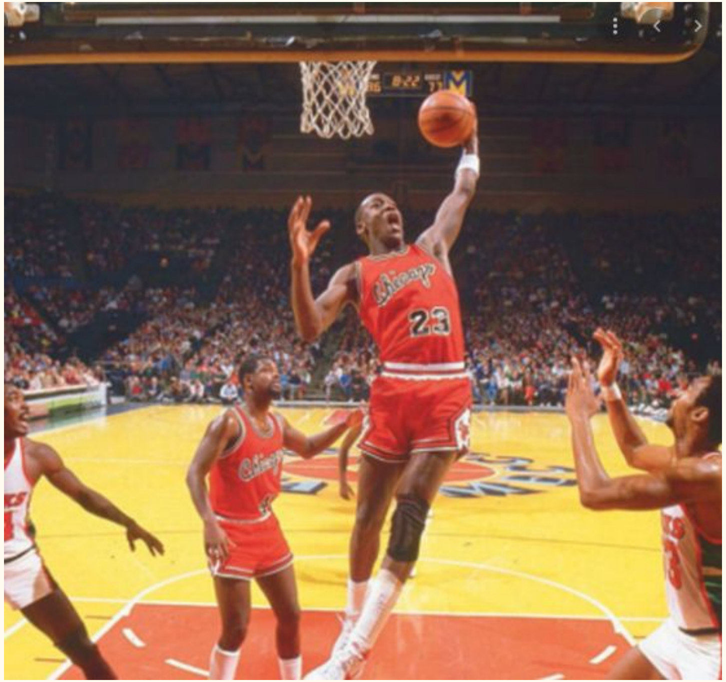
Predicting NBA Center Rebounds

Mario Hage



Introduction

This project is an attempt to answer the below questions, and to discover answers for currently unknown questions.

- Can compiling individual game data from all the NBA Centers in the league tell us something that can be used to our advantage whilst sports betting?
- What hints at players having big games?
- What hints at players falling short?
- Can I come up with data points that can be fed into a model that predicts accurate enough to profit whilst sports betting?
- What is the significance of a player or team “being on a roll?”
- What about the opposite of “being on a roll?”

Data

nba_api : https://github.com/swar/nba_api

Distance data: https://www.rostrum.blog/2018/12/24/nba-travel/

The nba_api contained player box score history, which was filtered for centers. It also contained the Match-Up column, which was split into Team & Opponent; later joined with the Distance data on Start/End Routes.

Data

(join process)

Nba_api original column

MATCHUP
LAC @ OKC

Nba_api (after data wrangling)

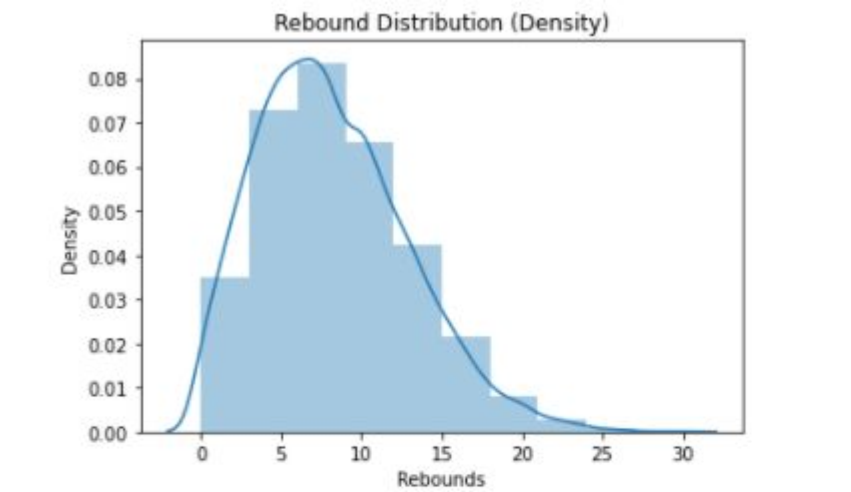
Home/Away	PreviousHome/Away	PreviousOpp	start	end
Away	Away	HOU	HOU	OKC
Away	Away	CHA	CHA	HOU
Away	Away	TOR	TOR	CHA
Away	Home	NYK	LAC	TOR

