DW
Assignment 2

By: Ryan Liam

Problem Statement

- To predict whether a driver will be in the Top Half

Hypothesis on feature importance:

- Grid position is very important
- Weather should play a role
- Competing in their own country

Datasets

- Results
- Race information
- Driver information
- Constructor information
- Circuit information
- Qualifying information
- Weather (Web Scraped from the races URL)

Weather dataset

	season	round	circuit	weather	weather_warm	weather_cold	weather_dry	weather_wet	weather_cloudy
0	1950	1	silverstone	Sunny, Mild, Dry	0	0	1	0	0
1	1950	2	monaco	Soleggiato	1	0	0	0	0
2	1950	3	indianapolis	Rainy	0	0	0	1	0
3	1950	4	bremgarten	Warm, dry and sunny	1	0	1	0	0
4	1950	5	spa	Warm, dry and sunny	1	0	1	0	0
...	...	...	...	...	...	...	...	...	...
1013	2019	17	suzuka	Sunny	1	0	0	0	0
1014	2019	18	rodriguez	Partly cloudy	0	0	0	0	1
1015	2019	19	americas	Sunny	1	0	0	0	0
1016	2019	20	interlagos	Sunny	1	0	0	0	0
1017	2019	21	yas_marina	Clear	1	0	0	0	0

1018 rows × 9 columns

Incorrect Values

Missing Values

Altitude

- 0.219% missing
- No information about the Losail International Circuit altitude
- Solution: Complete Case Analysis

Q_best, Q_worse, Q_avg

- Most of the null values were removed from the qualifying time columns by creating these columns
- Still has 1.577% missing
- One race with all the qualifying time missing
- All of the other missing values are at the last positions
- Solution: Complete Case Analysis

Weather

- 6.04% missing
- Tried replacing with mode weather based on weather OHE by location
- But Complete Case Analysis performed better generally
- Solution: Complete Case Analysis

Outliers

- There are outliers in altitude, Q_best, Q_worse and Q_avg columns
- All the methods worsens the model performance, so I did not use any of them

Winsorization

- Worsens ML model performance

Trimming based on boundaries found via Standard Deviation Method

Categorical Encoding

- We have two categorical columns to encode
- One of which is Circuits which has 41 unique values
- Another is Constructors which has 46 unique values
- Ordinal and label encoding makes no sense
- Target mean encoding will cause target leakage
- One Hot Encoding better than Rare Encoding+OHE
- Thus, I use One Hot Encoding

One Hot Encoding

- Creates a lot of features
- Consume a lot of memory
- But performed the best

Rare Encoding + OHE

- Encodes the values that have observations below 5% as Rare
- Makes model less prone to overfitting
- Led to test accuracy being better than train accuracy(Should not happen)

