GE Project Proposal

Northeastern University

Linmu Liu

Assignment: GE Project Proposal

Course Number: ALY 6980

Capstone

Second Summer

Instructor: Eugenia Bastos

Sponsor: Eric Gero, Staff Data Scientist, GE Digital – Aviation

Company: [https://www.geaviation.com/digital](https://nam05.safelinks.protection.outlook.com/?url=https://www.geaviation.com/digital&data=02%7C01%7Ce.bastos@northeastern.edu%7C05473649831546c5663808d6bc655284%7Ca8eec281aaa34daeac9b9a398b9215e7%7C0%7C0%7C636903540689291945&sdata=3VW4Y/yx0s5UptgwUwmesDvo2x/VhjYEswUsD3MSPzc%3D&reserved=0)

Date: Aug 24th, 2019

**Introduction**

The theme of this project is mainly associated with the utilization of digital guardian in terms of its effectiveness, business benefits. With the rapid technological advances and changes, many tech giants are competing against each other not only manifesting in profitability, but also data loss prevention, including preventing the leak of customer information, intellectual property, data visibility. Big data has become a common word nowadays which can be explored to create large gains for many organizations. GE Aviation applied ‘digital guardian’ to monitor employees computer activities to ensure there are no severe threats that are harmful to company’s regularly operating and performance to defend intellectual property. Digital guardian is a commercial tool and used to make sure data is not lost, misused, or accessed by unauthorized users. Data loss prevention(DLP) software has the ability to classify business critical data based on the organizations’ policies and rules. Once violations are identified, the system will react to create alert signals (Zhang, 2019). In the GE Aviation Company, the client will be installed on the GE asset and collect usage information. Each employee will have a typical heat scored recorded by the system, typically calculated by daily. There is a variety of indicators, or a combination of indicators (indicators pairs). Once the heat score pass certain threshold, the system will keep the record of date of alert created. Next, all of alerts will be investigated by analysts. Moreover, If any employees activities produce potential risk it will present atomic classified risk and multiple atomic indicators lead to heat alert. Even if the system didn’t perform with a 100% accuracy, it’s still somewhat effective in detecting heat alerts. Thanks to our honorable sponsor Eric Gero and professor Bastos, I have this valuable opportunity to dig into the dataset and tell a data story.

The main dataset consisting of 186558 entries in sum. The other tables are related to alerts\_by\_indicators, alerts\_by\_pairs, demographics, heat\_scores. The time span of the data ranges from 2018 to 2019. There are many numerical and categorical variables in the dataset. Malicious is the column that I am interested in the most. Meanwhile, This study intends to figure out whether the GE is using digital guardian in the most effective way.

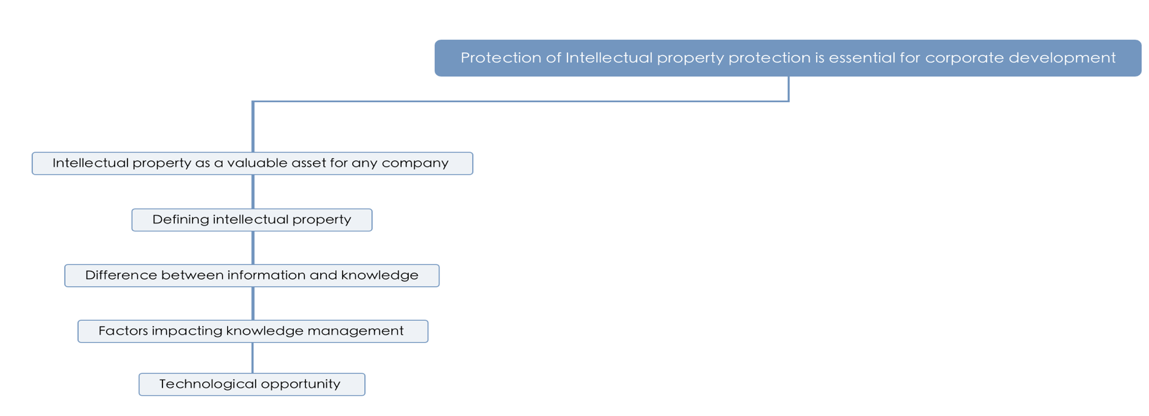
**Purpose**

The purpose is to help GE find drawbacks on the current system and present reasonable modeling to improve the digital guardian by increasing the accuracy rate. The mix of methods will be utilized to analyze the data. First of all, Simply looking at the dataset and search for trends, obvious pattern, or relationship between variables. Then moving on to quantitative analysis to produce charts, predictive modeling to further investigate the contribution of features and explore which model offers the best valid results. Business issues are waited to be solved, such as

* Is GE using their indicators in the most successful way?
* Are the analysts effective in classifying the alerts?
* Which variable needs to be considered while building predictive modeling?
* Is there a relationship between features in main table and variables in demographic table?
* Machine learning might be a good way to deal with lowering the high error rate of the classification of heat alerts, but which machine learning methods play a vital role in predicting malicious?
* From a broad perspective of GE business, how to reduce the costs and boost the profitability through the improved digital guardian.
* There are many extreme values which can be considered as outliers in the score column. These alerts which belong to FP will produce large biases. Should we include those values in our modeling?
* Which features can explain the variance of malicious the most? the least?

**Literature Review**

1. “Strategies for Managing Knowledge Assets: The Role of Firm Structure and Industrial Context”



Organizations nowadays are competing against each other in terms of technological perspective. Teece(2000) points out that the ability to create and innovate things are essential for business firms in order to survive and thrive in the market. Accordingly, the ownership and protection of Intellectual property become increasingly important. The goal of this paper is to inform readers of strategies for managing knowledge assets to sustain and strengthen the competitive advantage. Traditionally, physical assets are considered as key assets. In the modern society, the possess of knowledge and IP, like copyrights and patents are regarded as vital assets. Therefore, how to effectively take advantage of the knowledge assets becomes a common issue. Individual knowledge transaction happens every day as one gets hired or fired. However, organization knowledge consists of knowledge of a large group of people which is a different scenario. The paper illustrates the difficulties of the process of knowledge assets transfer and help correct a misleading conception that people believe knowledge transfer is cost free. In reality, that’s not true. Knowledge transfer is not simply transferring techniques, also including competitors, customers and suppliers. It is necessary to clarify the meaning of information and knowledge which are two different things. The posts and comments received by Instagram and Facebook are information. These raw data could easily be the cause of information overload. People who possess the ability to make use of big data to figure out business questions by filtering out data, building visualizations or modeling to interpret and integrate the data are knowledgeable. knowledge are codified and stored electronically in the company’s own database which increases the risk exposures that information/knowledge leaks out within seconds. It’s a tough task to protect the safety of the knowledge. Therefore, many organizations are seeking for an efficient tool to protect their IP, like GE using Digital Guardian. A decent company structure will lower the risk of information leak-out. The conflict between shareholders and managers is the main issue each large company confronts. The use of equity of pay is an effective way to gain confidence on making sure decisions employees made are working in the best interest of the company. Teece(2000) emphasizes the importance of understanding the relative role of knowledge assets, and dynamic capabilities to maintain the competitive advantage. Education is the least impacted sectors compared with agriculture, retailing industry as information technology develops. Moreover, a successful company must respond to adjustments quickly and be ready to implement strategic strategies comprehensively. When new technology is invented, organizations need to take actions and make predictions about future trends. For example, firms should respond in different ways with respect to 3D graphics accelerator chips and demand for caskets and coffins.

