**2017-2018 NBA**

**Job Salaries**

**Purpose of the Study**

This study analyzes the salaries of NBA players who played during the 2017 - 2018 season. This topic was interesting because NBA, basketball's most elite league, is one of the most famous and highest paid professions in the sports industry.

This project aims to determine the factors which mainly influence the salary of NBA players.

**Dataset**

This dataset was obtained from Kaggle and contains 13 variables, in addition to 1 response variable, which was created to identify if the salary is above or below a specified threshold value.

Additionally, this dataset comprises 485 rows of data.

Data Source: https://www.kaggle.com/datasets/aishjun/nba-salaries-prediction-in-20172018-season

The dependent variable, salary, represents the amount paid to a player during the 2017-2018 NBA season. For logistic regression, an additional dependent variable, Y will be created to identify if salaries are high or low based on a specified threshold value.

Y is ‘high’ if the salary is more than $4M and ‘low’ if below.

The independent variables (X) that will be used are:

Age - Age of the player

G - Number of games played

PER - Player efficiency rating

TS% - Player’s true shooting percentage

3PAr - Player’s 3 point attempt rate

FTr - Player’s free throw rate

TRB% - Player’s rebound percentage

STL% - Player’s steal percentage

BLK% - Player’s block percentage

WS% - Player’s win share percentage

**Logistic Regression Analysis**

During the initial stages of the data analysis process, the data was cleaned, removing records with missing values. As a result, 485 records remained. Next, the data was partitioned into training and validation, 60% and 40%, respectively. Finally, logistic regression was run using “High” as the success class and 0.5 as the cutoff. The results are displayed below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | **Estimate** | **Confidence Interval: Lower** | **Confidence Interval: Upper** | **Odds** | **Standard Error** | **Chi2-Statistic** | **P-Value** |
| **Intercept** | -5.6146974 | -8.296026223 | -2.933368538 | 0.003644 | 1.368050058 | 16.84413564 | 4.06E-05 |
| **Age** | 0.18600584 | 0.116545955 | 0.255465721 | 1.204429 | 0.035439367 | 27.54743949 | 1.53E-07 |
| **G** | -0.0092164 | -0.025840278 | 0.007407409 | 0.990826 | 0.008481709 | 1.180753297 | 0.277203 |
| **PER** | 0.01180762 | -0.036021494 | 0.059636728 | 1.011878 | 0.024403056 | 0.234118691 | 0.628487 |
| **TS%** | -1.1891723 | -5.001561933 | 2.62321737 | 0.304473 | 1.945132503 | 0.373758582 | 0.540963 |
| **3PAr** | 0.19802175 | -1.193054348 | 1.589097852 | 1.218989 | 0.709745746 | 0.077843114 | 0.780242 |
|  |  |  |  |  |  |  |  |
| **FTr** | -0.2924007 | -1.599795085 | 1.014993734 | 0.746469 | 0.667050221 | 0.192149686 | 0.661133 |
| **TRB%** | 0.0049536 | -0.071485406 | 0.081392609 | 1.004966 | 0.03900021 | 0.016132743 | 0.898929 |
| **STL%** | 0.06558953 | -0.193998785 | 0.325177849 | 1.067788 | 0.132445453 | 0.245242062 | 0.620445 |
| **BLK%** | 0.01485129 | -0.182481544 | 0.212184118 | 1.014962 | 0.100681866 | 0.021758337 | 0.882  732 |
| **WS** | 0.47356413 | 0.271391363 | 0.675736887 | 1.605707 | 0.103151264 | 21.07698475 | 4.41E-06 |

Table 1: Analysis of Coefficients

In table 1, the formula for the logistic regression can be derived. This equation can be represented in three formats, log, odds, and probability.

Log format:

Log(Odds) = -5.6146974 + 0.18600584 Age -0.0092164 G + 0.01180762 PER -1.1891723 TS% + 0.19802175 3PAr -0.2924007 FTr + 0.0049536 TRB% +0.06558953 STL% + 0.01485129 BLK% + 0.47356413 WS

Odds Format:

We substitute using the log equation above to derive:

Odds = =

Probability Format:

We substitute using the log equation above to derive:

Probability =

The significance level of the regressors can be determined using the P-value. When analyzing the significance of the regressors above at a 30% level, we see that three of the regressors are significant. Since the P-value of age (1.53E-07), G (games played,0.277203), and WS (win share, 4.41E-06) are all less than 0.3, they are considered to be statistically significant in determining the salary. On the other hand, the p-value for PER (performance efficiency rating), TS% (true shooting percentage), 3PAr (3 point attempt rate), FTr (free throw rate), TRB% (Total rebound percentage), STL% (steal percentage), and BLK% (block percentage) are all more than 0.3, therefore these regressors are not considered to be statistically significant at the 30% level.

