POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea

Machine learning e Intelligenza artificiale nel Direct Marketing

Relatore

Roberto Fontana

Candidato

Edoardo Roppolo

Anno 2020/2021

Ringraziamento

Il primo pensiero va alla mia famiglia che mi ha supportato e sopportato in questo mio percorso scolastico e di vita, non facendomi mai mancare nulla sotto tutti i punti di vista, senza di loro oggi non sarei qui.
In secondo luogo volevo ringraziare i miei amici con i quali mi sono svagato e divertito in questi anni di studio. Ringrazio la mia stupenda fidanzata ha avuto la pazienza di stare al mio fianco durante questo periodo non facile e mi ha aiutato a raggiungere questo traguardo che non avrei potuto raggiungere senza di lei. Infine colgo l'occasione per ringraziare l'azienda Msx International per avermi dato la possibilità di svolgere il tirocinio e il mio relatore Roberto Fontana per avermi seguito durante questo mio lavoro di tesi.

Edoardo

Indice

Introduzione

L'obiettivo di questa tesi è la presentazione di uno fra i progetti al quale mi sono dedicato durante il mio internship di otto mesi presso Msx International, un'azienda situata a Nanterre, presso Parigi, il cui scopo era verificare l'uso di analisi predittive in ambito customer care per migliorare le prestazioni attuali dell'azienda stessa e i suoi profitti.

Msx opera nel ramo automobilistico, relazionandosi con svariate case produttrici e spaziando in diversi ambiti: dall'assistenza clienti alla creazione di dashboard per concessionarie.

Il primo compito affidatomi, che è il focus della tesi, riguardava la verifica di fattibilità della creazione di un modello predittivo. Tale modello, ricevendo una lista clienti, deve essere in grado di ordinare la suddetta in base alla probabilità di risposta dei clienti alla telefonata di Msx. Questi ultimi, precedentemente selezionati tramite eventi all'interno di concessionarie sparse per tutte Francia, o, più in generale, sulla base dei loro dati vendita, hanno la possibilità di prenotare un appuntamento al fine di revisionare o sostituire componenti appartenenti ai propri veicoli. I clienti possono essere chiamati fino ad un massimo di quattro volte o fino al rifiuto definitivo dell'offerta proposta dall'azienda, ricordando che a monte vi è un processo di validazione in cui i candidati stessi vengono vagliati secondo le leggi vigenti in Francia e regole imposte a priori dal contratto.

Nello specifico mi è stato richiesto di creare un sistema, possibilmente automatizzato, che ricevesse dati grezzi da sottoporre ad un processo ETL e che fosse capace di identificare i clienti propensi ad accettare la proposta tramite chiamata; in seguito il sistema si occupa di creare una tabella oraria in cui i clienti sono suddivisi per agevolare l'assistenza. Il processo, sviluppato come progetto pilota per constatare la possibilità di utilizzo del predictive in Msx ha gettato le basi per partnership future.

Questo progetto rientra nell'ambito del Direct Marketing, una branca di comunicazione commerciale che ha riscosso nel corso del tempo maggior interesse grazie all'evoluzione delle tecnologie e la possibilità di sfruttare queste ultime.

È un tipo di comunicazione commerciale tramite la quale le aziende si interfacciano direttamente con clienti specifici, in questo caso dei privati proprietari di autoveicoli.

Gli albori del Direct Marketing risalgono al secolo scorso con le vendite porta a porta degli anni 50 che in seguito si è evoluto nelle televendite degli anni 70 passando per il telemarketing dei primi anni del nuovo millennio fino alla pubblicità mirata modellizzando un cliente target tramite analisi statistiche e predittive.

Questo mercato è divenuto nel corso del tempo talmente ampio da fondare una associazione, il DMA (Direct Marketing Association), che detta le linee guida deontologiche sull' utilizzo dei dati e della privacy dei clienti.

Nell'area del Direct Marketing è di fondamentale importanza selezionare in modo accurato il target a cui comunicare, per evitare sprechi di risorse economiche e massimizzare il ROI (Return Of Investment) della campagna. In particolare, l'obiettivo principale è quello di individuare, all'interno della Customer Table o di un Prospect Database opportunamente arricchito da informazioni socio-demografiche, gli individui o le imprese che presentano le caratteristiche più simili al consumatore tipo. Una procedura, che normalmente viene utilizzata per ottimizzare la selezione dei candidati ideali all'interno di una base dati, denominata Customer Table, individuando pattern che rappresentano il consumatore tipo della campagna di comunicazione da attivare. Questo procedimento risulta un'alternativa al più semplice approccio delle cosiddette liste verticali, ovvero la selezione di pattern all'interno del Prospect Database secondo criteri definiti a priori, che rispettano il più possibile le caratteristiche proprie o ideali del target.

L'impianto metodologico normalmente utilizzato è basato su modelli di regressione logistica, che sfruttano congiuntamente informazioni individuali e territoriali. La stima dei parametri della funzione logistica è resa possibile dal confronto dei livelli di analogia o differenza tra i caratteri espressivi del target obiettivo e quelli della popolazione nel suo complesso.

L'output del modello denominato score, assegnato a ciascun cliente, è un valore teorico che può essere interpretato come probabilità dei candidati di appartenere, per similitudine, al target obiettivo. In seguito si valuta la bontà del modello, intesa

come capacità di individuare le regole che discriminano l'appartenenza dei singoli al target obiettivo. Viene valutata preliminarmente sul campione di stima, o Training set e, successivamente, su un campione di anagrafiche non utilizzato per la stima, il Validation set. Infine il modello logistico precedentemente stimato e validato consente di qualificare completamente le anagrafiche presenti nel Prospect DB e di valutare il livello di precisione rispetto al target. La regola finale di selezione sarà quella di estrarre i candidati con più alti livelli di score, a garanzia di un maggior livello di similitudine con il target obiettivo tramite l'utilizzo di una soglia.

Questa analisi è stata condotta anche con altri modelli, quali bayesian classifier, support vector machine, neural network e random forest, che sono stati messi a confronto al fine di selezionare il più adatto.

Msx ha designato come perimetro del progetto l'utilizzo dei soli software presenti all'interno del proprio portfolio. A mia disposizione ho potuto sfruttare gran parte delle potenzialità di Alteryx, un software prodotto dall'omonima compagnia, che permette a qualsiasi data worker di preparare, manipolare e analizzare i dati, quindi distribuire e condividere le analisi su vasta scala in modo accessibile. Ha un'interfaccia user friendly e molto intuitiva basata sull'utilizzo di tool, ovvero linee di codice prefatte, che possono anche essere personalizzati o costruiti dagli utenti per successivi utilizzi. I tool vanno a costituire un cosiddetto workflow, un file di progetto. Tale applicativo è uno strumento molto duttile e di infinite potenzialità poiché si può implementare con diversi linguaggi informatici tra cui json, xlm, java, python, r, file command e sql, non ha nessun problema di compatibilità con qualsiasi tipo di database e server; inoltre si può integrare con metodi Apache sparks e api. Ad esempio al contrario di Sas enterprise il codice del workflow può essere tradotto in altri linguaggi oltre alla lingua natia del software.

Il suddetto prodotto, utilizzato da grosse firm a livello globale tra cui Coca Cola, 7 eleven e Levi's, spesso viene affiancato da Tableau, un software che esegue query su database relazionali, cubi di elaborazione analitica online, cloud database, fogli di calcolo per generare visualizzazioni di dati di tipo grafico; da semplici

grafici a molteplici dashboard interattive a scorrimento. Inoltre è possibile implementare il tutto con rappresentazioni geografiche delle informazioni tramite geocoding, creare grafici dinamici. Si possono anche estrarre, archiviare e recuperare dati da un motore di dati in memoria, come nell'esempio seguente.

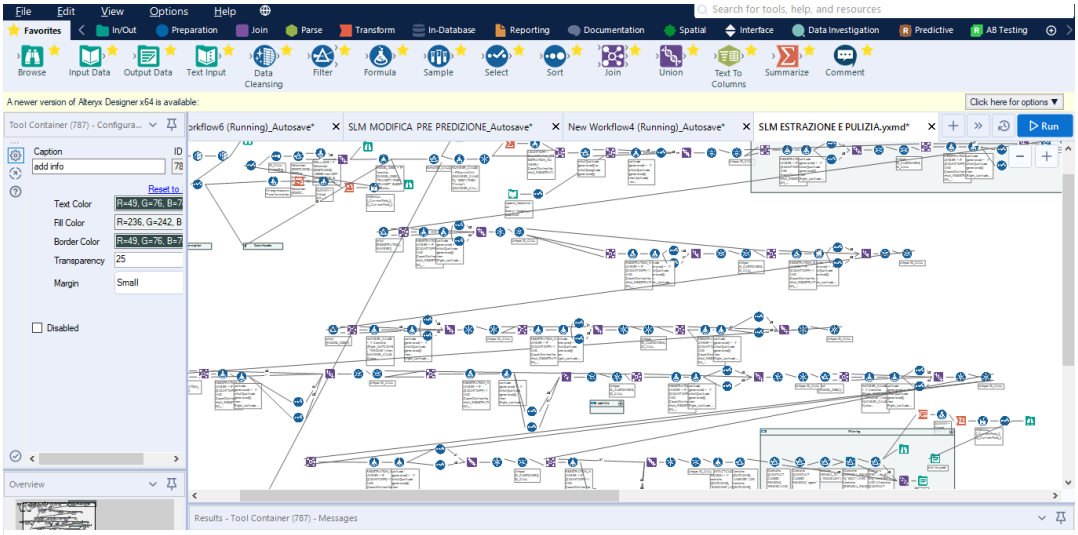
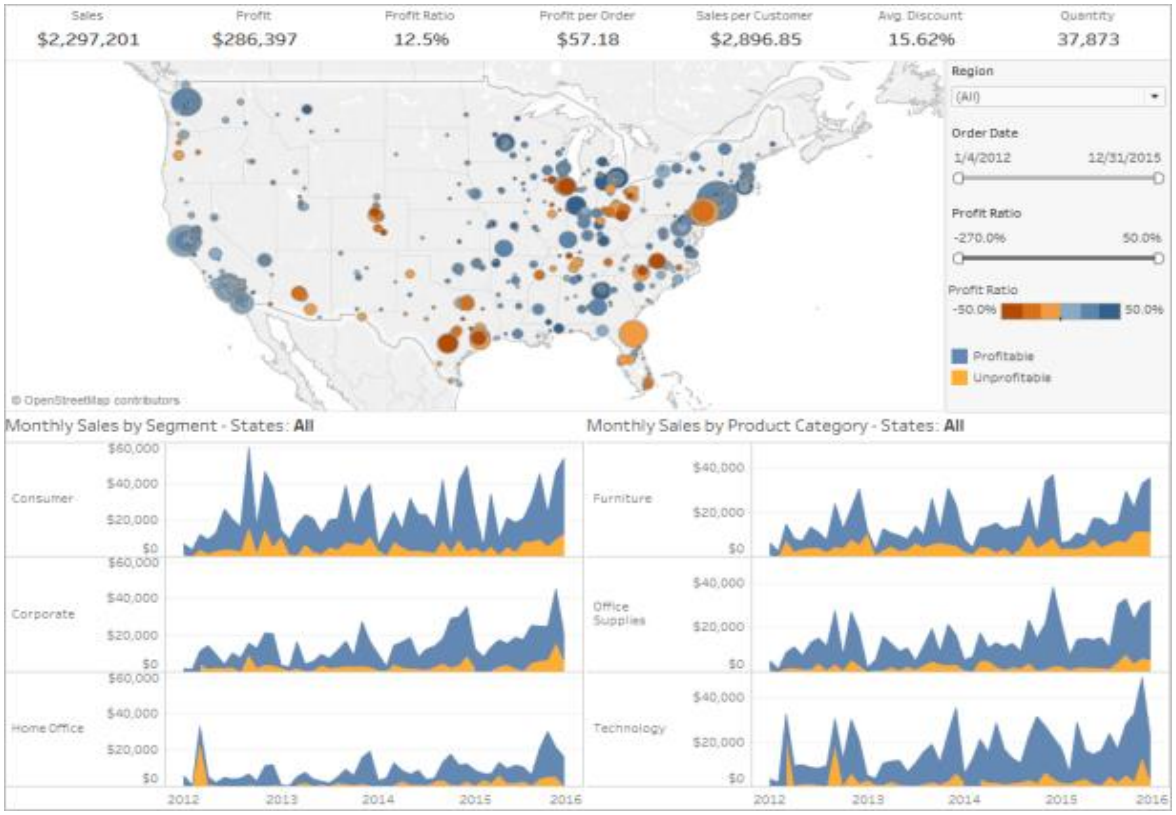


Fig. 1 Esempio di Workflow

Fig. 2 Dashboard[1] di Tableau



[1] Esempio standard: https://help.tableau.com/current/pro/desktop/it-it/dashboards_refine.htm

Inoltre ho adoperato Sharepoint, una piattaforma software di Content Management System (CMS) sviluppata da Microsoft, ovvero un programma che girando lato server permette la creazione e distribuzione di particolari siti web principalmente ad uso aziendale (Intranet), ma che possono anche essere distribuiti in Rete e quindi essere utilizzati come normali siti web. Lo scopo del software, completamente integrato con il pacchetto Microsoft Office, è condividere informazioni e/o documenti in diversi modi. È possibile creare liste, repository documentali, calendari sincronizzati con altri applicativi e offre soluzioni come il "versionamento" dei documenti.

Dal momento che tali documenti sono salvati su server, è possibile lavorare su di essi in collaborazione: più persone possono collegarsi da posti differenti e visualizzare o lavorare sullo stesso documento. In base all'architettura del software, un solo utente alla volta può modificare un certo documento mentre più persone possono visualizzarlo in contemporanea. Il blocco di un documento estratto riguarda solo la possibilità di modificarlo e serve infatti a impedire conflitti nelle modifiche. L'autenticazione avviene inserendo un nome utente e password al momento del login. Questa procedura viene agevolata dal single sign on, che nell'ambito delle tecnologie Microsoft viene spesso inserito al momento dell'accensione del proprio PC. Il sistema si preoccuperà pertanto di autenticare automaticamente l'utente nei siti in cui dispone delle credenziali di accesso.

