ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ
ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΜΕΘΟΔΟΙ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΑΝΑΛΥΣΗΣ ΕΠΙΧΕΙΡΗΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Παπούλια Καλλιόπη

Διπλωματική εργασία
που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Σεπτέμβριος 2020

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από το ΓΣΕΣ του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμόν ……. συνεδρίαση του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της επιτροπής ήταν:
- Μπερσίμης Σωτήριος (Επιβλέπων)
- Γεωργακέλλος Δημήτριος
- Τζαβελάς Γεώργιος

Η έγκριση της Διπλωματικής Εργασίας από το τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIREAUS

School of Finance and Statistics Science

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

METHODS FOR MULTIVARIATE ANALYSIS OF BUSINESS DATA

By
Papoulia Kalliopi

MSc Dissertation
submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Pireaus
September 2020

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαιτέρως τον επιβλέποντα καθηγητή μου κύριο Σωτήρη Μπερσίμη, ο οποίος έδειξε εξαιρετική υπομονή και επιμονή κατά τη διάρκεια της συγγραφής της παρούσας εργασίας, μεταδίδοντας γνώσεις και πράττοντας σημαντικές υποδείξεις και συμβουλές, ώστε όλη η προσπάθεια να ολοκληρωθεί επιτυχώς εντός των προβλεπόμενων χρονικών ορίων.

Επιπρόσθετα θα ήθελα να ευχαριστήσω τα μέλη της τριμελούς επιτροπής, κ. Γεωργακέλλο Δημήτριο και κ. Τζαβελά Γεώργιο, για το χρόνο που αφιέρωσαν στη μελέτη και τη διόρθωση της. Τέλος, θα ήταν σημαντική παράληψη να μην ευχαριστήσω τους γονείς μου για την υπομονή και τη στήριξη που μου παρείχαν.

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία αναφέρεται σε μεθόδους πολυμεταβλητής στατιστικής ανάλυσης. Η πολυμεταβλητή στατιστική ανάλυση αποτελεί ένα πολύτιμο εργαλείο για πολλές επιστήμες και βιομηχανίες. Περιλαμβάνει πολλές μεθόδους ανάλογα με το πρόβλημα που χρήζει αντιμετώπισης. Αποτελείται από πολλές μεθοδολογίες και ποικίλες διαδικασίες κατάλληλες για χρήση πολλών μεταβλητών με στόχο την εξαγωγή αποτελεσμάτων και την στατιστική συμπερασματολογία.

Συγκεκριμένα στη διπλωματική αυτή εργασία, αρχικά παρουσιάζεται το θεωρητικό υπόβαθρο των κυριότερων πολυμεταβλητών μεθόδων, δηλαδή βασικοί όροι, περιγραφή και προϋποθέσεις που απαιτούνται για την εφαρμογή τους. Μέθοδοι όπως είναι η λογιστική παλινδρόμηση, η ανάλυση συστάδων, η διαχωριστική ανάλυση και άλλα. Επιπρόσθετα γίνεται αναφορά και περιγραφή πραγματικών προβλημάτων που επιλύθηκαν με τις προαναφερθείσες μεθόδους σε διάφορες βιομηχανίες όπως τηλεπικοινωνίες, τουριστικές επιχειρήσεις, περιβάλλον και άλλα.

Τέλος, η εργασία ολοκληρώνεται με ανάλυση δεδομένων με χρήση των μεθόδων και με τη βοήθεια του στατιστικού πακέτου SPSS MODELER.

ABSTRACT

The present study refers to multivariate statistical analysis methods. Multivariate statistical analysis is a valuable tool for many science fields and industries. It includes many methods depending on problem that needs to be addressed. Multivariate statistical analysis consists of many methodologies and various procedures suitable for the use of many variables in order to extract results and statistical conclusions.

Specifically, in this thesis, firstly, the theoretical background of some multivariable methods such as accounting regression, cluster analysis, separation analysis and others is presented. Namely, basic terms, description and conditions required for their application.

Additionally, real problems that were solved with above-mentioned methods in various industries such as telecommunications, tourism companies or environment are reported and described.

Finally, this research is completed with a statistical data analysis and methods are used with assistance of statistical package SPSS MODELER.

Περιεχόμενα

Κεφάλαιο 3

Μέθοδοι Ανάλυσης: Ταξινομήσεις – Προβλέψεις

Κεφάλαιο 4

Μελέτες περιπτώσεων που αναλύθηκαν με τις παραπάνω μεθόδους

Κεφάλαιο 5

Ανάλυση Δεδομένων με τη χρήση του SPSS Modeler

Κεφάλαιο 1

Εισαγωγή

Στη σύγχρονη εποχή η ανάγκη ανάλυσης δεδομένων γίνεται όλο και μεγαλύτερη. Τα δεδομένα όμως που χρήζουν ανάλυσης κατά πλειοψηφία δεν αποτελούνται από ένα χαρακτηριστικό. Αντίθετα, αποτελούνται από μεγάλο αριθμό μεταβλητών–χαρακτηριστικών, τα οποία υπάρχει ανάγκη να αναλυθούν συγχρόνως. Αρχικός στόχος των μεθόδων είναι να βρεθεί μια σωστή δομή των δεδομένων ώστε να εξαχθεί η μεγαλύτερη δυνατή και σωστή πληροφορία από αυτά.

Οι μέθοδοι πολυμεταβλητών τεχνικών έχουν πολλές εφαρμογές: στις επιστήμες Υγείας που πιο πολύ σχετίζονται με διαγνώσεις και εμφανίσεις συμπτωμάτων, σε εταιρείες πληροφορικής για την εξόρυξη χρήσιμων πληροφοριών από τεράστιες και πολύπλοκες βάσεις δεδομένων, σε εταιρίες τηλεπικοινωνιών με πιο συχνές εφαρμογές για την ικανοποίηση και διατήρηση πελατών, στις κοινωνικές επιστήμες, παραδείγματος χάρη στο χώρο της εκπαίδευσης για εξέταση μεθόδων και συμπεριφορών και σε πολλές άλλες ακόμη.

Ως εκ τούτου, η ανάγκη πολλών επιχειρήσεων είναι, μέσω της ανάλυσης ενός κατάλληλου δείγματος να μπορέσει να καταλήξει σε κάποια **patterns** πληθυσμού και βασιζόμενος σ αυτά να εξαγάγει συμπεράσματα για ολόκληρο τον πληθυσμό. Αυτό μπορεί να γίνει εφικτό με τη χρήση πολυμεταβλητών τεχνικών. Διότι μέσω αυτών μπορούν να μελετηθούν συσχετίσεις και να ερμηνευτούν μεταβλητές-χαρακτηριστικά μέσω διάφορων επιλογών και κανόνων. Επιπρόσθετα, μπορούν να δημιουργηθούν ομάδες με κοινά χαρακτηριστικά. Είναι δυνατή επίσης, η μείωση των διαστάσεων των δεδομένων (συμπύκνωση πληροφορίας). Μια ακόμα βασική διαδικασία είναι η ποσοτικοποίηση μη παρατηρήσιμων ποσοτήτων όπως για παράδειγμα η ευφυΐα, η πρόβλεψη νέων τιμών και άλλα πολλά που μπορούν να βοηθήσουν μια επιχείρηση / οργανισμό να λάβει σωστές αποφάσεις. Τέτοιες μέθοδοι αναλύονται στο Κεφάλαιο 1. Στην παρούσα διπλωματική αναφέρονται τέτοια παραδείγματα αναλύσεων στο Κεφάλαιο 2.

Τέλος, ο λόγος που τα τελευταία χρόνια οι τεχνικές πολυμεταβλητής ανάλυσης γίνονται όλο και πιο διαδεδομένες είναι η ανάπτυξη πληθώρων στατιστικών πακέτων. Μερικά από αυτά είναι το Minitab, Matlab, SAS, STATA, R, Python, SPSS Statistics και SPSS Modeler. Με τη βοήθεια του τελευταίου έγινε στο Κεφάλαιο 3 μια ανάλυση ανάλυση δεδομένων με τη χρήση των παραπάνω τεχνικών.

Κεφάλαιο 2

Μέθοδοι Ανάλυσης: Παλινδρομήσεις - Μείωση Διαστάσεων

- Παλινδρομήσεις

2.1 Γραμμική Παλινδρόμηση

Σε διάφορα προβλήματα της στατιστικής ενδιαφερόμαστε για την ταυτόχρονη μελέτη δύο ή περισσότερων μεταβλητών, ώστε να προσδιοριστεί με ποιο τρόπο αυτές οι μεταβλητές σχετίζονται μεταξύ τους. Η γραμμική παλινδρόμηση υποθέτει ότι τα δεδομένα μπορούν να μοντελοποιηθούν με μία γραμμική σχέση. Συγκεκριμένα η γραμμική παλινδρόμηση είναι ένα χρήσιμο εργαλείο για την πρόβλεψη μιας ποσοτικής μεταβλητής (απόκρισης).

2.1.1 Απλή Γραμμική Παλινδρόμηση

Η πιο απλούστερη περίπτωση της γραμμικής παλινδρόμησης είναι η απλή γραμμική παλινδρόμηση. Η απλή γραμμική παλινδρόμηση είναι μία απλή προσέγγιση για την πρόβλεψη μιας ποσοτικής μεταβλητής Y, η οποία καλείται εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης και θεωρείται τυχαία, χρησιμοποιώντας μόνο μία μεταβλητή X ,η οποία καλείται ανεξάρτητη ή ερμηνευτική μεταβλητή και δεν θεωρείται τυχαία. Μαθηματικά αυτό εκφράζεται από τη σχέση:

$$Y=\beta_0+\beta_{1*} X + \varepsilon$$

όπου:

- β_0, β_1 παράμετροι ή συντελεστές του μοντέλου, σταθερά και κλίση αντίστοιχα. Στην πράξη είναι άγνωστα και αυτά και πρέπει να χρησιμοποιήσουμε τα δεδομένα ώστε να τα εκτιμήσουμε. Το β_0 είναι μία σταθερά, η τιμή της Y, όταν X=0 και β_1 είναι ο συντελεστής της X, που σημαίνει ότι μία μεταβολή της X κατά μία μονάδα μέτρησης επιφέρει αλλαγή στη μεταβλητή Y κατά β_1. Το ε είναι τα κατάλοιπα του μοντέλου, δηλαδή η διαφορά της προβλεπόμενης με την εκτιμώμενη τιμή.

Η πρόβλεψη της Y:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$$

Η τιμή της εκτιμήτριας $\hat{\beta}_0$ της παραμέτρου β_0 παριστάνει την τεταγμένη του σημείου στο οποίο η ευθεία τέμνει τον άξονα y'y. Όταν $\hat{\beta}_0 = 0$ τότε η ευθεία διέρχεται από την αρχή των αξόνων.

Η τιμή της εκτιμήτριας $\hat{\beta}_1$ παριστάνει τη μεταβολή της Y. Όταν αυξηθεί το X κατά μία μονάδα τότε το $\hat{y}$ αυξάνεται κατά $\hat{\beta}_1$, όταν $\hat{\beta}_1 > 0$, αλλιώς εάν $\hat{\beta}_1 < 0$ τότε το Y μειώνεται κατά $\hat{\beta}_1$ μονάδες.

2.1.1.1 Υποθέσεις

- $\varepsilon_i \sim N(0,\sigma^2)$
- $E(\varepsilon_i)=0$
- $Var(\varepsilon_i)=\sigma^2$
- $COV(\varepsilon_i, \varepsilon_j)=0$
- $Y_i \sim N(\beta_0 + \beta_1 X_1, \sigma^2)$
- $E(Y_i)= \beta_0 + \beta_1 X_1$
- $Var(Y_i)=\sigma^2$
- $COV(Y_i, Y_j)=0$

Έλεγχος Υποθέσεων

Για την εξαγωγή σωστών στατιστικών συμπερασμάτων πρέπει να ελέγξουμε αν ισχύουν για τα σφάλματα τα εξής: κανονικότητα, ομοσκεδαστικότητα και ανεξαρτησία των σφαλμάτων από τις προβλεπόμενες τιμές.

α. Κανονικότητα

- Γραφικά

Ένας τρόπος ελέγχου της κανονικότητας των σφαλμάτων είναι γραφικά. Θα μπορούσε να γίνει κάνοντας ένα ιστόγραμμα των καταλοίπων και προσαρμόζοντας την καμπύλη της κανονικής κατανομής για σύγκριση, ή με ένα scatter plot προσαρμόζοντας την ευθεία παλινδρόμησης και αυτό που θέλουμε είναι να είναι τα σημεία εκατέρωθεν της ευθείας.

- Στατιστικός έλεγχος

Έλεγχος υπόθεσης:

H_0: Τα σφάλματα ακολουθούν κανονική κατανομή έναντι

H_1: Τα σφάλματα δεν ακολουθούν κανονική κατανομή

Μπορούμε να χρησιμοποιήσουμε τον έλεγχο του Kolmogorov-Smirnov και αυτό που θέλουμε είναι το $\mathbf{p_{value}}$ του ελέγχου να είναι > α (επίπεδο σημαντικότητας) ώστε να μην απορρίψουμε την H_0.

β. Ομοσκεδαστικότητα

- Γραφικά

Μπορούμε να ελέγξουμε την ομοσκεδαστικότητα γραφικά κατασκευάζοντας ένα διάγραμμα διασποράς (scatter plot) των καταλοίπων συναρτήσει των τιμών που παίρνει η ανεξάρτητη μεταβλητή. Δε θέλουμε να υπάρχει κάποια συστηματικότητα.

- Στατιστικός έλεγχος

Έλεγχος υπόθεσης:

H_0: Τα σφάλματα είναι ομοσκεδαστικά έναντι

H_1: Τα σφάλματα δεν είναι ομοσκεδαστικά

Μπορούμε να χρησιμοποιήσουμε τον έλεγχο του Levene και αυτό που θέλουμε είναι το **p.value** του ελέγχου να είναι > α ώστε να μην απορρίψουμε την H_0.

γ. Ανεξαρτησία

- Γραφικά

Μπορούμε να ελέγξουμε την ανεξαρτησία των σφαλμάτων κατασκευάζοντας ένα διάγραμμα διασποράς (scatter plot) με τα κατάλοιπα και τις προσαρμοσμένες τιμές. Δε θέλουμε συστηματικότητα όσον αφορά τα σφάλματα συναρτήσει των προβλεπόμενων τιμών.

- Στατιστικός έλεγχος

Έλεγχος υπόθεσης:

H_0: Τα σφάλματα είναι ανεξάρτητα έναντι

H_1: Τα σφάλματα δεν είναι ανεξάρτητα

Μπορούμε να χρησιμοποιήσουμε το Runs Test και αυτό που θέλουμε είναι το p_{value} του ελέγχου να είναι > α ώστε να μην απορρίψουμε την H_0.

2.1.1.2 Έλεγχος γραμμικής σχέσης

Ένας αρχικός έλεγχος ώστε να ελέγξουμε αν υπάρχει γραμμική σχέση ανάμεσα στο Y και στο X είναι η δημιουργία ενός διαγράμματος διασποράς (scatter plot). Σ' ένα τέτοιο διάγραμμα οι τιμές της μεταβλητής Y τοποθετούνται στον κατακόρυφο άξονα και της X στον οριζόντιο. Επιπλέον, προσαρμόζουμε την ευθεία παλινδρόμησης στο διάγραμμα αυτό και θέλουμε τα σημεία να είναι όσο πιο κοντά στην ευθεία αυτή , η οποία θα είναι κ αυτή που προσαρμόζονται καλύτερα τα δεδομένα.

2.1.1.3 Μέτρο προσαρμογής

Μέθοδος ελαχίστων τετραγώνων

Η μέθοδος αυτή επιλέγει τα $\hat{\beta}_0$,$\hat{\beta}_1$ που ελαχιστοποιούν το άθροισμα των τετραγώνων των καταλοίπων ε_i ,δηλαδή:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 \cdot x_i)\right)^2$$

Οι εκτιμητές ελάχιστων τετραγώνων για τις παραμέτρους β_0, β_1 της ευθείας $y = \beta_0 + \beta_1 * x$ με βάση τα ζεύγη σημείων δίνονται από τους τύπους:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i^2}$$

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 - \frac{1}{n}\sum_{i=1}^{n} x_i$$

Η ευθεία που προκύπτει ονομάζεται ευθεία ελαχίστων τετραγώνων.

Πέρα από τα κριτήρια που προαναφέραμε θέλοντας να ποσοτικοποιήσουμε το μέγεθος στο οποίο η ευθεία προσαρμόζεται καλύτερα στα δεδομένα υπάρχει ένα μέτρο προσαρμογής που χρησιμοποιείται συνηθέστερα και είναι το R^2 ,το οποίο μετριέται πάνω σε μία τυποποιημένη κλίμακα .Η κλίμακα αυτή είναι 0-1, είναι ανεξάρτητη από την κλίμακα της Y, με 0 να σημαίνει μηδενική γραμμική συσχέτιση έως 1 τέλεια γραμμική συσχέτιση.

Το μέτρο αυτό μας δίνει το ποσοστό διασποράς των τιμών της Y η οποία εξηγείται μέσω της ευθείας παλινδρόμησης.

Τύπος:

$$R^2 = \frac{SSR}{SSTO} = \frac{SSTO-SSE}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Όπου:

- $R^2 = 0 \Leftrightarrow SSR=0$
- $R^2 = 1 \Leftrightarrow SSR=SSTO \Leftrightarrow SSE=0$

1.1.2 Πολλαπλή Παλινδρόμηση

Η πολλαπλή γραμμική παλινδρόμηση αποτελεί μία επέκταση της απλής. Αντί για μια ανεξάρτητη επιτρέπει τη χρήση παραπάνω. Δηλαδή αντί να προσαρμόζουμε ένα ξεχωριστό απλό μοντέλο γραμμικής παλινδρόμησης για κάθε προγνωστικό παράγοντα, μία καλή προσέγγιση είναι να επεκταθεί το απλό μοντέλο έτσι ώστε να μπορεί να λάβει και άλλους προγνωστικούς παράγοντες. Στην πολλαπλή παλινδρόμηση ως ανεξάρτητες μεταβλητές X μπορούν να χρησιμοποιηθούν και κατηγορικές (ποιοτικές) μεταβλητές.

Η μορφή του είναι:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p+\varepsilon$$

1.1.2.1 Υποθέσεις

- Y είναι συνεχής μεταβλητή
- X είναι συνεχής ή κατηγορική μεταβλητή
- $COV(X_i, X_j) = 0$
- ε_i είναι ανεξάρτητα
- $\varepsilon_i \sim N(0,\sigma^2)$

- **Έλεγχοι Υποθέσεων**

Στην πολλαπλή παλινδρόμηση με *p* προγνωστικούς παράγοντες πρέπει να κάνουμε τους εξής ελέγχους :

Έλεγχος υπόθεσης:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p$

H_1**:** έστω ένα β_i να μην είναι 0

Γι' αυτή την υπόθεση χρησιμοποιείται η στατιστική συνάρτηση F-statistics με τύπο:

$$F=\frac{(TSS-RSS)\,/\,p}{RSS\,/\,(n-p-1)}=\frac{MSR}{MSE}$$

Όπου:

- SSR= $\sum(y_i-\bar{y})^2$, $RSS=\Sigma(y_i-\hat{y})^2$

<u>Κανόνας απόφασης</u>:

- Αν $F > F_{p-1,\,n-p}(\alpha)$, απορρίπτουμε τη μηδενική υπόθεση
- Αν $F \leq F_{p-1,\,n-p}(\alpha)$, δεν απορρίπτουμε τη μηδενική υπόθεση

<u>Αν ο έλεγχος είναι αμφίπλευρος</u> :

$H_0: \beta_i = 0$ vs $H_1: \beta_i > 0$

τότε:

$$T=\frac{\hat{\beta}_i}{S(\hat{\beta}_i)}$$

Κανόνας απόφασης:

- Αν $T > t_{n-p}(\alpha)$, απορρίπτουμε τη μηδενική υπόθεση
- Αν $T \leq t_{n-p}(\alpha)$, δεν απορρίπτουμε τη μηδενική υπόθεση

Ενώ αν :

$H_0: \beta_i = 0$ vs $H_1: \beta_i < 0$

τότε:

$$T=\frac{\hat{\beta}_i}{S(\hat{\beta}_i)}$$

Κανόνας απόφασης:

- Αν $T < -t_{n-p}(\alpha)$, απορρίπτουμε τη μηδενική υπόθεση
- Αν $T \geq -t_{n-p}(\alpha)$, δεν απορρίπτουμε τη μηδενική υπόθεση

2.1.2.2 Επιλογή ανεξάρτητων μεταβλητών

Είναι πιθανό όλοι οι προγνωστικοί παράγοντες να συσχετίζονται με την απόκριση, αλλά συχνότερα συμβαίνει να σχετίζεται μόνο με ένα υποσύνολο αυτών. Ιδανικά λοιπόν, θέλουμε να επιλέγουμε μεταβλητές δοκιμάζοντας πολλά διαφορετικά μοντέλα που το καθένα θα περιέχει διαφορετικό υποσύνολο μεταβλητών.

Για παράδειγμα αν έχουμε δύο μεταβλητές μπορούμε να θεωρήσουμε τέσσερα μοντέλα.Τα οποία είναι:**1.** πού να μην περιέχει καθόλου μεταβλητές, **2.** να περιέχει τη μία μόνο X_1 , **3.** να περιέχει τη X_2 μόνο και **4.** να περιέχει και τις δύο ανεξάρτητες μαζί. Ώστε να αποφασίσουμε για την ποιότητα του μοντέλου, ποιο δηλαδή είναι το καλύτερο , μπορούν να χρησιμοποιηθούν διάφορα στατιστικά κριτήρια, όπως Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) , Adjusted R^2 ,Cp Mallow κλπ.

Όμως ανάλογα με το p, δεν είναι πάντα εφικτό να δοκιμάζεται κάθε πιθανό υποσύνολο μεταβλητών. Για παράδειγμα, όπως είπαμε αν p=2 υπάρχουν $2^2 = 4$ μοντέλα, αλλά αν p=25 τότε υπάρχουν $2^{25} = 33.554,432$ κι αυτό δεν είναι πρακτικό. Επομένως αν το p είναι πολύ μικρό μπορούμε να εξετάσουμε όλα τα πιθανά μοντέλα , αν όχι χρειαζόμαστε μία αυτοματοποιημένη και αποτελεσματική προσέγγιση ώστε να επιλέξουμε το σύνολο των μοντέλων που πρέπει να λάβουμε υπόψιν.

Αναφέρουμε 3 τέτοιες προσεγγίσεις:

1. **Forward selection**: Ξεκινάει με το μοντέλο χωρίς καμία ανεξάρτητη μεταβλητή μόνο με τη σταθερά και με βάση όποια μεταβλητή έχει το μικρότερο RSS προστίθεται στο μοντέλο. Αυτή η διαδικασία συνεχίζετε μέχρι να ικανοποιηθεί κάποιος κανόνας διακοπής.
2. **Backward selection:** Ξεκινάει με ένα μοντέλο με όλες τις μεταβλητές και βγάζει κάθε φορά τη μεταβλητή με το μεγαλύτερο p_{value} .Αυτή η διαδικασία συνεχίζετε μέχρι να ικανοποιηθεί κάποιος κανόνας διακοπής.
3. **Mixed selection:** Αυτή η μέθοδος είναι ένας συνδυασμός των δύο παραπάνω μεθόδων. Ξεκινάει με ένα μοντέλο με καθόλου μεταβλητές όπως η forward και προσθέτει τη μεταβλητή που έχει την καλύτερη προσαρμογή. Συνεχίζει και προσθέτει μία-μια τις μεταβλητές. Αν για παράδειγμα κάποιο p_{value} για κάποια μεταβλητή αυξηθεί πάνω από ένα όριο ενώ μπει μία καινούργια μεταβλητή στο μοντέλο, τότε την αφαιρεί. Αυτή η διαδικασία forward και backward συνεχίζεται έως ότου οι μεταβλητές έχουν όλες χαμηλή τιμή p_{value} .

Σχόλια

Η επιλογή backward δεν μπορεί να χρησιμοποιηθεί αν p > n ενώ η επιλογή forward μπορεί πάντα να χρησιμοποιηθεί. Η forward selection μπορεί να περιλαμβάνει μεταβλητές που αργότερα καθίστανται περιττές. Η mixed selection μπορεί να το διορθώσει αυτό.

2.1.2.3 Κατηγορικές ανεξάρτητες μεταβλητές

Στην πράξη πολλές φορές είναι ανάγκη να χρησιμοποιηθούν για την πρόβλεψη μία ή περισσότερες κατηγορικές (ποιοτικές) μεταβλητές. Ο πιο αποτελεσματικός τρόπος διαμόρφωσης είναι με τη χρήση δείκτριων μεταβλητών (ψευδομεταβλητές/dummy) της μορφής:

- Κατηγορικές μεταβλητές με 2 επίπεδα:

$$x_i = \begin{cases} 1, & \alpha\nu\ \iota\sigma\chi\acute{\upsilon}\epsilon\iota\ \kappa\acute{\alpha}\pi o\iota\alpha\ \sigma\upsilon\nu\theta\acute{\eta}\kappa\eta \\ 0, & \alpha\nu\ \acute{o}\chi\iota \end{cases}$$

$$\text{Άρα: } y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta o + \beta 1 + \varepsilon i, & \text{αν ισχύει} \\ \beta o + \varepsilon i, & \text{αν όχι} \end{cases}$$

Αν έχουμε k – κατηγορίες, χρησιμοποιούμε k-1 δείκτριες

- Κατηγορικές μεταβλητές με πάνω από 2 επίπεδα:

$$x_{i1} = \begin{cases} 1, & \alpha\nu\ \iota\sigma\chi\acute{\upsilon}\epsilon\iota\ \kappa\acute{\alpha}\pi o\iota\alpha\ \sigma\upsilon\nu\theta\acute{\eta}\kappa\eta\ \alpha \\ 0, & \alpha\nu\ \acute{o}\chi\iota \end{cases}$$

$$x_{i2} = \begin{cases} 1, & \alpha\nu\ \iota\sigma\chi\acute{\upsilon}\epsilon\iota\ \kappa\acute{\alpha}\pi o\iota\alpha\ \sigma\upsilon\nu\theta\acute{\eta}\kappa\eta\ \beta \\ 0, & \alpha\nu\ \acute{o}\chi\iota \end{cases}$$

$$\text{Άρα: } y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta o + \beta 1 + \varepsilon i, & \text{αν ισχύει η α} \\ \beta o + \beta 2 + \varepsilon i, & \text{αν ισχύει η β} \\ \beta o + \varepsilon i, & \text{αν οχι} \end{cases}$$

2.1.2.4 Πολυσυγγραμμικότητα

Η πολυσυγγραμμικότητα αναφέρετε στην κατάσταση στην οποία δύο ή περισσότερες ανεξάρτητες μεταβλητές συσχετίζονται. Λόγω αυτής μειώνεται η ακρίβεια των εκτιμήσεων των συντελεστών παλινδρόμησης διότι προκαλεί αύξηση του τυπικού σφάλματος για το $\hat{\beta}_i$. Αν σκεφτούμε ότι το t-statistics, για κάθε πρόβλεψη υπολογίζεται διαιρώντας το $\hat{\beta}_i$ με το τυπικό του σφάλμα, η πολυσυγγραμμικότητα οδηγεί σε μείωση της τιμής της συνάρτησης αυτής και αυτό σημαίνει ότι ενδέχεται να αποτύχουμε να απορρίψουμε την H_0.

Ένας άλλος τρόπος για την ανίχνευση πολυσυγγραμμικότητας είναι να δούμε τον πίνακα συσχετίσεων. Είναι πιθανό να δούμε μεγάλες απόλυτες τιμές της συσχέτισης μεταξύ 2 μεταβλητών, πιθανό είναι βέβαια και να υπάρχει συσχέτιση μεταξύ τριών ή περισσότερων μεταβλητών. Ένας καλύτερος λοιπόν τρόπος να εκτιμήσουμε την πολυσυγγραμμικότητα είναι ο παράγοντας διόγκωσης διακύμανσης (Variance Inflation Factor) με τύπο:

$$VIF_k=\frac{1}{1-R_k^2} \quad , k=1,2,\ldots,p-1$$

όπου:

R_k^2 : ο συντελεστής προσδιορισμού του μοντέλου που χρησιμοποιεί ως εξαρτημένη μεταβλητή τη $X_κ$ και ως ανεξάρτητες τις υπόλοιπες p-2 ανεξάρτητες X_i , i≠k.

<u>Κανόνες</u>:

- Αν $VIF_k \cong 1$, τότε η αντίστοιχη ανεξάρτητη μεταβλητή X_k δεν έχει πρόβλημα πολυσυγγραμμικότητας σε σχέση με τις υπόλοιπες
- Αν $VIF_k > 10$, τότε η αντίστοιχη ανεξάρτητη μεταβλητή X_k εμφανίζει πρόβλημα πολυσυγγραμμικότητας σε σχέση με τις υπόλοιπες

Ως ένα ομοιόμορφο κριτήριο έχει προταθεί η χρήση του μέσου όρου:

$$\overline{VIF_k} = \frac{1}{p-1}\sum_{k=1}^{P-1} VIF_k$$

Ο οποίος παίρνει τιμή 1 όταν ισχύει $VIF_k=1\ \forall\ k=1,2,\ldots,p-1$. Αλλιώς δεδομένου ότι ισχύει πάντοτε $VIF_k \geq 1\ \forall\ k=1,2,\ldots,p-1$, αν η ποσότητα $\overline{VIF_k}$ λάβει τιμές μεγαλύτερες του 1 αντιλαμβανόμαστε ότι κάποιος ή κάποιοι από τους δείκτες VIF_k έχουν απομακρυνθεί από το 1 οπότε έχουμε ένδειξη πολυσυγγραμμικότητας.

2.2 Απλή Λογιστική Παλινδρόμηση

Η απλή λογιστική παλινδρόμηση είναι ένα προβλεπτικό μοντέλο που προβλέπει ένα κατηγορικό πεδίο. Επιτρέπει στο πεδίο να έχει 2 κατηγορίες και χρησιμοποιείται για την πρόβλεψη πιθανότητας εμφάνισης ενός γεγονότος. Η απόκριση μπορεί να πάρει ακριβώς δύο τιμές τύπου ΝΑΙ/ΟΧΙ ,Επιτυχία/Αποτυχία. Αν μια δίτιμη απόκριση με P(Y=1)= π =E(Y) το μοντέλο της λογιστικής παλινδρόμησης εκφράζει:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = x'\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

όπου x' το διάνυσμα των ανεξάρτητων μεταβλητών.

Η συνάρτηση logit αναφέρεται στο **λογάριθμο της σχετικής πιθανότητας** του ενδεχομένου που μας ενδιαφέρει («επιτυχία»).

Αντιλογαριθμίζοντας τα δύο μέλη της εξίσωσης έχουμε:

$$\frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}$$

ενώ θέτοντας : $z = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$,έχουμε:

$$\frac{\pi}{1-\pi} = e^z$$

Ώστε να ελέγξουμε την επάρκεια ενός γενικευμένου γραμμικού μοντέλου, έχουμε δει ότι η απόκλιση του δεν ακολουθεί την κατανομή X^2, ούτε καν προσεγγιστικά γι' αυτό συνηθέστερος έλεγχος είναι αυτός των Hosmer – Lemeshow.

2.2.1 Πίνακες ταξινόμησης (classification tables)

Ένα άλλο χρήσιμο μέσο αξιολόγησης της προσαρμογής ενός μοντέλου είναι ένας **πίνακας ταξινόμησης**. Ο πίνακας ταξινόμησης είναι ένας πίνακας δύο διαστάσεων που μας δίνει τις **συχνότητες** για τις επιτυχίες και τις αποτυχίες ανάμεσα στις παρατηρούμενες και τις εκτιμώμενες τιμές.

		Παρατηρούμενες	
Εκτιμώμενες		Επιτυχία	Αποτυχία
	Επιτυχία	a	b
	Αποτυχία	c	d

Ισχύει προφανώς, $a + b + c + d = n$, (το συνολικό μέγεθος δείγματος).Διαισθητικά, όσο μεγαλύτερο είναι το άθροισμα $a + d$ σε σχέση με το άθροισμα $b + c$, τόσο μεγαλύτερη είναι η προβλεπτική αξία του υποδείγματος.

Υπάρχουν επίσης διάφορα μέτρα που έχουν προταθεί συνήθως είναι γνωστά ως pseudo-R^2 όπως:

α. McFadden

Είναι το πιο δημοφιλές ψευδο-R^2 . Ο τύπος είναι:

$$1 - \frac{log\ LM}{\log L0}$$

Τιμές μεταξύ 0.2 και 0.4 υποδηλώνουν καλή προσαρμογή.

β. McFadden adjusted

Μία παραλλαγή του παραπάνω, η οποία θέτει ποινή για το πλήθος k των παραμέτρων με τύπο:

$$1 - \frac{log\ LM - k}{\log L0}$$

γ. Cox and Snell

Ο τύπος είναι:

$$1 - (L0/LM)^{2/n}$$

δ. Nagelkerke /Cragg and Uhler

Τροποποίηση του παραπάνω, όταν διαιρεθεί με τη μέγιστη τιμή του. Ο τύπος του είναι :

$$1 - (L0/LM)^{2/n} / 1 - L0^{2/n}$$

2.3 Πολυωνυμική Παλινδρόμηση

Στην παραπάνω παράγραφο εστιάσαμε στη χρήση της λογιστικής παλινδρόμησής όταν επρόκειτο για δίτιμη μεταβλητή απόκρισης. Το μοντέλο αυτό μπορεί εύκολα να τροποποιηθεί προκειμένου να διαχειριστεί την περίπτωση όπου η μεταβλητή αυτή έχει περισσότερες από δύο κατηγορίες.

Έστω J > 2 το πλήθος των κατηγοριών της απόκρισης Y και π_1 , π_2 ,…, π_J οι αντίστοιχες πιθανότητες. Αυτές ικανοποιούν τη σχέση:

$$\sum_{j=1}^{J} \pi j = 1$$

Έστω ότι οι μεταβλητές $X_1, X_2, \ldots X_{k-1}$, έχουν (από κοινού) συνάρτηση πιθανότητας

$$P(X_1 = x_1 , X_2 = x_2 ,\ldots, X_{k-1}= x_{k-1}) = \frac{n!}{x_{1^1} \cdot x_{2^!} \ldots x_{(k-1)}! x_k!} \pi_1^{x1}\ \pi_2^{x2} \ldots \pi_k^{xk}$$

Τότε λέμε ότι η κατανομή των $X_1, X_2, \dots X_{k-1}$, είναι η πολυωνυμική με παραμέτρους n και π_i.

Συμβολίζεται ως εξής:

$$(X_1, X_2, \dots X_{k-1}) \sim \mathbf{Mn}(n, \pi_i)$$

Η μέση τιμή και διακύμανση των X_i είναι αντίστοιχα

- $E(X_i) = n \cdot \pi_i$
- $Var(X_i) = n \cdot \pi_i \cdot (1 - \pi_i)$

Ώστε να πραγματοποιήσουμε την ανάλυση, επιλέγουμε μία κατηγορία αναφοράς της Y.

- Μείωση Διαστάσεων

2.4 Ανάλυση κυρίων συνιστωσών (PCA Analysis)

Βασικές έννοιες

Η ανάλυση σε κύριες συνιστώσες είναι ένας γραμμικός μαθηματικός μετασχηματισμός ο οποίος πραγματοποιεί τη μετατροπή πιθανώς συσχετισμένων μεταβλητών σε μη συσχετισμένες μεταβλητές αλλά να περιέχουν όσο το δυνατόν μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών.

Τα οφέλη από μία τέτοια διαδικασία είναι:
1. Στις συσχετισμένες μεταβλητές οι εκτιμήσεις που προκύπτουν δεν είναι συνεπής όπως για παράδειγμα στο πρόβλημα της πολυσυγγραμμικότητας, ενώ αν πάρουμε ασυσχέτιστες το πρόβλημα αυτό δεν υφίσταται
2. Μείωση του όγκου των δεδομένων και έτσι κερδίζουμε χώρο και ταχύτητα επεξεργασίας
3. Επιτρέπει την ποσοτικοποίηση μη μετρήσιμων ποσοτήτων πχ αγάπη, ευφυΐα, κ.λ.π

Η μέθοδος αυτή αναπτύχθηκε από τον Karl Pearson (1901) και τα τελευταία χρόνια αναλυτές την χρησιμοποιούν για την περιγραφή δεδομένων που έχουν υποστεί βελτιστοποιήσεις χωρίς υποθέσεις κατανομών ή στατιστικά μοντέλα.

2.4.1 Περιγραφή μεθόδου

Η μέθοδος στηρίζεται στη φασματική ανάλυση ενός τετραγωνικού πίνακα. Έτσι μπορούμε να χρησιμοποιήσουμε είτε πίνακα διακυμάνσεων είτε πίνακα συσχετίσεων, που είναι ο πίνακας διακυμάνσεων των τυποποιημένων δεδομένων. Έστω ένα σύνολο συσχετισμένων μεταβλητών X_1 , $X_2,\ldots,X_p$ και ένα σύνολο ασυσχέτιστων $Y_1,Y_2,\ldots,Y_p$ μεταβλητών, αλλά να περιέχουν όσο το δυνατό γίνεται μεγαλύτερο μέρος της ολικής διακύμανσης των αρχικών μεταβλητών:

$$\sum_{i=1}^{p} Var(X_i) \approx \sum_{i=1}^{k} Var(Yi)$$

Οι μεταβλητές Y_i ονομάζονται κύριες συνιστώσες οι οποίες αργότερα θα αντικαταστήσουν τις αρχικές μεταβλητές. Οι κύριες συνιστώσες είναι γραμμικοί συνδυασμοί των $X_1,X_2,\ldots,X_p$ τους οποίους γράφουμε:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$
$$Y_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$
.
.
$$Y_p = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p$$

Υπό μορφή πίνακα γράφεται Y=AX όπου Y, X διανύσματα p x 1 και A είναι p x p με στοιχεία:

$$A = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1p} \\ a_{21} & a_{22} & \ldots & a_{2p} \\ . & . & . & . \\ . & . & . & . \\ a_{p1} & a_{p2} & \ldots & a_{jp} \end{bmatrix} = \left[a_1, a_{2,} \ldots, a_p\right]$$

όπου α_j *είναι το διάνυσμα στήλη με στοιχεία* : $\alpha_j^{'} = \left[a_{j1}, a_{j2,} \ldots, a_{jp}\right]$, $j = 1,2,\ldots,p$

Οι κύριες συνιστώσες προέκυψαν περιστρέφοντας τους άξονες αλλά διατηρώντας τους ορθογώνιους. Τα α_{ij} πρέπει να πληρούν τους κανόνες:

$$\sum_{i=1}^{p} a_{ij}^2 = 1 \ , j = 1,2,\ldots,p \ \kappa\alpha\iota \ \sum_{i=1}^{p} a_{ij}\, a_{ik} = 0 \ , j \neq k \ , j,k = 1,2,\ldots,p$$

Οι δύο περιορισμοί εξασφαλίζουν την ορθοκανονικότητα των συντελεστών. Το πρόβλημα ανάγεται στην εύρεση του πίνακα Α με τρόπο ώστε οι συνιστώσες που προκύπτουν να είναι σε σειρά φθίνουσας διακύμανσης και κάθε συνιστώσα να μην είναι συσχετισμένη με τις προηγούμενες. Υπάρχουν τυπικοί αλγόριθμοι ώστε να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα διακύμανσης **Σ** του διανύσματος $X=[X_1,X_2,\ldots,Xp]$.

Παρατηρήσεις

1. Η διακύμανση της i συνιστώσας : $Var(Yi) = \lambda_i$, i = 1,2,…,p με $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$
2. Μπορούμε να βρούμε ότι για όλες τις κύριες συνιστώσες τα α_j διανύσματα θα αντιστοιχούν στα ιδιοδιανύσματα της j σε φθίνουσα σειρά ιδιοτιμής

3. Η συνολική διακύμανση είναι: Var(Y)= $\sum_{i=1}^{p} Var(Yi) = \sum_{i=1}^{\rho} \lambda_i = \lambda_1 + \lambda_2 + \cdots + \lambda_\rho$

οπότε το ποσοστό συνολικής διακύμανσης που εξηγείται από τη j συνιστώσα είναι:

- $\frac{\lambda_j}{\lambda_1+\lambda_2+\cdots+\lambda_p}$

Και άρα το ποσοστό που εξηγείται από τις K συνιστώσες είναι:

- $\frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p}$

4. Συμφέρει να διατηρούμε τις πρώτες συνιστώσες που εξηγούν μεγαλύτερο ποσοστό διακύμανσης
5. Ο πίνακας διακύμανσης (Λ) των κυρίων συνιστωσών είναι διαγώνιος με στοιχεία τις ιδιοτιμές λ_j αφού οι κύριες συνιστώσες είναι ασυσχέτιστες
6. Η συνολική διακύμανση των αρχικών μεταβλητών είναι ίση με τη συνολική διακύμανση των κυρίων συνιστωσών αφού tr(Σ) = tr(Λ) (Ίχνος συμμετρικού και τετραγωνικού πίνακα).
7. Οι κύριες συνιστώσες δεν είναι μοναδικές αφού αλλάζοντας τα πρόσημα των ιδιοδιανυσμάτων αποτελούν πάλι αποδεκτή λύση. Επειδή οι τιμές των συντελεστών δεν αλλάζουν κατά απόλυτη τιμή μπορούμε να καταλάβουμε την επίδραση των αρχικών μεταβλητών στις κύριες συνιστώσες.

Πίνακας συσχετίσεων

Όταν χρησιμοποιούμε πίνακα συσχετίσεων οι συσχετίσεις δεν αλλάζουν όταν αλλάξουμε τις μονάδες μέτρησης ή την κλίμακα όπως διαφοροποιούνται με τον πίνακα διακύμανσης. Όλες οι μεταβλητές έχουν το ίδιο βάρος καθώς τα στοιχεία της διαγωνίου είναι 1. Στην πράξη η επιλογή πίνακα δεν είναι εύκολη και ξεκάθαρη η υπόθεση. Συνήθως αποφεύγουμε τον πίνακα διακύμανσης όταν υπάρχουν μεταβλητές με πολύ μεγαλύτερη διακύμανση από τις υπόλοιπες. Αν οι διακυμάνσεις είναι συγκρίσιμες ή μπορεί τα δεδομένα να μετασχηματιστούν ώστε να γίνουν συγκρίσιμες , τότε χρησιμοποιούμε πίνακα διακύμανσης. Η συσχέτιση ανάμεσα στην i κύρια συνιστώσα Y_i και στη j αρχική μεταβλητή X_j δίνεται από τον τύπο:

$$r(Y_i, X_j) = \frac{\alpha_{ij}\sqrt{\lambda_i}}{S_j}$$

όπου α_{ij} ο συντελεστής της μεταβλητής X_j στην κύρια συνιστώσα Y_i και S_j^2 η διακύμανση της μεταβλητής X_j.

- Αν $\alpha_{ij} = 0$ τότε δεν υπάρχει συσχέτιση
- Αν $\alpha_{ij} = \pm 1$ τότε η συσχέτιση γίνεται ± 1

Βήματα Μεθόδου

1. Αξιολόγηση συσχετίσεων

i) Σε πίνακα συσχετίσεων ένα μέτρο που μας επιτρέπει να συγκρίνουμε 2 σετ δεδομένων ως προς τις συσχετίσεις τους είναι το:

$$\Phi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{i=1}^{p} r_{ij}^2 - p}{p(p-1)}}$$

Όπου:

r_{ij} είναι το ij στοιχείο του πίνακα συσχετίσεων.

Το Φ παίρνει τιμές κοντά στο 1 αν υπάρχουν μεγάλες συσχετίσεις και 0 αν δεν υπάρχουν συσχετίσεις.

Πρακτικά ώστε να προχωρήσουμε στην ανάλυση θέλουμε συσχετίσεις της τάξης του 0,4 ή και μεγαλύτερες κατά απόλυτη τιμή.

ii) Σε πίνακα διακυμάνσεις ένα αντίστοιχο μέτρο είναι:

$$\Phi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{i=1}^{p} S_{ij}^2 - \sum_{i=1}^{p} S_{jj}^2}{\sum_{i=1}^{p}\sum_{i=1}^{p} S_{ii} S_{jj}}}$$

για το οποίο ισχύουν τα παραπάνω.

2. Επιλογή πίνακα. (Δηλαδή αν θα χρησιμοποιήσουμε πίνακα διακύμανσης ή πίνακα συσχετίσεων).

3. Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων

4. Απόφαση για το πλήθος των συνιστωσών που θα κρατήσουμε. Τα κριτήρια για την επιλογή του κατάλληλου αριθμού κυρίων συνιστωσών είναι αρκετά εμπειρικά και αφήνεται στον ερευνητή.

i)Επιλέγουμε τις k πρώτες συνιστώσες που εξηγούν ένα μεγάλο ποσοστό της συνολικής διακύμανσης έστω 70-80%

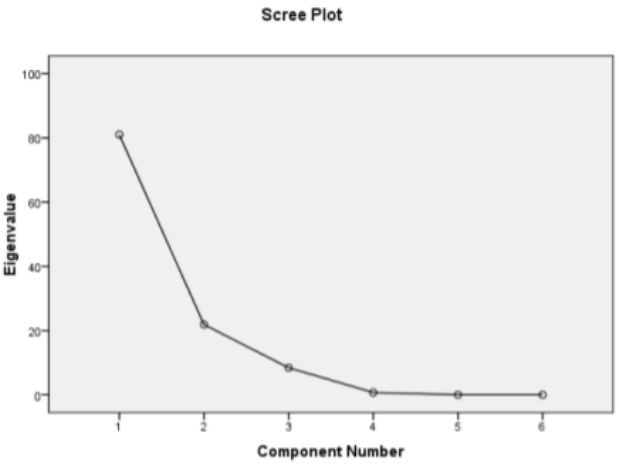
ii)Κριτήριο Kaiser, το οποίο λέει ότι αν χρησιμοποιήσουμε τον πίνακα διακύμανσης τότε θεωρούμε τόσες συνιστώσες όσες η ιδιοτιμή τους είναι μεγαλύτερη από τη μέση τιμή

$$\bar{\lambda} = \sum_{j=1}^{p} \frac{\lambda_j}{p}$$

Ενώ αν χρησιμοποιήσουμε πίνακα συσχετίσεων ο οποίος είναι μοναδιαίος, διαλέγουμε τόσες συνιστώσες όσες και οι ιδιοτιμές είναι μεγαλύτερες του 1.

iii) Scree plot.

