The statistical analysis of Telco customer churn

Luigi Mascolo

Contents

1 Abstract

In questo report viene analizzato il fenomeno dell'abbandono dei clienti (**churn**) utilizzando i dati di una compagnia di telecomunicazioni. L'obiettivo principale è individuare le caratteristiche dei clienti più associate alla decisione di interrompere il servizio. A tal fine, sono stati costruiti e confrontati tre modelli predittivi: una regressione logistica classica, un modello con regolarizzazione (Lasso-Ridge-Elastic Net), e una regressione logistica eseguita in ambiente distribuito con Spark. Ogni modello è stato valutato in base a metriche di accuratezza e capacità discriminante, ed anche al suo livello di interpretabilità. I risultati mettono in evidenza vantaggi e limiti di ciascun approccio, offrendo indicazioni utili per applicazioni pratiche, anche su larga scala, legate alla fidelizzazione dei clienti e alla riduzione del tasso di abbandono.

2 Introduction

Nel contesto attuale delle telecomunicazioni, la fidelizzazione del cliente è una delle sfide principali per le aziende del settore. Il fenomeno del churn, ovvero l'abbandono da parte del cliente, rappresenta una perdita economica rilevante e una minaccia alla stabilità del business. È dunque essenziale per le aziende prevedere in anticipo quali clienti sono a rischio di abbandono e quali sono le cause che li portano ad abbandonare, così da poter attuare strategie mirate per evitare ciò.

Il presente report analizza il *Telco Customer Churn Dataset*, un insieme di dati fornito da una compagnia telefonica operante in California. Il dataset raccoglie informazioni su **7043 clienti**, tra cui dati anagrafici, tipologia di servizi sottoscritti (telefonia, Internet, opzioni aggiuntive), modalità di pagamento, durata del contratto, livello di soddisfazione e valore stimato del cliente (Customer Lifetime Value) per un totale di **50 variabili**. Ecco alcune delle variabili più importanti:

- **Dependents** : Se il cliente ha delle persone a carico (Yes,No).
- **Referred a friend**: Indica se il cliente ha mai condigliato l'azienda a qualche amico/familiare (Yes,No).
- **Tenure in Months** : Il numero di mesi totali cui il cliente ha usufruito delle offerte dell'azienda.
- **Offer** : Indica il tipo di offerta scelta dal cliente .
- **Internet Service**: Se il cliente ha sottoscritto il servizio di Internet dell'azienda (Yes,No).
- **Online Backup** : Se il cliente usufruisce di un servizio di backup online (Yes,No).
- **Contract** : Indica la tipologia di contratto sottoscritta dal cliente.

- **Payment Method** : Indica la tipologia di pagamento.
- **Monthly Charge** : L'ammontare totale mensile speso dal cliente per i servizi dell'azienda.
- **Churn Label** : Indica se il cliente ha lasciato o meno l'azienda (Yes,No).
- **Churn Score** : Un punteggio che misura il potenziale abbandono del cliente, più è alto più è probabile che il cliente abbandonerà.
- **CLTV**: Customer lifetime Value. Più è alto il valore più è prezioso il cliente.
- **Churn Reason** : La ragione per cui il cliente ha abbandonato l'azienda.

Il dato di **churn_label**, indicato come variabile binaria, segnala se un cliente ha abbandonato il servizio o meno nel periodo di riferimento (Q3).Il primo obiettivo del report è quello di identificare le variabili che influenzano o meno il tasso d'abbandono del cliente e fare un analisi delle varie tipologie di clienti e delle varie offerte/contratti che sottoscrivono per identificare problemi relativi ad una determinata tipologia di offerta, analizzare anche i servizi che utilizzano i clienti per verificare che esiste una correlazione tra il tasso d'abbandono e uno specifico servizio (segnalando magari strategie non efficaci per quel servizio).

Atro obiettivo dell'analisi è costruire modelli predittivi capaci di stimare con buona accuratezza la probabilità di churn, aiutando così l'azienda a identificare con anticipo i clienti a rischio. Per raggiungere questo scopo, l'indagine adotta tre diversi approcci modellistici:

1. Una regressione logistica classica (**Generalized Linear Model**) stimata in ambiente R, che consente di valutare l'effetto delle singole variabili sulla probabilità di churn.
2. Un approccio penalizzato con **Lasso o Elastic Net**, utile per selezionare automaticamente le variabili più rilevanti e ridurre il rischio di overfitting.
3. L'utilizzo di un GLM in ambiente **Spark**, che simula l'impiego di strumenti di calcolo distribuito, tipici dei contesti aziendali con grandi moli di dati.

I dati vengono preventivamente suddivisi in un set di addestramento e uno di test, per garantire una valutazione onesta delle prestazioni predittive dei modelli. Questa indagine offre quindi una panoramica metodologica e pratica sull'uso di modelli statistici e strumenti di analisi dati per affrontare un problema reale e centrale per molte aziende: prevedere e prevenire il **churn** dei propri clienti.

3 Methods

L'analisi del dataset è stata condotta seguendo un approccio strutturato, articolato in diverse fasi, ciascuna finalizzata a garantire una preparazione adeguata dei dati e una valutazione accurata dei modelli predittivi. In particolare, il processo ha incluso:

- **Preparazione dei dati** (Data Pre-processing): trasformazione e pulizia delle variabili per assicurare che i dati siano nel formato corretto per l'analisi statistica e modellistica.
- **Analisi Esplorativa dei Dati** (EDA): esplorazione delle caratteristiche principali del dataset, individuazione di pattern, anomalie e relazioni tra le variabili.
- **Modelling**: stima di diversi modelli di regressione, con approcci classici e regolarizzati (Lasso, Elastic Net), sia in ambiente locale che distribuito.

3.1 Preparazione dei dati (Data Pre-processing)

La prima fase è quella di controllare che le tipologie delle variabili siano quelle adatte, per trattarle nella maniera più opportuna nei vari passaggi suggessivi dell'analisi. Nel nostro caso è stato fatto un lavoro di **rinomina** delle variabili e di cambio della **tipologia** delle variabili.

Ecco come si presentavano prima le variabili:

```
# A tibble: 5 x 50
  `Customer ID` Gender   Age `Under 30` `Senior Citizen` Married
  <chr>         <chr>  <dbl> <chr>      <chr>            <chr>
1 8779-QRDMV    Male      78 No         Yes              No
2 7495-OOKFY    Female    74 No         Yes              Yes
3 1658-BYGOY    Male      71 No         Yes              No
4 4598-XLKNJ    Female    78 No         Yes              Yes
5 4846-WHAFZ    Female    80 No         Yes              Yes
# i 44 more variables: Dependents <chr>,
#   `Number of Dependents` <dbl>, Country <chr>, State <chr>,
#   City <chr>, `Zip Code` <dbl>, Latitude <dbl>, Longitude <dbl>,
#   Population <dbl>, Quarter <chr>, `Referred a Friend` <chr>,
#   `Number of Referrals` <dbl>, `Tenure in Months` <dbl>,
#   Offer <chr>, `Phone Service` <chr>,
#   `Avg Monthly Long Distance Charges` <dbl>, ...
```

Possiamo notare come molte variabili nel nome hanno degli spazi e spesso la tipologia di variabile (dbl,chr,fct) non viene assegnata nella maniera corretta, questi sono problemi che possono invalidare l'analisi e portare a degli errori, quindi provvediamo col rinominare le variabili e assegnare la giusta tipologia alle varie variabili; inoltre notiamo che la variabile

