Michela Ropele

Data Mining: Warlog

1. Quesito d'analisi

L'analisi delle situazioni di guerra è certamente importante sia per motivi militari, che per motivi umanitari al fine di aiutare le popolazioni colpite da eventi e attacchi militari. Il dataset presente nel file warlog.csv contiene un sottoinsieme di rapporti militari relativi alla guerra in Iraq tra il 2004 e il 2009, pubblicato da WikiLeak nel 2010.

Sono disponibili le seguenti variabili:

report_key | testo: identificativo del rapporto

to_timestamp | timestamp: data di rilascio del rapporto (aggiornato al minuto)

Type | text: Classificazione degli eventi in ogni rapporto

category | text: Classificazione specifica di ogni rapporto

region | text: Area geografica dove si è verificato l'evento

attack_on | text: obiettivo dell'evento/attacco riportato nel report

coalition_forces_wounded | integer: n. unità delle forze di coalizione ferite nell'evento/attacco

coalition_forces_killed | integer: n. unità delle forze di coalizione uccise nell'evento/attacco

iraq_forces_wounded | integer: n. unità delle forze irachene ferite nell'evento/attacco

iraq_forces_killed | integer: n. unità delle forze irachene uccise nell'evento/attacco

civilian_wia | integer: n. di civili feriti nell'evento/attacco

civilian_kia | integer: n. di civili uccisi nell'evento/attacco

enemy_wia | integer: n. di unità nemiche ferite nell'evento/attacco

enemy_kia | integer: n. di unità nemiche ferite nell'evento/attacco

enemy_detained | integer: n. di unità nemiche catturate nell'evento/attacco

total_deaths | integer: numero totale di morti nell'evento/attacco

st_x | numeric: longitudine della posizione dell'evento/attacco

st_y | numeric: latitudine della posizione dell'evento/attacco

Tra le diverse analisi è di interesse capire cosa caratterizza i diversi tipi (Type) di eventi descritti nei rapporti, cercando di identificare, tra le altre caratteristiche, se in determinati intervalli temporali, in alcune aree geografiche o in specifiche posizione spaziali vi è maggior probabilità di osservare eventi criminali, azioni nemiche, azioni amiche o caratterizzate dal cosiddetto "fuoco amico", eventi di non combattimento, incidenti sospetti, minacce, ecc.

2. Dati

Il dataset presenta inizialmente 52048 osservazioni e 18 variabili. Le unità statistiche corrispondono ai singoli rapporti militari registrati. Non ci sono osservazioni ripetute.

La variabile risposta è il tipo di evento descritto nel rapporto, qualitativa con 8 modalità fortemente sbilanciate. Rinomino le modalità della risposta che risultano codificate in modo errato e unifico le azioni "Friendly Fire" nella categoria "Friendly Action". La distribuzione di frequenze relative della risposta è la seguente:

Criminal Event	**Enemy Action**	**Explosive Hazard**	**Friendly Action**
0.4680	0.2244	0.1898	0.0926
Non-Combact Event	**Other**	**Suspicious Incident**	**Threat Report**
0.0213	0.0028	0.0006	0.0004

L'obiettivo dell'analisi è individuare quali fattori caratterizzano i diversi tipi di eventi, verificando se la loro probabilità varia in funzione del tempo, dell'area geografica o della posizione spaziale.

Eseguo le seguenti operazioni di pulizia:

- Rimuovo l'identificativo di riga.
- La variabile Category presenta 86 modalità fortemente sbilanciate. Le raggruppo mantenendo 5 modalità, le 4 più numerose e 'Other'.
- La variabile Region ha 8 modalità, di cui una senza etichetta. La rinomino 'Unknown'.
- La variabile temporale restituisce informazioni sulla data e sull'ora del rilascio del rapporto. Estraggo anno e mese dalla data e raggruppo l'orario di produzione del rapporto in quattro fasce della giornata: "mattina", "pomeriggio", "sera" e "notte".

Inoltre, noto che c'è perfetta corrispondenza tra la somma delle variabili 'enemy_kia', 'civilian_kia', 'iraq_forces_killed', 'coalition_forces_killed' e 'total_deaths'. Mantengo tutte le variabili e lascio che siano i modelli a fare selezione. Questa collinearità potrebbe creare problemi nei modelli lineari, ma alcuni algoritmi la gestiscono automaticamente.

Ora il dataset presenta 52048 osservazioni e 19 variabili.

