PROGETTO DI MACHINE LEARNING

ANALISI PREDITTIVA APPLICATA AI POKÉMON LEGGENDARI

Alessi Luca[1], Bressan Matteo[1], D'Agostino Danilo Sebastian[1], Severino Fabrizio[1], Zangirolami Valentina[2]

I mostri tascabili (pocket monster) anche detti Pokémon, hanno invaso il mondo videoludico nel XX secolo. Lo slogan principale era di catturarli tutti, ma il vero obiettivo di ogni allenatore era catturare i Pokémon Leggendari, i più rari ed esclusivi di tutto il gioco. Sulla base di questa idea si sono adottate diverse tecniche di Machine Learning con l'obiettivo di prevedere se un Pokémon preso dal dataset in esame è leggendario. La previsione che si ottiene dal dataset utilizzato è influenzata dalla natura della variabile target, la cui classe positiva (Pokémon leggendari) è fortemente sbilanciata. L'analisi ha dunque lo scopo di trovare una soluzione, attraverso alcune metodologie per affrontare attivamente il problema degli eventi rari.

KEYWORDS
Machine Learning
Pokémon
Imbalanced Class

INDICE

1. INTRODUZIONE

Nel 1982 Satoshi Tajiri e Ken Sugimori, fondarono una rivista di videogiochi chiamata Game Freak. Tajiri, con i disegni di Sugimori, pubblicava giornalini della lunghezza media di circa 30 pagine ad un prezzo di 300¥ ossia l'equivalente di 3 dollari statunitensi. Inizialmente, la rivista era redatta a mano e solo in seguito Tajiri, sull'onda del successo, si affidò ad una società di stampa professionale. Nel 1991 Satoshi vide per la prima volta un Game Boy, una console portatile per videogiochi, e non appena notò il cavo Game Link immaginò degli insetti passare attraverso esso. Nacque così l'idea dei Pokémon: i più famosi mostricciattoli della storia. Successivamente Tajiri chiamò la sua software house Game Freak. Dapprima il gioco nacque solo in Giappone, ma presto si diffuse in tutto il mondo, rappresentando oggi un brand multimiliardario e tra i più conosciuti del pianeta.

[1] DataScience (DS)
[2] Scienze Statistiche ed Economiche (CLAMSES)

2. DESCRIZIONE DATASET

Il dataset, prelevato da Kaggle, è composto da 721 righe e 21 variabili che descrivono le caratteristiche dei Pokémon:

• *Name (String):* Nome del Pokémon;
• *Type1 (String):* Tipo primario del Pokémon (Fuoco, Acqua, Erba, Elettrico…);
• *Type2 (String):* Tipo secondario del Pokémon (Fuoco - Terra, Acqua - Erba, …), non tutti i Pokémon hanno un doppio tipo;
• *Total (Numerico):* Totale di tutte le statistiche di base;
• *HP (Numerico):* Punti vita;
• *Attack (Numerico):* Attacco del Pokémon;
• *Defense (Numerico):* Difesa del Pokémon;
• *Sp_atk (Numerico):* Attacco speciale;
• *Sp_def (Numerico):* Difesa speciale;
• *Speed:* Velocità;
• *Generation (String):* Generazione del Pokémon;
• *IsLegendary (String):* variabile dicotomica in cui TRUE indica che un pokemon è leggendario e FALSE altrimenti;
• *Color (String):* Colore del Pokémon indicato nel Pokédex;
• *HasGender (String):* Booleano che indica se il Pokémon può essere classificato come femmina o maschio;
• *Pr_male (Decimale):* Probabilità che sia maschio;
• *Egg_group_1 (String):* Gruppo di uova del Pokémon;
• *Egg_group_2 (String):* Gruppo di uova del Pokémon 2;
• *Has_MegaEvolution (String):* Booleano che indica se il Pokémon è in grado di evolversi o meno;
• *Height_m (Decimale):* Altezza;
• *Weight_kg (Decimale):* Peso;
• *Catch_Rate (Numerico):* Tasso di cattura;
• *Body_Style (String):* Stile del corpo del Pokémon.

