NFL Betting Analysis

Time Series Modeling to Predict Total Score

Dev Misra

Background

- Decided to investigate the betting lines set on the total points scored for games
    - From this data, I would be able to predict the total points scored for each game played and leverage the model to predict whether the Over or Under on total points will happen
- Some important factors that I'm interested in are:
    - Date of the game, Home Team, Away Team, Total Score, Total Score Open (Betting Line)
    - Weather Conditions: Average Temperature, Average Humidity, Average Wind Speed, and existence of Precipitation (type),
- At a high level, and after some research, I specifically became interested in detailing how environmental factors affect the total points scored in a game

Data Set

- Located scores and betting data of all NFL games starting from 2006 regular season to present
    - Grouped into columns representing Home Team, Away Team, Home Score, Away Score, Total Score, Total Score Open (which is the set Over/Under line), and Line Differential

- The most important external factors effecting total score identified from preliminary analysis were the weather conditions for each game

- To account for this, added weather data for all the NFL games from 2006 to 2016
    - Added columns showing Average Temperature, Average Humidity, Average Wind, Precipitation, and Fog/Haze to the applicable years within the data set

Preliminary Analysis

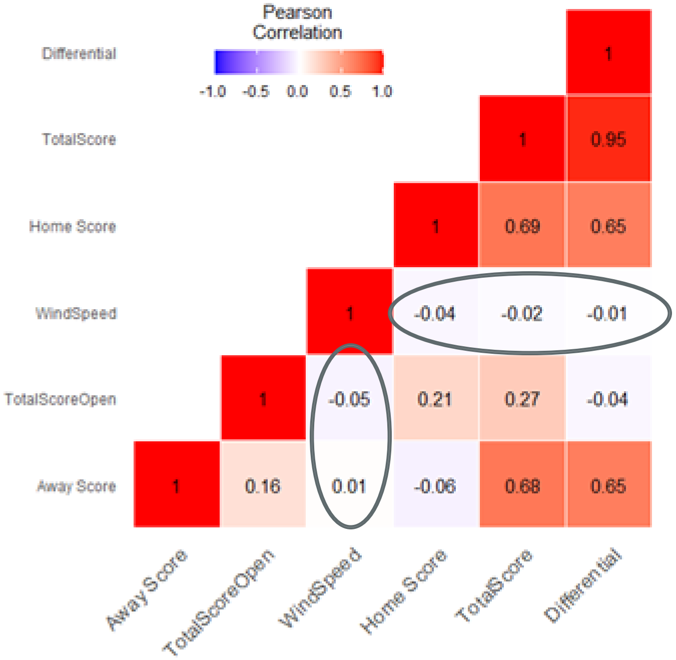
My preliminary analysis of historic over/under lines showed that taking the over is more appealing, yet when the over/under line is higher than the mean of all over/under bets, the under bet won with increasing frequency.

- Weather conditions that most influence the overall performance are temperature, humidity, wind speed, precipitation, and fog/haze

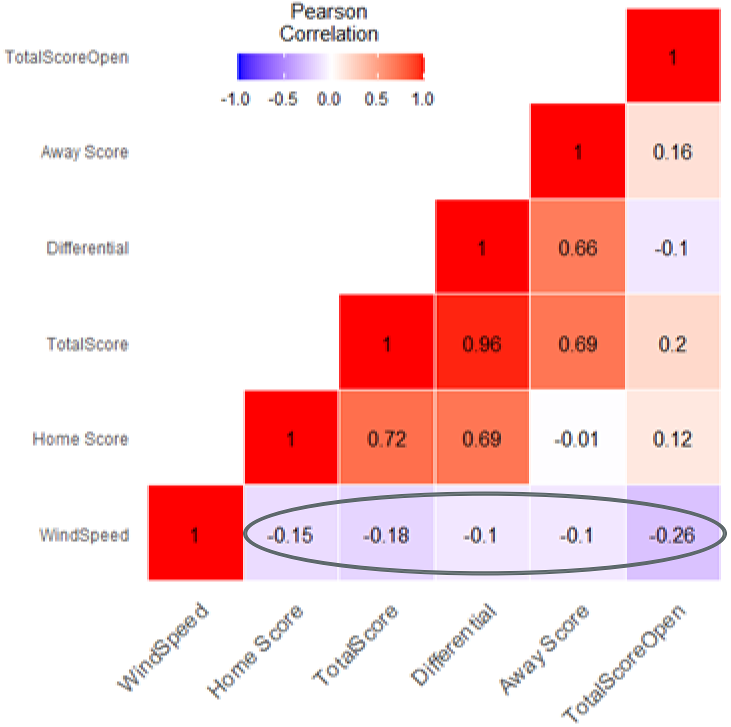
- If a game is played in a a closed dome stadium, all impacts of above weather conditions are rendered null, creating the optimal environment for teams to score more points

