ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

ΑΝΑΛΥΣΗ ΤΕΧΝΙΚΩΝ PREDICTIVE ANALYTICS ΣΤΗ ΣΧΕΔΙΑΣΗ ΠΡΟΙΟΝΤΩΝ ΚΑΙ ΥΠΗΡΕΣΙΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΡΕΝΟΥ ΜΩΡΑΙΤΗ

Επιβλέπων : Δημήτριος Ασκούνης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Ανάλυση Τεχνικών Predictive Analytics στη Σχεδίαση Προϊόντων και Υπηρεσιών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΡΕΝΟΥ ΜΩΡΑΙΤΗ

Επιβλέπων : Δημήτριος Ασκούνης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 22η Ιουλίου 2015.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Ασκούνης	Βασίλειος Ασημακόπουλος	Ιωάννης Ψαρράς
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2015

(Υπογραφή)

.....

ΡΕΝΟΣ ΜΩΡΑΙΤΗΣ

Προπτυχιακός Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Τα τελευταία χρόνια, με την άνοδο του Web 2.0 και των μέσων κοινωνικής δικτύωσης, εμφανίζεται η πρόκληση της αξιοποίησης του τεράστιου αριθμού δεδομένων που κυκλοφορούν σε αυτά, προς όφελος της επιχειρηματικότητας. Πλέον, για να είναι μια επιχείρηση ανταγωνιστική, θα πρέπει να είναι σε θέση να εξάγει πληροφορία από αυτά τα δεδομένα, και να την αξιοποιεί στην λήψη αποφάσεων. Η παρούσα διπλωματική εργασία ασχολείται με την εφαρμογή τεχνικών Predictive Analytics που αξιοποιούν αυτά τα δεδομένα, και αποσκοπούν στην βελτιστοποίηση της σχεδίασης προϊόντων και υπηρεσιών. Αρχικά, γίνεται μια επισκόπηση των διαδεδομένων και αποτελεσματικών εφαρμογών που έχουν τα Predictive Analytics σε τομείς όπως η προώθηση και οι πωλήσεις. Έπειτα, η διαδικασία της σχεδίασης αναλύεται ώστε να αποφανθούν οι μεταβλητές τις οποίες θα καλούνται να συσχετίσουν οι εφαρμογές των Predictive Analytics. Αναφέρονται παραδείγματα επιτυχημένων εφαρμογών Predictive Analytics από επιχειρήσεις στον τομέα των μέσων κοινωνικής δικτύωσης και της σχεδίασης. Στο επόμενο στάδιο αναλύεται η διαδικασία της δημιουργίας ενός προβλεπτικού μοντέλου, από τον καθορισμό του σκοπού του και την συλλογή δεδομένων, μέχρι και τις μετρικές και τεχνικές για την αξιολόγηση του. Τα εργαλεία και οι αλγόριθμοι αξιολογούνται με βάση ένα σύνολο από έρευνες που μελετήθηκαν, και εξάγονται συμπεράσματα για την αξιοποίηση των δεδομένων των μέσων κοινωνικής δικτύωσης σε αυτά. Τέλος, προτείνονται κάποια προβλεπτικά μοντέλα αποκλειστικά για τη σχεδίαση προϊόντων και υπηρεσιών, και συγκρίνονται με ήδη υπάρχοντα.

Λέξεις Κλειδιά: <<Predictive Analytics, Μέσα Κοινωνικής Δικτύωσης, Σχεδίαση Προϊόντων και Υπηρεσιών, Προβλεπτικά Μοντέλα>>

Abstract

During the last years, with the rise of Web 2.0 and the social media, a new challenge has arised; using the massive amount of data circulating in the social media in order to benefit business management. In our days, for a company to be competitive, it should be able to extract useful information from that data, and use it in order to drive better decisions. This diploma thesis deals with the use of Predictive Analytics techniques that utilise data from social media in order to optimise the procces of planning of products and services. Firstly, a survey on the most renowned and effective uses of Predictive Analytics in marketing and sales is conducted. The process of planning is analysed in order to decide the variables that the Predictive Analytics techniques will be demanded to correlate. Real examples of succesful Predictive Analytics uses in the social media and planning sectors are described. In the next part, the process of building a predictive model is analysed, from deciding its purpose and collecting the needed data, to the metrics and techniques used to evaluate the model. The tools and algorithms are then evaluated, based on a number of surveys that were studied, and conclusions about the use of social media data are made. Finally, a number of predictive models specifically for the process of planning products and services are proposed, and are compared to already existing ones.

