I'M HACKING YOUR SAFETY!

Implementazione di un modello per la classificazione dei siti internet

Relazione progetto finale per il corso di Modelli per Dati Categoriali
Tenuto dalla Prof.ssa Giordano Sabrina

Studente MICIELI OTTAVIO
Matricola 214209
Anno accademico 2019/2020

Lo sviluppo di tecnologie telematiche e informatiche, sempre più sofisticate ed efficienti, ha favorito l'insorgere di nuove forme d'illecito diffuse nella nostra società attraverso l'impiego di strumenti di utilizzo comune, quali computer, tablet, cellulari, smart tv, ecc.

Il **cybercrimine** è la minaccia invisibile che sta cambiando il mondo. Quando si parla di cybercrimine (detto anche crimine informatico) ci si riferisce generalmente a una attività criminosa caratterizzata dall'abuso di componenti informatiche, sia hardware che software.

Il sistema del World Wide Web agevola, infatti, la commissione di determinati crimini. Questo perché, i reati informatici sono denotati da peculiari caratteristiche che li differenziano rispetto ai reati, per così dire, offline. Esse sono quelle della de-localizzazione delle risorse, quella della de-temporizzazione delle attività, nonché quella della de-territorializzazione. In parole semplici, c'è un'estrema facilità con cui gli individui riescono a commettere tali reati da qualunque parte del mondo, senza un necessario collegamento fisico fra l'utente e il sistema.

Nel frattempo, il mercato della lotta al cybercrimine italiano cerca di tenere il passo dei vari attacchi hacker che si susseguono. A testimoniare ciò, basta analizzare il mercato della Cyber Security in Italia: un valore complessivo pari a **1,19 miliardi** di euro, con il 75% di questa cifra appartenente alle grandi imprese. Succede, quindi, che la poca sicurezza informatica delle piccole e medie imprese aumenti esponenzialmente il rischio di essere bersagli di un attacco informatico.

Stando ai dati di Fastweb, che partecipa ogni anno al rapporto Clusit, l'Associazione Italiana per la sicurezza informatica, fornendo informazioni relative agli attacchi rilevati dal Security Operations Center, **ogni cinque minuti** un utente italiano viene colpito da un evento di attacco informatico. Per tale ragione, la privacy di ciascuno di noi è messa costantemente a dura prova.

Nella pratica, in cosa consiste un attacco informatico?

La scheda riassuntiva sottostante mostra quali possano essere le principali tipologie di cybercrime.

I PRINCIPALI CRIMINI INFORMATICI

ATTACCHI DDOS Obiettivo è rendere invisibile un servizio online e collassarne la banda dei server sovraccaricando il sito. Una volta collassata la rete, l'hacker può accedere al sistema.

PUP I Programmi Potenzialmente non desiderati sono tipi di malware che installano software non richiesti nel sistema informatico della vittima. Tali programmi possono contenere spyware o adware.

BOTNET Sono reti di computer infetti, controllate in remoto da hacker, che le utilizzano per inviare spam o attaccare altri computer attraverso malware.

PHISHING attività criminose di hacker che inviano e-mail con allegati o link fraudolenti per ingannare gli utenti e ottenere accesso ai loro account o dispositivi informatici.

FURTI D'IDENTITÀ Consistono nell'accesso da parte di un criminale ai dati personali di un utente per rubare fondi, informazioni riservate o prendere parte a frodi fiscali o assicurative.

CONTENUTI VIETATI O ILLEGALI
Si basa sulla distribuzione e condivisione di contenuti inappropriati, considerati pericolosi e offensivi, come per esempio la pornografia, scene molto violente o di attività criminali.

STALKING INFORMATICO Il criminale sottopone a una serie di molestie online la vittima, mediante l'inoltro di moltissimi messaggi ed e-mail, utilizzando soprattutto i social network.

EXPLOIT KIT Questi ricercano un punto debole del sistema, come ad esempio un bug nel codice di un software, per ottenere il controllo del computer di un utente.

INGEGRERIA SOCIALE Il criminale contatta direttamente la vittima, per telefono o via e-mail. Cerca di guadagnarsi la fiducia dell'utente, spesso spacciandosi per un agente dell'assistenza al cliente di un'azienda, per ottenere informazioni personali.

TRUFFE ONLINE Spesso si tratta di annunci o e-mail di spam contenenti false promesse di premi o offerte di denaro spropositate. Ma dietro questo si nascondono malware che bloccano o rubano informazioni.

Tuttavia, le tecniche e i mezzi utilizzati per difendersi dagli attacchi dei criminali informatici non sono sufficienti, e i dati parlano chiaro.
Secondo le informazioni in possesso alla polizia postale, l'ente preposto al contrasto delle frodi postali e del crimine informatico, nel 2019 ci sono stati **4930 casi** di financial cybercrime. È stato registrato un aumento degli episodi di phishing, tecnica che è stata attuata soprattutto attraverso malware e siti web clonati, ma sono cresciuti anche i casi di *smishing* (tentativi di phishing veicolati attraverso gli sms) e *vishing* (phishing che fa uso dei servizi di telefonia vocale, quindi attraverso l'uso di telefonate).
Anche le truffe online sono in aumento. Sempre secondo i dati provenienti dal rapporto della polizia postale, nel 2019 sono pervenute **196mila segnalazioni**, specialmente in ambito di e-commerce e di trading online.

Perché si ha un aumento così considerevole di tali fenomeni?

Perché in primo luogo è aumentata sostanzialmente la popolazione soggetta a tali rischi, in quanto ognuno di noi è costantemente connesso a Internet. Inoltre, le protezioni tradizionali, quali antivirus e firewall, non risultano essere più sufficienti a bloccare le minacce, che sono sempre più sofisticate e sfuggono alla maggior parte dei sistemi di controllo. In aggiunta, nessuna piattaforma oggigiorno è immune alle minacce informatiche. Fino ad un paio di anni fa, le minacce si concentravano principalmente sui prodotti Microsoft, data la loro vastissima diffusione sia in ambito enterprise che nel settore privato. Oggi gli attacchi informatici avvengono con crescente frequenza anche verso piattaforme meno diffuse, come McOS o Linux.

Anche l'uso costante, ma soprattutto inconsapevole, dei social network non aiuta la diminuzione di tale fenomeno. La prepotente affermazione dei social non ha, infatti, coinciso né con una presa di coscienza da parte degli utenti dei rischi che si possono correre, né con l'adozione di particolari forme di protezione da parte delle piattaforme, per esempio applicando sistemi di autenticazione forte all'accesso, o monitorando i propri network per bloccare le minacce alla fonte.
Non è infatti raro sentire qualcuno lamentarsi del fatto che abbia subito un furto delle credenziali di accesso. Inoltre, i social network risultano essere il veicolo principale per attacchi di stalking, cyber bullismo, spionaggio, furti di identità e perdita dei dati personali.

Attenzione ai link

Gli hacker, solitamente, hanno bisogno della collaborazione involontaria delle vittime. Una delle tecniche di inganno più comuni, tra quelle impiegate dai cybercriminali, sono i link malevoli, che se aperti possono indirizzare il traffico degli utenti su siti pericolosi oppure scaricare direttamente sulle macchine dei malcapitati software dannosi.
Ad esempio, la maggior parte dei metodi di phishing usa degli exploit tecnici per far apparire i link malevoli nelle mail, alla stregua di quelli autentici. Altri trucchetti diffusi sono utilizzare URL scritti male, oppure usando dei sottodomini. Ad esempio, http://www.tuabanca.it.esempio.com/ può sembrare a prima vista un sito legittimo, ma in realtà sta puntando a un sottodominio di un altro sito. In questo tipo di truffe la veste grafica del sito fraudolento è assolutamente identica a quella del sito originale, pertanto risulta molto facile cadere in errore.
Insomma, dietro ad un link apparentemente innocuo possono nascondersi pericoli di ogni genere.

