Data Modeling

Contents

Overview

Analysis Methodology

A stratified analysis method will be implemented:

1. Test Overall Data
    - Iteratively create model step wise,
2. Split and Test Data by Voting Method
    - Can't stratify the data by each country due to a lack of data
    - Need a minimum of 10/20 observations per covariate for regression analysis
    - Split data by voting method, research televote shows more bias than the jury

Note:

- forward and step wise fitting will be utilized using AIC to determine model of best fit
- the models will be evaluated using the car package

Model Assumptions

Multiple Linear regression (MLR) requires the model residuals to be ~ IID N(0, sigma^2). The model residuals will be standardized for these assessments.

1. Normality Assumptions will be accessed using:
    - Normality tests from the nortest package
    - Visualizations such as histograms, QQ-plots, Residual Plots and Add Variable Plots
2. Constant Variance will be accessed using:

- non-constant variance test

3. Multi-collinearity will be accessed using:
    - variance inflation factors
4. Outliers will be accessed using:
    - Cooks Distance

Multiple Linear Regression Models

Overall

```
##
## Call:
## lm(formula = overall_final_model_form, data = processed_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5505 -2.3301 -0.2858  2.1846  7.8517
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.7244     0.6169   6.037 2.64e-09 ***
## Average_Points     0.4798     0.1253   3.830 0.000141 ***
## acousticness       0.6959     0.1302   5.344 1.26e-07 ***
## speechiness        0.6973     0.1362   5.119 4.05e-07 ***
## METRIC_Citizens    0.3251     0.1399   2.324 0.020438 *
## TC_PerfType_Solo   1.4412     0.5613   2.568 0.010457 *
## key_0              1.2923     0.4516   2.861 0.004353 **
## CAP_DIST_km        0.2956     0.1280   2.309 0.021260 *
## OOA                1.2837     0.4512   2.845 0.004579 **
## FC_NonCOB          0.3604     0.1391   2.592 0.009766 **
## ComSONGLAN         0.2760     0.1287   2.145 0.032338 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.074 on 647 degrees of freedom
## Multiple R-squared:  0.1762, Adjusted R-squared:  0.1635
## F-statistic: 13.84 on 10 and 647 DF,  p-value: < 2.2e-16
```

	vif(overall_final_model)
Average_Points	1.091139
acousticness	1.179141
speechiness	1.289927
METRIC_Citizens	1.360161
TC_PerfType_Solo	1.104542
key_0	1.246838
CAP_DIST_km	1.139690
OOA	1.188927
FC_NonCOB	1.344230
ComSONGLAN	1.150996

Televote

```
##
## Call:
## lm(formula = televote_final_model_form, data = televote_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3561 -1.9688 -0.0461  1.7443  6.7011
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.1314     0.3466  14.806  < 2e-16 ***
## METRIC_Citizens   0.5344     0.1555   3.436 0.000668 ***
## Average_Points    0.8126     0.1607   5.057 7.22e-07 ***
## TC_NumNeigh       0.7464     0.1742   4.286 2.42e-05 ***
## speechiness       0.5175     0.1656   3.125 0.001943 **
## acousticness      0.4804     0.1681   2.858 0.004550 **
## FC_NonCitzens     0.6452     0.1767   3.652 0.000304 ***
## VBlocs1_TC_13    -6.8165     2.1841  -3.121 0.001968 **
## OOA               0.8913     0.6028   1.479 0.140203
## CAP_DIST_km       0.3029     0.1726   1.755 0.080254 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.772 on 317 degrees of freedom
## Multiple R-squared:  0.3384, Adjusted R-squared:  0.3196
## F-statistic: 18.02 on 9 and 317 DF,  p-value: < 2.2e-16
```

	vif(televote_final_model)
METRIC_Citizens	1.440215
Average_Points	1.072688
TC_NumNeigh	1.440252
speechiness	1.231842
acousticness	1.059780
FC_NonCitzens	1.599009
VBlocs1_TC_13	1.233766
OOA	1.147115
CAP_DIST_km	1.291723

Jury

```
##
## Call:
## lm(formula = jury_final_model_form, data = jury_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.136 -2.494 -0.291  2.024  8.297
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.0865     0.4637   8.812  < 2e-16 ***
## CAP_DIST_km            0.6617     0.1854   3.568 0.000414 ***
## acousticness           0.5032     0.1747   2.880 0.004247 **
## speechiness            0.8932     0.2004   4.457 1.15e-05 ***
## TC_PerfType_Mixed     -9.6005     3.2765  -2.930 0.003632 **
## TC_LANGFAM_Armenian   -3.1767     0.9880  -3.215 0.001435 **
## VBlocs1_TC_1           3.0611     0.6177   4.956 1.17e-06 ***
## ComVBlocs1_y          -2.2750     0.6857  -3.318 0.001011 **
## VBlocs1_FC_1           0.8442     0.4283   1.971 0.049601 *
## VBlocs2_TC_1           1.5367     0.4794   3.205 0.001484 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.027 on 321 degrees of freedom
## Multiple R-squared:  0.2132, Adjusted R-squared:  0.1912
## F-statistic: 9.667 on 9 and 321 DF,  p-value: 4.405e-13
```

	vif(jury_final_model)
CAP_DIST_km	1.214251
acousticness	1.228278
speechiness	1.363699
TC_PerfType_Mixed	1.167945
TC_LANGFAM_Armenian	1.525300
VBlocs1_TC_1	3.277766
ComVBlocs1_y	2.648599
VBlocs1_FC_1	1.656485
VBlocs2_TC_1	2.070903