Bayesian Analysis Project Progress Report

Introduction

Phishing is a common form of scamming or social engineering that costs individuals and companies significant amounts of money each year. The FTC reports that billions of dollars are lost to scams every year, with $2.6 billion in losses from impersonator scams in 2022 alone (FTC). Phishing is an attempt to trick someone into believing that a fraudulent email or website is legitimate for the purpose of gaining access to data such as banking details, passwords, and other types of sensitive information. Phishing is initiated through a text or email message containing a link or attachment that the attacker wants the target to click. For example, the attacker claims that the user must take some type of action to protect their account or avoid fees. Once the target clicks the link, they may be taken to a fraudulent website mimicking a trusted brand such as Netflix or Amazon and asked for login credentials that the attacker will then steal and use for nefarious purposes.

Because of the frequency and severity of the consequences of phishing attacks, email providers employ filters to reduce the amount of potentially malicious emails one receives, although some still fall through the cracks. Email users are often aware of the risks of phishing and have some vague idea of what to look out for. I am not too familiar with the phishing landscape in general so I am interested in learning more to see what precautions are recommended to keep people safe online.

On a more personal note, I have been conducting qualitative research about vulnerable internet users' experiences online. Several interviewees have addressed experiences with phishing emails and their experiences vetting the trustworthiness of messages they receive. My

research studies more legal forms of online deception but I would like to take the opportunity to look at one of the more illegal forms, phishing, through this semester project.

Description of Data

To search for a dataset for this project, I looked at many, many potential topics, sources, and questions for analysis. I first wanted to come up with a topic that was meaningful or interesting to me and then find relevant data to analyze. I could not find much data related to the slightly niche topics I was interested in, so I decided to take the approach of first finding a dataset and then brainstorming an application. This worked out somewhat. I searched for trending datasets on Kaggle and came across the Dataset for Phishing Link Detection (Kaggle). I have been thinking about online scams a lot recently because of my independent research so this caught my attention.

The dataset is intended for use in phishing link detection, as in analyzing characteristics about the URL to predict whether the link is a phishing attempt or a link to a legitimate website. It contains 50% legitimate URLs and 50% phishing URLs as well as information about the composition of the URL and website information such as traffic and page rank.

To begin thinking of a more concrete application of this data, I did some preliminary Google searches for background information about phishing. I was curious about the estimated frequency that it occurs, so I first searched for that information. Several articles stated that an estimated 1.2% of all emails sent were phishing attempts, totaling to 3.4 billion phishing emails sent per day. I wanted to know how this number was determined so I looked for the original source of the statistic. After doing some digging, I found that the number came from a 2021 report from the Valimail anti-phishing company. They claim that in the general email landscape

an estimated one percent of all emails sent were phishing attempts (Valimail). Unfortunately, no further data about this statistic is publicly available.

In some of my further internet searches, I came across an article stating that an astounding 47.3% of all emails sent were phishing attempts (EarthWeb). After looking deeper into the article sources and related statistics, I found that the original claim is that about half of all emails sent are considered *spam*, which not only includes phishing emails but also other unwanted messages such as unsolicited marketing emails (EarthWeb). Nevertheless, the article does claim that about half of all emails sent are phishing attempts. This number in particular stuck in my mind because of how much higher it was than other estimates.

Analysis

To create an initial "word problem" for this project, I first thought about a hypothetical scenario about a company where approximately 50% of the incoming emails are phishing attempts, similar to the statistic given in the EarthWeb article. This company is trying to avoid phishing attempts by only opening emails and clicking URLs that are likely to be legitimate. They have collected a sample of 19,431 emails and categorized them as either phishing and legitimate emails.

status	n	percent
legitimate	9716	0.5000257
phishing	9715	0.4999743
Total	19431	1.0000000

These emails contain URL links and information about them that they would like to analyze to try to better predict whether future incoming emails are phishing attempts.

url_length <dbl>	hostname_length <dbl>	ip <dbl>	total_of. <dbl>	total_of- <dbl>	total_of@ <dbl>	total_of? <dbl>
46	20	0	3	0	0	1
128	120	0	10	0	0	0
52	25	0	3	0	0	0
21	13	0	2	0	0	0
28	19	0	2	0	0	0
128	50	1	4	1	0	1