This paper demonstrates that superior performance comes from the ability to innovate, to protect intellectual property. Knowledge management help companies list the priorities. Context is an essential part of considerations while making reform decisions. The target the paper aims to inform are related to technology-based companies, especially in Silicon Valley.

2.“Detecting Financial Fraud Using Data Mining Techiniques: A Decade Review from 2004 to 2015”

In the GE dataset, the purpose of implementing digital guardian is to monitor employees computer activities by measuring the heat score, once it cross certain threshold, the system will generate alert and pass it to analysts to proceed further analysis. The analysts categorized the threats into four types and record the results by confirmed or cleared. When I first thought about this case, logistics regression appears in my mind which is a widely used statistical model while dealing with a binary dependent variable, judging the alert: malicious(YES or NO).

The goal of this article is to illustrate the distribution of commonly used techniques of detecting financial fraud using data mining techniques, in which logistics regression is the representative one. Large organizations such as Facebook, Google invested billions of dollars into cyber security to prevent personal information leak and financial fraud. Albashrawi(2016) mentions that Bank of America even agreed to pay $16.5 billion to tackle with the financial fraud issue. Since Mr.Walker mislead more than thousands of investors and committed 56 million dollars fraud, the growing problem attracts the public attention and indicates that financial fraud cannot be ignored.

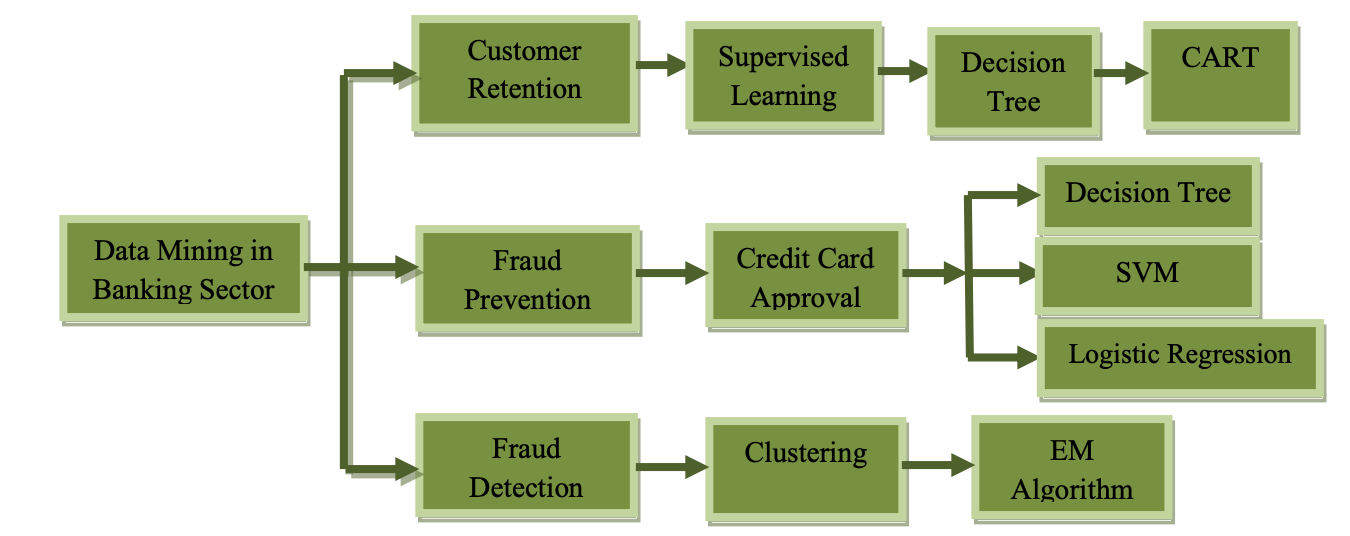
The article presents many types of fraud and various data mining tools used, such as Automobile insurance fraud with Probit model, Logit model; Credit card fraud with SVM; Fraudulent financial statements predicted by decision trees, Naïve Bayes. Except for applying machine learning methods listed above. Text mining techniques are employed and added into the forecasted model. Key words like “detecting financial fraud, financial fraud and data mining” are set up as signal words while searing for related articles(Albashrawi,2016). The distribution of articles by journals and conferences from 2004 to 2015 concentrates on Expert Systems with Applications. Moreover, the United States published the most number of articles about detecting financial fraud. Among the data mining techniques in these articles, it appears that logistic regression model is the leading data mining tool accounting for 13% of all techniques. On top of that, Neural network is the second common application tool (Albashrawi,2016).

This article has some limitations. Detecting financial fraud requires high priority to get access to the dataset. This article lacks up-to-date and cannot be used as standards to decide which factors should be taken into consideration while detecting financial fraud and aiding in choosing the appropriate technique. The combination of frequency of financial statement fraud and Bank fraud takes up 63%. Furthermore, the article makes a conclusion that there is a huge increase research conducted in next four years starting from 2008, because the 2008 financial crisis drew many economists and financial industry attention into preventing the next severe event from happening.

3. “Data Mining Techniques and its Applications in Banking Sector”

Until so far, I’ve found articles associated with the significance of protecting intellectual rights which are essential for the development of GE company, especially for technology-based companies. Moreover, the role of firm structure and strategies being applied to manage intangible assets become increasingly important. Besides, the article aims to introduce how to use data mining techniques to facilitate business environment by keeping a steady customer retention rate. This manifest in detecting fraud transactions and preventing the personal information from leaking and being manipulated by evil people. Similarly, GE company are making use of commercial data protection software to monitor their employees. Since the system is not one hundred percent correct, we can take advantage of the data mining methods to calculate the detection accuracy of the system and get the system tuned by lowering the false positive records.

CRM, short for customer relationship management, is regarded as a strongly useful strategy in bank industry. Low cost and high profits are two main goals. Chitra and Subashini (2013) points out that the Data mining workflow composes of several steps. First, understanding the problem, like bank’s priority concerns; Second, collect customer data and have a comprehensive understanding of datasets; Third, data processing, data filtering which makes sure your data is clean enough to be mined; Next, system modeling consists of study ways to keep a high customer retention and prevent fraud. On top of that, system evaluation covers testifying your modeling by using test data; Last, analyzing your results and performance and express your insights. Both data mining algorithms for supervised functions and unsupervised functions are introduced. In this article, it illustrates that decision tree, as a supervised learning method, is implemented using CART algorithm for customer retention. Credit approval process apply data mining tools, such as Decision tree, Support Vector Machine(SVM) and Logistic Regression. By the way, the category of Credit card has been the number one serious problem reported by banks in India(Chitra& Subashini ,2013). Likewise, In the GE dataset, the logistics regression can be used to determine to what degree that the alert is malicious (outcome variable).



