Using Machine Learning to Find the Best and Worst Value Contracts in the NBA

Introduction

Even the most casual sports fan has found themselves saying “I can’t believe we’re paying this guy so much. He stinks!”. Every fan turns into a professional analyst when discussing their favorite teams. Arguments regarding what players are overpaid and underpaid seem never-ending in sports circles, forums, and social media. In reality, the task of how much to pay their players is dictated by the player’s ability, the current market, and ultimately the team’s general manager. While hindsight might make the value of these decisions clear for some fans, for most contracts, the debate rages on.

The NBA is driven by stars. The best players make the most money, and they, hopefully, carry their team to success making the team owners' investment worth it. However, having one or two well-paid superstars is not a recipe for success. Building a team around your superstars is no easy task and often comes down to finding the right pieces that complement your team for the right price. The real pain of team building then becomes the opportunity cost: Could you have gotten better value for that money instead? Are you getting the best bang for your buck?

We set out to determine what players are over and underpaid using machine learning. The premise is; we use a machine learning algorithm to figure out how much a player should be making and compare that against their actual contract. The difference between these values determines how much the player is overpaid or underpaid. For this task, we’ll create 4 machine learning models to predict a player’s salary based on historical data. Note, this is not meant to be a predictive metric of how much a player’s next contract will be worth. This evaluates expected salary relative to real salary to see who’s overpaid and who’s underpaid.

History of the Cap and Salary Distribution

During the 1984-85 season, the NBA implemented a salary cap model. This cap was much simpler than the current cap system that is subject to a complex system of rules and exceptions. It was not until later that the current rules of rookie contracts, max contracts, free agency, etc. were gradually implemented. However, the main idea has remained the same since the beginning; the salary cap sets a limit to the total amount of money that National Basketball Association teams are allowed to pay their players and exists to promote cost-control and parity among teams. And while this cap or spending limit has increased as both inflation and the league’s financial success have risen, general managers still must divide the pie as best they can to construct a successful team.

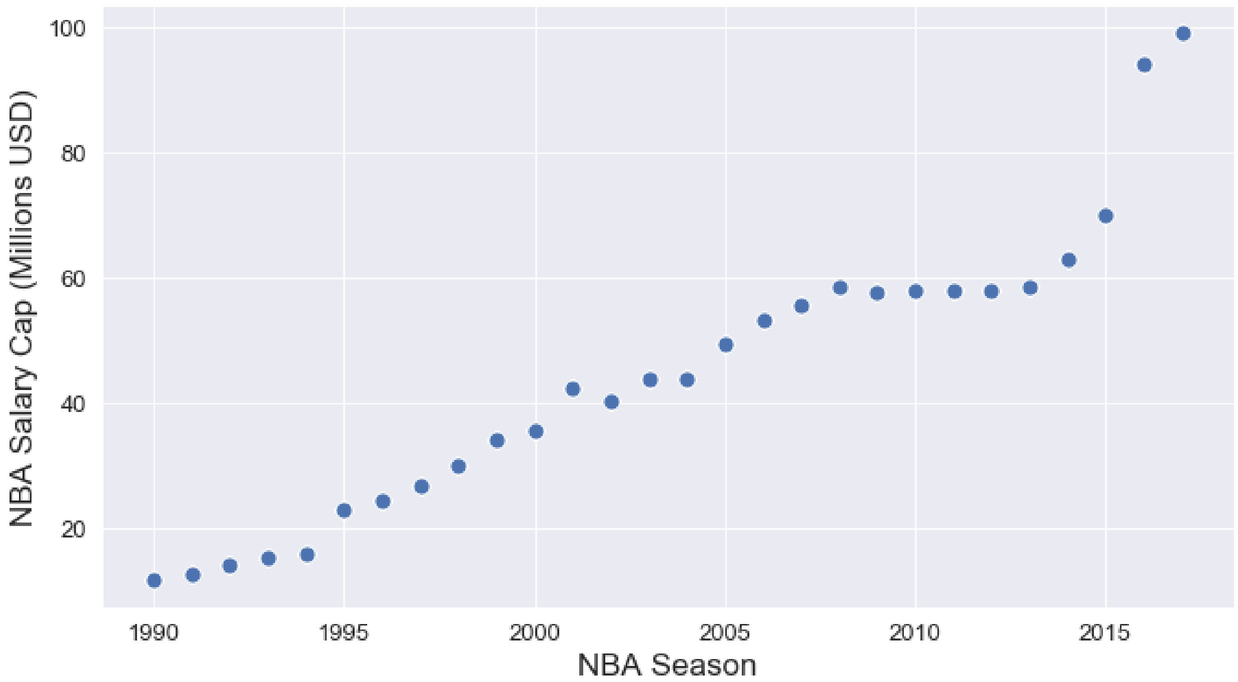

Figure 1. Total NBA Salary Cap in millions since 1990-91 season.

