Andrew Robbertz
May 9, 2022

Explainable AI For Baseball World Series Champion Classification

Abstract

Major League Baseball team managers face a number of challenges in building championship-winning teams. They must maximize their team's performance given the budget they have to hire players. This paper analyzes the factors that contribute most to a championship winning team through using multiple visualization techniques and machine learning methods. We also use explainable AI methods to visualize how the various machine learning methods make their predictions.

Introduction

Major League Baseball in the United States is one of the most watched sports series and draws the second highest salaries among all American sports leagues. Team managers, analysts, and enthusiasts have all developed unique strategies for building a championship-winning team. At the highest level, the task is simple; score more runs in each game than the opposing team. However, teams each take a different strategy for this, employ the best pitching staff, the best defensive players, the most powerful home run hitters, or the most consistent hitters at getting on-base. The number of factors involved in creating a championship winning team are innumerable. Additionally, there will always be upsets, unexpected results, injuries, and suspensions that disrupt the trajectory of an otherwise championship winning team.

This paper creates several visualizations to better understand what factors contribute to a championship winning team. Multiple statistics are generated to better capture the performance of a team's offensive and defensive capabilities Additionally, this paper analyzes how the salary of players and the budget of the overall championship winning teams compares to other teams. For example, it's no surprise that that the New York Yankees have won more championships than any other team but have the highest budget to spend on top-tier players compared to any other team. I'd like to see how different teams perform adjusted for money they spend on their players.

Machine learning and artificial intelligence can be used to better understand the factors that result in a championship winning team. Machine learning algorithms are often capable of identifying trends that a human observer would not normally recognize. However, our ability to understand and interpret how machine learning models make their predictions and classifications is extremely limited.

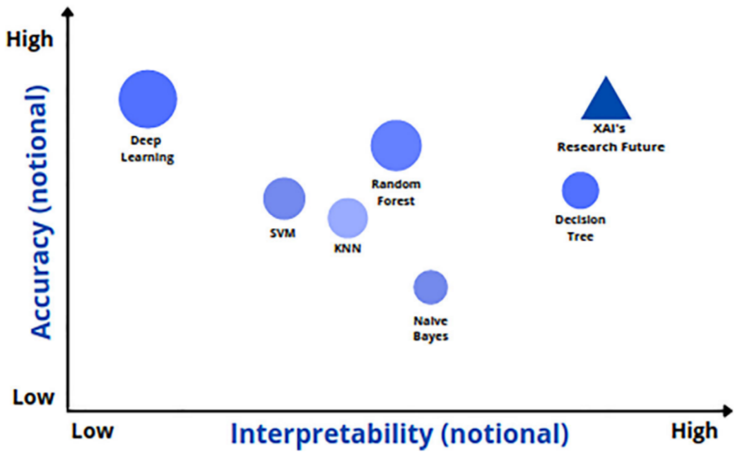

Figure 1. Accuracy vs. interpretability for different machine learning models [5]

Therefore, this paper uses state-of-the-art strategies within the field of explainable AI to better understand how machine learning models can predict world series championship winning teams.

Background

There are many unique data-visualization techniques for sports-related data [3]. The article provides insights and suggestions for three types of data including box-score data), live tracking data, and meta-data for summary statistics. For the box-score data, the paper provides multiple strategies for showing statistics over time, overlaying statistics to provide context, and showing relative distances between metrics rather than actual values to make them more apparent to an observer. These strategies were implemented in this paper to better show the relationships between features rather than raw statistics themselves.

There are also several difficulties in handling unbalanced classification data. The problem proposed is to determine whether a given team is a championship-winning team based on other characteristics. However, there are for more examples of non-championship-winning teams compared to winning teams. This makes visualization of model performance difficult since most common metrics and figures are designed for balanced datasets. Some research uses F-Curves to measure model learning and performance, compared to traditional accuracy and loss curves, since the F-Score metric much better represents the performance of a model on unbalanced data [4]. Implemented in this paper we use F-measures, Matthew's correlation coefficient, as well as visualizations of confusion matrices to show the performance of our machine learning models on the imbalanced datasets we have.

There are many current state of the art strategies for explainable AI including the SHAP library implementing Shapley values to measure feature interactions within multiple model types. The SHAP library combines multiple strategies including feature-oriented methods, global methods, concept methods, surrogate methods, local methods, and human-centric methods [5, 6]. The library models explanation functions to wrap model behavior and understand interdependencies within black-box machine learning models. This approach has been shown to work within linear and kernel-based models [6] as well as tree-based models [7].

Approach

The dataset used for this analysis is the Baseball Databank dataset featuring data from baseball seasons 1871 – 2021 [1]. The dataset has several CSVs for statistics about teams and players alike, with features covering each of the areas of analysis this project is concerned in. The Baseball Databank dataset has limited data for player salaries and team payrolls, therefore salary data is extended using yearly team payroll data from The Baseball Cube [2].

This paper and accompanying code uses various Python libraries to explore relationships within these datasets. The following features are generated to better visualize complex aggregate statistics for teams.

Batting Average (BA)	$AVG = \frac{H}{AB}$
On-Base Percentage (OPS)	$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}$
Slugging Percentage (SLG)	$\mathrm{SLG} = \frac{(1B) + (2 \times 2B) + (3 \times 3B) + (4 \times HR)}{AB}$
On-Base Plus Slugging (OPS)	$OPS = OBP + SLG$

Data is cleaned for use including removing rows where appropriate, removing unneeded feature columns, filling in missing values for relevant features, and normalizing all data before being used within the machine learning models. Plots are generated using the Matplotlib and Seaborn libraries to explore relationships within the data. Visualizations include scatter plots, bar charts, stacked bar charts, and heatmaps to show relationships between features and trends in the data.

A series of machine learning classifiers are then generated to predict whether a given team is a championship-winning team or not. Additionally, a series of visualizations are generated showing how the machine learning models detected trends and learned to predict championship-winning teams. We generate a support vector machine classifier to demonstrate the performance and visualization for kernel-based models. We generate a logistic regression classifier to demonstrate the performance and visualization for linear-based models. Finally, we generate a decision tree classifier to demonstrate the performance and visualization for a tree-based model. The SHAP library is used from previous discussion to generate explanations for each type of model structure. The SHAP library generates both static and interactive plots using HTML and JavaScript that can be interacted with in the accompanying code.

Results

The visualizations generated and results of our machine learning classification models are discussed in this section. The first visualizations compare offensive and defensive statistics between the American League and National league world-series winning teams and non-winning teams.