Η τεχνική αυτή είναι μία γραφική μέθοδος. Το γράφημα αυτό έχει στον οριζόντιο άξονα των x τον αριθμό των συνιστωσών και στον κάθετο άξονα y την τιμή κάθε ιδιοτιμής. Επιλέγονται τόσες συνιστώσες μέχρι το γράφημα να αρχίζει να αλλάζει κλίση.



iv) Cross validation

Η μέθοδος στηρίζεται σε επαναληπτικούς υπολογισμούς, όπου κάθε φορά αγνοούμε κάποιες τιμές των δεδομένων μας και εξετάζουμε τη συμπεριφορά των συνιστωσών προσπαθώντας να προβλέψουμε τα δεδομένα που δεν χρησιμοποιήσαμε στην ανάλυση. Επαναλαμβάνοντας τη διαδικασία πολλές φορές έχουμε ένα σκορ που μας δείχνει αν το μοντέλο με k- συνιστώσες δίνει καλά αποτελέσματα. Έτσι συγκρίνοντας τα αποτελέσματα για τις διάφορες τιμές του k, βρίσκουμε την τιμή για την οποία τα αποτελέσματα είναι καλύτερα.

5. Εύρεση κυρίων συνιστωσών

6. Ερμηνεία κυρίων συνιστωσών που είναι και το πιο δύσκολο κομμάτι αυτής της ανάλυσης. Συνήθως καταφεύγουμε στην περιστροφή των αξόνων πολλαπλασιάζοντας τον πίνακα συντελεστών με έναν ορθογώνιο πίνακα τέτοιον ώστε να υπάρχουν λίγες συνιστώσες με μεγάλες απόλυτες τιμές ενώ οι υπόλοιπες να έχουν συντελεστές που τείνουν στο μηδέν.

7.Δημιουργία νέων μεταβλητών όσες και οι κύριες συνιστώσες πού διατηρήσαμε. Ώστε να γίνει αυτό αρκεί να αντικαταστήσουμε στον τύπο της κάθε συνιστώσας τις τιμές που η παρατήρηση είχε για κάθε μεταβλητή.

Σχόλια

α. Δεν έχει νόημα να συμπεριλάβουμε στην ανάλυση μία μεταβλητή η οποία είναι ασυσχέτιστη με τις υπόλοιπες αφού αν παραμείνει κάποια από τις κύριες συνιστώσες θα ταυτιστεί μαζί της.

β. Αν δύο ιδιοτιμές προκύψουν ίδιες τότε αυτές αντιστοιχούν σε όμοιες κύριες συνιστώσες και αποτελεί πλεονασμό.

γ. Αν έχουμε μηδενικές ιδιοτιμές ή πρακτικά να τείνουν στο μηδέν τότε κάποιες μεταβλητές είναι γραμμικώς εξαρτημένες. Τέτοιες ιδιοτιμές αντιστοιχούν σε συνιστώσες με μηδενική σχεδόν διακύμανση και μπορούμε να τις αγνοήσουμε.

δ. Στην περίπτωση που έχουμε ίδια ιδιοδιανύσματα αλλά διαφορετικές ιδιοτιμές σημαίνει πως παίρνουμε ίδιες συνιστώσες αλλά σε κάθε περίπτωση η συνιστώσα εξηγεί διαφορετικό ποσοστό της διακύμανσης.

ε. Αν όλες οι συσχετίσεις είναι θετικές τότε η πρώτη κυρία συνιστώσα μπορεί να εκληφθεί ως σταθμικός μέσος των μεταβλητών με σταθμίσεις τους αντίστοιχους συντελεστές.

2.5 Παραγοντική Ανάλυση

Είναι μια στατιστική μέθοδος που έχει ως σκοπό να ανιχνεύσει την ύπαρξη κάποιων κοινών παραγόντων σε μια διαθέσιμη ομάδα μεταβλητών, οι οποίοι τις επηρεάζουν. Πιο συγκεκριμένα να συνοψίσει τις σχέσεις μεταξύ των υπό μελέτη μεταβλητών μ' έναν ακριβή τρόπο. Ο παράγοντας που μας βοηθά να εξηγήσουμε τις υπό μελέτη μεταβλητές δε θα πρέπει να είναι απαραίτητα μια υπαρκτή ποσότητα αλλά κατασκευάζεται με στόχο την απλοποίηση της δομής των αρχικών δεδομένων. Η ύπαρξη ισχυρής συσχέτισης μεταξύ των μεταβλητών που θα συμμετέχουν στην παραγοντική ανάλυση είναι απαραίτητη για την ποιότητα των αποτελεσμάτων της ανάλυσης. Η ύπαρξη μικρής συσχέτισης μεταξύ των αρχικών μεταβλητών, είναι ένδειξη ότι δεν υπάρχουν κοινοί παράγοντες που να ερμηνεύουν σημαντικό μέρος της διακύμανσης της κάθε μεταβλητής.

Με λίγα λόγια η παραγοντική ανάλυση χρησιμοποιείται:

1. Να εντοπίσουμε κρυφούς παράγοντες που περιγράφουν χαρακτηριστικά, σκέψεις, απόψεις και λοιπά που δεν είναι δυνατόν να παρατηρηθούν ή να μετρηθούν άμεσα.
2. Να συνοψίσουμε την πληροφορία μίας μεγάλης ομάδας μεταβλητών σε λιγότερους κοινούς παράγοντες.

Τα αποτελέσματα της εφαρμογής της ανάλυσης παραγόντων είναι τα εξής:

1. Μειώνονται οι διαστάσεις του προβλήματος, αντικαθιστώντας τις αρχικές μεταβλητές με λιγότερες αφού οι παράγοντες κατασκευάζονται ώστε να διατηρούν το μεγαλύτερο μέρος της πληροφορίας των αρχικών μεταβλητών.
2. Δημιουργούμε νέες μεταβλητές, τους παράγοντες που μπορούμε να τις αναγνωρίσουμε με υποκειμενικό τρόπο.

2.5.1 Γενικό μοντέλο ανάλυσης παραγόντων

Στο μοντέλο αυτό που ονομάζεται και ορθογώνιο, υποθέτουμε ότι οι συσχετίσεις μεταξύ των μεταβλητών οφείλονται στην ύπαρξη κάποιων κοινών παραγόντων που δεν γνωρίζουμε και θέλουμε να τους βρούμε.

Τύπος μοντέλου:

$$X-\mu-=LF+\varepsilon$$

όπου:

X: το διάνυσμα των αρχικών μεταβλητών (p x 1)

μ: το διάνυσμα των μέσων (p x 1)

L: είναι ένας πίνακας (p x k) όπου τα L_{ij} λέγονται φορτία του παράγοντα F στη μεταβλητή X_i

F: είναι ένα διάνυσμα τυχαίων μεταβλητών (k x 1) που λέγονται παράγοντες

ε: Τα ε_i λέγονται σφάλματα ή ειδικοί παράγοντες , είναι το μέρος που δεν μπορεί να εξηγηθεί από τους παράγοντες.

Αξίζει να σημειωθεί ότι το μοντέλο αν και μοιάζει με γραμμικό έχει δύο βασικές διαφορές:
1. Τα X_i δεν είναι παρατηρήσεις αλλά τυχαίες μεταβλητές.
2. Το δεξί μέλος της εξίσωσης είναι άγνωστο και πρέπει να εκτιμηθεί.

Υποθέσεις

- $E(F_k) = 0$
- $E(\varepsilon_i) = 0$
- $D(F)=I_m$
- $D(\varepsilon) = \psi \Leftrightarrow COV(\varepsilon_i,\varepsilon_j) = 0 , i \neq j$
- $COV(\varepsilon_i,F_j) = 0 , i \neq j.$

Από τις παραπάνω υποθέσεις μπορεί να δειχθεί ότι :

- $\Sigma=COV(X)=COV(LF+\varepsilon)=L\ COV(F)\ L'+COV(\varepsilon)=LL'+\psi$

Παρατηρούμε λοιπόν ότι ο πίνακας συνδιακύμανσης διασπάται σε δύο μέρη. Το πρώτο κομμάτι ονομάζεται εταιρικότητα (communality) και είναι αυτό που ερμηνεύεται από τους παράγοντες και το δεύτερο κομμάτι ονομάζεται ιδιαιτερότητα (specificity) και είναι αυτό που οφείλεται στους ειδικούς παράγοντες και το μοντέλο δεν μπορεί να ερμηνεύσει.
Σκοπός λοιπόν, είναι να εκτιμήσουμε τους πίνακες L και Ψ ώστε να αναπαραστήσουμε τον πίνακα Σ στη μορφή αυτή.

2.5.2 Εφαρμογή της παραγοντικής ανάλυσης

Η εφαρμογή της μεθόδου, μέσω των κοινών παραγόντων προσπαθεί να εξηγήσει τις συσχετίσεις ανάμεσα στις μεταβλητές. Για το λόγο αυτό λοιπόν πριν εφαρμοστεί η παραγοντική ανάλυση πρέπει να ελεγχθεί η ύπαρξη επαρκούς συσχέτισης μεταξύ των μεταβλητών που συμπεριλαμβάνονται στην ανάλυση. Συσχετίσεις μεγαλύτερες του 40% κατά απόλυτη τιμή είναι ικανοποιητικές. Τέλος, αν υπάρχουν μεταβλητές που είναι ασυσχέτιστες με τις υπόλοιπες, θα πρέπει να τις αφαιρέσουμε, ώστε να μην καταλήξουν να είναι απλά ξεχωριστοί παράγοντες.

Έλεγχοι συσχετίσεων

- **Συντελεστής συσχέτισης**

Ώστε να ελέγξουμε τη στατιστική σημαντικότητα ενός δειγματικού συντελεστή συσχέτισης χρειαζόμαστε έλεγχο για την υπόθεση:

1. Έλεγχος υπόθεσης

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho \neq 0$$

Χρησιμοποιούμε τη συνάρτηση που ακολουθεί, της t-κατανομής με n-2 βαθμούς ελευθερίας:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

όπου r: συντελεστής συσχέτισης Pearson
Ο παραπάνω όμως έλεγχος δεν μπορεί να χρησιμοποιηθεί για ελέγχους διαφορετικής υπόθεσης από ρ = 0.

2. Για ελέγξουμε υποθέσεις της μορφής:

$$H_0 : \rho = \rho_0 \text{ vs } H_1 : \rho \neq \rho_0$$

Χρησιμοποιούμε την ελεγχοσυνάρτηση Z:

$$Z = \frac{z(\rho) - z(\rho_0)}{\sqrt{\frac{1}{n - 3}}}$$

με $z(\alpha) = 0.5 \log\left(\frac{1+\alpha}{1-\alpha}\right)$

3. Ένας ακόμη έλεγχος που μπορούμε να χρησιμοποιήσουμε είναι ο έλεγχος σφαιρικότητας του Bartlett.

Η υπόθεση αυτού του ελέγχου είναι της μορφής:

$$H_0 : \Sigma = \sigma^2 I_p \text{ vs } H_1 : \Sigma \neq \sigma^2 I_p$$

Με στατιστική συνάρτηση:

$$L = -\left[n - \frac{1}{6\rho}(2p^2 + p + 2)\right] \cdot \left[log|S| - log \prod_{i=\perp}^{p} S_i^2\right]$$

Συγκρίνω την τιμή αυτής με το ποσοστιαίο σημείο της X^2 κατανομής με p(p-1)/2 βαθμούς ελευθερίας.

όπου:
S : Δειγματικός πίνακας διακυμάνσεων συνδιακυμάνσεων
S_i^2: Δειγματικός πίνακας διακυμάνσεων της i μεταβλητής

- **Μερικός συντελεστής συσχέτισης**

Αυτός ο συντελεστής συσχέτισης σε σχέση με τον απλό υπολογίζει τη συσχέτιση αφού αφαιρέσει την επίδραση των υπόλοιπων μεταβλητών. Θέλουμε οι μερικοί συντελεστές συσχέτισης να είναι μικροί. Ένα μέτρο ώστε να συγκρίνουμε το μέγεθος των συντελεστών συσχέτισης σχετικά με τους μερικούς συντελεστές είναι το Kaiser-Meyer-Olkin (KMO):

$$KMO = \frac{\sum \sum_{i \neq j} r_{ij^2}}{\sum \sum_{i \neq j} r_{ij^2} + \sum \sum_{i \neq j} a_{ij^2}}$$

όπου:

r_{ij}: δειγματικός συντελεστής συσχέτισης

a_{ij}: δειγματικός συντελεστής μερικής συσχέτισης

Τιμές κάτω από 0.5 χαρακτηρίζονται αρκετά μικρές-κακές τιμές, δηλαδή δε θα έχουμε ικανοποιητικά αποτελέσματα από την παραγοντική ανάλυση, ενώ τιμές γύρω στο 0.8 θεωρούνται αρκετά ώστε να προχωρήσουμε στην ανάλυση.

2.5.3 Εύρεση κατάλληλου αριθμού παραγόντων

Ο καθορισμός του αριθμού των παραγόντων που θα χρησιμοποιηθούν δεν είναι γνωστός. Θα μπορούσε να χρησιμοποιηθεί κάποια τεχνική όπως το scree plot που αναφέραμε και στην ανάλυση κυρίων συνιστωσών. Επίσης, θα μπορούσε ο ερευνητής να ξεκινήσει την ανάλυση αυξάνοντας διαδοχικά τον αριθμό παραγόντων και να κρατήσει το μοντέλο με βάση κάποιο κριτήριο καλής προσαρμογής.

<u>Κριτήρια τέτοιου είδους είναι:</u>

1. Από τον πίνακα επιβαρύνσεων μπορεί κάποιος να εκτιμήσει τον πίνακα **Σ.** Οι αποκλίσεις του πραγματικού πίνακα με τον εκτιμημένο θα πρέπει να είναι μικρές. Χωρίς βέβαια να υπάρχει κάτι που να προσδιορίζει το πόσο μικρές.
2. Έλεγχος λόγου πιθανοφανειών ,αν οι εκτιμήσεις έχουν γίνει με τη μέθοδο μεγίστης πιθανοφάνειας.

Κεφάλαιο 3

Μέθοδοι Ανάλυσης: Ταξινομήσεις – Προβλέψεις

3.1 Διαχωριστική Ανάλυση

Η διαχωριστική ανάλυση (discriminant analysis) είναι μία τεχνική η οποία, με τη βοήθεια ενός διαχωριστικού κανόνα έχει σκοπό να χωρίσει έναν πληθυσμό σε ευδιάκριτα σύνολα και να κατατάξει παρατηρήσεις (συνήθως πολυδιάστατες) στα παραπάνω σύνολα ώστε να λάβουμε αποφάσεις για το μέλλον. Η διαχωριστική ανάλυση είναι μία μέθοδος πολύπλοκη, καθώς δε γνωρίζουμε τις παραμέτρους για κάθε πληθυσμό και πρέπει να τις εκτιμήσουμε από ένα δείγμα. Η χρήση ηλεκτρονικών υπολογιστών επιτρέπει τη λύση πολύπλοκων υποθέσεων. Η διαχωριστική ανάλυση έχει εφαρμογές σε πολλές επιστήμες (στην πληροφορική αναφέρεται ως αναγνώριση προτύπων). Αναπτύχθηκε από τον R.Fisher ο οποίος πρότεινε την έννοια της γραμμικής διαχωριστικής ανάλυσης για την ταξινόμηση Ανθέων.

3.1.1 Κανόνες / Κριτήρια Διαχωρισμού Ομάδων

3.1.1.1 Κανόνας μέγιστης πιθανοφάνειας

Είναι το πιο απλό κριτήριο διαχωρισμού ομάδων και κατατάσσει κάθε νέα παρατήρηση στην ομάδα από την οποία είναι πιο πιθανό να έχει προέλθει.

Έχουμε:

$$R_1: \frac{f_1(x)}{f_2(x)} \geq 1 \text{ και } R_2: \frac{f_1(x)}{f_2(x)} < 1$$

όπου $f_i(x)$, η πιθανοφάνεια του i πληθυσμού

Άρα αν $f_1(x) \geq f_2(x)$ κατατάσσουμε τη νέα παρατήρηση στον πληθυσμό Π_1 διαφορετικά στον πληθυσμό Π_2.

3.1.1.2 Κανόνας του Bayes

Στην περίπτωση που ομάδες έχουν διαφορετικά μεγέθη μας ενδιαφέρει η πιθανότητα να πάρουμε παρατήρηση από κάθε ομάδα (π_i , i=1,2…). Θα πρέπει να βρούμε τις εκ των υστέρων πιθανότητες $P(x \mid \pi_i)$, η παρατήρηση αυτή να προέρχεται από τον πληθυσμό αυτό και τις εκ των προτέρων πιθανότητες $P(\pi_i)$, i=1,2...

Τύπος του Bayes :

$$P(\pi_i / x) = \frac{P(\pi_i, x)}{P(x)} = \frac{P(x / \pi_i)P(\pi_i)}{\sum_{i=1}^{2} P(x/\pi_i)f_i(x_i)}$$

Όμως:

$$P(\pi / x) = \frac{\pi\iota\theta\alpha\nu o\phi\acute{\alpha}\nu\epsilon\iota\alpha\ x\ \epsilon\kappa\ \tau\omega\nu\ \upsilon\sigma\tau\acute{\epsilon}\rho\omega\nu\ \pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha}{\gamma\epsilon\gamma o\nu\acute{o}\varsigma}$$

και αν θέσουμε: $P(\pi_1) = P_1$, $P(\pi_2) = P_2$

θα έχουμε:

$$P(\pi_1 / x) = \frac{f_1(x)P_1}{f_1(x)P_1+f_2(x)P_2} \ , \ P(\pi_2 / x) = \frac{f_2(x)P_2}{f_1(x)P_1+f_2(x)P_2}$$

Αν :

$P(\pi_1 / x) \geq P(\pi_2 / x) \Leftrightarrow f_1(x)\, P_1(x) \geq f_2(x)\, P_2(x) \Leftrightarrow$

$$\frac{\boldsymbol{f_1(x)}}{\boldsymbol{f_2(x)}} \geq \frac{\boldsymbol{P_2}}{\boldsymbol{P_1}}$$

τότε κατατάσσω την παρατήρηση x στον 1ο πληθυσμό, αλλιώς στο 2ο.

Σχόλια

1. Για κάθε x, οι τιμές των εκ των προτέρων πιθανοτήτων έχουν άθροισμα μονάδα : $P_1+P_2=1$.
2. Αν $P_1=P_2$ τότε η απόφαση θα εξαρτάται μόνο από τις συναρτήσεις πιθανοφάνειας, άρα έχω τον κανόνα μεγίστης πιθανοφάνειας.

3.1.1.3 Κανόνας Ελαχιστοποίησης του κόστους λανθασμένης κατάταξης

Μερικές φορές μπορεί η κατάταξη μιας παρατήρησης που θα έπρεπε να είναι στον πληθυσμό Π_1 αλλά τοποθετείται στον πληθυσμό Π_2 να είναι πιο σοβαρό σφάλμα από ότι η παρατήρηση να είναι του πληθυσμού Π_2 όμως να τοποθετηθεί λανθασμένα στον Π_1. Άρα θα πρέπει να συμπεριλάβουμε και το κόστος που θα έχουμε από την κάθε μία λανθασμένη κατάταξη. Πρέπει να ορίσουμε έναν τύπο που θα δίνει το αναμενόμενο κόστος λανθασμένης κατάταξης ώστε τελικά να κατατάξουμε την κάθε παρατήρηση στην ομάδα (πληθυσμό) με το μικρότερο αναμενόμενο κόστος.

Το κόστος λανθασμένης κατάταξης ECM_k δίνεται από τον τύπο:

$$ECM_k = P_k \sum_{m=1}^{k} C(m / k) \cdot P(m / k)$$

όπου:

- c(m/k) : Το κόστος να κατατάξουμε την παρατήρηση στην m ομάδα ενώ ανήκει στην k
- P(m/k) : Η πιθανότητα να κατατάξουμε την παρατήρηση στην m ομάδα ενώ ανήκει στην k
- P_k : Η εκ των προτέρων πιθανότητα να ανήκει μία παρατήρηση στην k ομάδα

Το συνολικό αναμενόμενο κόστος λανθασμένης κατάταξης θα είναι ίσο με το άθροισμα των επιμέρους ECM_k.

Οπότε:

$$ECM = ECM_1 + ECM_2 = P_1\, c(2/1)\, P(2/1) + P_2\, c(1/2)\, P(1/2)$$

Σχόλια

1. Το κόστος που προκύπτει όταν μία κατάταξη είναι σωστή είναι μηδέν. Άρα ,

   c(1/1) = c(2/2)=0, γενικά c(m/m)=0

2. i) $P(2 / 1) = P(x \epsilon R_2 / \pi_1) = \int_{R_2} f_1(x)\, dx$

   ii) $P(1 / 2) = P(x \epsilon R_1 / \pi_2) = \int_{R_1} f_2(x)\, dx$

3. Το πρώτο ολοκλήρωμα παριστάνει τον όγκο που σχηματίζεται από την $f_1(x)$ πάνω στην περιοχή R_2 και το δεύτερο ολοκλήρωμα παριστάνει τον όγκο που σχηματίζεται από την $f_2(x)$ πάνω στην περιοχή R_1.

ΘΕΩΡΗΜΑ 1

Αν,

$$\frac{f_1(x)}{f_2(x)} \geq \left(\frac{c(1/2)}{c(2/1)}\right) \cdot \left(\frac{P_2}{P_1}\right)$$

κατατάσσω την παρατήρηση στην ομάδα (πληθυσμό) Π_1 , διαφορετικά στην ομάδα (πληθυσμό) Π_2.

Σχόλια

1. Αν έχουμε $\frac{P_1}{P_2} = 1$ τότε ο κανόνας περιγράφεται ως εξής:

- $\frac{f_1(x)}{f_2(x)} \geq \frac{c(1/2)}{c(2/1)}$

2. Αν τα κόστη είναι ίσα $\frac{c(1/2)}{c(2/1)}$=1 τότε ο κανόνας κατάταξης γίνεται :

- $\frac{f_1(x)}{f_2(x)} \geq \frac{P_2}{P_1}$, κατέταξε στον πληθυσμό Π_1 αλλιώς στον Π_2

ο οποίος ταυτίζεται με τον κανόνα Bayes.

3. Αν, $\frac{P_2}{P_1} = \frac{c(1/2)}{c(2/1)} = 1$ τότε ο κανόνας είναι :

- $\frac{f_1(x)}{f_2(x)} \geq 1$

κατέταξε στον πληθυσμό Π_1 αλλιώς στον Π_2 , ο κανόνας αυτός ταυτίζεται με τον κανόνα μεγίστης πιθανοφάνειας.

3.1.1.3.1 Ελαχιστοποίηση της συνολικής πιθανότητας λανθασμένης κατάταξης

Υπάρχουν κανόνες διαχωρισμού δύο πληθυσμών που κατατάσσουν τις παρατηρήσεις σύμφωνα με την πιθανότητα λανθασμένης κατάταξης με σκοπό να ελαχιστοποιήσουν τη συνολική πιθανότητα λανθασμένων κατατάξεων (TPC) η οποία ορίζεται ως εξής:

$$TPM = P\left(\eta\ παρατήρηση\ κατατάσσεται\ στην\ \Pi_2\ ενώ\ ανήκει\ στην\ \Pi_1\right) +$$
$$+ P\left(\eta\ παρατήρηση\ κατατάσσεται\ στην\ \Pi_1\ ενώ\ ανήκει\ στην\ \Pi_2\right) =$$
$$= P_1 \cdot P(2/1) + P_2 \cdot P(1/2) = P_1 \int_{R_2} f_1(x)\,dx + P_2 \int_{R_1} f_2(x)\,dx$$

Αυτός ο κανόνας είναι ίδιος με την ελαχιστοποίηση του αναμενόμενου κόστους λανθασμένης κατάταξης στην περίπτωση που τα κόστη των λανθασμένων κατατάξεων είναι ίσα.

Σχόλια

1. Όλα τα κριτήρια είναι της μορφής:

$$\frac{\mathbf{f_1(x)}}{\mathbf{f_2(x)}} \geq \mathbf{c}$$

Παρουσιάζουν όμως μία κλιμάκωση όσον αφορά με τι παίρνουμε υπόψη για την κατάταξη των παρατηρήσεων. Στον κανόνα μεγίστης πιθανοφάνειας μας ενδιαφέρει απλά τι είναι πιο πιθανό. Στον κανόνα Bayes μας ενδιαφέρει και η πιθανότητα κάθε ομάδας, ενώ στον επόμενο κανόνα τα κόστη λανθασμένης κατάταξης.

2. Όλοι οι κανόνες στο ότι: η παρατήρηση κατατάσσεται ή όχι στον πληθυσμό άσχετα αν είναι δυνατό να καταταχθεί λανθασμένα.

3.1.2 Ταξινόμηση Πληθυσμών

3.1.2.1 Διαχωρισμός 2 πληθυσμών με τη χρήση κανονικής κατανομής

Στις πολυμεταβλητές κανονικές κατανομές, στηρίζονται στο μεγαλύτερο μέρος τους οι διαδικασίες ταξινόμησης, εξαιτίας της απλότητας τους και της υψηλής αποδοτικότητας τους. Έστω ότι έχουμε δύο πληθυσμούς (ομάδες) οι οποίοι προέρχονται από κανονικούς πληθυσμούς και έστω x ένα διάνυσμα $x' = [x_1, x_2, \ldots, x_p]$ για τον κάθε πληθυσμό .

Για τον πληθυσμό Π_1 είναι $X \sim N_p(\mu_1, \Sigma_1)$ με συνάρτηση πυκνότητας πιθανότητας $f_1(x)$.

Για τον πληθυσμό Π_2 είναι $X \sim N_p(\mu_2, \Sigma_2)$ με συνάρτηση πυκνότητας πιθανότητας $f_2(x)$.

όπου $\mu_i = E(X)$, $i=1,2,..$, $\Sigma_i = COV(X)$ $i=1,2,..$

i. Κανονικοί πληθυσμοί με $\Sigma_1 = \Sigma_2 = \Sigma = COV(X)$

Οι συναρτήσεις πυκνότητας πιθανότητας για τους πληθυσμούς Π_1 , Π_2 δίνονται από τη σχέση:

$$f_i(x) = \frac{1}{2\pi^{P/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\right]$$

Έστω μ_1, μ_2 ,Σ είναι γνωστά στην περίπτωση αυτή οδηγούμαστε σε απλή γραμμική στατιστική ταξινόμηση. Η ποσότητα $(x-\mu_i)' \Sigma^{-1}(x-\mu_i)$ ορίζει ένα μέτρο απόστασης γνωστό ως απόσταση Mahalanobis.

Θα ξεκινήσουμε από τον κανόνα που ελαχιστοποιεί το αναμενόμενο κόστος κατάταξης (ECM) οπότε ο κανόνας είναι:

Αν: $\frac{f_1(x)}{f_2(x)} = \frac{c(1/2)}{c(2/1)}\left(\frac{P_1}{P_2}\right)$, τότε κατέταξε την παρατήρηση στον Π_1 αλλιώς στον Π_2. Λογαριθμίζοντας και τα 2 μέλη παίρνουμε:

$$-\frac{1}{2}\left(x-\mu_1\right)'\Sigma^{-1}\left(x-\mu_1\right)+\frac{1}{2}\left(x-\mu_2\right)'\Sigma^{-1}\left(x-\mu_2\right)\geq \ln\left[\frac{c(1/2)}{c(2/1)}\cdot\left(\frac{P_2}{P_1}\right)\right]=k$$

Η οποία μετατρέπεται (δες Απόδειξη **) στη σχέση

$(\mu_1 - \mu_2)' \Sigma^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2) \geq k$

Θέτουμε $L = \Sigma^{-1}(\mu_1 - \mu_2)$ και έχουμε εάν $L'X - \frac{1}{2}L'(\mu_1 + \mu_2) \geq k$

τότε κατατάσσουμε την παρατήρηση στον 1ο πληθυσμό αλλιώς στο 2ο.

Η συνάρτηση $U(X)=L'X - \frac{1}{2}L'(\mu_1+\mu_2)$ λέγεται διαχωριστική συνάρτηση Fisher.

Αν πάρουμε τη μέση της τιμή $E[U(X)]=L' E(X) - \frac{1}{2}L'(\mu_1+\mu_2)$

Αν η παρατήρηση X προέρχεται από την 1η ομάδα τότε:

$$E[U(X)]=\frac{1}{2}\left(\mu_1-\mu_2\right)'\Sigma^{-1}\left(\mu_1-\mu_2\right)=\frac{1}{2}\alpha$$

ενώ αν προέρχεται από τη 2η ομάδα τότε:

$$E[U(X)]=-\frac{1}{2}\left(\mu_1-\mu_2\right)'\Sigma^{-1}\left(\mu_1-\mu_2\right)=-\frac{1}{2}\alpha$$

όπου α:

$$\alpha=\left(\mu_1-\mu_2\right)'\Sigma^{-1}\left(\mu_1-\mu_2\right)$$

είναι η απόσταση Mahalanobis μεταξύ των μέσων των 2 ομάδων.

Η διακύμανστη της U(X) υπολογίζεται ως εξής:

$$VAR[U(X)]=VAR[L'X]=L'VAR(X)L=\left(\mu_1-\mu_2\right)'\Sigma^{-1}\Sigma\,\Sigma^{-1}\left(\mu_1-\mu_2\right)=$$
$$=\left(\mu_1-\mu_2\right)'\Sigma^{-1}\left(\mu_1-\mu_2\right)=\alpha$$

Άρα:

$\circ\ U(X) \sim N\left(\frac{a}{2}, a\right)$ *αν η παρατήρηση προέρχεται από την 1η ομάδα*

$\circ\ U(X) \sim N\left(-\frac{a}{2}, a\right)$ *αν η παρατήρηση προέρχεται από τη 2η ομάδα*

Η πιθανότητα να κατατάξουμε λανθασμένα μια παρατήρηση στην 1η ομάδα ενώ ανήκει στη 2η ειναι:

$$P(\text{κατάταξη στην } 1\eta \text{ ομάδα} / \text{ανήκει στη } 2\eta) = P\left(U(X) \geq k / U(X) \sim N\left(-\frac{\alpha}{2}, \alpha\right)\right) =$$

$$= P\left(\frac{U(X) + a/2}{\sqrt{a}} \geq \frac{k + a/2}{\sqrt{a}}\right) = P\left(Z \geq \frac{k + a/2}{\sqrt{a}}\right) = 1 - P\left(Z < \frac{k + a/2}{\sqrt{a}}\right) = 1 - \Phi\left(\frac{k + a/2}{\sqrt{a}}\right)$$

Όπου Φ(X) η συνάρτηση κατανομής της τυποποιημένης κανονικής.

Η πιθανότητα να κατατάξουμε λανθασμένα μια παρατήρηση στη 2η ομάδα ενώ ανήκει στην 1η είναι:

$$P(\text{κατάταξη στην } 1\eta \text{ ομάδα} / \text{ανήκει στη } 2\eta) = P\left(U(X) < k / U(X) \sim N\left(\frac{a}{2}, a\right)\right) =$$

$$= P\left(\frac{U(X) - a/2}{\sqrt{a}} \geq \frac{k - a/2}{\sqrt{a}}\right) = P\left(Z < \frac{k - a/2}{\sqrt{a}}\right) = \Phi\left(\frac{k - a/2}{\sqrt{a}}\right)$$

Άρα η συνολική πιθανότητα λάθους είναι:

P(Λανθασμένης κατάταξης) = P(Λανθασμένη κατάταξη / Ανήκει στην 1^{η}) P(Ανήκει στην 1^{η}) + P(Λανθασμένη κατάταξη / ανήκει στη 2^{η}) P(Ανήκει στην 2^{η})

⇨ P(Λανθασμένης κατάταξης)$= P_1 . \Phi\left(\frac{k-\alpha/2}{\sqrt{\alpha}}\right) + P_2\left[1 - \Phi\left(\frac{k+\alpha/2}{\sqrt{\alpha}}\right)\right]$

Όπου:

$k = \ln c(1/2) \frac{c(1/2)}{c(2/1)} \frac{P_2}{P_1}$

$\alpha = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$

Σχόλιο

Όσο πιο μεγάλο είναι το α ,τόσο πιο πετυχημένος είναι ο διαχωρισμός. Όσο πιο κοντά είναι οι πληθυσμοί τόσο πιο δύσκολο να τους ξεχωρίσουμε.

ii. Ταξινόμηση κανονικών πληθυσμών όταν $\Sigma_1 \neq \Sigma_2$

Έστω ότι οι συναρτήσεις πυκνότητας πιθανότητας της τυχαίας μεταβλητής

$X'[x_1,x_2,\ldots,x_p]$ για τους πληθυσμούς $\Pi_1.\Pi_2$ δίνονται από τις σχέσεις:

$$f_1(x) = \frac{1}{2\pi^{P/2}|\Sigma_1|}\exp\left[-\frac{1}{2}(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)\right],$$

$$f_2(x) = \frac{1}{2\pi^{P/2}|\Sigma_2|}\exp\left[-\frac{1}{2}(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2)\right]$$

και $\mu_1,\mu_2,\Sigma_1,\Sigma_2$ με $\Sigma_1 \neq \Sigma_2$ γνωστά

Εάν $-\frac{1}{2}x'(\Sigma_1^{-1}-\Sigma_2^{-1})x + (\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})x - k \geq ln\left[\frac{c(1/2)}{c(2/1)}\cdot\left(\frac{P_2}{P_1}\right)\right]$

τότε κατέταξε την παρατήρηση x στον πληθυσμό Π_1 ,αλλιώς στον πληθυσμό Π_2 .

$$\text{όπου } k = \frac{1}{2}\ln\left(\left|\frac{\Sigma_1}{\Sigma_2}\right|\right) + \frac{1}{2}\left(\mu_1'\,\Sigma_1^{-1}\mu_1 - \mu_2'\,\Sigma_2^{-1}\mu_2\right)$$

Στην πράξη τα $\mu_1,\mu_2,\Sigma_1,\Sigma_2$ *αντικαθίστανται με τα* $\bar{x}_1, \bar{x}_2, S_1, S_2$ *οπότε παίρνουν τη μορφή* :

$$\text{Εάν } -\frac{1}{2}x'(S_1^{-1}-S_2^{-1})x + (\bar{x}_1'S_1^{-1}-\bar{x}_2'S_2^{-1})x - k \geq \ln\left[\frac{c(1/2)}{c(2/1)}\cdot\frac{P_2}{P_1}\right]$$

κατέταξε στον 1ο πληθυσμό.

Ο παραπάνω κανόνας λέγεται και τετραγωνικός κανόνας ταξινόμησης.

iii. Ταξινόμηση κανονικών πληθυσμών όταν μ_1, μ_2, Σ άγνωστα

Έστω ένα δείγμα με n_1 παρατηρήσεις για τον πληθυσμό Π_1 και ένα δείγμα n_2 με παρατηρήσεις για τον πληθυσμό Π_2 με:

$$\underset{pxn_1}{X_1}\left[x_{11}, x_{12}, \ldots, x_{1n_1}\right] \text{ για τον } \Pi_1$$

$$\underset{pxn_2}{X_2}\left[x_{21}, x_{22}, \ldots, x_{2n_2}\right] \text{ για τον } \Pi_2$$

$$\text{με } n_1 + n_2 - 2 \geq P$$

Οι εκτιμώμενες για τα μ_1, μ_2, Σ_1, Σ_2 θα είναι:

$$\overline{x_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j} \ , \ \overline{x_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j}$$

$$\underset{pxp}{S_1} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(x_{1j} - \overline{x_1})(x_{1j} - \overline{x_1})' \ , \ \underset{pxp}{S_2} = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(x_{2j} - \overline{x_2})(x_{2j} - \overline{x_2})'$$

Επειδή $\Sigma_1 = \Sigma_2 = \Sigma$, *οι δειγματικοί πίνακες* S_1, S_2 *συνδυάζονται με τον*

$$S_{pooled} = \frac{(n_1 - 1)\, S_1 + (n_2 - 1) S_2}{n_1 + n_2 - 2}$$

.

Άρα ο κανόνας θα γίνει:

$$Αν \ \left(\overline{x_1} - \overline{x_2}\right)' S_{pooled}^{-1} \cdot x - \frac{1}{2} \cdot \left(\overline{x_1} - \overline{x_2}\right)' S_{pooled}^{-1}\left(\overline{x_1} + \overline{x_2}\right) \geq \ln\left[\frac{c(1/2)}{c(2/1)} \cdot \left(\frac{P_2}{P_1}\right)\right]$$

τότε κατέταξε την παρατήρηση x στον πληθυσμό Π_1 αλλιώς στον Π_2.

Ειδική περίπτωση αν: $\frac{c(1/2)}{c(2/1)} \cdot \left(\frac{P_2}{P_1}\right) = 1$, *τότε ο λογάριθμος της ποσότητας αυτής είναι* 0,

οπότε ο κανόνας γίνεται:

Αν:

$$\left(\overline{x_1} - \overline{x_2}\right)' S_{pooled}^{-1} \cdot x \geq \frac{1}{2}\left(\overline{x_1} - \overline{x_2}\right)' S_{pooled}^{-1} \cdot \left(\overline{x_1} + \overline{x_2}\right)$$

τότε κατέταξε την παρατήρηση στον Π_1.

3.1.2.2 Διαχωριστική συνάρτηση Fisher

Ο Fisher πρότεινε τους γραμμικούς συνδυασμούς του X ώστε να δημιουργήσει τα Y. Η προσέγγιση του δεν υποθέτει ότι οι πληθυσμοί έχουν κανονικές κατανομές, υποθέτει όμως ότι έχουν ίσους πίνακες συνδιακύμανσης. Η μετατροπή των X σε Y γίνεται με μια διαχωριστική συνάρτηση , τα σκορ των 2 πληθυσμών θα πρέπει να είναι απομακρυσμένα ώστε να διευκολύνεται ο διαχωρισμός. Έστω ότι τα σκορ δίνονται ως :

$Y_{11}, Y_{12},\ldots,Y_{1n1}$ για τον Π_1 και $Y_{21}, Y_{22},\ldots,Y_{2n2}$ για τον Π_2

Ο Fisher πήρε ως μέτρο απόστασης των 2 ομάδων , την ποσότητα:

$$D = \frac{\bar{Y}_1 - \bar{Y}_2}{S_y}$$

όπου: $S_y^2 = \frac{\sum_{J=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{J=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$

Σκοπός είναι να μεγιστοποιήσουμε την D ή την απόσταση D^2 για να είναι τα σκορ των δύο ομάδων όσο γίνεται διαφορετικά. Έστω ο γραμμικός συνδυασμός $Y = L'X$ τότε πρέπει να μεγιστοποιήσουμε την ποσότητα :

$$D^2 = \frac{L'\left(\overline{X_1} - \overline{X_2}\right)^2}{L'\, S_{pooled} \cdot L} \text{ όπου } S_{pooled} = \frac{\sum_{i=1}^{k}\left(n_i - 1\right)\cdot S_i}{n-k}, n = n_1 + n_2, \; k = 2(\text{πλήθος ομάδων})$$

$$\text{Άρα για } L = c \cdot S_{pooled}^{-1}\left(\overline{X_1} - \overline{X_2}\right),\ c > 0 \text{ έχουμε } D^2 = \left(\overline{X_1} - \overline{X_2}\right)' \cdot S_{pooled}^{-1}\left(\overline{X_1} - \overline{X_2}\right)$$

τη μέγιστη απόσταση μεταξύ των μέσων και τον καλύτερο δυνατό διαχωρισμό. Το c είναι σταθερά και συνήθως παίρνουμε c=1. Η κρίσιμη τιμή είναι η ποσότητα :

$$m = \frac{\overline{Y_1} + \overline{Y_2}}{2} = \frac{L'\left(\overline{X_1} + \overline{X_2}\right)}{2}$$

Έτσι ο κανόνας είναι:

Μια παρατήρηση ανήκει στον πληθυσμό Π_1 αν:

$$Y_0 = \left(\overline{X_1} - \overline{X_2}\right) \cdot S_{pooled}^{-1} \cdot X_0 \geq m = \frac{1}{2}\left(\overline{X_1} - \overline{X_2}\right)' \cdot S_{pooled}^{-1} \cdot \left(\overline{X_1} + \overline{X_2}\right)$$

$$\Rightarrow Y_0 - m \geq 0 \text{ διαφορετικά ανήκει στον πληθυσμό } \Pi_2$$

Η ποσότητα $L'X - m$ είναι ίση με τη διαχωριστική συνάρτηση που βρήκαμε χρησιμοποιώντας τον κανόνα αποφάσεων υποθέτοντας κανονικές κατανομές.

3.1.3 Γενίκευση κανόνων διαχωριστικής ανάλυσης σε k-ομάδες

3.1.3.1 Γενίκευση της διαχωριστικής ανάλυσης του Fisher

Προτείνει τη χρήση k-1 γραμμικών συνδυασμών της μορφής $L_k'X$ με L_k να είναι τα ιδιοδιανύσματα του πίνακα $\Delta = (n-k)\cdot S_p^{-1}\cdot W$ με $L'S_PL = 1$ με σειρά που αντιστοιχεί στο μέγεθος των ιδιοτιμών.

- L_1 είναι το διάνυσμα που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή
- L_2 είναι το διάνυσμα που αντιστοιχεί στη δεύτερη μεγαλύτερη ιδιοτιμή , κλπ

και $W = \sum_{k=1}^{K} n_k(\bar{x}_k - \bar{x})\cdot(\bar{x}_k - \bar{x})'$

είναι ένα μέτρο της διακύμανσης των K ομάδων.

Η γεωμετρική ερμηνεία των παραπάνω διαχωριστικών συναρτήσεων είναι ότι :

- η 1η μεγιστοποιεί τις διαφορές των μέσων σε μια διάσταση
- η 2η μεγιστοποιεί την απόσταση των μέσων σε μια κατεύθυνση ορθογώνια στην πρώτη
- η 3η μας δείχνει την απόσταση σε μια τρίτη διάσταση ανεξάρτητη των άλλων δύο κτλπ.

Οι διαχωριστικές συναρτήσεις μπορούν να περιγραφούν σαν παράγοντες που διαχωρίζουν βέλτιστα τα κεντροειδή σε σχέση με τη διασπορά μέσα σε κάθε ομάδα.

Διαχωριστικός κανόνας με r-διαχωριστικές συναρτήσεις

Ταξινομούμε τη x παρατήρηση στην k ομάδα αν :

$$\sum_{m=1}^{r}\left[L_m\left(X-\overline{X_k}\right)\right]^2 \le \sum_{k=m}^{r}\left[L_m\left(X-\overline{X_i}\right)\right]^2 \gamma\iota\alpha\ i\neq k$$

3.1.3.2 Γενίκευση κανόνα πιθανοφάνειας

Έστω ότι έχουμε k πληθυσμούς $\Pi_1,\Pi_2,\ldots,\Pi_k$ k > 2 με γνωστές κατανομές. Αν η $f_i(x)$ πιθανοφάνεια του i πληθυσμού i=1,2...k τότε ορίζουμε ως R_i την περιοχή που κατατάσσουμε τον i πληθυσμό ως εξής:

$$R_i=\left\{x: f_i(x) > f_j(x)\ , j=1,2,..k\ \ \mu\varepsilon\ i\neq j\right\}$$

3.1.3.3 Γενίκευση κανόνα του Bayes

Αν συμβολίσουμε με P_j την πιθανότητα να πάρουμε μια παρατήρηση από τον j πληθυσμό τότε ο κανόνας του Bayes χρησιμοποιεί την εκ των υστέρων πιθανότητα η παρατήρηση να προέρχεται από τον πληθυσμό αυτό για να κατατάξει τις παρατηρήσεις . Η πιθανότητα αυτή είναι:

$$w_{ij}=\frac{P_j\cdot f_i(x_i)}{\sum_{j=k}^{k} P_j f_i(x_i)}$$

Οι περιοχές R_i που κατατάσσουμε τον i πληθυσμό είναι : $R_i=\{x: \rho_i\cdot f_i(x) > P_j\cdot f_j(x)$, j=1,2,…,k , $i\neq j$

3.1.3.4 Γενίκευση αναμενόμενου κόστους λανθασμένης κατάταξης

$$ECM = P_j \sum_{i=1}^{k} c(i/j) \cdot P(i/j)$$

όπου :

- c(i / j) το κόστος να κατατάξουμε την παρατήρηση x στην i ομάδα ενώ ανήκει στη j.
- P(j / i) η πιθανότητα να κατατάξουμε την παρατήρηση x στην j ομάδα ενώ ανήκει στη i
- P_j η εκ των προτέρων πιθανότητα να ανήκει μια παρατήρηση x στην j ομάδα

3.2 Ανάλυση κατά συστάδες (Cluster analysis)

Είναι μία μέθοδος που εφαρμόζεται με τέτοιο τρόπο ώστε να κατατάσσονται σε ίδιες συστάδες (ομάδες), παρατηρήσεις που είναι όσο περισσότερο όμοιες μεταξύ τους. Για να χαρακτηριστεί μία ανάλυση επιτυχημένη θα πρέπει:

1. Οι παρατηρήσεις μέσα σε κάθε συστάδα να είναι όσο γίνεται πιο ομοιογενείς, δηλαδή να έχουν κοινά χαρακτηριστικά και ιδιότητες.
2. Οι παρατηρήσεις διαφορετικών συστάδων να διαφέρουν όσο το δυνατό περισσότερο, δηλαδή οι τιμές μιας συστάδας θα πρέπει να διαφέρουν σε μέγεθος κλίμακας από τις τιμές άλλων συστάδων.

