CS 399 COMMUNICATION SECURITY

Call Detail Record Analytics

Furkan CANTÜRK

22.05.2020

Data Features

- There are 58143 call records for time range of 8 days - December 17-24 in 2018.
- 20% of callees are domestic and there are 48 different countries in the remaining non-domestic partition.

Identical Features: id_pk, call ID

Temporally Features: start time, end time, ring time, answer time, p_ras, p_crt

Numerical Features: duration, p_dlm

Categorical Features: caller, callee, callee country, callee prefix level, disposition, source IP, destination IP, error code, trunk ID, caller group ID, callee group ID, diverted_from, source port, destination port, caller indicator ID, callee indicator ID, is_forwarded, call_dir, call state

- Red ones are useless features which have only one.
- Blue ones are have similar information with start time feature.
- Purple ones have two values, but almost all of the records take one value.

Data Cleaning and Enrichment

- Start time is quantized to hours. -> call hour
- Day is extracted from start time. -> call day
- Caller-caller pair group ID is the combined form of caller and callee group IDs. -> pair group ID
- The colored features in the previous slide are discarded and 3 new features are added to the calls.
- More features will be added in the next slides.

Numerical Features: duration, call hour, call day

Categorical Features: caller, callee, callee country, callee prefix level, disposition, source IP, destination IP, error code, trunk ID, caller group ID, callee group ID, pair group ID

Data Cleaning and Enrichment

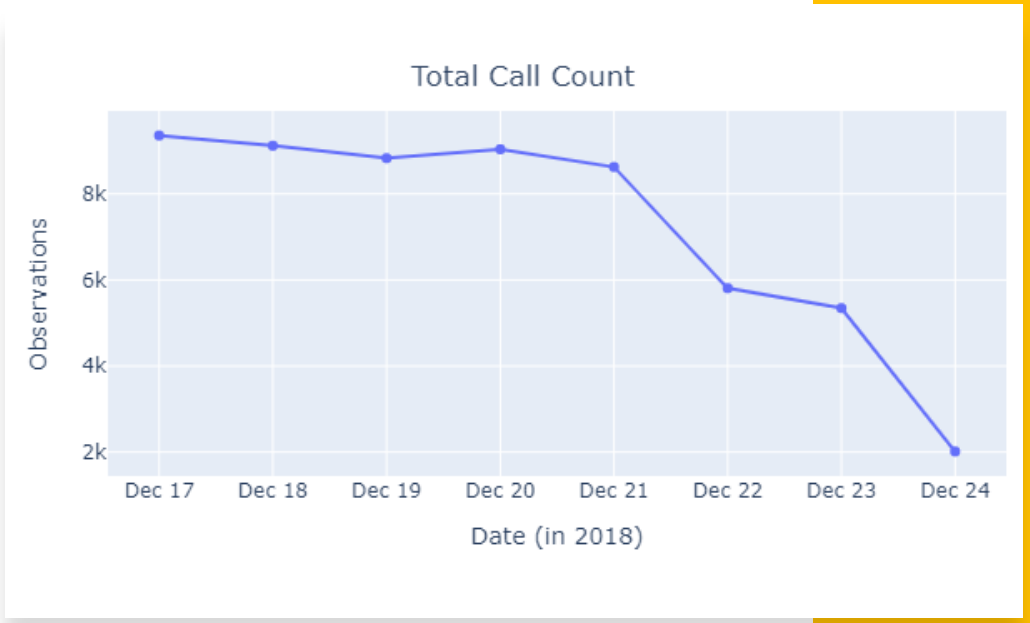
- There are too many class levels in some categorical features.
- Caller, callee, and callee country values should be segmented to be used in data learning effectively.
- Error codes would be segmented in some way.
- Calls will not be segmented based on callee categories (prefix level) although there are 10 class levels. See the next slides.

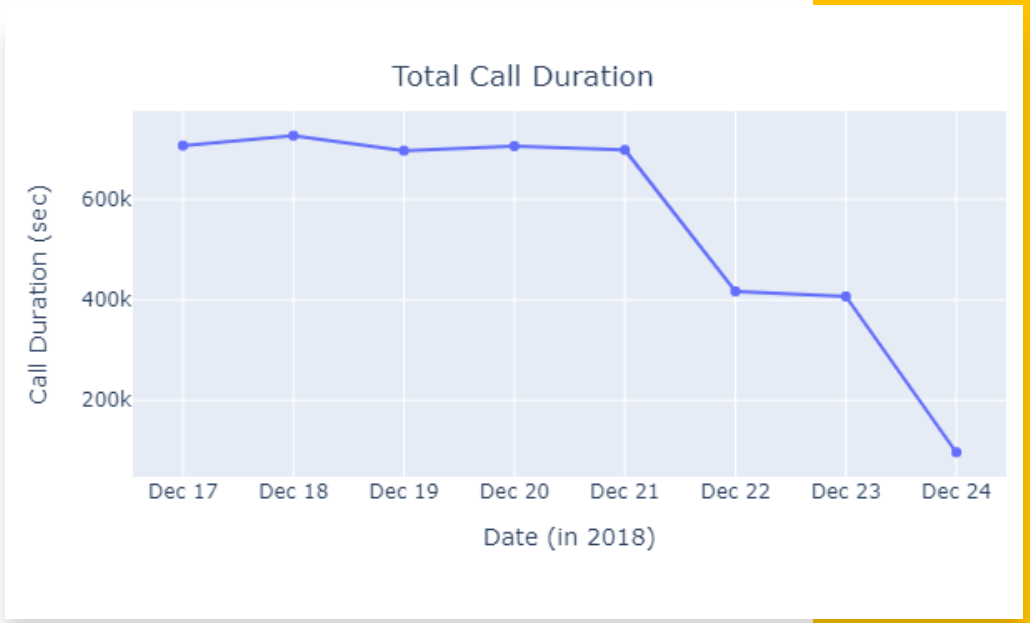
Class Levels

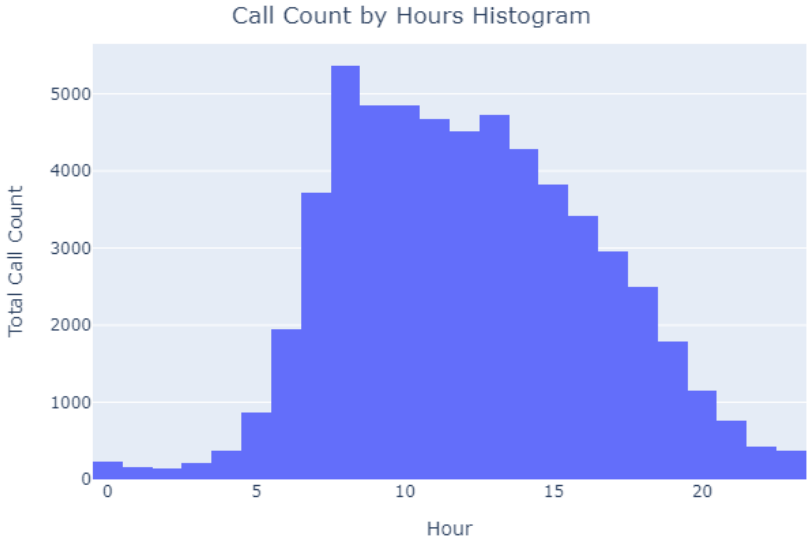
caller : 22039
callee : 32760
callee_country: 48
callee_prefix_level : 10
disposition : 4
src_ip : 5
dest_ip : 6
error_code : 18
trunk_id : 5
caller_group_id : 2
callee_group_id : 3
pair_group_id : 6

