Digital Futures

F1 DRIVER RETENTION PREDICTION

JAMES MURPHY

AGENDA

- PROJECT BACKGROUND
- IMPORTANCE OF DRIVER PREDICTION
- DATA CLEANING AND MANIPULATION
- MODEL TUNING
- MODEL OUTCOMES
- CONCLUSIONS

PROJECT BACKGROUND

- Formula 1 is the highest class of open wheel racing
- F1 World Championship contested for over 70 years
- Billions exchanged in television rights and team sponsorship
- Viewership at a joint all time high
- 2.3 million UK viewers for 2021 championship decider

IMPORTANCE OF DRIVER PREDICTION

- Nearly impossible to predict a race due to extenuating factors
- Making driver predictions is also important
  - Driver underperforming
  - Useful for teams to get best value
  - Betting markets driver transfer

CHALLENGES OF DRIVER PREDICTION

- Driver kept and released based on more than merit
- Internal and external politics often at play
- Data not shared in public domain

DATA CLEANING AND MANIPULATION

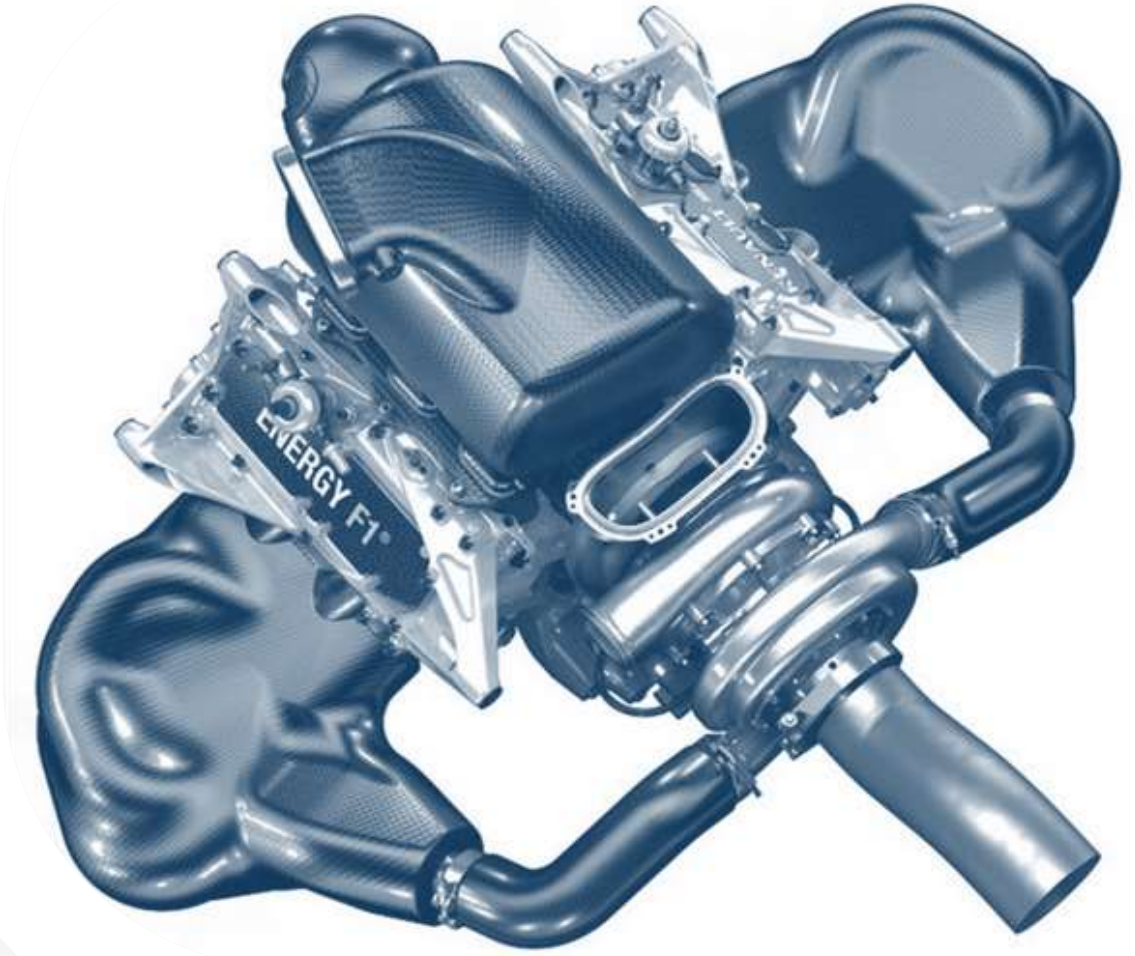
DATA TARGET

- Input data: drivers, races, results, standings, constructors
- Data for each season compiled
- Data that showed season results with multiple associated attributes
- DriverID and Year ‘composite key’

Year	DriverID	Attributes	driver_move
2002	1	…	?
2002	2	…	?
2003	1	…	?
2003	3	…	?

DATA OBSERVATIONS

- Earlier seasons of F1 have less uniform entry restrictions
- One race drivers and manufacturers
- > 100 drivers in 10 races, now 21 in 22 races in 2021
- Take last 20 years of data

