DS 325 Final Project Report: Analyzing Formula 2 Performance Metrics to Identify Future Formula 1 Talent

Author: Andry Rakotonjanabelo
Course: DS 325 Spring 2025
Instructor: Prof. Roth

Introduction/Abstract

Formula 1 (F1) represents the pinnacle of motorsport, and the FIA Formula 2 (F2) Championship serves as its primary feeder series. Graduating from F2 to F1 is incredibly competitive, often depending on a complex mix of talent, results, funding, and opportunity. This project explores whether quantitative performance metrics from F2 can help identify drivers with the potential to reach F1.

Motivated by the desire to apply data science techniques to a real-world prediction problem in motorsport, the objective was to analyze aggregated F2 driver statistics from the 2018-2019 seasons to uncover patterns related to F1 graduation. Using exploratory data analysis (EDA), feature engineering, and unsupervised K-Means clustering, I grouped drivers based on their performance profiles (average speed, finishing position, gap to winner, etc.).

Principal Component Analysis (PCA) was used for dimensionality reduction and visualization. The analysis revealed distinct performance clusters, with one group clearly representing consistent top performers. While this top cluster contained a higher proportion of drivers who eventually reached F1, the clustering results did not perfectly align with F1 graduation outcomes, indicating that F2 performance alone, as captured by these metrics and grouped by unsupervised methods, is not a sole predictor. This suggests the need for more sophisticated modeling or additional features to fully capture the nuances of F1 progression.

Methods

Data Description

The primary data source for this project was the "Formula 2 Dataset (2018-2019)" available on Kaggle, created by user alarchemn [1]. This dataset contains detailed race-by-race results, including lap times, final positions, time gaps to the winner, average speeds, driver names, and team affiliations for the 2018 and 2019 F2 seasons. For initial exploratory comparisons (presented in Prelim_EDA_Andry.ipynb), data related to Formula 3 teams and drivers was also briefly used, sourced from HTML tables and another Kaggle dataset.

The core analysis focused on the F2 data. A key step involved creating an aggregated dataset (f2_drivers_to_f1.csv) where each row represents a unique driver from the 2018-2019 seasons. This

aggregated dataset includes calculated performance metrics averaged across all races the driver participated in during that period. Additionally, a binary target variable, **REACHED_F1**, was manually added to this dataset. This label was assigned a '1' if the driver participated in at least one official F1 Grand Prix race after their F2 stint (up to the knowledge cutoff date), and '0' otherwise. This labeling was based on publicly available F1 records.

Pre-processing

Several pre-processing steps were performed, primarily within the EDA_formula_2.ipynb notebook, to prepare the data for analysis:

1. **Time Conversion:** Lap times and race times, initially stored as string objects (e.g., "MM:SS.fff"), were converted into numerical representations (total seconds) for calculation.

2. **Handling Missing/Invalid Data:** The 'GAP' column (time difference to the leader or winner) contained non-numeric entries (e.g., "DNF", "DNS", "+1 Lap"). These were handled appropriately – often treated as missing values or filtered out depending on the specific calculation, ensuring that averages were computed only from valid numerical gaps for finished races.

3. **Data Aggregation:** The race-by-race data was grouped by **PILOT NAME** to compute aggregate statistics for each driver over the two seasons. Key metrics calculated include: total **LAPS** completed, average finishing position (**AVG_POS**), average time gap to winner (**AVG_GAP**), average race speed (**AVG_KPH**), average total race time (**AVG_TIME_seconds**), and average best lap time (**AVG_BEST_seconds**).

4. **Feature Scaling:** Before applying clustering and PCA, the numerical performance features in the aggregated dataset were standardized using StandardScaler from scikit-learn. This ensures that features with larger values (like total laps) do not disproportionately influence the distance calculations compared to features with smaller values (like average gap).

