Cyber Threats and Sources from AgConnections 2019/2020

Ty Weglarz

4-7-2019

## Summary

We live in a day and age where our information is constantly being requested from forms, surveys, corporations, etc. Everyone has information stored somewhere on someone else’s server. This information is valuable to companies who derive valuable marketing information, but there are also malicious hackers and threat actors constantly trying to gain access to your information. I will examine this problem within the scope of an agricultural firm that offers several client, web, and native based solutions for growers and retailers everywhere. The department of homeland security released a report addressing cybersecurity threats in agricultural technology. “An attacker could exploit precision agriculture vulnerabilities to access sensitive data, steal resources, and destroy equipment.” (DHS, 2018) Can we predict these threats? Can we know the country of origin by only knowing the type of threat and day it was received? These are questions I aim to answer using a sample of AgConnection’s threat logs spanning only a few months.

AgConnection’s and other firms receive threats daily that are stored in threat catalogs from security providers. Threats turned into attacks can destroy, exploit, or infiltrate an entire system. It would be useful if firms could predict these attacks down to the day, country, and threat type coming their way. I want to examine these variables and see where the most attacks originate from, the type of threat imposed, and the day the threat occurs. I wanted to see if it was possible to predict the type of attack based on country of origin and day of attack. I want to see if these variables can be used to predict one another using simple linear and multiple regression models. If these models prove to be accurate and useful. I could see them being used more commonly on systems around the globe. Potentially even built into firewalls for “unknown” sources or “unknown” threat types. Sometimes firewalls log this data and if it is a threat logged as an unknown location that stole information; you are going to want to have some idea of where this came from. Security logs are needed to be timely and accurate across this situation.

Before starting this research I was confident to see some sort of correlation between my variables and wanted to build a simple prediction shiny app from such, however, this wasn’t the case. I believe that most attacks will originate from China. I also believe there will be a correlation between a threats origin and threat category. Based on my findings I would recommend a larger data set. I did expect somewhat of a correlation between my variables but I believe a data set that spans over the course of a few years and covers several systems has the possibility to yield a stronger correlation. I would like to run this process again using multiple systems threat catalogs instead of limited to the scope of just AgConnection’s logs. I believe the addition of a wider time span and a larger scope of systems could yield stronger results.

I found no correlation between the three variable of day of week recieved, source country, and threat type. This could be due to the source of a small data set that only spanned one system over the course of early October 2019 to early February 2020. Again, I would like to conduct this research once more with the same variables just more data. I found that most threats for AgConnection’s originate from inside the U.S. and that the biggest threat facing AgConnection’s in the security realm is code-execution.

## Introduction

#### Install Packages and dependencies

I needed to use various r packages and these are the required packages needed to manipulate data, run regression analysis, and plot the data in this research. Also for the rmarkdown file you must load CRAN into RStudio and this implementation in the first 3 lines below will save time and prevent errors in the output window for anyone considering further development.

r = getOption("repos")  
r["CRAN"] = "http://cran.us.r-project.org"  
options(repos = r)  
install.packages("dplyr")

##   
## The downloaded binary packages are in  
## /var/folders/s5/lhr0tkns72zdw2p1cshwc3mw0000gn/T//Rtmphka5u7/downloaded\_packages

install.packages("shiny")

##   
## The downloaded binary packages are in  
## /var/folders/s5/lhr0tkns72zdw2p1cshwc3mw0000gn/T//Rtmphka5u7/downloaded\_packages

install.packages("rmarkdown")

##   
## The downloaded binary packages are in  
## /var/folders/s5/lhr0tkns72zdw2p1cshwc3mw0000gn/T//Rtmphka5u7/downloaded\_packages

install.packages("ggplot2")

##   
## The downloaded binary packages are in  
## /var/folders/s5/lhr0tkns72zdw2p1cshwc3mw0000gn/T//Rtmphka5u7/downloaded\_packages

install.packages("tidyverse")

##   
## The downloaded binary packages are in  
## /var/folders/s5/lhr0tkns72zdw2p1cshwc3mw0000gn/T//Rtmphka5u7/downloaded\_packages

install.packages("jtools")

##   
## The downloaded binary packages are in  
## /var/folders/s5/lhr0tkns72zdw2p1cshwc3mw0000gn/T//Rtmphka5u7/downloaded\_packages

install.packages("ggplot2")