3. Bilanciamento, stima dei pesi e correlazioni

Suddivido il dataset in stima e verifica (70% e 30% delle osservazioni). Poiché l'insieme di stima è fortemente sbilanciato (la classe meno rappresentata ha una frequenza relativa dello 0,4%), scelgo di bilanciare parzialmente le classi e di assegnare pesi alle osservazioni.

Per bilanciare, in parte, le classi nell'insieme di stima, sotto-campiono quelle con una quota maggiore di osservazioni. Successivamente, per non perdere ulteriore informazione, applico dei pesi alle osservazioni.

La seguente tabella riporta la frequenza relativa di osservazioni dell'insieme di stima di partenza, la quota di osservazioni campionate, la frequenza relativa finale dell'insieme di stima e i pesi:

Modalità	Criminal Event	Enemy Action	Explosive Hazard	Friendly Action	Non-Combact Event	Other	Suspicious Incident	Threat Report
% iniziale	46.80	22.44	18.98	9.26	2.13	0.28	0.06	0.04
Quota	0.05	0.08	0.09	0.12	0.4	0.8	1	1
% finale	28.68	22.01	21.13	13.45	10.68	2.73	0.7	0.5
Pesi	0.009	0.012	0.012	0.020	0.024	0.096	0.339	0.487

Osservo due coppie di variabili con correlazione elevata (valore assoluto compreso tra 0.70 e 0.85): civilian_kia con total_deaths, e st_x con st_y (latitudine e longitudine).

Eseguo la standardizzazione separata delle variabili quantitative nei due sottoinsiemi, in modo da rendere i coefficienti confrontabili in termini di una deviazione standard (ad esempio nel modello lineare senza interazioni) ed evitare che alcune variabili influenzino eccessivamente l'analisi nei modelli sensibili alla scala.

4. Prestazioni del modello

Per valutare le diverse previsioni, utilizzo il tasso di errata classificazione, cioè il rapporto di unità scorrettamente classificate sull'insieme di verifica fratto il totale. Se necessario, utilizzerò il criterio AIC per confrontare modelli alternativi.

I parametri di regolazione e i modelli verranno stimati sull'insieme di stima, mentre la loro metrica finale utilizzata per il confronto verrà calcolata sull'insieme di verifica.

5. Modelli

5.1 Modello lineare multivariato

Inserisco l'interazione lineare tra latitudine e longitudine.

Metrica: 0.0673

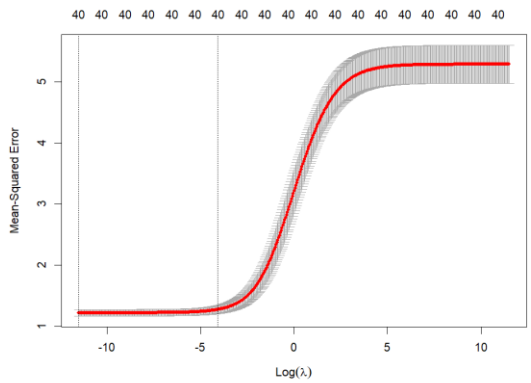
5.2 Multinomiale

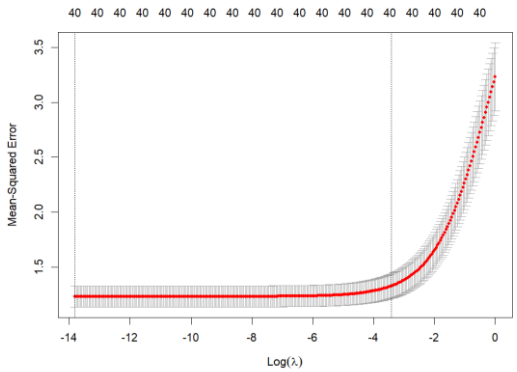
Inserisco l'interazione lineare tra latitudine e longitudine.

Metrica: 0.0269

5.3 Ridge

Stimo il modello tramite convalida incrociata con 10 folds. Uso una griglia di lambda con 500 valori compresi tra 10^-5 e 10^5. Non osservo un minimo. Uso un'altra griglia di 300 valori compresi tra 10^-6 e 1. Il lambda minimo rimane il valore estremo 10^-6 a significare una contrazione minima dei coefficienti del modello lineare. Di seguito i grafici di andamento dell'errore quadratico medio al variare dei valori di lambda.