Verificare se un link è sicuro prima di cliccarci di sopra è quindi di fondamentale importanza, soprattutto se si nutrissero dei dubbi su di esso in termini di provenienza e di veridicità.
In prima istanza, se ci sposteremo con il cursore del mouse sul link ci apparirà quello che sarà l'URL di destinazione. Se esso non ha nulla a che vedere con il test del link e soprattutto con il contenuto del messaggio in cui è presente tale link, c'è rischio che si tratti di un sito pericoloso. Inoltre, il nome a dominio che precede, ad esempio, .com, .it, .net, ecc., è quello che fa fede nei link. Non bisogna quindi farsi ingannare da eventuali domini di livello inferiore, come prevedeva l'esempio nella scheda superiore. Ovviamente prima di cliccare un link di cui non siamo sicuri, possiamo ispezionarlo attraverso l'utilizzo di servizi online di scansione. Ma ci sono alcuni accorgimenti che si possono utilizzare per riconoscere la tipologia di sito cui abbiamo di fronte?

La definizione del problema

Il lavoro di analisi svolto mira ad individuare un modello attraverso il quale poter predire se un sito sia pericoloso o meno. Esso farà utilizzo di variabili legate propriamente al sito web, oppure al server che lo ospita, e che potrebbero far aumentare o diminuire la propensione di un determinato sito web ad essere pericoloso o meno.
L'obiettivo finale è quello di creare un semplice, quanto intuitivo, modello di machine learning per la detezione della tipologia di pagina web, che potrà essere successivamente implementato e arricchito attraverso l'aggiunta di ulteriori variabili.

Descrizione del dataset

L'intero lavoro di analisi fa utilizzo delle informazioni presenti all'interno del dataset "Malicious and Benign Website", caricato online sul sito Kaggle.com da Christian Urcuqui, assistente presso l'Università Icesi in Colombia, e appassionato di cybersecurity e data science.

Tale dataset venne utilizzato per la stesura dell'articolo dal titolo "*Machine Learnig Classifiers to Detect Malicious Websites*", nel quale Urcuqui ed altri valutavano la capacità di classificazione di quattro algoritmi di machine learning dei siti web a seconda della loro tipologia.
Il file contenete i dati può essere scaricato dal seguente link (che assicuro non pericoloso!): https://www.kaggle.com/xwolf12/malicious-and-benign-websites.
Come da lui affermato, al fine di ottenere i dati sono stati utilizzati differenti risorse verificate contenenti liste sia di URL sicuri che pericolosi, insieme ad alcuni strumenti addizionali, utilizzati per estrarre altre informazioni di interesse, non visualizzabili normalmente.

Procediamo con la lettura del dataset da file csv, isoliamo le variabili di interesse e visualizziamo le prime dieci osservazioni.

	URL_LENGTH	NUMBER_SPECIAL_CHARACTERS	CHARSET	SERVER	WHOIS_COUNTRY	WHOIS_REGDATE	APP_BYTES	Type
M0_109	16	7	iso-8859-1	nginx	<NA>	10/10/2015 18:21	700	1
B0_2314	16	6	UTF-8	Apache/2.4.10	<NA>	<NA>	1230	0
B0_911	16	6	us-ascii	Microsoft-HTTPAPI/2.0	<NA>	<NA>	0	0
B0_113	17	6	ISO-8859-1	nginx	US	7/10/1997 4:00	3812	0
B0_403	17	6	UTF-8	<NA>	US	12/05/1996 0:00	4278	0
B0_2064	18	7	UTF-8	nginx	SC	3/08/2016 14:30	894	0
B0_462	18	6	iso-8859-1	Apache/2	US	29/07/2002 0:00	1189	0
B0_1128	19	6	us-ascii	Microsoft-HTTPAPI/2.0	US	18/03/1997 0:00	0	0
M2_17	20	5	utf-8	nginx/1.10.1	<NA>	8/11/2014 7:41	0	1
M3_75	20	5	utf-8	nginx/1.10.1	<NA>	8/11/2014 7:41	0	1

Figura 1: Le prime dieci osservazioni del dataset

Trattamento e ricodifica delle variabili

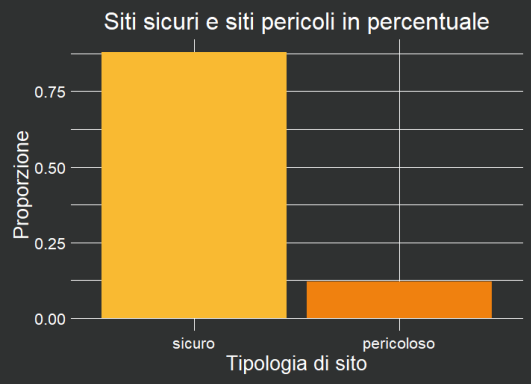
Per quanto riguarda le variabili di tipo quantitativo, quali URL_LENGTH, NUMBER_SPECIAL_CHARACTERS e APP_BYTES, non viene apportata alcuna modifica. In tali variabili, inoltre, non vi sono valori mancanti.
Per quanto riguarda la variabile CHARSET vi sono alcuni valori mancanti, che vengono catalogati come *"Other"*, insieme ad altre categorie con poche unità al loro interno. Vengono anche omogeneizzate le etichette delle categorie di tale variabile.
Nell'analisi del carattere statistico SERVER in prima istanza si sono raggruppate tutte quelle categorie dovute al solo fatto che si analizzavano software appartenenti sempre alla medesima software house ma con versioni differenti. Per quanto riguarda invece i valori mancanti in questa categoria, poiché abbastanza numerosi, sono stati etichettati come *"Unknow"*, mentre tutti i server minori, e quindi con poche unità statistiche, sono stati etichettati come *"Other"*.
Per quanto riguarda la variabile che indica il paese di provenienza del sito internet, tale informazione era fornita attraverso le sigle delle nazioni. Per prima cosa tali sigle sono state ricodificate con il nome per esteso della nazione, mentre i valori mancanti sono stati classificati come *"Unknow"*. Si hanno però troppe categorie per cui si preferisce raggruppare le nazioni per continenti di appartenenza, unendo i continenti Africa, Asia e Oceania in quanto vi sono poche osservazioni in essi.
Per la variabile che informa sulla data di registrazione del server, WHOIS_REGDATE, se tiene in considerazione solo dell'anno di attivazione del server e successivamente si categorizza, dividendo per decadi. Anche qui vi sono valori mancanti, classificati come *"Unknow"*.

