Τελική Εργασία Στατιστικής Μάθησης

Μελέτη Αιτήσεων Πιστωτικής Κάρτας, με κριτήριο το κοινωνικό-οικονομικό υπόβαθρο του δείγματος

Ανδρέας Πολυχρονίδης ΑΕΜ:17085

Περιεχόμενα

Εισαγωγή

Η **Στατιστική Μάθηση** (ή Μηχανική μάθηση όπως συναντάμε πιο συχνά) αποτελεί ένα από τα πεδία της επιστήμης των υπολογιστών, που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μάθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα. Η μηχανική μάθηση είναι στενά συνδεδεμένη με υπολογιστική στατιστική, ένας κλάδος, που επίσης επικεντρώνεται στην πρόβλεψη μέσω της χρήσης των υπολογιστών.

Σύμφωνα με το τελευταίο σχόλιο ενεργούμε παρακάτω, καθώς χρησιμοποιούμε τις μεθόδους εκπαίδευσης υπολογιστικών μονάδων με σκοπό την ανάλυση δεδομένων και την ταξινόμηση τους , προκειμένου να κάνουμε προβλέψεις σχετικά με τις μεταβλητές που αποτελούν το dataset μας. Στην παρακάτω παρουσίαση θα ασχοληθούμε με το σύνολο δεδομένων ***CreditCard*** της βιβλιοθήκης AER της R.

Το dataset ***CreditCard*** συνδέει το πιστωτικό ιστορικό ενός δείγματος ανθρώπων που έχουν αιτηθεί για την έκδοση μίας συγκεκριμένης πιστωτικής κάρτας. Το σύνολό μας αποτελείται από 1319 παρατηρήσεις και 12 μεταβλητές, 3 ποιοτικές και 9 ποσοτικές, τις οποίες θα αναλύσουμε στη συνέχεια.

```
$ card        : chr  "yes" "yes" "yes" "yes" ...
$ reports     : int  0 0 0 0 0 0 0 0 0 0 ...
$ age         : num  37.7 33.2 33.7 30.5 32.2 ...
$ income      : num  4.52 2.42 4.5 2.54 9.79 ...
$ share       : num  0.03327 0.00522 0.00416 0.06521 0.06705
...
$ expenditure: num  124.98 9.85 15 137.87 546.5 ...
$ owner       : chr  "yes" "no" "yes" "no" ...
$ selfemp     : chr  "no" "no" "no" "no" ...
$ dependents  : int  3 3 4 0 2 0 2 0 0 0 ...
$ months      : int  54 34 58 25 64 54 7 77 97 65 ...
$ majorcards  : int  1 1 1 1 1 1 1 1 1 1 ...
$ active      : int  12 13 5 7 5 1 5 3 6 18 ...
```

Περιγραφή συνόλου δεδομένων

Το σύνολο δεδομένων ***CreditCard*** με το οποίο θα ασχοληθούμε παρουσιάζει το πιστωτικό παρελθόν και γενικό υπόβαθρο των ατόμων που έχουν αιτηθεί έκδοση νέας πιστωτικής κάρτας. Οι μεταβλητές μας είναι οι εξής :

- card : εξετάζει αν το αίτημα για νέα πιστωτική κάρτα έγινε δεκτό
- owner : εξετάζει αν το άτομο έχει στην κατοχή του την προσωπική του κατοικία
- selfemp : εξετάζει αν το άτομο είναι αυτοαπασχολούμενος
- reports : ο αριθμός των αρνητικών αναφορών που έχει δεχθεί κάποιος
- age : η ηλικία του ατόμου + 12τα του έτους
- income : το ετήσιο εισόδημα του ατόμου
- share : αναλογία εξόδων πιστωτικών καρτών με το ετήσιο εισόδημα
- expenditure : μέσος όρος μηνιαίων εξόδων πιστωτικών καρτών
- dependents : αριθμός χρεών του ατόμου
- months : μήνες παραμονής στην ίδια διεύθυνση κατοικίας
- majorcards : αριθμός πιστωτικών καρτών που κατέχει το άτομο
- active : ο αριθμός των ενεργών λογαριασμών πιστωτικών καρτών

Οι μεταβλητές card, owner και selfemp είναι ποιοτικές, ενώ οι υπόλοιπες είναι ποσοτικές. Ο τρόπος ζωής όλων των ατόμων του δείγματος φαίνεται από την ηλικία, την επαγγελματική τους κατάσταση, το πιστωτικό τους ιστορικό, τα πιστωτικά τους έξοδα σε σχέση με το εισόδημα τους και τη διάρκεια της παραμονής τους στην ίδια διεύθυνση κατοικίας . Ως στόχο ,έχουμε την ανάλυση των μεταβλητών που περιγράφουν αυτά τα στοιχεία και την τελική εξαγωγή συμπερασμάτων για το δείγμα μας.

Τα αρχικά συμπεράσματά μας θα προκύπτουν από τις ιδιότητες της κάθε ποιοτικής μεταβλητής που περιγράφει το σύνολο δεδομένων και από την πιθανή συσχέτιση που έχουν αυτές μεταξύ τους. Αντίστοιχα, θα αναλύσουμε τα μέτρα κεντρικής τάσης και διασποράς και θα κρίνουμε με ορισμένα test, που βασίζονται στην ανάλυση διασποράς των ποσοτικών μεταβλητών, την ευστοχία των εκτιμήσεων μας. Ταυτόχρονα, θα υπάρχουν και διαγράμματα με σκοπό την κατανόηση των μεταβλητών όσον αφορά στις τιμές τους και στη συσχέτιση που μπορεί να υπάρχει μεταξύ τους.

Τελικά, η ταξινόμηση του συνόλου δεδομένων θα αποτελέσει την κύρια πηγή εξαγωγής συμπερασμάτων, καθώς είναι η πιο "έμπιστη" μορφή συσχέτισης όλων των μεταβλητών μας. Τα είδη ταξινόμησης που θα δούμε έχουν κάποιες διακριτές διαφοροποιήσεις τόσο στον διαχωρισμό του δείγματος όσο και στα αποτελέσματα τους.

Κύριο Μέρος

Ανάλυση ποιοτικών μεταβλητών

Οι ποιοτικές μεταβλητές είναι οι ακόλουθες :

- card : εξετάζει αν το αίτημα για νέα πιστωτική κάρτα έγινε δεκτό
- owner : εξετάζει αν το άτομο έχει στην κατοχή του την προσωπική του κατοικία
- selfemp : εξετάζει αν το άτομο είναι αυτοαπασχολούμενος

Μεταβλητή card

Για την μεταβλητή card, κατασκευάζουμε τον πίνακα συχνοτήτων και σχετικών συχνοτήτων, προκειμένου να δούμε σε πόσους και σε ποιο ποσοστό του δείγματος είχαμε θετική απάντηση και πόσα άτομα έλαβαν αρνητική απάντηση.

Στη συνέχεια, συνδυάζουμε του δύο πίνακες και προκύπτει ο πίνακας συχνοτήτων και σχετικών συχνοτήτων της μεταβλητής card.

	card	Frequency	Relative Frequency
1	no	296	0.2244124
2	yes	1023	0.7755876

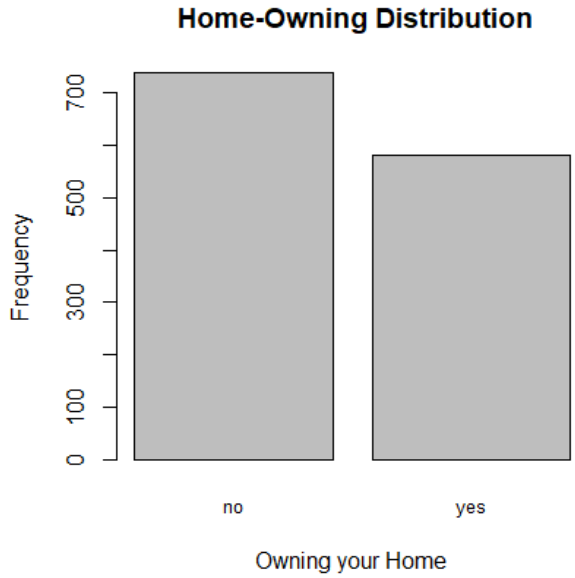
Με βάση τον παραπάνω πίνακα, καταλήγουμε στα συμπεράσματα μας. Δηλαδή, το 77,5% του δείγματος (1023 άτομα) έλαβε θετική απάντηση στην αίτηση του για νέα πιστωτική κάρτα, ενώ για το υπόλοιπο 22,5% περίπου του δείγματος (296 άτομα) δεν έγινε δεκτό το αίτημα τους.

Μεταβλητή owner

Συνεχίζουμε με την μεταβλητή owner, κατασκευάζοντας τους αντίστοιχους πίνακες συχνότητας και σχετικής συχνότητας, εξετάζοντας έτσι πόσοι και ποιο ποσοστό του δείγματος έχει στην κατοχή του την προσωπική του κατοικία.

Στη συνέχεια, συνδυάζουμε τους δύο πίνακες και προκύπτει ο πίνακας συχνοτήτων και σχετικών συχνοτήτων της μεταβλητής owner.

	owner	Frequency	Realitve Frequency
1	no	738	0.5595148
2	yes	581	0.4404852



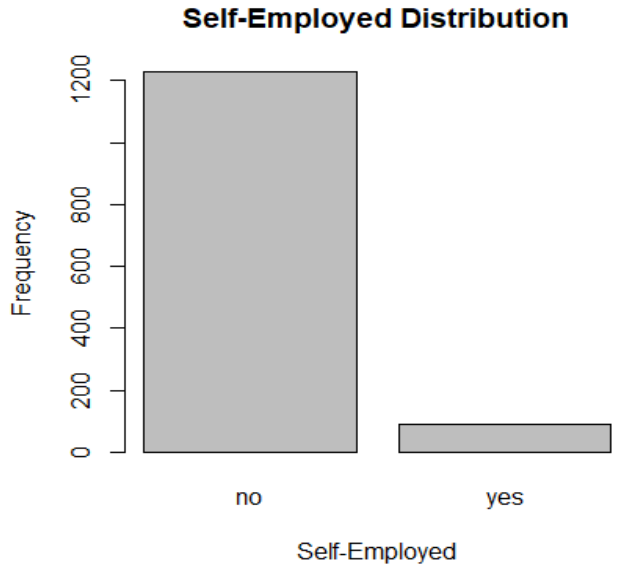
Έτσι, καταλήγουμε στα συμπεράσματά μας. Το 44% του δείγματος (581 άτομα) έχει στην κατοχή του την προσωπική του κατοικία, ενώ κάτι τέτοιο δεν ισχύει για το υπόλοιπο 56% του δείγματος (738 άτομα).

Μεταβλητή selfemp

Ολοκληρώνουμε την μελέτη των ποιοτικών μεταβλητών με την μεταβλητή selfemp, κατασκευάζοντας τους πίνακες συχνότητας και σχετικής συχνότητας, ούτως ώστε να δούμε ποιο ποσοστό του δείγματος αποτελείται από αυτοαπασχολούμενους.

Στη συνέχεια, συνδυάζουμε τους δύο πίνακες και προκύπτει ο πίνακας συχνοτήτων και σχετικών συχνοτήτων της μεταβλητής selfemp.

	selfemp	Frequency	Relative Frequency
1	no	1228	0.93100834
2	yes	91	0.06899166



Με αυτόν τον πίνακα, μπορούμε να βγάλουμε τα συμπεράσματα μας όσον αφορά στην μεταβλητή selfemp. Το 7% περίπου του δείγματος (91 άτομα) είναι αυτοαπασχολούμενοι, το οποίο δεν ισχύει για το 93% του δείγματος (1228 άτομα).

Έλεγχος εξάρτησης μεταβλητών-Pearson's Chi-Squared test

Ο έλεγχος x^2 (Pearson chi-square) αποτελεί επαγωγικό έλεγχο μέσω του οποίου ελέγχουμε την υπόθεση ότι οι δύο μεταβλητές του πίνακα συνάφειας είναι ανεξάρτητες μεταξύ τους.

Οι προς διερεύνηση υποθέσεις είναι οι ακόλουθες:

H_0 = Οι μεταβλητές που εξετάζουμε είναι ανεξάρτητες (Μηδενική υπόθεση)
H_1 = Οι μεταβλητές που εξετάζουμε ΔΕΝ είναι ανεξάρτητες (Εναλλακτική υπόθεση)

Εφαρμόζουμε x^2 test σε επίπεδο σημαντικότητας α=0,05
Αν το p-value για το x^2 test < 0,05=α, τότε απορρίπτουμε την H_0 και συνεπώς οι μεταβλητές ΔΕΝ είναι ανεξάρτητες.
Αν το p-value για το x^2 test > 0,05=α, τότε ΔΕΝ απορρίπτουμε την H_0 και συνεπώς οι μεταβλητές είναι ανεξάρτητες.

Αρχικά, εξετάζουμε τις μεταβλητές card και owner:

```
> chisq.test(dok1)

        Pearson's Chi-squared test with Yates' continuity
        correction

data:  dok1
X-squared = 28.114, df = 1, p-value = 1.144e-07
```

Από τον έλεγχο προκύπτει ότι p-value<0.05, συνεπώς η μηδενική υπόθεση απορρίπτεται και οι μεταβλητές card και owner δεν είναι ανεξάρτητες, δηλαδή η απάντηση του αιτήματος για νέα πιστωτική κάρτα εξαρτάται από την κατοχή ή μη της προσωπικής κατοικίας του ατόμου.

Συνεχίζουμε με τις μεταβλητές card και selfemp:

```
> chisq.test(dok2)

        Pearson's Chi-squared test with Yates' continuity
        correction

data:  dok2
X-squared = 3.3979, df = 1, p-value = 0.06528
```

Από τον έλεγχο προκύπτει ότι p-value>0.05, οπότε η μηδενική υπόθεση δεν απορρίπτεται, δηλαδή οι μεταβλητές card και selfemp είναι ανεξάρτητες, δηλαδή η απάντηση του αιτήματος για νέα πιστωτική κάρτα δεν εξαρτάται από το αν το άτομο είναι αυτοαπασχολούμενος (ελεύθερος επαγγελματίας) ή όχι.

