NBA Salary Prediction

Nayan Jani
njjani@umass.edu

UMassAmherst
Data Analytics and Computational Social Science Program

Motivation

Being a General Manager in the NBA comes with a ton of decisions. One of the most important decisions a GM can make is how much they pay the players on their team. It is so important to pay the players the right amount in order to build the strongest roster. Overpaying a player will hurt a teams cap space, meaning that the team will not be able to sign good players because they do not have enough money to afford them. My motivation for this project is to see if Machine Learning techniques can correctly predict a players' salary. The idea is if I am able to create a model that performs well enough, then it could be used as a tool to determine a players' salary for their next contract. Here I will perform different regression methods to predict players' salary and then use the best method for prediction.

Data & Methods

The dataset I am using comes from Kaggle. The dataset has 199 instances with 29 features. The features contain information about player names, time span of the contract, average salary per year, and all stats that players' accumulated during the NBA season before signing their next contract. The players' stats come from basketball-reference.com. The scope of the data is as follows:
- There are only contracts signed since 2010/2011 season to 2019/2020 season.
- Only includes players that are active in 2020/2021 season.
- Doesn't include rookie or retained contracts.
- Doesn't include contracts for players that haven't played year before the signing the contract.

This is a good scope because I want to use modern players contracts for future predictions. The limitation of only including players that are active in 20/21 means that these players were able to earn multiple contracts within the 10 year span, which validates them as players who are worth to continuing paying. Having this removes players who had massive contracts early in their career and then faded out quickly after their primes.

In this project, I used R and Python. In R, I conducted my exploratory data analysis. Here I created a density plot to check the distribution of my target variable and a correlation matrix to see which independent variables are related to my target variable. In Python, I conducted Grid Search cross-validation with 5 folds in order to tune the hyperparameters for each model using training and validation sets. After finalizing hyperparameter choices, I used the test set as the final evaluation for how each model does. The methods and hyperparameters I selected are listed below:

Method	Hyper parameters
Random Forest Regression	n_estimators, max_features, min_samples_split, min_sample_leafs
Support Vector Machines Regression	C, gamma, kernel
Ridge Regression	Alpha
Gradient Boosting Decision Trees Regression	n_estimators, learning_rate, max_depth

Evaluation Metric: Root Mean Squared Logarithmic Error (RMSLE). I am choosing RMSLE because my predicted and actual values of my target variable (salary) are large integers, so by taking the logs of them will remove any penalization of those huge differences between those values. My target variable has a skewed distribution and a large range, so RMSLE is a good fit because prediction errors of low and high salary will be treated evenly.

References

Current NBA Players Contracts History. (2021, February 8). Retrieved from https://www.kaggle.com/datasets/jarosawjaworski/current-nba-players-contracts-history

Preprocessing & EDA

Preprocessing Decisions:
- Removed the players names, their wins, their loses, their contract start year and contract end year from the dataset.
- Created variable called "c_duration" which is the difference between "contract_start" and "contract_end".
- Split data up : 70% training, 15% validation and 15% test
- Scaled my training, validation and test data using the StandardScaler() function for SVR and Ridge