Heat Maps - Wind

Low/No Wind (Under 6 miles per hour - sample of 443 games):

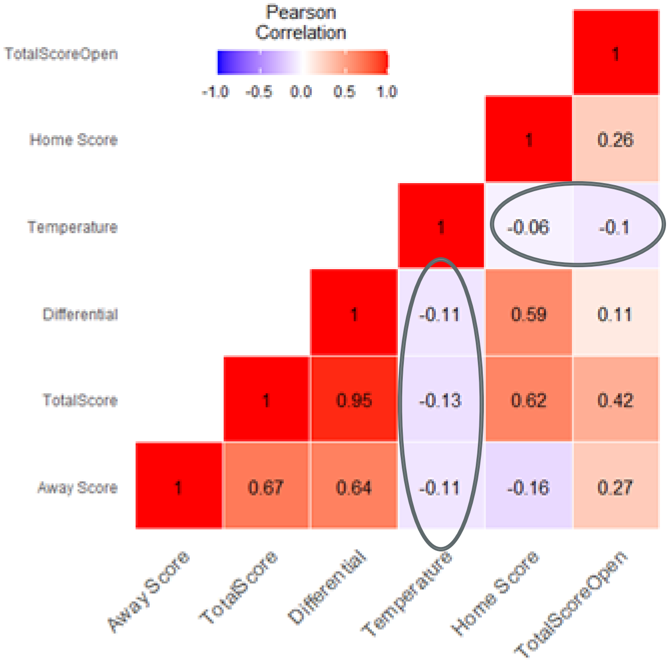


High Wind (Over 13 miles per hour - sample of 268 games):

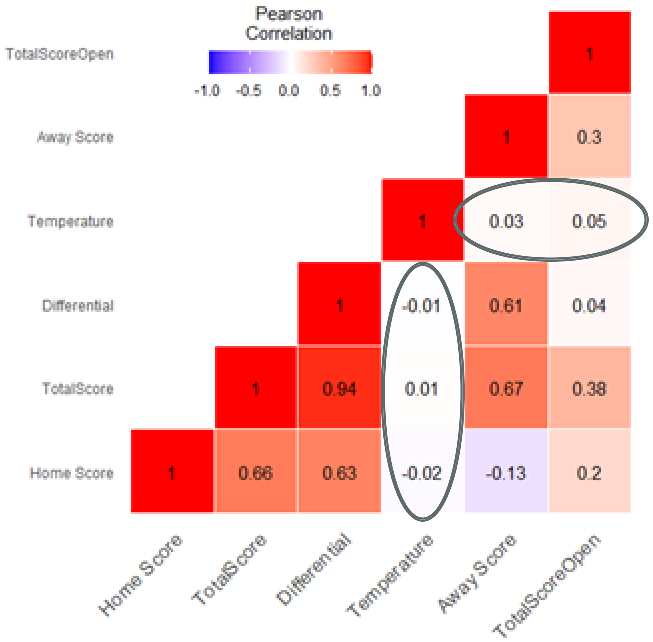


Heat Maps - Temperature

Very Hot Weather (Over 80°F - Sample of 153 games):



Very Cold Weather (Under 30°F - Sample of 89 Games):



