# 

**BC2407 – ANALYTICS II: ADVANCED PREDICTIVE TECHNIQUES**

**GROUP PROJECT REPORT**

**An Analytics-Based Approach to Reducing Phishing Scams on Carousell**

**Seminar Class 5, Team 4 (Thursday - 2.30PM)**

**Prepared for: Prof. Hyeokkoo Eric Kwon**

**Prepared By:**

|  |  |
| --- | --- |
| **Name** | **Matriculation Number** |
| Lim Shao Zhe Lenson | U2121230A |
| Phoebe Cheong Qian Ming | U2121506K |
| Syed Saad Muntasirul Islam | U2121273K |
| Tai Jia An | U2110916L |
| Tay Wei Ming | U2110998K |

# 

[**Executive Summary 4**](#_filr3shsjd99)

[**Introduction & Background 5**](#_kzugtq9j82ka)

[Significance of problem 5](#_rxkow5pzeqg)

[The Business Problem 5](#_f7wjvjl46uyw)

[Current Solutions and Limitations 6](#_jl18frhm1eyl)

[Sift 6](#_92lr9bpabcid)

[Limitations of Sift 6](#_h6sr1mhhtn5p)

[Carousell Protection 6](#_uzztlyl5w4so)

[Limitations of Carousell Protection 6](#_gc5qxcd2wuno)

[**Objective 7**](#_4ncb38es5876)

[Project Objective 7](#_ntwfrcbc1c8)

[Business Outcome 7](#_l1b89rizptr4)

[**Proposed Solution 8**](#_1hjif2cjx5oh)

[Overview 8](#_2pd4uc52ysbx)

[Implementation 9](#_ymspeutt09my)

[Data engineering 9](#_3e2tp1zb2ep0)

[Analytics Pipeline 9](#_qvc7sco8dah3)

[Auto user suspension and warning 10](#_m9nbm9li8bn9)

[Admin Dashboard 11](#_v8pqrkiv7i37)

[Models Required 13](#_fl0nz478dp70)

[Malicious URL Detection Model (First Layer) 13](#_fkb8jhgp7rmp)

[Input variables 13](#_1u41ylqo3y9w)

[Output variable 14](#_d7txhnd53rx5)

[Malicious Website Content Detection Model (Second Layer) 14](#_o9be9lb5f8yz)

[Input Variables 14](#_wcbmw7foolml)

[Output variable 14](#_5i269g2z1itx)

[**Dataset and Model Development 14**](#_521bbooxnqf2)

[Data Cleaning and Selection 14](#_selb2suc118k)

[Exploratory Data Analysis 15](#_4805qkoo7uru)

[Data balancing 16](#_39dv85ugydsc)

[Models 17](#_wj5r4lshovai)

[Malicious URL Detection 17](#_pf4akwt6nthx)

[Model Selection 18](#_urdjvehvu5f4)

[Tuning 18](#_aszux9qax2fk)

[Malicious Website Content Detection 19](#_uuurod18palr)

[Model Selection 19](#_1c9ukf7i6p03)

[**Evaluation 20**](#_8w6sipl35cpi)

[Models 20](#_gi6xue7wdknv)

[Key Insights 20](#_xgnbzjbub616)

[Feasibility 20](#_owgnakk71m0c)

[Limitations 21](#_3ml0p9csaruc)

[**Recommendation 22**](#_mj0gz34zo0s)

[Adoption of Proposed Solution 22](#_ytvcqw2gqfzu)

[URL Blacklist 22](#_ln4tgg9aknf6)

[Partnership with other E-commerce Platforms 23](#_voih2mtegqv5)

[Improving Public and Investor Relations 23](#_d539v8v5gdxf)

[Post Implementation Review 23](#_b6ssbxf4ao6h)

[**Conclusion 24**](#_4cssi8wkcbmh)

[**References 25**](#_vqdc5q2wuqnt)

[**Appendix A - Data Dictionary 27**](#_mev1k1hj4alm)

[**Appendix B - EDA 29**](#_n65zr6eewvu7)

[**Appendix C - Dashboard 36**](#_wnlz94e1kjul)

[**Appendix D - Models for Malicious URLs 38**](#_e1gjk04lamvx)

[**Appendix E - Models for Malicious Websites 49**](#_1fve3jvd14mc)

[**Appendix F - Finding Optimal Threshold For Neural Network 61**](#_2zzqy8oiuq7m)

# 

# Executive Summary

With tens of millions of users as of 2020, Carousell is one of the fastest growing e-commerce platforms in Southeast Asia. Despite the stiff competition in the e-commerce market, Carousell has managed to carve out its own niche by being more accessible to individual sellers as compared to the other platforms. However, as a consequence of this accessibility, Carousell has also become one of most scam prone platforms, accounting for 38% of all e-commerce scam occurrences in 2022.

Unsurprisingly, this has hampered its growth and is a significant source of its lost revenues, while also being a black spot on its reputation that could drive away users as well as investors; As Carousell is focused on profitability, and one of its primary business goals is to execute its exit strategy in 2024, clearly this issue is a significant detriment to them.

To address this problem, Carousell has two current solutions, Sift and Carousell Protection. However, the limitations of these solutions make them ineffective in detecting and preventing these scams. Hence, through this report, we aim to develop a robust analytics solution that is able to automatically detect and prevent scammers from sharing malicious URLs and QR codes and take appropriate action against offenders. This will be complemented by an admin dashboard that would provide valuable data and key insights to Carousell staff to allow them to further develop new initiatives and strategies to combat this issue.

By adopting our solutions, Carousell would be taking a strong stance against scams on its platform, acting as a catalyst for future growth, and positioning itself as a leader in the e-commerce industry in Southeast Asia for user experience and safety.

For Carousell to achieve further success in reducing phishing rates, additional recommendations were made. In parallel with the implementation of our proposed solution, Carousell should incorporate a black list of URLS, engage in partnership with other e-commerce platforms and invest in improving public relations. Incorporating these recommendations with the proposed solution can maximise Carousell’s ability to reduce phishing rates and to ultimately achieve its business goals.

# Introduction & Background

## Significance of problem

Scams involving e-commerce in Singapore have become more prevalent, with the number of cases doubling in a year from 1,057 cases in 2021 to 2,267 cases in 2022 (Chua, N, 2022). For four years in a row, e-commerce scams have been ranked as the top 5 scam categories in Singapore (Dazeinfo, 2023). Carousell in particular, has one of the lowest anti-scam scores rated by the government's E-Commerce Marketplace Transaction Safety Ratings (Mahmud, A. H., 2022), with nearly two out of every five (38%) e-commerce scams in 2022 involving Carousell (Hirschmann, R., 2023). In December 2022 alone, Carousell had more than 1,115 phishing scam victims, causing a loss of about $1 million and personal information involved (Lim, J., 2023). This calls for greater attention to Carousell’s current operations and interactions between buyers and sellers to improve the situation.

Phishing scams occurring on Carousell’s platform includes but not limited to, one party sending out phishing QR codes, or website links through the app to lure the other party to fraudulent websites with the purpose of obtaining sensitive personal data such as credit card and bank details.

## The Business Problem

In 2020, Carousell set a business goal to execute its exit strategy within 4 years in 2024

which includes an initial public offering (IPO), trade sale, secondary sale or share buyback

(Pillai, S., 2020). In order to do so, Carousell set its focus on profitability from 2020-2024

in order to secure its position in the other 8 markets it operates in (Choudhury, S. R., 2020).

However, the rapid rise in scams on its platform has been affecting Carousell's reputation among investors and customers and one of its largest revenue streams - commissions through transactions on their platform (Gilchrist, K., 2020). Phishing scams prevents transactions from executing through its platform, causing Carousell to lose potential commissions from these transactions. Furthermore, victims’ trust in the platform would diminish, leading to a decline in customer base and potential revenue. If left unaddressed, this could further worsen Carousell’s image which would cause a further reduction in its user base and its revenue stream, which is also detrimental to investors’ outlook for Carousell and its position in the markets where it operates in.

Therefore, to further Carousell’s goal in executing a successful exit strategy, preventing unnecessary loss of commissions, preserving Carousell’s image as a secure and safe platform and maintaining investors confidence is key. Hence, there is a need to reduce the rising number and severity of e-commerce scams occurring on Carousell’s platform.

## Current Solutions and Limitations

### Sift

Carousell currently uses Sift, an automation and machine learning technology, to actively detect fraudulent listings and sellers. It assigns a risk score according to the behavioural characteristics of sellers and automatically bars sellers who exceed a specified risk score. According to Tan, S. L (n.d.), Sift had proven effective by detecting 10% and 23% more fraudulent listings and sellers, respectively, and saving more than $350,000 per year while achieving a return on investment (ROI) of 2 times in 2019.

#### Limitations of Sift

While Sift helps Carousell detect fraudulent listings and sellers to protect buyers, sellers and buyers are not protected when it comes to phishing scams. Sift was trained to identify suspicious seller behaviour via their listings and as phishing is conducted via chat instead of listings, these activities essentially by-passes Sift checks. Furthermore, as bad actors are now sending phishing links as pretend buyers, Sift’s targeted nature towards sellers creates a larger gap for bad actors to thrive. Sift is clearly lacking in detecting shifts in e-commerce scams and is also evident in the increasing number and growing severity of e-commerce scams occurring on Carousell’s platform since 2019.

### Carousell Protection

Furthermore, Carousell offers Carousell Protection, a payment solution that protects both the sellers and buyers by serving as an escrow and holding a buyer's payment until the expected order is delivered as described and the transaction is deemed completed by both parties. However, many buyers do not favour Carousell Protection as a fee of 4.5% is charged to the buyers for using this payment solution (Carousell, n.d.). As a result, sellers are open to accepting other payment methods outside of the platform. However, this has led to scammers exploiting the situation by impersonating buyers and sharing fraudulent phishing links and QR codes with sellers and pretending to be facilitating payment (Channel News Asia, 2022). As a result, personal credit cards, banking information and money have fallen into the wrong hands. This problem is not exclusive to sellers as sellers may also scam buyers using the same method. Therefore, both the sellers and buyers are vulnerable to such phishing scams.

#### 

#### Limitations of Carousell Protection

Due to the high transaction fees, Carousell Protection is not enticing to buyers. Although a solution is to reduce these high transaction fees, these fees cannot be lowered further by Carousell as doing so may lead to operational losses. Currently, these fees act as a reserve for Carousell to cover most of the costs associated with incidents of fraud, misrepresentation, and lost items (Carousell, n.d.). Furthermore, their partnership with Stripe to handle transactions incur fees per transaction (Simonson, 2023). Hence, these fees ensure the sustainability of this feature. By reducing or eliminating such fees, Carousell would not be able maintain the service to effectively manage the risks associated with transactions.

# Objective

With a thorough understanding of the challenges Carousell is facing, we have identified a critical need for a solution that can effectively block malicious links or QR codes, while still allowing safe ones to pass through. This solution should seamlessly integrate into Carousell's existing operations without causing any disruption or negatively impacting user experience. Additionally, the solution should be cost-effective to align with Carousell's overall business goals.

To achieve these objectives, we have developed specific goals for our solution and identified the intended business outcomes it should deliver. By clearly defining these objectives and outcomes, we can ensure that our solution effectively addresses Carousell's needs and provides measurable value to the business.

## Project Objective

The objective of this project is to build an analytics solution to actively and accurately detect fraudulent links that are shared among buyers and sellers on the platform and present this information effectively to Carousell staff. Hence, we have the following objectives,

* An analytical framework that can detect malicious URLs or QR codes accurately.
* Highlight detected malicious URL or QR code effectively to users.
* Prevent repeat offenders from abusing the platform.
* Ensure minimal disruption to the current user flow and experience and be seamlessly integrated into Carousell’s platform.
* Create a user-friendly dashboard that presents our predicted information in a concise and actionable format for Carousell staff.

## Business Outcome

With the ability to detect and highlight fraudulent links to users, it will actively prevent users from accessing these phishing links which could lead to sensitive personal data, banking information and money getting stolen by scammers. Therefore, the likelihood of users who fall victim to malicious links would be reduced. Furthermore, being well integrated into Carousell’s current user flow ensures their current user base would not be affected which if not considered might have adverse impact on their current user base.

Furthermore, with the dashboard, it will remove any prerequisite analytical knowledge for Carousell’s staff to utilise results from our solution. This will allow Carousell’s staff to capitalise and utilise information produced by our solution and help staff to make data driven decisions in addressing phishing occurrences on their platform.

Ultimately, this solution aims to reduce the phishing rates on their platform, improve the overall image and reputation of Carousell and rebuild the trust that customers and investors have in Carousell, leading to greater profitability and improve Carousell’s market position.

# Proposed Solution

## Overview

To address the issue of malicious links and QR codes being sent through Carousell chat and achieve the objectives stated above, we propose implementing an analytics-based solution. This solution involves an analytics pipeline that screens every link and QR code sent through the chat, flag any potentially malicious URLs or QR codes and utilise the predicted information with two solutions - an auto user suspension and warning system and an admin dashboard.

|  |
| --- |
| Fig 1 - Solution overview |
|  |

The auto user suspension and warning system utilised predicted information from the analytics model and if a message is flagged, a warning appears to the receiver and sender. If a user repeatedly triggers these warnings, their account would be suspended.

Furthermore, the admin dashboard will present our predicted information in a concise and actionable format. This dashboard will be customised for Carousell's staff, allowing them to easily access and interpret data from our model to make informed business decisions.

## Implementation

### Data engineering

Upon receiving an url, the necessary features are first extracted and fed to our analytics pipeline. If the input is a QR code, the QR code will be converted to its URL, and it will be parsed and handled like a usual URL.

### Analytics Pipeline

Utilising a labelled dataset of features that is extracted from URLs, we would train and optimise a machine learning classifier that would be able to accurately classify a URL as malicious or not.

Due to the context of analysing malicious links, the consequence of false negatives (When a URL is flagged as legitimate when it is actually malicious) are much higher than false positives (When a legitimate URL is flagged as malicious) as in a false negative, an unsuspecting user may fall for a scam. Therefore, we would further tune our model to be more sensitive to phishing links to keep the number of false negatives to a minimum.

Although this extra cautiousness would lead to fewer false negatives and ensure more peoples safety, it would also cause more legitimate links to be falsely classified as malicious - resulting in more false warnings that could reduce the trust between two parties and damage a sellers reputation. Consequently, this would also increase the workload for Carousell as more users would be falsely suspended and would appeal to remove the suspension.

To counter this issue, there would be a second layer of checks if a URL is initially flagged as suspicious by our classifier model. In this case, the website would be visited and a second model would be used to analyse the content of the website itself to classify whether the website is legitimate or malicious. This added layer of checks would greatly increase the confidence of our predictions and ensure the most appropriate steps are taken for each case.

This 2-step method would create 3 scenarios:

|  |  |  |
| --- | --- | --- |
| Table 1- Possible scenarios | | |
| URL classification | Website classification | Actions taken |
| Legitimate | - | No actions taken |
| Malicious | Legitimate | No actions taken |
| Malicious | Malicious | Malicious URL warning |

Using these models, our proposed two layered approach will be able to categorise any links or QR codes sent within the application as malicious or not with high certainty and accuracy.

|  |
| --- |
| Fig 2 - 2 layer method |
|  |

### Auto user suspension and warning

To ensure that our model can be integrated into Carousell’s existing system with minimal changes, we identified the most suitable points in the user flow to implement our solution and that is through Carousell’s chat function. This point in the user flow has the highest frequency and probability where a malicious link is sent. Therefore, a solution at this point to warn or alert users would ensure that predictive information from our model can be utilised effectively to prevent users from falling victim to malicious URLs. As such our first solution to utilise our predicted information is the “Auto user suspension and warning” system.

When a URL or QR code is flagged as malicious, the flagged message would be hidden from the receiver and a warning will appear (Fig.3), indicating that the sender may be sending potentially malicious messages. If the receiver has a special reason to trust the sender or the link, they can still choose to bypass the warning and access the message to ensure that users still have full control over their experience at Carousell.

|  |  |  |
| --- | --- | --- |
| Fig 3 - Malicious URL warnings | | |
| Receiver warning | Sender warning | Sender suspension notification |
|  |  |  |

The sender will also receive a warning (Fig.3) that the items they are sending are potentially malicious and that they should refrain from such behaviour. The number of times a user has been flagged to have sent a malicious link would also be tracked along with the user details. If a user triggers such warnings more than 3 times, their accounts will be automatically suspended (Fig.3) and their posts will be hidden, as that user is a potential scam account. The user can also choose to appeal to remove the suspension if they believe that it was an error.

### Admin Dashboard

The admin dashboard (Appendix C) will provide a comprehensive overview of potential phishing activity on the Carousell platform by integrating our predicted information with their existing user data. This integration will include a counter into Carousell’s existing user information which indicates the frequency of warnings triggered by flagged messages, as well as the cause of each flagged URL or QR code. Utilising this integrated information, the dashboard will present the information into three main sections:

1. The first section will display an overview of phishing cases that occur on Carousell’s platform. This includes an overall percentage of phishing cases occurring on their platform, a trend of phishing case count over different periods, countries, categories and lists of suspended users. This overview would be useful in analysing the effects of any measures to reduce phishing on their platform.
2. The second section gives detailed information on the demographic of affected users (Users who received phishing URLs) (Fig.4). Asides the demographic trends, the country of origin, category of listing and an overall trend of phishing cases by demographic will also be displayed. This information will allow staff to gather insights from the different age groups and even categories of listings bad actors tend to target. Using this information, staff would be able to craft targeted warnings to different categories or age groups and even tune the sensitivity of our model towards malicious URL being sent to chats based on certain categories of a listing or age groups of users.
3. The third section will display an overview of suspended users displaying the user's geographical location, category they target, links they sent and even their transaction count. This information will allow staff to identify regions or states where scammers originate from and introduce targeted measures such as increasing the friction of account creation in those areas. These measures will make it difficult for scammers to create accounts en masse while maintaining a convenient user experience for legitimate users. Furthermore, knowing the transaction count of flagged users will allow staff to identify high value users efficiently (users that have contributed to Carousell with large numbers of transactions) and handle cases like this to prevent losing valuable users.

|  |
| --- |
| Fig 4 - Admin dashboard |
|  |

## Models Required

The two layered approach of the analytics pipeline requires a different model at each layer, each expecting different variable inputs. The first layer requires a model that is able to classify if a given url is malicious, analysing only the link itself while the second layer contains the model that would analyse the full website content of a given link.

For each of the 2 layers, 3 models will be tested - MARS, Random Forest, and Neural Network, and the most appropriate model will then be selected.

### Malicious URL Detection Model (First Layer)

#### Input variables

The input for our first layer includes 11 variables which are components that make up a URL:

* having\_IPhaving\_IP\_Address
* URLURL\_Length
* Shortining\_Service
* having\_At\_Symbol
* Double\_slash\_redirecting
* Prefix\_Suffix
* having\_Sub\_Domain
* SSLfinal\_State
* Domain\_registeration\_length
* Port
* HTTPS\_token

#### 

#### Output variable

After providing our model with a URL, the model should output a binary variable “Result” where a value of1 represents that the url provided is a phishing URL.

### Malicious Website Content Detection Model (Second Layer)

#### Input Variables

The following website content components were highlighted to be relevant in determining if a website content is malicious or not:

* Favicon
* Request\_URL
* URL\_of\_Anchor
* Links\_in\_tags
* SFH
* Submitting\_to\_email
* Abnormal\_URL
* Redirect
* On\_mouseover
* RightClick
* popUpWidnow
* Iframe
* Age\_of\_domain
* DNSRecord
* Web\_traffic
* Page\_Rank
* Google\_Index
* Links\_pointing\_to\_page
* Statistical\_report

#### 

#### Output variable

After providing our model with contents of a website, the model should output a binary variable “Result” where a value of 1 represents that the website content provided is a phishing website.

# Dataset and Model Development

“Phishing Website Dataset” from Kaggle was used as it contains the necessary data inputs required by our models to build both the URL and website detection models. It contains 11055 samples (11055 rows, 32 columns); 11 were used for malicious URL classification and the other 19 were used for malicious website content classification. The variable “Result” has been set as our Y variable to state whether the sample was legitimate or phishing.

## Data Cleaning and Selection

To ensure high quality inputs and reliable results, our team conducted data cleaning before going into data exploration and model building and found no empty or invalid values. We further re-leveled our variable Y, “Result” from “-1” as phishing, and “1” as legitimate to “0” as phishing and “1” as legitimate for the Y variable. The data types of each of the columns were also converted from integer to factor.