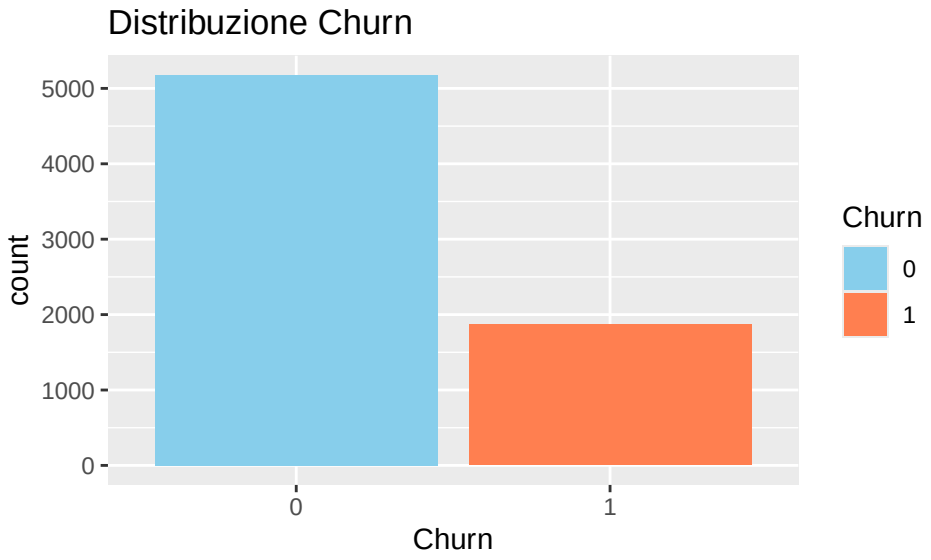
di nostro interesse **churn_label** non viene trattata come una variabile fattoriale, quindi creiamo una nuova variabile **churn** con valori 0 = non ha abbandonado l'azienda, 1 = ha abbandonato l'azienda. Infine rimuoviamo anche alcune variabili non utili all'analisi o ridondanti.

Ecco come si presentano le variabili dopo il pre-processing:

```
# A tibble: 5 x 46
  gender   age married dependents number_of_dependents country   state
  <fct>  <dbl> <fct>   <fct>                     <dbl> <chr>     <chr>
1 Male      78 No      No                            0 United S~ Cali~
2 Female    74 Yes     Yes                           1 United S~ Cali~
3 Male      71 No      Yes                           3 United S~ Cali~
4 Female    78 Yes     Yes                           1 United S~ Cali~
5 Female    80 Yes     Yes                           1 United S~ Cali~
# i 39 more variables: city <chr>, latitude <dbl>, longitude <dbl>,
#   quarter <chr>, referred_a_friend <fct>,
#   number_of_referrals <dbl>, tenure_in_months <dbl>, offer <fct>,
#   phone_service <fct>, avg_monthly_long_distance_charges <dbl>,
#   multiple_lines <fct>, internet_service <fct>,
#   internet_type <fct>, avg_monthly_gb_download <dbl>,
#   online_security <fct>, online_backup <fct>, ...
```

3.2 Analisi Esplorativa dei Dati (EDA)

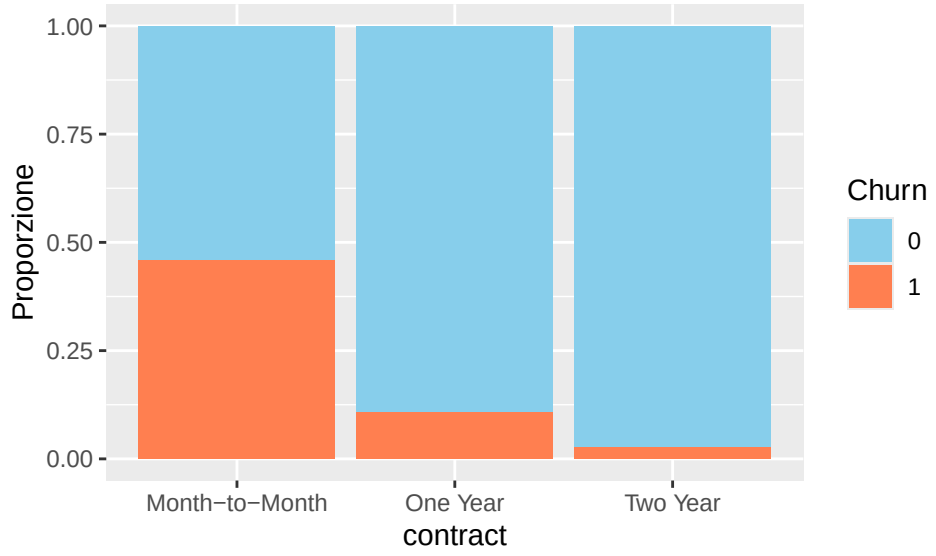
Fase cruciale è l'**EDA** per poter vedere quelle che sono le statistiche descrittive delle variabili, intercettare relazioni tra le variabili e sopratutto individuare le distribuzioni delle variabili, sopratutto della variabile target di nostro interesse. Nel nostro caso la variabile risposta è **churn** che ci dice se il cliente ha abbandonato o meno l'azienda di telecomunicazion. Dopo aver visto le statistiche descrittive delle variabili andiamo a visualizzare un bar plot della variabile churn.



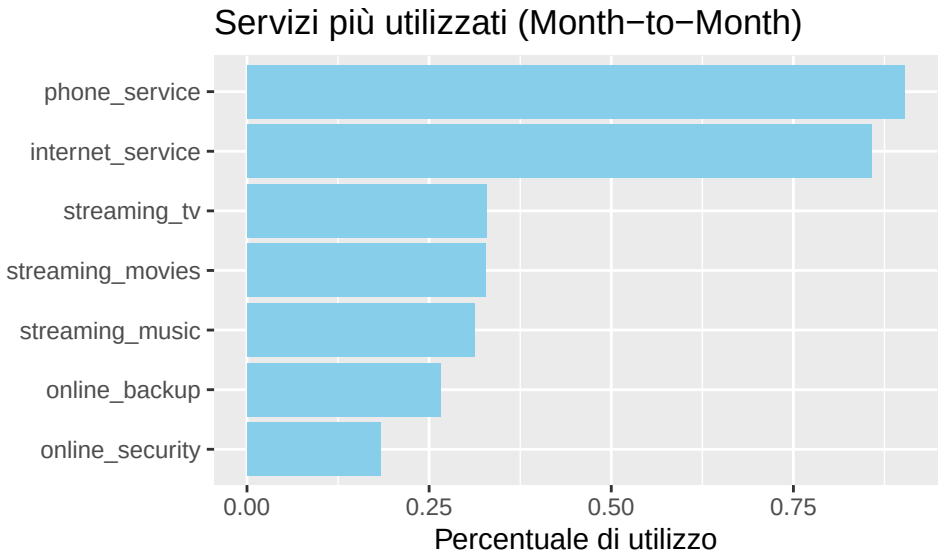
Abbiamo quindi circa 2000 clienti che hanno disdetto l'abbonamento con l'azienda e circa 5000 che non hanno abbandonato l'azienda. Andiamo ora a mettere in relazione il tasso di abbandono con altre variabili, in questa fase potremmo individuare pattern che portano i clienti ad abbandonare dovuti per esempio al metodo di pagamento, tipologia di contratto/offerta, l'avere o meno un servizio di backup.

3.2.1 Relazione tra tasso di abbandono e tipologia di contratto

Questa è la relazione tra il tasso di abbandono è la tipologia di contratto del cliente



Possiamo notare come i clienti che hanno un contratto più lungo sono meno propensi ad abbandonare la compagnia, mentre i clienti che hanno contratti mensili hanno più probabilità di lasciare l'azienda. Detto ciò una buona e semplice pratica di business potrebbe essere offrire contratti a lungo termine vantaggiosi ai clienti mensili per *ridurre il loro tasso di abbandono*, magari facendo leva su quelli che sono i servizi più utilizzati da questa tipologia di clienti.



Possiamo vedere come i servizi più utilizzati sono **phone_service** e **internet_service** ma questi sono i servizi più utilizzati per tutti i clienti a prescindere dal loro contratto, la vera

differenza dei clienti mensili sta in un maggior utilizzo dei **servizi di streaming** (mentre i clienti a lungo termine preferiscono i servizi di online_backup e online_security), detto ciò si potrebbero offrire contratti a lungo termine ai clienti mensili facendo leva sui servizi di streaming.

