Milestone 5: Individual Summary Analysis

Sachit Patel

In this individual analysis, I go over my individual contributions to the group towards our dataset during this step, as well as the larger picture of group work on our subject dataset. Our group is analyzing the 2022-2023 NBA Player Stats dataset. We decided to split into pairs to help tackle the multiple lines of research the project asks of us. Victor and I looked into building models that can help measure player performance.

Victor attempted an initial OLS model during the exploratory analysis with the dataset. However, it has a concerning issue. The OLS full model is highly susceptible to overfitting. Most of the beta values are extremely low (values with a scientific notation). Additionally, the r-squared is close to 1 (about .9952). These two pieces of evidence suggest the OLS model is highly susceptible to overfitting. Additionally, with the high amount of correlation shown between predictors within the corrplot, multicollinearity is a present issue within the model (this may be responsible for the inflated beta values). In order to utilize a relaxed lasso effectively, we must first fix the multicollinearity present in the model.

I found that a lot of the multicollinearity lies within “partial” variables within the dataset. What I mean by this is that there are some variables that are “pieces” of another. For example, there are 3 variables for 2 points. 2 points shots scored, 2 points shot attempted, and 2 point shot success rate percentage. This exists for 3 points, total field goal percent (as well as effective field goal%), free throws, and so on. Removing a lot of these overlapping variables was the key to removing the multicollinearity from the dataset, which allowed for a stronger relaxed lasso.

This is a point where the analysis step can divide a bit. What variables we choose to take out of the model for multicollinearity can influence the logic of what the model can predict. Since we are modeling for points per game, we would want to take out statistics that have a minimal effect on that goal, whilst reducing multicollinearity. Additionally, Victor identified through the NBA Stats Glossary how the percent statistics may not be a ratio of successful to failed attempts. Rather, it could be about a player’s contribution to that team of that specific stat, i.e one player may account for 3% of all 2-point goals scored from a team. If we remove the other two 2-point variables, then we’re removing the variables that

identify an individual's successful and failed scores. Through some further analysis in the corrplot, we found that some predictors had similar correlation profiles. There are a couple of obvious examples: the multiple predictors for 2 & 3 points (x2p with x2pa and x3p with x3pa). One interesting case was between offensive rebounds and blocks. These two had similar correlation profiles! Refer to figure 1.1 for which variables were left over, and the Kaggle dataset for the variable definitions.

With these changes, multicollinearity has been addressed. Running a second OLS shows that the adjusted r^2 no longer suggests overfitting, but the betas are still somewhat inflated. This is where a relaxed lasso can enter the picture as an approach. After trying several runs with cross-validated glmnet, different combinations of predictors and lambda 1se's were given. A specific output was settled on by using the lambda 1se value of 0.04653347 on the training split of the dataset. As for overall statistics, this gave an R^2 of around 93%, with an RMSE of roughly 0.18. The specific predictors and plot can be found in the appendix (figure 1.2 and figure 1.4). Additionally, the gamma for 1se is 0.00, which indicates that overfitting is not a problem. The coefficients of this fit after LASSO's variable selection were games played (negligible effect, could be removed), 3-point attempts, 2-point attempts, free throws, and defensive rebounds. These discovered outputs help build a groundwork for later factor analyses and discriminant analyses, which other members of the group are doing. My role in this project as it played out was to address the multicollinearity and to address the approach of building a model for points per game for individual players.

A few latent insights have been discovered through the coefficients of relaxed lasso. We know that points per game will be directly tied to the mechanisms that garner points (scoring field goals).Lambda is the measure of the penalty. As lambda increases, so does the inherent bias in the model. In these cases shown by the lasso regression, it took only a little addition of lambda to remove several predictors. Selection was quickly done with an extremely small lambda, drowning noise. 2 points, 3 points, free throws, and defensive rebounds are the mechanisms behind points per game. U The closest I could come to from interpreting this model is that defensive rebounds are surprisingly valuable for scoring points. But this just seems common sense in basketball, as it is a way for a team to establish control over the ball. Keep in mind that we used the attempt variables for two points and 3 points, not the actual points scored. This is probably

because the ball is in the hands of their most trusted players to score, or in the positions that score. The combination of predictors themselves are the types of scores that can be made, as well as the most important mechanism to upstart scoring points- defensive rebounds.

