Final Draft

Johnny Ferrara

2023-12-07

Introduction

This report analyzes a classification problem of the performance statistics of the Kansas City Chiefs football team to determine which factors are most influential in securing a win. The dataset contains a variety of game-related metrics, shown below. Prior to analysis, categorical variables were converted to factors, and non-essential and redundant variables were removed to streamline the dataset, ensuring a focused examination of the most impactful factors.

Variable Description

Results - Chiefs winning or losing the game (Win / Loss)
Day - What day the Chiefs played (Monday / Thursday / Saturday / Sunday)
Time - Time the Chiefs played (Noon / Afternoon / Night)
OT - Overtime games (Yes / No)
Location - Location of the game (Home / Away)
OFF1stD - Chiefs offensive first downs
OFFPassY - Chiefs offensive passing yards
OFFRushY - Chiefs offensive rushing yards
OFFTO - Chiefs offensive turnovers
DEF1stD - First downs given up by the Chiefs defense
DEFPassY - Passing yards given up by the Chiefs defense
DEFRushY - Rushing yards given up by the Chiefs defense
DEFTO - Turnovers that the Chiefs defense caused

Methods

The primary analytical method was logistic regression, chosen for its suitability in modeling binary outcomes like wins or losses. The regression formula included game time, location, offensive and defensive statistics, and turnovers. The model’s robustness was assessed through 10-fold cross-validation to mitigate overfitting and ensure generalizability. Additionally, stepwise regression using Akaike’s Information Criterion (AIC) was employed to refine the model by selecting the most significant predictors. Ridge regression was also conducted, introducing regularization to address multicollinearity and enhance model stability, with the optimal complexity parameter (lambda) determined through cross-validation. Bagging was applied to further stabilize the predictions, creating a number of models to reduce variance.

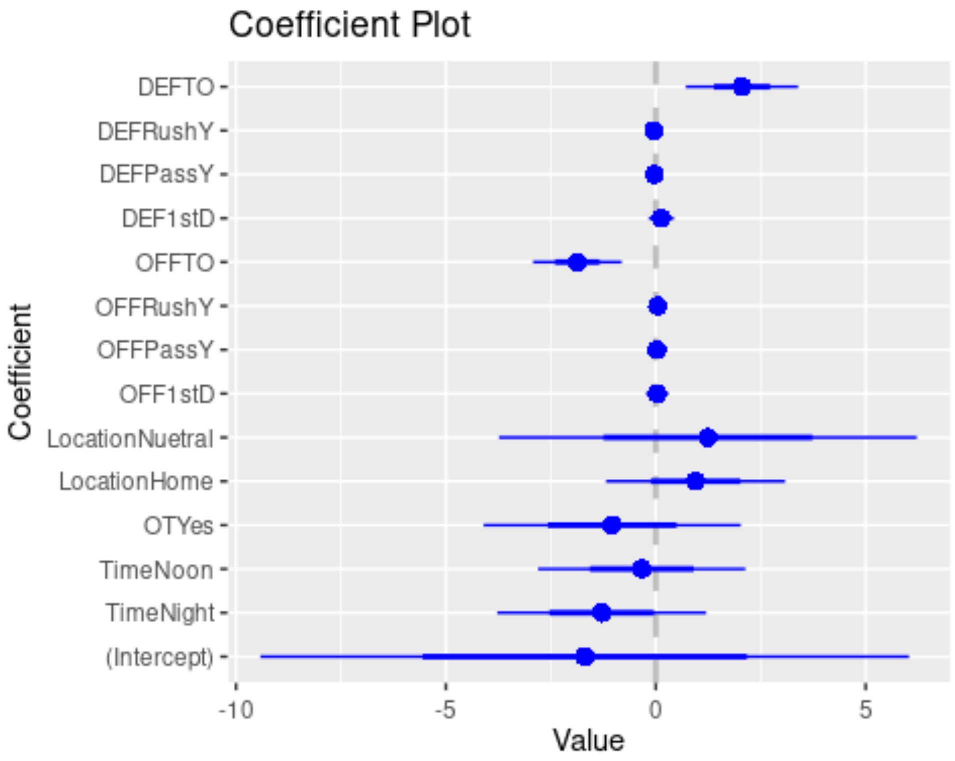
```r
logistic_model <- glm(Result ~ Time + OT + Location + OFF1stD + OFFPassY +
                        OFFRushY + OFFTO + DEF1stD + DEFPassY +
                        DEFRushY + DEFTO, family = "binomial", data =
chiefs_data)
summary(logistic_model)
```

```
##
## Call:
## glm(formula = Result ~ Time + OT + Location + OFF1stD + OFFPassY +
##     OFFRushY + OFFTO + DEF1stD + DEFPassY + DEFRushY + DEFTO,
##     family = "binomial", data = chiefs_data)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.69289    3.86314  -0.438 0.661230
## TimeNight        -1.28862    1.24424  -1.036 0.300357
## TimeNoon         -0.33489    1.23457  -0.271 0.786191
## OTYes            -1.03852    1.53158  -0.678 0.497726
## LocationHome      0.94714    1.06738   0.887 0.374891
## LocationNuetral   1.24276    2.48857   0.499 0.617508
## OFF1stD           0.03648    0.13400   0.272 0.785449
## OFFPassY          0.02822    0.01273   2.217 0.026616 *
## OFFRushY          0.04199    0.02042   2.057 0.039710 *
## OFFTO            -1.86822    0.52800  -3.538 0.000403 ***
## DEF1stD           0.13249    0.15099   0.877 0.380237
## DEFPassY         -0.02838    0.01244  -2.281 0.022571 *
## DEFRushY         -0.04495    0.01647  -2.730 0.006329 **
## DEFTO             2.05235    0.66845   3.070 0.002138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 100.862  on 95  degrees of freedom
## Residual deviance:  42.054  on 82  degrees of freedom
## AIC: 70.054
##
## Number of Fisher Scoring iterations: 7
```