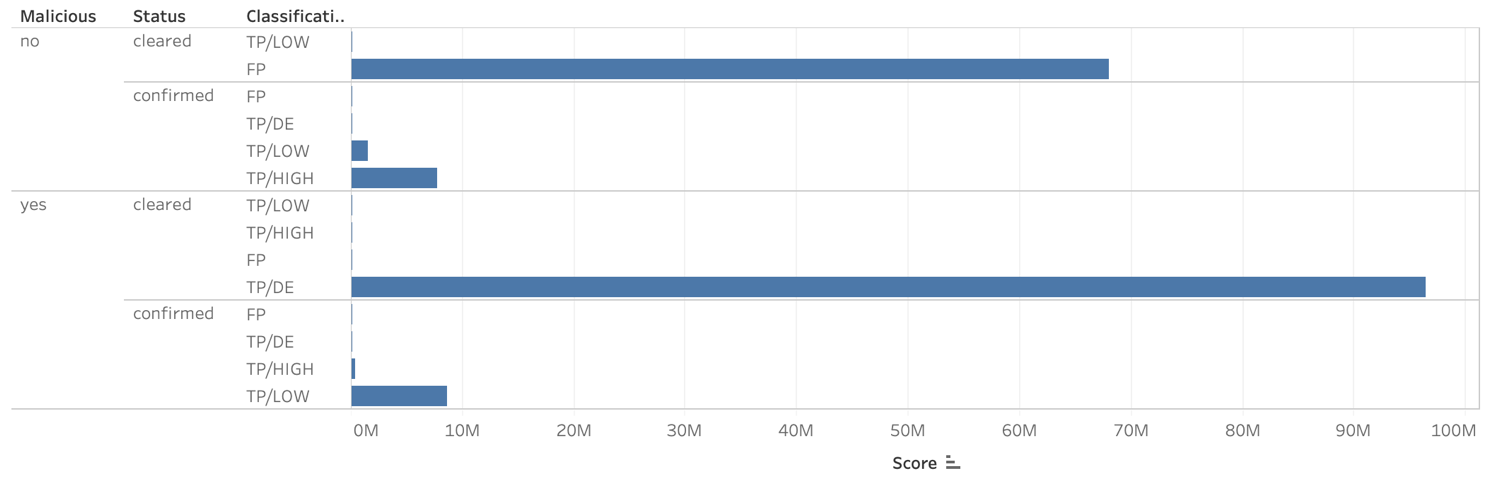
The tree analysis above shows three common problems with type of methods applied and machine learnings techniques match-up in the most effective way.

**Methods**

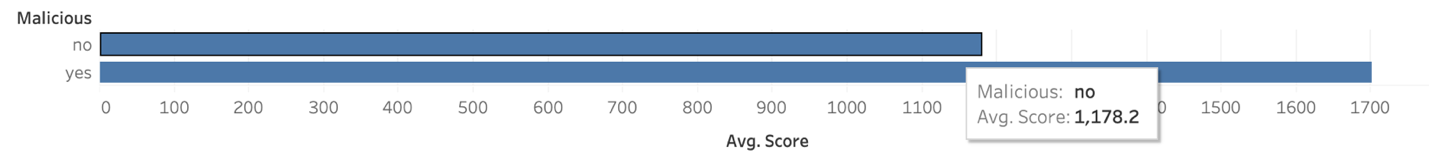
The tableau and python jupyter notebook are main tools used for this project. Shorten & Smith(2017) present that ‘Mixed methods’ is a research approach whereby researchers collect and analyse both quantitative and qualitative data within the same study. The mixed methods allows researchers to study various perspectives and discover relationship. In this study, starting with qualitative analysis looking at the observations in the dataset and find relationships through interpretation of evidence. Qualitative analysis instigate researchers to come up with excellent research questions. A mix of methods is a good fit since GE dataset cannot be dealt by quantitative or qualitative alone. The data dictionary provides us with brief explanation. Therefore, it’s necessary to further study the meaning of each column and find interrelation between variables, even columns between tables. The qualitative analysis is followed by quantitative analysis by showing charts, like histogram, pie charts. Lastly, building prediction modeling to compare which model shows the best performance.

Observing the processed\_alert dataset, the higher the score, the higher the risk factor, since the risk\_factor is calculated by score divided by ave.score. Malicious column has two categories marked as yes or no. Furthermore, the incoming alerts will be sorted as four categories, FP, TP/DE, TP/LOW,TP/HIGH. For the Owner\_name section, it can infer to a hypothesis that senior\_analysts could deal with more alerts and show high capability.

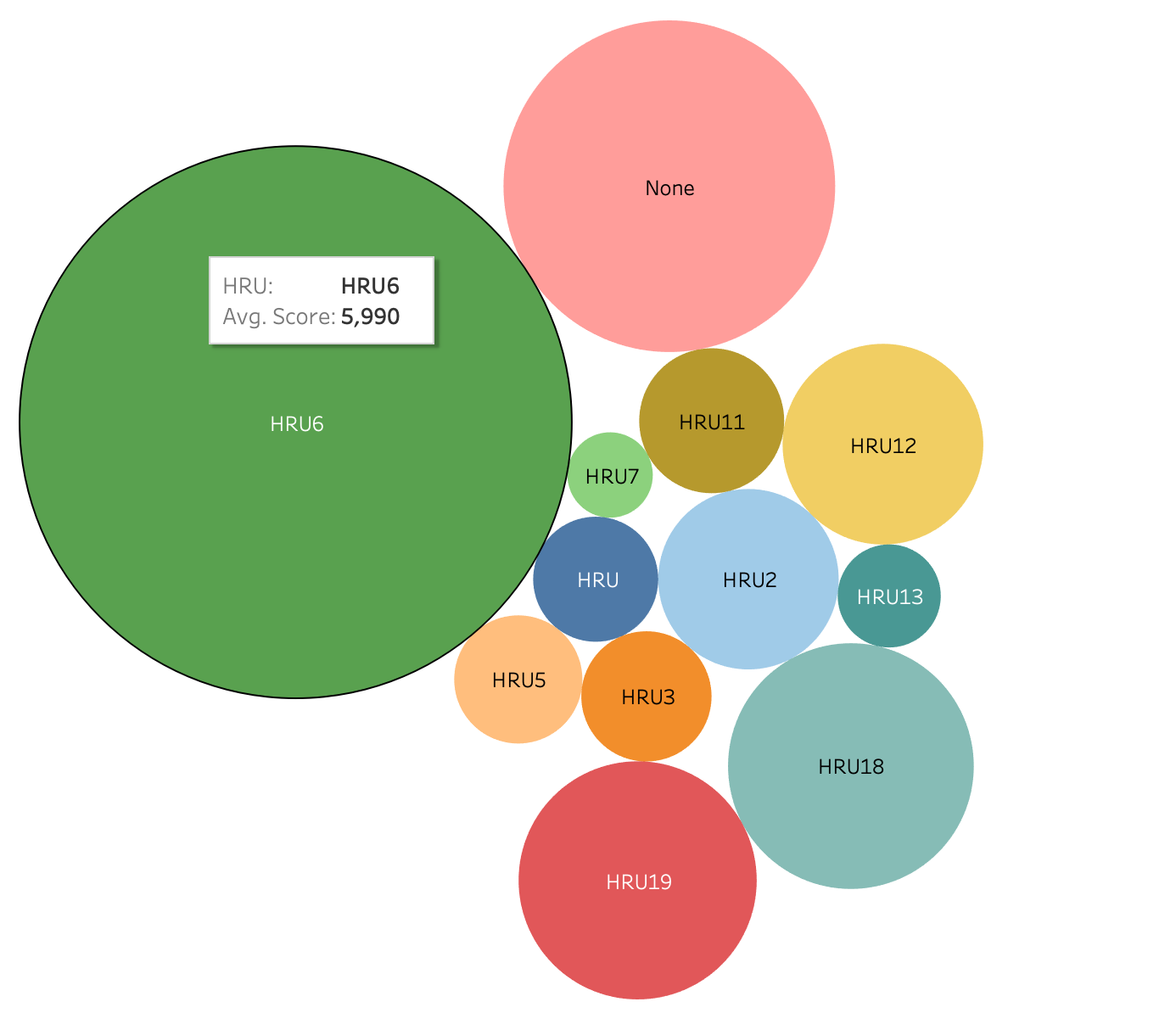
Moving on to EDA parts.



