McMaster University

PROJECT RAPTORS REBOUND

DAT 205 - CAPSTONE

**WINTER 2021 – GROUP 1**

Bhavika Patil

Cindy Guo

Dennis Hung

Table of Contents

[Executive Summary 3](#_Toc68726167)

[Business Goal 4](#_Toc68726168)

[Objectives 4](#_Toc68726169)

[Data Description 5](#_Toc68726170)

[Gathering data – nba api 5](#_Toc68726171)

[Gathering data – Salary Data 7](#_Toc68726172)

[Implementations 9](#_Toc68726173)

[Data Preparation and Analysis 9](#_Toc68726174)

[Phase 1: Feature Selection 9](#_Toc68726175)

[Phase 2: Player Selection 12](#_Toc68726176)

[Results 14](#_Toc68726177)

[Conclusions 16](#_Toc68726178)

[Recommendations 16](#_Toc68726179)

[Limitations 17](#_Toc68726180)

[References 18](#_Toc68726181)

[APPENDIX A: Project Planning 19](#_Toc68726182)

[APPENDIX B: Implementation Tools 20](#_Toc68726183)

[APPENDIX C: Correlation Matrix 21](#_Toc68726184)

[APPENDIX D: Feature Importance 23](#_Toc68726185)

**Figures**

[Figure 1 Example PlayerGameLogs data subset 6](#_Toc68641753)

[Figure 2 Available features for data between 2005 to 2020 (NBA Media Ventures, LLC., 2021) 8](#_Toc68641754)

[Figure 3 Example NBA Salary data subset 8](#_Toc68641755)

[Figure 4 Available features for salary data between 2015 to 2025 (ESPN and HoopHype) 8](#_Toc68641756)

[Figure 7 Process for Phase 1: Feature Selection 9](#_Toc68641757)

[Figure 9 Feature correlations with values greater than 0.85 11](#_Toc68641758)

[Figure 11 Accuracy / F1-Score / Sensitivity / Specificity by Model Comparison 12](#_Toc68641759)

[Figure 12 Plot of Feature Importance by Model 12](#_Toc68641760)

[Figure 13 Model Feature Importance 13](#_Toc68641761)

[Figure 14 Process for Phase 2: Player Selection 13](#_Toc68641762)

[Figure 15 TOR and Non-TOR Player Average Stats Comparison 14](#_Toc68641763)

# Executive Summary

HOW TO IMPROVE TORONTO RAPTORS’ CHANCES AT ANOTHER CHAMPIONSHIP RUN WHILE PRESERVING THE TEAM’S LONG-TERM FUTURE

**PROBLEM DESCRIPTION**

​Toronto​​ Raptors is the only Canadian franchise in NBA. After winning their first NBA Championship in 2019, their performance has been in decline since then. The team’s General Manager is seeking recommendations on improving the team’s chances at another championship run while preserving the team’s long-term future. The GM is also very keen on evaluating if, with the recommendations, Toronto Raptors can improve performance-to-cost effectiveness while retaining the public interest in the team.

**DATA**

​We collected the data hosted by various sources such as player performance data from stats.nba.com; and player salary data from ESPN and HoopsHype websites.

**ANALYSIS**

​This analysis was done in 2 phases:

Phase 1 - Players’ Feature Selection: We reviewed Toronto Raptors regular and playoff seasons (2014 to 2020). We compared various models and implemented Random Forest to determine the important features; Plus/Minus, Player Efficiency Rating (PER), and Player Impact Estimate (PIE) for the team to win a game.

Phase 2 - Player Selection: We developed a similar model as in Phase 1 using aggregated team data to predict a game outcome. We selected underperforming Toronto players and replaced them with better performing non-Toronto players. We predicted the wins on the newly generated data with the replacement players.

The model projected a 23%​ ​win improvement.

**RECOMMENDATIONS**

​For improvements to the Toronto Raptors’ roster, the organization is to look at non-Toronto players with higher values in PLUS\_MINUS, PER, and PIE compared to their current roster. These statistics hold potential values in producing team wins and are underutilized by their current teams.​ We recommend players with higher important feature values than the average TOR player and a salary below the average of a TOR player.

# Business Goal

The Toronto Raptors (TOR) is an NBA championship team with a $2.5 billion market capitalization. The franchise added key players in 2019 who contributed to critical wins during the season and championship playoff run. The team won their first-ever championship title in that season. The team’s performance has been in decline since being crowned 2019 NBA champions.

The team’s General Manager (GM) and owners are seeking recommendations on improving the team’s chances at another championship run while preserving the team’s long-term future. The recommendations will be in the form of marginal/minor improvements to the player roster. The stakeholders intend to either maintain or improve on the number of wins during the regular season and more importantly, during the playoffs.

The organization wants to retain a public interest in the team while improving the team’s performance without rebuilding the entire team. A spot in the NBA finals while maintaining the number of regular-season wins would be ideal.

