DATA INTEGRATION & TRANSFORMATION

Group 40 – Shaun & Nigel

Formula 1 DNF Prediction

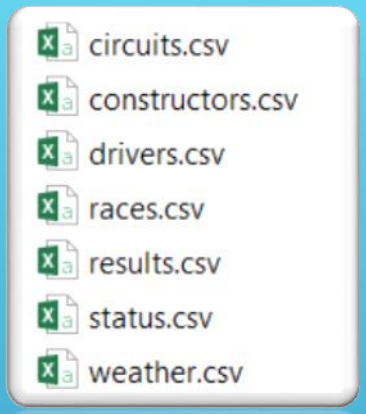
INTRODUCTION & GOAL

- A DNF is a fairly common scenario in which a driver fails to finish a race for any reason.
- We will be predicting the chances of this happening to a certain driver using various attributes
  - Team
  - Engine
  - Weather
  - Grid Position (i.e. what position they start the race in)
  - Circuit

Our dataset dates all the way back to 1950 onwards, although we will only be taking race results from 2003 onwards.

OUR DATASET

- Dataset initially downloaded from Kaggle
- Tables included
  - Race results; Drivers; Teams; Circuit Info; 'Status' IDs; Weather, etc…
  - We removed completely irrelevant tables like Lap Times, Pit stop info



- We are mainly concerned with the Race Results and Status ID tables, however we will be using them all

- The Race Results table *also* contained many attributes which wouldn't affect the outcome of a race. These were removed.

driverId	constructc	circuitId	grid	weatherIc	statusId
18	23	1	1	1	1
22	23	1	2	1	1
15	7	1	20	1	1
10	7	1	19	1	1
4	4	1	10	1	1
3	3	1	5	1	1
67	5	1	13	1	1
7	5	1	17	1	1
16	10	1	16	1	1
2	2	1	9	1	1
21	10	1	15	1	1
17	9	1	8	1	1
20	9	1	3	1	3
9	2	1	4	1	3
8	6	1	7	1	2

→

statusId	status
1	Finished
2	Car Failure/Retired
3	Crash

TRANSFORMATION

Aggregation, Generalisation

statusId	status
1	Finished
2	Disqualified
3	Accident
4	Collision
5	Engine
6	Gearbox
7	Transmission
8	Clutch
9	Hydraulics
10	Electrical
11	+1 Lap
12	+2 Laps
13	+3 Laps
14	+4 Laps
15	+5 Laps
16	+6 Laps
17	+7 Laps
18	+8 Laps
19	+9 Laps
20	Spun off
21	Radiator
22	Suspension
23	Brakes
24	Differential
25	Overheating
26	Mechanical
27	Tyre
28	Driver Seat
29	Puncture

statusId	status
1	Finished
2	Car Failure/Retired
3	Crash

Since statusID is actually the class itself that we are predicting, having 100s of them was just silly.

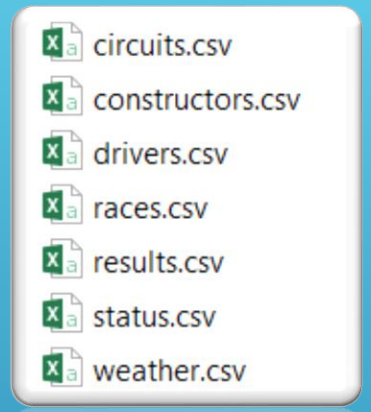
weatherIc	weather
1	Dry
2	Intermitting
3	Light Rain
4	Heavy Rain
5	Torrential

weatherIc	weather
1	Dry
2	Wet

You might ask why we'd throw 4 different weather conditions under the 'Wet' umbrella...

INTEGRATION

resultId	raceId	driverId	constructo	circuitId	grid	weatherIc	statusId
7554	1	18	23	1	1	1	1
7555	1	22	23	1	2	1	1
7556	1	15	7	1	20	1	1
7557	1	10	7	1	19	1	1
7558	1	4	4	1	10	1	1
7559	1	3	3	1	5	1	1
7560	1	67	5	1	13	1	1
7561	1	7	5	1	17	1	1
7562	1	16	10	1	16	1	1
7563	1	2	2	1	9	1	1
7564	1	21	10	1	15	1	1
7565	1	17	9	1	8	1	1
7566	1	20	9	1	3	1	3
7567	1	9	2	1	4	1	3
7568	1	8	6	1	7	1	2
7569	1	13	6	1	6	1	2
7570	1	12	4	1	14	1	3
7571	1	6	3	1	11	1	3
7572	1	5	1	1	12	1	3
7573	1	1	1	1	18	1	2
7574	2	18	23	2	1	2	1
7575	2	2	2	2	10	2	1
7576	2	10	7	2	3	2	1
7577	2	15	7	2	2	2	1
7578	2	22	23	2	8	2	1
7579	2	17	9	2	5	2	1

