**Memo**

**To: GE Aviation**

**From: Jie Wang**

**Date: 03-29-2020**

**Re: Ge Aviation – IP Theft and Fraud Detection System Proposal**

**Project Purpose**

GE Aviation provides worldwide business, commercial, military and aviation jet, turboprop engines etc. services to their global customer network (GE Official Website). Therefore, it is important for GE to prevent their internal information from hackers, external personnel even their own staff. In addition, GE has executed an alert system to monitor employees’ computers to confirm they would not do any actions that could threat company’s intellectual property. In addition, I will perform unsupervised learning on atomic alert and threshold alert which I will use k-means to get the clustering and former information. As for the current alert system, the most of classification is TP/DE which give meaningless effort so we should find the optimized threshold of probabilities to trigger the alerts and give some recommendations to the sponsor how to save the cost of this IP protection program. Also, increasing the percent of TP/HIGH and TP/LOW is a way to increase the accuracy of effective alerts. Moreover, I will give the predictive models of different types of alert in different risk groups and analyze the models to obtain the results.

**Project Recommendations**

1. The atomic alert in both high risk and notable risk as well as notable risk in daily heat need improvement immediately because they have low accuracy of alerts in the system.
2. The probability of threshold in high risk and notable risk of atomic alert as well as notable risk in daily heat alert should be decreased to 0.0001, 0.001 and 0.05 respectively. Moreover, high risk and notable risk in weekly heat alert and monthly heat alert also high risk in daily heat alert should be increased to 0.99
3. Analyst\_2, Analyst\_3 and Analyst\_5 are not important in the alert program and they put few efforts to company. So, the company should consider whether to fire these three persons to save the employee cost.
4. Senior\_Analyst\_1 plays an important role in atomic alert and notable risk in monthly heat alert, so the company should consider whether increase the salary for the person or divide the workload more evenly between Senior\_Analyst\_1 and Senior\_Analyst\_2.
5. The indicator App\_3\_Heat\_Email\_CAD is important because it has high proportion in high risk and notable risk in these alerts more times so GE should have further investigation on this indicator and reasons why employees tend to make related mistake and trigger the indicator.
6. CAREER\_BAND\_10, CAREER\_BAND\_4, JOB\_FUNCTION\_16 and hru5 are important, GE should take more attention on these factors.

**Project Methods**

In this project, I used KNIME and Tableau to conduct the EDA analysis and unsupervised learning for four types of alerts also I used K-means clustering to do the preliminary observation for each alert. Furthermore, I did decision trees of TP/HIGH and TP/HIGH & TP/LOW for each model to do prediction of the four alerts.

**Project Analysis and Insights**

1. **Exploratory Data Analysis**

**A screenshot of a cell phone

Description automatically generated**

Figure1. Workload of Analysts in Classification

From figure 1, we can know that Analyst\_2, Analyst\_3 and Analyst\_5 have the least workload towards different types of classification. However, we still need to have deeper analysis of these analysts to make the final decision. For example, which of them put more effort in classification of TP/HIGH or TP/LOW.

A screenshot of a cell phone

Description automatically generated

Figure 2. Classification in Different Types of Alert.

From figure 2, TP/HIGH classification occupies the lowest part in all of alerts especially in the atomic alert. In the following, I will have deeper analysis and insights in the TP/HGIH and TP/LOW.

1. **K-means of Different Types of Alerts**

I use k-means clustering to do the unsupervised learning and did 3 to 12 clusters for each model which include atomic alert, daily heat alert, weekly heat alert and monthly heat alert. From the output, I will decide how many clusters are the optimal one and do further observation.

1. **Atomic Alert**

**A picture containing text, newspaper

Description automatically generated**

Figure 3. Partial Cluster Output of Atomic Alert.