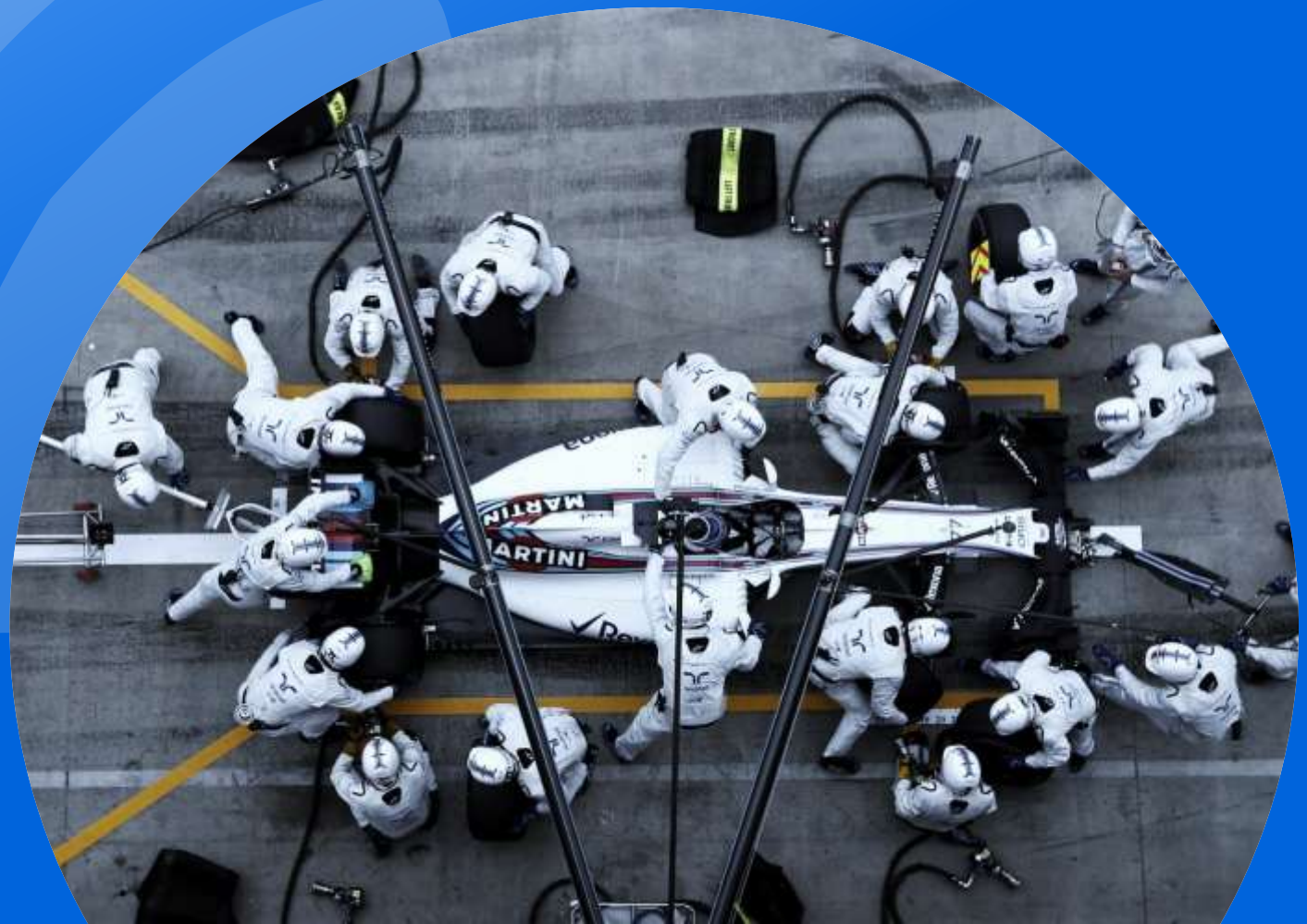


GENERATING “driver_move” COLUMN

- No given value, can be calculated
- Modal constructor for each year
- Matching driver and constructor for consecutive seasons
- Leaves = 1; Remains = 0
- Final year data accounted for manually
- 481 data entries

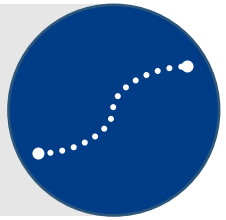
Leave?	No. of Drivers	%
0	256	53.2
1	225	46.8