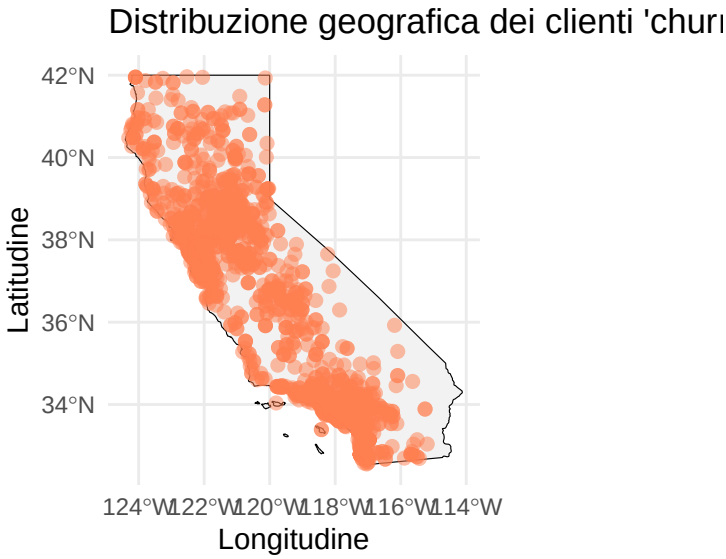
3.2.2 Tasso di Abbandono e Offerte dell'azienda

Andiamo ora ad esplorare le varie offerte della compagnia e vediamo il relativo tasso di abbandono.

Notiamo subito un tasso di abbandono preoccupante per l'offerta E e per l'offerta D, questo potrebbe significare malfunzionamenti in servizi collegati a queste offerte e deve portarci ad intervenire per abbassare il tasso di abbandono per questi clienti.

3.2.3 Distribuzione Geografica del tasso di abbandono

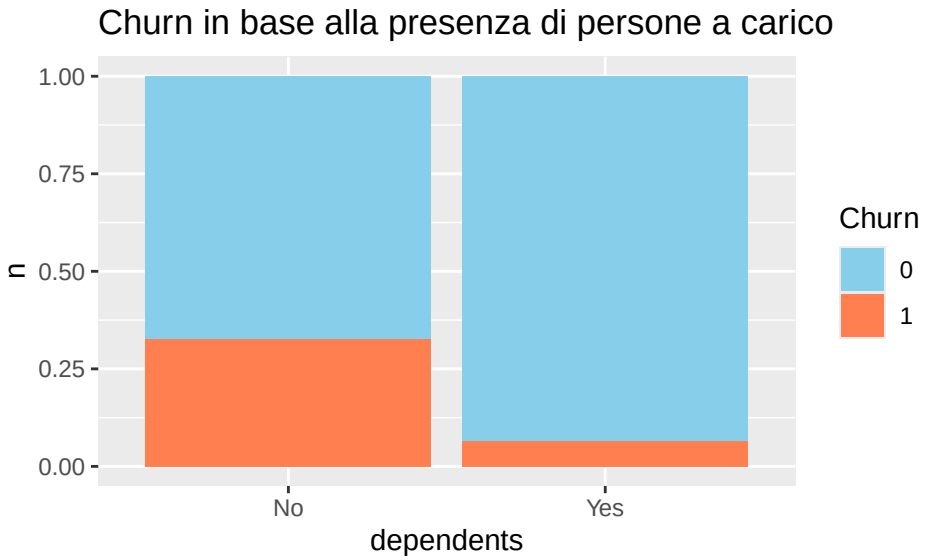
Altro spunto di analisi relativa alle strategie di business è quello di identificare le zone geografiche con maggiore presenza di clienti che hanno abbandonato, per fare ciò andiamo a graficare una mappa della california e usiamo la coordinate dei clienti che hanno abbandonato; quindi usiamo le variabili **latitude** e **longitude** solo per i clienti che hanno **churn** = 1.



La aree più dense corrispondono alle zone più abitate della California, ma questo può essere uno spunto ad approfondire quest'analisi per identificare una possibile correlazione tra il tasso di abbandono e la zona/città dove il cliente vive, questo per identificare possibili guasti o malfunzionamenti di servizi specifici in determinate zone dello stato.

3.2.4 Persone a carico e tasso di abbandono

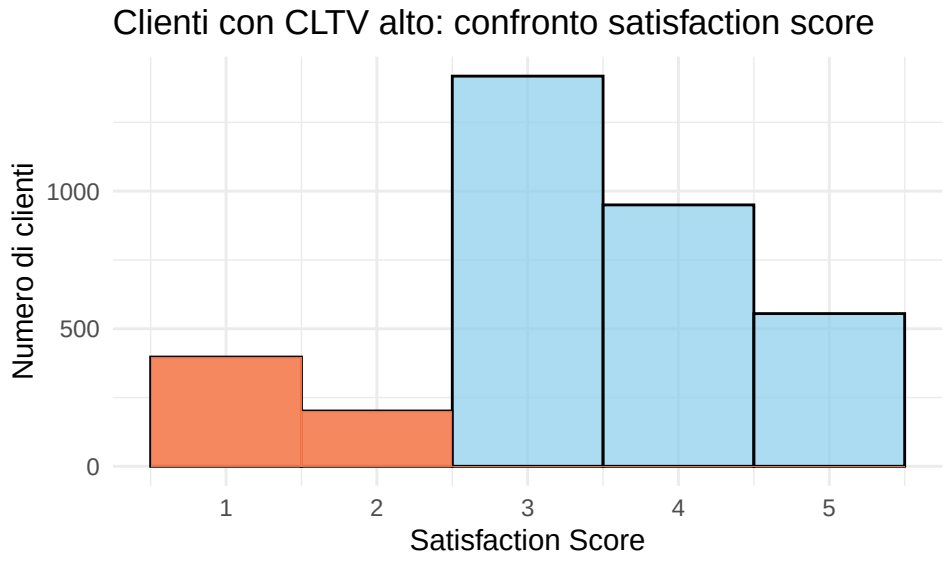
Altra variabile che potrebbe influenzare l'abbandono dei clienti è l'avere o meno delle persone a carico, abbiamo a nostra disposizione in tal senso due variabili: **dependets** (binaria) e **number_of_dependents** (numerica). L'idea è che chi ha persone a carico è meno incline a disdire il contratto con la compagnia, questo perchè ci sono più persone che utilizzano i servizi in uno stesso nucleo familiare e di conseguenza diventa più difficile cambiare provider per internet,telefonia,streaming.



Dal grafico si evince che i clienti con persone a carico sono meno propense ad abbandonare l'azienda rispetto ai clienti che non hanno persone a carico, questo potrebbe essere molto utile sia per la fase successiva di *modelling* sia per la creazione di strategie ad hoc per fidelizzare maggiormente i clienti che non hanno persone a carico.

3.2.5 Clienti preziosi e satisfaction score

Come ultima analisi nella nostra analisi esplorativa andiamo a monitorare i clienti con un alto valore di **CLTV** (Customer Lifetime Value), più questo valore è alto più è prezioso il cliente di conseguenza bisogna prestare attenzione a questi clienti perchè costituiscono un bene per l'azienda. Andiamo ad investigare sul numero di clienti ad alto valore in relazione al **satisfaction_score**, per monitorare potenziali clienti preziosi ma insoddisfatti. I clienti ad alto valore sono tutti quei clienti che hanno almeno un valore di circa 4500 CLTV (valore mediano per la variabile).



I risultati suggeriscono che ci sono pochi clienti ad alto valore con un satisfaction_score basso (1 o 2), mentre la maggior parte dei clienti preziosi si trova nella fascia intermedia e nelle fasce alte (4 o 5). Questa situazione suggerisce un contesto non preoccupante, bisogna comunque monitorare quei pochi clienti insoddisfatti magari andando ad esplorare variabili come **churn_reason**, **churn_score** , per intercettare possibili abbandoni che nel caso di clienti preziosi possono far del male all'azienda.

3.3 Regressione Logistica

Passiamo ora alla parte di creazione dei modelli per interpretare al meglio quelle che sono le influenze delle variabili indipendenti sulla variabile dipendente **"churn"**. L'approccio scelto passa prima per la creazione di un modello di **Regressione Logistica** con link logit, verrà poi creato un modello di regressione penalizzata usando tecniche come **Lasso** dato il numero considerevole di variabili presenti nel nostro dataset, infine useremo anche un approccio distribuito usando **spark** e creando sempre un modello di regressione binaria. Infine i modelli verranno valutati e confrontati seguendo diverse metriche e verranno usati anche per supportare le strategie di business legate al far abbassare il tasso di abbandono attraverso la gestione delle variabili che più lo influenzano (positivamente e negativamente).

