Project Shift, Don't Lift

An investigation into position prediction in Formula 1

Author: Nick Lucido

1.) Problem Statement

This project aims to help Formula 1 Teams make the best data driven decision when looking to hire a driver by using machine learning on historical individual race results to predict finishing position. This would allow teams to objectively rank available drivers and predict where they potentially would finish.

2.) Background

2021 is a big year in Formula 1 with the first-ever set of Financial Regulations in the sport's history. The FIA has imposed a new cost cap to deliver a more competitive championship that reduces to $140m in 2022 and $135m from 2023 onwards. The increase in the number of competitive teams means the increase in the number of competitive drivers.

With the increase in driver competition comes a greater need for teams to be able to objectively select their drivers. As of now, the majority of F1 drivers being selected are from avenues that are "often surprisingly non-scientific[1] " and lack advanced data analytics as seen below :

- 'Pay Drivers' - Drivers that have money and can buy a seat with a team
    - Financially weak teams rely heavy on pay drivers
    - But long-term they won't need to be with the ability to be more competitive and demand more competitive drivers.
- Right Place, Right Time
    - They know the team owner/manager/main sponsor
    - There was nobody else available with F1 experience
- Driver impression in each new championship.
    - How fast can a driver adapt to new car/tracks/team
    - Main metric is % of podium finishes (1-3) against number of races entered.
    - "If a driver had finished second in a championship at the end of their first year that would score them higher than a driver who won the championship at the end of their third year in that championship" - Peter McCool, Head of Design Mercedes Formula E Team

3.) Data

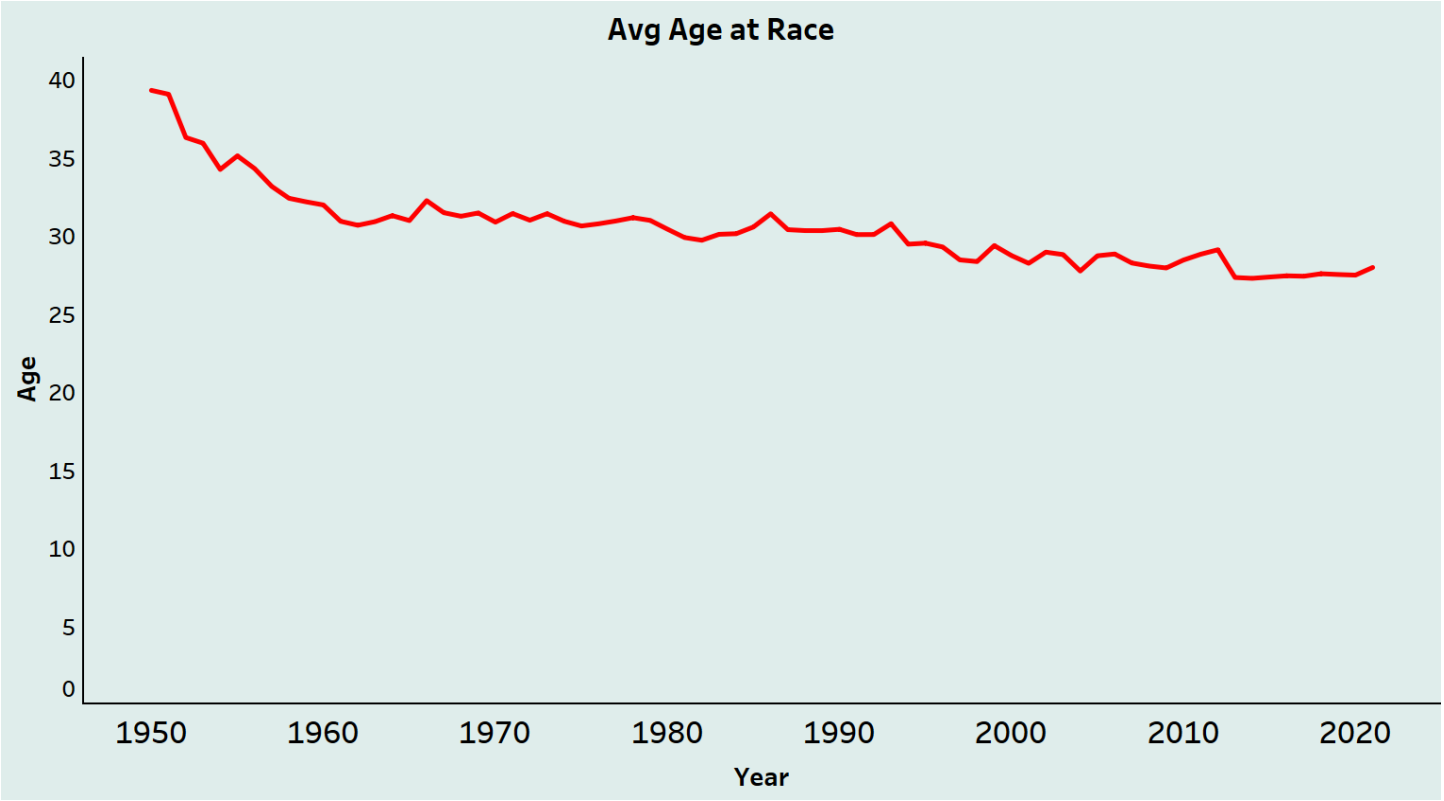
The data used to explore and tackle this problem was discovered in kaggle and consisted of individual datasets containing information on Formula 1 races, drivers, constructors, qualifying, circuits, lap times, pit stops, as well as championships from 1950 till the latest 2021 season. These datasets were all formatted in .csv file types, holding both numeric and categorical values. There were over 25K race results, for 853 unique drivers and 210 unique teams. The lap times data set had over 500K laps of data.