Early during the implementation of the salary cap, owners would divide the pie more or less equally among their players. By the early ‘90s, as the league shifted to being increasingly star-driven, several players started to earn a larger portion of their team’s salary cap. Eventually, the rules for the cap were re-written to include the max contract, dictating the maximum percentage of the cap a single player can make. Before max contracts were implemented, Michael Jordan earned 120% of the total cap for two years after returning from retirement and his stint in baseball to lead the Chicago Bulls to three consecutive championships. Although the rules allow for exceptions, teams exceeding the salary cap are penalized by the league by a luxury tax payment system which is something owners for the most part want to avoid.

Today, due to the current cap rules based on exceptions and restrictions, teams structure their cap room to get star players and contend for a championship. Most contending teams often have a few players on a maximum contract, a couple guys in the $10 million range, and lots of minimums contracts to fill out the roster. Additionally, teams are required to spend at least 90 percent of the salary cap each year. So, if we’re going to predict salary, we must first understand its distribution.

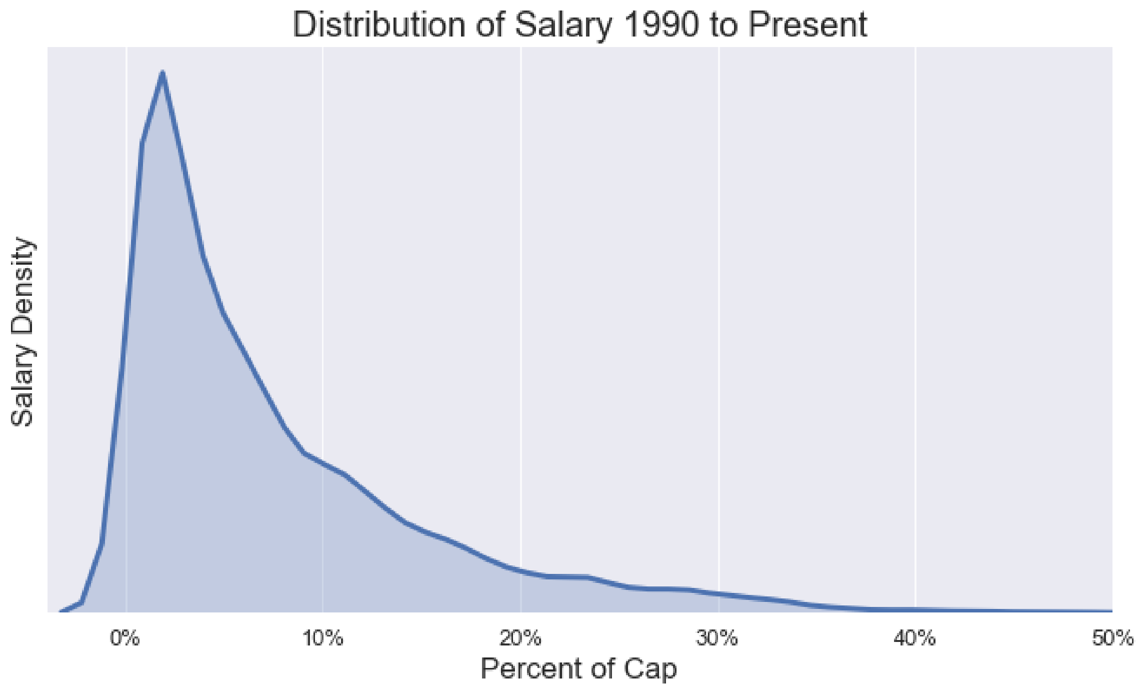

Figure 2. Distribution of percentage of the cap earned by players since the 1990-91 season.

As expected, we can see that historically most players have earned a small fraction of a team's total cap. With 12 to 15 roster spots for a team to fill, most players find themselves in the lower spectrum of the salary range. Given that teams must use at least 90 percent of the cap, teams will usually have one or two players with maximum contracts. For contending teams, these are their star players carrying the team to success. For teams at the bottom of the standings, these maximum contracts can be players taken in return for valuable draft picks, salary dumps from other teams or failed experiments to contend. Because of this, in this exercise we will use a player’s individual metrics and not the team’s success to measure how much they should be paid.

Data

Our dataset starts with the 1990-91 NBA season and spans the period until the 2017-18 season. The NBA cap rules added unrestricted free agency in the 1989 season completely changing the way contracts were handed out to players. This makes the period before that too different to have any predictive power for contracts in the modern era where a player’s first chance at a significant payday comes once they enter free agency. Before entering free agency for the first time, players are on rookie contracts for three years with a fixed, non-negotiable, salary based on their draft position. This makes rookie contracts particularly a good value for players that can positively contribute to the team early on their careers. We will account for this in our model as explained later.

For our dataset we combine a player's salary information, the familiar basketball counting stats, like points, assists and rebounds, along with several modern advanced metrics. This information is present for all players in the league during the seasons from

1990 to 2018. In total the database consists of over 10,000 entries and 28 tracked metrics. For our machine learning model, we will trim the total number of metrics to predict a player's salary to the following:

1. Points per Game
2. Rebounds per Game
3. Assists per Game
4. Steals per Game
5. Blocks per Game
6. True Shooting Percentage
7. Win Shares
8. Years Pro

Generally, these factors paint a picture of a player's performance and often are the main statistical categories brought up by fans when discussing players' salaries. We will be using these to predict a player's salary to compare with their actual contract. Win shares is a player statistic calculated using player, team and league-wide metrics, which attempts to divide the credit for team success to the individuals on the team. It is probably the most popular of the so-called advanced statistics.