# Objectives

To attain the requirements outlined by the management, we have set 2 objectives. we are looking to determine what features affect team wins/losses and predict how they can be improved/reduced through changes in the team’s roster.

1.       The primary objective of this analysis is to determine which features are important to the Toronto Raptors team to produce more wins in the regular season and the playoffs.

2.       The secondary objective is to review team players’ performance-to-cost effectiveness.

The scope of this analysis will be to focus on the historical player data from 2005 to 2020 as this closely represents the players of the current era. The performance baseline of the team and players will be established using the data for the Toronto Raptors 2019-2020 season. These statistics will be used to indicate performance features for the team and the individual players.

To simplify the analysis the following assumptions were made:

* Changes are only applied to the Toronto Raptors; the statistics from other teams are not affected or changed.
* Obtaining any players is up to the Raptors’ GM to determine if they will pursue the player via free agency, trading of players, etc.

# Data Description

### Gathering data – nba api

For the primary objective, we have consumed the data hosted by [nba.com](http://www.nba.com); It is made available via web services through [stats.nba.com](http://stats.nba.com). A player or team’s base statistics are updated simultaneously with the game and in real-time. nba\_api is an open-access API client library for Python developed by Swar Patel. There are more than 250 endpoints (the method to request information through the API) available.

#### Understanding the Data - PlayerGameLogs

PlayerGameLogs endpoint gathered the statistical data of each player in each game by season (figure 1). Input to this endpoint is the season year (For example:’ 2014-15’). The response from the API will be 34 parameters. Each row represents an entry for a player log for a game. This consists of features from the year 1946 to the present day for roughly 1.3 million records.



Figure 1 Example PlayerGameLogs data subset

Feature WL in the above figure indicates the win or loss of the game. As an exercise of this analysis, the win or loss is to be predicted based on players’ performances, which makes WL the target feature. Classification models such as Logistic Regression, Decision Tree, or Random Forest will be applied for the analysis.

List of Output variables from the API endpoint

|  |  |  |  |
| --- | --- | --- | --- |
| **Attributes** | **Data Type** | **Example value** | **Definitions** |
| **SEASON\_YEAR** | STRING | 2015-16 | NBA season |
| **PLAYER\_ID** | INT | 202710 | Player Unique ID |
| **PLAYER\_NAME** | STRING | Jimmy Butler | Player name |
| **TEAM\_ID** | BIG INT | 1610612741 | Team Unique ID |
| **TEAM\_ABBREVIATION** | STRING | CHI | Team 3-character abbreviation |
| **TEAM\_NAME** | STRING | Chicago Bulls | Team name |
| **GAME\_ID** | STRING | 0011500103 | Unique ID for the game |
| **GAME\_DATE** | DATETIME | 2015-10-23T00:00:00 | Date/Time game is played |
| **MATCHUP** | STRING | CHI vs. DAL | Teams featured in the game |
| **WL** | STRING | W | The player who was on the team who won/lost |
| **MIN** | FLOAT | 24.25 | Minutes Played The number of minutes played by a player |
| **FGM** | INT | 2 | Field Goals Made The number of field goals that a player has made. This includes both 2 pointers and 3 pointers |
| **FGA** | INT | 7 | Field Goals Attempted The number of field goals that a player has attempted. This includes both 2 pointers and 3 pointers |
| **FG\_PCT** | FLOAT | 0.286 | Field Goal Percentage The percentage of field goal attempts that a player makes Formula (FGM)/(FGA) |
| **FG3M** | INT | 0 | 3 Point Field Goals Made The number of 3-point field goals that a player has made |
| **FG3A** | INT | 0 | 3 Point Field Goals Attempted The number of 3-point field goals that a player has attempted |
| **FG3\_PCT** | FLOAT | 0 | 3 Point Field Goal Percentage The percentage of 3-point field goal attempts that a player makes Formula (3PM)/(3PA) |
| **FTM** | INT | 0 | Free Throws Made The number of free throws that a player has made |
| **FTA** | INT | 0 | Free Throws Attempted The number of free throws that a player has attempted |
| **FT\_PCT** | FLOAT | 0 | Free Throw Percentage The percentage of free throw attempts that a player has made Formula (FTM)/(FTA) |
| **OREB** | INT | 0 | Offensive Rebounds The number of rebounds a player has collected while they were on offense |
| **DREB** | INT | 3 | Defensive Rebounds The number of rebounds a player has collected while they were on defense |
| **REB** | INT | 3 | Rebounds A rebound occurs when a player recovers the ball after a missed shot. This statistic is the number of total rebounds a player has collected on either offense or defense |
| **AST** | INT | 6 | Assists The number of assists -- passes that lead directly to a made basket -- by a player |
| **TOV** | INT | 1 | Turnovers A turnover occurs when the player on offense loses the ball to the defense |
| **STL** | INT | 1 | Steals Number of times a defensive player takes the ball from a player on offense, causing a turnover |
| **BLK** | INT | 1 | Blocks A block occurs when an offensive player attempts a shot, and the defense player tips the ball, blocking their chance to score |
| **BLKA** | INT | 0 | Blocks Against The number of shots attempted by a player that is blocked by a defender |
| **PF** | INT | 3 | Personal Fouls The number of personal fouls a player committed |
| **PFD** | INT | 0 | Personal Fouls Drawn The number of personal fouls that are drawn by a player |
| **PTS** | INT | 4 | Points  The number of points scored |
| **PLUS\_MINUS** | INT | 11 | Plus-Minus The point differential when a player is on the floor |
| **DD2** | INT | 0 | Double Doubles The number of double-doubles (double-digit number total in two of the five categories in a game) a player achieves |
| **TD3** | INT | 0 | Triple Doubles The number of triple-doubles (double-digit number total in three of the five categories in a game) a player achieves |

