

"MONITORAGGIO ENERGETICO ED AMBIENTALE"

STUDIO DI FATTIBILITA' PIATTAFORMA METERING

Data Science e Machine Learning

Indice

Indice	2
Competenze tecniche e Infrastruttura informatica necessarie.....	3
Processo di Data Science	3
Framework logico	3
Attori coinvolti e principali utilizzatori	4
Piattaforma	4
Caratteristiche tecniche	4
Scenari da gestire.....	5
Tipi di Algoritmi.....	5
Manutenzione Predittiva	6
Obiettivi	6
Dati necessari.....	6
Tecniche di modellazione.....	6
Anomaly Detection	6
Obiettivi	7
Dati necessari.....	7
Tecniche di modellazione.....	7
Machine Vision.....	7
Obiettivi	7
Dati necessari.....	7
Tecniche di modellazione.....	7
Analisi avanzata dei dati.....	8
Obiettivi	8
Dati necessari.....	8
Tecniche di modellazione.....	8

Competenze tecniche e Infrastruttura informatica necessarie

Processo di Data Science

Framework logico

Le infrastrutture informatiche che verranno installate dovranno garantire la possibilità di lavorare nel rispetto della metodologia CRISP-DM (CRoss Industry Standard Process for Data Mining) o eventualmente TDSP (Team Data Science Process).



Figura 1: Tipica raffigurazione del processo CRISP-DM

Si ricordano a scopo puramente indicativo le fasi che compongono il processo:

- Analisi dei requisiti (Business Understanding);
- Analisi preliminare dei dati (Data Understanding);
- Trasformazione e pulizia dei dati (Data Preparation);
- Modellazione (Modeling);
- Valutazione delle performance (Evaluation);
- Distribuzione dei modelli (Deployment).

Una volta che i dati sono stati raccolti, sarà compito del data scientist analizzarli e capire le relazioni che intercorrono tra le variabili raccolte, in modo da definire un modello in grado di replicare i dati generati nell'edificio, e dunque poterne prevedere o simulare i comportamenti.

Sarà compito del data scientist documentare rigorosamente il processo di selezione del modello in modo da ottenere quello che garantisce le performance migliori, date le metriche di accuratezza.

La strutturazione di sistemi di machine learning è frutto dell'interazione tra uno o più Data Scientist e gli esperti di facility management identificati da COTRAL S.p.A.. Tale interazione è necessaria per definire in modo operativo ed estremamente preciso il perimetro applicativo delle applicazioni. Nonostante possa sembrare marginale, la definizione operativa degli obiettivi dei modelli di machine learning è essenziale per

poter strutturare al meglio gli obiettivi finali e minimizzare gli sforzi in termini di tempo-macchina per l'addestramento degli algoritmi.

Attori coinvolti e principali utilizzatori

Il sistema dovrà garantire ad una serie di attori la possibilità di collaborare tra loro:

- Project Manager: verifica lo stato di avanzamento degli sviluppi e certifica l'eventuale raggiungimento delle performance attese;
- Business Analyst e Product Owner: monitorano le performance raggiunte e supportano i Data Scientist nella definizione dei requisiti di base;
- Data Scientist: la persona delegata in prima istanza all'analisi dei dati e all'addestramento dei Modelli;
- Data Engineer: la persona delegata alla trasformazione dei dati;
- Sviluppatore: la persona che dovrà integrare le previsioni con eventuali applicazioni esterne.

Piattaforma

La piattaforma di data science da adottare deve consentire la gestione integrata di tutto il ciclo di vita dei modelli di machine learning in modo aderente a paradigmi come CRISP-DM e TDSP.

La creazione e la gestione di modelli di machine learning deve basarsi su infrastrutture dedicate il cui compito finale è:

- Agevolare le attività del Data Scientist, sollevandolo dai compiti di gestione delle infrastrutture e della distribuzione delle previsioni;
- Agevolare le attività degli sviluppatori, facilitando l'integrazione delle previsioni con le altre applicazioni di gestione degli edifici.

