Integrated Experiential Learning  
  
  
ALY6080

**XN Individual Project Proposal**

**GE Aviation Analysis**

Spring Term (2020)

SUBMITTED BY: KEYUR SHAH

NUID: 001089242

SUBMITTED TO: Prof. Dr. Nathaniel Lin

COMPLETION DATE: 06/24/2020

**INTRODUCTION**

Intellectual Property (IP) rights secure the innovations, processes and ideas of your company that are vital to keep the brand growth and competitive advantage alive. It also helps to enhance the business and lead the market. Preventing IP theft is a difficult task as it requires analysing user characteristics, monitoring devices etc. The misuse of IP directly intimidates the market share, competitiveness of GE aviation and keeps at risk the critical data of GE aviation. GE's new method is the detection of fraud based on statute, which does not require large data sets of training and it is built on a Bayesian network. But for efficiency and dynamic based, the system needs to be migrated to predictive based like various machine learning algorithms for achieving higher accuracy and also cost efficiency. The rule-based system brings a huge workload to many senior analysts as many false alerts are generated and it can cost millions of dollars for each and every single IP theft alert.

**APPROACH**

Rule based approach isa System that uses human-made rules for data storage, sorting and manipulation of data. It also requires a predefined threshold for generating alerts. On other hand to that, the predictive based approach works on data and predicts the future outcome and also they are known as Machine learning Algorithms. This would help to reduce human efforts for identifying the alerts and also help to dig major insights of hidden patterns. The algorithm which GE should implement is Decision Tree and Random Forest. For analysing, I had performed exploratory data analysis for digging major insights from GE aviation by using tableau and performed unsupervised learning like k-means clustering for positioning and identifying risk groups using python pandas data frame. This helped me to identify the Number of false positives, thresholds of firing the alerts etc. Finally, I applied predictive models like random forest and decision tree for classifying notable risk and high-risk alerts and also set a threshold for saving cost to the company.

**ANALYSIS**

GE Aviation has provided 5 different datasets for our analysis which comprises all the details regarding the employees, heat score, indicators and etc. We evaluated all the datasets and found the employee and BN indicator to be important as the majority of columns are related. These 2 datasets i.e. “obfuscated\_data” which has demographic information of employees and “BN\_Data” that contains the heat score of all alerts triggered. The dataset merged using pandas’ libraries by matching the employee id of all employees. We then identified the null values and replaced them with relevant values for inconsistency of data.

Next, we perform Exploratory Data Analysis in tableau. We found that the majority of classification were TP/DE count as 45274, which are True Positive and that does not require any escalation counts as for the alert and this led to investigating many false alerts. Also we identify maximum alerts are Monthly heats and it counts as 28651.

Furthermore, I had performed an unsupervised learning algorithm for identifying the critical risk group. I had cleaned missing values and also performed one hot encoding method for conversion to integer data. Then using Knime, I had merged datasets and then performed K-means Clustering on different alert systems by filtering as Daily\_Heat, Weekly\_Heat, Monthly\_Heat. We found that optimal clusters for Atomic, Daily, Monthly and Weekly are 10,11,10,12 respectively. By profiling the different clusters for various alerts, we analyse that “JOB\_FUNCTION\_6”, “JOB\_FUNCTION\_7”, “JOB\_FUNCTION\_8”, “JOB\_FUNCTION\_9”, “CAREER\_BAND\_2”, “CAREER\_BAND\_6”, “CAREER\_BAND\_8”, “Senior\_Analyst\_1”, “Analyst\_5”, “Analyst\_6”, “Analyst\_7”, “Analyst\_8” , “HRU11”, “HRU5”, “indicators\_App\_13\_Heat\_USB\_IND” ,”BUSINESS\_SEGMENT\_4 etc.

Finally, I performed supervised learning for different alert types. I had used Machine Learning Algorithms like Decision tree and Random Forest for different alert types. I had classified the alerts as High Risk group (“TP/HIGH ->0) and Notable Risk group as (“TP “ ,”FP/DE” ,”TP/LOW”) for better classifying the data. As per results, we found that the accuracy was not upto efficient as data imbalance and oversampling was implemented for data distribution. The FNs were found to less with Random Forest in case of Daily\_Heat which was found to 12% and that of Monthly\_Heat was found to be 11% where as decision tree tuned out to perform better in case of Atomic and Weekly\_Heat with FNs as 12% and 9% respectively.

**RECOMMENDATION AND DELIVERABLES**

By analyzing, we realized that the system efficiency can be easily improved if the system shifts itself over predictive models like the random forests and decision tree that gave us results with higher accuracy as compared to that achieved by the rule based system on which GE currently works. Also, other models such as GVM, Gradient Boost etc. can also be implemented to check which model generates results with higher accuracy. Here, we applied multiple values of threshold to observe which one is responsible for obtaining more accurate results. Threshold value as 0.8 gave me the best output and so we can try other values also to observe the varying accuracy of the system.

**REFERENCES: -**

* Brownlee, J. (2019, August 12). Supervised and Unsupervised Machine Learning Algorithms. Retrieved June 24, 2020, from <https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>
* Sunil RayI am a Business Analytics and Intelligence professional with deep experience in the Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years. (2020, April 15). Commonly Used Machine Learning Algorithms: Data Science. Retrieved June 24, 2020, from <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>
* ), M. (2018, October 23). Chapter 4: Decision Trees Algorithms. Retrieved June 24, 2020, from <https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1>