Classifying Pass-types in NFL using Statistical Learning Methods

Presented by :
Rahul Kamal Pandey
Karthika Madhusudhana
Sharath Tamvada

Motivation

This project is inspired from the state of the art technologies like NFL Next Gen Stats that were invested in National Football League data exploration. The NFL big data bowl was an initiative taken in pursuit of advancing data and sports analytics. Prior to these technologies, keeping track of real time data had not been possible. But with more and more new inventions, radical techniques in finding new and creative was to collect data in real time from the field has made machine learning applicable in sports. Idealistically, there is immense scope in improving performances of real time updates and insights from this data so managers can make decisions that shape the team over the course of the season.

Problem Statement

Coverage schemes, quarterback positioning, and team orientations are some of the critical decision making aspects that managers in the NFL have long explored. The NFL big data bowl this year aims to use data analytics to better understand the schemes and players that make for a successful defense against passing plays. With the Data at hand, this project aims to utilize predictive analysis to infer the effectiveness of an offensive team's play by classifying the type of pass and determine a successful pass over an incomplete pass.

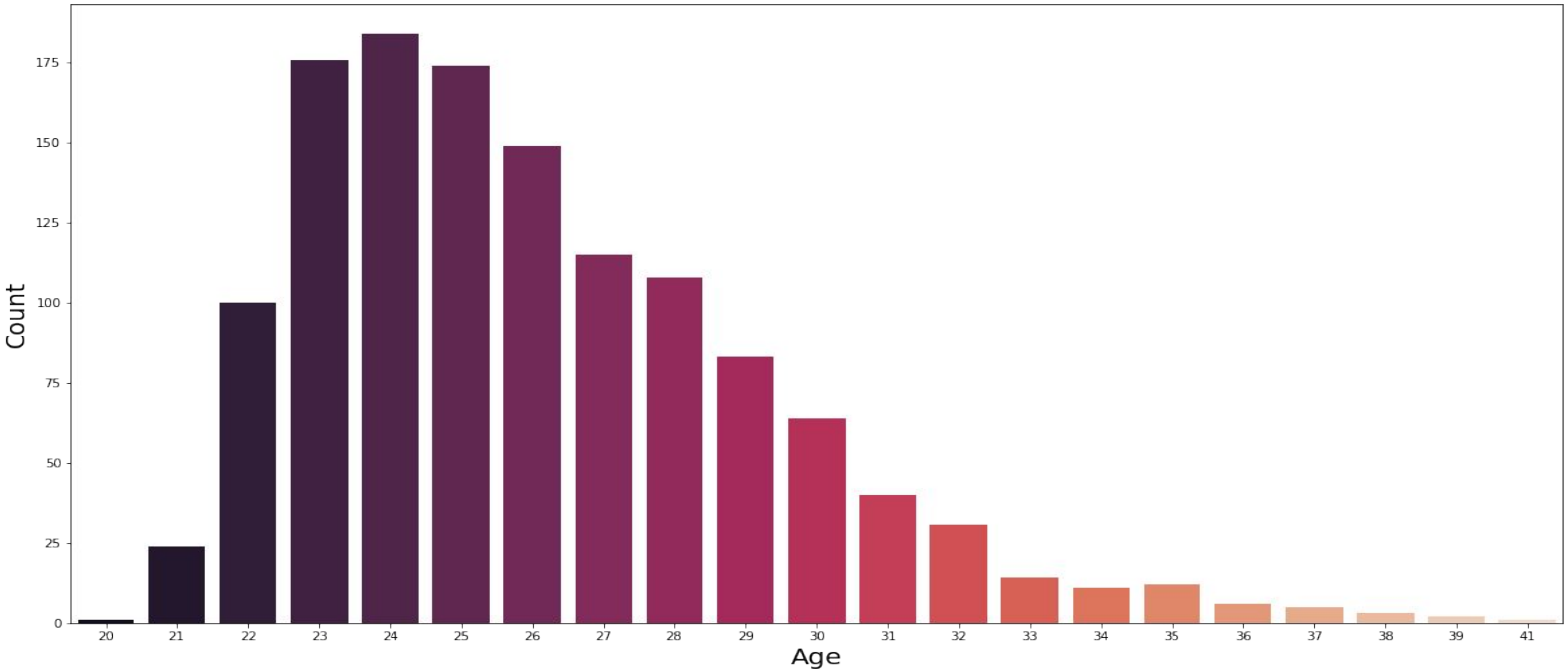
Data Description

- The 2021 Big Data Bowl contains player tracking, play, game, and player level information for all possible passing plays during the 2018 regular season

