WGU C964: Task 2

Machine Learning Project Proposal for the Memphis GRIZZLIES

**Nikolas Butalid**

Student ID: 001531753

28 April 2022

2022

Contents

[A. Project Proposal for Business Executives 4](#_Toc101787606)

[A.1 Letter of Transmittal 4](#_Toc101787607)

[A.2 Project Recommendation 7](#_Toc101787608)

[A.2.1 Project Summary 7](#_Toc101787609)

[A.2.2 Application Benefits 7](#_Toc101787610)

[A.2.3 Application Description 8](#_Toc101787611)

[A.2.4 Data Description 8](#_Toc101787612)

[A.2.5 Objectives and Hypothesis 8](#_Toc101787613)

[A.2.6 Methodology 9](#_Toc101787614)

[A.2.7 Funding Requirements 10](#_Toc101787615)

[A.2.8 Data Precautions 10](#_Toc101787616)

[A.2.9 Developer’s Expertise 11](#_Toc101787617)

[B. Project Proposal 11](#_Toc101787618)

[B.1 Problem Statement 11](#_Toc101787619)

[B.2 Client Summary 12](#_Toc101787620)

[B.3 Existing System Analysis 12](#_Toc101787621)

[B.4 Data 13](#_Toc101787622)

[B.5 Project Methodology 13](#_Toc101787623)

[B.6 Project Outcomes 14](#_Toc101787624)

[B.7 Implementation Plan 14](#_Toc101787625)

[B.8 Evaluation Plan 15](#_Toc101787626)

[B.9 Resources and Costs 16](#_Toc101787627)

[B.10 Timeline and Milestones 16](#_Toc101787628)

[C. Application 17](#_Toc101787629)

[C.1 List of Submitted Files 17](#_Toc101787630)

[D. Post-implementation Report 18](#_Toc101787631)

[D.1 Business Vision 18](#_Toc101787632)

[D.2 Datasets 19](#_Toc101787633)

[D.3 Data Product Code 21](#_Toc101787634)

[D.4 Objective Verification 22](#_Toc101787635)

[D.5 Effective Visualization and Reporting 23](#_Toc101787636)

[D.6 Accuracy Analysis 25](#_Toc101787637)

[D.7 Application Testing 26](#_Toc101787638)

[D.8 Application Files 26](#_Toc101787639)

[D.9 User Guide 28](#_Toc101787640)

[D.10 Summation of Learning Experience 29](#_Toc101787641)

[E. Sources 30](#_Toc101787642)

# A. Project Proposal for Business Executives

## A.1 Letter of Transmittal

April 25, 2022

Zach Kleiman

General Manager

Memphis Grizzlies

191 Beale Street

Memphis, TN 38103

Dear Mr. Kleiman,

Being successful in the NBA is no easy feat. The success of the past regular season was a major achievement and calls for much to be celebrated. The Memphis Grizzlies seem to be on track on becoming a formidable dynasty in the West for years to come. There is, however, always room for improvement until a championship is won. It has come to my attention that there is a desire in the organization to further build upon the preexisting success and create a championship-contending team. The current core team has been great, but there are some areas of the team that could be better optimized. As you know, after this next season there will be some flexibility in the salary cap and I believe that we can use that to the Grizzlies’ advantage to fill out some of pieces that the team needs to make a deep playoff run, possibly including a championship appearance. To do this to the best of our ability, I have a data product solution I would like to propose. I believe that my application will be able to find the best players for our given budget and help build the most effective team available to us.

This proposed application will use machine learning models to find an appropriate salary for a player based on their past performances. Simply put, this solution will analyze past player statistics and their salaries. The data product will then be trained to predict what a prospective player’s yearly salary should be based on their past statistics, such as number of games started, box plus minus, and win shares. Using this information, the Memphis Grizzlies would be able to come up with appropriate contracts that are fair to the player but also maximize the potential of our salary cap. Drafting contracts using this solution would not only help the Grizzlies further build a dominate team, but it would also help with negotiating contracts and scouting reports.

