MLB Game Predictor Machine Learning Model

4/18/2023: Michael Brown

Objective

Create a machine learning model that is optimized to predict the outcome of an Major League Baseball (MLB) game based on historical data.

Motivation

Sports provide a relevant and tangible application of machine learning in everyday life. Sports and statistics are heavily integrated in 2023, and will continue to develop as technology advances. Both organizations or spectators could utilize machine learning models drive decisions and make predictions.

Why MLB?

MLB was chosen due to the large amount of historical data available. 30 teams each play 162 games a season. Utilizing data across 20 season results in 48,600 data points.

Additionally, there exists a plethora of supplemental statistics that could be incorporated to build a robust machine learning mode.

Data Source

Retrosheet game logs - includes all game events from 2000 to 2023

(https://www.retrosheet.org/gamelogs/index.html)

Sean Lahman baseball database - Contains specific player statistics. This could be used to calculate team and player statistics.

(https://www.kaggle.com/datasets/freshrenzo/lahmanbaseballdatabase?select=Pitching.csv)

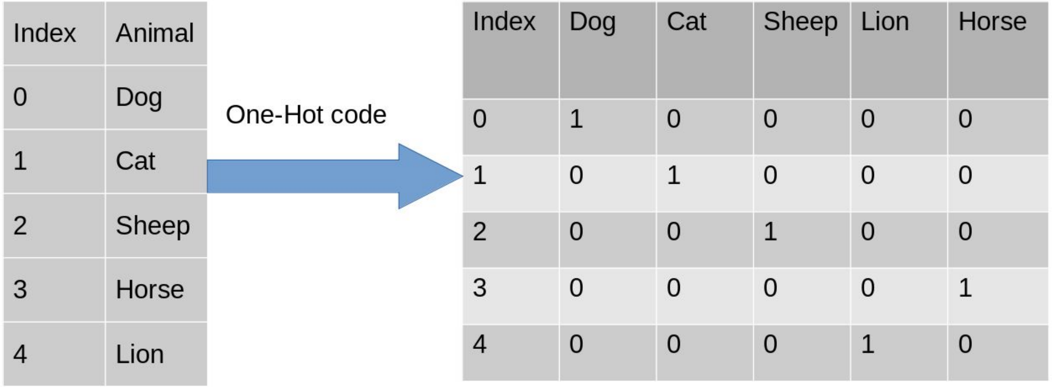
Data Target and Features

Features

Target

dt	away_team	home_team	away_pitcher	home_pitcher	away_1_id	away_2_id	away_5_id	away_6_id	outcome
2020-07-23	SFN	LAN	cuetj001	may-d003	yastm001	florw001	pench001	mccaj002	1
2020-07-23	NYA	WAS	coleg001	schem001	hicka001	judga001	gardb001	sancg002	0
2020-07-24	COL	TEX	marqg001	lynnl001	dahld001	stort001	murpd006	mcmar001	1

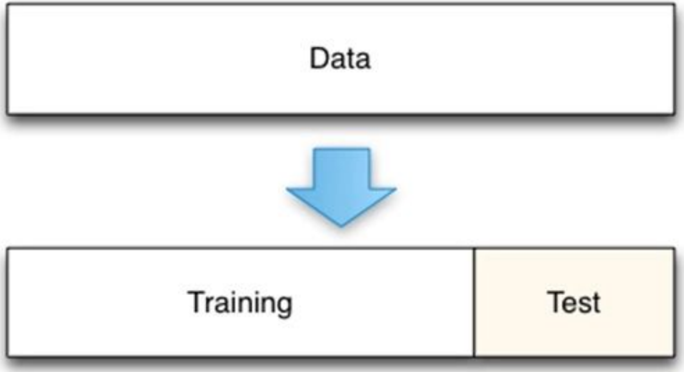
Data Processing | Data was split, encoded, and scaled.



Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

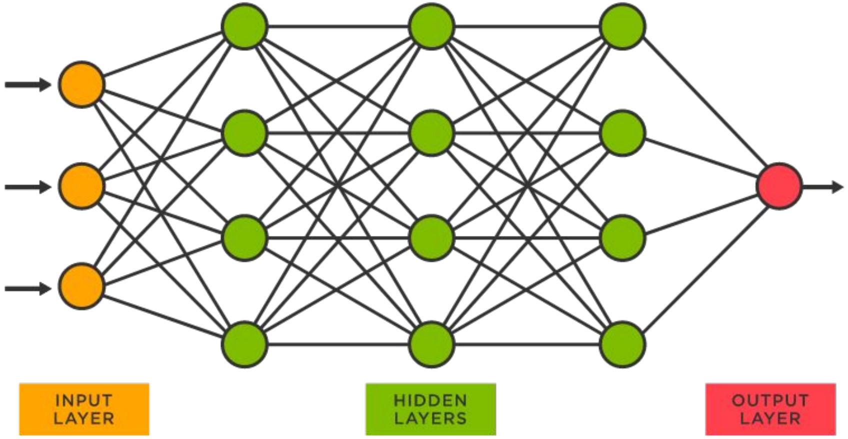
Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0



Machine Learning Model | Logistic Regression Neural Network

Layer Configuration

- Inputs: 9986
- 1st Hidden Layer: 40 nodes
- 2nd Hidden Layer: 13 nodes
- Output Layer: 1 node



Results | Neural Network

Accuracy: 0.5372

Loss: 4.829

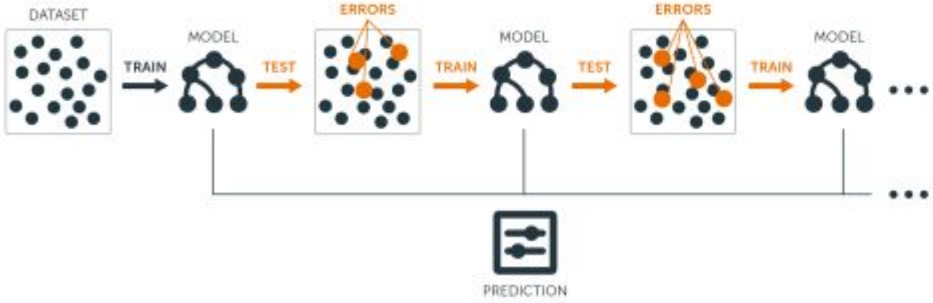
The neural network model showed promising results. Accuracy is better than 50% (expected average if randomly selecting outcomes), but the model does not appear to be overfit. The model could likely be further optimized by running additional iterations on the number of nodes, layers, and epochs.

Machine Learning Model | Gradient Boosting Machine

GBM is a machine learning algorithm that combines multiple weak models (e.g., decision trees) to create a strong predictive model. Each model is trained on the residuals of the previous model, so the final model can correct the errors of the previous models.

GBM is particularly suitable for predicting complex and nonlinear relationships between variables, which is often the case in baseball analytics. For example, the performance of a pitcher or batter can depend on many factors such as the opposing team, the weather, the location of the game, etc. GBM can capture these complex interactions and provide accurate predictions.

The same dataset was used for for the GBM model as for the neural network.



Results | Gradient Boosting Method

Accuracy: .995

The gradient boosting machine appears to be overfit to the data. An accuracy of 99% is extremely unlikely for predicting sports outcomes and is not to be believed. The model should be validated and revised until a more accurate

Next Steps

- Address overfitting of GBM

- Add in supplementary team statistics to the model including:
  - Batting average
  - Earned run average (ERA)
  - On base percentage (OBS)
  - On base plus slugging (OPS)

Questions?