The original dataset was then split into 2 datasets using the respective input variables for each layer. The full data dictionary can be found in the appendix (Appendix A).

## Exploratory Data Analysis

Exploratory data analysis was conducted to visualise important relationships and patterns that can assist us in building our model.

Since all of the columns are categorical, the Cramer’s V values for each variable were calculated. Cramer’s V gives the association between two variables and has an output between 0 and 1; A higher value means the two variables are more highly correlated with each other (Appendix B). Based on the Cramer’s V values, there is weak association between most of the variables and the Result variable, except for “URL\_of\_Anchor” and “web\_traffic” for website content, and “having\_sub\_domain” and “SSLfinal\_state” for URL data

|  |  |
| --- | --- |
| Fig 5 - Notable URL features | |
|  |  |

From the histograms, we could conclude that URLs with subdomains seem to involve more phishing activities. It is interesting to see URLs with single subdomains having higher phishing activities than URLs with multiple subdomains.

The legitimacy of a URL is dependent on whether a trusted issuer exists and if HTTPS is included. The presence of HTTPS in a URL has been shown to reduce the number of phishing URLs, while having a trusted issuer increases the number of legitimate URLs.

|  |  |
| --- | --- |
| Fig 6 - Notable website content features | |
|  |  |

From the figure, we can see that having more than 67% of anchor URLs within a webpage greatly increases the chances that a website is phishing, and that the majority of legitimate websites are ranked above 100,000 in terms of traffic.

## Data balancing

The original dataset had a total of 11055 samples, where 6157 samples represent legitimate and the remaining 4898 samples represent phishing records. In order to ensure a balanced dataset for model training, we decided to down sample the train dataset. It is important to balance a dataset because it helps to reduce the machine learning models’ bias towards the majority class and improves performance on the minority class.

In this case, instead of upsampling the data, we chose to down sample the majority class to balance the data as upsampling the data might lead to overfitting due to the introduction of duplicate data. We also understood the consequences of information loss due to down sampling. However, bearing in mind that we have a large dataset, dropping some of the data will still allow us to have sufficient data for accurate model predictions. In this scenario, rather than risking issues from upsampling, down sampling would be a better option.

After we conduct downsampling on the train data set, the variable Y now contains an equal amount of data points for both phishing and legitimate records, allowing us to move on to model building.

## Models

For each of the 2-layer models, we conducted a train-test split of the respective datasets in the ratio 70:30. The Accuracy, False Positive Rate (FPR), False Negative Rate (FNR), Precision, and Recall of balanced test datasets were recorded and compared.

### Malicious URL Detection

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 2 - Malicious URL model performance | | | | | |
| Model | Accuracy | False Positive Rate | False Negative Rate | Precision | Recall |
| MARS | 0.8875 | 0.08259 | 0.1159 | 0.8920 | 0.8489 |
| Random Forest | 0.9011 | 0.06659 | 0.1396 | 0.9113 | 0.8604 |
| Neural Network | 0.9017 | 0.07456 | 0.09895 | 0.9010 | 0.8741 |

#### 

|  |  |
| --- | --- |
| Fig 7- Statistically Significant URL Variables | |
| Random Forest Model | Neural Network Model |
|  |  |

From Fig. 7, we see that SSLfinal\_state, prefix\_suffix, and having\_sub\_domain are some of the most important features in determining if a URL is malicious.

#### Model Selection

Based on the comparison metrics table (Table 2), it shows that Neural Network has a higher accuracy (90.17%) and recall (87.41%) among the other 2 models. Even though the Neural Network model has a slightly higher false positive rate (7.47%) and lower precision (90.10%) than Random Forest model, having a lower false negative rate (9.90%) and higher accuracy is the main priority when detecting malicious URLs. With that said, the Neural Network model produces better performance in all other metrics as compared to the other 2 models. Therefore, Neural Network would be the optimal model selection for the detection of malicious URLs.

#### Tuning

As mentioned in our proposed solution, the false negative rate (FNR) must be minimised. To reduce the FNR, we adjusted the threshold for the Neural Network model which changes the minimum value required to classify an output as phishing or legitimate. We found an optimal threshold value by testing different threshold values for the Neural Network model and found that the threshold value of 0.1 gives one of the lowest FNR while having the highest accuracy for this cutoff value (Appendix F). Thus, a threshold of 0.1 will be used to build the final Neural Network model and the following shows the resulting metrics of our final model,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 3 - Malicious URL Neural Network Model performance (Optimal Threshold) | | | | | |
| Model | Accuracy | False Positive Rate | False Negative Rate | Precision | Recall |
| Neural Network | 0.8450 | 0.01708 | 0.03775 | 0.9622 | 0.6767 |

### Malicious Website Content Detection

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 4 - Malicious Website Content model performance | | | | | |
| Model | Accuracy | False Positive Rate | False Negative Rate | Precision | Recall |
| MARS | 0.8706 | 0.04129 | 0.2393 | 0.9355 | 0.7604 |
| Random Forest | 0.9180 | 0.05739 | 0.1130 | 0.9248 | 0.8870 |
| Neural Network | 0.9041 | 0.08545 | 0.1077 | 0.8923 | 0.8911 |

|  |  |
| --- | --- |
| Fig 8 - Statistically Significant Website Content Variables | |
| Random Forest Model | Neural Network Model |
|  |  |

From Fig. 8, we see that URL\_of\_Anchor is consistently ranked as the most statistically important variable in predicting whether a website is malicious.

#### Model Selection

Upon comparison (Table 4), Random Forest was selected as the best model choice for detecting malicious websites. Random Forest has the highest overall accuracy (91.80%) among the 3 models while still having a relatively high precision (92.48%) and recall (88.70%). In comparison with the MARS model, it has a much lower false negative rate (11.30%), which is crucial when it comes to detecting malicious websites. Even though Neural Network models have a slightly lower FNR (10.77%), it is noteworthy that Random Forest outperforms Neural Network models in other metrics, such as lower FPR (5.74%), thus Random Forest model is being claimed as the best model for malicious website detection.

The model was not further tuned to reduce its FNR as for the second layer, the overall accuracy is deemed to be the most important factor, as this final layer is the main determinant of what action should be taken. Therefore, it cannot be too sensitive towards false negatives, as it would lead to too many unfair suspensions nor can it be too sensitive towards false positives, as it would let too many malicious URLs pass through without warnings.

# Evaluation

## Models

From the above results, we see that both chosen models are viable in achieving their respective goals. The decreased cut-off for the Neural Network classifier to classify a URL as malicious ensures that there is a low probability of false negatives with a FNR of only 3.8%. The added checks by the second layer further ensures that there are few false positives overall as the second model is highly accurate and has a FPR of only 5.7%. The two models combined would ensure that the vast majority of truly malicious URLs are blocked and all users have a smooth experience with minimal disruptions.

## Key Insights

From the EDA of the URL data, we see that of all the features, SSLfinal\_state has the most significant relationship with whether a URL is malicious. SSLfinal\_state is a technical feature that indicates whether the website is secured and has the trusted certificates used by legitimate websites. This is an important finding and presents an added opportunity for Carousell to quickly eliminate a large number of malicious websites from being shared among users by simply preventing any website that does not meet this requirement from being sent. This is a quick and easy way for Carousell to reduce the total number of checks they have to conduct and the overhead involved in the process. This also ensures that innocent users that may have unknowingly shared a malicious URL due to their ignorance are notified early and are reminded to check the authenticity of websites they share.

## Feasibility

The performance of our models prove the feasibility of the proposed solutions, as the auto user suspension and warning system would be able to receive the output from the models and take the appropriate actions. The automated system’s response would be rapid and ensure that all parties receive the correct warnings and are aware of potential attempts at phishing. Scammers would also very quickly have their accounts suspended and be prevented from causing any further harm.

Deploying machine learning and analytical solutions have never been easier with the emergence of cloud providers. This will require an upfront investment from Carousell in order to purchase the cloud infrastructure needed to run the models at the required scale. However, after this initial investment, the models can be scaled up easily as multiple instances can be deployed concurrently ensuring that there is sufficient bandwidth for Carousell to screen all of the URLs and QR codes sent by users.

Carousell already has existing customer service and content moderation teams that would be able to pick up the new tasks required such as managing user appeals. Although some new staff may need to be hired to keep up with the increased overhead, this should be minimal due to the automated nature of the solution and the efficiency gains from the dashboard.

The software will also need to be updated to enable the auto suspension system and connect with the analytics models, however, this is not an issue given Carousell’s experienced software team.

Carousell has also already pledged to allocate more of its budget towards reducing phishing on its platform, ensuring that there is sufficient funding for such expansions to occur (Lim, 2022). The additional funds along with the abundance of data that Carousell already has plus what it will be able to collect over time will allow it to implement this solution effectively.

## Limitations

One of the limitations of this solution is that there is a trade-off in the accuracy of our Neural Network model and its false negative rate, and we chose to reduce the false negative rate in favour of increasing the accuracy. Ideally, we would be able to achieve a high accuracy with the lowest possible false negative rate. As our solution was developed using publicly available data on phishing websites, it may be imperfect; If Carousell were to collect its own data and further train and tune the model, this limitation could be mitigated, and the models could achieve a higher accuracy as well as a low false negative rate.

Another limitation is that with only 2 layers of checks, it is still possible that some number of malicious URLs may pass through both layers and be sent to unsuspecting users. One way to ensure the highest safety would be to use more models per layer, however, this method is expensive as it requires more computational power to train the models, as well as to deploy them. It will also increase the lag time for users that send URLs or QR codes, as it would have to pass through more checks before it gets sent. Carousell can also supplement our technical solution with a people-centric solution aimed at increasing their users’ scam detecting capabilities, to ensure that they are kept safe even from those few malicious links that do pass through.

A possible issue Carousell may encounter is that implementing this solution requires analysis of user’s chats to be able to detect links and QR codes. This may pose a privacy concern for users, and many may be uncomfortable with this, however, if Carousell is transparent of its processes and strictly abides by the Personal Data Protection Act (PDPA), users would have little reason to be concerned.

# Recommendation

## Analytical Solution

### Adoption of Proposed Solution

The analytics pipeline forms the basis of our solutions and from our model development, we have successfully developed a model for each layer of our analytics pipeline, achieving a reliable analytics model that minimises both false negatives and positives.

The prediction made by our analytics model can be effectively employed to reduce occurrences of scams in Carousell by limiting the types of URLs or QR codes that can be sent. Through this process, users would be warned automatically to be extra cautious when dealing with potentially malicious links with the added benefit that they would still have the freedom to bypass these warnings and have full control over their experience at Carousell. Senders of such malicious links would also receive warnings that such behaviour will not be tolerated and repeat offenders will be automatically suspended. We recognise that there are cases where improper suspensions may occur, so suspended users also have the option of appealing to Carousell.

The automated nature of this solution means that it can easily be deployed at scale without much need for Carousell to greatly increase its workforce. Furthermore, combined with the admin dashboard that would provide valuable information in real-time, Carousell would be able to take rapid action based on scam trends and create new initiatives and awareness campaigns to allow their users to navigate their platform more confidently and safely.

By tackling phishing rates effectively, Carousell’s investors will also benefit as less investments have to be poured into using inefficient methods to tackle this issue and compensating users who fall victims to these scams. This will allow investments to be directed to more profitable projects that have net positive values and increase Carousell’s cash flow. This in turn will result in better return of investment for investors. This will improve Carousell’s outlook as compared to the past several years which further increases investors’ confidence and willingness to invest in them which will be highly beneficial especially when Carousell executes their exit strategy.

### Post Implementation Review

Once this solution has been fully implemented, we recommend that Carousell revisit the model at fixed intervals, such as every quarter, to review its performance and improve the models. This is important as it is impossible to successfully identify and block every single phishing attempt sent by scammers, and although our models have been tuned to be able to detect the large majority of them, inevitable, there will still be some few cases that slip through the cracks. By conducting frequent reviews, Carousell would be able to identify what is unique about those cases that are not being caught by the models and can retrain the models to be able to detect such cases; over several reviews, the solution would be much more highly refined and there would be fewer cracks. This would also enable Carousell to identify key trends in e-commerce scams in the industry, allowing them to create their own separate strategies, and share this valuable information with their partners in the industry.

## Non-Analytical Solution

### URL Blacklist

Furthermore, from the key insight derived from this paper, Carousell can enact a simple step that would automatically eliminate a significant number of scams occurring on the platform by preventing any links that do not meet the proper SSL final state requirements from being shared. Not only would this reduce the number of scams, but it would also reduce the number of URLs that need to be processed using the models, reducing the overhead and increasing the cost efficiency of the system. This can be further extended to future insights gathered from our dashboard, by analysing trends in types of malicious links sent by users, these links can be added to a blacklist, allowing for Carousell to immediately block such malicious URL without the need for the model to process them.

### Partnership with other E-commerce Platforms

Carousell could also enter strategic partnerships with other e-commerce platforms due to the prevalence of phishing cases across all e-commerce platforms. This would not only allow access to more phishing related data to train more reliable and accurate models, but also encourage knowledge exchange between the different companies, allowing knowledge like different approaches and measures taken by different companies to be shared. Overall, this will not only reduce phishing incidence on Carousell’s platform, but the whole industry itself, creating a safe environment for all users to utilise these platforms confidently.

### Improving Public Relations

As Carousell aims to reduce phishing rates on its platform, its effort and success needs to be continuously communicated to users and investors, allowing Carousell users to be aware that Carousell has been putting a lot of effort in caring and protecting their users. The persistent problem of phishing occurring on its platform in the past several years resulted in damage to its brand and reputation. Hence, by communicating its anti-phishing efforts through multi communication channels like social media, press releases and marketing strategies, Carousell can rebuild its brand image and establish itself as a platform that takes user security seriously. This will also help to increase users and investors' confidence in the platform, which is crucial in achieving Carousell’s goal of executing a successful exit strategy.

# Conclusion

In this report, we have identified key steps Carousell can take to reduce the number of phishing scams that occur on its platform, as well as ways for them to continuously improve on these steps to achieve the best possible results. We have also identified key insights derived from our EDA that have allowed us to expand on our original solution and provide a simple and cost-effective way for Carousell to quickly reduce the number of current scam links shared that would also improve the efficiency of the proposed solution.

By implementing this solution, Carousell users would be able to safely use the platform and have the peace of mind in knowing that there is a lower chance that they are dealing with a bad actor on the platform. This would also increase their trust in the platform, and the likelihood of them continuing to use the platform. The reduction of scams and boost in its reputation would help Carousell to attract more users and improve its investment outlook; this would allow them to generate more revenue through its usual business, and make it more attractive among investors.

The admin dashboard would also allow the Carousell staff to derive their own key actionable insights and create new initiatives and strategies to further reduce e-commerce scams on their platform.

Aside from implementing our original solution, Carousell could further implement a blacklist to increase the efficiency of detecting malicious URLs, partner with other e-commerce platforms to tackle the issue of phishing industry-wide, and capitalise on its new efforts to improve public relations. With these recommendations, Carousell would thereby cross a major hurdle in their goal to achieve their main business goal - achieving profitability and executing its exit strategy.

# References

Carousell. (n.d.). [Singapore] What is the platform fee? Retrieved April 1, 2023, from https://support.carousell.com/hc/en-us/articles/360000970147--Singapore-What-is-the-platform-fee-

Channel News Asia. (2022, May 24). *Police warn of scammers impersonating Carousell buyers, sending victims phishing links*. CNA. https://www.channelnewsasia.com/singapore/phishing-scam-carousell-payment-caroupay-singapore-police-force-2668031

Choudhury, S. R. (2020, October 7). *Southeast Asia’s classifieds start-up Carousell aims for profitability in 4 years, potential exit thereafter*. CNBC. https://www.cnbc.com/2020/10/07/carousell-aims-for-profitability-in-3-to-4-years-potential-exit-after.html

Chua, N. (2022, December 27). *Young people biggest group hauled up for e-commerce scams*. The Straits Times. https://www.straitstimes.com/singapore/young-people-biggest-group-hauled-up-for-e-commerce-scams

Dazeinfo. (2023, February 23). *Singapore saw a record number of scams and cybercrime cases in 2022: Lost over half a billion dollars [REPORT]*. https://dazeinfo.com/2023/02/23/singapore-saw-a-record-number-of-scams-and-cybercrime-cases-in-2022-lost-over-half-a-billion-dollars-report/

Gilchrist, K. (2020, March 12). *How 3 friends turned their unwanted items into a $550 million business*. CNBC. https://www.cnbc.com/2019/10/08/carousell-how-3-friends-turned-used-goods-into-a-550-million-startup.html

Hirshmann, R. (2023, February 23). *E-commerce platforms scam: E-commerce platforms with the largest share of scams reported in Singapore in 2022*. Statista. https://www.statista.com/statistics/1107023/singapore-leading-platforms-used-e-commerce-scams/

Lim, J. (2023, January 3). *Carousell, banks take steps to tackle spike in phishing scams*. The Straits Times. https://www.straitstimes.com/singapore/courts-crime/carousell-working-with-police-to-tackle-phishing-scams-banks-to-disrupt-scammers

Mahmud, A. H. (2022, May 14). *Facebook Marketplace, Carousell get lowest anti-scam scores in new government e-commerce rating system*. CNA. https://www.channelnewsasia.com/singapore/carousell-facebook-e-commerce-scam-rating-2684741

Pillai, S. (2020, September 25). *Carousell aims for exit within 4 years, possibly at value of US$1.1b or more*. TheBusinesstimes. <https://www.businesstimes.com.sg/startups-tech/startups/carousell-aims-exit-within-4-years-possibly-value-us11b-or-more>

Simonson, J. (2023, January 2). *Stripe Pricing And Fees (2023 Guide)*. Forbes Advisor. https://www.forbes.com/advisor/business/services/stripe-pricing-fees/#:~:text=Stripe%27s

Tan, S. L. (n.d.). *How Carousell keeps fraudulent listings off of their platform.* Sift + Carousell. https://resources.sift.com/wp-content/uploads/2019/09/Sift\_carousell\_casestudy-1.pdf

Lim, J. (2022, May 17). *'It's an arms race': Carousell spends five-digit sum per month, hires more staff to fight scams*. The Straits Times. https://www.straitstimes.com/singapore/courts-crime/its-an-arms-race-carousell-spends-five-digit-sum-per-month-hires-more-staff-to-fight-scams

# 

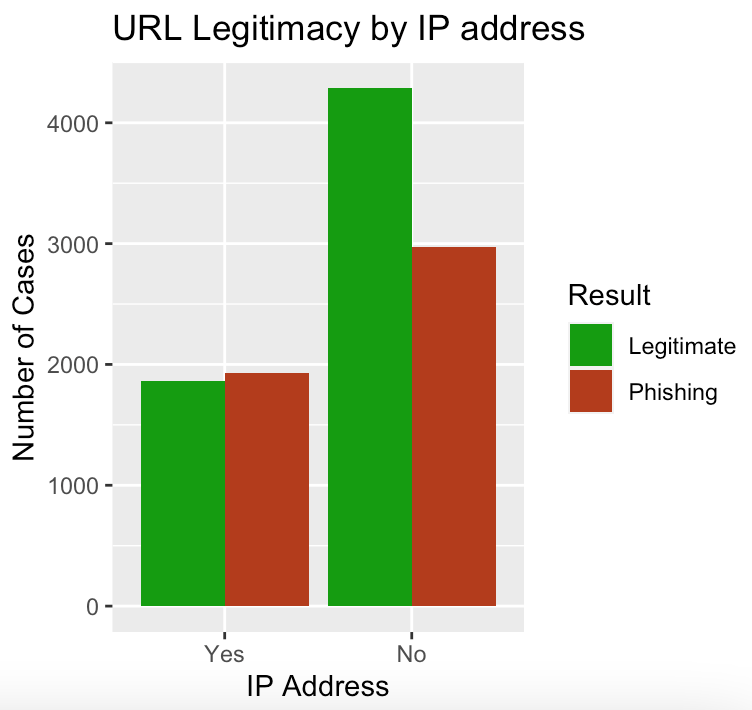
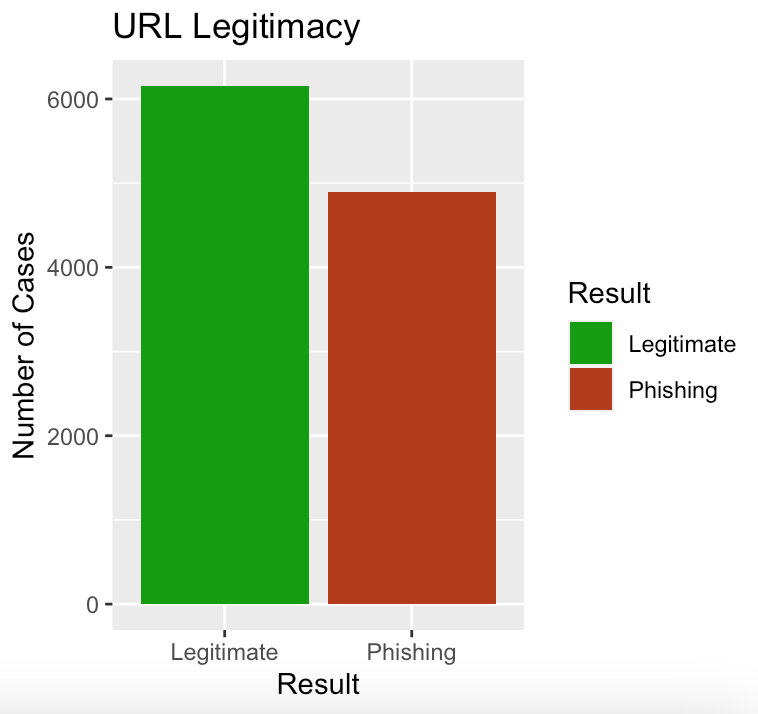
# Appendix A - Data Dictionary

