Predicting Offensive Rebounds

Analyst Intern, Data Science &
Solutions Project - Deliverables

The goal of this project is to create a model to predict the chance a given shot is rebounded by the offensive team (i.e. an offensive rebound) for a given shot.

Predicting Offensive Rebounds

Analyst Intern, Data Science & Solutions
Project - Deliverables

1) Describe, generally, from start to finish how you approached, executed, and completed the project. Include all relevant materials (e.g. code)

2) Include and describe a visualization from the project. The visualization should highlight a feature or insight from your model. Explain the decisions you made in constructing the visualization.

3) A general manager of a team wants you to discuss the findings of your rebounding project with the team's head coach. Please write a short email to the head coach introducing your project and summarizing one or two key findings from your research.

1) Describe, generally, from start to finish how you approached, executed, and completed the project. Include all relevant materials (e.g. code)

The coding was done using Jupyter notebook in Python. I have described my approach and my execution of the project briefly below:

Step1: Data Exploration & Research

I spent the first few days exploring and researching the datasets and the data definition sheet provided to gain a deeper understanding of the fields and columns. This gave me a very good idea of what data is presented to me and how can I use this data to build the model.

Step2: Explanatory Data Analysis (EDA) – Pandas

After spending some time researching the data, I imported the data and did some EDA using pandas. This helped me to identify the number of features and attributes, distribution of each columns and how many null values were present in the data.

Below are the few steps that I did part of EDA:

- Checking the shape and datatypes of the datasets
- Checking for missing values in the dataset
- Histogram plots to identify the distribution of the columns

Step3: Data Preparation and Preprocessing

This is the task that took majority of my time during this project. Since we had multiple datasets and a huge number of missing values data cleaning was imperative before I proceeded to the further steps.

Approach – Data Preparation:

I want to merge these datasets to utilize both location data and play by play data to make predictions.

- Merge *testing_data_loc.csv* with *testing_data_pbp.csv*
- Merge *training_data_loc.csv* with *training_data_pbp.csv*

3.1 Handling missing values - Imputation process:

Training dataset:

- Since it contained a huge number of missing values, imputing with mean or median would have introduced a bias when running the algorithm. Therefore, Missing values were removed from the dataset.

Testing dataset:

- Mean imputation was done to the testing dataset since most of the columns had a normal distribution.

3.2 Datasets with huge missing values:

a) testing_data_loc.csv:

Since in the project deliverables it was mentioned that predictions were expected for each play (*playbyplayorder_id),* I did not want to remove any of the missing values from the dataset as it will reduce the length of the dataset.

Missing values by each column:

AtShot_loc_x_off_player_1	42689	AtShot_loc_x_def_player_1	42689
AtShot_loc_y_off_player_1	42689	AtShot_loc_y_def_player_1	42689
AtRim_loc_x_off_player_1	57537	AtRim_loc_x_def_player_1	57537
AtRim_loc_y_off_player_1	57537	AtRim_loc_y_def_player_1	57537
AtShot_loc_x_off_player_2	42688	AtShot_loc_x_def_player_2	42688
AtShot_loc_y_off_player_2	42688	AtShot_loc_y_def_player_2	42688
AtRim_loc_x_off_player_2	57536	AtRim_loc_x_def_player_2	57536
AtRim_loc_y_off_player_2	57536	AtRim_loc_y_def_player_2	57536
AtShot_loc_x_off_player_3	42689	AtShot_loc_x_def_player_3	42689
AtShot_loc_y_off_player_3	42689	AtShot_loc_y_def_player_3	42689
AtRim_loc_x_off_player_3	57537	AtRim_loc_x_def_player_3	57536
AtRim_loc_y_off_player_3	57537	AtRim_loc_y_def_player_3	57536
AtShot_loc_x_off_player_4	42689	AtShot_loc_x_def_player_4	42686
AtShot_loc_y_off_player_4	42689	AtShot_loc_y_def_player_4	42686
AtRim_loc_x_off_player_4	57537	AtRim_loc_x_def_player_4	57534
AtRim_loc_y_off_player_4	57537	AtRim_loc_y_def_player_4	57534
AtShot_loc_x_off_player_5	42714	AtShot_loc_x_def_player_5	42720
AtShot_loc_y_off_player_5	42714	AtShot_loc_y_def_player_5	42720
AtRim_loc_x_off_player_5	57557	AtRim_loc_x_def_player_5	57565
AtRim_loc_y_off_player_5	57557	AtRim_loc_y_def_player_5	57565

b) training_data_loc.csv

Missing values by each column:

