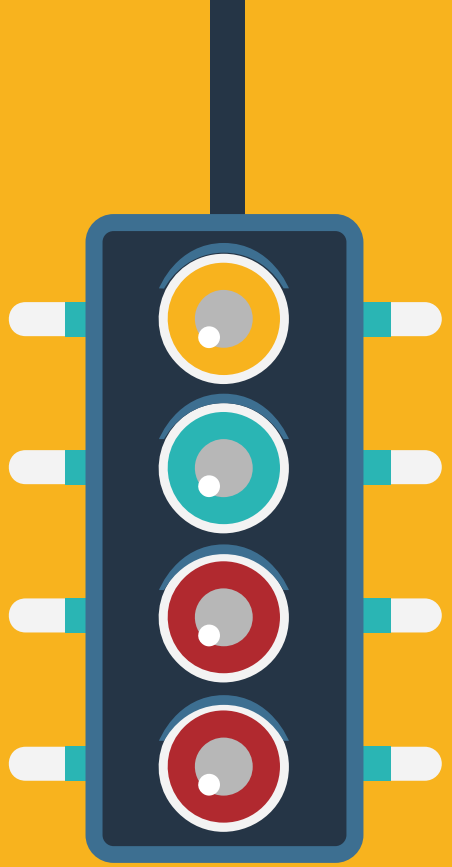
Formula 1: Drive to Supervised Learning

Braden Taack

TABLE OF CONTENTS

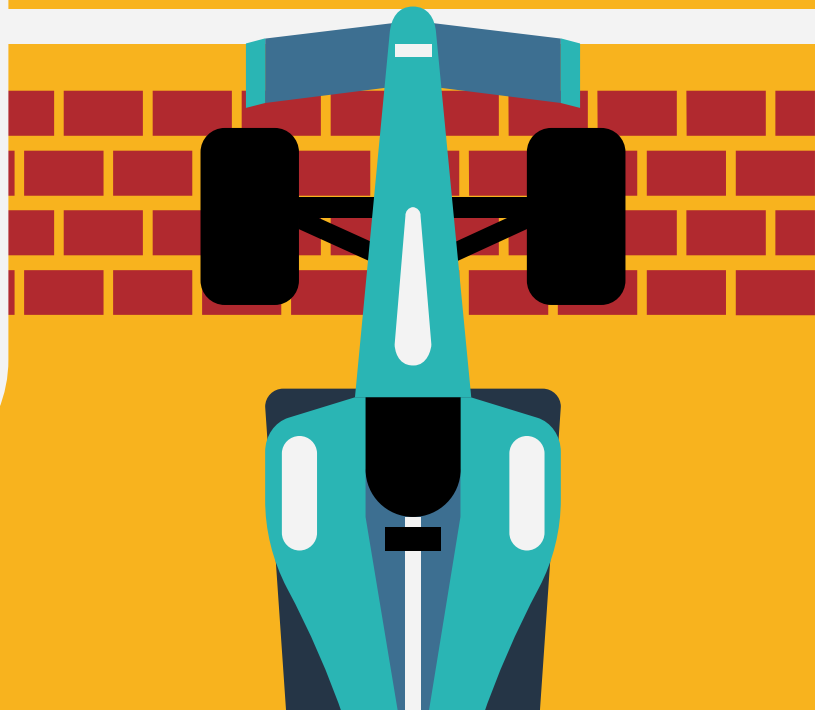
01

Introduction



Motivation: What does it take to be a winner? Where can teams focus in during practice trials for the best performance?