Numerical Transformation

Transform Distribution To Normal

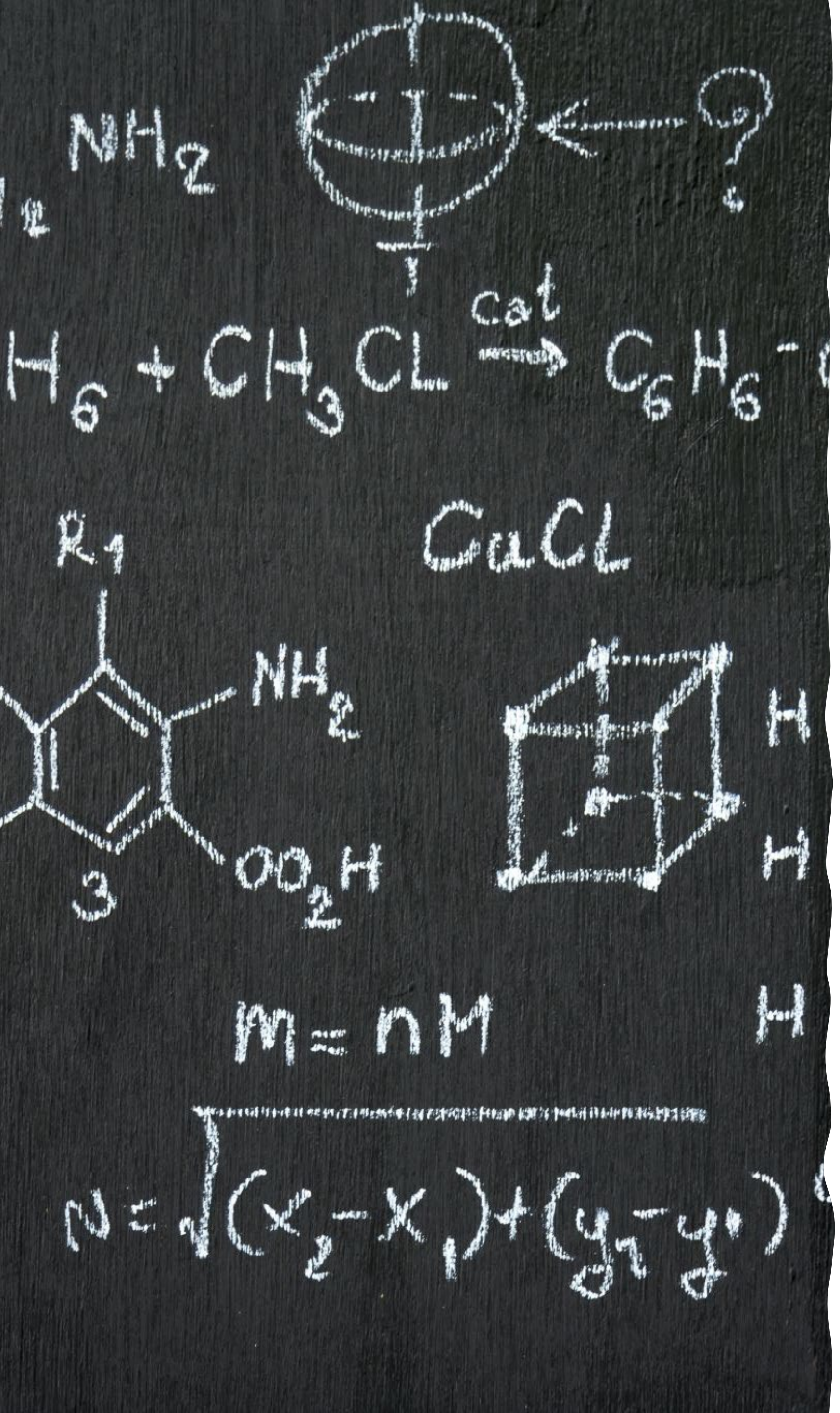
- Q_best, Q_worse and Q_avg distribution is slightly right-skewed
- Transforming it to normal may improve the model
- Log Transformation, Reciprocal Transformation, Square Cube Root, Yeo Johnson, and Box-Cox.
- All made the Linear Regression model perform worse
- **None of the transformation methods are used**

Discretization/Binning

- Discretize or Bin Q_best, Q_worse and Q_avg to reduce impact of outliers
- All made the Logistic Regression model perform worse
- **None of the transformation methods are used**

Feature Engineering (Create new Columns)

- Season (Derived from cleaned date column)
- Driver’s Full Name
- Driver’s Age
- Driver’s home (boolean)
- Constructors’ home (boolean)
- Best, worse and average qualifying time
- Top Half (target)



Scaling

- To have features within a similar scale
- Tried Standardization, Mean Normalization, MInMax Normalization, Mean Absolute Scaling and Robust scaling
- MinMax Normalization performed the best
- Thus, I use MinMax Normalization

