FINAL PROJECT REPORT

Win Shares, NBA All Stars
CSC 532 2020
student id: chun1

Raymond Chun

ABSTRACT

Win Shares attempts to measure how much an individual player contributes to the wins of a team. This statistic has a lofty goal. When a team wins, it wants to assign credit to the player with the most coveted goal in sports: Wins. To predict Win Shares, this project uses several regression models and calculated the error of each one. This project will demonstrate the connection between traditional and advanced statistics with Win Shares. As an added goal, I included a measurement on how to predict future All Star selections from players' rookie season. Classification models were implemented in order to predict these All Stars.

PROJECT DEFINITION AND GOALS

As an advanced statistic, Win Shares shows that many of the most heralded basketball players in NBA history are at the top. Nearly all of the top 1% of Win Shares leaders are in the hall of fame or headed to it. Players such as Michael Jordan, Kareem Abdul-Jabbar, Lebron James, are all on this list. Predicting this statistic can be invaluable for player evaluation and development. Win Shares originated from baseball stats made by the godfather of Sabermetrics, Bill James, and was translated to basketball.

A key point of Win Shares: Win Shares is not a direct combination of other stats, it has its own formula. While many advanced stats are a combination of basic stats such as points, rebounds etc, Win Shares is slightly different. Team and league statistics are factored in and scaled in with offensive and defensive rating numbers as well. This is important because we are trying to predict the Win Share column from other variables.

As an added measurement, I worked on how to predict if a rookie player can be measured as an All Star caliber player. All Stars are annually voted in as one of the best 24 players in the NBA.

Goals:

- Predict and demonstrate the current season's selected Win Shares Statistic
- Predicting the NEXT year season's Win Shares Statistic
- Bonus Measurement: Predicting All Stars based off their first rookie season

RELATED WORKS

Basketball Analytics: Predicting Win Shares
https://towardsdatascience.com/basketball-analytics-predicting-win-shares-7c155651e7cc
Summary: Using 2016-17 and 2017-2018 NBA Season to predict Win Shares in Python. Only focuses on top players.

Rookie NBA Data Analysis
https://github.com/rykwan/rookieNBA-data-analysis-R
Summary: Using logistic regression trying to predict All-Star potential for 2017

Without having to be said, neither works were ever viewed before the project began. Upon nearly completing the final project, did I even search for related works. By all means, the reader may examine the works cited to compare and contrast if need be.

PRE-PROCESSING

There went a lot of attempts to study and understand the date before actually implementing the models.

1. Correlations: I studied the positively correlated variables for the NBA, Win Shares, and the next years' Win Shares very carefully.
2. Lasso: Since Lasso has a built in variable selection process, that was also considered carefully.
3. For this particular basketball data set, there were some historical narratives that were important to consider. In the 1950's, pro basketball was in its infancy. Many statistics were never considered or measured. In 1977, the NBA merged with the ABA league for financial and practicality reasons. Furthermore, every year there is a 'competition committee' that convenes and decides on basic rule changes to the game. Comparing 1977 to 2020 is not realistic. For section 2 and 3 of my project, I filtered out the NBA from 1999 to 2019, using a long list of player data but also using 1999 as a starting point for modern basketball. This is further explained in my PDF presentation file.
4. For section 3, I created a CSV file of NBA All Stars from basketballreference.com and wikipedia in order to merge it with my main NBA data set. This was used to predict All Stars from their rookie seasons in Section 3 of my project.
5. Skewness: An important detail to note is that many, many NBA players do not contribute at all to playing time or statistics. Histogram analysis will show how skewed this data set really is. Many players are often signed temporarily as backups. There were thousands of entries who never played or played very little. I increased the minutes and games played threshold to adjust for this. The NBA is truly a league of outliers.

RESULTS

Section 1: Predicting Current Season Win Shares

LINEAR REGRESSION

```
predictionsLR=
predict(linearRegression,
NBA_test)
RMSE(predictionsLR,
NBA_test$WinShares)
R2(predictionsLR,
NBA_test$WinShares)
```

RMSE: 0.5852
R2: 0.9618

RANDOM FOREST

```
predictionsRF <- predict(rf,
NBA_test)
RMSE(predictionsRF,
NBA_test$WinShares)
R2(predictionsRF,
NBA_test$WinShares)
```

RMSE: 0.6065193
R2: 0.9596143

LASSO

```
coef(lasso$finalModel,
lasso$bestTune$lambda)
predictionsL <- predict(lasso,
NBA_test)
RMSE(predictionsL,
NBA_test$WinShares)
R2(predictionsL,
NBA_test$WinShares)
```

RMSE: 0.6153303
R2: 0.9581337

Section 2: Predicting the NEXT Season's Win Shares

LASSO

```
coef(lasso$finalModel,
lasso$bestTune$lambda)
predictionsL <- predict(lasso,
NBA_test)
RMSE(predictionsL,
NBA_test$WinShares_Next_Year)
R2(predictionsL,
NBA_test$WinShares_Next_Year)
```

RMSE: 2.172565
R2: 0.5673374

LINEAR REGRESSION

```
#Checking the RMSE of the
WinShares Next Year
predictionsLR=
predict(linearRegression, NBA_test)
RMSE(predictionsLR,
NBA_test$WinShares_Next_Year)
R2(predictionsLR,
NBA_test$WinShares_Next_Year)
```