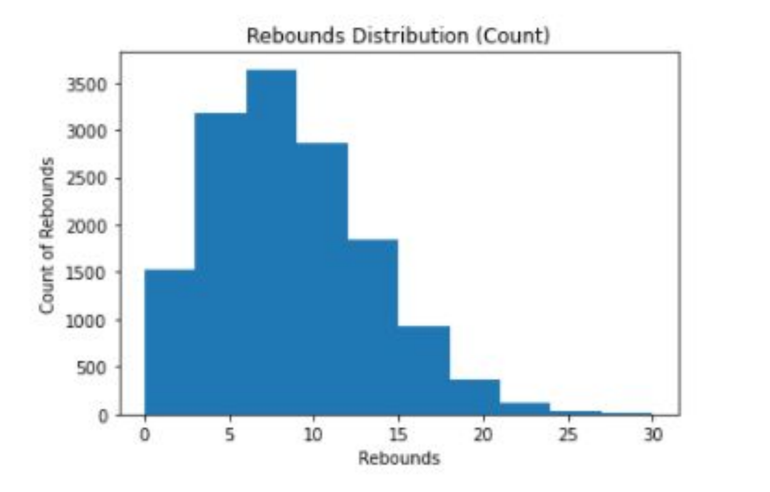
Left Join On Start and End Columns

Distance Data

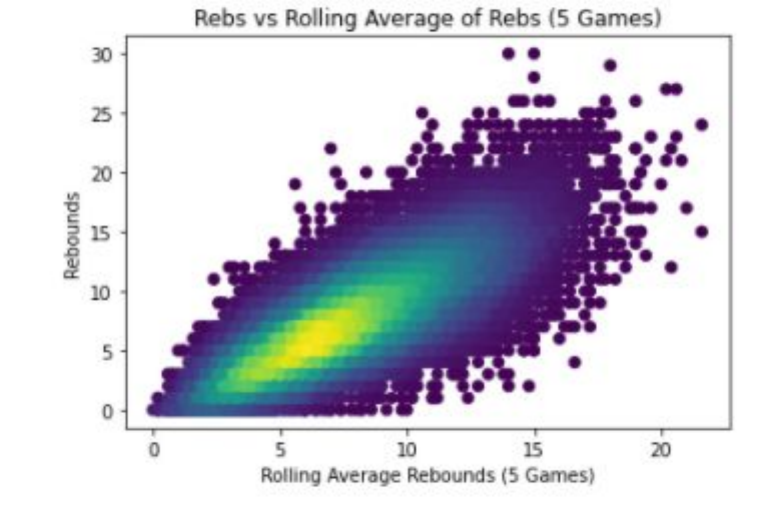
Start	End	Duration (mins)
MIA	POR	3459
POR	MIA	3457
GSW	BOS	3311
BOS	GSW	3307
POR	BOS	3302

