Soybean Data Grid Comparison

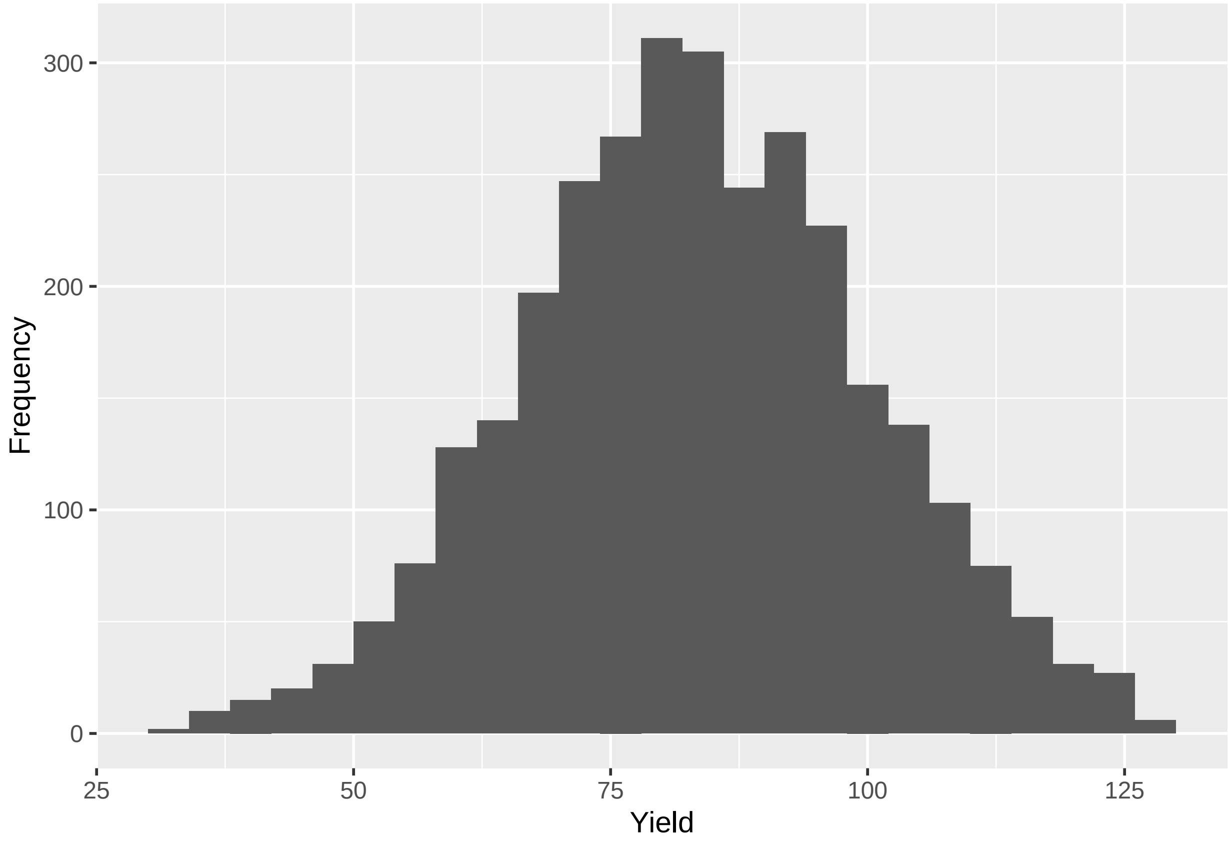
Dina Dinh

2023-12-18

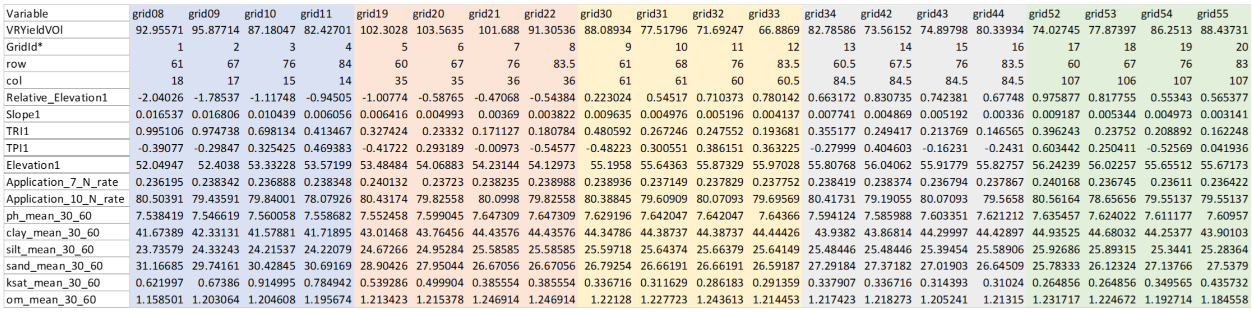
```r
library(easypackages)
libraries("tidyverse","boot","randomForest","psych","AUC","MASS","car",
          "viridis","caret","ggplot2", "corrplot", "gridExtra", "mlbench", "neuralnet",
          "rpart")
```