Τέλος, εξετάζουμε τις μεταβλητές owner και selfemp:

```
> chisq.test(dok3)

        Pearson's Chi-squared test with Yates' continuity
        correction

data:  dok3
X-squared = 1.9714, df = 1, p-value = 0.1603
```

Από τον έλεγχο προκύπτει ότι p-value>0.05, επομένως η μηδενική υπόθεση δεν απορρίπτεται, δηλαδή οι μεταβλητές owner και selfemp είναι ανεξάρτητες, δηλαδή η κατοχή της προσωπικής κατοικίας του ατόμου δεν εξαρτάται από την κατηγορία του εργασιακού του προφίλ (αυτοαπασχολούμενος ή μη).

Ανάλυση ποσοτικών μεταβλητών

Οι ποσοτικές μεταβλητές του συνόλου δεδομένων είναι οι ακόλουθες :

- reports : ο αριθμός των αρνητικών αναφορών που έχει δεχθεί κάποιος
- age : η ηλικία του ατόμου + 12τα του έτους
- income : το ετήσιο εισόδημα του ατόμου
- share : αναλογία εξόδων πιστωτικών καρτών με το ετήσιο εισόδημα
- expenditure : μέσος όρος μηνιαίων εξόδων πιστωτικών καρτών
- dependents : αριθμός χρεών του ατόμου
- months : μήνες παραμονής στην ίδια διεύθυνση κατοικίας
- majorcards : αριθμός πιστωτικών καρτών που κατέχει το άτομο
- active : ο αριθμός των ενεργών λογαριασμών πιστωτικών καρτών

Προκειμένου να κατανοήσουμε το σύνολο δεδομένων ***CreditCard***, θα χρειαστεί να αναλύσουμε τις ποσοτικές παραμέτρους, δίνοντας κάποιες σημαντικές τιμές, με σκοπό να τις περιγράψουμε επαρκώς.

Μέτρα Κεντρικής Τάσης

Κεντρική τάση μιας κατανομής είναι η τάση που εμφανίζουν οι τιμές της κατανομής να συσσωρεύονται γύρω από κάποιο κεντρικό σημείο της. Τα μέτρα κεντρικής τάσης στοχεύουν στον προσδιορισμό αυτής της τάσης. Οι τιμές που χαρακτηρίζουν το μέτρο κεντρικής τάσης είναι η μέση τιμή, η διάμεσος, η επικρατούσα τιμή (αριθμός που εμφανίζεται περισσότερες φορές στο δείγμα) και τα εκατοστημόρια.

Μεταβλητές	Μέση τιμή	Διάμεσος	Επικρατούσα τιμή (συχνότητα)	Εκατοστημόρια (25%-75%)
reports	0.4564064	0	0 (1060 φορές)	0-0
age	33.2131	31.25	28.16667 (13 φορές)	25.41667-39.41667
income	3.365376	2.9	3 (61 φορές)	2.24375-4.00000
share	0.06873217	0.03882722	0.00048 (12 φορές)	0.002315922-0.093616825
expenditure	185.0571	101.2983	0 (317 φορές)	4.583333-249.035800
dependents	0.9939348	1	0 (659 φορές)	0-2
months	55.26763	30	12 (101 φορές)	12-72
majorcards	0.8172858	1	1 (1078 φορές)	1-1
active	6.996967	6	0 (219 φορές)	2-11

Μέτρα Διασποράς

Τα μέτρα διασποράς στοχεύουν στο προσδιορισμό της μεταβλητότητας (ή ετερογένειας) που παρουσιάζει ένα σύνολο μετρήσεων. Τα μέτρα αυτά χρησιμοποιούνται σε συνδυασμό με τα μέτρα θέσης και από κοινού περιγράφουν τις κατανομές δεδομένων με τρόπο συμπληρωματικό. Οι τιμές που θα χρειαστεί να υπολογίσουμε είναι το εύρος, η διασπορά και η τυπική απόκλιση.

Μεταβλητές	Εύρος (min-max)	Διασπορά	Τυπική Απόκλιση
reports	0 - 14	1.809745	1.345267
age	0.1666667 - 83.5000000	102.8761	10.14278
income	0.21 - 13.50	2.869303	1.693902
share	0.0001090909 - 0.9063205000	0.008959676	0.09465557
expenditure	0 - 3099.505	74103.14	272.2189
dependents	0 - 6	1.556868	1.247745
months	0 - 540	4391.944	66.27175
majorcards	0 - 1	0.149443	0.3865786
active	0 - 46	39.76327	6.305812

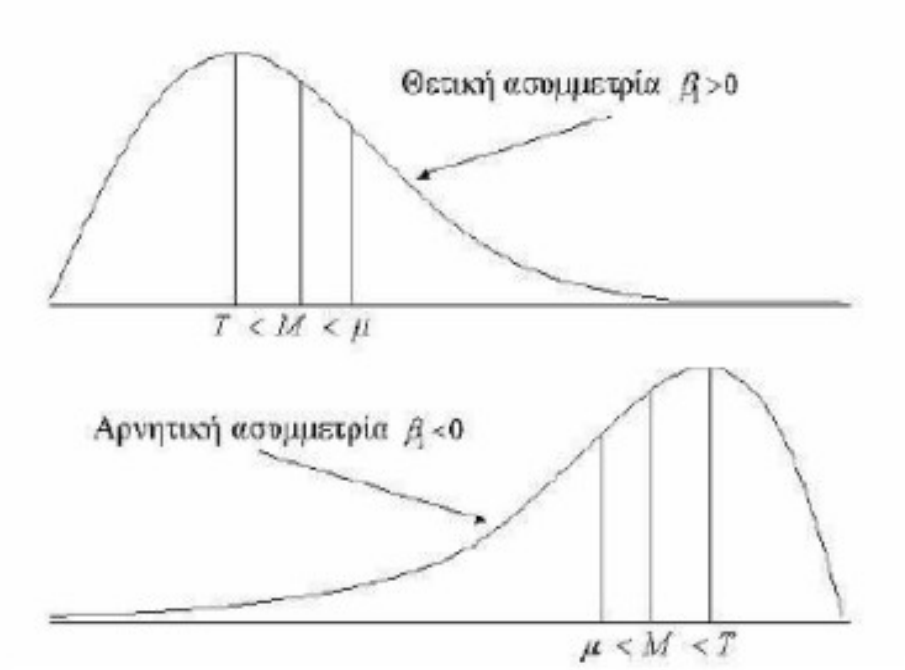
Συντελεστές Λοξότητας-Κύρτωσης

Ο συντελεστής λοξότητας (skewness) δείχνει την έκταση στην οποία μια κατανομή τιμών αποκλίνει από τη συμμετρία, γύρω από το μέσο όρο, ενώ ο συντελεστής κύρτωσης (kyrtosis) δείχνει τον βαθμό συγκέντρωσης των τιμών γύρω από το κέντρο ή το μέσο μιας κατανομής.

Μεταβλητές	Λοξότητα	Κύρτωση
reports	19.17176	143.0051
age	1920.281	57741.3
income	16.13382	78.15537
share	3.163933	16.1602
expenditure	95049017	143623162292
dependents	5.566261	13.38471
months	1037481	258848405
majorcards	-1.677252	0.795753
active	555.4975	10021.95

Η τιμή του συντελεστή λοξότητας μετράει το βαθμό της συμμετρίας των δεδομένων ως προς τη συχνότητά – κατανομή τους γύρω από τη μέση τιμή. Ταυτόχρονα, μας βοηθάει στην επαλήθευση των προηγούμενων αποτελεσμάτων, καθώς καθορίζει την ασυμμετρία της κατανομής κάθε μεταβλητής και χρησιμοποιείται συχνά για τη σύγκριση της μέσης τιμής με τη διάμεσο και την επικρατούσα τιμή, δηλαδή

- Αν η κατανομή μίας μεταβλητής έχει <u>αρνητική ασυμμετρία</u> (λοξότητα) ,τότε ισχύει :
  Μέση τιμή (δειγματικός μέσος) < Διάμεσος < Επικρατούσα τιμή
- Αν η κατανομή μίας μεταβλητής έχει <u>μηδενική ασυμμετρία</u> (λοξότητα), τότε είναι συμμετρική και ισχύει ότι :
  Μέση τιμή (δειγματικός μέσος) = Διάμεσος = Επικρατούσα τιμή
- Αν η κατανομή μίας μεταβλητής έχει <u>θετική ασυμμετρία</u> (λοξότητα), τότε ισχύει ότι:
  Μέση τιμή (δειγματικός μέσος) > Διάμεσος > Επικρατούσα τιμή

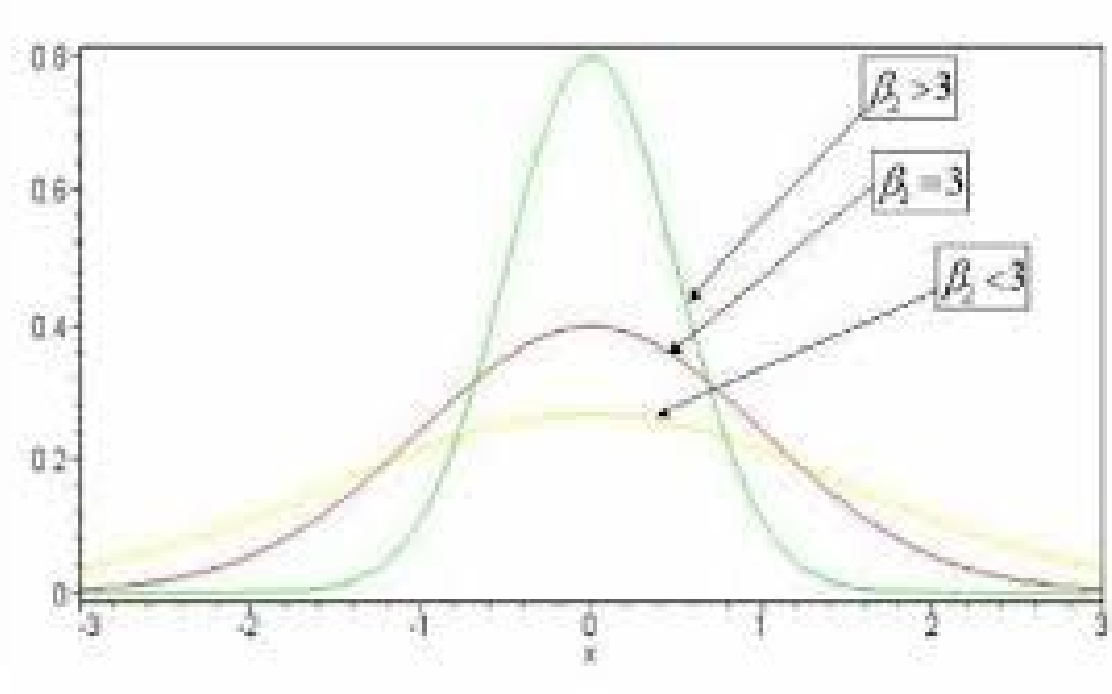


Στην παραπάνω εικόνα έχουμε

i. *β1 = συντελεστής λοξότητας*
ii. *μ = μέση τιμή (δειγματικός μέσος)*
iii. *M = διάμεσος*
iv. *T = επικρατούσα τιμή*

Η τιμή της κύρτωσης μετράει το βαθμό συγκέντρωσης των δεδομένων γύρω από τη μέση τιμή Επιπλέον, μας βοηθάει στην κατασκευή της γραφικής απεικόνισης της κατανομής των ποσοτικών μεταβλητών, καθώς δείχνει την αιχμηρότητα ή την πλάτυνση της κατανομής. Η γραφική απεικόνιση της χωρίζεται σε 3 μορφές :

- Αν **ο συντελεστής κύρτωσης** μίας μεταβλητής έχει τιμή μικρότερη του 3, η κατανομή λέγεται πλατύκυρτη,
- αν **ο συντελεστής κύρτωσης** μίας μεταβλητής έχει τιμή ίση με 3, η κατανομή λέγεται μεσόκυρτη, ενώ
- αν **ο συντελεστής κύρτωσης** μίας μεταβλητής έχει τιμή μεγαλύτερη του 3, η κατανομή λέγεται λεπτόκυρτη.



Στην παραπάνω εικόνα έχουμε

i. *β2 = συντελεστής κύρτωσης*

Διαγράμματα Διασποράς-Συντελεστής Συσχέτισης

Η παρουσίαση των ποσοτικών μεταβλητών σε διαγράμματα διασποράς είναι χρήσιμη, καθώς συγκεντρώνουμε όσες και όποιες μεταβλητές χρειαζόμαστε από το dataset ***CreditCard***, με σκοπό την ομαδοποίηση τους. Έτσι, εξετάζουμε την πιθανή συσχέτισή τους.

Εκτός από το διάγραμμα διασποράς, χρήσιμο εργαλείο αποτελεί και ο συντελεστής συσχέτισης ρ. Ο συντελεστής γραμμικής συσχέτισης ρ δίνει ένα μέτρο του μεγέθους της γραμμικής συσχέτισης μεταξύ δύο μεταβλητών και παίρνει τιμές στο κλειστό διάστημα [-1 , 1]. Επίσης, ισχύουν οι ακόλουθες ιδιότητες :

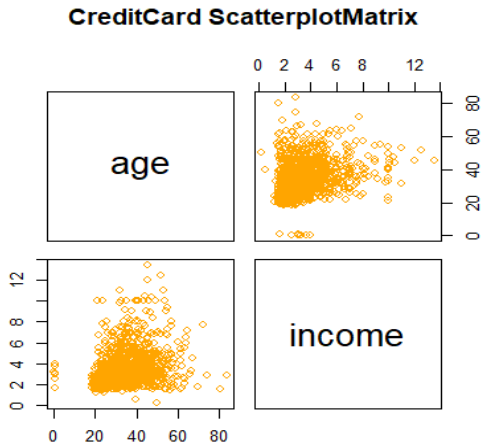
- Αν $\rho = \pm 1$, υπάρχει τέλεια γραμμική συσχέτιση
- Αν $-0,3 \leq \rho < 0,3$ δεν υπάρχει γραμμική συσχέτιση
- Αν $-0,5 < \rho \leq -0,3$ ή $0,3 \leq \rho < 0,5$ υπάρχει ασθενής γραμμική συσχέτιση
- Αν $-0,7 < \rho \leq -0,5$ ή $0,5 \leq \rho < 0,7$ υπάρχει μέση γραμμική συσχέτιση
- Αν $-0,8 < \rho \leq -0,7$ ή $0,7 \leq \rho < 0,8$ υπάρχει ισχυρή γραμμική συσχέτιση
- Αν $-1 < \rho \leq -0,8$ ή $0,8 \leq \rho < 1$ υπάρχει πολύ ισχυρή γραμμική συσχέτιση

Θετικές τιμές του ρ δεν υποδηλώνουν, κατ' ανάγκην μεγαλύτερο βαθμό γραμμικής συσχέτισης από το βαθμό γραμμικής συσχέτισης που υποδηλώνουν αρνητικές τιμές του ρ. Το πρόσημο του ρ καθορίζει το είδος, μόνο, της συσχέτισης (θετική ή αρνητική). Μας πληροφορεί δηλαδή για το αν αύξηση της μιας μεταβλητής αντιστοιχεί σε αύξηση ή σε μείωση της άλλης μεταβλητής.

