CS 1090A: Predicting F1 Driver Lap Times

Henry Huang, Hannah Kim, Cristiana Murgoci, Ethan Wang

01

Problem Statement

Problem Statement

Can a Formula 1 driver's **average race lap time** be accurately predicted within a given regulatory era using qualifying pace, constructor, circuit, and contextual race factors such as driver age and weather conditions? This research question aims to quantify **how pre-race indicators explain in-race performance** and assess the extent to which car engineering (constructor) versus external variables (weather, track) drive lap-time variance.

02

EDA + Analysis

Data Summary

F1 Kaggle Dataset

- Race data containing information about races, drivers, constructors, circuits, lap times, and qualifying sessions from 1950 onwards.
- 13 Subdatasets

F1 Weather Data

- 3 Features
- Added after identifying factors not addressed in the primary dataset

F1 Dataset

F1 Kaggle Dataset

- Circuits
- Constructors
- Teams
- Drivers
- Lap Times
- Results
- Qualifiers
- Races
- Pitstops
- Sprints
- Seasons
- Status Codes

F1 Weather Data

- 3 Features
- Added after identifying factors not addressed in the primary dataset

F1 Dataset

F1 Kaggle Dataset

- Circuits
- Constructors
- Teams
- Drivers
- Lap Times
- Results
- Qualifiers
- Races
- Pitstops
- Sprints
- Seasons
- Status Codes

