Final Report:
Formula 1 Race Win Predictor

Alex Lewis
August 2022

1. Background

Formula 1 is one of the fastest growing, and most ardently traditional, sports in the world. A true global spectacle, its teams and drivers push the bounds of technology and physical endurance as they vye for both individual and team championships each season. As with any race, being able to accurately predict the winner of a given Formula 1 “grand prix” is immensely valuable to both competitors and fans.

But the sheer volume of technical and incidental factors that determine who’s fastest on a given race weekend makes accurate prediction a tall order. The ever-present threats of crashes and mechanical failures make every race a unique proposition, even though most F1 seasons are dominated by a handful of clear favorites. A model that can ingest these factors and render an accurate classification of a driver’s most probable finishing position can certainly improve the fortunes of weekend gamblers - but more importantly, it can help team principles and race engineers tune their race day strategy to account for, and beat, the odds.

My approach to modeling was to treat the final finishing positions of each driver in a given race as a classification problem, where each potential grid position (1-20) was a categorical variable precipitated by factors such as the driver’s team, age, performance through qualifying, performance at the track in past seasons, and more.

2. Data Wrangling

The data utilized comes from the Ergast F1 API, an impressive collection of F1 race, team, and driver data going back to the inception of the Driver’s Championship in 1951.

I queried the Ergast API and concatenated several tables containing data on: race results, including driver finishing times and position, and whether or not the driver experienced a premature race ending event; qualifying performance before the final race; the historical performance of the driver and his team; and number and duration of pit stops for each driver in each race.

Notably absent here is any deep technical data about car construction and configuration that is usually considered very material to race results. For example: we can’t tell from the Ergast

database what Sebastian Vettel's tire compound selection was in Monaco, or how Lewis Hamilton set his brake bias at Silverstone. Future versions of this project may benefit from the inclusion of such technical data.

Some cleaning and feature engineering was required to situate the data in a fashion that made sense for a hypothetical scenario in which one was attempting to predict a race winner after qualifying was complete but before the race began. Specifically, I removed all drivers older than the current oldest active driver, to ensure the model would only predict drivers who could actually win a race in the new season.

Drivers fail to finish races all the time, for many different reasons. In the dataset, non-finishers received NaN values in deveral important features. Dropping these observations entirely would have been detrimental to modeling, since a DNF result is important for gauging driver performance. So I interpolated max or zero values, depending on what was most appropriate, for several features.

In terms of feature engineering, I averaged all pit stops for a given driver in a given race to reduce the pit stop feature set down to a single column that summarized team performance on that front. I also reformatted the Ergast 'time' feature, which presented the winning driver's race time as a datetime-esque string and all others as the number of seconds beyond the winning time it took them to complete the race. By reformatting the winner's time as zero and converting all values to floats, I created a machine-readable feature that I suspected would closely correlate to finishing position.

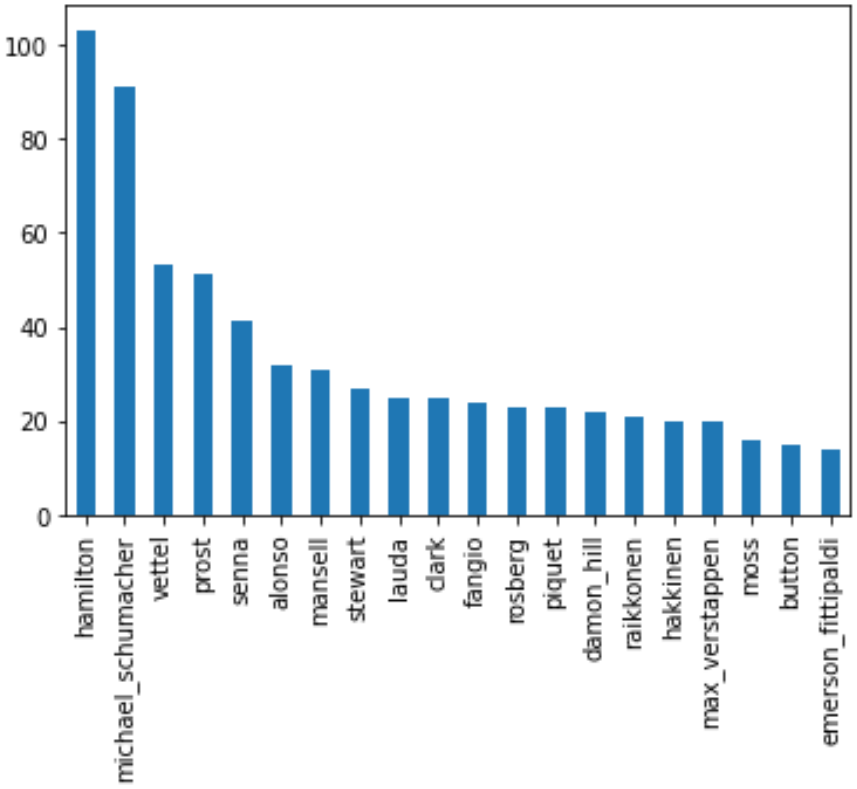
The final shape of my dataset was 4,965 rows x 16 columns prior to encoding categorical variables.

