Predicting the performance of prospective Division 2 basketball players at a Division 1 level

Sarah Borsotto,* Nick Bettencourt,* Sean Chan,* Juan Castro,* Mingjun Sun,* and Xuan Jiang*

E-mail: sborsott@ucsd.edu; nbettencourt@ucsd.edu; swchan@ucsd.edu; jdcastro@ucsd.edu; mis005@ucsd.edu; xuj001@ucsd.edu

Abstract

In this project, the problem we are addressing revolves around assessing the potential transferability of basketball players to the Division 1 UCSD basketball team. Successfully recruiting basketball players for a college team can be very challenging. Evaluating a player's skill set, such as shooting, offense, defense, and teamwork, requires comprehensive knowledge and experience from the recruiters. Limited scouting resources, uncertainties regarding long-term development, and competition from other colleges make the recruitment process highly competitive and prone to errors. By leveraging past statistics of both the players under consideration and those who have successfully transitioned to D1 teams, we aim to develop a predictive model that can be used to ease these concerns. This model will help evaluate the suitability of prospective transfers based on their performance metrics, thereby aiding in the selection process and ensuring that the chosen players are a good fit for the team.

1. Source of the data

1.1 Web Scraping

Acquiring data for Division 1 basketball teams was straightforward. We acquired CSVs from barttorvik.com for the years 2018 to 2023. In addition, we acquired a XLSX file from ncsasports.org on the Division 1 schools and their respective basketball conferences. On the other hand, acquiring data for Division 2 Basketball teams proved to be more complex as we needed to scrape data from individual D2 team websites. The main problem we had

to solve was accounting for the different urls and different formats in which each website presented their players' statistics. Using the Selenium and BeautifulSoup packages for the web scraping effort, we were able to obtain 314 Division 2 schools that we compiled into one CSV file (d2_basketball_schools.csv)

Once we had obtained the list of Division 2 schools, we then needed to obtain each school's team roster, which consisted of all the player's individual statistics. However, each school in the list did not utilize the same notation when labelling player positions. For example, some schools would use 'F', 'small', or 'power' to label the position of a player that was a 'Forward'. To ensure this would not result in many variations of positions when performing our analyses, we cleaned this data to include only three main positions in basketball: 'forward', 'guard', 'center'; other positions were labeled to 'other'. We also noted that some lines of our data contained only zeros or NaN values, which we removed from our dataset. Once data cleaning was completed, we were left with a dataset with 13028 entries, which was sufficient for our studies.

Table 1: Description of Basketball School/Player Datasets

Files	**Description**
division1schools.xlsx	XLSX file of Division 1 basketball schools.
d1data.csv	CSV file of Division 1 basketball players' stats.
d2_basketball_schools.csv	CSV file of Division 2 basketball schools.
data.csv	CSV file of Division 2 basketball players' stats.

Table 2: Description of Division 1 Basketball Schools Dataset

Column Name	**Description**
Conference	The athletic conference in which the school competes.
Division	The NCAA division level of the school. In this dataset, all schools are in Division 1.
School	The name of the school.

Table 3: Description of Division 2 Basketball Schools Dataset

Column Name	**Description**
Name	The name of the school.
Conference	The athletic conference in which the school competes.
Division	The NCAA division level of the school (Division 2).
Reclass Division	Information on whether the school is reclassifying to a different division.
Public/Private	Whether the school is a public or private institution.
HBCU	Indicates if the school is a Historically Black College or University.
State	The U.S. state in which the school is located.

