F1 Rank Prediction Analysis

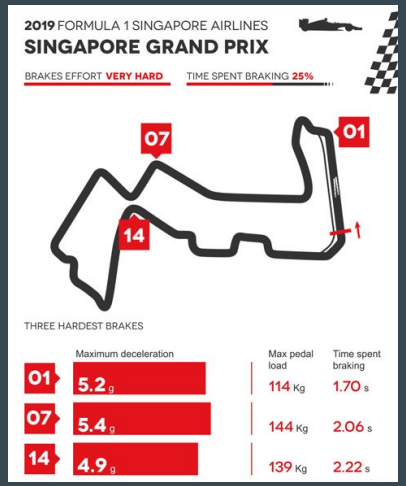
Team:
Rishabh Salwan
Ruolin Zhao
Sai venkat
Suraj Singh
Victoria Gonzalez

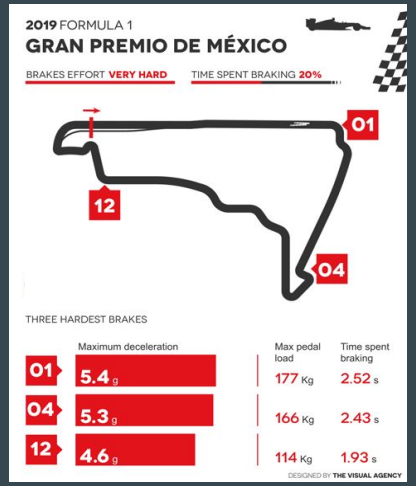
Table of Contents

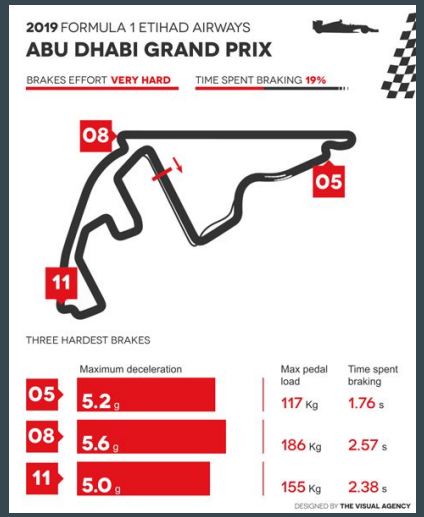
1. Intro to F1
2. Data
3. Feature Selection
4. Model Selection
5. Results
6. Season Prediction

Intro to F1

1. Circuits
2. Grid
3. Points System
4. Podium







Circuits

68 Circuits

22 Grand Prix every season Ex: Singapore, Mexico, Abu Dhabi, Azerbaijan etc.

Source: https://www.brembo.com/en/company/news/brembo-brake-analysis-abu-dhabi-formula-1

Home Advantage

Max Verstappen fans in the Dutch Grand Prix

Ferrari fans in the Italian Grand Prix

Grid

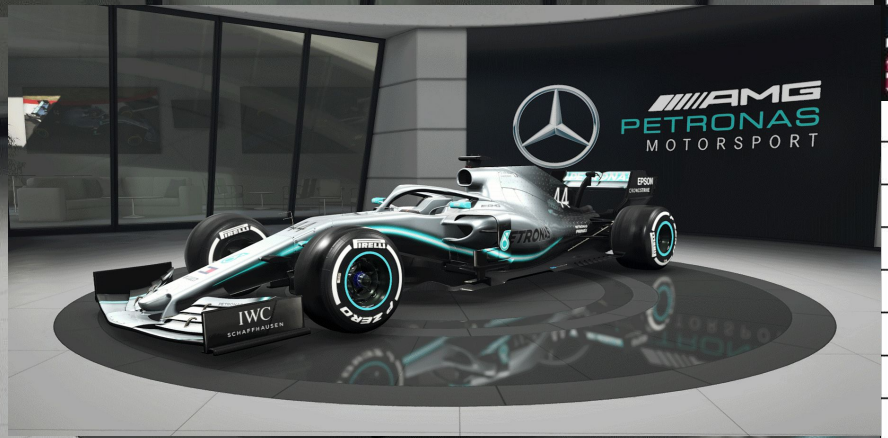
20 Drivers

20 positions

Pole position

Qualifying race

Points system



2022 DRIVER STANDINGS

AFTER JAPANESE GRAND PRIX

VERSTAPPEN	366
PEREZ	253
LECLERC	252
RUSSELL	207
SAINZ	202
HAMILTON	180
NORRIS	101
OCON	78
ALONSO	65
BOTTAS	46

2022 TEAM STANDINGS

AFTER JAPANESE GRAND PRIX

RED BULL RACING	619
FERRARI	454
MERCEDES-AMG	387
ALPINE	143
McLAREN	130
ALFA ROMEO	52
ASTON MARTIN	45
HAAS	34
ALPHATAURI	34
WILLIAMS	8

