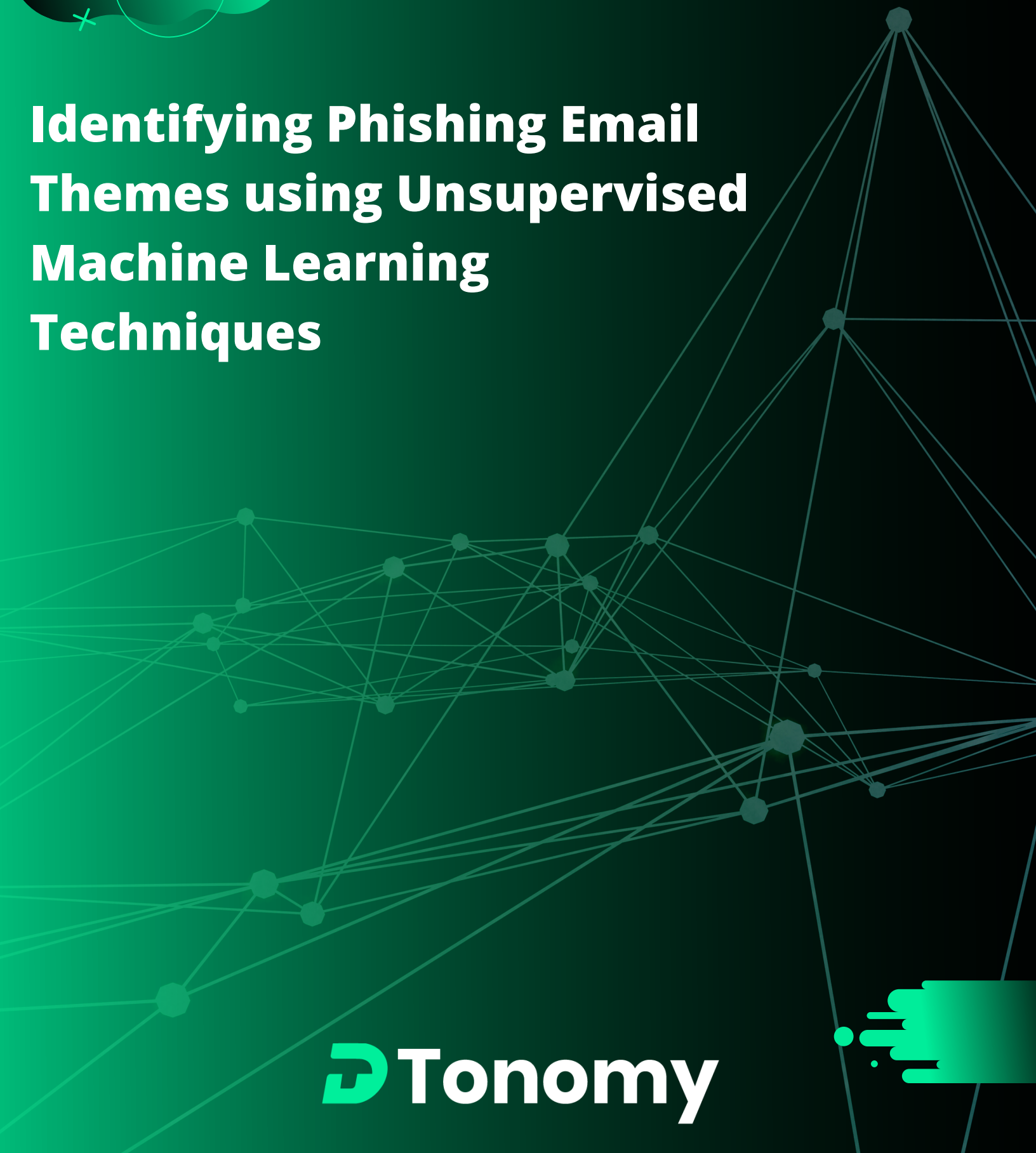
Identifying Phishing Email Themes using Unsupervised Machine Learning Techniques