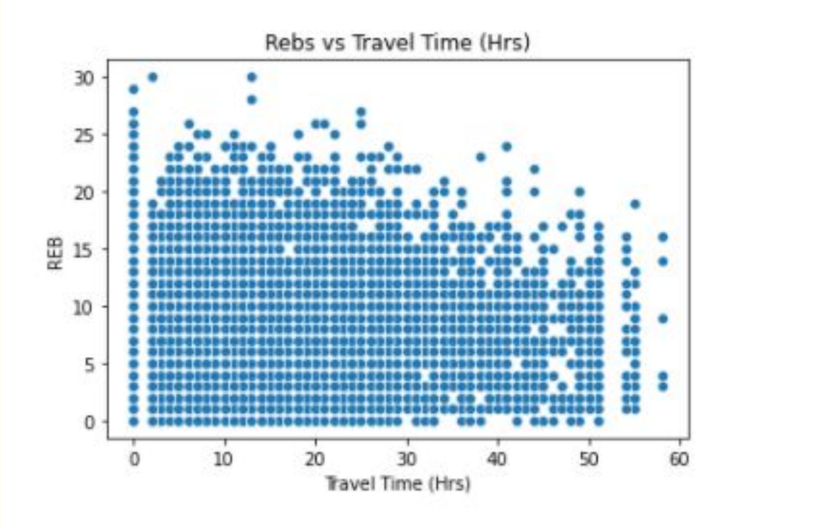
Rebounds Distribution



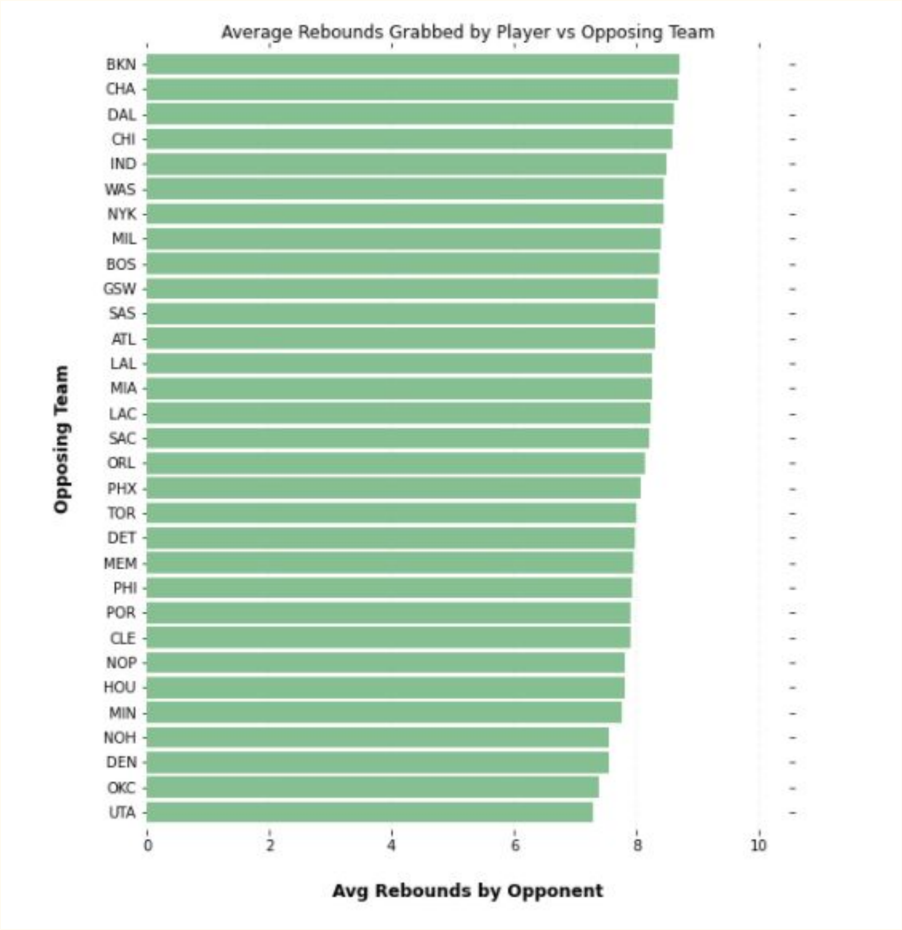


Rebounds vs Rolling Average Rebounds & vs Travel Time

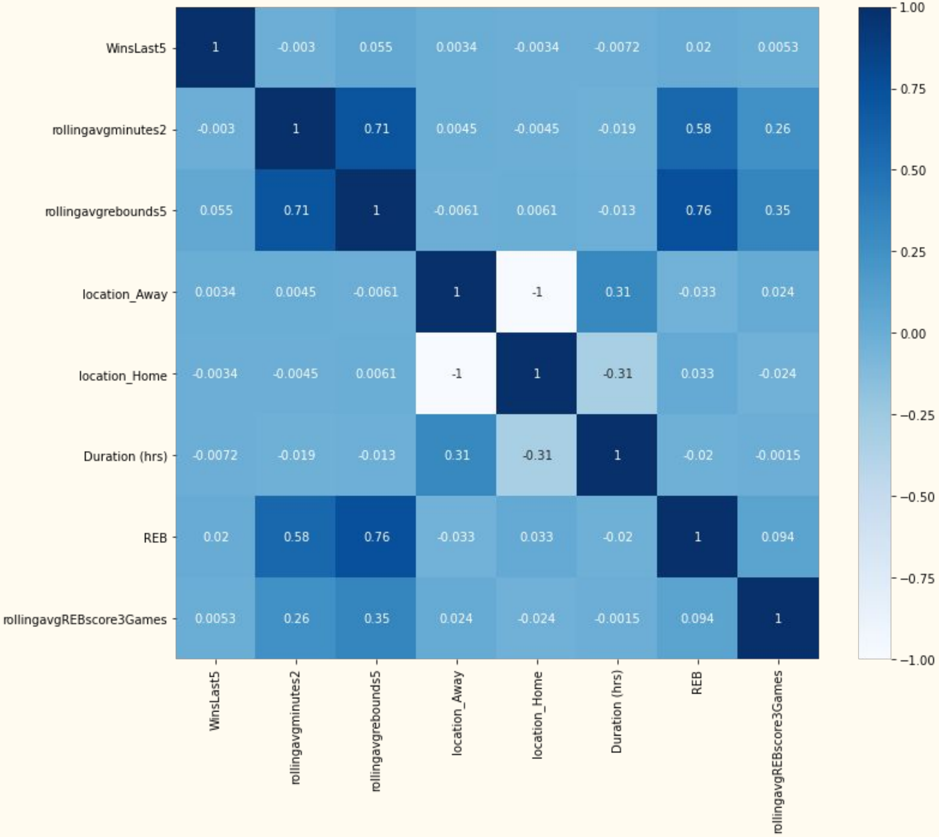




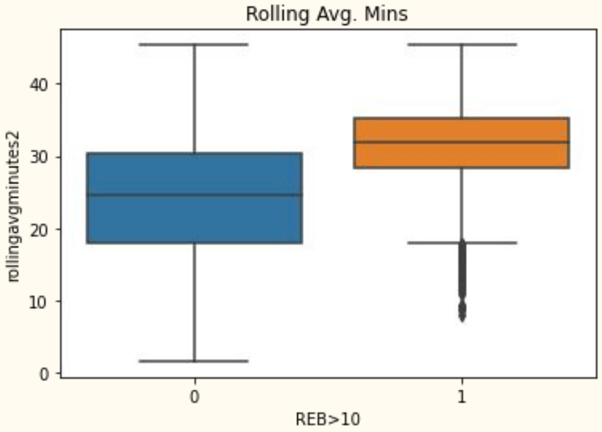
Average Player Rebounds by Opponent

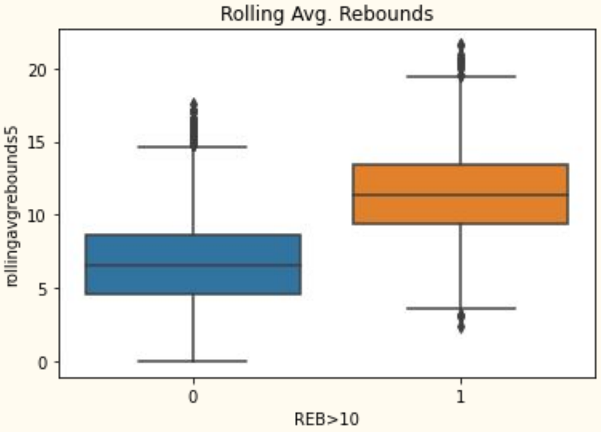


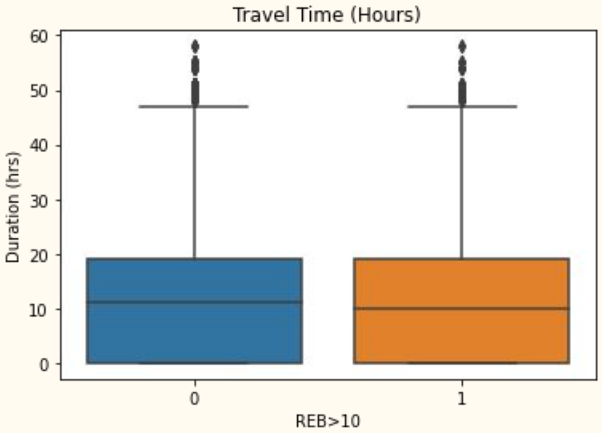
Feature Correlation

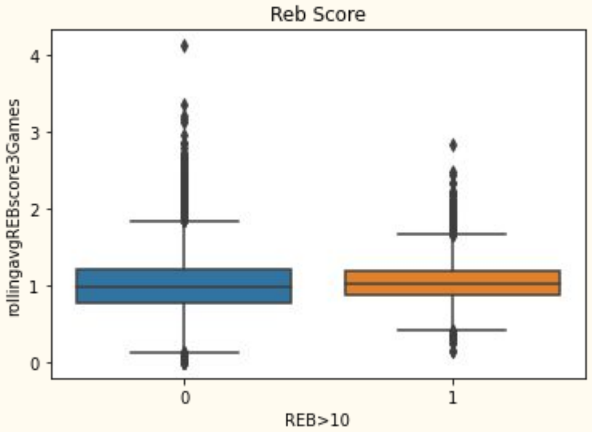


Feature Target Analysis









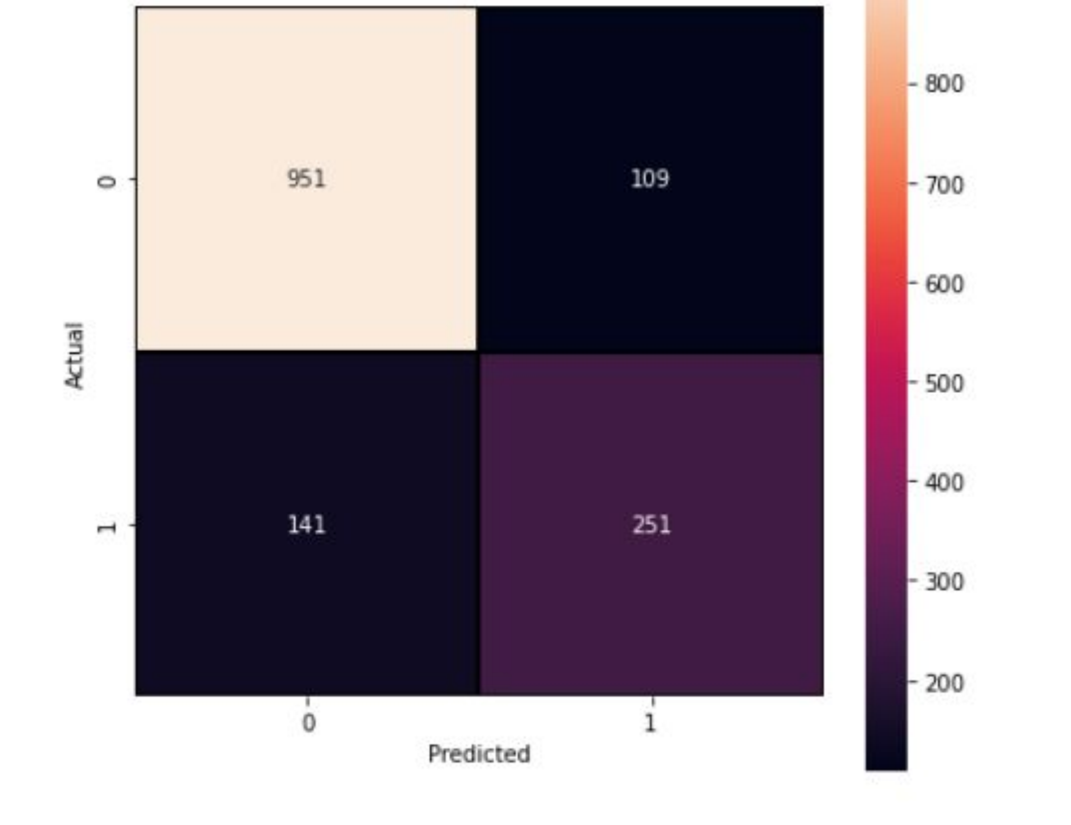
Modeling

Grid Search CV was used on all models to find optimal parameters

