Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica e Tecnologie Informatiche

RELAZIONE FINALE

ANALISI DELL'OUTPUT DI UN SISTEMA DI TICKETING: COSA CI DICONO I DATI?

Relatore Prof. Mariangela Guidolin
Dipartimento di Scienze Statistiche

Laureando: Flavio Boschiggia
Matricola N. 575195

Anno Accademico 2017/2018

Introduzione

I sistemi di ticketing, utilizzati speso negli ambienti helpdesk, sono dei sistemi informatizzati in grado di permettere ad un cliente di essere assistito, per ogni eventuale problema o richiesta, attraverso uno staff in grado di gestire velocemente il lavoro. Un sistema di ticketing è uno strumento web, appoggiato a un database, accessibile dagli utenti, che permette una facile introduzione e descrizione della propria necessità attraverso l'apertura di un ticket, componente principale di un sistema di ticketing. Lo staff tecnico dell'azienda, dall'altra parte, è in grado di tracciare la richiesta, inserire particolari file e note in modo da arrivare alla soluzione di quanto chiesto dall'utente. Alla fine l'utente ha una visione completa del proprio ticket, potendo visualizzare l'avanzamento della soluzione step by step e con ampia documentazione. Inoltre, in ogni momento, è possibile raggiungere i dati di tutte le proprie richieste per una consultazione, oppure per vedere se una richiesta precedente risolta può aiutare anche in un caso presente.

Obiettivo dello studio

Una delle caratteristiche basilari di ogni sistema di ticketing è quelli di poter generare dei dataset che permettono di studiare come sta andando il servizio di helpdesk.

Non avendo a disposizione una situazione reale ho trovato una buona base di dati in internet, disponibile in bibliografia.

L'obiettivo è quello di partire dai dati, in questo specifico caso un file excel, per trarre diverse conclusioni sulla soddisfazione dei clienti e sull'efficienza del servizio. A tal proposito, ho pensato di strutturare il mio lavoro in tre parti.

La prima parte prevede la descrizione dei dati, del sistema di ticketing in esame, la selezione delle variabili, la pulizia dei dati e diverse statistiche descrittive che possano darci l'idea di come sono fatte le singole variabili e di quale sia la relazione tra loro.

La seconda parte invece prevede la valutazione del personale in cui vaso a studiare il comportamento degli operatori in base al problema che devono risolvere.

Ogni operatore infatti va ad assegnare un grado di priorità ad ogni ticket che prende in carico. Cosa influenza l'attribuzione di un livello di priorità? Forse il tipo di cliente? O il tipo di problema?

Ogni ticket viene risolto in un lasso di tempo. Cosa influenza i tempi di chiusura del ticket? Ci sono fattori che spingono gli operatori a concludere più o meno in fretta il lavoro? Per rispondere a queste due domande, farei uso di alcuni modelli lineari generalizzati.

La terza parte prevedere lo studio del grado di soddisfazione della clientela. A tal proposito voglio porre una serie modelli di previsione e valutare quello che meglio prevedere il grado di soddisfazione del prodotto consegnato. Proporrei qui qualche modello di previsione

utilizzando degli opportuni metodi di selezione delle variabili, come per esempio la divisione dei dati in stima e verifica.

Seguirà poi un capitolo con le conclusioni e un riassunto dei risultati ottenuti. Al fine di rendere chiaro quello che sto facendo, ho anche intenzione in ogni sezione di descrivere gli strumenti utilizzati a livello teorico prima di farne uso.

Capitolo 1

1.1 Il sistema di ticketing

Questo elaborato si pone l'obiettivo di analizzare la qualità del sistema di ticketing di un'anonima azienda informatica che fornisce consulenza a livello hardware e software.

Più nello specifico, il flusso di informazioni relativo ad un ticket è il seguente. Un generico cliente con codice identificativo Requestor richiede alla nostra azienda un'intervento attraverso un ticket. Di questo cliente è noto il suo ruolo nell'azienda in cui lavora, identificato dalla variabile RequestorSeniority, e la severità che attribuisce al problema per cui ha chiesto l'intervento, registrata come Severity.

Un operatore della nostra azienda con codice identificativo ITOwner prende in carico il problema, e nell'ordine procede con le seguenti operazioni:

1. classifica il problema sulla base della sua natura (hardware, software, sistema, etc…) e registra l'informazione nella variabile FiledAgainst;
2. controlla nel database di sistema se un problema simile è mai stato affrontato precedentemente in azienda e registra l'informazione in TicketType;
3. assegna una priorità al ticket (Priority), che può coincidere oppure no con quella fornita dal cliente.

Una volta classificato, si procede alla risoluzione del problema. Il numero di giorni in cui il ticket rimarrà aperto fino alla sua risoluzione viene registrato nella variabile daysOpen. Una volta che il lavoro è stato ultimato, il cliente ha la possibilità di esprimere il suo grado di soddisfazione per il servizio ricevuto, contenuto in Safisfaction. Una sintesi grafica di questa procedura viene riportata in Figura 1, dove viene mostrato il flusso di informazioni relativo ad un singolo ticket dal momento della richiesta fino alla consegna al cliente.

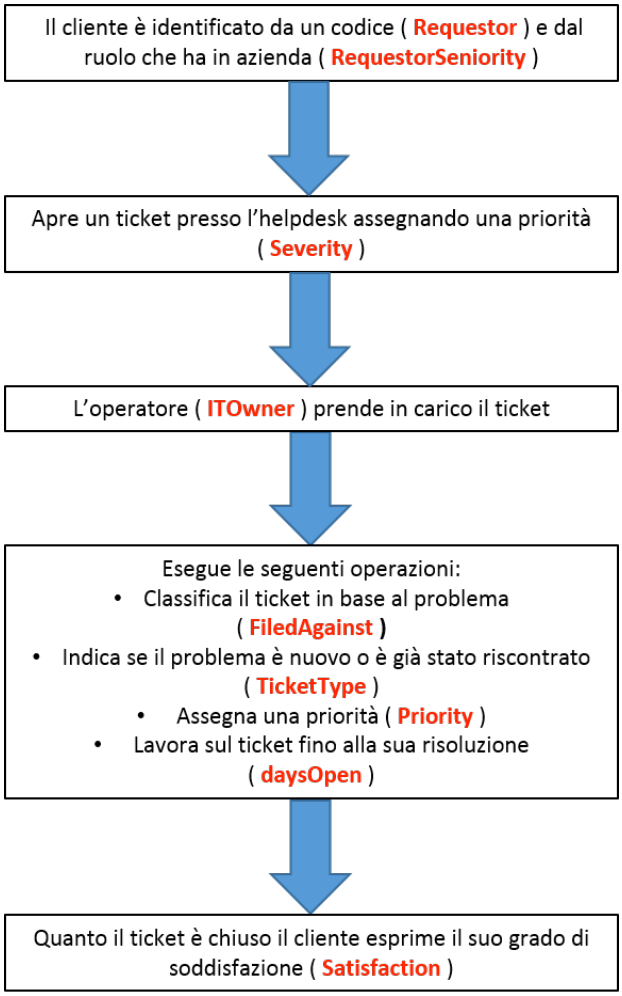

Figura 1: flusso di gestione di un singolo ticket richiesto da un cliente all'helpdesk dell'azienda.

Procediamo ora con l'analisi descrittiva delle singoli variabili, discutendo su quali possono risultare interessanti per un'analisi statistica dei dati al fine di comprendere meglio quali la qualità del servizio erogato dalla nostra azienda e quali sono i punti su cui concentrarsi maggiormente.

1.2 Analisi descrittiva delle variabili

In questa sezione procediamo ad analizzare le variabili contenute nell'insieme di dati analizzato, composto da 100mila osservazioni.

Per prima cosa eliminiamo i codici identificativi dei clienti e degli operatori, in quanto non risultano utili ai fini dell'analisi statistica che svolgeremo.

La prima variabile che analizziamo è RequestorSeniority, relativa al ruolo dei richiedenti nella azienda in cui lavorano. Questa variabile è divisa in quattro classi che rappresentano il grado dei clienti, dal più piccolo (Junior) al più grande (Managment). Dal grafico in Figura 2 emerge che le richieste vengono maggiormente fatte da clienti di livello Regular; le altre tre classi hanno invece un comportamento molto simile.