3.3.1 Definizione Modello

Il primo modello utilizzato è una Regressione Logistica dato che la variabile di nostro interesse (**churn**) è una variabile binaria. La Regressione logistica è un GLM dove la random component segue una distribuzione di Berouille o **Binomiale**:

$$Y|x \sim Ber(\mu)$$

E la link function è una funzione logit :

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \text{logit}(\mu)$$

Quindi avremo che:

$$\mu = \frac{\exp(x'\beta)}{1+\exp(x'\beta)}$$

Il primo passo utile alla successiva analisi dei modelli è quello di dividere il nostro dataset in **train set** (70%) e **test set**(30%), questo per valutare l'abilità predittiva dei modelli sui dati test che ovviamente non fanno parte dei dati su cui il modello è stato addestrato.

Successivamente, costruiamo un primo modello che comprende tutte le variabili presenti nel nostro dataset escludendo soltanto le variabili che non possono essere trattate e le variabili sicuramente legate a churn che potrebbero forviare l'analisi del modello (churn_category, churn_reason, …).

Il warning del modello *"si sono verificate probabilità stimate numericamente pari a 0 o 1"* suggerisce un pericolo di **overfitting** del modello, questo perchè il modello riesce ad attribuire una probabilità molto vicina allo 0 o all'1 a molte osservazioni presenti nel nostro dataset di training, in questo caso possiamo ridurre il numero di predittori per evitare questo pericolo e in questa maniera rendiamo il modello più semplice, abbassando il numero di variabili cui bisogna interpretare l'impatto sulla variabile target **churn**. Inoltre c'è la presenza di coefficienti "NA" ciò sta a significare una possibile collinearità tra i predittori, infatti

il coefficiente "NA" della variabile **total_revenue** è dovuto al fatto che total_revenue = monthly_charges * tenure_in_months, quindi è una combinazione di altre variabili e va eliminata per non invalidare l'analisi. Costruiamo un *secondo modello* andando ad includere soltanto le variabili significative individuate nel primo modello completo, questo secondo modello ridotto risulterà molto più facile da interpretare e non sarà a rischio overfitting.

In questo modello ridotto abbiamo meno variabili e la maggior parte appaiono significative, ma possiamo creare un nuovo modello che escluda alcune variabili che risultano ancora non significative e confrontare i due modelli coducendo un test "LRT" per valutare il modello da preferire, usiamo il test LRT perchè i due modelli sono *"nested"* uno è costruito su un subset delle variabili dell'altro.

3.3.2 Validazione modello

Il test **Likelihood ratio test (LRT)** va a comparare la log-likelihood dei due modelli per verificare se l'aggiunta di alcune variabili al modello migliora o meno significativamente il fit del modello ai dati. La statistica likelihood ratio test è :

$$\mathrm{LR} = 2\log\left(\frac{L_{\text{full}}}{L_{\text{resticted}}}\right) = 2(\ell_{\text{full}} - \ell_{\text{restricted}})$$

```
Analysis of Deviance Table

Model 1: churn ~ age + dependents + referred_a_friend + tenure_in_months +
    number_of_referrals + offer + phone_service + internet_service +
    +avg_monthly_gb_download + online_security + online_backup +
    premium_tech_support + streaming_movies + contract + monthly_charge +
    payment_method
Model 2: churn ~ age + dependents + referred_a_friend + tenure_in_months +
    number_of_referrals + offer + phone_service + online_security +
    online_backup + premium_tech_support + contract + monthly_charge +
    payment_method
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      4907     3371.9
2      4910     3377.6 -3  -5.6381   0.1306
```

Possiamo notare che il p-value del test è **0.1306**, in tal caso possiamo concludere col dire che non rifiutiamo l'ipotesi nulla secondo cui il modello ridotto è adeguato. Pertanto, *scegliamo il modello ridotto* in quanto più parsimonioso e più facilmente interpretabile, senza compromettere significativamente la bontà del fit. Quindi il modello scelto avrà questa formula:

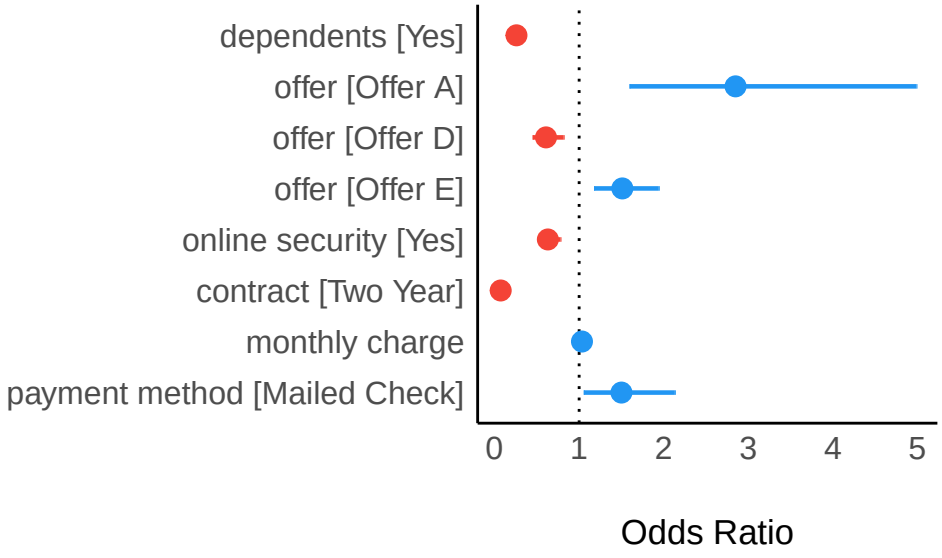
$$\begin{aligned}\text{churn} \sim\ & \text{age} + \text{dependents} + \text{referred a friend} + \text{tenure in months} \\ & + \text{number of referrals} + \text{offer} + \text{phone service} \\ & + \text{online security} + \text{online backup} + \text{premium tech support} \\ & + \text{contract} + \text{monthly charge} + \text{payment method}\end{aligned}$$

3.3.3 Interpretazione del modello

Avendo scelto il modello più adatto passiamo ora ad interpretare i **coefficienti del modello**, che per la regressione logistica vanno letti in termini di odds, quindi i coefficienti delle variabili andranno ad impattare sugli odds di abbandonare la compagnia. Andiamo a vedere quelli che sono i coefficienti in termini di Odds:

```
# Fixed Effects

Parameter             | Coefficient |       SE |       95% CI
-----------------------------------------------------------------
(Intercept)           |        0.32 |     0.06 | [0.21, 0.47]
age                   |        1.01 | 2.63e-03 | [1.01, 1.02]
dependentsYes         |        0.26 |     0.04 | [0.19, 0.34]
referred_a_friendYes  |        6.62 |     0.96 | [5.01, 8.83]
tenure_in_months      |        0.97 | 3.35e-03 | [0.97, 0.98]
number_of_referrals   |        0.53 |     0.03 | [0.47, 0.58]
offerOffer A          |        2.85 |     0.83 | [1.60, 4.99]
offerOffer B          |        0.87 |     0.17 | [0.60, 1.26]
```



Ricordiamo che il modello creato predice la probabilità di churn (cioè di abbandono), questo è utile ricordarlo affinchè non si commettano errori di interpretazione degli odds, la regola generale è che *OR > 1 –> Aumenta la probabilità di abbandono, OR < 1 –> Riduce la probabilità di abbandono.* Ecco l'interpretazione in termini di odds di alcune variabili:

- **Dependents [Yes]**: I clienti con persone a carico hanno odds di abbandono inferiori del 74% rispetto a chi non ha persone a carico, a parità delle altre variabili.
- **Offer A**: L'offerta A è associata a odds di churn significativamente più alto (OR =2.85), cioè i clienti che la ricevono sono più propensi ad abbandonare.

- **Offer D**: L'offerta D avendo un odds ratio minore di 1 (0.61) porterà a ridurre la probabilità di abbandono.
- **Online Security**: I clienti che utilizzano il servizio di sicurezza online hanno meno probabilità di abbandonare la compagnia.
- **Contract Two Year**: I clienti che firmano un contratto biennale avendo un OR minore di 1, sono più propensi a rimanere in azienda.