Figure Available features for data between 2005 to 2020 (NBA Media Ventures, LLC., 2021)

### Gathering data – Salary Data

For the secondary objective, we have used multiple data sources. This data is not provided by NBA. We have retrieved historical player salary data from [espn.com/nba/salaries](http://www.espn.com/nba/salaries). We have merged the future salary data from [hoopshype.com/salaries/](https://hoopshype.com/salaries/) to the salary dataset. This was done by web-scraping using python libraries.

#### Understanding the Data – NBA player salary data

The Salary data consists of 16 possible features from the year 2015 to 2025(If contracted) for approximately 500 players.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rk** | **Player Name** | **Tm** | **2015-2016** | **2016-2017** |
| 1 | Stephen Curry | GSW | $11,370,786 | $12,112,359 |
| 2 | Russell Westbrook | WAS | $16,744,218 | $26,540,100 |
| 3 | Chris Paul | PHO | $21,468,695 | $22,868,827 |
| 4 | John Wall | HOU | $15,851,950 | $16,957,900 |
| 5 | James Harden | BRK | $15,756,438 | $26,540,100 |

Figure 3 Example NBA Salary data subset

| **Feature** | **Data Type** | **Example value** | **Definitions** |
| --- | --- | --- | --- |
| **Rk** | INT | 1 | Current NBA rank of the player |
| **PLAYER NAME** | STRING | Stephen Curry | Player name |
| **Tm** | STRING | GSW | Team’s 3-character abbreviation |
| **YYYY-YYYY** | FLOAT | 11,370,786 | Salary (in US dollars) for the given NBA season year  Column YYYY-YYYY represents the NBA season year |

Figure 4 Available features for salary data between 2015 to 2025 (ESPN and HoopHype)

# Implementations

The model from this analysis will be used by Business or Data analysts working for The Toronto Raptors basketball team to improve the performance looking at any set of years. The analyst can choose which performance factors they would like to focus on.

Additional details can be found in the Appendix for the:

* Project plan - Appendix A
* Tools used - Appendix B

# Data Preparation and Analysis

For the Primary objective (Phase 1), we are identifying the most important features for Toronto Raptors to win.

For the secondary objective (Phase 2), we have strategized to identify the underperformers (based on the features identified in the Primary Objective) and replace them with some non-Toronto players who rate better for those features.

### Phase 1: Feature Selection

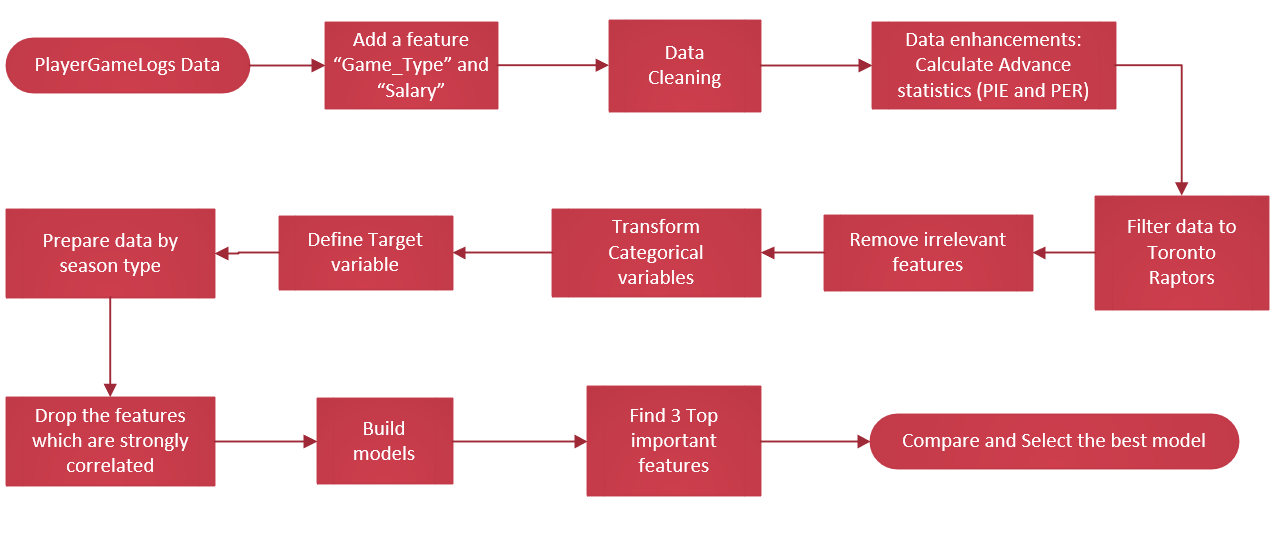


Figure 7 Process for Phase 1: Feature Selection

**Reviewing/Cleansing the data**

The initial data set is gathered through the nba\_api using the endpoint PlayerGameLogs. New features were added