Using the dataset with most outliers removed in the response variable

```
##
##  Shapiro-Wilk normality test
##
## data:  df1$VRYieldVOl
## W = 0.99819, p-value = 0.001254
```



Discovering Any Patterns in the Dataset



Variable	grid08	grid09	grid10	grid11	grid19	grid20	grid21	grid22	grid30	grid31	grid32	grid33	grid34	grid42	grid43	grid44	grid52	grid53	grid54	grid55
VRYieldVOl	92.95571	95.87714	87.18047	82.42701	102.3028	103.5635	101.688	91.30536	88.08934	77.51796	71.69247	66.8869	82.78586	73.56152	74.89798	80.33934	74.02745	77.87397	86.2513	88.43731
GridId*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
row	61	67	76	84	60	67	76	83.5	61	68	76	83.5	60.5	67.5	76	83.5	60	67	76	83
col	18	17	15	14	35	35	36	36	61	61	60	60.5	84.5	84.5	84.5	84.5	107	106	107	107
Relative_Elevation1	-2.04026	-1.78537	-1.11748	-0.94505	-1.00774	-0.58765	-0.47068	-0.54384	0.223024	0.54517	0.710373	0.780142	0.663172	0.830735	0.742381	0.67748	0.975877	0.817755	0.55343	0.565377
Slope1	0.016537	0.016806	0.010439	0.006056	0.006416	0.004993	0.00369	0.003822	0.009635	0.004976	0.005196	0.004137	0.007741	0.004869	0.005192	0.00336	0.009187	0.005344	0.004973	0.003141
TRI1	0.995106	0.974738	0.698134	0.413467	0.327424	0.23332	0.171127	0.180784	0.480592	0.267246	0.247552	0.193681	0.355177	0.249417	0.213769	0.146565	0.396243	0.23752	0.208892	0.162248
TPI1	-0.39077	-0.29847	0.325425	0.469383	-0.41722	0.293189	-0.00973	-0.54577	-0.48223	0.300551	0.386151	0.363225	-0.27999	0.404603	-0.16231	-0.2431	0.603442	0.250411	-0.52569	0.041936
Elevation1	52.04947	52.4038	53.33228	53.57199	53.48484	54.06883	54.23144	54.12973	55.1958	55.64363	55.87329	55.97028	55.80768	56.04062	55.91779	55.82757	56.24239	56.02257	55.65512	55.67173
Application_7_N_rate	0.236195	0.238342	0.236888	0.238348	0.240132	0.23723	0.238235	0.238988	0.238936	0.237149	0.237829	0.237752	0.238419	0.238374	0.236794	0.237867	0.240168	0.236745	0.23611	0.236422
Application_10_N_rate	80.50391	79.43591	79.84001	78.07926	80.43174	79.82558	80.0998	79.82558	80.38845	79.60909	80.07093	79.69569	80.41731	79.19055	80.07093	79.5658	80.56164	78.65656	79.55137	79.55137
ph_mean_30_60	7.538419	7.546619	7.560058	7.558682	7.552458	7.599045	7.647309	7.647309	7.629196	7.642047	7.642047	7.64366	7.594124	7.585988	7.603351	7.621212	7.635457	7.624022	7.611177	7.60957
clay_mean_30_60	41.67389	42.33131	41.57881	41.71895	43.01468	43.76456	44.43576	44.43576	44.34786	44.38737	44.38737	44.44426	43.9382	43.86814	44.29997	44.42897	44.93525	44.68032	44.25377	43.90103
silt_mean_30_60	23.73579	24.33243	24.21537	24.22079	24.67266	24.95284	25.58585	25.58585	25.59718	25.64374	25.66379	25.64149	25.48446	25.48446	25.39454	25.58906	25.92686	25.89315	25.3441	25.28364
sand_mean_30_60	31.16685	29.74161	30.42845	30.69169	28.90426	27.95044	26.67056	26.67056	26.79254	26.66191	26.66191	26.59187	27.29184	27.37182	27.01903	26.64509	25.78333	26.12324	27.13766	27.5379
ksat_mean_30_60	0.621997	0.67386	0.914995	0.784942	0.539286	0.499904	0.385554	0.385554	0.336716	0.311629	0.286183	0.291359	0.337907	0.336716	0.314393	0.31024	0.264856	0.264856	0.349565	0.435732
om_mean_30_60	1.158501	1.203064	1.204608	1.195674	1.213423	1.215378	1.246914	1.246914	1.22128	1.227723	1.243613	1.214453	1.217423	1.218273	1.205241	1.21315	1.231717	1.224672	1.192714	1.184558