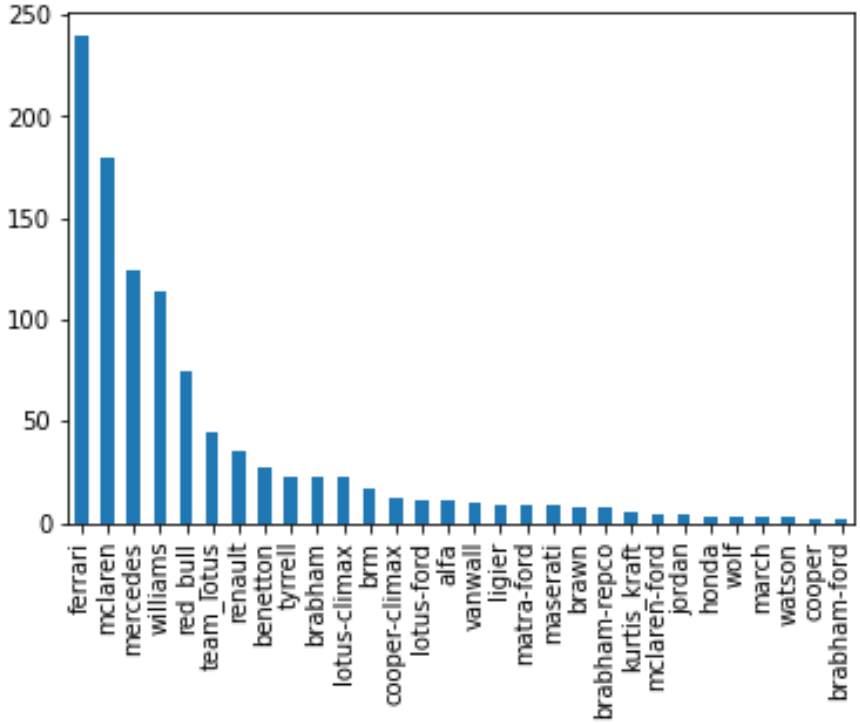
3. Exploratory Data Analysis

If you asked even casual Formula 1 fans which features are most likely to predict race wins, you're likely to get similar answers that all make good, intuitive sense. For example: it's generally accepted that drivers who perform well in qualifying and start high on the grid have a far better chance of winning the race than those at the back. It's also the case, in recent years anyway, that between one and three teams are much stronger than the midfield. If you were to back a driver from one of the favorites, you'd necessarily stand a good chance of successfully backing the winner.

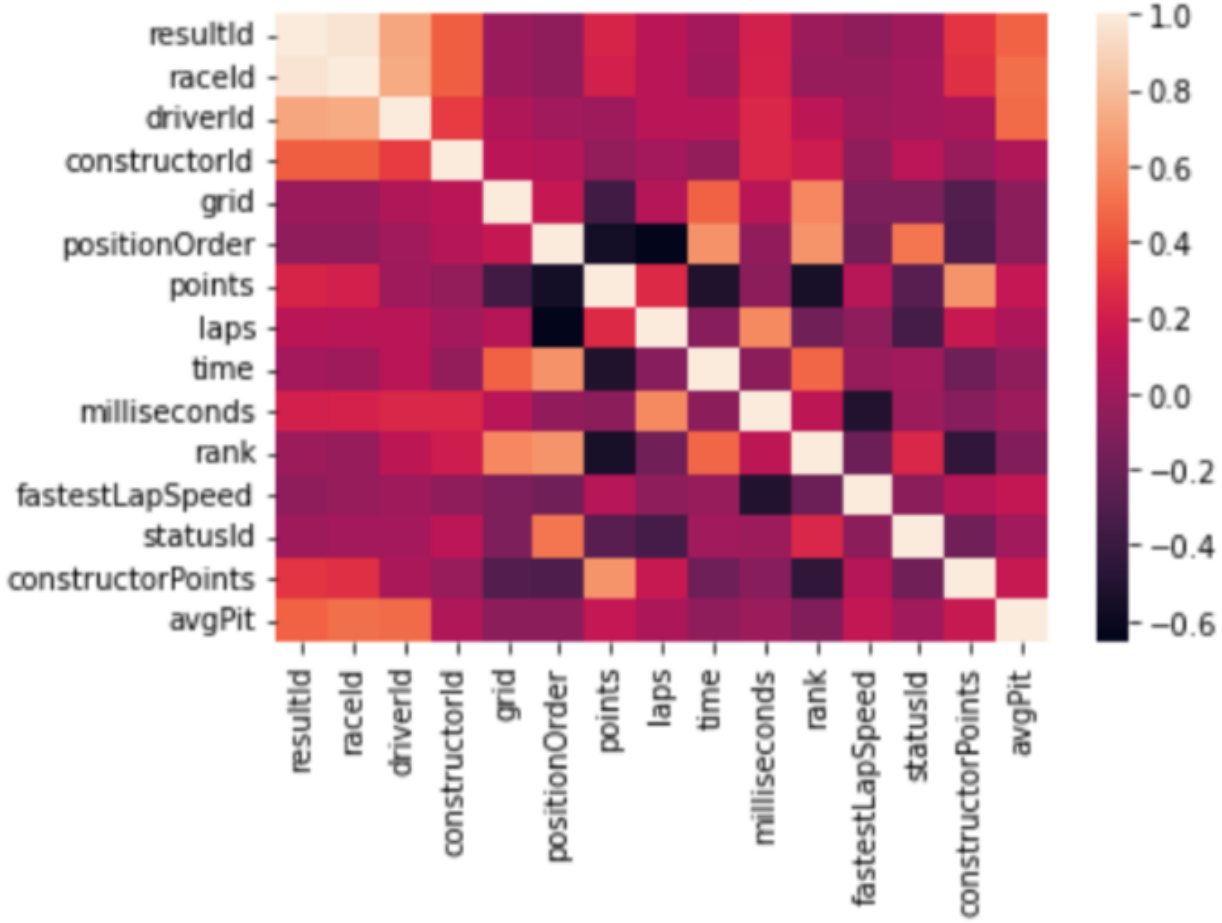
My exploratory data analysis focused on examining which factors seem to be correlated, negatively or positively, with a low finishing position and whether or not these correlations match the conventional wisdom.