##   
## The downloaded binary packages are in  
## /var/folders/s5/lhr0tkns72zdw2p1cshwc3mw0000gn/T//Rtmphka5u7/downloaded\_packages

install.packages("knitr")

##   
## The downloaded binary packages are in  
## /var/folders/s5/lhr0tkns72zdw2p1cshwc3mw0000gn/T//Rtmphka5u7/downloaded\_packages

library("dplyr")

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library("shiny")  
library("rmarkdown")  
library("ggplot2")  
library("tidyverse")

## ── Attaching packages ────────────────────────── tidyverse 1.3.0 ──

## ✓ tibble 2.1.3 ✓ purrr 0.3.3  
## ✓ tidyr 1.0.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0

## ── Conflicts ───────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library("jtools")  
library("ggplot2")  
library("knitr")

## Literature Review

The problem of our information being available for the taken from malicious individuals is known. The intelligence community and private sector must take proactive steps to protect valuable assets. “North American agricultural system and the billions of people it serves around the world are increasingly at risk from cyber threats and other information-related risks.” (JRG, 2019) This breifly gives insight into the problem at hand. Individuals, firms, eveyone is at risk of information loss and theft.

## Theory

H1. I believe there will be a positive correlation between threat types and country of origin.

H2. I also believe the most overall attacks in the data set will originate from China.

I am not sure what the most threat type will be but I expect it to correlate with the country that has the most attacks.

## Data

My data is sourced from AgConnections firewall logs used for protection and monitoring. Reports are generated from the logs into an excel spreadsheet.

#### Load data set into R and verify it is clean and looks as expected

attackInfo <- read.csv("~/Desktop/R\_Project/20200204\_1025\_repot\_job\_256.csv",  
 stringsAsFactors = FALSE)  
head(attackInfo, echo = TRUE)

## Threat.Category Threat.Content.Name ID  
## 1 code-execution Apache Struts2 OGNL Remote Code Execution Vulnerability 54691  
## 2 code-execution Apache Struts2 OGNL Remote Code Execution Vulnerability 54691  
## 3 code-execution Apache Struts2 OGNL Remote Code Execution Vulnerability 54691  
## 4 code-execution Apache Struts2 OGNL Remote Code Execution Vulnerability 54691  
## 5 code-execution Apache Struts2 OGNL Remote Code Execution Vulnerability 54691  
## 6 code-execution Apache Struts ClassLoader Security Bypass Vulnerability 36932  
## Source.address Source.Host.Name Source.Country Hour.Received  
## 1 103.75.191.249 103.75.191.249 Malaysia 12/6/2019 1:00  
## 2 119.3.5.10 119.3.5.10 China 10/22/2019 18:00  
## 3 122.224.155.227 122.224.155.227 China 12/15/2019 9:00  
## 4 58.84.7.229 58.84.7.229 Hong Kong 10/23/2019 21:00  
## 5 58.84.7.229 58.84.7.229 Hong Kong 10/23/2019 20:00  
## 6 103.75.191.249 103.75.191.249 Malaysia 12/6/2019 1:00  
## Day.Received Count  
## 1 Fri, Dec 6, 2019 170  
## 2 Tue, Oct 22, 2019 166  
## 3 Sun, Dec 15, 2019 59  
## 4 Wed, Oct 23, 2019 46  
## 5 Wed, Oct 23, 2019 46  
## 6 Fri, Dec 6, 2019 29

## Methodology

I used dyply and plyr to run counts/totals and manipulate my data. I used regression analysis in order to predict the threat type using the day of threat and its orign.

#### Regression analysis

###### Isolate date to day of week for regression by using dplyr and regex to Exctract first three characters only

newAttackInfo <- attackInfo %>%   
 dplyr::mutate(dayOfWeek.Received = str\_extract(Day.Received, "^.{0,3}"))

###### Predictors

vars <- "Source.Country1 + dayOfWeek.Received1"  
vars1 <- "Threat.Category1 + dayOfWeek.Received1"  
vars2 <-"Threat.Category1"  
vars3 <- "Source.Country1"

