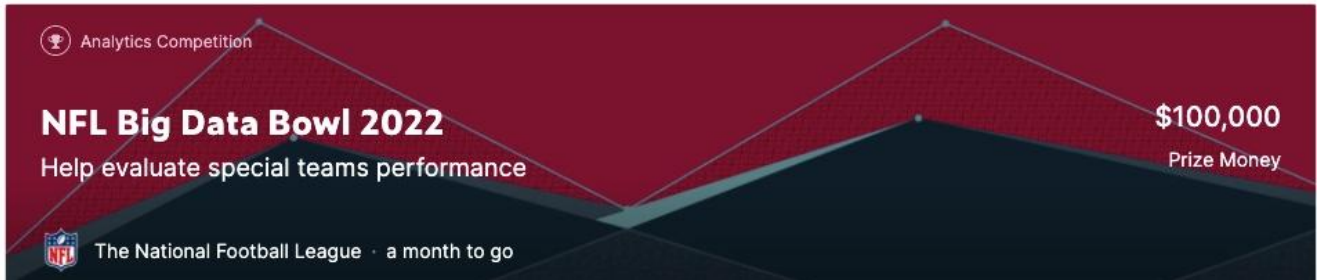
NFL Data Analytics

Darron Kotoyan, Nic Vamis, & Obada Yosef

NV

Motivation

- Our Motivation to pick this specific dataset is that it we are entering into a data analytics competition with this dataset for a cash prize.
- We are also interested in working in the sports analytics industry and this project will be a great opportunity to gain some experience in this field.
- Overall goal: how does the kick distance and hang time of a kick effect the outcome of a kick return. This can be used by NFL teams to evaluate how their special teams units are performing and how they will perform in the future.



Data Cleaning

- Step one: Choosing the datasets to merge.
- Step two: Deciding wanted variables and removing unwanted variables in preparation for merge
- Step three: Merge datasets respective of “gameId”
- Step four: Delete all rows with N/A in important variable information.
- Step five: Remove all duplicate rows from the dataset.

Variables

- **Game ID:** Specific Identification for each game
- **Play ID:** Specific Identification for each play
- **Play Description:** Type of play, yardage, and how it was received
- **Quarter:** What quarter the game is currently in
- **Special Teams Play Type:** Kickoff or Punt
- **Kick Length:** distance of the kick in yards
- **Kick Return Yardage:** distance of the kick return in yards
- **Play Result:** Net yards gained by kicking team
- **Hangtime:** The amount of time in seconds the ball spends in the air
- **Kick Contact Type:** Details on how the punt was fielded
- **Return Type:** Classified kick return yardage into short, medium, or long distance (Short<20 , 20<Medium<50, 50<Long)