Questa tesi si struttura con un iniziale insight sull’ estrazione e manipolazione dello storico disponibile al fine di preparare una base dati adatta alla creazione e verifica di diversi modelli predittivi. In seguito si svolgerà un focus sui suddetti che verranno presi in considerazione e comparati. In conclusione verrà presentato un breve excursus sulla pianificazione del sistema automatizzato del processo e la “messa a terra” di questo.

Capitolo 1

Overview

1.1 Insight

Il database di riferimento per questo progetto era costituito da un insieme di dati relativi a tutti i progetti dell'azienda automotive per cui ho sviluppato il sistema. Essendo un file di diversi gygabite ho estratto tramite una query sql tutte le tuple inerenti al progetto e le relative variabili, arrivando ad un file di circa 500 mb costituito da 307.070 contatti definiti chiusi e a 75 variabili, cioè allo storico delle precedenti campagne.
Non erano direttamente presenti i target che d'ora in poi verranno chiamati BOOK, la prenotazione, e CALL, la risposta alla chiamata, ma altre variabili da cui, tramite il reverse engineering, si sono potuti creare.
Successivamente ho analizzato tutti i predittori contenuti nello storico, con l'aiuto di coloro i quali si erano occupati in precedenza della manipolazione dei dati, la quale ha prodotto come risultato il database attuale. Tale analisi mi ha permesso dunque di eliminare le variabili superflue e ridondanti come nel seguente caso:

Esempio
ID_CUSTOMERS, CUPID_ID, ID_CUSTOMER_CONTACT, ID_CUSTOMER_CONTACTS sono tutte variabili che dovrebbero identificare in modo univoco un singolo cliente, per individuarla ho creato una tabella in cui ho riportato il conteggio dei diversi codici identificativi sia come singoli sia in funzione degli altri valori ed ho così selezionato quello che non fosse ripetitivo.

Ho ristretto il numero dei predittori ai seguenti 17:

- ID_CALL: identificativo della chiamata.
- CALL_COMPLETED_DATE: data e ora della chiamata che viene effettuata.

- CONTACT CLOSED REASON: motivo della chiusura del contatto. Se non viene effettuata la chiamata ne restituisce comunque il motivo.
- REFUSAL_REASON: motivo del rifiuto dell'offerta.
- OUTCOME: il risultato della chiamata
- DEALER_CODE: codice della concessionaria a cui si riferisce il cliente.
- EVENT_CUSTOM_1: la campagna a cui si fa riferimento per il cliente:
- ID_CUSTOMERS: identificativo del cliente
- ADDRESS: indirizzo di residenza del cliente.
- POSTCODE: codice postale del cliente
- NUMBER_CALLED: numero di telefono chiamato.
- LAST_VISIT_DATE: ultima visita del veicolo a cui si fa riferimento durante la campagna.
- MODEL_DESC: modello del veicolo.
- ODOMETER: chilometraggio del veicolo.
- REGISTRATION_NUMBER: targa del veicolo
- MONTHS_FROM_LAST_VISIST: mesi dall'ultima verifica, successivamente rinominato Months
- VEHICLE_MONTH_AT_AGE_CALL: mesi di vita del veicolo, di seguito chiamato Vehicle.Month

REFUSAL_REASON e OUTCOME sono le due variabili che mi hanno permesso di creare i target BOOK e CALL, mentre il predittore MODEL_DESC è stato trasformato in un insieme di variabili binarie che rappresentato il segmento del veicolo.

Non sono stati presi in considerazioni ulteriori elementi poiché le suddette racchiudono le informazioni essenziali, sia del cliente sia del relativo veicolo per il nostro caso di studio; inoltre costituiscono la parte dei dati forniti dall'azienda i quali possono essere utilizzati senza violare la privacy per la previsione mentre le informazioni relative alla chiamata servono per l'analisi procedurale ed una valutazione e miglioramento delle performance.

1.2 Missing values

Poiché in diversi casi vi sono dei missing values ho cercato, ove possibile e sensato, di riempire le lacune prendendo ad esame tutti i database relativi all'assistenza clienti, in quanto alcuni tra i candidati potevano figurare anche in database relativi a progetti di altre aziende. In seguito, il risultato da me ottenuto è stato utilizzato come base dati per tutti i progetti relativi al call center. Ciò è stato possibile grazie all'utilizzo di un codice Apache per l'estrazione dai vari database[2] con l'utilizzo delle funzioni fornite da MapReduce[3], che si possono implementare all'interno di workflow in Alteryx tramite un apposito tool.

Per quanto concerne i veicoli invece, dopo aver mappato i diversi modelli nei rispettivi segmenti con una categoria a parte per suv e veicoli commerciali, ho creato differenti variabili binarie per ogni taglia dei veicoli tramite il metodo OneHotEncoder, disponibile nella libreria sklearn di pyton, concentrandomi poi esclusivamente sulla stima del chilometraggio, se assente.

Il seguente processo è stato effettuato escludendo gli outlier da esso; questi ultimivengono descritti nel seguente capitolo insieme a quelli degli altri attributi.

Ho creato una simulazione della distribuzione del suddetto, utilizzando i mesi del veicolo dall'immatricolazione, la taglia e i chilometri già percorsi, sulla base delle considerazioni formulate da Patrick Plötza, Niklas Jakobssonb e Frances Sprei[4] (nel loro articolo cercano una distribuzione per il chilometraggio giornaliero dei veicoli), giungendo ad una Weibull con $\lambda = 1$ e $k = 0.8$, (Fig.1).

[2]Script Apache per l'estrazione dei dati: https://spark.apache.org/docs/2.0.0/mllib-feature-extraction.html
[3] MapReduce: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
[4] On the distribution of individual daily driving distances, July 2017,links:
https://www.sciencedirect.com/science/article/pii/S0191261516309067?via%3Dihub

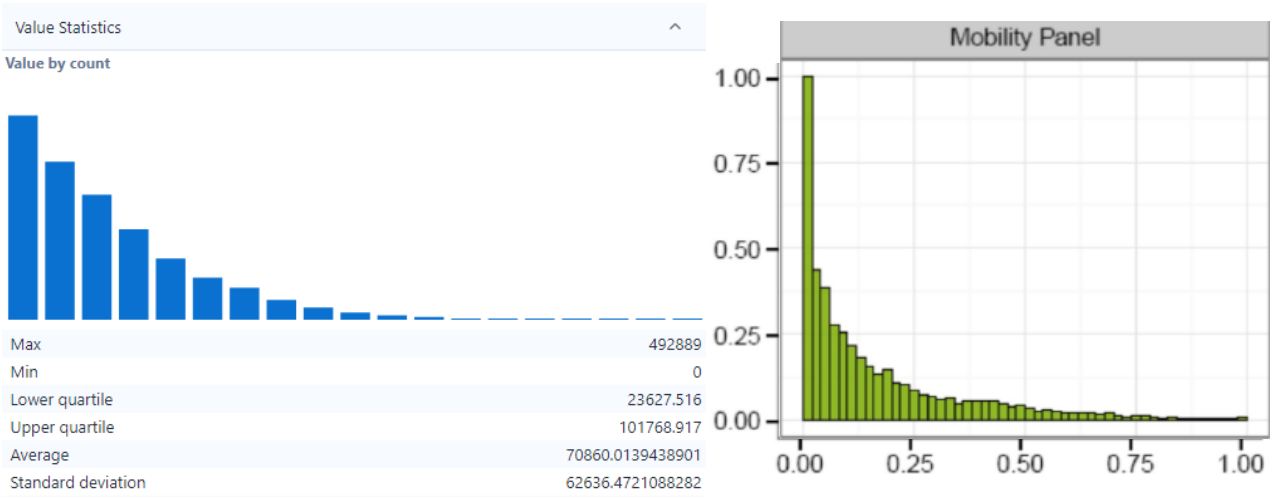


Fig 1. Confronto della distribuzione dei dati in esame e quelli utilizzati nell'articolo di nota 3

	AIC	RMSE	χ^2
Godness of Fit	63.24 %	87.425 %	91.357 %

Tale risultato si discosta dalle valutazioni presenti nell'articolo, in cui si predilige una distribuzione log – Normale, ma essa, in questo caso, fornisce risultati peggiori.

La distribuzione di Weibull[5], che prende il nome dal matematico svedese Waloddi Weibull, ma trattata in precedenza dal matematico francese Maurice Fréchet nel 1927, è una distribuzione di probabilità continua definita sui numeri reali positivi e descritta da due parametri strettamente positivi: λ (parametro di scala o vita caratteristica) e k (parametro di forma).

Tale distribuzione fornisce un'interpolazione tra la distribuzione esponenziale (per λ = 1), la distribuzione di Rayleigh (per k = 2).

Ha in questo caso una funzione di ripartizione pari a:

$$F(x) = 1 - e^{\left(\frac{x}{1.1}\right)^{1.2}}$$

[5] Horst R., *The Weibull Distribution*: A Handbook, CRC Press (2008)

Questa previsione è stata affiancata da un control paramater binario, opportunamente eliminato prima dell'analisi dei dati, per permettere un'iterazione del processo di arricchimento del database tenendo in considerazione soltanto i dati che avevano un diretto precedente dato storico di riferimento che non fosse una previsione.

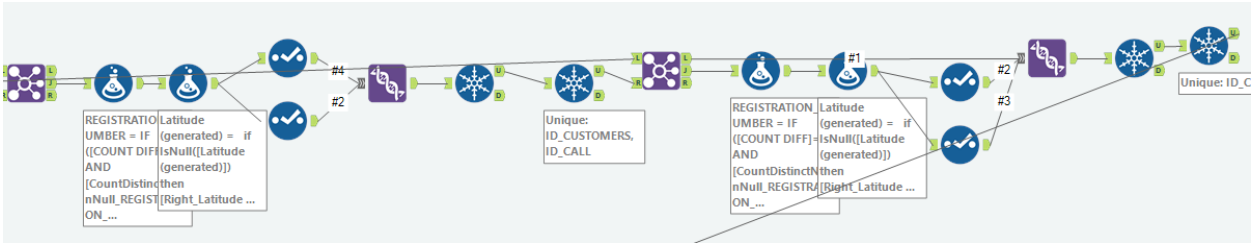


Fig 2. Blocco di workflow per l'arricchimento del database

1.3 Mappatura ed arricchimento dei dati

Poiché le campagne di selezione dei candidati si distinguevano in base alla nomenclatura delle suddette, specificando inoltre se fossero assistenza-garanzia e assistenza semplice, ho deciso di mapparle solo per il tipo di intervento da effettuare sui veicoli nel seguente modo:

TIPOLOGIA CAMPAGNA						
VUSEPT	IC	BS	MOT	OEW	WL	RS
altro	sostituzioni	manutezione	motore	ricambi	meccanica	bollo

Per quanto riguarda i numeri di telefono mi è stato richiesto di diversificarli in base al prefisso telefonico in quanto in Francia quest'ultimo determina se si tratta di un apparecchio fisso o di un dispositivo mobile, non considerando la portabilità del numero nel seguente schema:

Fisso	Mobile
03, 04	05, 06, 07

Proseguendo nella modifica delle variabili ho scaricato tramite il sito ufficiale delle poste francesi il database dei comuni[6] con relativo codice postale, in modo da implementare la posizione delle concessionarie e ho generato tramite l'API di Google Maps le coordinate geografiche dell'indirizzo di casa o il codice postale del comune, in assenza del primo, e di conseguenza ho calcolato la distanza dei clienti dalle concessionarie. Il risultato ottenuto è stato utilizzato sia nella costruzione del modello sia nel miglioramento delle performance.

Inoltre, dal momento che nell'elenco delle chiamate ne sono presenti alcune effettuate a vecchi proprietari di veicoli e a società, che per contratto non possono aderire alle offerte, è stato creato un database a sé stante che verrà utilizzato durante il processo di selezione per eliminare tali candidati prima della predizione e non sono stati considerati nel database per la creazione dei modelli.

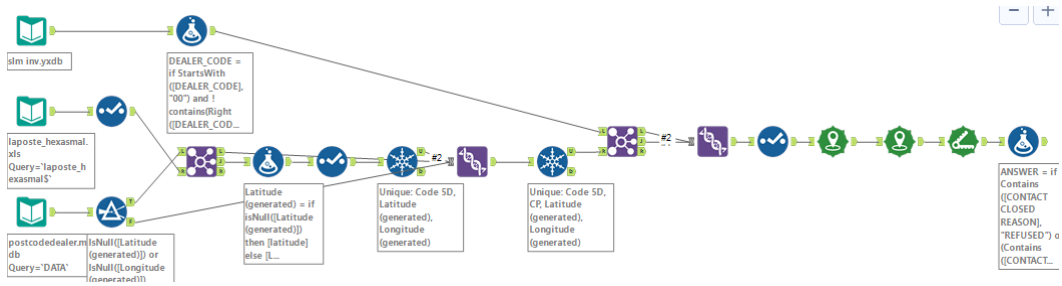


Fig 3. Workflow per la creazione della distanza

[6] Link di riferimento del database dei comuni-codici postali sul sito delle poste francesi: https://datanova.laposte.fr/explore/embed/dataset/laposte_hexasmal/table/?disjunctive.code_commune_insee&disjunctive.nom_de_la_commune&disjunctive.code_postal&disjunctive.ligne_5

Capitolo 2

Analisi dei dati

2.1 Outlier

Prima di creare i modelli predittivi ho analizzato i vari predittori in funzione delle due variabili BOOK e CALL ed effettuato diversi test che verranno presentati suddivisi tra continue e discrete.

Per poter effettuare tale analisi ho analizzato il database alla ricerca di outlier negli attributi riferiti al veicolo e la distanza tra i clienti e la concessionaria di riferimento eliminando un totale di circa cinque mila osservazioni, pari a quasi il 3% dei dati totali, settando come soglie la distanza tra la residenza del cliente e la concessionaria non superiore a 800 chilometri, un chilometraggio dei veicoli non superiore a 500.000 chilometri ed un età dei veicoli non superiore a 10 anni. I dati più anomali riguardo alla prima soglia si riferiscono a veicoli immatricolati in Francia e situati oltre oceano, mentre per la seconda veicoli commerciali con circa 1.000.000 di chilometri percorsi, la terza soglia è stata posta come parametro di interesse a priori dall'azienda.

