STAT5525 Final Project

Group 1003

2017-12-07

Abstract

This paper discusses the methods and results of various statistical learning techniques applied to unstructured data. We discuss our feature development and model performance in a business analytics setting. Also included are computational challenges and compromises in our analyses. We conclude with interpretatations and usefulness of results.

Introduction

Companies with many employees face challenges monitoring and evaluating the behavior of their employees. By utilizing data collected from employees, such as web traffic, email contents, file contents, and logon information, a company can expect to gain valuable insight into the the potential for employee attrition. Predicting employee attrition is a mulifaceted issue, involving behavior both inside and outside the office. Using all available information for both ACME and DTAA, variables to predict employee attrition were created and modeled using various classification techniques. In addition to these models, both data sets were searched for signs of unscrupulous behavior such as spam, data theft, or personal conflict. These are traits of substandard employees. By integrating these findings within and across company information, some interesting trends and important variables emerged that each company can find useful to monitor employee activity and ensure quality work.

Data Handling

Because the combined data sets were too large for local processing, we used the ARC clusters for data storage and processing. For intial data exploration and local processing, the larger data sets were aggregated by month. These smaller data sets then became the training sets for jobs that would later be processed over the entire data sets using the ARC machines. This way we were able to process data more efficiently with shorter runtimes than local processing.

The separate data sets were combined several different ways for additional visualization and summarization. The timestamped data sets were each merged together into seveal different data structures for different types of processing. For getting a sense of a user's daily activity, we created a single data structure consisting of user timestamps from logon information, USB device connections and file downloads, receiving and sending emails with and without attachments, and HTTP usage. This data was then visualized for a given user in a `gif` over time. These timestamps were augmented with other useful information such as the weekday, employee's first logon of that day, indicator of before/after business hours, and holiday indicator. From this information, we found that several DTAA employees regularly worked after-hours and on holidays, whereas ACME employees did not work on holidays. In each data set, there were several instances of consecutive user logons due to screen-locks, as well as USB connections without a matching disconnect. These lonely logons or USB disconnects were inserted into the data with the average logon to logoff or connect to disconnect time of the respective user.

A separate data structure was also created with the contents of the file downloads, emails, and HTTP visits. Word clouds were then generated for each content type per user and across the company. Here we noticed an interesting trend that the words "prince", "ahmose", and "ankh" had by far the highest frequencies across the different contents. We determined this type of content to be spam and the result of malware spreading through the company. HTTP content containing these terms was determined to contain the substring "pnerreqriry". Additionally, from analyzing sentiment scores from email content within roles, we found that the IT administrators had lower scores than average. Looking closer, we found several instances

of employee-supervisor emails complaining about working weekends, after-hours, and not feeling appreciated with replies to the supervisor rebuttle such as "company will suffer". Furthermore, the day after was the respective employee's last. Further exploration shows each of these IT administrators to have downloaded files containing malicious keylogging software. This software is then installed on the supervisor machine between 5:30-8:00pm. The guilty IT administrators work one more day before logging into the supervisor's primary machine after hours to presumably extract this information. We focused on variable creation with instances like these in mind.

Variable Creation

Using all available information, we wanted to create several variables that would accurately predict attrition. We defined employee attrition as those employees who do not appear in the final LDAP data set. Our approach was twofold; create variables that reflect uncommon work habits and others that reflect malicious behavior. We flagged any user activity that was after hours in two different ways, 5:30-11:59pm and 12:00-6:00am. These two different classifications of user traffic allowed us to differentiate between working late and suspicious behavior. Additionally, we made note of how many different PCs a user is logging onto, connecting USB devices, and downloading files from. We would expect IT administrators to have high traffic here for updating or repairing machines, however we would find it suspicious for a janitor to logon to over 60 different machines over the course of their employment or a technician with no USB traffic to download a file off of someone else's machine late at night. USB file download types were also aggregated, with particular focus on `.exe` files. We then wanted to also highlight supervisor machines, so we created indicators for logging onto these machines and downloading files. We would expect these variables to explain data theft or malware installation. Another form of possible data theft consisted of email attachments to non-DTAA users.

