Formula One: Analysis and Predictions

Statistical Learning

G13: Alberico Emanuele, Ludovico Lentini, Camilla Lombardi, Matteo Saba

Università La Sapienza di Roma

Contents

1
Introduction

This project aims at forecasting the finish grid of Formula 1 races. We would like to be able to predict the podiums, the drivers that arrived between the 4th and 10th position, and the 10th and 20th position. Formula One world is as fascinating as complex.

A Formula One weekend is usually scheduled in the following way:

- On Fridays there are two sessions of one hour each of Free Practices in which the drivers ”learn” the circuit and try all the customizing elements (tyres, balance, wing) of the car ahead of the race.
- On Saturdays, we have the third and last Free Practice Session in the morning and the Qualfyings in the afternoon. This kind of qualifying was introduced in 2006 and is known as ”knock-out” qualifying.It is split into three periods, known as Q1, Q2, and Q3. In each period, drivers run qualifying laps to attempt to advance to the next period, with the slowest drivers being ”knocked out” of qualification (but not necessarily the race) at the end of the period and their grid positions set within the rearmost five based on their best lap times. After each period, all times are reset, and only a driver’s fastest lap in that period (barring infractions) counts. The number of cars eliminated in each period is dependent on the total number of cars entered into the championship.[59] Currently, with 20 cars, Q1 runs for 18 minutes, and eliminates the slowest five drivers. In Q2, the 15 remaining drivers have 15 minutes to set one of the ten fastest times and proceed to the next period. Finally, Q3 lasts 12 minutes and sees the remaining ten drivers decide the first ten grid positions. Until 2022 each car had to start the race with the tyre used to qualify in Q3 but from season 2022 drivers are given free choice of tyre to use at the start of the race. In 2021 a new type of qualifying known as ”Sprint Qualify” was experimented. The sprint change all the order of the weekend. On Friday morning we have a Free Practice Session, while in the afternoon we have the usual qualifying. On Saturday morning we have another Free Practice Session, while in the afternoon a mini-race takes place with the starting grid obtained on the friday afternoon. The race starts and the ending grid of it will be the starting grid of the official race of the Sunday.
- On Sunday, the race takes place.

FIA, the *International Automobile Federation*, is the association, established in 1904, which represents the interests of motoring organisations and motor car users and it is the governing body for auto racing events, including Formula One. In Formula One things change as fast as Formula One cars are. Many of these changes are driven by technical evolutions, so in average every five years we have a new ”Formula One Era” with new types of engine requested from FIA to partecipate. However, innovations are not limited to engines, but also to fuels, tyres, car design which includes aerodynamics, or electronic aids, or moreover the introduction of new rules regarding the general aspects. Of course, these innovations together with the business aspect, leads FIA to change rules very quickly. For example, in 2021 a new type of qualifying was introduced in order to involve more the audience from the Fridays, which is the least attendended day. In 2022 they modified this experimentation changing the number of points gained in the mini-race or the fact that the ”Pole Man” is the Fridays’ one and not the Saturday one.

For this reason, it’s not easy to implement an analysis in which the same assumptions are valid over the whole period. Hence, we decided to focus on a fixed timeframe 2014-2021 in which the same engine was maintained during the years and it is not too ”short”.

2
Data Collection

The data used in this project was gathered from an online database which stored all the information regarding Formula 1, from 1950 to 2022, and it is updated after each race.
The data is stored in 14 different tables, each containing specific information:

- **Circuits:** A table containing features of all the circuits in which a Formula 1 race took place such as the name of the circuit, the geographic coordinates, the country and the location;
- **Constructor Results:** A table containing the results regarding the teams for each races to which they participated. For each team we can find the position and the points gained;
- **Constructors:** A table containing the information of each team such as their nationality, their ID, names and number of the car;
- **Constructor Standings:** A table containing features of all the circuits in which a Formula 1 race took place;
- **Drivers:** A table containing socio-demographic features associated to every driver which partecipated at least at one race;
- **Driver Standings:** A table containing data regardind the drivers' standings;
- **Lap Times:** A table storing for each race, for each driver and for each lap, the time in which the lap was completed;
- **Pit Stops:** A table containing for each race and for each driver, the number of pit stops done, the laps in which they occured, the total time it tooked to pit and the position in which the driver was in the race;
- **Qualifying:** A table containing data regarding the qualifyings, so for each race and for each driver we have the qualifying ID, the starting grid for the race as the final positions obtained during qualifyings, the detail of the lap times in Q3, Q2 and finally in Q1;
- **Races:** A table containing all the features of the race sucg as the race ID, the date and time, the circuit and there are also the results of all the sessions of Free Practices and of the Sprint Qualifyings;
- **Results:** A table storing the results of each race, in particular we have detailed data for each driver and for each team regarding the position in which they finished, the status which is a code that can assume different values representing the status of the driver at the end of the race, the points that the driver gained, the fastest lap, the speed of the fastest lap;
- **Season:** A table containing just the year of the season;
- **Sprint Results:** A table containing the results of the Sprint Qualifying which was introduced in 2021. Also in this case for each driver, team and race we have information regarding the starting position, the final position, the points gained, laps, time, fastest lap, the status;
- **Status:** A table which explains each status ID;

2.1. Data Transformation

Once the tables were downloaded, we started manipulating the data on SQL in order to obtain a complete dataset with all the useful information for our analysis.

