Progetto Data Mining

Agostino Tassan Mazzocco, 833933,
Gianluca Borchielli, 833003,
Francesco Pio Sacco 837029

6 febbraio 2023

1 Introduzione

Il problema in esame riguarda la previsione della durata, in secondi, delle chiamate in uscita effettuate dai clienti di un'azienda di telecomunicazioni, riferite al mese successivo rispetto a quello corrente, per cui sono disponibili le informazioni.

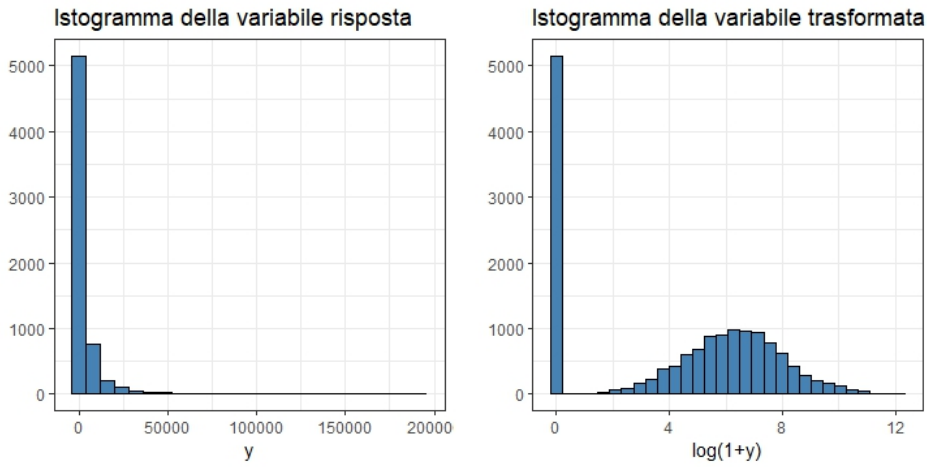

Figura 1: Istogramma della variabile risposta e della variabile risposta trasformata.

Nella figura 1 sono rappresentati gli istogrammi della variabile risposta in scala reale e in scala logaritmica. È evidente come la distribuzione sia sbilanciata verso lo zero. Infatti, nel training set, il 33.6% delle osservazioni presenta un valore della risposta pari a zero.

Alla luce di quanto appena affermato, questo progetto considera due diversi approcci:

- il primo intende prevedere la risposta per il test set, facendo direttamente regressione considerando tutto il dataset.
- il secondo, invece, si pone l'obiettivo di mitigare l'effetto dell'elevata numerosità di osservazioni che presentano un valore della risposta pari a zero, applicando in fase preliminare un algoritmo di classificazione binaria, in cui la risposta (y) viene ricodificata nel seguente modo:

$$y_{bin} = 0, \;\; se \;\; y = 0$$

$$y_{bin} = 1, \;\; se \;\; y > 0$$

per poi analizzare separatamente i due gruppi.

2 Data Preparation

Il dataset analizzato non presentava dati mancanti, tuttavia, è stato necessario risolvere alcuni problemi. Innanzitutto la variabile categoriale *activ.area* presentava un livello rappresentato da una sola osservazione, il quale è stato accorpato alla categoria più rappresentata all'interno del dataset.

Le variabili *q03.in.dur.tot* e *q09.out.dur.peak* presentavano rispettivamente tre e due valori negativi. Essendo variabili riferite a una durata, esse non possono assumere valori minori di zero, pertanto, i suddetti valori, sono stati sostituiti dalle medie, per singola osservazione, delle stesse variabili riferite agli altri otto mesi.

Infine, la varibile risposta e le variabili numeriche riferite ai nove mesi, sono state trasformate secondo la seguente trasformazione logaritmica:

$$y = log(1 + x)$$

3 Modelling Process

In questa fase dell'analisi verrano presentati i modelli previsivi utilizzati. Dapprima seguendo il primo dei due approcci presentati nel paragrafo 1 e successivamente adottando un algoritmo di classificazione preliminare. Le valutazioni dei modelli sono state eseguite mediante 5-folds cross validation.

3.1 Regressione

Nella figura 2 è presentato il flusso di lavoro effettuato. Dopo aver ottimizzato il Random Forest e il Gradient Boosting basato sugli alberi di regressione, è stato eseguito un ensemble tra i due.

