Using Data Science Techniques in Formula 1

Asala Aljazaery

12 December 2022  
  
University of Sunderland

Word count: 1,542

Table of Contents

[Introduction 1](#_Toc121582409)

[Data Preprocessing 2](#_Toc121582410)

[Machine Learning Techniques 2](#_Toc121582411)

[Patterns Detection 3](#_Toc121582412)

[Conclusions 4](#_Toc121582413)

[References 4](#_Toc121582414)

# Introduction

Formula 1 (F1) is a racing sport with the fastest cars, well known as a data-driven sport. F1 collects data from hundreds of sensors placed into each vehicle. The reaction speed of the driver and the system for the received data can determine the win or loss. The data gathered can estimate the results of future races and can support the team in building faster cars that are more efficient and remain beneath the budget at the same time. The auto industry’s future depends on Data analysis and evolving increasingly ([Wodehouse, 2021](#ref-Wodehouse2021)). Data has to be synchronized and accessed by the race support room as quickly as possible during the race. Many decisions are made based on this data. Data latency plays a significant role in these decisions. Furthermore, the data reading must be correct from the first time because the race can not be repeated. Data Scientists team process the data during the race or afterward ([Media, 2022](#ref-Media2022)), because analyzing the Data on the spot will give insights into the race strategies and decisions affecting the results ([Accenture, 2022](#ref-Accenture2022)).

This report will present how F1 gathers different types of data and how to preprocess these data types for further analysis, discussing numerous machine learning techniques with the used infrastructure. Examine diverse areas where data play a significant role in F1. Then, explain the data’s effect on the technical side and mark some discoveries associated with different patterns and their impact. Finally, Investigate future technologies that contribute to sustainable solutions in the auto industry.

# Data Preprocessing

F1 management built a standardized communication network that gives access points to the teams to send real-time data between the car and the garage ([Wodehouse, 2021](#ref-Wodehouse2021)). Edge computing reduces latency by processing the data near the source using edge devices which refers to networks or devices near the users, such as the sensors ([Accenture, 2022](#ref-Accenture2022)). Sensors can measure physical quantities like temperatures and pressures or system operations like gearbox units and brake temperature. All these factors keep the car stick to the road, balanced, and at high speed. There are more than 250 sensors on each car. Each car generates around one terabyte of data during the race. It is about 30 megabytes per lap and can be more by two or three times when the car returns to the pitstop ([Media, 2022](#ref-Media2022)). F1 uses Amazon web services S3 cloud for data storage and data analytics ([Miller, 2021](#ref-Miller2021)). F1 uses a hybrid distributed relational database system to provide the availability and high scalability of NoSQL and the consistency of the traditional database ([Shute *et al.*, 2013](#ref-Shute2013)). F1 Datasets have attributes such as F1 driver, race number, number of laps, laps time, seasons data, and pitstop status ([Wyawahare, 2021](#ref-Wyawahare2021)).

In preprocessing, data types can be nominal, ordinal, or continuous ([Team, 2022](#ref-GreatLearningTeam2022)). The missing values must be detected because they will affect the machine learning models’ reliability and performance. Missing values can be due to crushes or mechanical failure ([Sicoie, 2022](#ref-Sicoie2022)). These values can be filled by predicting the probabilities using the regression or removing the row. Then, outliers are detected by clustering the data into groups using unsupervised classification. If inconsistent data or anomalies are presented, they can be removed depending on the analysis’s purposes. For noisy data issues, we can apply filtering methods to minimize them. The Data must be normalized by reducing redundancy and dependence. Then all Correlations will be tested to choose which attributes will be valuable for data mining ([Alghamdi and Javaid, 2022](#ref-Alghamdi2022)). Natural language processing will be used for audio collected from the team radio. Data will be normalized by breaking the sentences into words, symbols, or characters to identify each keyword’s meaning, and unnecessary keywords will be removed. Then, a standard text version will be created for keywords in multiple variances. Word frequency and the driver’s main concerns while driving will be detected and compared ([Saxena, 2021](#ref-Saxena2021)). Finally, sentiment analysis labels the sentences as positive, negative, or neutral ([Kulkarni, 2022](#ref-Kulkarni2022)).