Βασικές έννοιες

Ένα χαρακτηριστικό της μεθόδου είναι ότι δεν απαιτεί καμία εκ των προτέρων υπόθεση για να ξεκινήσει η διαδικασία, έτσι λοιπόν δεν απαιτείται η εφαρμογή στατιστικών ελέγχων για τη σημαντικότητα των αποτελεσμάτων. Η ανάλυση κατά συστάδες πραγματοποιείται με τη χρήση αλγορίθμων με ίσως και εντελώς διαφορετικές ιδιότητες μεταξύ τους ως προς την απόδοσή τους και τον τρόπο λειτουργίας. Επεξεργάζονται συστάδες (ομάδες) οι οποίες εννοιολογικά σημαίνουν αποστάσεις μεταξύ των παρατηρήσεων, ειδικές κατανομές των παρατηρήσεων και λοιπά.

Δύο βασικές έννοιες για την ανάλυση κατά συστάδες, είναι οι έννοιες της απόστασης και τις ομοιότητας. Οι δύο αυτές έννοιες είναι αντίθετες, διότι οι παρατηρήσεις που είναι όμοιες θα έχουν μεγάλη τιμή στο μέτρο ομοιότητας και μικρή τιμή στο μέτρο απόστασης. Οι έννοιες αυτές χρησιμεύουν στο να βρούμε τις όποιες παρατηρήσεις και να τις τοποθετήσουμε στην ίδια συστάδα (ομάδα).

3.2.1 Αποστάσεις

Ορισμός

Σκοπός της απόστασης είναι να μετρήσει πόσο απέχουν δύο παρατηρήσεις, αν δηλαδή η τιμή της απόστασης που δίνεται είναι πολύ μικρή, οι παρατηρήσεις θα μοιάζουν πολύ μεταξύ τους. Άρα λοιπόν θα τοποθετηθούν στην ίδια συστάδα (ομάδα). Θέλουμε δηλαδή, οι αποστάσεις μέσα στην ίδια συστάδα να ελαχιστοποιούνται και ανάμεσα στις συστάδες να μεγιστοποιούνται.

Ευκλείδεια Απόσταση

Το πιο γνωστό και απλό μέτρο απόστασης ανάμεσα σε δύο συνεχή δεδομένα (παρατηρήσεις): $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ και $x_j = (x_{j1}, x_{j2}, \ldots, x_{ip})$ είναι η ευκλείδεια απόσταση. Ο τύπος της:

$$d_{ij}=d(x_i, x_j) = \sqrt{\sum_{r=\perp}^{P} (x_{ir} - x_{jr})^2}$$

και ικανοποιεί τις εξής τρεις ιδιότητες:

1. $d_{ij} \geq 0$ για κάθε i,j και $d_{ij} = 0 \Leftrightarrow i=j$
2. $d_{ij} \leq d_{is} + d_{sj}$ (Τριγωνική ανισότητα)
3. $d_{ij} = d_{ji}$ (Συμμετρική ιδιότητα)

Μειονεκτήματα

1. Η ευκλείδεια απόσταση εξαρτάται από την κλίμακα μέτρησης, οπότε αλλάζοντας την κλίμακα μπορούμε να πάρουμε τελείως διαφορετικά αποτελέσματα.
2. Ακόμη, οι μεταβλητές με μεγάλες απόλυτες τιμές συνεισφέρουν περισσότερο σε σχέση με τις μεταβλητές με μικρές απόλυτες τιμές. Αυτό έχει ως αποτέλεσμα η απόσταση ανάμεσα στις παρατηρήσεις να καθορίζεται σχεδόν μόνο από τις πρώτες.

Συμπερασματικά λοιπόν η ευκλείδεια απόσταση δεν μοιάζει καλό μέτρο απόστασης όταν η κλίμακα στις παρατηρήσεις είναι διαφορετική.

Απόσταση του Pearson

Παρατηρώντας λοιπόν το μειονέκτημα της ευκλείδειας απόστασης όσον αφορά στην κλίμακα, ένας τρόπος να φέρουμε όλες τις μεταβλητές σε συγκρίσιμη κλίμακα είναι να διαιρέσουμε κάθε μία μεταβλητή με την τυπική της απόκλιση. Έτσι αν συμβολίσουμε με S_r την τυπική απόκλιση της r μεταβλητής έχουμε:

$$S_r=\left[\frac{1}{n-1} \sum_{i=1}^{n} (x_{ir} - \bar{x}_r)^2\right]^2 \quad , \overline{x_r} = \frac{1}{n}\sum_{i=1}^{n} x_{in}$$

Παίρνουμε μία απόσταση της μορφής:

$$d_{ij} = d(x_i, x_j) = \sqrt{\sum_{r=1}^{P} \frac{(x_{ir} - x_{jr})^2}{S_r^2}} = \sqrt{\sum_{r=1}^{\rho} \left(\frac{x_{ir} - x_{jr}}{s_r}\right)^2}$$

Και είναι γνωστή ως απόσταση του Pearson. Συμπερασματικά, η παραπάνω απόσταση επιτρέπει πιο καλές συγκρίσεις ανάμεσα στις μεταβλητές.

Απόσταση του Mahalanobis

Οι προαναφερθείσες αποστάσεις έχουν ένα ακόμη μειονέκτημα. Δε λαμβάνουν υπόψιν το κατά πόσο δύο μεταβλητές είναι πολύ συσχετισμένες μεταξύ τους ή όχι. Αυτό έχει ως αποτέλεσμα αν η απόσταση των παρατηρήσεων οφείλεται στη μία από αυτές η άλλη να την ακολουθεί εξαιτίας της συσχέτισης. Μία απόσταση λοιπόν που να λαμβάνει υπόψιν τις συνδιακυμάνσεις είναι η απόσταση του Mahalanobis.
Ο τύπος της:

$$d_{ij}^2 = (x_i - x_j)' \Sigma^{-1} (x_i - x_j)$$

Όπου Σ είναι ο δειγματικός πίνακας διακύμανσης - συνδιακύμανσης πού αντιστοιχεί στα διανύσματα $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ και $x_j = (x_{j1}, x_{j2}, \ldots, x_{ip})$.

Απόσταση Manhattan ή City-Block Metric

Η μόνη διαφορά της απόστασης Manhattan με την ευκλείδεια απόσταση είναι ότι εδώ χρησιμοποιούνται απόλυτες αποκλίσεις. Τα αποτελέσματα που δίνονται είναι περίπου ίδια εκτός από την περίπτωση που υπάρχουν outliers, επειδή τους δίνει μικρότερο βάρος τα αποτελέσματα τείνουν να είναι ανθεκτικότερα.
Ο τύπος της:

$$d_{ij} = \sum_{r=1}^{P} |x_{ir} - x_{jr}|$$

Απόσταση Minkowski

Αυτή η απόσταση ουσιαστικά γενικεύει την απόσταση Manhattan και την ευκλείδεια απόσταση. Παρατηρώντας τον τύπο της:

$$d_{ij} = \sum_{r=1}^{P} \left[\left(|x_{ir} - x_{jr}| \right)^k \right]^{1/k}$$

όπου $k \geq 1$, βλέπουμε ότι για $k = 1$ προκύπτει η απόσταση Manhattan και για $k=2$ η ευκλείδεια απόσταση.

Σχόλια

Πέραν των αποστάσεων που περιγράφηκαν παραπάνω υπάρχουν και άλλα μέτρα απόστασης ανάμεσα σε συνεχή δεδομένα, ενδεικτικά αναφέρω ονομαστικά την απόσταση Max ή Chebyshev, απόσταση του Gower, απόσταση Bhattacharyya, απόσταση Canberra metric και λοιπά.

Σημαντικό είναι να αναφερθεί ότι όταν χρησιμοποιούμε διαφορετικό τύπο απόστασης, οι αποστάσεις μεταξύ των παρατηρήσεων δεν είναι ίδιες όμως συνήθως η σχετική διάταξη δεν μεταβάλλεται χωρίς βέβαια αυτό να αποκλείεται.

3.2.2 Μεταβλητές προς μελέτη

Για τον υπολογισμό της απόστασης υπάρχει διαφοροποίηση ανάλογα με το αν τα δεδομένα μας περιέχουν αριθμητικές, δυαδικές ή κατηγορικές τιμές.

3.2.2.1 Αποστάσεις για δίτιμες μεταβλητές (Binary Variables)

Τα δυαδικά δεδομένα είναι δεδομένα / μεταβλητές που μπορούν να πάρουν μόνο δύο πιθανές τιμές (καταστάσεις), 0 και 1 σύμφωνα με το δυαδικό σύστημα. Η τιμή της καταγραφής 1 δηλώνει την παρουσία του χαρακτηριστικού και το 0 την απουσία. Για να ορίσουμε ένα μέτρο απόστασης d_{ij} μεταξύ δύο ατόμων i και j , φτιάχνουμε ένα πίνακα συνάφειας 2 x 2.

		παρατήρηση j		
		1	0	
παρατήρηση i	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	p

Όπου a, b, c, d το πλήθος των συνδυασμών (1,1) (1,0) (0,1) (0,0) αντίστοιχα και p = a + b + c +d. Το (1,1) δηλώνει τα χαρακτηριστικά που είναι παρόντα και στις δύο παρατηρήσεις, το (1,0) τα χαρακτηριστικά που είναι παρόντα στην i - παρατήρηση και απόντα στην j, το (0,0) τα χαρακτηριστικά που είναι απόντα και στις δύο παρατηρήσεις και το (0,1) τα χαρακτηριστικά που είναι παρόντα στην j-παρατήρηση και απόντα στην i - παρατήρηση. Οι πιο συνηθισμένες αποστάσεις που χρησιμοποιούμε σε δυαδικά δεδομένα είναι οι εξής:

1. Simple matching distance με τύπο:

- $d_{ij} = \frac{b+c}{a+b+c+d}$

Αυτό το μέτρο απόστασης, μετράει το ποσοστό των περιπτώσεων για τις οποίες οι δύο παρατηρήσεις δεν συμφωνούν (κοινή παρουσία ή απουσία ενός χαρακτηριστικού).

2. Rogers and Tanimoto distance με τύπο:

- $d_{ij} = \frac{2(b+c)}{(a+d)+2(b+c)}$

Αυτό το μέτρο απόστασης δίνει διπλάσιο βάρος στις ασυμφωνίες.

3. Sokal and Sneath distance με τύπο:

- $d_{ij} = \frac{b+c}{2(a+d)+2(b+c)}$

Το μέτρο απόστασης αυτό δίνει διπλάσιο βάρος στις συμφωνίες.

4. Jaccard distance με τύπο:

- $d_{ij} = \frac{b+c}{a+b+c}$

Το μέτρο απόστασης αυτό δε λαμβάνει καθόλου υπόψιν του την απουσία και των δύο χαρακτηριστικών

5. Dice and Sorensen distance με τύπο:

- $d_{ij} = \frac{b+c}{2a+b+c}$

Το μέτρο απόστασης αυτό δίνει διπλάσιο βάρος στην παρουσία και των δύο χαρακτηριστικών.

Πίνακας Αποστάσεων

Όταν έχουμε n - παρατηρήσεις (άτομα) , τις αποστάσεις τους $d_{ij}=d(x_i, x_j)$, i,j=1,2,…,n, τις τοποθετούμε σε ένα πίνακα $D=[d_{ij}]$ με n-γραμμές και n-στήλες ο οποίος ονομάζεται πίνακας αποστάσεων. Λόγω της πρώτης ιδιότητας όλα τα διαγώνια στοιχεία του πίνακα θα είναι μηδέν και από την τρίτη ιδιότητα προκύπτει ότι ο πίνακας θα είναι συμμετρικός.

3.2.2.2 Κατηγορικές μη διατάξιμες

Στην περίπτωση αυτή είναι δύσκολο να υπολογίσουμε την απόσταση. Μπορούμε να κατασκευάσουμε ψευδομεταβλητές, μία για κάθε επίπεδο κάθε μεταβλητής και ύστερα να υπολογίσουμε την αντίστοιχη απόσταση για τα δίτιμα δεδομένα που θα προκύψουν. Συνήθως η απόσταση που χρησιμοποιείται είναι Simple matching distance :

$d_{ij} = \frac{p-u}{p}$, όπου u είναι ο αριθμός των μεταβλητών που έχουν την ίδια τιμή και p ο συνολικός αριθμός των μεταβλητών.

3.2.2.3 Κατηγορικές διατάξιμες μεταβλητές

Στην περίπτωση αυτή συνήθως θεωρούμε τις μεταβλητές ως συνεχείς και χρησιμοποιούμε μία κατάλληλη απόσταση. Σε τέτοιες περιπτώσεις φροντίζουμε να χρησιμοποιείται η ίδια κλίμακα σε όλες τις ερωτήσεις. Εναλλακτικά για να μην υπάρχει πρόβλημα μπορούμε να μετασχηματίσουμε την κλίμακα για να παίρνει τιμές για παράδειγμα στο διάστημα (0,1).

3.2.3 Μέτρα ομοιότητας

Τα μέτρα ομοιότητας μπορεί να χρησιμοποιηθούν για να μας δείξουν αν δύο παρατηρήσεις είναι όμοιες ή ανόμοιες μεταξύ τους. Δηλαδή θέλουμε μεγάλη τιμή στο μετρό ομοιότητας για τις παρατηρήσεις που μοιάζουν πολύ, ενώ για τις ανόμοιες παρατηρήσεις θέλουμε μικρή τιμή.

Ας υποθέσουμε ότι για κάθε ζεύγος παρατηρήσεων $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ και $x_j = (x_{j1}, x_{j2}, \ldots, x_{ip})$ ορίζεται ένας πραγματικός αριθμός $S_{ij} = S(x_i, x_j)$ που είναι η ομοιότητα ανάμεσα στην i και j παρατήρηση, έτσι ώστε να ισχύουν οι παρακάτω τρεις ιδιότητες:

1. $S_{ij} \geq 0$ για κάθε i,j και i=j → $S_{ij}=1$
2. $S_{ij} \leq 1$
3. $S_{ij} = S_{ji}$ (συμμετρική ιδιότητα)

Τότε θα λέμε ότι η συνάρτηση $S_{ij} = S(x_i, x_j)$ δίνει ένα μέτρο ομοιότητας. Το πιο γνωστό μετρό ομοιότητας για ποσοτικές παρατηρήσεις είναι ο δειγματικός συντελεστής συσχέτισης. Ο τύπος:

$$S_{ij} = \frac{\sum_{r=1}^{\rho} (x_{ir} - x_i \cdot)(x_{jr} - \bar{x}_j \cdot)}{\left(\sum_{r=1}^{p} (x_{ir} - \bar{x}_i \cdot)^2 \sum_{r=1}^{p} (x_{j_r} - \bar{x}_j \cdot)^2 \right)^{1/2}}$$

όπου:

$$\bar{x}_i \cdot = \frac{1}{p} \sum_{r=1}^{\rho} x_{ir} \quad , \quad \bar{x}_j \cdot = \frac{1}{p} \sum_{r=1}^{\rho} x_{jr}$$

1.5.3.1 Δίτιμες Μεταβλητές (Binary Variables)

Για να ορίσουμε ένα μέτρο ομοιότητας για δυαδικά δεδομένα, δηλαδή μεταβλητών που μπορούν να πάρουν μόνο μία από δύο τιμές έστω 0 και 1, δημιουργούμε πάλι ένα πίνακα συνάφειας 2 x 2 όπου a, b, c, d είναι το πλήθος των συνδυασμών (1,1) (1,0) (0,1) (0,0) αντίστοιχα και p = a + b + c + d.

Τα πιο συνηθισμένα μέτρα ομοιότητας που χρησιμοποιούμε για την ομαδοποίηση παρατηρήσεων που παρατηρούνται δυαδικά δεδομένα είναι τα εξής:

1. Simple matching με τύπο: $S_{ij} = \frac{a+b}{a+b+c+d}$, χρησιμοποιεί ίσα βάρη για συμφωνίες

2.Rogers and Tanimoto με τύπο: $S_{ij} = \frac{a+d}{(a+d)+2(b+c)}$, χρησιμοποιεί διπλάσιο βάρος για τις ασυμφωνίες.

3.Sokal and Sneath με τύπο: $S_{ij} = \frac{2(a+d)}{2(a+d)+2(b+c)}$, χρησιμοποιεί διπλάσιο βάρος για τις συμφωνίες.

4.Jaccard coefficient με τύπο: $S_{ij} = \frac{a}{a+b+c}$, δε χρησιμοποιεί καθόλου ,ούτε σε παρανομαστή ούτε σε αριθμητή τις συμφωνίες (0,0).

5.Russel and Rao με τύπο: $S_{ij} = \frac{a}{a+b+c+d}$, δε χρησιμοποιεί στον αριθμητή τις συμφωνίες (0,0).

6.Dice and Sorensen με τύπο: $S_{ij} = \frac{2a}{2a+b+c}$, χρησιμοποιεί διπλάσιο βάρος στις συμφωνίες (1,1) και δε χρησιμοποιεί καθόλου ,ούτε σε παρανομαστή ούτε σε αριθμητή τις συμφωνίες (0,0).

7.Kulczynski με τύπο: $S_{ij} = \frac{a}{b+c}$,δε χρησιμοποιεί καθόλου τις συμφωνίες (0,0).

Συμπερασματικά λοιπόν υπάρχουν πολλά είδη ομοιότητας / αποστάσεων που μας δίνουν τη δυνατότητα να επιλέγουμε κάθε φορά τον κατάλληλο ανάλογα με τον τύπο δεδομένων που εξετάζουμε για την ανάλυση μας. Η επιλογή του μέτρου απόστασης / ομοιότητας είναι καθοριστικό βήμα στην ανάλυση συστάδων, αφού επηρεάζει το σχηματισμό ομάδων.

3.2.4 Μέθοδοι Ομαδοποίησης

Οι μέθοδοι ομαδοποίησης μπορούν να χωριστούν σε δύο διαφορετικές κατηγορίες ανάλογα με τον τρόπο που προχωρούν στη διαμόρφωση των ομάδων. Χωρίζονται στις **ιεραρχικές** (hierarchical methods), ίσως η πιο δημοφιλής κατηγορία μεθόδων ανάλυσης συστάδων και η δεύτερη κατηγορία είναι η **μη ιεραρχικές**.

3.2.4.1 Ιεραρχικές Μέθοδοι

Στις ιεραρχικές μεθόδους οι ομάδες προκύπτουν από συγχώνευση μικρότερων ομάδων μέχρι να φτάσουμε να έχουμε όλα τα δεδομένα σε μία ομάδα και ονομάζονται συσσωρευτικές μέθοδοι, είτε με διαίρεση ομάδων σε μικρότερες μέχρι να φτάσουμε κάθε παρατήρηση να είναι μία ομάδα , αυτές ονομάζονται διαιρετικές μέθοδοι. Αυτές οι μέθοδοι χρησιμοποιούν σε κάθε βήμα ένα πίνακα αποστάσεων (δηλαδή τις αποστάσεις όλων των παρατηρήσεων από τις υπόλοιπες) και έτσι απαιτούν πολύ χρόνο και χώρο στον υπολογιστή. Συμπερασματικά λοιπόν καλύτερα να μη χρησιμοποιούνται για μεγάλο πλήθος δεδομένων.

Συσσωρευτικές Μέθοδοι

Στο πρώτο βήμα ενός συσσωρευτικού αλγορίθμου θεωρούμε ότι η κάθε παρατήρηση αποτελεί μία ομάδα και στη συνέχεια βρίσκουμε τις δύο πλησιέστερες παρατηρήσεις εντοπίζοντας στον πίνακα D (αποστάσεων) πού εμφανίζεται η μικρότερη απόσταση και τα συγχωνεύουμε για να αποτελέσουν μία ομάδα. Έτσι καταλήγουμε σε μία ομαδοποίηση που αποτελείται από n-1 ομάδες, μία ομάδα με δύο στοιχεία και n-2 ομάδες του ενός στοιχείου. Συνεχίζεται ο αλγόριθμος έως ότου όλα τα άτομα θα βρεθούν σε μία και μοναδική ομάδα αποτελούμενη από n- στοιχεία. Παρακάτω παρουσιάζεται ο αλγόριθμος με μορφή βημάτων.

Βήματα

Βήμα 1

Ξεκινώντας θεωρούμε καθεμία από τις παρατηρήσεις σα μία ξεχωριστή ομάδα και υπολογίζουμε τον πίνακα αποστάσεων (εναλλακτικά ομοιότητας) μεταξύ τους (για όλες τις ομάδες).

Βήμα 2

Εντοπίζουμε στον πίνακα τη μικρότερη δυνατή απόσταση (ή εναλλακτικά τη μεγαλύτερη τιμή ομοιότητας).

Βήμα 3

Συνενώνουμε τις παρατηρήσεις με τη μικρότερη απόσταση (ή μεγαλύτερη τιμιότητας) μειώνοντας έτσι τον αριθμό των ομάδων κατά 1.Υπολογίζουμε ξανά τον πίνακα με τις ομάδες που έχουν προκύψει και βρίσκουμε ξανά τη μικρότερη απόσταση και ενώνουμε τις δύο ομάδες που αντιστοιχούν σ' αυτή τη απόσταση.

Βήμα 4

Αν δεν έχουν μπει όλες οι παρατηρήσεις σε μία ομάδα επανέλαβε τα βήματα 2 και 3 αλλιώς σταμάτα.

Κριτήρια συνένωσης

Αυτά τα κριτήρια τα χρειαζόμαστε διότι αν και έχουμε ορίσει μέτρα απόστασης μεταξύ στοιχείων δεν έχουμε ορίσει μέτρα απόστασης μεταξύ ομάδων. Υπάρχουν πολλές μέθοδοι μερικές παρουσιάζονται παρακάτω:

1. Μέθοδος της απλής συνένωσης (Single Linkage Method)

Η μέθοδος αυτή γνωστή και ως μέθοδος του πλησιέστερου ή κοντινότερου γείτονα υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως τη μικρότερη απόσταση από μία παρατήρηση μέσα στη μία ομάδα με κάποια παρατήρηση στην άλλη ομάδα. Παρόλο που έχει κάποιες χρήσιμες μαθηματικές ιδιότητες, συνήθως δημιουργεί ομάδες που δεν είναι συμπαγείς, μερικές πολύ μεγάλες ομάδες και κάποιες άλλες πολύ μικρές. Προκειμένου να ενώσουμε δύο ομάδες υπολογίζουμε την ελάχιστη απόσταση $\min d_{ij}$.

2.Μέθοδος της πλήρους συνένωσης (Complete Linkage Method)

Η μέθοδος αυτή γνωστή και ως μέθοδος του μακρινότερου γείτονα υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως τη μεγαλύτερη απόσταση από μία παρατήρηση μέσα σε μία ομάδα με κάποια παρατήρηση στην άλλη ομάδα $\max d_{ij}$. Οι ομάδες που δημιουργούνται με αυτή τη μέθοδο είναι συνήθως συμπαγείς και μεγάλες, όμως η μέθοδος αρκετά συχνά αποτυγχάνει να ξεχωρίσει κάποιες πολύ συμπαγείς μικρές ομάδες.

3. Average Between Groups

Στην περίπτωση αυτή η απόσταση είναι ο μέσος όλων των αποστάσεων που προκύπτουν όταν ενώσουμε τις δύο ομάδες.

4. Μέθοδος των κέντρων βάρους

Σε αυτή τη μέθοδο η απόσταση των δύο ομάδων είναι ίση με την απόσταση του κέντρου βάρους των ομάδων. Οι ομάδες που δημιουργούνται είναι συμπαγείς και ελλειπτικές. Ένα μειονέκτημα της μεθόδου είναι ότι εφαρμόζεται μόνο σε ποσοτικά δεδομένα λόγω της χρήσης της ευκλείδειας απόστασης.

5.Μέθοδος του Ward

Η μέθοδος αυτή διαφέρει από τις υπόλοιπες διότι δεν υπολογίζει απόσταση ανάμεσα στις ομάδες αλλά είναι σχεδιασμένη να ελαχιστοποιεί τη διακύμανση μέσα στις ομάδες. Ένα μέτρο ομοιογένειας που χρησιμοποιείται είναι το άθροισμα των τετραγώνων των σφαλμάτων. Κι αυτή η μέθοδος εφαρμόζεται μόνο σε ποσοτικά δεδομένα, δημιουργεί ομάδες με παρόμοιο αριθμό παρατηρήσεων και συνήθως δίνει τα καλύτερα αποτελέσματα για αυτό και πολύ συχνά χρησιμοποιείται στην πράξη.

Συμπληρωματικά αναφέρουμε κι άλλες δύο μεθόδους τη μέθοδο του Gower (Gower's method) και τη μέθοδο ευέλικτης στρατηγικής (flexible strategy method).

Σχόλιο

Για όλες τις μεθόδους χρειαζόμαστε έναν πίνακα αποστάσεων με βάση τον οποίο υπολογίζουμε σε κάθε βήμα τις καινούργιες αποστάσεις. Αυτό δεν ισχύει για την Centroid method όπου χρειαζόμαστε το κέντρο της ομάδας. Από πειράματα προσομοίωσης για σύγκριση των μεθόδων, οι μέθοδοι με την καλύτερη επίδοση είναι η μέθοδος του Ward και η Average Between groups και η μέθοδος με τη χειρότερη επίδοση είναι αυτή του κοντινότερου γείτονα. Παρόλα αυτά σε πολλά προβλήματα δεν είναι ξεκάθαρο ποια μέθοδος είναι προτιμότερη, θα πρέπει πάντα ο ερευνητής να μπορεί να κρίνει διότι ανάλογα με τη μορφή των δεδομένων κάθε μέθοδος δουλεύει καλύτερα.

Διαιρετικές Μέθοδοι

Αυτή η κατηγορία ιεραρχικών μεθόδων εκτελεί ακριβώς την αντίστροφη διαδικασία από τις συσσωρευτικές. Θεωρούμε όλες τις παρατηρήσεις σαν μέλη μιας ενιαίας ομάδας. Στη συνέχεια η αρχική αυτή ομάδα διαιρείται σε δύο υποομάδες οι οποίες θα έχουν τη μεγαλύτερη ανομοιότητα. Η διαδικασία αυτή διαδοχικού διαχωρισμού των ομάδων επαναλαμβάνετε έως ότου φτάσουν στο σημείο όλες οι ομάδες να περιέχουν μία μόνο παρατήρηση.

Κριτήρια επιλογής πλήθους ομάδων

Δενδροδιάγραμμα

Το δενδροδιάγραμμα αρχίζει με τόσες ομάδες όσες είναι οι παρατηρήσεις και τελειώνει σε μία ομάδα η οποία περιλαμβάνει όλες οι παρατηρήσεις, διότι προκύπτει από μία ιεραρχική συσσωρευτική μέθοδο. Είναι ένας πολύ απλός και πρακτικός τρόπος καθορισμού του βέλτιστου πλήθους των ομάδων.

Σε αυτό οι κάθετες γραμμές δηλώνουν συνδυασμούς ομάδων παρατηρήσεων, ενώ στον οριζόντιο άξονα καταγράφεται η ποσότητα (απόσταση ή μέτρο ομοιότητας) κατά τα οποία οι ομάδες συνδυάζονται. Ουσιαστικά, κάθε επίπεδο ενός δενδρογράμματος ορίζει ένα βήμα του αλγορίθμου. Στο σημείο εκείνο του δενδροδιαγράμματος που παρατηρείται η μεγαλύτερη μεταβολή της ποσότητας μπορούμε να φέρουμε μία παράλληλη γραμμή προς τον κάθετο άξονα και να δούμε σε πόσα σημεία τέμνει δενδροδιάγραμμα. Το πλήθος k, για το οποίο παρατηρούμε μεγάλες αποστάσεις συνένωσης σε σχέση με το προηγούμενο (k-1 ομάδες) αποτελεί μία λογική τιμή για το βέλτιστο πλήθος των ομάδων.

Όμως τα αποτελέσματα που προκύπτουν από μία τέτοια διαδικασία υπόκεινται στην κρίση του ερευνητή, με αποτέλεσμα να μην αντιστοιχεί πάντα στο σωστό τρόπο για την εύρεση του βέλτιστου πλήθους των ομάδων.

Μειονεκτήματα ιεραρχικών μεθόδων

Όταν το πλήθος των δεδομένων που μελετάμε είναι μεγάλο η ανάλυση του με ιεραρχικές μεθόδους είναι συνήθως ασύμφορες. Αυτό συμβαίνει διότι πρέπει κανείς να σχηματίσει και να αποθηκεύσει στη μνήμη του υπολογιστή ολόκληρο τον πίνακα αποστάσεων των παρατηρήσεων ακόμα και η απόσταση είναι συμμετρική. Δηλαδή χρειάζεται να αποθηκευτούν n(n-1)/2 αποστάσεις και πρέπει να ανανεώνεται σε κάθε βήμα. Αυτό έχει ως αποτέλεσμα να οδηγεί την ανάλυση σε χρονοβόρες υπολογιστικές διαδικασίες. Επίσης οι αρχικές ομάδες που δημιουργούνται δεν μπορούν να χωρίσουν με αποτέλεσμα οι παρατηρήσεις που ενώνονται σε αρχικά βήματα μένουν για πάντα μαζί.

Πολύ συχνά καταλήγει στη δημιουργία ενός μικρού πλήθους ομάδων με πολλές παρατηρήσεις και αφήνει πολλές παρατηρήσεις να είναι από μόνες τους ανεξάρτητες ομάδες. Οι μέθοδοι είναι ευαίσθητες στην ύπαρξη ακραίων τιμών (outliers). Από την άλλη πλευρά οι διαιρετικές μέθοδοι δεν είναι αρκετά διαδεδομένες στην πράξη διότι απαιτούν πολύ περισσότερους υπολογισμούς από τις συσσωρευτικές μεθόδους. Στο πρώτο στάδιο ενός διαιρετικού αλγορίθμου οι πιθανοί διαμερισμοί του συνόλου n των παρατηρήσεων σε δύο μη κενά σύνολα είναι $(2^n - 2) = 2^{n-1} - 1$, έχουμε δηλαδή εκθετική αύξηση του πλήθους.

3.2.4.2 Μη Ιεραρχικές Μέθοδοι

Μία άλλη κατηγορία μεθόδων είναι οι λεγόμενες μη ιεραρχικές μέθοδοι. Εδώ θεωρείται ότι ο αριθμός των ομάδων είναι από πριν γνωστός. Ο στόχος αυτών των μεθόδων είναι να ομαδοποιήσουν το n-πλήθος παρατηρήσεων που έχουμε σε k ομάδες, όπου το k είναι καθορισμένο από την αρχή από τον ερευνητή. Χρησιμοποιούμε έναν επαναληπτικό αλγόριθμο για να τοποθετήσουμε τις παρατηρήσεις στις ομάδες ανάλογα με το ποια ομάδα είναι πιο κοντά στην εκάστοτε παρατήρηση.

Ουσιαστικά ο τρόπος λειτουργίας ενός τέτοιου αλγόριθμου είναι είτε να θεωρεί k συγκεκριμένες παρατηρήσεις (μητρικά σημεία) και γύρω από αυτά να ταξινομούν τις υπόλοιπες παρατηρήσεις έως ότου διαμορφωθούν οι επιθυμητές ομάδες είτε να ξεκινούν με ένα αρχικό διαμερισμό των παρατηρήσεων σε k-ομάδες και στη συνέχεια να μετακινούν τα άτομα μεταξύ των ομάδων μέχρι να πετύχει ο καλύτερος διαμερισμός. Αυτές οι μέθοδοι ενώ δουλεύουν ικανοποιητικά με μεγάλα δείγματα επηρεάζονται αρκετά από τις αρχικές τιμές που θα χρησιμοποιήσουμε.

Για τον τρόπο δημιουργίας μητρικών σημείων υπάρχουν διάφορες μέθοδοι. Ενδεικτικά αναφέρουμε τρεις:

1. Αριθμούμε τις παρατηρήσεις από το 1 έως το n, δημιουργούμε k διαφορετικούς τυχαίους αριθμούς από το 1 έως το n και επιλέγουμε τις παρατηρήσεις που αντιστοιχούν σε αυτούς τους αριθμούς.
2. Επιλέγουμε τις πρώτες k στη σειρά παρατηρήσεις.

3. Διαχωρίζουμε με κάποιο υποκειμενικό συνήθως τρόπο τις παρατηρήσεις σε k- ομάδες και θεωρούμε τα κέντρα βάρους των ομάδων μητρικά σημεία.

3.2.4.2.1 Αλγόριθμοι ομαδοποίησης μη ιεραρχικών μεθόδων

Από τις πιο γνωστές μεθόδους ανάλυσης συστάδων μη ιεραρχικών μεθόδων είναι αυτή που προτάθηκε από τον Forgy και ο αλγόριθμος k-means (μέθοδος Mac Queen).

K-means

Περιγραφή Αλγορίθμου

Βήμα 1

Καθορισμός από τον ερευνητή ενός αρχικού συνόλου από k μητρικά σημεία χρησιμοποιώντας k από τις n παρατηρήσεις που είναι διαθέσιμες.

Βήμα 2

Κατάταξε σε καθεμία από τις εναπομείναντες n-k παρατηρήσεις στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την παρατήρηση.

Βήμα 3

Μετά από κάθε τοποθέτηση παρατήρησης υπολόγισε ξανά τα μητρικά σημεία (κέντρα) της καινούργιας πλέον ομάδας.

Βήμα 4

Αν τα μητρικά σημεία δεν είναι διαφορετικά από τα παλιά σταμάτα αλλιώς πήγαινε στο Βήμα 2.

Πλεονεκτήματα

Ο αλγόριθμος του Mac Queen στην πράξη τερματίζει συνήθως μετά από σχετικά λίγες επαναλήψεις, με αποτέλεσμα να επιλέγεται για ομαδοποίηση στις περιπτώσεις σε μεγάλα σύνολα δεδομένων. Επιπλέον δεν χρειάζεται να κρατά στη μνήμη πολλά στοιχεία άρα δεν απαιτεί μεγάλη υπολογιστική ισχύ και χωρητικότητα. Συνήθως η τελική ομαδοποίηση που δημιουργείται περιέχει ομάδες με ίσο περίπου αριθμό παρατηρήσεων. Είναι απλός και εύκολος στην εφαρμογή του και περιλαμβάνεται στα περισσότερα στατιστικά πακέτα.

Μειονεκτήματα

Εφαρμόζεται μόνο σε ποσοτικές μεταβλητές και εξαρτάται από τα μητρικά σημεία και τις αρχικές διαμερίσεις τα οποία αν δεν είναι σωστά επιλεγμένα μπορεί να οδηγήσουν σε εντελώς διαφορετική ομαδοποίηση από τη φυσική. Ένα ακόμη μειονέκτημα είναι ότι η ύπαρξη έκτροπων παρατηρήσεων μπορεί να οδηγήσει στη δημιουργία ομάδων με πολύ διεσπαρμένα άτομα. Ακόμη είναι ευαίσθητος στην κλίμακα μέτρησης των δεδομένων, για παράδειγμα η κανονικοποίηση των δεδομένων μπορεί να αλλάξει ριζικά την ομαδοποίηση.

Συμπέρασμα

Σκοπός της ανάλυσης συστάδων είναι να ομαδοποιεί παρατηρήσεις σε ομάδες έτσι ώστε κάθε ομάδα να είναι όσο το δυνατόν ομοιογενής σε σχέση με τις μεταβλητές ομαδοποίησης. Το πρώτο βήμα στην ανάλυση συστάδων είναι να επιλέξουμε ένα μέτρο ομοιότητας ή απόστασης με το οποίο μετράμε τη σχέση μεταξύ των αντικειμένων. Στη συνέχεια εξετάζουμε το είδος της τεχνικής ομαδοποίησης που θα χρησιμοποιήσουμε ιεραρχική ή μη ιεραρχική. Έπειτα επιλέγουμε το είδος της μεθόδου ομαδοποίησης για την τεχνική που επιλέξαμε και τέλος γίνεται μία συζήτηση όσον αφορά τον αριθμό των ομάδων και ερμηνεύεται η λύση τους.

3.3 Δέντρα Αποφάσεων

Τα δέντρα απόφασης (decision trees) είναι από τα πιο γνωστά μοντέλα κατηγοριοποίησης και αποτελούν μια μέθοδο που βοηθάει στην επιλογή της βέλτιστης απόφασης. Τα δέντρα απόφασης είναι διαμορφωμένες δομές που αντιπροσωπεύουν σύνολα αποφάσεων. Αυτές οι αποφάσεις παράγουν τους κανόνες για την ταξινόμηση ενός συνόλου δεδομένων. Η κεντρική ιδέα στην οποία βασίζονται είναι ότι η επίλυση των προβλημάτων περνάει από τη διαδοχική λήψη αποφάσεων μέχρι να δημιουργηθούν οι κατάλληλες προϋποθέσεις για την επίλυση. Έτσι με δεντροειδή μορφή παρουσιάζονται οι διαδοχικές αποφάσεις.Τα δέντρα αποφάσεων έχουν διάφορα πλεονεκτήματα, όπως το ότι είναι εύκολο να τα καταλάβουμε, μπορούν να μετασχηματιστούν σε κανόνες και πειραματικά έχει αποδειχθεί ότι λειτουργούν πολύ καλά.

Για να χτίσουμε ένα δέντρο απόφασης, πρέπει αρχικά να επιλέξουμε ένα υποσύνολο περιπτώσεων από το σύνολο των δεδομένων που θα χρησιμοποιηθούν στην εκπαίδευση (training set). Αυτό το υποσύνολο (training set) χρησιμοποιείται έπειτα από τον αλγόριθμο για να κατασκευάσει το δέντρο απόφασης. Τα υπόλοιπα δεδομένα, τα δεδομένα (testing set), χρησιμοποιούνται στην εξέταση της ακρίβειας του κατασκευασμένου δέντρου. Εάν το δέντρο απόφασης ταξινομεί τις περιπτώσεις σωστά, η διαδικασία ολοκληρώνεται. Το διάγραμμα δένδρου είναι μια χρονολογική απεικόνιση όλων των πιθανών ακολουθιών ενεργειών και γεγονότων που οδηγούν στο τελικό αποτέλεσμα. Η γραφική απεικόνιση των δέντρων αποφάσεων αποτελείται από κλαδιά και κόμβους.

Ένα δέντρο απόφασης αποτελείται:

- έναν αρχικό κόμβο, τη ρίζα
- τους εσωτερικούς κόμβους και
- τους εξωτερικούς κόμβους, τα φύλλα.

Σε κάθε εσωτερικό κόμβο αντιστοιχεί ένα χαρακτηριστικό που χρησιμοποιείται για περαιτέρω διαχωρισμό του δέντρου. Στις ακμές που εξέρχονται από τη ρίζα ή κάθε εσωτερικό κόμβο, αντιστοιχεί μια συνθήκη ελέγχου με βάση το διαχωριστικό χαρακτηριστικό, ενώ τα κλαδιά που προέρχονται από έναν κόμβο απόφασης αντανακλούν όλες τις εναλλακτικές αποφάσεις που μπορεί να ληφθούν.

Βήματα

Αρχικά, επιλέγουμε ένα χαρακτηριστικό, το οποίο αναφέρεται στη ρίζα του δέντρου, και, στη συνέχεια, κατασκευάζουμε μια ακμή και έναν κόμβο για καθεμία από τις διακριτές τιμές του χαρακτηριστικού. Αυτά τα δύο βήματα επαναλαμβάνονται συνεχώς, μέχρις ότου όλα τα χαρακτηριστικά να εισαχθούν στους κόμβους του δέντρου.

3.3.1 Μέθοδοι κατασκευής δέντρων αποφάσεων

- C4.5 είναι ο πιο πρόσφατος αλγόριθμος κατασκευής δέντρων αποφάσεων του Quinlan. Είναι μια επέκταση του προηγούμενου αλγορίθμου ID3 Quinlan. Παρέχει καλύτερη διαχείριση ελλειπουσών τιμών. Επίσης γίνεται χρήση βελτιωμένης συνάρτησης για την αποφυγή overfitting. Τέλος, διαχειρίζεται τόσο διακριτά όσο και με συνεχή Δεδομένα.
- CART (Breiman) ήταν το πρώτο σύστημα που εισήγαγε τα δέντρα παλινδρόμησης (regression trees). Ουσιαστικά, τα δέντρα παλινδρόμησης παίρνουν τη μορφή των δέντρων αποφάσεων, όπου οι κόμβοι είναι αριθμητικές αντί για κατηγοριακές τιμές. Είναι ένα δέντρο που δημιουργείται διαχωρίζοντας έναν κόμβο σε δύο θυγατρικούς κόμβους επανειλημμένα, ξεκινώντας με τον ριζικό κόμβο που περιέχει ολόκληρο το δείγμα εκμάθησης.
- C5.0 προσφέρει μια σειρά βελτιώσεων στο C4.5. Μερικά από αυτά είναι:
  Ταχύτητα, Χρήση μνήμης , Μικρότερα δέντρα απόφασης
- Random forests Ένα μεγάλο πλεονέκτημα του αλγόριθμου Random Forest είναι ότι μπορεί να χρησιμοποιηθεί τόσο για προβλήματα ταξινόμησης όσο και για παλινδρόμησης. Συνδυάζει την ιδέα του “bagging” και την τυχαία επιλογή χαρακτηριστικών, προκειμένου να κατασκευαστεί μια συλλογή δέντρων αποφάσεων με ελεγχόμενη διακύμανση. Προσθέτει επιπλέον τυχαιότητα στο μοντέλο, ενώ καλλιεργεί τα δέντρα. Αντί να αναζητά το πιο σημαντικό χαρακτηριστικό ενώ διαχωρίζει έναν κόμβο, αναζητά το καλύτερο χαρακτηριστικό σε ένα τυχαίο υποσύνολο χαρακτηριστικών. Αυτό οδηγεί σε μια μεγάλη ποικιλία που γενικά οδηγεί σε ένα καλύτερο μοντέλο. Μόνο ένα τυχαίο υποσύνολο των χαρακτηριστικών λαμβάνεται υπόψη από τον αλγόριθμο για τον διαχωρισμό ενός κόμβου.

3.4 Νευρωνικά Δίκτυα (Neural Networks)

Ένα νευρικό δίκτυο είναι ένα απλοποιημένο μοντέλο του τρόπου με τον οποίο ο ανθρώπινος εγκέφαλος επεξεργάζεται πληροφορίες, ουσιαστικά αποτελεί μια προσπάθεια προσέγγισης της λειτουργίας του ανθρώπινου εγκεφάλου. Στόχο του είναι να μαθαίνουν να αναγνωρίζουν μαθηματικά πρότυπα σε συγκεκριμένα δεδομένα. Είναι ένα δίκτυο που αποτελείται από απλούς υπολογιστικούς κόμβους (νευρώνες) και υπάρχουν συνήθως τρία μέρη σε ένα νευρωνικό δίκτυο:

- ένα επίπεδο εισόδου, με μονάδες που αντιπροσωπεύουν τις ανεξάρτητες μεταβλητές,
- ένα ή περισσότερα κρυφά στρώματα (υπολογιστικοί νευρώνες)
- και ένα επίπεδο εξόδου, με μια μονάδα ή μονάδες που αντιπροσωπεύουν τα πεδία-στόχους (εξαρτημένη μεταβλητή).

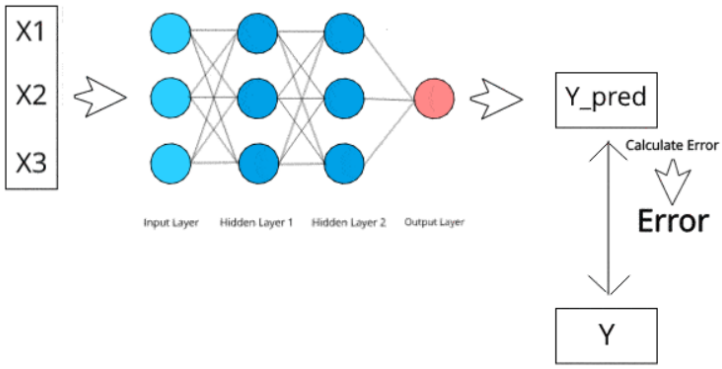
Οι μονάδες συνδέονται με κάποια βάρη (weights).

Οι νευρώνες εισόδου (input layer) περιέχουν όλα τα πεδία που θα χρησιμοποιηθούν για να προβλέψουμε την ανεξάρτητη μεταβλητή (Y), οι νευρώνες εξόδου (output layer) περιέχουν μόνο ένα πεδίο, το προς εκτίμηση πεδίο (Y) και οι υπολογιστικοί νευρώνες ή κρυμμένο στρώμα (hidden layer) περιέχει έναν αριθμό νευρώνων που συνδυάζουν τα αποτελέσματα από προηγούμενες νευρώνες. Ουσιαστικά οι υπολογιστικοί νευρώνες πολλαπλασιάζουν τις εισόδους τους με τα συνοπτικά βάρη και υπολογίζουν το άθροισμα του γινομένου.

Ως εκ τούτου είναι ένα κύκλωμα διασυνδεδεμένων νευρώνων, που κάθε νευρώνας μπορεί να θεωρηθεί ως ένας επεξεργαστής στοιχείων. Οι συνδέσεις μεταξύ των νευρώνων παρέχουν στο δίκτυο την ικανότητα να μαθαίνει μοτίβα και σχέσεις μεταξύ των δεδομένων και των αποτελεσμάτων. Όλοι οι νευρώνες σε ένα στρώμα του δικτύου συνδέονται με όλους τους νευρώνες του επόμενου στρώματος. Κάθε νευρώνας έχει πολλές εισόδους αλλά μόνο μια έξοδο η οποία αποτελεί είσοδο για άλλους νευρώνες. Η μάθηση περιλαμβάνει αλλαγές στα συνοπτικά βάρη (όποτε κάνει μια λανθασμένη πρόβλεψη), που περιέχονται μεταξύ των νευρώνων.

