Δένδρα αποφάσεων στη μηχανική μάθηση με χρήση της R

Δρ. Σωτήριος Δ. Νικολόπουλος

Big Data & Analytics

Πανεπιστήμιο Πελοποννήσου

Τμήμα Λογιστικής & Χρηματοοικονομικών

s.nikolopoulos@go.uop.gr

Δένδρα αποφάσεων στη μηχανική μάθηση με χρήση της R

1. Δένδρα αποφάσεων στη μηχανική μάθηση με χρήση της R
2. Πλεονεκτήματα και μειονεκτήματα των δέντρων απόφασης
3. Δόμηση μοντέλων δέντρων απόφασης βήμα προς βήμα στην R
4. Μοντελοποίηση δέντρου απλής απόφασης
5. Αξιολόγηση της απόδοσης του δέντρου αποφάσεων
6. Πρόβλεψη αποτελεσμάτων με χρήση του μοντέλου δέντρου αποφάσεων
7. Ερμηνεία του Δέντρου Αποφάσεων

Δένδρα αποφάσεων στη μηχανική μάθηση με χρήση της R

1. Δένδρα αποφάσεων στη μηχανική μάθηση με χρήση της R
2. Πλεονεκτήματα και μειονεκτήματα των δέντρων απόφασης
3. Δόμηση μοντέλων δέντρων απόφασης βήμα προς βήμα στην R
4. Μοντελοποίηση δέντρου απλής απόφασης
5. Αξιολόγηση της απόδοσης του δέντρου αποφάσεων
6. Πρόβλεψη αποτελεσμάτων με χρήση του μοντέλου δέντρου αποφάσεων
7. Ερμηνεία του Δέντρου Αποφάσεων

Δένδρα αποφάσεων στη μηχανική μάθηση με χρήση της R

Ένας ολοκληρωμένος οδηγός για τη δημιουργία, την οπτικοποίηση και την ερμηνεία μοντέλων δέντρων αποφάσεων με το R.

Δένδρα αποφάσεων στη μηχανική μάθηση με χρήση της R

Φανταστείτε τον εαυτό σας να περιηγείστε σε έναν λαβύρινθο.

Με κάθε βήμα, αντιμετωπίζεις μια απόφαση που σε οδηγεί πιο κοντά στην έξοδο ή πιο βαθιά στον λαβύρινθο.

Αυτό είναι παρόμοιο με έναν αλγόριθμο δέντρου αποφάσεων, μια ισχυρή και διαισθητική μέθοδο μηχανικής μάθησης που μας βοηθά να κατανοήσουμε πολύπλοκα δεδομένα και να επιλέξουμε την καλύτερη πορεία δράσης.

Δένδρα αποφάσεων στη μηχανική μάθηση με χρήση της R

Ένας αλγόριθμος δέντρου αποφάσεων αναλύει ένα σύνολο δεδομένων σε όλο και μικρότερα υποσύνολα με βάση ορισμένες συνθήκες.

Όπως ένα διακλαδούμενο δέντρο με φύλλα και κόμβους, ξεκινά με έναν μόνο κόμβο ρίζας και επεκτείνεται σε πολλαπλούς κλάδους, καθένας από τους οποίους αντιπροσωπεύει μια απόφαση που βασίζεται στην τιμή ενός χαρακτηριστικού.

Τα τελικά φύλλα του δέντρου είναι τα πιθανά αποτελέσματα ή προβλέψεις

Δένδρα αποφάσεων στη μηχανική μάθηση με χρήση της R

Το σημερινό μάθημα θα σας μυήσει στον κόσμο των δέντρων αποφάσεων που χρησιμοποιούν τη γλώσσα προγραμματισμού R.

Θα συζητήσουμε τα βασικά, θα εξετάσουμε σε δημοφιλείς τύπους αλγορίθμων δέντρων αποφάσεων, θα εξερευνήσουμε μεθόδους που βασίζονται σε δέντρα και θα υλοποιήσουμε ένα παράδειγμα βήμα προς βήμα.

Στο τέλος, θα μπορείτε να αξιοποιήσετε τη δύναμη των δέντρων αποφάσεων για να λάβετε καλύτερες αποφάσεις βάσει δεδομένων.

Τι είναι το Δέντρο Αποφάσεων στη Μηχανική Μάθηση;

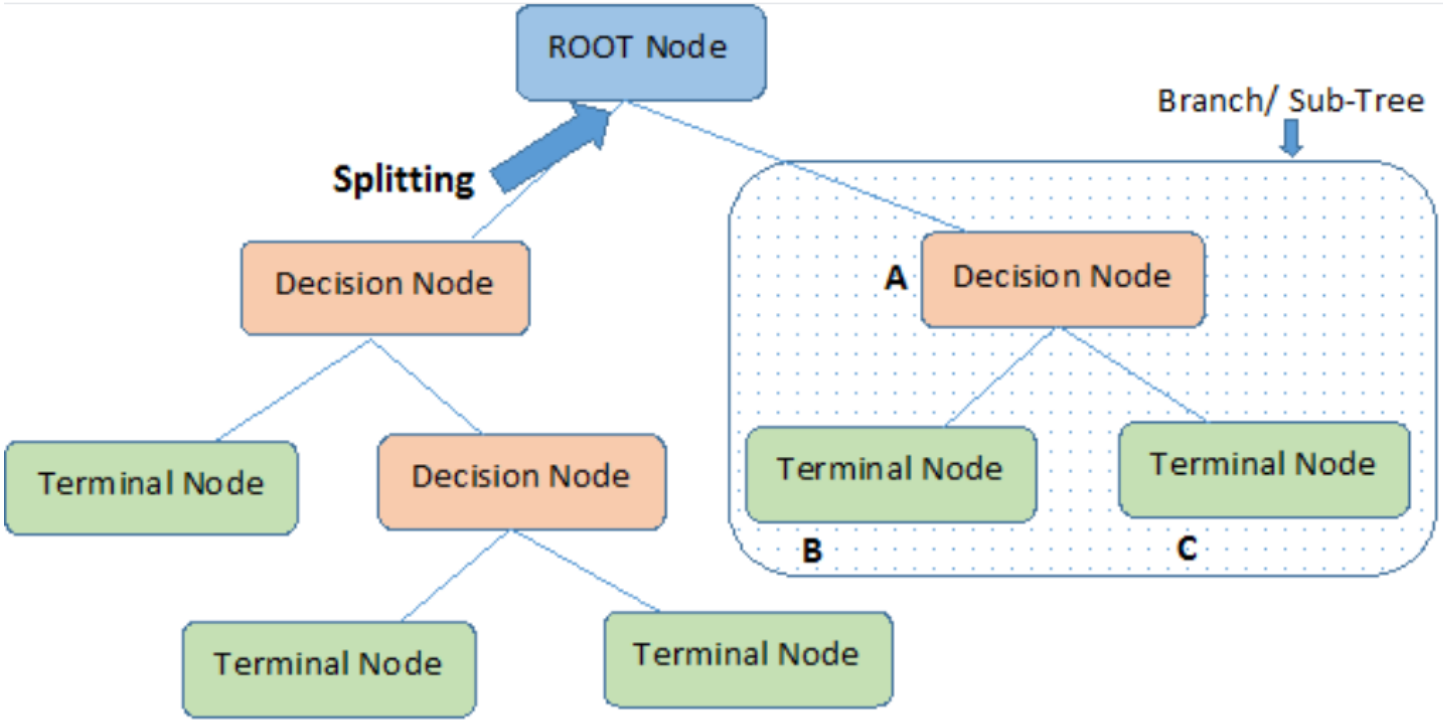
Τα δέντρα αποφάσεων είναι δημοφιλή στη μηχανική μάθηση λόγω της απλότητας, της ερμηνευσιμότητας και της ευελιξίας τους.

