Università degli Studi di Milano-Bicocca

Data Science Lab Final Project

Marco Ferrario - 795203

Giorgio Ottolina - 839017

Contents

Dataset - Introduction

- Il dataset iniziale contiene informazioni sui visitatori del parco acquatico Nettebad in Germania.
- In aggiunta, si sono presi in esame due ulteriori dataset contenenti informazioni relative al meteo nella zona interessata.
- Ci si è prefissati come obiettivo quello di prevedere il numero di visitatori al parco acquatico utilizzando modelli di Machine Learning di regressione e classificazione.

Dataset – Features Description

I 3 Datasets

1 Dataset Visitatori Nettebad

- Periodo: Marzo 2005 – Dicembre 2012
- Numero Visitatori per giorno
- Prezzi
- Festività
- Apertura dello scivolo ad alta velocità (sloop)
- Chiusure per eventi/ispezioni

2 Dataset Postazione Meteo 1

- Periodo: Gennaio 2005 – Dicembre 2012
- Temperatura massima, media, minima
- Umidità dell'aria
- Precipitazioni
- Quantità di neve
- Eliofania
- Velocità massima del vento

3 Dataset Postazione Meteo 2

- Periodo: Giugno 2009 – Dicembre 2012
- Umidità e pressione dell'aria
- Intensità delle radiazioni solari
- Temperatura
- Direzione ed intensità media e massima del vento

Contents

Preprocessing

Passaggi

Analisi Correlazione Variabili

Per poter realizzare I modelli, si è scelto di non utilizzare tutti i dati a disposizione, eliminando le ridondanze e le caratteristiche meteo «poco informative» ai fini della previsione.

Trasformazione Feature Data

Le date sono state suddivise in “anno”, “mese” e “giorno” in modo tale da poter tener traccia del giorno settimanale (feriale o festivo) e del mese/periodo dell’anno (esempio: stagioni).

Range Temporale

Il periodo temporale in comune ai tre dataset è quello che va da Giugno 2009 a Dicembre 2012, quindi l’insieme di osservazioni risultanti al termine del preprocessing abbraccia questo range.

Preprocessing

Passaggi

Gestione Valori Missing

I records contenenti valori mancanti si registrano esclusivamente nel dataset della postazione meteo DWD.

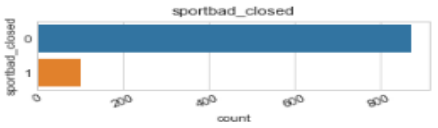
In particolar modo, ci si è accorti che alcune features ne presentavano una grande quantità: quindi la decisione è risultata essere quella di rimuovere le colonne interessate.

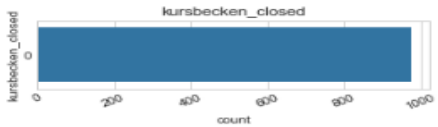
Aggregazione Dati Rimanenti

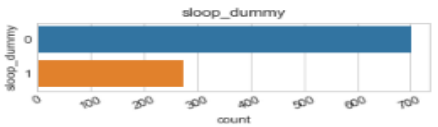
A questo punto si è creato un dataset complessivo dove ogni record rappresenta una giornata e le relative informazioni.

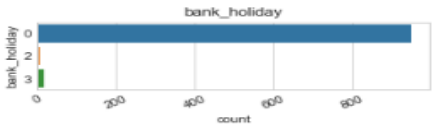
Per I dati meteo, le colonne in comune sono state imputate utilizzando la media (es: temperatura).

