Machine Learning Report

Studio della ricchezza negli USA: Una ricerca predittiva della popolazione benestante

TEAM 50

Marco Branciforti 796670 M.Sc Data Science

Emanuele Marnati 812503 M.Sc Data Science

Valerio Schips 872954 M.Sc Data Science

ABSTRACT

Secondo le statistiche della World Bank in America il reddito medio degli individui (nel 1994) corrispondeva a 27.694,85 dollari, meno della metà del reddito medio del 2019. Eppure molti economisti sostengono oggi che la distribuzione della ricchezza negli anni Novanta fosse più equa e ci fosse una popolazione più eterogeneamente benestante. Secondo www.laleggepertutti.it il reddito che in quegli anni permetteva di essere considerato benestante in America ammontava a circa 50.000,00 dollari, contro I 60.000,00 dollari di oggi. La crescita è meno che proporzionale rispetto al reddito medio in quanto gli stipendi sono cresciuti più del costo medio della vita. Alla luce di quanto riportato ci si vuole porre il seguente interrogativo: è possibile prevedere, analizzando i dati dell'ufficio anagrafe americana, il reddito annuale di un cittadino americano al fine di poter comprendere quali siano gli americani considerabili benestanti? Per rispondere a questa interessante tematica si è deciso di analizzare dati del 1994 e cercare di individuare il modello predittivo che meglio riesca a riconoscere le condizioni economiche di un individuo date alcune sue caratteristiche. I risultati ottenuti potranno essere di interesse ai fini della allocazione di supporti economici per le classi non benestanti. Dopo alcune analisi preliminari, tramite l'utilizzo di un Workflow Knime è stato eseguito un processo in grado di sviluppare un modello predittivo il più possibile performante.

Machine Learning Report 1
Dataset 1
Preprocessing 2
- Missing Imputation 2
- Rimozione attributi e record poco rilevanti 2
- Raggruppamento variabili in classi 2

Modelli e misure 3
- Modelli Adoperati 3
- Misure di Valutazione Adoperate 4

Classificazione 4
- Sbilanciamento del Dataset 4
- Metodo dell'Hold Out e Features Selection 5
- Cross Validation 6
- Confronto Modelli e Intervalli di Confidenza 7
- Cost Sensitive Approach 7

Conclusioni 8
Bibliografia 8

Dataset

La base dati citata nell'abstract è interpretabile come un tradizionale censimento della popolazione che raccoglie variabili socio-economiche con metodi campionari nei 51 stati americani. Ai fini della nostra analisi si è deciso di predire il valore della variabile *income*, attributo dicotomico che, per via della sua distribuzione, è considerato positivo se assume valore >50k e negativo se assume valore <=50k.

Variabile	Descrizione	Tipologia
Age	Eta` del Cittadino in anni compiuti al censimento	Interval
Workclass	Classe lavorativa di appartenenza	Nominal
Education	Livello di scolarizzazione	Ordinal

Variabile	Descrizione	Tipologia
Marital.status	Stato civile	Nominal
Occupation	Occupazione	Nominal
Education.num	Numero di anni di scolarizzazione	Interval
Fnlwgt	Grado di rappresentanza di una classe sociale da parte di un cittadino	Ratio
Relationship	Ruolo nel nucleo familiare se presente	Nominal
Race	Razza	Nominal
Sex	Sesso	Nominal
Capital.gain	Guadagno in conto capitale	Ratio
Capital.loss	Perdita in conto capitale	Ratio
Hours.per.week	Ore settimanali salariate	Interval
Native.country	Nazione natia	Nominal
Income	Stipendio annuale	Nominal

Tabella 1

Preprocessing

Missing Imputation

La prima importante valutazione che viene compiuta è attinente ai missing values la cui locazione ci ha permesso di effettuare una scelta ben misurata. Dei 4262 valori mancanti nell'intero dataset 1836 appartengono al campo *Workclass*, 1843 al campo *Occupation* e infine 583 occorrono in *Native.country*. Analizzando più nel dettaglio i singoli records che presentano valori mancanti, si è riscontrata un'assenza ricorsiva nell'attributo *Workclass* esattamente nelle stesse istanze in cui il valore *Occupation* risulta assente.

Per approfondire questa ricorrenza, si è notato che la distribuzione dell'attributo *age* all'interno delle sole osservazioni che presentano una co-assenza in *Workclass* e *Occupation* (figura 1) ha permesso di affermare che c'è un elevato numero di queste proprio nelle fasce d'età inferiori ai 25 anni e superiori ai 60.