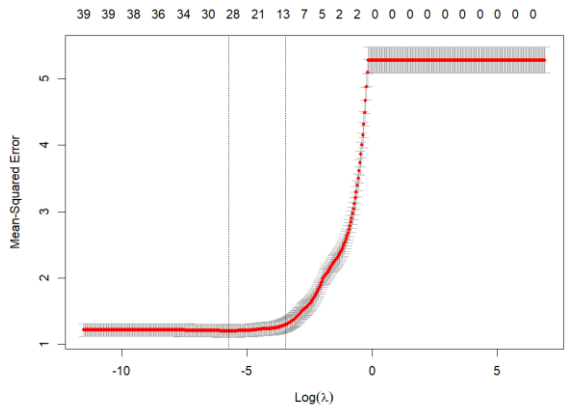


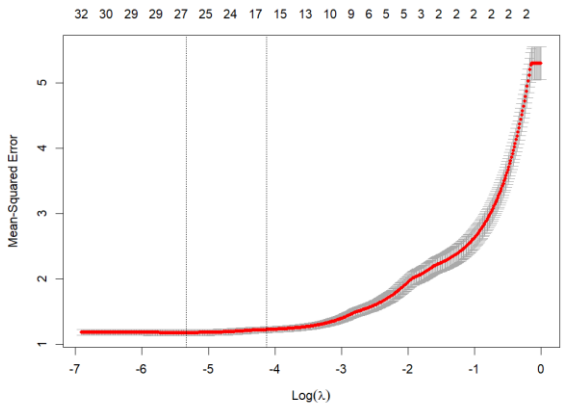
Metrica: 0.2226

5.4 Lasso

Stimo il modello tramite convalida incrociata con 10 folds. Uso una griglia di lambda con 500 valori compresi tra 10^-5 e 10^5. C'è un punto di minimo ben visibile. uso una seconda griglia di 500 valori compresi tra 10^-3 e 1, ed individuo il punto di minimo in corrispondenza di lambda = 0.0048. Di seguito i grafici di andamento dell'errore quadratico medio al variare dei valori di lambda.

Ristimo il modello con tutti i dati di stima ed osservo che questo mi seleziona 28 coefficienti su 41.





Metrica: 0.2230

5.5 LDA

Lo stimo sulle sole variabili quantitative (13 su 18, riduco molto l'informazione). Il modello viene stimato correttamente, ma segnala collinearità tra alcune variabili e produce un tasso di errore elevato.

Metrica: 0.4123223

Nota: QDA

Non stimabile neanche con le sole variabili quantitative a causa di collinearità e carenza di rango nel gruppo 1.

5.6 Modello Additivo con variabili selezionate dal Lasso

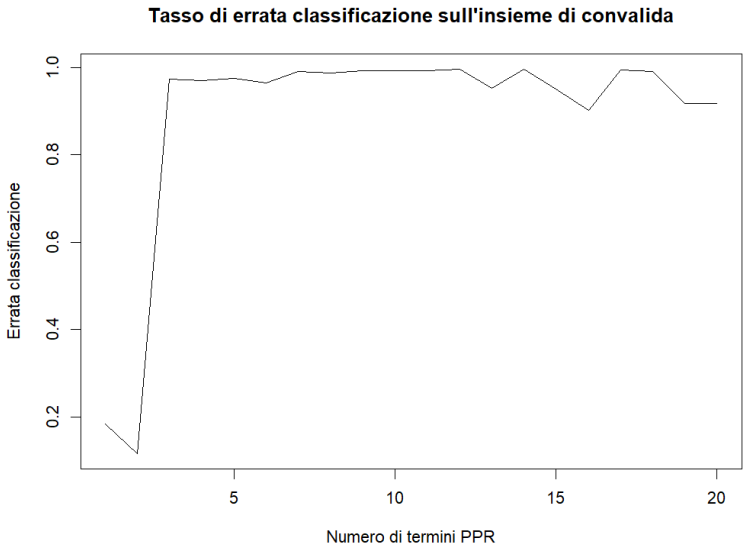
Uso le sole variabili selezionate dal Lasso e una spline bivariata latitudine e longitudine. Imposto il grado del lisciatore a 3.

Metrica: 0.0553

5.7 Regressione Projection Pursuit

Utilizzo un numero di funzioni dorsali (parametro di regolazione) variabile da 2 a 20.

Il grafico dell'errore di classificazione stimato tramite stima-convalida al variare delle funzioni dorsali è il seguente, con un minimo in corrispondenza di 2.

