Formula E

Racing Lap Prediction

Team SmithRacers101

Agenda

- Business Case Introduction
- Data Preprocessing
- Graphical Data Analysis
- Data Modelling
- Model Evaluation and Results
- Conclusion and Future Recommendations

1.

Business Case Introduction

Formula E Racing

- World's first fully electric, international one-seater, street racing championship
- First Global Sport with net zero carbon footprint

“

Predict the number of laps a driver will need to complete the race

Why should we solve this problem?

Real Time Prediction will optimize race performance and result in better energy optimization strategy for the racer

If number of laps left = 30, battery status = Low :
Strategy : Drive conservatively and save energy

If number of laps left = 30, battery status = High :
Strategy : Drive aggressively and expend energy

2.

Data Preprocessing

Data Merging

1. Concatenating datasets across all seasons
2. Keeping the common columns

```
Index(['TOTAL_TIME', 'DRIVER_FIRSTNAME', 'DRIVER_HOMETOWN', 'DRIVER_SHORTNAME',
       'LAP_TIME', 'GROUP', 'Unnamed: 25', 'KPH', 'S3_LARGE', 'ï»¿POSITION',
       'S1_LARGE', 'S2', 'MANUFACTURER', 'NUMBER', 'S1', 'file_loc', 'S3',
       'TIRES', 'STATUS', 'DIVISION', 'LAP_IMPROVEMENT', 'HOUR', 'TOP_SPEED',
       'PIT_TIME', 'S3_IMPROVEMENT', 'CLASS', 'DRIVER_COUNTRY',
       'CROSSING_FINISH_LINE_IN_PIT', 'GAP_PREVIOUS', 'S1_IMPROVEMENT',
       'DRIVER_SECONDNAME', 'VEHICLE', 'S2_LARGE', 'FL_LAPNUM', 'LAP_NUMBER',
       'LAPS', 'DRIVER_LICENSE', 'S2_IMPROVEMENT', 'GAP_FIRST', 'FL_KPH',
       'ï»¿NUMBER', 'TEAM', 'ELAPSED', 'DRIVER_NAME', 'FL_TIME',
       'DRIVER_NUMBER'],
      dtype='object')

(66321, 46)
```

Defining New Parameters

Creation of new columns :

- Location
- Match Type
- Total Lap Number
- Total Pit Count

```
28191            Berlin
33092          HongKong
40606          Montreal
66243            Zurich
            Long Beach
                Berlin
                 Punta
                Mexico
           Mexico city
                 Punta
location, dtype: object
```

```
           Q
         FP2
         FP2
         FP1
        Race
        Race
         FP1
         FP1
         FP2
        Race
match_type, dtype: object
```

Aggregation of Data

- Aggregation of data by driver name, match type, group and team
- Including data of KPH, S1, S2 and S3 summary statistics
- Conversion of Lap Number into total laps done
- Summation of all pit time into total pit time taken

mean_KPH	std_KPH	med_KPH	max_KPH	min_KPH	range_KPH	skew_KPH	mean_S1
103.200000	27.378660	114.20	123.5	14.2	109.3	-2.195524	4.994981e+04
109.641176	28.468537	118.50	127.0	9.7	117.3	-3.125489	7.058582e+04
55.950000	74.034080	55.95	108.3	3.6	104.7	0.000000	1.205626e+06
118.713889	12.634514	121.95	139.3	68.7	70.6	-3.401031	2.122811e+04

