F1 Winner Prediction – ISTE 780 Final Project

Tejas Parvathappa, Arish Bhayani, Rohan Parkar, Shravan Kanchan

1 Introduction

Formula-1 stands as the pinnacle of single-seater auto racing, overseen by the Fédération Internationale de l'Automobile (FIA) and owned by the Formula One Group. Renowned as one of the most popular global sporting events, this demanding sport involves intense racing, necessitating thorough behind-the-scenes analysis to grasp numerous factors influencing the race outcome [1]. Like many sports, Formula 1 is subject to many variables during the event, introducing unpredictability to the results. Recognizing these factors and assessing each player's likelihood of success aids in predicting the race winner. Viewers often have their favorites, and franchises predict potential winners. This proposal focuses on forecasting the race winner, considering factors like race conditions, weather, and past player performances. The primary objective is to analyze race conditions to facilitate companies and race crews in making informed decisions.

To achieve these predictions, the project employs various classification algorithms, comparing their predictions to select the core algorithm with the highest accuracy. Classification algorithms, including Xgboost, KNN, Random Forest, Decision Tree, and Logistic Regression, are utilized and their accuracies are tested. The dataset, sourced from Kaggle, encompasses player details, race conditions, past winners, and other relevant factors, providing substantial information for training and testing models. With over 300,000 data entries, there is ample data to feed these models, enabling them to discern intricate patterns and relationships within the complex dataset associated with Formula-1 races [2].

2 Dataset Preprocessing

The data for this project was acquired from Kaggle and incorporates information about the F1 World Championships and races from 1950 until 2023. The data set consists of the race details, results, statistical data, driver profiles, constructor information, driver standings, victories, race circuits, and lap times for each driver for the respective races. The dataset is available in separate CSV files for specific types of data. The different datasets were merged into a unified dataset, creating a single data frame for model generation. The finalized dataset consists of 21 predictors, and the target label is the driver’s name, who the model predicts to win the race. We have more than 300k data points in the dataset for training, validation and testing the models.

2.1 Pre-processing Steps

The preprocessing phase commenced with the consolidation of disparate datasets into a unified data frame. The amalgamation was executed using the pandas merge operation based on the unique identifiers of the corresponding datasets. After this integration, a thorough examination of the data was conducted to identify any instances of null values; however, no null values were detected across any columns. Given that the presence of null values can compromise model accuracy, it is imperative to address such occurrences to uphold data integrity. To streamline the dataset for

analysis, columns deemed irrelevant to the analytical objectives were omitted, thereby simplifying the dataset, and potentially enhancing the computational efficiency of the model. Recognizing that certain column names lacked specificity in conveying their intended purpose, these were subsequently renamed to facilitate a clearer understanding of the data.

Further, careful attention was paid to ensuring the appropriateness of data types for each column, facilitating seamless analysis. For instance, the column containing dates of birth underwent a conversion from a string to a datetime datatype to enable more comprehensive analysis.

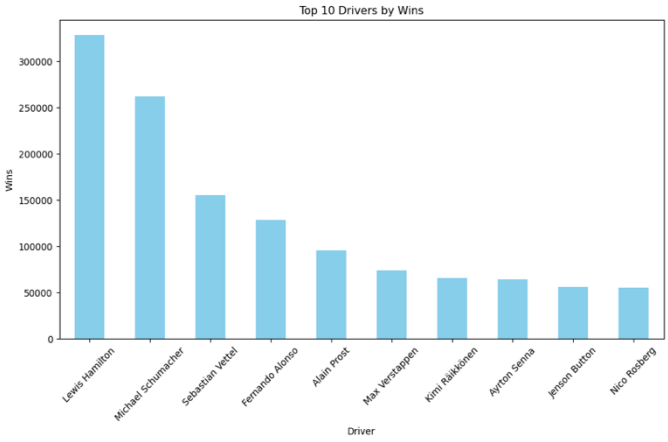

Figure 1: Histogram of historically top performing drivers

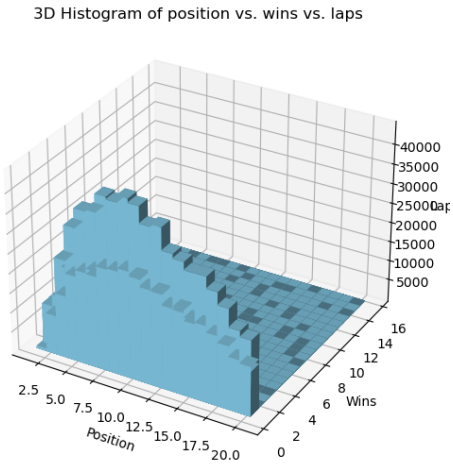

Figure 2: 3D Histogram position vs wins vs laps

Figure 1 is essential for understanding the historical dominance of specific drivers based on the number of victories. Identifying drivers with a consistent track record of high wins provides valuable insights for potential focal points in predictive modeling. Figure 2's 3-D histogram displays the starting position, victories, and matching number of laps completed. It indicates the total number of victories based on the starting position and laps completed.