MODEL TUNING



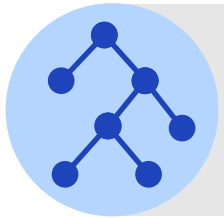
MODEL COMPARISON

LOGISTIC REGRESSION



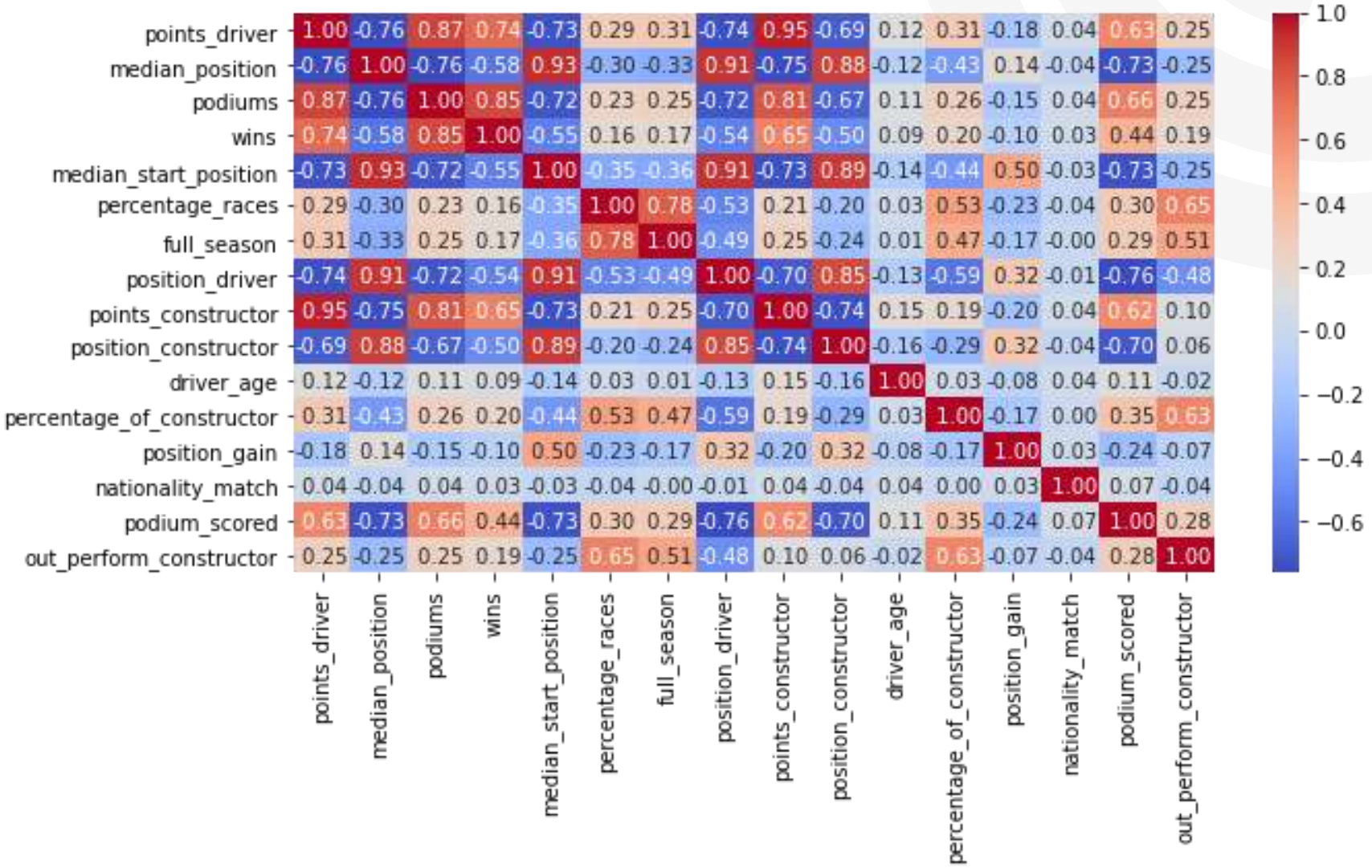
- Divides into discrete binary targets
- Interpret results as probabilities of the outcome
- Single decisions boundary
- Data manipulation needed for continuous data