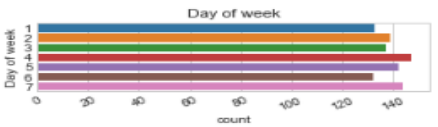
Preprocessing – Variables Distribution

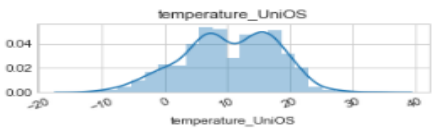


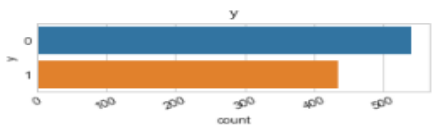


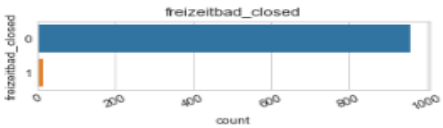


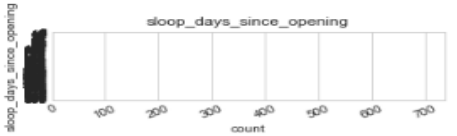


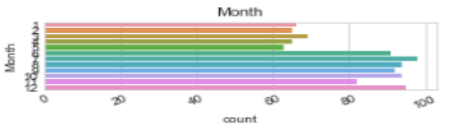


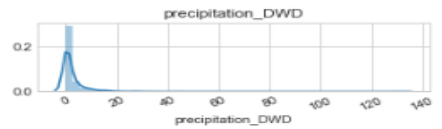


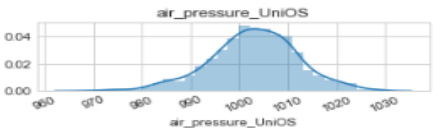


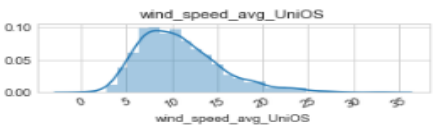


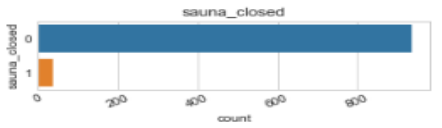


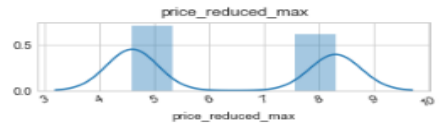


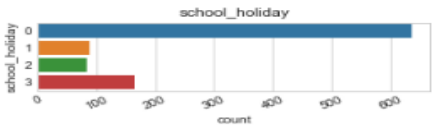


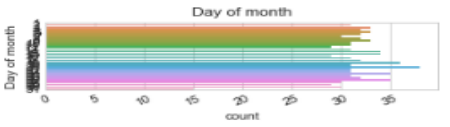


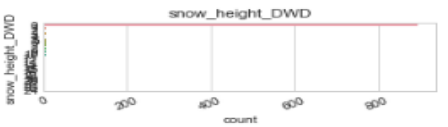


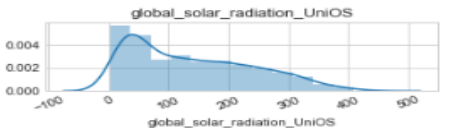


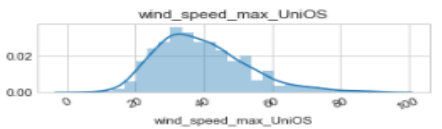












Contents

Linear regression

- Come osservato in precedenza, il dataset presenta una grande quantità di features anche dopo la prima scrematura.

- Per individuare quelle più significative, è stata effettuata un'analisi tramite regressione lineare, con un focus particolare sulle variabili esplicative.

- Dopo una prima regressione tra numero visitatori e la totalità delle rimanenti variabili, da un'iniziale analisi dei risultati ottenuti si è notato come non tutte le variabili considerate siano significative. Per la scelta delle features ci si è serviti di una funzione basata sull'**AIC** (Akaike's information criterion), la quale fa una cernita delle variabili che contribuiscono a dare maggiore significatività al modello.

Linear regression - Results

- Le features rimosse sono principalmente quelle che riguardano il prezzo dell'ingresso ed alcune relative ai dati meteo, come l'intensità del vento e l'umidità dell'aria.

- Risultano invece rilevanti la temperatura, l'apertura/chiusura di una particolare attrazione e se il giorno in questione è di vacanza scolastica.

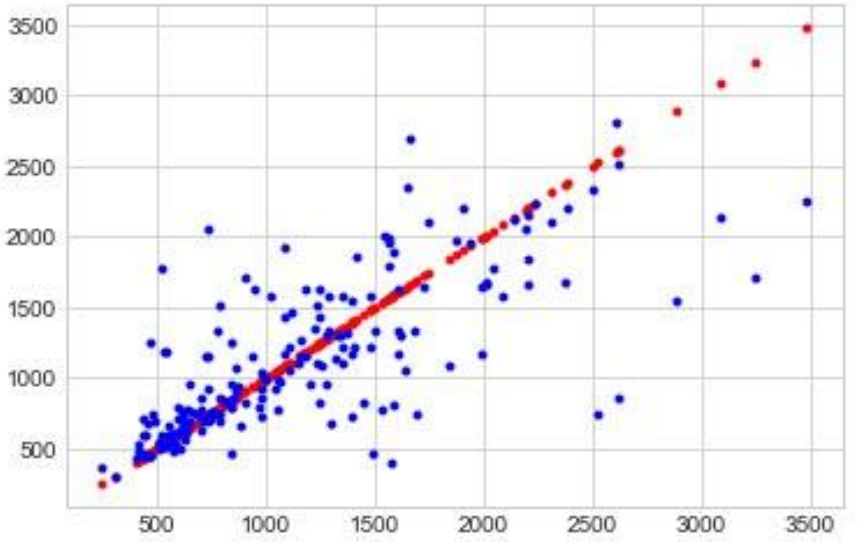
- Utilizzando quindi le variabili risultate maggiormente significative (16 delle 23), con il modello di regressione lineare si è ottenuto un **errore quadratico medio** pari a 196583.