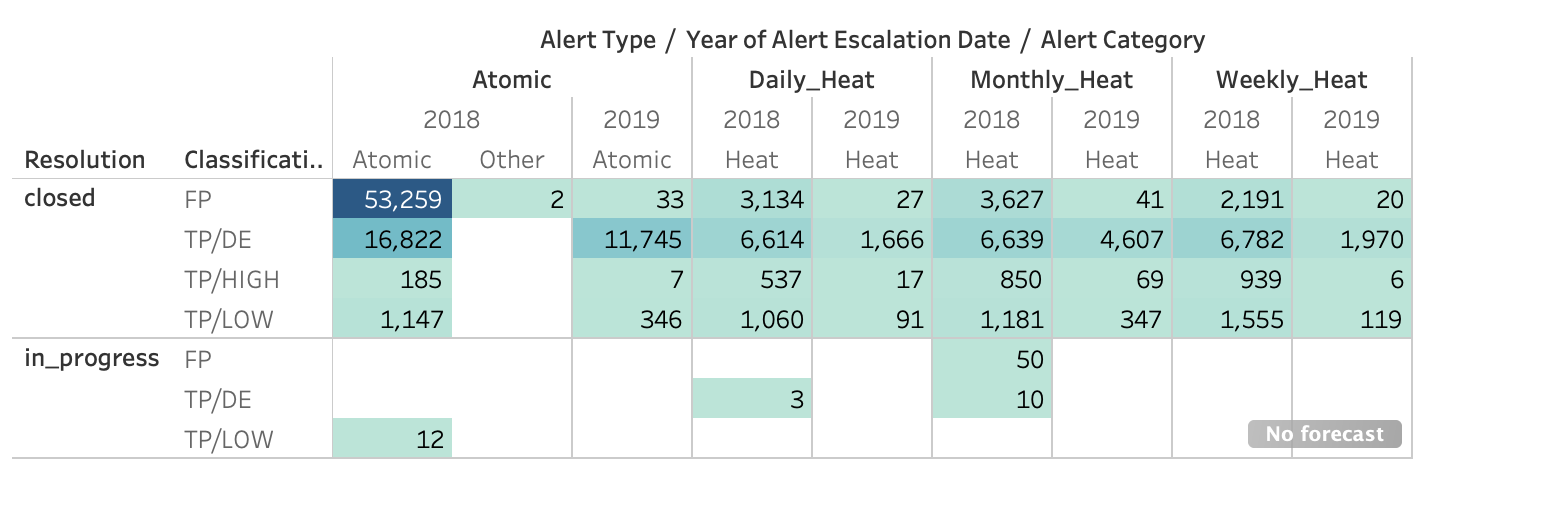
* The horizontal chart is made through Tableau tool. The length of the blue column indicates the sum of score for each category. There are four types of alert assigned by analysts. FP represents that there is no risk and the indicator misfired. Most alerts which are FP are considered as not malicious and was cleared. Among positive malicious cases, a majority of the records belong to TP/DE saying that employee tend to pass the threshold but with trial risk. Interestingly, there are a few records which involve confirmed high risks(TP/HIGH) but defined as no malicious. The system might incorrectly ascribe these TP/HIGH to no malicious.



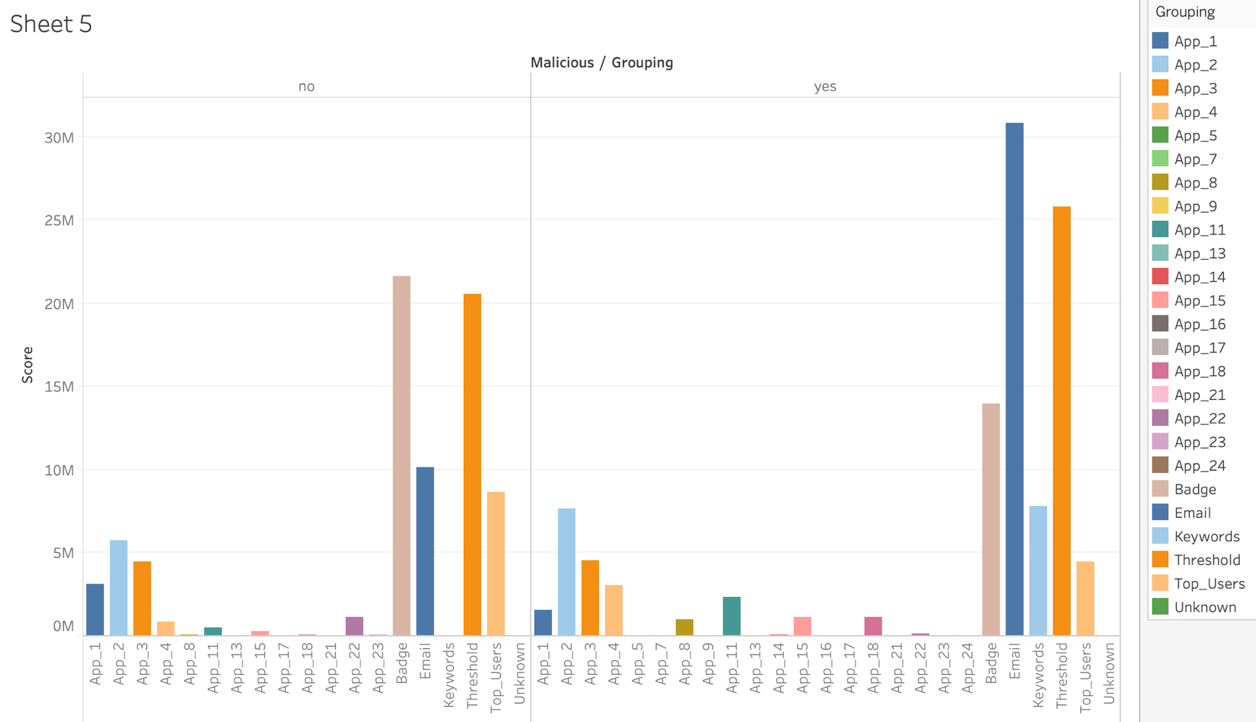
* The higher the heat score, the more likely employee computer activities associate with more risks. The average heat score for no malicious is 1178.2 which is lower than average score of malicious(1701.9).



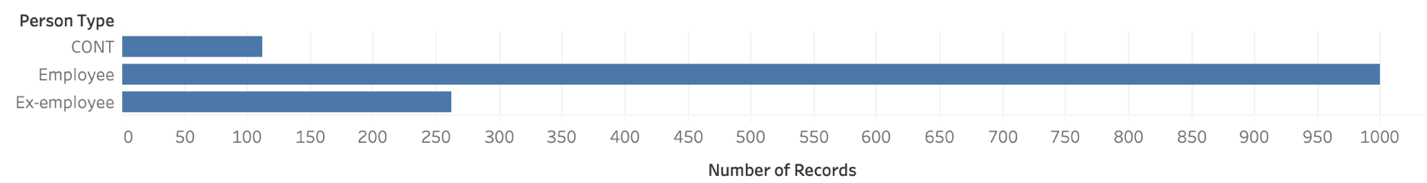
* As you can see, there are many circles shown above in different colors. The size of the circle is proportional to the average amount of heat score the system generated. Most employees who are associated with high risks are belong to HRU6 group. The second largest risky group of employees don’t belong to any group. HRU7 seems to perform relatively well with the lowest score.



