MSc. in Business Administration and Data Science 2022

Data Mining, Machine Learning, and Deep Learning

(CDSCO1004E) - Oral exam based on written product (IC)

Formula 1: Predicting Grand Prix Winners with Logistic Regression and Gradient Boosting

Final Paper

Submitted by:

Aleksandra Durska: 158370

Anh Dung Nguyen: 158391

Maria Zdanowicz: 158285

Yuechen Zhao: 158794

Submission deadline: 12:00 noon CEST 19/05/2023 (e-submission on Digital Exam)

Submission date: 19/05/2023

Supervisor name: Somnath Mazumdar

Number of pages: 15

Abstract

In recent years, Formula 1 has gained lots of popularity due to its highly unpredictable nature that generates a thrilling atmosphere and provides an exciting spectacle for the fans of motorsports. Since the results of each race can be affected by various factors, we decided to utilise supervised machine learning techniques to research the predictability. Therefore, we chose to examine how accurately Gradient Boosting algorithms can predict Formula 1 Grand Prix winners in the 2022 season compared to Logistic Regression models. We analysed the performance of Multinomial Logistic Regression, Ordinal Logistic Regression, Gradient Boosting Classifier and XGBoost in classifying a driver as a winner.

The dataset contained 16 columns and 9269 records representing race results and driver performance statistics from years 1999 - 2022. After thorough pre-processing, transformation, and visual exploration of the dataset, we trained and tested our four classifiers on records before 2022 and generated predictions for the 2022 season, we evaluated them using the accuracy, precision, recall and F1 scores as well as the running time.

Based on the research conducted, both Gradient Boosting classifiers performed significantly better compared to Logistic Regression models. Furthermore, XGBoost outstanded other classifiers in terms of computational efficiency. Since its accuracy was comparable to the Gradient Boosting Classifier, XGBoost appears to be the overall best classifier for predicting Formula 1 Grand Prix winners. The future work may involve exploring other variables that could be relevant to the problem, developing the Deep Neural Network (DNN), and utilising Support Vector Machines (SVM).

Keywords

Formula 1, Gradient boosting, GradientBoostingClassifier, XGBoost, Ordinal Logistic Regression, supervised learning, Multinomial Logistic Regression, Winner Prediction, Classification

1. Introduction

In light of the unpredictable nature of sports driven by many different variables that influence the results of a competition or a specific situation, intelligent software has gained popularity in the field of sports analytics. The introduction of Deep Learning (DL) and Machine Learning (ML) approaches has resulted in advancements in many sports, such as Formula 1. It continues to be a sports innovation pioneer, with a cooperation between Formula 1 and Amazon Web Services (AWS), bringing data-driven solutions and insights that have revolutionised the racing industry (Amazon, 2023). The use of ML approaches and High-Performance Cloud Computing (HPCC) have further enhanced race strategic decision-making and car effectiveness in terms of aerodynamics and downforce loss.

In addition, these improvements have made it possible for racing teams to do thorough opponent analysis, as well as car and driver evaluation, in order to exploit flaws and develop ideal driving approaches. Jenkins and Floyd (2001) pioneered in the examination of technology trajectories and trends in the sport, advocating particular case studies centred on Formula 1 teams for rapid advancement. Apart from specific internal data, there are many different external factors, such as human error, mechanical failure, weather or track condition that could affect the race results.

Ultimately, the use of artificial intelligence and other machine learning approaches in sports analytics, has shaped the way Formula 1 is practised and managed, resulting in substantial advances in vehicle performance metrics, competitiveness, and race strategy decision-making processes.

2. Motivation

Reasoning for forecasting Formula 1 champions can be approached from a variety of aspects. Estimating sports standings and results is a challenging and exciting study issue that involves the development and application of complex approaches and models. As a result, the purpose of this study is to investigate, explain, and compare supervised learning algorithms capable of effectively projecting racing winners in the 2022 season of Formula 1. The outcomes of this study will not only benefit the scientific community by enhancing current understanding of ranking prediction, but will also serve as a practical basis for future research in similar disciplines.

Within Formula 1, forecasts of racing winners and provisional rankings can aid teams and drivers by giving them useful data to enhance their performance. Determining the possibility of winning a given race or ending in a specific place, for instance, can assist teams and drivers in developing successful strategies to attain their objectives. Furthermore, precise forecasts of provisional standings can assist teams in making educated resource allocation decisions, such as where to focus on improvement or which races to prioritise.

In addition, betting on sporting events is becoming a popular economic activity, and many people are employing predictive analytics to get an upper hand over bookmakers. Formula 1 is no exception, since the sport's competitiveness and uncertainty make it an interesting choice for those who bet. People that engage in sports betting can get helpful insights and make more informed decisions that could possibly end up in cash rewards by accurately predicting race winners and provisional rankings. As a result, this research can help to design profitable methods for betting on sports.

In conclusion, estimating the winners and standings in Formula 1 is a difficult study issue that can assist academics, sports betting fans, and the competing teams and drivers themselves. The findings from this study, by properly forecasting race results, can help to design profitable sports betting strategies as well as offer useful information that will assist teams and drivers further develop their performance.

