Iteration 1: Random forest preliminary

Panacea

Used random forest to predict Superbowl\_Winner by using the entire dataset. Replaced missing values with -99999, split into 75-25 train test.

Where it went wrong: 1450 rows for losers, 50 rows for winners. This led to the model predicting everyone as losers because it has been trained too much on loser data.

Iteration 2: Logistic Regression

Removed pre 1990 data because too many missing columns. Ran Information value function to determine the top 40 variables which predict superbowl winner. Narrowed down to the below ones.

Wins

Losses

Offense\_field\_goals\_XPM

Offense\_scoring\_Rk

Offense\_touchdowns\_Rk

Offense\_field\_goals\_XP\_Att

Offense\_kicking\_KO

Defense\_scoring\_Rk

Offense\_touchdowns\_Total

Offense\_touchdowns\_Rsh

Offense\_game\_stats\_TO

Defense\_rushing\_Rk

Offense\_game\_stats\_Rk

Defense\_touchdowns\_Rk

Defense\_game\_stats\_1st.G

Offense\_touchdowns\_Rec

Offense\_passing\_Avg

Defense\_scoring\_XPM

Defense\_touchdowns\_Total

Offense\_game\_stats\_3rd\_Pct

Defense\_game\_stats\_Pts.G

Defense\_game\_stats\_4th\_Att

Defense\_game\_stats\_Yds.P

Offense\_game\_stats\_Pts.G

Offense\_rushing\_Att.G

Defense\_passing\_1st.

Defense\_rushing\_Att.G

Defense\_tackles\_Sck

Defense\_passing\_Sck

Offense\_rushing\_1st

Defense\_rushing\_1st

Defense\_interceptions\_PDef

Offense\_game\_stats\_3rd\_Att

Offense\_offensive\_line\_Exp

Defense\_game\_stats\_Rk

Offense\_game\_stats\_Yds.P

Defense\_passing\_Att.G

Offense\_punting\_Punts

Offense\_punting\_Rk

Offense\_passing\_1st

Created a correlation matrix to further reduce these 40 to 15 by removing the highly correlated variables, then removed the below ones,

temp$Losses <- NULL

temp$Offense\_scoring\_Rk <- NULL

temp$Offense\_touchdowns\_Rk <- NULL

temp$Offense\_field\_goals\_XP\_Att <- NULL

temp$Offense\_kicking\_KO <- NULL

temp$Offense\_touchdowns\_Total <- NULL

temp$Defense\_touchdowns\_Rk <- NULL

temp$Defense\_scoring\_XPM <- NULL

temp$Offense\_game\_stats\_Rk <- NULL

temp$Defense\_passing\_Sck <- NULL

temp$Offense\_game\_stats\_Pts.G <- NULL

Call:

glm(formula = Superbowl\_Winner ~ ., family = binomial(link = "logit"),

data = dataset)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.20086 -0.29786 -0.09632 -0.02257 2.58855

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -13.985719 5.865483 -2.384 0.01711 \*

Wins 0.889663 0.118935 7.480 7.42e-14 \*\*\*

Offense\_touchdowns\_Rsh 0.041990 0.045511 0.923 0.35620

Offense\_touchdowns\_Rec 0.035113 0.032223 1.090 0.27586

Offense\_passing\_Avg 0.563956 0.338847 1.664 0.09604 .

Defense\_touchdowns\_Total -0.035743 0.034799 -1.027 0.30436

Offense\_game\_stats\_3rd\_Pct -0.144582 0.055888 -2.587 0.00968 \*\*

Defense\_passing\_1st. -0.077072 0.069686 -1.106 0.26873

Offense\_rushing\_1st 0.008057 0.010097 0.798 0.42488

Defense\_rushing\_1st 0.014474 0.010606 1.365 0.17233

Defense\_interceptions\_PDef -0.001124 0.001890 -0.595 0.55215

Offense\_game\_stats\_3rd\_Att 0.039309 0.016616 2.366 0.01799 \*

Offense\_punting\_Punts -0.046917 0.021757 -2.156 0.03105 \*

Call:

glm(formula = Superbowl\_Winner ~ ., family = binomial(link = "logit"),

data = dataset)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.15866 -0.29923 -0.09431 -0.02232 2.62156

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.477e+01 5.815e+00 -2.539 0.0111 \*

Wins 9.402e-01 1.109e-01 8.477 <2e-16 \*\*\*

Offense\_touchdowns\_Rsh 1.834e-02 3.984e-02 0.460 0.6453

Offense\_passing\_Avg 6.713e-01 3.244e-01 2.069 0.0385 \*

Defense\_touchdowns\_Total -1.683e-02 3.004e-02 -0.560 0.5752

Offense\_game\_stats\_3rd\_Pct -1.205e-01 5.085e-02 -2.369 0.0178 \*

Defense\_passing\_1st. -8.428e-02 6.921e-02 -1.218 0.2234

Offense\_rushing\_1st 8.170e-03 1.012e-02 0.808 0.4193

Defense\_rushing\_1st 1.263e-02 1.040e-02 1.215 0.2244

Defense\_interceptions\_PDef -8.773e-04 1.874e-03 -0.468 0.6397

Offense\_game\_stats\_3rd\_Att 3.583e-02 1.624e-02 2.206 0.0274 \*

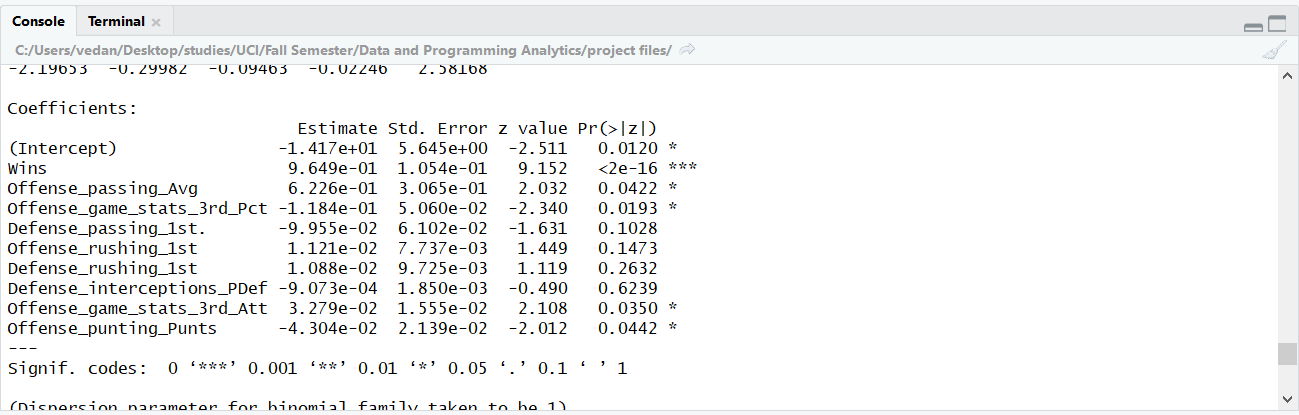
Offense\_punting\_Punts -4.448e-02 2.156e-02 -2.063 0.0391 \*

FALSE TRUE

0 28 0

1 2 2

**POST 1990 DATA, LOGISTIC REGRESSION MODEL VARIABLE SIGNIFICANCE**



**1967 ONWARDS DATA LOGISTIC REGRESSION MODEL VARIABLE SIGNIFICANCE**