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | | |
| **Actual\Predicted** | **High** | **Low** |
| **High** | 71 | 52 |
| **Low** | 30 | 137 |

Table 2: Confusion Matrix for Training data

Table 2 shows the confusion matrix for the training data after the logistic regression is performed. We see that the true positive is 71 and the true negative is 137.Also, the false positive is 30 and the false negative is 52.

|  |  |  |  |
| --- | --- | --- | --- |
| Error Report | | | |
| **Class** | **# Cases** | **# Errors** | **% Error** |
| **High** | 123 | 52 | 42.27642276 |
| **Low** | 167 | 30 | 17.96407186 |
| **Overall** | 290 | 82 | 28.27586207 |

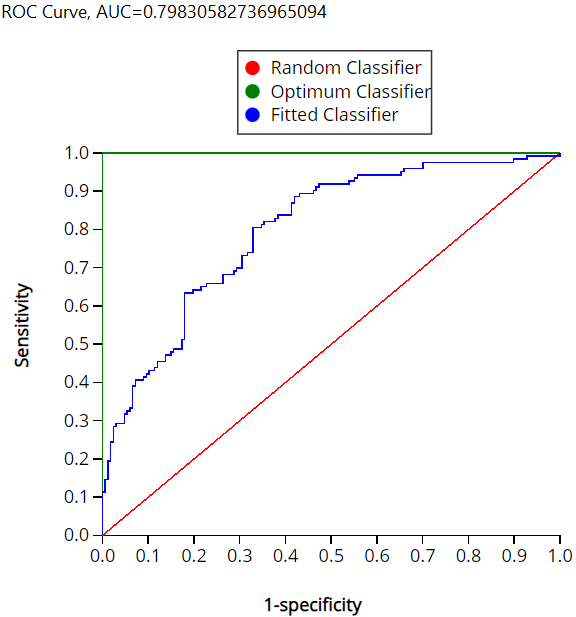
Table 3: Error Report for Training Data

Table 3 shows the error report for the training data. Here we see there are 52 errors in the success class (high) which makes the false negative percentage 52/123 = 42.28%. In the low (0) class, there are only 30 errors which makes the false positive percentage 30/167 = 17.96%, which is almost two times less than the false negative percentage. Overall, when combined they produce a total misclassification error of 82/290 = 28.28%.

|  |  |
| --- | --- |
| Metrics | |
| **Metric** | **Value** |
| **Accuracy (#correct)** | 208 |
| **Accuracy (%correct)** | 71.72414 |
| **Specificity** | 0.820359 |
| **Sensitivity (Recall)** | 0.577236 |
| **Precision** | 0.70297 |
| **F1 score** | 0.633929 |
| **Success Class** | High |
| **Success Probability** | 0.5 |

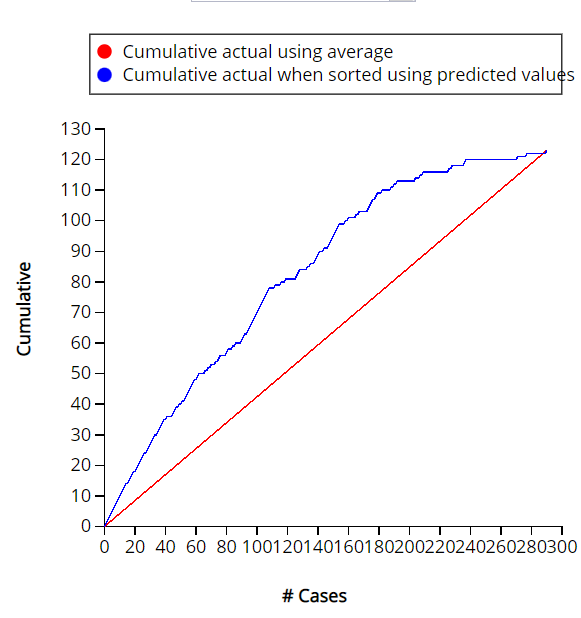
Table 4: Metrics for Training Data

Table 4 shows the metrics for the training data. In total, there are 208 (290 cases - 82 errors) classifications that are correctly classified, resulting in an accuracy classification percentage of 71.72%. The specificity, which represents the true negative percentage, of the training data is 137 (true negative cases) / 167 (number of low/negative cases) = 0.82. This means that 82% of cases classified as low were correctly classified. The sensitivity, which represents the true positive percentage, of the training data is 71 (true positive cases) / 123 (number of high/positive cases) = 0.58. This means that 58% of cases classified as high were correctly classified.



