

Data Analytics

Assignment -3

Name : R Siva Girish

SRN : PES1201700159

Dataset :White Wine Quality Prediction

Dataset link :

https://www.kaggle.com/sgus1318/winedata#winequality_white.csv

❖ About the Dataset

➤ Dataset contains 4898 rows and 12 columns

```
> head(wine)
fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol quality
1      7.0         0.27      0.36      20.7      0.045          45          170 1.0010 3.00      0.45      8.8      6
2      6.3         0.30      0.34       1.6      0.049          14          132 0.9940 3.30      0.49      9.5      6
3      8.1         0.28      0.40       6.9      0.050          30           97 0.9951 3.26      0.44     10.1      6
4      7.2         0.23      0.32       8.5      0.058          47          186 0.9956 3.19      0.40      9.9      6
5      7.2         0.23      0.32       8.5      0.058          47          186 0.9956 3.19      0.40      9.9      6
6      8.1         0.28      0.40       6.9      0.050          30           97 0.9951 3.26      0.44     10.1      6
>
```

❖ Goal

➤ Use different models to predict the data accurately.

➤ Models I have used on the white wine dataset are-:

- Linear Regression model(30% R^2)
- Multivariate Regression(86% and 20% R^2)
- Non Linear regression(27% R^2)
- Binomial Logistic Regression(High R^2)

❖ Linear Regression

- Statistical approach to model the relationship between one scalar response and a couple of explanatory variables.
- In our case we try to predict the quality of white wine using other parameters such as acidity, density, pH, alcohol content etc.
- Code

```
■ LinearModel<-lm(wine$trans.quality~.,data=
wine)
summary(LinearModel)
```

```
> summary(LinearModel)

Call:
lm(formula = wine$trans.quality ~ ., data = wine)

Residuals:
    Min       1Q   Median       3Q      Max
-3.8288 -0.4663 -0.0292  0.4744  3.0806

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.85762    0.01213  482.864 < 2e-16 ***
trans.fixed.acidity  0.06491    0.01875   3.462 0.000543 ***
trans.volatile.acidity -0.15785    0.01351 -11.687 < 2e-16 ***
trans.citric.acid    0.05696    0.01540   3.699 0.000220 ***
trans.residual.sugar  0.34003    0.02950  11.526 < 2e-16 ***
trans.chlorides     -0.07147    0.01712  -4.175 3.05e-05 ***
trans.free.sulfur.dioxide 0.13932    0.01667   8.359 < 2e-16 ***
trans.total.sulfur.dioxide -0.05098    0.01883  -2.707 0.006821 **
trans.density       -0.45035    0.05028  -8.957 < 2e-16 ***
trans.pH            0.12271    0.01646   7.457 1.10e-13 ***
trans.sulphates      0.08239    0.01294   6.368 2.15e-10 ***
trans.alcohol       0.16674    0.03148   5.297 1.25e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7394 on 3738 degrees of freedom
Multiple R-squared:  0.3052,    Adjusted R-squared:  0.3032
F-statistic: 149.3 on 11 and 3738 DF,  p-value: < 2.2e-16
```

■ R Squared Value is very Less = ~30%.

- Therefore based on P score values we successively remove trans.fixed.acidity, trans.citric.acid,trans.denstiy.

Code:

- `LinearModel<-lm(wine$trans.quality~.-trans.density- trans.fixed.acidity - trans.citric.acid,data=wine)`
`summary(LinearModel)`

```
> summary(LinearModel)
```

```
Call:
```

```
lm(formula = wine$trans.quality ~ . - trans.density - trans.fixed.acidity - trans.citric.acid, data = wine)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-3.8106 -0.4630 -0.0319  0.4760  3.0287
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.85756	0.01223	479.043	< 2e-16	***
trans.volatile.acidity	-0.15563	0.01332	-11.688	< 2e-16	***
trans.residual.sugar	0.11026	0.01432	7.698	1.76e-14	***
trans.chlorides	-0.09142	0.01718	-5.321	1.09e-07	***
trans.free.sulfur.dioxide	0.15754	0.01659	9.496	< 2e-16	***
trans.total.sulfur.dioxide	-0.07330	0.01861	-3.939	8.33e-05	***
trans.pH	0.05837	0.01302	4.483	7.57e-06	***
trans.sulphates	0.06150	0.01280	4.802	1.63e-06	***
trans.alcohol	0.41095	0.01620	25.360	< 2e-16	***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.7484 on 3741 degrees of freedom
```

```
Multiple R-squared:  0.2878,    Adjusted R-squared:  0.2862
```

```
F-statistic: 188.9 on 8 and 3741 DF,  p-value: < 2.2e-16
```

Conclusion:

Therefore we can conclude that Linear regression is not a very good model .

❖ Multivariate Regression

Code For creating model :

#Remove free.Sulphur.dioxide

```
mymodel<-lm(cbind(wine$alcohol,wine$quality)~wine$fixed
.acidity+wine$volatile.acidity+wine$citric.acid+wine$residua
l.sugar+wine$chlorides+wine$total.sulfur.dioxide+wine$den
sity+wine$pH+wine$sulphates,data=wine)
```

```
summary(mymodel)
```

```
coef(mymodel)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-3.4205 -0.2559 -0.0325  0.2177 16.1715

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.855e+02  5.032e+00 136.221 < 2e-16 ***
wine$fixed.acidity  5.263e-01  9.838e-03  53.493 < 2e-16 ***
wine$volatile.acidity  9.480e-01  6.491e-02 14.606 < 2e-16 ***
wine$citric.acid  3.613e-01  5.656e-02  6.388 1.84e-10 ***
wine$residual.sugar  2.396e-01  2.753e-03  87.013 < 2e-16 ***
wine$chlorides -2.706e-01  3.242e-01  -0.835  0.404
wine$total.sulfur.dioxide -2.749e-04  1.806e-04  -1.522  0.128
wine$density -6.930e+02  5.177e+00 -133.869 < 2e-16 ***
wine$pH  2.463e+00  5.157e-02  47.753 < 2e-16 ***
wine$sulphates  1.028e+00  5.773e-02  17.808 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4459 on 4888 degrees of freedom
Multiple R-squared:  0.869,    Adjusted R-squared:  0.8687
F-statistic: 3601 on 9 and 4888 DF,  p-value: < 2.2e-16

Response wine$quality :

Call:
lm(formula = `wine$quality` ~ wine$fixed.acidity + wine$volatile.acidity +
  wine$citric.acid + wine$residual.sugar + wine$chlorides +
  wine$total.sulfur.dioxide + wine$density + wine$pH + wine$sulphates,
  data = wine)

Residuals:
    Min       1Q   Median       3Q      Max
-3.5303 -0.4933 -0.0501  0.4725  6.0060

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.886e+02  8.545e+00  33.773 <2e-16 ***
wine$fixed.acidity  1.617e-01  1.670e-02  9.679 <2e-16 ***
wine$volatile.acidity -1.785e+00  1.102e-01 -16.197 <2e-16 ***
wine$citric.acid  1.036e-01  9.604e-02  1.079  0.2806
wine$residual.sugar  1.312e-01  4.675e-03  28.071 <2e-16 ***
wine$chlorides -2.260e-01  5.504e-01  -0.410  0.6815
wine$total.sulfur.dioxide  6.557e-04  3.066e-04  2.138  0.0325 *
wine$density -2.901e+02  8.789e+00 -33.006 <2e-16 ***
wine$pH  1.161e+00  8.757e-02  13.257 <2e-16 ***
wine$sulphates  8.241e-01  9.802e-02  8.407 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7571 on 4888 degrees of freedom
Multiple R-squared:  0.2705,    Adjusted R-squared:  0.2692
F-statistic: 201.4 on 9 and 4888 DF,  p-value: < 2.2e-16
```


- ❖ Based on the above stats observed above we notice that eliminating the variable from our model based on significance of pillai score is not giving us proper results. Hence a multivariate regression model in this case is unable to solve our problem. We must use a different model.