Descrizione delle variabili del dataset

Tipologia	Descrizione
	URL-LENGTH
Quantitativa discreta	Tale variabile indica il numero di caratteri complessi contenuti nell'URL del sito in analisi.
	NUMBER_SPECIAL_CHARACTERS
Quantitativa discreta	Tale variabile indica il numero di caratteri speciali presenti all'interno dell'URL. Sono esempi di caratteri speciali / % & # . = ecc.
	CHARSET
Qualitativa multinomiale	Tale variabile indica la tipologia di codifica dei caratteri utilizzata all'interno della pagina web. Una codifica non è altro che un codice che associa un insieme di caratteri (ossia tutti quei grafemi normalmente utilizzati per la comunicazione) ad un insieme di altri oggetti quali numeri, specialmente nell'ambito informatico, oppure pulsazioni elettriche, con lo scopo di facilitare la memorizzazione di un testo in un computer o la sua trasmissione attraverso una rete di telecomunicazione. Esempi comuni sono il Codice Morse e la codifica ASCII.
	SERVER
Qualitativa multinomiale	Tale variabile indica il server utilizzato per la gestione del sito internet, anche definito come *server web*. Esso gestisce tutte le richieste di trasferimento di pagine web che arrivano dal client, che di norma è un web browser.
	WHOIS_COUNTRY
Qualitativa multinomiale	Tale variabile indica la nazione di provenienza del sito web richiesto dal client.
	WHOIS_REGDATE
Qualitativa multinomiale	Tale variabile indica la data in cui il server è stato inizializzato nel World Wide Web.
	APP_BYTES
Quantitativa discreta	Indica il numero di pacchetti IP che vengono generati durante la comunicazione tra server e client. Tali pacchetti vengono utilizzati dal protocollo IP per la trasmissione dei dati, utile all'accesso a alla visualizzazione del sito. è una variabile numerica che non richiede alcuna codifica.
	Type
Qualitativa dicotomica	Indica se il sito è pericolo oppure sicure. Se tale variabile assume il valore 1 per un determinato sito allora tale sarà un sito web pericoloso.

Analisi univariata delle variabili e in relazione con la variabile risposta

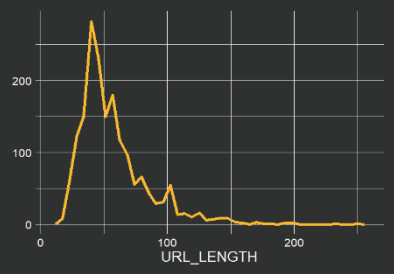
Vengono ora proposte una serie di informazioni sia a livello grafico, che sotto forma di indice statistico, in relazione alle variabili utilizzate nell'analisi, al fine di familiarizzare maggiormente con l'argomento e iniziare a produrre le prime ipotesi sul problema in esame.

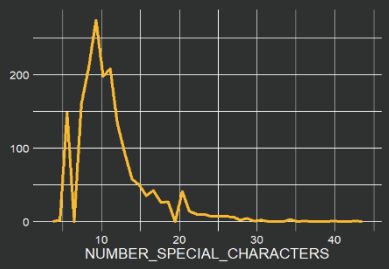


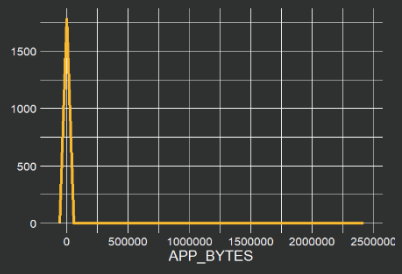
Poniamo inizialmente la nostra attenzione su quella che poi sarà la nostra variabile risposta, ossia la tipologia di sito internet, indicata nel dataset come *Type*.
Si può osserva, anche dal grafico a barre posto lateralmente, come nel campione vi sia una percentuale nettamente maggiore di siti internet sicuri (che rappresentano circa l'88% del totale delle unità statistiche) rispetto ai siti ritenuti pericolosi (che rappresentano, ovviamente, il restante 12% della numerosità campionaria).

Analisi descrittiva dei regressori quantitativi

	URL_LENGTH	NUMBER_SPECIAL_CHARACTERS	APP_BYTES
MINIMO	16,00	5,00	0,00
1° QUARTILE	39,00	8,00	0,00
MEDIANA	49,00	10,00	672,00
3° QUARTILE	68,00	11,11	2328,00
MASSIMO	249,00	13,00	2362906,00
MEDIA	56,96	43,00	2982,00
DEVIAZIONE STANDARD	27,55559	4,549896	56050.57

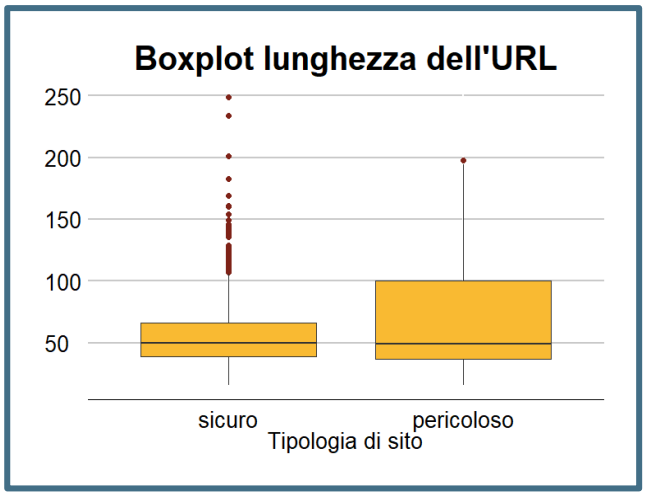




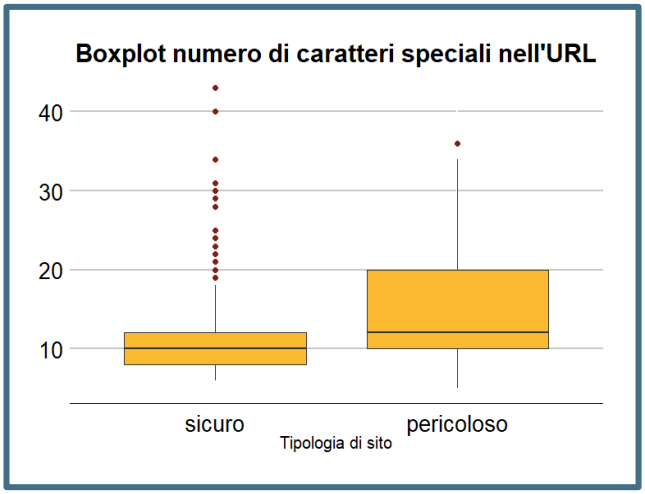


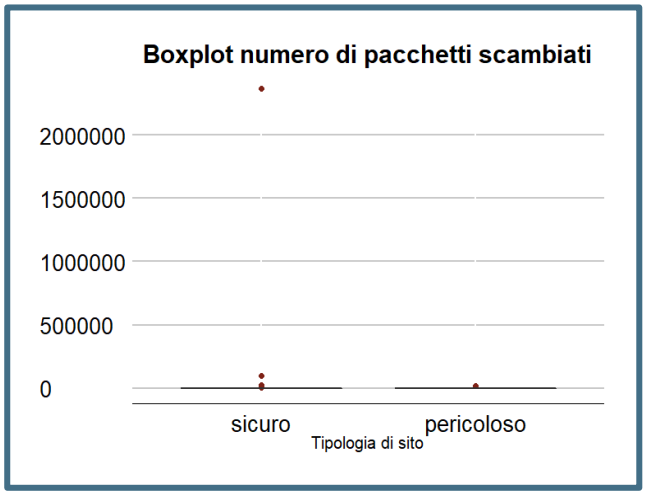
Come si può notare facilmente dai grafici sopra riportati, le tre variabili quantitative presentano tutte quante una distribuzione asimmetrica, con una coda pesante a destra. Tale situazione è indice di valori singolari verso le quantità più grandi, che potrebbero anche essere considerati come outlier. Lo si può anche notare osservando gli indici di posizione: in tutte e tre le variabili vi è una distanza relativamente molto più consistente tra il terzo quartile e il massimo della distribuzione rispetto alla stessa distanza tra minimo e primo quartile.
Una attenzione particolare deve essere rivolta verso la variabile APP_BYTES. In essa vi è, infatti, una forte presenza di valori 0, precisamente 657 unità statistiche presentano tale valore in questo carattere statistico, ossia quasi il 37% della numerosità totale del campione. Questo significa che in molti casi non c'è stato nessun invio di dati tra il server e il client utilizzato per recuperare i dati dei siti, molto probabilmente poiché non vi è stato bisogno di accedervi per rintracciare le informazioni necessarie.