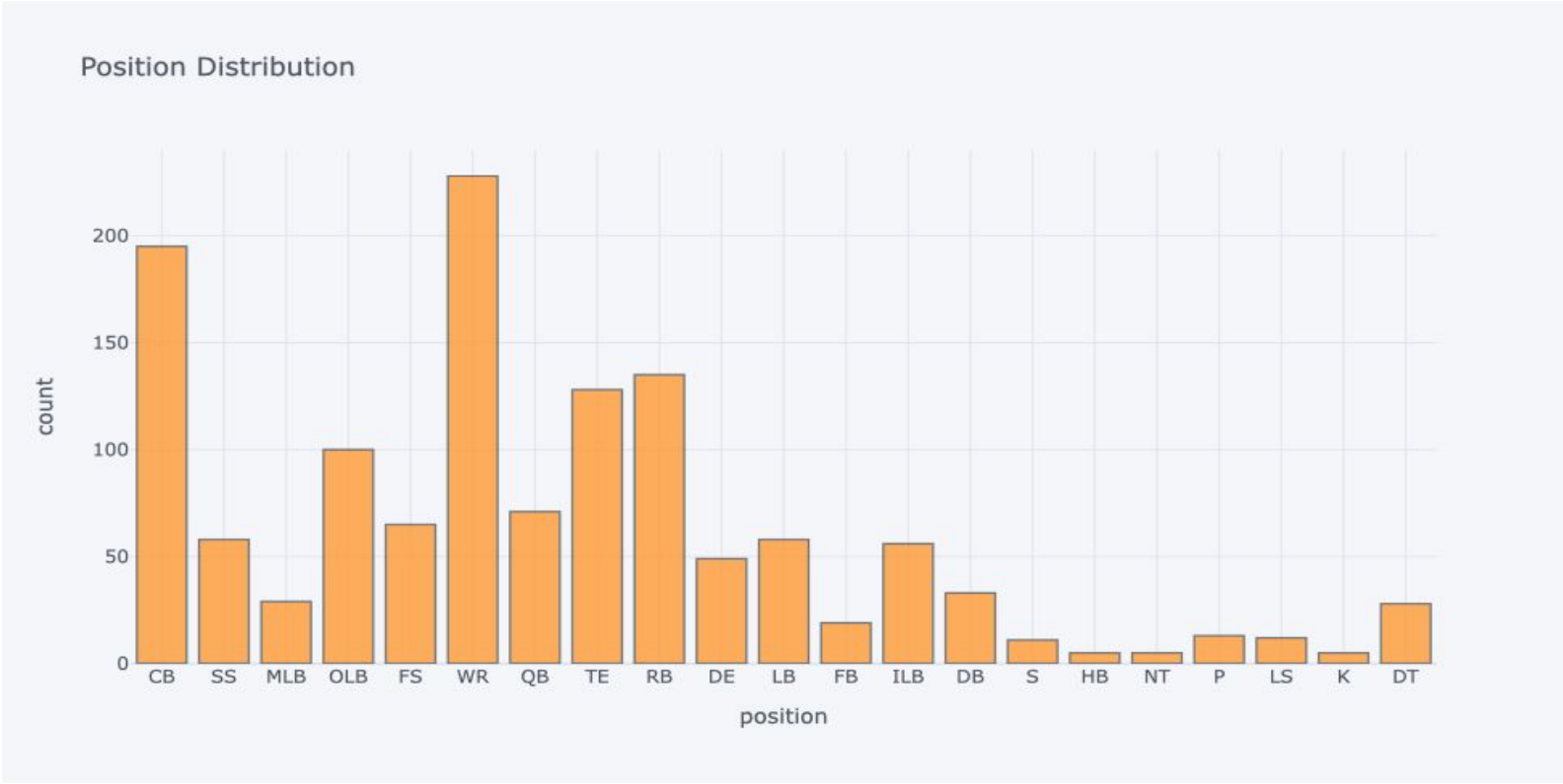
- Also, there are 17 weeks to a standard NFL Regular season and each of these week's files of tracking data contains the player tracking data from all passing plays
  - Game Data : Team playing in each game
  - Player Data : Player level Information of all players participated in tracking data
  - Play Data : Play-level Information for each game
  - Tracking Data : Player Tracking Data for all games in the weeks

