

BASEBALL WINNING PREDICTION MODEL



Submitted by:
UJJWAL PRATIK

Contents

ACKNOWLEDGMENT	3
INTRODUCTION	4
ABOUT BASEBALL	4
PROBLEM STATEMENT	4
DATA ANALYSIS	5
COLUMNS IN THE DATAFRAME.....	5
OBJECTIVE	5
DATA DESCRIPTION.....	6
METHODOLOGY	6
METRIC USAGE	6
SYSTEM REQUIREMENTS	7
APPROACH	8
IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES.....	8
TESTING OF IDENTIFIED APPROACH(Algorithms)	8
KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION	9
CONCLUSION	10
LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE	10
REFERENCES	11

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. their suggestions and directions have helped in the completion of this project. I had also taken help from YouTube & online videos.

INTRODUCTION

ABOUT BASEBALL

Baseball is a bat-and-ball game played between two opposing teams who take turns batting and fielding. The game proceeds when a player on the fielding team, called the pitcher, throws a ball which a player on the batting team tries to hit with a bat. The objective of the offensive team (batting team) is to hit the ball into the field of play, allowing its players to run the bases, having them advance counter clockwise around four bases to score what are called "runs". The objective of the defensive team (fielding team) is to prevent batters from becoming runners, and to prevent runners' advance around the bases. A run is scored when a runner legally advances around the bases in order and touches home plate (the place where the player started as a batter). The team that scores the most runs by the end of the game is the winner.

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)

	W	R	AB	H	2B	3B	HR	BB	SO	SB	RA	ER	ERA	CG	SHO	SV	E
0	95	724	5575	1497	300	42	139	383	973	104	641	601	3.73	2	8	56	88
1	83	696	5467	1349	277	44	156	439	1264	70	700	653	4.07	2	12	45	86
2	81	669	5439	1395	303	29	141	533	1157	86	640	584	3.67	11	10	38	79
3	76	622	5533	1381	260	27	136	404	1231	68	701	643	3.98	7	9	37	101
4	74	689	5605	1515	289	49	151	455	1259	83	803	746	4.64	7	12	35	86

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>)

METHODOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learning 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.
- Zscore: 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 zcore method.
- Removed skewness present in the dataset.
- Done Standard Scaling.
- Cross validate the r^2 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.

```
loaded_model=pickle.load(open('baseball_win_pred.pkl','rb'))
result=loaded_model.score(x_test, y_test)
print(result)
```

```
0.94610173985087
```

```
conclusion=pd.DataFrame([loaded_model.predict(x_test)[:],pred_decision[:]],index=['predicted','original'])
conclusion
```

	0	1	2	3	4	5	6
predicted	91.01	76.41	68.67	81.52	84.05	77.94	85.44
original	91.01	76.41	68.67	81.52	84.05	77.94	85.44

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.