6 rows | 3-9 of 87 columns

I first looked at URL length in relation to legitimacy status. The statistics show that phishing emails have a higher mean and median URL length as well as a larger range of the length.

status <chr>	mean_length <dbl>	median_length <dbl>	min_length <dbl>	max_length <dbl>
legitimate	47.40191	41	12	557
phishing	74.86855	55	15	1641

I created a probability model for assessing whether an incoming email containing a link is a phishing attempt based on the length of the URL. I somewhat arbitrarily chose a length of 100 characters as the threshold for being considered a "long URL."

```
          length_category     n   percent
Less than or equal to 100 17083 0.8791622
           More than 100  2348 0.1208378
```

If we look at the status of these long URLs, a large number of them are phishing links.

status <chr>	count <int>
legitimate	335
phishing	2013

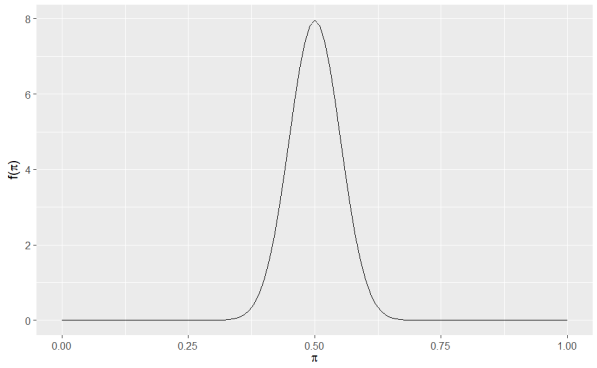
With this information, I started by asking the question: If we believe that 50% of all emails received are phishing attempts, what is the probability that a URL is a phishing given that its length is greater than 100?

	Phishing	Legitimate	total

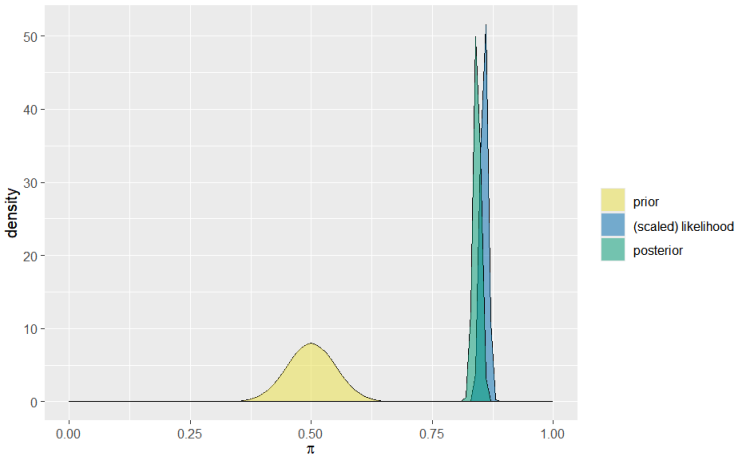
Prior probability	0.50	0.50	1
Likelihood	.2072	.0345	
Posterior probability	.8576	.1428	1

The chance that the >100 character URL is phishing is 0.8576.

Now instead of representing the prior probability as a single value 0.5, we can represent it as a range of plausible values. Let's say that the company believes that 0.5 of emails received are phishing attempts, but that the number could reasonably range from 0.4 to 0.6. Instead of asking the probability that a single email containing a long URL is a phishing attempt, we can ask what the expected proportion of incoming emails containing long URLs are phishing attempts. This can be represented by the Beta-Binomial model that can provide an estimate for a proportion $\pi \in [0,1]$. We first tune the hyperparameters α and β to suit the expected value and variability of the company's prior understanding of π. This gives us a reasonable prior probability model of Beta(50,50):