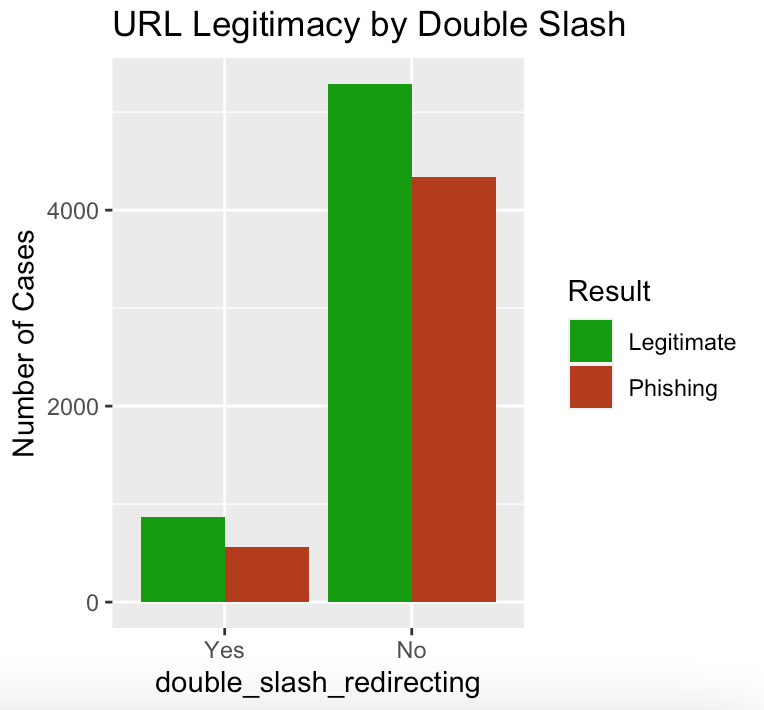
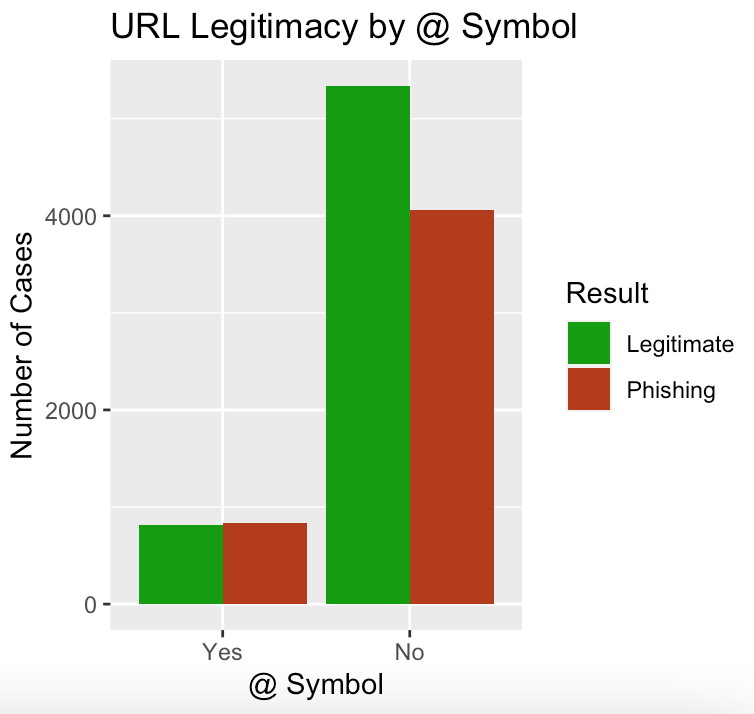
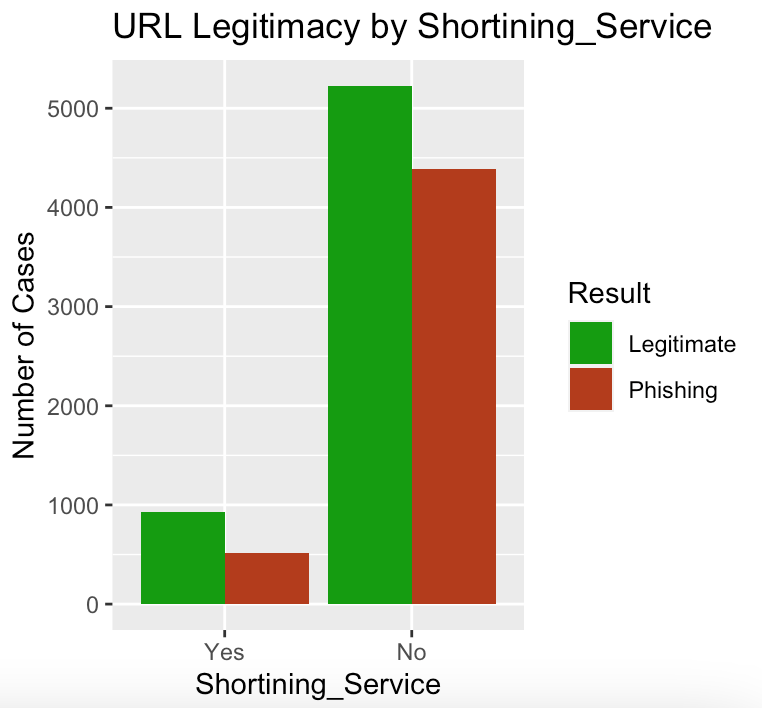
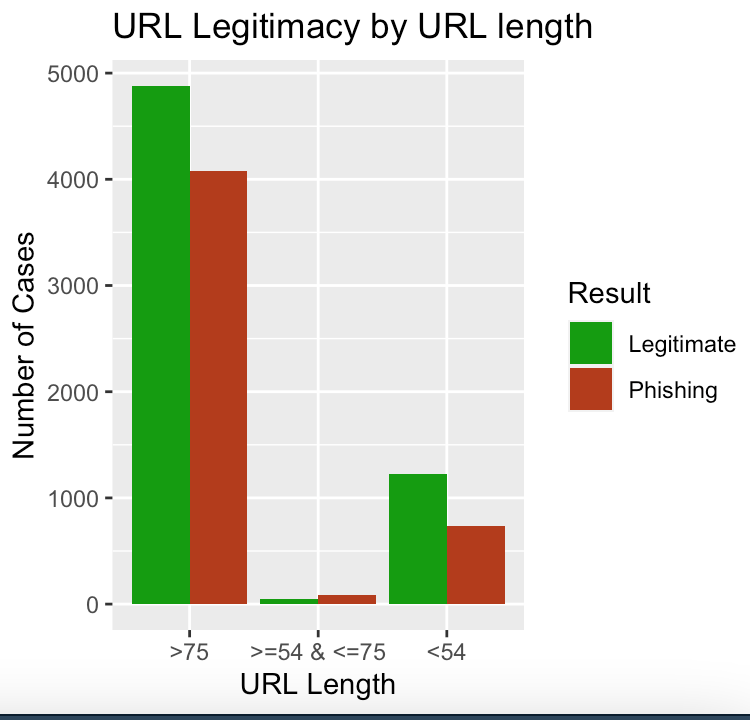
|  |  |
| --- | --- |
| **Variable** | **Description** |
| index | Indexing on the number of rows in the data |
| having\_IPhaving\_IP\_Address | Usage of IP address as an alternative of the domain name |
| URLURL\_Length | Measures the length of the URL |
| Shortining\_Service | Shortened URL using shortening service |
| having\_At\_Symbol | Any existence of @ symbol in the URL |
| double\_slash\_redirecting | Location where double slashing occurs in an URL |
| Prefix\_Suffix | Any dash symbol included in the URL |
| having\_Sub\_Domain | Amount of sub domain that an URL contains |
| SSLfinal\_State | Certificate details assigned with HTTPS, see if the certificate is trusted and valid |
| Domain\_registeration\_length | Validity period of a trustworthy domains |
| Favicon | A graphic image that is linked to a specific webpage |
| port | Validate if a particular service is up based on a preferred server status |
| HTTPS\_token | If HTTPS token happens to be in the domain area of an URL |
| Request\_URL | Percentage of external object shared within the same domain |
| URL\_of\_Anchor | Difference between domain name with <a> tag |
| Links\_in\_tags | Linkage of <> tags to the same domain |
| SFH | Whether the Server Form Handler (SFH) contains empty or blank string and if refers from different domain |
| Submitting\_to\_email | Any redirecting of information to other email |
| Abnormal\_URL | Identity of the URL, if the host’s name is included in the URL |
| Redirect | Frequency of a website being redirected |
| on\_mouseover | Any changes on the status bar |
| RightClick | Whether right click function have been disabled or not |
| popUpWidnow | Any pop up windows from the website that contains text fields |
| Iframe | A HTML tag that allows the display of an extra webpage within the current one being viewed |
| age\_of\_domain | Time length (life) of an existing domain |
| DNSRecord | Information that associates a domain name to an IP address |
| web\_traffic | The popularity of the website |
| Page\_Rank | Importance of web page on the internet |
| Google\_Index | Whether the webpage is in google database |
| Links\_pointing\_to\_page | Quantity of hyperlinks on other websites that direct users to a specific webpage |
| Statistical\_report | External report that shows the analyse if a host belongs to “Top 10 phishing domains” or “Top 10 phishing IP” |
| Result | The final prediction that determines if an URL or website is legitimate or phishing |

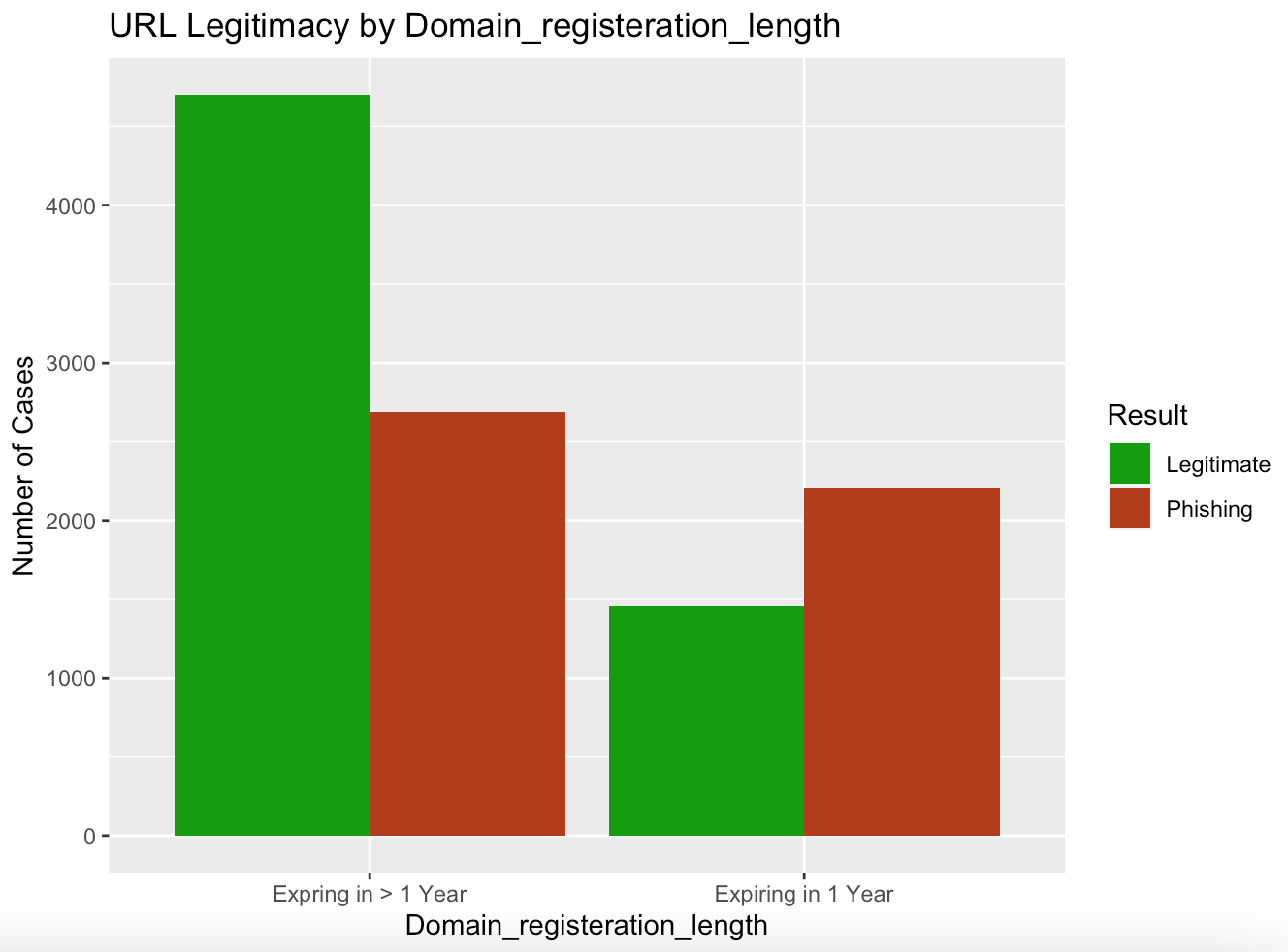
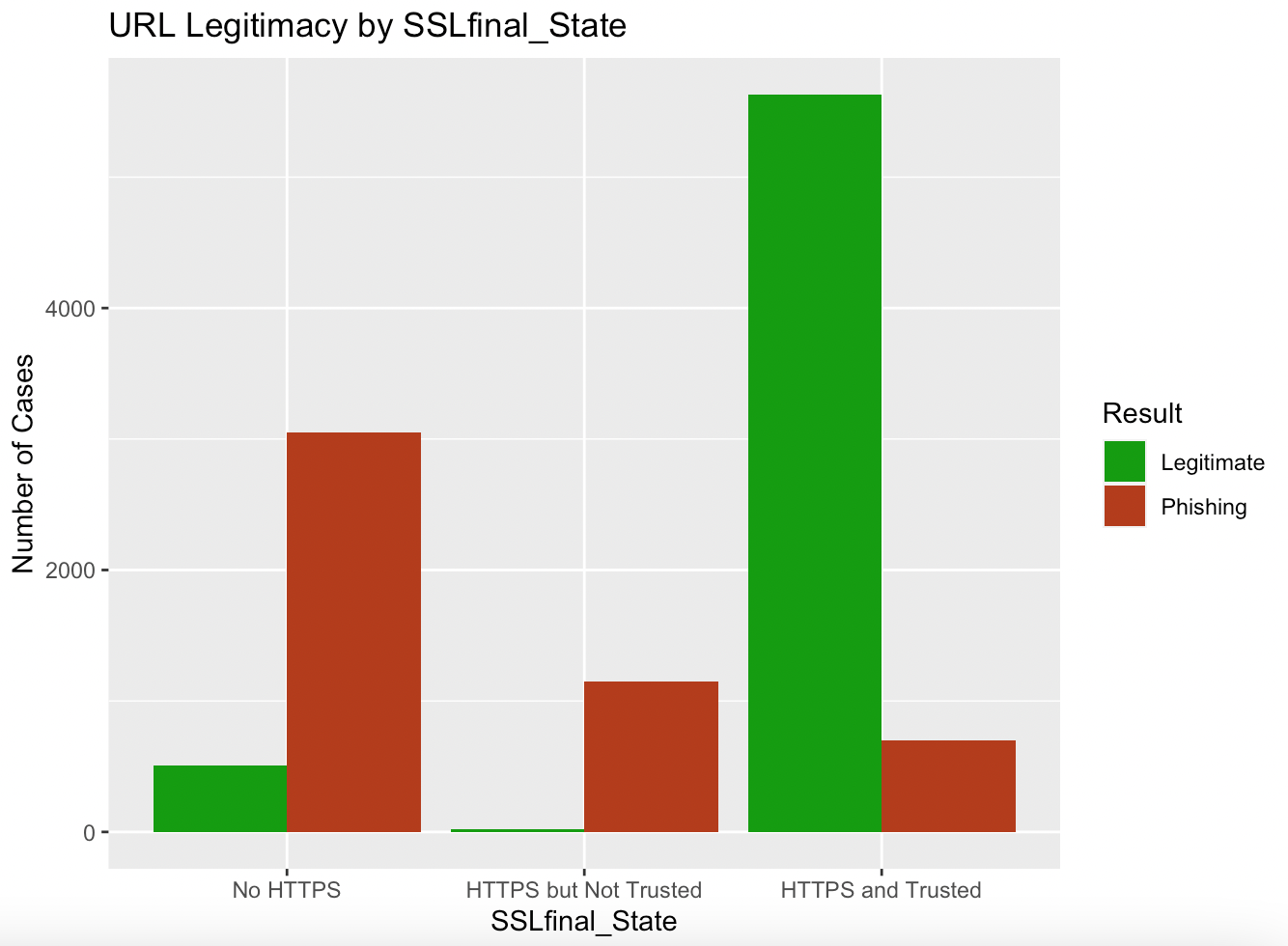
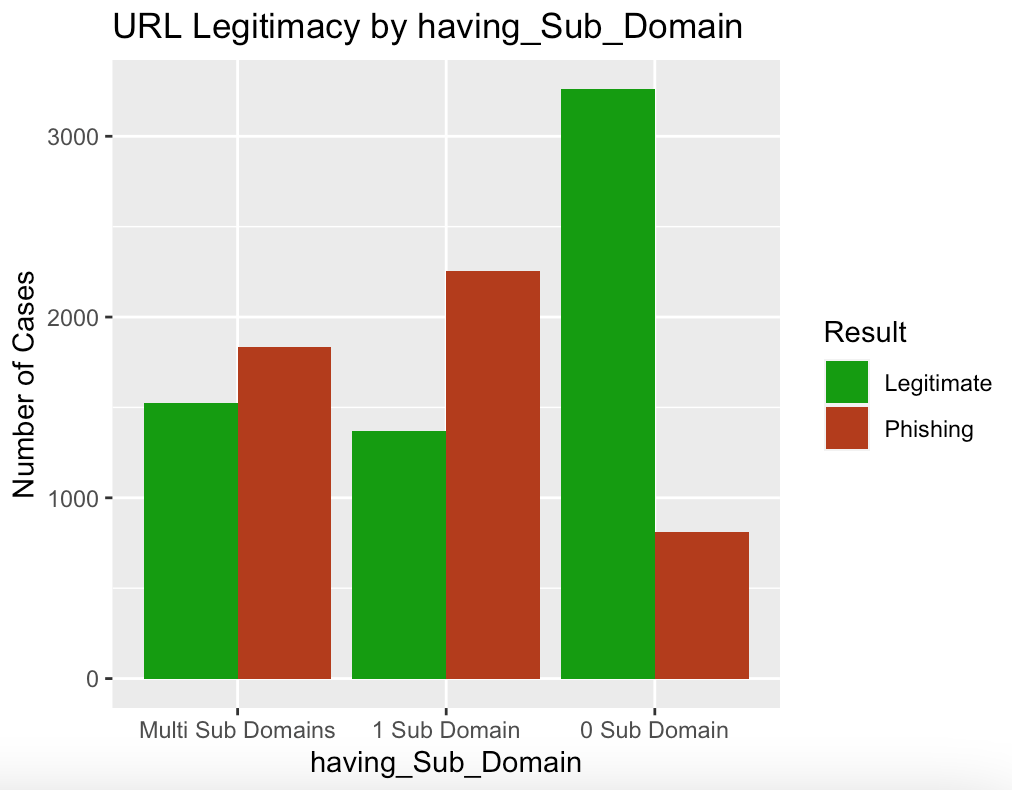
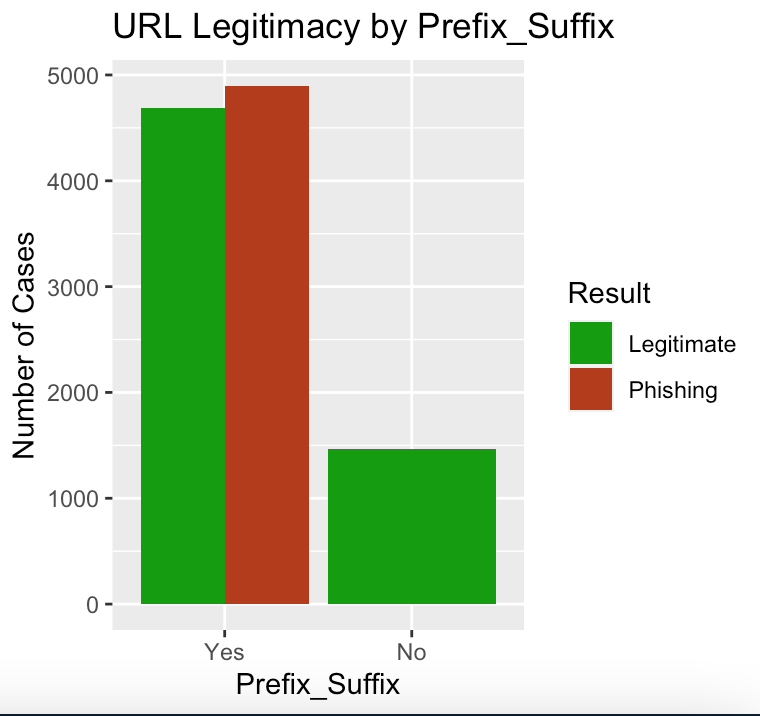
Data source - <https://www.kaggle.com/datasets/akashkr/phishing-website-dataset>

