Predicting F1 Pilots' Performance

Maurice Chiu
chiu93@purdue.edu
Purdue Computer Science
West Lafayette, Indiana, USA

Nishtha Aggarwal
aggarwn@purdue.edu
Purdue Computer Science
West Lafayette, Indiana, USA

Daniel Castro
dcastrom@purdue.edu
Purdue Computer Science
West Lafayette, Indiana, USA

Jiangqiong Liu
liu3328@purdue.edu
Purdue Computer Science
West Lafayette, Indiana, USA

Mayesha Monjur
monjur@purdue.edu
Purdue Computer Science
West Lafayette, Indiana, USA

ABSTRACT

Formula 1 (F1) is a highly competitive motorsport characterized by complex interactions among factors such as driver skill, team performance, car specifications, track conditions, and weather. Accurate prediction of F1 race winners is of significant interest to racing enthusiasts, teams, and betting communities. This study presents a comprehensive comparative analysis of various machine learning models employed for F1 race winner prediction, including Naive Bayes Classifier, Logistic Regression, and Neural Network.

ACM Reference Format:
Maurice Chiu, Nishtha Aggarwal, Daniel Castro, Jiangqiong Liu, and Mayesha Monjur. 2024. Predicting F1 Pilots' Performance. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Predicting the F1 champion based on the data at hand can be framed as either a classification or a regression problem. Our aim is to predict the performance of the drivers during each race of a season, and consequently, the season champion. As we intend to implement models learned during this course, we have focused on supervised learning methods. Using these models, we predicted the top 10 positions for each race throughout an entire season and identified the top podium finishers for each season (2015-2023). Given the stochastic nature of the problem and the multitude of factors influencing the outcome, probabilistic models could be particularly advantageous. In this project, we have compared different models using multiple metrics. The list of models we employed includes the following:

- Naive Bayes Classifier:
  Naive Bayes Classifier (NBC) is a supervised machine learning algorithm, that seeks to model the distribution of inputs of a given class or category. This model is characterized by its "naive" assumption of conditional independence in predictions, and for the equal contribution assumption of all features to the outcome. It usually performs well with small sample sizes.
- Logistic Regression:
  Logistic Regression is also another supervised machine learning algorithm. This algorithm is used to predict a dependent categorical target variable. In this model, our objective is to consider multiple classes into which an item can be classified, which are ordered. This approach will assist us in predicting the top 10 drivers of each race.
- Neural Network:
  Deep Learning techniques, such as Multi-Layer Perceptrons (MLP) is suitable given the considerable amount of data we have. MLPs, a class of feed-forward artificial neural networks can effectively capture non-linear interactions, adding to our model's robustness. These methods can capture complex relationships and patterns in the data. In F1 racing predictions, various factors such as driver skills, team strategies, car performance, and track conditions play significant roles. MLPs can analyze historical data and learn the intricate patterns that often dictate race outcomes.

Our goal for this project was to produce a standings table for each race by associating each racer with the points earned by their position. Using these tables, we determined who would win the 2023 championship, and we compared our results with the actual data to assess our model's performance by comparing our predicted rankings against the actual rankings.

GitHub Repository: https://github.com/MauriceChiu7/573-f1

2 PRIOR WORK

The realm of sports analytics has seen substantial evolution with the use of diverse analytical tools to forecast and examine performance results. Given its complexity, Formula 1 racing makes a suitable candidate for such strategies.

The first paper [5] talks about using Artificial Neural Network (ANN) as a tool for sports result prediction. Even though it doesn't really provide any specific algorithm, it gave us insight into general framework of ML based sports-predictor tools.

The use of AI in sports increased dramatically in 2021, and research increasingly incorporated deep learning methods. Although not specifically focused on F1 racing, studies like those by Zhou et al. (2021) [8] on applying deep learning to forecast sports results may provide approaches and frameworks that can be applied to motorsport settings.

As we know, there are various types of racing related sports, each with their own unique twists and each vastly different between them. However, they all share some common concepts (i.e. all sports racing involves some form of competitors competing for the top spots and the probability of a competitor winning the race depends

Conference'17, July 2017, Washington, DC, USA
2024. ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Data collected from Ergast API

Race Results	Qualifying Results
Driver Standings	Driver Experience
Constructor Information	Circuit Information

on skills/experience, and the race environment/conditions). Therefore, we also looked into other specific ML-based racing-sports predicting researches like [7], which describes a ML based Learn-to-Rank approach for Road-cycling race, and [6], which describes a ANN based approach to predict horse-racing results. These papers gave us valuable insight into exactly how to approach these type of problems and how to model and solve them. We learned that for racing-sport prediction problems, data analysis and feature engineering is more important than model construction. Engineering effective feature-sets and creating a high-quality dataset contributes for a significant portion of the effort required to address the entire problem. In cases like these, poor feature sets and noisy data present challenging obstacles for even the most excellent models.

Furthermore, a study done by Bopaiah and Samuel [4] used 2019 regulations as a baseline to optimize Formula One (F1) car performance for 2021. It highlighted important performance indicators including MGU-K deployment speed using actual driver data and engine-powertrain modelling. The model's accuracy was validated against On-Board films and International Automobile Federation (FIA) regulations, which produced suggested tactics for better race and qualification outcomes.

In conclusion, there are many different variables that affect how well F1 racing performances are predicted and analyzed, ranging from race strategy to car engineering. The literature cited above provides a framework for understanding these processes, but there is still much room for research in this area.

3 DATASETS

The final version of our dataset were scraped and merged from four data sources that includes data recorded between 2015 and 2023.

The four data sources are:

(1) Ergast Developer API
(2) www.formula1.com
(3) Wikipedia
(4) www.autosport.com

We sourced data pertaining to the details outlined in Table 1 from the Ergast Developer API [1]. Information regarding qualifying results was obtained from www.formula1.com [2]. Weather information for each race was gathered from Wikipedia. Additionally, www.autosport.com [3] provided other vital data corresponding to the information in Table 2. We also manually added additional important information such as elevation of the circuit that were not present in any of the datasets available.