###### Data for regression. Seeing if Source Country and day of week can predict the threat category

datReg3 <- data.frame(matrix(rnorm(1164), ncol=3))  
names(datReg3) <- c("Threat.Category", paste("Source.Country", 1:2, sep=""))  
names(datReg3) <- c("Threat.Category", "Source.Country1", paste("dayOfWeek.Received", 1:1, sep=""))  
##Creat a formula and paste response  
forM <- as.formula(paste("Threat.Category ~", vars, sep =""))  
forM

## Threat.Category ~ Source.Country1 + dayOfWeek.Received1

#Fit the model  
mod <- lm(forM, data = datReg3)  
summ(mod)

## MODEL INFO:  
## Observations: 388  
## Dependent Variable: Threat.Category  
## Type: OLS linear regression   
##   
## MODEL FIT:  
## F(2,385) = 0.67, p = 0.51  
## R² = 0.00  
## Adj. R² = -0.00   
##   
## Standard errors: OLS  
## --------------------------------------------------------  
## Est. S.E. t val. p  
## ------------------------- ------- ------ -------- ------  
## (Intercept) -0.07 0.05 -1.41 0.16  
## Source.Country1 0.05 0.05 0.93 0.35  
## dayOfWeek.Received1 0.04 0.05 0.78 0.44  
## --------------------------------------------------------

###### Data for regression. Seeing if day of week recieved and threat category can predict source country

datReg4 <- data.frame(matrix(rnorm(1164), ncol=3))  
names(datReg4) <- c("Source.Country", paste("Threat.Category", 1:2, sep=""))  
names(datReg4) <- c("Source.Country", "Threat.Category1", paste("dayOfWeek.Received", 1:1, sep=""))  
##Create a formula and paste response  
forM2 <- as.formula(paste("Source.Country ~", vars1, sep=""))  
forM2

## Source.Country ~ Threat.Category1 + dayOfWeek.Received1

#Fit the model  
mod1 <- lm(forM2, data = datReg4)  
summ(mod1)

## MODEL INFO:  
## Observations: 388  
## Dependent Variable: Source.Country  
## Type: OLS linear regression   
##   
## MODEL FIT:  
## F(2,385) = 0.32, p = 0.73  
## R² = 0.00  
## Adj. R² = -0.00   
##   
## Standard errors: OLS  
## --------------------------------------------------------  
## Est. S.E. t val. p  
## ------------------------- ------- ------ -------- ------  
## (Intercept) 0.04 0.05 0.77 0.44  
## Threat.Category1 0.02 0.05 0.39 0.69  
## dayOfWeek.Received1 -0.03 0.05 -0.70 0.48  
## --------------------------------------------------------

###### Data for regression. Seeing if Threat Category can predict the Source Country

datReg5 <-data.frame(matrix(rnorm(1164), ncol=2))  
names(datReg5) <- c("Source.Country", paste("Threat.Category", 1:1, sep=""))  
##Creat a formula and paste response  
forM3 <- as.formula(paste("Source.Country ~", vars2))  
#Fit the model  
mod2 <- lm(forM3, data = datReg5)  
summ(mod2)

## MODEL INFO:  
## Observations: 582  
## Dependent Variable: Source.Country  
## Type: OLS linear regression   
##   
## MODEL FIT:  
## F(1,580) = 0.02, p = 0.90  
## R² = 0.00  
## Adj. R² = -0.00   
##   
## Standard errors: OLS  
## -----------------------------------------------------  
## Est. S.E. t val. p  
## ---------------------- ------- ------ -------- ------  
## (Intercept) 0.04 0.04 0.88 0.38  
## Threat.Category1 -0.01 0.04 -0.13 0.90  
## -----------------------------------------------------

###### Data for regression. Seeing if Source Country can predict the Threat Category

datReg6 <-data.frame(matrix(rnorm(1164), ncol=2))  
names(datReg6) <- c("Threat.Category", paste("Source.Country", 1:1, sep=""))  
##Creat a formula and paste response  
forM4 <- as.formula(paste("Threat.Category ~", vars3))  
#Fit the model  
mod3 <- lm(forM4, data = datReg6)  
summ(mod3)

