PREDICTING FORMULA ONE RACE RESULTS

Joel Calm

Dec 2024, Aprenentatge Computacional

F1 Introduction

- Season Format:
  - 20+ rounds per season
  - 20 drivers
  - 10 teams
- Grand Prix Format:
  - 3 Practice sessions
  - 1 Qualifying session
  - 1 Race
- Race Format:
  - 50-75 laps
  - Top 10 score points

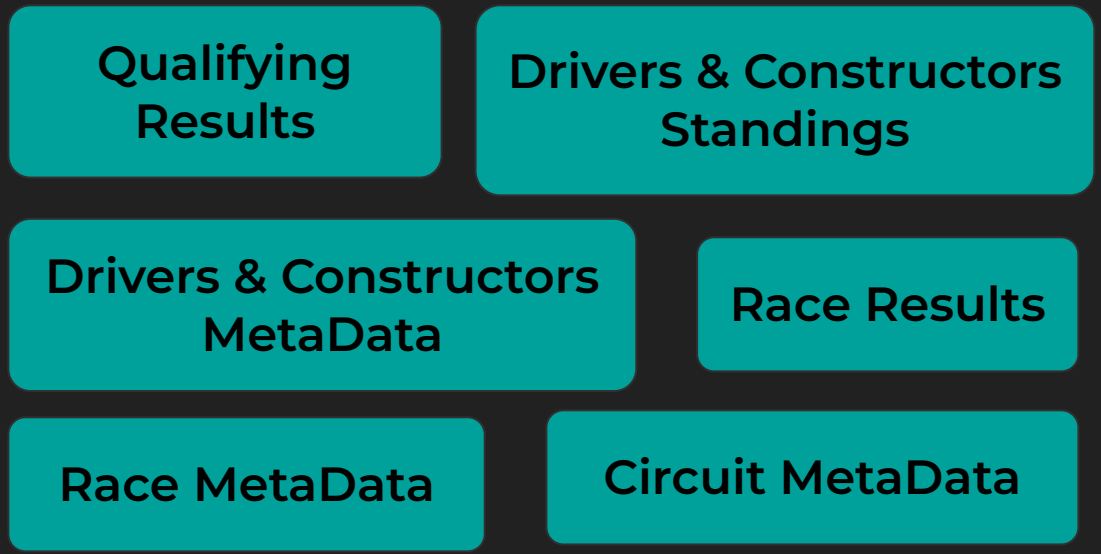
Project Overview

- Goals: Predict race winners and group positions (podium, points, no-points) for 2020-2024 seasons.
- Challenges: Imbalance (1 winner per race)
- Classification or regression task?

