Formula 1 Racing: What Makes A Successful Driver?

Justin Amgott, Oliver Law, Thomas Wallison

December 5th, 2021

1 Introduction

Formula 1 racing (F1) is a class of racing for single-seater cars, each of which must be constructed according to a set of guidelines, or a 'formula.' Each race is made up of 20 drivers with each of the 10 official F1 teams fielding two drivers. Formula 1 vehicles can reach higher speeds than any other regulated vehicle, with max velocities of roughly 215 mph. As such, the piloting of such a vehicle is reserved only for the most elite drivers, and the construction of them is reserved for only the top engineers.

An interesting question that arises when analyzing F1 racing is, 'Does the car have an effect on a driver's placement in a given race, and if so, to what degree?' In a hypothetical setting in which each racer had identical cars, we would expect that the car would not effect the outcome of the race. However, in F1 racing each of the 10 teams fields a different car model each with a different chassis, and engine (some teams use the same engines as they don't design those). With 10 different models of cars on the track, one would assume that differences in quality of car, if indeed such differences exist, would come down to disparities in technological features such as engine and chassis (the frame of the car). It would then be reasonable to infer that these differences may directly affect how well a racer does. In, this project, it is our hope to determine whether or not there is a causal effect of the car on the race. That is, which variables may have more or less significance on the driver's outcome. We are secondarily interested in finding how well F1 racing can be predicted with the somewhat limited data resources available to the layperson.

The machine learning techniques we have learned in class lend themselves well to exploring these questions. We will first apply naive techniques to form a baseline, and we will then apply more complex techniques for comparison.

2 Background and Review

2.1 Overview

Generally, being able to predict the results of a race is of great interest to the sports betting market. To this end, machine learning has been implemented in the past to try to generate predictive models for the winners of various races. These algorithms have shown potential for being able to beat the bookkeepers' odds. However, limited work has been done towards identifying causal factors for the success of one racer over another.

2.2 Prior Literature

One very similar project to ours was done in June 2020 by Veronica Nigro, who used data from many years of F1 racing to develop machine learning models and applied them to the posted betting odds for the 2019 season[2]. They used various models to both arrive at regression based and classification based results for the winner of a particular race. One difficulty the author had was due to the wide stretch of time associated with their data, while the F1 regulations for qualifying have varied over time. Nevertheless, the author determined they could have used their model to make a profit given previous bookkeeper odds, with a 62% prediction accuracy.

Another related project was done in June 2021 by Piotr Borowski and Marcin Maciej Chlebus, who used a variety of machine learning models to try to predict horse racing results in Poland [1]. Previous work on horse racing had determined there was likely an inefficiency in the betting market, meaning effective predictive models could be used to find a winning betting strategy. Their data included information on not just the horses, but also the jockeys and trainers, which had never been done before. The authors found that under certain circumstances they could have a correct bets ratio of 41% for the Win bet, which is choosing the horse in first place, and 36% for the Quinella bet, which is choosing the horses in first and second place in either order, demonstrating the existence of an effective betting technique and reinforcing the market inefficiency claims.

Previous literature is almost completely focused on the predictive ability of models. Some researchers report measures of importance of variables, but this is not the focus of their work [2]. Additionally, their goal is to predict the winner of a race, though the definition of ”winner” depends on the subject. For example, Borowski and Chlebus had two definitions of winner based on the way a particular betting system works, either Win or Quinella bets as described above.

2.3 Predictive vs Causal

Here we detail the differences of predictive and causal models.

A predictive algorithm uses information on past outcomes and attempts to arrive at a prediction for future outcomes. It assumes there is a function of the given variables on the outcome of interest, and tries to estimate that function with various methods

A causal algorithm focuses on particular variables of interest, called ”treatment,” and attempts to isolate the effect of a modification of that treatment on the outcome, holding all else fixed. This effect is typically called the Average Treatment Effect (ATE), which causal algorithms attempt to estimate

The key difference between predictive and causal algorithms is the object of interest: a predictive algorithm attempts to minimize the prediction error of the estimated function, while a causal algorithm tries to isolate the effect of a specific treatment on the outcome.

In previous literature, the algorithms employed were almost entirely predictive. This makes sense, because the question that was posed was generally in regards to betting strategy. In that case, the effectiveness of the model at providing an accurate output is all that matters, since the predicted winner would be where one places their bet. The potential causal factors on success aren't important in this paradigm.