Esistono diversi software di mercato capaci di coprire tutta la filiera di cui sopra, utilizzando sistemi più o meno automatizzati, partendo dalle funzioni di auto-machine-learning per arrivare a codici custom scritti in linguaggi come Python o R. Questi linguaggi di programmazione costituiscono ormai uno standard de facto per la gestione del ciclo di vita di applicazioni di machine learning.

Sistemi che garantiscono l'utilizzo dei linguaggi di programmazione come Python o R sono preferibili a sistemi chiusi (non a programmazione) in modo da garantire portabilità dei progetti di data science.

Caratteristiche tecniche

La piattaforma tipicamente risiede su server fisici, virtuali o cloud (sia essa PaaS, o IaaS) interamente dedicati allo scopo.

Dato che gli algoritmi di machine learning spesso necessitano di molte risorse infrastrutturali (HD, RAM, CPU), questo capitolo mette in particolare evidenza la scalabilità della soluzione da implementarsi.

- Scalabilità: le fasi di addestramento possono utilizzare grandi quantità di risorse fisiche, soprattutto RAM e CPU. A tale scopo, sistemi scalabili come quelli cloud possono facilitare la prima prototipazione dei modelli di machine learning e consentono di calcolare le risorse da rendere disponibili on premise.

Si segnalano due casi ulteriori relativi alla scalabilità:

- La scalabilità deve comprendere la possibilità (compatibilità del software e fisica) di utilizzare processori del tipo graphics processing unit (GPU), per garantire la possibilità di addestrare algoritmi di deep learning.

- La possibilità (compatibilità del software e fisica) di lanciare processi batch su dati storici in modalità di calcolo distribuito.
- Sicurezza: la piattaforma dovrà essere accessibile solamente alle persone autorizzate a farlo. Si noti come la piattaforma ospiterà dati potenzialmente sensibili (ad es. immagini).
- Portabilità: i modelli da addestrare dovranno essere quanto più indipendenti dalla tecnologia della piattaforma utilizzata, in modo da garantire un eventuale sostituzione della piattaforma, senza dover riscrivere gli algoritmi di addestramento e distribuzione dei modelli.

Scenari da gestire

- Creazione e gestione di Progetti: la piattaforma deve consentire la creazione e la gestione di diversi progetti, ognuno dei quali deve poter accedere unicamente agli elementi software ad esso dedicati (dataset, pipeline, endpoint, ...). La definizione dei progetti deve seguire flussi che siano aderenti alla metodologia CRISP-DM.

I Progetti devono essere accessibili da persone non tecniche, con la finalità di monitorarne l'avanzamento e le performance attese.

- Creazione e gestione di Dataset: la piattaforma deve consentire la creazione e la gestione di dataset, in modo da garantire un'estrazione solamente periodica di dati dai sistemi di salvataggio dati (NB: queste estrazioni possono essere molto onerose, e dunque giustificano una ridondanza dei dati).
- Analisi preliminare dei dati: la piattaforma deve garantire la possibilità di manipolazione dei dati tramite linguaggi di programmazione come Python o R, supportati da framework di visualizzazione come Jupyter Notebook.
- Creazione e gestione di Modelli: uno dei semilavorati dei Progetti sono i Modelli, oggetti che vanno salvati su supporti appositi, storicizzati e resi disponibili ai Data Scientist e Sviluppatori per la distribuzione delle previsioni.
- Creazione e gestione di Endpoint: il flusso di interrogazione dei Modelli deve poter essere quanto più automatico possibile. La piattaforma deve occuparsi in modo automatico di rendere i Modelli interrogabili da servizi esterni tramite web service o pipeline temporizzate (per previsioni in modalità batch).
- Monitoraggio degli sviluppi e delle performance: le previsioni erogate devono poter essere continuamente confrontate con dati reali presi dal campo con la finalità di monitorare l'accuratezza delle previsioni stesse ed individuare eventuali derive di precisione/variabilità.