AtShot_loc_x_off_player_1	39798	AtShot_loc_x_def_player_1	39789
AtShot_loc_y_off_player_1	39798	AtShot_loc_y_def_player_1	39789
AtRim_loc_x_off_player_1	54265	AtRim_loc_x_def_player_1	54256
AtRim_loc_y_off_player_1	54265	AtRim_loc_y_def_player_1	54256
AtShot_loc_x_off_player_2	39791	AtShot_loc_x_def_player_2	39792
AtShot_loc_y_off_player_2	39791	AtShot_loc_y_def_player_2	39792
AtRim_loc_x_off_player_2	54257	AtRim_loc_x_def_player_2	54258
AtRim_loc_y_off_player_2	54257	AtRim_loc_y_def_player_2	54258
AtShot_loc_x_off_player_3	39789	AtShot_loc_x_def_player_3	39789
AtShot_loc_y_off_player_3	39789	AtShot_loc_y_def_player_3	39789
AtRim_loc_x_off_player_3	54256	AtRim_loc_x_def_player_3	54256
AtRim_loc_y_off_player_3	54256	AtRim_loc_y_def_player_3	54256
AtShot_loc_x_off_player_4	39790	AtShot_loc_x_def_player_4	39791
AtShot_loc_y_off_player_4	39790	AtShot_loc_y_def_player_4	39791
AtRim_loc_x_off_player_4	54257	AtRim_loc_x_def_player_4	54258
AtRim_loc_y_off_player_4	54257	AtRim_loc_y_def_player_4	54258
AtShot_loc_x_off_player_5	39813	AtShot_loc_x_def_player_5	39817
AtShot_loc_y_off_player_5	39813	AtShot_loc_y_def_player_5	39817
AtRim_loc_x_off_player_5	54274	AtRim_loc_x_def_player_5	54277
AtRim_loc_y_off_player_5	54274	AtRim_loc_y_def_player_5	54277

c) training_data_pbp.csv

- F.oreb column: 187834 missing values
- reb_player_id: 205556 missing values

Upon inspecting this dataset, I found that the indicator for the offensive rebound column and the Rebound player id column for *"Made Shots"* was also missing.

F.oreb column: Introducing a third class in the dataset *NRO – No Rebound Opportunity* for all the rows with shot = 'Made shot' seemed like a meaningful data imputation process.

3.3 Feature engineering

- Merging the play by play and location datasets resulted in having a huge number of columns in the resultant dataset. Our goal is to identify important meaningful features and try to reduce the number of features as much as possible before building the model.

Some important features created using existing data for this model:

Feature/ Column created	**Meaning**	**Calculation**	**Usage in model**	**Dataset used**
ORB/g – Player 1 – Player 5 in Offensive and Defensive team	Offensive rebounds per game	Off_rebs/games	Used in place of player id in the final dataset	player_reb_data.csv
DRB/g - Player 1 – Player 5 in Offensive and Defensive team	Defensive rebounds per game	def_rebs/games	Used in place of player id in the final dataset	player_reb_data.csv
Shot_distance	Distance from where the shot was attempted	Distance between rim (-41.75,0) and the Atshot_loc (x,y) coordinates for the shooter id	Used in place of shooter id in the final dataset	training_data_pbp.csv testing_data_pbp.csv training_data_loc.csv testing_data_loc.csv

Dist_from_rim - Player 1 – Player 5 in Offensive and Defensive team	Distance of each player from rim when the ball hits the rim	Distance between rim (-41.75,0) and the AtRim_loc (x,y) coordinates for each player id	Used in place of player id in the final dataset	training_data_pbp.csv testing_data_pbp.csv training_data_loc.csv testing_data_loc.csv

3.4 Feature Transformation

- All the categorical values in the dataset was converted to binary values using OneHot-encoding technique to make it suitable for model building purposes.

3.5 Dropping unnecessary columns

- After feature engineering and transformation, unnecessary columns were dropped to reduce the number of features.
- Number of features were reduced from more than 70 to 30 after this process.