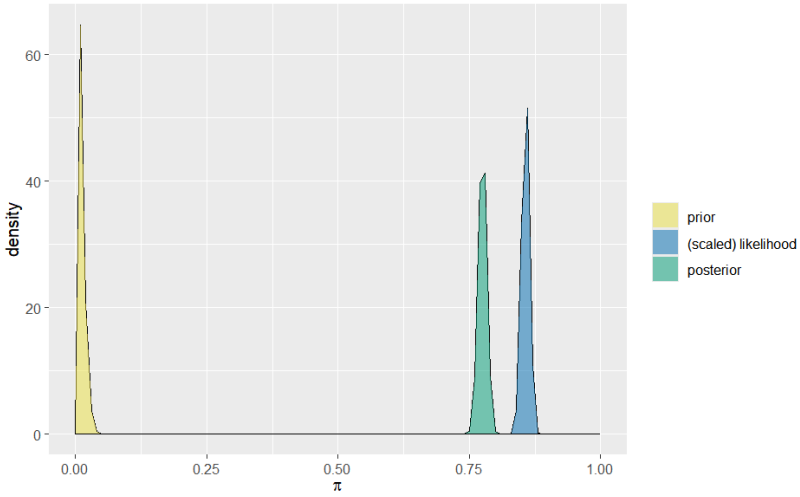
Using the Phishing Link dataset, we find that y=2013 of the n=2348 long URLs are phishing attempts. We use this to update our prior model Beta(α + y, β + n - y) = Beta(50 + 2013, 50 + 2348 - 2013) = Beta(2063,385).



model <chr>	alpha <dbl>	beta <dbl>	mean <dbl>	mode <dbl>	var <dbl>	sd <dbl>
prior	50	50	0.5000	0.500	0.0025	0.0498
posterior	2063	385	0.8427	0.843	0.0001	0.0074

The company can expect that about 84% of incoming emails with URLs longer than 100 characters are phishing attempts, compared to the prior understanding that about 50% of them would be.

This hypothetical scenario is just one example of how the dataset can be used to update our prior understanding of the email phishing landscape. We can change our prior to reflect these different understandings and use the data to come to different posterior probabilities. For instance, instead of the company expecting 50% of incoming emails to be phishing attempts, we can model for the expected 1.2% of all emails sent are phishing. If we are confident with this number, we can represent this prior with Beta(3,250). Keeping the threshold of long URLs at a length of 100, we can update our posterior given our data that 2013 of 2348 long URLs are phishing attempts.



model <chr>	alpha <dbl>	beta <dbl>	mean <dbl>	mode <dbl>	var <dbl>	sd <dbl>
prior	3	250	0.0119	0.0080	0e+00	0.0068
posterior	2016	585	0.7751	0.7753	1e-04	0.0082

Future Direction

There is a lot of potential to explore more interesting and real-world applications of this type of data. I'd like to ask more meaningful questions about phishing and how we can use data to make decisions to keep people safe. My application of this data is limited by my basic knowledge of statistical methods and Bayesian analysis. Hopefully as I progress through the course my learning will support more sophisticated analyses.

Additionally, as suggested by my peers, I may be able to explore different datasets or look at different characteristics of phishing URLs for example the number of real words or misspelled words. Websites such as openphish.com and phishtank.org provide open source data about reported phishing attacks that I would like to explore for further research.

Sources

EarthWeb. “How Many Phishing Emails Are Sent Daily in 2024? 11+ Statistics - EarthWeb.” Retrieved March 8, 2024 (https://earthweb.com/how-many-phishing-emails-are-sent-daily/).

EarthWeb. “How Many Spam Emails Are Sent Per Day in 2024? - EarthWeb.” Retrieved March 8, 2024 (https://earthweb.com/how-many-spam-emails-are-sent-per-day/).

Federal Trade Commission. “The Top Scams of 2022.” *Consumer Advice*. Retrieved March 8, 2024 (https://consumer.ftc.gov/consumer-alerts/2023/02/top-scams-2022).

Kaggle. “Dataset for Link Phishing Detection.” Retrieved March 8, 2024 (https://www.kaggle.com/datasets/winson13/dataset-for-link-phishing-detection).

Valimail. 2021. “Valimail Report Reveals 3 Billion Spoofed Emails Are Sent Every Day - Valimail.” Retrieved March 8, 2024 (https://www.valimail.com/newsroom/valimail-report-reveals-3-billion-spoofed-emails-are-sent-every-day/).