Ιδιαίτερα σημαντικό στην κατανόηση της μεθόδου που χρησιμοποιούμε για την εξαγωγή αποτελεσμάτων όσον αφορά στις ποσοτικές μεταβλητές είναι ότι συσχέτιση δε σημαίνει αιτιότητα. Όταν σε μια μη πειραματική έρευνα (δειγματοληψία) δύο μεταβλητές βρίσκονται συσχετισμένες αυτό σημαίνει μόνο ότι οι μεταβλητές αυτές συνδέονται με κάποια σχέση. Δε συνεπάγεται, κατ' ανάγκη, αιτιότητα.

Λαμβάνοντας υπόψιν αυτές τις σχέσεις, εργαζόμαστε για τις ποσοτικές μεταβλητές του συνόλου μας ανά ομάδες.

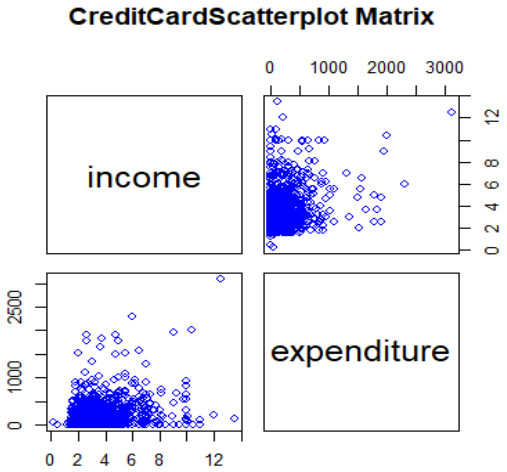
Αρχικά, εξετάζουμε τις μεταβλητές age και income :



Βρίσκουμε ρ(age, income) = 0.3246532 , άρα έχουμε ασθενή και θετική γραμμική συσχέτιση των μεταβλητών μας. Αυτό σημαίνει ουσιαστικά πως η αύξηση της ηλικίας των ατόμων του δείγματος συνεπάγεται αύξηση του ετήσιου εισοδήματος.

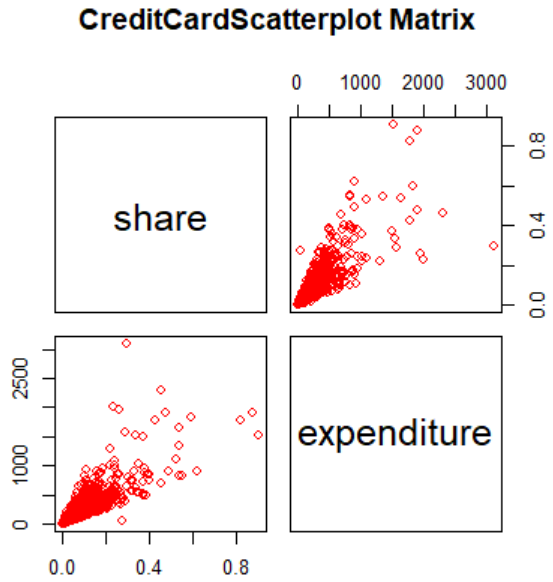
Αυτό φαίνεται πιο ξεκάθαρα με την ευθεία γραμμικής παλινδρόμησης που συνδέει τις δύο μεταβλητές, y= 1.56460 + 0.05422x , με y = ετήσιο εισόδημα (income) και x = ηλικία του ατόμου (age).

Συνεχίζουμε με τις μεταβλητές income και expenditure :



Βρίσκουμε ρ(income, expenditure) = 0.281104 , δηλαδή δεν έχουμε κάποια μορφή γραμμικής συσχέτισης μεταξύ των συγκεκριμένων μεταβλητών.

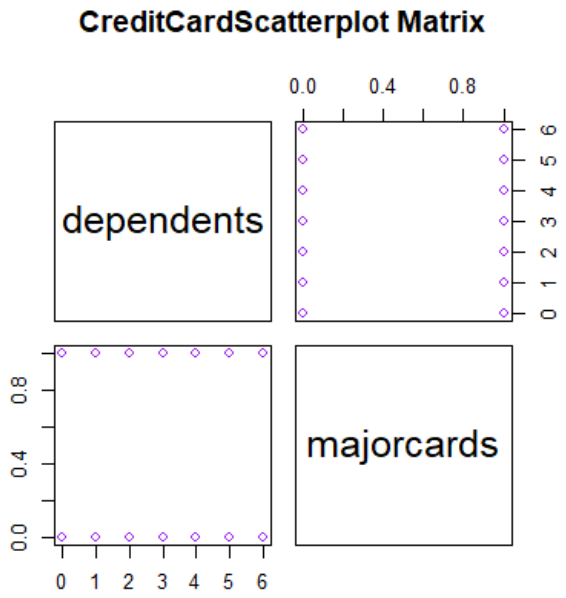
Σειρά έχουν οι μεταβλητές share και expenditure :



Βρίσκουμε ρ(share, expenditure) = 0.8387793 , δηλαδή έχουμε πολύ ισχυρή γραμμική συσχέτιση μεταξύ των συγκεκριμένων μεταβλητών. Αυτό σημαίνει ότι η αύξηση της αναλογίας μηνιαίων εξόδων με χρήση πιστωτικής κάρτας - ετήσιου εισοδήματος του ατόμου συνεπάγεται την αύξηση του μέσου όρου των μηνιαίων εξόδων των πιστωτικών καρτών του .

Κάτι που αποτυπώνεται και στην ευθεία γραμμικής παλινδρόμησης που συνδέει τις δύο μεταβλητές, y =19.26 + 2412.24x , με y = μέσος όρος των μηνιαίων εξόδων των πιστωτικών καρτών (expenditure) και x = αναλογία μηνιαίων εξόδων με χρήση πιστωτικής κάρτας - ετήσιου εισοδήματος του ατόμου (share).

Τέλος, θα εξετάσουμε τις μεταβλητές dependents και majorcards :



Βρίσκουμε ρ(dependents, majorcards) = 0.01028454, δηλαδή δεν έχουμε κάποια μορφή γραμμικής συσχέτισης μεταξύ των συγκεκριμένων μεταβλητών.

Ανάλυση Διασποράς/Διακύμανσης

Η ανάλυση διασποράς ή ανάλυση διακύμανσης (**ANOVA**) αποτελεί μια από τις πιο σημαντικές μεθόδους για την ανάλυση δεδομένων. Ορίζεται ως μία στατιστική μέθοδος με την οποία η μεταβλητότητα που υπάρχει σε ένα σύνολο δεδομένων διασπάται στις επιμέρους συνιστώσες της με στόχο την κατανόηση της σημαντικότητας των διαφορετικών πηγών προέλευσης της. Για να δοθεί απάντηση στο συγκεκριμένο ερώτημα, κατασκευάζουμε έναν έλεγχο υποθέσεων με μηδενική υπόθεση H_0 ότι όλα τα δείγματα προέρχονται από πληθυσμούς με την ίδια μέση τιμή έναντι μιας εναλλακτικής υπόθεσης ότι τουλάχιστον δύο μέσες τιμές είναι διαφορετικές. Ουσιαστικά πρόκειται για μια γενίκευση του T-test που εφαρμόζεται σε δύο πληθυσμούς. Θεωρητικά, θα μπορούσαν να εφαρμοστούν πολλαπλοί ανεξάρτητοι έλεγχοι, αλλά η συγκεκριμένη μεθοδολογία δεν ενδείκνυται καθότι με αυτό τον τρόπο αυξάνεται η πιθανότητα να οδηγηθούμε σε σφάλμα τύπου Ι. Συνεπώς, η **ANOVA** είναι η κατάλληλη μεθοδολογία διότι, πρόκειται για συντομότερη διαδικασία ανάλυσης ενώ έχει και ακρίβεια διάγνωσης

Οι προϋποθέσεις που απαιτούνται για την εφαρμογή της ανάλυση διακύμανσης, είναι οι εξής :

- η κατανομή των τιμών να είναι κανονική,
- τα δείγματα να είναι αντιπροσωπευτικά και οι παρατηρήσεις ανεξάρτητες μεταξύ τους,
- οι πληθυσμοί από τους οποίους επελέγησαν τα δείγματα να έχουν την ίδια διακύμανση

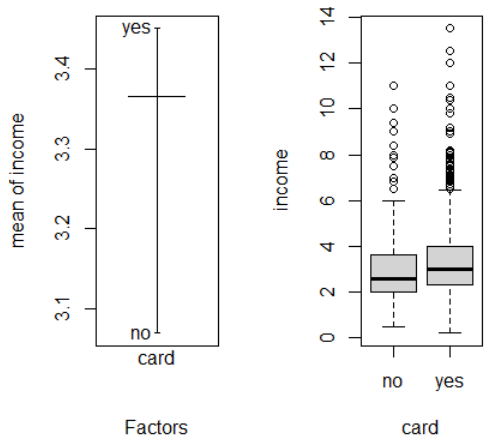
Η μορφή ελέγχου **ANOVA** που θα δούμε παρακάτω ονομάζεται ανάλυση διακύμανσης κατά ένα παράγοντα (**one-way-ANOVA**). Ουσιαστικά, πρόκειται για ένα πλήρως τυχαιοποιημένο σχεδιασμό, σύμφωνα με τον οποίο εργαζόμαστε με n ανεξάρτητα τυχαία δείγματα, ένα από κάθε πληθυσμό (έναν από κάθε στάθμη του παράγοντα) κάτι που αποτελεί γενίκευση του ελέγχου των μέσων τιμών που προκύπτουν από ανεξάρτητα τυχαία δείγματα.

Θα εξετάσουμε αν η μεταβλητή card (η απάντηση για την αίτηση έκδοσης πιστωτικής κάρτας) επηρεάζεται από την μεταβλητή income (ετήσιο εισόδημα του κάθε ατόμου).

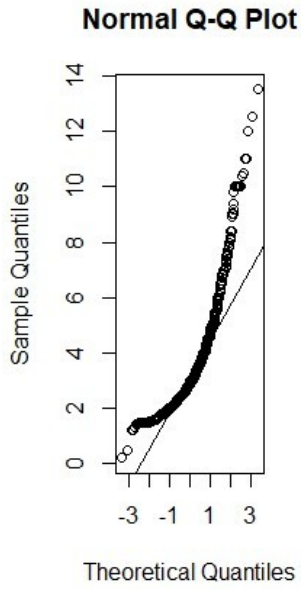
Αρχικά, δημιουργούμε ένα πλαίσιο δεδομένων και αναλύουμε γραφικά το νέο σύνολο που περιέχει τις παρατηρήσεις για τις δύο μεταβλητές που επιλέξαμε.

```
> income.df<-data.frame(card,income)
> par(mfrow=c(1,2))
> plot.design(income.df)
> boxplot(income~card,income.df)
```

Το γράφημα που κατασκευάζουμε απεικονίζει τα στατιστικά μέτρα των δύο μεταβλητών μας με παράμετρο τη μέση τιμή της ποσοτικής μεταβλητής income.



Έπειτα, ελέγχουμε αν το πλαίσιο δεδομένων το οποίο μελετάμε ακολουθεί κανονική κατανομή με τον έλεγχο Kolmogorov-Smirnov.




```
> ks.test(income.df$income,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  income.df$income
D = 0.92789, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Τελικά, βγάζουμε το συμπέρασμα με την ανάλυση διασποράς και τον πίνακα ANOVA.

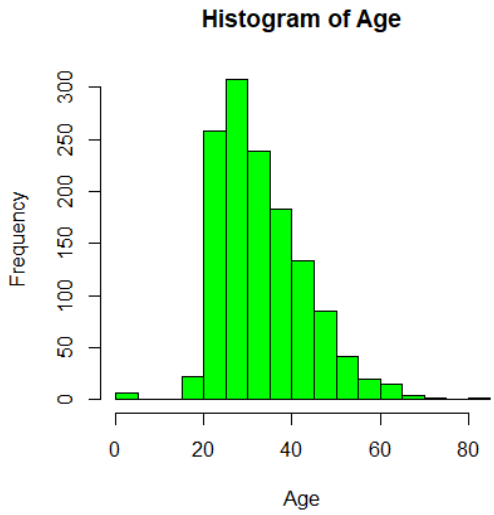
```
> aov.income<-aov(income~card,income.df)
> aov.income
Call:
   aov(formula = income ~ card, data = income.df)

Terms:
                    card Residuals
Sum of Squares    33.634  3748.107
Deg. of Freedom        1      1317

Residual standard error: 1.686992
Estimated effects may be unbalanced
> summary(aov.income)
              Df Sum Sq Mean Sq F value   Pr(>F)
card           1     34   33.63   11.82 0.000605 ***
Residuals   1317   3748    2.85
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