The starting point in SQL was to perform some exploratory queries in order to understand which kind of data was stored in each table, identify the useful tables and drop the unncessary ones.
In order to be coherent with the goal of our project, not all these tables were useful. This for two main reasons. The first one is that, as already described before, rules in Formula 1 change very quickly and data not coherent with the period considered may affect the results of the predictions. The second one is that some of these tables simply don't contain useful data.
We decided to drop the season table since it didn't store any useful data and we also decided to drop the Sprint Results table. This choice was driven by the fact that the Sprint Qualifying was only introduced in 2021 and this type of qualifying was used only in three races (Silverstone, Monza, Rio).

Once defined the tables and the field we were interested in, we procedeed with the creation of the EER Diagram to work on, shown in *Figure 1*. The considered the race as the main table and joining all the other tables with the unique keys. As a result we obtain a dataset with 20 rows, as the number of drivers which takes part to each race, for each race, each one representing a driver with all the associated features.

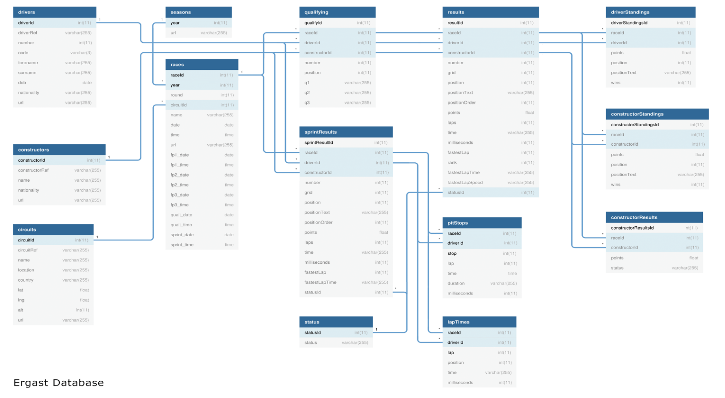

Figure 1: EER Diagram

As cited before, the dataset was filtered by the season, since we are just considering the "F1 era" in which the same rules were applied. For this reason, the dataset was composed of every race ran between 2014 and 2021. We will use as a training set the years between 2014 and 2020, and as a test set 2021 season. Finally, we had a dataset composed of 3267 observations and 38 variables.

2.2. Data Manipulating

Our analysis continued by extracting the data from SQL and importing it into R tool. In this phase we performed an exploratory data analysis and implemented some customization that we considered important.
First of all, we focused on drivers and teams. Of course we observed that not all drivers were present during the considered timeframe and that the same driver could have changed team during this period. The same observation can be done for the teams: some of the constructors changed their name over the years. We decided to manage these teams by replacing old names with their current name. In particular:

- Racing Point and Force India became Aston Martin
- Sauber became Alfa Romeo
- Lotus and Renault became Alpine
- Toro Rosso became AlphaTauri

We then focused on some features that could be useful for prediction. We introduced some new variables such as the percentage of wins per driver and not finish races per driver and per team. We constructed the percentage of wins as the number of wins per driver over the total number of races for that driver. This decision was made because we thought that the experience of a driver could have an impact on its performance. Instead, we implemented the percentage of not finished race per driver only considering the races in the time frame analysed. This decision was driven by the assumption that in a specific time period there is a particular engine, and it won't be useful to use previous data collected on old engines. Of course, both of the new variables are continously updated race per race.

The creation of the not finish ratio per team has been much more complicated since in the same race there are two drivers per Team, where potentially both could ritire and not finish the race. For this reason we are considering one race of the team as two, because two drivers are racing. This new measure updates race in race, but of course if one of the cars retires duting a race, this would not impact the race of the other driver of the same team.

Finally, the dataset is filtered by the drivers, the teams and the circuits that are active in 2021, dropping all the other observations since they are not relevant for the classification. We also consider only the features that could be useful for our predictions and dropping the useless ones.

We end the data manipulating process with a dataset composed of 1462 observations and 17 features.

3

Analysis and Modeling

3.1. Descriptive Analysis

We proceeded by analyzing the final dataset we are going to use for the predictions.