2.4 Aside on Neural Network Models

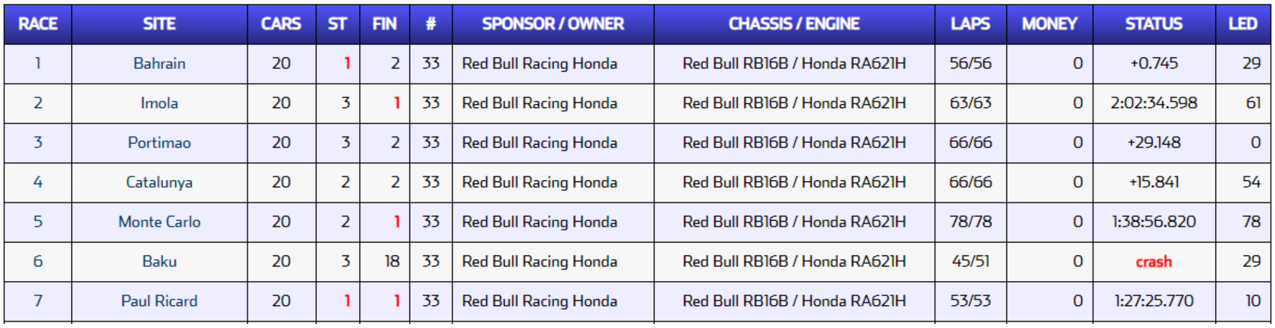
Much of the previous literature has found that Neural Network models are very effective for use within sports prediction, and as such have become very commonly used [3]. Part of the reason for this is the fact that such models can be essentially endlessly iterated on, without having to completely retrain the model on the dataset, since another ”neuron” can be trained at any point.

Though they see a lot of use in these settings, we have chosen to not use Neural Networks in our project. The reason for this is twofold: one is that in similar settings, researchers have found that a neural network model does not have the lowest prediction error [1]. Second, our project doesn't require an iterable model, since we don't plan to introduce new data, nor would a potential use case call for it, unlike the betting models in other projects.

3 Data Description

3.1 Initial Data set

The initial set of data was pulled from the F1 section of `https://www.racing-reference.info`. Its time frame is from 2017-2021 and it includes data for each F1 driver for each race. Although there were more years of available data, there were significant changes to the parameters for car construction in 2017, and we believe these changes may have led to red herrings and false flags. Additionally, we are interested in the modern causal effects of the racecars, which is why we did not look at a set of years under the old rules. A brief snapshot of the data for one driver, Max Verstappen, from the first 7 races of the 2021 season is included below:



RACE	SITE	CARS	ST	FIN	#	SPONSOR / OWNER	CHASSIS / ENGINE	LAPS	MONEY	STATUS	LED
1	Bahrain	20	1	2	33	Red Bull Racing Honda	Red Bull RB16B / Honda RA621H	56/56	0	+0.745	29
2	Imola	20	3	1	33	Red Bull Racing Honda	Red Bull RB16B / Honda RA621H	63/63	0	2:02:34.598	61
3	Portimao	20	3	2	33	Red Bull Racing Honda	Red Bull RB16B / Honda RA621H	66/66	0	+29.148	0
4	Catalunya	20	2	2	33	Red Bull Racing Honda	Red Bull RB16B / Honda RA621H	66/66	0	+15.841	54
5	Monte Carlo	20	2	1	33	Red Bull Racing Honda	Red Bull RB16B / Honda RA621H	78/78	0	1:38:56.820	78
6	Baku	20	3	18	33	Red Bull Racing Honda	Red Bull RB16B / Honda RA621H	45/51	0	crash	29
7	Paul Ricard	20	1	1	33	Red Bull Racing Honda	Red Bull RB16B / Honda RA621H	53/53	0	1:27:25.770	10

This data gave a good foundation to build from, but it did not have everything we wanted, nor did we want everything it had. We proceeded to add and drop several columns in addition to pre-processing the data to get it in an ideal form for our usage.

3.2 Edits to the Initial Data

We started by deleting several columns that we found to be of little use, including the site of the race which seemed to have no impact on how well the drivers did relative to any other locales, the number given to the racecar, the money won in the race (which was 0 in all cases in the initial data set), and the status of the driver when the first place driver finished. This last variable was deleted since it gave time stamps, and we were more interested in the position a racer got than the time they did it in, especially since some races are longer than others. We also decided to delete the LED variable which gave how many laps the racer led the race, because it is not something we would know before the race, and thus is not necessarily useful for prediction. Additionally, we were interested in what could cause a racer to lead for much of the race and win, more so than we were interested in the fact that it happened.
We added a year column for our use during preprocessing, and proceeded to delete it at the end of that process. We did manual research on each of the racers to determine how long each had been driving in F1 and how old each was. We made the decision to ignore drivers who had driven less than five races in each season as they were usually back up drivers with little experience, and we therefore considered them to be outliers. We were also interested in the family wealth of each of the drivers, as becoming an F1 driver has notoriously high costs in its pre-professional phases, and it requires immense wealth or immense sponsorship to pay these costs. Finally, it became necessary to deal with the racing vehicle. We found that the vehicles varied primarily by chassis (the frame of the vehicle) and engine. It was also the case that each team usually only fielded one vehicle, provided to both of its drivers. By far the most substantial differences between any two teams is the car they field, and the drivers they have under contract. Controlling for the drivers, the effect of a car will be uniquely determined by the team at any point in time. Therefore, regardless of the year, we consider a vehicle from a given team to be the same.

