FORMULA 1 DATA ANALYSIS

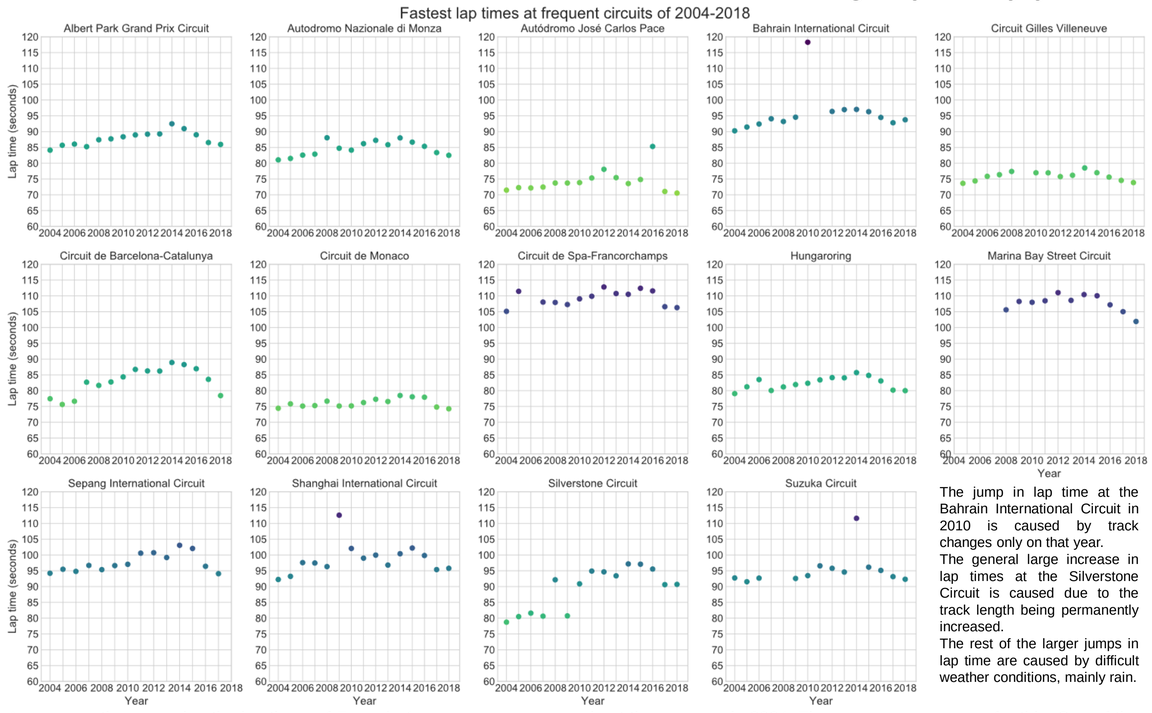
Peeter Paal, Aveli Klaos, Siim Tanel Laisaar

Introduction

Formula 1 (F1) is the highest class of single-seater auto racing sanctioned by the Fédération Internationale de l'Automobile (FIA). The FIA Formula One World Championship has been one of the premier forms of racing around the world since its inaugural season in 1950. Every F1 race generates a huge amount of data — we used a large, free, and non-official dataset with data from the first season to the last complete season of 2018. We set 4 main goals which are further explained in their according sections.

2004-2018 lap times

The lap times are taken from racing circuits that have been used 11 or more times during the 2004-2018 period. This restriction was placed because the set of circuits for an F1 season often changes year by year.