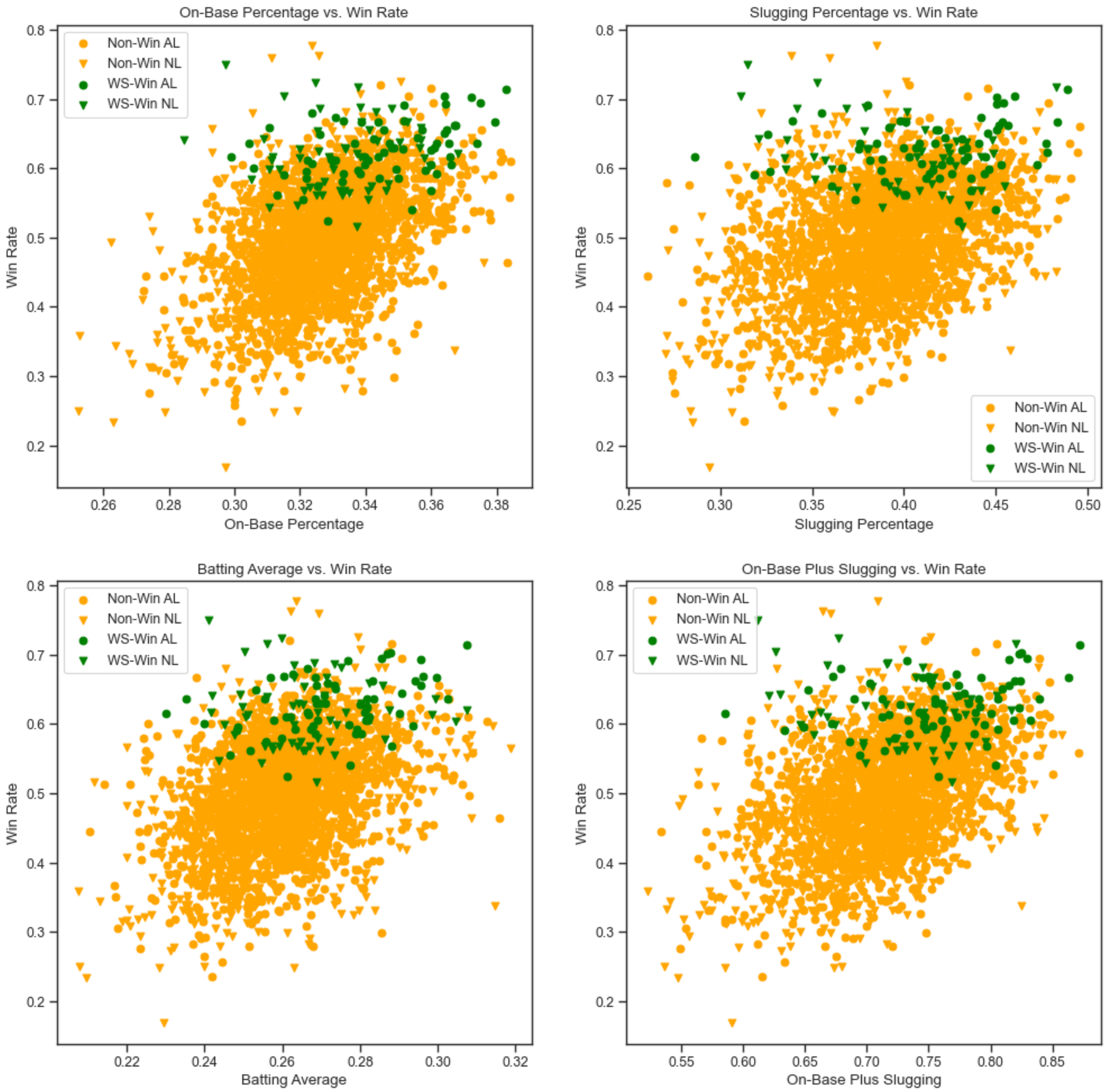

Figure 2. Offensive Statistics for AL and NL Champions

Figure 2 compares various offensive statistics to the win-rate of each team, or the ratio of games they won compared to games played. Its clear from the visualization that teams that performed better in the given offensive area, getting on base and slugging (hitting for power), win a higher ratio of games compared to other teams. Additionally, we can see that the championship-winning teams, colored in green, both perform best in the offensive metrics categories as well as have the highest win-ration of games compared to other teams. This figure also confirms the idea in baseball that on-base percentage is more important than batting average for hitters. There is a greater correlation between the on-base percentage and the ratio of games won compared to the teams overall batting average and the ratio of games won.