We note here that one thing that won't be included in any of our models, due to a lack of available data, is tires. Tires are also beholden to rules and teams must strategically pick tires to use. Each team is allowed to choose from a set of tire types provided by the exclusive F1 manufacturer Pirelli. Since each team is given the same options for tires and the same information about them, we believe that tires will only account for a small portion of the causal effect of a car on placement, and so we think our estimates of the causal effect may still hold weight. It is additionally the case that we were unable to find any sort of practical way to collect data on the whether conditions, which may have effected road speeds in the case of rain, or driver fatigue in the case of heat. That being said, since the whether is more or less random, we don't think this will negatively effect our results in a significant way.

3.3 Variable Summary

Following the edits and preprocessing mentioned in the past subsection, we had the following variables:

NAMES is a factor and represents each racer.

RACE is the number of the race in the context of a given year i.e. 1 is the first race of a given year.

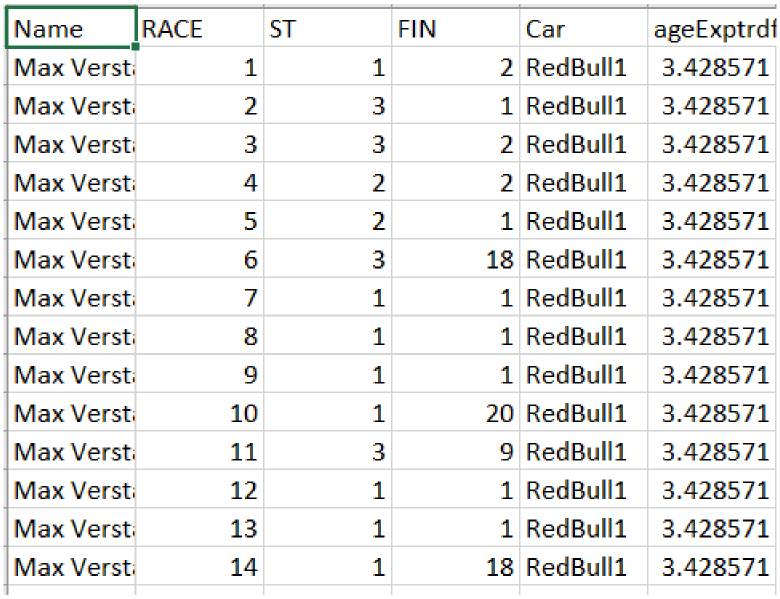
ST indicates the starting position of the given racer.

Exp is the years of experience in F1 the driver has prior to the year of the given race. So if they are a rookie, year is 0.

fWealth is whether or not the driver came from a wealthy family. We were ultimately unable to use this variable in any meaningful way.

ageExptrdf is a measure of the tradeoff we believe exists between age and experience, and is measured by
age / (Exp + 1).

Car is what car the driver is driving, determined by their team (which is the label on the factor level). A brief snapshot of the data after pre-processing is given below, and shows a subset of the data for F1 racer, Max Verstappen in the 2021 season:



Name	RACE	ST	FIN	Car	ageExptrdf
Max Verst	1	1	2	RedBull1	3.428571
Max Verst	2	3	1	RedBull1	3.428571
Max Verst	3	3	2	RedBull1	3.428571
Max Verst	4	2	2	RedBull1	3.428571
Max Verst	5	2	1	RedBull1	3.428571
Max Verst	6	3	18	RedBull1	3.428571
Max Verst	7	1	1	RedBull1	3.428571
Max Verst	8	1	1	RedBull1	3.428571
Max Verst	9	1	1	RedBull1	3.428571
Max Verst	10	1	20	RedBull1	3.428571
Max Verst	11	3	9	RedBull1	3.428571
Max Verst	12	1	1	RedBull1	3.428571
Max Verst	13	1	1	RedBull1	3.428571
Max Verst	14	1	18	RedBull1	3.428571

4 Methods

4.1 A Fortuitous Early Insight

An initial look at placement by team is graphically displayed below:

Right away, we can see that some teams finish better on average. The question becomes whether or not this occurs because the teams have better technology. Fortunately, Red Bull owns two F1 teams, one which is a junior team with less skilled drivers, and one which is a senior team with more skilled drivers. These teams use different cars, and in the 2019 season, Pierre Gasly spent the first 12 races of the season on the senior team and the last 9 on the junior team. Similarly, Alexander Albon spent the first 12 races of the season on the junior team and the next 9 on the senior team. For both drivers, we calculated their average finishing place when they were on the junior and senior teams respectively, in order to estimate the effect of switching cars.