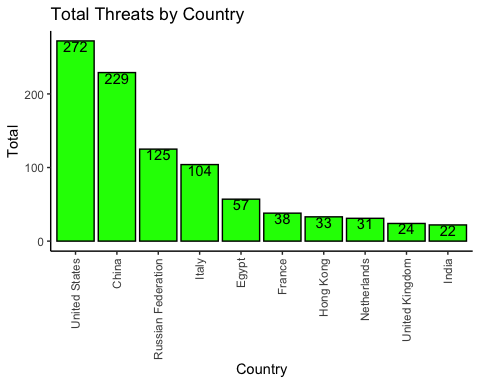
## MODEL INFO:  
## Observations: 582  
## Dependent Variable: Threat.Category  
## Type: OLS linear regression   
##   
## MODEL FIT:  
## F(1,580) = 0.12, p = 0.73  
## R² = 0.00  
## Adj. R² = -0.00   
##   
## Standard errors: OLS  
## ---------------------------------------------------  
## Est. S.E. t val. p  
## --------------------- ------ ------ -------- ------  
## (Intercept) 0.02 0.04 0.39 0.69  
## Source.Country1 0.01 0.04 0.34 0.73  
## ---------------------------------------------------

## Results

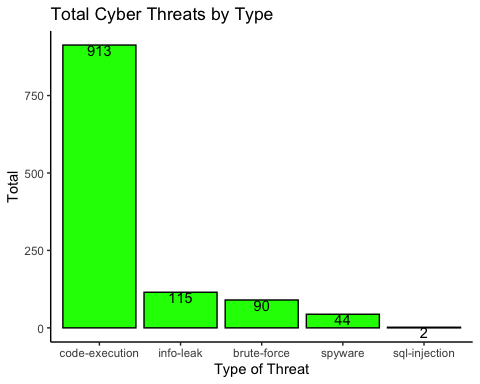
As shown in the tables above, the regression models don’t show much correlation, as the highest r-squared value was 0.02. Shown below are charts of most attacks by type and origin.

#### Plotting and Sorting Data

####Calculating frequencies  
tab <- table(attackInfo$Source.Country)  
####Sorting  
tab\_s <- sort(tab)  
####Extract Top 10 Countries  
top10 <- tail(names(tab\_s), 10)  
####Subsetting dataframe  
d\_s <- subset(attackInfo, Source.Country %in% top10)  
####Ordering factor by level   
d\_s$Source.Country <- factor(d\_s$Source.Country, levels = rev(top10))  
  
##Plotting  
ggplot(d\_s, aes(x = Source.Country, fill=I("green"), col=I("black"))) +  
 geom\_bar() +  
 theme\_classic() +  
 theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5)) +  
 ggtitle("Total Threats by Country") +  
 geom\_text(stat='count',aes(label=..count..),vjust=1) +  
 xlab("Country") +  
 ylab("Total")



####Calculating frequencies  
tab1 <- table(attackInfo$Threat.Category)  
####Sorting  
tab\_s1 <- sort(tab1)  
####Extracting Top 10 Threat Categories  
top5 <- tail(names(tab\_s1), 5)  
####Subsetting dataframe  
d\_s1 <- subset(attackInfo, Threat.Category %in% top5)  
####Ordering factor by level  
d\_s1$Threat.Category <- factor(d\_s1$Threat.Category, levels = rev(top5))  
##Plotting  
ggplot(d\_s1, aes(x = Threat.Category, fill=I("green"), col=I("black"))) +  
 geom\_bar() +  
 theme\_classic() +  
 ggtitle("Total Cyber Threats by Type") +  
 xlab("Type of Threat") +  
 geom\_text(stat='count',aes(label=..count..),vjust=1) +  
 ylab("Total")



## Implications

Cyber threats and attacks are a complex area of variables. While, I have been unsuccesful in finding a combination of variables to predict a threat type or source of a threat, I believe that a wider scope of systems and longer time frame could yield different results. This project was limited in scope by comparison, to a project spanning several systems with a less limited time frame.

## Conclusion

With a coefficient of determination; or R-squared value of less than 0.02 it is evident that the threat category is not a good predictor of the source country. The vice-verse is also true, as source country of a threat is not a good predictor of the type of threat. In conculstion, the location of a threats origin and the type of threat thrown at AgConnections do not correlate. The same is to be said when adding in an extra predictor of the day of week the threat was recieved. In my first two regression models I used the day of week recieved and source/category to predict the other. This yielded equally non-correlating results with the addition of the day of week receieved variable.

## References

Jahn Research Group (2019). Cyber Risk and Security Implications in Smart Agriculture and Food Systems ***University of Wisconsin-Madison College of Agriculture and Live Sciences***