DECISION TREE AND RANDOM FOREST

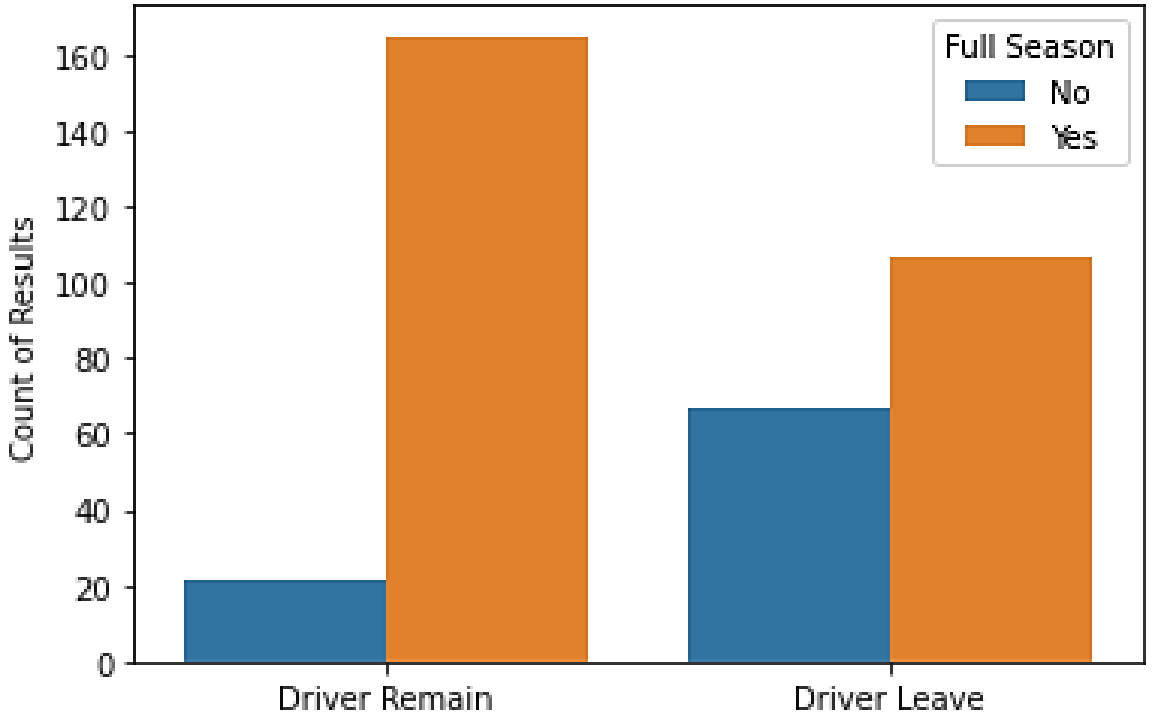


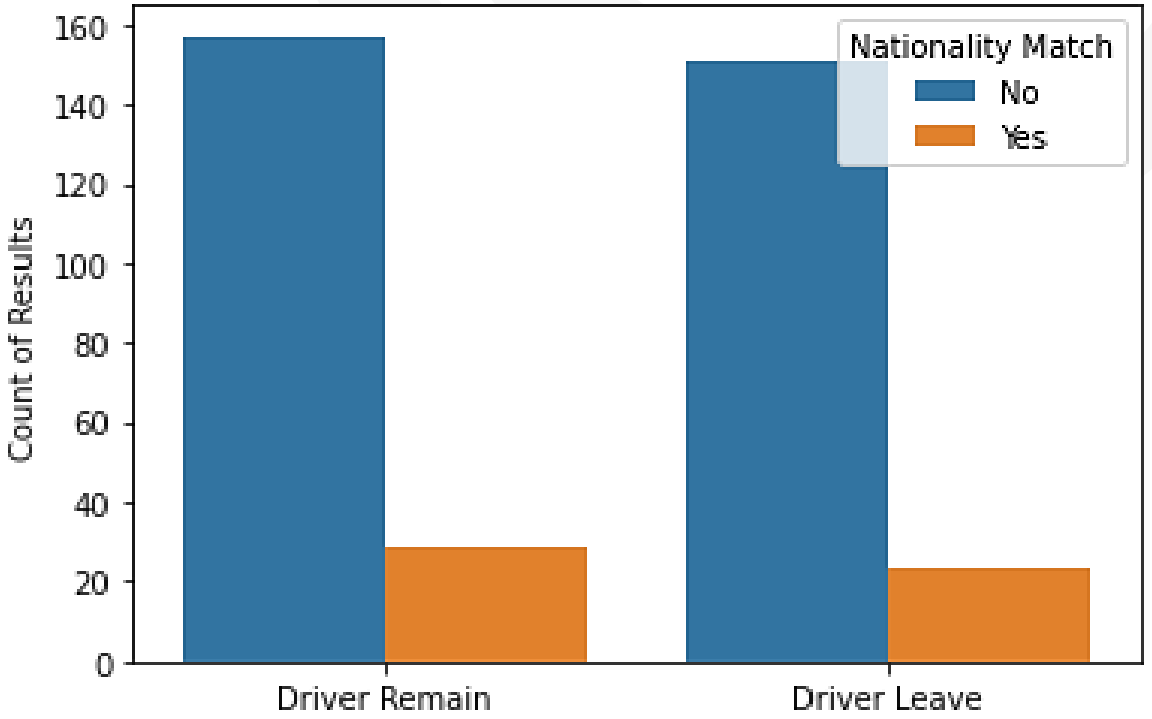
- Decision trees partition features using boundaries
- Node splits the data into branches
- Leaf represents a decision
- Random forest use decisions from multiple decision trees for output
- Data overfitting common