* Game\_Type is applied to each record to categorized by Pre Season, Regular Season, or Playoffs games from 2004 to 2020. Pre Season games were excluded as these held little value. Veteran or marquee players would not play to their typical abilities, unlike new players who need to prove their worth to the coaching staff.
* The salary was appended for each player by season.

Initial analysis of this data indicated minimum data cleansing was required under the feature PLAYER\_NAME. Null values were found for 740 records. No data cleansing was required as these records were under the Pre Season.

**Data Enhancements**

Additional features were appended to enhance the data. These were advanced NBA statistics analysts used to distinguish player utilization or effectiveness which we will use later in our modeling and analysis. We selected to advanced statistics:

* Player Efficiency Rating (PER)
  + The player efficiency rating (PER) is a rating of a player's per-minute productivity.
  + The formula is defined as

PER = (FGM x 85.910 + STL x 53.897 + FG3M x 51.757 + FTM x 46.845 + BLK x 39.190 + OREB x 39.190 + AST x 34.677 + DREB x 14.707 - PF x 17.174 – (FTA-FTM) x 20.091 – (FGA-FGM) x 39.190 - TOV x 53.897) \* (1 / Minutes)

* Player Impact Estimate (PIE)
  + PIE is a metric to gauge a player's all-around contribution to the game.
  + The formula is defined as

PIE = [PTS + FGM + FTM - FGA - FTA + DREB + (0.5 \* OREB) + AST + STL + (0.5 \* BLK) - PF – TOV] / [GmPTS + GmFGM + GmFTM - GmFGA - GmFTA + GmDREB + (0.5 \* GmOREB) + GmAST + GmSTL + (0.5 \* GmBLK) - GmPF – GmTOV]

Note: Statistics prepend with “Gm” refers to the sum of the value for that feature for the entire game

**Feature preparation**

We are focusing on which features will help Toronto Raptors win games.

Features were removed based on experience and initial analysis of the data. Since these typically would not affect the target feature Win/Loss (WL)

* Unwanted Numerical Features
  + PLAYER\_ID
  + TEAM\_ID
* Unwanted Categorical Features
  + PLAYER\_NAME
  + TEAM\_ABBREVIATION
  + TEAM\_NAME
  + GAME\_ID
  + GAME\_DATE
  + MATCHUP

The data was separated into the TARGET feature (WL) and the PREDICTIVE features.

Correlation Matrices (heat maps were plotted to determine any possible correlation among the remaining features.

From the initial correlational matrix projection (Appendix C - Initial Matrix), any values above 0.85 are designated to be too closely related. We found several feature combinations (Figure 9) where the features (PTS, FGA, FTM, REB) added no benefit to the model analysis and were removed.

|  |  |
| --- | --- |
| Feature combination | Correlation Matrix Value |
| FGM vs PTS | 0.96 |
| FGA vs PTS | 0.88 |
| FGM vs FGA | 0.97 |
| FTM vs FTA | 0.97 |
| REB vs DREB | 0.93 |

Figure 9 Feature correlations with values greater than 0.85

The correlation matrix was replotted (Appendix C - Final Matrix) to ensure no other redundant features were in the dataset.

**Model Analysis**

Once the data was split into 70% training and 30% test data, we processed the data using 3 tuned algorithms: Logistic Regression, Decision Tree, and Random Forest. Random Forest showed the best results scoring the highest in all 4 analysis categories (Figure [*Figure 11 Accuracy / F1-Score / Sensitivity / Specificity by Model Comparison*]). The model placed the most importance on the features PLUS\_MINUS, PER, and PIE. (Figure [*Figure 12 Feature Importance by Model*]). These 3 features consisted of 60% of the importance. (Appendix D: Feature Importance)