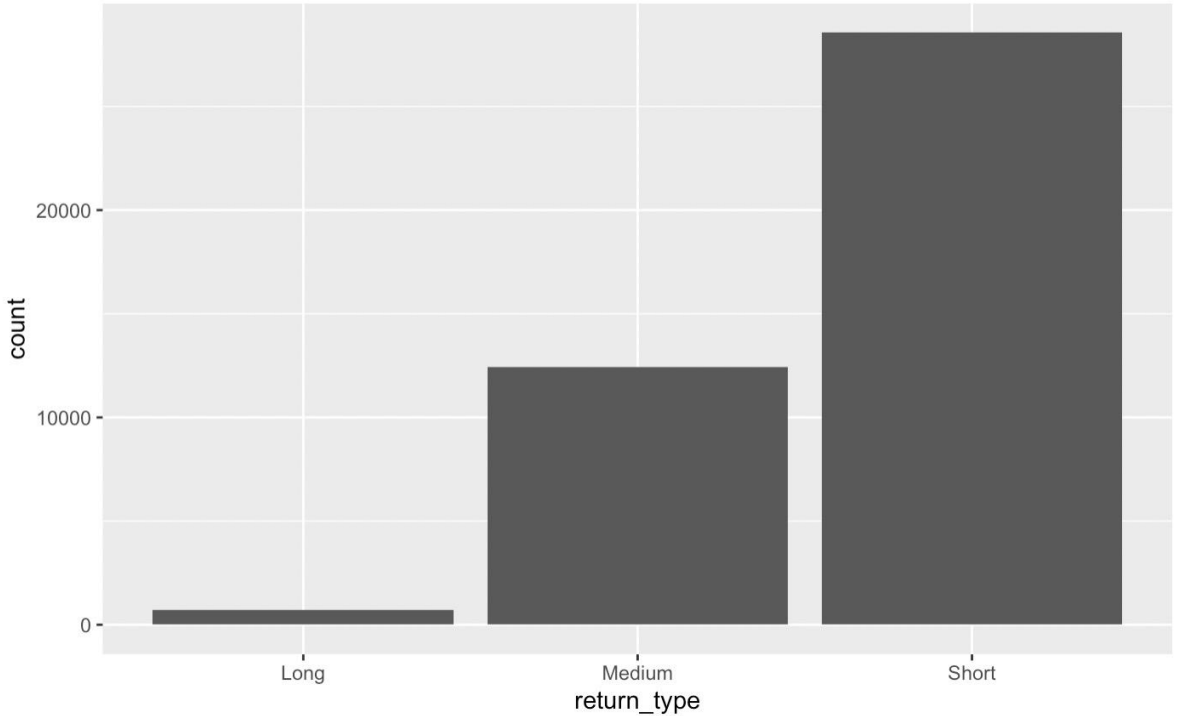
DK

Summary Statistics

kickLength
Min: 2
Max: 90
Mean: 54.58

kickReturnYardage
Min: -16
Max: 104
Mean: 15.08

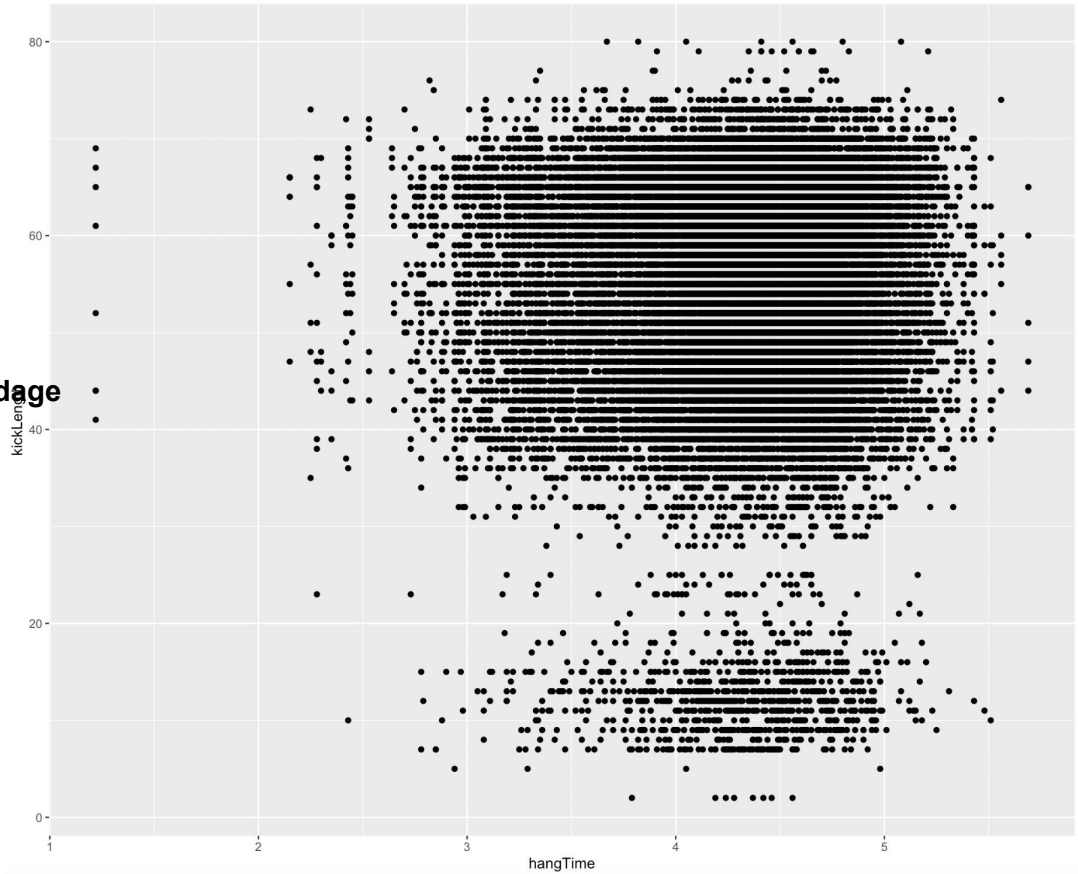
hangTime
Min: 1.22
Max: 5.69
Mean: 4.317