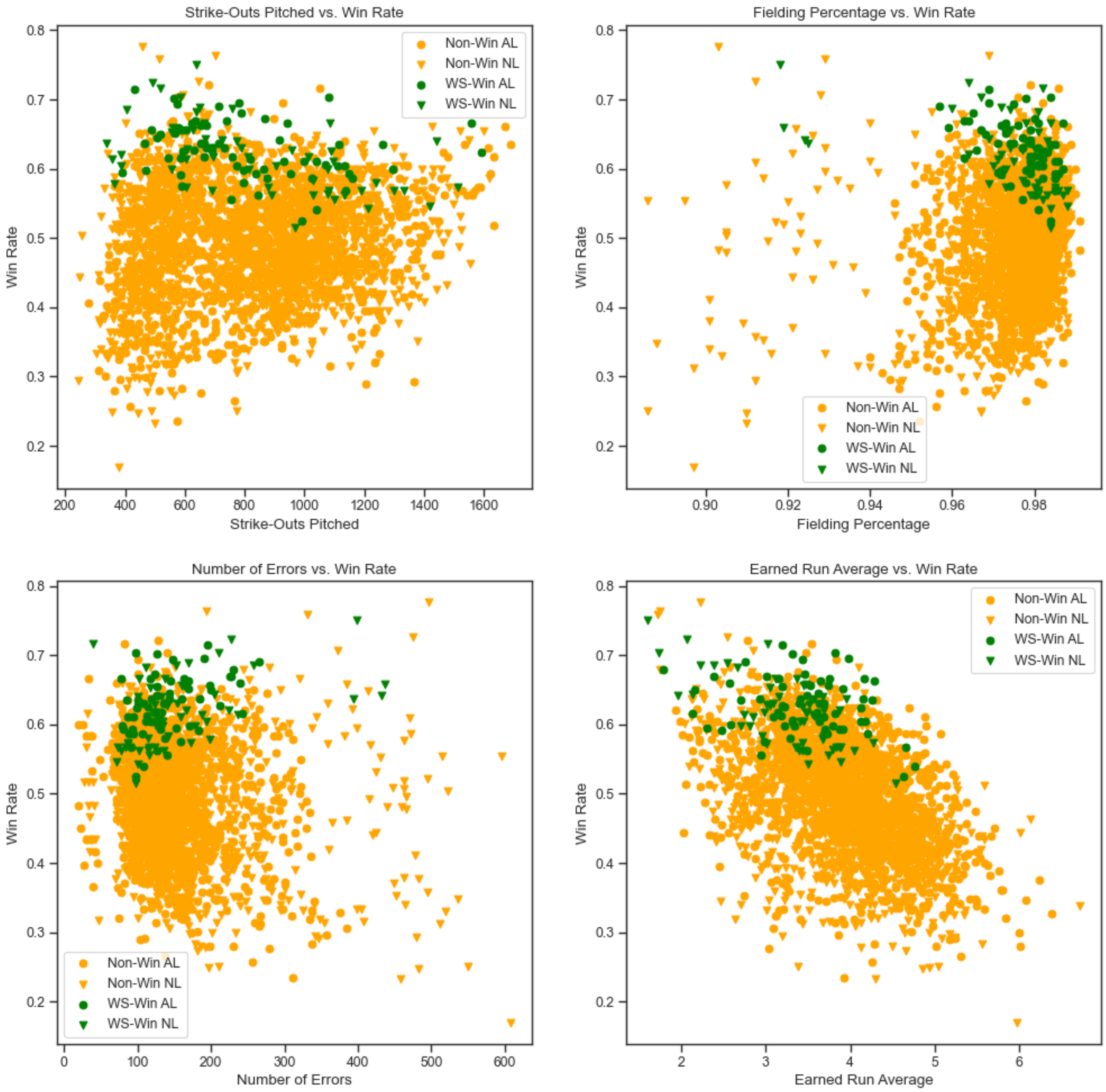

Figure 3. Defensive Statistics for AL and NL Champions

Figure 3 compares various defensive statistics to the win-rate of each team. Its clear from the visualization that teams that performed better in the given defensive area, pitching more strikeouts, cleanly fielding more balls, making fewer errors in the field, and having fewer ear-ed runs per game, win a higher ratio of games compared to other teams. Additionally, we can see that the championship-winning teams, colored in green, both perform best in the defensive metrics categories as well as have the highest win-ration of games compared to other teams. These visualizations also confirm the belief in baseball that its less important for pitchers to strike out batters as is for the pitcher to have the ball put in play and the rest of the team support in getting the out. We can see that there is a greater correlation between the fielding percentage and the number of errors with the win rate compared to the number of strikeouts pitched and the win rate.

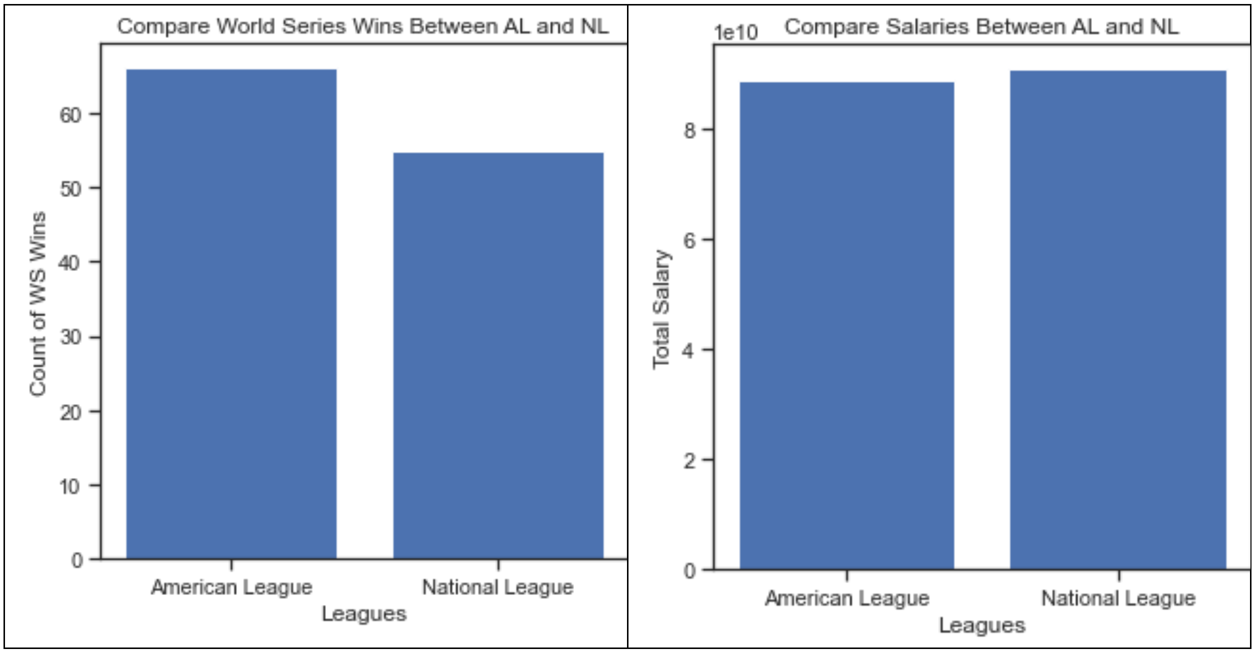

Figure 4. Comparison of American and National League Teams

Figure 4 compares American and National league teams within MLB. From the figure its interesting to see that the American League teams have won more world-series titles compared to the national league despite the National League teams having a greater overall budget to spend on their players.