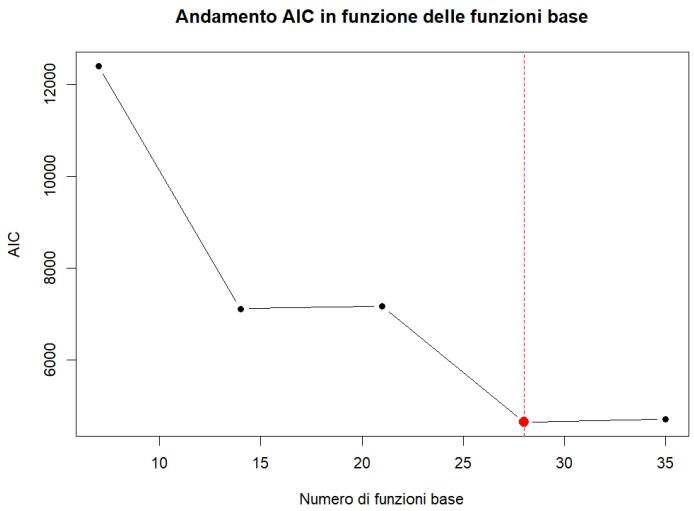


Metrica: 0.04498

5.8 MARS

Testo un numero di funzioni base pari a 7, 14, 21, 28, e 35. Inserisco nel modello solo gli effetti principali, e la possibilità di inserire interazioni tra variabili.

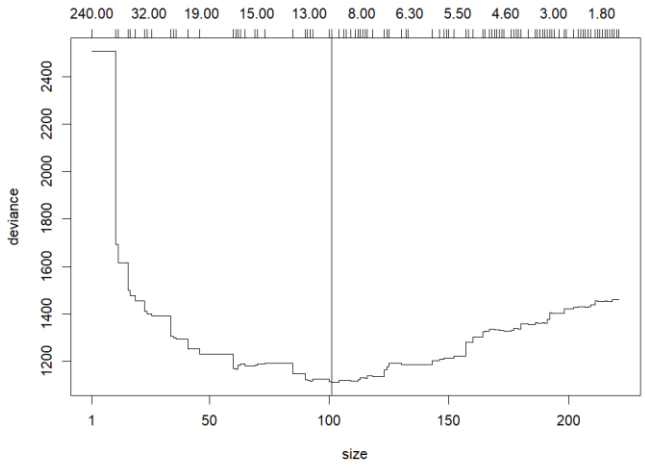
Il seguente grafico mostra l'andamento dell'AIC al variare del numero di funzioni base (complessità), con una linea verticale in corrispondenza del numero ottimo di funzioni base. Il valore scelto è 28.



Metrica: 0.1398

5.9 Albero di classificazione

Stimo alberi per un numero di split variabili da 1 a 221. Di seguito il grafico della devianza al variare del numero di split, con una linea verticale in corrispondenza del valore ottimo 101. Eseguo la potatura in corrispondenza di 102 foglie. Uso l'indice di Gini.

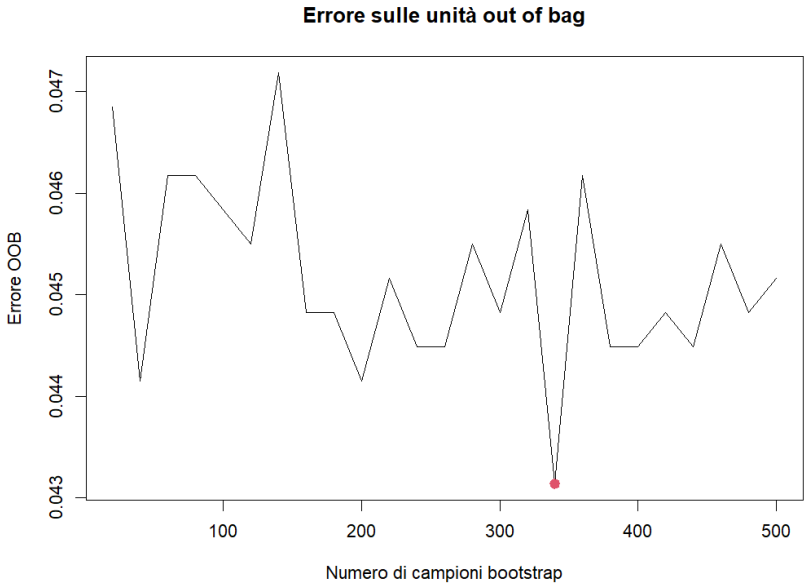


Metrica: 0.1997

5.10 Bagging

Il parametro di regolazione è il numero di campioni bootstrap. Calcolo l'errore su un insieme di valori che va da 20 a 500 con passo 20.

