Προβλέψεις Χρονοσειρών χρησιμοποιώντας το Prophet του Facebook

Δρ. Σωτήριος Δ. Νικολόπουλος

Big Data & Analytics

Πανεπιστήμιο Πελοποννήσου

Τμήμα Λογιστικής & Χρηματοοικονομικών

s.nikolopoulos@go.uop.gr

Προβλέψεις Χρονοσειρών χρησιμοποιώντας το Prophet του Facebook

Προβλέψεις Χρονοσειρών χρησιμοποιώντας το Prophet του Facebook

Προβλέψεις Χρονοσειρών χρησιμοποιώντας το Prophet του Facebook

Η κατανόηση των χρονικών προτύπων είναι κρίσιμη για οποιαδήποτε επιχείρηση.

Ερωτήσεις όπως πόσο απόθεμα να διατηρήσετε, πόση επισκεψιμότητα αναμένετε στο κατάστημά σας έως πόσοι άνθρωποι θα ταξιδέψουν με μια αεροπορική εταιρεία - όλα αυτά είναι σημαντικά προβλήματα χρονοσειρών που πρέπει να επιλυθούν.

Αυτός είναι ο λόγος για τον οποίο η πρόβλεψη χρονοσειρών είναι μία από τις τεχνικές που πρέπει να γνωρίζει κάθε επιστήμονας δεδομένων.

Προβλέψεις Χρονοσειρών χρησιμοποιώντας το Prophet του Facebook

Το Prophet είναι μια βιβλιοθήκη ανοιχτού κώδικα που δημοσιεύτηκε από το Facebook και βασίζεται σε αποσυνθέσιμα (decomposable) μοντέλα (τάση + εποχικότητα + αργίες).

Μας παρέχει τη δυνατότητα να κάνουμε προβλέψεις χρονοσειρών με καλή ακρίβεια χρησιμοποιώντας απλές και εύκολα προσαρμόσιμες παραμέτρους και υποστηρίζει τη συμπερίληψη της επίδρασης της προσαρμοσμένης εποχικότητας και των αργιών!

Σε αυτό το μάθημα, θα αποκτήσουμε το βασικό υπόβαθρο για το πώς το Prophet καλύπτει τα υπάρχοντα κενά στη δημιουργία γρήγορων και αξιόπιστων προβλέψεων.

Προβλέψεις Χρονοσειρών χρησιμοποιώντας το Prophet του Facebook

Το Prophet είναι μια βιβλιοθήκη ανοιχτού κώδικα που δημοσιεύτηκε από το Facebook και βασίζεται σε αποσυνθέσιμα (decomposable) μοντέλα (τάση + εποχικότητα + αργίες).

Μας παρέχει τη δυνατότητα να κάνουμε προβλέψεις χρονοσειρών με καλή ακρίβεια χρησιμοποιώντας απλές και εύκολα προσαρμόσιμες παραμέτρους και υποστηρίζει τη συμπερίληψη της επίδρασης της προσαρμοσμένης εποχικότητας και των αργιών!

Σε αυτό το μάθημα, θα αποκτήσουμε το βασικό υπόβαθρο για το πώς το Prophet καλύπτει τα υπάρχοντα κενά στη δημιουργία γρήγορων και αξιόπιστων προβλέψεων.

Τι νέο υπάρχει στο Prophet;

Όταν ένα μοντέλο πρόβλεψης δεν λειτουργεί ικανοποιητικά, θα πρέπει να ρυθμίσουμε τις παραμέτρους του σε σχέση με το συγκεκριμένο πρόβλημα που αντιμετωπίζουμε.

Η ρύθμιση αυτών των μεθόδων απαιτεί μια βαθιά κατανόηση του πώς λειτουργούν τα υποκείμενα μοντέλα χρονοσειρών.

Για παράδειγμα, σε μη εποχιακό μοντέλο ARIMA (Αυτοπαλίνδρομο μοντέλο κινητού μέσου) θα πρέπει να προσδιοριστούν ο αριθμός των διαφοροποιήσεων και οι μέγιστες τάξεις τόσο για το αυτοπαλίνδρομο μέρος. όσο και τον κινητό μέσο όρο του μοντέλου.

