ALY 6080 – INTEGRATED EXPERIENTIAL LEARNING
Winter 2020 Quarter

Fraud Detection

Goal

To build a ML fraud detection model that is cost effective and dynamic in nature to be used by GE to save the analyst from heavy work while not miss any high-risk incidents.

Focus

Our area of focus is to study the various factors leading to IP Theft and using it build a dynamic robust ML model.

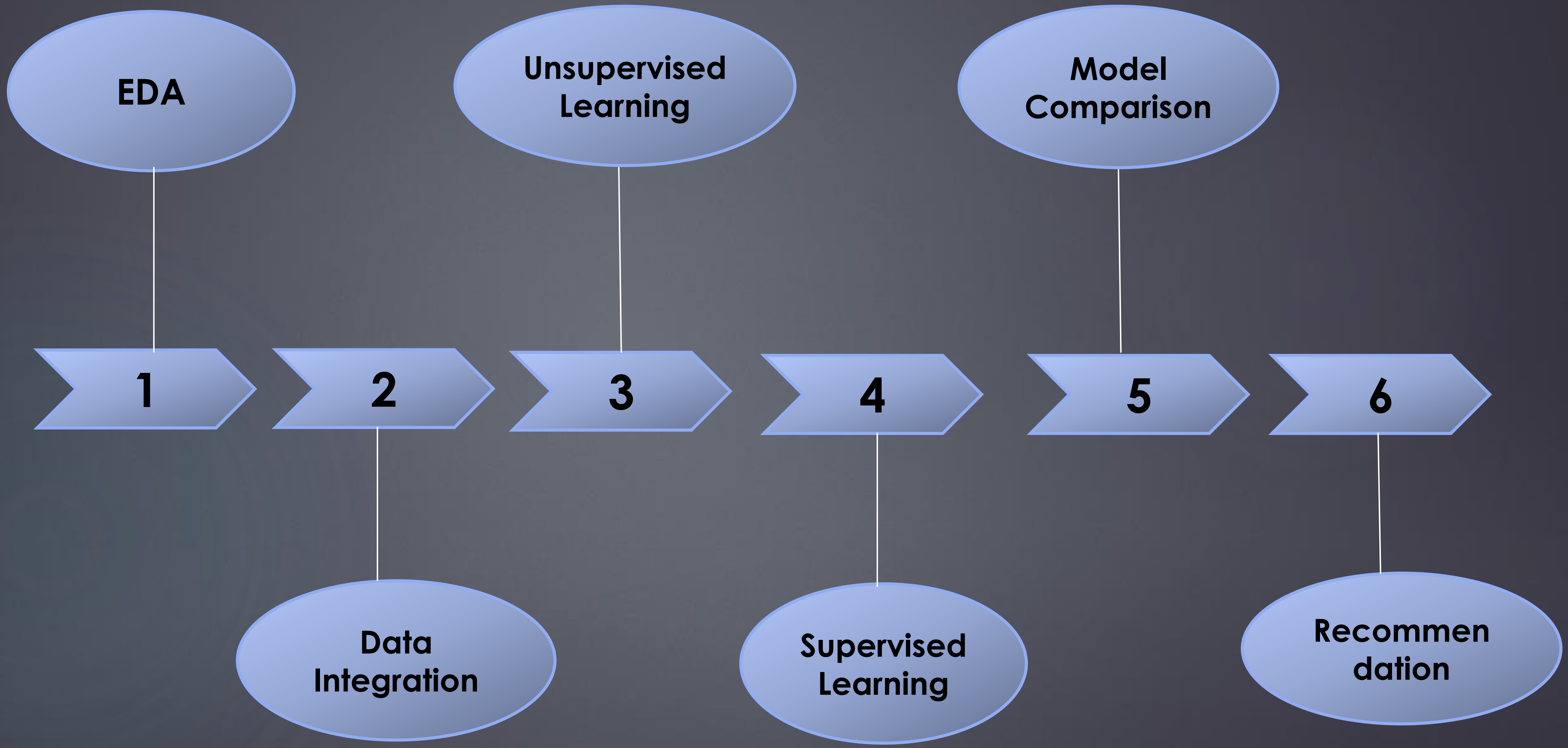
Techniques

Unsupervised Learning
- K-Mode

Sampling
- Over Sampling
- Under Sampling

Supervised Learning
- Logistic Regression
- Decision Tree
- Naive Bayes
- Random Forest
- XGBoost

Roadmap Of The Project



Data Summary

Data Files	Description	No of Rows	No of Columns	Missing Values	Important Columns
Heat_Score.csv	▪ static table along with indicator and values	176	6	32	Heat Value SHARED_INDICATION_NAME SHARED_INDICATOR_TYPE
Indicators_data.csv	▪ Details of different type of alerts generated for the employees with escalated date(2018-2019) classified as TP/HIGH, TP/LOW,TP/DE,FP by analyst. ▪ It contains heat scores for employee ids with escalated date(2018-2019, risk factors, alert type (weekly, monthly, atomic and daily), alert category(atomic, heat), average scores. ▪ Classification, alert category and indicators are assigned by analyst and all rest variables are assigned by the system. ▪ It contains duplicate records of alert being re-classified by analyst	99246	14	None	Classification Owner_name Score risk_factor Avg_score hru alert_category alert_type indicator
Indicators_pair_updated.csv	▪ This is similar to indicators_Data file with updated data of employees with pairs of indicators fired for an employee . ▪ Has no duplcate data ▪ there are mainly two types of alert category – 1. Atomic - This is a single indicator when fired send an immediate alert 2.. Heat – These are the indicators when fired, a heat score value is added to the employee's accumulated heat score. Once the accumulated heat score reaches a threshold, an alert is raised. It is further classified into Daily, Weekly, and monthly alerts.	132079	14	2 in alert type 258 in owner name.	Classification Owner_name Score risk_factor Avg_score hru alert_category alert_type indicator
Obfuscated_data.csv	▪ It gives us the general employee information such as type(contractor, functional, employee), id, hire date, status(active, inactive), employee type, job function, function group(function enabling, production), career band and industry focus name with country details.	2356	11	403 inge_hire_date 102 city/state/country 101 - function_group	Employee_id Job_function Career_Band

Data Integration

Heat_Score

Indicators_pairs

Obfuscated_demo_data

Join Column : Indicators

Join Column : Employee_id

Resultant Table Columns

score
owner_name
employee_id
risk_factor
avg_score
classification
hru
alert_category
alert_type
indicators
person_type
person_status
ge_hire_date
country_name
function_group
job_function
career_band
industry_focus_name
ge_hire_year
employee_tenure (today date – ge_hire_date)

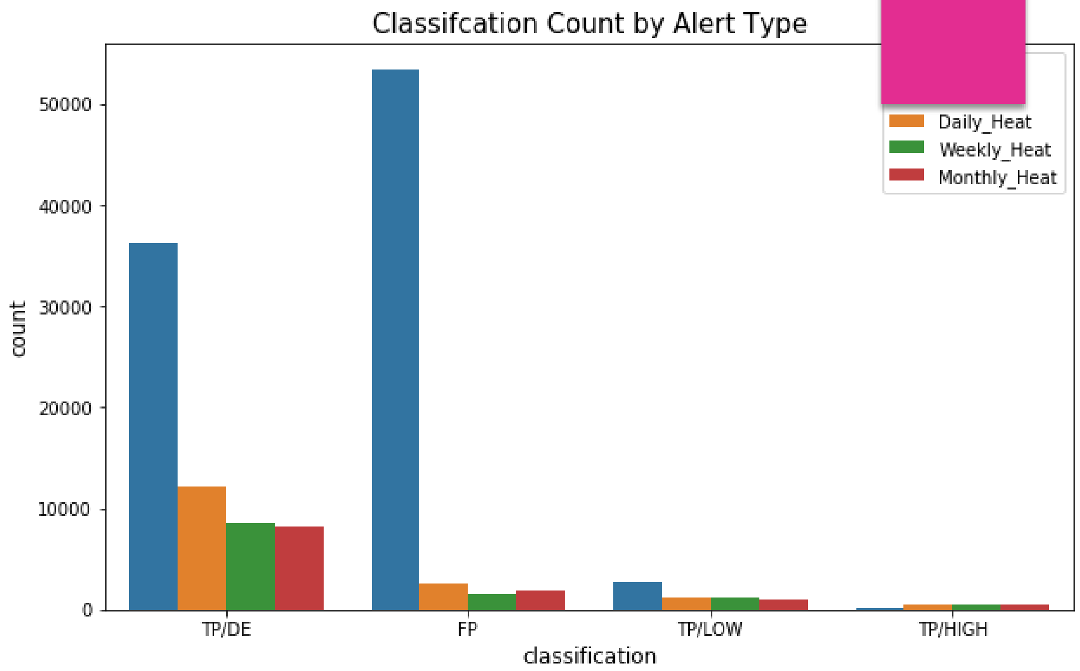
Exploratory Data Analysis

- 94% of alert fired are FP or TP/DE, we have class imbalance.
- More Atomic alerts are fired as compared to Heat alerts.
- Heat Score threshold for firing the alerts for an employee

