# F1 Race Outcome Prediction

Joshua Lee Nicholas Pfeifer

## Table of contents



## 1 F1 Podium Finisher Prediction

F1 is the pinnacle of motorsport. It features the fastest drivers in the world, driving the fastest cars ever built, on the most storied and beloved tracks around world. Not only is F1 itself a competitive sport, but so is the practice of attempting to predict its associated outcomes. Just like any popular sport, F1 boasts high volume betting markets where highly accurate outcome predictions can lead to a major advantage, and better returns. In addition, F1 officially sponsors fantasy team selection, with some leagues offering cash prizes and other incentives.

Needless to say, there is a lot riding on predicting who wins any given race, or aces qualifying on a given weekend. However, if it were so easy, then everyone would be making money. Although F1 outcomes for the 2022 and 2023 season were "boring" for many viewers (Max Verstappen won 19 out of 21 races in 2023), there was still a great deal of variability in who (else) would end up on the podium. This year takes this type of variability to a new extreme. Although McLaren appears to have the strongest overall package over the course of the season, recent races have marked a resurgence for Ferrari who had won earlier in the season as well. Moreover, Mercedes and Red Bull have also experienced winning runs of 3 or more races over the course of the season.

Despite the high variance of the problem, machine learning models have proved quite effective at modeling the outcome prediction problem.

However, there are several critical issues with these prior approaches:

1. Failure to correctly model point valuations for drivers and constructors
2. Prior methods have not successfully incorporated track features for meaningful track-oriented prediction
3. Data window limitations

Although each of these are important issues, the last is arguably the most significant. In prior work, training data has included the entirety of race results spanning across seasons from 2013 to 2021 (Franssen 2021) and from 2022 to 2023 in the case of our previous work (Lee 2024) on F1 podium prediction. Additionally, (Cheteles 2024) performs evaluations for the entirety of the 2018 season. Of course, these methods have achieved reasonable success at their respective prediction tasks (multi-class prediction and finishing position regression). However, because patterns change so quickly in F1, the underlying driver and constructor performance distribution can change significantly within a relatively short period of time.

For instance, utilizing data from the beginning of 2022 to the end of 2023, our prior research into podium position forecasting placed a significant weight on the Red Bull constructor, and for good reason. From 2022 to 2023, Red Bull won 37 of the 47 races which occurred. As such, training and evaluating on this data (80-20 train test split) resulted in reasonably strong performance and an F1 score of 0.65. Yet, this performance does not carry over into 2024. Because the constructor and driver performance rankings have changed so significantly, it would be fair to say that our methods no longer work at all.

This lack of generalizability is what we aim to address here. We hypothesize that limiting the training data window to the set of $n$ races before the race to be predicted will improve prediction performance on a per-race basis. Additionally, we expect that scaling raw point values heading into a race will assist the model in understanding the points distribution more effectively. This is important since even if a driver has accumulated 75 points over the first three races, their record will appear the same as a driver who has 75 points after 24 races. However, these results are entirely incomparable. The diminished utility of this information is reflected in the fact that our prior modeling reduced the coefficient of this feature to nearly 0. (Lee 2024)

## 1.1 Problem Description

The problem we would like to tackle is the classification of finishing positions for each driver (namely podium or non-podium). In addition, we aim to determine the optimal training size for fitting a model to predict the results of the next race in sequence. In prior analyses, overfitting to the training set has been a common source of error. This is due to how quickly the F1 teams can make significant adjustments to the cars. Many models will struggle to adapt to these changes. By playing around with various training sizes, we will determine the number of

races (leading up to the prediction target race) most useful for fitting a model. The problem here is that of classification and supervised learning.

Additionally, we will explore several modeling methods, in addition to logistic regression. Namely, we will analyze the effectiveness of tree-based methods including decision trees, random forest classifiers, and extreme gradient boosting. One issue uncovered by our prior work was the difficulty of encoding feature interactions. For example, for each track specific variable, such as minimum corner speed, one would need to interact the dummy variables for driver, constructor, or both to generate a meaningful feature. Otherwise, the value would appear exactly the same for all drivers in a race. Because decision trees can implicitly encode some interactions, we may be able to reduce the sparsity of our data construction and reduce the number of interaction encodings required.

## 1.2 Dataset Description

The dataset we use consists of result records for each race, where a result record is given for each driver who participated in a given race. So, for example, if a race has 20 participating drivers, then there will be 20 records associated with that specific event. Our data is collected from F1 seasons spanning from 2012 to 2024 (18 to 24 races per year) and so the total number of data points is roughly 5700 (this will be augmented through our proposed work) with 72 features. These features provide relevant information about each driver's result record - including the circuit where the record comes from, the driver who corresponds to the record, windspeed, precipitation, starting position, and a wide variety of other variables. We also include track specific features such as the mean straight length, the minimum corner speed (how fast do cars go at the slowest corner on the track), the number of slow and fast corners (slow corners involve speeds less than the first quartile for corner speeds across different tracks). We also include data about the prior races from a given season, including the number of previous wins, constructor standing, and driver standing (before the race). These are extremely useful predictors as shown by the prior work (Franssen 2021) and so we include them as well.

For many of these features, we engineer interactions to explicitly encode the relationship between constructor and driver performance, and the characteristics of different tracks. These track differences help to make for more interesting predictions than would be the case if only point tallies and standings were used for podium prediction.

A list of the features we have previously identified as important can be found here

## 1.3 Methods

Once the data has been gathered, we plan on applying synthetic minority oversampling (Chawla et al. 2002) in order to balance the training set. The point of this is to improve the balance of positive and negative samples in the training set since our binary classifier (Podium Finish or not) naturally has many more observations in the non-podium class. From

there, conducting forward selection should be effective in finding the best features for the model. That process would only include data from a specified number of prior races $n$. Then, the predictions can be made on the $n+1$th race. Lastly to optimize $n$, the results can be aggregated over all subsets of $n$ sequential races.

Additionally, we will study several methods to model the prediction problem. First, we will investigate the performance of decision tree and random forest classifiers. Second, we will compare the performance of these models to logistic regression which was used in our prior work (Lee 2024).

## 2 Citations


Chawla, Niteesh V., Kevin V. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. "SMOTE: Synthetic Minority over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–57.

Cheteles, Octavania-Alexandra. 2024. "Feature Importance Versus Feature Selection in Predictive Modeling for Formula 1 Race Standings." University of Twente, The Netherlands.

Franssen, Kike. 2021. "Comparison of Neural Network Architectures in Race Prediction." Master's thesis, School of Humanities; Digital Sciences of Tilburg University.

Lee, Joshua. 2024. "F1-Race-Predictions." https://yoshi234.github.io/f1-fanatasy-analysis/about.html.