Per analizzare la relazione che sussiste tra le variabili quantitative e la tipologia di sito vengono proposti tali boxplot, che suddividono le pagine web in sicure e pericolose.
Per quanto riguarda la variabile URL_LENGTH, si può subito notare come i valori mediani di entrambi i box siano più o meno sullo stesso livello, ma che la distanza tra tale valore e il terzo quartile, ossia la parte alta della "scatola", sia maggiore nei siti pericolosi rispetto ai siti sicuri. Questo ci porta ad affermare che, generalmente, i siti pericolosi siano caratterizzati da una lunghezza maggiore dell'URL rispetto ai siti protetti. Bisogna anche prestare attenzione ai valori che il boxplot individua come anomali, che nei siti sicuri sono in quantità superiore rispetto ai siti web rischiosi.
Molto probabilmente si tratta di particolari pagine web il cui accesso deriva da altre pagine sovrastanti, e per tal ragione dovrebbero essere prive di rischi, proprio perché difficilmente accessibili.



Lo stesso ragionamento si può applicare alla variabile NUMBER_SPECIAL_CHARACTERS, anche se qui il valore mediano nei due sottogruppi è leggermente differenti. Anche in questo caso i siti rischiosi tendono ad avere un maggior numero di caratteri speciali, molto probabilmente utilizzati per confondere l'utente che legge l'URL e non riuscire a comprendere quale sia il vero indirizzamento. Anche qui vi sono alcuni valori anomali per quanto riguarda i siti sicuri, molto probabilmente per lo stesso ragionamento fatto in precedenza.



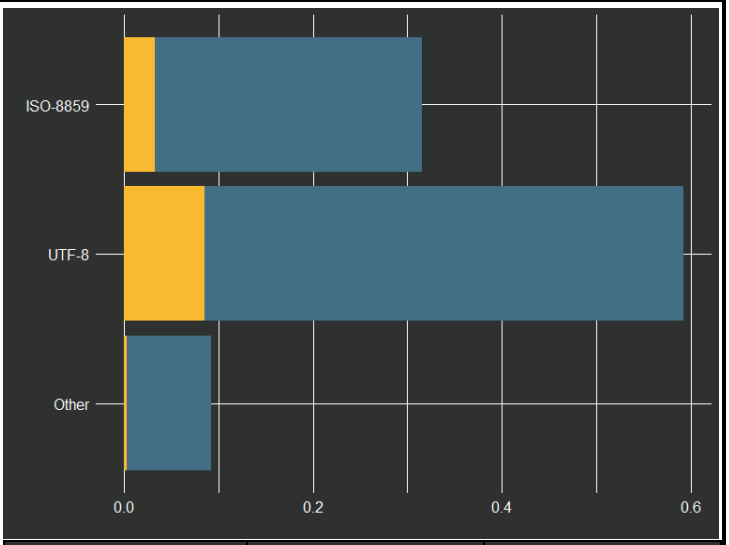


Come si è già potuto notare attraverso le misure di posizione, per quanto riguarda la variabile APP_BYTES, essa presenta valori molto schiacciati intorno allo 0. Vi sono comunque dei casi molto particolari nei siti sicuri, in cui il valore risulta essere anomalo.
Si è preferito comunque non trattare tali valori, in quanto potrebbero proprio essere degli indicatori attendibili: spesso, infatti, un sito internet sicuro contiene in sé alcuni certificati di sicurezza che ne validano tale condizione. La ricezione di tali certificati, composti ovviamente da pacchetti di dati, ne aumentano sensibilmente il numero di bytes scambiati tra client e server.

Analisi delle variabili qualitative

Variabile CHARSET

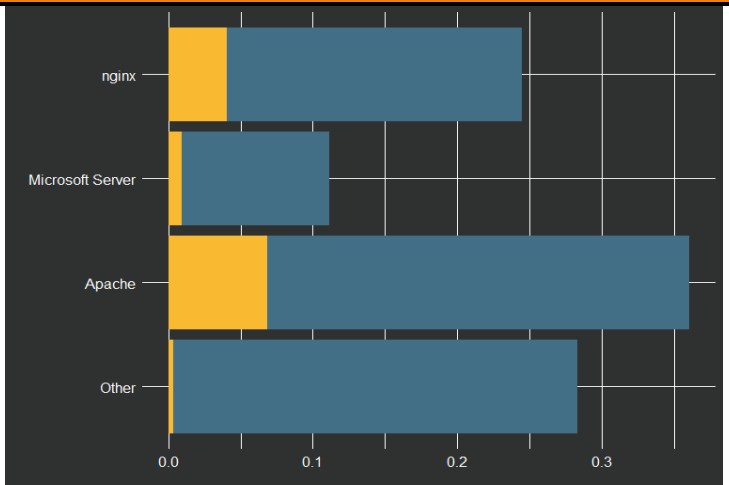
La prima variabile a carattere qualitativo che si vuole analizzare è CHARSET. Inizialmente, a seguito della semplice ricodifica descritta precedentemente, tale carattere statistico comprendeva quattro differenti categorie (*ISO-8859*, *UTF-8*, *us-ascii*, *Other*). Una di esse (*us-ascii)* è stata incorporata alla categoria *Other* poiché conteneva poche unità statistiche, non consentendone un'adeguata analisi.



	Pericoloso	Sicuro
ISO-8859	58	504
UTF-8	152	903
Other	6	158

Come si può notare dal bar plot a destra, la maggior codifica utilizzata dalle pagine web presenti nel data set è *UTF-8*, la stessa categoria che detiene la proporzione maggiore di siti web pericolosi (il 14,4% circa rispetto al totale dei siti nella medesima categoria).
Entrambi i test per la verifica dell'indipendenza effettuati su tale variabile in relazione alla tipologia di pagina web mostrano che vi è una certa dipendenza tra i due caratteri statistici. Tale dipendenza verrà indagata a seguito nella formulazione del modello, ma già si può affermare che l'essere un sito web codificato con *UTF-8* rispetto ad altro fa aumentare la propensione ad essere un sito pericoloso.

Variabile SERVER

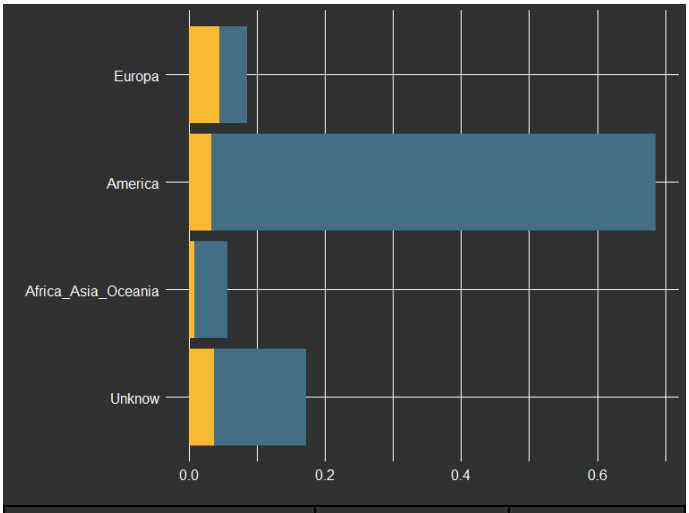


	Pericoloso	Sicuro
Nginx	72	364
Microsoft Server	16	182
Apache	122	521
Other	6	498

Anche per quanto riguarda la variabile SERVER, inizialmente dopo la codifica presentava un maggior numero di categorie, che sono state accorpate in quanto non presentavano unità statistiche in esse, in particolar modo siti di tipologia pericolosa.
Si può osservare che vi è un'alta percentuale di siti gestiti da altri server, di cui quasi la totalità di essi sono sicuri. La più elevata percentuale di siti pericolosi deriva dal web server *Apache*, che risulta essere anche il più utilizzato. Di 643 pagine web gestite, il 19% circa sono siti non sicuri.
La propensione per il server Apache di gestire un sito pericolo è ben 19 volte e mezzo la stessa propensione per un altro server.
Questa situazione si ha, molto probabilmente, poiché Apache è la piattaforma server Web più diffusa ed in grado di operare su una grande varietà di sistemi operativi.
Per quanto riguarda la dipendenza tra la tipologia di sito web, essa è verificata essere presente da entrambe le statistiche test utilizzate.