Dopo un'attenta valutazione, si è considerato che inferire a priori se i giovani di età minore ai 25 anni siano o meno lavoratori, possa essere un errore.

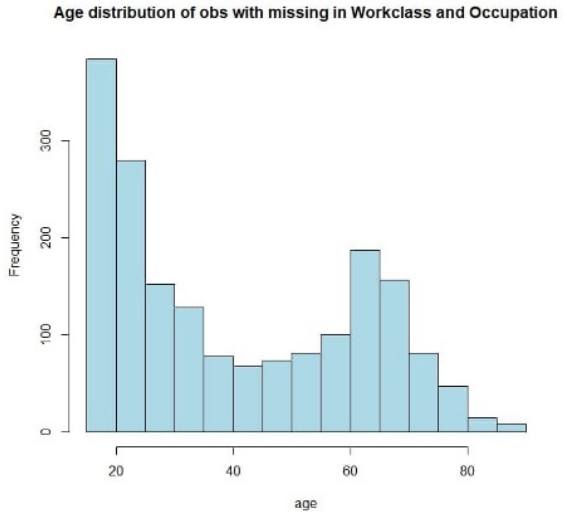

Figura 1

Inoltre, ipotizzando che difficilmente superino i 50 mila dollari di stipendio, si è optato per la rimozione dei records corrispondenti.

Per quanto riguarda gli anziani, si è supposto che una eliminazione mediante lo stesso criterio potesse essere lesivo del potere esplicativo dei modelli. Per imputare i valori mancanti sono stati addestrati due modelli Naive-Bayes utilizzando come variabili target prima l'attributo *Occupation*, successivamente *Workclass* e tutte le restanti come variabili di input.

Rimozione attributi e record poco rilevanti

Dai 15 campi di partenza viene effettuata una selezione di variabili da adoperare nell'analisi per la quale *Fnlwgt* e *education.num* sono eliminate: la prima viene scartata in quanto il processo di calcolo è poco chiaro e probabilmente frutto di combinazioni delle variabili già presenti nel data set, la seconda è una variabile perlopiù equivalente a *education*, la quale invece viene mantenuta.

Volendo adoperare la logica più appropriata possibile, è stato convenuto che tenere conto delle istanze di *Workclass* come "Never-worked" e "Without_pay" sia poco utile ai fini dell'analisi, perché poco attinenti alla variabile *income*.

Raggruppamento variabili in classi

La prima aggregazione è stata effettuata per la variabile *native country*, nella quale il 90% delle osservazioni ha come modalità "United-States". Per non procedere all'esclusione di tutta la variabile, si è pensato di aggregare tutti gli stati non americani nella classe "not USA", in modo da binarizzare

l'attributo e tentare di ribilanciare il più possibile le frequenze delle classi.

Un ulteriore raggruppamento è stato effettuato per *education*, ottenendo dalle 16 classi iniziali solo 4 gruppi: Pre & Elementary school, Middle & High school, Major school & College e Further Specialization.

Infine, per la variabile *Occupation* è stato più complesso effettuare un'aggregazione logica e quindi ai fini di una suddivisione il più adeguata possibile, è stata effettuata una valutazione basata sul fatto che il lavoro degli individui analizzati fosse più o meno intellettuale. Questa ratio è stata adottata considerando l'impatto che può avere l'uso dell'intelletto nella propria mansione lavorativa in corrispondenza di un reddito più o meno elevato. Distribuendo ordinatamente i valori di classe in tal modo è stato possibile discriminare 3 classi dalle 15 di partenza (vedi Tabella 2).

Gruppi	Classi
highly-intellectual-occupation	Exec-managerial
	Prof-speciality
specialized-occupation	Tech-support
	Sales
	Adm-clerical
	Craft-repair
	Protective-serv
	Transport-moving
low-intellectual-occupation	Armed-forces
	Machine-op-inspct
	Farming-fishing
	Other-service
	Priv-house-serv
	Handlers-cleaners

Tabella 2

Dovendo etichettare alcune professioni come quelle clericali, le quali racchiudono figure che hanno una "expertise" certamente avanzata, l'approccio è stato volutamente conservativo, dividendo solo in tre classi.