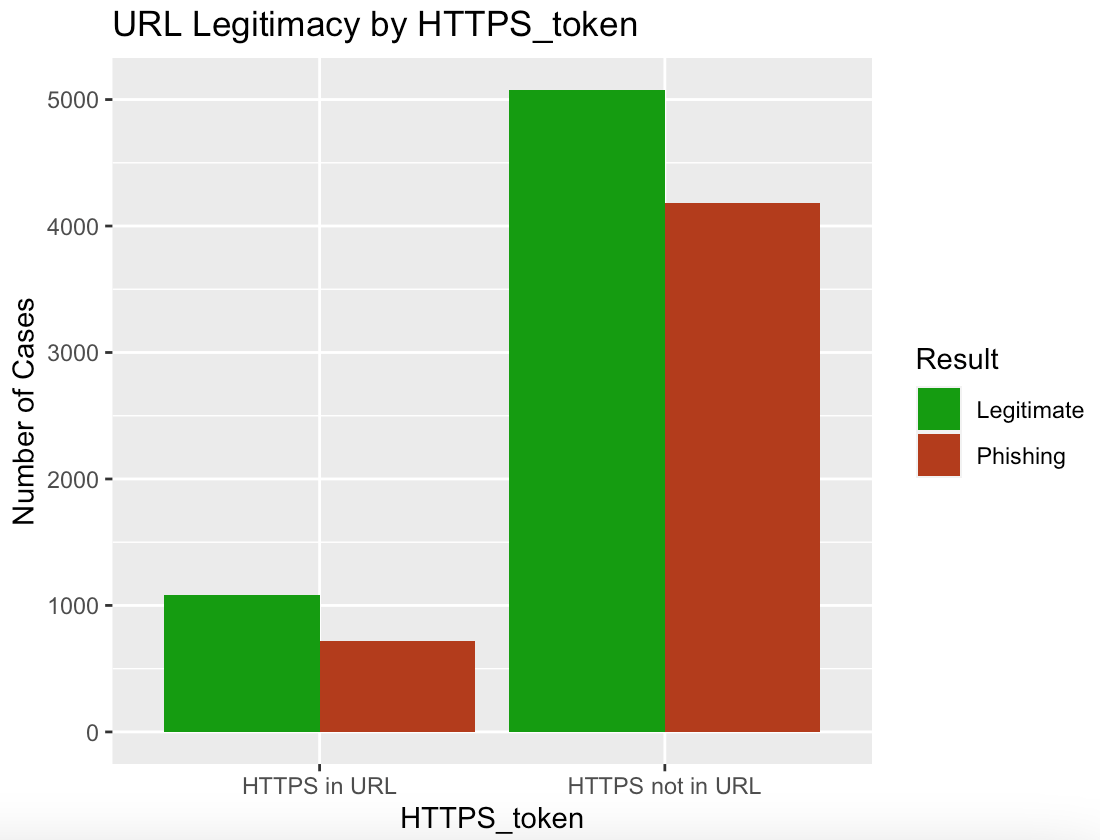
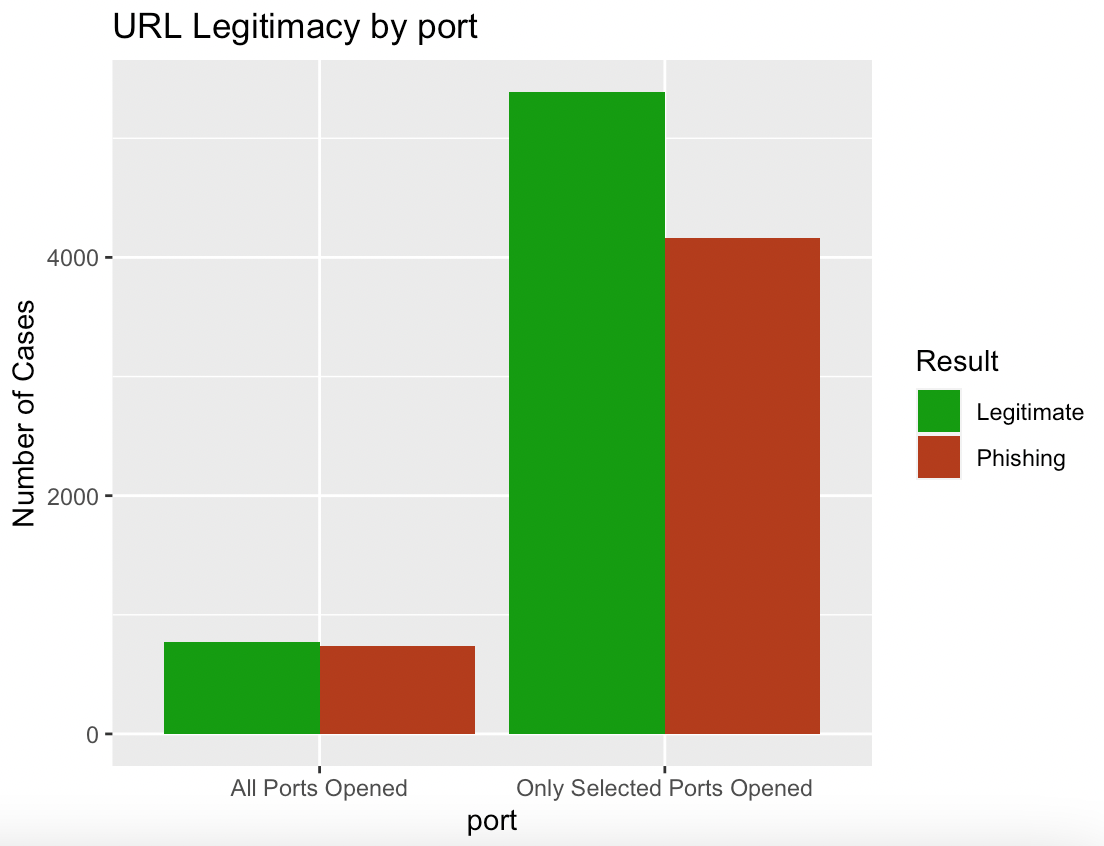
# Appendix B - EDA