ARIMA είναι ένα αρκτικόλεξο του AutoRegressive Integrated Moving Average (Αυτο-παλίνδρομος Ολοκληρωμένος Κινητός Μέσος)

Τι νέο υπάρχει στο Prophet;

Ειδικές περιπτώσεις ARIMA:

Μοντέλο	**ARIMA**
Λευκός θόρυβος	ARIMA(0,0,0) χωρίς σταθερά
Τυχαίος περίπατος	ARIMA(0,1,0) χωρίς σταθερά
Τυχαίος περίπατος με περιπλάνηση	ARIMA(0,1,0) με σταθερά
Αυτοπλαινδρόμηση	ARIMA(p,0,0)
Κινητός μέσος	ARIMA(0,0,q)

Πίνακας: Διαφορετικά μοντέλα ARIMA

Ο προσδιορισμός των σωστών παραμέτρων σε ένα μοντέλο ARIMA δεν είναι μια εύκολη διαδικασία.

Το Μοντέλο Πρόβλεψης Prophet

Χρησιμοποιεί ένα αποσυνθέσιμο μοντέλο χρονοσειράς με τρία κύρια στοιχεία μοντέλου: τάση, εποχικότητα και αργίες. Αυτά συνδυάζονται στην παρακάτω εξίσωση:

$$y(t) = g(t) + s(t) + h(t) + \epsilon(t)$$

Όπου:

- $g(t)$: κατακερματισμένη γραμμική ή λογιστική καμπύλη για τη μοντελοποίηση μη περιοδικών αλλαγών στις χρονοσειρές
- $s(t)$: περιοδικές αλλαγές (π.χ. εβδομαδιαία/ετήσια εποχικότητα)
- $h(t)$: επιπτώσεις από αργίες (παρέχονται από τον χρήστη) με ακανόνιστη περιοδικότητα
- ε_t: όρος σφάλματος που λαμβάνει υπόψη τυχόν ασυνήθιστες αλλαγές που δεν καλύπτονται από το μοντέλο

Το Μοντέλο Πρόβλεψης Prophet

Χρησιμοποιώντας τον χρόνο ως μεταβλητή πρόβλεψης, το Prophet προσπαθεί να δημιουργήσει/προσαρμόσει διάφορες γραμμικές και μη γραμμικές συναρτήσεις.

Η μοντελοποίηση της εποχικότητας ως προσθετικό στοιχείο είναι η ίδια προσέγγιση που λαμβάνεται από την εκθετική εξομάλυνση στην τεχνική Holt-Winters.

Στην πραγματικότητα, διαμορφώνουμε το πρόβλημα πρόβλεψης ως άσκηση προσαρμογής μιας καμπύλης.

Το Μοντέλο Πρόβλεψης Prophet

Τάση

Η τάση μοντελοποιείται με την προσαρμογή μιας κατακερματισμένης γραμμικής καμπύλης πάνω στην τάση ή το μη περιοδικό μέρος της χρονοσειράς.

Η γραμμική προσαρμογή εξασφαλίζει ότι επηρεάζεται ελάχιστα από αιχμές ή ελλιπή δεδομένα.

Το Μοντέλο Πρόβλεψης Prophet

Κορεσμός Ανάπτυξης

Ένα σημαντικό ερώτημα που πρέπει να θέσουμε εδώ είναι - Αναμένουμε η μεταβλητή που θέλουμε να προβλέψουμε να συνεχίσει να αυξάνεται/μειώνεται για ολόκληρο το διάστημα πρόβλεψης;

Πολύ συχνά, υπάρχουν περιπτώσεις μη γραμμικής τάσης.

Ας υποθέσουμε ότι προσπαθούμε να προβλέψουμε τον αριθμό λήψεων μιας εφαρμογής κινητού τηλεφώνου σε μια περιοχή για τους επόμενους 12 μήνες.

Ο μέγιστος αριθμός λήψεων περιορίζεται πάντα από τον συνολικό αριθμό χρηστών smartphone στην περιοχή.

Ωστόσο, ο αριθμός των χρηστών smartphone θα αυξάνεται επίσης με την πάροδο του χρόνου.

Βήμα 1: Φόρτωση των απαραίτητων βιβλιοθηκών

Αρχικά, θα φορτώσουμε τα απαραίτητα πακέτα για αυτό το παράδειγμα. Για αυτό το απλό παράδειγμα, χρειαζόμαστε μόνο ένα πακέτο:

```
library(randomForest)
```

Βήμα 2: Προσαρμογή του μοντέλου τυχαίου δάσους

Για αυτό το παράδειγμα, θα χρησιμοποιήσουμε ένα ενσωματωμένο σύνολο δεδομένων R που ονομάζεται airquality, το οποίο περιέχει μετρήσεις ποιότητας αέρα στη Νέα Υόρκη σε 153 μεμονωμένες ημέρες.

```
#view structure of airquality dataset
str(airquality)
'data.frame':   153 obs. of  6 variables:
$ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
$ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
$ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
$ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
$ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
$ Day    : int  1 2 3 4 5 6 7 8 9 10 ...

#find number of rows with missing values
sum(!complete.cases(airquality))
[1] 42
```