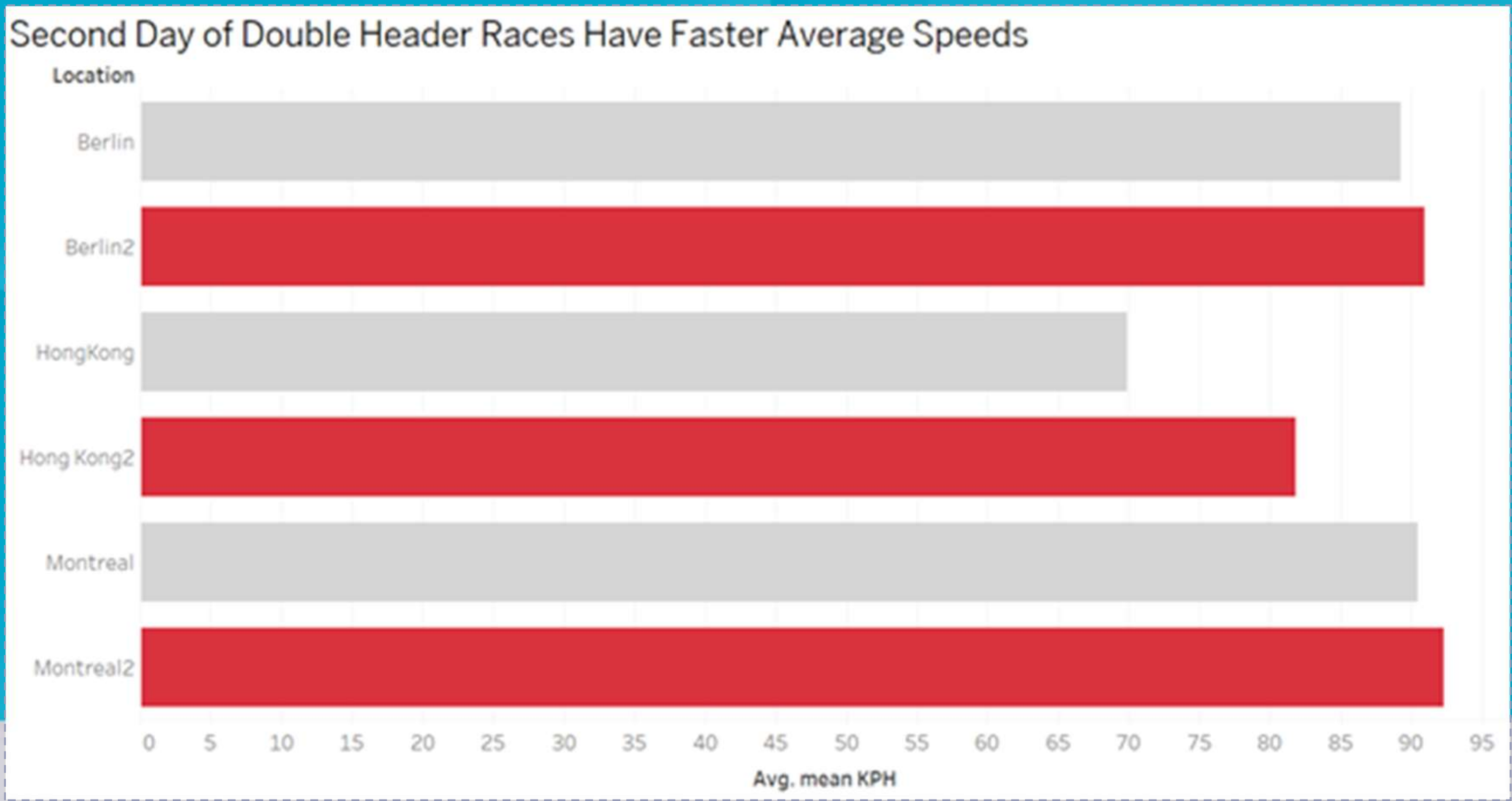
Aggregation of Weather Data

mean_air_temp	std_air_temp	med_air_temp	max_air_temp	min_air_temp	range_air_temp	skew_air_temp	mean_track_temp
27.101836	0.574968	27.4444	27.7222	26.3889	1.3333	-0.406788	21.574100
16.888880	1.024915	17.3333	19.2222	15.0556	4.1666	-0.614885	21.913584
25.546313	0.219368	25.6667	25.7778	25.1667	0.6111	-0.918498	20.555600
27.262177	0.175923	27.2778	27.5000	26.9444	0.5556	-0.645954	23.245589
27.817464	0.145223	27.8889	27.9444	27.5556	0.3888	-1.383638	22.222200