Max	492889
Min	1
Lower quartile	21842.531
Upper quartile	74702.262
Average	53339.1381042475
Standard deviation	42054.2641875547

Fig. 1 Chilometraggio post rimozione Outlier

2.2 Variabili categoriche

2.2.1 Cramer's V

Per quanto riguarda le variabili categoriche ho utilizzato la Cramer's V[7], uno strumento, creato da Harald Cramer nel 1946, utile per il confronto di più variabili nominali che si basa sul test χ^2 di Pearson. Non essendo influenzato dalle dimensioni del campione è molto utile in situazioni in cui si sospetta che un chi-quadrato statisticamente significativo sia il risultato delle dimensioni del campione anziché da qualsiasi relazione sostanziale tra le variabili.

Viene interpretato come una misura della relativa forza di un'associazione tra due variabili. Il coefficiente varia da 0 a 1, che rappresenta l'associazione perfetta, ed un valore superiore a 0.1 fornisce una buona soglia minima per suggerire che esiste una relazione sostanziale tra due variabili. Orientativamente se il valore ottenuto è compreso tra 0.1 e 0.3 si ha una bassa connessione, da 0.3 a 0.6 si ha una buona connessione, da 0.6 a 1 si ha un'ottima connessione. Inoltre il p-value di questo test coincide con quello del χ^2di Pearson poiché è basato su di esso. La formula per ottenere il valore desiderato è il seguente:

$$V = \sqrt{\frac{\chi^2}{n\,q}}$$

dove n è la dimensione del campione, la χ^2 deriva appunto dal test di Pearson e q la dimensione minima tra il numero di righe e colonne meno 1. Può essere uno stimatore influenzato dal bias per quanto riguarda la popolazione e tenderà a sovrastimare la forza dell'associazione, per questo si può usare una correzione del bias, sostituendo a q il valore calcolato come

$$\hat{q} = q - \frac{q^2}{n - 1}$$

[7] Wu B., Zhang L. and Zhao Y., *Feature Selection via Cramer's V-Test Discretization for Remote-Sensing Image Classification*, Transactions on Geoscience and Remote Sensing, vol. 52, no. 5, 2014

Prenotazione	
Field	**Cramer's V**
FFRAVUSEPT	**0,0187**
A	**0,0149**
B	0,0069
C	**0,0112**
COMMERCIAL	0,0038
D	0,0038
SUV	0,0027
ANSWER	**0,0434**
IC	**0,0125**
BS	**0,0649**
MOT	**0,0227**
OEW	**0,0286**
WL	**0,0221**
RSNF	0,0011
telephone	0.0048

Risposta	
Field	**Cramer's V**
FFRAVUSEPT	0,0007
A	0,0007
B	0,0030
C	0,0002
COMMERCIAL	0,0001
D	0,0049
SUV	0,0043
IC	0,0005
BS	**0,0119**
MOT	**0,0218**
OEW	0,0085
WL	0,0066
RSNF	0,0036
Telephone	**0.0653**

Come si può osservare dalle tabelle soprastanti esistono alcune relazioni, evidenziate in grassetto, che spiccano rispetto ad altre per un fattore di dieci volte superiore, ma nessuna di esse permette di individuare un legame quantomeno significativo tra i predittori e le variabili di interesse.

Non sono state testate le interazioni tra i predittori poiché non è stata raggiunta la soglia minima per permettere l'ipotesi di una interazione tramite questo strumento.

2.2.2 Welch T-test[8]

Ho scelto poi di concentrarmi sulla variabile categorica che rappresenta la taglia dei veicoli effettuando un' analisi tramite il t-test, uno strumento utile per verificare se la differenza fra le medie dei gruppi fosse significativa tramite il confronto di un' ipotesi nulla, $\mathbf{H}_0 \coloneqq \mu_a - \mu_b = 0$, ovvero che le medie dei gruppi siano uguali contro l'ipotesi alternativa che non lo siano, $\mathbf{H_a}$, tramite la statistica:

[8] Desiderio V. J, *Handbook of Trace Evidence Analysis*, John Wiley & Sons, 2020

$$\boldsymbol{T}_{\text{test}} = \frac{\bar{X}_{\text{a}} - \bar{X}_{\text{b}}}{\sqrt{\frac{S^2{}_{\text{a}}}{n_{\text{a}}} + \frac{S^2{}_{\text{b}}}{n_{\text{b}}}}}$$

Questi test vengono tipicamente applicati quando le unità statistiche sottostanti i due campioni confrontati non si sovrappongono.

Il t-test di Welch è più robusto del t-test di Student e mantiene tassi di errore di tipo I vicini al nominale per varianze disuguali e per dimensioni del campione disuguali in condizioni di normalità. Inoltre, la potenza del t-test di Welch si avvicina a quella del t-test di Student, anche quando le varianze della popolazione sono uguali e le dimensioni del campione sono bilanciate. Il t-test di Welch può essere generalizzato a più di 2 campioni, come in questo caso, ed è più robusto dell'analisi della varianza a una via.

Secondo [9] non è consigliabile eseguire un test preliminare per varianze uguali e quindi scegliere tra il t-test di Student o di Welch. Piuttosto, il t-test di Welch può essere applicato direttamente e senza alcuno svantaggio sostanziale rispetto al t-test di Student come indicato sopra. Il t-test di Welch rimane robusto per distribuzioni distorte e campioni di grandi dimensioni. L'affidabilità diminuisce per distribuzioni distorte e campioni più piccoli.

Attraverso questo processo ho potuto individuare le differenze significative tra i vari segmenti di veicoli, in grassetto nella tabella sottostante, insieme al valore del t-test e ai gradi di libertà.

Mi sono solamente soffermato ad osservare i risultati ottenuti senza modificare ulteriormente i dati, bensì tenendoli in considerazione per un'eventuale modifica o eliminazione futura dei gruppi dopo l'elaborazione del modello.

Lo stesso procedimento è stato applicato alle varie campagne di selezione sotto riportate:

[9] Zimmerman, D. W., & Zumbo, B. D., *Rank transformations and the power of the Student t test and Welch t' test for non-normal populations with unequal variances*. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, (1993).

BOOK				
Group1	Group2	t_test	df	p_value
A	D	4.9126	19778	9.0556e-07
A	B	4.0987	17105	4.1731e-05
A	COMM	5.0698	23264	4.0122e-07
A	SUV	4.1491	21573	3.3500e-05
A	C	6.6459	17462	3.1018e-11
A	UKWN	6.1034	12827	1.0676e-09
B	D	2.3615	11525	0.0182
B	COMM	2.4234	15756	0.0153
B	SUV	0.5163	58916	0.6056
B	C	4.5827	105287	4.5947e-06
B	UKWN	4.0136	6981	6.0408e-05
C	D	-0.1087	11757	0.9133
C	COMM	-0.3041	16122	0.7610
C	SUV	-3.2733	59317	0.0010
C	UKWN	1.8517	7091	0.0640
D	COMM	0.1352	18414	0.8924
D	SUV	-1.9034	14492	0.0570
D	UKWN	1.5713	12406	0.1161
SUV	COMM	-1.9038	20334	0.0569
SUV	UKWN	3.5447	8420	0.0003
COMM	UKWN	1.7692	11872	0.0768

CALL				
Group1	Group2	t_test	df	p_value
WL	RS	-7.8846	68525	3.2009e-15
WL	BS	-24.7832	54085	7.7062e-135
WL	OEW	3.9498	26675	7.8408e-05
WL	MOT	-3.7351	61045	0.0001
WL	IC	2.3655	1950	0.0181
WL	USEPT	-17.5567	5294	4.0789e-67
MOT	RS	5.6900	106168	1.2736e-08
MOT	BS	-29.2704	121950	1.1081e-187
MOT	OEW	-6.8492	19320	7.6445e-12
MOT	IC	3.2824	1852	0.0010
MOT	USEPT	16.7764	3713	6.0394e-61
RS	BS	-19.8275	85955	2.6999e-87
RS	OEW	9.6927	21644	3.5922e-22
RS	IC	4.3779	1882	1.2636e-05
RS	USEPT	13.2885	4175	1.6925e-39
BS	OEW	-20.7846	18406	7.3154e-95
BS	IC	-8.1337	1841	7.5694e-16
BS	USEPT	3.3965	3546	0.0006
OEW	IC	0.8231	2216	0.4105
OEW	USEPT	17.8693	9426	2.9259e-70
IC	USEPT	8.8992	2222	1.1393e-18

2.3 Variabili continue

2.3.1 Kullback-Leibler

Nel caso delle variabili continue mi sono incentrato su altri tipi di test; il primo è stato la divergenza di Kullback–Leibler[10], o in ambito machine learning detto comunemente information Gain. É un metodo statistico che misura l'aumento delle informazioni su una variabile, rappresentato dalla funzione entropia di questa ultima e dall'osservazione di un'altra, calcolata grazie all'entropia della prima condizionata dalla seconda, tramite la formula nella sua versione discreta

$$information\ gain = H(Y) - H(Y|\,X)$$
$$H(Y|X) = -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x) \log_2\big(p(y|x)\big)$$
$$\mathrm{H(Y)} = -\sum_{\mathrm{y}\in Y} \mathrm{p}(y) \log_2 p(y)$$

2.3.2 Gain Ratio

Il gain ratio[11] è il rapporto tra l'information gain e l'informazione intrinseca, in questo frangente, dell'intero dataset in considerazione. É uno strumento usato per ridurre il bias verso attributi multi-valued prendendo in considerazione il numero e la grandezza dei possibili valori durante la scelta di un attributo.
Il rapporto di guadagno di informazioni distoglie l'albero decisionale dal considerare attributi con un numero elevato di valori distinti risolvendo lo svantaggio del guadagno di informazioni, vale a dire, il guadagno di informazioni applicato ad attributi che possono assumere un gran numero di valori distinti potrebbe creare un modello overfitted nel training. Ad esempio, supponiamo di creare un albero decisionale per alcuni dati che descrivono i clienti di un'azienda. Il guadagno di

[10] Hughes Anthony William, An Iterative Approach to Variable Selection Based on the Kullback-Leibler Information, School of Economics, 1997

[11] Desiderio Vincent J, Handbook of Trace Evidence Analysis, John Wiley & Sons, 2020

informazioni viene spesso utilizzato per decidere quali attributi sono i più rilevanti, in modo che possano essere testati vicino alla radice dell'albero. Uno degli attributi di input potrebbe essere il numero di carta di credito del cliente. Questo attributo ha un elevato guadagno di informazioni, perché identifica in modo univoco ogni cliente, ma non vogliamo includerlo nell'albero decisionale: è improbabile che decidere come trattare un cliente in base al numero della sua carta di credito sia generalizzato a clienti che non abbiamo visto prima.

L'informazione intrinseca viene calcolata come l'entropia ma sui predittori tramite la seguente formula:

$$\mathrm{IV}(\mathrm{X}) \coloneqq -\sum_{x \in \mathrm{X}} p(x) \log_2 p(x)$$

2.3.3 Symmetrical uncertainty

La Symmetrical uncertainty[12], ovvero la normalizzazione dell'information gain rispetto alle dimensioni al fine di avere una misura più accurata dell'importanza dei predittori é ottenuta dalla formula seguente:

$$\mathrm{SU} = 2.0 \ \frac{\mathrm{gain}}{\mathrm{H(Y)} + \mathrm{H(X)}}$$

2.3.4 Relief

Infine il Relief[13], un algoritmo sviluppato nel 1992 che adotta un metodo di filtro per la selezione delle variabili che è particolarmente sensibile alle interazioni tra di esse. È stato originariamente progettato per l'applicazione a problemi di classificazione binaria con caratteristiche discrete o numeriche calcolando uno score per ogni variabile che può quindi essere applicato per ordinare e selezionare le mi-

[12] Mark A. Hall, *Correlation-based Feature Selection for Machine Learning*, 1999,

[13] Algoritmo disponibile al link: https://pypi.org/project/ReliefF/#files,
Sammut Claude, *Machine Learning: Proceedings of the Nineteenth International Conference,* Morgan Kaufmann Publishers, 2002

gliori. In alternativa, questi punteggi possono essere applicati come pesi per i predittori con lo scopo di guidare la costruzione del modello. Il punteggio si basa sull' identificazione delle differenze di valore delle feature tra le coppie di istanze più vicine; nel caso si osservi una differenza di istanze vicine con la stessa classe, il punteggio diminuisce, mentre se appartengono a classi diverse il punteggio aumenta. L'algoritmo non dipende dall'euristica, "corre" in tempi polinomiali di basso ordine, è tollerante al rumore ed è robusto per caratterizzare le interazioni, tuttavia non discrimina tra variabili ridondanti e un basso numero di istanze di training ingannano l'algoritmo. Le seguenti tabelle includono i valori ottenuti dai test effettuati, come si può notare nessuno tra essi restituisce valori significativi.

	BOOK			
Field	**Information Gain**	**Gain Ratio**	**SU**	**Relief**
KM	0,0014	0,0012	0,0020	0.0345
Months	0,0025	0,0014	0,0025	0.0563
Vehicle.Month	0,0021	0,0013	0,0023	0.0169
Distance	0,0006	0,0007	0,0012	0.0277

	CALL			
Field	**Information Gain**	**Gain Ratio**	**SU**	**Relief**
KM	0	0	0	0.0345
Months	0	0	0	0.0563
Vehicle.Month	0	0	0	0.0169
Distance	0.0001	0.0007	0.0007	0.0277

2.4 Correlazioni

2.4.1 Pearson Coefficient

Ho utilizzato il calcolo del Pearson correlation[14] coefficient, strumento utile per misurare la correlazione lineare tra due variabili. Il quale assumendo valori nell'intervallo [-1, +1]; è il rapporto tra la covarianza delle due variabili e il prodotto delle relative deviazioni standard; ovvero si calcola tramite la formula:

$$\rho_{xy} = \frac{\mathrm{cov}_{xy}}{\sigma_x \sigma_y}$$

2.4.2 Spearman's rank

Inoltre ho calcolato la Spearman's rank correlation[15] coefficient o Spearman's ρ, è una grandezza non parametrica che permette di stabilire quanto bene una relazione tra due variabili può essere descritta usando una funzione monotona, è un caso particolare del coefficiente di correlazione di Pearson dove i valori degli attributi vengono convertiti in ranghi prima di calcolare il coefficiente.

