PREDICTING

NBA ROOKIE LONGEVITY

Based on first-year stats

Sandra Tran
Metis 2022

Intro

According to SportsRec (2018), team rosters included players who had:

- 4 or fewer years: 52.1%
- > 4 years: 10.6%

Long-term career in NBA a long shot!

NBA rookie contract
1 to 4 years

Objective

- Predict whether an NBA player will play beyond 4 years based on their first-year stats

- Use classification algorithm with best F1 score
  - Harmonic balance between precision & recall

Data

NBA Official Website:

www.nba.com

- Regular Seasons
  - 1996-97 to 2021-22
- Player Stats
  - Name
  - Age
  - Games Played (Wins, Losses)
  - Minutes Played
  - Points made
  - FG%, 3P%, FT%
  - Etc.
- Over 12K data points
- Over 20 features

- **Filter only first-year stats for each player**
- **Add Target_4UP**
  - **0 = 4 or fewer years**
  - **1 = > 4 years**
- **Add Num_Yrs**
  - **Total # of years in NBA**

~2500 data points

Workflow

Partition Dataset

- Train/Test set
- Stratification
  - Equal proportion in both train/test
- Stratified KFold Cross-Validation
  - Equal proportion in each fold

Baseline Models

- kNN
- Logistic Regression
- Gaussian NB
- Decision Tree
- Random Forest
- XGBoost

Class Imbalance

- Random Oversampling on minority (target) class

Feature Engineering

- Log transformation of features
- Gaussian NG, Logistic Regression

F1 SCORE

Random Forest

Hyperparameter Tuning

- RandomizedSearchCV
- GridSearchCV

Feature Selection

- SelectFromModel
  - Selects features based on importance weights
- 11 out of 20 features

Baseline Scores

	kNN	Logistic	Gaussian	Decision Tree	Random Forest	XGBoost
Precision	0.715	0.735	0.745	0.576	0.718	0.683
Recall	0.584	0.632	0.562	0.592	0.621	0.631
F1	0.643	0.679	0.641	0.583	0.670	0.656

Resampling

	kNN	Logistic	Gaussian	Decision Tree	Random Forest	XGBoost
Precision	0.748	0.754	0.807	0.727	0.812	0.787
Recall	0.676	0.718	0.585	0.789	0.818	0.825
F1	0.710	0.735	0.678	0.756	0.818	0.805

Baseline Scores

	kNN	Logistic	Gaussian	Decision Tree	Random Forest	XGBoost
Precision	0.715	0.735	0.745	0.576	0.718	0.683
Recall	0.584	0.632	0.562	0.592	0.621	0.631
F1	0.643	0.679	0.641	0.583	0.670	0.656

Log Transform Features

	kNN	Logistic	Gaussian	Decision Tree	Random Forest	XGBoost
Precision	0.749	0.755	0.759	0.723	0.800	0.791
Recall	0.641	0.722	0.690	0.786	0.818	0.826
F1	0.691	0.738	0.723	0.753	0.814	0.808