The resulting dataset comprises 3,740 data points, each characterized by 48 features and a corresponding label.

Table 2: Data collected from www.autosport.com

Practice Results	Sprint Results
Initial Tyres	Best Lap

3.1 Pre-processing

The dataset we collected presented significant challenges for immediate use with our models. Below, we enumerate the key factors that contributed to these challenges.

The extensive regulations of Formula One, established and implemented by the FIA and subsequently by the FISA, have undergone significant transformations since the inaugural Formula One World Championship in 1950. Whenever significant changes were implemented, new metrics and scoring methods were also introduced. As a result, these changes needed the collection of new important features, which only a small subset of races would possess. Consequently, this left the value of these features absent for the majority of the races. Other factors derive from the unpredictability of the sport. Sessions could be cancelled due to weather conditions, major accidents could prevent completion of a race, and individuals could be absent from qualifying or practice rounds. All of the above lead to 21 features in our dataset with missing values. Also a significant amount of data points had at least one feature with missing value, so removing all instances with missing values wasn't a viable solution.

Features with missing values and the respective solutions implemented are listed next:

- **initial_tyre:** This feature shows what was the choice for the initial tyre of a racer. As some racers didn't start races due to malfunction in their cars, the solution to this missing value is filling it with the type of tyre that is used mostly at the beginning of this track by all the drivers. Usually there is a set of tyre strategies depending on the different racetracks, so assuming this car would have used the most common initial tyre makes sense.
- **qualifying_position, qualifying_tyre:** Throughout a weekend of Formula 1 there are 3 consecutive qualifying rounds, which end up deciding what are the starting positions in the grid for the drivers involved in the race. During the first qualifying session, all drives participate, and the 25% of drivers with worse times are eliminated in this first round. This happens as well in the second round where the same amount of drivers that where eliminated in round 1 are also eliminated in round 2. For round 3 as it is the last one, no one is eliminated and at the end the best times are computed. If a drivers appears without a qualifying position is due to him not participating in any of the sessions. So the decision is to give him the worst position possible and the worst time of the drivers who did compete.
- **fp_pos_1, fp_time_1, fp_pos_2, fp_time_2, fp_pos_3, fp_time_3:** For almost all weekend of races there are 3 free practice sessions, where drivers get to practice in the track before competing in the qualifying session. Sometimes some drivers don't participate in a practice session due to illness, or because they want a substitute racer practicing (this usually

happens throughout the last races of the season). In the case they don't participate in one of the 3 practice sessions of the week, the solution is to take the worse time between the other sessions they did that weekend. It can also happen that a whole practice session is cancelled due to weather. In this case the results of the session were taken as the worst times of the drivers in the other practice sessions. Also when a weekend has a sprint race there could either be only 2 or 3 practice sessions, so again we just decided to take the worst times drivers did during the weekend to fill 3 columns with free practice results. After duplicating this times, we ranked again all the column for an specific event in order to get the actual positions of the drivers.

- **sprint_qualifying_position, sprint_qualifying_time, sprint_fl_pos, sprint_fl_time, sprint_position, sprint_time, sprint_laps:** In all of our data we only have 12 weekends with sprint races, which were introduced in the 2021 season. A sprint race is a short race done the day before the actual race, and also gives points for the final standings table. This race has a significant amount of fewer laps, so the prediction for weekends without this result was a duplication of all of this values for the actual race, but for sprint_time. For this feature the final time considered was proportional to the amount of laps the sprint race had. At the end we only used this columns when there was a weekend with a sprint race on it. For all the other races this columns weren't considered in the model. Also at the end sprint_laps feature wasn't considered.
- **num_o_ps:** If there wasn't any value for amount of pit stops done during the race, it meant that the racer did 0 stops. Empty values where then just replaced by this value.
- **time:** If there wasn't any value for final time done during the race, it meant that the racer didn't finish the race. Empty values where then just replaced by the sum of the worst time done during that race and the fastest lap value that racer who didn't finish had.
- **fl_pos, fl_time:** If there wasn't any value for fastest lap position and time done during the race, it meant that the racer didn't finish a single lap during the race. The missing value was then replaced with the worst lap time of all the racers, and the worst position possible.
- **points:** For this feature there where only a few amount of rows with this problem. There where actual values for this column, and the solution was manually adding this values.

As it can be noticed, the process of replacing the missing values could only be done with having sufficient expertise in this field. After this successful process, the data was completely ready for analysis and modeling.

4 EXPLORATORY DATA ANALYSIS

4.1 Number of times a Formula 1 driver achieved Podium 1 finish

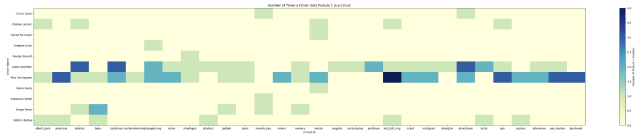

Figure 1: Heatmap representing the number of times a Formula 1 driver achieved a Podium 1 finish (2019 - 2023).

Fig. 1 is a heatmap representing the number of times a Formula 1 driver achieved a Podium first position finish in various circuits. The y-axis lists the drivers' names, while the x-axis identifies different race circuits. The color intensity reflects the frequency of a driver's top finish at a particular circuit; darker shades indicate more first-place finishes, and lighter ones signify fewer or none. For example, Lewis Hamilton has secured the top spot multiple times at the 'americas' and 'baku' circuits. The color key on the right quantifies the color intensity, ranging from 0 to 4 or more top finishes.

