Predicting Colts Run-Play Direction with KNN Model

Sean G. Bradford

Fundamentals of Machine Learning 64060

Data Source: https://www.kaggle.com/competitions/nfl-big-data-bowl-2020/data?select=train.csv

Data Source

The data utilized in this project is real-world data because it was gathered during live NFL games from the 2017, 2018 and 2019 seasons (as opposed to being gathered in an experimental setting). It contains Next Gen Stats tracking data for running plays executed during the aforementioned time frame. Each row represents a single player's involvement in a single play. Link to the data source is shown on the title page.

Objectives

The Cleveland Browns Analytics Team is challenged to create a machine-learning model that can aid the team's run-defense against their next opponent, the Indianapolis Colts. The model must be equipped for different game-time scenarios, classify the direction of the play, and generate accurate results (accuracy of 80 percent or greater).

Methodology

Run direction is a critical element for a defense. By accurately predicting the direction of a running play, a defensive coordinator can stunt levels of defensive pressure into the running back's path and enhance the team's run-stopping ability. To accomplish this goal, the data was filtered down to only include Colts running plays. Variables "X", "Y" and "Orientation" were utilized in the model because they determine the running back's field position and enable situational model usage.

Predictor	Definition
X	Player position along the x axis of the field
Y	Player position along the y axis of the field
Orientation	Degree of the player position

Table 1. Definitions of predictors

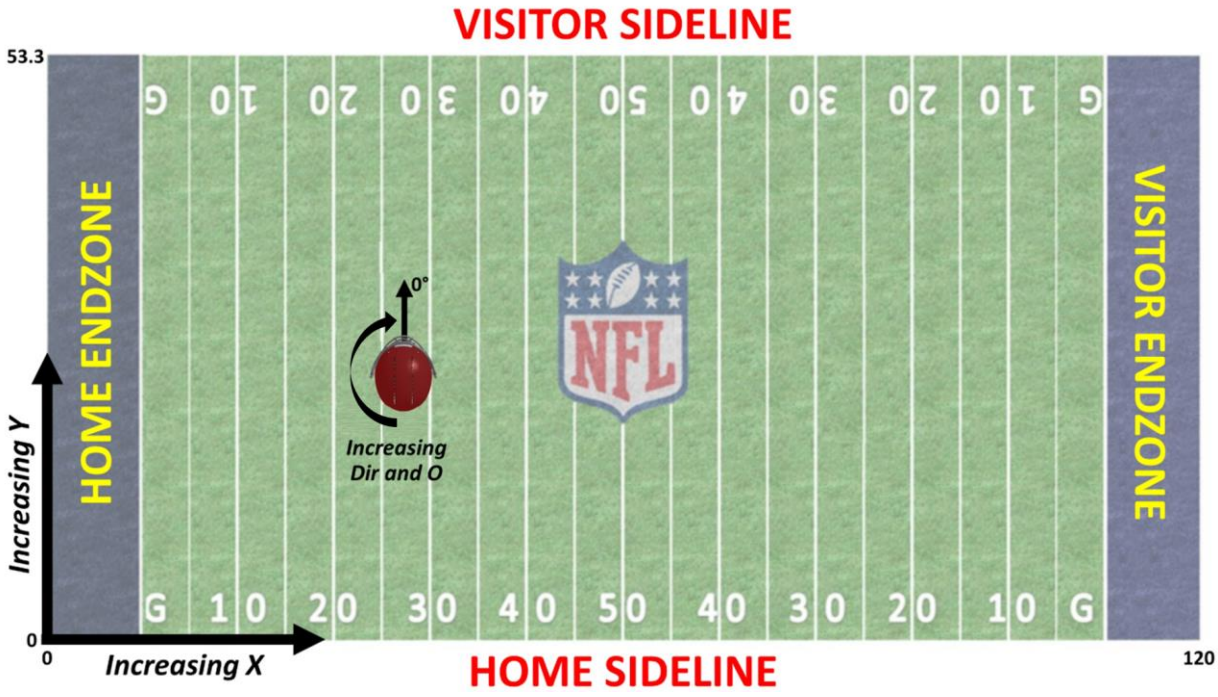

Figure 1. Visualization of predictors

Source: https://www.kaggle.com/competitions/nfl-big-data-bowl-2020/data

The k-nearest neighbor algorithm was implemented based on its classification reliability, and its capacity to store complete datasets in training (Tsypin & Roder, 2007). This allows the model to take the field position of a play, compare it against the run-directions of previous Colts plays from similar field positions, then predict the direction of the current play. The model was trained using 80 percent of the original data and tested using 20 percent of the original data. Training with 80 percent of the data allows the model to store more information for future analyses and leaves enough data for accurate testing. During model tuning, the 'train' function in R was used to determine the optimal value for k, then the 'CrossTable' function was utilized for calculating the model's performance.