Είναι ένας εποπτευόμενος αλγόριθμος μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί τόσο για προβλήματα παλινδρόμησης (πρόβλεψη συνεχών τιμών) όσο και για προβλήματα ταξινόμησης (πρόβλεψη κατηγορικών τιμών).

Επιπλέον, χρησιμεύουν ως το θεμέλιο για πιο προηγμένες τεχνικές, όπως το bagging, το boosting και τα τυχαία δάση.

Το παρακάτω διάγραμμα θα απεικονίσει τις ορολογίες πίσω από τα δέντρα αποφάσεων:

Τι είναι το Δέντρο Αποφάσεων στη Μηχανική Μάθηση;



Δημοφιλείς τύποι αλγορίθμων δέντρων αποφάσεων

Ας κατανοήσουμε διαισθητικά τόσο τα δέντρα απόφασης παλινδρόμησης όσο και ταξινόμησης, τι είναι παρόμοιο και διαφορετικό σε καθένα, καθώς και τις συναρτήσεις σφάλματος.

Η παρακάτω εικόνα, βοηθά στην οπτικοποίηση της φύσης της κατάτμησης που πραγματοποιείται από ένα δέντρο παλινδρόμησης.

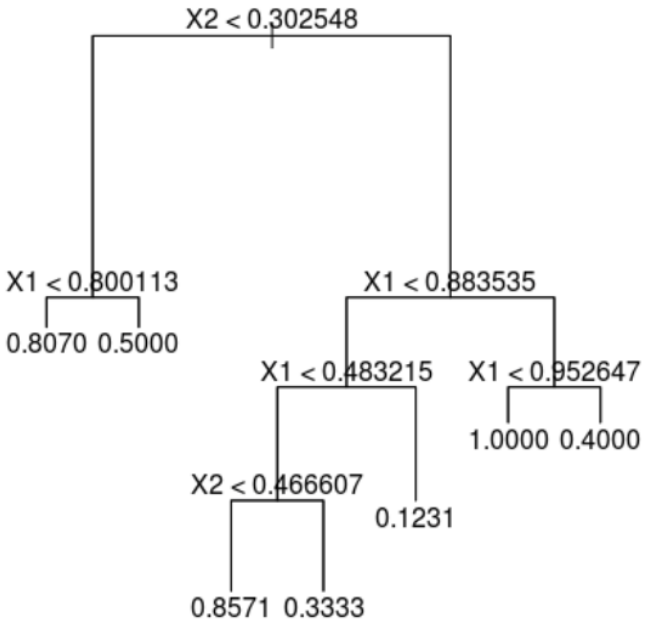
βλέπουμε ένα δέντρο που δεν έχει κλαδευτεί και ένα δέντρο παλινδρόμησης ταιριάζει σε ένα τυχαίο σύνολο δεδομένων.

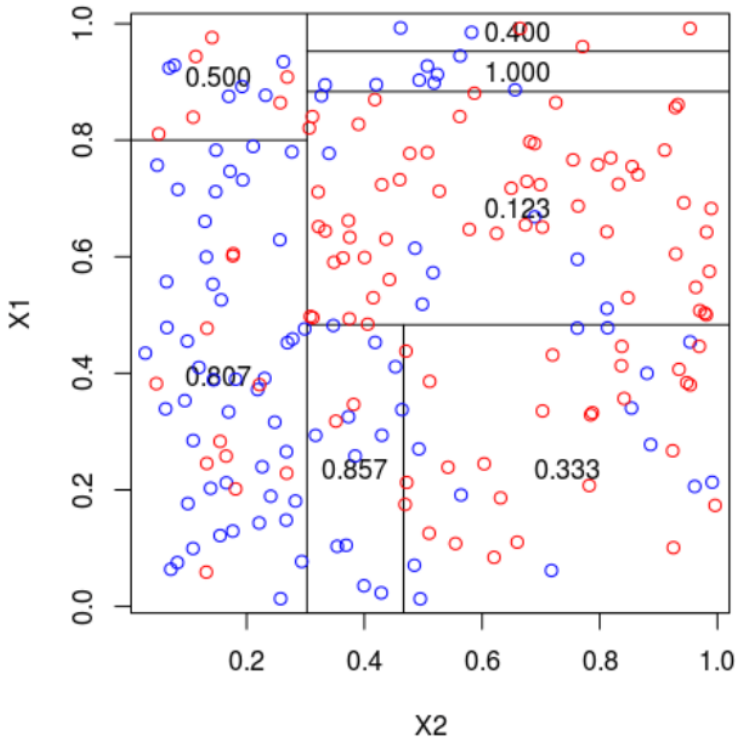
Και οι δύο απεικονίσεις δείχνουν μια σειρά από κανόνες διαχωρισμού, ξεκινώντας από την κορυφή του δέντρου.

Παρατηρήστε ότι κάθε διαχωρισμός του τομέα είναι ευθυγραμμισμένος με έναν από τους άξονες χαρακτηριστικών.

Η έννοια της παράλληλης διαίρεσης άξονα γενικεύεται ευθέως σε διαστάσεις μεγαλύτερες από δύο.

Για έναν χώρο χαρακτηριστικών μεγέθους p, ένα υποσύνολο $\mathbb{R}^p$, ο χώρος χωρίζεται σε περιοχές M, R_m, καθεμία από τις οποίες είναι p-διάσταση "υπερμπλοκ".





Για να δημιουργήσετε ένα δέντρο παλινδρόμησης, χρησιμοποιείτε πρώτα αναδρομικό δυαδικό διαχωρισμό για να αναπτύξετε ένα μεγάλο δέντρο στα δεδομένα εκπαίδευσης, σταματώντας μόνο όταν κάθε τερματικός κόμβος έχει λιγότερους από κάποιο ελάχιστο αριθμό παρατηρήσεων.

Ο αναδρομικός δυαδικός διαχωρισμός είναι ένας άπληστος (greedy algorithms) και από πάνω προς τα κάτω αλγόριθμος που χρησιμοποιείται για την ελαχιστοποίηση του υπολειπόμενου αθροίσματος τετραγώνων (RSS), ένα μέτρο σφάλματος που χρησιμοποιείται επίσης σε ρυθμίσεις γραμμικής παλινδρόμησης.

Το RSS, στην περίπτωση ενός διαμερισμένου χώρου χαρακτηριστικών με M κατατμήσεις δίνεται από:

$$RSS = \sum_{M=1}^{M} \sum_{i \in R_m} (y_i - \hat{y}_{R_m})^2$$

Ξεκινώντας από την κορυφή του δέντρου, το χωρίζετε σε 2 κλαδιά, δημιουργώντας ένα διαμέρισμα 2 χώρων.

Στη συνέχεια, πραγματοποιείτε αυτή τη συγκεκριμένη διαίρεση στην κορυφή του δέντρου πολλές φορές και επιλέγετε τη διαίρεση των χαρακτηριστικών που ελαχιστοποιεί το (τρέχον) RSS.

Στη συνέχεια, εφαρμόζετε το κλάδεμα πολυπλοκότητας κόστους στο μεγάλο δέντρο για να λάβετε μια ακολουθία από τα καλύτερα υποδέντρα, ως συνάρτηση του α.

Η βασική ιδέα εδώ είναι να εισαγάγουμε μια πρόσθετη παράμετρο συντονισμού, που συμβολίζεται με α που εξισορροπεί το βάθος του δέντρου και την καλή προσαρμογή του στα δεδομένα εκπαίδευσης.

Μπορείτε να χρησιμοποιήσετε τη διασταυρούμενη επικύρωση στο K-fold για να επιλέξετε α.

Αυτή η τεχνική περιλαμβάνει απλώς τη διαίρεση των παρατηρήσεων εκπαίδευσης σε πτυχές K για να εκτιμηθεί το ποσοστό σφάλματος δοκιμής των υποδέντρων.

Ο στόχος σας είναι να επιλέξετε αυτό που οδηγεί στο χαμηλότερο ποσοστό σφάλματος.