We were interested in some descriptive analysis, the ration of wins, and not finish races per team, per driver and over the years. In particular, we began by visualizating the winners in between 2014 and 2021, both per team (*Figure 2*) and per driver (*Figure 3*).

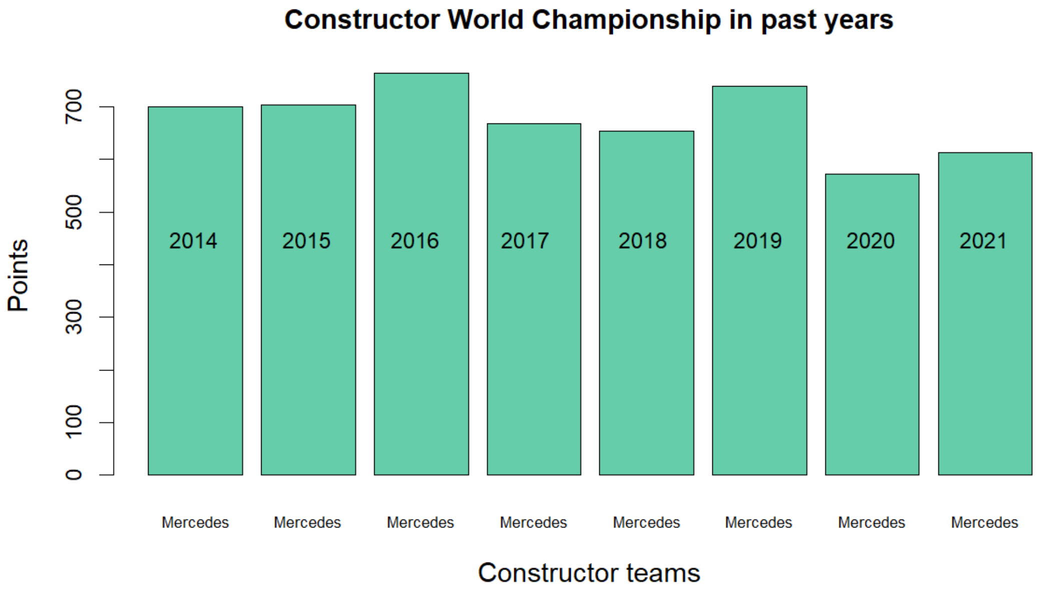

Figure 2: Formula One World Title Team Winners in past years

From the above plot is clear that all the Team championship over the period was won by Mercedes Team..nothing more to say.

We were then interested in comparing this team results with driver results, to check whether the Team Championship is related to the drivers one. We actually expect that the Word Title Driver winners are drivers of the same team which won the Team Championship.

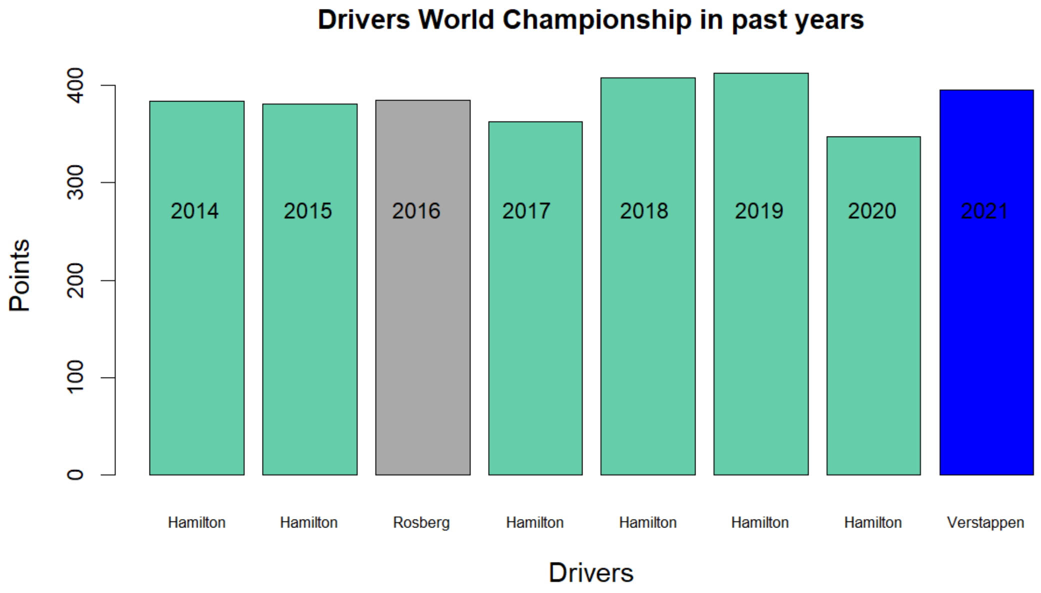

Figure 3: Formula One World Title Winners in past years

Analyzing the plot we observe that what we expected, actually happens. Infact; in between 2014 and 2020 the winner of the World Title is a Mercedes driver! The most successful is Hamilton. In 2021 we have the only expectation: a RedBull driver won the driver Championship while Mercedes won the Constructor Championship.