Molti pokemon sono costituiti da una forma alternativa (ovvero la capacità di tramutarsi rispetto alla forma originale), altri invece hanno la possibilità di potenziare la loro ultima evoluzione grazie a delle pietre. Questo ulteriore potenziamento viene definito mega- evoluzione. Le trasformazioni, appena citate, permettono ai Pokémon di cambiare il loro aspetto ma anche la loro potenza. In questo dataset sono state considerate le forme e le statistiche standard, senza considerare questi cambiamenti. Inoltre, i Pokémon leggendari, generalmente, sono caratterizzati da statistiche più elevate rispetto a quelli non leggendari.

3. ESPLORAZIONE DATI

La prima fase dell'analisi consiste nell'esplorazione dei dati, in particolar modo si analizza la distribuzione della variabile IsLegendary. Al fine di valutare le proporzioni di osservazioni appartenenti alle due classi, si considera il seguente istogramma.

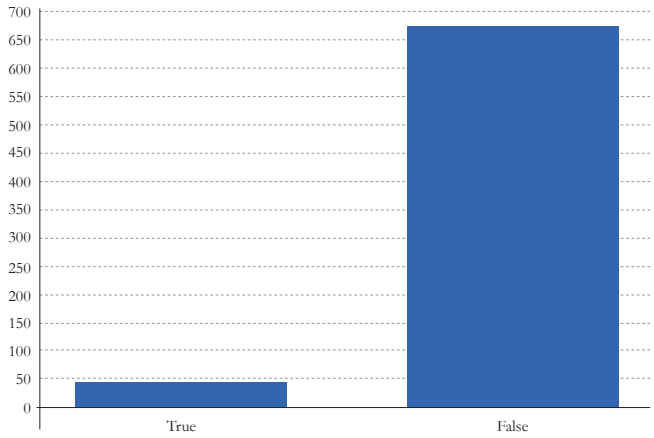

Figura 1: *Grafico classe sbilanciata*

Dal grafico, quindi, si evince che la variabile IsLegendary risulta sbilanciata. Pertanto si rende necessario utilizzare delle tecniche adeguate, prima di applicare un modello di classificazione, al fine di eliminare eventuali distorsioni sulla previsione.

4. PRE-PROCESSING

4.1 MISSING VALUE

La fase iniziale dell'analisi comprende l'applicazione di tecniche di data cleaning e data pre-processing, con lo scopo di attenuare il rumore dei dati, selezionando le informazioni necessarie per generare i modelli in una fase successiva. Le uniche tre variabili che presentano valori mancanti sono *egg_group_2, Type2 e Pr_male.*

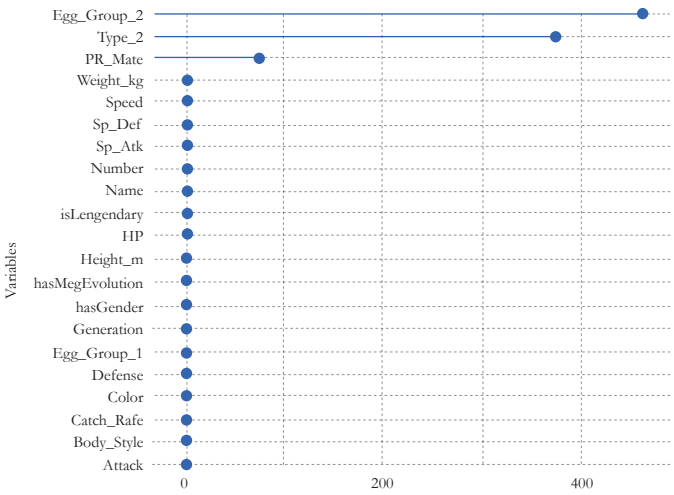

Figura 2: *Grafico Missing value*

Come evidenziato nel grafico è possibile notare che le variabili *egg_group_2* e *Type2* presentano una percentuale di valori mancanti maggiore (circa del 50%) rispetto le altre variabili, per cui quest'ultime verranno eliminano dal dataset. Mentre la variabile *Pr_male* presenta il 10% di valori mancanti, per cui vengono sostituiti dalle previsioni ottenute mediante un modello di regressione lineare, che permette di calcolare con maggiore esattezza il valore previsto, rispetto alle altre metodologie (sostituzione con media, mediana o moda) sulla base di una combinazione lineare di coefficienti stimati e regressori:

$$PrMale_i = 0.797 - 0.001 * CatchRate_i + 0.147 * IsLegendary_i - 1.22 * e + 14 * Speed_i - 1.22 * e + 14 * Spdef_i - 1.22 * e + 14 * Spatk_i - 1.22 * e + 14 * Defense_i - 1.22 * e + 14 * Attack_i - 1.22 * e + 14 * HP_i + 1.22 * e + 14 * Total_i$$

