

Cat Checker – Elimination Classification

Pet care and food space company

Spring Board Project

Topics/Agenda

01 | Objective

02 | Data Preparation

03 | Data exploration and Hypothesis Validation

04 | Feature Engineering

05 | Train model to predict the elimination and non-elimination activity

06 | Train model to predict the urination and defecation

07 | Business Insights

08 | Conclusion



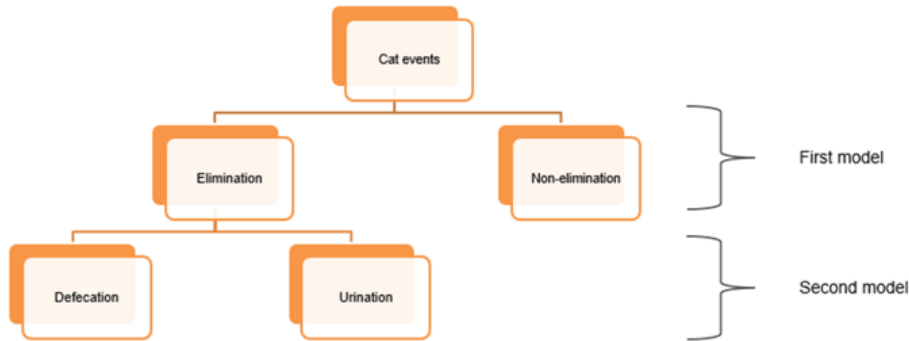
01

Objective

01. Objective

Smart Litter Sytem

objective of this project is to develop a solution approach to correctly classify the events as elimination or non-elimination and further classify the elimination events into urination and defecation.



Pet Care and food space

Relation between

- Cat Activity vs Senson signals
- Cat Activity vs time
- Cat Activity Vs Cat Power use
- Cat weight Vs Activity
- Type of Cat vs Activitv



02. Data Preparation

- a) Decisions on data selection
- b) Decisions on merging the data, joining key
- c) Decisions on the target variable

Is Provided data is sufficient ?

Elimination is depending on cat behavior:

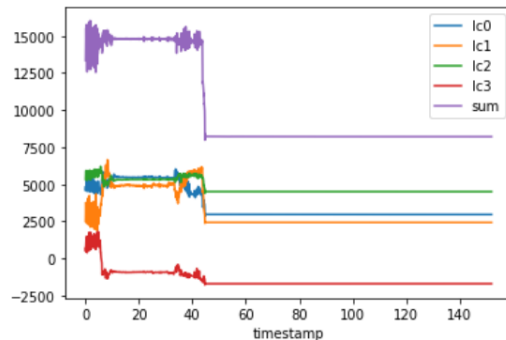
- 1) Cat Activity/Behavior
- 2) Duration of Activity
- 3) Cat power/energy used by cat
- 4) Cat Type
- 5) Cat Weight
- 6) Load sensor data(lc0 , lc1, lc1 , lc3)
- 7) Time & Season

02 – Data Prepration

Convert CSV File to Data frame of each load sensor reading

Duration of Activity

	t	lc0	lc1	lc2	lc3	timestamp	sum	filename
0	0 days 00:00:00	4814	2504	5337	701	0.000	13356	Raw_data\VA000000000000017\RXZlbnRJbmZvcn1hdGlv...
1	0 days 00:00:00.025000	4656	2789	5481	497	0.025	13423	Raw_data\VA000000000000017\RXZlbnRJbmZvcn1hdGlv...
2	0 days 00:00:00.050000	4566	3137	5583	483	0.050	13769	Raw_data\VA000000000000017\RXZlbnRJbmZvcn1hdGlv...
3	0 days 00:00:00.075000	4640	3449	5724	628	0.075	14441	Raw_data\VA000000000000017\RXZlbnRJbmZvcn1hdGlv...
4	0 days 00:00:00.100000	4741	3531	5806	808	0.100	14886	Raw_data\VA000000000000017\RXZlbnRJbmZvcn1hdGlv...
5	0 days 00:00:00.125000	4731	3768	5903	853	0.125	15255	Raw_data\VA000000000000017\RXZlbnRJbmZvcn1hdGlv...
6	0 days 00:00:00.150000	4794	3720	5919	983	0.150	15416	Raw_data\VA000000000000017\RXZlbnRJbmZvcn1hdGlv...
7	0 days 00:00:00.175000	4841	3750	5896	1050	0.175	15537	Raw_data\VA000000000000017\RXZlbnRJbmZvcn1hdGlv...
8	0 days 00:00:00.200000	4851	3451	5782	1104	0.200	15188	Raw_data\VA000000000000017\RXZlbnRJbmZvcn1hdGlv...
9	0 days 00:00:00.225000	4772	2904	5659	1137	0.225	14472	Raw_data\VA000000000000017\RXZlbnRJbmZvcn1hdGlv...



We have 13075 csv file. Each has 4 load sensor reading with time gap of 0.025 sec

Each csv has json file which has property of the event occurs.

Data Selection / Extraction from load sensors

- Load sensor merge at statistical measure

Lc sum Mean

Lc sum Min

Lc sum Median

Lc sum Max.

Lc sum Std

Lc sum Range

Lc sum entropy

Lc sum rms

Lc sum skew

Lc sum kurtosis

10 Features

From **4 sensors** value of :
Mean, Max, Min, std, log energy entropy

Mean

Minimum

Std

Maximum

4*4 = 16 Features

- Duration of Activity

Total = 26 Features Extract for statistical measure

02 – Data Prepration

Convert JSON File to Data frame

	Device_ID	Event_ID	Event_start_time	Tags	Free_text	Name_of_cat	Weight_of_cat	Tare_weight
0	VA00000000000017	RXZlbnRJbmZvcmlhdGlvbjo0OWQ4M2l3MS1kNDZmLTRiYj...	2020-09-05T10:56:11-05:00	[urination, cat in box]	None	Ryan	6589	None
1	VA00000000000017	RXZlbnRJbmZvcmlhdGlvbjo0Yml3ZDc5YS0wNzMwLTQ0MT...	2020-07-18T14:41:37-05:00	[cat in box, synapse_labeled, urination]	14:42:15t14:43:09tConfirmed activity, cat in...	Ryan	6935	None
2	VA00000000000017	RXZlbnRJbmZvcmlhdGlvbjo0ZTBhMDVhOC0wZml2LTQ2MG...	2020-09-03T15:43:43-05:00	[cat in box, urination]	None	Ryan	5139	None
3	VA00000000000017	RXZlbnRJbmZvcmlhdGlvbjo1NGEwMTg2My01NDhjLTRhYT...	2020-09-01T12:02:05-05:00	[non_elimination, cat in box]	None	Ryan	6983	None
4	VA00000000000017	RXZlbnRJbmZvcmlhdGlvbjo1YzQwMzE4MS1lNzhkLTQ3Nz...	2020-08-08T06:47:11-05:00	[cat in box, synapse_labeled, urination]	6:47:16t6:48:59tConfirmed activity, cat in t...	Ryan	7017	None
5	VA00000000000017	RXZlbnRJbmZvcmlhdGlvbjo1ZWQyODcxZS0wYTl1LTQxMW...	2020-09-07T04:33:29-05:00	[cat in box, urination]	None	Ryan	6105	None
6	VA00000000000017	RXZlbnRJbmZvcmlhdGlvbjo2MGlxMzM3My03NmlwLTRhZT...	2020-09-01T10:24:09-05:00	[cat in box, defecation]	None	Ryan	6019	None
7	VA00000000000017	RXZlbnRJbmZvcmlhdGlvbjo2MWE0NzMzS0Z02NTkzLTRlNm...	2020-09-03T06:57:17-05:00	[cat in box, urination]	None	Ryan	6139	None
8	VA00000000000017	RXZlbnRJbmZvcmlhdGlvbjo2ZGEzODIiYy02M2l1LTRmNz...	2020-07-08T02:34:30-05:00	[urination, cat in box]	None	Ryan	7147	None
9	VA00000000000017	RXZlbnRJbmZvcmlhdGlvbjo2ZjYwZTlhNC05OWNhLTRlOD...	2020-07-29T06:28:40-05:00	[cat in box, synapse_labeled, urination]	6:29:16t6:30:11tConfirmed activity, cat in t...	Ryan	7127	None

Extract Diff. Activity from Tags and free text

Elimination

Defecation

Urination

Cat In box

Synapse labelled

Vomit

Jumped

Cat Rubbing

Drift

Covering & Digg.

Event Time features

Year Month

Day Time

Features that may not impact

Device ID

Tare Weight

Data Merging with Joining key

JSON FILE:
EVENT_ID

CSV FILE:
File Name

Final
Dataset

02 – Data Prepration

Merged Dataset

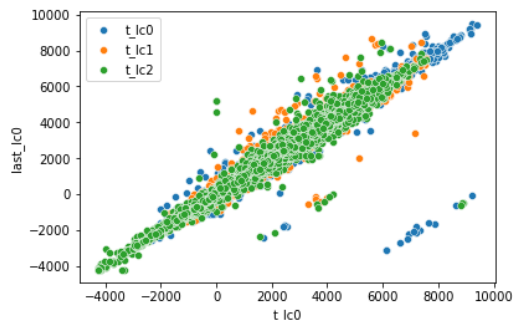
	Device_ID	Event_ID	Name_of_cat	Weight_of_cat	Tare_weight	index	lc_sum_mean	lc_sum_std	time_span	min_total_load	max_total_load	median_total_load	std_total_load	skew_total_load	kourtosis_total_load	range_total_load	log_e_entro_total_load	lc_max_4_sensor
0	VA000000000000017	RXZlbnRlbnZvcmlhdGlvbjo0WQ4M2l3MS1kNDZmLTRlYj...	Ryan	6589	NaN	0	10112.07	2961.45	151.97	7982	16035	8214.0	2961.45	0.93	-1.11	8053	-8.55	4777.92
1	VA000000000000017	RXZlbnRlbnZvcmlhdGlvbjo0YmI3ZDc5S0wNzIwLTQ0MT...	Ryan	6935	NaN	1	10049.85	3453.25	111.97	6796	14962	6825.0	3453.25	0.13	-1.97	8166	-8.20	5344.23
2	VA000000000000017	RXZlbnRlbnZvcmlhdGlvbjo0ZTBhMDVhOC0wZmI2LTQ2MG...	Ryan	5139	NaN	2	11050.50	2572.06	102.37	8739	15311	8875.0	2572.06	0.31	-1.84	6572	-8.21	4815.72
3	VA000000000000017	RXZlbnRlbnZvcmlhdGlvbjo1NGEwMTg2My01NDhjLTRhYT...	Ryan	6983	NaN	3	9855.52	2384.83	66.37	8786	16092	8795.0	2384.83	1.89	1.69	7306	-7.76	3067.45

lc_min_4_sensor	std_of_mean_4_sensor	...	PSD_15_to_20Hz_mean	PSD_15_to_20Hz_std	power_by_cat	non_elimination	urination	defecation	synapse_labeled	cat_on_edge	cat_rubbing_box	partial_cat_in_box	vomit	Act_jumped	Act_paws	Act_drift	Act_feet5mm	Act_digging_cover	Date	Time	EVENT_MONTH	hour
-1406.29	2349.53	...	97.96	72.72	2358.4	0	1	0	0	0	0	0	0	0	0	0	0	0	2020-09-05	10:56:11	9	10
-2520.61	3097.21	...	80.68	63.16	1630.7	0	1	0	1	0	0	0	0	0	0	0	0	0	2020-07-18	14:41:37	7	14
921.63	1501.45	...	16.64	15.70	2432.5	0	1	0	0	0	0	0	0	0	0	0	0	0	2020-09-03	15:43:43	9	15
1248.88	714.40	...	231.68	63.10	923.2	1	0	0	0	0	0	0	0	0	0	0	0	0	2020-09-01	12:02:05	9	12