1.2 Description of the data

Our data includes statistics on basketball player performance at Division 1 and Division 2 levels. We ensured that each of datasets had the same columns to prevent any merge conflicts. Each row represents an observation for the performance of one player during a game. The quantitative metrics are GP, GS, MIN, MIN/G, FGM, FGA, FG%, 3PT, 3PTA, 3PT%, FT, FTA, FT%, PTS, AVG, OFF REB, DEF REB, REB, REB/G, PF, AST, TO, STL, BLK, and Year. The categorical metrics are Position and Team. A summary of these variables is provided below:

Table 4: Description of Division 1 and 2 Basketball Player Stats Dataset

Column Name	**Description**
GP	Games Played.
GS	Games Started.
MIN	Minutes Played.
FGM	Field Goals Made.
FGA	Field Goals Attempted.
FG%	Field Goal Percentage.
FTM	Free Throws Made.
FTA	Free Throw Attempts.
FT%	Free Throw Percentage.
3PT	Three-Point Field Goals Made.
3PTA	Three-Point Field Goals Attempted.
3PT%	Three-Point Field Goal Percentage.
REB	Rebounds.
OFF REB	Offensive Rebounds.
DEF REB	Defensive Rebounds.
PF	Personal Fouls.
AST	Assists.
STL	Steals.
BLK	Blocks.
TO	Turnovers.
Position	Basketball Position.
Team	College Team Affiliation.
Year	The season or year in which games were played.

2. Exploratory data analysis

2.1 Data Cleaning

For the Division 1 and 2 basketball player stats datasets, we handled missing data by dropping rows in which more than half of the columns are zero. Having many zeros indicates that a player has information missing or have played a scarce amount of games, thus are outliers for the dataset.

2.2 Figures

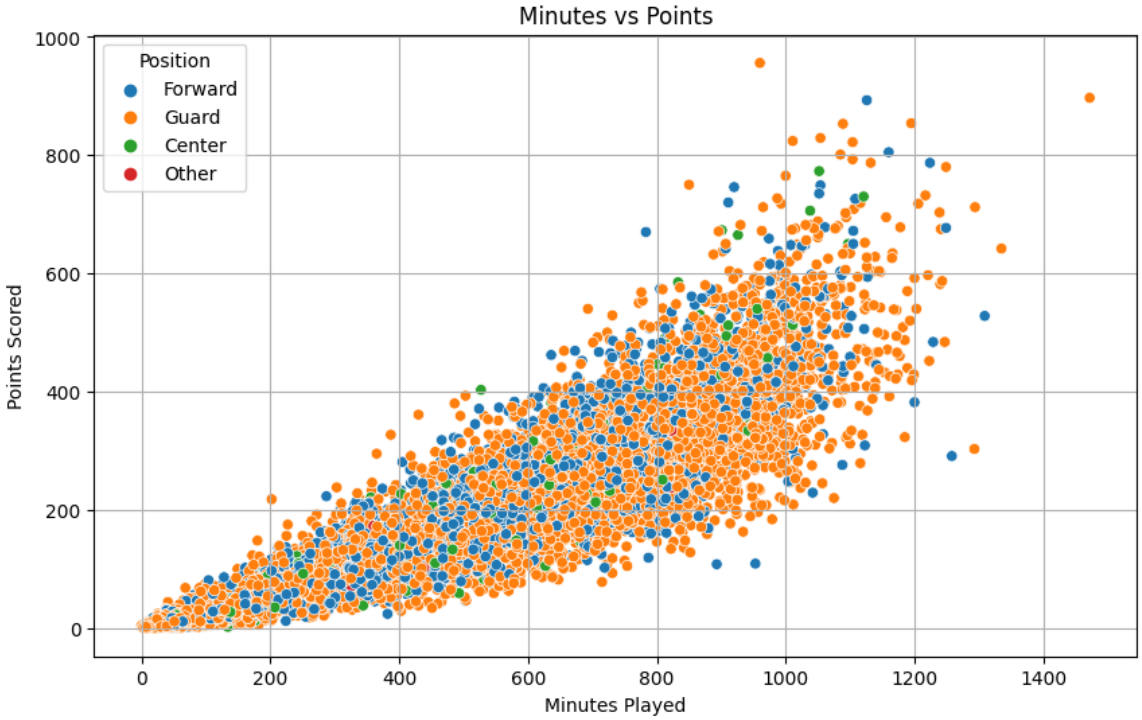

Figure 1: Analyzing how many points were scored by each position as playtime changed.