3.3.4 Predictions

Passiamo ora a quelle che sono le predictions del modello creato, useremo il test set per valutare il modello con metriche come l'**Accuray** e **AUC**, indagheremo anche sugli effetti marginali per rendere ancora più chiare le relazioni tra i vari predittori e la variabile target.

L'accuracy misura la percentuale di osservazioni correttamente classificate:

$$\text{Accuracy} = \frac{\text{Numero di predizioni corrette}}{\text{Totale delle osservazioni}}$$

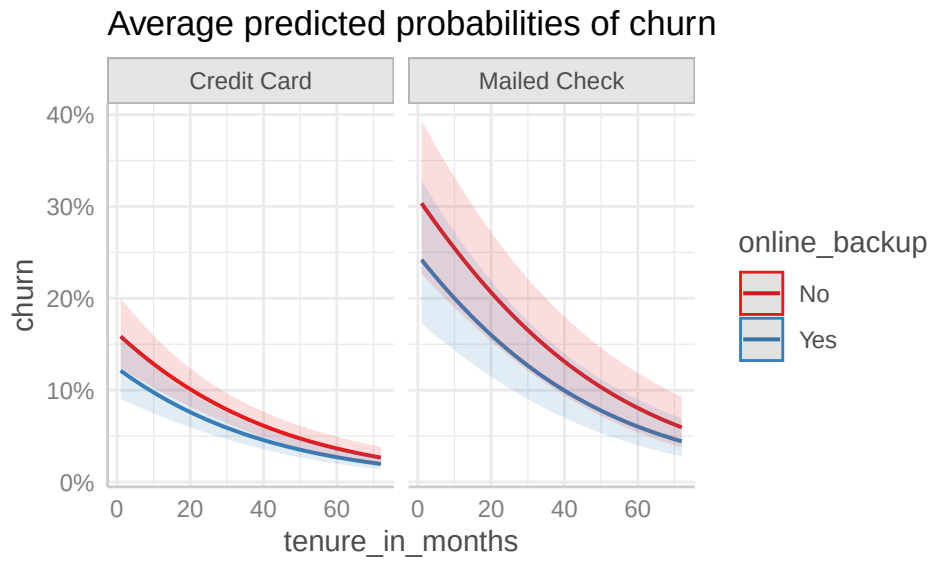
L'accuratezza del modello è stata calcolata confrontando le classi predette con quelle osservate nel test set. Il modello ha ottenuto un'**accuracy pari al 83 %**, indicando la percentuale di clienti correttamente classificati come "churn" o "non churn".

```
[1] "Accuracy: 0.832"
```

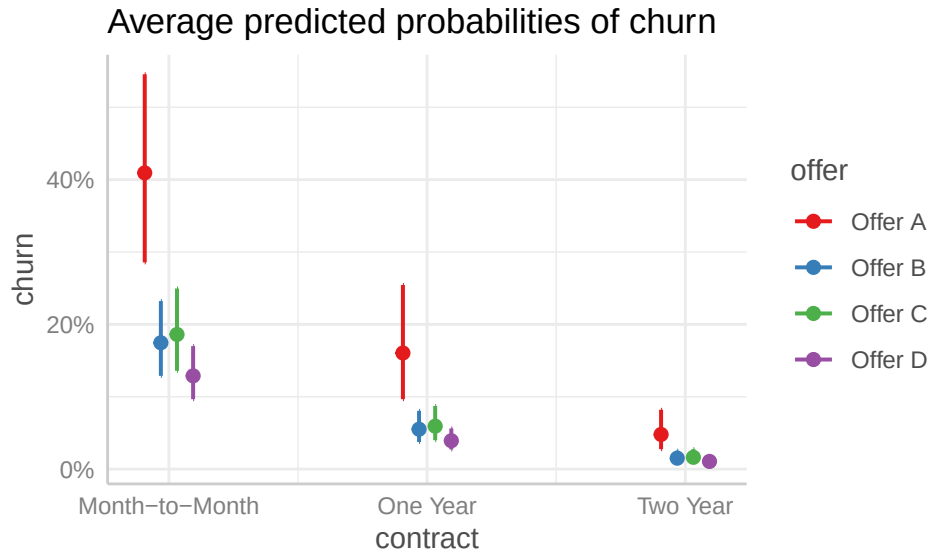
L'AUC è una metrica che misura la capacità del modello di distinguere tra le classi, l'AUC deriva dalla curva ROC (Receiver Operating Characteristic) che rappresenta con l'asse X i falsi positivi e con l'asse Y i veri positivi. Per valutare le performance predittive del modello, è stata calcolata l'area sotto la curva ROC (AUC). Il valore ottenuto è pari a **0.9018**, il che indica una **eccellente capacità discriminativa** del modello nel distinguere tra clienti che abbandonano (churn = 1) e quelli che rimangono (churn = 0). Questo risultato suggerisce che il modello è altamente affidabile per la classificazione del comportamento dei clienti.

```
Area under the curve: 0.9018
```

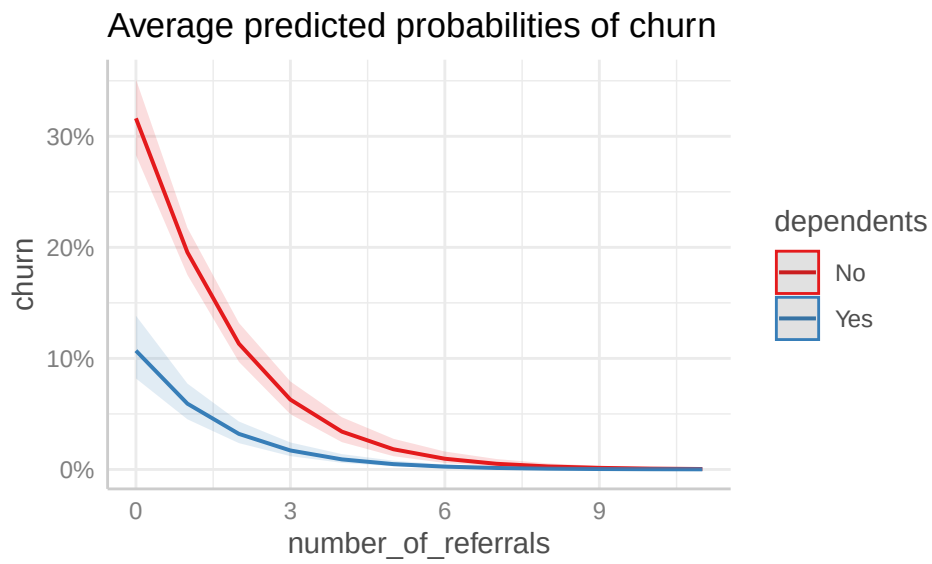
Passiamo ora all'analisi degli effetti marginali per quantificare come una variazione in una variabile indipendente influenzi il valore atteso della variabile dipendente, gli effetti marginali aiutano ad interpretare i coefficienti sopratutto quando le relazioni tra predittori e outcome non sono lineari.



Dal grafico notiamo che la probabilità di churn è in generale più bassa per i clienti che pagano con **credit_card** e continua ad abbassarsi con l'aumentare dei mesi del cliente con la compagnia(**tenure in months**); i clienti che usufruiscono del servizio **online backup** hanno più probabilità di rimanere con la compagnia, sopratutto se pagano con credit card, ed in generale sia per i metodi di pagamento sia per il servizio usato i clienti più mesi passano con la compagnia più sono propensi a rimanere.



Andando ad investigare tra le relazioni delle variabili contract e offer notiamo come l'**offerta A** è quella più problematica per tutte le tipologie di contratto perchè ha una probabilità di churn maggiore, inoltre dal grafico possiamo vedere come i clienti che hanno l'**offerta A** e il tipo di contratto **Month-to-month** hanno una probabilità di abbandono che tocca quasi il 50 %, questo è molto preoccupante in termini di business e fa capire che "accoppiare" il contratto month-to-month con l'offerta A è sicuramente un pericolo per la compagnia.