We will examine how these different factors correlate with salary. Theoretically, better players should always earn higher salaries. We then expect the correlation between something like points and salary to be positive. This is because better scorers are often better players and, in turn, should earn more money. On the other hand, if we're looking at performance relative to salary, rookie contracts are by far the best value in the league and would dominate a calculation predicting the best value contracts. To account for this, we add how long a player has been in the league (years pro) as one of the features of our machine learning algorithm. We expect then players in their first years in the league to make less relative to similar performing players outside their rookie contracts.

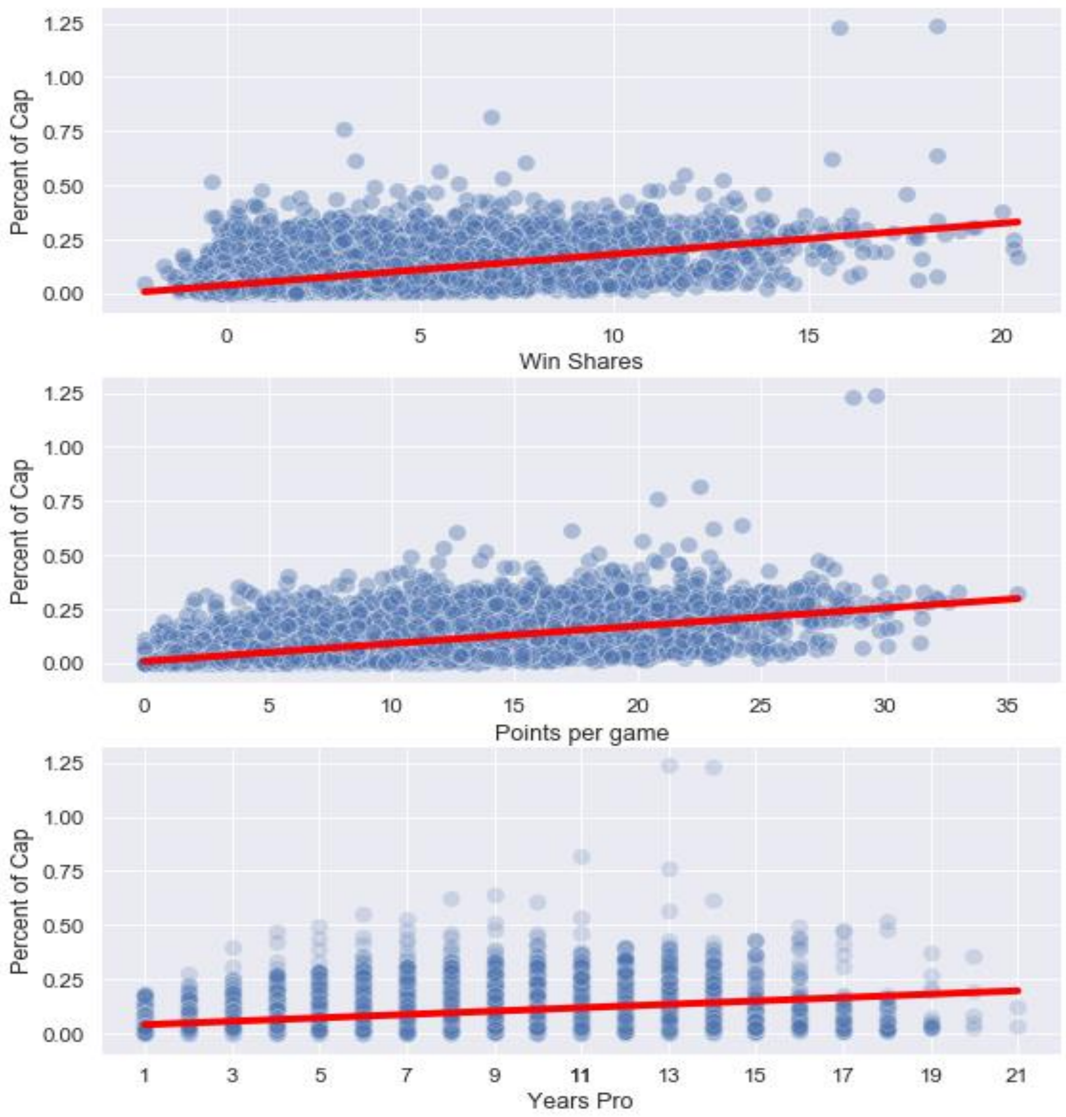

Figure 3. Percent of the cap earned as a function of Win Shares, Points per Game and Years Pro.

As expected, win shares and points per game positively correlate with salary. Furthermore, the plot of the percentage of the cap players earn as a function of how many years in the league shows the effect of rookie contracts. With few exceptions before rookie-scale contracts were implemented, no players within the first three years in the league earned more than 25% of the cap.

