Dipartimento di Informatica
Corso di Laurea Magistrale in Informatica
Esame di Architetture Dati

Impatto di Outliers e Valori Nulli sui Modelli di Deep Learning e GLM di H2O AutoML: Analisi della Completezza e Consistenza del Dataset

Pallini Vincenzo - 907303
Nicolò Nicholas Zagami - 829888

Anno Accademico 2023 - 2024

Indice

1 Introduzione

Questo elaborato ha come obiettivo quello di presentare il progetto svolto per il corso di Architettura dati. Abbiamo condotto un'analisi dettagliata del dataset "Customer Analytics" con l'obiettivo di comprendere l'impatto degli outliers e dei valori nulli sulle **dimensioni di qualità** di un dataset e sulle **performance** di alcuni modelli di machine learning. Il dataset, disponibile su Kaggle, include dati rilevanti sui clienti e sulle loro transazioni, come il tipo di spedizione, il costo del prodotto e l'importo della sconto, tra gli altri. Tali informazioni sono essenziali per analizzare il comportamento dei clienti nell'e-commerce e sviluppare strategie di marketing efficaci.

Il nostro approccio ha coinvolto diverse fasi, partendo dalla comprensione e alla valutazione di modelli predittivi. Abbiamo implementato tecniche di preprocessing per convertire variabili categoriali in numeriche e normalizzare i dati, garantendo così la coerenza e la comparabilità delle variabili. Successivamente, abbiamo introdotto artificialmente outliers e valori nulli in percentuali variabili (5%, 10%, 15%, 20%) per valutare l'effetto di queste modifiche sui modelli di machine learning.

Abbiamo utilizzato diverse librerie Python, come *Pandas* per la manipolazione dei dati, *Scikit-Learn* per la costruzione dei modelli di machine learning, *Matplotlib* e *Seaborn* per la visualizzazione dei dati, e *H2O AutoML* e *TensorFlow* per l'automazione e l'ottimizzazione del processo di machine learning.

Infine sono stati analizzati i risultati dei modelli allenati con i dataset originali e quelli modificati, traendo conclusioni sull'importanza della gestione dei dati anomali e mancanti. Questo studio fornisce insight utili per migliorare le strategie di preprocessing e la gestione dei dati nelle applicazioni di data science, contribuendo a sviluppare modelli predittivi più robusti ed efficaci.

1.1 Cos'è un outlier?

Un outlier rappresenta un valore anomalo in un insieme di osservazioni. Gli outliers sono importanti perché possono avere un impatto significativo sui risultati delle analisi statistiche e dei modelli di machine learning.

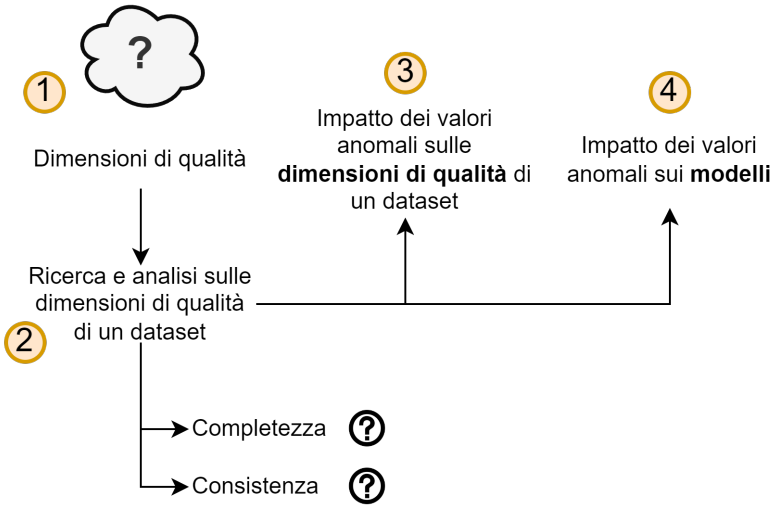
2 Descrizione del dataset

Il dataset di Customer Analytics disponibile su Kaggle contiene informazioni dettagliate sui clienti e sulle loro transazioni. È composto da varie colonne che rappresentano diverse caratteristiche rilevanti per l'analisi del comportamento dei clienti nel contesto dell'e-commerce. Le principali colonne del dataset includono:

- **ID**: Identificativo unico per ogni cliente.
- **Warehouse block**: Il blocco di magazzino da cui l'ordine è stato spedito.
- **Mode of shipment**: Il metodo di spedizione utilizzato (nave, volo o strada).
- **Customer care calls**: Numero di chiamate al servizio clienti fatte dal cliente.
- **Customer rating**: Valutazione data dal cliente al servizio ricevuto.
- **Cost of the product**: Costo del prodotto acquistato.
- **Prior purchases**: Numero di acquisti precedenti effettuati dal cliente.
- **Product importance**: Importanza del prodotto (bassa, media, alta).
- **Gender**: Genere maschile o femminile del cliente.
- **Discount offered**: Sconto offerto sul prodotto.
- **Weight in gms**: Peso del prodotto in grammi.
- **Reached on time**: Indicatore se il prodotto è arrivato in tempo (1) o in ritardo (0).

