Formula 1 Analysis

Neo Kok
University of Michigan - Ann Arbor

11/1/23

https://images.app.goo.gl/iDL7XaXd2DELrqW78

Project Outline

- Data
- Motivation
- Goals
    1. What effect has changing the point scoring system had?
    2. How has competitiveness changed over time?
    3. Predicting race results
- Methodology
- Results/Discussion

Project Outline

- Data
- Motivation
- Goals
  1. What effect has changing the point scoring system had?
  2. How has competitiveness changed over time?
  3. Predicting race results
- Methodology
- Results/Discussion

Data

- Related datasets on a variety of information about Formula 1 drivers, races, and results:
  - Driver standings
  - Lap times
  - Pit stops
  - Races
  - Race Results
- Data range varies between May 1950 - July 2023
- Sourced from Kaggle[1] on 10/25/23

[1] https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020/

Project Outline

- Data
- Motivation
- Goals
    1. What effect has changing the point scoring system had?
    2. How has competitiveness changed over time?
    3. Predicting race results
- Methodology
- Results/Discussion

Motivation

- Showcase data science skills
- Copious amounts of data in F1
- Real-world predictability
- Massive F1 fan for most of my life

Project Outline

- Data
- Motivation
- Goals
    1. What effect has changing the point scoring system had?
    2. How has competitiveness changed over time?
    3. Predicting race results
- Methodology
- Results/Discussion

Goals

- What effect has changing the point scoring system had?
    - Retrospectively re-adjusting points based on original and modern scoring systems
    - Comparing difference in leaderboards between scoring systems
- How has competitiveness changed over time?
    - Driver Championship standings
    - Constructors Championship standings
    - Teammates
- Predicting race results
    - Simple linear models
    - Logistic regression model

Project Outline

- Data
- Motivation
- Goals
  1. What effect has changing the point scoring system had?
  2. How has competitiveness changed over time?
  3. Predicting race results
- Methodology
- Results/Discussion

Methodology - Changing Point Systems

Original (1950s)

- Top 5 drivers get points
- 8 points for 1st
- 2 points for 5th
- 1 point for fastest lap
  - Only if driver is in top 5

Modern (2021-now)

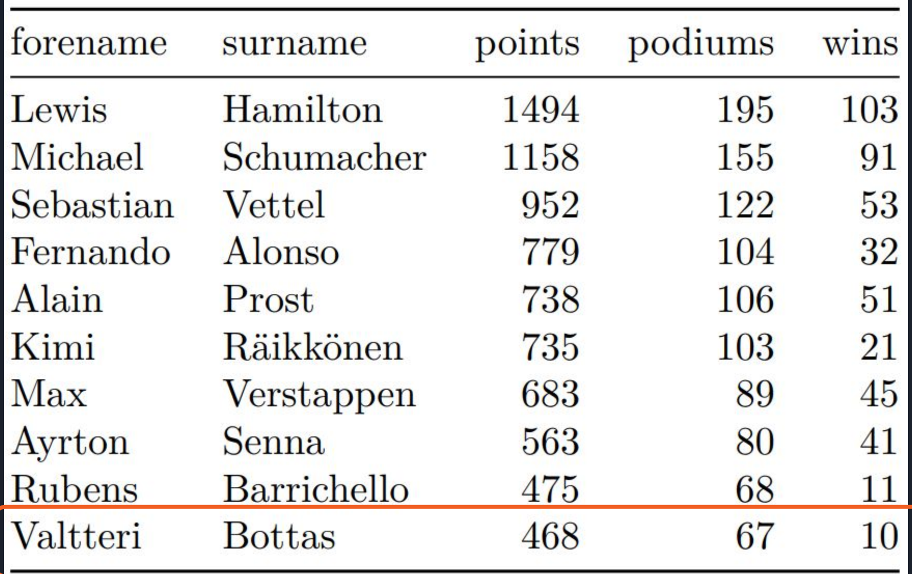
- Top 10 drivers get points
- 25 points for 1st
- 1 points for 5th
- 1 point for fastest lap
  - Only if driver is in top 10

- Re-calculated points for each driver in history with each point system
- Only analyzed original and modern systems - not anything in between
- Arranged drivers in order of respectively re-adjusted points
- Compared positions differences for top 50 drivers
- Compared driver inclusion for top 50 drivers

https://simple.wikipedia.org/wiki/List_of_Formula_One_World_Championship_points_scoring_systems