To understand the correlation between all the selected features and the percentage of the cap earned, it is useful to visualize all relationships together in a correlation plot as shown below.

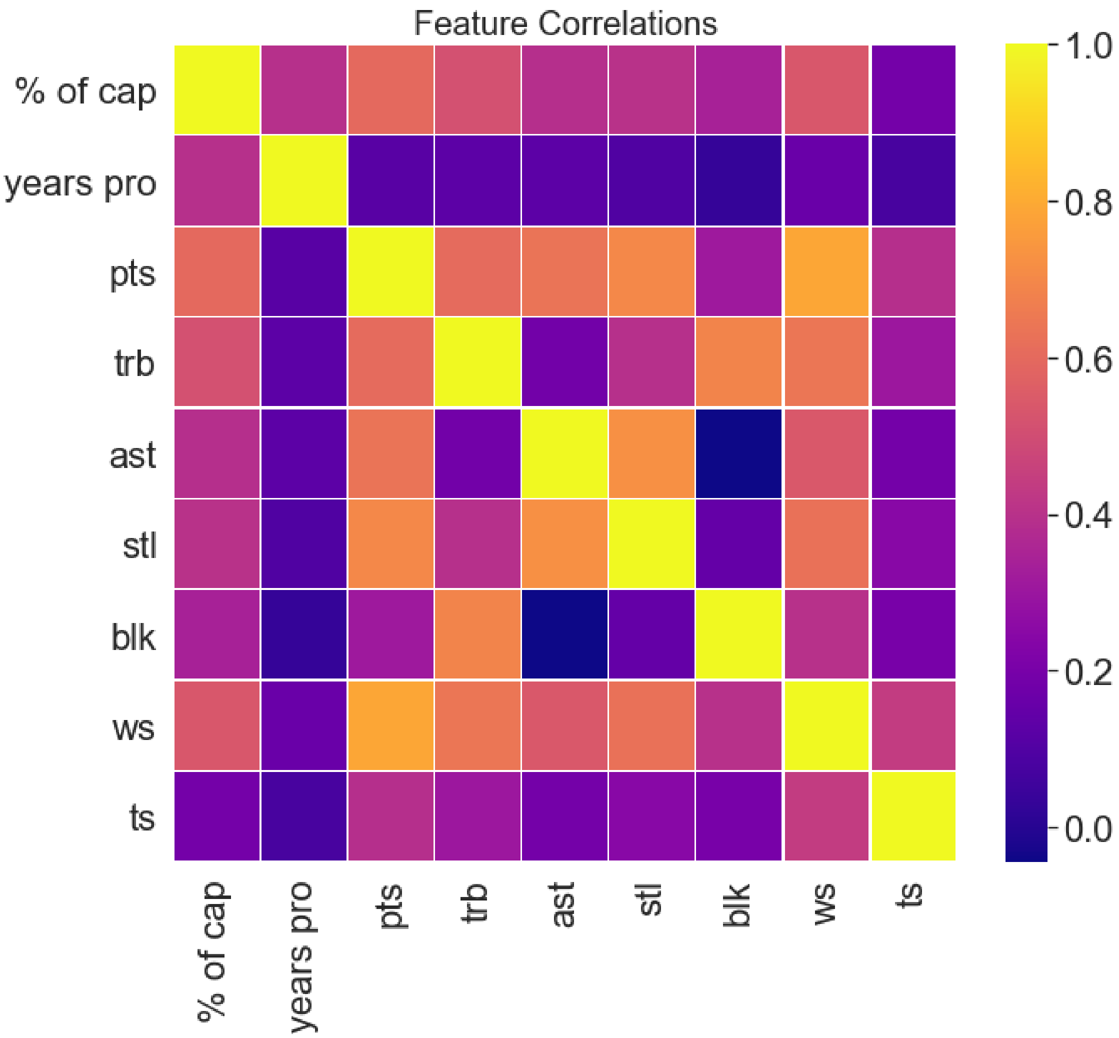

Figure 4. Correlation plot of all selected features. The percentage of the cap is shown along the first row. The lighter the color, the more intense the correlation between features. The average of the correlation coefficients of our features in 0.42.

Methodology

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks, in this case calculate how much an NBA player should be paid relative to their performance in key metrics. With our 10,000-plus samples, consisting of statistical and salary data for players since the 1990-91 NBA season, we are ready to train a machine learning model to make salary predictions. We will randomly split the data using a traditional 70/30 split to train/test our model. That is, we randomly select 70% of our total data to train the machine learning model, and the remaining 25% to test it for accuracy.

The types of machine learning algorithms differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve. In this case we can compare the predicted salary from the model with the actual salary a player received. Since our model is built of a set of data that contains both the inputs and the desired outputs, we can say this is a type of supervised learning algorithm. Additionally, because we are predicting a continuous variable, rather than a discrete one, this is a regression problem.

