# Executive Summary

This report presents an analysis of the Daily cell ranking whereas the raw data of cell incidences are obtained from Ericsson Expert Analytics (EEA). This report presents a method of ranking cells based on incident data associated with the cell. With this ranking we intend to shortlist top cells with high chances of needing some correction. The idea is to combine various incidents to obtain a score giving a performance measure for the cell. We adopted two approaches Principal Component Analysis and Auto-encoders to model the behavior of cells of various bands and use the internal correlation and the deviation in behavior of the cells from the modeled distribution as a measure of performance for the cells.

The goal of this analysis is to develop a classifier which can differentiate human operated devices from non-human operated devices.

The results presented are based on the analysis of 97 301 unique records collected over a single day. These records provide a wide variety of subscriber information including traffic related measures for audio, file sharing and email, as well as Service Level Index (SLI), Average Revenue Per User (ARPU) and so on.

The dataset contains only 5.1% of records associated to machine, while the rest (94.9%) are associated to human users. Analysis showed that many of the feature distributions appear to be distinctive of their class, suggesting their relevance to the classification problem.

A classifier has been trained against the provided data. The performance appears to be in the expected range, based on previous studies, with an F1 score of 80.2 (98.2% accuracy, 94.4% precision, 69.7% recall).

Contents

[1 Use Cases and Problem Statement 2](#_Toc10641071)

[2 Exploratory Data Analysis 3](#_Toc10641072)

[2.1 Dataset Overview 3](#_Toc10641073)

[2.1.1 Correlation Analysis: 4](#_Toc10641074)

[2.1.2 Defining Ranking Window: 9](#_Toc10641075)

[3 Ranking Methodology & Approach: 14](#_Toc10641076)

[3.1 PCA Approach 15](#_Toc10641077)

[3.2 PCA Result & Validation 15](#_Toc10641078)

[3.2.1 Score comparison across geography 17](#_Toc10641079)

[3.2.2 Summary & Conclusion 17](#_Toc10641080)

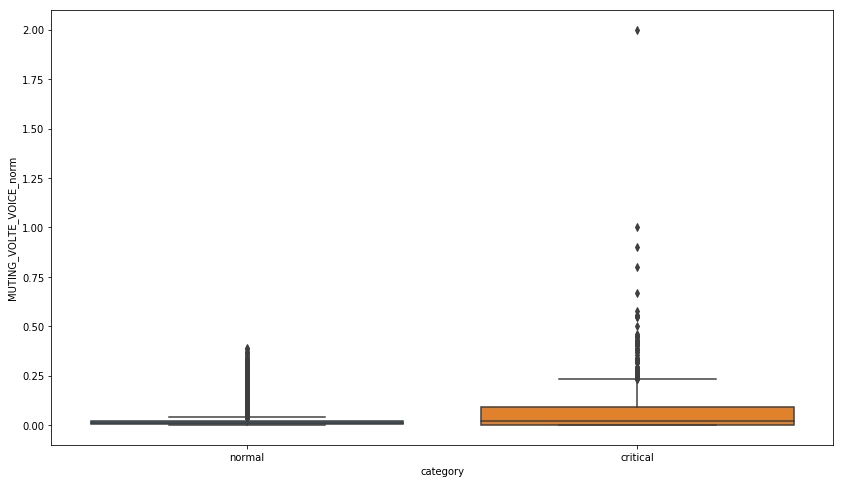
[3.3 Auto encoder Approach 17](#_Toc10641081)

[3.3.1 Reconstruction error 18](#_Toc10641082)

[3.3.2 Mapping reconstruction error to probability score 20](#_Toc10641083)

[3.3.3 Exponential weighted moving average of probability scores 21](#_Toc10641084)

[3.3.4 Result & Validation 21](#_Toc10641085)

[ 27](#_Toc10641086)

[3.3.5 Summary & Conclusion 30](#_Toc10641087)

[4 Appendix A 30](#_Toc10641088)

[4.1 Notebook 30](#_Toc10641089)

[4.2 Field Definitions 31](#_Toc10641090)

[4.3 Month wise record detail 32](#_Toc10641091)

[4.4 Correlation analysis of the incidence parameters across months 32](#_Toc10641092)

[4.5 Rolling window to determine 33](#_Toc10641093)

[4.6 Brief math behind PCA approach 33](#_Toc10641094)

[4.7 Acronyms 34](#_Toc10641095)

# Use Cases and Problem Statement

In this business problem, we would like to rank or score the cells based on the 4G VoLTE call incidences data obtained from EEA. For this study, six months’ historical data (Aug 2018 – Feb 2019) from NY Manhattan area are analyzed. Data collected from more than 9000 cells in daily time granularity are analyzed to figure out the daily ranking of the cells. Better the rank of a cell better the cell performance and eventually lower the recorded call incidence parameters. Among the various call incidence parameters; muting, garbling, soft drop, hard drops are physically known to be very influential parameters and the cells with recorded higher value of these parameters would drive lower rankings.

The goal of this activity is therefore to develop a ranking mechanism to rank the cells on daily basis based on historical call incidences. As the hardware and software configuration of the cells also do vary across the bands, the cells working under different bands would be ranked differently. We are considering three frequency bands at which the cells are operating for our analysis. They are:

1. Band 2100
2. Band 700
3. Band 1900

The ability to distinguish if a device is human operated or machine operated is of value to mobile service operator. Indeed, accurate information on user types can be used to: ​

* Provide a better Quality of Experience (QoE) to users.
* Offer more targeted plans and packages​ to customer.
* Detect fraudulent usage of “IoT Plan” by human users. ​

# Exploratory Data Analysis

## Dataset Overview

The dataset used for this study is maintained in a single csv file for each month’s call incidence records. In a month’s data, we have the call incidence parameters to be analyzed in the study. We have described the brief description of the fields in the appendix section.

Same fields are also present across all the months’ data and number of cells expected to be same across the months.

Six months’ (Aug 2018- Feb 2019) of data sets are used for this study. For each cell level, data is collected in daily time granularity. The actual variation of a cell performance within a day is not available to study. The data received from EEA is aggregated and rolled out in daily level. In the exploration, it is observed that few days’ data are also missing across all the bands. In the study, “band 0” and “band 1900\_2” have been ignored considering those records were erroneous. In fact, for these two bands, data are not available for all the days within a month. In the following table, we have summarized the days for which the call incidences are missing in the data sets across the months and bands. For example, in the month of Jan 2019, data set named as “incidents\_IncidentId\_Jan2019” 5 days (17th Jan to 22nd Jan) data are missing across all the bands. Consequently, the number of records also varies across the months. The summary of the number of records that were present across all the months are presented in the Appendix.

The reason behind the missing data for few days is possibly data was not extracted from EEA in those days. In the practical scenario when the model would go into the production system, it is expected that data would be available for all the days within a month and the ranking of a cell in a day would not be affected due to historical missing days of data.

### Correlation Analysis:

Each cell is functioning under a bandwidth. The values of all the call incidence parameters are not likely to be very much isolated, rather they do occur simultaneously. To validate the hypothesis, we study the correlation coefficient of all the call incidence parameters. The correlation matrix is derived by considering all the bands together and separately across all the bands.

In the study, it is observed that there is significant correlation among the various call incidence parameters. In the following figures, the month wise correlation among all the variables across the three bands (2100, 700 and 1900) are depicted.



Figure 1: Correlation matrix for all the bands together in August 2018

Considering all the bands together, we observed that call incidence parameter “DROP\_SERVICE\_TERMINATE\_VOLTE\_IMS” does not have any significant correlation with other parameters. This is to mention that correlations above 0.80 are colored with deep green; correlations between 0.40 to 0.80 are colored with light green; and correlations below 0.40 are left with no color. In the above table, correlation coefficients of all the call incidence parameters are described by considering all the bands together for the month of August 2018. The values of all the parameters (muting, garbling, soft drop etc.) are summed up for a month in each cell level and then the correlation numbers are derived. By this process of data role up, each cell has monthly aggregated value for all call incidence parameters. The detailed correlation coefficient numbers of all the other months along with the month of August 2018 are described in the embedded excel sheet attached in the appendix section.



Figure 2: Correlation matrix of band 2100 cells for the month of Aug 2018

In the table above, we have described the similar correlation numbers for the cells functioning under band 2100. These numbers are also derived for the month of Aug 2100. We observe the similar fact that “DROP\_TERMINATE\_VOLTE\_IMS” does not have much correlation with other call incidence parameters. Similar behavior is observed for all the other bands also. The color coding carries the same meaning as described. The detailed correlation coefficient numbers of all the other months along with the month of August 2018 are described in the embedded excel sheet attached in the appendix section.