# Machine Learning Techniques

In F1, the regression and classification methods are used for supervised learning, while clustering methods are for unsupervised learning. The chosen method depends on the data characteristics and the goal of the analysis:  1. Classification is used to classify the data into specified categories and label it.  2. Regression is an algorithm used to predict future values.   3. Clustering is used when unlabelled data needs to be categorized and labeled depending on the similarities or differences ([IBM, 2022](#ref-IBM2022)).  For methods evaluation, the precision score metric is the most important for this type of test to get a high rate of correct predictions ([scikit-learn, 2022](#ref-scikit-learn2022)).

Time series analysis is a method to analyze data points over specific periods. In F1, time plays a significant role and will help us with race progress. The statistical methods for time series forecasting are Naive models, Exponential Smoothing Models, Autoregressive integrated moving averages (ARIMA), and linear regression. There is plenty of machine learning techniques for time series forecasting, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Decision Trees and Gradient Boosting variations, and Extreme Learning Machines (ELM). These methods recognize the characteristics and patterns of data to make predictions ([CodeIT, 2022](#ref-CodeIT2022)). Gradient boosting algorithms are very powerful for classifying ([Brownlee, 2016](#ref-Brownlee2016)).

Artificial Neural Networks (ANN) such as CNN and RNN are feedforward neural network models that can detect any non-linearity in the data. ANN architecture is based on the weights and biases inside multiple hidden layers set by backpropagation learning algorithms that feed the neural network. ELM has the same ANN concept, but they take less training time and contain only one hidden layer, and parameters are set randomly without a backpropagation algorithm ([Chaudhuri, 2021](#ref-Chaudhuri2021)). These algorithms can be implemented using R or python ([Brownlee, 2016](#ref-Brownlee2016)). For model evaluation, the Mean Absolute Percent Error (MAPE) will compare the forecast accuracy and the error percentage ([Chaudhuri, 2021](#ref-Chaudhuri2021)). Cross-validation will tune the different dataset hyperparameters ([Brownlee, 2018](#ref-Brownlee2018)).

Amazon SageMaker trains deep-learning models on real-time data to make forecasts and provide insights into the driver or team strategies and a better fan experience ([Elucidate, 2022](#ref-Elucidate2022)). SageMaker uses an open-source library in python called SageMaker Python SDK to deploy models using popular deep-learning models frameworks such as Apache MXNet and TensorFlow ([GitHub, 2022](#ref-GitHub2022)).  Cloud computing is critical for data mining as a massive amount of data in categorized feeds will be sent to the mechanics and car engineers in the operation rooms and factories ([Dey, 2022](#ref-Dey2022)). Amazon Elastic Compute Cloud (Amazon EC2) helped F1 build multi-car simulations to assess and test the car’s aerodynamics. This project crunch around 550 million data points to model the effect of one car’s aerodynamics on another, helping them to produce cost analysis, predict production issues, and makeup decisions before getting the car onto the track ([Miller, 2021](#ref-Miller2021)).

# Pattern Detection

The data scientist’s job in the company is to skim the data and find crucial information for the team ([Schutt and O’Neil, 2013](#ref-Schutt2013)). A correlation between the driver’s starting position and the chances of winning was spotted ([Nigro, 2020](#ref-Nigro2020)). Driver’s age and performance were significantly correlated ([Sicoie, 2022](#ref-Sicoie2022)). Winning predictions using classification were 66% correct ([Nigro, 2020](#ref-Nigro2020)).

For Strategy patterns, Data gave insights into tyre performance and which tyre types to pick based on ([Amazon, 2022](#ref-Amazon2022)) 1. The number of laps, which affects the number of pitstops and the wasted time 2. Car speed, Because hard types decrease the car speed immensely. 3. Circuit conditions such as weather ([F1, 2022](#ref-MercedesAMGF12022)). In the latest race this year, the predicted strategy for efficient tyres in the case of three pitstops is to use the medium type up for 20 laps, switch to hard ones, and then go back to the medium type after finishing 44 laps. For two pitstops, if the driver can use the medium at the beginning of the race up to 30 laps, there is a high chance for the driver to keep the hard types and make it to the end. Taking into consideration that the temperature will be over 30°C at the race beginning, but the conditions will change as soon as the sun sets, and the temperature will drop eventually ([Formula1, 2022a](#ref-Formula12022a)).