I have great confidence that this solution would propel the Grizzlies into the next tier of teams in the NBA. I have been analyzing and working with NBA data for the past decade. I have worked with other professional basketball teams and helped their organization increase their winning percentages by at least 10%. Using my degree in Computer Science, I have also been keeping up to date with the latest emergences in machine learning and artificial intelligence technology. I have multiple certifications in Project Management and would be able to confidently lead a small team in developing this application. The cost of development for this proposal would be $29,800 with an additional $1,500 contingency fund which would be a total of $31,300. This investment would cover the cost of labor, and the tools and technology need to create this application.

Thank you for your time and consideration. I hope to hear from you soon, and I look forward to working with you and getting the Grizzlies to the Finals one day.

Sincerely,

Graphical user interface, application

Description automatically generated

Nikolas Butalid

## A.2 Project Recommendation

### A.2.1 Project Summary

One of the things that NBA teams must deal with every season is how to handle their player contracts and how to make the most of their payroll. The data application described in this proposal will aim to ease the burden of assessing a player’s value. Many NBA teams in the past have suffered from creating bad contracts that do not match the value of the involved player. There are many examples, even today, of players with contracts they do not deserve that are greatly hindering their team. This data product solution will enable the Memphis Grizzlies to accurately and fairly evaluate players based on their statistics. It will use machine learning techniques to generate an estimated salary for a given player. If implemented, the proposed solution will allow the Grizzlies to make better educated decisions with how they pay players. Doing so will maximize the potential of the team that the salary cap would allow.

### A.2.2 Application Benefits

Many good NBA teams in the past have experienced success because of good contracts. The proposed application will enable the Grizzlies to find the best available players within their budget. There are players in the NBA whose market value is skewed heavily from their actual value. The data product will be able to reveal which players are being under or overvalued in the NBA market and the Grizzlies will be able to take advantage of that. Using this information, the Grizzlies will be able to surround their already excellent core with perfect-fitting role players that will turn the Grizzlies from a good team into a championship-contending team.

### A.2.3 Application Description

The proposed data application will utilize machine learning methodologies to study past NBA player data. It will look at data from the past 20 years (2000-2020). This data will include player statistics and their salaries. Machine learning models will be used to find the relationship between certain stats and salaries. Using these techniques will allow the application to predict what a player’s salary should be based on their given statistics.

### A.2.4 Data Description

The data that will be used for this proposal will be from Basketball-Reference which is one of the most comprehensive basketball data sources available to the public. The data was scraped by a contributor on Kaggle.com. The data will include both basic and advanced NBA statistics. This includes features like points, rebounds, win shares, player efficiency rating, and many more. The definitions of these categories can be found on the Basketball-Reference website. Another dataset that will be used will contain information about a player’s salary for a given season. The data that will be used in this proposal will be from 2000 to 2020. It will only include NBA data and no other professional basketball leagues. There are also numerous players that played on multiple teams in a single season, and their data will be consolidated into a single entry for one season.

### A.2.5 Objectives and Hypothesis

The aim of this solution will be to accurately access the monetary value of an NBA player based on their statistics for a given year. This information will hopefully give an accurate representation of what kind of salary that certain player should expect. This proposal also seeks to find which statistics tend to have a greater impact on a player’s salary. If the data application evaluates NBA data from the past, then it will be able to make accurate predictions of a player’s salary. This proposal would like to achieve at least a 0.8 coefficient of determination which would suggest the model has found a strong correlation between the independent variables (player stats) and dependent (player salary) variable.

### A.2.6 Methodology

This proposal will use the SEMMA (Sample, Explore, Modify, Model, and Assess) methodology. SEMMA was developed by the SAS Institute and is widely used as a data mining approach. It can be used in many different implementations to solve business problems. Because of its data orientated design, SEMMA fits well for this proposal.

**Sample**

The data that will be used will be from Kaggle.com datasets. Data from Basketball-Reference was scraped by an online contributor. It will include data about player stats per season. It will include both basic box score stats and advanced metrics. The application will also use data about player salary. The data will be from the 2000 NBA season to the 2020 NBA season. Cross validation will be used so the data will not be manually split into training, validation, or testing sets.