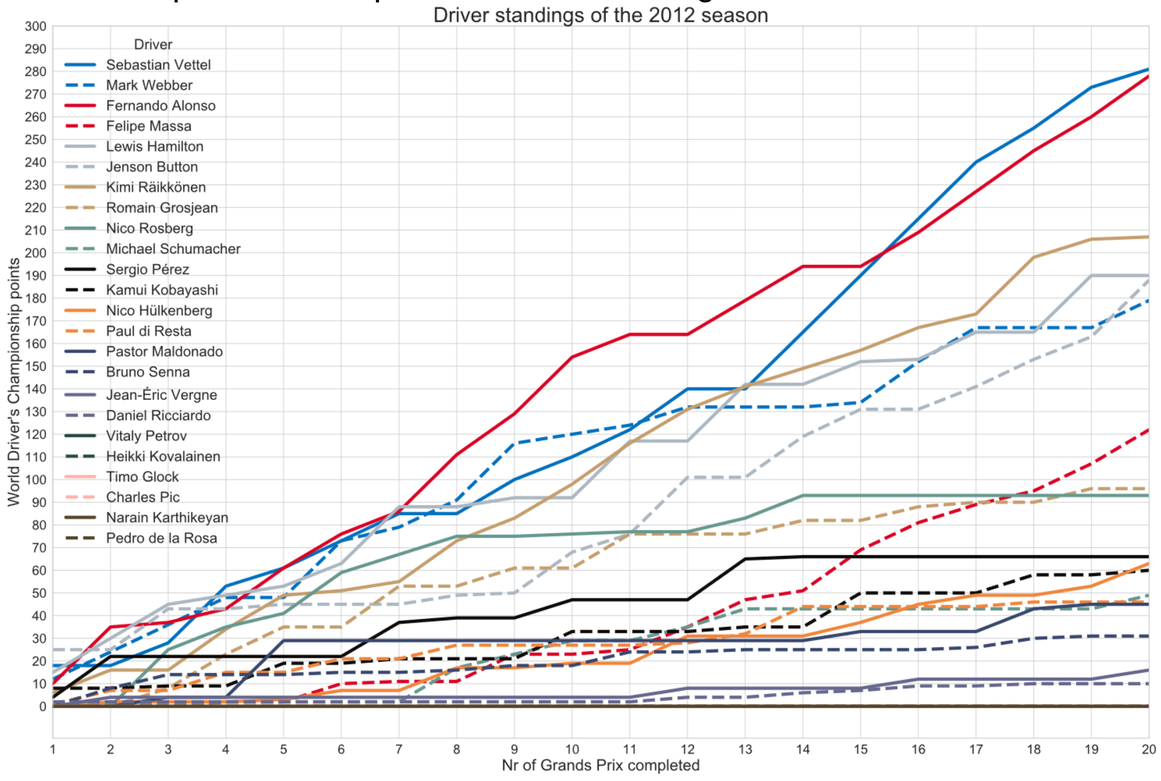


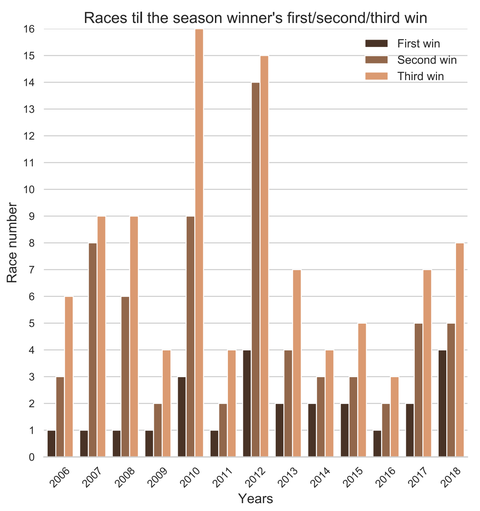
The jump in lap time at the Bahrain International Circuit in 2010 is caused by track changes only on that year.
The general large increase in lap times at the Silverstone Circuit is caused due to the track length being permanently increased.
The rest of the larger jumps in lap time are caused by difficult weather conditions, mainly rain.

As seen on these graphs, the lap times of Formula 1 cars are on average about the same as in 2004. There are many reasons for this. One of the main ones is the ban of refuelling during the race. Before 2010, the cars used to carry light fuel loads, refuelling when necessary. Nowadays, the cars have to carry fuel for the whole race which slows them down noticeably. The increasing focus on safety in F1 also comes at the cost of some speed. Furthermore, the engines have been made more reliant on electrics and to be more efficient in favour of raw power.

Analysis of the 2012 season

The 2012 season was one of the most competitive seasons of the last decade. Thus we decided to analyse the data from it in order to visualise it and compare it to the previous and following six seasons.

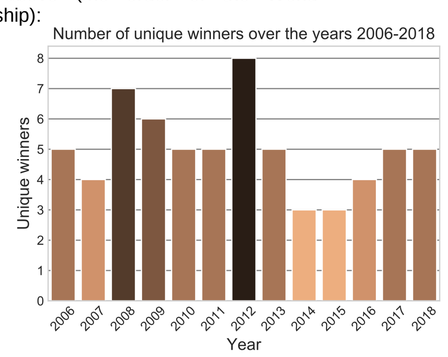




Every colour on the graph above represents a team and its two drivers. The colours are chosen by the main colour of the team on that season. In every pair, the better performing drivers' results are visualised by a regular line while their teammates' results are shown by a dashed line.