4.2 Number of constructors a Formula 1 driver has been associated with

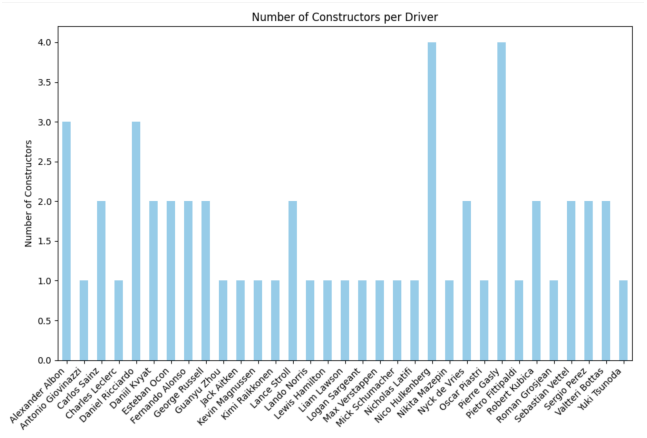

Figure 2: Bar Plot representing the number of constructors a Formula 1 driver has been associated with (2019 - 2023).

Fig. 2 illustrates the distribution of how many racing teams (constructors) each driver has been associated with. The x-axis lists individual driver names, while the y-axis quantifies the number of constructors. The vertical bars represent each driver's association with a particular number of constructors. For instance, some drivers have been with just one constructor, while others have switched between multiple teams. The height of the bars corresponds to the number of teams each driver has been with. This graph provides insights into driver mobility, loyalty, and potential versatility in their racing careers.

4.3 Number of constructors per country

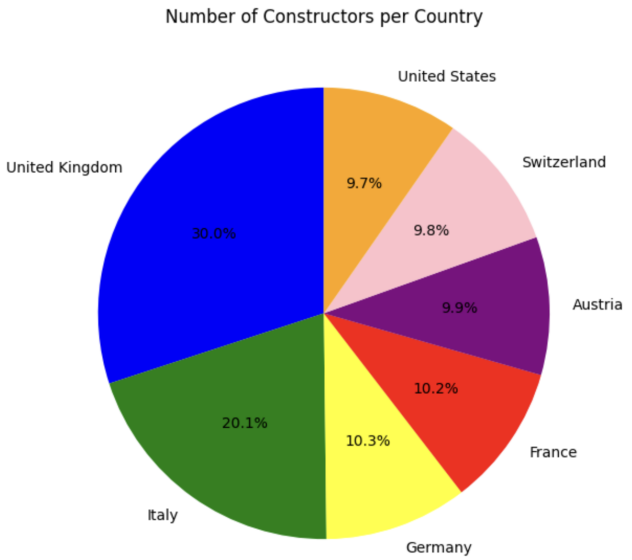

Figure 3: Pie Chart representing the number of constructors per country (2019 - 2023).

Fig. 3 depicts the distribution of "Number of Constructors per Country". The United Kingdom dominates the chart with 30.1% of the constructors. Italy follows at 20.1%, with Germany and France closely behind at 10.2% each. The United States, Switzerland, and Austria each contribute 9.8% to 9.9%. The chart visualizes the prevalence of constructors from various countries, highlighting the dominance of certain nations in this context. The colors serve to differentiate each country, and the percentages offer a precise representation of their respective shares. Overall, the chart provides a clear overview of the geographical distribution of constructors.

4.4 Most dangerous Formula 1 circuit

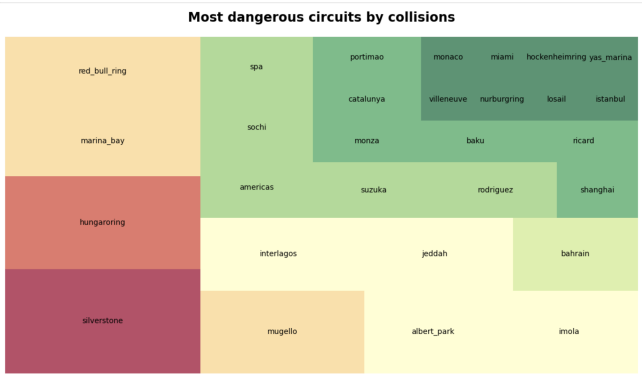

Figure 4: Treemap representing the most dangerous Formula 1 circuit (2019 - 2023).

Fig. 4 visually represents the relative danger of various racing circuits based on the frequency of collisions. Using a treemap layout, each circuit is represented by a colored rectangle, with the size of the rectangle indicative of the number of collisions. Larger areas suggest a higher frequency of accidents. Circuits like "hungaroring", "silverstone", and "marina_bay" have more considerable portions, indicating they witness more collisions than smaller areas like "imola" or "albert_park". Different shades might represent distinct data ranges or categories. The graph offers an intuitive way to understand and compare the danger levels of each circuit.

4.5 Highest number of crashes by a Formula 1 driver

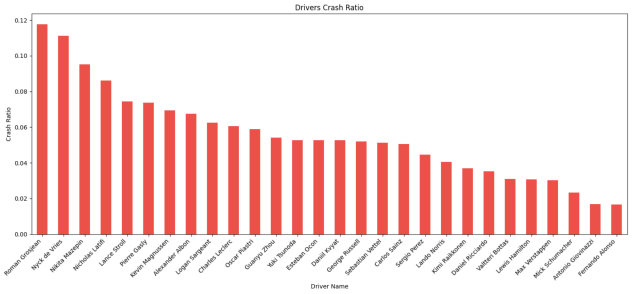

Figure 5: Bar Plot representing the highest number of crashes by a Formula 1 driver (2019 - 2023).

Fig. 5 visualizes the crash ratios of various race car drivers. The vertical axis measures the crash ratio, while the horizontal axis lists the names of the drivers. Displayed as vertical red bars, the lengths represent the frequency at which each driver has crashes. A higher bar indicates a higher crash ratio. Some drivers have notably higher crash ratios than others. For instance, the first driver - Roman Grossjean has the highest ratio, whereas the last driver - Fernando Alonso has one of the lowest. The data provides a comparative insight into the driving styles of these drivers.

