Applied Artificial Intelligence

Semester Project

Course Instructor
Ma’am Shahela Saif

Submitted By
Muhammad Ali (22i-2685)
Taha Ather Awan (22i-2416)
Ahmed Saeed (22i-2445)

Date
09-05-2025

Spring 2025

Department of Computer Science

FAST – National University of Computer & Emerging Sciences

F1 Race Finishing Position Prediction Report

1. Problem Statement

Predicting the outcome of Formula 1 races is a complex task influenced by numerous factors including driver skill, car performance, team strategy, track characteristics, and unpredictable race events. Accurate prediction can be valuable for various applications, from sports analytics to betting markets.

The objective of this project is to develop and compare different machine learning models to predict the finishing positions of drivers in a Formula 1 race. Specifically, the project focuses on predicting the finishing order for the 2025 Saudi Arabian Grand Prix using historical data and pre-race information.

2. Methodology

The project follows a standard machine learning workflow, including data collection, feature engineering, data preparation, modeling, training, and evaluation.

2.1. Data Collection

Race data was collected using the fastf1 Python library. The dataset comprises:

- Historical race data from 2018 up to the year preceding the target race (2024). All races within these years were collected.
- Data from the current season (2025) for races before the target race

For each race and driver, raw data such as qualifying times, grid positions, and race results (finishing positions) were obtained.

2.2. Feature Engineering

Based on the collected data, several features were engineered to capture relevant information for predicting finishing positions:

- Points Index: A normalized score representing the driver's standing in the championship before the race. This was calculated as:
  Points Index= Points of the championship leader before race/Driver's points before race
  This feature aims to capture the driver's form and success throughout the season leading up to the race.
- Constructor Points Index: A normalized score representing the constructor's standing in the championship before the race. This was calculated as:
  Constructor Points Index=Points of the leading constructor before race/Constructor's points before race
  This feature aims to capture the team's overall performance and car competitiveness.

- **Driver Characteristics:** Predefined scores for each driver across several attributes: Wet Weather Skill, Qualifying Pace, RaceCraft, Consistency, Aggression, and Tire Management. These scores were manually defined based on expert assessment and aim to quantify inherent driver abilities.

The target variable for prediction is the FinishingPosition in the race. For the LambdaMART model, a RelevanceScore was engineered as 21−FinishingPosition, where a higher score indicates a better finishing position, suitable for ranking tasks.

2.3. Data Preparation

1. All collected race data from historical and current (pre-target race) seasons, along with the engineered features, were combined into a single comprehensive dataset.
2. Missing QualifyingTime (s) values (e.g., for drivers who didn't set a qualifying time) were imputed using the mean qualifying time from the training data to handle incomplete records.
3. For the target race prediction , qualifying data from the actual qualifying session was collected. The same set of features (Qualifying Time, Grid Position, Points Index, Constructor Points Index, Driver Characteristics) were prepared for the drivers participating in this specific race. Grid Position for the prediction set was determined by sorting the qualifying times.
4. For the LambdaMART model, a QueryID was assigned to each unique race weekend (Year+EventName) to group data points belonging to the same race for the ranking algorithm.

2.4. Modeling

Three different machine learning models were selected and implemented for predicting race finishing positions:

1. **Gradient Boosting Regressor (GBR):** An ensemble model that builds decision trees sequentially, with each tree attempting to correct the errors of the preceding ones. It is a powerful model for regression tasks.
2. **Random Forest Regressor (RFR):** Another ensemble model that constructs a multitude of decision trees during training and outputs the average prediction of the individual trees. It is known for its robustness and ability to handle non-linear relationships.
3. **LambdaMART (LightGBM Ranker):** A specialized learning-to-rank algorithm based on gradient boosting. It is designed to optimize ranking metrics directly, making it suitable for predicting the ordered list of finishers.

2.5. Training and Evaluation

The combined historical and current (pre-target race) dataset was split into training and test sets (80% for training, 20% for testing) using a random split, while ensuring that for the LambdaMART model, the split was done such that races (QueryIDs) were not split between training and testing sets to maintain the integrity of the ranking task.

The selected models were trained on the training dataset. After training, each model was used to predict the finishing positions (or relevance scores for LambdaMART) for the drivers in the 2025 Saudi Arabian Grand Prix using the prepared prediction dataset for that specific race.

Model performance was evaluated on the held-out test set using appropriate metrics:

- **Mean Absolute Error (MAE):** Used for the regression models (GBR and RFR) to measure the average absolute difference between the predicted numerical finishing position and the actual finishing position.
- **Mean NDCG@20:** Used for the LambdaMART model to evaluate the ranking accuracy of the top 20 predicted finishers. NDCG (Normalized Discounted Cumulative Gain) is a standard metric for evaluating ranked lists, where a higher value indicates a better ranking.

Feature importance was calculated for each trained model to understand the relative influence of each input feature on the model's predictions.

3. Results

This section presents the predicted finishing orders for the 2025 Saudi Arabian Grand Prix from each model, the evaluation metrics on the test set, and the feature importance analysis.

