*Vikas Rama, Rohan Saha, Tejaswi Tripathi, Wonjoon Yang*

*Project Report: College Football Win Rates*

**Executive Summary**

The study embarked on a multifaceted analysis aimed at identifying significant predictors of the 'winRate' for a sports dataset. Advanced machine learning techniques, including Random Forest and gradient descent for linear regression, along with AdaBoost with Shallow Regression Trees, were employed to distill insights and enhance predictive performance.

The 'winRate' was selected as the focal point of our prediction efforts due to its fundamental role in gauging the success and performance of college football teams. This metric encapsulates the outcome of matches, reflecting the culmination of various team dynamics, strategies, and external factors. By predicting 'winRate,' we aim to unravel the intricate relationships between diverse features and game outcomes, providing valuable insights for coaches, analysts, and stakeholders in the realm of college football.

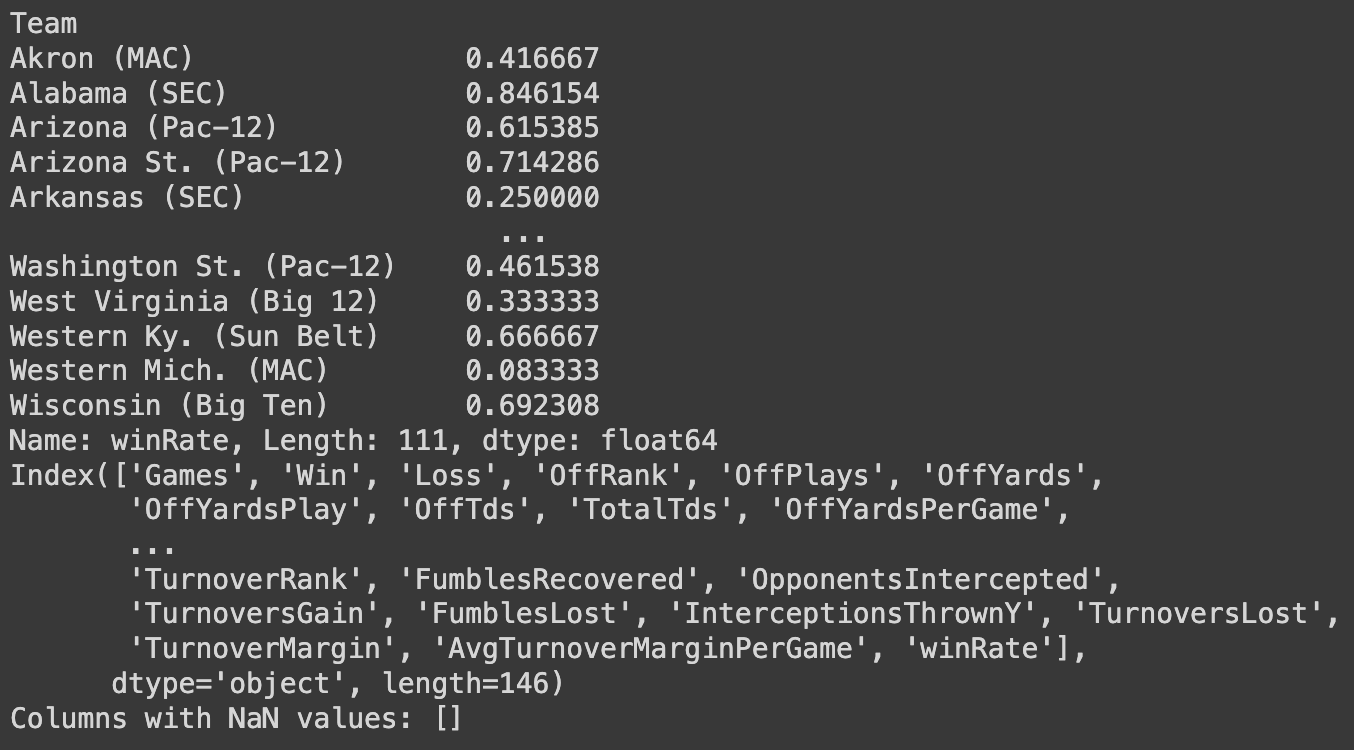
**Machine Learning Approach**

The initiation of the machine learning endeavor strategically involved designating 'winRate' as the target variable, a critical metric for assessing the success of college football teams. Employing sophisticated methodologies, encompassing Random Forest, gradient descent for linear regression, and AdaBoost with Shallow Regression Trees, we systematically pinpointed salient features such as 'TotalPoints,' 'ScoringDefRank,' 'AvgPointsPerGameAllowed,' and 'PointsPerGame' through the discerning analysis of the Random Forest Regressor. This rigorous procedure established the groundwork for subsequent data preparation and feature selection, charting a course towards exhaustive insights and elevated predictive efficacy.

