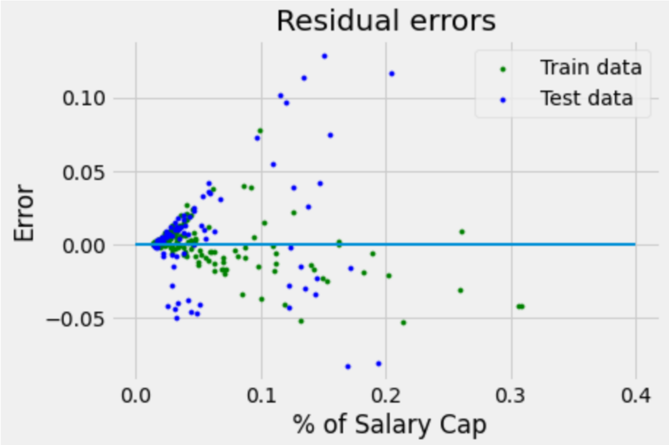
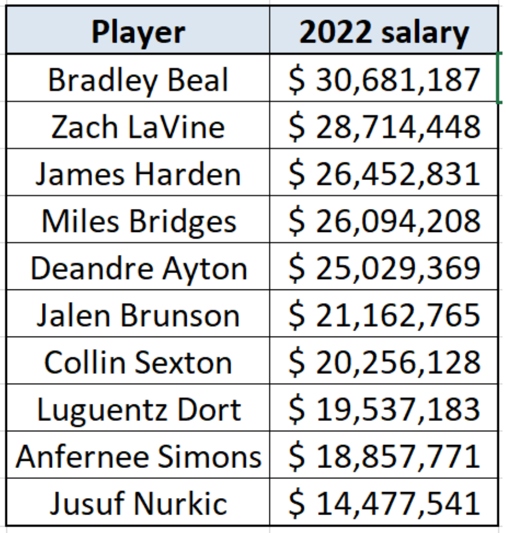
My first step towards completing this project was planning what data I would need and where to access it. I needed player statistics, contract history, and award history. I accessed these at [https://www.basketball-reference.com , https://www.spotrac.com/nba/free-agents , https://basketball.realgm.com/nba/info/salary_cap] respectively. The contract data was only available for free back to 2020 so I had a limited dataset to train the model on. For the statistics, I utilized basic box score stats as inputs to the models.

Once all of the data was scraped, cleaned, formatted and merged, it was necessary to do some planning about model objectives. Since the salary cap changes year to year, predicting the actual dollar amount would not achieve the best results and it would contribute to underestimations since the cap is increasing every season. To avoid this issue, the model was built to estimate the percent of a team's total cap space the contract would take in the first season of the contract. By using the percent of the cap, it allowed for a normalized output that accounted for the differing cap limits. Another issue that presented itself this year were players who didn't play full seasons (Beal) but would be getting large contracts. I tried to counter this by using player per game averages instead of season totals for the stats being fed into the model. It seemed to work out well in the contract ranking order.

I used a linear regression to create a baseline model. It had an **R^2** = 0.48 and a **Mean Absolute Error** = 0.05. This would work well as a baseline model to improve on. I decided on a Random Forest model for my final submission. I used the same data and same input features to train and test the random forest on 2020 and 2021 data. It showed some improvements with an **R^2** = 0.51 and a **Mean Absolute Error** = 0.02. Below is a plot of the residual errors for the random forest model. I think it highlights the biggest issue of the model which is that the larger contracts are not being estimated with high accuracy. The smaller contracts that account for ~1-5% of the cap are predicted with much higher accuracy.



There is still much more room for improvement on this model and I have included my thoughts on the immediate areas of improvement at the end of this report. The high end of contracts are being vastly under valued (Beal, received about 50M a year but my model only estimated 30M; even though it was still ranked as the top contract in my model output). The results from my model are below.



Player	2022 salary
Bradley Beal	$ 30,681,187
Zach LaVine	$ 28,714,448
James Harden	$ 26,452,831
Miles Bridges	$ 26,094,208
Deandre Ayton	$ 25,029,369
Jalen Brunson	$ 21,162,765
Collin Sexton	$ 20,256,128
Luguentz Dort	$ 19,537,183
Anfernee Simons	$ 18,857,771
Jusuf Nurkic	$ 14,477,541

Areas of Improvement:

- The use of advanced statistics as input features
- Incorporate a model to predict the contract length and feed that into the contract $ prediction model as a feature
- Using multiple seasons of data for the statistics instead of just the statistics from the season prior to the contract year
- Use contract data for seasons prior to 2020 to train the model
- Incorporate features that designates players who are eligible for super max, vet, and other contract types
- Exploring more advanced models

The project code is available here:

https://app.box.com/s/pgchtrz6xb3kpo12c7k67cb1ipvcsw27