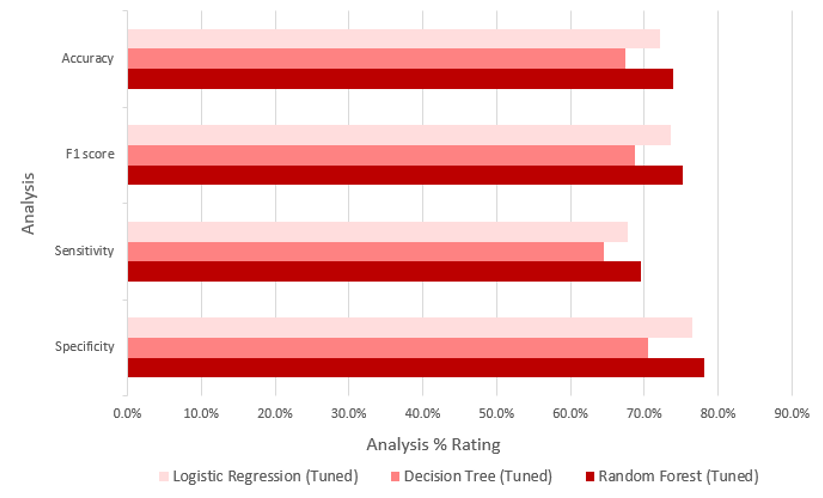


Figure 11 Accuracy / F1-Score / Sensitivity / Specificity by Model Comparison

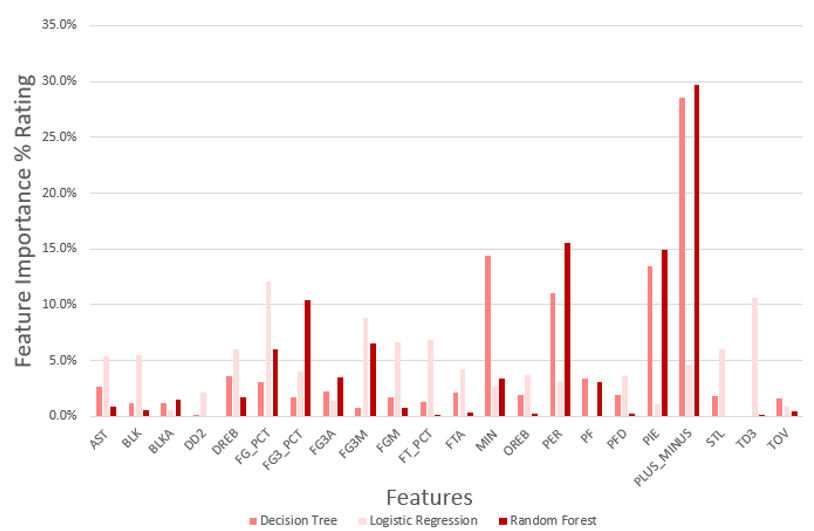


Figure 12 Feature Importance by Model

### Phase 2: Player Selection

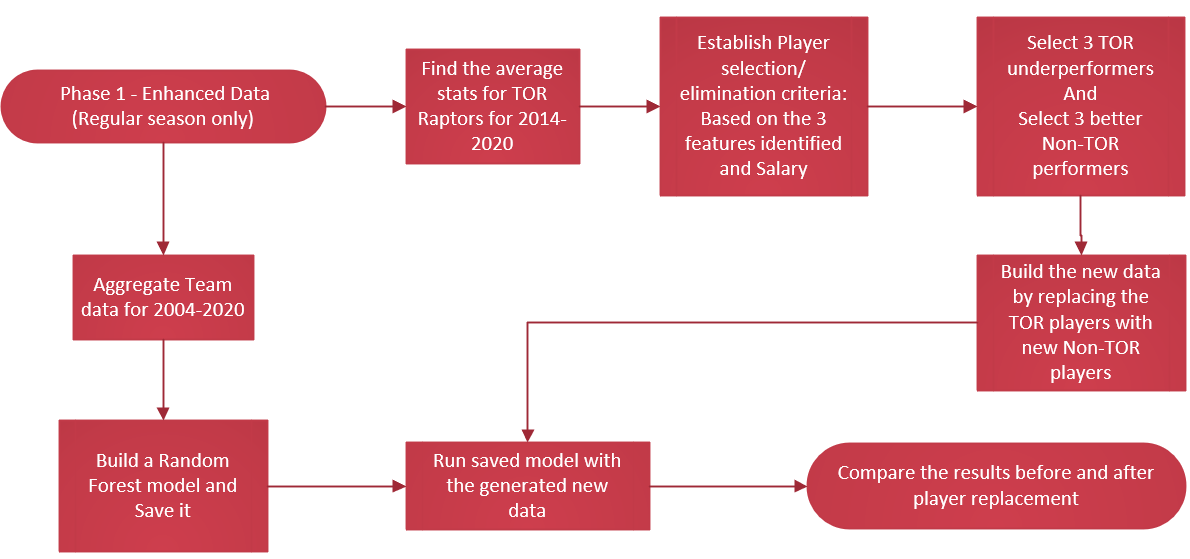


Figure 14 Process for Phase 2: Player Selection

In preparation for showing Team Win improvements, the transformed data from Phase 1 were aggregated by the team/game level for the entire dataset (EDS file) and for one regular season (ORS file based on 2019-2020). The EDS file was processed in the same as the initial PlayerGameLogs data where we only focused on the Random Forest to develop our Team Wins Prediction model. Following are the model results.

|  |  |
| --- | --- |
| **Accuracy** | 96.1% |
| **F1 score** | 96.1% |
| **Sensitivity** | 95.1% |
| **Specificity** | 97.1% |

