Points Prediction in Formula1 Race

# Aim of the project

This study examines if a driver can earn a point after the race based on his various type of racing statistics. While investigating the dataset, variables of data is visualized with different techniques and statistical test and modelling techniques applied for the related research questions.

# Source of the data

The dataset is taken from the Ergast ([http://ergast.com/mrd/db/)](http://ergast.com/mrd/db/) and is about the Formula1 statistics which includes races, drivers, circuits, constructers, lap times etc. from 1950 to 2022. However, for the sake of this project, statistics from 2010 to 2021 is taken.

# Data description

After removing and tidying the unwanted columns, and creating new column, the data has 4761 observations with 15 variables. In this study, the “points” variable is chosen as the dependent variable.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| raceId | Continuous Variable: Id of the race |
| grid | Categorical Variable: Starting grid position |
| position | Categorical Variable: Official classification, driver position |
| positionText | Categorical Variable: Driver position string |
| positionOrder | Categorical Variable: Driver position for ordering purposes |
| points | Continuous Variable: Driver points for race |
| laps | Continuous Variable: Number of completed laps |
| time | Continuous Variable: Finishing time or gap |
| milliseconds | Continuous Variable: Finishing time in milliseconds |
| fastestLap | Continuous Variable: Lap number of fastest lap |
| rank | Categorical Variable: Fastest lap rank, compared to other drivers |
| fastestLapTime | Continuous Variable: Fastest lap time |
| fastestLapSpeed | Continuous Variable: Fastest lap speed(km/h) |
| StatusId | Continuous Variable: Id of finishing status |
| Pitduration(Created after) | Continuous Variable: Total time spent during pit stops in one race |

# Data Cleaning and Tidying

Graphical user interface, application, table, Excel

Description automatically generatedGraphical user interface, application, table, Excel

Description automatically generatedIn this chapter, the data is cleaned and tidied with different types of procedures to be able to available for data analysis parts.

A picture containing text, orange

Description automatically generated It is seen that from the above graph, the dataset contains NA values. Most of the NA values come from the variable time, then second most is milliseconds. The time variable has nearly 70% of missing data. It can be considered to remove the variable. Yet, the exploratory data analysis was done with these variables where NA values in it.

# Exploratory data analysis

Firstly, descriptive statistics for variables are shown below.

Table

Description automatically generatedFor the continuous variables:

Table

Description automatically generatedFor the categorical variables:

For the categorical variables, it can be seen that there is 24 unique seeds for races and there are missing values in position and positionText.

Chart, histogram

Description automatically generatedChart

Description automatically generatedLet’s visualize the variables and see their distributions.

Chart

Description automatically generated

Diagram

Description automatically generatedFrom looking the graphs above, it can be seen that Pit Duration Times is right skewed, on the other hand, Fastest Lap speed has left skewed distribution. The distribution of Fastest lap variable seems normal but must check with statistical methods. Lastly, for Fastest Lap times, there is outlier values but other than that it looks normally distributed, but again must check with tests.

When we examine the scatter plots of continuous variables, it seems none of them have a strong relationship between each other.

Chart, bar chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedAs a last, let’s look at the categorical variables of dataset.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

From above graphs of categorical variables, there is huge amounts of retired cars, which means the accident, or they had problem with car. Again, for the same reasons, there is huge amounts of withdrew cars too.

With that plots of categorical variables, they were the last plots for univariate examination of dataset. Now let us look the relationship between variables according to research questions with the dependent variable as points.

**Chart, waterfall chart, box and whisker chart

Description automatically generatedDoes number of the fastest lap affect have a point?**

It is seen that from the boxplot above, the ones who earn a point from the race generally have the fastest lap after middle of the race.

**Chart

Description automatically generatedDoes time spent during pitstops has an effect having a point?**

For this violin boxplot, there seems not a huge difference between earning a point from race or not, pitstop wise. It can be said, pitstops has not a huge impact on earning a point. Since the variation of violin boxplot seems good with right skewed shape.

Chart, box and whisker chart

Description automatically generated**What is the relationship between The Fastest Lap Speed and Point?**

From this boxplot, we can conclude that having fastest lap speeds has a positive effect on the earning a point. If the driver can reach more maximum speed than the other, he has more chance to earning a point.

**Does grid position influence earn a point from the race?**

Chart

Description automatically generated

As it can be seen from the graph above, the grid position affects the total laps done by driver which is reasonable. The reason for that when the race finished by the winner of the race, remaining drivers only have to finish the lap they in.

**Dimension Reduction**

Dimension reduction techniques is applied, and result is below.

Chart, histogram

Description automatically generated

From the graph, dimension 8 can be reduced. However, for this sake of report, dimension reduction will not be use.

**Data Manipulation and Feature Engineering**

While this dataset was creating some of the unnecessary columns removed. The dataset has over 50 variables before reduction, after reduction it became 14 variables. To add on that 14 variables, pitstopduration variable created from every race pit stops and summed for this variable.

**Handling With The Missing Values**

The dataset has approximately 8% of Na values after ignoring the least important columns which have over 40 percent of NA values. With using mice packages` multiple imputation algorithm the missing value problem is solved. After the imputation, the characteristics of the dataset remained same.