4.) Preprocessing & EDA

Each individual dataset was checked for nulls, duplications, as well **\N** values for what was assumed to be null or missing data points found in certain data sets. From here, certain features were selected from each individual dataset in what eventually became the final data frame. Specific motivations for selecting each feature are recorded in the notebook. There was one particular challenge that arrived from the **grid** feature from the results dataset. It contained a grid spot of "0". This most likely meant a driver was starting at the back for a variety of reasons, but for whatever reason was encoded as '0' for these races. This was going to affect predicting power & correlation as 6% of all grid values were "0" and would skew results by numerically misrepresenting the order of grid position. The solution found the average grid count for these races which contained a "0" grid position, and simply substituted the numbers out. This analysis was executed in the Results - Grid tableau workbook, where it appears the average grid count is 27.64, therefore encode all '0' grid positions as '27.64'. It is noted that this wasn't a perfect solution, however should have provided greater modeling accuracy results compared to not changing at all.

[1] https://the-race.com/formula-e/secrets-of-teams-driver-selection-methods/

From here, EDA was then carried out in an effort to help gain a better understanding of overall progression in F1 as well as identify potential insights between the different variables that might help predict finishing position. An interesting observation was discovered when plotting driver's age at race over time - the average age at each race has slowly decreased from ~39 years of age in 1950 to ~28 in 2021. That's more than a 10 year difference!



This new feature was added to the dataframe in part II of the notebooks in hopes of providing greater predictive power for finishing position.

Feature engineering was then needed to transform all object type data into numeric data, as well as to binarize the target variable to predict finishing in the top 5 positions or not. Top 5 finishing positions were chosen based on insights derived from the average finishing visual from EDA as well as previous research on how teams look for overall impressions into new championships.

5.) Modeling & Results

As mentioned, given this is a binary classification problem (will the driver finish in the top 5 or not?), a LogisticRegression model was the first model of choice with all parameters set to default. The train and test scores were observed based on different scalers.

- Standard Scaler provided highest train score, but lowest test score.
- MinMax Scalar provided smallest difference between scores.
- Robust Scalar provided highest test score with differences in scores between the two scalers.

The final modeling process sought to optimize various scalers, dimensionality reduction, and classifiers. This was achieved by creating and fitting a GridSearchCV model. Below are all models selected with a brief explanation for choosing each model and their unique qualities for tackling this classification problem.

- Logistic Regression
    - Provides probability of an outcome occurring and expands on current model by allowing machine learning to optimize parameters.
- Support vector machines (SVMs)
    - maximizes the distance(margin) between the decision boundary and the closest points from the training data.
- K Nearest Neighbors (KNN)
    - Dealing with multi-class data using KNN is simpler. The decision rule for a given point is just the most common class amongst the K nearest neighbor.