4.6 Formula 1 car failures by constructor

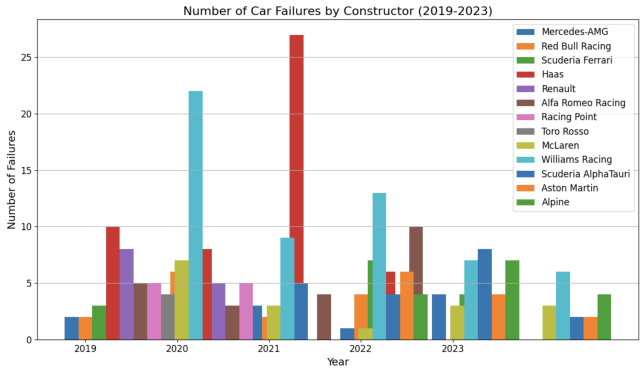

Figure 6: Bar Plot representing the Formula 1 car failure data by a constructor (2019- 2023).

Fig. 6 depicts car failure data over a five-year period for various Formula 1 racing teams. Each year is plotted on the x-axis, with the count of failures on the y-axis. Different constructors, like Mercedes-AMG, Red Bull Racing, and Scuderia Ferrari, are represented by distinct colored bars. For example, in 2020, Mercedes-AMG experienced a notably high number of failures. Contrarily, in 2021, Red Bull Racing had a significant spike. The visual suggests varied reliability and performance of teams over the years, providing insights into their mechanical robustness and areas for potential improvement.

4.7 Correlation Matrix

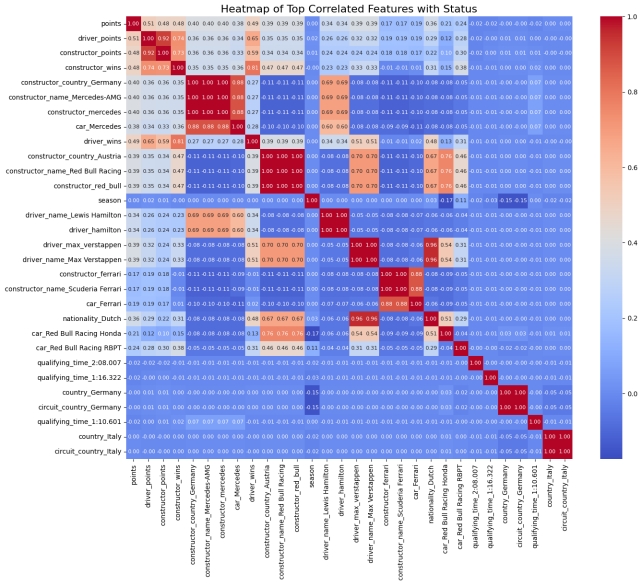

Figure 7: Correlation Matrix representing important features.

Fig. 7 depicts the correlation matrix, which has been utilized to understand the linear relationships between variables. It provides insights into which features are related to each other and, potentially, to the target variable in supervised learning tasks. Each cell in the heatmap represents the correlation coefficient between two features, with values close to 1 or -1 indicating strong positive or negative correlations, respectively. For this task, we have used the "status" attribute as the target.

However, since our dataset includes both categorical and numerical data, we had to one hot encode the categorical attributes in order to generate the heatmap. The one hot encoding lead to a significant increase in the number of columns, which made the analysis more complex and resulted in less insights.

4.8 Top 10 most important features

In order to get better insights about the features, we used a tree based model - Random Forest Classifier, which has inherent method for feature calculation that is easier to interpret. We also used ordinal encoding instead of one hot encoding. We replaced NaN values in "time" and "qualifying_time" with a placeholder value 'Unknown'.

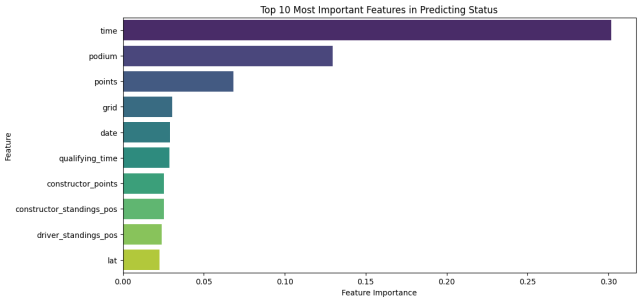

Figure 8: Top 10 most important features

For NaN values in "points", we used median imputation. Fig. 8 depicts the 10 most important features in predicting status. As we can see, "time" has been predicted to be the most important column in predicting "status". The second most important is "podium", and the third most important is "points".

5 METHODOLOGY

Utilizing the selected probabilistic models, we computed the probability of individual racers securing a top-10 finish status in each round of a racing season. Given a list of participating racers, the probabilistic models take into account variables such as the past performance of them, characteristics of the circuit, geographic location, etc. pertinent to each round when making the prediction. These racers are then ranked in descending order based on the calculated probabilities. The racer with the highest computed probability is positioned to secure the first-place position for that specific round. Points are allocated to the top ten racers for each round, and the summation of these points across all rounds is employed to determine the overall champion for the season. The position to points mapping could be found in Tab. 3

position	points
1st place	25
2nd place	18
3rd place	15
4th place	12
5th place	10
6th place	8
7th place	6
8th place	4
9th place	2
10th place	1
11th place onwards	0

Table 3: Points Table

5.1 Hyperparameter Tuning

We tuned our models using the validation set to find out the optimal combination of parameters to use. For logistic regression, we tried different penalty functions (l1, l2), solvers (saga, liblinear), and

inverse of regularization strengths C (np.logspace(-3,1,20)). For neural network, we tested it against different hidden layer sizes ((80,20,40,5), (75,25,50,10)), activation functions (identity, logistic, tanh, relu), solvers (adam, sgd, lbfgs), and penalty values alpha (np.logspace(-4,2,20)). Lastly, for naive bayes classifier, we tried different values for var smoothing (np.logspace(-4,2,20)).