Figure 1 demonstrates the points that were scored over minutes played for each basketball position. As seen in the figure, there seemed to be more points scored by Guards in the overall minutes played compared to Forward and Center players.

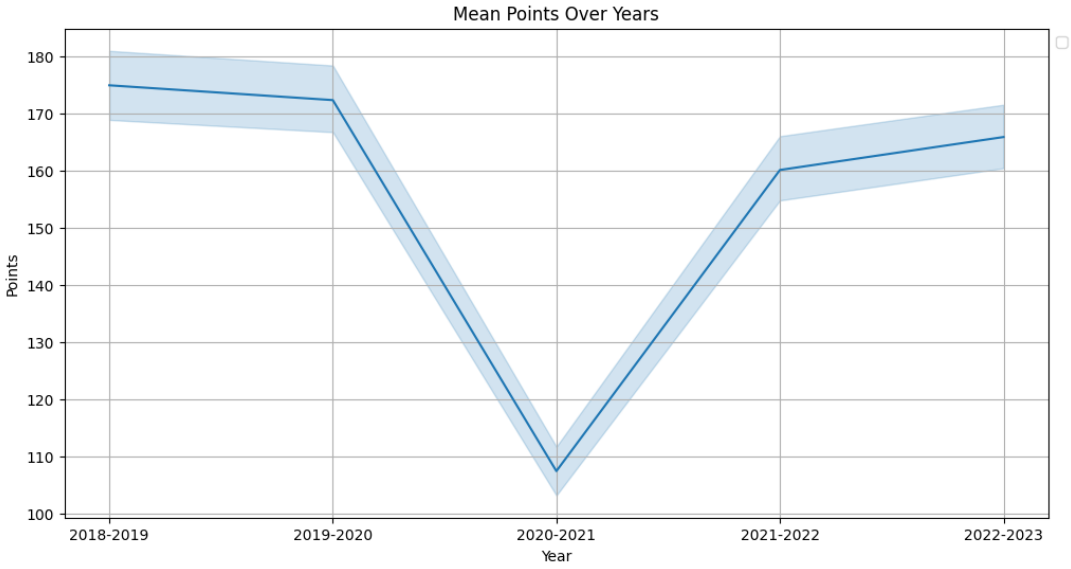

Figure 2: Analyzing average points per year.

As we see in Figure 2, there seems to be an even distribution of years in our dataset. The number of observations for each year appear to be about the same. The dip that is seen in the figure for the number of points in 2020-2021 may be correlated with some other factor, such as the COVID-19 pandemic.