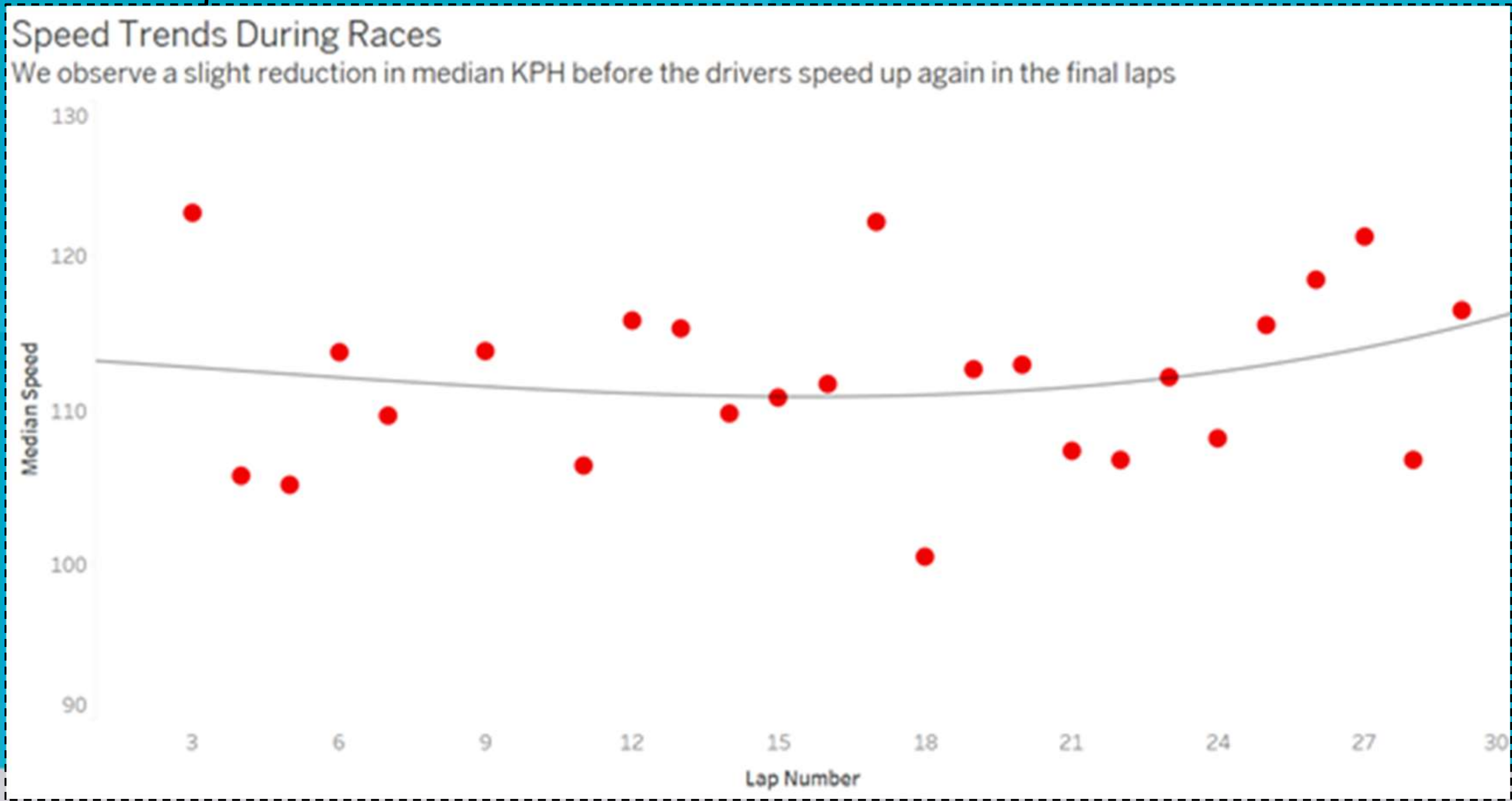
3.