From all models, we have chosen the best performing combinations to use for our final tests. The optimal parameters combination for logistic regression is (l1, saga, 10.0), neural network is ((80, 20, 40, 5), relu, adam, 1.2742749857031321), and naive bayes classifier is var smoothing = $2.848035868435799e-05$.

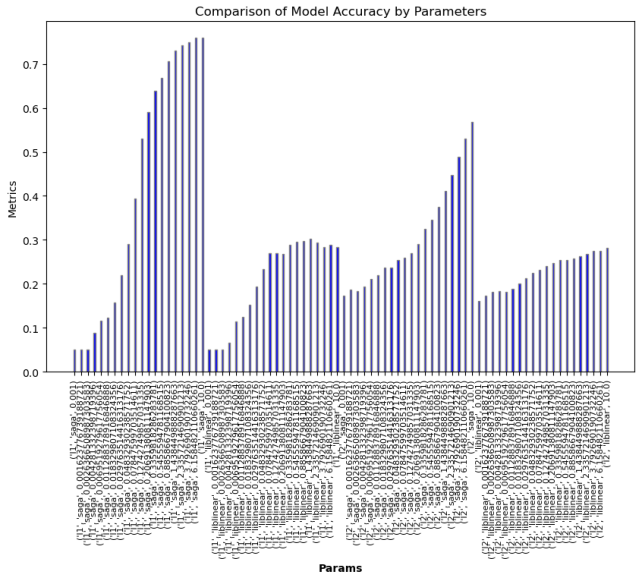

Figure 9: Hyperparameter Tuning for Our Logistic Regression Model.

6 EVALUTAION & ANALYSIS

6.1 Evaluation Metrics

Besides the standard accuracy metric, we compare the Normalized Discounted Cumulative Gain (NDCG) across the different models. For all metrics, for each of the seasons we predict, we use all available past data for training.

6.1.1 Accuracy by Achieving Top 10. To compute this accuracy, we found out the racers who finished in the top 10 places for each round and stored this information as a binary variable. In our model, we predict whether a racer would end up in the top 10 places, also as a binary variable. This accuracy score is then calculated by computing the number of matching binary values over the total number of participating racers.

6.1.2 Accuracy by Podium Ranking. For this accuracy score, we get the actual top-10 ranked racers from a round and store this ranking for comparison in the later step. Then, our model would compute the predicted ranking based on the method described in the first paragraph of Sec. 5. We then sort the predicted ranking in ascending order so that the predicted first place is on top of the list. Lastly, we compute this accuracy score by calculating the number of matching racers over 10.

6.1.3 Normalized Discounted Cumulative Gain (NDCG). NDCG helps in comparing different rankings by considering how close they are to the best possible ranking. NDCG is the DCG value of the current result divided by the DCG of the ideal result list, that is the Ideal DCG (IDCG). This normalization ensures that NDCG lies in the range between 0 and 1, with 1 being the perfect ranking.

$$NDCG@p = \frac{DCG@p}{IDCG@p}$$

Calculating the DCG for predictions involves summing up the relevance scores for each position in the predicted ranking, discounted at each position according to its rank. The formula for DCG at a given rank p is

$$DCG@p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+2)}$$

where rel_i is the relevance of the result at position i, and p is the rank position up to which we are are evaluating. Computing IDCG, which represents the best possible DCG that could be achieved with perfect predictions, involves rearranging the actual finishing positions to create the best ranking and calculating its DCG.

6.2 Analysis

The comprehensive analysis of the machine learning models applied to Formula 1 (F1) race winner predictions in this study highlights several key findings. The three models, Naive Bayes Classifier (NBC), Logistic Regression (LR), and Neural Networks (NN), each demonstrate unique strengths and limitations when applied to the complex and dynamic environment of F1 racing. The performance summary of all models could be found in Tab. 4. The predicted ranking for season 2019-2023 are summarized in Tab. 5-9. For each season's prediction, we trained the model using all of the data we have on hand up to the year before. I.e., to make a prediction for season 2022, we used data from 2015-2021, and for season 2023, we used data from 2015-2022.

The bar plot seen in Fig. 13 and Fig. 14 is the visualization of Tab. 4. Looking at just the model accuracies, it is really difficult to evaluate how well the models have performed. Even when more training data were given for each subsequent seasons, it's still difficult to observe any relationships between dataset sizes versus the accuracy of our models. This conjecture is further validated when observing Fig. 11. The statistics shown in Fig. 11 was obtained by computing the accuracy of our model using 10% more of our training data for each iteration.

These poor and inconsistent accuracies, however, is not too discouraging when we realize that it is not a suitable indicator for showing how good our models are. Since these accuracies were computed as explained in Sec. 6.1.2. This means that, for every wrong position our models predicted, the accuracy is heavily impacted. When looking at the actual predicted ranking, as can be seen in Tab. 5-9, almost all models could correctly predict the first place, if not the top three.

year	LR Acc Podium	LR Acc Top-10	LR NDCG	NN Acc Podium	NN Acc Top-10	NN NDCG	NBC Acc Podium	NBC Acc Top-10	NBC NDCG
2019	0.100000	0.976190	0.651573	0.600000	0.976190	0.821400	0.400000	0.938095	0.835956
2020	0.000000	0.970588	0.627772	0.300000	0.970588	0.824303	0.300000	0.976471	0.834290
2021	0.200000	0.977273	0.695603	0.100000	0.977273	0.808804	0.500000	0.940909	0.839411
2022	0.500000	0.981818	0.832180	0.500000	0.977273	0.846201	0.700000	0.929545	0.844071
2023	0.200000	0.993182	0.884320	0.200000	0.979545	0.896677	0.400000	0.954545	0.892651