Our overall model will consist of the average prediction by four separate supervised regression models:

1. K-nearest Neighbors Regressor (KNN)
2. Random Forest Regressor (RF)
3. Linear Model Regressor (LM)
4. Support Vector Regressor (SVR)

We will be training our models using all NBA seasons from 1990 to 2018. We will then use the models to predict salaries for the 2018-19 season and compare them to the observed salaries. Our final model salary output will consist of the average predicted salary of our four individual regression models above. As a reminder, we are predicting the percentage of the cap a player deserves, instead of their raw salary. For all models we performed a grid search of the hyperparameters to optimize the model. This means we tested different combinations for factors that determine how our models fit the data and select the combination that results in the best score.

Model Scoring

For our regression models, we will use two metrics to measure their accuracy. First, R-squared (R^2) which is a statistical measure that represents the proportion of the variance for a dependent variable percentage of the cap) that's explained by independent variables (points, rebounds, win shares, etc.). The value of R-squared is between 0 and 1, with 1 being the best possible value. Second, root-mean-squared error (RMSE), which measures the difference between the predicted and observed values. Unlike R-squared, a lower RMSE is better, with a value of 0 indicating a perfect fit to the data. We can then interpret the RMSE as how close our predictions are to the real value on average.

In machine learning overfitting occurs when a model corresponds too closely or exactly to a dataset and may therefore fail to fit additional data or predict future observations reliably. To prevent overfitting, we'll resample our data by performing *k*-fold cross-validation (CV). In *k*-fold cross-validation, the dataset is split into *k* groups reserving one group to test the model and the remaining *k*-1 groups are used to train the model. The process is repeated for every combination of groups. This means that each group is given the opportunity to be used as the testing dataset one time and used to train the model *k*-1 times. This gives us an estimate of how our models perform on different splits of the data. If the cross-validated model scores close to our initial random split model, it indicates the model performs almost the same on the different splits. This is an indication that our model is not overfitting.

Model	R-Squared	RMSE	CV R-Squared	CV RMSE
KNN	0.572	0.548	0.551	0.548
RF	0.595	0.548	0.589	0.548
LM	0.524	0.548	0.517	0.548
SVR	0.584	0.548	0.566	0.548

All the models have an identical RMSE and cross-validated RMSE of 0.548. This indicates that our models are not overfitting the data. Additionally, we can interpret the value of 0.548 to mean that, on average, our models predicted salary cap percentage are 5.48% off from the observed value.

We saw before that the average correlation coefficient (R) for our features and the salary cap percentage regression is 0.42. This is equivalent to an r-squared of 0.18. Our models all have r-squared above 0.50, so they greatly outperform simple methods to predict salary.

Another tool we can use to score our models are the residuals, which are defined by the difference between the predicted and observed value at a point. For a model to be of value, it’s residuals should share two characteristics. First, they should follow the normal distribution with 95% of a model’s normalized residuals falling within 2 standard deviations of the mean. Second, they should have no autocorrelation or trend to show that the model isn’t biased.

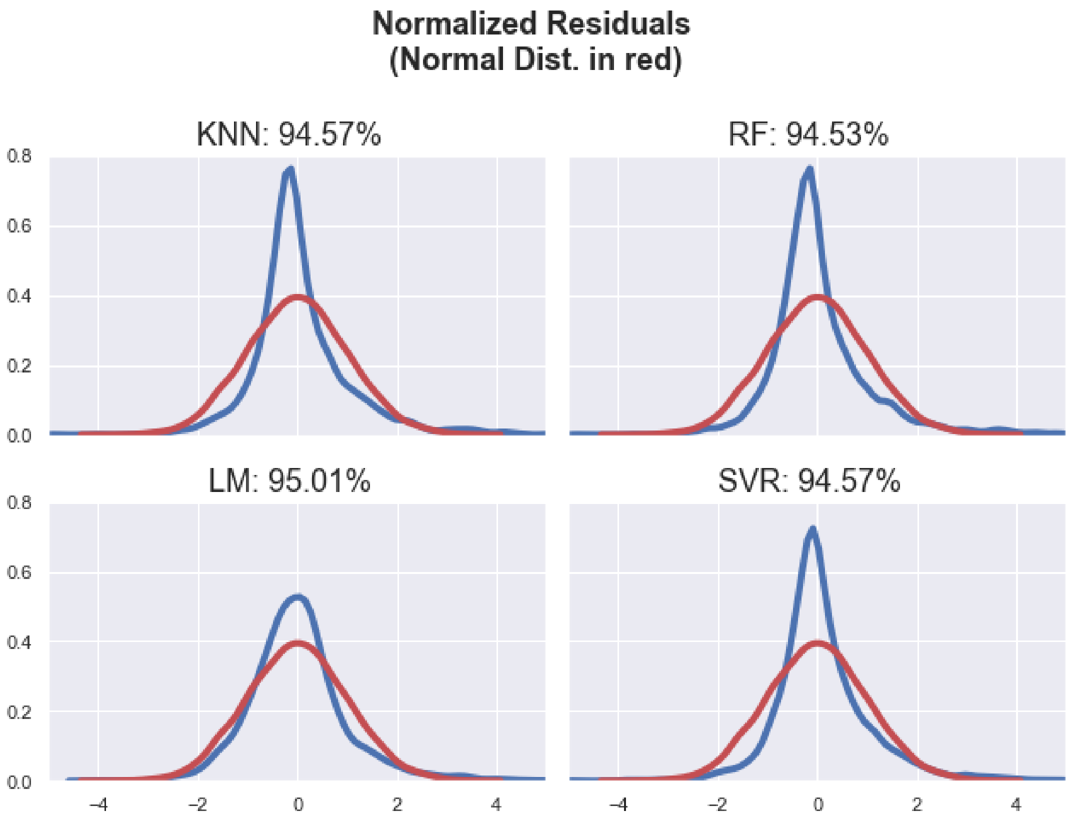

Figure 5. Plot of residual distributions of our models compared to the normal distribution (red line).