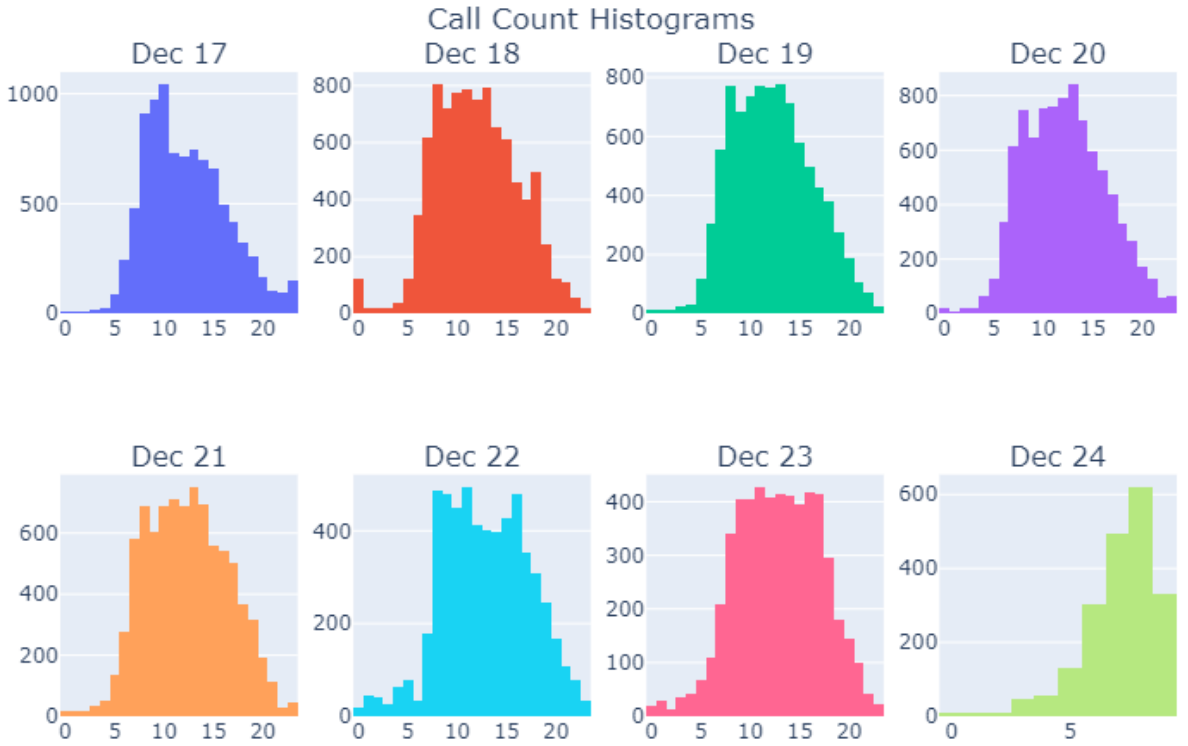
Call Counts and Durations by Days

- In first five week days no anomaly is observed.
- In the next two days call counts and mean call durations are less as expected in weekends.
- In the last day, there are much less records compared to December 17. The reasons may be that data sharing was done this way or the Christmas holiday week.







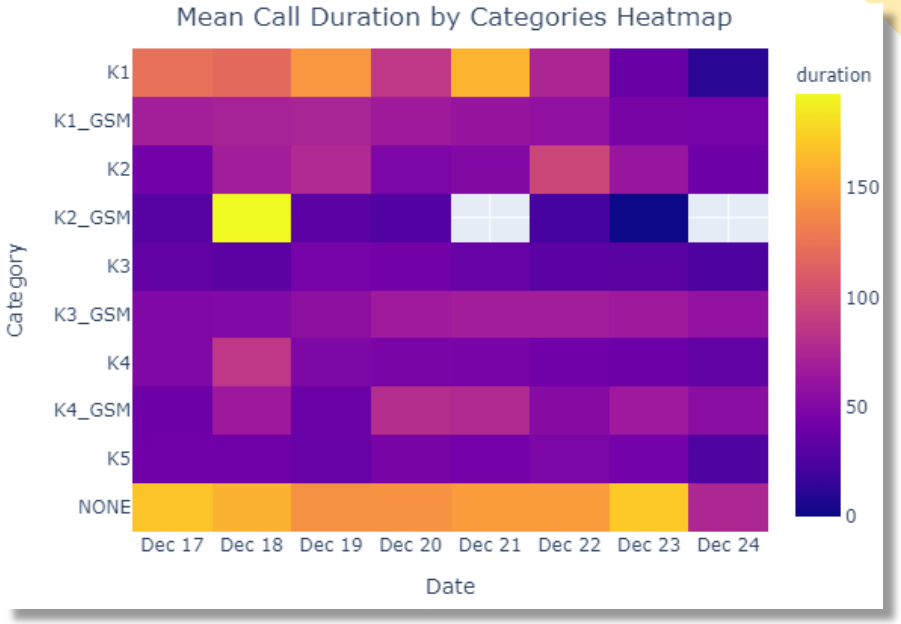


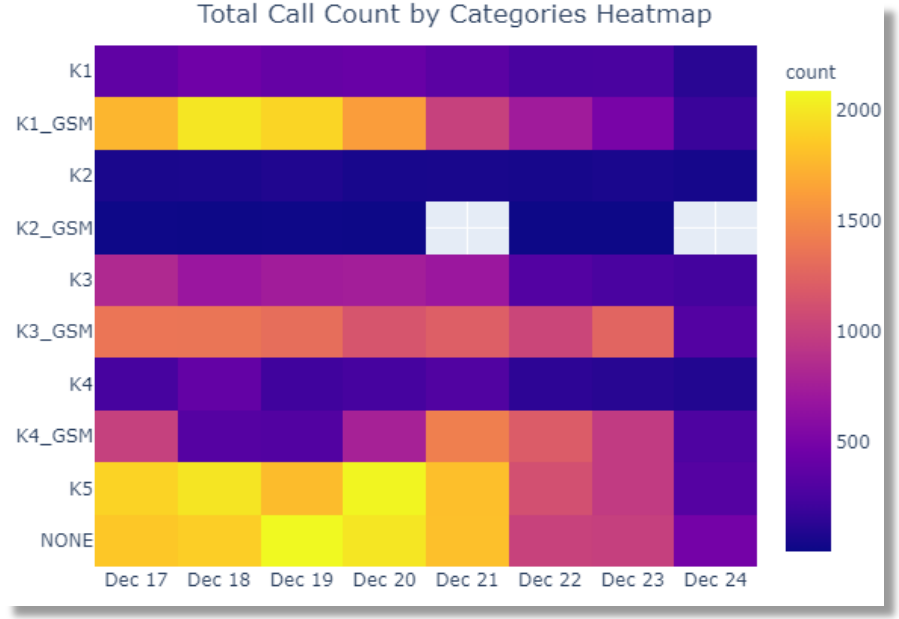
Call Count Distribution

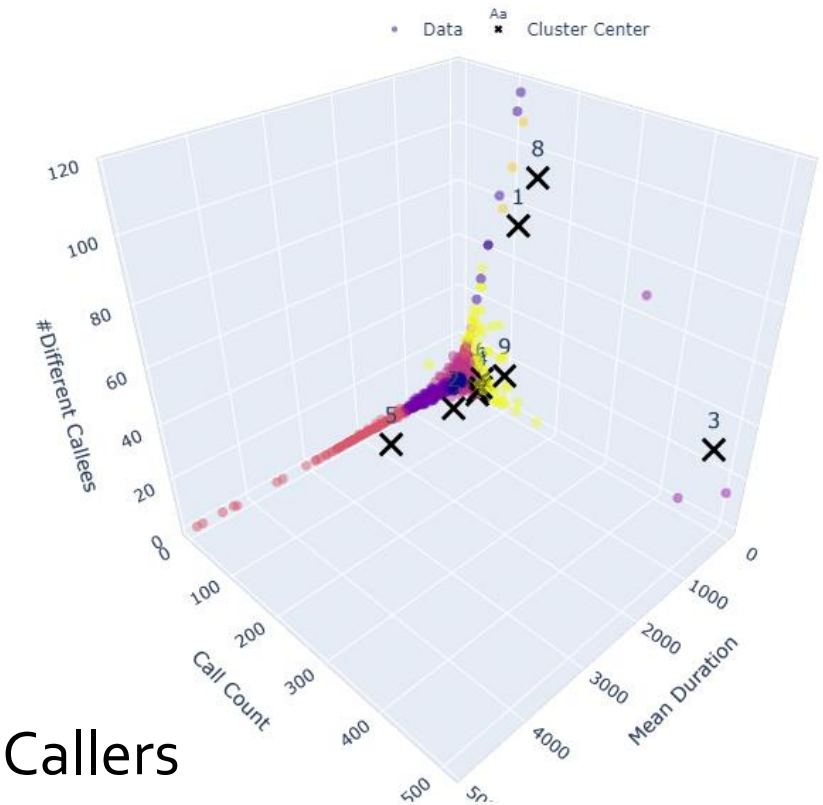
- Number of calls in 24 hours are normally distributed.
- No remarkable difference in call distribution in each day except the last day. There is no record after 9 a.m. in December 24.