Questo dataset è particolarmente utile per analizzare vari aspetti del comportamento dei clienti, come la loro soddisfazione, l'efficacia del servizio clienti, l'impatto degli sconti, e la puntualità delle spedizioni. Attraverso l'analisi di queste variabili, è possibile ottenere insight significativi per migliorare le strategie di marketing e ottimizzare le operazioni aziendali.

3 Ipotesi dello studio



L'obiettivo del nostro studio è esaminare come gli outliers e i valori nulli influenzano alcune dimensioni di qualità di un dataset e i modelli presi in esame. In particolare, andremo ad analizzare:

- **Completezza** (di un insieme di dati) che descrive tutte le osservazioni e gli elementi necessari al raggiungimento di uno scopo, sia esso di analisi o meno.
- **Consistenza**, la quale può assumere due significati: la consistenza definita da uno schema (ad esempio quello definito per un database) oppure la consistenza nella rappresentazione dell'oggetto.

Passiamo quindi alle ipotesi:

- Q1: *Qual è l'impatto di outliers e valori nulli sulle performance dei modelli?*
- Q2: *La presenza di outliers e valori nulli hanno lo stesso impatto?*
- Q2.1: *Se no, perché uno impatta più dell'altro?*
- Q3: *Qual è l'impatto sulle dimensioni di qualità di un dataset invece?*

Ognuna di queste ipotesi troverà risposta all'interno della relazione, nel prossimo capitolo verrà illustrata la metodologia di lavoro utilizzata e relativi esperimenti fatti.

4 Metodologia di lavoro

In questo capitolo descriveremo la metodologia adottata per analizzare il dataset "Customer Analytics", con particolare attenzione agli outliers e ai valori nulli, e come questi influenzano le dimensioni di qualità del dataset e le performance dei modelli di machine learning.

4.1 Schema logico

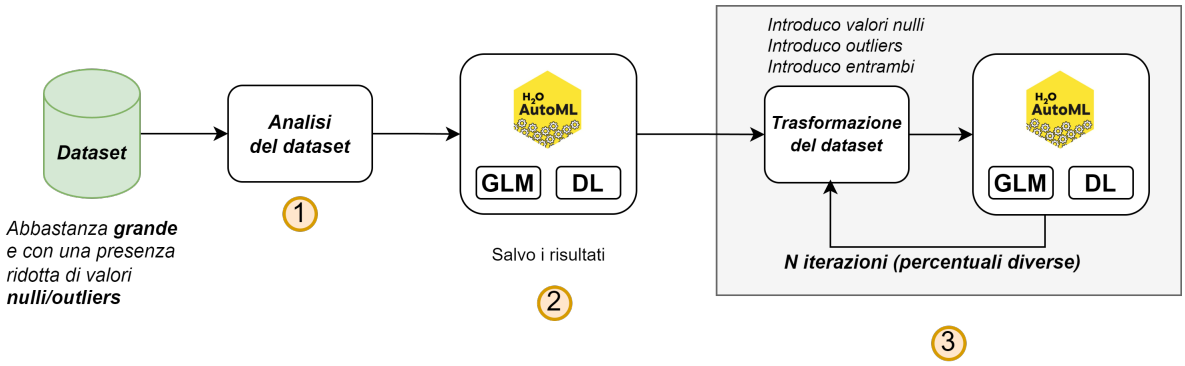


Fig. 1: Schema logico della metodologia di lavoro applicata

4.2 H2O autoML

Tra gli strumenti utilizzati sfruttiamo la funzionalità di "AutoML" dello strumento open source chiamato H2O: la quale permette a chi la utilizza di allenare un numero configurabile di modelli di ML; nel nostro caso ne abbiamo scelti due in particolare, i quali verranno approfonditi nel prossimo capitolo.

Tutto questo ci ha portato a creare una procedura automatizzata il cui unico obiettivo era quello di studiare gli effetti delle "trasformazioni" applicate al dataset di partenza.