**Explore**

Different visualization techniques will be used to help understand the data better. Various statistics will be plotted against player salaries to visually show the relationship between different stats and salary.

**Modify**

The datasets will be cleaned and trimmed. Extraneous categories will be removed from the preexisting data so that the model is not overwhelmed with superfluous information. Any incomplete or irrelevant pieces of data will be appropriately dealt with.

**Model**

After the data is cleaned, a machine learning model will be fed the cleaned-up data. It will look through the different statistics and analyze their impact on a player’s monetary value. The model will be trained to predict what a player’s salary is based on a set of specific statistics.

**Assess**

After the model is trained, the results will be assessed. Different metrics such as coefficient of determination and mean absolute error will be used to evaluate the accuracy of the model.

### A.2.7 Funding Requirements

The project will initially require $500 for access to various tools and licensed software. It will also require $500 for the server that will be used. The solution will be developed over the course of 30 business days. There will be two data scientists working 8 hours a day on this project at $60/hour. The total cost of labor would be $28,800. Including a contingency fund of $1,500, the total cost of the project will be $31,300.

### A.2.8 Data Precautions

All the involved raw data will already be accessible to the public. The information that will be used can be found on Kaggle.com and will not need to be protected. Any findings of this proposed solution, however, should only be accessible by those who need it in the Memphis Grizzlies Organization.

### A.2.9 Stakeholders Impact

First and foremost, if the proposed application proves to be worthwhile, the success of the Memphis Grizzlies will be enjoyed by our loyal fans. Winning more games would also be a great victory for the shareholders and the ownership group of the Memphis Grizzlies. Success tends to bring in more fans and the data product could certainly help in achieving that. In addition, any advancements made in the machine learning realm could greatly inspire even more data-related projects in the future. The Grizzlies’ analytics team could apply some of the findings of this project to their own work and further help the team achieve success.

### A.2.10 Developer’s Expertise

The lead engineer in this project brings a lot of NBA data analytical experience. They have various degrees and certifications related to data science, analytics, project management, and statistics. They have over a decade’s worth of experience working with professional basketball teams. They have helped those teams increase their winning percentages by at least 10%. They are also very involved in the data science community and have worked on numerous machine learning and artificial intelligence projects. This experience guarantees that the methodologies used in this application will be reliable and up to industry standards. Their experience in project management will also help in the completion of the solution in a timely manner.

# B. Project Proposal

## B.1 Problem Statement

Recently, the Memphis Grizzlies have been successful in the regular NBA season. Although they have a great young core of players, there are still a few missing pieces that could turn them into a championship-contending team for the foreseeable future. Every year, the team must make decisions about how to get the best players during the offseason using their available money from the salary cap. To acquire the best available players, the Grizzlies are looking to find better ways of analyzing player value and how to make more informed financial decisions. Doing this would likely lead to a more well-rounded team and greater success in the playoffs.

## B.2 Client Summary

The proposed solution would be beneficial to those in the Memphis Grizzlies’ front office. It would enable them to evaluate players and make better decisions regarding contracts and trades. It would also be helpful to those in the scouting department. The application could aid the scouts in determining what kind of metrics to pay attention to when surveying prospective players. The data product will not be difficult to use and will not require much technological experience.

## B.3 Existing System Analysis

The Memphis Grizzlies does use machine learning techniques in some areas of the organization, but there are currently no data science related products being used to estimate player value. The Grizzlies mostly rely on traditional box score analysis to examine player performance. They do have access to a large database of statistics from past seasons, but there is no machine learning implementation utilizing those stats. When determining a player’s monetary worth, there seems to be a heavy reliance on human assessment. While there is great information to be gleaned from these methods, a solution could be created to do these same tasks but on a much grander scale. The proposed application will use machine learning to conduct similar analyses but with thousands and thousands of data points. The application will also be able to incorporate more categories of statistics than what would be feasible for a human to consider.

## B.4 Data

The data that will be used for this project will primarily come from Kaggle.com. These datasets will be .csv files that contain a variety of statistics and information about individual players. A contributor on Kaggle.com scraped this data from Basketball-Reference which is a website that contains comprehensive information about everything NBA related. Most of the data will not need to be cleaned but there some entries that have been left empty and must be dealt with appropriately.