Model	Params	AUC Score	Precision	Recall	F1
Random Forest	criterion: entropy, max_depth: 8, n_estimators: 500	0.87	0.78	0.75	0.76
KNN	n_neighbors: 24	0.86	0.78	0.75	0.76
Logistic Regression	C: 0.05179	0.76			

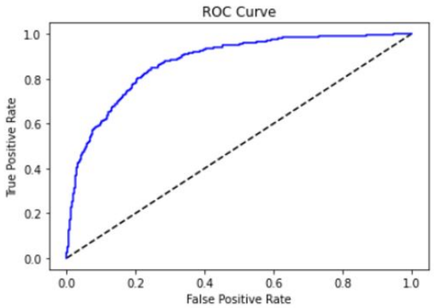
Modeling - RF Confusion Matrix

All of the models performed worse at accurately classifying scenarios where a player grabs more than 10 rebounds, however the RF model was the most accurate out of the 3



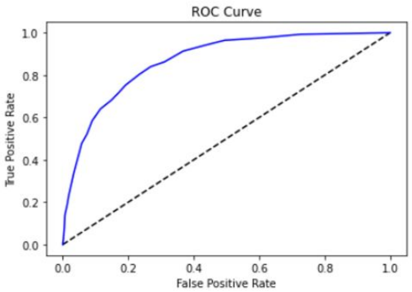
Modeling - ROC_AUC curves

Random Forest



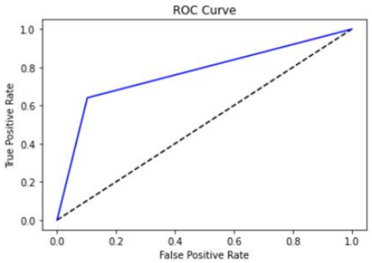
ROC_AUC: 0.8748315363881402

KNN



ROC_AUC: 0.8681772237196765

Logistic Regression



ROC_AUC: 0.7687379668848672

Conclusion

Mario Hage

Takeaways

- Random Forest Model is the optimal choice between the 3 models. All 3 struggle at predicting positive classifications, however the RF Model can accurately predict Negative classifications (91%)

- Model can be utilized to bet under scenarios

- EDA portion provided useful notifications/alerts

Looking Forward

- Spend time collecting data/ quantifying more features

- Integrate betting odds

- Test and optimize