Dataset prepare after statistical measure on the load sensor and activity extract from the tags and free text

Data Cleaning : Missing Value

A) Impute Missing value :Tare value



Linear relation between tare value and last value of the sensor

B) Impute Missing value : Cat Name

```
list = {}
for i in range(0, len(dataset)):
    if dataset['Name_of_cat'][i] != 'None':
        list.update({dataset['Device_ID'][i] : dataset['Name_of_cat'][i]})

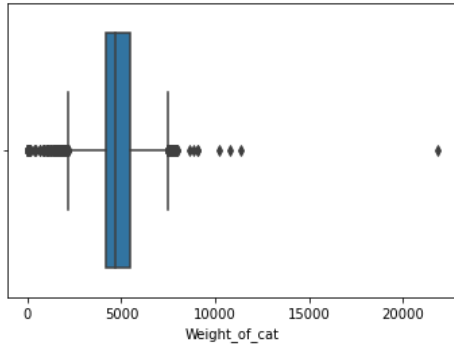
dataset['Name_of_cat'].fillna('0' , inplace = True)

for i in range(0, len(dataset)):
    if dataset['Name_of_cat'][i] == '0' :
        dataset['Name_of_cat'][i] = list[dataset['Device_ID'][i]]
```

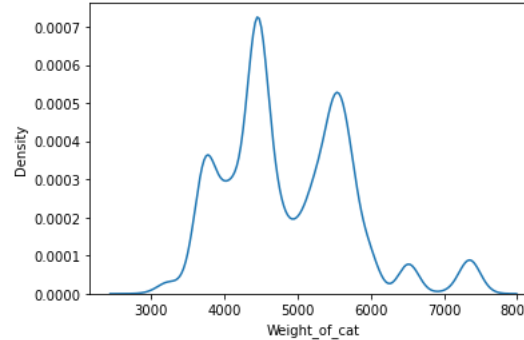
Each **Device ID** has unique **Cat name** so impute by same

02 – Data Prepration

Data Cleaning : Treated Outlier Value



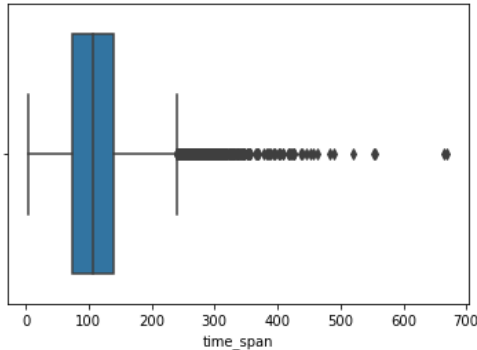
- Box plot of weight of all cat



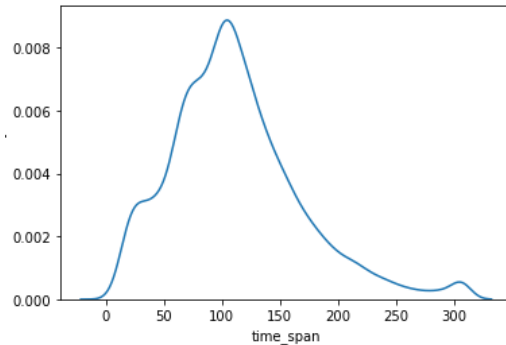
- Distribution after treating outlier of weight

After Exploring Cat Weight so, find some cat name that has more or less weight, so Treat the Outlier of by Cat Name median weight value

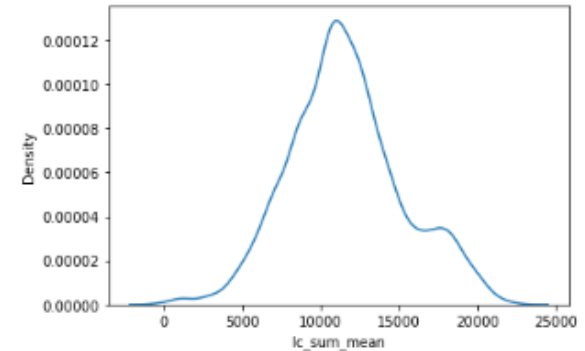
By Exploring Time span, find no particular reason for the large duration so treat by upper bound value of IQR method.



- Box plot of time span by cat



- Distribution after treating outlier of weight



- Distribution of load sensor sum mean

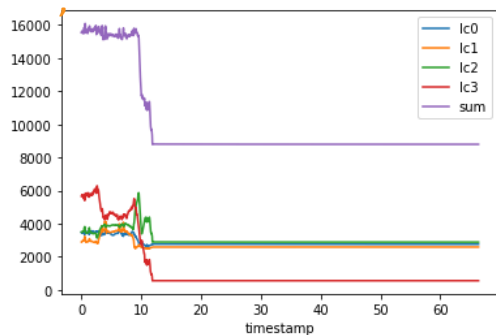


03. Data Exploration

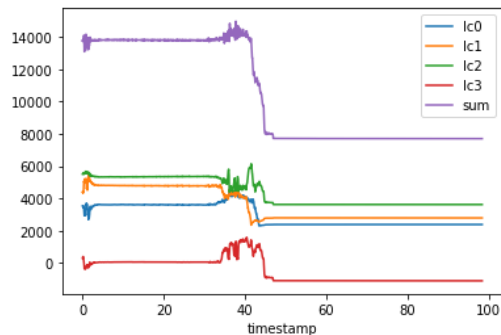
1. Time duration of activities
2. Statistical measures of load sensor data differ for each activity
3. Activities' behaviour visualization of raw signals for different activities
4. For elimination (defecation and urination) usually the cat digs up the litter, eliminates and covers it up. Relation between digging up and covering activity differ between urination and defecation
5. Summary of activity patterns

03 – Exploratory data Analysis(Bi-variate analysis)

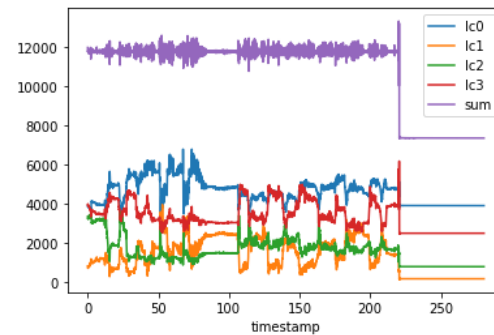
Activity : Non-Elimination, Urination and Defecation



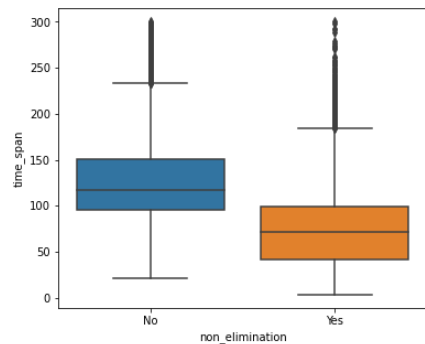
In Non-Elimination there is no pattern but happens in short time



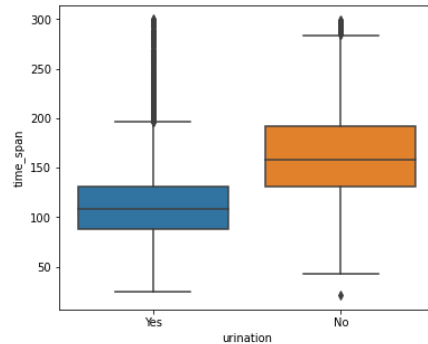
In Elimination(Urination) there is no pattern that happens .



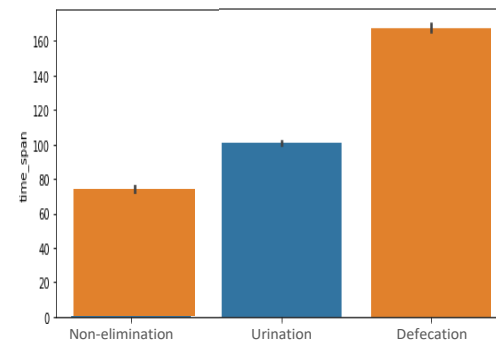
In Elimination(defecation) there is more disturbance or less stable in signal



In Non-Elimination there is pattern that happens in short time



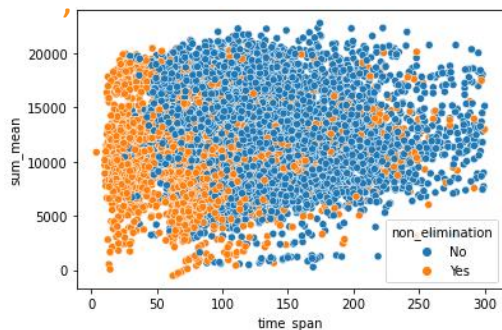
In Elimination(Urination) there is pattern that happens in short time with non-Eli.



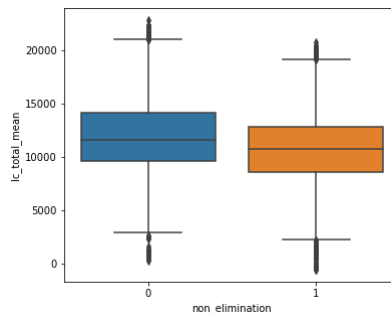
Clearly shows how different activity takes time for non-elimination

03 – Exploratory data Analysis(Bi-variate analysis)

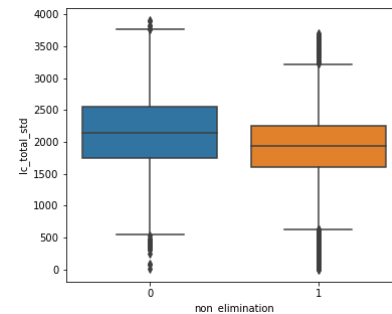
Activity : Non-Elimination, Urination and Defecation



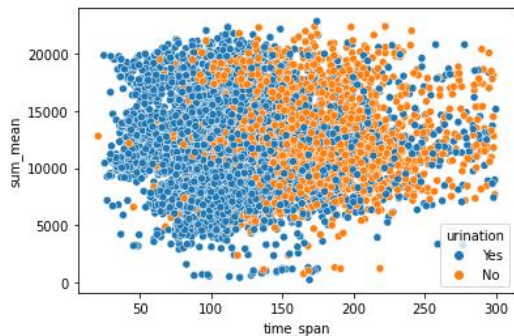
There is no pattern in the total load sensor measure in the activity