Modelli e misure

Modelli Adoperati

Sono stati addestrati 7 modelli differenti: J48, Random Forest, Logistic Regression, Multilayer Perceptron, SPegasos, Naïve Bayes e Naïve Bayes Tree. La scelta è stata dettata dalla volontà e dalla curiosità di applicare modelli di tipo diverso passando dai classici metodi euristici e di regressione, fino ai modelli di separazione e probabilistici. Di seguito una breve spiegazione dei vari classificatori implementati:

- **J48:** è l'implementazione offerta da Weka di un decision tree, nel caso in questione il C4.5. L'albero decisionale segmenta tramite la misura di entropia i soggetti in maniera mutualmente esclusiva in base al target, in modo da avere nodi finali più puri possibili. Un nodo che presenta modalità ugualmente frequenti non è puro. Purtroppo, questo modello è molto sensibile alle variazioni seppur contenute degli attributi di input, però è molto facile da interpretare ed è in grado di ricevere in input sia dati numerici che categoriali senza particolari trasformazioni. Tra i suoi settaggi c'è la possibilità di decidere il tipo di splitting, in numero oggetti per ogni nodo e l'eventualità di una tecnica di pruning, ovvero la possibilità di poter arrestare ad una certa profondità il processo (meccanismo conosciuto come potatura). Il numero minimo di elementi per ogni foglia è stato lasciato a 2, invece la funzione di split è settata su multipla per una maggiore separazione in ciascun nodo.
- **Random Forest:** questo secondo modello euristico implementa più alberi di decisione assieme selezionando casualmente per ognuno K variabili esplicative. È stato lasciato in prima istanza invariato il setting di default, dunque il modello ha adoperato 10 alberi decisionali complessivamente.
- **Logistic Regression:** modello di regressione, è adatto solamente al caso in cui la variabile output sia dicotomica e implementa una particolare funzione di collegamento lineare chiamata "logit". L'obiettivo del modello è stabilire la probabilità con cui un'osservazione può generare uno o l'altro valore della variabile target.
- **Neural Network:** il tipo particolare di rete neurale implementata è il percettrone a multilivello (MLP), appartenente alla famiglia dei classificatori di separazione. È un modello privo di funzionalità di back-propagation, cioè il flusso ponderato procede solamente in una direzione (forward) ed è composto da una struttura tripartita. Variabili di input, uno o più strati di neuroni nascosti e uno strato finale di neuroni output. In questa architettura solamente lo strato composto dai neuroni nascosti è oggetto di decisione: all'aumentare di layers e neuroni aumenterà la complessità del modello che modificherà l'output. Alla base della connessione è necessario che ogni neurone comunichi con tutti i neuroni dello strato precedente e successivo (struttura fully-connected) ma che la connessione sia solo tra strati contigui e non nel medesimo. Tra gli svantaggi, la complessità del modello stesso e la sua difficile interpretazione. La configurazione utilizzata prevede un numero di strati nascosi pari a

3 i quali rispettivamente contengono 2, 3 e 5 neuroni.

- **SPegasos:** variante stocastica del metodo Pegasos tradizionale, implementa una Support Vector Machine che adotta una funzione kernel, cioè un metodo supervisionato per la regressione e il riconoscimento di pattern. Riesce a sostituire globalmente tutti i valori mancanti e trasforma gli attributi nominali in binari. Inoltre, è in grado di normalizzare tutti gli attributi, quindi i coefficienti nell'output sono basati sui dati normalizzati. Per via della struttura dei dati, il tipo di separazione non sarà lineare e in quest'ottica le SVM sono molto efficienti. Il setting è rimasto quello di default.
- **Naïve Bayes:** primo dei due modelli probabilistici, parte dalla forte ipotesi che gli attributi siano indipendenti tra loro. Implementa un modello Bayesiano semplice (fondato sul teorema della probabilità di Bayes) il quale è un metodo che si adatta a tutti i tipi di dati e che necessita solo la conoscenza delle probabilità condizionate a priori (stimabili). La configurazione è lasciata di default.
- **NBTree:** secondo modello probabilistico, è anch'esso fondato sul teorema di Bayes, tuttavia implementa questa probabilità ad una struttura ad albero decisionale. L'albero Naïve Bayes utilizza l'albero decisionale come struttura generale e distribuisce classificatori bayesiani sulle foglie. L'idea alla base è che i classificatori bayesiani funzionino meglio degli alberi decisionali quando il set di dati del campione è piccolo. Pertanto, dopo diversi split, è meglio usare questo classificatore piuttosto che continuare a suddividere gli attributi.