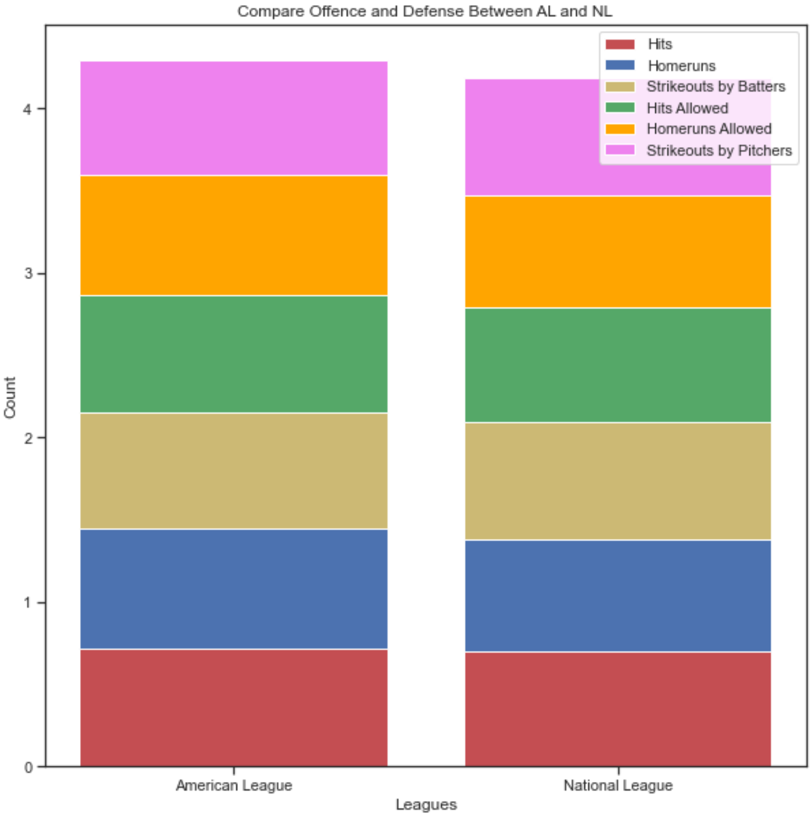

Figure 5. Comparison of AL and NL Offensive and Defensive Stats.

Figure 5 compares American League and National League offensive and defensive statistics. From the visualization we can see that the National League teams hit fewer homeruns compared to their American League counterparts. This can also be seen in Figure 2 where the majority of teams with the highest slugging percentage are from the American League and the majority of teams with the lowest slugging percentage are from the National League. This may put into perspective the results from Figure 4 that the AL teams have won more World Series since slugging percentage is so highly-correlated with a high win-ratio and World Series titles.

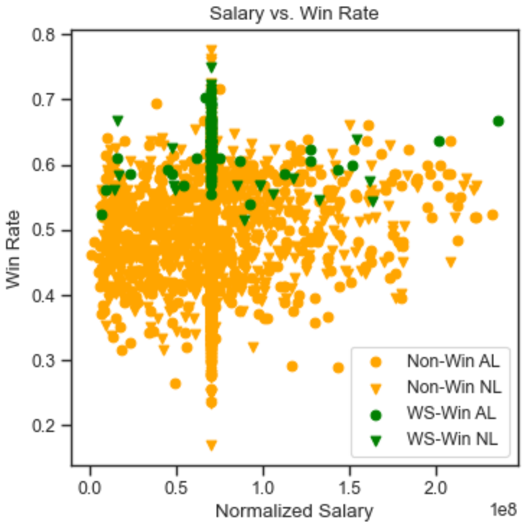

Figure 6. Normalized Salary vs. Win Rate

Figure 6 shows the normalized salaries of AL and NL teams compared to ratio of games the team won as well as whether they won the world series. Salary data for teams was missing from 1871-1985 resulting in the majority of teams having the ‘average’ normalized salary. However, we can still see from the plot that teams with a higher normalized salary won a greater ratio of games and won a greater number of World Series compared to lower payroll teams.

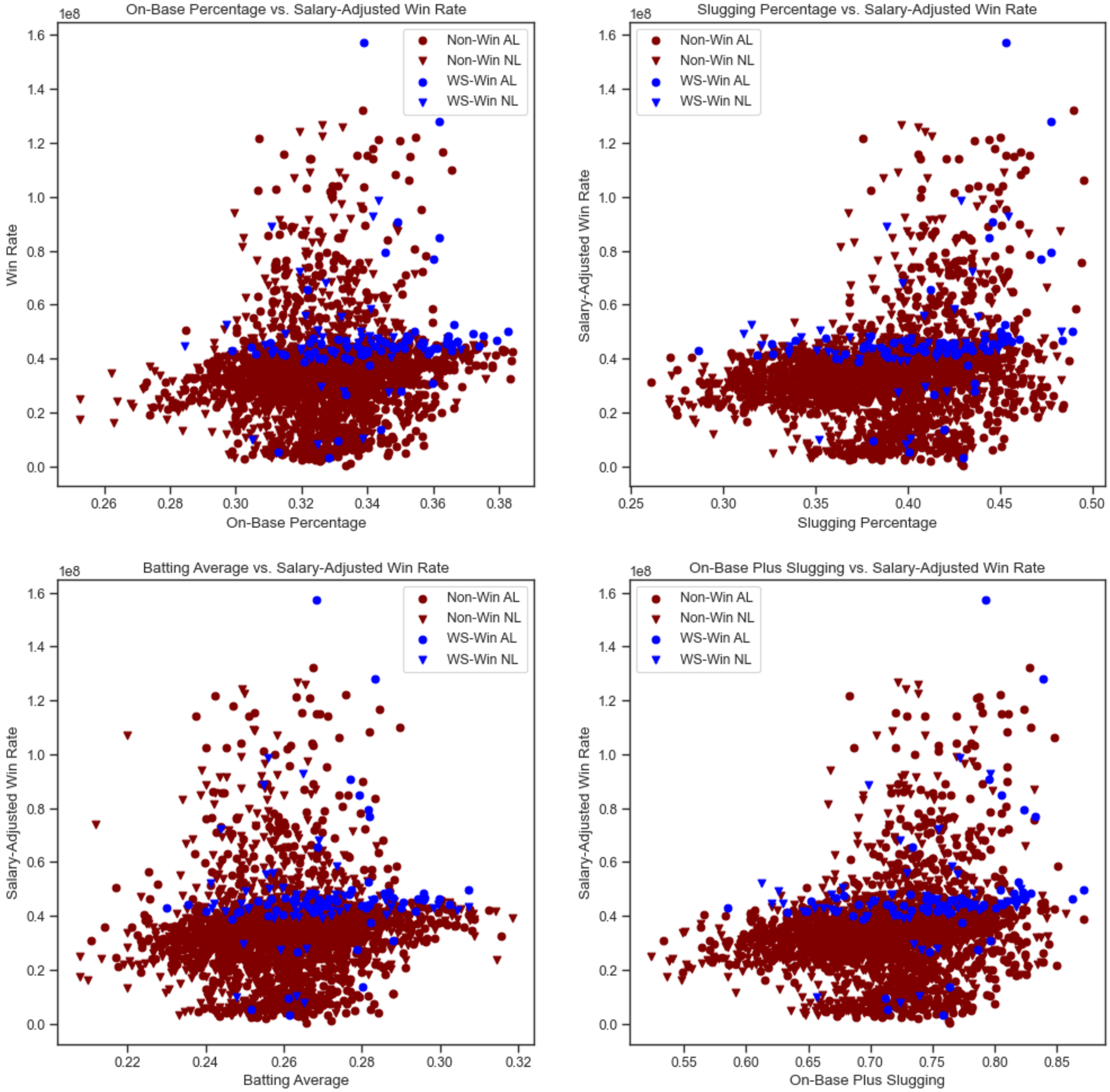

Figure 7. Salary Adjust Offensive Statistics.

