EXPERIMENT REPORT

Student Name	Anton Domini Sta. Cruz
Project Name	NBA Analysis - Tenure
Date	
Deliverables	stacruz_anton_X_week2_EDA.ipynb stacruz_anton_X_week3_Model_Crude.ipynb stacruz_anton_X_week4_modeling.ipynb https://github.com/adcstacruz/nba_uts

---

1. EXPERIMENT BACKGROUND	
Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach.	
1.a. Business Objective	Explain clearly what is the goal of this project for the business. **Understand if a player would stay longer for more than 5 years in the NBA.** **This could be done through building a model that estimates the longevity of an nba player based on the player stats.** How will the results be used? **From a team management standpoint, insights from these can be used for roster building / player selection.** What will be the impact of accurate or incorrect results? **Given the information that I have right now, I don't have any answer. However, this depends on the use case.**
1.b. Hypothesis	Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it, **Through feature engineering, specifically, adding a feature Player Efficiency Rating (PER) and feature selection would result in a less overfitted model.**

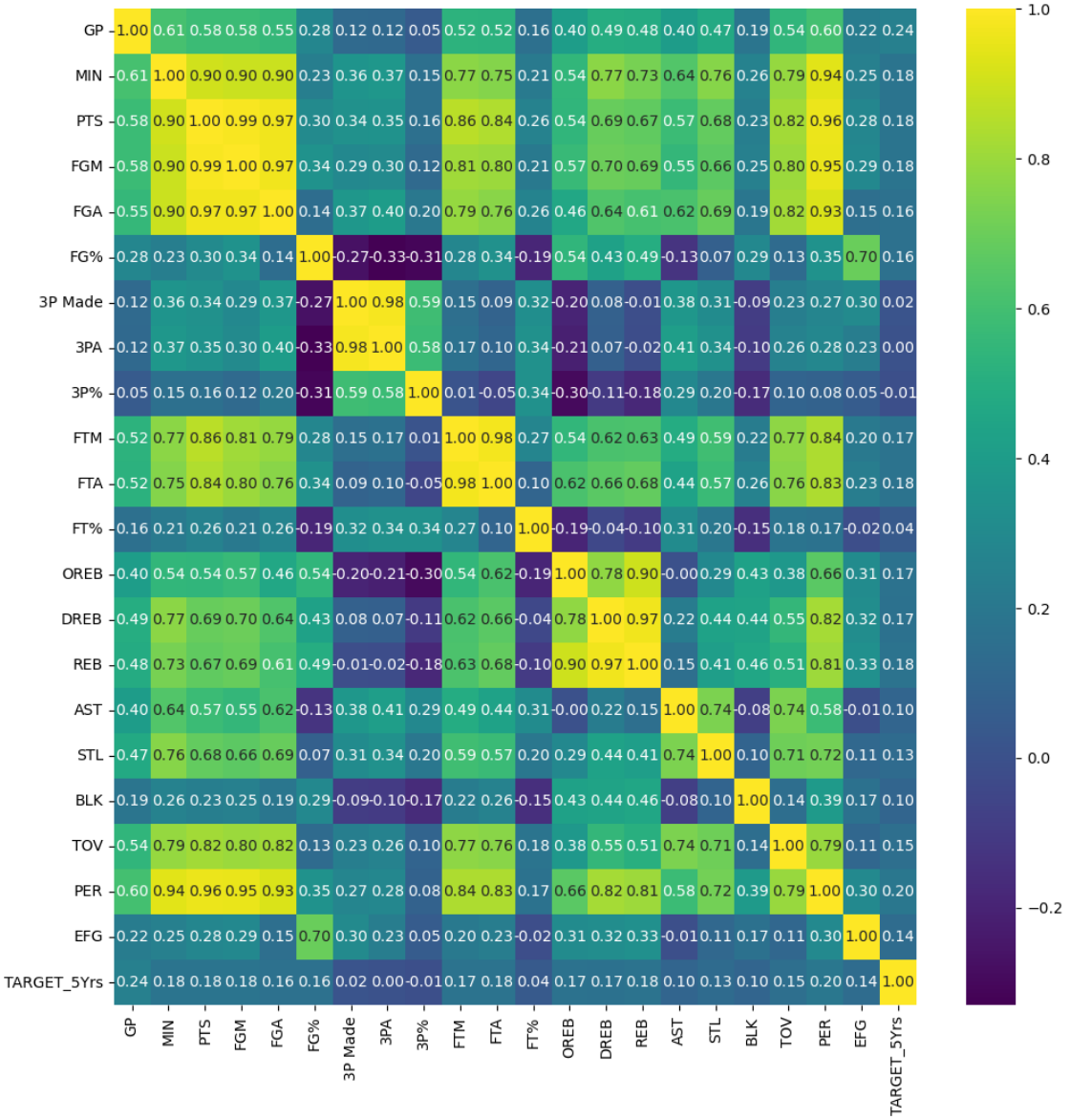
1.c. Experiment Objective	Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment. **Two RF models will be compared: Tuned RF model using all given features (RF1) and tuned RF using selected features and PER (RF2).** **Goal is to compare the gap between the training and test ROCAUC from 5-fold CV experiments.**