We were then interested in the average points per race scored by the top 3 teams in Formula One, that are Mercedes, Ferrari and RedBull.

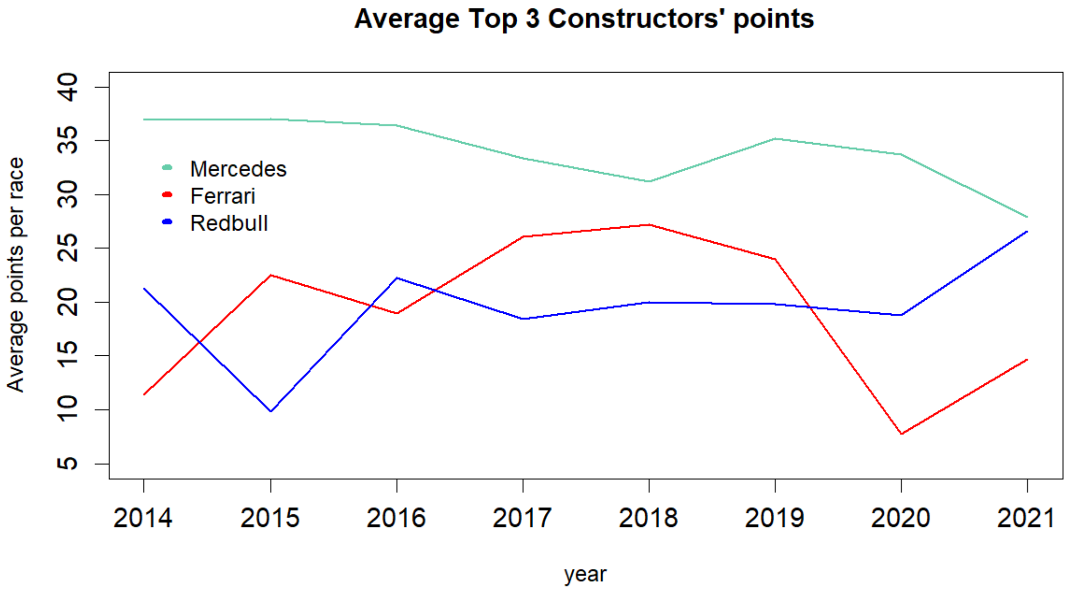

Figure 4: Race points average between 2014-2021

In *Figure 4* we plot the average points per race scored by the two drivers of the same team. We observe that Mercedes has an extraordinary race points average and it's constant over all the period. Ferrari and RedBull instead, have a very fluctuating trend. RedBull, however, after the first years, managed to import developments which improved the car perfomance. On the other hand, Ferrari was able to introduce great innovations between 2017 and 2019, but it was then discovered taht some components of Ferrari's car didn't completely respect the rules, and they had to change them. As we see from the figure, this actually penalized the performance of the car with a worsening of the results.

These three figures highlight a Mercedes supremacy. This is in line with the assumption that we made at the beginning of the project. Infact, in this period we had new rules for the engine and Mercedes was the team that better developed a performing car which allowed them to win everything. To add a note to this point, in 2022, year in which the engine and the car structure were changed by the FIA, Mercedes it is not a leading team in the championship; while Ferrari,has reversed its situation by producing a quite performative engine.

We then focused on some of the measures we implemented and decided to visualize them in order to understand their trend over all the period.