Figure 3: correlation matrix of band 1900 cells for the month of Aug 2018



Figure 4: Correlation matrix of band 700 cells for the month of Aug 2018

It is observed that “DROP\_TERMINATE\_VOLTE\_IMS” does not have much correlation with other parameters across all the considered bands, we have removed this variable from our study and decided to perform the dimension reduction technique “Principal Component Analysis” on the following list of call incidence parameters:

|  |
| --- |
| VARIABLES |
| ACCESS\_FAILURE\_CC\_503(1:223)\_VOLTE\_IMS |
| CALL\_SETUP\_DELAY\_VOLTE\_IMS |
| DROP\_CC\_408\_VOLTE\_IMS |
| DROP\_CC\_481\_VOLTE\_IMS |
| DROP\_CC\_503\_VOLTE\_IMS |
| DROP\_CC\_OTHER\_VOLTE\_IMS |
| GARBLING\_VOLTE\_VOICE |
| IMMEDIATE\_VOLUNTARY\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE |
| MUTING\_VOLTE\_VOICE |
| SOFT\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_IMS |
| CALL\_COUNT |
| EST\_COUNT\_HANDOVER |

Table 1::List of call incidence parameters considered for PC

From the above list of variables, “CALL\_COUNT” and “EST\_COUNT\_HANDOVER” are the variables that have a significant correlation with each other across all the bands and similar behavior is also observed across all the months. These two variables are quite similar by nature. It is a known fact that if more number of calls are passed through a cell, then more number of call handover incidences are likely to happen. Hence these two parameters can be used to normalize the core call incidence parameters. By choice, we used “CALL\_COUNT” variable to normalize all the other call incidence parameters and ignore the “EST\_CALL\_HANDOVER” variable. The normalized variables are defined as follows.

Normalized (X)= Sum of the incidences for the variable (X) / sum of the CALL\_COUNT

Where X is the any call incidence parameter apart from “CALL-COUNT” and “EST\_CALL\_HANDOVER”

The sum of “DROP\_CC\_481\_VOLTE\_IMS” parameter and “DROP\_CC\_503\_VOLTE\_IMS” parameter is defined to be a new parameter named as “HARD-DROP\_CC\_VOLTE-IMS”. Hence, by summing up these two call incidence parameter, we have defined the new parameter.

Consequently, we choose the following 9 variables in its’ normalized state to perform PCA analysis and rank the cells based on the derived PCA score.

|  |
| --- |
| VARIABLES |
| ACCESS\_FAILURE\_CC\_503(1:223)\_VOLTE\_IMS |
| CALL\_SETUP\_DELAY\_VOLTE\_IMS |
| DROP\_CC\_408\_VOLTE\_IMS |
| DROP\_CC\_OTHER\_VOLTE\_IMS |
| GARBLING\_VOLTE\_VOICE |
| IMMEDIATE\_VOLUNTARY\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE |
| MUTING\_VOLTE\_VOICE |
| SOFT\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_IMS |
| HARD\_DROP\_CC\_VOLTE\_IMS |

Table 2: List of normalized variables used for PCA

### Defining Ranking Window:

Data of all incidence parameters are collected on daily basis for each cell level. This is the minimum time granularity data is available for this study. Within a day how the cell performances vary is not captured whereas business demands to rank or score the cells on daily basis. As a result, historical behavior of the cells up to certain historical time period is to be considered to rank them i.e. the current ranking of a cell would depend on its own historical performance and relative performance of the other cells. For daily ranking, how many days to consider is to be determined analytically. For this analysis, we have considered 1 to 15 days of rolling window and for each chosen rolling time window we have determined the standard deviation of the call incidence parameters. As the number of days are increasing to calculate the moving average, the standard deviation of the call incidence parameter is decreasing.

For band 2100, we have observed the decreasing standard deviation with increasing number of days of moving averages.

Figure 5: For band 2100 decreasing standard deviation of access failure for increasing rolling moving average

In the graph above, x axis represents the number of days for defining moving average and y axis represents the measured standard deviation of all the cells within the band 2100. We have observed the break-even point at 7 or 8 days rolling moving average. It emphasizes even for the increasing number of historical days for rolling moving average, the overall standard deviation of “Access failure “is getting stabilized around 7-8 days.

For all the other incidence parameters, we have observed similar pattern. In the following figures it is illustrated that similar pattern is observed for all the call incidence parameters.

Figure 6: For band 2100 decreasing standard deviation of garbling for increasing rolling moving average

Same behavior is observed for “Garbling” i.e. standard deviation is getting stabilized for rolling moving average of 7-8 days.

Figure 7: For band 2100 decreasing standard deviation of muting for increasing rolling moving average

Figure 8: For band 2100 decreasing standard deviation of soft drop for increasing rolling moving average

Figure 9: For band 2100 decreasing standard deviation of call count for increasing rolling moving average

Figure 10: For band 2100, decreasing standard deviation of call hand over for increasing rolling moving average

Figure 11: For band 2100 decreasing standard deviation of hard drop for increasing rolling moving average

It is decided in this study that we will consider 7 days of rolling moving window to rank the cells. For the other bands of cells, we have observed similar behavior for all the call incidence parameters. The detail is embedded in an excel sheet in the appendix.

# Ranking Methodology & Approach:

To rank the cells, we have adopted two different approaches.

* Approach1: Principal Component Analysis
* Approach2: Autoencoder

Using the correlation analysis, we adopt the dimension reduction technique “Principal Component Analysis”. A comparative study based on unsupervised neural network based dimension reduction technique “Autoencoder” also implemented in this study. In the following subsections we have described both the approaches to describe the cell ranking procedure.

## PCA Approach

Principal components analysis is one of the simplest multivariate dimension reduction methods. The objective of this technique is to take p variables () which are call incidences in our study; and find linear combinations of those to produce indices () that are uncorrelated in order of their importance, and that best describe the variation in the data. As described in section 2.1.1 in correlation analysis, we have observed that 9 call incidence variables are highly correlated with each other. We use those variables to get the top principal components that capture majority of variance in the raw data.

The principal components with Eigen value more than equal to 1 are usually captures larger proportion of variability and are considered only for ranking the cells. The detail math behind this technique is briefly described in the Appendix section.

Once the top principal components (Eigen value >=1) are derived, we have the Z values of each cell. These Z values are multiplied by a weight which is the proportion of variance captured by the individual components. The summed up value of these weight multipliers is the score of a cell.

For example, the score of the *ith* cell is described as follows:

Where,

are the Z values of the ith cell considering K number of components having Eigen value more than equal to 1

values are the weights of the components which are the individual captured variances

Once we get the score of all the cells, we ranked them as per the obtained scored value. Higher the score of a cell would receive lower rank. High score of a cell means that the cell has exhibited more number of call incidences in its normalized form and it physically defined as bad performance of a cell.

## PCA Result & Validation

For illustration purpose, we ranked the cells on few days in the month of January 2019. In the following table we described the Eigen values of the principal components when the cells are ranked for few days in the month of January 2019

|  |  |  |
| --- | --- | --- |
| Date | Eigen values | Weights |
| Jan 13th, 2019 | ([2.56120171, 1.76781547, 1.04898184, 0.96431293, 0.91430946, 0.83421584, 0.66453685, 0.15525484, 0.09346754]) | ([0.29213777, 0.18177305, 0.12373641, 0.10693235, 0.10562497,0.09336365, 0.0706242 , 0.0152733 , 0.0105343 ]) |
| Jan 14th, 2019 | ([2.68338312, 1.80359006, 1.0959253, 0.98359786, 0.91287219, 0.81498877, 0.47145332, 0.14891445, 0.08937143]) | ([0.28859215, 0.18528832, 0.12397474, 0.10799208, 0.10499487,0.09458555, 0.06904465, 0.01475701, 0.01077063]) |
| Jan 15th, 2019 | ([2.72103399, 1.78999849, 1.10471642, 0.9935853, 0.91389443, 0.81705229, 0.42114328, 0.15006653, 0.09260577]) | [(0.29236346, 0.18454164, 0.12334509, 0.10831636, 0.10516932, 0.09287914, 0.06922328, 0.01366119, 0.01050053]) |
| Jan 16th, 2019 | ([2.69919073, 1.76702291, 1.15797827, 1.002595, 0.92139068,0.86492285,0.3438426 , 0.15241928, 0.09473417]) | ([0.29777019, 0.18682227, 0.12324721, 0.10690092, 0.10188966, 0.09542039, 0.06515162, 0.01312732, 0.00967043]) |

Table 3: Eigen values and proportion of variability captured by the principal components when cells are ranked for few consecutive days in Jan 2019 for band 1900 cells

We can observe that when the band 1900 cells are ranked on 13th January, 2019, first three components have Eigen value more than equal to 1 and cumulative way they capture more than 60% variability of the call incidences for this band. To score the cells on 13th January, 2019, we have considered 0.29, 0.18 and 0.12 as the associated weights and these weights are multiplied with the corresponding Z values to derive the score of a cell. Similarly, we can also observe that on January 16th, 2019, four principal components and their weights would be considered as the Eigen value of the top four components exceed the numeric value 1. These top four components capture more than 70% variability of the raw call incidence parameters in its normalized form when the cells of 1900 bands are ranked on 16th January, 2019.

For the other band of cells, similar table depicting the Eigen values and the proportion of captured variability is presented in the Appendix section.