Team Wins Prediction Model Performance

We adjusted the ORS file to remove 3 underperforming TOR players with selected 3 potential Non-TOR players based on PLUS\_MINUS, PIE, PER, and Salary criteria. (Figure [*Figure 15 TOR and Non-TOR Player Average Stats Comparison*])

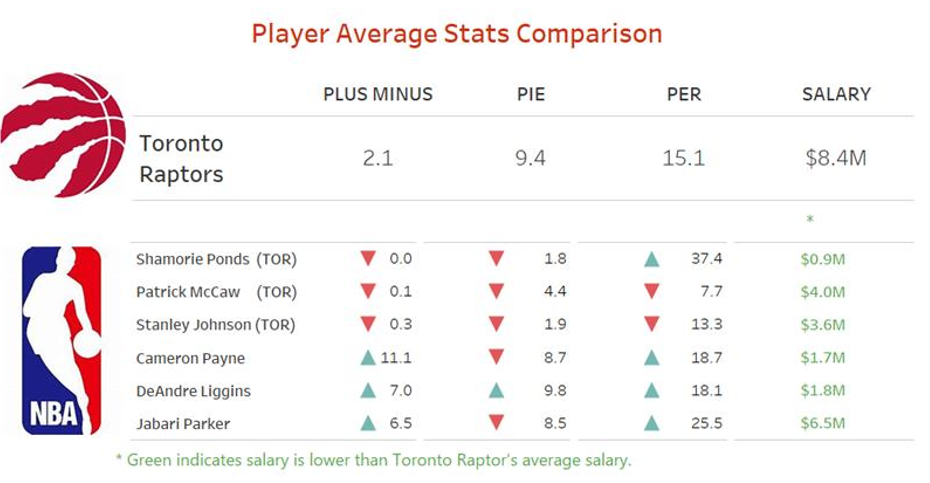
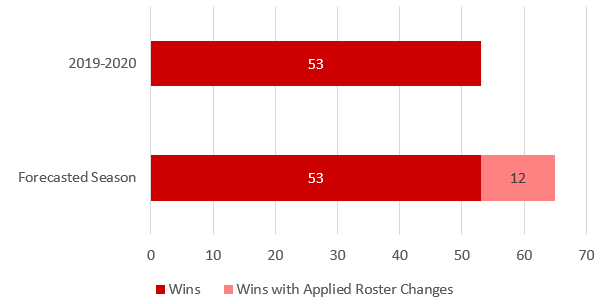


Figure 15 TOR and Non-TOR Player Average Stats Comparison

The model developed from the EDS file was applied to predict the Win/Loss (WL) of the modified ORS file.

# Results

From our model analysis, we forecast an improvement of 22.6% (12 wins) in a 72-game regular season.



Win Prediction

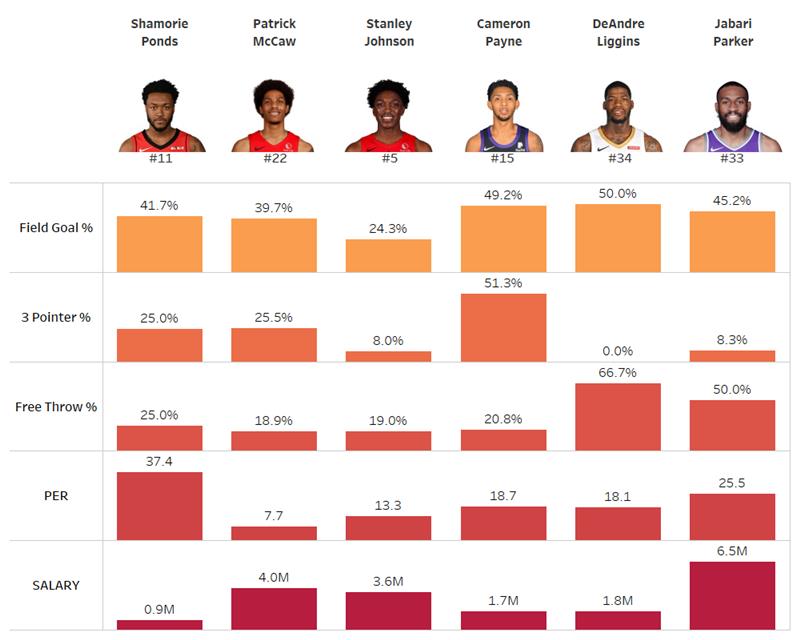


Team Performance Improvement

# Conclusions

### Recommendations

* For improvements to the Toronto Raptors’ roster the organization is to look at non-Toronto players with higher values in PLUS\_MINUS, PER, and PIE compared to their current roster. These hold potential values in producing team Wins at and are underutilized by their current teams.
* Alternatively, Field Goal % (FG\_PCT) and 3-pointers % (FG3\_PCT) are to look out for player efficiency.
* We recommend players with higher important feature values than the average TOR player and a salary below the average of a TOR player.
* There is an evidence of underutilization of a selected TOR player (Ponds). We advocate exploring his utilization or replacing him.



