ML Final Project Extra Credit

Dutt Thakkar

2023-05-07

For Regression model, using different data subset for evaluating the regression results. Selecting those variables that acutally affects the analysis

#loading reqiured pacakges

```
library(caret)
```

```
## Loading required package: ggplot2
```

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

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.0     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ lubridate 1.9.2     ✔ tibble    3.1.8
## ✔ purrr     1.0.1     ✔ tidyr     1.3.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ purrr::lift()   masks caret::lift()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conf
licts to become errors
```

```
library(tidyr)
library(dplyr)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve
3WBa
```

```
library(cluster)
```

#importing dataset and viewing summary

```
Data_set <- read.csv("/Users/duttthakkar/Desktop/fuel_receipts_costs_eia923(1).csv")
summary(Data_set)
```

```
##      rowid          plant_id_eia   plant_id_eia_label report_date
##  Min.   :     1   Min.   :    3   Length:608564      Length:608564
##  1st Qu.:152142   1st Qu.: 2712   Class :character   Class :character
##  Median :304282   Median : 6155   Mode  :character   Mode  :character
##  Mean   :304282   Mean   :18290
##  3rd Qu.:456423   3rd Qu.:50707
##  Max.   :608564   Max.   :64020
##
##  contract_type_code contract_type_code_label contract_expiration_date
##  Length:608564      Length:608564            Length:608564
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
##  energy_source_code energy_source_code_label fuel_type_code_pudl
##  Length:608564      Length:608564            Length:608564
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
##  fuel_group_code     mine_id_pudl    mine_id_pudl_label supplier_name
##  Length:608564      Min.   :   0     Min.   :   0       Length:608564
##  Class :character   1st Qu.:  42     1st Qu.:  42       Class :character
##  Mode  :character   Median : 972     Median : 972       Mode  :character
##                     Mean   :1577     Mean   :1577
##                     3rd Qu.:3121     3rd Qu.:3121
##                     Max.   :4562     Max.   :4562
##                     NA's   :391946   NA's   :391946
##  fuel_received_units fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct
##  Min.   :       1    Min.   :   0.000    Min.   : 0.0000    Min.   : 0.000
##  1st Qu.:    3700    1st Qu.:   1.025    1st Qu.: 0.0000    1st Qu.: 0.000
##  Median :   21565    Median :   1.061    Median : 0.0000    Median : 0.000
##  Mean   :  242967    Mean   :   8.839    Mean   : 0.5145    Mean   : 3.606
##  3rd Qu.:  106164    3rd Qu.:  17.809    3rd Qu.: 0.4900    3rd Qu.: 5.800
##  Max.   :48159765    Max.   :1049.000    Max.   :11.0100    Max.   :72.200
##
##  mercury_content_ppm fuel_cost_per_mmbtu primary_transportation_mode_code
##  Min.   :0.00        Min.   :   -71.9    Length:608564
##  1st Qu.:0.00        1st Qu.:     2.3    Class :character
##  Median :0.00        Median :     3.3    Mode  :character
##  Mean   :0.01        Mean   :    14.2
##  3rd Qu.:0.00        3rd Qu.:     4.8
##  Max.   :1.82        Max.   :562572.2
##  NA's   :289482      NA's   :200240
##  primary_transportation_mode_code_label secondary_transportation_mode_code
##  Length:608564                          Length:608564
##  Class :character                       Class :character
```

```
##  Mode  :character                      Mode  :character
##
##
##
##
##  secondary_transportation_mode_code_label natural_gas_transport_code
##  Length:608564                             Length:608564
##  Class :character                          Class :character
##  Mode  :character                          Mode  :character
##
##
##
##
##  natural_gas_delivery_contract_type_code moisture_content_pct
##  Length:608564                           Min.   :  0.0
##  Class :character                        1st Qu.:  6.6
##  Mode  :character                        Median : 11.9
##                                          Mean   : 15.6
##                                          3rd Qu.: 26.8
##                                          Max.   :247.0
##                                          NA's   :516588
##  chlorine_content_ppm data_maturity      data_maturity_label
##  Min.   :   0.0       Length:608564      Length:608564
##  1st Qu.:   0.0       Class :character   Class :character
##  Median :   0.0       Mode  :character   Mode  :character
##  Mean   :  59.2
##  3rd Qu.:   0.0
##  Max.   :3747.0
##  NA's   :516588
```

#data cleaning

#selecting attributes

