Evan Chen

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

Research Question: Using a dataset of the 2021 – 2022 NBA season's players and salaries, can we determine if the data better fits a numerically predictive regression model or a categorical classification model?

Dataset: The dataset used in this project will be one consisting of NBA players' statistics and salary data for the most recent NBA season (2021 – 2022). I have already obtained this data from an existing repository on Kaggle. I have passed through the data and applied some simple cleaning and merging in Pandas to obtain a CSV with only the most recent NBA season's players and salaries at https://github.com/zuyouchen/is-327-proj. In this personal GitHub repository, the file data/cleaned2021-2022.csv is the one I will be using for analysis. It contains various features describing players, from free throw percentage (FP%) to points per game (PTS). The label or target is each player's salary.

Method to Apply / Investigate: I would like to apply multiple machine learning methods practiced previously in the course for this project. One is a regression task and the other is a classification task.

First, I will multi-variate regression analysis to identify if correlations exist between player performance statistics and their salaries. By utilizing cross-validation folds and sk-learn's linear regression package, I can evaluate the performance of a model that tries to predict player salary based on statistics using the coefficient of determination (R^2). It will be multivariate to account for the wide range of features in the dataset.

Following this, I will use random forest clustering to determine if there are identifiable patterns in the dataset that can group players by position or salary ranges. By utilizing cross-validation folds and sk-learn's random forest package, I can evaluate the performance of a model that tries to cluster players into categories (either by salary bucket or position) with Cohen's Kappa.

After applying both types of machine learning, I can utilize both quantitative and qualitative forms of analysis on their respective measures of effectiveness to answer my research question.

Hypothesis: I hypothesize that a numerically predictive regression model will suit this dataset of NBA players and salaries better. Logically, it would make sense that players with better performance statistics receive higher salaries. Thus, the correlation between player performance statistics and salary would be presumably high and simple to predict numerically. Despite this, many confluent factors such as player fame or past performance might change salary – giving credence to the project's inquiry into potential clustering classification models as well.

Learning Goals: As I complete this project, I hope to gain and solidify my existing experience using sk-learn's various packages to perform basic machine learning tasks on real world data. Aside from answering my research questions and practicing my evaluative skills in the domain of machine learning, I think completing this project will supplement my future viewing of NBA games and the NBA draft. If the project is successful, I can use the takeaways to frame my understanding of salaries given to the most popular NBA players.