## B.5 Project Methodology

The proposed solution will follow the SEMMA methodology. During the **Sample** phase, the data will be gathered and downloaded from Kaggle.com. The dataset will be loaded onto a local machine and be prepped to be used. Information about basic and advanced statistics from 20 NBA seasons (2000-2020) will be collected. A dataset about player salary will also be downloaded from github.com. During the **Explore** phase, the data will go through an initial examination to find which data points and categories will be most useful to the machine learning model. Some of the data will visualized through graphs and charts to see what kind of relationship certain stats have with player salary. During the **Modify** phase, the data will be cleaned up and imputation will be done on the datasets. Some entries have null values which the model is not fit to handle. These values will be replaced with zeroes to give the model a more accurate representation. The features of the dataset will be analyzed, and the stat categories the model will consider will be selected. During the **Model** phase, the dataset will be fed into the chosen machine learning model. The model that will be used for this solution will be a linear regression model. The Random Forest Regressor from the scikit-learn library will be trained to predict a player’s salary based on a variety of statistics. A base model using the default hyperparameters will first be used to establish a baseline performance. During the **Assess** phase, the model will be evaluated for its accuracy. The baseline model will be assessed by finding its R-squared value, mean absolute error, mean squared error, and mean absolute percentage error. After this evaluation, the model will be tuned by randomly changing certain hyperparameters such as n\_estimators, min\_samples\_split, and max\_depth. A more optimal model using the new parameters will be created and evaluated. This process can be repeated until an acceptable model is found.

## B.6 Project Outcomes

Once the proposed solution has been developed and is ready for production, the application will be delivered to the client. The data product will be a web-application that will include a dashboard report of the findings of the project as well as an interactive GUI. The interface will allow a user to enter the statistics of a prospective player and the application will give an estimate of that player’s monetary value. The user will be able to view several graphs and charts that detail and explain the project datasets. A user guide will be provided that will outline how to operate the application and explain the provided visuals.

## B.7 Implementation Plan

Data will be gathered and downloaded from Kaggle.com. The data will include data about basic and advanced statistics. A dataset about player salary will also be included. The most useful stat categories will be selected for use. Some of the data will visualized through graphs and charts to see what kind of relationship certain stats have with player salary. Some entries have null values which the model is not fit to handle. These values will be replaced with zeroes to give the model a more accurate representation. The dataset will be fed into the chosen machine learning model. The Random Forest Regressor model from the scikit-learn library will be trained to predict a player’s salary based on a variety of statistics. A base model using the default hyperparameters will first be used to establish a baseline performance. The model will be evaluated for its accuracy. After this evaluation, the model will be improved by experimenting with randomly selected hyperparameters. Once a suitable model is found, it will be incorporated into the data application and an interactive GUI will be created. Any changes that must be made will be done on the server that the application is hosted on. The model could be further improved upon in the future if desired.

## B.8 Evaluation Plan

The data application will be evaluated during development by looking at its performance metrics. These assessments will adhere to industry standards and will be dependent on the kind of machine learning model that will be used. Because the model in use is a linear regression model, four methods of grading will be used. They will be the coefficient of determination (R-squared), mean absolute error (MAE), mean squared error (MSE), and mean absolute percentage error (MAPE). The r-squared value of a model explains how strong the correlation between its independent and dependent variables is. In the case of this proposal, a higher r-squared value means that the machine learning model is accurate. MAE is the average difference between the actual values and the values predicted by the model. MSE shows how close the regression line is to the dataset. MAPE is also another method of measuring the accuracy of the model. The main metric that will be observed is the coefficient of determination. This proposal will strive to achieve a coefficient of determination of at least 0.8. If this is not reached, the model will be reassessed and will be further worked upon.