Graphical Data Analysis

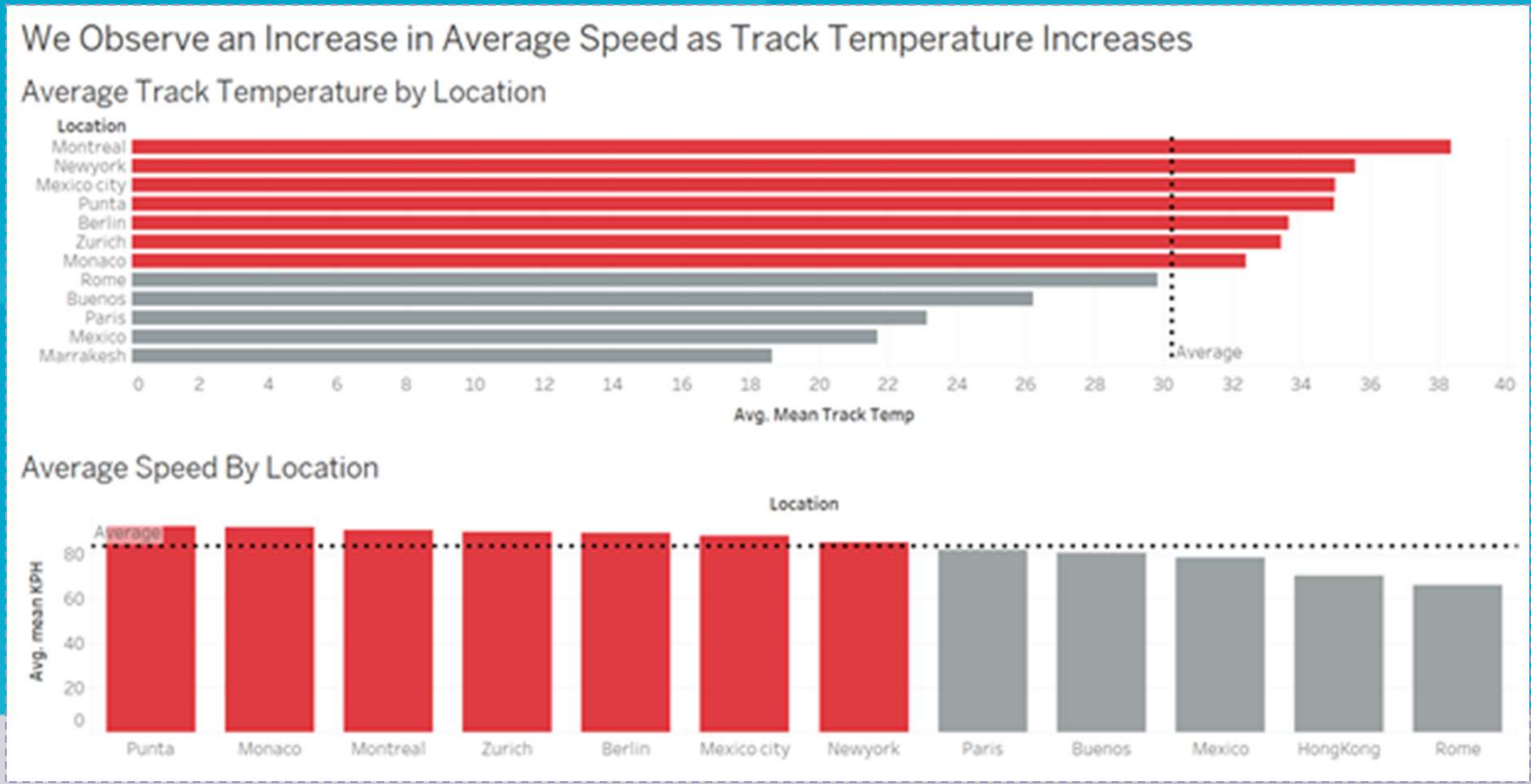
Faster Races on a Particular Day?



Overall Speed Trends



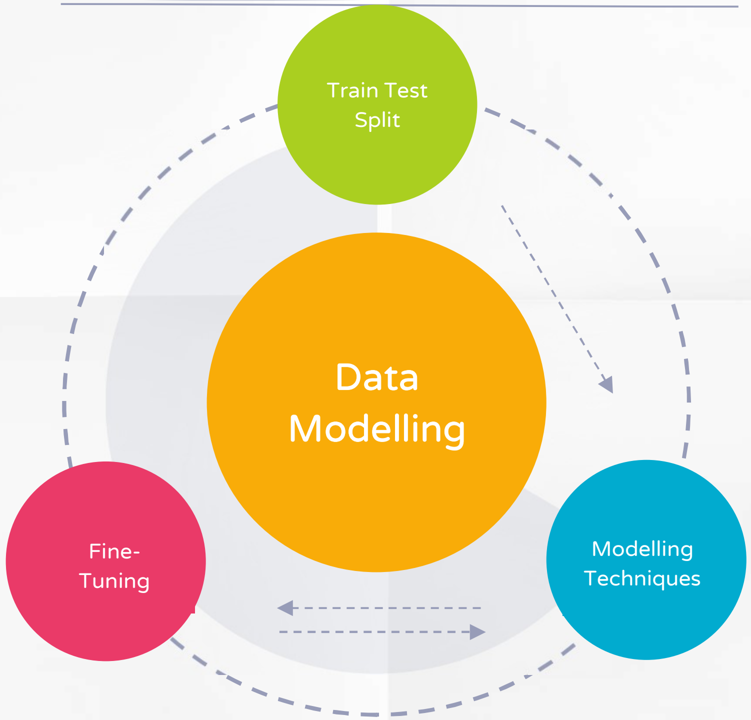
Does the wind help race faster?