For technical patterns, car simulations using AWS enabled the F1 to build a design that minimized the downforce effect, preventing the car from overtaking by making stopping or changing sides very difficult and causing tyres to wear. AWS cloud used over 550 million data points affecting downforce, which has helped them build 7000 car models and find the best design to reduce the downforce effect from 50% to 15% ([Review, 2021](#ref-AustralianFinancialReview2021)). Thus, it reduced the testing time by 80% and lowered the simulation cost by up to 30% ([Services, 2022](#ref-AmazonWebServices2022)). For performance patterns, the driver could review their performance to learn from their mistakes and perform better next race ([Media, 2022](#ref-Media2022)). For Financial patterns, the teams can determine the amount of carbon fiber they need, avoid unnecessary orders, and cut costs, because carbon fiber is very expensive ([F1, 2022](#ref-MercedesAMGF12022)).

# Conclusions

F1 data analysis plays a critical role in their decisions and delivers insights that were impossible before. This revolution has supported strategies simultaneously. Using AWS in F1 changed the competition strategies and fan experience drastically. AWS supports predictive analytics and increases prescriptive intelligence. The prescriptive side is discovered by applying complex algorithms to our data that inform the company what decisions they should make. The company aims to be more sustainable by 2030 and started carbon reduction projects. The data analysis presents a tremendous opportunity to overcome this global issue. Therefore, F1 has adopted new net zero-carbon technologies for Heating, ventilation, and air conditioning (HVAC) ([Formula1, 2022b](#ref-Formula12022)). HVAC technology is based on energy management using data science to control their systems and reduce the environmental impact ([Latinoamérica, 2022](#ref-ACRLatinoamerica2022)). This innovation has positively contributed to carbon emissions in F1 ([Formula1, 2022b](#ref-Formula12022)).

# References

Accenture (2022) *Edge Computing | Accenture*. Accenture. Available at: <https://www.accenture.com/ae-en/insights/cloud/edge-computing-index> (Accessed: 2 December 2022).

Alghamdi, T.A. and Javaid, N. (2022) ‘A Survey of Preprocessing Methods Used for Analysis of Big Data Originated From Smart Grids’, *IEEE Access*, 10, pp. 29149–29171. Available at: <https://doi.org/10.1109/ACCESS.2022.3157941>.

Amazon (2022) *F1 Insights powered by AWS | Formula 1 uses Amazon Web Services*. Amazon Web Services, Inc. Available at: <https://aws.amazon.com/sports/f1/> (Accessed: 3 December 2022).

Brownlee, J. (2016) *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*. MachineLearningMastery.com. Available at: <https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/> (Accessed: 8 December 2022).

Brownlee, J. (2018) *A Gentle Introduction to k-fold Cross-Validation*. MachineLearningMastery.com. Available at: <https://machinelearningmastery.com/k-fold-cross-validation/> (Accessed: 8 December 2022).

Chaudhuri, K.D. (2021) *Time Series Forecasting with Extreme Learning Machines -*. Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/12/time-series-forecasting-with-extreme-learning-machines/> (Accessed: 8 December 2022).

CodeIT (2022) *Using Machine Learning for Time Series Forecasting Project*. CodeIT. Available at: <https://codeit.us/blog/machine-learning-time-series-forecasting> (Accessed: 8 December 2022).

Dey, V. (2022) *Why Are F1 Teams Turning To Big Data Analytics?* Datatechvibe. Available at: <https://datatechvibe.com/data/why-are-f1-teams-turning-to-big-data-analytics/> (Accessed: 3 December 2022).

Elucidate (2022) *Formula 1 and Machine Learning*. Available at: <https://www.elucidate.ai/post/formula-1-and-machine-learning> (Accessed: 2 December 2022).

F1, M.A. (2022) *INSIGHT: Five Examples Why F1 Is Accelerating the Future*. Available at: <https://www.mercedesamgf1.com/en/news/2018/10/insight-five-examples-why-f1-is-accelerating-the-future/> (Accessed: 3 December 2022).

Formula1 (2022a) *STRATEGY GUIDE: What are the possible race strategies for the 2022 Abu Dhabi Grand Prix? | Formula 1*. Available at: <https://www.formula1.com/en/latest/article.strategy-guide-what-are-the-possible-race-strategies-for-the-2022-abu-dhabi.1kuZv7pc5ED8wmELjMKI8s.html> (Accessed: 8 December 2022).