Tonomy

Overview

Phishing emails are a type of online scam where criminals send an email that appears to be from a legitimate company and ask you to provide sensitive information. This is usually done by including a link that will appear to take you to the company's website to fill in your information – but the website is a clever fake and the information you provide goes straight to the person behind the scam.
In this research paper, unsupervised machine learning and the K-means Clustering technique is applied to identify the theme of phishing on a phishing-email dataset which is scraped from various sources spanning from 2010 till 2019.
Unsupervised machine learning uncovers previously unknown patterns in data and is best applied when there is no data on desired outcomes or for problems that the business has not seen before.
K-means clustering is a popular unsupervised machine learning algorithm that groups similar data points together based on certain similarities to discover underlying patterns.

This research paper will explain:·

- The purpose of the study·
- The analysis of the data set·
- The process of cleaning the data·
- Identification of a phishing theme using clustering·
- Our conclusions

Structure of the dataset

The dataset consists of 26 columns and around 9k rows. The topics include:

- 'ID',
- 'Email Body',
- 'Links_in_body',
- 'Return-Path',
- 'X-Original-To',
- 'Delivered-To',
- 'Received',
- 'Message-ID',
- 'From',
- 'Reply-To',
- 'To',
- 'Subject',
- 'Date',
- 'MIME-Version',
- 'Content-Type',
- 'X-Priority',
- 'X-CS-IP',
- 'Status',
- 'X-Status',
- 'X-Keywords',
- 'X-UID',
- 'month',
- 'year',
- 'day',
- 'body',
- 'subject'

Purpose of the Study

The goal of this study was to learn how phishing techniques and themes have changed over the years, and how phishing emails are distributed over year, month and day of the week.

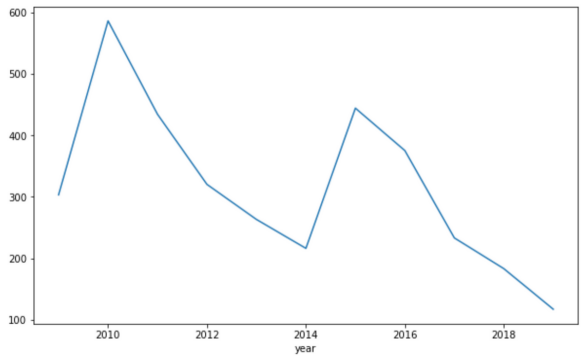
Analysis of the Dataset

First, we analyze the phishing email dataset to see the distribution of Phishing emails according to:

- Year
- Month
- Day of the Week

By Year

From this graph, we can see the distribution of phishing emails according to year. We can see that most of the phishing emails present in that dataset are from 2010, 2011 and 2015. We were not able to collect a lot of emails for the year 2018 and 2019 so going forward to make the dataset more balanced, data from 2018 and 2019 will be combined and represented as one.

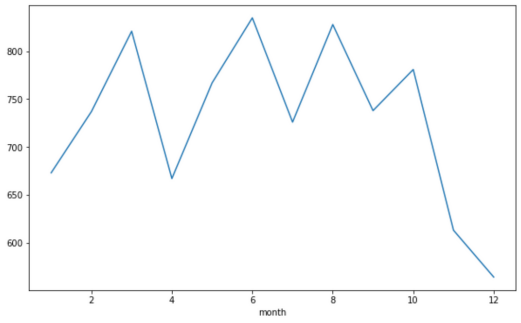


year
2009.0 303
2010.0 586
2011.0 434
2012.0 320
2013.0 263
2014.0 216
2015.0 444
2016.0 375
2017.0 233
2018.0 183
2019.0 117
Name: year,
dtype: int64

Based on our dataset, after the initial spike of phishing emails in 2010, there was a significant decrease, likely because security professionals had become familiar and started developing techniques to block suspect emails. As attackers developed new and unique methods of phishing, there was another increase, but again a decrease as security professionals become familiar with the techniques.

