FINTECH BOOTCAMP

Formula One Race Prediction

BY AARON SIMPSON,PAARAS DHALIWAL, CRAIG BRAGANZA, GAETANO BUONGIORNO , LUKE EVANS

The Concept

Optimal F1 Race Prediction Model

1

The Idea

We realised that Formula One was an appropriate sport where machine learning models could be used as there were many objective measures of performance.

There was an Average of 20 Grand Prix per F1 season with data available to us from 1950.

2

Motivation for Development

To formulate effective machine learning models to predict race outcome allowing for better race betting outcomes.

3

User Story

Lorenzo was watching the F1 in Melbourne and being an innovative data scientist he thought to himself.

"Could I predict the outcome of the race and beat the system?" So he created a model with a team to see if he could.

Data Techniques

- How can machine learning be used to accurately predict the results of the formula one races during the season.
- Evaluated the predicted vs actual race outcomes to find the optimal model
- We can apply this model to gain an edge when betting on the outcome of the races

Data Techniques

Data Sources

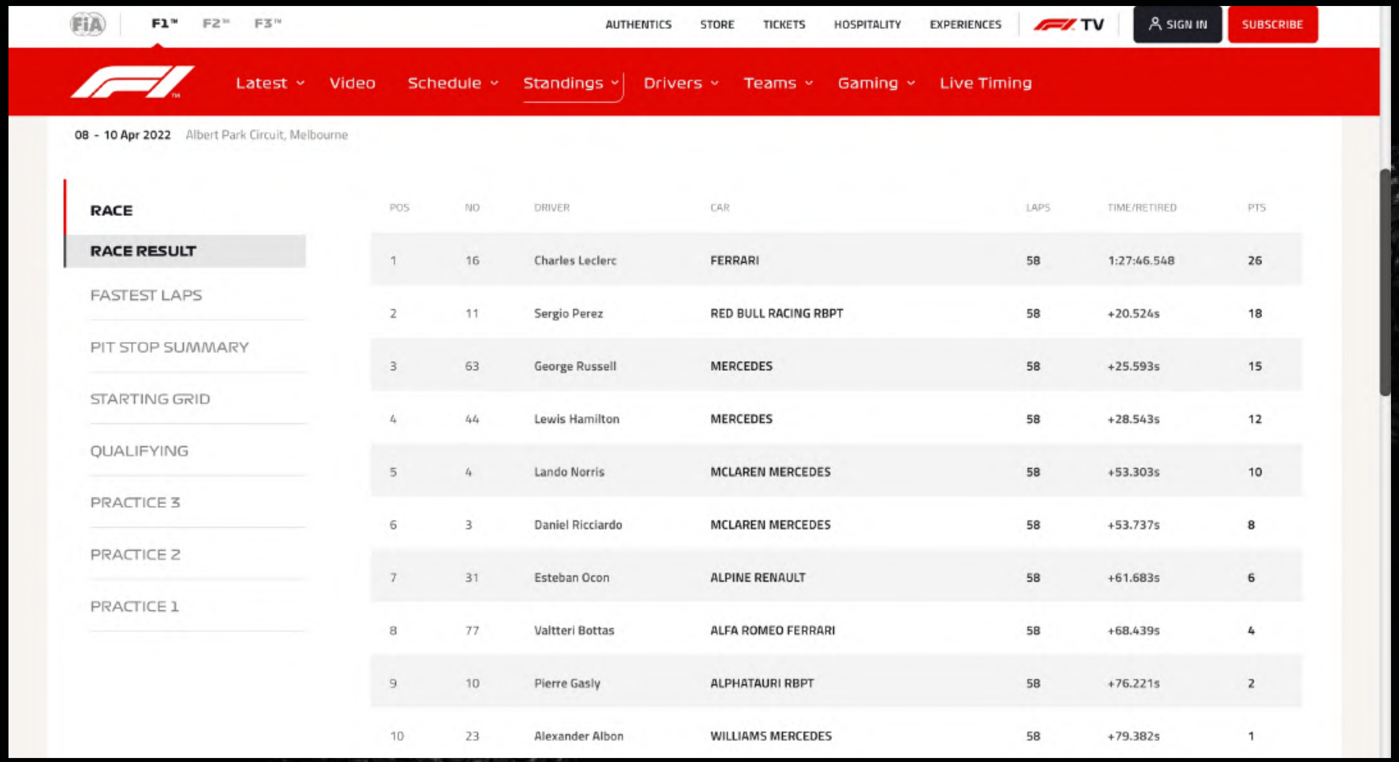
We sourced an Open Source Formula One 'Ergast' API as well as web-scraping from the official formula.com website.

