Relevant Alarms Detection

Relevant Alarm Detection Using Machine Learning for Alarm Reduction Systems

Dataset

The dataset used in this study has been collected from the ACE Fault Management platform. We are utilizing historical data to analyze and develop a machine learning model capable of predicting whether an alarm is relevant or irrelevant.

Irrelevant alarms are those considered to be transient alarms.

According to the **Fault Management Documentation**, transient (irrelevant) alarms can be defined as follows:

- **Maintenance Mode**: Alarms triggered when systems are in maintenance mode.
- **Low Priority Alarms**: Alarms classified with low priority levels.
- **Repetitive Alarms**: Alarms of the same type from the same Source or Site that occur more than a specified number of times within a defined period (e.g., more than three times in 24 hours) should be labeled as irrelevant.
- **Maintenance Period Alarms**: Alarms that occur during a scheduled maintenance window should be labeled as irrelevant.

Since our dataset does not come pre-labeled with **relevant** and **irrelevant** alarms, we need to employ a heuristic to label the dataset.

We will label an alarm as irrelevant if it is cleared within a short period of time, denoted as n. The value of n should ideally be chosen by a domain expert. For the purpose of this study, we will use $n = 7$ minutes.

The dataset contains **32,326 rows** representing alarm events and **28 columns** representing various features.

Features in the Dataset:

- Severity
- Status
- Alarm Name
- Technical ID
- Site Name
- Site ID
- Region
- Zone
- NE Name
- Source
- Cell Name
- NE Type
- Vendor
- Technology
- Location Information
- First Occurrence
- Last Occurrence
- Event Time
- CSN
- Clearance Time
- Clearance User
- Acknowledgement Time
- Acknowledgement User
- FM Receive Time
- Is Active
- Ticket ID
- Alarm Category
- Parent Node
- Target

After labeling the alarms as relevant or irrelevant according to the previously outlined rule, an additional column **Target** was added with binary values **relevant** or **irrelevant**.

Columns **Status**, **Cell Name**, and **Ticket ID** contain no data and will be removed from the study as they provide no information. Similarly, the **Vendor** column contains only one value, "Huawei," and will be excluded for lack of variability.

The columns **Region** and **Zone** contain only "R0" and "Z0," respectively. Since they do not offer useful information beyond these constant values, **Region** will be dropped from the study.

Data Exploration

Datetime Features Analysis

Analysis of the **First Occurrence** and **Last Occurrence** columns revealed alarms spanning an unrealistic range from **January 1, 1990**, to **February 4, 2037**. Further investigation uncovered three events with a **First Occurrence** in the year 2037, deemed outliers, and subsequently removed from the dataset.

CSN	First Occurence	Severity	Alarm Name	NE Type
326763515	2037-04-02 03:08:24	Critical	ETH_LOS	Optix RTN 950
326685958	2037-04-02 03:07:28	Cleared	SWDL_INPROCESS	Optix RTN 950
326763279	2037-04-02 03:05:00	Warning	SWDL_INPROCESS	Optix RTN 950

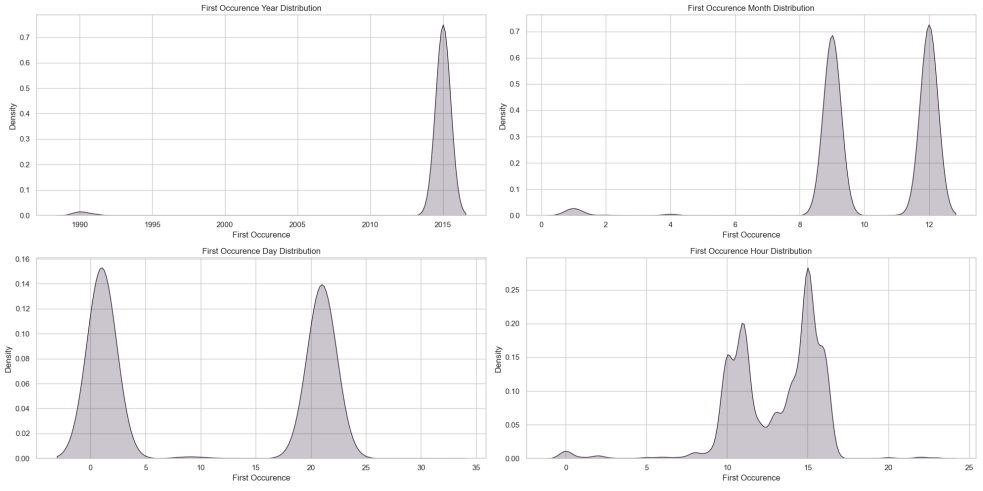
Further exploration revealed missing values in the following columns:

Columns	Missing values
Clearance Time	10456
Clearance User	10456
Acknoledgement Time	32299
Acknoledgement User	32299

Notably, the missing values in **Clearance Time** corresponded with identical missing values in **Clearance User**, and similarly for the **Acknowledgement** columns.

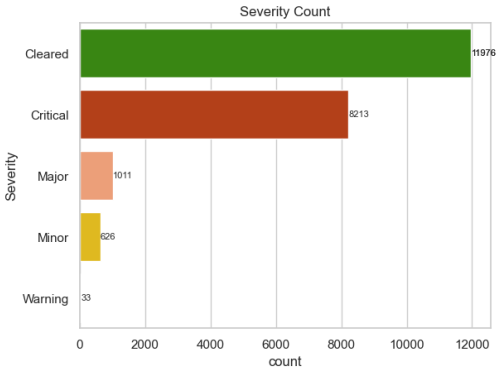
Additionally, there were 25 rows where both **Acknowledgement Time** and **Clearance Time** were null simultaneously. To maintain data integrity and facilitate labeling, rows with null values in **Clearance Time** were dropped.

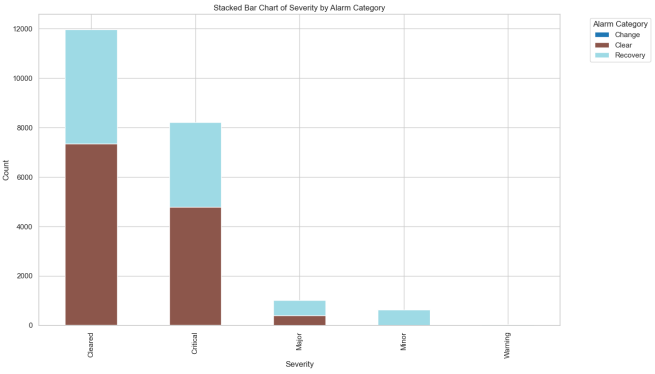
First Occurence Distributions



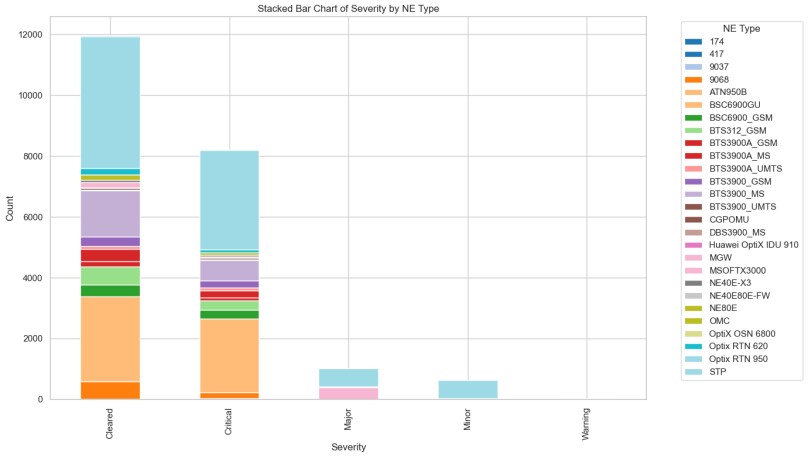
By examining this plot, we can conclude that most alarms first occurred in **2015**, with notable peaks in **September** and **December**. Specifically, the first occurrences are concentrated on the **1st**, **2nd**, **20th**, and **21st** days of these months, particularly between **10:00 AM** and **4:00 PM**.

