Machine Learning-Based Analytical and Predictive Study on Formula 1 and Its Safety

S. Dhanvanth, Rohith Rajesh, S. S. Samyukth, and G. Jeyakumar

Abstract *Formula 1*, in short *F1,* is an international racing sport. It is of the highest class among the category. It is a premier sporting event right from the time it started in 1950. It consists of a series of races, known as Grand Prix, taking place across the world on a wide range of circuits. There is enormous amount of data being captured, analyzed, and used to design, build, and drive the *F1* cars. The primary goal of this paper is to increase the safety in *F1* races by predicting if a race can result in accidents based on the conditions and factors under which the race will be held. This paper also proposes to use different racing datasets from various sources in order to analyze, visualize, and make fruitful predictions along with some kind of analysis using different types of plots, based on the requirements. The data are visualized using different plots after doing required preprocessing. A few conclusions were made based on the patterns observed through comparing different plots. This paper also presents the comparative performance analysis of cars over the years. The performance measurement considered in this study (based on the race results) is the time taken for laps/races. Apart from this, a comparative analysis of two drivers on different race tracks was depicted using plots. Finally, the paper presents a machine learning model to predict the winner of the next race based on the previous results including important factors like the weather conditions.

Keywords Data visualization · Data analysis · Regression model · Machine learning · Formula 1

1 Introduction

Formula1 being a sport where the cars are racing at extremely high speeds, it is a risky sport unless proper precautions are taken. The *FIA* has come up with stringent rules to ensure the safety of drivers. In spite of this, there have been many dark hours

S. Dhanvanth · R. Rajesh · S. S. Samyukth · G. Jeyakumar (✉)
Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
e-mail: g_jeyakumar@cb.amrita.edu

A. Tomar et al. (eds.), *Proceedings of 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication*, Lecture Notes in Electrical Engineering 915, https://doi.org/10.1007/978-981-19-2828-4_25

in this sport. Many of these accidents have resulted in fatalities making this a matter of deep concern. This paper attempts to make the sport safer by predicting in advance the possibility of accidents in an upcoming race by considering factors like the driver, his grid position, the weather, and the past records of accidents at that circuit. This prediction could indeed result in avoiding fatalities and accidents in races to a great deal [1–22].

Data visualization is a vast field dealing with the graphic representation of data. It is a very useful way to depict and communicate especially when the data are entirely numeric. It provides an accessible way to note different trends, outliers, and patterns in data. Translating information into a visual context, such as a map or graph makes it easier to get much better insights. Effective visualization helps users analyze and reason about data with evidence. Effective visualization implies making the right choice from the huge pool of choices like scatterplots, line plot, pie charts, histograms, and a lot more.

Machine learning provides the luxury of training models where the models learn by experience. It can be used in different applications where algorithms are used to train models. This learning capability of models is used in the work presented in this paper. In this work, a model is trained to predict the race winners after training the model using the previous results. This paper is an attempt to demonstrate the various visualization techniques by analyzing the *Formula 1* datasets using a machine learning model. The proposed model is used to predict the winner of the next *Formula 1* race. The sources of the datasets used in this study and the conclusive plots are presented in this paper.

Some insights and rules in Formula 1 are 'for a year, in a particular circuit, the race held is called a Grand prix', 'there are many teams as per the Federation Internationale de l'Automobile (FIA) rules', 'one team can send only 2 drivers for a race', 'the teams are called constructors', and 'each driver belongs to a team (constructor)'. The *F1 Grids* determine the start position of a racer, and it is based on their performance in the qualifying round.

Rest of the paper is organized as follows—Sect. 2 narrates few important related works, Sect. 3 discusses the details of the proposed work, Sect. 4 explains the results and discussion of the machine learning models, and the Sect. 5 concludes the paper.

2 Related Works

Formula 1 undoubtedly strives on the results from various kinds of data. The impact of analyzing data is huge in this sport. There are numerous data scientists working for each team in Formula 1. Few interesting related works are highlighted in this section.

There have been many attempts to visualize the data. The work presented in [1] shows a year wise overview in a calendar representation and the lap time overview of drivers using visualization tools. [2] is another work where data are collected from scratch, then analyzed using a few plots and finally a prediction of winning races

using a regression model is done. In [3], the *Formula 1* data have been interpreted in the best way possible which could help to arrive at conclusions. Tools like *Tableau* have been used in [3] to explore the data in the form of plots. The importance of word cloud is well written in [4]. The correct use of the type of regression is vital in getting better predictions. Hence, choosing the right learning model is important. The articles [5, 6] have sighted the use cases where the logistics regression should be used.