Misure di Valutazione Adoperate

Come anticipato in precedenza, si ha a che fare con un dataset sbilanciato dettato dall'eccessivo divario tra le frequenze di una modalità della variabile target rispetto all'altra. Per questo motivo l'accuratezza non è sufficiente di per sé per giudicare l'adattamento dei classificatori e valutarne l'errore, perciò si sono tenuti in considerazione altri indicatori come Precision, F1 Measure, AUC (Area Under Curve) e più in particolare, si è prestata maggior attenzione all'indice di Recall. Nello specifico gli errori che si possono commettere sono due: classificare i cittadini con reddito superiore ai 50.000 come se fosse inferiore e viceversa. L'obiettivo principale è quello di ottenere modelli nei quali il falso negativo abbia maggiore rilievo rispetto al falso positivo, nell'ottica di un'allocazione di budget volta a non dilapidare denaro pubblico a chi è classificabile benestante. Di seguito una concisa spiegazione delle caratteristiche delle varie misure adoperate:

- **Precision:** L'indice di recall misura la frazione di record positivi correttamente predetti dal modello. Recall prossima all'unità è indice di come il modello abbia un errore sulla classe positiva verosimilmente basso e viceversa. Al contrario della precision, questo indice tiene conto dei falsi negativi.

$$precision, p = \frac{TP}{TP + FP}$$

- **Recall:** La Precision determina la frazione di record che appartiene realmente alla classe positiva nel gruppo tra tutti i casi definiti come tali. Il valore di precision prossimo all'uno, indica come nel modello ci siano pochi falsi positivi e viceversa.

$$recall, r = \frac{TP}{TP + FN}$$

- **Accuracy:** L'accuracy, principale indice di performace, valuta la bontà del discernimento del modello non solo rispetto alla classe rara ma considerando anche la classe abbondante. A differenza degli altri indicatori descritti, al denominatore presenta la somma di tutte le possibili combinazioni, pari quindi al numero totale di records.

$$Accuracy = \frac{TN + TP}{TN + FN + FP + TP}$$

- **F1-measure:** La precision e la recall possono essere tuttavia valori che, specie se letti insieme, non danno particolare significato per via di un conflitto logico: quanto si è disposti a migliorare la recall se questo implica un aumento consequenziale di falsi positivi? È a tale scopo che s'introduce la F1 che è una media armonica dei due indici ed è quindi manifestazione della bontà generale.

$$F_1 = \frac{2rp}{r + p}$$

- **Area Under Curve (AUC):** Dalla rappresentazione bidimensionale tra sensitivity e specificity i quali in letteratura sono indicati come True Positive Rate e False Positive Rate (TPR, FPR) si ottiene la curva Receiver Operating Characteristic (più comunemente conosciuta come ROC), la cui area sottesa definisce il valore della Area Under Curve (nota come AUC). Questo indice, ugualmente ai precedenti, acquisisce bontà quanto più si avvicina all'unità.

Classificazione

Sbilanciamento del Dataset

Già in una prima fase di data exploration è chiaro come la variabile da prevedere fosse sbilanciata: la classe rara, che per convenzione si definisce positiva, rappresenta circa un quarto del dataset (24%) contrariamente alla necessità di avere un dataset quanto più bilanciato per una massima efficacia d'apprendimento. Si sono dunque adoperate alcune tecniche di campionamento tra cui l'oversampling, l'undersampling e successivamente è stato fatto un ulteriore tentativo con l'approccio cost-sensitive. La

prima tecnica permette, tramite un campionamento casuale con reinserimento dalla classe meno rappresentata, di pareggiare la numerosità della classe più frequente. La seconda intuizione è speculare alla prima: ridurre l'impatto dell'abbondanza della classe maggioritaria campionando casualmente un numero di istanze pari a quelle della classe rara ed eliminando le restanti. Quest'ultimo metodo porta comunque alla possibilità concreta di perdita di informazione, mentre la presenza di record identici per via del campionamento con reinserimento potrebbe portare ad un rischio di overfitting. Per questo motivo è stata introdotta una terza modalità che provi a simulare l'approccio che il governo americano adotterebbe per valutare il problema e tradurlo in termini di matrice di costo.