Βήμα 2: Προσαρμογή του μοντέλου τυχαίου δάσους

Το σύνολο δεδομένων αυτό έχει 42 γραμμές με ελλείψεις, επομένως πριν προσαρμόσουμε ένα μοντέλο τυχαίου δάσους, θα συμπληρώσουμε τις ελλείψεις σε κάθε στήλη με τις διάμεσες τιμές κάθε στήλης.

```
#replace NAs with column medians
for(i in 1:ncol(airquality)) {
        airquality[ , i][is.na(airquality[ , i])] <- median(airquality[ , i], na.rm=TRUE)
}
```

Βήμα 2: Προσαρμογή του μοντέλου τυχαίου δάσους

Ο παρακάτω κώδικας δείχνει πώς να προσαρμόσετε ένα μοντέλο τυχαίου δάσους στην R χρησιμοποιώντας τη συνάρτηση randomForest() από το πακέτο randomForest.

```
#make this example reproducible
set.seed(1)

#fit the random forest model
model <- randomForest(
formula = Ozone ~ .,
data = airquality
)

#display fitted model
model
Call:
randomForest(formula = Ozone ~ ., data = airquality)
Type of random forest: regression
Number of trees: 500
No. of variables tried at each split: 1

Mean of squared residuals: 327.0914
% Var explained: 61
```

Βήμα 2: Προσαρμογή του μοντέλου τυχαίου δάσους

```
#find number of trees that produce lowest test MSE
which.min(model$mse)

[1] 82

#find RMSE of best model
sqrt(model$mse[which.min(model$mse)])

[1] 17.64392
```

Από την έξοδο μπορούμε να δούμε ότι το μοντέλο που παρήγαγε το χαμηλότερο μέσο τετραγωνικό σφάλμα (MSE) δοκιμής χρησιμοποίησε 82 δέντρα.

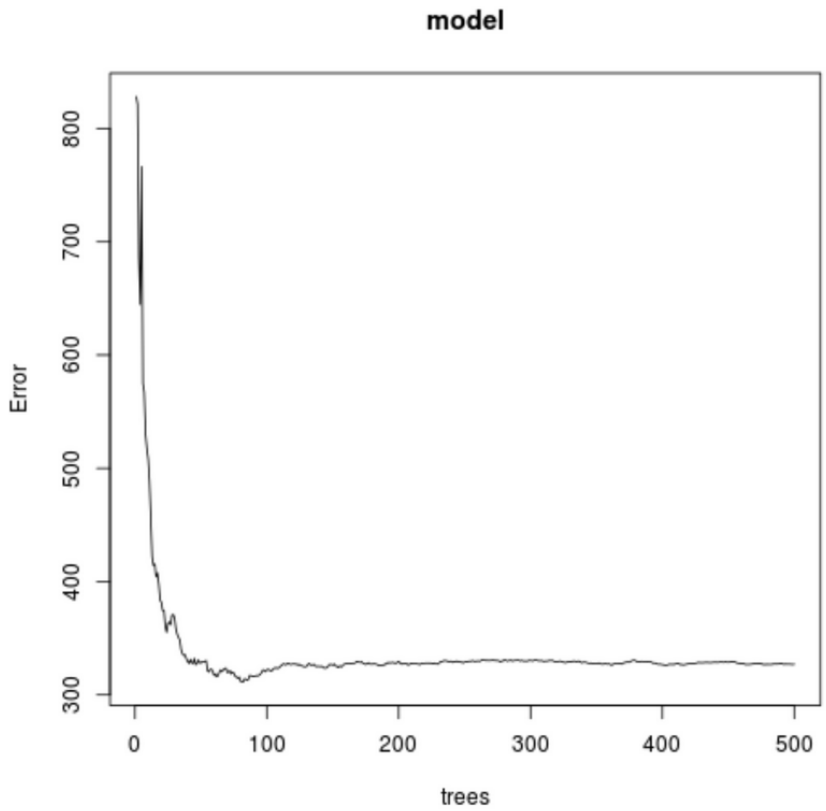
Μπορούμε επίσης να δούμε ότι το μέσο τετραγωνικό σφάλμα (RMSE) αυτού του μοντέλου ήταν 17.64392.

Μπορούμε να το σκεφτούμε ως τη μέση διαφορά μεταξύ της προβλεπόμενης τιμής για το Όζον και της πραγματικά παρατηρηθείσας τιμής.