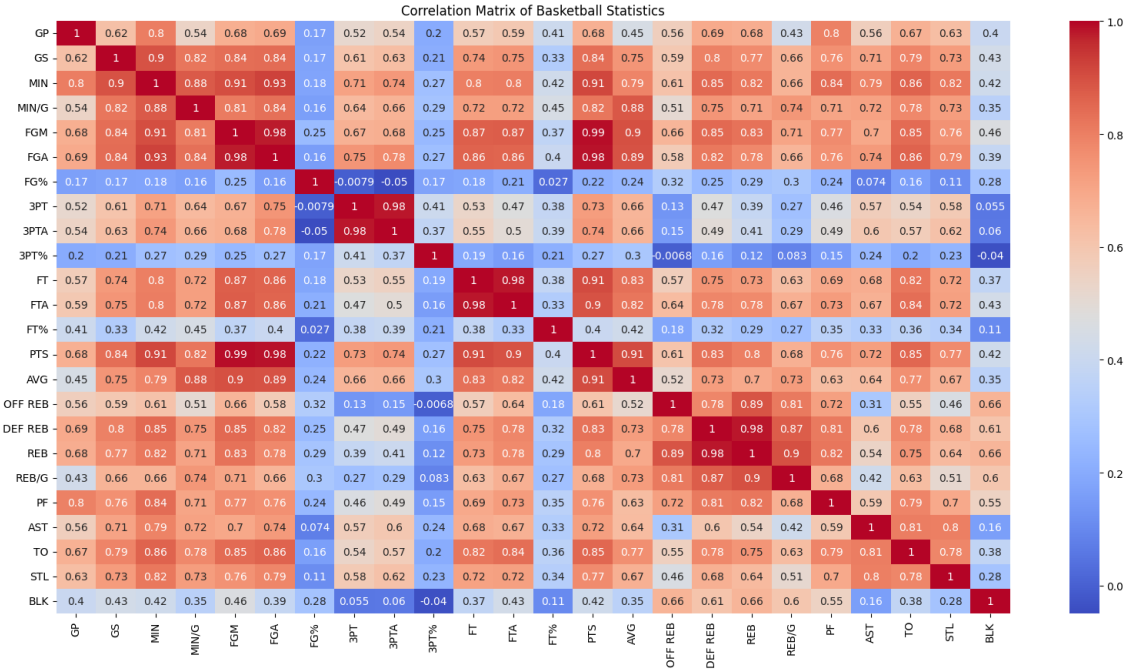

Figure 3: Correlation Matrix of Basketball Statistics

In Figure 3, there seems to be a large variability in the strength of the correlations. One correlation that appears to be relatively high is PTS and FGA. Let's graph these variables against each other to see what their relationship looks like.

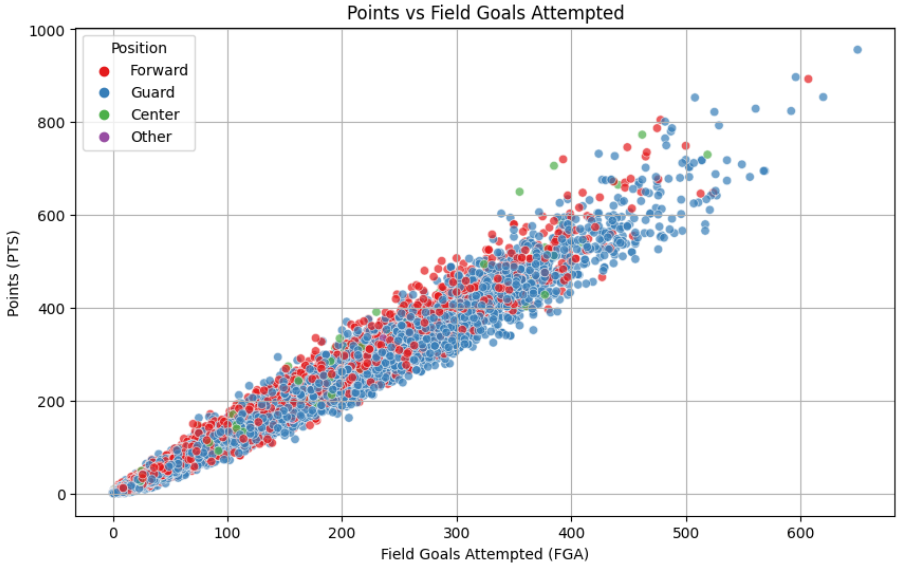

Figure 4: Analyzing the relationship between Points per FGA

In Figure 4, The relationship is strongly linear, with points increasing as FGA increases. This demonstrates the players' skills; they were able to obtain points for most of the field goal attempts they made. Some other relationships that are commonly explored in basketball include FG% (Field Goals Percentage) and PTS (Points), AST (Assists) and TO (Turnovers), MIN (Minutes Played) and Rebounds (REB), Three-Point Attempts (3PTA) and Three-Point Percentage (3PT%), and Offensive Rebounds (OFF REB) and Defensive Rebounds (DEF REB). These relationships can provide insight into the performances of the players and the team.