SML Project

2024-03-05

Data glimpse

```
head(df)
```

```
## # A tibble: 6 x 87
##    ...1 url   url_length hostname_length    ip total_of. `total_of-` `total_of@`
##   <dbl> <chr>      <dbl>           <dbl> <dbl>     <dbl>       <dbl>       <dbl>
## 1     0 http~         46              20     0         3           0           0
## 2     1 http~        128             120     0        10           0           0
## 3     2 http~         52              25     0         3           0           0
## 4     3 http~         21              13     0         2           0           0
## 5     4 http~         28              19     0         2           0           0
## 6     5 http~        128              50     1         4           1           0
## # i 79 more variables: `total_of?` <dbl>, `total_of&` <dbl>, `total_of=` <dbl>,
## #   total_of_ <dbl>, `total_of~` <dbl>, `total_of%` <dbl>, `total_of/` <dbl>,
## #   `total_of*` <dbl>, `total_of:` <dbl>, `total_of,` <dbl>, `total_of;` <dbl>,
## #   `total_of$` <dbl>, total_of_www <dbl>, total_of_com <dbl>,
## #   total_of_http_in_path <dbl>, https_token <dbl>, ratio_digits_url <dbl>,
## #   ratio_digits_host <dbl>, punycode <dbl>, port <dbl>, tld_in_path <dbl>,
## #   tld_in_subdomain <dbl>, abnormal_subdomain <dbl>, nb_subdomains <dbl>, ...
```

Summary stats

```
summary_stats <- df %>%
  group_by(status) %>%
  summarize(
    mean_length = mean(url_length),
    median_length = median(url_length),
    min_length = min(url_length),
    max_length = max(url_length)
  )
summary_stats
```

```
## # A tibble: 2 x 5
##   status     mean_length median_length min_length max_length
##   <chr>            <dbl>         <dbl>      <dbl>      <dbl>
## 1 legitimate        47.4            41         12        557
## 2 phishing          74.9            55         15       1641
```

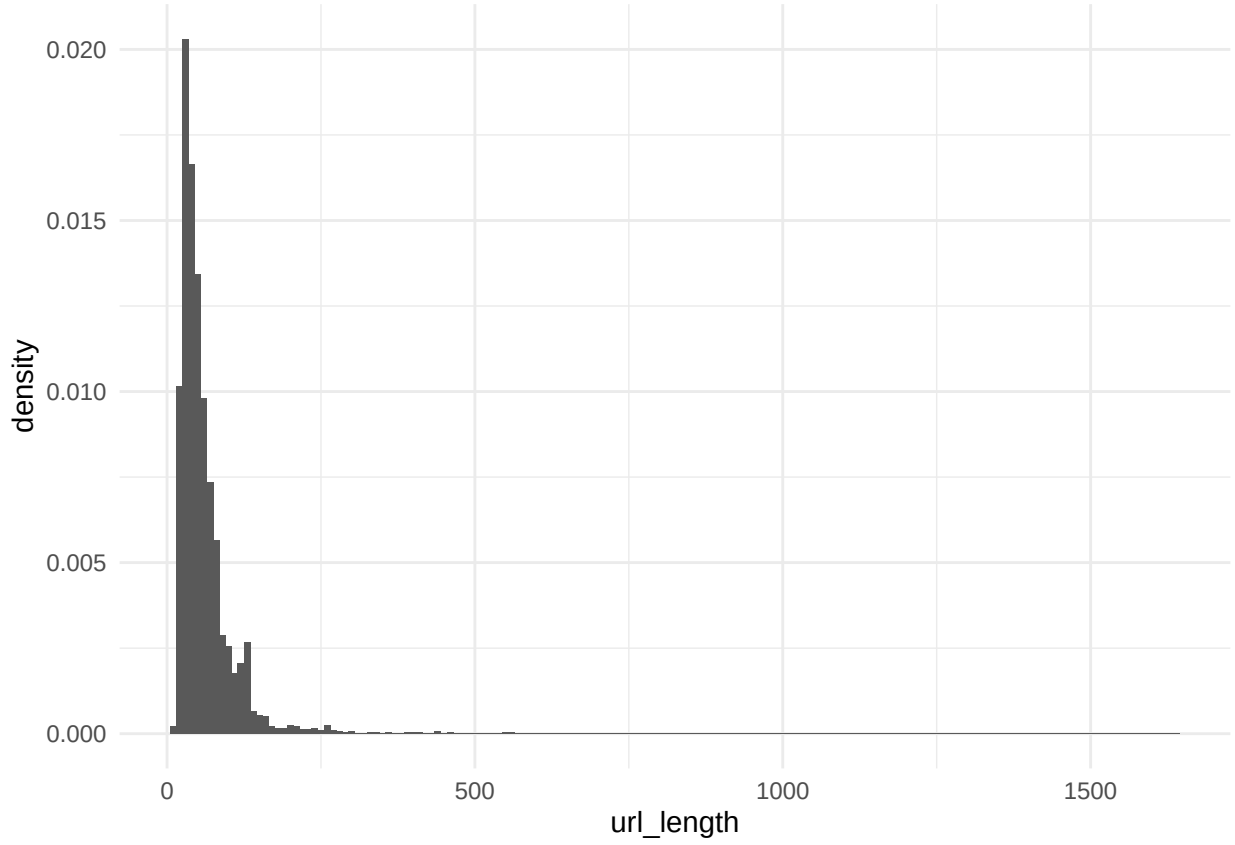
Phishing status distribution

```r
#data by phishing status
df %>%
  tabyl(status) %>%
  adorn_totals("row")
```

```
##      status     n   percent
##  legitimate  9716 0.5000257
##    phishing  9715 0.4999743
##       Total 19431 1.0000000
```

