PREDICTING NBA CAREER DURATION BASED ON ROOKIE SEASON

Andy Chuang & Kevin Nguyen

THE DATA

Where did the data come from?

- National Basketball Association (NBA)

Data Overview

- In-game statistics for players in the NBA during their rookie season
- 1,294 observations and 12 variables
- 0 missing values

Data Cleaning

- Duplicate players → Aggregate duplicate rows together using mean
- Variables are on per-game basis
    - Potential correlation
    - Centered and scaled data
    - Standardized variables to per-minute basis

MOTIVATION

Motivation

- NBA franchises want to sign promising rookies while they are still cheap
- Very useful if we can predict a rookie's career duration using their in-game statistics
- Rookie contracts are generally about 4 years in length (with team options)

Research Question

- Which in-game statistics from a player's rookie season are most strongly associated with a career duration of more than 4 years?

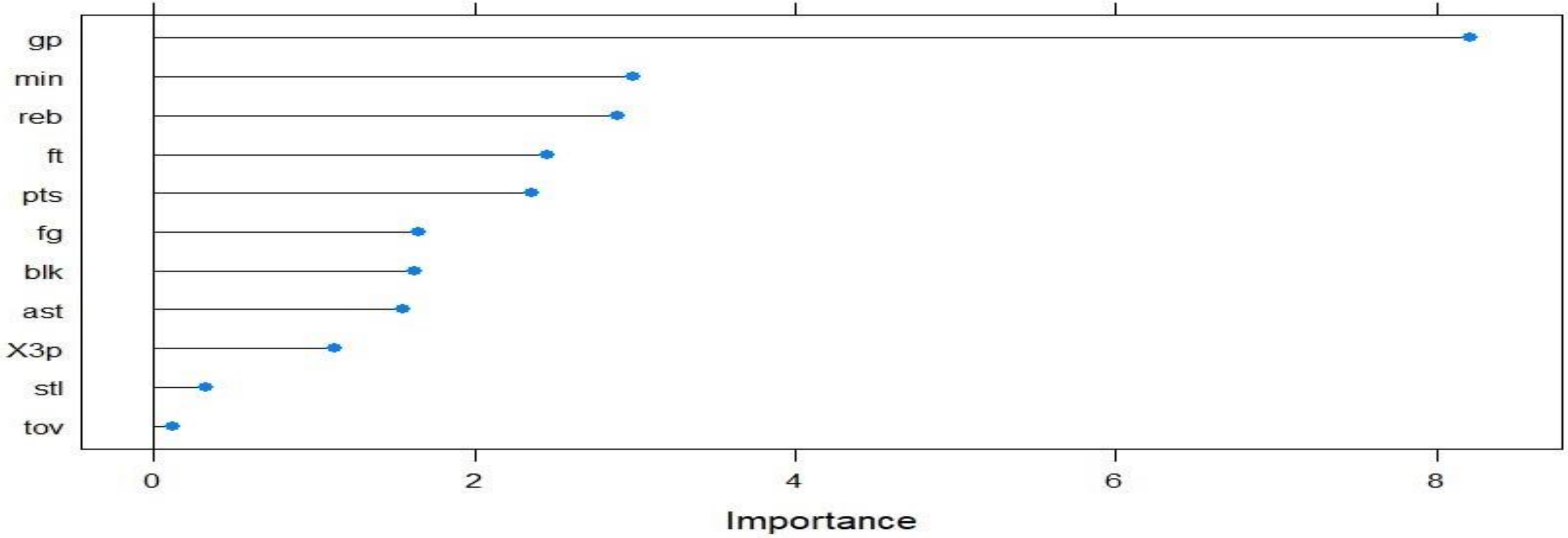
METHODOLOGY

Modeling Procedure

- Split data into training and test sets
- Run models using all predictors and 10-fold cross validation (if applicable)
- Calculate accuracy, false positive rate, and false negative rate for each model
- Compare models using metrics listed above

LOGISTIC REGRESSION

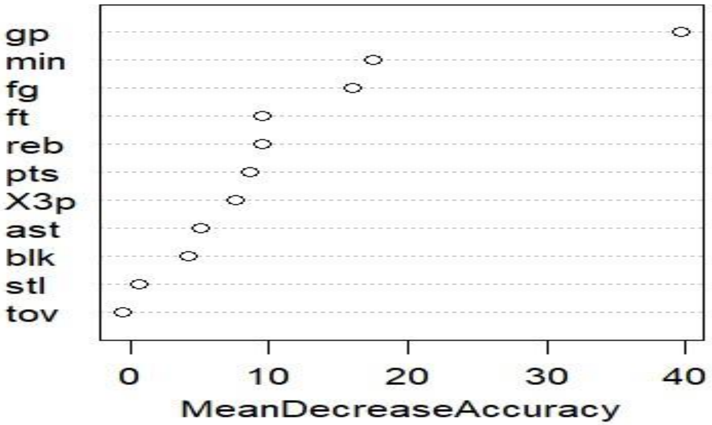
Accuracy Rate	False Positive Rate	False Negative Rate
72.9730%	45.0550%	17.2619%