**Checking the Balance of the Dataset**

Chart

Description automatically generated

From the chart above, the balance of the response variable seems good since their count near to each other. So, there is no need for further work for balancing the dataset.

**Confirmatory Data Analysis for the questions that raised in EDA**

**Does lap number of the fastest lap affect have a point?**

|  |  |
| --- | --- |
| Welch Two Sample t-test | p-value < 2.2e-16 |

Since the p-value is less than 0.05 which says there is a significant relationship.

**Does time spent during pitstops has an effect having a point?**

|  |  |
| --- | --- |
| Welch Two Sample t-test | p-value < 2.2e-16 |

Since the p-value is less than 0.05 which says there is a significant relationship.

**What is the relationship between The Fastest Lap Speed and Point?**

|  |  |
| --- | --- |
| Welch Two Sample t-test | p-value < 2.2e-16 |

Since the p-value is less than 0.05 which says there is a significant relationship.

**Does grid position influence earn a point from the race**

|  |  |
| --- | --- |
| Pearson's Chi-squared test |  |
| p-value < 2.2e-16 | X-squared = 1629.7 |

Since the p-value is less than 0.05 which says there is a significant relationship.

# Modelling

After imputation of missing values and controlling the assumptions, data set is ready for conducting a model to predict if the driver can get a point or not. However, before starting to conduct the models, the dataset divided into two parts which are train and test set with 70% and 30%, respectively. The dimension of train set is 3809 observations with 12 variables, and for test set is 952 observations with 12 variables.

## Logistic Regression

Logistic regression is a statistical model that uses logistic functions in its basic form to model binary dependent variables, but there are many more complex extensions.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | *Coefficients* | | | | | | *Death* | **Estimate** | **Std. Error** | **z value** | **Pr(>|z|)** | | *(Intercept)* | -15.70 | 2.79 | -5.620 | 0.001 | | *grid* | 3.162 | 0.626 | 5.050 | 0.00042 | | *laps* | 0.204 | 0.015 | 13.141 | < 2e-16 | | *time* | -0.006 | 0.0023 | -2.817 | 0.0048 | | *fastestlap* | -0.003 | 0.005 | -0.695 | 0.486 | | *fastestlaptime* | 0.1287 | 0.01005 | 12.805 | < 2e-16 | | *fastestlapspeed* | 0.0404 | 0.0041 | 9.801 | < 2e-16 | | *pitduration* | -0.022 | 0.0021 | -10.787 | < 2e-16 | | *rank* | -4.907 | 0.930584 | -5.273 | 0.0000134 | | Null deviance: 6585.8 on 4760 degrees of freedom / AIC: 3227.2 | | | | | | Residual deviance: 3117.2 on 4706 degrees of freedom / Number of Fisher Scoring iterations: 17 | | | | | |

Table 4 The Results of Logistic Regression

The significance of the model is checked taking into account the difference between zero and the residual deviation, and the model is found to be significant. (P <0.05). It can be seen that from the result table above all the variable are significant expect fastestlap variable.

1. *Support Vector Machine*

Support vector machines are supervised learning algorithms that can be used for both classification and regression problems.

Figure 4 Classification Performance of SVM

1. *Artificial Neural Networks*

Artificial Neural Network is another supervised learning algorithm that can be used for both classification and regression problem.

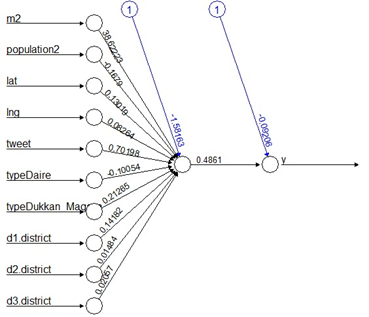


Figure 5 Representation of Neural Network for One Hidden Layer

1. *Random Forests*

Random forest is another supervised learning based on a tree algorithm and can be applied to both regression and classification problems.

Chart

Description automatically generated

1. *XgBoost*

Xgboost is a machine learning method based on tree algorithm and can be used for both regression and classification problems.

# RESULTS

The 6 different model were conducted. Those are Logistic Regression, Support Vector Machine, Artificial Neural Network, Random Forest and XgBoost.

|  |  |  |  |
| --- | --- | --- | --- |
|  | ACC | SN | SP |
| LR | 0.859 | 0.804 | 0.908 |
| SVM | 0.546 | 0.55 | 0.52 |
| ANN | 0.88 | 0.80 | 0.35 |
| RF | 0.997 | 1 | 0 |
| XGBoost | 0.996 | 1 | 0 |

# CONCLUSION

In this paper, exploratory data analysis is performed using a graphical approach and descriptive statistics. Then apply outlier analysis, data cleaning of some spurious variables, and basic data transformations to improve the quality of your data. Then try to predict the target you want using various techniques. The previous chapter shows those results. For drivers to earn point rank, grid, position, fastestlap, fastestlaptime, fastestlapspeed are found to be the most important thing according to the models.