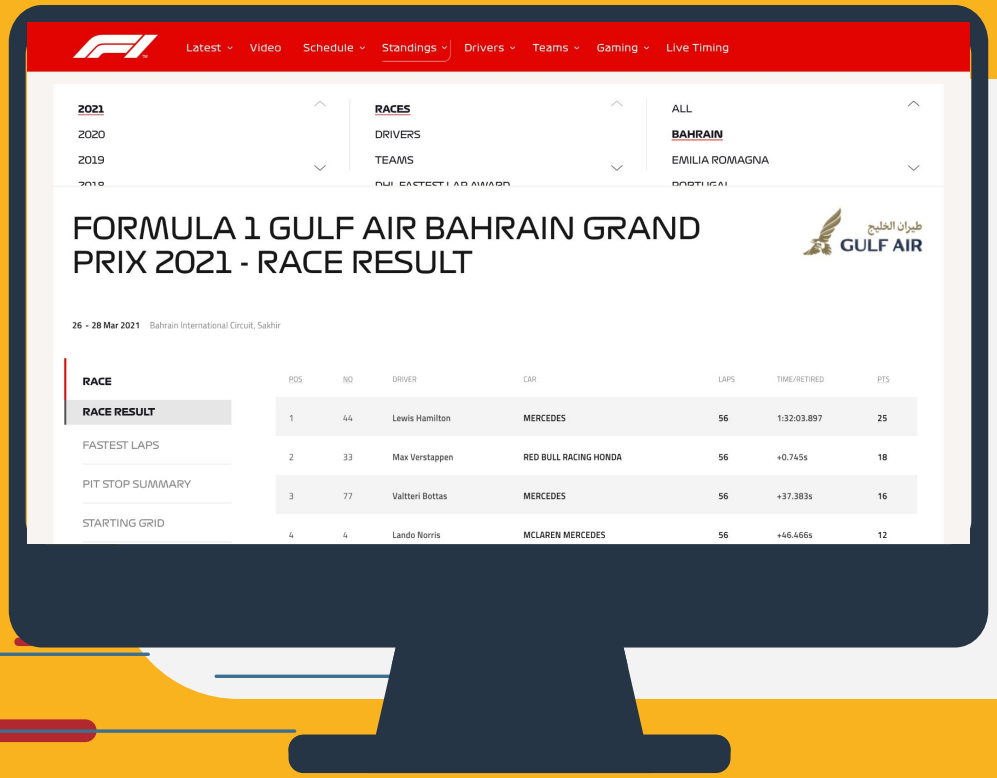
Goal: Model race day results based on practice data and pit stop performance



02

Data

Data Web Scraping



~2500 Pages

Scraped from Forumla1.com

Data Collected

4200

Unique Race Performances

26

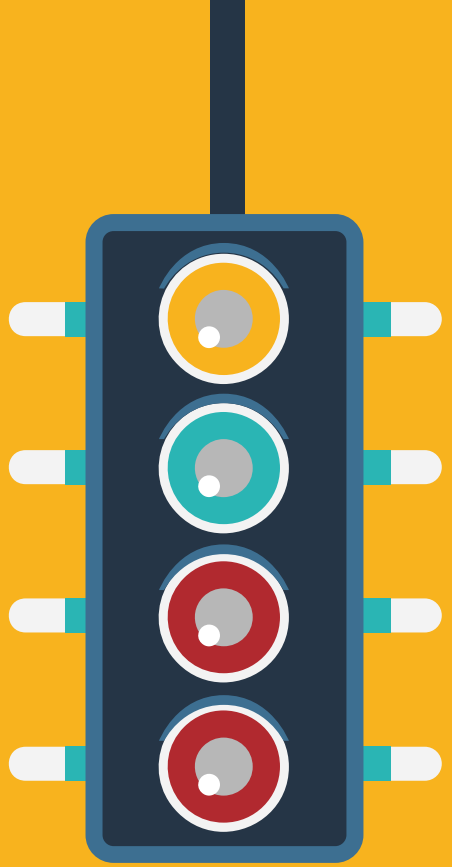
Initial Features

2006 - 2021

All races for last 15 years

03

Methodology



Pipeline

1

Web Scraping

- Selenium and BeautifulSoup

2

Data to CSV

- Saved for use in modeling notebook

3

Remove Duplicate Features

- Car driver #'s
- Time of Day Columns
- Practice Gaps

Pipeline

4

Data Cleanup

- Fix No Finish (disqualified, crash, etc)
- Position Strings to Numbers
- Strings to time
- Fill in Missing Qualifying Times