Variabile Continente

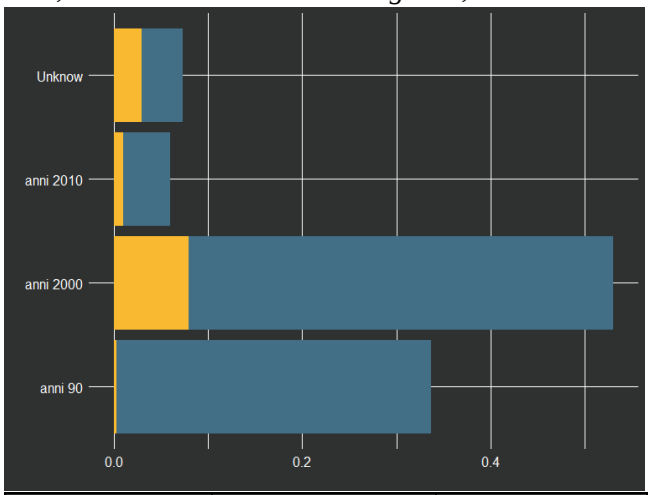
Per quanto riguarda la variabile continente, artificialmente creata dal carattere statistico indicante la Nazione di provenienza del sito web, essa ci mostra come la maggior parte dei siti web provenga dall'America, precisamente il 68% circa.
Per quanto riguarda la propensione ad avere siti pericolosi, l'America non è il continente peggiore. In effetti, l'Europa possiede una propensione ad avere siti pericolosi maggiore del 20,39% rispetto alla propensione ad avere siti pericolosi dell'America (con un minimo del 20% ed un massimo del 20,8%). Questo significa che c'è un'alta probabilità che navigando su siti europei si può incappare in qualche pericolo.
Indagando in profondità, si può osservare che tale percentuale elevata di siti pericolosi per l'Europa deriva dalla Spagna, che presenta 62 siti pericolosi contro uno solo sicuro.
Comunque, applicando i test per l'indipendenza si ha che ci indicano una certa relazione tra il continente di provenienza del sito e la sua tipologia.



	Perico-loso	Sicuro
Europa	79	74
America	58	1164
Africa Asia Oceania	14	86
Unknow	65	241

Variabile RegDate

Infine, analizziamo la variabile *RegDate*, che ci indica la data di registrazione del server. Essa viene utilizzata come indicatore di sicurezza del server web: si ipotizza, infatti, che i server messi in rete negli anni '90 abbiano delle patch di sicurezza poco efficace sulle nuove minacce cibernetiche sorte in quest'ultimi anni rispetto ai server di nuova data. È vero anche che i server vengono aggiornati, ma spesso si preferisce immettere sul mercato nuove versioni del sistema. I dati, però, non confermano del tutto tale ipotesi, poiché la percentuale di siti pericolosi rispetto al totale gestiti da server degli anni '90 è bassa (non raggiunge nemmeno l'1%). In questo caso la percentuale maggiore di siti pericolosi sul totale per quella categoria è presente per i server di cui non se ne conosce la data di registrazione, anche se guardando le 4 categorie, la maggior percentuale di siti pericolosi appartengo ai server degli anni 2000.
Anche in questo caso, comunque, la relazione presente tra tale variabile e la tipologia di sito internet è confermata dal p-value delle statistiche test.



	Pericoloso	Sicuro
Unknow	52	78
Anni 2010	18	88
Anni 2000	141	804
Anni '90	5	595

Costruzione del modello per la previsione della tipologia di sito

Questa fase dell'analisi che si sta conducendo verte sulla costruzione del modello previsionale che successivamente potrà essere applicato, in eventuali lavori futuri, in un algoritmo di machine learning per la classificazione della tipologia di sito.
Per la costruzione di tale modellistica è stata implementata la tecnica di *Forward selection*, ossia vengo aggiunti termini in qualità di regressori in sequenza e ne viene testata la significatività all'interno del modello. Se essa risulta esserci, allora la variabile viene mantenuta all'interno del predittore lineare, altrimenti viene eliminata.

Prima fase

Inizialmente il modello prevede come variabile risposta la tipologia di sito web, indicata dalla variabile *Type*, mentre avevano ruolo di regressore il numero di caratteri presenti all'interno dell'URl e il numero di caratteri speciali, indicati rispettivamente dalle variabili *URL_LENGTH* e *NUMBER_SPECIAL_CHARACTERS*.

```
Type ~ URL_LENGTH + NUMBER_SPECIAL_CHARACTERS
```

Entrambi i coefficienti risultavano essere statisticamente significativi, ma insospettiva il fatto che senza il numero di caratteri speciali, il numero di caratteri totali all'interno dell'URL avevano un effetto positivo sulla propensione ad essere un sito pericoloso (un URL con un carattere in più aveva maggior probabilità di indicizzare ad un sito pericoloso rispetto ad un URL con un carattere in meno), mentre inserendo come regressore la variabile che indica il numero di caratteri speciali, il coefficiente associato a *URL_LENGTH* diventava negativo (quindi adesso un sito internet con un URL con un carattere in più aveva minor probabilità di essere un sito pericoloso rispetto ad un sito web con URL di un carattere in meno).
Si è quindi indagato sul grado di correlazione tra queste due variabili, che risulta essere elevato, con un indice di correlazione di Pearson pari a circa 0,92. Applicando i test per la multicollinearità nel modello, si scopre, effettivamente che essa è presenta, quindi un regressore spiega anche l'altro, in quanto sono linearmente dipendenti tra di loro. Questo comporta problemi a livello dell'inferenza che si può applicare su tale modello.
Si decide, quindi, di modificare la variabile indicante il numero di caratteri speciali presenti all'interno dell'URL, rapportandoli al numero di caratteri totali.
Utilizzando tale carattere statistico, indicato con *prop.spec*, al posto di *NUMBER_SPECIAL_CHARACTERS*, innanzitutto si ha che tale regressore risulta essere statisticamente significativo, successivamente viene risolto il problema di multicollinearità presente all'interno del modello. Inoltre, vi è un miglioramento generale della bontà di adattamento del modello, anche se quest'ultima risulta essere scarsa.