La procedura di AutoML di H2O può essere riassunta nei seguenti step:

1. **Inizializzazione di H2O**: in questa fase importiamo e inizializziamo il cluster di H2O, il cui compito è quello di mettere in piedi "l'infrastruttura".

2. **Preparazione del dataset**: fase in cui dividiamo il training set dal test set e convertiamo il dataset in un frame di H2O.

3. **Addestramento AutoML**: infine configuriamo in base alle nostre esigenze il numero di modelli di cui abbiamo bisogno per poi svolgere gli esperimenti

4. **Estrazione delle metriche**: sia che il problema sia di classificazione che di regressione, H2O permette l'esportazione sottoforma di leaderboard delle metriche dell'addestramento, in modo da fornire una chiara panoramica delle performance dei modelli.

Il nostro caso tratta un problema di classificazione binaria, di conseguenza verranno valutate le seguenti metriche:

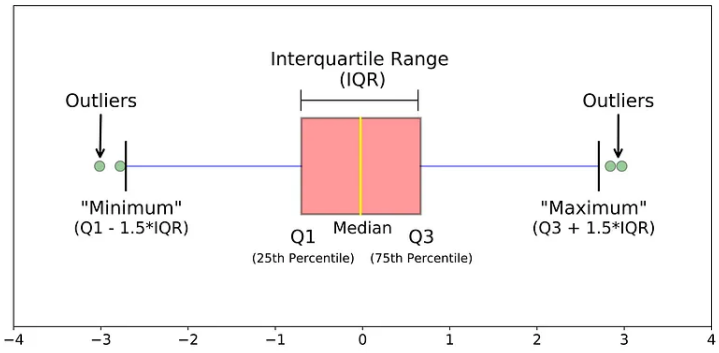
- AUC

- Logloss
- AUCPR
- mean per class error
- RMSE
- MSE

4.3 Dettagli implementativi

Approfondimento delle scelte implementative fatte all'interno del progetto.

4.3.1 Generazione degli outliers



- Il *metodo del range interquartile (IQR)* viene utilizzato per identificare gli outliers all'interno di un dataset; noi lo abbiamo applicato in maniera inversa per cercare tutti gli indici dei valori non-outliers.
- Attraverso una percentuale P variabile introduciamo nuovi outliers prendendo solo i valori che fanno parte della lista di valori non outliers.

4.3.2 Generazione dei valori nulli

Per generare dei valori nulli all'interno del dataset basta sostituire dei valori casuali all'interno del dataset con il valore nullo; anche in questo caso utilizziamo una percentuale P variabile.

4.3.3 Label Encoding

Height
Tall
Medium
Short

Height
0
1
2

Per trasformare le variabili categoriche in variabili numeriche abbiamo utilizzato la classe `LabelEncoder` del package `sklearn.preprocessing`; la trasformazione delle feature è pressoché obbligatoria poiché i modelli di machine learning operano principalmente con valori numerici.