  Daily_Heat : 171

  Weekly Heat : 451

  Monthly_heat : 721

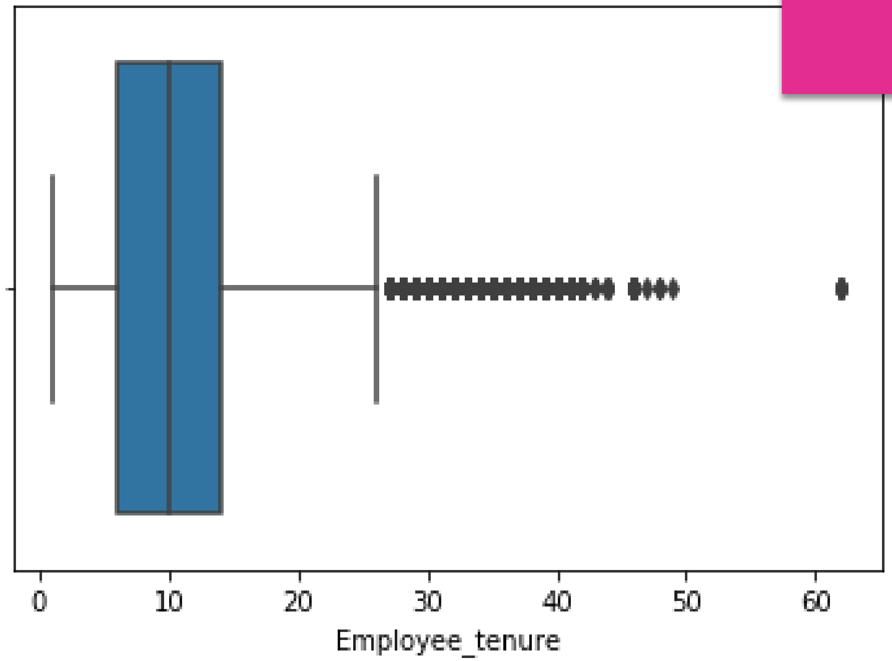


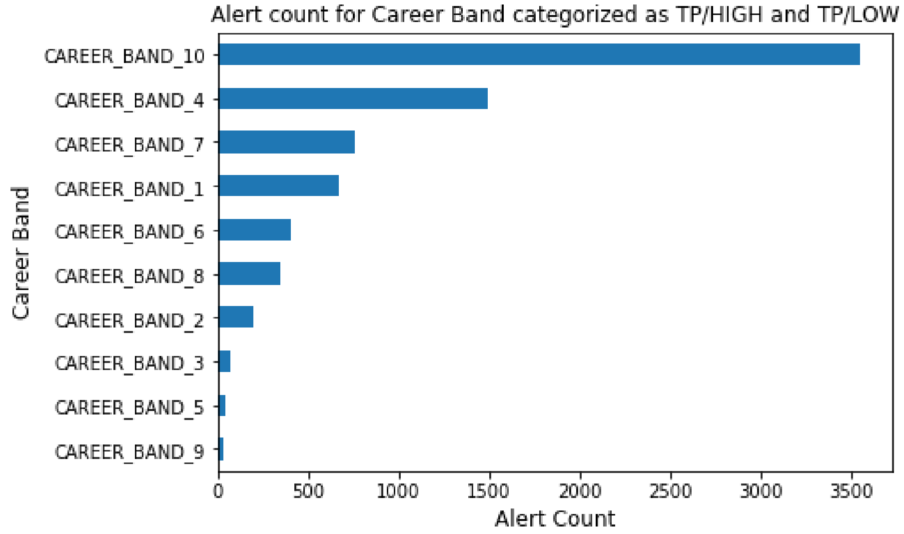
Heat Score Statistics

Alert Category	Alert Type	Min. Score	Max. Score	Min. Avg Score	Min. Risk Factor	Count of Number of ..
Heat	Daily_Heat	171	432,440	100	2	16,379
	Monthly_Heat	721	1,147,968	90	9	11,428
	Weekly_Heat	451	606,739	90	6	11,681
Other	NA	3	3	-1	-1	2

