

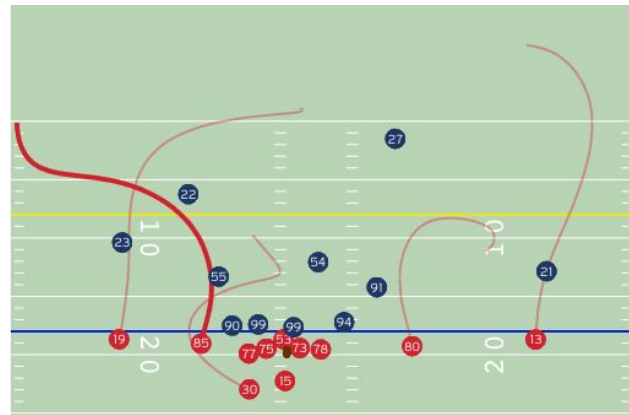
2020 NFL Big Data Bowl: Utilizing Pyspark to Predict Achieving a First Down on American Football Plays


Group 3:
Anoop Nath, Lauren O'Donnell, OC Ofoma




Introduction

- NFL started Next Gen Stats in 2013
 - RFID chips in player pads and footballs
 - Data on all players on the field during every play of every game since 2017
 - Opened up possibility for more advanced data analytics to be conducted
 - Captures player data such as location, speed, distance traveled and acceleration at a rate of 10 times per second
- NFL Big Data Bowl started in 2019 on Kaggle
 - Annual Competition open to the public





What model type will best predict whether a play will achieve a first down?



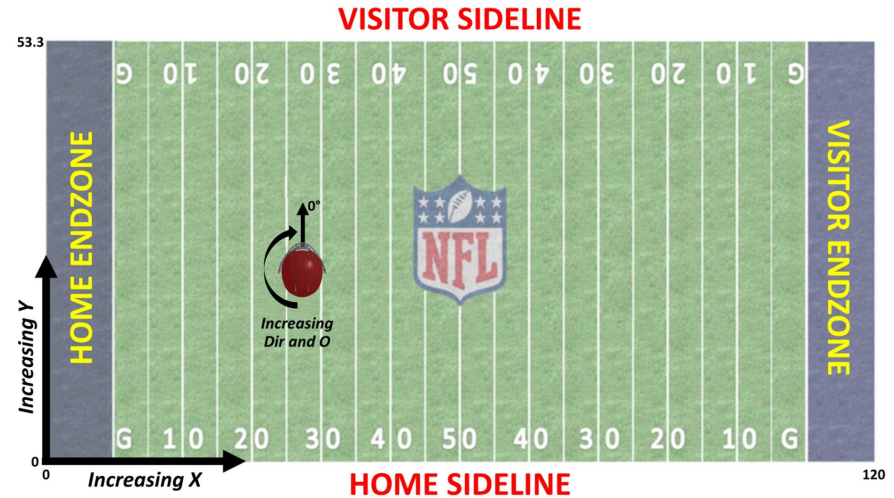
Data Introduction

- **Dataset:**

- The dataset contains Next Gen Stats tracking data for running plays in NFL games.
- Grouped by 'PlayId'.
- Each row corresponds to a single player's involvement in a single play.

- **Features:**

- Includes 49 features:
 - Player's position, speed, and acceleration.
 - Game identifiers, teams, quarter, down, and distance needed for a first down.
 - Environmental factors such as stadium, location, weather, temperature, humidity, and wind speed/direction.
 - Player-specific details such as name, height, weight, and college.