Random Forest

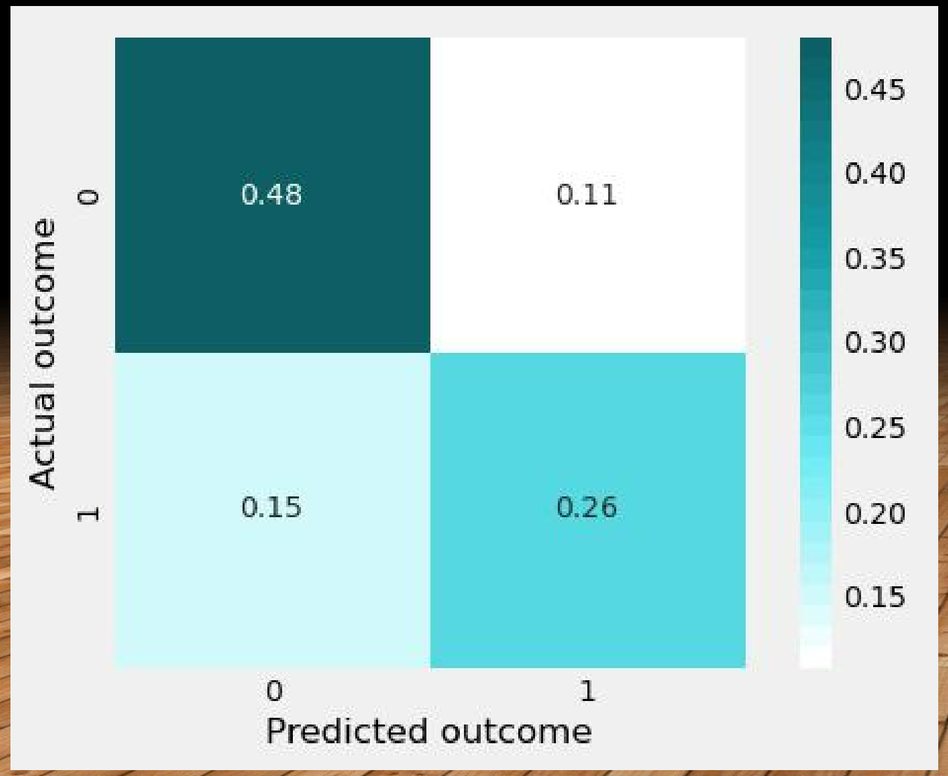
	F1 Score
Baseline	0.670
Resampling	0.818
Feature Transformation	0.814
Hyperparameter Tuning	0.810
Feature Selection	0.817

Random Forest
-Final Score-

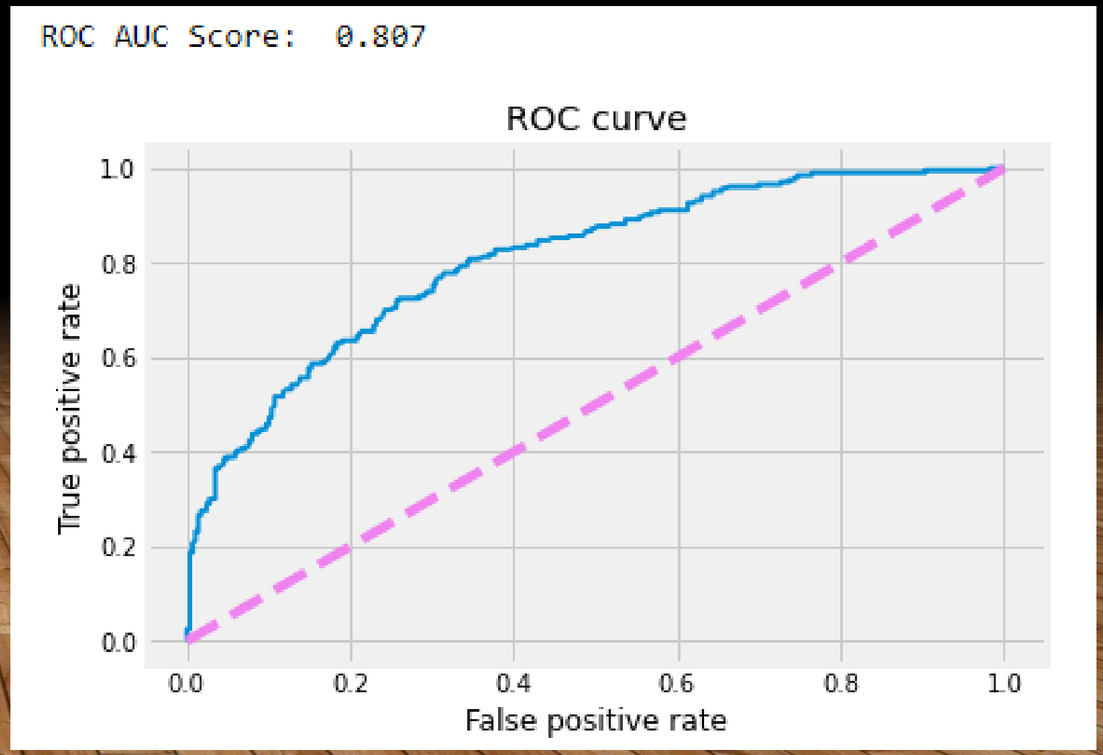
Train	
Accuracy	0.814
Precision	0.805
Recall	0.830
F1	0.817
ROC-AUC	0.890