4.4 Esperimenti fatti

4.4.1 H2O AutoML sui dati originali e dataset con valori nulli

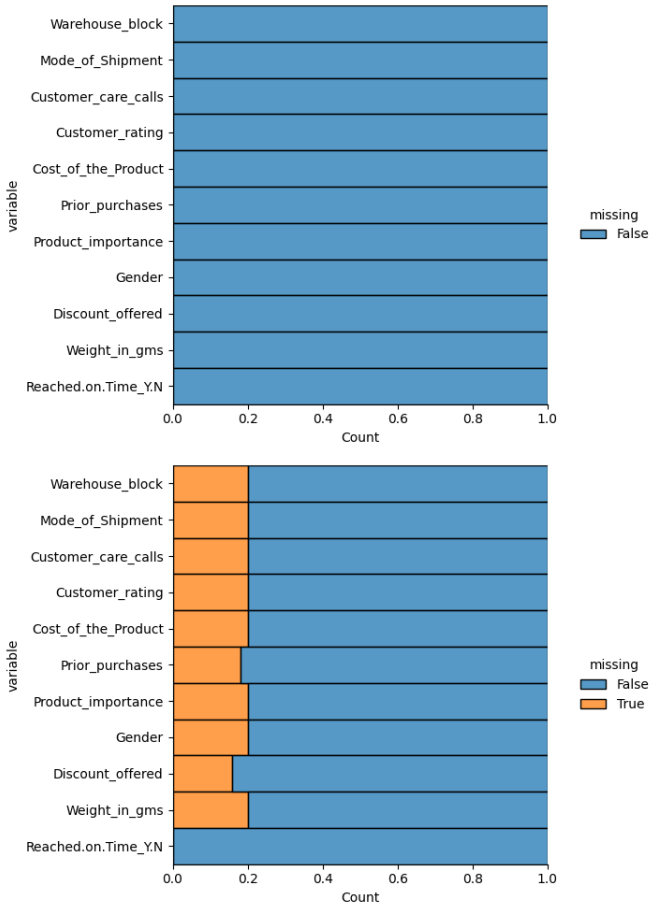


Fig. 2: Progressione del numero di valori nulli nel dataset di partenza

- Creazione di 4 dataset con diverse percentuali di valori nulli al suo interno: 5%, 10%, 15%, 20%.
- Addestramento di H2O AutoML con il dataset originale.
- Addestramento di H2O AutoML con i dataset alterati.
- Analisi e valutazione delle performance attraverso la leaderboard dei modelli.

4.4.2 H2O AutoML sui dati originali e dataset con outliers

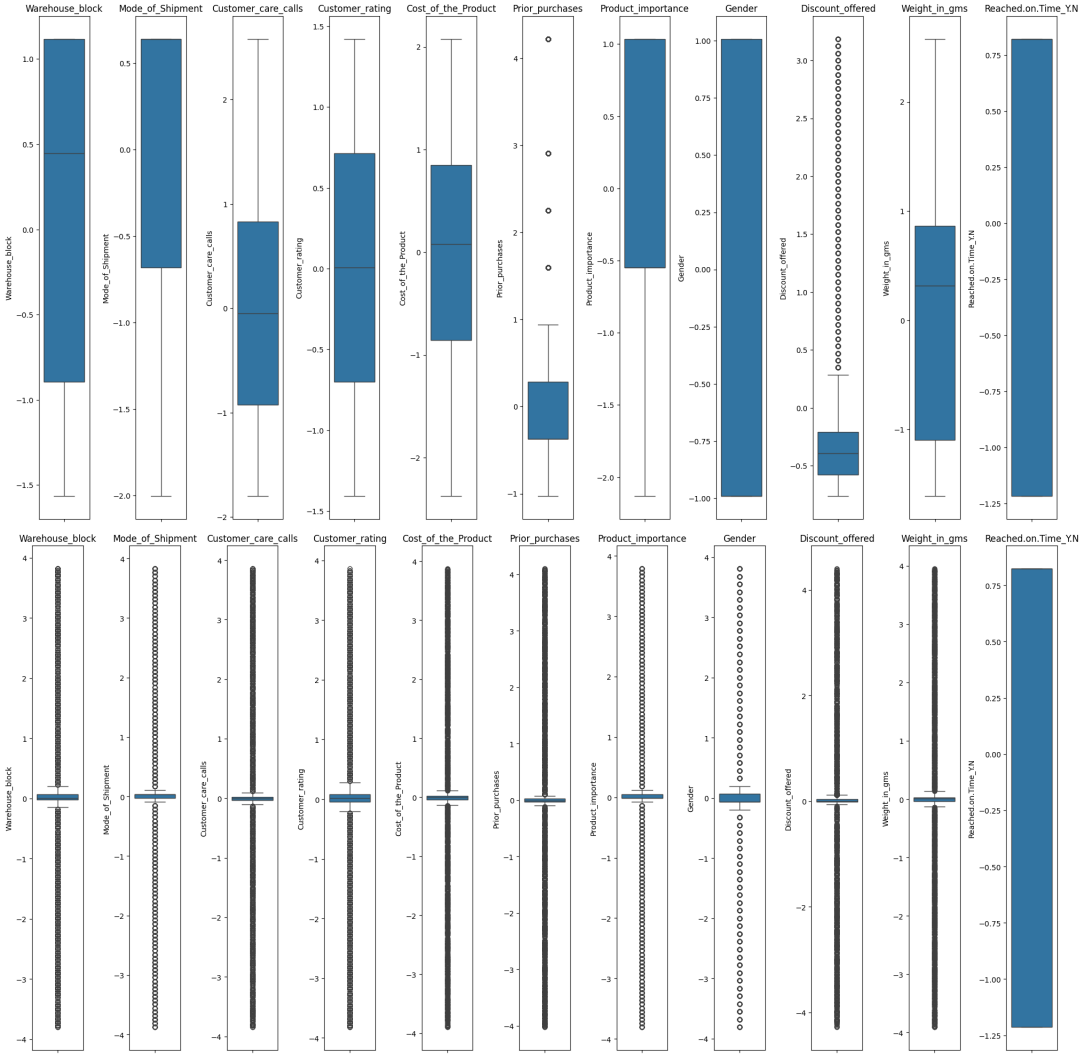


Fig. 3: Dataset prima e dopo l'inserimento di outliers

- Creazione di 4 dataset con diverse percentuali di outliers al suo interno: 5%, 10%, 15%, 20%.
- Addestramento di H2O AutoML con il dataset originale.
- Addestramento di H2O AutoML con i dataset alterati.
- Analisi e valutazione delle performance attraverso la leaderboard dei modelli.