Le tabelle che seguono mostrano i rispettivi i coefficienti di correlazione dei predittori rispetto alla prenotazione e alla risposta alla chiamata insieme alla significatività statistica.

Per tutti gli altri coefficienti si vedano le tabella nell'appendice A.

[14] Sammut Claude, *Machine Learning: Proceedings of the Nineteenth International Conference*, Morgan Kaufmann Publishers, 2002

[15] Zill Dennis G. , *Advanced Engineering Mathematics*, Jones & Bartlett Learning, 2020

BOOK Pearson coefficient			
Field	**Coefficient**	**p-value**	
Answer	0.1338	0.0000e+00	***
BS	-0.0644	0.0000e+00	***
Months	-0.0483	0.0000e+00	***
Vehi-cle.Month	-0.0471	0.0000e+00	***
KM	-0.0437	0.0000e+00	***
OEW	0.0288	0.0000e+00	***
MOT	0.0279	0.0000e+00	***
Distance	-0.0277	0.0000e+00	***
FFRAVUSEPT	-0.0196	6.8834e-15	***
WL	0.0185	1.7231e-13	***
A	0.0144	1.1506e-08	***
IC	0.0119	2.2539e-06	***
C	-0.0089	3.6922e-04	***
B	0.0049	5.1601e-02	.
telephone	0.0048	5.5122e-02	.
RSNF	-0.0033	1.8338e-01	
COMMERCIAL	-0.0018	4.5222e-01	
D	-0.0014	5.7429e-01	
SUV	-0.0010	6.7827e-01	

BOOK Spearman rank			
Coefficient	**p-value**		
0.1338	0.0000e+00	***	
-0.0644	0.0000e+00	***	
-0.0492	0.0000e+00	***	
-0.0461	0.0000e+00	***	
-0.0423	0.0000e+00	***	
0.0288	0.0000e+00	***	
0.0279	0.0000e+00	***	
-0.0275	0.0000e+00	***	
-0.0196	6.8834e-15	***	
0.0185	1.7231e-13	***	
0.0144	1.1506e-08	***	
0.0119	2.2539e-06	***	
-0.0089	3.6922e-04	***	
0.0049	5.1601e-02		
0.0048	5.5122e-02		
-0.0033	1.8338e-01		
-0.0018	4.5222e-01		
-0.0014	5.7429e-01		
-0.0010	6.7827e-01		

CALL Spearman rank			
Field	**Coefficient**	**p-value**	
telephone	0.0653	0.0000e+00	***
MOT	-0.0215	0.0000e+00	***
BS	0.0119	2.0917e-06	***
OEW	0.0075	2.8002e-03	**
WL	0.0073	3.6168e-03	**
KM	0.0056	2.5208e-02	*
D	-0.0040	1.0727e-01	
Months	0.0030	2.3370e-01	
B	-0.0029	2.4012e-01	
Distance	0.0029	2.4965e-01	
RSNF	0.0028	2.6189e-01	
SUV	0.0025	3.1296e-01	
COMMERCIAL	0.0010	6.8471e-01	
C	-0.0009	7.1857e-01	
Vehicle.Month	-0.0007	7.6845e-01	
A	0.0005	8.2597e-01	
IC	0.0001	9.3766e-01	
FFRAVUSEPT	0.0001	9.3890e-01	

CALL Pearson rank			
Field	**Coeffiecient**	**p-value**	
telephone	0.0653	0.0000e+00	***
MOT	-0.0215	0.0000e+00	***
BS	0.0119	2.0917e-06	***
Distance	0.0114	5.3502e-06	***
OEW	0.0075	2.8002e-03	**
WL	0.0073	3.6168e-03	**
KM	0.0054	3.0814e-02	*
D	-0.0040	1.0727e-01	
B	-0.0029	2.4012e-01	
RSNF	0.0028	2.6189e-01	
SUV	0.0025	3.1296e-01	
Months	0.0018	4.6795e-01	
Vehicle.Month	-0.0011	6.4939e-01	
COMMERCIAL	0.0010	6.8471e-01	
C	-0.0009	7.1857e-01	
A	0.0005	8.2597e-01	
IC	0.0001	9.8766e-01	
FFRAVUSEPT	0.0001	9.9890e-01	

2.4.3 Hoeffding D

Ho calcolato la Hoeffding D statistic[16], che permette di calcolare la differenza tra i ranghi congiunti di due variabili e il prodotto dei loro ranghi marginali. A differenza delle già utilizzate misure può cogliere relazioni non lineari. Può essere utilizzata per testare l'indipendenza, dovrebbe essere applicato solo a dati che hanno una distribuzione continua a causa dei problemi di calcolo della distribuzione cumulativa F_{xy}.

Si ottiene un valore appartenente all'intervallo [-0.5; 1] se non ci sono ranghi in parità, con valori più grandi che indicano una relazione più forte tra le variabili mentre cifre prossime allo zero sono sintomo di un legame tra gli attributi.

Questo test rispetto agli altri è molto costoso in termini di tempo, infatti per ottenere dei risultati ci sono volute circa 8 ore.

$$HG = \int \left(F_{xy} - F_x F_y\right)^2 dF_{xy}$$

Le seguenti tabelle includono i valori ottenuti dai test effettuati, come si può notare nessuno tra essi restituisce valori significativi.

	HG	
Field	**CALL**	**BOOK**
KM	0.0027	0.0021
Months	0.0021	0.0020
Vehicle.Month	-0.0018	-0.0018
Distance	0.0012	0.0019

	HG tra variabili continue		
Field	**Months**	**Vehicle**	**Distance**
KM	0.0009	0.0021	0.0010
Months		0.0003	0.0005
Vehicle.Month			0.0089

[16] Hoeffding (1948). A non-parametric test of independence
https://projecteuclid.org/download/pdf_1/euclid.aoms/1177730150

Capitolo 3

Costruzione e comparazione dei modelli

Dato che si richiedeva di fare una previsione dei clienti che avrebbero risposto alle chiamate e una stima della probabilità di accettare l'offerta, inizialmente ho provato a costruire un modello unico che mi permettesse di ottenere entrambi le previsioni di interesse ma questo si è dimostrato non efficiente e con un periodo di calcolo alquanto esiguo, per cui la struttura che ho scelto è costituita da due diversi modelli:
Il primo modello che ho creato prevede se un cliente accetti l'offerta supponendo che abbia risposto alla chiamata, in tal modo ho potuto implementare i predittori usati in esso, come un kernel, per costruire il secondo modello che prevede la risposta alla chiamata.
Ho considerato diversi tipi di modelli:

3.1 Logistic Regression

Il modello logistico[17] o logit prende il nome dalla funzione che viene utilizzata per stimare la probabilità di una certa classe o evento binario.
La regressione logistica nella sua forma base utilizza appunto la funzione logistica

$$\text{logit}(x) = \log\left(\frac{e^{\beta x}}{1 - e^{\beta\, x}}\right)$$

che permette di modellare e convertire le probabilità logaritmiche in probabilità o score della variabile dipendente.
A ognuno dei due attributi della classe viene assegnata una probabilità compresa tra 0 e 1, con somma delle due probabilità di uno che permette, tramite l'utilizzo di una soglia, la stima della classe, poiché non è un metodo di classificazione statistica.

[17] Marwick P., *Implementation of the N-dimensional Logit Model*, Mitchell & Co (1972)

Possono essere utilizzati anche modelli analoghi con una diversa funzione sigmoidea al posto della funzione logistica, come il modello probit; la caratteristica che definisce il modello logistico è che aumentando una delle variabili indipendenti si scalano moltiplicativamente le probabilità del risultato dato ad un tasso costante, con ogni variabile indipendente che ha il proprio parametro; per una variabile dipendente binaria questo generalizza l'odds ratio.

I coefficienti generalmente non sono calcolati da un'espressione in forma chiusa, a differenza dei minimi quadrati lineari.

La regressione logistica come modello statistico generale è stata originariamente sviluppata e resa popolare principalmente da Joseph Berkson, che ha coniato il termine "logit".

Per migliorare il risultato di questo modello sono stati utilizzate e messe a confronto tre diverse tecniche:

- Step-wise Selection[18]
- Akaike Information Criterion[19]
- Bayesian Information Criterion[20]

3.1.1 Step-wise Selection

In statistica, la regressione step-wise è un metodo di adattamento dei modelli di regressione in cui la scelta delle variabili predittive viene eseguita mediante una procedura automatica. In ogni iterazione, una variabile viene considerata per l'addizione o la sottrazione dall'insieme di variabili esplicative in base a un criterio prestabilito che di solito assume la forma di una sequenza di test F o t, ma sono possibili altre tecniche che in questo caso sono

[18] Meesad Phayung, *Recent Advances in Information and Communication Technology*, Springer, (2017)
[19] Hughes Anthony William, *An Iterative Approach to Variable Selection Based on the Kullback-Leibler Information, School of Economics*, (1997)
[20] Meesad Phayung, *Recent Advances in Information and Communication Technology*, Springer (2017)

Akaike e Bayesian Information Criterion

L' AIC (Akaike Information Criterion) è uno stimatore dell'errore di previsione e quindi della qualità relativa dei modelli statistici per un dato insieme di dati. Data una raccolta di modelli per i dati, l'AIC stima la qualità di ciascun modello, rispetto a ciascuno degli altri modelli attraverso la quantità relativa di informazioni perse da un dato modello: meno informazioni perde un modello, maggiore è la qualità di quel modello.
Nella stima della quantità di informazioni perse da un modello, l'AIC si occupa del trade-off tra la bontà di adattamento del modello e la semplicità del modello. In altre parole, l'AIC si occupa sia del rischio di overfitting sia del rischio di under-fitting. Costituisce la base di un paradigma per i fondamenti della statistica ed è anche ampiamente utilizzato per l'inferenza statistica.
Questo valore si calcola tramite la formula:

$$AIC = 2k - 2\ \ln(\hat{\mathcal{L}})$$

Dove $\hat{\mathcal{L}}$ è la funzione di massima verosimiglianza e k corrisponde al numero di parametri nel modello.
Il BIC (Bayesian Information Criterion) può creare problemi quando k^2 risulta molto inferiore alla grandezza del dataset ma permette di scegliere un modello meglio tra una serie di candidati.

$$BIC = k\ln(n) - 2\ \ln(\hat{\mathcal{L}})$$

Il BIC è stato utilizzato per verificare se i modelli con varie trasformazioni delle variabili continue quali radice quadrata, elevamento a potenza e logaritmo performassero meglio delle corrispettive inalterate, mentre l'AIC è stato utilizzato come criterio del Backward Step-wise usufruendo di interazioni tra le variabili fino al 2° tipo.
I seguenti modelli hanno permesso di creare uno score il quale è stato sottoposto ad una soglia pari a 0.2 per mappare i risultati e confrontarli con i valori reali.

Book

Report for Logistic Regression Model

Basic Summary

Call: glm(formula = Book ~ Months+Vehicle.Month+ DistanceKm + IC + BS + MOT + OEW + WL, family = binomial("log it"), data= the.data)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.277976	0.0455889	-49.968	< 2.2e-16 ***
Months.from.last.visit	-0.031193	0.0021917	-14.232	< 2.2e-16 ***
Vehicle.Month.Age.at.Call	-0.004904	0.0003946	-12.429	< 2.2e-16 ***
DistanceKilometers	-0.002996	0.0004354	-6.880	5.99e-12 ***
IC	0.510059	0.1406326	3.627	0.00029 ***
BS	-0.569019	0.0662964	-8.583	< 2.2e-16 ***
MOT	0.376106	0.0446525	8.423	< 2.2e-16 ***
OEW	0.349879	0.0605257	5.781	7.44e-09 ***
WL	0.575468	0.0526132	10.938	< 2.2e-16 ***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 36595 on 86877 degrees of freedom
Residual deviance: 35596 on 86869 degrees of freedom
McFadden R-Squared: 0.02729, Akaike Information Criterion 35614
Number of Fisher Scoring iterations: 6

Test		**Predicted**	
		No	**Yes**
Actual	**No**	80650	1566
	Yes	148	4662

Call

Report for Logistic Regression Model

Basic Summary

Call: glm(formula = Call ~ Months+Vehicle.Month+B+C+ COMMERCIAL+SUV DistanceKm + IC + BS + MOT + OEW + WL, family = binomial("log it"), data= the.data)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.129930	0.0561537	-37.930	< 2.2e-16 ***
Months.from.last.visit	-0.031286	0.0022177	-14.107	< 2.2e-16 ***
Vehicle.Month.Age.at.Call	-0.004980	0.0004015	-12.403	< 2.2e-16 ***
B	-0.120386	0.0462578	-2.603	0.00925 **
C	-0.165912	0.0485981	-3.414	0.00064 ***
COMMERCIAL	-0.210480	0.0758698	-2.774	0.00553 **
SUV	-0.223501	0.0534214	-4.184	3e-05 ***
DistanceKilometers	-0.002917	0.0004322	-6.750	1.47e-11 ***
IC	0.583465	0.1432813	4.072	5e-05 ***
BS	-0.572288	0.0672333	-8.512	< 2.2e-16 ***
MOT	0.374681	0.0459879	8.147	3.71e-16 ***
OEW	0.339177	0.0613479	5.529	3.22e-08 ***
WL	0.570016	0.0537907	10.597	< 2.2e-16 ***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 36595 on 86877 degrees of freedom
Residual deviance: 35566 on 86864 degrees of freedom
McFadden R-Squared: 0.02811, Akaike Information Criterion 35594

Test		**Predicted**	
		No	**Yes**
Actual	**No**	11601	913
	Yes	570	74762

Il secondo modello presentato è il migliore per quanto riguarda la previsione della risposta alla chiamata ed è stato implementato all'interno del workflow; provando a modificare i mesi e l'età del veicolo tramite una radice quadrata o un logaritmo non si ottengono miglioramenti significativi.