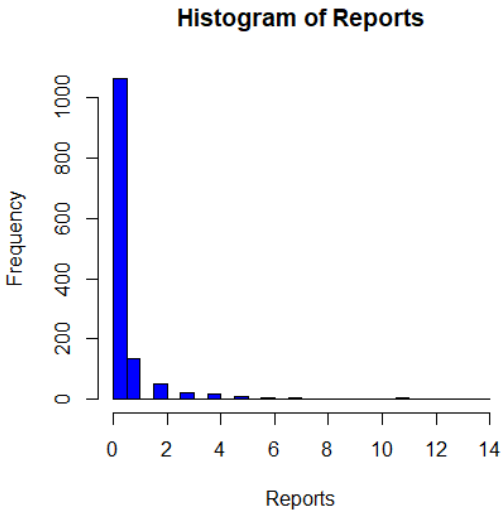
Άρα, έχουμε p-value<0.05 ,επομένως η μηδενική υπόθεση (ισότητα για τις μέσες τιμές) απορρίπτεται.

```
> anova(aov.income)
Analysis of Variance Table

Response: income
            Df Sum Sq Mean Sq F value    Pr(>F)
card         1   33.6  33.634  11.818 0.0006048 ***
Residuals 1317 3748.1   2.846
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

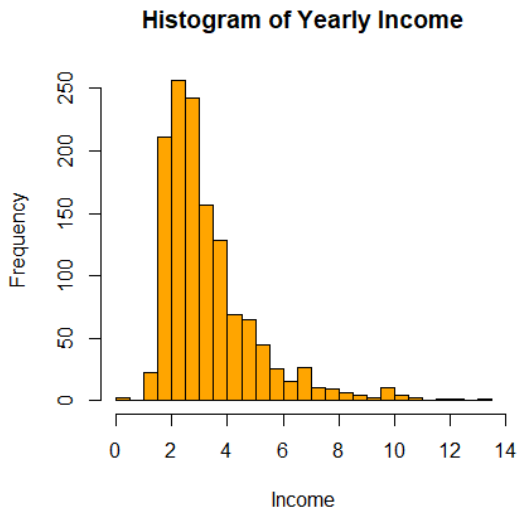
Ραβδογράμματα ποσοτικών μεταβλητών

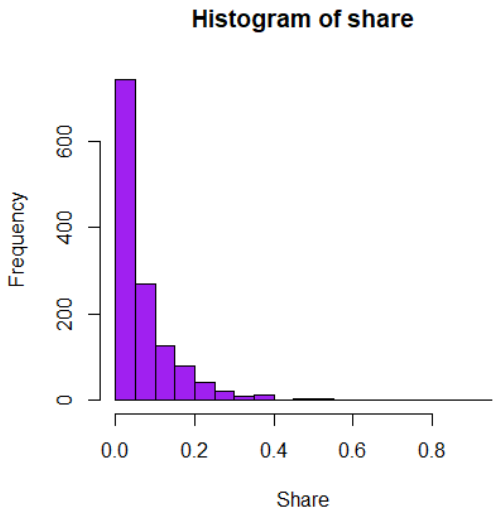
Η απεικόνιση των ποσοτικών μεταβλητών σε ραβδογράμματα μας παρουσιάζει τη συχνότητα εμφάνισης όλων των τιμών που παίρνει η κάθε μεταβλητή.



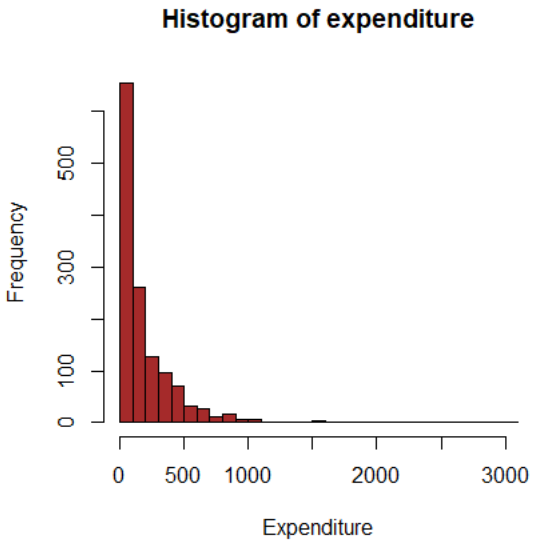


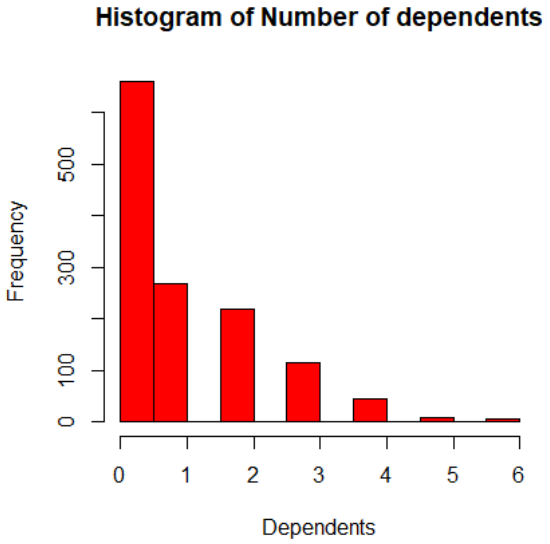
Από το ραβδόγραμμα της μεταβλητής **age**, φαίνεται ότι η ηλικία των περισσότερων ατόμων του δείγματος κυμαίνεται μεταξύ των 25 με 35 χρονών. Αντίστοιχα, τα περισσότερα άτομα δεν έχουν δεχτεί καμία αρνητική αναφορά, όπως υποδεικνύει το ραβδόγραμμα της μεταβλητής **reports**.



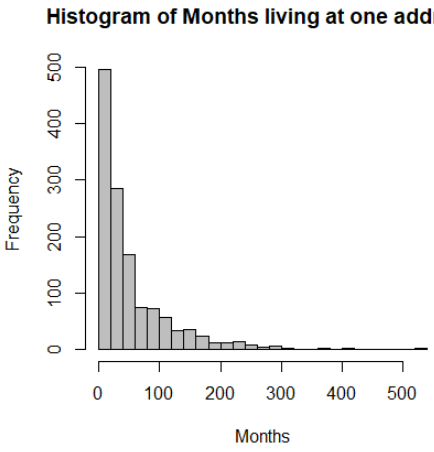


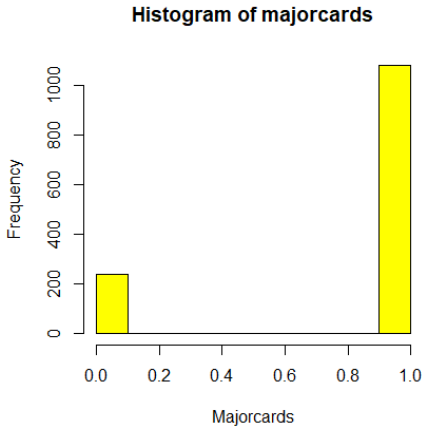
Από το ραβδόγραμμα της μεταβλητής **income**, φαίνεται ότι τα περισσότερα άτομα του δείγματος έχουν ετήσιο εισόδημα 20000 με 40000 USD (αμερικάνικα δολάρια). Αντίστοιχα, τα περισσότερα άτομα ξοδεύουν ελάχιστα ως και καθόλου χρήματα με τις πιστωτικές τους κάρτες, κάτι που υποδεικνύει το ιστόγραμμα της μεταβλητής **share** (αναλογία μηνιαίων εξόδων πιστωτικών καρτών και ετήσιου εισοδήματος).



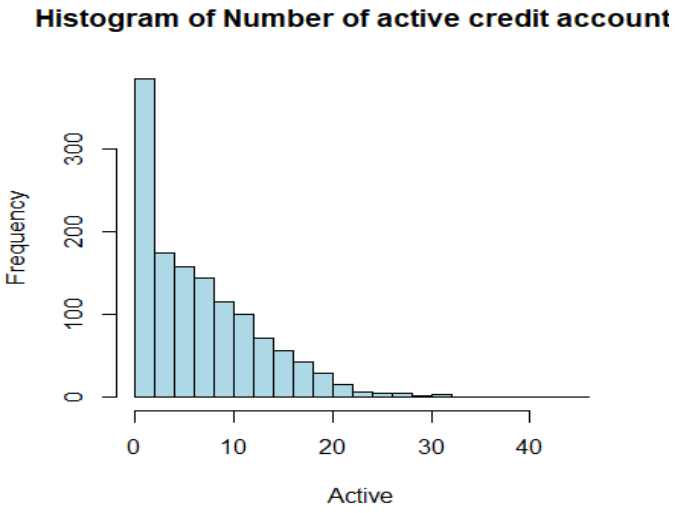


Από το ραβδόγραμμα της μεταβλητής **expenditure** (μέσος όρος μηνιαίων εξόδων των πιστωτικών καρτών του χρήστη) , φαίνεται ότι τα περισσότερα άτομα του δείγματος δεν ξοδεύουν ελάχιστα ως και καθόλου χρήματα με την πιστωτική τους κάρτας ανά μήνα. Αντίστοιχα, τα περισσότερα άτομα δεν έχουν πρόβλημα με πιστωτικά χρέη, όπως είναι ξεκάθαρο στο ραβδόγραμμα της μεταβλητής **dependents**.





Από το ραβδόγραμμα της μεταβλητής **months** (μήνες παραμονής στην ίδια διεύθυνση κατοικίας) , φαίνεται πως το μεγαλύτερο πλήθος των ατόμων του δείγματος δεν μένει στην ίδια διεύθυνση κατοικίας για μεγάλο διάστημα μηνών-λιγότερο από έτος. Αντίστοιχα, η πλειονότητα των ατόμων που αποτελούν το δείγμα μας έχει μόνο μία πιστωτική κάρτα στην κατοχή του, καθώς είναι ξεκάθαρο στο ραβδόγραμμα της μεταβλητής **majorcards** (πλήθος πιστωτικών καρτών που έχει το άτομο στην κατοχή του-όχι απαραίτητα υπό χρήση).



Το τελευταίο ραβδόγραμμα ,αυτό της μεταβλητής **active** (πλήθος ενεργών πιστωτικών λογαριασμών) μας δείχνει ότι ένα μεγάλο μέρος του δείγματος δεν έχει κανέναν κανένα ενεργό πιστωτικό λογαριασμό. Γενικά, τα περισσότερα άτομα έχουν μικρό αριθμό ενεργών πιστωτικών λογαριασμών, πιο συγκεκριμένα λιγότερο από 10.

Με βάση τις παρατηρήσεις που κάναμε για τις ποσοτικές μεταβλητές του δείγματος ***CreditCard*** που μελετάμε, μπορούμε να εξάγουμε ένα γενικό συμπέρασμα για την πλειονότητα των ατόμων που αποτελούν το σύνολο δεδομένων μας.

Το σύνολο μας αποτελείται κατά κύριο λόγο από νέους ανθρώπους (25- 35 χρονών), με ετήσιο εισόδημα 20000-40000 USD (αμερικάνικα δολάρια) και θετικό ιστορικό συναλλαγών, δηλαδή χωρίς κάποια αρνητική αναφορά . Επιπλέον, τα έξοδα τους είναι ελάχιστα ως και μηδενικά με χρήση πιστωτικών καρτών, κάτι που δικαιολογείται άμεσα από το γεγονός πως η συντριπτική πλειονότητα του δείγματος δεν έχει στην κατοχή κάποια πιστωτική κάρτα (ενεργή ή και μη). Τέλος, οι περισσότεροι δεν παρουσιάζουν μακροχρόνια σταθερότητα στην κατοικίας τους, αφού το μεγαλύτερο μέρος του συνόλου μένει στην τωρινή διεύθυνση κατοικίας τους για μικρό χρονικό διάστημα.

Μετρικές Αξιολόγησης

Το σύνολο δεδομένων είναι αρκετά μεγάλο, όπως έχουμε δει , κάτι που το καθιστά συχνά δύσκολο στην μελέτη των μεταβλητών του. Γι' αυτό, θα υπολογίσουμε το μέσο τετραγωνικό σφάλμα των ποσοτικών μεταβλητών. Αυτό θα μας δώσει την "απόσταση" των πραγματικών τιμών που παίρνει η κάθε μεταβλητή στο σύνολο σε σχέση με τη μέση τιμή της.

```
> mse(CreditCard$reports,mean(CreditCard$reports))
[1] 1.808373
> mse(CreditCard$age,mean(CreditCard$age))
[1] 102.7981
> mse(CreditCard$income,mean(CreditCard$income))
[1] 2.867128
> mse(CreditCard$share,mean(CreditCard$share))
[1] 0.008952883
> mse(CreditCard$expenditure,mean(CreditCard$expenditure))
[1] 74046.96
> mse(CreditCard$dependents,mean(CreditCard$dependents))
[1] 1.555687
> mse(CreditCard$months,mean(CreditCard$months))
[1] 4388.615
> mse(CreditCard$majorcards,mean(CreditCard$majorcards))
[1] 0.1493297
> mse(CreditCard$active,mean(CreditCard$active))
[1] 39.73312
```