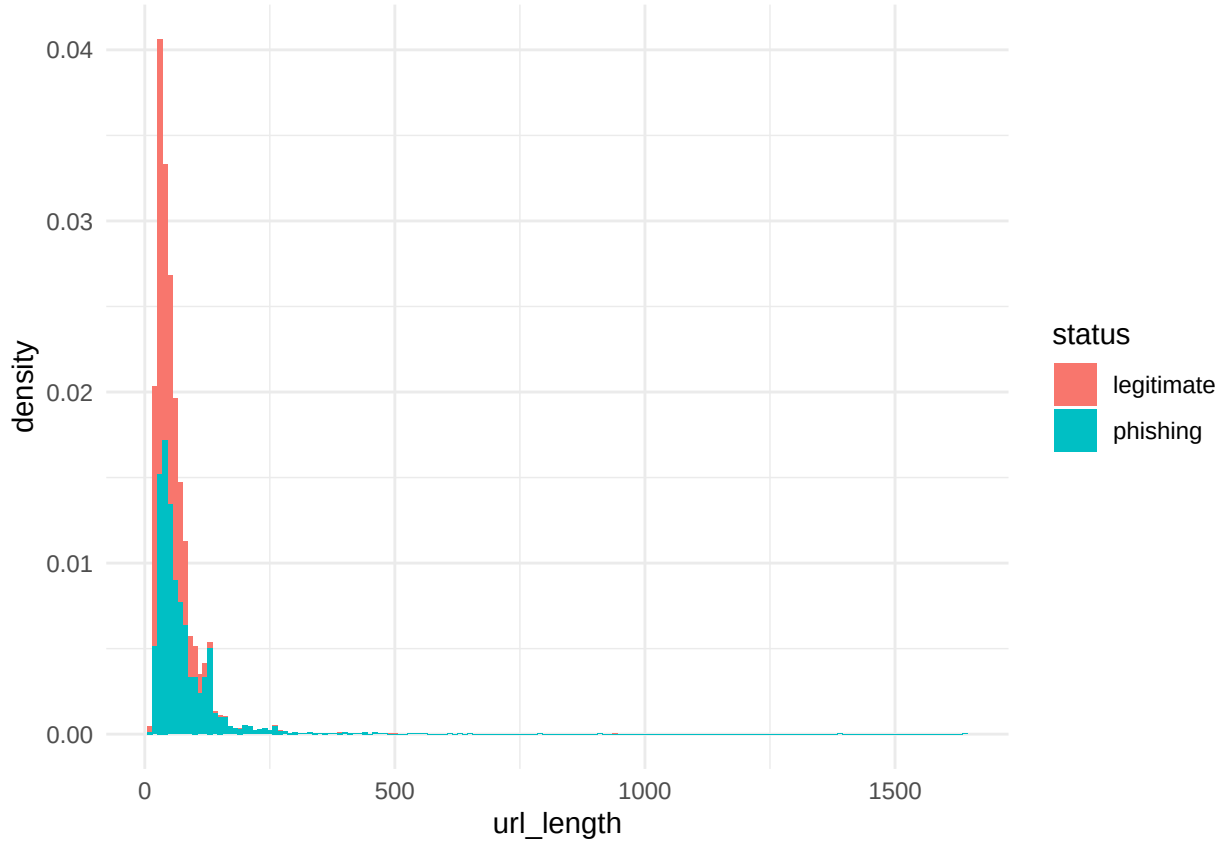
URL length distribution

```r
#histogram
ggplot(df, aes(x = url_length)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 10) +
  theme_minimal()
```

