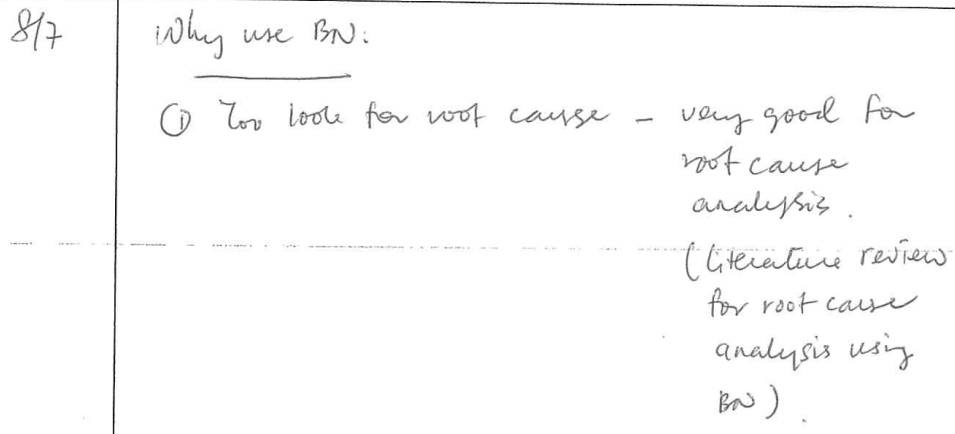
PHD - Progress update 14/10/2015

Fauzy Bin Che Yayah

October 14, 2015

* Why use Bayesian Net ?

1. To look for the root cause - the dataset is in factor type which is status not a number.
2. To find out the probabilistic relationship between the symptom error code and the resolution



1

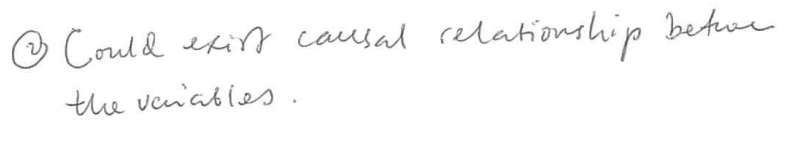
|  |  |  |  |
| --- | --- | --- | --- |
| Citation | Method | Output | Conclusions |
| TroubleMiner: Mining network trouble tickets Medem, A. ; Akodjenou, M.-I ; Teixeira, R. 20091 | Trouble tickets classification | \* Automation process on clustering the free text inside the description of the trouble tickets. \* Choosing the correct keywords for the analysis | Using term frequency distance between trouble tickets and similarity between clusters |
| Knowledge Discovery from Trouble Ticketing Reports in a Large Telecommunication Company Temprado, Y. ; Garcia, C. ; Molinero, F.J. 2009 | Data Mining , Text Mining and Machine Learning , Bayes Net, Naïve Bayes | Prediction on the next action of trouble tickets ,Different snapshots were added to the machine learning algorithm for training | Combination of multiple method to construct the recommendation , Using Bayesian for prediction |
| A Bayesian Approach To Stochastic Root Finding 2011 | x | x | x |
| A Fully Bayesian Approach For Unit Root Testing 2011 | x | x | x |
| Online Root-Cause Analysis Of Alarms In Discrete Bayesian 2014 | x | x | x |
| Documents Categorization Based On Bayesian Spanning Tree 2006 | x | x | x |
| Benefits of a Bayesian Approach to Anomaly and Failure 2009 | x | x | x |

List of literature review regarding Bayesian Net :-

1.A real-life application of multi-agent systems for fault diagnosis in the provision of an Internet business service

2.A Bayesian Network approach to diagnosing the root cause of failure

3.sss



# Process on gathering the dataset

* Acquiring dataset for 100 records, for each zone , randomize , selective year ; ie . 2015
* Rules :-

|  |  |
| --- | --- |
| Rules | Description |
| status = 'Closed' | Dataset must be closed for complete information |
| network\_tt\_id is NULL | Dataset must be not related to Network Trouble Ticket |
| trouble ticket type <> PASSIVE | Trouble Ticket must related to the Active elements such as routers, switches , modem , etc |
| installed\_date is NOT NULL | This field must have value |
| created\_date is NOT NULL | This field must have value |
| closed\_date is NOT NULL | This field must have value |
| closed\_date is NOT NULL | This field must have value |
| product is NOT NULL | This field must have value |
| sub\_product is NOT NULL | This field must have value |
| length description > 10 | This field is useful for text analysis |
| rand() | Record selection is in random mode |
| zone | Should selective from different zone , sparse |

For sample purpose - selecting dataset from ZONE KEPONG for the analysis due to this zone has the highest records inside the Trouble Ticket dataset.

* Using Impala for the data retrieval :-

Documentation - <https://github.com/piersharding/dplyrimpaladb>

* Data processing using DplyrImpalaDb
* Package installation manual below :-

install.packages(c("RJDBC", "devtools", "dplyr"))  
devtools::install\_github("jwills/dplyrimpaladb")  
install.packages("dplyrimpaladb")

* Basic notes why choosing Impala.