By Month

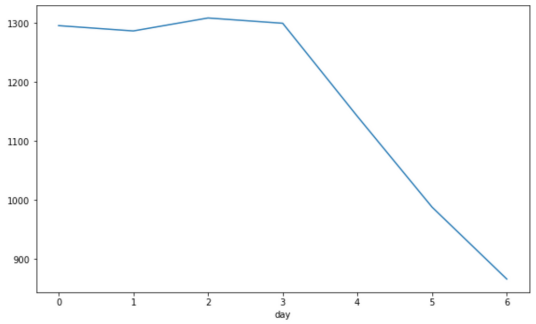
From the analysis we can see that phishing-email scam was very active during Feb, Mar, June and Aug but drastically reduced during November and December. Though the reason is unknown, we suspect this may be due to the holiday season between November and December.



month
1.0 673
2.0 737
3.0 821
4.0 667
5.0 767
6.0 835
7.0 726
8.0 828
9.0 738
10.0 781
11.0 613
12.0 564
Name: month, dtype: int64

By Day of the Week

From the analysis we can also see that the phishing emails are mostly active throughout the week but drastically reduced during weekends. This could be because attackers may suspect that people don't read email as often on the weekends and may be lost in the inbox as the week starts.



day
0.0 1296
1.0 1287
2.0 1309
3.0 1300
4.0 1142
5.0 988
6.0 866
Name: day,
dtype: int64

Data Cleaning

Data cleaning is one of the most important tasks that needs to be performed before applying a machine learning algorithm.

Data cleaning is also important because it improves data quality and in doing so, it increases overall productivity. When data is cleaned, all outdated or incorrect information is removed which leaves us with the highest quality information, assuring a better and more accurate output.

Since we are working primarily with text data, some of the tasks involved in cleaning are removing Stop words, converting the entire text to lowercase, and removing punctuations and special characters.

When cleaning the dataset, stop words need to be removed from the dataset because they don't help us to find the context or the true meaning of a sentence. These are words that can be removed without any negative consequences to the final model that we are training. We define custom Stop words by taking some of the repeated words specific to this dataset which doesn't make much difference to the model and adding those to nltk's Stop words.

Word Cloud

Word clouds act as an alternative for analyzing text data and it is also much faster than coding. Essentially, word clouds generators work by breaking the text down into component words and counting how frequently they appear in the body of the text.

We run the word cloud to quickly find some of the most common words in the dataset. We can see that lot of emails in our dataset contains "Paypal", "ebay" along with "congratulations" so we can assume that they might be gift card themed phishing attacks. We can also associate security related phishing attacks such as links to change your password as we can see frequent references to words such as user and security.

Clustering

Clustering is the process of grouping similar contents to determine the theme of the attacks. We apply clustering to group phishing-emails into categories and check if we can associate or draw conclusions to any type of phishing attacks from the words present in the clusters.

Prior to clustering, we start by applying the Elbow method to identify the number of clusters(K) and then perform K-means clustering on the dataset.

We have divided the dataset based on years to find the common theme in those particular years. In this research paper, the primary aim is to find the theme of phishing emails and check if there has been any change between the year 2010 and 2017 and later.

2010 Analyzed

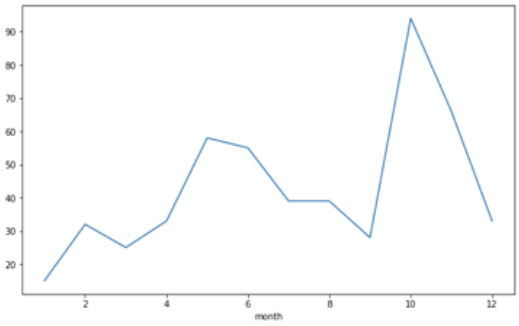
As mentioned above, we start by finding the optimal number of clusters using the Elbow method and then use this to perform clustering.

From the results of the Elbow method graph below we can see that there is a steep slope at 3 (on X-axis), so we consider the number of clusters to be 3 while applying K-means clustering.

Data Analysis on 2010 - According to Month

Given that there was a high number of phishing attacks in 2010, we explored the number of phishing emails that were reported in 2010 based on months. We can see that the highest number of phishing emails reported were in Oct and Nov. There was also a small spike in May and June as well.