**URL Data plots**

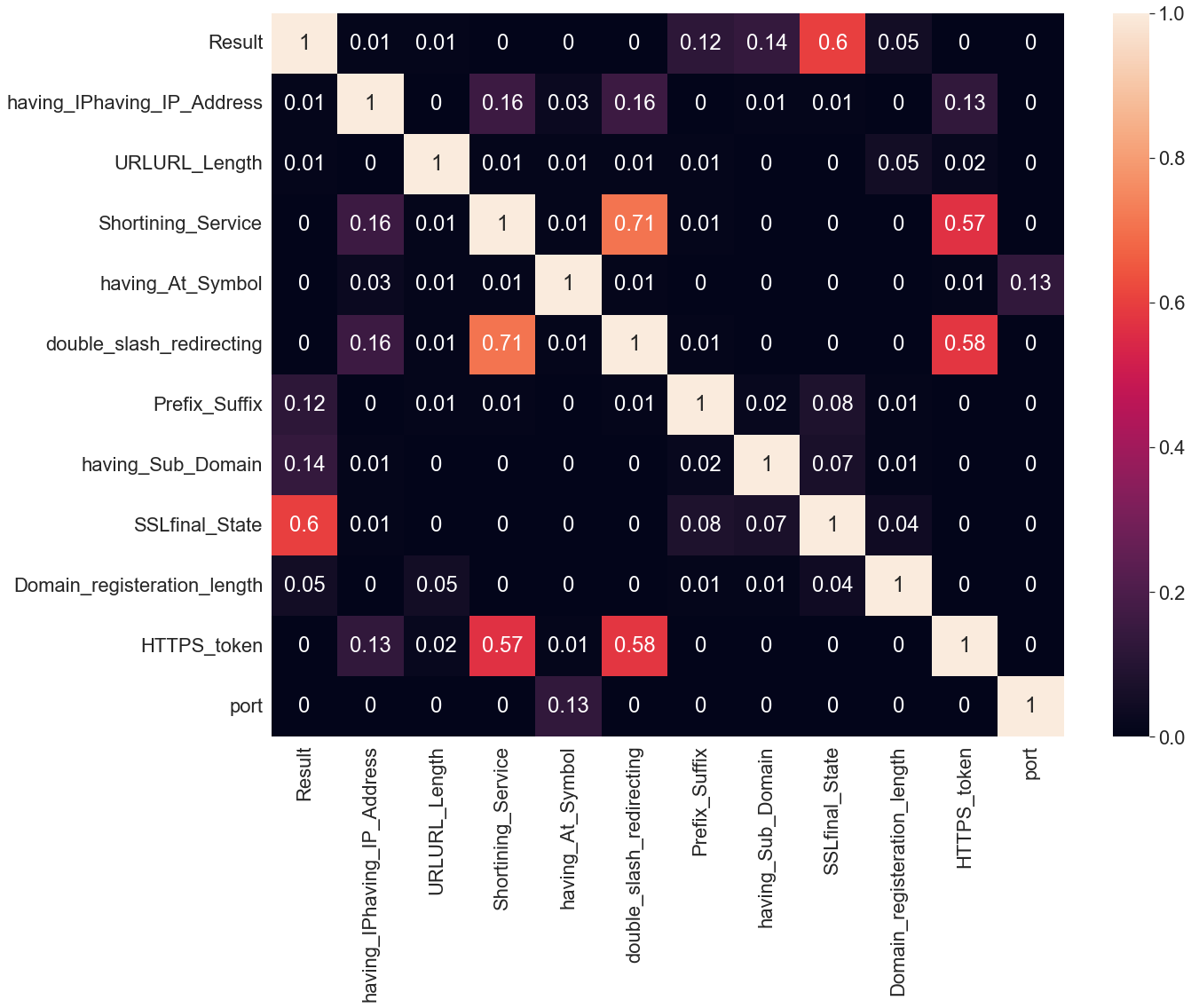








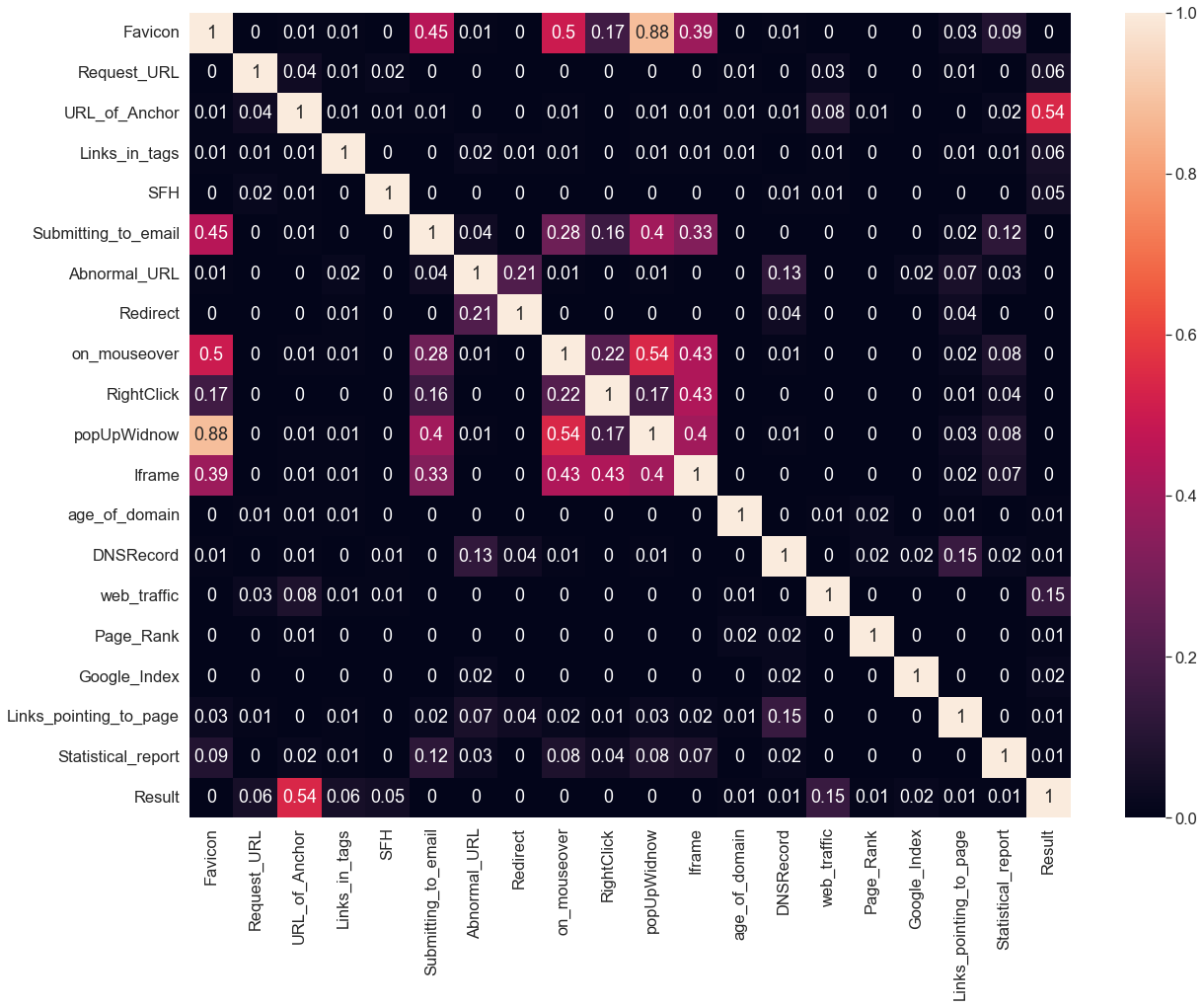
**Cramer’s V of URL data**



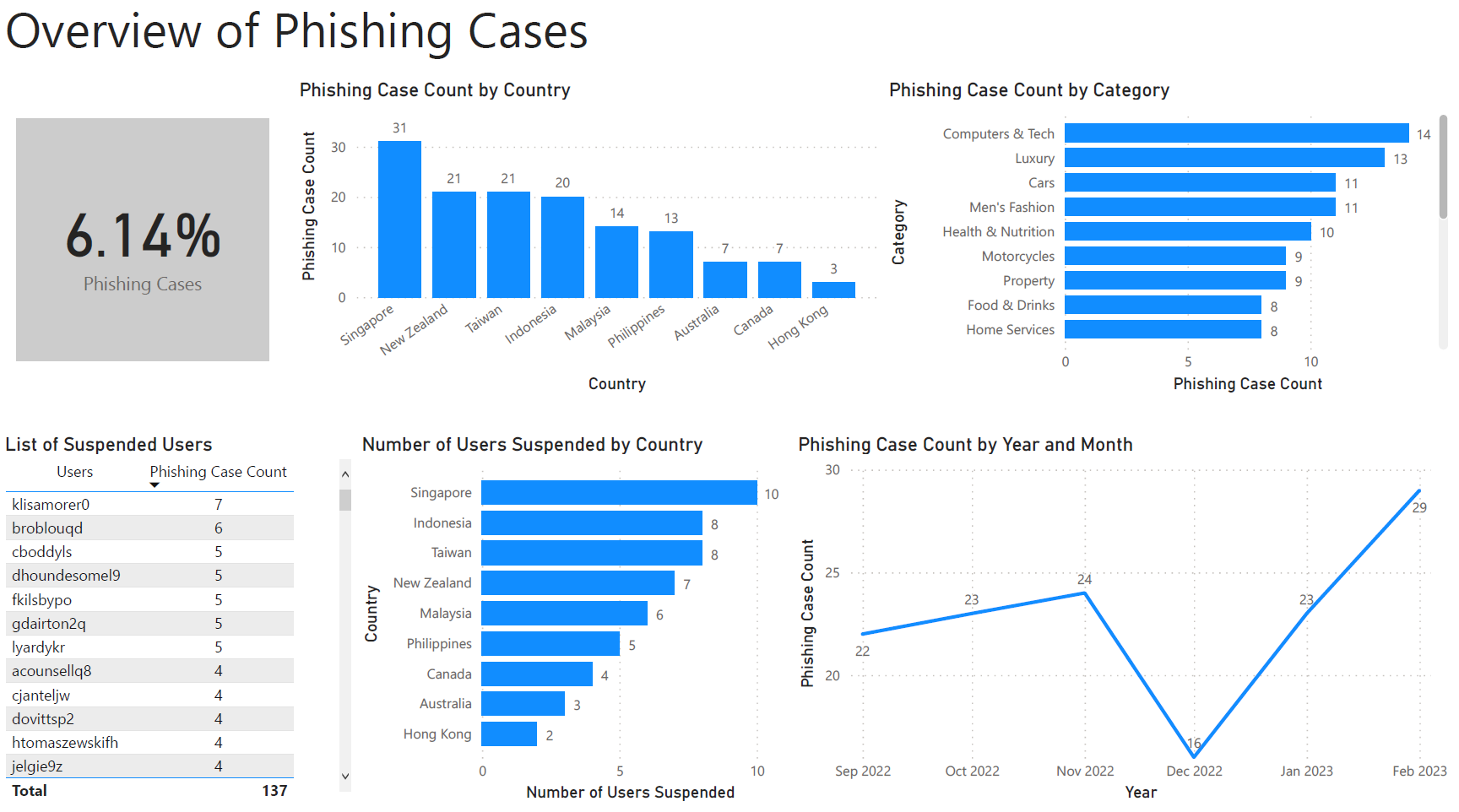
**Website Data plots**

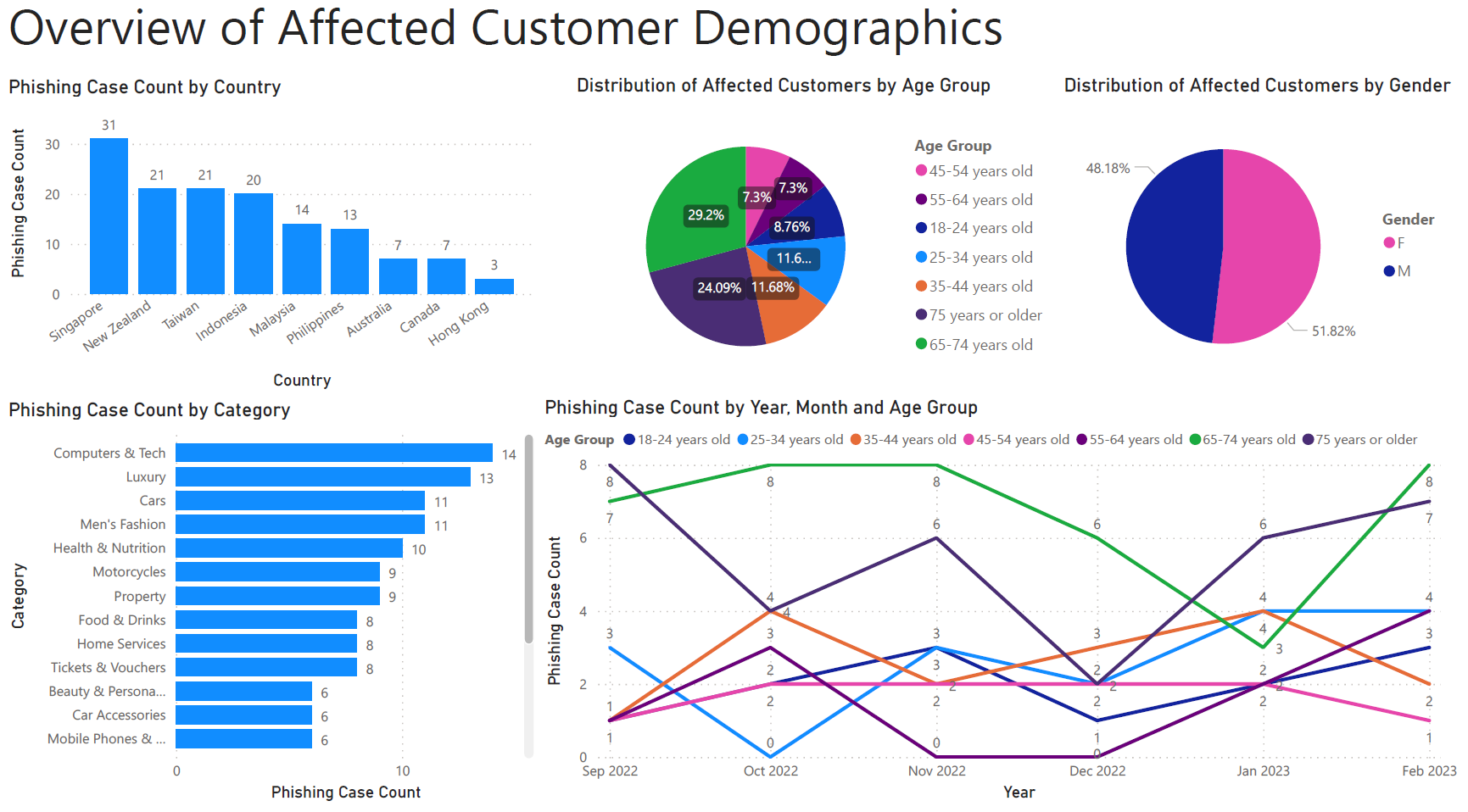
|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

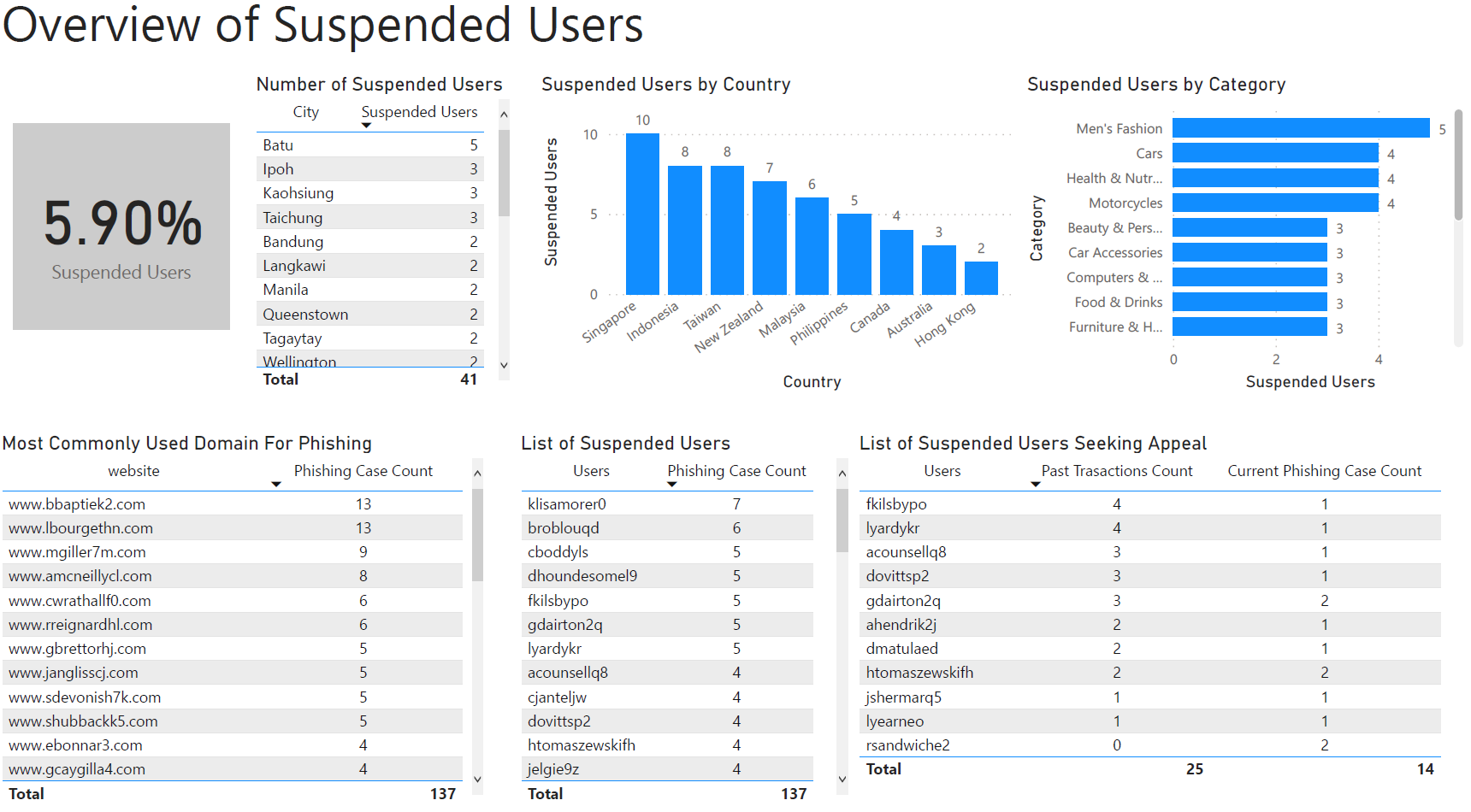
**Cramer’s V of Website Content**



# Appendix C - Dashboard







# 

# 

# 

# 

# 

# 

# 

# 

# Appendix D - Models for Malicious URLs

**Random Forest Models**

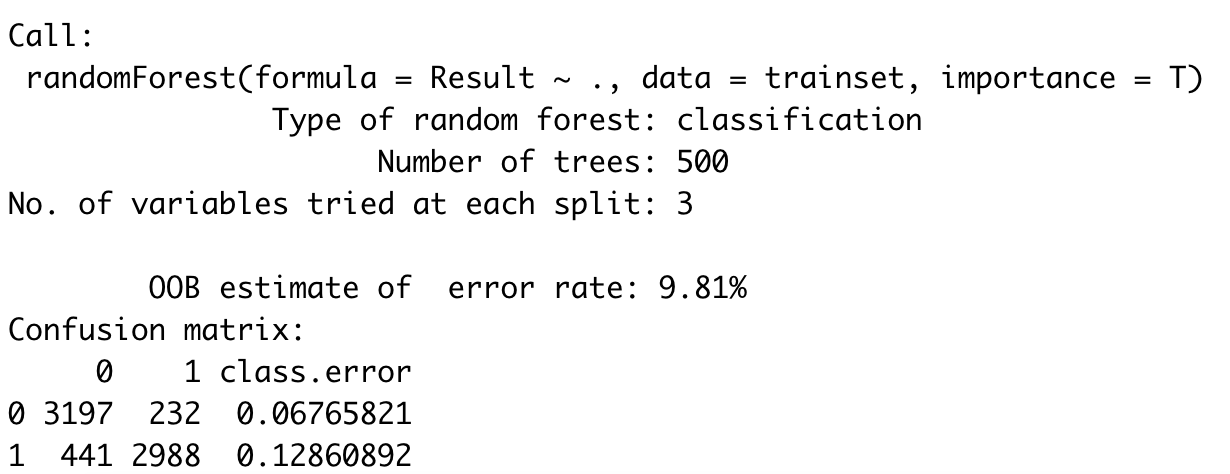
Random Forest Model with Balanced Dataset

*Train dataset*

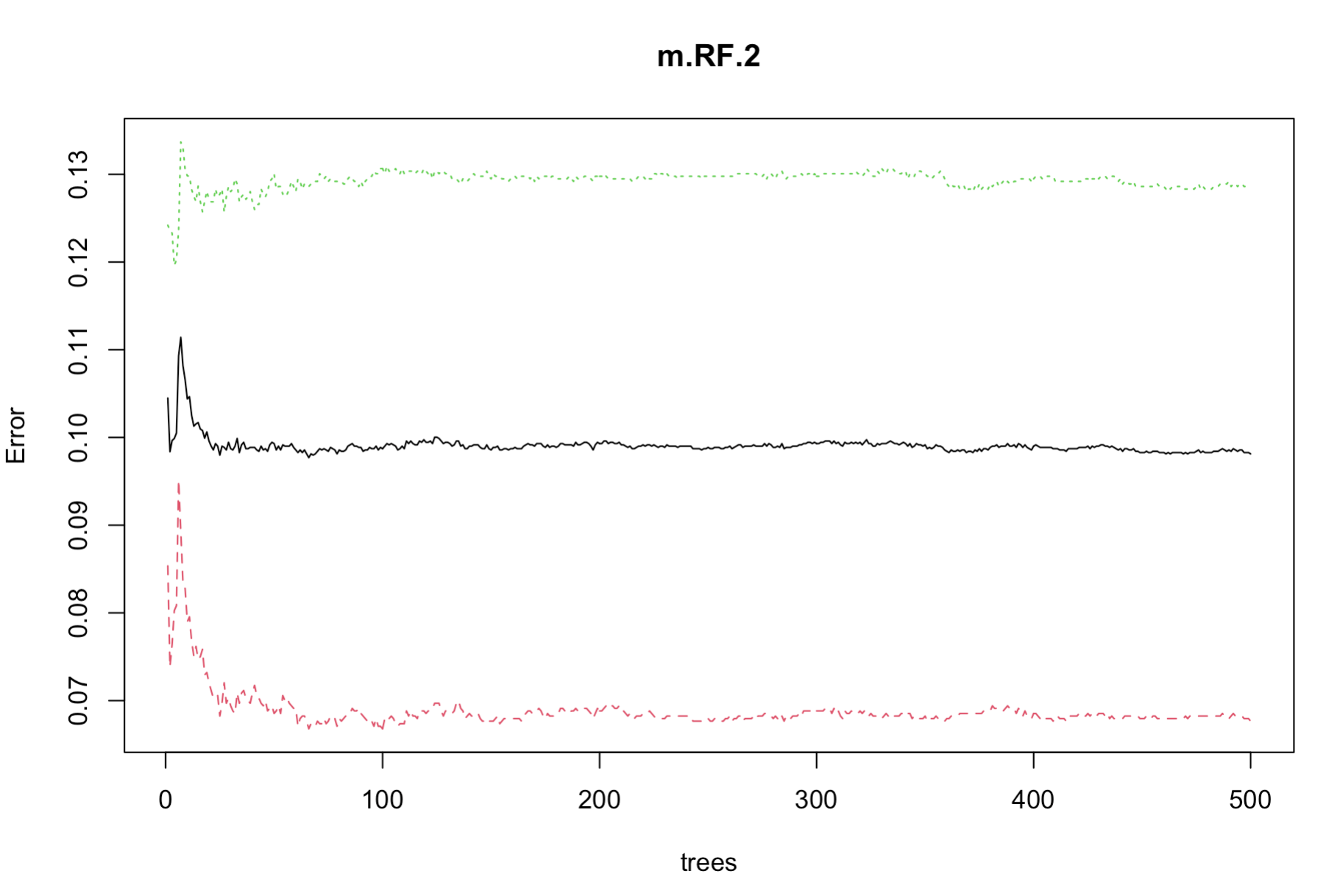
After performing 70-30 train-test split and balancing the trainset by downsampling the majority class, the balanced trainset used to train the Random Forest model for URLs has 3429 samples for both phishing and legitimate URLs.



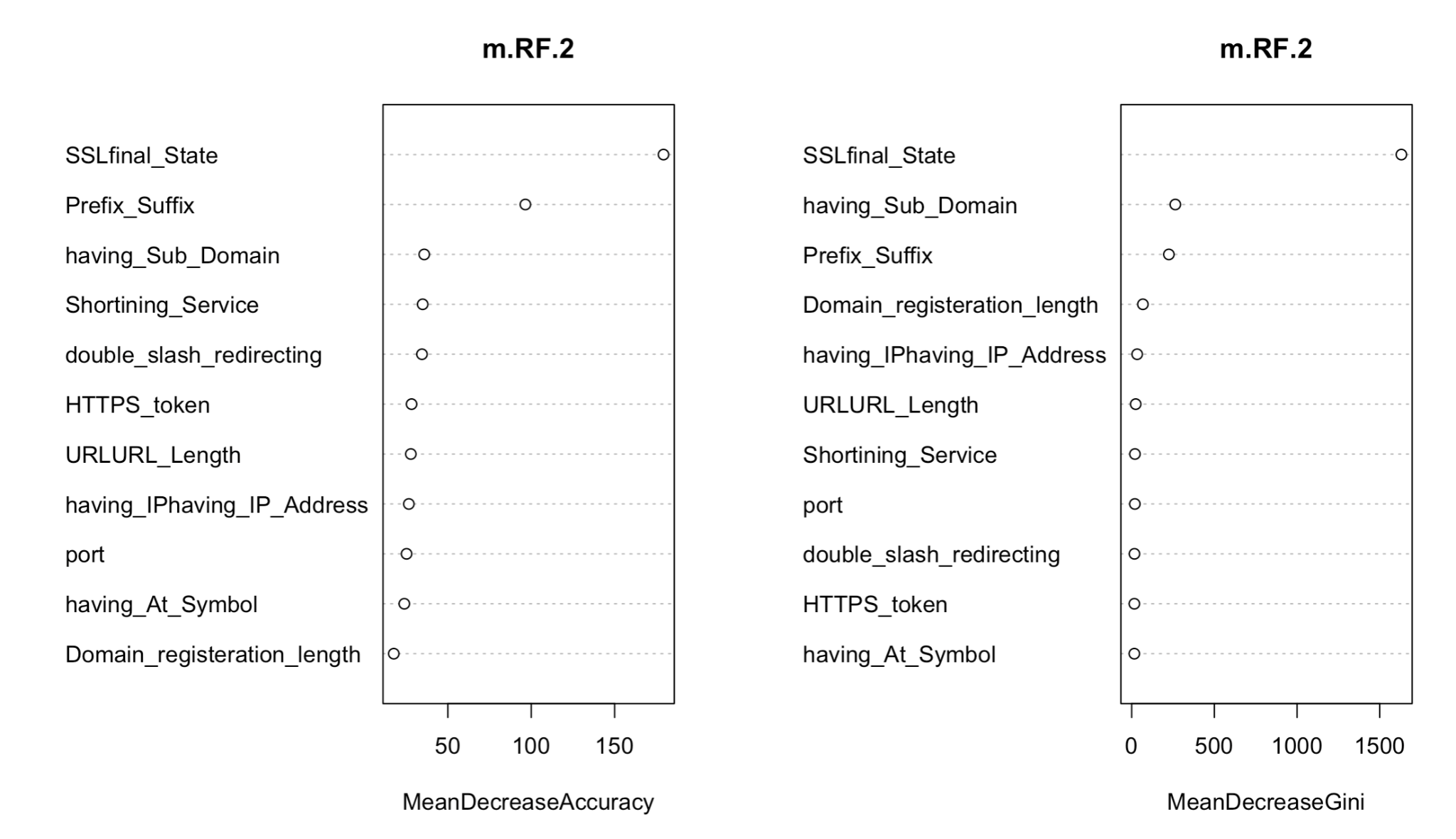
Output of Random Forest model built using the balanced trainset:

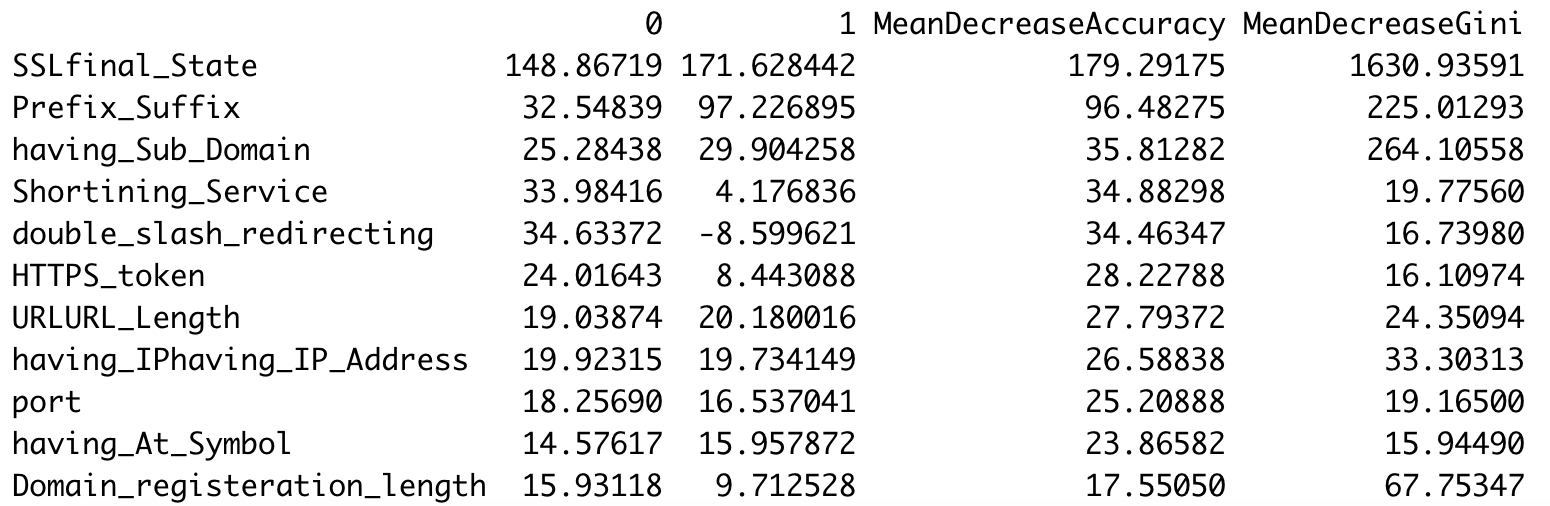


The OOB error rate is 9.81% for the Random Forest model built with the balanced trainset for predicting the URLs. The default number of trees in the Random Forest model is 500 and the number of variables at each split is 3.

