linreg

2024-12-07

Comparison to linear model

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
train_data_reg <- read.csv("xgb_train.csv")
test_data_reg <- read.csv("xgb_test.csv")
test_y <- read.csv("xgb_test_y.csv")
df <- read.csv("match_data_clean.csv")
#Linear regression model
lin_model <- lm(total.goals~., data=as.data.frame(train_data_reg))
#Predictions
lin_pred <- predict(lin_model, test_data_reg)
```

```
## Warning in predict.lm(lin_model, test_data_reg): prediction from rank-deficient
## fit; attr(*, "non-estim") has doubtful cases
```

```
write.csv(lin_pred, "lin_pred.csv")
#Evaluate performance
#MSE
mean((test_y - lin_pred)^2)
```

```
## Warning in mean.default((test_y - lin_pred)^2): argument is not numeric or
## logical: returning NA
```

```
## [1] NA
```

```
#MAE
MAE(test_y, lin_pred)
```

```
## Warning in mean.default(abs(pred - obs), na.rm = na.rm): argument is not
## numeric or logical: returning NA
```

```
## [1] NA
```

```
#RMSE
RMSE(test_y, lin_pred)
```

```
## Warning in mean.default((pred - obs)^2, na.rm = na.rm): argument is not numeric
## or logical: returning NA
```

```
## [1] NA
```

```
mean(df$total.goals)
```

```
## [1] 2.6875
```

As suspected, the model's performance is poor. All three of the evaluation metrics are horrendous considering that the mean number of total goals is 2.6875. Furthermore, the fit is rank-deficient since there are more predictors than observations. This results in very poor predictions. This example illustrates the importance of reducing multicollinearity and variable selection.