3. Research question

The aim of this report is to establish which supervised machine learning model is the most likely to accurately predict the winner of each Formula 1 Grand Prix in the 2022 season. The four models covered by this study are Multinomial Logistic Regression, Ordinal Logistic Regression, Gradient Boosting Classifier, and XGBoost, which were evaluated in terms of accuracy score and precision.

4. Related work

Extensive scientific study has been conducted in the subject matter of sports analytics, offering a solid foundation for addressing research problems having a high degree of credibility. Bunker and Thabtah (2019) provided an empirical foundation for forecasting sports winners via unsupervised learning techniques that outperformed domain experts. But supervised models also play a significant role in making sport predictions. Ofoghi et al. (2010) created a framework that closely matches our project's goal. Their research revolves around supervised learning, specifically using Naive Bayes Classifier to categorise instances into separate groups that indicate final winners and standings. Their method was successfully adapted to omnium cycling, a sport with a structure comparable to Formula 1.

Additionally, Haghighat et al. (2013) examined data mining methods for predicting results across different sports, emphasising the need of acquiring relevant data from legitimate sport websites via online scraping approaches. To improve predicted accuracy, hybrid modelling strategies such as ensemble methods were suggested. Drucker (1997) presented Gradient Boosting as an algorithm, which has significant potential as a model selection tool. Gradient boosting combines groups of weak learners to build stronger models while minimising training errors. Moreover, Horvat and Job (2020) most recently highlighted the necessity of feature selection ahead of model implementation, suggesting a general strategy based on expert experience and traditional approaches such as PCA or iterative feature reduction to decrease dimensionality and increase prediction scores.

Eventually, the academic field of sports analytics provides a solid foundation for investigating research issues through the use of unsupervised as well as supervised machine learning techniques.

5. Methodology

5. 1. Dataset description:

On Kaggle we have found a data source that has needed variables and a possibility of creating ones that we thought were needed for the models. In the appendix, a description and properties of all selected datasets can be found.

5. 2. Data pre-processing and filtering:

We have selected five separate datasets: ‘driver_details’, ‘race_details, ‘driver_standings’, ‘constructor_standings; and ‘starting_position’. The name of the variable ‘Pos’ in the starting_positon dataset was changed for further use. We performed outer merges on the datasets, firstly, driver_standings and race_details, where we merged them on ‘Driver Code’, ‘Driver’ and ‘PTS’ as we have seen some of the repeating ‘Driver Codes’. Secondly, we have merged that dataset with ‘driver_details’ based on ‘Driver’, ‘Year’ and ‘Grand Prix’. Then renamed some of the variables and dropped the not needed ones. Next, we sorted the data in an ascending order in terms of the “Year” variable and output was exported to Excel. Years before 1998 were dropped for efficiency since those drivers were no longer competing. Missing values were found in multiple columns.

Next, we have adjusted the starting position so it does not contain missing values so we adjusted it for the last place and also created a function which changes the ‘Time/Retired’ variable into a numerical one, in that function we account for ‘+ (x) laps’ type of values and ‘DNF’ cases. Another variable that was added was the number of Grand Prix each driver has participated in the truncated data set. We have also decided to drop ‘Race Position’ and ‘Driver Code’ which were duplicated. Another outer join was performed on the existing dataset with ‘constructor_standings’ as we considered it crucial to train the model with the ‘Team Position’. Unnecessary variables were dropped and renamed. After careful consideration we have also created a ‘Total_Points_Per_Season’ variable for each driver. The rows in the dataset that were showing these values were removed at that point. Moreover, we have also dropped any row that had missing values in more than 3 features. Next variable created was ‘Pos_Change’ which as the name suggests reflects the amount of places a driver has gained/lost throughout the race. In the case of non-numerical variables we have assigned 24 as a finish position. It was done for the training purposes to represent a non-finished race and punish the driver in the predictions. At last, ‘Won’ variable was created, which is a binary and represents whether the driver has won that race. Because of efficiency reasons we have exported the created dataset to excel and adjusted the few missing columns in the ‘Date’ variable.

In the latest stage of processing the data we have identified and removed the duplicates. Moreover, we have corrected variables that were incorrectly coded before exporting the data. We transformed the ‘Time’ variable into a numerical one by deleting ‘+ … s’ from the data points, both in terms of laps and time measured. We have also transformed ‘Team_Pos’ by assigning position 13 to the team that was excluded from classification in 2007. The 'Won' function was adjusted. In the ‘Pos’ variable we have also converted the non-numerical inputs into position 25 as they represent drivers who did not finish, were not classified, etc. Lastly, we have reset the index of the data source.

5.3. Data Exploration:

After the pre-processing, we conducted a few exploratory analyses on the dataset to get a general idea of the distributions of essential variables. Firstly, two bar charts of the total points (Figure 1) and the total number of wins for each driver (Figure 2) are plotted and shown below.

From these two plots we can see that there are only slight differences between number of wins and number of total points per driver. For example, Michael Schumacher ranks second for number of wins, however, the total points earned is significantly lower. This could be due to the fact that his races spanned from 1991 to 2012, during which the point system underwent several changes, impacting the points accumulated.