5

Feature Engineering

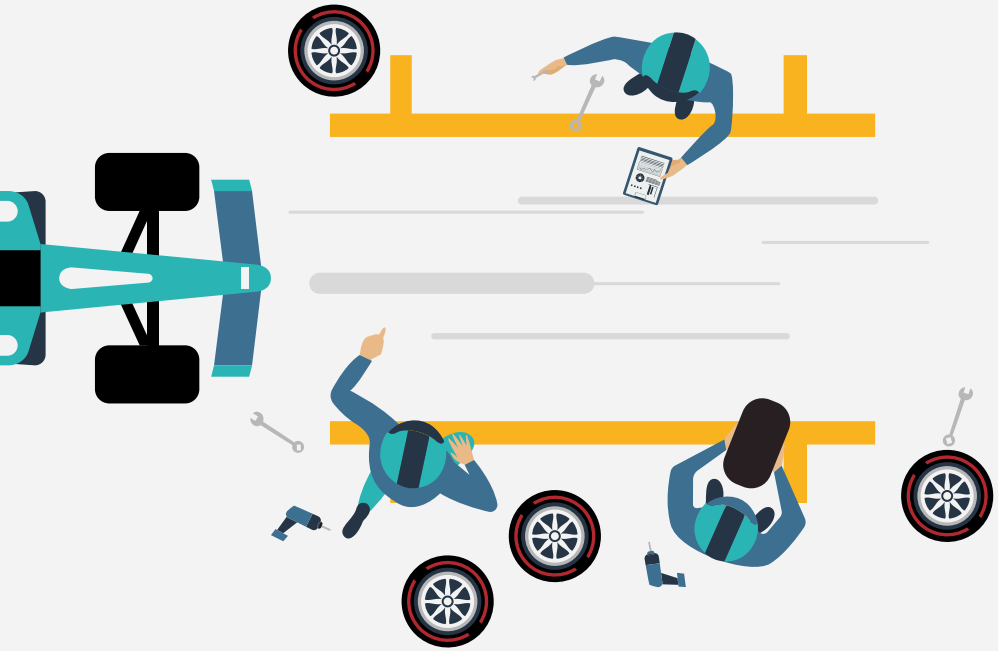
- Leverage Lasso Reg & VIF to drop bad features
- Practice time Deltas
- Interaction Features