Generalmente questa tecnica è molto onerosa da un punto di vista computazionale, ma in questo caso il dataset è costituito da numero di records ridotto che permette di utilizzare questa procedura. Attraverso il test di significatività dei coefficienti, con le seguenti ipotesi:

$$H0 : \beta_i = 0$$
$$H1 : \beta_i \neq 0$$

La cui statistica test, sotto H0, è data:

$$\frac{\hat{\beta}}{SE(\hat{\beta})} \sim t_{n-2}$$

Dove $\hat{\beta}$ è la stima OLS dei coefficienti β e $SE(\hat{\beta})$ è la deviazione standard di $\hat{\beta}$. Tutti i regressori presenti nel modello hanno coefficienti significativi per spiegare la variabile risposta *Pr_male*. Infatti, il valore del *p-value* di ogni coefficiente, valutato singolarmente, è minore di ogni soglia critica e porta al rifiuto dell'ipotesi nulla.

4.2 CAMPIONAMENTO

Il primo metodo utilizzato per equilibrare la variabile sbilanciata è il campionamento stratificato. Applicando questa tecnica, nel seguente istogramma si può notare che la variabile *isLegendary* risulta bilanciata.

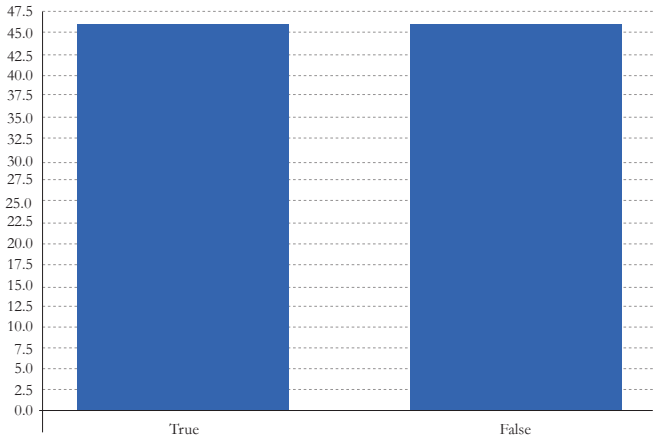

Figura 3: *Grafico classe bilanciata con equal size sampling*

Le criticità relative all'applicazione di questo metodo riguardano la riduzione della dimensionalità perché il dataset considerato contiene poche osservazioni; per cui successivamente quando viene suddiviso in traning e test set il secondo sottoinsieme è formato da poche osservazioni che inficerebbero sulle performance del modello (errori piccoli/ridotti).

4.3 FEATURE SELECTION

Al fine di migliorare l'interpretabilità dei dati e di selezionare le variabili più significative, riducendo il numero di attributi in input, è stata applicata la feature selection. Confrontando il metodo filter con il metodo wrapper, si riscontra che la prima tecnica sia più performante. Infatti il metodo wrapper includerebbe nell'analisi solamente due regressori: *Total* e *PrMale*. Inoltre per valutare i criteri filter si sono considerati il *CfsSubsetEval* e l'*InfoGainAttributeEval* attraverso nodo *AttributeSelectiveClassifier*. Il classificatore LMT, valutato con l'informazione di Gain, presenta l'errore più basso, per cui viene selezionato il seguente sottoinsieme di variabili: *Catch_Rate, Total, Egg_group_1, Pr_male, hasGender*.