### Limitations

This analysis is limited to Toronto Raptors only. As the teams change, the important features for those teams can sway. There are qualitative factors that affect the performance of a player and are beyond the scope of this project. These include but are not limited to the player’s:

* Health
* Rapport with the team
* Satisfaction with the coaching
* Placement in the games
* Happiness with the contract
* Personal life

There is one player that is currently selected to be replaced on the Toronto roster. Ponds who is new to the roster with only 4 regular season games is seen as a low-value player in terms of PLUS\_MINUS and PIE but has a very high PER. This mainly is due to a lack of usage, but additional data is required.

# References

Bleacher Report, Inc. (2021). *Advanced NBA Stats for Dummies: How to Understand the New Hoops Math*. Retrieved from bleacherreport.com: https://bleacherreport.com/articles/1813902-advanced-nba-stats-for-dummies-how-to-understand-the-new-hoops-math

ESPN. (2020-2021). *NBA Player Salaries 2020-2021*. Retrieved from ESPN.com: http://www.espn.com/nba/salaries

HoopsHype. (2020-2021). *NBA Player Salaries*. Retrieved from hoopshype.com: https://hoopshype.com/salaries/players/

Jee, K. (2019, December 22). *How to Get NBA Data Using the nba\_api Python Module (Beginner).* Retrieved from Playing Numbers: https://www.playingnumbers.com/2019/12/how-to-get-nba-data-using-the-nba\_api-python-module-beginner/

NBA Media Ventures, LLC. (2021). *NBA Advanced Stats*. Retrieved from NBA.com: https://www.nba.com/stats/

NBA Media Ventures, LLC. (2021). *NBA Advanced Stats - Glossary*. Retrieved from nba.com: https://www.nba.com/stats/help/glossary/

Patel, S. (2020, August 19). *swar / nba\_api.* Retrieved from GitHub: https://github.com/swar/nba\_api/blob/master/docs/table\_of\_contents.md

SportRac. (2021). *Toronto Raptors 2020 Salary Cap*. Retrieved from spotrac.com: https://www.spotrac.com/nba/toronto-raptors/cap/

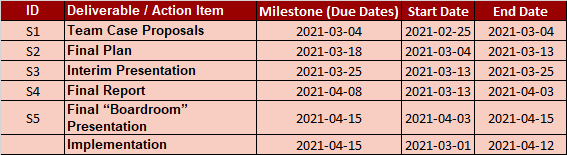
Sports Reference. (2020-2021). *2020-21 NBA Player Contracts*. Retrieved from Basketball Reference: https://www.basketball-reference.com/contracts/players.html

Sports Reference. (2020-2021). *Basketball Reference*. Retrieved from Basketball Reference: https://www.basketball-reference.com/

# APPENDIX A: Project Planning

The overall implementation will be in 5 phases with a deliverable provided at the end of each phase. Completion of the final report is targeted for Apr 15th along with a presentation to summarizing the recommendations outlined in the report.