3.6 Resampling data

- Since the prediction class variables were skewed and highly imbalanced, I decided to do resampling to obtain a more balanced dataset for training the model.

3.7 Feature Scaling

- MinMaxScaler was used to normalize and scale the data for model training.

Step 4. Model Selection and Training:

Model Selection:

Considering the length and features of the dataset, I have used Random forest and XGBoost decision tree classification models to train the test dataset and make predictions. Logistic regression, K-NN and SVM models were also considered but were later dropped considering performance issues.

Model Training:

70-30 split was used to on the training dataset to see the model fit on known dataset before using for predictions on the unknown test dataset.

Model Tuning:

Grid search and cross validation was used to find the best model parameters and performance tuning of the model.

Step 5: Prediction

- After finding the best model parameters, the model was predicted using both XGB and Random forest algorithms.
- The results are stored in an csv file for evaluation as requested in the project deliverables.

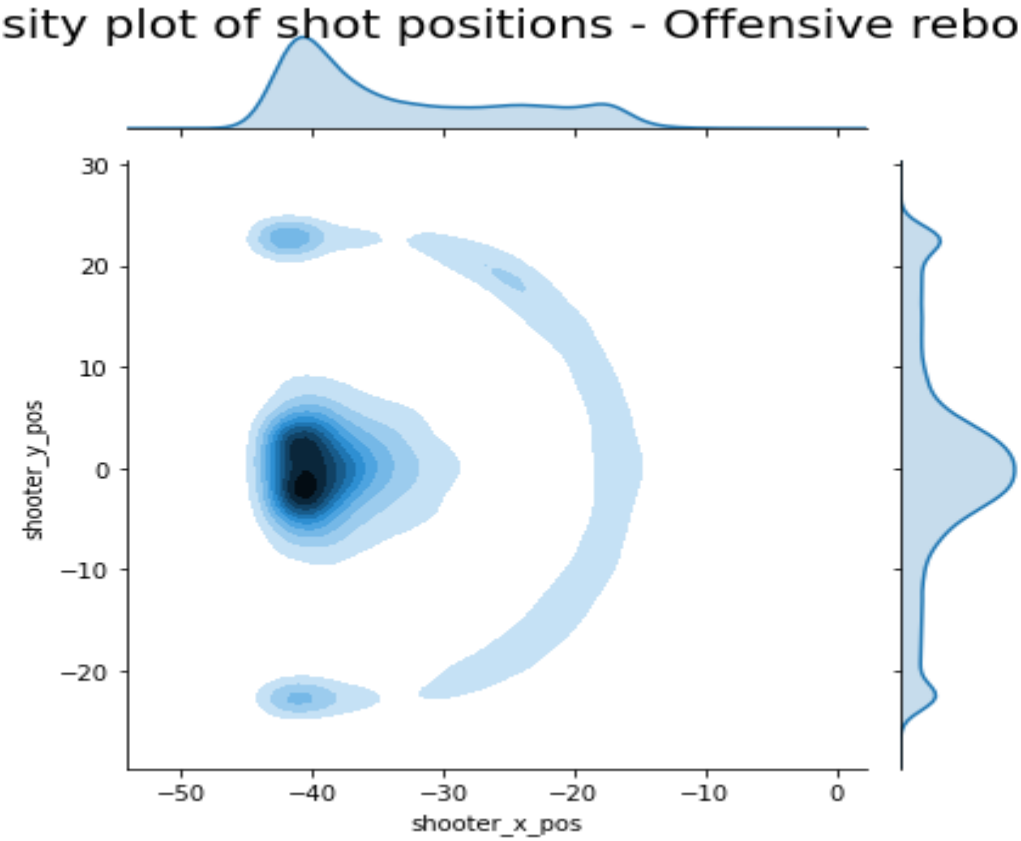
2) Include and describe a visualization from the project. The visualization should highlight a feature or insight from your model. Explain the decisions you made in constructing the visualization.

Visualization 1 – Effect of Shot attempt distance on rebounds:

Approach:

The idea was to see the effect of shot distance on offensive and defensive rebounds. I used a density pot to determine the area where the most number offensive and defensive rebounds were taken. The result was not so surprising as shots attempted closer to the basket and missed were offensively rebounded more compared to longer field goals.