First, I compared the number of race wins between drivers and teams to confirm that both are likely to be strong indicators of future performance.


100
80
60
40
20
0
hamilton
michael_schumacher
vettel
prost
senna
alonso
mansell
stewart
lauda
clark
fangio
rosberg
piquet
damon_hill
raikkonen
hakkinen
max_verstappen
moss
button
emerson_fittipaldi


250
200
150
100
50
0
ferrari
mclaren
mercedes
williams
red_bull
team_lotus
renault
benetton
tyrrell
brabham
lotus-climax
brm
cooper-climax
lotus-ford
alfa
vanwall
ligier
matra-ford
maserati
brawn
brabham-repco
kurtis_kraft
mclaren-ford
jordan
honda
wolf
march
watson
cooper
brabham-ford

In addition to supporting that prima facie nation, these visualization also reinforced the need to filter our old drivers and teams. Ayrton Senna in a McLaren may once have been a formidable combination, but it's not applicable in present day.



Based on degree of correlation, it was clear there were some promising relationships between the target variable, position order, and rank, time, driver and team points to date, and starting position after qualifying.

Again, most of these relationships would be fairly intuitive to even a lay fan. The main challenge for this analysis was to discover if the relationships, and their magnitude, could be useful for predictive modeling.

4. Model Selection & Evaluation

I considered three algorithms - Linear Regression (mainly as a hedge against my assumption that classification was better suited to the problem at hand), Random Forest, and Gradient Boosting - and weighed their performance based on accuracy score.

Data was split into train and test groups, with the target feature of finishing position set as the Y feature. After scaling, a Linear Regression and Random Forest classifier were each tuned via grid search to determine optimal parameters. A Gradient Boost classifier was also fit to the train data with a stock set of parameters. All three models were then cross validated on a 5-fold split.

I compared accuracy scores on the test data across all three models to reveal a significant improvement in performance from the Gradient Boost:

	Algorithm	Model accuracy score
0	Logistic Regression	0.351007
1	Random Forest	0.364430
2	Gradient Boost	0.512081

While the best model here significantly outperforms random choice (roughly 1 in 20), it's unclear if an accuracy of 51% on average would outperform a human selecting between *likely winners*, i.e. the top 6 after qualifying, for a given race. A quick review of existing literature shows that a predictive model for Formula 1 winners that outperforms betting odds earned accuracy scores closer to .60.

5. Recommendations & Improvements

It's likely that the gradient boost algorithm utilized in this analysis could improve its accuracy score after more parameter and hyperparameter tuning. Additionally, the modeling exercise may be improved overall by altering the predictive task at hand.

Currently, I task the model with correctly categorizing the finishing position (a discrete category between 1 and ~20) of all drivers. The model doesn't know that only one driver per race can hold the number 1 position. It may also be better to utilize a model that assigns a probability of finishing position 1 to all drivers, and then rank-orders the probabilities from 1 to ~20.