Key variables: gameId, nflId, playId

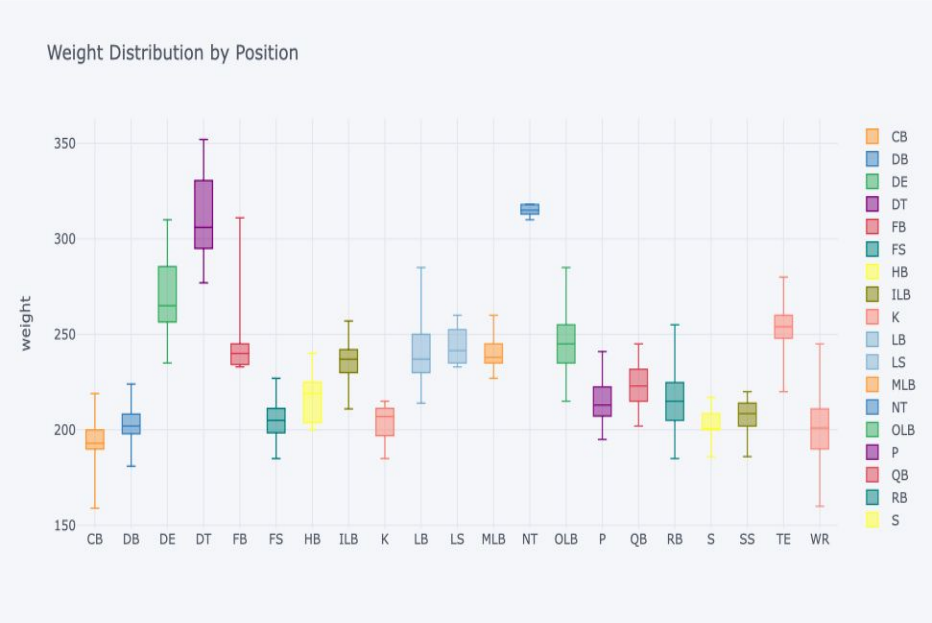
Exploratory Data Analysis

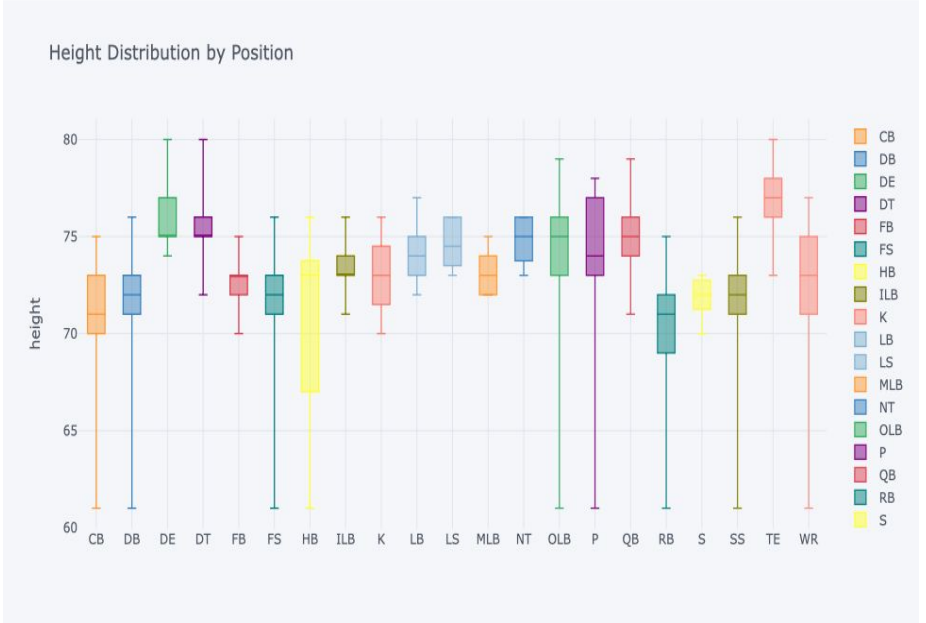


Exploratory Data Analysis



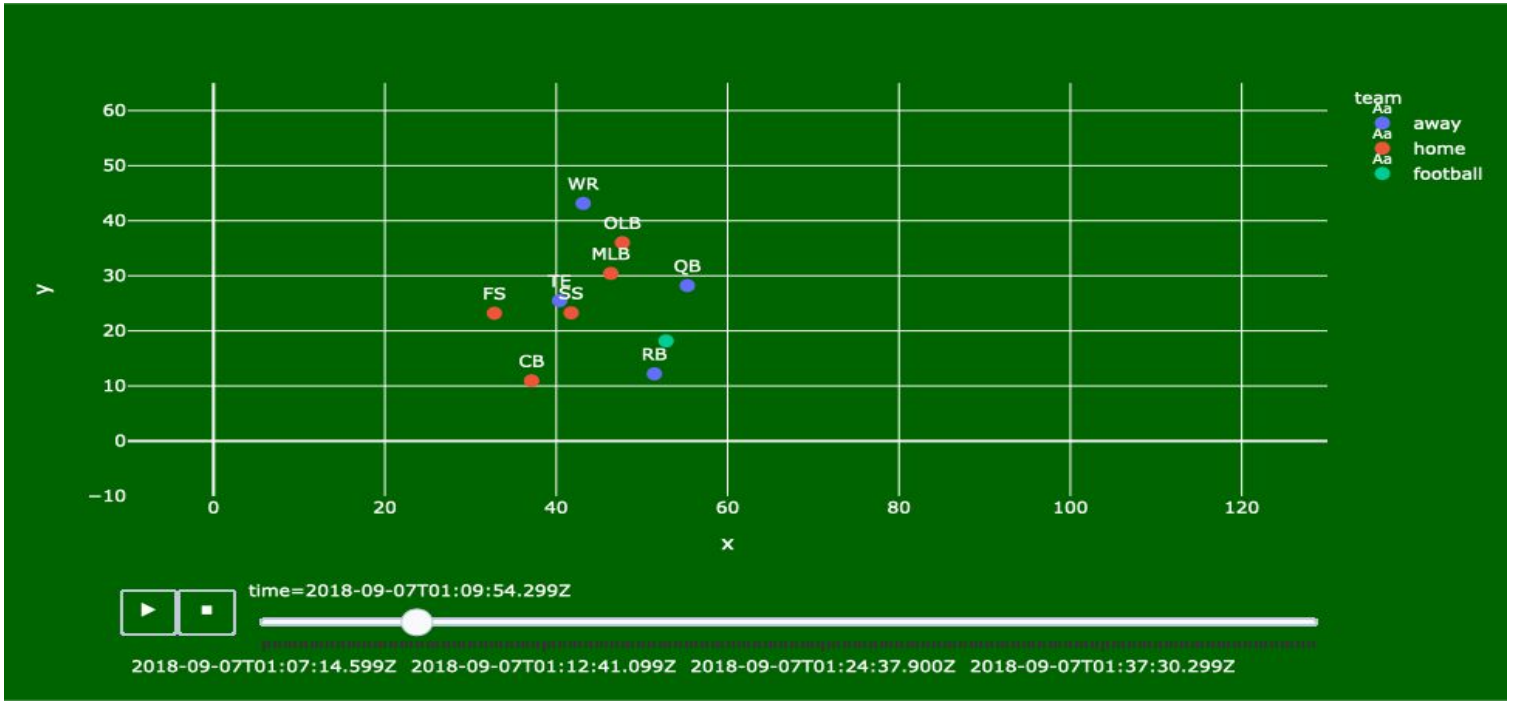
Exploratory Data Analysis





Note: Library plotly

Exploratory Data Analysis

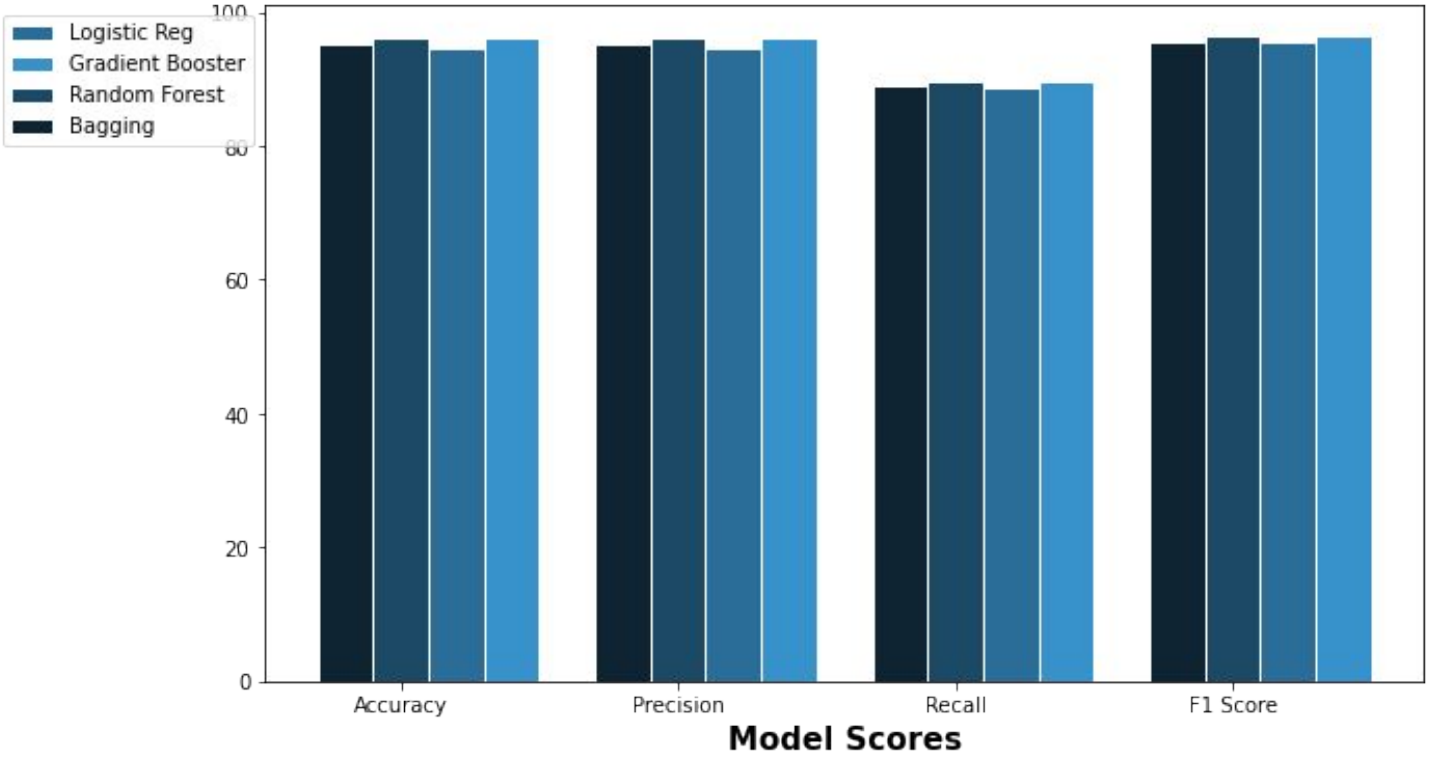


Classification

For classifying “pass-result” into Complete pass (C) , Incomplete pass (I) , Quarterback sack (S) , Intercepted pass (IN), we have used following classification models:

1. Logistic Regression
2. Gradient Boosting Regressor
3. Random Forest
4. Bagging

Comparison of Performance Measures



Comparison of Classification models

Models	F1-score