Metodo dell'Hold Out e Features Selection

Il primo metodo implementato è stato l'Holdout: questo metodo è basato sulla sezione del dataset in due partizioni disgiunte (record esaustivi ed esclusivi), solitamente utilizzando la consueta suddivisione 2/3 + 1/3. È stato utilizzato il campionamento casuale stratificato per la variabile target per assegnare ciascuna istanza alle due partizioni Training Set, la partizione maggiore, e Test Set. Successivamente ai dati di Training è stato applicato il metodo dell'under sampling. Con i dati ottenuti si è proceduto alla fase di Feature Selection: tramite la tecnica Cfs Subset, sono state selezionati gli attributi da sottoporre all'apprendimento dei clasificatori. Questo metodo valuta il valore di un sottoinsieme di attributi considerando la capacità predittiva individuale di ogni caratteristica insieme al grado di ridondanza tra di esse. Il modello scelto per la selezione è una Random Forest con 100 alberi, la quale ha individuato 8 attributi fondamentali alla classificazione della variabile target. Le features selezionate sono:

- *Age*
- *Education*
- *Marital.Status*
- *Relationship*
- *Hours.per.week*
- *Occupation*
- *Capital.loss*
- *Capital.gain*

Le variabili *Workclass*, *Race*, *Native.Country* e *Sex* sono state escluse dal classificatore e quindi scartate.

Dopo aver addestrato il modello sul Training Set si è proceduto con la validazione sul Test Set ottenendo così la Tabella 3, in grado di confrontare tutte le misure precedentemente menzionate.

MODELLO	RECALL	PRECISION	ACCURACY	F1-MEASURE	AUC
J48	0,825	0,578	0,809	0,68	0,892
Random Forest	0,815	0,522	0,77	0,636	0,868
Logistic	0,823	0,554	0,793	0,662	0,889
Neural Network	0,852	0,527	0,775	0,651	0,886
Spegasos	0,821	0,527	0,774	0,642	0,79
Naïve Bayes	0,474	0,678	0,815	0,558	0,878
NBTree	0,851	0,575	0,808	0,686	0,908

Tabella 3

In generale tutti i classificatori hanno buone misure di performance, in particolare per quanto riguarda l'Accuracy e AUC. Inoltre è possibile notare come in quasi tutti la Recall tenda ad essere più alta della Precision anche in maniera consistente, ciò mostra quanto i modelli abbiano una tendenza generale a classificare records come positivi, cioè appartenenti alla classe rara. Tra tutti spiccano Neural Network e NBTree, con alte misure di Accuracy e Recall, mentre Naïve Bayes si mostra particolarmente sofferenti a livello di Recall ma leggermente migliore in termini di Precision.

Per effettuare un confronto grafico tra i modelli si è scelto di utilizzare le curve ROC ottenute dalla combinazione delle percentuali di TP e FP:

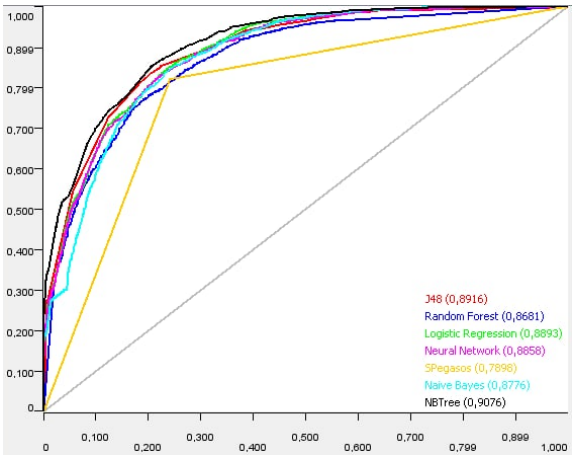

Figura 2

La figura 2 conferma una bontà generale dei modelli adottati. L'unica eccezione è SPegasos, il quale, individuata una qualsiasi soglia di records positivi correttamente classificati, presenta una percentuale di FP quasi sempre superiore agli altri. Quindi nel caso in questione, è preferibile un classificatore la cui

curva ROC "sovrasti" il più possibile le altre. Data la sovrapposizione, c'è la necessità di indagare più nel dettaglio le due curve che sembrano avere un andamento leggermente migliore rispetto alle altre: NBTree e Neural Network. Quindi sarà fatto un confronto con le Cumulative Gains, le quali illustrano graficamente il guadagno in termini di records positivi in percentuale, all'aumentare della porzione di dati presi in considerazione.