The ranking snapshot of top 10 and bottom 10 cells on January 13th, 2019 are depicted in the following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cell\_Name | Band | Score | Date | Rank |
| BNYB0301A11 | band1900\_1 | -0.48618 | 1/13/2019 | 1 |
| BBQ04139A11 | band1900\_1 | -0.47811 | 1/13/2019 | 2 |
| BBQ04181B21 | band1900\_1 | -0.4771 | 1/13/2019 | 3 |
| BBQ04271A31 | band1900\_1 | -0.46751 | 1/13/2019 | 4 |
| BBQ04289B21 | band1900\_1 | -0.46706 | 1/13/2019 | 5 |
| BBK04025A21 | band1900\_1 | -0.46692 | 1/13/2019 | 6 |
| BBQ04855B11 | band1900\_1 | -0.46192 | 1/13/2019 | 7 |
| BNYY8119A11 | band1900\_1 | -0.45928 | 1/13/2019 | 8 |
| BNYY8322A21 | band1900\_1 | -0.45916 | 1/13/2019 | 9 |
| BBQ06083B21 | band1900\_1 | -0.45748 | 1/13/2019 | 10 |
| BBQM6022A11 | band1900\_1 | 4.163921 | 1/13/2019 | 2189 |
| BBX03007A31 | band1900\_1 | 4.510133 | 1/13/2019 | 2190 |
| BNY01639A21 | band1900\_1 | 4.643222 | 1/13/2019 | 2191 |
| BNY02796A31 | band1900\_1 | 5.121074 | 1/13/2019 | 2192 |
| BNY01996A11 | band1900\_1 | 5.13928 | 1/13/2019 | 2193 |
| BBQY0208A11 | band1900\_1 | 5.539576 | 1/13/2019 | 2194 |
| BNY01639A41 | band1900\_1 | 5.744874 | 1/13/2019 | 2195 |
| BBQY0207A21 | band1900\_1 | 6.618061 | 1/13/2019 | 2196 |
| BBQY0205A11 | band1900\_1 | 7.035408 | 1/13/2019 | 2197 |
| BNY01724A21 | band1900\_1 | 7.04665 | 1/13/2019 | 2198 |

Table 4: Top 10 and Bottom 10 cells ranked on Jan 13th, 2019 for band 1900

The PCA scores and ranks of the all cells across all the three bands for consecutive 4 days (Jan 13th- Jan 16th) are illustrated in the embedded excel sheet in the Appendix section.

To compare the good and bad cells as per the derived score and the ranks, a comparative analysis of the normalized form of cell incidence parameters are described below table. In this analysis we calculate the average value of the normalized call incidence parameters for top 10 good cells and bottom 10 bad cells. We the ratio of bad and good cells average values of all the parameters and observed that bad cells have recorded higher number of the call incidences than the good cells. For example, when we compare for band 1900, we have observed that Garbling is 97 time more in bad cells in comparison to good cells whereas average muting also 175 times more is being recorded in bad cells. The detail comparison of good and bad cells of 1900 bands for the other parameters are described in the below table. In the appendix section we have described the similar comparison for other bands also.

|  |  |  |  |
| --- | --- | --- | --- |
| Band 1900 | | | |
| Call Incidence Parameters | Good cells | Bad Cells | Ratio(Bad/Good) |
| ACCESS\_FAILURE\_CC\_503(1:223)\_VOLTE\_IMS | 0.000215 | 0.031194 | 145.24 |
| CALL\_SETUP\_DELAY\_VOLTE\_IMS | 0.002326 | 0.021128 | 9.08 |
| DROP\_CC\_408\_VOLTE\_IMS | 0.000000 | 0.002774 | - |
| DROP\_CC\_OTHER\_VOLTE\_IMS | 0.000181 | 0.000121 | 0.67 |
| GARBLING\_VOLTE\_VOICE | 0.020830 | 2.027047 | 97.31 |
| IMMEDIATE\_VOLUNTARY\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE | 0.000000 | 0.000252 | - |
| MUTING\_VOLTE\_VOICE | 0.001289 | 0.226284 | 175.50 |
| SOFT\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_IMS | 0.002648 | 0.008860 | 3.35 |
| hard\_drop\_call\_503\_481 | 0.008424 | 0.014672 | 1.74 |

Table 5: Comparison of normalized call incidence parameters for top 10 good cells and bottom 10 bad cells of band 1900

To get a comprehensive understanding of the best performing cells and the worst performing cells we prepared the mean score for the cells with respect to the whole period i.e. from 01 sept. 2018 till 28 Feb 2019 for the 1900 band . We arranged these values in the following table displaying the top 10 and the bottom 10 cells .

|  |  |  |
| --- | --- | --- |
| Rank | cellname | score |
| 1 | BBK04025A21 | -0.48 |
| 2 | BBQ06174B31 | -0.48 |
| 3 | BBK04011A11 | -0.48 |
| 4 | BBQ04289B21 | -0.48 |
| 5 | BBQ04288A11 | -0.47 |
| 6 | BBQ04770D31 | -0.46 |
| 7 | BBQ06174B11 | -0.46 |
| 8 | BBQ06111E31 | -0.46 |
| 9 | BBQ04796C11 | -0.46 |
| 10 | BBQ04254A21 | -0.46 |
| 2191 | BNY01724A21 | 3.25 |
| 2192 | BBQ06007B31 | 3.31 |
| 2193 | BNY01639A41 | 3.32 |
| 2194 | BNY01996A11 | 3.57 |
| 2195 | BNY02796A31 | 4.06 |
| 2196 | BNY01639A21 | 4.39 |
| 2197 | BBQY0207A21 | 5.12 |
| 2198 | BBQY0208A11 | 5.81 |
| 2199 | BBQY0205A11 | 7.28 |

### Score comparison using rank percentile

The ranking procedure described in this approach produces daily score of a cell based on its performance i.e recorded normalized call incidences of last 7 days. The score of cell may vary on day to day basis as the historical performance of the cells also likely to vary. To find a benchmark based on the score of good and bad cells, we have done the percentile analysis of the scores.

For doing this analysis, we consider more than three months of data from 01st September, 2018 to 28th February, 2019. We score the cells on daily basis for continuous 301 days of time span. Considering all the derived PCA scores of all the cells, the various percentile values starting from 0th to 100th percentiles are calculated. In the following figure, we have presented the score vs percentile graph for band 1900 cells. From this figure, we can derive for a given score what would be percentile value.

Figure 12:PCA score and percentile values for band 1900 cells.

From this percentile value, we can sense the relative performance of the cell. The data of various percentile values for the PCA scores across other bands are presented in the Appendix section. The minimum value and maximum value of the scores are considered as the lower cap and upper cap thresholds. If a new score of a cell in a day lies beyond this range, then it would be capped to the thresholds.

### Summary & Conclusion

* There is significant correlation among the call incidence parameters except “”. The call incidence parameter values do depend on the total number of call counts and call hand over parameters. Hence, the incidence parameters are normalized with the help of call count variables to have a comparison of cell performances. These normalized values are used to perform principal component analysis. The idea of using PCA is to combine the various call incidences with some weights to transform them into a single score.
* Analytically, it is estimated that ranking of cell would majorly depends on last seven days’ historical performance.
* Only the top principal components that captures significant proportion of variability of the raw data are considered for cell ranking.
* The derived PCA scores of good and bad cells are validated with the actual call incidences. Cells with good ranks are observed with significant lower normalized incidences in comparison to the cells with bad ranks.

## Auto encoder Approach

In the approach we have taken we make a very reasonable assumption that, most cells do not function abnormally most of the time. What we mean by abnormal is, those cells that need corrective action. Thus we can use historical data to build a model of the normal behavior of the cells. We have decided to use Auto-encoders to model the behavior of the cells.

Auto-encoders are an unsupervised learning technique where we make neural networks are used for learning probability distributions. The neural network architecture is designed such that information bottleneck in the network results in a compressed knowledge representation of the original input. A compressed representation of the input forces the network to learn patterns present in the data.

As visualized above, we can take an unlabeled dataset and frame it as a supervised learning problem tasked with taking as input and outputting  a reconstruction of the original input . This network can be trained by minimizing the *reconstruction error*,  , which measures the differences between our original input and the consequent reconstruction. The bottleneck is a key attribute of our network design; without the presence of an information bottleneck, our network could easily learn to simply memorize the input values by passing these values along through the network.

We have used Auto-encoders because, the use of non-linear activation functions and the flexibility of choosing the number of hidden layers and nodes help us model non-linear relationships between the variables. We use the incident data observed at a cell level to train the auto-encoder. We have built 3 different models for the three different bands that we are considering. After the model has been trained

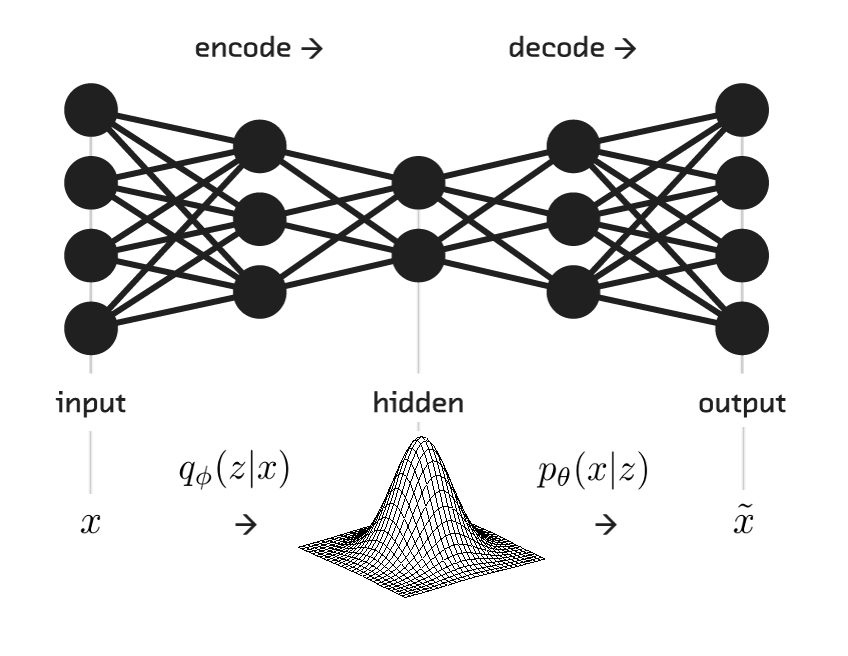


Figure 13:

### Reconstruction error

After the auto-encoder model has been trained with the historical incident data occurring at every cell (in our case the data is at a day level), when presented with any new incident data captured at a cell, it attempts to reconstruct this data based on patterns it has observed in the historical data. For example, if we present a record of a cell with muting, garbling, dropping values represented by the vectors then it is mapped to a lower dimensional vector and then from that it attempts to reconstruct the input to get the values . As the vector is of lower dimension it is forced to identify relationships present in the data. The difference between the input and reconstructed input gives us the reconstruction error.The larger the reconstruction error, more is the chances that this input does not conform to the relationships or patterns learnt from the historical data, i.e., this input may indicate abnormal behavior. Below is presented the reconstruction errors for the three Bands.