Ένα δέντρο ταξινόμησης είναι πολύ παρόμοιο με ένα δέντρο παλινδρόμησης, εκτός από το ότι χρησιμοποιείται για την πρόβλεψη μιας ποιοτικής απόκρισης και όχι μιας ποσοτικής.

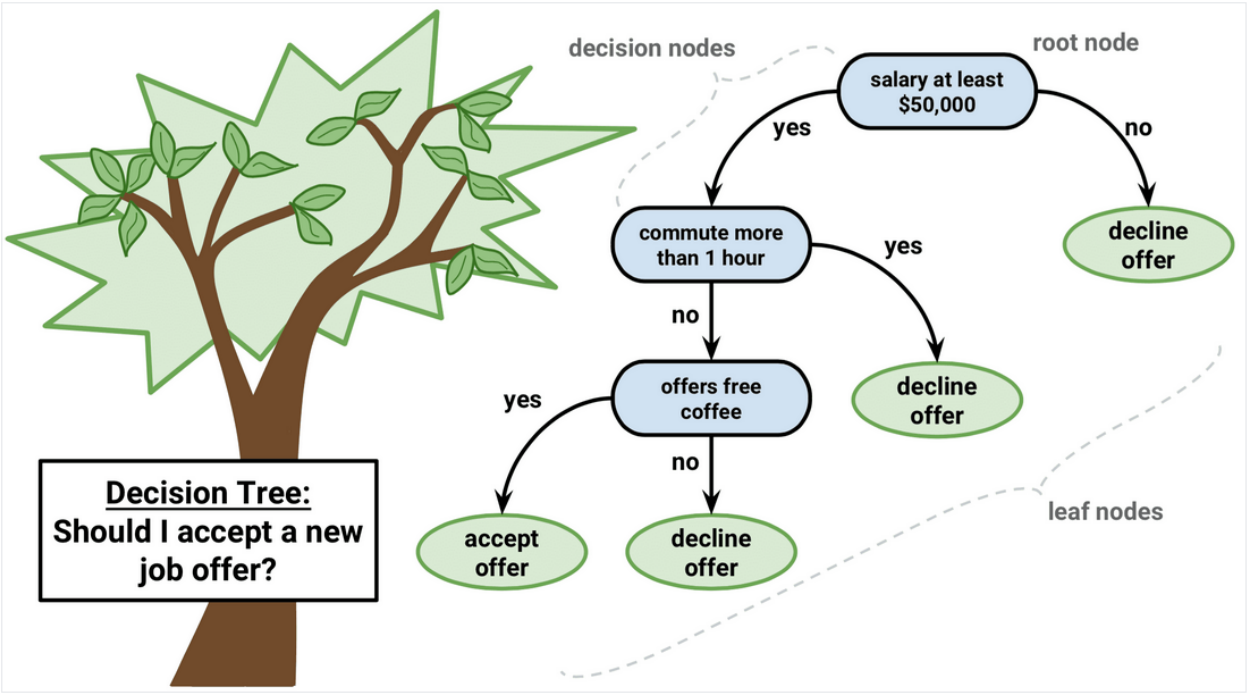
Θυμηθείτε ότι για ένα δέντρο παλινδρόμησης, η προβλεπόμενη απόκριση για μια παρατήρηση δίνεται από τη μέση απόκριση των εκπαιδευτικών παρατηρήσεων που ανήκουν στον ίδιο τερματικό κόμβο.

Αντίθετα, για ένα δέντρο ταξινόμησης, προβλέπετε ότι κάθε παρατήρηση ανήκει στην πιο συχνά εμφανιζόμενη κατηγορία εκπαιδευτικών παρατηρήσεων στην περιοχή στην οποία ανήκει.

Δημοφιλείς τύποι αλγορίθμων δέντρων αποφάσεων

Δέντρα Ταξινόμησης

Κατά την ερμηνεία των αποτελεσμάτων ενός δέντρου ταξινόμησης, συχνά ενδιαφέρεστε όχι μόνο για την πρόβλεψη κλάσης που αντιστοιχεί σε μια συγκεκριμένη περιοχή τερματικού κόμβου, αλλά και για τις αναλογίες κλάσεων μεταξύ των εκπαιδευτικών παρατηρήσεων που εμπίπτουν σε αυτήν την περιοχή.



Το έργο της ανάπτυξης ενός δέντρου ταξινόμησης είναι αρκετά παρόμοιο με το έργο της ανάπτυξης ενός δέντρου παλινδρόμησης.

Ακριβώς όπως στη ρύθμιση παλινδρόμησης, χρησιμοποιείτε αναδρομικό δυαδικό διαχωρισμό για να αναπτύξετε ένα δέντρο ταξινόμησης.

Ωστόσο, στη ρύθμιση ταξινόμησης, το Υπολειπόμενο άθροισμα τετραγώνων δεν μπορεί να χρησιμοποιηθεί ως κριτήριο για την πραγματοποίηση των δυαδικών διαχωρισμών.

Αντίθετα, μπορείτε να χρησιμοποιήσετε μία από αυτές τις 3 μεθόδους παρακάτω:

Ποσοστό σφάλματος ταξινόμησης: Αντί να βλέπετε πόσο απέχει μια αριθμητική απόκριση από τη μέση τιμή, όπως στη ρύθμιση παλινδρόμησης, μπορείτε να ορίσετε το ”ποσοστό επιτυχίας” ως το κλάσμα των παρατηρήσεων εκπαίδευσης σε μια συγκεκριμένη περιοχή που δεν ανήκει η πιο ευρέως εμφανιζόμενη τάξη. Το σφάλμα δίνεται από αυτή την εξίσωση: E = 1 - argmaxc($\hat{\pi}_{mc}$)

στην οποία το $\hat{\pi}_{mc}$ αντιπροσωπεύει το κλάσμα των δεδομένων εκπαίδευσης στην περιοχή Rm που ανήκουν στην κατηγορία c.

Δείκτης Gini: Ο δείκτης Gini είναι μια εναλλακτική μέτρηση σφάλματος που έχει σχεδιαστεί για να δείξει πόσο ”καθαρή” είναι μια περιοχή.

”Καθαρότητα” σε αυτή την περίπτωση σημαίνει πόσα από τα δεδομένα εκπαίδευσης σε μια συγκεκριμένη περιοχή ανήκουν σε μια κατηγορία.

Εάν μια περιοχή R_m περιέχει δεδομένα που προέρχονται ως επί το πλείστον από μια κατηγορία c, τότε η τιμή του δείκτη Gini θα είναι μικρή:

$$G = \sum_{c=1}^{C} (1 - \hat{\pi_{mc}})$$

Διασταυρούμενη Εντροπία (Cross-Entropy): Μια τρίτη εναλλακτική, η οποία είναι παρόμοια με τον Δείκτη Gini, είναι γνωστή ως Διασταυρούμενη Εντροπία ή Απόκλιση:

$$G = \sum_{c=1}^{C} (1 - \hat{\pi_{mc}})$$

Η διασταυρούμενη εντροπία θα λάβει μια τιμή κοντά στο μηδέν εάν τα $\hat{\pi}_{mc}$ είναι όλα κοντά στο 0 ή κοντά στο 1.

Επομένως, όπως ο δείκτης Gini, η διασταυρούμενη εντροπία θα λάβει ένα μικρή τιμή εάν ο $m_t h$ κόμβος είναι καθαρός.

Στην πραγματικότητα, αποδεικνύεται ότι ο δείκτης Gini και η διασταυρούμενη εντροπία είναι αρκετά παρόμοια αριθμητικά.

Κατά την κατασκευή ενός δέντρου ταξινόμησης, είτε ο δείκτης Gini είτε η διασταυρούμενη εντροπία χρησιμοποιούνται συνήθως για την αξιολόγηση της ποιότητας ενός συγκεκριμένου διαχωρισμού, καθώς είναι πιο ευαίσθητα στην καθαρότητα του κόμβου από το ποσοστό σφάλματος ταξινόμησης.

