What task will you address, and why is it interesting? This can be as simple as a couple of sentences.

We are aiming to forecast the winner of the upcoming Formula One Grand prix races using a data-driven approach. Our goal is to create a predictive model which will be able to determine the winning driver of a race. We will also be exploring real-time prediction possibilities.

Formula one has garnered a lot of interest in the past few years especially with the launch of the “Driver to Survive” series on Netflix. The increasing interest in the world championship was reflected in a claimed overall fan attendance of 5.7 million, which represents a rise of 36% since 2019. With multiple new circuits added in the United States, the past few years have recorded maximum viewerships both in live attendance and on television.

Human predictions for F1 races are becoming increasingly common particularly on social media platforms such as X (formerly Twitter) and reddit. It is common for enthusiasts to try to predict the race outcomes having previous information on the qualifiers and free practices. Moreover, there exists, "F1 Fantasy", an official game provided by Formula One, where fans can immerse themselves even deeper into the sport by creating and managing their own F1 team using a virtual budget. Betting on Formula One race results have gained significant traction in the past few years.

Our goal is to harness these factors conducted by fans and expert teams as domain knowledge in constructing a sophisticated predictive model.

How will you acquire your data? This element is intended to serve as a check that your project is doable -- so if you're inventing a new data set, be as specific as possible here.

We are currently utilizing an API called the “Ergast API” (http://ergast.com/mrd) which provides us the race data. This is a link to the Kaggle dataset which has pulled data from the Ergast API and stored it in 14 csv files conveniently.

https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020

On initial EDA, we found that we have a lot of data such as lap times per driver for each race, the pit stop strategies, geographical location of the circuits, pre-race data etc which is already pretty clean. Therefore, we have a good starting point in terms of data availability.

The datasets within these CSV files, while comprehensive, lack crucial weather information - a pivotal factor influencing race outcomes. We intend to extract this data from the information already available in the csv files which mention geographical location of the circuits and the starting times of each race. We might engineer more features to the existing data as and when required. We intend to extract and utilize the weather data which could prove to be an important metric to predict the outcome.

Which features/attributes will you use for your task?

We are planning a two step model approach to solve this problem. Each model will be using a different set of features / attributes.

We intend to use two types of feature groups for each kind of model:

1. **On track** (**real time**) - lap times per driver for each race, number of laps covered, number of pit stops, time spent at each pitstop, initial grid position, safety car deployed or not etc.

   Rationale:

   The features chosen for the on-track model focus on the performance and strategy aspects of the race itself. These features directly relate to how a driver and their team manage the race on the track, which is crucial for predicting race outcomes.

2. **Off track** - weather data i.e. wind speed, temperature and fluctuations in the same, geographical location of the circuits etc.

   Rationale:

   The features chosen for the off-track model consider external factors that can influence race performance. Weather conditions and geographical location can have a significant impact on how a race unfolds. For instance, rain or extreme temperatures can affect tire choices and race strategies.

The above features are just a few to begin initial analysis with and is not an exhaustive list of the kind of attributes that will be modeled on.

What will your initial approach be? What data pre-processing will you do, which machine learning techniques (decision trees, KNN, K-Means, Gaussian mixture models, etc.) will you use, and how will you evaluate your success (Note: you must use a quantitative metric)? Generally you will likely use mean-squared error for regression tasks and precision-recall for classification tasks. Think about how you will organize your model outputs to calculate these metrics.

We envision a two step method to solve this problem.

Obtain Driver Ability to Win Estimation

This variable will capture intrinsic driver skill, or ability to win.

1. This variable attempts to hold all else constant, and asks how much relative skill (“winning ability”) each driver has

    a. The variables that will make up the “environment” will include characteristics about each race, including when it was, the number of laps, our weather variables, and where the race took place
        i. The environment features will exclude any information about each driver’s performance in the race (i.e. lap times)
    b. The target variable will be if the driver’s performance in the race
        i. We plan on using a logistic regression with the target variable being if the driver won
            1. If particular drivers consistently win, we will consider a linear regression that predicts the place the driver finished in
2. We will use the weights from the model as a feature in our next model
    a. Each driver is one-hot-encoded
    b. The environment variables make up the rest of the features1
    c. The weights for each driver’s feature vector from the fitted model will be recorded

The weights generated for each driver will be used as a feature in our prediction model.

Prediction Model With Driver Ability Feature and Time Series Data

In the second stage of the model, we are considering time series data. The second stage of the model deals with on track data i.e. lap times in particular. We will engineer the data as a part of data preprocessing. Some of the preprocessing will include :

1. Relative lap times - [fastest time in a lap - time of all drivers] which will be a performance measure
2. Cumulative lap times - capturing race progression of the drivers.
3. Rolling statistics - rolling mean or standard deviation (moving window) for smoothing and noise reduction and volatility analysis.
4. Lag features - capturing temporal dependencies improves the analysis of the driver’s performance. It's also useful in predicting the pitstop timing of the particular race.

We will initially try to regress on the data using a simple model such as linear regression. We will quantify the results using mean squared error. From there we will have some insights on how to best classify the winning driver.