Project 4

Monte Carlo Simulation: LSTM, Random Forest, and Linear Regression

This Monte Carlo simulation compares the performance of three regression models: - Linear Regression - Ridge Regression - Random Forest in predicting WAR, using 30 synthetic datasets (10 per correlation level: 0.0, 0.5, 0.99). The models will be evaluated on Mean Squared Error (MSE). The results will be summarized using the mean and standard deviation of the MSE for each combination of model and correlation structure.

Scientific or Statistical Question

How do Linear Regression, Random Forest, and Ridge Regression perform in predicting a baseball player's WAR under varying levels of predictor correlation (none, mild, high)?

Data

The generate_data function creates synthetic data that mimics our data using the original data's means and standard values.

Estimates

From there, we calculate

WAR = Sum(X + c), where

weight array = [0.1, 0.1, −0.1, 0.1, −0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1]

X = Multivariate Normal(mean vector (from dataset.describe()), covariance matrix)

c = noise. follows standard normal distribution

Methods

We are using three models to compare: Linear Regression, Ridge Regression, and Random Forest. Ridge Regression will have an alpha (parameter) value of 0.1 and Random Forest will be restricted to max_depth = 12. Ridge Regression is expected to be similar to Linear Regression in performance but should be better than Linear Regression in higher correlation because of the L2 regularizer. Random Forest is expected to perform much better in higher correlation settings but it is hard to say if it will be better than Ridge Regression without testing.

Performance Criteria

We are estimating the MSE (mean squared error) and r^2 from the data above. They measure model performance on the test set for each synthetic dataset.

Simulation Plan

Number of simulations: 30 simulations, 10 for each correlation level. 100 samples per simulation. 70% of the data will be for training and the rest for testing.

Parameter settings: Ridge alpha = 0.1, Random Forest max_depth = 12

What will be recorded: MSE and r^2

Any changes from Project III's code: We removed LSTM model

Anticipated Challenges or Limitations

Variables such as age are difficult to factor in because some players can peak in their late 20s and fall off very quickly while others are able to maintain their high performance throughout their 30s and sometimes even into their 40s. Also, there are factors such as: a player happened to be on a team where the position that they played was already filled by a hall-of-fame level player so they did not get any playing time early on in their career.