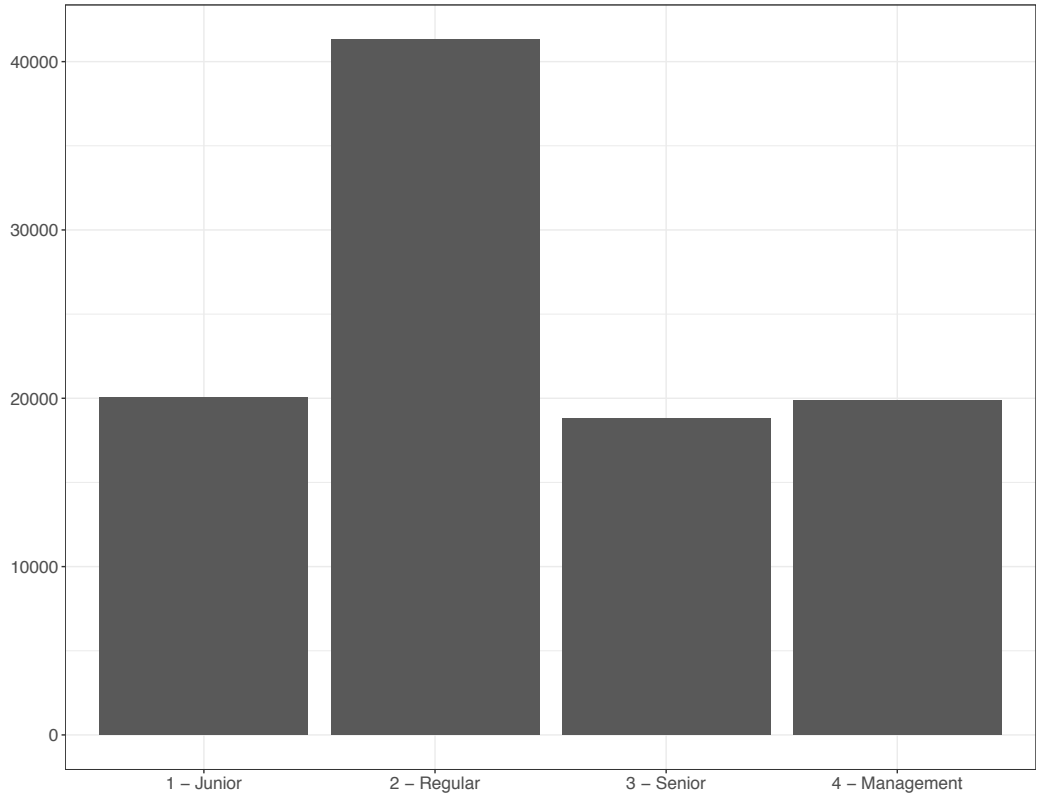

Figura 2: frequenze assolute della variabile RequestorSeniority.

La seconda variabile di interesse è FiledAgainst, che riporta il tipo di problema. Generalmente, un ticket viene classificato in una delle quattro possibili classi selezionate: Hardware se vi è un problema relativo ad una macchina o a una componente fisica, Software se il problema è legato invece ad un programma, Access/Login se il richiedente ha un problema di autenticazione all'interno della piattaforma informatica aziendale e Systems se il problema è legato invece ai sistemi operativi delle macchine.

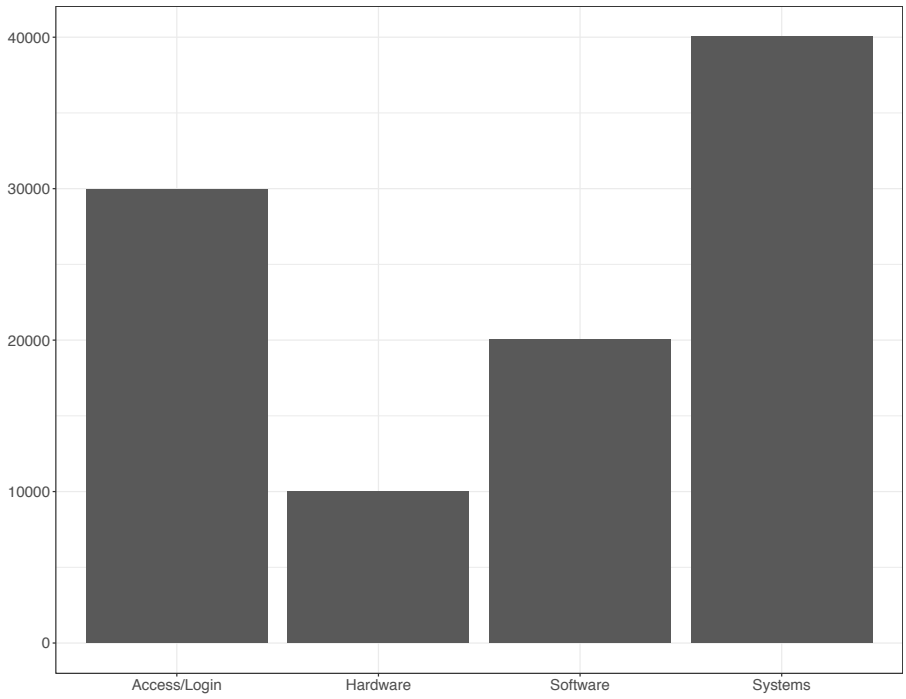

Figura 3: frequenze assolute della variabile FiledAgainst.

Dalla Figura 3 emerge che i maggiori problemi sono legati all'accesso (40%) e ai sistemi operativi (30%). Gli interventi software occupano invece il 20% del totale, e il restante 10% è occupato dai problemi hardware.

Riportiamo ora la tabella delle frequenze relative congiunte per le variabili appena analizzate.

	Access/Login	Hardware	Software	Systems
1 - Junior	0.05980	0.02001	0.04079	0.07980
2 - Regular	0.12291	0.04064	0.08315	0.16633
3 - Senior	0.05678	0.01934	0.03827	0.07362
4 - Management	0.05972	0.01977	0.03847	0.08060

Tabella 1: frequenze relative congiunte delle variabili RequestorSeniority e FiledAgainst.

Attraverso il test del chi-quadrato per l'indipendenza, possiamo testare l'indipendenza delle variabili. In particolare, vale che due variabili possono essere considerate come indipendenti se, per ogni riga i e per ogni colonna j vale che

$$n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n},$$

ovvero la singola frequenza congiunta è ottenibile come prodotto delle frequenze marginali e diviso per la numerosità totale. La distribuzione asintotica del test è un chi-quadro con 9 gradi di libertà; il p-value risulta pari a 0.0657 e ci spinge ad accettare l'ipotesi nulla di indipendenza delle variabili. Non sembra quindi che statisticamente vi sia un legame tra il ruolo del richiedente e la tipologia di problema.

La variabile Severity contiene la priorità che il cliente associa al ticket sottomesso. Notiamo che essa è composta da quattro livelli di grado crescente, più un livello denominato Unclassified che ci mostra come alcuni clienti hanno preferito non specificare il grado di severità associato alla richiesta effettuata.

Dal grafico in Figura 4 notiamo che alla maggior parte delle richieste (91%) è stato associato un grado di importanza standard, noto come Normal. Soltanto lo 0.3% dei richiedenti non ha espresso un grado di priorità per il ticket sottomesso.

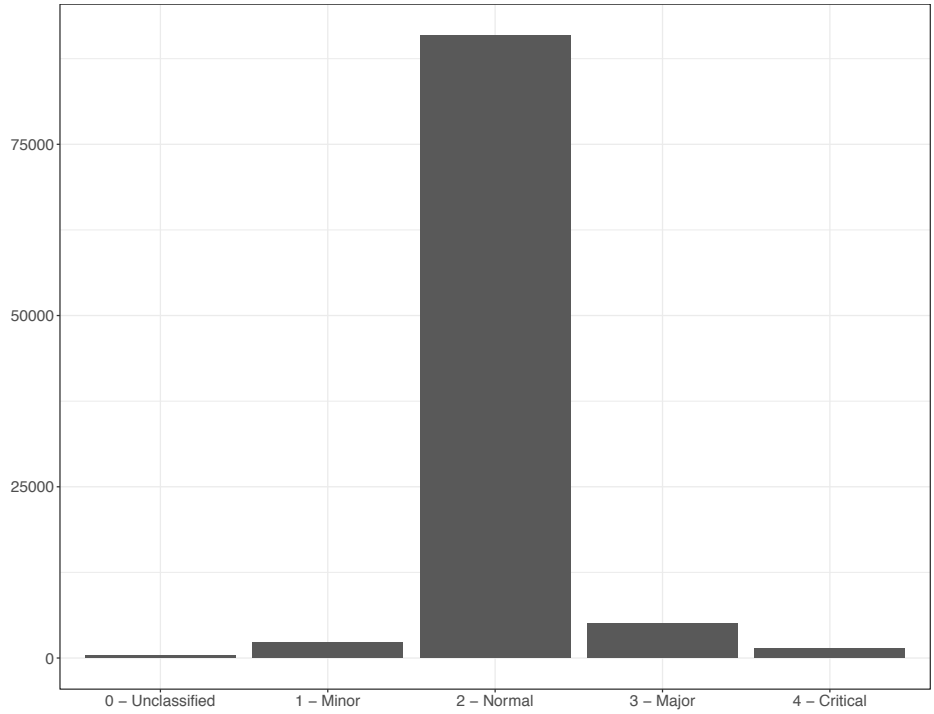

Figura 3: frequenze assolute della variabile Severity.

