E' possibile prevedere lo spopolamento a 10 anni delle città Italiane?

Tripicchio Giovanni, Manoccio Mariachiara

Master Data Science and Artificial Intelligence

Anno Accademico 2023/2024

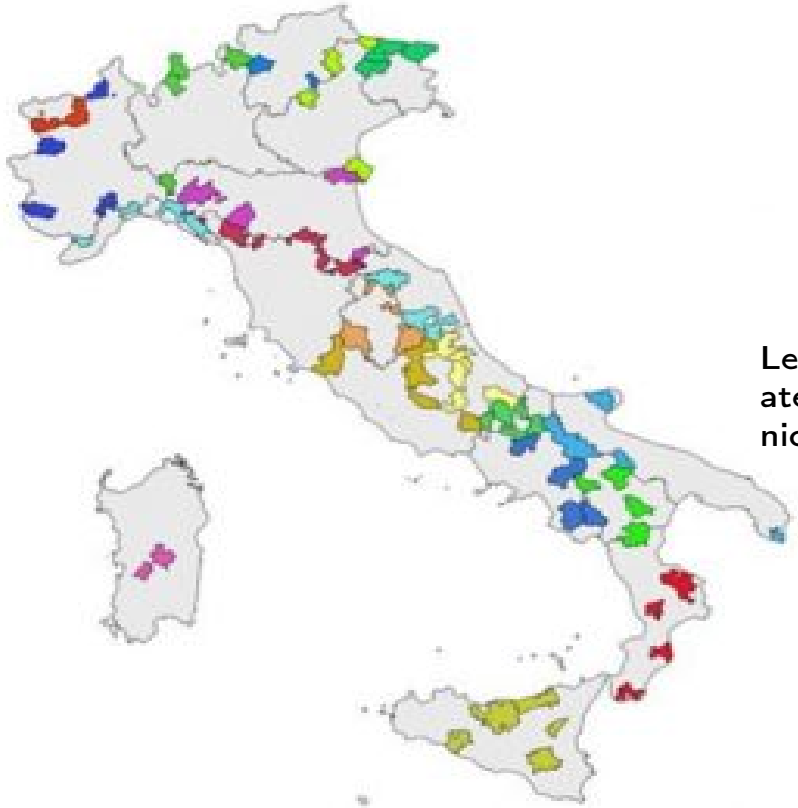
Sommario

Introduzione

In questo progetto, sono stati utilizzati differenti modelli di Machine Learning per investigare le principali features che potrebbero determinare lo spopolamento dei comuni in 10 anni, partendo dalle aree selezionate della Strategia Nazionale per le Aree Interne (SNAI). La SNAI è un'iniziativa del governo Italiano avviata nel 2016 per sviluppare e sostenere le regioni interne del Paese e soprattutto invertire le tendenze dello spopolamento e dell'immigrazione e favorire lo sviluppo locale.

Introduzione

Obbiettivo SNAI: Coinvolgere politiche e le risorse a livelo nazionale e locale per favorire lo sviluppo integrato delle aree interne concentrandosi su economia locale, ambiente, servizi sociali e culturali.

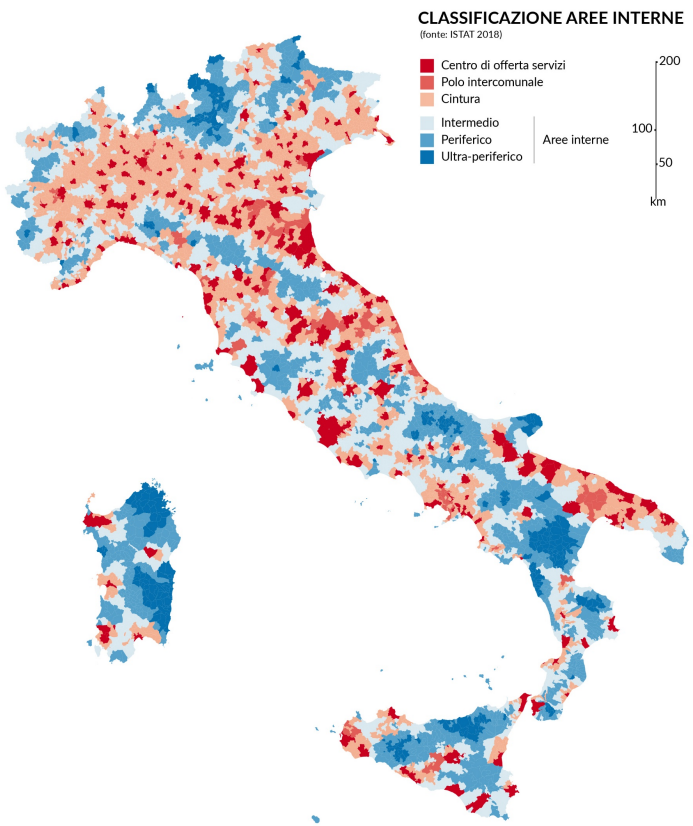


Le 72 aree progetto selezionate. Fonte: Comitato Tecnico Aree Interne, 2019.

