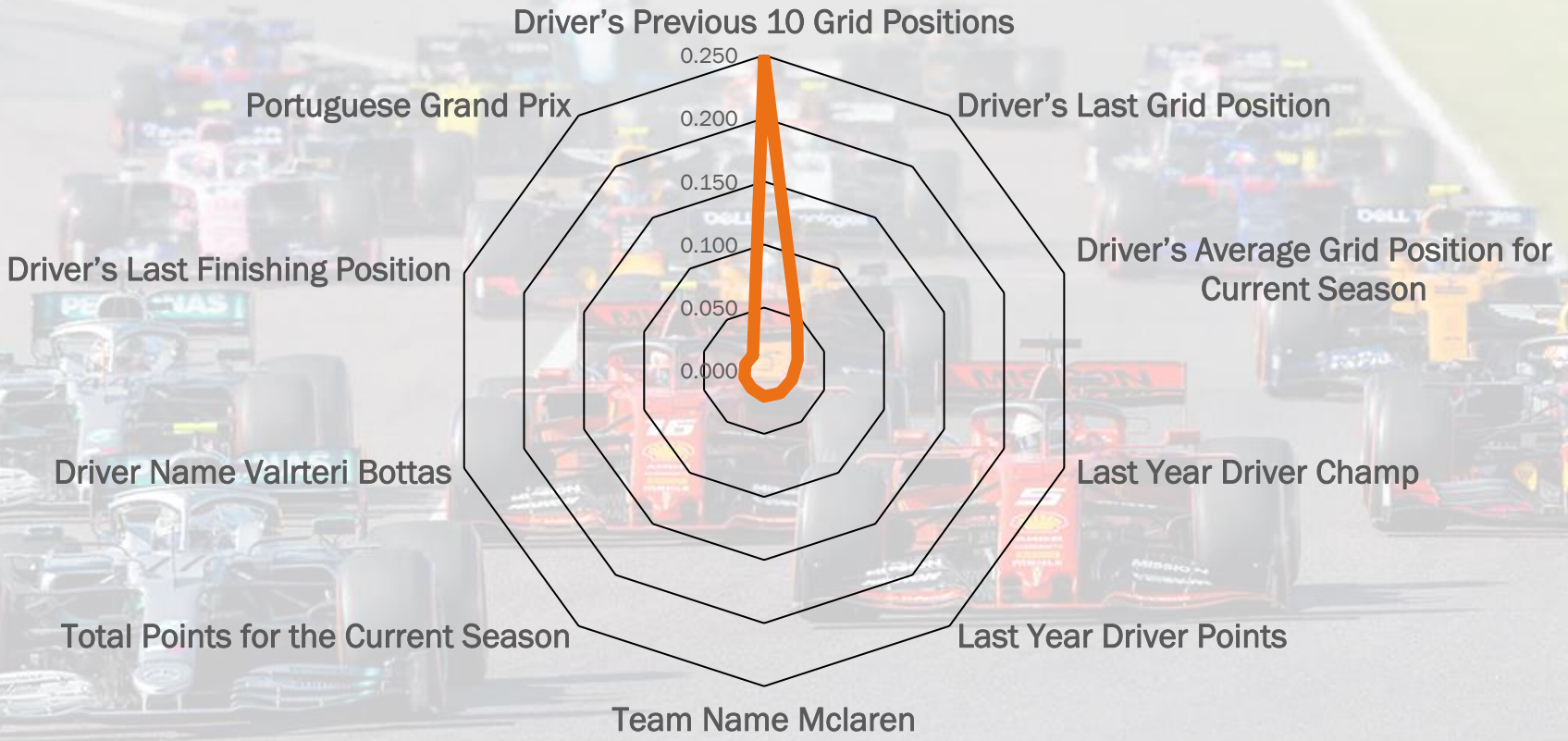




Further XGBoost Development

- **Remove Qualifying Data**
- Run Model for all three binary targets
- XGBoost Parameters:
  - Max Depth: 9
  - Estimators 200
  - RandomOverSampler Strategy: 1
- Measure results for all three target outputs + combined result output:
  - Accuracy of 1st place predictions
  - Accuracy of predicting a podium finish
  - Accuracy of all finishing place predictions