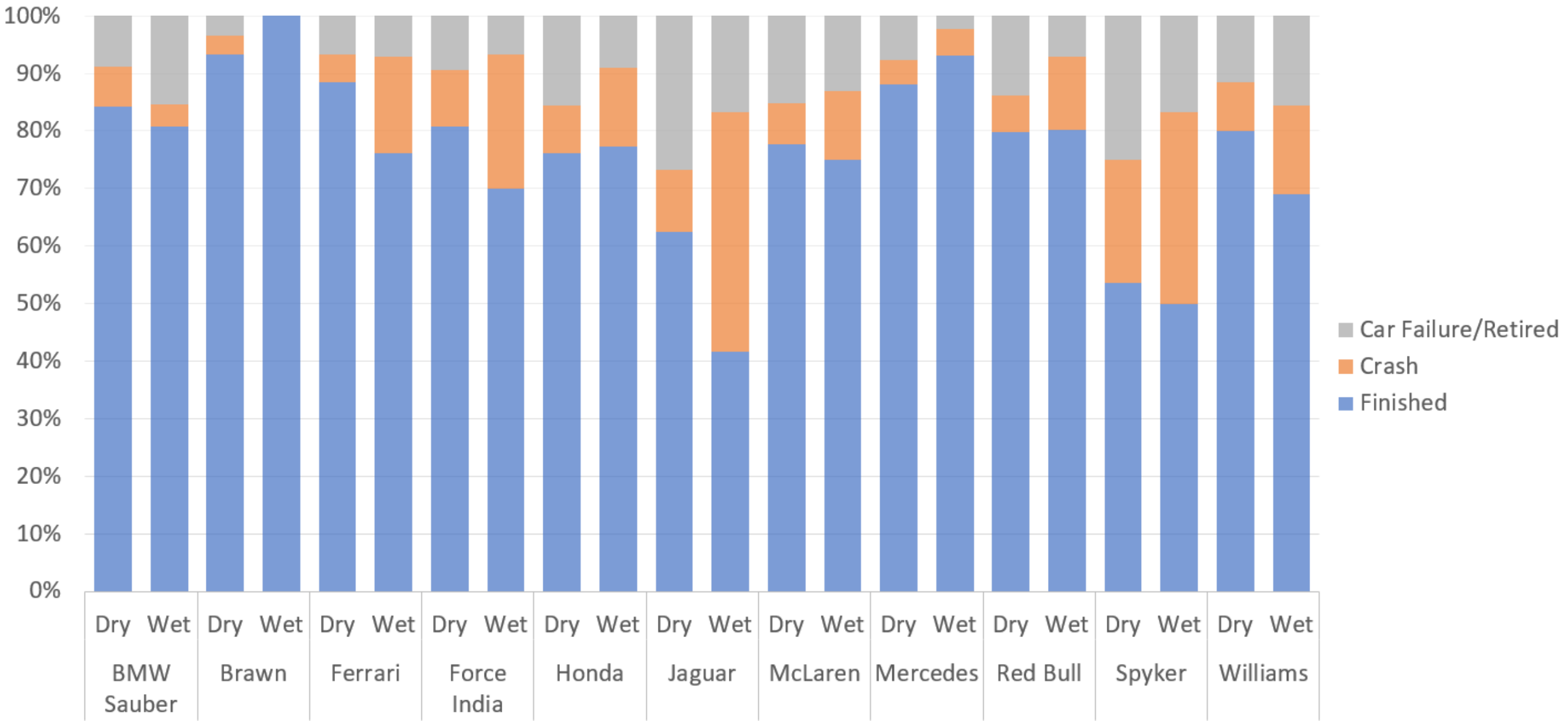


- Too many IDs, hard to interpret?
- We can integrate the data tables. Why?
  - View the data as categorical not numerical
  - Perform data mining over a universal set
  - The table to the left is before integration took place.

INTEGRATED VIEW

Driver	Team	Start_Pos	Circuit	Weather	Result
Michael Schumacher	Ferrari	1	A1-Ring	Dry	Finished
Kimi Raikkonen	McLaren	2	A1-Ring	Dry	Finished
Juan Pablo Montoya	Williams	3	A1-Ring	Dry	Car Failure/Retired
Nick Heidfeld	Sauber	4	A1-Ring	Dry	Car Failure/Retired
Rubens Barrichello	Ferrari	5	A1-Ring	Dry	Finished
Jarno Trulli	Renault	6	A1-Ring	Dry	Finished
Jenson Button	BAR	7	A1-Ring	Dry	Finished
Antonio Pizzonia	Jaguar	8	A1-Ring	Dry	Finished
Giancarlo Fisichella	Jordan	9	A1-Ring	Dry	Car Failure/Retired
Ralf Schumacher	Williams	10	A1-Ring	Dry	Finished
Olivier Panis	Toyota	11	A1-Ring	Dry	Car Failure/Retired
Jacques Villeneuve	BAR	12	A1-Ring	Dry	Finished
Cristiano da Matta	Toyota	13	A1-Ring	Dry	Finished
David Coulthard	McLaren	14	A1-Ring	Dry	Finished
Heinz-Harald Frentzen	Sauber	15	A1-Ring	Dry	Car Failure/Retired
Ralph Firman	Jordan	16	A1-Ring	Dry	Finished
Mark Webber	Jaguar	17	A1-Ring	Dry	Finished
Justin Wilson	Minardi	18	A1-Ring	Dry	Finished
Fernando Alonso	Renault	19	A1-Ring	Dry	Car Failure/Retired
Jos Verstappen	Minardi	20	A1-Ring	Dry	Car Failure/Retired

SAMPLE SET OF TEAMS



WHAT WE LEARNED

- Derived conclusions that:
  - Driver is responsible for crashes, team responsible for mechanical etc.
  - A single race is either Dry or Wet and not a mix of both together

- Removal of attributes how does this affect our results?
  - Advantage: Removal of data not affecting our goal state of DNF.
  - Disadvantage: Possibly miss out on valuable information.
- Transformation techniques gave us a clear idea of a goal state.

- Integrating our dataset - provides us with a unified view to produce analytics and statistics.
- Human readable while still machine readable
- Possible Ramifications of Pre-processing. How will our trained set match test set?


Thanks for listening!
7UP
Brooksfield
Ireland
Brooksfield