F1 Race Predictor

ENRIQUE GONZALEZ

ALEXIS MARTINEZ

Project Overview

- Predict the finishing positions of F1 drivers using machine learning.
- Comprehensive F1 race data (4,626 races, 70 features)
- Multiple regression models with some custom evaluation metrics
- Accurate prediction of race positions to within 1-3 positions

What is F1 and why it's hard to predict.

- Formula 1 is the highest class of international single seater auto racing, including:
  - 20 of the world's best drivers
  - 10 teams (constructors) competing across 24 global races
  - High performance cars reaching speeds over 215 mph
  - Why F1 is hard to predict?
    - Complex race dynamics, weather variability, team strategy decisions, safety cars, car performance variations, driver form fluctuations and mid-season technological developments

Data Overview

- Historical F1 race data starting from 2014 –2024 seasons
- Feature categories include:
  - Driver data(ID)
  - Constructor/team data
  - Circuit characteristics
  - Historical performance metrics
  - Race specific information

Feature Engineering

- Driver form metrics: Last 3 races average position, last 3 races average points and championship position and points
- Constructor performance: Team championship standings and recent team performance metrics
- Circuit specific performance: Driver's history at specific circuits and team history at specific circuits
- Custom features:
  - Grid to finish potential: grid / last 3 races average position
  - Recent form : last 3 races average position / championship position

Modeling Approach

- Models evalutated
  - Linear Regression
  - K-Nearest Neighbors(KNN)
  - Decision Tree
  - Random Forest
  - XGBoost
  - Support Vector Regression(SVR)

  Data Split:
  - 80% Training / 20% Testing
  - 5-fold cross validation on best model

