AML Final Project: Formula 1 Race Anaylsis

Anvita Bhagavathula(akb249) Nikhil Chenna Reddy Cheemalavagupalli(nc463)

Jiaying Wang(jw2737)

Abstract

This final project focuses on employing a multimodal machine learning (ML) approach to forecast Formula 1 (F1) driver positions during races using the kaggle F1 dataset[1]. Formula 1, renowned for its dynamism and competitiveness, encompasses diverse variables that significantly influence driver performance. Our study aims to gauge the relevance of F1 race constraints in predicting driver performance. Our approach begins by analyzing different models like Logistic Regression, DecisionTreeClassifier, RandomForestClassifier, SVC, GaussianNB and KNeighborsClassifier using the data of active drivers and constructors and comparing them to pick a model to move forward with. We then train the model on data using both drivers and constructors, only drivers and only constructors. Finally, we incorporate race track data by using their images and leveraging pretrained VGG19 to generate feature maps of these images. Through the utilization of RF-feature importance guided dimensionality reduction, we obtain our highest-performing model which yields a testing accuracy of 0.85. Furthermore, employing feature importance analysis has enabled us to pinpoint the most informative features crucial in predicting race outcomes. This strategic mapping back to these influential features enhances our understanding of their impact on race predictions.

1. Introduction

Formula 1 is a highly dynamic and competitive motorsport where numerous variables can impact a driver's performance. The objective of the project is to assess the significance of Formula 1 race constraints in predicting driver performance. By analyzing various race-related variables and constraints, such as driver confidence, track characteristics, and qualifying race position, and pitstop times, to evaluate which features are most important when it comes to predicting driver standings in a given race.

Utilizing a machine learning framework, our approach to predicting F1 driver standings encompasses a multifaceted strategy. We aim to compile historical race data while integrating features extracted from a convolutional neural network analyzing race track images, embracing a multimodal methodology. Through an ablation study, we seek to discern the most influential features crucial in predicting these variables. Comparing our outcomes with RF feature importance generators will allow us to pinpoint any disparities and refine our model.

2. Related Work

The F1 Predictor tool[2] has served as a foundational baseline model for assessing race outcomes. Noteworthy for its initial attempts at forecasting race positions, the existing predictor, however, falls short in its predictive scope by neglecting crucial factors such as pitstop data, lap times and track-specific characteristics.. Our research aims to build upon this foundation by incorporating these critical variables into the prediction model. By accounting for the strategic dynamics of pit stops, the nuanced performance captured through lap times, and the unique challenges posed by individual tracks, our enhanced model seeks to provide a more comprehensive and accurate depiction of Formula 1 race outcomes. Furthermore, whereas the current predictor delivers position ranges, our improved model strives to predict exact positions, thereby refining the precision and reliability of race predictions in F1.

3. Methods

3.1. Data Exploration

Our research utilizes an extensive dataset encompassing Formula 1 race data from 1950 to 2023. This dataset includes detailed records of races, drivers, constructors, qualifying sessions, circuits, lap times, pit stops, and championship outcomes. To enhance the dataset's richness, we integrated geometric data of race tracks, capturing the unique shapes and characteristics of each circuit. The integration process involved merging information from various datasets, which presented challenges due to inconsistencies

in column names. We addressed this by standardizing the column names to create a coherent and unified dataset.

A significant aspect of our preprocessing was the normalization of constructor(team) names, as some teams underwent name changes over time (e.g., 'Aston Martin' previously known as 'Racing Point'). We manually standardized these names to maintain continuity in our analysis. In terms of feature engineering, we added several key attributes: 'Driver Age', calculated from drivers' birthdates and the race dates; 'Driver Confidence', a ratio of unfinished races to total races for each driver; and 'Constructor Reliability', mirroring the Driver Confidence metric for teams. Additionally, we focused our analysis on active participants by adding indicators for active drivers and constructors, filtering out non-active entities.