4.4.3 H2O AutoML sui dati originali e dataset con outliers/valori nulli

- Creazione di 4 dataset con diverse percentuali di outliers/valori nulli al suo interno: 5%, 10%, 15%, 20%.
- Addestramento di H2O AutoML con il dataset originale.

- Addestramento di H2O AutoML con i dataset alterati.
- Analisi e valutazione delle performance attraverso la leaderboard dei modelli.

5 Modelli presi in esame

In questo capitolo, descriviamo due modelli di machine learning utilizzati per la nostra analisi: il Generalized Linear Model (GLM) e un modello di Deep Learning. Questi modelli sono stati selezionati tramite H2O AutoML, che automatizza il processo di selezione e tuning dei modelli. I modelli GLM e Deep Learning sono noti per la loro sensibilità ai valori nulli e agli outliers, che possono influenzare significativamente le loro prestazioni.

5.1 GLM: Generalized Linear Model

Il *Generalized Linear Model (GLM)* è stato utilizzato per costruire un modello di regressione lineare che può gestire vari tipi di distribuzioni della variabile di risposta. Nel nostro caso trattandosi di un problema di classificazione binaria l'autoML di H2O adotta la **regressione logistica binaria**, la quale risponde in termini di 0 e 1.

5.1.1 Caratteristiche principali del GLM

1. **Funzione di Link:** la funzione di link serve ad associare variabili numeriche (con valori reali) con una variabile di risposta binaria attraverso una combinazione lineare; nel caso della *regressione logistica* si utilizza la *funzione di logit.*

2. **Combinazione lineare:** Essendo il GLM basato sul modello di regressione lineare ordinario mantiene quella che è la combinazione lineare classica, che viene successivamente combinata alla funzione di link.

3. **Distribuzione della famiglia:** Descrive la "forma" matematica che viene utilizzata dal modello per descrivere i dati; nel caso della *regressione logistica* utilizziamo la *distribuzione binomiale.*

5.2 Deep Learning

Il modello di Deep Learning fornito dallo strumento di AutoML di H2O è basato su una *FNN (Feedforward Neural Network)*; la quale è composta da più strati di neuroni e viene allenata attraverso un algoritmo di **backpropagation** basato sulla **discesa del gradiente**.

5.2.1 Caratteristiche di una FNN

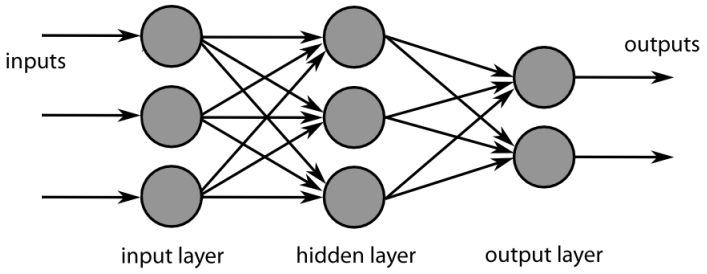


Fig. 4: Strati di una feedforward neural network

1. **Multi-layer Perceptron (MLP)**: Le FNN sono composte da diversi strati di neuroni interconnessi fra di loro, e sono composte dai seguenti strati:

- **Input Layer**: È il primo strato della rete che riceve i dati di input. Ogni neurone nell'input layer rappresenta una caratteristica del dataset.
- **Hidden Layers**: Questi strati intermedi eseguono la maggior parte del calcolo e delle trasformazioni sui dati. Ogni neurone in un hidden layer prende input dai neuroni dello strato precedente, applica una funzione di attivazione e passa il risultato allo strato successivo.
- **Output Layer**: È l'ultimo strato della rete che produce il risultato finale. Il numero di neuroni nell'output layer dipende dal tipo di problema (ad esempio, classificazione binaria o multiclasse).

2. **Algoritmo di backpropagation**: si occupa di aggiornare i pesi dei neuroni presenti nella rete in modo da minimizzare l'errore riscontrato durante l'allenamento; questo algoritmo si basa su due concetti chiave:

   - **Loss Function**, misura della discrepanza tra l'output della rete e l'output desiderato; di fatto verifica la qualità della predizione della rete determinando se sia necessario qualche accorgimento in fase di allenamento della rete.
   - **Gradient of the loss function**, misura l'impatto necessario a minimizzare l'errore, esplicitando "quanto" sia necessario intervenire sul **peso del neurone** e sulla **direzione** che esso deve prendere.