Model Creation Processes

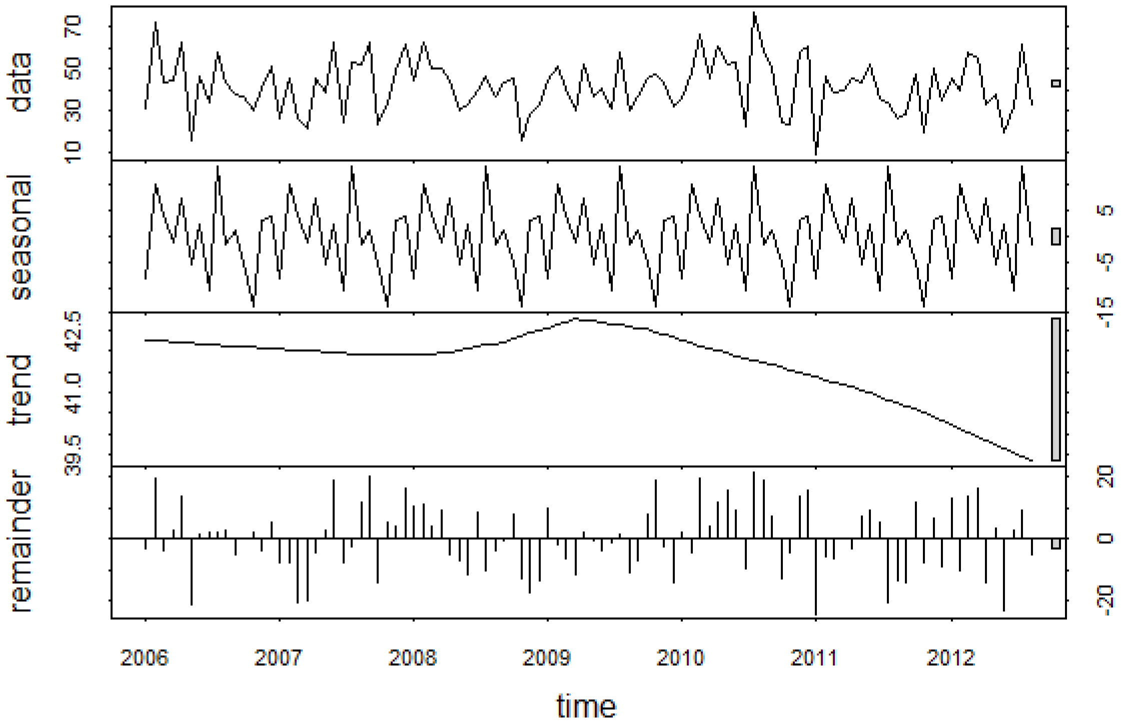
- I began with a decomposition of the response variable (TotalScore), followed by the creation and analysis of a Holt-Winters model and an ARIMA model

- Decided to use one team's data (the Seattle Seahawks) in order to build my preliminary models
    - Split into a training set and test set for each model

- Created a data frame containing Total Score, Average Wind Speed, Average Temperature, Average Humidity, and Precipitation

- Compared values forecasted by training set to actual values from test set to determine model accuracy