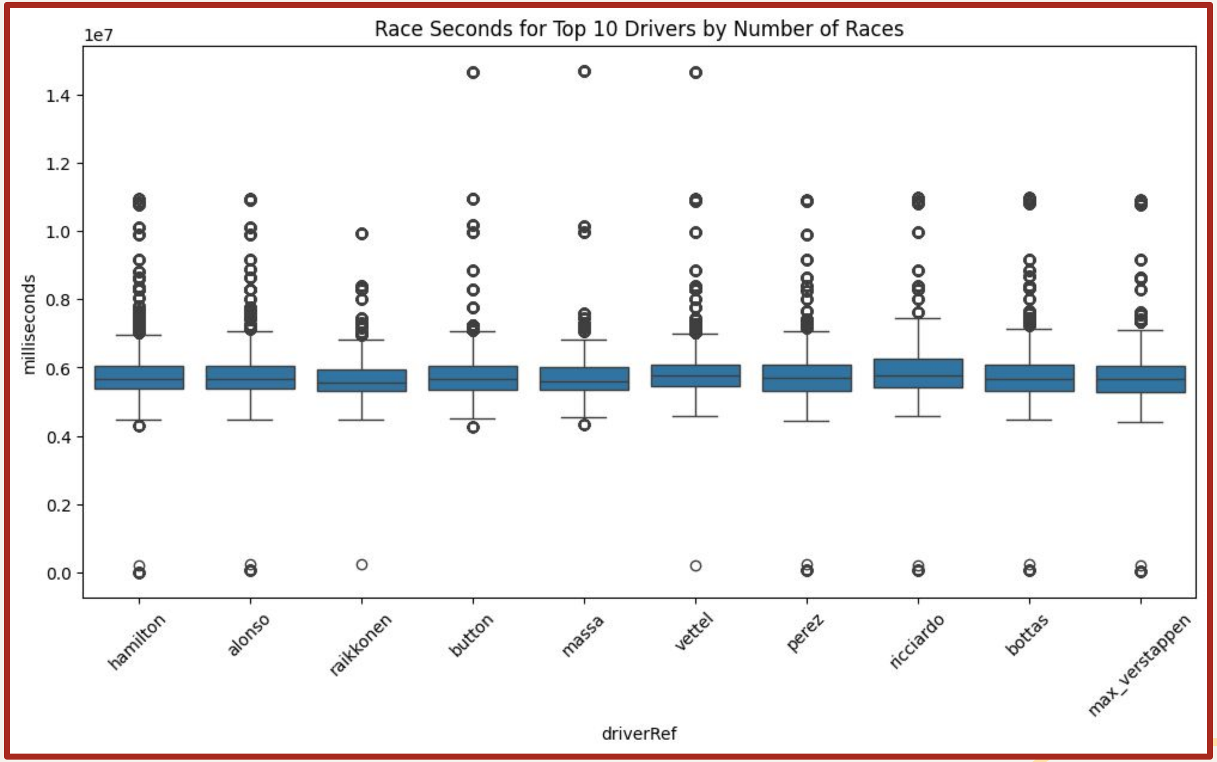
F1 Weather Data

- Temperature
- Precipitation
- Windspeed

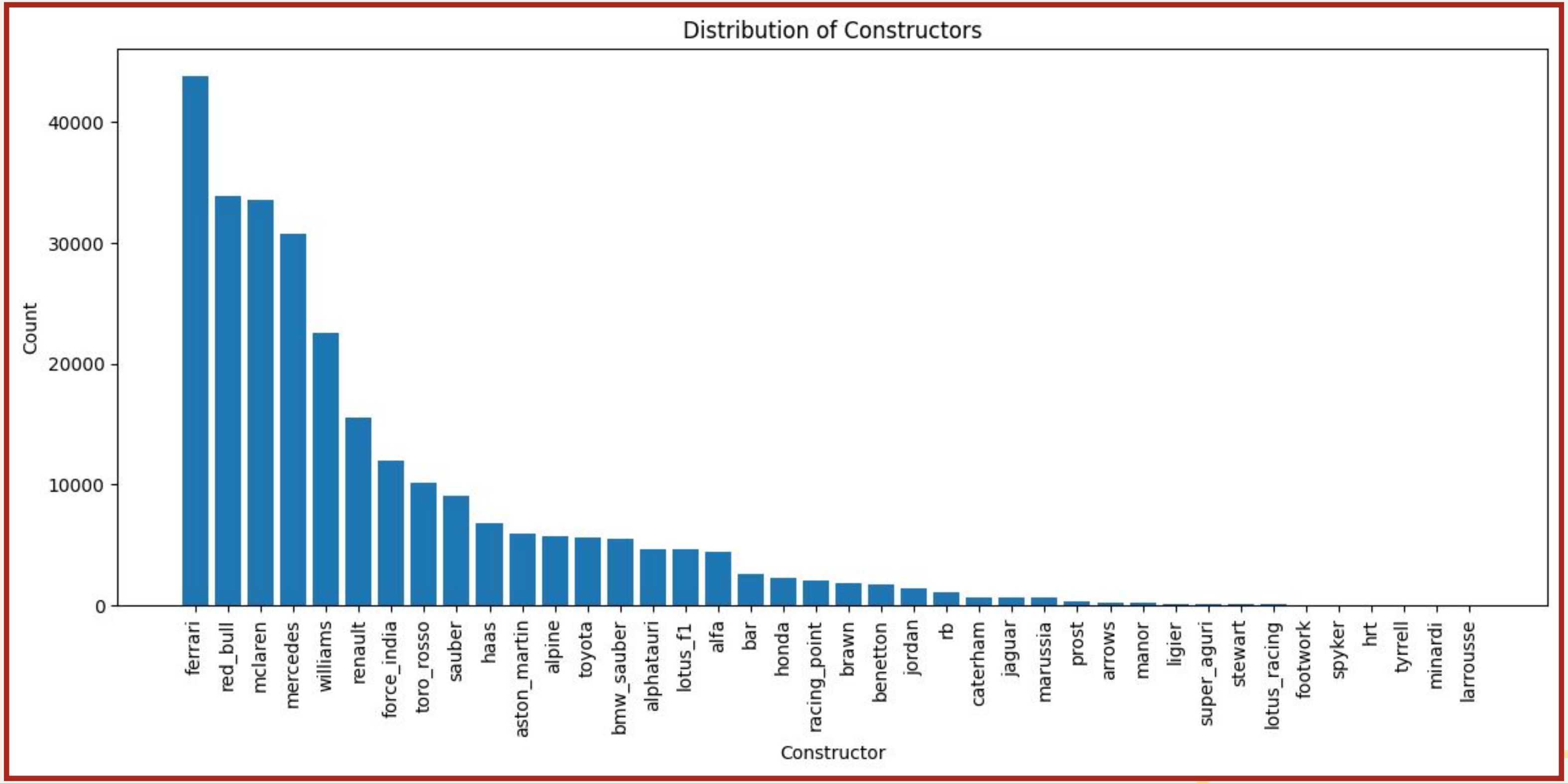
F1 Dataset

- F1 Kaggle Dataset
  - Circuits
  - Constructors
  - Teams
  - Drivers
  - Lap Times
  - Results
  - Qualifiers
  - Races
  - Pitstops
  - Sprints
  - Seasons
  - Status Codes
- F1 Weather Data
  - Temperature
  - Precipitation
  - Windspeed

Data Summary



Data Summary



Data Summary

Final Dataset

Temperature	Precipitation	Windspeed
Circuits	Constructors	Teams
Drivers	Lap Times	Results
Qualifiers	Races	Pitstops
Sprints	Seasons	Status Codes

- ~270k observations
- 29 Features
- Merged on race and driver id.
- Pre-Processing:
  - Standardization
  - One-hot encoding
  - Best session substitution for missing data

03

Modeling

Basic Model: Regression

One-hot Encoded Categorical Features

↓

$$\widehat{\text{lap\_time\_sec}}_i = \beta_0 + \beta_1\,\text{grid}_i + \beta_2\,\text{laps}_i + \beta_3\,\text{best\_quali\_sec}_i + \beta_4\,\text{age}_i + \beta_5\,\text{windspeed}_i + \beta_6\,\text{precipitation}_i + \beta_7\,\text{temperature}_i$$