Graph 1: Specificity vs Sensitivity Chart for Training Data

The Receiver Operating Characteristic Curve or the ROC Curve is a graphic which shows the performance of a classification model at all classification thresholds. In the graphic above we notice that each line corresponds to a color. The red line represents the random classifier, blue represents the fitted classifier and lastly the green represents the optimum classifier. On the Y and X axis respectively the sensitivity or true positive and 1 - specificity or false positive was plotted. An ideal performance of a model would have the fitted classified line (blue) overlaying the optimum classified line (green) denoting that the model has 100% area under the curve. On the other hand, in our model the area under the curve was 79.8% which is still fairly accurate.



Graph 2: Lift Chart for Training Data

The Lift curve shows the relation between the number of instances which were predicted positive and those that are indeed positive and thus measures the performance of a chosen classifier against a random classifier. The blue line in the chart indicates the lift of the model or the cumulative actual when sorted using the predicted values. After partitioning the dataset, 290 rows of data was assigned for training. Out of the 290 entries, 208 predicted and actually derived as positive from running the model which is 71.72% accurate. The higher the chart the better the performance of the model therefore an ideal chart in this instance would have the 208 entries consecutively then plateau until it intersects with the cumulative actual using average (red) line.

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | | |
| **Actual\Predicted** | **High** | **Low** |
| **High** | 54 | 41 |
| **Low** | 15 | 83 |

Table 5: Confusion Matrix for Validation Data

Table 5 shows the confusion matrix for the validation data after the logistic regression is performed. We see that the true positive is 54 and the true negative is 83. Also, the false positive is 15 and the false negative is 41.

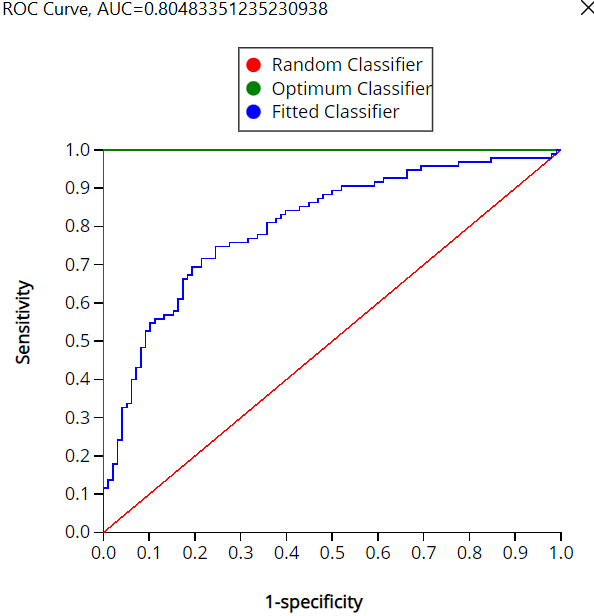
|  |  |  |  |
| --- | --- | --- | --- |
| Error Report | | | |
| **Class** | **# Cases** | **# Errors** | **% Error** |
| **High** | 95 | 41 | 43.15789474 |
| **Low** | 98 | 15 | 15.30612245 |
| **Overall** | 193 | 56 | 29.01554404 |

8

|  |  |
| --- | --- |
| Metrics | |
| **Metric** | **Value** |
| **Accuracy (#correct)** | 137 |
| **Accuracy (%correct)** | 70.98446 |
| **Specificity** | 0.846939 |
| **Sensitivity (Recall)** | 0.568421 |
| **Precision** | 0.782609 |
| **F1 score** | 0.658537 |
| **Success Class** | High |
| **Success Probability** | 0.5 |