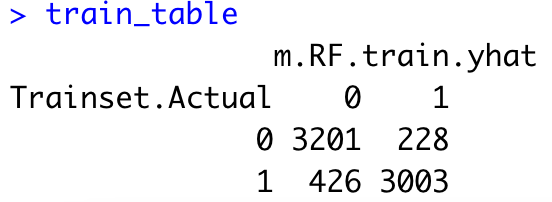
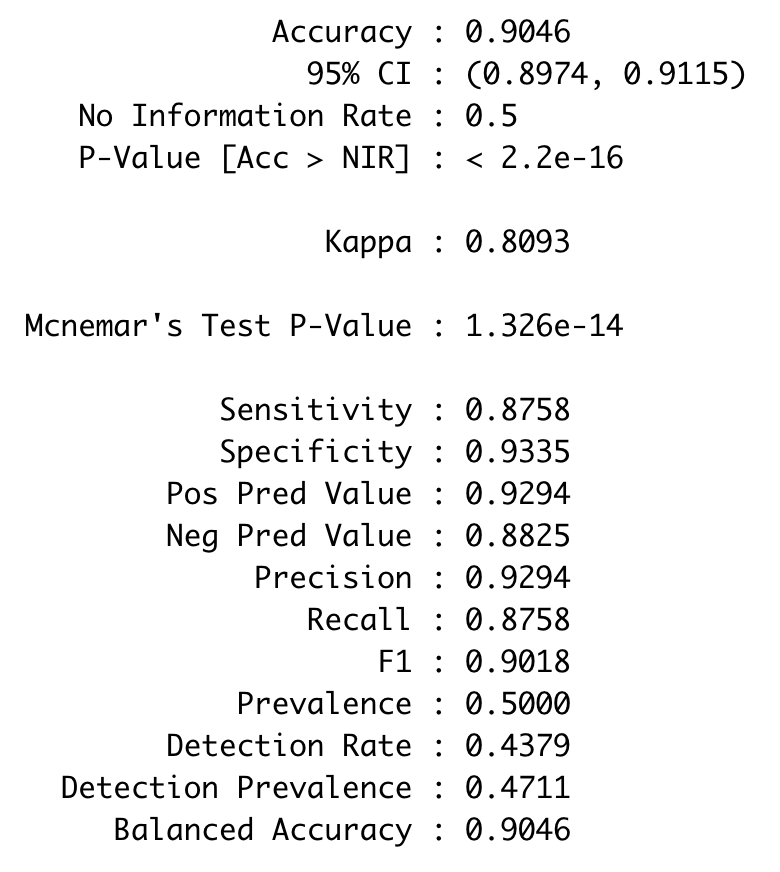


The errors stabilised at about 500 trees.





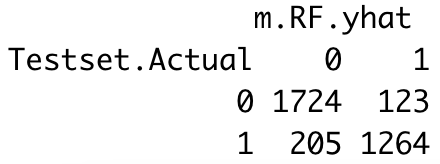
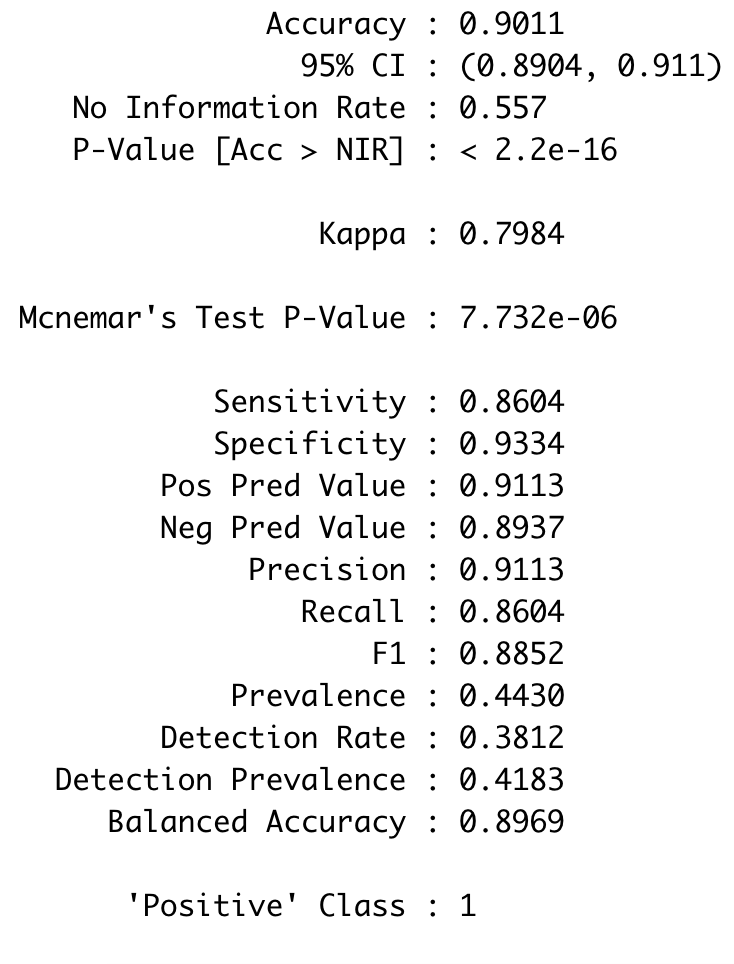
Mean Decrease Accuracy (MDA) was set more than 0 for the model and all variables have Mean Decrease Accuracy and Mean Decrease Gini of more than 0, which means all the variables being selected are important and no variables are eliminated, suggesting that they are useful to some extent in the prediction. From the plot, we could see that SSLfinal\_State has the highest score for both Mean Decrease Accuracy and Mean Decrease Gini. Something interesting is that SSLfinal\_State has a score 2 times the second important variable, “Prefix\_Suffix”. This means that SSLfinal\_State is largely significant to the result of whether an URL is legitimate or phishing.



Trainset Accuracy: 90.46%, FNR: 12.42%, FPR: 6.65%, Precision: 92.94%, Recall: 87.58%.

*Test dataset*

After training the model, the trained model predicts the testset to give the testset metrics.



Testset Accuracy: 90.11%, FNR: 13.96%, FPR: 6.66%, Precision: 91.13%, Recall: 86.04%.

**Performance of** **Random Forest Model Balanced Test Datasets:**

|  |  |
| --- | --- |
| **Malicious URL Detection** | **Balanced Test Dataset** |
| Confusion Matrix |  |
| Overall Accuracy | 0.9010856 |
| False Positive Rate (FPR) | 0.0665945 |
| False Negative Rate (FNR) | 0.1395507 |
| Precision | 0.9113 |
| Recall | 0.8604 |

**Neural Network Models**

Neural Network Model with Balanced Dataset

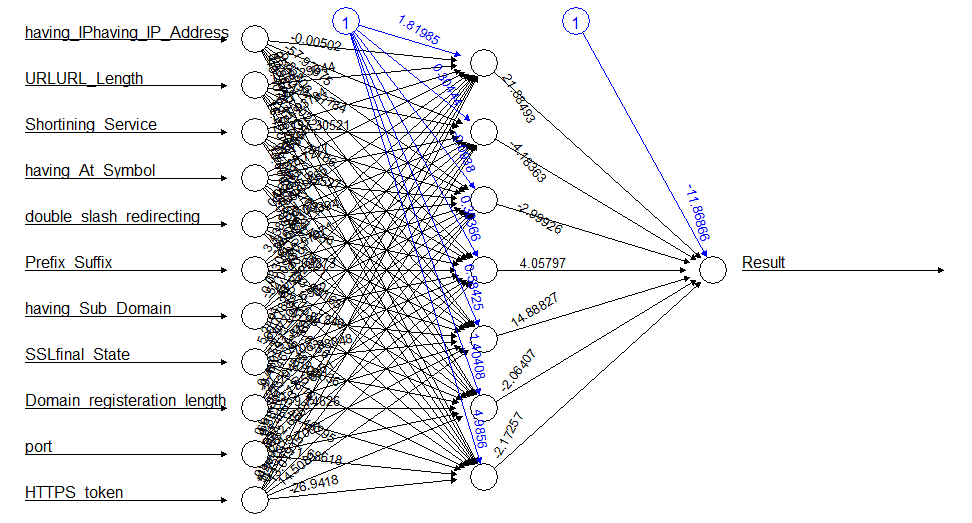
The neural network's Result variable is different from the rest. Phishing is 0 and Legitimate is 1. When using the same result variable as the rest, the models could not converge.



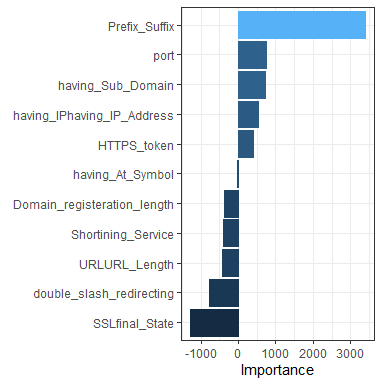
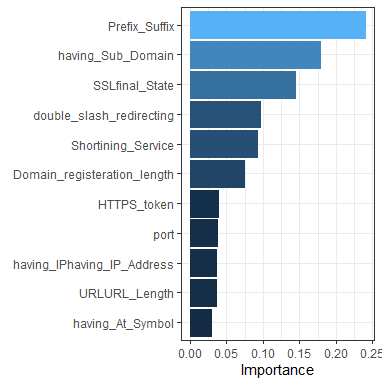
The unbalanced and balanced dataset is used to run 2 separate neural network models with 1 hidden layer of 7 nodes.



The final Neural Network Diagram using the balanced dataset



Using Garson() and Olden() on the model developed with the balanced dataset to calculate the feature importance. We can see that the Prefix\_Suffix is the most significant feature with the highest relative importance in determining whether the link is a phishing link or a legitimate link.



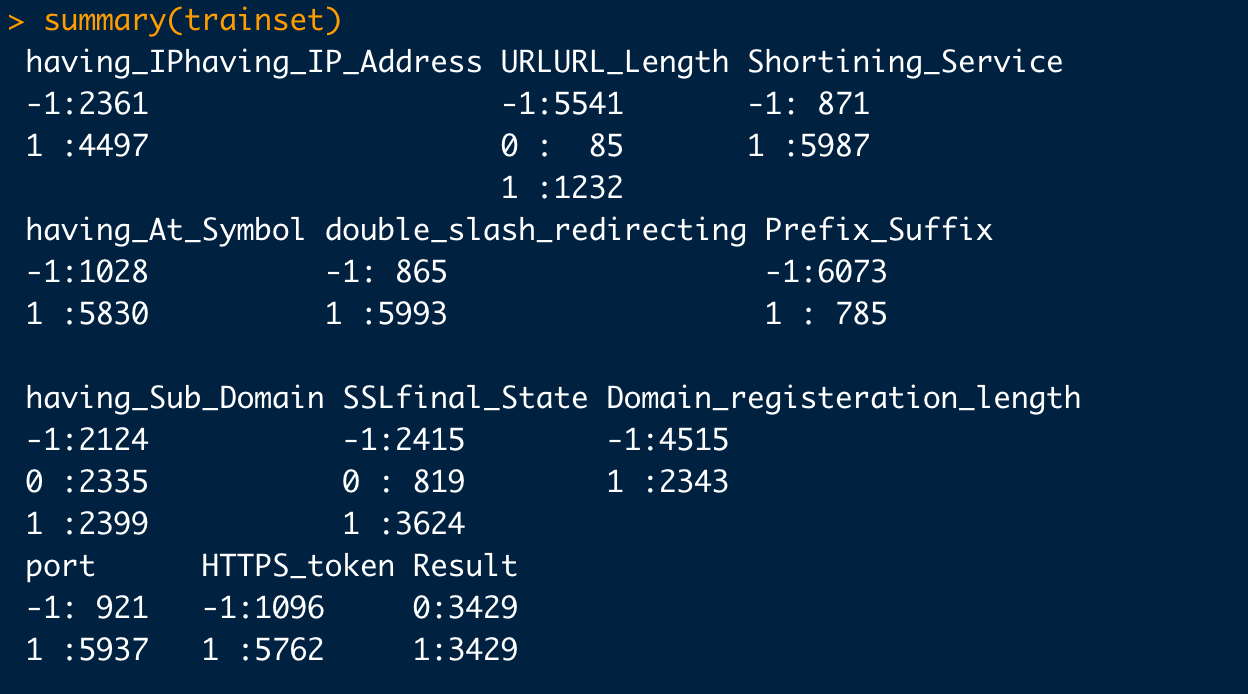
**Performance of Neural Network Model Balanced Test Datasets:**

|  |  |
| --- | --- |
| **Malicious URL detection** | **Balanced Test Dataset** |
| Confusion Matrix |  |
| Overall Accuracy | 0.90168878 |
| False Positive Rate | 0.07456372 |
| False Negative Rate | 0.09894737 |
| Precision | 0.90105263 |
| Recall | 0.87406399 |

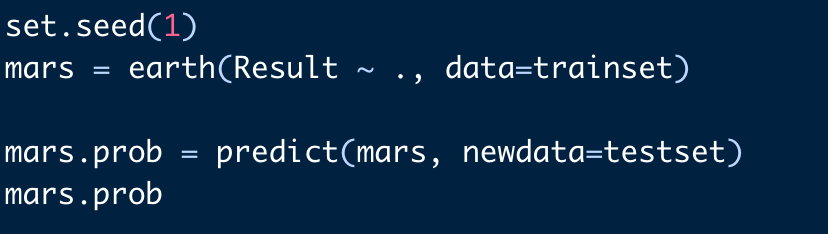
**MARS Model**

MARS Model with Balanced Dataset

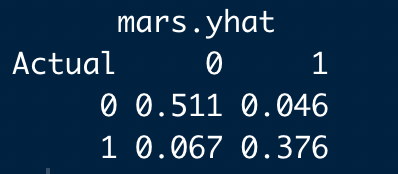
After performing a 70-30 train-test split and balancing the trainset by downsampling the majority class, the balanced trainset used to train the MARS model for malicious URLs has 3429 samples for both phishing and legitimate URLs.



We fit a MARS model to the training dataset to predict the testset data.



The probability threshold for classification is set to 0.5. A probability value that is higher than the threshold value will be deemed as phishing. Next, we create a confusion matrix and find the respective performance metrics.



Testset Accuracy: 88.75%, FNR: 11.59%, FPR: 8.26%, Precision: 89.20%, Recall: 84.89%.

**Performance of** **MARS Model Balanced Test Datasets:**

|  |  |
| --- | --- |
| **Malicious URL detection** | **Balanced Test Dataset** |
| Confusion Matrix |  |
| Overall Accuracy | 0.8875 |
| False Positive Rate | 0.0825853 |
| False Negative Rate | 0.1159170 |
| Precision | 0.8920 |
| Recall | 0.8489 |

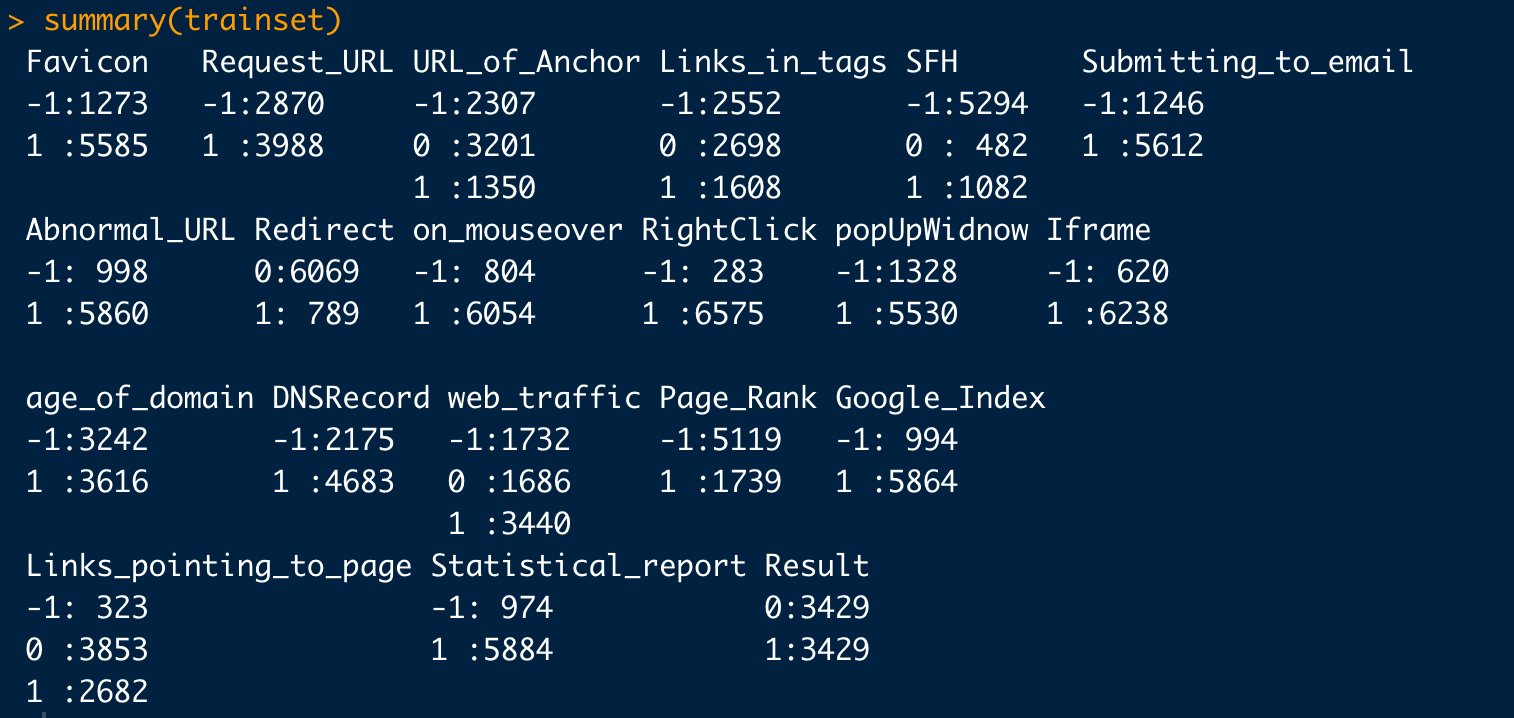
# Appendix E - Models for Malicious Websites

**Random Forest Models**

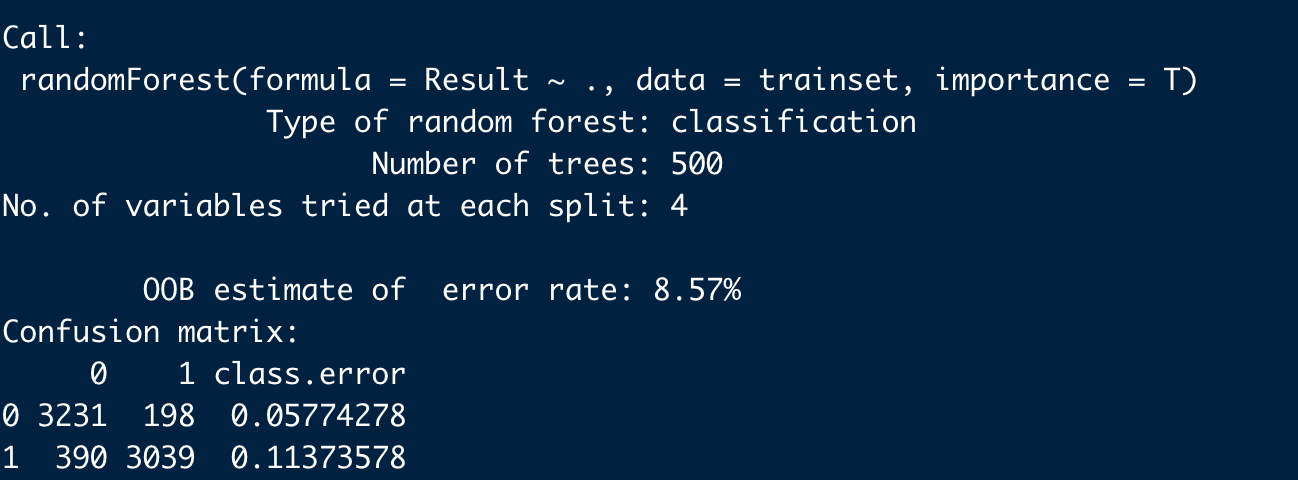
Random Forest Model with Balanced Dataset