1. Cloudera 'Impala', which is a massively parallel processing (MPP) SQL query engine runs natively in Apache Hadoop
2. Impala's Place in the Big Data Ecosystem
3. Flexibility for Big Data Workflow
4. High-Performance Analytics

# Connection to Impala

Basic Impala drivers can be downloaded from <https://github.com/Mu-Sigma/RImpala/blob/master/impala-jdbc-cdh5.zip>

Below is the components required and how to set the class path for the Impala drivers , RJava , RJDBC and dplyr

suppressWarnings(suppressMessages(library("rJava")))  
suppressWarnings(suppressMessages(library("RJDBC")))  
suppressWarnings(suppressMessages(library("dplyr")))  
suppressWarnings(suppressMessages(library("caret")))  
suppressWarnings(suppressMessages(library("corrplot")))  
suppressWarnings(suppressMessages(library("lazy")))  
suppressWarnings(suppressMessages(library("dplyrimpaladb")))  
suppressWarnings(suppressMessages(library("rpart")))  
  
  
.jaddClassPath(c(list.files(paste(getwd(),"/lib",sep = ''),pattern="jar$",full.names=T)))  
  
.jinit(classpath = c(list.files(paste(getwd(),"/lib",sep = ''),pattern="jar$",full.names=T)))  
  
dplyr.jdbc.classpath = c(list.files(paste(getwd(),"/lib",sep = ''),pattern="jar$",full.names=T))  
  
conn <- src\_impaladb(dbname='nova', host='10.54.1.151')

## Loading required package: testthat

## [1] "here:"  
## [1] FALSE

* Zone list

result <- tbl(conn, sql("select zone from nova.nova\_trouble\_ticket where zone <> 'null' group by zone order by zone limit 1000"))  
as.data.frame(result)

## zone  
## 1 ZONE AIR ITAM  
## 2 ZONE BANGI  
## 3 ZONE BANGSAR  
## 4 ZONE BANTING  
## 5 ZONE BATU  
## 6 ZONE BATU PAHAT  
## 7 ZONE BAYAN BARU  
## 8 ZONE BINTULU  
## 9 ZONE BUKIT ANGGERIK  
## 10 ZONE BUKIT MERTAJAM  
## 11 ZONE BUKIT RAJA  
## 12 ZONE BUTTERWORTH  
## 13 ZONE CYBERJAYA  
## 14 ZONE GOMBAK  
## 15 ZONE IPOH  
## 16 ZONE KAJANG  
## 17 ZONE KEPONG  
## 18 ZONE KERAMAT  
## 19 ZONE KINRARA  
## 20 ZONE KL CENTRAL  
## 21 ZONE KLANG  
## 22 ZONE KOTA KINABALU SELATAN  
## 23 ZONE KOTA KINABALU UTARA  
## 24 ZONE KUCHING  
## 25 ZONE KULIM  
## 26 ZONE LANGKAWI  
## 27 ZONE MALURI  
## 28 ZONE MELAKA UTARA  
## 29 ZONE MIRI  
## 30 ZONE N. SEMBILAN UTARA  
## 31 ZONE PANDAN  
## 32 ZONE PELANGI  
## 33 ZONE PERLIS  
## 34 ZONE PETALING JAYA  
## 35 ZONE PUCHONG  
## 36 ZONE SEBERANG JAYA  
## 37 ZONE SENAI  
## 38 ZONE SG PETANI  
## 39 ZONE SHAH ALAM  
## 40 ZONE SIBU  
## 41 ZONE SKUDAI PONTIAN  
## 42 ZONE STAMPIN  
## 43 ZONE SUBANG JAYA  
## 44 ZONE TAMAN PETALING  
## 45 ZONE TAMPOI  
## 46 ZONE TAR  
## 47 ZONE TASEK  
## 48 ZONE TASIK AMPANG  
## 49 ZONE TDI  
## 50 ZONE TELUK INTAN  
## 51 ZONE TERENGGANU SELATAN  
## 52 ZONE TERUNTUM

* Trouble Ticket Data Dictionary

result <- tbl(conn, sql("select \* from nova\_trouble\_ticket where zone <> 'null' limit 1"))  
as.data.frame(apply(as.data.frame(result),2,class))

## apply(as.data.frame(result), 2, class)  
## tt\_row\_id character  
## tt\_num character  
## tt\_type character  
## tt\_sub\_type character  
## status character  
## severity character  
## important\_message character  
## appointment\_flag character  
## nova\_account\_name character  
## nova\_subscriber\_num character  
## nova\_account\_num character  
## package\_row\_id character  
## created\_by character  
## category character  
## symptom\_error\_code character  
## priority character  
## product character  
## sub\_product character  
## package\_name character  
## network\_tt\_id character  
## swap\_order\_num character  
## cause\_category character  
## cause\_code character  
## resolution\_code character  
## closure\_category character  
## resolution\_team character  
## service\_affected character  
## service\_order\_num character  
## btu\_type character  
## owner character  
## owner\_name character  
## group\_owner character  
## owner\_position character  
## btu\_platform character  
## dp\_location character  
## created\_date character  
## pending\_verify\_date character  
## closed\_by character  
## closed\_date character  
## source character  
## installed\_date character  
## description character  
## repeat\_ticket\_count character  
## follow\_up\_ticket\_count character  
## fdp\_device\_name character  
## fdp\_site\_name character  
## olt\_site\_name character  
## exchange character  
## timestamp character  
## contact\_id character  
## contact\_name character  
## contact\_office\_phone character  
## contact\_mobile\_phone character  
## contact\_home\_phone character  
## contact\_email\_addr character  
## due\_date character  
## part\_num character  
## network\_layer character  
## network\_row\_id character  
## asset\_id character  
## ptt character  
## zone character  
## service\_point\_id character

# Getting the dataset from Impala

Sample dataset - Selection trouble tickets only from Zone Kepong. The SQL is define by :-

* Why Kepong zone ?