Tripicchio Giovanni, Manoccio Mariachiara

Framework SNAI

I Comuni considerati sono quelli marginali, concentrati prevalentemente nelle aree interne del Paese e distanti dai servizi di cittadinanza (ex. Scuola superiore, stazione ferroviaria, ospedali).



Dati

Il dataset utilizzato è quello contenente gli indicatori del censimento ISTAT del 2011 e la popolazione residente per comune nel 2021 ricavata da fonti istat.
Il dataset è basato su 99 indicatori che comprendono diversi campi:
Popolazione: Variazione intercensuaria, Incidenza superficie centri e nuclei abitati, Indice di vecchiaia ecc.
Famiglie: Ampiezza media delle famiglie, Incidenza di anziani soli, ecc.
Condizioni abitative ed insediamenti: Incidenza delle abitazioni in proprietà, Mobilità residenziale, ecc.
Istruzione: Incidenza di adulti con diploma o laurea, Incidenza di analfabeti, ecc.
Mercato del lavoro: Tasso di occupazione, Partecipazione al mercato del lavoro, ecc.
Mobilità: Mobilità giornaliera per studio o lavoro, Mobilità occupazionale, ecc.
Vulnerabilità materiale e sociale: Indice di vulnerabilità sociale e materiale, Incidenza delle famiglie con potenziale disagio economico, ecc.

Features nei modelli

Le features indipendenti considerate nei modelli sono rappresentate dagli indicatori Istat.
La feature target è rappresentata dalla differenza tra la popolazione censita nel 2011 e quella censita nel 2021 per ogni comune identificando la variazione demografica dei comuni in 10 anni.
Alla feature viene assegnata la variabile categorica Yes se c'è stato spopolamento (differenza <0), NO altrimenti.

Prediction Task

La prediction task per ogni i-esima città italiana, è basata sull'applicare al set delle 99 variabili $P1, P2, P3, ..., V9$ una funzione, dipendente dal modello di machine learning scelto, che possa predire lo spopolamento (Abs_i)

$$f(\{P1, P2, P3, ..., V9\}_{i,t-1}) - - > Abs_{i,t} \qquad (1)$$

dove $P1, P2, P3, ..., V9$ sono le variabili esplicative del modello, e Abs è la variabile risposta che indica se c'è stato o meno lo spopolamento.

Modelli ML

- **Elastic Net:** metodo statistico di regressione lineare che include nel modello di regressione statistica classica (OLS) 2 termini di penalizzazione rispettivamente L1 (Lasso) e L2 (Ridge) per la selezione delle caratteristiche del modello, in maniera tale da aumentarne l'accuratezza e ridurne gli errori;
- **Gradient Boosting Machine (GBM):** si basa su una combinazione (ensemble) di predittori deboli, chiamati alberi di decisione, dove i predittori successivi nell'iterazione apprendono dai predecessori. La riduzione della parte di errore commesso sulla varianza e la riduzione del rischio di selezionare l'ipotesi sbagliata, sono i maggiori vantaggi.
- **Random Forest (RF):** metodo ensamble di una famiglia di alberi decisionali di classificazione basati su alberi randomizzati che utilizza sottoinsiemi casuali diversi di caratteristiche ad ogni divisione dell'albero.
- **Neural Network (NN):** modello che utilizza un insieme di unità di input/output connesse, chiamati neuroni e collegati in diversi strati, in cui ogni connessione ha un peso associato, e apprende regolando i pesi per prevedere l'insieme delle caratteristiche associate al modello.

Metodi di sampling

Poiché i dati sono sbilanciati (la categoria YES è 127% maggiore rispetto alla categoria NO), per bilanciare il dataset di train per ogni modello si utilizzano 3 diversi metodi di sampling: Class weights, Up sampling e Rose.

