RACE FINISH PREDICTION IN CAR RACING

YASHWANTH THONUKUNURU

IESEG SCHOOL OF MANAGEMENT, LILLE

BIG DATA TOOLS

Contents

1. Introduction

This report analyzes Peugeot Sports' €5 billion investment in car racing and racing academy over five years. It employs machine learning to predict race completion probability, aiding investment optimization.

2. Executive Summary

2.1 Competitive Dynamics

Formula One is booming. Over the period of 67 years from 1950- 2017, number of car races conducted have significantly increased. The count of races has increased from 30 at the beginning to 80 by the year 2017. This presents a promising opportunity for an entry into the car racing and car development business. This upward trend signifies a rising demand for racing events and associated products like advertisements.

Examining rivals offers crucial insights to enhance strategies and secure a competitive advantage. Among constructors, Ferrari, McLaren, and Williams stand out as the top 3 in wins. Similarly, Lewis Hamilton, Sebastian Vettel, and Nico Rosberg lead in driver victories. Driver expertise significantly influences race outcomes, making it advantageous to choose drivers with over 150 races, as they boast a 30%+ win rate and adeptly maintain race positions.

2.2 Impact of Weather Conditions

Considering the driver's importance, factoring in weather conditions that impact race outcomes is vital. An observed trend shows reduced race success during hot summer races due to longer lap times and extended pit stops caused by high temperatures damaging tires. Crafting strategies to navigate harsh weather conditions is optimal.

2.3 Car Safety Effecting Race Outcome

Car safety significantly impacts race outcomes. While around 3500 races were completed successfully, others were left unfinished due to incidents such as Engine issues, Spun-off incidents, Gearbox-off instances, suspension problems and clutch complications. These incidents frequently occur when cars are driven between speeds of 140 and 240 km/h. Moreover, focusing on this aspect during car development proves advantageous, as drivers often secure Top 3 positions when they achieve speeds around 220 to 250 km/h.

2.4 Importance of Rapid Pit Stops

Effective and swift pit stop strategies play a crucial role in race differentiation. Teams and drivers utilizing fewer average pit stops have demonstrated enhanced position gains in post-pit stop laps, highlighting the efficacy of efficient pit stop utilization

3. Technical Overview

3.1 Data Preparation

Provided including circuit details, constructor results, driver attributes, race outcomes, lap times, and pit stops. When these tables are joined together they provide a comprehensive foundation for our predictive analysis of race completion dynamics, encapsulating the multifaceted aspects of car racing.

As a part of part of the data cleaning and preparation process for modeling, I undertook actions to rectify data types, revised column names for clarity, eliminated extraneous and redundant columns, addressed missing and null values by substituting with mean or mode as appropriate for each column's nature. I'd like to highlight some notable columns I've generated to provide a better grasp of the data.

I. **Likelihood of winning a race**
Winning Rate rate for each driver in Formula 1 races is calculated by analyzing their race results. It is done by calculating the total number of races won by a driver and dividing it by their total number of races participated in. This winning rate metric can be used to evaluate and compare drivers' performances over time, aiding in driver selection and performance assessment.

II. **Performance Change Index**
This metric helps in evaluating driver's performance based on both the average positions changed and the direction of change. The index ranges from -1 to 1:
    - Positive values reflect improved average positions.
    - Negative values signify a decrease in average positions.
    - 0 indicates minimal change.

III. **Pit Stop Analysis Metrics**
Generated variables to assess total and average pit stop duration, total pit stops, and average laps between pit stops, offering insights into pit stop strategies. Additionally, introduced the "pitStop_efficiency" metric that gauges driver pit strategies' effectiveness by comparing pit stop time with laps between stops.

IV. **Time related Variables**
Utilized time, date, and year columns to comprehend how weather conditions and race timings influence race outcomes.

V. **Performance Metrics and Insights**
Generated columns within the 'results' to understand race dynamics through average lap times, position groups, driver experience and constructor success.

A basetable was created by merging all the tables. In order to use the base table in our modeling, Status column has different types of statuses, so replaced status values other than 'Finished' with 'Not Finished' and replaced resulting values with 1 and 0, where 1 indicting 'Finished' and 0 indicating 'Not Finished'

Afterward, I used label encoding to encode the values of categorical columns and I have divided the base table into training and test data sets in such a way that data 30 % of data randomly divided as test and remaining as training data set. Used RFormula transformer to create features and target variables for training and test data.

3.2 Baseline Modelling

Initiated modeling with very basic configuration and with no variables selection and hyperparameter tuning. I have used Logistic Regression, Support Vector Machine and Decision Tree modes and measured models performance using metrics like AUC, accuracy, F1 score, precisian and recall.

Based on the information presented in the following table, Performance of Logistic Regression and SVM models show that Logistic Regression outperforms SVM across various metrics. Specifically, Logistic Regression demonstrates higher Test AUC, Test Accuracy, Test F1 Score, Test Precision, and Test Recall compared to SVM. This suggests that the Logistic Regression model provides a more balanced and accurate classification of data.

Model	Test AUC	Test Accuracy	Test F1 Score	Test Precision	Test Recall
SVM	0.942371	0.849879	0.849934	0.850478	0.849879
Logistic Regression	0.96268	0.891041	0.891059	0.891119	0.891041
Decision Tree	0.997669	0.997579	0.997579	0.997591	0.997579

Decision Tree model shows impressively high performance scores, with Test AUC and Test Accuracy close to 1, this is a potential sign of potential overfitting. This high accuracy is a result of the model memorizing the training data rather than generalizing well to new, unseen data. This would lead to poor performance on real-world data and new scenarios. This is why I did not consider Decision Tree model despite having good performance metrics.

As a next step, I have created challenger model that are aimed to do better than the best model in base line models. I have tried to improve models performance by using variable selection methods, hyperparameter tuning and cross-validation techniques.

I have used the top 12 features from the Logistic Regression model, incorporated these selected variable to retrain the data. However, this approach led to a decrease in model effectiveness due to the loss of information. Herein, am providing an overview of the Logistic Regression model created using the chosen features.

Model (Top-N Features)	Test AUC	Test Accuracy	Test F1 Score	Test Precision	Test Recall
Logistic Regression	0.762002	0.848668	0.84844	0.854225	0.848668

References:

Spark - The Definitive Guide - Big data processing made simple.pdf

Bayesian_Analysis_of_Formula_One_Race_Results_Dise.pdf

Formula 1 World Championship (1950 - 2023) | Kaggle

How does weather impact Formula 1? Find out here (redbull.com)

F1 Insights powered by AWS | Formula 1 uses AWS for Sports (amazon.com)

svm-pyspark/svm-spark.ipynb at master · tonifuc3m/svm-pyspark · GitHub

https://www.kaggle.com/code/carlkirstein/ham-vs-ver-2021-laps-pitstops-conversions

F1, Netflix and Cigarette Company Advertising Executive Summary (exposetobacco.org)

In full: Abu Dhabi GP executive summary of the analysis and clarification exercise - Pitpass.com

PySpark Tutorial - YouTube