Keywords: <<Predictive Analytics, Social Media, Planning of Products and Services, Predictive Modeling>>

Πίνακας περιεχομένων

1 Εισαγωγή

1.1 Εισαγωγή στα Predictive Analytics

Καθώς διανύουμε την εποχή της πληροφορίας, ξεπερνάμε τα ιστορικά βιβλία και τις εγκυκλοπαίδειες, και χρησιμοποιούμε πλέον συστήματα που καταγράφουν κάθε κλικ, πληρωμή, κλήση, ατύχημα, έγκλημα και αρρώστια. Αλλά αυτή η φαινομενικά άπειρη ποσότητα πληροφορίας δεν περιέχει τα γεγονότα, η επίγνωση των οποίων θα ήταν πολύτιμη, τα γεγονότα δηλαδή που δεν έχουν συμβεί ακόμα. Χρησιμοποιώντας Predictive Analytics, θα ήταν υπερβολή να ισχυριστούμε ότι μπορούμε να προβλέψουμε το μέλλον με ακρίβεια. Όμως, αυτό που ισχύει αναμφίβολα είναι ότι έστω και μία αιτιολογημένη πρόβλεψη μικρού βαθμού, μπορεί να μας ωφελήσει πολύ περισσότερο από το να προσπαθούμε απλά να "μαντέψουμε" τι πρόκειται να συμβεί. Δεν απαιτείται να πετύχουμε το ακατόρθωτο και να αποκτήσουμε αψεγάδιαστη εικόνα του μέλλοντος. Η πιθανολόγηση των διαφόρων ενδεχομένων, με σκοπό απλά να ξεδιαλύνουμε την εικόνα που έχουμε για το μέλλον, μεταφράζεται σε άμεσα κέρδη. Με αυτό τον τρόπο, τα Predictive Analytics αντιμετωπίζουν το οικονομικό ρίσκο, ενισχύουν την αντιμετώπιση παθήσεων, βοηθάνε την αντιμετώπιση εγκληματικής συμπεριφοράς, και αυξάνουν της πωλήσεις. Απαντώντας δηλαδή, σε ερωτήματα όπως: πόσες πιθανότητες έχει ο ασθενής να κάνει επιτυχής χειρουργική επέμβαση; ο πελάτης να ανταποκριθεί θετικά αν του σταλθεί ένα διαφημιστικό φυλλάδιο; ο κάτοχος οικείας να αντιμετωπίσει μια χαμηλή υποθήκη;

Χτισμένα πάνω στην Επιστήμη Υπολογιστών και στην στατιστική, και ενισχυμένα από αποκλειστικά συνέδρια και πανεπιστημιακά πτυχιακά προγράμματα, τα Predictive Analytics έχουν αναδειχθεί σε ένα ξεχωριστό κλάδο. Εκτός όμως από επιστημονικό κλάδο,

αποτελούν και ένα νέο φαινόμενο που ασκεί πολύ ισχυρό αντίκτυπο. Εκατομμύρια αποφάσεις καθημερινά, αναδεικνύουν ποια άτομα πρέπει να κληθούν, να εγκριθούν, να διαγνωστούν, να προειδοποιηθούν και να ερευνηθούν. Τα Predictive Analytics είναι το μέσο με το οποίο οδηγούμε ατομικές εμπειρικές αποφάσεις, αξιοποιώντας δεδομένα. Απαντώντας το σύνολο που αποτελείται από όλες αυτές τις μικρότερες ερωτήσεις, τα Predictive Analytics τελικά είναι ικανά να απαντήσουν το πολύ σημαντικό ερώτημα του πώς μπορούμε να βελτιώσουμε την αποτελεσματικότητα όλων αυτών των μαζικών λειτουργιών στην κυβέρνηση, στην πρόνοια, στην επιχειρηματικότητα, και στο νομικό σύστημα.