Per quanto riguarda il rapporto tra la variabile **number of referrals** e **dependents** si evince che con poche persone portate in azienda la probabilità è molto alta di abbandono, probabilità che si abbassa quando il numero di persone portate in azienda da parte del cliente sale, contribuisce all'abbassamento della probabilità di churn anche l'avere persone a carico; notiamo che i clienti che hanno portato almeno 3 persone ed hanno persone a carico hanno una probabilità di abbandono quasi pari a zero, mentre per chi non ha persone a carico la probabilità arriva a zero soltanto quando abbiamo clienti che hanno portato in compagnia almeno 6 persone. Questo può sicuramente far riflettere sul corretto funzionamento della pratica "porta un'amico" dato che i clienti che portano più persone sono anche più fedeli, bisogna però agire sui clienti che non hanno nessun a carico e non hanno portato persone in azienda, perchè la loro probabilità di churn è pari al 30 %, in tal senso si potrebbe pensare ad ideare strategie/promozioni ad hoc per questo segmento di clienti, perchè non gestirli può portare un danno all'azienda.

3.4 Regressione Logistica Penalizzata

Nella costruzione di un modello predittivo, uno degli aspetti fondamentali da bilanciare è il **compromesso tra bias e varianza**:

- Un modello **troppo semplice** (alto bias) tende a sottostimare la complessità delle relazioni tra le variabili → *underfitting*.
- Un modello **troppo complesso** (alta varianza) si adatta eccessivamente ai dati di training, ma generalizza male su nuovi dati → *overfitting*.

L'obiettivo è quindi quello di trovare un modello sufficientemente flessibile da cogliere le relazioni reali, ma non così flessibile da adattarsi al rumore dei dati. Per bilanciare bias e varianza usiamo la regolarizzazione, la **regolarizzazione** è una tecnica che consente di controllare la complessità del modello aggiungendo un termine di penalizzazione alla funzione obiettivo, questo può essere molto utile quando abbiamo: molte variabili predittive, possibile multicollinearità tra le variabili, dataset con molti dati. A questo punto dovendo aggiungere un termine di penalizzazione alla regressione parliamo di regressione Penalizzata.

La **Regressione Penalizzata** aggiungendo un termine di penalizzazione (λ) riesce a limitare la complessità del modello e migliorare le generalizzazione su nuovi dati, può prestarsi anche ad una selezione automatica delle variabili; abbiamo tre tipi di regressione penalizzata:

- **Ridge Regression**: Applica una penalizzazione proporzionale al **quadrato** dei coefficienti. È utile in presenza di multicollinearità, ma **non azzera mai i coefficienti**.
- **Lasso Regression**: Applica una penalizzazione proporzionale al **valore assoluto** dei coefficienti. È in grado di **azzerare** quelli meno importanti, facendo una selezione automatica delle variabili.
- **Elastic Net Regression**: Combina le penalizzazioni di Ridge e Lasso. Viene controllata dal parametro α, con $\alpha = 1$ avremo Lasso, con $\alpha = 0$ avremo Ridge.

Ecco la formula generale della funzione obiettivo della regressione penalizzata (Elastic Net):

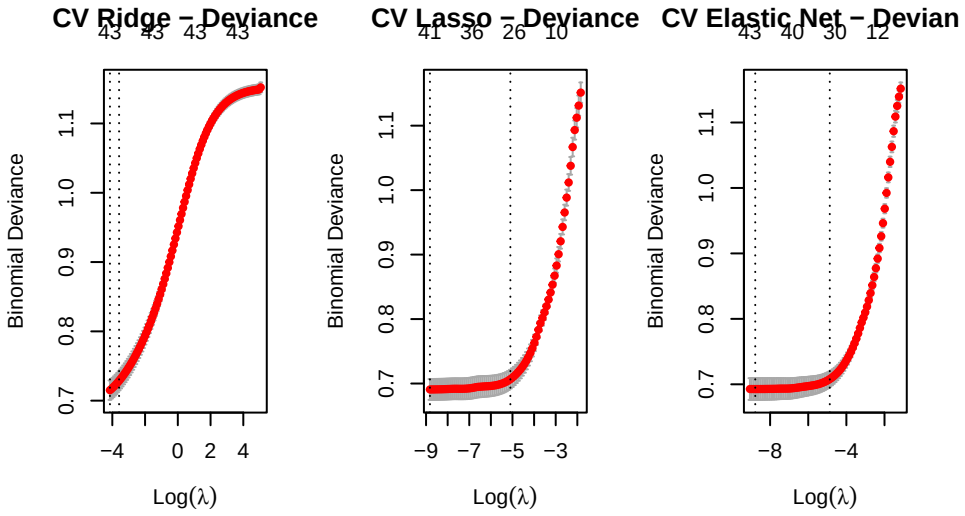
$$\mathcal{L}(\beta) = -\sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] + \lambda \left(\alpha \sum_{j=1}^{k} |\beta_j| + (1 - \alpha) \sum_{j=1}^{k} \beta_j^2 \right)$$

Dove p_i è la probabilità stimata .Cambiando i valori di α daremo più importanza a Lasso o a Ridge; inoltre, il parametro di penalizzazione λ verrà scelto tramite **Cross validation**.

3.4.1 Definizione Modello

In questa fase andremo a definire quale tra i modelli di regressione penalizzata (**Lasso, Ridge, Elastic Net**) è più adatto ai nostri dati, quindi verranno definiti più modelli andando a testare per quale valore di α riceveremo le performance migliori, per performance in questa fase intendiamo Deviance, AUC ed Accuracy, prestando sempre attenzione alla semplicità del modello. L'utilizzo pratico della Regressione Penalizzata in R ha un approccio diverso rispetto agli altri modelli, in questo caso andiamo a definire prima i nostri predittori e la nostra variabile target (per train set e test set), questo perchè la funzione cv.glmnet non accetta come parametro direttamente la formula del modello (anche in questo caso andiamo ad escludere variabili problematiche come city,state,satisfaction score…).

Abbiamo definito quindi un modello di Ridge regression ($\alpha = 0$), un modello Lasso Regression ($\alpha = 1$), ed un modello Elastic Net Regression ($\alpha = 0.5$). Andiamo a graficare la devianza dei modelli in funzione dei diversi valori di λ.



Possiamo notare come la devianza dei tre modelli sia intorno allo 0.7, il fattore che può farci scegliere un modello piuttoso che un altro è il numero di variabili prese in considerazione in relazione al valore di λ e deviance, quindi scegliere anche tra **lambda.min**, che assicura prestazioni ottimali di previsione a discapito della semplicità del modello, e tra **lambda.1se** che migliora quella che è la semplicità del modello andando a selezionare meno variabili penalizzando la capacità predittiva del modello.

3.4.2 Validazione Modello

Per valutare l'efficacia dei tre modelli di regressione penalizzata sviluppati (Lasso, Ridge ed Elastic Net), sono state calcolate diverse metriche di performance sul dataset di test. Come prima metrica andiamo a visualizzare la **deviance** dei tre modelli, utilizzando sia lambda.min che lambda.1se:

```
[1] "Deviance Ridge (lambda.1se): 0.7286"
```

```
[1] "Deviance Lasso (lambda.1se): 0.7068"
```

```
[1] "Deviance Elastic (lambda.1se): 0.7079"
```

```
[1] "Deviance Ridge (lambda.min): 0.7151"
```

```
[1] "Deviance Lasso (lambda.min): 0.6908"
```

```
[1] "Deviance Elastic (lambda.min): 0.6927"
```

Vediamo come i valori per la deviance $Deviance = -2 * Log(Likelihood(model))$ sono piuttosto simili, con i modelli Lasso ed Elastic più performanti sia in termini di lambda.min che in termini di lambda.1se. A questo punto dato che le deviance sono molto simili andremo ad utilizzare **lambda.1se** per favorire una buona interpretabilità del modello andando a selezionare meno variabili, detto ciò andiamo a visualizzare le altre metriche dei modelli usando **lambda.1se** come parametro di penalizzazione. Ecco l'AUC dei tre modelli:

```
[1] "AUC Ridge: 0.8945"
```

```
[1] "AUC Lasso: 0.899"
```

```
[1] "AUC Elastic: 0.8996"
```