Categorical Features analysis

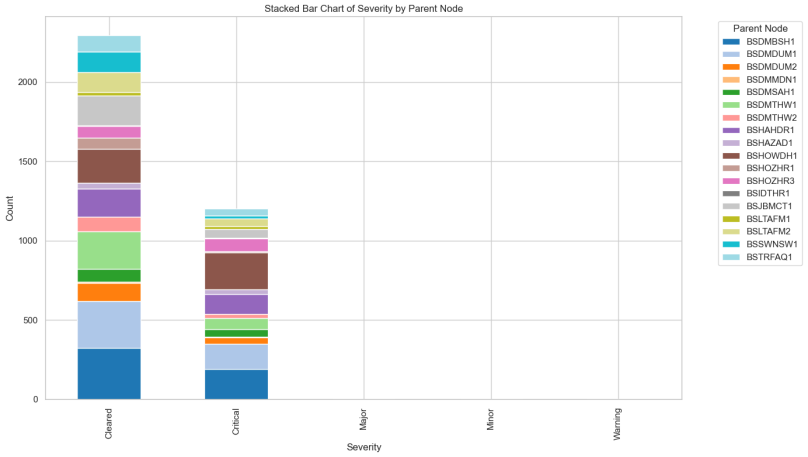




NE Type	Count
Optix RTN 950	8819
BSC6900GU	5212
BTS3900_MS	2176
BTS312_GSM	897
9068	781
BSC6900_GSM	675



Parent Node	Count
BSDMBSH1	513
BSDMDUM1	456
BSHOWDH1	447
BSDMTHW1	304
BSHAHDR1	303
BSJBMCT1	241



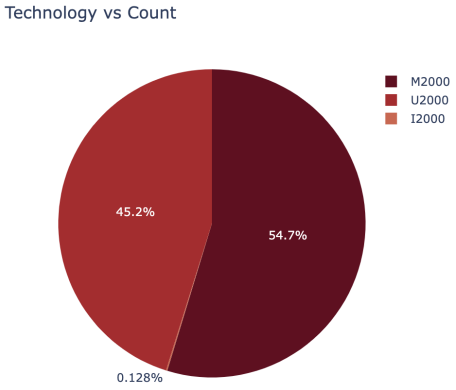


Figure - Technology distribution

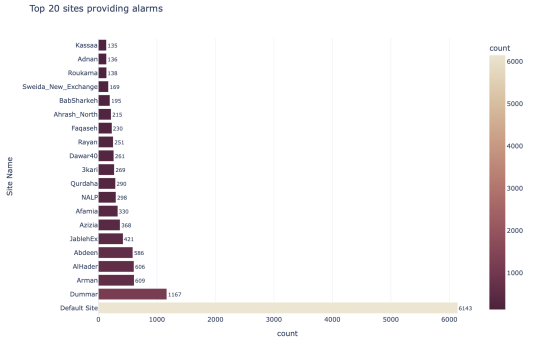


Figure - Top 20 sites providing alarms

In this figure, it's evident that the site with the highest number of alarms is labeled as the "Default Site". However, it's essential to discuss the nature of the sites and whether they are fixed categories in this column.

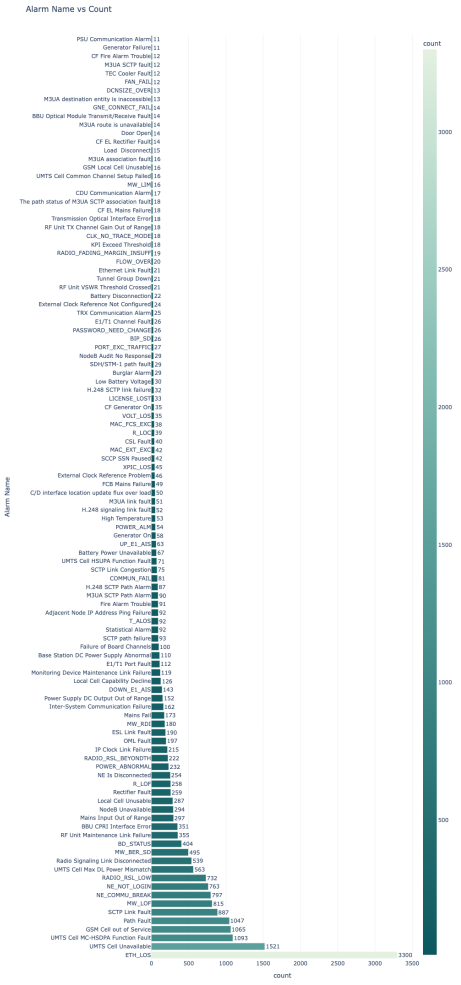


Figure - Alarm Names distribution

Feature Engineering

In this section, we aim to derive additional insights from the original features (columns) and create new features accordingly.