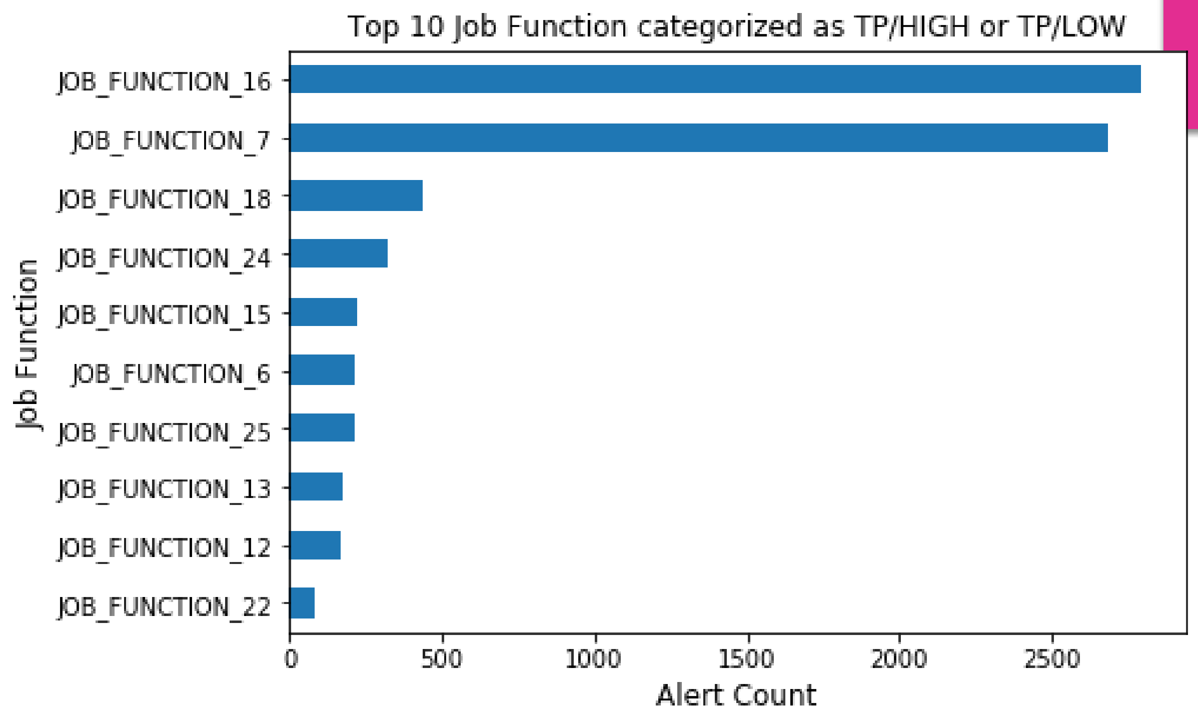
Exploratory Data Analysis

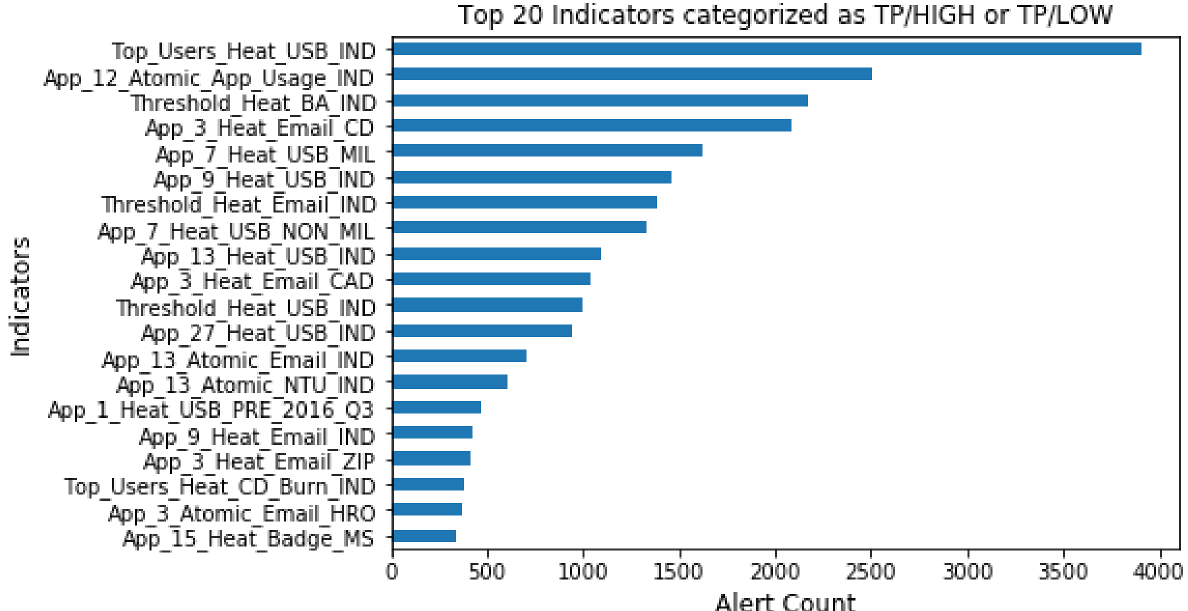
- Most of the employee's tenures working with GE is 5 to 15 years.
- They are few employees with tenure > 30.
- Most of the notable alerts(TP/HIGH +TP/LOW) came from Career Band 10,4,7





Exploratory Data Analysis





- Employee with job function as 16,7,18,24 have generated more notable alerts.
- Although highest count alert fired is Top_users_Heat_USB_IND is categorized as TP/HIGH or TP/LOW, its heat value is 0.
- The heat alerts which are fired the most have low heat value ranging from 0-5

Unsupervised Learning

Unsupervised Model

- Unsupervised learning is a type of machine learning algorithm which is used to draw inferences like finding hidden pattern or grouping in input data without labeled responses.

Why Unsupervised Learning

- Analyzing large datasets manually is very costly.
- To discover groups of similar examples within the data and thus getting an insight on features contributing more towards the cluster.

Since we are dealing with majority of categorical variables, among the many methods available for unsupervised learning, we have used the **K-Modes** clustering technique.

K-Mode Clustering

K-Mode is an extension to K-Means but Instead of distances, it uses dissimilarities (that is, quantification of the total mismatches between two objects: the smaller this number, the more similar the two objects). And instead of means, it uses modes.

Steps performed for building K-Mode cluster

- Imputed employee tenure using a decision tree model for the missing values using the variables function group and career band
- Split the indicator_pairs column to multiple rows with unique indicators and dropped the duplicate rows.
- Performed one hot encoding on the categorical variables.
- Performed K-Mode clustering on the resultant data for 4,5,6,8,10 clusters.
- Analyzed all the clusters and selected the one which has highest percentage of TP/High and TP/LOW using the below formulae

$$\%Attribute, cluster\ i = \Sigma(Attribute\ value, cluster\ i)/\Sigma(Attribute\ value\ for\ cluster\ i..n)$$

Where,attribute is the variables/columns in the resultant dataset

i is the particular cluster

n is the total cluster count

Clustering Insights for Atomic Alerts

- 1. By using the visual inspection method, the best insights were obtained using 5 clusters as this cluster has the highest % of TP/HIGH _TP/LOW (93%)clusterd .
- 2. The table depicts the features that account for more than 30% of its total.
- 3. It's quite interesting to observe that, most of the High+Low alert came from the contractor and functional persons who are inactive now, were as employee who are functional were clustered more towards non-risk alerts.
- 4. An employee working with Higher Risk Unit are more prone to theft.
- 5. It seems, the missing data for HRU, the country is also clustered to TP/HIGH and TP/LOW, GE aviation team should try to gather more data for the missing information.

Cluster	FP	TP/DE	TP/HIGH	TP/LOW	Owner Name	HRU	person_type	country	career_band
0	3%	32%	63%	30%	Analyst_1,Analyst_4 ,Analyst7,Analyst 8, No Data(Missing Value)	HRU_19, HRU_18 ,No_Data,HRU_13	Contractor, Functional	Brazil, Mexico, No_Data, India,Australia ,US	1,5,2,6,9,7
1	13%	51%	30%	16%	Analyst_1,Analyst_4 ,Analyst_6,Analyst 8(Missing Value)	HRU_12, HRU_13 ,HRU_18	Employee	Canada, China , Australia, Poland, Emirates,US	4,7,2
2	35%	1%	0	0.03	Analyst_5, Sr Analyst_1	No Data	Employee, Functional	US	8
3	48%	14%	7%	5%	Analyst_5, Sr Analyst_1	HRU_11, HRU_1	Employee	US	10
4	1%	3%	0	46%	Analyst_6, Analyst_7	HRU_2	NA	Singapore	7

Cluster	indicators	Job_function	Person_Status
0	App_13_Atomic_Email_IND,App_3_Atomic_Email_SS,App_13, Atomic_CD_Burn_IND, indicators_App_3_Atomic_Email_HRO, Threshold_Heat_Email_IND,App_13_Atomic_DVD_Burn_IND	13,25,22,19,12,5 ,24,10,6,16,8,14	I
1	App_1_Atomic_Email_PRE_2016_Q3 App_9_Threshold_Access_>8, App_9_Threshold_Access_Confidential>15,App_9_Threshold_Access_Confidential>5	3,17,10,7,8	I, A
2	App_9_Atomic_DVD_Burn_IND,App_13_Atomic_DVD_Burn_IND	7,8,19	A
3	App_9_Atomic_DVD_Burn_IND, App_13_Atomic_DVD_Burn_IND, App_9_Threshold_Access_Confidential>15, App_9_Threshold_Access_Confidential>5	7,1	A
4	App_12_Atomic_App_Usage_IND, indicators_App_25_Atomic_TT_IND	18	I

Clustering Insights for Daily Heat Alerts

Daily Heat Alerts

Cluster	FP	TP/DE	TP/HIGH	TP/LOW	Owner Name	HRU	Person_Type	Job Function	Career_Band	Indicators	Insights
0	0	0.21	0.04	0.08	Analyst_7, Analyst_8,NoData, Analyst6	8,18,19	Employee	14,10,17,16,18,19,25,21	7,8,3,2,10		
1	0.06	0.23	**0.05**	**0.22**	Analyst_6,Analyst_8,Analyst_7, Analyst1	None	Contractor	10,15,12,22,5,23	6	App_11_Heat_Access_IPI,Top_Users_Heat_CD_Burn_IND	
2	0.16	0.19	0.21	0.21	No Data, Sr. Analyst 1, Sr. Analayst 2, Analyst_6,Analyst_7,Analsyt_8,Analayst_4	2,11,13,12	Employee	13,10,12,19	5,8	App_11_Threshold_Access_100_Day,App_10_Threshold_Access_30_Day, App_10_Threshold_Access_90_Day, App_11_Heat_Access_IPI, App_11_Heat_Access_Pool,App_11_Threshold_Access_10_Day	
3	0.05	0.06	0.01	0.03	Analyst_8,Analyst_4,Analyst_6	0.19	Employee	16,3,8	3,7,8		
4	0.58	0.01	0	0.02	Sr.Analyst1, Sr.Analyst 2	11,12	Employee	4,5,12,7,8,2	4,5		
5	0	0.18	0	0.2	No Data,Analyst_8, Analyst_6, Analyst_7 , Analyst_4, Analyst1	5,7,11,12	Employee	1, 14, 12,18,21	2,8	App_11_Heat_Access_Restricted,App_10_Threshold_Access_7_Day , App_10_Threshold_Access_30_Day,App_10_Threshold_Access_90_Day	
6	0	0	**52%**	0	Analyst_6, Analyst_7, Sr.Analyst1, Analyst_8, Analyst_4	19,13	Employee	17,18,8,6,16,5	5,4,,7,8	App_27_Heat_NTU_IND,App_7_Heat_USB_MIL,Threshold_Heat_USB_IND,App_1_Heat_USB_PRE_2016_Q3,App_13_Heat_USB_IND,App_13_Heat_Box_IND,Top_Users_Heat_CD_Burn_IND ,App_9_Heat_USB_IND	
7	0.05	0.12	**0.06**	**0.15**	Analyst1, Analyst_4, Sr.Analyst 2	None, 19	Functional,Contractor	13,15,12	1	App_13_Atomic_CD_Burn_IND, App_13_Heat_Print_IND, App_2_Heat_CD_Burn_MIL,Threshold_Heat_NTD_EXE_IND, Top_Users_Heat_CD_Burn_IND Threshold_Heat_Print_IND App_2_Heat_CD_Burn_NON_MIL	
8	0.08	0	**0.15**	**0.09**	Analyst_4, Sr.Analyst 1	5,12	Employee	16,8,10	8	il_IND, App_14_Heat_Terminal_IND, App_3_Heat_Email_CAD,App_9_Atomic_DVD_Burn_IND	

1. Most of the indictaors in the cluster (highlighted in Green) are the ones with lower higher heat value in the range from 1-5. Since the minimum score to fire daily and weekly alert is 170, this suggests that these alerts are fired multiple times. If we could keep the track of the frequency of alerts being fired, we could accordingly add weights to the indicators which our model can then emphasize.