❖ Non linear Regression

- Polynomial regression model
- Eliminating attributes based on high p squared values we reach a set of attributes.
- Upon increasing the power of a few attributes and making the model non linear we don't notice any significant or useful change in the R^2 values.
- Upon increasing the polynomial power to four an increase in the adjusted R^2 is observed. But this is also not a very good method as it could lead to overfitting.
- Even upon increasing the power till four the R^2 value we get is around 27%.
- Hence there is no point in increasing the power any further.
- As even if we increase the power anymore it would result in overfitting and will not lead to any useful model.
- Hence the non linear model is not a very useful fit for this model.

➤ Hence we must resort to other models such as Logistic regression models to accurately fit this model.

```
call:
lm(formula = wine$trans.quality ~ poly(wine$trans.volatile.acidity,
  4) + poly(wine$trans.citric.acid, 4) + poly(wine$trans.residual.sugar,
  4) + poly(wine$trans.chlorides, 4) + poly(wine$trans.free.sulfur.dioxide,
  4) + poly(wine$trans.total.sulfur.dioxide, 4) + poly(wine$trans.pH,
  4) + poly(wine$trans.sulphates, 4), data = wine)

Residuals:
    Min       1Q   Median       3Q      Max
-2.94664 -0.50506 -0.02586  0.49024  2.94680

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   5.86507     0.01236  474.400 < 2e-16 ***
poly(wine$trans.volatile.acidity, 4)1 -3.47188     0.86311   -4.023 5.87e-05 ***
poly(wine$trans.volatile.acidity, 4)2  1.31861     0.79758    1.653 0.098361 .
poly(wine$trans.volatile.acidity, 4)3 -1.30907     0.76661   -1.708 0.087790 .
poly(wine$trans.volatile.acidity, 4)4 -2.18509     0.76354   -2.862 0.004236 **
poly(wine$trans.citric.acid, 4)1       3.64288     0.80573    4.521 6.34e-06 ***
poly(wine$trans.citric.acid, 4)2      -7.29084     0.81784   -8.915 < 2e-16 ***
poly(wine$trans.citric.acid, 4)3       0.94625     0.77707    1.218 0.223411
poly(wine$trans.citric.acid, 4)4       2.52848     0.77011    3.283 0.001036 **
poly(wine$trans.residual.sugar, 4)1    1.10940     0.87667    1.265 0.205781
poly(wine$trans.residual.sugar, 4)2    -5.89452     0.79311   -7.432 1.32e-13 ***
poly(wine$trans.residual.sugar, 4)3     3.12497     0.77985    4.007 6.27e-05 ***
poly(wine$trans.residual.sugar, 4)4    -0.19445     0.77036   -0.252 0.800732
poly(wine$trans.chlorides, 4)1        -12.86790     0.83787  -15.358 < 2e-16 ***
poly(wine$trans.chlorides, 4)2         0.48204     0.76918    0.627 0.530897
poly(wine$trans.chlorides, 4)3         3.25935     0.77905    4.184 2.93e-05 ***
poly(wine$trans.chlorides, 4)4         0.88829     0.76948    1.154 0.248410
poly(wine$trans.free.sulfur.dioxide, 4)1 12.08445     1.04751   11.536 < 2e-16 ***
poly(wine$trans.free.sulfur.dioxide, 4)2 -9.71179     1.11825   -8.685 < 2e-16 ***
poly(wine$trans.free.sulfur.dioxide, 4)3  1.42079     1.30516    1.089 0.276403
poly(wine$trans.free.sulfur.dioxide, 4)4 -0.50276     0.90724   -0.554 0.579499
poly(wine$trans.total.sulfur.dioxide, 4)1 -11.29508     1.14189   -9.892 < 2e-16 ***
poly(wine$trans.total.sulfur.dioxide, 4)2 -3.52482     0.97447   -3.617 0.000302 ***
poly(wine$trans.total.sulfur.dioxide, 4)3  2.09614     1.22440    1.712 0.086985 .
poly(wine$trans.total.sulfur.dioxide, 4)4  1.00317     1.17234    0.856 0.392220
poly(wine$trans.pH, 4)1               4.27295     0.80629    5.299 1.23e-07 ***
poly(wine$trans.pH, 4)2                0.02985     0.77119    0.039 0.969126
poly(wine$trans.pH, 4)3               -0.70806     0.76377   -0.927 0.353955
poly(wine$trans.pH, 4)4              -3.69995     0.76471   -4.838 1.36e-06 ***
poly(wine$trans.sulphates, 4)1         2.46949     0.79688    3.099 0.001957 **
poly(wine$trans.sulphates, 4)2         3.15968     0.77786    4.062 4.97e-05 ***
poly(wine$trans.sulphates, 4)3        -0.77573     0.76770   -1.010 0.312339
poly(wine$trans.sulphates, 4)4         0.88368     0.76412    1.156 0.247562
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7571 on 3717 degrees of freedom
Multiple R-squared:  0.2758,    Adjusted R-squared:  0.2695
F-statistic: 44.23 on 32 and 3717 DF, p-value: < 2.2e-16
```