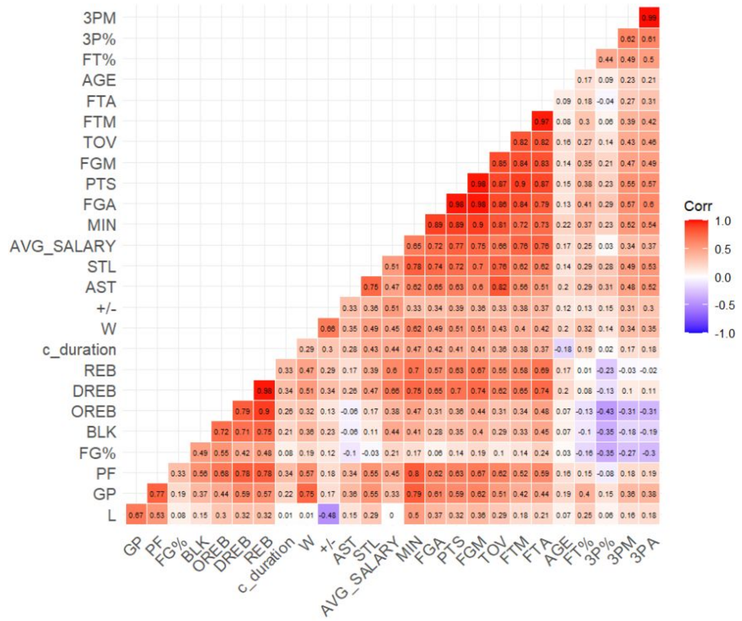

Figure 1: Correlation matrix of all features and the target variable.

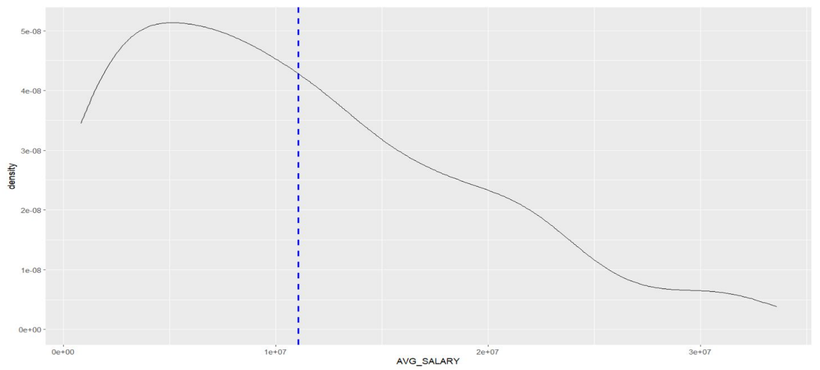

Figure 2: Distribution of target variable, Avg_salary. The distribution is not normal and skewed right. This suggests that more flexible methods will work better.

Results

Below are my results for hyperparameter tuning using Grid Search cross-validation with 5 folds and training, validation and test errors for each method.

Method	Hyperparameter values
Random Forest Regression	n_estimators = 200 max_features = 1.0 min_samples_split = 4 min_sample_leafs = 1
Support Vector Machines Regression	C = 0.1 gamma = 0.001 kernel = poly
Ridge Regression	Alpha = 100.0
Gradient Boosting Decision Trees Regression	n_estimators = 100 learning_rate = 1.0 max_depth = 6

Training error Random Forest (RMSLE): 0.569804892934335
Training error SVR (RMSLE): 1.0135970157480874
Training error Ridge (RMSLE): 0.5579904916871834
Training error GBR (RMSLE): 0.7894211344728965
Validation error Random Forest (RMSLE): 0.46377953249672954
Validation error SVR (RMSLE): 0.6805718407616901
Validation error Ridge (RMSLE): 0.5200462148165859
Validation error GBR (RMSLE): 0.6046343223863266
Test error Random Forest (RMSLE): 0.4913785330380669
Test error SVR (RMSLE): 1.0137954511265557
Test error Ridge (RMSLE): 0.5376660192539188
Test error GBR (RMSLE): 0.7315004460794416

Conclusion

Based on my results, Random Forest Regression performs the best out of all my methods. I believe it performs the best due to the fact that the hyperparameters are dealing with overfitting. The hyperparameter min_samples_split = 4 reduces the number of splits that happening in the decision trees, which shortens the depth of the tree. Having shorter trees allows for the model to generalize better because it will not rely on the structure of the training data as much. Though this method performs the best, it is not the best for interpretability.

Limitations and Ideas for Further Study:
- More data from the 2021-2022 and 2022-2023 seasons that follows the scope of the data discussed earlier would be useful to make my model generalize better for future predictions.
- Add more features that relate to the advanced statistics that are not covered in the base statistics.
- There will always be external factors that have to due with salary market in the NBA that cannot be covered in the model. For example, players personal accolades, such as first team all NBA, All star team, defensive team of the year etc. can greatly increase ones salary if they earn one of these honors.