$$+ \sum_c \gamma_c\,\mathbf{1}\{\text{constructorRef}_i = c\} + \sum_r \delta_r\,\mathbf{1}\{\text{circuitRef}_i = r\}$$
$$+ \sum_d \eta_d\,\mathbf{1}\{\text{driverRef}_i = d\} + \sum_n \theta_n\,\mathbf{1}\{\text{nationality}_i = n\} + \varepsilon_i$$

Advanced Regression

One-hot Encoded Categorical Features + Standardized Numeric Features

$\downarrow$ Cross Validation with L2 Regularization

$$\widehat{\text{lap\_time\_sec}}_i = \beta_0 + \sum_{k=1}^{m} \left[\beta_{1,k}\,\text{grid}_i^k + \beta_{2,k}\,\text{laps}_i^k + \beta_{3,k}\,\text{best\_quali\_sec}_i^k + \beta_{4,k}\,\text{age}_i^k + \beta_{5,k}\,\text{windspeed}_i^k + \beta_{6,k}\,\text{precipitation}_i^k + \beta_{7,k}\,\text{temperature}_i^k\right]$$
$$+ \sum_c \gamma_c\,\mathbf{1}\{\text{constructorRef}_i = c\} + \sum_r \delta_r\,\mathbf{1}\{\text{circuitRef}_i = r\}$$
$$+ \sum_d \eta_d\,\mathbf{1}\{\text{driverRef}_i = d\} + \sum_n \theta_n\,\mathbf{1}\{\text{nationality}_i = n\} +$$
$$+ \sum_{i,j \in P} \zeta_{ij} x_{ij} + \varepsilon_i$$

Final Model: RandomForestRegressor

01

n_estimators

Swept from 1 to 200, with step_size = 1 and patience = 3 (final n = 3)

02

max_depth

Set to None!