Figure 7 shows salary-adjusted offensive statistics for AL and NL teams. From the figure we can see that there is still a very high correlation between teams having higher slugging percentages and on-base plus slugging percentages and winning more world series championships. This shows that, all salaries being equal, the best offensive team is more likely to win a World Series.

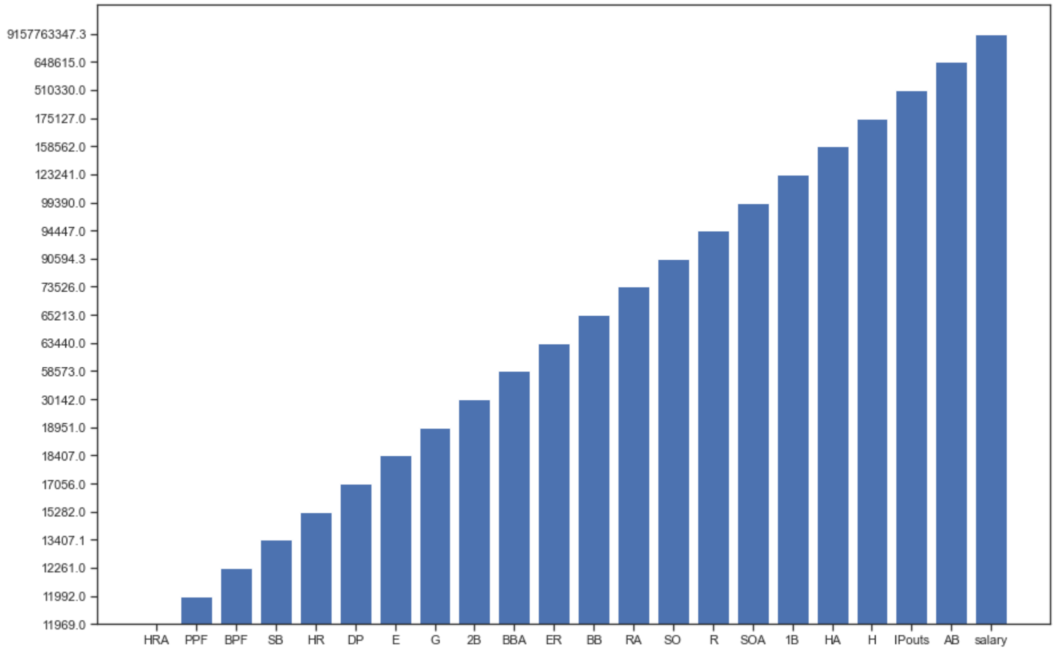

Figure 8. Direct Covariance of Features to World Series Wins

Figure 8 shows the top 20 feature covariance to winning Worlds Series championships. From this figure we can see that most related feature is the team's over salary to pay its players. After this interesting features include this (H), hits allowed (HA), strikeouts (SO), and strikeouts allowed (SOA).

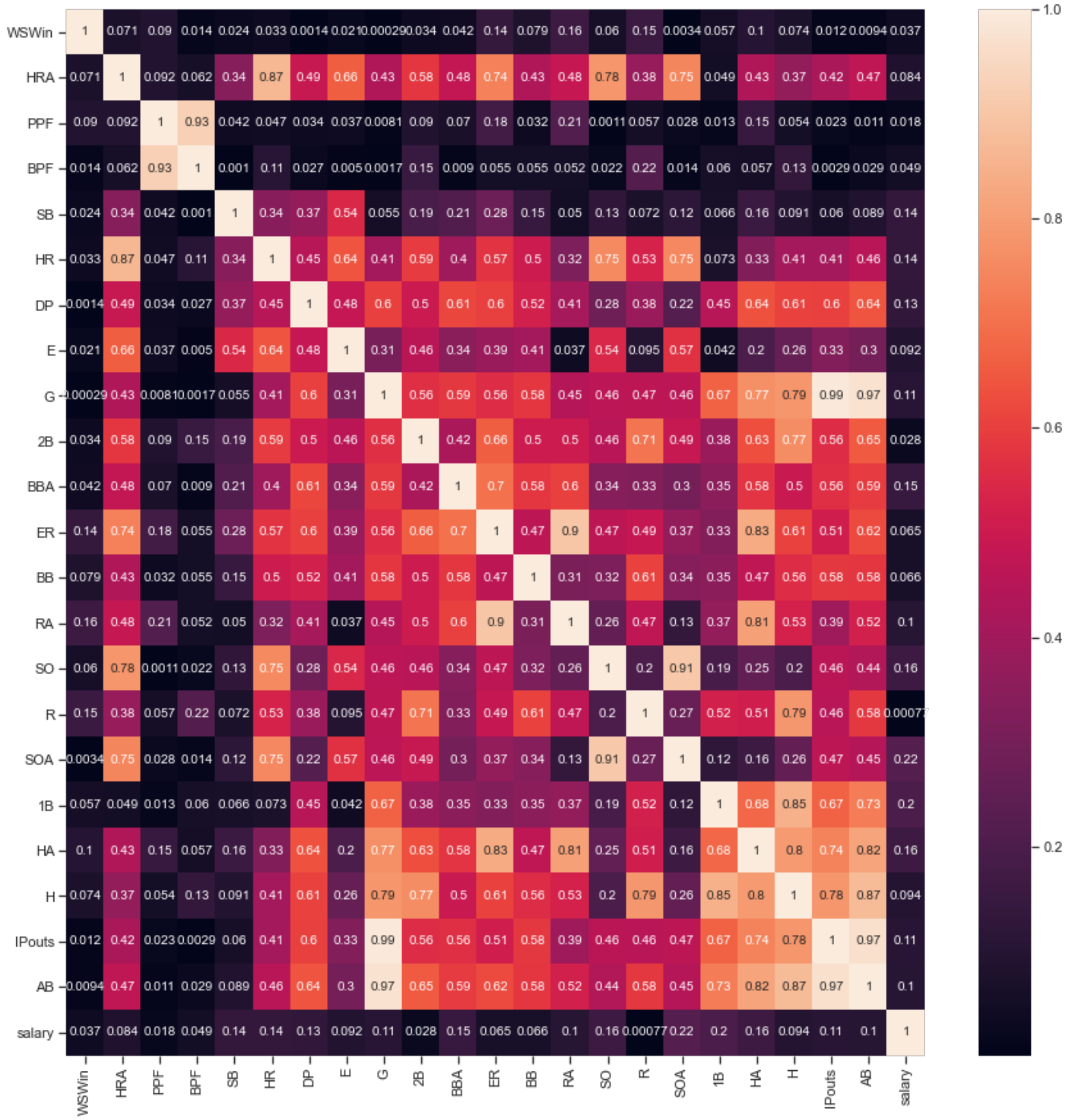

Figure 9. Heatmap of Feature Correlations

Figure 9 shows a heatmap of the top 20 feature correlations to each other. From this we can see that the most correlated features are runs allowed (RA), runs (R), earned runs (ER), and hits allowed (HA) similar to the previous figure as well.

