IS327 – Final Project Report – Evan Chen

Project Name: Exploring NBA Player Statistics and Salaries with Machine Learning

Public Project GitHub Repository: https://github.com/zuyouchen/is-327-proj.

Folder structure:

- */data* contains cleaning scripts and CSV used
- */notebooks* contains the Python scripts for model training and evaluation
- */visualizations* contains the matplotlib visualizations exported as PNGs

Modification to Proposal:

In comparison to the proposal, my final project had one significant change: a clustering model was trained and evaluated rather than a classification model. The project maintained the focus of determining which model (regression or clustering) would perform better on the NBA dataset.

I opted to conduct KMeans clustering rather than Random Forest because I realized that a decision tree ensemble model like Random Forest did not truly fit the dataset due to the continuous label (player salary). Instead, I found it intriguing to investigate the performances of a supervised versus unsupervised model.

Outside of training methodology, I did not initially anticipate the extent of data cleaning necessary. For instance, I had to account for string cash values and rows containing NaN values. I also conducted some pre-training and post-training work with visualizations that are unmentioned in the project proposal.

Research Question

My final research question is the original proposal question with additions or modifications highlighted:

Using a dataset of the 2021 – 2022 NBA season's players and salaries, can I determine if the data better fits a numerically predictive supervised regression model or an unsupervised clustering model?

Methods and Process

My project process began with data acquisition and data cleaning. The original Kaggle dataset is credited in the README markdown document of the project's GitHub repository. After downloading the separate CSV files of player salaries and statistics, I took various cleaning steps (dropping rows, merging data, renaming columns, and changing column datatypes) in a data cleaning script acquire a single, cleaned CSV containing both player salary and performance statistics from the 2021 – 2022 season.

Following the acquisition of this cleaned CSV, I began scripting exploratory data analysis, model training, model evaluation, and model visualizations. This code can be found here: [hyperlink].

Upon importing the cleaned dataset, I defined a list of columns named `eda_columns` that contained some of the interesting numerical columns from the dataset. Following this, I utilized Seaborn to generate a square heatmap of the correlation matrix:

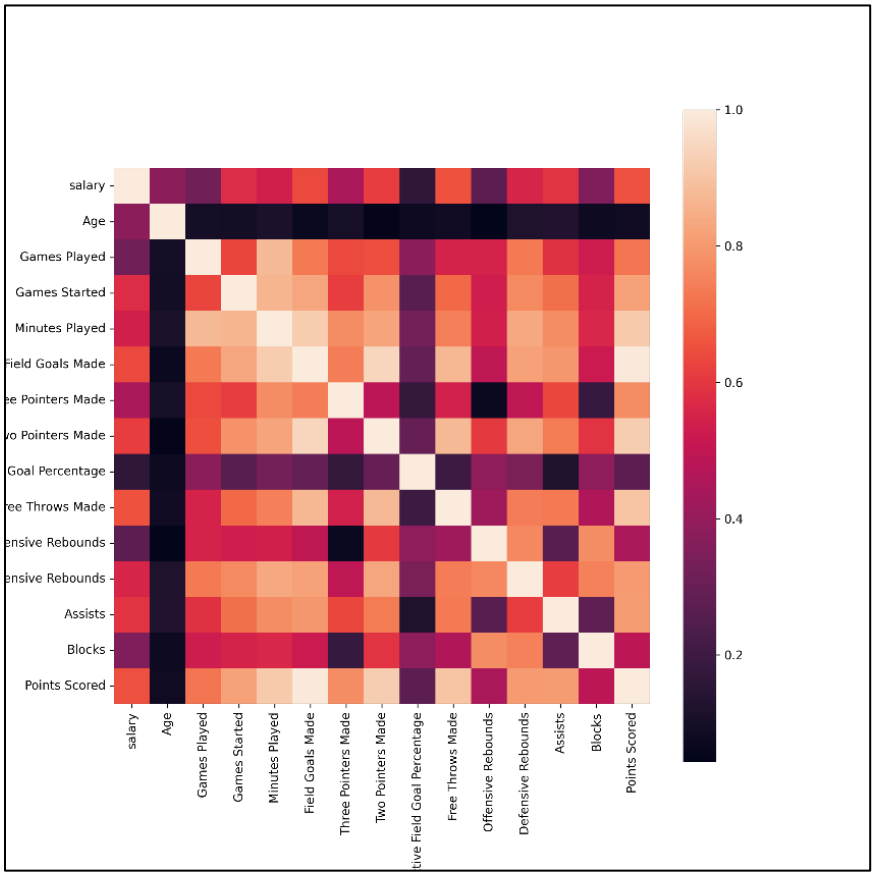

Figure 1: Heatmap of NBA Dataset Correlation Matrix

Observing the heatmap above, I noticed that a fair number of features had significant correlation (>0.5) with the target label (salary). This reinforced my confidence in the prediction of a regression model's success. I then trained a 10-fold cross-validation linear regression model with a list of columns named `numerical_features` that contained all numerical columns except for the label column (salary). I printed out some aggregated training results, recorded in the table below:

Performance Statistic	**Observed Result**
Mean Testing R^2	0.5218547421849219
Standard Deviation of Testing R^2	0.18505937612905649
Mean Training R^2	0.6368914236801234

Table 1: Average Linear Regression Model Performance

I found the "best model" by iterating through results and taking a simple average of test and train R^2. The average of train/test R^2 for this best model was approx. **0.688495**. With this model, I populated a column of numerical predictions in the DataFrame. I generated a scatterplot of the observed salary values alongside the predicted values and a red line at y=x:

Figure 2: Regression Model Predictions

I then utilized the best model to generate single predictions of the salary of multiple popular NBA players by extracting a single row matching a player name from the original DataFrame. Here is a table of some model predictions alongside real salaries:

Player Name	**Predicted Salary**	**Real Salary**	**Difference (R – P)**
Stephen Curry	$31,541,796	$45,780,966	- $14,239,170
LeBron James	$37,738,212	$41,180,544	-$3,442,332
Giannis Antetokounmpo	$41,620,874	$39,344,900	-$2,275,974
Kyrie Irving	$15,187,757	$35,328,700	+$20,140,943
DeMar DeRozan	$32,690,290	$26,000,000	+$6,690,290
Zach LaVine	$24,248,323	$19,500,000	+$4,748,323
Nikola Jokic (MVP)	$36,457,294	$32,480,000	+$3,977,294

Table 2: Player Salary Predictions

Next, I began taking steps towards the training, evaluation, and visualization of a KMeans clustering model of the NBA Salary and Statistics data. I imported the dataset again to reset any changes to the data from regression. I generated a list of all numerical columns (including salary) and used sklearn StandardScaler to normalize the numeric data.

Once the data was transformed, I employed the elbow method to determine the optimal number of clusters. I populated a list containing the sum of squared errors (SSE) values for

KMeans models containing between 1 and 10 clusters fit to the NBA data. Then, I plotted these SSE values against the number of clusters for each value of k, determining the "elbow point" of the graph to be at k=3 clusters:

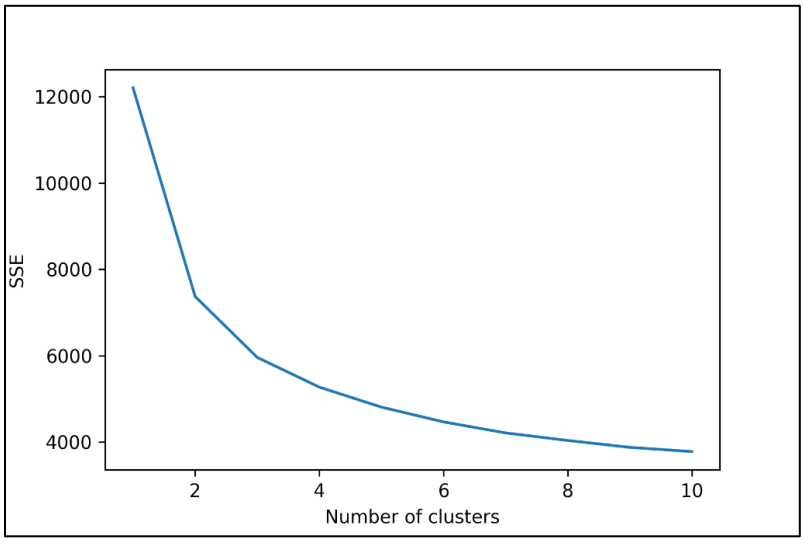

Figure 3: Clustering Elbow Method SSEs

Moving forward, I fit a KMeans model with three clusters to the scaled numerical data. I added a column `cluster` to the DataFrame indicating the cluster that the model classified each row (player) into. For future evaluation, I also output the silhouette score of the model: **0.258848**.

Finally, I generated three scatterplot visualizations where two dataset features would be represented by the two axes' values and the color of each plot point would represent the cluster each instance belonged to:

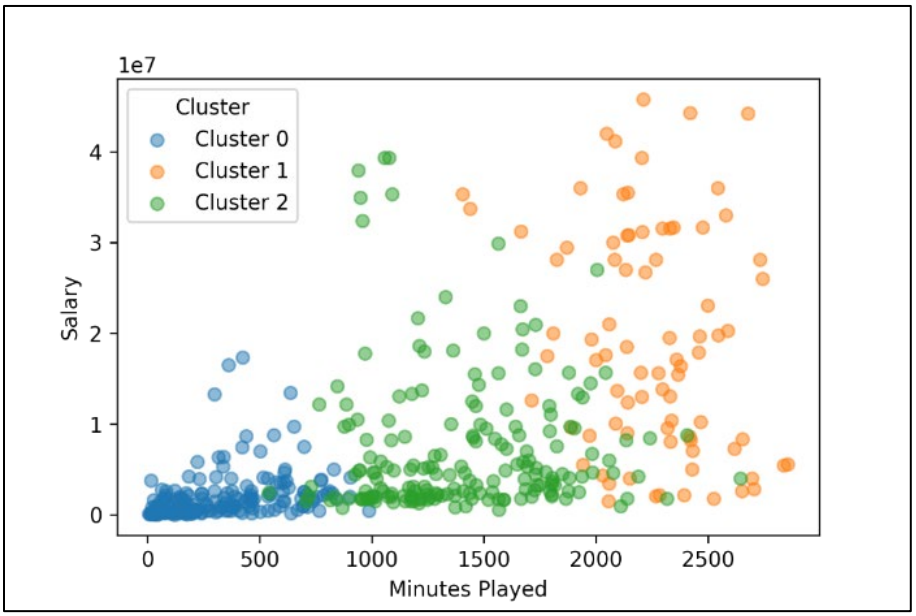

Figure 4: Clustering Scatterplot - Minutes Played and Salary

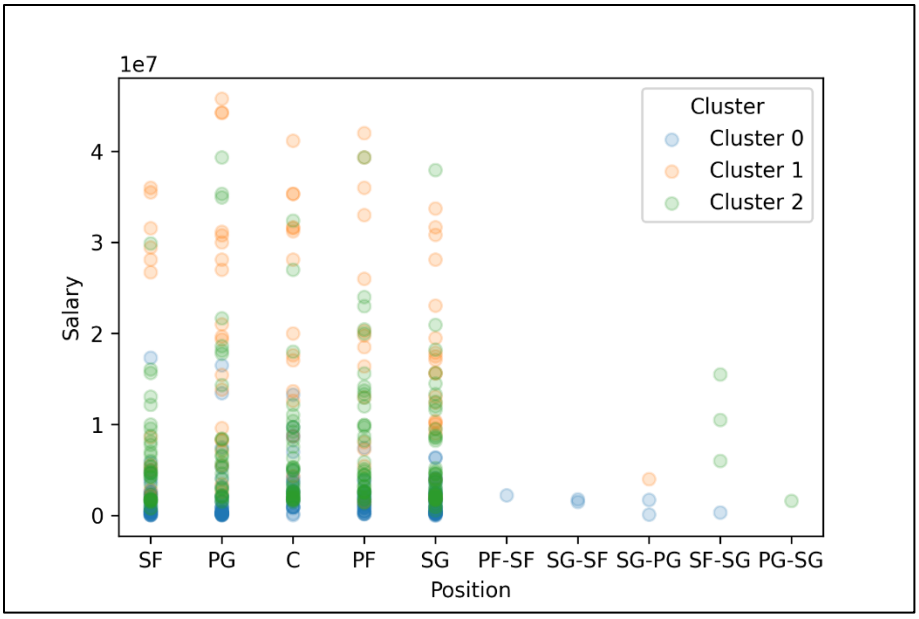

Figure 5: Clustering Scatterplot - Position and Salary

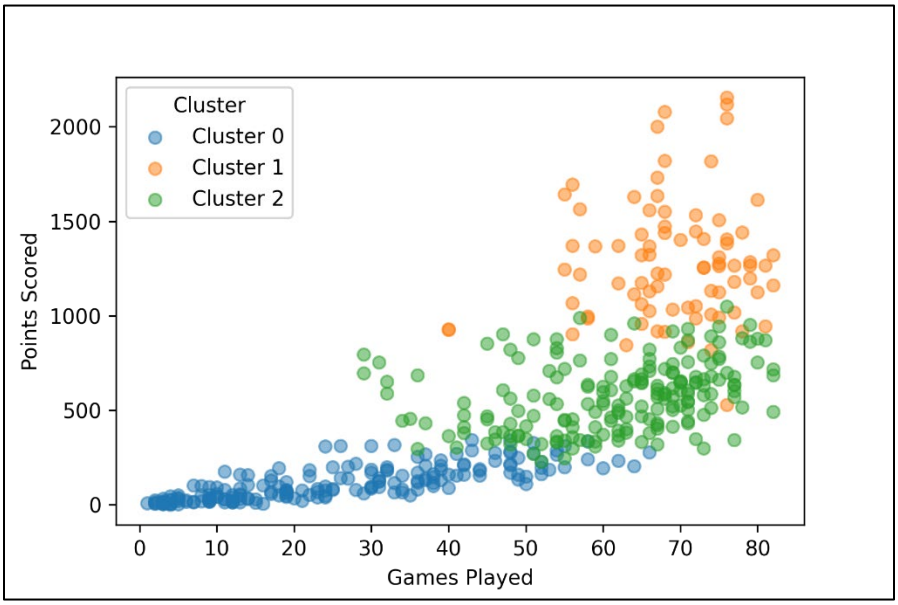

Figure 6: Clustering Scatterplot - Games Played and Points Scored

Analysis of Results

Prior to completing this project, I hypothesized that a numerically predictive regression model would suit this dataset of NBA players and salaries better than an unsupervised clustering model. Logically, it would make sense that players with better performance statistics receive higher salaries. Generating and observing the correlation matrix heatmap (Figure 1) reinforced this logic.

After completing the project, I can answer with fair confidence that the regression model performed better. Clustering did not produce any extremely meaningful categorical takeaways from the three clusters of players, while regression allowed for a moderately accurate numerical prediction of player salary.

The coefficient of determination (R^2) of the best linear regression model was approximately 0.688. I can interpret this as the model explaining about 68.8% of the data. The average regression model R^2 on the training datasets (0.63) signifies that the model was not overfitting on the dataset.