Table 4: Performance Comparison

#	Actual Standing 2019	LR 2019	NN 2019	NBC 2019
1	lewis hamilton	lewis hamilton	lewis hamilton	lewis hamilton
2	valtteri bottas	robert kubica	valtteri bottas	valtteri bottas
3	max verstappen	george russell	max verstappen	sebastian vettel
4	charles leclerc	valtteri bottas	charles leclerc	charles leclerc
5	sebastian vettel	max verstappen	sebastian vettel	max verstappen
6	carlos sainz	antonio giovinazzi	pierre gasly	pierre gasly
7	pierre gasly	charles leclerc	carlos sainz	carlos sainz
8	alex albon	sebastian vettel	alex albon	alex albon
9	daniel ricciardo	romain grosjean	lando norris	lando norris
10	sergio perez	alex albon	kimi raikkonen	kimi raikkonen

Table 5: Predicted Podiums for 2019

#	Actual Standing 2020	LR 2020	NN 2020	NBC 2020
1	lewis hamilton	nicholas latifi	lewis hamilton	lewis hamilton
2	valtteri bottas	george russell	valtteri bottas	valtteri bottas
3	max verstappen	romain grosjean	max verstappen	max verstappen
4	sergio perez	lewis hamilton	alex albon	alex albon
5	daniel ricciardo	kevin magnussen	lando norris	lando norris
6	alex albon	kimi raikkonen	charles leclerc	sergio perez
7	carlos sainz	valtteri bottas	pierre gasly	charles leclerc
8	charles leclerc	antonio giovinazzi	sergio perez	lance stroll
9	lando norris	daniil kvyat	daniel ricciardo	daniel ricciardo
10	pierre gasly	max verstappen	carlos sainz	carlos sainz

Table 6: Predicted Podiums for 2020

#	Actual Standing 2021	LR 2021	NN 2021	NBC 2021
1	lewis hamilton	max verstappen	max verstappen	max verstappen
2	max verstappen	sergio perez	lewis hamilton	lewis hamilton
3	valtteri bottas	valtteri bottas	sergio perez	valtteri bottas
4	sergio perez	lewis hamilton	valtteri bottas	sergio perez
5	carlos sainz	lando norris	lando norris	lando norris
6	lando norris	charles leclerc	daniel ricciardo	charles leclerc
7	charles leclerc	daniel ricciardo	charles leclerc	carlos sainz
8	daniel ricciardo	pierre gasly	esteban ocon	daniel ricciardo
9	pierre gasly	carlos sainz	carlos sainz	pierre gasly
10	fernando alonso	fernando alonso	pierre gasly	fernando alonso

Table 7: Predicted Podiums for 2021

The NDCG scores, as shown in Tab. 4, Fig. 10 and Fig. 14 also demonstrate that our models are relatively powerful in determining how good each racer would perform at the end of a season in relations to one another.

Furthermore, if we compute accuracy by using methods explained in Sec. 6.1.1, our models showed consistently high accuracies, except for a case where NBC had a drop in performance when training data size is small.

#	Actual Standing 2022	LR 2022	NN 2022	NBC 2022
1	max verstappen	max verstappen	max verstappen	max verstappen
2	charles leclerc	charles leclerc	charles leclerc	charles leclerc
3	sergio perez	sergio perez	sergio perez	sergio perez
4	george russell	carlos sainz	carlos sainz	carlos sainz
5	lewis hamilton	george russell	george russell	george russell
6	carlos sainz	lewis hamilton	lewis hamilton	lewis hamilton
7	lando norris	lando norris	lando norris	lando norris
8	esteban ocon	esteban ocon	esteban ocon	esteban ocon
9	fernando alonso	valtteri bottas	valtteri bottas	fernando alonso
10	valtteri bottas	kevin magnussen	fernando alonso	valtteri bottas

Table 8: Predicted Podiums for 2022

#	Actual Standing 2023	LR 2023	NN 2023	NBC 2023
1	max verstappen	max verstappen	max verstappen	max verstappen
2	sergio perez	sergio perez	sergio perez	sergio perez
3	lewis hamilton	fernando alonso	fernando alonso	lewis hamilton
4	fernando alonso	lewis hamilton	lewis hamilton	fernando alonso
5	charles leclerc	carlos sainz	carlos sainz	carlos sainz
6	lando norris	george russell	charles leclerc	george russell
7	carlos sainz	charles leclerc	george russell	charles leclerc
8	george russell	lando norris	lance stroll	lance stroll
9	oscar piastri	lance stroll	lando norris	lando norris
10	lance stroll	pierre gasly	oscar piastri	oscar piastri