5. TECNICA DI BILANCIAMENTO

Un ulteriore metodo per bilanciare la variabile *IsLegendary* è il *CostSensitiveLearning*, che permette di bilanciare il dataset applicando pesi specifici ad ogni valore della matrice di confusione (TP, TN, FP, FN) generando una matrice di costo, al fine di aumentare la rilevanza della classe rara nella valutazione delle performance finali. Con il nodo *CostSensitiveClassifier* [3]

[3] Nodo di Knime

è stata eseguita questa tecnica. La matrice dei costi considerata è la seguente:

AC / IP	-1	+1
- 1	0	1.0
+1	100	-1.0

Figura 4: *Tabella Cost Matrix*

6. MODELLI

In questo studio sono state implementate diverse tecniche di classificazione con lo scopo di individuare la più adatta, sulla base dei dati disponibili:

• Modelli di regressione tra cui regressione logistica che si basa sul concetto di odds ratio in cui la relazione tra la variabile risposta e le esplicative è valutata dalla funzione logit;

• Modelli probabilistici come il naive bayes basato sul teorema di bayes che permette di prevedere le classi attraverso una regola decisionale basata sulle probabilità a priori e a posteriori;

• Modelli euristici di cui ne fanno parte i classificatori J48 e randomforest che prevedono l'attributo di classe attraverso la costruzione di alberi di decisione.

7. PERFORMANCE

I criteri adottati per valutare le performance dei classificatori utilizzati sono: recall, f-measure, accurancy, precision e AUC (Area Under Curve). Quest'ultima è calcolata attraverso la rappresentazione della curva di ROC (Receiver Operating Characteristic). L'accuracy indica la percentuale di osservazioni previste correttamente e permette di selezionare il modello che garantisce una percentuale maggiore di stanze previste corrette:

$$Accuracy = \frac{TP+TN}{N}$$

dove TP rappresenta il numero di osservazioni con classe positiva previste correttamente; TN numero di osservazioni di classe negativa previste correttamente e N il numero di osservazioni totali.

Tuttavia la sola stima dell'accuracy non basta a identificare un buon classificatore. Soprattutto in presenza di classe sbilanciate è opportuno utilizzare altri criteri per ottenere una valutazione più completa dello classificatore. L'indicatore recall rappresenta la porzione di record positivi classificati correttamente dal modello. In particolare la sua formula è:

$$Recall = \frac{TP}{TP+FN}$$

Dove FN sono le osservazioni classificate erroneamente come classe negativa. Un alto valore della recall indica che la maggior parte delle osservazioni con classe positiva sono state previste nel modo corretto.

L'indicatore precision descrive la frazione di record che sono effettivamente positivi tra tutti quelli previsti, nel dettaglio la sua formula è:

$$Precision = \frac{TP}{FP+TP}$$

Dove FP sono le osservazioni classificate erroneamente come classe positiva. Un buon classificatore deve avere un alto valore simultaneamente delle due tecniche appena elencate. Per questa ragione si considera la f-measure ovvero la media armonica tra precision e recall, che consente di fornire un'interpretazione delle due metriche più ragionevole (un valore levato garantisce che sia la recall che la precision siano alte):

$$F_{measure} = \frac{2 * Recall * Precision}{Recall + Precision}$$

Un'altra tecnica utilizzata per valutare i metodi di classificazione, anzidetta è la curva di ROC che permette di rappresentare la percentuale di osservazioni con classe positiva e sull'asse delle ascisse la percentuale di falsi positivi. Da questo grafico si estrapola l'area al di sotto della curva (AUC), il classificatore più performante presenta un AUC più elevato.

8. ANALISI E RISULTATI

L'obiettivo principale dell'analisi effettuata consiste nel prevedere i Pokémon leggendari. A partire dal dataset iniziale si procede con la suddivisione in training e test al fine di applicare gli algoritmi di classificazione. Ulteriormente, si confrontano i modelli valutati sul dataset iniziale e quelli testati sul sottoinsieme di variabili selezionate a seguito dell'applicazione della tecnica feature selection. Di seguito vengono affrontanti nel dettaglio i diversi approcci impiegati.

8.1 TECNICHE DI PARTIZIONAMENTO

A causa della ridotta dimensionalità del dataset, invece di rispettare le divisioni ottimali prescritte dalla teoria (2/3 training set e 1/3 test set) si è proceduto associando al training set il 60% delle osservazioni e il restante 40% al test set. Tra le tecniche di campionamento considerate per procedere alla suddivisione ci sono: Draw randomly, Linear sampling, stratified sampling e take from the top. Tra queste tecniche quelle che presentano la percentuale maggiore della classe non rara è Draw randomly. Al fine di valutare le altre metodologie, si è ricercata la configurazione migliore che rispettasse la struttura inziale del dataset.