3.1 Predicted Finishing Order (2025 Saudi Arabian GP)

Finishing Order	Driver (Gradient Boosting)	Driver (Random Forest)	Driver (LambdaMART)	Driver(Actual Finishing Order in Race)
1	Oscar Piastri	Oscar Piastri	Oscar Piastri	Oscar Piastri
2	Lando Norris	Lando Norris	Max Verstappen	Max Verstappen
3	Kimi Antonelli	George Russell	Lando Norris	Charles Leclerc
4	George Russell	Max Verstappen	George Russell	Lando Norris
5	Charles Leclerc	Charles Leclerc	Charles Leclerc	George Russell
6	Lewis Hamilton	Kimi Antonelli	Kimi Antonelli	Kimi Antonelli
7	Max Verstappen	Lewis Hamilton	Lewis Hamilton	Lewis Hamilton
8	Yuki Tsunoda	Pierre Gasly	Yuki Tsunoda	Carlos Sainz
9	Carlos Sainz	Yuki Tsunoda	Carlos Sainz	Alexander Albon
10	Pierre Gasly	Carlos Sainz	Pierre Gasly	Isack Hadjar
11	Alexander Albon	Fernando Alonso	Alexander Albon	Fernando Alonso
12	Liam Lawson	Alexander Albon	Oliver Bearman	Liam Lawson
13	Fernando Alonso	Isack Hadjar	Esteban Ocon	Oliver Bearman
14	Oliver Bearman	Oliver Bearman	Liam Lawson	Esteban Ocon
15	Esteban Ocon	Liam Lawson	Isack Hadjar	Nico Hulkenberg
16	Isack Hadjar	Esteban Ocon	Fernando Alonso	Lance Stroll
17	Nico Hulkenberg	Jack Doohan	Lance Stroll	Jack Doohan
18	Lance Stroll	Gabriel Bortoleto	Jack Doohan	Gabriel Bortoleto
19	Jack Doohan	Lance Stroll	Nico Hulkenberg	
20	Gabriel Bortoleto	Nico Hulkenberg	Gabriel Bortoleto	
DNF				Yuki Tsunoda
DNF				Pierre Gasly

3.2 Predicted Finishing Order (2025 Japanese GP)

Finishing Order	Driver (Gradient Boosting)	Driver (Random Forest)	Driver (LambdaMART)	Driver(Actual Finishing Order in Race)
1	Lando Norris	Lando Norris	Max Verstappen	Max Verstappen
2	Oscar Piastri	Oscar Piastri	Lando Norris	Lando Norris
3	George Russell	George Russell	Oscar Piastri	Oscar Piastri
4	Kimi Antonelli	Kimi Antonelli	George Russell	Charles Leclerc
5	Max Verstappen	Charles Leclerc	Kimi Antonelli	George Russell
6	Isack Hadjar	Lewis Hamilton	Alexander Albon	Kimi Antonelli
7	Charles Leclerc	Isack Hadjar	Lewis Hamilton	Lewis Hamilton
8	Alexander Albon	Alexander Albon	Charles Leclerc	Isack Hadjar
9	Esteban Ocon	Esteban Ocon	Esteban Ocon	Alexander Albon
10	Lewis Hamilton	Oliver Bearman	Lance Stroll	Oliver Bearman
11	Oliver Bearman	Max Verstappen	Nico Hulkenberg	Fernando Alonso
12	Pierre Gasly	Carlos Sainz	Isack Hadjar	Yuki Tsunoda
13	Carlos Sainz	Fernando Alonso	Oliver Bearman	Pierre Gasly
14	Lance Stroll	Pierre Gasly	Pierre Gasly	Carlos Sainz
15	Liam Lawson	Lance Stroll	Carlos Sainz	Jack Doohan
16	Yuki Tsunoda	Yuki Tsunoda	Fernando Alonso	Nico Hulkenberg
17	Gabriel Bortoleto	Liam Lawson	Yuki Tsunoda	Liam Lawson
18	Fernando Alonso	Gabriel Bortoleto	Liam Lawson	Esteban Ocon
19	Jack Doohan	Nico Hulkenberg	Gabriel Bortoleto	Gabriel Bortoleto
20	Nico Hulkenberg	Jack Doohan	Jack Doohan	Lance Stroll

4. Comparison of Models

This section compares the performance of the three models based on the evaluation metrics and feature importance analysis.

4.1. Performance Comparison

The models were evaluated on a held-out test set to assess their ability to generalize to unseen data. The key performance metrics are summarized below:

Model	**Metric**	**Value**	**Interpretation**
Gradient Boosting Regressor	Mean Absolute Error	3.30 positions	Average error in predicted position
Random Forest Regressor	Mean Absolute Error	3.45 positions	Average error in predicted position
LambdaMART (LightGBM Ranker)	Mean NDCG@20	94.47%	Ranking accuracy (higher is better)

Comparing the regression models, the Gradient Boosting Regressor achieved a slightly lower Mean Absolute Error (3.30) compared to the Random Forest Regressor (3.45), indicating that on average, its predictions were closer to the actual finishing positions on the test set.

The LambdaMART model, being a ranking model, was evaluated using Mean NDCG@20. It achieved a high score of 94.47%, suggesting that it is effective at ranking the drivers in an order that is highly correlated with the actual finishing order, particularly for the top positions.

4.2. Feature Importance Comparison

Analyzing the feature importance scores from each model provides insight into which factors were most influential in their predictions:

Feature Importance Rankings:

Rank	**Gradient Boosting Regressor**	**Random Forest Regressor**	**LambdaMART (LightGBM Ranker)**
1	GridPosition	GridPosition	GridPosition
2	ConstructorPointsIndex	PointsIndex	ConstructorPointsIndex
3	PointsIndex	ConstructorPointsIndex	PointsIndex
4	QualifyingTime (s)	QualifyingTime (s)	QualifyingTime (s)
5	RaceCraft		
6	TireManagement		