Table 9: Predicted Podiums for 2023

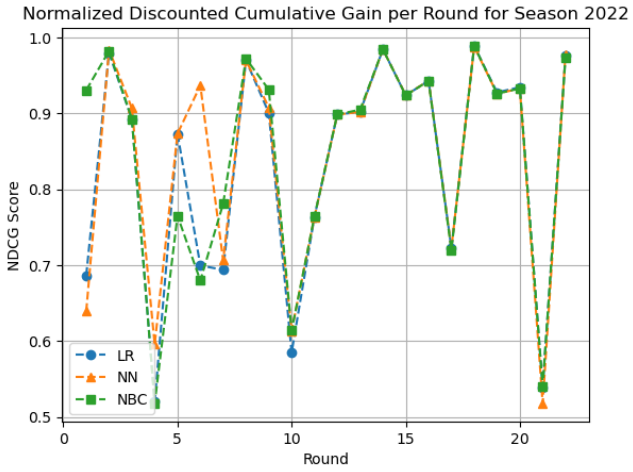

Figure 10: Normalized Discounted Cumulative Gain per Round for Season 2022

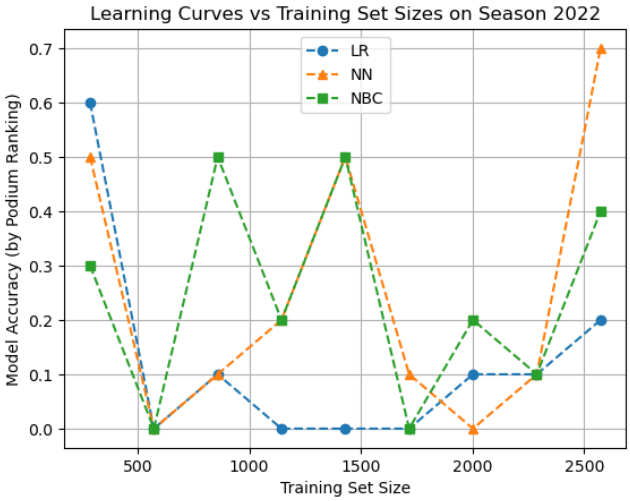

Figure 11: Learning Curves vs Training Set Sizes on Season 2022 - Podium Ranking

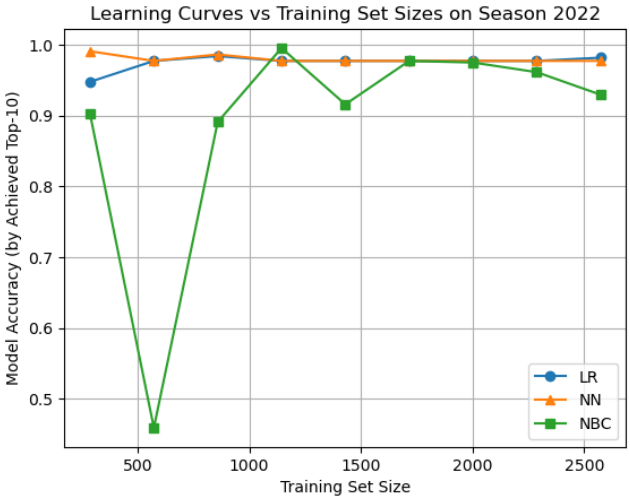

Figure 12: Learning Curves vs Training Set Sizes on Season 2022 - Achieved Top 10

6.3 Did We Achieve Our Goals for the Final Project As We Set In the Proposal?

We believe that we have delivered all of the items proposed in our original proposal.

7 REAL-LIFE APPLICATIONS

- **Fantasy Sports Integration:** The system can be integrated into fantasy sports platforms, allowing users to make more informed decisions when selecting their fantasy F1 teams. This could involve real-time updates based on practice and qualifying sessions, weather conditions, and driver historical performance.

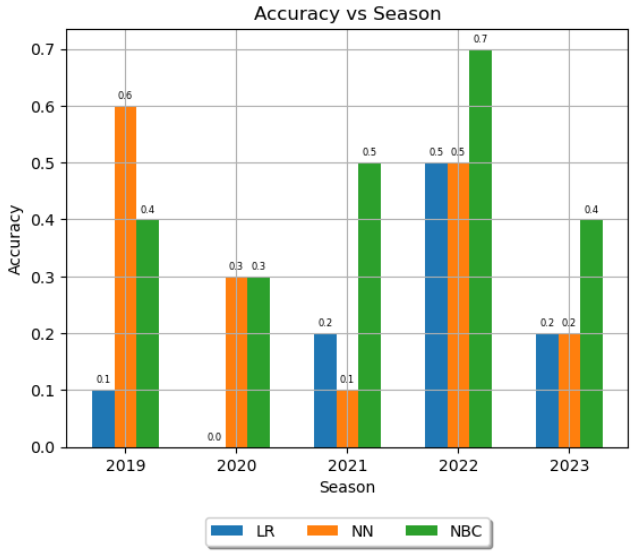

Figure 13: Accuracy by Podium Ranking for Season 2019-2023

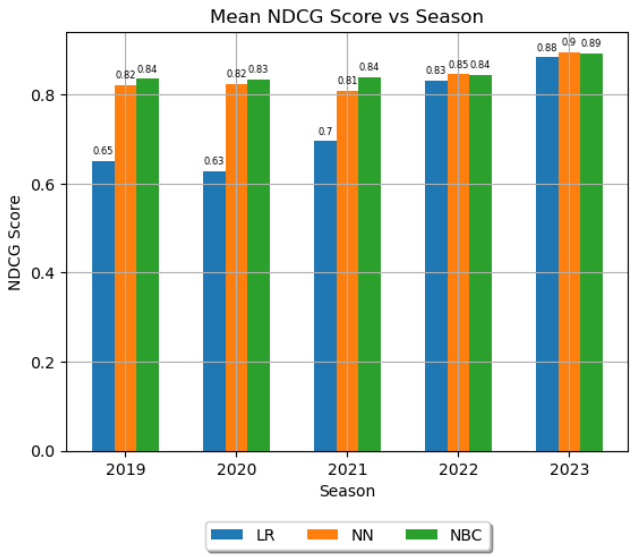

Figure 14: Mean NDCG by Podium Ranking for Season 2019-2023

- **Betting and Gambling Platforms:** The predictive model could be utilized by betting companies to set odds more accurately for each race. This could also involve creating a specialized betting platform that uses the system's predictions to offer unique betting opportunities.
- **Driver and Team Performance Analysis:** Teams could use the system to analyze their drivers' performances and strategize for upcoming races. This might include optimizing car setups, pit stop strategies, and driver training programs.

