STAT5525 Final Project

Group 1003

2017-12-05

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

Becausee the combined data sets were too large for loca 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 in 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 substrig "pnerreqriry". Additionally, from analyzing sentiment scores from email content within roles, we found that the IT administrators had lower scores than averaage. Looking closelier, 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

Data Aggregation

Explanatory/Predictive Modeling

Model Justification

Computational Considerations And Demands

Computational Modeling Choices

Conclusions