## B.9 Resources and Costs

The application will be hosted using Heroku. The cost of using setting up a server that meets the performance requirements of the application will be $500. The project will also use tools such as JetBrains’ Datalore and PyCharm. Usage of these tools will cost approximately $500 during the duration of the project. Two data scientists will be needed for the development of the data product. The development of the application is estimated to take 30 days. The two data scientists will be working 8 hours day at $60 an hour. The total cost of labor will be $28,800, Other libraries such as scikit-learn and Tensorflow will be used but will not require any additional costs. The cost of tools and labor will be a total of $29,800. Another $1,500 (about 5% of project cost) will required as a contingency fund. The estimated cost of the entire project is $31,300.

## B.10 Timeline and Milestones

**Proposed Timeline**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Activity** | **Staff Assigned** | **Hours required** | **Start Date** | **End Date** |
| 1 | Proposal Accepted | Project Management Office | 4 | 5/2/22 | 5/2/22 |
| 2 | Data gathering | Data Scientists (2) | 8 | 5/2/22 | 5/3/22 |
| 3 | Data examination | Data Scientists (2) | 12 | 5/3/22 | 5/4/22 |
| 4 | Data visualization | Data Scientists (2) | 16 | 5/5/22 | 5/6/22 |
| 5 | Data imputation | Data Scientists (2) | 8 | 5/9/22 | 5/9/22 |
| 6 | Data features selection | Data Scientists (2) | 4 | 5/10/22/ | 5/10/22 |
| 7 | Baseline model creation and training | Data Scientists (2) | 28 | 5/10/22 | 5/13/22 |
| 8 | Baseline model examination | Data Scientists (2) | 16 | 5/16/22 | 5/17/22 |
| 9 | Model tuning and reassessment | Data Scientists (2) | 40 | 5/18/22 | 5/24/22 |
| 10 | Final model training and evaluation | Data Scientists (2) | 16 | 5/25/22 | 5/26/22 |
| 11 | GUI development | Data Scientists (2) | 32 | 5/27/22 | 6/1/22 |
| 12 | Application deployment | Data Scientists (2) | 32 | 6/2/22 | 6/7/22 |
| 13 | Application approved | Data Scientists (2) | 16 | 6/8/22 | 6/9/22 |
| 14 | Project delivery | Data Scientists (2) +  Project Manager Office | 8 | 6/10/22 | 6/10/22 |

# C. Application

## C.1 List of Submitted Files

A link to the web-application will be provided:

<https://butalid-c964.herokuapp.com/>

# D. Post-implementation Report

## D.1 Business Vision

Building a championship-contending team is a difficult task to achieve. One of the obstacles any NBA team must face is selecting a group of players they believe can bring success to the franchise. In addition, the team must do so within the NBA salary cap. The Memphis Grizzlies is an example of a team that has good foundational players that will likely stay with the organization for years to come. They have proved this in the recent regular season with their ranking in the western conference. The Grizzlies, however, still need additional supplemental players to have the best chance of reaching the NBA Finals. The proposed solution has been built to address the need of recruiting the most effective players in the upcoming offseason. The application, using machine learning, is able to estimate what a player’s salary should be based on their statistics. This enables the Grizzlies to scout out potential players and find the largest return on investment possible.

For example, if the Grizzlies are looking at a potential player, they can enter that player’s statistics from the previous season. The application has looked at decades’ worth of information about player statistics and salary. Using this knowledge, it will give an estimated salary based on what the model thinks that player is worth. The Grizzlies front office would then be able to use this information to negotiate a possible contract that would be advantageous to both the player and the team.

Graphical user interface, application, Word

Description automatically generated

Figure 1: The model estimates the player should be given a $18,715,371 contract based on their stats

## D.2 Datasets

The data used in this application came from Kaggle.com and github.com. The data was scraped from Basketball-Reference.com. Basketball-Reference is a great source of NBA and contains a vast amount of statistics about the NBA. The data came in the form of .csv files and were mostly numeric in nature. Some data had to imputed because some of the values were null. These entries were unreadable to the model so the null values were properly replaced with zeroes. Multiple datasets were combined to include both basic and advanced statistics. The resulting dataset was further A picture containing text, screenshot, computer

Description automatically generatedtrimmed down from containing 68 categories to 25.