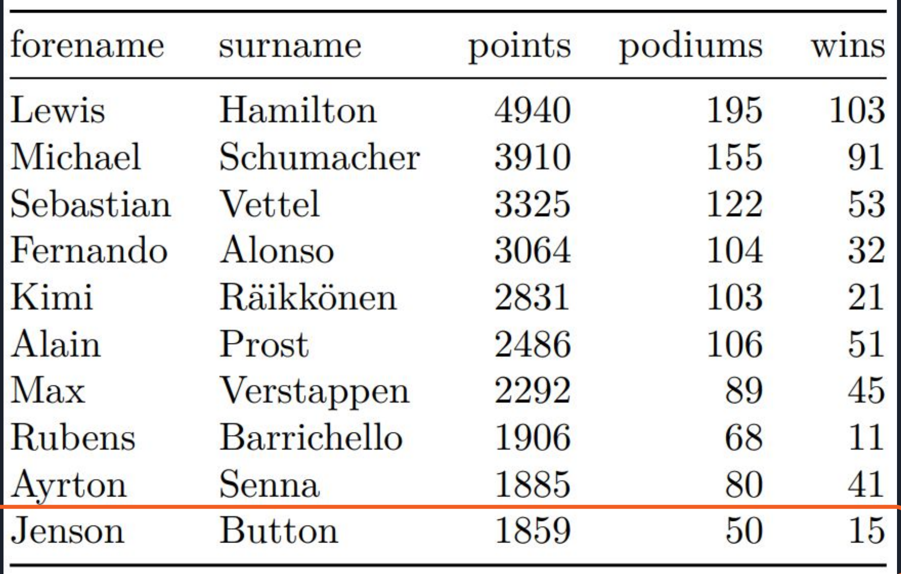
Project Outline

- Data
- Motivation
- Goals
    1. What effect has changing the point scoring system had?
    2. How has competitiveness changed over time?
    3. Predicting race results
- Methodology
- Results/Discussion



forename	surname	points	podiums	wins
Lewis	Hamilton	1494	195	103
Michael	Schumacher	1158	155	91
Sebastian	Vettel	952	122	53
Fernando	Alonso	779	104	32
Alain	Prost	738	106	51
Kimi	Räikkönen	735	103	21
Max	Verstappen	683	89	45
Ayrton	Senna	563	80	41
Rubens	Barrichello	475	68	11
Valtteri	Bottas	468	67	10

Top 10 with original scoring



forename	surname	points	podiums	wins
Lewis	Hamilton	4940	195	103
Michael	Schumacher	3910	155	91
Sebastian	Vettel	3325	122	53
Fernando	Alonso	3064	104	32
Kimi	Räikkönen	2831	103	21
Alain	Prost	2486	106	51
Max	Verstappen	2292	89	45
Rubens	Barrichello	1906	68	11
Ayrton	Senna	1885	80	41
Jenson	Button	1859	50	15

Top 10 with modern scoring

Results/Discussion - Adjusted Points Systems

- 10% of the top 50 drivers would not be in the top 50 if the point system stayed the same from 1950 to today
- 84% of the top 50 drivers would be in a different position in the leaderboards if the point system stayed the same from 1950 to today
- Differing point systems change rankings quite significantly