2. Most of the High and low alerts were analyzed by Analyst 6, 7,8 and Senior Analyst 1

3. An employee working with None, 19,13,12 Higher Risk Unit are more prone to theft.

Clustering Insights for Weekly Heat Alerts

Weekly Heat Alerts

Cluster	FP	TP/DE	TP/HIGH	TP/LOW	Owner Name	HRU	Person_Type	Job Function	Career_Band	Indicators	Insights
0	0	17%	**7%**	**2%**	Analyst7,	19,2,None	Employee	14,18,19,22	10,5,7	App_16_Threshold_Box_A, App_10_Threshold_Access_90_Day , App_11_Heat_Access_Pool, App_13_Atomic_Email_IND	1. Most of the indictaors in the cluster (highlighted in Green) are the ones with lower higher heat value in the range from 1-5. Since the minimum score to fire eeekly alert is 450 this suggests that these alerts are fired multiple times. If we could keep the track of the frequency of alerts being fired, we could accordingly add weights to the indicators which our model can then emphasize. 2. Most of the High and low alerts were analyzed by Analyst 6, 8,7,1. 3. An employee working with career Band 3,7,8 are more prone to theft.
1	4%	9%	**37%**	**27%**	Analyst_6, Analyst_7,Analyst_8, Sr.Analyst1, Analyst_4	18,13,None	Employee	14,18,19,21,24,25	3,4,7,8	App_13_Heat_Chrome_IND,App_13_Heat_Box_IND, App_3_Heat_Email_CD,Keywords_Heat_Service_MIL, Threshold_Heat_Print_IND App_13_Atomic_Email_IND,App_14_Heat_HR_IND,App_28_Heat_HR_Action, App_27_Heat_NTU_IND App_15_Threshold_Badge_MS , App_13_Heat_Box_IND	
2	3%	8%	0	11%	Analyst_4	13,12	Employee	12,19	5,8	App_10_Threshold_Access_30_Day, App_10_Threshold_Access_7_Day,App_10_Threshold_Access_90_Day App_13_Atomic_Email_IND	
3	0	7%	1%	4%	Analyst7, Analyst_8	19	Employee	19,21,26	3,7,8	Heat_Email_ZIP,App_1_Atomic_Email_PRE_2016_Q3, App_20_Threshold_App_Usage_30_Day	
4	4%	18%	**27%**	**8%**	Analyst_1, Analyst_6,Analyst_8	12,13,5	Employee	12,14	4,7,8	App_11_Heat_Access_Pool, App_11_Threshold_Access_10_Day ,App_20_Threshold_App_Usage_7_Day	
5	67%	0	0	2%	Analyst_6,Analyst_4,Analyst_1	5,13,12	Functional,Employee	14,22	8,1	App_1_Heat_Print_PRE_2016_Q3, App_1_Heat_Terminal_PRE_2016_Q3, App_11_Heat_Access_Pool	
6	20%	19%	0	**33%**	Senior_Analyst_2,Analyst_1, Analyst_4, Analyst_6	3	Contractor	5,15,10,21,3,26,25	1	App_1_Heat_USB_PRE_2016_Q3	
7	0	6%	**12%**	**4%**	Analyst_4,Analyst_7, Analyst_6	19	Functional	12,13,18,21,22,8,6	2	App_13_Atomic_Email_IND, App_1_Heat_USB_PRE_2016_Q3	
8	2%	16%	**16%**	**10%**	Analyst_4,Analyst_7	11,18,2	Employee	18,3,7	10,2,5	App_10_Threshold_Access_30_Day, App_10_Threshold_Access_7_Day,App_10_Threshold_Access_90_Day, App_11_Threshold_Access_30_Day App_11_Threshold_Access_10_Day App_11_Heat_Access_Pool, App_11_Heat_Access_Restricted	
9	0	1%	0	0	Analyst7	5	Employee	7	5,4	App_9_Threshold_Access_7_Day	

Clustering Insights for Monthly Heat Alerts

Monthly Heat Alerts

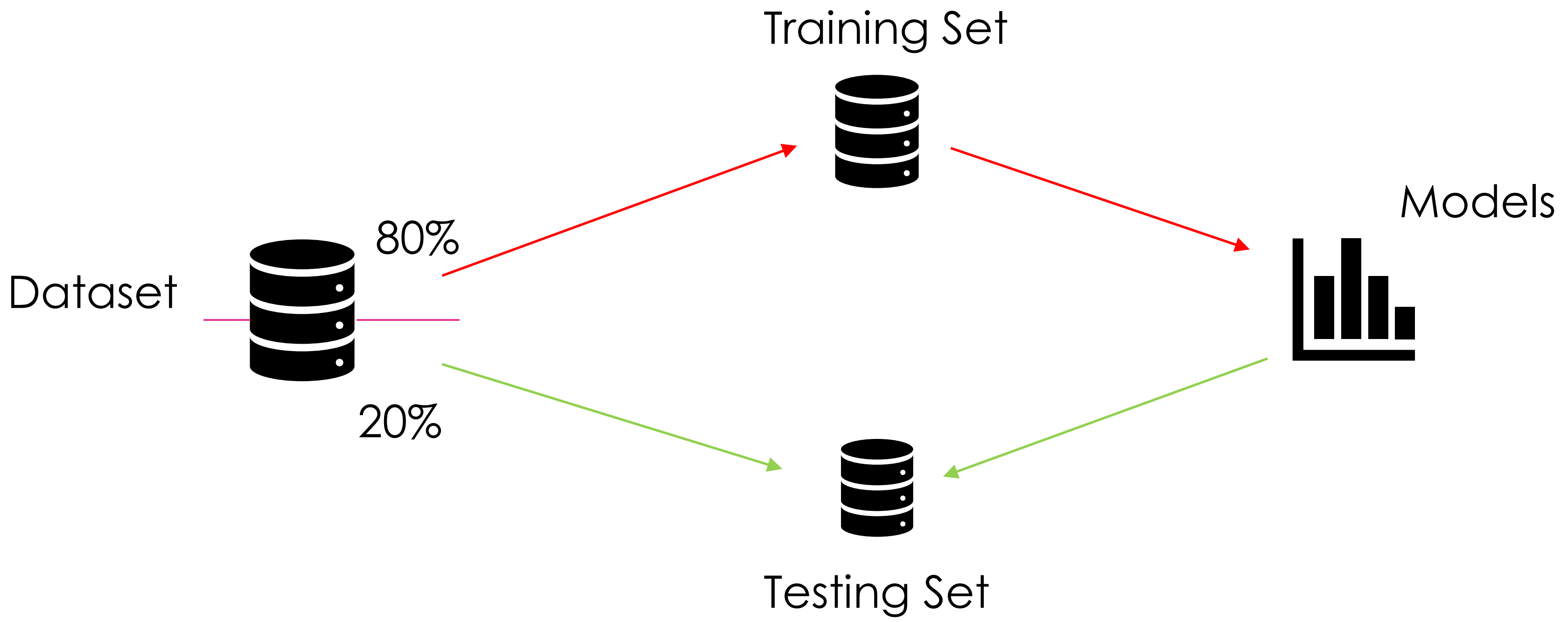
Cluster	FP	TP/DE	TP/HIGH	TP/LOW	Owner Name	HRU	Person_Type	Job Function	Career_Band	Indicators	Insights
0	8%	32%	**24%**	**0.13**	Analyst_1, Analyst_4,Analyst_7 ,Analyst_8,Senior_Analyst_2	5,13,12,19	Employee	12,14,21,1,6	4,3,1	App_13_Heat_Chrome_IND, App_13_Heat_Print_IND, App_27_Heat_NTU_IND	
1	15%	11%	4%	12%	Senior_Analyst_2, Analyst_1	19	Contractor	13,15	1	App_2_Heat_CD_Burn_NON_MIL,Top_Users_Heat_CD_Burn_IND	
2	2%	25%	**28%**	**29%**	Analyst_6,Analyst_4 ,Analyst_7,Analyst_8,Analyst_1	18,11,21	Employee	14,3,18	2,10	App_10_Threshold_Access_30_Day,App_22_Threshold_App_Usage_90_Day,App_22_Threshold_App_Usage_30_Day, App_13_Heat_Chrome_IND,Top_Users_Heat_CD_Burn_IND, App_3_Heat_Email_NON_MIL	
3	0	0.17	**0.11**	**0.12**	No_Data, Analyst_6,Analyst_1,Analyst_7	None	Employee	12,18,21	10	Threshold_Heat_NTD_EXE_IND,App_27_Heat_Email_IND,App_10_Threshold_Access_30_Day, App_10_Threshold_Access_7_Day	
4	0.15	0.02	**0.32**	**0.23**	Senior_Analyst_1, Analyst4, Analyst_1,Analyst_8	12,19,7,Non	Functional, Employee	16,1,22,24,25	3,7,1,8	App_14_Heat_HR_IND,App_13_Atomic_Email_IND,App_13_Atomic_NTU_IND,App_13_Heat_Chrome_IND, App_13_Heat_Print_IND, App_14_Heat_Terminal_IND, App_1_Heat_Print_PRE_2016_Q3,App_28_Heat_HR_Action,App_28_Heat_HR_Confirmed	1. Most of the High and low alerts were analyzed by Analyst 6, 4,7,1.
5	61%	0	0	2%	Senior_Analyst_1	12	Functional	18,8	8	App_4_Heat_CAD_IND,App_28_Heat_HR_Action,App_28_Heat_HR_confirmed ,App_28_Heat_HR_Action, App_14_Heat_Terminal_IND	2. An employee working with career Band 3,1,10 are more prone to theft.
6	0	14%	0	9%	Analyst_4	18,19,21	Employee	19,1,21	5,10,7	App_10_Threshold_Access_30_Day,App_13_Atomic_NTU_IND,App_27_Heat_NTU_IND, App_27_Heat_Box_IND	3. It seems Functional Employee could also be involved in theft.