Zone Kepong contains very rich information especially for the textual analysis and also one of the largest composition of the cause code & the resolution code which is good for the supervised learning.

|  |  |
| --- | --- |
| Rules | Description |
| a.status like '%Closed%' | Dataset must be closed for complete information |
| network\_tt\_id = 'null' | Dataset must be not related to Network Trouble Ticket |
| trouble ticket type <> PASSIVE | Trouble Ticket must related to the Active elements such as routers, switches , modem , etc. Excluding for now if related to the 3rd party causes , customer behavior and Passive elements |
| installed\_date is NOT NULL | This field must have value |
| created\_date is NOT NULL | This field must have value |
| closed\_date is NOT NULL | This field must have value |
| closed\_date is NOT NULL | This field must have value |
| product is NOT NULL | This field must have value |
| sub\_product is NOT NULL | This field must have value |
| length description > 10 | This field is useful for text analysis |
| rand() | Record selection is in random mode |
| zone | Should selective from different zone , sparse control |

Generated SQL :-

select \* from nova\_trouble\_ticket a join active\_code b on (trim(a.cause\_code) = trim(b.cause\_code)) join exchange\_zone c ON (trim(a.exchange)=trim(c.building\_id)) and (b.code <> 'PASSIVE' ) where c.zone\_name like '%ZONE KEPONG%' and a.status like '%Closed%' and length(a.cause\_category) > 1 and length(a.created\_date) > 6 and length(a.closed\_date) > 6 and length(a.installed\_date) > 6 and a.package\_name not like '%null%' and a.product not like '%null%' and a.sub\_product not like '%null%' and length(a.description) > 10 and network\_tt\_id = 'null' order by rand() limit 10000 "

# Datset filtering

Removing non-related fields such as trouble ticket key , trouble ticket number , trouble ticket date etc.

conn <- src\_impaladb(dbname='nova', host='10.54.1.151')

## [1] "here:"  
## [1] FALSE

result <- tbl(conn, sql("select a.tt\_row\_id,a.tt\_num,a.tt\_type,a.tt\_sub\_type,a.status,a.severity,a.important\_message,a.appointment\_flag,a.nova\_account\_name,a.nova\_subscriber\_num,a.nova\_account\_num,a.package\_row\_id,a.created\_by,a.category,a.symptom\_error\_code,a.priority,a.product,a.sub\_product,a.package\_name,a.network\_tt\_id,a.swap\_order\_num,a.cause\_category,a.cause\_code,a.resolution\_code,a.closure\_category,a.resolution\_team,a.service\_affected,a.service\_order\_num,a.btu\_type,a.owner,a.owner\_name,a.group\_owner,a.owner\_position,a.btu\_platform,a.dp\_location,a.created\_date,a.pending\_verify\_date,a.closed\_by,a.closed\_date,a.source,a.installed\_date,a.description,a.repeat\_ticket\_count,a.follow\_up\_ticket\_count,a.fdp\_device\_name,  
a.fdp\_site\_name,a.olt\_site\_name,a.exchange,a.`timestamp`,a.contact\_id,a.contact\_name,a.contact\_office\_phone,a.contact\_mobile\_phone,a.contact\_home\_phone,a.contact\_email\_addr,a.due\_date,a.part\_num,a.network\_layer,a.network\_row\_id,a.asset\_id,a.ptt,a.zone,a.service\_point\_id , c.zone\_name, c.district,c.state, c.region from nova\_trouble\_ticket a join active\_code b on (trim(a.cause\_code) = trim(b.cause\_code)) join exchange\_zone c ON (trim(a.exchange)=trim(c.building\_id)) and (b.code <> 'PASSIVE' ) where c.zone\_name like '%ZONE KEPONG%' and a.status like '%Closed%' and length(a.cause\_category) > 1 and length(a.created\_date) > 6 and length(a.closed\_date) > 6 and length(a.installed\_date) > 6 and a.package\_name not like '%null%' and a.product not like '%null%' and a.sub\_product not like '%null%' and length(a.description) > 10 and a.network\_tt\_id = 'null' order by rand() limit 100"))  
  
result <- as.data.frame(result)

Close the connection from Impala

x <- conn$con  
class(x) <- c('JDBCConnection')  
dbDisconnect(x)

## [1] TRUE

Save the class as the data.frame

df <- as.data.frame(result)  
df$contact\_name <- NULL  
df$contact\_home\_phone <- NULL  
df$contact\_email\_addr <- NULL  
df$contact\_office\_phone <- NULL  
df$contact\_mobile\_phone <- NULL  
df$`tt\_row\_id` <- NULL  
df$`tt\_num` <- NULL  
df$tt\_type <- NULL  
df$`created\_date` <- NULL  
df$`closed\_date` <- NULL  
df$`installed\_date` <- NULL  
df$timestamp <- NULL  
df$service\_point\_id <- NULL  
df$contact\_id <- NULL  
df$owner\_position <- NULL  
df$tt\_sub\_type <- NULL  
df$severity <- NULL  
df$status <- NULL  
df$important\_message <- NULL  
df$network\_tt\_id <- NULL  
df$swap\_order\_num <- NULL  
df$appointment\_flag <- NULL  
df$nova\_account\_name <- NULL  
df$nova\_subscriber\_num <- NULL  
df$nova\_account\_num <- NULL  
df$repeat\_ticket\_count <- NULL  
df$follow\_up\_ticket\_count <- NULL  
df$service\_order\_num <- NULL  
df$source <- NULL  
df$owner\_name <- NULL  
df$description <- NULL  
df$due\_date <- NULL  
df$part\_num <- NULL  
df$zone <- NULL  
df$ptt <- NULL  
df$asset\_id <- NULL  
df$network\_layer <- NULL  
df$network\_row\_id <- NULL  
df$pending\_verify\_date <- NULL  
df$package\_row\_id <- NULL  
df$priority <- NULL  
summary(df)

## created\_by category symptom\_error\_code  
## Length:100 Length:100 Length:100   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## product sub\_product package\_name   
## Length:100 Length:100 Length:100   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## cause\_category cause\_code resolution\_code   
## Length:100 Length:100 Length:100   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## closure\_category resolution\_team service\_affected   
## Length:100 Length:100 Length:100   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## btu\_type owner group\_owner   
## Length:100 Length:100 Length:100   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## btu\_platform dp\_location closed\_by   
## Length:100 Length:100 Length:100   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## fdp\_device\_name fdp\_site\_name olt\_site\_name   
## Length:100 Length:100 Length:100   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## exchange zone\_name district   
## Length:100 Length:100 Length:100   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## state region   
## Length:100 Length:100   
## Class :character Class :character   
## Mode :character Mode :character

