Project nr. D15 Report

Silver Vaino

"Patterns of Excellence: Formula 1 Dataset Analysis and Prediction of Constructors' Winners in the Next 10 Years"

This project focuses on a historical data analysis of Formula 1 from 1950 to 2023, dividing this period into three equal segments. The goal is to conduct a comprehensive analysis and prediction in each period to understand team performance and highlight potential development patterns. The project emphasizes the exploration of teams' success and performance.

Task 2. Bussiness understanding

I Identifying your business goals:

- Background: The project focuses on analyzing Formula 1 dataset to understand team performance over the periods 1950-2023 and predict winners for the next decade.
- Business goals: The clear objective is to comprehend historical development and predict the top team of the next decade.
- Business success criteria: Success is measured through the accuracy of predictions and identification of possible trends.

II Assessing your situation:

- **Inventory of resources:** The dataset comprises multiple files interconnected by a unique identifier (ID). It includes:
  All Formula 1 constructors throughout history.
  All teams' results in various races, including the total points accumulated over time.

- **Requirements, assumptions, and constraints:** Some teams have changed their constructor team name over time. This factor must be considered in the data analysis to ensure accurate and historically correct information about teams.
- **Risks and contingencies**:
- **Data inadequacy or low quality:** Since the project aims to cover Formula 1 data history from 1950-2023, some data may be incomplete or inaccurate.
- **Constructor team name changes:** As some teams have changed names over time, this may affect the accuracy and historical consistency of the analysis.
- **Statistical biases:** Data collection may introduce statistical biases that could impact analysis results.

Terminology:

- **Constructor**: A team participating in Formula 1 races and responsible for constructing the car.

- **Constructor Standing**s: The constructor's point table indicates how different F1 teams accumulate points during a season based on their drivers' achievements, determining the overall team standings.
- **Chassis Performance**: Chassis performance refers to the efficiency of the car's frame and aerodynamics in racing, encompassing stability, handling in corners, and aerodynamic effectiveness.
- **Win-Loss Ratio**: The ratio expresses how many times a team has won compared to losses during a specific period, indicating the overall team capability.
- **Podium Finishes**: Achieving a podium means a team's driver secured a top-three position in a race, signifying the team's strength and competitive ability
- **Technical Failures**: Technical failures indicate mechanical or technical issues during a race, potentially influencing team results and overall performance.

III Costs and Benefits:

Costs:

- **Data Collection and Processing:** Resources necessary for collecting, cleaning, and processing the extensive Formula 1 dataset spanning from 1950 to 2023.
- **Time:** The time investment required for thorough data analysis, developing a prediction model, and the overall project timeline.

Benefits:

- **More Accurate Predictions:** Possible financial benefits and other advantages when more accurate predictions enable informed decisions regarding team strategies, collaboration opportunities, and sponsorships.
- **Long-Term Understanding:** A deeper understanding of team dynamics can lead to long-term benefits in strategic planning and decision-making, enabling teams to better adapt to changing circumstances.

Defining data-mining goals:

Data-mining goals:

Pattern Identification: Identify significant patterns in the Formula 1 dataset that could provide insights into the dynamics of team performance.

Building a Prediction Model: Develop a prediction model that accurately forecasts the results of future Formula 1 races, especially regarding team performance.

Identifying Trends and Development Patterns: Analyze long-term trends and development patterns that may indicate success or decline of Formula 1 teams during specific periods.

Determining Influencing Factors: Highlight essential influencing factors that may impact team results, such as technical innovations, driver changes, or other variables.

Data-mining success criteria:

Accuracy of Predictions: Measure the accuracy of predictions compared to actual results to assess the model's reliability.

Explanatory Power of Patterns: Evaluate how well the model can explain discovered patterns and **their practical significance.**

Supporting Practical Decision-Making: Verify whether data analysis and the prediction model can provide practical insights supporting informed decisions in team strategies, collaborations, and sponsorships.

Persistence Over Time: Ensure that the model maintains its effectiveness and accuracy over time, especially when data volume or nature undergoes changes.