Other Regression Models

- Come ulteriori modelli di confronto si è passati ad utilizzare **Support Vector Regression** e **Random Forest Regression.**

- Gli score migliori sono stati ottenuti con Random Forest (**MSE** = 185639 - grafico a lato), mentre con Support Vector i risultati rilevati si sono dimostrati più bassi:

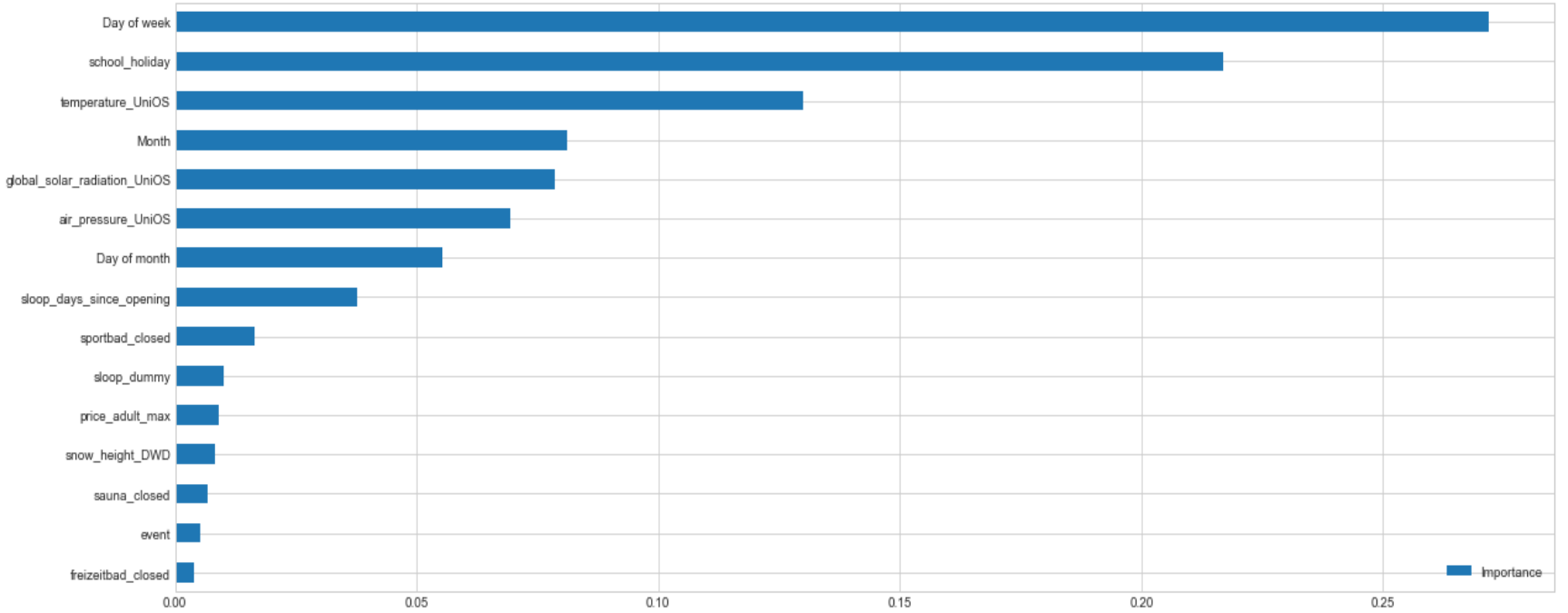
  (**MSE** = 384081).