4.2 ML Models Used

4.2.1 Linear Regression

The simplest way to measure a causal effect is through linear regression, although this makes many assumptions about the hypothesis class and the presence of all controls. Nevertheless, we include this naive model as a baseline which we hope to improve upon.

We began by running linear models with the car as the independent variable and finishing position as the dependent variable. We assessed this for confounders and added interaction terms to account for them, and the remainder of the relevant variables to control. We also ran linear models with family wealth as the independent variable (treatment) and finishing position as the dependent variable. And, we similarly improved upon this model.

4.2.2 Generalized Linear Models

We used both LASSO and Ridge regression models to predict the outcomes of the racers.

4.2.3 Double Machine Learning

We ran a double machine learning model to estimate the treatment effect of the car on the driver's performance.

4.2.4 Random Forests

We used random forest models to generate a basic prediction estimate for the finishing position of a driver.

4.2.5 Logistic Regression for Classification

In order to perform classification modeling on our dataset we first were required to convert the finishing position variable to a factor (/categorical) variable. We used a logistic regression model to classify the drivers into either first place finish or non-first place finish. In order to perform this analysis, we were first required to define a new binary factor variable: 1 if the driver won a given race, else 0. We decided to classify for just the winner, because our previous model was significantly better at classifying first position over all other positions.

4.2.6 Instrumental Variables

As we have learned in class, causal effects can be measured utilizing IV techniques under the correct circumstances. Knowing this, we attempted to come up with instruments to use. A particularly interesting one was family wealth, represented by the fWealth variable. As mentioned before, there are incredibly high costs to racing even prior to joining F1, and many, though they may be talented, cannot afford to do it. As a result, many F1 drivers come from quite a bit of wealth, Those that do not are a select few most of who earn sponsorships after demonstrating superb skill and talent. We believed that this followed the exclusion restriction because we did not see any way the family wealth of a driver would effect placement in a race besides through the team and therefore, the car. Furthermore, we did not see any variable that could affect both family wealth and finishing placement, so the IV regression assumptions seemed to be satisfied.

5 Results

5.1 Early Insight

The result of the early insights showed that Pierre Gasly's average placing was 8.5 when he was on the Red Bull senior team for the beginning of the 2019 season, and 10.5 when he was on the Red Bull junior team for the end of the 2019 season. Meanwhile, Alexander Albon's average placing was 6.2 when he was on the senior team for the latter part of the season, and 10.6 when he was on the junior team in the early part of the season. Both drivers did much better in the car provided by the Red Bull senior team that year, which would seem to point to the existence of a causal effect of the car. After all, one would expect that the quality of the driver wouldn't have changed much during the transition to another team.

5.2 Linear Regression

The linear regression model we used included the factor variable for each of the drivers, the factor variable for each of the cars (uniquely determined by team), and several controls ageExptrdf, RACE, and ST. This regression revealed the estimates given in table 1. The coefficients on car can be interpreted causally if we assume that the true model is in fact linear, and that we have all control variables. Under these assumptions, it would seem that some cars such as those used on the Ferrari and Red Bull junior teams, do have a causal effect on the placement of their respective racers, while others do not.

Table 1: Results: Simple Linear Regression on All Variables

	Dependent variable:
	FIN
NameAntonio Giovinazzi	4.264^{***}
NameBrendon Hartley	2.855^{**}
NameCarlos Sainz Jr.	1.417
NameCharles Leclerc	3.175^{**}
NameDaniel Ricciardo	1.022
NameDaniil Kvyat	1.297
NameEsteban Ocon	1.424
NameFelipe Massa	1.075
NameFernando Alonso	2.862^{**}
NameGeorge Russell	3.723^{**}
NameJolyon Palmer	3.429^{**}
NameKevin Magnussen	4.252^{***}
NameKimi Raikkonen	3.001^{**}
NameLance Stroll	2.161^{*}
NameLando Norris	0.508
NameLewis Hamilton	-2.618
NameMax Verstappen	0.079
NameMick Schumacher	6.420^{***}
NameNicholas Latifi	3.633^{**}
NameNico Hulkenberg	3.035^{**}
NameNikita Mazepin	6.420^{***}
NamePascal Werhlein	4.434^{**}
NamePierre Gasly	0.779
NameRobert Kubica	5.077^{***}
NameRomain Grosjean	4.925^{***}
NameSebastian Vettel	2.417^{*}
NameSergey Sirotkin	3.793^{**}
NameSergio Perez	0.332
NameStoffel Vandoorne	2.682^{**}
NameValtteri Bottas	-0.498
NameYuki Tsunoda	1.935
RACE	-0.005
ST	0.330^{***}
CarMercedes	-0.292
CarRacingPoint	0.875
CarMcLaren	0.650
CarFerrari	-2.484^{**}
CarRedBull2	1.852^{**}
CarSauber	-0.024
CarWilliams	1.101
CarRenault	0.886
ageExptrdf	0.033
Constant	4.632^{***}
Observations	1,872
R^2	0.433
Adjusted R^2	0.420
Residual Std. Error	4.397 (df = 1829)
F Statistic	33.203^{***} (df = 42; 1829)
Note:	*p<0.1; **p<0.05; ***p<0.01