RMSE: 2.140996
R2: 0.6062071

RANDOM FOREST

```
rf<- train(WinShares_Next_Year ~ ., data =
NBA_train, importance=T, method = "rf", trControl =
ctrl, tuneGrid = grid_rf)
varImp(rf)
predictionsRF <- predict(rf, NBA_test)
RMSE(predictionsRF,
NBA_test$WinShares_Next_Year)
R2(predictionsRF,
NBA_test$WinShares_Next_Year)
```

RMSE: 2.214575
R2: 0.5792119

RIDGE

```
predictionsRidge <-
predict(ridge,NBA_test)
RMSE(predictionsRidge,
NBA_test$WinShares_Next_Year
)
R2(predictionsRidge,
NBA_test$WinShares_Next_Yea
r)
```

RMSE: 2.160676
R2: 0.599431

RESULTS CONT'D

Section 2: Predicting the NEXT Season's Win Shares

ELASTIC NET

```
predictionsElasticNet <- predict(enet, NBA_test)
RMSE(predictionsElasticNet,
NBA_test$WinShares_Next_Year)
R2(predictionsElasticNet,
NBA_test$WinShares_Next_Year)
```

RMSE: 2.146052
R2: 0.6046081

GBM

```
lpredictionsRF <- predict(rf, NBA_test)
RMSE(predictionsRF, NBA_test$WinShares)
R2(predictionsRF, NBA_test$WinShares)
```

RMSE: 2.20075
R2: 0.5833839

ANN

```
library(keras)
rmse= function(x,y){
return((mean((x - y)^2))^0.5)
}
rmse(predictions, NBA_testy)
R2(predictions, NBA_testy)
```

Best Numbers

RMSE: 2.020652
R2: 0.6382309

Section 3: Predicting All Stars from their Rookie Seasons

LASSO

```
table(AllStarPred_test$AllStar)
predict.lasso = predict(lasso, AllStarPred_test)
confusionMatrix(predict.lasso, AllStarPred_test$AllStar)

lasso_predictions_prob=predict(lasso, AllStarPred_test,
type="prob")
head(lasso_predictions_prob)

pred_lasso = prediction(lasso_predictions_prob$`1`,
AllStarPred_test$AllStar)
performance(pred_lasso, measure = "auc")@y.values
perf <- performance(pred_lasso, measure = "tpr", x.measure =
"fpr")

#Plotting the ROC Curve
plot(perf, col = "blue")
```

AUC: 0.8371736
Accuracy : 0.94

RIDGE

```
ridge_predictions_prob=predict(ridge, AllStarPred_test,
type="prob")

pred_ridge = prediction(ridge_predictions_prob$`1`,
AllStarPred_test$AllStar)
performance(pred_ridge, measure = "auc")@y.values

perfR <- performance(pred_ridge, measure = "tpr",
x.measure = "fpr")

#Plotting the ROC Curve
plot(perfR, col = "green")
```

AUC: 0.8379416
Accuracy : 0.935

RESULTS CONT'D

Section 3: Predicting All Stars from their Rookie Seasons

ENET

```
enet_predictions_prob=predict(enet,
AllStarPred_test, type="prob")

pred_enet = prediction(enet_predictions_prob$`1`,
AllStarPred_test$AllStar)
performance(pred_enet, measure =
"auc")@y.values


perfE <- performance(pred_enet, measure = "tpr",
x.measure = "fpr")

#Plotting the ROC Curve
plot(perfE, col = "purple")
```

AUC: 0.8379416
Accuracy : 0.935

LOGISTIC

```
library(pROC)

predict.logistic <- predict(model.logistic,
AllStarPred_test, type="response")
predict.logistic.label =
factor(ifelse(predict.logistic > .1, "Yes", "No"))
actual.label = AllStarPred_test$AllStar
table(actual.label, predict.logistic.label)
ROC <- roc(AllStarPred_test$AllStar,
predict.logistic)

#Plotting the ROC Curve
ROCplot = plot(ROC, col = "red")

#AUC= The area under the curve
auc(ROC)
```

AUC: 0.8015
Accuracy: 0.865

*Full codes are in the R notebook and html preview

CONCLUSIONS

The Main Conclusions:
Based off the error and accuracy calculations, we can reasonably conclude that the models for predicting Win Shares for the next season and predicting All Stars from their rookie seasons, **both had acceptable-to-good performances**. Predicting Win Shares was based off regression and predicting All Stars was based off of classification. Predicting All Stars preformed better overall, but had a smaller sample size.

Interesting details I learned:
The NBA is truly a league of elite outliers. It appears that the top 60-80 players matter so much more than everybody else. Considering the league can have upwards of 400 players at any season, it is a very top heavy league.

In the past, data in sports was very clean and simple. Not so much anymore. Only recently have sports attempted to add more advanced metrics to their tools. Personally, I can't tell if the analytics are complicating sports, or if they are simplifying them. It's hard to tell with sports sometimes.

The best part of the project for me was applying our skills to a field of our choice. One of the most enjoyable things about data is that it can reach almost any field or industry, seeking out detail and nuance in interesting ways. This project allowed us to do so and in a creative way.

-Raymond Chun

REFERENCES

Data set from kaggle.com: https://www.kaggle.com/lancharro5/seasons-stats-50-19

Win Shares: https://www.basketball-reference.com/about/ws.html

Basketball Information: https://www.basketball-reference.com/

List of All Stars: https://en.wikipedia.org/wiki/List_of_NBA_All-Stars