CORRELATION HEATMAP

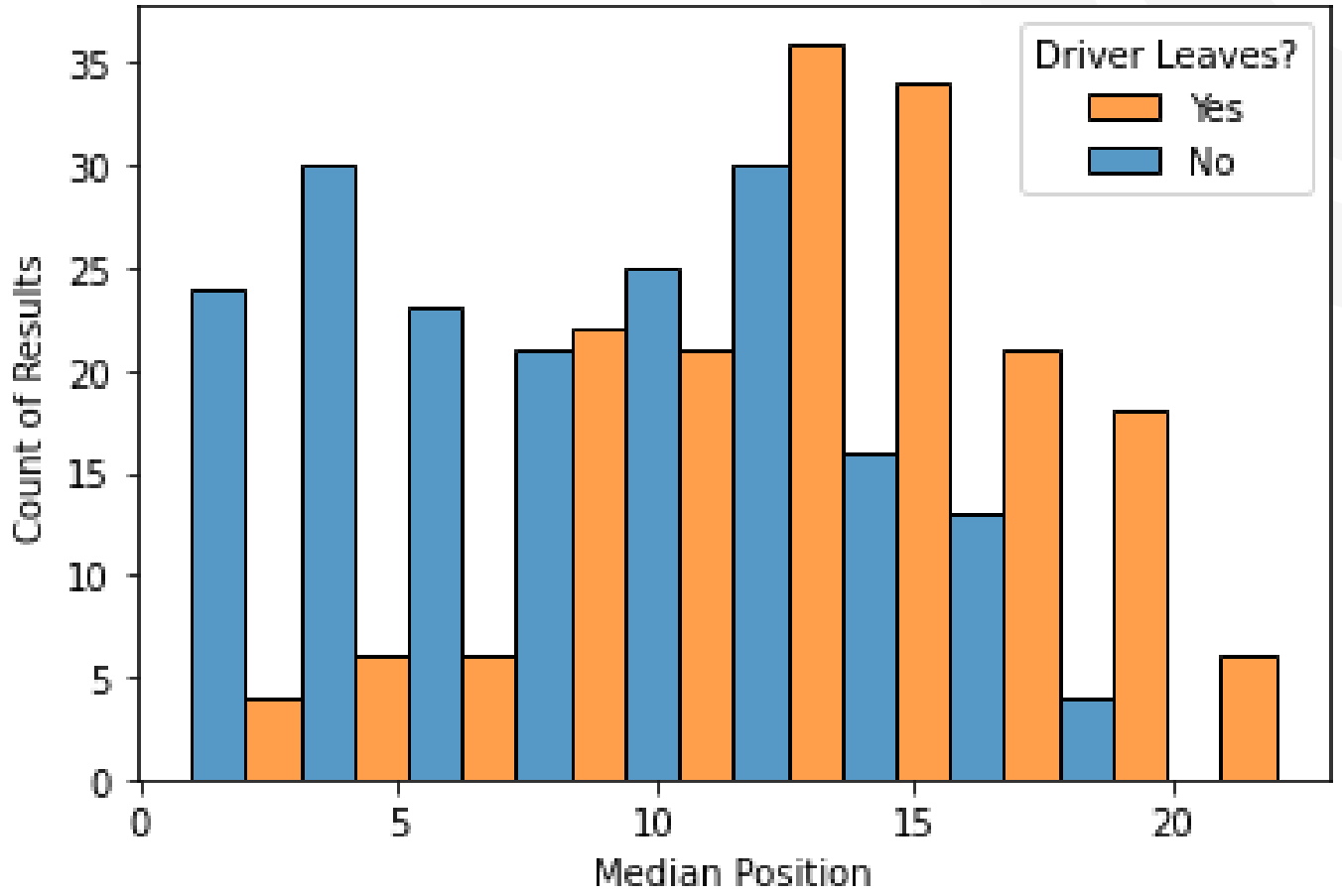


DISCRETE BINARY VARIABLES





CONTINUOUS VARIABLES

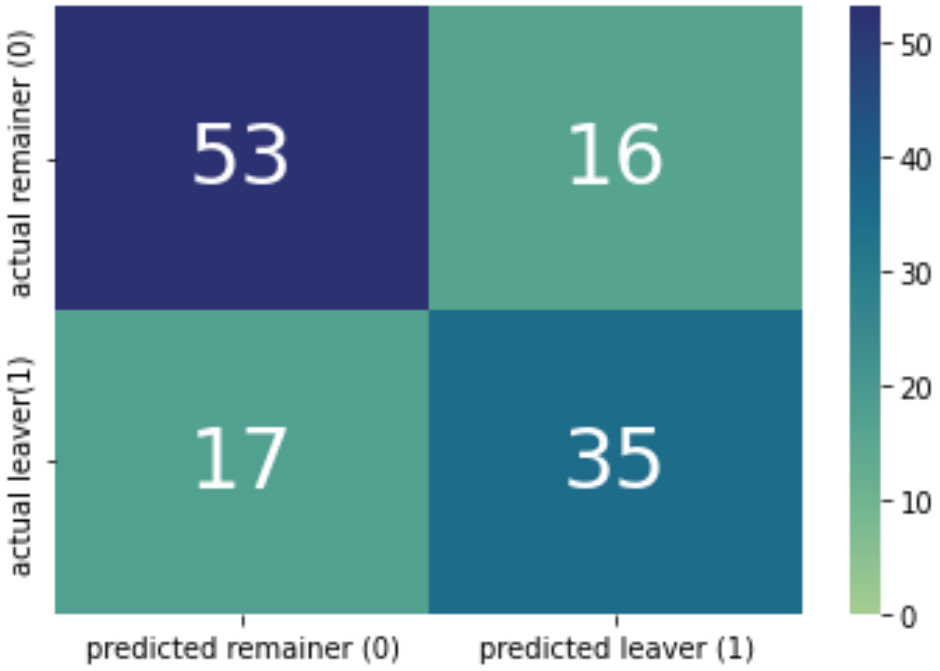


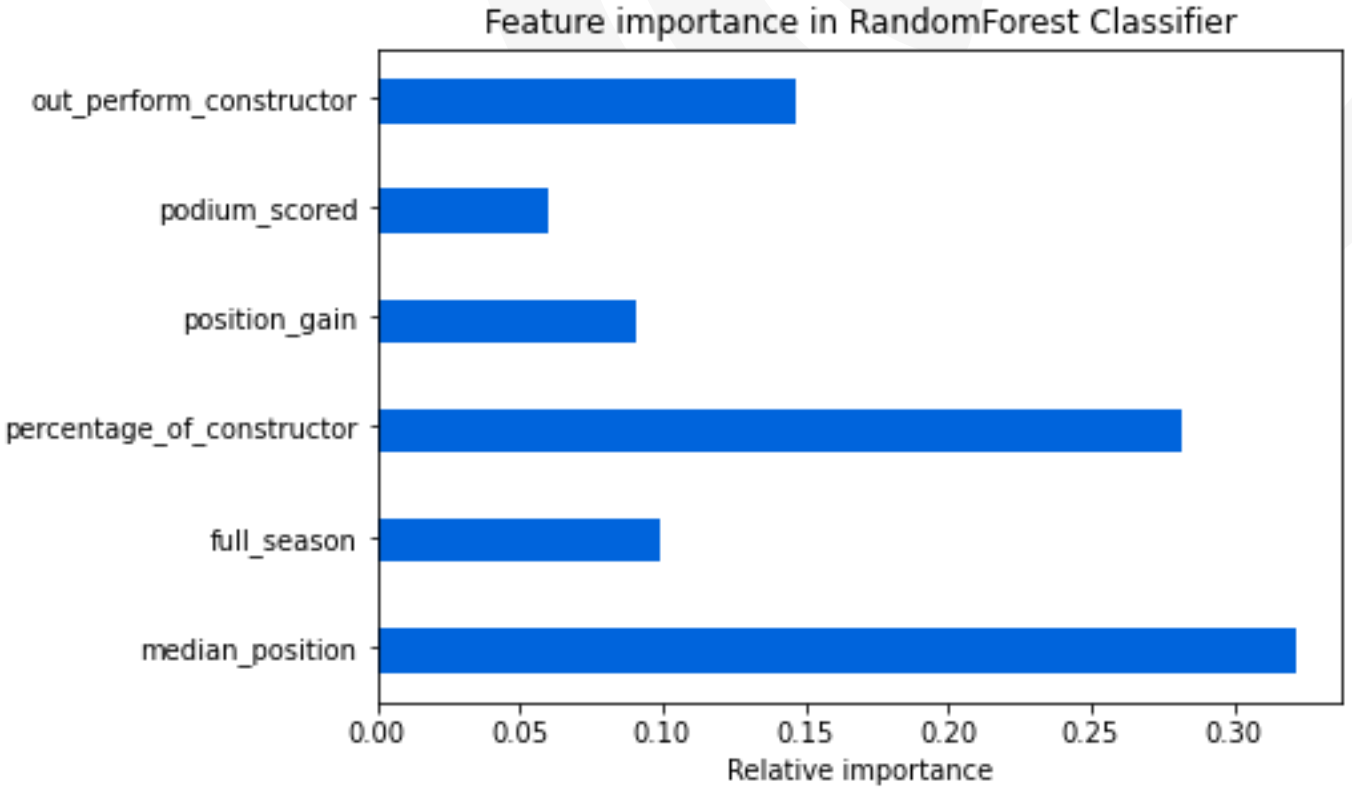
MODEL OUTCOMES

MODEL SCORES

Model	Train Accuracy	Test Accuracy	Test Precision	Test Recall	Test F1 Score
Starting LogReg	69.2	65.1	63.7	52.8	57.7
LogReg w/ FE	68.3	71.1	64.8	68.6	66.7
Decision Trees	71.9	69.4	65.2	58.8	61.9
Random Forest	**78.3**	**72.7**	**67.3**	**68.6**	**68.0**
IMPROVEMENTS	+ 9.1	+ 7.6			

MATRIX AND FEATURE IMPORTANCE





CONCLUSIONS

MODEL SUCCESSES

- 72.7% accuracy, aiming for around 75%
- Generate positive predictions despite limited data
- Solution to new data and problem
