Project Proposal CS 412

Group members:

Shalin Patel, Qasim Mir and Ansar Khan and Ajitesh Bansal

Project Goal:

We plan on scrapping basketball reference and collecting all NBA players stats for the 2019-2020 season. With these stats we will try to learn if players are worthy of their current salaries by calculating various statistics. We can also perhaps try to predict given a set of stats what a player might be worth from a pure statistical point of view.

Data Collection:

We plan on collecting the data from multiple sources:

- **Basketball reference:** We plan on collecting/scraping data for players' contracts and NBA players stats per game. This will be done for multiple seasons to get a better analysis and predictions
    - https://www.basketball-reference.com/
- **Goldsberry Package:** We plan on using this python package which contains NBA statistics directly related to the data that is provided by the officially on stats.nba.com
    - https://pypi.org/project/py-goldsberry/

Next steps:

- We will first scrape the raw data from the sources provided above
- Then we will apply data processing methods to clean the scraped data and prepare the data for implementing and performing analysis using prodigious machine learning concepts and algorithms
- We iterate multiple machine learning algorithms and iterate those models over our cleaned data and train it.
- The models we obtain from the previous step, can then be tested on the test set to calculate the MSE for each model
- We then chose a model that is compelling as per our requirements and learnings for predictions based on the best MSE
- We can then deploy that particular model to make further predictions that are not already present.

Machine learning Tasks / Techniques

- We want to apply machine learning tasks to our datasets to get models and infer the statistics and predict our labels to calculate the accuracy of those specific models.
- If we can predict on our test data and get good enough mean squared error or accuracy, then perhaps we can go further and use the same models and algorithms to predict the future of certain games, predict the future performances of certain players that is if they are young prospects we can even predict their potential future potential. By using linear regression, this can be achieved.
- More in depth of linear regression:
    - We want to use Linear regression to model the relationship between each player's salary and their game statistics to determine if they really deserve that or they are being overpaid.
    - Also apart from checking if they are overpaid, we can use linear regression to predict their next season salary based on the salary they have been receiving past seasons.
- Over time after analyzing and cleaning the dataset we will add more machine learning tasks if necessary but as of now this is our game plan.