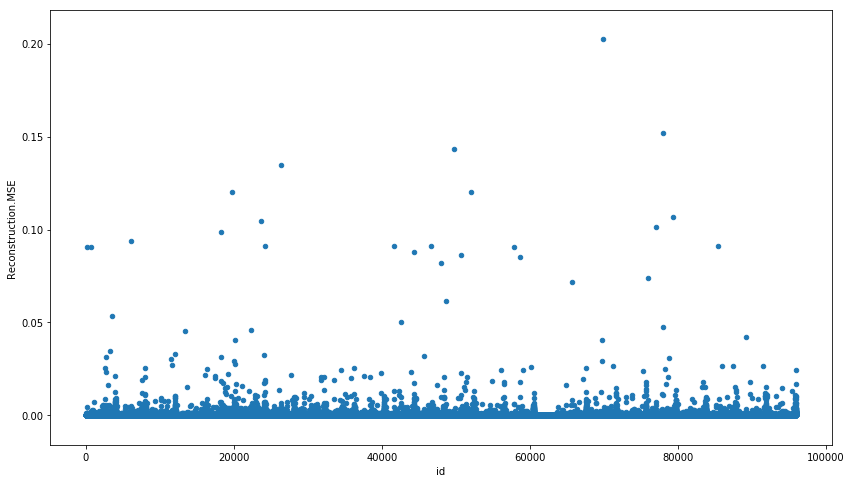


Fig 2.2

Reconstruction error for Band 2100

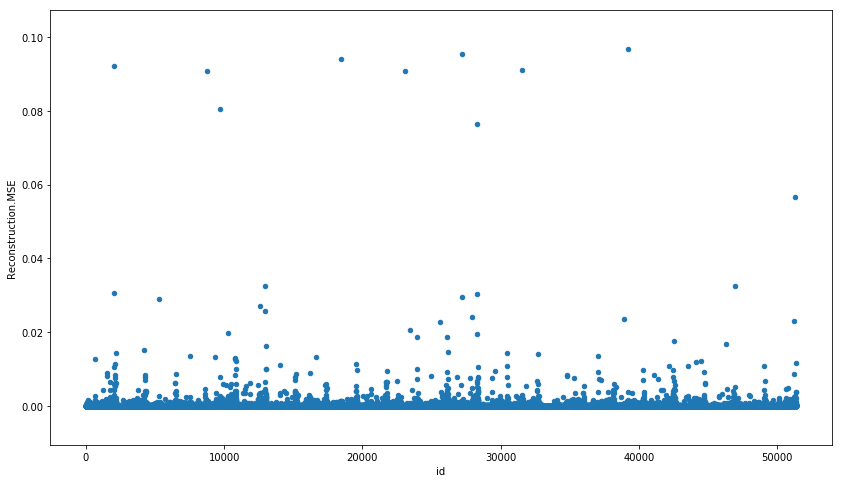


Fig 2.3

Reconstruction error for Band 1900\_1

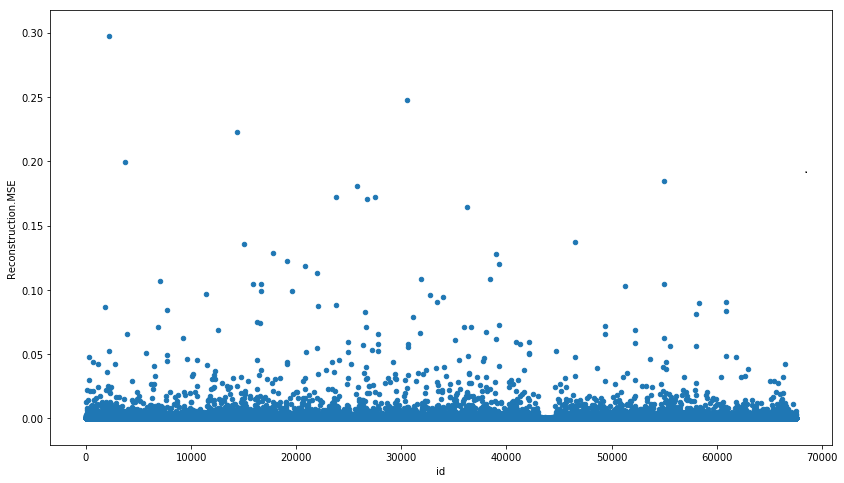


Fig 2.4

Reconstruction error for Band 700

### Mapping reconstruction error to probability score

The reconstruction errors serve as a measure of how deviated the behavior of the cell is from normal behavior, but it becomes difficult to compare the errors of cells across bands. Hence propose to map the reconstruction error to the probability that this cell is not coming from the probability distribution of the normally functioning cells. This is done by applying the empirical distribution function of the reconstructions errors for any given value of the reconstruction error. Thus for every incident recording of a cell, we get a corresponding reconstruction error and for that reconstruction error we get a corresponding probability of this error coming from an anomalous or abnormal cell, i.e.

gives the probability that a cell is abnormal when the reconstruction Error for that cell is

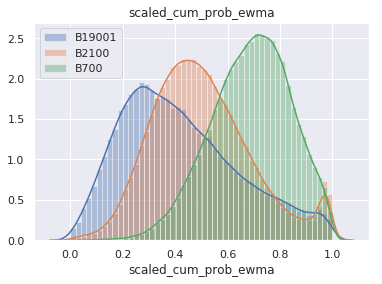
This is illustrated in Fig 2.4 below. Here we have the empirical distribution of the errors and two records of errors for two cells shown by the orange and green line. The cell depicted by the green line has less reconstruction error and thus lesser probability of having been observed in an anomalous/abnormally functioning cell. The cell depicted by the orange line has more reconstruction error and has a higher probability of having been observed in an abnormal cell.

Thus by mapping the reconstruction error to a probability score, we can now compare these scores across bands. gives the……??? We can now say which cell irrespective of the band it belongs to has a higher probability of needing our attention (i.e., behaving abnormally) and we can now rank the cells based on this probability scores.

We have decided to scale the reconstruction errors (for each bands separately) using min-max scaling to limit the range of values that can be taken and thus have more points falling into the bins used for getting the empirical distribution.

### Exponential weighted moving average of probability scores

It can be seen that some cells behave abnormally for a short time and then get back to normal, such cells require lesser attention than the ones which have been behaving abnormally for a sustain periods of time. Hence we have used exponential weighted average of the probability scores for the past 7 days, to arrive at the final score for ranking/comparing cells. The reason for using this approach is, more weight is given to the recent measurements than the ones in the past.



### Result & Validation

Two different Auto-encoder architectures were experimented with. One with the architecture (10, 7, 5, 7,10), 10 input nodes, followed 7, 5, 7 hidden nodes and then 10 output nodes. Then with the architecture (10,7,10). Tanh has been used as the activation function for the Auto-encoders. This was implemented using the H2O library in Python.

As input to the Auto-encoders, the normalized values of the indicators have only been used, i.e., the input to the models were:

|  |
| --- |
| **Variables** |
| MUTING\_VOLTE\_VOICE\_norm |
| GARBLING\_VOLTE\_VOICE\_norm |
| ACCESS\_FAILURE\_CC\_503.1.223.\_VOLTE\_IMS\_norm |
| DROP\_CC\_481\_VOLTE\_IMS\_norm |
| CALL\_SETUP\_DELAY\_VOLTE\_IMS\_norm |
| SOFT\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_IMS\_norm |
| DROP\_CC\_503\_VOLTE\_IMS\_norm |
| DROP\_CC\_408\_VOLTE\_IMS\_norm |
| DROP\_CC\_OTHER\_VOLTE\_IMS\_norm |
| IMMEDIATE\_VOLUNTARY\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_norm |
| DROP\_SRVCC\_TERMINATE\_VOLTE\_IMS\_norm |

To verify if the approach we have taken to rank cells, we have performed three sets of tests.