- **Class weights:** alla categoria maggiormente rappresentata assegna un peso maggiore rispetto alla categoria meno rappresentata;
- **Up-sampling:** Aumenta artificiosamente la classe meno rappresentata, anche nella cross validation;
- **Rose:** aumenta il numero dei positivi (o negativi), facendo si che i nuovi positivi(o negativi) siano combinazione lineare dei positivi reali (o negativi);

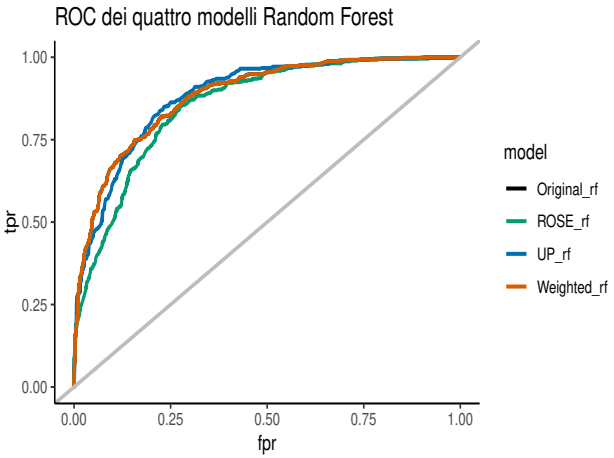
Modello di Train e Test

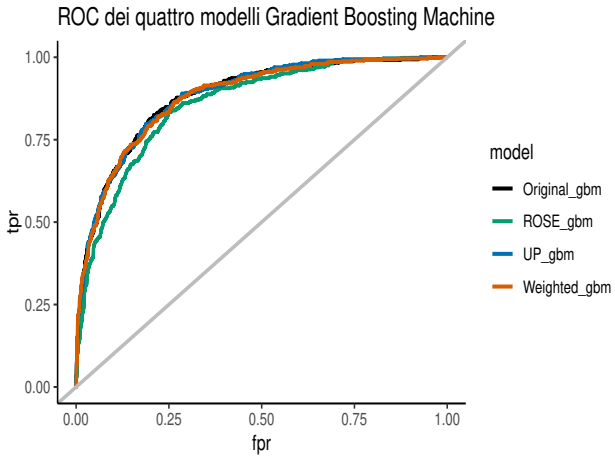
Il dataset originale è stato suddiviso in train (70%) e test out-of-sample (30%) utilizzando il metodo Repeated Cross Validation 10 fold ripetuti 5 volte. La performance del modello e la predizione dello spopolamento entro 10 anni vengono valutate identificando le caratteristiche cruciali tramite feature importance e partial plots.

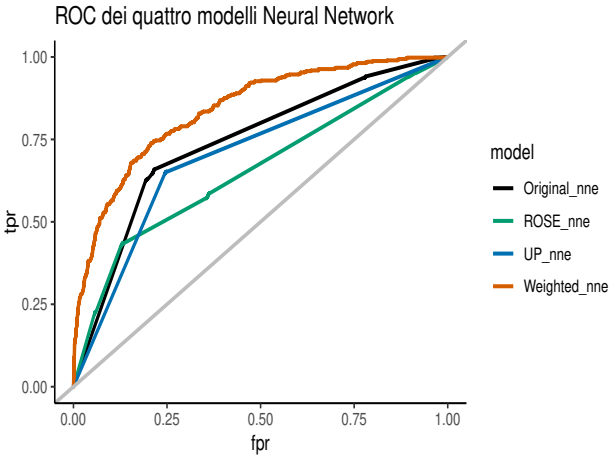
Performance modelli

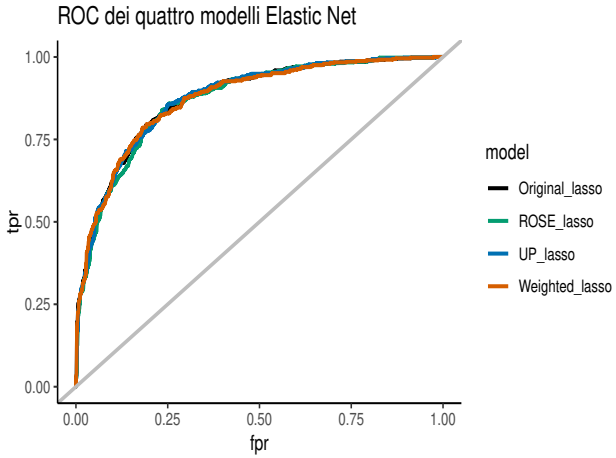
Nelle slide successive sono mostrate le curve ROC e le curve PROC per ogni modello e per i metodi di sampling applicati. Le curve ROC visualizzano la trade-off tra sensibilità e specificità di un modello di classificazione binaria, mentre le curve PROC mostrano la relazione tra precision e recall. In entrambi i casi le curve seguono un trend simile tranne per il modello NN, che mostra delle differenze tra i vari sampling considerati. Per le reti neurali, il modello migliore risulta essere quello con il Class weight sampling.