To facilitate the processing of categorical variables in our machine learning models, we employed label encoding. This technique converts categorical data into a numeric format, assigning unique integers to each category. This transformation is crucial for integrating categorical data, like team names and driver nationalities, into our predictive models.

As shown in Figure 1, some of the active drivers are very recent and hence, their data is scarce in comparison to more experienced drivers which leads to an imbalance in their data and will probably lead to poor predictions of their race outcomes.

As shown in Figure 2, the constructor data is more equally balanced barring one constructor (Haas F1 team) which is a more recent entry and hence the predictions corresponding to the team might not be as accurate as the others.

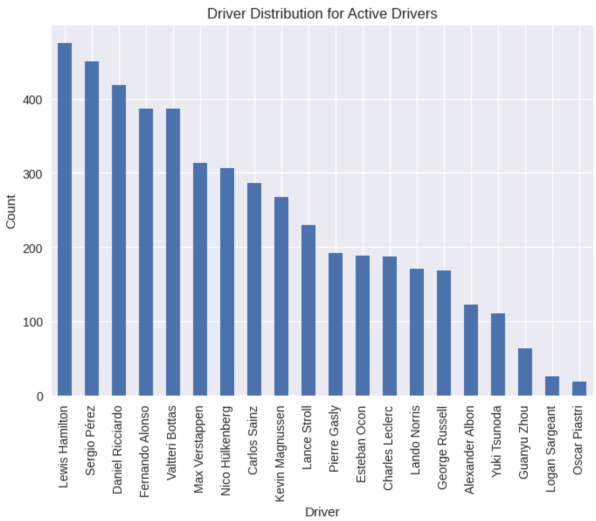

Figure 1. Driver distribution for active drivers

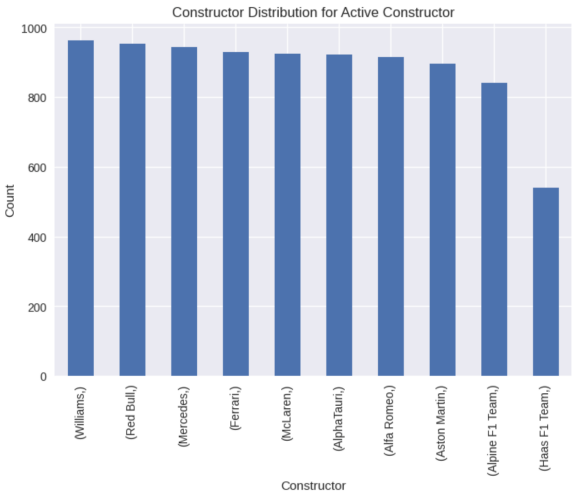

Figure 2. Driver distribution for active drivers

3.2. Experimentation

3.2.1 Baseline Model

In this initial phase of our analysis, we concentrated on predicting the race positions of drivers within the realm of active constructors and drivers. Our dataset was constructed to include only active participants, ensuring a more accurate and relevant analysis. We also divided the dataset into training and testing sets.

Our approach involved leveraging data from active Formula 1 drivers and constructors, employing a cross-validation strategy with a selection of diverse models, including Logistic Regression, Decision Tree Classifier, RandomForest Classifier, Support Vector Classifier (SVC), Gaussian Naive Bayes, and K-Nearest Neighbors. The results are shown in the Table 1. Among these models, RandomForest emerged as the most effective with an accuracy of 0.5739. We also implemented the RandomForest model separately on active drivers and active constructors with both their features. We set the hyperparameter `"n_estimators"` to 100 after experimenting with various other values, lower values yielded a lower accuracy while higher values led to a saturation of accuracy. The isolated application of the RandomForest model on active drivers along with only driver relevant features yielded an accuracy of 0.6331, while with solely active constructors along with only constructor relevant features it scored an accuracy of 0.6393. The baseline accuracy was low across all approaches and so for our next approach we decided to incorporate pitstop and lap time data.