Modeling/Model Selection

To identify natural groupings within the driver performance data without relying on the pre-defined F1 graduation label, an unsupervised learning approach was chosen. K-Means clustering was selected due to its simplicity and effectiveness in partitioning data into distinct spherical clusters based on feature similarity. Based on preliminary analysis (visual inspection of PCA plots and considering the nature of performance tiers), k=3 clusters were chosen to represent potential groups like 'top performers', 'midfield', and 'lower performers/less experienced'.

Notably, a custom K-Means implementation using pure NumPy was utilized (as detailed in EDA_formula_2.ipynb), primarily to ensure compatibility and avoid potential issues with optimized linear algebra libraries (like BLAS) in certain environments.

Principal Component Analysis (PCA) was employed for dimensionality reduction, transforming the standardized multi-dimensional feature space into two principal components. This facilitated the visualization of the clusters and the distribution of drivers in a 2D scatter plot.

Results

The K-Means clustering algorithm successfully partitioned the 93 drivers into three distinct clusters based on their aggregated F2 performance metrics from 2018-2019. Visualization using PCA (Figure 1) shows the separation of these clusters in a reduced two-dimensional space.

- **Cluster 1 (Top Performers):** Generally characterized by lower average finishing positions, smaller average gaps to the winner, and higher average speeds.
- **Cluster 2 (Midfield):** Occupied an intermediate space in terms of performance metrics.
- **Cluster 3 (Lower Performers/Inconsistent):** (labeled '0' on figure below) Typically associated with higher average finishing positions, larger gaps, and potentially fewer total laps completed.