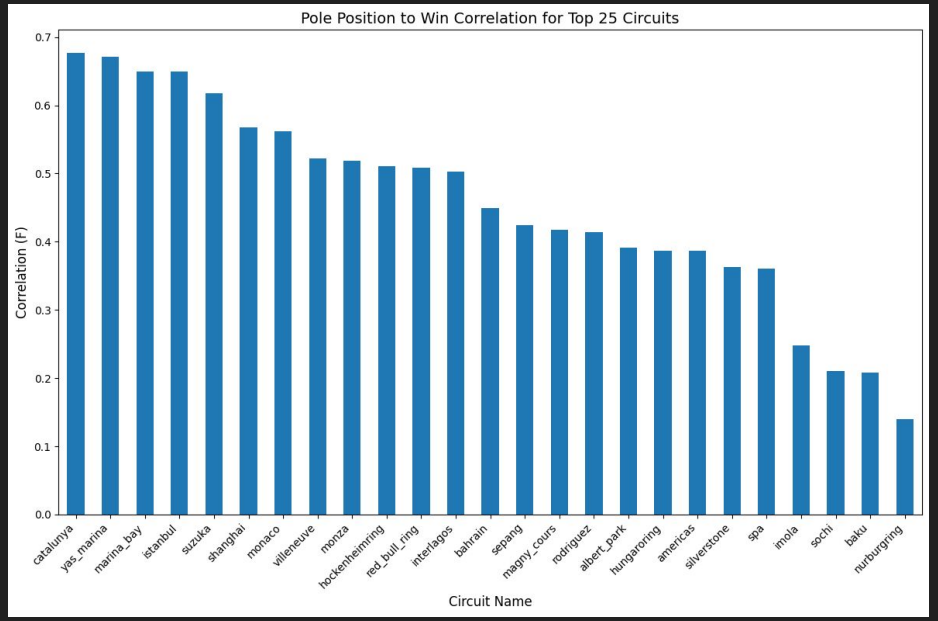
Data Collection

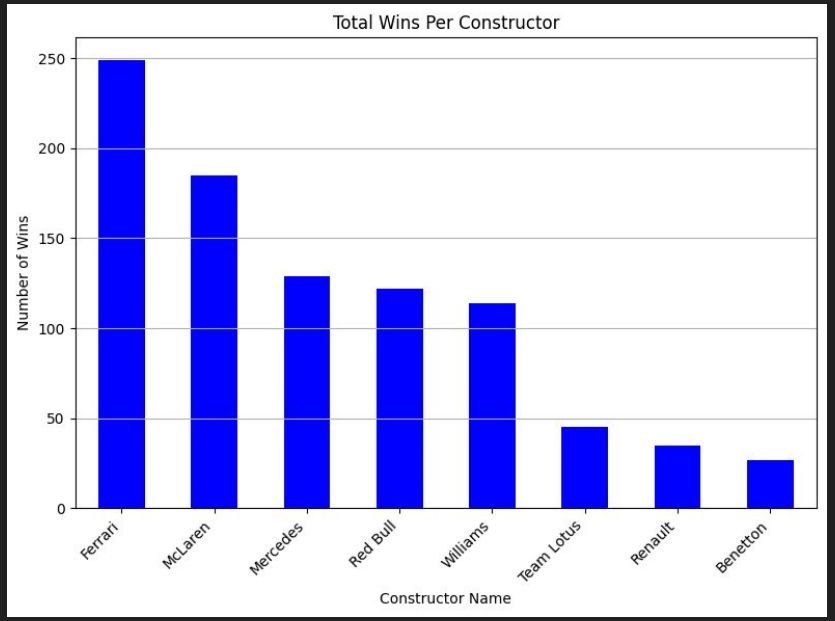
Kaggle Datasets:
(1950-mid 2024)