An attempt was made to calculate the time taken to detect alarms by computing the difference between the Event Time and First Occurrence columns. However, this resulted in a value of 0 for each event, indicating that the system detects alarms in real-time. Consequently, the **Event Time** column will be dropped from the study as it duplicates information already present in the **First Occurrence** column.

Furthermore, we propose calculating the **Clearance Duration Time** as a feature to aid in labeling alarm events as **relevant** or **irrelevant**. As previously discussed, if the Clearance Duration Time for an event is less than $n = 7$ *minutes*, it will be labeled as **irrelevant**; otherwise, it will be labeled as **relevant**.

The **Clearance Duration Time** is calculated by determining the difference between the **Clearance Time** and **First Occurrence** columns.

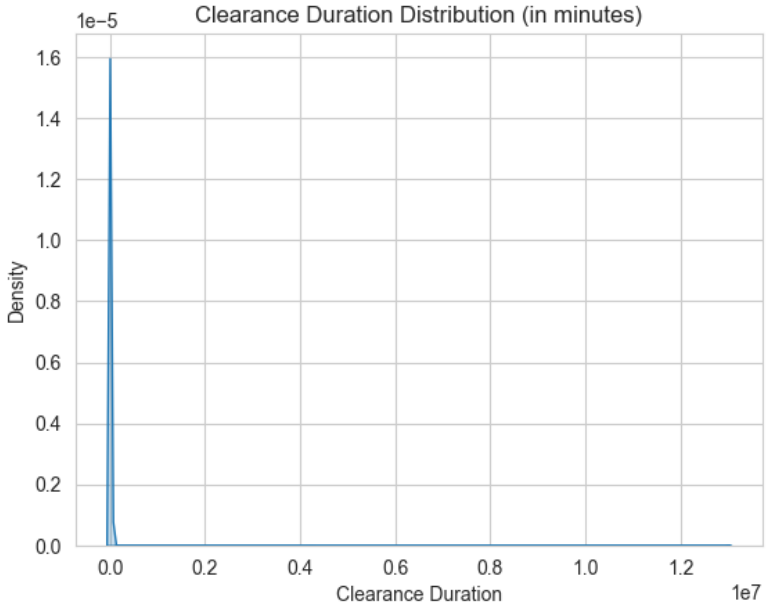
During this calculation, anomalies were identified, with six events exhibiting a negative **Clearance Duration Time**. This indicates that these alarms were cleared before they occurred, a scenario that is logically impossible.

CSN	Severity	First Occurence	Clearance Time	Alarm Name	NE Type	Alarm Category
326764282	Major	1990-01-01 04:02:18	1990-01-01 02:04:15	POWER_ABNORMAL	Optix RTN 950	Recovery
326764281	Major	1990-01-01 04:02:18	1990-01-01 02:04:15	POWER_ABNORMAL	Optix RTN 950	Recovery
215729941	Critical	1990-01-01 00:05:07	1990-01-01 00:03:58	ETH_LOS	Optix RTN 950	Recovery
215718959	Critical	1990-01-01 00:25:07	1990-01-01 00:04:13	ETH_LOS	Optix RTN 950	Recovery
215717587	Critical	1990-01-01 05:15:26	1990-01-01 00:04:59	BUS_ERR	Optix RTN 950	Recovery
215716711	Major	1990-01-01 05:05:23	1990-01-01 00:04:59	BD_STATUS	Optix RTN 950	Recovery

To ensure data integrity, these anomalous events will be investigated further and potentially removed from the dataset.