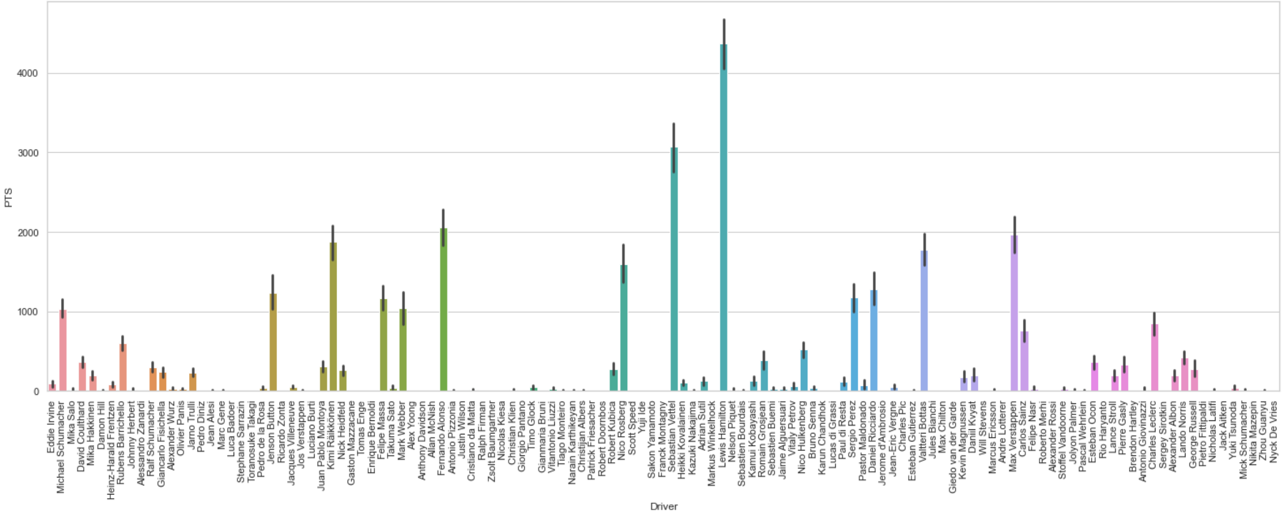


Figure 1: Bar Chart of Total Points Earned by Each Driver

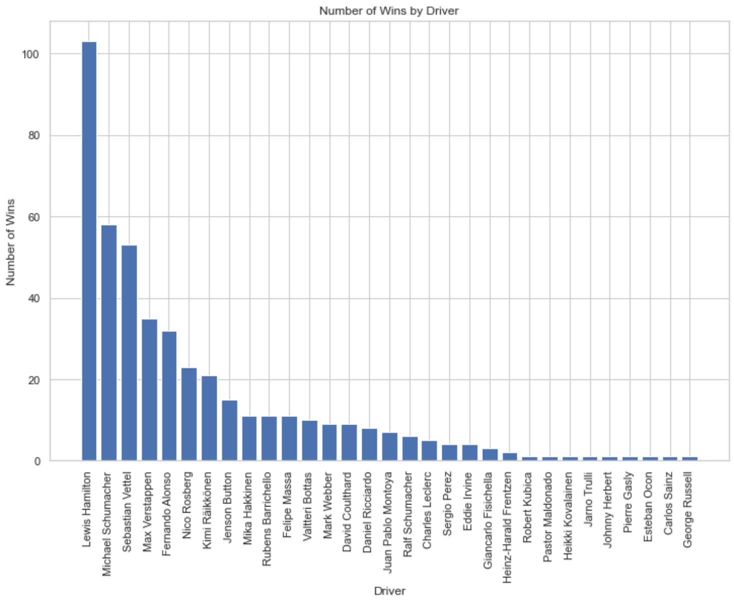
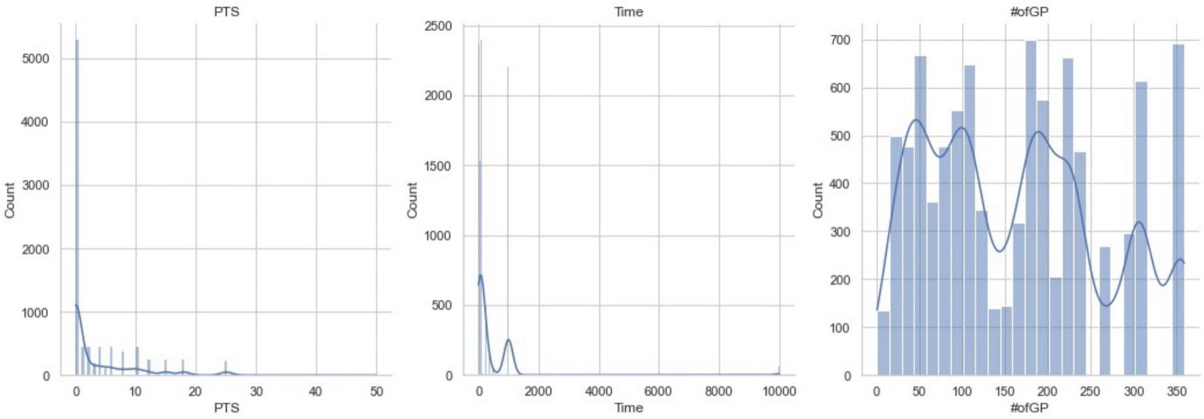