*Train Dataset*

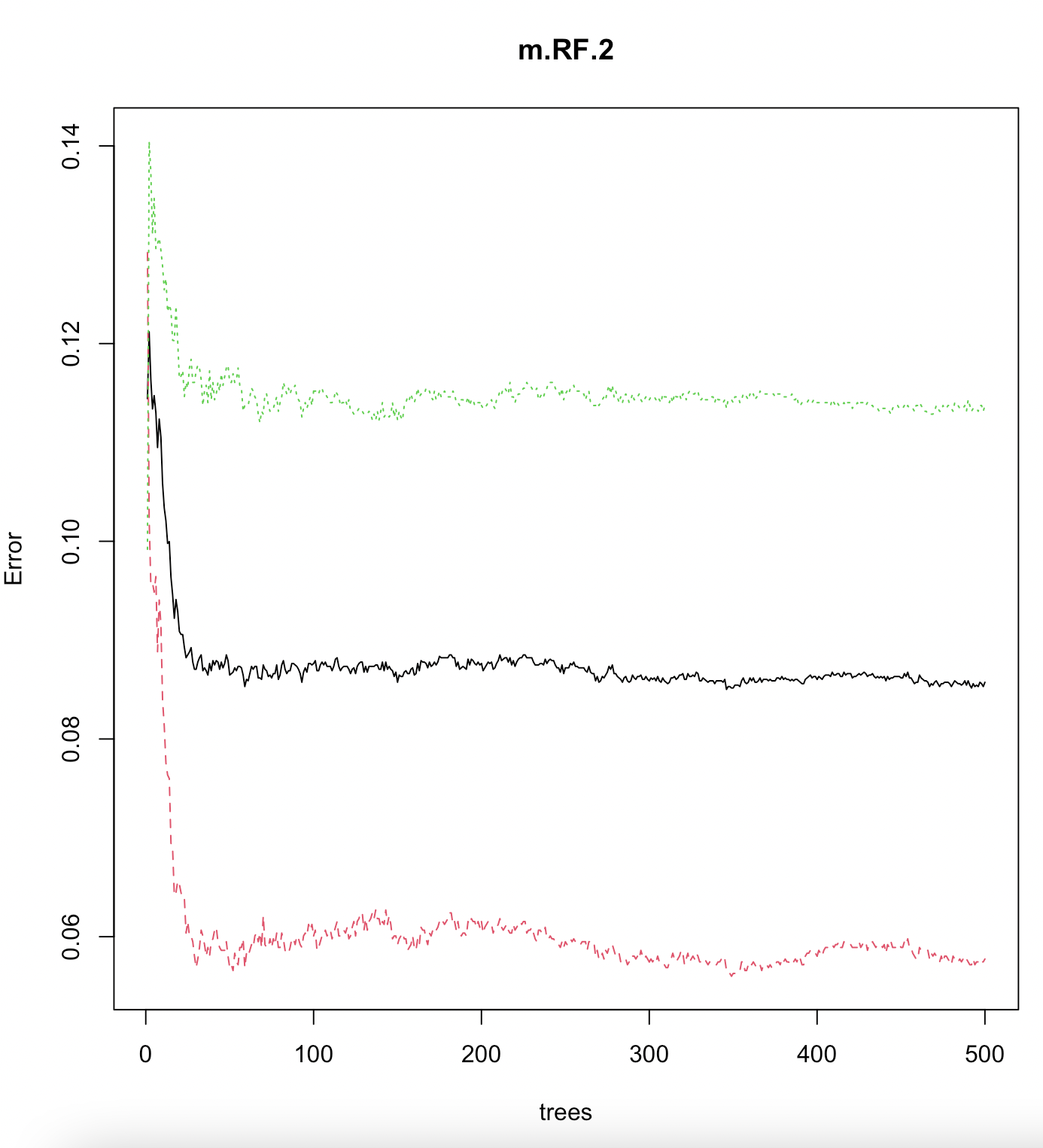
The balanced trainset used to build the Random Forest model for predicting website content has 3429 samples for both phishing and legitimate website content, similar to the one used for predicting malicious URLs.

****

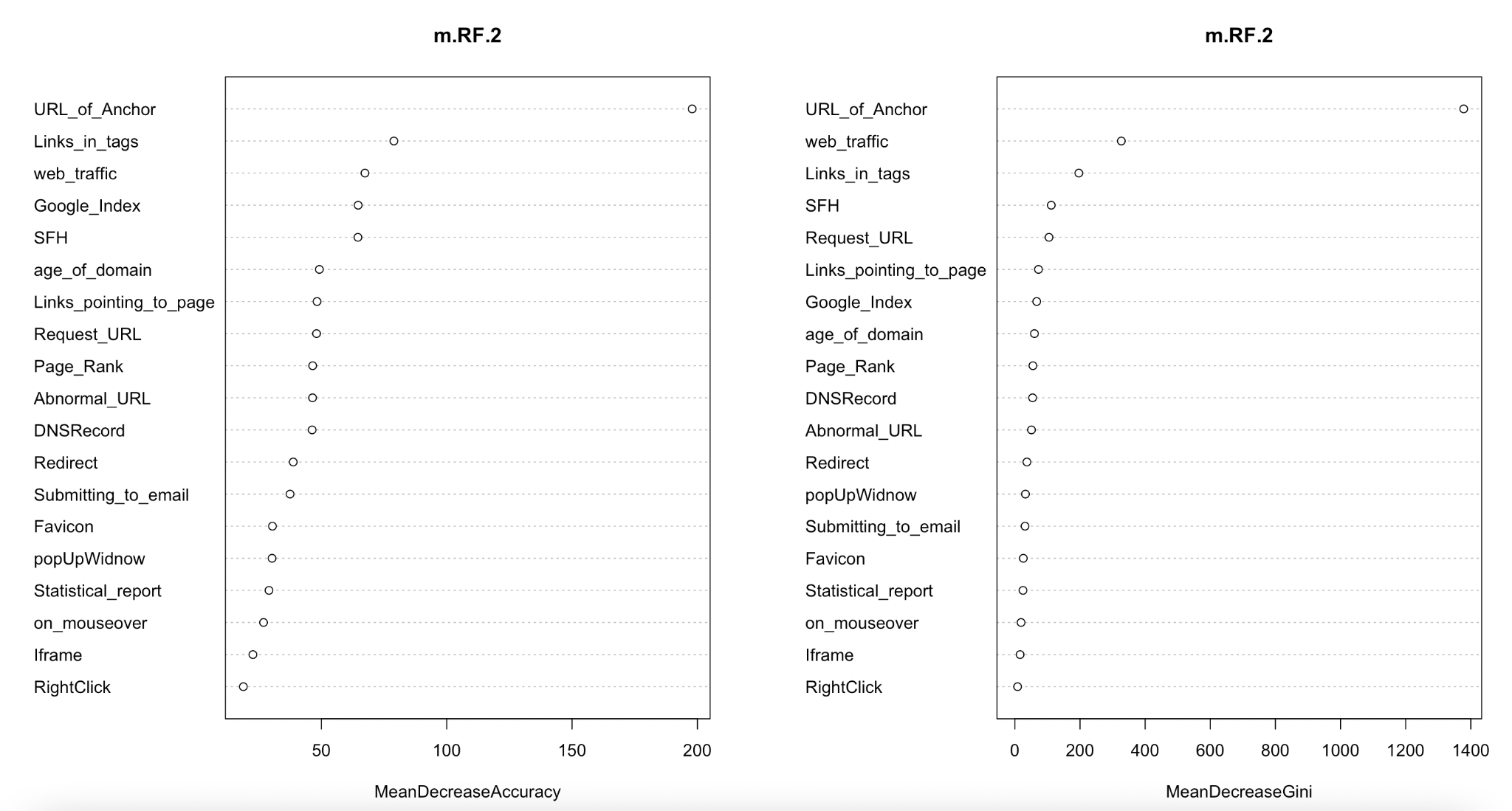
Output of Random Forest model built using the balanced trainset:

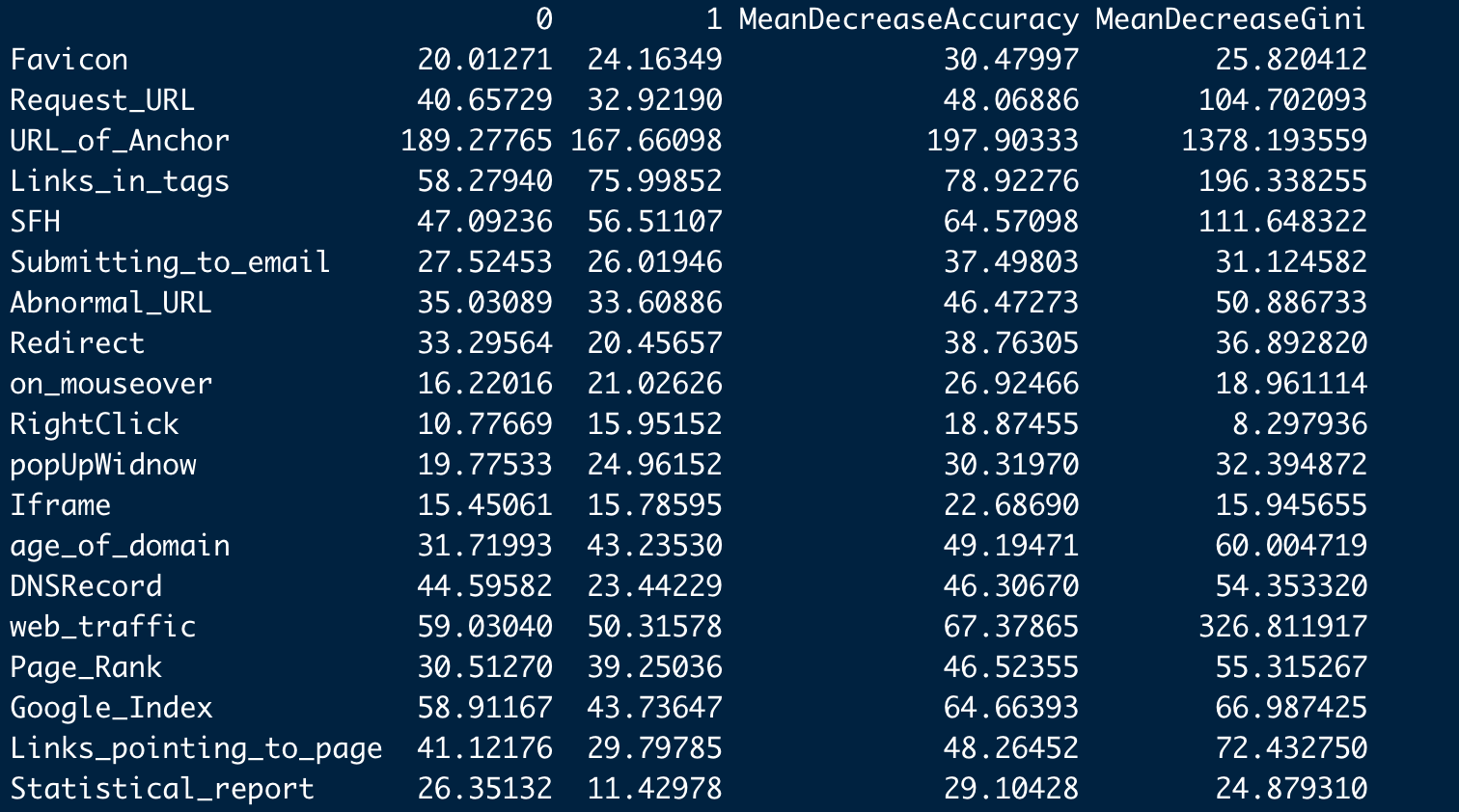
****

The OOB error rate is 8.57% for the Random Forest model built with the balanced trainset for predicting malicious website content. The default number of trees in the Random Forest model is 500 and the number of variables at each split is 4, which is slightly different from the Random Forest used for malicious website content detection.

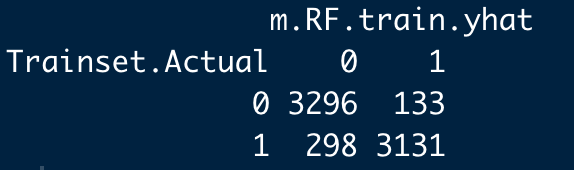
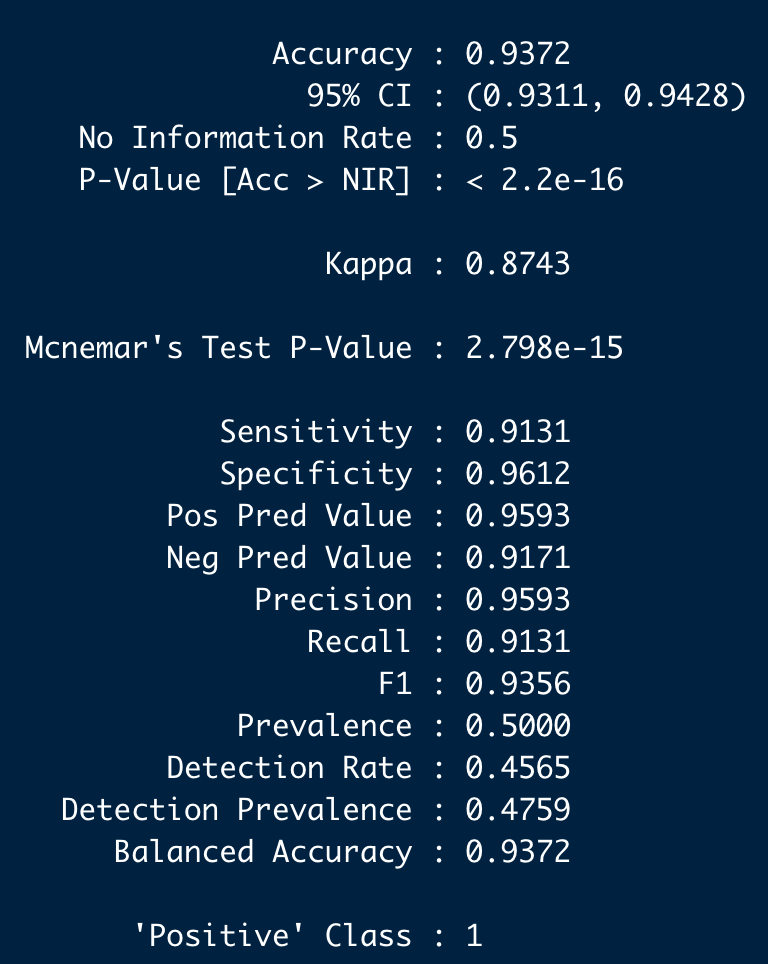


The errors stabilised at about 500 trees.

****

****

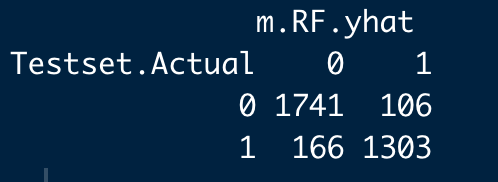
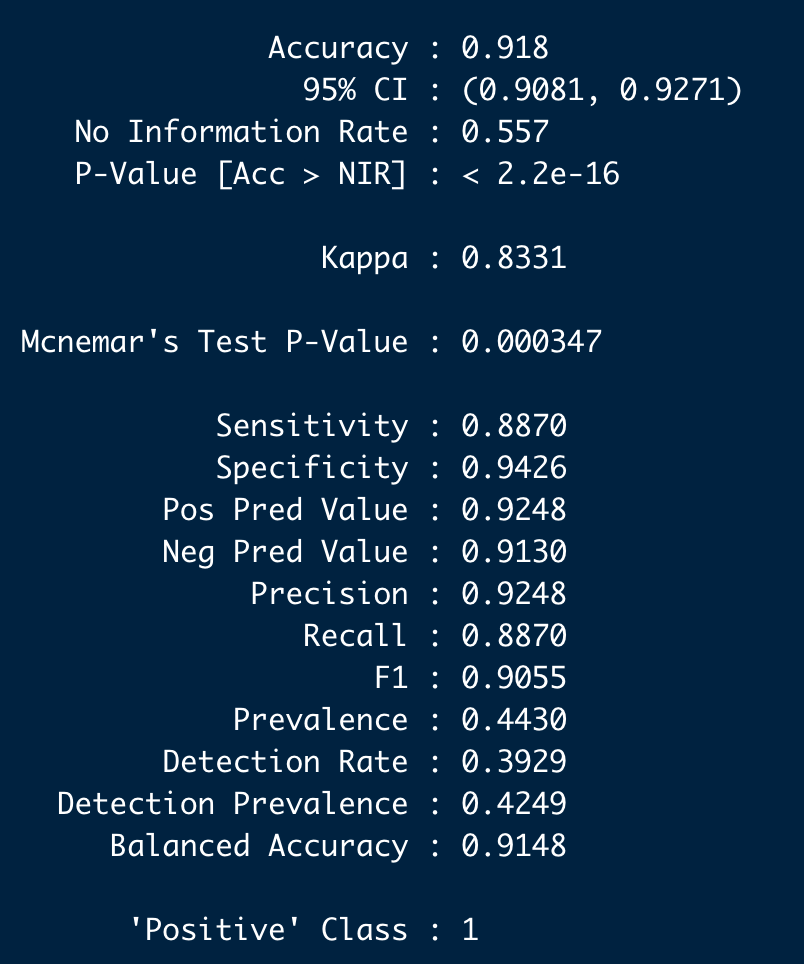
Mean Decrease Accuracy (MDA) was set more than 0 for the model and all variables have Mean Decrease Accuracy and Mean Decrease Gini more than 0, which means all the variables being selected are important and no variables are eliminated, suggesting that they are useful to some extent in the prediction. From the plot, we could see that URL\_of\_Anchor has the highest score for both Mean Decrease Accuracy and Mean Decrease Gini. This means that URL\_of\_Anchor is largely significant to the result of whether a website content is malicious or legitimate.



Trainset Accuracy: 93.72%, FNR: 8.69%, FPR: 3.88%, Precision: 95.93%, Recall: 91.31%.

*Test Dataset*

After training the model, the model predicts the testset using the trained model. The model will then be put into the testset to give the testset metrics.



Testset Accuracy: 91.80%, FNR: 11.30%, FPR: 5.74%, Precision: 92.48%, Recall: 88.70%.

**Performance of** **Random Forest Model Balanced Test Datasets:**

|  |  |
| --- | --- |
| **Malicious Website Content Detection** | **Balanced Test Dataset** |
| Confusion Matrix |  |
| Overall Accuracy | 0.9179735 |
| False Positive Rate (FPR) | 0.0573904 |
| False Negative Rate (FNR) | 0.113002 |
| Precision | 0.9248 |
| Recall | 0.8870 |

**Neural Network Models**

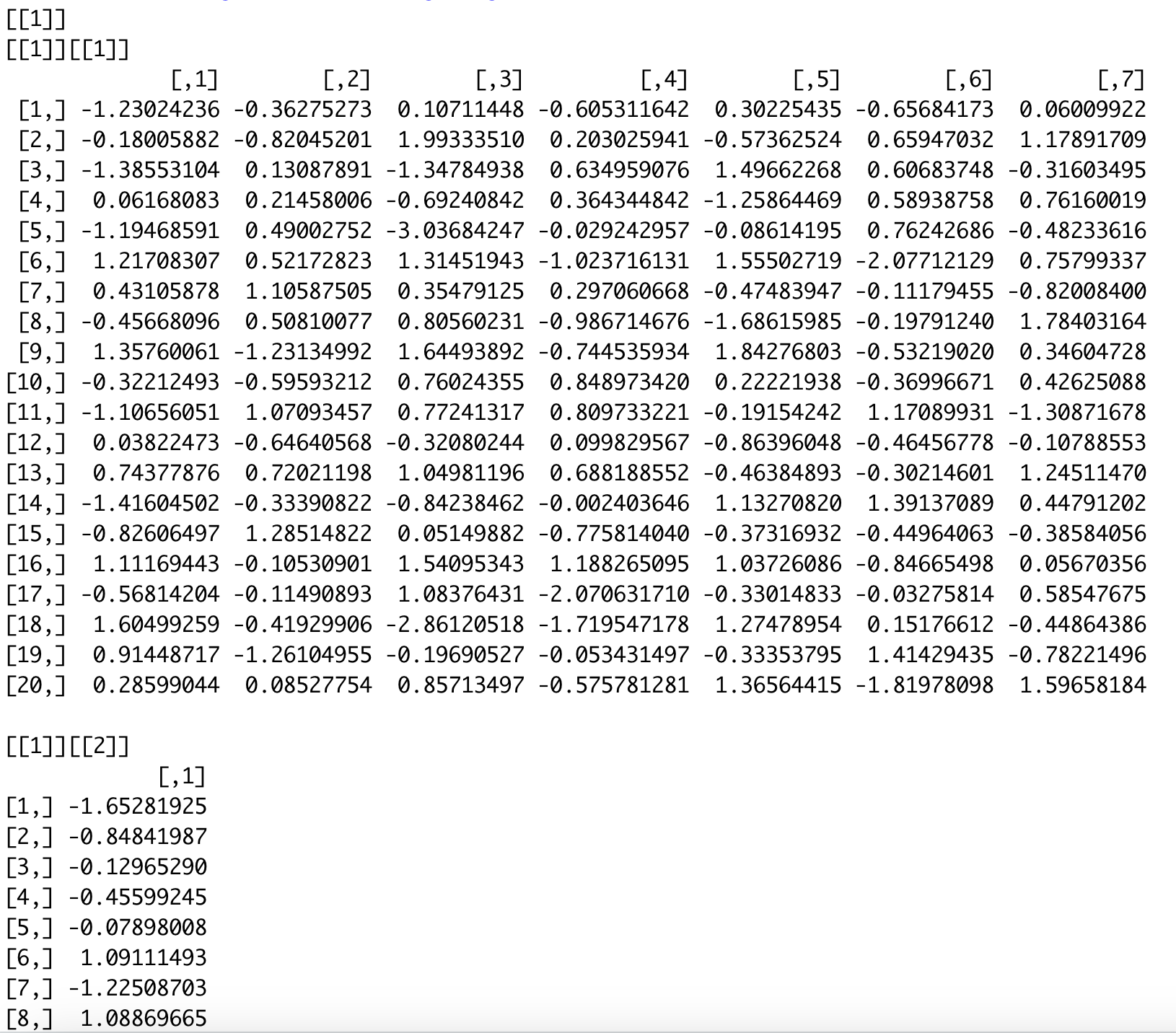
Neural Network Model with Balanced Dataset

The unbalanced and balanced dataset is used to run 2 separate neural network models with 1 hidden layer of 7 nodes.