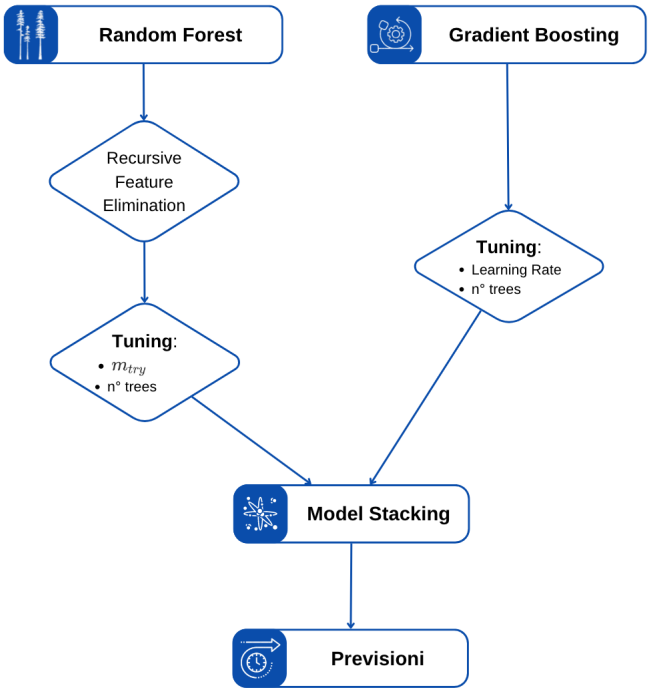

Figura 2: Diagramma a flusso del processo di modellizzazione.

- **Random Forest**: Recursive Feature Elimination: per selezionare un subset delle variabili, basato sul random forest, è stato implementato il seguente algoritmo:

  1. **Begin**: Tuning del modello pieno
  2. Calcolo dell'*importance score* delle variabili, basato sull'impurità, e ranking delle stesse
  3. Eliminazione del 10% delle variabili, a partire da quella con minore importanza
  4. Tuning del modello con le variabili selezionate dall'iterazione precedente
  5. Ripetere ricorsivamente dallo step 2

6. **stop**: numero di variabili pari a 5.

- **Gradient Boosting**: tuning dei seguenti parametri: numero di alberi e learning rate.
- **Model Stacking**:
  - Calcolo delle previsioni dei singoli modelli tramite 5-folds Cross Validation.
  - Stima dello *stacking model*, tramite regressione lineare delle previsioni ottenute al punto precedente sulle vere risposte.

3.2 Classificazione + Regressione

Come suggerito nell'analisi presentata da Azzalini e Scarpa [1], il processo di modellizzazione seguito, presentato nella figura 3, si compone di due fasi principali:

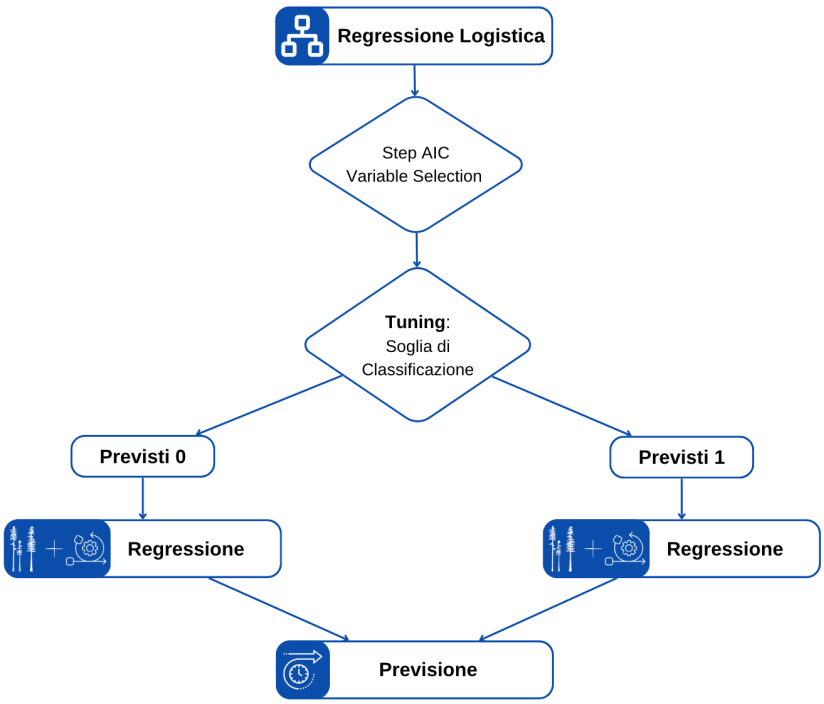

Figura 3: Diagramma a flusso del processo di modellizzazione

- **Classificazione**: l'obiettivo è quello di attribuire ad ogni osservazione la probabilità di assumere un valore della risposta diverso da zero. Sulla

base dei risultati ottenuti, i dati di train e di test sono stati divisi in due gruppi, secondo una soglia ottimizzata in fase di tuning.

- **Regressione**: considerando separatamente i due subset ottenuti in fase di classificazione, vengono addestrati due modelli di regressione differenti. I modelli utilizzati fanno riferimento a quanto descritto nel paragrafo 3.1.

La fase di classificazione si compone dei seguenti passaggi:

- **Variable Selection:** la selezione delle varibili è stata eseguita mediante l'algortimo *step AIC backward*.
- **Tuning:** sono state valutate alcune soglie, per ciascuna delle quali è stato diviso il dataset considerando le probabilità previste di appartenere ai due gruppi. L'obbiettivo consiste nel di minimizzare l'errore di previsione finale ottenuto tramite regressione su entrambi i dataset.