From the cluster output of atomic alert, I think 11 clusters is the most ideal number. Since in the 11 clusters, there are the same numbers of effective statistics in TP/LOW with the statistics in 12 clusters so 11 clusters are the most optimize one. From 11 clusters, I conclude that Senior\_Analyst\_1 put the most effort; Analyst\_2 and Analyst\_3 don’t have any workload; TP/DE and FP occupy the most of classification but there still has some of statistics about TP/HIGH especially TP/LOW that we can analyze; HRU 11 and App\_9\_Atomic\_DVD\_Burn\_IND are important.

1. **Daily Heat Alert**

A screenshot of a computer

Description automatically generated

Figure 4. Partial Cluster Output of Daily Heat Alert.

From figure 4, I think 5 clusters is enough to observe the results because since the clusters are 6, the output of TP/LOW is only 1. Then, I observe that Senior\_Analyst\_1 still put the most effort and Analyst\_3 has the least effort; the most classification is still TP/DE but there are a lot TP/LOW can be do deeper analysis; HRU13, App\_1\_Heat\_Box\_PRE\_2016\_Q3, App\_13\_Heat\_Print\_IND are important.

1. **Weekly Heat Alert**

A picture containing wooden

Description automatically generated

Figure 5. Partial Cluster Output of Weekly Heat Alert.

From figure 5, I think 7 clusters are enough to observe because from 8 clusters, the effective statistics of TP/HIHG become less. I conclude that Senior\_Analyst\_1 still put the most effort and Senior\_Analyst\_2 is also important; TP/DE is the most classification and TP/HIGH is second;HRU11,App\_1\_Heat\_Box\_PRE\_2016\_Q3, App\_3\_Heat\_Email\_CAD/App\_3\_Heat\_Email\_CD/Threshold\_Heat\_NTD\_EXE\_IND are important.

1. **Monthly Heat Alert**

A close up of a green screen

Description automatically generated

Figure 6. Partial Cluster Output of Monthly Heat Alert.

From figure 6, I think 8 clusters are the most ideal for monthly heat alert and Seinior\_Analyst\_1, Senior\_Analyst\_2, TP/HIGH, HRU 11 and App\_3\_Heat\_Email\_CAD/App\_3\_Heat\_Email\_CD/Threshold\_Heat\_NTD\_EXE\_IND are important factors.

1. **Decision Tree Models in Different Types of Alerts**

In this part, I will show the decision tree model, confusion matrix and lift chart of atomic alert, daily heat alert, weekly heat alert and monthly alert. Also, the decision tree models will be conducted in two risk groups that are high risk group (TP/HIGH) and notable risk group (TP/HIGH & TP/LOW) in these four alerts. From the results of decision tree, we can know which factors are important for different cases. In addition, in all of decision tree models, the first branch is the most important factor for the model and has the highest correlation.

1. **Atomic Alert – High Risk**

**A close up of text on a white background

Description automatically generated**

Figure 7. Decision Tree

**A screenshot of a cell phone

Description automatically generated**

Figure 8. Important Branches.

Figure 7 and figure 8 are the decision tree and important branches of the tree in the high risky group of atomic alert. We can see that career\_band\_2 is highly related to the model with percent of 0.5. It means that the portion of high risk in the atomic alert is pretty low. In addition, the branch of App\_13\_Atomic\_NTU\_IND&career\_band\_2 has 78.4%, senior\_analyst\_1 has 87.3% and the branch of job\_function\_7&App\_13\_Atomic\_NTU\_IND&career\_band\_2 occupies 93.9% of detection probability respectively. Therefore, the factors that App\_13\_Atomic\_NTU\_IND, job\_function\_7 and senior\_analyst\_1 are important in this case.

1. **Atomic Alert – Notable Risk**

A picture containing screenshot

Description automatically generated

Figure 9. Decision Tree

A screenshot of a cell phone

Description automatically generated

Figure 10. Important Branches.

From figure 9 and figure 10, variable that job\_function\_7 is highly correlated to the model and hru5, career\_band\_4, senior\_analyst\_1 and App\_12\_Atomic\_App\_Usage\_IND are important.