Valutiamo ora la presenza di un possibile legame statistico tra la variabile appena descritta e Priority, ovvero la priorità assegnata dai nostri operatori ai ticket ricevuti. Attraverso il medesimo test del chi-quadro descritto precedentemente vogliamo valutare la presenza di un possibile legame fra le due variabili di valutazione; l'accettazione dell'ipotesi comunicherebbe in questo caso che la valutazione degli operatori non è influenzata dalla priorità assegnata dal cliente, mentre un rifiuto dell'ipotesi nulla suggerirebbe la presenza di un legame.

La Tabella 2 mostra le frequenze relative congiunte delle due variabili.

	0 - Unclassified	1 - Minor	2 - Normal	3 - Major	4 - Critical
0 - Unassigned	0.00120	0.00642	0.27478	0.01469	0.00418
1 - Low	0.00082	0.00564	0.15670	0.00626	0.00175
2 - Medium	0.00057	0.00420	0.14834	0.00740	0.00207
3 - High	0.00108	0.00691	0.32930	0.02139	0.00630

Tabella 2: frequenze relative congiunte delle variabili Priority e Severity.

L'applicazione del test del chi-quadro con 12 gradi di libertà porta ad un p-value che rasenta lo zero e quindi ad un rifiuto dell'ipotesi nulla in favore dell'ipotesi alternativa. Le due variabili non possono essere considerate come indipendenti; questo si può tradurre come un segnale del fatto che la valutazione degli operatori è in qualche modo legata al grado di priorità dato dai clienti.

Studiamo ora il comportamento della variabile daysOpen, ovvero il numero di giorni spesi per risolvere i singoli ticket. Essa è una variabile quantitativa discreta; il grafico in Figura 4 mostra una struttura asimmetrica concentrata maggiormente attorno a valori bassi.

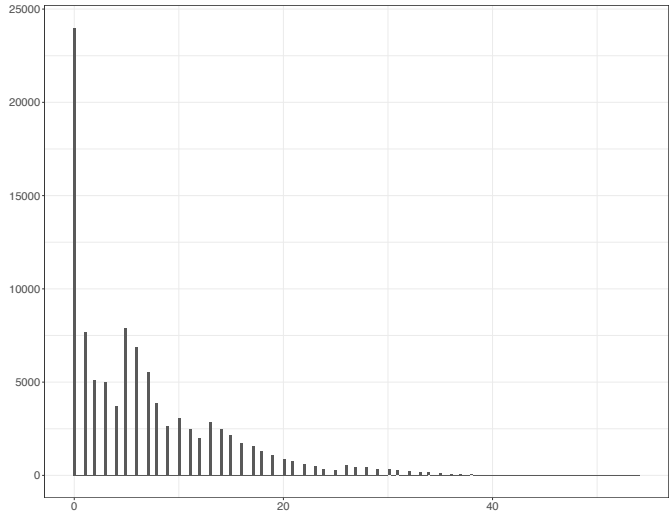

Figura 4: grafico in frequenze assolute della variabile daysOpen.

Vogliamo ora analizzare graficamente la relazione fra il numero di giorni in cui un ticket rimane aperto e il grado aziendale del richiedente. Questa operazione può già darci un'idea della presenza o meno di disparità nel trattamento dei clienti e ci può quindi aiutare a comprendere se il servizio offerto varia a seconda delle classi dei richiedenti.

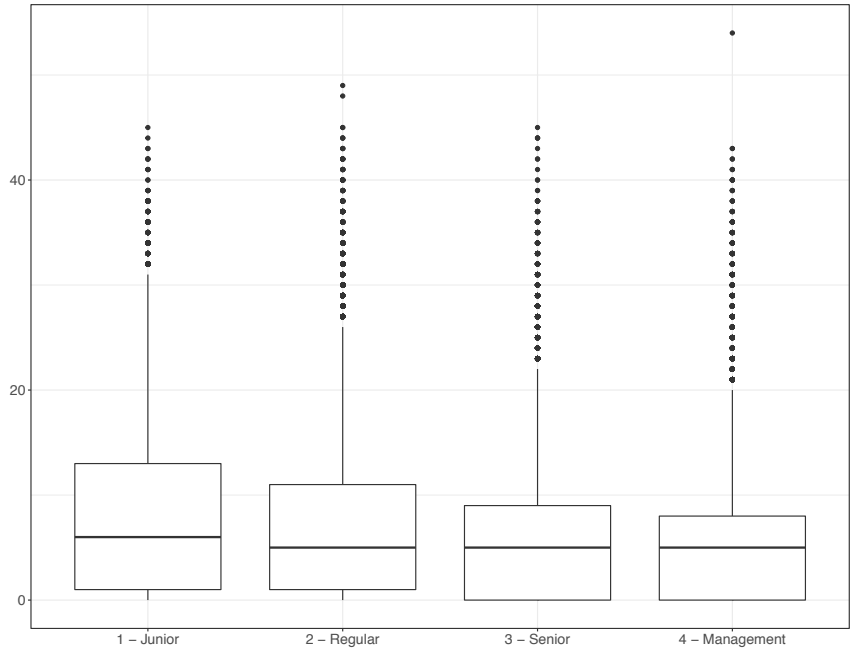

Figura 5: distribuzione del numero di giorni spesi per risolvere i ticket aperti in base al richiedente.

Dal grafico in Figura 5, che mostra i boxplot della variabile daysOpen diversificati per classi dei richiedenti, emergono distribuzioni più concentrate attorno allo zero all'aumentare del grado del cliente. Da una prima analisi sembra quindi che l'importanza del cliente vada ad influenzare l'operato dei nostri addetti, spingendoli a risolvere i problemi in minor tempo per i manager aziendale e in più tempo per i dipendenti Junior. Questa supposizione dovrà comunque essere confermata da una più approfondita analisi, che valuterà se questo andamento decrescente risulta statisticamente significativo o se può essere attribuito invece a fluttuazioni casuali dei dati.

Riportiamo infine le frequenze relative della variabile Satisfaction, che riporta il grado di soddisfazione dei clienti una volta chiuso il ticket. La variabile si compone di tre livelli di grado crescente, più un quarto che riguarda la mancata assegnazione di un voto. Dalla Tabella 3 osserviamo che più del 30% degli utenti non ha espresso una valutazione finale, ma non possiamo dire se questo sia attribuibile ad una insoddisfazione o semplicemente ad una mancanza di interesse nell'esprimere un giudizio.

0 - Unknown	1 - Unsatisfied	2 - Satisfied	3 - Highly satisfied
0.302	0.211	0.196	0.291

Tabella 3: frequenze relative della variabile Satisfaction.

Una medesima percentuale di clienti è risultata insoddisfatta o parzialmente soddisfatta, mentre il 30% è stato altamente soddisfatto.

Capitolo 2

In questo capitolo vogliamo studiare, attraverso l'utilizzo di opportuni metodi statistici, il comportamento degli operatori dell'azienda che prendono in carico i ticket. In particolare, vogliamo rispondere a due domande:

1. quali sono i fattori che inducono un certo operatore ad assegnare una determinata priorità ad un ticket?
2. quali sono i fattori che impattano sul tempo di chiusura di un certo ticket?

Procediamo quindi per ordine in modo da tratte delle conclusioni su questi due quesiti.

2.1 Studio dei criteri di assegnazione delle priorità

In questa sezione vogliamo studiare il criterio con cui gli operatori assegnano le priorità ai ticket. Da quanto emerge in Tabella 2, la variabile Priority è composta da quattro livelli, uno dei quali mostra che alcuni ticket non sono stati valutati. Scegliamo di eliminare tutte quelle osservazioni con valore Unassigned.

