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, etc.
  ❖ 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
Retrieved via the Ergast API through the FastF1 Python package

- Years Available
  - Race results & weather 1951-2022
  - Qualifying results 1994-2022
  - 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
Data
❖ Retrieved via the Ergast API through the FastF1 Python package
❖ Full Data Available from 2018-2022
❖ Qualifying and Race Results available 1951-2022
❖ Available Data includes:
  ❖ Session Results – Qualifying and Race finishing order
  ❖ Lap Results – Lap Time, Sector Time, Tire Compound, Speed
  ❖ 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
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:
- British Grand Prix
- Italian Grand Prix
- Monaco Grand Prix
- Belgian Grand Prix
- French Grand Prix
- Spanish Grand Prix
- United States Grand Prix
- Hungarian Grand Prix
- Austrian Grand Prix
- Dutch Grand Prix

Most Dangerous Tracks By Average Accident Frequency:
- Saudi Arabian Grand Prix
- Monaco Grand Prix
- Azerbaijan Grand Prix
- Belgian Grand Prix
- Portuguese Grand Prix
- Spanish Grand Prix
- Dutch Grand Prix
- Austrian Grand Prix
- Hungarian Grand Prix
- British Grand Prix
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%

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

<table>
<thead>
<tr>
<th>Team</th>
<th>Driver</th>
<th>Prob.</th>
<th>Rank</th>
<th>Finishing Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercedes</td>
<td>Lewis Hamilton</td>
<td>0.870893</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Red Bull Racing</td>
<td>Max Verstappen</td>
<td>0.843082</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Mercedes</td>
<td>Valtteri Bottas</td>
<td>0.396337</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>McLaren</td>
<td>Lando Norris</td>
<td>0.394122</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Ferrari</td>
<td>Charles Leclerc</td>
<td>0.341299</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Ferrari</td>
<td>Carlos Sainz</td>
<td>0.225045</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>AlphaTauri</td>
<td>Pierre Gasly</td>
<td>0.196246</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>Red Bull Racing</td>
<td>Sergio Perez</td>
<td>0.172109</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>McLaren</td>
<td>Daniel Ricciardo</td>
<td>0.079183</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Alpine</td>
<td>Fernando Alonso</td>
<td>0.07317</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Aston Martin</td>
<td>Sebastian Vettel</td>
<td>0.068455</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>AlphaTauri</td>
<td>Yuki Tsunoda</td>
<td>0.068271</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>Williams</td>
<td>Nicholas Latifi</td>
<td>0.067565</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Aston Martin</td>
<td>Lance Stroll</td>
<td>0.067502</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Alfa Romeo Racing</td>
<td>Antonio Giovinazzi</td>
<td>0.067844</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>Haas F1 Team</td>
<td>Nikita Mazepin</td>
<td>0.067484</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Alfa Romeo Racing</td>
<td>Kimi Räikkönen</td>
<td>0.067484</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>Haas F1 Team</td>
<td>Mick Schumacher</td>
<td>0.067484</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Alpine</td>
<td>Esteban Ocon</td>
<td>0.067073</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>Williams</td>
<td>George Russell</td>
<td>0.066678</td>
<td>20</td>
<td>12</td>
</tr>
</tbody>
</table>

*** 2021 British Grand Prix
## Model Results

<table>
<thead>
<tr>
<th>Year</th>
<th>Target</th>
<th>With Qualifying</th>
<th>Without Qualifying</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; Place Target</td>
<td>0.761</td>
<td>0.571</td>
</tr>
<tr>
<td>2020</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; Place Target</td>
<td>0.75</td>
<td>0.6875</td>
</tr>
<tr>
<td><strong>2021</strong></td>
<td><strong>1&lt;sup&gt;st&lt;/sup&gt; Place Target</strong></td>
<td><strong>0.818</strong></td>
<td><strong>0.772</strong></td>
</tr>
<tr>
<td>Average</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; Place Target</td>
<td>0.776</td>
<td>0.677</td>
</tr>
</tbody>
</table>

❖ 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

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

- 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/](https://theoehrly.github.io/Fast-F1/)
❖ Github Repo: [https://github.com/SpencerStaub/Capstone](https://github.com/SpencerStaub/Capstone)