ROC

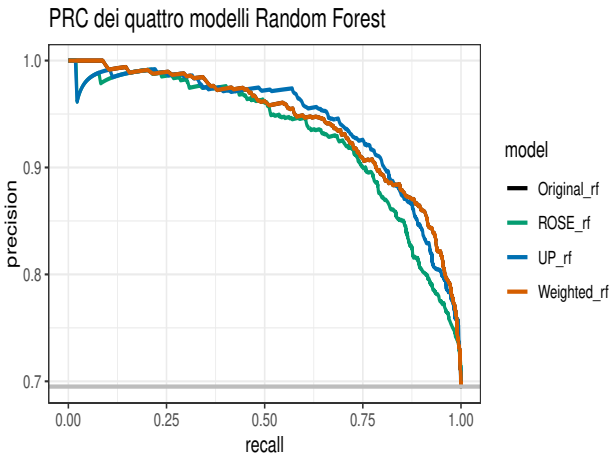

ROC dei quattro modelli Random Forest
tpr
fpr
0.00
0.25
0.50
0.75
1.00
model
Original_rf
ROSE_rf
UP_rf
Weighted_rf

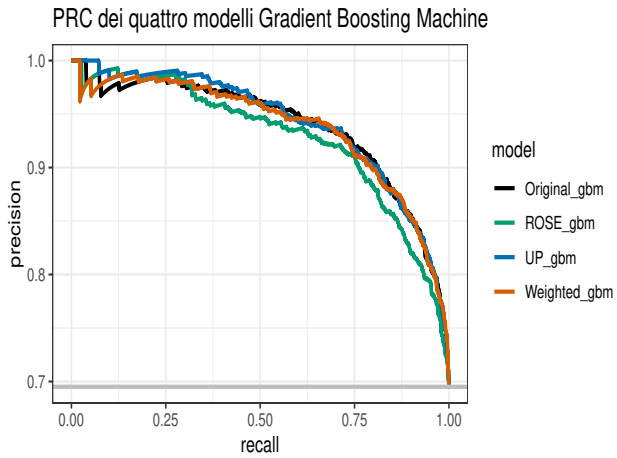

ROC dei quattro modelli Gradient Boosting Machine
tpr
fpr
0.00
0.25
0.50
0.75
1.00
model
Original_gbm
ROSE_gbm
UP_gbm
Weighted_gbm

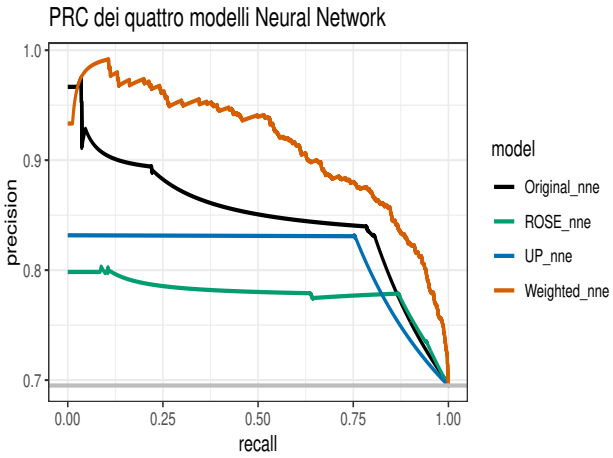

ROC dei quattro modelli Neural Network
tpr
fpr
0.00
0.25
0.50
0.75
1.00
model
Original_nne
ROSE_nne
UP_nne
Weighted_nne

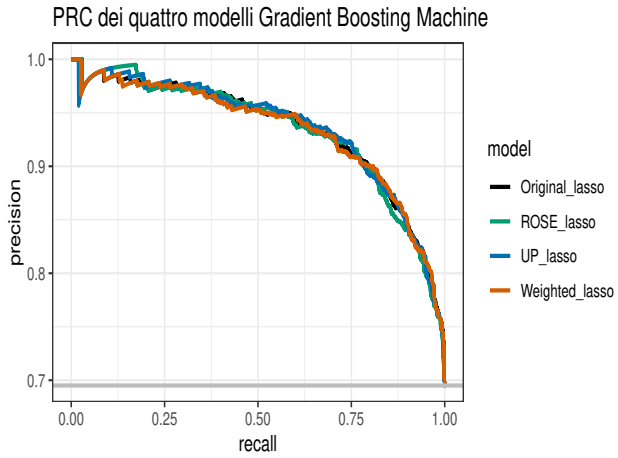

ROC dei quattro modelli Elastic Net
tpr
fpr
0.00
0.25
0.50
0.75
1.00
model
Original_lasso
ROSE_lasso
UP_lasso
Weighted_lasso