Με αυτή την έννοια, τα Predictive Analytics αποτελούν μια τελείως διαφορετική οντότητα από τις Τεχνικές Προβλέψεων. Οι Τεχνικές Προβλέψεων εκτελούν συναθροιστικές προβλέψεις σε μακροσκοπικό επίπεδο. Ενώ οι Τεχνικές Προβλέψεων παράγουν εκτιμήσεις για τον αριθμό των προϊόντων που πρόκειται να πουλήσουμε, τα Predictive Analytics μας δείχνουν ποιοι καταναλωτές είναι πιο πιθανόν να αγοράσουν το προϊόν μας σε ατομικό επίπεδο. Τα Predictive Analytics ακολουθούν την αυξανόμενη τάση σύμφωνα με την οποία οι αποφάσεις πρέπει να είναι πιο “data driven”, δηλαδή να βασίζονται λιγότερο στο ένστικτό μας, και περισσότερο σε απτές, εμπειρικές αποδείξεις.

Μία οργάνωση ή επιχείρηση, μπορεί να θεωρηθεί ως ένα “μεγάλο” άτομο, επομένως θα πρέπει να μαθαίνει με τον ίδιο ρυθμό. Μία ομάδα δημιουργείται με σκοπό το ομαδικό κέρδος των μελών της και αυτών που υπηρετεί, είτε είναι μια εταιρεία, μια κυβέρνηση, ένα νοσοκομείο ή ένα πανεπιστήμιο. Αφού δημιουργηθεί, επωφελείται από την εργασία, τις ικανότητες, και την αποτελεσματικότητα της μαζικής παραγωγής. Όπως ένας πωλητής μαθαίνει με την πάροδο του χρόνου από τις αρνητικές και θετικές εμπειρίες του, τις επιτυχίες και τις αποτυχίες του, τα Predictive Analytics αποτελούν την διαδικασία μέσω της οποίας μία επιχείρηση μαθαίνει από την εμπειρία που έχει συλλέξει από τα μέλη της και τα υπολογιστικά της συστήματα. Συγκεκριμένα, μία επιχείρηση που δεν αξιοποιεί τα δεδομένα της με αυτό τον τρόπο, μπορεί να παρομοιαστεί με ένα άτομο που έχει φωτογραφική μνήμη αλλά δεν την αξιοποιεί.

Με λίγες εξαιρέσεις, οι επιχειρήσεις, ωφελούνται από την χρήση των Predictive Analytics. Αυτό συμβαίνει επειδή οι επιχειρήσεις είναι αυτές που καλούνται να λάβουν πληθώρα αποφάσεων, για τις οποίες υπάρχει μεγάλο περιθώριο βελτίωσης. Γενικότερα, οι επιχειρήσεις και οι οργανισμοί είναι έμφυτα μη αποδοτικές και σπάταλες σε μεγάλο βαθμό. Για παράδειγμα οι ενέργειες προώθησης μέσω μαζικών email χαρακτηρίζονται ως σπαταλημένο κεφάλαιο αφού πλέον εκτιμάται ότι 80% όλων των email κατηγοριοποιείται αυτόματα στην κατηγορία spam. Σε άλλες περιπτώσεις δίνεται μεγάλη αξία σε επικίνδυνους οφειλέτες, ενώ βάσιμες αιτήσεις για κρατικά επιδόματα καθυστερούν ή αγνοούνται.

Στον εμπορικό τομέα, το κέρδος είναι κινητήρια δύναμη. Είναι εύκολο να φανταστούμε τα κίνητρα πίσω από το να κάνουμε τις καθημερινές διαδικασίες πιο αποδοτικές, την προώθηση πιο ακριβής, να αναγνωρίζουμε πιο εύκολα τις απάτες, να αποφεύγουμε κακούς οφειλέτες και να ελκύουμε περισσότερους πελάτες μέσω του διαδικτύου. Βελτιστοποιώντας δηλαδή τις διάφορες λειτουργίες, όπου κάνει την μεγαλύτερη διαφορά, τα Predictive Analytics αναβαθμίζουν την επιχειρηματικότητα.