```
fuel_data<-Data_set[,c(11,16,17,18,20)]
summary(fuel_data)
```

```
##  fuel_group_code    fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct
##  Length:608564      Min.   :   0.000    Min.   : 0.0000    Min.   : 0.000
##  Class :character   1st Qu.:   1.025    1st Qu.: 0.0000    1st Qu.: 0.000
##  Mode  :character   Median :   1.061    Median : 0.0000    Median : 0.000
##                     Mean   :   8.839    Mean   : 0.5145    Mean   : 3.606
##                     3rd Qu.:  17.809    3rd Qu.: 0.4900    3rd Qu.: 5.800
##                     Max.   :1049.000    Max.   :11.0100    Max.   :72.200
##
##  fuel_cost_per_mmbtu
##  Min.   :   -71.9
##  1st Qu.:     2.3
##  Median :     3.3
##  Mean   :    14.2
##  3rd Qu.:     4.8
##  Max.   :562572.2
##  NA's   :200240
```

```
str(fuel_data)
```

```
## 'data.frame':    608564 obs. of  5 variables:
##  $ fuel_group_code    : chr  "coal" "coal" "natural_gas" "coal" ...
##  $ fuel_mmbtu_per_unit: num  23.1 22.8 1.04 24.61 24.45 ...
##  $ sulfur_content_pct : num  0.49 0.48 0 1.69 0.84 1.54 0 2.16 1.24 1.9 ...
##  $ ash_content_pct    : num  5.4 5.7 0 14.7 15.5 14.6 0 15.4 11.9 15.4 ...
##  $ fuel_cost_per_mmbtu: num  2.13 2.12 8.63 2.78 3.38 ...
```

#checking for Na's

```
colMeans(is.na(fuel_data))
```

```
##     fuel_group_code fuel_mmbtu_per_unit  sulfur_content_pct     ash_content_pct
##           0.0000000           0.0000000           0.0000000           0.0000000
## fuel_cost_per_mmbtu
##           0.3290369
```

#Data imputing

```
fuel_data$fuel_cost_per_mmbtu[is.na(fuel_data$fuel_cost_per_mmbtu)] <- mean(fuel_dat
a$fuel_cost_per_mmbtu, na.rm = TRUE)
colMeans(is.na(fuel_data))
```

```
##     fuel_group_code fuel_mmbtu_per_unit  sulfur_content_pct     ash_content_pct
##                   0                   0                   0                   0
## fuel_cost_per_mmbtu
##                   0
```

#all Na's have been imputed using the mean

#Data partition

```
library(caTools)
set.seed(2299)
# Sample about 2% of data
sample_size <- round(0.02 * nrow(fuel_data))
sample_indices <- sample(nrow(fuel_data), sample_size, replace = FALSE)

# Split sampled data into training and test sets
train_data <- fuel_data[sample_indices[1:round(0.75*sample_size)], ]
test_data <- fuel_data[sample_indices[(round(0.75*sample_size) + 1):sample_size], ]

nrow(train_data)
```

```
## [1] 9128
```

```
nrow(test_data)
```

```
## [1] 3043
```

#normalization of the data

```
cluster_data <- train_data %>% select( 'ash_content_pct', 'sulfur_content_pct','fuel_
mmbtu_per_unit','fuel_cost_per_mmbtu')

cluster_train <- preProcess(cluster_data, method = "range")
cluster_predict <- predict(cluster_train, cluster_data)

summary(cluster_predict)
```

```
##  ash_content_pct   sulfur_content_pct fuel_mmbtu_per_unit fuel_cost_per_mmbtu
##  Min.   :0.00000   Min.   :0.00000    Min.   :0.00000     Min.   :0.0000000
##  1st Qu.:0.00000   1st Qu.:0.00000    1st Qu.:0.03262     1st Qu.:0.0005627
##  Median :0.00000   Median :0.00000    Median :0.03377     Median :0.0009984
##  Mean   :0.05566   Mean   :0.04540    Mean   :0.28503     Mean   :0.0019052
##  3rd Qu.:0.09121   3rd Qu.:0.03815    3rd Qu.:0.58220     3rd Qu.:0.0029716
##  Max.   :1.00000   Max.   :1.00000    Max.   :1.00000     Max.   :1.0000000
```

#Applying K-means Algorithm