PRC


PRC dei quattro modelli Random Forest
precision
recall
model
Original_rf
ROSE_rf
UP_rf
Weighted_rf


PRC dei quattro modelli Gradient Boosting Machine
precision
recall
model
Original_gbm
ROSE_gbm
UP_gbm
Weighted_gbm


PRC dei quattro modelli Neural Network
precision
recall
model
Original_nne
ROSE_nne
UP_nne
Weighted_nne


PRC dei quattro modelli Gradient Boosting Machine
precision
recall
model
Original_lasso
ROSE_lasso
UP_lasso
Weighted_lasso

Misure di performance di predizione

Valori di area sotto la curva (AUC) per PRC e ROC

	GBM		Random Forest		Elastic Net		NN	
	PRC	ROC	PRC	ROC	PRC	ROC	PRC	ROC
Original	0.938	0.882	**0.942**	**0.883**	0.934	0.875	0.848	0.74
Weighted	0.937	0.879	0.942	0.883	0.933	0.874	0.91	0.84
Rose	0.928	0.857	0.929	0.855	0.934	0.87	0.777	0.656
Up-sampling	0.941	0.882	0.942	0.883	0.935	0.875	0.812	0.703

Il Random Forest emerge come il modello migliore poiché il suo valore di AUC si avvicina maggiormente a 1, suggerendo una maggiore efficacia. Al contrario, l'utilizzo di metodi con sampling non ha fornito informazioni aggiuntive o migliorato le prestazioni.

Performance dei modelli

Di seguito si riportano altre metriche ricavate per tutti i modelli.

	Accuracy	NoInfoRate	Kappa	McnemarPValue	Sensitivity	Specificity
Original RF	0.83	0.695	0.587	0.000	0.90	0.659
Weighted RF	0.83	0.695	0.587	0.000	0.906	0.659
Up RF	0.81	0.695	0.546	0.003	0.889	0.64
Rose RF	0.73	0.695	0.462	0.000	0.66	0.887
Original GBM	0.82	0.695	0.572	0.001	0.89	0.66
Weighted GBM	0.80	0.695	0.48	0.000	0.95	0.47
Up GBM	0.80	0.695	0.57	0.000	0.80	0.79
Rose GBM	0.76	0.695	0.50	0.000	0.73	0.84
Original Elastic	0.81	0.695	0.54	0.000	0.90	0.61
Weighted Elastic	0.79	0.695	0.43	0.000	0.97	0.39
Up Elastic	0.79	0.695	0.54	0.000	0.79	0.79
Rose Elastic	0.79	0.695	0.54	0.000	0.78	0.80
Original Neural	0.74	0.695	0.42	0.004	0.80	0.63
Weighted Neural	0.69	0.695	0.00	0.000	1.00	0.00
Up Neural	0.72	0.695	0.38	0.000	0.75	0.65
Rose Neural	0.73	0.695	0.32	0.000	0.87	0.43

Tripicchio Giovanni, Manoccio Mariachiara

Delong's ROC Test

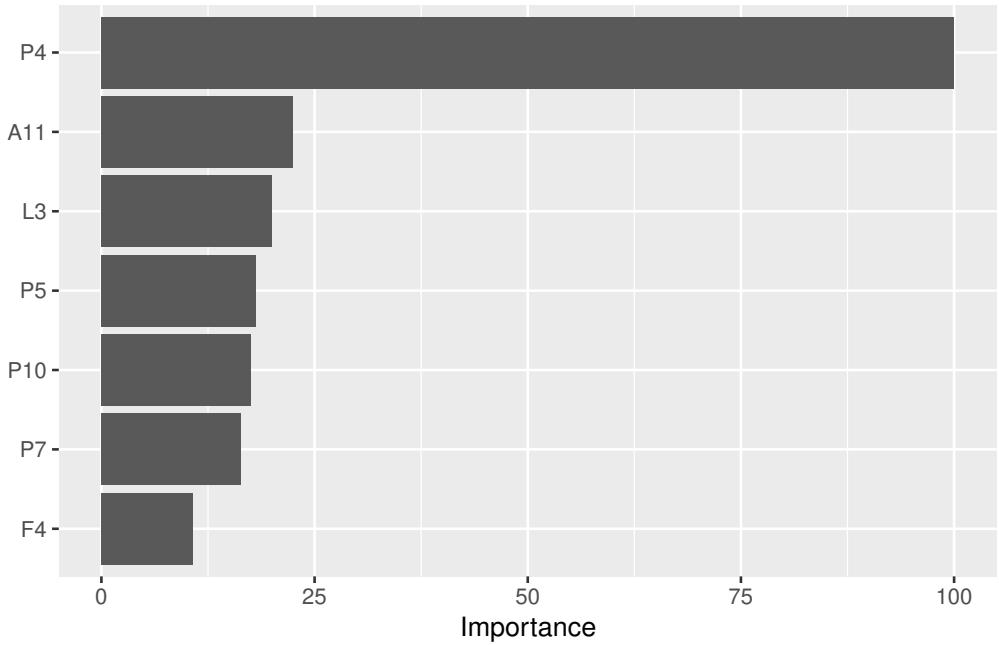
Il Delong's ROC Test confronta le curve ROC di due modelli per valutare differenze significative nell'AUC, utile per determinare se il modello Random Forest ha una performance significativamente diversa rispetto agli altri modelli considerati.