In addition to identifying behavior with possible malicious intent, we also created several variables that would model non-malicious behaviors. Sentiment scores were calculated for the contents of each user's emails, file downloads, and HTTP visits. From this information, coupled with the psychometric data set, we were able to get an overall look into a user's personality. The email sentiment scores were separately calculated for emails to and from a user's supervisor. We would expect that a high frequency of low sentiment emails between the user and their supervisor to be reflective of personal conflict and be an intuitive predictor of attrition. We created similar scores for emails that contain employee names with a similar thought process. Counts were also aggregated for user's sending and receiving non-work emails. While we would expect that some users will check their non-professional email each day, a majority of email communication with non-coworkers during the work day is an unproductive use of company time. Spam indicators were also created to track the spread of the "prince" emails, both sending and receiving. Figure 1 shows the number of unique users sending the spam email over time. An employee sending 30% of these spam emails could be the result of a virus or bot traffic. Spam indicators were also created for HTTP visits and file downloads. Indicators were also created for users who visit career search or dating

websites. We expect career searching to be a reliable indicator of user attrition if a user is actively searching for a new job. We also considered users visiting more than 15 websites in under one minute to be reflective of potential bot traffic. These variables could explain user attrition for non-malicious user behavior.

Data Aggregation

The HTTP data set proved useful in forming various clusters of employees. The first series of analyses focuses on the times of day at which employees access the internet. Via preliminary data exploration and visualization, we noticed that employees tend to access websites most frequently at the beginning of their workday and at the end of the workday. For some employees, we see a substantial dip in internet usage around lunchtime while for other employees, internet usage appears to be relatively steady between the two modes. In order to classify employees by their internet usage throughout the day, we decided to mine parameters of interest from approximated *probability distribution functions* (PDF) obtained via *Kernel Density Estimation* (KDE). Figure 2 shows a KDE for the PDF of internet usage throughout the day for a given employee. This figure identifies five key points on the shape of the PDF which were used for further feature extraction. `D1` and `D2` are points where the PDF crosses a threshold of 0.02. These points serve as a biased proxy for average start time and average end time, respectively, for an employee's work day.

The objective of the first clustering analysis was to classify employees into groups based on the shape of their internet usage over time PDF. The metric used to describe the shape of each estimated PDF is the depth of the trough relative to the average height of the two crests (modes), scaled by the average of the two crests. Assuming the data come from a Gaussian mixture model, we used the *EM* algorithm to classify employees into two different groups based on the shape of their internet usage throughout the workday estimated PDF. The two sets of distributions are shown in Figures 3a & 3b.

Apart from the shape of an individual's internet usage (time of day) distribution, we found clear groups of individuals based on biased estimates of start time of their workday and length of their workday. Clusters are clear for each of these parameters, but they are most clear when considered together as shown in Figure 4. Due to the compactness of these clusters and the large number of clusters, we thought the most efficient clustering method would be a density-based method such as *DBSCAN*. The clusters shown in Figure 4 were identified with the *DBSCAN* algorithm.

Explanatory & Predictive Modeling

To accurately model user attrition over time, feature data sets were generated for each time period from start date to each employee attrition date. This way our feature space was not biased with the lesser traffic from a user who left the company within the first month, for example. After each employee attrition date, those employees were then removed for model

fits on later attrition dates. In addition to being fit over all employee attrition rates, 10-fold cross-validation was also performed with the overall classification error being reported.

Our first model was a non-malicious behavioral-based attrition classifier. Logistic LASSO, CART, and Random Forest models were fit using total non-work emails, negative sentiment scores, employee gossip, supervisor email sentiments, spam indication, and psychometric scores. We found that spam indication was a strong indicator of no attrition. That is, sending and receiving spam emails had a high probability of current employment. Our interpretation suggests that these emails are being sent without the users knowledge. Potentially, the original spammers are targeting higher fixtures in the company to make the spam appear more credible. If true, this would explain the low attrition rates for this type of behavior.

Assuming this spam correlation is not useful, after removal, we achieved similar prediction accuracy. This suggests that labeling spam is a confounding variable. Interestingly, this model gives high importanc to negative sentiment scores for user emails. Adding positive sentiment scores to our model and refitting, the results of this classifier are shown in Figure 5.