- **Media and Broadcast Enhancement:** Media outlets covering F1 can use the predictions to enhance their pre-race and post-race analysis. This could involve interactive segments where analysts discuss the system's predictions and compare them with actual outcomes.
- **Data-Driven Marketing for Sponsors:** Sponsors could use the predictive data to tailor their marketing strategies around races. For example, if a driver is predicted to do well, sponsors might increase advertising spending or promotions linked to that driver.
- **Simulation and Training Tools for Drivers:** The predictive model could be integrated into training simulators to help drivers prepare for races. By simulating different race scenarios based on the predictions, drivers can be better prepared for various race conditions.
- **Development of Advanced Analytics Tools:** The underlying technology of the system could be further developed to create more advanced analytics tools for F1 and potentially other motorsports.

8 CHALLENGES

Below were some of the challenges we have encountered and resolved while working on this project.

- **More Data Regarding Driver Standing Required:** Originally, we thought that the Ergast Developer API might have all of the information we need, however, some crucial information were still missing from this source. For example, we have determined that how each racer performs during the practice rounds is a strong indication of how they would perform in the actual race. To gather all the required data, we now merge two datasets to create our own. Additionally, we originally determined that data from the year of 2019 to 2023 would be enough for our purposes. We found out that, even though the current racers did not participate in previous seasons, historical data still contains valuable information on circuits and constructors' performance.
- **Change in Models of Choice:** During our team discussions, we realized that we need to change our approach in predicting the champion. In order to model the rules of F-1, it is important for our model to be able to produce a ranking. The two models we previously pitched, Random Forests and Support Vector Machine, produces categorical class labels which would not give us the required information for us to form our ranking.

9 INSIGHTS & CONCLUSION

The research undertaken in this study has led to several key insights regarding the prediction of Formula 1 (F1) race outcomes using machine learning models. These insights not only highlight the potential of such models in sports analytics but also shed light on the complexities and challenges inherent in predicting outcomes in a highly dynamic and multifactorial environment like F1 racing.

9.1 Importance of Comprehensive Data and Feature Engineering

One of the primary insights from this study is the crucial role of comprehensive data collection and meticulous feature engineering. The success of the machine learning models, especially in a sport as complex as Formula 1, heavily relies on the quality and comprehensiveness of the dataset. Factors such as racer performance, car specifications, circuit characteristics, and even external conditions like weather play a significant role in influencing race outcomes. The effectiveness of the predictive models is directly tied to how well these myriad factors are captured and engineered into the dataset.

9.2 Suitability of Different Machine Learning Models

The comparative analysis of Naive Bayes, Logistic Regression, and Neural Network models in this study reveals interesting insights into the suitability of different machine learning approaches for F1 race predictions. Even though, Neural Networks, with their ability to capture complex patterns and relationships in data, seem particularly apt for predicting outcomes in a data-rich and variable-heavy domain like F1 racing, Naive Bayes, with its assumption of feature independence, showed more promising results in terms of achieving equal or better accuracies across models for four out of the five seasons evaluated. Naive Bayes also demonstrated commendable performance in terms of the NDCG metric. This underscores the efficacy of simpler models like Naive Bayes and Logistic Regression when applied with well-engineered features.

9.3 Broader Implications for Sports Analytics

Finally, the insights from this study have broader implications for the field of sports analytics. The methodologies and lessons learned can be applied to other sports disciplines, especially those that involve a high degree of variability and complexity. The approach of combining diverse data sources and applying machine learning models could offer new perspectives in understanding and predicting sports outcomes beyond Formula 1.

In conclusion, this study not only contributes to the growing body of knowledge in sports analytics but also paves the way for future research in this exciting and evolving field. As data becomes increasingly abundant and machine learning technologies continue to advance, the potential to enhance our understanding and prediction of sports events through analytics seems boundless.

10 WORK DISTRIBUTION

(1) Prior Work: Jiangqiong(Joan) and Mayesha
(2) Dataset
  (a) Dataset Research: Daniel and Nishtha
  (b) Dataset Collecting, Joining, and Pruning: Maurice, Nishtha, and Daniel
(3) Exploratory Data Analysis: Nishtha and Mayesha
(4) Model Research and Implementation
  (a) Naive Bayes: Maurice
  (b) Logistic Regression: Daniel and Nishtha

(c) Neural Network: Mayesha and Jiangqiong (Joan)
(5) Evaluation and Analysis: All Members

REFERENCES

[1] [n. d.]. Ergast Developer API – A public open source Formula One API. http://ergast.com/mrd/

[2] 2023. F1 - The Official Home of Formula 1® Racing. https://www.formula1.com/en.html

[3] 2023. Latest Formula 1 News, Analysis, Results and More. https://www.autosport.com/f1/

[4] K. Bopaiah and S. Samuel. 2020. Strategy for Optimizing an F1 Car's Performance Based on FIA Regulations. *SAE International Journal of Advances and Current Practices in Mobility* 2 (2020), 2516–2530. https://doi.org/10.4271/2020-01-0545

[5] Rory P. Bunker and Fadi Thabtah. 2019. A Machine Learning Framework for Sport Result Prediction. *Applied computing informatics* 15.1 (2019), 27–33. https://www.sciencedirect.com/science/article/pii/S2210832717301485

[6] Alireza. Davoodi, Elnaz Khanteymoori. 2010. Horse racing prediction using artificial neural networks. *Recent Adv. Neural Netw. Fuzzy Syst.Evol. Comput.* (2010), 155–160. https://www.researchgate.net/publication/228847950_Horse_racing_prediction_using_artificial_neural_networks

[7] Leonid et al. Kholkine. 2021. A Learn-to-Rank Approach for Predicting Road Cycling Race Outcomes. *Frontiers in sports and active living* 3 (2021), 714107–714107. https://pubmed.ncbi.nlm.nih.gov/34693282/

[8] Chai W. Hao S. Hu W. Wang G. Cao S. Song M. Hwang J. Wang G. Zhao, Z. 2023. A Survey of Deep Learning in Sports Applications: Perception, Comprehension, and Decision. *arXiv preprint arXiv:2307.03353* (2023). https://arxiv.org/abs/2307.03353