Logistic Regression	89.0256
Gradient Boosting Regressor (GBR)	**95.5083**
Random Forest	95.3030
Bagging	95.2015

The best model classifier we is **Gradient Boosting Regressor**

F1-score is the harmonic mean of precision and recall, so it provides the better score for any other performance measures.

Gradient Boosting Regressor

Confusion matrix:

	C	I	IN	S
C	4456	82	2	0
I	8	2486	5	1
IN	1	164	4	0
S	0	13	0	472

C - Complete pass
I - Incomplete pass
IN - Intercepted pass
S - Quarterback Sack

Result

Classification report:

	Precision	Recall	F1-score	Support
C	1.00	0.98	0.99	4540
I	0.91	0.99	0.95	2500
IN	0.36	0.02	0.04	169
S	1.00	0.97	0.99	485
Accuracy			0.96	7694
Macro avg	0.82	0.74	0.74	7694
Weighted avg	0.95	0.96	0.96	7694

Conclusion

- We found that among the models that outperformed the others, Gradient Boosting Regression model indicated the highest F-1 score for the dataset.

- Gradient Boosting Regressor (GBR) classified pass-types with an **F1 score** of **95.5083**

- It is evident from this project that indeed classification models can be used to determine whether or not a play will result in a yard line gain during that particular play. General managers can then make decisions based on these insights to better improve gameplay strategies.

References

1. https://operations.nfl.com/the-game/big-data-bowl/

2. https://nextgenstats.nfl.com

3. nfl-big-data-bowl-2021