4.

Data Modelling

In a Nutshell



Train Test Split

- Since we concatenated train and test data together in the beginning for data preprocessing purposes, eventually, we separated them out into train and test respectively.

- Train - 85%
- Test - 15%

```
Feature Training Dataset Size :  (2377, 182)
Predictor Variable Training Size :  (2377,)
Feature Testing Dataset Size :  (793, 182)
Predictor Variable Testing Size :  (793,)
```

Training Data : Independent Variables

```
Index(['GROUP', 'mean_KPH', 'std_KPH', 'med_KPH', 'max_KPH', 'min_KPH',
       'range_KPH', 'skew_KPH', 'mean_S1', 'std_S1',
       ...
       'TEAM_20', 'TEAM_21', 'TEAM_22', 'TEAM_23', 'TEAM_24', 'TEAM_25',
       'TEAM_26', 'TEAM_27', 'TEAM_28', 'TEAM_29'],
      dtype='object', length=182)
```

Training Data : Predictor Variable = Total Number of Laps

```
final_df['Total_Lap_Num'].tail()
```

```
3165    36.0
3166    19.0
3167    13.0
3168     4.0
3169    34.0
Name: Total_Lap_Num, dtype: float64
```

“

Remaining Laps =

Total Predicted Laps - Laps Already Completed

Data Models

Random Forest

- Number of estimators = 1000

K-Nearest Neighbours

- Number of neighbours= 5
- Algorithm = “kd_tree”
- Weights = “Distance”