kickLength
Min: 2
Max: 90
Mean: 54.58

kickReturnYardage
Min: -16
Max: 104
Mean: 15.08

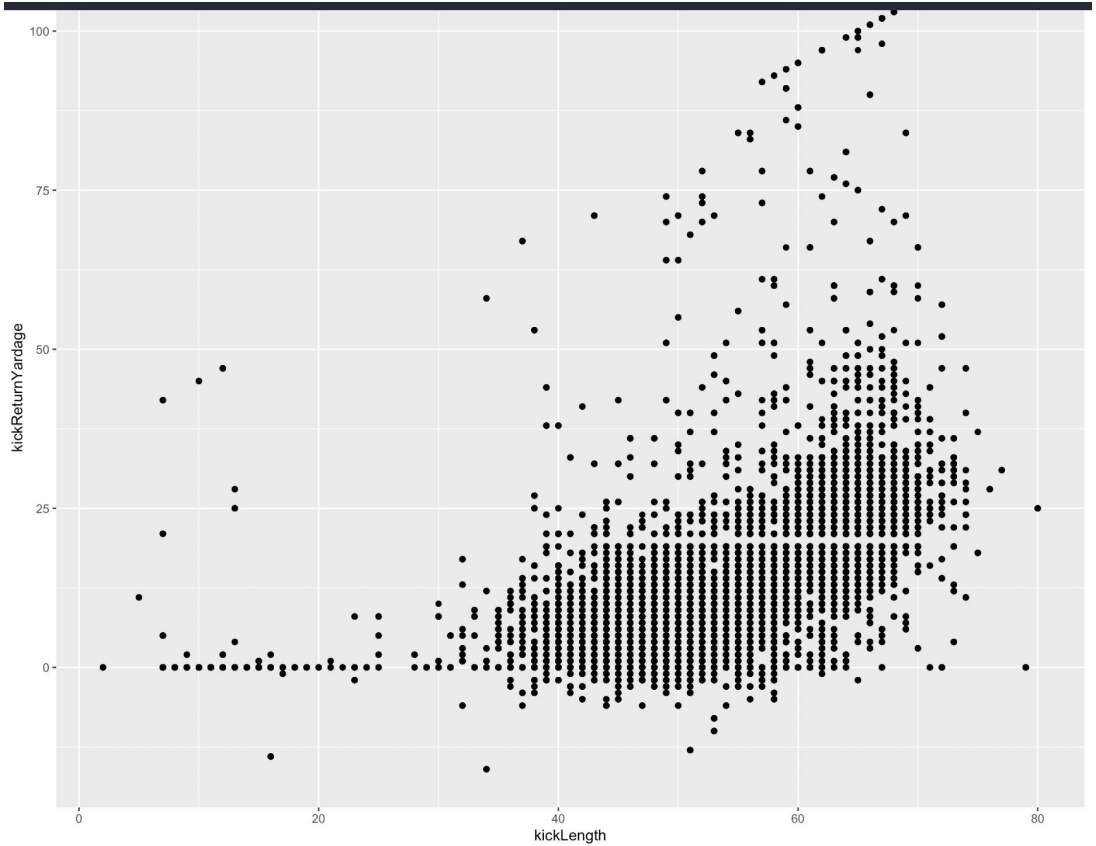
hangTime
Min: 1.22
Max: 5.69
Mean: 4.317