We next decided to include an interaction term for the car and the starting position as we identified the starting position to be a confounder in our initial linear regression model. This inclusion lowered the variance of coefficients across the cars: the effect of each car on finishing position was smaller (positive or negative). See table below for estimates:

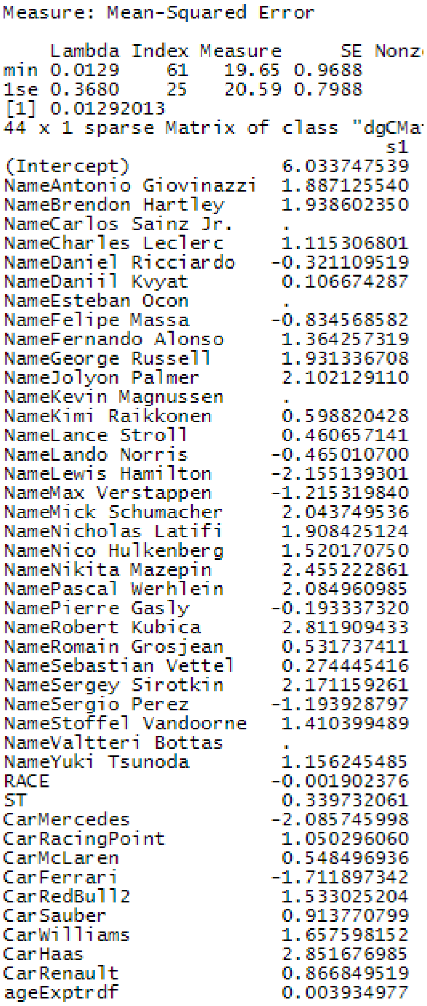
Table 2: Results: Linear Regression Model w/ Interaction Term for Age and Experience and for Car and Starting Position

	Dependent variable:
	FIN
CarMercedes	-4.725*
	(2.491)
ST	0.402**
	(0.190)
Exp	-0.433**
	(0.169)
Age	0.221***
	(0.074)
Exp:Age	0.005
	(0.005)
Constant	37.566
	(160.476)
Observations	1,852
R^2	0.421
Adjusted R^2	0.412
Residual Std. Error	4.425 (df = 1823)
F Statistic	47.394*** (df = 28; 1823)
Note:	*p<0.1; **p<0.05; ***p<0.01

5.3 LASSO Prediction Model For Finishing Position

We implemented a LASSO Prediction model for finishing position using all relevant factors from our dataset. We identified our minimizing lambda to be 0.0129, for which the mean-squared error, estimated through cross validation, was 19.65. Four of the drivers were removed from the model. In the coefficients for the team cars given below, we can see estimates for the treatment effect of each, using the Red Bull senior team car as the control. These estimates are similar to those given in the regression model, and would once again point to a causal effect being present between car and placement.