6

Model Choice

- Chose simple linear regression
- Attempted Ridge, Lasso, 2 Deg Polynomial

Picking a Model



Simple OLS Reg

R^2 - 0.710
STD - 0.0178

LASSO

R^2 - 0.711
STD - 0.0177

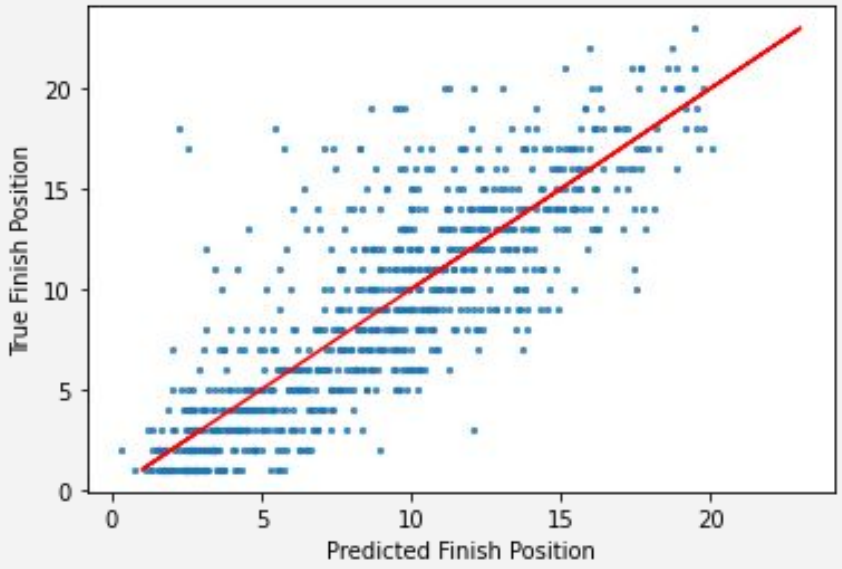
RIDGE

R^2 - 0.711
STD - 0.0174

04

Results

Model Results



0.684

R2 Scored

2.19

MAE

1. **# Pit Stops**
2. **Fastest Lap Pos**
3. **Starting Grid Pos**

3 most important features

Example Predictions

YEAR	RACE	DRIVER	PREDICTED FINISH	ACTUAL FINISH
2017	Great Britain	Lewis Hamilton	0.773	1st
2012	Malaysia	Pedro de la Rosa	19.49	22nd
2011	Japan	Paul di Resta	12.4	12th

Conclusions

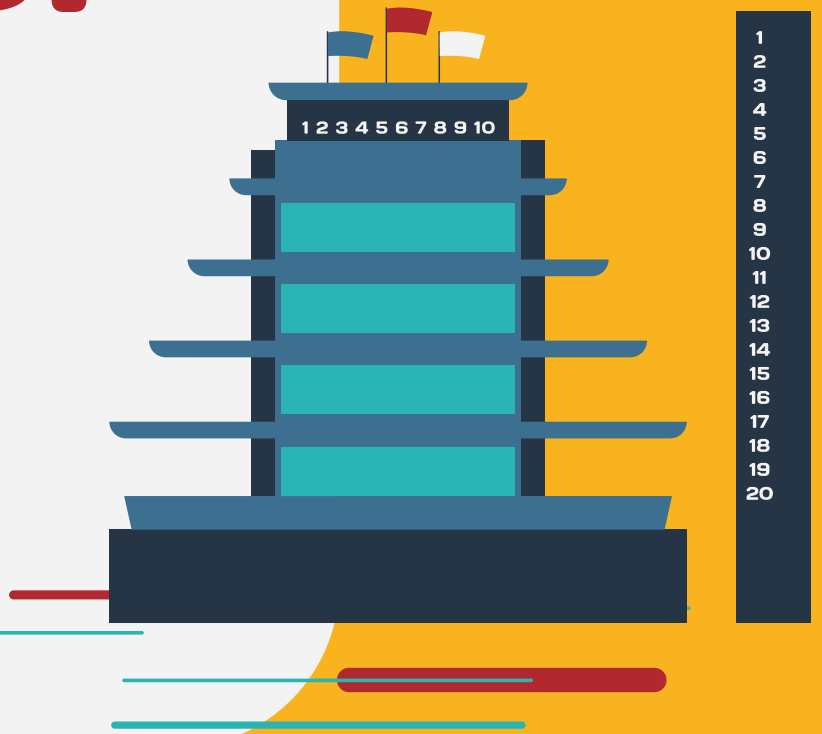
- Teams to focus on reducing # of pit stops, maximize speed through qualifying
- Future work
  - Incorporate more features and data
  - Attempt Poisson Regression and Tree based models

THANKS!