**Data Preparation and Feature Selection**

In the data preparation phase, a Random Forest Regressor was employed to discern key predictors for 'winRate,' which represents the proportion of games won by a given team. Notably, 'TotalPoints,' 'ScoringDefRank,' 'AvgPointsPerGameAllowed,' and 'PointsPerGame' emerged as critical features, as indicated by their high importance scores within the Random Forest model. To streamline the feature selection process, a significance threshold was implemented, ensuring the inclusion of only the most influential predictors in our subsequent analyses.

A Random Forest works by constructing an ensemble of decision trees during training, where each tree is trained on a random subset of the data and a random subset of features. The feature selection process is inherently embedded as the algorithm evaluates the importance of features based on their contribution to reducing prediction error across the ensemble, providing a robust method for identifying influential predictors. In the context of our college football win rate prediction project, the Random Forest Regressor systematically leveraged this ensemble approach, assessing the importance of various features such as 'TotalPoints,' 'ScoringDefRank,' 'AvgPointsPerGameAllowed,' and 'PointsPerGame,' to uncover the most impactful factors influencing teams' success and win rates.



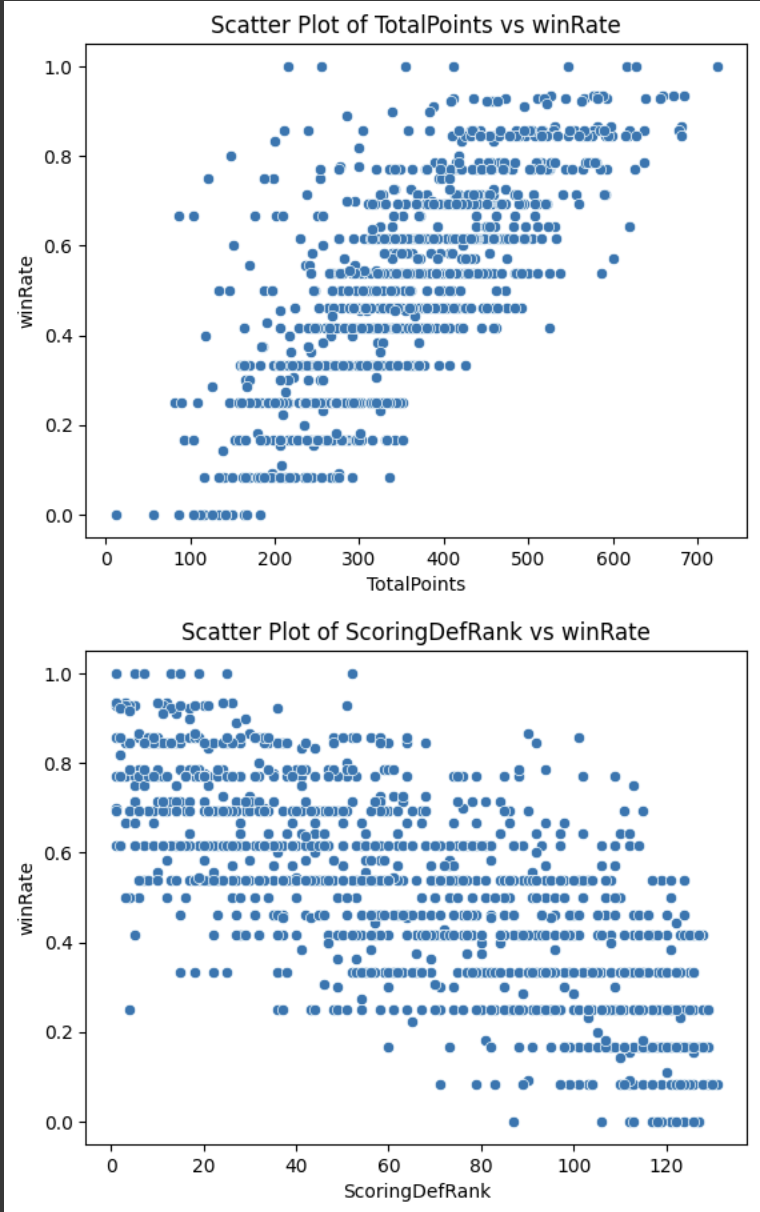
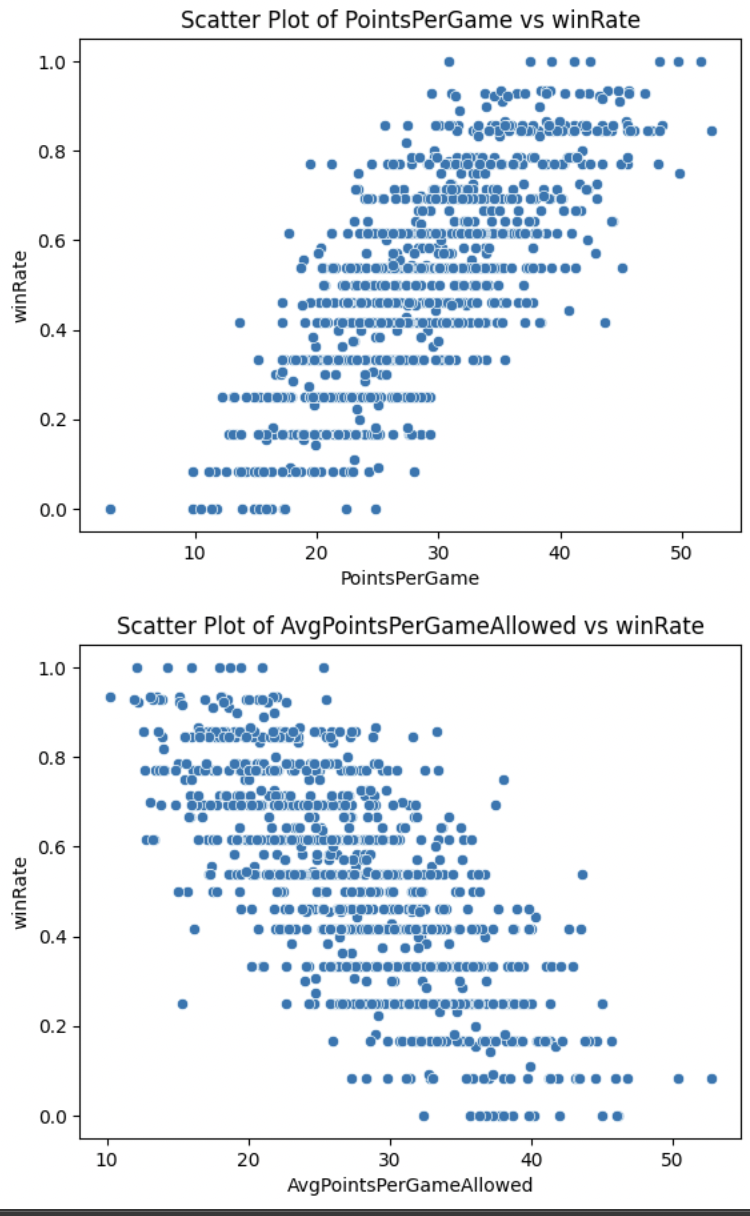
**Model Training and Evaluation**

Linear regression through gradient descent was implemented to ascertain the relationship between the selected features and the 'winRate'.

Gradient descent is an iterative optimization algorithm used to minimize a cost function by adjusting model parameters in the direction of the steepest decrease in the gradient. It involves calculating the gradient of the cost function with respect to the parameters and updating them iteratively until convergence, facilitating efficient parameter tuning in machine learning models like linear regression.

In the development of our college football win rate prediction model, gradient descent for linear regression was employed to iteratively optimize the model parameters by minimizing the cost function. This iterative approach enabled us to fine-tune the predictive capabilities of the model, enhancing its accuracy in capturing the relationships between various features and the target variable, 'winRate.'