Looping the columns name and rename it to [column name]+1 as the factor name

for(i in names(df)){  
  
 num <- as.numeric(as.factor(df[,i]))-1  
 df <- cbind(df,num)  
 names(df)[names(df)=="num"] <- paste(names(df[i]),"\_factor",sep = "")  
 print(paste(names(df[i]),"1",sep = ""))  
}

## [1] "created\_by1"  
## [1] "category1"  
## [1] "symptom\_error\_code1"  
## [1] "product1"  
## [1] "sub\_product1"  
## [1] "package\_name1"  
## [1] "cause\_category1"  
## [1] "cause\_code1"  
## [1] "resolution\_code1"  
## [1] "closure\_category1"  
## [1] "resolution\_team1"  
## [1] "service\_affected1"  
## [1] "btu\_type1"  
## [1] "owner1"  
## [1] "group\_owner1"  
## [1] "btu\_platform1"  
## [1] "dp\_location1"  
## [1] "closed\_by1"  
## [1] "fdp\_device\_name1"  
## [1] "fdp\_site\_name1"  
## [1] "olt\_site\_name1"  
## [1] "exchange1"  
## [1] "zone\_name1"  
## [1] "district1"  
## [1] "state1"  
## [1] "region1"

df <- df[27:52]  
names(df)

## [1] "created\_by\_factor" "category\_factor"   
## [3] "symptom\_error\_code\_factor" "product\_factor"   
## [5] "sub\_product\_factor" "package\_name\_factor"   
## [7] "cause\_category\_factor" "cause\_code\_factor"   
## [9] "resolution\_code\_factor" "closure\_category\_factor"   
## [11] "resolution\_team\_factor" "service\_affected\_factor"   
## [13] "btu\_type\_factor" "owner\_factor"   
## [15] "group\_owner\_factor" "btu\_platform\_factor"   
## [17] "dp\_location\_factor" "closed\_by\_factor"   
## [19] "fdp\_device\_name\_factor" "fdp\_site\_name\_factor"   
## [21] "olt\_site\_name\_factor" "exchange\_factor"   
## [23] "zone\_name\_factor" "district\_factor"   
## [25] "state\_factor" "region\_factor"

Remove the predictors column which might have one unique value which can leads to zero variance result

The list below is non-zero variance variables

df <- df[,-nearZeroVar(df)]   
names(df)

## [1] "created\_by\_factor" "sub\_product\_factor"   
## [3] "package\_name\_factor" "cause\_category\_factor"   
## [5] "cause\_code\_factor" "resolution\_code\_factor"   
## [7] "closure\_category\_factor" "resolution\_team\_factor"   
## [9] "service\_affected\_factor" "btu\_type\_factor"   
## [11] "owner\_factor" "group\_owner\_factor"   
## [13] "btu\_platform\_factor" "dp\_location\_factor"   
## [15] "closed\_by\_factor" "fdp\_device\_name\_factor"   
## [17] "fdp\_site\_name\_factor" "olt\_site\_name\_factor"   
## [19] "exchange\_factor"

Find the correlation between the variables using Pearson.

correlations <- cor(df, use="pairwise.complete.obs", method="pearson")  
print(correlations)