Analysis

Upon completion of model tuning, the optimal value for k was found at 13, which yielded an accuracy of 88.03 percent.

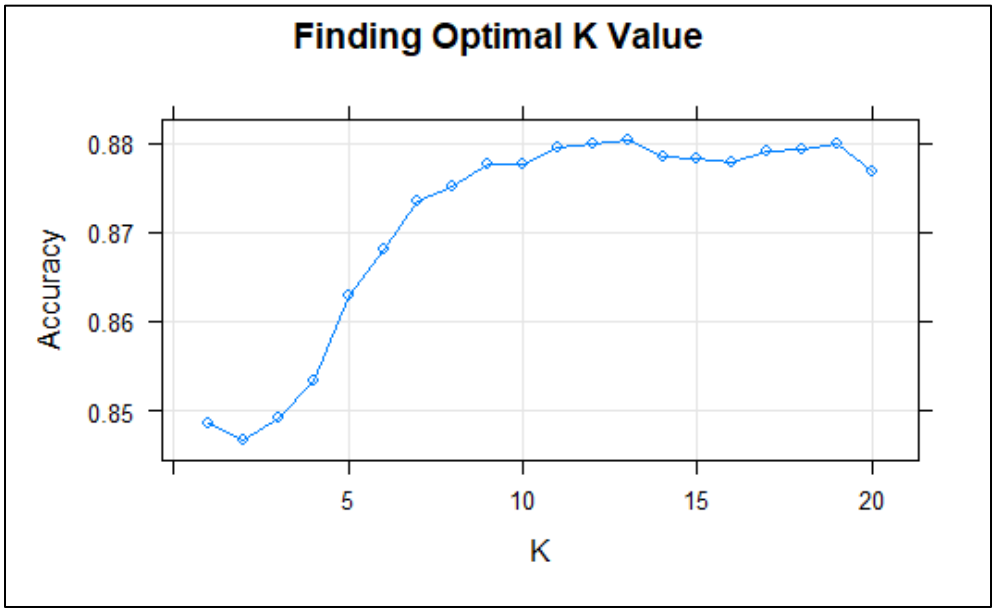

Figure 2. Output from using train() to determine the best k value for the model

After setting k equal to 13, the cross table showed 96 true positives, 10 false positives, 94 true negatives and 11 false negatives. Out of 211 observations, only 21 were misclassified by the model. Performance calculations also reflected high-quality results for the model: 90.05 percent accuracy, 89.72 percent recall, 90.57 percent precision and 90.38 percent specificity.

		Predicted Run Direction	
		Left	Right
Actual Run Direction	Left	94	10
	Right	11	96

Table 2. Confusion matrix results after setting k = 13

Finally, the model was run through a mock-game situation. In this case, the Colts were on the Browns' five yard-line (X=105), closer to the home sideline (Y=25), and the running back was oriented at 21 degrees (Orientation=21). The model determined that the running play would be directed to the left.

Conclusions

The findings from the analysis can provide an advantage for the defense in obvious run-play situations. By classifying the direction of a running play, the defense can apply pressure toward the path of the running back. This disrupts offensive blocking assignments and creates a better opportunity for tackles to be made in the backfield. The findings could also benefit the offense. Analyzing the team's past running tendencies could uncover predictable trends. Identifying these trends could empower the offense to call different plays in certain parts of the field and disrupt opposing defenses. The evaluation of the model showed that it meets the accuracy requirements from management and has a high percentage in recall, precision and specificity. These metrics account for discrepancies between different ratios true positives, true negatives, false positives and false negatives.

References

NFL Big Data Bowl. (2020). Kaggle.com. Retrieved December 1, 2022, from https://www.kaggle.com/competitions/nfl-big-data-bowl-2020/data?select=train.csv

NFL Tendency Analysis and Basic Play Type Prediction – MSiA Student Research. (2020). Sites.northwestern.edu. https://sites.northwestern.edu/msia/2020/01/31/nfl-tendency-analysis-and-basic-play-type-prediction/

Shmueli, G., Bruce, P., Yahav, I., Patel, N., Lichtendahl, K. (2018). Data Mining for Business Analytics: Concepts, techniques, and applications in R. Wiley.

Tsypin, Maxim & Roder, Heinrich. (2007). On the Reliability of kNN Classification. Lecture Notes in Engineering and Computer Science. 2167.

Uprety, Dambar. (2022). *k*-Nearest Neighbor (kNN) Classification. Retrieved December 1, 2022, from https://kent.instructure.com/courses/25514/pages/view-k-nn-tuning-5-27-minutes?module_item_id=2118251