3 Training and Testing

Recognizing that the dataset might lack certain requisite columns for model training, calculations were performed to derive necessary variables. Specifically, the age of the driver was computed from the date of birth column. Skewness, indicating asymmetry in data distribution, was addressed, recognizing that outliers could be a contributing factor. To mitigate this concern, outliers were identified and subsequently removed from the dataset to mitigate their impact, particularly on certain models. Additionally, categorical variables within the dataset underwent Label Encoding, wherein each category was replaced with a unique numerical value. Lastly, the transformed dataset was divided into predictors (X) and class labels (Y) using the train_test_split function.

The whole dataset was split into 60% training, 20% validation, and 20% testing data. Random Forest and Decision Tree models demonstrate impressive predictive capabilities with minimal errors, achieving a test accuracy of 99.72% [4]. As the core algorithm, XG Boost maintains robust performance with a test accuracy of 97.82%, showcasing reliability in predicting Formula 1 race winners [3]. KNN sustains strong predictive capabilities with a high accuracy of 99.24%.

Meanwhile, Logistic Regression records the lowest accuracy at 16.46%, highlighting its limitations in capturing the intricate dynamics of Formula 1 race outcomes. These results offer valuable insights into each algorithm's strengths and limitations, guiding the next phase—enhancing the XG Boost model using feature scaling techniques like Robust Scaler and Hyperparameter tuning to further refine its predictive capabilities.

3.1 Algorithm Tuning

During exploration and preprocessing, we detected potential data leaks in both Random Forest and Decision Tree models. Data leaks arise when features inadvertently expose information about the target variable during training. Feature importance analysis on these models identified features with unexpectedly high importance, indicating possible information leakage.

In optimizing our Xgboost model for predicting Formula 1 race winners, we systematically tested various learning rates (0.01, 0.1, 0.2) and maximum depths (3, 5, 7) using GridSearchCV. Simultaneously, the impact of feature scaling, specifically Robust Scaler, was assessed to enhance the model's resilience to outliers. This inclusive approach allowed for an exhaustive search of hyperparameters, the best-performing combination of hyperparameters and scaling techniques was then used to train the final Xgboost model, resulting in an optimized predictor for Formula 1 race winners. This streamlined hyperparameter tuning, combined with robust scaling, improves the model's adaptability and predictive capabilities.

Robust Scaler on all the algorithms along with the Core Algorithm (XG boost) shows improved accuracy:

Algorithm	Test Accuracy	Training error	Validation error	Testing Error
XG Boost	**100 %**	**0.0000**	**0.0000**	**0.0000**
KNN	99.74 %	0.0043	0.0060	0.0026
Random Forest	99.76 %	0.0041	0.0038	0.0024
Decision Tree	100 %	0.0000	0.0000	0.0000
Logistic Regression	95.46 %	0.0425	0.0473	0.0454

4 Summary

In summary, this project concludes with a thorough analysis of multiple datasets, and the meticulous examination of many measures, including accuracy, precision, and recall, provides vital information into the efficacy of the original methodology. The testing results demonstrate the Xg boost model's promising abilities in tackling categorical prediction. Based on obvious trends and outcomes, a preliminary forecast positions our proposed model as the F1 winner, representing an optimal balance of precision and recall. However, it is critical to recognize that real-world circumstances may include additional complexities, needing constant refining to ensure the long-term efficacy of any prediction model. This study acts as a foundational step, encouraging further analysis and refinement into how using the right algorithm and fine tuning can prove to be an efficient way to forecast and understand insights better. Finally, ongoing model monitoring and adaptation to accommodate evolving trends in Formula 1, such as rule changes or team dynamics,

will be essential for ensuring the continued relevance and effectiveness of our predictive framework in the dynamic landscape of Formula 1 racing.

References:

[1] A. Patil, N. Jain, R. Agrahari, M. Hossari, F. Orlandi, and S. Dev, "A Data-Driven Analysis of Formula 1 Car Races Outcome," in *Artificial Intelligence and Cognitive Science*, L. Longo and R. O'Reilly, Eds., in Communications in Computer and Information Science. Cham: Springer Nature Switzerland, 2023, pp. 134–146. doi: 10.1007/978-3-031-26438-2_11.

[2] "Formula 1 World Championship (1950 - 2023)." Accessed: Nov. 16, 2023. [Online]. Available: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020

[3] H. T. T. Nguyen, L.-H. Chen, V. S. Saravanarajan, and H. Q. Pham, "Using XG Boost and Random Forest Classifier Algorithms to Predict Student Behavior," in 2021 Emerging Trends in Industry 4.0 (ETI 4.0), May 2021, pp. 1–5. doi: 10.1109/ETI4.051663.2021.9619217.

[4] A. Gupta, K. Gusain, and B. Popli, "Verifying the value and veracity of extreme gradient boosted decision trees on a variety of datasets," in 2016 11th International Conference on Industrial and Information Systems (ICIIS), Dec. 2016, pp. 457–462. doi: 10.1109/ICIINFS.2016.8262984.

[5] W. Alawad, M. Zohdy, and D. Debnath, "Tuning Hyperparameters of Decision Tree Classifiers Using Computationally Efficient Schemes," in 2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Sep. 2018, pp. 168–169. doi: 10.1109/AIKE.