- Decision Tree
    - Learns highly non-linear decision boundaries on multi-class data. Unlike KNN, it is not a distance-based classifier that is constrained to learn using notions of closeness. Instead it works by chaining together simple binary classifiers.

6.) Findings

It appeared that throwing loads of computer power through a GridSearchCV simply resulted in reinforcing that Logistic Regression is still the best model. The Robust Scaler LR model with no dimensionality reduction and default c-value of 1.0 provided slightly better test accuracy compared to the rest.

Model	Scaler	Parameters	Dim_Reduction	Test Accuracy
Logistic Regression	None	C = 1.0 / P=None	None	86.63%
Logistic Regression	Robust	C = 1.0	None	87.03%
GridSearch LR	Robust	C = 0.1	PCA(n_components=10)	86.56%

Things became a bit unclear however when interrupting the predictor's coefficients to help determine their impact on the odds of a driver finishing in the Top 5. A noticeable difference in odds was observed when comparing the different scaler versions of the first LogisticRegression model to each other, ultimately leading to questions about the effects from the scalers.

Another observed difference was penalty of the LR model. By default, penality is 'L2' in sklearn logistic regression model which distorts the value of coefficients (regularization), but if penality='none', the odds ratios slightly increase. Using intuition about the sport, the None-Scaled LR model where penalty = 'None', with no dimensionality reduction, and default c-value of 1.0 model was chosen and it's predictor's odds were used for final insights as seen below.

Model I - Best Results

Model	Scaler	Parameters	Dim_Reduction	Test Accuracy	Precision	Recall
Logistic Regression	None	C = 1.0 / P=None	None	86.63%	71.27%	61.15%

<u>Top Odds Ratios</u>

- **Ferrari** & **Mercedes** drivers are over 2 times more likely to finish in the Top 5 compared to all other drivers racing for different constructors having all other features the same.
- A driver on a **German** or **Austrian** team is about 2 times more likely to finish in the Top 5 compared to all other teams of different nationalities having all other features the same.

7.) The Future

Business Applications

This model would equip F1 racing teams to predict a Top 5 finishing position or not for drivers they are considering hiring during the driver selection process. This would allow teams to objectively rank available drivers in effort to make the best data driven employment decision given the new competitive financial regulations.

Model II

Next steps in the short-term will focus on additional modeling, specifically on the Lap time dataset in effort to engineer lap time features for each race, and then join this to Model I in hopes of increasing predictive power.

The Lap time dataset contains 501,586 laps, but only for years 1996-2021 which only makes up 44% of all races & 16% of all drivers unfortunately (reason for not including in Model 1). An **average lap time** feature is then created and joined to the final data frame that carries out the same modeling process in Model I.

- There were other engineered lap time features such as *Std_Deviation_of_laps* per race and *Fastest_Lap* that were considered, however not implemented due to:

- **Std_Deviation_of_laps:** null values are produced from drivers with 0 completed laps (crash or retire car on opening lap).
- **Fastest_Lap:** assumed to produce high multicollinearity with avg_lap_time.

It appeared again that using loads of computer power via a GridSearch resulted in reinforcing that LogisticRegression is still the best model. This time however, a Standard Scaler version with no dimensionality reduction, a default c-value of 1, and penalty set to none provided highest test accuracy and better precision/recall compared to all models observed in part 6. This model's top odds ratios were also different as seen below.

Model II - Best Results

Model	Scaler	Parameters	Dim_Reduction	Test Accuracy	Precision	Recall
Logistic Regression	None	C = 1.0 / P=None	None	87.01%	74.26%	71.80%

Top Odds Ratios

- A driver who has or shares the most **completed_laps** is over **4** times more likely to finish in the Top 5 compared to all other drivers finishing the race with fewer completed laps and yet having all other features the same.
- A driver with the lowest **avg_lap_time** is over **1.4** times more likely to finish in the Top 5 compared to all other drivers finishing the race with slower avg laps times and yet having all other features the same.

Long-term

The **avg_lap_time** engineered feature ended up becoming a significant predictor in determining Top 5 finishing position. This leads to the assumption that additional lap time features may increase overall predictive power. In the long-term, the goal is to apply this final model with additional lap time features to various motorsports for greater business opportunities.