Before giving the data to any model, the data have to be preprocessed in such a way that it helps the model to make near-accurate predictions. [7] explains, in detail, the library called pandas which is very vital in the various stages of data preprocessing. [8, 9] explain how the data visualization plays a major role not just in understanding the data but also for presenting data in an attractive way.

There are also many interesting articles in the literature presenting the aerodynamic design of race cars [10] and the mechanical and technical aspects for *Formula 1* technology [11].

Using machine learning approaches for predicting various factors in the real-world problems around us is a common trend of research nowadays. In [12], the basic machine learning models have been used in the ideal way possible to get good predictions of placement possibility of students. A supervised machine learning algorithm-based system to process and classify the images of the license plates is presented in [13]. The authors of [14] have done bitcoin price prediction using time series and roll over. The comparison of random forest regression and logistic regression and the ideal use cases for the techniques is presented in [15]. The idea of using neural networks to make strategical changes in races is well presented in [16]. An efficient way of analyzing the presence of gender bias in text data systems in terms of the occupations, using a pre-trained model, is proposed by authors of [17]. Ref. [18] presents a model to translate the sign language words to English word, using deep learning techniques. In [19], the paper presents an approach where artificial neural networks are used to predict race results.

Considering the study presented above, this paper presents an experimental work which focuses to use these resources to assimilate the necessary information and use the impactful parameters in the analysis and design of a machine learning model.

3 The Proposed Work

The work presented in this paper aims at predicting accidents in an upcoming race in order to enhance safety of driver, carrying out visualization to bring out meaningful observations, and then using the best suitable machine learning model to predict the winner of *Formula 1* race.

The primary goal of increasing safety in the sport by predicting the accidents in an upcoming race is carried out by preprocessing the data to extract the correlated attributes and then using the new processed dataset with an appropriate learning model in order to predict the accidents. Three models using random forest, *KNN,*

and logistic regression are proposed to be tested and the one better with the better performance has to be chosen.

The following are the observations that are intended to be found out by the visualization techniques.

1. How the speeds of the cars have changed over the years? Has the speed of cars increased continually due to the tweaks in the car by the constructors?
2. Is the grid position a major deciding factor in the race?
3. A comparison between two rivals of this era.

This work also proposes a machine learning model to predict the winner of a race based on the correlated factors. To reach the goal of using visual techniques to analyze the data, the data preparation and exploration were done first.

3.1 Data Preparation and Exploration

The datasets for this study are taken from following sources.

- A group of datasets from Kaggle [20] giving details of Formula 1 race from 1950 to 2017. This dataset gives the information about constructors, race drivers, lap times, status, pit stops, and race results.
- In order to get more recent data, few datasets were formed after scraping details from Websites which also gives additional information like the weather condition during the race.

After preparing these datasets, the next task of preprocessing is done as and when required based on the requirements. For the purpose of prediction of accidents, the preprocessing steps before training the model involved:

- Scaling the data using min–max scaler in order to restrict each feature to a specific range.
- *SMOTE*, which is an oversampling technique, is used in order to generate synthetic values for minority class, thus making the dataset balanced.

On observing the final preprocessed data frame used for the prediction of accidents, it was observed that the data are imbalanced as the number of records available for accident data is much smaller than the non-accident data. This imbalance can lead to a biased model which will result in many false negatives (predicting that no accidents have occurred in a race where accidents actually occurred).

To handle such an imbalance, synthetic minority oversampling technique (SMOTE) is used to oversample the minority class.

3.2 Treating Missing Value

The missing values in the dataset were minimal, and in order to avoid discrepancies by using methods to fill missing values, the missing values are dropped.

4 Results and Discussion

The experimental results are presented in three phases. The Phase I focuses on the proposed model to predict accidents in an upcoming race. The Phase II focuses on the analyzing the observations made through a set of data visualization plots. The Phase III is discussing the proposed machine learning model to predict winner of a race.