Συγκεκριμένα ένα μέρος της εκπαίδευσης αποτελεί τη διαδικασία προσδιορισμού των καταλλήλων συντελεστών βάρους το οποίο πραγματοποιείται με την βοήθεια καταλλήλων αλγορίθμων. Αυτή η διαδικασία επαναλαμβάνεται πολλές φορές και το δίκτυο συνεχίζει να βελτιώνει τις προβλέψεις του έως ότου ικανοποιηθεί ένα ή περισσότερα από τα κριτήρια διακοπής. Ουσιαστικά έχουμε ένα σύνολο εισόδων και ένα σύνολο τιμών στόχων και προσπαθούμε να λάβουμε προβλέψεις που να ταιριάζουν με αυτές τις τιμές στόχου, όσο το δυνατόν πιο κοντά.

Το δίκτυο μαθαίνει μέσω της εκπαίδευσης. Αποτελέσματα για τα οποία είναι γνωστή η μεταβλητή στόχος παρουσιάζονται επανειλημμένα στο δίκτυο και οι απαντήσεις που δίνει συγκρίνονται με τα γνωστά αποτελέσματα. Οι πληροφορίες από αυτήν τη σύγκριση μεταφέρονται μέσω του δικτύου, αλλάζοντας σταδιακά τα βάρη. Καθώς προχωρά η εκπαίδευση, το δίκτυο γίνεται όλο και πιο ακριβές στην αναπαραγωγή των γνωστών αποτελεσμάτων. Μόλις εκπαιδευτεί, το δίκτυο μπορεί να εφαρμοστεί σε μελλοντικές περιπτώσεις όπου το αποτέλεσμα είναι άγνωστο.



Παρατηρώντας το παραπάνω σχήμα βλέπουμε την είσοδο των ανεξάρτητων μεταβλητών σαν πρώτο βήμα οι οποίες κατά την πρώτη φάση, κάθε είσοδος πολλαπλασιάζεται με το βάρος που της αντιστοιχεί στα hidden layer.

Άρα ένα νευρωνικό δίκτυο, έχει δύο βασικές λειτουργίες:

- Εκπαίδευση
- Πρόβλεψη

3.4.1 Περιγραφή Μεθόδου

Όπως αναφέραμε και παραπάνω το πρώτο στάδιο είναι η εκπαίδευση του νευρωνικού δικτύου. Στο στάδιο αυτό δημιουργείται ένα σύνολο μάθησης (training set), δηλαδή ένα σύνολο από διανύσματα εισόδου και εξόδων – αποτελεσμάτων. Αυτά ονομάζονται πρότυπα εκπαίδευσης. Χρησιμοποιώντας το σύνολο μάθησης και κατάλληλο αλγόριθμο, το νευρωνικό δίκτυο εκπαιδεύεται, δηλαδή υπολογίζει τα βάρη του.Τελικός σκοπός της εκπαίδευσης του νευρωνικού δικτύου είναι η ελαχιστοποίηση του σφάλματος πρόβλεψης στο σύνολο μάθησης.

Μετά την εκπαίδευση του νευρωνικού δικτύου ακολουθεί το στάδιο της πρόβλεψης. Στο στάδιο αυτό δημιουργείται ένα σύνολο ελέγχου (testing set), δηλαδή ένα σύνολο από πρότυπα ελέγχου. Στο στάδιο αυτό δίνονται μόνο τα διανύσματα εισόδου στο νευρωνικό δίκτυο και αυτό υπολογίζει τα προβλεπόμενα διανύσματα εξόδου. Ο υπολογισμός αυτός γίνεται, χρησιμοποιώντας τις τιμές των βαρών που υπολογίστηκαν κατά το στάδιο της εκπαίδευσης. Το σφάλμα πρόβλεψης στο σύνολο ελέγχου προκύπτει από το σφάλμα των προβλεπόμενων εξόδων του νευρωνικού δικτύου ως προς τις επιθυμητές εξόδους για κάθε ένα από τα πρότυπα ελέγχου.

Υπάρχουν πολλές μέθοδοι που μπορούν να χρησιμοποιηθούν για την εκπαίδευση ενός νευρωνικού δικτύου. Όμως δεν υπάρχει τρόπος για να προσδιοριστεί εκ των προτέρων ποια μέθοδος εκπαίδευσης λειτουργεί καλύτερα σε μία συγκεκριμένη εφαρμογή.

Σχόλιο

Ένα μεγάλο πλεονέκτημα των νευρωνικών δικτύων είναι ότι δεν έχουν ευαισθησία στην παρουσία δεδομένων με θόρυβο. Επιπρόσθετα μπορούν να αποθηκεύσουν γνώση και εμπειρία από το περιβάλλον, την οποία μπορεί στη συνέχεια να ανακαλέσει.

Κεφάλαιο 4

Μελέτες περιπτώσεων που αναλύθηκαν με τις παραπάνω μεθόδους

4.1 Διαχωριστική Ανάλυση

4.1.1 Πρόβλεψη τελικών αποτελεσμάτων μαθητών

Στόχος

Ο στόχος αυτής της έρευνας είναι να εξακριβώσει ποια θέματα καθορίζουν τα τελικά αποτελέσματα μαθητών. Ένας ακαδημαϊκός σύμβουλος αναμένεται να είναι σε θέση να προβλέψει τα τελικά αποτελέσματα των μαθητών. Για την επίτευξη αυτού του στόχου, μπορεί να εφαρμοστεί διαχωριστική ανάλυση.

Αυτή η έρευνα στοχεύει στον εξοπλισμό των ακαδημαϊκών συμβούλων με γνώσεις και εργαλεία που απαιτούνται για την παροχή καθοδήγησης. Ως μελέτη περίπτωσης, επιλέχθηκε μια συγκεκριμένη ερευνητική αρένα. Για λόγους εμπιστευτικότητας, η αρένα ονομάζεται Σχολή Πληροφορικής, Πανεπιστήμιο Χ στο Bandung, Δυτική Ιάβα, Ινδονησία. Τα ακαδημαϊκά αντίγραφα ορισμένων αποφοίτων χρησιμεύουν ως δεδομένα εισαγωγής για μια τη διαχωριστική ανάλυση.

Σε αυτήν την έρευνα:

• Αντικείμενα που μελετώνται είναι οι προπτυχιακοί φοιτητές

• Με βάση τα αποτελέσματα των βαθμών τους, υπάρχουν τρεις πιθανές ομάδες μαθητών:

1. Έκτακτα

2. Πολύ ικανοποιητικά και

3.Ικανοποιητικά

Τα χαρακτηριστικά των μαθητών είναι οι τελικοί βαθμοί σε ορισμένα μαθήματα. Στο προπτυχιακό εκπαιδευτικό σύστημα της Ινδονησίας, οι τελικοί βαθμοί ενός μαθήματος ταξινομούνται σε πέντε ομάδες, ως εξής: A (High Distinction), B (Distinction) C (Credit), D (Pass) και E (fail). Τα αποτελέσματα βαθμολογούνται ως 4 (A), 3 (B), 2 (C), 1 (D) και 0 (E), αντίστοιχα.

Με βάση τα διαθέσιμα χαρακτηριστικά των μαθητών, η τεχνική της διαχωριστικής ανάλυσης προσδιορίζει εκείνα που διακρίνονται μεταξύ των ομάδων. Τα αντίστοιχα χαρακτηριστικά είναι μαθήματα στα τέσσερα πρώτα εξάμηνα που σχετίζονται με την επιστήμη της Πληροφορικής.

Δείγμα

Υπάρχουν 146 αντίγραφα διαθέσιμα για αυτήν την έρευνα. Κάθε αντίγραφο περιείχε τους τελικούς βαθμούς 31 μαθημάτων από το 1ο έως το 8ο εξάμηνο. Επιλέχθηκαν τα πρώτα τέσσερα εξάμηνα ως χαρακτηριστικά αποφοίτων ή φοιτητών.

Τα μαθήματα αυτά ήταν IF102 (Εισαγωγή στην Εφαρμογή Υπολογιστών), IF103 (Εισαγωγή στην Πληροφορική), IF104 (Αλγόριθμοι και Προγραμματισμός), IF105 (Βασικός Προγραμματισμός), IF106 (Μαθηματικά Πληροφορικής), IF201 (Αγγλικά), IF202 (Γραμμική Άλγεβρα και πίνακες) , IF203 (Δίκτυο Υπολογιστών) και IF205 (Αρχεία και Σύστημα Πρόσβασης).

Αποτελέσματα

Τα αποτελέσματα της διαχωριστικής ανάλυσης είναι τα εξής:

Canonical Discriminant Function

Subjects	Function	
	1	2
IF102 (Introduction to Computer Application)	0.481	1.767
IF103 (Introduction to Information Technology)	0.858	-0.638
IF104 (Algorithms and Programming)	0.663	-0.334
IF105 (Basic Programming)	0.589	-0.128
IF202 (Linear Algebra and Matrices)	0.223	0.831
(Constant)	-9.255	-5.865

ΠΙΝΑΚΑΣ 1

Ο παραπάνω πίνακας υποδεικνύει ότι οι τελικοί βαθμοί των μαθητών στα ακόλουθα πέντε μαθήματα καθόρισε σημαντικά τα αποτελέσματα επιτυχίας τους: IF102 (Εισαγωγή στην Εφαρμογή Υπολογιστών), IF103 (Εισαγωγή στην Πληροφορική), IF104 (Αλγόριθμοι και Προγραμματισμός), IF105 (Βασικός Προγραμματισμός) και IF202 (Γραμμική Άλγεβρα και πίνακες).

Οι εξισώσεις των δύο διαχωριστικών συναρτήσεων είναι:

$$\text{Function\_1} = -9.255 + 0.481*\text{IF102} + 0.858*\text{IF103} + 0.663*\text{IF104} + 0.589*\text{IF105} + 0.223*\text{IF202}$$

$$\text{Function\_2} = -5.865 + 1.767*\text{IF102} - 0.638*\text{IF103} - 0.334*\text{IF104} - 0.128*\text{IF105} + 0.831*\text{IF202}$$

Για τους μελλοντικούς μαθητές, μόλις οι ερευνητές αποκτήσουν τους τελικούς βαθμούς των πέντε σημαντικών μαθημάτων, κάθε μαθητής θα αποκτήσει δύο βαθμολογίες από τις δύο διαχωριστικές συναρτήσεις.

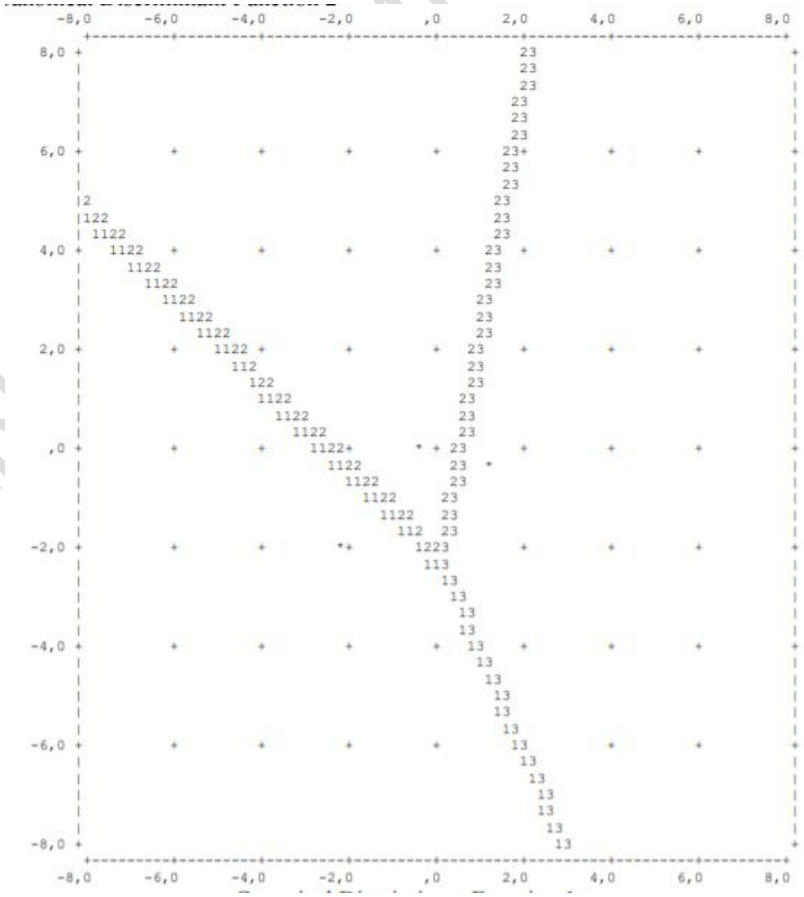
Μόλις σχεδιαστεί το σημείο με τη συντεταγμένη (Function_1, Function_2), λαμβάνεται η θέση αυτού του σημείου στον χάρτη. Αυτή η τοποθεσία δείχνει σε ποια από τις 3 κατηγορίες θα βρίσκεται ένας μελλοντικός μαθητής. Αυτό επιτρέπει σε έναν ακαδημαϊκό σύμβουλο να προβλέψει τα τελικά αποτελέσματα του φοιτητή, και αυτός είναι ο κύριος στόχος αυτής της έρευνας.

Για παράδειγμα:

Ας υποθέσουμε ότι ένας μαθητής έχει λάβει τα ακόλουθα αποτελέσματα: A, B, B, C και C στα ακόλουθα μαθήματα: IF102, IF103, IF104, IF105 και IF202, αντίστοιχα. Ως εκ τούτου, αυτός ο μαθητής θα αποκτήσει τις ακόλουθες δύο βαθμολογίες:

Function_1 = - 9.255 +0.481(4) + 0.858(3) + 0.663(3) + 0.589(2) + 0.223(2) = -1.144
Function_2 = - 5.865 +1.767(4) - 0.638(3) - 0.334(3) - 0.128(2) + 0.831(2) = 11.423

Στον παρακάτω χάρτη που απεικονίζεται στο Σχήμα 1, το σημείο (-1.144, 11.423) βρίσκεται στην Περιοχή 2, το οποίο δείχνει ότι μόνο από το ακαδημαϊκό αποτέλεσμα των τριών πρώτων εξαμήνων του, αυτός ο μαθητής προβλέπεται να λάβει τα τελικά αποτελέσματα του Πολύ Ικανοποιητικού.



ΣΧΗΜΑ 1

		Results	Predicted Group Membership			Total
			1.00	2.00	3.00	
Cross-validated	Count	1.00	3	1	0	4
		2.00	10	65	24	99
		3.00	0	5	38	43
	%	1.00	75.0	25.0	0.0	100.0
		2.00	10.1	65.7	24.2	100.0
		3.00	0.0	11.6	88.4	100.0

ΠΙΝΑΚΑΣ 2

Στον παραπάνω ***ΠΙΝΑΚΑΣ 2*** παρατηρούμε ότι:

Τέσσερις απόφοιτοι με τελικό αποτέλεσμα επιτυχίας 1 (Έκτακτα), 3 από αυτούς ταξινομήθηκαν σωστά και 1 εσφαλμένα ταξινομήθηκε ως αποτελέσματα 2 (Πολύ ικανοποιητικό), 99 απόφοιτοι με τελικά αποτελέσματα επιτυχίας 2 (Πολύ ικανοποιητικά), 65 από αυτούς ταξινομήθηκαν σωστά και 10 και 24, ταξινομήθηκαν λανθασμένα ως 1 (Έκτακτα) και 3 (Ικανοποιητικά), αντίστοιχα και τέλος, 43 απόφοιτοι με τελικό αποτέλεσμα επιτυχίας 3 (Ικανοποιητικό), 38 από αυτούς ταξινομήθηκαν σωστά και 5 εσφαλμένα ως αποτελέσματα 2 (Πολύ ικανοποιητικό).

Ως εκ τούτου, από τις 146 περιπτώσεις, 106 ταξινομήθηκαν σωστά, πράγμα που σημαίνει ότι η ακρίβεια της διαχωριστικής ανάλυσης είναι 72,6%.

Συμπέρασμα έρευνας

Αυτή η έρευνα βοηθά τους ακαδημαϊκούς συμβούλους να προβλέψουν τα τελικά αποτελέσματα ενός φοιτητή με βάση την απόδοσή του σε ορισμένα θέματα στο αρχικό στάδιο (τα πρώτα τέσσερα εξάμηνα) της σπουδής τους στην τριτοβάθμια εκπαίδευση. Αυτό το είδος διευκόλυνσης επιτρέπει στους ακαδημαϊκούς συμβούλους να βοηθήσουν τους μαθητές να καταρτίσουν τα σχέδια σπουδών τους κάθε εξάμηνο, προκειμένου να αποδώσουν στο μέγιστο.

4.2 Δέντρα αποφάσεων

4.2.1 Προσδιορισμός μοτίβων συμπεριφορών προστασίας από τον ήλιο και ευαισθησίας στον ήλιο που σχετίζονται με το ηλιακό έγκαυμα

Στόχος

Ο καρκίνος του δέρματος αντιπροσωπεύει μια σημαντική ανησυχία για τη δημόσια υγεία, με σχεδόν 100.000 Αμερικανούς να γιγνώσκονται με κακοήθη μελάνωμα το 2018. Επιπλέον, περισσότεροι από 5 εκατομμύρια Αμερικανοί γιγνώσκονται με καρκίνο του δέρματος χωρίς μελάνωμα κάθε χρόνο. Οι περισσότεροι καρκίνοι του δέρματος προκαλούνται από υπερβολική έκθεση στην υπεριώδη ακτινοβολία UV και θα μπορούσε να προληφθεί μέσω της χρήσης συμπεριφορών προστασίας από τον ήλιο. Οι συστάσεις για την προστασία από τον ήλιο περιλαμβάνουν την αποφυγή του ήλιου, την αναζήτηση σκιάς και τη χρήση αντηλιακού, καθώς και προστατευτικά ρούχα, καπέλα και γυαλιά ηλίου. Παρά τις προσπάθειες παρέμβασης, ωστόσο, τα ποσοστά καρκίνου του δέρματος συνεχίζουν να αυξάνονται.Ο σκοπός της έρευνας είναι να θέσει τους περιορισμούς των μετρήσεων της ηλιακής προστασίας χρησιμοποιώντας μια προσέγγιση δέντρων αποφάσεων.

Μεταβλητές

- Sun Sensitivity

Ζητήθηκε από τους συμμετέχοντες να δηλώσουν τι θα συνέβαινε στο δέρμα τους εάν βρισκόταν στον ήλιο για 1 ώρα μετά από αρκετούς μήνες που δεν εκτέθηκαν. Οι αποκρίσεις κωδικοποιήθηκαν ως ευαισθησία στον ήλιο (δηλ., «σοβαρό έγκαυμα με φουσκάλες», «μέτριο έγκαυμα με ξεφλούδισμα» και «κάψιμο ελαφρώς με λίγο ή χωρίς μαύρισμα») ή μη ευαισθησία στον ήλιο (δηλαδή, «μαύρισμα χωρίς έγκαυμα »και« τίποτα δεν θα συμβεί »).

- Sunburn

Οι συμμετέχοντες δήλωσαν πόσες φορές είχαν υποστεί ηλιακό έγκαυμα τους τελευταίους 12 μήνες. Οι περισσότεροι συμμετέχοντες (66,3%) ανέφεραν ότι δεν είχαν περιστατικά ηλιακού εγκαύματος τον τελευταίο χρόνο. Μεταξύ εκείνων που ανέφεραν ηλιακά εγκαύματα (33,7%), το 16,3% ανέφερε 1 ηλιακό έγκαυμα, το 9,9%, 2 ηλιακά εγκαύματα και 5,8%, 3 έως 5 ηλιακά εγκαύματα. Λιγότερο από το 1% του δείγματος ανέφερε 6 έως 360 ηλιακά εγκαύματα. Αυτή η μεταβλητή κωδικοποιήθηκε για να δημιουργήσει μια δυαδική μεταβλητή αποτελέσματος για το ηλιακό έγκαυμα (0 έναντι ≥1).

- Sun Protection

Έξι στοιχεία αξιολόγησαν μέτρα προστασίας από τον ήλιο. Ζητήθηκε από τους ερωτηθέντες να δηλώσουν σε μια ζεστή, ηλιόλουστη μέρα, πόσο συχνά φορούσαν αντηλιακό, αναζητούσαν σκιά, φορούσαν καπέλο, φορούσαν μακρυμάνικο και φορούσαν μακρύ παντελόνι. Οι επιλογές απάντησης περιλάμβαναν «πάντα», «τις περισσότερες φορές», «μερικές φορές», «σπάνια» και «ποτέ».

Δείγμα και Μέθοδος Chaid

Εντοπίστηκαν 28.558 περιπτώσεις με πλήρη δεδομένα. Η ευαισθησία στον ήλιο (Sun Sensitivity) και οι 6 μεταβλητές προστασίας από τον ήλιο (Sun Protection) χρησιμοποιήθηκαν ως ανεξάρτητες μεταβλητές και η επίπτωση του ηλιακού εγκαύματος (Sunburn) χρησιμοποιήθηκε ως εξαρτημένη. Η μέθοδος CHAID έχει σχεδιαστεί για να λειτουργεί με κατηγορικές μεταβλητές και χρησιμοποιήθηκε για την ανάλυση. Η ανάλυση καθορίζει ποιες μεταβλητές έχουν στατιστικά σημαντική σχέση με την εξαρτημένη και διαιρεί το δείγμα με βάση αυτήν τη μεταβλητή.

Το δέντρο συνεχίζει να διακλαδίζεται έως ότου δεν βρεθούν περαιτέρω στατιστικά σημαντικές διασπάσεις ή δεν πληρούνται τα κριτήρια διακοπής. Ο τελικός κόμβος μετά τον οποίο δεν εμφανίζεται περαιτέρω διαχωρισμός ονομάζεται child κόμβος. Χρησιμοποιήθηκε test Pearson X^2. Τα κριτήρια διακοπής ήταν τουλάχιστον 100 περιπτώσεις σε έναν parent κόμβο και 50 περιπτώσεις σε έναν child κόμβο. Η μέθοδος CHAID χρησιμοποιεί μια διόρθωση Bonferroni για να χωρίσει τους κόμβους και προσπαθεί να ελέγξει το μέγεθος του δέντρου (δηλαδή, αποφεύγετε η υπερβολική τοποθέτηση) διαχωρίζοντας έναν κόμβο μόνο εάν πληρείται το κριτήριο σημασίας. Από προεπιλογή, το CHAID επιλέγει τη σειρά των μεταβλητών εισαγωγής με σχετική σημασία (δηλαδή, την υψηλότερη τιμή χ^2).

Δημιουργήθηκαν ξεχωριστά σύνολα δεδομένων για εκπαίδευση (training) και επικύρωση (testing). Ένα τυχαίο δείγμα 29,3% των περιπτώσεων διατηρήθηκε ως αναμονή για επικύρωση, το υπόλοιπο δείγμα χρησιμοποιήθηκε για εκπαίδευση. Το δέντρο που είχε ως αποτέλεσμα την απλούστερη δομή με το χαμηλότερο σφάλμα πρόβλεψης επιλέχθηκε ως το τελικό μοντέλο.

Αποτελέσματα

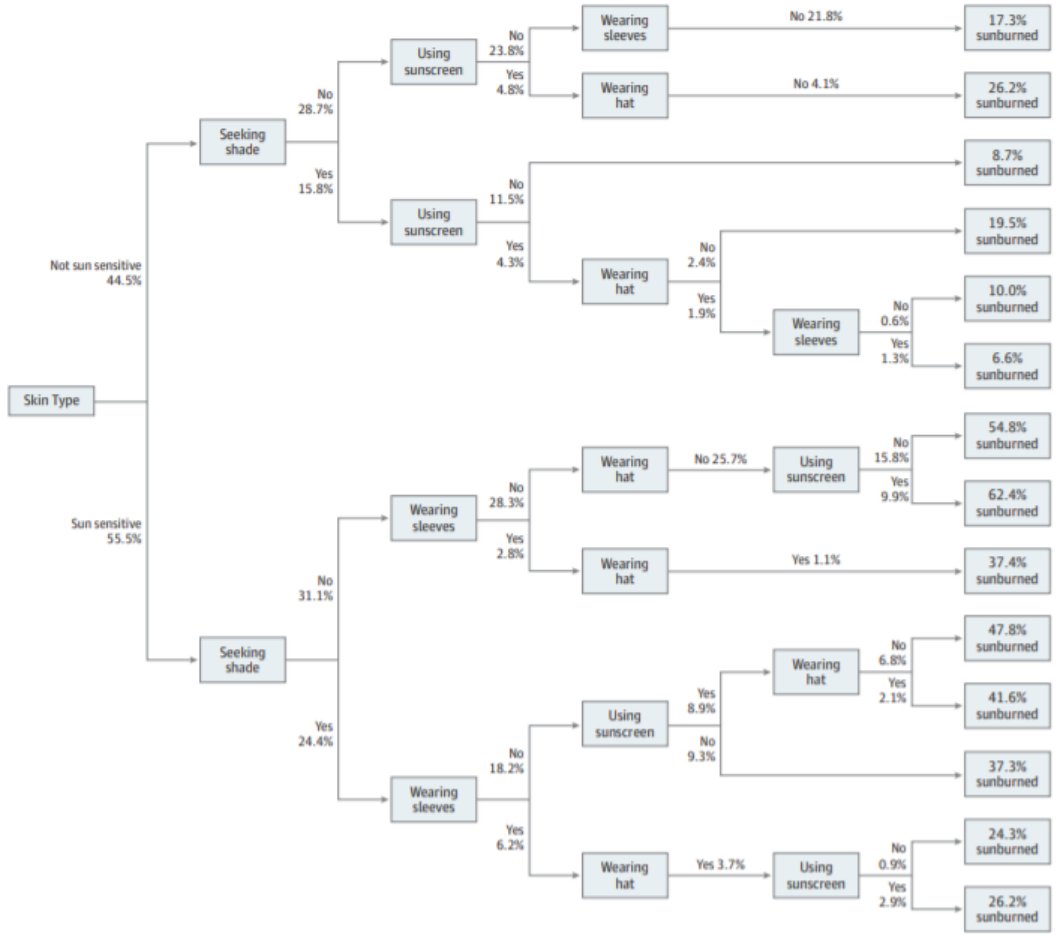
Χαρακτηριστικά συμμετεχόντων

Μεταξύ των 28.558 ερωτηθέντων με πλήρη στοιχεία, 13 104 (45,9%) ήταν άνδρες και 15 454 (54,1%) ήταν γυναίκες. Οι συμμετέχοντες κυμαίνονταν σε ηλικία από 18 έως 85 ετών, με μέσο όρο ηλικίας 49. Το μεγαλύτερο μέρος του δείγματος 22.285 (78,0%) ήταν λευκοί, 3.614 συμμετέχοντες (12,7%) ήταν μαύροι και 2.575 (9,0%) Ασιάτης, Ιθαγενής Ινδός ή Αλάσκα, ή περισσότερες από 1 φυλή.

Το μεγαλύτερο μέρος του δείγματος, 15.992 συμμετέχοντες (56,0%) είχε ευαίσθητο δέρμα στον ήλιο, 12.566 συμμετέχοντες (44,0%) δεν είχαν. Τα περισσότερα άτομα, 22.110 συμμετέχοντες (77,4%) έπαιρναν έστω 1 μέτρο προστασίας για τον ήλιο. Η αναζήτηση σκιάς ήταν το πιο συνηθισμένο μέτρο, 11.369 περιπτώσεις (39,8%) και αυτοί που φορούσαν μακρυμάνικο ήταν το λιγότερο κοινό μέτρο, 4.502 περιπτώσεις (15,8%).
Στο δείγμα εκπαίδευσης (20.185), η εκτίμηση κινδύνου, δηλαδή το ποσοστό των περιπτώσεων που ταξινομήθηκαν λανθασμένα ήταν 30,5%.

Το μοντέλο ταξινόμησε σωστά 14.023 περιπτώσεις (69,5%) συνολικά από τις 13.362 περιπτώσεις χωρίς ηλιακό έγκαυμα, 11.052 (82,7%) ταξινομήθηκαν σωστά και από τις 6.823 περιπτώσεις με ηλιακό έγκαυμα, 2.971 (43,5%) ταξινομήθηκαν σωστά. Στο δείγμα επικύρωσης (8.373), η εκτίμηση κινδύνου ήταν 0,296.Το μοντέλο ταξινόμησε σωστά 5.898 περιπτώσεις (70,4%) συνολικά από τις 5.565 περιπτώσεις χωρίς ηλιακό έγκαυμα, 4.655 (83,6%) ταξινομήθηκαν σωστά και από τις 2.808 περιπτώσεις με ηλιακό έγκαυμα, 1.243 (44,3%) ταξινομήθηκαν σωστά.

Επιλέχθηκε για απεικόνιση μερική διαδρομή δέντρου απόφασης για 8.373 άτομα με και χωρίς ευαισθησία στον ήλιο



Περιγραφή μερικών αποτελεσμάτων που διακρίνονται στο παραπάνω δέντρο:

Μεταξύ των ευαίσθητων στον ήλιο ατόμων, εκείνοι που έπαιρναν και τα 4 προστατευτικά μέτρα (αναζητώντας σκιά, φορώντας μακριά μανίκια, φορώντας καπέλο και χρησιμοποιώντας αντηλιακό) είχαν 26,2% πιθανότητα ηλιακού εγκαύματος. Οι συμμετέχοντες που ανέφεραν μόνο αναζήτηση σκιάς είχαν πιθανότητα 37,3% ηλιακού εγκαύματος. Ομοίως, όσοι ανέφεραν ότι φορούσαν μακριά μανίκια και φορούσαν καπέλο είχαν 37,4% πιθανότητα ηλιακού εγκαύματος.

Παρόλο που οι συμμετέχοντες που δεν χρησιμοποίησαν αντηλιακό, δεν αναζητούσαν σκιά ή φορούσαν προστατευτικά ρούχα είχαν μεγαλύτερη πιθανότητα ηλιακού εγκαύματος (54,8%), η ομάδα με την υψηλότερη πιθανότητα ηλιακού εγκαύματος συνίστατο σε εκείνους που χρησιμοποίησαν μόνο αντηλιακό (62,4%). Η ομάδα με τη χαμηλότερη πιθανότητα του ηλιακού εγκαύματος δεν ανέφεραν χρήση αντηλιακού αλλά ανέφεραν ότι χρησιμοποιούσαν τα άλλα 3 μέτρα (24,3%).

Μεταξύ των ατόμων που δεν είναι ευαίσθητα στον ήλιο, τα μονοπάτια αποφάσεων ήταν λιγότερο περίπλοκα. Ωστόσο, προέκυψαν παρόμοια μοτίβα. Όσοι έπαιρναν και τα 4 προστατευτικά μέτρα είχαν τη χαμηλότερη πιθανότητα ηλιακού εγκαύματος (6,6%). Όσοι δεν έπαιρναν κανένα μέτρο προστασίας είχαν 17,3% πιθανότητα ηλιακού εγκαύματος. Ωστόσο, η υψηλότερη πιθανότητα ηλιακού εγκαύματος ήταν μεταξύ εκείνων που ανέφεραν μόνο χρήση αντιηλιακών (26,2%). Όσοι αναφέρουν ότι αναζητούν σκιά αλλά δεν χρησιμοποιούν αντηλιακό είχαν πιθανότητα 8,7% ηλιακού εγκαύματος.

4.3 Γραμμική Παλινδρόμηση

4.3.1 Πρόβλεψη πωλήσεων ηλεκτρικών αυτοκινήτων

Στόχος

Τα ηλεκτρικά οχήματα είναι μια νέα μορφή οχήματος, φιλικά προς το περιβάλλον και γίνονται δημοφιλής τα τελευταία χρόνια. Είναι σημαντικό τόσο για τους καταναλωτές όσο και για τους παραγωγούς να αναζητήσουν παράγοντες που θα μπορούσαν να επηρεάσουν τις πωλήσεις τους. Σε αυτή τη μελέτη, εξετάζονται τρία μοντέλα γραμμικής παλινδρόμησης για τον προσδιορισμό παραγόντων που έχουν σημαντική επίδραση στις πωλήσεις αυτών. Τα ηλεκτρικά οχήματα είναι σε θέση να αντικαταστήσουν τα συμβατικά οχήματα που καταδικάζονται από όλους τους τομείς της κοινότητας για υψηλές εκπομπές διοξειδίου του άνθρακα (ένας σημαντικός παράγοντας στην υπερθέρμανση του πλανήτη). Λόγω της περιορισμένης πρόσβασης στην τεχνολογία και στην ανώριμη αγορά, ορισμένοι παράγοντες (όπως η ανταγωνιστική τιμή) εξακολουθούν να επηρεάζουν δυσμενώς τις πωλήσεις των ηλεκτρικών οχημάτων. Έτσι, αυτή η μελέτη στοχεύει να εξετάσει παράγοντες που επηρεάζουν τις πωλήσεις των οχημάτων αυτών. Σε αυτή τη μελέτη, εξετάζονται τρία μοντέλα γραμμικής παλινδρόμησης για τον προσδιορισμό παραγόντων που έχουν σημαντική επίδραση στις πωλήσεις των ηλεκτρικών οχημάτων.

Προτεινόμενες μεταβλητές

Οι μεταβλητές που προτείνονται για χρήση είναι οι εξής:

Oil Price (USD/barrel), Oil Volume (per barrel), CPI, Disposable Income (per person), Exchange Rate (USD/CNY), Number of Charging Station (per unit), Subsidies (billion), Literacy level (%).

Περιγραφή μεταβλητών

Oil Price (USD/barrel), Oil Volume (per barrel): Η τιμή της βενζίνης και ο όγκος της θα μπορούσαν να είναι ένας κρίσιμος παράγοντας για την πώληση ηλεκτρικών (Kah, 2018). Είναι σημαντικό να εξεταστεί η σχέση μεταξύ τιμής βενζίνης, όγκου βενζίνης και ηλεκτρικού οχήματος.

Υπόθεση 1: Η τιμή της βενζίνης έχει θετική επίδραση στις πωλήσεις των ηλεκτρικών οχημάτων

Υπόθεση 2: Ο όγκος της βενζίνης έχει αρνητικές επιπτώσεις στις πωλήσεις των ηλεκτρικών οχημάτων

CPI (Consumer Price Index): Ένας σημαντικός δείκτης που παρουσιάζει εάν οι άνθρωποι μπορούν να αντέξουν οικονομικά τις δαπάνες τους (Burns et al., 2008). Είναι ένα σημείο αναφοράς που δημιουργήθηκε από την κυβέρνηση για τη μέτρηση του πληθωρισμού. Οι άνθρωποι μπορεί να είναι λιγότερο πρόθυμοι να δοκιμάσουν ηλεκτρικά οχήματα με υψηλό πληθωρισμό (Zhang, 2017).

Υπόθεση 3: Ο CPI έχει αρνητικές επιπτώσεις στις πωλήσεις των ηλεκτρικών οχημάτων

Disposable Income (per person): Το διαθέσιμο εισόδημα θα μπορούσε να είναι ένας άλλος σημαντικός παράγοντας. Ένα υψηλό διαθέσιμο εισόδημα εγγυάται ότι οι άνθρωποι μπορούν να αντέξουν οικονομικά το κόστος αγοράς ενός ηλεκτρικού οχήματος.

Υπόθεση 4: Το διαθέσιμο εισόδημα έχει θετική επίδραση στις πωλήσεις Ηλεκτρικών οχημάτων

Exchange Rate (USD/CNY): Θεωρείται η αξία του νομίσματος μιας χώρας σε σχέση με ένα άλλο νόμισμα. Η επιλογή CNY και USD ως αντικειμένου της έρευνας οφείλεται στο γεγονός ότι όλο και περισσότερα ανταλλακτικά αυτοκινήτων κατασκευάζονται στην Κίνα και εξάγονται στην Αμερική (Nykvist et al., 2015).

Υπόθεση 5: Η συναλλαγματική ισοτιμία έχει θετική επίδραση στις πωλήσεις ηλεκτρικών οχημάτων

Number of Charging Station (per unit): Η ποσότητα σταθμών φόρτισης σε αυτή τη μελέτη είναι ο αριθμός των σταθμών φόρτισης στην Αμερική. Ο αυξανόμενος αριθμός σταθμών φόρτισης αντικατοπτρίζει αποτελεσματικά την ανάπτυξη ηλεκτρικών οχημάτων (Li, 2016). Αυτή η μελέτη προτείνει ότι ο αριθμός των σταθμών φόρτισης μπορεί να είναι ένας παράγοντας που επηρεάζει τις πωλήσεις των ηλεκτρικών οχημάτων.

Υπόθεση 6: Ο αριθμός των σταθμών φόρτισης έχει θετική επίδραση στις πωλήσεις των ηλεκτρικών οχημάτων

Subsidies (billion): Οι επιδοτήσεις είναι μια μορφή χρηματοδοτικής βοήθειας ή στήριξης που επεκτείνεται σε έναν οικονομικό τομέα γενικά με σκοπό την προώθηση της οικονομικής και κοινωνικής πολιτικής (Qian, 2018). Το ότι οι άνθρωποι μπορούν να λάβουν χρήματα είναι ένα γεγονός που τονώνει το ενδιαφέρον τους για την αγορά ηλεκτρικών οχημάτων (Holtsmark et al., 2014).

Υπόθεση 7: Οι επιδοτήσεις έχουν θετική επίδραση στις πωλήσεις ηλεκτρικών οχημάτων

Literacy level (%): Το επίπεδο παιδείας θα μπορούσε να επηρεάσει την προθυμία των ανθρώπων να αγοράσουν ηλεκτρικά οχήματα. Το υψηλότερο επίπεδο προσφέρει στους ανθρώπους πλήρη κατανόηση της νέας ενέργειας και ενισχύει τα ενδιαφέροντά τους για την αγορά ηλεκτρικών οχημάτων (Kortland, 1993).

Υπόθεση 8: Το επίπεδο παιδείας έχει θετική επίδραση στις πωλήσεις ηλεκτρικών οχημάτων

Μεθοδολογία

Στο μοντέλο, το Y αντιπροσωπεύει τις πωλήσεις ηλεκτρικών οχημάτων, ενώ τα X αντιπροσωπεύουν όλους τους παράγοντες που θα μπορούσαν να επηρεάσουν τις πωλήσεις των ηλεκτρικών οχημάτων.

Οι μεταβλητές ορίζονται οι εξής:

Y: sales of New-energy Vehicle (per unit)
X1: Oil Price (USD/barrel)
X2: Oil Volume (per barrel)
X3: CPI
X4: Disposable Income (per person)
X5: Exchange Rate (USD/CNY)
X6: Number of Charging Station (per unit)
X7: Subsidies (billion)
X8: Literacy level (%)

Όλα τα δεδομένα που χρειάστηκαν σ αυτή την έρευνα, για κάθε μεταβλητή που χρησιμοποιήθηκε συλλέχτηκαν από αρμόδιες εταιρείες.(αναφέρονται στο paper).

Αναλυση

Η ανάλυση πραγματοποιήθηκε με την R. Αρχικά έγινε μια περιγραφική ανάλυση των μεταβλητών:

Distribution	Car sale	Oil Price	Oil Volume	CPI	Disposable Income	Charging Station	subsidies	Literacy	Currency
Maximum	286,367	113.93	5,300,000	252.9	14,640	61,067	85.3	86.4	6.976
Minimum	345	33.62	3,590,000	219.2	12,027	3,394	56.01	84.56	6.054
Mean	50,141	73.94	8,129,072	237	13,119	27,251	60.29	85.78	6.424
Median	13,038	73.25	7,170,000	237.1	13,139	25,602	59.39	85.81	6.357
Interquartile	73,676	46.71	5,580,000	10.4	1,327	28,637	2.4	0.89	0.388
Range	286,022	80.31	1,710,000	33.7	2,613	57,673	29.29	1.84	0.922

Παρατηρήθηκε ότι οι πωλήσεις ηλεκτρικών οχημάτων (ανά μονάδα) κυμαίνονται από 345 έως 286.367 με μέσο όρο 50.141. Ο αριθμός αυξάνεται με το χρόνο. Η τιμή της βενζίνης (USD / BBL) κυμαίνεται από 33,62 έως 113,93 με μέσο όρο 73,94, μετά από μια σταθερή αύξηση των πωλήσεων κάθε χρόνο. Ο όγκος της βενζίνης (ανά BBL) κυμαίνεται από 3.590.000 έως 8.129.072 με μέσο όρο 530.000. Υπήρξε μια μικρή αύξηση των πωλήσεων κατά τη διάρκεια του έτους. Ο δείκτης CPI κυμαίνεται από 219,2 έως 252,9 με μέσο όρο 237. Ο αριθμός αυξάνεται σταθερά τα τελευταία χρόνια.

Το διαθέσιμο εισόδημα (ανά άτομο) κυμαίνεται από 12.027 έως 14.640 με μέσο όρο 13.119. Υπήρχε βελτίωση των αριθμών κάθε χρόνο. Ο αριθμός των σταθμών φόρτισης κυμαίνεται από 3394 έως 61.067 με μέσο όρο 27.251. Ο αριθμός αυξάνεται με το χρόνο. Όλες οι επιδοτήσεις (ανά δισεκατομμύριο) κυμαίνονται από 56,01 έως 85,3 με μέσο όρο 60,29, με σταθερά αυξανόμενο αριθμό κάθε χρόνο. Το επίπεδο παιδείας (%) κυμαίνεται από 84,56% έως 86,4% με μέσο όρο 85,78. Ο αριθμός αυξάνεται με το χρόνο. Η συναλλαγματική ισοτιμία (CNY / USD) κυμαίνεται από 6.054 έως 6.976 με μέσο όρο 6.424. Ο αριθμός αυξάνεται σταθερά κατά τη διάρκεια του έτους

Μοντέλο Παλινδρόμησης

Το 1ο μοντέλο χρησιμοποιεί όλες τις ανεξάρτητες μεταβλητές. Τα αποτελέσματα ακολουθούν στον ***ΠΙΝΑΚΑΣ 1.***

	Estimate	Std. Error	t-value	Pr(>|t|)	
(intercept)	3.274e+06	9.072e+05	3.608	0.000511	***
Oil Price	2.432e+02	1.928e+02	1.261	0.21049	
Oil Volume	-1.350e-03	1.235e-03	-1.093	0.277345	
CPI	4.536e+02	1.497e+03	0.303	0.762537	
Disposable Income	5.073e+01	1.342e+01	3.781	0.000284	***
Currency	1.467e+04	1.168e+04	1.256	0.212473	
Charging Station	2.239e+00	8.376e-01	2.672	0.008971	**
All subsidies	3.045e+03	4.601e+02	6.617	2.77e-09	***
Literacy	-5.062e+04	1.086e+04	-4.663	1.10e-05	***

ΠΙΝΑΚΑΣ 1

Σύμφωνα με τα αποτελέσματα με επίπεδο σημαντικότητας α=0,05,βλέποντας τα p_{value} παρατηρούμε ότι οι μεταβλητές oil price (p_{value} = 0,21), oil volume (p_{value} = 0,28),CPI (p_{value} = 0,76) και Currency (p_{value} = 0,21) είναι τα $p_{value} > \alpha$, επομένως δεν είναι στατιστικά σημαντικές. Το R^2 του μοντέλου είναι 92,06%.

Επειδή οι προαναφερθείσες μεταβλητές είχαν p_{value} κοντά στο 0,1, αποφασίστηκε να αφαιρεθεί μόνο η μεταβλητή CPI με p_{value} = 0,76 για την εφαρμογή του 2ου μοντέλου. Τα αποτελέσματα του μοντέλου αυτού παρουσιάζονται στον ***ΠΙΝΑΚΑΣ 2.***

	Estimate	Std. Error	t-value	Pr(>|t|)	
(intercept)	3.264e+06	9.021e+05	3.618	0.000492	***
Oil Price	2.773e+02	1.558e+02	1.780	0.078466	
Oil Volume	-1.338e-03	1.228e-03	-1.089	0.278954	
Disposable Income	5.162e+01	1.302e+01	3.964	0.000149	***
Currency	1.358e+04	1.106e+04	1.228	0.222649	
Charging Station	2.407e+00	6.248e-01	3.852	0.000221	***
All subsidies	3.064e+03	4.532e+02	6.762	1.39e-09	***
Literacy	-4.941e+04	1.004e+04	-4.920	3.95e-06	***

ΠΙΝΑΚΑΣ 2

Τα αποτελέσματα δείχνουν ότι οι μεταβλητές oil price (p_{value} = 0,08), Currency (p_{value} = 0,22) και Oil volume (p_{value} = 0,28) συνεχίζουν να μην είναι στατιστικά σημαντικές για το μοντέλο. Το R^2 του μοντέλου αυτού είναι 92,14%.Σύμφωνα λοιπόν με τα αποτελέσματα εξαιρούνται οι τρεις παραπάνω μεταβλητές και εφαρμόζεται ένα περαιτέρω μοντέλο, εξαιρουμένων αυτών των τριών.

Το αποτέλεσμα του 3^{ου} μοντέλου δείχνει ότι όλοι οι παράγοντες επηρεάζουν σημαντικά τις πωλήσεις των ηλεκτρικών αυτοκινήτων. Και το R^2 του μοντέλου είναι 92%.

Anova

Τέλος, εκτελείται ανάλυση ANOVA για να ελεγχτεί η εγκυρότητα του μοντέλου και να συγκριθεί η επίδραση των τριών μοντέλων που αποκτήθηκαν.