Figure 2: Bar Chart of Number of Wins by Each Driver

We also plotted the distributions for each driver per race: points earned ("PTS"), finished time ("Time"), number of Grand Prixes the driver has participated in ("#ofGP"), total points earned per season ("Total_Points_Per_Season"), position change after finishing the race ("Pos_Change"), and final position ("Pos"). Plots for these distribution are illustrated in Figure 3:



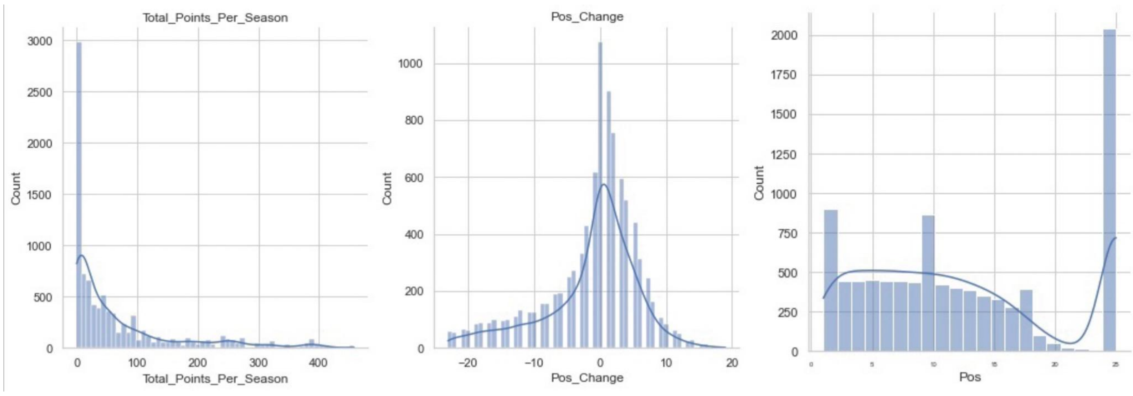


Figure 3: Distribution graph of points earned ("PTS"), finished time ("Time"), number of Grand Prixes the driver has participated in ("#ofGP"), total points earned per season ("Total_Points_Per_Season"), position change after finishing the race ("Pos_Change"), and final position ("Pos") for each driver per race

From the distribution plots we can derive some general ideas about the dataset and its background. Firstly, it is normal for drivers to receive 0 points for a race since only 10 best drivers receive points in each race, and that also leads to the left-skewed distribution for total points per season for each driver. The position change distribution roughly follows a normal distribution around 0, which means normally the starting position and the final position are the same or very close, indicating the significance of a good starting position. The distribution of final position also shows that it is very often for drivers to not complete the race due to possible mechanical or other issues.

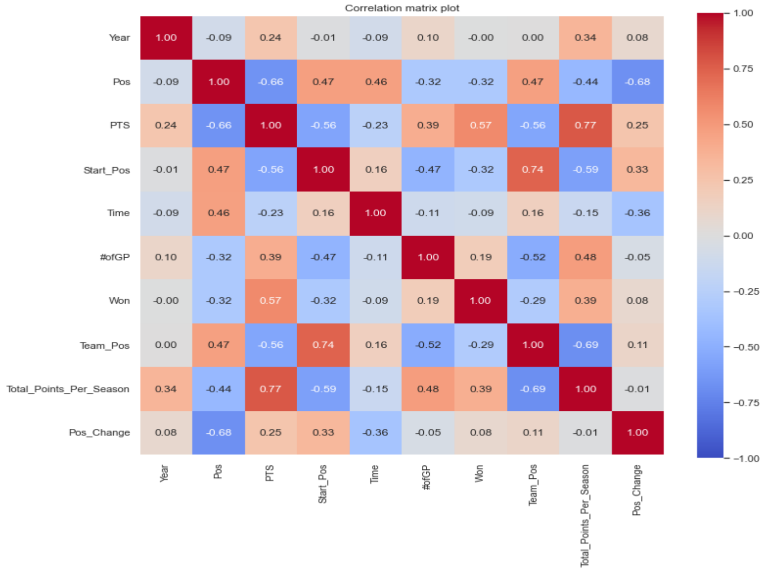


Figure 4: Correlation Matrix of Variables

Lastly, the correlation matrix is plotted and shown in Figure 4. From the correlation matrix we can observe some significant positive correlations between total points per season and points for each driver per race, which is intuitive because we calculated the former based on the latter. The same goes for position change and final position, as well as final position and points earned, except for the correlation is negative. There is also a notable positive correlation between start position and team points, because start positions are usually determined by qualifications prior to the actual race and drivers who take better start positions are more likely to earn points for their teams.

5.4. Data analytics: Modelling, methods and tools

We have decided to implement four models to predict a winner of each Grand Prix in the 2022 season. Two of them are different types of Logistic Regressions and the other two are different types of Gradient Boosting. Based on the results, we select a driver with the highest probability of winning each of the races during that season.