- Data not overly fit

MODEL LIMITATIONS

- Makes predictions based on a whole season’s results
- Limited data to model on
  - 481 total, 360 for train set
  - Reliability and schedule consistency too low pre 2000
- Points system changes, affects a reliable driver success metric
- New team branding recorded as driver leaving

PROJECT IMPROVEMENTS

- Identify a team name change
- Further metrics to measure driver
- Have a clustering or multiple outcome result
  - Retirement, promotion, relegation

ACKNOWLEDGEMENTS

- Project training and assistance
  - Lisa Carpenter – Data Science Training Lead
  - Blair Young – Data Science Instructor and Engagement Manager
- Data Origin
  - kaggle.com/rohanrao/formula-1-world-championship-1950-2020?select=results.csv
  - Compiled from ergast.com/mrd
- Photo credits
  - Title Slide: Clive Mason/Getty Images; media.bleacherreport.com
  - Slide 1: Bryn Lennon/Getty Images
  - Slide 2: clickandraces.com
  - Slide 3, 4: vanityfair.it
  - Slide 5, 6: wonderfulengineering.com
  - Slide 7: 2.bp.blogspot.com
  - Slide 9: wired.co.uk
  - Slide 14: Alex Treinitz/Motorsport Images
  - Slide 17: Motorsport Images
  - Slide 19: Peter Kohalmi/Reuters
  - Slide 20: Rainer Schlegelmilch/Motorsport Images
  - Slide 22: David Phipps/Motorsport Images

Digital Futures

THANK YOU

jemurphyuk@gmail.com

www.linkedin.com/in/james-murphy-96082b173/