Callee Categories

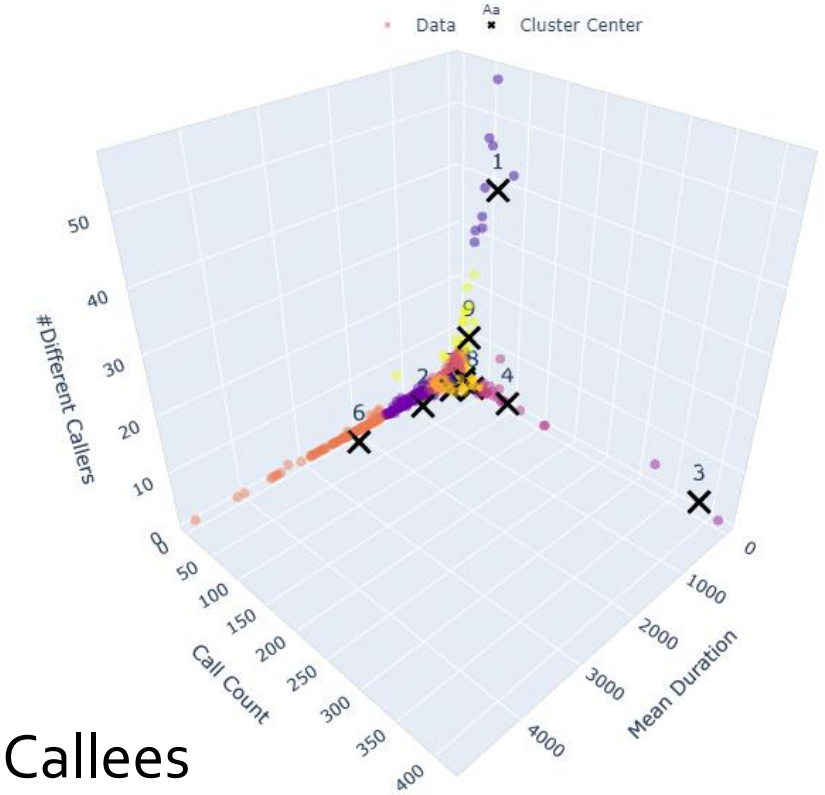
- There are lots of records whose callee category value is null.
- Mean call durations for K1 type callees are more than other types.
- There are more calls for K1 GSM and K5 numbers than other callee categories.
- No call records for K2 GSM callees in December 21 and December 24.
- In December 18, K2 GSM callees have much more mean duration than other days and other callee categories.
- There is no remarkable change in call durations by days.
- There is no enough similarity within similarly named categories, for example, K1-K1 GSM, K2-K2 GSM, …; or K1-K2, K2-K3, … So, calls should not be segmented within callee categories.







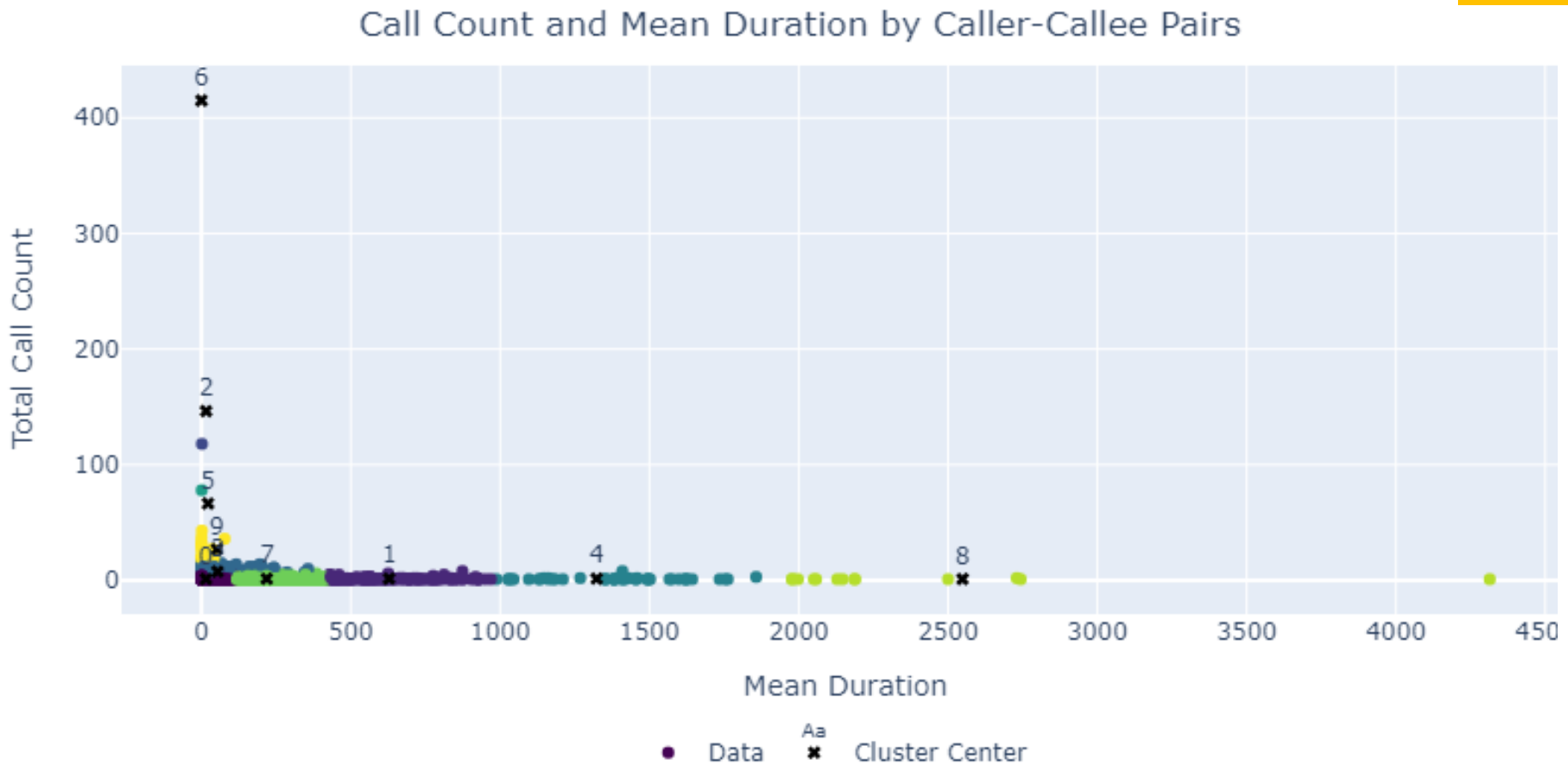
Callers



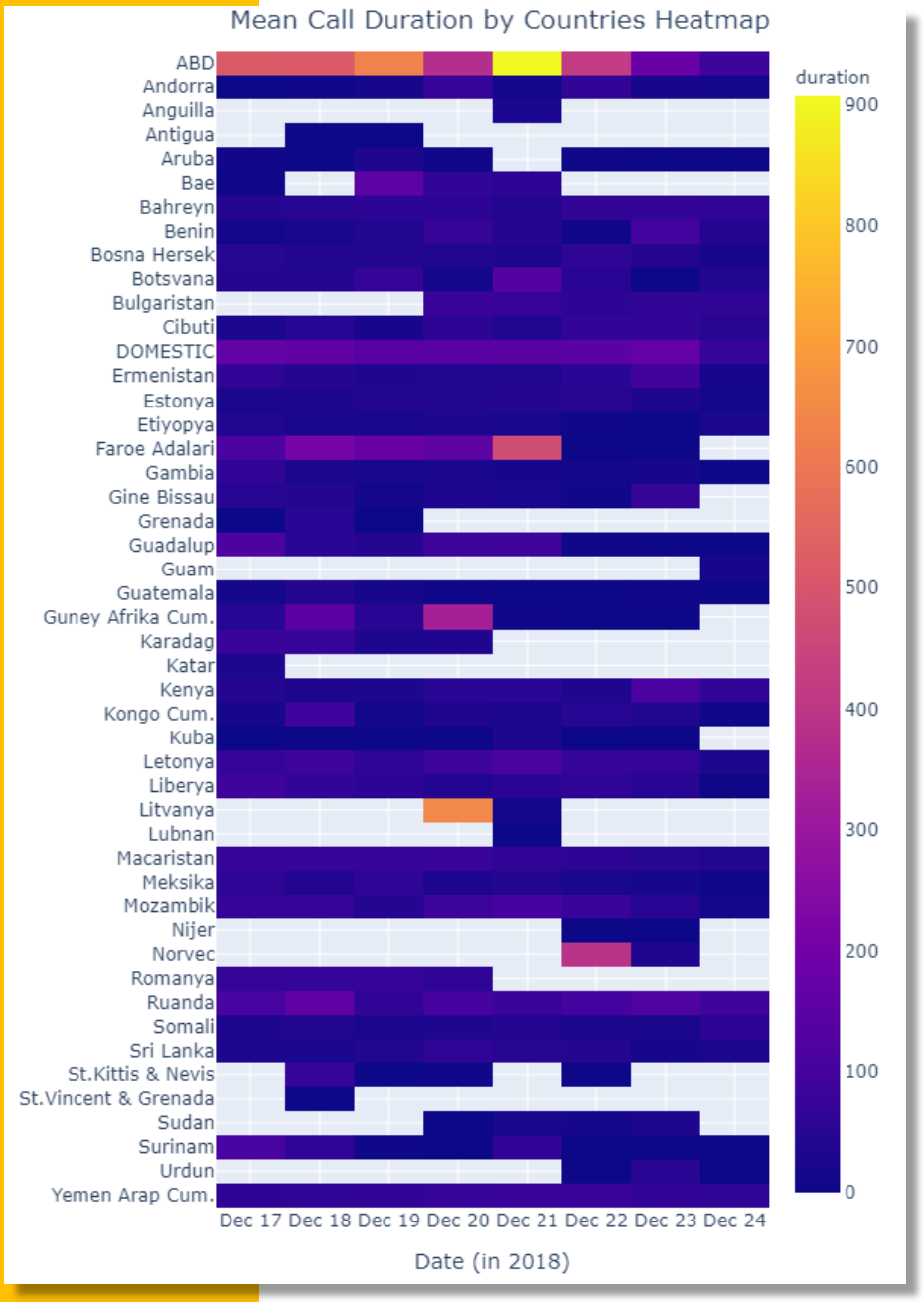
Callees

Segmentation of Callers and Callees

- 4 variables for caller segmentation: mean duration, total call count, number of different callees, and number of different callee countries
- 3 variables for callee segmentation: mean duration, total call count, and number of different callers