La fase di regressione segue la stessa procedura descritta nel paragrafo 3.1.

4 Risultati

4.1 Regressione

La metrica utilizzata per valutare l'errore di previsione di ciascun modello è la seguente:

$$\sum_{i=1}^{n}(log(1+y) - log(1+\hat{y}))^2$$

dove y rappresenta il vero valore della variabile risposta e $\hat{y}$ indica il valore della risposta, stimato dal modello.

Nella tabella 1 sono risportati i risultati dell'algoritmo di Recursive Feature Elimination. Se ne può evincere che il migliore subset è composto da x variabili e i rispettivi iperparametri saranno: $m_{try} = 15$ e $n^o alberi = 2000$.

n° variabili	$\mathbf{m_{try}}$	**n° alberi**	**Training Error**	**Validation Error**
99	20	2000	9672	11860
89	20	2000	9484	11845
80	20	2000	9367	11803
72	20	2000	9217	11802
65	20	1000	9196	11797
59	15	2000	9466	11770
53	15	1500	9774	11845
48	10	1500	10370	11843
43	10	2000	10697	11832

Tabella 1: Risultato dell'algoritmo di Recurive Feature Elimination basato sul Random Forest. Sono riportati i primi 9 risultati.

Per quanto riguarda il modello di Gradient Boosting, nella tabella 2 sono riportati i risultati del grid tuning basato su una griglia 24x7, che fa variare il Learning Rate, il numero di alberi considerati e la percentuale di covariate da estrarre per ogni albero.

Learning Rate	**Vars_bytree**	**n°alberi**	**Validation Error**
0.005	0.50	4000	11838.64
0.005	0.50	3000	11843.71
0.005	0.50	5000	11853.37
0.005	0.50	6000	11873.23
0.01	0.50	3000	11873.48
0.001	0.50	6000	11890.27
0.005	1	3000	11909.90
0.001	0.50	5000	11915.63
0.01	0.50	4000	11919.06
0.001	1	6000	11927.92

Tabella 2: Risultato del tuning effettuato sugli iperparametri del Gradient Boosting basato sugli alberi. Sono riportate le migliori 10 combinazioni.

Il risultato migliore considera come iperparametri del modello un valore pari a 0.005 per il Learning Rate, un valore pari a 4000 per il numero di alberi e ogni albero verrà costruito estraendo il 50% delle variabili.

Infine nella tabella 3 sono riportati i valori finali degli iperparametri del modello finale, costruito tramite questo primo approccio.

Random Forest		**Gradient Boosting**			**Model Stacking**		**Errore**	
m_{try}	*n° alberi*	*Learning Rate*	*Vars_bytree*	*n°alberi*	β_{rf}	β_{boost}	*Training Error*	*Validation Error*
15	2000	0.005	0.5	5000	0.5854	0.4243	13425.29	11662.51

Tabella 3: Iperparametri e valutazione del modello finale.

dove β_{rf} e β_{boost} rappresentano i pesi attribuiti dal meta-algoritmo di ensemble ai due modelli.

4.2 Classificazione + Regressione

Come descritto nel paragrafo 3.2, un punto cruciale di questa procedura è la scelta della soglia per la quale dividere i dati in due subset. Il risultato del tuning effettuato in questa fase è disponibile nella tabella 4, dove il risultato migliore conduce alla scelta di una soglia pari a 0.3.

Soglia	Validation Error 0	Validation Error 1	Validation Error Tot
0.3	2490.069	9197.481	11687.55
0.4	2862.264	8843.451	11705.72
0.5	2669.107	9047.001	11716.11
0.6	3664.358	8087.302	11751.66
0.7	4604.013	7027.362	11631.38

Tabella 4: Risultato del tuning sulle soglie.

Il modello a due stadi finale (classificazione + regressione) è presentato nella tabella 5, nella quale sono riportati gli iperparametri risultato del tuning.

Regressione Logistica	**Random Forest**		**Gradient Boosting**			**Model Stacking**		**Errore**	
$soglia = 0.3$	m_{try}	n° *alberi*	*Learning Rate*	*Vars_bytree*	n°*alberi*	β_{rf}	β_{boost}	*Validation Error*	*TOT Validation Error*
Gruppo 0	15	1000	0.001	0.5	3000	0.5259	0.4497	2490.069	11687.55
Gruppo 1	20	2000	0.005	0.5	2000	0.5624	0.4478	9197.481	

Tabella 5: Iperparametri e valutazione del modello finale con classificazione.

4.3 Conclusione

Nella tabella 6 è possibile confrontare i due modelli finali, sulla base dell'errrore di training e di validation prodotti dalle previsioni. Quello che sembra essere il modello migliore è quello basato solo sulla regressione.

Modello	Training Error	Validation Error
Regressione	13425.29	11662.51
Classificazione + Regressione	13713.41	11687.55

Tabella 6: Confronto tra i due modelli finali.

Riferimenti bibliografici

[1] Azzalini Scarpa. Analisi dei dati e data mining. pages 105–117, 2004.