GMMA 860 Team Project Proposal

Team Melbourne – April 2020

Executive Summary

Formula 1 racing is a sport that generates an incredible amount of data around driver performance, car performance and race results. The volume of data is large in both the variables that are measured regularly as well as the time series of data that dates back decades. We believe that an analysis of this data could be used to predict potential winners, or winning race times for future races, which could be used by companies to inform their marketing techniques and/or sponsorship choices. Historical data will be analyzed in a variety of ways in order to make estimates on who might be the most probable winners of the next races (Canadian Grand Prix on June 14, French Grand Prix on June 28 and Austrian Grand Prix on July 5) as well as which car manufacturers are most likely to place.

Analysis

In order to succeed in predicting the top three race times and constructors (manufacturers) for the races we completed 5 pieces of analysis: cleaning data, merging/joining of data, data visualization, feature engineering, and predictive modeling.

1) Cleaning Data

Data cleaning involves 8 steps:

1) Variable Identification – to create data dictionary
2) Data Structuring – each column is a single variable, each observation is a row, and each value has its own cell
3) Missing Value Analysis – understanding what data is missing and the pattern of that missing data.
4) Outlier Analysis – reviewing data that is at least three standard deviations from the mean to understand if it needs to be treated differently
5) Variable Transformation – modifying any necessary variables to capture non-linear patterns or violations of assumptions.

Before we can do any modelling, we must first clean the data to ensure the data is in useable form (Alex Scott & Keith Rogers, 2020). Basically, if we put garbage in, we get garbage out.

We may have a hard time identifying missing data since we are not Formula 1 experts. In addition, we may need data from different data sources to support the model and as a result we may not have access to additional data source to come up with an accurate model.

2) Merging/Joining Data

The objective for this operation is to make necessary data frames available to model and visualize efficiently. The data has multiple data frames, multiple ID (primary keys), and multiple data attributes

Approach

1) Understand all analysis requirements
2) Identify primary keys in each data frame
3) Create separate data frames based on analysis requirement
4) Validate data integrity post join and merge

Some limitation might include ability to create a valid data frame by joining or merging due to missing data entries and ability to do full outer join and right join are not available in SQLDF.

3) Data Visualization

We will use a variety of visualizations to describe the data:

- Time-series line charts: our dataset has data ranging from 1950 to 2020 which we will use to demonstrate trends through the years
- Bar charts: use bar charts to visualize the race rankings for each circuit, each driver and each car maker
- Frequency distributions: to show a proportion of car brands driven per race. We may use a histogram, bar chart or boxplot to help visualize the key statistics and data distribution.
- Scatterplots: we can show correlations between two variables (X = race ranking and Y = car brands, or X = car brands and Y = top three winning racers) to determine if they tend to move in the same or opposite directions.

Data visualization is the art of turning data into graphs, plots, and other graphics. Effective visualization can aid a story, reveal insights about our data, and help us explore particularly large data where summary statistics alone are not enough. (Alex Scott & Keith Rogers, 2020).

Data visualization may lose credibility, accuracy and effectiveness due to the following limitations:

1) Data visualization is more of a descriptive tool not, predicative.
2) Oversimplification of the data may impact our conclusions
3) We may consider some pieces of data important, whereas we may choose to throw out the other pieces entirely. We may not account for all situations by ignoring data outliers or unique situations for further analysis.
4) Visualizations only show a selection of variables from the data set. If we choose the wrong variables to show it can impact how an audience might interpret the results

4) Feature Engineering

Our proposed approach involves the use of indicators, interactions, representation features and external data. An initial review identified the variable “status” as a suitable candidate for developing indicator and representation features. The status variable provides data on the state of finish (or lack of) of a race. It includes 136 unique status descriptions which we propose to manipulate into groups for flexibility in our final model. As an example, we may combine the status data dealing with engine problems, into one group representing a new feature. We could also apply dummy variables to the status, with finish = 1 and the alternatives = 0. We will also identify candidates for feature interactions and external data imports.

Feature engineering is an appropriate model development technique in this context because our raw data doesn't contain all the information required to predict the desired dependent variables - race and or constructor times. However, it does include a variety of independent variables such as, year, time of day and duration - all of which are good candidates for feature engineering.

A major challenge is access to external data. Engine specification data unique to constructors may be confidential. We propose to work around this by identifying and using proxy data as appropriate.

5) Predictive Modeling

We will be using a multiple linear regression model to identify the factors that influence the race times for each driver. Approach:

1) Identify variables that might have missing values (MAR, MNAR, or MCAR) that need to be dropped, or transformed further using imputation, dummy variables, and/or other methods.
2) Build our regression model. Race time (in milliseconds) ~= B0 + B1 (Grid) + B2 (Fastest Lap time) + B2 (Fastest Lap Speed) + B3 (Qualifying Round 1 time) + B4 (Qualifying Round 2 time) + B5 (Qualifying Round 3 time)+ B6 (Pitstop Duration) + B7 (Pitstop #) + B8 (Driver Age) + B9 (Driver Nationality) + B10 (total # of Races competed since 2009) + B11(Constructor)………
3) Assess and refine model through hypothesis tests
4) Remove all variables that should not be included in the model.

There are a lot of values across the data set which could be limiting in our ability to analyze all of them. We will need to do some research on the formula one racing process to understand the real-life implications of how this data could influence racing times. Several trials of manual variable dropping or adding might be needed to fine tune the model and get the highest adjusted R-square value.

References

Alex Scott & Keith Rogers, 2020. From Data to Insight: A Concise Guide to Practical Analytics

Larry Alton, 2016, Data Science Central Blog Post: 4 Potential Problems With Data Visualization