	Res.Df	RSS	DF	Sum of Sq	F	Pr(>F)
1	88	2.6405e+10				
2	89	2.6432e+10	-1	-27563732	0.0919	0.7625
3	92	2.8500e+10	-3	-2067590171	2.2969	0.08314

ΠΙΝΑΚΑΣ 3

Και οι δύο τιμές (p_{value}= 0,76 και p_{value} = 0,08) είναι μεγαλύτερες από 0,05, πράγμα που σημαίνει ότι δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση σημαντικών διαφορών μεταξύ των μοντέλων. Υποθέτοντας ότι αυτό είναι ένα τυχαίο και ανεξάρτητο δείγμα που επιλέγεται από έναν κανονικό πληθυσμό και η διακύμανση του δείγματος είναι η ίδια, μπορούμε να συμπεράνουμε ότι και τα τρία μοντέλα δεν έχουν σημαντική διαφορά. Είμαστε σίγουροι ότι θα αποδεχτούμε το 3^{ο} μοντέλο ως το τελικό μοντέλο για να αποφασιστεί ποιες μεταβλητές θα επηρεάσουν τις πωλήσεις των ενεργειακών οχημάτων.

Συμπέρασμα

Τα αποτελέσματα δείχνουν ότι το διαθέσιμο εισόδημα, ο αριθμός των σταθμών φόρτισης, οι επιδοτήσεις και το επίπεδο αλφαβητισμού επηρεάζουν θετικά τις πωλήσεις των ηλεκτρικών οχημάτων.

4.4 Παραγοντική Ανάλυση

3.4.1 Μοντελοποίηση των παραγόντων που σχετίζονται με τη χρήση της ζώνης ασφαλείας από τους νέους οδηγούς στην Αθήνα

Στόχος - Δείγμα

Ο κύριος στόχος αυτής της μελέτης είναι να προσδιορίσει και να αποσαφηνίσει τη σχέση μεταξύ των προθέσεων των νέων οδηγών (για χρήση ζώνης ασφαλείας/ μη χρήση ζώνης ασφαλείας) και της συμπεριφοράς τους. Επιπλέον, ο σκοπός αυτής της μελέτης είναι να αξιολογήσει τα ποσοστά νεαρών οδηγών που φορούν ζώνη ασφαλείας (SB) σε σχέση με το είδος μετακίνησης τους. Το δείγμα αποτελούνταν από 200 νέους Έλληνες οδηγούς και των δύο φύλων. Το δείγμα αυτής της μελέτης αποτελούνταν από 127 άνδρες και 73 γυναίκες.

Μεταβλητές

Στο ερωτηματολόγιο συμπεριλήφθηκαν 115 διαφορετικές μεταβλητές, οι οποίες χωρίστηκαν σε τέσσερις ενότητες:

Η 1η ενότητα αναφέρεται στα κοινωνικο-δημογραφικά χαρακτηριστικά του δείγματος, συμπεριλαμβανομένου του φύλου, της ηλικίας, του τόπου προέλευσης, της οικογενειακής κατάστασης, του εκπαιδευτικού επιπέδου και του επαγγέλματος

Η 2η ενότητα χωρίστηκε σε πέντε μέρη. Το 1ο και το 2ο μέρος αφορούσαν την εμπειρία των νέων οδηγών. Το 3ο αναφέρεται στη συχνότητα με την οποία ένας νεαρός οδηγός οδηγεί σε διαφορετικό είδος μετακίνησης, π.χ. να πάει σε ένα μπαρ, για επαγγελματικούς λόγους κ.λπ. Το 4ο μέρος της δεύτερης ενότητας αναφέρθηκε στη συχνότητα με την οποία οι νεαροί οδηγοί του δείγματος οδηγούν σε διαφορετικές ημέρες και ώρες. Το τελευταίο 5ο μέρος αναφέρεται στην κατανάλωση αλκοόλ, κάθε μέρα, αρκετές φορές την εβδομάδα, μερικές φορές το μήνα, σπάνια και ποτέ.

Στην 3η ενότητα, χρησιμοποιήθηκαν 30 στοιχεία για τη μέτρηση των κινήτρων (προθέσεων) για τη χρήση ζώνης ασφαλείας (SB). Κάθε στοιχείο εμφανίστηκε σε μορφή δήλωσης. Παραδείγματα δηλώσεων ήταν: «Φοράω SB για να αποφύγω τραυματισμό» ή «… για να αποφύγω ποινές» ή λόγω: «… υπακοή στους κρατικούς κανόνες», «… εκπαίδευση και πληροφορίες σχετικά με τα οφέλη» ή «… πρώην τραυματισμός», «… Καιρικές συνθήκες ή οδικές συνθήκες».

Για τη μέτρηση των κινήτρων μη χρήσης SB, χρησιμοποιήθηκαν 36 αντικείμενα και κατατάχθηκαν με τον ίδιο τρόπο όπως στο προηγούμενο μέρος. Αυτές περιλάμβαναν δηλώσεις όπως «Δεν φοράω SB επειδή οδηγώ αργά» ή «… γιατί ξεχάσω να το κάνω» ή «… Φοβάμαι ότι παγιδεύομαι ή πνίγω σε ατύχημα», «Δεν φοράω ζώνη ασφαλείας λόγω μειωμένης οδηγικής απόλαυσης »ή«… γιατί τίποτα δεν θα συμβεί σε μένα ».

Η 4η και τελευταία ενότητα εφάρμοσε την μέθοδο μέσω συνεντεύξεων για τη συλλογή δεδομένων. Το ποσοστό που φορά ζώνη ασφαλείας (SB) μετρήθηκε σε σχέση με επτά είδη μετακίνησης: κατά τη διάρκεια της εργασίας / του σχολείου, προς ή από την εργασία / το σχολείο, πηγαίνοντας σε μπαρ, ντίσκο, πάρτι, χορό ή κάτι παρόμοιο, επιστροφή από μπαρ, ντίσκο, πάρτι, χορό ή κάτι άλλο παρόμοιο, απλά οδηγώ προς ή από μια λέσχη, μια ένωση ή κάτι παρόμοιο και προς ή από άλλο συμβάν. Η χρήση SB σχετικά με τον τύπο μετακίνησης σε μπαρ, κλαμπ κ.λπ. (η οποία, σύμφωνα με τη βιβλιογραφία, μπορεί να περιλαμβάνει κατανάλωση αλκοόλ) αξιολογήθηκε με δύο διαφορετικές μεταβλητές (μετάβαση και επιστροφή από), έτσι ώστε μια καλύτερη κατανόηση μιας πιθανής διαφοροποίησης. Η μεταβλητή SB χρησιμοποιείται κατά την επιστροφή από ένα μέρος ψυχαγωγίας με μέρος αναφοράς το σπίτι.

Παραγοντική Ανάλυση

Η στατιστική ανάλυση πραγματοποιήθηκε με ανάλυση παραγόντων και η μοντελοποίηση δεδομένων κατασκευάστηκε μέσω πολλαπλής γραμμικής παλινδρόμησης. Η ανάλυση παραγόντων χρησιμοποιώντας περιστροφή με τη μέθοδο Varimax χρησιμοποιήθηκε για τα 30 αντικείμενα που μετρούν τη χρήση SB και για τα 36 αντικείμενα που μετρούν τη μη- χρήση SB. Επιλέχθηκε για δύο βασικούς λόγους. Πρώτον, για να βρεθούν οι μεταβλητές (παράγοντες) που επηρεάζουν και δεύτερον, σε μια προσπάθεια μείωσης του μεγάλου αριθμού των μεταβλητών.

Αποτελέσματα Παραγοντικής Ανάλυσης

Τα αποτελέσματα των δύο παραγοντικών αναλύσεων μετά την περιστροφή με τη μέθοδο Varimax πληρούσαν τα ακόλουθα κριτήρια:

1. Βασίστηκε σε παράγοντες με ιδιοτιμή > 1.0
2. Κάθε μεμονωμένο στοιχείο συσχετίστηκε με τον σχετικό παράγοντα σε επίπεδο 0,40 ή παραπάνω
3. Κάθε στοιχείο που περιείχε κάθε παράγοντας δεν είχε σημαντική συσχέτιση με άλλο παράγοντα και
4. Επιλέχθηκαν μόνο στοιχεία με communality > 0,50

1. Βασικά κίνητρα των νέων οδηγών για χρήση της ζώνης ασφαλείας

	Factor						
	1	2	3	4	5	6	7
Environment							
Bad weather conditions	0.71						
Unknown area	0.86						
Bad road surface	0.83						
Narrow roads	0.78						
Heavy traffic	0.70						
Imitation							
Set an example to others		0.65					
Education		0.70					
Imitate his/her family		0.65					
Imitate his/her friends		0.69					
To wipe off the hesitations of the co-driver		0.66					
Confidence/consistency		0.45					
Self-protection							
Avoidance of injury			0.77				
Feeling of security			0.72				
Driver's stabilisation			0.56				
Protection in high speed			0.70				
Avoidance of fatal accident			0.72				
Fear							
Lack of trust in the driver				0.48			
Fear in general				0.41			
Fear due to inexperience				0.81			
Feeling less stressed				0.68			
Experience							
Personal accident in the past					0.82		
Accident as co-driver					0.60		
Accident of a relative/friend					0.73		
Witness of an accident					0.63		
Financial issues							
Loss of working hours						0.72	
Being uninsured						0.80	
His/her car plates will be taken away						0.69	
Legality							
Avoidance of law penalties							0.68
Compliance with traffic regulations							0.81
Compliance with state rules							0.80

Παρατηρούμε ότι προέκυψαν 7 παράγοντες για τον προσδιορισμό των κινήτρων για χρήση SB, οι οποίοι αντιστοιχούσαν στο 64% της συνολικής διακύμανσης. Οι μεταβλητές που περιέχει κάθε ένας από τους επτά παράγοντες έχουν προκύψει σχετικά με τα σημαντικά φορτία.

Ο πρώτος παράγοντας αντιπροσώπευε το 29% της διακύμανσης αποτελούμενο από πέντε στοιχεία που σχετίζονται με περιβαλλοντικές συνθήκες, οπότε χαρακτηρίστηκε ως κίνητρο που επικαλούσε θέματα «περιβάλλοντος» (π.χ. «κακές καιρικές συνθήκες», «άγνωστη περιοχή»), όπως φαίνονται και στον παραπάνω πίνακα. Ο δεύτερος παράγοντας που αντιστοιχούσε στο 8% της διακύμανσης ονομάστηκε «μίμηση» επειδή φαίνεται να αντικατοπτρίζει τη μίμηση ενός θέματος (π.χ. «παράδειγμα για την οικογένεια», «μίμηση της οικογένειας», «μίμηση κάποιου φίλου»).

Ο τρίτος (6,2%) και ο τέταρτος (6%) παράγοντες έλαβαν σημαντικά φορτία σε αντικείμενα που σχετίζονται αντίστοιχα με την «αυτοπροστασία» και τον «φόβο» (π.χ. «αποφυγή τραυματισμού», «αίσθημα ασφάλειας», «έλλειψη εμπιστοσύνης στον οδηγό «,« φόβος λόγω απειρίας »,« αίσθημα λιγότερο άγχους »). Ο πέμπτος παράγοντας (5,6%) περιλάμβανε τέσσερα στοιχεία που σχετίζονται με την «εμπειρία» (π.χ. «προσωπικό ατύχημα στο παρελθόν», «ατύχημα ως συν-οδηγός»), ενώ οι επόμενοι δύο παράγοντες περιελάμβαναν στοιχεία που σχετίζονται με «οικονομικά ζητήματα» (5%) και «νομιμότητα» (4,2%).

2. Βασικά κίνητρα των νέων οδηγών για μη - χρήση της ζώνης ασφαλείας

	Factor			
	1	2	3	4
Risky behaviour				
Not always compliant with the regulations	0.50			
Going against the mainstream	0.55			
Not being the well-behaved type	0.57			
Risky personality	0.62			
Being a man of action	0.63			
Not being afraid of death	0.66			
Losing prestige	0.66			
Incompatible with high speed	0.65			
Incompatible with a smart behaviour	0.71			
Free to jump in case of accident	0.48			
Belt cannot protect	0.46			
Does not apply in my case	0.44			
Discomfort				
Claustrophobia		0.75		
It's tiring		0.80		
Restricts movements		0.78		
Feeling pressure		0.79		
Tiresome experience		0.68		
Potential risk of being trapped		0.40		
Not being accustomed to wearing		0.77		
Negligence		0.65		
Underestimation of danger				
Driving slowly			0.46	
Having a safe car			0.55	
Being a good driver			0.47	
Availability of airbags			0.48	
Sitting in the back seat			0.60	
Being a co-driver			0.62	
Belt is useless			0.56	
Wrinkles and dirt on clothes			0.50	
Keeps somebody warm			0.41	
Cannot happen to me			0.48	
Waste of time				
Frequent stops				0.74
Being in a hurry				0.71
Driving short distances				0.52
Dislike of delays				0.49
Unconventional for one's job				0.42

Όσον αφορά τον προσδιορισμό των κινήτρων για μη χρήση SB, 4 παράγοντες εμφανίστηκαν με ιδιοτιμές > 1. Οι παράγοντες που εξήχθησαν,συλλογικά, αντιπροσώπευαν το 49,3% της κοινής διακύμανσης. Ο πρώτος παράγοντας αντιπροσώπευε το 30% της συνολικής διακύμανσης που περιέχει δώδεκα στοιχεία που υποδηλώνουν ένα κίνητρο «επικίνδυνης συμπεριφοράς» (π.χ. «δεν συμμορφώνεται πάντα με τους κανονισμούς»).

Ο δεύτερος παράγοντας αντιπροσώπευε το 7,8% της διακύμανσης που περιείχε οκτώ αντικείμενα. Αυτά τα οκτώ αντικείμενα αφορούσαν «δυσφορία» (π.χ. «η ζώνη ασφαλείας περιορίζει τις κινήσεις»). Ο τρίτος παράγοντας (6,4%) έλαβε σημαντικά φορτία σε δέκα αντικείμενα σχετικά με την «υποτίμηση του κινδύνου» (π.χ. «έχοντας ασφαλές αυτοκίνητο», «διαθεσιμότητα αερόσακων»).

Ο τέταρτος παράγοντας (5,3%) περιείχε πέντε στοιχεία που αναγνωρίστηκαν ως σχετικά με το «χάσιμο χρόνου» (π.χ. «συχνές στάσεις», «βιαστικά»).

Τέλος, αναπτύχθηκε ένα μοντέλο πολλαπλής παλινδρόμησης για να εκτιμηθεί η σχέση μεταξύ κάποιων δημογραφικών μεταβλητών και των παραγόντων που βρέθηκαν με την εξαρτημένη μεταβλητή (αθροιστικό αποτέλεσμα των ποσοστών χρήσης της ζώνης ασφαλείας), μια συνεχής μεταβλητή. Κάποιοι παράγοντες βρέθηκαν θετικά και άλλοι αρνητικά συσχετισμένοι με τη μεταβλητή απόκρισης.

4.4.2 Η αντίληψη των πελατών απέναντι στην ποιότητα υπηρεσιών της εταιρείας ασφάλισης ζωής στην Ινδία

Στόχος

Η παρούσα μελέτη στοχεύει στη μέτρηση της αντίληψης των πελατών για την ποιότητα των υπηρεσιών ασφάλισης ζωής. Η μελέτη διεξήχθη σε πελάτες της Life Insurance Corporations που βρίσκονται στις μεγάλες πόλεις, δηλαδή, Amritsar, Jalandhar και Ludhiana, στο Punjab, μια προοδευτική πολιτεία της Ινδίας.

Στοιχεία Δείγματος

Το τελικό σύνολο που συμπλήρωσαν σωστά τα ερωτηματολόγια από κάθε άποψη είναι 337. Το δείγμα αποτελείται από μια σημαντική υπεροχή (72,1%) ανδρών ερωτηθέντων έναντι γυναικών (27,9%) ερωτηθέντων. Οι ερωτηθέντες κυμαίνονται κυρίως μεταξύ των ηλικιών 21 έως 40 ετών (57,3%) και 41 έως 60 (38,6%). Όσον αφορά την οικογενειακή κατάσταση, μια σημαντική πλειοψηφία των ερωτηθέντων (82,5%) είναι παντρεμένοι, ενώ το 17,5% των ερωτηθέντων είναι άγαμοι.

Η πλειοψηφία (94,4%) των ερωτηθέντων ανήκει σε αστικές περιοχές, ενώ μόνο 5,6% κατοικούν σε αγροτικές περιοχές. Οι περισσότεροι από τους ερωτηθέντες (43,3%) προέρχονται από το Amritsar, ακολουθούμενο από τον Jalandhar (33,8%) και τη Ludhiana (22,9%). Όσον αφορά τα ακαδημαϊκά προσόντα, η πλειοψηφία (39,8%) των ερωτηθέντων είναι απόφοιτοι ακολουθούμενη από μεταπτυχιακούς (32,3%), επαγγελματίες (19,3%), πτυχίο δευτεροβάθμιας εκπαίδευσης (6,5%) και πτυχιούχους (1,2%) .

Επίσης το 46,3% των ερωτηθέντων εμπίπτουν στο εύρος εισοδήματος των από 15.001 έως 30.000 ακολουθούμενοι από εκείνους (19,9%) που παίρνουν μεταξύ 30.001 έως 45.000. Ωστόσο, το 18,1% των ερωτηθέντων ανήκουν στην ομάδα εισοδήματος 15.000 ή περίπου το 15,7% ξεπερνούν τα 45.000. Όσον αφορά τον τρόπο πληρωμής των ασφαλίστρων, η πλειοψηφία (46,3%) των ερωτηθέντων προτιμούν την ετήσια πληρωμή, το 23,1% προτιμά το εξάμηνο, ακολουθούμενο από το 17,5% ανά τρίμηνο και το 7,1% μηνιαίως.

Η τελική κλίμακα περιλάμβανε 42 στοιχεία. Προκειμένου ναι είναι πιο εύκολη η ερμηνεία των αποτελεσμάτων, η κλίμακα 42 στοιχείων αναλύθηκε με παραγοντική ανάλυση χρησιμοποιώντας περιστροφή με τη μέθοδο Varimax. Ωστόσο, πριν από την εφαρμογή της ανάλυσης παραγόντων, τα δεδομένα ελέγχθηκαν για την καταλληλότητά τους. Οι έλεγχοι που εφαρμόστηκαν για την καταλληλότητα των δεδομένων για ανάλυση παραγόντων ήταν ο Kaiser-Meyer-Oklin (KMO) Measure of Sampling Adequacy (MSA) και Bartlett's test Sphricity.

Παραγοντική Ανάλυση

Η τιμή του KMO βρέθηκε 0,918 και ο έλεγχος του Bartlett ήταν σημαντικός (p_{value} <0,001). Τα αποτελέσματα έδειξαν έτσι ότι το δείγμα που ελήφθη ήταν κατάλληλο για να προχωρήσει σε μια διαδικασία ανάλυσης παραγόντων. Εκτός από το Bartlett's Test και το KMO Measure, παρατηρήθηκαν επίσης τα communalities όλων των μεταβλητών. Οι τιμές όλων των communalities όλων των μεταβλητών ήταν αρκετά πάνω από 0,50 εκτός από μια μεταβλητή, η οποία αφαιρέθηκε.

Περαιτέρω, για τον καθορισμό των παραγόντων με σαφήνεια, χρησιμοποιήθηκαν κριτήρια:

1. Να διαγραφεί οποιαδήποτε μεταβλητή με φορτία κάτω από ± 0,50
2. Ένας παράγοντας πρέπει να αναγνωρίζεται από τουλάχιστον δύο μεταβλητές.
3. Ιδιοτιμές > 1

Μετά από κάθε ανάλυση, στοιχεία που δεν πληρούσαν τα κριτήρια διαγράφηκαν από την ανάλυση. Τελικά, η λύση του τελικού, η οποία πληρούσε τα κριτήρια, περιλάμβανε 34 μεταβλητές.

Factors	Loading	Eigen Value	Percentage of Variance	Cronbach Alpha
F1 Proficiency		14.893	37.233	0.9143
Willingness to help customers and the readiness to respond to customers' requests	0.767			
Giving caring and individual attention to customers by having the customers' best interests at heart	0.763			
Agents and employees who instill confidence in customers by proper behaviour	0.756			
Agents and employees who understand the specific needs of their customers	0.754			
Apprising the customers of the nature and schedule of services available in the organization	0.751			
Providing prompt service to customers	0.707			
Agents and employees who have the proper knowledge and competence to answer customers' specific queries and requests	0.651			
Effective customers' grievance redressal procedures and processes	0.600			
F2 Media and Presentations		3.435	8.588	0.8508
Attractive and informative media, theme layout, and language of the advertisement	0.753			
Visually appealing materials and facilities associated with the service	0.732			
Easy to get information about insurance policies through T.V., newspaper, Internet etc. rather than agents	0.722			
Staff appeared neat and professional	0.638			
Modern looking updated equipment, fixtures, and facilities	0.611			
F3 Physical and Ethical Excellence		2.553	6.383	0.8714
Provides proper drinking water and sanitary facilities	0.710			
Branch layout has been designed to give more space to the customers to transact business	0.702			
Providing visually appealing signs, symbols, advertisement boards, pamphlets and other artifacts in the branch offices	0.691			
Comfortable physical layout of premises, furnishings, and ambient conditions (e.g. temperature, ventilation, noise, odor) for the customers to interact with official staff	0.635			
Promotes ethical conduct in everything it does	0.615			
High rate of return on insurance products as compared to the other saving instruments (fixed deposit in banks, national saving certificates etc.)	0.516			
F4 Service Delivery Process and Purpose		1.787	4.469	0.8638
Adequate and necessary personnel/agents for good customer services	0.677			
Timely revival of lapsed policies, change of nominations, addresses and mode of premium payment etc.	0.660			
Speedy documentation and processes from the time of issue of policies up to the settlement of claims (e.g. premium and default notices etc.)	0.626			
Number of regular meetings with agents, discussion on each and every aspect of the policy, analysis of various tax aspects etc. in order to buy life insurance policy	0.606			
Performing services right the first time	0.600			
Ability of agents to give truthful advice on investments /tax benefits etc.	0.535			
F5 Security and Dynamic Operations		1.435	3.587	0.7711
Convenient to pay premium on due date	0.739			
Flexible products/ new products that meet customers' needs	0.652			
Making customers feel safe and secure in their transactions	0.638			
Enhancement of technological capability (e.g. computerization, networking of operation, etc.) to serve customers more effectively	0.559			
F6 Credibility		1.273	3.183	0.7309
Adequate and necessary facilities for good customer services	0.656			
Wide use of modern and alternate mode of premium payment, such as electronic clearing system, payment through Internet etc.	0.598			
Appropriate behaviour of the concerned staff	0.504			
F7 Functionality		1.193	2.982	0.4814
Convenient location of the branch offices	0.719			
Availability of top officials in case of need	0.507			

Note: Factor loadings below 0.50 are not shown in this Table.

ΠΙΝΑΚΑΣ 1

Στον παραπάνω πίνακα εμφανίζονται οι μεταβλητές που αποτελούνται οι 7 παράγοντες και παρέχει πληροφορίες για τα φορτία, τις ιδιοτιμές, τις τιμές Cronbach alpha και το ποσοστό της διακύμανση που εξηγείται από κάθε παράγοντα. Η λύση των 7 παραγόντων εξηγεί το 66,42% της συνολικής διακύμανσης.

Όλοι οι παράγοντες ονομάστηκαν με βάση το περιεχόμενο των μεταβλητών που αποτελούσαν τον καθέναν. Οι παράγοντες που δημιουργήθηκαν έτσι είχαν ιδοτιμές μεταξύ 1,193 έως 14,893. Οι τιμές των communalities κυμαίνονταν από 0,549 έως 0,801 για διάφορες δηλώσεις. Αυτό σήμαινε ότι η ανάλυση παραγόντων εξήγαγε μια καλή διακύμανση στις δηλώσεις. Η λύση των 7 παραγόντων θα μπορούσε να προταθεί για τον κλάδο ασφάλισης ζωής για τη μέτρηση της ποιότητας των υπηρεσιών.

Τέλος πραγματοποιήθηκε ένα μοντέλο γραμμικής παλινδρόμησης το οποίο θεώρησε τους επτά παράγοντες ως ανεξάρτητες μεταβλητές και τη συνολική ποιότητα υπηρεσιών ως εξαρτημένη μεταβλητή. Οι συμμετέχοντες στην ινδική αγορά ασφάλισης ζωής πρέπει να προσδιορίσουν το βάρος κάθε παράγοντα που επηρεάζει την αντίληψη του πελάτη για την ποιότητα των υπηρεσιών ασφάλισης ζωής και με βάση τη συνάφεια καθενός από αυτούς τους παράγοντες, ο κλάδος ασφάλισης ζωής μπορεί να προτείνει κατάλληλα σχέδια δράσης για να ανταποκριθούν στις προσδοκίες των πελατών.

4.5 Λογιστική Παλινδρόμηση

4.5.1 Ανάλυση των καθοριστικών παραγόντων για αποχώρηση πελατών VIP σε εταιρεία τηλεπικοινωνιών

Πληροφορίες

Η μείωση της αποχώρησης πελατών(churn) είναι μια από τις μεγαλύτερες ανησυχίες για τις παγκόσμιες εταιρείες τηλεπικοινωνιακών. Σε πολλές από τις παγκόσμιες εταιρείες παροχής υπηρεσιών κινητής τηλεφωνίας, ο ετήσιος ρυθμός μετακίνησης πελατών κυμαίνεται από 20% έως 40% (Berson et al., 1999; Madden et al., 1999 , Kim and Jeong, 2004 , Parks Associates, 2003). Η μετακίνηση-αποχώρηση πελατών επηρεάζει σοβαρά αυτές τις εταιρείες επειδή τείνουν να χάνουν πολλά ασφάλιστρα, μειώνοντας τα επίπεδα κέρδους.

Επιπλέον, το κόστος διατήρησης ενός υπάρχοντος πελάτη είναι χαμηλότερο από την απόκτηση ενός νέου πελάτη (Siber, 1997). Σε σύγκριση με το συνηθισμένο κύκλο πελατών, ο VIP πελάτης προσελκύει περισσότερη προσοχή των εταιρειών τηλεπικοινωνιών επειδή οι πελάτες VIP φέρνουν περισσότερη χρηματική ροή και κέρδη από τον απλό πελάτη. Οι VIP πελάτες αντιπροσωπεύουν το 20% του συνόλου των πελατών, οι οποίοι αποφέρουν το 80% των συνολικών κερδών. Από μια εσωτερική αναφορά του μεγαλύτερου φορέα κινητής τηλεφωνίας στην Κίνα: China Mobile Communication Corporation (CMCC), το κορυφαίο 20% των πελατών συνεισφέρει το 50% των εσόδων (Huang, 2010). Ο Zhu (2008) ανέλυσε πελάτες VIP της κινητής επικοινωνίας της Κίνας. Διαπίστωσε VIP πελάτες με υψηλές πληρωμές, μεγαλύτερη διάρκεια και υψηλή πίστωση. Έτσι, η μείωση του VIP πελάτη είναι πολύ σημαντική για την επιβίωση και την ανάπτυξη της επιχείρησης.

Προκειμένου να διαχειριστούν καλύτερα την πρόθεση αποχώρησης των πελατών VIP, οι εταιρείες πρέπει να κατανοήσουν πλήρως τους παράγοντες που σχετίζονται με αυτή. Ωστόσο, το πρόβλημα δεν έχει αντιμετωπιστεί πλήρως στις υπάρχουσες αναλύσεις για τους ακόλουθους λόγους: οι προηγούμενες μελέτες επικεντρώθηκαν κυρίως στο συνηθισμένο κύκλο πελατών. Ωστόσο, υπάρχουν οι διαφορές της συμπεριφοράς μεταξύ VIP πελάτη και απλού πελάτη. Με τη διερεύνηση εμπειρογνωμόνων στον τομέα των τηλεπικοινωνιών, υπάρχουν τρεις διαστάσεις διαφορών μεταξύ VIP και απλών πελατών: το διαφορετικό μοτίβο απόφασης αγοράς, το διαφορετικό επίπεδο πληρωμής και η διαφορετική προσδοκία υπηρεσίας. Εκτός από αυτό, πολλές έρευνες σχετικά με την πρόβλεψη της μετακίνησης - αποχώρησης των πελατών διαπίστωσαν ότι οι πελάτες με υψηλή πληρωμή έχουν διαφορετικό μοτίβο συμπεριφοράς μετακίνησης-αποχώρησης. Επομένως, η μελέτη των καθοριστικών παραγόντων της κινητής τηλεφωνίας VIP πελατών έχει μεγάλη σημασία.

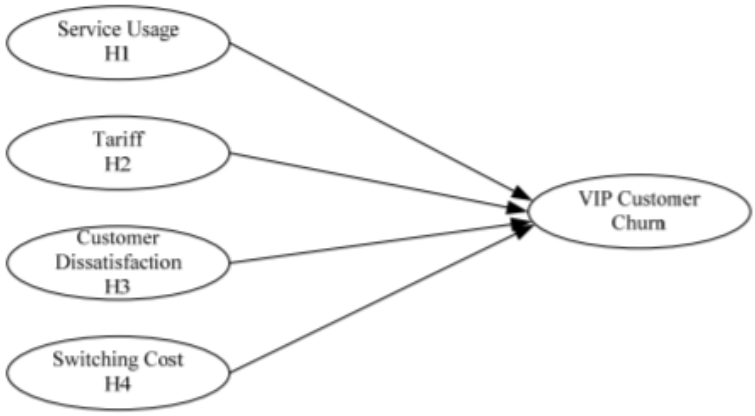
Στόχος

Σε σύγκριση με τις προηγούμενες μελέτες, αυτή η ανάλυση έχει επικεντρωθεί σε πελάτες VIP αφού φέρνουν το 80% των συνολικών κερδών και διαδραματίζουν σημαντικό ρόλο. Αναπτύσσεται ένα ολοκληρωμένο μοντέλο churn (πελατών με πρόθεση αποχώρησης) και δοκιμάζεται εμπειρικά χρησιμοποιώντας ένα μεγάλο δείγμα πραγματικών συναλλαγών πελατών και δεδομένων χρέωσης, το οποίο σχετίζεται άμεσα με τις πραγματικές αποφάσεις σχετικά με τη μετακίνηση-αποχώρηση πελατών.

Προκειμένου να αναλυθεί η διχοτομική μεταβλητή - απόφαση των πελατών κινητής τηλεφωνίας σχετικά με το αν θα φύγουν ή όχι, οι Kim και Yoon (2004) πρότειναν μια οικονομετρική μεθοδολογία η οποία είναι ένα διωνυμικό μοντέλο logit που βασίζεται στη θεωρία διακριτών επιλογών.

<u>Καθοριστικοί παράγοντες για το VIP πελάτη:</u>

Το ***ΣΧΗΜΑ 1*** παρουσιάζει τέσσερις μεγάλες κατηγορίες μεταβλητών που υποτίθεται ότι επηρεάζουν την αποχώρηση πελατών VIP



ΣΧΗΜΑ 1

<u>Μεταβλητές/Υποθέσεις</u>

- *H1: Χρήση υπηρεσίας*

Στις περισσότερες έρευνες, η χρήση υπηρεσιών συσχετίζεται αρνητικά με την πιθανότητα αποχώρησης πελατών. Επομένως, έχουμε:

H1 (α) : Ο τύπος πληρωμής σχετίζεται με την πιθανότητα αποχώρησης πελατών VIP

H1 (β) : Το εύρος επαφών σχετίζεται αρνητικά με την πιθανότητα αποχώρησης πελατών VIP

H1 (γ) : Ο αριθμός των υπηρεσιών σχετίζεται αρνητικά με την πιθανότητα αποχώρησης πελατών VIP.

- *Η2: Πάγιο-Φόρος*

Η μηνιαία χρέωση είναι ένας από τους πιο δημοφιλείς προγνωστικούς παράγοντες συμπεριφοράς της αποχώρησης (Buckinx and Poel, 2005; Kim and Yoon, 2004; Mozer et al., 2000), η οποία συνεπάγεται ότι οι μηνιαίες χρεώσεις και τα ποσά χρήσης συνδέονται με την πρόθεση αποχώρησης. Ωστόσο, δεν είναι ακόμη σαφές εάν η σχέση μεταξύ μηνιαίας χρέωσης και αναχώρηση πελάτη είναι πραγματικά θετική ή αρνητική. Για την ακριβή μέτρηση της μηνιαίας χρέωσης, συμπεριλαμβάνουμε τη μηνιαία χρέωση και τη μηνιαία χρέωση δεδομένων υπηρεσιών και λαμβάνουμε την ακόλουθη υπόθεση:

Η2 (α) : Το μέσο μηνιαίο κόστος σχετίζεται θετικά με την πιθανότητα αποχώρησης πελατών VIP

Η2 (β) :Το μέσο μηνιαίο κόστος των υπηρεσιών σχετίζεται θετικά με την πιθανότητα αποχώρησης VIP πελατών

- *Η3: Δυσαρέσκεια πελατών*

Η3: Τα παράπονα VIP πελατών σχετίζονται θετικά με την πιθανότητα αποχώρησης

- *Η4: Κόστος αλλαγής*

Η4 (α) : Οι συσσωρευμένοι πόντοι επιβράβευσης σχετίζονται αρνητικά με την πιθανότητα αποχώρησης πελατών.

Η4 (β) : Ένα συμβόλαιο με τερματικό όριο σχετίζεται αρνητικά με την πιθανότητα αποχώρησης πελατών VIP.

Η4 (γ) : Η διάρκεια της συνδρομής σχετίζεται αρνητικά με την πιθανότητα αποχώρησης πελατών VIP.

Η4 (δ) : Ο συνδρομητής δικτύου ομάδας VPMN σχετίζεται αρνητικά με την πιθανότητα αποχώρησης πελατών VIP.

Η4 (ε) : Το οικογενειακό πακέτο σχετίζεται αρνητικά με την πιθανότητα αποχώρησης πελατών VIP.

<u>Μεταβλητή στόχος:</u>

Η εξαρτημένη μεταβλητή είναι δυαδική όπου δηλώνει:

1: "churn" και

0: "non churn"

Στους παρακάτω πίνακες περιγράφονται οι ανεξάρτητες μεταβλητές:

Dimensions	*Variable name*	*Level*	*Description*
Demographics	Gender	1 and 2	'1' represents male and '2' represents female
	Age	Non-negative integers	Subscriber's age, in years
Service usage	Data service amounts	Non-negative integers	The amounts of data service used by subscriber
	Contact range	Non-negative integers	The number of the distinct individuals that the subscriber contact
	Payment type	0 and 1	'0' represents the subscriber's payment type is post-payment and '1' represents prepayment

ΠΙΝΑΚΑΣ 1

Dimensions	*Variable name*	*Level*	*Description*
Tariff	Monthly average cost	Non-negative	Average monthly average costs for mobile service
	Monthly average data services costs	Non-negative	Average monthly average data services costs
Customer satisfaction	VIP customer complaints	0 and 1	'1' represents this VIP customer makes complaints to customer service centre in 3 months and '0' represents no complaints in this period
Switching costs	Accumulated loyalty points	Non-negative integers	accumulated loyalty points provided by mobile service provider for incentives of using service
	Terminal bounded contract	0 and 1	'1' represents terminal bounded contract without expired and '0' represents a bounded contract with expired
	Duration of subscription	Non-negative	Duration of subscription with the present carrier (years). '0' means the tenure is between 0.3 and 1 year
	Group VPMN network subscriber	0 and 1	'1'represents that the subscriber is one of group V network subscribers and '0' represents not belonging to any group
	The subscriber with family package	0 and 1	'1'represents that the subscriber is using family package and '0'represents that the subscriber isn't using family package

ΠΙΝΑΚΑΣ 2

Για την ανάλυση, τα δεδομένα λαμβάνονται από τη βάση δεδομένων ενός επαρχιακού υποκαταστήματος CMCC με 13 εκατομμύρια συνδρομητές. Ο αριθμός των VIP πελατών είναι περίπου 5,64% της συνολικής πελατειακής βάσης. Οι λογαριασμοί ήταν ενεργοί από τον Αύγουστο του 2009 έως τον Οκτώβριο του 2009, που σημαίνει ότι ήταν ενεργοί τουλάχιστον τρεις μήνες για να συμπεριληφθούν στο δείγμα. Για τους σκοπούς της ανάλυσης, το «churn» ορίστηκε ως το συμβάν στο οποίο τερματίστηκε μια συνδρομή μέχρι το τέλος Νοεμβρίου 2009.

Χρησιμοποιείται η μέθοδος likelihood ratio για να ελεγχθεί η σημαντικότητα της λογιστικής παλινδρόμησης.

Τα likelihood statistics ακολουθούν περίπου την κατανομή X^2 και εάν το p_{value} του ελέγχου είναι < του α , είναι σημαντικά, άρα μπορούμε να απορρίψουμε την υπόθεση
(H0: β1 = β2 =… = βm = 0) και να θεωρήσουμε ότι οι πληροφορίες που λαμβάνονται από τις ανεξάρτητες μεταβλητές είναι χρήσιμες για να προσδιορίσουμε την τάση αποχώρησης του VIP πελάτη. Σύμφωνα με το αποτέλεσμα στον ***ΠΙΝΑΚΑΣ 3***, μπορούμε να ανακαλύψουμε ότι το μοντέλο είναι σημαντικό.

	Chi-square	*Df*	*Sig*
Model	4,098.439	14	0.000

ΠΙΝΑΚΑΣ 3

Ο ***ΠΙΝΑΚΑΣ 4*** , ταξινόμησης εμφανίζει την ακρίβεια πρόβλεψης του μοντέλου λογιστικής παλινδρόμησης.

Classification table		*Predicted*		
		Label		*Percentage correct*
	Observed	*0*	*1*	
Label	0	4,673	1,108	62.3
	1	629	5,152	89.1
Accuracy				75.7

ΠΙΝΑΚΑΣ 4

Η προβλεπτική ακρίβεια αυτού του λογιστικού μοντέλου είναι 75,7% και για την κλάση μηδέν (non-churn), αυτή η τιμή είναι 62,3% και για την κατηγορία 1(churn), αυτή η τιμή είναι 89,1%. Η ικανότητα πρόβλεψης για το churn είναι καλύτερη από ό, τι για τον non-churn. Η ακρίβεια των προβλέψεών του μοντέλου είναι καλή. Το να βρεθεί σωστά ένας churner είναι πιο σημαντικό από το να βρεθεί ένας churner ενώ δεν είναι. Έτσι, το μοντέλο μας είναι αποτελεσματικό για την πρόβλεψη αποχώρησης του VIP πελάτη.

Αποτελέσματα Μοντέλου:

Independent variable		*Coefficient*	*Significance level*	*Exp (coefficient)*	*Wals*
Gender (1)	(X_1)	–0.020	0.708	0.980	0.140
Age	(X_2)	–0.012	0.000***	0.988	45.994
Payment type (0)	(X_3)	0.244	0.000***	1.276	19.736
Data service amounts	(X_4)	–0.077	0.000***	0.926	20.047
Monthly average costs	(X_5)	0.002	0.000***	1.002	80.529
Monthly average data costs	(X_6)	0.002	0.194	1.002	1.686
Complaints (0)	(X_7)	–0.045	0.729	0.956	0.120
Loyalty points	(X_8)	0.000	0.747	1.000	0.104
Terminal bounded contract (0)	(X_9)	–0.022	0.722	0.978	0.126
Duration of subscription	(X_{10})	–0.090	0.000***	0.914	441.454
Contact range	(X_{11})	–0.011	0.000***	0.989	947.899
Group VPMN network (0)	(X_{12})	0.315	0.000***	1.370	70.920
The subscriber with family package (0)	(X_{13})	0.208	0.000***	1.231	13.374

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

ΠΙΝΑΚΑΣ 5

Dependent variable	*Independent variable*	*Hypothesis*	*Result*
VIP customer churn	Payment type	H1(a)	Accepted
	Contact range	H1(b)	Accepted
	Data service amounts	H1(c)	Accepted
	Monthly average costs	H2(a)	Accepted
	Monthly average data-costs	H2(b)	Rejected
	VIP customer complaints	H3	Rejected
	Accumulated loyalty points	H4(a)	Rejected
	Terminal bounded contract	H4(b)	Rejected
	Duration of subscription	H4(c)	Accepted
	Group VPMN network subscriber	H4(d)	Accepted
	The subscriber with family package	H4(e)	Accepted

ΠΙΝΑΚΑΣ 6

Η υπόθεση 1 (α) (H1(a)) δείχνει ότι οι post-payment έχουν σημαντικό θετικό αντίκτυπο στην πιθανότητα αποχώρησης, πράγμα που σημαίνει ότι ο συνδρομητής τείνει να αποχωρεί εάν αυτός ή αυτή είναι συνδρομητής post-payment. Δεδομένου ότι οι post-payment είναι όφελος μόνο για πελάτες VIP, το αποτέλεσμα υπενθυμίζει στους διαχειριστές μάρκετινγκ ότι θα πρέπει να βρουν νέο τρόπο να ωθήσουν τους πελάτες VIP να υιοθετήσουν τρόπο prepayment.

Επιπλέον παρατηρείται ότι ο αριθμός των υπηρεσιών έχει αρνητική επίδραση στην πιθανότητα αποχώρησης. Με τον αυξανόμενο αριθμό υπηρεσιών δεδομένων, μειώνεται η πιθανότητα αποχώρησης. Τέλος, όταν εξαπλώνεται το εύρος επαφών των πελατών, τείνουν να παραμένουν με τον πάροχο υπηρεσιών, επομένως η οικοδόμηση μιας καλής επικοινωνίας πελατών VIP είναι αποτελεσματική για τη διατήρησή τους.

Η υπόθεσης 2 (α) δείχνει ότι το μηνιαίο μέσο κόστος φαίνεται να σχετίζεται θετικά με την πιθανότητα αποχώρησης. Ωστόσο, ο συντελεστής είναι 0,002 υποδηλώνει ότι το μηνιαίο μέσο κόστος έχει πολύ μικρή επίδραση στην αποχώρηση. Σύμφωνα με την υπόθεση 2 (b), το μέσο μηνιαίο κόστος δεδομένων δεν σχετίζεται σημαντικά με την πιθανότητα εμφάνισης αποχώρησης(p_{value}=0.194 > α =0,05) . Έτσι, δεν επηρεάζουν σημαντικά την αποχώρηση του πελάτη VIP.

Το τεστ της υπόθεσης 3 αποκαλύπτει ότι οι πελάτες VIP με παράπονα δεν σχετίζονται σημαντικά με την πιθανότητα αποχώρησης (p_{value}=0.729 > α =0,05. Αυτό ίσως υποδηλώνει ότι οι εταιρείες κάνουν καλά προγράμματα για την αντιμετώπιση των καταγγελιών πελατών VIP.

Η δοκιμή της υπόθεσης 4 (α) και της υπόθεσης 4 (β) δείχνει ότι ούτε οι συσσωρευμένοι πόντοι επιβράβευσης ούτε ένα συμβόλαιο με τερματικό όριο σχετίζεται σημαντικά με την πιθανότητα αποχώρησης.

Όσον αφορά την υπόθεση 4 (α) και 4(β) φαίνεται ότι ούτε οι συσσωρευμένοι πόντοι επιβράβευσης ούτε ένα συμβόλαιο με τερματικό όριο σχετίζεται σημαντικά με την πιθανότητα αποχώρησης, (p_{value}=0,747 ,p_{value}=0,722, αντίστοιχα).

Το τεστ των υποθέσεων 4(γ), 4(δ) και 4(ε) δείχνουν ότι η διάρκεια της συνδρομής, οι συνδρομητές της ομάδας VPMN και οι συνδρομητές με οικογενειακά πακέτα σχετίζονται αρνητικά με την πιθανότητα αποχώρησης. Η ομάδα δικτύου VPMN είναι ένα εικονικό δίκτυο για επιχειρηματική μονάδα. Οι χρήστες που έχουν εγγραφεί στο δίκτυο έχουν κάποια πλεονεκτήματα, π.χ. δωρεάν λεπτά επικοινωνίας μεταξύ οποιουδήποτε ατόμου στο ίδιο δίκτυο. Εάν αποχωρήσει κάποιος από το δίκτυο VPMN, όχι μόνο αυτός αλλά και όλοι σε αυτό το δίκτυο VPMN πρέπει να πληρώσουν περισσότερα όταν επικοινωνεί μαζί τους, αυτό αυξάνει σε μεγάλο βαθμό το κόστος, καθιστώντας έτσι το κόστος αλλαγής πολύ υψηλό. Το οικογενειακό πακέτο λειτουργεί ακριβώς όπως το δίκτυο ομάδας VPMN, αλλά μόνο για τους συνδρομητές και τα μέλη της οικογένειάς τους.

<u>Συμπέρασμα:</u>

Οι πιο σημαντικές συνεχείς μεταβλητές είναι «Duration of subscription» (διάρκεια συνδρομής) και «Data services amount» (ποσά υπηρεσιών δεδομένων), των οποίων η τιμή β (στήλη coefficient) είναι –0,09 και –0,077 αντίστοιχα. Αυτό σημαίνει ότι εάν τα ποσά των υπηρεσιών δεδομένων ενός πελάτη VIP αυξάνονται κατά 1, τότε η πιθανότητα αποχώρησης μειώνεται κατά 0,926 φορές ($e^{-0,077}$=0,926). Το «contact range» (εύρος επαφών) και το «monthly average costs» (μηνιαίο μέσο κόστος) είναι επίσης σημαντικά για τον προσδιορισμό της εξέλιξης αποχώρησης των VIP πελατών. Ωστόσο, οι επιπτώσεις τους είναι πολύ αδύναμες. Οι πιο σημαντικοί καθοριστικοί παράγοντες είναι «Group V network subscriber». Εάν ένας πελάτης VIP δεν ενταχθεί σε δίκτυο V, η πιθανότητα αποχώρησης του θα αυξηθεί 1,37 φορές. Ομοίως, ένας VIP χωρίς οικογενειακό πακέτο θα αυξήσει την πιθανότητα αποχώρησης του (churn) 1.231 φορές, και ένας πελάτης post-paid θα αυξήσει την πιθανότητα αποχώρησης του 1.276 φορές.

Συμπερασματικά λοιπόν τα αποτελέσματα δείχνουν ότι το κόστος αλλαγής έχει σημαντικά αρνητική επίδραση στην αποχώρηση πελατών VIP. Το ομαδικό δίκτυο VPMN και το οικογενειακό πακέτο είναι ο πιο αποτελεσματικός τρόπος διατήρησης VIP πελατών.

