First Down

Sam Hughes

1/8/2022

An important part of a fourth down model is to check what will happen after the fourth down, because a missed field goal, punt, successful conversion attempt, and failed conversion attempt all result in a first down. Therefore, I devised a model to predict the expected points given the yardline on every first and 10/goal. While this is given to me via the play by play data I used, I felt as if using that data would be against the spirit of the project because I wanted to do everything from simple play by play data that could be found anywhere. Therefore, I created a long formula that would check every single first down in between 2010 and 2020, and return the next score (positive if the original offense scores, negative if the original defensive team scores). Below is the work to create a model based off that acquired data.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.3
```

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

```
# load in data from csv
fd = read.csv("first_down_expected_points.csv")[2:3]

### Make model that predicts the expected points on a first down
```

```r
# Split data
set.seed(123)
training_samples_fd = createDataPartition(fd$points_scored, p = 0.8, list = FALSE)
fd_train = fd[training_samples_fd, ]
fd_test = fd[-training_samples_fd, ]

# aggregate test data for graph later
fd_test_agg = fd_test %>%
  group_by(distance) %>%
  summarise("mean" = mean(points_scored), "count" = length(points_scored))

fd_train$sq = fd_train$distance^2
fd_train$cu = fd_train$distance^3

summary(lm(points_scored ~ ., data = fd_train))
```

```
##
## Call:
## lm(formula = points_scored ~ ., data = fd_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9192  -2.6373   0.9323   3.6922   7.1247
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.072e+00  6.326e-02  95.984  < 2e-16 ***
## distance    -9.229e-02  5.222e-03 -17.672  < 2e-16 ***
## sq           5.128e-04  1.215e-04   4.221 2.43e-05 ***
## cu          -2.212e-06  8.246e-07  -2.683  0.00731 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.511 on 128574 degrees of freedom
## Multiple R-squared:  0.09977,    Adjusted R-squared:  0.09975
## F-statistic:  4750 on 3 and 128574 DF,  p-value: < 2.2e-16
```

```r
anova(lm(points_scored ~ ., data = fd_train))
```

```
## Analysis of Variance Table
##
## Response: points_scored
##               Df  Sum Sq Mean Sq    F value    Pr(>F)
## distance       1  288168  288168 14158.7303 < 2.2e-16 ***
## sq             1    1689    1689    82.9677 < 2.2e-16 ***
## cu             1     146     146     7.1963  0.007306 **
## Residuals 128574 2616824      20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

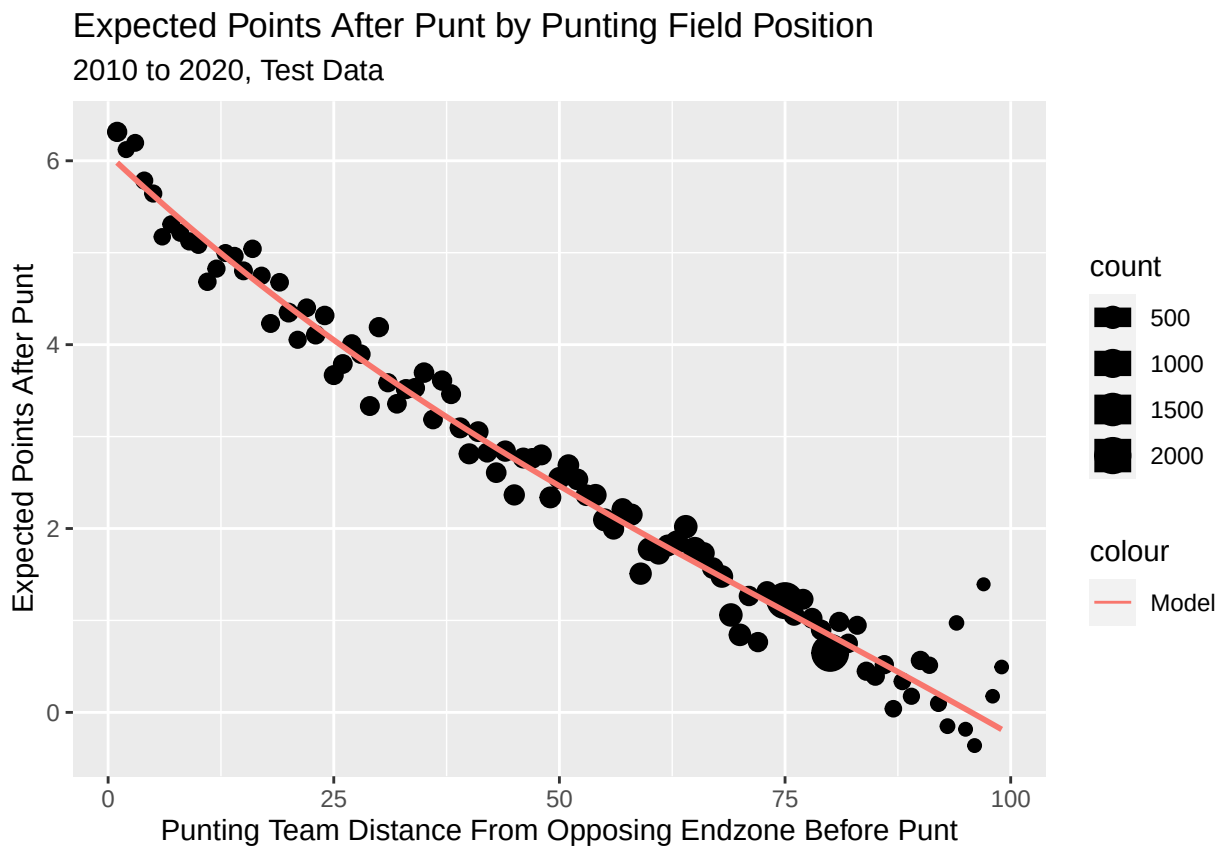
```r
# All values are significant
fd_model = lm(points_scored ~ ., data = fd_train)


fd_equation = function(yardline) {
  fd_model$coefficients[1] + fd_model$coefficients[2] * yardline + fd_model$coefficients[3] * yardline^
}

fd_equation(10)
```

```
## (Intercept)
##    5.197894
```

```r
ggplot(data = fd_test_agg,
       mapping = aes(x = distance,
                     y = mean))+
  ggtitle("Expected Points After Punt by Punting Field Position", subtitle = "2010 to 2020, Test Data")+
  xlab("Punting Team Distance From Opposing Endzone Before Punt")+
  ylab("Expected Points After Punt")+
  geom_point(aes(size = count))+
  geom_line(aes(y = fd_equation(distance),
                size = 3, colour = "Model"))
```



The fit is fantastic on this model, especially considering that this is the test data, so I do not feel as if it is necessary to make any other models.