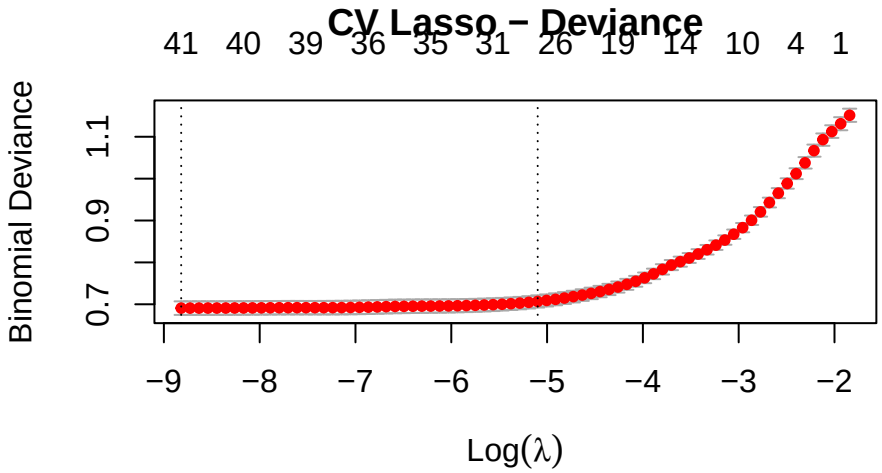
In termini di AUC i modell miglior sono Elastic Net e Lasso, anche se abbiamo performance simili per i tre modelli, questa metrica è utile per misurare la capacità discriminativa del modello di classificazione binaria. Passiamo ora all'Accuracy dei tre modelli:

```
[1] "Accuracy Ridge: 0.831"
```

```
[1] "Accuracy Lasso: 0.832"
```

```
[1] "Accuracy Elastic: 0.832"
```

Anche in termini di Accuracy i modelli migliori appaiono essere i modelli Lasso ed Elastic, a questo punto visto che secondo le metriche utilizzate i modelli appaiono equivalenti andremo a scegliere il modello più adatto in base alla semplicità e quindi in base a quante variabili seleziona (tra Elastic e Lasso), questo perchè a parità di performance è giusto preferire un modello più semplice. Ritorniamo al grafico della devianza dei modelli in funzione dei valori di λ:



Possiamo notare come usando un modello Lasso e utilizzando come termine di penalizzazione **lambda.1se** il modello andrà a selezionare 28 tra le 50 variabili iniziali, dimezzando quasi le variabili da analizzare e rendendo l'analisi dei coefficienti molto più semplice. Una volta scelto il modello di penalizzazione più adatto ai nostri dati (**Lasso**), il parametro di penalizzazione in modo da bilanciare performance e interpretabilità andiamo ad interpretare come influiscono i predittori sulla variabile target **churn**.

3.4.3 Interpretazione del modello

Mentre nel modello di regressione logistica classica i coefficienti rappresentano stime imparziali della variazione attesa nelle log-odds per un incremento unitario del predittore, nel modello Lasso logistico i coefficienti subiscono una contrazione (shrinkage) verso lo zero a causa della penalità L1. Pertanto:

- I valori assoluti dei coefficienti **non sono direttamente confrontabili** (tra il modello classico ed il modello Lasso).
- L'**interpretazione qualitativa** (direzione dell'effetto) rimane valida: segni positivi/negativi indicano rispettivamente aumento o diminuzione della probabilità dell'evento.

Bisogna ricordare che per come è costruito il modello di Lasso regression diventa difficile interpretare i coefficienti alla maniera della regressione logistica classica, l'interpretazione resta la stessa (impatto dei predittori sui log-odds della variabile target) ma data l'introduzione del termine di penalizzazione e la presenza di alcuni coefficienti uguali a zero andando ad interpretare i coefficienti nella maniera fatta in precedenza col modello classico avremo delle **stime distorte**, possiamo sempre interpretare i segni dei coefficienti, che ci dicono se quella variabile contribuisce all'abbassamento o meno della probabilità di churn. A questo punto andiamo a visualizzare i coefficienti del modello Lasso:

```
                   Variabile Coefficiente
1           contractTwo Year   -2.0014344
2       referred_a_friendYes    1.1303508
3              dependentsYes   -1.0690186
4           contractOne Year   -1.0668117
5        number_of_referrals   -0.4178935
6  payment_methodCredit Card   -0.4072694
7                offerOffer E    0.3825047
8        internet_serviceYes    0.3636069
9         online_securityYes   -0.3266104
10          phone_serviceYes   -0.3253446
```

Vedendo i coefficienti del modello penalizzato possiamo dire (in termini qualitativi):

- **Contract Two Year**: Avere un contratto di due anni fa abbassare la probabilità di abbandonare la compagnia.
- **Dependents Yes**: Avere persone a carico riduce le probabilità di churn, e quindi favorisce la fidelizzazione del cliente.
- **Payment method Credit card**: Pagare con carta di credito abbassa le probabilità di churn.
- **Offer E**: Chi usufruisce dell'offerta E ha più possibilità di lasciare l'azienda.
- **Online security Yes**: Avere il servizio di sicurezza online riduce la probabilità di churn del cliente.

Volendo fare un paragone tra l'interpretazione dei coefficienti del modello logistico classico e i coefficienti del modello logistico Lasso,troviamo molti *predittori in comune* che influenzano alla stessa maniera la variabile target nei due modelli distinti, confermando che sopratutto le variabili in comune tra i due modelli contribuiscono alla variazione della probabilità di churn.

3.5 Regressione Logistica Distribuita

Infine, è stato sviluppato un altro approccio all'analisi del nostro dataset che utilizza tecnologie come **Apache Spark** attraverso l'interfaccia **sparklyr** in R. Sebbene il dataset preso in esame può essere trattato anche in ambiente locale tramite l'utilizzo del solo R, in questa sezione è stato utilizzato Spark per proporre un metodo di lavoro distribuito che può essere utilizzato anche qualora i dati avessero presentato molte più osservazioni e molte più variabili. Andremo quindi a caricare il dataset in ambiente Spark, dopodichè verrà attuato un pre-processing apposito per la logica di spark; infine utilizzeremo la tecnica **Divide and Recombine** andando a confrontare i coefficienti calcolati con l'intero dataset in spark e le medie dei coefficienti calcolati a seguito della creazione di diversi modelli a partire da diversi subset del dataset originale.

Il primo passo da fare è aprire una connessione in Spark e copiarci il dataset **telco**, in precedenza già trattato, e visualizzare come spark tratta le nostre variabili:

```
# Source:   SQL [6 x 46]
# Database: spark_connection
  gender   age married dependents number_of_dependents country   state
  <chr>  <dbl> <chr>   <chr>                     <dbl> <chr>     <chr>
1 Male      78 No      No                            0 United S~ Cali~
2 Female    74 Yes     Yes                           1 United S~ Cali~
3 Male      71 No      Yes                           3 United S~ Cali~
4 Female    78 Yes     Yes                           1 United S~ Cali~
5 Female    80 Yes     Yes                           1 United S~ Cali~
6 Female    72 No      Yes                           1 United S~ Cali~
# i 39 more variables: city <chr>, latitude <dbl>, longitude <dbl>,
#   quarter <chr>, referred_a_friend <chr>,
#   number_of_referrals <dbl>, tenure_in_months <dbl>, offer <chr>,
#   phone_service <chr>, avg_monthly_long_distance_charges <dbl>,
#   multiple_lines <chr>, internet_service <chr>,
#   internet_type <chr>, avg_monthly_gb_download <dbl>,
#   online_security <chr>, online_backup <chr>, ...
```

3.5.1 Preparazione dati Spark

Possiamo notare come le variabili che in locale avevamo come **factor** con diversi livelli ora in spark sono salvate come variabili **chr** o **dbl**, questo perchè spark non ammette la tipologia factor a dispetto di R in locale, bisogna quindi trasformare le variabili che useremo nel nostro modello utilizzando la codifica one_hot_encoding. La **one-hot encoding** è una tecnica di codifica delle variabili categoriche in un formato numerico, dove solo una colonna alla volta assume il valore 1 mentre le altre restano a 0, preservando l'informazione categoriale. Nel pacchetto sparklyr la codifica one-hot avviene in due fasi distinte:

1. **Indicizzazione delle categorie** (ft_string_indexer): Le variabili categoriche di tipo stringa vengono prima convertite in interi (es. “No” -> 0, “Yes” -> 1). Questo passaggio è necessario perché Spark ML accetta solo input numerici per la codifica one-hot.