Data Selection

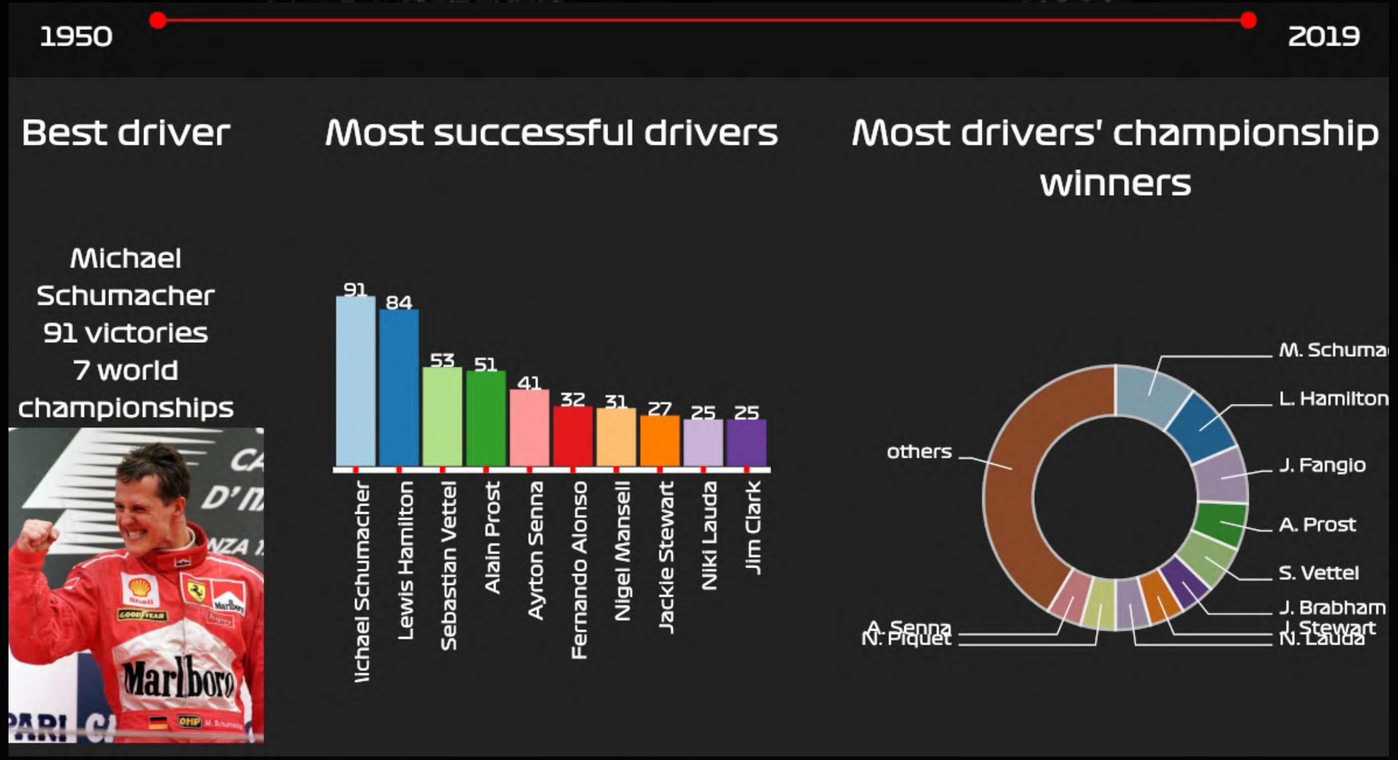
We selected all of the data sourced besides using constructor standings as this not affect the accuracy of the model for betting purposes.

Collection, Exploration and Cleaning Process

The data was cleaned and explored using different methods such as:



POS	NO	DRIVER	CAR	LAPS	TIME/RETIRED	PTS
1	16	Charles Leclerc	FERRARI	58	1:27:46.548	26
2	11	Sergio Perez	RED BULL RACING RBPT	58	+20.524s	18
3	63	George Russell	MERCEDES	58	+25.593s	15
4	44	Lewis Hamilton	MERCEDES	58	+28.543s	12
5	4	Lando Norris	MCLAREN MERCEDES	58	+53.303s	10
6	3	Daniel Ricciardo	MCLAREN MERCEDES	58	+53.737s	8
7	31	Esteban Ocon	ALPINE RENAULT	58	+61.683s	6
8	77	Valtteri Bottas	ALFA ROMEO FERRARI	58	+68.439s	4
9	10	Pierre Gasly	ALPHATAURI RBPT	58	+76.221s	2
10	23	Alexander Albon	WILLIAMS MERCEDES	58	+79.382s	1