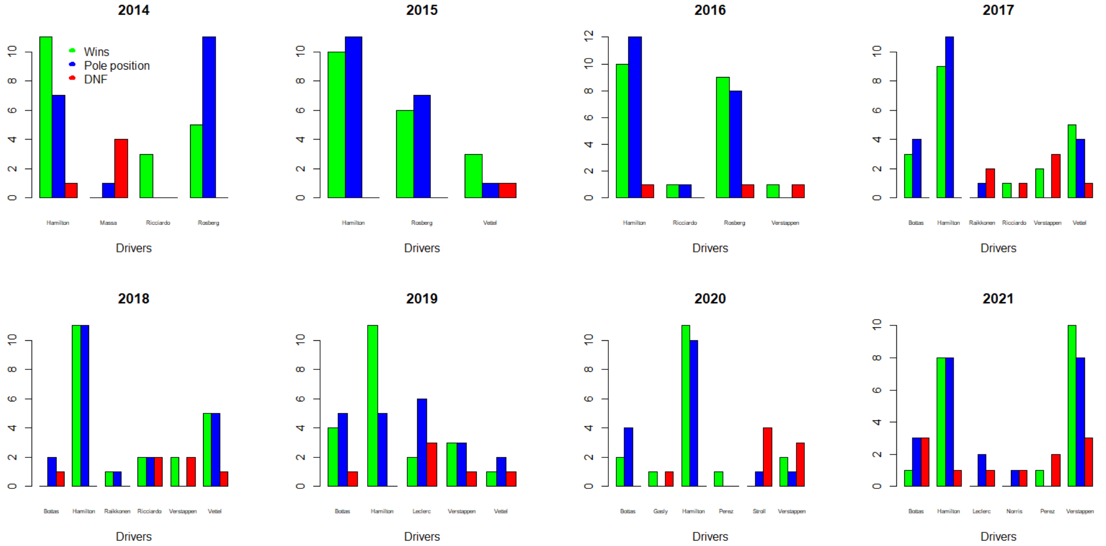

Figure 5: Wins, Pole positions, and DNF details per Driver

In the above plots we represent the number of wins, pole positions and DNF gained for the most successful drivers in the specific year. In particular, we show all the drviers which have at least one win, or won pole or DNF. It seems incredible that in 2015 just the two drivers of Mercedes obtain most of the wins and pole position. This trend is maintained over all the years analyzed. The Mercedes drivers in this year are Rosberg and Hamilton and Bottas and Hamilton. We can observe that most of the wins, and pole position were of course obtained by Mercedes' drivers such as Hamilton and Rosberg before, and Hamilton and Bottas after. A nice observation to make is that Hamilton in 2018 obtain as many wins as poleposition which of course does not mean that every pole position was transformed in a win on the sunday. Of course, it is also necessary to highlight the performance of Verstappen in 2021, which was able to win the Championship even with more DNF than Hamilton.

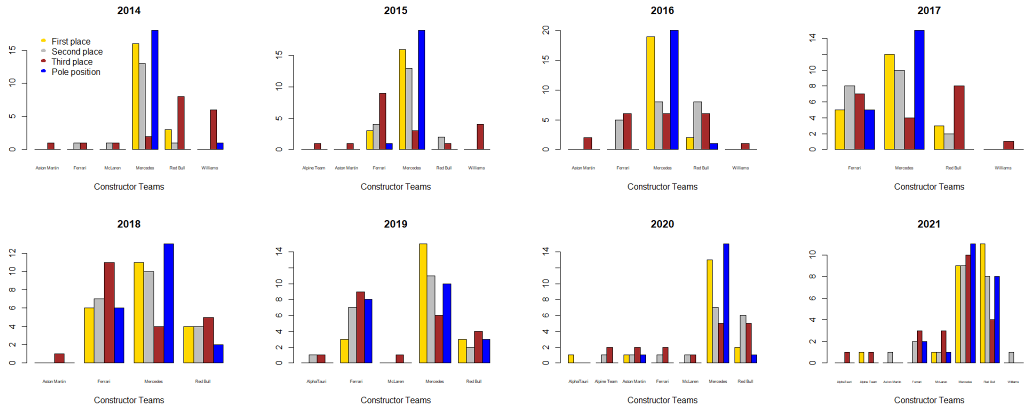

Figure 6: Podium positions and Pole positions details per Team

In *Figure 6* we plot the number of first, second and third position, the number of pole position. Of course, over the years the trend is the same showed in the previous plots. It is easy to observe that Mercedes team is actually the leading team over all the years analyzed having a huge number of wins and pole position, and a very low number of DNF with respect to all the other teams.

3.2. Data Modeling

The aim of our project is to predict whether the driver has finished the race on podium, between 4-10 position, or between 10th and 20th.To do so, we converted the positionText variable in such a way that it assumes 3 values: 1 for podium, 2 for position 4-10, 3 for position 10-20 and DNFs.

We selected the following independent variables as regressors:

- Driver
- Qualifying position
- Seconds of the fastest lap
- Speed of the fastest lap
- Team
- The circuit where the race took place
- Win percentage per driver
- DNF percentage per driver
- DNF percentage per team

To perfrom our prediction we implemented four different models: Ordered Logistic Regression, Random Forest, Naive Bayes Classifier and Support Vector Machine.

We then perfomed a test to check the correlation between the independent variables chosen.