Figure 1: Medians of each variable for each grid color-coded by columns

From previous discoveries, GridId and col are the two most important variables for predictions in Random Forest followed by ksat and row. We can see that ksat around 0.5 had the highest yield as seen in grids 19 and 20. However, ksat under 0.30 had the lowest yields seen in grids 32 and 33. Grids with negative relative elevation are consistently better than grids with positive relative elevation with yields greater than 80 for those grids. This makes me believe that the soybeans in this dataset prefer lower elevation which causes water to pool in these areas. Maybe these soybeans require more water than given.

Since the medians of each variables are not very different among the grids, this makes me believe there's an unobserved underlying effect causing high yields in some grids. However, location is proven to be important for predicting crop yield.

Linear Regression (LR)

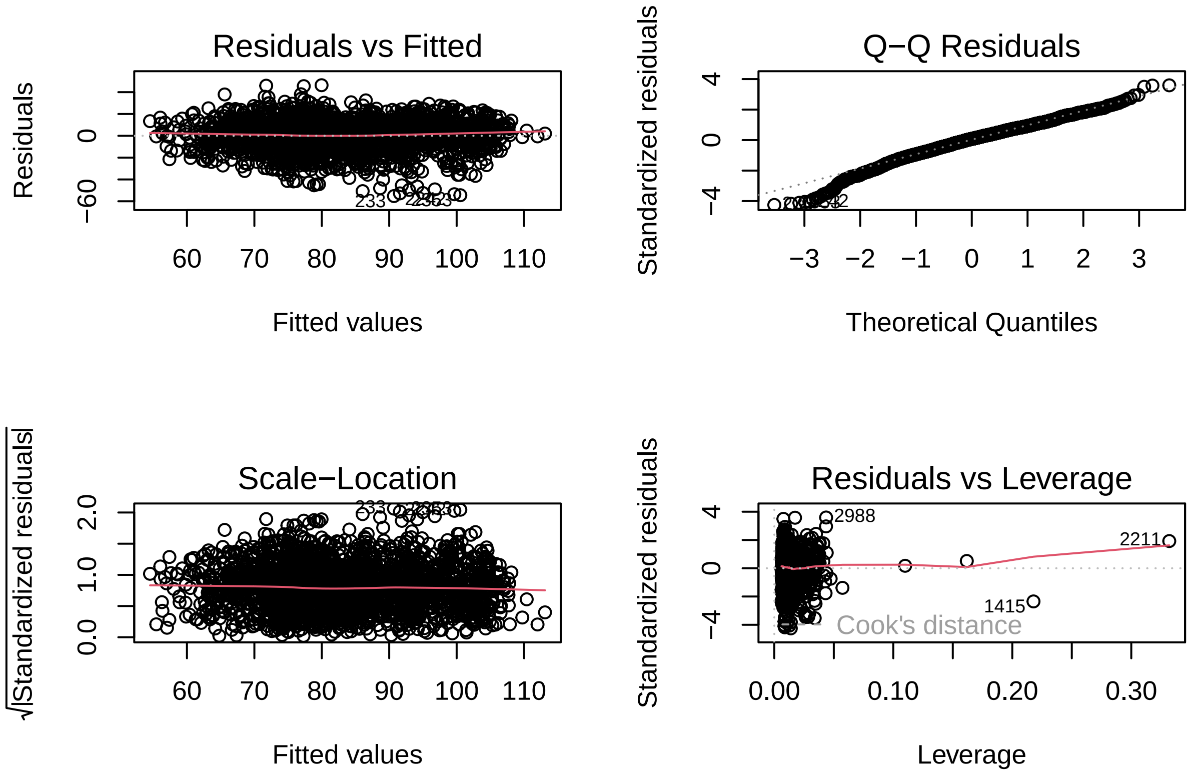
LR of Full Model

```
##
## Call:
## lm(formula = VRYieldVOl ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.097  -8.020   1.038   8.644  45.700
##
## Coefficients: (1 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2381.67879  473.67445   5.028 5.31e-07 ***
## GridId9                 14.23289    2.59696   5.481 4.67e-08 ***
## GridId10                11.06544    3.09674   3.573 0.000359 ***
## GridId11                12.36345    3.83097   3.227 0.001266 **
## GridId19                 8.61694    2.74501   3.139 0.001715 **
## GridId20                17.37024    2.84953   6.096 1.26e-09 ***
## GridId21                24.55229    3.42476   7.169 9.93e-13 ***
## GridId22                21.10444    4.01260   5.260 1.57e-07 ***
## GridId30                 4.41366    3.22657   1.368 0.171464
## GridId31                 2.20656    3.35696   0.657 0.511042
## GridId32                -1.20628    3.75907  -0.321 0.748315
## GridId33                -3.91406    4.35799  -0.898 0.369201
## GridId41                -8.48704    3.85551  -2.201 0.027809 *
## GridId42                  -7.44976    3.84874  -1.936 0.053026 .