We trained and tested the Logistic Regression models on the 1999 - 2021 data, which was randomly divided into the training set (80%) and the test set (20%) for logistic regressions, and training set (70%), validation set (15%) and test set (15%) for Gradient Boosting. This is because adding a validation set improved the performance of two boosting algorithms, but not the case for regressions. For the purpose of prediction, we used the 2022 data, as the model had not been exposed to it before.

5.4.1. Multinomial Logistic Regression

Applying Multinomial Logistic Regression is an applicable solution when predicting the winners of each race. This specific classification algorithm is based on the assumption that there are more than two possible outcomes and that it performs well in this type of environment (Hoffman, 2019).

This model is based on the standard implementation of the Logistic Regression from the sklearn.linear_model library, with the specification for the multinomial regression. The solver chosen is ‘lbfgs’, it is well suited for problems with a large number of features which are present in our model. However, we have also tested ‘sag’ solver and the results have not improved at all. On top of that, we have tested the code with different amounts of iterations and the results did not change, which leads us to a conclusion that the model is reaching the prediction before reaching the maximal amount. Therefore, we have decided to input 1000. Moreover, the code has also implemented ‘l2’ penalty, and we have tested different scales of the C, regulatory parameter, and arrived at a conclusion that the strength over 4 does not improve the results or accuracy any further and we have selected C=5 for a minimally better accuracy. In the model itself we have also reversed ‘Start_Pos’, ‘Time’ and ‘Team_Pos’ because in the original state the code was punishing the drivers with the lower values in these variables. The variables 'Driver’ and ‘Grand Prix’ are being encoded. We have also decided to drop the ‘Date’ variable as we have seen it does not have any valuable input in the code. Moreover, the (encoded) variable ‘Car’ was also dropped as the names of the teams change, sometimes even during the season, so there is no long term training provided.

5.4.2. Ordinal logistic regression

The Ordinal Logistic Regression is a classification method that can be used to examine the relationship between one hierarchical variable with multiple category levels and many explanatory variables (Parry, 2020). Our goal of implementing the Ordinal Logistic Regression was to prepare a classifier that predicts the final position of each driver in the Formula 1 Grand Prix and therefore, reduces the risk of overestimating the probability of winning that may happen if the binary logistic classifier was used instead. The output of the model are the probabilities of taking each final position, ranging from 1 to 25, for each driver. In the next step, we select the highest probability of finishing first, which is then linked to the driver whose name becomes a predicted label for a particular Grand Prix.

To create the model, we used OrderedModel from statsmodels.miscmodels.ordinal_model module in statsmodels library. We applied ‘logit’ as the distribution type and ‘bfgs’ as the fitting algorithm with 500 iterations. We chose the final position (Pos) as the ordinal variable Y, which has 20 possible variants J in 2022 (Parry, 2020). Initially, we set ‘Year’, ‘PTS’, ‘Start_Pos’, ‘Time’, ‘Pos_Change’, ‘#ofGP’, ‘Total_Points_Per_Season’, ‘Team_Pos’, ‘Driver_Encoded’, ‘Car_Encoded’, ‘GP_Encoded’ as explanatory variables. As logistic regression models may overfit when multiple explanatory variables are included in the model (Géron, 2019, p. 29), we experimented with removing some explanatory variables to observe if the z-score of the remaining ones is improving. As a result, we decided to drop ‘#ofGP’ to tune the model and continue with the following explanatory variables: ‘Year’ (x_1), ‘PTS’ (x_2), ‘Start_Pos’ (x_3), ‘Time’ (x_4), ‘Pos_Change’ (x_5), ‘Total_Points_Per_Season’ (x_6), ‘Team_Pos’ (x_7), ‘Driver_Encoded’ (x_8), ‘Car_Encoded’ (x_9), ‘GP_Encoded’ (x_{10}). The notation of the Model 1 follows the one proposed by Parry (2020).

$$log(\frac{P(Y \leq j)}{P\ (Y > j)}) \ = \ log(\frac{P(Y \leq j)}{1 - P\ (Y \leq j)}) \ = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{10} x_{10}$$

where j = 1, 2, … J - 1

Model 1: The notation of the Ordinal Logistic Regression Model

5.4.3. Gradient Boosting and XGBoost

We will use two Gradient Boosting algorithms in order to classify the final positions of each race for the 2022 Formula 1 season, namely traditional Gradient Boosting and XGBoost. The features chosen to train the model were: ‘PTS’, ‘DriverCode’, ‘Start_Pos’, ‘#ofGP’, ‘Team_Pos’, ‘Total_Points_Per_Season’, ‘Pos_Change’, ‘GP_Code’, ‘Car_Code’. Compared to the regression models, we aimed to avoid indicating straightly who are the winners, by not including ‘Won’ and ‘Time' variables.

The Gradient Boosting method combines several weak learners, often decision trees, to produce a robust classification or prediction model. It uses a Gradient Boosting framework in which each weak learner sequentially corrects the errors and overlooked patterns of preceding learners (Géron, 2019). XGBoost is an advanced implementation of Gradient Boost which performs significantly faster than many existing algorithms with better scalability. This is realised by innovations in both system and algorithm, such as parallel and distributed computing and a learning algorithm dedicated to handle sparse data (Chen & Guestrin, 2016). In our work, we decided to create two models, based on the built-in functions: GradientBoostingClassifier and XGBoost from sklearn.ensemble package.