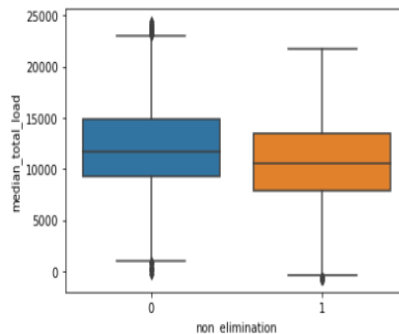
There more total load mean in the elimination activity



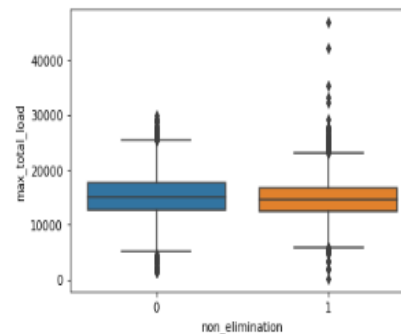
There more total load std in the elimination activity



In elimination, low time span there is Urination



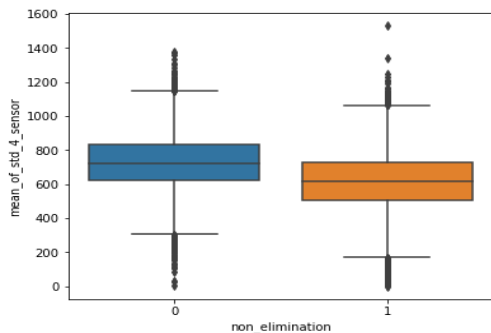
There more total load median in the **elimination** activity



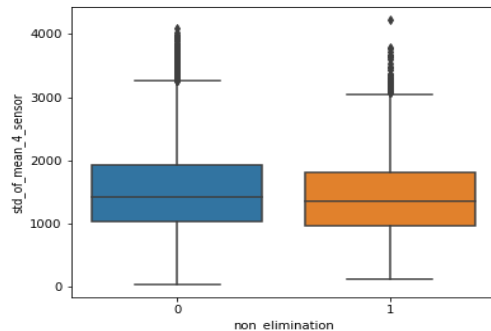
Std deviation of load has no pattern for elimination activity

03 – Exploratory data Analysis(Bi-variate analysis)

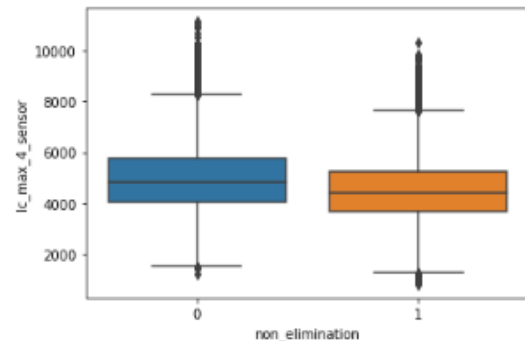
Load signal statistics vs Target Variable



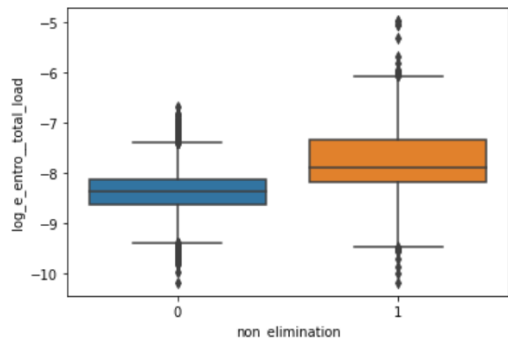
Mean of four load has higher in case of elimination overall



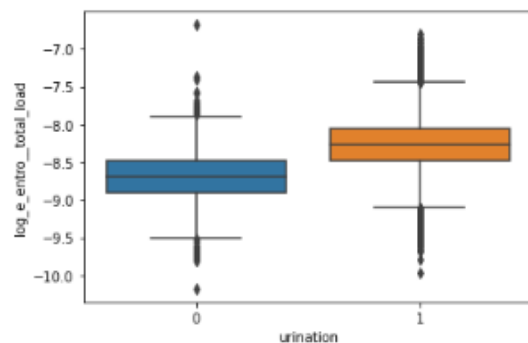
Std of four mean load sensor has more in elimination overall



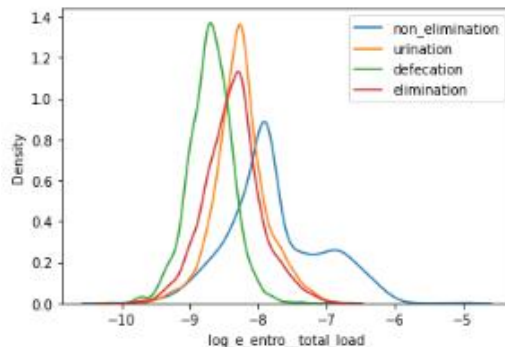
Max of four load mean sensor has more in elimination overall



Log energy entropy of total load has lower in elimination overall



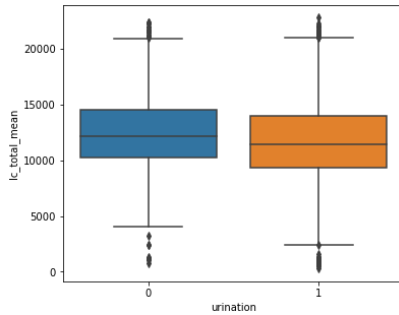
Log energy entropy of total load has lower in defecation overall



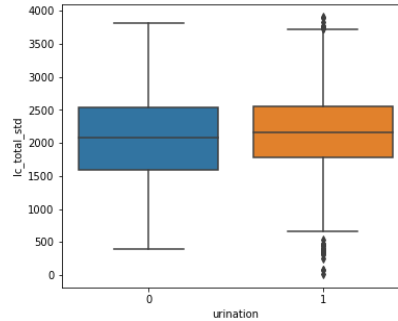
Activity takes to more on elimination

03 – Exploratory data Analysis(Bi-variate analysis)

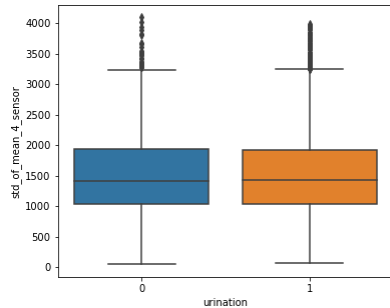
Activity : Non-Elimination, Urination and Defecation



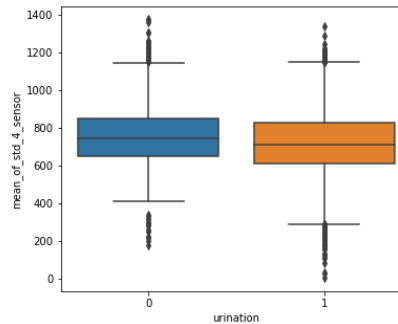
There is no pattern in the total load sensor mean in the activity



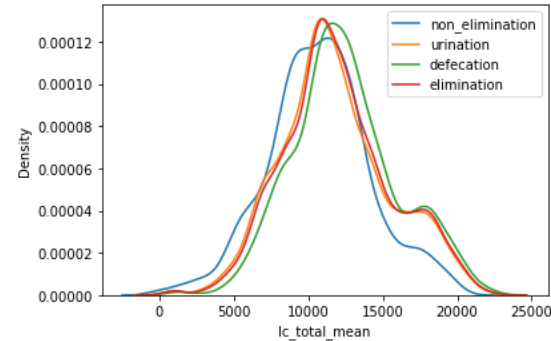
There is no pattern in the total load std measure in the activity



Mean of std of load has no pattern for elimination activity



Std deviation of load has no pattern for elimination activity



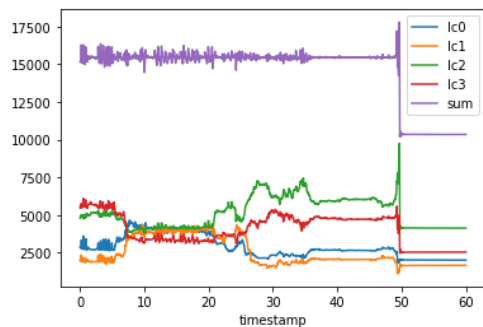
There is no pattern in the defecation and urination activity

Observation:

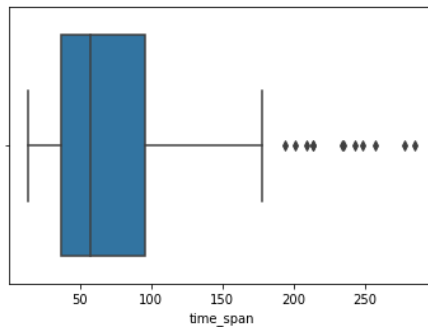
- Time span of Elimination and non-elimination was more and also for log energy entropy in elimination
- Mean of std of load each sensor is higher for elimination
- Total load mean has lower in non-elimination activity
- There is no pattern in Urination vs Defecation

03 – Exploratory data Analysis(Bi-variate analysis)

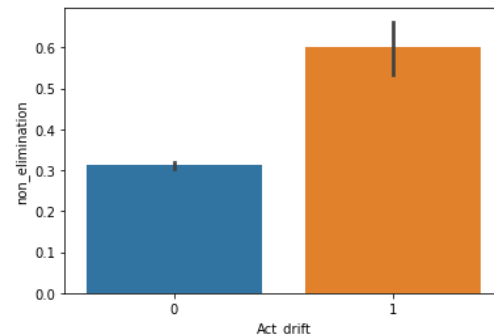
Activity A) Drift



Drift activity load sensor change/drop at last 10 second only with less stable load

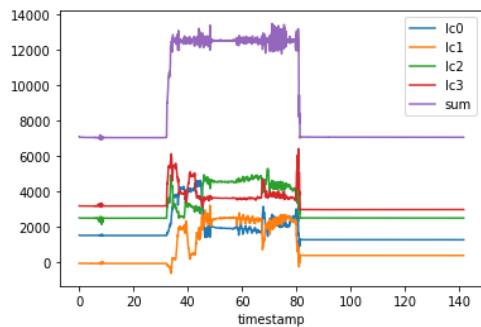


Time span of activity is usually takes minimum 40seconds

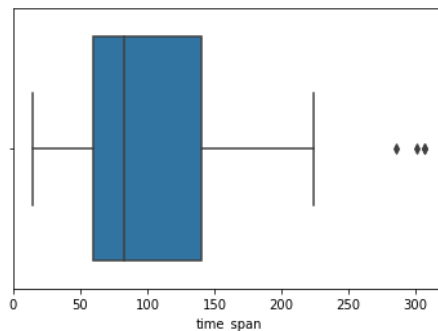


Drift activity takes to less on elimination

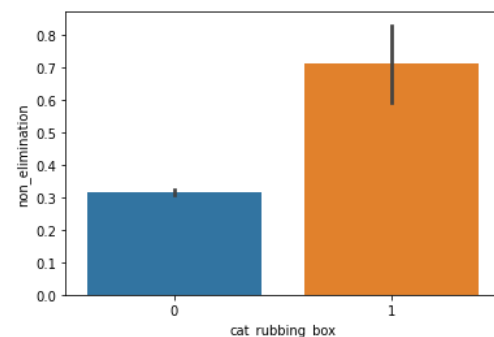
Activity B) Cat Rubbing Box



Starting and ending load was same that means activity happen between



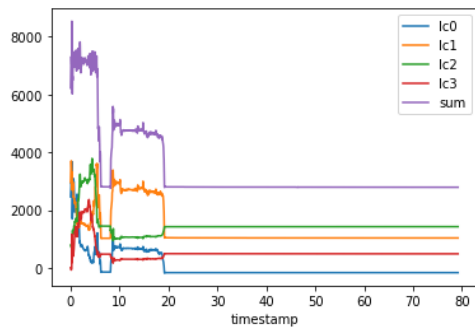
Time span of activity is no pattern with time duration