## created\_by\_factor sub\_product\_factor  
## created\_by\_factor 1.00000000 0.15527975  
## sub\_product\_factor 0.15527975 1.00000000  
## package\_name\_factor -0.02169162 0.54092685  
## cause\_category\_factor 0.08695198 0.13883471  
## cause\_code\_factor 0.06709369 0.05714152  
## resolution\_code\_factor 0.14138960 0.13799330  
## closure\_category\_factor -0.02963631 0.08515177  
## resolution\_team\_factor -0.03770784 0.13548991  
## service\_affected\_factor -0.25701646 0.03433917  
## btu\_type\_factor -0.09380821 0.10097992  
## owner\_factor -0.19666128 0.04029265  
## group\_owner\_factor -0.15149773 0.01206344  
## btu\_platform\_factor -0.10156021 0.09926766  
## dp\_location\_factor 0.12466975 -0.12279996  
## closed\_by\_factor -0.25431439 -0.07507889  
## fdp\_device\_name\_factor 0.09238575 -0.04956953  
## fdp\_site\_name\_factor 0.09238575 -0.04956953  
## olt\_site\_name\_factor 0.15128625 -0.09028615  
## exchange\_factor 0.14029797 -0.10699648  
## package\_name\_factor cause\_category\_factor  
## created\_by\_factor -0.021691625 0.086951975  
## sub\_product\_factor 0.540926848 0.138834714  
## package\_name\_factor 1.000000000 0.067470644  
## cause\_category\_factor 0.067470644 1.000000000  
## cause\_code\_factor 0.047368883 0.712988290  
## resolution\_code\_factor 0.119704464 0.500391755  
## closure\_category\_factor -0.089130468 0.120060296  
## resolution\_team\_factor 0.014645039 0.131335423  
## service\_affected\_factor 0.052907590 -0.016466843  
## btu\_type\_factor -0.001137447 -0.203827348  
## owner\_factor -0.028714259 0.011768807  
## group\_owner\_factor -0.014372786 0.061018044  
## btu\_platform\_factor -0.003706492 -0.197956422  
## dp\_location\_factor -0.015828188 0.158303552  
## closed\_by\_factor -0.024573945 0.001055907  
## fdp\_device\_name\_factor -0.144243743 0.057241455  
## fdp\_site\_name\_factor -0.144243743 0.057241455  
## olt\_site\_name\_factor -0.168453604 0.076975091  
## exchange\_factor -0.147199700 0.178258462  
## cause\_code\_factor resolution\_code\_factor  
## created\_by\_factor 0.067093689 0.1413896042  
## sub\_product\_factor 0.057141516 0.1379933048  
## package\_name\_factor 0.047368883 0.1197044645  
## cause\_category\_factor 0.712988290 0.5003917552  
## cause\_code\_factor 1.000000000 0.7753208563  
## resolution\_code\_factor 0.775320856 1.0000000000  
## closure\_category\_factor -0.049244255 0.0066694846  
## resolution\_team\_factor 0.224932484 0.1552000940  
## service\_affected\_factor -0.070053606 -0.1011122000  
## btu\_type\_factor -0.136602336 -0.0738782931  
## owner\_factor 0.058108970 0.0004631452  
## group\_owner\_factor 0.008566294 0.0083778041  
## btu\_platform\_factor -0.126848776 -0.1003084843  
## dp\_location\_factor 0.147880741 0.1438055330  
## closed\_by\_factor 0.041448754 -0.0177078401  
## fdp\_device\_name\_factor 0.155037743 0.1973165650  
## fdp\_site\_name\_factor 0.155037743 0.1973165650  
## olt\_site\_name\_factor 0.179745623 0.2079004206  
## exchange\_factor 0.284582802 0.3193373604  
## closure\_category\_factor resolution\_team\_factor  
## created\_by\_factor -0.029636311 -0.037707840  
## sub\_product\_factor 0.085151769 0.135489910  
## package\_name\_factor -0.089130468 0.014645039  
## cause\_category\_factor 0.120060296 0.131335423  
## cause\_code\_factor -0.049244255 0.224932484  
## resolution\_code\_factor 0.006669485 0.155200094  
## closure\_category\_factor 1.000000000 0.050122538  
## resolution\_team\_factor 0.050122538 1.000000000  
## service\_affected\_factor -0.030970231 0.133778200  
## btu\_type\_factor -0.075035135 -0.006358404  
## owner\_factor 0.186655414 0.398690724  
## group\_owner\_factor 0.134472432 0.480491139  
## btu\_platform\_factor -0.082874803 0.010648015  
## dp\_location\_factor 0.004661405 -0.125459645  
## closed\_by\_factor 0.200659204 0.372504975  
## fdp\_device\_name\_factor -0.105392953 -0.016519320  
## fdp\_site\_name\_factor -0.105392953 -0.016519320  
## olt\_site\_name\_factor -0.096596540 -0.107205649  
## exchange\_factor -0.035628014 -0.115819388  
## service\_affected\_factor btu\_type\_factor  
## created\_by\_factor -0.25701646 -0.093808213  
## sub\_product\_factor 0.03433917 0.100979918  
## package\_name\_factor 0.05290759 -0.001137447  
## cause\_category\_factor -0.01646684 -0.203827348  
## cause\_code\_factor -0.07005361 -0.136602336  
## resolution\_code\_factor -0.10111220 -0.073878293  
## closure\_category\_factor -0.03097023 -0.075035135  
## resolution\_team\_factor 0.13377820 -0.006358404  
## service\_affected\_factor 1.00000000 0.208818622  
## btu\_type\_factor 0.20881862 1.000000000  
## owner\_factor 0.11577434 -0.129381614  
## group\_owner\_factor 0.13410639 -0.050493541  
## btu\_platform\_factor 0.21778393 0.979861006  
## dp\_location\_factor -0.23535696 -0.868914656  
## closed\_by\_factor 0.10233429 -0.160300006  
## fdp\_device\_name\_factor 0.16440462 0.542856840  