Approach

Technologies Used

- Logistic Regression
- SGD
- KNeighbours
- Decision Tree
- Random Forest

Breakdown of Tasks and Roles

- Ideation
- Data Sourcing
- Data Cleaning
- Model Formation
- Presentation
- Worked Collabratively on all tasks

Technical Challenges

- The Ideation phase took longer than expected as we realised many sports we wanted to use machine learning methods with had insufficient data
- Data cleaning took longer than the time allocated with models having limitations on the amount of data types used

Successes

- We successfully predicted the outcome of the race by upto 89% with a Random Forest Classifier

FORMULA ONE PREDICTION

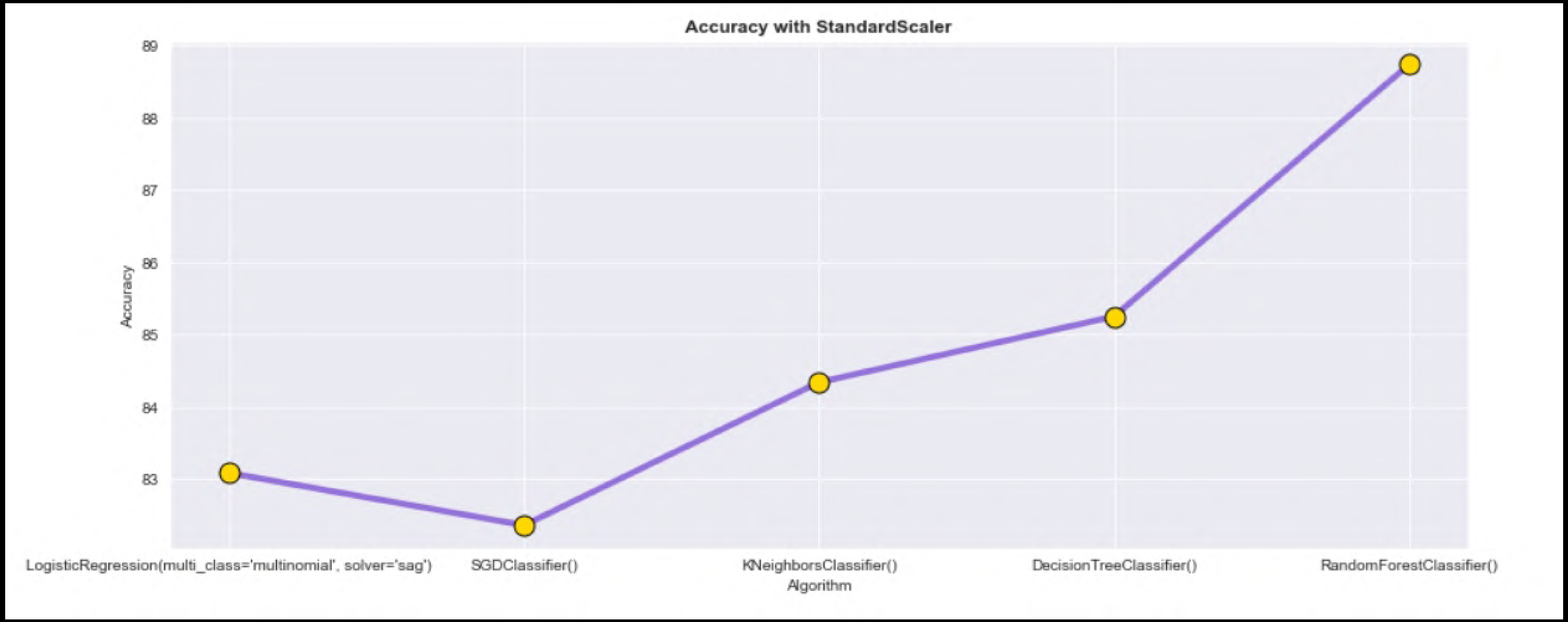
DEMO

GITHUB

NOTEBOOK VIEWER

Link

Model Comparison



Pitch

Challenges Faced

Ideation

+ Ideation phase took longer than anticipated

+ Lead to limited amount of time to working on fully comprehensive models.

Data Sourcing

+ Data sourcing took longer than expected

+ Technical learning curve for the use of potential APIs.