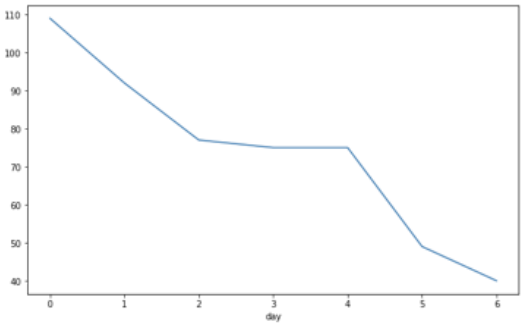
Interestingly, in reviewing the emails from October and Novembers, the subjects were mainly with respect to travel as the holiday season approached.




```
month
1.0    15
2.0    32
3.0    25
4.0    33
5.0    58
6.0    55
7.0    39
8.0    39
9.0    28
10.0   94
11.0   66
12.0   33
Name: month,
dtype: int64
```

Data Analysis on 2010 - According to Day of the week

We explore further to look at the number of phishing emails that were reported in 2010 based on day of the week. According to the data we collected, the attacks were more during the working days and significantly reduced during weekends.

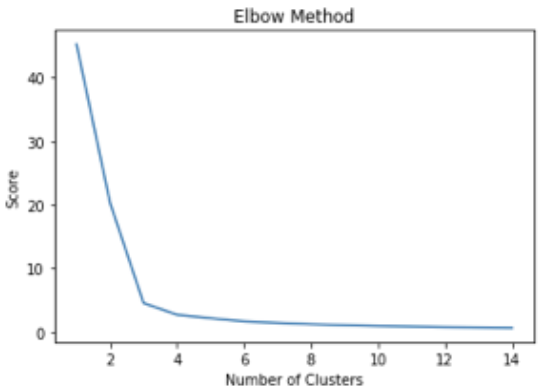



```
day
0.0    109
1.0    92
2.0    77
3.0    75
4.0    75
5.0    49
6.0    40
Name: day,
dtype: int64
```

Data Analysis on 2010 - Elbow Method

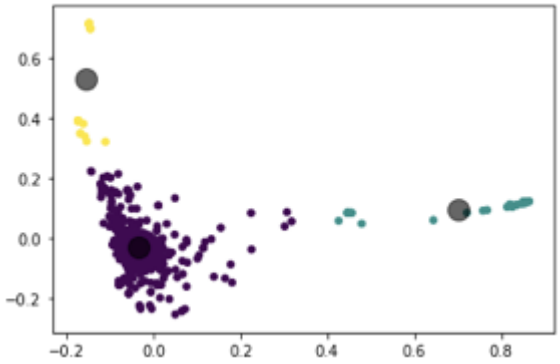
K-means clustering unsupervised machine learning algorithms group the data into a specified number of clusters, but it may not be the right number of clusters. The Elbow Method helps to determine if the person doing the analysis is using the right number of clusters. If the line chart looks like an arm, the “elbow” on the arm is the value that is the best.

This elbow method analysis was done to choose the right number of clusters. This analysis indicates that the current number of clusters to proceed is 3.



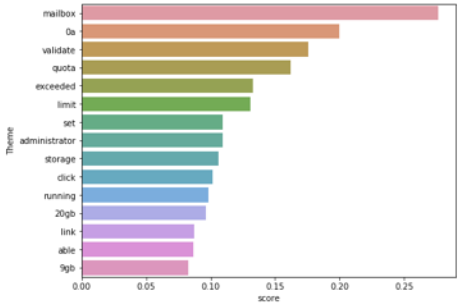
Data Analysis on 2010 - Plotting the clusters

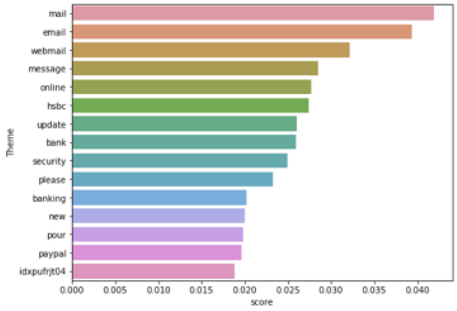
The graph to the right shows how the clusters are distributed with the black circle marking their centers.