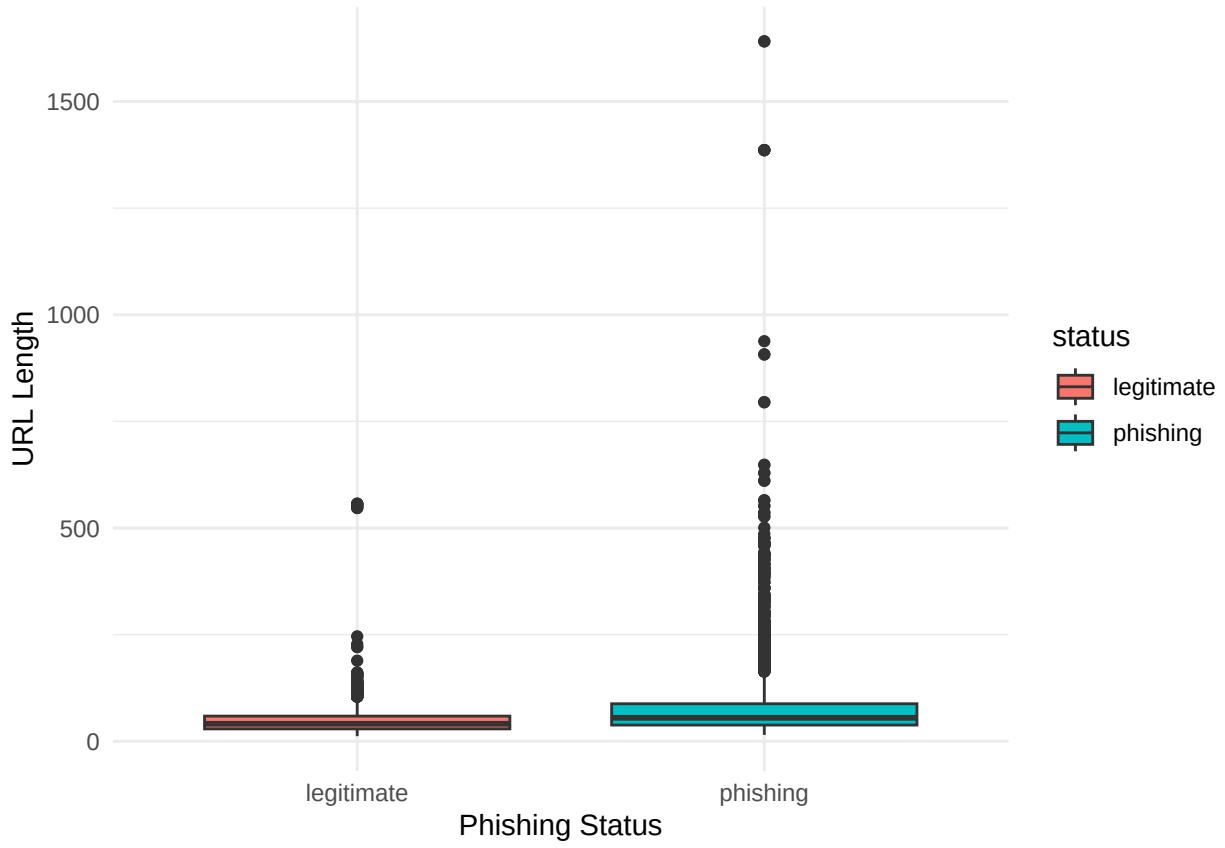


URL length by phishing status

```r
#histogram
ggplot(df, aes(x = url_length,fill=status)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 10) +
  theme_minimal()
```




```r
#boxplot
ggplot(df, aes(x = status, y = url_length, fill = status)) +
  geom_boxplot() +
  labs(x = "Phishing Status", y = "URL Length") +
  theme_minimal()
```



Long URLs and phishing status

```
#What length is considered a long URL?
long_length <- 100

df<-df %>%
  mutate(long_url = ifelse(url_length > long_length, TRUE, FALSE))

#long URLs in dataset
df %>%
  tabyl(long_url)
```

```
##  long_url     n   percent
##     FALSE 17083 0.8791622
##      TRUE  2348 0.1208378
```

```
#long URLs by phishing status
df %>%
  tabyl(long_url, status) %>%
  adorn_totals(c("row","col"))
```