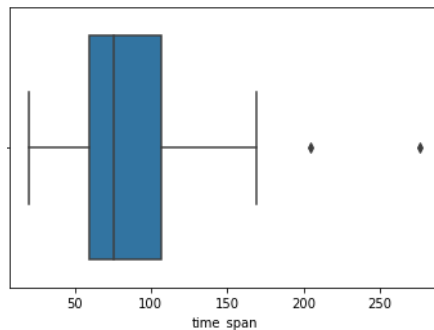
Rubbing activity takes to less on elimination

03 – Exploratory data Analysis(Bi-variate analysis)

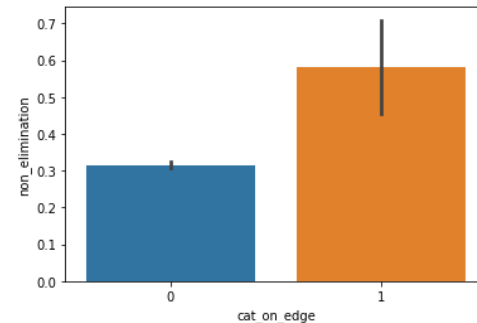
Activity C : Cat on the Edge



Cat on edge activity load sensor has no pattern in raw signals with time

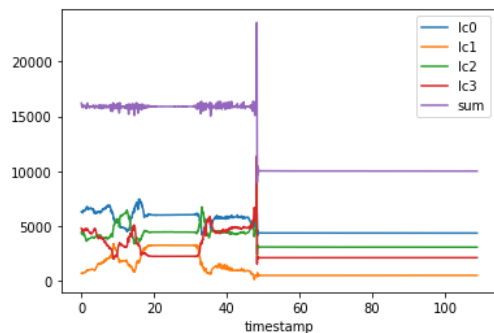


Time span of activity is usually takes minimum 40seconds

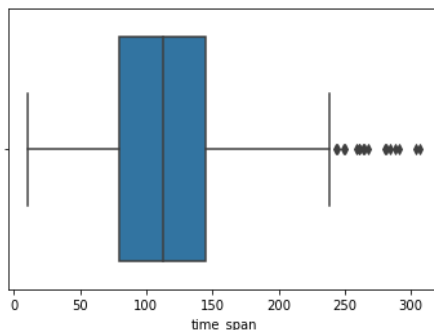


Cat on the edge activity takes to less on elimination

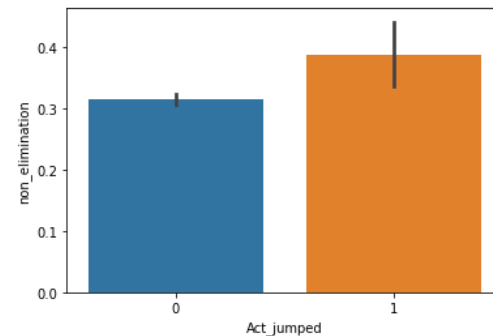
Activity D: Jumped



Jumped activity load sensor change very high at any point of time



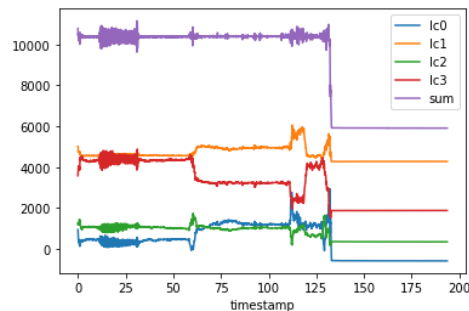
Time span of activity is usually takes more time(60seconds)



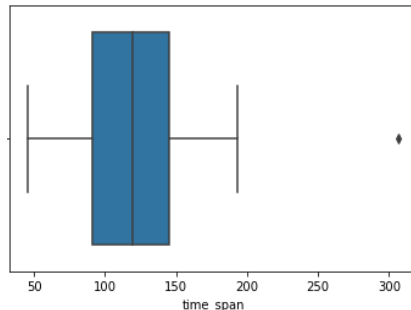
Jumped takes to almost same on elimination vs non_elimination

03 – Exploratory data Analysis(Bi-variate analysis)

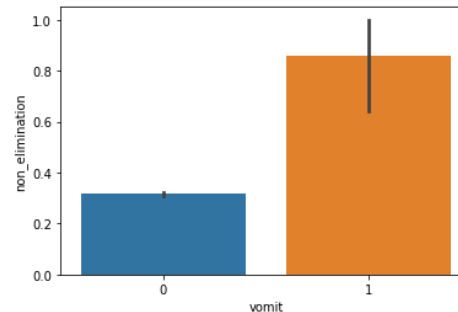
Activity E : Vomit



Vomit activity load sensor has no pattern in raw signals with time

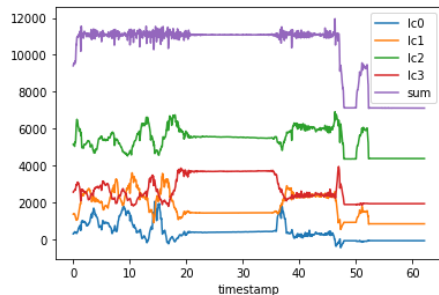


Time span of activity is usually takes minimum 50seconds

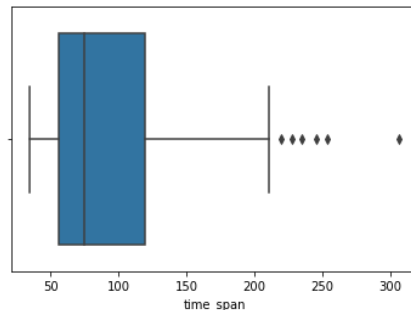


Vomit activity takes to less on elimination

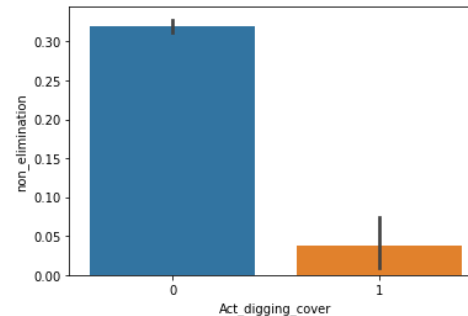
Activity F: Digging & Covering



Digging activity load sensor change adaptively before 10 sec. for 20 sec

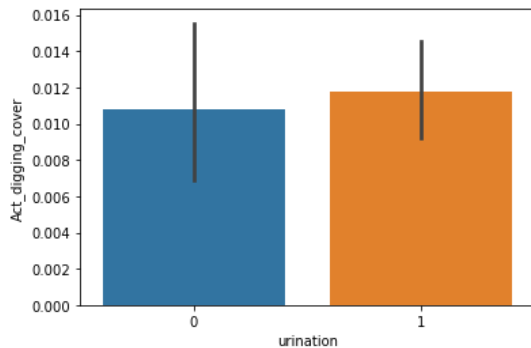


Time span of activity is usually takes minimum 40seconds



Digging activity takes to more on elimination

03 – Exploratory data Analysis(Bi-variate analysis)



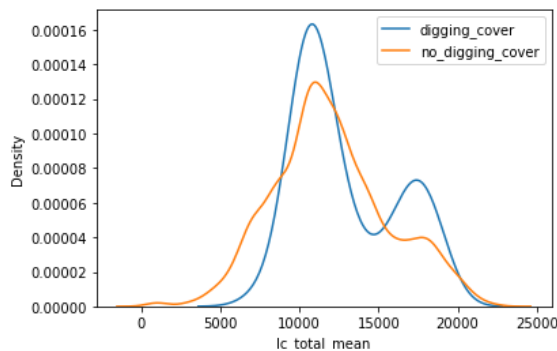
There is no major difference in activity digging and urination or defecation

		Activity	p_value
0	Act_jumped	0.026071	
1	cat_on_edge	0.059106	
2	Act_drift	0.198091	
3	vomit	0.437733	
4	Act_digging_cover	0.698185	
5	cat_rubbing_box	0.820135	

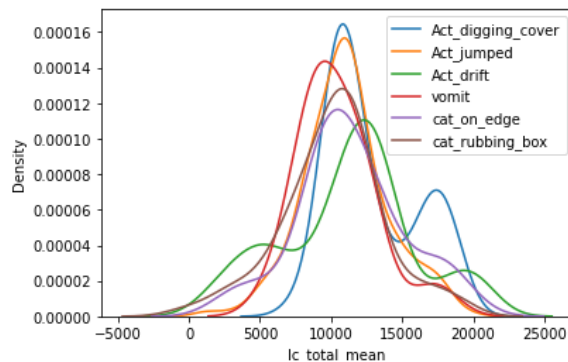
Chi2 test between Urination and activity

		Activity	p_value
0	Act_drift	7.769518e-20	
1	cat_rubbing_box	7.286858e-11	
2	Act_digging_cover	5.006966e-10	
3	vomit	1.415266e-05	
4	cat_on_edge	2.470266e-05	
5	partial_cat_in_box	1.035731e-04	
6	Act_jumped	5.895888e-03	
7	Act_paws	1.298164e-02	
8	Act_feet5mm	1.278497e-01	

Chi2 test between elimination and activity



Digging activity load sensor higher in digging compare to no digging



There is less difference in other activity vs digging activity of load sensor

Observation:

- All activity has no impact less impact on elimination compare to digging and covering
- Almost all activity takes less time than digging and covering with minimum time of 40 second
- Digging and covering has more load compare to the other activity
- There is no relation between behaviour of digging in urination and defecation activity
- Total load mean has lower in non-elimination activity
- There is relation between activity and elimination

04. Feature Engineering

Features & Cat behavior

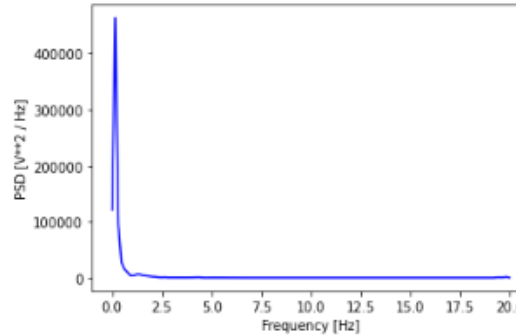
1. Perform appropriate data transformation/aggregation (Look for time domain, frequency domain features)
2. cat have a specific behaviour pattern?
3. Are the patterns similar across different cat morphologies?
4. the relationship of features with target variable

04 – Feature Engineering

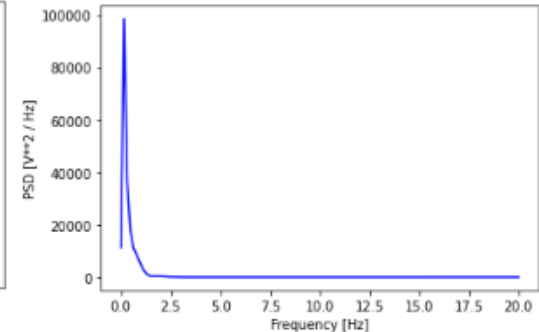
Feature A: Signal Processing (Power Spectral Density)

Transformations between time and frequency-domain

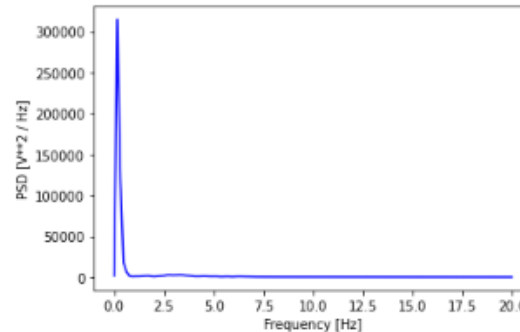
- **Fast Fourier Transform (FFT)** : It is an efficient algorithm for calculating the Discrete Fourier Transform (DFT) and is the de facto standard to calculate a Fourier Transform.
- **Power Spectral Density** : Similar to the FFT, it describes the frequency spectrum of a signal. But in addition to the FFT it also takes the power distribution at each frequency (bin) into account.
- **Autocorrelation** : The auto-correlation function calculates the correlation of a signal with a time-delayed version of itself.