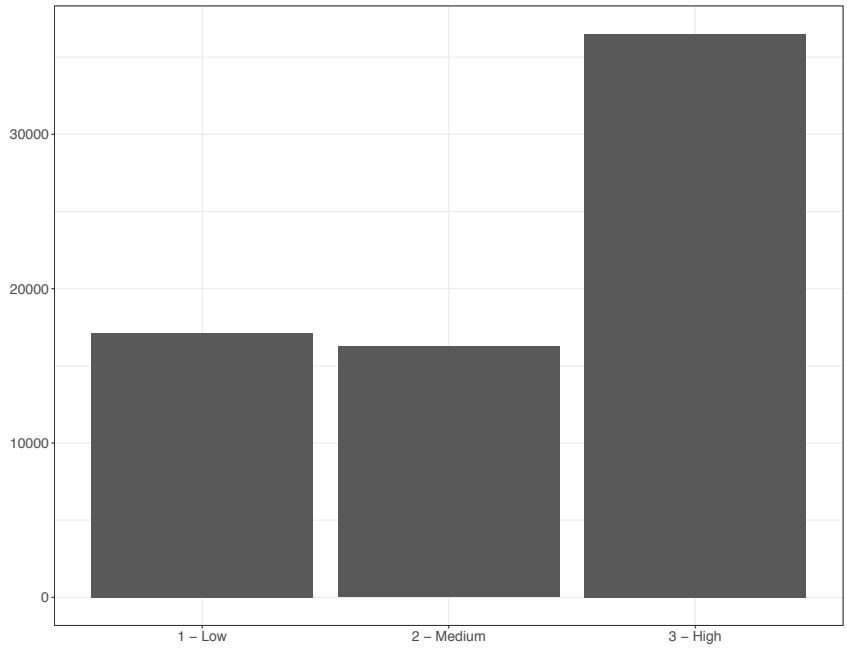

Figura 6: frequenze assolute della variabile Priority.

Il grafico in Figura 6 mostra la distribuzione della variabile di interesse. La numerosità all'interno della classe più alta è pari a circa la somma delle altre due, quella intermedia e quella bassa. Per questa ragione, ci concentriamo sul ripondere alla seguente domanda: cosa spinge un operatore ad assegnare priorità alta? La variabile analizzata è quindi dicotomica, dove 1 indicherà l'assegnazione di livello High, mentre 0 indicherà il livello Low/Medium.

Per la costruzione di un modello statistico, selezioniamo un insieme di variabili che potrebbero essere utilizzate come covariate, e valutiamone la significatività. Per ragioni legate al flusso di gestione descritto in Figura 1, escludiamo a priori le variabili daysOpen e Satisfaction, in quanto il processo di

assegnazione della priorità avviene a monte, prima che l'assistenza vera e propria abbia inizio.
Passiamo ora all'introduzione del modello statistico di cui varemo uso per analizzare la variabile di interesse. Sia $Y_i \in \{0, 1\}, i = 1, \dots, n$ una collezione di variabili dicotomiche osservate sul campione di n osservazioni. Un modello statistico per la variabile causale Y_i può essere

$$Y_i \sim Be(p_i),$$

dove $Be(p_i)$ rappresenta la distribuzione di Bernoulli con probabilità p_i di essere uguale a 1. Al fine di studiare un legame fra la variabile aleatoria considerata e una serie di altre variabili incluse nell'insieme di dati, possiamo ipotizzare il seguente legame:

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik},$$

che in letteratura statistica è conosciuto come legame logit e che viene utilizzata per garantire che, qualunque sia il valore delle variabili x_{i1}, la probabilità di successo p_i sia inclusa fra 0 e 1. Il vettore dei coefficienti $\beta = (\beta_0, \dots, \beta_k)$ ha una distribuzione asintotica normale multivariata, e questo ci permette di procedere facilmente con l'inferenza.

Il modello statistico che vogliamo stimare è quindi un modello lineare generalizzato con distribuzione di Bernoulli e legame logit. Le variabili utilizzate per descrivere la probabilità di successo p_i sono: il grado del cliente nella sua azienda, la tipologia di problema, la presenza/assenza di un problema nuovo per

l'azienda e il grado di importanza assegnato dal cliente. I risultati del modello sono riportati in Tabella 4.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.211345	0.153831	-14.375	< 2e-16
RequestorSeniority2 - Regular	1.463573	0.026674	54.868	< 2e-16
RequestorSeniority3 - Senior	2.887057	0.031344	92.109	< 2e-16
RequestorSeniority4 - Management	3.803051	0.035817	106.179	< 2e-16
FiledAgainstHardware	-0.001142	0.032507	-0.035	0.97197
FiledAgainstSoftware	0.001731	0.025688	0.067	0.94628
FiledAgainstSystems	0.008294	0.021527	0.385	0.70003
TicketTypeRequest	0.023776	0.020728	1.147	0.25138
Severity1 - Minor	-0.191920	0.162439	-1.181	0.23741
Severity2 - Normal	0.406166	0.151877	2.674	0.00749
Severity3 - Major	0.971379	0.156726	6.198	5.72e-10
Severity4 - Critical	1.036441	0.168670	6.145	8.01e-10

Tabella 4: coefficienti, standard error e p-value ottenuti dal modello lineare generalizzato sulla variabile Priority a due classi.

Dai risultati si possono trarre alcune interessanti conclusioni circa i nostri operatori:

- La tendenza ad assegnare alta priorità ad un ticket è fortemente legata al ruolo del richiedente. In particolare, dal modello si evince che tutti i livelli della variabile RequestorSeniority impattano in modo statisticamente rilevante sull'assegnazione della priorità rispetto al livello Junior; le stime dei coefficienti aumentato considerevolmente all'aumentare del livello, segno che la probabilità di ricevere High è più alta al crescere del proprio ruolo in azienda al netto delle altre variabili;
- La tipologia di problema tecnico non sembra essere in alcun modo legata alla priorità assegnata, così come il fatto di

trovarsi ad affrontare un problema nuovo rispetto a quelli fatti fino ad allora;

- L'assegnazione di un alto livello di severità da parte del cliente contribuisce positivamente all'aumento della probabilità di classificazione del ticket come urgente, mentre l'assegnare un livello Minor non sembra avere una certa rilevanza statistica. Il confronto viene fatto rispetto al livello Unclassified. Per esempio, assegnare grado di severità Normal rispetto a non classificare equivale a un incremento di 0.41 sul log-rapporto delle probabilità, di 0.97 se il cliente assegna priorità Major e di 1.04 se si assegna Critical.

Dalle stime ottenute emerge quindi un importante concetto chiave: la valutazione di un problema da parte dei nostri operatori, e la conseguente classificazione che emerge da parte loro nei confronti di un ticket, dipende esclusivamente dal cliente che ha presentato quel ticket, mentre la natura del problema non ha statisticamente alcuna valenza.
Analizziamo ora i grafici dei residui di regressione che ci permettono di comprendere la qualità del modello proposto.

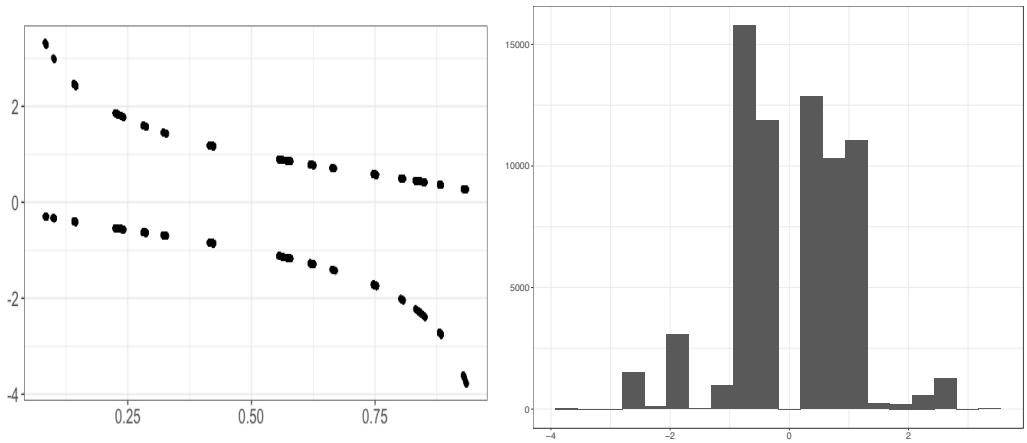

Figura 7: sinistra, grafico di dispersione dei residui di Pearson rispetto ai valori predetti dal modello; destra, istogramma dei residui.

