F1 Predictive Analytics

Brainstation Capstone - Sprint 1

Manny Brar

Agenda

- Project Summary
- Project Objective
- SWOT Analysis
- Data Snapshot
- Data Processing
- ML Model Exploration

As Christian Horner, CEO of Oracle Red Bull Racing, says, “Data is in the team’s lifeblood. Every element of performance – how we run a race, how we develop a car, how we select and analyze drivers – it’s all driven by data.”

Project Summary

Back to Agenda Page

Predictive Analytics in Formula 1

This project aims to explore and develop predictive analytics models that utilizes historical race data, weather conditions, and car performance metrics to forecast race outcomes in Formula 1. By analyzing patterns and correlations within the data, the model seeks to provide actionable insights that can help F1 teams make informed decisions regarding race strategies, car setups, and driver management to maximize their competitive edge.

Big Business, Big Risks, Big Data Driven Decissions

With Formula 1 being one of the most valuable sports empires in the world at a staggering $18 billion dollars... This is a big business, where every decision has a major dollar figure attached.

Throughout the history of F1, teams have utilized a combination of machine learning techniques and statistical analysis to process and analyze a comprehensive dataset comprising lap times, tire usage, pit stop strategy, weather conditions, and other relevant variables from past F1 races.

Project Objective

Race Placement Predictive Modelling

This project will utilize machine learning algorithms to make predictions on top 5 race placements and will explore an extensive dataset that includes lap times from past races and qualification lap results, tire usage, pit stop strategies, weather conditions, and additional pertinent variables from historical Formula 1 races. The predictive model developed from this analysis will undergo thorough training, testing, and validation phases using these datasets, to confirm its precision and dependability in predicting outcomes.

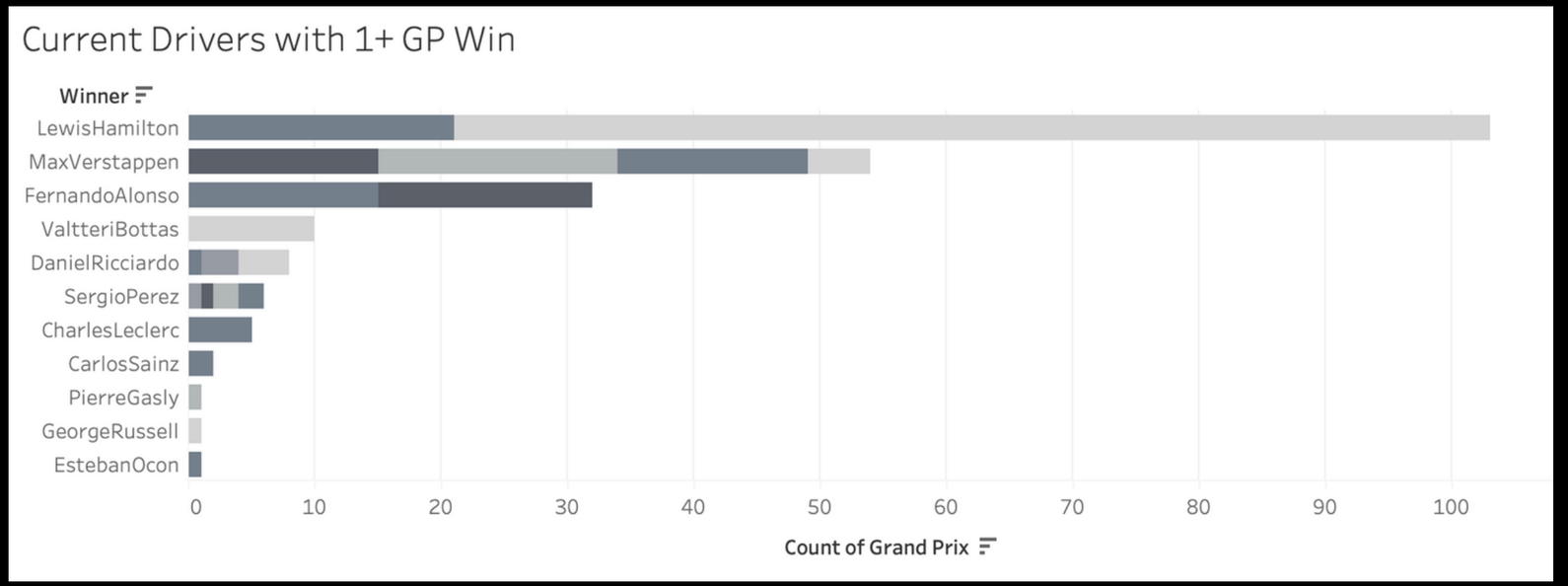
S	W	O	T
Strengths	Weaknesses	Opportunities	Threats
• I will be using a number of data resources and with a combination of these datasets, I will have access to a lot of valuable and intricate data points, that can lead to a stronger predictive model	• Although I have access to a range of data points, I do have a lot of missing values and metrics that will need to be cross referenced and verified as I proceed.	• If the project is executed according to the plan, and I don't run into any unforeseen hurdles, the overall value of these predictions could be huge in terms of sports betting.	• Important factors to keep in mind for this project would be, the amount of time required to get everything completed, any further issues with data collection and data integrity could cause further delays. So I will need to be mindful and very calculated as I progress.