```
> mean(CreditCard$reports)
[1] 0.4564064
> mean(CreditCard$age)
[1] 33.2131
> mean(CreditCard$income)
[1] 3.365376
> mean(CreditCard$share)
[1] 0.06873217
> mean(CreditCard$expenditure)
[1] 185.0571
> mean(CreditCard$dependents)
[1] 0.9939348
> mean(CreditCard$months)
[1] 55.26763
> mean(CreditCard$majorcards)
[1] 0.8172858
> mean(CreditCard$active)
[1] 6.996967
```

Μέθοδος Βελτιστοποίησης Μοντέλου

Προκειμένου να βελτιστοποιήσουμε το μοντέλο μας, προκύπτει το ζήτημα ελαχιστοποίησης αυτής της απόστασης. Η απάντηση δίνεται με την εκπαίδευση ενός μέρους του δείγματος (επιλέγουμε το 80%) και διακρίνουμε το σύνολο που έχει τιμές με την μικρότερη διαφορά από τη μέση τιμή.

Η διαδικασία διαίρεσης του dataset γίνεται με τα εξής βήματα :

- Χωρίζουμε τα δεδομένα σε training validation και test,
- Επιλέγουμε μέθοδο στατιστικής μάθησης και παραμέτρους εκπαίδευσης,
- Εκπαιδεύουμε το σύστημα με το training set,
- Αξιολογούμε το σύστημα με το validation set,
- Επαναλαμβάνουμε τα βήματα 2-4 για διαφορετικές μεθόδους στατιστικής μάθησης και παραμέτρους,
- Επιλέγουμε το καλύτερο μοντέλο και το εκπαιδεύουμε με τα training και τα validation sets,
- Αξιολογούμε το σύστημα με το test set

Μέθοδοι Αναδειγματοληψίας

Η αναδειγματοληψία του συνόλου δεδομένων μας γίνεται είτε με τη μέθοδο Hold-out , είτε με την διασταυρωμένη επικύρωση Cross-Validation, είτε με τη μέθοδο Leave-one-out Cross-Validation

Μέθοδος Hold-out

Η μέθοδος Hold-out διαχωρίζει το σύνολο δεδομένων σε δύο τυχαία και ανεξάρτητα υποσύνολα. Στη συνέχεια, ένα υποσύνολο χρησιμοποιείται για την εκπαίδευση του μοντέλου και ονομάζεται σύνολο εκπαίδευσης (training set) και το άλλο υποσύνολο ονομάζεται σύνολο επικύρωσης (validation set), που χρησιμοποιείται για την επικύρωση του μοντέλου.

Πρώτα, το μοντέλο εκπαιδεύεται με το σύνολο εκπαίδευσης και στη συνέχεια δοκιμάζεται η απόδοσή του στο δεύτερο υποσύνολο επικύρωσης. Η αξιολόγηση της απόδοσης του μοντέλου πραγματοποιείται μέσω της σύγκρισης των μέσων τετραγωνικών σφαλμάτων, δηλαδή επιλέγουμε το υποσύνολο με το μικρότερο μέσο τετραγωνικό σφάλμα.

Μέθοδος K-Fold Cross-Validation

Για την εφαρμογή της μεθόδου K-Fold Cross Validation διαιρούμε το αρχικό δείγμα σε k ίσα υποσύνολα (k-folds). Κάθε φορά ένα υποσύνολο χρησιμοποιείται για έλεγχο και τα υπόλοιπα για εκπαίδευση. Επαναλαμβάνουμε αυτή τη διαδικασία k φορές με κάθε υποσύνολο να ελέγχεται ακριβώς μία φορά. Όλες οι παρατηρήσεις χρησιμοποιούνται για εκπαίδευση και για έλεγχο, άρα κάθε παρατήρηση ελέγχεται ακριβώς μία φορά. Τελικά το αποτέλεσμα που προκύπτει θα είναι ο μέσος όρος των test errors των k υποσυνόλων.

Μέθοδος Leave-one-out Cross-Validation

Αυτή η μέθοδος διαφέρει αρκετά από τις προηγούμενες, καθώς αρχικά χωρίζουμε τα δεδομένα σε περισσότερα από δύο υποσύνολα, δημιουργώντας παράλληλα περισσότερα από ένα σύνολα εκπαίδευσης (αντίθετα από την Hold-out). Μετά, όμως, όλα τα υποσύνολα εκτός ένα μοντέλο θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου στατιστικής μάθησης, το οποίο θα προβλέψει την τιμή της παρατήρησης που δεν συμμετείχε στην εκπαίδευση.

ΠΡΟΣΟΧΗ : Το αποτέλεσμα ΔΕΝ θα είναι αξιόπιστο αν το τρέξουμε χωρίς διασταύρωση.

Εφαρμογή για τις μεταβλητές age και income

Θα χρησιμοποιήσουμε και τις τρεις μεθόδους παρακάτω με σκοπό να δούμε τη διεργασία που απαιτείται για την κάθε μία ξεχωριστά, προκειμένου να καταλήξουμε στο επιθυμητό αποτέλεσμα. Θα εργαστούμε για τις ποσοτικές μεταβλητές age (ηλικία του ατόμου) και income (ετήσιο εισόδημα) του dataset ***CreditCard***.

Μέθοδος Hold-out

Αρχικά, ορίζουμε το 50% των παρατηρήσεων του dataset μας ως σύνολο εκπαίδευσης (training set). Οι παρατηρήσεις που επιλέγονται για το σύνολο εκπαίδευσης είναι τυχαίες, καθώς το καθορίζουμε με την εντολή set.seed.

```
Call:
lm(formula = age ~ income, data = CreditCard, subset = train)

Residuals:
    Min      1Q  Median      3Q     Max
-33.332  -6.710  -1.955   5.333  51.117

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.3529     0.8659  31.590  < 2e-16 ***
income        1.7513     0.2262   7.743 3.69e-14 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.03 on 657 degrees of freedom
Multiple R-squared:  0.08362,   Adjusted R-squared:  0.08222
F-statistic: 59.95 on 1 and 657 DF,  p-value: 3.694e-14
```

Στη συνέχεια, εκτιμούμε την απάντηση για όλες τις παρατηρήσεις του dataset και αφαιρούμε από τη μεταβλητή age τις παρατηρήσεις που χρησιμοποιήθηκαν για την εκπαίδευση. Έτσι, βρίσκουμε το μέσο τετραγωνικό σφάλμα για το 50% των παρατηρήσεων του συνόλου επικύρωσης (validation set).

```
> predicted<-predict(lm.fit,CreditCard)[-train]
> mse(predicted,age[-train])
[1] 83.84295
```

Το μέσο τετραγωνικό σφάλμα του βέλτιστου μοντέλου που βρήκαμε είναι **MSE=83.84295**.

Μέθοδος Leave-one-out Cross-Validation

Για αυτή τη μέθοδο, θα χωρίσουμε το σύνολό μας σε σύνολο εκπαίδευσης (training set) που θα περιέχει το 80% των παρατηρήσεων του dataset που μελετάμε, ενώ το σύνολο επικύρωσης (validation set) θα περιέχει το υπόλοιπο 20% των παρατηρήσεων του dataset ***CreditCard***.

Έπειτα, εκπαιδεύουμε το σύνολο εκπαίδευσης με τη μέθοδο Cross-Validation.

```
Call:
glm(formula = age ~ income, data = CreditCard)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-33.863   -6.585   -2.092    5.173   51.245

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.6709     0.5879   45.36   <2e-16 ***
income        1.9440     0.1561   12.46   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 92.1028
3)

    Null deviance: 135591  on 1318  degrees of freedom
Residual deviance: 121299  on 1317  degrees of freedom
AIC: 9712.9

Number of Fisher Scoring iterations: 2
```

Άρα η εκτίμηση μας για το σφάλμα της μέτρησης είναι περίπου 92.25

Συνεχίζουμε όμως την εργασία για την εύρεση του βέλτιστου μοντέλου, καθώς ψάχνουμε το μοντέλο με το ελάχιστο σφάλμα.

```
> cv.error2=rep(0,5)
> for (i in 1:5){glm.fit2=glm(age~poly(income,i),data=CreditCa
rd)
+       cv.error2[i]=cv.glm(CreditCard,glm.fit2)$delta[1]}
> cv.error2
[1] 92.25824 91.40203 91.57591 91.61246 90.68029
```

Βλέπουμε ότι το βέλτιστο μοντέλο είναι το 5ο αφού έχει το μικρότερο σφάλμα.

Το μέσο τετραγωνικό σφάλμα του βέλτιστου μοντέλου που βρήκαμε είναι **MSE=92.14519**.

Μέθοδος K-Fold Cross-Validation

Η τρίτη και τελευταία μέθοδος που θα χρησιμοποιήσουμε μοιάζει με την προηγούμενη , απλώς ορίζουμε συγκεκριμένη τιμή για την παράμετρο K. Πιο συγκεκριμένα, θα ορίσουμε K=10.

```
Call:
glm(formula = age ~ income, data = CreditCard)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-33.863   -6.585   -2.092    5.173   51.245

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.6709     0.5879   45.36   <2e-16 ***
income        1.9440     0.1561   12.46   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 92.1028
3)

    Null deviance: 135591  on 1318  degrees of freedom
Residual deviance: 121299  on 1317  degrees of freedom
AIC: 9712.9

Number of Fisher Scoring iterations: 2

> cv.err0=cv.glm(CreditCard,glm.fit0,K=10)
> cv.err0$delta
[1] 92.11612 92.10811
```

Άρα η εκτίμηση μας για το σφάλμα της μέτρησης είναι περίπου 92.1 .

Έπειτα, συνεχίζουμε την διεργασία για την εύρεση του βέλτιστου μοντέλου.

```
> cv.error1=rep(0,10)
> for (i in 1:10){lm.fit=glm(age~poly(income,i),data=CreditCar
dTrain)
+       cv.error1[i]=cv.glm(CreditCardTrain,lm.fit)$delta[1]}
> cv.error1
 [1]  93.65192  92.82372  93.07266  93.52861  91.69624
 [6]  92.66825  92.00946 109.14015 448.96087 390.04039
> min(cv.error1)
[1] 91.69624
```

Εφόσον το 5ο μοντέλο έχει το μικρότερο σφάλμα, τότε είναι το βέλτιστο.

```
> glm.fit=glm(age~poly(income,5),data=CreditCardTest)
> predicted<-predict(glm.fit0,CreditCardTest)
> mse(predicted,CreditCardTest$age)
[1] 86.65453
```

Το μέσο τετραγωνικό σφάλμα του βέλτιστου μοντέλου που βρήκαμε είναι **MSE=86.65453**

Δέντρα ταξινόμησης

Τα δέντρα ταξινόμησης συνοψίζουν όλες τις διεργασίες που δείξαμε παραπάνω, καθώς γίνεται η κατηγοριοποίηση των παρατηρήσεων και με βάσει των κριτηρίων βελτιστοποίησης, κατασκευάζουμε το δέντρο ταξινόμησης για το σύνολο των δεδομένων.

Κατηγοριοποίηση

Η κατηγοριοποίηση των μεταβλητών αποτελεί ένα ακόμα πολύ σημαντικό κεφάλαιο στο πλαίσιο βελτιστοποίησης του αρχικού μας μοντέλου. Η διαδικασία που ακολουθούμε προϋποθέτει την κατανομή κάθε παρατήρησης σε ένα καθορισμένο σύνολο κατηγοριών.

Ο σκοπός αυτής της διαδικασία είναι η ανάπτυξη ενός μοντέλου, το οποίο μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση άγνωστων δεδομένων.

Οι κατηγορίες όπου αναθέτουμε τις παρατηρήσεις του συνόλου είναι προκαθορισμένες και συνήθως τις αναφέρουμε ως κλάσεις. Γενικά, το μοντέλο κατηγοριοποίησης χρησιμοποιείται ως περιγραφικό μοντέλο (descripitve modeling) ,για την επεξήγηση του συνόλου, και ως μοντέλο πρόβλεψης (predictive modeling), για την πρόβλεψη της κλάσης άγνωστων εγγραφών.

Το ζήτημα που καλούμαστε να αντιμετωπίσουμε κάθε φορά που επιχειρούμε να ταξινομήσουμε τα δεδομένα σε κατηγορίες είναι η επιλογή της συνθήκης διαχωρισμού. Δηλαδή, χρειάζεται να συγκρίνουμε τις κλάσεις στις οποίες ταξινομούνται τα δεδομένα του δείγματος, προκειμένου να επιλέξουμε την τελική. Συνήθως, οι κόμβοι με ομοιογενείς κατανομές είναι προτιμότεροι, με ιδανικό σενάριο να έχουμε ΟΛΕΣ τις εγγραφές στην ίδια κλάση.

Η κατανομή των εγγραφών σε μία κλάση ορίζεται και από τον βαθμό μη καθαρότητας, που υποδεικνύουν το **ευρετήριο Gini (Gini Index)**, **η εντροπία (Entropy)** και **το λάθος ταξινόμησης**. Όταν το μεγαλύτερο (ή σχεδόν όλο) το πλήθος των εγγραφών κατανέμεται σε μία κλάση, τότε έχουμε μεγάλο βαθμό μη καθαρότητας ενός κόμβου.

Τελικά, καθορίζονται ως προτιμότερες μορφές κόμβων αυτοί που μπορούν να χαρακτηριστούν, βάσει των παραπάνω ορισμών, ως ομοιογενείς, με μεγάλο βαθμό μη καθαρότητας.

Κατασκευή Δέντρου

Ο αλγόριθμος που ακολουθούμε για την κατασκευή του δέντρου έχει ως εξής :

1. Υπολογίζουμε το πληροφοριακό κέρδος κάθε μεταβλητής.
2. Θέτουμε ως ρίζα του δέντρου τη μεταβλητή με το μεγαλύτερο πληροφοριακό κέρδος.
3. Δημιουργούμε τόσα κλαδιά όσες και οι διακριτές τιμές της μεταβλητής.
4. Χωρίζουμε το σύνολο δεδομένων σε τόσα υποσύνολα όσα και οι διακριτές τιμές της μεταβλητής που επιλέχθηκε.
5. Επιλέγουμε μια τιμή-υποσύνολο, που δεν έχει ήδη επιλεχθεί. Αν στην τρέχουσα τιμή υποσύνολο αντιστοιχεί μόνο μια τιμή πηγαίνουμε στο βήμα 6 ,αλλιώς στο βήμα 7.
6. Βάζουμε την τιμή κλάσης ως φύλλο και προχωρούμε στην επόμενη τιμή μεταβλητή-υποσύνολο και επαναλαμβάνουμε το βήμα 5.
7. Υπολογίζουμε το πληροφοριακό κέρδος των υπολοίπων μεταβλητών για το συγκεκριμένο υποσύνολο.
8. Επιλέγουμε τη μεταβλητή με το μεγαλύτερο πληροφοριακό κέρδος και προσθέτουμε έναν νέο κόμβο στον κλάδο που αντιστοιχεί στην τρέχουσα τιμή-υποσύνολο.
9. Επαναλαμβάνουμε τη διαδικασία από το βήμα 3, μέχρι να μην μπορούν να δημιουργηθούν νέα φύλλα.

Το πληροφοριακό κέρδος ορίζεται ως **G(S , A) = E(S) - I(S , A)**.

- Όπου $I(S,A) = \sum_j \frac{|S_j|}{|S|} E(S_j)$,
  - όπου S_j είναι τα δείγματα με τιμή **j** για το χαρακτηριστικό **A**,
  - Όπου $|S_j|$ το πλήθος των δειγμάτων με τιμή **j** για το χαρακτηριστικό **A**
  - Όπου **S** όλα τα δείγματα και
  - Όπου $|S|$ το πλήθος όλου του δείγματος
  - $E(S_j)$ είναι η εντροπία για το υποσύνολο δειγμάτων του συνόλου δεδομένων με τίμη **j** για το χαρακτηριστικό A
  - Αν έχουμε k κλάσης στο σύνολο των δειγμάτων **S** η εντροπία δίνεται από

$$\mathrm{E(S)} = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

  - Όπου p_i είναι η πιθανότητα της κλάσης **i** στο **S**

Συνεχίζουμε με το ευρετήριο Gini. Το ευρετήριο Gini (Gini Index) υπολογίζει την ανισότητα μεταξύ των τιμών μιας κατανομής συχνοτήτων. Οι τιμές του είναι ανάμεσα στο 0 και 1, με το 0 να δηλώνει πλήρη ισότητα και το 1 να δηλώνει πλήρη ανισότητα.

Για ένα σύνολο δεδομένων S με m δείγματα και k κλάσεις, το gini(S) ορίζεται ως,

$$gini(S) = 1 - \sum_{j=1}^{k} p_j^2$$

όπου p*j* είναι η πιθανότητα εμφάνισης της κλάσης j στο σύνολο δεδομένων S.

Στην περίπτωση που το S διαχωριστεί σε S*1* και S*2* ,τότε το gini(S) ορίζεται ως,

$$gini(S) = \frac{n_1}{n} gini(S_1) + \frac{n_2}{n} gini(S_2)$$

όπου n1 και n2 είναι το σύνολο των δειγμάτων στο S_1 και S_2 αντίστοιχα.

Το πλεονέκτημα της μεθόδου αυτής είναι ότι για τον υπολογισμό απαιτείται μόνο ο διαχωρισμός των κλάσεων σε κάθε υποσύνολο. Το καλύτερο χαρακτηριστικό είναι εκείνο με τη μικρότερη τιμή Gini.

Τέλος, ορίζουμε το λάθος ταξινόμησης (classification error) ως **Error(t) = 1 – maxP(i|t)** ,ως ένα τύπο σφάλματος μέτρησης με το οποίο ο ερωτώμενος δεν δίνει αληθινή απάντηση σε ένα στοιχείο της έρευνας. Για δεδομένα κατηγορικών μεταβλητών, αυτό μπορεί να συμβεί είτε με ψευδώς αρνητικό ισχυρισμό είτε με ψευδώς θετικό ισχυρισμό.

Κατασκευή δέντρου για τη νέα μεταβλητή young

Το δέντρο ταξινόμησης που θα κατασκευάσουμε είναι , ουσιαστικά, το δέντρο πρόβλεψης για τη νέα μεταβλητή young. Η μεταβλητή young είναι κατηγορική και προέρχεται από τη μεταβλητή age (ηλικία των ατόμων του δείγματος).

Η ηλικία του κάθε ατόμου πλέον θα αποτελεί το κριτήριο διαχωρισμού των παρατηρήσεων του δείγματος ως young (νέοι) ή όχι. Ο αριθμός που ορίζεται ως σημείο διαχωρισμού είναι τα 45 έτη και η μεταβλητή young έχει δύο κλάσεις : **Yes** και **No** .

Αυτό σημαίνει ότι όσοι έχουν ηλικία μικρότερη από 45 έτη θα θεωρούνται νέοι (young) και θα κατανέμονται στην κλάση **Yes** ,ενώ για όσους δεν ισχύει αυτός ο ισχυρισμός θα κατανέμονται στην κλάση **No.**