High-level Milestones: Summary of important deliverable dates



Detailed Project Plan: Detailed project schedule



# APPENDIX B: Implementation Tools

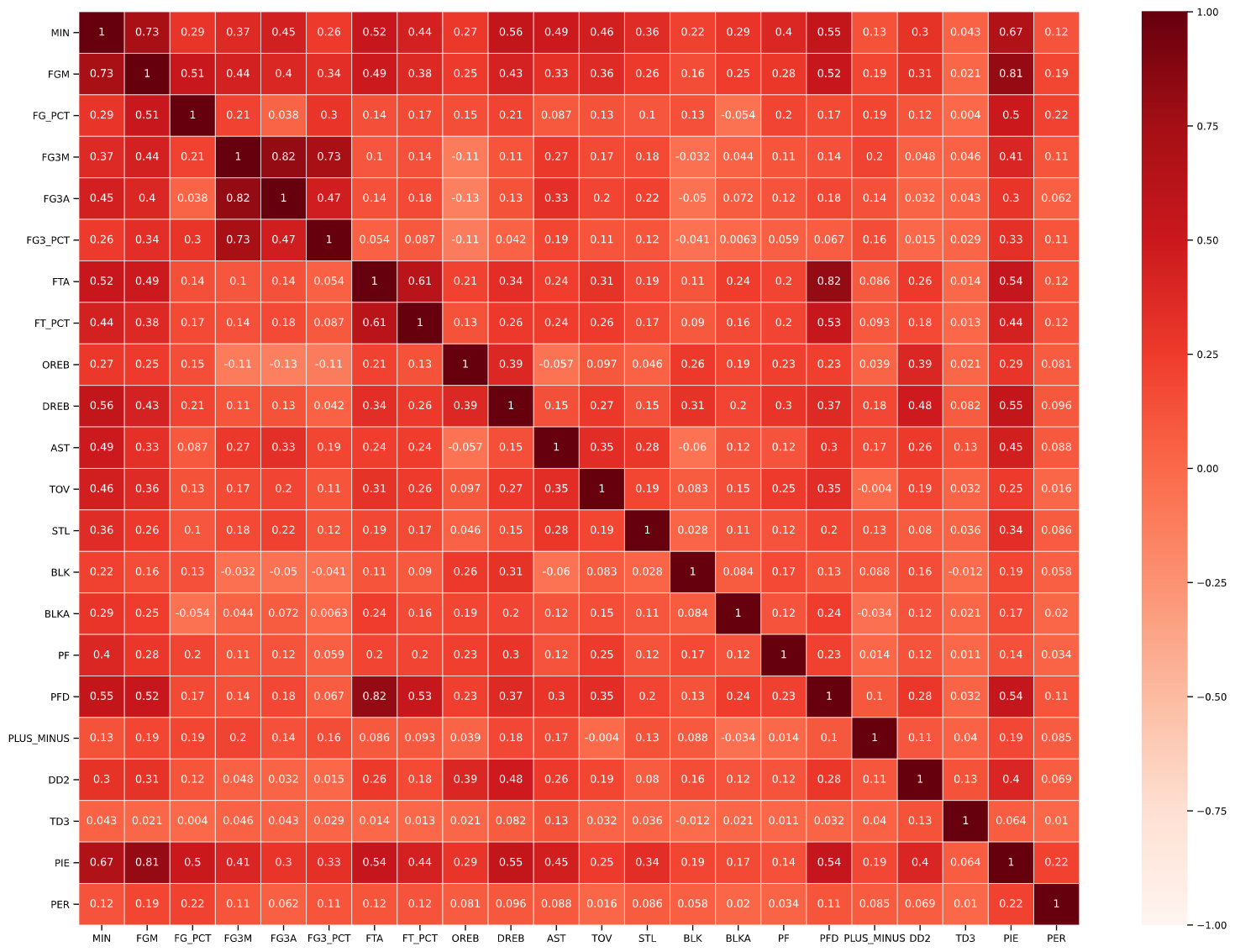
The tools used:

* Visual studio code as an Integrated Development Environment
* Python as the programming language
* Jupyter notebook
* Microsoft Access, Microsoft Excel, and csv as file formats for scraping data.
* Git as a collaboration repository: <https://github.com/bhavikapatil/Capstone> (currently, a private repository)
* Google drive for large files (Private Github doesn’t support files larger than 100 MB)
* GitHub desktop for a Git client
* Microsoft Office as reporting tools
* Tableau desktop for data visualization
* Microsoft Visio

# APPENDIX C: Correlation Matrix

Initial Correlation Matrix for Toronto Raptors (Regular Seasons 2004 to 2020)

Final Correlation Matrix for Toronto Raptors (Regular Seasons 2004 to 2020) after removal of features with values > 0.85



# APPENDIX D: Feature Importance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | | **Decision Tree** | | **Random Forest** | |
| **Feature** | **Feature Importance** | **Feature** | **Feature Importance** | **Feature** | **Feature Importance** |
| FG\_PCT | 12.1% | PLUS\_MINUS | 28.6% | PLUS\_MINUS | 29.7% |
| TD3 | 10.6% | MIN | 14.4% | PER | 15.5% |
| FG3M | 8.9% | PIE | 13.4% | PIE | 15.0% |
| FT\_PCT | 6.8% | PER | 11.0% | FG3\_PCT | 10.4% |
| FGM | 6.6% | DREB | 3.6% | FG3M | 6.5% |
| DREB | 6.0% | PF | 3.4% | FG\_PCT | 6.0% |
| STL | 6.0% | FG\_PCT | 3.1% | FG3A | 3.5% |
| BLK | 5.5% | AST | 2.7% | MIN | 3.4% |
| AST | 5.4% | FG3A | 2.3% | PF | 3.1% |
| PLUS\_MINUS | 4.6% | FTA | 2.1% | DREB | 1.7% |
| FTA | 4.3% | PFD | 1.9% | BLKA | 1.5% |
| FG3\_PCT | 4.0% | OREB | 1.9% | AST | 0.9% |
| OREB | 3.7% | STL | 1.9% | FGM | 0.7% |
| PFD | 3.6% | FG3\_PCT | 1.7% | BLK | 0.6% |
| PER | 3.1% | FGM | 1.7% | TOV | 0.5% |
| MIN | 2.7% | TOV | 1.6% | FTA | 0.3% |
| DD2 | 2.1% | FT\_PCT | 1.3% | PFD | 0.2% |
| FG3A | 1.5% | BLKA | 1.2% | OREB | 0.2% |
| PIE | 1.0% | BLK | 1.2% | FT\_PCT | 0.2% |
| TOV | 0.9% | FG3M | 0.8% | TD3 | 0.0% |
| BLKA | 0.6% | DD2 | 0.1% | STL | 0.0% |
| PF | 0.0% | TD3 | 0.0% | DD2 | 0.0% |