BASEBALL WINNING PREDICTION MODEL

Submitted by:

Abhilekh Verma

Contents

ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. Under their guidance I learned a lot about this project. I had also taken help from YouTube & online videos.

INTRODUCTION

ABOUT BASEBALL

Baseball, game played with a bat, a ball, and gloves between two teams of nine players each on a field with four white bases laid out in a diamond (i.e., a square oriented so that its diagonal line is vertical). Teams' alternate positions as batters (offense) and fielders (defence), exchanging places when three members of the batting team are "put out." As batters, players try to hit the ball out of the reach of the fielding team and make a complete circuit around the bases for a "run." The team that scores the most runs in nine innings (times at bat) wins the game.

PROBLEM STATEMENT

This dataset utilizes data from 2014 Major League Baseball seasons to develop an algorithm that predicts the number of wins for a given team in the 2015 season based on several different indicators of success. There are 16 different features that will be used as the inputs to the machine learning and the output will be a value that represents the number of wins.

DATA ANALYSIS

COLUMNS IN THE DATAFRAME

- Runs
- At Bats
- Hits
- Doubles
- Triples
- Homeruns
- Walks
- Strikeouts
- Stolen Bases
- Runs Allowed
- Earned Runs
- Earned Run Average (ERA)
- Shutouts
- Saves
- Complete Games
- Errors
- Wins (Target Column)

OBJECTIVE

- Our main objective is to predict the number of wins using machine learning algorithms.
- All the parameters will be analysed through Machine Learning algorithms like Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression etc which will help to predict the Number of wins.

DATA DESCRIPTION

- The dataset contains Runs, At Bats, Hits, Doubles, Triples, Homeruns, Walks, Strikeouts, Stolen Bases, Runs Allowed, Earned Runs, Earned Run Average (ERA), Shutouts, Saves, Complete Games and Errors
- The source of data is taken from GitHub. (https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/baseball.csv)

METHEDOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the sales because it gives the best outcome.
- The dependent variable is Wins, whereas other variables i.e., Runs, At Bats, Hits, Doubles, Triples, Homeruns, Walks, Strikeouts, Stolen Bases, Runs Allowed, Earned Runs, Earned Run Average (ERA), Shutouts, Saves, Complete Games and Errors are independent variables.

METRIC USAGE

a. Linear Regression.
b. Lasso Regression.

c. Ridge Regression.
d. Elastic Net Regression.
e. Support Vector Regressor.
f. Random Forest Regressor.

SYSTEM REQUIREMENTS

Hardware and Software Requirements and Tools Used

a) Hardware Requirement:
   i. Intel core i5
   ii. 8 GB Ram

b) Software Requirement:
   i. Python 3.x with packages:
      1. Pandas: Data analysis and manipulation tool
      2. NumPy: Provide support for mathematical functions, random number etc.
      3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
      4. Seaborn: is a library mostly used for statistical plotting in python.
      5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

APPROACH

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- R2 score: is used to evaluate the performance of a linear regression mode.
- Linear Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- Lasso: The Lasso is a linear model that estimates sparse coefficients with l1 regularization.
- Ridge: Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model.
- Elastic Net: is a linear regression model trained with both l1 and l2 -norm regularization of the coefficients.
- Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- Mean Squared Error: this metric gives an indication of how good a model fits a given dataset.
- Root Mean Squared error : is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.
- Z-score: Z-score is also known as standard score gives us an idea of how far a data point is from the mean.
- Label Encoder: Label Encoding refers to converting the labels into numeric form.
- Random Forest Regressor: A Random Forest is an ensemble technique capable of performing both regression and classification tasks.
- Standard Scaler: Standard Scaler. Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance).
- Support Vector Regressor: is a regression algorithm that supports both linear and non-linear regressions. This method works on the principle of the Support Vector Machine.

TESTING OF IDENTIFIED APPROACH(Algorithms)

a. Train Test Split
b. Linear Regression
c. Lasso Regression
d. Ridge Regression
e. Elastic Net Regression
f. Grid Search CV
g. Cross Validation

KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

- Analysed data for any unique values.
- Analysed data for distribution.
- Checked and removed outliers through z-score method.
- Removed skewness present in the dataset.
- Done Standard Scaling.
- Cross validate the r2 score from overfitting.
- Hyper Parameter tuning.

CONCLUSION

As our conclusion we proclaim that, after checking r2 score, cross validation, Ensemble Techniques, we declare Linear Regression predicting 100% accuracy is best suited model for our purpose of predicting Number of win.

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc.

It also helped me to learn how to apply various model techniques on data and enable predications.

REFERENCES

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.