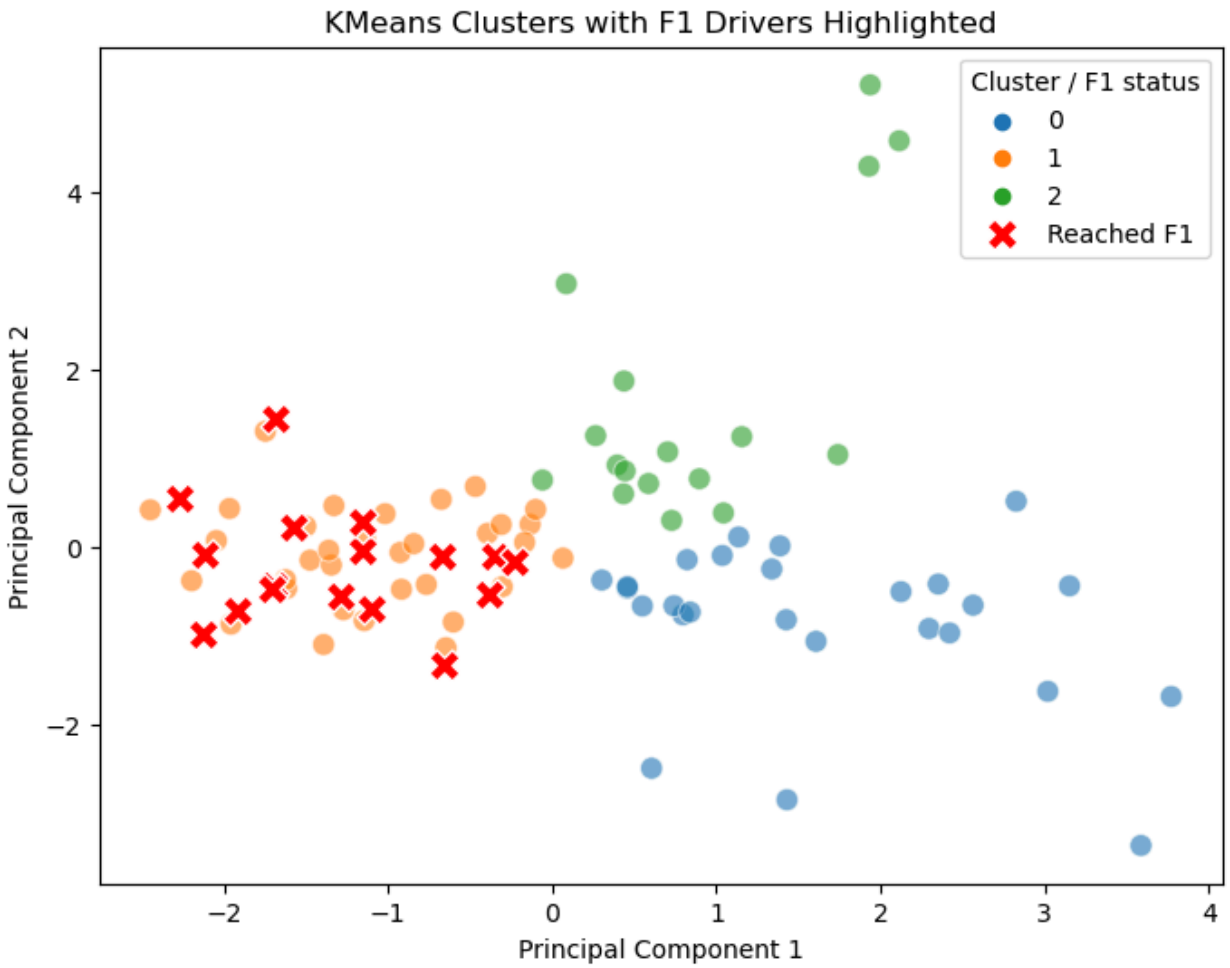

FIGURE 1: PCA visualization of F2 drivers based on 2018-2019 performance metrics. Points are colored by K-Means cluster assignment (k=3). Drivers who subsequently raced in F1 are marked with an 'X'. PC1 and PC2 capture the principal directions of variance in the scaled performance data. Cluster 1 (Orange) tends towards one side, representing top performers, and contains a higher concentration of F1 graduates.

To quantitatively assess the clustering quality, several metrics were calculated:

- Silhouette Score: 0.289
- Calinski-Harabasz Index: 29.8
- Davies-Bouldin Index: 1.443

These scores indicate moderate cluster separation; while distinct groups exist, there's overlap at the boundaries.

When comparing the cluster assignments to the actual **REACHED_F1** labels, the results were:

- Homogeneity: 0.213
- Completeness: 0.100
- V-measure: 0.136

These low scores confirm the visual observation from Figure 1: the unsupervised clusters, while identifying performance tiers, do not cleanly map onto the binary outcome of F1 graduation. Figure 2 further illustrates this distribution.