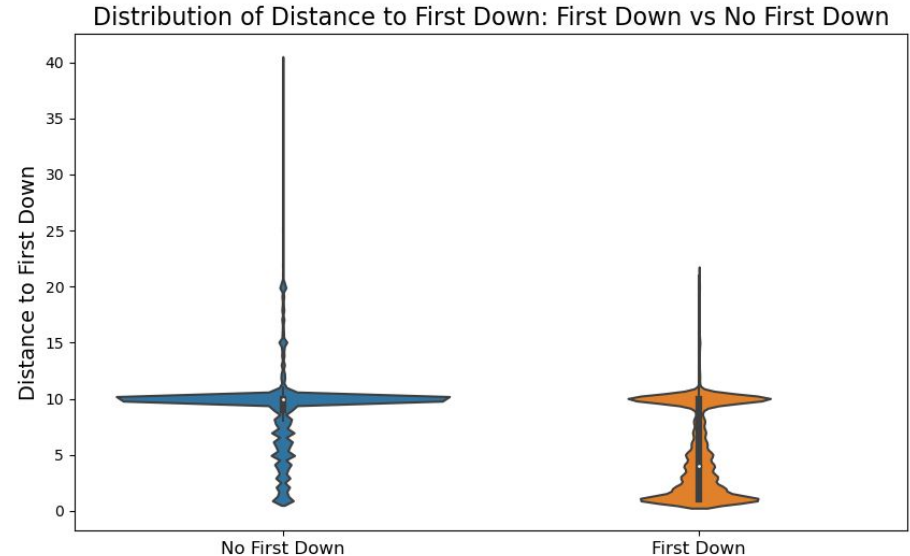


Data Preprocessing

- Create new variables to add as features:
 - Distance to End Zone
 - Distance of Closest Defender
- Create response variable:
 - First Down Gained
- Assembled feature variable vector containing:
 - Distance to End Zone
 - Distance to Closest Defender
 - Number of Defenders in the Box
 - Down
 - Distance to first down
- Remove rows with null values
- Data was split into training, validation, and holdout datasets

Exploratory Data Analysis

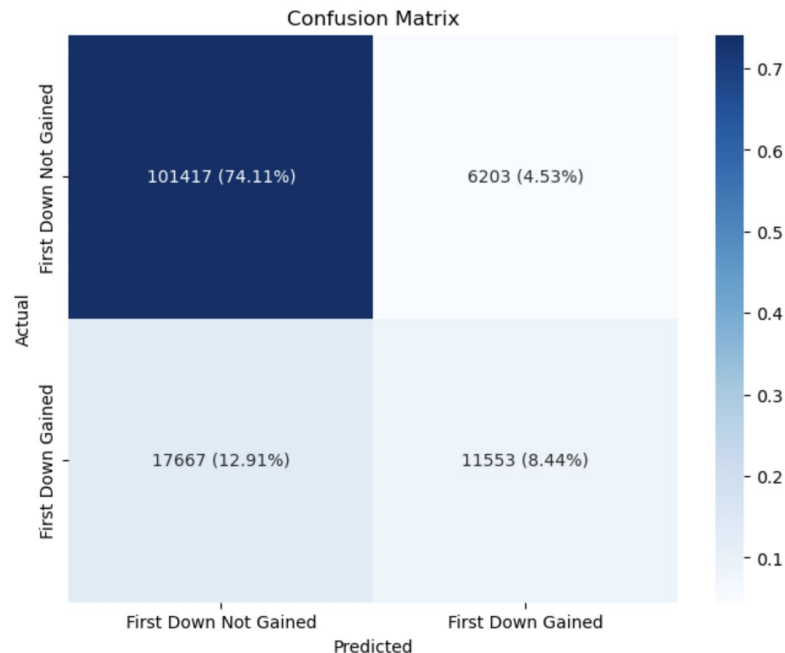
- 682,000 rows
 - 31,000 rushing plays from Sep 2017 to Nov 2019
 - 22 players per play
- Response Variable = first_down_gained
 - 21% of plays resulted in first downs
- Five Feature Variables
 - Distance to first down was strongest



Model Performance/Evaluation

Logistic Regression (Benchmark Model)