3.2.2 Adding pitstop and lap time data

In iteration 2, we add pitstop data and lap time data. On integrating pitstop and lap time data, the sheer volume of the

Table 1. Baseline Model Accuracies

Model	Accuracy
LogisticRegression	0.1715
DecisionTreeClassifier	0.5731
RandomForestClassifier	0.5739
GaussianNB	0.3328
SVC	0.0739
KNeighborsClassifier	0.0808

laptime of every driver for every lap presented a challenge, leading to overfitting and resulting in both training and testing accuracies converging towards 1.0. To address this, we filtered the data by specifically retaining rows where laptime lap and pitstop lap coincided. This approach enabled us to harness the informative aspects of these variables while mitigating overfitting concerns. The refined model achieved a testing accuracy of 0.7973 for both active drivers and constructors. The model focused exclusively on active drivers, yielded an accuracy of 0.8373, and active constructors, with a testing accuracy of 0.6906.

3.2.3 Using track features extracted using a CNN

To refine predictive models for Formula 1 race outcomes, we identified the unique geometric characteristics of race tracks as a crucial factor influencing race strategies and outcomes. To integrate this aspect into our analysis, we collected a dataset of 35 race tracks[4], each represented as a geojson file (sample track image in Figure 4). These files detailed the tracks' layouts and geometries, including essential metadata like altitude and length, providing a detailed view of each circuit's spatial features.

Recognizing the potential of Convolutional Neural Networks (CNNs) in image data analysis, we employed the VGG19 architecture[3], a deep CNN known for its effectiveness in image recognition. We used this network to extract features from the images of the race tracks. The process involved obtaining features from the final fully connected layer of VGG19, prior to the softmax activation. This step resulted in a high-dimensional dataset, initially sparse due to the predominance of white pixels in the race track images.

To address this sparsity and enhance model performance, we initially trained a Random Forest model with 600 estimators. Although the initial performance was modest, this model was instrumental in identifying the most critical features from our dataset. Notably, our analysis revealed that two image features from the race tracks were among the top 10 influential features. This discovery prompted us to construct a new dataset focused on these top 10 features, encompassing key driver, constructor, and image attributes.

To capture the geometric information of these race tracks, we used a pre-trained CNN with a VGG19 backbone (Figure 5). We chose a CNN because utilizing a convolution would provide the proper inductive bias for the structured geometric features of the race-track, compared to an architecture without built-in inductive bias, such as a Vision Transformer. After ingesting the data, we extracted the feature map from the final connected layer, and flattened the feature vector. We then concatenated this feature vector, along with the meta-data (i.e. altitude and length of the race track), with the F1 dataset, filtered for only active drivers. We trained a RF on this augmented F1 dataset. However, we found that the sparsity of features extracted through the CNN and geo-json, proved to be noisy, resulting in poor performance from this RF model. Hence, we took the top 10 features from the feature importance of the RF model. We observe that 2 of those 10 features were features extracted from the CNN, implying that encoding geometric information yielded higher performance. We then retrained a RF only on the top 10 features across driver, constructor, and