Modelling

STEPS

- Data Preparation for modeling.
- Converting the classification label to 1 for TP/HIGH and 0 for TP/LOW+TP/DE+FP.
- Exclude the features which are not required.
- Convert the categorical variable in the form of 1 and 0 i.e. one-hot encoding on all the categorical variables.
- Separate data into training data and test data such that 80% randomly goes into training and 20% into test data set
- Use training data set to fit the prediction models.
- Use the model parameters from above to predict the values of the outcome for the test data and then for the whole data again.
- Assess the accuracy of the models.
- Repeat the above steps (2-4) for Notable Risks after converting the classification label to 1 for TP/HIGH +TP/LOW and 0 for FP +TP/DE.

Model Building



Models

1. Logistic Regression
2. Decision Tree
3. Naïve's Bayes
4. Random Forest
5. XGBoost

Model accuracy for Atomic Alerts

To **measure the accuracy of the model** we have focused on the following parameters

- Confusion matrix
- Recall for High Risk and notable– It is the percentage of True positives predicted by the model out of the actual True positives.
- $Recall = (True\ Positive)/ (Total\ Actual\ Positives)$

- Both the models performed poorly in predicting the 1(TP/HIGH). The reason for the poor performance is the class imbalance in the data. 0.2% of total atomic alerts are classified as TP/HIGH and hence the model is overfitting to the majority class i.e. non-TP/HIGH.
- To deal with class imbalance would be performing the following sampling techniques.
- **Synthetic Minority Oversampling Technique** (SMOTE) – In this method, we would oversample the minority class. It works by creating synthetic observations based upon the existing minority observations (Chawla et al., 2002).
- **Under-sampling** – In this majority class samples randomly and uniformly so that the majority class is 10 times more than the minority class, instead of using the entire majority class. This can potentially lead to loss of information. But if the examples of the majority class are near to others, this method might yield good results.
- **Oversampling and Undersampling** – In this we would first oversample the minority class and then undersample the majority class.

Model	No Sampling		
	Accuracy	Recall (% of TP High Correctly Predicted)	Confusion Matrix
Logistic Regression	0.99	0	pred:0 pred:1 true:0 92381 0 true:1 223 0
Decision Tree	0.99	0.26	pred:0 pred:1 true:0 92371 10 true:1 165 58

Atomic Alerts : Model Accuracy for High Risk Alert (TP/HIGH)

Model	Over Sampling			Under Sampling			Over Sampling Under Sampling		
	Accuracy	Recall	Confusion Matrix	Accuracy	Recall	Confusion Matrix	Accuracy	Recall	Confusion Matrix
Logistic Regression	0.97	0.83	pred:0 pred:1 true:0 89432 2949 true:1 39 184	0.99	0	pred:0 pred:1 true:0 92309 72 true:1 222 1	0.99	0.57	pred:0 pred:1 true:0 91619 762 true:1 96 127
Decision Tree	0.99	0.83	pred:0 pred:1 true:0 91658 723 true:1 37 186	0.99	0.7	pred:0 pred:1 true:0 92142 239 true:1 67 156	0.99	0.76	pred:0 pred:1 true:0 92044 337 true:1 54 169
Naïve Bayes	0.94	0.75	pred:0 pred:1 true:0 86973 5408 true:1 56 167	0.96	0.33	pred:0 pred:1 true:0 88852 3529 true:1 150 73	0.94	0.75	pred:0 pred:1 true:0 86973 5408 true:1 56 167
Random Forest	0.99	0.85	pred:0 pred:1 true:0 91652 729 true:1 34 189	0.99	0.74	pred:0 pred:1 true:0 92120 261 true:1 57 166	0.99	0.78	pred:0 pred:1 true:0 92046 335 true:1 49 174
XGBoost	0.94	0.92	pred:0 pred:1 true:0 87453 4928 true:1 18 205	0.99	0.43	pred:0 pred:1 true:0 92240 141 true:1 126 97	0.99	0.65	pred:0 pred:1 true:0 91887 494 true:1 78 145

As can be observed from the above statistics, XGBoost with SMOTE(Oversampling) outperformed all the other models with a recall of 92% i.e. the model was able to correctly predict 92% of the alerts as TP/HIGH and hence we have decided to move forward this model for predicting notable risks (TP/HIGH and TP/LOW) as well.

The recall for notable alerts (TP/HIGH +TP/LOW) is 91% which suggests that the model can correctly classify 91% of the total notable risks which is pretty good.

```
******** For Threshold = 0.5 ******
Accuarcy score =0.920597382402488
Gradient Boost Model Accuracy
              precision    recall  f1-score   support

           0       1.00      0.92      0.96     89679
           1       0.27      0.91      0.42      2925

    accuracy                           0.92     92604
   macro avg       0.64      0.92      0.69     92604
weighted avg       0.97      0.92      0.94     92604

        pred:0  pred:1
true:0   82581    7098
true:1     255    2670
```