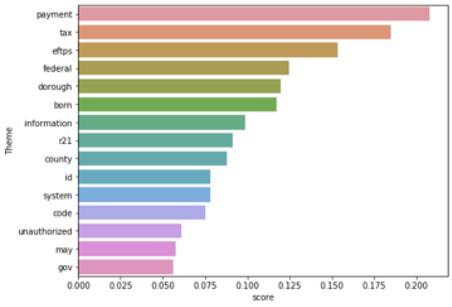


Below we plot the top words belonging to each cluster and from these words we try to determine the theme of the phishing attack.

From the plots below we can see that some of the phishing themes used in 2010 were related to email validation, tax, Paypal, payment related to tax or any items bought online, Government related like mails from Federal department, bank accounts etc.





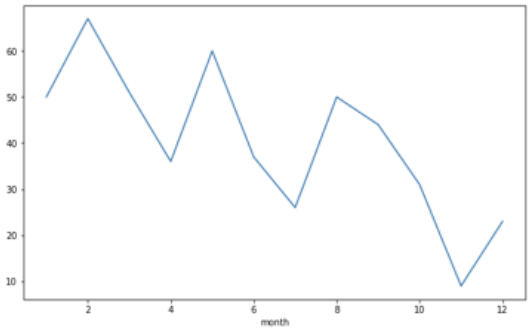


2017 and later

We follow the exact same procedure we followed for 2010 phishing emails I.e., we get a number of clusters by plotting the elbow graph and then applying clustering and plot the top words in each cluster to see if we can determine the theme of phishing attack

Data Analysis on 2017 and later - According to Month

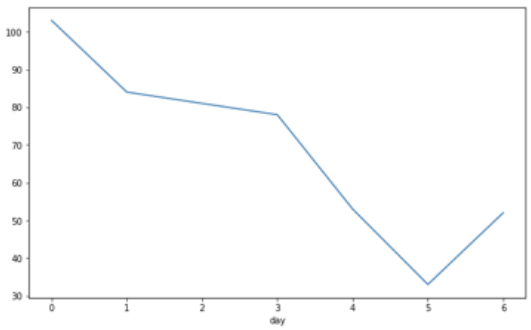
To the right is the number of phishing emails that were reported in 2017, 2018 and 2019 based on months. We can see that the highest number of phishing emails reported were in Feb, March, May and August. There was also a small spike in May and June as well.



month
1.0 50
2.0 67
3.0 51
4.0 36
5.0 60
6.0 37
7.0 26
8.0 50
9.0 44
10.0 31
11.0 9
12.0 23
Name: month, dtype: int64

Data Analysis on 2017 and later - According to Day of the Week

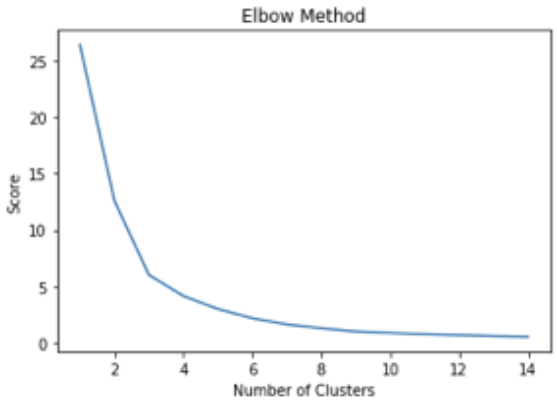
Below is the number of phishing emails that were reported in 2017, 2018 and 2019 based on day of the week. According to the data we collected, the attacks were consistent during the working days and significantly reduced on Saturday. One thing we can see is that the phishing attacks increased on Sunday which we had not seen in previous years.