Ergast API

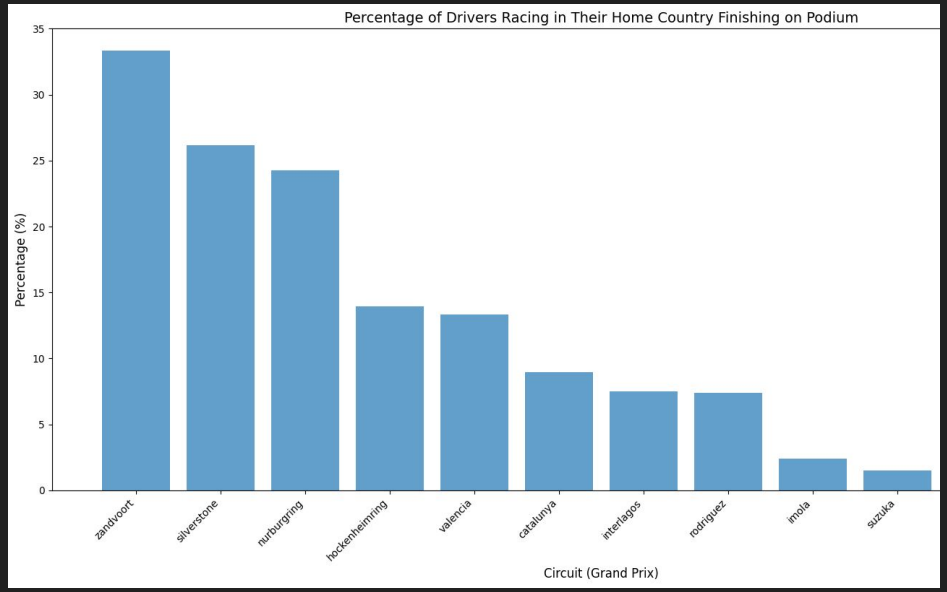


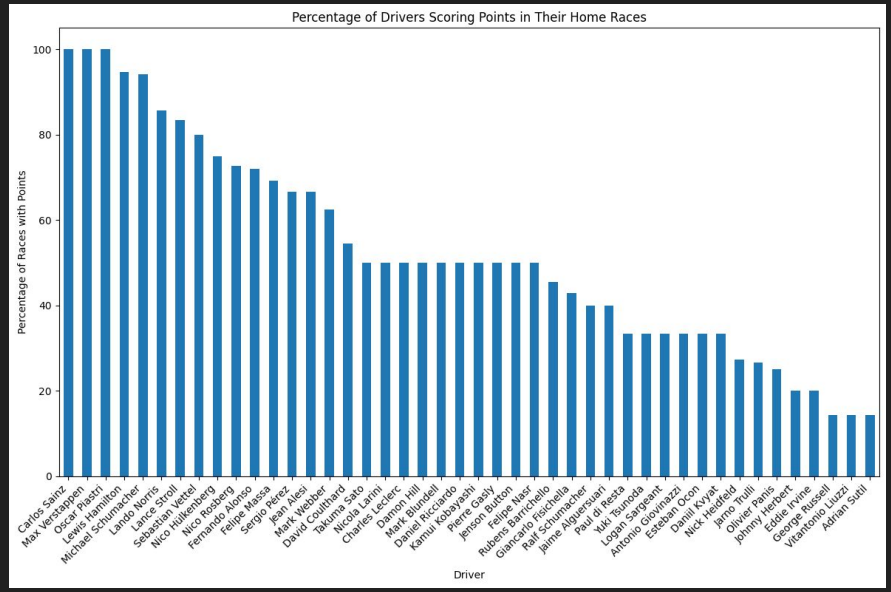
Exploratory Data Analysis





Exploratory Data Analysis





Data Preprocessing

Merging Data: Combined datasets into a single one

Removed unnecessary columns: URL's, Id's, number, times..

Added new columns: driver_name, prev_points, prev_wins...

Encoding: Applied one-hot encoding for categorical data

Scaling: Standardized numerical features for uniform scaling

10,242 rows and 24 features.

Metric Selection

- 1 winner per race
- Race-wise accuracy
- Group accuracy

Target?

Driver	Probability (0)	Probability (1)	Actual	Predicted
V. Bottas	0.6926	0.3074	1	1
M. Verstappen	0.8482	0.1518	0	0
L. Norris	0.8618	0.1382	0	0
A. Albon	0.9241	0.0759	0	0
L. Hamilton	0.9932	0.0068	0	0

Machine Learning Models

Models Used:

- Classification: Logistic Regression, Random Forest, SVM i XGboost
- Regression: Linear Regression, RF Regressor, SVM Regressor i XBG Regressor

Cross Validation

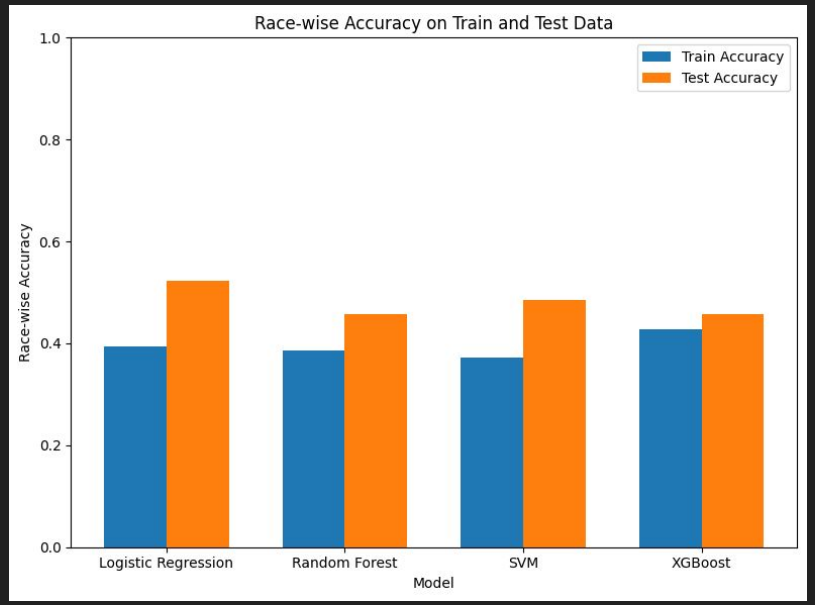
- TimeSeriesSplit

Imbalance:

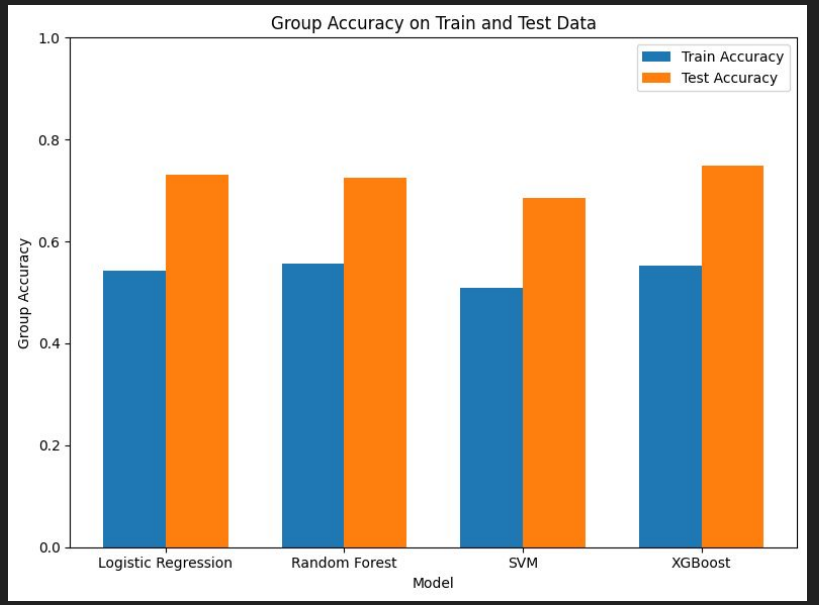
- Class weighting
- Metrics

Baseline Model

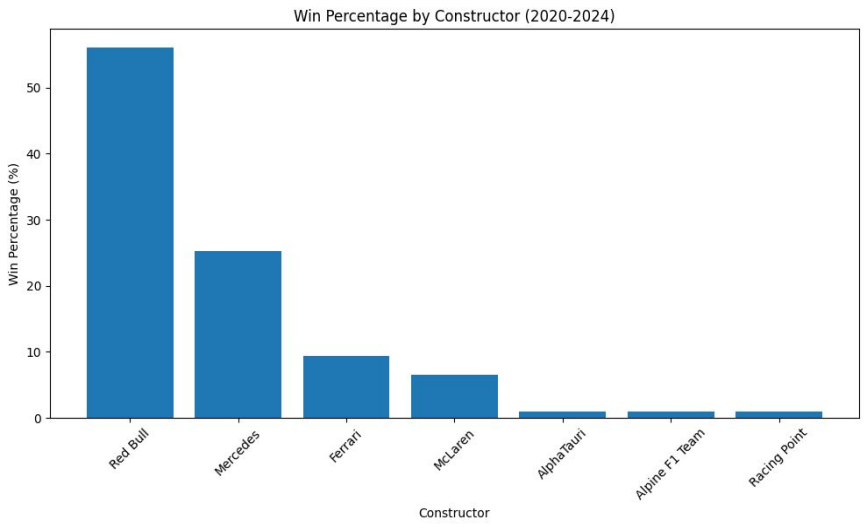
winner

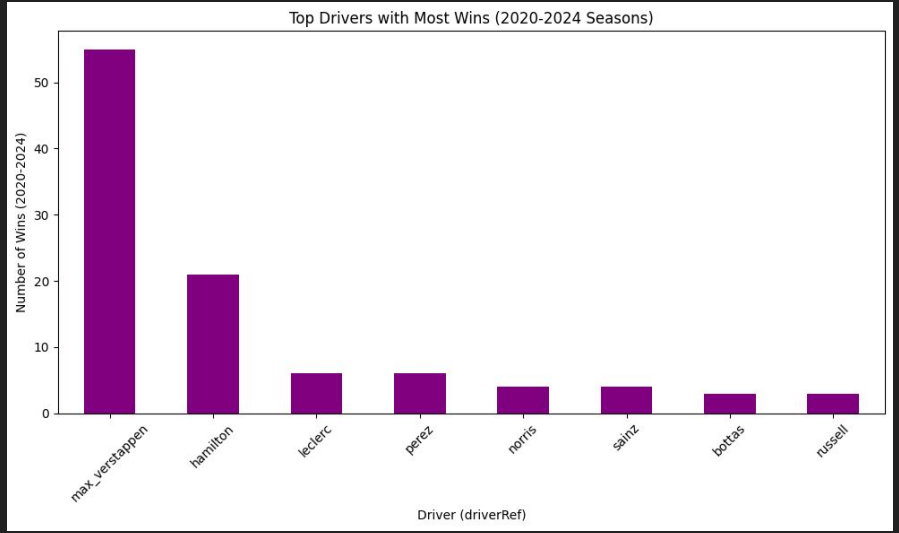