Seconda fase

Al modello precedente, si è aggiunta l'informazione riguardante il continente di provenienza del sito web.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1872  -0.3716  -0.2272  -0.1452   2.9916

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -12.240227   0.855198 -14.313  < 2e-16 ***
URL_LENGTH                      0.037561   0.003548  10.586  < 2e-16 ***
prop.spec                      38.600811   2.945734  13.104  < 2e-16 ***
ContinenteAfrica_Asia_Oceania  -0.120975   0.386914  -0.313    0.755
ContinenteAmerica              -1.153297   0.229459  -5.026 5.00e-07 ***
ContinenteEuropa                1.433461   0.270701   5.295 1.19e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1316.04  on 1780  degrees of freedom
Residual deviance:  823.12  on 1775  degrees of freedom
AIC: 835.12

Number of Fisher Scoring iterations: 6
```

Superiormente è possibile visualizzare l'output derivante dalla stima dei coefficienti appartenenti a tale modello, che ricordiamo essere

```
Type ~ URL_LENGTH + prop.spec + Continente
```

Si ha, tuttavia, che il coefficiente associato alla dummies inerente al macro-continente creato artificialmente dall'unione di Africa, Asia e Oceania, risulti essere statisticamente non significativo a spiegare la propensione ad essere un sito pericoloso. Questo significa che conoscere che un sito provenga da tale unione di continenti piuttosto che non avere informazioni a riguardo non cambia la propensione ad essere una pagina web potenzialmente pericolosa.

Analizzando però la significatività della variabile Continente, essa risulta essere statisticamente valida.

```
Analysis of Deviance Table

Model 1: Type ~ URL_LENGTH + prop.spec
Model 2: Type ~ URL_LENGTH + prop.spec + Continente
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1778     941.80
2      1775     823.12  3   118.69 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Si decide quindi di unire il macro-continente già formato con tutti quei siti di cui non sappiamo la provenienza geografica, rinominando tale categoria come *Other*.
Operando tale trasformazione si ha che i regressori risultano tutti essere statisticamente significativi, con un miglioramento della bontà di adattamento del modello rispetto a quello con solo il numero di caratteri nell'URL e la percentuale di caratteri speciali nello stesso. Si ha, quindi, che la percentuale di bontà di adattamento raggiunge il 37% circa, ancora un po' scarsa, per cui si inseriscono altri regressori.

Terza fase

In questa fase è stata aggiunta al modello la variabile *CHARSET*, che indica il tipo di codifica utilizzata all'interno della pagina web. Tale carattere statistico risulta non essere significativo, non apportando nuova informazione all'interno del modello, come viene confermato anche dal confronto tra il nuovo modello e quello privato di tale variabile.

```
Model 1: Type ~ URL_LENGTH + prop.spec + Continente
Model 2: Type ~ URL_LENGTH + prop.spec + Continente + CHARSET
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1776     823.21
2      1774     819.26  2   3.9491   0.1388
```

Tale variabile, pertanto, non viene più presa in considerazione.
Successivamente, vi è stata l'aggiunta delle variabili inerenti al server web che si occupa della gestione dei siti. Inizialmente entrambe le variabili, *SERVER* e *RegDate*, sono state inserite nel modello considerando unicamente i loro effetti principali.
Tale situazione ha dato degli esiti soddisfacenti, in quanto tutti i regressori risultano statisticamente significativi, tranne per una dummies inerente al continente di provenienza del sito web. Prima di trattare tale problema si è pensato di inserire una interazione tra il server e la sua data di registrazione sul World Wide Web. Si ipotizza, infatti, considerare le versioni più recenti di uno stesso server, rispetto a quelle più datate, avesse un qualche effetto sulla propensione dei propri siti serviti ad essere pericolosi.

```
Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -1.800e+01  1.831e+00  -9.835  < 2e-16 ***
URL_LENGTH                             4.178e-02  4.355e-03   9.595  < 2e-16 ***
prop.spec                              4.039e+01  3.653e+00  11.058  < 2e-16 ***
ContinenteAmerica                      2.971e-01  2.870e-01   1.035  0.30052
ContinenteEuropa                       2.120e+00  3.511e-01   6.036 1.58e-09 ***
SERVERApache                           9.025e-01  1.655e+00   0.545  0.58559
SERVERMicrosoft Server                -1.210e+01  7.136e+02  -0.017  0.98648
SERVERnginx                            1.880e+00  1.453e+00   1.294  0.19553
RegDateanni 2000                       1.986e+00  1.457e+00   1.363  0.17286
RegDateanni 2010                      -1.202e+01  2.025e+03  -0.006  0.99526
RegDateUnknow                          5.429e+00  1.562e+00   3.476  0.00051 ***
SERVERApache:RegDateanni 2000          2.196e+00  1.783e+00   1.232  0.21813
SERVERMicrosoft Server:RegDateanni 2000  1.548e+01  7.136e+02   0.022  0.98269
SERVERnginx:RegDateanni 2000           1.657e-02  1.618e+00   0.010  0.99183
SERVERApache:RegDateanni 2010          1.653e+01  2.025e+03   0.008  0.99349
SERVERMicrosoft Server:RegDateanni 2010  2.711e+01  2.147e+03   0.013  0.98992
SERVERnginx:RegDateanni 2010           1.636e+01  2.025e+03   0.008  0.99355
SERVERApache:RegDateUnknow            -1.809e+00  1.936e+00  -0.934  0.35010
SERVERMicrosoft Server:RegDateUnknow  -6.621e+00  1.511e+03  -0.004  0.99650
SERVERnginx:RegDateUnknow             -5.156e-01  1.735e+00  -0.297  0.76634
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Osservando unicamente i livelli di significatività dei livelli, si scopre che l'interazione porta un sostanziale peggioramento della significatività dei regressori. Non solo perdono di significato gli effetti principali di tali variabili, ma nemmeno l'interazione risulta essere statisticamente informativa.
Sorge, quindi, il dubbio che tra la variabile *Server* e la variabile *RegDate* vi sia un certo livello di dipendenza. Tale relazione viene effettivamente confermata, per cui, inserendo anche l'interazione vi è un problema di multicollinearità che precedentemente non esisteva.
Si decide, per tale motivo, di non tenere in considerazione l'interazione di queste due variabili, ma di mantenere nel modello unicamente i loro effetti principali.

L'ultima variabile inserita nel modello, *APP_BYTES*, si scopre essere statisticamente non significativa all'interno del modello, causando addirittura un peggioramento nella bontà di adattamento. Tale variabile non viene quindi ad essere inserita nel modello finale.

```
Call:
glm(formula = Type ~ URL_LENGTH + prop.spec + Continente + SERVER +
    RegDate + APP_BYTES, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7430  -0.3210  -0.1273  -0.0420   3.5079

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.855e+01  1.315e+00 -14.109  < 2e-16 ***
URL_LENGTH           4.166e-02  4.115e-03  10.123  < 2e-16 ***
prop.spec            3.999e+01  3.382e+00  11.826  < 2e-16 ***
ContinenteAmerica    1.876e-01  2.830e-01   0.663 0.507397
ContinenteEuropa     2.072e+00  3.369e-01   6.149 7.79e-10 ***
SERVERApache         2.287e+00  4.853e-01   4.713 2.44e-06 ***
SERVERMicrosoft Server  1.951e+00  5.627e-01   3.466 0.000527 ***
SERVERnginx          2.163e+00  4.919e-01   4.398 1.09e-05 ***
RegDateanni 2000     3.279e+00  5.325e-01   6.158 7.36e-10 ***
RegDateanni 2010     3.825e+00  6.173e-01   6.197 5.76e-10 ***
RegDateUnknow        4.221e+00  6.101e-01   6.918 4.57e-12 ***
APP_BYTES           -1.472e-06  6.268e-06  -0.235 0.814345
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1316.04  on 1780  degrees of freedom
Residual deviance:  679.48  on 1769  degrees of freedom
AIC: 703.48

Number of Fisher Scoring iterations: 8
```

Quarta fase: la costruzione del modello finale e la diagnostica

Termini, quindi, i regressori da poter inserire all'interno del modello, quello che risulta essere il miglior per quanto riguarda significatività dei regressori e bontà di adattamento risulta essere