❖ Binomial Logistic Regression

- Logistic Regression is a classification model used for predicting categorical variables.
- Due to the absence of categorical variables we will convert the quality variable into a categorical variable by assigning levels as good and bad.
- If quality level is less than 5 then we classify it as bad wine otherwise good fine.

Code:

```
LogisticModel <- glm(Lwine$Category ~ ., data = Lwine,  
family=binomial(link = "logit"))
```

```
summary(LogisticModel)
```

```
LogisticModelStepwise <- step(LogisticModel)
```

```
> LogisticModel <- glm(Lwine$Category ~ ., data = Lwine, family=binomial(link = "logit"))  
warning messages:  
1: glm.fit: algorithm did not converge  
2: glm.fit: fitted probabilities numerically 0 or 1 occurred  
> summary(LogisticModel)  
  
Call:  
glm(formula = Lwine$Category ~ ., family = binomial(link = "logit"),  
    data = Lwine)  
  
Deviance Residuals:  
    Min       1Q   Median       3Q      Max   
-6.333e-06 -4.409e-06  2.110e-08  4.665e-06  6.252e-06   
  
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)   
(Intercept)   -2.766e+02  4.182e+04  -0.007    0.995   
trans.fixed.acidity  -5.427e-02  5.156e+03   0.000    1.000   
trans.volatile.acidity -1.522e-01  4.011e+03   0.000    1.000   
trans.citric.acid    -3.485e-02  4.187e+03   0.000    1.000   
trans.residual.sugar -5.956e-02  8.354e+03   0.000    1.000   
trans.chlorides       7.634e-03  4.791e+03   0.000    1.000   
trans.free.sulfur.dioxide -3.798e-03  4.829e+03   0.000    1.000   
trans.total.sulfur.dioxide 1.965e-03  5.254e+03   0.000    1.000   
trans.density        1.701e-01  1.376e+04   0.000    1.000   
trans.pH             -5.210e-02  4.679e+03   0.000    1.000   
trans.sulphates       1.745e-02  3.891e+03   0.000    1.000   
trans.alcohol         1.924e-01  8.486e+03   0.000    1.000   
trans.quality         5.029e+01  7.393e+03   0.007    0.995   
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 4.7876e+03 on 3749 degrees of freedom  
Residual deviance: 6.3967e-08 on 3737 degrees of freedom  
AIC: 26  
  
Number of Fisher Scoring iterations: 25
```

Upon calculating the adjusted R^2 for this model we notice that it is ~ 1 by Mcfadden's Pseudo R^2 method which is very high. Hence this model is the best.

P value also when calculated is 0.

Hence The Logistic regression model is the best fit model.