group position



Train vs Test

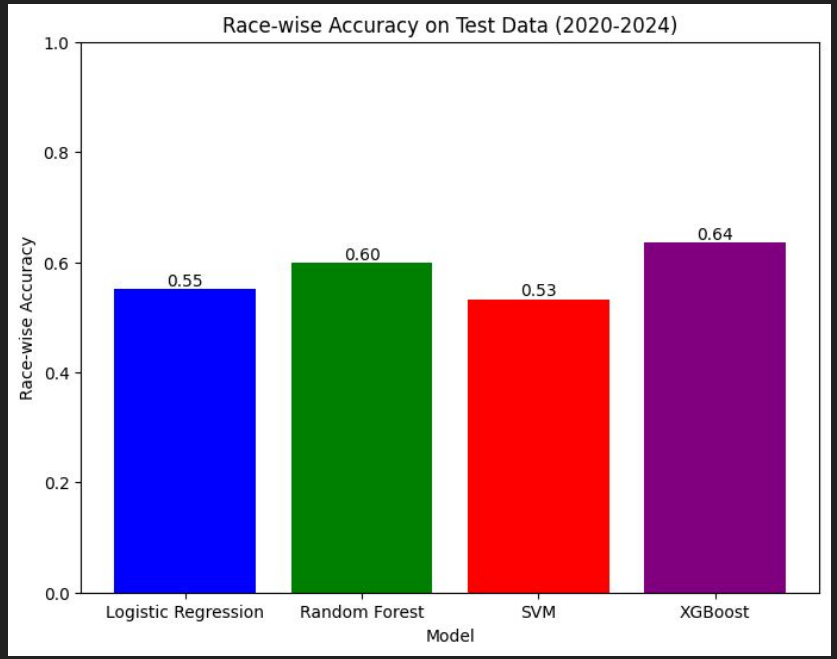




Feature Engineering

- Driver Reliability
- Constructor Reliability
- Average Performance
- Home Advantage