	Z	p-value
Original RF vs *Weighted RF*	0	1
Original RF vs *Up RF*	-0.13	0.89
Original RF vs *Rose RF*	4.7	0.000
Original RF vs *Original GBM*	0.28	0.77
Original RF vs *Weighted GBM*	0.82	0.4
Original RF vs *Up GBM*	0.14	0.88
Original RF vs *Rose GBM*	4.48	0.00
Original RF vs *Original Elastic*	1.51	0.129
Original RF vs *Weighted Elastic*	1.76	0.07
Original RF vs *Up Elastic*	1.54	0.12
Original RF vs *Rose Elastic*	2.42	0.01
Original RF vs *Original Neural*	12.77	0.000
Original RF vs *Weighted Neural*	6.01	0.000
Original RF vs *Up Neural*	15.86	0.000
Original RF vs *Rose Neural*	15.50	0.000

Tripicchio Giovanni, Manoccio Mariachiara

Features importance del modello Random Forest Original

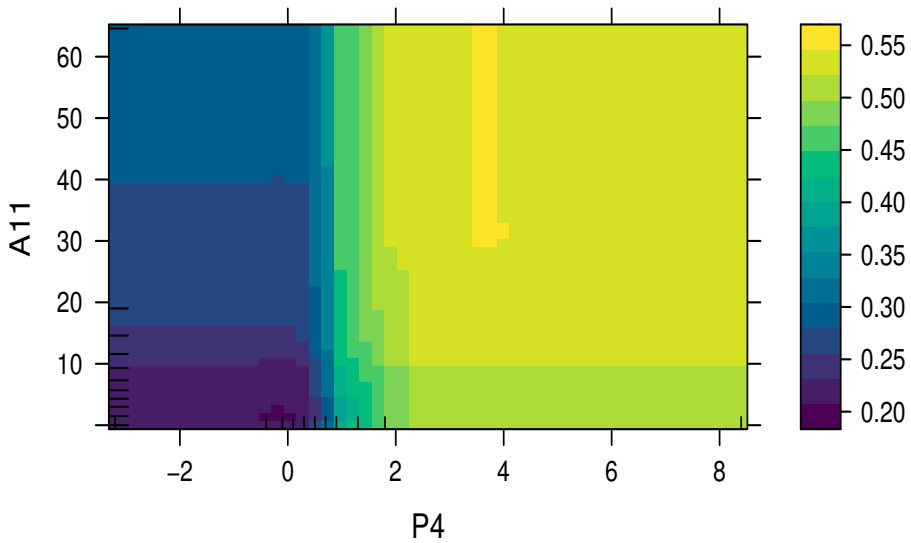
Il seguente grafico riporta i valori delle features che hanno un impatto maggiore nella previsione dello spopolamento.



- **Dinamica demografica e struttura della popolazione:** P4 (Variazione intercensuaria popolazione con 15 anni ed oltre), P5 (Incidenza superficie centri e nuclei), P10(Incidenza popolazione residente di 75 anni e più), P7 (Densità demografica)
- **Condizioni abitative ed insediamenti:** A11 (Indice di espansione edilizia nei centri e nuclei abitati);
- **Attività della popolazione:** L3 (Partecipazione al mercato del lavoro);
- **Famiglie:** F4 (Incidenza di giovani che vivono da soli);