Η πρόβλεψη συνήθως ξεκινάει σε μικρό επίπεδο. Το θεμελιώδες στοιχείο των Predictive Analytics είναι η προγνωστική μεταβλητή, μια απλή τιμή, που μετράμε για κάθε άτομο ή παρατήρηση. Για παράδειγμα, η προσφατότητα (recency), δηλαδή ο αριθμός των εβδομάδων που έχουν περάσει από την τελευταία φορά που ένα άτομο έκανε μια αγορά, διέπραξε ένα έγκλημα, ή εξέφρασε ένα σύμπτωμα υγείας, συχνά αποκαλύπτει τις πιθανότητες να ξαναεμφανίσει την ίδια συμπεριφορά στο κοντινό μέλλον. Σε πολλά πεδία, έχει περισσότερο νόημα να ξεκινήσουμε με τους πιο πρόσφατα δραστήριους ανθρώπους, παρά με έρευνα και αξιολόγηση όλων των πιθανών περιπτώσεων.

Παρόμοια, η συχνότητα, δηλαδή ο αριθμός των φορών που το άτομο έχει εμφανίσει την συμπεριφορά, είναι ένα αποτελεσματικό μέγεθος που χρησιμοποιείται σε πολλές περιπτώσεις. Η κοινή λογική λέει ότι άνθρωποι που έχουν κάνει κάτι πολλές φορές, έχουν μεγάλες πιθανότητες να το ξανακάνουν.

Ειδικότερα, είναι συνηθισμένο να προβλέπεται πώς θα συμπεριφερθεί ένα άτομο, με βάση το πώς έχει συμπεριφερθεί στο παρελθόν. Έτσι τα Predictive Analytics αξιοποιούν τα δεδομένα που λανθασμένα δεν θεωρούνται προσοδοφόρα, όπως τοποθεσία και φύλο, για να δημιουργήσουν προβλέψεις συμπεριφοράς όπως η συχνότητα, η προσφατότητα (recency), οι μελλοντικές αγορές, η οικονομική δραστηριότητα και η χρήση προϊόντων και διαδικτύου. Αυτές οι συμπεριφορές είναι συχνά οι πιο πολύτιμες, καθώς στην πλειοψηφία των περιπτώσεων, η συμπεριφορά είναι αυτό που προσπαθούμε να προβλέψουμε, και είναι γεγονός ότι η παρελθοντική συμπεριφορά προβλέπει την μελλοντική. Τα Predictive Analytics αντλούν την δύναμή τους συνδυάζοντας δεκάδες προβλέψεις. Σε τελικό στάδιο, ένα μοντέλο παίρνει σαν είσοδο όλα τα δεδομένα που έχουμε για τις παρατηρήσεις μας, και σαν έξοδο μας δίνει πιθανότητες για όλα τα ενδεχόμενα.

Μερικές αμφιλεγόμενες προβλέψεις παρουσιάζουν πολύ μεγάλο ενδιαφέρον. Μελέτες έχουν δείξει ότι οι πελάτες είναι πιο προσοδοφόροι όταν δεν σκέφτονται, ότι η εγκληματική δραστηριότητα αυξάνεται μετά από αθλητικές εκδηλώσεις, ότι η πείνα μπορεί να επηρεάσει την απόφαση κάποιου δικαστή, ακόμα και ότι μία προαγωγή μπορεί να προκαλέσει παραίτηση. Για όλους αυτούς τους συσχετισμούς υπάρχει μια προτεινόμενη εξήγηση. Συγκεκριμένα, για την σχέση προαγωγής με παραίτηση, προτεινόμενη εξήγηση

είναι ότι οι αυξημένες ευθύνες γίνονται αντιληπτές ως παραπάνω φορτίο εφ' όσον δεν ανταμείβονται ανάλογα. Σε όλες αυτές τις περιπτώσεις όμως, η εξήγηση είναι πάντα "προτεινόμενη" αποτελούν δηλαδή εικασίες, χωρίς απτές αποδείξεις.