Lasso Coefficients



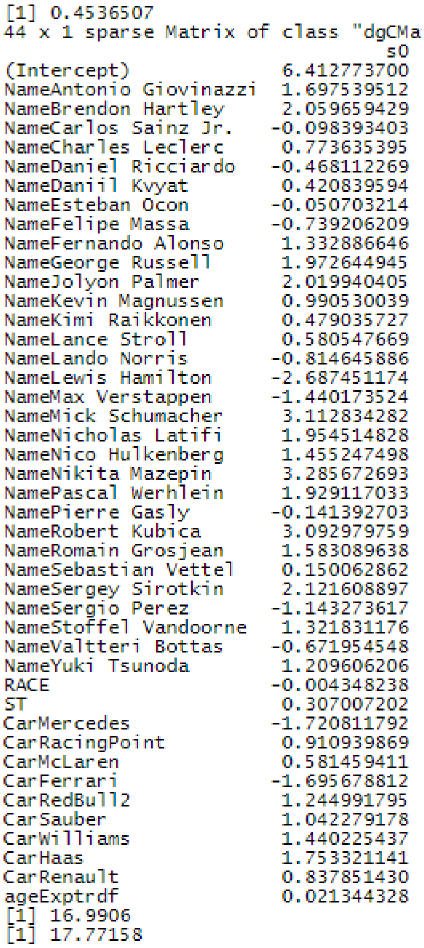
```
Measure: Mean-Squared Error

    Lambda Index Measure     SE Nonz
min 0.0129    61    19.65 0.9688
1se 0.3680    25    20.59 0.7988
[1] 0.01292013
44 x 1 sparse Matrix of class "dgCMa
                                  s1
(Intercept)                6.033747539
NameAntonio Giovinazzi     1.887125540
NameBrendon Hartley        1.938602350
NameCarlos Sainz Jr.       .
NameCharles Leclerc        1.115306801
NameDaniel Ricciardo      -0.321109519
NameDaniil Kvyat           0.106674287
NameEsteban Ocon           .
NameFelipe Massa          -0.834568582
NameFernando Alonso        1.364257319
NameGeorge Russell         1.931336708
NameJolyon Palmer          2.102129110
NameKevin Magnussen        .
NameKimi Raikkonen         0.598820428
NameLance Stroll           0.460657141
NameLando Norris          -0.465010700
NameLewis Hamilton        -2.155139301
NameMax Verstappen        -1.215319840
NameMick Schumacher        2.043749536
NameNicholas Latifi        1.908425124
NameNico Hulkenberg        1.520170750
NameNikita Mazepin         2.455222861
NamePascal Werhlein        2.084960985
NamePierre Gasly          -0.193337320
NameRobert Kubica          2.811909433
NameRomain Grosjean        0.531737411
NameSebastian Vettel       0.274445416
NameSergey Sirotkin        2.171159261
NameSergio Perez          -1.193928797
NameStoffel Vandoorne      1.410399489
NameValtteri Bottas        .
NameYuki Tsunoda           1.156245485
RACE                      -0.001902376
ST                         0.339732061
CarMercedes               -2.085745998
CarRacingPoint             1.050296060
CarMcLaren                 0.548496936
CarFerrari                -1.711897342
CarRedBull2                1.533025204
CarSauber                  0.913770799
CarWilliams                1.657598152
CarHaas                    2.851676985
CarRenault                 0.866849519
ageExptrdf                 0.003934977
```

5.4 Ridge Regression Model for Finishing Position

We implemented a Ridge Regression Prediction model for finishing position using all relevant factors from our dataset. We identified our optimal lambda to be 0.45, for which the mean-squared error reported was 16.99. As we seek to minimize MSE in predictive modeling, this represents an improvement on our Lasso Model at least in terms of prediction. As we would expect, the estimates on the treatment effects for each car are similar, but slightly larger in magnitude than those in the LASSO model. We note that while we have been calling these treatment effects, it is likely they suffer from large biases. Therefore, our best estimate of the treatment effect will likely come through double machine learning. See below for Ridge Regression Coefficients.

Ridge Coefficients



```
[1] 0.4536507
44 x 1 sparse Matrix of class "dgCMa
                                  s0
(Intercept)              6.412773700
NameAntonio Giovinazzi   1.697539512
NameBrendon Hartley      2.059659429
NameCarlos Sainz Jr.    -0.098393403
NameCharles Leclerc      0.773635395
NameDaniel Ricciardo    -0.468112269
NameDaniil Kvyat         0.420839594
NameEsteban Ocon        -0.050703214
NameFelipe Massa        -0.739206209
NameFernando Alonso      1.332886646
NameGeorge Russell       1.972644945
NameJolyon Palmer        2.019940405
NameKevin Magnussen      0.990530039
NameKimi Raikkonen       0.479035727
NameLance Stroll         0.580547669
NameLando Norris        -0.814645886
NameLewis Hamilton      -2.687451174
NameMax Verstappen      -1.440173524
NameMick Schumacher      3.112834282
NameNicholas Latifi      1.954514828
NameNico Hulkenberg      1.455247498
NameNikita Mazepin       3.285672693
NamePascal Werhlein      1.929117033
NamePierre Gasly        -0.141392703
NameRobert Kubica        3.092979759
NameRomain Grosjean      1.583089638
NameSebastian Vettel     0.150062862
NameSergey Sirotkin      2.121608897
NameSergio Perez        -1.143273617
NameStoffel Vandoorne    1.321831176
NameValtteri Bottas     -0.671954548
NameYuki Tsunoda         1.209606206
RACE                    -0.004348238
ST                       0.307007202
CarMercedes             -1.720811792
CarRacingPoint           0.910939869
CarMcLaren               0.581459411
CarFerrari              -1.695678812
CarRedBull2              1.244991795
CarSauber                1.042279178
CarWilliams              1.440225437
CarHaas                  1.753321141
CarRenault               0.837851430
ageExptrdf               0.021344328
[1] 16.9906
[1] 17.77158
```