Decomposition of Total Score



Holt-Winters Output

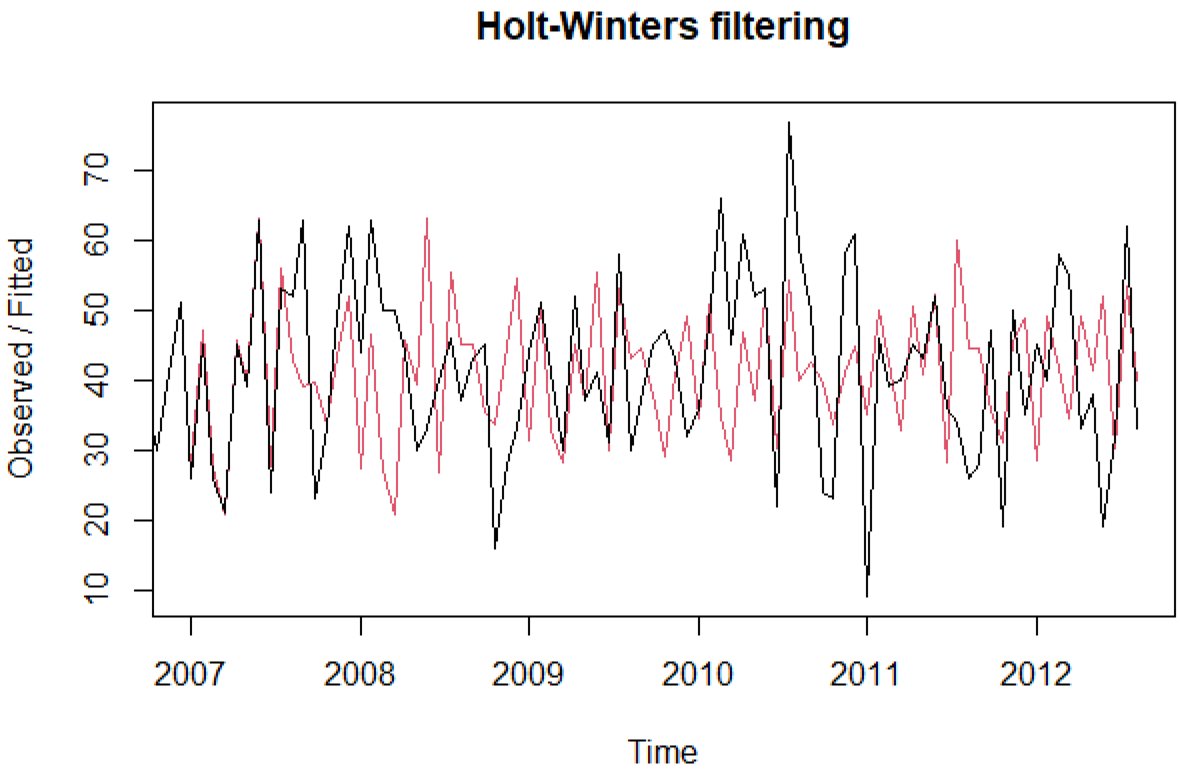
```
> scorecast
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = ts.total)

Smoothing parameters:
 alpha: 0
 beta : 0
 gamma: 0.2640427
```

- Based on the decomposition done in the previous slide, a pattern in the seasonality component can be observed
- Running the Holt-Winters function indicates that there is seasonality (as shown by the gamma)
- To account for seasonality in the model, the gamma parameter is set to .26

Holt-Winters Filtering on Total Score



Holt-Winters Forecast and Accuracy

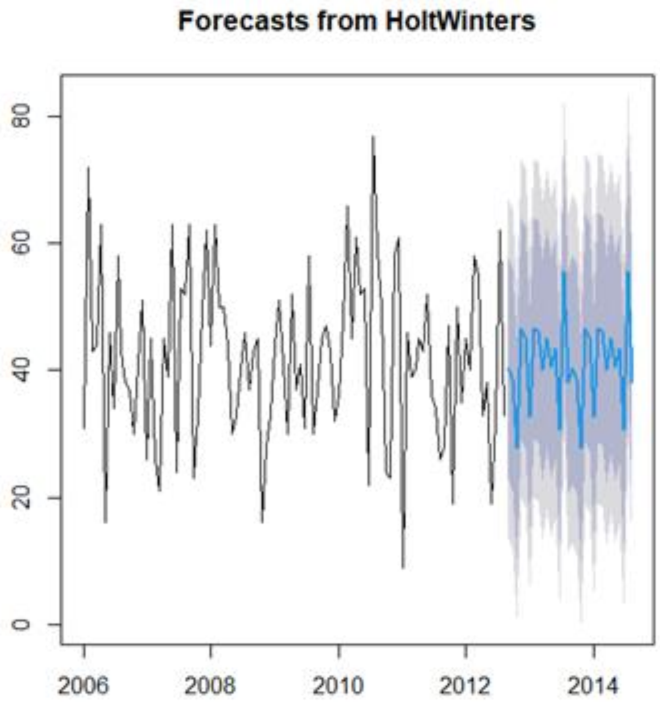
```
projectedScores <- scorecast2$mean

realLine <- data[101:130, 9]
projectedDifferential = projectedScores - realLine
realDifferential <- data[101:130, 11]

productDiff <- projectedDifferential * realDifferential
posProductDiff <- productDiff[which(productDiff >= 0)]

length(posProductDiff)/length(projectedScores)
```