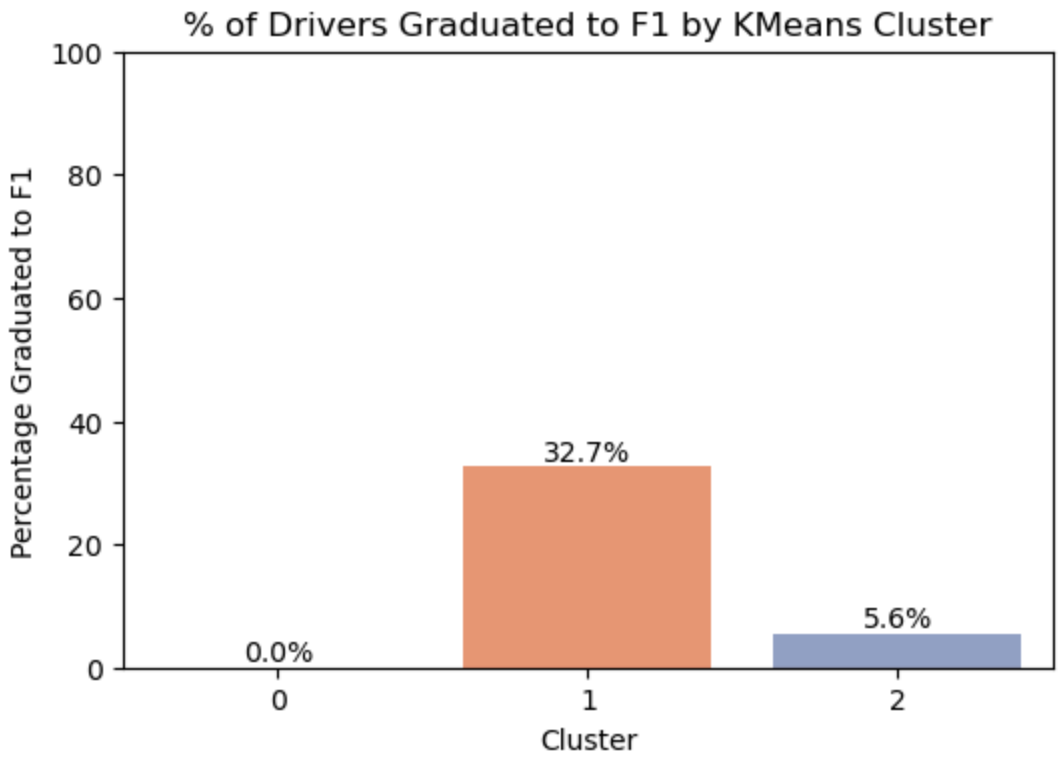

FIGURE 2: Percentage of drivers within each K-Means performance cluster who graduated to F1. While Cluster 1 shows a significantly higher proportion of F1 graduates compared to Clusters 2 and 3, no cluster consists exclusively of F1 graduates or non-graduates, highlighting the imperfect prediction based solely on these unsupervised groupings.

Discussion

Major Conclusions

While F2 performance metrics (average position, speed, gap, etc.) from the 2018-2019 seasons contain sufficient signal to group drivers into distinct performance tiers using unsupervised clustering, these clusters alone are not reliable predictors of F1 graduation. The top performance cluster identified (Cluster 1) did indeed contain a majority of the drivers who eventually raced in F1, confirming that strong F2 results are correlated with F1 opportunity.

However, the presence of F1 graduates in other clusters and non-graduates in the top cluster underscores the complexity of the F1 pathway. Factors beyond the scope of this dataset, such as financial backing, team connections, specific F1 team needs, and performance consistency over longer periods, likely play crucial roles.

Hurdles and Solutions

Data preparation presented minor hurdles, particularly converting time strings and handling inconsistent 'GAP' entries, which required careful parsing and imputation logic. Choosing the optimal number of clusters (k) for K-Means is often subjective; while k=3 was chosen based on domain intuition and visual inspection, formal methods like the elbow method or silhouette analysis across different k values could provide further justification. Using a NumPy-based K-Means implementation circumvented potential environment-specific library issues.

Caveats or Limitations

This analysis is based on only two F2 seasons (2018-2019). Performance trends might differ in other years. The **REACHED_F1** label is a simplification; it doesn't distinguish between drivers who secured a full-time race seat versus those who only participated in a single race or served as reserve drivers.

The unsupervised nature of K-Means means it optimizes for cluster compactness and separation based on the input features, not explicitly for predicting the **REACHED_F1** outcome. Crucially, many external factors influencing F1 graduation are not captured in this dataset.

Future Work

The clear next step is to employ supervised machine learning models. Training models like Logistic Regression, Support Vector Machines (SVM), or Random Forests directly on the **REACHED_F1** label, using the aggregated performance metrics as features, could yield better predictive accuracy.

Feature engineering could be expanded to include metrics like qualifying performance, race start performance, consistency measures (e.g., standard deviation of finishing positions), and potentially incorporating data from earlier junior series like F3. Expanding the dataset to include more seasons would also strengthen the analysis and model robustness.

Citations

[1] Alarchemn. (2023). *Formula 2 Dataset (2018-2019)*. Kaggle. Retrieved May 5, 2025, from https://www.kaggle.com/datasets/alarchemn/formula-2-dataset

[2] Gemini (Large Language Model by Google). (2025, May 5). Assistance in generating final project report structure and text based on provided Jupyter notebooks (Prelim_EDA_Andry.ipynb, EDA_formula_2.ipynb), project guidelines (ProjectReportGuidelines.pdf),