Uso l'errore OOB, calcolato per mezzo della split validation e peso ciascuna classificazione errata 0.1, e scelgo una frazione di dati Out Of Bag di 0.5. Di seguito il grafico dell'errore Out Of Bag al variare del numero di campioni bootstrap.

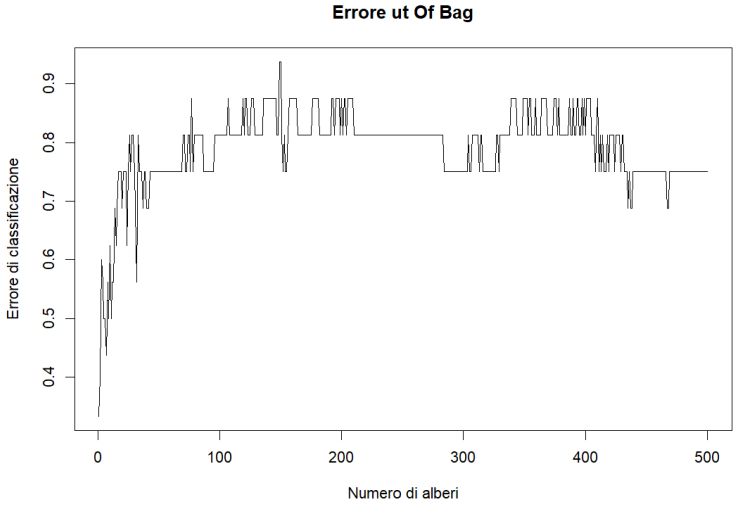


Osservo un minimo in corrispondenza di 340, ma noto una stabilizzazione dell'errore già a 200. Ristimo il modello con tutti i dati e calcolo la metrica d'errore.

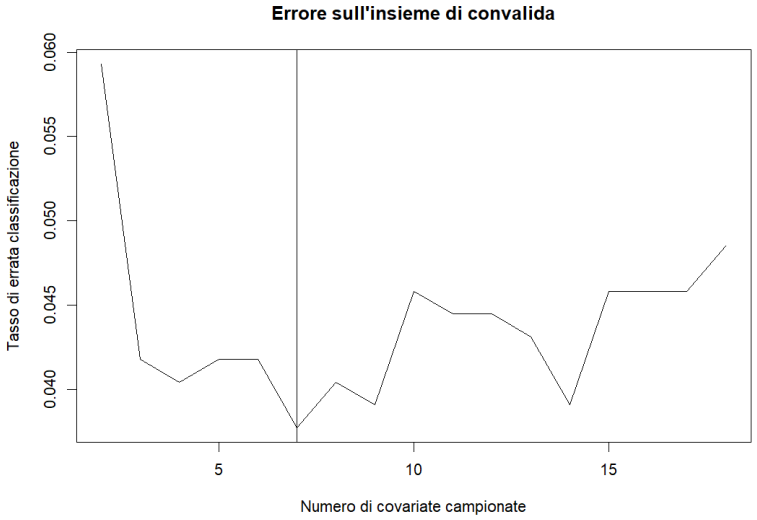
Metrica: 0.02376

5.11 Foresta Casuale

Stimo, tramite convalida incrociata, la foresta per un numero di alberi variabile. Dal grafico sottostante osservo una stabilizzazione dell'errore in corrispondenza del valore 100. Scelgo come valore 200 per restare cautelativa.



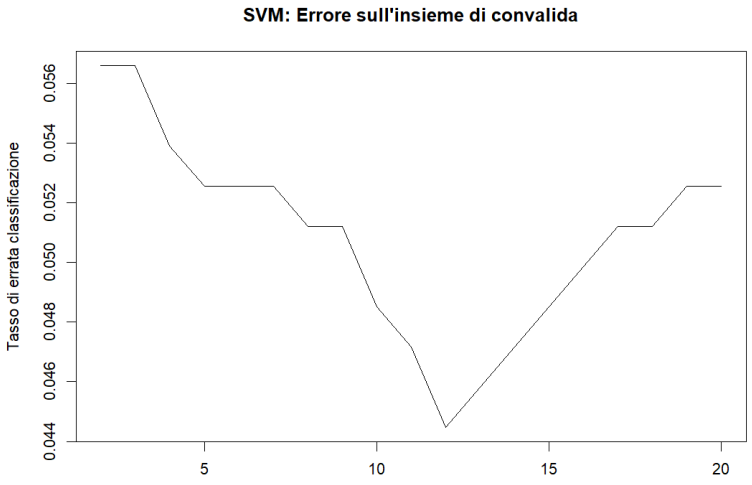
Ora stimo, sempre tramite convalida incrociata, foreste casuali al variare del numero di covariate campionate. Il grafico sottostante rappresenta il tasso di errata classificazione Out Of Bag ottenuto con una linea verticale in corrispondenza del minimo pari a 7.