## fdp\_site\_name\_factor 0.16440462 0.542856840  
## olt\_site\_name\_factor 0.14728802 0.480266874  
## exchange\_factor 0.13089858 0.086755774  
## owner\_factor group\_owner\_factor  
## created\_by\_factor -0.1966612763 -0.151497728  
## sub\_product\_factor 0.0402926540 0.012063441  
## package\_name\_factor -0.0287142590 -0.014372786  
## cause\_category\_factor 0.0117688065 0.061018044  
## cause\_code\_factor 0.0581089701 0.008566294  
## resolution\_code\_factor 0.0004631452 0.008377804  
## closure\_category\_factor 0.1866554136 0.134472432  
## resolution\_team\_factor 0.3986907237 0.480491139  
## service\_affected\_factor 0.1157743353 0.134106393  
## btu\_type\_factor -0.1293816138 -0.050493541  
## owner\_factor 1.0000000000 0.218619903  
## group\_owner\_factor 0.2186199033 1.000000000  
## btu\_platform\_factor -0.1552807996 -0.045880965  
## dp\_location\_factor 0.0498728115 -0.012547233  
## closed\_by\_factor 0.9432061031 0.265214327  
## fdp\_device\_name\_factor -0.2767225750 -0.081284309  
## fdp\_site\_name\_factor -0.2767225750 -0.081284309  
## olt\_site\_name\_factor -0.2852806971 -0.108822328  
## exchange\_factor -0.2545881347 -0.102514775  
## btu\_platform\_factor dp\_location\_factor  
## created\_by\_factor -0.101560212 0.124669752  
## sub\_product\_factor 0.099267663 -0.122799960  
## package\_name\_factor -0.003706492 -0.015828188  
## cause\_category\_factor -0.197956422 0.158303552  
## cause\_code\_factor -0.126848776 0.147880741  
## resolution\_code\_factor -0.100308484 0.143805533  
## closure\_category\_factor -0.082874803 0.004661405  
## resolution\_team\_factor 0.010648015 -0.125459645  
## service\_affected\_factor 0.217783927 -0.235356963  
## btu\_type\_factor 0.979861006 -0.868914656  
## owner\_factor -0.155280800 0.049872812  
## group\_owner\_factor -0.045880965 -0.012547233  
## btu\_platform\_factor 1.000000000 -0.844521313  
## dp\_location\_factor -0.844521313 1.000000000  
## closed\_by\_factor -0.185751967 0.073315502  
## fdp\_device\_name\_factor 0.510803184 -0.391897466  
## fdp\_site\_name\_factor 0.510803184 -0.391897466  
## olt\_site\_name\_factor 0.441287156 -0.303638041  
## exchange\_factor 0.058843162 0.087077824  
## closed\_by\_factor fdp\_device\_name\_factor  
## created\_by\_factor -0.254314388 0.09238575  
## sub\_product\_factor -0.075078886 -0.04956953  
## package\_name\_factor -0.024573945 -0.14424374  
## cause\_category\_factor 0.001055907 0.05724145  
## cause\_code\_factor 0.041448754 0.15503774  
## resolution\_code\_factor -0.017707840 0.19731656  
## closure\_category\_factor 0.200659204 -0.10539295  
## resolution\_team\_factor 0.372504975 -0.01651932  
## service\_affected\_factor 0.102334286 0.16440462  
## btu\_type\_factor -0.160300006 0.54285684  
## owner\_factor 0.943206103 -0.27672257  
## group\_owner\_factor 0.265214327 -0.08128431  
## btu\_platform\_factor -0.185751967 0.51080318  
## dp\_location\_factor 0.073315502 -0.39189747  
## closed\_by\_factor 1.000000000 -0.31948586  
## fdp\_device\_name\_factor -0.319485862 1.00000000  
## fdp\_site\_name\_factor -0.319485862 1.00000000  
## olt\_site\_name\_factor -0.321278219 0.94847595  
## exchange\_factor -0.280545533 0.81460447  
## fdp\_site\_name\_factor olt\_site\_name\_factor  
## created\_by\_factor 0.09238575 0.15128625  
## sub\_product\_factor -0.04956953 -0.09028615  
## package\_name\_factor -0.14424374 -0.16845360  
## cause\_category\_factor 0.05724145 0.07697509  
## cause\_code\_factor 0.15503774 0.17974562  
## resolution\_code\_factor 0.19731656 0.20790042  
## closure\_category\_factor -0.10539295 -0.09659654  
## resolution\_team\_factor -0.01651932 -0.10720565  
## service\_affected\_factor 0.16440462 0.14728802  
## btu\_type\_factor 0.54285684 0.48026687  
## owner\_factor -0.27672257 -0.28528070  
## group\_owner\_factor -0.08128431 -0.10882233  
## btu\_platform\_factor 0.51080318 0.44128716  
## dp\_location\_factor -0.39189747 -0.30363804  
## closed\_by\_factor -0.31948586 -0.32127822  
## fdp\_device\_name\_factor 1.00000000 0.94847595  
## fdp\_site\_name\_factor 1.00000000 0.94847595  
## olt\_site\_name\_factor 0.94847595 1.00000000  
## exchange\_factor 0.81460447 0.86376024  
## exchange\_factor  
## created\_by\_factor 0.14029797  
## sub\_product\_factor -0.10699648  
## package\_name\_factor -0.14719970  
## cause\_category\_factor 0.17825846  
## cause\_code\_factor 0.28458280  
## resolution\_code\_factor 0.31933736  
## closure\_category\_factor -0.03562801  
## resolution\_team\_factor -0.11581939  
## service\_affected\_factor 0.13089858  
## btu\_type\_factor 0.08675577  
## owner\_factor -0.25458813  
## group\_owner\_factor -0.10251477  
## btu\_platform\_factor 0.05884316  
## dp\_location\_factor 0.08707782  
## closed\_by\_factor -0.28054553  
## fdp\_device\_name\_factor 0.81460447  
## fdp\_site\_name\_factor 0.81460447  
## olt\_site\_name\_factor 0.86376024  
## exchange\_factor 1.00000000