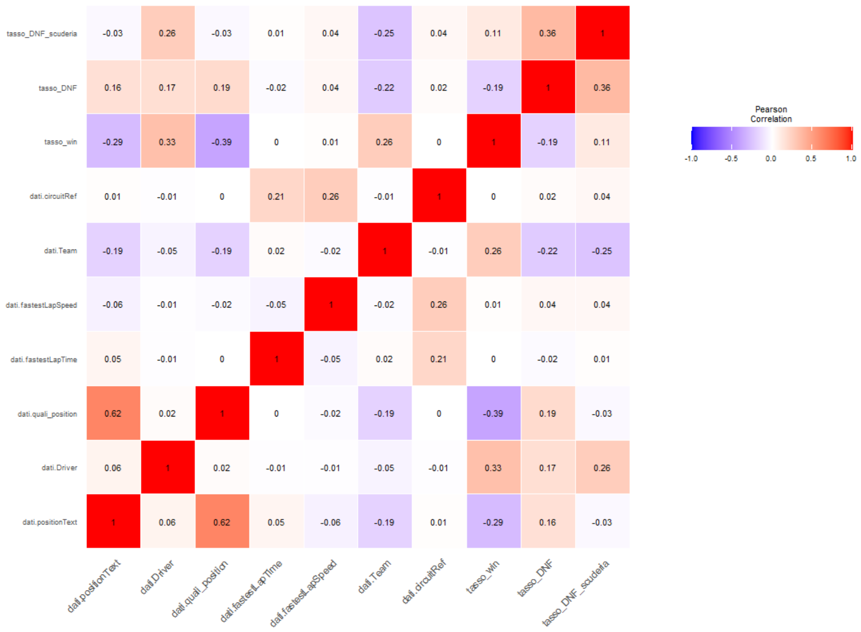

Figure 7: Correlation heatmap of independent variables

In *Figure 7* we represent a correlation heatmap between the independent variables in order to check whether there are some dependencies. From the figure we actually notice that there aren't any kind of relationship between these variables and we can proceed with the implementation of the models.

3.2.1. Ordered Logistic Regression

The Ordered Logistic Regression is the first model we applied since it is the one that best meets our needs. The assumptions behind this model are:

- The dependent target variable has to be ordered
- One or more of the independent variables are either continuous, categorical or ordinal
- No multi-collinearity.

All the assumptions are verified since our target variable is ordered: it assumes values 1,2 or 3 based on the positions classes gained during the race. We have a ordered variable which is the Qualifying position and there is no multi-collinearity as checked through the correlation heatmap.

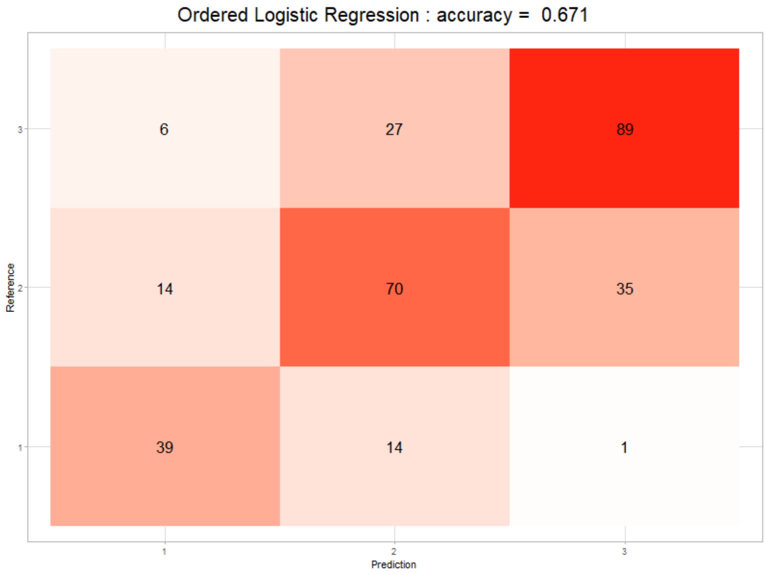

Figure 8: Confusion Matrix and Accuracy for the Ordered Logistic Regression

The ordered logistic regression accuracy is equal to 0.671, which is not an optimal value, but at the same time it's not so bad. From the confusion matrix we observe that the model actually performs well in predicting the exact classes, and at the same time it doesn't make many mistake when predicting the two most distant classes (e.g. 1-3, 3-1). However, there is a consistent number of mistake when predicting the class '2' with respect to the truth value of class '3'.

3.2.2. Random Forest

A Random Forest Classifier is an algorithm that with an ensemble technique builds different decision trees on bootstrapped data observation and a subset of features. Each model is trained independently and generates a result: then the final output is based on majority voting after combining the results of all models.