After labelling and splitting the data, we defined the models. In the first model we set: n_estimators=100 (number of weak learners), learning rate=0.1 (it controls the contribution of each weak learner in the overall ensemble. The lower learning rate, the slower learning, but it can improve generalisation), max_depth=3 (higher values can result in more complex models and may make them prone to overfitting) and random_state=0. This hyperparameter ensures reproducibility of results. By setting it to 0, we will get identical results every time the code is performed. In the case of the XGBoost model, based on the nature of this study, the objective is set to “multi:softmax” and the evaluation metric is “mlogloss”. We chose the same maximum depth since it didn’t influence the performance significantly, while for learning rate, we decided on the 0.7 since it produces the highest accuracy.

Finally, to explore and compare the robustness of both models, we ran the classification reports that allowed us to retrieve validation metrics across different classification classes.

5.5. Model complexity analysis

As shown in Table 1, the best computing time was achieved by XGBoost because of its innovations in both system and algorithm.

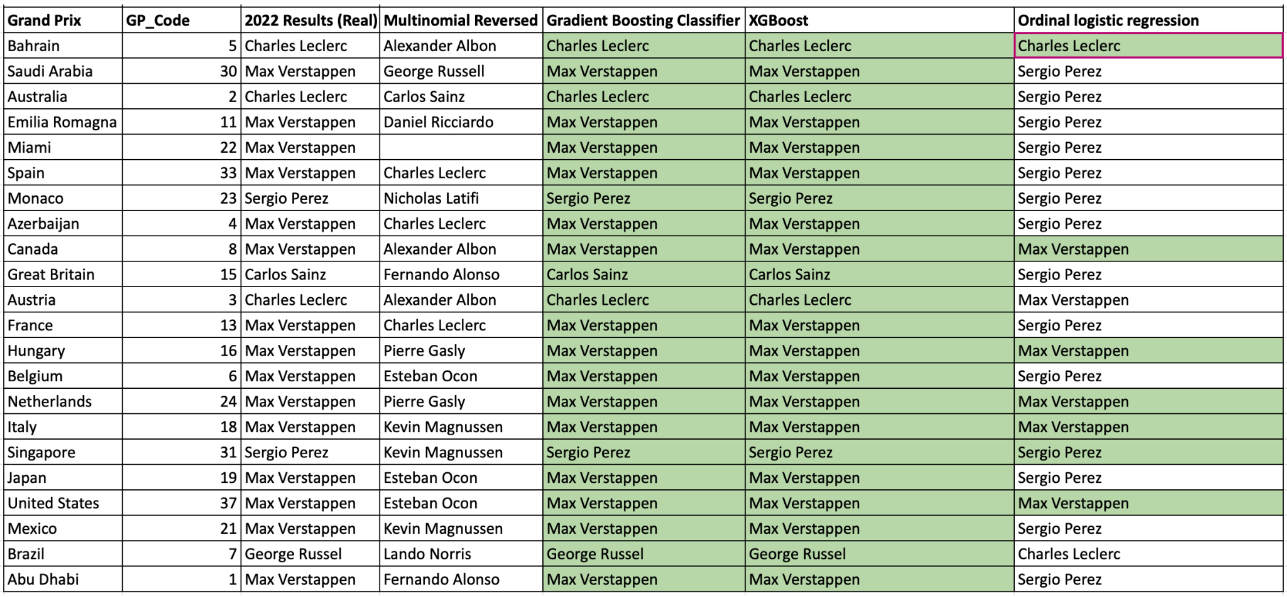
Model	Multinomial Regression	Ordinal logistic regression	Gradient Boosting Classifier	XGBoost
Running time of training the model	4:52:02 s	17:18:35 s	15:61:03 s	0:24:13 s
Elapsed time of producing the prediction	0.0207 s	0.047 s	0.042 s	0.0016s

Table 1 Models’ Running Time

6. Results

After testing all the models on the test sets, we also decided to compare the four models against the actual results of the 2022 season. The final results are presented in Table 2. The accuracy and precision, recall, F1-score of the prediction on the test sets as well as the accuracy of the 2022 season prediction of each model are compared and shown in Table 3. As it can be seen the Gradient Boosting and XGBoost are performing the best with all of the predictions being correct. The Logistic Regressions are performing poorly compared to the GBs. In the Ordinal Logistic Regression there are 32% of the predictions correct, while the Multinomial does not predict any of the winners correctly. However, when looking at further places there are drivers who actually won predicted in the Top 5 in certain cases. The pattern of people like Alexander Albon or Kevin Magnussen being predicted to be winners gives us a reason to believe that the Pos_Change variable has been influential in the code quite significantly.

The best scores are reached by the Gradient Boosting Classifier in all categories, however, the XGBoost is also performing very well and reaches values of over 80%. The Multinomial Regression is performing slightly better in terms of the evaluation metrics, having values over 70%, where the Ordinal Regression has values of 65% and above.