We next modeled features we derived to describe malicious user behavior, such as after hour file downloads, email attachments to outside users, logging onto non-primary machines after-hours, and the like. Logistic LASSO, CART, and Random Forest methods were applied to these features to predict attrition. We see an overall improvement in attrition prediction in each method compared to the non-malicious models fit previously. We find that important features in these models are supervisor PC access and after-hour file downloads. The distribution of error rates are shown in Figure 6.

Model Justification

As the misclassification figures show, accuracy was comparable across methods for each model fit. This was likely the case of CART and LASSO identifying similar importance of features. We would also expected the Random Forest method to perform similar to CART because Random Forest is a bootstrapped version of CART. From these results, we choose the LASSO for it's flexibility in adding additional variables and quick training and testing runtimes. We also choose the LASSO for its performance with correlated features like above. We also choose the feature set modeled by malicious user behavior. These features have clear interpretations and should be carefully monitored in each company.

Computational Considerations & Demands

Initial computational challenges were primarily data exploration and aggregation. Our solution was to use various software packages to group data monthly for individual processing and before combining the results. Making `gif`s of user logon and logoff activity over time was also an expensive CPU process. ARC's resources allowed us to make these visualizations in parallel much faster than local processing. These movies were our first steps in finding research questions and the additional processing power from the ARC was especially important for the additional size of the DTAA information. An additional step in our visualization process

were network graphs of user interactions. For ACME, these networks were easily created and visualized across PC connections. However, for the DTAA email information, the networks were time consuming to create with the ARC machines, but practically infeasible based on the large number of emails users send and receive. Also searching for employee names in email contents was very computational intensive and processed in chunks similar to above. The most computationally difficult part of this project was the actual model fitting. Fitting multiple models to feature sets with cross-validation that were regenerated at every employee attrition date made it almost impossible to tune model parameters. Our solution was to impliment subroutines that performed grid searches to find the optimal model parameters for each model across every employee attrition date. We thus elected to not fit neural networks to save computational time.

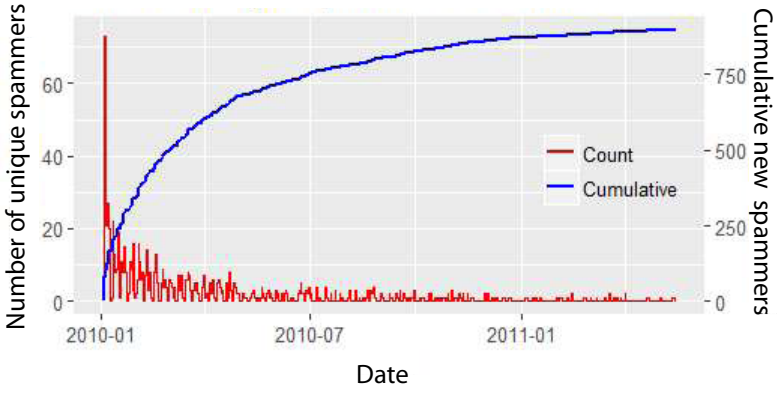
Conclusions

The goal of these analyses was to predict user turnover through detection of suspicious behavior, other abnormal behaviors, and sentiment analysis. We have highlighted several factors that suggest these and other suspicious behaviors have malicious intent and it is in a company's best interest to identify these individuals as potentially dangerous. Modeling the email behavior of employees showed a clear trend towards reduced chance of attrition with more positive sentiment in emails. Further analysis of the email contents allowed us to discover a spam email that has spread throughout the company, including into both file download contents and HTTP contents. By clustering using *DBSCAN*, we were able to cluster user HTTP traffic into two different groups; those with high traffic in the middle of the day versus those with high traffic at the beginning/end of the workday. While this clustering is not highly correlated with employee attrition, it is important for a company to identifiy periods of low productivity. By also monitoring these activities, companies can flag users with suspicious activity in real time.