kickLength
Min: 2
Max: 90
Mean: 54.58

kickReturnYardage
Min: -16
Max: 104
Mean: 15.08

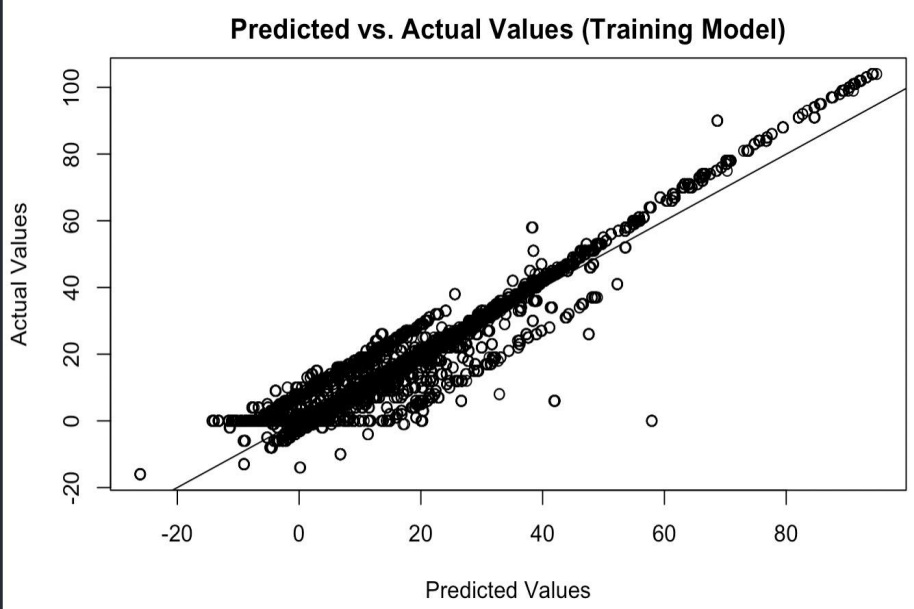
hangTime
Min: 1.22
Max: 5.69
Mean: 4.317