Project Outline

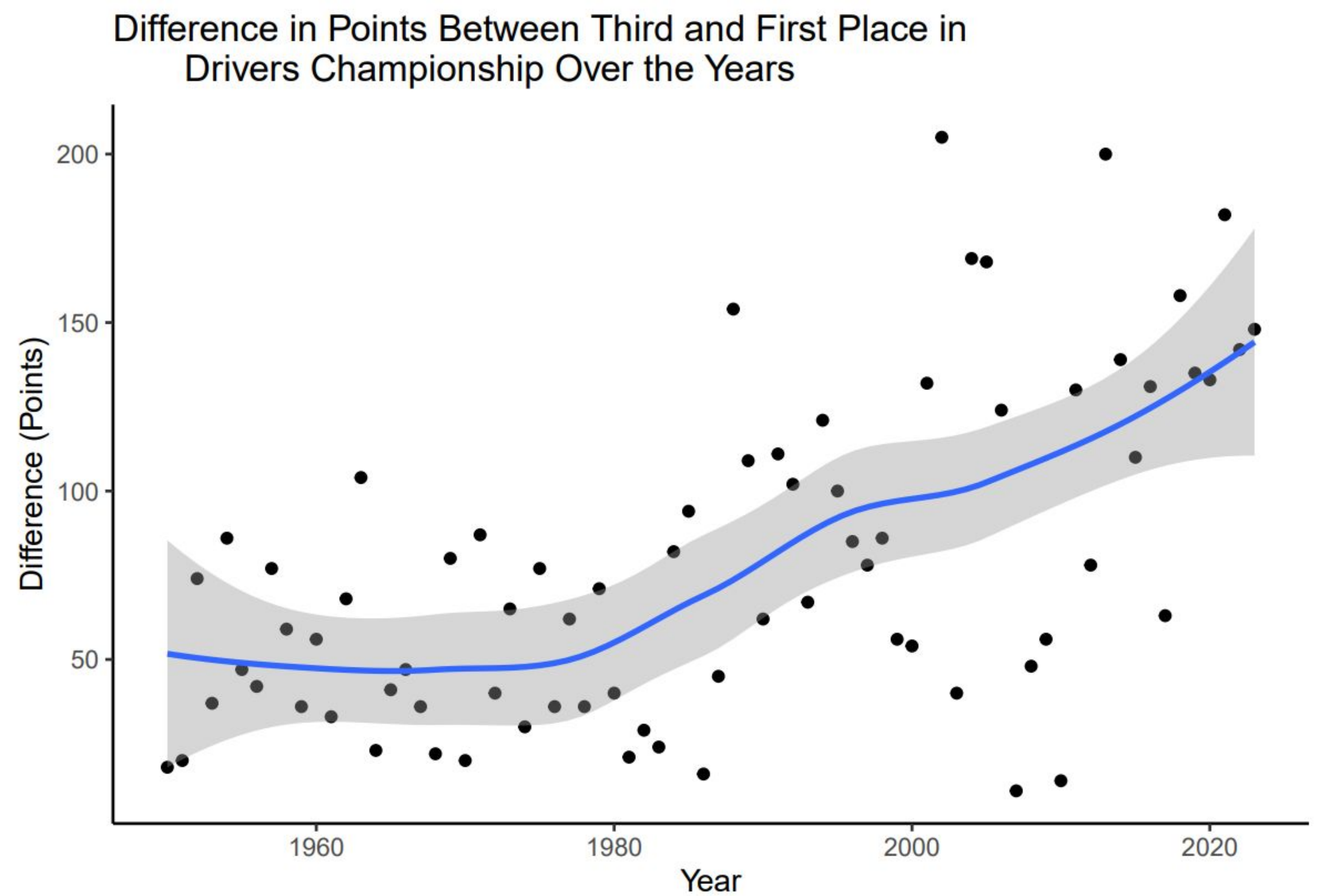
- Data
- Motivation
- Goals
    1. What effect has changing the point scoring system had?
    2. How has competitiveness changed over time?
    3. Predicting race results
- Methodology
- Results/Discussion

