Week 6:-

Question 2:-

1. Start the data profiling:-

############## Assignment 2 ################

############################################

########### Read and validate data #########

############################################

data <- read.table("https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/52b89051b9bd37a91a54e8570b963719/asset-v1:GTx+ISYE6501x+2T2017+type@asset+block/breast-cancer-wisconsin.data.txt",sep=",", header=FALSE)

head(data)

summary(data)

> summary(data)

V1 V2 V3 V4

Min. : 61634 Min. : 1.000 Min. : 1.000 Min. : 1.000

1st Qu.: 870688 1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 1.000

Median : 1171710 Median : 4.000 Median : 1.000 Median : 1.000

Mean : 1071704 Mean : 4.418 Mean : 3.134 Mean : 3.207

3rd Qu.: 1238298 3rd Qu.: 6.000 3rd Qu.: 5.000 3rd Qu.: 5.000

Max. :13454352 Max. :10.000 Max. :10.000 Max. :10.000

V5 V6 V7 V8

Min. : 1.000 Min. : 1.000 1 :402 Min. : 1.000

1st Qu.: 1.000 1st Qu.: 2.000 10 :132 1st Qu.: 2.000

Median : 1.000 Median : 2.000 2 : 30 Median : 3.000

Mean : 2.807 Mean : 3.216 5 : 30 Mean : 3.438

3rd Qu.: 4.000 3rd Qu.: 4.000 3 : 28 3rd Qu.: 5.000

Max. :10.000 Max. :10.000 8 : 21 Max. :10.000

**(Other): 56**

V9 V10 V11

Min. : 1.000 Min. : 1.000 Min. :2.00

1st Qu.: 1.000 1st Qu.: 1.000 1st Qu.:2.00

Median : 1.000 Median : 1.000 Median :2.00

Mean : 2.867 Mean : 1.589 Mean :2.69

3rd Qu.: 4.000 3rd Qu.: 1.000 3rd Qu.:4.00

Max. :10.000 Max. :10.000 Max. :4.00

**Key Observation: -** Column V7 has “other :56”. Have to research that.

1. **Validate V7.**

> levels(data$V7)

[1] "?" "1" "10" "2" "3" "4" "5" "6" "7" "8" "9"

table(data$V7)

? 1 10 2 3 4 5 6 7 8 9

16 402 132 30 28 19 30 4 8 21 9

**Key Observation: -** Column V7 has “16 of ?”. Have to research that.About 2%. Since it less than 5% it good for Data imputation

1. **Go with Mode(as the data is categorical) or Data imputation**
   1. Mode:- By looking data the mode is 1.

Plug in 1 in V7 with value 1 (mode)

data$V7[data$V7=='?'] <- 1

> table(data$V7)

? 1 10 2 3 4 5 6 7 8 9

0 418 132 30 28 19 30 4 8 21 9

> typeof(data$V7)

[1] "integer"

1. Go with Linear regression

Call:

lm(formula = V7 ~ V2 + V3 + V4 + V5 + V6 + V8 + V9 + V10, data = data\_modified)

Residuals:

Min 1Q Median 3Q Max

-9.7316 -0.9426 -0.3002 0.6725 8.6998

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.616652 0.194975 -3.163 0.00163 \*\*

V2 0.230156 0.041691 5.521 4.83e-08 \*\*\*

V3 -0.067980 0.076170 -0.892 0.37246

V4 0.340442 0.073420 4.637 4.25e-06 \*\*\*

V5 0.339705 0.045919 7.398 4.13e-13 \*\*\*

V6 0.090392 0.062541 1.445 0.14883

V8 0.320577 0.059047 5.429 7.91e-08 \*\*\*

V9 0.007293 0.044486 0.164 0.86983

V10 -0.075230 0.059331 -1.268 0.20524

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.274 on 674 degrees of freedom

Multiple R-squared: 0.615, Adjusted R-squared: 0.6104

F-statistic: 134.6 on 8 and 674 DF, p-value: < 2.2e-16

1. Re run the model by picking low Factors that have low P Value

Call:

lm(formula = V7 ~ V2 + V4 + V5 + V8, data = data\_modified)