= 12/30 = 0.4 = 40%



Box Test on Holt-Winters Forecast

```
> Box.test(scorecast2$residuals, lag = 20, type = "Ljung-Box")

        Box-Ljung test

data:  scorecast2$residuals

X-squared = 30.892, df = 20, p-value = 0.05664
```

- According to the box test, the p-value is .05664, which is greater than .05
- This indicates that the ACFs of the residuals are not equal to 0
- However, it is very close to .05 so we will attempt to improve the forecast

ARIMA Model

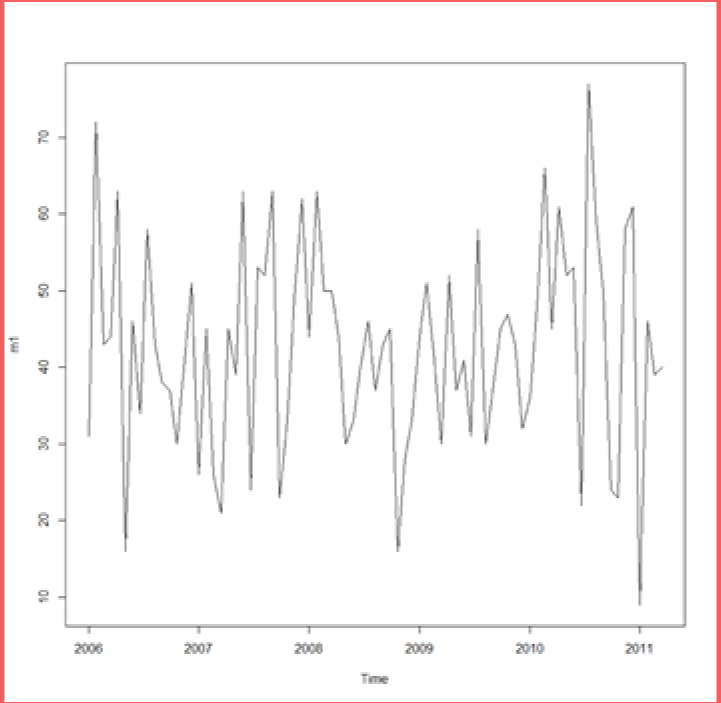
```
t1 <- train[,7]     #total score column
wind <- train[,9]
temp <- train[,10]
humidity <- train[,11]
t <- cbind(wind,temp,humidity)

m1 <- ts(t1, start = c(2006,38),
end = c(2011,41), frequency=15)

fitm2 <- auto.arima(m1, xreg = t)

f1 <- forecast(fitm2, xreg = t)
```

plot(m1)



ARIMA Model Output (cont.)

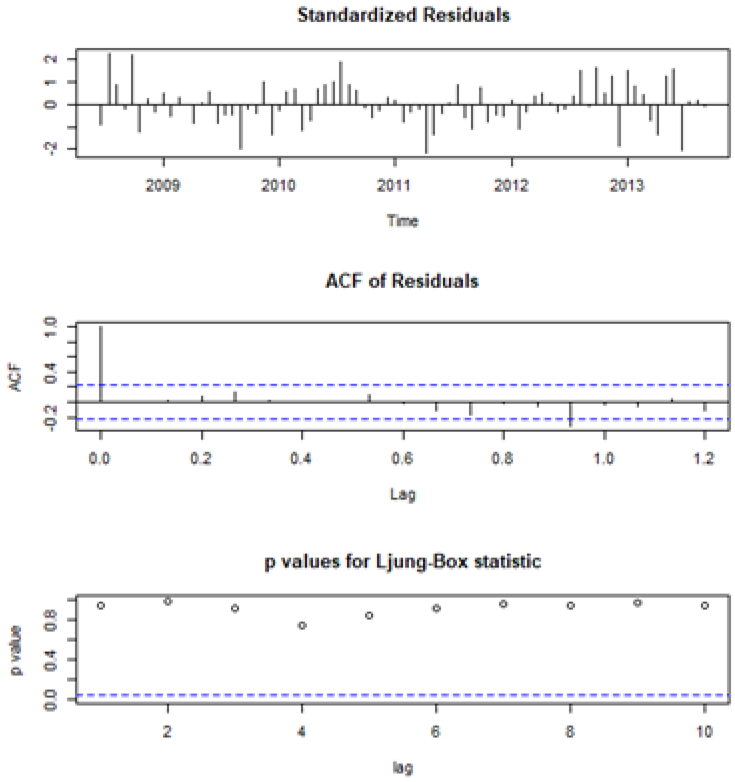
```
> fitm2
Series: m1
Regression with ARIMA(3,0,1) errors

Coefficients:
          ar1      ar2      ar3     ma1   intercept
      -0.8275  -0.1271  -0.1135  0.6140     72.2015
s.e.   0.2896   0.1604   0.1285  0.2751     10.9099

           wind     temp          humidity
        -0.9186  -0.2602  -0.1303
S.e      0.3307   0.1064   0.0952

sigma^2 estimated as 169.3:  log likelihood=-310.68
AIC=639.36   AICc=641.97   BIC=660.69
```

```
> tsdiag(fitm2)
```