Dai grafici riportati in Figura 7possiamo osservare la distribuzione dei residui di Pearson ottenuti dal modello. I residui di Pearson sono un ottimo strumento per valutare i residui di regressione di un modello lineare generalizzato che non faccia uso della distribuzione Normale per modellare la relazione tra la variabile risposta e le variabili esplicative. Questa tipologia di residui si ottiene attraverso una procedura di standardizzazione, ovvero vengono prima sottratti alla media e successivamente divisi per lo scarto quadratico medio osservato. Dai grafici in Figura 7 emerge una struttura simmetrica dei residui attorno allo zero. La variabilità sembra aumentare per valori predetti attorno a 0 e a 1. La particolare struttura dei residui che emerge dal grafico a destra, che sembra mostrare un andamento non casuale dei dati, è prettamente dovuta alla natura discreta di tutta l'informazione utilizzata nel modello. Sia la variabile risposta che le variabili esplicative utilizzate risultano infatti discrete con un numero di livelli non superiore a 5. A causa di questo, il modello

stimato è in grado di identificare soltanto un numero finito di possibili combinazioni in base ai valori assunti dalle osservazioni. Questo effetto si ripercuote sui valori previsti dal modello in maniera analoga a quanto presentato in Figura 7.

2.2 Analisi dei tempi di risoluzione dei ticket

Spostiamo ora la concentrazione su una seconda variabile di grande importanza dalla quale possiamo trarre informazioni circa l'operato dei nostri tecnici, ovvero il tempo di risoluzione dei ticket.
Così come abbiamo già discusso precedentemente, una volta che il ticket è stato classificato sulla base della sua priorità e del tipo di problema, si passa al processo di risoluzione dello stesso.
L'insieme di dati analizzato ci comunica quanti giorni sono trascorsi dalla presa in carico di ciascun ticket alla sua risoluzione finale. L'analisi presentata in questo paragrafo è volta pertanto a comprendere quali sono i fattori che impattano sui tempi di risoluzione dei ticket inviati all'azienda, per eventualmente intervenire e correggere l'operato dei nostri tecnici.

La variabile di interesse è nota nell'insieme di dati come daysOpen e conta il numero di giorni trascorsi prima della chiusura di ciascun ticket. Il grafico della variabile analizzata è riportato in Figura 4 e mostra un andamento prettamente asimmetrico verso sinistra.

Essendo che la variabile risposta è un conteggio senza alcun limite massimo, proponiamo di analizzare daysOpen attraverso un modello lineare generalizzato con distribuzione Poisson. Più dettagliatamente, consideriamo una collezione di variabili aleatorie $Y_i \in \mathbb{N}, i = 1, \dots, n$ osservate sul campione di n osservazioni. Assumiamo quindi che la variabile aleatoria Y_i può essere descritta come

$$Y_i \sim Poi(\lambda_i),$$

dove $Poi(\lambda_i)$ rappresenta la distribuzione di Poisson con media λ_i. La funzione più usata per legare questo parametro a una collezione di altre variabili è la funzione logaritmo; più nel dettaglio, assumiamo che

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik},$$

in modo da mantenere λ_i sempre positivo. Anche in questo caso, i coefficienti sono asintoticamente normali.

I regressori che possono essere presi in considerazione sono tutte quelli che sono stati mantenuti nel capitolo 1, fatta eccezione naturalmente per Satisfaction, ovvero il grado di soddisfazione del cliente ad intervento ultimato, la cui inclusione condurrebbe ad un errore logico.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.111947	0.028394	-74.380	<2e-16
RequestorSeniority2 - Regular	-0.057213	0.003222	-17.757	<2e-16
RequestorSeniority3 - Senior	-0.116254	0.004158	-27.962	<2e-16
RequestorSeniority4 - Management	-0.144699	0.004251	-34.035	<2e-16
FiledAgainstHardware	4.165638	0.011431	364.417	<2e-16
FiledAgainstSoftware	3.129336	0.011534	271.313	<2e-16
FiledAgainstSystems	3.571685	0.011292	316.313	<2e-16
TicketTypeRequest	0.751537	0.003556	211.370	<2e-16
Severity1 - Minor	0.064977	0.027495	2.363	0.0181
Severity2 - Normal	0.246699	0.025952	9.506	<2e-16
Severity3 - Major	0.387027	0.026445	14.635	<2e-16
Severity4 - Critical	0.444243	0.027894	15.926	<2e-16
Priority1 - Low	0.184892	0.003596	51.417	<2e-16
Priority2 - Medium	0.049699	0.003692	13.463	<2e-16
Priority3 - High	-0.119165	0.003162	-37.688	<2e-16

Tabella 5: coefficienti, standard error e p-value ottenuti dal modello lineare generalizzato sulla variabile daysOpen.

Le variabili risultato tutte fortemente significative. Unica eccezione viene fatta per la label Minor della variabile Severity, che sembra avere sì un impatto significativamente diverso rispetto al livello base, ovvero Unclassified, ma meno evidente rispetto alle altre modalità. Prima di procedere con l'interpretazione dei risultati, è necessario valutare la presenza di un fenomeno comune quando si effettua la regressione di Poisson, ovvero la sovra dispersione.
Ciò che accade con questo tipo di regressione è che la media condizionata di ciascuna osservazione viene posta uguale alla varianza condizionata. Questa assunzione in realtà non è sempre rispettata dai dati e può condurre ad un'errata valutazione della significatività dei regressori del modello.
Una soluzione per ovviare a questo problema descritta in Ver Hoef & Boveng (2007) è quella di considerare una variabile aleatoria Y_i

con valore atteso λ_i e varianza $\kappa\lambda_i$, dove κ è un parametro di dispersione. Questo tipo di distribuzione è in realtà una generalizzazione del modello di Poisson, che assume invece $\kappa = 1$ ed è noto in letteratura come modello quasi-Poisson.
Facendo così, la variabilità in eccesso che non può essere spiegata dalla regressione di Poisson viene inclusa nella stima del parametro di dispersione, permettendo così una corretta stima della variabilità dei regressori del modello.
L'utilizzo del modello quasi-Poisson non altera la stima dei coefficienti rispetto a quelli individuati precedentemente, bensì ne cambia gli errori standard riportati. Rispetto ai valori in Tabella 4, ci aspettiamo che la significatività dei regressori non cambi, fatta eccezione per quello associato al livello Minor della variabile Severity.
Il parametro di dispersione viene stimato pari a 2.476, confermando la presenza di sovra dispersione nei dati e l'inacuratezza del modello precedente. Come previsto, il coefficiente associato a Minor assume ora uno standard error pari a 0.043 e un p-value di 0.133. Passiamo ora all'interpretazione dei risultati, sintetizzando il tutto attraverso alcuni punti chiave.

- Così come per il modello nel paragrafo 2.1 per la variabile Priority, il grado del cliente che sottomette il ticket all'helpdesk dell'azienda sembra avere un impatto significativo sul tempo di risoluzione del problema. In particolare, dai risultati in Tabella 5 emerge che il tempo medio di risposta per un dipendente Regular è pari circa al 94.46% rispetto a quello per un dipendente Junior, per un Senior è circa l'89.05% e per un Manager l'86.53%, il tutto al

netto del valore delle altre variabili. Queste percentuali possono essere facilmente ricavate calcolando l'esponenziale dei valori riportati nella colonna Estimate in Tabella 5; per esempio, 0.9446 è l'esponenziale di -0.0572.

- I problemi che vengono risolti più facilmente sono quelli legati all'accesso. Tenendo come riferimento questo livello, i problemi software richiedono un tempo medio di risoluzione pari a 22.87 volte rispetto a quello speso per risolvere i problemi di accesso, 35.59 per i problemi di sistema e 64.39 volte in più per l'hardware. Il tempo medio di risoluzione tende anche ad aumentare quando si presenta un problema fino ad ora sconosciuto in azienda.
- I tempi di risoluzione medi aumentano all'aumentare della severità comunicata dal cliente. Dai risultati emerge che non risultano esserci sostanziali differenze per i problemi classificati come Minor rispetto a quelli non classificati; potrebbe essere quindi di uso comune per chi sottomette il ticket di evitare di classificare un problema nel livello più basso, probabilmente per paura di essere poco presi in considerazione, prediligendo quindi la scelta di non esprimere alcun giudizio. I coefficienti associati alle classi Normal, Major o Critical indicano invece un reale aumento della difficoltà nella risoluzione del problema all'aumentare del livello della classe. Questo indica che la classe Normal contiene effettivamente problemi meno complessi, o per lo meno più veloci da risolvere, rispetto a quelli contenuti nella classe Major, e lo stesso valore per Major con Critical. Anche questo risultato è interessante, in quanto evidenza

come non vi sia una tendenza da parte di chi richiede assistenza a “sovrastimare” il proprio problema.

