## GR5069 - Group Project

The F1 dataset that we have been using during the semester will be the basis for the group project. Your final project consists of answers to three questions below (25pts each) and a GitHub repo per team where all team members commit and collaborate that follows best practices that we have discussed during the semester.

## QUESTIONS TO ANSWER

The F1 dataset contains a number of features available for you to use. Construct your dataset from all available datasets, and select the features that make the most sense to use to answer the questions below.

(1) [25pts] 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?

(2) [25pts] Now we move on to prediction. Fit a model using data from 1950:2010, and predict drivers that come in second place between 2011 and 2017. [Remember, this is a predictive model where variables are selected as the subset that is best at predicting the target variable and not for theoretical reasons. This means that your model should not overfit and most likely be different from the model in (1).]

From your fitted model:

- describe your model, and explain how you selected the features that were selected
- provide statistics that show how good your model is at predicting, and how well it performed predicting second places in races between 2011 and 2017
- the most important variable in (1) is bound to also be included in your predictive model. Provide marginal effects or some metric of importance for this variable and make an explicit comparison of this value with the values that you obtained in (1). How different are they? Why are they different?

(3) [25pts] This task is inferential. You are going to try to explain why a constructor wins a season 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 constructors that win seasons
- does it make sense to think of it as an "explanation" for why a constructor wins a season? or is it simply an association we observe in the data?

(4) [25pts] Back to prediction. Fit a model using data from 1950:2010, and predict constructors success. [Remember, this is a predictive model where variables are selected as the subset that is best at predicting the target variable and not for theoretical reasons. This means that your model should not overfit and most likely be different from the model in (3).]

From your fitted model:

- describe your model, and explain how you selected the features that were selected
- provide statistics that show how good your model is at predicting, and how well it performed predicting constructors success between 2011 and 2017
- the most important variable in (3) is bound to also be included in your predictive model. Provide marginal effects or some metric of importance for this variable and make an explicit comparison of this value with the values that you obtained in (3). How different are they? Why are they different?

(5) [25pts] This task is inferential. You are going to try to explain why a driver's performance improves 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 a drivers performance improvements
- does it make sense to think of it as an "explanation" for why a driver's performance improved? or is it simply an association we observe in the data?

(6) [25pts] Back to prediction. Fit a model using data from 1950:2010, and predict a driver's performance improvements. [Remember, this is a predictive model where variables are selected as the subset that is best at predicting the target variable and not for theoretical reasons. This means that your model should not overfit and most likely be different from the model in (5).]

From your fitted model:

- describe your model, and explain how you selected the features that were selected
- provide statistics that show how good your model is at predicting, and how well it performed predicting a drivers performance improvement between 2011 and 2017
- the most important variable in (5) is bound to also be included in your predictive model. Provide marginal effects or some metric of importance for this variable and make an explicit comparison of this value with the values that you obtained in (5). How different are they? Why are they different?

## WHAT SHOULD YOUR REPO LOOK LIKE

What should be in your repo and your AWS S3 Bucket? [50pts]

# Github Repo- a well-structured project:

project\

|

| -- src

| |-- data <- Code to read/munge raw data.

| |-- features <- Code to transform/append data.

| |-- models <- Code to analyze the data.

| |-- visualizations <- Code to generate visualizations.

|

| -- reports

| |-- documents <- Documents synthesizing the analysis.

| |-- figures <- Images generated by the code.

|

| -- references <- Data dictionaries, explanatory materials.

|

| -- README.md

(Links to an external site.)

<- Project description.

Make sure to have:

- a very informative landing page that guides you through the project

- well structured and modularized code

- well-commented code

- well commented commits

Above all, any person should be able to pick up your project and run it / build on it seamlessly

Also, please provide documentation of your model tracking with screenshots of your model experiments and explaining how you selected your best model as part of your explanation and story

## AWS S3 bucket

- absolutely no raw data in your S3 bucket

- transformed data used in your models on your processed folder