4.5.2 Έρευνα σχετικά με την πρόβλεψη εκκένωσης έκτακτης ανάγκης

Στόχος

Αυτό το μοντέλο θέλει να προβλέψει την πιθανότητα όλων των τρόπων εκκένωσης των κατοίκων, ώστε να αποκτήσουν το καλύτερο σχέδιο διαχείρισης της εκκένωσης έκτακτης ανάγκης σε σεισμούς.

Μεταβλητές

Η εξαρτημένη μεταβλητή είναι εκκένωση (η τιμή είναι 1) ή μη-εκκένωση (η τιμή είναι 0). Τα παρακάτω στοιχεία επηρεάζουν την εξαρτημένη μεταβλητή:

• Τύπος κτιρίου (βίλα ή διαμέρισμα) **(X_2)**
• Ιδιοκτησία κατοικίας (ιδιοκτησία ή όχι) **(X_3)**
• Εμπειρία έκτακτης ανάγκης του κατοίκου (εμπειρία έκτακτης ανάγκης και σοβαρή απώλεια, εμπειρία έκτακτης ανάγκης και μικρή απώλεια, ή όχι). **(X_4)**
• Υποχρεωτική εντολή εκκένωσης (έχει ληφθεί ή όχι). **(X_5)**
• Απόσταση από πληγείσα περιοχή (στο κέντρο της περιοχής που πλήττεται από καταστροφές, σε περιοχή που επηρεάστηκε από καταστροφές ή σε ασφαλή περιοχή) **(X_6)**
•Ζει μόνος **(X_7)**
•Κατάσταση γάμου κατοίκου που συμμετείχε στην εκκένωση (άγαμος και χωρίς ιστορικό γάμου ή παντρεμένος και συγκατοίκηση). **(X_8)**
• Ηλικία κατοίκου που συμμετείχε στην εκκένωση . **(X_9)**
• Εκπαιδευτικό επίπεδο κατοίκων που συμμετείχαν στην εκκένωση (επίπεδο δευτεροβάθμιας εκπαίδευσης ή κάτω από αυτό, επίπεδο δευτεροβάθμιας εκπαίδευσης, προπτυχιακό επίπεδο, μεταπτυχιακό επίπεδο ή πάνω από αυτό). **(X_{10})**

Λογιστική Παλινδρόμηση

Χρησιμοποιείται μέθοδος Stepwise για την επιλογή ανεξάρτητων μεταβλητών:

			Score	*df*	*Sig.*
step 0	variables	x2	6.781	1	0.009
		x3	0.170	1	0.680
		x4	0.166	1	0.683
		x5	5.345	1	0.021
		x6	15.941	1	0.000
		x7	65.988	1	0.000
		x8	70.940	1	0.000
		x9	11.880	4	0.018
		x91	0.647	1	0.421
		x92	3.079	1	0.079
		x93	11.562	1	0.001
		x94	2.342	1	0.126
		x10	12.936	3	0.005
		x101	0.246	1	0.620
		x102	9.006	1	0.003
		x103	12.852	1	0.000
	Overall Statistics		118.703	14	0.000

ΠΙΝΑΚΑΣ 1

Στον ***ΠΙΝΑΚΑΣ 1*** εμφανίζεται το step 0 της μεθόδου Stepwise που εμφανίζονται όλες οι μεταβλητές. Να αναφέρουμε ότι οι X_{91}, X_{92}, X_{93}, X_{94}, X_{101}, X_{102}, X_{103} είναι dummy μεταβλητές των μεταβλητών X_9 και X_{10}.

		B	*S.E.*	*Wald*	*df*	*Sig.*	*Exp(B)*
step 6[f]	x2	1.365	0.658	4.297	1	0.038	3.916
	x4	1.021	0.510	4.007	1	0.045	2.777
	x5	1.859	0.477	15.160	1	0.000	6.415
	x6	2.582	0.538	23.002	1	0.000	13.225
	x8	4.517	0.607	55.453	1	0.000	0.011
	x91	0.080	1.380	0.003	1	0.954	0.923
	x92	2.188	0.713	9.407	1	0.002	0.112
	x93	0.221	0.610	0.132	1	0.717	0.802
	x94	0.836	0.645	1.682	1	0.195	0.433
	Constant	0.659	0.846	0.606	1	0.436	0.518

ΠΙΝΑΚΑΣ 2

Στον ***ΠΙΝΑΚΑΣ 2*** εμφανίζεται το step 6 και τελευταίο βήμα της μεθόδου Stepwise που εμφανίζονται οι τελικές σημαντικές ανεξάρτητες μεταβλητές. Βλέπουμε ότι έχουν επιλεχθεί η X_2, X_4, X_5, X_6, X_8, X_{91}, X_{92}, X_{93}, X_{94}.

Το μοντέλο είναι της μορφής:

$ln\left(\frac{p}{1-p}\right)$= 0.659 + 1.365* X_2 + 1.021* X_4 + 1.859* X_5 + 2.582* X_6 + 4.517* X_8 + 0.08* X_{91} +2.188* X_{92} + 0.221* X_{93} + 0.836* X_{94}

Παρατηρούμε από τον ***ΠΙΝΑΚΑΣ 2*** για παράδειγμα, ότι όσο πιο κοντά στην πληγείσα περιοχή (X_6) είσαι τόσο μεγαλύτερη πιθανότητα εκκένωσης, ταυτόχρονα, εάν ο ερωτώμενος είναι παντρεμένος ή συμβιώνει (X_9), η πιθανότητα εκκένωσης είναι χαμηλότερη.

Ακρίβεια μοντέλου

Η ακρίβεια της πρόβλεψης του μοντέλου είναι μια από τις μεθόδους για την αξιολόγηση ενός μοντέλου.

Observed			*Predicted*		
			Y Evacuation or not		Percentage
			0	1	Correct
step 6	*Y* Evacuation or not	0	79	27	74.5
		1	21	123	85.4
	Overall Percentage				80.8

ΠΙΝΑΚΑΣ 3

Λαμβάνοντας υπόψη τον ***ΠΙΝΑΚΑΣ 3***, όταν επάγονται 6 ανεξάρτητες μεταβλητές, η ακρίβεια της πρόβλεψης εκκένωσης θα ήταν 85,4%, την ίδια στιγμή, η ακρίβεια της πρόβλεψης για μη-εκκένωση θα ήταν 74,5%, η συνολική ακρίβεια του μοντέλου είναι 80,8%.

step	*-2 Log likelihood*	*Cox&Snell R Square*	*Nagekerke R Square*
1	265.034	0.261	0.351
2	232.360	0.352	0.473
3	214.773	0.396	0.532
4	197.113	0.437	0.587
5	193.116	0.446	0.599
6	188.918	0.455	0.612

ΠΙΝΑΚΑΣ 4

Στον παραπάνω ***ΠΙΝΑΚΑΣ 4*** υπάρχουν κριτήρια καλής προσαρμογής του μοντέλου, -2Log likelihood, Cox&Snell R^2 , Nagekerke R^2. Ένα καλό μοντέλο έχει χαμηλή τιμή , -2Log likelihood.Όπως φαίνεται στον πίνακα όταν μια μεταβλητή μπεί στο μοντέλο τιμή -2LL likelihood μειώνεται και η τιμή των Cox & Snell R^2 και Negelkerke R^2 αυξάνεται.Η καλύτερη εφαρμογή του μοντέλου φαίνεται να είναι στο step 6.

Συμπέρασμα

Άρα, καταλήγουμε στο συμπέρασμα ότι παράγοντες που επηρεάζουν την εκκένωση είναι η κατοικία, η εμπειρία έκτακτης ανάγκης, η λήψη υποχρεωτικής εντολής εκκένωσης, η απόσταση από την πληγείσα από την καταστροφή περιοχή, παντρεμένος ή συμβίωση και ηλικία. Επιπλέον, εάν οι κάτοικοι είναι πιο κοντά στην πληγείσα περιοχή, η πιθανότητα εκκένωσής τους είναι υψηλότερη και από την άλλη πλευρά, η διαμονή με άλλους όπως η οικογένεια θα οδηγούσε σε χαμηλότερη πιθανότητα εκκένωσης.

4.6 Ανάλυση συστάδων

4.6.1 Προσδιορισμός κρίσιμων παραγόντων επιτυχίας που μεγιστοποιούν την ικανοποίηση των πελατών

Η μελέτη αυτή βασίστηκε στον εντοπισμό κρίσιμων παραγόντων επιτυχίας (ΚΠΣ) που μεγιστοποιούν την ικανοποίηση των πελατών της εταιρείας Futurlab. Σύμφωνα με τον Rockart (1979), ο ορισμός των ΚΠΣ είναι σαφείς δείκτες που μπορούν να οδηγήσουν τις επιχειρήσεις στην επιτυχία. Για το σκοπό αυτό εξετάστηκαν , από συνολικά λίστα με 1.055 πελάτες Futurlab οι οποίοι ήταν τυχαίοι, από τους οποίους μόνο 225 απάντησαν στο ερωτηματολόγιο. Αυτό αντιπροσωπεύει το 21% της μελέτης πληθυσμού, 225 πελάτες της εταιρείας. Όλες οι αναλύσεις βασίστηκαν στις απαντήσεις των διευθυντών και μανατζερ εταιρειών που είναι πελάτες της Futurlab. Για να προσδιοριστούν τα ΚΠΣ, χρησιμοποιήθηκε ανάλυση παραγόντων. Αυτή η μελέτη επίσης προσπάθησε να εντοπίσει ομοιογενείς ομάδες πελατών για το οποίο χρησιμοποιήθηκε cluster ανάλυση.

Δείγμα

Table 1 - Population vs sample

Sector	Population		Sample	
	N	%	n	%
Pharmaceutical	166	16%	29	17%
Industrial Control	129	12%	17	13%
Cryopreservation	10	1%	0	-
Food Industry	140	13%	40	29%
External Laboratory	40	4%	40	100%
Education/Research	147	14%	53	36%
Others	423	40%	46	11%
Total	**1,055**	**100%**	**225**	**21%**

Source: Authors.

ΠΙΝΑΚΑΣ 1

Οι παράγοντες που κατέληξαν είναι 7 και είναι οι εξής:
Παράγοντας 1 - Στρατηγική τιμολόγησης και δωρεάν υπηρεσίες
Παράγοντας 2 - Πίστη
Παράγοντας 3 - Εικόνα
Παράγοντας 4 - Προμήθεια και απόθεμα
Παράγοντας 5 - Πληροφορίες
Παράγοντας 6 - Logistics
Παράγοντας 7 - Εικονικά κανάλια

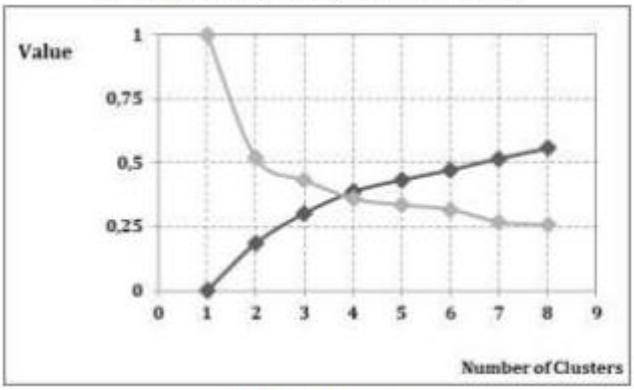
Προκειμένου να συμπληρωθεί η εμπειρική μελέτη και να εντοπιστούν ομοιογενείς ομάδες πελατών με βάση το πόση σημασία δίνεται στα ΚΠΣ της Futurlab,επιλέχτηκε να εκτελεστεί μια ανάλυση κατά συστάδες, όπου χωρίζει το αρχικό σύνολο απαντήσεων των πελατών της Futurlab σε διάφορες ομάδες.

Μέθοδος

Χρησιμοποιήθηκε ιεραρχική ανάλυση κατά συστάδες και εφαρμόστηκε χρησιμοποιώντας την Ευκλείδεια απόσταση μεταξύ των ερωτηθέντων και χρησιμοποιήθηκε η μέθοδος του μακρινότερου γείτονα (δηλαδή πλήρη σύνδεση).

Για να οριστεί ο βέλτιστος αριθμός ομάδων που θα διατηρηθεί, χρησιμοποιήθηκε το κριτήριο R^2 . Εξετάστηκαν λύσεις μεταξύ δύο και οκτώ ομάδων. Μετά την εφαρμογή των κριτηρίων, μια βέλτιστη λύση των τεσσάρων ομάδων , επιλέχθηκε και εξηγεί το 39% της συνολικής μεταβλητότητας .Όπως φαίνεται και στην παρακάτω γραφική παράσταση :

Figure 3 - Optimal number of clusters



Source: Authors.

ΣΧΗΜΑ 1

Με βάση την πραγματοποιηθείσα ομαδοποίηση, ο αριθμός των εταιρειών / πελατών που εμπίπτουν σε κάθε σύμπλεγμα φαίνεται στον παρακάτω πίνακα:

Table 5 - Number of Futurlab customers by cluster

Cluster 1	49 customers
Cluster 2	94 customers
Cluster 3	69 customers
Cluster 4	13 customers

ΠΙΝΑΚΑΣ 2

Περιγραφή Ομάδων

Ομάδα 1

Σημασία που δόθηκε στους παράγοντες «Πληροφορίες» και «πίστη». Αυτή η ομάδα αποτελείται από 49 πελάτες, οι οποίοι είναι διευθυντές και διαχειριστές. Από αυτούς, 17 είναι άνδρες και 32 γυναίκες, οι περισσότεροι είναι ηλικίας μεταξύ 26 και 35 ετών και έχουν ως επί το πλείστον διαθέτουν πανεπιστημιακό ή / και μεταπτυχιακό τίτλο. Αυτοί οι πελάτες είναι τεχνικοί εργαστηρίων και ερευνητές. Οι εταιρείες τους διδάσκουν ή / και είναι ερευνητικά και ξένα εργαστήρια τα οποία βρίσκονται στο κέντρο της χώρας, στην περιοχή της Λισαβόνας / Ταζού.

Ομάδα 2

Σημασία που δόθηκε στον παράγοντα «εικόνα». Αυτή η ομάδα αποτελείται από 94 πελάτες. Από τους διευθυντές / διαχειριστές που απάντησαν στο ερωτηματολόγιο, 34 είναι άνδρες και 60 γυναίκες, κυρίως ηλικίας μεταξύ 31 και 50 ετών, και έχουν ως επί το πλείστον πτυχία πανεπιστημίων, μεταπτυχιακών και διδακτορικών σπουδών. Πρόκειται για ένα σύνολο πελατών που είναι τεχνικοί εργαστηρίων, τεχνικοί αγορών και διευθυντές εργαστηρίων. Οι εταιρείες τους είναι οργανώσεις τροφίμων, φαρμακευτικοί, εκπαιδευτικοί ή / και ερευνητικοί οργανισμοί που βρίσκονται στο κέντρο της χώρας και στις περιφέρειες Λισσαβώνας / Ταζού και Πόρτο και Βόρεια.

Ομάδα 3

Σημασία που δόθηκε στους παράγοντες «προσφορά και το απόθεμα» και στα «εικονικά κανάλια». Αυτή η ομάδα αποτελείται από 69 πελάτες. Από αυτούς τους διευθυντές και τους διαχειριστές, 33 είναι άνδρες και 36 γυναίκες, οι περισσότεροι ηλικίας μεταξύ 31 και 50 ετών, και κυρίως πτυχιούχοι πανεπιστημίων με μεταπτυχιακά και διδακτορικά διπλώματα. Αυτό είναι ένα σύνολο πελατών που είναι τεχνικοί εργαστηρίων, εκπαιδευτικοί και τεχνικοί αγορών. Οι συνδεδεμένες/σχετιζόμενες εταιρείες είναι εξωτερικά εργαστήρια και εκπαιδευτικοί ή / και ερευνητικοί οργανισμοί που βρίσκονται στις περιφέρειες Λισσαβώνας / Ταζού και Πόρτο και Βόρεια.

Ομάδα 4

Σημασία που δόθηκε στους παράγοντες «στρατηγική τιμολόγησης και τις δωρεάν υπηρεσίες» και στην «εφοδιαστική». Αυτή η ομάδα αποτελείται από 13 πελάτες. Μεταξύ αυτών, 4 διευθυντές ή διαχειριστές είναι άνδρες και 9 γυναίκες. Οι περισσότεροι είναι ηλικίας μεταξύ 26 και 55 ετών και οι περισσότεροι έχουν μεταπτυχιακό. Αυτά είναι άτομα που είναι ερευνητές και καθηγητές. Οι εταιρείες είναι εκπαιδευτικοί ή / και ερευνητικοί οργανισμοί που βρίσκονται στην περιοχή Λισαβόνας / Ταζού.

Κεφάλαιο 5

Ανάλυση Τραπεζικών Δεδομένων

Στο κεφάλαιο αυτό θα γίνει μια ανάλυση ενός συνόλου δεδομένων χρησιμοποιώντας τεχνικές που περιγράφηκαν στο Κεφάλαιο 1. Για την ανάλυση θα χρησιμοποιηθεί το λογισμικό **SPSS MODELER**.

Στην παρακάτω ανάλυση εξετάζεται μια έρευνα εκτίμησης κινδύνου στην οποία πελάτες έχουν εκχωρηθεί σε μία από τις τρείς παρακάτω κατηγορίες:

- **good risk**
- **bad risk-profitable** (απώλεια κάποιων πληρωμών ή εμφάνιση άλλων προβλημάτων, αλλά παρέμειναν επικερδή)
- **bad risk-loss**

Αυτές οι κατηγορίες εμπεριέχονται στη μεταβλητή **risk**, η οποία είναι και η μεταβλητή στόχος.

ΠΕΔΙΟ ΣΤΟΧΟΣ	ΠΕΡΙΓΡΑΦΗ
RISK	Credit risk : 1= bad risk-loss, 2 = bad risk-profitable, 3 = good risk

Εκτός από το πεδίου ρίσκου, υπάρχει ένας αριθμός δημογραφικών μεταβλητών που θα χρησιμοποιηθούν στην ανάλυση οι οποίες είναι:

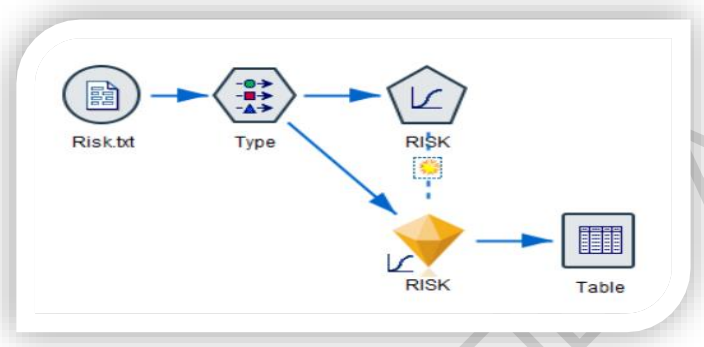
ΠΕΔΙΑ	ΠΕΡΙΓΡΑΦΗ
Age	Ηλικία σε χρόνια
Income	Εισόδημα (σε Βρετανικές λίρες(χιλιάδες))
Gender	f = γυναίκα , m = άντρας
Marital Status	single =ανύπαντρος/η, married = παντρεμένος/η, divsepwid = χωρισμένος/η
Numkids	Αριθμός παιδιών
Numcards	Αριθμός πιστωτικών καρτών
Howpaid	Χρόνος πληρωμής (εβδομαδιαία, μηνιαία)
Mortgage	Ύπαρξη στεγαστικού δανείου (y=ναι, n=όχι)
Storecar	Αριθμός καρτών για καταστήματα
Loans	Αριθμός δανείων

Στο dataset έχουμε διαθέσιμες 4117 εγγραφές

3.1 Πολυωνυμική Παλινδρόμηση

Το πεδίο πρόβλεψης είναι το credit risk (RISK), με 3 κατηγορίες όπως αναφέραμε, γι΄αυτό και θα χρησιμοποιηθεί πολυωνυμική παλινδρόμηση.

Παρακάτω παρουσιάζεται η ροή που δημιουργείται στο SPSS MODELER για να πραγματοποιηθεί η πολυωνυμική παλινδρόμηση και τα αποτελέσματα.



Αποτελέσματα:

Case Processing Summary

		N	Marginal Percentage
RISK	bad loss	906	22,0%
	bad profit	2407	58,5%
	good risk	804	19,5%
GENDER	f	2077	50,4%
	m	2040	49,6%
MARITAL	divsepwid	873	21,2%
	married	2089	50,7%
	single	1155	28,1%
HOWPAID	monthly	2026	49,2%
	weekly	2091	50,8%
MORTGAGE	n	917	22,3%
	y	3200	77,7%
Valid		4117	100,0%
Missing		0	
Total		4117	
Subpopulation		4117[a]	

ΠΙΝΑΚΑΣ 1

Στον ***ΠΙΝΑΚΑ 1*** παρατηρούμε μια σύντομη περιγραφή των έγκυρων και των ελλειπουσών τιμών των δεδομένων. Συγκεκριμένα έχουμε 4117 εγγραφές για την ανάλυση και καμία ελλείπουσα τιμή.

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	7953,278			
Final	5699,521	2253,757	20	,000

ΠΙΝΑΚΑΣ 2

Τα αποτελέσματα που παρατηρούμε στον ***ΠΙΝΑΚΑΣ 2*** αντιστοιχούν στον έλεγχο υπόθεσης:

$$\mathbf{H_0: \beta_i = 0 \text{ vs } H_1: \beta_i \neq 0}$$

Η στατιστική συνάρτηση $X^2 = 2388,908$ έχει 18 βαθμούς ελευθερίας, όσες και οι παράμετροι του μοντέλου. Κοιτώντας το $p_{value}= 0,000$, απορρίπτουμε την H_0 και συμπεραίνουμε ότι τουλάχιστον μια μεταβλητή είναι σημαντική για το μοντέλο.

Pseudo R-Square

Cox and Snell	,422
Nagelkerke	,493
McFadden	,283

ΠΙΝΑΚΑΣ 3

Στον ***ΠΙΝΑΚΑΣ 3*** εμφανίζονται οι έλεγχοι Pseudo R^2 που χρησιμοποιούνται για την αξιολόγηση του λογιστικού μοντέλου. Τιμές από 0,2 - 0,4 υποδηλώνουν καλή προσαρμογή. Επομένως, συμπεραίνουμε ότι το μοντέλο έχει καλή προσαρμογή.

Likelihood Ratio Tests

Effect	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC of Reduced Model	BIC of Reduced Model	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	3967,212	4098,192	3923,212[a]	,000	0	.
AGE	4063,472	4182,545	4023,472	100,260	2	,000
INCOME	4026,055	4145,129	3986,055	62,844	2	,000
MARITAL	4121,918	4229,085	4085,918	162,707	4	,000
NUMKIDS	3982,207	4101,281	3942,207	18,996	2	,000
NUMCARDS	4013,450	4132,524	3973,450	50,239	2	,000
HOWPAID	3975,097	4094,170	3935,097	11,885	2	,003
MORTGAGE	3972,128	4091,201	3932,128	8,916	2	,012
STORECAR	3987,798	4106,872	3947,798	24,587	2	,000
LOANS	4030,874	4149,948	3990,874	67,663	2	,000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

ΠΙΝΑΚΑΣ 4

Η στήλη Model Fitting Criteria του ***ΠΙΝΑΚΑΣ 4*** παρέχει έναν σύνολο ελέγχων αποτελεσματικότητας του μοντέλου (AIC, BIC, Likelihood).Παρατηρούμε κοιτώντας τα p_{value} ότι όλες οι επιδράσεις είναι στατιστικά σημαντικές.

Parameter Estimates

RISK[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
bad loss	Intercept	-1,412	,377	14,051	1	,000			
	AGE	,007	,009	,520	1	,471	1,007	,989	1,025
	INCOME	,000	,000	17,393	1	,000	1,000	1,000	1,000
	[MARITAL=divsepwid]	-4,460	,437	104,291	1	,000	,012	,005	,027
	[MARITAL=married]	,345	,211	2,665	1	,103	1,412	,933	2,137
	[MARITAL=single]	0[b]	.	.	0	.	.	.	.
	NUMKIDS	,402	,088	21,082	1	,000	1,495	1,259	1,774
	NUMCARDS	,473	,078	36,946	1	,000	1,605	1,378	1,869
	[HOWPAID=monthly]	-,587	,207	8,047	1	,005	,556	,370	,834
	[HOWPAID=weekly]	0[b]	.	.	0	.	.	.	.
	[MORTGAGE=n]	,266	,166	2,581	1	,108	1,305	,943	1,805
	[MORTGAGE=y]	0[b]	.	.	0	.	.	.	.
	STORECAR	,391	,070	31,456	1	,000	1,479	1,290	1,695
	LOANS	,628	,122	26,494	1	,000	1,874	1,475	2,380
bad profit	Intercept	4,521	,335	182,093	1	,000			
	AGE	-,064	,008	57,926	1	,000	,938	,923	,954
	INCOME	,000	,000	87,293	1	,000	1,000	1,000	1,000
	[MARITAL=divsepwid]	-3,004	,404	55,387	1	,000	,050	,022	,109
	[MARITAL=married]	-,060	,174	,119	1	,730	,942	,669	1,325
	[MARITAL=single]	0[b]	.	.	0	.	.	.	.
	NUMKIDS	,204	,082	6,184	1	,013	1,227	1,044	1,441
	NUMCARDS	,048	,067	,516	1	,472	1,049	,920	1,197
	[HOWPAID=monthly]	-,627	,182	11,853	1	,001	,534	,374	,763
	[HOWPAID=weekly]	0[b]	.	.	0	.	.	.	.
	[MORTGAGE=n]	,581	,150	15,085	1	,000	1,788	1,333	2,397
	[MORTGAGE=y]	0[b]	.	.	0	.	.	.	.
	STORECAR	,044	,062	,509	1	,476	1,045	,925	1,181
	LOANS	1,019	,113	81,291	1	,000	2,770	2,219	3,456

a. The reference category is: good risk.

b. This parameter is set to zero because it is redundant.

ΠΙΝΑΚΑΣ 5

Στον ***ΠΙΝΑΚΑΣ 5*** βλέπουμε ότι υπάρχουν 2 σετ δεδομένων. Να αναφέρουμε ότι έχουμε θέσει ως κατηγορία αναφοράς την κατηγορία «good risk». Έτσι λοιπόν το ένα σετ είναι η αναλογία πιθανότητας του «bad risk / loss» με «good risk», το οποίο ονομάζεται «bad loss» και το άλλο σετ είναι η αναλογία του «bad risk – profit» με «good risk» και ονομάζεται «bad profit».

Για κάθε ένα από τα δύο σετ δεδομένων παρουσιάζονται οι μεταβλητές, οι συντελεστές (Β) και η σταθερά. Επίσης, τα τυπικά τους σφάλματα, ένα τεστ σημαντικότητας του στατιστικού ελέγχου Wald ,οι εκθετικές τιμές των συντελεστών (Exp(B)) και ένα 95% διάστημα εμπιστοσύνης τους.

Κοιτώντας τη μεταβλητή age στο 1ο σετ βλέπουμε ότι δεν έχει επίδραση στην αναλογία πιθανότητας του "bad loss" σε σχέση με το "good risk". Κοιτάζοντας τώρα τη μεταβλητή age στο 2ο σετ παρατηρούμε ότι Exp(B) = 0.938. Αυτό σημαίνει ότι αυξάνοντας την ηλικία κατά μια μονάδα μειώνεται η αναμενόμενη πιθανότητα να είναι κάποιος «bad profit» σε σχέση με το να είναι «good risk».

Πρέπει να αναφέρουμε ότι για κάθε κατηγορική μεταβλητή συγκεκριμένα εδώ τις Marital και Mortgage, ότι η τελευταία κατηγορική τιμή γίνεται η κατηγορία αναφοράς και οι υπόλοιποι συντελεστές μεταφράζονται ως αντισταθμιστές προς την κατηγορία αναφορά. Εξετάζοντας το πίνακα βλέπουμε ότι οι τελευταίες κατηγορίες για τις μεταβλητές Marital (single) και Mortgage (y) έχουν συντελεστές (B coefficients) ίσο με μηδέν. Αφού οι συντελεστές για τις κατηγορίες αναφοράς ορίζονται με 0, δεν υπάρχουν αντίστοιχοι στατιστικοί έλεγχοι ή διαστήματα εμπιστοσύνης.

Κοιτώντας τη μεταβλητή Marital στο σετ 1 βλέπουμε ότι ο εκτιμώμενος συντελεστής Exp(B) στην κατηγορία divsepwid είναι 0,012 και στην κατηγορία married είναι 1,412. Έτσι λοιπόν περνώντας από την ομάδα των singles στην ομάδα των divsepwid τα αποτελέσματα μειώνονται (0.120) για την αναμενόμενη αναλογία πιθανότητας του να είναι κάποιος «bad loss» σε σχέση με το να είναι «good risk». Συμπερασματικά η ομάδα των divsepwid αναμένεται να έχει λιγότερους «bad loss» σε σχέση με τους «good risk» από την ομάδα των singles. Ενώ, η ομάδα των married αναμένεται να έχει περισσότερα άτομα «bad loss» σε σχέση με τους «good risk» από την ομάδα των singles.

Εξετάζοντας την κατηγορική μεταβλητή Mortgage, παρατηρούμε ότι το να έχει κάποιος στεγαστικό δάνειο αποτελεί την κατηγορία αναφοράς και εξετάζοντας τις τιμές στη στήλη Exp(B) οι συντελεστές δείχνουν ότι συγκρίνοντας τις δύο ομάδες, αυτοί που δεν έχουν στεγαστικό δάνειο έχουν μεγαλύτερη αναμενόμενη πιθανότητα,σχεδόν διπλάσια να είναι "bad profits" (1.788) σε σύγκριση με το να είναι "good risks". Παρ' όλα αυτά, η κατοχή στεγαστικού δανείου δεν έχει επίδραση στην αναλογία πιθανότητας του "bad loss" σε σχέση με το "good risk".

Κοιτώντας τη μεταβλητή Numkids και εξετάζοτας τη στήλη Exp(B) για τη μεταβλητή Numkids στο κομμάτι "bad loss" του πίνακα, ο συντελεστής εκτιμάται στο 1,495. Αυτό σημαίνει ότι για κάθε επιπρόσθετο παιδί, η αναμενόμενη αναλογία πιθανότητας να είναι κάποιος "bad loss" σε σχέση με το να είναι "good risk" είναι σχεδόν διπλάσια. Άρα, κρατώντας σταθερούς τους άλλους εκτιμητές, η προσθήκη ενός παιδιού σχεδόν διπλασιάζει την αναμενόμενη πιθανότητα να είναι κάποιος "bad loss" με το να είναι κάποιος "good risk". Στο 2ο σετ (bad profit), ο συντελεστής εκτιμάται στο 1,227. Άρα, κρατώντας σταθερούς τους άλλους εκτιμητές, η προσθήκη ενός παιδιού η αναμενόμενη πιθανότητα να είναι κάποιος "bad profit" με το να είναι κάποιος "good risk" είναι μεγαλύτερη.

Για τη μεταβλητή Howpaid παρατηρούμε ότι είναι σημαντική και για τις δύο αναλογίες "bad loss" και "bad profit".Ο συντελεστής στο σετ 1 εκτιμάται στο 0,556. Αυτό σημαίνει ότι η αναμενόμενη αναλογία πιθανότητας αυτών που πληρώνουν μηνιαία να είναι κάποιος "bad profit" σε σχέση με το να είναι "good risk" μειώνεται. Αντίστοιχα μειώνεται και η πιθανότητα να είναι κάποιος "bad loss" σε σχέση με το να είναι "good risk".

Τέλος, να αναφέρουμε ότι ενώ κάποιες μεταβλητές δε βγήκαν σημαντικές και στα 2 σετ, επειδή όμως στον ***ΠΙΝΑΚΑΣ 4*** δείχνουν ότι είναι σημαντικές, θα τις διατηρήσουμε.

Classification

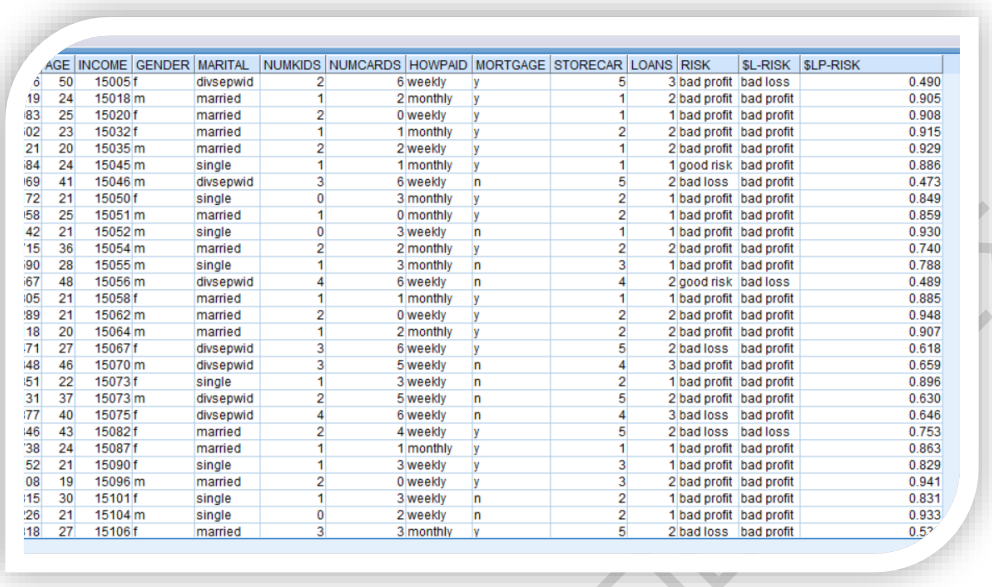
Observed	Predicted			
	bad loss	bad profit	good risk	Percent Correct
bad loss	411	355	140	45,4%
bad profit	141	2088	178	86,7%
good risk	71	219	514	63,9%
Overall Percentage	15,1%	64,7%	20,2%	73,2%

ΠΙΝΑΚΑΣ 6

Στον παραπάνω πίνακα ταξινόμησης, οι γραμμές αντιπροσωπεύουν τις παρατηρούμενες προς πρόβλεψη κατηγορίες ενώ οι στήλες αποτελούν τις προβλεπόμενες προς εκτίμηση κατηγορίες. Παρατηρούμε ότι η συνολική προβλεπτική ακρίβεια του μοντέλου είναι: 73,2%. Όπως βλέπουμε από τις εγγραφές η πιο κοινή κατηγορία είναι η bad profit με 2088 εγγραφές. Αυτό είναι το 86,7% των 4117 εγγραφών.

Εξετάζοντας τις κατηγορίες ξεχωριστά, το γκρουπ "bad risk – profit" προβλέπεται με μεγαλύτερη ακρίβεια (86,7%), ενώ οι άλλες κατηγορίες, "bad risk – loss" (45,4%) και "good risk" (63,9%) προβλέπονται με μικρότερη ακρίβεια.

Στον παρακάτω πίνακα παρουσιάζονται και 2 καινούριες στήλες που αφορούν τις προβλεπόμενες πιθανότητες για το μοντέλο και που έχουν κατανεμηθεί οι εγγραφές.



	AGE	INCOME	GENDER	MARITAL	NUMKIDS	NUMCARDS	HOWPAID	MORTGAGE	STORECAR	LOANS	RISK	$L-RISK	$LP-RISK
6	50	15005	f	divsepwid	2	6	weekly	y	5	3	bad profit	bad loss	0.490
19	24	15018	m	married	1	2	monthly	y	1	2	bad profit	bad profit	0.905
83	25	15020	f	married	2	0	weekly	y	1	1	bad profit	bad profit	0.908
02	23	15032	f	married	1	1	monthly	y	2	2	bad profit	bad profit	0.915
21	20	15035	m	married	2	2	weekly	y	1	2	bad profit	bad profit	0.929
84	24	15045	m	single	1	1	monthly	y	1	1	good risk	bad profit	0.886
69	41	15046	m	divsepwid	3	6	weekly	n	5	2	bad loss	bad profit	0.473
72	21	15050	f	single	0	3	monthly	y	2	1	bad profit	bad profit	0.849
58	25	15051	m	married	1	0	monthly	y	2	1	bad profit	bad profit	0.859
42	21	15052	m	single	0	3	weekly	n	1	1	bad profit	bad profit	0.930
15	36	15054	m	married	2	2	monthly	y	2	2	bad profit	bad profit	0.740
90	28	15055	m	single	1	3	monthly	n	3	1	bad profit	bad profit	0.788
67	48	15056	m	divsepwid	4	6	weekly	n	4	2	good risk	bad loss	0.489
05	21	15058	f	married	1	1	monthly	y	1	1	bad profit	bad profit	0.885
89	21	15062	m	married	2	0	weekly	y	2	2	bad profit	bad profit	0.948
18	20	15064	m	married	1	2	monthly	y	2	2	bad profit	bad profit	0.907
71	27	15067	f	divsepwid	3	6	weekly	y	5	2	bad loss	bad profit	0.618
48	46	15070	m	divsepwid	3	5	weekly	n	4	3	bad profit	bad profit	0.659
51	22	15073	f	single	1	3	weekly	n	2	1	bad profit	bad profit	0.896
31	37	15073	m	divsepwid	2	5	weekly	n	5	2	bad profit	bad profit	0.630
77	40	15075	f	divsepwid	4	6	weekly	n	4	3	bad loss	bad profit	0.646
46	43	15082	f	married	2	4	weekly	y	5	2	bad loss	bad loss	0.753
38	24	15087	f	married	1	1	monthly	y	1	1	bad profit	bad profit	0.863
52	21	15090	f	single	1	3	weekly	y	3	1	bad profit	bad profit	0.829
08	19	15096	m	married	2	0	weekly	y	3	2	bad profit	bad profit	0.941
15	30	15101	f	single	1	3	weekly	n	2	1	bad profit	bad profit	0.831
26	21	15104	m	single	0	2	weekly	n	2	1	bad profit	bad profit	0.933
18	27	15106	f	married	3	3	monthly	y	5	2	bad loss	bad profit	0.5?

ΠΙΝΑΚΑΣ 7

Οι πιθανότητες και για τις τρεις προς εκτίμηση κατηγορίες πρέπει να αθροίζουν στο 1. Ας πάρουμε για παράδειγμα τα δεδομένα του παραπάνω πίνακα για να δείξουμε πως ακριβώς γίνεται ο υπολογισμός. Έχουμε ένα άτομο που έχει τα εξής χαρακτηριστικά: είναι 44 ετών,είναι παντρεμένος, έχει στεγαστικό δάνειο, έχει 1 παιδί, έχει 2 κάρτες, 2 πιστωτικές κάρτες, καθόλου δάνεια, πληρώνει μηνιαία και έχει εισόδημα 59.994 λίρες.Η προβλεπόμενη πιθανότητα του συγκεκριμένου ατόμου να βρεθεί σε μια από τις τρεις κατηγορίες υπολογίζεται ως εξής:
Έχουμε:

Τύπος:

$$\pi(j) = g(j) / \Sigma\, g(i)$$

$\hat{g}(1) = e^{-1,412 + 0,007*age + 0,000*INCOME + 0,345*Marital(Married) + 0,405*Numkids + 0,473*Numcards - 0,587*Howpaid(Μηνιαία) + 0*Mortgage(y) + 0,391*Storecar + 0.628*Loans}$

⇨ $\hat{g}(1) = e^{0,787} = 2,196$

$\hat{g}(2) = e^{4,521 - 0,064*Age + 0,000*INCOME - 0,060*Marital(Married) + 0,204*Numkids + 0,048*Numcards - 0,627*Howpaid(Μηνιαία) + 0*Mortgage(y) + 0,044*Storecar + 1,019*Loans}$

⇨ $\hat{g}(2) = e^{1,406} = 4,079$

⇨ $\hat{g}(3) = 1$

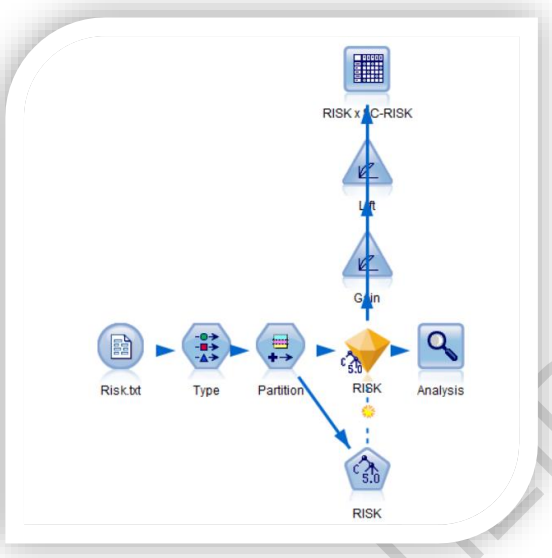
Άρα οι εκτιμώμενες πιθανότητες για το άτομο που επιλέξαμε είναι

- **π(1)= 2,196 / (2,196 + 4,079 +1)=0,301 = 30,1%**
- **π(2)= 4,079/ (2,196 + 4,079 +1)= 0,560 = 56%**
- **π(3)= 1 / (2,196 + 4,079 +1)= 0,137 = 13,7%**

Παρατηρούμε ότι το 2ο γκρουπ (bad profit) έχει τη μεγαλύτερη αναμενόμενη πιθανότητα (56%), το μοντέλο προβλέπει ότι το συγκεκριμένο άτομο ανήκει σε αυτό το γκρουπ. Το επόμενο πιο πιθανό γκρουπ στο οποίο θα μπορούσε να ταξινομηθεί το συγκεκριμένο άτομο είναι το 1ο (bad risk) επειδή έχει τη δεύτερη μεγαλύτερη αναμενόμενη πιθανότητα (0.301).

3.2 Δέντρα Αποφάσεων

- **Αλγόριθμος $C_{5.0}$**

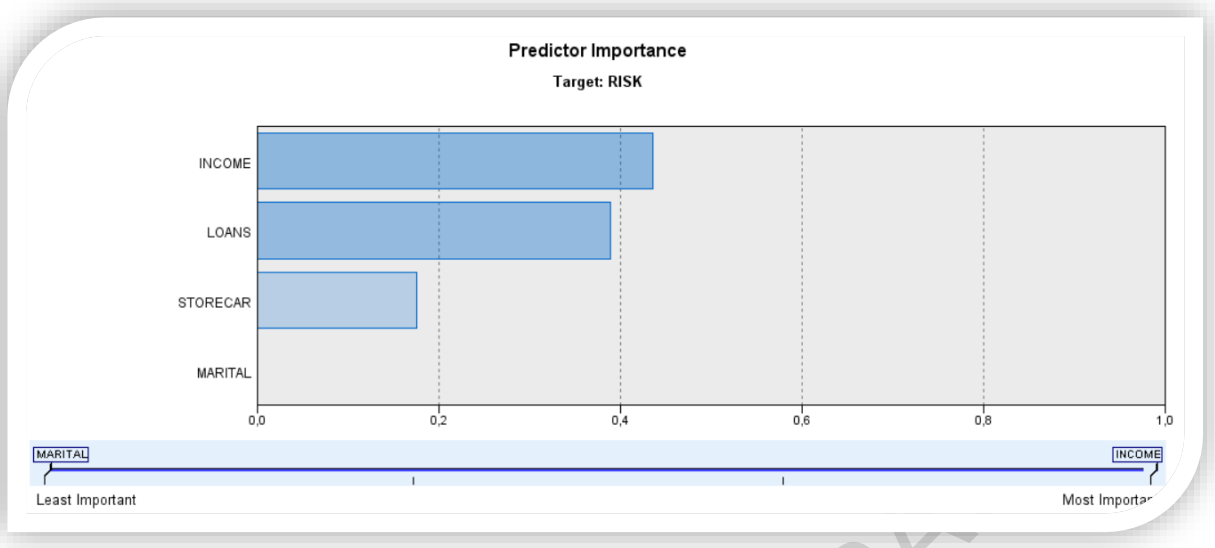


Αποτελέσματα

	INCOME	GENDER	MARITAL	NUMKIDS	NUMCARDS	HOWPAID	MORTGAGE	STORECAR	LOANS	RISK	Partition	$C-RISK	$CC-RISK
44	59944	m	married	1	2	monthly	y	2	0	good risk	1_Training	good risk	0.711
35	59692	m	married	1	1	monthly	y	1	0	bad loss	1_Training	good risk	0.711
34	59508	m	married	1	1	monthly	y	2	1	good risk	1_Training	good risk	0.526
34	59463	m	married	0	2	monthly	y	1	1	bad loss	2_Testing	good risk	0.526
39	59393	f	married	0	2	monthly	y	1	0	good risk	1_Training	good risk	0.711
41	59276	m	married	1	2	monthly	y	1	1	good risk	1_Training	good risk	0.526
42	59201	m	married	0	1	monthly	y	2	0	good risk	1_Training	good risk	0.711
31	59193	f	married	1	2	monthly	y	1	1	good risk	1_Training	good risk	0.526
28	59179	m	married	1	1	monthly	y	2	1	bad loss	1_Training	good risk	0.526
30	59036	m	married	1	1	monthly	y	2	1	good risk	1_Training	good risk	0.526
38	58914	m	married	0	1	monthly	y	1	1	bad profit	1_Training	good risk	0.526
36	58878	f	married	1	1	monthly	y	1	0	bad profit	1_Training	good risk	0.711
42	58785	f	married	0	2	monthly	y	1	0	good risk	2_Testing	good risk	0.711
44	58529	m	married	0	1	monthly	y	1	0	bad loss	1_Training	good risk	0.711
33	58505	f	married	0	2	monthly	y	1	0	good risk	2_Testing	good risk	0.711
45	58381	m	married	1	1	monthly	y	1	0	good risk	2_Testing	good risk	0.711
34	58026	m	married	0	1	monthly	y	2	0	good risk	1_Training	good risk	0.711
32	57718	m	married	1	2	monthly	y	1	1	bad profit	1_Training	good risk	0.526
35	57689	m	married	1	1	monthly	y	2	1	good risk	1_Training	good risk	0.526
38	57683	f	married	1	1	monthly	y	2	1	bad loss	1_Training	good risk	0.526
28	57623	m	married	1	1	monthly	y	1	1	bad loss	1_Training	good risk	0.526
43	57598	f	married	1	1	monthly	y	1	1	good risk	2_Testing	good risk	0.526
41	57520	f	married	1	1	monthly	y	1	0	bad loss	1_Training	good risk	0.711
43	57388	f	married	0	1	monthly	y	1	0	bad loss	1_Training	good risk	0.71
44	57376	m	married	0	2	monthly	y	2	1	good risk	1_Training	good risk	0.F
37	57004	f	married	1	1	monthly	y	2	0	good risk	1_Training	good risk	

ΠΙΝΑΚΑΣ 1

Αρχικά στον ***ΠΙΝΑΚΑΣ 1*** παρατηρούμε στο αρχείο μας να έχουν προστεθεί 2 μεταβλητές . Η μεταβλητή $C-RISK αντιπροσωπεύει τις προβλέψεις για κάθε εγγραφή-πελάτη και η μεταβλητή $CC-RISK τη σχετική πιθανότητα .