Methodology - Competitiveness Over Time

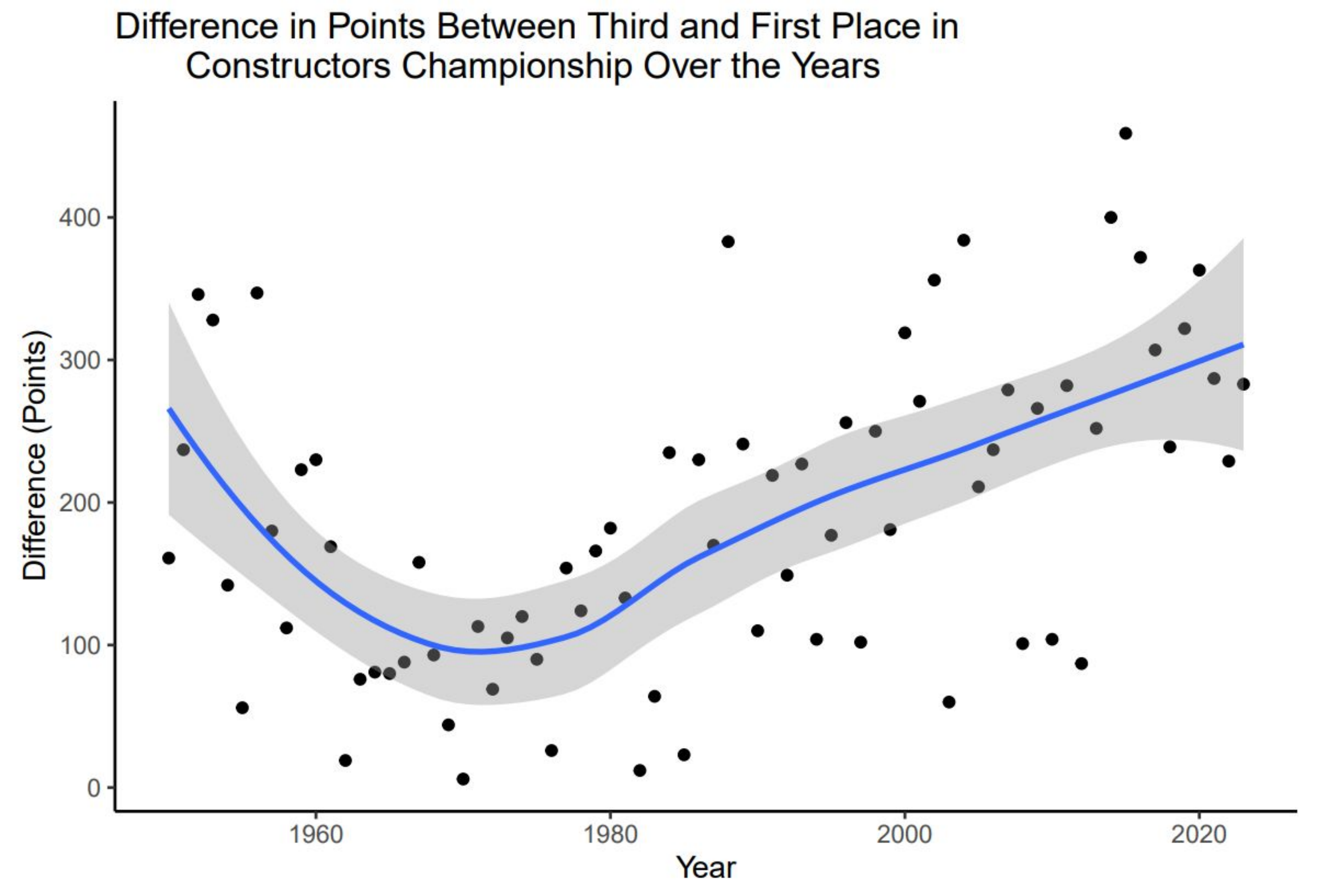
- Driver Championship standings
  - Calculate the difference in points between the top three drivers in the final Driver Championship standings for each year
- Constructors Championship standings
  - Calculate the difference in points between the top three constructors (teams) in the final Constructors Championship standings for each year
- Teammates
  - Calculate the average difference in points between the top two teammates for each constructor for each year
- Overall
  - Increased competitiveness defined as decreasing points differences over time
  - Plot point differences over time
  - All points are adjusted to reflect modern scoring system for consistency