1. we picked up a few cells ranked highly as abnormal and have checked what are the different incident rates for these cells. We also checked what are the incident rates for those cells ranked lowly. Below in Figures. 3.1 to 3.6 we have taken the distribution of these incidents for all the cells (can be seen by the histogram in green), the incident rates for the highly ranked cells (marked by the dashed red lines) and the incident rates for the lowly ranked cells (indicated by the green dashed lines). we can see that those sites rated highly (having higher chances of abnormality) have higher incident rates compared to the incident rates normally seen. We can also see that the green lines indicating the lowly ranked cells are always well within the normal range. From these figures we can see that cells ranked have highly has one or more incident types having higher than normal values, but the cells ranked lowly never have any incident types taking higher than normal values.
2. we divided the cells into two categories, those that are normal and those that are critical. The critical cells are the top 5% of the highly ranked(abnormal) cells and the rest are considered normal. We check the distribution of the different types of incidents. We can notice that the distribution for the critical cells clearly have a higher mean and more extreme values for different types of incidents, this can be observed in the figures 3.7 to 3.10.
3. We picked the top 20 highly ranked cells and obtained their mean values for all the normalized values for different incident types, this was compared to the top 20 lowly ranked cells and obtained their mean values for all the normalized values for different incident types. We can observe that normalized incident values for the highly ranked or termed Bad cells are significantly higher than lowly ranked or Good cells.

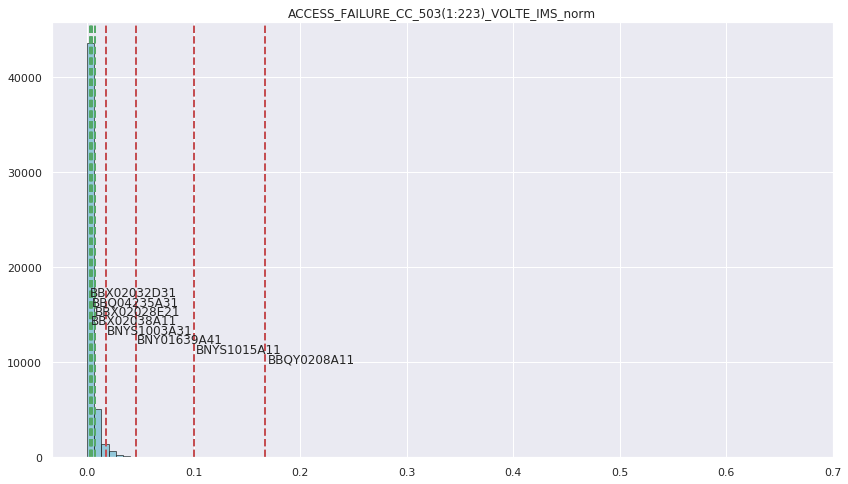


Fig 3.1

Access Failure incidents for highly and lowly ranked cells for Band 1900\_1

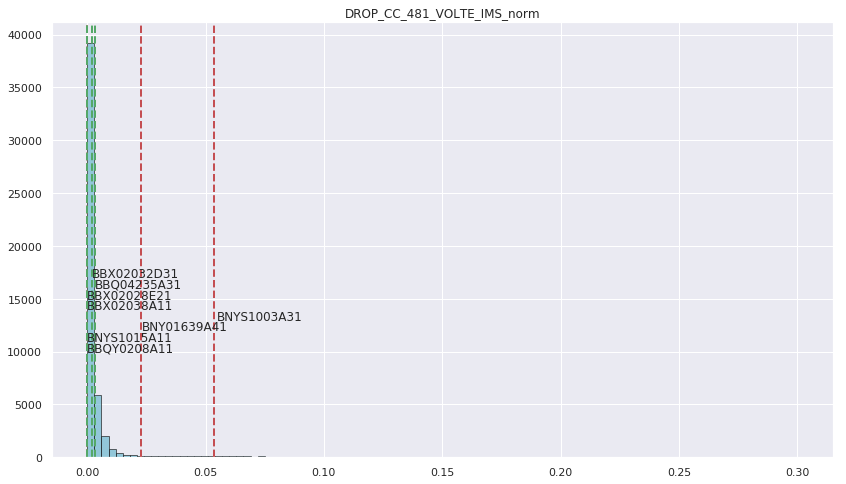


Fig 3.2

Drop call incidents for highly and lowly ranked cells for Band 1900\_1

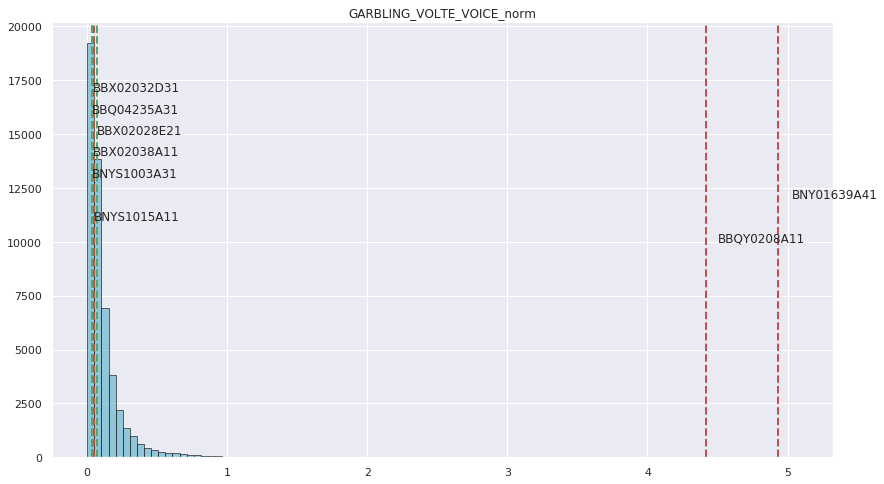


Fig 3.3

Garbling incidents for highly and lowly ranked cells for Band 1900\_1

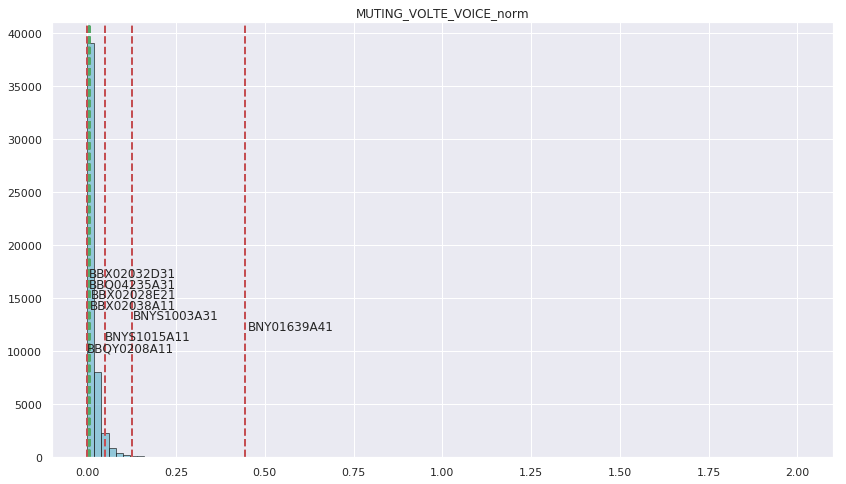


Fig 3.4

Muting incidents for highly and lowly ranked cells for Band 1900\_1

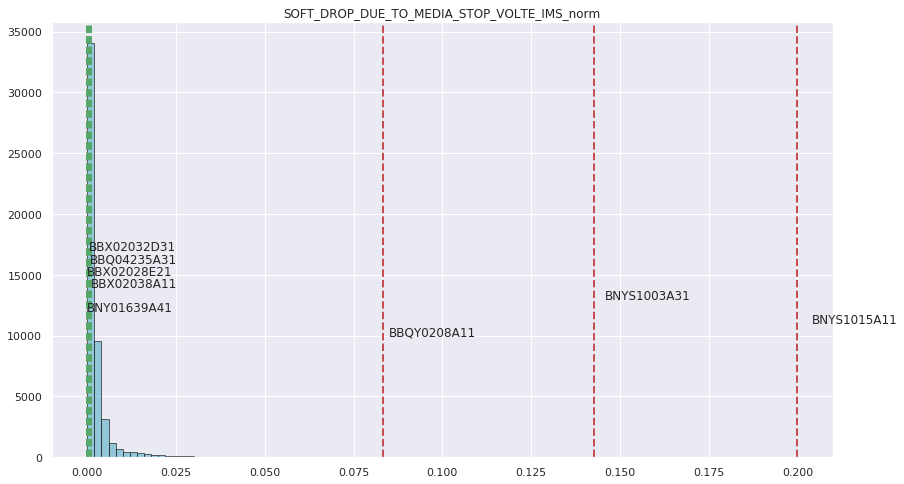


Fig 3.5

Soft Drop incidents for highly and lowly ranked cells for Band 1900\_1

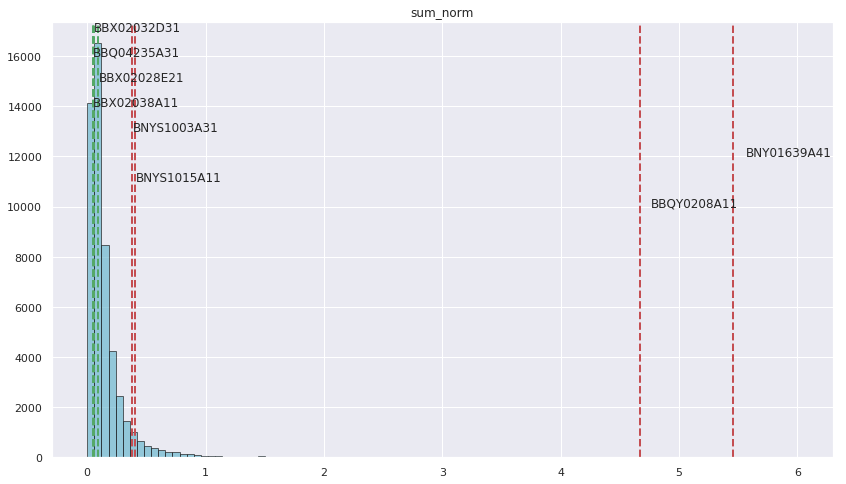


Fig 3.6

Sum\_Nom for highly and lowly ranked cells for Band 1900\_1

Sum\_norm was not used for training the model, but we can observe that highly ranked cells have higher sum\_norm. sum\_norm is the sum of all the normalized incident values for a given cell and day.