3.2 Neural Network

Le reti neurali artificiali [21]sono sistemi informatici ispirati alle reti neurali biologiche che costituiscono il cervello degli animali. Tali sistemi mirano a eseguire compiti considerando esempi, archiviati in un training set, generalmente senza essere programmati con compiti o regole specifiche. Ad esempio, nel riconoscimento delle immagini, potrebbero imparare a identificare le immagini che contengono una classe specifica come automobili, gatti, persone semplicemente analizzando immagini di esempio che sono state etichettate manualmente per indicare se la classe è presente o meno e per utilizzare i risultati per identificarlo in altre immagini. Sono in grado di farlo senza alcuna conoscenza preliminare della classe o delle sue caratteristiche. Una rete neurale si basa su un insieme di unità o nodi collegati chiamati neuroni artificiali o più semplicemente neuroni che rappresentano la controparte nel cervello naturale. Invece le sinapsi, cioè le connessioni nel cervello biologico, sono rappresentate nelle reti neurali come frecce e permettono la trasmissione dei segnali tra i neuroni. Un neurone artificiale che riceve un segnale poi lo elabora e può inviarne uno in uscita ai neuroni ad esso collegati, infatti i neuroni hanno solitamente una soglia che stabilisce se il segnale può essere inviato o meno. Nelle implementazioni delle reti, il segnale in corrispondenza di una connessione è un numero reale e l'output di ciascun nodo viene calcolato applicando una funzione non lineare alla somma dei suoi input. Inoltre le connessioni hanno un peso associato che si adatta al processo di apprendimento che aumenta o diminuisce l'importanza del segnale in una connessione. In genere, i neuroni vengono aggregati in strati. Livelli diversi possono applicare trasformazioni diverse al loro input. Un segnale viaggia dal primo livello, quello di input, ad uno strato interno denominato nascosto, costituito da più passaggi, fino all'ultimo, lo strato di output.

È importante sottolineare che questo modello di cervello biologico è molto grossolano, in effetti nel nostro cervello ci sono molti diversi tipi di neuroni ciascuno

[21] Wlodzislaw D., Erkki O., Slawomir Z., *Artificial Neural Networks: Formal Models and Their Applications*, (2005)

con proprietà diverse e le sinapsi non sono solo un unico peso, ma sono un complesso sistema dinamico non lineare. Inoltre, è noto che la tempistica esatta dei picchi di uscita in molti sistemi è importante, suggerendo che l'approssimazione del codice potrebbe non reggere.

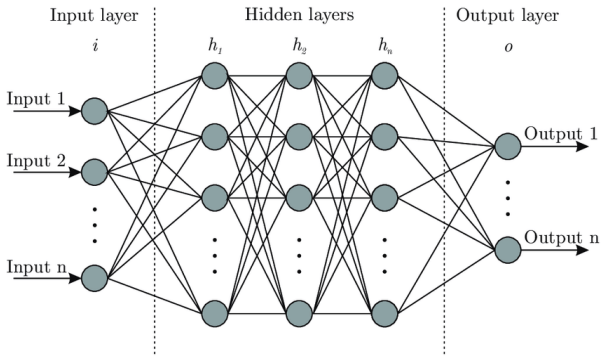


Figure 1.1. Un esempio di Artificial Neural Network con tre strati nascosti.

In origine, le reti neurali artificiali avrebbero dovuto risolvere i problemi allo stesso modo di un cervello umano, tuttavia nel corso del tempo l'attenzione nel corso si è spostata sull'esecuzione di compiti specifici, portando a deviazioni dalla biologia. Alcuni campi comuni in cui vengono utilizzate le neural network sono: visione artificiale, riconoscimento vocale, traduzione automatica, social network filtering, giochi da tavolo e videogiochi.

Il primo grande passo verso la nascita e lo sviluppo delle artificial neural network è stato fatto nel 1943 da Warren Sturgis McCulloch, un neurofisiologo, e Walter Pitts, un matematico, nella loro pubblicazione: *"A logical calculus of the ideas immanent in nervous activity"*. Hanno provato a creare un primo modello di neurone artificiale e, di conseguenza, è stato possibile calcolare alcune semplici funzioni booleane.

Successivamente un passo importante è avvenuto con il primo scheletro di rete neurale, precursore delle attuali reti, denominato di Perceptron. Era una rete con un livello di input e uno di output e con una regola di apprendimento intermedia, un algoritmo per l'apprendimento supervisionato di classificatori binari. È un tipo di classificatore lineare che fa le sue previsioni sulla base di una funzione predittiva

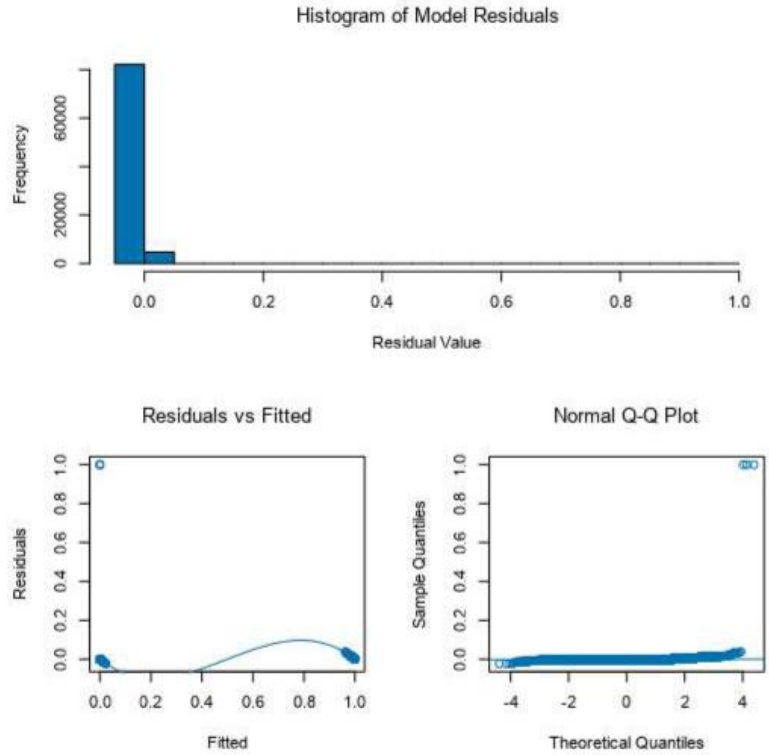
lineare combinando un insieme di pesi con il vettore delle caratteristiche. Purtroppo il perceptron era solo in grado di apprendere schemi separabili linearmente, non era in grado di calcolare la funzione XOR.
Successivamente è stato introdotto il terzo strato delle reti neurali che ha preso il nome di nascosto; questo nuovo schema ha consentito la costruzione di modelli per l'addestramento delle reti perceptron multistrato (MLP).
In seguito sono state apportate delle migliorie tra le quali l'implementazione dell'algoritmo di backpropagation, che è ancora alla base delle reti, o il max-pooling, per aiutare a ridurre al minimo l'invarianza di spostamento e la tolleranza alla deformazione fino ad arrivare alle odierne reti che attraverso il pre-training non supervisionato e l'aumento della potenza di calcolo delle nuove tecnologie e l'elaborazione distribuita hanno consentito l'uso di reti più grandi per problemi più complessi che sono diventati noti come deep learning.

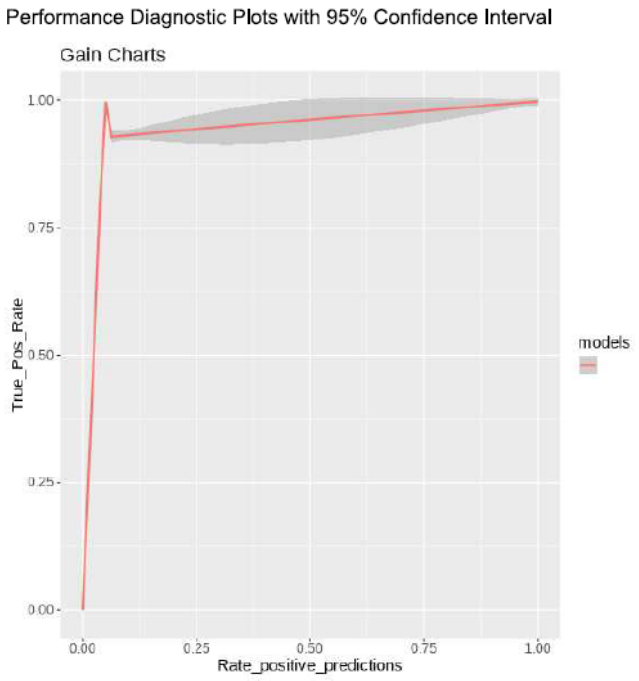
Book
Report for Neural Network Model
Basic Summary
Call: nnet.formula (formula = Book ~ Months+Vehicle.Month+A+B+C+ COMMERCIAL+SUV+D+DistanceKm+IC+BS+MOT+OEW+WL+RSNF, data=the.data, size= 10, linout= FALSE, RANG= c(0,9), decay=0,55 , MaxNWts= 10000, maxit= 100)
Structure: A 19-10-1 network with 211 weights

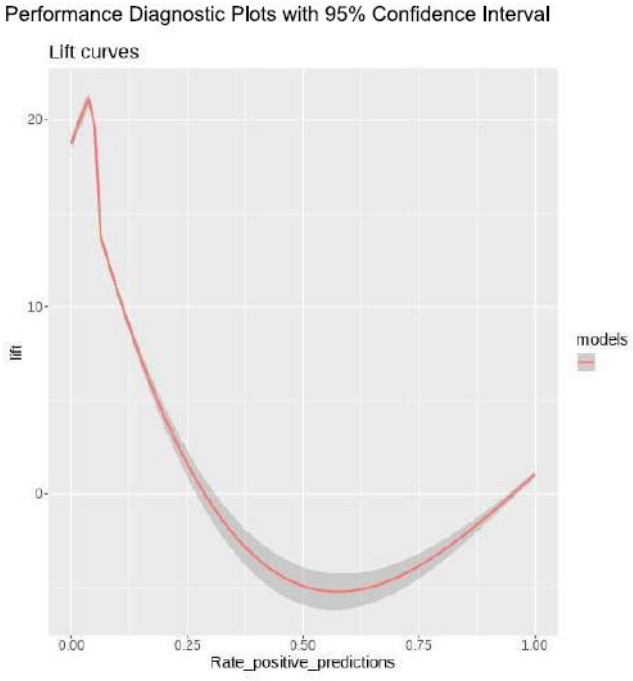
Test		**Predicted**	
		No	**Yes**
Actual	**No**	80650	3
	Yes	0	4662

Questo modello è il migliore per quanto riguarda la previsione della prenotazione ed è stato implementato all'interno del workflow, nonostante presenti problemi nei quantili dovuti ai veicoli commerciali. Questi ultimi presentato chilometraggi ed

età dei veicoli fuori norma ma non possono essere eliminati poichè sono parte dei parametri impostati dall'azienda, seppur siano una quantità esigua all'interno del database. Con la loro eliminazione scomparirebbero i problemi nei reisidui.