Βήμα 2: Προσαρμογή του μοντέλου τυχαίου δάσους

Μπορούμε επίσης να χρησιμοποιήσουμε τον ακόλουθο κώδικα για να δημιουργήσουμε ένα γράφημα του μέσου τετραγωνικού σφάλματος δοκιμής (test MSE) βάσει του αριθμού των δέντρων που χρησιμοποιήθηκαν.

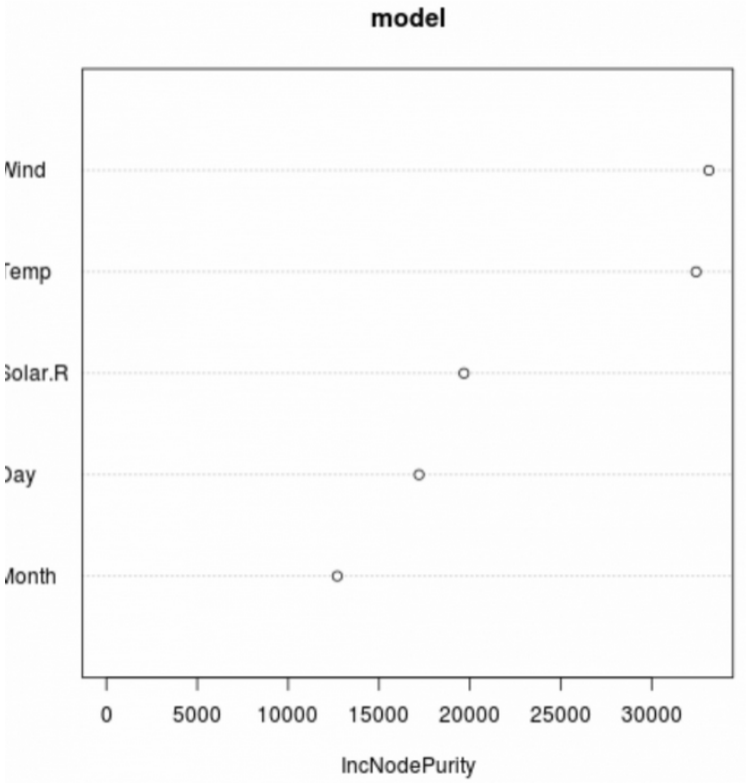
```
#plot the test MSE by number of trees
plot(model)
```



Βήμα 2: Προσαρμογή του μοντέλου τυχαίου δάσους

Επίσης, μπορούμε να χρησιμοποιήσουμε τη συνάρτηση varImpPlot() για να δημιουργήσουμε ένα γράφημα που απεικονίζει τη σημαντικότητα κάθε μεταβλητής πρόβλεψης στο τελικό μοντέλο.

```
#produce variable importance plot
varImpPlot(model)
```



Βήμα 2: Προσαρμογή του μοντέλου τυχαίου δάσους

Ο άξονας των x απεικονίζει τη μέση αύξηση στην καθαρότητα των κόμβων των δέντρων παλινδρόμησης με βάση τον διαχωρισμό με βάση τις διάφορες μεταβλητές πρόβλεψης που εμφανίζονται στον άξονα των y.

Από το γράφημα μπορούμε να δούμε ότι η μεταβλητή Wind είναι η πιο σημαντική μεταβλητή πρόβλεψης, ακολουθούμενη στενά από τη μεταβλητή Temp.

Βήμα 3: Συντονισμός του μοντέλου

Από προεπιλογή, η συνάρτηση **randomForest()** χρησιμοποιεί 500 δέντρα και (σύνολο προβλεπτικών μεταβλητών / 3) τυχαία επιλεγμένες προβλεπτικές μεταβλητές ως υποψήφιες για διαχωρισμό σε κάθε κόμβο.

Μπορούμε να προσαρμόσουμε αυτές τις παραμέτρους χρησιμοποιώντας τη συνάρτηση **tuneRF()**.