5.3 Sensibilità ai valori nulli e agli outliers

Facendo riferimento alla ipotesi Q1 parliamo delle performance dei modelli in presenza di valori nulli o outliers all'interno del dataset:

1. **Outliers:** Gli outliers distorcono direttamente la distribuzione dei dati. Nei GLM, alterano i coefficienti delle variabili indipendenti, causando una relazione inaccurata tra le variabili.
   Nelle FNN, gli outliers causano gradienti elevati durante l'addestramento, portando a pesi dei neuroni inaccurati e a un modello che non generalizza bene. Questo effetto è amplificato con l'aumento della percentuale di outliers generati.

2. **Valori Nulli:** I valori nulli portano bias nelle stime dei GLM, inoltre introducendo valori anomali riduciamo di fatto il campione di informazione da cui il modello può attingere. Nelle FNN invece hanno un effetto di distorsione sui pesi assegnati ai neuroni presenti nella rete, rendendo di fatto più difficile l'allenamento.

6 Risultati

In generale è evidente che la presenza di una forte percentuale di valori nulli/outlier influenzi negativamente l'apprendimento dei modelli di machine learning, specialmente se essi sono più sensibili verso questi valori: come il GLM e Deep learning.

Tuttavia, facendo riferimento all'ipotesi Q2 possiamo dire che gli outliers hanno un **impatto maggiore** rispetto ai soli valori nulli durante la fase di addestramento per questi due modelli:

Dataset	Model	auc	logloss	aucPR	mean per class error	RMSE	MSE
Normal train	GLM_1_AutoML_14_20240615_90502	0.739519	0.533020	0.851551	0.499290	0.435129	0.189337
	DeepLearning_1_AutoML_14_20240615_90502	0.733618	0.518559	0.849923	0.499887	0.433452	0.187880
5% of outliers	DeepLearning_1_AutoML_15_20240615_90517	0.722281	0.534782	0.838494	0.498598	0.437948	0.191799
	GLM_1_AutoML_15_20240615_90517	0.719631	0.549873	0.837647	0.498321	0.442546	0.195847
10% of outliers	DeepLearning_1_AutoML_16_20240615_90529	0.720422	0.543186	0.836598	0.5	0.441522	0.194942
	GLM_1_AutoML_16_20240615_90529	0.718281	0.554785	0.835100	0.497255	0.444302	0.197404
15% of outliers	GLM_1_AutoML_17_20240615_90542	0.713219	0.561013	0.831052	0.4973	0.446711	0.199551
	DeepLearning_1_AutoML_17_20240615_90542	0.712569	0.552939	0.831114	0.4972	0.445434	0.198412
20% of outliers	GLM_1_AutoML_18_20240615_90551	0.709492	0.566118	0.827668	0.499487	0.448688	0.201321
	DeepLearning_1_AutoML_18_20240615_90551	0.709412	0.551341	0.829018	0.496933	0.444673	0.197734

Tabella 1: Dataset con soli valori nulli

Dataset	Model	auc	logloss	aucPR	mean per class error	RMSE	MSE
Normal train	DeepLearning_1_AutoML_1_20240615_84039	0.739713	0.507185	0.852709	0.496186	0.427178	0.182481
	GLM_1_AutoML_1_20240615_84039	0.739519	0.533020	0.851551	0.499290	0.435129	0.189337
5% of outliers	DeepLearning_1_AutoML_3_20240615_84935	0.680315	0.617313	0.787194	0.498512	0.466121	0.217269
	GLM_1_AutoML_3_20240615_84935	0.546675	0.674151	0.623057	0.499814	0.490465	0.240555
10% of outliers	DeepLearning_1_AutoML_4_20240615_84947	0.574089	0.676133	0.663078	0.5	0.491184	0.241262
	GLM_1_AutoML_4_20240615_84947	0.496617	0.674648	0.589564	0.5	0.490720	0.240807
15% of outliers	DeepLearning_1_AutoML_5_20240615_85001	0.520841	0.681724	0.604508	0.5	0.493613	0.243654
	GLM_1_AutoML_5_20240615_85001	0.499471	0.674616	0.594662	0.499673	0.490705	0.240791
20% of outliers	DeepLearning_1_AutoML_6_20240615_85012	0.524245	0.681355	0.608127	0.499719	0.493785	0.243824
	GLM_1_AutoML_6_20240615_85012	0.497622	0.674657	0.594767	0.499859	0.490725	0.240811