5.5 Double Machine Learning

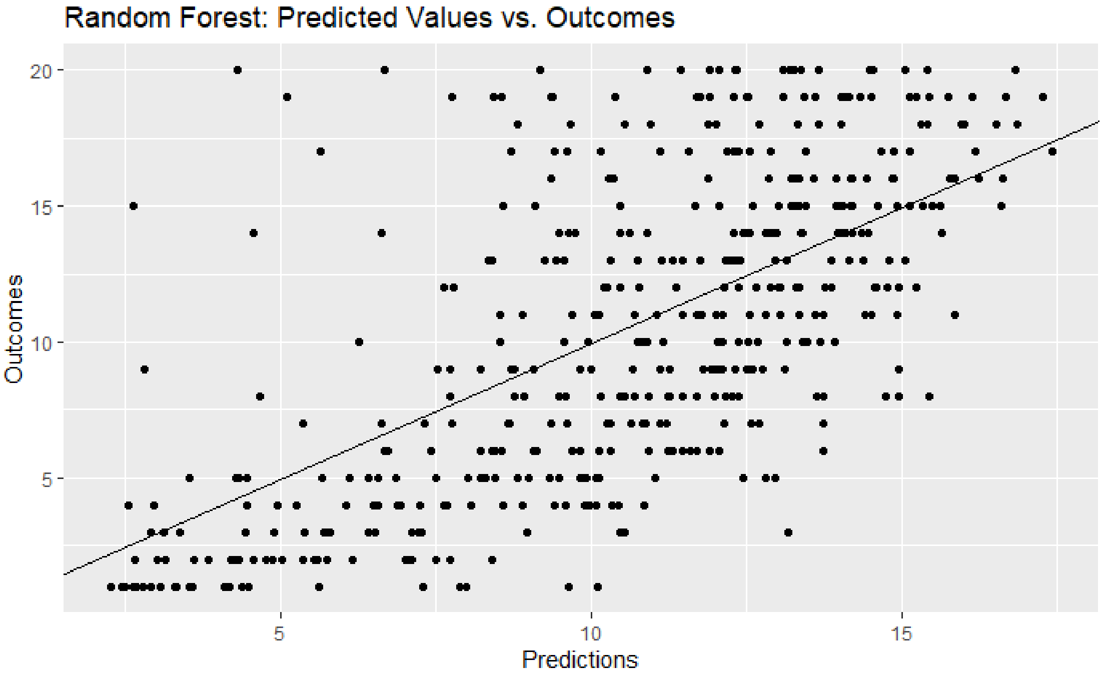
We implemented Double machine learning by first using a random forest model to predict the finishing place using ageExptrdf, RACE, and ST. We then predicted Car using the same three variables, before doing the second machine learning regression step. This final step yielded the causal estimate of car on finishing place, as shown in table 3. This result is not significant, and might lead one to believe that there is no causal effect of car on the finishing place of a driver. We finally note here, that we did assume homogeneity of the treatment effect, that is, we assumed that all drivers would have the same benefit or detriment when switching cars. The car left out of all models is the RedBull1 racing car, and so this serves as the control.

Table 3: Double Machine Learning

	Dependent variable:
	e_y
e_d	0.111 (0.196)
Constant	0.012 (0.104)
Observations	1,872
R^2	0.0002
Adjusted R^2	−0.0004
Residual Std. Error	4.474 (df = 1870)
F Statistic	0.317 (df = 1; 1870)
Note:	*p<0.1; **p<0.05; ***p<0.01

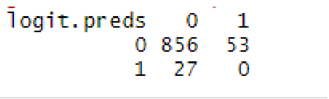
5.6 Random Forest Prediction Model for Finishing Position

We implemented a Random Forest Prediction model for finishing position using all relevant factors from our dataset. We identified our optimal mtry to be 4, for which the mean-squared error reported was 20.59. This indicates an improvement on the Lasso model. However, our model with the lowest MSE is still the Ridge Regression model. See below for a graphical representation of predicted values versus true outcomes:



5.7 Logistic Regression Classification Model for Winner

We reported a cross-validation error of 0.96 from this classification model. This is very high and indicates that our model is very poor at classifying winners. However, our model was strong at predicting non-winners; this is expected as there are 19 positions for which a driver can be classified as a non-winner versus just the single position for a winner. See below for predictions versus outcomes:



```
logit.preds   0   1
          0 856  53
          1  27   0
```

5.8 Instrumental Variables Regression

To estimate the treatment effect, we performed instrumental variables using all of the possible controls at our disposal. This can be seen below in table 4. We note that only two of the cars have coefficients with significance, once again casting some doubt on whether or not the Car has a causal effect on how well a driver places.