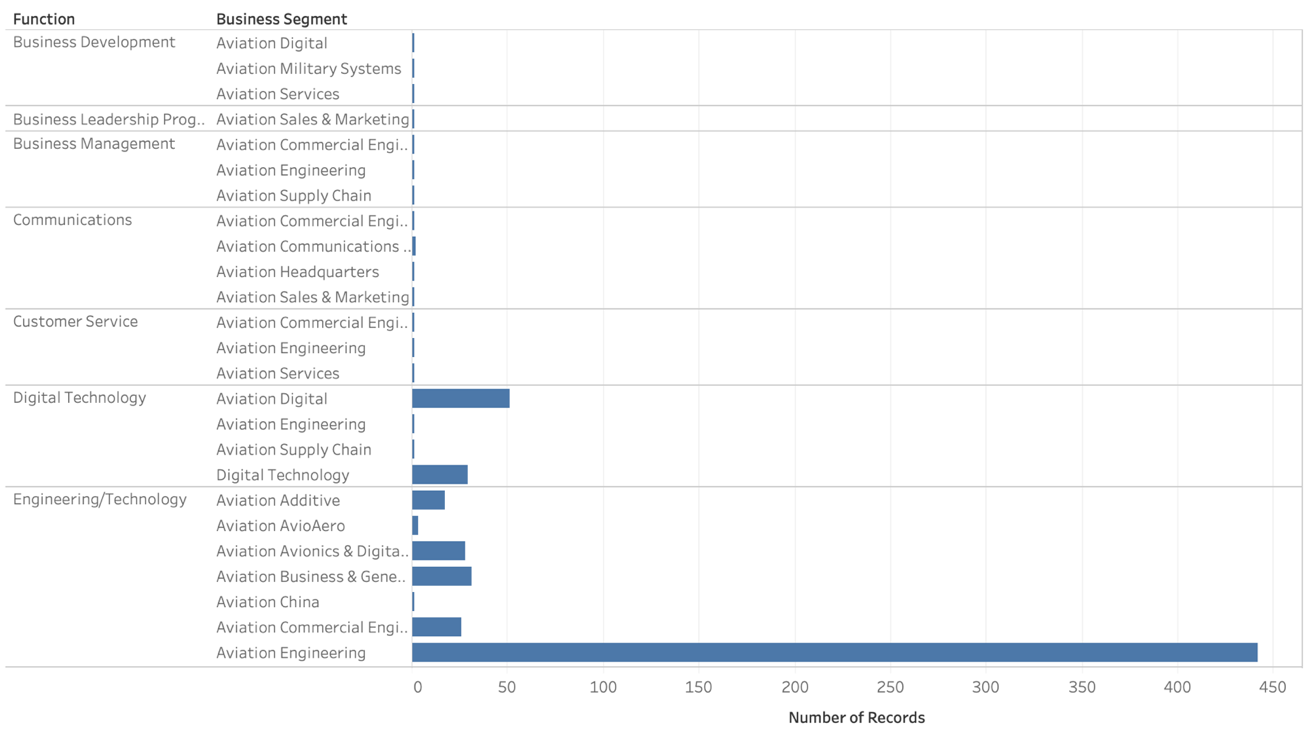
* The numbers stand for number of records accordingly for each detailed category. Only a little records settle in\_progress workflow. For the atomic alert, the average score per record is 10. The TP/DE of Daily heat alert for 2018 and 2019 differs with average score of 556 and 1173.



* The grouping column includes many variables which are used to group alerts by business or risk function. For positive malicious alerts, the blue bar representing email may be considered as a way to impair GE’s intellectual property. The threshold is the second possible variable that causes more risks.



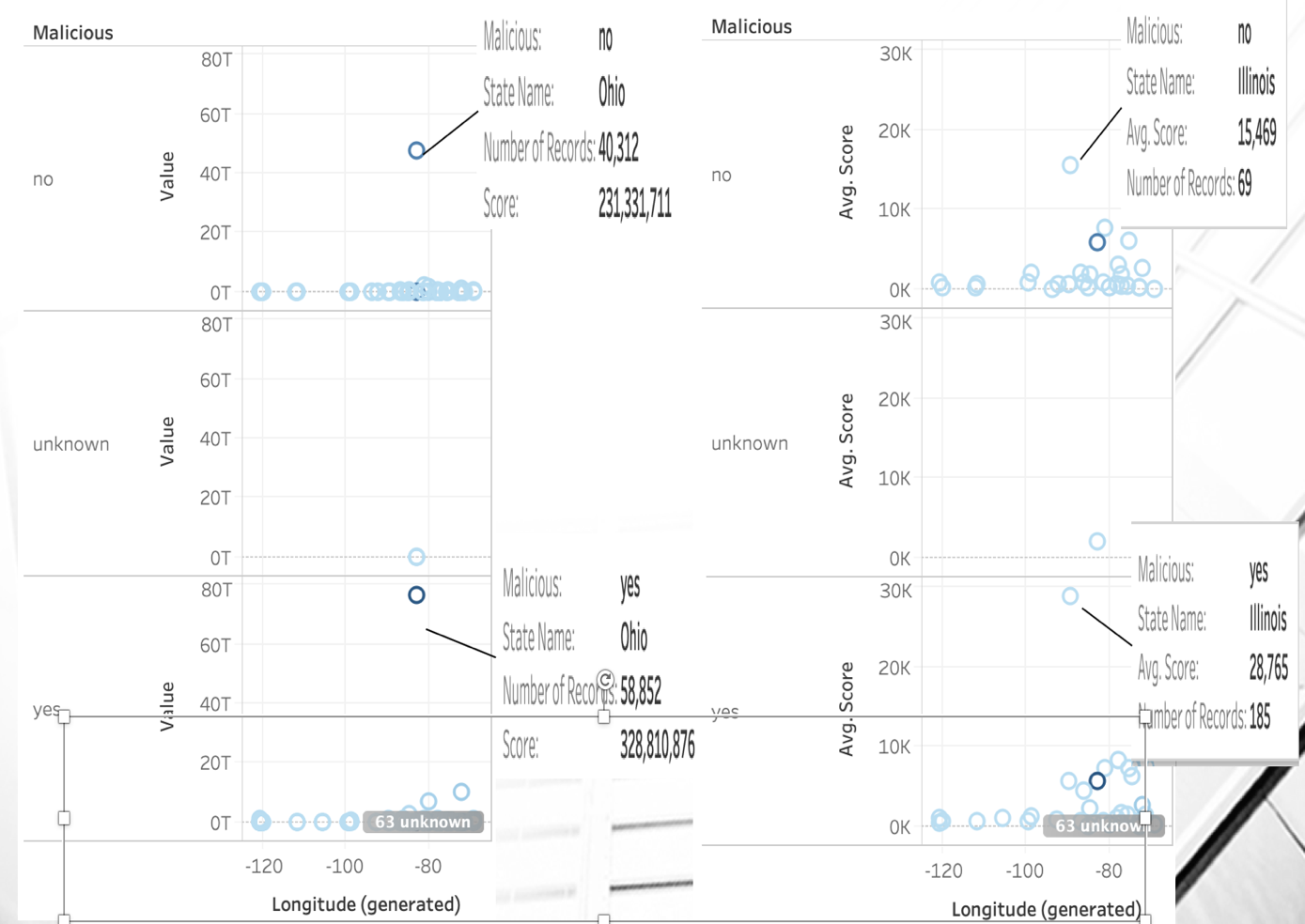
* Among the 1374 employees , there are 1000 who are currently working for GE company, 262 employees are ex-employee. 112 people are categorized as neither employee nor ex-employee.