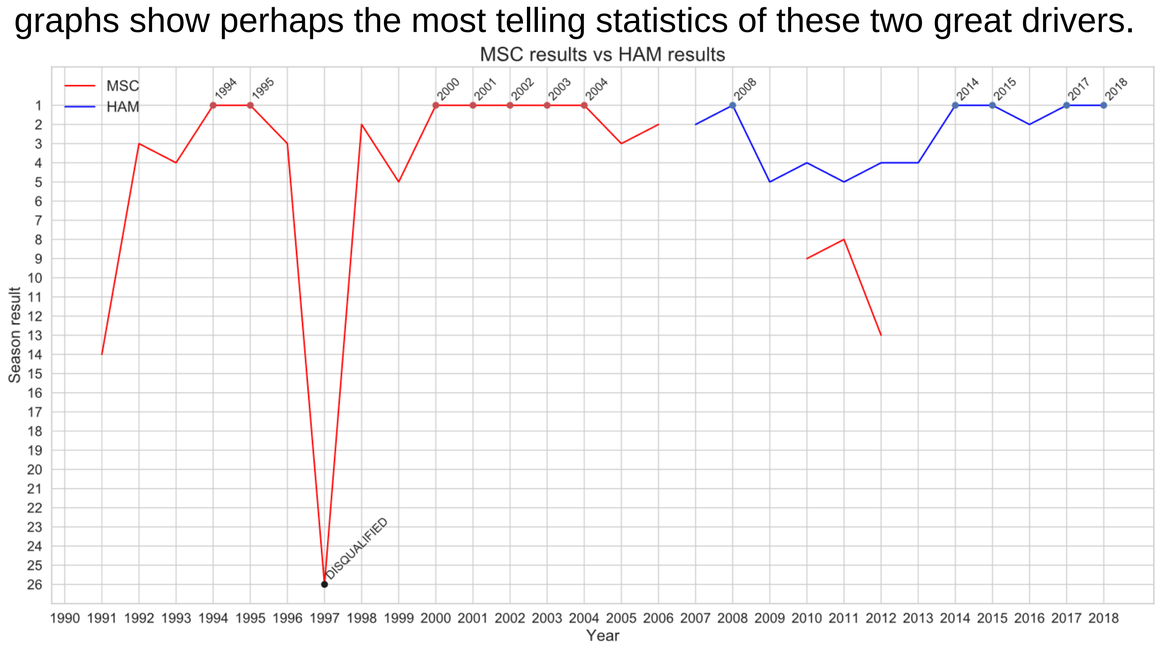
Teams of the 2012 season in the same order as on the legend of the graph above (the order of the World Constructor's Championship):

1. Red Bull-Renault
2. Ferrari
3. McLaren-Mercedes
4. Lotus-Renault
5. Mercedes
6. Sauber Ferrari
7. Force India-Mercedes
8. Williams-Renault
9. Toro Rosso-Ferrari
10. Caterham-Renault
11. Marussia-Cosworth
12. HRT-Cosworth

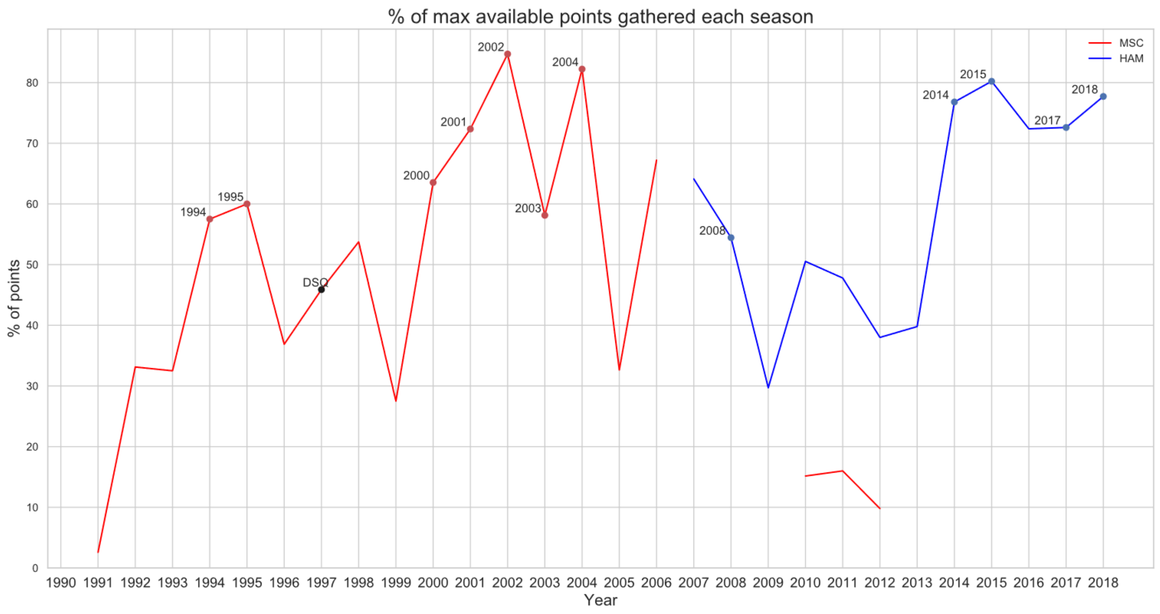


Michael Schumacher vs Lewis Hamilton

Michael Schumacher (MSC) and Lewis Hamilton (HAM) are two of the most successful drivers in Formula 1 by the number of World Driver's Championship titles. While Michael Schumacher has been considered the greatest for a long time, Lewis Hamilton's recent success has brought him into contention for the title of the greatest of all time. While this was a difficult subject to visualise due to large amounts of missing data, these graphs show perhaps the most telling statistics of these two great drivers.