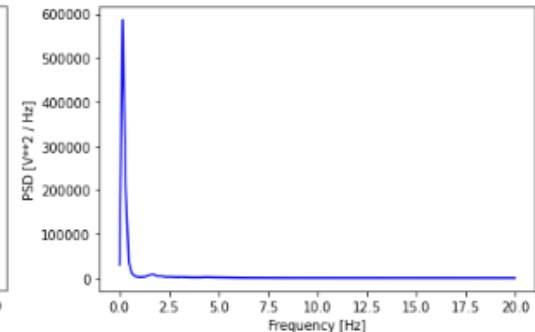
PSD Distribution for elimination



PSD Distribution for elimination



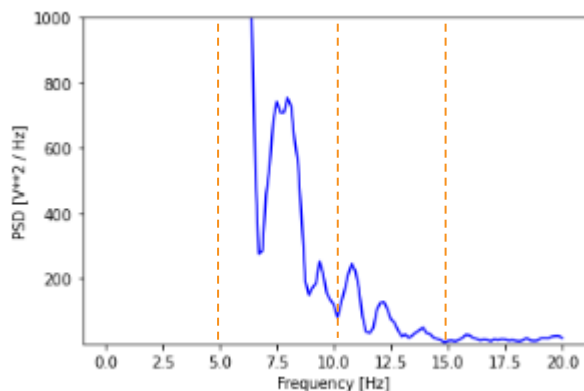
PSD Distribution for urination



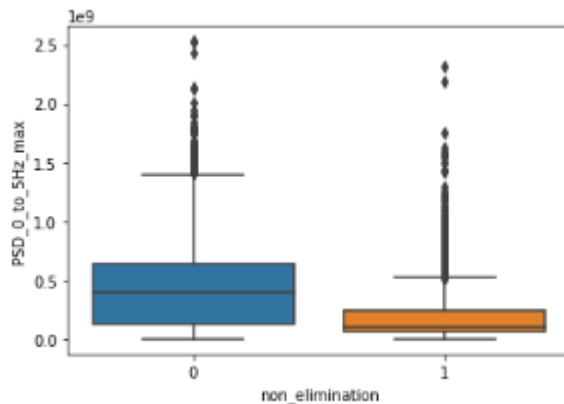
PSD Distribution for defecation

04 – Feature Engineering

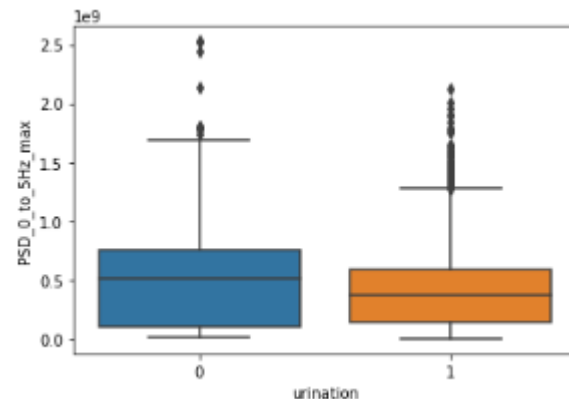
Feature A: Power Spectrum Density



We can divide the whole frequency in 4 parts



PSD Having impact on the elimination



PSD Having impact on the urination

Feature that are Extracted from the data

PSDF_0_5Hz_Mean

PSDF_5_10Hz_Mean

PSDF_10_15Hz_Mean

PSDF_15_20Hz_Mean

PSDF_0_5Hz_Median

PSDF_5_10Hz_Median

PSDF_10_15Hz_Median

PSDF_15_20Hz_Median

PSDF_0_5Hz_Maximum

PSDF_5_10Hz_Maximum

PSDF_10_15Hz_Maximum

PSDF_15_20Hz_Maximum

PSDF_0_5Hz_Minimum

PSDF_5_10Hz_Minimum

PSDF_10_15Hz_Minimum

PSDF_15_20Hz_Minimum

PSDF_0_5Hz_Std

PSDF_5_10Hz_Std

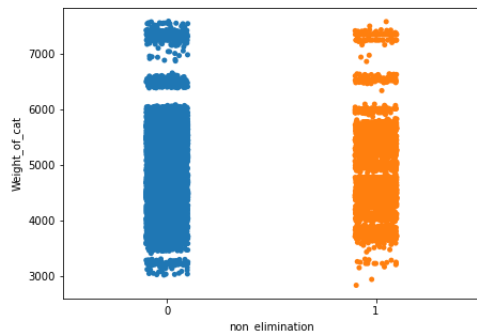
PSDF_10_15Hz_Std

PSDF_15_20Hz_Std

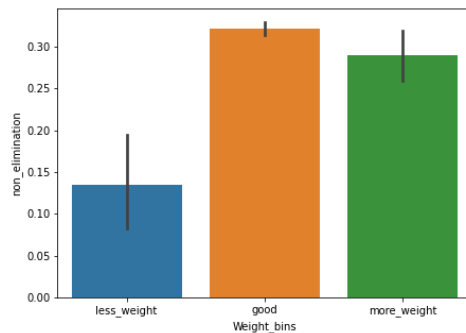
20 Features

04 – Feature Engineering

Feature B: Cat Weight



Weight of the Cat Breed can be Bin



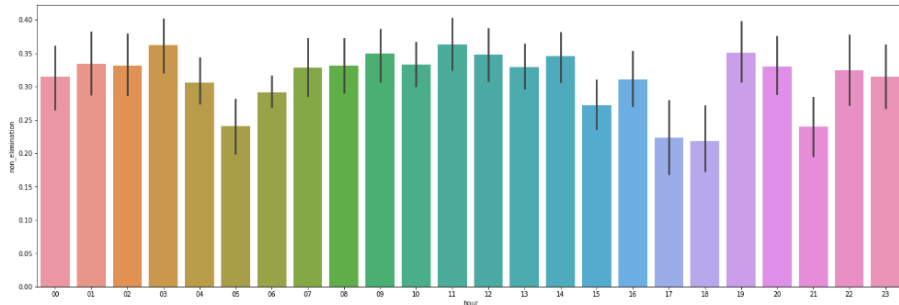
Low less weight cat breed has less elimination compare to more weight

Decision on Bins:

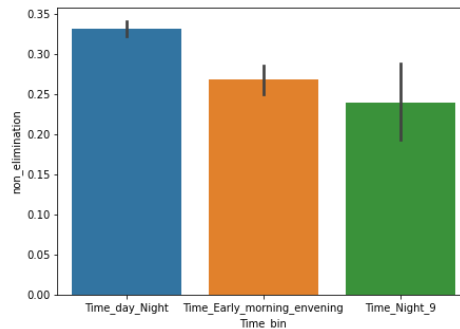
On the basis of Weight there is pattern and type of cat that are occurs :

- Weight less than 3400 - 'Marlowe', 'Meera'
- Weight more than 6500 weight - Ryan', 'Lexi', 'Zeb', 'Simon'

Feature C: Cat elimination event time



In elimination, low time span there is Urination

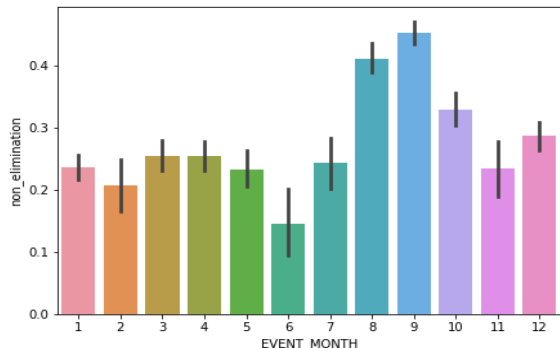


Usually elimination happens in early morning(05-07), evening(17-19) and night(21-22)

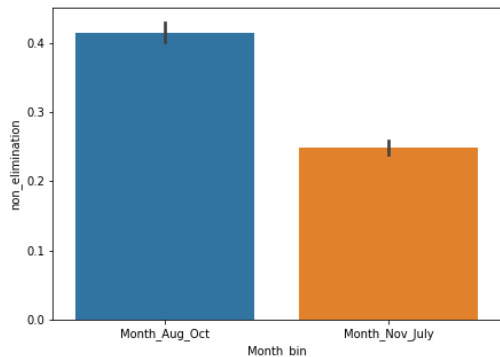
In elimination, low time span there is Urination