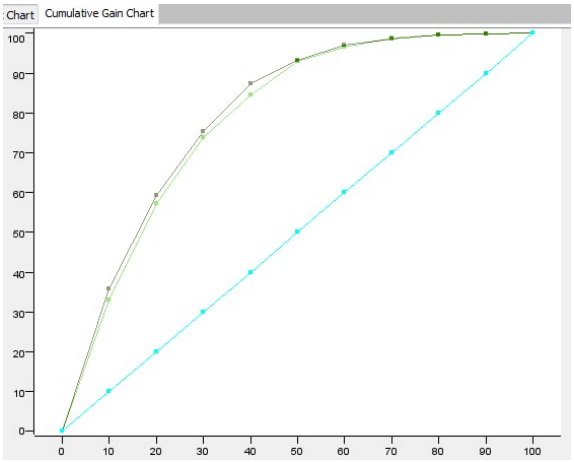

Figura 3

La figura 3 mostra come il modello NBTree (linea grigia) abbia maggiore affidabilità rispetto al Neural Network (linea verde) perché prendendo casualmente un numero di record fino al 50% dell'intero insieme dei dati, garantisce una maggiore percentuale di soggetti positivi correttamente individuati. Oltre al 50% i due modelli risultano equivalenti.

Cross Validation

L'approccio di Cross-Validation, anche chiamato K-fold, è strutturato su un partizionamento del dataset in K sottoinsiemi: è eseguita un'iterazione ripetuta per K volte, in ciascuna delle quali si utilizza come Test Set una delle partizioni create e come Training Set le rimanenti. Prima di approcciare tale tecnica è stato eseguito un processo di normalizzazione degli input numerici e di binarizzazione che ha interessato le variabili *education*, *marital.status*, *relationship* ed *occupation*, con l'obiettivo di valutare delle performance utilizzando attributi trasformati. Per ogni partizione ricavata è stata eseguita la tecnica di Over sampling che, aumentando di molto il numero di istanze iniziali, influisce sulla potenza computazionale. Da qui la scelta di mantenere un numero esiguo di fold per ogni modello addestrato, impostando K=3.

MODELLO	RECALL	PRECISION	ACCURACY	F1-MEASURE	AUC
J48	0,827	0,582	0,811	0,683	0,872
Random Forest	0,737	0,587	0,807	0,653	0,862
Logistic	0,835	0,561	0,798	0,671	0,895
Neural Network	0,744	0,586	0,807	0,655	0,885
Spegasos	1	0,262	0,306	0,415	0,539
Naïve Bayes	0,705	0,636	0,828	0,669	0,885
NBTree	0,892	0,53	0,778	0,665	0,906

Tabella 4

La Tabella 4 riconferma delle buone performance generali di quasi tutti i modelli. NBTree appare ancora tra i migliori in termini di Recall e AUC, seguito dal modello logistico e dal decision tree. Il modello SPegasos mostra come l'approccio complesso non sempre sia la chiave di lettura giusta: il valore di recall del metodo che implementa le SVM è massimo, ciò vuol dire che riesce a classificare correttamente quasi tutte le osservazioni di classe positiva; un valore molto basso di Precision indica tuttavia che sbaglia nella classificazione della classe negativa. Queste considerazioni, rapportate alla scadente Accuracy e ad altri indici, non permettono di reclamare la bontà del modello.

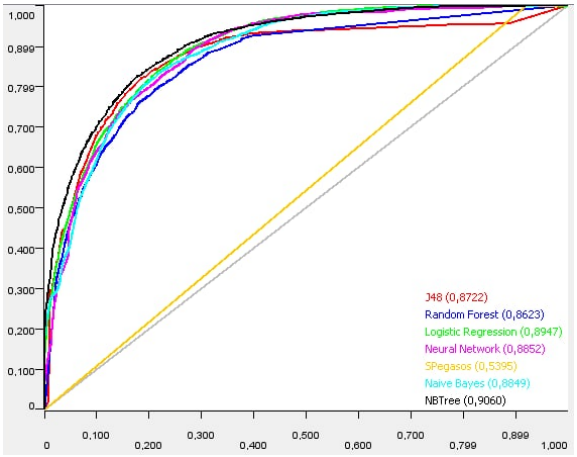

Figura 4

Le curve ROC in figura 4 confermano le buone performance del modello NBTree e mettono in evidenza la scarsità del modello SPegasos, il quale si distacca di poco dalla bisettrice che indica il modello nullo.

Selezionando 2 tra i modelli con Recall più elevata, si può notare dalle Cumulative Gain riportate in figura 5 che anche in questo caso NBTree (rosso) è sempre

preferibile al J48 (blu), soprattutto per un campionamento superiore al 50% dei dati a disposizione.

