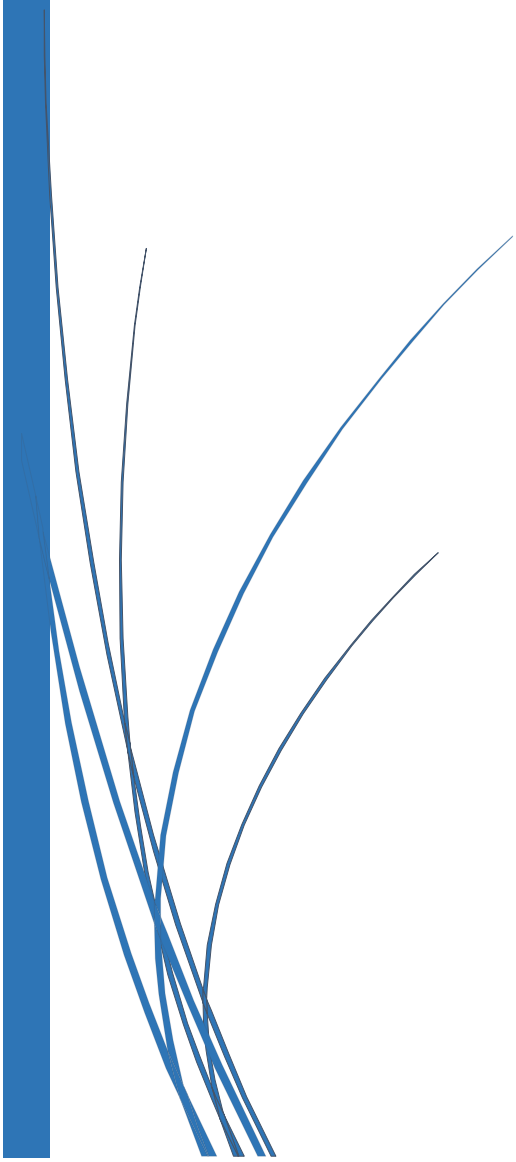
6/12/2022

#Datathon 2022

Time Series Forecasting



Συντελεστές Εργασίας:

ΚΑΡΚΑΝΗΣ ΕΥΣΤΡΑΤΙΟΣ – Π19064

ΧΡΙΣΤΟΦΟΡΙΔΗΣ ΧΑΡΑΛΑΜΠΟΣ – Π19188

1. Εισαγωγή

Στο συγκεκριμένο πρόβλημα μηχανικής μάθησης, καλούμαστε να διαχειριστούμε ένα πρόβλημα χρονοσειράς. Πιο συγκεκριμένα, δοθέντος δεδομένων κατανάλωσης ηλεκτρικής ενέργειας (μέσω έξυπνων μετρητών μέτρησης κατανάλωσης), ως data scientists πρέπει να προβλέψουμε ποια θα είναι η κατανάλωση του ηλεκτρικού ρεύματος σε επόμενα χρονικά διαστήματα, δηλαδή σε χρόνους για τους οποίους δεν έχουμε δεδομένα.

2. Το σύνολο δεδομένων

Το σύνολο δεδομένων, το οποίο χρησιμοποιήσαμε, αποτελείται από μετρήσεις κατανάλωσης ηλεκτρικής ενέργειας 370 πελατών. Κάθε μέτρηση καθενός από τους 370 πελάτες καταγράφεται ανά 15 λεπτά. Το σύνολο δεδομένων δεν έχει ελλιπείς τιμές (κενά) και ουσιαστικά αποτελείται από 371 στήλες.

Dataset's Link: https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014

Σύμφωνα με τα παραπάνω, το πρόβλημα που διατυπώθηκε ανάγεται σε πρόβλημα χρονοσειράς. Ουσιαστικά, οι τιμές κατανάλωσης που καταγράφονται είναι σε σχέση με τον χρόνο. Σε ένα πρόβλημα χρονοσειράς, ο χρόνος θεωρείται βασικό στοιχείο και αποτελεί την ανεξάρτητη μεταβλητή των δεδομένων μας.

3. Προεπεξεργασία των δεδομένων

Στη συγκεκριμένη περίπτωση ακολουθήσαμε τα ακόλουθα βήματα:

- Μετατρέψαμε τους τύπους δεδομένων των μετρήσεων κατανάλωσης ηλεκτρικής ενέργειας κάθε πελάτη σε δεκαδικό αριθμό (float)
- Μετατρέψαμε το πεδίο της ημερομηνίας σε DateTime format.
- Διαιρέσαμε κάθε τιμή μέτρησης με το 4, ώστε να πάρουμε την ωριαία κατανάλωση ηλεκτρικής ενέργειας (kWh)
- Επίσης, θεωρήσαμε καλύτερο να έχουμε μία στήλη στο σύνολο δεδομένων, η οποία θα περιλαμβάνει το άθροισμα κατανάλωσης ηλεκτρικής ενέργειας

όλων των πελατών, ανά μονάδα χρόνου. Επομένως, όλες τις στήλες που αντιπροσωπεύουν πελάτες τις συμπτίξαμε σε μία στήλη με όνομα «Consumption_Sum».