Test	
Accuracy	0.728
Precision	0.673
Recall	0.660
F1	0.667
ROC-AUC	0.718

Confusion Matrix



ROC Curve



Feature Importance

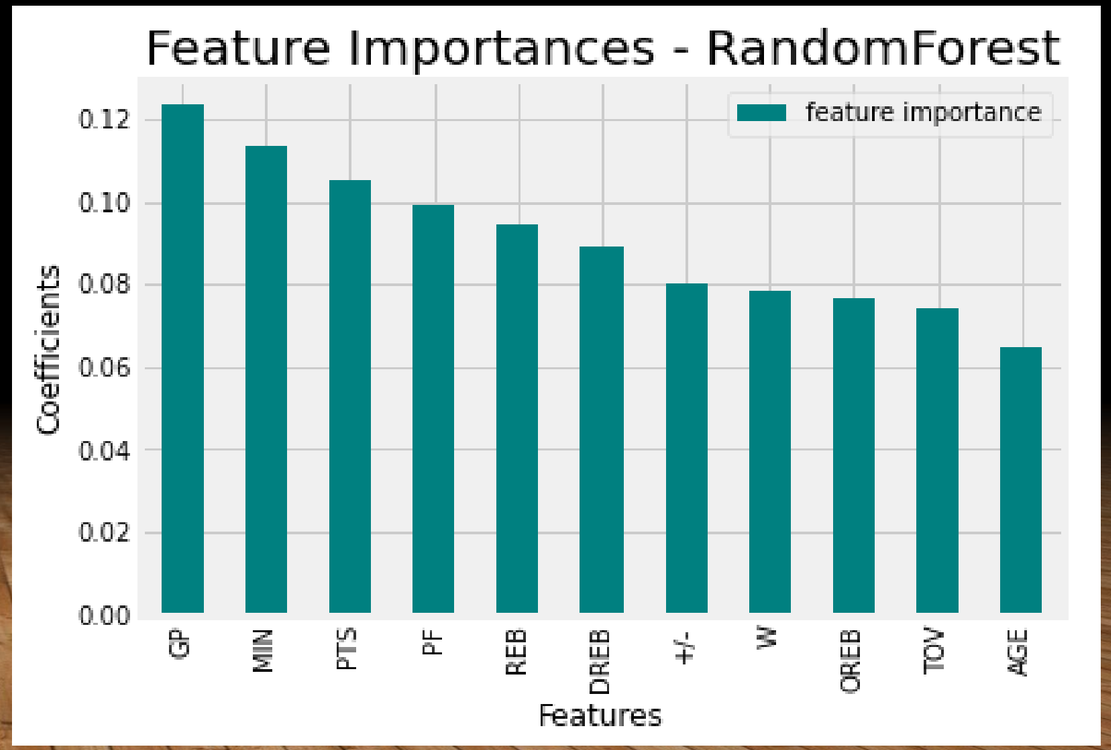
Top 3 Features

Games Played

Minutes Played

Total Points Scored

	feature importance
GP	0.123442
MIN	0.113761
PTS	0.105089
PF	0.099525
REB	0.094673
DREB	0.089062
+/-	0.080334
W	0.078269
OREB	0.076761
TOV	0.074122
AGE	0.064962



Future Work

- Filter dataset more
  - Remove "rookies" from '96-97 season
  - Explore whether active players <5 years contribute to model's inaccuracies
- Need bigger dataset
- Delve deeper into XGBoost
- Tune hyperparameters using GridSearch for all models and compare
- Create interactive web app

REFERENCES

1. Front page image: https://images.designtrends.com/wp-content/uploads/2015/12/21133823/Basketball-Background2.jpg
2. Page background: pdUdA3Q.jpg (2560×1600) (wallpapercave.com)
3. Contract image: 552542.jpg (960×720) (clipart-library.com
4. Basketball player silhouette with orange ball: sylhouettes-of-basketball-players-in-action-vector-id166008764 (609×612) (istockphoto.com)
5. Stephen Curry before and after: curry-titles.jpg (1200×900) (fadeawayworld.net)