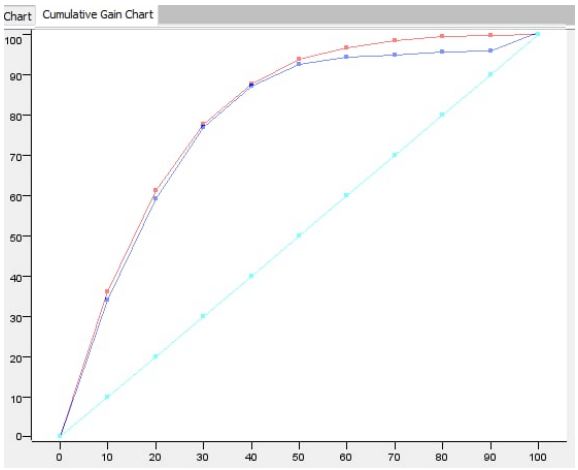

Figura 5

Confronto Modelli e Intervalli di Confidenza

Per verificare se la differenza del valore di recall sia dovuta alla variazione della composizione del test set e quindi la misura di performance non sia adattiva si ricorre agli intervalli di confidenza di Wilson. Ai fini di implementare questa analisi ci si concentrerà solo su una misura: la recall, che è di nostro interesse. Il metodo adottato è stato quello delle doppie partizioni. La Partizione A è composta dall'80% delle osservazioni e la Partizione B invece dalle rimanenti. La Partizione A viene poi suddivisa nuovamente nel rapporto 2/3 - 1/3 come nel metodo Holdout. Quindi nel Training Set della Partizione A è stato applicato una tecnica di over sampling per bilanciare le classi della variabile target e addestrati i modelli che verranno validati sia sulla Test Set della Partizione A sia sulla Partizione B.

MODELLO	RECALL	RECALL_B
J48	0,857	0,859
Random Forest	0,716	0,727
Logistic	0,866	0,852
Neural Network	0,726	0,73
Naïve Bayes	0,502	0,512
NBTree	0,862	0,851

Tabella 5

In tabella 5 si evidenzia che NBTree, J48 e Logistic Regression hanno un ottimo valore di Recall, Neural Network e Random Forest sono discretamente posizionati e Naive Bayes si colloca in ultima posizione. Interessante notare come i valori siano tutti simili tra la Recall calcolata nel Test Set della Partizione A e quella calcolata in B. Il fatto che le validazioni del modello sulla Partizione A e su B non si discostino poi tanto in valore indica che i modelli siano abbastanza robusti alla variazione di dati.

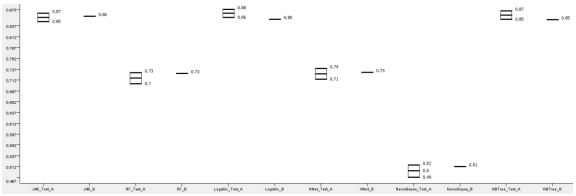

Figura 6

In figura 6 è rappresentato l'intervallo di confidenza sulla Recall di Wilson sulla Partizione A e il corrispettivo valore della misura sulla Partizione B. Questa grandezza assume significato nell'ottica di ripetere la validazione un buon numero di volte (ad esempio N= 100, per ottenere le proprietà della curva gaussiana) su test set diversi. il livello di confidenza è stato posto al 99%. Non è stato considerato il modello SPegasos: nonostante abbia ottimi valori di recall ha una pessima accuracy. Tutti i modelli presi in analisi, eccetto la logistic regression, possono essere annoverati come robusti al cambiamento del test set secondo gli assunti dell'intervallo di confidenza: il valore puntuale di recall sulla Partizione B cade sempre nell'intervallo di recall sulla Partizione A. Nel caso della logistic regression, tuttavia, il valore di recall sulla Partizione B non è compreso nell'intervallo di confidenza, questo modello viene dunque scartato.

Cost Sensitive Approach

Si parte dall'assunto che, basandosi sugli esiti della classificazione, il governo intervenga per stanziare fondi e in quale misura a difesa di chi percepisce un reddito inferiore a 50.000 dollari o per lo meno si adoperi per valutare un budget preventivo o la sua allocazione. La supposizione, elaborata seguendo un approccio economico sociale al problema è la seguente: il governo americano come la società stessa, sono impregnati di una ideologia liberista la quale può tradursi in una politica sociale dove la spesa pubblica passa in secondo piano, rispetto ad una spesa come quella militare o strategica. Quindi la matrice dei costi rispecchia una mentalità restia alla spesa pubblica per le classi sociali meno abbienti. Questa tecnica, più in generale, permette di associare un costo di classificazione specifico ad ogni cella della confusion matrix. Questo approccio descritto inizialmente a livello teorico e logico ora si tramuta nell'individuazione una matrice di costo A, qui presentata.