---

2. EXPERIMENT DETAILS	
Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them.	

2.a. Data Preparation	Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments
	
	1. **Performed EDA**
	  a. **Check the data quality**
	  i. **Null values**
	  ii. **Bad data**
	  iii. **Outliers**
	  iv. **Distribution**
	  b. **Check the correlated features**
	  c. **Explore the feature space through visualization – using t-SNE**
	2. **Searched for new possible features**
	  a. **Explored indices used in NBA player analytics**
	  b. **Understood the feasibility of usage given the data that I have and tried executing it.**

2.b. Feature Engineering	Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments
	
	The main processes in feature engineering considered were feature selection and feature generation.
	
	1. **For feature generation:**
	  a. Player efficiency rating (PER) and effective field goal percentage (EFG) were generated from the existing data – having these may increase the stability of the model due to outliers in the other features.
	  b. Combinations of the original features and the generated features were used for modeling.
	2. **For feature selection:** The approach was to initially explore the highly correlated features together with the features generated. Correlation plot is shown below:



Through the process of modeling the features: 'MIN', 'PTS', 'FGM', 'FGA', 'FTA', 'REB', '3PA'

2.c. Modelling

Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments

There are 3 modeling stages that were performed: 1) Building crude models in order to understand baseline performances (week 3 output) 2) more building and comparing more refined models (week 4), and 3) building the final model.

1. **Building crude models: Without feature engineering, models were developed using different algorithms just to understand what would be a good modeling algorithm to work on. Below are the results of the models together with different performance metrics:**

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT
et	Extra Trees Classifier	0.8354	0.6693	0.9946	0.8381	0.9097	0.0508	0.1046	0.1
rf	Random Forest Classifier	0.8341	0.6664	0.9925	0.8383	0.9089	0.0526	0.1076	0.1
lda	Linear Discriminant Analysis	0.8339	0.7056	0.9914	0.8387	0.9087	0.0572	0.1105	0.0
ridge	Ridge Classifier	0.8336	0.0000	0.9996	0.8338	0.9092	0.0028	0.0162	0.0
dummy	Dummy Classifier	0.8336	0.5000	1.0000	0.8336	0.9092	0.0000	0.0000	0.0
lr	Logistic Regression	0.8334	0.6927	0.9983	0.8344	0.9090	0.0096	0.0444	0.0
gbc	Gradient Boosting Classifier	0.8334	0.6947	0.9904	0.8389	0.9083	0.0591	0.1157	0.2
lightgbm	Light Gradient Boosting Machine	0.8298	0.6723	0.9822	0.8405	0.9058	0.0737	0.1175	0.4
ada	Ada Boost Classifier	0.8295	0.6797	0.9884	0.8366	0.9062	0.0340	0.0688	0.1
knn	K Neighbors Classifier	0.8127	0.5691	0.9591	0.8392	0.8951	0.0538	0.0674	0.0
dt	Decision Tree Classifier	0.7391	0.5503	0.8333	0.8507	0.8419	0.0964	0.0967	0.0
svm	SVM - Linear Kernel	0.7309	0.0000	0.8409	0.7558	0.7851	0.0188	0.0304	0.0
qda	Quadratic Discriminant Analysis	0.7054	0.6667	0.7562	0.8734	0.8103	0.1637	0.1727	0.0
nb	Naive Bayes	0.5632	0.6740	0.5334	0.9028	0.6705	0.1352	0.1832	0.0