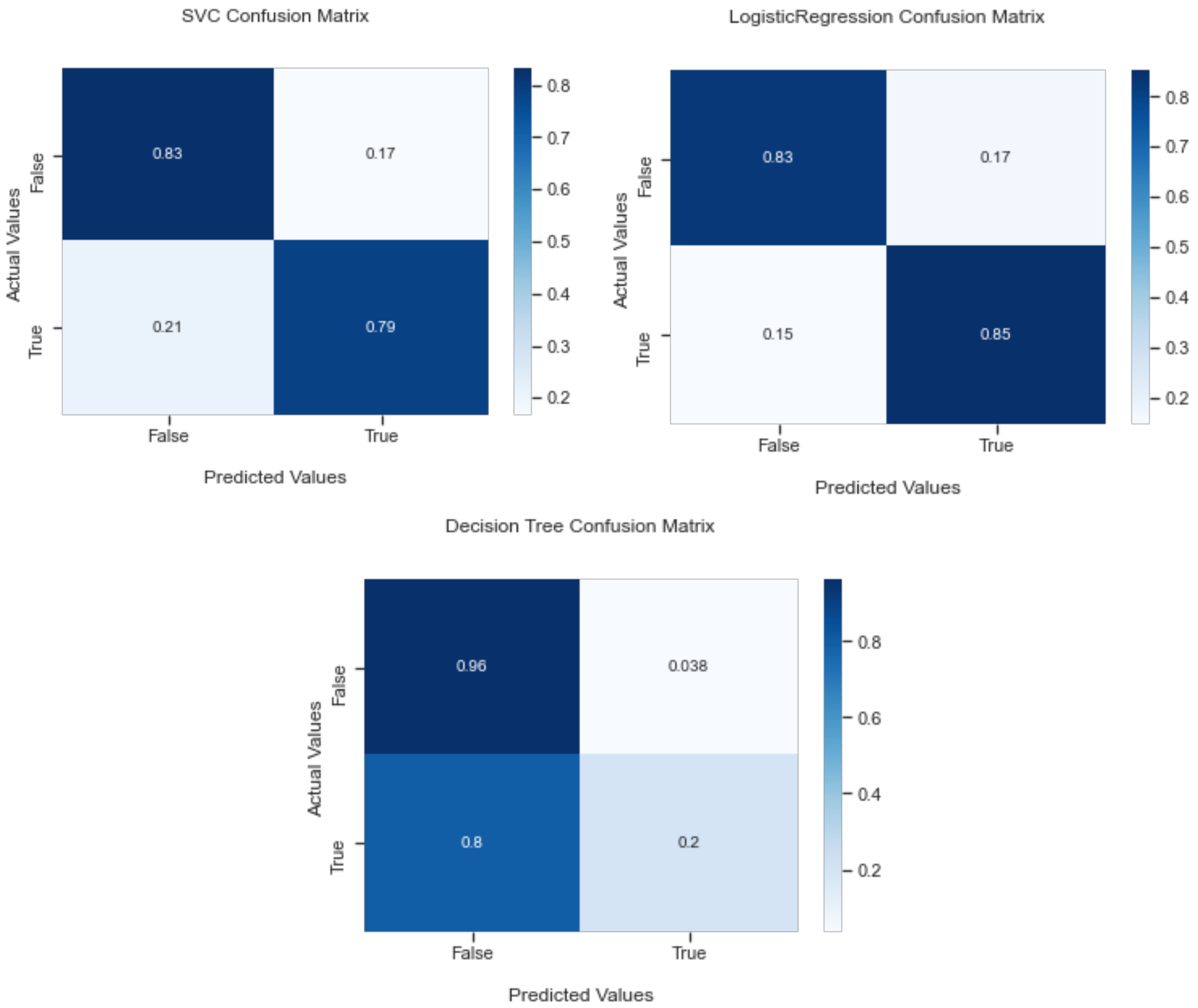

Figure 10. ML Classifier Confusion Matrices

Figure 10 shows the classification performance of the three machine learning models used. The Logistic Regression model achieves the best overall performance with 84% balanced accuracy between the majority and minority classes. The decision tree did not achieve very good performance, however we included it to show the SHAP libraries ability to explain their feature interactions.

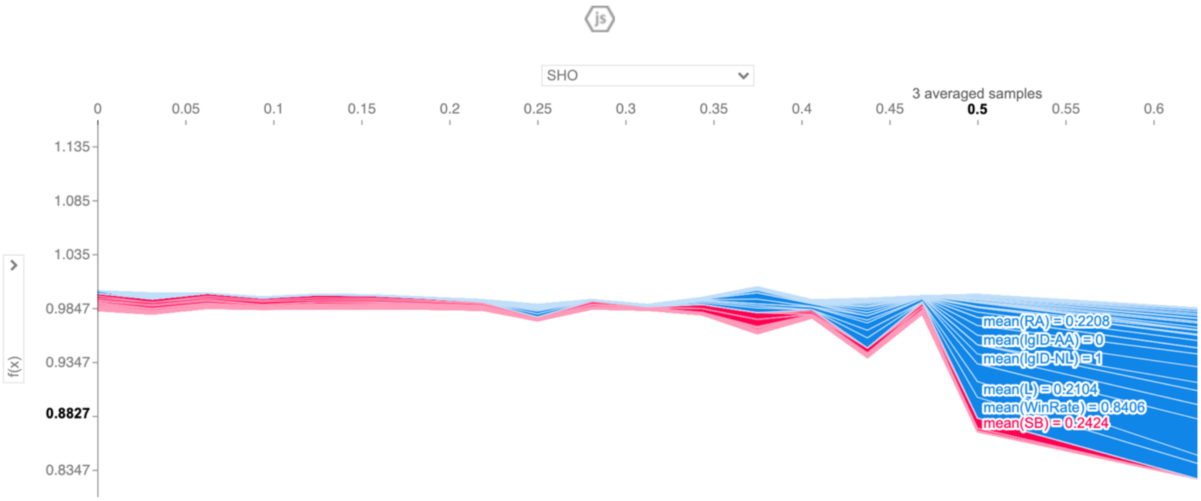

Figure 11. Interactive SHAP Values for SVC

Figure 11 shows a screenshot of an interactive plot generated by the SHAP library. The interactive plot allows the user to visualize the effect of various features on the output of SVC model. The user can filter based on the features they're interested in visualizing. In this screenshot we can visualize the effects of the runs allowed (RA), the league, the number of losses (L), the win-rate, and the number of stolen bases (SB).