Figure 2: The initial combined dataset

A picture containing text, screenshot, computer, computer

Description automatically generated

Figure 3: The final dataset used for the application

## D.3 Data Product Code

The resulting dataset was mostly numeric so there little cleaning that had to be done. Null data entries were imputed with zeroes. Multiple datasets were joined together using the player’s name and the year of the season. Some of the major stat categories were plotted against a player’s salary to show a visual representation of their relationship. The data was split into independent variables and a single dependent variable. The independent variables included all statistics except for player-identifying information and salary. The dependent variable was the player’s salary.

The data was fed into a Random Forest Regressor model from the scikit-learn library. A baseline model was trained after the dataset was split into training and testing sets. The Random Forest Regressor was chosen because it is a comprehensive model that handles large datasets well. It also works with many independent variables. After the baseline model was trained, it was evaluated for accuracy. After evaluation, the model was tuned by experimenting with random hyperparameters. This was done by using scikit-learn’s Randomized Search CV. The function was able to find more optimal hyperparameters for the model. The found parameters were used to create and train a new model. This model was indeed more accurate than the baseline model and is the model being used in the application. The evaluations were done being using scikit-learn’s coefficient of determination, mean absolute error, mean square error, and mean absolute percentage error functions.

Graphical user interface, application, Word

Description automatically generated

Figure 4: The hyperparameters were changed in the model and resulted in a slightly better model

## D.4 Objective Verification

The original hypothesis of the proposal aimed to understand the relationship between player statistics and their salary. This was attempted by training a machine learning model and evaluating the salary predictions of that model. The hypothesis looked to achieve a coefficient of determination of at least 0.8. The resulting model from the proposal could only achieve a coefficient of determination of 0.473 which is far from the targeted value. In conclusion, the original hypothesis was not verified. Additional research and tuning will need to be done in order to reach the original objective.

## D.5 Effective Visualization and Reporting

Chart, histogram

Description automatically generatedTo visualize the dataset, various graphs were created to better understand the data. The first visual is a histogram showing the distribution of salaries in the given dataset. This gives the user an idea of how money was spread out among NBA contracts. Another graph shows the feature importance of the machine learning model. This tells the user that the model found that the number of games started had the largest impact on a player’s salary. After that image, another scatterplot showing salary and the number of games started can be found. The next scatterplot visualizes the relationship between points scored and salary. The graph also distinguishes between total points in a season and points per 100 possessions. Per 100 possessions stats are largely accepted as more accurate metrics than per game stats. Lastly, there are four scatterplots showing four of the major basketball stat categories versus player salary. It shows a slightly positive correlation for each category, but it does not appear to be a strong correlation.

Figure 5: Salary Histogram

Chart, bar chart

Description automatically generated

Figure 6: Feature Importance Bar Graph

Chart, scatter chart

Description automatically generated

Figure 7: Games Started Scatterplot

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 8: Major Stat Category Scatterplots

## D.6 Accuracy Analysis

Because a linear regression model was used, accuracy was graded by using four different functions from the scikit-learn library. These include coefficient of determination (R-squared), mean absolute error (MAE), mean squared error (MSE), and mean absolute percentage error (MAPE). The main objective of the proposal was to achieve a coefficient of determination, or r-squared value, of 0.8. As shown in Figure 4, the r-squared value of the baseline model was 0.4695. The model was then improved and scored a r-squared value of 0.473. This evaluation shows that the model did not find a strong correlation between player statistics and their salary. In Figure 1, the model made a player’s salary prediction of $18,715,371 based on the given statistics. The accuracy evaluation of the model shows that this prediction may not be very accurate. For example, Nikola Jokic’s 2020-2021 statistics were entered into the model predicted a salary of $23,387,288. His actual salary from that season was $28,542,009.

## D.7 Application Testing

Testing and verification was done at each stage of the development process. This was essential to do because each subsequent step was dependent on the previous step. The datasets were tested and verified with official NBA data to assure the data was merged correctly. Immense testing was done to assure the application runs on its web server properly. Over five different hosing services were tested to find the best and most cost-effective solution. The model was also evaluated for its accuracy and its hyperparameters were tuned to increase its accuracy. In order for the GUI and the application to function properly, all the preceding code and datasets have to work as planned. This further verifies the working condition of the application.