```
 $ card       : chr  "yes" "yes" "yes" "yes" ...
 $ reports    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ age        : num  37.7 33.2 33.7 30.5 32.2 ...
 $ income     : num  4.52 2.42 4.5 2.54 9.79 ...
 $ share      : num  0.03327 0.00522 0.00416 0.06521 0.06705
 ...
 $ expenditure: num  124.98 9.85 15 137.87 546.5 ...
 $ owner      : chr  "yes" "no" "yes" "no" ...
 $ selfemp    : chr  "no" "no" "no" "no" ...
 $ dependents : int  3 3 4 0 2 0 2 0 0 0 ...
 $ months     : int  54 34 58 25 64 54 7 77 97 65 ...
 $ majorcards : int  1 1 1 1 1 1 1 1 1 1 ...
 $ active     : int  12 13 5 7 5 1 5 3 6 18 ...
> young=ifelse(CreditCard$age>45,"No","Yes")
```

Έπειτα, αντικαθιστούμε την ποσοτική μεταβλητή **age** με τη νέα ποιοτική μεταβλητή **young** που προκύπτει από αυτή. Η διαδικασία που ακολουθούμε είναι αρχικά η αφαίρεση της μεταβλητής **age** από το σύνολο δεδομένων ***CreditCard*** ,ενώ παράλληλα προσθέτουμε τη μεταβλητή **young** που δημιουργήσαμε προηγουμένως. Ωστόσο, απαραίτητη είναι και η αφαίρεση των ποιοτικών μεταβλητών του συνόλου δεδομένων για την αποφυγή λανθασμένων προβλέψεων στην κατασκευή του δέντρου ταξινόμησης.

Έτσι, προκύπτει το νέο σύνολο δεδομένων, το οποίο ονομάζουμε ***CreditCardD*** .

Θα εργαστούμε με το νέο σύνολο δεδομένων για την κατασκευή του δέντρου πρόβλεψης της μεταβλητής **young**.

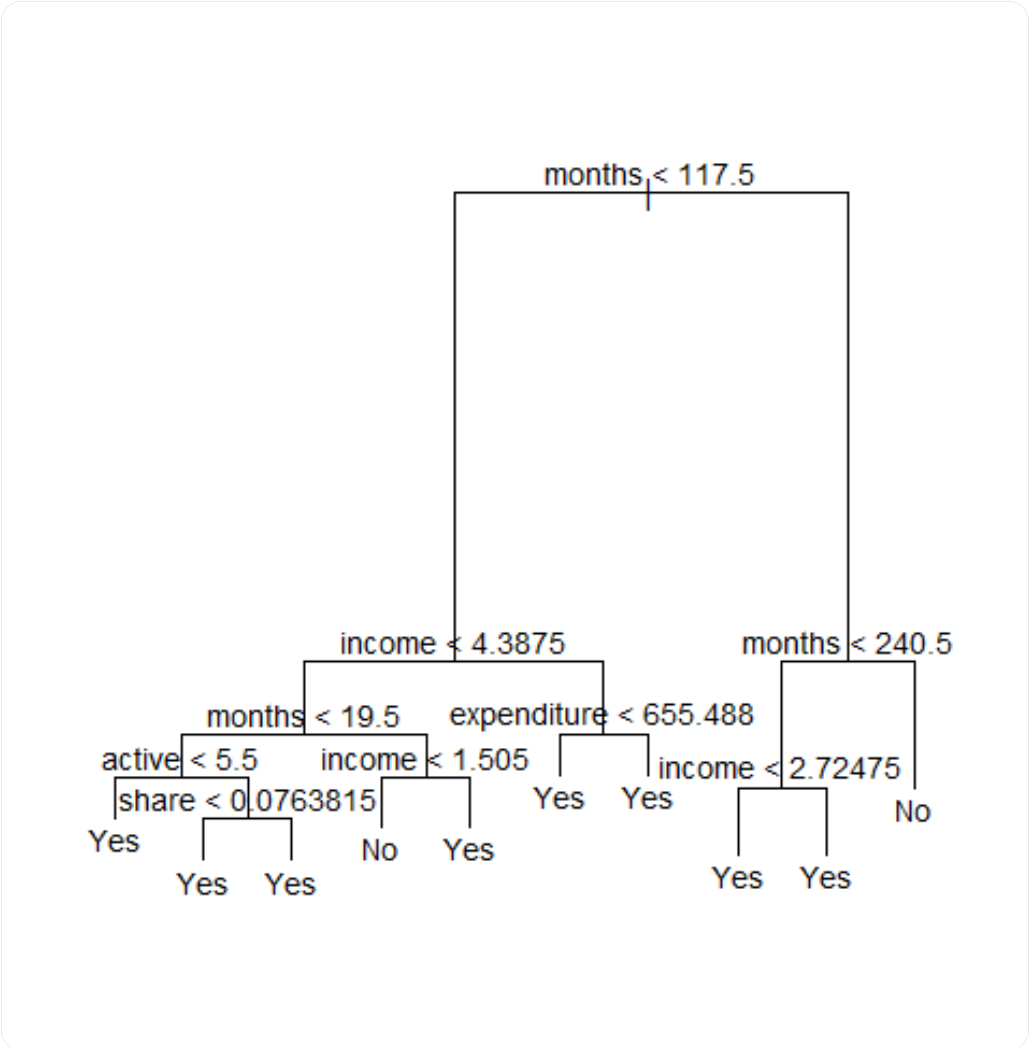
```
> str(CreditCardD)
'data.frame':   1319 obs. of  9 variables:
 $ reports    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ income     : num  4.52 2.42 4.5 2.54 9.79 ...
 $ share      : num  0.03327 0.00522 0.00416 0.06521 0.06705
 ...
 $ expenditure: num  124.98 9.85 15 137.87 546.5 ...
 $ dependents : int  3 3 4 0 2 0 2 0 0 0 ...
 $ months     : int  54 34 58 25 64 54 7 77 97 65 ...
 $ majorcards : int  1 1 1 1 1 1 1 1 1 1 ...
 $ active     : int  12 13 5 7 5 1 5 3 6 18 ...
 $ young      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2
 2 2 ...
```

Χωρίζουμε τις παρατηρήσεις σε σύνολο εκπαίδευσης (training set) και σύνολο επικύρωσης (validation set) και ορίζουμε ως δέντρο πρόβλεψης της μεταβλητής **young** από όλες τις μεταβλητές του συνόλου train, που θέτουμε ως σύνολο εκπαίδευσης (training set).

```
> s_size <- floor(0.8 * nrow(CreditCardD))
> set.seed(1)
> train_index <- sample(1:nrow(CreditCardD),size=s_size)
> train1 <- CreditCardD[train_index,]
> test1 <- CreditCardD[-train_index,]
> tree.young <- tree(young~.,train1)
> summary(tree.young)

Classification tree:
tree(formula = young ~ ., data = train1)
Variables actually used in tree construction:
[1] "months"      "income"      "active"      "share"
[5] "expenditure"
Number of terminal nodes:  10
Residual mean deviance:  0.5516 = 576.4 / 1045
Misclassification error rate: 0.1062 = 112 / 1055
```

Το ποσοστό σφάλματος εκπαίδευσης (misclassification error) είναι 10.62%.

Οι μεταβλητές που χρησιμοποιούνται από τον αλγόριθμο για την τελική μορφή του δέντρου ταξινόμησης είναι οι ποσοτικές μεταβλητές ***income***, ***months, active, share*** και ***expenditure*** . Αυτό σημαίνει πως το δέντρο πρόβλεψης προκύπτει από κόμβους που βασίζονται στο διαχωρισμό των τιμών που παίρνουν αυτές οι μεταβλητές και μας δίνεται ότι θα έχουμε 10 τερματικούς κόμβους.



Για την αξιολόγηση του δέντρου πρόβλεψης χρειαζόμαστε τον πίνακα σύγχυσης, καθώς θα εξετάσουμε πόσες τιμές κατανεμήθηκαν σωστά. Έτσι, βρίσκουμε την ακρίβεια της πρόβλεψης.

```
> confusionMatrix(ConMat)
Confusion Matrix and Statistics


tree.pred1  No Yes
       No    5   7
       Yes  31 221

               Accuracy : 0.8561
                 95% CI : (0.8078, 0.8961)
    No Information Rate : 0.8636
    P-Value [Acc > NIR] : 0.6792434

                  Kappa : 0.1504

 Mcnemar's Test P-Value : 0.0001907

            Sensitivity : 0.13889
            Specificity : 0.96930
         Pos Pred Value : 0.41667
         Neg Pred Value : 0.87698
             Prevalence : 0.13636
         Detection Rate : 0.01894
   Detection Prevalence : 0.04545
      Balanced Accuracy : 0.55409

       'Positive' Class : No
```

Ο πίνακας σύγχυσης (**Confusion Matrix**) μας δίνει και τα μέτρα εκτίμησης :

True positive rate or Sensitivity : Το ποσοστό των θετικών παραδειγμάτων που ταξινομούνται σωστά .

True negative rate or Specificity : Το ποσοστό των αρνητικών παραδειγμάτων που ταξινομούνται σωστά .

False positive rate : Το ποσοστό των αρνητικών παραδειγμάτων που ταξινομούνται λάθος (δηλαδή ως θετικά) .

False negative rate : Το ποσοστό των θετικών παραδειγμάτων που ταξινομούνται λάθος (δηλαδή ως αρνητικά) .

Precision (ακρίβεια) : Πόσα από τα παραδείγματα που ο χρήστης έχει ταξινομήσει ως θετικά είναι πράγματι θετικά .

Recall (σύγχυση) : Πόσα από τα θετικά παραδείγματα κατάφερε ο χρήστης να βρει .

Θα παρουσιάσουμε τα αποτελέσματα μέσω του πίνακα σύγχυσης, προκειμένου να δούμε την ακρίβεια της μελέτης μας μέχρι τώρα.

Tree.young	**Positive Condition**	**Negative Condition**	
Predicted Condition Positive	**True Positive -TP (Αληθώς θετικό) 221**	**False Positive-FP (Ψευδώς θετικό) 31**	**Precision (Positive Predictive Value)**
Predicted Condition Negative	**False Negative-FN (Ψευδώς αρνητικό) 7**	**True Negative-TN (Αληθώς αρνητικό) 5**	**Negative Predictive Value**
	Recall (Sensitivity) (True Positive Rate) 0.96930	**True Negative Rate (Specificity) 0.13889**	**Accuracy 0.8561**

Από τις τιμές του πίνακα σύγχυσης έχουμε και τα μέτρα εκτίμησης :

- **TP = 221**
- **FP = 31**
- **FN = 7**
- **TN = 5**
- **Sensitivity = TPR = 0.96930**
- **Specificity = TNR = 0.13889**
- **FPR = 0.86111**
- **FNR = 0.003**
- **Accuracy = 0.8561**
- **Precision = 0.8770**
- **Recall = TPR = 0.96930**

Έχουμε ορθότητα (accuracy) των προβλέψεων μας ως 85.61% από την ταξινόμηση της μεταβλητής **young**.

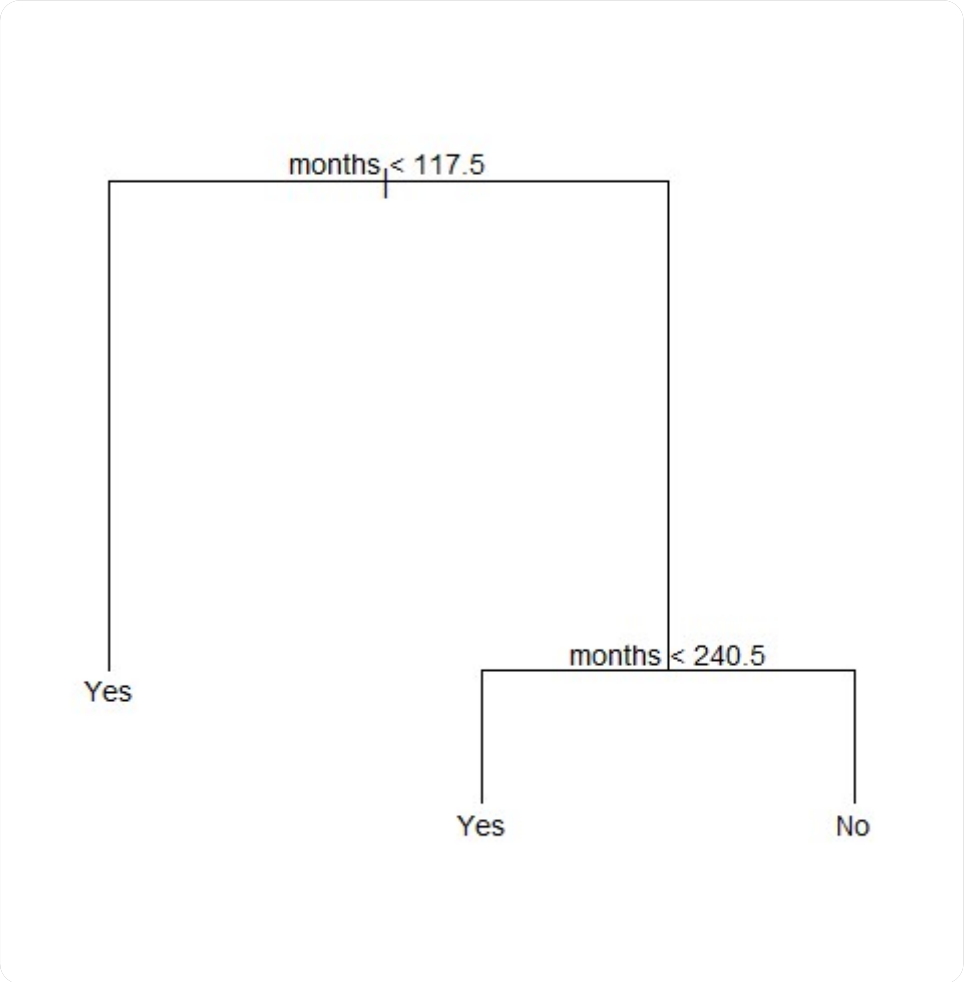
Δεν αρκούμαστε όμως μόνο σε αυτό το δέντρο ταξινόμησης για τη διεξαγωγή συμπερασμάτων όσον αφορά στην ακρίβεια της πρόβλεψης της κατανομής των δεδομένων μας. Γι' αυτό, δουλεύουμε με σκοπό τη βελτιστοποίηση του υπάρχοντος μοντέλο "κλαδεύοντας" το δέντρο μας, χρησιμοποιώντας το σφάλμα ταξινόμησης ως κριτήριο του Cross-Validation.