Tabella 2: Dataset con soli outliers

Con riferimento all'ipotesi Q2.1 tra le ragioni principali troviamo:

- *Maggiore variazione nei dati*, gli outliers tendono ad aumentare significativamente **varianza** e **deviazione standard**, due misure che descrivono le dispersione dei dati attorno alla media. Inoltre la **media** è particolarmente influenzata dagli outliers, poiché dei valori molto grandi portano inevitabilmente la media a crescere
- *Sensibilità agli outliers da parte dei modelli di machine learning*, infatti alcuni modelli risultano essere più sensibili alla presenza di outliers: come nel caso del **GLM** e **Deep Learning**, questo è dovuto all'influenza che questi valori hanno verso la funzione **MSE**; infatti è possibile apprezzarne un aumento nei risultati una volta introdotti degli outliers nel dataset di partenza.

Per quanto riguarda le metriche rimanenti invece:

- **AUC (Area Under the ROC Curve)**, la quale rappresenta il grado di separabilità tra le classi del dataset, il trend calante di questa metrica ci mostra come il modello per colpa degli outliers perde la capacità di predizione verso le classi del problema.
- **AUCPR (Area Under the Precision-Recall Curve)**, anche in questo caso la precisione potrebbe essere compromessa se valori anomali fossero classificati come positivi.

- La **Logloss** misura l'incertezza delle probabilità del modello confrontandole alle etichette vere (quelle presenti nel dataset); gli outliers contribuiscono a fare in modo che la logloss cresca producendo previsioni errate, allo stesso modo i valori nulli poiché portano il dataset ad avere sempre meno informazioni su cui i modelli possono fare riferimento.

Per rispondere invece all'ipotesi Q3 possiamo dire che:

- **Rispetto alla completezza** possiamo sviluppare più ragionamenti, intanto è importante sottilineare quanto gli outliers non compongano un reale problema verso la completezza di un dataset rispetto ai valori nulli, i secondi infatti introducono problemi in fase di analisi e rendono i dataset più incompleti; di conseguenza anche meno usabili. Per ovviare a questo problema e ottenere delle performance migliori spesso si preferisce riempire i valori nulli con altri valori.

- **Nel caso della consistenza** gli outliers giocano un ruolo fondamentale: in quanto possono rompere la consistenza con cui vengono definiti gli oggetti.
  Pensiamo ad esempio al sesso di un individuo, ad una valutazione definita attraverso una rigida scala di valori o alle modalità di spedizione di un'azienda di trasporti; attraverso l'uso di valori anomali è possibile alterare tutte quelle variabili che seguono uno schema ben definito.
  A questo si aggiunge la correlazione fra le informazioni, come ad esempio il codice fiscale con il nome, il cognome e l'età di una persona; tutte informazioni che, se alterate con un valore anomalo (outlier o valore nullo) andrebbero a compromettere la coerenza di più dati all'interno del dataset.

Non sono stati rilevati risultati interessanti producendo dei dataset misti (valori nulli e outliers insieme) rispetto ai risultati ottenuti con **soli outliers**, analizzando quanto illustrato fino ad ora si può giungere alla conclusione che l'impatto degli outliers rispetto a quello dei valori nulli fa in modo che non si rivelino dei risultati significativi.

Dataset	Model	auc	logloss	aucPR	mean per class error	RMSE	MSE
Normal train	GLM_1_AutoML_20_20240615_90654	0.739519	0.533020	0.851551	0.499290	0.435129	0.189337
	DeepLearning_1_AutoML_20_20240615_90654	0.734448	0.522624	0.850692	0.499887	0.435276	0.189465
5% of outliers	DeepLearning_1_AutoML_21_20240615_90704	0.686576	0.611356	0.795604	0.499392	0.463976	0.215273
	GLM_1_AutoML_21_20240615_90704	0.501054	0.674670	0.594843	0.499859	0.490731	0.240817
10% of outliers	DeepLearning_1_AutoML_22_20240615_90715	0.557979	0.674849	0.645331	0.5	0.490630	0.240717
	GLM_1_AutoML_22_20240615_90715	0.510647	0.674541	0.597554	0.5	0.490666	0.240753
15% of outliers	DeepLearning_1_AutoML_23_20240615_90724	0.542366	0.679794	0.619417	0.5	0.492837	0.242888
	GLM_1_AutoML_23_20240615_90724	0.506417	0.674444	0.601848	0.499859	0.490620	0.240708
20% of outliers	DeepLearning_1_AutoML_24_20240615_90732	0.526666	0.680464	0.614958	0.499814	0.493418	0.243461
	GLM_1_AutoML_24_20240615_90732	0.483020	0.674780	0.574496	0.499859	0.490785	0.240870

Tabella 3: Dataset con valori misti (nulli e outliers)

7 Conclusione e sviluppi futuri

L'analisi condotta sugli effetti degli outliers e dei valori nulli sui modelli di Deep Learning e GLM di H2O AutoML ha messo in luce diversi aspetti critici riguardanti la qualità del dataset in termini di completezza e consistenza.