Το δίλημμα που προκύπτει εδώ, είναι αν τελικά ο συσχετισμός αποδεικνύει αιτιότητα. Τελικά όμως η απόδειξη μιας σχέσης ανάμεσα σε δύο γεγονότα, δεν σημαίνει ότι το ένα προκαλεί το άλλο, είτε άμεσα είτε έμμεσα. Αυτό προκύπτει από το γεγονός ότι και τα δύο αυτά συμβάντα μπορεί να προκαλούνται από το ίδιο αίτιο, αλλά το γεγονός που μπορεί λανθασμένα να θεωρηθεί σαν αίτιο, να παρατηρείται πριν από το δεύτερο. Για παράδειγμα, ας υποθέσουμε ότι έχει παρατηρηθεί προβλεπτική σχέση ανάμεσα στις πωλήσεις γυαλιών ηλίου και την αύξηση πωλήσεων σε ακτοπλοϊκά εισητήρια. Μία σαφώς λανθασμένη υπόθεση-εξήγηση θα ήταν ότι τα γυαλιά ηλίου προωθούνται συνεργατικά με τα ακτοπλοϊκά εισητήρια, ή ότι προσφέρονται πακέτα προσφοράς που συμπεριλαμβάνουν και τα δύο. Σε αυτό το παράδειγμα είναι εύκολο να καταλάβουμε ότι τελικά δεν υπάρχει σχέση ανάμεσα σε αυτά τα δύο γεγονότα, αλλά είναι αποτελέσματα των καλών καιρικών συνθηκών. Τελικά όμως, δεδομένου ότι η αρχική σχέση που αναφέρθηκε στο παράδειγμα είναι αληθής, είναι εφικτό να προβλέψουμε αύξηση πωλήσεων ακτοπλοϊκών εισητηρίων με βάση την αύξηση των γυαλιών ηλίου, ανεξάρτητα από το αν τελικά το ένα γεγονός προκαλεί το άλλο.

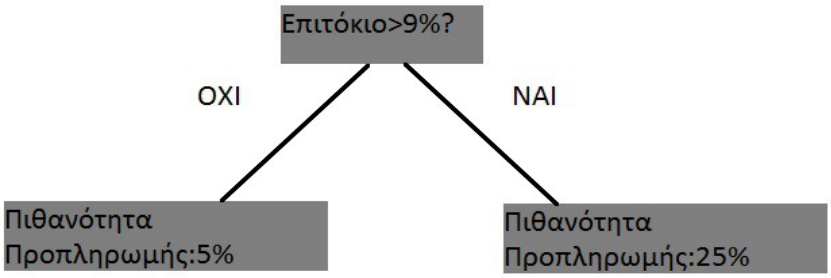
Όταν εφαρμόζουμε Predictive Analytics, στις περισσότερες των περιπτώσεων δεν καταλαβαίνουμε ξεκάθαρα την αιτιότητα, και συχνά δεν μας ενδιαφέρει. Για τις εφαρμογές των Predictive Analytics, ο σκοπός είναι περισσότερο να προβλέψουμε, παρά να εξηγήσουμε το πώς λειτουργεί το αντίστοιχο περιβάλλον και να εξηγήσουμε ποιο γεγονός προκαλεί ποιο. Η αιτιότητα είναι δύσκολο να συλληφθεί και να οριστεί ξεκάθαρα. Γενικότερα υποθέτουμε ότι κάποια πράγματα επηρεάζονται μεταξύ τους με κάποιο τρόπο, και εξηγούμε αυτά τα φαινόμενα με φυσικούς, χημικούς, ιατρικούς, οικονομικούς ή ψυχολογικούς όρους. Αντίθετα, στον χώρο των Predictive Analytics, η πρόβλεψη έχει μεγαλύτερη αξία από την αιτιολόγηση, αφού τα μοντέλα σχεδιάζονται με απόλυτο γνώμονα την λύση του προβλήματος, και μας ενδιαφέρει μόνο να φέρουν αποτελέσματα. Αν και τελικά σε πολλές περιπτώσεις τα Predictive Analytics προσφέρουν εξηγήσεις αιτιότητας παρόμοιες με αυτές διαφόρων κοινωνικών επιστημών, αυτές αποτελούν υποπροϊόντα της λειτουργίας τους, και όχι τον κύριο σκοπό τους.