- I ticket che non vengono classificati in nessuno dei tre livelli della variabili Priority hanno un tempo di risoluzione minore rispetto a quelli classificati come Low o Medium, che risultano rispettivamente 1.203 e 1.0501 volte più alti; quando invece un problema è classificato con proprità alta, il tempo di risoluzione si riduce all’88.7% rispetto a quelli senza alcun livello di priorità assegnata.

Capitolo 3

In questo capitolo vogliamo concentrarci sull'unica variabile a disposizione utile ad analizzare il punto di vista del cliente, ovvero il grado di giudizio espresso una volta che il lavoro è stato ultimato. Nello specifico, vogliamo studiare un modello che ci premetta di prevedere, date le caratteristiche di un certo cliente e del tipo di lavoro richiesto, quale sarà il suo grado di soddisfazione finale. In questo modo, possiamo direzionare il metodo di lavoro per ciascun nuovo cliente che sottometterà un ticket all'helpdesk in modo da rendere la sua esperienza con la nostra azienda più positiva possibile.

3.1 Preparazione del dataset

La variabile Satisfaction che racchiude il grado di soddisfazione dei clienti è divisa in tre livelli ordinati che esprimono rispettivamente una totale insoddisfazione, una parziale soddisfazione o una completa soddisfazione nei confronti del lavoro svolto. In aggiunta è presente anche un quarto livello, quello di coloro che non hanno espresso un giudizio sul lavoro svolto. Essendo che non è possibile in alcun modo recuperare questa informazione, eliminiamo tutte quelle osservazioni con questo livello della variabile risposta, che corrispondono a circa il 30% dell'informazione presente nel dataset.
Abbiamo inoltre deciso, al fine di semplificare l'analisi e puntare esclusivamente ad identificare i fattori che rendono i clienti

veramente soddisfatti, di inglobare in un'unica classe i livelli di bassa e media soddisfazione. L'analisi che stiamo per condurre si baserà quindi su una variabile risposta di tipo categoriale composta da due livelli: non soddisfatto (NS) e soddisfatto (S).

Come descritto in Azzalini & Scarpa (2009), la costruzione di un modello di previsione deve essere effettuata avendo cura di separare i dati in due parti, una delle quali verrà utilizzata per costruire e stimare i modelli, mentre la seconda servirà per valutarne la bontà di classificazione.
Questo tipo di differenziazione deve essere fatta al fine di stimare correttamente la capacità predittiva degli strumenti utilizzati.
Essendo che il nostro insieme di dati è composto da un abbondante numero di osservazioni (circa 70mila), possiamo pensare di dividerlo in due parti: la prima, composta dal 70% delle osservazioni, costituirà l'insieme di dati su cui i modelli saranno stimati e selezionati, mentre il restante 30% fungerà da test dove validare le loro capacità predittive.
Nell'insieme di dati di stima, la percentuale di osservazioni nella classe NS è circa il 60%: scegliamo quindi di ridurre le osservazioni in questa prima classe in modo di bilanciare la quantità di osservazioni in ciascuna nelle due. Il dataset di riduce così ad avere circa 40500 osservazioni.
Lo strumento con cui valuteremo i risultati di un modello è il tasso di errata classificazione, ovvero la percentuale di osservazione che sono state assegnate dal modello ad una classe della variabile risposta diversa da quella realmente osservata.

Facendo uso del flusso di gestione dei ticket descritto in Figura 1, possiamo selezionare a monte le variabili che possono essere considerate per valutare i modelli di previsione. Essendo che il modello è atto ad immedesimarsi nel cliente in modo da comprendere cosa lo spinge ad assegnare un certo grado di soddisfazione, le variabili utilizzate saranno le sole che il cliente conosce o può osservare: il suo grado in azienda, il grado di severità da lui assegnato al problema, la tipologia di problema e il numero di giorni spesi per risolvere il problema.

3.2 I modelli di classificazione

In questa sezione, proponiamo tre modelli di classificazione per la nuova variabile Satisfaction a due livelli appena definita.

3.2.1 modello lineare generalizzato binomiale

Il primo modello statistico di classificazione che utilizziamo è il modello lineare generalizzato con famiglia binomiale; le variabili che verranno prese in considerazione solo quella sopra descritte; alcuni dettagli teorici sono già stati dati in Sezione 2.1.
Presentiamo prima le stime dei coefficienti ottenute.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.606211	0.172792	-3.508	0.000451
RequestorSeniority2 - Regular	-0.083293	0.027695	-3.008	0.002634
RequestorSeniority3 - Senior	0.011852	0.032826	0.361	0.718058
RequestorSeniority4 - Management	-0.058675	0.032282	-1.818	0.069123
FiledAgainstHardware	1.274926	0.050615	25.189	< 2e-16
FiledAgainstSoftware	0.391693	031158	12.571	< 2e-16
FiledAgainstSystems	0.667024	030663	21.753	< 2e-16
Severity1 - Minor	0.091906	0.184760	0.497	0.618882
Severity2 - Normal	0.657105	0.171331	3.835	0.000125
Severity3 - Major	1.161911	0.177107	6.561	5.36e-11
Severity4 - Critical	1.047961	0.190265	5.508	3.63e-08
daysOpen	-0.074933	0.002114	-35.439	< 2e-16

Tabella 6: modello lineare generalizzato sulla variabile Satisfaction a due classi.

Dai risultati ottenuti possiamo vedere che diverse variabili hanno un impatto significativo sulla valutazione espressa. Per prima cosa, sembra che i dipendenti di fascia Regular abbiano una tendenza minore ad esprimere un buon giudizio sul lavoro volto rispetto ai lavoratori Junior; questo risultato è riscontrabile dal coefficiente di regressione, pari a -0.083. La probabilità di esprimere un giudizio positivo sembra dipendere anche dalla tipologia di problema: i coefficienti associati alle etichette Hardware, Software e Systems risultano essere significativamente diversi dal livello di base Login e vengono tutti stimati come positivi. Da questo risultato possiamo affermare che i clienti ritengono il problema dell'accesso alle proprie risorse informatiche di grande rilevanza, e questo li spinge ad essere fortemente critici nei confronti dell'assistenza (la positività dei coefficienti legata alle altre tre etichette ci spinge a pensare che,

quando il problema non è di accesso, la probabilità di esprimere un giudizio positivo aumenta considerevolmente).
Notiamo infine come coloro che assegnato un grado di severità medio o alto abbiamo una tendenza maggiore ad essere soddisfatti alla fine del lavoro. L'ultimo fattore di rilevante importanza è naturalmente il tempo speso per risolvere il problema; così come potevamo aspettarci, il coefficiente ad esso associato è negativo, segno che al crescere dei giorni la probabilità di essere soddisfatti diminuisce.
Passiamo ora alla previsione sull'insieme di test. Dai risultati in Tabella 7 possiamo notare come il modello fatichi a distinguere le due classi della variabile. In particolar modo sembra che la percentuale di falsi positivi, ovvero di osservazioni classificate come S che in realtà hanno etichetta NS, è molto alta (26.01%); meno elevato è invece il tasso di falsi negativi, pari al 16.78%. Il tasso di errata classificazione complessivo è di 42.8%.
L'obiettivo dei prossimi modelli è quello di migliorare il più possibile questa percentuale.

	NS	S
NS	0.3201	0.1678368
S	0.2601614	0.2518030

Tabella 7: risultati di previsione del modello lineare generalizzato binomiale sull'insieme di test.