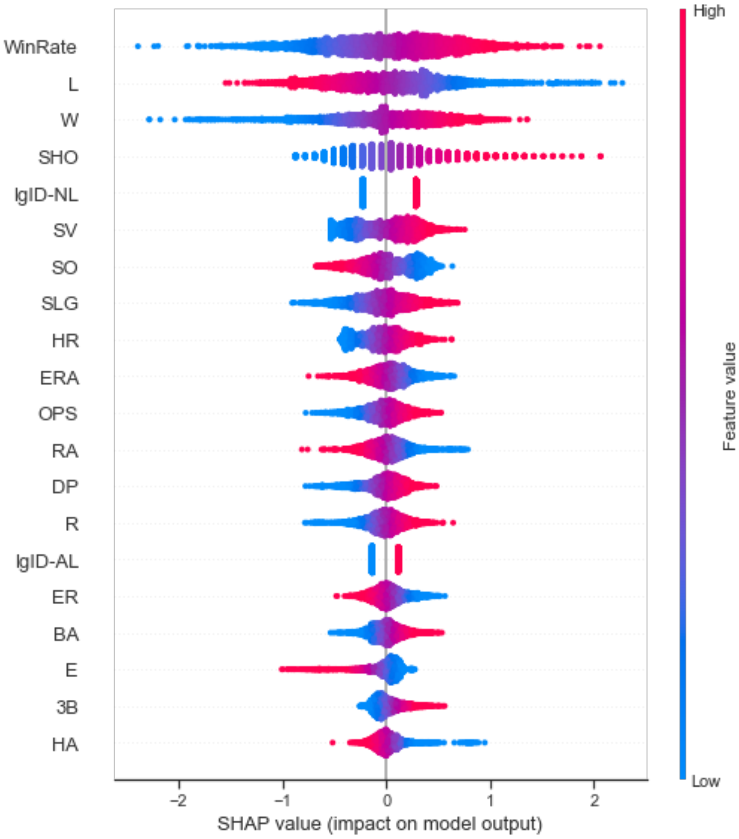

Figure 12. SHAP Values for Logistic Regression Model

Figure 12 shows the SHAP values for the most influential features in the Logistic Regression model. From this we can see how each features value(in color) impacts its SHAP value for importance within the model. The figure also shows the distribution of each of the features used in the model.

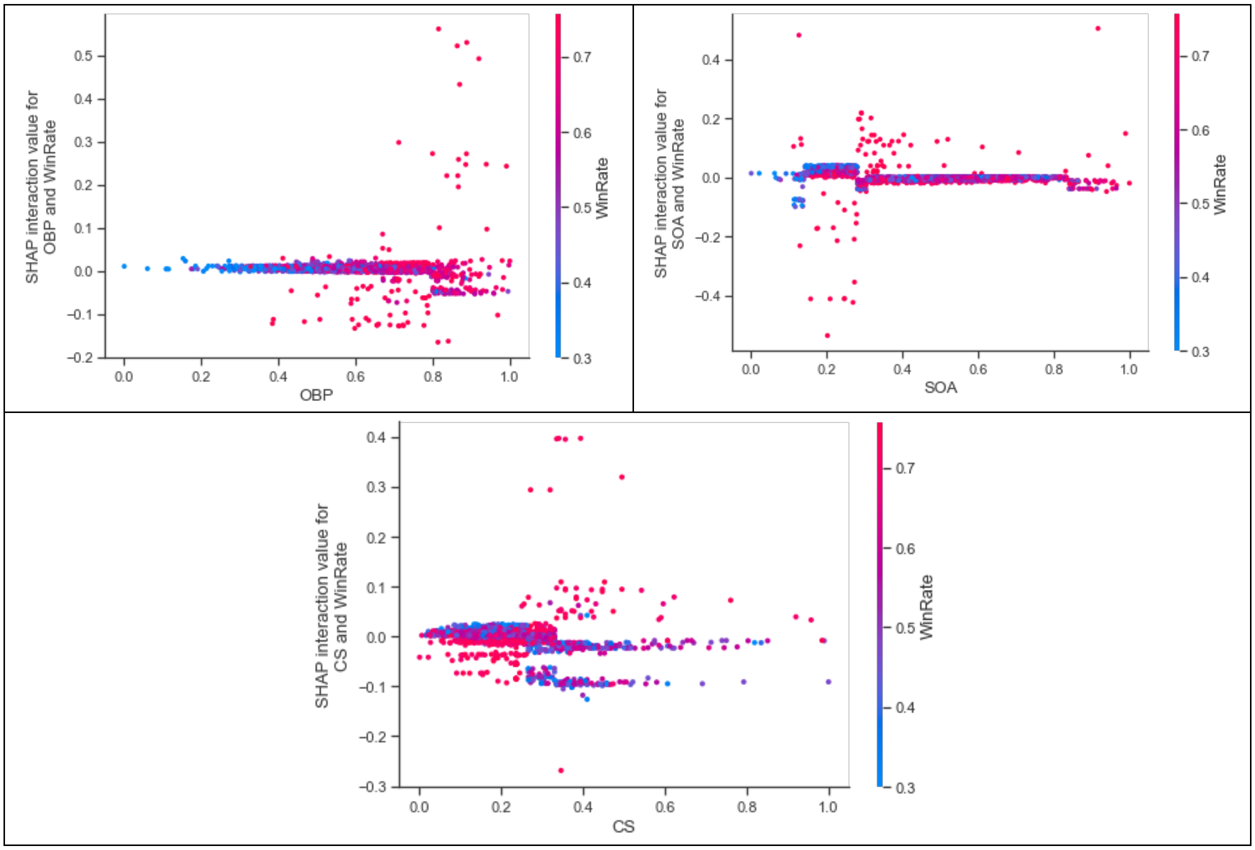

Figure 13. SHAP Interaction Dependency Plots for Decision Tree

Figure 13 shows SHAP interaction depencency plots for three features within the decision tree. Her we can see that on-base-percentage has a positive increasing effect within the decision tree, showing within the decision tree that teams with higher on-base percentages are more likely to be classified as winning the World Series. An interesting plot on the bottom show the interaction of the number of times a team was caught stealing (CS) a base. Here we can see that there is little dependency on the SHAP interaction value based on whether the team was caught many or few times.