Contents

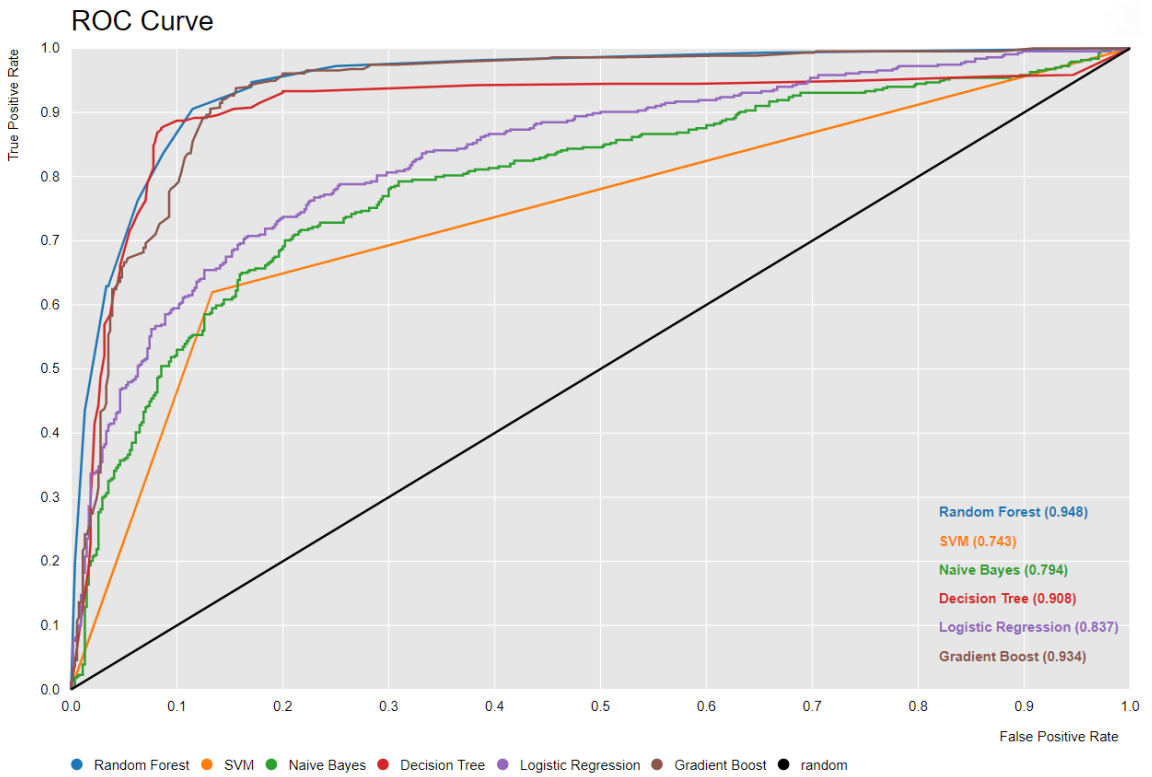
Classification Models – Preliminary Steps

- Innanzitutto, la feature target “visitors_pool_total” è stata trasformata in una variabile binaria per impostare il problema di classificazione: nel caso i visitatori fossero 1000 o più è stato assegnato il valore 1, altrimenti 0.
- Un’ulteriore analisi tramite **Random Forest Classifier**, di cui viene mostrata qui sotto una visualizzazione grafica, ha confermato le deduzioni precedenti riguardo la feature importance delle variabili a nostra disposizione:



Classification Models – ROC Curves and Accuracy Scores

Di seguito il confronto tra le curve ROC e i risultati relativi all'accuracy: anche stavolta **Boosting, Random Forest** e **Decision Tree** sono stati i modelli migliori. E' stata anche implementata una **Neural Network** con TensorFlow/Keras, la quale ha ottenuto un'**accuracy score** pari a 82.55%.



Model	Score
Gradient Boosting Trees	90.37
Random Forest	87.93
Decision Tree	87.29
Logistic Regression	77.66
Naive Bayes	73.94
SVM	71.00
Neural Network (Keras)	82.55

Contents

Evaluation and Conclusion

Considerazioni Finali

1 Risultati ottenuti

L'algoritmo di classificazione che ha offerto le prestazioni migliori è stato il gradient boosting (90 % di accuratezza), seguito dal Random forest.

Il modello in questione può essere utilizzato per individuare le condizioni che hanno determinato maggiormente l'affluenza al parco acquatico e prevedere in quali giornate applicare sconti ai visitatori o offrire un certo numero di servizi / eventi apprezzati dal pubblico.

2 Possibili sviluppi e miglioramenti

I modelli di regressione potrebbero essere migliorati per offrire previsioni più accurate, inoltre,essendo ogni riga del dataset la rappresentazione di una singola giornata, sarebbe possibile effettuare un'analisi basata sullo studio di serie temporali.