04 – Feature Engineering

Feature D: Cat Elimination Month wise



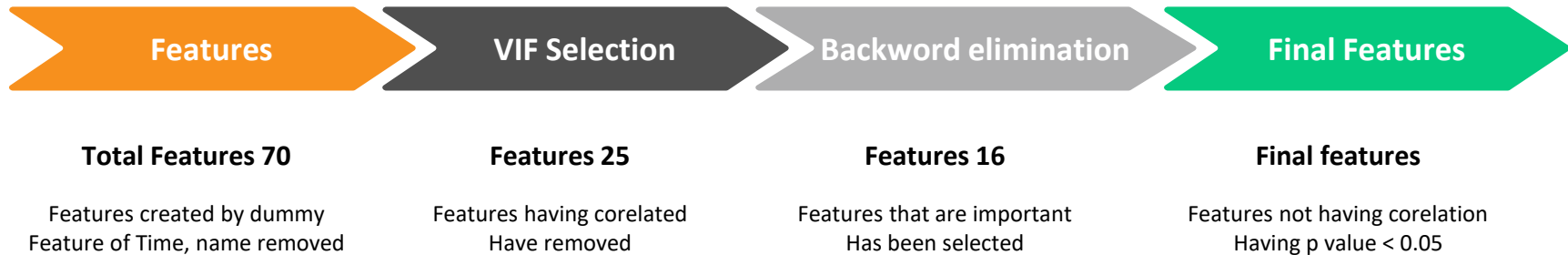
Month wise non-elimination bar plot



Binning month wise plot

August to October non-elimination process is more likely be done

Feature Selection Process



04 – Feature Engineering

Feature Selection

0	Device_ID	31	PSD_5_to_10Hz_mean
1	Event_ID	32	PSD_5_to_10Hz_std
2	Event_start_time	33	PSD_10_to_15Hz_max
3	Name_of_cat	34	PSD_10_to_15Hz_min
4	Weight_of_cat	35	PSD_10_to_15Hz_median
5	lc_sum_mean	36	PSD_10_to_15Hz_mean
6	lc_sum_std	37	PSD_10_to_15Hz_std
7	time_span	38	PSD_15_to_20Hz_max
8	min_total_load	39	PSD_15_to_20Hz_min
9	max_total_load	40	PSD_15_to_20Hz_median
10	median_total_load	41	PSD_15_to_20Hz_mean
11	skew_total_load	42	PSD_15_to_20Hz_std
12	kurtosis_total_load	43	power_by_cat
13	range_total_load	44	non_elimination
14	log_e_entrop_total_load	45	urination
15	lc_max_4_sensor	46	defecation
16	lc_min_4_sensor	47	ls_cat_Box
17	std_of_mean_4_sensor	48	cat_on_edge
18	max_of_std_4_sensor	49	cat_rubbing_box
19	min_of_std_4_sensor	50	mixed
20	mean_of_std_4_sensor	51	partial_cat_in_box
21	std_of_std_4_sensor	52	vomit
22	last_sum	53	Act_jumped
23	PSD_0_to_5Hz_max	54	Act_drift
24	PSD_0_to_5Hz_min	55	Act_digging_cover
25	PSD_0_to_5Hz_median	56	cat_on_top
26	PSD_0_to_5Hz_mean	57	Date
27	PSD_0_to_5Hz_std	58	Time
28	PSD_5_to_10Hz_max	59	EVENT_MONTH
29	PSD_5_to_10Hz_min	60	hour
		61	Weight_bins
		62	Time_bin
		63	Month_bin
		64	std_total_power

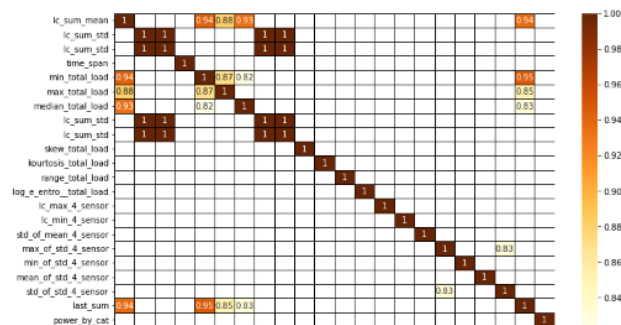
Features that are extracted

0	PSD_10_to_15Hz_median
1	PSD_5_to_10Hz_median
2	mean_std_sensor_power
3	PSD_0_to_5Hz_median
4	last_sum
5	median_total_load
6	log_e_entrop_total_load
7	power_by_cat
8	PSD_15_to_20Hz_median
9	PSD_15_to_20Hz_min
10	PSD_0_to_5Hz_max
11	range_total_load
12	lc_max_4_sensor
13	PSD_10_to_15Hz_min
14	PSD_5_to_10Hz_max
15	lc_min_4_sensor
16	PSD_10_to_15Hz_max
17	max_of_std_4_sensor
18	skew_total_load
19	min_of_std_4_sensor
20	PSD_5_to_10Hz_min
21	PSD_0_to_5Hz_min
22	PSD_15_to_20Hz_max
23	kurtosis_total_load
24	non_elimination

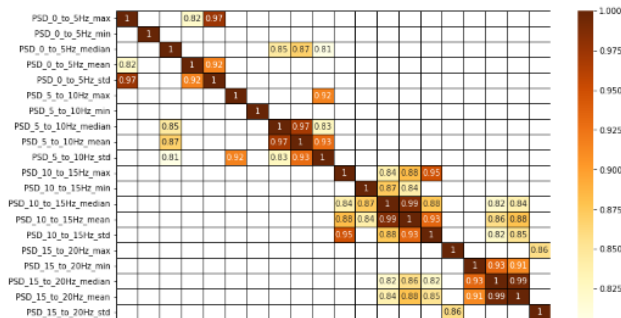
Features by reducing co-linearity

0	PSD_10_to_15Hz_median
1	PSD_5_to_10Hz_median
2	mean_std_sensor_power
3	last_sum
4	median_total_load
5	log_e_entrop_total_load
6	PSD_15_to_20Hz_median
7	PSD_0_to_5Hz_max
8	range_total_load
9	PSD_10_to_15Hz_min
10	PSD_5_to_10Hz_max
11	PSD_10_to_15Hz_max
12	skew_total_load
13	PSD_15_to_20Hz_max
14	kurtosis_total_load
15	non_elimination

Final Features



Load values Features that are co-related



PSD values Features that are co-related

05. Train Model to Elimination & Non- Elimination

Right Modeling Technique

- Different modelling technique for prediction
- Justification of the models based on
 - Hold out validation : Create train, test and validation datasets
 - Model Performance using the evaluation metrics: Confusion Matrix, Multiclass overall and class wise metrics
- Residual analysis
 - Summarize the cause/patterns in misclassifications

06 –Model : Elimination and Non-Elimination

Class imbalance & Test Train Split

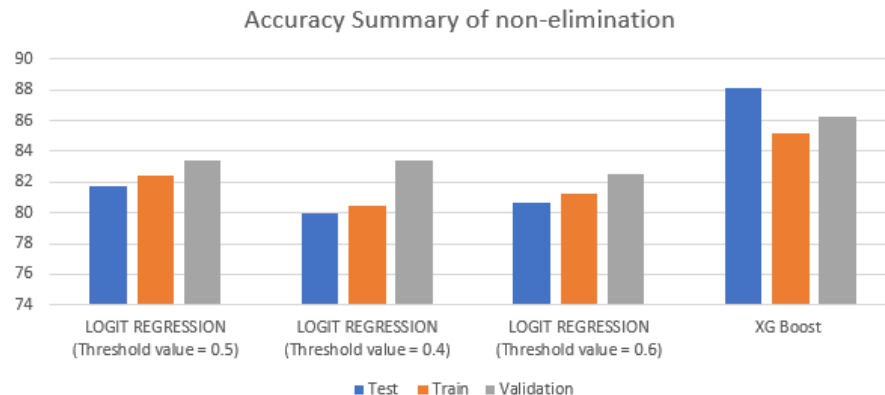
	Non elimination	Class %
0	8358	0.67
1	4073	0.33

		Class %
Train	7458	0.60
Test	2486	0.20
Validation	2486	0.20

Result Summary of Models

	LOGIT REGRESSION (Threshold value = 0.5)			LOGIT REGRESSION (Threshold value = 0.4)			LOGIT REGRESSION (Threshold value = 0.6)			XG Boost (Hyperparameter by Grid search)		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Accuracy	81.7	82.4	83.4	79.97	80.45	83.4	80.7	81.2	82.5	88.12	85.28	86.33
Recall	0.92	0.93	0.92	0.85	0.85	0.92	0.96	0.97	0.97	0.97	0.95	0.94
Precision	0.83	0.83	0.85	0.85	0.86	0.85	0.79	0.80	0.81	0.87	0.85	0.87
Fi-Score	0.87	0.88	0.88	0.85	0.85	0.88	0.87	0.87	0.88	0.92	0.90	0.90

Accuracy of different model



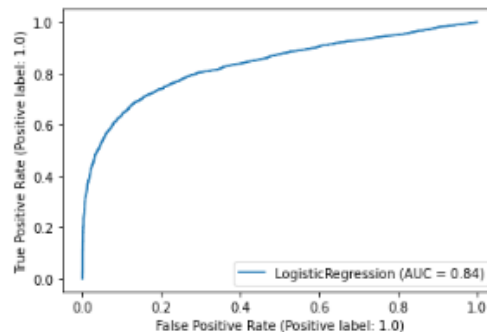
05 – Model : Elimination and Non-Elimination

Train Test: Logistic regression

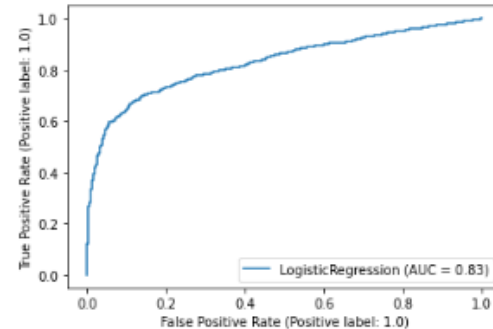
Logit Regression Results

Dep. Variable:	non_elimination	No. Observations:	7458
Model:	Logit	Df Residuals:	7447
Method:	MLE	Df Model:	10
Date:	Sun, 11 Dec 2022	Pseudo R-squ.:	0.2642
Time:	01:48:41	Log-Likelihood:	-3482.3
converged:	True	LL-Null:	-4732.4
Covariance Type:	nonrobust	LLR p-value:	0.000

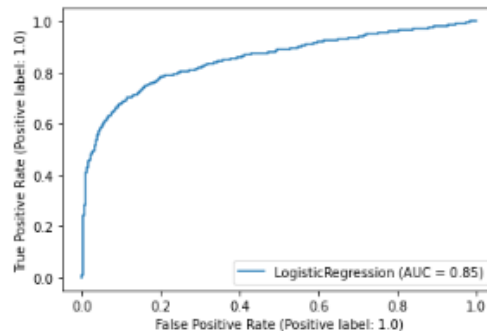
	coef	std err	z	P> z	[0.025	0.975]
const	-7.3019	0.541	-13.494	0.000	-8.362	-6.241
time_span	6.8023	0.580	11.734	0.000	5.666	7.938
log_e_entro_total_load	17.1176	0.943	18.160	0.000	15.270	18.965
min_of_std_4_sensor	-1.5052	0.249	-6.044	0.000	-1.993	-1.017
mean_of_std_4_sensor	-2.6806	0.433	-6.196	0.000	-3.529	-1.833
PSD_0_to_5Hz_max	-3.9483	0.511	-7.721	0.000	-4.951	-2.946
power_by_cat	-0.3478	0.189	-1.839	0.066	-0.718	0.023
EVENT_MONTH	0.0945	0.110	0.857	0.391	-0.122	0.310
Weight_bins_more_weight	0.9345	0.137	6.840	0.000	0.667	1.202
Time_bin_Time_day_Night	0.3538	0.076	4.669	0.000	0.206	0.502
Month_bin_Month_Nov_July	0.3471	0.088	3.958	0.000	0.175	0.519