```
Type ~ ULR_LENGTH + prop.spec + Continente + Server + RegDate
```

Dal suo computo otteniamo l'output riportato sottostante.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7437  -0.3222  -0.1275  -0.0419   3.5074

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -18.555402   1.315015 -14.110  < 2e-16 ***
URL_LENGTH                0.041703   0.004115  10.135  < 2e-16 ***
prop.spec                39.995128   3.382279  11.825  < 2e-16 ***
ContinenteAmerica         0.184209   0.282996   0.651 0.515096
ContinenteEuropa          2.072643   0.336870   6.153 7.62e-10 ***
SERVERApache              2.284229   0.485214   4.708 2.51e-06 ***
SERVERMicrosoft Server    1.951468   0.562742   3.468 0.000525 ***
SERVERnginx               2.163762   0.491901   4.399 1.09e-05 ***
RegDateanni 2000          3.278965   0.532760   6.155 7.52e-10 ***
RegDateanni 2010          3.827883   0.617505   6.199 5.68e-10 ***
RegDateUnknow             4.220132   0.610331   6.915 4.70e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1316.04  on 1780  degrees of freedom
Residual deviance:  679.71  on 1770  degrees of freedom
AIC: 701.71
```

Rimane da affrontare il problema riguardante la non significatività statistica del regressore collegato alla dummies che indica il continente America. Si potrebbe, come già fatto in precedenza, accorpare anche l'America all'insieme degli *Other*. Riprendendo, però, le considerazioni fatte in precedenza sulla nazione di provenienza dei siti, abbiamo potuto constatare come una percentuale cospicua di siti pericolosi per l'Europa provengono dalla Spagna.

Si sceglie, quindi, di inserire come variabile indicante il luogo geografico di provenienza del sito internet un carattere statistico binaria, che assumerà valore "Spagna" oppure "Other". Si vuole, quindi, mettere in confronto l'effetto che comporta il fatto che il sito provenga dalla Spagna piuttosto che da altre nazioni del mondo sulla propensione di tale sito ad essere pericoloso o meno. Il modello che si ottiene, insieme al suo output, viene proposto qui di fianco.

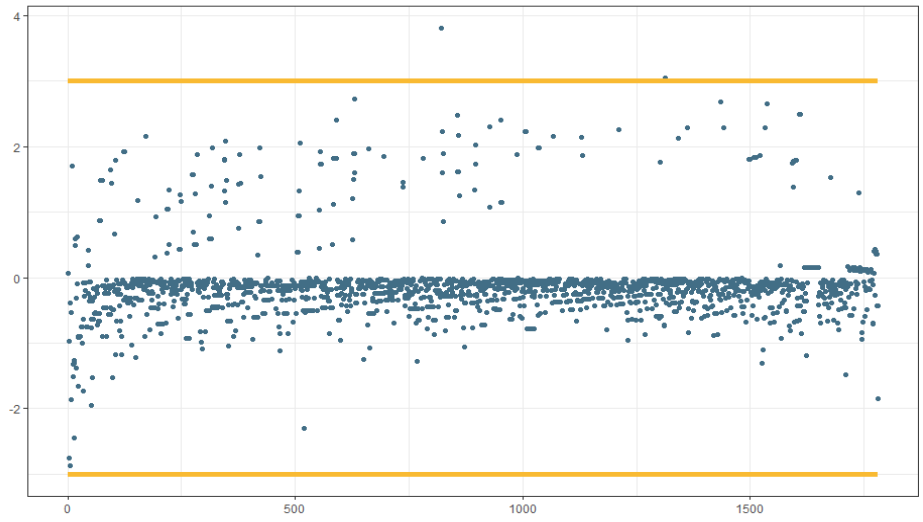
Nel modello, adesso sono presenti unicamente regressori statisticamente significativi. Prima di commentare tali risultati, poniamo il modello sotto alcuni test diagnostici per verificarne validità e adattamento.

```
Call:
glm(formula = Type ~ URL_LENGTH + prop.spec + SERVER + RegDate +
    IS_NATION, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8762  -0.3334  -0.1436  -0.0591   3.8089

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -16.676901   1.253444 -13.305  < 2e-16 ***
URL_LENGTH                0.028641   0.004819   5.943 2.79e-09 ***
prop.spec                38.634638   3.401251  11.359  < 2e-16 ***
SERVERApache              2.067017   0.464214   4.453 8.48e-06 ***
SERVERMicrosoft Server    2.061322   0.536171   3.845 0.000121 ***
SERVERnginx               2.044437   0.469728   4.352 1.35e-05 ***
RegDateanni 2000          2.776011   0.509339   5.450 5.03e-08 ***
RegDateanni 2010          3.212477   0.586579   5.477 4.33e-08 ***
RegDateUnknow             3.451874   0.541224   6.378 1.80e-10 ***
IS_NATIONSpagna           5.579515   1.078129   5.175 2.28e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1316.04  on 1780  degrees of freedom
Residual deviance:  644.48  on 1771  degrees of freedom
AIC: 664.48

Number of Fisher Scoring iterations: 7
```

La diagnostica del modello inizia verificandone il livello della bontà di adattamento. Secondo la statistica test dello pseudoR^2, il modello spiegherebbe un 51% della variabilità totale presente nel dataset. Si tratta di un valore non tanto soddisfacente, ma nemmeno demoralizzante. Diciamo che il modello riesce ad avere un'accettabile capacità predittiva, che poi andrà indagata ulteriormente con altri strumenti. Per il test di hosmerlemeshow vi è comunque una differenza statisticamente significativa tra i valori osservati e i valori stimati, per cui non si ha una precisione di distinzione tale da poter essere utilizzato come classificatore dei siti internet, ma data la semplicità di applicazione di tale modello, può essere utilizzato come punto di partenza.
Applicando il test che involve la statistica test della deviance, si può osservare nuovamente come aggiungendo di seguito i regressori, vi sia sempre una preferenza maggiore verso il modello con un regressore in più.

	Df	Deviance	Resid. Df	Resid. Dev	p_value
NULL	NA	NA	1780	1316.0447	NA
URL_LENGTH	1	39.70348	1779	1276.3413	2.95598e-10
prop.spec	1	334.53912	1778	941.8021	0.00000e+00
SERVER	3	89.26595	1775	852.5362	0.00000e+00
RegDate	3	121.95509	1772	730.5811	0.00000e+00
IS_NATION	1	86.10086	1771	644.4802	0.00000e+00

Per quanto riguarda le considerazioni che si possono fare sulla componente erratica del modello, già dall'output del modello possiamo notare come gli errori si distribuiscano in una banda che va da un valore di -3 a un valore di circa 3 (vi è giusto qualche punto al di fuori di tale banda).

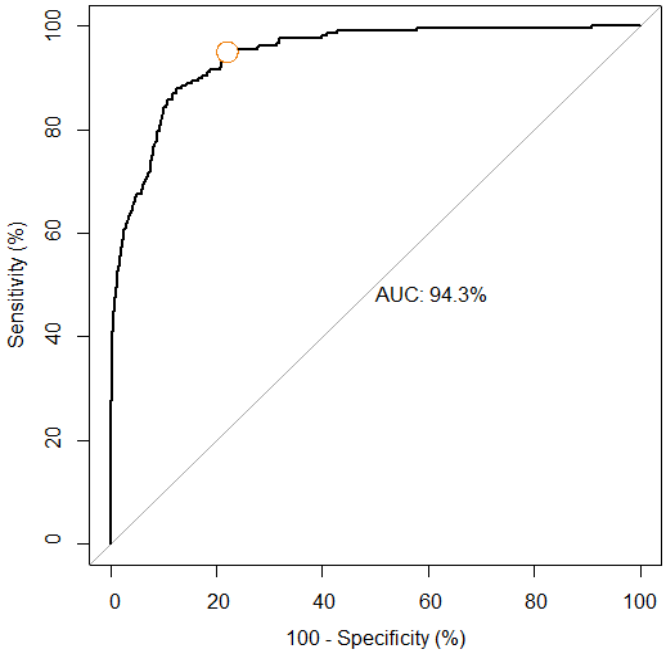


La tabella presentata qui di fianco mostra la capacità del modello di prevedere quali sono i siti sicuri e quali quelli pericolosi sulla base di un determinato valore di cut-off della stima della probabilità che si ottiene. Per il cut-off scelto, se tale probabilità risulta essere maggiore di 0,5, allora il sito verrà indicato come pericoloso, altrimenti sarà ritenuto sicuro.