Λαμβάνοντας αυτά υπ' όψη, τα Predictive Analytics μπορούν να θεωρηθούν ως μία ξεχωριστή επιστήμη, η οποία εφαρμόζεται σε δεύτερο χρόνο, υπερπηδώντας τα σύνορα των φυσικών και κοινωνικών επιστημών, μαθαίνοντας και συνδυάζοντας πληροφορίες από πολλές και διαφορετικές πηγές δεδομένων που κανονικά θα υπάγονταν στον δικό τους επιστημονικό κλάδο, είτε αυτός είναι ο βιολογικός, ο εγκληματικός, ο οικονομικός, ο

φαρμακευτικός, ο ψυχολογικός η ο κοινωνιολογικός. Τελικά, τα Predictive Analytics έχουν ως αποστολή την δημιουργία λύσεων, και αξιοποιούν οποιαδήποτε πληροφορία μπορούν για να την πετύχουν.

Η τεχνολογία των Predictive Analytics πλέον εφαρμόζεται σε πολλά πεδία και μας επηρεάζει όλους, καθημερινά. Επιδρά στις εμπειρίες μας, στις περισσότερες των περιπτώσεων χωρίς να το καταλαβαίνουμε, καθώς οδηγούμε, ψωνίζουμε, μελετάμε η επικοινωνούμε. Η εξαγωγή προβλέψεων αποτελεί μια μεγάλη πρόκληση. Κάθε πρόβλεψη βασίζεται σε πολλούς παράγοντες: τα διάφορα γνωστά χαρακτηριστικά του ασθενή, του κάτοχου οικίας, και κάθε mail. Εμφανίζεται το ερώτημα του πώς θα αντιμετωπίσουμε το άμεσο πρόβλημα του να συνδυάσουμε όλα αυτά τα κομμάτια δεδομένων ώστε να καταφέρουμε να εξάγουμε πρόβλεψη. Ενώ η ιδέα αρχικά εμφανίζεται απλή, δεν είναι εύκολη στην υλοποίηση. Το πρόβλημα τελικά αντιμετωπίζεται με ένα επιστημονικό και συστηματικό μέσο το οποίο αναπτύσσει και συνεχώς βελτιώνει προβλέψεις, μαθαίνει δηλαδή κυριολεκτικά, πώς να προβλέπει. Αυτή η λύση ονομάζεται μηχανική εκμάθηση, η χρήση δηλαδή αλγορίθμων με σκοπό να αναπτύσσουν νέα γνώση και δυνατότητες αυτόματα, έχοντας ως τροφή την πιο δυναμική και αστείρευτη πηγή της σύγχρονης κοινωνίας, η οποία είναι τα δεδομένα.

Για να δείξουμε ένα σχετικά απλό παράδειγμα λειτουργίας, θα υποθέσουμε ότι διευθύνουμε μία τράπεζα η οποία διαχειρίζεται έναν μεγάλο αριθμό από δάνεια. Θεωρούμε το ενδεχόμενο της προπληρωμής, την περίπτωση δηλαδή ο πελάτης να κάνει μια απρόβλεπτη πληρωμή ολόκληρου του χρέους του, στερώντας την τράπεζα από μελλοντικές εισπράξεις τόκων. Μελετώντας το ιστορικό των δανείων μας, παρατηρούμε ότι στα δάνεια που το επιτόκιο ήταν μικρότερο του 9%, μόνο το 5% των πελατών μας προπλήρωσε το χρέος του. Αντίστοιχα, στα δάνεια με επιτόκιο μεγαλύτερο του 9%, οι πελάτες που έκαναν προπληρωμή αποτελούσαν το 25%.Αυτόματα, δεδομένου ότι μας ενδιαφέρει πάντα το ενδεχόμενο προπληρωμής, με βάση το ιστορικό της τράπεζας, μπορούμε να χωρίσουμε το παροντικό και το μελλοντικό σύνολο των δανείων σε δύο μεγάλες ομάδες, με την μία να παρουσιάζει 5 φορές μεγαλύτερο κίνδυνο:



Αυτή η ανακάλυψη είναι πολύτιμη, αν και προφανής καθώς εμπειρικά μπορούμε να καταλάβουμε ότι όσο μεγαλύτερο επιτόκιο αναγκάζεται να πληρώνει ο δανειολήπτης, τόσο μεγαλύτερο κίνητρο έχει να προπληρώσει το χρέος του. Ανεξάρτητα όμως από το αν είναι προφανής, πλέον έχει ελεγχθεί με βάση το ιστορικό της τράπεζας, και το φαινόμενο πλέον είναι ποσοτικοποιημένο με ακρίβεια. Αυτό μπορεί να θεωρηθεί ως το πρώτο βήμα της μηχανικής εκάθησης στο πρόβλημα της τράπεζάς μας.