While, by definition, the normal distribution has 95% of its data within 2 standard deviations of the mean, we see that only the LM meets this condition. The other models are close to 95% but do not quite follow the normal distribution. Additionally, we see that the residuals peak close to 0 far above the expected density for a normal distribution. Though our residuals do not follow the normal distribution exactly, the models are still useful. It is important to note that the residuals have no autocorrelation, meaning the models are not biased.

Results

First, we will try to understand if our models are accurately capturing the effect of how years as a pro affect salary. As a reminder, rookie contracts are on non-negotiable, fixed scales for a players' first four years in the league. This makes these contracts of particularly high value when contrasted with regular contracts. To examine this, we will look at the effect the years pro metric has on Luka Doncic, the Rookie of the Year award winner for the 2018-19 season. We expect his salary to be low, despite his very high level of play, 21.2 PPG, 7.8 RPG, and 6 APG on good efficiency, because he is a rookie.

Our model predicts Luka should earn 5.4% of the cap compared to the actual observed value of 6.6%. We still consider his contract a bargain, as a player with these stats would earn far more in the open market. We will test this by predicting what Luka's salary would be from 1 to 15 years of experience by leaving all other metrics unchanged and using the model to predict the expected salary.

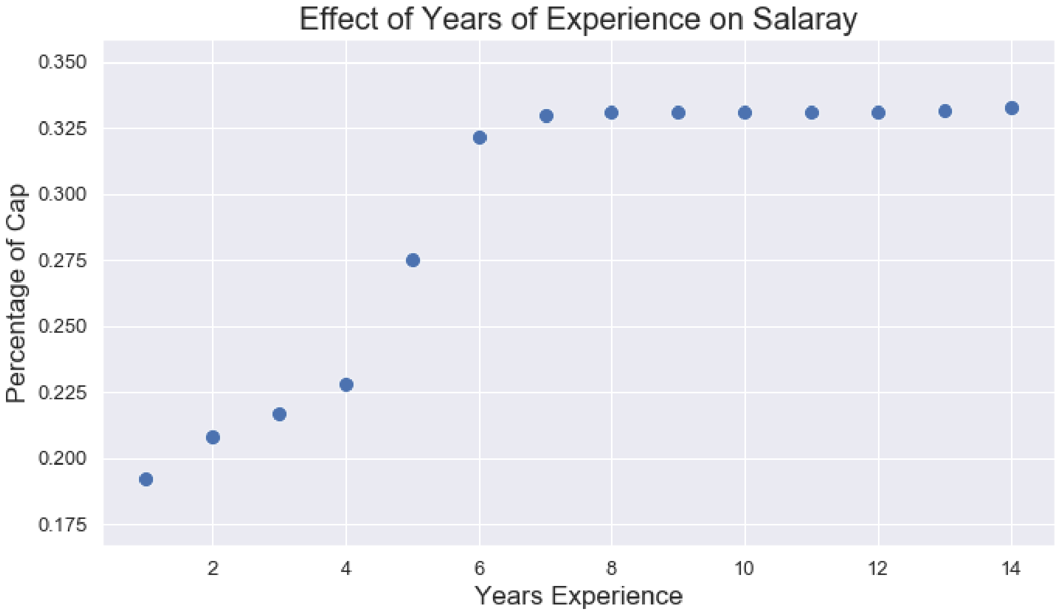

Figure 6. Predicted effect of years of experience on salary using Luka Doncic's stats for the 2018-19 NBA season.

The model predicts Luka's salary relative to the cap will increase in a linearly for his first four years in the league, until leveling off on subsequent years at a max contract. This is the behavior we expect years of experience will have on a player's salary and we can say the model accurately captures the effect of how long a player has been a professional. This shows that years pro has a stronger impact on salary earlier on a player's career as rookie contracts keep salaries low.

Now that we know our model is correctly accounting for experience, we can examine the results. We'll look at both player-by-player and team-by-team differences between predicted and observed percent of the cap. The higher the value, the better value the player is as he's earning below his predicted salary. The graph below shows the top-10 best value contracts as predicted by the average of our four models.