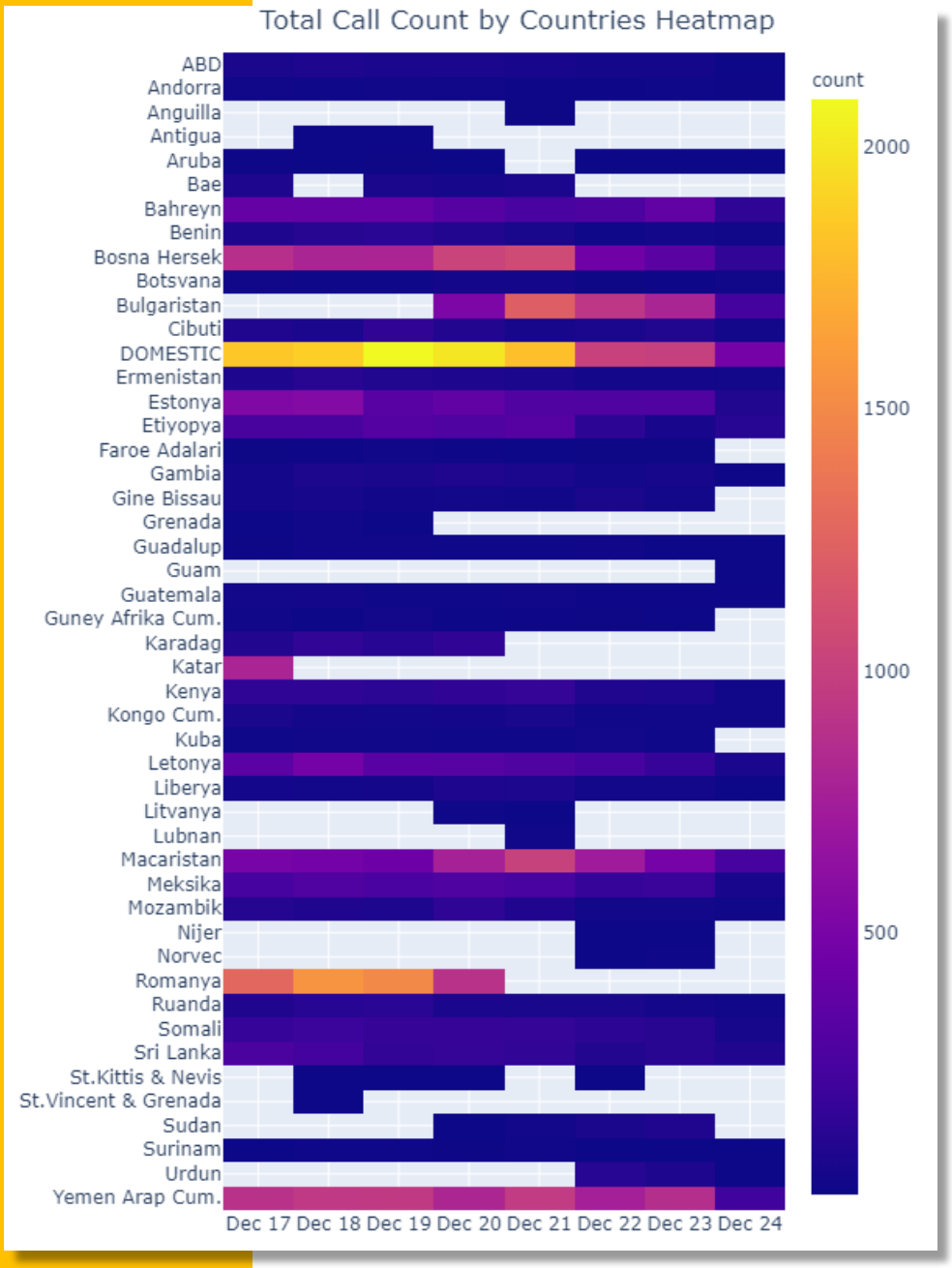


- Some outliers are observed when count and mean duration of calls between same caller and callee are scattered.
- So, those pairs are segmented into 10 clusters and it is introduced to the records as a new feature: pair_seg.



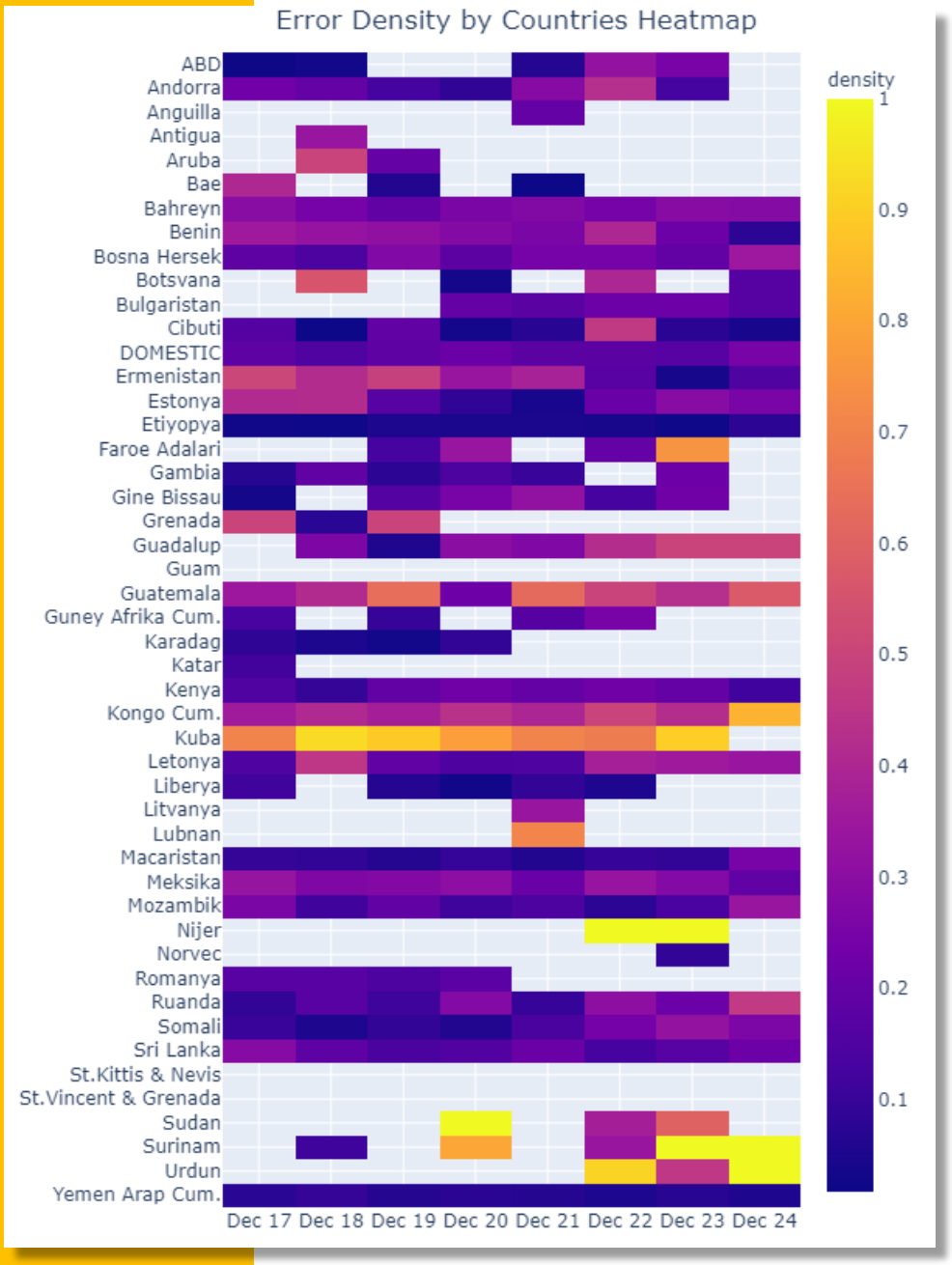
Mean Duration of Callee Countries by Days

- The USA calls took more longer.
- Faroe Islands, South Africa, Lithuania, and Norway calls in some days have more duration than others.