		IP	
		-1	1
AC	-1	-1	3
	1	9	0

Tabella 6

L'approccio è stato il seguente: il costo è stato valutato superiore per i falsi negativi, rispetto ai falsi positivi e aggiunto un costo negativo (cioè un profitto) per i veri negativi. Si assume che sia grave classificare la classe rara come abbondante e sperperare il poco denaro per il Welfare e che addirittura ci sia un guadagno per lo stato se chi percepisce un reddito superiore a 50.000 dollari sia individuato correttamente.

MODELLO	RECALL	PRECISION	ACCURACY	F1-MEASURE	AUC
J48	0,703	0,681	0,845	692	0,797
Random Forest	0,731	0,599	0,813	0,659	0,786
Logistic	0,78	0,607	0,821	0,683	0,807
Neural Network	0,791	0,59	0,813	0,676	0,805
Spegasos	0,505	0,735	0,833	0,599	0,723
Naïve Bayes	0,507	0,687	0,821	0,583	0,716
NBTree	0,761	0,656	0,843	0,705	0,815

Tabella 7

Questa tecnica sensibile al costo mostra valori di Accuracy ottenuti nei risultati in tabella 7 abbastanza elevati. Indagando un'altra metrica di interesse, si evince che la Recall sia particolarmente alta nella rete neurale e nella regressione logistica. Si notano valori di AUC leggermente più bassi rispetto alle misure di performance ricavate nelle analisi precedenti.

Una generale tendenza individuabile in questo modello è tuttavia che l'uso delle matrici di costo aumenta le performance di alcuni indicatori rispetto ad altre. Nel caso specifico è stato volutamente assegnato un maggior costo ai FN per cercare un miglioramento in termini di Recall.

Conclusioni

Durante l'analisi si sono seguiti più approcci e più tecniche tra loro differenti ma si è riscontrato un pattern ricorrente per quanto concerne gli indicatori di performance.

In un primo momento tramite l'approccio Holdout sono stati individuati, come migliori, i modelli come NBTree, Neural Network e Logistic Regression. La cross-validation ha messo in luce buone performance da parte dei modelli Naïve Bayes Tree, J48 e Logistic. L'analisi degli intervalli di confidenza secondo Wilson ha mostrato i modelli più robusti per quanto riguarda la misura di Recall permettendo di escludere la logistic regression. Adoperando la matrice di costo si è notato come si riconfermino i modelli NBTree, Neural Network e Logistic.

Le possibili decisioni scaturite dal lavoro svolto sono due: la previsione per un miglior processo di budgeting federale oppure una migliore ripartizione del capitale sulla base della previsione. Esemplificando la seconda via il metodo Naïve Bayes Tree utilizzato nel cost-sensitive approach permetterebbe con un errore del 15,7% di stabilire un'allocazione di risorse per i cittadini non benestanti fatta pro-quota, partendo appunto dalla stima del classificatore. Se gli individui classificati come non abbienti sono numerosi la quota ad individuo sarà ridotta e viceversa nel caso opposto.

Un approccio migliorato consisterebbe nell'aggiungere ulteriori variabili significative, come il costo della vita per l'individuo oppure gli anni di lavoro effettivi e non solo il titolo di studio, potendo così analizzare anche l'importanza dell'esperienza o dell'anzianità di servizio ai fini del guadagno. In seconda ipotesi, si potrebbe procedere con un raggruppamento dai dati in cluster con l'ausilio di un esperto di dominio, con l'obiettivo di individuare quali siano le caratteristiche più ricorrenti della classe povera e riuscire a prendere i dovuti accorgimenti (come agevolazioni per lo studio), per poter garantire prospettive più favorevoli a tutti i cittadini.

Bibliografia

I. https://www.rdocumentation.org/packages/arules/versions/1.6-6/topics/Adult
II. https://rpubs.com/Net/IncomeLevelClassification
III. https://weka.sourceforge.io/doc.stable/weka/classifiers/functions/SPegasos.html
IV. https://www.kaggle.com/uciml/adult-census-income
V. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/