Call

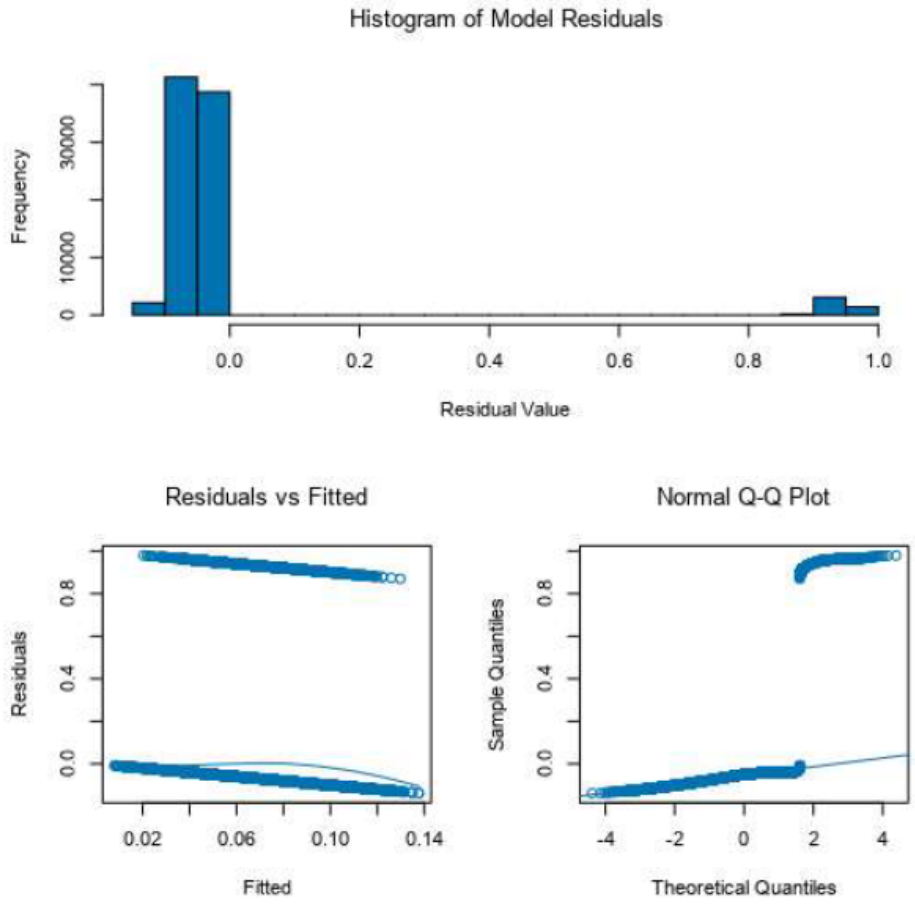
Report for Neural Network Model

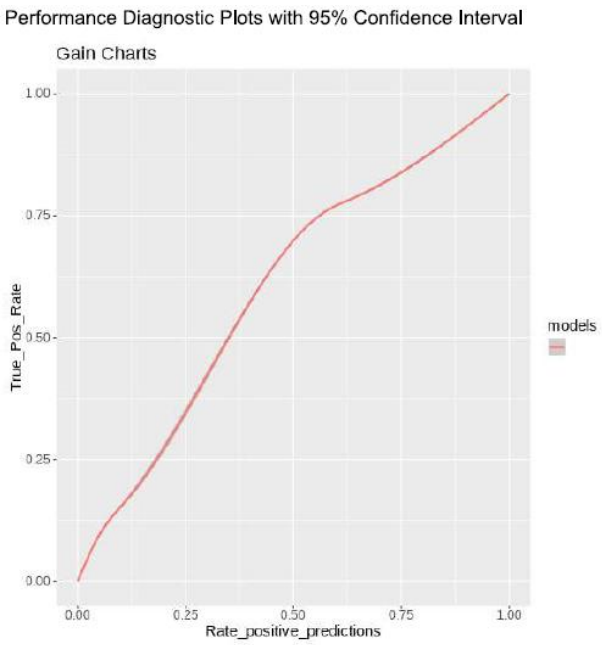
Basic Summary

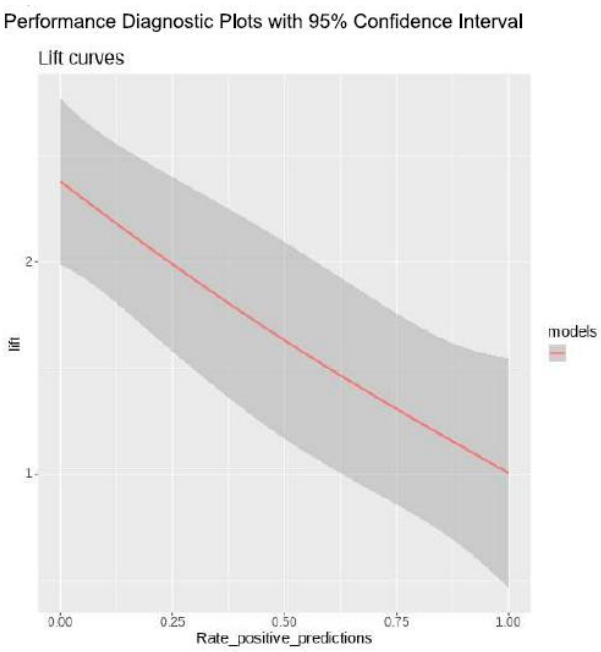
Call: nnet.formula (formula = Call ~ Months+Vehicle.Month+A+B+C+ COMMER-CIAL+SUV+D+DistanceKm+IC+BS+MOT+OEW+WL+RSNF, data=the.data, size= 10, linout= FALSE, RANG= c(0,9), decay=0,5, MaxNWts= 10000, maxit= 100)

Structure: A 18-10-1 network with 201 weights

Test		Predicted	
		No	**Yes**
Actual	**No**	10650	1215
	Yes	850	74662







3.3 Random Forest

Le foreste casuali[22] o le foreste decisionali casuali sono un metodo di apprendimento che operano costruendo una moltitudine di alberi decisionali al momento dell'addestramento e fornendo la classe che è la modalità delle classi dei singoli alberi. Le foreste decisionali casuali correggono l'abitudine degli alberi decisionali di adattarsi eccessivamente al loro set di addestramento, il cosiddetto overfitting. Le foreste casuali generalmente superano gli alberi decisionali, ma la loro accuratezza è inferiore rispetto agli alberi potenziati dal gradiente. Tuttavia, le caratteristiche dei dati possono influenzare le loro prestazioni.

Le foreste casuali sono modelli blackbox, poiché generano previsioni ragionevoli su un'ampia gamma di dati, richiedendo al contempo una configurazione ridotta.

Questo metodo utilizza un algoritmo di apprendimento dell'albero modificato che seleziona, ad ogni divisione del candidato nel processo di apprendimento, un sottoinsieme casuale delle caratteristiche. Viene adoperata questa tecnica per evitare la correlazione degli alberi in un normale campione se una o poche caratteristiche sono predittori molto forti per la variabile di risposta, queste caratteristiche saranno selezionate in molti degli alberi, causando una correlazione. Come parte della loro

[22] Yu L., *Random Forest*, Pavlov, (2019)

costruzione, i predittori casuali delle foreste portano naturalmente a una misura di diversità tra le osservazioni. Si può anche definire una misura di diversità di foresta casuale tra dati non etichettati: l' idea è di costruire un predittore di foresta casuale che distingua i dati "osservati" da dati sintetici opportunamente generati. Una diversità casuale tra foreste può essere utile perché gestisce molto bene le variabili miste, è invariante alle trasformazioni monotone delle variabili di input ed è robusta agli outlier. La dissomiglianza casuale della foresta si occupa facilmente di un gran numero di variabili semicontinue a causa della sua intrinseca selezione di variabili.

In generale, per un problema di classificazione con caratteristiche p, vengono utilizzate le caratteristiche $\sqrt{p}$ arrotondato per difetto in ciascuna divisione, per i problemi di regressione invece è preferibile un $^{p}/_{3}$ con una dimensione minima del nodo di 5 come default ma in pratica i valori migliori per questi parametri dipenderanno dal problema e dovrebbero essere trattati come parametri di regolazione.
Le inoltre le foreste possono essere utilizzate per classificare l'importanza delle variabili.

Il primo passo per misurare l'importanza delle variabili in un set di dati $D_n = \{X_i, Y_i\}_{i=1}^{n}$ serve ad adattare una foresta ai dati. Durante tale processo, l'errore out-of-bag per ogni dato viene registrato e calcolato come media sulla foresta

Per misurare l'importanza della funzione j-esima dopo il training i valori dell'elemento j-esimo sono permutati tra i dati di traning e l'errore out-of-bag viene nuovamente calcolato su questo insieme di dati perturbato. Il punteggio di importanza per l'elemento j viene calcolato calcolando la media della differenza nell'errore out-of-bag prima e dopo la permutazione su tutti gli alberi. Il punteggio è normalizzato dalla deviazione standard di queste differenze. Le caratteristiche che producono valori grandi per questo punteggio sono classificate come più importanti delle caratteristiche che producono valori piccoli.

Tuttavia questo presenta alcuni inconvenienti, ad esempio per i dati che includono variabili categoriali con un diverso numero di livelli, le foreste casuali sono orientate a favore di quegli attributi con più livelli. Per risolvere il problema possono essere utilizzati metodi come permutazioni parziali o unbiased alberi crescenti.

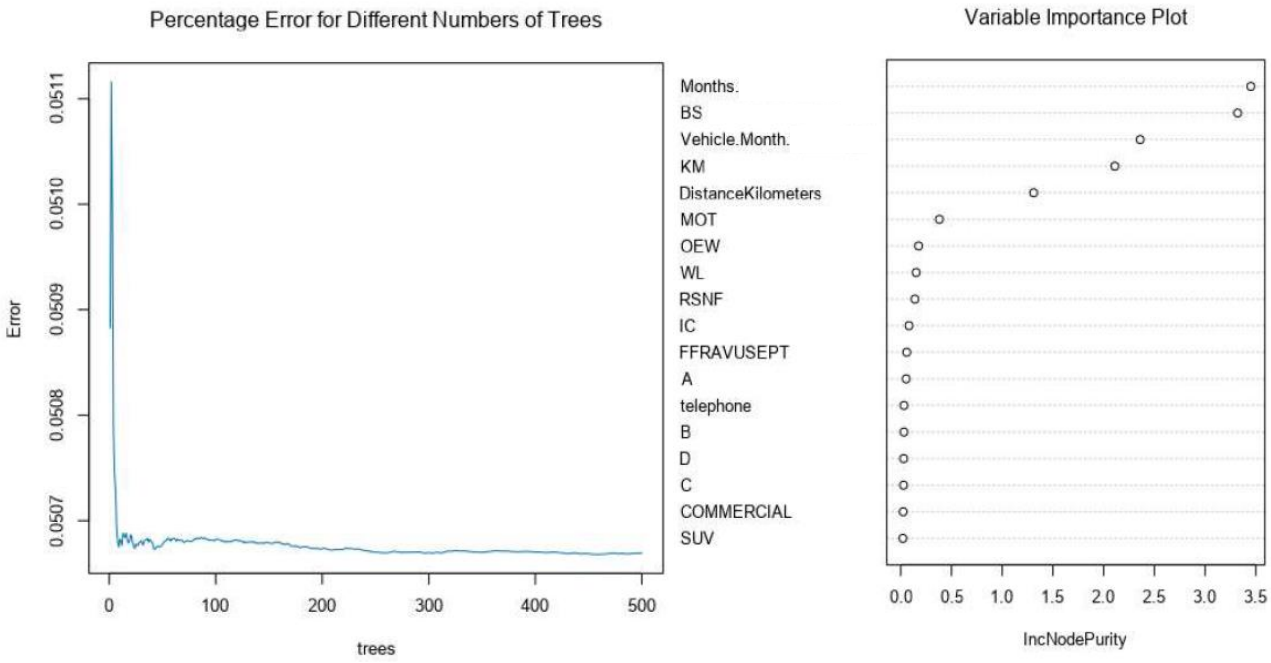
Questo algoritmo presenta una similarità con l'algoritmo K-NN (K-Nearest Neighbour). Questi sono modelli costruiti da un set $\{x_i, y_i\}_{i=1}^{n}$ che fa previsioni $\hat{y}$ per i nuovi punti x' osservando i punti vicini, tramite l'uso di una funzione peso W non negativa, inoltre per ogni particolare x', i pesi dei punti devono sommarsi a uno:

$$\hat{y} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} W_j(x_i, x')y_i$$

Ciò mostra che l'intera foresta è di uno schema di vicini ponderato, con pesi che fanno la media di quelli dei singoli alberi. I vicini di x' in questa interpretazione sono i punti che condividono la stessa foglia in qualsiasi albero. In questo modo, l'intorno di x' dipende in modo complesso dalla struttura degli alberi, e quindi dalla struttura del training set.

Call
Report for Random Forest Model
Basic Summary
Call: randomforest(formula = Call ~ Km+Months+Vehi-
cle.Month+VUSEPT+A+B+C+ COMMERCIAL+SUV+D+Dis-
tanceKm+IC+BS+MOT+OEW+WL+RSNF, data = the.data, mtry = 5, maxnodes = 10
replace = TRUE, size = 26063)
Type of forest: classification
Number of trees: 500
Number if variables tried at each split: 5
Mean of the squared residuals: 0.05
Percentage of variance explained: 0.9
Structure: A 18-10-1 network with 201 weights

Test		Predicted	
		No	**Yes**
Actual	**No**	11665	1100
	Yes	635	73762



La foresta creata per predire l'output BOOK ha la medesima precisione di un classificatore casuale, per questo motivo non viene riportato alcun report o script.

3.4 Cross validation

La cross validation talvolta chiamata stima della rotazione o test fuori campione, è una delle varie tecniche di convalida del modello per valutare come i risultati di un' analisi verrà generalizzata rispetto a un set di dati indipendente. Viene utilizzato principalmente in contesti in cui l'obiettivo è la previsione e si desidera stimare con quanta precisione un modello predittivo si esibirà nella pratica.

In un problema di previsione, a un modello viene solitamente fornito un set di dati noti su cui viene eseguito l'addestramento di questo e un altro set di dati rispetto al quale viene testato il modello. L'obiettivo della convalida incrociata è testare la

capacità del modello di prevedere nuovi dati che non sono stati utilizzati per la stima, al fine di segnalare problemi come l'overfitting o il bias di selezione dei predittori e per dare un'idea di come il modello si generalizzerà a un dataset indipendente e sconosciuto.

Un ciclo di convalida incrociata prevede il partizionamento di un campione di dati in sottoinsiemi complementari, l'esecuzione dell'analisi su un sottoinsieme e la convalida dell'analisi sull'altro sottoinsieme. Per ridurre la variabilità, nella maggior parte dei metodi vengono eseguiti più cicli di convalida incrociata utilizzando diverse partizioni e i risultati della convalida vengono nei cicli per fornire una stima delle prestazioni predittive del modello.

In questo progetto per la costruzione dei modelli ho utilizzato come parametri standard un partizionamento in 5 sottoinsiemi ed un numero di cicli pari a 3.

Capitolo 4

Costruzione e automatizzazione del processo

A questo punto ho progettato una struttura organizzativa più snella che permette di ottenere dei risultati in modo efficiente; ho inizializzato una base clienti in cui ho immagazzinato tutte le informazioni personali e i dati relativi ai veicoli raggruppata per targa.

Ho creato inoltre un database delle concessionarie e carrozzerie anche per altri progetti, in modo tale da poter poi estrarre per questo workflow solo l'indirizzo per la geolocalizzazione.

Nel caso in cui la distanza del cliente dal relativo centro fosse superiore a 50 km, il sistema cerca se ci sono delle location più vicine che possano essere disponibili per offrire il medesimo servizio.

Tale scelta è stata dettata dalla percentuale di persone che hanno aderito all'offerta in base alla distanza.