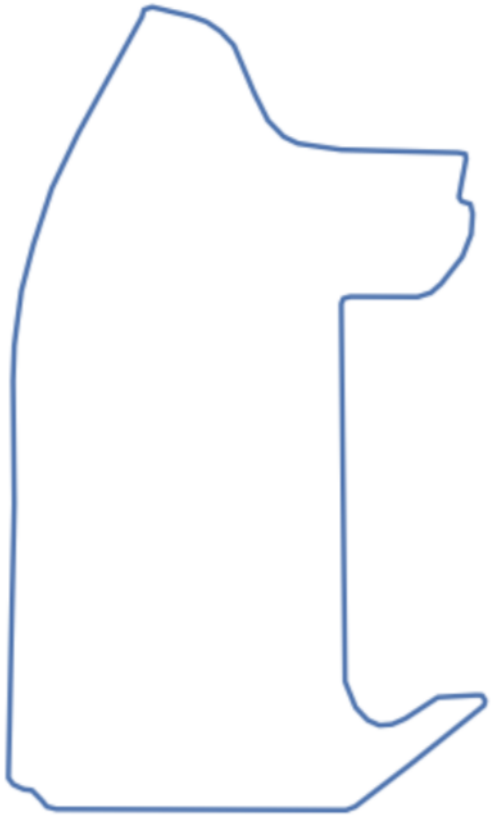

Figure 3. Race Track Example: Las Vegas Strip Circuit

Table 2. Basline vs Augment Model Accuracy

Training Data	Baseline Acc.	Aug. Model Acc.
Driver + Constructor	0.5704	0.7973
Driver	0.6331	0.8373
Constructor	0.6393	0.6906

image extraction features, yielding our highest performance model with a test accuracy of 0.85.

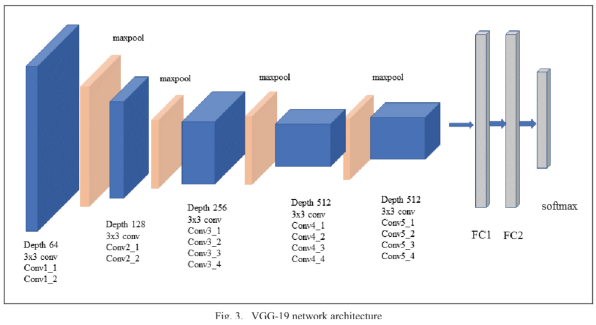

Fig. 3. VGG-19 network architecture

Figure 4. Pre-trained CNN with a VGG19 backbone

4. Results

The results of our initial approach of adding pitstop and lap time (corresponding to the pitstop lap) features to the training data compared to the baseline model are provided in Table 2.

The improved accuracy of the augmented model suggests that pitstop data and lap time data are relevant features to predicting the final position and the driver only dataset performing the better than the constructor only dataset suggests that driver related features are more relevant in predicting the final position thus providing an answer to the age old question, does the driver matter more or the car?

The feature importances of the models using the three datasets are provided in Figure 7, Figure 8 and Figure 9.

As seen in Figure 8 and 9, qualifying position (i.e. position where the driver starts the race) is the most significant feature in determining the outcome of a race. The next most important criteria for a driver is his age while for a constructor, it's their pitstop duration. Interestingly a driver's age and his laptime are more important than the pitstop duration as seen in the combined model (Figure 7), which also indicates the importance of the driver.

For the final model, we wanted to also observe if track characteristics also play a major role in the race outcome and hence, we added race track related features including race geometry extracted from the CNN along with altitude and length of the track and utilized the overall top ten important features to train a model with the driver only dataset