03

min_samples_split

Set to 5

04

random_state

Set to 42

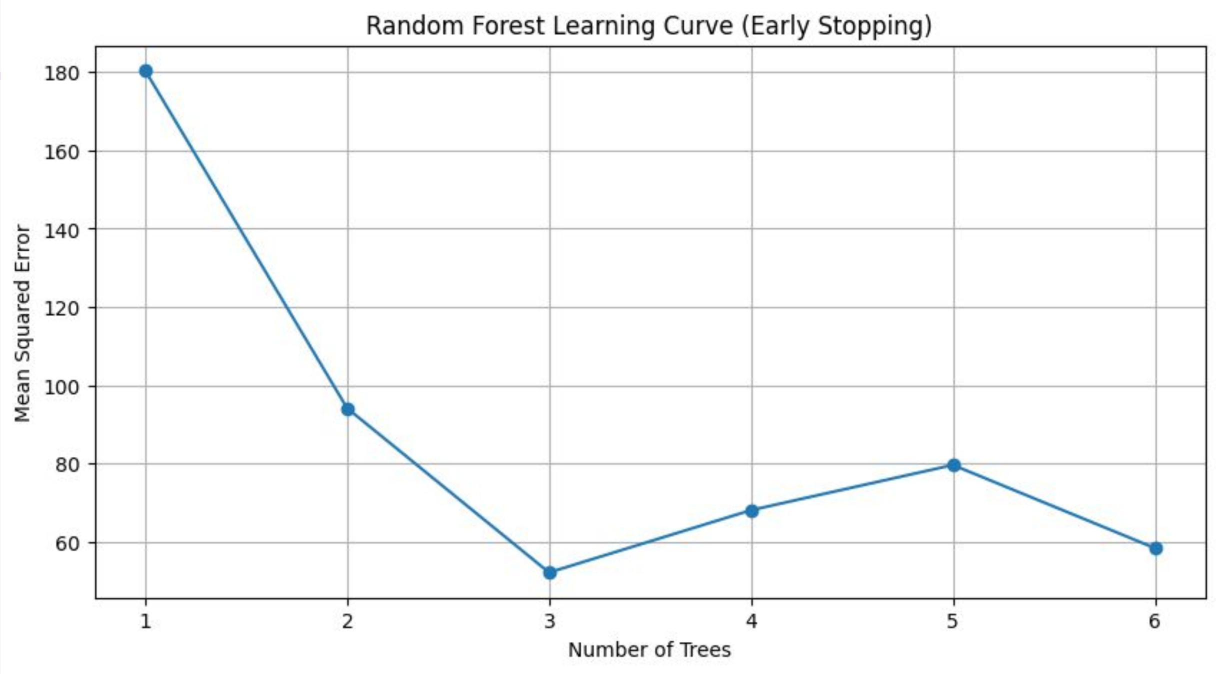
04

Results

Random Forest Outperforms Regularized Linear Regression

Metric	Regularized Linear Regression	Random Forest
MSE	234 sec^2	52 sec^2
RMSE	~ 15 sec	~ 7 sec
R^2	0.66	0.92

Tuning the Random Forest



Final Random Forest Model:

- 3 trees
- R^2 = 0.92
- MSE = 52

Feature Importance

feature	importance	score
1	laps	0.54
2	best_quali_sec	0.17
3	windspeed	0.07
4	temperature	0.06
5	precipitation	0.03

05

Conclusion

Final Pipeline

Data Work

- EDA
- Data Cleaning
- Feature Selection

Modeling

- Baseline linear regression
- Optimized linear regression
- Random Forest

Answer to Question

- Most important factor was the course and driver
- External factors like weather were next, but not nearly as important

06

Future Work

Next Steps

<u>Data Improvements</u>

- Feature Engineering
  - Weight data on recency
- More data
  - Granular weather data (e.g. track temp)

<u>Model Improvements</u>

- Gradient boosting
- Models/methods past 1090a
  - SVM, neural networks, etc.

Thanks!

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