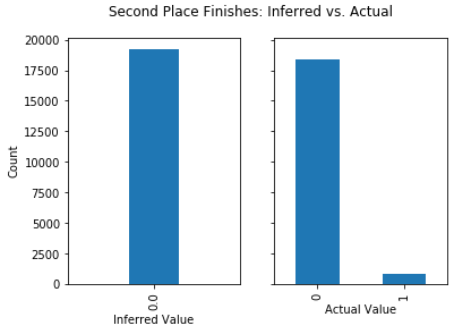
1. Your first task is inferential. You are going to try to explain why a driver arrives in second place in a race between 1950 and 2010. Fit a model using features that make theoretical sense to describe F1 racing between 1950 and 2010. Clean the data, and transform it as necessary, including dealing with missing data. [Remember, this will almost necessarily be an overfit model where variables are selected because they make sense to explain F1 races between 1950 and 2010, and not based on algorithmic feature selection]

From your fitted model:

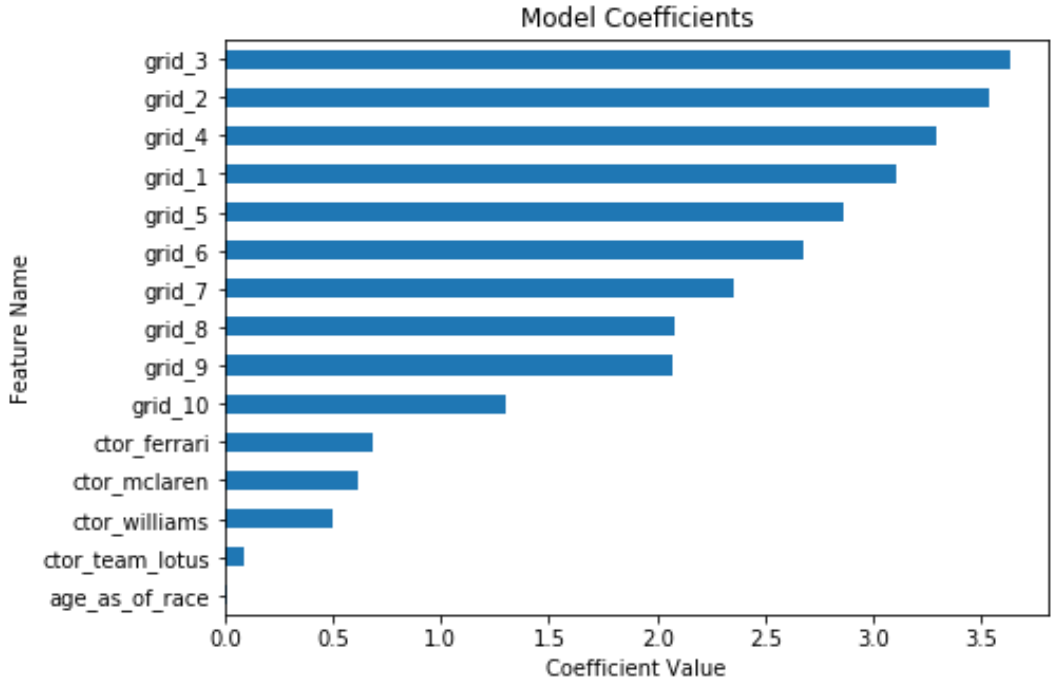
- describe your model, and explain why each feature was selected
- provide statistics that show how well the model fits the data
- what is the most important variable in your model? How did you determine that?
- provide some marginal effects for the variable that you identified as the most important in the model, and interpret it in the context of F1 races: in other words, give us the story that the data is providing you about drivers that come in second place
- does it make sense to think of it as an "explanation" for drivers arriving in second place? or is it simply an association we observe in the data?

To explain why a driver arrives in second place in a race between 1950 and 2010, I used a logistic regression model with a binary “second place” variable as my outcome of interest. In terms of features, I focused on the race of driver (as of the date of a given race), certain constructors (Ferrari, Mclaren, Williams, and Team Lotus), and the first 10 starting grid positions (binary variables for each position). These features were chosen because they could potentially impact a driver’s performance and their likelihood of arriving in second place. The four specific constructors were chosen because they were decently active from 1950 through 2010 and are top constructors in terms of total race wins. Starting positions are important in all sorts of races, so it should not be any different for F1 races. Additionally, I chose to break out the some of the positions individually because I theorized that the 2nd – 4th starting position would be particularly impactful. 1st place might be more indicative of a 1st place finish, but starting positions beyond the 4th position might not be competitive enough for a 2nd place finish.

For model fit, I took a look at model accuracy and what the model predicted. I would argue that accuracy is not the typical metric for explanatory models, but I was interested in seeing it anyway. It turns out that the model ends up labeling every driver as not finishing in second place. This is not surprising given that most drivers do not end up finishing 2nd, so the model predicting everyone as not finishing in 2nd place would be a pretty safe bet.



The most important variable ended up being “grid_3”, a binary variable indicating a 3rd starting position. I determined this based on the coefficients from the logistic regression model. The coefficient for “grid_3” is around 3.6, which means the logit for finishing 2nd goes up by 3.6 on average for a driver starting in the 3rd position compared to driver starting in the 11th position or later, net of other factors. In other words, a driver is more likely to finish 2nd if they are starting in the 3rd position, compared to a driver starting the 11th position or later.



Ultimately, it does not seem to make sense to think of this as an “explanation” for drivers arriving in second place, at least not a complete one. I think that the driver’s constructor and starting grid position are definitely factors for determining whether a driver comes in second place, but there are probably too many variables that are either unobservable or random in the context of F1 races. At the same time, trying to explain 2nd place finishes specifically might be too nuanced or difficult. What exactly separates 2nd place from 1st or 3rd?