Probability Threshold Variation for XGBoost Model for Atomic High-Risk Alerts

	TP/HIGH							
Threshold	TP(Pred =1 and actual =1	FP(pred =1, actual -0)	FN(pred =0 and actual =1)	Expected Loss(10 Million per FN) in $	Reduction in Loss	Increase in FP	Labor Cost (FP*$14.84)	Total Cost = Expected Loss +Labor Cost
0.5	203	4275	20	20000000	0	0	0	20000000
0.45	209	4954	14	14000000	6000000	679	73517.36	14073517.36
0.4	215	5806	8	8000000	6000000	852	86161.04	8086161.04
0.35	215	6903	8	8000000	0	1097	102440.52	8102440.52
0.3	217	7531	6	6000000	2000000	628	111760.04	6111760.04
0.25	217	8281	6	6000000	0	750	122890.04	6122890.04
0.2	217	8281	6	6000000	0	0	122890.04	6122890.04
0.15	219	10673	4	4000000	2000000	2392	158387.32	4158387.32
1	219	11511	4	4000000	0	838	170823.24	4170823.24
0.05	220	27944	3	3000000	1000000	16433	414688.96	3414688.96
0.04	221	28288	2	2000000	1000000	344	419793.92	2419793.92
0.03	221	31366	2	2000000	0	3078	465471.44	2465471.44
0.02	222	46943	1	1000000	1000000	15577	696634.12	1696634.12
0.01	223	47700	0	0	1000000	757	707868	707868
132079/616 days/8 analyst =27 alerts per day per analyst								
400(anlayst saary per day)/27 = $14.84 per alert								

At Probability threshold .01, The model can classify all the High Alerts with total cost for it being 700K

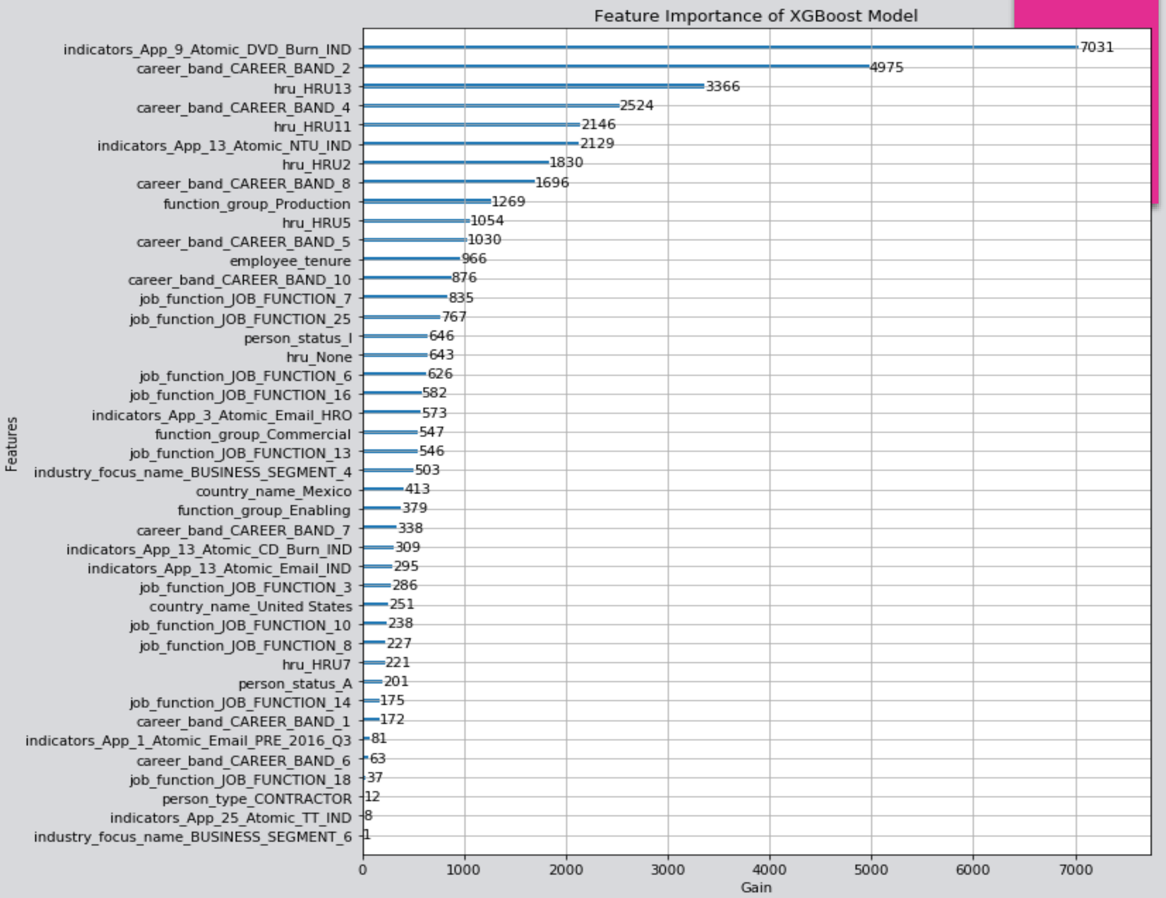
Probability Threshold Variation for XGBoost Model for Atomic Notable Alerts

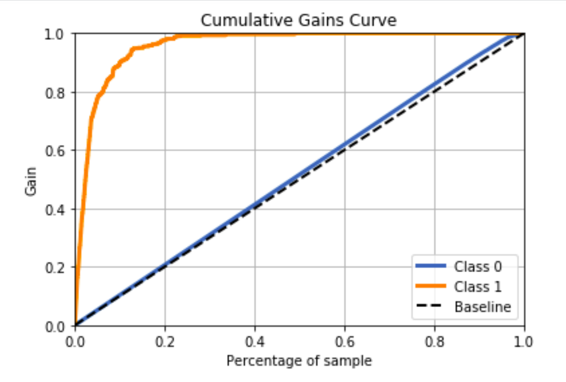
	Notable alerts(TP/HIGH+TP/LOW)							
Threshold	TP(Pred =1 and actual =1	FP(pred =1, actual -0)	FN(pred =0 and actual =1)	Expected Loss(10 Million per FN) in $	Reduction in Loss	Increase in FP	Labor Cost (FP*$14.84)	Total Cost = Expected Loss +Labor Cost
0.5	2670	7098	255	2550000000	0	0	105334.32	2550105334
0.45	2711	8588	214	2140000000	410000000	1490	127445.92	2140127446
0.4	2744	8828	181	1810000000	330000000	240	131007.52	1810131008
0.35	2759	9269	166	1660000000	150000000	441	137551.96	1660137552
0.3	2777	11000	148	1480000000	180000000	1731	163240	1480163240
0.25	2791	12037	134	1340000000	140000000	1037	178629.08	1340178629
0.2	2891	19114	34	340000000	1000000000	7077	283651.76	340283651.8
0.15	2904	22952	21	210000000	130000000	3838	340607.68	210340607.7
1	2913	27416	12	120000000	90000000	4464	406853.44	120406853.4
0.05	2924	42289	1	10000000	110000000	14873	627568.76	10627568.76
0.04	2922	44904	3	30000000	-20000000	2615	666375.36	30666375.36
0.03	2922	44917	3	30000000	0	13	666568.28	30666568.28
0.02	2925	71310	0	0	30000000	26393	1058240.4	1058240.4
0.01	2925	71313	0	0	0	3	1058284.92	1058284.92
	132079/616 days/8 analyst =27 alerts per day per analyst							
	400(anlayst saary per day)/27 = $14.84 per alert							

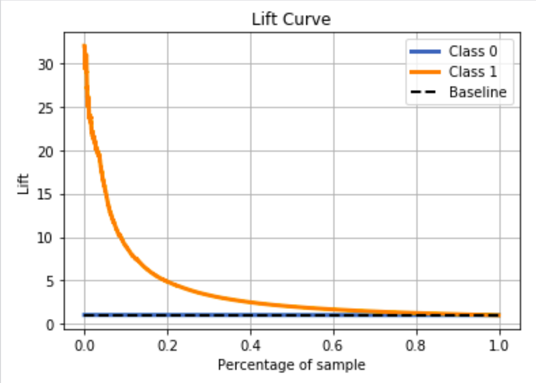
At Probability threshold .02, for notable alerts the model can classify all the High Alerts with total cost for it being 1Million

Model Insights for Atomic Alerts

- From feature importance, we can see that employee data like employee tenure, job function, function Group and career band play an important part in predicting the alerts, which is currently not integrated with the GE aviation model. GE aviation could use this feature importance to analyze which predictors needs to be focused on.
- Employee with HRU 13, HRU 11 are more prone to theft, which was also concluded from clustering, GE aviation could monitor who all are in the higher risk unit to get better insights of why employees in these units are prone to theft.
- Tenure is an important factor too in classifying the alerts. There are quite missing values for tenure, HRU and country, which if known would make the prediction more robust. GE aviation could try to gather more data regarding the hire date and background of employee.
- From the lift chart, we can state that our model is pretty good. Using the model, the GE team can predict possible thefts 30 times as compared to if they use no model.