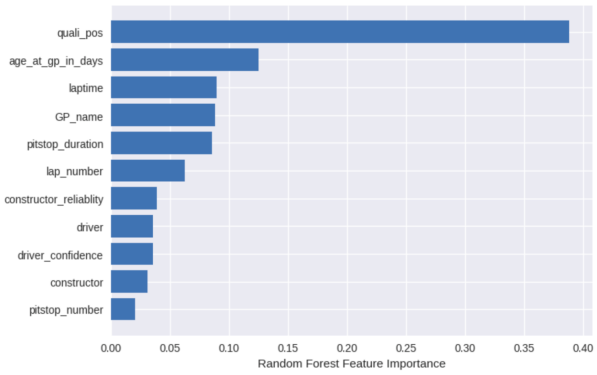

Figure 5. Feature importances of both active drivers and constructors dataset

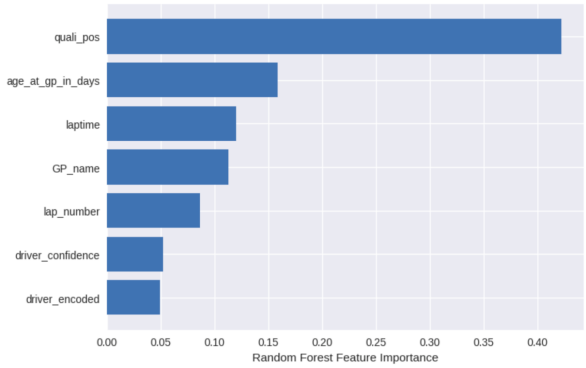

Figure 6. Feature importances of only active drivers dataset

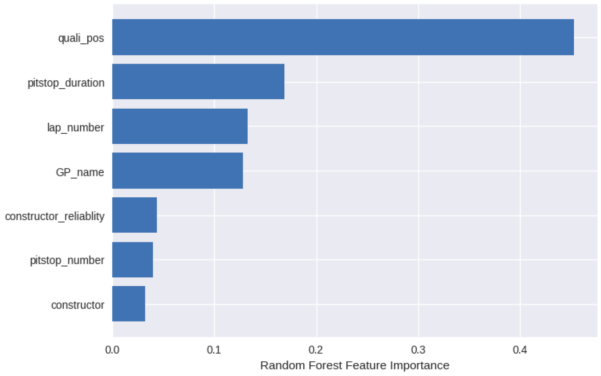

Figure 7. Feature importances of only active constructors dataset

to achieve the best accuracy of 0.85. The feature importance chart is provided in Figure 10.

The final two features are the track features extracted by the CNN which suggests that track layout is more important than the altitude and length of the track but is also significantly less important than driver and constructor features which finally determine the outcome of the race.

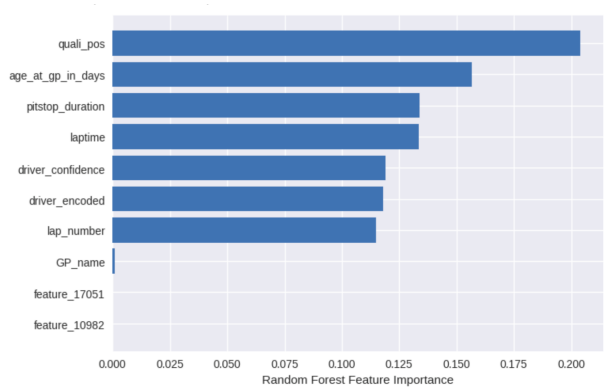

Figure 8. Feature importance chart with track features added

5. Discussion

The incorporation of pitstop and laptime data significantly improved the accuracy of our predictive models, emphasizing the relevance of these features in forecasting Formula 1 race outcomes. The refined model, focused on active drivers and constructors, demonstrated notable enhancements in testing accuracy, with the driver-specific dataset proving more influential than its constructor-oriented counterpart. This underscores the preeminence of driver-related attributes in shaping race results. Moreover, our exploration into race track data, utilizing a Convolutional Neural Network (CNN) for feature extraction, yielded our highest-performing model with a testing accuracy of 0.85. While track characteristics were identified as contributors to race predictions, they remained subordinate to the dominance of driver and constructor attributes, revealing a nuanced hierarchy in the factors influencing Formula 1 outcomes.

Despite these successes, our analysis acknowledges certain limitations. The absence of key features such as tyre data, track banking, and weather conditions poses opportunities for further refinement to enhance predictive accuracy. These elements, integral to the nature of Formula 1 races, could be crucial additions in future iterations. Additionally, careful consideration is needed in addressing the sparsity and noise introduced by certain image-based features and also the scarce data of newer drivers. In conclusion, our multimodal machine learning approach provides valuable insights into the dynamics of Formula 1 races, helping us analyze key features and their importance in playing a role in the outcome of a race and brings us a step closer in answering the age-old question is the driver more important or the car.

References

[1] Formula 1 World Championship. Formula 1 world championship (1950-2023). https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020.

[2] Jaideep Guntupalli, Ritvik Pendyala, Tejdeep Chippa. F1 predictor. https://f1-predictor.gjd.one.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[4] Tomo Bacinger. f1-circuits. https://github.com/bacinger/f1-circuits.