3.2.2 Analisi discriminante quadratica

L'analisi discriminante quadratica è una tecnica di classificazione di una variabile a due o più livelli. Come descritto in Azzalini &

Scarpa (2009), essa è un'estensione dell'analisi discriminante lineare e si basa sull'ipotesi di normalità condizionata delle classi, ovvero presuppone che le osservazioni all'interno delle singole classi siano state generate da una normale multivariata di dimensione pari al numero di variabili considerate. Questo concetto può essere espresso in formule come segue.
Sia $x = (x_1, \dots, x_k)$ un vettore di k covariate, e Y_i una variabile aleatoria qualitativa a due o più. L'analisi discriminante quadratica assume quindi che

$$x|Y_i \sim N_k\left(\mu_{Y_i}, \Sigma_{Y_i}\right),$$

ovvero, condizionatamente alla classe, il vettore x è una realizzazione di una variabile aleatoria gaussiana k-variata con media e matrice di covarianza che dipendono dalla classe.
Rispetto alla sua versione lineare, essa assume che ogni livello della variabile risposta identifica una distribuzione normale multivariata con la propria media e la propria matrice di varianze e covarianze, invece di fissare quest'ultima come unica per tutte le categorie. L'analisi discriminante quadratica risulta per tanto più flessibile.
La classificazione di ciascuna osservazione con vettore di covariate x_i avviene attraverso il valore ottenuto dalle funzioni discriminanti quadratiche, ovvero, per ogni livello della variabile Y, si definisce la funzione

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k(x-\mu_k) + \log\pi_k,$$

dove π_k è la probabilità a priori di appartenere alla classe k, che qui assumiamo essere pari a 0.5. Per ogni osservazione valuteremo quindi le k funzioni discriminanti (2 in questo caso) e

classificheremo sulla base di quale delle classi ha ottenuto il punteggio più alto.

I risultati di classificazione sui dati di test sono riportati in Tabella 8, dalla quale notiamo che la capacità classificativa del modello proposto è addirittura minore del modello lineare generalizzato.

	NS	**S**
NS	0.17055930	0.06944643
S	0.40980083	0.35019344

Tabella 8: risultati di previsione dell'analisi discriminante quadratica sull'insieme di test.

In particolar modo, il tasso di falsi positivi è anche qui molto elevato. Questo modello è in grado di cogliere con più precisione i veri positivi, ovvero quei clienti che sono stati classificati come soddisfatti che lo sono veramente, ma contemporaneamente risulta molto debole nell'individuare coloro che non lo sono. Il tasso di errata classificazione finale è del 47.93%.

3.2.3 Albero di classificazione

L'ultimo modello che qui proponiamo è un albero di classificazione, ovvero un modello non parametrico molto flessibile. Così come descritto in Azzalini & Scarpa (2009), lo spazio delle variabili viene partizionato iterativamente in una collezione di regioni $R_1, \dots, R_K$ in modo da identificare gruppi di osservazioni sempre più omogenee, così fino ad ottenere una suddivisione molto estesa (da qui il concetto di albero che si ramifica fino alle foglie).

All’interno di ciascun gruppo identificato viene poi applicata la regola del voto di maggioranza: di tutte le osservazioni presenti, si valuta qual è la classe della variabile risposta che compare più volte. A quella partizione dello spazio verrà quindi associata l’etichetta più presente, e nuove osservazioni che cadranno in quella partizione verranno quindi classificate con quella modalità. Essendo che le partizioni ottenute fanno uso di un molteplice numero di variabili, il modello ad albero è molto efficiente perché riesce anche a valutare l’interazione fra diverse variabili, cosa che fino ad ora non era stata fatta con i modelli precedenti.
Ipoteticamente è possibile costruire un albero con un numero di foglie pari al numero di osservazioni: in questo caso il numero di partizioni dello spazio delle variabili ottenuto sarebbe pari al numero di righe della matrice di dati. Questo tipo di modello non avrebbe in realtà alcun tipo di utilità predittiva, in quanto la sua struttura eccessivamente articolata non lo renderebbe flessibile a nuovi dati. Per questa ragione, il numero di foglie ottimale per la previsione può essere selezionato attraverso diversi criteri in modo da identificare un modello che sia in grado di spiegare l’informazione proveniente dai dati su cui è stato stimato, e contemporaneamente classificare quanto più correttamente possibile le nuove osservazioni che diverranno disponibili in futuro.
Il metodo che qui proponiamo di utilizzare per stimare il modello è quello della crescita e potatura attraverso due diversi insieme di dati. Nello specifico, l’insieme di dati di stima descritto all’inizio di questo capitolo verrà diviso in due parti della stessa dimensione: sulla prima, l’albero verrà fatto crescere in modo da adattarsi

quanto più possibile ai dati, mentre il secondo verrà utilizzato per ridurne la dimensione in base all'errore di classificazione che l'albero commette su di esso.

Passiamo alla stima. I due sottoinsiemi di dati, di crescita e di potatura, contengono ciascuno più di 20200 osservazioni. L'albero di classificazione è stato fatto crescere usando come criterio di split su ciascuna variabile l'indice di Gini, che è definito come segue. Sia R_m una regione con N_m osservazioni al suo interno. Si definisce quindi

$$\hat{p}_{km} = \frac{1}{N_m} \sum_{x_i \in R_m} I(Y_i = k),$$

la proporzione di osservazioni della classe *k* osservate nel nodo *m*. L'indice di Gini è quindi definito come

$$\sum_{k=1}^{K} \hat{p}_{mk} \, (1 - \hat{p}_{mk}).$$

Più nel dettaglio, ad ogni iterazione viene scelta la variabile e il punto ideale in modo tale che la differenza dell'indice di Gini fra quello dell'insieme padre e la somma di quelli ottenuti dai due insiemi figli sia massima rispetto a tutti gli altri possibili nodi.

Osserviamo ora i risultati ottenuti dall'albero sul secondo insieme di dati, quello di potatura. Il grafico in Figura 8 mostra l'andamento dell'errore di errata classificazione sui dati di potatura al variare della dimensione del modello stimato sui dati di crescita.

La linea rossa rappresenta il numero di foglie ottimale, pari a 50, ottenuto attraverso la minimizzazione dell'errore di classificazione sul secondo insieme di dati, quello di potatura. Al

fine di costruire un modello più semplice, scegliamo comunque di costruire un albero con 21 foglie, che nel grafico corrisponde ad un punto di minimo locale della funzione di perdita (linea blu).

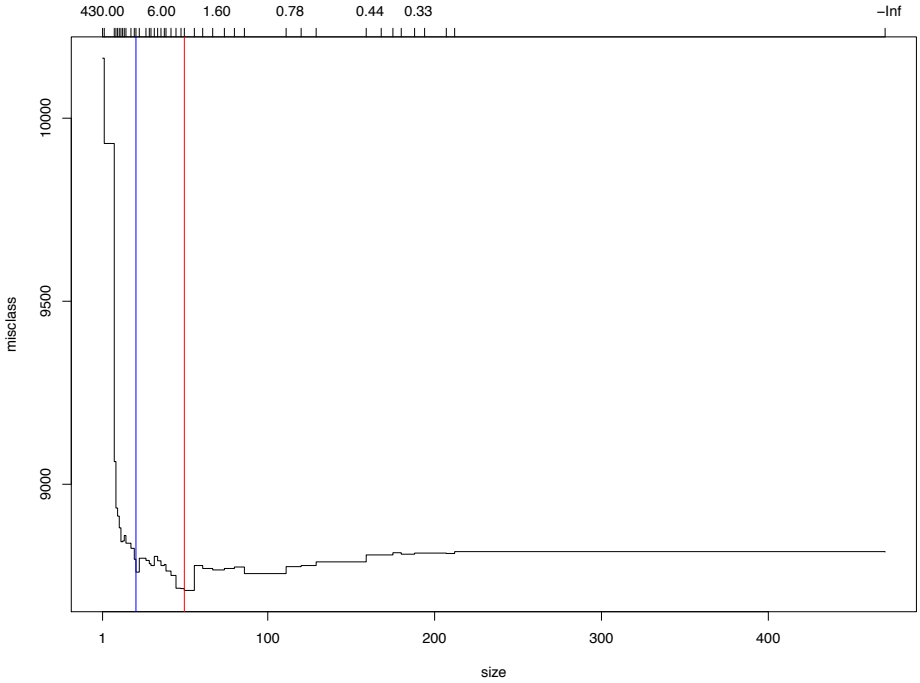

Figura 8: variazione dell'errore di errata classificazione sui dati di potatura al variare della dimensione del modello; la linea in rosso rappresenta la dimensione con la quale di ottiene il minimo assoluto (50 foglie), mentre la linea in blu rappresenta un punto di minimo locale ma che fornisce un albero nettamente più piccolo (21 foglie).

Il modello finale è riportato in Figura 9, dalla quale possiamo vedere come la variabile di maggiore impatto risulti essere il numero di giorni spesi fino alla risoluzione del problema. Questa variabile sembra inoltre essere fortemente legata alla tipologia del problema: come avevamo già visto infatti precedentemente con il

modello lineare generalizzato, coloro che hanno problemi di accesso tendono alla fine del lavoro ad esprimere un giudizio negativo con maggiore probabilità rispetto a coloro che hanno un diverso tipo di necessità. La parte destra dell'albero ci mostra che negli split che interessano la variabile daysOpen nei livelli 4.5, 5.5, 6.5 e 8.5 sono legati alla variabile ticketType, e in particolar modo se il ticket riguarda un problema di accesso allora il modello classifica il cliente come non soddisfatto, altrimenti lo ritiene soddisfatto. Questo tipo di comportamento riflette molto bene, e anzi spiega con maggiore dettaglio quanto abbiamo già osservato in precedenza.