Οποιαδήποτε από αυτές τις 3 προσεγγίσεις μπορεί να χρησιμοποιηθεί κατά το κλάδεμα του δέντρου, αλλά το ποσοστό σφάλματος ταξινόμησης είναι προτιμότερο εάν ο στόχος είναι η ακρίβεια πρόβλεψης του τελικού κλαδευμένου δέντρου.

Πλεονεκτήματα και μειονεκτήματα των δέντρων απόφασης

Όσο κι αν θέλουμε να κατανοήσουμε τον αλγόριθμο και τα δυνατά του σημεία, είναι σημαντικό να κατανοήσουμε τα μειονεκτήματά του.

Η αλήθεια είναι ότι τα δέντρα αποφάσεων δεν ταιριάζουν καλύτερα σε όλους τους τύπους αλγορίθμων μηχανικής μάθησης, κάτι που ισχύει και για όλους τους αλγόριθμους μηχανικής μάθησης.

Πλεονεκτήματα

Εύκολο στην κατανόηση και την ερμηνεία. Τα δέντρα αποφάσεων είναι πολύ διαισθητικά και μπορούν εύκολα να οπτικοποιηθούν.

Η ιεραρχική τους δομή μοιάζει με την ανθρώπινη λήψη αποφάσεων, καθιστώντας τα προσβάσιμα ακόμη και σε μη ειδικούς.

Αυτή η ερμηνευσιμότητα είναι ζωτικής σημασίας σε καταστάσεις όπου η κατανόηση της διαδικασίας απόφασης είναι εξίσου σημαντική με την ίδια την πρόβλεψη.

Πλεονεκτήματα

Μπορεί να χειριστεί τόσο συνεχείς όσο και κατηγορηματικές μεταβλητές. Τα δέντρα αποφάσεων είναι ευέλικτα και μπορούν να διαχειρίζονται σύνολα δεδομένων με συνδυασμό συνεχών και κατηγορικών χαρακτηριστικών, καθώς και μεταβλητών στόχων οποιουδήποτε τύπου.

Αυτή η ευελιξία επιτρέπει στα δέντρα αποφάσεων να εφαρμόζονται σε ένα ευρύ φάσμα προβλημάτων.

Πλεονεκτήματα

Απαιτείται ελάχιστη προεπεξεργασία δεδομένων. Τα δέντρα απόφασης δεν απαιτούν κλιμάκωση ή κανονικοποίηση χαρακτηριστικών, καθώς είναι αμετάβλητα σε μονοτονικούς μετασχηματισμούς. Μπορούν επίσης να χειριστούν εύκολα τιμές και ακραίες τιμές που λείπουν, καθιστώντας τα κατάλληλα για ακατέργαστα και θορυβώδη δεδομένα.

Μπορεί να χρησιμοποιηθεί για την επιλογή χαρακτηριστικών και τον προσδιορισμό σημαντικών μεταβλητών. Τα δέντρα αποφάσεων εκτελούν φυσικά την επιλογή χαρακτηριστικών επιλέγοντας τα πιο ενημερωτικά χαρακτηριστικά για τον διαχωρισμό των δεδομένων.

Η σημασία ενός χαρακτηριστικού μπορεί να προσδιοριστεί με βάση το πόσο νωρίς εμφανίζεται στο δέντρο και πόσο συχνά χρησιμοποιείται για διάσπαση (splitting).

Μειονεκτήματα

Επιρρεπής σε υπερβολική προσαρμογή. Τα δέντρα αποφάσεων μπορεί να γίνουν υπερβολικά πολύπλοκα όταν μεγαλώνουν βαθιά, οδηγώντας σε υπερβολική προσαρμογή των δεδομένων εκπαίδευσης. Αυτό οδηγεί σε κακή γενίκευση των αόρατων δεδομένων.

Τεχνικές όπως το κλάδεμα ή ο καθορισμός ενός μέγιστου βάθους μπορούν να βοηθήσουν στην άμβλυνση αυτού του προβλήματος.
Ευαίσθητο σε μικρές αλλαγές στα δεδομένα. Μια μικρή αλλαγή στα δεδομένα εκπαίδευσης μπορεί να οδηγήσει σε μια εντελώς διαφορετική δομή δέντρου, καθιστώντας τα δέντρα αποφάσεων ασταθή.

Αυτή η ευαισθησία μπορεί να μειωθεί με μεθόδους συνόλου, όπως η συσσώρευση, η ενίσχυση ή τα τυχαία δάση, που χτίζουν πολλά δέντρα και συγκεντρώνουν τα αποτελέσματά τους.

Μειονεκτήματα

Τα μοντέλα δέντρων αποφάσεων συχνά δεν είναι τόσο ακριβή όσο άλλες μέθοδοι μηχανικής μάθησης. Λόγω της απλότητάς τους και της άπληστης φύσης της κατασκευής τους, τα δέντρα αποφάσεων μπορεί να μην παράγουν πάντα τα πιο ακριβή μοντέλα.

Ωστόσο, μπορούν να χρησιμεύσουν ως καλό σημείο εκκίνησης ή βασικό μοντέλο για σύγκριση με πιο εξελιγμένες μεθόδους.

Μειονεκτήματα

Μεροληπτούν προς τις κυρίαρχες τάξεις. Τα δέντρα αποφάσεων μπορεί να είναι προκατειλημμένα προς την κλάση πλειοψηφίας σε μη ισορροπημένα σύνολα δεδομένων, οδηγώντας σε μη βέλτιστες διαχωρισμούς και εσφαλμένη ταξινόμηση των περιπτώσεων μειοψηφίας.

Τεχνικές όπως η **υπερδειγματοληψία**, η **υποδειγματοληψία** ή η χρήση μάθησης με ευαισθησία στο κόστος μπορούν να βοηθήσουν στην αντιμετώπιση αυτού του ζητήματος.

Παρά αυτά τα μειονεκτήματα, τα δέντρα αποφάσεων παραμένουν δημοφιλής επιλογή σε πολλές εφαρμογές λόγω της απλότητας, της ερμηνευσιμότητας και της ευελιξίας τους.

Για τα ακόλουθα παραδείγματα, θα χρησιμοποιήσουμε το δημοφιλές σύνολο δεδομένων Boston Housing.

Το σύνολο δεδομένων Boston Housing περιέχει πληροφορίες σχετικά με την αγορά κατοικίας στη Βοστώνη της Μασαχουσέτης τη δεκαετία του 1970.

Έχει 506 παρατηρήσεις και 14 μεταβλητές, συμπεριλαμβανομένων 13 χαρακτηριστικών και 1 μεταβλητής στόχου.

Τα χαρακτηριστικά στο σύνολο δεδομένων Boston Housing είναι:

- **CRIM**: Κατά κεφαλήν ποσοστό εγκληματικότητας ανά πόλη
- **ZN**: Αναλογία οικιστικής γης που χωροθετείται για οικόπεδα άνω των 25.000 τ.μ.
- **INDUS**: Αναλογία στρέμματα επιχειρήσεων μη λιανικής ανά πόλη
- **CHAS**: εικονική μεταβλητή του ποταμού Charles (= 1 εάν η οδός οριοθετεί το ποτάμι, 0 διαφορετικά)
- **NOX**: Συγκέντρωση νιτρικού οξειδίου (μέρη ανά 10 εκατομμύρια)
- **RM**: Μέσος αριθμός δωματίων ανά κατοικία
- **ΗΛΙΚΙΑ**: Ποσοστό ιδιοχρησιμοποιούμενων μονάδων που κατασκευάστηκαν πριν από το 1940
- **DIS**: Σταθμισμένες αποστάσεις σε πέντε κέντρα απασχόλησης της Βοστώνης

Τα χαρακτηριστικά στο σύνολο δεδομένων Boston Housing είναι:

- **RAD**: Δείκτης προσβασιμότητας σε ακτινωτούς αυτοκινητόδρομους
- **TAX**: Συντελεστής φόρου ακίνητης περιουσίας πλήρους αξίας ανά 10.000$
- **PTRATIO**: Αναλογία μαθητών-δασκάλων ανά πόλη
- **Black**: αναλογία Μαύρων ανά πόλη
- **LSTAT**: Ποσοστό της κατώτερης θέσης του πληθυσμού.