Tipi di Algoritmi

L'apprendimento automatico e gli approcci statistici basati principalmente sui dati di certificazione energetica degli edifici sono stati identificati e analizzati in due gruppi:

- (1) valutazione automatica delle prestazioni energetiche degli edifici
- (2) previsione di un retrofit efficiente dal punto di vista energetico

L'obiettivo del presente capitolo è fornire una serie di approcci (tra i più utilizzati e più appropriati) che la piattaforma di data science dovrà consentire di sviluppare per analizzare le prestazioni energetiche di diversi tipi di edifici.

Manutenzione Predittiva

Le performance energetiche di un edificio dipendono da una serie di dispositivi come impianti di condizionamento e deumidificazione, sensori, pannelli solari, etc. Eventuali malfunzionamenti dei dispositivi possono impattare negativamente sull'impronta energetica dell'edificio. Dunque, anticipare quanto possibile le manutenzioni da effettuare può garantire, sul lungo periodo, performance energetiche migliori.

L'output degli algoritmi di Manutenzione Predittiva consiste in allarmi sollevati verso gli Operatori di area SFE e/o figure preposte.

Obiettivi

Per ogni tipologia di asset potranno essere sviluppati uno o più modelli di machine learning. Si suggeriscono i seguenti:

- Failure probability: fissato un tipo di guasto e un intervallo di tempo (ad es. una settimana), fornisce la probabilità che per l'asset in questione si verifichi il guasto selezionato nell'intervallo di tempo.
- Remaining Useful Life: fissato un tipo di guasto, fornisce le ore di utilizzo mancanti per l'asset prima di un possibile guasto. Eventualmente può anche essere fornita una curva di probabilità (analisi di sopravvivenza).
- Number of failures: dato un intervallo di tempo fissato (ad es. 1 mese), questo modello prevede il numero di guasti che potenzialmente potranno coinvolgere l'asset.

Questi modelli consentono sia una consultazione “AS IS” delle previsioni per l'asset, sia la fruizione in modalità “WHAT IF”, ovvero simulando certe condizioni di utilizzo (pertinenti alla telemetria disponibile).

Dati necessari

L'addestramento di questi modelli necessita di una serie di dati come:

- Dati di telemetria derivanti dai sensori IoT;
- Dati di impostazione e/o eventuale gestione degli asset;
- Dati di fabbricazione degli asset (caratteristiche, età, ...);
- Dati di manutenzione (guasti storicizzati, manutenzioni ordinarie effettuate).

Tecniche di modellazione

A seconda del tipo di approccio da seguire, verranno utilizzate tecniche di regressione (Remaining Useful Life, Number of failures), classificazione (Failure probability).

Anomaly Detection

Per ogni asset o per gruppi di asset è possibile stabilire se l'asset si stia comportando in continuità con il suo passato, o se stia affrontando periodi di particolare stress di utilizzo o se sia sottoposto a condizioni ambientali modificate.

Eventuali discontinuità nelle performance energetiche di un edificio (ad es. un rapporto alterato tra temperatura esterna, umidità, temperatura interna e impostazioni dell'impianto di riscaldamento) possono essere sintomo di un eventuale malfunzionamento degli isolamenti di un edificio. Come possiamo immaginare, individuare il rapporto tra un alto numero di grandezze necessita dell'applicazione di algoritmi di Machine Learning.

L'output degli algoritmi di Anomaly Detection consiste in allarmi sollevati verso gli Operatori di area SFE e/o figure preposte allo scopo.

Obiettivi

Si possono impostare diversi algoritmi di Machine Learning, per monitorare diversi fenomeni all'interno e all'esterno degli edifici monitorati.