Campionamento	Minimo	Massimo	Media	Deviazione standard	Variazione
Take from Top	0.1	950	58.906	91.365	8347.493
Stratified Sampling	0.1	683	56.645	87.726	7695.763
Linear Sampling	0.1	650	56.143	83.001	6889.111
Draw Randomly	0.1	650	54.913	80.309	6449.598
Dataset	0.1	950	56.773	89.096	7938.038

Figura 5: *Performance del metodo di partizionamento*

Il metodo migliore di partizionamento risulta essere il campionamento stratificato ovvero un campiomento che permette di ottenere le stesse proporzioni di osservazione per classe rispetto al dataset originale.

8.2 MODELLI DI CLASSIFICAZIONE

Il metodo valutato per ovviare il problema del dataset sbilanciato si basa sull'analisi dei costi. Attraverso il nodo *CostSensitiveClassifier*, applicando la matrice di costo descritta nel capitolo precedente, sono stati inizializzati i classificatori. Quindi si è proceduto a suddividere il dataset in training test con campionamento stratificato il quale è risultato il più performante tra tutte le tecniche considerate. Successivamente, si sono valutati sei modelli di classificazione: naive bayes, naive bayes tree (cost), naive bayes tree, random forest, logistic e J48.
Per valutare quale modello fosse il più adatto per l'analisi in esame si sono considerate, inizialmente, le misure di performance (accuracy, precisio, recall e f-measure) relative alla previsione della classe rara, verificando che fosse equamente performante come la classe relativa ai non leggendari:

Classificatore	Recall	Precision	F-measure
NB - False	0.948	1	0.973
NB - True	1	0.562	0.720
NB Tree - False	0.996	0.996	0.996
NB Tree - True	0.944	0.944	0.944
NB Cost - False	0.959	1	0.979
NB Cost - True	1	0.621	0.766
RandomForest - False	0.948	1	0.973
RandomForest - True	1	0.562	0.720
Logistic - False	0.974	0.996	0.985
Logistic - True	0.944	0.708	0.810
J48 - False	0.978	1	0.989
J48 - True	0.1	0.75	0.857

Figura 6: *Performance per le due classi*

Gli algoritmi di classificazione che risultano meno performanti rispetto alla classe rara sono il random forest e naive bayes cost. Successivamente, si valuta l'accuratezza:

Modello	Accuracy
Naive Bayes	0.975
NB Tree	0.99
NB Cost	0.96
RandomForest	0.95
Logistic	0.97
J48 - True	0.98

Figura 7: *Performance classificatori*

In questi termini, i due modelli più perfomanti sono il naive bayes tree e il J48.
Al fine di selezionare il classificatore con il minore costo, si valuta la seguente tabella:

Modello	Cost
Naive Bayes	1657
NB Tree	369
NB Cost	1360
RandomForest	1657
Logistic	963
J48 - True	865

Figura 8: *Costi classificatori*

Il modello che garantisce costi ridotti e performance elevate è il naive bayes tree.

8.2.1 FEATURE SELECTION

Considerando il sottoinsieme di variabili selezionate attraverso la tecnica feature selection, si analizzano le misure di performance dei modelli di classificazione.
Come nel caso precedente, si verificano per entrambi le classi della variabile *IsLegendary*:

Classificatore	Recall	Precision	F-measure
NB - False	0.974	0.996	0.985
NB - True	0.944	0.708	0.810
NB Tree - False	0.989	0.996	0.993
NB Tree - True	0.944	0.850	0.895
NB Cost - False	0.945	1	0.972
NB Cost - True	1	0.545	0.706
RandomForest - False	0.970	1	0.985
RandomForest - True	1	0.692	0.818
Logistic - False	0.985	0.996	0.991
Logistic - True	0.944	0.810	0.872
J48 - False	0.978	0.993	0.985
J48 - True	0.889	0.727	0.800

Figura 9: *Performance modelli per le due classi*

I classificatori random forest e naive bayes cost, come nel caso precedente, presentano delle performance meno elevate rispetto agli altri modelli. Successivamente, si valuta l'accuratezza:

Modello	Accuracy
Naive Bayes	0.975
NB Tree	0.99
NB Cost	0.96
RandomForest	0.95
Logistic	0.97
J48 - True	0.98

Figura 10: *Performance modelli*

In questi termini, i modelli naive bayes tree e logistic risultano migliori. In secondo luogo, si è calcolato il costo sostenuto dal modello di classificazione:

Modello	Cost
Naive Bayes	1657
NB Tree	369
NB Cost	1360
RandomForest	1657
Logistic	963
J48 - True	865

Figura 11: *Costo dei classificatori con feature selection*

Il classificatore che presenta un costo ridotto e perfomance soddisfacenti è l'albero naive bayes.