Project Outline

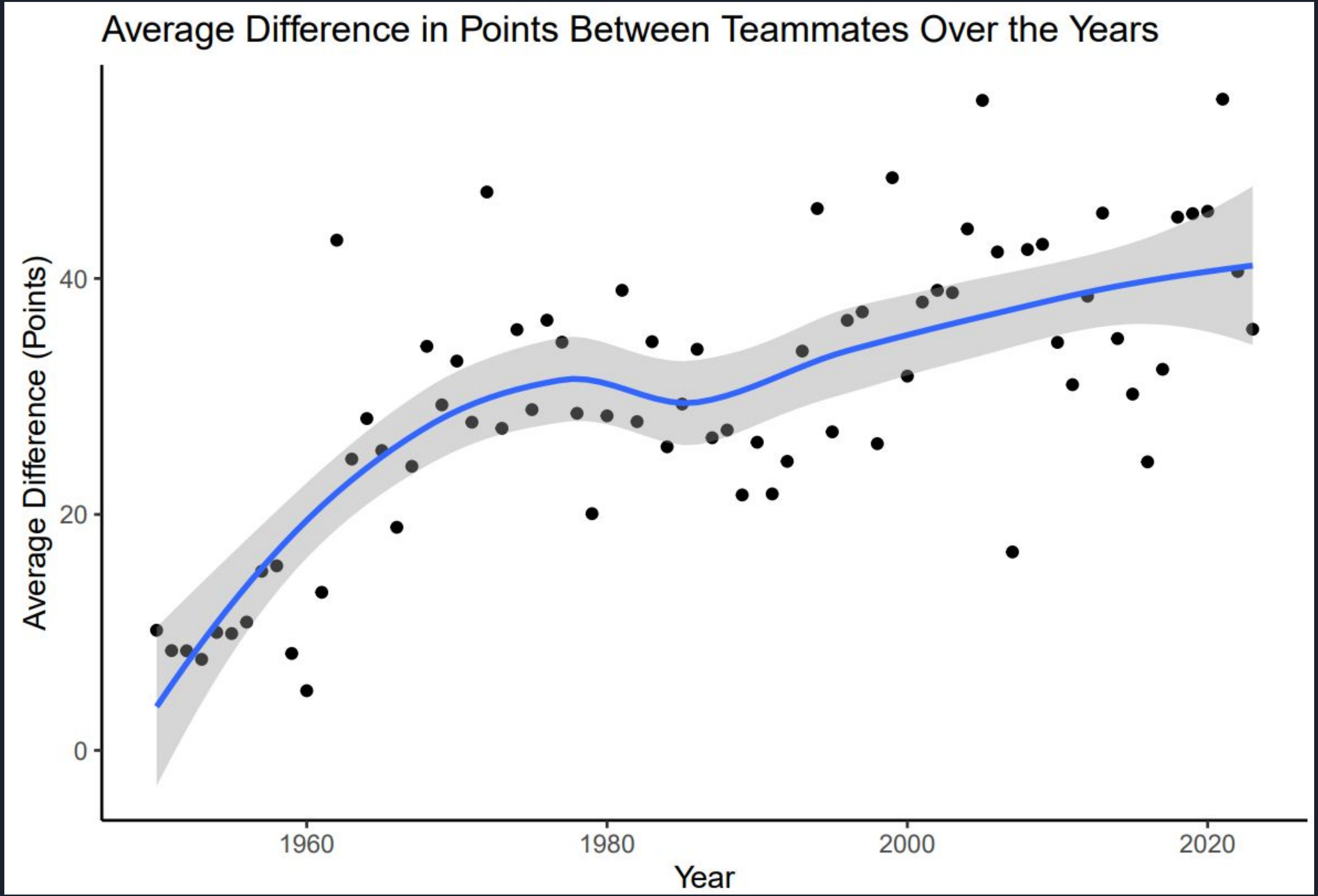
- Data
- Motivation
- Goals
    1. What effect has changing the point scoring system had?
    2. How has competitiveness changed over time?
    3. Predicting race results
- Methodology
- Results/Discussion



Relatively consistent decreasing trend in competitiveness in Drivers Championship over time



Initial increase followed by consistent decreasing trend in competitiveness in Constructors Championship over time



Initial sharp decrease followed by gradual decreasing trend in competitiveness between teammates over time

Result/Discussion - Competitiveness Over Time

- Increasing trend for point differences indicating reduced competitiveness across all three metrics
- Drivers and Constructors Championships loss in competitiveness likely stemming from increased reliance on vehicle performance
    - Single constructors’ vehicle dominance has been common in recent years
    - More gradual decrease in competitiveness between teammates in recent years supports this idea
- Indicates that F1 should focus on increasing competitiveness

Project Outline

- Data
- Motivation
- Goals
  1. What effect has changing the point scoring system had?
  2. How has competitiveness changed over time?
  3. Predicting race results
- Methodology
- Results/Discussion