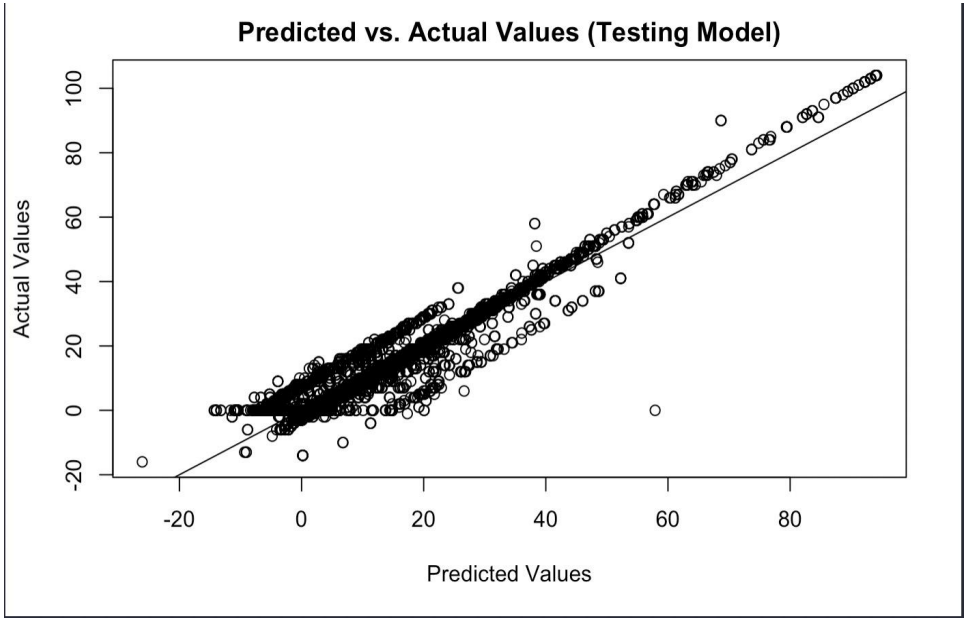


Yard Returned by Hangtime

- Linear Regression to predict kick return yardage
- X Variables: Hangtime, Kick distance, Kick length, Play result, Kick contact type
- Y Variables: Kick return yardage
- Conclusion: All though the plot between X and Y has a nice linear pattern and has a high R-squared, we estimated that the model is underfit.
- MSE of Training: 0.28 MSE of Testing: 0.92
- For Predicted kick returns in yards, 68.7% < 20 yds, 29.9% between 20-50 yds, 1.4% > 50 yds

OY


Predicted vs. Actual Values (Training Model)
Actual Values
Predicted Values
-20
0
20
40
60
80
100


Predicted vs. Actual Values (Testing Model)
Actual Values
Predicted Values
-20
0
20
40
60
80
100

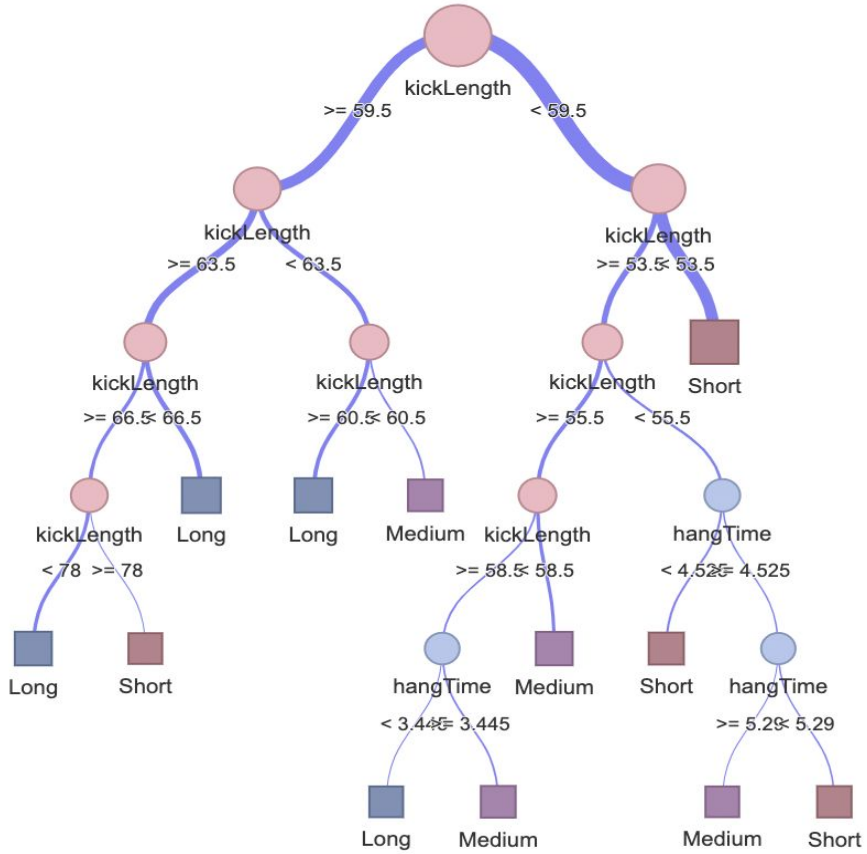
Predicting Short, Medium, or Long Return Distance

- Decision tree to determine whether a kick return will be classified as short, medium or long.
- X Variables: Hangtime and kick length,
- Y Variables: Return Type
- Conclusion: The most significant split is with the kick distance. If the hangtime was between 4.27 and 4.59, it had more of an impact on the kick return yardage.