Η μεταβλητή στόχος είναι το **MEDV** που αντιπροσωπεύει τη διάμεση αξία των κατοικιών σε 1000$.

Ο στόχος είναι να προβλεφθεί η διάμεση αξία των ιδιοκατοικούμενων κατοικιών (σε χιλιάδες δολάρια) με βάση τα δεδομένα.

Στο R, τα δεδομένα παρέχονται σε ένα πακέτο που ονομάζεται ”MASS”. Θα πρέπει να εγκαταστήσετε πολλά πακέτα για αυτό το σεμινάριο και να τα φορτώσετε.

Εφόσον πρόκειται για επανάληψη, ας δείξουμε αυτή τη διαδικασία με το πακέτο MASS μία φορά και θα το επαναλαμβάνετε κάθε φορά που βλέπετε ένα νέο πακέτο που χρησιμοποιείται σε αυτόν τον οδηγό.

Δόμηση μοντέλων δέντρων απόφασης βήμα προς βήμα στην R

```
# install the package
install.packages("MASS")

# Load the MASS package
library(MASS)

# Load the Boston Housing dataset
data(Boston)
```

Είναι συχνά απαραίτητο να εξερευνήσετε τα δεδομένα μέσω οπτικοποιήσεων και να εκτελέσετε βήματα προεπεξεργασίας δεδομένων πριν προχωρήσετε στη μοντελοποίηση. Ας δούμε την κατανομή των μεταβλητών μέσω ιστογραμμάτων.

Δόμηση μοντέλων δέντρων απόφασης βήμα προς βήμα στην R

```r
# Load the library
library(tidymodels)
library(tidyr)

# Prepare the dataset for ggplot2
boston_data_long <- Boston %>%
pivot_longer(cols = everything(),
names_to = "variable",
values_to = "value")

# Create a histogram for all numeric variables in one plot
boston_histograms <- ggplot(boston_data_long, aes(x = value)) +
geom_histogram(bins = 30, color = "black", fill = "lightblue") +
facet_wrap(~variable, scales = "free", ncol = 4) +
labs(title = "Histograms of Numeric Variables in the Boston Housing
     Dataset",
x = "Value",
y = "Frequency") +
theme_minimal()

# Plot the histograms
print(boston_histograms)
```

Δόμηση μοντέλων δέντρων απόφασης βήμα προς βήμα στην R

Είναι συχνά απαραίτητο να εξερευνήσετε τα δεδομένα μέσω οπτικοποιήσεων και να εκτελέσετε βήματα προεπεξεργασίας δεδομένων πριν προχωρήσετε στη μοντελοποίηση.

Ας δούμε την κατανομή των μεταβλητών μέσω ιστογραμμάτων.

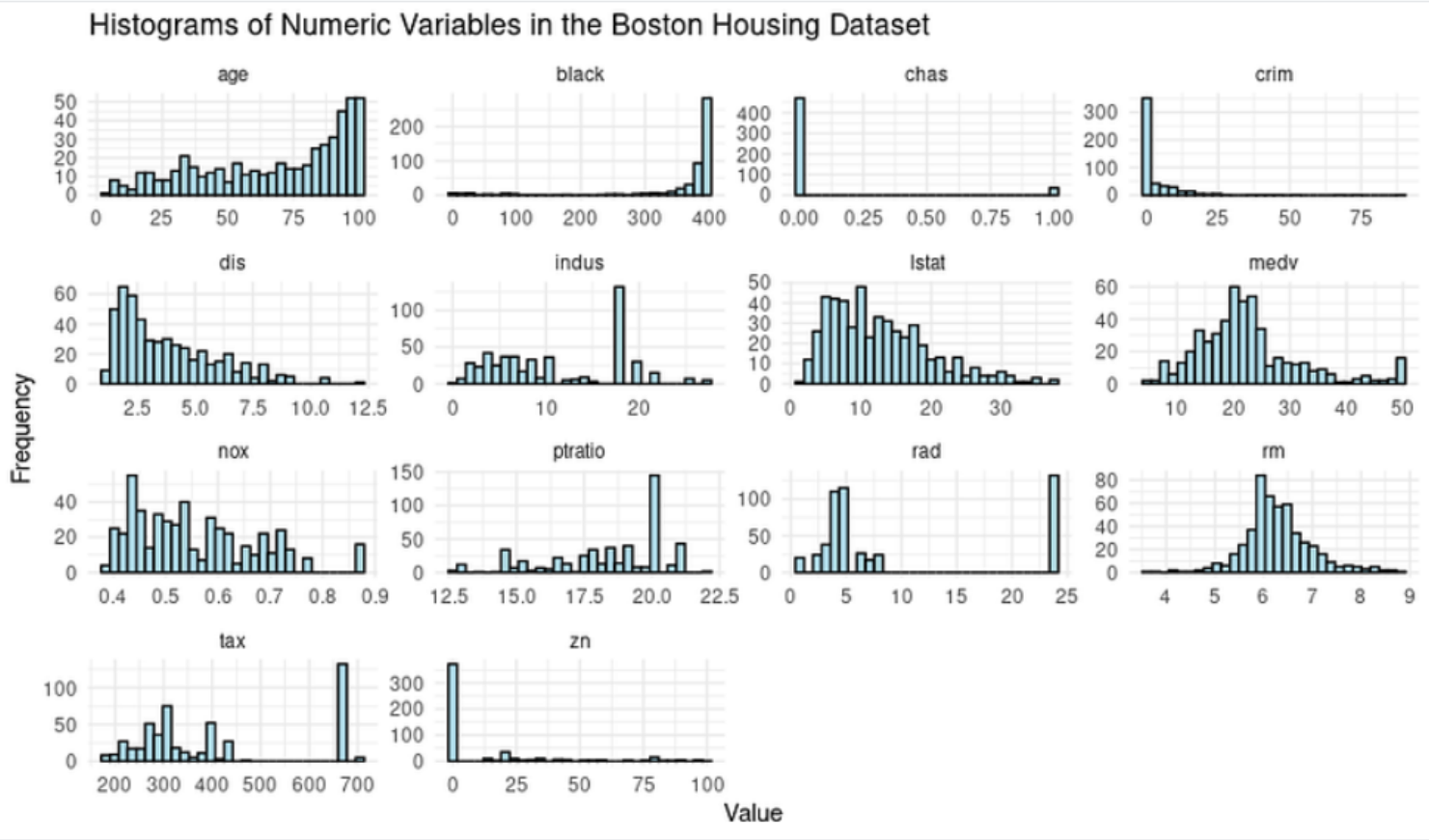
Δόμηση μοντέλων δέντρων απόφασης βήμα προς βήμα στην R

Κώδικας για τη δημιουργία τους:

```
# Load the library
library(tidymodels)
library(tidyr)

# Prepare the dataset for ggplot2
boston_data_long <- Boston %>%
pivot_longer(cols = everything(),
names_to = "variable",
values_to = "value")

# Create a histogram for all numeric variables in one plot
boston_histograms <- ggplot(boston_data_long, aes(x = value)) +
geom_histogram(bins = 30, color = "black", fill = "lightblue") +
facet_wrap(~variable, scales = "free", ncol = 4) +
labs(title = "Histograms of Numeric Variables in the Boston Housing
    Dataset",
x = "Value",
y = "Frequency") +
theme_minimal()

# Plot the histograms
print(boston_histograms)
```

Δόμηση μοντέλων δέντρων απόφασης βήμα προς βήμα στην R

Και η έξοδος μοιάζει με αυτό:



Δόμηση μοντέλων δέντρων απόφασης βήμα προς βήμα στην R

Παρατηρούμε κάποιες ακραίες τιμές, ειδικά στις στήλες όπως RAD, TAX και NOX. Ο στόχος μας για αυτό το σεμινάριο είναι να επικεντρωθούμε στη φάση της μοντελοποίησης του δέντρου αποφάσεων.