Μετά από την προεπεξεργασία, το dataset μεταμορφώθηκε, όπως φαίνεται στην παρακάτω εικόνα:

	Consumption_Sum
Timestamps	
2011-01-01 00:15:00	17128.278835
2011-01-01 00:30:00	17295.076090
2011-01-01 00:45:00	17341.212643
2011-01-01 01:00:00	17087.620165
2011-01-01 01:15:00	16541.718576

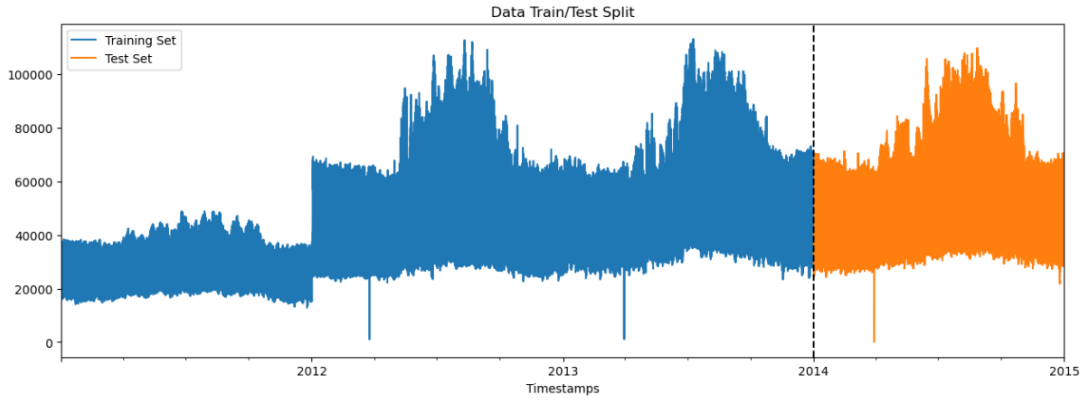
4. Οπτικοποίηση των δεδομένων

Με την οπτικοποίηση μπορούμε να διαβάσουμε ευκολότερα και γρηγορότερα τα δεδομένα μας. Οι παρατηρήσεις μας είναι οι ακόλουθες:

- Κατά την θερινή περίοδο, η αθροιστική κατανάλωση ενέργειας είναι μεγαλύτερη, ενώ κατά την χειμερινή περίοδο, η αθροιστική κατανάλωση ενέργειας είναι μικρότερη.
- Το χρονικό διάστημα 2011-2012, η αθροιστική κατανάλωση ενέργειας είναι μικρότερη, καθώς οι περισσότεροι από τους 370 πελάτες δεν κατανάλωναν ενέργεια ή δεν συμμετείχαν στον συγκεκριμένο πάροχο ηλεκτρικής ενέργειας.

5. Δημιουργία του μοντέλου πρόβλεψης

Για την δημιουργία του μηχανισμού πρόβλεψης χρονοσειράς, χρησιμοποιήσαμε μία τεχνική regression. Αφού χωρίσαμε τα δεδομένα μας σε train και test σύνολα, όπως φαίνεται στην παρακάτω εικόνα:

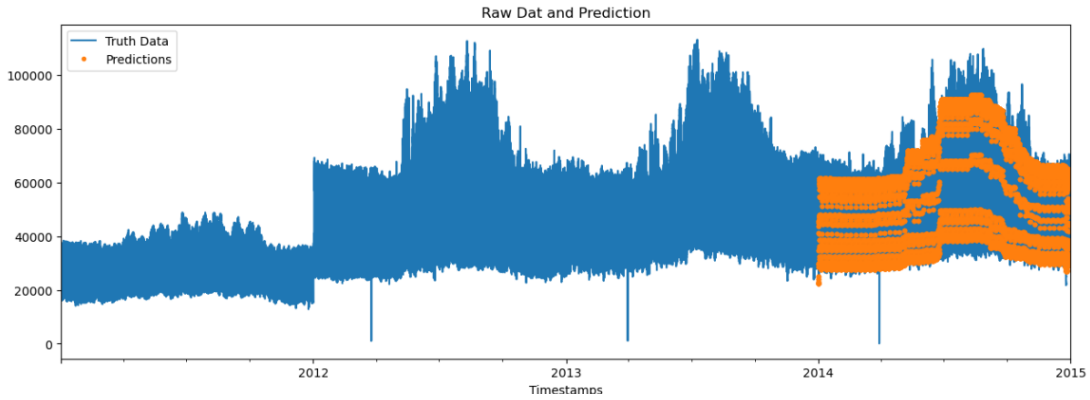


Για τα δεδομένα εκπαίδευσης, χρησιμοποιήσαμε όλα τα δεδομένα κατανάλωσης από το 2011 έως και το 2013. Για τα δεδομένα ελέγχου χρησιμοποιήσαμε τα υπόλοιπα δεδομένα.

αναλύσαμε την χρονοσειρά σε περισσότερα συστατικά (Χρόνος, Ημέρα της εβδομάδας, ώρα, ημέρα του χρόνου κτλ.).

Χρησιμοποιώντας ως είσοδο στο μοντέλο μας (Features) την ημέρα του χρόνου, την ώρα, την ημέρα της εβδομάδας και τον χρόνο, προσπαθήσαμε να προβλέψουμε (Target) την συνολική κατανάλωση ηλεκτρικής ενέργειας (Consumption_Sum).

Τέλος, σε ένα τελικό γράφημα αποτυπώνονται οι προβλέψεις του μοντέλου μας σε σχέση με τις πραγματικές τιμές. Οι πορτοκαλί τιμές είναι οι προβλέψεις που έγιναν και βρίσκονται ακριβώς επάνω από τις πραγματικές (μπλε) τιμές, ώστε να γίνεται ευκολότερα η σύγκριση μεταξύ τους.



Σύμφωνα με το παραπάνω γράφημα, το μοντέλο μηχανικής μάθησης που διαλέξαμε δεν είναι και τόσο καλό στην διαδικασία πρόβλεψη ή θα μπορούσε να συμπεριφέρεται καλύτερα. Ένας τρόπος για να γίνει αυτό είναι είτε να επιλεγεί ένα άλλο μοντέλο, ή να γίνει διαφορετικό κούρδισμα παραμέτρων του συγκεκριμένου αλγορίθμου.