NV

NV


hangTime
kickLength
Long
Medium
Short


kickLength
>= 59.5
< 59.5
kickLength
>= 63.5
< 63.5
kickLength
>= 53.5
< 53.5
kickLength
>= 66.5
< 66.5
kickLength
>= 60.5
< 60.5
kickLength
>= 55.5
< 55.5
Short
kickLength
< 78
>= 78
Long
Long
Medium
kickLength
>= 58.5
< 58.5
hangTime
< 4.525
>= 4.525
Long
Short
hangTime
Medium
Short
hangTime
< 3.445
>= 3.445
>= 5.29
< 5.29
Long
Medium
Medium
Short

Predicting Kick Contact Type

- Random Forest model to predict the kick contact type
- X Variables: Hangtime and kick length
- Y Variables: Kick contact type
- Conclusion: Kick length seems to be the more significant of the two X variables. For each of the 13 possible categorical outcomes, we can conclude that Class CC (Clean Catch from air) has the highest prevalence out of all the other outcomes. In the confusion matrix CC has the highest number of true positives. Our model has 66% accuracy.

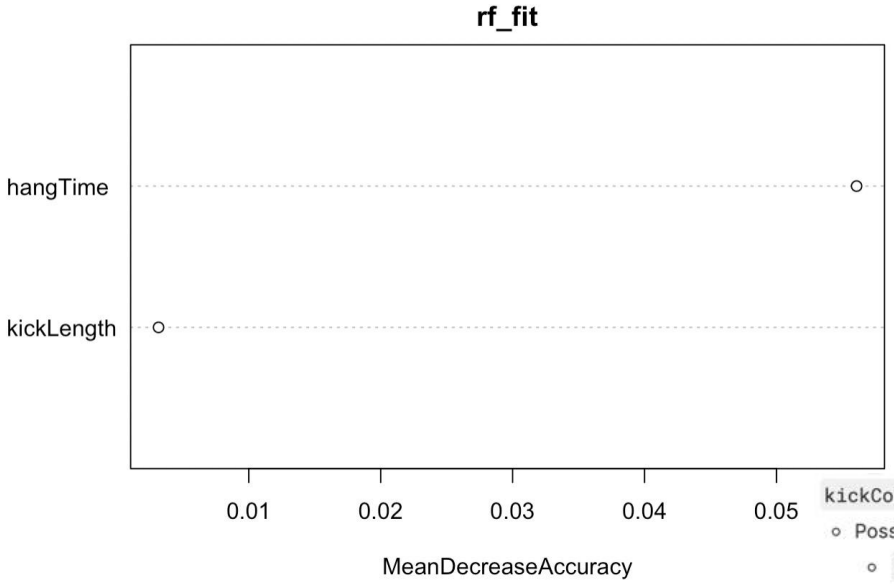
Confusion Matrix and Statistics

Prediction \ Reference	BB	BC	BF	BOG	CC	CFFG	DEZ	ICC	KTB	KTC	KTF	MBC	MBDR	OOB
BB	6	0	18	0	471	5	0	0	0	0	0	0	0	0
BC	0	1	10	0	40	0	0	0	0	0	0	0	0	0
BF	7	1	224	0	1540	25	0	0	0	0	0	0	1	2
BOG	0	0	1	0	9	0	0	0	0	0	0	0	0	0
CC	4	0	140	0	6616	20	0	0	0	0	0	0	1	3
CFFG	3	0	76	0	211	27	0	0	0	0	0	0	1	2
DEZ	1	0	3	0	106	1	0	0	0	0	0	0	0	0
ICC	0	0	0	0	32	0	0	0	0	0	0	0	0	0
KTB	0	0	1	0	30	0	0	0	0	0	0	0	0	0
KTC	0	0	1	0	61	0	0	0	0	0	0	0	0	0
KTF	0	0	0	0	19	0	0	0	0	0	0	0	0	0
MBC	1	0	1	0	11	0	0	0	0	0	0	0	0	0
MBDR	0	0	5	0	231	1	0	0	0	0	0	0	0	0
OOB	1	0	16	0	429	6	0	0	0	0	0	0	0	4

Statistics by Class:

	Class: BB	Class: BC	Class: BF	Class: BOG	Class: CC	Class: CFFG	Class: DEZ	Class: ICC	Class: KTB	Class: KTC
Sensitivity	0.2608696	0.50000000	0.45161	NA	0.6747	0.317647	NA	NA	NA	NA
Specificity	0.9525137	0.99520338	0.84129	0.9990409	0.7290	0.971666	0.98935	0.996931	0.997027	0.994053
Pos Pred Value	0.0120000	0.01960784	0.12444	NA	0.9752	0.084375	NA	NA	NA	NA
Neg Pred Value	0.9982873	0.99990361	0.96847	NA	0.1241	0.994261	NA	NA	NA	NA