8.3 CURVA ROC E CUMULATIVE GAIN

Per valutare la performance dei classificatori binari è stata affrontata l'analisi della curva ROC, effettuando un confronto tra tutti i modelli.

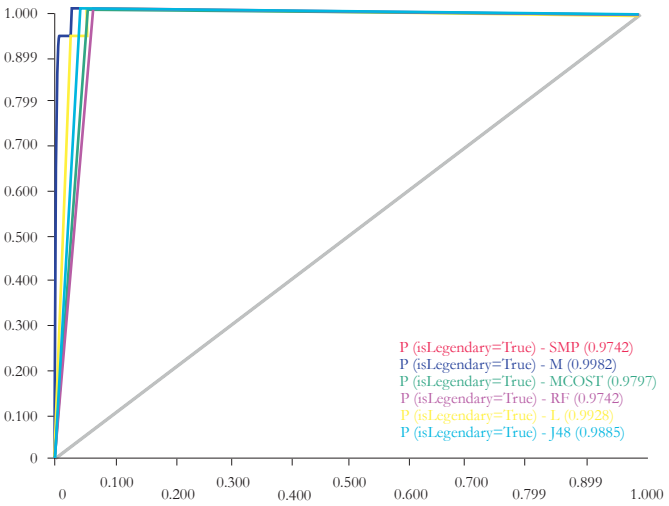

Figura 12: *ROC modelli senza feature selection*

Dal grafico emerge che il classificatore naive bayes tree presenta un valore dell'AUC molto elevato, quindi porta a confermare le ipotesi precedente sviluppate.
La seguente ROC curve è valutata nei modelli aventi come variabili il sottoinsieme selezionato con la tecnica feature selection.

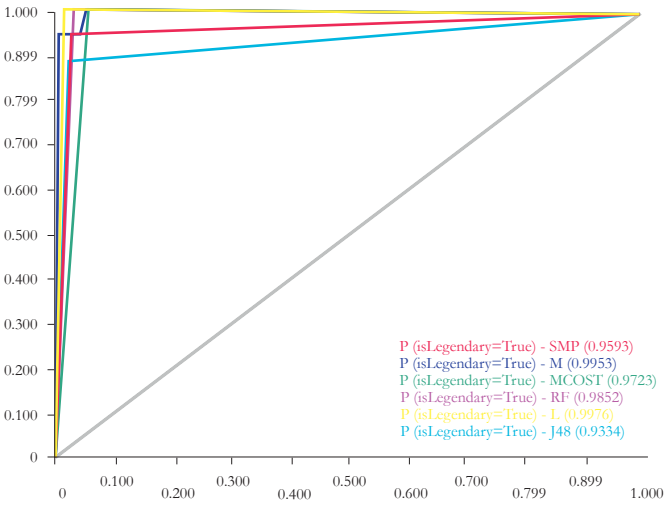

Figura 13: *ROC modelli con feature selection*

Dalla figura il classificatore più performante è il modello logistico.
Un ulteriore misura per valutare le prestazioni del modello è il cumulative gain chart (in egual modo è possibile considerare anche il lift chart), attraverso il quale è possibile misurare il guadagno di efficienza di un classificatore.
Questa tecnica ordina tutte le osservazioni in base alla probabilità più alta di ottenere la classe 'TRUE' della variabile *IsLegendary*, successivamente si misura il guadagno di efficienza che risulta essere maggiore quando si ottiene un'alta percentuale di osservazioni con classe positiva corrispondenti ad un sottoinsieme

molto piccolo del dataset.