Heat Alerts : Model Accuracy for High Risk Alert (TP/HIGH)

High Risk Alert(TP/HIGH Only)											
		Random Forest					XGBoost				
Heat_Type	Threshold	Accuracy	Recall	Confusion Matrix			Accuracy	Recall	Confusion Matrix		
					pred:0	pred:1				pred:0	pred:1
				true:0	6828	110			true:0	3907	3031
Daily	0.2	0.97	0.86	true:1	36	225	0.83	0.52	true:1	0	261
					pred:0	pred:1				pred:0	pred:1
				true:0	6390	280			true:0	2866	3804
Weekly	0.05	0.98	0.96	true:1	8	359	0.45	1	true:1	0	367
					pred:0	pred:1				pred:0	pred:1
				true:0	7698	415			true:0	355	7758
Monthly	0.05	0.95	0.99	true:1	5	366	0.08	1	true:1	0	371

Notable Risk Alert(TP/HIGH +TP/LOW)											
		Random Forest					**XGBoost**				
Heat_Type	Threshold	Accuracy	Recall	Confusion Matrix			Accuracy	Recall	Confusion Matrix		
					pred:0	pred:1				pred:0	pred:1
				true:0	5653	972			true:0	150	6475
Daily	0.2	0.85	0.92	true:1	44	530	0.1	1	true:1	0	574
					pred:0	pred:1				pred:0	pred:1
				true:0	5551	656			true:0	187	6020
Weekly	0.05	0.9	0.99	true:1	11	819	0.14	1	true:1	0	830
					pred:0	pred:1				pred:0	pred:1
				true:0	7081	531			true:0	315	7297
Monthly	0.1	0.93	0.98	true:1	18	854	0.13	0.98	true:1	0	872

Although for XGBoost FN is 0 but at the same time count of FP has also increased sharply, which suggest XGBoost is overfitting the class 1, on other hand for Random Forest (RF) count of FP is very low. On furthering reducing the threshold for RF we were able to capture more TP and thus this is the best model for detecting heat alerts.

Probability Threshold Variation for Random Forest Model for Daily Heat Alerts

	Daily Notable alerts (TP/HIGH +TP/LOW)							
Threshold	TP(Pred =1 and actual	FP(pred =1, actual -0)	FN(pred =0 and actual =1)	Expected Loss(10 Million per FN) in $	Reduction in Loss	Increase in FP	Labor Cost (FP*$14.84	Total Cost = Expected Loss +Labor Cost
0.5	2711	234	231	2310000000	0	0	3466.67	2310003467
0.4	2776	329	166	1660000000	650000000	95	4874.07	1660004874
0.3	2820	498	122	1220000000	440000000	169	7377.78	1220007378
0.2	2859	862	83	830000000	390000000	364	12770.37	830012770.4
0.1	2884	1671	58	580000000	250000000	809	24755.56	580024755.6
0.09	2886	1814	56	560000000	20000000	143	26874.07	560026874.1
0.07	2890	2253	52	520000000	40000000	439	33377.78	520033377.8
0.05	2896	2921	46	460000000	60000000	668	43274.07	460043274.1
0.04	2902	3422	40	400000000	60000000	501	50696.3	400050696.3
0.03	2912	4120	30	300000000	100000000	698	61037.04	300061037
0.02	2917	5055	25	250000000	50000000	935	74888.89	250074888.9
0.01	2923	6802	19	190000000	60000000	1747	100770.37	190100770.4
132079/616 days/8 analyst =27 alerts per day per analyst								
400(anlayst saary per day)/27 = $14.84 per alert								

At Probability threshold .01, the model can classify maximum TP, while reducing the loss by 60Million and increase in FP is not much compared to total current FP(31701)

Probability Threshold Variation for Random Forest Model for Weekly Alert

Threshold	Weekly Notable alerts (TP/HIGH +TP/LOW)							
	TP(Pred =1 and actual	FP(pred =1, actual -0)	FN(pred =0 and actual =1)	Expected Loss(10 Million per FN) in $	Reduction in Loss	Increase in FP	Labor Cost (FP*$14.84	Total Cost = Expected Loss +Labor Cost
0.5	4012	84	98	980000000	0	0	1244.44	980001244.4
0.4	4046	136	64	640000000	340000000	52	2014.81	640002014.8
0.3	4061	221	49	490000000	150000000	85	3274.07	490003274.1
0.2	4075	408	35	350000000	140000000	187	6044.44	350006044.4
0.1	4094	921	16	160000000	190000000	513	13644.44	160013644.4
0.09	4094	1027	16	160000000	0	106	15214.81	160015214.8
0.07	4098	1313	12	120000000	40000000	286	19451.85	120019451.9
0.05	4099	1803	11	110000000	10000000	490	26711.11	110026711.1
0.04	4101	2126	9	90000000	20000000	323	31496.3	90031496.3
0.03	4104	2644	6	60000000	30000000	518	39170.37	60039170.37
0.02	4104	3440	6	60000000	0	796	50962.96	60050962.96
0.01	4107	4929	3	30000000	30000000	1489	73022.22	30073022.22

132079/616 days/8 analyst =27 alerts per day per analyst

400(anlayst saary per day)/27 = $14.84 per alert

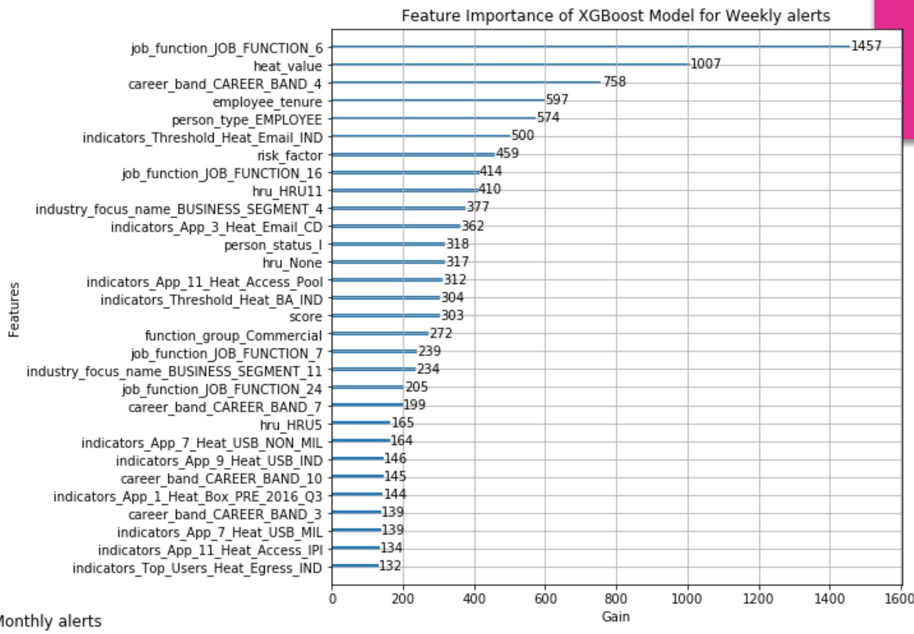
At Probability threshold .01, The model can classify almost all the alerts except 3, while reducing the loss by 30Million and increase in FP is not much compared to total current FP(4110)

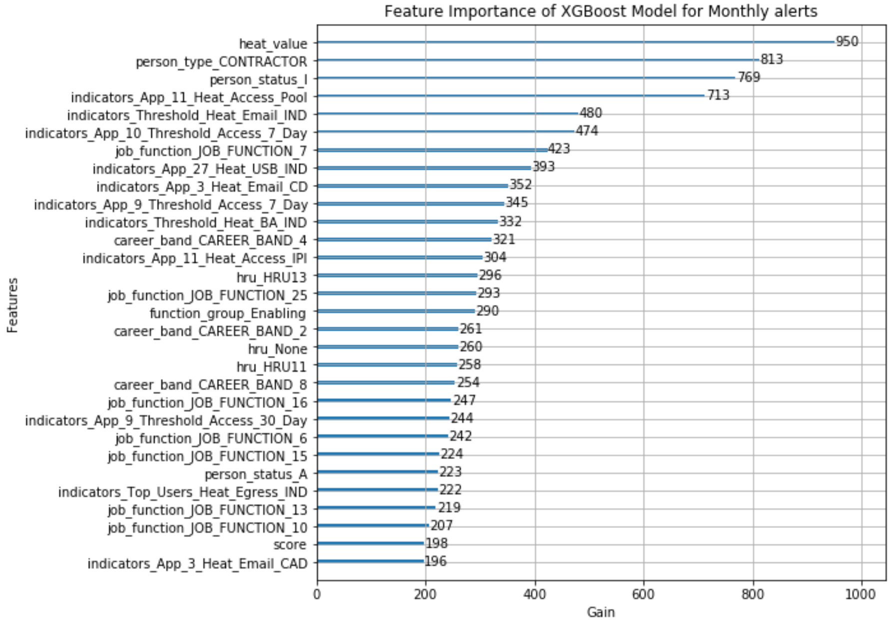
Probability Threshold Variation for Random Forest Model for Monthly Heat Alert