Επομένως, ας χωρίσουμε το σύνολο δεδομένων σε σύνολα εκπαίδευσης και δοκιμών.

```
# Split the data into training and testing sets

set.seed(123)
data_split <- initial_split(Boston, prop = 0.75)
train_data <- training(data_split)
test_data <- testing(data_split)
```

Τώρα θα προχωρήσουμε στη μοντελοποίηση και την αξιολόγηση της απόδοσης του μοντέλου.

Μοντελοποίηση δέντρου απλής απόφασης

Χρησιμοποιώντας τη συνάρτηση decision_tree από το πακέτο Tidymodels στο R, είναι εύκολο να δημιουργήσετε πρώτα μια προδιαγραφή μοντέλου δέντρου αποφάσεων και μετά να προσαρμόσετε το μοντέλο στα δεδομένα εκπαίδευσης.

Εδώ χρησιμοποιούμε το μοντέλο regression «παλίνδρωσης».

Για ένα δέντρο αποφάσεων ταξινόμησης, θα πρέπει να χρησιμοποιήσουμε τη λειτουργία classification «ταξινόμηση».

```
# Create a decision tree model specification
tree_spec <- decision_tree() %>%
set_engine("rpart") %>%
set_mode("regression")

# Fit the model to the training data
tree_fit <- tree_spec %>%
fit(medv ~ ., data = train_data)
```

Αξιολόγηση της απόδοσης του δέντρου αποφάσεων

Για να αξιολογήσουμε την απόδοση του μοντέλου, θα χρησιμοποιήσουμε το πακέτο Tidymodels για να υπολογίσουμε το ριζικό μέσο τετράγωνο σφάλμα (RMSE) και την τιμή R-τετράγωνο για το μοντέλο δέντρου αποφάσεών μας στα δεδομένα δοκιμής.

```
# Make predictions on the testing data
predictions <- tree_fit %>%
predict(test_data) %>%
pull(.pred)

# Calculate RMSE and R-squared
metrics <- metric_set(rmse, rsq)
model_performance <- test_data %>%
mutate(predictions = predictions) %>%
metrics(truth = medv, estimate = predictions)

print(model_performance)
```

Αξιολόγηση της απόδοσης του δέντρου αποφάσεων

Θα λάβετε ένα αποτέλεσμα που παρουσιάζει δύο μετρήσεις απόδοσης: Σφάλμα ριζικού μέσου τετραγώνου (RMSE) και R-τετράγωνο (R^2).

```
.metric  .estimator  .estimate
chr>     <chr>           <dbl>
 rmse     standard        5.22
 rsq      standard        0.689
```

Ανάλυση των αποτελεσμάτων του μοντέλου

Μπορούμε επίσης να βελτιστοποιήσουμε τις hyper-parameters για να συμπιέσουμε την απόδοση ή να χρησιμοποιήσουμε πιο σύνθετα μοντέλα όπως τα Random Forests και το XGBoost με κόστος την ερμηνευτικότητα του μοντέλου.
ω

Πρόβλεψη αποτελεσμάτων με χρήση του μοντέλου δέντρου αποφάσεων

Μόλις είστε ευχαριστημένοι με το μοντέλο, ήρθε η ώρα να αφήσετε το μοντέλο να κάνει προβλέψεις.

Αυτό είναι το ίδιο με το πώς κάναμε με τα δεδομένα δοκιμής χρησιμοποιώντας τη συνάρτηση predict(), αλλά θα πρέπει να παρέχουμε ένα νέο σύνολο δεδομένων που μιμούνται πληροφορίες για ένα νέο σπίτι στη Βοστώνη.

Είναι ένα πιθανό σενάριο όταν το μοντέλο κυκλοφορεί σε περιβάλλον παραγωγής.

```
# Make predictions on new data
new_data <- tribble(
~crim, ~zn, ~indus, ~chas, ~nox, ~rm, ~age, ~dis, ~rad, ~tax, ~ptratio,
      ~black, ~lstat,
0.03237, 0, 2.18, 0, 0.458, 6.998, 45.8, 6.0622, 3, 222, 18.7, 394.63,
     2.94
)
predictions <- predict(tree_fit, new_data)
print(predictions)
```

Πρόβλεψη αποτελεσμάτων με χρήση του μοντέλου δέντρου αποφάσεων

Και θα λάβετε την προβλεπόμενη μεσαία αξία (σε $1000) αυτού του συγκεκριμένου σπιτιού:

```
# A tibble: 1 × 1
  .pred
  <dbl>
1  37.8
```

Έως εδώ, μάθαμε τα βήματα δημιουργίας ενός μοντέλου δέντρου αποφάσεων—ας εστιάσουμε τώρα στο πώς μπορούμε να ερμηνεύσουμε τι συμβαίνει μέσα στο μοντέλο για εμάς και τους ενδιαφερόμενους που χρησιμοποιούν τη λύση που μόλις δημιουργήσαμε.

Ερμηνεία του Δέντρου Αποφάσεων

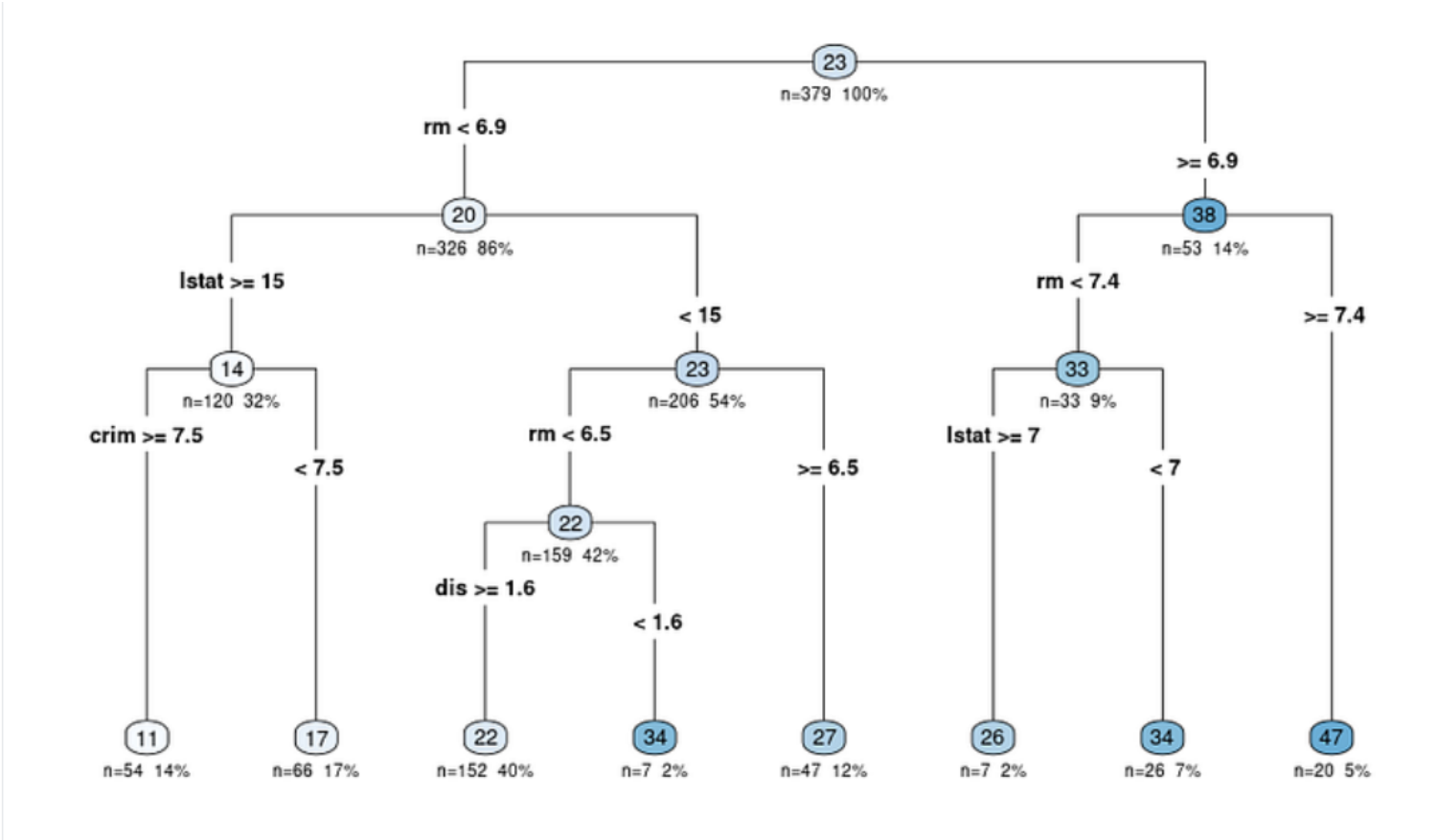
Το πιο σημαντικό πλεονέκτημα, όπως αναφέραμε προηγουμένως, είναι η ερμηνευτικότητα των μοντέλων δέντρων αποφάσεων.