From the plot, it's evident that the Clearance Duration varies between 0 minutes and 1.3×10^7 minutes *(approximately 24.7 years)*

Clearance User	count
<SYSTEM>	21863
saurabh	1



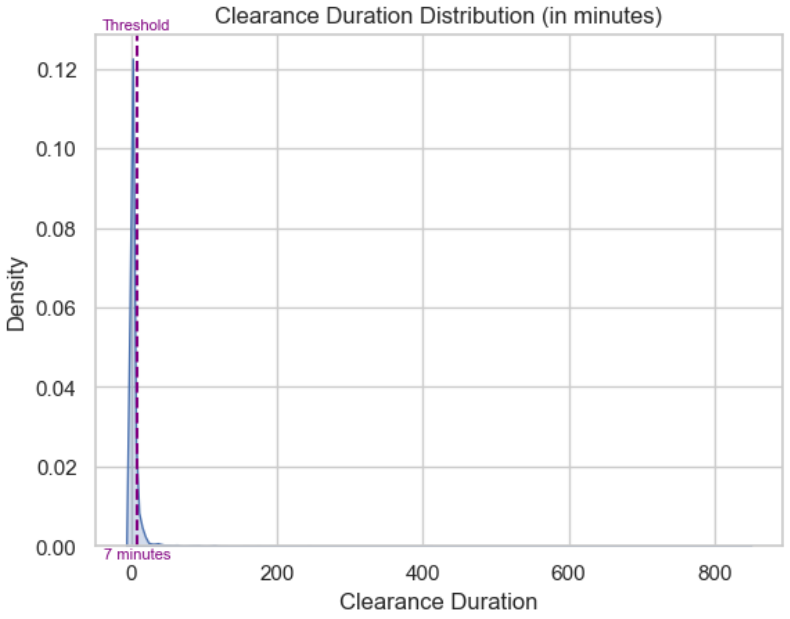
The plot visually confirms the presence of outliers in the dataset.

CSN	Severity	First Ocurrence	Clearance Duration (years)	NE Type	Alarm Category
136003519	Critical	2015-12-01 16:34:14	**7.994**	NE Is Disconnected	OMC
215718927	Major	1991-01-14 22:51:15	**24.683**	POWER_ABNORMAL	Optix RTN 950
215718921	Major	1991-01-14 22:51:06	**24.683**	PASSWORD_NEED_CHANGE	Optix RTN 950
215718920	Minor	1991-01-14 22:55:21	**24.683**	CLK_NO_TRACE_MODE	Optix RTN 950
215716520	Major	1991-01-14 22:42:35	**24.683**	POWER_ABNORMAL	Optix RTN 950

These outliers will be removed from the study to avoid biasing the data, which will lead to better results and more accurate predictions when using machine learning models.

Statistics

events	21859
mean	3.690100
std	17.995587
min	0
25%	0
50%	0
75%	0.816667
max	842.9333

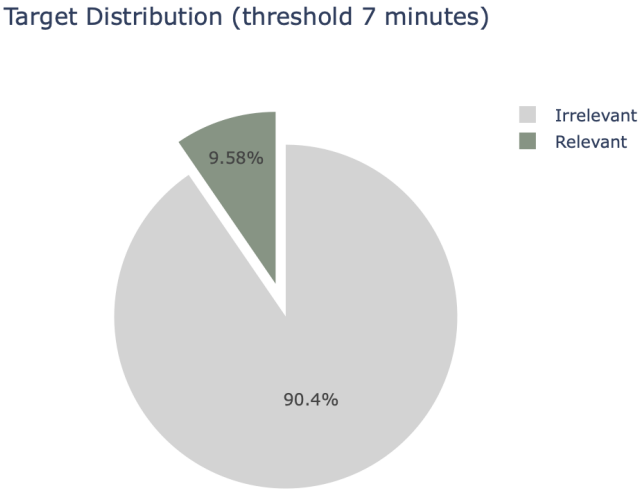


After removing the outliers, the Clearance Duration now ranges between 0 minutes and 800 minutes (13.3 hours), which is more logical.

Using a threshold of 7 minutes to label our dataset, we have the following distribution:

Relevant	2093
Irrelevant	19764

This dataset is clearly imbalanced, making it challenging for machine learning models to accurately detect irrelevant alarms.



To address this imbalance, we need to collect more "Relevant" alarms. If that is not feasible, we can employ advanced algorithms to generate synthetic "Relevant" alarms. *(This approach will be discussed in the following sections.)*

Modeling Approach