Relazioni tra variabili

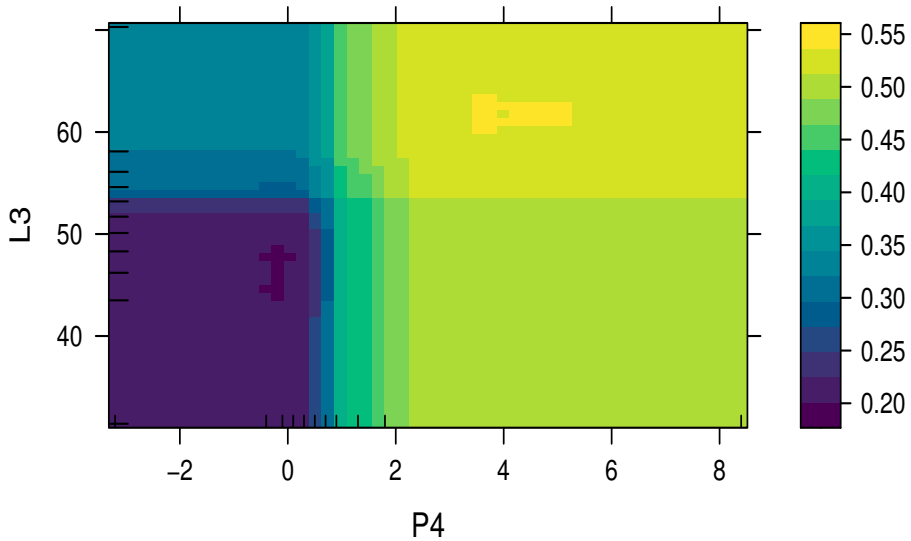
Per esaminare le relazioni tra le precedente feature importance, sono realizzati dei partial plot. In particolare, è confrontata la variabile P4 (Variazione intercensuaria popolazione con 15 anni ed oltre) la cui feature importance risultante è molto alta, con le altre.



La probabilità di spopolamento potrebbe aumentare con l'espansione edilizia nei comuni dove la variazione intercensuaria dei giovani con 15 anni ed oltre è ad oggi positiva.

Relazioni tra variabili

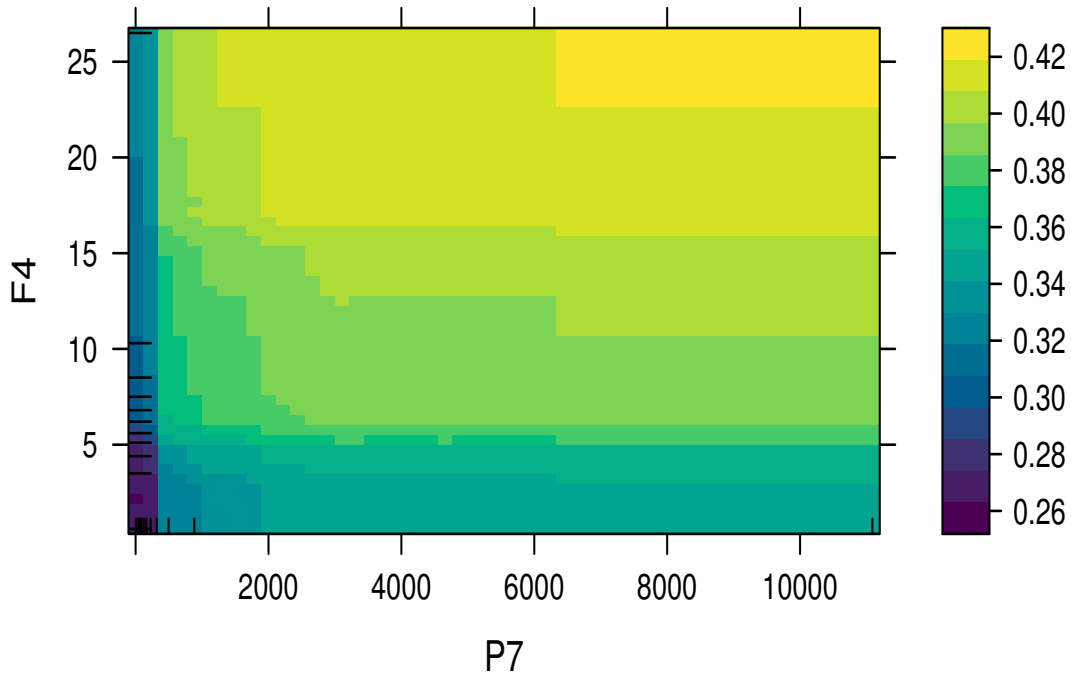
Partial plot della variazione intercensuaria popolazione con 15 anni ed oltre (P4) vs la partecipazione al mercato del lavoro (L3)



I comuni che attualmente stanno subendo un calo della variazione intercensuaria dei giovani con 15 anni ed oltre potrebbero avere minore probabilità di spopolamento per causa della partecipazione al mercato del lavoro.