ElasticNet

- Max Iterations = 10000

Weighted Average Ensemble Model

- Weights assigned to each baseline model using Linear Regression

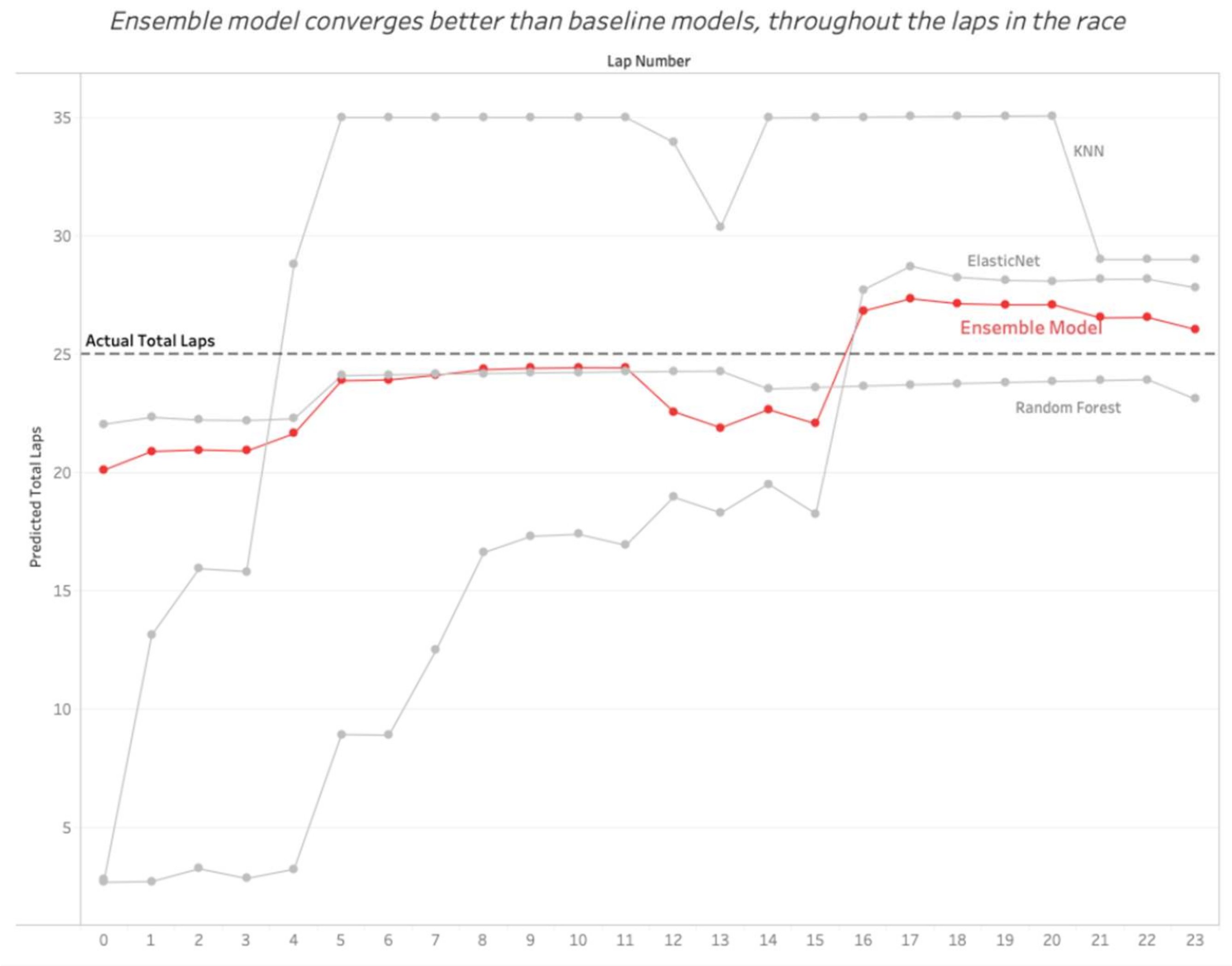
5.