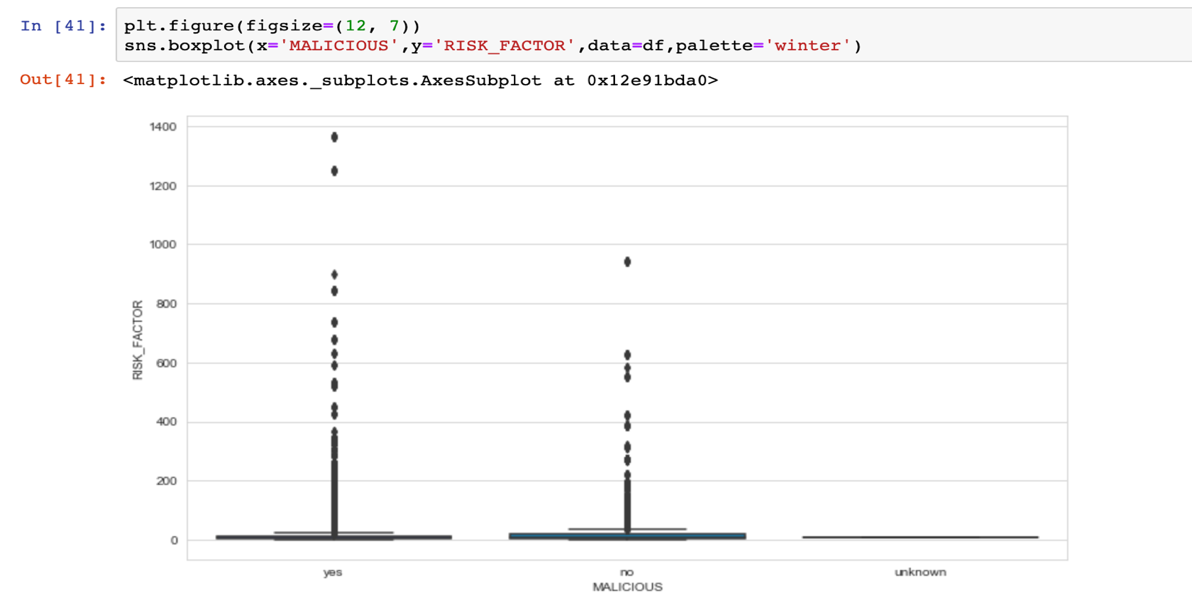
* Group by the function and business segment, we can see that 442 employees belong to the aviation Engineering business segment. Digital Technology and Engineering technology are two main functioning parts of GE company. Aviation Engineering is the most high demanded positions.



* Speaking of the composition of employees’ nationality. The majority of employees are from United States. The second largest group of employee come from China. There are 112 missing data in terms of their nationality.



* Combing the demographics and processed\_alerts table, I surprisingly found out the a majority of heat alert happened in the state of Ohio with 40312 no\_malicious and 58852 malicious records. The reason being most frequent is that many of GE Aviation headquarters are located in Evendale, Ohio, outside Cincinnati. Meanwhile, state of Ohio has the biggest sum of scores among states. On the other hand, Judged by Avg.Score, there is a noticeable point can be seen in the Illinois State with 69 records and 15469 average score which delivers a message that high risks tend to be happening in the state of Illinois.



* Here is a box plot I draw using maclicious and risk\_factor. Only a few of unknow records in the dataset which will not present any meaningful information. It’s obvious to see that malicious alert tend to have higher risk\_factor values than no\_malicious alert. There are some extreme risk\_factor values above 1200 which might be incorrectly fired alerts.

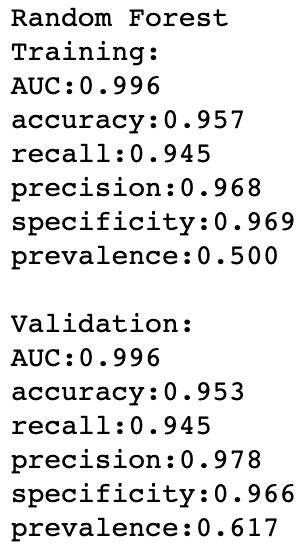
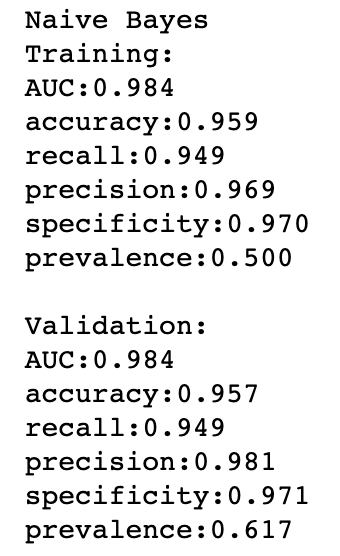
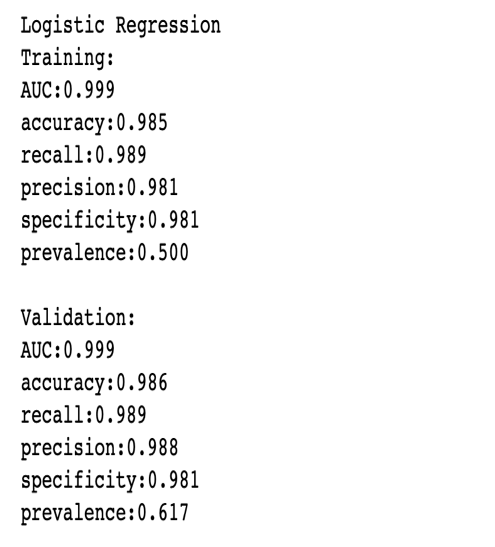
**Predictive Modeling**

Numerical variable: SCORE

Categorical variable: Status, Classification, Resolution, Owner\_name, HRU, Alert\_category, Shared\_indicator\_application