Grand Prix	GP_Code	2022 Results (Real)	Multinomial Reversed	Gradient Boosting Classifier	XGBoost	Ordinal logistic regression
Bahrain	5	Charles Leclerc	Alexander Albon	Charles Leclerc	Charles Leclerc	Charles Leclerc
Saudi Arabia	30	Max Verstappen	George Russell	Max Verstappen	Max Verstappen	Sergio Perez
Australia	2	Charles Leclerc	Carlos Sainz	Charles Leclerc	Charles Leclerc	Sergio Perez
Emilia Romagna	11	Max Verstappen	Daniel Ricciardo	Max Verstappen	Max Verstappen	Sergio Perez
Miami	22	Max Verstappen		Max Verstappen	Max Verstappen	Sergio Perez
Spain	33	Max Verstappen	Charles Leclerc	Max Verstappen	Max Verstappen	Sergio Perez
Monaco	23	Sergio Perez	Nicholas Latifi	Sergio Perez	Sergio Perez	Sergio Perez
Azerbaijan	4	Max Verstappen	Charles Leclerc	Max Verstappen	Max Verstappen	Sergio Perez
Canada	8	Max Verstappen	Alexander Albon	Max Verstappen	Max Verstappen	Max Verstappen
Great Britain	15	Carlos Sainz	Fernando Alonso	Carlos Sainz	Carlos Sainz	Sergio Perez
Austria	3	Charles Leclerc	Alexander Albon	Charles Leclerc	Charles Leclerc	Max Verstappen
France	13	Max Verstappen	Charles Leclerc	Max Verstappen	Max Verstappen	Sergio Perez
Hungary	16	Max Verstappen	Pierre Gasly	Max Verstappen	Max Verstappen	Max Verstappen
Belgium	6	Max Verstappen	Esteban Ocon	Max Verstappen	Max Verstappen	Sergio Perez
Netherlands	24	Max Verstappen	Pierre Gasly	Max Verstappen	Max Verstappen	Max Verstappen
Italy	18	Max Verstappen	Kevin Magnussen	Max Verstappen	Max Verstappen	Max Verstappen
Singapore	31	Sergio Perez	Kevin Magnussen	Sergio Perez	Sergio Perez	Sergio Perez
Japan	19	Max Verstappen	Esteban Ocon	Max Verstappen	Max Verstappen	Sergio Perez
United States	37	Max Verstappen	Esteban Ocon	Max Verstappen	Max Verstappen	Max Verstappen
Mexico	21	Max Verstappen	Kevin Magnussen	Max Verstappen	Max Verstappen	Sergio Perez
Brazil	7	George Russel	Lando Norris	George Russel	George Russel	Charles Leclerc
Abu Dhabi	1	Max Verstappen	Fernando Alonso	Max Verstappen	Max Verstappen	Sergio Perez

Table 2: Prediction results

Model	Multinomial Regression	Ordinal logistic regression	Gradient Boosting Classifier	XGBoost
Accuracy on Test Set	0.7361	0.7692	0.8524	0.8116
Precision	0.72	0.7	0.85	0.82
Recall	0.73	0.65	0.85	0.81
F1-Score	0.73	0.65	0.85	0.81
Accuracy on the 2022 dataset	0	0.32	1.00	1.00

Table 3: Models' Accuracy

The linearity of predictions in the case of the regressions can be an explanation of why their performance is much lower than the other type of classifiers. Moreover, the main reason for overly high accuracy for the predictions of classification algorithms on the 2022 dataset is that we only considered the winners in this case. During the training, the accuracy for predicting first position is also 100% for both classification algorithms, and overall excellent for top 5 positions. This could be due to the underlying information leading to the actual winner in the variables that we failed to discover and eliminate, or the fact that more variables are needed for the algorithms to build more comprehensive models. Also the explanatory variables that we were able to find are correlated to some extent, when they are expected to be independent. Because of the black-box nature of Gradient Boosting and our limited understanding of the model, it is difficult to interpret which features are important to the prediction and what can be added to improve. Nevertheless, Gradient Boosting Classifier and XGBoost, are proven to be a wise choice when it comes to predicting F1 race winners, compared to regressions.

7. Discussion

Overall, Gradient Boosting seems like the most suitable approach for predicting the winners of Formula 1, namely the XGBoost model. It is also in line with the academic literature. We decided to choose this model as it resulted in similar values in terms of evaluation metrics to the Gradient Boosting Classifier, but achieved faster running time. Additionally, the regression models are not performing as well, with only a few wins being predicted correctly.

However, there are various limitations and constraints to consider when employing Gradient Boosting and regressions for Formula 1 race forecasts. Insufficient number of variables, may limit the models' capacity to capture all the complexities of races. Furthermore, the relationships between various aspects in Formula 1, like driver’s skills, car performance, and weather conditions, can be nonlinear, which linear models may not effectively reflect. Also, labelling characteristics “DNF”, "NC”, "DQ”, and "EX" (indicating disqualification or not being counted) as 25, might cause data imbalance and bias in predictions. This imbalance may have an impact on the models' capacity to generalise well. Also, limiting projections to a particular type of model reduces general predictive power because different models have varied strengths and weaknesses. In addition, due to constraints in data collecting, variability in different sources, and potential discrepancies, data quality might have been affected. Finally, Gradient Boosting and regressions can be difficult to understand, making it difficult to figure out how individual attributes contribute to predictions. Also, Gradient Boosting models may be prone to overfit and fail to generalise successfully new data, as in our case they resulted in 100% accuracy in predicting the winners of 2022 season.