Models Performance

Naïve Baseline

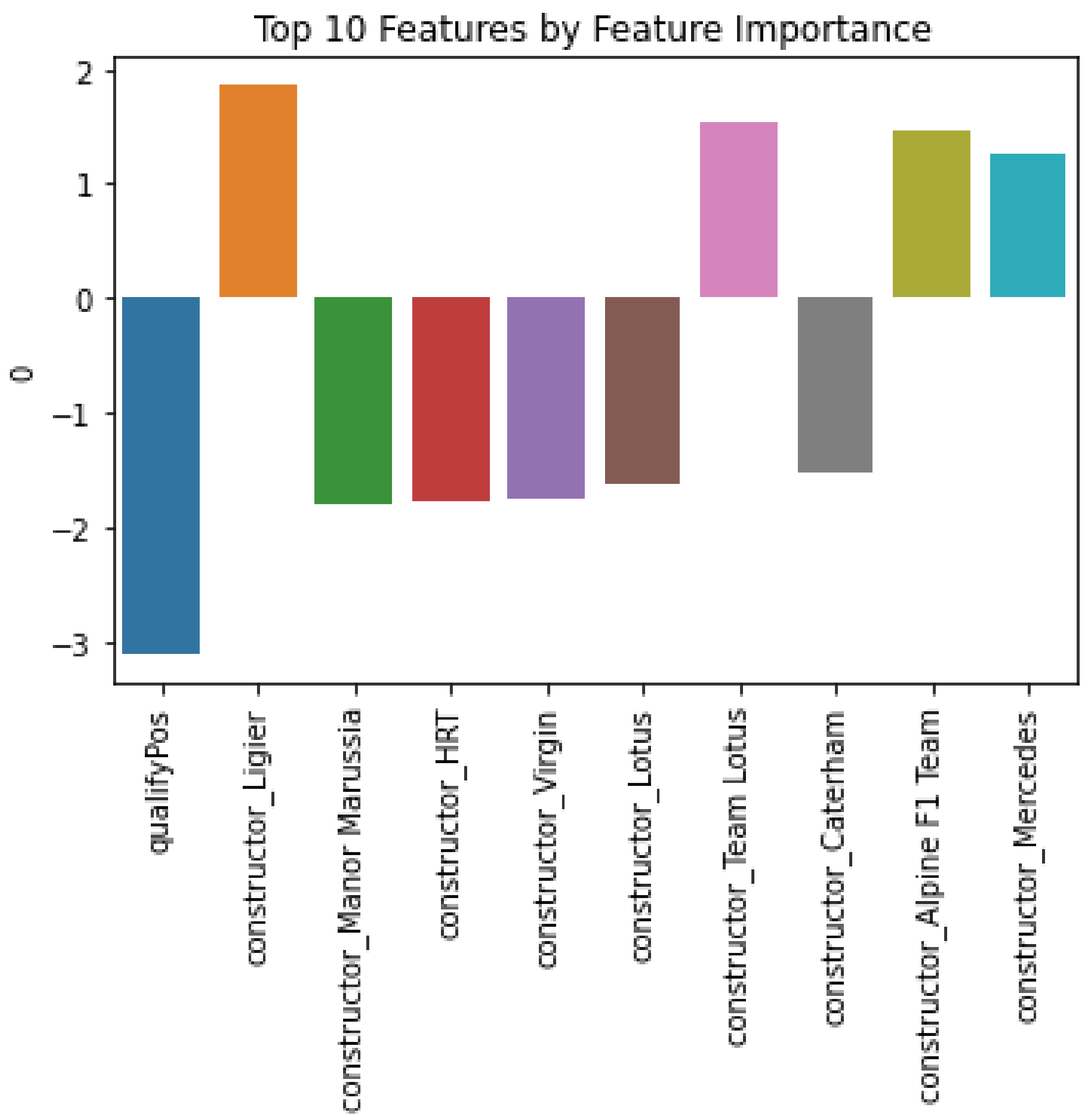
```
The Naive Baseline Model's accuracy on train data is 50.33%.
The Naive Baseline Model's accuracy on test data is 49.15%.
```

Logistic Regression

```
The LogReg Model's accuracy on train data is 73.59%.
The LogReg Model's accuracy on test data is 72.32%.
```

Feature Importance

- Most important feature is Qualifying Position
- Most of the constructors mentioned in the Top 10 Feature by importance have very little observations.



Possible Improvements

01

Further experiment with different combinations of encoding or transformation methods

02

Find better methods of imputing the missing values instead of just removing them

03

Look into the availability of more granular hourly weather data, which we can use.

04

Investigate why the constructors with lesser observations have high feature importance

The End


Santander
Santander
Santander
Santander
HUBLOT
TNT
KASPERSKY
PIRELLI
5