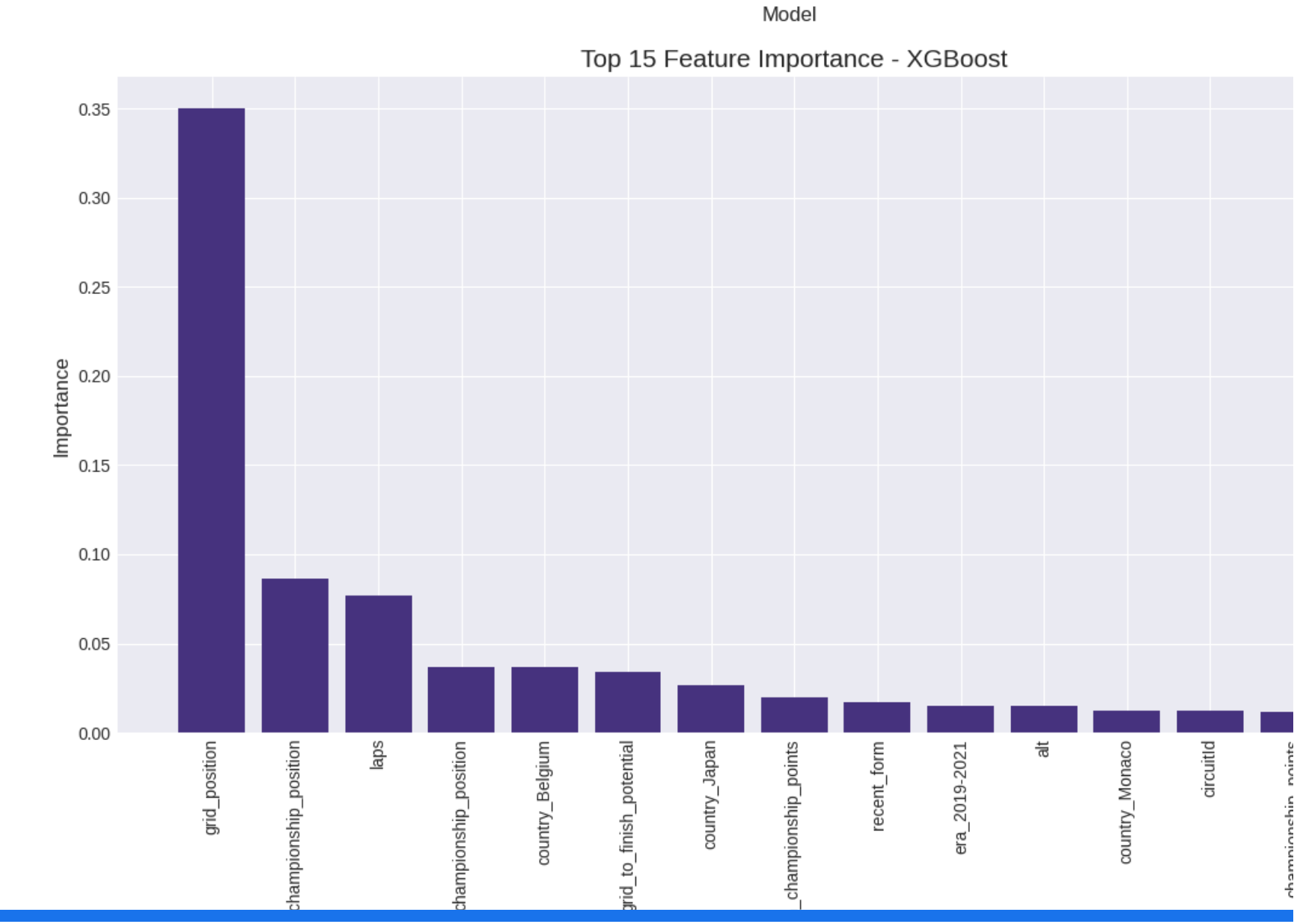
Evaluation Metrics

- R^2 Score
- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- % predictions within 1 and 3 positions

```
Model Performance Summary:
                Model        R²       MAE      RMSE  Within 1 Position
0   Linear Regression  0.692837  2.528522  3.216642           0.222462
1                 KNN  0.464711  3.319870  4.246320           0.222462
2       Decision Tree  0.471857  3.172786  4.217878           0.347732
3       Random Forest  0.750295  2.212451  2.900228           0.291577
4             XGBoost  0.793742  2.011144  2.635871           0.319654
5                 SVR  0.501338  3.221706  4.098468           0.198704

   Within 3 Positions
0            0.685745
1            0.555076
2            0.636069
3            0.750540
4            0.786177
5            0.548596

Best model based on R²: XGBoost

Performing cross-validation on
Cross-validated R2 for XGBoost: 0.7666 ± 0.0197
```