## GridId43                  -2.48643    4.10389  -0.606 0.544656
## GridId44                  12.02965    4.64572   2.589 0.009671 **
## GridId52                  -3.74634    4.53074  -0.827 0.408390
## GridId53                   0.81225    4.45614   0.182 0.855380
## GridId54                  10.73250    4.66409   2.301 0.021469 *
## GridId55                  14.28988    5.16158   2.769 0.005673 **
## row                       -0.65123    0.12306  -5.292 1.32e-07 ***
## col                       -0.14574    0.05280  -2.760 0.005818 **
## Relative_Elevation1       -2.59971    1.37584  -1.890 0.058936 .
## Slope1                  -387.82554   64.47183  -6.015 2.06e-09 ***
## TRI1                      -1.35805    2.35262  -0.577 0.563823
## TPI1                      -0.31663    0.53417  -0.593 0.553398
## Elevation1                      NA         NA      NA       NA
## Application_7_N_rate       5.98244    7.44268   0.804 0.421589
## Application_10_N_rate      0.07066    0.03021   2.339 0.019420 *
## ph_mean_30_60           -198.60416   28.15784  -7.053 2.26e-12 ***
## clay_mean_30_60           -8.56196    5.38133  -1.591 0.111727
## silt_mean_30_60            7.87081    5.49827   1.432 0.152411
## sand_mean_30_60           -7.70101    5.05895  -1.522 0.128074
## ksat_mean_30_60          -21.02565    3.38157  -6.218 5.91e-10 ***
## om_mean_30_60           -282.95425   39.19038  -7.220 6.89e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.09 on 2467 degrees of freedom
## Multiple R-squared:  0.4122, Adjusted R-squared:  0.4044
## F-statistic: 52.43 on 33 and 2467 DF,  p-value: < 2.2e-16
```