Refer to our Jupyter Notebook for the following analyses: "Field Goal Percentage vs. Points, "Assists vs. Turnovers", "Minutes Played vs. Rebounds", "Three-Point Attempts vs. Three-Point Percentage", and "Offensive Rebounds vs. Defensive Rebounds".

2.3 What (if any) analyses have already been performed on this data (or another similar dataset)?

For our particular dataset, there have been no previous studies conducted on this data since we sourced our data from web scraping. However, one previous study that was similar to our study was "Predicting NBA Player Performance" by Kevin Wheeler. In this study, the goal was to be able to predict NBA player performance using machine learning techniques. The specific player performance this study was predicting was the number of points a player scored against an opponent. This was similar to our study when using player statistics to make predictions, though their study differed in what the statistics were used to predict. We initially built on the ideas and methods of Wheeler's study, which included linear regression, Naïve Bayes, and support vector machine (SVM) models, in order to conduct our own analysis on our data.

3. Analyses performed for the project

Below we introduce our process for variable selection for our models, as well why we went from a multivariate linear regression model to a neural network.

3.1 Methodology

3.1.1 Similarity Function

In order to obtain accurate predictions for a player's predicted stats at a Division 1 level, we decided to find the most similar players that transferred from Division 2 to Division 1 schools to a specific player at a Division 2 school. The method we decided to use for this task was Cosine Similarity, as our data per player had mostly integer values that Cosine Similarity could handle well.

$$Cosine(x, y) = \frac{x \cdot y}{|x||y|}$$

By finding former Division 2 players with similar Division 2 statistics to the prospective player, we can use the similar players' known stats from the Division 2 level to predict the future stats of the specific player. We also ensured that the player's stats were normalized when running the Cosine similarity function so that we would not encounter any extreme outliers when doing comparisons between players. Once we had this function finalized, we utilized this in our multivariate regression model.

3.1.2 Feature Selection/Classification

For our initial selection of variables, we chose to include all columns in our dataset, which was a total of 23 features. However, we quickly realized that this led to very inaccurate results. Using the exploratory data analysis we performed, we saw a general overview of all our player statistics and their correlation with other stats from our correlation matrix. To select the features best suited for our dataset, we decided to use a Random Forest Classification to select the top 10 most correlated player statistics from the correlation matrix, which were: FGM, FGA, MIN, AVG, FT, FTA, TO, GS, DEF REB, and MIN/G. However, once we had finalized the model, our highest resulting accuracy for predicting player performance was 13.9%. This was not the accuracy we had hoped for, which resulted in us changing our predictive model.

3.1.3 EDA for Neural Network

After viewing the results from the feature selection, we redirected our approach to the predictions with a neural network model. To compare the performance of different players from different divisions, we assigned conference score/grade values to account for the different average skill ratings. The conference score values for Division 2 schools ranged from [0.75, 1.1], while those for Division 1 schools ranged from [1, 2]. For each Division 1 and Division 2 basketball team, we assign each of the players their conference grades and we combine the D1 and D2 player datasets. In addition to the conference score, we calculated the per-game statistic of each player because it would be a more accurate representation of their performance:

Table 5: Per-game Performance Metrics for College Basketball Players

Metric	**Description**
PPG	Points Per Game
TO/G	Turnovers Per Game
PF/G	Personal Fouls Per Game
STL/G	Steals Per Game
BLK/G	Blocks Per Game
OFF_REB/G	Offensive Rebounds Per Game
DEF_REB/G	Defensive Rebounds Per Game
AST/G	Assists Per Game

The last feature we added to our data is the "Occurrence" column which tracks the number of years a player has been with a team. As a player stays on the same team, it can lead to an increase or decrease in their performance.