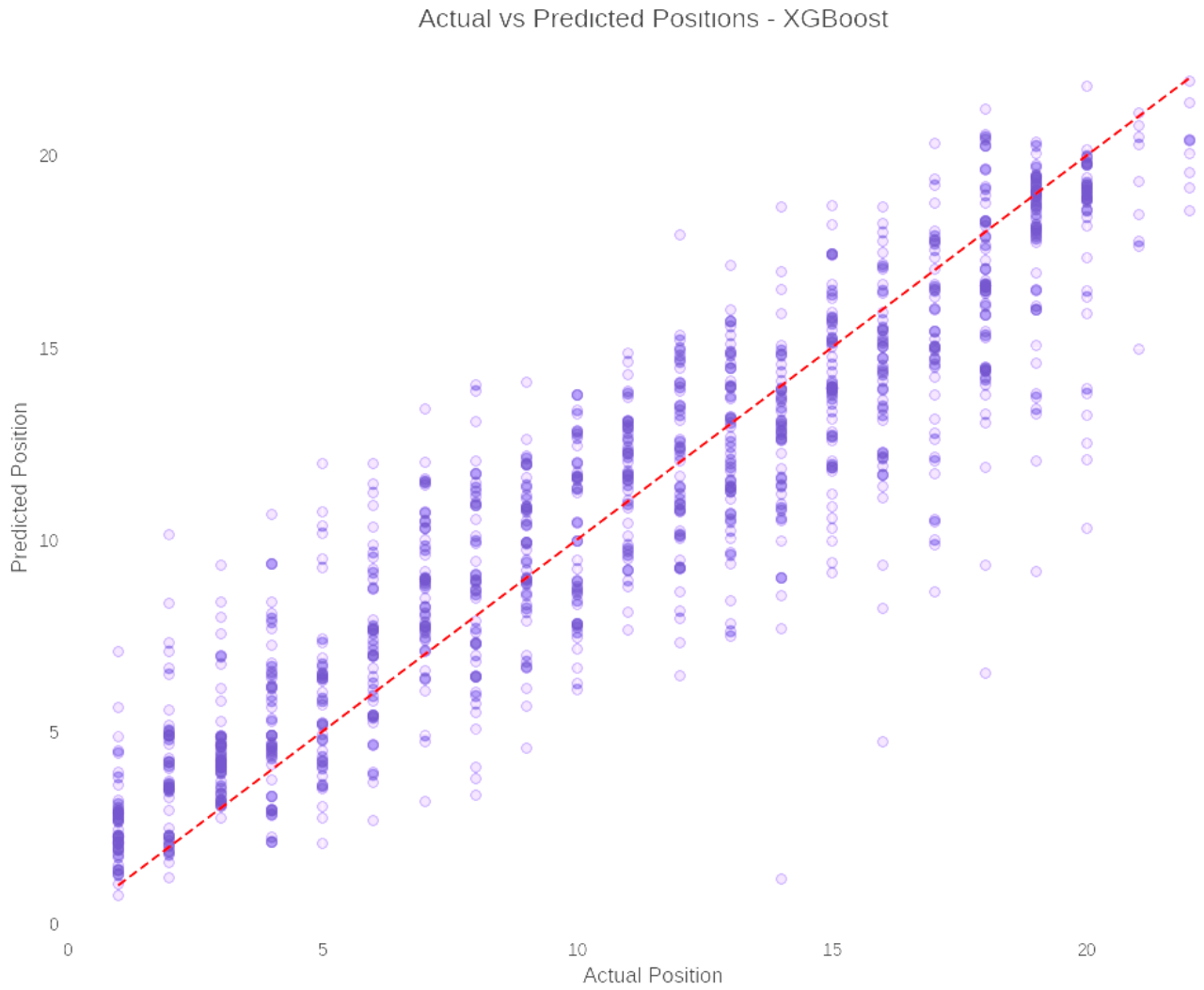
Model Performance

- XGBoost explains 79% of the variance in race finishing positions
- Can predict finishing positions within 1 position about 32% of the time and 76% within 3 positions
- MAE values around 2-3 positions suggest predictions are typically off by 2-3 grid positions.



Feature Importance

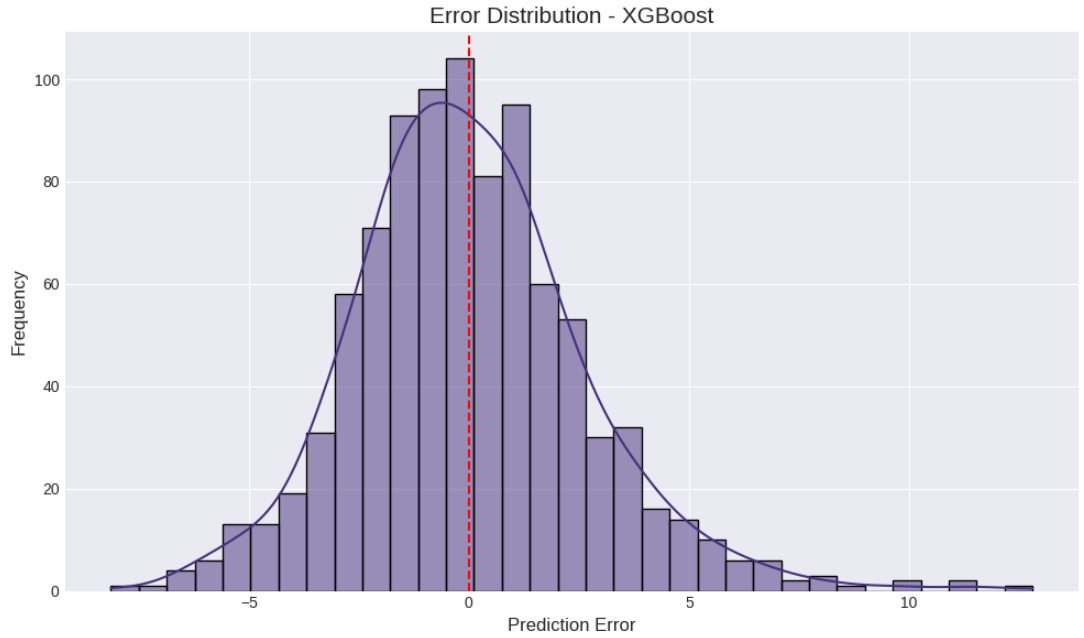
- Grid position dominates – qualifying position strongly impacts race results
- Championship position reflects a driver's consistent performance throughout the season
- Laps captures race experience and reliability



Model Predictions

- Points follow the diagonal line which confirms good R2 value.
- Front position(5) – good accuracy with tighter clustering
- Mid-field positions(6-15) – wider spread indication more variability
- Back positions – mixed performance with notable outliers

Error analysis



- Peak distribution is slightly to the left of the line, indicates that there is a tendency to predict better positions than driver achieve
- Bulk of errors fall within ±5 positions

Challenges and future improvements

- Challenges:
  - Data leakage in feature engineering
  - Complex race dynamics difficult to capture
  - Data preprocessing and gathering

- Improvements:
  - Additional feature engineering: weather conditions, qualifying performance details, driver/team dynamics and car metrics
- Focus on preprocessing to avoid underfitting
- Model alternatives

Thank you!