Offensive rebounds:

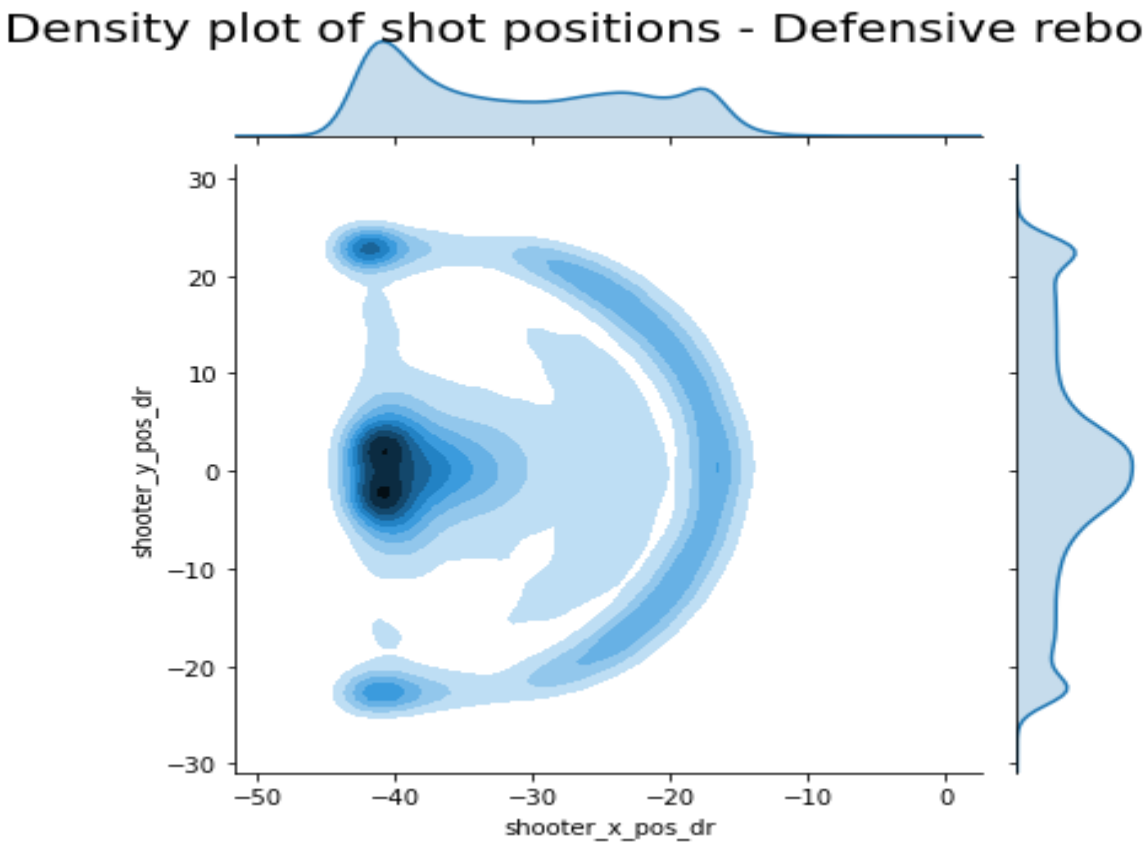


Conclusion:

Attempting a shot in the paint (such as a layup) provides more chances to offensively rebound the ball.

Defensive rebounds:

Even though a greater number of defensive rebounds are taken near the basket, we can also see the number of rebounds taken from shots attempted from the 3 point line is much higher compared to the previous chart.