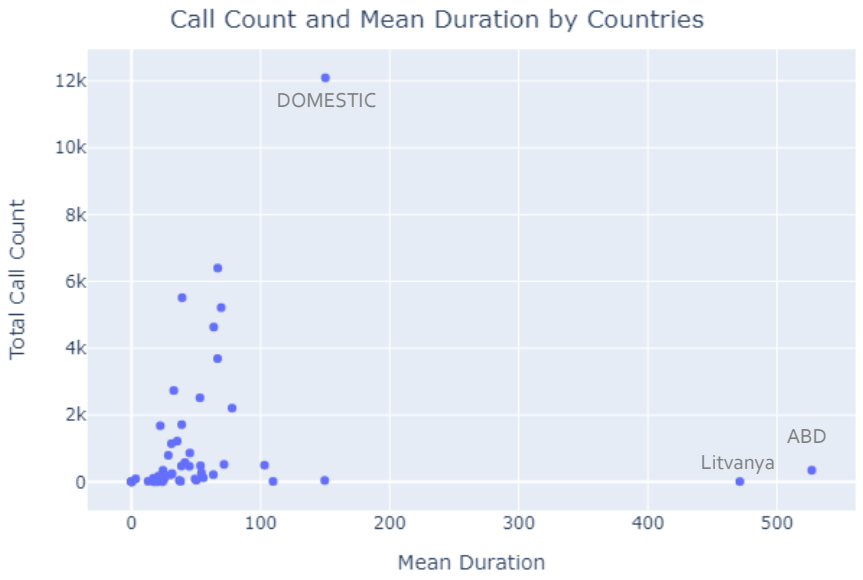
Call Counts of Callee Countries by Days

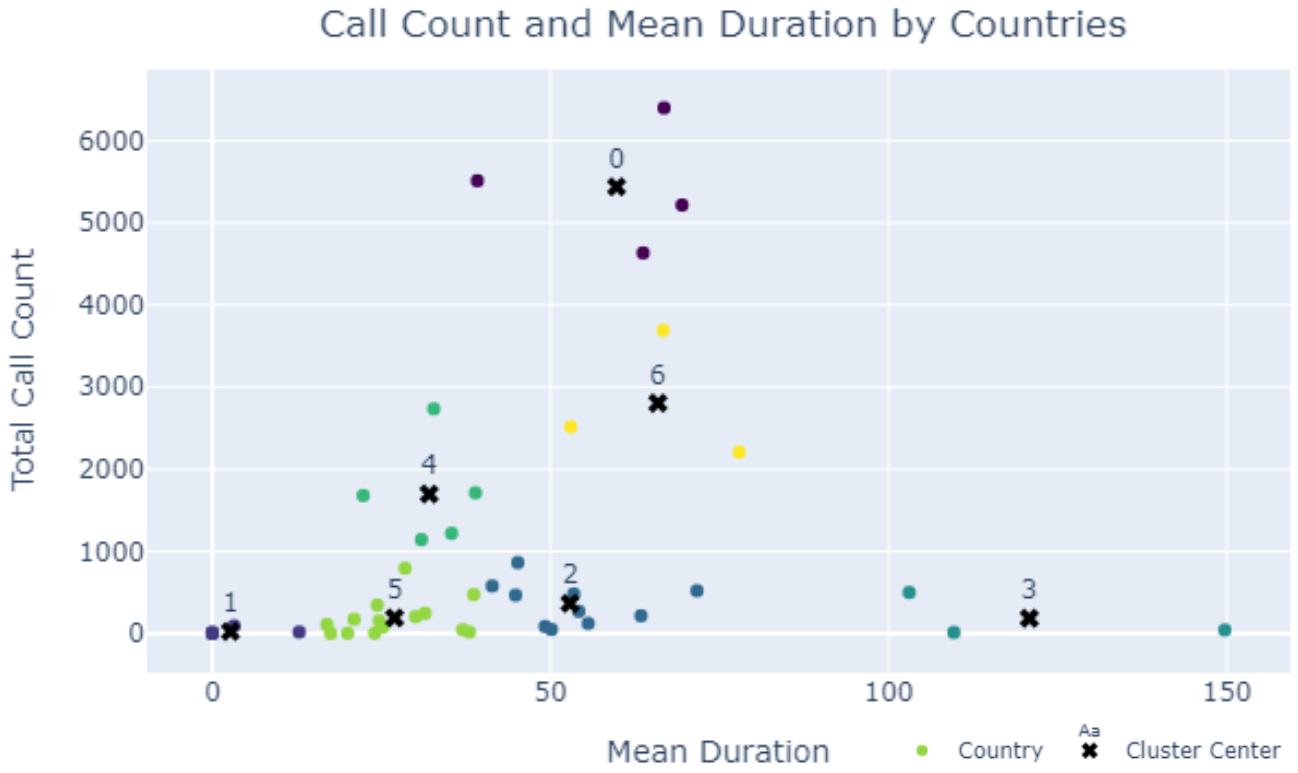
- There much more domestic (inbound) calls than outbound calls. So, inbound calls should be inspected separately than outbound calls.
- There more calls for Bosnia and Herzegovina, Romania, Georgia, and Yemen than other countries.
- Call records is very rare in most of the countries called.



Error Density of Callee Countries by Days

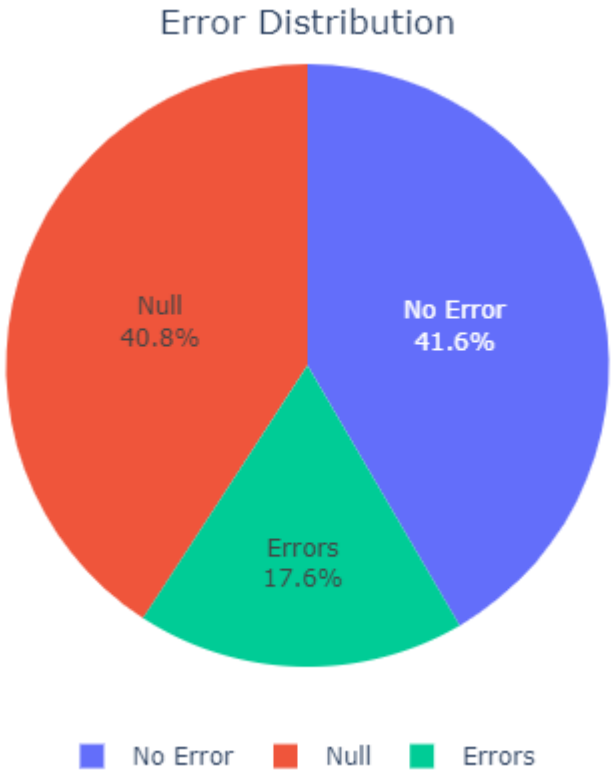
- In some countries in some days any error is encountered during calls.
- Guatemala, Congo and Cuba calls are encountered with high rates of errors in all days.
- Nigeria, Sudan, Surinam and Jordan calls in some days are almost always encountered with errors.