ΠΙΝΑΚΑΣ 2

Στον παραπάνω ***ΠΙΝΑΚΑΣ 2*** βλέπουμε να εμφανίζονται οι πιο σημαντικές μεταβλητές, με πιο σημαντική τη μεταβλητή Income.

ΠΙΝΑΚΑΣ 3

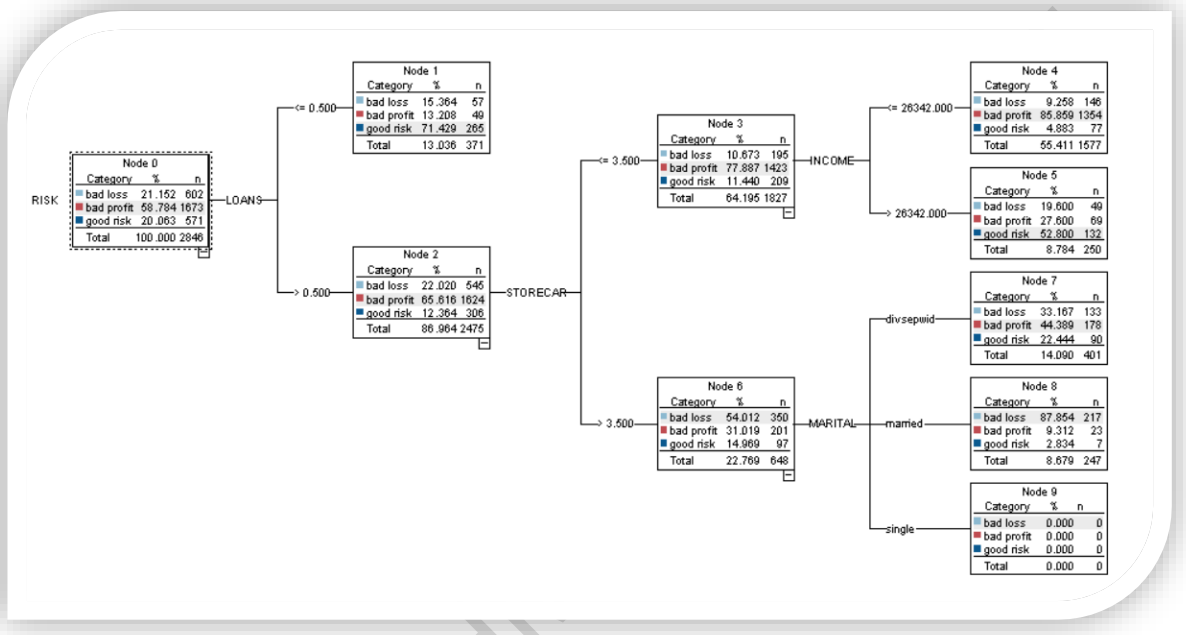
Στον παραπάνω πίνακα βλέπουμε να εμφανίζονται τα αποτελέσματα, που έχουν τη μορφή ενός δένδρου αποφάσεων. Η μεταβλητή LOANS αποτελεί τη πρώτη μεταβλητή στη διακλάδωση.

Για παράδειγμα εάν ενδιαφερόμαστε για bad profit πελάτες , μια ομάδα που θα προβλέπαμε ως πελάτες είναι αυτή που εάν έχουν LOANS > 0,500 , STORECAR ≤ 3,500 και

INCOME ≤ 26.342. Σε αυτές τις περιπτώσεις που το Mode αποτελεί και την πρόβλεψη θεωρούνται τερματικά Nodes.

Μέσα στην παρένθεση βλέπουμε ότι 1.577 άτομα πληρούν αυτά τα κριτήρια που αναφέραμε. Ο αριθμός 0,859 ονομάζεται confidence και είναι η αναλογία των εγγραφών που έχουν ταξινομηθεί σωστά, πιο συγκεκριμένα αυτοί που προβλέφθηκαν ως bad risk και πραγματικά ήταν bad risk. Αντίστοιχα και στα υπόλοιπα.

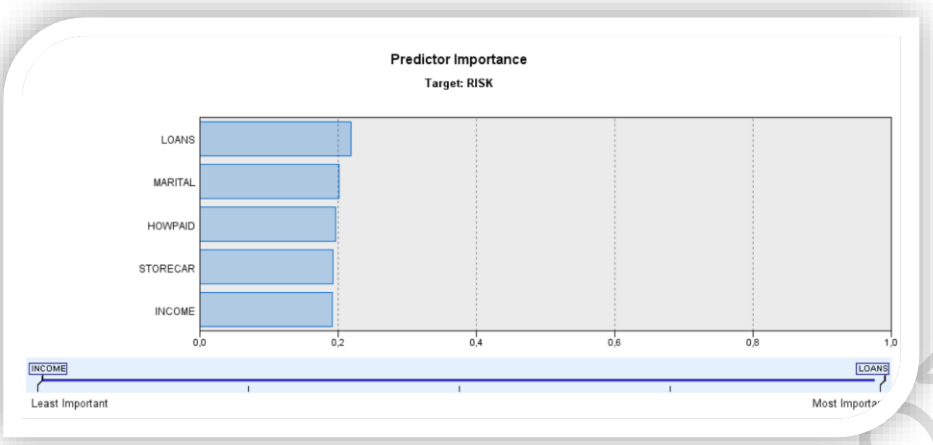
Μια άλλη επιλογή παρουσίασης του παραπάνω δέντρου (***ΠΙΝΑΚΑΣ 3***) είναι αυτή :



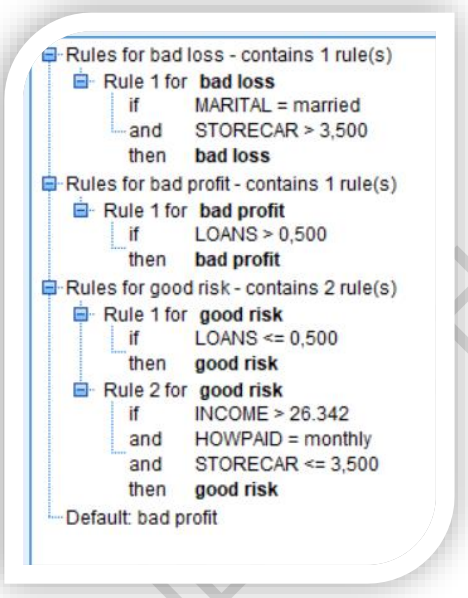
ΠΙΝΑΚΑΣ 4

Η αρχή του δένδρου μας δείχνει το συνολικό αριθμό και τα ποσοστά και των τριών κατηγοριών της μεταβλητής RISK. Όπως είδαμε και προηγουμένως η μεταβλητή για την πρώτη διακλάδωση είναι η LOANS. Τα σκιαγραφημένα είναι η κατηγορία στην οποία καταλήγει κάθε κανόνας, αν αποτελεί και την πρόβλεψη βλέπουμε ότι κόβεται και δεν υπάρχει άλλη διακλάδωση.

Ο αλγόριθμος αυτός έχει και την επιλογή να δούμε ένα σετ κανόνων αντί να αναλύουμε ολόκληρο το δέντρο:



ΠΙΝΑΚΑΣ 5



Για παράδειγμα ο κανόνας 1 για να είναι ένα άτομο bad loss πρέπει να είναι παντρεμένος και να έχει STORECAR >3,500

ΠΙΝΑΚΑΣ 6

Από το να αναλύουμε λοιπόν όλο το δέντρο η μορφή αυτή των κανόνων με If και Then εστιάζει περισσότερο σ΄ ένα συγκεκριμένο συμπέρασμα.

Προβλεπτική ακρίβεια του μοντέλου

Για να ελέγξουμε εάν ένα δέντρο απόφασης είναι καλό μοντέλο θα χρησιμοποιήσουμε το Matrix Node, συγκεκριμένα στο Modeler, το Analysis Node και το Evaluation Charts. Παρακάτω παρουσιάζονται τα αποτελέσματα.

- Analysis Node

esults for output field RISK

Comparing $C-RISK with RISK

'Partition'	1_Training		2_Testing	
Correct	2.146	75,4%	943	74,19%
Wrong	700	24,6%	328	25,81%
Total	2.846		1.271	

Coincidence Matrix for $C-RISK (rows show actuals)

'Partition' = 1_Training	bad loss	bad profit	good risk
bad loss	217	279	106
bad profit	23	1.532	118
good risk	7	167	397
'Partition' = 2_Testing	**bad loss**	**bad profit**	**good risk**
bad loss	116	142	46
bad profit	10	665	59
good risk	3	68	162

ΠΙΝΑΚΑΣ 7

Κοιτάζοντας τα αποτελέσματα του ***ΠΙΝΑΚΑΣ 7***, παρατηρούμε ότι στο Training set, το μοντέλο προβλέπει σωστά το 75.4% συνολικά και αντίστοιχα στο Testing Set το 74.1%.Όπως παρατηρούμε κινούνται στο ίδιο επίπεδο τα δύο set, το οποίο δηλώνει ότι θα εκτελείται καλά το μοντέλο με νέα δεδομένα.

- *Matrix Node*

Training set

$C-RISK

RISK		bad loss	bad profit	good risk
bad loss	Count	217	279	106
	Row %	36.047	46.346	17.608
	Column %	87.854	14.105	17.069
bad profit	Count	23	1532	118
	Row %	1.375	91.572	7.053
	Column %	9.312	77.452	19.002
good risk	Count	7	167	397
	Row %	1.226	29.247	69.527
	Column %	2.834	8.443	63.929

Testing set

$C-RISK

RISK		bad loss	bad profit	good risk
bad loss	Count	116	142	46
	Row %	38.158	46.711	15.132
	Column %	89.922	16.229	17.228
bad profit	Count	10	665	59
	Row %	1.362	90.599	8.038
	Column %	7.752	76.000	22.097
good risk	Count	3	68	162
	Row %	1.288	29.185	69.528
	Column %	2.326	7.771	60.674

ΠΙΝΑΚΑΣ 8

Κοιτάζοντας τα αποτελέσματα του πίνακα Training, το μοντέλο προβλέπει σωστά το 36.04% της κατηγορία bad loss, το 91.57% της κατηγορίας bad profit και το 69.52% της κατηγορίας good risk. Τα αποτελέσματα του πίνακα Testing set όπως παρατηρούμε κινούνται περίπου στο ίδιο επίπεδο με το αυτά του πίνακα Training set, το οποίο δηλώνει ότι θα εκτελείται καλά το μοντέλο με νέα δεδομένα.

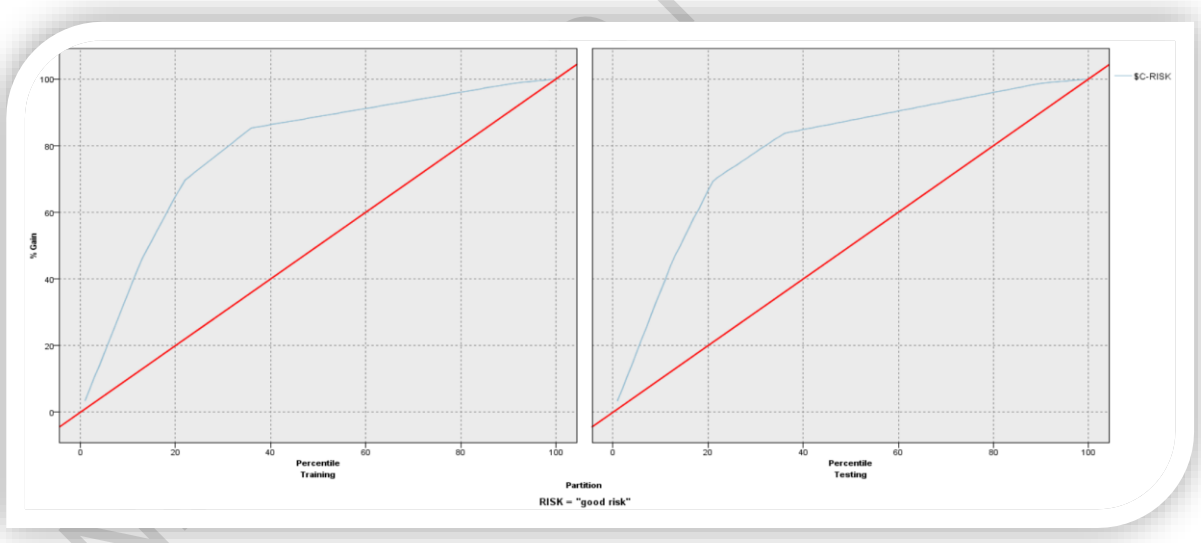
Αξιολόγηση του μοντέλου

Επιλέγουμε ως κατηγορία αναφοράς :"good risk"

- Evaluation chart (Gain)

Τα αθροιστικά διαγράμματα gain ξεκινούν πάντα στο 0% και τελειώνουν στο 100% καθώς πηγαίνετε από αριστερά προς τα δεξιά. Για ένα καλό μοντέλο, το διάγραμμα κερδών θα αυξηθεί απότομα στο 100% και στη συνέχεια θα μειωθεί.

Training set ***Testing set***



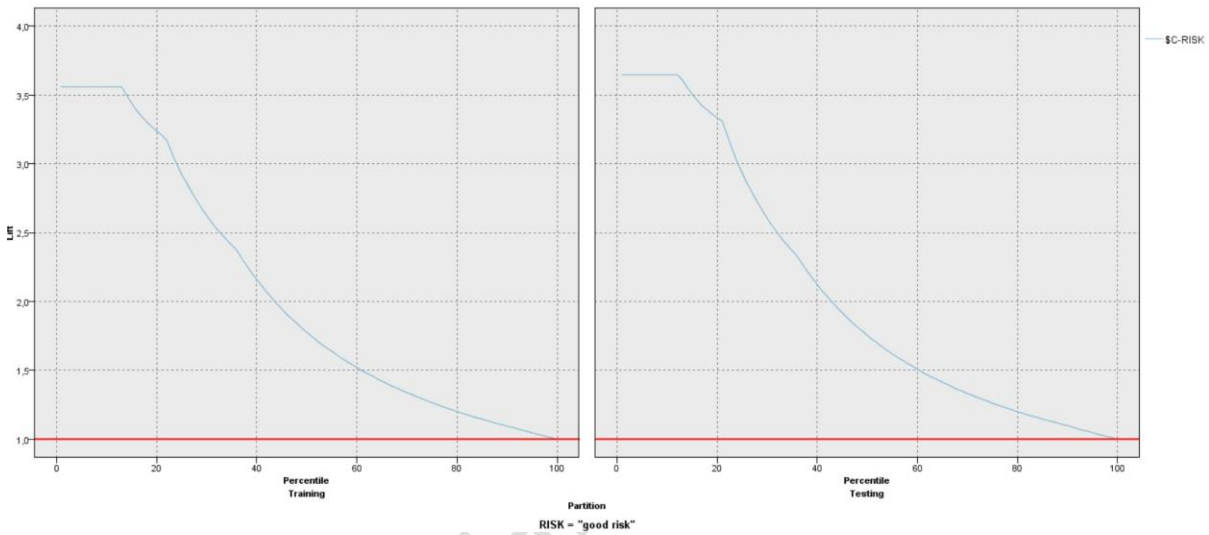
ΠΙΝΑΚΑΣ 9

Τα παραπάνω διαγράμματα αντιπροσωπεύουν το gain chart. Κοιτώντας για παράδειγμα το ποσοστό 5% βλέπουμε το gain να είναι ίσο με 17,763 αυτό σημαίνει ότι το ότι με τη χρήση του μοντέλου επιλέγοντας το 5% των πελατών θα έχουμε το 17,7% των good risk. Τα γραφήματα των Testing και Training δεδομένων είναι αρκετά όμοια, γεγονός το οποίο δηλώνει ότι το συγκεκριμένο μοντέλο μπορεί να χρησιμοποιηθεί αξιόπιστα για τη πρόβλεψη good risk πελατών με νέα δεδομένα.

Πιο αναλυτικά, σημαίνει ότι αν σκοράρουμε ένα σύνολο δεδομένων και ταξινομήσουμε όλες τις περιπτώσεις με την προβλεπόμενη πιθανότητα «good risk», θα περιμένουμε το κορυφαίο 5% των πελατών με τις μεγαλύτερες πιθανότητες να περιέχει περίπου το 17,7% όλων των περιπτώσεων που στην πραγματικότητα ανήκουν στην κατηγορία good risk (προεπιλεγμένοι).

- Evaluation chart (Lift)

Training set ***Testing set***



ΠΙΝΑΚΑΣ 10

Το διάγραμμα lift προέρχεται από το γράφημα αθροιστικών κερδών. Οι τιμές στον άξονα y αντιστοιχούν στην αναλογία του αθροιστικού κέρδους για κάθε καμπύλη προς τη γραμμή βάσης. Έτσι, το lift στο 5% για την κατηγορία Good Risk στο training test είναι :

$\frac{gain\%}{5\%} = \frac{0.17763}{0.05}$ =3,56. Παρέχει έναν άλλο τρόπο εξέτασης των πληροφοριών στο γράφημα αθροιστικών κερδών.

Το αντίστοιχο Lift στο 5% είναι 3,5602 στο Training set και 3,6482 στο Testing set.

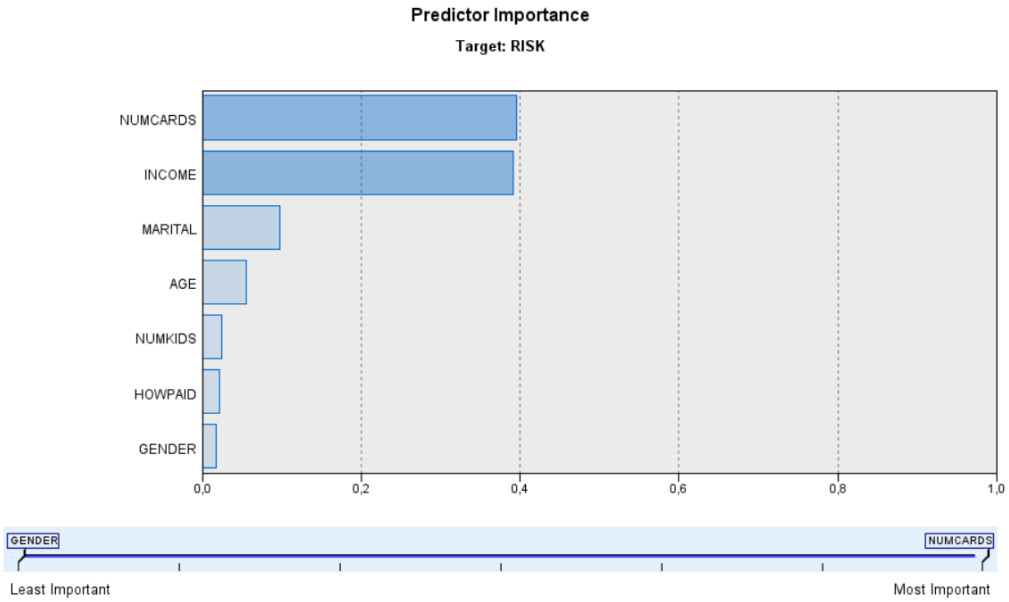
- **<u>Αλγόριθμος CHAID</u>**

Αποτελέσματα

NCOME	GENDER	MARITAL	NUMKIDS	NUMCARDS	HOWPAID	MORTGAGE	STORECAR	LOANS	RISK	Partition	$R-RISK	$RC-RISK
22930	f	married	3	4	weekly	y	5	1	bad loss	2_Testing	bad loss	0.730
23240	f	married	2	4	weekly	y	5	1	bad loss	2_Testing	bad loss	0.895
20863	m	married	2	4	monthly	y	4	2	bad loss	1_Training	bad loss	0.895
16473	f	married	2	3	monthly	y	5	2	bad loss	1_Training	bad loss	0.921
16475	m	married	2	3	monthly	y	4	2	bad loss	1_Training	bad loss	0.697
19013	f	married	3	4	monthly	y	4	1	bad loss	2_Testing	bad loss	0.730
23240	f	married	3	3	monthly	y	4	2	bad loss	2_Testing	bad loss	0.921
21002	f	married	2	3	weekly	y	4	2	bad loss	2_Testing	bad loss	0.921
17088	f	married	2	4	weekly	y	5	2	bad loss	1_Training	bad loss	0.895
17697	m	married	3	3	monthly	y	4	2	bad loss	1_Training	bad loss	0.697
16492	m	married	3	4	weekly	y	4	2	bad loss	1_Training	bad loss	0.730
16494	f	married	2	4	weekly	y	5	1	bad loss	1_Training	bad loss	0.895
16504	f	married	2	3	monthly	y	4	2	bad loss	1_Training	bad loss	0.921
21716	m	married	2	4	weekly	y	5	1	bad loss	1_Training	bad loss	0.895
15800	m	married	3	3	weekly	y	4	2	bad loss	2_Testing	bad loss	0.697
23361	m	married	2	3	monthly	y	5	1	bad profit	1_Training	bad loss	0.697
16520	m	married	3	3	monthly	y	5	1	bad loss	1_Training	bad loss	0.697
18586	m	married	2	4	weekly	y	5	2	bad loss	2_Testing	bad loss	0.895
17101	f	married	3	3	weekly	y	4	2	bad loss	1_Training	bad loss	0.921
17705	m	married	2	3	weekly	y	4	2	bad loss	2_Testing	bad loss	0.697
20150	m	married	2	4	monthly	y	5	1	bad loss	1_Training	bad loss	0.895
15082	f	married	2	4	weekly	y	5	2	bad loss	2_Testing	bad loss	0.895
22446	m	married	2	4	weekly	y	5	2	bad loss	2_Testing	bad loss	0.895
21835	f	married	3	3	weekly	y	5	2	bad loss	1_Training	bad loss	0.921
22140	f	married	2	4	monthly	y	4	1	bad loss	2_Testing	bad loss	0.895
21019	f	married	3	4	weekly	y	4	1	bad loss	1_Training	bad loss	0.730
19288	f	married	2	4	monthly	y	4	1	bad loss	1_Training	bad loss	0.8
15106	f	married	3	3	monthly	y	5	2	bad loss	1_Training	bad loss	0
23251	m	married	2	3	monthly	y	4	1	bad loss	1_Training	bad loss	

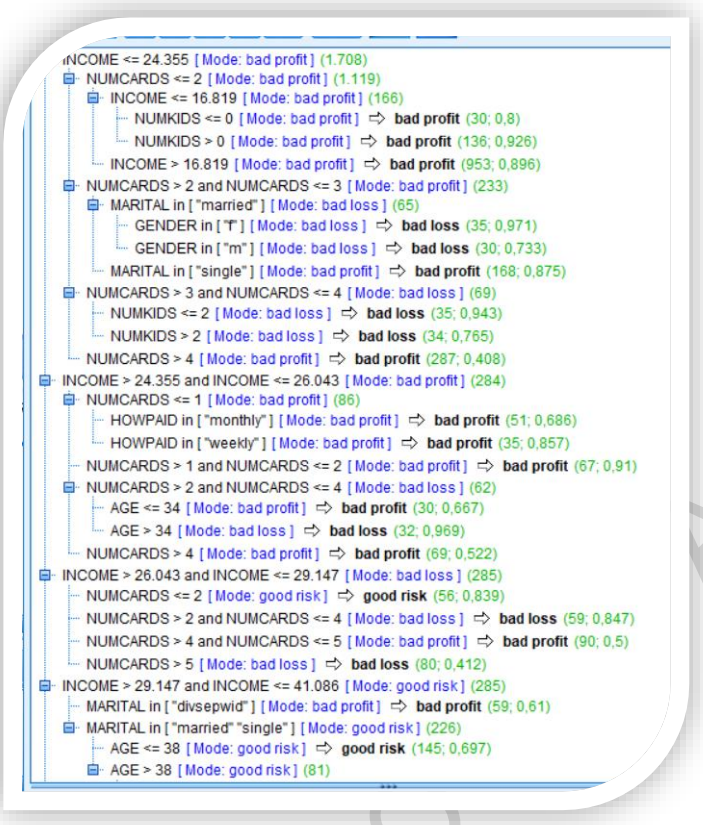
ΠΙΝΑΚΑΣ 1

Παρατηρούμε και εδώ τις δυο καινούριες μεταβλητές στο dataset .



ΠΙΝΑΚΑΣ 2

Παρατηρούμε ότι βάσει σημαντικότητας η 1η μεταβλητή είναι η NUMCARDS.



ΠΙΝΑΚΑΣ 3

Στον παραπάνω πίνακα βλέπουμε ότι η μεταβλητή INCOME αποτελεί τη πρώτη μεταβλητή στη διακλάδωση. Συνεχίζοντας, για παράδειγμα εάν ενδιαφερόμαστε για good risk πελάτες , μια ομάδα που θα προβλέπαμε ως πελάτες είναι αυτή που θα έχουν εισόδημα ανάμεσα από INCOME > 26.043 και INCOME>29.147 και NUMCARDS ≤ 2. Σε αυτές τις περιπτώσεις που το Mode αποτελεί και την πρόβλεψη θεωρούνται τερματικά Nodes.

Μέσα στην παρένθεση βλέπουμε ότι 145 άτομα πληρούν αυτά τα κριτήρια που αναφέραμε και την αναλογία των εγγραφών που έχουν ταξινομηθεί σωστά, πιο συγκεκριμένα το ποσοστό που των good risk που προβλέφθηκαν σωστά 0,697.

- Analysis Node

Results for output field RISK

Comparing $R-RISK with RISK

'Partition'	1_Training		2_Testing	
Correct	2.139	75,16%	941	74,04%
Wrong	707	24,84%	330	25,96%
Total	2.846		1.271	

Coincidence Matrix for $R-RISK (rows show actuals)

'Partition' = 1_Training	bad loss	bad profit	good risk
bad loss	252	274	76
bad profit	54	1.531	88
good risk	45	170	356
'Partition' = 2_Testing	**bad loss**	**bad profit**	**good risk**
bad loss	127	141	36
bad profit	26	663	45
good risk	12	70	151

Κοιτάζοντας τα αποτελέσματα του πίνακα,παρατηρούμε ότι στο Training set, το μοντέλο προβλέπει σωστά το 75.16% συνολικά και αντίστοιχα στο Testing Set το 74.04%.Όπως παρατηρούμε και εδώ κινούνται στο ίδιο επίπεδο τα δύο set , το οποίο δηλώνει ότι θα εκτελείται καλά το μοντέλο με νέα δεδομένα.

ΠΙΝΑΚΑΣ 4

- Matrix Node

Training set

$R-RISK

RISK		bad loss	bad profit	good risk
bad loss	Count	252	274	76
	Row %	41.860	45.515	12.625
	Column %	71.795	13.873	14.615
bad profit	Count	54	1531	88
	Row %	3.228	91.512	5.260
	Column %	15.385	77.519	16.923
good risk	Count	45	170	356
	Row %	7.881	29.772	62.347
	Column %	12.821	8.608	68.462

Testing set

$R-RISK

RISK		bad loss	bad profit	good risk
bad loss	Count	127	141	36
	Row %	41.776	46.382	11.842
	Column %	76.970	16.133	15.517
bad profit	Count	26	663	45
	Row %	3.542	90.327	6.131
	Column %	15.758	75.858	19.397
good risk	Count	12	70	151
	Row %	5.150	30.043	64.807
	Column %	7.273	8.009	65.086

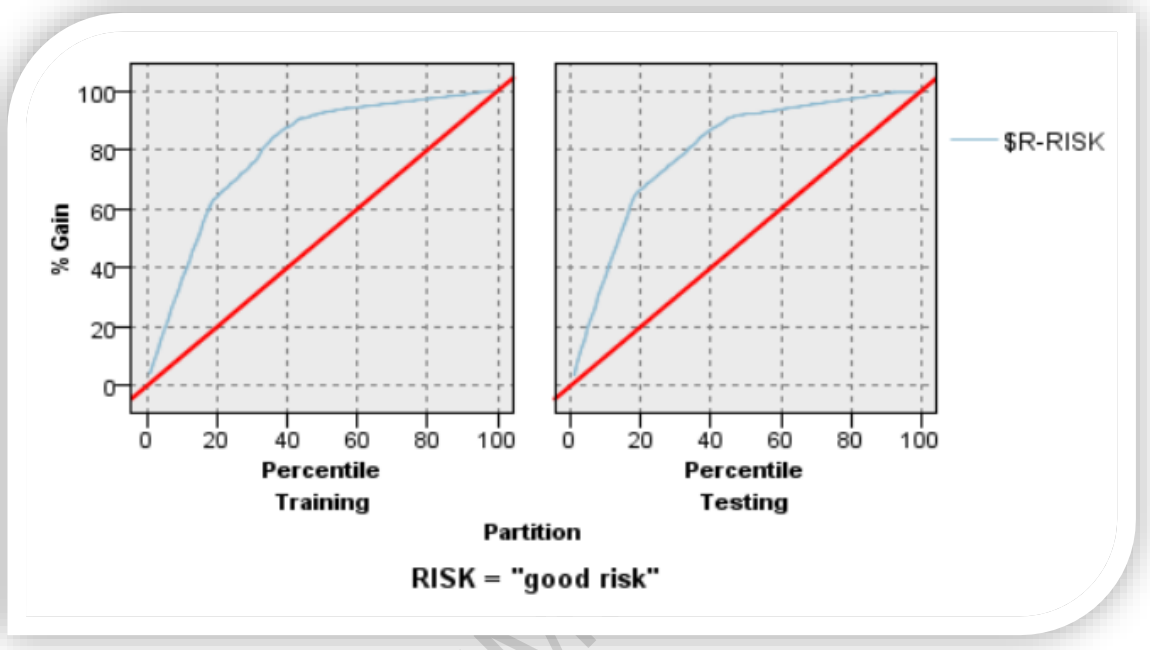
ΠΙΝΑΚΑΣ 5

Κοιτάζοντας τα αποτελέσματα του πίνακα στο Training test, το μοντέλο προβλέπει σωστά το 41.86% της κατηγορία bad loss, το 91.512% της κατηγορίας bad profit και το 62.347% της κατηγορίας good risk. Τα αποτελέσματα στο Testing set όπως παρατηρούμε κινούνται περίπου στο ίδιο επίπεδο με αυτά του Training set, το οποίο δηλώνει ότι θα εκτελείται καλά το μοντέλο με νέα δεδομένα.

Επιλέγουμε και εδώ ως κατηγορία αναφοράς :"good risk"

- *Evaluation chart (Gain)*

Training set ***Testing set***

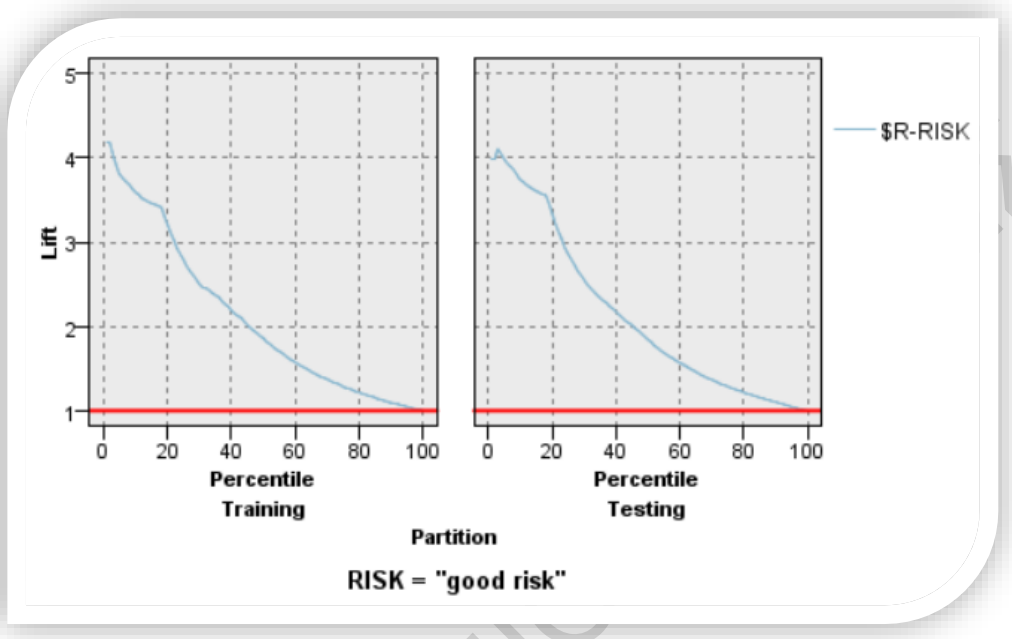


ΠΙΝΑΚΑΣ 6

Κοιτώντας τον ***ΠΙΝΑΚΑΣ 6*** έστω ότι παίρνουμε το 40% των πελατών με τις υψηλότερες πιθανότητες βλέπουμε το gain να είναι ίσο με 87,7 αυτό σημαίνει ότι με τη χρήση του μοντέλου επιλέγοντας το 40% των πελατών με τις μεγαλύτερες πιθανότητες, θα έχουμε βρει το 87,7 % των good risk. Τα γραφήματα Testing set και Training set είναι αρκετά όμοια, γεγονός το οποίο δηλώνει ότι το συγκεκριμένο μοντέλο μπορεί να χρησιμοποιηθεί αξιόπιστα για την πρόβλεψη good risk πελατών με νέα δεδομένα.

- *Lift chart*

Training set ***Testing set***

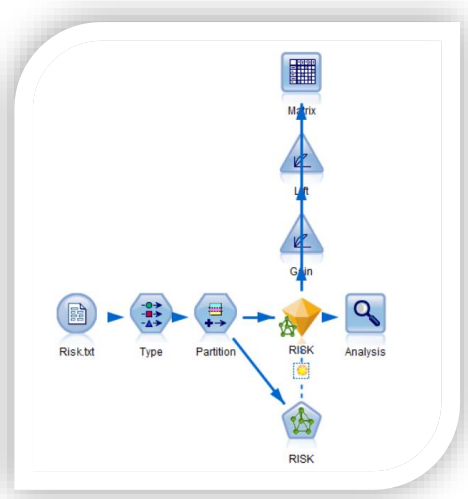


ΠΙΝΑΚΑΣ 7

Το αντίστοιχο Lift στο 40% είναι 2,1944 στο Training set και 2,1722 στο Testing set.

Συμπερασματικά, και τα δυο μοντέλα έχουν την ίδια σχεδόν προβλεπτική ικανότητα.

3.3 Νευρωνικά Δίκτυα



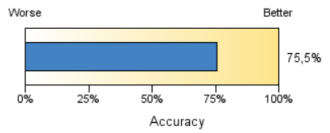
Το SPSS MODELER δίνει ως επιλογή δύο τύπους μοντέλων που καθορίζουν πώς το δίκτυο συνδέει τις ανεξάρτητες μεταβλητές με τους στόχους μέσω των κρυφών επιπέδων. Το multilayer perceptron (MLP) επιτρέπει πιο περίπλοκες σχέσεις, με το πιθανό κόστος αυξημένο χρόνο προπόνησης των δεδομένων και του scoring. Ο άλλος τύπος radial basis function (RBF) μπορεί να έχει χαμηλότερους χρόνους εκπαίδευσης και scoring, αλλά αντίστοιχα πιθανό κόστος μειωμένης προγνωστικής ισχύος σε σύγκριση με το MLP.

Στις επιλογές του νευρωνικού δικτύου το IBM SPSS Modeler σταματάει να εκπαιδεύει ένα μοντέλο όταν φαίνεται ότι η ακρίβεια στα δεδομένα ελέγχου δεν βελτιώνεται παραπάνω. Διαφορετικά μπορούμε να ορίσουμε μια τιμή ακρίβειας, ένα όριο στον αριθμό των κύκλων των δεδομένων ή ένα χρονικό όριο σε λεπτά.

Αποτελέσματα

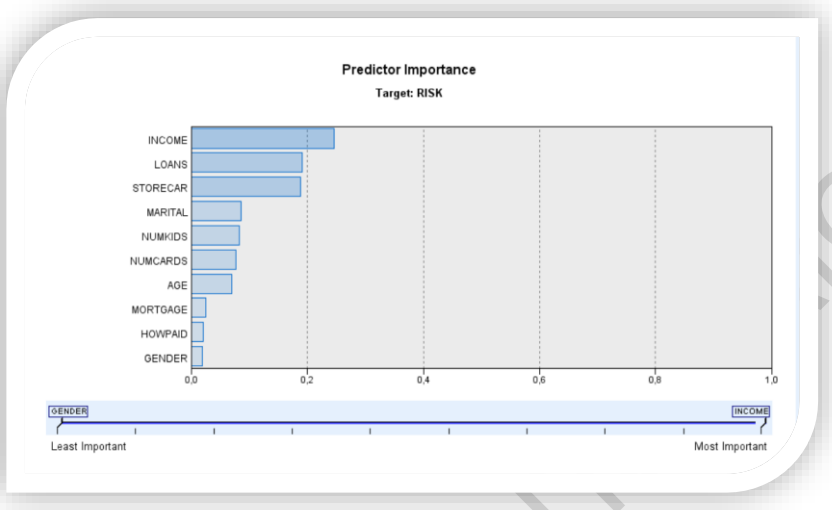
Model Summary

Target	RISK
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	6



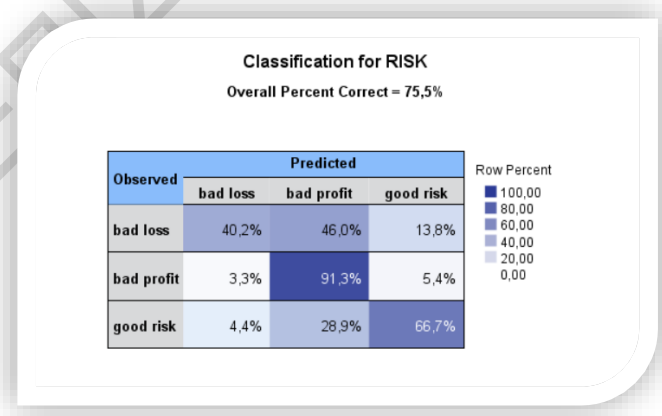
ΠΙΝΑΚΑΣ 1

Στον παραπάνω πίνακα φαίνεται η προβλεπτική ακρίβεια του νευρωνικού δικτύου που είναι στο 75%, δηλαδή η αναλογία που έχει προβλεφθεί σωστά. Επίσης παρατηρούμε ότι το στρώμα εισόδου (input layer) έχει δημιουργηθεί από ένα νευρώνα για συνεχή ή δίτιμα πεδία. Τα Set πεδία έχουν ένα νευρώνα για κάθε κατηγορία της μεταβλητής.



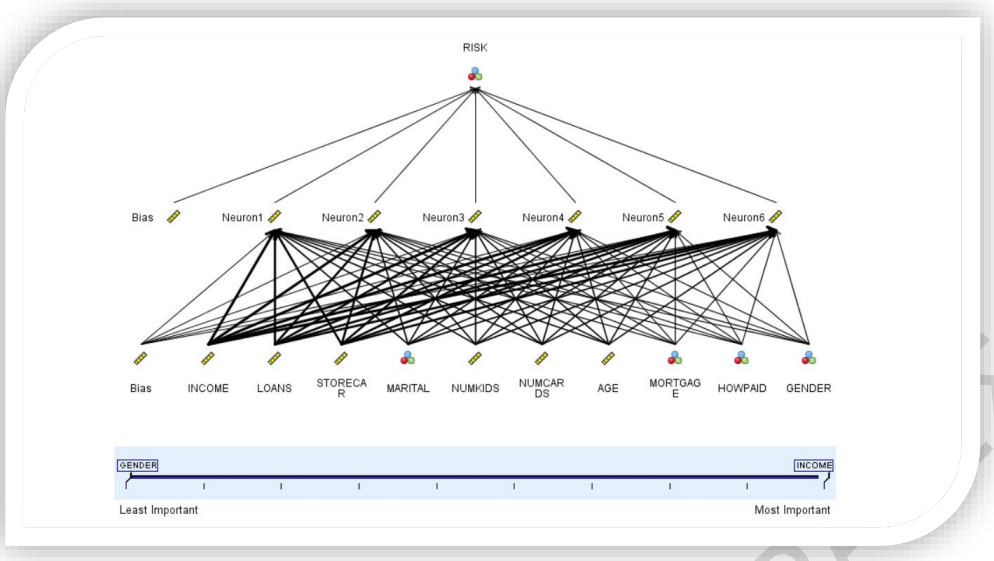
ΠΙΝΑΚΑΣ 2

Στον παραπάνω πίνακα βλέπουμε τους εκτιμητές που χρησιμοποιήθηκαν στο μοντέλο σε φθίνουσα σειρά με βάση τη σημαντικότητα τους. Η πρώτη πιο σημαντική μεταβλητή είναι η INCOME και δεύτερη ακολουθεί η LOANS.



ΠΙΝΑΚΑΣ 3

Παραπάνω παρουσιάζεται ένας classification table και βλέπουμε ότι το συνολικό ποσοστό σωστής ταξινόμησης είναι 75,5% όπως είδαμε και παραπάνω. Επίσης παρατηρούμε ότι οι σωστές ταξινομήσεις ανά κατηγορία είναι : 40,2% bad loss , 91.3% bad profit και 66,7% good risk.



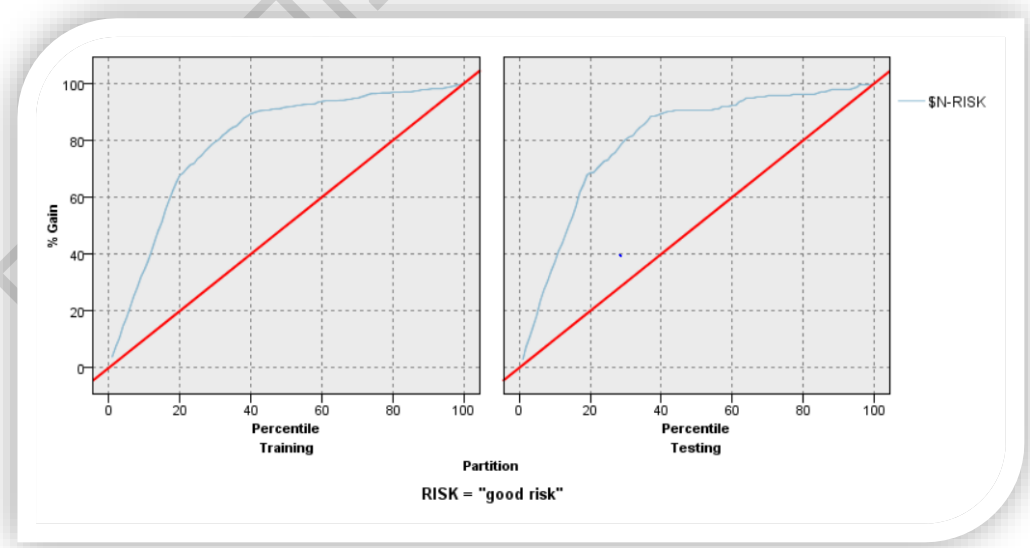
ΠΙΝΑΚΑΣ 4

Στην παραπάνω εικόνα εμφανίζεται το νευρωνικό δίκτυο που δημιουργήθηκε. Παρατηρούμε τις μεταβλητές του μοντέλου με πιο σημαντική τη μεταβλητή INCOME και λιγότερο σημαντική τη μεταβλητή GENDER.

Αξιολόγηση του μοντέλου

- *Evaluation chart (Gain)*

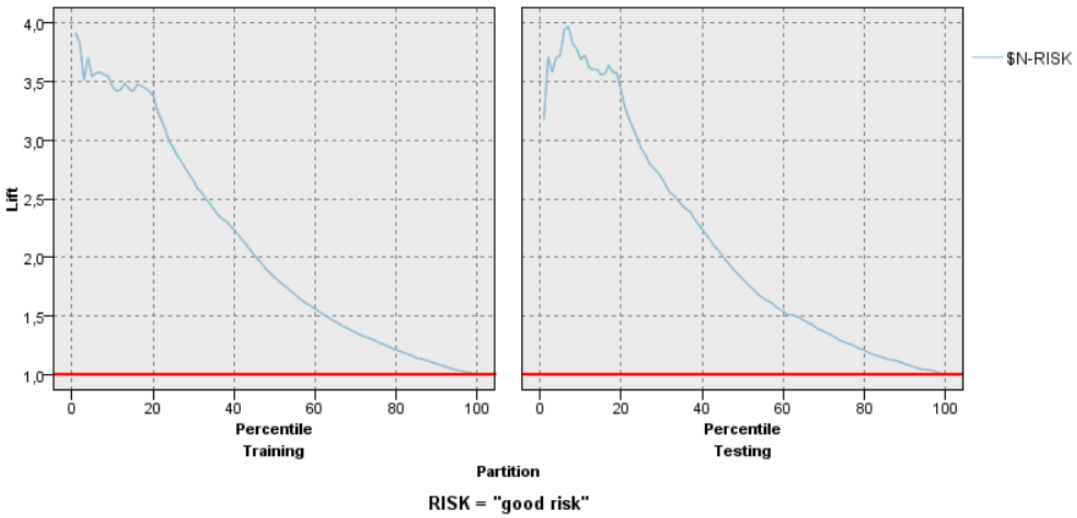
Training set ***Testing set***



Κοιτώντας για παράδειγμα το ποσοστό 30% των πελατών με τις μεγαλύτερες πιθανότητες (χρησιμοποιώντας το ποντίκι πάνω στο γράφημα) βλέπουμε το gain να είναι ίσο με 79,51, αυτό σημαίνει ότι με τη χρήση του μοντέλου επιλέγοντας το 30% των πελατών με τις μεγαλύτερες πιθανότητες, θα έχουμε το 79,51 % των good risk. Η καμπύλη gain στο γράφημα με το Testing set είναι αρκετά όμοιο με του training set το γράφημα, γεγονός το οποίο δηλώνει ότι το συγκεκριμένο μοντέλο μπορεί να χρησιμοποιηθεί αξιόπιστα για τη πρόβλεψη good risk πελατών με νέα δεδομένα.