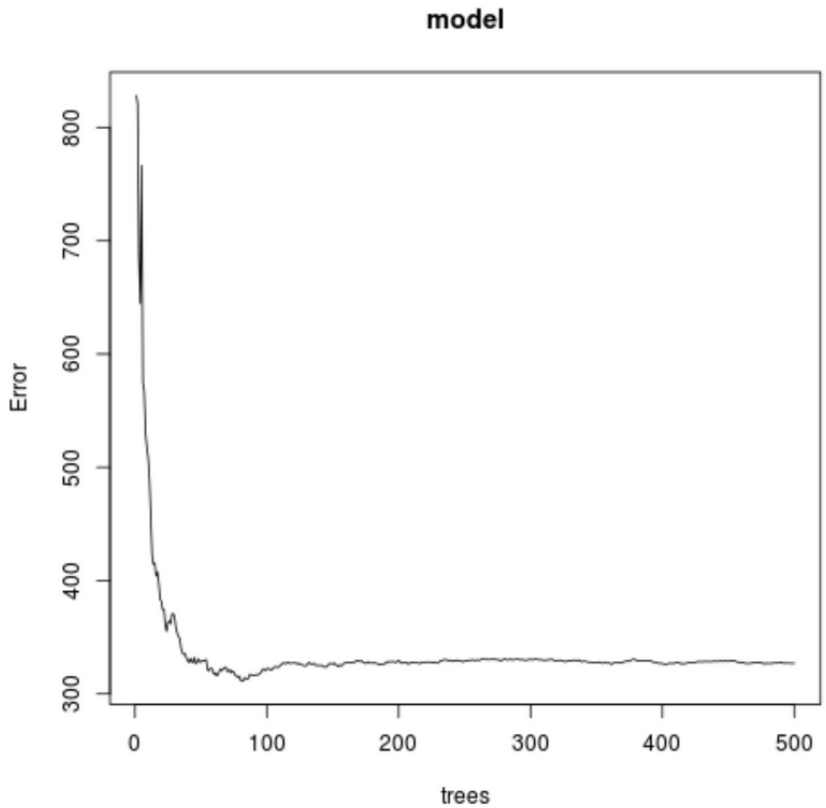
Ο παρακάτω κώδικας δείχνει πώς να βρούμε το βέλτιστο μοντέλο χρησιμοποιώντας τις ακόλουθες προδιαγραφές:

Βήμα 3: Συντονισμός του μοντέλου

ntreeTry (αριθμός δέντρων δοκιμής): Ο αριθμός των δέντρων που θα δημιουργηθούν για να βρεθεί το βέλτιστο μοντέλο.
mtryStart (αρχικός αριθμός μεταβλητών δοκιμής): Ο αρχικός αριθμός μεταβλητών πρόβλεψης που θα εξεταστούν ως υποψήφιες για διαχωρισμό σε κάθε κόμβο.
stepFactor (παράγοντας διαβάθμισης): Ο παράγοντας με τον οποίο θα αυξάνεται ο αριθμός των μεταβλητών πρόβλεψης που εξετάζονται σε κάθε διαδοχική δοκιμή.
improve (βελτίωση): Το ελάχιστο ποσοστό βελτίωσης του εκτός σάκου (out-of-bag) σφάλματος που απαιτείται για να συνεχιστεί η αύξηση του αριθμού των μεταβλητών πρόβλεψης που εξετάζονται.

Βήμα 3: Συντονισμός του μοντέλου

Η συνάρτηση tuneRF() παράγει το ακόλουθο γράφημα, το οποίο απεικονίζει τον αριθμό των μεταβλητών πρόβλεψης που χρησιμοποιούνται σε κάθε διαχωρισμό κατά τη δημιουργία των δέντρων στον άξονα των x και το εκτός δείγματος (out-of-bag) εκτιμώμενο σφάλμα στον άξονα των y.



Βήμα 3: Συντονισμός του μοντέλου

Παρατηρούμε ότι το χαμηλότερο σφάλμα εκτός δείγματος εκπαίδευσης, επιτυγχάνεται με τη χρήση 2 τυχαία επιλεγμένων προβλεπτικών μεταβλητών σε κάθε διαχωρισμό κατά τη δημιουργία των δέντρων.

Αυτό στην πραγματικότητα συμπίπτει με την προεπιλεγμένη παράμετρο (σύνολο προβλεπτικών μεταβλητών / 3 = 6/3 = 2) που χρησιμοποιείται από την αρχική συνάρτηση randomForest().

Βήμα 4: Χρήση του τελικού μοντέλου για Πρόβλεψη

Τέλος, μπορούμε να χρησιμοποιήσουμε το προσαρμοσμένο μοντέλο τυχαίου δάσους για να κάνουμε προβλέψεις σε νέες παρατηρήσεις.

```
#define new observation
new <- data.frame(Solar.R=150, Wind=8, Temp=70, Month=5, Day=5)

#use fitted bagged model to predict Ozone value of new observation
predict(model, newdata=new)
```

```
27.19442
```

Με βάση τις τιμές των μεταβλητών πρόβλεψης, το προσαρμοσμένο μοντέλο τυχαίου δάσους προβλέπει ότι η τιμή του Όζον θα είναι 27.19442 αυτή την συγκεκριμένη ημέρα.