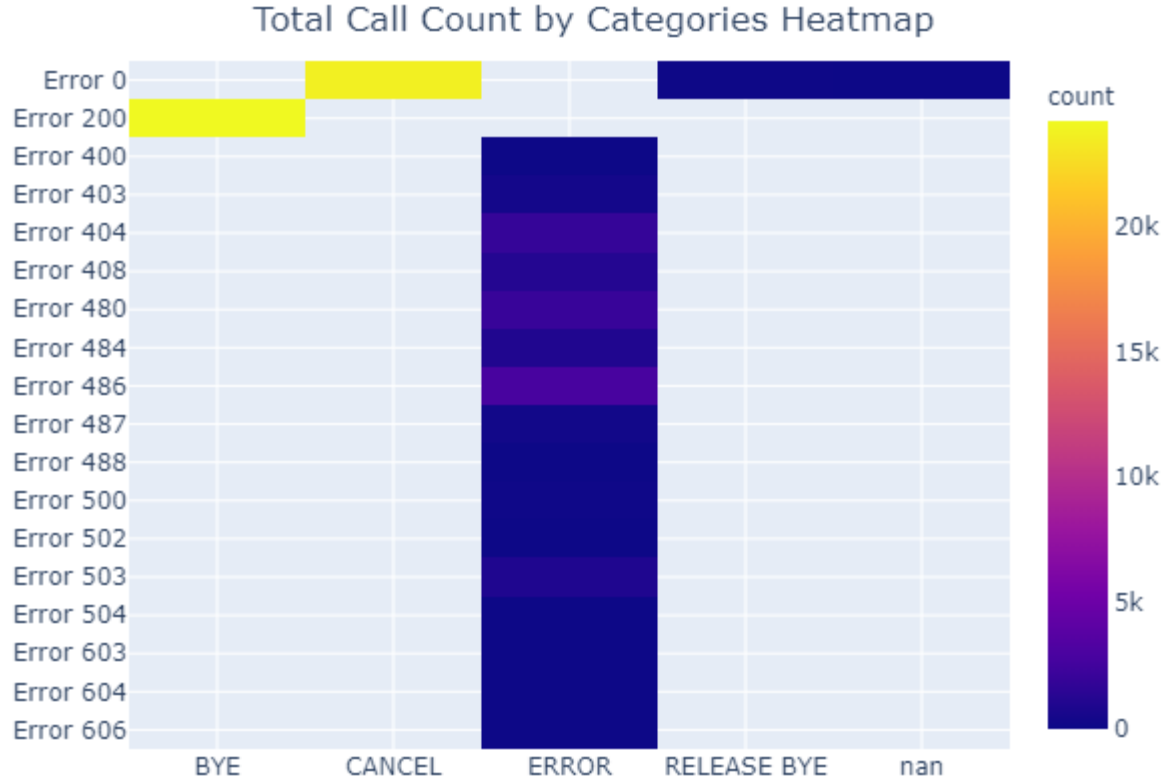




Segmentation of Callee Countries

- Callee country feature has too much class levels for data learning. Therefore 48 different countries are segmented into 9 clusters.
- Lithuania ,the USA calls are in the same segment, domestic calls are seperate, and other coutries are segmented as seen in the right scatter plot.





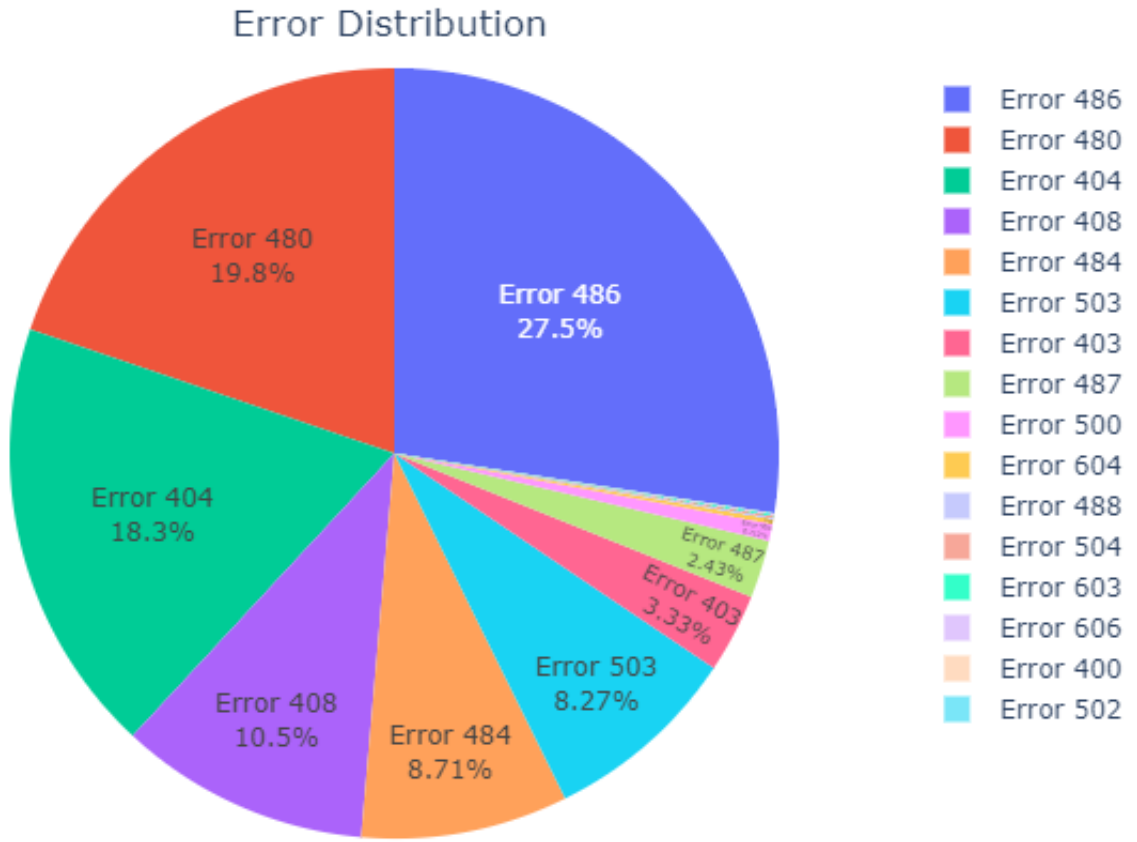
Error Distribution

- Huge part of the data has null values (Error 0) for error code. They can be imputed when the error code-disposition heatmap is inspected.
- There are some calls ended by CANCEL or RELEASE BYE. So, their null error codes are replaced with error code 200, which means no error in call.
- Also, nan values in disposition is replaced with ERROR since their end time is null as in errored calls.

Error Distribution

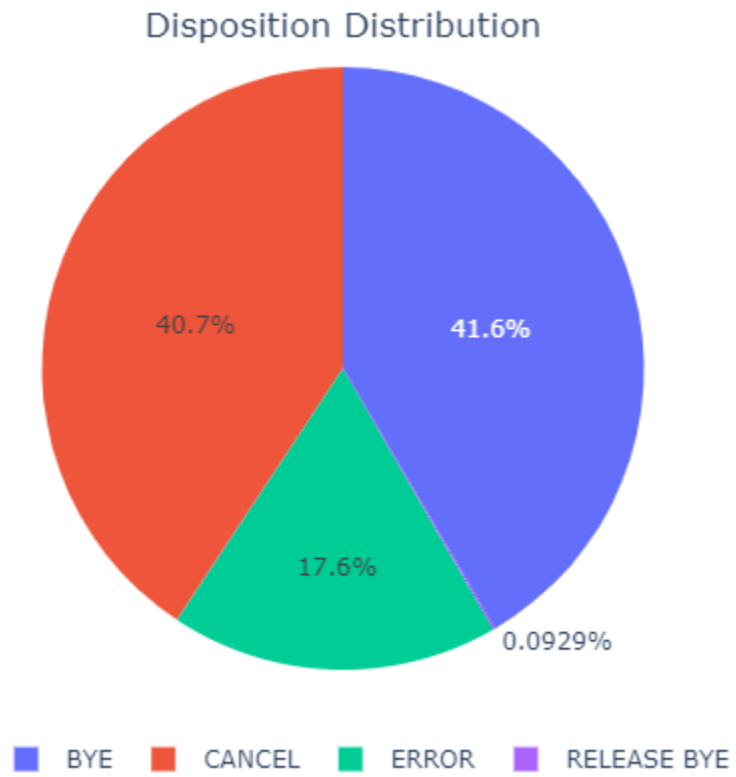
Value Counts

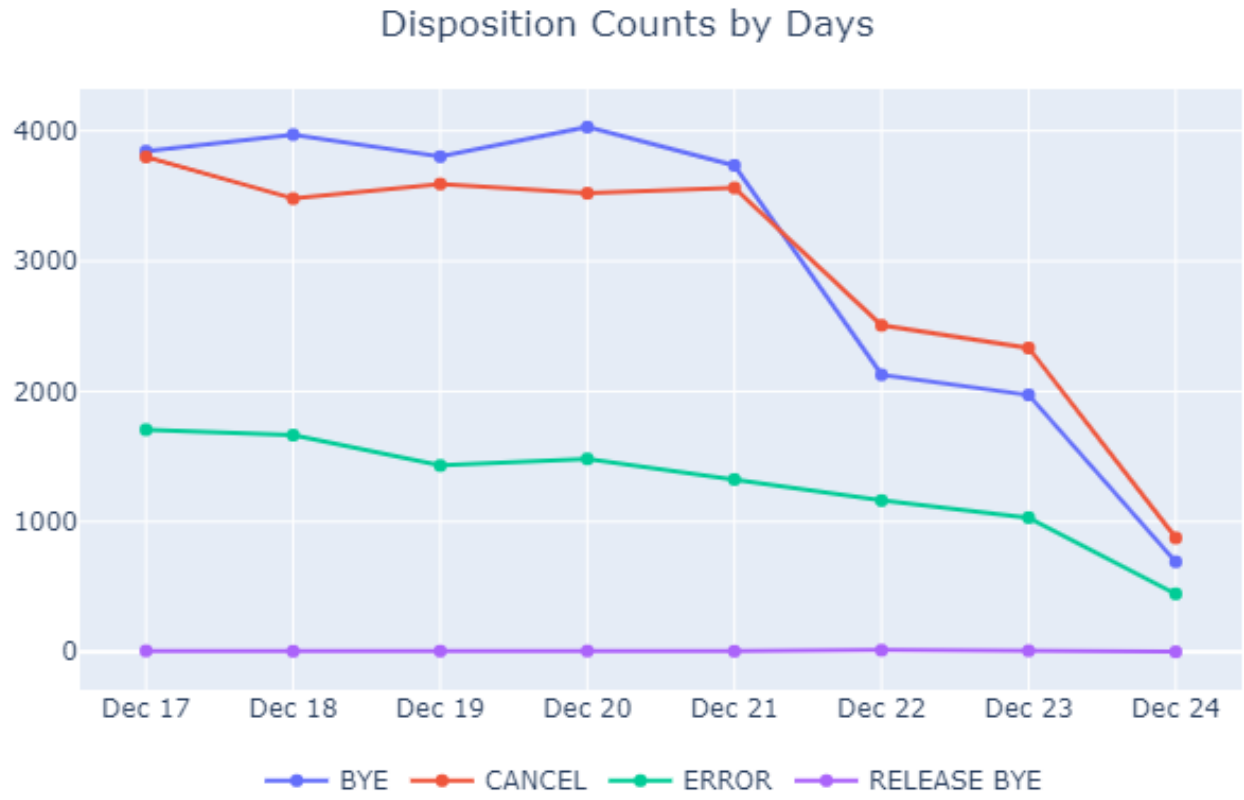
error_code	
502	1
400	2
606	4
603	5
488	7
504	7
604	20
500	73
487	249
403	341
503	847
484	892
408	1079
404	1876
480	2027
486	2816