8. Conclusion and Future Work

Due to the dynamic nature of Formula 1, accurately predicting the winner of each Grand Prix is a highly challenging task. Various unpredictable factors, such as weather conditions, mechanical failure, or human errors can significantly influence results of the race. In order to predict Formula 1 Grand Prix winners in the 2022 season, we examined the accuracy and performance of Logistic Regression and Gradient Boosting classifiers. The study revealed that Gradient Boosting is more suitable for this task, as we managed to achieve 100% accuracy in the prediction with the Gradient Boosting Classifier and XGBoost algorithm. Although the Gradient Boosting Classifier reached the highest accuracy, precision, recall, and F1-scores, XGBoost was far more effective in terms of computational efficiency. In conclusion, XGBoost seems to be the best model overall.

There are several possibilities to improve the current approach in the future work. Firstly, it could be beneficial to explore additional variables to identify ones with higher relevance to the problem. This can involve variables representing drivers' fastest lap, best qualifying time, crash likelihood, teams' position in constructor standings and the average pit stop time. Additionally, exploring different models such Deep Neural Networks, Support Vector Machines and other ensemble methods may provide better results, as they capture more complex classification problems.

References:

Amazon.com, (2023), Featured AWS Sports and Entertainment partnerships, accessed (04.05.2023), https://aws.amazon.com/sports/f1/

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. Applied computing and informatics, 15(1), 27–33.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Drucker, H. (1997). Improving regressors using boosting techniques. Monmouth University, NJ07764.

Formula 1 Official Data (1950-2022). (2022, November 30). Kaggle. https://www.kaggle.com/datasets/debashish311601/formula-1-official-data-19502022?resource=download

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media, Inc.

Haghighat, M., Rastegari, H., Nourafza, N., Branch, N., & Esfahan, I. (2013). A review of data mining techniques for result prediction in sports. Advances in Computer Science: an International Journal, 2(5), 7–12.

Hoffman, J. I. E. (2019). Logistic Regression. In *Elsevier eBooks* (pp. 581–589). https://doi.org/10.1016/b978-0-12-817084-7.00033-4

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(5), e1380.

Jenkins, M., & Floyd, S. (2001). Trajectories in the evolution of technology: A multi-level study of competition in formula 1 racing. Organization studies, 22(6), 945–969.

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Dwyer, D. (2010). A machine learning approach to predicting winning patterns in track cycling omnium. In Ifip international conference on artificial intelligence in theory and practice (pp. 67–76).

Parry, S. (2016, June). *Ordinal Logistic Regression models and Statistical Software: What You Need to Know, Statnews #91* (2020, August). Cornell Statistical Consulting Unit. https://cscu.cornell.edu/wp-content/uploads/91_ordlogistic.pdf

Appendix

variable name	**description**
Driver	The name of the driver
Grand Prix	The name of the Grand Prix
Date	The date of the Grand Prix
Year	The year of the Grand Prix
Pos	The ending position in the Grand Prix
Time/Retired	Time required to finish the race
PTS	Points earned during the race
DriverCode	The code of the driver
Car	The name of the team the driver belongs to
Start_Pos	The starting position of the driver during the race
Time	Time required to finish the race; Time/Retired adjusted
#ofGP	The number of Grands Prix in which the driver competed
Won	The fact of winning (1) or not winning (0) in the Grand Prix
Team_Pos	The position of the team
Total_Points_Per Season	Points earned during the season
Pos_Change	The number of positions gained/lost during the race

Name of the dataset	Variables and information
driver_details	0 Car 19807 non-null object 1 Date 19814 non-null object 2 Driver 19814 non-null object 3 Grand Prix 19814 non-null object 4 PTS 19804 non-null float64 5 Race Position 19798 non-null object 6 Year 19814 non-null int64
driver_standings	0 Pos 1618 non-null object 1 Driver 1618 non-null object 2 Nationality 1618 non-null object 3 Car 1607 non-null object 4 PTS 1618 non-null float64 5 DriverCode 1618 non-null object 6 Year 1618 non-null int64
starting_postion	0 Car 22527 non-null object 1 Detail 22529 non-null object 2 Driver 22529 non-null object 3 DriverCode 22529 non-null object 4 Grand Prix 22529 non-null object 5 No 22529 non-null int64 6 Pos 22529 non-null int64 7 Time 15657 non-null object 8 Year 22529 non-null int64
race_details	0 Pos 23978 non-null object 1 No 23978 non-null int64 2 Driver 23978 non-null object 3 Car 23952 non-null object 4 Laps 23771 non-null float64 5 Time/Retired 23970 non-null object 6 PTS 23978 non-null float64 7 Year 23978 non-null int64 8 Grand Prix 23978 non-null object 9 Detail 23978 non-null object 10 DriverCode 23978 non-null object
constructor_standings	0 Pos 675 non-null object 1 Team 675 non-null object 2 PTS 675 non-null float64 3 Year 675 non-null int64