- *Lift*

Training set ***Testing set***



Το αντίστοιχο Lift στο 30% είναι 2,6528 στο Training set και 2,6774 στο Testing set.

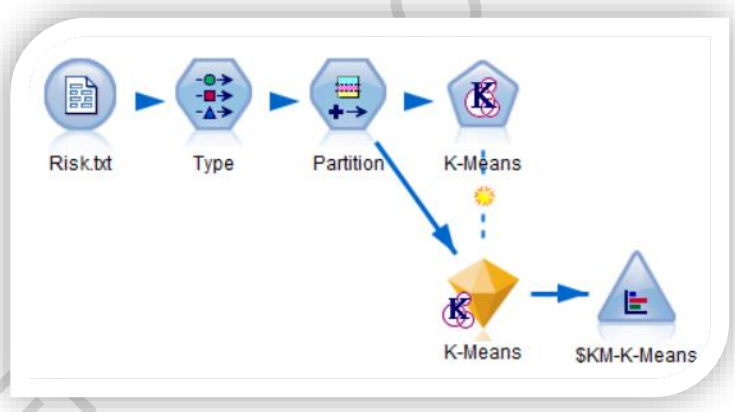
3.4 Ανάλυση κατά συστάδες

Θα χρησιμοποιήσουμε τη μέθοδο ανάλυση κατά συστάδες για να προσπαθήσουμε να βρούμε φυσικές ομάδες πελατών, για να διαπιστώσουμε για παράδειγμα κατά πόσο θα μπορούσε η τράπεζα να χρησιμοποιήσει διαφορετικές προωθητικές ενέργειες στοχευμένες στη κάθε ομάδα πελατών, και να διερευνήσουμε αν οι διαφορετικές ομάδες σχετίζονται με τη κατάσταση του πελάτη.

Όπως και στη μοντελοποίηση έτσι και στην ομαδοποίηση μπορούμε να χρησιμοποιήσουμε δύο σετ δεδομένων ένα για να δημιουργήσουμε τη λύση και το άλλο για να προσπαθήσουμε να επαναλάβουμε την ίδια λύση με μια δεύτερη ανάλυση. Αν οι βασικές ομάδες εμφανιστούν ξανά τότε μπορούμε να είμαστε πιο σίγουροι ότι η λύση θα ισχύει και στα μελλοντικά δεδομένα.

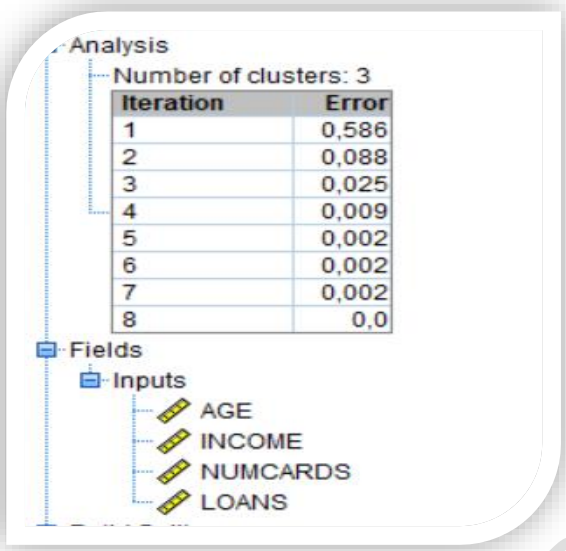
- **Αλγόριθμος K-means**

Ο αριθμός των συστάδων που καταλήξαμε μετά από πολλές δοκιμές, είναι 3 και οι μεταβλητές που θα χρησιμοποιηθούν είναι η ηλικία , ο αριθμός καρτών που κατέχει ο κάθε πελάτης, το εισόδημα του και αριθμός δανείων που έχει.



Δημιουργείται αυτόματα ένα πεδίο $KM - K- Means που καταγράφει τη συστάδα που ανήκει η κάθε εγγραφή, εάν επιλέξουμε και την επιλογή Generate distance field θα δημιουργήσει ένα πεδίο με το όνομα $KMD-K-Means, που δείχνει την απόσταση μεταξύ της εγγραφής και το κέντρο της αντίστοιχης συστάδας. Ο αλγόριθμος χρησιμοποιεί την Ευκλείδεια απόσταση.

Η default επιλογή είναι ο αλγόριθμος να τερματίσει όταν το άθροισμα όλων των αλλαγών στα κέντρα των συστάδων (Change tolerance) είναι λιγότερο από ένα συγκεκριμένο όριο (default 0.000001) ή εναλλακτικά όταν ο αλγόριθμος έχει ήδη πραγματοποιήσει ένα μέγιστο αριθμό επαναλήψεων (Maximum Iterations). Η προεπιλογή είναι 20 επαναλήψεις. είναι συνήθως ικανοποιητική γι' αυτό θα την αφήσουμε.

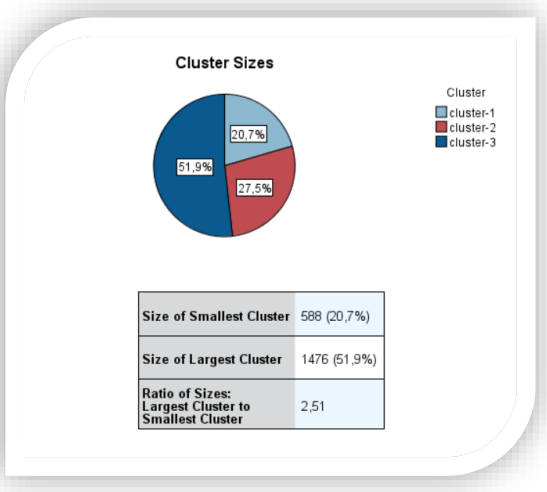


ΠΙΝΑΚΑΣ 1

Παρατηρούμε ότι δημιουργήθηκαν 3 συστάδες όπως ζητήθηκε και ότι μετά από 8 επαναλήψεις του αλγόριθμου ικανοποιήθηκε το κριτήριο της ελάχιστης μεταβολής 0.000001 και ο αλγόριθμος ολοκληρώθηκε.

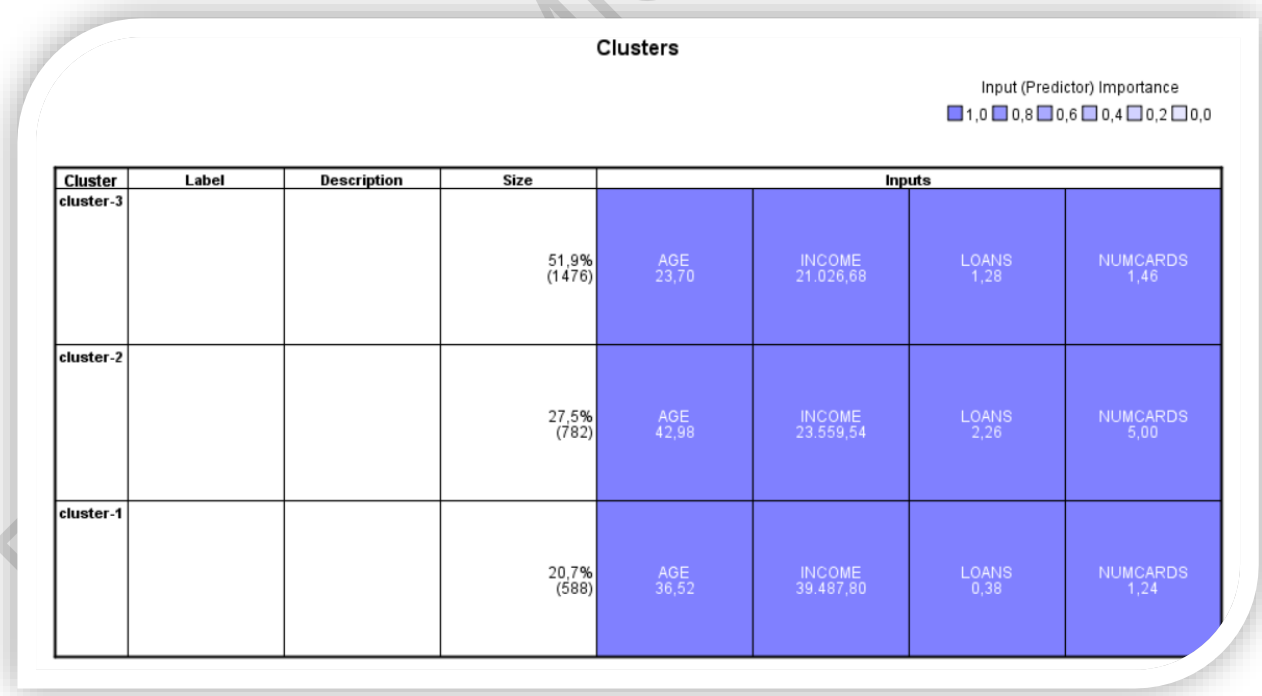
ΠΙΝΑΚΑΣ 2

Στον παραπάνω πίνακα βλέπουμε πληροφορίες για τον αριθμό των πεδίων που έχουν χρησιμοποιηθεί και των συστάδων που έχουν δημιουργηθεί και επιπλέον μας δίνεται και ένα μέτρο της ποιότητας της λύσης (silhouette measure) = 0.6 (good).



ΠΙΝΑΚΑΣ 3

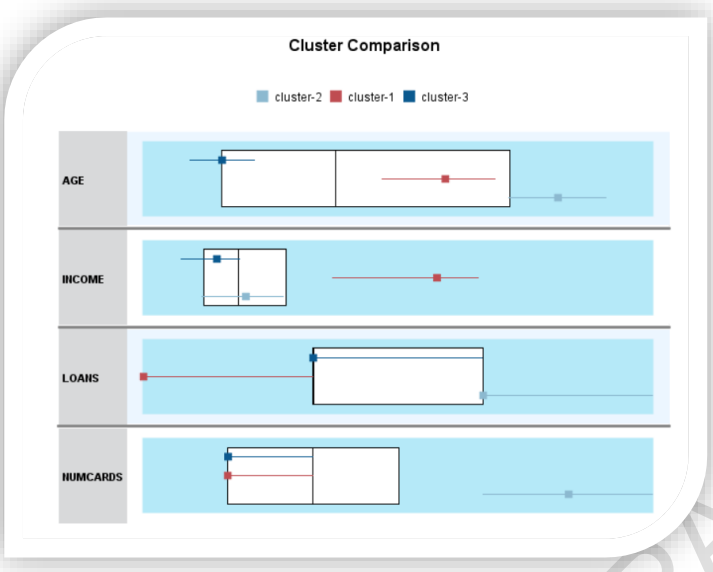
Στο παραπάνω γράφημα (pie-chart) παρατηρούμε πως κατανέμονται οι εγγραφές στις ομάδες και ακριβώς από κάτω το μέγεθος της μικρότερης και της μεγαλύτερης ομάδας. Πιο συγκεκριμένα παρατηρούμε ότι στην ομάδα 1 είναι το 20,7% των παρατηρήσεων , στην ομάδα 2 το 27,5% και στην ομάδα 3 το 51,9%.Η μεγαλύτερη ομάδα είναι η 3.



Cluster	Label	Description	Size	Inputs			
cluster-3			51,9% (1476)	AGE 23,70	INCOME 21.026,68	LOANS 1,28	NUMCARDS 1,46
cluster-2			27,5% (782)	AGE 42,98	INCOME 23.559,54	LOANS 2,26	NUMCARDS 5,00
cluster-1			20,7% (588)	AGE 36,52	INCOME 39.487,80	LOANS 0,38	NUMCARDS 1,24

ΠΙΝΑΚΑΣ 4

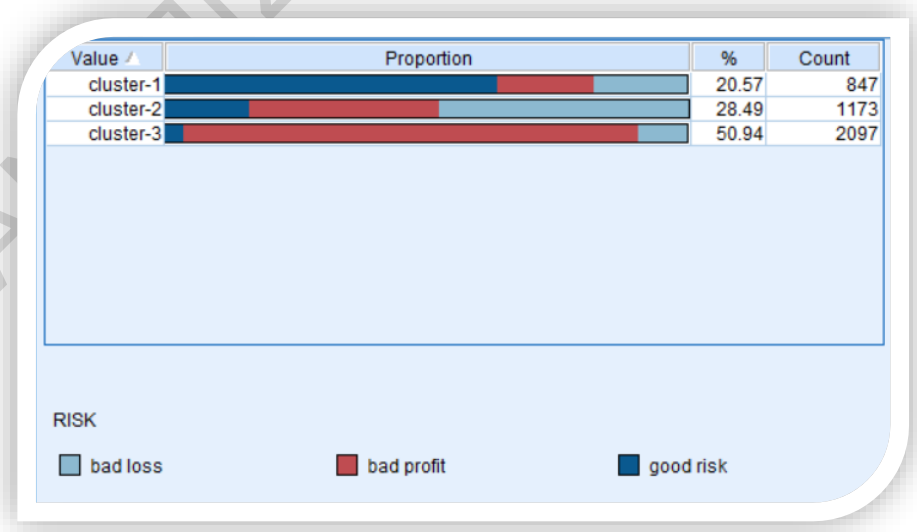
Στον παραπάνω πίνακα περιέχονται πληροφορίες για κάθε μεταβλητή που περιέχεται σε κάθε ομάδα (cluster). Για παράδειγμα το Cluster 1 αποτελείται από πελάτες με μέση ηλικία 36 χρονών, μέσο εισόδημα 39.48, μέσο όρο μια πιστωτική κάρτα και καθόλου δάνεια.



ΠΙΝΑΚΑΣ 5

Με τον παραπάνω πίνακα μπορούμε να κάνουμε μια εύκολη σύγκριση των ομάδων. Βλέπουμε ότι η ομάδα 3 περιέχει τους μικρότερους σε ηλικία και με το μικρότερο εισόδημα. Η ομάδα 1 περιέχει πελάτες μέσης ηλικίας με το υψηλότερο εισόδημα και με λίγα ή σχεδόν καθόλου δάνεια. Τέλος η ομάδα 2 εμπεριέχει πελάτες με πολλές κάρτες , πολλά δάνεια και μεγάλους σε ηλικία.

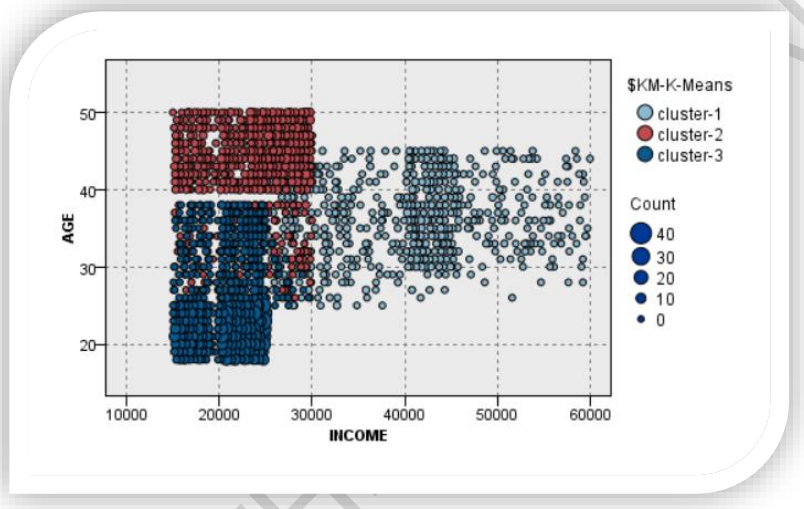
Για να εξετάσουμε τη σχέση των συστάδων με το πεδίο στόχο Risk :



ΠΙΝΑΚΑΣ 6

Παρατηρούμε ότι στην ομάδα 1 ανήκουν οι good risk, αυτή η συστάδα περιέχει πελάτες με μέσες ηλικίες, μεγάλα εισοδήματα, σχεδόν καθόλου δάνεια και μέσο όρο 1-2 κάρτες στην κατοχή τους. Στην ομάδα 2 φαίνεται να υπερτερούν οι bad loss, αυτή η συστάδα περιέχει πελάτες με μέση ηλικία 43 χρονών, μέσο εισόδημα 23.559 , κατά μέσο όρο 2 δάνεια και 5 κάρτες. Στην ομάδα 3 φαίνεται να υπερτερούν οι bad profit με μέση ηλικία 23 χρονών, μέσο εισόδημα 21.026, κατά μέσο όρο 1 δάνειο και 1 κάρτα.

Κάνοντας και ένα Scatter Plot από την παλέτα Graphs , χρησιμοποιώντας τις μεταβλητές age στο Y field, income στο X field και $KM-K-Means στο Overlay color έχουμε το παρακάτω διάγραμμα :



ΔΙΑΓΡΑΜΜΑ 1

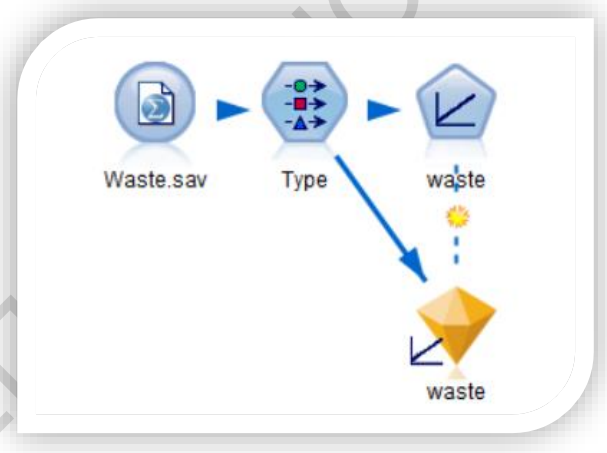
Από το παραπάνω διάγραμμα παρατηρούμε ότι οι συστάδες δεν συμπίπτουν ιδιαίτερα μεταξύ τους. Αυτό συμβαίνει γιατί οι μεταβλητές είναι σημαντικές για την κατασκευή των ομάδων. Έτσι βλέπουμε και γραφικά ότι η ομάδα 1 (που εμπεριέχει τους περισσότερους Good Risk) αποτελείται από άτομα με μεσαία προς μεγάλα εισοδήματα και αντίστοιχα μέσες ηλικίες.

Για να μπορέσουμε να χρησιμοποιήσουμε πολλαπλή παλινδρόμηση και παραγοντική ανάλυση, θα χρησιμοποιήσουμε το παρακάτω dataset.

ΠΕΔΙΟ - ΣΤΟΧΟΣ	ΠΕΡΙΓΡΑΦΗ
waste	Στερεά απόβλητα

ΠΕΔΙΑ	ΠΕΡΙΓΡΑΦΗ
indust	Βιομηχανική γη
metals	Περιοχή εξόρυξης μετάλλων
trucks	Χονδρεμπόριο
retail	Λιανεμπόριο
restr	Εστιατόρια και ξενοδοχεία

3.5 Γραμμική Παλινδρόμηση



Αποτελέσματα

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,921[a]	,848	,826	,150801	1,686

a. Predictors: (Constant), restr, metals, indust, trucks, retail

ΠΙΝΑΚΑΣ 1

Το μοντέλο παλινδρόμησης παρατηρούμε ότι εξηγεί περίπου το 85% της εξαρτημένης μεταβλητής waste.

Correlations

		indust	metals	trucks	retail	restr	waste
Pearson Correlation	indust	1,000	,393	,424	,138	,233	,186
	metals	,393	1,000	,893	,282	,199	,484
	trucks	,424	,893	1,000	,467	,384	,645
	retail	,138	,282	,467	1,000	,920	,767
	restr	,233	,199	,384	,920	1,000	,821
	waste	,186	,484	,645	,767	,821	1,000
Sig. (1-tailed)	indust	.	,006	,003	,198	,074	,126
	metals	,006	.	,000	,039	,109	,001
	trucks	,003	,000	.	,001	,007	,000
	retail	,198	,039	,001	.	,000	,000
	restr	,074	,109	,007	,000	.	,000
	waste	,126	,001	,000	,000	,000	.
N	indust	40	40	40	40	40	40
	metals	40	40	40	40	40	40
	trucks	40	40	40	40	40	40
	retail	40	40	40	40	40	40
	restr	40	40	40	40	40	40
	waste	40	40	40	40	40	40

ΠΙΝΑΚΑΣ 2

Βλέπουμε αρχικά τον πίνακα συσχετίσεων του Pearson. Αν παρατηρήσουμε βλέπουμε ότι όλες οι μεταβλητές συσχετίζονται θετικά, με κάποιες να έχουν υψηλές συσχετίσεις, όπως οι μεταβλητές metals-trucks (0.893) και μεταβλητές retail-restrnts (0.920). Οι υψηλές συσχετίσεις αυτές μπορεί να μας δημιουργήσουν προβλήματα σταθερότητας του μοντέλου (μεγάλα standard errors) λόγω πολυσυγραμμικότητας.

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,123	,032		3,872	,000
	indust	-5,241E-5	,000	-,232	-2,919	,006
	metals	3,946E-5	,000	,041	,256	,799
	trucks	,000	,000	,497	2,854	,007
	retail	-,001	,000	-,436	-2,263	,030
	restr	,013	,002	1,078	5,805	,000

ΠΙΝΑΚΑΣ 3

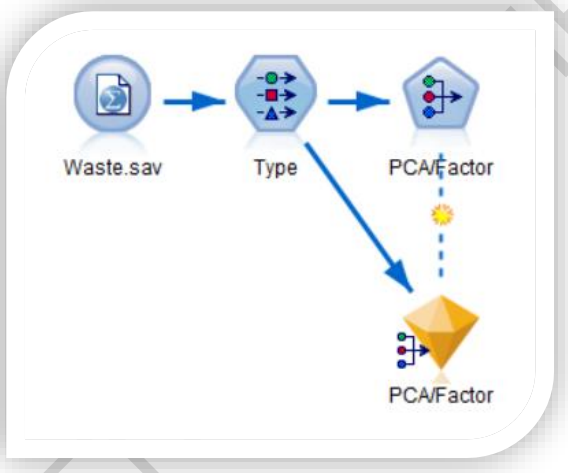
Βάσει της καρτέλας summary στα αποτελέσματα του μοντέλου και στον πίνακα 3 το μοντέλο έχει τη μορφή:

Y= 0,123-0,00005241*indust+0,00003946*metals+0,0002533*trucks+0,01327*restr

Δύο σημαντικές για το μοντέλο μεταβλητές (indust, retail) ενώ έχουν θετική συσχέτιση με την waste, παρατηρούμε ότι έχουν αρνητικά coefficients για κάποιο λόγο, δεδομένου ότι η μεταβλητή retail είναι υψηλά συσχετισμένη με μια άλλη μεταβλητή.

Έτσι καταλαβαίνουμε ότι το σετ των input μεταβλητών παρουσιάζουν σημάδια πολυσυγραμμικότητας και έτσι θα προχωρήσουμε σε Παραγοντική ανάλυση για να βελτιώσουμε τη κατάσταση.

3.6 Παραγοντική Ανάλυση



Επιλέξαμε να διαλέξει όσους παράγοντες έχουν ιδιοτιμή μεγαλύτερη της μονάδας.

Αποτελέσματα

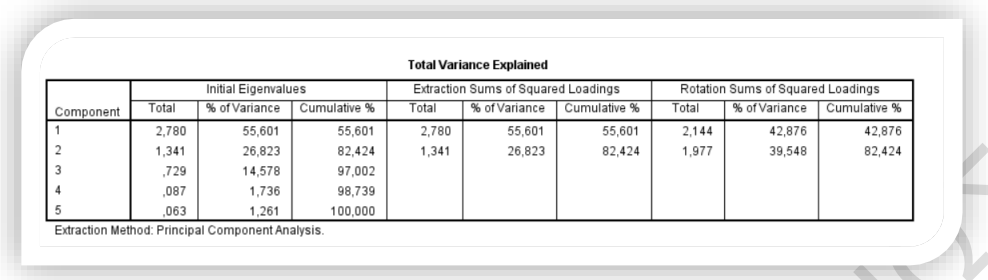
Communalities

	Initial	Extraction
indust	1,000	,441
metals	1,000	,870
trucks	1,000	,893
retail	1,000	,964
restr	1,000	,953

Extraction Method: Principal Component Analysis.

ΠΙΝΑΚΑΣ 1

Ο παραπάνω πίνακας, περιέχει τα Communalities τα οποία αντιπροσωπεύουν το ποσοστό της μεταβλητότητας των πεδίων που εξηγείται από τους παράγοντες. Όλες εκτός τη μεταβλητή Indust έχουν ένα μεγάλο ποσοστό μεταβλητότητας που εξηγείται από τους δύο παράγοντες.Της μεταβλητής indust εξηγείται μόνο το 44.1% της μεταβλητότητας της.



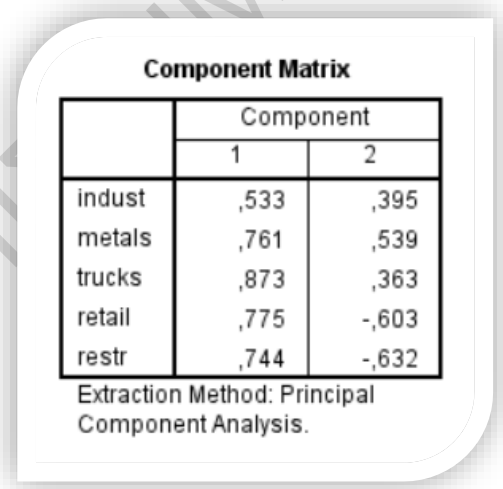
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,780	55,601	55,601	2,780	55,601	55,601	2,144	42,876	42,876
2	1,341	26,823	82,424	1,341	26,823	82,424	1,977	39,548	82,424
3	,729	14,578	97,002						
4	,087	1,736	98,739						
5	,063	1,261	100,000						

Extraction Method: Principal Component Analysis.

ΠΙΝΑΚΑΣ 2

Στις στήλες του Initial Eigenvalues βλέπουμε και τις πέντε ιδιοτιμές καθώς και το ποσοστό που ερμηνεύει η κάθε μία. Επιλέγονται αυτές οι οποίες είναι μεγαλύτερες της μονάδας. Αυτές είναι οι πρώτες δυο με τη πρώτη να είναι διπλάσια από την δεύτερη. Επιπρόσθετα παρατηρούμε ότι οι 2 παράγοντες που δημιουργήθηκαν εξηγούν σχεδόν το 82% της μεταβλητότητας των 5 ανεξάρτητων μεταβλητών.



Component Matrix

	Component	
	1	2
indust	,533	,395
metals	,761	,539
trucks	,873	,363
retail	,775	-,603
restr	,744	-,632

Extraction Method: Principal Component Analysis.

ΠΙΝΑΚΑΣ 3

Στον ***ΠΙΝΑΚΑΣ 3*** βλέπουμε τις ανεξάρτητες μεταβλητές και τους παράγοντες. Οι τιμές που βλέπουμε αποτελούν τα φορτία (loadings) που σε όλές τις μεταβλητές είναι θετικά
Θα χρησιμοποιήσουμε τη μέθοδο περιστροφής Varimax για καλύτερα αποτελέσματα.

Rotated Component Matrix

	Component	
	1	2
indust	,661	,059
metals	,927	,104
trucks	,893	,309
retail	,178	,966
restr	,135	,967

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 3 iterations.

ΠΙΝΑΚΑΣ 4

Παρατηρούμε ότι ο 1ος παράγοντας αποτελείται από τις μεταβλητές: μέταλλα, χονδρεμπόριο και βιομηχανική γη. Θα μπορούσε να αντιπροσωπεύει τη χρήση γης από βαριά βιομηχανία. Ενώ ο 2ος παράγοντας αποτελείται από τα εστιατόρια ξενοδοχεία και λιανικό εμπόριο και ουσιαστικά αντιπροσωπεύει τη χρήση γης από τομείς παροχής υπηρεσιών.

- Ξανατρέχοντας την παλινδρόμηση, αυτή τη φορά με ανεξάρτητες τους 2 παράγοντες, έχουμε:

Αποτελέσματα

Correlations

		$F-Factor-1	$F-Factor-2	waste
Pearson Correlation	$F-Factor-1	1,000	,000	,396
	$F-Factor-2	,000	1,000	,765
	waste	,396	,765	1,000
Sig. (1-tailed)	$F-Factor-1	.	,500	,006
	$F-Factor-2	,500	.	,000
	waste	,006	,000	.
N	$F-Factor-1	40	40	40
	$F-Factor-2	40	40	40
	waste	40	40	40

ΠΙΝΑΚΑΣ 1

Παρατηρούμε ότι δεν υπάρχει συσχέτιση μεταξύ των παραγόντων αλλά βλέπουμε και τη συσχέτιση που έχει η μεταβλητή στόχος με τους 2 παράγοντες.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,861[a]	,742	,728	,188330

a. Predictors: (Constant), $F-Factor-2, $F-Factor-1

ΠΙΝΑΚΑΣ 2

Το μοντέλο παλινδρόμησης με τους δύο παράγοντες επεξηγεί το 73% της μεταβλητότητας της waste. Χάσαμε 10% περίπου σε σχέση με το προηγούμενο μοντέλο αλλά κερδίσαμε πιο σταθερά coefficients και πιθανά ευκολότερη επεξήγηση του μοντέλου.

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,381	,030		12,786	,000
	$F-Factor-1	,143	,030	,396	4,743	,000
	$F-Factor-2	,276	,030	,765	9,161	,000

ΠΙΝΑΚΑΣ 3

Παρατηρώντας τα p_{value} του ελέγχου, βλέπουμε ότι και οι 2 παράγοντες είναι στατιστικά σημαντικοί.

Ο πρώτος παράγοντας μας δείχνει ότι όσο αυξάνει η χρήση γης από βαριά βιομηχανία αυξάνουν και τα απορρίμματα (waste). Ο δεύτερος παράγοντας (ο οποίος έχει αρνητικό πρόσημο) μας δείχνει ότι όσο αυξάνει η χρήση γης από τομείς παροχής υπηρεσιών μειώνεται η παραγωγή απορριμμάτων.

Τέλος είναι σημαντικό να δούμε ότι από μια παλινδρόμηση με 5 μεταβλητές καταλήξαμε σε μία πιο σταθερή παλινδρόμηση με δύο μεταβλητές

Κεφάλαιο 6

Συμπεράσματα

Ο πιστωτικός κίνδυνος ενός πελάτη της τράπεζας είναι ο κίνδυνος απώλειας λόγω αποτυχίας ή αδυναμίας του να αποπληρώσει ένα καταναλωτικό πιστωτικό προϊόν, όπως ένα στεγαστικό δάνειο, ένα προσωπικό δάνειο, πιστωτική κάρτα, υπερανάληψη κ.λπ. Ο πιστωτικός κίνδυνος είναι από τα σημαντικότερα θέματα που πρέπει να διαχειριστεί μια τράπεζα και με όσο περισσότερη ακρίβεια είναι εφικτό. Μια τράπεζα με αυξημένο αριθμό επισφαλών δανείων διατρέχει κίνδυνο στο μέλλον να εμφανίσει μεγάλες ζημιές, δανειακές αλλά και μη.

Στη σημερινή εποχή η τεράστια ποσότητα δεδομένων πελατών που συλλέγουν οι τράπεζες τους δίνει ένα πλεονέκτημα στην κατανόηση των κινδύνων που παρουσιάζουν οι πελάτες, στο σχεδιασμό προϊόντων για αυτούς και στο μάρκετινγκ σε αυτούς. Κάθε τράπεζα μπορεί να έχει τη δική της προσέγγιση για τη δημιουργία μοντέλων διαχείρισης πιστωτικού κινδύνου. Φυσικά, πρέπει να καθορίζουν στόχους, χαρακτηριστικά και στρατηγικές που σχετίζονται με την αποτελεσματική διαχείριση του πιστωτικού κινδύνου.

Οι τράπεζες ενδιαφέρονται να εντοπίσουν 'καλούς' και 'κακούς' πελάτες πριν τη χορήγηση δανείου ή γενικά κάποιου πιστωτικού προϊόντος. Με χρήση ιστορικών στοιχείων σχετικά με άτομα που έλαβαν στο παρελθόν πιστωτικό προϊόν από την τράπεζα, η τράπεζα θα ήθελε να σχηματίσει κανόνες ώστε να μπορεί να κατατάξει έναν καινούριο πελάτη σε μια από τις κατηγορίες που έχει κατατάξει το πελατολόγιό της.

Οι στατιστικές τεχνικές που χρησιμοποιούνται για τη μέτρηση πιστωτικού κινδύνου είναι κυρίως, η διαχωριστική ή διακριτική ανάλυση, τα Δένδρα απόφασης, η πολλαπλή παλινδρόμηση, το λογιστικό μοντέλο, η Ανάλυση επιβίωσης.

Στο dataset που χρησιμοποιήθηκε στη διπλωματική οι πελάτες είναι χωρισμένοι σε 3 κατηγορίες, σε good risk πελάτες,bad risk-loss και bad risk-profit. Τα αποτελέσματα από την πολυωνυμική παλινδρόμηση που χρησιμοποιήθηκε, ουσιαστικά μπορούν να βοηθήσουν μια τράπεζα να διακρίνει χαρακτηριστικά τα οποία θα τη βοηθήσουν να κατατάξει έναν νέο πελάτη, σε μια από τις 3 αυτές κατηγορίες.

Για παράδειγμα αν η κατηγορία που ενδιέφερε την τράπεζα ήταν οι πελάτες good risk θα εστίαζε στα χαρακτηριστικά που ήταν σημαντικά γι΄αυτήν. Ένα τέτοιο σημαντικό χαρακτηριστικό που θα μπορούσε να κατατάξει ένα νέο πελάτη σ αυτή την κατηγορία θα ήταν, ο τρόπος πληρωμής του δανείου (πιστωτικού προϊόντος). Από την ανάλυση συμπεραίνουμε ότι αυτοί που πληρώνουν μηνιαία έχουν μεγαλύτερη πιθανότητα να είναι good risk απ’ ότι bad loss ή bad profit. Ακόμη μια τέτοια μεταβλητή είναι η οικογενειακή κατάσταση. Παρατηρείται ότι από τους χωρισμένους και σε σχέση με τους ελεύθερους αναμένεται να είναι περισσότεροι οι good risk από τους bad loss.

Στο ίδιο dataset χρησιμοποιήσαμε κι άλλες τεχνικές για την αναγνώριση χαρακτηριστικών και γενικά προτύπων για την κατάταξη πελατών σε μια από τις 3 κατηγορίες της μεταβλητής Risk. Μία από αυτές είναι και η ανάλυση κατά συστάδες. Παρατηρώντας τι χαρακτηριστικά έχει η κάθε ομάδα και συγκρίνοντας τα με τη μεταβλητή στόχο μπορέσαμε να δούμε για παράδειγμα από τι μέσο όρο ηλικίας και εισοδήματος αποτελείται η κάθε ομάδα. Επίσης ποια κατηγορία ανήκει σε κάθε ομάδα με μεγαλύτερο ποσοστό με αποτέλεσμα να μπορούν να διεξαχθούν συμπεράσματα πάλι για το προφίλ των πελατών.

Στους αλγορίθμους των δέντρων απόφασης και του νευρωνικού δικτύου είδαμε ένα πολύ καλό ποσοστό σωστής ταξινόμησης. Επιπρόσθετα χρησιμοποιήσαμε για την αξιολόγηση των μοντέλων gain και lift charts . Τα διαγράμματα αυτά είναι αρκετά δημοφιλή στην εξόρυξη δεδομένων για εφαρμογές μάρκετινγκ και γενικότερα σωστής επιλογής δείγματος για καμπάνιες. Για παράδειγμα ίσως ήθελε μια τράπεζα να αναγνωρίσει τους bad risk-profit πελάτες της και να πάρει ένα μεγάλο ποσοστό αυτών για να τους προτείνει κάποιο επιπρόσθετο προϊόν με σκοπό την αντιστάθμιση κινδύνου.

Συμπερασματικά λοιπόν, ο πιστωτικός κίνδυνος αποτελεί τον πιο σημαντικό κίνδυνο για τις τράπεζες και έτσι οι τράπεζες πρέπει να διερευνούν συνεχώς και να αναλύουν δεδομένα, με στόχο τη διασφάλιση πιστοληπτικά αξιόπιστων πελατών. Άρα οι τεχνικές πολυμεταβλητής ανάλυσης φαίνονται να είναι πολύ χρήσιμα εργαλεία στον τραπεζικό τομέα.

Όχι όμως μόνο σ αυτό τον τομέα, ούτε μόνο στην περίπτωση πιστωτικού ελέγχου, αλλά σε πολλές διαφορετικές περιπτώσεις, εταιρίες, οργανισμούς, βιομηχανίες. Αυτό μπορεί να διαπιστωθεί από τις περιπτώσεις που αναφέρονται στο Κεφάλαιο 4, αλλά και από το $2^{ο}$ dataset που χρησιμοποιήσαμε για ανάλυση, το οποίο αφορά την πρόβλεψη στερεών αποβλήτων, σε σχέση ουσιαστικά με λιανεμπόριο, χονδρεμπόριο, χώρους εξόρυξης μετάλων, εστιατόρια-ξενοδοχεία. Χρησιμοποιήσαμε πολλαπλή παλινδρόμηση με στόχο την ανακάλυψη σημαντικών παραγόντων που επηρεάζουν την αύξηση των στερεών αποβλήτων. Σε $2^{η}$ φάση εφαρμόσαμε παραγοντική ανάλυση και καταλήξαμε σε 2 σημαντικούς παράγοντες τη χρήση γης από βαριά βιομηχανία και τη χρήση γης από τομείς παροχής υπηρεσιών, οι οποίοι μας έδωσαν ένα πιο σταθερό μοντέλο.

Βιβλιογραφία

Ελληνική

Ηλιόπουλος Γ. (2017). Γραμμικά και Γενικευμένα Γραμμικά Μοντέλα (Σημειώσεις)

Καρλής Δ. (2005). Πολυμεταβλητή στατιστική ανάλυση

Κούτρας Μ. (2017). Εφαρμοσμένη Πολυμεταβλητή Ανάλυση (Σημειώσεις)

Μηλιώνης Α. (2014). Ανάλυση Παλινδρόμησης Πανεπιστήμιο Αιγαίου, Σάμος

Πετρίδης Δ. (2015). Ανάλυση πολυμεταβλητών τεχνικών

Ξένη

http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf
Aungkana Wungwanitchakorn, Adoption intention of banks' customers on internet, September – December 2002, pp. 63-80

Dr. H. S. Sandhu, Ms. Neetu Bala (2011).Customers' Perception towards Service Quality of Life Insurance Corporation of India: A Factor Analytic Approach

Fanliang Bu, Qingmei XieResearch on Emergency Evacuation Traffic Trip Generation Forecasting Based on Logistic Regression, University Beijing, China

Hanudin Amin Ricardo Baba Mohd Zulkifli Muhammad. An analysis of mobile banking acceptance by malaysian customers, University of Malaysia

Joannes El. Chliaoutakis , Charalambos Gnardellis , Ismini Drakou , Christina Darviri , Vickey Sboukis (2000). Modelling the factors related to the seatbelt use by the young drivers of Athens p. 815–825

Julianti Kasih, Sani Susanto (2012). Predicting students' final results through discriminant analysis

Kasey L. Morris, Frank M. Perna (2018). Decision Tree Model vs Traditional Measures to Identify Patterns of Sun-Protective Behaviors and Sun Sensitivity Associated With Sunburn.

Lingxiao Tang and Jia Sun (2019). Predict the sales of New-energy Vehicle using linear regression analysis

Paula Odete Fernandes, Hélder Pires Ferreira, Tourism & Management Studies (2015), p. 164-172, Identification of critical success factors that maximise customers' satisfaction: Multivariate Analysis

Qian Su, Peiji Shao and Quanfu Ye (2012). The analysis on the determinants of mobile VIP customer churn: a logistic regression approach

ScienceDirect, Factor Analysis

Wikipedia

Παράρτημα

ΑΠΟΔΕΙΞΗ *

Για p=1

ΘΕΩΡΗΜΑ 1

Αν,

$$\frac{f_1(x)}{f_2(x)} \geq \left(\frac{c(1/2)}{c(2/1)}\right)\cdot\left(\frac{P_2}{P_1}\right)$$

κατατάσσω την παρατήρηση στην ομάδα Π_1 , διαφορετικά στην ομάδα Π_2.

ΑΠΟΔΕΙΞΗ

Είναι :

$$ECM = P_1 \cdot c(2/1) \cdot P(2/1) + P_2 \cdot C(1/2) \cdot P(1/2)$$

$$ECM = P_1 \cdot c(2/1) \cdot \int_{R_2} f_1(x)\,dx + P_2 \cdot C(1/2) \cdot \int_{R_1} f_2(x)\,dx$$

Επειδή:

$$\Omega = R_1 \cup R_2 \ \text{έχουμε} \int_{R_2} f_1(x)\,dx = 1 - \int_{R_1} f_1(x)\,dx$$

άρα:

$$ECM = P_1 \cdot c(2/1) \cdot \left[1 - \int_{R_1} f_1(x)\,dx\right] + P_2 \cdot c(1/2) \cdot P(1/2)$$

$$ECM = P_1 \cdot c(2/1) - P_1 \cdot c(2/1) \cdot \int_{R_1} f_1(x)\,dx + P_2 \cdot c(1/2) \cdot \int_{R_1} f_2(x)\,dx$$

$$ECM = P_1 \cdot c(2/1) - \int_{R_1} P_1 \cdot c(2/1) \cdot f_1(x)\,dx + \int_{R_1} P_2 \cdot c(1/2) \cdot f_2(x)\,dx$$

$$ECM = P_1 \cdot c(2/1) + \int_{R_1} \left[P_2 \cdot c(1/2) \cdot f_2(x) - P_1 \cdot c(2/1) \cdot f_1(x)\right] dx$$

Επειδή :

P_1 , P_2 ,c(2/1) ,c(1/2) μη αρνητικά και $f_1(x)$, $f_2(x)$ μοναδικές ποσότητες στο ECM που εξαρτώνται από το x και μη αρνητικές ,άρα το ECM θα ελαχιστοποιηθεί όταν η περιοχή R_1 πάρει εκείνες τις τιμές του διανύσματος x για τις οποίες το παραπάνω ολοκλήρωμα θα γίνει μη αρνητικό.

Δηλαδή:

$$\int_{R_1} \left[P_2 \cdot c(1/2) \cdot f_2(x) - P_1 \cdot c(2/1)\, f_1(x) \right] dx \le 0$$

$$\Rightarrow P_2 \cdot c(1/2) \cdot f_2(x) \le P_1 \cdot c(2/1)\, f_1(x)$$

$$\Rightarrow \frac{f_1(x)}{f_2(x)} \ge \left[\frac{c(1/2)}{c(2/1)} \right] \left(\frac{P_2}{P_1} \right) \quad (1)$$

Οπότε αν ισχύει η (1) θα κατατάξουμε την παρατήρηση στην 1η ομάδα αλλιώς στη 2η.

*ΑΠΟΔΕΙΞΗ***

Την παρατήρηση x_0 την κατατάσσουμε στον 1ο πληθυσμό εάν:

$$-\frac{1}{2}(x_0-\mu_1)'\,\Sigma^{-1}(x_0-\mu_1)+\frac{1}{2}(x_0-\mu_2)'\,\Sigma^{-1}(x_0-\mu_2)\ge k$$

$$\Rightarrow -\frac{1}{2}(x_0-\mu_1)'\,\Sigma^{-1}x_0+\frac{1}{2}(x_0-\mu_1)'\Sigma^{-1}\mu_1+\frac{1}{2}(x_0-\mu_2)'\Sigma^{-1}x_0-\frac{1}{2}(x_0-\mu_2)'\Sigma^{-1}\mu_2\ge k$$

$$\Rightarrow \frac{1}{2}\left[-(x_0-\mu_1)'+(x_0-\mu_2)'\right]\Sigma^{-1}x_0+\frac{1}{2}(x_0-\mu_1)'\Sigma^{-1}\mu_1-\frac{1}{2}(x_0-\mu_2)'\Sigma^{-1}\mu_2\ge k$$

$$\Rightarrow \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}x_0+\frac{1}{2}x_0'\Sigma^{-1}\mu_1-\frac{1}{2}\mu_1'\Sigma^{-1}\mu_1-\frac{1}{2}x_0'\Sigma^{-1}\mu_2+\frac{1}{2}\mu_2'\Sigma^{-1}\mu_2\ge k$$

$$\Rightarrow \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}x_0+\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}x_0-\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)\ge k$$

$$\Rightarrow (\mu_1-\mu_2)'\Sigma^{-1}x_0-\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)\ge k$$

Θέτουμε $L=\Sigma^{-1}(\mu_1-\mu_2)$ *και έχουμε εάν* $L'X-\frac{1}{2}L'(\mu_1+\mu_2)\ge k$