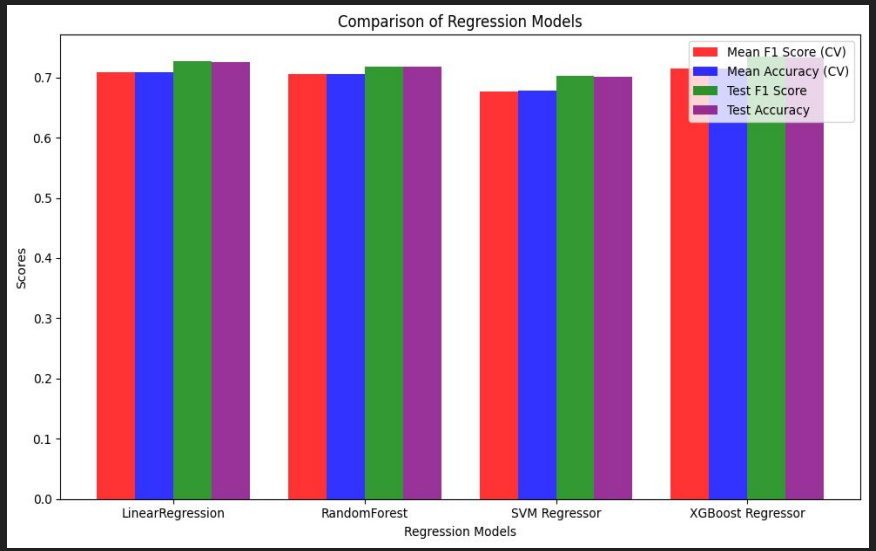
winner

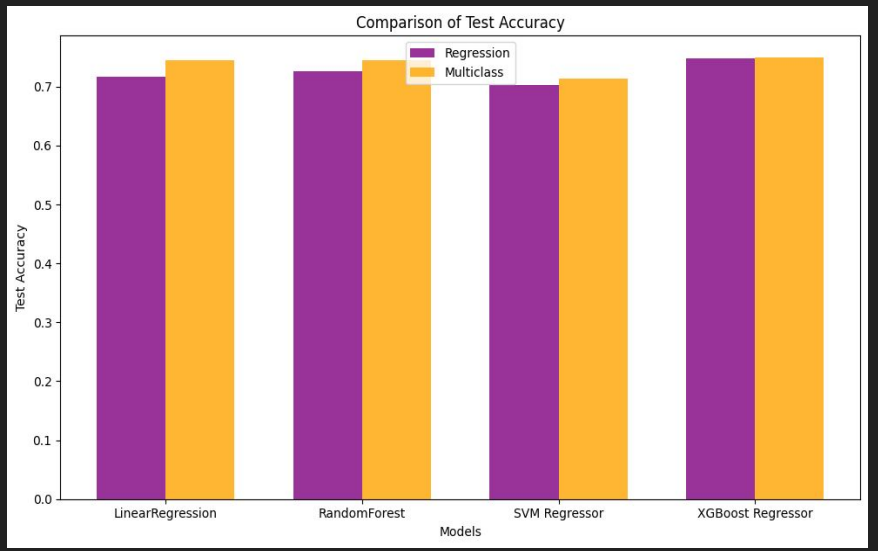


Feature Selection & Hyperparameter Tuning

RFE, SelectKBest (feature importance)
learning rate, n estimators, max depth..

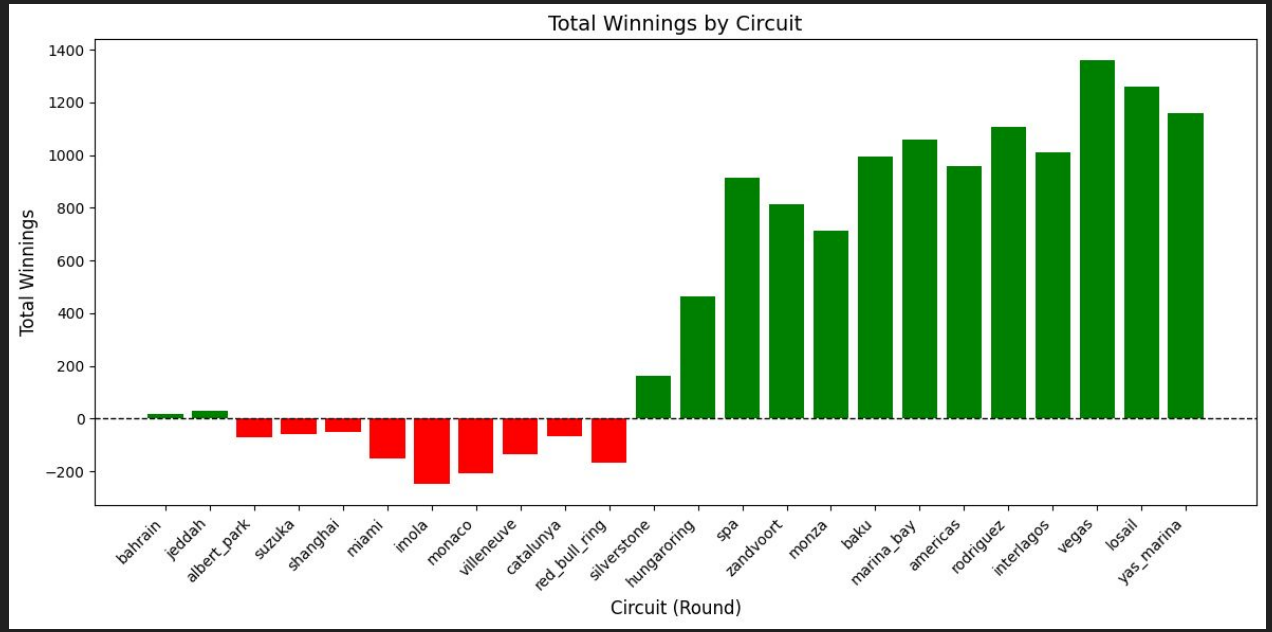
Multiclass vs Regression





Betting Analysis 2024 (post-qualifying)

ROI: 35.79%



Conclusions

- Xgboost
- Future Work

Year	Race Accuracy
2020	0.647059
2021	0.454545
2022	0.590909
2023	0.909091
2024	0.583333
Average	**0.636988**