Methodology - Predicting Race Results

- Data preparation
  - Joining multiple related datasets together
  - Manipulating and selecting variables of interest
    - Points won, win/no win, qualifying position, number of pit stops during race, average time of pit stops, Driver Championship position, fastest lap ranking
- Create simple linear models with all/some predictors of interest
  - Evaluate model fit
- Create logistic regression model to predict winner of an F1 race
  - Only use variables known before race start
  - Binary response variable: win/no win
  - Evaluate accuracy with train/test sets
  - Evaluate accuracy on 10/29/23 F1 race vs. expert predictions

Project Outline

- Data
- Motivation
- Goals
  1. What effect has changing the point scoring system had?
  2. How has competitiveness changed over time?
  3. Predicting race results
- Methodology
- Results/Discussion

Results/Discussion - Prediction using Simple Linear Models

- Full model using all variables - qualifying position, number of pit stops, average pit stop time, driver standing, and fastest lap ranking to predict points won in a race
  - *Points = 18.8 - 0.23 * Quali Position - 0.39 * # of Stops - 0.08 * Avg Stop Time - 0.44 * Driver Standing - 0.35 * Fastest Lap Rank*
  - Adjusted R^2 of 0.593
  - All variables p-value < 0.01
- Partial model using most significant variables - qualifying position, driver standing, and fastest lap ranking to predict points won in a race
  - *Points = 16.2 - 0.23 * Quali Position - 0.45 * Driver Standing - 0.33 * Fastest Lap Rank*
  - Adjusted R^2 of 0.589
- Useable model using variables known prior to race start - qualifying position and driver standing to predict points won in a race
  - *Points = 15.2 - 0.29 * Quali Position - 0.63 * Driver Standing*
  - Adjusted R^2 of 0.552

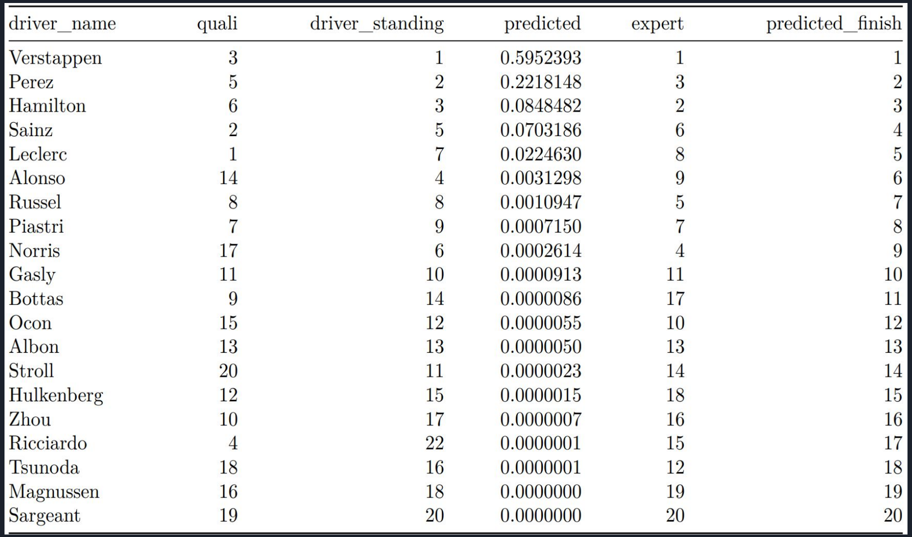
Results/Discussion - Prediction using Logistic Regression Model

- _Win = 1.54 - 0.33 * Quali Position - 0.75 * Driver Standing_
  - Log odds of winning
  - Probability > 50% considered win
- Model evaluation
  - In-sample accuracy using 70% train set = 96%
  - Out-of-sample accuracy using 30% test set = 96.5%
  - Seems super accurate
- Misleading overall accuracy
  - Accurately predicts winner 54% of the time
  - Exceptional at predicting non-winners (99%)
  - Overall, very good considering only two predictors

	True Win	True Loss	% Correct
Predicted Win	45	14	76.3%
Predicted Loss	38	1390	97.3%
% Correct	54.2%	99%	96.5%

Results/Discussion - Real-Life Application of Logistic Regression Model

- Model prediction of chance of winning
  - Verstappen 59.5% chance
  - Predicted finish based on chance of winning
- Predictions ran on 10/28/23 post-qualifying