Positions	Points
1st Position	25 Points
2nd Position	18 Points
3rd Position	15 Points
4th Position	12 Points
5th Position	10 Points
6th Position	8 Points
7th Position	6 Points
8th Position	4 Points
9th Position	2 Points
10th Position	1 Point

Podium

Top 3 drivers in a race get the podium

Most Podiums - Lewis Hamilton

F1 Season, 2022

Driver Championship: Max Verstappen

Constructors Championship: Redbull

Data

- Data source: Formula 1 World Championship from 1950 till the latest 2021 season (Source: Kaggle)
- Races, drivers, constructors, qualifying, circuits, lap times, pit stops, championships
- Data cleaning and feature engineering: Race results, lap times, previous performances, home advantage etc. to predict **positions**.
- Most relevant feature in prediction: Win rate, past avg pos, Constructor home advantage

1	Race Performance Result	• Podium • Points • Pole Position • Milliseconds • Position
2	Historical Performances (Absolute)	• Past_n_year_points • Past_n_year_wins • Past_n_year_pole • Past_n_year_races • Past_n_year_podium_driver • **Past_n_year_avg_pos**
3	Historical Performances (Rates)	• **Win Rate** • Podium Rate • Points Rate • Pole Rate
4	Home Advantage	• Driver Home Advantage • **Constructor Home Advantage**

Our Model - Feature Selection

OLS Regression Results

Dep. Variable:	position	R-squared:	0.444
Model:	OLS	Adj. R-squared:	0.444
Method:	Least Squares	F-statistic:	3881.
Date:	Tue, 06 Dec 2022	Prob (F-statistic):	0.00
Time:	00:46:23	Log-Likelihood:	-39330.
No. Observations:	14599	AIC:	7.867e+04
Df Residuals:	14595	BIC:	7.870e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3499	0.094	14.398	0.000	1.166	1.534
win_rate	-1.1061	0.239	-4.635	0.000	-1.574	-0.638
Past_n_year_avg_pos	0.8363	0.010	85.168	0.000	0.817	0.856
constructor_home_adv	0.2083	0.106	1.966	0.049	0.001	0.416

Omnibus:	1020.288	Durbin-Watson:	1.069
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1523.579
Skew:	0.577	Prob(JB):	0.00
Kurtosis:	4.082	Cond. No.	73.4

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Modelling and Scoring

- Implement Multiple ML Algorithms
- Getting in-sample and out-of-sample RMSE
- Improvement in R2 or adjusted R2 value

Data Manipulation and Sport Understanding

- Creating meaningful features by data manipulation
- Adding features contingent on performance and conditions

	date	year	raceId	constructorId	driverId	Past_n_year_races	points	position	win_rate	Past_n_year_avg_pos	constructor_home_adv
0	2008-03-16	2008	18	5	1	16	10.0	1	0.250000	3.000000	0
1	2008-03-16	2008	18	8	2	15	8.0	2	0.000000	5.333333	0
2	2008-03-16	2008	18	...	3	14	6.0	3	0.000000	9.142857	0
3	2008-03-16	2008	18	...	4	16	5.0	4	0.250000	2.937500	0
4	2008-03-16	2008	18	5	5	16	4.0	5	0.000000	7.875000	0
...	...	...	...	...	...	...	...	...	...	...	...
14594	2022-03-20	2022	1074	5	846	21	0.0	15	0.000000	6.619048	0
14595	2022-03-20	2022	1074	9	849	19	0.0	16	0.000000	15.105263	0
14596	2022-03-20	2022	1074	7	807	0	0.0	17	0.000000	10.000000	0
14597	2022-03-20	2022	1074	3	815	20	0.0	18	0.050000	6.600000	0
14598	2022-03-20	2022	1074	3	830	19	0.0	19	0.526316	1.842105	0

14599 rows × 11 columns

	Past_n_year_podium_driver	Past_n_year_wins	Past_n_year_pole	points	win_rate	podium_rate	points_rate	pole_rate	winner
Past_n_year_podium_driver	1.000000	0.845604	0.817489	0.749963	0.847912	0.985815	0.944407	0.815696	0.496339
Past_n_year_wins	0.845604	1.000000	0.936935	0.616576	0.992022	0.811022	0.784615	0.923964	0.533398
Past_n_year_pole	0.817489	0.936935	1.000000	0.597625	0.926986	0.782466	0.750097	0.992328	0.522727
points	0.749963	0.616576	0.597625	1.000000	0.630430	0.754921	0.788201	0.604642	0.641595
win_rate	0.847912	0.992022	0.926986	0.630430	1.000000	0.830729	0.802294	0.927662	0.540928
podium_rate	0.985815	0.811022	0.782466	0.754921	0.830729	1.000000	0.953083	0.796467	0.490391
points_rate	0.944407	0.784615	0.750097	0.788201	0.802294	0.953083	1.000000	0.760986	0.471896
pole_rate	0.815696	0.923964	0.992328	0.604642	0.927662	0.796467	0.760986	1.000000	0.524595
winner	0.496339	0.533398	0.522727	0.641595	0.540928	0.490391	0.471896	0.524595	1.000000