- Rarely encountered, less than 50, error types are grouped as 'other' so that number of class levels in error code feature is reduced to 10.

Disposition Distribution





No anomaly or problem is observed.

Anomaly Detection in CDR

- ML models for inbound and outbound calls data is seperately trained.
- Training – test partition: **70% - 30%**
- For outbound calls:

  # variables after encoding categorical variables: 84

  #variables after removing correlated variables: 71

- For inbound calls:

  #variables after encoding categorical variables: 69

  #variables after removing correlated variables: 62

#variables: 16

Numerical Features

duration
call hour
call day

Categorical Features

callee country
callee prefix level
disposition
source IP
destination IP
error code
trunk ID
caller group ID
callee group ID
pair group ID
callee country segment
caller segment
callee segment
caller-callee pair segment

Unsupervised and Supervised Models for Anomaly Detection

- **Isolation Forest** models are deployed to detect anomalous outbound and inbound calls since they perform well in high dimensions compared to One-Class SVM and other one-class classifier models.
- **Extreme Gradient Boosting, Random Forest, and Extra Trees** models are trained based on the labels by Isolation Forest.
- They have similar scores but they have many different predictions for the test data. So, these 3 models are combined into a new model computing average prediction probabilities of them, called **Mean Classifier**.
- **Recall** score is more important in anomaly phenomena since false negative prediction is more costly.

For Outbound Calls

Number of different predictions between models:

	train	test
XGB-RF	1.0	43.0
XGB-ExT	2.0	67.0
RF-ExT	0.0	42.0

For Inbound Calls

Number of different predictions between models:

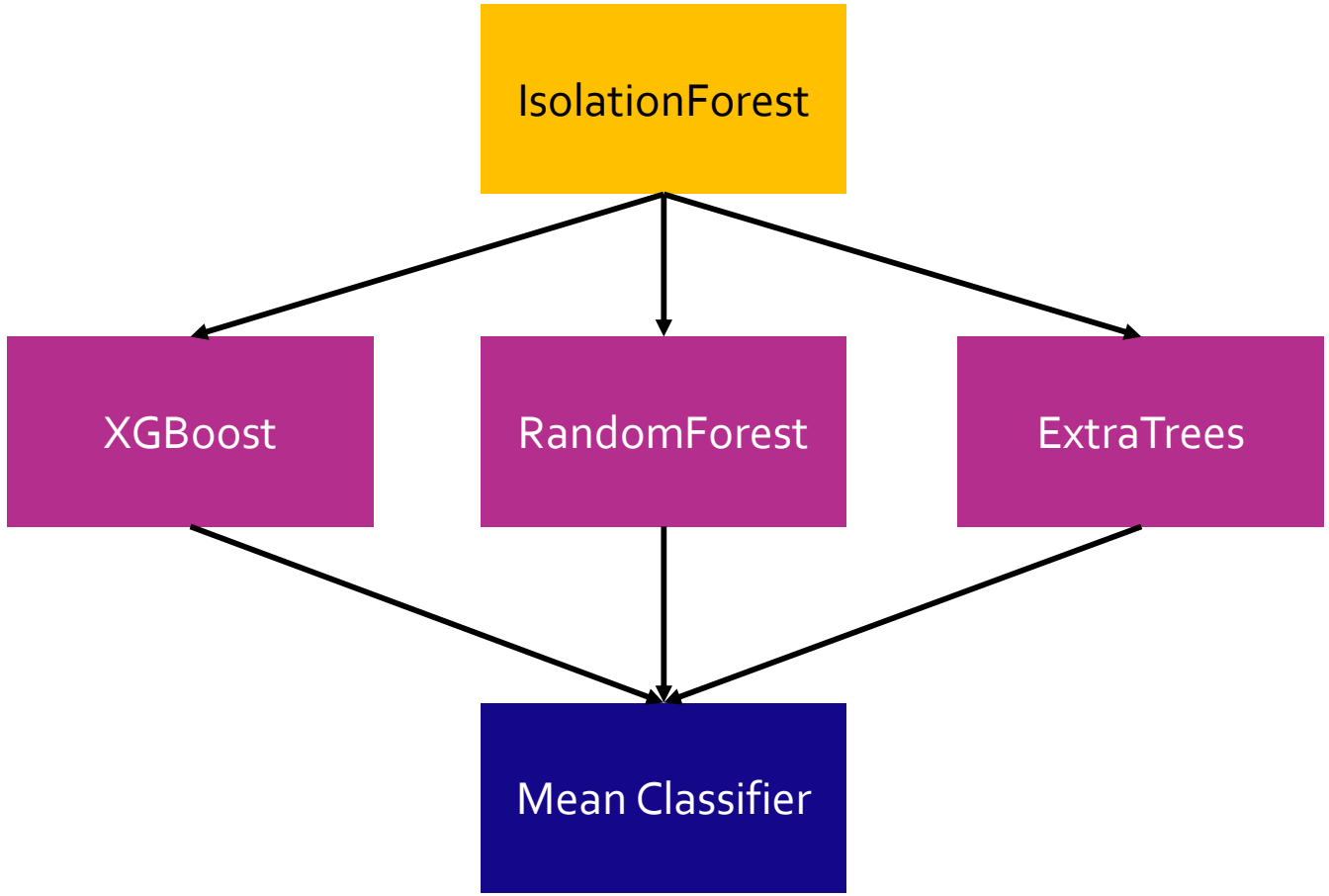
	train	test
XGB-RF	0.0	31.0
XGB-ExT	0.0	33.0
RF-ExT	0.0	18.0

Composed ML Model for Anomaly Detection

Unsupervised Learner

Supervised Learners

Combined Predictor



Scoring Supervised Learners

Classification Report for XGBoost Classifier on Test Outbound Calls

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13261
1	0.96	0.92	0.94	555
accuracy			1.00	13816
macro avg	0.98	0.96	0.97	13816
weighted avg	1.00	1.00	1.00	13816

Classification report for RandomForest Classifier on Test Outbound Calls

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13261
1	0.97	0.89	0.93	555
accuracy			0.99	13816
macro avg	0.98	0.95	0.96	13816
weighted avg	0.99	0.99	0.99	13816

Classification Report for ExtraTrees Classifier on Test Outbound Calls

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13261
1	0.95	0.89	0.92	555
accuracy			0.99	13816
macro avg	0.97	0.94	0.96	13816
weighted avg	0.99	0.99	0.99	13816

Classification Report for the Mean Classifier on Test Outbound Calls

	precision	recall	f1-score	support
0	0.99	1.00	0.99	3425
1	0.93	0.89	0.91	203
accuracy			0.99	3628
macro avg	0.96	0.94	0.95	3628
weighted avg	0.99	0.99	0.99	3628

Models for Outbound Calls

Classification Report for XGBoost Classifier on Test Inbound Calls

	precision	recall	f1-score	support
0	0.99	1.00	1.00	3425
1	0.93	0.90	0.92	203
accuracy			0.99	3628
macro avg	0.96	0.95	0.96	3628
weighted avg	0.99	0.99	0.99	3628

Classification report for RandomForest Classifier on Test Inbound Calls

	precision	recall	f1-score	support
0	0.99	1.00	0.99	3425
1	0.97	0.83	0.89	203
accuracy			0.99	3628
macro avg	0.98	0.91	0.94	3628
weighted avg	0.99	0.99	0.99	3628