3.1.4 Neural Network Training

We use multi-layer perceptron (MLP) model architecture because it allows us to build a neural network and map the connection between Division 2 player stats(input) to Division 1 player stats(output).

Before we start training our neural network models, we select the numerical columns we want to predict and select the features we want to input into our model.

Table 6: Features in the Dataset

Feature
Player
GP
GS
MIN/G
FG%
3PT%
FT%
PPG
OFF_REB/G
DEF_REB/G
REB/G
AST/G
TO/G
PF/G
STL/G
BLK/G
Position
Team
Conference
Conference_Grade
Occurrence

Table 7: Numerical Columns

Numerical Output Predictions
GP
GS
MIN/G
FG%
3PT%
FT%
PPG
REB/G
OFF_REB/G
DEF_REB/G
PF/G
AST/G
TO/G
STL/G
BLK/G

After we choose the features and numerical columns, we preprocess our data: filtering rows with missing data, handling categorical data using one-hot encoding, and standardizing the numerical features. Our dataset is then split into training and testing set where the training set is all seasons before "2022-2023" and the testing set is the "2022-2023" season. Using the preprocessed data, we convert them into PyTorch Tensors so that they can be used for our neural network.

For each target column, we create a MLP class to transform our inputs. The loss function criterion to evaluate our model's performance is mean squared error loss and the optimizer is the Adam optimizer. For each model we train, we report the testing and training error and save the model in a dictionary with the column name as its key.

3.1.5 Inference Augmentation

On run time, we made a few adjustments to ensure we were getting the most out of our model(s). Our Neural Network takes a many-to-one mapping of inputs to output, wherein a metric like 'GP' (Games Played) is predicted only by knowing all other features of a player (i.e. 'GS', 'MIN/G', 'FG%', '3PT%', etc. goes into the input, with 'GP' being the output - this being the case for every predicted variable). Upon inference, there are only five features we know beforehand: 'Player', 'Position', 'Team', 'Conference', and 'Conference_Grade' - meaning we had to find someway of determing these other features to make our prediction. Our solution was to get the top 150 most similar players to the previous year of a player's data (cosine similarity), filter that by players which had a subsequent year of data, get the median of each feature from this filtered group of players' data from the subsequent year upon inference (make-shift way of getting a rough estimate of the predicting player's other stats), and using these estimates to feed into the model(s).