Residuals:

Min 1Q Median 3Q Max

-9.8115 -0.9531 -0.3111 0.6678 8.6889

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.53601 0.17514 -3.060 0.0023 \*\*

V2 0.22617 0.04121 5.488 5.75e-08 \*\*\*

V4 0.31729 0.05086 6.239 7.76e-10 \*\*\*

V5 0.33227 0.04431 7.499 2.03e-13 \*\*\*

V8 0.32378 0.05606 5.775 1.17e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.274 on 678 degrees of freedom

Multiple R-squared: 0.6129, Adjusted R-squared: 0.6107

F-statistic: 268.4 on 4 and 678 DF, p-value: < 2.2e-16

1. After running regression here is the table of V7. Looks much better than mode where we default to 1.

table(data\_reg\_imp$V7)

2 3 4 5 6 7 8 9 10 11

402 132 30 28 19 30 4 8 21 9

1. Lets randomize our addition to do perturbation.

V7p = round(runif(16, 2, 11))

dataper[missing,]$V7 = V7p

1. Run KNN:-
   1. For Values removed:-
      1. KNN list 1 thro 20

[1] 0.5241581 0.5300146 0.5139092 0.5021962 0.4890190 0.4729136 0.4641288

[8] 0.4582723 0.4553441 0.4597365 0.4597365 0.4626647 0.4626647 0.4714495

[15] 0.4758419 0.4846266 0.4846266 0.4802343 0.4890190 0.4978038

* + 1. Findings:-

KNN is 2 having highest which is 0.53.

* 1. For Mode Value:-
     1. KNN 1 thro 20

[1] 0.5241581 0.5300146 0.5139092 0.5021962 0.4890190 0.4729136 0.4641288

[8] 0.4582723 0.4553441 0.4597365 0.4597365 0.4626647 0.4626647 0.4714495

[15] 0.4758419 0.4846266 0.4846266 0.4802343 0.4890190 0.4978038 0.4890190

* + 1. Findings:-

KNN is 2 having highest which is 0.53. Similar to data with out “?”

* 1. For perturbated Value:-
     1. KNN 1 thro 20

[1] 0.5080527 0.5095168 0.4612006 0.4494876 0.4685212 0.4597365 0.4421669

[8] 0.4275256 0.4245974 0.3909224 0.4202050 0.4143485 0.4114202 0.4216691

[15] 0.4055637 0.4084919 0.4070278 0.4040996 0.4055637 0.4114202

* + 1. Findings:-

Accuracy is lowering when perturbation is done.

1. For Imputation Value:-
   * 1. KNN 1 thro 20

[1] 0.5064378 0.5207439 0.5193133 0.4849785 0.4806867 0.4649499 0.4592275

[8] 0.4592275 0.4549356 0.4721030 0.4577969 0.4721030 0.4678112 0.4706724

[15] 0.4749642 0.4692418 0.4663805 0.4692418 0.4663805 0.4721030

* + 1. Findings:-

Accuracy is better than perturbation but not as good as Mode. This might true only for this case

Assigning binary Values

data$V7[(data1$V7=="?")] <- 0

data$V7[!(data1$V7=="?")] = 1

KNN Value are

[1] 0.9542203 0.9542203 0.9527897 0.9513591 0.9713877 0.9713877 0.9771102

[8] 0.9771102 0.9728183 0.9771102 0.9771102 0.9771102 0.9771102 0.9771102

[15] 0.9771102 0.9771102 0.9771102 0.9771102 0.9771102 0.9771102

KNN is 9. Much better accuracy. Conclusion: KNN appears to work better for binary or we need more data if there are more classifications.

**Question3:-**

We get hundreds of production tickets per day. I want to optimize the resource as the tickets go through Analysis team, development team and testing team. Also I want assign priorities based on the severity of the tickets. Additionally if a ticket is reopened that will get a higher priority. Additionally I want to limit the support tickets to 10 per department if the queue is busy. If the ticket is from VP and above then that will get highest priority irrespective of severity. We to come up correct staffing plan such a way that there are no more 100 open tickets and less than 2 critical tickets at the end of the day.

Question 1:-