Model Drawbacks

+ Models has restriction on number of different data types

+Models could be underfit or overfit

Other Challenges

+ Micro-issues relating to the technical requirements i.e. installing packages etc.

+ Team co-ordination could of been improved by using project management tools such as Jira.

Conclusion and Next Steps

#1

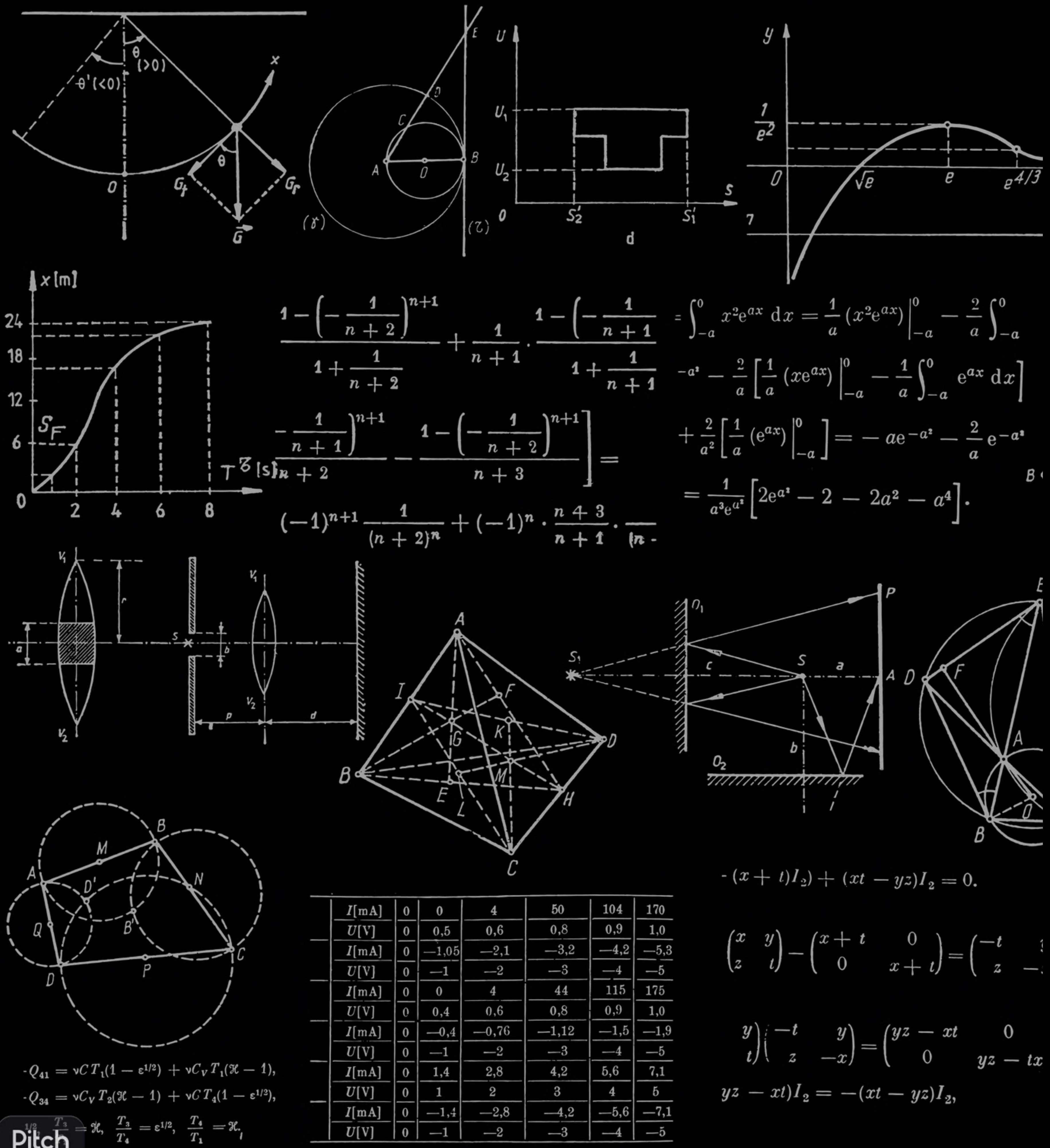
Conclusion

We can conclude that using machine learning models allows us to predict the outcome of a F1 race by up to 89%.

#2

Next Steps

- More Data: sets for some of the possible other influences such as wet weather, temperature as well as things like pit stops and other human error related matters.
- Refine our technical skillset
- Use Project management tools
- In the future we would combine multiple models to have a fully comprehensive one



Questions?