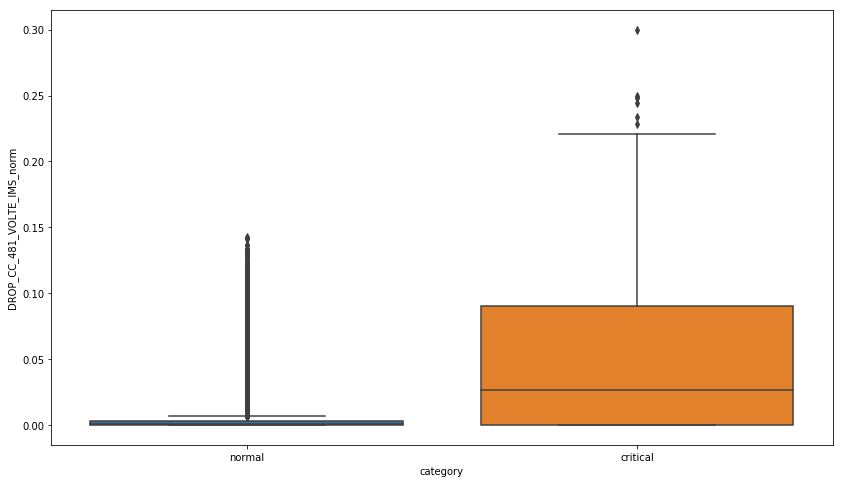


Fig 3.7

Drop incidents distribution for normal and critical cells for Band 1900\_1

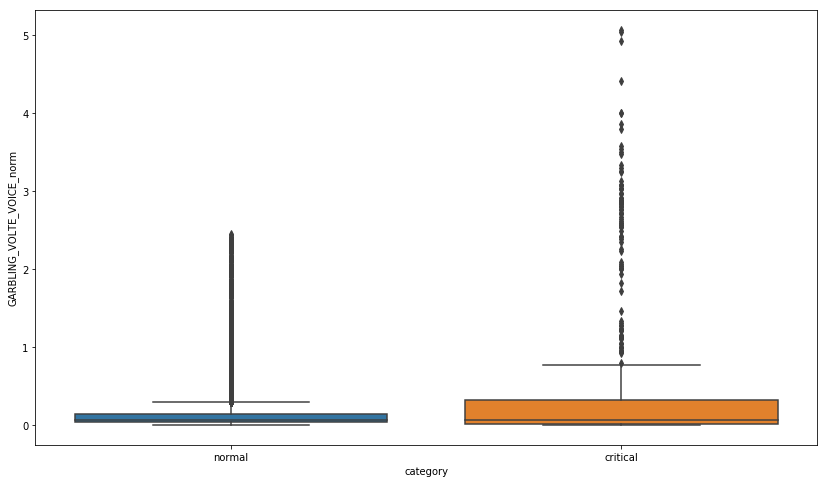


Fig 3.8

Garbling incidents distribution for normal and critical cells for Band 1900\_1

# C:\Projects\EEA TMO cell\Results_24May\boxplot_b19001_muting.png

Fig 3.9

Muting incidents distribution for normal and critical cells for Band 1900\_1

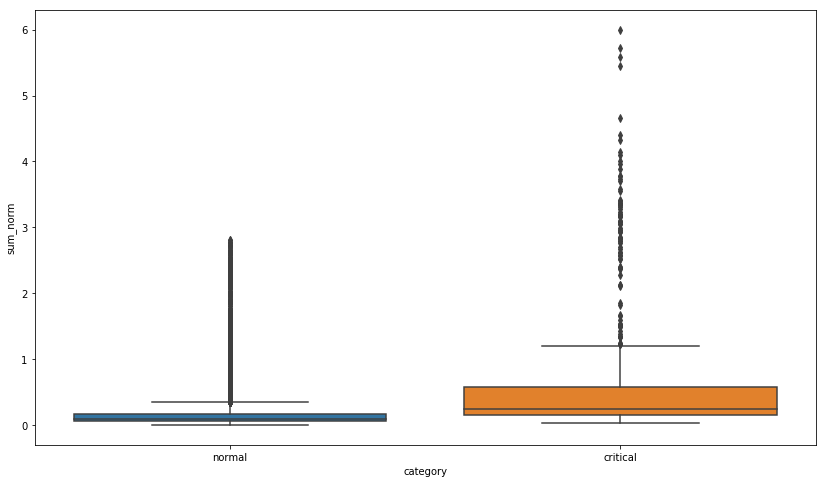


Fig 3.10

Sum\_norm distribution for normal and critical cells for Band 1900\_1

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Mean Values** | **Mean Values** | **Ratio of Bad to Good cells** |
|  | **Good Cells** | **Bad cells** |  |
| **band** | 1900\_1 | 1900\_1 |  |
| **date** | Jan-19 | Jan-19 |  |
| **GARBLING\_VOLTE\_VOICE\_norm** | 0.0233904 | 0.22318785 | 9.541862415 |
| **MUTING\_VOLTE\_VOICE\_norm** | 0.0027082 | 0.01639588 | 6.054058909 |
| **DROP\_CC\_481\_VOLTE\_IMS\_norm** | 0.0002974 | 0.05789732 | 194.6683453 |
| **CALL\_SETUP\_DELAY\_VOLTE\_IMS\_norm** | 0.0041992 | 0.00420408 | 1.001167936 |
| **ACCESS\_FAILURE\_CC\_503(1:223)\_VOLTE\_IMS\_norm** | 0.0013672 | 0.00952175 | 6.964499038 |
| **SOFT\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_IMS\_norm** | 0.0006229 | 0.01836464 | 29.48137023 |
| **DROP\_CC\_503\_VOLTE\_IMS\_norm** | 0.0003166 | 0.0019952 | 6.30207926 |
| **DROP\_CC\_408\_VOLTE\_IMS\_norm** | 4.038E-05 | 0.00019647 | 4.865909525 |
| **DROP\_CC\_OTHER\_VOLTE\_IMS\_norm** | 4.805E-05 | 0.0003283 | 6.831946788 |
| **IMMEDIATE\_VOLUNTARY\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_norm** | 3.339E-05 | 6.8776E-05 | 2.059700872 |
| **scaled\_cum\_prob\_ewma** | 0.1228627 | 0.99369139 | 8.087818263 |

Table 3.1

Sum\_norm distribution for normal and critical cells for Band 1900\_1

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Mean Values** | **Mean Values** | **Ratio of Bad to Good cells** |
|  | **Good Cells** | **Bad Cells** |  |
| **band** | B2100 | B2100 |  |
| **date** | Jan-19 | Jan-19 |  |
| **GARBLING\_VOLTE\_VOICE\_norm** | 0.018264117 | 0.23164876 | 12.683272 |
| **MUTING\_VOLTE\_VOICE\_norm** | 0.002286207 | 0.022067285 | 9.6523549 |
| **DROP\_CC\_481\_VOLTE\_IMS\_norm** | 0.000252127 | 0.064070417 | 254.11983 |
| **CALL\_SETUP\_DELAY\_VOLTE\_IMS\_norm** | 0.004286528 | 0.007643998 | 1.7832608 |
| **ACCESS\_FAILURE\_CC\_503(1:223)\_VOLTE\_IMS\_norm** | 0.001635242 | 0.025145773 | 15.377403 |
| **SOFT\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_IMS\_norm** | 0.000492974 | 0.015216456 | 30.866677 |
| **DROP\_CC\_503\_VOLTE\_IMS\_norm** | 0.000268387 | 0.001052199 | 3.9204602 |
| **DROP\_CC\_408\_VOLTE\_IMS\_norm** | 1.10405E-05 | 0 | 0 |
| **DROP\_CC\_OTHER\_VOLTE\_IMS\_norm** | 2.54039E-05 | 0.000171999 | 6.7705477 |
| **IMMEDIATE\_VOLUNTARY\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_norm** | 1.39811E-05 | 0.000180245 | 12.892057 |
| **DROP\_SRVCC\_TERMINATE\_VOLTE\_IMS\_norm** | 0 | 0 |  |
| **scaled\_cum\_prob\_ewma** | 0.136051888 | 0.993860344 | 7.3050096 |