Table. Performance of different classifiers without fine-tuning and feature engineering.

Evidently, there is no model that would perform best in all of the performance metrics. Also, it is worth noting that the differences of the model performances were very marginal.

Together with the results above, Random Forest (RF) classifier was considered as the modeling algorithm for the next modeling stage since it would be significantly easier to fine-tune. The process is more data-centered rather than model-centered. This means that in this experiment focused on performing feature engineering rather than finding the best modeling algorithm which can be performed later on.

2. **Building fine-tuned RF classifiers: Different RF models were built using different feature sets. The performance considered for optimization is ROC AUC score since the data is imbalanced. Hyperparamteres were determined through 5-fold CV. In addition, F1-score and Accuracy were also monitored. Below are the results:**

Model	Feature Set	Train ROC AUC	Test ROC AUC
RF1	**All raw features**	**0.914**	**0.695**
RF2	**All raw features + PER + EFG**	**0.913**	**0.697**
RF3	**Feature selected features**	**0.898**	**0.703**
RF4	**Feature selected features + PER**	**0.903**	**0.705**

	NOTE: Other metrics such as accuracy, precision, recall, and F1-score are shown in the notebook.

---

3. EXPERIMENT RESULTS

Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified.

3.a. Technical Performance

Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.

The performance improvements are quite marginal, as shown in the performance table below. However, it was seen that removing the correlated features and adding indices decreased the overfitting slightly. By using indices, the model might have become more robust to highly variable and noisy features since this is one way to aggregate multiple features into one.

Model	Feature Set	Train ROC AUC	Test ROC AUC
RF1	**All raw features**	**0.914**	**0.695**
RF2	**All raw features + PER + EFG**	**0.913**	**0.697**
RF3	**Feature selected features**	**0.898**	**0.703**
RF4	**Feature selected features + PER**	**0.903**	**0.705**

In terms of the modeling problem, there are a lot of overlapping features in the feature space as shown in the week3 notebook. It might be possible that there should be more features that need to be gathered like the number of incidents and accidents. There might be other possible features related to demographics.

3.b. Business Impact

Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others)

Given my understanding in the NBA context, I may not have much insight here.

Incorrect results may affect the roster building process, leading to contingencies, such as changing a player, that should be resolved by the team management. This would be more costly than selecting a more expensive player.

3.c. Encountered Issues

List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments.

Unresolved: Understanding the data set further, there may be other player indices that can be considered.

4. FUTURE EXPERIMENT	
Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective.	
4.a. Key Learning	Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end. **There were very marginal improvements. I think that the process was more on the modeling and feature engineering. However, I believe that collecting more data with new features would be more relevant.**
4.b. Suggestions / Recommendations	Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production. **Potential next steps:** ● **Collect more data with new features** ● **Remove noise in the data set** ● **Explore other indices** ● **Note: having time data might matter, since there might be a data drift during the whole duration of data collection.** **Recommendation for production:** ● **[Development]: Store the pipeline, setup CI/CD for new data, web app for interactive sessions** ● **[Monitoring]: Make a monitoring system for performance, data drifts, and model drifts, and etc.**