## D.8 Application Files

No installation is needed to use the application. The secure HTTPS web-app can be found at:

<https://butalid-c964.herokuapp.com/>

On the webpage, the user can look at the provided data and visuals. Near the bottom of the page is the interactive component that allows them to utilize the machine learning model to make a salary prediction.

**Submitted Files**

The original source code and datasets will also be provided but are not required to run the web-application. This is the hierarchy of the submitted files **(Bolded font are folders)**:

**Butalid\_C964 Folder**

-.**ipynb\_checkpoints**

**-data**

**-ipynb\_checkpoints**

**-data\_since\_1947**

**-**Advanced.csv

-All-Star Selections.csv

-End of Season Teams (Voting).csv

-End of Season Teams.csv

-Opponent Stats Per 100 Poss.csv

-Opponent Stats Per Game.csv

-Opponent Totals.csv

-Per 36 Minutes.csv

-Per 100 Poss.csv

-Player Award Shares.csv

-Player Career Info.csv

-Player Per Game.csv

-Player Play By Play.csv

-Player Season Info.csv

-Player Shooting.csv

-Player Totals.csv

-Team Abbrev.csv

-Team Stats Per 100 Poss.csv

-Team Stats Per Game.csv

-Team Summaries.csv

-Team Totals.csv

-nba-salaries.csv

-player\_data.csv

-Players.csv

-Season\_Stats.csv

**-user\_data**

advanced\_data\_2000-2020.csv

per\_100\_2000-2020

totals\_2000-2020

-butalid-nba-capstone.ipynb

-ideal\_model.joblib

-requirements.txt

If desired, the application can be run on a machine by installing the Python modules listed under “requirements.txt”. The application was developed using Python 3.9.12 and Jupyter Notebook. This is not necessary to run the application as it can be found at the previously mentioned website.

## D.9 User Guide

**How to access and use the web application:**

1) Go to <https://butalid-c964.herokuapp.com/>

2) A screen saying “Executing (some number) of 59” should appear. The model is executing and loading the application.

3) This web app may take a minute or two to load.

3) If it does not load, try refreshing the page.

4) You will be presented with the involved data and the provided visuals.

5) Scroll down to near the bottom of the page to the section called “**Predict a player's salary based on their stats**”.

6) Enter the statistics of the prospective player. The stats should be for one year’s worth of play. Make sure the statistics you enter are totals for an entire season, not per game.

7) Click the “Predict Salary” button and it should give an estimated salary.

8) If it does not work, try pressing the button again. If that does not work either, try refreshing the page and go back to step 5.

## D.10 Summation of Learning Experience

This capstone project was quite enjoyable. I got the opportunity to work with data that I am very passionate about as I am an avid NBA fan. I learned a lot about the mechanical workings of machine learning and data science. Outside of school-related material, I have not had much hands-on experience with machine learning models. To prep myself for this project, I went through a data science bootcamp on Udemy that was very helpful in teaching me machine learning methods. I spent a lot of time working in and getting familiar with Python and Jupyter Notebook. I did run into some trouble with hosting my application. In the end, I used Heroku’s free services to host my application. Through these difficulties, I have a better understanding of server hosting and cloud services. I had to resort to a lot of official documentation on scikit-learn, Jupyter Notebook, Heroku, and Python. I did not reach other to any individuals for assistance but sometimes used online forums to find solutions to issues I ran into. I really enjoyed working with data and this project further confirmed that this is an industry I am very interested in and passionate about. The kind of information you can glean from analyzing data is astonishing and can even be life-changing for some.

# E. Sources

*Basketball statistics and history*. Basketball. (n.d.). Retrieved April 28, 2022, from <https://www.basketball-reference.com/>

Datta, S. (2021, April 28). NBA Data Since 1947. Kaggle.com

Webb, E. (2020). datasets (Version 1.0.0) [Computer software]. <https://github.com/erikgregorywebb/datasets>