Predicting NFL Field Goal Outcome

Quang Nguyen

Spring 2021 CDA

Background

- Field goal scoring is important in football
- There are many factors that could affect the outcome of a field goal attempt: players, field position, game situation, coaching, playing condition...
- Use statistical learning to predict field goal success in the NFL
- Methods:
  - Logistic Regression
  - Classification Tree
  - Random Forest

Data

- All field goals from 2005 to 2015
- Obtained from Michael Lopez's GitHub repository

(`https://github.com/statsbylopez/StatsSports/blob/master/Data/nfl_fg.csv`)

Team	Year	GameMinute	Kicker	Distance	ScoreDiff	Grass	Temp	Success
PHI	2005	3	Akers	49	0	FALSE	72	0
PHI	2005	29	Akers	49	-7	FALSE	72	0
PHI	2005	51	Akers	44	-7	FALSE	72	1
PHI	2005	14	Akers	43	14	TRUE	82	0
PHI	2005	60	Akers	23	0	TRUE	75	1

- Created additional features:
  - `Leading`: Whether or not the kicking team is leading
  - `Period`: Quarters 1, 2, 3, 4, overtime
  - `ScoreType`: 1 possession, 2 possession, 3+ possession
  - `Foot`: Kicker's dominant foot

Goals

- Predict all field goal observations in 2015 from prior data (2005-2014)
- Examine and compare different prediction methods
  - Logistic Regression
  - Classification Tree
  - Random Forest