Classification Report for ExtraTrees Classifier on Test Inbound Calls

	precision	recall	f1-score	support
0	0.99	1.00	0.99	3425
1	0.95	0.87	0.91	203
accuracy			0.99	3628
macro avg	0.97	0.93	0.95	3628
weighted avg	0.99	0.99	0.99	3628

Classification Report for the Mean Classifier on Test Inbound Calls

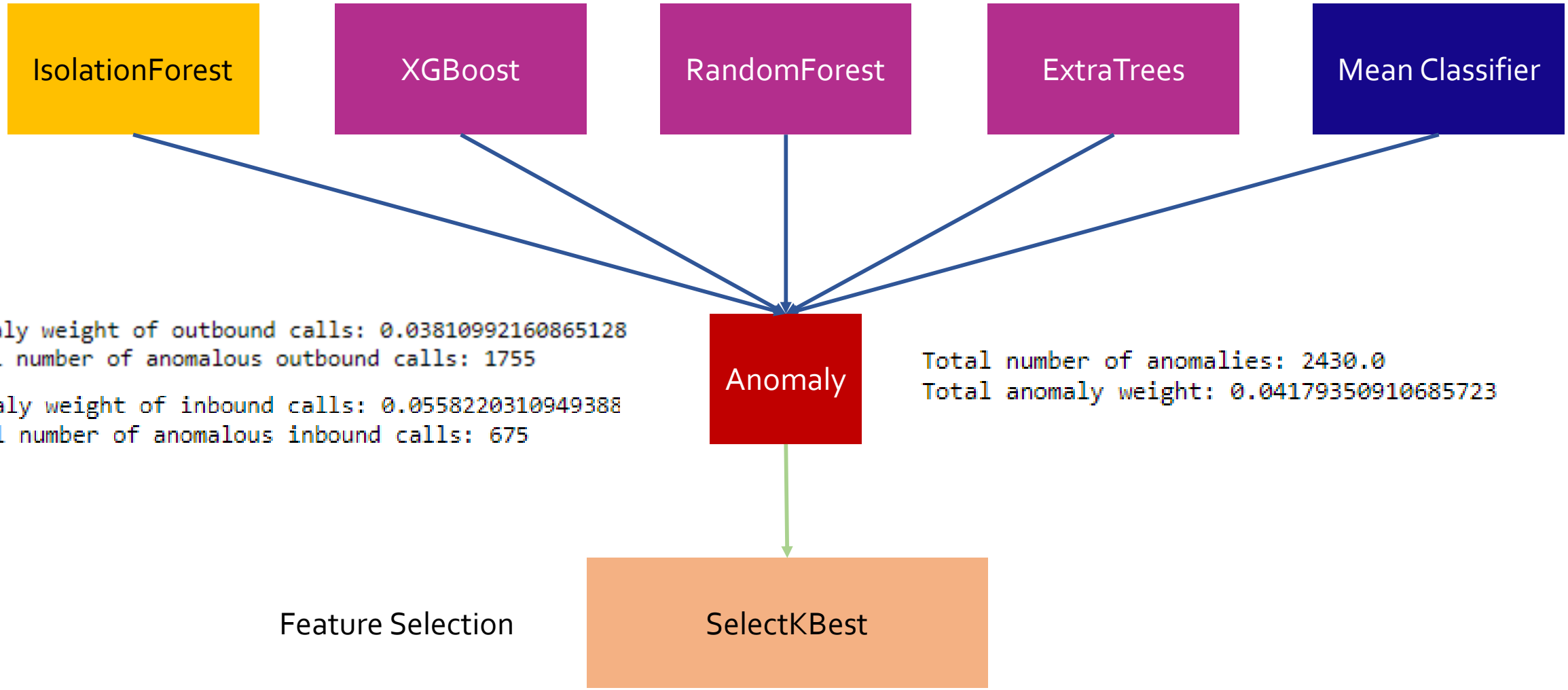
	precision	recall	f1-score	support
0	0.99	1.00	0.99	3425
1	0.93	0.89	0.91	203
accuracy			0.99	3628
macro avg	0.96	0.94	0.95	3628
weighted avg	0.99	0.99	0.99	3628

Models for Inbound Calls

Labelling Calls Based on the Anomaly Score and Feature Selection

- **Anomaly score** = sum of the predictions by 5 models.
- If anomaly score >= 3 then the **final anomaly label** is 1, else 0.
- Feature importances of 3 supervised models may not be reliable since target variable represents a unbalanced phenomena and the most of the variables are one-hot encoded categorial variables. See feature importance rates of the 3 models in Jupyter Notebook.
- **SelectKBest** algorithm is used to extract important features over their **ANOVA F-value**s.

Labelling Calls Based on the Anomaly Score and Feature Selection



Important Features for Detection of Anomalous Calls

- When we used the best **30** features scored by SelectKBest for training a Gradient Boosting model, we can get similar scores compared to the oubound models using all **71** features, the inbound models using all and **62** features.

Classification Report for XGBoost Classifier-2
with the 30 features on Test Outbound Calls

	precision	recall	f1-score	support
0	0.99	1.00	0.99	13261
1	0.90	0.84	0.87	555
accuracy			0.99	13816
macro avg	0.94	0.92	0.93	13816
weighted avg	0.99	0.99	0.99	13816

Classification Report for XGBoost Classifier-2
with the 30 features on Test Inbound Calls

	precision	recall	f1-score	support
0	0.99	0.99	0.99	3425
1	0.90	0.80	0.85	203
accuracy			0.98	3628
macro avg	0.94	0.90	0.92	3628
weighted avg	0.98	0.98	0.98	3628

FOR OUTBOUND CALL ANOMALY

The first 15 important features out of the best 30

	importance
caller_seg_2.0	14019.622376
caller_seg_3.0	9142.689643
pair_seg_1	7712.341539
callee_seg_3.0	4673.796916
duration	4118.809213
pair_seg_0	3669.522568
callee_seg_2.0	3623.260199
callee_seg_0.0	3508.723363
caller_seg_0.0	2286.963138
callee_seg_7.0	1873.699140
error_code_486	1706.373701
pair_seg_4	1342.889768
disposition_CANCEL	1300.026282
callee_country_seg_7.0	1271.378313
disposition_ERROR	1207.125732

How much reliable or consistent important features are based on the results?

Anomaly densities in each class level of the top important categorical variables:

```
3.0    0.833333
2.0    0.600000
5.0    0.215517
4.0    0.037187
7.0    0.029864
9.0    0.027144
8.0    0.010417
6.0    0.009690
1.0    0.003795
0.0    0.002939
Name: caller_seg,
```

```
2.0    0.540373
3.0    0.492578
6.0    0.228261
7.0    0.184658
4.0    0.106454
9.0    0.100000
8.0    0.070927
5.0    0.015341
0.0    0.001386
1.0         NaN
Name: callee_seg,
```

```
6    0.874699
1    0.574365
4    0.514286
5    0.096226
9    0.087262
7    0.078626
8    0.039216
3    0.024819
0    0.001016
2         NaN
Name: pair_seg,
```

```
7.0    0.391667
3.0    0.236559
4.0    0.067358
5.0    0.054034
6.0    0.047896
2.0    0.026732
0.0    0.012179
1.0         NaN
Name: callee_country_seg,
```

FOR INBOUND CALL ANOMALY

The first 15 important features out of the best 30

	importance
callee_group_id_1	0.022138
callee_group_id_4	0.019040
error_code_200	0.017748
caller_seg_3.0	0.016536
callee_seg_0.0	0.016027
pair_seg_0	0.015298
pair_group_id_44	0.015194
disposition_ERROR	0.014965
src_ip_GoEiEsiGGiGE	0.014477
src_ip_GoEiEsiEiJB	0.013012
dest_ip_GoEiEsiEiJB	0.012697
duration	0.011957
caller_seg_0.0	0.011766
caller_seg_5.0	0.010831
pair_group_id_11	0.009862

How much reliable or consistent important features are based on the results?

Anomaly densities in each class level of the top important categorical variables:

```
4    0.210830
1    0.035771
Name: callee_group_id,

44    0.489627
14    0.152098
41    0.044400
11    0.033811
Name: pair_group_id,

5.0    0.492537
3.0    0.251455
8.0    0.099174
9.0    0.085586
2.0    0.051869
4.0    0.038337
0.0    0.018040
7.0         NaN
Name: caller_seg,

4    0.314516
5    0.282528
2    0.222910
9    0.135819
8    0.119403
1    0.052009
7    0.045866
3    0.045089
0    0.019644
Name: pair_seg,

4.0    0.311973
6.0    0.229167
8.0    0.129236
5.0    0.118005
2.0    0.072848
7.0    0.055863
0.0    0.033442
1.0         NaN
9.0         NaN
Name: callee_seg,
```