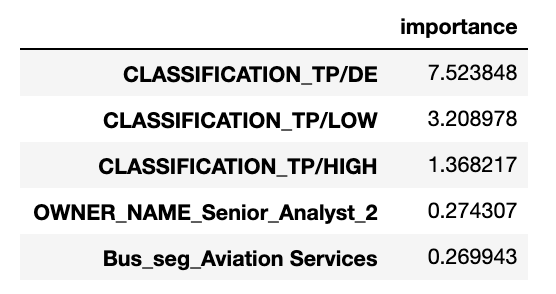
Dependent variable: Malicious

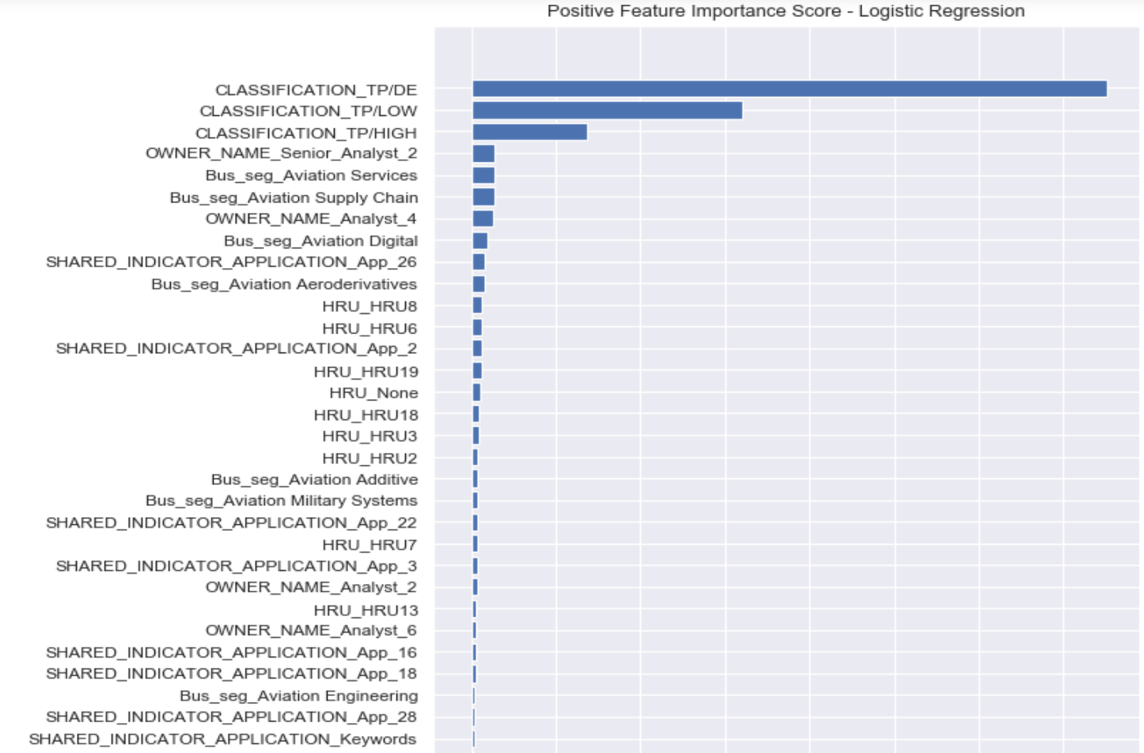
After I’ve done data importing, data cleaning, feature engineering, data split(training, testing, validation) , I manage to generate predictive modeling with Logistic regression , Random Forest, Naïve Bayes.



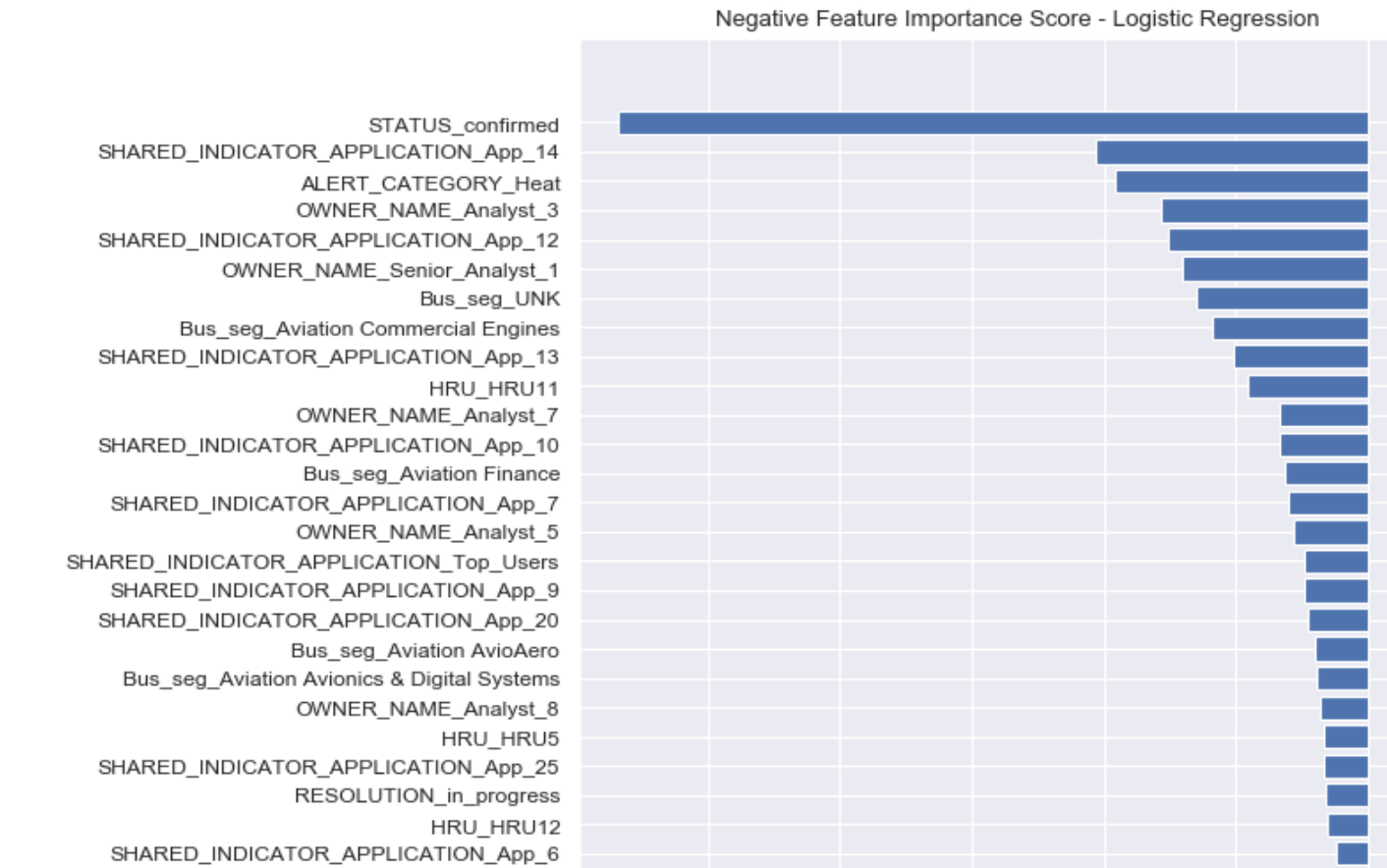
AUC is an abbreviation for area under the curve. It’s used in classification analysis to determine which of the models give you the best prediction. Logistic regression is superior to the other two models with AUC: 0.999. Besides, logistic regression give us a higher accuracy, recall, precision, specificity score in which the higher the value, the better the prediction results. Naïve Bayes and Random Forest are similar in terms of predicting malicious with chosen independent variables.

There is no overfitting problem, because the validation shows closely similar results compared with training dataset which indicate my model fits new dataset as well as it fits the data used to estimated the model. My predictive model in validation part also produces high AUC (0.999) score with 0.986 accuracy score.