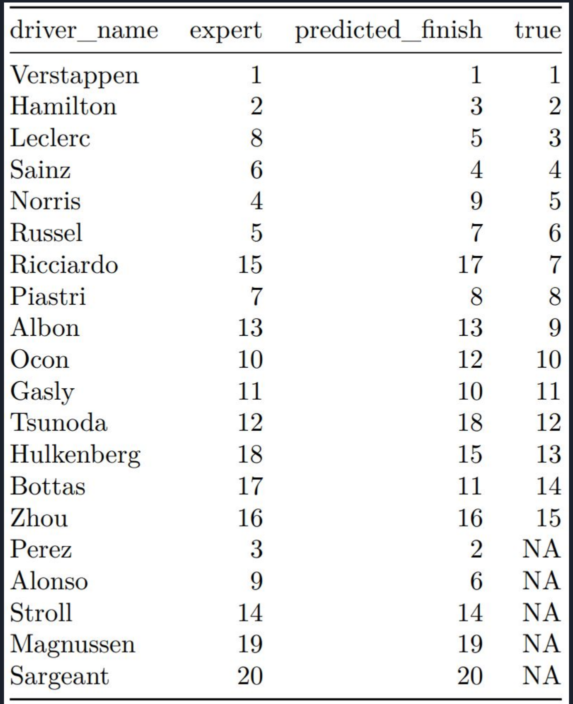
driver_name	quali	driver_standing	predicted	expert	predicted_finish
Verstappen	3	1	0.5952393	1	1
Perez	5	2	0.2218148	3	2
Hamilton	6	3	0.0848482	2	3
Sainz	2	5	0.0703186	6	4
Leclerc	1	7	0.0224630	8	5
Alonso	14	4	0.0031298	9	6
Russel	8	8	0.0010947	5	7
Piastri	7	9	0.0007150	7	8
Norris	17	6	0.0002614	4	9
Gasly	11	10	0.0000913	11	10
Bottas	9	14	0.0000086	17	11
Ocon	15	12	0.0000055	10	12
Albon	13	13	0.0000050	13	13
Stroll	20	11	0.0000023	14	14
Hulkenberg	12	15	0.0000015	18	15
Zhou	10	17	0.0000007	16	16
Ricciardo	4	22	0.0000001	15	17
Tsunoda	18	16	0.0000001	12	18
Magnussen	16	18	0.0000000	19	19
Sargeant	19	20	0.0000000	20	20

Driver name, qualifying position, driver standings, model prediction, expert predictions[2], and predicted finish. Arranged in order of model predicted chance of winning.

[2] https://www.cbssports.com/motor-sports/news/formula-1-picks-odds-race-time-2023-mexico-city-grand-prix-predictions-f1-best-bets-from-computer-model/

Results/Discussion - Real-Life Application of Logistic Regression Model

- Accurately predicted winner of 10/29/23 Mexico Grand Prix
- Raw predictions
  - Expert correct predictions: 25%
  - Model correct predictions: 15%
  - Expert average absolute error: 1.55
  - Model average absolute error: 1.85
- Removing DNF drivers
  - Expert correct predictions: 20%
  - Model correct predictions: 20%
  - Expert average absolute error: 2.27
  - Model average absolute error: 2
- For such a simple model, performed quite well relative to expert predictions



driver_name	expert	predicted_finish	true
Verstappen	1	1	1
Hamilton	2	3	2
Leclerc	8	5	3
Sainz	6	4	4
Norris	4	9	5
Russel	5	7	6
Ricciardo	15	17	7
Piastri	7	8	8
Albon	13	13	9
Ocon	10	12	10
Gasly	11	10	11
Tsunoda	12	18	12
Hulkenberg	18	15	13
Bottas	17	11	14
Zhou	16	16	15
Perez	3	2	NA
Alonso	9	6	NA
Stroll	14	14	NA
Magnussen	19	19	NA
Sargeant	20	20	NA

Driver name, expert predictions, model predictions, and actual finish. Arranged in order of actual finish.

Recap

- Differing point systems change leaderboard rankings quite significantly
- Increasing trend for point differences indicating reduced competitiveness across all three metrics
- Logistic regression model can predict race winners with relatively good accuracy using only two predictors

Future Analysis

- Exploring how the differing point scoring systems would have changed the winner of Drivers and Constructors Championships
- Investigating which factors influence competitiveness in Formula 1
- Testing more/different predictors to increase performance of predicting race winners using logistic regression models