Feature Importance



Model Scoring with Binary Classification

- Create Probabilities of a True Classification
- Shuffle Dataframe and remove index
- Sort results by event date
- Sort results by probability
- Rank drivers from highest probability to lowest
  - Highest = first place
  - Lowest = last place
  - If probabilities tie: Order by Grid Position or best performance for that specific track
- Calculate Accuracy
  - First Place
  - Podium Prediction
  - Total Accuracy

Team	Driver	Prob.	Rank	Finishing Position
Mercedes	Lewis Hamilton	0.870893	1	1
Red Bull Racing	Max Verstappen	0.843082	2	20
Mercedes	Valtteri Bottas	0.396337	3	3
McLaren	Lando Norris	0.394122	4	4
Ferrari	Charles Leclerc	0.341299	5	2
Ferrari	Carlos Sainz	0.225045	6	6
AlphaTauri	Pierre Gasly	0.196246	7	11
Red Bull Racing	Sergio Perez	0.172109	8	16
McLaren	Daniel Ricciardo	0.079183	9	5
Alpine	Fernando Alonso	0.07317	10	7
Aston Martin	Sebastian Vettel	0.068455	11	19
AlphaTauri	Yuki Tsunoda	0.068271	12	10
Williams	Nicholas Latifi	0.067565	13	14
Aston Martin	Lance Stroll	0.067502	14	8
Alfa Romeo Racing	Antonio Giovinazzi	0.067484	15	13
Haas F1 Team	Nikita Mazepin	0.067484	16	17
Alfa Romeo Racing	Kimi RÃ¤ikkÃ¶nen	0.067484	17	15
Haas F1 Team	Mick Schumacher	0.067484	18	18
Alpine	Esteban Ocon	0.067073	19	9
Williams	George Russell	0.066678	20	12

*** 2021 British Grand Prix

Model Results

Year	Target	With Qualifying	Without Qualifying
2019	1st Place Target	0.761	0.571
2020	1st Place Target	0.75	0.6875
2021	1st Place Target	0.818	0.772
Average	1st Place Target	0.776	0.677

- First Place Target is best for predicting across all three tested years
- Including qualifying data is better for prediction by about 10% (1-3 races per season)

Betting Results for 2021 Season

Race	Name	Prediction	Result	Odds
Abu Dhabi Grand Prix	Max Verstappen	1	1	3
Austrian Grand Prix	Max Verstappen	1	1	1.6
Azerbaijan Grand Prix	**Max Verstappen**	**1**	**18**	**N/A**
Bahrain Grand Prix	Lewis Hamilton	1	1	2.6
Belgian Grand Prix	Max Verstappen	1	1	2.1
British Grand Prix	**Max Verstappen**	**1**	**20**	**N/A**
Dutch Grand Prix	Max Verstappen	1	1	2.1
Emilia Romagna Grand Prix	Lewis Hamilton	1	1	2.25
French Grand Prix	Max Verstappen	1	1	2.4
Hungarian Grand Prix	**Lewis Hamilton**	**1**	**2**	**N/A**
Italian Grand Prix	**Max Verstappen**	**1**	**18**	**N/A**
Mexico City Grand Prix	**Lewis Hamilton**	**1**	**2**	**N/A**
Monaco Grand Prix	Max Verstappen	1	1	2.4
Portuguese Grand Prix	Lewis Hamilton	1	1	2.25
Qatar Grand Prix	Lewis Hamilton	1	1	1.65
Russian Grand Prix	Lewis Hamilton	1	1	1.6
São Paulo Grand Prix	Lewis Hamilton	1	1	3.25
Saudi Arabian Grand Prix	Lewis Hamilton	1	1	1.5
Spanish Grand Prix	Lewis Hamilton	1	1	2.1
Styrian Grand Prix	Max Verstappen	1	1	2.25
Turkish Grand Prix	**Max Verstappen**	**1**	**2**	**N/A**
United States Grand Prix	Max Verstappen	1	1	2.4

- Qualifying data removed
- Red: Incorrect Prediction
- Accuracy: 77%
- Return on Investment: 25.1%

Continued Development Post Graduation:

- Neural Network Classification
- Custom Loss/Accuracy Function
- Additional target variables? Combinations?
- Additional Feature Engineering
  - Telemetry Data
  - Lap Data
  - Practice Data

Thank You

Questions?

References

- FastF1 Api: https://theoehrly.github.io/Fast-F1/
- Github Repo: https://github.com/SpencerStaub/Capstone