Introduction

This report explores the question: what are determinants and predictors of success or improvement in F1 races? The analyses use the F1 race data which includes information on F1 drivers, constructors, and races for the years 1950 to 2017. Our analyses answer 3 core questions:

1. What determines constructors' success?
2. Why do F1 drivers come in second place in races?
3. What are explanations of driver improvement in performance?

We will explore these topics further below through statistical analysis, visualization, inference, and prediction.

Inferring Constructor Success

Initial Thoughts

When thinking about the potential factors that contribute to a constructor's success in F1 races, there are a few ideas that immediately come to mind. When comparing these ideas with the variables available in the dataset, we found the most logical three to be average points in a season, average starting position, and geographical region. All of these variables can be made from the F1 datasets available to us. We started by merging the constructors, constructor results, constructor standings, and races datasets together. We then filtered out any data from years outside of our target range, 1950 to 2010. After this we selected our variables of interest, including the constructor ID as well as our dependent variable, number of wins. We also chose to leave out any observations that contained empty values, to balance the data and lower the risk of faulty results. After this we grouped the data by constructor ID and number of wins, taking the

average points and starting position for each constructor at each number of wins. Finally, to reduce the levels in our nationality variable, we created a continent variable that includes whatever continent the nation is in. Finally, we were ready to run our models.

Model Comparison

Using our three independent variables and our dependent variable, we decided to compare the accuracy of a linear model and a random forest model. When running the linear model, both the points and the position variables are statistically significant at the 0.05 level. The R-squared value is also quite high at 0.81, meaning that about 81% of the variability in the number of wins can be explained by our predictors. We ran the random forest model nine times, with the number of trees being 50, 75, and 100 while the number of variables available for splitting were 1, 2, and 3. When looking at the sensitivity and specificity of these random forest models, we found that they were both maximized when there were 50 trees and 2 variables available for splitting. Next, we created prediction vectors using the linear regression and the best random forest model to predict the number of wins for each observation. Next, we took the average difference between the predicted and actual values and found that the linear model predicts much more accurately. After this we moved into marginal effects.

Most Important Variable and Marginal Effect

Since we decided to use the linear model, we used the t-statistics for each variable to see which one is the most significant contributor to constructor success. We see that the largest explanatory variable in our model is the average number of points. We went on to calculate the marginal effect of the points variable, which shows that constructors are expected to have about two wins when they have fifty points. This is in contrast to when they have five hundred points,

they are expected to record close to eighteen wins. These predictions are made while holding both continent of origin and average starting position constant.

Application into real-life F1 racing

There are many different factors that can contribute to a constructor's success in F1 racing. Through our models and analysis, we have found that the biggest determining factor in constructor success is their average number of points accrued. Starting position is also very important. This means that these are the particular aspects of the races that constructors should try to hone in on and put their effort into in order to maximize their success. Luckily for constructors, it does not seem to matter what country or continent they come from when looking at overall success. Since this is a factor that constructors cannot control, this is important information. We have also determined that these variables account for about 81% of variability in the number of wins for constructors, which shows that these variables truly are the most important predictors of success.

Looking at the marginal effect of the number of points on the number of wins for constructors, the t-statistics and p-values, as well as the R-squared for our model, it makes sense that average points is a huge determining factor in the number of wins that a constructor has. While it may not explain 100% of the variability in number of wins, average points definitely shows to be a very influential variable in the process. We are confident in our final model and we are pleased with our results.

Predicting second place drivers

Since the question was a classification task (predicting a categorical variable - the second place driver), we decided to go with a decision tree for our machine learning model. Decision trees have the advantage of being easy to implement, and also are relatively easy to understand and to explain to others, which is useful in the context of data-driven storytelling.

In selecting the variables that were used to train the model, we first isolated all the variables that could plausibly have a predictive use from our previously combined and merged dataset. This included the variables for raceId, driverId, the driver date of birth, the driver nationality, circuitId, grid, position, and the # of laps that were completed. Using ML flow, then we tested and logged the accuracy of the model when removing one or more parameters. The model that we eventually selected included just the variables for raceId, driverId, driver birthday, and driver nationality. We found that we could remove the other parameters without suffering a loss of accuracy, and decided to do so in order to minimize the risk of overfitting. Our final model resulted in an accuracy of about 67%. Curiously, the model had a near 90% accuracy in the first five years of the test set (2011-2015), about a 50% accuracy in the years from 2016 to 2019, and just a 5% accuracy in 2020. This suggests that there may be other parameters that we could include to try and add more explanatory power in the latter half of the training dataset.

While we did not tackle question 1, the most important variable in our model was, perhaps unsurprisingly, the driverId. Removing the driverId from our model reduced our predictive accuracy by about 12%, which was the largest single predictive power that we found from a single variable.

Explaining Driver Performance Improvement

Based on online posts of what makes great F1 drivers (example), building blocks of excellence comes from drivers' knowledge and teamwork skills. Using the variables that are available in the dataset, we chose the number of wins, points, position in the previous race (created feature), and age (created feature) to explain improvement. The position in the previous race was created to understand where the driver's ranking was in the group of drivers in the dataset. Age was calculated based on birth date and race date. The dependent variable improvement was a binary variable based on whether the driver had a position improvement from the current race compared to their previous race.

We used a linear regression model and a random forest model to analyze the data. The linear regression model had very low predictive power using the variables described above, predicting less than 1% of the variance in improvement. Despite the low R-squared value, age (-0.001***), wins (0.008***), and points (0.0002*) were all statistically significant predictors of improvement. Age and wins are statistically significant with a p-value less than 0.01, and points has a p-value below 0.1. The negative relationship between age and improvement could be interpreted as: with the increase of age, our data predicts less improvement in performance. Another interpretation may be: there are records of older drivers and they only continue racing because they become more consistent in their races (whether that is in a top rank or lower rank). The position relationship between wins/ points and improvement can be interpreted as: drivers who have more wins/ points are predicted to have more improvements.

To see if there is improvement in explanatory power, we used the random forest model to run all of the same variables described above. After tuning, the random forest model was able to achieve an R-squared of 14%, much higher than the linear regression model, but still not a very high prediction of improvement. This could be because of the definition of the dependent

variable as an improvement of position from one race from the previous race. There is likely greater variance between race to race than, for example, aggregating the best performance from each season and comparing those from the previous season. In additional analysis of this data, we would suggest using a broader dependent variable that aggregates the data over a longer period to attempt to lower the variance in the data.