There are a couple lessons I took away from this project. For one. I learned how to do relaxed lasso specifically for this project from the supplemental lecture. I learned some of the intuition and process in group-work, particularly through Victor's emphasis on direction and oversight of if something is truly an insight or not. Some smaller details I learned include specific coding bits from Alaa and Victor, group communication and the importance of multimedia in sharing findings (messaging simply isn't enough). For now, it feels as if I have a smaller role in this group, but I hope I am able to contribute more by the end.

Appendix

Figure 1.1: VIF output after removing all desired variables.

```
> vif(fit1)
       Age          G         GS       X3PA       X2PA         FT        DRB
  1.148365   2.153664   3.852235   2.033714   5.669877   4.073425   4.928944
       AST        STL        BLK         PF Position_dummy
  4.646091   2.407818   2.225933   2.996943   2.796241
> |
```

Figure 1.2: Sample Cross-Validated Relaxed Lasso plot output.

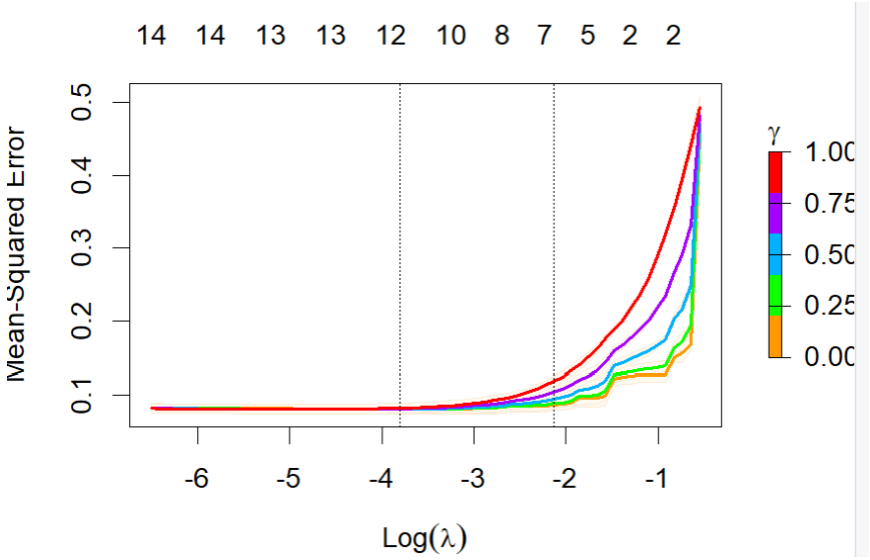

Figure 1.3: GLMNET call shown to list %Dev and %Dev R.

```
Call:  glmnet(x = xTrain, y = yTrain, lambda = 0.0322274, relax = T,      data = dsTrain)
Relaxed

  Df  %Dev %Dev R  Lambda
1 11 83.87   84.6 0.03223
```

Figure 1.4: Coefficient output for Relaxed Lasso with the set lambda1se.

```
> coef(fitLasso2, s="lambda.1se")
13 x 1 sparse Matrix of class "dgCMatrix"
                            s1
(Intercept)    0.5363707794
Age            .
G              0.0009251931
GS             .
X3PA           0.3010356751
X2PA           0.4751484473
FT             0.2707712163
DRB            0.2206258663
AST            .
STL            .
BLK            .
PF             .
Position_dummy .
```

Works Cited

Haefner, J. (n.d.). *What is effective field goal percentage? and why you should use it*. Welcome to BREAKTHROUGH BASKETBALL. https://www.breakthroughbasketball.com/stats/effective-field-goal-percentage.html#:~:text=Effective%20Field%20Goal%20Percentage%20is,shots%20are%20given%20extra%20weight

Stat Glossary | Stats | NBA.com. (n.d.). https://www.nba.com/stats/help/glossary

Vinco, V. (2023, May 14). *2022-2023 NBA player stats*. Kaggle. https://www.kaggle.com/datasets/vivovinco/20222023-nba-player-stats-regula