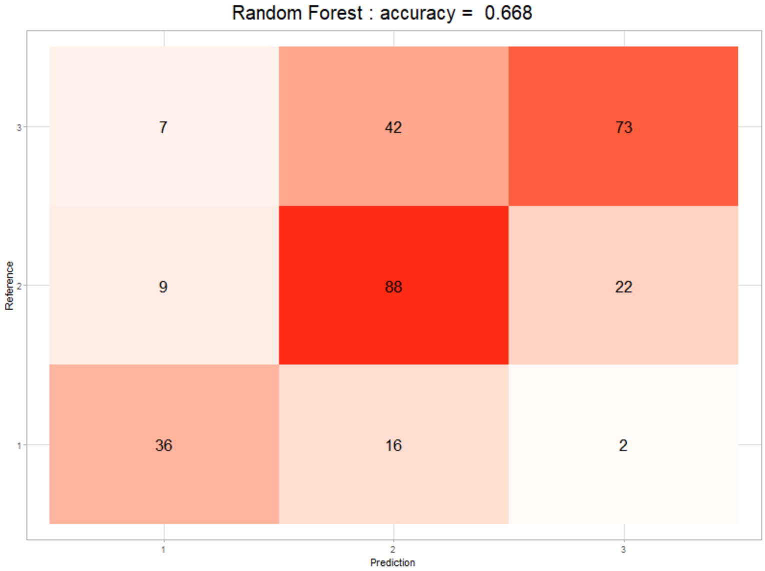

Figure 9: Confusion Matrix and Accuracy for the Random Forest

The random forest model performs similarly to the ordered logistic regression, even though it has a slightly lower accuracy. It is very performing when classifying the second class. It slightly increases the number of miss classified labels, especially the ones we would like to avoid (eg. 1-3,3-1) but also the label for which the true value is 3 and it predicts 2. Overall, also this model is not so bad.

3.2.3. Naive Bayes Classifier

The Naive Bayes Classifier is an algorithm based on Bayes' Theorem with strong hypothesis of independence between the independent variables. It is used for classification tasks, both binary and multiclass. In our analysis, it is quite useful since the independent variables don't have strong relationships, as shown by the correlation heatmap in Figure 7. Hence, they classify independently the target dependent variable.

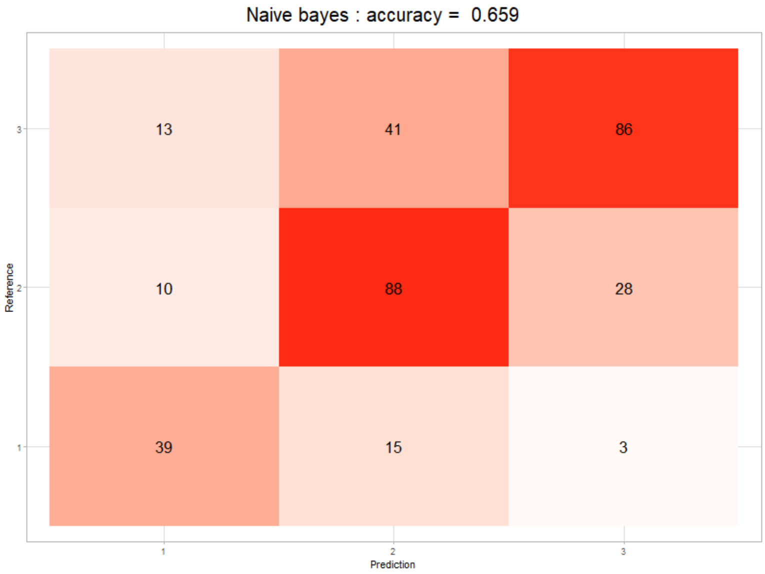

Figure 10: Confusion Matrix and Accuracy for Naive Bayes Classifier

In *Figure 9* we can observe the confusion matrix obtained through the Naive Bayes Classifier. Also in this case the accuracy is very similar to the previous algorithms analyzed, and it is equal to 0.659. It is interesting to highlight that it classifies very well the labels '2' and '3'. It's performance is worsening with respect to the previous models since it miss classifies much more labels.

3.2.4. Support Vector Machine

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The algorithm creates a line or a hyperplane which separates the data into classes.The best hyperplane is the hyperplane whose distance to the nearest element of each tag is the largest.