Observing the scatterplot of predicted salaries and salaries (Figure 2) further demonstrates the quality of the linear regression model. The red line on this scatterplot is y=x, representing the theoretical outcome of a model that could perfectly predict salary. I notice that most of the points are relatively close to this line, suggesting the model is decent at predicting true salary. For the average player, the model seems equally likely over or undershoot predictions. For superstar players on the higher end of true salary (right end of the scatterplot), I observe the model tending to underpredict their salary. This makes sense because franchise-level players add value to a team aside from raw statistics and performance. Investigating the individual player salary predictions (Table 2) amplifies my prior observations. Upon extracting individual rows of players, I observed predictions usually not far off more than an absolute value of five million. Particularly high-end players seemed to give the model a lot of difficulty, such as Stephen Curry. Players like Curry can carry weight in their name or branding potential alone – possibly explaining their high salaries that are inaccurately predicted by the model. The model also grossly underpredicted the salary of Kyrie Irving, a star player offered a large contract who played a mere 29 games in the regular season due to vaccine compliance mandates. This could suggest that features like games played or total points are heavily weighted by the regression model.

As for the clustering model, it produced a silhouette score of approximately 0.258. This value is somewhat low, indicating the data points are not especially well-separated across the three clusters. It suggests that some of our clusters have overlapping data, dampening the effectiveness of distinguishing points between the three clusters. As a more interpretable evaluation of the clustering model, I turned to the generated scatterplots representing various dataset features and being colored by cluster:

- From (Figure 4), I notice the clusters are heavily dependent on the minutes played, as they have clean cluster boundaries perpendicular to the x-axis (minutes played). Unfortunately, this means that the cluster of a player does not provide much insight into a player's salary – merely how much the player plays.
- From (Figure 5), I notice the clustering model does a very poor job retaining any level of cluster separation. It seems that all player positions [PG (point guard), PF (power forward), C (center), etc.] contain a mixture of all three clusters. Thus, the clustering model does not make identifiable groups of players by position.

- From (Figure 6), I notice the clustering model does well in forming three distinct clusters influenced primarily by points scored. This is an interestingly well-modeled relationship by the clustering model but is not particularly informative given that we could only group players into low, medium, and high-scoring buckets from said clusters.

Generally, the clustering model only seems to produce good clusters for continuous linear variables that are better utilized as features for salary prediction by the regression model. Therefore, I can state with fair confidence that this NBA Salaries and Statistics dataset fits a supervised regression model slightly better.

Overall, it makes sense that the results of this project aligned with my hypothesis. Given that salary is a continuous variable and that most of our numerical features are positively correlated with salary, you would expect a fair linear relationship between them and salary. Potential categorical takeaways from clustering such as player position may be too complex to group by pure statistics alone, resulting in semi-poor performance of the clustering model. This may have been a result of the dataset or the nature of NBA statistics. There is no guarantee that player positions suggest different statistics. Even if there was, those relationships may have been overshadowed by the larger relationship between obvious correlations such as games played, points scored, and salary.

Learning and Takeaways

One of the biggest takeaways from this project was solid coding practice. I was able perform some data cleaning and reproduce slight variations of the model boilerplate provided in lectures and. Both these forms of practice are extremely valuable, as I employ similar data cleaning principles and basic machine learning models in my current work and internships.

I was able to come to a solid answer to my hypothesis given my observations of model metrics and visualizations. As I hypothesized, the dataset fits a linear regression model predicting player price well. Despite this straightforward conclusion, I learned it was not easy to compare models utilizing different metrics. The scales of the coefficient of determination and silhouette score were different, and I was not comfortable concluding without some visual analysis. Although the linear model was easier to interpret and produced a better metric, I still believe there is some value to the clustering approach that is difficult to measure. From my analysis, I also take away the fact that the machine learning techniques I applied in this project are novel and do not reflect the full extent or potential of applying machine learning to the dataset.

It will be exciting to take some of these concepts from the course into my leisurely viewership of the NBA playoff games soon. While it is easy to claim that player performance logically reflects in higher salary, it is more fulfilling to back these claims with evidence and numerical values generated by a simple machine learning model. After completing this project, I am inspired to investigate the world of sports analytics further.