Find the highest correlated variables.

|  |  |
| --- | --- |
| Rules | Description |
| - +.70 or higher | Very strong relationship |
| - +.40 to +.69 | Strong positive relationship |
| - +.30 to +.39 | Moderate relationship |
| - +.20 to +.29 | weak relationship |
| - +.01 to +.19 | No or negligible relationship |

# Choose 0.7 Very strong relationship  
highlyCorrelated <- findCorrelation(correlations, 0.7 ,verbose = FALSE,names = TRUE)  
highlyCorrelated

## [1] "olt\_site\_name\_factor" "fdp\_device\_name\_factor"  
## [3] "fdp\_site\_name\_factor" "btu\_type\_factor"   
## [5] "btu\_platform\_factor" "closed\_by\_factor"   
## [7] "cause\_code\_factor"

Summary of the correlated variables.

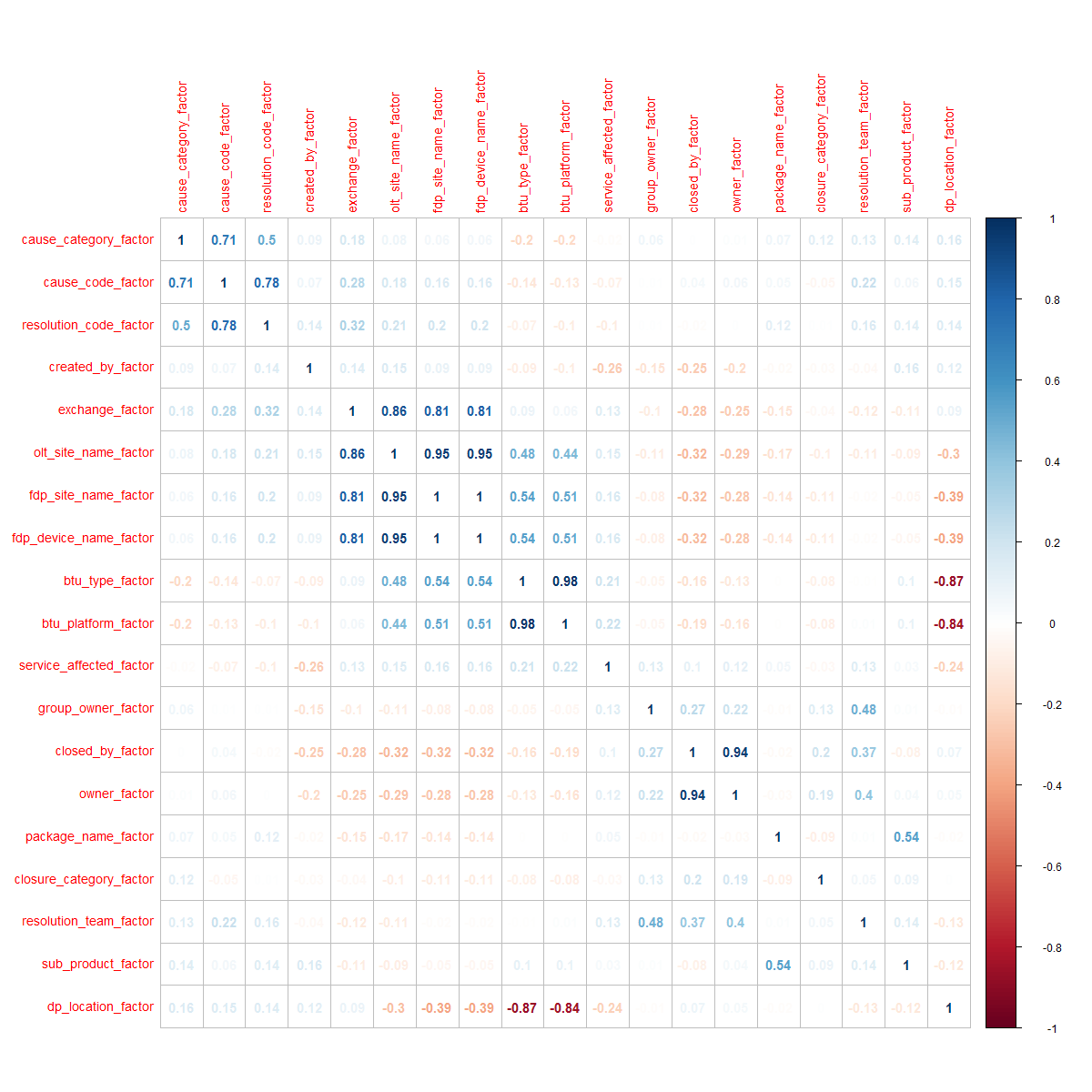
summary(correlations)

## created\_by\_factor sub\_product\_factor package\_name\_factor  
## Min. :-0.25702 Min. :-0.12280 Min. :-0.16845   
## 1st Qu.:-0.09768 1st Qu.:-0.04957 1st Qu.:-0.05892   
## Median : 0.06709 Median : 0.05714 Median :-0.01437   
## Mean : 0.04778 Mean : 0.10755 Mean : 0.05472   
## 3rd Qu.: 0.13248 3rd Qu.: 0.13674 3rd Qu.: 0.05014   
## Max. : 1.00000 Max. : 1.00000 Max. : 1.00000   
## cause\_category\_factor cause\_code\_factor resolution\_code\_factor  
## Min. :-0.20383 Min. :-0.13660 Min. :-0.101112   
## 1st Qu.: 0.03451 1st Qu.: 0.02501 1st Qu.: 0.003566   
## Median : 0.07698 Median : 0.06709 Median : 0.141390   
## Mean : 0.15482 Mean : 0.18592 Mean : 0.190431   
## 3rd Qu.: 0.14857 3rd Qu.: 0.20234 3rd Qu.: 0.202608   
## Max. : 1.00000 Max. : 1.00000 Max. : 1.000000   
## closure\_category\_factor resolution\_team\_factor service\_affected\_factor  
## Min. :-0.10539 Min. :-0.12546 Min. :-0.25702   
## 1st Qu.:-0.07895 1st Qu.:-0.01652 1st Qu.:-0.02372   
## Median :-0.02964 Median : 0.05012 Median : 0.11577   
## Mean : 0.05729 Mean : 0.14117 Mean : 0.09978   
## 3rd Qu.: 0.10261 3rd Qu.: 0.19007 3rd Qu.: 0.15585   
## Max. : 1.00000 Max. : 1.00000 Max. : 1.00000   
## btu\_type\_factor owner\_factor group\_owner\_factor   
## Min. :-0.868915 Min. :-0.28528 Min. :-0.151498   
## 1st Qu.:-0.111595 1st Qu.:-0.17597 1st Qu.:-0.065889   
## Median :-0.006358 Median : 0.01177 Median : 0.008378   
## Mean : 0.112771 Mean : 0.07474 Mean : 0.088118   
## 3rd Qu.: 0.344543 3rd Qu.: 0.15121 3rd Qu.: 0.134289   
## Max. : 1.000000 Max. : 1.00000 Max. : 1.000000   
## btu\_platform\_factor dp\_location\_factor closed\_by\_factor   
## Min. :-0.844521 Min. :-0.86891 Min. :-0.32128   
## 1st Qu.:-0.114205 1st Qu.:-0.26950 1st Qu.:-0.22003   
## Median :-0.003707 Median :-0.01255 Median :-0.01771   
## Mean : 0.104453 Mean :-0.08017 Mean : 0.05480   
## 3rd Qu.: 0.329535 3rd Qu.: 0.10587 3rd Qu.: 0.15150   
## Max. : 1.000000 Max. : 1.00000 Max. : 1.00000   
## fdp\_device\_name\_factor fdp\_site\_name\_factor olt\_site\_name\_factor  
## Min. :-0.39190 Min. :-0.39190 Min. :-0.3213   
## 1st Qu.:-0.09334 1st Qu.:-0.09334 1st Qu.:-0.1080   
## Median : 0.09239 Median : 0.09239 Median : 0.1473   
## Mean : 0.21568 Mean : 0.21568 Mean : 0.2086   
## 3rd Qu.: 0.52683 3rd Qu.: 0.52683 3rd Qu.: 0.4608   
## Max. : 1.00000 Max. : 1.00000 Max. : 1.0000   
## exchange\_factor   
## Min. :-0.28055   
## 1st Qu.:-0.10476   
## Median : 0.08708   
## Mean : 0.19662   
## 3rd Qu.: 0.30196   
## Max. : 1.00000

Plot correlated variables.

png(height=1200, width=1200, pointsize=15, file="corrplot.png")  
corrplot(correlations, method = "number",tl.cex = 0.9 ,addCoef.col="grey", order = "AOE")  
dev.off()

## png   
## 2



Feature selection process to confirm which variable does become the independent and resolution code is the dependent variable via GBM (Stochastic Gradient Boosting).