4.1 Phase I—Predicting Accidents in an Upcoming Race

In this phase, the preprocessed datasets were used to train three different models in order to predict accidents in an upcoming race. The three different machine learning algorithms used for training were logistic regression, random forest, and *KNN*. Logistic regression is a parametric model used for linear solutions. *KNN* is a non-parametric model which supports non-linear solutions although it takes more time than logistic regression. Random forest follows the principle of ensemble learning wherein it functions by constructing multiple decision trees to carry out predictions.

The features that were chosen for training the models were decided based on correlation analysis. The features chosen were the driver *ID*, the grid position of the driver, the weather condition where the race is going to be held and the past record of accidents on that particular circuit for a particular constructor.

The results upon training models using the mentioned features are compared using accuracy as the measure. Accuracy is a suitable measure here as the data have been made balanced in the preprocessing phase. The logistic regression-based model gave an accuracy of only 44.74% while the models based on random forest and *KNN* gave accuracy of 72.20% and 85.93%, respectively. The poor accuracy of logistic regression is because it does not work well for non-linear predictions. *KNN* is working with the nearest neighbors, and hence, it tends to give better results in some particular cases like in our model, thereby giving a higher accuracy.

This model with an accuracy of 85.93% can indeed help in prediction of accidents and thereby influence the safety factor of the sport. The *FIA* could in fact use these predictions to make key decisions like rescheduling a race which could in turn avoid fatal accidents.

4.2 Phase II—'Data Visualization and Analysis'

This section discusses the findings in the order of the observations.

Observation 1: 'There is no constant increase in speed of cars over the years'

A part of the plot used to bring out this observation is shown in Fig. 1. 27 such subplots were generated. This plot consists of subplots for each race track, initially the result and the race datasets were merged. Then, the required preprocessing was done in order to obtain uniformity in the lap times for plotting the values. Then, the fastest lap time for each track was taken out year wise and stored separately. This was then plotted using subplots where each subplot is for a different Grand Prix, and the x and y axes denote the year and fastest lap time, respectively.

On observing the line plotted in each subplot, the speed of the cars has reduced during the middle years showing the variations in the line plot. This variation occurred as a result of the decision by the officials to enforce safety in the sport and thereby bring in more stringent rules which have to be enforced by the constructors.

Observation 2—'The grid position plays a deterministic role in the final race position'

This observation is made with the help of a scatterplot (shown in Fig. 2) using the *seaborn* library. *Seaborn* is a library in *Python* for making statistical graphs to understand and explore data.

Upon using the alpha parameter in the scatter plot, the points where maximum occurrences happen are darker. In the plot, shown in Fig. 2, the darker points are clearly visualized around $x = y$. This indicates that the grid position and the final position are approximately the same in majority of the cases. This makes perfect sense as the person who has an advantage in the grid position has less traffic ahead of him and hence, he has high chances of finishing near the same position from where he started the race.

