FormulaAI Hackathon 2022 Data Analytics Report version α

Abyoso Hapsoro Nurhadi, Je-Mé Kruger-Baartjes, Manan Thakral, Muhammad Asif, Sai Ganesh Manda,
Sardor Abdirayimov, Thirulok Sundar

6 steps for data-driven decision-making

1. Ask questions and define the problem.
2. Prepare data by collecting and storing the information.
3. Process data by cleaning and checking the information.
4. Analyze data to find patterns, relationships, and trends.
5. Share data with your audience.
6. Act on the data and use the analysis results.

Introduction

Weather forecasting takes a big part on the possible outcome of a race in Formula 1. Race engineers and technicians need to track the current weather and make predictions during the race. Race sometimes are won and lost based on making sense of what the weather is going to do during a race. Being prepared as a team to act accordingly help drivers to win.

The Challenge

F1 2021, the official Formula 1 videogame uses a physics engine that behaves like the real world. The competition presents historical weather data from the RedBull Racing eSports team with a dataset size of 3,572,328 rows and 58 columns. We are required to develop an AI model that is able to make accurate weather forecasts. The target variables are the categorical `'M_WEATHER'` and the numeric `'M_RAIN_PERCENTAGE'` distributed as follows.

`'M_WEATHER'` : 0 = clear, 1 = light cloud, 2 = overcast, 3 = light rain,
4 = heavy rain, 5 = storm

`'M_RAIN_PERCENTAGE'` : 0 to 100

As inputs, we are free to use as many variables to train the model. The output must be a dictionary or similar representation indicating the predicted weather at 5, 10, 15, 30, and 60 minutes after a timestamp with the rain percentage probability at that time, e.g.,

```
{
  '5':{
    'type': 3,
    'rain_percentage': 0.89
  },
  '10':{
    'type': 3,
    '3': 0.64
  }
  '15':{
    'type': 3,
    '3': 0.52
  }
  '30':{
    'type': 2,
    '2': 0.19
  }
  '60':{
    'type': 3,
    '3': 0.94
  }
}
```

Problem Identification

Predicting `'M_WEATHER'` and `'M_RAIN_PERCENTAGE'` is a classification and a regression problem respectively. The target `'M_WEATHER'` only has values of 0, 1, 2, and 5 while the target `'M_RAIN_PERCENTAGE'` only has values from 0 to 93. There is not much issue with the latter, however this is a severe limitation on the former as our model cannot predict 3 or 4 because it does not have any information to learn it. However as with any general data problem, we proceed with what we are given and this is the data given by the game.

Compulsory Cleaning

We are to filter out rows that provide no value to the AI model. With the considerations directive, we remove observations where `'M_NUM_WEATHER_FORECAST_SAMPLES' = 0` or `'M_SESSION_TYPE' = 0`. After doing this, we reduce the dataset rows size to 2,745,117. However, with this process, we reduce our classification problem even further as follows.

0	2,664,421
1	763,609
2	101,486
5	42,812

→

0	1,880,022
1	763,609
2	101,486

So unfortunately, our AI model will not be able to predict light rain, heavy rain, or storm. Suppose we do not do this cleaning; we can base predictions for storm by simply using a check of `'M_NUM_WEATHER_FORECAST_SAMPLES' = 0` or `'M_SESSION_TYPE' = 0` which is not informative to real world setting. We remove these cases in the first place because the former implies zero weather samples and the latter implies unknown race session. Therefore, we proceed our research and analysis with these limitations by game data.

Further Cleaning Considerations

We proceed with removing features that provide no value to the AI model starting with checking if a feature has all the same values. We find 7 features match this criterion, i.e., `'M_PACKET_FORMAT'`,`'M_GAME_MAJOR_VERSION'`,`'M_PACKET_VERSION'`,`'M_PACKET_ID'`, `'M_SECONDARY_PLAYER_CAR_INDEX'`,`'M_SLI_PRO_NATIVE_SUPPORT'`, and `'M_SAFETY_CAR_STATUS'`. Continuing, we remove irrelevant features considering real world setting as follows.

	Feature	Reason
1		

The Solution

The Outcome