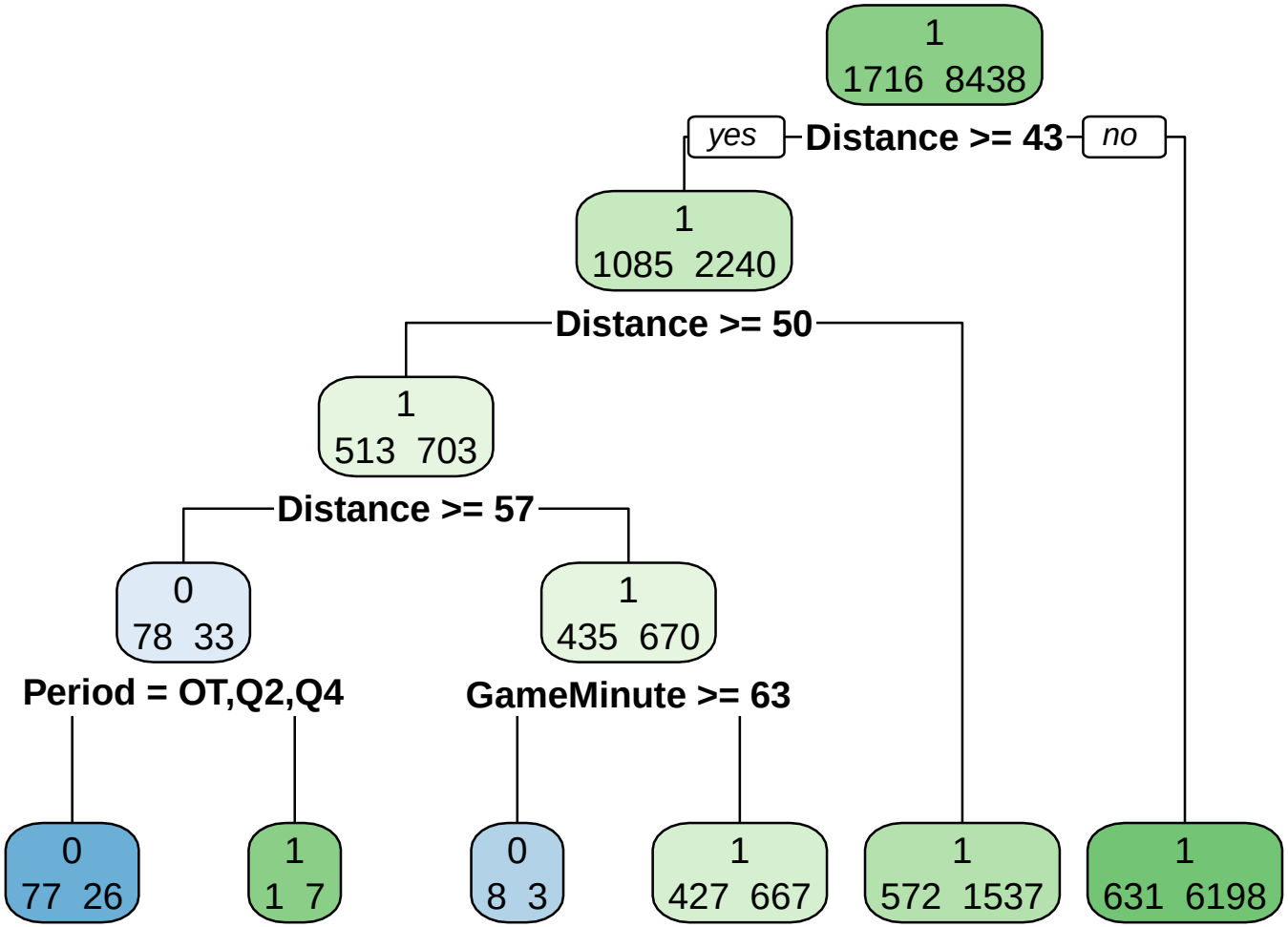
Logistic Regression

- Use stepwise AIC to get the best subset of predictors
  - {`Distance`, `Grass`}
- Model: $\textit{logit}(\hat{\pi}) = 5.7776 - 0.1024\ \textit{Distance} - 0.1624\ \textit{Grass}$

	Estimate	OR	95% CI	Std. Error	z value	Pr(>|z|)
(Intercept)	5.7776	322.9859	(242.9107,432.2403)	0.1470	39.3052	0.0000
Distance	-0.1024	0.9027	(0.8969,0.9085)	0.0033	-31.2961	0.0000
GrassYes	-0.1624	0.8501	(0.7597,0.9508)	0.0572	-2.8385	0.0045