The R^2 value of the full model is 0.4383011 and the correlation between the predicted and actual value of the test set is 0.663234 which is not a very high accuracy. (I ran the LR without removing the outliers in the response variable and the results were cor = 0.58, and $R^2 = 0.34$. The results were the same for the reduced model. Therefore, removing the outliers improved the model.)

LR of Reduced Model

```
## Start:  AIC=12896.92
## VRYieldVOl ~ GridId + row + col + Relative_Elevation1 + Slope1 +
##     TRI1 + TPI1 + Elevation1 + Application_7_N_rate + Application_10_N_rate +
##     ph_mean_30_60 + clay_mean_30_60 + silt_mean_30_60 + sand_mean_30_60 +
##     ksat_mean_30_60 + om_mean_30_60
##
##
## Step:  AIC=12896.92
## VRYieldVOl ~ GridId + row + col + Relative_Elevation1 + Slope1 +
##     TRI1 + TPI1 + Application_7_N_rate + Application_10_N_rate +
##     ph_mean_30_60 + clay_mean_30_60 + silt_mean_30_60 + sand_mean_30_60 +
##     ksat_mean_30_60 + om_mean_30_60
##
##                         Df Sum of Sq    RSS   AIC
## - TRI1                   1        57 422564 12895
## - TPI1                   1        60 422567 12895
## - Application_7_N_rate   1       111 422617 12896
## <none>                               422507 12897
## - silt_mean_30_60        1       351 422858 12897
## - sand_mean_30_60        1       397 422903 12897
## - clay_mean_30_60        1       434 422940 12898
## - Relative_Elevation1    1       611 423118 12898
## - Application_10_N_rate  1       937 423443 12900
## - col                    1      1305 423811 12903
## - row                    1      4796 427303 12923
## - Slope1                 1      6197 428704 12931
## - ksat_mean_30_60        1      6621 429128 12934
## - ph_mean_30_60          1      8520 431027 12945
## - om_mean_30_60          1      8928 431434 12947
## - GridId                19     86914 509420 13327
##
## Step:  AIC=12895.25
## VRYieldVOl ~ GridId + row + col + Relative_Elevation1 + Slope1 +
##     TPI1 + Application_7_N_rate + Application_10_N_rate + ph_mean_30_60 +
##     clay_mean_30_60 + silt_mean_30_60 + sand_mean_30_60 + ksat_mean_30_60 +
##     om_mean_30_60
##
##                         Df Sum of Sq    RSS   AIC
## - TPI1                   1        84 422648 12894
## - Application_7_N_rate   1       103 422667 12894
## <none>                               422564 12895
## - sand_mean_30_60        1       352 422916 12895
## - clay_mean_30_60        1       388 422952 12896
## - silt_mean_30_60        1       426 422990 12896
## - Relative_Elevation1    1       555 423119 12896
## + TRI1                   1        57 422507 12897
## - Application_10_N_rate  1       907 423471 12899
## - col                    1      1334 423898 12901
## - row                    1      4750 427313 12921
## - ksat_mean_30_60        1      6579 429143 12932
## - Slope1                 1      7704 430267 12938
## - ph_mean_30_60          1      8760 431323 12945
## - om_mean_30_60          1      9138 431701 12947
## - GridId                19     98758 521322 13382
##
## Step:  AIC=12893.75
## VRYieldVOl ~ GridId + row + col + Relative_Elevation1 + Slope1 +
##     Application_7_N_rate + Application_10_N_rate + ph_mean_30_60 +
##     clay_mean_30_60 + silt_mean_30_60 + sand_mean_30_60 + ksat_mean_30_60 +
##     om_mean_30_60
##
##                         Df Sum of Sq    RSS   AIC
## - Application_7_N_rate   1       102 422750 12892
## <none>                               422648 12894
## - sand_mean_30_60        1       362 423010 12894
## - clay_mean_30_60        1       398 423046 12894
## - silt_mean_30_60        1       455 423103 12894
## + TPI1                   1        84 422564 12895
## + TRI1                   1        81 422567 12895
## - Application_10_N_rate  1       894 423542 12897
## - col                    1      1265 423913 12899
## - Relative_Elevation1    1      1982 424630 12904
## - row                    1      4672 427320 12919
## - ksat_mean_30_60        1      6527 429175 12930
## - Slope1                 1      7784 430432 12937
## - ph_mean_30_60          1      8897 431545 12944
## - om_mean_30_60          1      9377 432025 12947
## - GridId                19    102511 525159 13399
##
## Step:  AIC=12892.35
## VRYieldVOl ~ GridId + row + col + Relative_Elevation1 + Slope1 +
##     Application_10_N_rate + ph_mean_30_60 + clay_mean_30_60 +
##     silt_mean_30_60 + sand_mean_30_60 + ksat_mean_30_60 + om_mean_30_60
##
##                         Df Sum of Sq    RSS   AIC
## <none>                                422750 12892
## - sand_mean_30_60        1       367 423117 12892
## - clay_mean_30_60        1       399 423149 12893
## - silt_mean_30_60        1       455 423205 12893
## + Application_7_N_rate   1       102 422648 12894
## + TPI1                   1        83 422667 12894
## + TRI1                   1        72 422677 12894
## - Application_10_N_rate  1       962 423711 12896
## - col                    1      1255 424004 12898
## - Relative_Elevation1    1      2055 424804 12902
## - row                    1      4674 427423 12918
## - ksat_mean_30_60        1      6430 429180 12928
## - Slope1                 1      7785 430535 12936
## - ph_mean_30_60          1      8954 431704 12943
## - om_mean_30_60          1      9444 432193 12946
## - GridId                19    102411 525161 13397
```



The R^2 value of the reduced model is 0.4380181, and the correlation between the predicted and actual value of the test set is 0.6631017. The results of the reduced model is almost the same as the full model.

In the diagnostic plots, we can check some assumptions for the linear regression such as homogeneous variance seen in the "Residuals vs Fitted" plot and normality of residuals seen in the Q-Q plot. We can see that the assumptions are met.