3.2 Interpretation of the results

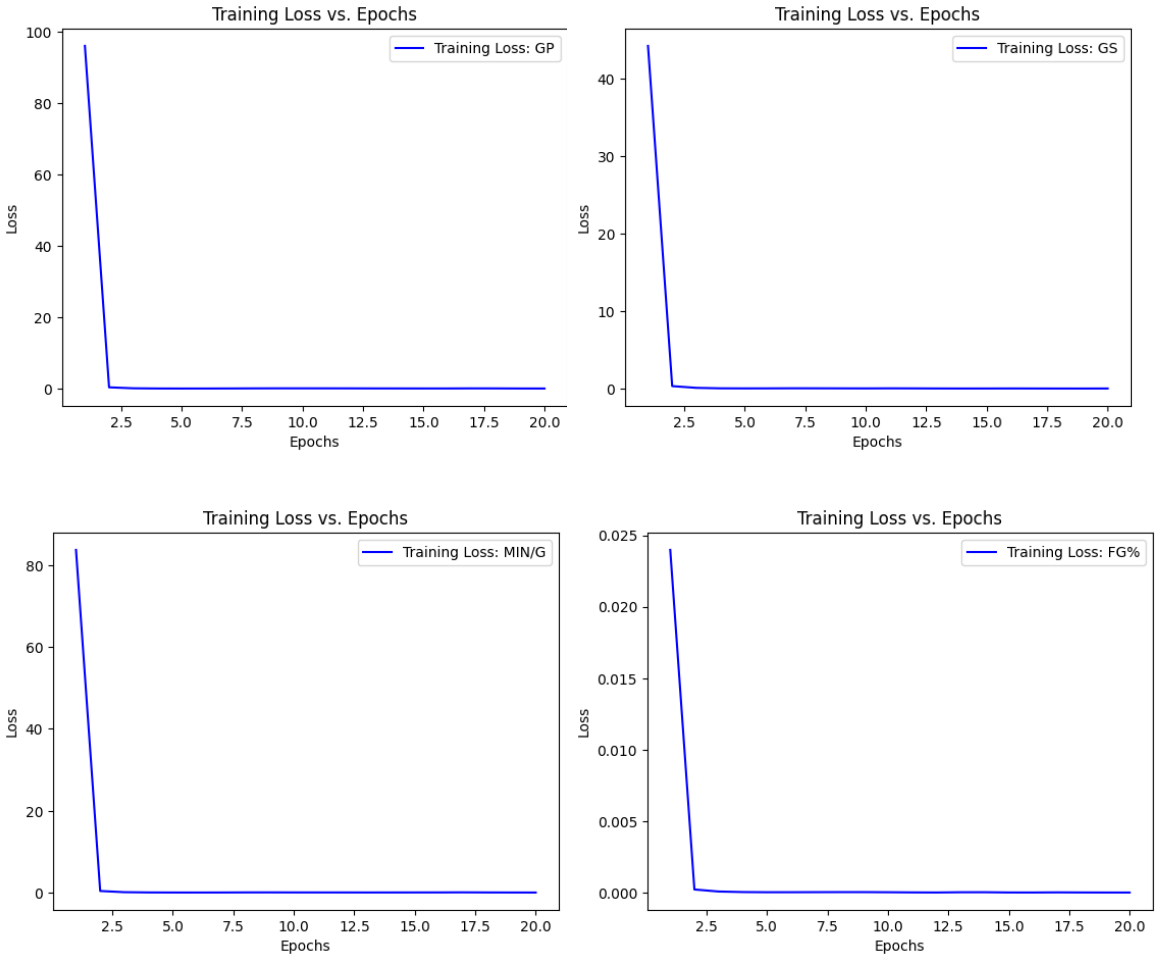

Figure 5: Training Loss vs. Epochs for 4 different columns (GP, GS, MIN/G, FG%)

Once we had finalized the neural network model, we input in some testing data from our dataset to train our neural network. We trained the network over a total of 20 epochs, which resulted in accurate predictions for each player statistic. We selected a few of the feature columns, these being GP, GS, MIN/G, FG% to graph that demonstrated the convergence of loss of our neural network. By measuring the training loss, the error between the model's predictions and the actual data, we were able to gauge how well our model performed when making predictions about a player's statistics. As seen in the figure above, these four graphs converged somewhat rapidly to 0, indicating that our model was learning well from the data and not overfitting the data.