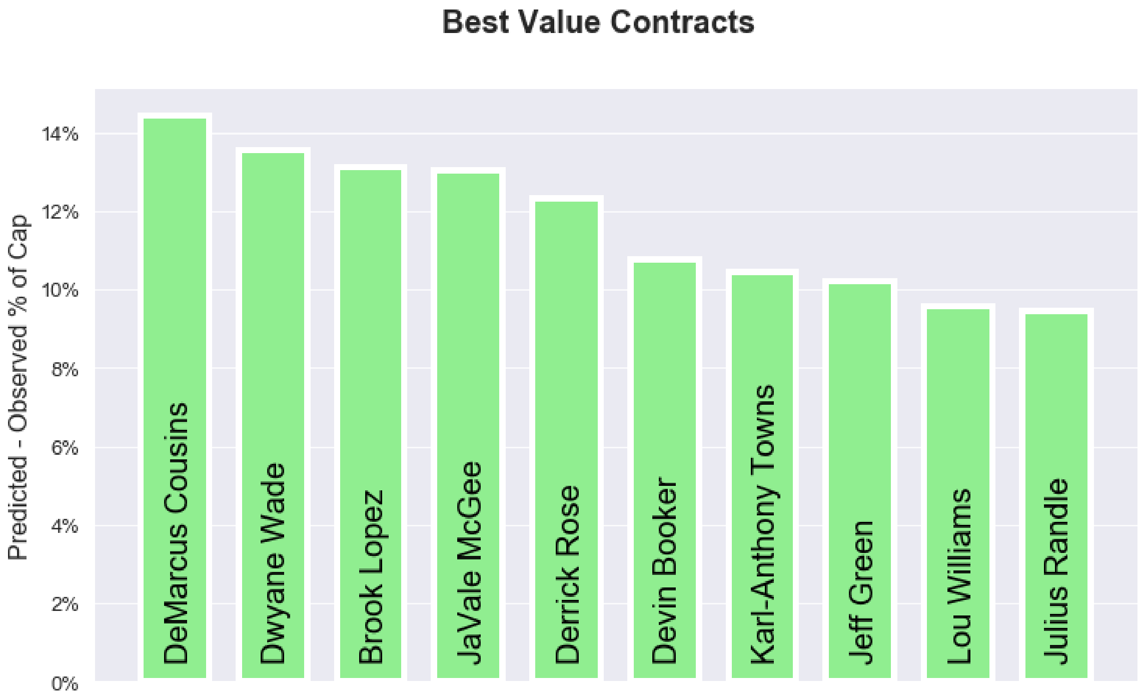

Figure 7. Top 10 best value contracts in the NBA during the 2018-19 season.

This list has some players that we would expect. First on the list is former All-Star Demarcus Cousins. Cousins, coming off a severe knee injury the year prior, easily outperformed his minimum contract signed with the Golden State Warriors to contend for a championship. Veteran players like Wade, Lopez, McGee and Rose are all on minimum contracts or exceptions, and all were solid contributors to their teams and as such considered a bargain. Despite our earlier example of how the models account for rookie contracts, there are still two rookie contract players in the top ten that are greatly outperforming their rookie contracts. Both Devin Booker and Karl Anthony-Towns are on the last year of their rookie contract, and both are projected to be superstars and expected offered maximum contracts during their next season.

Now let us examine the worst value contracts. Names on this list should not come as a surprise to the average NBA fan.

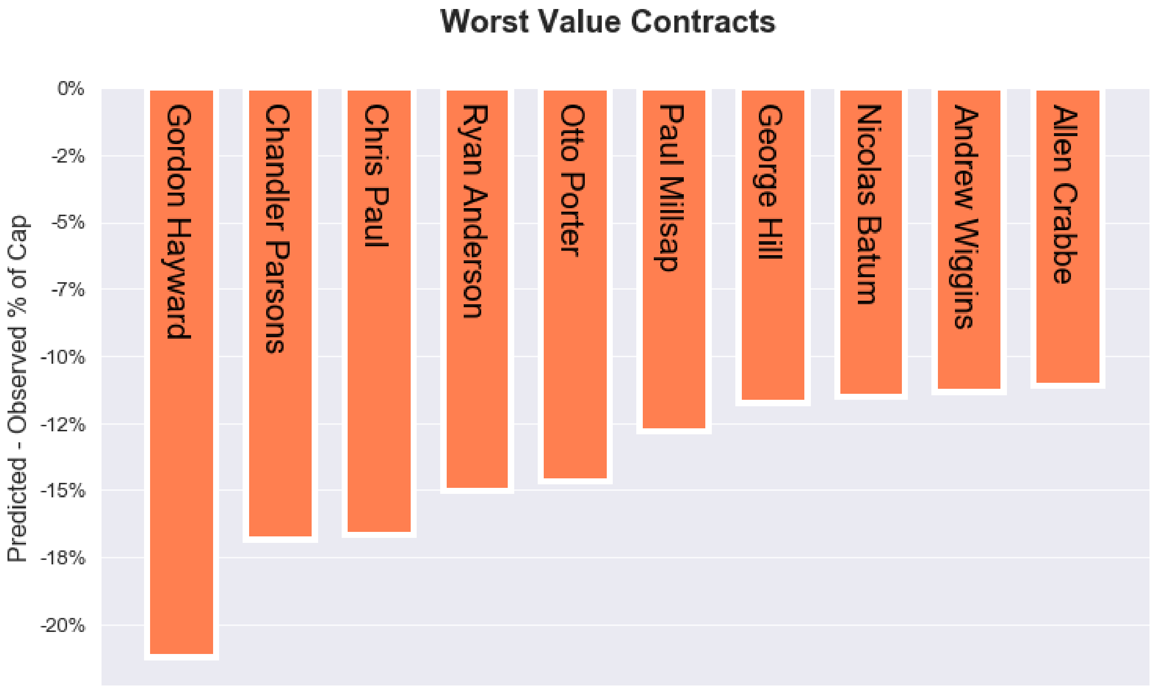

Figure 8. Top 10 worst value contracts in the NBA during the 2018-19 season.