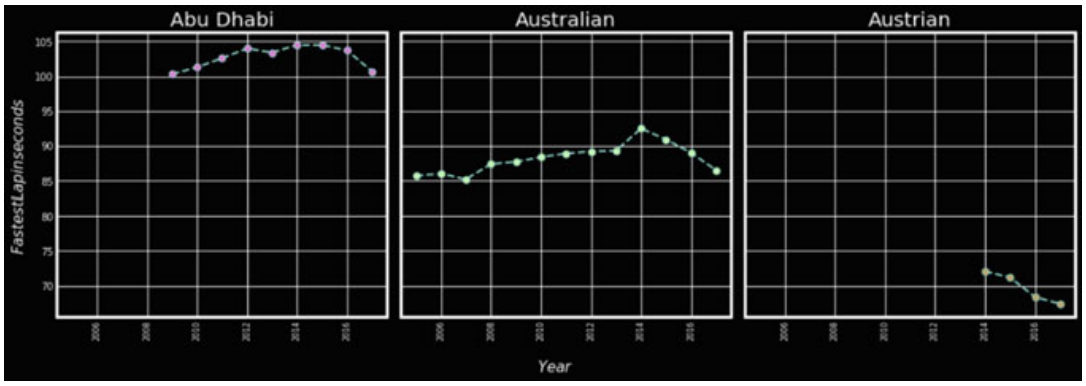

Fig. 1 Year wise plots of fastest lap times in different tracks

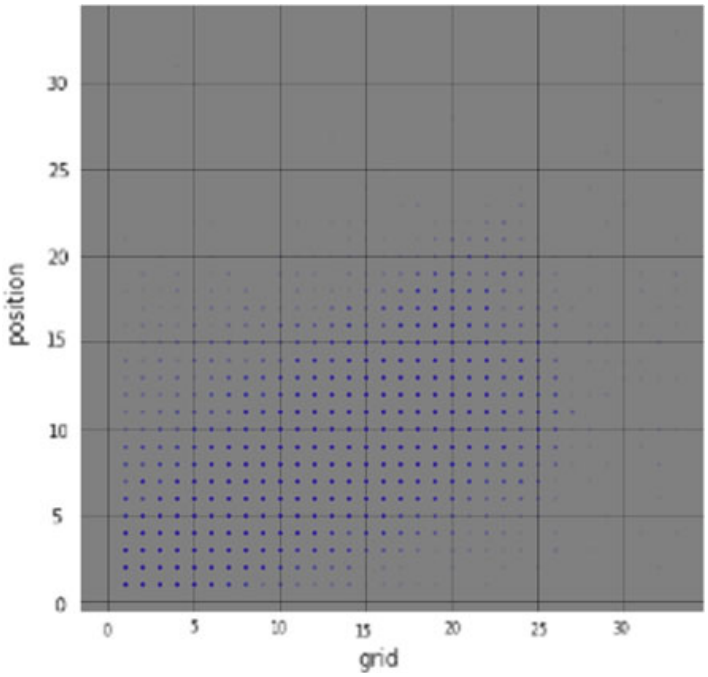

Fig. 2 Scatterplot for the grid versus position using alpha parameter to emphasize the majority of the values

Observation 3—Comparison

This is a comparative plot between two greats of the current era *Vettel* and *Hamilton*. The neck-to-neck competition of these two drivers is shown using two-line plots (one for each driver) on the same plot having year in the x axis and fastest lap time in the y axis. Upon observing the plots shown in Fig. 3, there is a clear change in performance of the drivers based on tracks. Such plots were generated for 26 different circuits. The performance depends upon factors that differ in a track like the ground force. Thus, a driver may have an advantage in a particular track over the other driver.

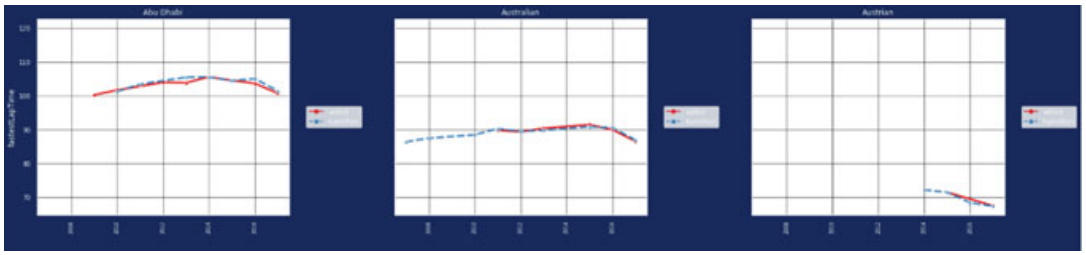

Fig. 3 Comparative performance of the two drivers in different tracks over the years

4.3 Phase III—Predicting the Winner of the Race

In Phase III, a machine learning model was designed to predict the winner of a race. This prediction was done using logistic regression. Regression is to determine the strength and character of relationship between a dependent variable and one more independent variable. There are different types of regression like linear regression, polynomial regression, logistic regression, and many more. Logistic regression is used when the dependent variable is of binary nature. This implies that the dependent variable falls into either of two specific categories. Hence, logistic regression was used in this experiment to determine who is going to win the next race. To perform logistic regression, the required datasets were collected from reliable sources.

If the driver and constructor are included in the set of independent variables, the model tends to be biased toward *Hamilton* and *Mercedes* as they have been very successful for the past 7 years. Hence, instead of taking these two variables, the current form of the driver and the current performance of his car were measured by taking into account the number of wins in the last 5 races and the number of podiums in the last 5 races.

Then after checking the correlation, only the attributes with higher correlation were chosen as dependent variables for the model. The attributes chosen for dependent variables were '*grid position*', '*number of wins in last 5 races*', '*number of podiums in last 5 races*', '*circuit*', and '*weather*'. The '*weather'* was classified as *wet* and *dry* as these are two extremes and tend to affect the result. '*Circuit'* is a categorical variable, and hence, it has to be encoded before using it in the regression model. The '*circuit*' was encoded using one hot encoding, and then, the model was trained.