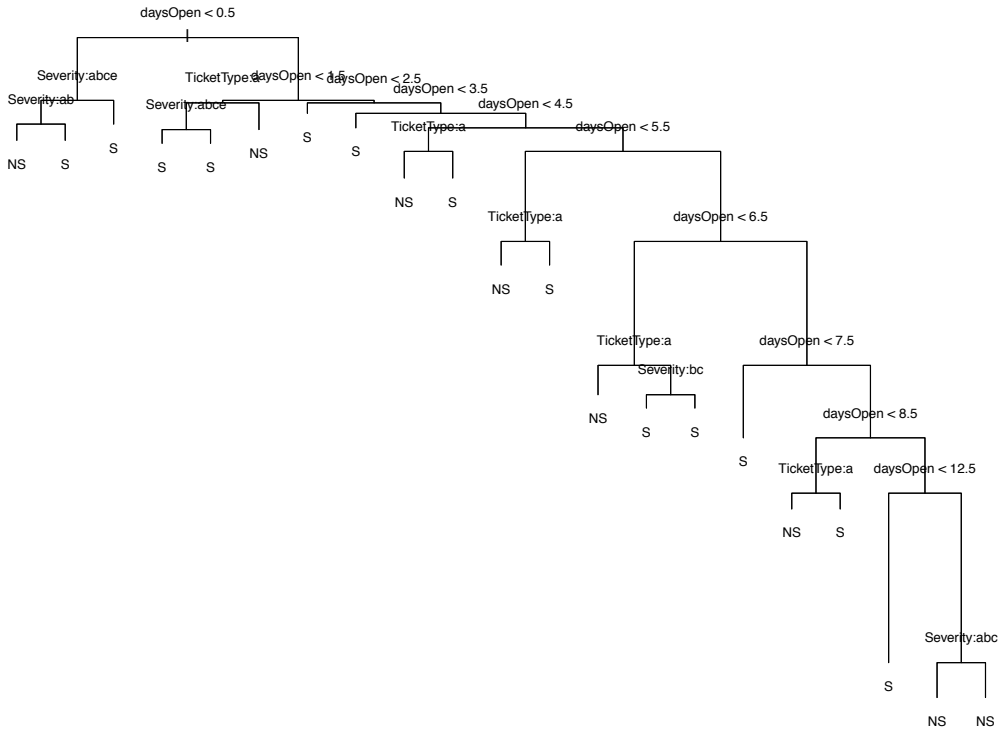

Figura 9: albero di classificazione con 21 nodi.

Andiamo ora a valutare il modello proposto sui dati di test. Dai risultati in Tabella 9 notiamo che il modello, così come gli altri, commette un alto errore di falsi positivi, mentre il numero di falsi negativi resta contenuto. L'errore di classificazione totale è di 0.4536.

	NS	S
NS	0.2580121	0.1312987
S	0.3223480	0.2883412

Tabella 9: risultati di previsione del modello ad albero sull'insieme di test.

Per riassumere, mostriamo in Tabella 10 la percentuale di errata classificazione per i tre modelli proposti. Ciò che emerge alla conclusione di questo capitolo è che la variabile Satisfaction risulta particolarmente difficile da prevedere utilizzando le variabili fornite. La presenza della classe centrale, che esprime un parziale grado di soddisfazione, sembra in qualche modo risultare un confondente per gli altri due livelli; questo spiegherebbe l'elevato tasso di falsi positivi osservati. D'altra parte, abbiamo costruito un albero di classificazione per studiare la variabile Satisfaction a tre livelli; i risultati, che qui non riportiamo, mostrano come il modello proposto non riesca minimamente a riconoscere il livello centrale, corrispondente ad una parziale soddisfazione del cliente.

Ciò che possiamo quindi affermare, in conclusione di questo capitolo, è che l'informazione contenuta in questo insieme di dati

non sembra essere sufficiente per costruire un buon modello di previsione del grado di soddisfazione dei clienti.

Modello Lineare Generalizzato	42.8%
Analisi Disciminante Quadratica	47.9%
Albero di Classificazione	45.4%

Tabella 10: percentuale di errata classificazione ottenuta con i modelli proposti.

Conclusioni

Il seguente elaborato ha avuto lo scopo di analizzare un insieme di dati di un sistema di ticketing in cui il personale, contattato al fine di risolvere problemi informatici, registra le informazioni nel database e procede alla risoluzione dello stesso.
Nel primo capitolo sono state presentate le variabili dell'insieme di dati sulle quali sono state poi effettuate delle analisi descrittive. Da qui abbiamo potuto comprendere che vi è una classe di dipendenti di livello intermedio che tende a fare maggiore richiesta di assistenza rispetto alle altre e che, al di fuori di quanto ci si possa aspettare, la maggior parte degli utenti pone un grado di urgenza di medio livello per i ticket richiesti. Abbiamo inoltre riscontrato che non vi è un alcun legame statistico fra il grado aziendale di un cliente e la tipologia di problema per cui si richiede assistenza.
Il secondo capitolo ha analizzato due aspetti legati al comportamento degli operatori della nostra azienda tramite opportuni modelli statistici. In particolare, nel primo caso abbiamo studiato quali sono le ragioni che spingono un operatore ad attribuire un certo livello di priorità ad un problema. Ne è emerso che i fattori di maggiore rilevanza sono la classe aziendale del richiedente e il grado di severità associata dal cliente al problema. Il secondo aspetto analizzato è stato invece il tempo di risoluzione dei ticket, ed in particolare ci siamo chiesti quali fossero i fattori che più hanno rilevanza sull'intervallo temporale che intercorre fra la presa in carico di un ticket da parte di un

tecnico e la conclusione del lavoro. Dal modello proposto è emerso che questa variabile è influenzata da molti fattori, ed in particolar modo abbiamo riscontrato che maggiore è il grado aziendale del richiedente e maggiore è la priorità assegnata al momento della presa in carico, minore sarà in media il tempo di risposta. Sembra inoltre che i problemi di accesso ai sistemi informatici siano i problemi più veloci da risolvere, mentre i più lunghi sono i problemi hardware.

Nel terzo ed ultimo capitolo abbiamo invece proposto una serie di modelli statistici atti a prevedere il grado di soddisfazione di un cliente una volta completato il lavoro. Sono stati proposti tre modelli di classificazione ed è stato selezionato un modello lineare generalizzato sulla base del minor tasso di errata classificazione. Le variabili più interessanti per la previsione risultano essere il tempo di consegna del lavoro finito e la tipologia di problema. In particolare, risulta che maggiore è il tempo speso per chiudere il ticket e minore è la probabilità di assegnare un giudizio positivo; contemporaneamente, i clienti tendono ad assegnare una valutazione positiva quando i problemi richiesti riguardano l'hardware, il software o il sistema rispetto a quando richiedono assistenza per un problema di accesso, probabilmente in virtù dei disagi che problemi legati a quest'ultima categoria possono portare.

Dobbiamo specificare inoltre che i modelli proposti non riescono ad ottenere un'accuratezza di classificazione sopra al 60% e pertanto riteniamo che vi sia la necessità di avere a disposizione un maggior numero di variabili per condurre un'analisi di tipo predittiva.

Bibliografia

Azzalini, A., & Scarpa, B. (2009). *Analisi dei dati e data mining*. Springer Science & Business Media.

Pace, Luigi, and Alessandra Salvan (2001). *Introduzione alla statistica: Inferenza, verosimiglianza, modelli.* Cedam, 1996.

Stanghellini, Elena. "L'analisi discriminante." *Introduzione ai metodi statistici per il credit scoring*. Springer, Milano, 2009. 87-107.

R Core Team. "R: A language and environment for statistical computing." (2013).

Ver Hoef, Jay M., and Peter L. Boveng. "Quasi-poisson Vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data?." *Ecology* 88.11 (2007): 2766-2772.

Wickham, Hadley. *ggplot2: elegant graphics for data analysis*. Springer, 2016.

Zani, Sergio, and Andrea Cerioli. *Analisi dei dati e data mining per le decisioni aziendali*. Giuffrè editore, 2007.