```
KMean_ <- kmeans(cluster_predict, centers = 2, nstart = 30)
```

#centers

```
KMean_$centers
```

```
##   ash_content_pct sulfur_content_pct fuel_mmbtu_per_unit fuel_cost_per_mmbtu
## 1    0.1549870761        0.122442140          0.69788018         0.001233876
## 2    0.0001137895        0.002311313          0.05413857         0.002280612
```

#The final cluster

```
fcluster<- KMean_$cluster
f_cluster<- cbind(train_data, fcluster)
f_cluster$fcluster<-as.factor(f_cluster$fcluster)
head(f_cluster)
```

```
##        fuel_group_code fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct
## 449867     natural_gas               1.078                  0               0
## 522907     natural_gas               1.011                  0               0
## 278462       other_gas               0.959                  0               0
## 81603      natural_gas               1.000                  0               0
## 557543     natural_gas               1.030                  0               0
## 487065     natural_gas               1.029                  0               0
##        fuel_cost_per_mmbtu fcluster
## 449867             2.96700        2
## 522907             2.26800        2
## 278462            14.18426        2
## 81603             14.18426        2
## 557543             1.84600        2
## 487065            14.18426        2
```

#We find the mean of all the quantitative variables

```
f_cluster%>%group_by(fcluster)%>%
  summarize(
    fuel_mmbtu_per_unit=mean(fuel_mmbtu_per_unit),
            fuel_cost_per_mmbtu=mean(fuel_cost_per_mmbtu),
            sulfur_content=mean(sulfur_content_pct),
    ash_content=mean(ash_content_pct))
```

```
## # A tibble: 2 × 5
##   fcluster fuel_mmbtu_per_unit fuel_cost_per_mmbtu sulfur_content ash_content
##   <fct>                  <dbl>               <dbl>          <dbl>       <dbl>
## 1 1                      21.3                 5.95         1.35       9.52
## 2 2                       1.68               10.9          0.0254     0.00699
```

#Use multiple-linear regression to determine the best set of variables to predict fuel_cost_per_mmbtu

#training data

```
reg_df<- f_cluster
fuel<-reg_df[,-c(1)]
fuel_ML<- preProcess(fuel, method = "range")
fuel_predict <- predict(fuel_ML, fuel)
head(fuel_predict)
```

```
##        fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct
## 449867          0.03436279                  0               0
## 522907          0.03216384                  0               0
## 278462          0.03045719                  0               0
## 81603           0.03180282                  0               0
## 557543          0.03278742                  0               0
## 487065          0.03275460                  0               0
##        fuel_cost_per_mmbtu fcluster
## 449867        0.0006055711        2
## 522907        0.0004581332        2
## 278462        0.0029715927        2
## 81603         0.0029715927        2
## 557543        0.0003691221        2
## 487065        0.0029715927        2
```

#performing multiple linear regression model on training data

```
k<-fuel_predict$fuel_cost_per_mmbtu
D1<- fuel_predict$fuel_mmbtu_per_unit
D2<- fuel_predict$sulfur_content_pct
D3<- fuel_predict$ash_content_pct
model_check <- lm(fuel_cost_per_mmbtu~.,data=fuel_predict)
summary(model_check)
```

```
##
## Call:
## lm(formula = fuel_cost_per_mmbtu ~ ., data = fuel_predict)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.00229 -0.00127 -0.00040  0.00076  0.99778
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.0014530  0.0010715  -1.356 0.175128
## fuel_mmbtu_per_unit  0.0032130  0.0015353   2.093 0.036401 *
## sulfur_content_pct  -0.0008769  0.0019624  -0.447 0.654993
## ash_content_pct      0.0035611  0.0017297   2.059 0.039537 *
## fcluster2            0.0035613  0.0010055   3.542 0.000399 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01098 on 9123 degrees of freedom
## Multiple R-squared:  0.002959,   Adjusted R-squared:  0.002522
## F-statistic: 6.769 on 4 and 9123 DF,  p-value: 1.952e-05
```

#Use the anova analysis