Relazioni tra variabili

Partial plot della densità demografica (P7) vs l'incidenza di giovani che vivono da soli (F4)

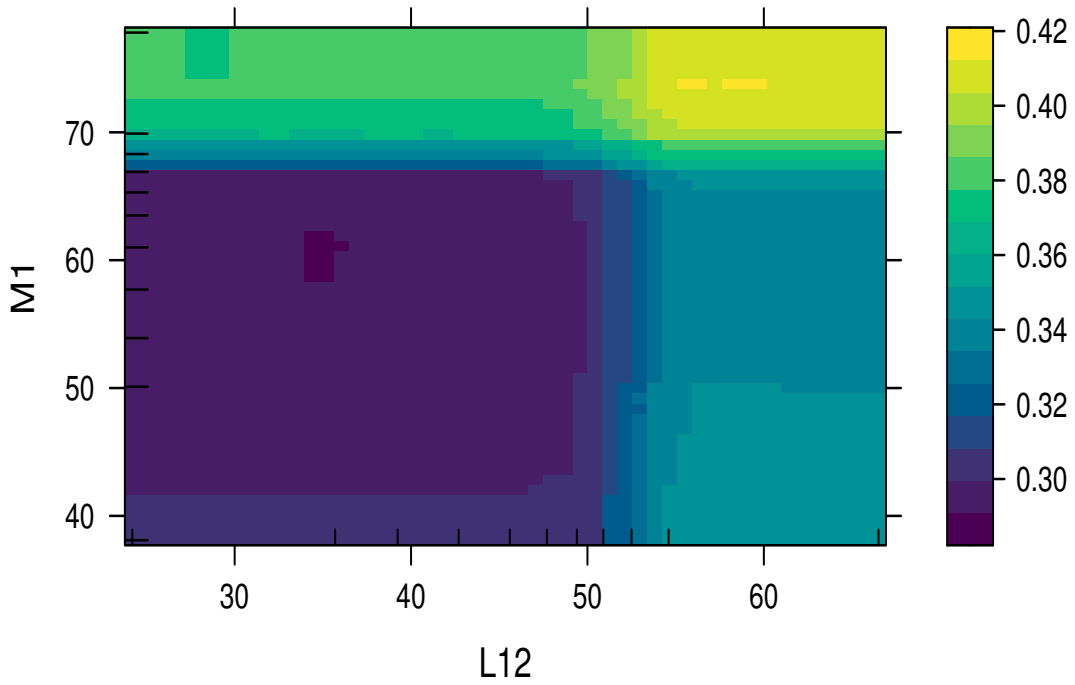


Comuni con una densità demografica più alta nei quali c'è un'elevata presenza di giovani che vivono da soli hanno un'alta probabilità di spopolarsi.

Tripicchio Giovanni, Manoccio Mariachiara

Relazioni tra variabili

Partial plot del tasso di occupazione (L12) vs la mobilità giornaliera per studio o lavoro (M1).



Comuni con un tasso di occupazione più alto e nei quali la mobilità giornaliera per studio o lavoro è maggiore hanno una più alta probabilità di spopolarsi.

Tripicchio Giovanni, Manoccio Mariachiara

Logit

I risultati ottenuti attraverso il modello RF sono stati confrontati con i risultati ottenuti attraverso un approccio statistico come la regressione logistica, comunemente impiegata nei modelli di classificazione. Nonostante le prestazioni dei due modelli siano simili, si osserva che le performance del modello RF rimangono superiori.

	Accuracy	NoInfoRate	Kappa	McnemarPValue	Sensitivity	Specificity	AUC-ROC	AUC-PRC
Logistic	0.79	0.695	0.548	0.000	0.79	0.79	0.87	0.93
Original RF	0.83	0.695	0.587	0.000	0.90	0.659	0.88	0.94

Sia l'accuratezza che l'AUC-ROC e PRC, risultano essere inferiori al modello Random Forest, ragion per cui un approccio di tipo machine learning risulta avere performance migliori rispetto ad un approccio di tipo statistical learning.

Conclusioni

- Sono stati valutati quattro modelli ML (Random Forest, Elastic Search, Neural Network e Gradient Boosting Machine), utilizzando diversi metodi di sampling, per prevedere le cause dello spopolamento per i piccoli comuni italiani.
- Dai risultati delle metriche ROC e AUC, è emerso che Random Forest ha fornito le migliori previsioni, con un'accuratezza dell'88% e AUC superiore al 90%.
- L'analisi della feature importance e i partial plot hanno rivelato che la variazione intercensuaria della popolazione con 15 anni e oltre ha un impatto significativo sullo spopolamento, suggerendo che le maggiori cause del declino demografico potrebbero essere legate ai giovani adulti soli, alle opportunità lavorative, alla vicinanza dalle scuole, università o luoghi di lavoro e allo sviluppo edilizio.

Questi risultati suggeriscono la necessità di sviluppare strategie socio economiche alternative e politiche mirate per contrastare lo spopolamento e promuovere lo sviluppo sostenibile.