Impatto degli outliers: La performance dei modelli è stata influenzata in modo significativo dagli outliers. L'aumento dell'*errore quadratico medio (MSE)* e della logloss a seguito della loro presenza mostra che le previsioni del modello sono più incerte. Inoltre, le metriche di AUC e $AUCPR$ sono diminuite, il che indica che il modello non è in grado di distinguere correttamente tra le classi. Questo risultato mostra come gli outliers possono compromettere la consistenza del dataset, influenzando variabili importanti e compromettendo la coerenza delle informazioni.

Effetti dei valori nulli: I valori nulli, invece, hanno un impatto principalmente sulla completezza del set di dati. La loro presenza riduce la quantità di dati disponibili per i modelli, il che significa che le performance dei modelli sono meno accurate. È comune sostituire i valori nulli con stime appropriate per risolvere questo problema. Ciò migliora la qualità del dataset e le prestazioni dei modelli.

Confronto tra outliers e valori nulli: L'analisi comparativa tra dataset contenenti solo outliers, solo valori nulli, e una combinazione di entrambi ha mostrato che l'impatto degli outliers è generalmente più pronunciato rispetto a quello dei valori nulli. I valori nulli causano principalmente una perdita di dati, mentre gli outliers distorcono maggiormente i risultati. Questa distinzione è fondamentale per la creazione di strategie di preprocessing che mirano a migliorare le prestazioni dei modelli e la qualità del dataset.

Sviluppi futuri: Al fine di ottenere risultati più consistenti è necessario approfondire lo studio dei modelli di machine learning, di rilevamento e gestione degli outliers e dei valori nulli; questo non porta beneficio solo alle performance dei modelli ma anche alle dimensioni di qualità del dataset. In questa relazione ci siamo focalizzati esclusivamente sulla *completezza* e la *consistenza* dei dati, tuttavia ci sarebbero altre dimensioni da esplorare, le quali potrebbero portare alla formulazione di nuove ipotesi e, di conseguenza, a nuovi risultati.

References

[1] AI ML Analytics. *Categorical Encoding: Label Encoding*. 2022. URL: https://shorturl.at/OPYu4.

[2] Mustafa Celik. *Outlier Analysis in Python*. 2022. URL: https://medium.com/@mstffclkk/outlier-analysis-in-python-6c21a6183004.

[3] CK-12 Foundation. *How Do Outliers Influence Measures of Central Tendency and Dispersion?* 2024. URL: https://www.ck12.org/flexi/math-grade-6/spreaddispersion-range-range-of-spreaddispersion/how-do-outliers-influence-measures-of-central-tendency-and-dispersion/#:~:text=Mean%3A%20Outliers%20can%20pull%20the,values%20can%20lower%20the%20mean..

[4] V7 Labs. *Performance Metrics in Machine Learning*. 2024. URL: https://www.v7labs.com/blog/performance-metrics-in-machine-learning.

[5] Prachi13. *Customer Analytics Dataset*. 2024. URL: https://www.kaggle.com/datasets/prachi13/customer-analytics/data.

[6] Sasirekha Rameshkumar. *Deep Learning Basics (Part 10): Feed Forward Neural Networks (FFNN)*. 2023. URL: https://medium.com/@sasirekharameshkumar/deep-learning-basics-part-10-feed-forward-neural-networks-ffnn-93a708f84a31.

[7] Towards Data Science. *Generalized Linear Models*. 2024. URL: https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab.

[8] Università degli Studi di Milano-Bicocca. *Data Quality*. 2024. URL: https://elearning.unimib.it/pluginfile.php/1711254/mod_resource/content/1/9%20Data%20Quality.pdf.

[9] Turing. *Mathematical Formulation of Feed Forward Neural Network*. 2023. URL: https://www.turing.com/kb/mathematical-formulation-of-feed-forward-neural-network.

[10] Wikipedia. *Logit*. 2024. URL: https://it.wikipedia.org/wiki/Logit.