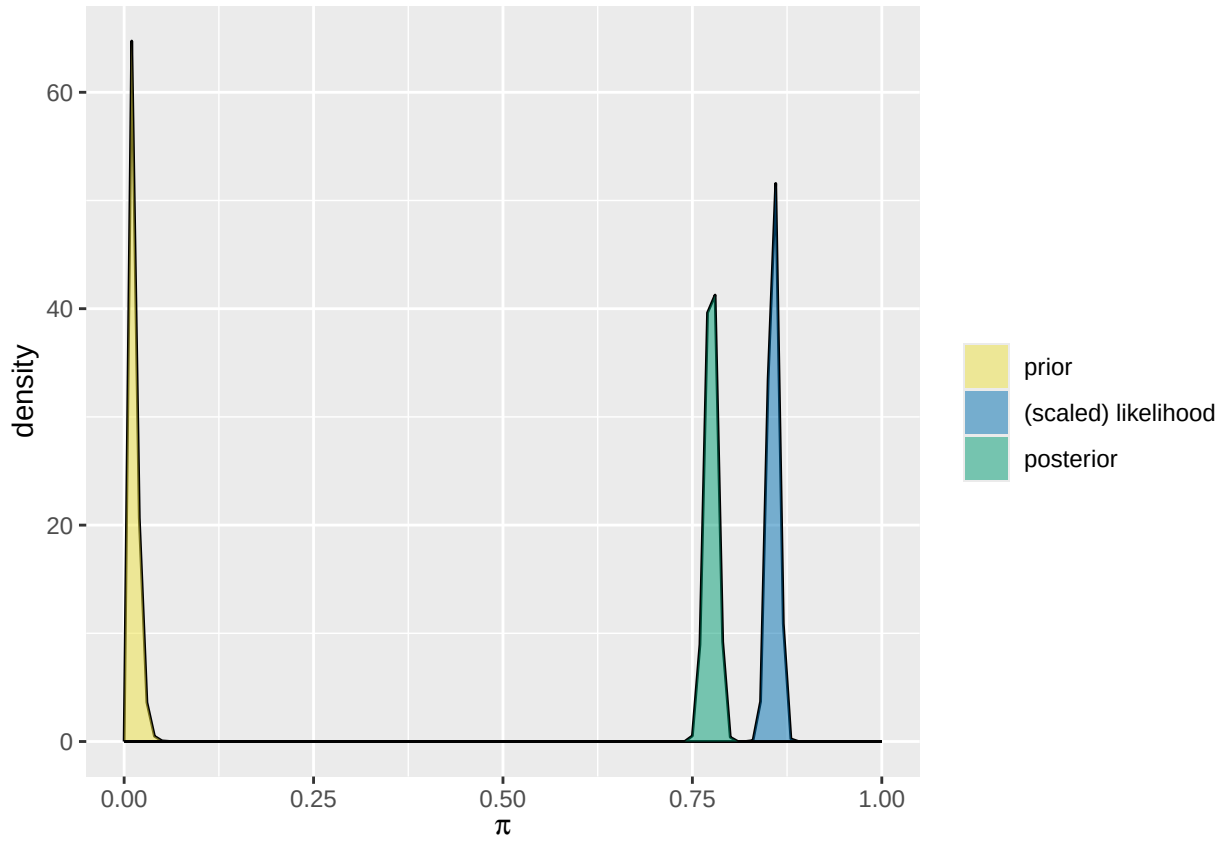
```
##  long_url legitimate phishing Total
##     FALSE       9381     7702 17083
##      TRUE        335     2013  2348
##     Total       9716     9715 19431
```

Beta-Binomial model

```
#tune beta prior
alpha <- 3
beta <- 250

y <- sum(df$long_url & df$status == "phishing")
n <- sum(df$long_url)

#posterior
plot_beta_binomial(alpha,beta,y,n)
```




```
summarize_beta_binomial(alpha,beta,y,n) |>
   mutate_if(is.numeric,round,digits=4)
```

```
##       model alpha beta   mean   mode   var     sd
## 1     prior     3  250 0.0119 0.0080 0e+00 0.0068
## 2 posterior  2016  585 0.7751 0.7753 1e-04 0.0082
```