ROC-AUC Curve for the Train set



ROC-AUC Curve for the validation set



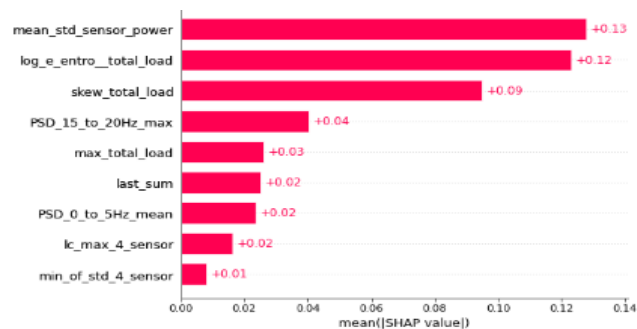
ROC-AUC Curve for the Test set

Features that can be related by negative coef. sign have negative impact on the elimination of cat

Log energy entropy, PSD and time has high in coef. So more important

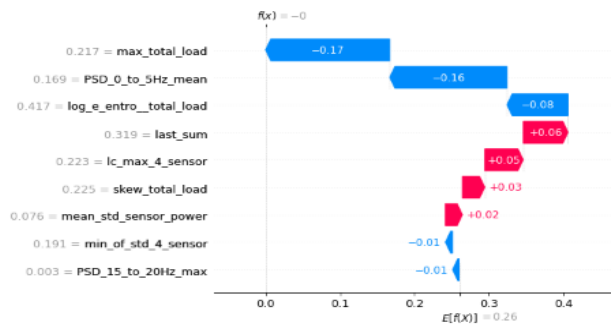
05 – Model Interoperability : Elimination and Non-Elimination

Shap value Graph



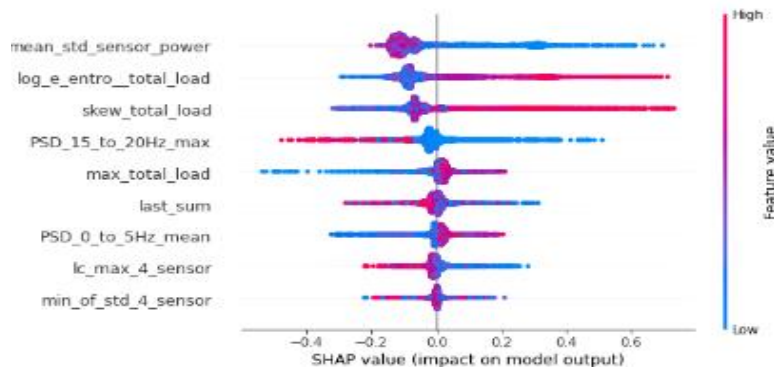
Mean Shap value from the XG Boost model

Shap value impact on the model

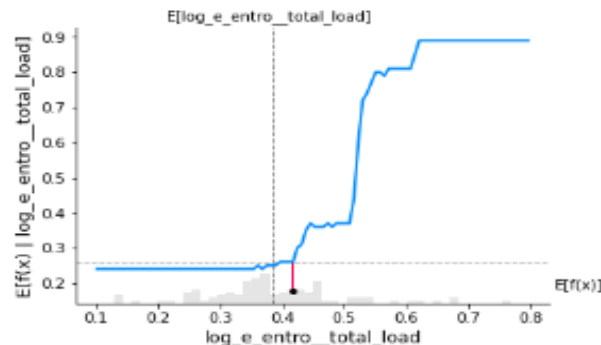


Shap value impact on the XG Boost model

Feature	Importance
mean_std_sensor_power	0.432966
log_e_entro_total_load	0.153114
skew_total_load	0.107625
PSD_15_to_20Hz_max	0.074057
last_sum	0.053374
max_total_load	0.051115
PSD_0_to_5Hz_mean	0.049244
lc_max_4_sensor	0.041573
min_of_std_4_sensor	0.036933



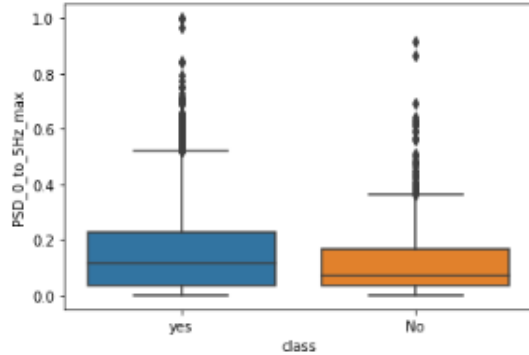
ROC-AUC Curve for the Test set



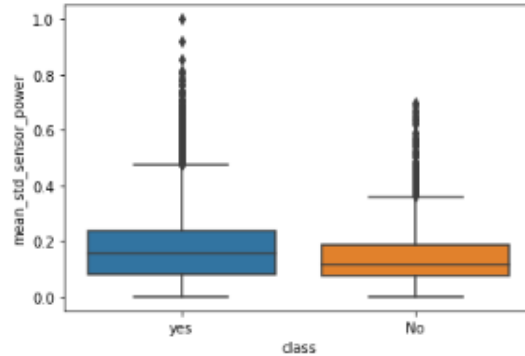
ROC-AUC Curve for the validation set

05 – Model Misclassification : Elimination and Non-Elimination

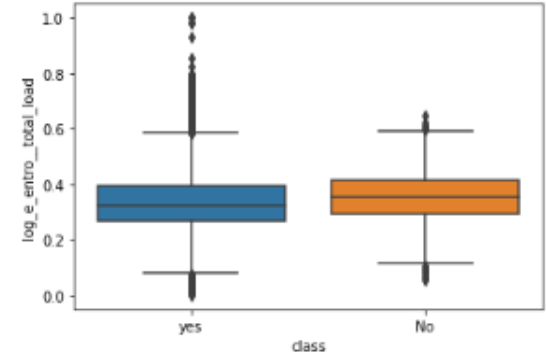
Co



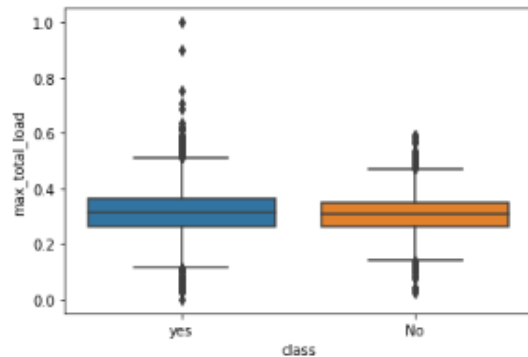
For lower PSD 0-5Hz there is misclass.



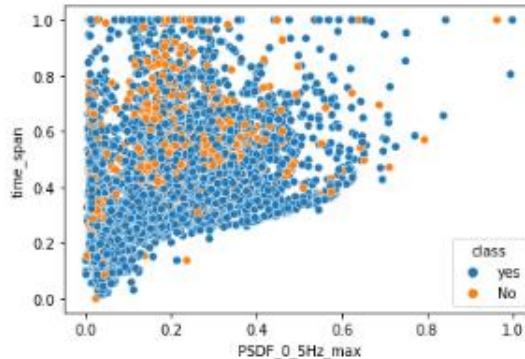
For low mean std there is misclass.



No impact of misclass on log energy entropy



No impact of misclass on Max. load value



No impact of misclass on Med. load value

PSD and mean of std that causes the misclassification of activity

All the mis classification activity are equally in distribution

06. Train Model to Urination & Deflamation

Right Modeling Technique

- Identify the right modelling technique
- Justify the models based on
 - Hold out validation : Create train, test and validation datasets
 - Model Performance using the evaluation metrics: Confusion Matrix, Multiclass overall and class wise metrics
- Residual analysis
 - Summarize the cause/patterns in misclassifications

06 – Model : Urination and Defecation

Class imbalance & Test Train Split

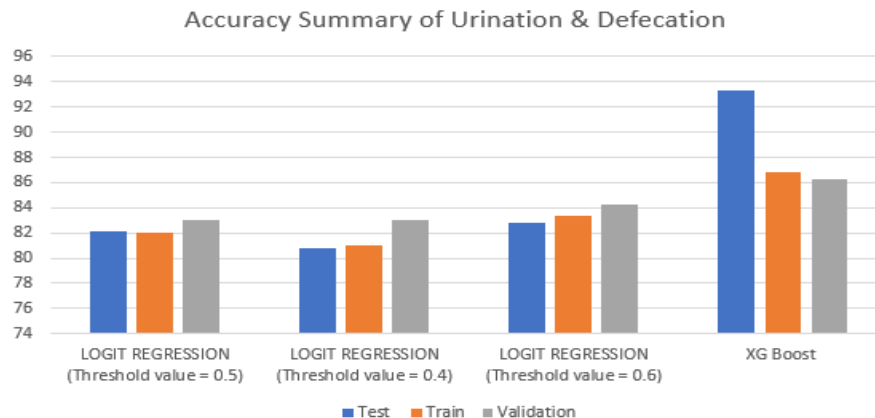
	Non elimination	Class %
Urination	6367	0.76
Defecation	1991	0.24

		Class %
Train	7458	0.60
Test	2486	0.20
Validation	2486	0.20

Result Summary of Models

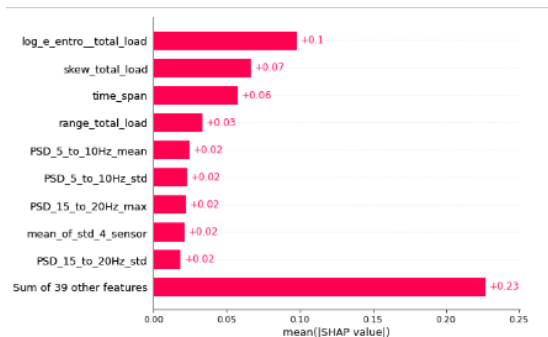
	LOGIT REGRESSION (Threshold value = 0.5)			LOGIT REGRESSION (Threshold value = 0.4)			LOGIT REGRESSION (Threshold value = 0.6)			XG Boost (Hyperparameter by Grid search)		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Accuracy	82.11	82.0	83.0	80.81	80.98	83.00	82.83	83.31	84.21	93.32	86.86	86.24
Recall	0.41	0.41	0.44	0.28	0.29	0.44	0.58	0.57	0.61	0.80	0.62	0.62
Precision	0.70	0.73	0.76	0.74	0.79	0.76	0.65	0.69	0.70	0.91	0.76	0.77
Fi-Score	0.52	0.53	0.56	0.41	0.43	0.56	0.61	0.62	0.65	0.85	0.68	0.69