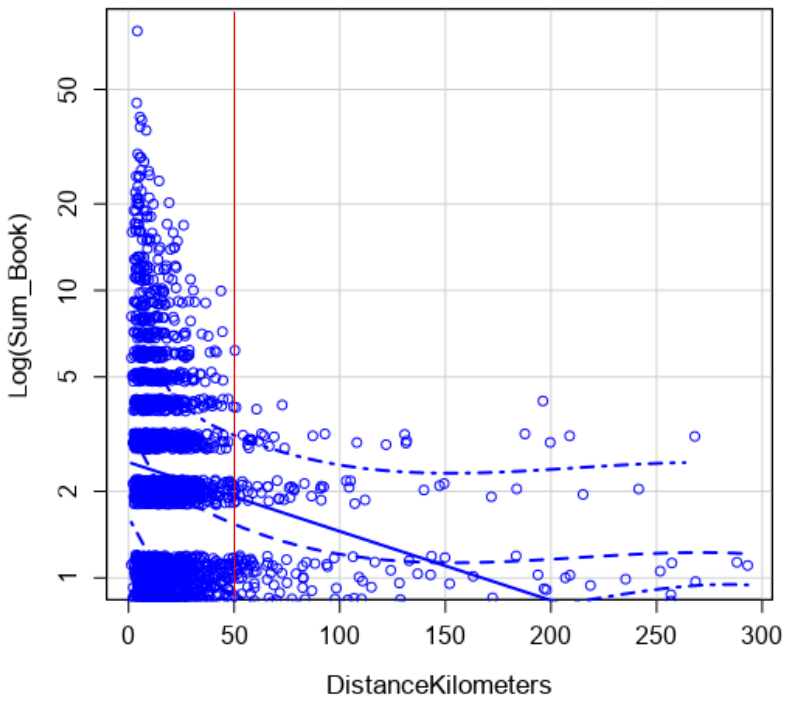


Figura 1 Distribuzione del logaritmo della somma delle prenotazioni con le curve di tendenza

Inoltre tutte le informazioni relative agli operatori dei call center sono state archiviate in un ulteriore database.

Per le chiamate effettuate in precedenza ho creato uno storico da utilizzare in futuro stornate dei valori che sono stati archiviati in altri database.

Lo schema seguente rappresenta la nuova struttura di archiviazione dei dati e le relazioni tra i vari database.

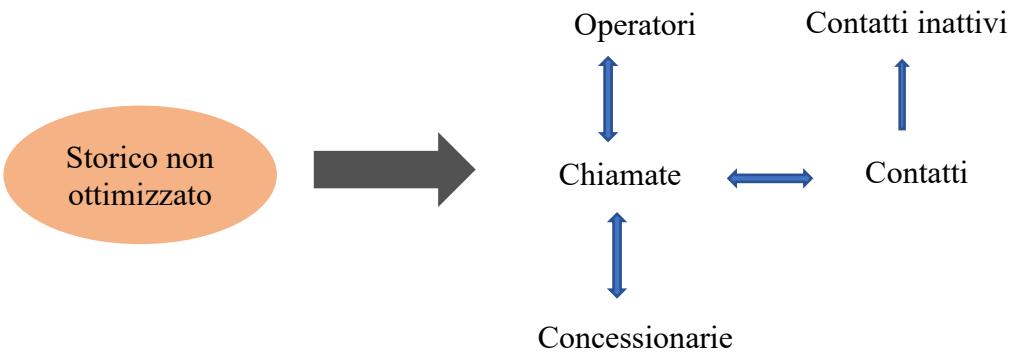


Figura 2 Nuova organizzazione del database

Tramite Rss (Really Simple Syndication) Feed ho ottenuto la lista Bloctel, un elenco di aziende e privati che non desidera essere chiamata per campagne di telemarketing ed è ottenibile sul sito governativo[23], secondo i decreti[24] vigenti in Francia. Esso viene aggiornato una volta al mese ed arricchisce i dati per tutti i processi riguardanti i servizi di call center.

A questo punto, il problema risulta essere l'individuazione di un canale che non richieda l'intervento umano per la trasmissione dei dati e che sia totalmente sicuro da attacchi informatici e che permetta lo sgravo di responsabilità per la manomissione o perdita dei dati.

La prima opzione vagliata per risolvere tale problema riguarda l'invio dei dati tramite mail criptate, il canale però può rivelarsi poco sicuro a causa della possibile alterazione delle informazioni, richiedendo comunque una componente umana e, per questo motivo, è stata scartata a priori.

[23] www.bloctel.gouv.fr

[24] Decreto n° 2015-556 del 19/05/2015, Legge n° 2014-344 del 17/03/2014

La seconda e decisiva scelta ricade sull'utilizzo di Sharepoint, che permette di creare un gateway sicuro tra il proprio dispositivo ed il cloud Microsoft sia in entrata sia in uscita.

In comune accordo tra l'azienda che richiede supporto e Msx sono stati creati due gateway indirizzati al sito di Msx su Sharepoint; uno permetterà all'azienda di caricare il database da una cartella designata in locale, tramite l'upload automatico e l'altro consentirà a Msx di scaricarlo in modo automatico attraverso le funzionalità presenti sul portale, utilizzando canali protetti forniti da Microsoft.

Dopo l'analisi di sicurezza del database, esso entra in un processo ETL classico, in cui i dati vengono estratti, puliti e catalogati tramite una serie di parametri di controllo; qualora non riescano ad essere catalogati correttamente, le tuple subiscono una serie di check per capire se possono essere ugualmente trasformati tramite un'analisi inversa delle componenti e, in caso contrario, vengono inviati al responsabile del progetto tramite una mail interna.

Inoltre, durante il processo di validazione, vengono richiamati due database per evitare di avere contatti inutili: Contatti inattivi e Bloctel.

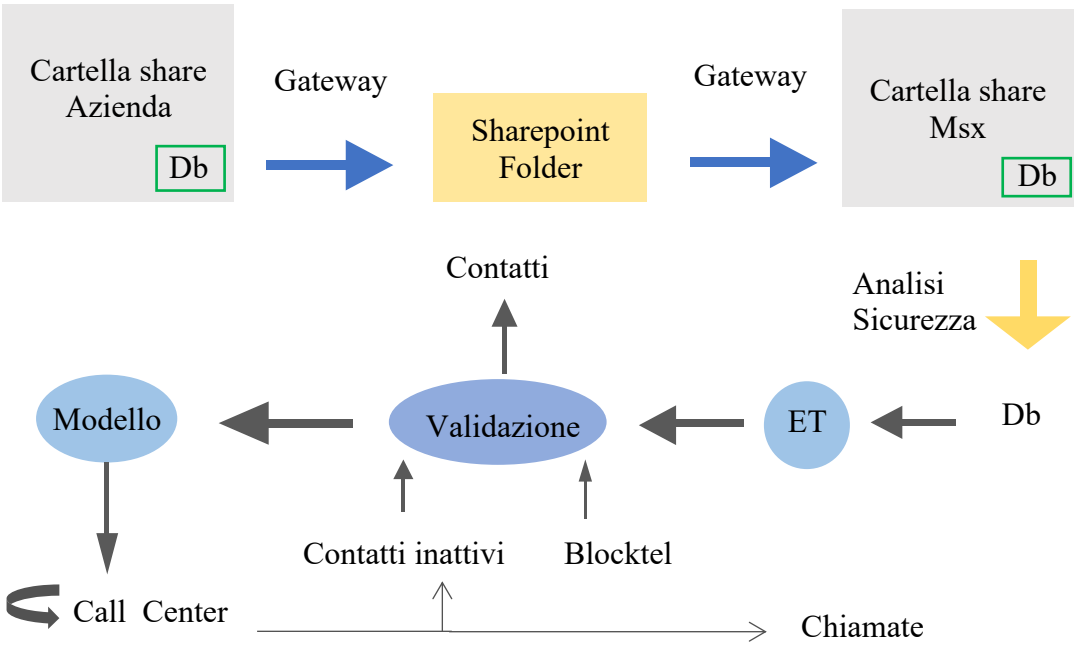


Figura 3 Schema riepilogativo del progetto

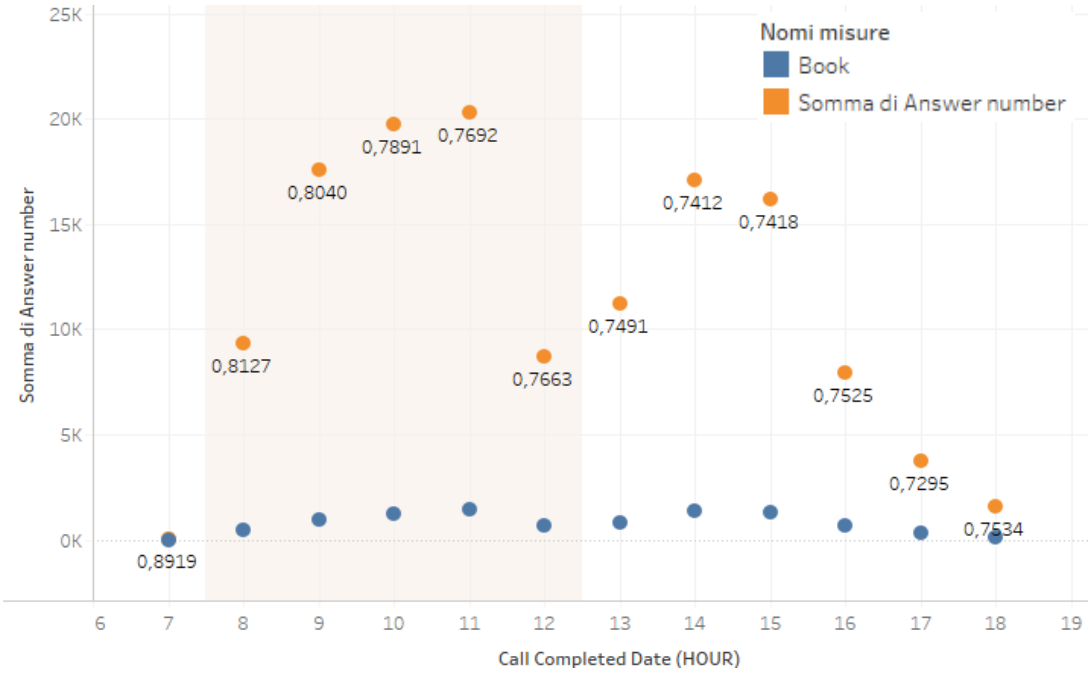


Figura 4 Volume di risposte e di prenotazioni con il rapporto rispetto al numero di chiamate totali

I contatti, una volta vagliati passati dal processo per definirne la priorità, vengono ordinati e catalogati, sulla base della suddetta, in una tabella oraria che agevoli il Call Center; secondo l'osservazione statistica delle precedenti chiamate, come mostrato nell'immagine sottostante, vengono inseriti nella fascia del mattino i candidati con uno score più alto per la prenotazione e poi per le chiamate e infine quelli con basso score nelle ore verso il tardo pomeriggio.

Inoltre dopo aver effettuato tutte le chiamate poi, esse vengono immagazzinate nel relativo file, mentre i contatti a seconda dell'outcome della telefonata subiscono diverse azioni:

OUTCOME	AZIONE
Non risponde	Reinserito nel chiamate da effettuare (fino a 4 totali) con una probabilità inferiore del 15%
Rifiuto della chiamata	Reinserito nel chiamate da effettuare (fino a 4 totali) con una probabilità inferiore del 25%
Rifiuto dell' offerta o non risponde al 4° tentativo	Rimosso dalla tabella
Vecchio proprietario/ Società	Immesso nei contatti inattivi
Prenotazione appuntamento	Prenotazione appuntamento e rimosso dalla tabella

Infine prima che il workflow si concluda avviene un check tra la previsione ed il tentativo di chiamata al candidato. In caso la percentuale di errori salga sopra una soglia prefissata del 15% viene inviata un mail al referente del progetto al fine di ricalibrare il modello o costruirne uno migliore.

Nonostante sia possibile creare dei modelli più complessi e performanti di quelli utilizzati la capacità disponibile della virtual machine e del server dedicato non permette ulteriori miglioramenti, poiché questo rischia il crash. Inoltre le disposizioni aziendali non permettevano l'utilizzo di tool complessi che non potessero essere modificati direttamente dall'interfaccia di Alteryx.

Questo progetto è funzionante con una percentuale di errore inferiore al 3% ed è stato preso come core per sviluppare altri processi in relazione all'assistenza clienti.

Bibliografia

Desiderio V. J, Chris E. Taylor and Niamh Nic Daéid, *Handbook of Trace Evidence Analysis*, John Wiley & Sons, (2020)

Hughes A. W., *An Iterative Approach to Variable Selection Based on the Kullback-Leibler Information*, School of Economics, (1997)

Horst R., *The Weibull Distribution: A Handbook*, CRC Press (2008)

Köppen M., *Advances in Neuro-Information Processing*, Springer Science & Business Media, (2009)

Meesad P., *Recent Advances in Information and Communication Technology*, Springer, (2017)

Marwick P., *Implementation of the N-dimensional Logit Model*, Mitchell & Co (1972)

Plötz P., *On the distribution of individual daily driving distances,* CRC Press, (2005)

Rinne H., *The Weibull Distribution: A Handbook,* CRC Press, 2008

Sammut C., *Machine Learning: Proceedings of the Nineteenth International Conference*, Morgan Kaufmann Publishers, (2002)

Stone B., *Direct marketing. I metodi e le tecniche vincenti*, Il Sole 24 Ore, (2005)

Wlodzislaw D., Erkki O., Slawomir Z., *Artificial Neural Networks: Formal Models and Their Applications*, Springer (2005)

Wu B., Zhang L. and Zhao Y., *Feature Selection via Cramer's V-Test Discretization for Remote-Sensing Image Classification*, Transactions on Geoscience and Remote Sensing, vol. 52, no. 5, (2014)

Yu L.*, Random Forest*, Pavlov, (2019)

Zill D. G. , *Advanced Engineering Mathematics*, Jones & Bartlett Learning, (2020)

Appendice A

Pearson Correlation Analysis

Full Correlation Matrix

	Months	Vehicle Month	FFRAVU	A	B
KM	-0.00678199	-0.00152996	-0.000708	-0.00336629	0.00985851
Months		0.01164731	0.117816	0.01208666	0.05164912
Vehicle Month			0.021382	-0.13441221	-0.01605740
VUSEPT				-0.03066488	-0.06810145
A					-0.19706422