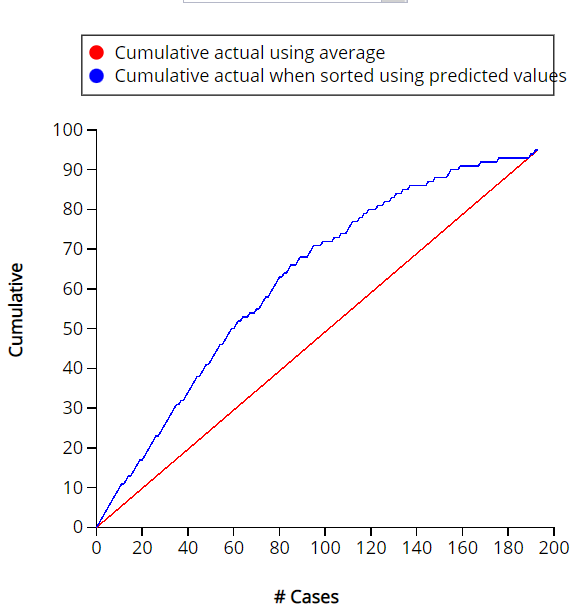
Table 7: Metrics for Validation Data

Table 7 shows the metrics for the validation data. In total, there are 137 (193 cases - 56 errors) classifications that are correctly classified, resulting in an accuracy classification percentage of 70.98%. The specificity, which represents the true negative percentage, of the training data is 83 (true negative cases) / 98 (number of low/negative cases) = 0.85. This means that 85% of cases classified as low were correctly classified. The sensitivity, which represents the true positive percentage, of the training data is 54 (true positive cases) / 95 (number of high/positive cases) = 0.57. This means that 57% of cases classified as high were correctly classified. Just as the error table for both training and validation data are relatively close, so are the metrics for training and validation data.



Graph 3: Specificity vs Sensitivity Chart for Validation Data

In graph 3, the area under the curve was 80.4% which is still fairly accurate. When compared to the training data the values of the models are quite similar.



Graph 2: Lift Chart for Validation Data

The partitioning allocated 40% of the total data to be validated which 185 rows of data. Out of the 185 entries, 137 predicted and actually derived as positive from running the model which is 70.98% accurate. The higher the chart the better the performance of the model therefore an ideal chart in this instance would have the 137 entries consecutively then plateau until it intersects with the cumulative actual using average (red) line. The accuracies of the training and validation data are also very similar with both values being approximately 71%.

**Additional Analysis**

There are a number of ways to increase the accuracy of these models and/or the accuracy of the high and low classes. Below, we rerun the previous logistic analysis using cutoffs of 0.7 and 0.3 to see how this change affects the classification.

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | | |
| **Actual\Predicted** | **High** | **Low** |
| **High** | 42 | 81 |
| **Low** | 11 | 156 |

Table 8: Confusion Matrix for Training data (0.7)

Table 8 shows the confusion matrix for the training data with the cutoff of 0.7. We see that the true positive is 42 which is lower than that of 0.5 cutoff, and the true negative is 156 which is higher than the 0.5 cutoff in Table 2. Similarly, the false positive is 11 which represents a decrease by almost half when compared to Table 2, and the false negative is increased to 81.

|  |  |  |  |
| --- | --- | --- | --- |
| Error Report | | | |
| **Class** | **# Cases** | **# Errors** | **% Error** |
| **High** | 123 | 81 | 65.85365854 |
| **Low** | 167 | 11 | 6.586826347 |
| **Overall** | 290 | 92 | 31.72413793 |

Table 9: Error Report for Training Data (0.7)

Table 9 shows the error report for the training data with a cutoff of 0.7. Here we see there are 81 errors compared to 52 making the error percentage, 65.85%, higher as well. In the low (0) class, there are 11 errors which is less than the 30 errors in Table 3, giving a lower false negative percentage of 6.58%. Overall, when combined they produce a total misclassification error of 31.72% which is higher.

|  |  |
| --- | --- |
| Metrics | |
| **Metric** | **Value** |
| **Accuracy (#correct)** | 198 |
| **Accuracy (%correct)** | 68.27586 |
| **Specificity** | 0.934132 |
| **Sensitivity (Recall)** | 0.341463 |
| **Precision** | 0.792453 |
| **F1 score** | 0.477273 |
| **Success Class** | High |
| **Success Probability** | 0.7 |

Table 10: Metrics for Training Data (0.7)

Table 10 shows the metrics for the training data with a cutoff of 0.7. In total, there are 198 correct classifications which is less than the model in Table 4. This results in a slightly lower accuracy classification percentage of 68.27%. The specificity, which represents the true negative percentage, increased to 93% while the sensitivity, which represents the true positive percentage, decreased to 34%.