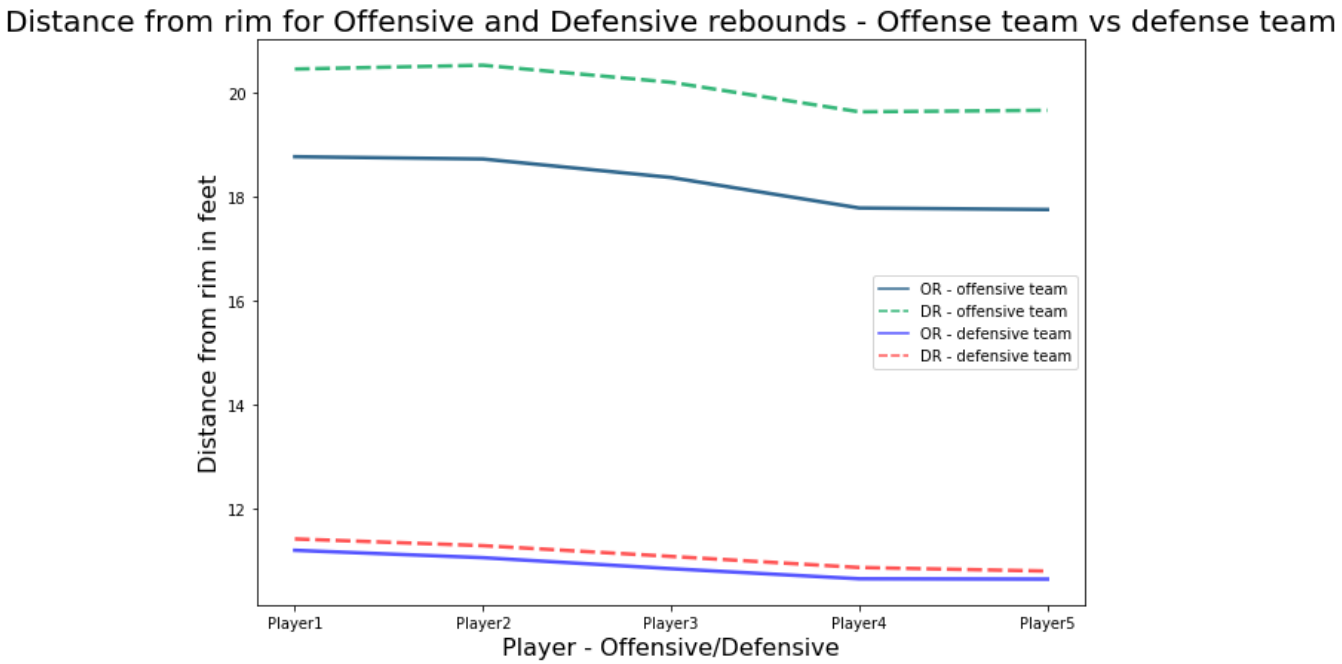
Conclusion:

The greater the distance of the shot attempt, greater the chance of it being defensively rebounded.

Visualization 2 – Effect of Player rebound positioning for an attempted shot:

Approach:

The players from both teams position themselves for the rebound for each play. Hence, the idea was to see if the position of players had any impact on the number of offensive or defensive rebounds collected by the team.



- This graph gives an interesting insight as to position or distance of players for both offensive and defensive team from the rim.

Defensive team:

- The defending teams, on an average stayed between 10 – 12 feet from the rim to collect the rebound irrespective of offensive or defensive rebounds. There is no visible difference between the 2 lines.
- The teams on the offensive side on an average stayed between 18 for offensive rebounds and 20 feet from the rim for the defensive rebounds.

Conclusion:

The offensive players were 2 feet closer to the rim when collecting an offensive rebound compared to defensive rebounds.

3) A general manager of a team wants you to discuss the findings of your rebounding project with the team's head coach. Please write a short email to the head coach introducing your project and summarizing one or two key findings from your research.

Email:

Hello Head Coach,

Hope you are doing great. The general manager of the team wanted me to get in touch with you regarding a project (Predicting Offensive rebounds) that I have been working on for the past couple of weeks. I want to take this opportunity to introduce and discuss some interesting findings with you.

Project Abstract: The goal of the project is to Predict the chance that an Offensive rebound will be taken in a play, using play by play and player location data.

PFA – I have attached a detailed report of my findings and research.

I have a couple of insights that I wanted to share with based on my research which I found interesting. Couple of the research points was to see the effect of shot distance of the attempted shot on offensive rebounds and the effect the player location on the court when the ball hits the rim after the attempted shot. Please find a brief summary below:

Effect of player position during shot attempt: There was a significant difference in the player positioning when collecting offensive and defensive rebounds. While collecting offensive rebounds the players were on an average 2 feet closer to the rim compared to collecting defensive rebounds.

Effect of shot attempt distance: Taking the shot closer to the rim (inside the paint) provided more offensive rebounding opportunities compared to long field goals.

While I really enjoyed working in this project, I also wanted to bring my findings to your knowledge and see if we can implement these findings to improve our offensive rebound rate and increase our possession. I would love to discuss about my project and findings in detail. Looking forward to hearing from you.

Regards,

Guru Prasad Kumar

Thanks