```
> cv.young
$size
[1] 10  6  3  1

$dev
[1] 124 124 122 133

$k
[1]       -Inf  0.0000000  0.3333333 10.0000000

$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
```

Οι τιμές των μεταβλητών που δίνονται από την εντολή cv.young που πραγματοποιεί Cross-Validation για το δέντρο δηλώνουν το πλήθος των τερματικών κόμβων του πιθανού δέντρου (size), το αντίστοιχο ποσοστό σφάλματος (dev) και τις τιμές της παραμέτρου κόστους πολυπλοκότητας Cross-Validation error rate (k).

Η μορφή που θεωρείται βέλτιστη (με το μικρότερο ποσοστό σφάλματος dev) είναι το δέντρο ταξινόμησης με πλήθος τερματικών κόμβων (size) 3, όπως φαίνεται παραπάνω . Συνεπώς, επιλέγουμε για τη συνέχεια για τη μορφή με τους 3 τερματικούς κόμβους.

Το νέο δέντρο πρόβλεψης είναι το ακόλουθο :



Η νέα μορφή του δέντρου είναι ξεκάθαρα πιο απλή από την πρώτη που είδαμε προηγουμένως και μας επιφέρει καλύτερα αποτελέσματα ,καθώς έχει μικρότερο ποσοστό σφάλματος dev. Για τη σύγκριση και την τελική αξιολόγηση της ταξινόμησης, δημιουργούμε τον πίνακα σύγχυσης για την βελτιωμένη μορφή του δέντρου ταξινόμησης για τη μεταβλητή young.

Prune.tree	Positive Condition	Negative Condition	
Predicted Condition Positive	**True Positive-TP (Αληθώς θετικό) 225**	**False Positive-FP (Ψευδώς θετικό) 32**	**Precision (Positive Predictive Value)**
Predicted Condition Negative	**False Negative-FN 3**	**True Negative-TN 4**	**Negative Predictive Value**
	Recall (Sensitivity) (True Positive Rate) 0.98684	**True Negative Rate (Specificity) 0.11111**	**Accuracy 0.8674**

```
> confusionMatrix(ConMat2)
Confusion Matrix and Statistics


tree.pred2  No Yes
       No    4   3
       Yes  32 225

               Accuracy : 0.8674
                 95% CI : (0.8205, 0.9059)
    No Information Rate : 0.8636
    P-Value [Acc > NIR] : 0.4729

                  Kappa : 0.1482

 Mcnemar's Test P-Value : 2.214e-06

            Sensitivity : 0.11111
            Specificity : 0.98684
         Pos Pred Value : 0.57143
         Neg Pred Value : 0.87549
             Prevalence : 0.13636
         Detection Rate : 0.01515
   Detection Prevalence : 0.02652
      Balanced Accuracy : 0.54898

       'Positive' Class : No
```

Παρατηρούμε ότι η ορθότητα της ταξινόμησης είναι μεγαλύτερη, καθώς από 85.61% που είχαμε πριν , τώρα φτάνουμε στο 86.74% ποσοστό ευστοχίας. Άρα ο στόχος μας έχει επιτευχθεί!!!

Συμπεράσματα

Η ταξινόμηση των τιμών του συνόλου δεδομένων αποτελεί μια δύσκολη διαδικασία, αλλά ταυτόχρονα είναι επιτακτική ανάγκη η εκπαίδευση των υποσυνόλων που ορίζουμε για την κατανόηση των σχέσεων μεταξύ των μεταβλητών. Η δημιουργία δέντρου πρόβλεψης βοηθάει, καθώς απεικονίζονται τα δεδομένα μας ,βάσει των συσχετίσεων των μεταβλητών του dataset. Με αυτό τον τρόπο επιτυγχάνεται και η δημιουργία του πίνακα σύγχυσης και βρίσκουμε την ευστοχία της πρόβλεψης , σε συνδυασμό με το ποσοστό επιτυχημένων προβλέψεων.

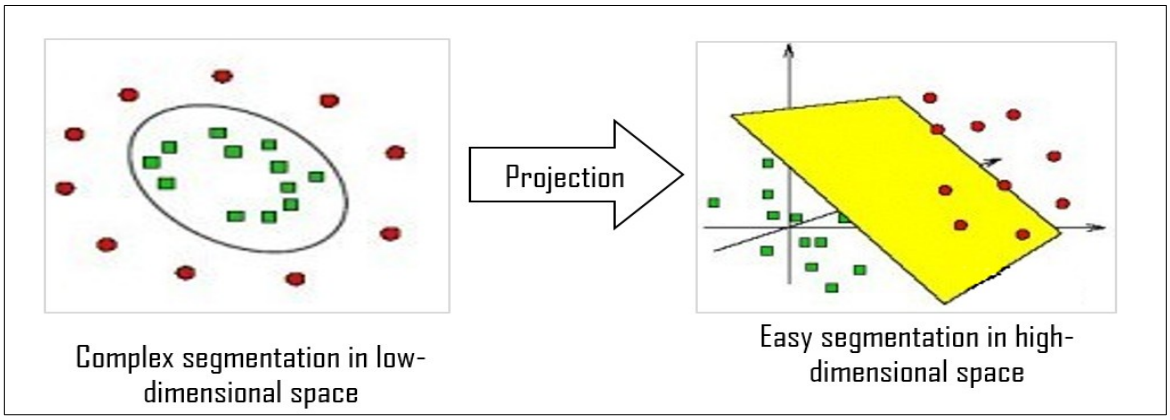
Ωστόσο, έχουμε τη δυνατότητα βελτιστοποίησης του μοντέλου που κατασκευάσαμε, εξετάζοντας όλες τις πιθανές μορφές δέντρων πρόβλεψης που ταιριάζουν στην περίπτωσή μας. Έτσι, επιλέγουμε το μοντέλο με το μικρότερο δυνατό σφάλμα και αυτό συνεπάγεται στην επανάληψη της διαδικασίας που ακολουθήσαμε προηγουμένως, με ορισμένες διαφοροποιήσεις. Ουσιαστικά, δημιουργούμε ένα νέο δέντρο πρόβλεψης (κατά πάσα πιθανότητα με λιγότερους τερματικούς κόμβους) και βρίσκουμε τον πίνακα σύγχυσης που αντιστοιχεί σε αυτό. Στη συνέχεια, ελέγχουμε την ακρίβεια των προβλέψεων με τη βοήθεια του πίνακα σύγχυσης και εφόσον έχουμε αύξηση της ακρίβειας, το αποτέλεσμα είναι επιτυχές.

Η διαδικασία που περιγράψαμε παραπάνω είναι τα βήματα που κάναμε για το dataset ***CreditCard*** , για το οποίο εργαζόμαστε και καταλήξαμε στη βελτιστοποίηση του μοντέλου μας.

Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine)

Στη μηχανική μάθηση, οι μηχανές διανυσμάτων υποστήριξης (SVM) είνα εποπτευόμενα μοντέλα μάθησης με σχετικούς αλγορίθμους που αναλύουν δεδομένα σχετικά με μοντέλα ταξινόμησης και παλινδρόμησης.

Πιο συγκεκριμένα, χρησιμοποιούνται σε προβλήματα ταξινόμησης και για την προσέγγιση της μορφής της συνάρτησης σε προβλήματα παλινδρόμησης, καθώς προβάλουν τα σημεία του συνόλου εκπαίδευσης σε έναν χώρο Ν διαστάσεων και βρίσκουν το υπερεπίπεδο που διαχωρίζει βέλτιστα τα σημεία των δύο τάξεων.
Τώρα, τα διανύσματα που ορίζουν το υπερεπίπεδο το οποίο χωρίζει τις δύο τάξεις ονομάζονται διανύσματα υποστήριξης. Όταν γίνεται η δοκιμή, οι παρατηρήσεις ελέγχου του συνόλου συγκρίνονται με αυτά τα διανύσματα , αφού είναι τα σημαντικότερα και με αυτόν τον τρόπο αποφασίζεται από τον αλγόριθμο σε ποιά κατηγορία θα τοποθετηθούν.



Στην εικόνα που τίθεται παραπάνω φαίνεται πως μεταβαίνουμε από μια πολύπλοκη κατανομή των σημείων – παρατηρήσεων του συνόλου στο επίπεδο των δύο διαστάσεων, σε μια πιο “εύπεπτη” και ταυτόχρονα καλύτερη κατανομή των παρατηρήσεων σε χώρο πολλών διαστάσεων (παραπάνω των 2).

Υπάρχουν πολλά πιθανά υπερεπίπεδα που μπορούν να σχηματιστούν, γι’ αυτό ο αλγόριθμος επιλέγει αυτό που προσφέρει το μέγιστο περιθώριο υεπερεπιπέδου (maximal margin hyperplane), δηλαδή τη μέγιστη απόσταση του επιπέδου διαχωρισμού απο τα σημεία-παρατηρήσεις που απεικονίζονται στο υπερεπίπεδο. Εν τέλει, το υπερεπίπεδο χωρίζει τα σημεία των δύο κλάσεων, έτσι ώστε να προκύπτει η μέγιστη απόσταση μεταξύ τους.

Η συνάρτηση προβολής των σημείων στο χώρο περισσότερων διαστάσεων ονομάζεται συνάρτηση πυρήνας. Το πλεονέκτημα που εμφανίζουν είναι ότι δεν χρειάζεται να γνωρίζουν τη συνάρτηση απεικόνισης φ αφού μπορούν και υπολογίζουν τα εσωτερικά γινόμενα των διανυσμάτων, πληροφορία που αρκεί για τους πυρήνες. Ο ρόλος του πυρήνας είναι σημαντικός, διότι καθορίζουν τη μορφή του διαχωρίζοντος υπερεπιπέδου και έτσι επηρεάζουν την απόδοση.

Γραμμικά διαχωρίσιμες παρατηρήσεις

Μας δίνεται ένα σύνολο δεδομένων εκπαίδευσης με n το πλήθος γραμμικά διαχωρίσιμων σημείων-πρατηρήσεων της μορφής :

(x1,y1),(x2,y2),…,(xn.yn), όπου y=1 ή y=-1, το οποίο ορίζει την κλάση του αντίστοιχου x.

Κάθε σημείο x αποτελεί πραγματικό διάνυσμα p διάστασης. Σκοπός μας είναι να βρεθεί το μέγιστο περιθώριο του υπερεπιπέδου που διαχωρίζει τα σημεία των δύο κλάσεων. Στην περίπτωση των γραμμικά διαχωρίσιμων παρατηρήσεων έχουμε την παρακάτω εξίσωση :

$\mathbf{w}^T\mathbf{x} - b = 0$, με w,x πραγματικά διανύσματα διάστασης p και b πραγματικός

Το x ονομάζεται διάνυσμα εισόδου, το w είναι ένα ρυθμιζόμενο διάνυσμα βαρών και το b αποτελεί το κατώφλι.

Στην περίπτωση των γραμμικά διαχωρίσιμων παρατηρήσεων, μπορούμε να επιλέξουμε δυο παράλληλα υπερεπίπεδα που χωρίζουν τις δύο κλάσεις των δεδομένων, έτσι ώστε η απόσταση μεταξύ τους να μεγιστοποιείται. Η περιοχή που μας δίνει το επιθυμητό αποτέλσμα για αυτή τη διαδικασία περιγράφεται από τις εξισώσεις

$\mathbf{w}^T\mathbf{x} - b = 1$,με οτιδήποτε **“πάνω”** από αυτό το όριο να ανήκει στην κλάση +1

$\mathbf{w}^T\mathbf{x} - b = -1$,με οτιδήποτε **“κάτω”** από αυτό το όριο να ανήκει στην κλάση -1

Αν χρησιμοποιήσουμε τη φόρμουλα απόστασης σημείου από ευθεία εύκολα προκύπτει πως η απόσταση των δύο υπερεπιπέδων είναι $\frac{2}{\|\mathbf{w}\|}$

Επομένως , το ζήτημα της μεγιστοποίησης της απόστασης των δύο υπερεπιπέδων αποτελεί ένα πρόβλημα ελαχιστοποίησης του μέτρου του διανύσματος w. Αυτό με τη σειρά του υπόκειται στη σχέση,

$y_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1 \text{ for } i = 1, \ldots, n.$"

Οι τιμές των w και b που αποτελούν λύσεις του προβλήματος καθορίζουν τον ταξινομητή

$\mathbf{x} \mapsto \operatorname{sgn}(\mathbf{w}^T\mathbf{x} - b)$,όπου **sgn** η συνάρτηση προσήμου

Συνεπώς, η μέγιστη απόσταση καθορίζεται πλήρως από τα διανύσματα x που βρίσκονται πιο κοντά στον ταξινομητή. Αυτά ονομάζονται διανύσματα υποστήριξης.

Μη γραμμικά διαχωρίσιμες παρατηρήσεις

Συνεχίζουμε τώρα για την περίπτωση που οι παρατηρήσεις του συνόλου μας δεν είναι γραμμικά διαχωρίσιμες. Χρησιμοποιούμε πάλι n το πλήθος παρατηρήσεις και εφαρμόζουμε το κόλπο του πυρήνα (kernel trick).

Για όλα τα x,x΄ του χώρου εισόδου X, ορισμένες συναρτήσεις k(x,x΄) που εκφράζονται ως εσωτερικό γινόμενο σε έναν άλλο χώρο V. Η συνάρτηση k αναφέρεται συχνά ως πυρήνας ή ως συνάρτηση πυρήνα και εκφράζει μια συνάρτηση στάθμισης για ένα σταθερό άθροισμα ή ολοκλήρωμα.

Ουσιαστικά, ο αλγόριθμος που προκύπτει είναι σχεδόν ίδιος με τον αρχικό, με τη διαφορα πως κάθε γινόμενο σημείων αντικαθιστάται από μια μη γραμμική συνάρτηση πυρήνα. Το αποτέλεσμα είναι πως ο αλγόριθμος έχει τη δυνατότητα να ταιριάξει το μέγιστο περι΄θωριο των υπερεπιπέδων σε έναν διαφορετικό χώρο. Ο μετασχηματισμός μπορεί να είναι μη γραμμικός όπως και ο χώρος διάστασης p, αλλά ο ταξινομητής είναι ένα υπερεπίπεδο στον νέο χώρο χαρακτηριστικών, αντίθετα με τον αρχικό χώρο εισόδου.

Να τονίσουμε εδώ πως καθώς μεγαλώνει η διάσταση του χώρου στον οποίο δουλεύουμε , αυτό συνεπάγεται την αύξυση του σφάλματος γενίκευσης των διανυσμάτων υποστήριξης. Βέβαια, κάτι τέτοιο δεν μειώνει την απόδοση του αλγορίθμου.

Άξιο αναφοράς αποτελούν τα είδη πυρήνων που χρησιμοποιούνται σε τέτοιες εφαρμογές :

Πολυωνυμικός πυρήνας (ομοιογενής) : $k(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j})^d$

Για d=1 έχουμε γραμμικό πυρήνα

Πολυωνυμικός πυρήνας (ανομοιογενής) : $k(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j} + r)^d$

Συνάρτηση ακτινικής βάσης Gauss : $k(\vec{x_i}, \vec{x_j}) = \exp(-\gamma \|\vec{x_i} - \vec{x_j}\|^2)$

Ορίζεται γ>0 και μερικές χρησιμοποιούμε την παραμετρικοποίηση γ=1/(2σ^2)

Σιγμοειδής συνάρτηση (υπερβολική εφαπτομένη) : $k(\vec{x_i}, \vec{x_j}) = \tanh(\kappa \vec{x_i} \cdot \vec{x_j} + c)$

Ορίζεται για κάποιο κ>0 και c<0

Ο πυρήνας σχετίζεται με τον μετασχηματσιμό φ από την εξίσωση $k(\vec{x_i}, \vec{x_j}) = \varphi(\vec{x_i}) \cdot \varphi(\vec{x_j})$

Εφαρμογή στο **σύνολο δεδομένων** CreditCardD

Το σύνολο δεδομέων CreditCardD αποτελεί παράγωγο του αρχικού μας dataset, χωρίς τις ποιοτικές μεταβλητές *card*, *owner*, *selfemp* καθώς έχουμε μετατρέψει και την ποσοτική μεταβλητή *age* στην ποιοτική μεταβλητή *young.*

Αρχικά, χωρίζουμε το σύνολο δεδομένων σε υποσύνολα εκπαίδευσης (train.data) και δοκιμής (test.data) για τις προβλέψεις της μεθόδου ταξινόμησης SVM. Στη συνέχεια, θα χρησιμοποιήσουμε τους δύο διαφορετικούς ταξινομητές γραμμικού και μη γραμμικού πυρήνα και θα συγκρίνουμε τα αποτελέσματα που μας δίνουν ως προς την ευστοχία της πρόβλεψης.

Χρησιμοποιούμε τον γραμμικό πυρήνα πρώτα με κόστος 100 για την ποιοτική μεταβλητή young του συνόλου δεδομένων.