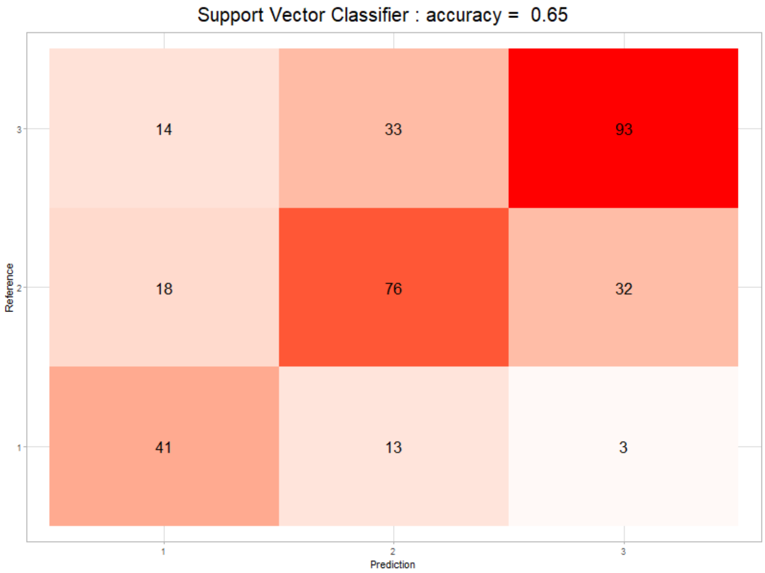

Figure 11: Confusion Matrix and Accuracy for Support Vector Machine

By analyzing the confusion matrix obtained by the predictions of support vector machine is interesting to notice that is the model which best classifies the label '1' and label '3'. The accuracy is not so high, but in line with the previous models. It doesn't improve the miss classification of the most far labels, which is the measure we would like to minimize.

3.3. Comparing the models

In this section we will perform a comparison between the different models analyzing the measures of Sensitivity, Specificity and Precision of the different models. Sensitivity and specificity describe the accuracy of the a classification. We could define the sensitivity as the ratio of the well predicted labels over the true labels, while the specificity is ratio of the miss classified labels. A high specificity means that there are less mistakes in that class.

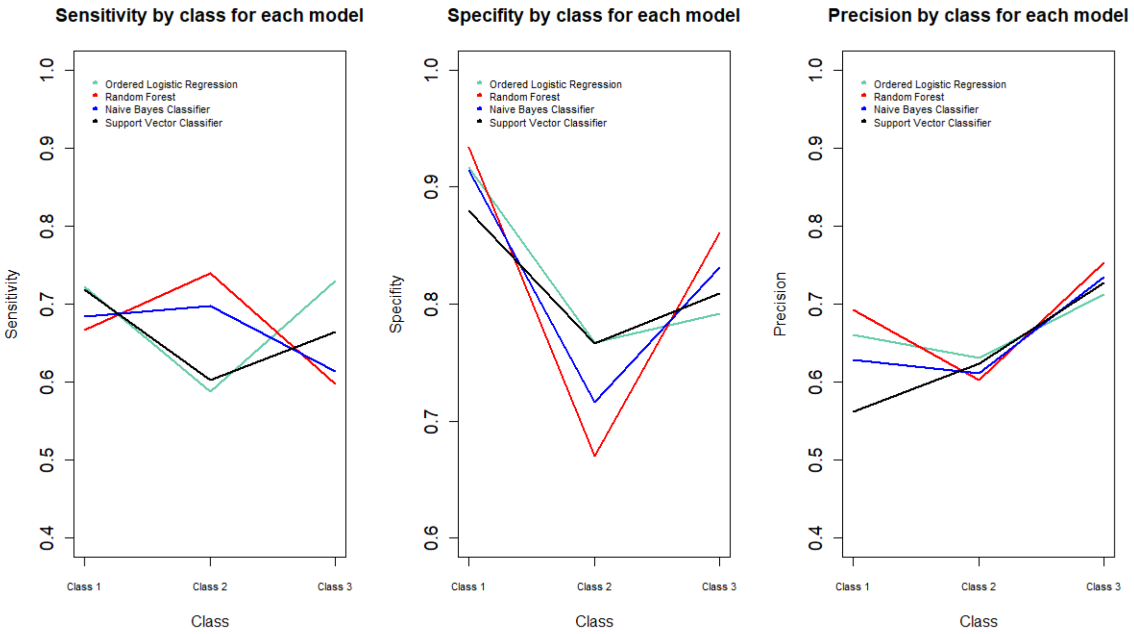

Figure 12: Comparison of the models by class

These plots are very interesting since they show the sensitivity, the specifity and the precision of the classification for each class and for each model. We can observe that the models slightly differ when predicting the different classes. For example, Random Forest is the model which best predict the class ’2’ with a very high sensitivity with respect to the other models. Analyzing the specificity of the second class, the OLR has the highest value for the class ’2’ because it is the one that predicts the worst class ’2’ but at the same time it makes less mistakes than the other models.

4

Conclusion

In conclusion, we implemented four different models to classify the finishing position of Formula One races. We observed that all the models actually have very similar perfomance, with an accuracy that is almost the same. We still believe that the best model is the Ordered Logistic Regression since it has a very sensitivity when predicting class '1' and class '3'. In addition where it predicts badly, class '2', it still has a very low specificity which means that it makes less mistakes in predicting the labels of class '2'.