Table 3.2

Sum\_norm distribution for normal and critical cells for Band 2100

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Mean Values** | **Mean Values** | **Ratio of Bad to Good** |
|  | **Good Cells** | **Bad cells** |  |
| **band** | B700 | B700 |  |
| **date** | Jan-19 | Jan-19 |  |
| **GARBLING\_VOLTE\_VOICE\_norm** | 0.2171965 | 1.3971174 | 6.43250252 |
| **MUTING\_VOLTE\_VOICE\_norm** | 0.0318301 | 0.2500258 | 7.855005157 |
| **DROP\_CC\_481\_VOLTE\_IMS\_norm** | 0.0036771 | 0.066309 | 18.03280666 |
| **CALL\_SETUP\_DELAY\_VOLTE\_IMS\_norm** | 0.0045359 | 0.0062155 | 1.370286935 |
| **ACCESS\_FAILURE\_CC\_503(1:223)\_VOLTE\_IMS\_norm** | 0.004224 | 0.0229389 | 5.430656334 |
| **SOFT\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_IMS\_norm** | 0.0030129 | 0.0399327 | 13.25393342 |
| **DROP\_CC\_503\_VOLTE\_IMS\_norm** | 0.0018462 | 0.0119872 | 6.492930372 |
| **DROP\_CC\_408\_VOLTE\_IMS\_norm** | 8.821E-05 | 0.0022552 | 25.56492287 |
| **DROP\_CC\_OTHER\_VOLTE\_IMS\_norm** | 0.000113 | 0.0003717 | 3.288670484 |
| **IMMEDIATE\_VOLUNTARY\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_norm** | 1.588E-05 | 0.0003981 | 25.06692442 |
| **DROP\_SRVCC\_TERMINATE\_VOLTE\_IMS\_norm** | 0 | 0 |  |
| **scaled\_cum\_prob\_ewma** | 0.2928825 | 0.9879721 | 3.373271557 |

Table 3.3

Sum\_norm distribution for normal and critical cells for Band 2100

### Summary & Conclusion

From the incident distributions we observed that the occurrences of high incident values are very low in number and most observations are concentrated in the region of low incident values. This confirms our assumption that most cells do not function abnormally most of the time.

In the three sets of tests that we conducted we observed that the higher reconstruction errors (and the mapped cumulative probability) was associated with higher incident values for one or generally more than one incident types. This was seen across all the three bands. We were thus able to come up with a scoring method which combines the observed incident values for a cell to provide a score indicating how much it is deviated from normal behavior.

This score is also comparable across the bands because the score is the probability of the cell having deviated from normal behavior for that band.

# Appendix A

## Notebook

The Jupyter Notebook used for this report can be found at this [location](https://ericsson.sharepoint.com/:u:/r/sites/EEA_Data_Science/Shared%20Documents/Misc/DataCleanTutorialNotebook.html?csf=1&e=nd64oe).

## Field Definitions

|  |  |
| --- | --- |
| **Variables** | **Description** |
| ACCESS\_FAILURE\_CC\_503.1.223.\_VOLTE\_IMS |  |
| CALL\_SETUP\_DELAY\_VOLTE\_IMS |  |
| DROP\_CC\_408\_VOLTE\_IMS |  |
| DROP\_CC\_481\_VOLTE\_IMS |  |
| DROP\_CC\_503\_VOLTE\_IMS |  |
| DROP\_CC\_OTHER\_VOLTE\_IMS |  |
| DROP\_SRVCC\_TERMINATE\_VOLTE\_IMS |  |
| GARBLING\_VOLTE\_VOICE |  |
| IMMEDIATE\_VOLUNTARY\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE |  |
| MUTING\_VOLTE\_VOICE |  |
| SOFT\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_IMS |  |
| Sum | Sum of all the incidences |
| call\_count | Total number of call counts |
| est\_count\_handover | Total number of calls handover |
| sum\_drop\_cc\_503\_481 | Total number of hard drops |
| GARBLING\_VOLTE\_VOICE\_norm | Normalized value of Garbling over the total number of call counts |
| MUTING\_VOLTE\_VOICE\_norm | Normalized value of muting over the total number of call counts |
| ACCESS\_FAILURE\_CC\_503.1.223.\_VOLTE\_IMS\_norm |  |
| DROP\_CC\_481\_VOLTE\_IMS\_norm |  |
| CALL\_SETUP\_DELAY\_VOLTE\_IMS\_norm |  |
| SOFT\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_IMS\_norm |  |
| DROP\_CC\_503\_VOLTE\_IMS\_norm |  |
| DROP\_CC\_408\_VOLTE\_IMS\_norm |  |
| DROP\_CC\_OTHER\_VOLTE\_IMS\_norm |  |
| IMMEDIATE\_VOLUNTARY\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_norm |  |
| DROP\_SRVCC\_TERMINATE\_VOLTE\_IMS\_norm |  |
| sum\_norm |  |
| Band | Bands on which the cell is configured |
| start\_date | Date on which the data is collected |

Table 6: data field definitions of cell incidence parameters

## Month wise record detail

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data sets** | **# Days band 2100** | **# Days**  **band 1900** | **# Days band 700** | **# Records** | **Comments/Missing days** |
| incidents\_IncidentId\_Aug2018 | 31 | 31 | 31 | 308,883 | No missing days |
| incidents\_IncidentId\_Sep2018 | 29 | 29 | 29 | 287,352 | 8th September missing across all bands including band "0" and band "1900\_2" |
| incidents\_IncidentId\_Oct2018 | 28 | 28 | 28 | 287,654 | 29th, 30th and 31st Oct data is missing across all the bands including "0" and "1900\_2" |
| incidents\_IncidentId\_Nov2018 | 29 | 29 | 29 | 306,692 | 28th Nov data is missing across all the bands including "0" and "1900\_2" |
| incidents\_IncidentId\_Dec2018 | 30 | 30 | 30 | 298,499 | 22nd December data is missing across all the bands including "0" and "1900\_2" |
| incidents\_IncidentId\_Jan2019 | 25 | 25 | 25 | 235,078 | 17th to 22nd Jan data is missing across all bands including "0" and "1900\_2" |
| incidents\_IncidentId\_Feb2019 | 16 | 16 | 16 | 158,520 | 9th, 10th,17th,20th…… all data is missing |

Table 7: Summary of records for all the months across the various bands

## Correlation analysis of the incidence parameters across months

The detail correlation of all the call incidences across the months are presented in the embedded excel sheet.



## Rolling window to determine

The detail of the ranking window determination as described in section 2.1.2 is embedded in the excel sheet



## Brief math behind PCA approach

Principal components analysis is one of the simplest multivariate methods. The objective of the technique is to take p variables () which are call incidences in the study; and find linear combinations of these to produce indices () that are uncorrelated in order of their importance, and that describe the variation in the data. The lack of correlation means that the indices are measuring different “dimensions” of the data, and the ordering is such that , where denotes the variance of the ith component . When doing principal components analysis, there is always the hope that the variances of most of the indices will be as low as to be negligible. In that case, most of the variation in the call incidence records can be adequately described by the few Z components with variances that are not negligible, and some degree of economy is then achieved. Often the significant variances explained by the Z variables have a dominant load factor associated with the original actual X variables and Z describe a degree of quantitative or qualitative nature of the X attributes.

Principal components analysis does not always work, in the sense that many original variables are reduced to a small number of transformed variables. Indeed, if the original variables are uncorrelated, then the analysis achieves nothing. The best results are obtained when the original variables are very highly correlated, positively, or negatively. If that is the case, then it is quite conceivable that call incidence records are also highly correlated and they can be summarized by three to four principal components that are uncorrelated by nature and captures most of the variance in the raw data. The important principal components will be of some interest as measures of the underlying dimensions in the data. It will also be of value to know that there is a good deal of redundancy in the original call incidence variables, with most of them measuring similar things.

The first principal component is then the linear combination of the raw variables call incidences and represented as

that varies as much as possible for the individuals, subject to the condition that

Thus , the variance of , is as large as possible given this constraint on the constants . The constraint is introduced because if this is not done, then can be increased by simply increasing any one of the values. Similarly, we can form the 2nd, 3rd and so on the p-th component in such a way so that is the second largest variance, is the 3rd largest variance and so on.

Subject to the constraint

Subject to the constraint

Subject to the constraint

As described in section 2.1.1 in correlation analysis, we observe that 9 variables are highly correlated with each other. We use those variables to get the principal components.