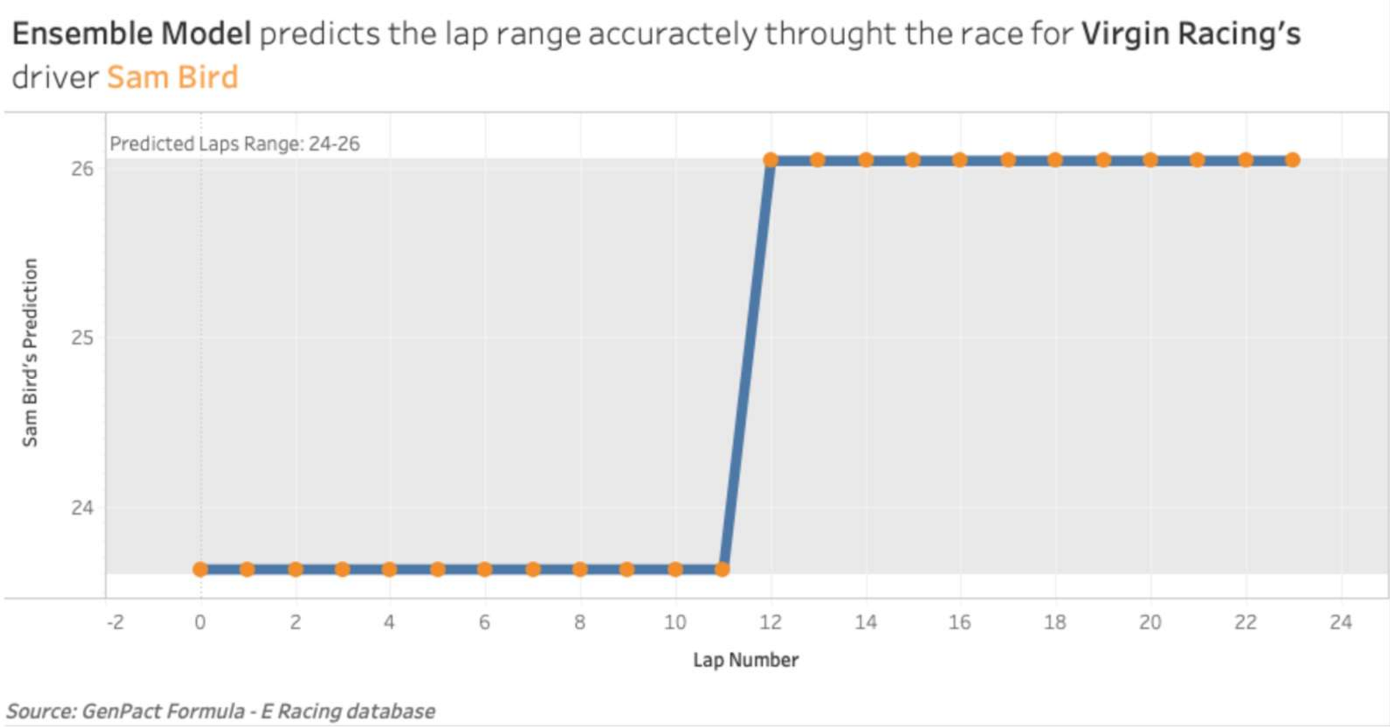
Model Evaluation and Results

Models with their accuracies (Excluding Weather Data)

	R2 values	MSE Score
Random Forest	0.9698	5.489
K Nearest Neighbours	0.8811	20.184
ElasticNet	0.7948	29.03

Model Output





Customized Predictions

6.

Conclusion & Future Recommendations

Conclusive Insights

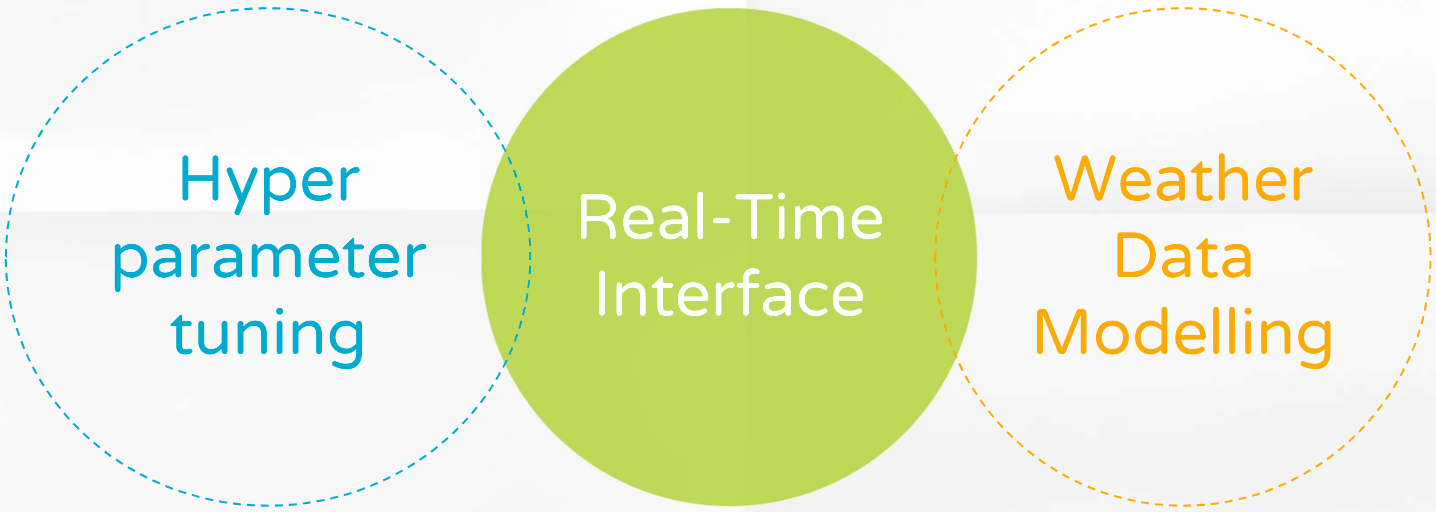
Insight 1

Weather definitely has an effect on the average speed of a racer. Specifically, a higher wind and temperature would lead to a slightly faster speed.

Insight 2

Out of the models we tried, Ensemble Model seems to fit the data best. At any given point, we can thus, predict the total laps in the race

Future Recommendations



Thank you

Any questions?