**PROJECT REPORT**

1. **INTRODUCTION**

The National Basketball Association (NBA) is a professional basketball league in North America. The league is composed of 30 teams and is one of the four major professional sports leagues in the United States and Canada. The NBA Data set is all about the performance stats of each team gathered in the year 2017-18 season, the data set consists of 30 rows and 19 columns these columns represent average factors calculated per game. The objective of the project is to build a supervised machine learning algorithm that can predict the winning percentage of the teams playing in the NBA and to highlight the major factors which play an important role in winning the matches.

1. **DISCUSSION**

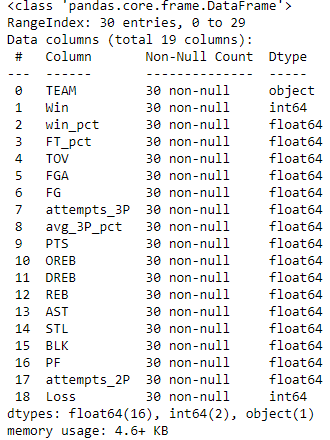
* **Exploratory data analysis:**
* Head of the data set.

**A picture containing application

Description automatically generated**

* Firstly, checking for null values, data types, and the shape of the data.

Below is the output of the info function indicating the data types and non-null count of each variable.



* Checking for outliers in each column and mentioning here some columns that consist of outliers that are negligible. Below mentioned boxplots indicating that there are some outliers present in the variables, but these outliers are not typically outliers because they are nearby to values of upper Whiskers and lower Whiskers.

Chart, box and whisker chart

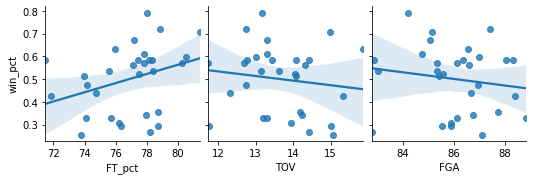
Description automatically generatedChart, box and whisker chart

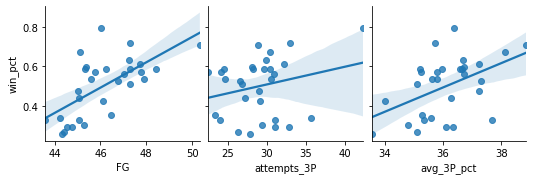
Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

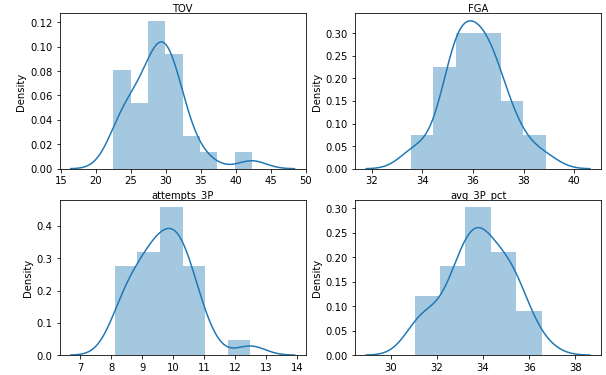
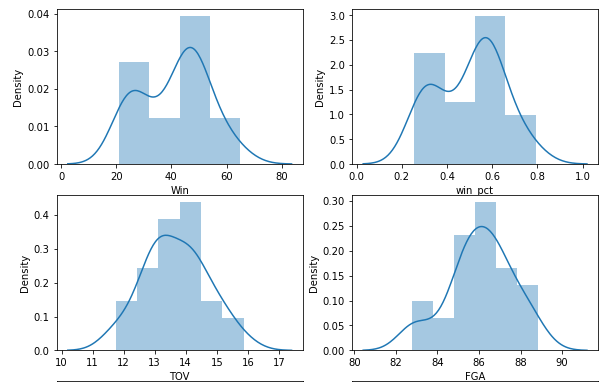
Description automatically generated

* Checking for a linear relationship between the dependent and independent variables using a pair plot, here the linear relationship is found to be average for all the independent variables. Mentioning some plots below:

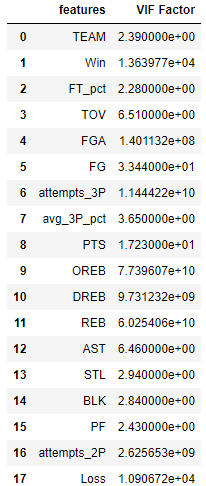




* Checking for the distribution of the data in each column of the overall dataset, there is a normal distribution among all the variables. mentioning some plots below:



* Checking for multicollinearity between independent variables using variance inflation factor, the VIF value of the independent variable less than 5 indicates no multicollinearity and greater than 5 indicates multicollinearity is present among the variables which need to be eliminated. Highlighted arrows indicating such multicollinearity variables are below:



**The arrows pointing the vif values more than 5, must be eliminated**

* **Comparing Algorithms**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Name** | **R2 Score** | **Adjusted R2** | **RMSE Score** |
| Linear Regression | 0.99 | 0.99 | 0.0003 |
| Ridge Regression | 0.99 | 0.99 | 0.0002 |
| Lasso Regression | 0.69 | 0.44 | 0.067 |
| Decision Tree | 0.88 | 0.79 | 0.040 |
| Random Forest | 0.97 | 0.95 | 0.020 |

COmparo

|  |  |
| --- | --- |
| **The minimum value of Y** | **0.305** |

Comparing Both values

* **R2 score: -** R2 score represents the percentage accuracy of the model.
* **Adjusted R2 Score:** - Adjusted r2 score as the name suggests it’s an adjusted accuracy score that must be closer to the r2 score to be a good model.
* **RMSE:** - Root Mean Square error value is compared with the minimum value of the Y test data, The RMSE value must be closer to the minimum value of the Y data which indicates the prediction is more accurate.
* **Factors Influencing Team Success**

****

**NOTE:** \_Ignore win column that’s a winning count column and look at the highlighted column values indicating their contribution to the model accuracy.

1. **REB: -** Average total rebounds per game is the first major factor for the team's success.
2. **FT\_PCT: -** Average free throw percentage per game is the second most factor for team success.
3. **PTS: -** Average points per game is the third most factor for team success.
4. **CONCLUSION**

Each model predicting the winning percentage of Teams differently is visible in the above table. comparing all the models, linear and Ridge models are performing good with high accuracy scores but poor RMSE value than lasso, decision tree, and random forest whereas the Decision tree and Random Forest are performing better with good RMSE value and accuracy score. Considering the RMSE value as a key evaluation metric Decision Tree model is performing better than any algorithm although it has less accuracy score than random forest, correct predictions are more important.

Overall, the Decision Tree Model with an accuracy of 88% and 0.04 RMSE value is best for predicting the winning percentage of NBA teams.