2. **One-hot encoding vero e proprio** (ft_one_hot_encoder): Gli interi prodotti dallo string_indexer vengono poi trasformati in vettori sparsi binari. Ad esempio, una categoria indicizzata come 2 in una variabile con 4 categorie sarà trasformata in un vettore tipo [0 ,0 ,1 ,0].

3.5.2 Definizione Modello

Avremo quindi all’interno del nostro dataset in spark due nuove colonne per ogni variabile categoriale trattata, che verranno utilizzate per il nostro modello. Visto che la variabile target resta **churn** anche in questo caso useremo una regressione logistica, andando ad includere le variabili più significative viste in precedenza, di conseguenza il nostro modello sarà:

$$\begin{aligned}\text{churn} \sim\ & \text{age} + \text{dependents onehot} + \text{referred a friend onehot} + \text{tenure in months} \\ & + \text{number of referrals} + \text{offer onehot} + \text{phone service onehot} \\ & + \text{online security onehot} + \text{online backup onehot} + \text{premium tech support onehot} \\ & + \text{contract onehot} + \text{monthly charge} + \text{payment method onehot}\end{aligned}$$

3.5.3 Divide and Recombine

Abbiamo strutturato il modello per l’intero dataset che abbiamo in Spark, ora la cosa da fare è applicare la tecnica **Divide and Recombine** per valutare possibili differenze marcate tra i coefficienti del modello che lavora con tutto il dataset ed i coefficienti prodotti dai diversi modelli che lavorano con subset del dataset. Quindi è stata definita una funzione per creare i subset e su questi costruire un modello che ha le stesse specifiche del modello sull’intero dataset, nello specifico la funzione di Resampling crea 100 subset, ognuno con il 10 % dei dati presi dal dataset originale, e per ogni subset verra creato un modello.

3.5.4 Statistics recombination

Divide and Recombine (D&R) è una strategia di analisi statistica che consiste nel dividere il dataset in più sottocampioni, eseguire analisi indipendenti su ciascun sottocampione e poi **combinare (recombine)** i risultati per ottenere stime globali. Questo approccio è particolarmente efficace in ambienti di calcolo distribuito come Spark, dove il dataset non viene elaborato tutto in una sola volta, ma è suddiviso tra diversi nodi. Applicata la funzione di Resampling avremo una matrice con i coefficienti derivanti dai 100 subset creati, per confrontare i coefficienti derivanti dalle diverse prove con i coefficienti dell modello intero andiamo a fare la medie dei diversi coefficienti per le 100 prove ed infine confrontiamo le medie dei coefficienti con i coefficienti del modello che lavora con l’intero dataset:

```
# A tibble: 20 x 3
   Coefficient                          Coef.subsets Coef.whole.dataset
   <chr>                                       <dbl>              <dbl>
 1 (Intercept)                                -3.18              -3.02
 2 age                                         0.0130             0.0124
 3 dependents_onehot_No                        1.53               1.47
 4 referred_a_friend_onehot_No                -2.01              -1.89
 5 tenure_in_months                           -0.0261            -0.0264
 6 number_of_referrals                        -0.668             -0.597
 7 offer_onehot_None                           0.312              0.254
 8 offer_onehot_Offer B                        0.0989             0.0990
 9 offer_onehot_Offer E                        0.670              0.584
10 offer_onehot_Offer D                       -0.196             -0.245
11 offer_onehot_Offer A                        0.438              1.18
12 phone_service_onehot_Yes                   -1.09              -1.01
13 online_security_onehot_No                   0.519              0.479
14 online_backup_onehot_No                     0.296              0.277
15 premium_tech_support_onehot_No              0.586              0.513
16 contract_onehot_Month-to-Month              1.48               1.37
17 contract_onehot_Two Year                   -1.43              -1.32
18 monthly_charge                              0.0357             0.0339
19 payment_method_onehot_Bank Withdra~        -0.724             -0.611
20 payment_method_onehot_Credit Card          -1.16              -1.03
```

Grazie alla tabella possiamo notare come le medie dei coefficienti ottenuti dai sottocampioni sono risultate molto simili ai coefficienti stimati sull'intero dataset (differenze < 0.2), dimostrando la **consistenza delle stime**, l'**assenza di dipendenza da specifici sottoinsiemi di dati** e la validità del modello anche in un contesto distribuito.

4 Conclusions

Il presente lavoro intende offrire una visione complessiva del problema dell'abbandono dei clienti nel settore delle telecomunicazioni, analizzando i principali fattori che influenzano tale fenomeno e proponendo diversi approcci per il trattamento dei dati e la costruzione dei modelli.

Dall'analisi condotta emergono alcuni fattori determinanti nella decisione dei clienti di **disdire i propri abbonamenti:**

- **Tipologia di offerta**: i clienti che sottoscrivono l'**offerta A** mostrano una maggiore propensione all'abbandono. Questo potrebbe indicare problematiche specifiche legate a tale offerta, che sarebbe opportuno approfondire e correggere per ridurre il tasso di *churn*.
- **Tipologia di contratto**: come prevedibile, i **contratti mensili** sono associati a una maggiore probabilità di abbandono, mentre i contratti **annuali e biennali** tendono a fidelizzare maggiormente i clienti.

Al contrario, altri fattori sembrano contribuire a **ridurre il rischio di abbandono:**

- **Online Backup**: i clienti che usufruiscono del servizio di backup online risultano meno inclini a disdire il contratto. Questo potrebbe essere dovuto al valore aggiunto del servizio o alla difficoltà di migrazione dei dati, rendendo la permanenza più conveniente. In quest'ottica, offrire il backup anche a tariffe agevolate potrebbe aumentare la fidelizzazione.
- **Presenza di persone a carico**: i clienti con almeno una persona a carico tendono ad abbandonare meno. Questo potrebbe essere legato al desiderio di garantire continuità nei servizi ai membri del nucleo familiare.
- **Metodo di pagamento**: l'analisi mostra che chi paga con **carta di credito** è più fedele rispetto a chi utilizza l'**assegno postale**. La compagnia potrebbe incentivare l'uso di carte di credito, ad esempio con sconti o promozioni, per migliorare la retention.

I principali fattori predittivi del *churn* sono risultati consistenti tra i **modelli utilizzati**. Il lavoro ha esplorato **tre approcci distinti**:

- **Regressione logistica classica**: si tratta dell'approccio più noto e diretto per la modellazione di variabili binarie. È particolarmente efficace quando il numero di osservazioni non è eccessivamente elevato e il numero di variabili è contenuto. Questo modello fornisce **coefficenti interpretabili** e permette un'analisi dettagliata del ruolo di ciascuna variabile indipendente. Tuttavia, può soffrire in presenza di molteplici variabili poco informative.

- **Regressione logistica penalizzata**: questo approccio introduce un termine di penalizzazione (LASSO) che consente di **controllare l'overfitting**, migliorare la generalizzazione del modello e, soprattutto, effettuare **selezione automatica delle variabili**; è particolarmente utile in contesti dove sono presenti molte variabili, anche potenzialmente ridondanti.
- La **Regressione logistica distribuita**, adatta a contesti *big data*, dove non è possibile trattare l'intero dataset in memoria. L'approccio *Divide and Recombine*, applicato in ambiente distribuito con Spark, ha evidenziato come la media dei coefficienti stimati su più sottocampioni sia **coerente con quelli ottenuti sull'intero dataset**, confermando così **la stabilità e l'affidabilità dei risultati**.

In conclusione, l'integrazione di tecniche classiche, penalizzate e distribuite ha permesso non solo di comprendere meglio le dinamiche del *churn*, ma anche di dimostrare la flessibilità degli strumenti statistici nel trattare **problemi reali e complessi**, anche in presenza di **grandi volumi di dati**.