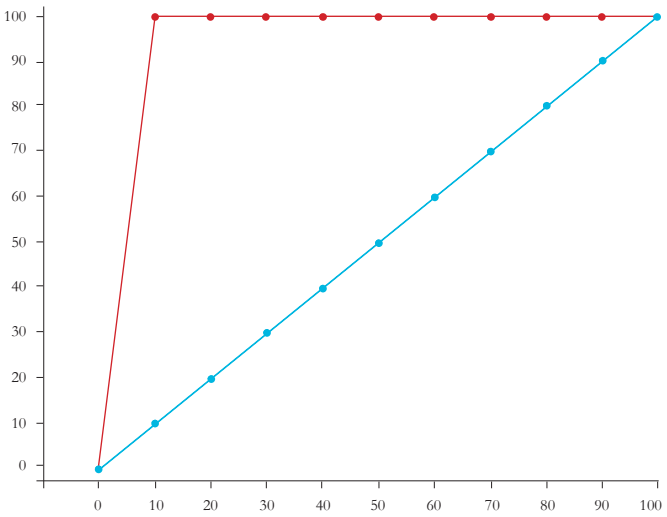

Figura 14: *Cumulative Gain - Naive Bayes Tree*

Il grafico soprastante misura la cumulative gain in favore del classificatore Naive Bayes Tree valutato nel dataset iniziale. In questo caso si può osservare che prendendo un sottoinsieme dei dati pari al 10% è possibile estrapolare il 100% delle osservazioni con classe positiva. Il guadagno di efficienza, quindi, risulta essere alto.

9. CONCLUSIONI E SVILUPPI FUTURI

A seguito di questo lavoro si è voluto identificare la tecnica di machine Learning migliore per la classificazione dei Pokemon leggendari. Il problema principale che si è cercato di affrontare è quello delle classi sbilanciate. Infatti se si applicasse un modello di classificazione senza le opportune tecniche migliorative, questo non permetterebbe di avere delle previsioni corrette perché si avrebbe la maggior parte di osservazioni previste per la classe "FALSE". Tra le soluzioni consigliate dalla letteratura, sono state impiegate la tecnica di campionamento equal size sampling, che in questo caso si è rilevata non adatta per la presenza di un dataset con un numero di osservazioni molto ridotte, e analisi dei costi. Con questa analisi si sono valutati diversi modelli di classificazione, confrontando quest'ultimi su due sottoinsiemi di dati diversi: il dataset iniziale e l'insieme selezionato dalla feature selection. Tra i due insiemi di dati non evincono particolari differenze in termini di performance, per entrambi la previsione migliore si riscontra con l'albero di classificazione naive bayes. In termini di costo, considerando il dataset inziale, rispetto al sottoinsieme valutato con la feature selection, si ha un costo nettamente minore. In conclusione, si evince che le previsioni in merito alla variabile *IsLegendary* migliori si hanno con il classificatore naive bayes tree, non incorrendo ad una riduzione della dimensionalità delle variabili esplicative mediante la tecnica descritta. Da questa analisi si può notare la complessità del mondo Pokémon e la vasta gamma di variabili presenti. Dalle basse probabilità di trovare un Pokémon leggendario si comprende come sia particolarmente difficile trovarne uno. I pokémon oltre che essere trovati nell'erba alta è possibile trovarli nelle uova, ma nonostante questo fattore , la probabilità di trovare un leggendario resta comunque bassa. Per finire l'analisi ci permette di evidenziare la possibilità di scovare un leggendario ogni qualvolta viene inserito un nuovo Pokémon.

10. REFERENZE

10.1 SITOGRAFIA

https://community.tibco.com/wiki/gains-vs-roc-curves-do-you-understand-difference

https://cran.r-project.org/web/packages/naniar/vignettes/naniar-visualisation.html

https://digidownload.libero.it/msa.economia/Lezione_2.pdf

https://towardsdatascience.com/meaningful-metrics-cumulative-gains-and-lyft-charts-7aac02fc5c14#:~:text=The%20cumulative%20gains%20curve%20is,target%20according%20to%20the%20model

https://www.kaggle.com/

https://www.knime.com/

https://www.kaggle.com/alopez247/pokemon (dataset)

https://www.pokemon.com/it/

https://pokemongohub.net/

https://www.rstudio.com/

10.2 BIBLIOGRAFIA

Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, *Introduction to Data Mining,* Pearson College Div, 2005;

Kevin P. Murphy, *Machine Learning - A Probabilistic Perspective,* The MIT Press Cambridge, 2012.