List of other available model - <http://topepo.github.io/caret/modelList.html>

set.seed(777)  
suppressWarnings(suppressMessages(library(mlbench)))  
control <- trainControl(method = "repeatedcv", number = 10, repeats = 3)  
model <-  
train(  
resolution\_code\_factor ~ ., data = df, method = "gbm", preProcess = "scale", trControl =  
control , verbose = FALSE  
)

## Loading required package: gbm

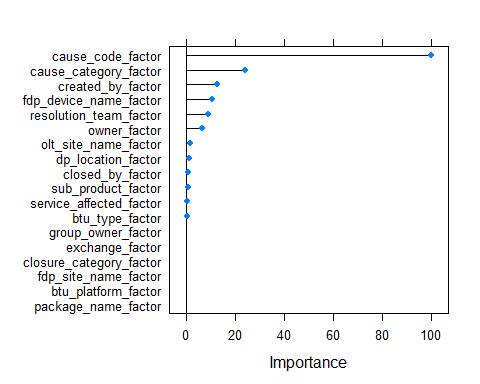
## Warning: package 'gbm' was built under R version 3.2.2

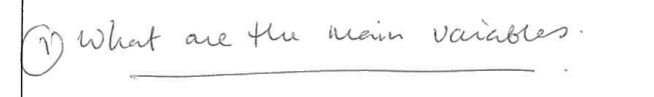
## Loading required package: survival  
##   
## Attaching package: 'survival'  
##   
## The following object is masked from 'package:caret':  
##   
## cluster  
##   
## Loading required package: splines  
## Loading required package: parallel  
## Loaded gbm 2.1.1  
## Loading required package: plyr  
## -------------------------------------------------------------------------  
## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)  
## -------------------------------------------------------------------------  
##   
## Attaching package: 'plyr'  
##   
## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

importance <- varImp(model, scale = TRUE)  
print(importance)

## gbm variable importance  
##   
## Overall  
## cause\_code\_factor 100.0000  
## cause\_category\_factor 24.0089  
## created\_by\_factor 12.6455  
## fdp\_device\_name\_factor 10.4811  
## resolution\_team\_factor 9.2062  
## owner\_factor 6.7560  
## olt\_site\_name\_factor 1.5966  
## dp\_location\_factor 1.3601  
## closed\_by\_factor 1.0719  
## sub\_product\_factor 0.8453  
## service\_affected\_factor 0.4593  
## btu\_type\_factor 0.2994  
## btu\_platform\_factor 0.0000  
## exchange\_factor 0.0000  
## fdp\_site\_name\_factor 0.0000  
## package\_name\_factor 0.0000  
## group\_owner\_factor 0.0000  
## closure\_category\_factor 0.0000

plot(importance)

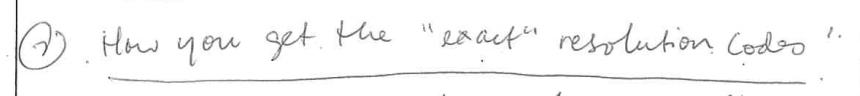




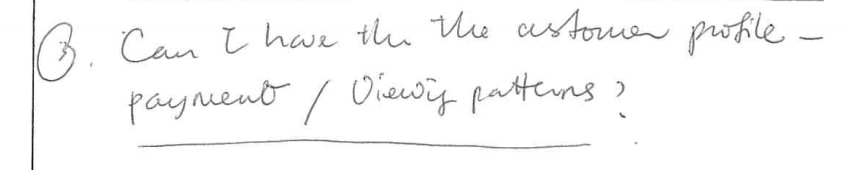
The main variables / factors are :-

* cause\_code\_factor
* resolution\_team\_factor
* cause\_category\_factor
* fdp\_device\_name\_factor
* owner\_factor
* created\_by\_factor
* service\_affected\_factor
* dp\_location\_factor
* btu\_type\_factor

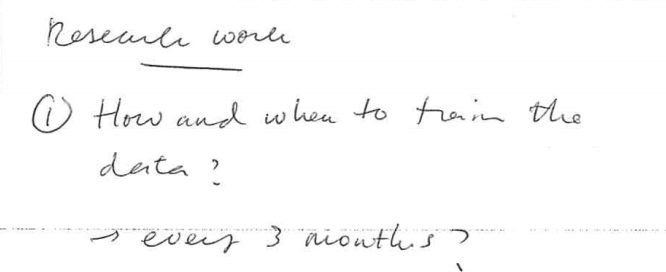
as listed from the importance plot



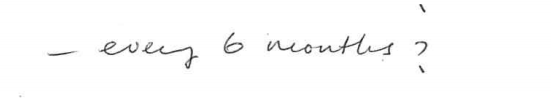
sss



sss



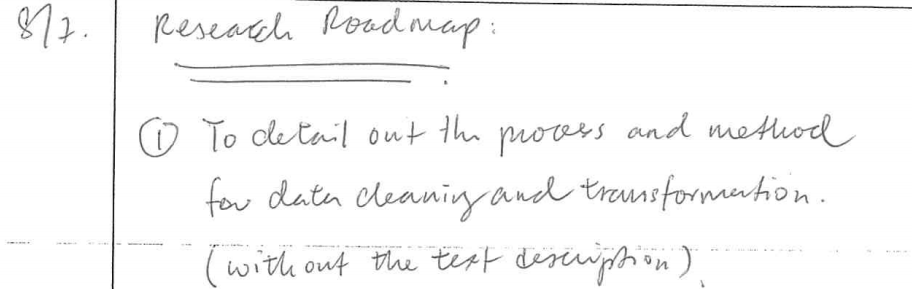
sss



sss



sss



1