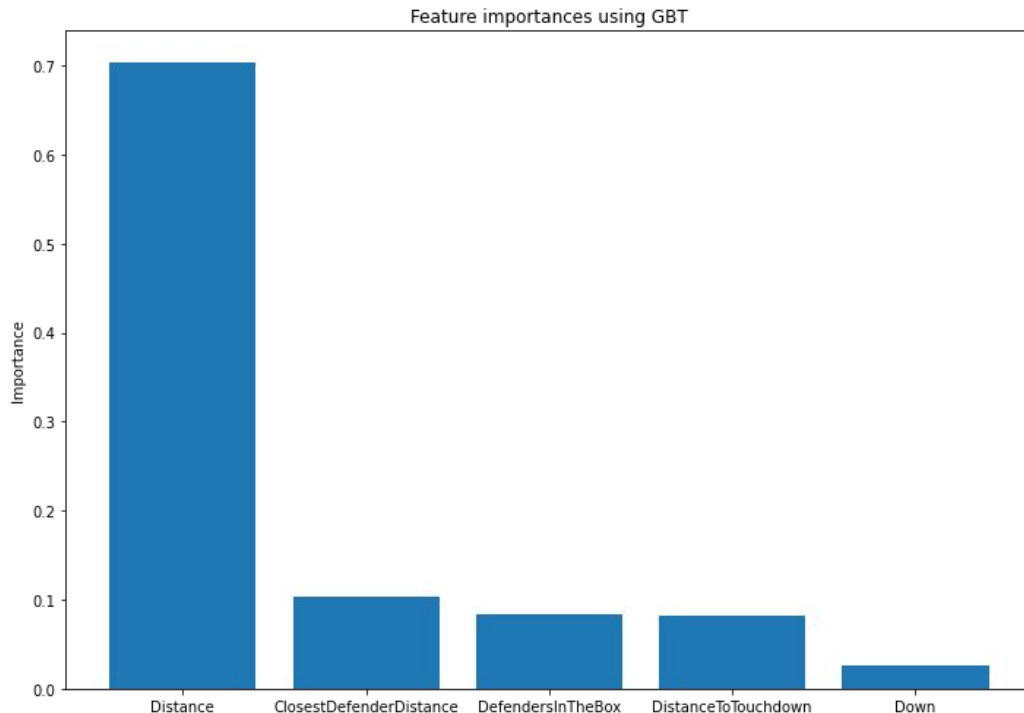
- Accuracy: 82.56%
- Precision: 80.87%
- Recall: 82.56%
- F1 Score: 0.8087
- AUC: 0.7628



Model Performance/Evaluation

Gradient Boosting

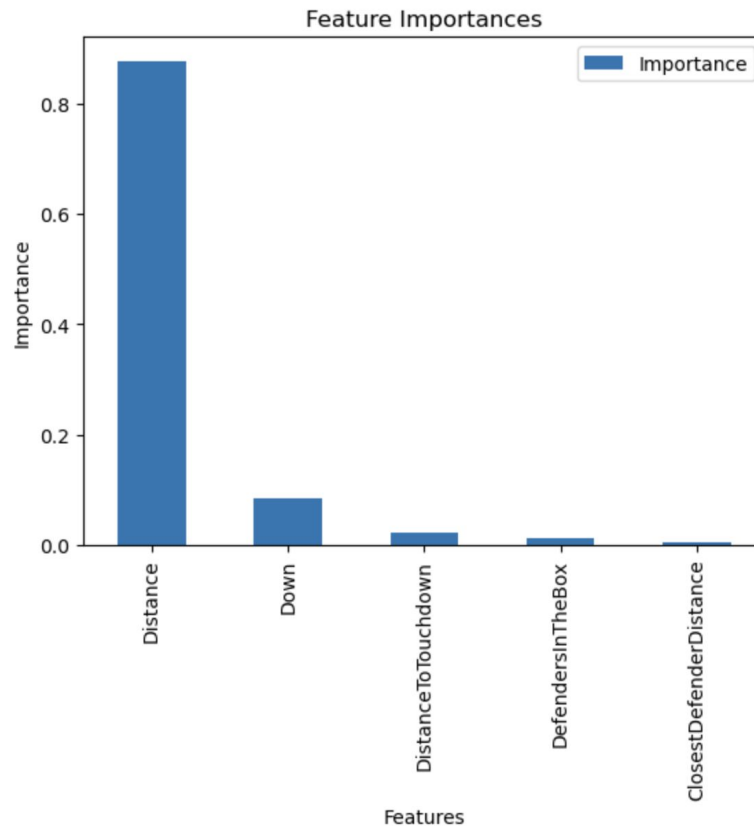
- Accuracy: 82.86%
- Precision: 81.26%
- Recall: 82.86%
- F1 Score: 0.812
- AUC: 0.7695



Model Performance/Evaluation

Random Forest (Champion Model)

- Accuracy: 83.24%
- Precision: 81.67%
- Recall: 83.24%
- F1 Score: 0.8166
- AUC: 0.7513



Model Performance/Evaluation Comparison

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	82.56%	80.97%	82.56%	0.8087	0.7628
Gradient Boosting	82.86%	81.26%	82.86%	0.8122	0.7695
Random Forest	83.24%	81.67%	83.24%	0.8166	0.7513

Conclusions and Future Research

- The Random Forest performed the strongest of the three models.
- Future research:
 - How other environmental factors impact play, i.e., sun positioning, time of day, stadium altitude, etc.
 - Deep Learning Techniques
 - Investigate further relationships between variables to explore modifications to fitting simple models