Accuracy of different model

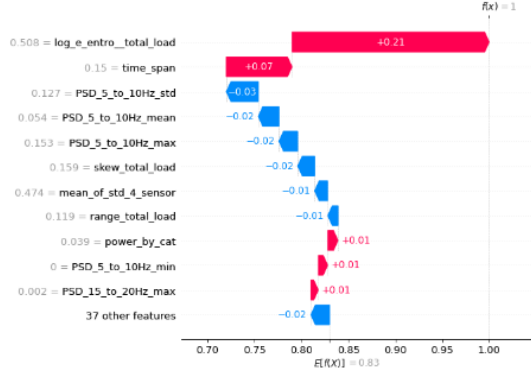


06 – Model Interpretability : Urination and Defecation

Shap value Graph

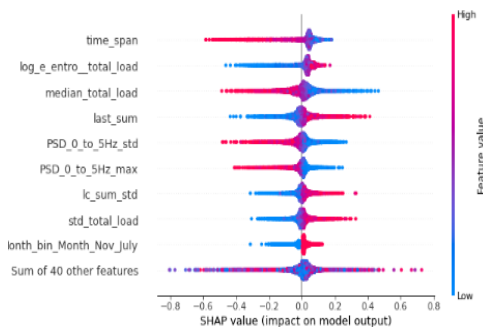


ROC-AUC Curve for the Train set

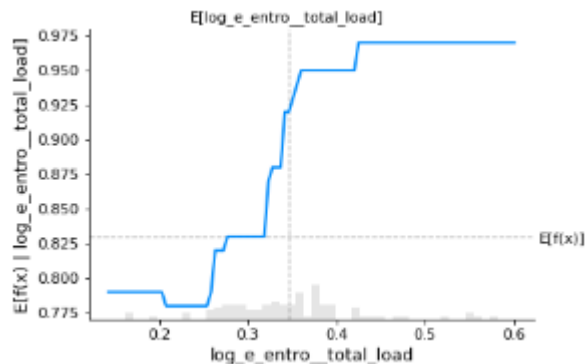


ROC-AUC Curve for the Test set

Shap value impact on the model



ROC-AUC Curve for the validation set

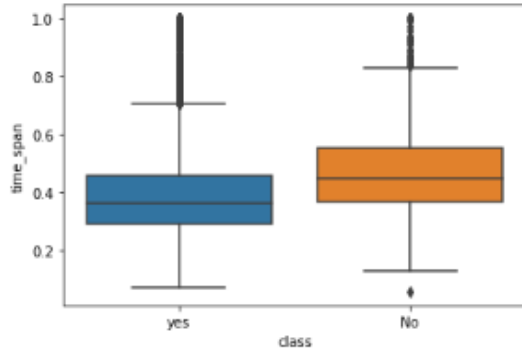


ROC-AUC Curve for the validation set

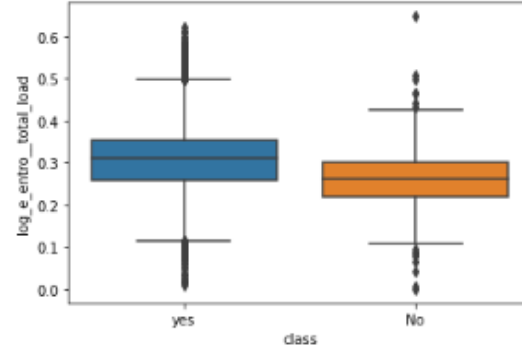
	cols	importance
0	log_e_entro_total_load	0.147310
1	time_span	0.084075
2	skew_total_load	0.054337
3	Weight_bins_good	0.029810
4	PSD_5_to_10Hz_std	0.028865
5	PSD_0_to_5Hz_mean	0.028398
6	mean_std_sensor_power	0.023475
7	Weight_bins_more_weight	0.022785
8	kourtosis_total_load	0.022681
9	PSD_15_to_20Hz_max	0.020380
10	lc_sum_std	0.020286
11	PSD_0_to_5Hz_std	0.019341
12	PSD_5_to_10Hz_mean	0.018064
13	Time_bin_Time_Night_9	0.018480
14	std_total_power	0.018160
15	PSD_15_to_20Hz_std	0.018083

Feature importance of the variables

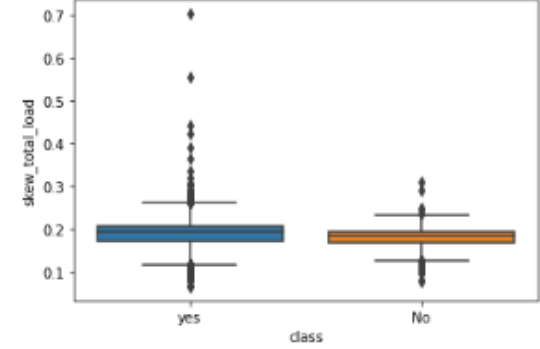
06 – Model Misclassification : Urination and Defecation



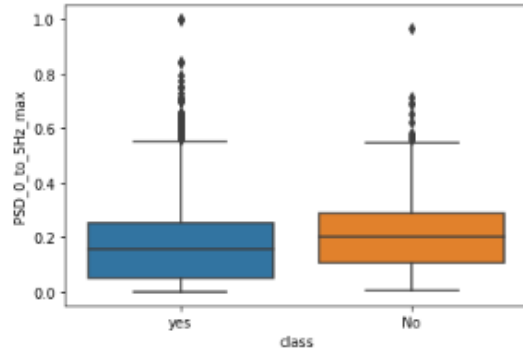
For higher time span then there is misclass.



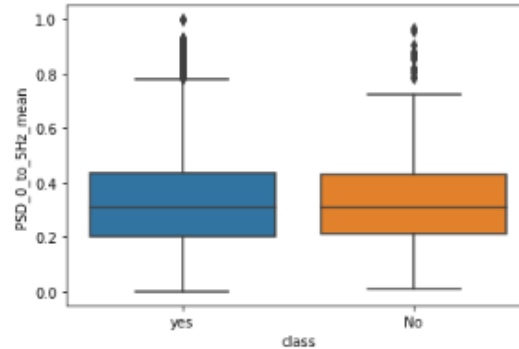
For lower log entropy of std then there is misclass.



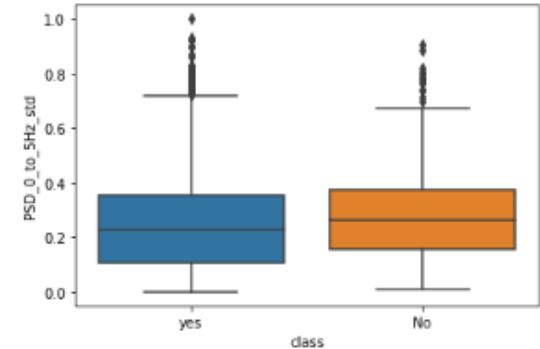
No impact of misclass on skew of load



Higher PSDF tends to miss class.

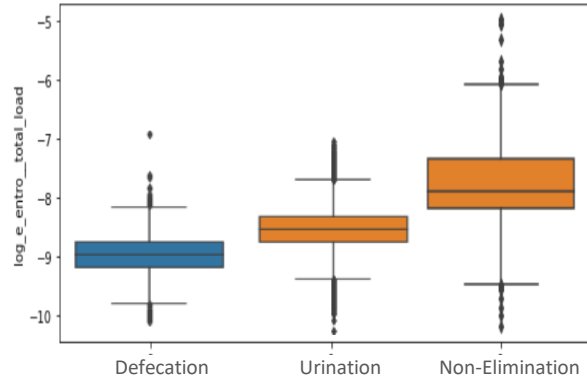


No impact of misclass on PSDF mean value

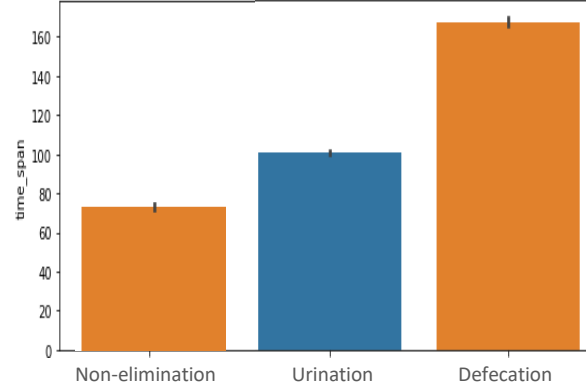


Higher std of PSD on misclass

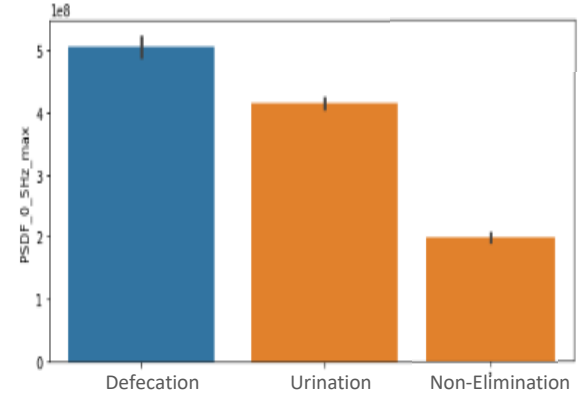
07. Business Insight



A) Log energy entropy is directly says that more entropy will cause the non-elimination that is directly impacted on the health of Cat

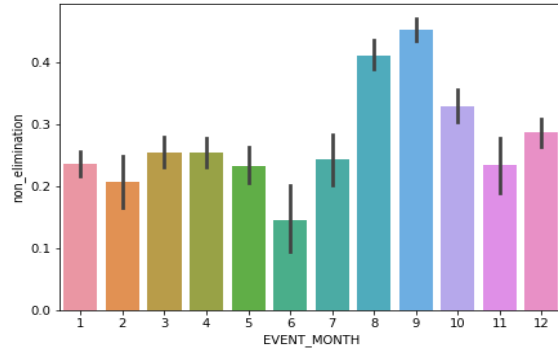


B) Time taken is directly impact on the type of activity may be used to describe the health that product used

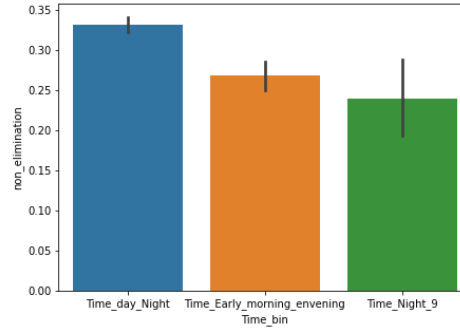


C) PSD OF Freq 0-5Hz is directly impact on the type of activity may be used to behaviour and health

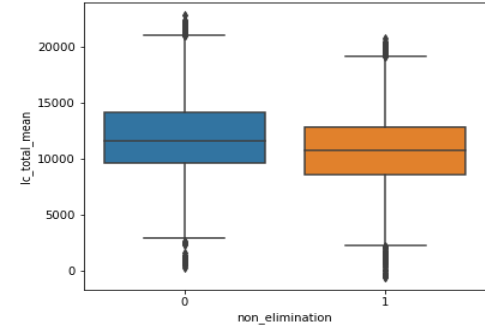
07. Business Insight



D) Event Month can say the August to October has less impact on the elimination



E) Early morning and evening time have more likely to eliminate and



F) Overall load value of elimination is more compare to the non-elimination

08. Conclusion

- A. If there is an elimination activity done by the cat, then there would be more time, but the log energy entropy would be less.
- B. Energy Entropy of the load sensor can determine the different activities, which we can also use to detect the health of the Cat.
- C. Activity day/month timing also plays an essential part in defining the activity, like the month and day part of the activity.
- D. Cat breed will also impact the urination activity in our case good weight cat has more importance than others.
- E. Each activity(drift, digging, and covering) also play a good part in detecting the elimination but has no impact on defining elimination and defecation.
- F. Starting power density received by the sensor is a lower case of cat elimination.
- G. Distribution of load also plays a part in determining the activity



Thank You

Do you have any questions?

Shubham Verma

Shubham.verma@tigeranalytics.com

+91-8077085794

www.tigeranalytics.com