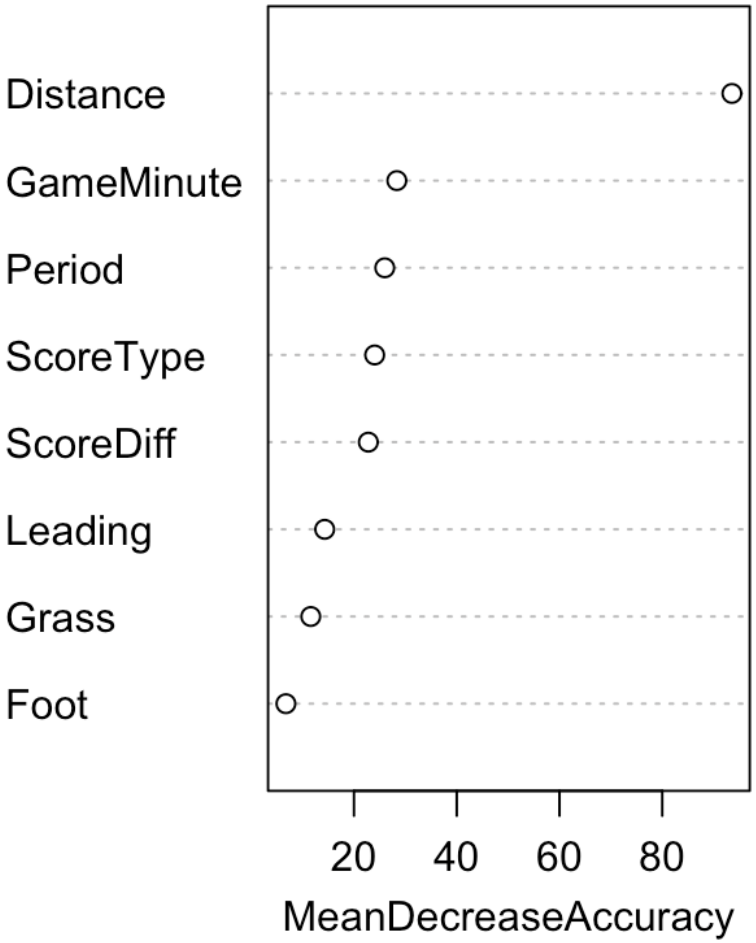
Classification Tree

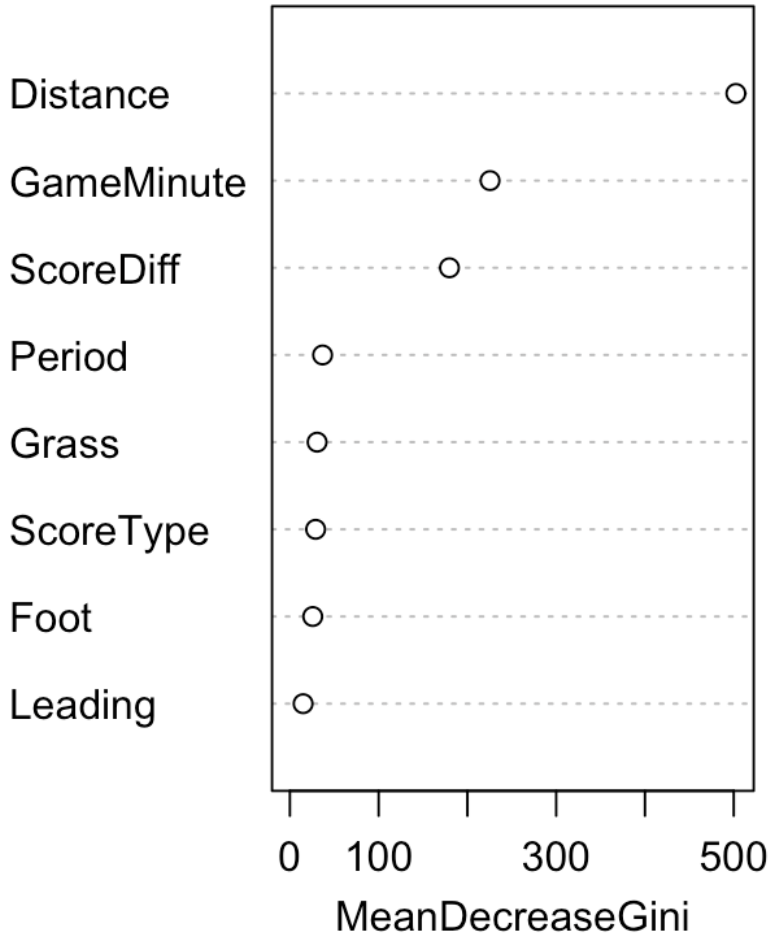
- Cross validation: 75-25 train-validation split



Random Forest

- Also perform cross-validation to obtain the final model
- Variable importance





Comparison

- Classification tables for predicting field goal success

	Logistic			Tree			RF	
	Actual			Actual			Actual	
Predicted	0	1		0	1		0	1
0	10	10		7	9		10	6
1	146	867		149	868		146	871

- How does each method perform?
  - Logistic regression, with accuracy $(10 + 867)/1033 = 0.8490$
  - Decision tree, with accuracy $(7 + 868)/1033 = 0.8470$
  - Random forest, with accuracy $(10 + 871)/1033 = 0.8529$

Discussion

- Random forest gave the best prediction
- Distance is important (unsurprising!)
- Future work:
  - Build better models with higher prediction accuracy
  - Extract more variables from play-by-play data
  - Try out other statistical learning methods

References

- Agresti, A. (2019). *An Introduction to Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

Cheers!