Formula1 (2022b) *Sustainability Strategy*. Available at: <https://corp.formula1.com/wp-content/uploads/2019/11/Environmental-sustainability-Corp-website-vFINAL.pdf> (Accessed: 7 December 2022).

GitHub (2022) *Aws/sagemaker-python-sdk*. Amazon Web Services. Available at: <https://github.com/aws/sagemaker-python-sdk> (Accessed: 3 December 2022).

IBM (2022) *Supervised vs. Unsupervised Learning: What’s the Difference?* Available at: <https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning> (Accessed: 9 December 2022).

Kulkarni, A. (2022) *Understand Text Mining Using No Code Tool Orange*. Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2022/05/understand-text-mining-using-no-code-tool-orange/> (Accessed: 8 December 2022).

Latinoamérica, A. (2022) *Data Science: Strategy for Energy Management in HVAC Systems*. Available at: <https://www.acrlatinoamerica.com/en/2022051717722/news/enterprises/data-science-strategy-for-energy-management-in-hvac-systems.html> (Accessed: 7 December 2022).

Media (2022) *How Much Data Does An F1 Car Generate? | F1 Data Explained*. Available at: [//f1chronicle.com/how-much-data-does-an-f1-car-generate/](https:////f1chronicle.com/how-much-data-does-an-f1-car-generate/) (Accessed: 30 November 2022).

Miller, R. (2021) *How Cloud Data-Crunching Power Accelerates the F1 Racing Experience*. Data Center Frontier. Available at: <https://www.datacenterfrontier.com/cloud/article/11427867/how-cloud-data-crunching-power-accelerates-the-f1-racing-experience> (Accessed: 2 December 2022).

Nigro, V. (2020) *Formula 1 Race Predictor*. Medium. Available at: <https://towardsdatascience.com/formula-1-race-predictor-5d4bfae887da> (Accessed: 2 December 2022).

Review, A.F. (2021) *How Formula 1 uses machine learning to spice up races*. Australian Financial Review. Available at: <https://www.afr.com/technology/how-formula-1-uses-machine-learning-to-spice-up-races-20210716-p58abx> (Accessed: 2 December 2022).

Saxena, S. (2021) *Tokenization and Text Normalization*. Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/03/tokenization-and-text-normalization/> (Accessed: 8 December 2022).

Schutt, R. and O’Neil, C. (2013) *Doing data science*. First edition. Beijing ; Sebastopol: O’Reilly Media.

scikit-learn (2022) *Sklearn.metrics.precision\_score*. scikit-learn. Available at: <https://scikit-learn/stable/modules/generated/sklearn.metrics.precision_score.html> (Accessed: 7 December 2022).

Services, A.W. (2022) *Formula 1 Case Study | HPC | AWS*. Amazon Web Services, Inc. Available at: <https://aws.amazon.com/solutions/case-studies/formula-1-graviton2/> (Accessed: 5 December 2022).

Shute, J. *et al.* (2013) ‘F1: A distributed SQL database that scales’, *Proceedings of the VLDB Endowment*, 6(11), pp. 1068–1079. Available at: <https://doi.org/10.14778/2536222.2536232>.

Sicoie, H. (2022) ‘Machine Learning framework for Formula 1 race winner and championship standings predictor’, p. 24.

Team, G.L. (2022) *4 Types Of Data - Nominal, Ordinal, Discrete and Continuous*. Great Learning Blog: Free Resources what Matters to shape your Career! Available at: <https://www.mygreatlearning.com/blog/types-of-data/> (Accessed: 10 December 2022).

Wodehouse, C. (2021) *How Formula 1 Car Sensors Create Data at Every Turn*. Pure Storage Blog. Available at: <https://blog.purestorage.com/perspectives/how-formula-1-car-sensors-create-data-at-every-turn/> (Accessed: 3 December 2022).

Wyawahare, C. (2021) *Formula 1 Grand Prix Analysis*. Medium. Available at: <https://towardsdatascience.com/formula-1-grand-prix-analysis-d05d73b1e79c> (Accessed: 29 November 2022).