	False	True
Sicuro	1541	24
Pericoloso	101	115

Misuriamo innanzitutto l'accuratezza totale del modello, che risulta esse pari a 0,9298. Questo significa che l'algoritmo di classificazione basato su tale modello riesce a classificare correttamente 93 siti ogni 100. Si tratta di un buon risultato, bisogna poi indagare sulla percentuale dei siti che risultano positivi al test i siti pericolosi e sulla percentuale di siti sicuri che invece risultano negativi al test.
Utilizzando tale sistema, sono stati individuati come pericolosi solo 115 siti rispetto al totale dei siti pericolosi, che è di 216 pagine non sicure. Si ha quindi che la sensibilità di tale strumento di diagnosi sia circa 0,53. Questo significa che nel 53% dei casi un sito pericoloso verrà riconosciuto come tale da l'algoritmo di classificazione che implementa tale modello. Non abbiamo un risultato alquanto soddisfacente. Ci rincuora, però, il livello di specificità di tale strumento di diagnosi, che è pari a 0,9847. Si ha, quindi, che nel 98% dei casi un sito sicuro venga riconosciuto come tale. Questo è un ottimo risultato, ma non risulta soddisfacente, in quanto il nostro scopo è quello di salvaguardarci da un sito pericoloso. Dal livello di sensibilità abbiamo che il 47%

dei siti realmente pericolosi vengano riconosciuti dall'algoritmo come sicuri. Vi è quindi un alto rischio di finire vittime di qualche cybercriminale. Dobbiamo aumentare, quindi, il numero di veri positivi, e di conseguenza il livello di sensibilità del modello. Per far questo dobbiamo spostare il livello di cut-off, aiutandoci tramite la visualizzazione grafica delle curve di ROC.



Dalla curva di ROC si può notare come un accettabile livello di sensibilità si possa raggiungere nel punto indicato dal pallino arancione. Analizzando i dati utilizzati per la costruzione di tale curva di individua come cut-off il livello di 0,07. Applicando tale valore nei test di sensibilità, specificità ed accuratezza otteniamo tali risultati.

```
Sensibilità 0,9537
Specificità 0,7649
Accuratezza 0,7878
```

Si è riusciti ad ottenere un sostanzioso aumento nel livello di sensibilità, a discapito ovviamente della specificità. Ora il modello riesce ad individuare come pericolosi il 95% dei siti che sono tali. Vi è quindi un 5% di siti pericolosi che non vengono individuati come tali, ma risulta essere più accettabile rispetto ad un quasi 50%. Ovviamente, questo ha richiesto una diminuzione notevole del livello di cut-off, che potrebbe essere anche discutibile. Effettivamente con un cut-off pari a 0,07, già al minimo segnale di pericolo il sito viene indicato come non sicuro. Bisogna quindi solo vedere se per non rischiare siamo disposti a rinunciare a siti che invece erano sicuri, e questo creda che dipenda tutto solo dal contenuto di tali pagine web. Se proprio non ne possiamo fare a meno, se ben protetti un piccolo rischio si può correre!

Considerazioni finali

Possiamo ora trarre alcune conclusioni dalla stima del modello.

- Bisogna prestare attenzione ai quei siti internet che presentano una lunghezza dell'URL consistente. Effettivamente, un aumento di 10 caratteri nel URL genera un aumento della propensione del sito ad essere pericoloso del 33%.
- Bisogna anche prestare attenzione ai siti spagnoli. Dai dati risultano essere i meno sicuri, con una probabilità di almeno 262 volte maggiore, fino ad un massimo di 267 volte, rispetto ai siti internet provenienti dalle altre nazioni.
- I server più utilizzati sono anche quelli che veicolano il maggior numero di siti internet pericolosi. Navigando su un sito web gestito dal server Apache vi è una propensione che esso sia pericoloso pari a 7,9 volte la stessa propensione che si avrebbe navigando su siti derivanti da altri Server. Bisogna, anche, notare che la probabilità di trovarsi di fronte ad un sito pericoloso è simili considerando tutti e tre i maggiori server, con una differenza di pochi punti percentuali, anche al di sotto dell'1%.
- Per quanto riguarda l'ipotesi che i server più datati siano anche i più vulnerabili alle minacce, essa risulta essere sfatata. Effettivamente la propensione di trovarsi di fronte un sito pericoloso nel caso in cui si navighi attraverso un server web degli anni 2000 è il 65% della stessa propensione nel caso in cui, invece, il sito provenga da un server degli anni 2010.

In conclusione, per evitare di essere vittime di qualche cyber criminale, bisogna prestare attenzione ai siti che presentano URL lunghi e con una percentuale alta di caratteri speciali, ma soprattutto bisogna evitare i siti spagnoli!

Simulazioni del funzionamento dell'algoritmo di classificazione

In ultimo, andiamo a simulare manualmente la funzione dell'algoritmo di classificazione. Prenderemo in considerazione alcuni prototipi di sito, non presenti all'interno del dataset, e vedremo se essi risulteranno pericolosi oppure no.

- Il primo sito che andiamo a considerare possiede una lunghezza dell'URL pari a 30 caratteri, con una percentuale di caratteri speciali pari al 50%. Il sito è gestito da un server della famiglia Apache, di cui però non se ne conosce la data di registrazione. Infine, sappiamo che il sito non è spagnolo, ma proveniente da una nazione americana.

In questo caso la probabilità stimata di essere un sito pericoloso sarà pari a 0,999, ossia si può dire pari a 1. Il sito in questione viene classificato come pericoloso. Tale valore sarà dovuto molto probabilmente all'alta percentuale di caratteri speciali presenti nel suo URL.

- Il secondo è un sito spagnolo, con un URL di 70 caratteri di cui il 10% sono caratteri speciali. Viene gestito da un server della famiglia nginx, con data di registrazione appartenente agli anni '90

0,4 è il risultato che si ottiene stimando la probabilità di essere un sito pericoloso. Per cui con un cut-off di 0,07, il sito viene considerato pericoloso.

- Il terzo è un sito con un URL formato da 50 caratteri, di cui 10 sono speciali. Tale sito appartiene ai server Microsoft, in maniera particolare ad un software registrato negli anni 2000. È un sito internet proveniente dalla Cina.

In questo caso la probabilità di essere un sito pericoloso è pari al 0,0641. Si tratta di un valore molto vicino al livello di cut-off. Calcolando l'intervallo di confidenza possiamo vedere che tale probabilità ha un minimo di 0.0347 fino ad un massimo di 0.1155. Il livello di cut-off risulta, quindi, compreso in tale intervallo. Vi è quindi il rischio che una classificazione come sito sicuro risulti poi errata. Molto probabilmente per tale ragione sarebbe meglio perdere un sito sicuro piuttosto che correre il pericolo.