Back to Agenda Page

Data Snapshot

Back to Agenda Page

Date	Grand_Prix	Winner	Team	Laps	RaceTime_sec	AVG_Lap_Time_sec
2023-11-26	Abu Dhabi	MaxVerstappen	Red Bull Racing Honda RBPT	58	5222	90.034483
2023-11-18	Las Vegas	MaxVerstappen	Red Bull Racing Honda RBPT	50	5348	106.960000
2023-11-05	Brazil	MaxVerstappen	Red Bull Racing Honda RBPT	71	7008	98.704225
2023-10-29	Mexico	MaxVerstappen	Red Bull Racing Honda RBPT	71	7350	103.521127
2023-10-22	United States	MaxVerstappen	Red Bull Racing Honda RBPT	56	5721	102.160714
...	...	...	...	...	...	...
1950-06-18	Belgium	Juan ManuelFangio	Alfa Romeo	35	10046	287.028571
1950-06-04	Switzerland	NinoFarina	Alfa Romeo	42	7373	175.547619
1950-05-30	Indianapolis 500	JohnnieParsons	Kurtis Kraft Offenhauser	138	10015	72.572464
1950-05-21	Monaco	Juan ManuelFangio	Alfa Romeo	100	11598	115.980000
1950-05-13	Great Britain	NinoFarina	Alfa Romeo	70	8003	114.328571



Data Processing

Accessing Raw API Data

On the left is a snapshot of the raw data in a nested JSON format that shows the qualifier results for a race in 2023, and on the right we have the data flattened into a pandas dataframe.

Back to Agenda Page

```
[{'season': 2023,
  'round': 1,
  'url': 'https://en.wikipedia.org/wiki/2023_Bahrain_Grand_Prix',
  'raceName': 'Bahrain Grand Prix',
  'Circuit': {'circuitId': 'bahrain',
   'url': 'http://en.wikipedia.org/wiki/Bahrain_International_Circuit',
   'circuitName': 'Bahrain International Circuit',
   'Location': {'lat': 26.0325,
    'long': 50.5106,
    'locality': 'Sakhir',
    'country': 'Bahrain'}},
  'date': datetime.datetime(2023, 3, 5, 0, 0),
  'time': datetime.time(15, 0, tzinfo=datetime.timezone.utc),
  'QualifyingResults': [{'number': '1',
    'position': '1',
    'Driver': {'driverId': 'max_verstappen',
     'permanentNumber': '33',
     'code': 'VER',
     'url': 'http://en.wikipedia.org/wiki/Max_Verstappen',
     'givenName': 'Max',
     'familyName': 'Verstappen',
     'dateOfBirth': '1997-09-30',
     'nationality': 'Dutch'},
    'Constructor': {'constructorId': 'red_bull',
     'url': 'http://en.wikipedia.org/wiki/Red_Bull_Racing',
     'name': 'Red Bull',
     'nationality': 'Austrian'},
    'Q1': '1:31.295',
    'Q2': '1:30.503',
    'Q3': '1:29.708'},
```