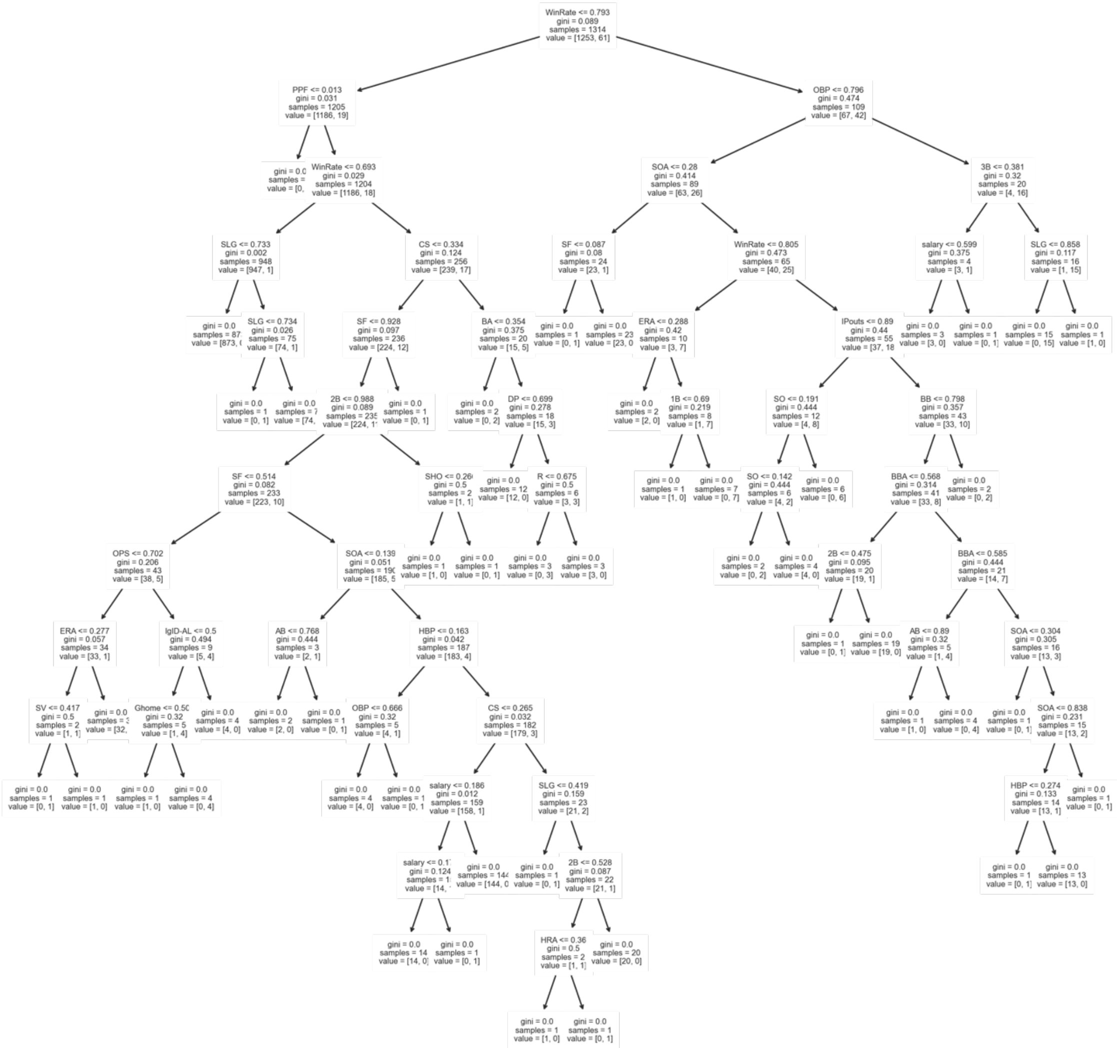

Figure 14. Decision Tree Plot

Finally Figure 14 shows a complete output of the decision tree model. Seen in Figure 1, decision trees are relatively simple to visualize and explain by themselves. This tree is a good example where the depth of the tree is not too severe, and the branching factor is low. This visualization allows us to see the number of total samples at each node as well as the number of positive and negative samples that appear at each node.

Conclusion

From this we can see that there are many different features that are strong indicators of a team's World Series chances. Offensive and defensive statistics both play an equal role in the determination of a World Series champion showing that a well-balanced team is required. Additionally we saw that the various machine learning models emphasized some different features in determining their prediction of World Series championship winning teams. The accompanying code has additional visualizations that were not shown here to save space in the paper.

References

[1]S. Lahman, "Baseball Databank", *SeanLahman.com*, 2021. [Online]. Available: https://www.seanlahman.com/baseball-archive/statistics/. [Accessed: 30- Apr- 2022].

[2]"MLB Payrolls By Season", *TheBaseballCube.com*, 2022. [Online]. Available: https://www.thebaseballcube.com/page.asp?PT=payroll_year&ID=2022. [Accessed: 30- Apr- 2022].

[3]C. Perin, R. Vuillemot, C. Stolper, J. Stasko, J. Wood and S. Carpendale, "State of the Art of Sports Data Visualization", *Computer Graphics Forum*, vol. 37, no. 3, pp. 663-686, 2018. Available: 10.1111/cgf.13447 [Accessed 16 April 2022].

[4]R. Soleymani, E. Granger and G. Fumera, "F-measure curves: A tool to visualize classifier performance under imbalance", *Pattern Recognition*, vol. 100, p. 107146, 2020. Available: 10.1016/j.patcog.2019.107146 [Accessed 16 April 2022].

[5]P. Angelov, E. Soares, R. Jiang, N. Arnold and P. Atkinson, "Explainable artificial intelligence: an analytical review", *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 5, 2021. Available: 10.1002/widm.1424 [Accessed 16 April 2022].

[6]S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017. Available: https://arxiv.org/abs/1705.07874. [Accessed 14 April 2022].

[7]S. Lundberg et al., "From local explanations to global understanding with explainable AI for trees", *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56-67, 2020. Available: 10.1038/s42256-019-0138-9 [Accessed 15 April 2022].