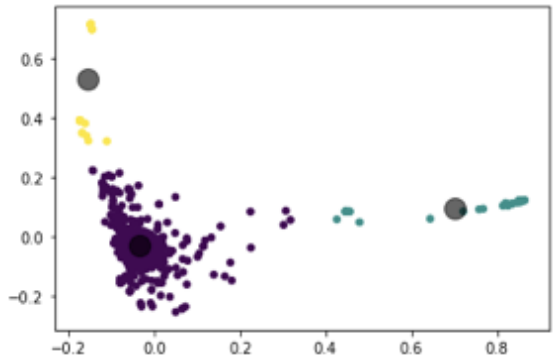
day
0.0 103
1.0 84
2.0 81
3.0 78
4.0 53
5.0 33
6.0 52
Name: day, dtype: int64

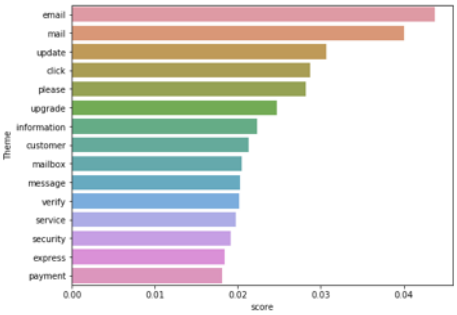
We noted that the phishing emails were reduced on weekends, suspecting that it is because people do not access their emails as often on weekends.

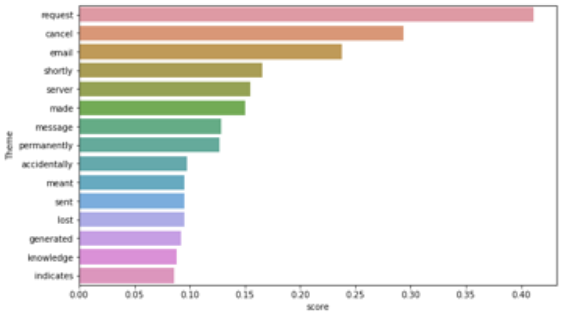
Data Analysis on 2017 and later - Elbow Method

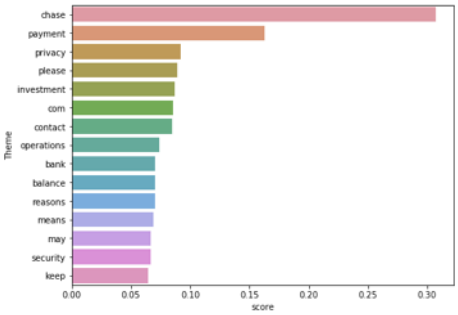


Data Analysis on 2017 and later- Plotting the clusters









From the above plots we can see that some of the phishing themes used in the year 2017 and later were related to banks like Chase, mails related to verifying an account, security related phishing attacks etc.

Conclusion

From the above analysis we could find the overall theme of phishing attacks by applying unsupervised machine learning techniques and also comparing the phishing attack themes between years. Another advantage of this method is that we could easily find the theme of the phishing attack without even reading the entire mail using clustering techniques. This method also helps in analyzing how the theme of attack has changed over years as we can see from the above plots, some of the common phishing attacks in 2010 such as eBay and PayPal have drastically reduced over years.

About DTonomy

DTonomy is hyper-focused on security orchestration, automation and response (SOAR). DTonomy solves a critical problem facing Security teams today: too many alerts that exceed the capacity of skilled professionals to investigate and resolve them. Organizations today have hundreds to thousands of daily alerts from hundreds of sources and these numbers will only continue to grow. Most organizations are short staffed which results in inconsistent investigation processes, high mean time to response, increased risk and analyst burnout. It is necessary for organizations to adopt a modern SOAR platform backed by artificial intelligence, machine-learning, and automated workflows.

DTonomy's AI Assisted Incident Response (AIR) platform manages alerts from multiple security tools and infrastructure and automates manual time-consuming and repetitive tasks. AIR is powered by DTonomy's adaptive learning engine which continuously learns and provides contextual insights that are not easily discoverable. The platform uses the insights to make relevant recommendations and automated workflows to guide security teams through steps and procedures. DTonomy's AIR platform resolves incidents up to 10 times quick which leads to decreased downtime and reduced alert fatigue for staff.

DTonomy