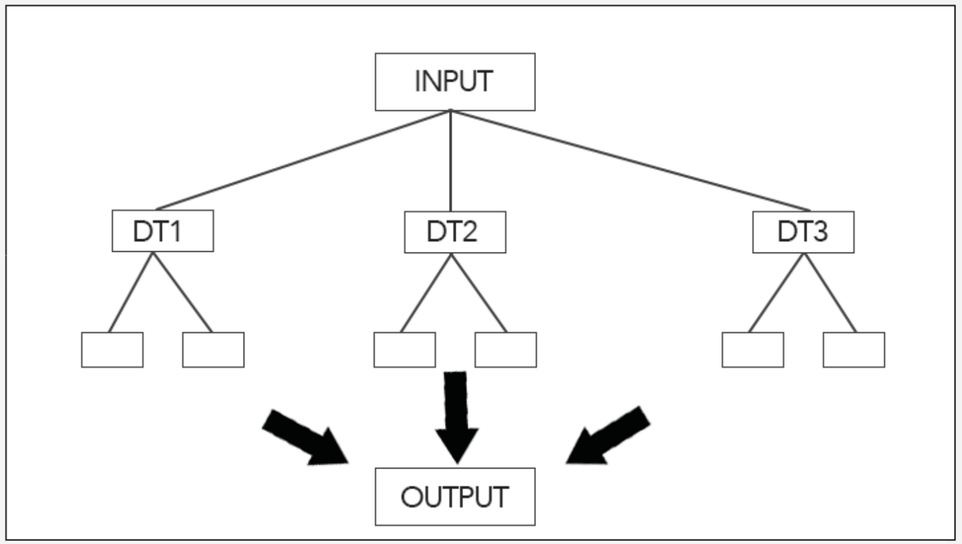
	number	position	Q1	Q2	Q3	Driver.driverId
0	1	1	1:31.295	1:30.503	1:29.708	max_verstappen
1	11	2	1:31.479	1:30.746	1:29.846	perez
2	16	3	1:31.094	1:30.282	1:30.000	leclerc
3	55	4	1:30.993	1:30.515	1:30.154	sainz
4	14	5	1:31.158	1:30.645	1:30.336	alonso
5	63	6	1:31.057	1:30.507	1:30.340	russell
6	44	7	1:31.543	1:30.513	1:30.384	hamilton
7	18	8	1:31.184	1:31.127	1:30.836	stroll
8	31	9	1:31.508	1:30.914	1:30.984	ocon
9	27	10	1:31.204	1:30.809	NaN	hulkenberg
10	4	11	1:31.652	1:31.381	NaN	norris
11	77	12	1:31.504	1:31.443	NaN	bottas
12	24	13	1:31.615	1:31.473	NaN	zhou
13	22	14	1:31.400	1:32.510	NaN	tsunoda
14	23	15	1:31.461	NaN	NaN	albon

ML Model Exploration

Back to Agenda Page

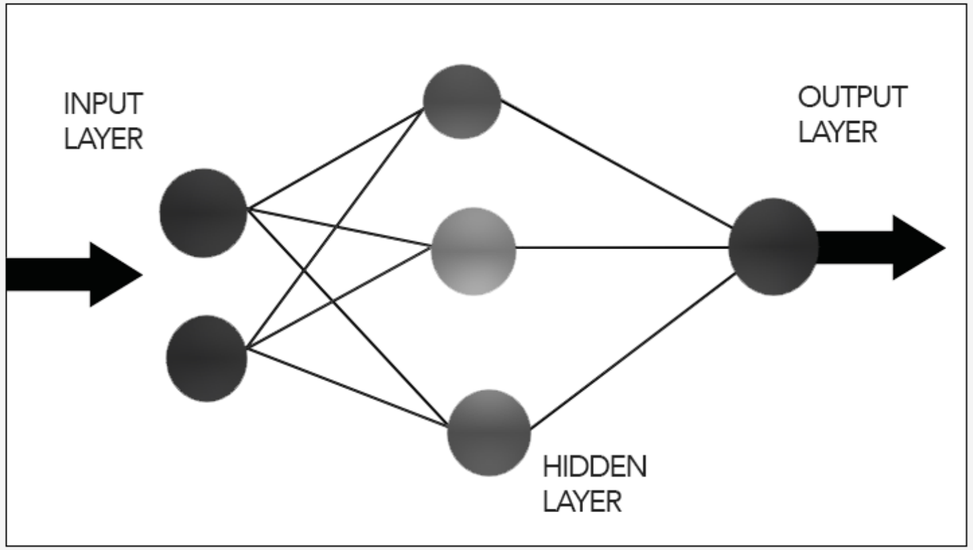
Random Forest Regression:

- This ensemble method combines multiple decision trees, leading to more robust predictions. It can handle non-linear relationships and complex interactions between features.



Multi-Layer Perceptron (MLP):

- This is a basic neural network architecture that can learn complex, non-linear relationships between features. It can be used for both regression (predicting finishing positions) and classification (predicting categories).



Thank You

Back to Agenda Page