Ας οπτικοποιήσουμε το δέντρο αποφάσεων για να κατανοήσουμε καλύτερα το μοντέλο:

```
# Load the library
library(rpart.plot)

# Plot the decision tree
rpart.plot(tree_fit$fit, type = 4, extra = 101, under = TRUE, cex =
    0.8, box.palette = "auto")
```

Ερμηνεία του Δέντρου Αποφάσεων

Θα δείτε το παρακάτω διάγραμμα:



Ερμηνεία του Δέντρου Αποφάσεων

Το διάγραμμα εξόδου της συνάρτησης rpart.plot δείχνει μια αναπαράσταση δέντρου αποφάσεων του μοντέλου.

Σε αυτό το διάγραμμα, κάθε κόμβος αντιπροσωπεύει μια διαίρεση στο δέντρο αποφάσεων με βάση τις μεταβλητές πρόβλεψης.

Το διάγραμμα εξόδου περιλαμβάνει αρκετές πληροφορίες που μπορούν να μας βοηθήσουν να ερμηνεύσουμε το δέντρο αποφάσεων:

Τα δενδρικά διαγράμματα ξεκινούν με τον κόμβο
Οι κόμβοι αντιπροσωπεύονται από κύκλους και συνδέονται με γραμμές, που δείχνουν την ιεραρχική δομή του δέντρου αποφάσεων.

Το δέντρο ξεκινά με έναν κόμβο ρίζας στην κορυφή και διακλαδίζεται σε εσωτερικούς κόμβους, οδηγώντας τελικά στους τερματικούς κόμβους ή στα φύλλα στο κάτω μέρος.

Ερμηνεία του Δέντρου Αποφάσεων

Τα κριτήρια διαχωρισμού για κάθε εσωτερικό κόμβο

Κάθε εσωτερικός κόμβος εμφανίζει το κριτήριο διαχωρισμού, το οποίο είναι η μεταβλητή πρόβλεψης και η τιμή που χρησιμοποιείται για τον διαχωρισμό των δεδομένων σε δύο υποσύνολα.

Για παράδειγμα, ένας κόμβος μπορεί να δείχνει ”RM < 6,8”, υποδεικνύοντας ότι οι παρατηρήσεις με μέσο αριθμό δωματίων ανά κατοικία (RM) μικρότερο από 6,8 θα ακολουθούν τον αριστερό κλάδο, ενώ παρατηρήσεις με RM μεγαλύτερο ή ίσο με 6,8 θα ακολουθούν το δεξιό κλαδί.

Ερμηνεία του Δέντρου Αποφάσεων

N: Αριθμός παρατηρήσεων μετά τη διάσπαση

Η τιμή n σε κάθε κόμβο αντιπροσωπεύει τον αριθμό των παρατηρήσεων στο σύνολο δεδομένων που εμπίπτουν σε αυτόν τον συγκεκριμένο κόμβο.

Για παράδειγμα, εάν ένας κόμβος εμφανίζει ”n = 100”, σημαίνει ότι 100 παρατηρήσεις στο σύνολο δεδομένων πληρούν τα κριτήρια των γονικών κόμβων αυτού του κόμβου.

X%: Το ποσοστό των συνολικών παρατηρήσεων δεδομένων που έφτασαν στον συγκεκριμένο κόμβο.

Η ποσοστιαία τιμή σάς βοηθά να κατανοήσετε το σχετικό μέγεθος κάθε κόμβου σε σύγκριση με ολόκληρο το σύνολο δεδομένων, δείχνοντας πώς τα δεδομένα χωρίζονται και κατανέμονται στο δέντρο.

Ένα υψηλότερο ποσοστό σημαίνει ότι ένα μεγαλύτερο ποσοστό των δεδομένων έχει ακολουθήσει τη διαδρομή απόφασης που οδηγεί στον συγκεκριμένο κόμβο, ενώ ένα χαμηλότερο ποσοστό υποδηλώνει μικρότερο ποσοστό των δεδομένων που φτάνουν σε αυτόν τον κόμβο.

Η προβλεπόμενη τιμή για κάθε κόμβο

Η προβλεπόμενη τιμή σε κάθε κόμβο εμφανίζεται ως αριθμός σε έναν έγχρωμο κύκλο (κόμβο).

Σε ένα δέντρο παλινδρόμησης, αυτή είναι η μέση τιμή μεταβλητής στόχου για όλες τις παρατηρήσεις που εμπίπτουν σε αυτόν τον κόμβο.

Για παράδειγμα, ο τελευταίος κύριος κόμβος που δείχνει 47 σημαίνει ότι η μέση τιμή μεταβλητής στόχου (στην περίπτωσή μας, η διάμεση τιμή των κατοικιών κατοικιών) για όλες τις παρατηρήσεις σε αυτόν τον κόμβο είναι 47.

Ερμηνεία του Δέντρου Αποφάσεων

Ξεκινήστε από τον ριζικό κόμβο και ακολουθήστε τους κλάδους

Έτσι, όταν ερμηνεύετε τυχόν αποτελέσματα, ξεκινάτε από τον ριζικό κόμβο και ακολουθείτε τους κλάδους με βάση τα κριτήρια διαχωρισμού μέχρι να φτάσετε σε έναν τερματικό κόμβο.

Η προβλεπόμενη τιμή στον τερματικό κόμβο δίνει την πρόβλεψη του μοντέλου για μια δεδομένη παρατήρηση και το σκεπτικό πίσω από την απόφαση.

Ερμηνεία του Δέντρου Αποφάσεων

Εξαγωγή κανόνων από το δέντρο αποφάσεων

Εάν εξακολουθείτε να προτιμάτε να εξάγετε τους κανόνες σε μορφή κειμένου (αντί να διασχίζετε το διάγραμμα)—μπορείτε να το κάνετε επίσης, χρησιμοποιώντας την ίδια βιβλιοθήκη που χρησιμοποιήσαμε για να σχεδιάσουμε το διάγραμμα.