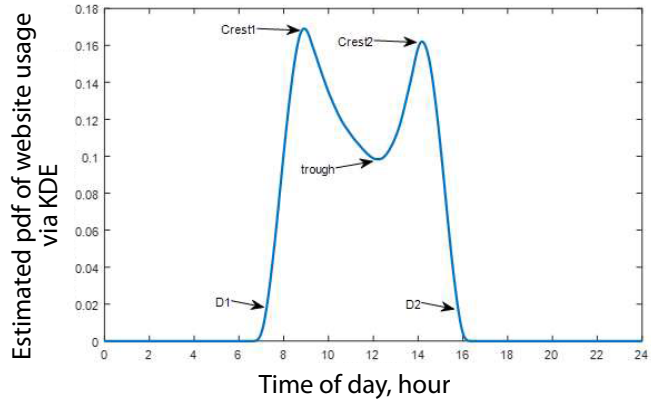
Through our analysis of these companies, we noticed that supervisors were much less likely to leave the company compared to other roles. In the DTAA data, no supervisors left over the entire time period. One thing this company could do moving forward is look more closely at supervisors with similar traits to many fired employees. While this is not proof that this employee should be fired, by identifying potential employees worth firing, we are illustrating how this predictive analysis can be used. This analysis is beneficial to a company for removing malicious employees that could cost the company money, time, or both. Identifying employees likely to leave is a great advantage to companies as it is very difficult to find replacements in a timely manner when employees quit suddenly. By allowing a head-start in the recruiting and firing processes, strain on the company and potential confusion can be avoided or limited.

Appendix

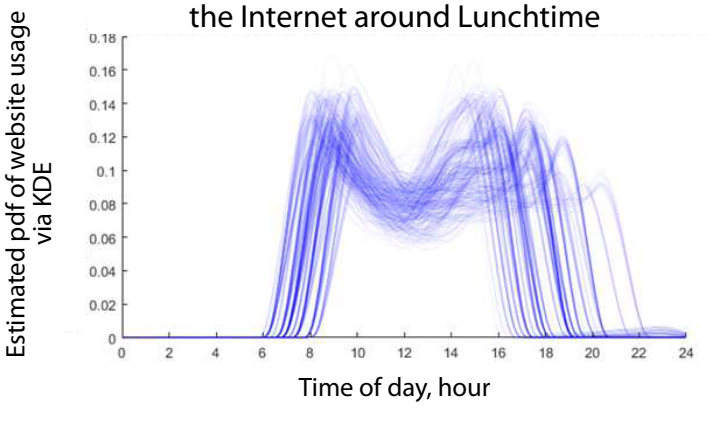
(1) Growth of unique spammers over time



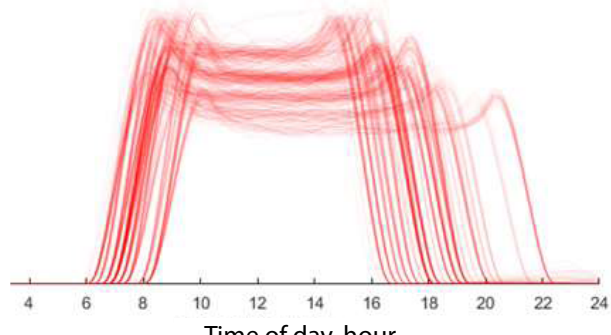
(2) Kernel density estimates for PDF of internet usage



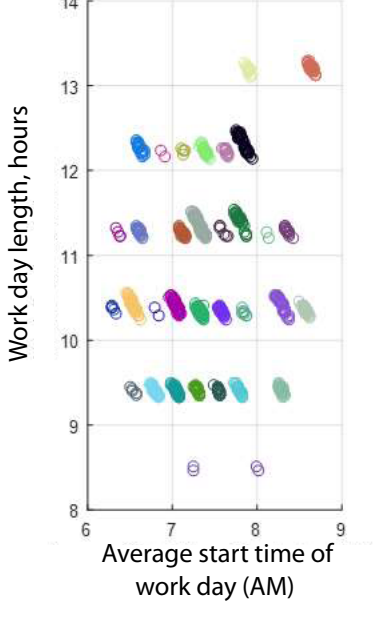
(3a) Estimated Distributions Users Abstain from the Internet around Lunchtime



(3b) Estimated distributions of all users using internet through lunchtime



(4) DBSCAN clusters of internet usage



(5) Error Rates For Email Model Fits

(6) Error Rates For Malicious Activity