	C	COMMERCIAL	D	SUV	Distance
KM	-0.0035	-0.00206317	0.00036328	-0.004307	-0.00203122
Months	-0.0892	0.06741627	-0.02146841	-0.027125	0.05737589
Vehicle Month	0.1421	-0.00277759	0.16240115	-0.211233	-0.02202436
VUSEPT	-0.0690	0.33309077	-0.02523683	-0.003849	0.01552356
A	-0.1722	-0.07162307	-0.06295457	-0.124479	-0.02402317
B	-0.4438	-0.18451303	-0.16218152	-0.320679	-0.02974222
C		-0.16130524	-0.14178256	-0.280345	0.00034663
COMMERCIAL			-0.05894494	-0.116551	0.03928463
D				-0.102445	0.01308130
SUV					0.01842452

	telephone	IC	BS	MOT	OEW
KM	0.0027	-0.00109792	0.01350161	-0.002773	-0.00075077
Months	0.0002	-0.04323868	-0.06086266	-0.075460	-0.04266612
Vehicle.Month	-0.0792	-0.08143118	0.24202940	0.167533	-0.20641470
FFRAVUSEPT	0.0338	-0.00970798	-0.05332461	-0.088241	-0.02971446
A	-0.0092	-0.02421704	-0.04418822	-0.084087	0.03516975
B	-0.0261	-0.06238716	-0.01865601	0.051614	0.05806257
C	0.0019	-0.05454019	0.11758560	0.031152	-0.04897567
COMMERCIAL	0.0108	0.06329333	-0.05600408	-0.001736	0.01509981
D	-0.0147	-0.01993034	0.04630462	0.049553	-0.04015730
SUV	0.0411	0.13880098	-0.05717632	-0.037047	-0.00229708
Distance	0.0346	0.03725974	-0.04042928	0.024122	0.01387707
telephone		0.01658735	-0.02317499	-0.008709	0.01409512
IC			-0.04211217	-0.069687	-0.02346647
BS				-0.382782	-0.12889820
MOT					-0.21330079

	WL	RSNF		WL	RSNF
KM	-0.00363969	-0.0059	**D**	0.00108930	-0.0710
Months	0.21317205	-0.0417	**SUV**	-0.02157665	0.0971
Vehicle.Month	0.04235764	-0.3530	**Distance**	-0.00402321	-0.0096
FFRAVUSEPT	-0.04720131	-0.0538	**telephone**	-0.00210430	0.0116
A	0.00699573	0.1342	**IC**	-0.03727640	-0.0425
B	0.04365331	-0.0891	**BS**	-0.20475427	-0.2337
C	-0.01403475	-0.0828	**MOT**	-0.33882744	-0.3507
COMMERCIAL	0.00648162	-0.0637	**OEW**	-0.11409672	-0.1302
			WL		-0.2068

Matrix of Corresponding p-values

	Months	Vehicle.Month	FFRAVU	A	B
KM	4.6984e-03	5.2364e-01	7.6768e-01	1.6056e-01	3.9653e-05
Months		1.2027e-06	0.0000e+00	4.6955e-07	0.0000e+00
Vehicle.Month			0.0000e+00	0.0000e+00	2.1751e-11
FFRAVUSEPT				0.0000e+00	0.0000e+00
A					0.0000e+00

	C	COMMERCIAL	D	SUV	Distance
KM	1.431601	3.8979e-01	8.7964e-01	7.2539e-02	3.9717e-01
Months	0.000000	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00
Vehicle.Month	0.000000	2.4694e-01	0.0000e+00	0.0000e+00	0.0000e+00
FFRAVUSEPT	0.000000	0.0000e+00	0.0000e+00	1.0862e-01	9.7234e-11
A	0.000000	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00
B	0.000000	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00
C		0.0000e+00	0.0000e+00	0.0000e+00	8.8511e-01
COMMERCIAL			0.0000e+00	0.0000e+00	0.0000e+00
D				0.0000e+00	4.9525e-08
SUV					1.5987e-14

	telephone	IC	BS	MOT	OEW
KM	2.449901	6.4720e-01	1.8206e-08	2.4766e-01	7.5432e-01
Months	9.211701	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00
Vehicle.Month	0.000000	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00
FFRAVUSEPT	0.000000	5.1935e-05	0.0000e+00	0.0000e+00	0.0000e+00
A	1.164504	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00
B	0.000000	0.0000e+00	7.5495e-15	0.0000e+00	0.0000e+00
C	4.129301	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00
COMMERCIAL	6.197006	0.0000e+00	0.0000e+00	4.6919e-01	3.0834e-10
D	8.010810	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00
SUV	0.000000	0.0000e+00	0.0000e+00	0.0000e+00	3.3831e-01
Distance	0.000000	0.0000e+00	0.0000e+00	0.0000e+00	7.2616e-09
telephone		4.6885e-12	0.0000e+00	2.8293e-04	4.2115e-09
IC			0.0000e+00	0.0000e+00	0.0000e+00
BS				0.0000e+00	0.0000e+00
MOT					0.0000e+00

	WL	RSNF
KM	1.2923e-01	1.345302
Months	0.0000e+00	0.000000
Vehicle.Month	0.0000e+00	0.000000
FFRAVUSEPT	0.0000e+00	0.000000
A	3.5442e-03	0.000000
B	0.0000e+00	0.000000
C	4.9010e-09	0.000000
COMMERCIAL	6.8963e-03	0.000000

	WL	RSNF
D	6.4978e-01	0.000000
SUV	0.0000e+00	0.000000
Distance	9.3536e-02	5.861705
telephone	3.8040e-01	1.145006
IC	0.0000e+00	0.000000
BS	0.0000e+00	0.000000
MOT	0.0000e+00	0.000000
OEW	0.0000e+00	0.000000
WL		0.000000

Spearman Correlation Analysis

Full Correlation Matrix

	Months	Vehicle Month	FFRAVU	A	B	C
KM	-0.15430493	0.78719462	0.037588	-0.24369201	-0.10542352	0.2446
Months		0.04226148	0.088388	0.01429005	0.04827609	-0.0891
Vehicle Month			0.053293	-0.17246721	0.00038920	0.1763
FFRAVUSEPT				-0.03066488	-0.06810145	-0.0690
A					-0.19706422	-0.1722
B					1.00000000	-0.4438
C					-0.10542352	0.2446

	C	COMMERCIAL	D	SUV	Distance	telephone
KM	0.2446	-0.02139007	0.16747896	-0.109150	-0.05867580	0.0338
Months	-0.0891	0.05489276	-0.01772281	-0.031828	0.07311874	-0.0016
Vehicle Month	0.1763	-0.01280934	0.13293943	-0.217291	-0.14567377	-0.0588
FFRAVUSEPT	-0.0690	0.33309077	-0.02523683	-0.003849	0.02409918	0.0338
A	-0.1722	-0.07162307	-0.06295457	-0.124479	-0.04548780	-0.0092
B	-0.4438	-0.18451303	-0.16218152	-0.320679	-0.06504703	-0.0261
C		-0.16130524	-0.14178256	-0.280345	-0.01081403	0.0019
COMMERCIAL			-0.05894494	-0.116551	0.05798444	0.0108
D				-0.102445	0.00173603	-0.0147
SUV					0.09709710	0.0411
Distance						0.0524

	IC	BS	MOT	OEW	WL	RSNF
KM	-0.07318500	0.30590064	0.180834	-0.21773997	-0.02743121	-0.3682
Months	-0.04862928	-0.04133923	-0.131038	-0.09380943	0.28481719	-0.0156
Vehicle Month	-0.10167998	0.32912495	0.204590	-0.23164522	0.02703101	-0.4587
FFRAVUSEPT	-0.00970798	-0.05332461	-0.088241	-0.02971446	-0.04720131	-0.0538
A	-0.02421704	-0.04418822	-0.084087	0.03516975	0.00699573	0.1342
B	-0.06238716	-0.01865601	0.051614	0.05806257	0.04365331	-0.0891
C	-0.05454019	0.11758560	0.031152	-0.04897567	-0.01403475	-0.0828
COMMERCIAL	0.06329333	-0.05600408	-0.001736	0.01509981	0.00648162	-0.0637
D	-0.01993034	0.04630462	0.049553	-0.04015730	0.00108930	-0.0710
SUV	0.13880098	-0.05717632	-0.037047	-0.00229708	-0.02157665	0.0971
Distance	0.04242897	-0.05921012	-0.015209	0.03354806	0.00667881	0.0315
telephone	0.01658735	-0.02317499	-0.008709	0.01409512	-0.00210430	0.0116
IC		-0.04211217	-0.069687	-0.02346647	-0.03727640	-0.0425
BS			-0.382782	-0.12889820	-0.20475427	-0.2337
MOT				-0.21330079	-0.33882744	-0.3507
OEW					-0.11409672	-0.1302
WL						-0.2068

Matrix of corresponding p_value

	Months	Vehicle Month	FFRAVU	A	B	C
KM	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	0.000000
Months		0.0000e+00	0.0000e+00	2.5702e-09	0.0000e+00	0.000000
Vehicle Month			0.0000e+00	0.0000e+00	8.7112e-01	0.000000
FFRAVUSEPT				0.0000e+00	0.0000e+00	0.000000
A					0.0000e+00	0.000000
B						0.000000

	COMMERCIAL	D	SUV	Distance	telephone
KM	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	0.000000
Months	0.0000e+00	1.4877e-13	0.0000e+00	0.0000e+00	5.047401
Vehicle.Month	9.3136e-08	0.0000e+00	0.0000e+00	0.0000e+00	0.000000
FFRAVUSEPT	0.0000e+00	0.0000e+00	1.0862e-01	0.0000e+00	0.000000
A	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	1.164504
B	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	0.000000
C	0.0000e+00	0.0000e+00	0.0000e+00	6.5496e-06	4.129301
COMMERCIAL		0.0000e+00	0.0000e+00	0.0000e+00	6.197006
D			0.0000e+00	4.6928e-01	8.010810
SUV				0.0000e+00	0.000000
Distance					0.000000

	IC	BS	MOT	OEW	WL	RSNF
KM	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	0.000000
Months	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	6.301811
Vehicle.Month	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	0.000000
FFRAVUSEPT	5.1935e-05	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	0.000000
A	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	3.5442e-03	0.000000
B	0.0000e+00	7.5495e-15	0.0000e+00	0.0000e+00	0.0000e+00	0.000000
C	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	4.9010e-09	0.000000
COMMERCIAL	0.0000e+00	0.0000e+00	4.6919e-01	3.0834e-10	6.8963e-03	0.000000
D	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	6.4978e-01	0.000000
SUV	0.0000e+00	0.0000e+00	0.0000e+00	3.3831e-01	0.0000e+00	0.000000
Distance	0.0000e+00	0.0000e+00	2.2908e-10	0.0000e+00	5.3691e-03	0.000000
telephone	4.6885e-12	0.0000e+00	2.8293e-04	4.2115e-09	3.8040e-01	1.145006
IC		0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00	0.000000
BS			0.0000e+00	0.0000e+00	0.0000e+00	0.000000
MOT				0.0000e+00	0.0000e+00	0.000000
OEW					0.0000e+00	0.000000
WL						0.000000

Appendice B

Codice ReliefF

```
from __future__ import print_function
import numpy as np
from sklearn.neighbors import KDTree
class ReliefF(object):
    """Feature selection using data-mined expert knowledge.
    Based on the ReliefF algorithm as introduced in:
    Kononenko, Igor et al. Overcoming the myopia of inductive learning
    algorithms with RELIEFF (1997), Applied Intelligence, 7(1), p39-55
    """
    def __init__(self, n_neighbors=100, n_features_to_keep=10):
        """Sets up ReliefF to perform feature selection.
        Parameters
        ----------
        n_neighbors: int (default: 100)
            The number of neighbors to consider when assigning feature
            importance scores.
            More neighbors results in more accurate scores, but takes longer.
        Returns
        -------
        self.feature_scores = None
        self.top_features = None
        self.tree = None
        self.n_neighbors = n_neighbors
        self.n_features_to_keep = n_features_to_keep
    def fit(self, X, y):
        """Computes the feature importance scores from the training data.
        Parameters
        ----------
        X: array-like {n_samples, n_features}
            Training instances to compute the feature importance scores from
        y: array-like {n_samples}
            Training labels
        }
        Returns
        -------
        None
        """
        self.feature_scores = np.zeros(X.shape[1])
        self.tree = KDTree(X)
        for source_index in range(X.shape[0]):
            distances, indices = self.tree.query(
                X[source_index].reshape(1, -1), k=self.n_neighbors+1)
            # Nearest neighbor is self, so ignore first match
            indices = indices[0][1:]
            # Create a binary array that is 1 when the source and neighbor
            #  match and -1 everywhere else, for labels and features..
            labels_match = np.equal(y[source_index], y[indices]) * 2. - 1.
            features_match = np.equal(X[source_index], X[indices]) * 2. - 1.
            # The change in feature_scores is the dot product of these  arrays
            self.feature_scores += np.dot(features_match.T, labels_match)
        self.top_features = np.argsort(self.feature_scores)[::-1]
    def transform(self, X):
        """Reduces the feature set down to the top `n_features_to_keep` features.
        Parameters
        ----------
        X: array-like {n_samples, n_features}
            Feature matrix to perform feature selection on
        Returns
        X_reduced: array-like {n_samples, n_features_to_keep}
            Reduced feature matrix
        """
        return X[:, self.top_features[:self.n_features_to_keep]]
    def fit_transform(self, X, y):
        """Computes the feature importance scores from the training data, then
        reduces the feature set down to the top `n_features_to_keep` features.
        Parameters
        ----------
        X: array-like {n_samples, n_features}
            Training instances to compute the feature importance scores from
        y: array-like {n_samples}
            Training labels
        Returns
        -------
        X_reduced: array-like {n_samples, n_features_to_keep}
            Reduced feature matrix
        """
        self.fit(X, y)
        return self.transform(X)
```

Appendice C

Codice nucleo Deep Learning NN

```
N_model = Sequential()

# The Input Layer :
NN_model.add(Dense(128, kernel_initial-
izer='normal',input_dim = train.shape[1], acti-
vation='relu'))

# The Hidden Layers :
NN_model.add(Dense(256, kernel_initial-
izer='normal',activation='relu'))
NN_model.add(Dense(256, kernel_initial-
izer='normal',activation='relu'))
NN_model.add(Dense(256, kernel_initial-
izer='normal',activation='relu'))

# The Output Layer :
NN_model.add(Dense(1, kernel_initial-
izer='normal',activation='linear'))

# Compile the network :
NN_model.compile(loss='mean_absolute_er-
ror', optimizer='adam', metrics=['mean_abso-
lute_error'])
NN_model.summary()
```