The 20018-19 season marked Gordon Hayward's return to the court after a gruesome leg injury just one game into the previous season in which he had signed a maximum contract with the Boston Celtics. The former All-Star has not yet returned to form and captured his former glory, and some question if he ever will. Similarly, Chandler Parsons suffered a series of injuries shortly after getting his first big payday, which has kept him largely off the court. Players like Chris Paul, Ryan Anderson and Nicolas Batum are veteran players whose levels of play have declined drastically over the years now on the tail end of multi-year contracts from their better playing days. Finally, we examine Andrew Wiggins, who has one of the worst contracts in the league relative to his production. His signing by the Minnesota Timberwolves was immediately panned by fans as a bad move and is not surprising to find his name on the list.

Finally, let us look at the distribution of our predicted salary for the 2018-19 season relative to the observed salaries.

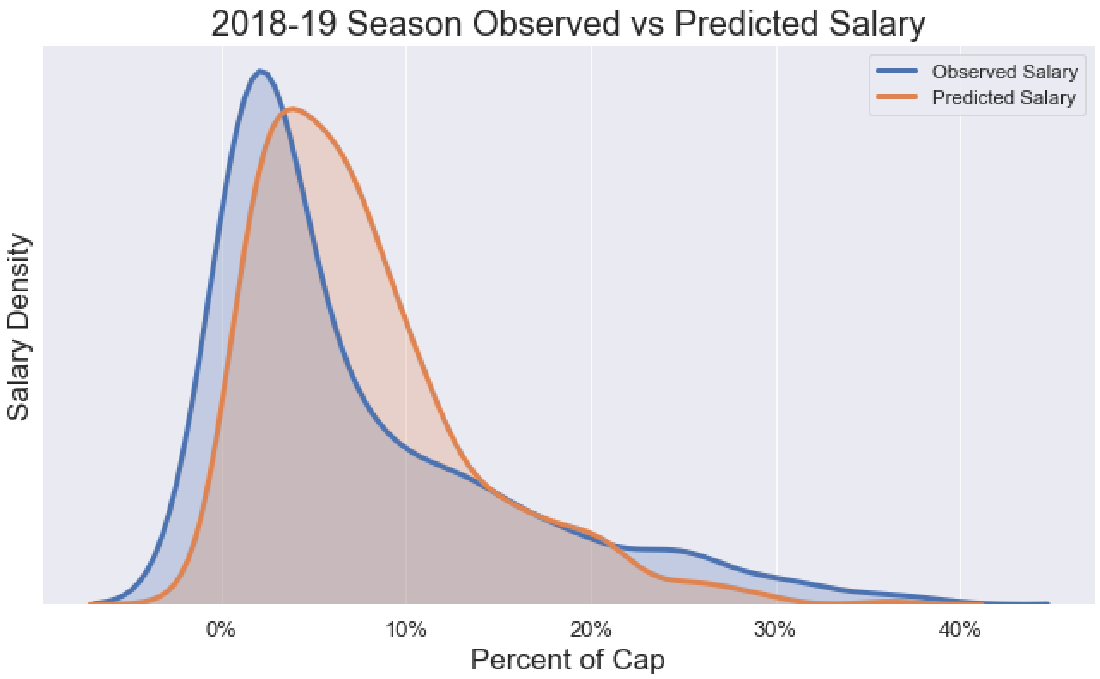

Figure 9. Observed and predicted salary densities for the 2018-19 NBA season.

We see that the predicted distribution has a right-shift relative to the observed distribution. This means that our model, on average, is predicting salaries higher than those observed. That is, it overpays players. This can be partly explained by good players taking minimum contracts and exceptions to play for contenders. Additionally, the rookie scale contracts the way they exist today did not exist for the entirety of the time of our training data, so it may not be fully capturing this effect.

Conclusion

When offering players contracts, general managers try to lure better players for their teams by offering lucrative contracts. By giving our machine learning model basic indicators of player performance, we can fairly accurately determine what a player should earn.

We could expand this to predict a player's salaries in advance by asking the question - Given a player's performance this year, what should he be earning the next? This would allow us to predict how much a player should earn before they start their contract or pick their team and it could be a tool used by both players and general managers during negotiations. However, this analysis would be much less accurate than what we did here. It's hard to predict how players will progress between seasons or how different schemes and environments affect their production. During a single season, a player's performance is much easier to predict so a retrospective look would be more accurate than a future prediction.