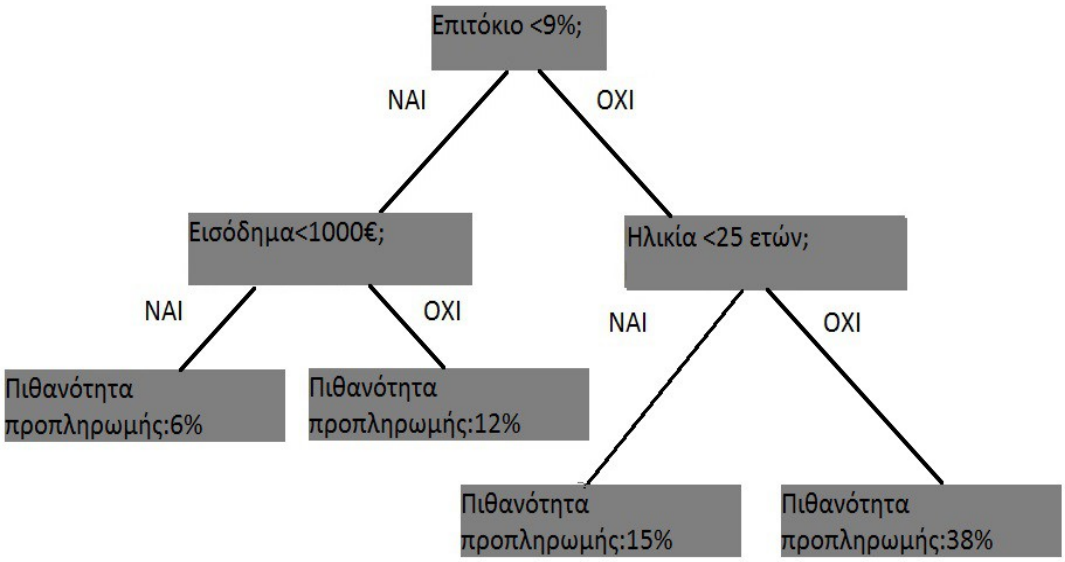
Μέχρι τώρα η διορατικότητα που έχουμε αποκτήσει είναι ότι το επιτόκιο μπορεί να προβλέψει το ρίσκο, και το μοντέλο μας είναι πολύ βασικό. Τοποθετεί το κάθε δάνειο σε μία από τις δύο κατηγορίες υψηλού και χαμηλού ρίσκου. Καθώς το μοντέλο μας λαμβάνει υπ' όψη του μόνο μία μεταβλητή, μπορούμε να το χαρακτηρίσουμε ως μονομεταβλητό. Προφανώς όλα τα παραδείγματα εφαρμογών Predictive Analytics που έχουν αναφερθεί μέχρι τώρα και πρόκειται να αναφερθούν, λαμβάνουν υπ' όψη τους πολύ μεγαλύτερο αριθμό μεταβλητών, για να μπορέσουν να προβλέψουν ότι τους ζητείται. Επομένως για να είναι και το δικό μας μοντέλο αποτελεσματικό, θα πρέπει να μετατραπεί σε πολυμεταβλητό. Ανατρέχοντας στα στοιχεία του πελάτη μας, μπορούμε να αντλήσουμε διάφορες ενδιαφέρουσες πληροφορίες, όπως το εισόδημά του, την συνδυασμένη αξία του, τον αριθμό των φορών που έχει καθυστερήσει την δόση του δανείου του, και άλλες πληροφορίες σχετικά με το επάγγελμά του. Όλα αυτά τα στοιχεία μπορούν να χαρακτηριστούν ως μεταβλητές για το μοντέλο μας. Τελικά ο σκοπός του μοντέλου θα είναι να αναλύσει όλες αυτές τις μεταβλητές και να τις συνυπολογίσει σε ένα τελικό νούμερο πιθανότητας προπληρωμής. Αυτή η διαδικασία αποτελεί βασικά και την πρόκληση της μηχανικής εκμάθησης. Ο σκοπός μας είναι να προγραμματίσουμε το σύστημά μας να αναλύει σε βάθος τα δεδομένα που