1. **Daily Heat Alert – High Risk**

**A close up of a piece of paper

Description automatically generated**

Figure 11. Decision Tree

A screenshot of a cell phone

Description automatically generated

Figure 8. Important Branches

From figure 11 and figure 12, factor that App\_3\_Heat\_Email\_CAD is highly related to the model and career\_band\_10, App\_9\_Heat\_USB\_IND and App\_27\_Heat\_USB\_IND are important

1. **Daily Heat Alert – Notable Risk**

A picture containing screenshot

Description automatically generated

Figure 13. Decision Tree.

A screenshot of a cell phone

Description automatically generated

Figure 14. Important Branches

From figure 13 and figure 14, factor that App\_27\_Heat\_USB\_IND is highly related to the model and App\_9\_Heat\_Email\_IND, App\_7\_Heat\_USB\_MIL and App\_3\_Heat\_Email\_CD etc. are important.

1. **Weekly Heat Alert – High Risk**

**A picture containing screenshot

Description automatically generated**

Figure 15. Decision Tree

A screenshot of a cell phone

Description automatically generated

Figure 16. Important Branches

In this model, App\_3\_Heat\_Email\_CAD is highly related and factors that career\_band\_10 and App\_7\_Heat\_USB\_MIL are important.

1. **Weekly Heat Alert – Notable Risk**

A screenshot of text

Description automatically generated

Figure 17. Decision Tree

A screenshot of a cell phone

Description automatically generated

Figure 18. Important Branches

In this model, indicator JOB\_FUNCTION\_16 is highly related and analyst\_4, App\_27\_Heat\_USB\_IND and App\_3\_Heat\_Email\_CAD/TOP-User etc. are important.

1. **Monthly Heat Alert – High Risk**

A close up of text on a white background

Description automatically generated A screenshot of a cell phone

Description automatically generated

Figure 19. Decision Tree Figure 20. Important Branches

In this model, App\_3\_Heat\_Email\_CAD is highly related and career\_band\_10, App\_13\_Heat\_USB\_IND and App\_3\_Heat\_Email\_CAD etc. are important.

1. **Monthly Heat Alert – Notable Risk**

A screenshot of text

Description automatically generated A screenshot of a cell phone

Description automatically generated

Figure 19. Decision Tree Figure 20. Important Branches

1. **Confusion Table**

A screenshot of a cell phone

Description automatically generated

Figure 21. Confusion Table of Threshold Alert.

From figure 21, we can know that notable risk in weekly heat has the most accuracy which is 591. In the high risky group, daily heat alert has the biggest number of accurate alerts which is 574. We need to increase the threshold probability of the model to test what probability is ideal to increase accurate alerts.

1. **Lift Chart**

A close up of a map

Description automatically generated

Figure 22. Lift Chart of Threshold Alert.

From figure 22, notable risk in weekly heat alert is the best predictive model because the difference between lift line (red line) and baseline (green line) is the biggest.

1. **Threshold Probability**

A screenshot of a cell phone

Description automatically generated

Figure 23. Table of Threshold Probability.

From figure 23, we can know that in the high risk of atomic alert, the probability should be 0.0001 then the accurate alerts would be 25165 and it will increase 25061 of accurate alerts compared with probability of 0.5 and the cost will be 371,274.2 dollars. Moreover, for the notable risk of atomic alert, the optimized probability is 0.001 and there will increase 19329 accurate alert and the cost will be 286,069.2. In addition, other types of alerts in high risky group and notable group don’t increase so much after optimization except notable risk in daily heat whose probability should be 0.05 and increased alerts are 12635. Although action of increasing accurate alerts cost amount of money, it will save huge amount of wasting money and increase the security of the system.

**Conclusion**

From above, GE IP Theft and Fraud Detection System has problems that should be solved to increase the accuracy of alerts and the important factors in each model should have deeper investigation inside the company then find the best solution. For example, indicators that have bigger probability to trigger the alerts, analysts who have lower effort to the company also group of job junction which includes the most employees in high risk and notable risk etc.

**Reference:**

GE Aviation. Retrieved from: <https://www.geaviation.com/>