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | | |
| **Actual\Predicted** | **High** | **Low** |
| **High** | 107 | 16 |
| **Low** | 70 | 97 |

Table 11: Confusion Matrix for Training data (0.3)

Table 11 shows the confusion matrix for the training data with the cutoff of 0.3. We see that the true positive is 107 which is higher than that of 0.5 cutoff, and the true negative is 97 which is lower than the 0.5 cutoff in Table 2. Similarly, the false positive is 70 which represents an increase when compared to Table 2, and the false negative is increased to 16 .

|  |  |  |  |
| --- | --- | --- | --- |
| Error Report | | | |
| **Class** | **# Cases** | **# Errors** | **% Error** |
| **High** | 123 | 16 | 13.00813008 |
| **Low** | 167 | 70 | 41.91616766 |
| **Overall** | 290 | 86 | 29.65517241 |

Table 12: Error Report for Training Data (0.3)

Table 12 shows the error report for the training data with a cutoff of 0.3. Here we see there are 16 errors, less than the 0.5 cutoff model, in the success class (high) which makes the false negative percentage 13% which is lower than Table 3. In the low (0) class, there are 70 errors which increases the false positive percentage to 41.92%. Finally, when the overall misclassification error 29% is compared to Table 3, we see that the misclassification error of the 0.3 cutoff model is slightly higher.

|  |  |
| --- | --- |
| Metrics | |
| **Metric** | **Value** |
| **Accuracy (#correct)** | 204 |
| **Accuracy (%correct)** | 70.34483 |
| **Specificity** | 0.580838 |
| **Sensitivity (Recall)** | 0.869919 |
| **Precision** | 0.60452 |
| **F1 score** | 0.713333 |
| **Success Class** | High |
| **Success Probability** | 0.3 |

Table 13: Metrics for Training Data (0.3)

Table 13 shows the metrics for the training data with a cutoff of 0.3. In total, there are 204 correct classifications which is less than the model in Table 4. This results in a slightly lower accuracy classification percentage of 70.34%. The specificity, which represents the true negative percentage, decreased to 58% while the sensitivity, which represents the true positive percentage, increased to 86%.

|  |  |
| --- | --- |
| Metrics | |
| Metric | Value |
| Accuracy (#correct) | 207 |
| Accuracy (%correct) | 71.37931 |
| Specificity | 0.820359 |
| Sensitivity (Recall) | 0.569106 |
| Precision | 0.7 |
| F1 score | 0.627803 |
| Success Class | High |
| Success Probability | 0.5 |

Table 14: Metrics for Training Data (0.5, significant Regressors only)

Table 14 shows the metrics for the training data with 0.5 cutoff and only significant regressors. When compared to Table 4 the metrics are almost identical with Table 4 having 1 more correct classification than this model. In total, the accuracy classification percentage of 71.38% with less than a 0.4% difference. Since all of these values are comparable to Table 4, we select this model since it is much simpler.

**Summary and Conclusion**

This report used both training and validation data to build logistic regression models. We were able to derive a logistic equation in the log, odds and probability formats. The log equation was -5.6146974 + 0.18600584 Age -0.0092164 G + 0.01180762 PER -1.1891723 TS% + 0.19802175 3PAr -0.2924007 FTr + 0.0049536 TRB% +0.06558953 STL% + 0.01485129 BLK% + 0.47356413 WS.

After exploring the logistic model with a 0.5 cutoff, other models, 0.7 and 0.3, respectively and 0.5 with only significant regressors, were also explored. After analyzing these models, we concluded that the 0.7 cutoff increases the accuracy classification for the low class, while the 0.3 cut off model increases the accuracy classification for the high class. Despite these increases in these classes, the 0.5 cutoff model has highest classification accuracy. Even though this was the case, we chose the model with 0.5 cutoff and only significant regressors due to simplicity since the accuracy difference was less than 1%.In conclusion, this analysis allowed us to see the factors which mainly influenced the salaries of NBA players. Based on the results obtained throughout this analysis we conclude that age, win shares, and games played the greatest impact on the salaries which were paid.