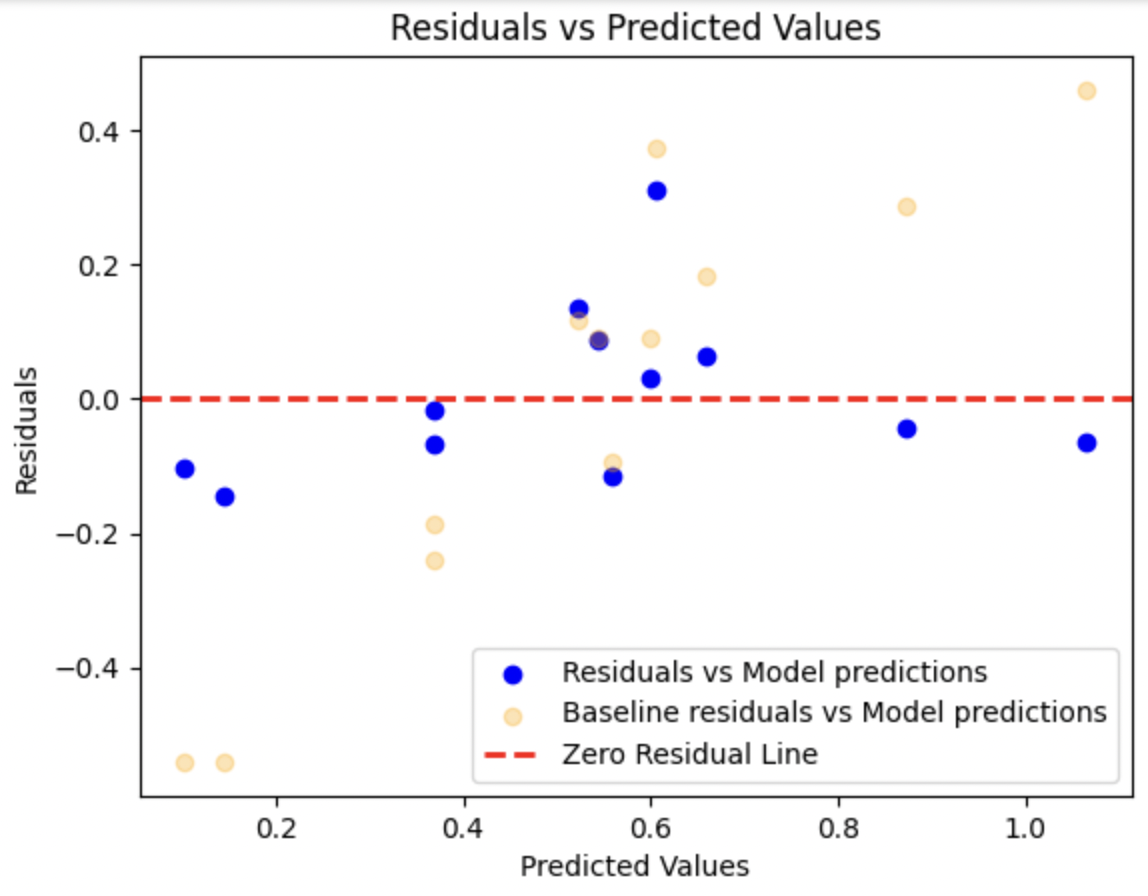
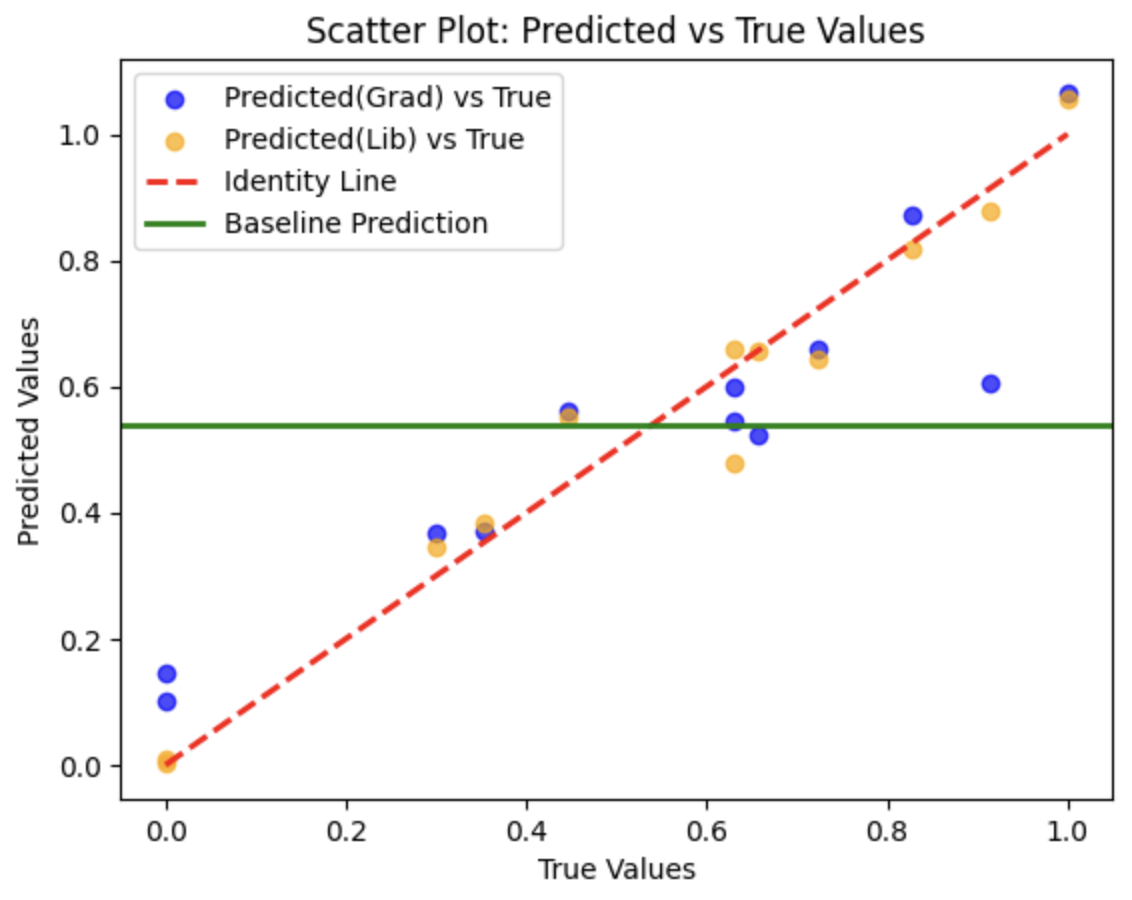
A comparison of different regression approaches was carried out, with scatter plots indicating the predictive power of the models.