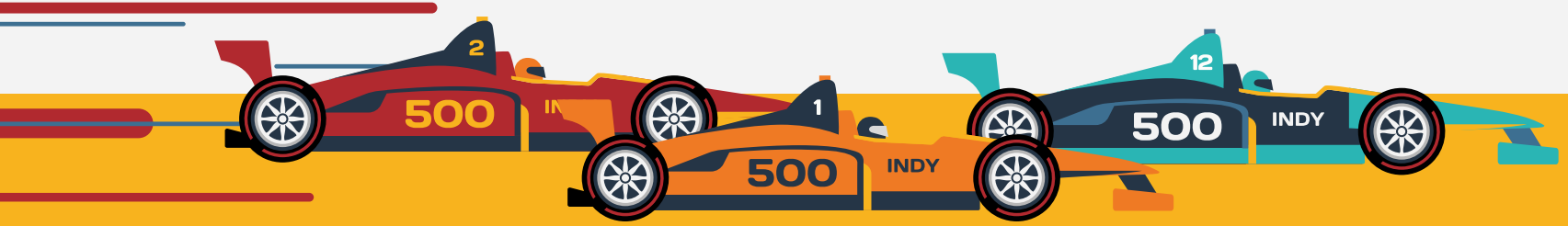
Do you have any questions?

https://github.com/braden-taack/Metis_Regression_Project

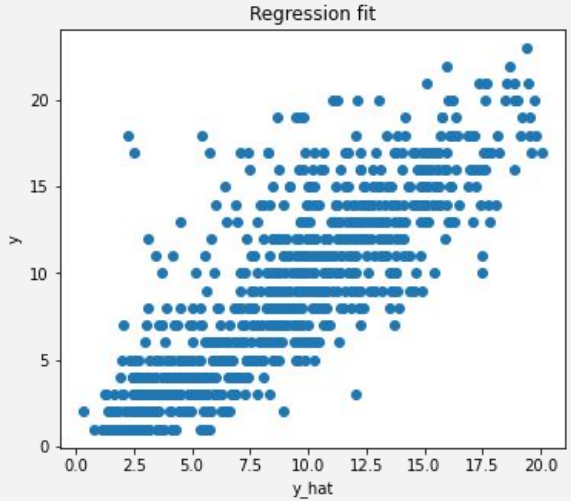
CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**

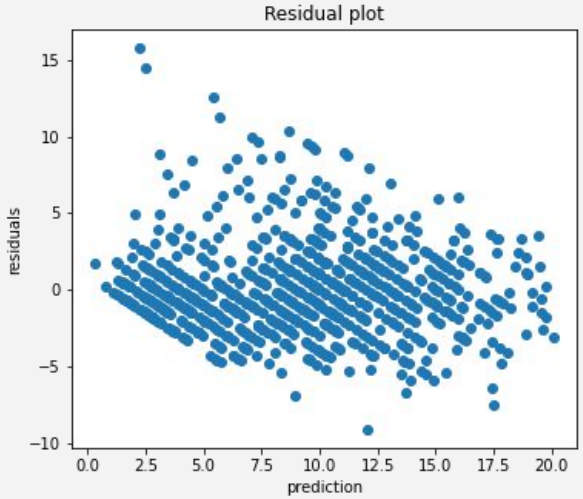


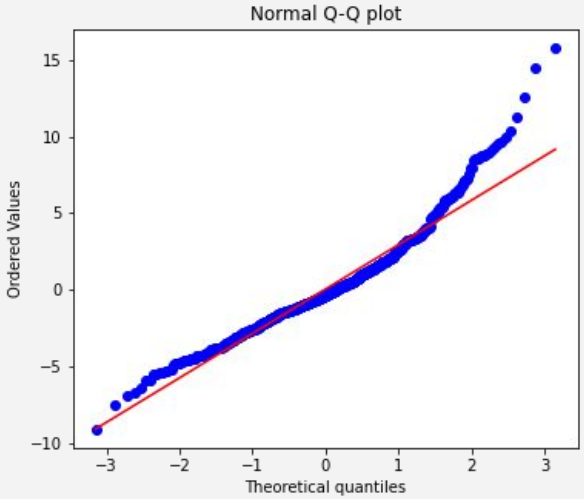
Appendix



Appendix


Regression fit
y
y_hat
0
5
10
15
20
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Residual plot
residuals
prediction
−10
−5
0
5
10
15
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Normal Q-Q plot
Ordered Values
Theoretical quantiles
−10
−5
0
5
10
15
−3
−2
−1
0
1
2
3