Sistemi di supporto alle decisioni

M9Museum Project

Svolto da: Achilli Mattia, Rettaroli Andrea.

Obbiettivo

L'obbiettivo di questo progetto è quello di riuscire a fare delle previsioni sulle serie storiche dell'M9Museum di Venezia. Le serie storiche che compongono il dataset monitorano l'affluenza al museo nelle varie sale. Tramite modelli statistici e modelli neurali si vuole prevedere l'afflusso di visitatori al museo.

Analisi

Analisi del problema

Il problema in se ci chide di prevedere l'afflusso all' M9Museum partendo dai dati raccolti sull'afflusso al museo. Una volta analizzati i dati va determinata la serie storica che essi costituiscono al fine di applicare modelli statiscici e neurali per fare forecasting. Da subito è chiaro che lavorando con dati reali si potrebbe incombere in problematiche legate ad essi come:

- Inconsistenza dei dati;
- Errori di misurazioni;
- Eventuali dati mancanti

Inoltre da subito risulta chiaro che la mole di dati in questione è molto vasta, l'elaborazione degli stessi è costosa in termini di tempo, sarà determinante lavorare sull'ottimizzazione dei processi di lettura e in seguito di addestramento.

Analisi dei dati

Al fine di riuscire ad applicare i modelli statistici e i modelli neurali, risulta di cruciale importanza analizzare e comprendere i dati in nostro possesso. In seguito si riporta la struttura del file CSV.

timestamp,"area_code","totals","alarms","date"
1632669666838,"floor_3","0","0","2021-09-26"
1632669667149,"front_desk","0","0","2021-09-26"
1632669667226,"floor_1","17","0","2021-09-26"
1632669667600,"floor_2","9","0","2021-09-26"

Dal CSV si individuano i dati timestamp, area_code e totals come i dati di rilievo su cui si costruisce la serie storica del numero di presenti nelle varie sale. Il timestamp indica il momento temporale in cui il dato è stato acquisito. il seguente esempio ci aiuta a capire il formato:

Epoch timestamp: 1640692068
Timestamp in milliseconds: 1640692068000
Date and time (GMT): Tuesday 28 December 2021 11:47:48
Date and time (your time zone): martedì 28 dicembre 2021 12:47:48 GMT+01:00

A seguito di questa conversione si nota che le misurazioni avvengono ogni circa 2secondi-1minuto a seconda del piano. Si decide che per costruire la serie si prendono i riferimenti definendo un lasso temporale sensato. Il grafico seguente mostra la serie costruita dall'intero dataset.

AGGIUNGERE IMMAGINE GENERALE.PNG

In seguito si vede il grafico utilizzando i riferimenti ogni 10minuti.

AGGIUNGERE IMMAGINE 3florCorrect-ordered-totals.png

Dal secondo grafico è facilmente rilevabile che all'interno della settimana vi sono dei giorni in cui l'afflusso è 0, ci siamo interfacciati con il Sig.re Luca Agatensi, coordinatore del progetto, per capire se questi dati fossero dovuti a delle chiusure causate dal Covid19 o a degli errori di misurazione o ad altro. In quei giorni il Museo non è aperto al pubblico, ciò prova la correttezza dei dati letti. La serie risulta continua, non si rileva la presenza di outlier.

Ci è stato esplicitamente richiesto di ignorare il front_desk in quanto relativo al personale.

In questa fase si leggono i dati da file, essendo 11.4M i record si pensa che sia meglio passare tramite DB e ottenerli già filtrati per sala al fine di ottimizzare i tempi.

Si ragiona anche alla possibilità di eseguire i metodi neurali su Colab al fine di migliorare le prestazioni soprattutto in termini di tempi di addestramento; risultano però delle complicazioni legate alla memoria delle macchine gratuitamente disponibili e al caricamento dei dati sulle stesse.

Scelte implementative

Una volta letti e compresi i dati si decide di ottimizzare il processo di lettura al fine di migliorare le prestazioni; a tale scopo si decide di utilizzare un database, nello specifico MongoDB che permette l'import di dati da CSV in maniera molto semplice. Si è deciso di creare un file config.py che non viene pushato nel repository ma viene tenuto in locale con la stringa di connessione. Si decide di ottimizzare anche le letture filtrando già i dati per piano in modo da lavorare su un singolo piano alla volta, così facendo i tempi necessari a questa fase si sono ridotti drasticamente. Successivamente si decide di rimuovere anche le colonne non utili dal dataset per alleggerire le esecuzioni. I dati vengono raccolti in intervalli differenti di tempo di conseguenza per ogni piano si sceglie di considerare una misurazione ogni tot tempo nello specifico:

- floor 1: misurazione ogni 2 secondi, intervallo scelto 10 minuti.
- floor 2: misurazione ogni 1 minuto, intervallo scelto 30 minuti.
- floor 3: misurazione ogni x , intervallo scelto x minuti.

Successivamente è deciso di fare preprocessing per andare a vedere trend, stagionalità, autocorrelazione, media, varianza e stazionarietà. Solo una volta determinati questi aspetti caratteristici della serie avremmo deciso con che algoritmi approcciare; si è però scelto di provare metodi statistici e metodi neurali in modo da confrontare i risutalti. Si decide di utilizzare Git, con Git flow per il verisonamento del codice.

Preprocessing

Questa è una delle fasi principali e più importanti per l'elaborazione dati. In questa fase si determinano le caratteristiche della serie che permettono di individuare i metodi migliori per andare a fare forecasting su di esse. In particolare ci interessa individuare eventuali componenti di trend e di stagionalità nella serie. Si applica il seasonal decompose per verificare la presenza di queste componenti. La figura seguente mostra un esempio di seasonal decompose nella serie legata al floor 1.

E' evidente che non vi sono componenti di stagionalità, vi è solo una componente di trend.

In seguito scegliamo di calcolare media e varianza e verificare che queste siano pressochè uguali. La figura mostra un esempio dei risultati ottenuti:

Aggiungere figura media e varianza

Si eseguono anche due test statistici al fine di determinare se la serie è stazionaria. Il test ADF Dickey-Fuller aumentato i cui risultati sono visibili nella figura seguente.

figura screen adf

In seguito viene effettuato anche il test statistico kpss, che prende il nome dai sui creatori Kwiatkowski, Phillips, Schmidt e Shin, il cui risultato viene mostrato nell'immagine seguente.

figura screen kpss

Ricordiamo che una serie è stazionaria se:

- Its mean is constant over time (stationarity in means);
- Its variance is constant over time (stationarity in variance);
- The ratio between values separated by k time periods depends only on k, not on time (stationarity in covariance).

Viene anche determinata l'autocorrelazione, l'immagine seguente mostra l'autocorrelazione presente nella serie storica del primo piano.

Immagine autocorrelazione

Determinare questi aspetti è essenziale al fine di capire se è necessario andare a fare delle operazioni alla serie prima di applicare i modelli. Una volta valutati i risultati si sceglie di non effettuare modifiche alla serie con le apposite tecniche. Inoltre, la serie non mostra outlier. Il fatto che la serie non sia stagionale ci porta a scegliere ARIMA come metodo statistico da applicare.