Ερμηνεία του Δέντρου Αποφάσεων

```
rules <- rpart.rules(tree_fit$fit)
print(rules)
```

Και θα δείτε το αποτέλεσμα με τις προβλεπόμενες τιμές και τους κανόνες που ακολουθεί για να φτάσετε σε αυτήν την τιμή όπως παρακάτω:

```
medv
11 when rm <  6.9          & lstat >= 15 & crim >= 7.5
17 when rm <  6.9          & lstat >= 15 & crim <  7.5
22 when rm <  6.5          & lstat <  15                 & dis >= 1.6
26 when rm is 6.9 to 7.4   & lstat >=  7
27 when rm is 6.5 to 6.9   & lstat <  15
34 when rm <  6.5          & lstat <  15                 & dis <  1.6
34 when rm is 6.9 to 7.4   & lstat <   7
47 when rm >=        7.4
```

Τώρα που βλέπετε τους κανόνες, ίσως αναρωτηθείτε πώς μπορεί να ληφθεί η απόφαση με 3-4 μεταβλητές όταν τροφοδοτούμε πολλές περισσότερες μεταβλητές στο δέντρο αποφάσεων.

Κατανόηση των σημαντικών μεταβλητών από τα δέντρα αποφάσεων

Έχουμε ήδη αποκαλύψει το δεντροδιάγραμμα και τον τρόπο λειτουργίας του μοντέλου.
Μια τελευταία πτυχή της ερμηνείας είναι η κατανόηση των σημαντικών μεταβλητών από το σύνολο δεδομένων.

Γιατί είναι κρίσιμο;
Επιλογή χαρακτηριστικών. Ο εντοπισμός σημαντικών μεταβλητών μπορεί να σας βοηθήσει να εστιάσετε σε σχετικές λειτουργίες και ενδεχομένως να αφαιρέσετε λιγότερο σημαντικές, απλοποιώντας το μοντέλο σας και μειώνοντας τον θόρυβο.
Πεδίο γνώσης. Η απόκτηση πληροφοριών σχετικά με τα χαρακτηριστικά που είναι πιο σημαντικά μπορεί να σας βοηθήσει να κατανοήσετε καλύτερα τις σχέσεις μεταξύ των μεταβλητών και της μεταβλητής στόχου στο πλαίσιο του συγκεκριμένου προβλήματός σας.
Μοντέλο ερμηνεία. Η κατανόηση των βασικών χαρακτηριστικών που οδηγούν τις προβλέψεις του μοντέλου μπορεί να προσφέρει πολύτιμες πληροφορίες για τη διαδικασία λήψης αποφάσεων.

Κατανόηση των σημαντικών μεταβλητών από τα δέντρα αποφάσεων

Στα δέντρα απόφασης, η σημαντικότητα των μεταβλητών καθορίζεται συνήθως από τα χαρακτηριστικά που χρησιμοποιούνται για τη διαίρεση στους κόμβους.

Τα χαρακτηριστικά που χρησιμοποιούνται για τον διαχωρισμό ψηλότερα στο δέντρο ή που χρησιμοποιούνται πιο συχνά μπορούν να θεωρηθούν πιο σημαντικά.

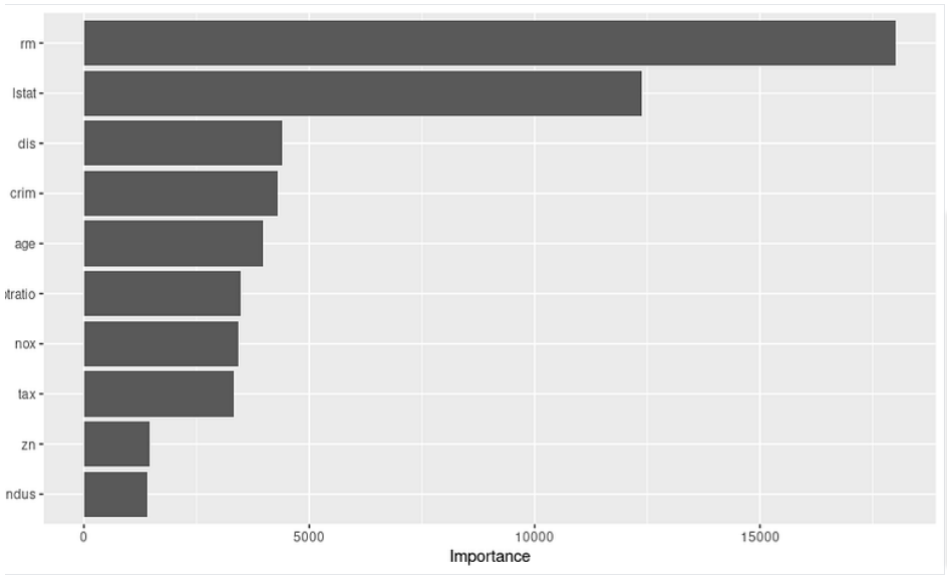
Η σημασία μιας μεταβλητής μπορεί να ποσοτικοποιηθεί με τη μείωση της μέτρησης της ακαθαρσίας (π.χ. δείκτης Gini ή μέσο τετράγωνο σφάλμα) που φέρνει όταν χρησιμοποιείται για διαχωρισμό.

Το πακέτο ”VIP” στο R έχει αφαιρέσει όλη την πολυπλοκότητα υπολογισμού και η απάντηση μπορεί να ληφθεί μέσω του παρακάτω κώδικα:

Κατανόηση των σημαντικών μεταβλητών από τα δέντρα αποφάσεων

```
# Load the necessary library
library(vip)

# Create a variable importance plot
var_importance <- vip::vip(tree_fit, num_features = 10)
print(var_importance)
```

Και θα λάβετε το διάγραμμα σημαντικότητας των μεταβλητών:



Κατανόηση των σημαντικών μεταβλητών από τα δέντρα αποφάσεων

Από το διάγραμμα, θα μπορούσατε να κάνετε περαιτέρω έρευνα σχετικά με το γιατί αυτές οι μεταβλητές είναι σημαντικές, σε συνεργασία με ειδικούς του τομέα.

Για παράδειγμα, με βάση την παραπάνω γραφική παράσταση, μπορούμε να συμπεράνουμε τις 3 πρώτες σημαντικές μεταβλητές και τη λογική:

RM: Ο μέσος αριθμός δωματίων ανά κατοικία είναι ένα εξαιρετικά κρίσιμο χαρακτηριστικό, υποδηλώνοντας ότι τα σπίτια με περισσότερα δωμάτια τείνουν να έχουν υψηλότερες μέσες τιμές.

Κατανόηση των σημαντικών μεταβλητών από τα δέντρα αποφάσεων

LSTAT: Το ποσοστό του πληθυσμού με χαμηλότερη κοινωνικοοικονομική κατάσταση είναι μια άλλη σημαντική μεταβλητή, υποδεικνύοντας ότι οι γειτονιές χαμηλότερης θέσης είναι πιθανό να έχουν χαμηλότερες μέσες τιμές κατοικίας.

DIS: Η σταθμισμένη απόσταση από τα κέντρα απασχόλησης είναι μια ζωτική μεταβλητή, δείχνοντας ότι τα σπίτια που βρίσκονται πιο κοντά σε μεγάλα κέντρα απασχόλησης είναι πιθανό να έχουν υψηλότερες διάμεσες τιμές.

Επομένως, θυμηθείτε να ελέγξετε το διάγραμμα μεταβλητής σημασίας πριν ολοκληρώσετε το μοντέλο σας. αυτό μπορεί να σας βοηθήσει να δημιουργήσετε και να επιλέξετε καλύτερες λειτουργίες για να βελτιστοποιήσετε την απόδοση.

Συμπέρασμα

Σε αυτή την ενότητα του μαθήματος, διερευνήσαμε τις θεμελιώδεις έννοιες των δέντρων αποφάσεων και προσεγγίσαμε όχι μόνο τη δημιουργία μοντέλων αλλά και την ερμηνεία τους.

Τα δέντρα αποφάσεων είναι ισχυρά και ερμηνεύσιμα μοντέλα τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης, καθιστώντας τα ένα ουσιαστικό εργαλείο στο οπλοστάσιο ενός επιστήμονα δεδομένων.