KU LEUVEN

Faculty of Engineering Technology, Campus Group T

Formula1 Race Predictions

Vivian Tsang & Dario Nucibella

Introduction

Formula 1 is a premier motorsport where 20 drivers from 10 teams compete for the World Championship across 25 unique circuits, with points awarded based on the finishing positions.

This study uses machine learning to predict race outcomes by applying neural networks and logistic regression to explore the relationship between key features and race results.

The goal is to provide deeper insights into F1 dynamics while showcasing the potential of machine learning to optimize team strategies and enhance the fan experience.

Dataset and Features

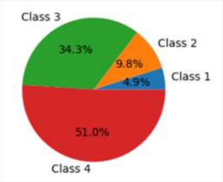
The Kaggle dataset titled "Formula 1 World Championship (1950 - 2024)" was used, containing data for over 1,000 races across various circuits. The dataset includes a vast number of features such as race results, drivers and circuit data, rankings, etc.

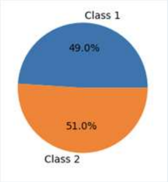
Features used:

- Circuit
- Driver
- Constructor
- Starting position
- Drivers home race (Yes/No)
- Constructors home race (Yes/No)

Challenges:

- Class Imbalance: A reweighting techniques was applied to balance the classes.
- Normalization: Data normalization was necessary for features with inconsistent formatting.
- Feature Selection: Manually selected the most pertinent features.





Methods

The models have been used to predict race outcomes:

- Neural Networks (Multiclass Classification):
  - 4 classes (Win, Podium, Points, No Points)
- Neural Networks (Binary Classification):
  - 2 classes (Points, No Points)
- Logistic Regression (Binary Classification):
  - 2 classes (Points, No Points)

The training and data test set have been defined as follows:

- Training: “Hybrid Era” (2014-2021)
- Test: 2021-2023 Championship

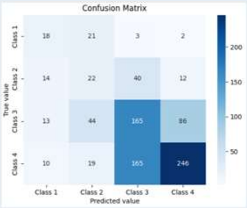
Cross-validation was used on the binary classification models to verify their accuracy across different subsets and minimize overfitting.

Results/Discussion

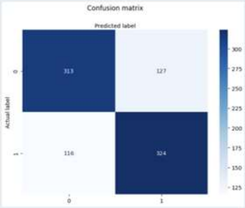
The trained models produced the following results:

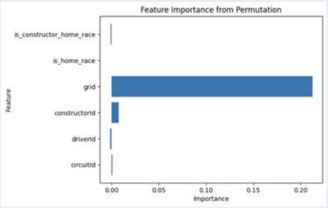
Model Type	Training Loss	Test Accuracy	Cross-Validation
Neural Networks (Multiclass)	0.553	54.66%	N.A.
Neural Networks (Binary)	0.516	72.27%	72.27%
Logistic Regression (Binary)	0.584	72.39%	72.39%

- Neural Networks (Multiclass Classification):

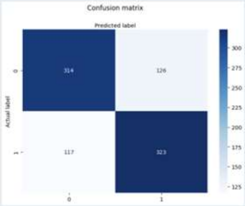


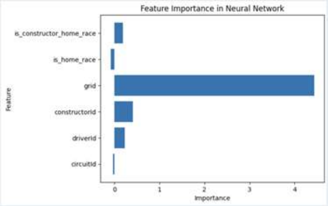
- Neural Networks (Binary Classification):





- Logistic Regression (Binary Classification):





Conclusion/Future Work

The multiclass neural network was excluded form further analysis due to poor performance, likely from task complexity and limited samples.

Both binary models achieved similar accuracies (~72%), with grid position emerging as the most influential feature, indicating potential bias.

A training loss of ~0.5 reflects optimization challenges rather than insufficient features. Cross-validation confirmed model stability, consistently maintaining accuracy.

The binary models demonstrated the potential of machine learning for predicting race outcomes, the following can be implemented:

- Improve accuracy and reduce loss
- Incorporate additional features
- Expand the training set
- Refine the Multiclass Classification model