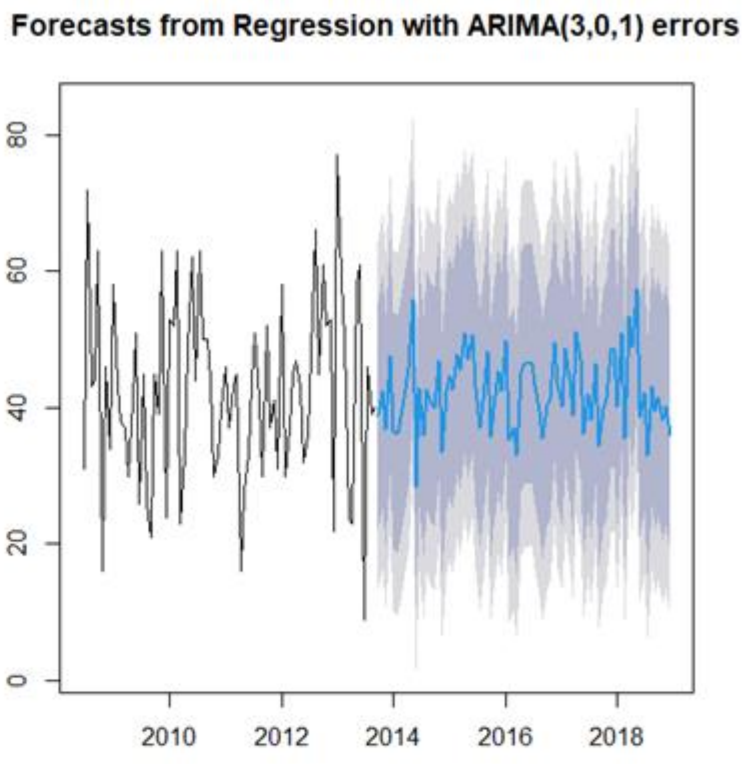


ARIMA Forecast

x = 1 at location where predicted values begin

x	y	x	y
1	38.81489	**11**	28.41210
2	42.39381	**12**	42.89350
3	36.78633	**13**	35.98835
4	47.60939	**14**	42.60429
5	36.73060	**15**	40.70436
6	36.22181	**16**	39.99184
7	38.56453	**17**	46.85284
8	42.72520	**18**	33.40350
9	45.74917	**19**	42.48318
10	55.92364	**20**	44.65914

```
> plot(f1)
```



ARIMA Forecast Accuracy

```
projectedScores <- f1$mean

realLine <- data[80:158, 9]
projectedDifferential = projectedScores - realLine
realDifferential <- data[80:158, 11]

productDiff <- projectedDifferential * realDifferential
posProductDiff <- productDiff[which(productDiff >= 0)]

length(posProductDiff)/length(projectedScores)
```

= 45/79 = 0.5696203 = 56.96%

Possible reasons for model inaccuracy:

- Roster differences over the course of multiple seasons
- Differing levels of opposing teams' defenses, both over the course of a season and over multiple seasons

Possible improvements to address model inaccuracy:

- Account for Average Points Scored by the Seahawks (for the season) and Average Points Allowed by opposing team (for the season)

Questions