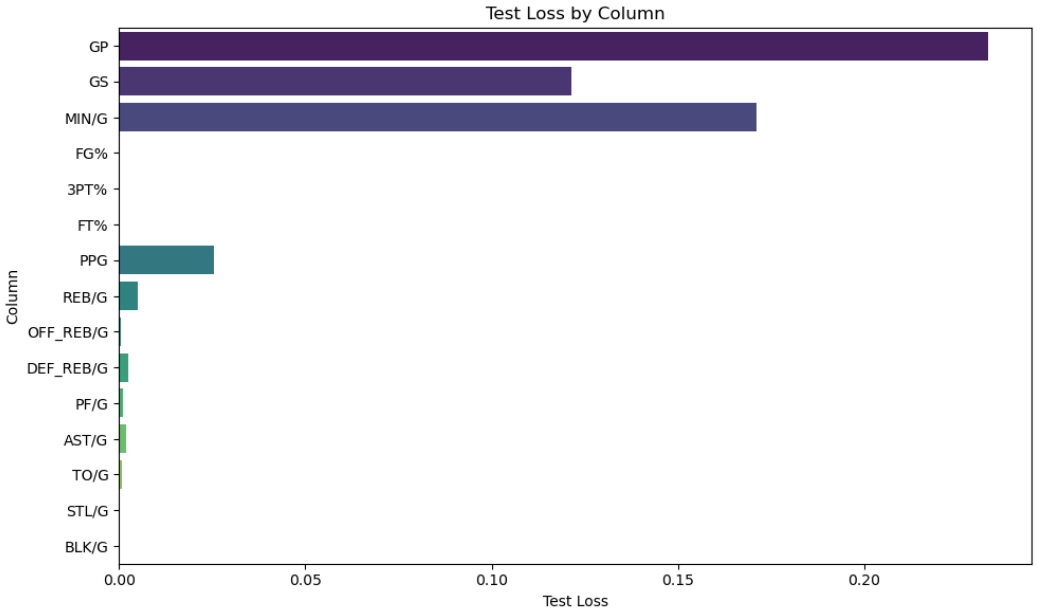

Figure 6: Test Loss by Numerical Features

By analyzing the test loss of each of our columns, we are able to determine the model's predictive performance on each of our target columns. It seems that 'GP'(Games Played), 'GS'(Games Started), and 'MIN/G'(Minutes Played Per Game) were more difficult for our model to predict accurately as they had the top 3 test losses. On the other hand, it seems that PPG(Points Per Game) had a significantly lower test loss than the 3 largest test losses. This indicates that the model predicts the PPG with moderate accuracy. For all the other target columns, their test losses are lower or are close to zero which indicates that our model accurately predicts those target columns as well.

4. Conclusion and discussion for future work

In conclusion, this project addresses the intricate challenge of assessing a prospective player's "transferability" as a D1 basketball player. Recruiting suitable players for college teams requires a deep understanding of the various skill sets needed to succeed, as well as a keen eye for talent amidst competition and limited scouting resources. Not to mention, human judgment can introduce biases into the evaluation process that may limit opportunities for certain players. In an effort to mitigate these challenges, we developed a predictive model that forecasts player performance by leveraging past player statistics. Unlike previous models that focused on aggregated estimates of performance, such as player efficiency rating (PER) and performance index rating (PIR), our model predicts 15 distinct statistics, providing a comprehensive representation of player capabilities. This depth allows coaches to more accurately gauge a player's potential fit within the team. By utilizing a multi-layer perceptron (MLP) model architecture, we established a neural network to map the relationship between Division 2 player stats and Division 1 player expectations. In doing so, our model offers a benchmark for projecting a player's future performance and their potential integration into a Division 1 team. With the predictive analytics produced by our model, coaches can make more informed decisions, optimizing the team's composition and increasing the likelihood of success on the court.

On the other hand, predicting individual player statistics can have various limitations. Some of the player statistics in our dataset may have been altered by outside factors, such

as coaches not letting a certain player play often or players being tired leading to lower game performance. This presents challenges when predicting individual player performance statistics as these kinds of factors are not easily accounted for in either a regression or neural network model.

While the predictions made in this project appear to show fairly accurate results, further testing is needed to determine the model's effectiveness. Given the long training time, we were unable to explore additional hyperparameters. For example, we implemented our own scoring system to encode the conference category based on approximate level. The subjectivity of this task may have lead to different training outcomes. The number of epoches (20), the chosen activation function (ReLU), and the loss function (MSELoss) could also greatly influence the outputs of the neural network. Further research is necessary to establish the optimized combination of network features that would lead to the lowest testing loss. Moreover, future work may incorporate a similar neural network to predict a player's performance for different positions, allowing current players to potentially diverge from their current positions and grow as well-rounded players.

References

Our final report submission folder contains source code and datasets used for our project.

[1] Website: barttovik.com

[2] Website: ncsasports.org

[3] Related Study: "Predicting NBA Player Performance" by Kevin Wheeler