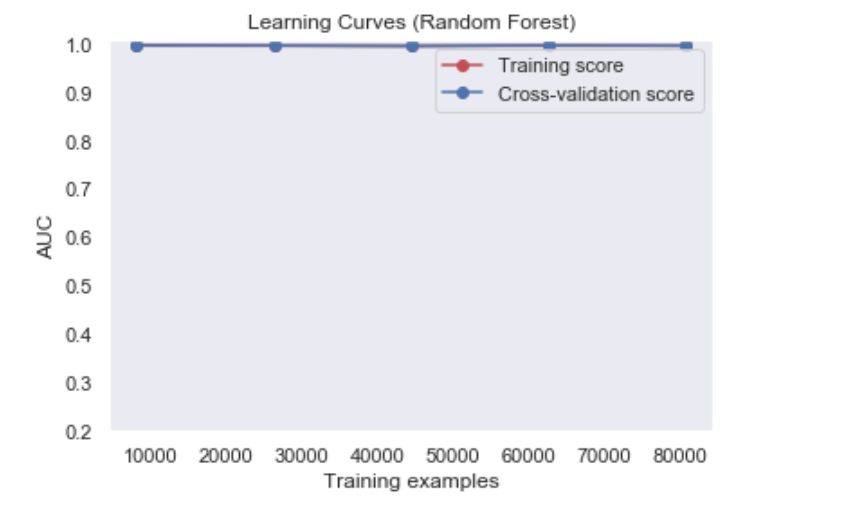




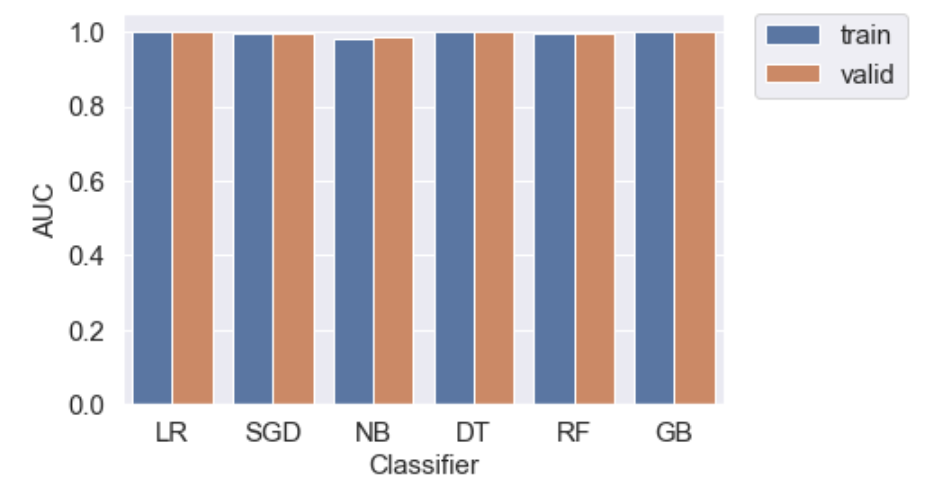
Here is a chart presenting positive feature importance score. For logistic regression, the top 3 variables including Classification\_TP/DE, Classification\_TP/LOW and Classification\_TP/HIGH with highest positive coefficients are predictive of malicious.



Likewise, the negative feature importance score with highest negative coefficients are predictive of non-malicious.



The Random Forest learning Curves show that the model has low variance and low bias. The training score and Cross-validation score are extremely close to 1.



Here is a bar chart summarizing all of the model results based on the AUC score. The naïve Bayes model performs relatively weak compared with the rest of models.

**Conclusion**

The digital guardian is approximately 99.976% accurate in terms of the FP alert. However, it lacks of accuracy speaking of the TP/HIGH. The possibility of incorrectly assigning high risky alert to no malicious is about 92%. Likewise, the system incorrectly assign TP/DE which involuves little to no risk as malicious with 99.9%. In addition,the TP/LOW is correctly categorized as malicious with 86.22% accuracy.This system is not as much as effective in predicting the TP/DE and TP/HIGH.

The digital guardian system performs badly on predicting the TP/HIGH with 92% error rate, TP/LOW with 18.65% error rate. Among the three predictive models, logistic regression is the best model we have which give us 0.985 accuracy score, 0.981 precision score. There is no overfitting or underfitting problem.

Status\_confirmed is the least important variable in predicting the malicious. With the logistic regression model, GE can quickly predict the malicious of heat alerts with a high precision and accuracy which save human costs.

**Personal impact/ Social impact**

Thanks again for the sponsor Eric and dear professor Bastos. I’ve done a lot of work on this project. Personal immediate impact includes gaining hands-on experience by finishing a applied project in the real-world. I systematically learned many machine learning methods. The data cleaning, processing process takes up the majority of my time spent on programming. The eight-weeks learning provokes that I am determined and become more confident to master the coding skills and consider it as my career goal. I am glad to show my insights to audience and acquire their applause.

Speaking of the social impact, big data has become a valuable asset for organizations. The world is connected by the Internet. Thus, it becomes crucial to protect the customer information and prevent attacks from hackers. Cybersecurity is vital to giant companies like Facebook who holds information of 2 billion users. Even if Facebook shows an extremely high sales for every year, the cybersecurity take off a large portion of the profits. Facing security breaches many times, Facebook even employed a private army to protect its own privacy and safety(Leetaru, 2019). It’s worthwhile to spend time on analyzing the ways to improve accuracy of the digital guardian system. It will be a honor if my model contributes to the improvement of the prediction, even as little as the sand. The website Kaggle provides data loves with many interesting datasets so that everyone has access to do analysis. It’s worth promoting and encouraging public to give it a try, together with contributing to the work. I sincerely admire the action that Netflix awarded 1 million prize to developer team in 2009 for an algorithm that increased the accuracy of the company’s recommendation engine by 10 percent(Casey & Johnson, 2017). It’s a mutual benefits saying that motivation for students, gains for companies.

**References**

Zhang, E. (2019, July 15). What is Data Loss Prevention (DLP)? A Definition of Data Loss Prevention. Retrieved from https://digitalguardian.com/blog/what-data-loss-prevention-dlp-definition-data-loss-prevention

Teece, D. J. (2000, July 03). Strategies for Managing Knowledge Assets: The Role of Firm Structure and Industrial Context. *Long Range Planning,33(1)*, February 2000, pp.35-54.

Albashrawi, M. (2016). Detecting financial fraud using data mining techiniques: a decade review from 2004 to 2015. *Journal of Data Science,* 553-570. Retrieved from <file:///Users/atple/Downloads/%E6%94%B910-Detecting%20Financial%20Fraud%20Using%20Data%20Mining%20Techniques-JDS_V3.pdf>

Founder of Bixby Energy Systems Sentenced to 25 Years in Prison for Stealing More Than $56 Million from Investors . (2014, September 30). Retrieved from https://www.fbi.gov/contact-us/field-offices/minneapolis/news/press-releases/founder-of-bixby-energy-systems-sentenced-to-25-years-in-prison-for-stealing-more-than-56-million-from-investors

Chitra, K., & Subashini, B. (2013, August). Data Mining Techniques and its Applications in Banking Sector. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.413.9171&rep=rep1&type=pdf

Shorten A, Smith J. Mixed methods research: expanding the evidence base *Evidence-Based Nursing*2017;**20:**74-75.

Leetaru, K. (2019, March 23). Facebook Succeeded In Killing Cybersecurity Like It Did Privacy. Retrieved from <https://www.forbes.com/sites/kalevleetaru/2019/03/23/facebook-succeeded-in-killing-cybersecurity-like-it-did-privacy/#3cd380c04549>

Casey Johnston, A. T. (2017, June 3). Netflix Never Used Its $1 Million Algorithm Due To Engineering Costs. Retrieved from https://www.wired.com/2012/04/netflix-prize-costs/