**Predict F1 Results**

**Background**

Formula 1 is the highest class of international single-seater auto racing, regulated by the Fédération Internationale de l'Automobile (FIA). The FIA Formula One World Championship has been one of the premier forms of racing around the world since its inaugural season in 1950. A Formula 1 season consists of a series of races, known as Grands Prix, held on purpose-built circuits or closed public roads in different countries and continents. Each race typically lasts around 90 minutes, and the championship usually features between 20 and 24 events per year (the number of races increased in the past few years). Teams develop highly advanced cars, capable of reaching speeds over 300 km/h (about 186 mph), and drivers must obtain a special license from the FIA to compete.

**Dataset Description**

The dataset that I used in this project was obtained in the following link: <https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020?resource=download>

It consists in data from all the races between 1950-2024, and also contains data from each circuit, constructors, qualifying, etc. Further on you’ll see other data that I worked on.

**Objectives**

Build a model to predict the race results, qualifying results, sprint results. Based on those predicts is possible to predict the Drivers and Constructors Champion.

**Data**

* **Data Import**
* **Data Analysis:** analyse the plots and explore data
  + Circuits stats
  + Pilots stats
* **Data Cleaning:** missing values, outliers, duplicates, convert categorical variables (model only can process numerical), feature scaling and dates transformations
* **Data Visualization:** histograms, pie charts, box plots, correlation plots
  + **Questions:**
    - Quais os pilotos mais titulados de sempre? – Horizontal Bar Plot
    - Quais os pilotos que ganharam mais corridas num determinado circuito? – Vertical Bar Plot with na interactive filter
    - Como a altura do circuito impacta a performance dos pilotos? – Scatter Plot
    - Qual o impacto que tem do piloto ser do país do circuito? – Comparative Table
    - Quais as equipas mais tituladas?- Horizontal Bar Plot
    - Como a performance do sprint impacta a performance do piloto na corrida? - Scatter Plot
    - Quantos pontos faz em média cada piloto? - Vertical Bar Plot
    - Como as questões geográficas interferem na performance do pilot? - Radar Chart
  + Choropleth Map (circuits geographic data)
  + Line Chart (comparison of pilots performance through the years, in selected years and specific pilots)
  + Treemap (comparison of points won in a specific year)
* **Train and Test Split:** temporal split without libraries

**Model**

* **Goal:** Predict the winner and the drivers that will achieve points
* **Feature Engineering:** Create one final dataset where each row represents a driver in an specific race
  + Main CSV files are races.csv and results.csv
  + Merge informations: drivers.csv, constructors.csv, circuits.csv and qualifying.csv
  + Create new features: pilot momentum (mean of the last 5 races – points and positions), team momentum (mean of the last 5 races – points), performance historical (pilot performance on each circuit in the past) and championship context (position and points of the pilot and team before the actual race – driver\_standings.csv and constructor\_standings.csv) – attention to the information that wasn’t available at that race
* **Starter Model –** Random Forest (Classification)
* **Advanced Model –** LightCBM, XGBoost and CatBoost
* **Model Validation:** Train set – all years until 2018, Validation set – 2018-2020 and Test Set – 2020-2024 – I didn’t use cross validation because this is a time-series dataset
* **Model Evaluation:** Classification report (accuracy, precision, recall, f1-score), Log-loss, confusion matrix and personalized metrics (?)

**Libraries and Language**

* Plotly, pandas, numpy, datetime

**Results and Discussion**