```
Call:
svm(formula = train.data$young ~ ., data = train.data,
    kernel = "linear", cost = 100, scale = FALSE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  100

Number of Support Vectors:  236

 ( 126 110 )


Number of Classes:  2

Levels:
 No Yes
```

Παρατηρούμε ότι προκύπτουν 236 διανύσματα υποστήριξης για τις δύο κλάσεις.
Ο βαθμός ευστοχίας που επιτυγχάνεται στο σύνολο πρόβλεψης είναι **0.7338403.**

Τώρα, χρησιμοποιούμε μη γραμμικό πυρήνα με κόστος 100 και γάμμα ίσο με 1 για την ποιοτική μεταβλητή young του συνόλου δεδομένων.

```
Call:
svm(formula = train.data$young ~ ., data = train.data,
    kernel = "radial", gamma = 1, cost = 100)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  100

Number of Support Vectors:  501

 ( 372 129 )


Number of Classes:  2

Levels:
 No Yes
```

Παρατηρούμε ότι προκύπτουν 501 διανύσματα υποστήριξης για τις δύο κλάσεις.
Ο βαθμός ευστοχίας που επιτυγχάνεται στο σύνολο πρόβλεψης είναι **0.8326996.**

Αυτό σημαίνει πως για τις συγκεκριμένες τιμές κόστους και τιμής γάμμα, ο μη γραμμικός πυρήνας παρουσιάζει μεγαλύτερη ευστοχία.

Ωστόσο, μπορούμε να δοκιμάσουμε πολλαπλές τιμές κόστους και τιμής γάμμα για τους δύο ταξινομητές προκιμένου να βγάλουμε πιο ασφαλές συμπέρασμα όσον αφορά στην αποτελεσματικότητα των δύο μεθόδων . Το πρόγραμμα εκτελεί τη μέθοδο 10 - cross validation και συγκρίνει μοντέλα SVM είτε με την συνάρτηση γραμμικού πυρήνα είτε με τη συνάρτηση του μη γραμμικού πυρήνα, δίνοντας ένα εύρος για την παράμετρο του κόστους.

Για τον γραμμικό ταξινομητή , θα χρησιμοποιήσουμε τα κόστοι (1,10,100,1000,10000) και θα βρούμε ανάμεσα σε αυτά το βέλτιστο μοντέλο.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost
 1000

- best performance: 0.1298203

- Detailed performance results:
   cost     error dispersion
1     1 0.1307637 0.04196200
2    10 0.1307637 0.04196200
3   100 0.1307637 0.04196200
4  1000 0.1298203 0.04250311
5 10000 0.1298203 0.04250311
```

Βλέπουμε πως για κόστος 1000 έχουμε το μικρότερο cross validation error rate.

Χρησιμοποιώντας το βέλτιστο μοντέλο κάνουμε προβλέψεις για την κλάση των παρατηρήσεων του συνόλου ελέγχου, ενώ στη συνέχεια κατασκευάζουμε τον πίνακα σύγχυσης που μας δίνει μια ξεκάθαρη εικόνα όσον αφορά την ευστοχία του μοντέλου μας.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No    0   0
       Yes  34 229

               Accuracy : 0.8707
                 95% CI : (0.8241, 0.9088)
    No Information Rate : 0.8707
    P-Value [Acc > NIR] : 0.5456

                  Kappa : 0

 Mcnemar's Test P-Value : 1.519e-08

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.8707
             Prevalence : 0.1293
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : No
```

Αντίστοιχα για τον μη γραμμικό ταξινομητή, θα χρησιμοποιήσουμε τα ίδια κόστοι και τις τιμές γάμμα(1,2,3,4,5) και θα βρούμε ανάμεσα σε αυτά το βέλτιστο μοντέλο.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 gamma cost
     3    1

- best performance: 0.1298473

- Detailed performance results:
   gamma  cost     error dispersion
1      1     1 0.1307996 0.03829086
2      2     1 0.1326864 0.03932323
3      3     1 0.1298473 0.03882732
4      4     1 0.1298473 0.03882732
5      5     1 0.1298473 0.03882732
6      1    10 0.1629650 0.03942694
7      2    10 0.1459479 0.03657084
8      3    10 0.1450135 0.03923813
9      4    10 0.1431357 0.04254536
10     5    10 0.1374394 0.04002754
11     1   100 0.1734142 0.04637698
12     2   100 0.1535400 0.04094561
13     3   100 0.1507008 0.04162275
14     4   100 0.1459838 0.04481868
15     5   100 0.1383827 0.04000296
16     1  1000 0.1847529 0.04925877
17     2  1000 0.1554358 0.04130727
18     3  1000 0.1497574 0.04148574
19     4  1000 0.1459838 0.04481868
20     5  1000 0.1383827 0.04000296
21     1 10000 0.1847529 0.04925877
22     2 10000 0.1554358 0.04130727
23     3 10000 0.1497574 0.04148574
24     4 10000 0.1459838 0.04481868
25     5 10000 0.1383827 0.04000296
```

Παρατηρούμε πως για κόστος 1 και γάμμα 3 έχουμε το μικρότερο cross validation error rate.

Ακολουθώντας την ίδια διαδικασία που πράξαμε και παραπάνω, κατασκευάζουμε τον πίνακα σύγχυσης και με αυτό τον τρόπο αποκτούμε πιο σαφή εικόνα των προβλέψεων για την ταξινόμηση των σημείων του συνόλου ελέγχου.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No    0   2
       Yes  34 227

               Accuracy : 0.8631
                 95% CI : (0.8156, 0.9023)
    No Information Rate : 0.8707
    P-Value [Acc > NIR] : 0.6834

                  Kappa : -0.0146

 Mcnemar's Test P-Value : 2.383e-07

            Sensitivity : 0.000000
            Specificity : 0.991266
         Pos Pred Value : 0.000000
         Neg Pred Value : 0.869732
             Prevalence : 0.129278
         Detection Rate : 0.000000
   Detection Prevalence : 0.007605
      Balanced Accuracy : 0.495633

       'Positive' Class : No
```

Συμπέρασμα ταξινόμησης SVM

Εφόσον ολοκληρώσαμε τη διαδικασία βελτιστοποίησης των μοντέλων πρόβλεψης ταξινόμησης για τους δύο πυρήνες, πλέον είμαστε σε θέση να συγκρίνουμε την αποτελεσματικότητά τους και να διεξάγουμε τα τελικά συμπεράσματα.

Αρχικά, για την περίπτωση του γραμμικού πυρήνα :

Βλέπουμε πως οι 229 από το πλήθος των 263 παρατηρήσεων του συνόλου ελέγχου ταξινομήθηκαν σωστά. Αυτό σημαίνει πως έχουμε ευστοχία περίπου **87 %**.

Ακολουθεί η περίπτωση του μη γραμμικού πυρήνα :

Παρατηρούμε πως οι 227 από το πλήθος των 263 παρατηρήσεων του συνόλου ελέγχου ταξινομήθηκαν σωστά. Αυτό σημαίνει πως έχουμε ευστοχία περίπου **86%**.

Το τελικό συμπέρασμα για την περίπτωση μας είναι πως ο γραμμικός πυρήνας αποτελεί την καλύτερη επιλογή για την πρόβλεψη ταξινόμησης των σημείων-παρατηρήσεων .

Βιβλιογραφία

Σημειώσεις Στατιστικής Μάθησης 1, Χ. Μπράτσας, Ι. Αντωνίου

Σημειώσεις Στατιστικής Μάθησης 2, Χ. Μπράτσας, Ι. Αντωνίου

Σημειώσεις Θεωρίας Πιθανοτήτων Ι, Ι. Αντωνίου

Σημειώσεις Θεωρίας Πιθανοτήτων ΙΙ, Γ. Αφένδρας

Ορισμός και πληροφορίες για την στατιστική/μηχανική μάθηση :

https://el.wikipedia.org/wiki/%CE%9C%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE_%CE%BC%CE%AC%CE%B8%CE%B7%CF%83%CE%B7

Ορισμός και πληροφορίες για το X^2 Pearson's test :

https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

Πληροφορίες και εικόνες για τα μέτρα κεντρικής τάσης, μέτρα διασποράς και συντελεστές λοξότητας και κύρτωσης :

https://eclass.upatras.gr/modules/document/file.php/MST135/%CE%95%CE%9D%CE%9F%CE%A4%CE%97%CE%A4%CE%91%2004.pdf

https://eclass.upatras.gr/modules/document/file.php/DEAPT167/%CE%98%CE%B5%CF%89%CF%81%CE%AF%CE%B1/10%CE%B7%20%CE%94%CE%B9%CE%AC%CE%BB%CE%B5%CE%BE%CE%B7.pdf

https://openclass.teiwm.gr/modules/document/file.php/BA-G106/prototype_template_TEI_enotita_2.pdf

Πληροφορίες για τον συντελεστή συσχέτισης :

https://www.aua.gr/gpapadopoulos/files/sisxetisi091.pdf

Πληροφορίες για την ανάλυση διασποράς :

https://el.wikipedia.org/wiki/%CE%91%CE%BD%CE%AC%CE%BB%CF%85%CF%83%CE%B7_%CE%B4%CE%B9%CE%B1%CE%BA%CF%8D%CE%BC%CE%B1%CE%BD%CF%83%CE%B7%CF%82

Πληροφορίες για τις μηχανές διανυσμάτων υποστήριξης (SVM) :

https://en.wikipedia.org/wiki/Support-vector_machine

https://en.wikipedia.org/wiki/Kernel_method#Mathematics:_the_kernel_trick

https://aibook.gr/wp-content/uploads/sites/2/2020/06/chapter18.pdf

http://www.sthda.com/english/articles/36-classification-methods-essentials/144-svm-model-support-vector-machine-essentials/

https://rpubs.com/Kushan/296706

https://stackoverflow.com/questions/27398517/svm-classification-using-r-variable-length-differ-error