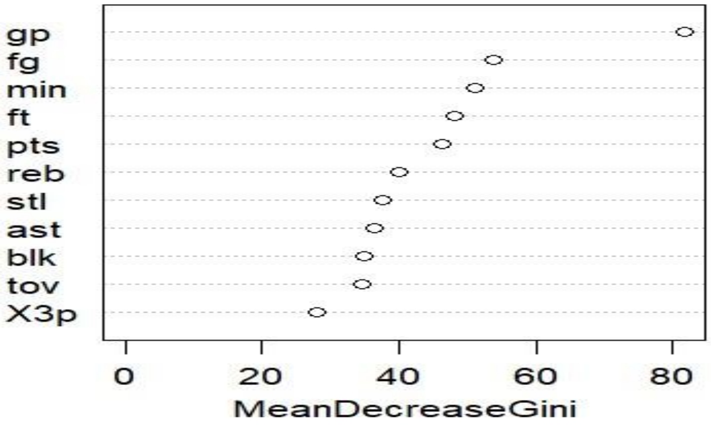


RANDOM FORESTS

Hyperparameter tuning: mtry = 5 and ntree = 400

Accuracy Rate	False Positive Rate	False Negative Rate
71.8147%	37.3626%	23.2143%

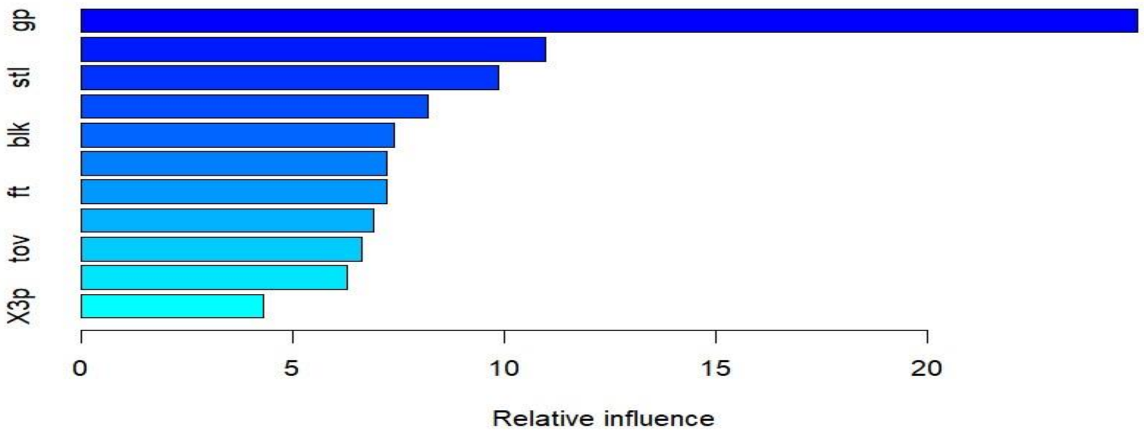


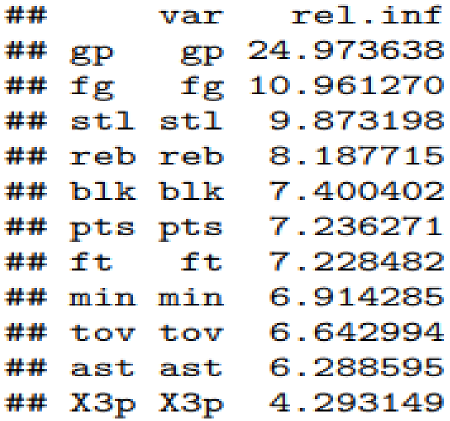


GRADIENT BOOSTING

Hyperparameter tuning: ntree = 400, shrinkage = 0.05, and interaction.depth = 4

Accuracy Rate	False Positive Rate	False Negative Rate
72.2008%	38.4615%	22.0238%





```
##       var   rel.inf
## gp     gp 24.973638
## fg     fg 10.961270
## stl   stl  9.873198
## reb   reb  8.187715
## blk   blk  7.400402
## pts   pts  7.236271
## ft     ft  7.228482
## min   min  6.914285
## tov   tov  6.642994
## ast   ast  6.288595
## X3p   X3p  4.293149
```

CONCLUSION

Summary

- Logistic regression is the “best” model
- All models identified *games played* is the most important predictor
- Random forests and gradient boosting models identified *field goals* as significant

Limitations

- Unbalanced response variable
- Computational power

Future Work

- Stratified sampling
- Logistic regression feature selection
- Raise classification threshold

THANK YOU FOR LISTENING TO OUR PRESENTATION!