Stimo il modello finale su tutti i dati di stima ed ottengo la metrica finale.

Metrica: 0.01947

5.12 Support Vector Machine



Stimo il modello al variare della misura di costo usando un nucleo radiale. Testo tramite stima-convalida tutti i valori compresi tra 2 e 20 ed ottengo il minimo del tasso in corrispondenza di 12. Ottengo il modello finale su tutti i dati di stima ed ottengo la metrica sui dati di verifica.

Metrica: 0.02811

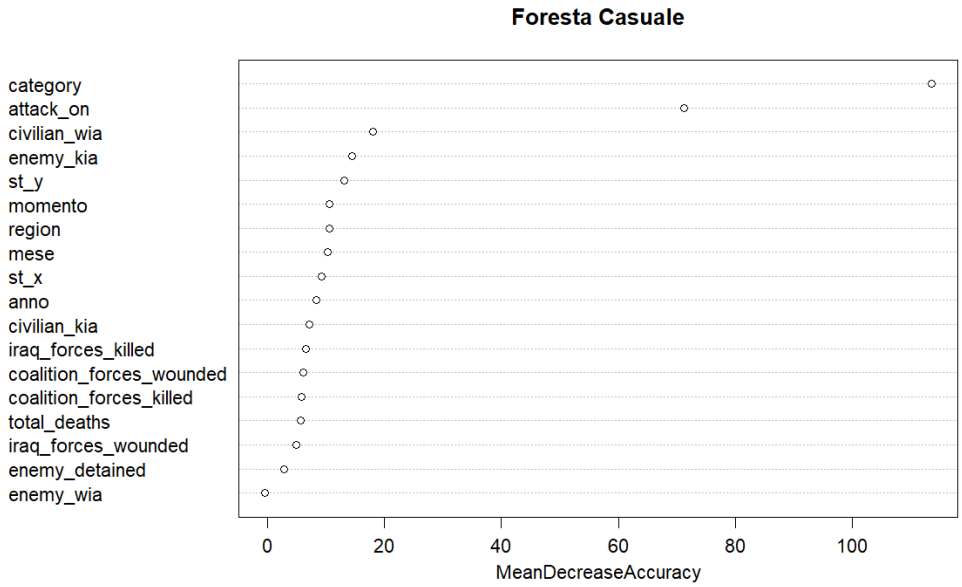
6. Conclusioni

Tutti i modelli sono stati stimati inserendo l'effetto di variazione congiunta tra latitudine e longitudine in modo esplicito, dove possibile, oppure in modo implicito.

Di seguito la tabella dei tassi di errata classificazione per i modelli stimati. Il migliore appare essere la foresta casuale, seguita dal bagging. Al terzo posto il modello parametrico multinomiale.

Modello	**Tasso di errata classificazione**
Foresta Casuale	0.0194
Bagging	0.0236
Modello multinomiale	0.0269
Support Vector Machine	0.0281
Regressione Projection Pursuit	0.0450
Modello additivo con variabili selezionate dal Lasso	0.0553
Lineare multivariato	0.0672
MARS	0.1400
Albero di classificazione	0.1994
Regressione Ridge	0.2226
Regressione Lasso	0.2230
Analisi discriminante lineare	0.4123

Di seguito è riportato il grafico che mostra l'importanza delle variabili secondo la foresta casuale:



Questo grafico mostra quanto cala l'accuratezza del modello se i valori della variabile vengono permutati a caso. Osservandolo è possibile capire cosa caratterizza maggiormente i diversi tipi di eventi descritti.

L'importanza massima è assegnata alla variabile Category, che indica la classificazione specifica di ogni rapporto (attacco, fuoco diretto, omicidio,…) mentre al secondo posto si trova la variabile categoriale relativa al tipo di vittima seguita dal numero di unità civili ferite e dal numero di unità nemiche uccide.

Tutte e variabili successive sono quelle di interesse: latitudine, momento della giornata di produzione del rapporto, regione, mese, longitudine e anno.