Table 4: Instrumental Variables

	Dependent variable:
	FIN
NameAntonio Giovinazzi	4.475***
NameBrendon Hartley	2.887**
NameCarlos Sainz Jr.	1.533
NameCharles Leclerc	3.402**
NameDaniel Ricciardo	0.796
NameDaniil Kvyat	1.267
NameEsteban Ocon	1.553
NameFelipe Massa	0.477
NameFernando Alonso	2.454
NameGeorge Russell	3.749**
NameJolyon Palmer	3.561**
NameKevin Magnussen	4.183***
NameKimi Raikkonen	2.524
NameLance Stroll	2.152
NameLando Norris	0.763
NameLewis Hamilton	−3.000
NameMax Verstappen	0.068
NameMick Schumacher	6.404***
NameNicholas Latifi	3.556**
NameNico Hulkenberg	2.932**
NameNikita Mazepin	6.443***
NamePascal Werhlein	4.708***
NamePierre Gasly	0.853
NameRobert Kubica	4.598**
NameRomain Grosjean	4.631***
NameSebastian Vettel	2.161
NameSergey Sirotkin	3.722**
NameSergio Perez	0.234
NameStoffel Vandoorne	2.985**
NameValtteri Bottas	−0.625
NameYuki Tsunoda	1.904
RACE	−0.004
ST	0.329***
Exp	0.050
CarMercedes	−0.293
CarRacingPoint	0.748
CarMcLaren	0.365
CarFerrari	−2.658***
CarRedBull2	1.804**
CarSauber	−0.264
CarWilliams	1.129
CarRenault	0.773
ageExptrdf	0.044
Constant	4.425***
Observations	1,852
R^2	0.430
Adjusted R^2	0.417
Residual Std. Error	4.408 (df = 1808)
Note:	*p<0.1; **p<0.05; ***p<0.01

6 Conclusion

As the previous literature suggests, predictive models are useful in this setting, and we have found very statistically significant coefficients for some of the variables, suggesting importance of those factors. Our best model was ridge regression, which we know uses all variables. This implies a lack of sparsity of the predictors.

Our causal models were somewhat inconclusive. While many indicators and models seemed to point to their being a causal effect with the cars and the placement of the driver's double machine learning, typically thought of as the best model for finding a causal effect, rejected that one existed.

It is possible that this problem stems from insufficient data on our part. Our dataset contains information on a number of things, but some of the key components such as tires and weather conditions are left out. These may be more important to the model then we thought they were in our first examination of the data. At the same time our treatment of the car is only considering the fact the cars are different, but doesn't consider what makes them different. Numerical measures of that, such as weight or max velocity, could allow for more specific models where the coefficients are more perhaps more interpretable. Unfortunately, data such as this is often unknown with complete precision, and estimated by viewers of the sport.

In practice, if we were to identify causal factors, their use case would be limited. Likely, coaches/engineers on teams would be the only ones who would find it particularly useful for performance enhancement. For a layperson, the predictive algorithms are suitable for their use case, because to them it's not particularly important precisely who wins. They would only need to know who wins for the purposes of a bet.

To build on this work, one would really want to look into much more detailed data on the vehicle specs. This could be difficult to acquire however, as many aspects of the car are kept secret in the case that they give their drivers some sort of competitive edge. It is additionally the case that more details about the drivers would be helpful since the information we used led to inconclusive results. Some drivers seem to predict high placements, but we have not been able to tell all of the exact reasons for this.

References

[1] Piotr Borowski and Marcin Chlebus. "MACHINE LEARNING IN THE PREDICTION OF FLAT HORSE RACING RESULTS IN POLAND". In: (June 2021). DOI: 10.13140/RG.2.2.22254.95043.

[2] Veronica Nigro. *Formula 1 race predictor*. June 2020. URL: https://towardsdatascience.com/formula-1-race-predictor-5d4bfae887da.

[3] Robert P Schumaker. "Data mining the harness track and predicting outcomes". In: *CSUSB ScholarWorks* (2013). URL: https://scholarworks.lib.csusb.edu/jitim/vol22/iss2/6/.

Declaration of Contributions

Justin Amgott:

- Collected F1 data from internet
- Pre-processed said data in R studio
- Assisted with methods and results

Oliver Law:

- Data Modeling
- Methods and Results

Thomas Wallison:

- Past literature collection and review.
- Background section in writeup.
- Writeup formatting.