Gli algoritmi possono essere basati sulla continuità dei segnali, o sul rapporto che intercorre tra una molteplicità di segnali. Questi algoritmi sono addestrati appositamente per riconoscere una performance energetica “accettabile”, e segnalare in modo automatico eventuali performance energetiche che non lo sono.

Per gli scopi specifici del progetto di COTRAL Carbon Neutrality, si identificano alcuni casi di interesse:

- Monitoraggio degli “spike” ovvero di situazioni anomale di consumi (che possono essere causati ad esempio da guasti a qualche componente idrico che determina perdite d'acqua)
- Sistemi di “raccomandazione” legati alla misurazione di “distanza” tra un modello ideale di “comfort” inteso come combinazione di temperatura, qualità dell'aria etc. rispetto alle rilevazioni reali, incrociate con l'occupazione reale dei locali e/o la rilevazione dei parametri di monitoraggio
- Ottimizzazione comfort vs risparmio energetico. Variante del caso precedente dove grazie alla definizione di obiettivi di comfort è possibile produrre in tempo “quasi-reale” le azioni di modifica e micro regolazione agli impianti di edificio (illuminazione, circolazione aria, riscaldamento/raffrescamento, etc.)

Dati necessari

Tipicamente questi algoritmi necessitano di dati di telemetria acquisiti dai sensori IoT. Eventualmente, questi dati possono essere integrati da dati relativi alle impostazioni e/o eventuali gestioni degli asset.

Tecniche di modellazione

Le tecniche di modellazione possono essere diverse. Tra i metodi più utilizzati possiamo annoverare:

- Tecniche di clustering;
- Analisi delle serie storiche;
- Regressioni (specialmente Support Vector Machine).

Machine Vision

Tecniche di Machine Vision consentono a sistemi di visione di riconoscere eventuali elementi (ad es. forme, oggetti, persone) all'interno degli ambienti dell'edificio monitorato.

Obiettivi

Un'applicazione tipica di machine vision nell'ambito di gestione energetica di edifici è il monitoraggio di dispersioni di calore su pareti, siano esse interne o esterne. Il riconoscimento di tali dispersioni può azionare un sistema di allarme, andando anche eventualmente a fare una valutazione preliminare dell'entità della dispersione.

Dati necessari

L'addestramento di modelli di riconoscimento di immagini richiede, appunto, immagini acquisite tramite apposite telecamere.

Tecniche di modellazione

Le tecniche di modellazione che garantiscono risultati migliori sono reti neurali, in diverse implementazioni: Recurrent Neural Network (RNN), Convolutional Neural Network (CNN).

Si noti come l'applicazione di queste tecniche trovi particolari vantaggi computazionali dalla disponibilità di GPU.

Analisi avanzata dei dati

Il caso più generale di applicazione di Data Science è l'analisi avanzata dei dati. Lo scopo di queste analisi (diversamente dai casi precedenti) non è distribuire un servizio automatico, ma è quello di realizzare reportistica che coinvolge trasformazioni dei dati non ottenibili tramite strumenti classici di business intelligence (fogli di calcolo o altre tipologie di supporto informatico).

Queste analisi si concentrano soprattutto nell'individuazione di cause che hanno portato al verificarsi di certi fenomeni, piuttosto che nella formulazione di previsioni sul futuro.

Obiettivi

L'obiettivo è condividere una serie di analisi avanzate con particolari stakeholder. Tipico esempio: clustering dei dati, decomposizione di serie storiche, analisi asimmetriche (ad es. Analisi delle Componenti Principali).

Dati necessari

Queste analisi possono essere applicate su qualsiasi tipologia di dati.

Tecniche di modellazione

Le tecniche di modellazione sono multiple e dipendono dal contesto.

In questo caso, l'elemento più importante riguarda le tecnologie che si utilizzano per la presentazione dei risultati e l'eventuale ripetibilità delle analisi. Dunque, si raccomanda l'utilizzo di sistemi quali Jupyter notebook.