Introduction

* Sport that everyone knows
* The performance is measured with hard data
* Hard data is created with every single race

Questions to answer

* Main elements are car, pilot and team
* How relevant is the driver vs the car to determine a race output?
* How relevant are the number of pit stops to a race output or winner?
* Are mechanical failures significant to determine a race output or winner?
* How relevant are the number of pole positions to determine a race output or winner?
* Which factors are the determinant to get a F1 race winner?

Overall questions

What is the more relevant factor? CAR, MACHINE

Sergio Gaytan Part:

Hi, Sergio Gaytan here. Regarding technologies used on this project we can find as tools AWS S3 and RDS (Simple Storage Service and Relational Database Service), PG admin, Jupyter Notebook, Google Colab, Tableau and GitHub. The languages used are, Python for data analysis and SQL for ETL process.

For database processing, on ETL process, data was extract from static files that came from Kaggle website; all these files were reviewed to validate they could be stored as a relational database, after this inspection, we had to transform these files to clean them up and, once these files passed the inspection, they were loaded to the final database in AWS.

Additional verifications were needed to validate data accuracy and quantity to perform the complete analysis. Since this is a sport on continuous improvement, data were limited to the last 4 years of championships to have small variation on drivers and same rules.

The complete process followed to integrate the database have 5 steps.

First, 15 csv were extracted from Kaggle website and stored in AWS S3, then, in Google Colab notebook with PySpark, data were extracted from AWS S3 and transformed to clean it for the database integration, in parallel, with PG Admin 4, database schema and relationships between the final tables were created, lastly, data were load into a RDS instance back in AWS. The complete database schema has 16 tables loaded and ready to be further analyzed.

In the preliminary analysis, we define the value from all the data that has the output we need to define a race win which was PositionOrder, this will be one of the main dependent variables used on this analysis that will start with some basic correlation’s validation between our independent variables with the dependent chosen.

Now, Jorge can you please deep dive on the analysis made?

MACHINE LEARNING MODELS

First, to answer the research question a regression model was run to calculate the significance of the variables. Then as the data set have a column result with the final position for every race we use supervised machine learning models to helps us to predict, based on data from previous observation, the possibility of a driver to win a race.

For this project, four different classification model were run, logistic regressions, Support Vector Machines, Random Forest Classifier and Gradient Boosting Classifier. As we can observe Random Forest Classifier and Gradient Boosting Classifier models are the two models with the highest F1-Score result.

MACHINE LEARNING MODELS IMPROVEMENTS

In order to improve the model performance, variables with less importance were removed to find out if some variables could make noise into the model, we pass from seventeen to nine variables. Additionally, considering the imbalance in the classification data set, the SMOTEENN technique was applied to balance the data set. As a result, the F1-Score increased to 0.72 (zero point seventy-two).

RESULTS OF ANALYSIS

To test our hypothesis, some prefabricated sets were built to compare the number of predicted wins with the real wins.

For driver comparison we use different drivers who had at least won one race with data from the car with the greatest number of victories, in other words, different drivers with same car.

On the other side, we perform a constructor comparison to simulate if the most winning driver could have the same perform with different car data. Same driver to test different car data.

The results of the analysis are that drivers with the most winning car, could in average get more victories. Also, a good driver performance could be very affected by the car he is driving. So as a conclusion we can said that the car has more impact than the driver.

NEXT STEPS

The models precision probably could be upgraded if some others relevant variables could be added into the models, variable as aerodynamics, temperature or investments are not considered in the model.

Also, we can perform a Principal Component Analysis, to transform a large set of variables into a smaller one that contains most of the information.

Complementing the results presented by Jorge, we have the next dashboard, here we have some information regarding our two principal characters, the constructor and the

driver.

Regarding the constructor, as you can see, Mercedez, Red bull and Ferrari are our three principal constructors, the three of them are the ones with more points in the competition. Mercedes as the leader.

But the points are not the only aspect that we analyzed, one factor that affect the race results, is the constructor failure.

In this graph Subaru, Mclaren, and Renault, are the constructors with a higher average failure.

Now regarding the driver, our principals competitors are Hamilton, Verstappen and Bottas.

The Average race position for Hamilton is 2.63, between the second and third place, and Verstappen and Bottas have an average position in 5.4, closer to the fifth position.

Now, taking the information regarding the constructor and the driver, we took one constructor, Mercedes, and one driver, Hamilton to analyzed what affect more a race result, the constructor or the driver.

What we wanted to see was how a good driver work in a different constructor and how does a good constructor works with different drivers.

(We run different types of model, linear regression, logistic regression, random forest classifier, etc, being random forest classifier fx4 the best one, with an accuracy of .963, a precision of .576)

The beauty of this is project that we can keep playing with the data, we can use more data, and we can make emphasis in variables as the weather, or more specific, like the car equipment, or change the machine learning and use the unsupervised instead.

And that brings us to the end of our project presentation. We would like to thank you for your time and attention today.