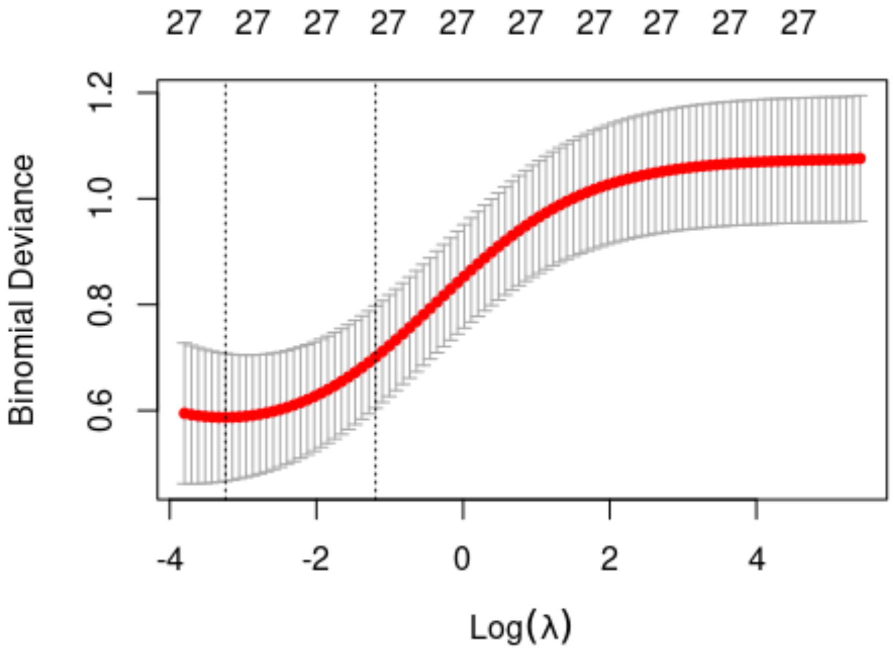
```
stepwise_model <- stepAIC(logistic_model, direction = "both", trace = FALSE)
summary(stepwise_model)
```

```
##
## Call:
## glm(formula = Result ~ OFFPassY + OFFRushY + OFFTO + DEFPassY +
##     DEFRushY + DEFTO, family = "binomial", data = chiefs_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.665770   3.051807   0.218 0.827308
## OFFPassY    0.023218   0.009006   2.578 0.009937 **
## OFFRushY    0.034643   0.012334   2.809 0.004974 **
## OFFTO       -1.602344   0.436499  -3.671 0.000242 ***
## DEFPassY    -0.019219   0.006726  -2.857 0.004270 **
## DEFRushY    -0.034767   0.011114  -3.128 0.001759 **
## DEFTO        1.670835   0.536013   3.117 0.001826 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 100.862  on 95  degrees of freedom
## Residual deviance:  45.279  on 89  degrees of freedom
## AIC: 59.279
##
## Number of Fisher Scoring iterations: 7
```

```
# Coefficients Plot for Logistic Regression
coefplot(logistic_model)
```



```
plot(cv_ridge_model)
```



```
# Bagging with Ridge Regression
print(bagged_ridge)

##
## Bagging classification trees with 100 bootstrap replications
##
## Call: bagging.data.frame(formula = Result ~ ., data = chiefs_data,
##     nbagg = 100, coob = TRUE)
##
## Out-of-bag estimate of misclassification error:  0.1354
```

Results

The logistic regression revealed key statistics significantly associated with game outcomes. Offensive passing yards (OFFPassY) and rushing yards (OFFRushY) had positive coefficients, indicating their contribution to a higher probability of winning. Conversely, offensive turnovers (OFFTO) negatively impacted the chances of winning. Defensive passing yards (DEFPassY) and rushing yards (DEFRushY) negatively influenced the probability of winning, while defensive turnovers (DEFTO) had a positive effect. The stepwise regression refined the model to six significant predictors: OFFPassY, OFFRushY, OFFTO, DEFPassY, DEFRushY, and DEFTO. This final model is interesting because for most teams, location of the games matter, but that variable was not included. Also at what time the Chiefs play, it is not significant in predicting wins or losses if they play at noon, afternoon, or night. The cross-validated Ridge regression supported the regularization of coefficients, emphasizing similar predictors. Bagging with 1000 bootstrap replications demonstrated an out-of-bag misclassification error of 12.5%, and bagging with 100 bootstraps showed an out-of-bag misclassification error of 13.54%.

Discussion

The analysis suggests that while offensive efforts in both passing and rushing are crucial, minimizing offensive turnovers is the most important factor in determining wins. Defensive performance, particularly the ability to force turnovers, also plays a crucial role in the team's success. The regularization techniques used in Ridge regression and the approach in bagging prove the logistic regression findings, highlighting consistent factors across different statistical methods.

Conclusion

This study identifies offensive and defensive yardages, along with turnover metrics, as critical predictors of the Kansas City Chiefs' winning outcomes. Future research could explore time-series analyses to evaluate performance trends or apply machine learning techniques for potentially uncovering nonlinear relationships. Extending the dataset to include more seasons, player-level data, or situational variables could provide a better understanding of the determinants of the Chiefs winning a football game.