Task 3. Data understanding

I Data Collection:

The data has been downloaded from the Kaggle website as a compressed file containing various Formula 1 datasets throughout history. Among the data are different CSV files, which I will use as a starting point for data cleaning. The goal is to select the necessary CSV files that are crucial for my project analysis.

Data Requirements:

Will select crucial CSV files to serve as the foundation for my analysis and proceed to process them thoroughly.

By dividing the dataset's period from 1950 to 2023 into three equal sectors, I plan to analyze and process each segment separately to later create my predictive model.

Since the dataset spans multiple time periods, and some constructors have updated their names over time, I intend to create a new dataset. In this, I will identify all distinct teams and merge those that have essentially been the same over time but under different names.

Will verify the availability and accessibility of the selected data, ensuring it aligns with my analysis requirements and objectives.

Defining Data Selection Criteria:

The data I require must encompass constructor results throughout history, featuring various constructors and their achievements.

- Will exclude any unnecessary information, focusing on key factors such as constructor names, race results, and other essential related data.
- Cleaning and selecting the dataset ensures that only the data crucial to achieving my research objectives remains.

- Finally, I plan to divide the dataset into three equal parts to facilitate analysis and the creation of my own predictive model. Each part represents a specific timeframe, allowing for in-depth analysis and a better understanding of Formula 1 constructor results.

II Describing Data:

- The dataset includes various CSV files, such as circuits, constructor_results, constructor_standings, driver_standings, drivers, lap_times, pit_stops, qualifying, races, results, seasons, sprint_results, and status.
- To complete the project, I have chosen to focus primarily on the following dataset components: constructor_results, constructor_standings, and constructors.
- The dataset is substantial in size, but through analysis, I have identified that these three CSV files contain the necessary information to fulfill the project's objectives effectively.

III Exploring Data:

- The datasets crucial for my research are interconnected through IDs, ensuring internal adaptability of the dataset.
- The dataset comprises 12,291 rows of data, with each row playing a significant role in conducting my analysis and research.
- The primary dataset I utilize is constructor_results, providing detailed information on Formula 1 constructor results, serving as the main focal point for exploration and analysis.

IV Verifying Data Quality:

- The researcher has employed two approaches to ensure the data quality in the dataset.
- First, the researcher independently cross-referenced and verified data from reliable online sources, prioritizing accuracy and up-to-date information.
- Additionally, a complementary method involved utilizing the "URL" column in the constructors.csv file. By accessing Wikipedia links provided in this column, the researcher aimed to confirm, correct, or augment dataset entries, relying on the credibility of Wikipedia for its continuous updates and accuracy assurance.
- These dual verification strategies were implemented to guarantee the reliability and currency of the Formula 1 constructor data used in the analysis.

Task 4. Planning your Project

Project Plan: Analyzing Formula 1 Constructor Data

Task 1: Data Collection and Preparation

- Method: Download and extract relevant CSV files from Kaggle.
- Tools: Python (pandas, numpy).
- Hours: 10

Task 2: Exploratory Data Analysis (EDA)

- Method: Explore key statistics, distributions, and visualize relevant data points.
- Tools: Python (matplotlib, seaborn).
- Hours: 15

Task 3: Data Cleaning and Integration

- Method: Clean and integrate the selected datasets (constructor_results, constructor_standings, constructors).
- Tools: Python (pandas).
- Hours: 20

Task 4: Constructing Predictive Model

- Method: Develop a predictive model to forecast future constructor performance.
- Tools: Python (scikit-learn).
- Hours: 10

Task 5: Documentation and Reporting

- Method: Prepare a detailed report outlining findings, methodology, and model performance.
- Tools: Jupyter Notebook, Microsoft Word.
- Hours: 5

Comments:

- The project is executed individually, with a target completion date of 11.12.2023.

- Regular progress checks and adjustments to the plan will be made to ensure timely completion.
- The chosen tools and methods are standard in the field of data analysis and machine learning, ensuring efficiency and reliability.
- Adequate time is allocated for documentation to communicate findings effectively.