```
anova(model_check)
```

```
## Analysis of Variance Table
##
## Response: fuel_cost_per_mmbtu
##                       Df  Sum Sq    Mean Sq F value    Pr(>F)
## fuel_mmbtu_per_unit    1 0.00162 0.00161889 13.4367 0.0002481 ***
## sulfur_content_pct     1 0.00010 0.00009507  0.7891 0.3743898
## ash_content_pct        1 0.00004 0.00003679  0.3054 0.5805489
## fcluster               1 0.00151 0.00151142 12.5448 0.0003993 ***
## Residuals           9123 1.09916 0.00012048
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#Test data

```
Check_df<- test_data
fuel<-Check_df[,-c(1)]
fuel_chk<- preProcess(fuel, method = "range")
fuel_check <- predict(fuel_chk, fuel)
head(fuel_check)
```

```
##        fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct
## 520985          0.03130121         0.00000000      0.00000000
## 489215          0.20409323         0.04178273      0.00000000
## 355382          0.56552854         0.03899721      0.08732171
## 41081           0.83279270         0.13370474      0.15847861
## 289686          0.56689964         0.03760446      0.08240887
## 284956          0.03106712         0.00000000      0.00000000
##        fuel_cost_per_mmbtu
## 520985         0.002666383
## 489215         0.015325218
## 355382         0.015325218
## 41081          0.003910319
## 289686         0.001360307
## 284956         0.003790558
```

#performing multiple linear regression model on test data

```
M<-fuel_check$fuel_cost_per_mmbtu

T1<- fuel_predict$fuel_mmbtu_per_unit
T2<- fuel_predict$sulfur_content_pct
T3<- fuel_predict$ash_content_pct
model_check1 <- lm(fuel_cost_per_mmbtu~.,data=fuel_check)
summary(model_check1)
```

```
##
## Call:
## lm(formula = fuel_cost_per_mmbtu ~ ., data = fuel_check)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.01070 -0.00618 -0.00354  0.00462  0.98930
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.0109820  0.0005523  19.885  < 2e-16 ***
## fuel_mmbtu_per_unit -0.0090075  0.0019953  -4.514  6.6e-06 ***
## sulfur_content_pct   0.0052125  0.0041127   1.267    0.205
## ash_content_pct      0.0063958  0.0050882   1.257    0.209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02234 on 3039 degrees of freedom
## Multiple R-squared:  0.009014,   Adjusted R-squared:  0.008036
## F-statistic: 9.215 on 3 and 3039 DF,  p-value: 4.572e-06
```

#Use the anova analysis to predict the model

```
anova(model_check1)
```

```
## Analysis of Variance Table
##
## Response: fuel_cost_per_mmbtu
##                        Df  Sum Sq   Mean Sq F value    Pr(>F)
## fuel_mmbtu_per_unit     1 0.01198 0.0119751 23.9891 1.019e-06 ***
## sulfur_content_pct      1 0.00104 0.0010357  2.0748    0.1499
## ash_content_pct         1 0.00079 0.0007887  1.5800    0.2089
## Residuals            3039 1.51704 0.0004992
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion

#It appears that the predictors in your model have varying levels of significance. The intercept and fuel_mmbtu_per_unit have statistically significant coefficients, while sulfur_content_pct and ash_content_pct do not appear to have a statistically significant effect on fuel_cost_per_mmbtu. Additionally, the fcluster2 variable also has a statistically significant effect on fuel_cost_per_mmbtu. In the test data, only the fuel_mmbtu_per_unit predictor appears to be statistically significant, while sulfur_content_pct and ash_content_pct do not appear to have a significant effect on fuel_cost_per_mmbtu. Overall, it seems that fuel_mmbtu_per_unit is the most important predictor in your model for predicting fuel_cost_per_mmbtu, with the other predictors having limited impact.