Prevalence	0.0022060	0.00019183	0.04757	0.0000000	0.9405	0.008153	0.00000	0.000000	0.000000	0.000000
Detection Rate	0.0005755	0.00009591	0.02148	0.0000000	0.6346	0.002590	0.00000	0.000000	0.000000	0.000000
Detection Prevalence	0.0479570	0.00489162	0.17265	0.0009591	0.6507	0.030692	0.01065	0.003069	0.002973	0.005947
Balanced Accuracy	0.6066916	0.74760169	0.64645	NA	0.7019	0.644657	NA	NA	NA	NA

	Class: KTF	Class: MBC	Class: MBDR	Class: OOB
Sensitivity	NA	NA	0.0000000	0.3636364
Specificity	0.998178	0.998753	0.9772618	0.9566011
Pos Pred Value	NA	NA	0.0000000	0.0087719
Neg Pred Value	NA	NA	0.9997056	0.9992979
Prevalence	0.000000	0.000000	0.0002877	0.0010551
Detection Rate	0.000000	0.000000	0.0000000	0.0003837
Detection Prevalence	0.001822	0.001247	0.0227316	0.0437368
Balanced Accuracy	NA	NA	0.4886309	0.6601187



`kickContactType` : Detail on how a punt was fielded, or what happened when it wasn't fielded (text).

- Possible values:
  - `BB` : Bounced Backwards
  - `BC` : Bobbled Catch from Air
  - `BF` : Bounced Forwards
  - `BOG` : Bobbled on Ground
  - `CC` : Clean Catch from Air
  - `CFFG` : Clean Field From Ground
  - `DEZ` : Direct to Endzone
  - `ICC` : Incidental Coverage Team Contact
  - `KTB` : Kick Team Knocked Back
  - `KTC` : Kick Team Catch
  - `KTF` : Kick Team Knocked Forward
  - `MBC` : Muffed by Contact with Non-Designated Returner
  - `MBDR` : Muffed by Designated Returner
  - `OOB` : Directly Out Of Bounds

Final Thoughts/Solution

Based off our analysis of the dataset we can conclude that the largest overall factor in kick return yardage is kick length. This hypothesis was of course expected as it makes the most logical sense.

Our goal was to discover the the relationship between hang time of kick and kick distance affect the outcome and return of the kick. Based on the decision tree, kicks that are less than 53 yards have a very high chance of having a short return, longer kicks vary in outcome.