Correlation Analysis & Statistical Significance

- Checking for multicollinearity
- Checking for statistical significance

Final Data Set

	date	year	raceId	constructorId	driverId	Past_n_year_races	points	position	win_rate	Past_n_year_avg_pos	constructor_home_adv
0	2008-03-16	2008	18	5	1	16	10.0	1	0.250000	3.000000	0
1	2008-03-16	2008	18	8	2	15	8.0	2	0.000000	5.333333	0
2	2008-03-16	2008	18	[illegible]	3	14	6.0	3	0.000000	9.142857	0
3	2008-03-16	2008	18	4	4	16	5.0	4	0.250000	2.937500	0
4	2008-03-16	2008	18	5	5	16	4.0	5	0.000000	7.875000	0
...	...	...	...	...	...	...	...	...	...	...	...
14594	2022-03-20	2022	1074	5	846	21	0.0	15	0.000000	6.619048	0
14595	2022-03-20	2022	1074	9	849	19	0.0	16	0.000000	15.105263	0
14596	2022-03-20	2022	1074	7	807	0	0.0	17	0.000000	10.000000	0
14597	2022-03-20	2022	1074	3	815	20	0.0	18	0.050000	6.600000	0
14598	2022-03-20	2022	1074	3	830	19	0.0	19	0.526316	1.842105	0

14599 rows × 11 columns

Model Selection

Random Model - Made a model to randomly rank drivers in each race

Train (2000 - first 5 races of 2021) and Test set (remaining races of 2021)

Model	Z Score Data Set	Absolute Value Data Set
Random Model	5.86	5.86
Linear Regression	2.70	2.70
Random Forest	3.54	3.54
KNN	3.57	3.57

P-value: **Linear regression** (2.78x10^28) > Random Forest > KNN (1.6x10^26)

Final Model Summary

OLS Regression Results

Dep. Variable:	position	R-squared:	0.444
Model:	OLS	Adj. R-squared:	0.444
Method:	Least Squares	F-statistic:	3881.
Date:	Tue, 06 Dec 2022	Prob (F-statistic):	0.00
Time:	00:46:23	Log-Likelihood:	-39330.
No. Observations:	14599	AIC:	7.867e+04
Df Residuals:	14595	BIC:	7.870e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3499	0.094	14.398	0.000	1.166	1.534
win_rate	-1.1061	0.239	-4.635	0.000	-1.574	-0.638
Past_n_year_avg_pos	0.8363	0.010	85.168	0.000	0.817	0.856
constructor_home_adv	0.2083	0.106	1.966	0.049	0.001	0.416

Omnibus:	1020.288	Durbin-Watson:	1.069
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1523.579
Skew:	0.577	Prob(JB):	0.00
Kurtosis:	4.082	Cond. No.	73.4

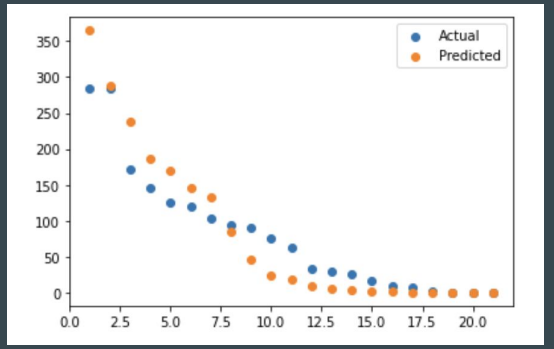
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

2021 Season Prediction Comparison

Pos	Our Prediction	Original Results
1	M. Verstappen	M. Verstappen
2	L. Hamilton	L. Hamilton
3	V. Bottas	V. Bottas
4	C. Sainz Jr	S. Perez
5	L. Norris	C. Sainz Jr
6	S. Perez	L. Norris
7	C. Leclerc	C. Leclerc
8	D. Ricciardo	D. Ricciardo
9	P. Gasly	P. Gasly
10	E. Ocon	F. Alonso

Prediction Accuracy: 80%

S. Perez changed teams from Racing Point to Redbull. This gave him a car advantage resulting in a better performance compared to his previous season, giving him Position 4 in original results and Position 6 in our prediction.



Thank You