## Significant components for scoring band 2100 and band 700

|  |  |  |
| --- | --- | --- |
| Date | Eigen values | Weights |
| Jan 13th, 2019 | ([2.62988894, 1.63636128, 1.1139026 , 0.96262873, 0.95085938,0.84048025, 0.63577473, 0.13749362, 0.09483215]) | ([0.29213777, 0.18177305, 0.12373641, 0.10693235, 0.10562497,0.09336365, 0.0706242 , 0.0152733 , 0.0105343 ]) |
| Jan 14th, 2019 | ([2.59797049, 1.66800654, 1.11604812, 0.97216865, 0.94518707,0.85148009, 0.62155522, 0.13284592, 0.09695957]) | ([0.28859215, 0.18528832, 0.12397474, 0.10799208, 0.10499487,0.09458555, 0.06904465, 0.01475701, 0.01077063]) |
| Jan 15th, 2019 | ([2.63192072, 1.66128475, 1.11037981, 0.97508784, 0.94675753,0.83611857, 0.62316327, 0.12298106, 0.09452813]) | ([0.29236346, 0.18454164, 0.12334509, 0.10831636, 0.10516932,  0.09287914, 0.06922328, 0.01366119, 0.01050053]) |
| Jan 16th, 2019 | ([2.68059324, 1.68181546, 1.1094987 , 0.96234579, 0.91723331,0.85899547, 0.58650935, 0.11817505, 0.08705531]) | ([0.29777019, 0.18682227, 0.12324721, 0.10690092, 0.10188966,0.09542039, 0.06515162, 0.01312732, 0.00967043]) |

Table 8: Eigen values and proportion of variability captured by the principal components when cells are ranked for few consecutive days in Jan 2019 for band 2100 cells

|  |  |  |
| --- | --- | --- |
| Date | Eigen values | Weights |
| Jan 13th, 2019 | ([3.38663178, 1.1463876 , 1.04669449, 1.01733114, 0.89996914,0.79299553, 0.39492024, 0.22487204, 0.09332738]) | ([0.29213777, 0.18177305, 0.12373641, 0.10693235, 0.10562497, 0.09336365, 0.0706242 , 0.0152733 , 0.0105343 ]) |
| Jan 14th, 2019 | ([3.3644861 , 1.14697601, 1.06176884, 0.97169794, 0.94200898,  0.80329533, 0.39597278, 0.22337252, 0.09355085]) | ([0.28859215, 0.18528832, 0.12397474, 0.10799208, 0.10499487,  0.09458555, 0.06904465, 0.01475701, 0.01077063]) |
| Jan 15th, 2019 | ([3.39870208, 1.17680906, 1.03826993, 0.95273939, 0.95129804,  0.78629826, 0.3870451 , 0.21866398, 0.0933035 ]) | ([0.29236346, 0.18454164, 0.12334509, 0.10831636, 0.10516932,  0.09287914, 0.06922328, 0.01366119, 0.01050053]) |
| Jan 16th, 2019 | ([3.37116318, 1.20129997, 1.02888171, 0.95735141, 0.9323223 ,  0.79191781, 0.39354947, 0.2370992 , 0.0895443 ]) | ([0.29777019, 0.18682227, 0.12324721, 0.10690092, 0.10188966,  0.09542039, 0.06515162, 0.01312732, 0.00967043]) |

Table 9: Eigen values and proportion of variability captured by the principal components when cells are ranked for few consecutive days in Jan 2019 for band 700 cells

## Ranking of the cells



## Comparison of good and bad cells for band 2100 and band 700 cells

|  |  |  |  |
| --- | --- | --- | --- |
| Band 700 | | | |
| Call Incidence Parameters | Good cells | Bad Cells | Ratio(Bad/Good) |
| ACCESS\_FAILURE\_CC\_503(1:223)\_VOLTE\_IMS | 0.000540 | 0.041198 | 76.34 |
| CALL\_SETUP\_DELAY\_VOLTE\_IMS | 0.000673 | 0.008531 | 12.67 |
| DROP\_CC\_408\_VOLTE\_IMS | 0.000000 | 0.009723 | - |
| DROP\_CC\_OTHER\_VOLTE\_IMS | 0.000000 | 0.003509 | - |
| GARBLING\_VOLTE\_VOICE | 0.094435 | 1.730126 | 18.32 |
| IMMEDIATE\_VOLUNTARY\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE | 0.000000 | 0.000724 | - |
| MUTING\_VOLTE\_VOICE | 0.017252 | 0.282225 | 16.36 |
| SOFT\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_IMS | 0.000628 | 0.027629 | 43.97 |
| hard\_drop\_call\_503\_481 | 0.003793 | 0.054744 | 14.43 |

Table 10: Comparison of normalized call incidence parameters for top 10 good and bottom 10 bad cells of band 700

|  |  |  |  |
| --- | --- | --- | --- |
| Band 2100 | | | |
| Call Incidence Parameters | Good cells | Bad cells | Ratio (Bad/Good) |
| ACCESS\_FAILURE\_CC\_503(1:223)\_VOLTE\_IMS | 0.001031 | 0.030226 | 29.31 |
| CALL\_SETUP\_DELAY\_VOLTE\_IMS | 0.003528 | 0.007060 | 2.00 |
| DROP\_CC\_408\_VOLTE\_IMS | 0.000249 | 0.000429 | 1.72 |
| DROP\_CC\_OTHER\_VOLTE\_IMS | 0.000990 | 0.000000 | 0.00 |
| GARBLING\_VOLTE\_VOICE | 0.013632 | 0.978682 | 71.79 |
| IMMEDIATE\_VOLUNTARY\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE | 0.000026 | 0.000167 | 6.45 |
| MUTING\_VOLTE\_VOICE | 0.001631 | 0.113238 | 69.44 |
| SOFT\_DROP\_DUE\_TO\_MEDIA\_STOP\_VOLTE\_IMS | 0.000359 | 0.006980 | 19.42 |
| hard\_drop\_call\_503\_481 | 0.000552 | 0.006631 | 12.02 |
|  |  |  |  |

## 

## Rank percentile

We created the PCA scores for all the cells fo 2100 band in the period 01st September, 2018 to 28th February, 2019 by applying the algorithm. We have plotted the percentiles of these scores in the below table :

|  |  |
| --- | --- |
| percentile | score |
| 0 | 38.07717 |
| 1 | 1.929078 |
| 2 | 1.258483 |
| 3 | 0.947161 |
| 4 | 0.738766 |
| 5 | 0.600308 |
| 10 | 0.287817 |
| 15 | 0.169871 |
| 20 | 0.101552 |
| 25 | 0.053361 |
| 30 | 0.015909 |
| 35 | -0.01522 |
| 40 | -0.04241 |
| 45 | -0.06716 |
| 50 | -0.09026 |
| 55 | -0.11275 |
| 60 | -0.13503 |
| 65 | -0.15752 |
| 70 | -0.18142 |
| 75 | -0.20736 |
| 80 | -0.23662 |
| 85 | -0.27089 |
| 90 | -0.31331 |
| 95 | -0.37609 |
| 96 | -0.3944 |
| 97 | -0.41789 |
| 98 | -0.45037 |
| 99 | -0.50598 |
| 99.1 | -0.50986 |
| 99.2 | -0.51886 |
| 99.3 | -0.52889 |
| 99.4 | -0.54089 |
| 99.5 | -0.55544 |
| 99.6 | -0.57116 |
| 99.7 | -0.59158 |
| 99.8 | -0.62277 |
| 99.9 | -0.67064 |
| 100 | -0.8186 |

*Table 11 : Comparison of various ranks and percentiles associated for the 2100 band*

For doing the analysis, we consider the data from 01st September, 2018 to 28th February, 2019. We score the cells on daily basis for continuous 301 days of time span. Considering all the derived PCA scores of all the cells, the various percentile values starting from 0th to 100th percentiles are calculated. In the following figure, we have presented the score vs percentile graph for band 2100 cells. From this figure, we can derive for a given score what would be percentile value.

*Fig 4.1 The comparison of the values of scores and their corresponding variables for the 2100 band*

We created the PCA scores for all the cells fo 2100 band in the period 01st September, 2018 to 28th February, 2019 by applying the algorithm. We have plotted the percentiles of these scores in the below table :

|  |  |
| --- | --- |
| percentile | score |
| 100 | -1.13 |
| 99.9 | -1.05 |
| 99.8 | -1.01 |
| 99.7 | -0.99 |
| 99.6 | -0.97 |
| 99.5 | -0.95 |
| 99.4 | -0.94 |
| 99.3 | -0.92 |
| 99.2 | -0.91 |
| 99.1 | -0.90 |
| 99 | -0.90 |
| 98 | -0.83 |
| 97 | -0.79 |
| 96 | -0.75 |
| 95 | -0.72 |
| 90 | -0.62 |
| 85 | -0.55 |
| 80 | -0.49 |
| 75 | -0.43 |
| 70 | -0.38 |
| 65 | -0.33 |
| 60 | -0.28 |
| 55 | -0.23 |
| 50 | -0.17 |
| 45 | -0.11 |
| 40 | -0.05 |
| 35 | 0.02 |
| 30 | 0.11 |
| 25 | 0.20 |
| 20 | 0.33 |
| 15 | 0.49 |
| 10 | 0.74 |
| 5 | 1.25 |
| 4 | 1.45 |
| 3 | 1.75 |
| 2 | 2.18 |
| 1 | 2.95 |
| 0 | 19.54 |

*Table 12 : percentiles of the average scores of cells for the 700 band*

For doing the analysis, we consider the data from 01st September, 2018 to 28th February, 2019. We score the cells on daily basis for continuous 301 days of time span. Considering all the derived PCA scores of all the cells, the various percentile values starting from 0th to 100th percentiles are calculated. In the following figure, we have presented the score vs percentile graph for band 700 cells. From this figure, we can derive for a given score what would be percentile value.

*Fig 4.1 The comparison of the values of scores and their corresponding variables for the 700 band*

From the aforementioned graphs and tables it can be easily observed what the scores of various percentiles of cells would be for the 2100 band and 700 band. Conclusions can be drawn from them regarding the nature of cells in the network.

## Acronyms

ARPU Average Revenue Per User

DPA Daily Profile Attributes

EEA Ericsson Expert Analytics

ESR Extended Session Record

IMEI International Mobile Equipment Identity

IM Instant Messages

IMSI International Mobile Subscriber Identity

QoE Quality of Experience

SLI Service Level Index

TAC Type Allocation Code