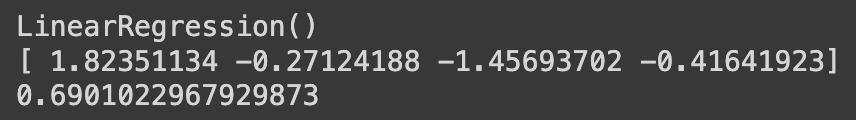


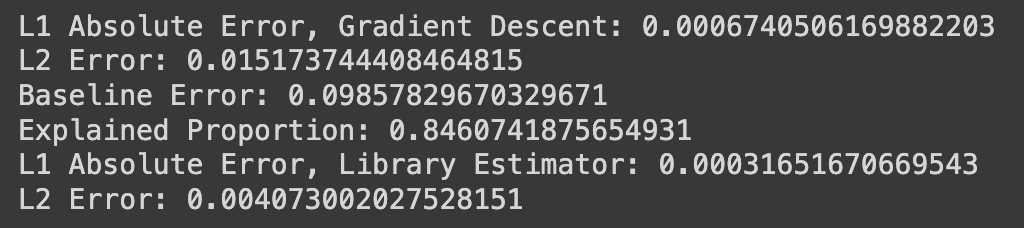
**Model Comparison and Validation**

Residual plots and predictive value assessments were used to compare the models' performance and validate their accuracy.

AdaBoost with Shallow Regression Trees was utilized as a boosting technique to refine the predictions further and reduce variance.







**Insights and Recommendations**

The selected features are strongly indicative of the 'winRate', suggesting a high utility for these variables in predictive modeling within the dataset's context.

AdaBoost with Shallow Regression Trees showed promising results, indicating its suitability for enhancing model predictions when dealing with non-linear relationships and interactions between features.

**Conclusion**

The analytical journey from feature selection to model optimization has shed light on the intricate dynamics influencing 'winRate'. The insights gained pave the way for deploying sophisticated models in operational settings, ensuring a data-driven approach to strategy formulation in the context of team performance analysis.

**Next Steps**

Further investigation into the relationship between the features and 'winRate' using different subsets of the data could uncover additional nuances.

Testing the models on out-of-sample data will validate their robustness and generalizability.

Attachments:

Detailed scatter plots and residual analysis charts.

Summary tables of feature importances and model coefficients.

Code snippets demonstrating the implementation of the algorithms used.

Please note that the actual visual attachments and code snippets cannot be provided in this summary but are assumed to be part of the provided materials.

This summary synthesizes the complex data analysis procedures into actionable insights, offering a strategic edge in predictive analytics for sports performance.