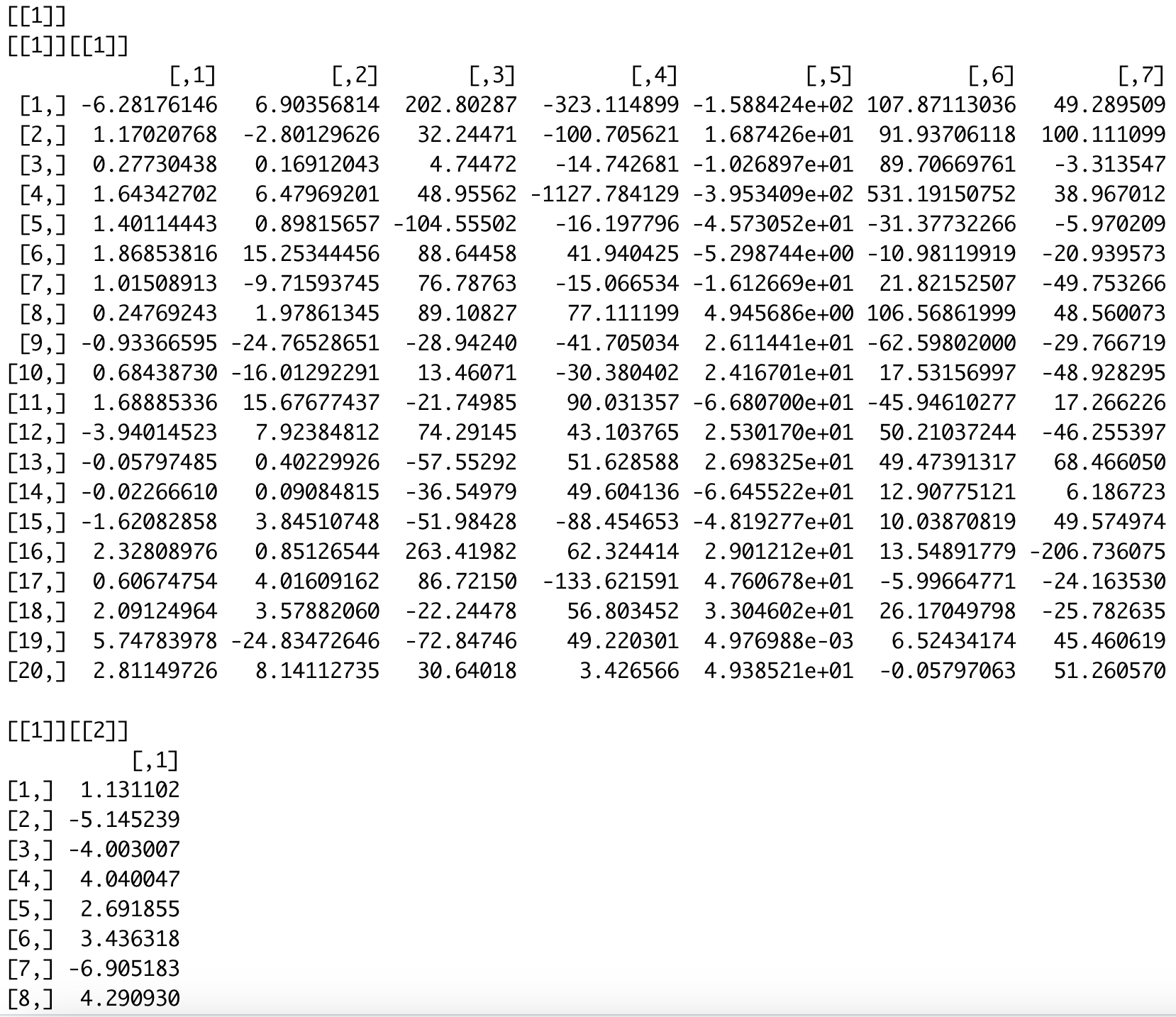


Balanced Dataset

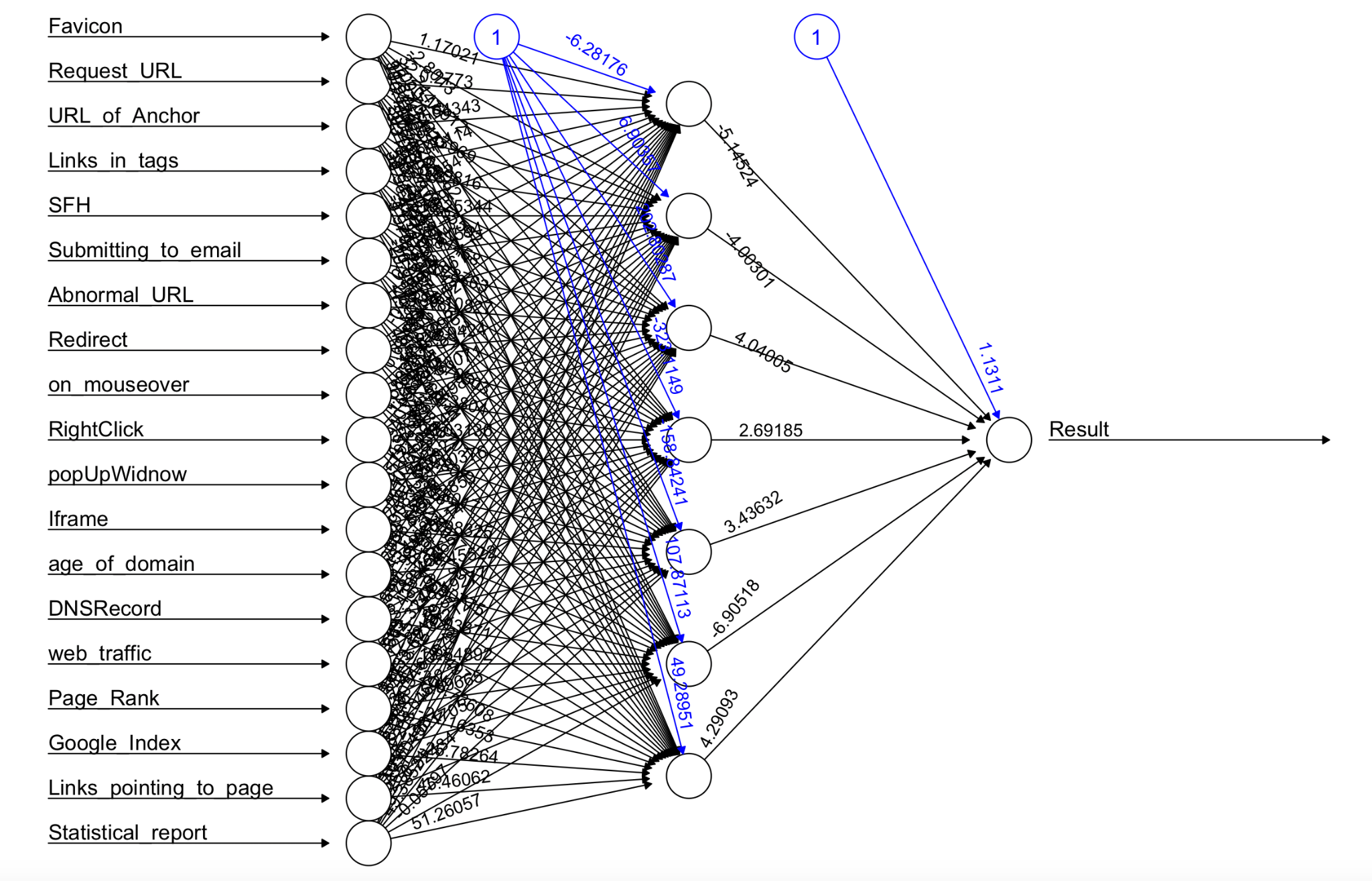
The startweights of the neural network model with the balanced dataset:



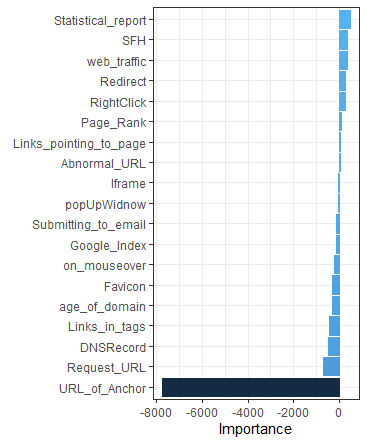
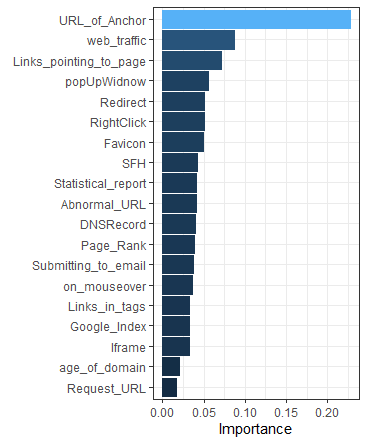
The final optimised weights of the neural network model with the balanced dataset:



The final Neural Network diagram using the balanced dataset:



Using Garson() and Olden() on the model developed with the balanced dataset to calculate the feature importance. We can see that the URL\_of\_Anchor is the most significant feature with the highest relative importance in determining whether the website is legitimate.



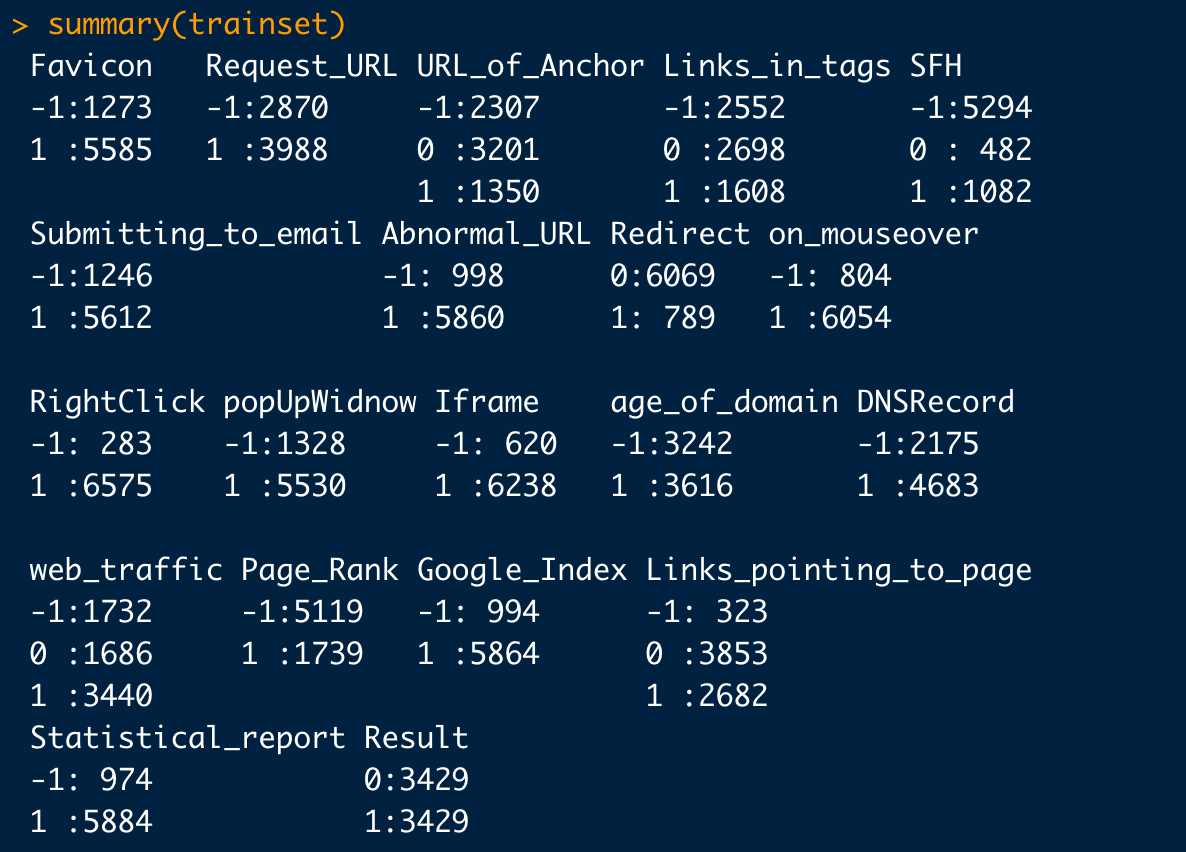
**Performance of** **Neural Network Model Balanced Test Datasets:**

|  |  |
| --- | --- |
| **Malicious Website Content Detection** | **Balanced Test Dataset** |
| Confusion Matrix |  |
| Overall Accuracy | 0.9041013 |
| False Positive Rate | 0.0854516 |
| False Negative Rate | 0.1077028 |
| Precision | 0.8922972 |
| Recall | 0.8910824 |

**MARS Models**

MARS Model with Balanced Dataset

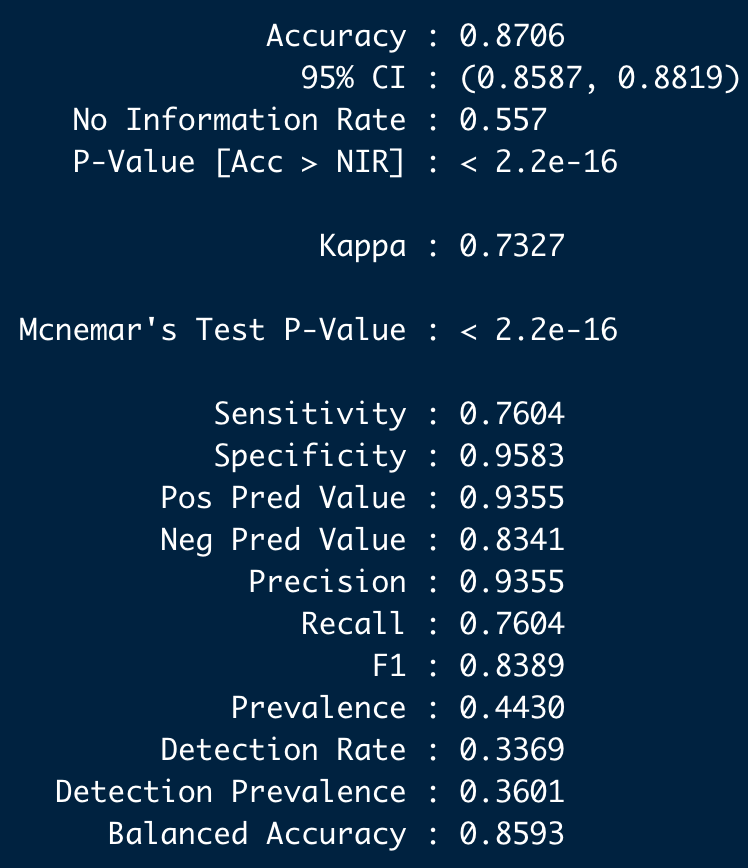
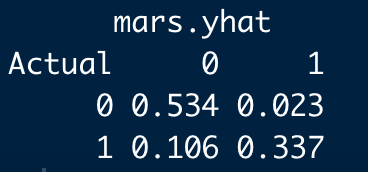
After performing a 70-30 train-test split and balancing the trainset by downsampling the majority class, the balanced trainset used to train the MARS model for malicious websites has 3429 samples for both phishing and legitimate website content.

****

We fit a MARS model to the training dataset and put the trained model to predict the testset data.

****

The probability threshold for classification is set to 0.5. A probability value that is higher than the threshold value will be deemed as phishing. Nex, we create a confusion matrix and find the respective performance metrics.

****

Testset Accuracy: 87.06%, FNR: 23.93%, FPR: 4.13%, Precision: 93.55%, Recall: 76.04%.

**Performance of** **MARS Model Balanced Test Datasets:**

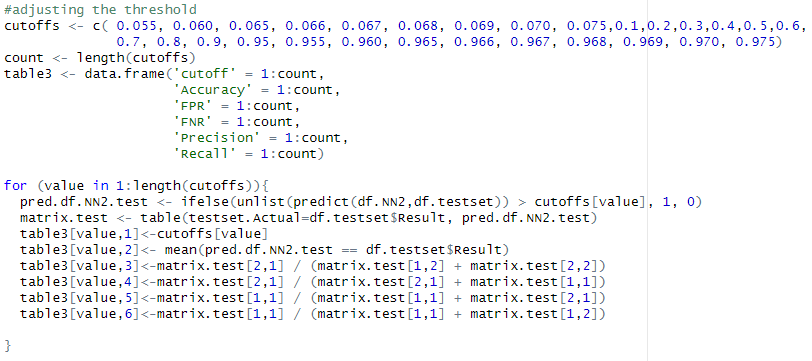
|  |  |
| --- | --- |
| **Malicious Website Content Detection** | **Balanced Test Dataset** |
| Confusion Matrix |  |
| Overall Accuracy | 0.8706 |
| False Positive Rate | 0.0412926 |
| False Negative Rate | 0.2392777 |
| Precision | 0.9355 |
| Recall | 0.7604 |

# 

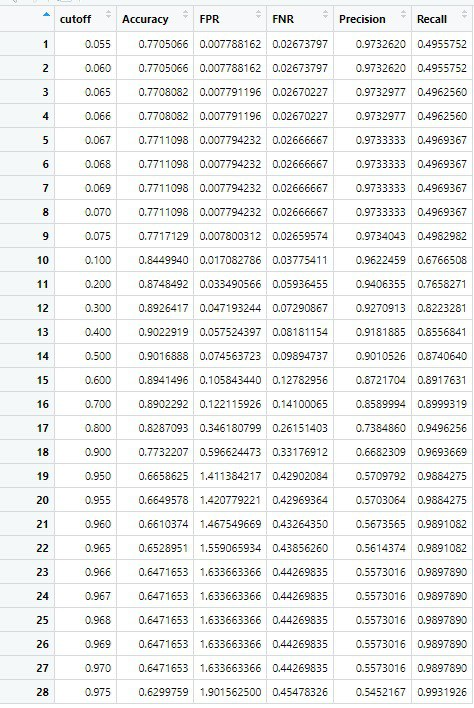
# 

# Appendix F - Finding Optimal Threshold For Neural Network

Utilising a for loop, we develop multiple models with different threshold values, after which the model metrics are stored and displayed.



After looping through all threshold values, we analyse the following table of performance metrics for each cutoff and choose a cutoff value that minimises our false negative rate while achieving an acceptable accuracy.



0.1 was chosen as the optimal threshold as we believe that it provides the best balance between accuracy and low false negative rate, as well as achieving good results for the other metrics.