After training the model, the correctness of the model was checked using the races in the year 2020. Upon observing the results of the prediction, it was found that the correct winner in a race was predicted in 11 out of the 17 races, held in 2020. This gives an accuracy of 64.7%. When the actual races were checked, it was found that the model still managed to predict winner. This prediction was successful if unforeseen circumstances like accidents and penalties were not considered.

5 Conclusion

In this paper, the prediction of accidents in an upcoming race, use of data visualization using plots to draw meaningful inferences and using logistic regression to predict the winner of a race has been presented. This study was performed using multiple datasets with the Formula 1 data from 1950 to 2020. The comparative study of models to predict the accidents was helpful in finding an ideal model to predict if accidents could occur in a race and thereby the model will positively influence the sport by enhancing the safety aspects. Using line plots, scatter plots, word cloud, and subplots, the data were interpreted. The meaningful conclusions obtained were

(i) there is no constant increase in speed of cars over the years. (ii) The grid position plays a deterministic role in the final race position. (iii) A comparative plot between two greats of the current era—Vettel and Hamilton shows that racers have advantages based on tracks. The machine learning model with a 64.7% accuracy predicted who wins the race based on the past performances. The accuracy is reduced to 64.7% as the model cannot take into account the unforeseen circumstances.

For future work, we plan to take more aspects into consideration like adding more features into the dataset which will lead to better accuracy in the model. The future work also has scope in predicting additional variables related to the sport like how often a driver crashes or get a penalty.

References

1. Tobias Lampprecht, David Salb, Marek Mauser, Huub van de Wetering, Michael Burch Tobias Lampprecht, David Salb, Marek Mauser, Huub van de Wetering, Michael Burch and Uwe Kloos, "Visual analysis of formula one races", in the proceeding of 23rd International Conference Information Visualisation, (2019).
2. Veronica Nigro, "Formula 1 race predictor", towards datascience, weblink: https://towardsdatascience.com/formula-1-race-predictor-5d4bfae887da, (2020).
3. Chinmay Wyawahare, "Formula 1 grand prix analysis", towards datascience, weblink: https://towardsdatascience.com/formula-1-grand-prix-analysis-d05d73b1e79c, (2020).
4. Depaolo C, Wilkinson K (2014) Get your head into the clouds: using word clouds for analyzing qualitative assessment data. TechTrends 58(3):38–44
5. Joanne Peng, Kuk Lida Lee and Gary M. Ingersoll, "An introduction to logistic regression analysis and reporting", The Journal of Educational Research, Vol. 96., No. 1., pp. 3–14., (2002).
6. Saishruthi Swaminathan, "Logistic regression — detailed overview", towards datascience, weblink: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc, (2018).
7. Wes Mckinney, "pandas: a foundational python library for data analysis and statistics", weblink: https://www.researchgate.net/publication/265194455_pandas_a_Foundational_Python_Library_for_Data_Analysis_and_Statistics/citations, (2011).
8. Rahul Raoniar, "Generate publication-ready plots using seaborn library", towards datascience, weblink: https://towardsdatascience.com/generate-publication-ready-plots-using-seaborn-library-part-1-f4c9a6d0489c, (2020).
9. Badreesh Shetty,"Data Visualization using Matplotlib", towards datascience, weblink: https://towardsdatascience.com/data-visualization-using-matplotlib-16f1aae5ce70, (2018).
10. Agathangelou B and Gascoyne M, "Aerodynamic design considerations of a Formula 1 racing car", SAE Technical Paper 980399, (1998).
11. Wright P, Matthews T (2001) Formula 1 Technology. Premier Series Books, SAE International
12. Tadi Aravind, Bhimavarapu Sasidhar Reddy, Sai Avinash and Jeyakumar G, "A comparative study on machine learning algorithms for predicting the placement information of under graduate students", In proceedings of Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), (2020).
13. Anjali Suresan, Divyaa Mahalakshmi G, Meenakshi Venkatraman, Shruthi Suresh and Supriya, "Comparison of machine learning algorithms for smart license number plate detection system", Image Processing and Capsule Networks, ICIPCN 2020 - Advances in Intelligent Systems and Computing, Vol 1200., (2021).