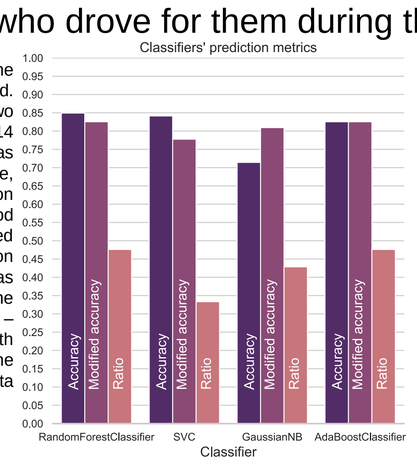
At the final race of the 1997 season Schumacher was disqualified from the World Drivers' Championship (WDC) for causing an avoidable accident. At the moment of the accident Schumacher was about to lose his leading position and thus the WDC title. It is also noteworthy that at this point Lewis Hamilton has won the 2019 season's WDC title (his 6th).



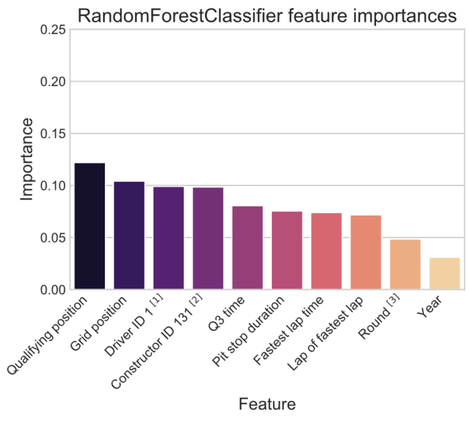
Predicting winners of the 2018 season

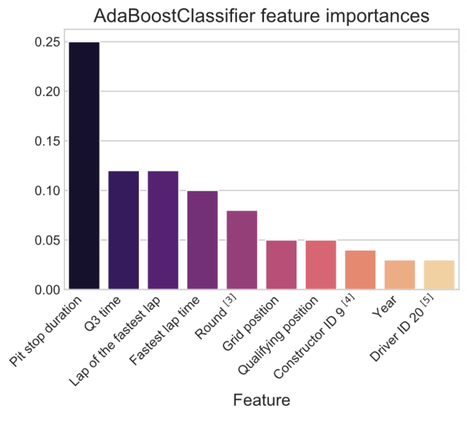
We used RandomForestClassifier, Support Vector Classifier, GaussianNB and AdaBoostClassifier to predict the winners of the 2018 season. We used the seasons 2014-2016 as the training set, 2017 as the validation set and 2018 as the test set. Because during this period only the drivers from the top 3 teams (Red Bull, Mercedes, Ferrari) won, we only used data of those drivers who drove for them during the year 2018.

Due to the nature of the sport, the corresponding data is heavily unbalanced. In different eras of Formula 1, one or two teams tend to dominate. After the 2014 regulation changes, the Mercedes team has been the most dominant. Furthermore, Mercedes' driver Lewis Hamilton has won the championship 4 times during this period with the 5th title belonging to now retired Mercedes driver Nico Rosberg. One reason why the classifiers are not performing as well as they could is that one of the Mercedes' drivers of the 2018 season – Valtteri Bottas – had an unlucky season with bad results, meaning he won 0 times. The classifiers, however, had trained on data where both Mercedes drivers often won.



The 4 scikit-learn library classifiers visualised on the graph to the left – RandomForestClassifier, SVC, GaussianNB and AdaBoostClassifier – generally performed the best compared to some other classifiers.
The graph displays the classifiers' different metrics, two of them created by us.
Accuracy – the accuracy of the predictions on the test set.
Modified accuracy – the accuracy of modified predictions on the test set. Every races' predictions are altered to only have one predicted winner. The classifiers don't know that exactly one driver must win, so they may predict 0 or more than 1 winners. We forced only the prediction with the highest probability to be a win.
Ratio – number of wins predicted correctly (after modifying) divided by the number of races.





[1] Driver ID 1 determines whether the driver was Lewis Hamilton or not.
[2] Constructor ID 131 determines whether the driver raced for the Mercedes team or not.
[3] Round determines which race of the year it was (for example the 4th race).
[4] Constructor ID 9 determines whether the driver raced for the Red Bull team or not.
[5] Driver ID 20 determines whether the driver was Sebastian Vettel or not.