	Monthly Notable alerts (TP/HIGH +TP/LOW)							
Threshold	TP(Pred =1 and actual	FP(pred =1, actual -0)	FN(pred =0 and actual =1)	Expected Loss(10 Million per FN) in $	Reduction in Loss	Increase in FP	Labor Cost (FP*$14.84	Total Cost = Expected Loss +Labor Cost
0.5	3994	159	154	1540000000	0	0	2355.56	1540002356
0.4	4031	193	117	1170000000	370000000	34	2859.26	1170002859
0.3	4057	334	91	910000000	260000000	141	4948.15	910004948.2
0.2	4085	629	63	630000000	280000000	295	9318.52	630009318.5
0.1	4113	1429	35	350000000	280000000	800	21170.37	350021170.4
0.09	4117	1610	31	310000000	40000000	181	23851.85	310023851.9
0.07	4120	1996	28	280000000	30000000	386	29570.37	280029570.4
0.05	4121	2575	27	270000000	10000000	579	38148.15	270038148.2
0.04	4126	3044	22	220000000	50000000	469	45096.3	220045096.3
0.03	4128	3726	20	200000000	20000000	682	55200	200055200
0.02	4130	4803	18	180000000	20000000	1077	71155.56	180071155.6
0.01	4134	6772	14	140000000	40000000	1969	100325.93	140100325.9
132079/616 days/8 analyst =27 alerts per day per analyst								
400(anlayst saary per day)/27 = $14.84 per alert								

At Probability threshold .01, The model can correctly classify 98% of the total monthly alerts, while reducing the loss by 40Million and increase in FP is not much compared to total current FP(4118)


Feature Importance of XGBoost Model for Daily alerts
heat_value 997
indicators_Threshold_Heat_Email_IND 910
indicators_App_1_Heat_Box_PRE_2016_Q3 837
indicators_App_3_Heat_Email_CD 509
job_function_JOB_FUNCTION_24 383
indicators_App_20_Threshold_App_Usage_30_Day 368
job_function_JOB_FUNCTION_16 331
career_band_CAREER_BAND_6 292
job_function_JOB_FUNCTION_11 283
indicators_Threshold_Heat_Print_IND 283
indicators_Threshold_Heat_BA_IND 268
job_function_JOB_FUNCTION_10 256
indicators_App_7_Heat_USB_NON_MIL 255
job_function_JOB_FUNCTION_13 244
job_function_JOB_FUNCTION_7 243
career_band_CAREER_BAND_7 240
indicators_Top_Users_Heat_Egress_IND 239
function_group_Commercial 234
indicators_App_9_Heat_Email_IND 234
indicators_App_7_Heat_USB_MIL 232
person_type_CONTRACTOR 231
indicators_App_13_Atomic_Email_IND 227
indicators_App_15_Heat_Badge_IND 225
career_band_CAREER_BAND_3 224
indicators_App_11_Heat_Access_Pool 222
indicators_App_4_Heat_CAD_IND 217
indicators_App_13_Heat_USB_IND 214
function_group_Enabling 206
indicators_App_11_Threshold_Access_30_Day 201
person_status_A 198
Features
0 200 400 600 800 1000
Gain


Feature Importance of XGBoost Model for Weekly alerts
job_function_JOB_FUNCTION_6 1457
heat_value 1007
career_band_CAREER_BAND_4 758
employee_tenure 597
person_type_EMPLOYEE 574
indicators_Threshold_Heat_Email_IND 500
risk_factor 459
job_function_JOB_FUNCTION_16 414
hru_HRU11 410
industry_focus_name_BUSINESS_SEGMENT_4 377
indicators_App_3_Heat_Email_CD 362
person_status_I 318
hru_None 317
indicators_App_11_Heat_Access_Pool 312
indicators_Threshold_Heat_BA_IND 304
score 303
function_group_Commercial 272
job_function_JOB_FUNCTION_7 239
industry_focus_name_BUSINESS_SEGMENT_11 234
job_function_JOB_FUNCTION_24 205
career_band_CAREER_BAND_7 199
hru_HRU5 165
indicators_App_7_Heat_USB_NON_MIL 164
indicators_App_9_Heat_USB_IND 146
career_band_CAREER_BAND_10 145
indicators_App_1_Heat_Box_PRE_2016_Q3 144
career_band_CAREER_BAND_3 139
indicators_App_7_Heat_USB_MIL 139
indicators_App_11_Heat_Access_IPI 134
indicators_Top_Users_Heat_Egress_IND 132
Features
0 200 400 600 800 1000 1200 1400 1600
Gain


Feature Importance of XGBoost Model for Monthly alerts
heat_value 950
person_type_CONTRACTOR 813
person_status_I 769
indicators_App_11_Heat_Access_Pool 713
indicators_Threshold_Heat_Email_IND 480
indicators_App_10_Threshold_Access_7_Day 474
job_function_JOB_FUNCTION_7 423
indicators_App_27_Heat_USB_IND 393
indicators_App_3_Heat_Email_CD 352
indicators_App_9_Threshold_Access_7_Day 345
indicators_Threshold_Heat_BA_IND 332
career_band_CAREER_BAND_4 321
indicators_App_11_Heat_Access_IPI 304
hru_HRU13 296
job_function_JOB_FUNCTION_25 293
function_group_Enabling 290
career_band_CAREER_BAND_2 261
hru_None 260
hru_HRU11 258
career_band_CAREER_BAND_8 254
job_function_JOB_FUNCTION_16 247
indicators_App_9_Threshold_Access_30_Day 244
job_function_JOB_FUNCTION_6 242
job_function_JOB_FUNCTION_15 224
person_status_A 223
indicators_Top_Users_Heat_Egress_IND 222
job_function_JOB_FUNCTION_13 219
job_function_JOB_FUNCTION_10 207
score 198
indicators_App_3_Heat_Email_CAD 196
Features
0 200 400 600 800 1000
Gain

Model Insights for Heat Alerts

Daily Heat Alert Insights	Weekely Heat Alert Insights	Monthly Heat alert Insiights
1.indicators contributing to high and low-risk clusters are the ones with lower higher heat value in the range from 1-5. Since the minimum score to fire daily and weekly alert is 170 and 450 respectively, this suggests that these alerts are fired multiple times. If we could keep the track of the frequency of alerts being fired, we could accordingly add weights to the indicators which our model can then emphasize. 2.Employee with job Function 16 and 24 are major contributors. 3. Need to check on function group Commercial and Enabling	1.Employee Tenure is having higher importance for weekly alert model as compared to the score(which currently is the base for firing alert in the current system) which clearly indicates how important it is to have the correct value of this feature and therefore GE aviation should try to gather more data regarding the hire date and background of employee. 2. Employee with job Function 6 , 16 are major contributors. 3.Need to monitor the activities of employee with career band 4 and 7	1. Most of the monthly alerts came for employee who are contractor. 2. Person Status as Inactive is also a major contributor to the alerts. 3. Need to monitor employee working in HRU 13, 11, None.

Conclusion

- Our model would outperform the rule-based GE aviation's current model in terms of number of analysts and time spent by them in manually classifying the alerts and Cost GE pays for each theft.
- By varying the thresholds of the best model, GE aviation can judiciously utilize its resources depending on the availability of analysts, priority of their work.
- Lastly, the model can be easily retrained by adding more data and features making the model more dynamic in nature which is quite a complex task to accommodate with current rule-based system.

Recommendation

- Integrate Employee details.
- Missing Value for Ge_hire_Date, HRU, Owner_Type needs to be captured.
- Store the Frequency of the indicators being fired.
- Most of the High and low alerts were analyzed by Analyst 6, 7,8 and Senior Analyst 1, could gather more insights into the potential steps needed to improve the accuracy of the current system by collaborating with these analysts.

Thank You for your time