έχουμε για κάθε άτομο ξεχωριστά, και αυτόματα να δημιουργεί ένα πολυμεταβλητό προβλεπτικό μοντέλο.

Αυτή τη στιγμή έχουμε δύο ομάδες δανειοληπτών. Θα προσπαθήσουμε να βρούμε μία μεταβλητή με βάση την οποία η ομάδα χαμηλού κινδύνου μπορεί να διαιρεθεί σε δύο υποομάδες, και θα κάνουμε το ίδιο και για την ομάδα υψηλού κινδύνου. Ύστερα μπορούμε να εφαρμόσουμε την ίδια τακτική στις υποομάδες μας, εφαρμόζοντας δηλαδή μια τεχνική "διαίρει και βασίλευε".Όμως οι παρελθοντικές εφαρμογές έχουν δείξει ότι όσον αφορά τα Predictive Analytics, είναι καλύτερα να σταματάμε την υποδιαίρεση τον ομάδων από ένα σημείο και μετά.

Αυτή η μέθοδος, είναι γνωστή ως δέντρα αποφάσεων, και δεν αποτελεί την μόνη μέθοδο δημιουργίας προβλεπτικών μοντέλων, αλλά γενικότερα θεωρείται από τις πιο πρακτικές, αφού συνδυάζει απλότητα και αποτελεσματικότητα σε ικανοποιητικό επίπεδο. Συνήθως δεν δημιουργεί τα πιο ακριβή προβλεπτικά μοντέλα, αλλά καθώς η απεικόνισή τους είναι πολύ πιο κατανοητή από τις συμπαγείς μαθηματικές μεθόδους που χρησιμοποιούνται σε άλλες περιπτώσεις, είναι καλό να ξεκινήσουμε από αυτήν.

Επιστρέφοντας στο παράδειγμά μας, ας αναζητήσουμε μια μεταβλητή με βάση την οποία μπορούμε να χωρίσουμε τις ομάδες μας σε υποομάδες. Ανατρέχοντας πάλι στο ιστορικό των δανείων μας, διαπιστώνουμε ότι στα δάνεια με επιτόκιο μικρότερο του 9%, και στα οποία οι δανειολήπτες είχαν μηνιαίο εισόδημα μικρότερο των 1000€, έγινε προπληρωμή σε ποσοστό 6%, ενώ οι δανειολήπτες που είχαν μηνιαίο εισόδημα μεγαλύτερο των 1000€, προπλήρωσαν σε ποσοστό 12%.Αντίστοιχα, στην ομάδα δανείων με επιτόκιο μεγαλύτερο του 9%, οι δανειολήπτες ηλικίας κάτω των 30 ετών προπλήρωσαν σε ποσοστό 15%, ενώ οι δανειολήπτες άνω των 30 ετών σε ποσοστό 38%.Εδώ μπορεί να γίνει μια εξήγηση ότι όσο μεγαλύτερο μηνιαίο εισόδημα έχουν οι δανειολήπτες, τόσο μεγαλύτερη δυνατότητα προπληρωμής έχουν, και ότι η μεγαλύτερη ηλικία των δανειοληπτών συνεπάγεται καλύτερη κατανόηση του τραπεζικού συστήματος, επομένως και αναγνώριση του προσωπικού κέρδους που συνεπάγεται η προπληρωμή. Αξίζει να σημειωθεί ότι και οι δύο αυτές προτεινόμενες εξηγήσεις έχουν αρκετά μικρότερες πιθανότητες να είναι αληθείς από την εξήγηση που δώσαμε στην πρώτη τμηματοποίηση με βάση το επιτόκιο, αλλά όπως προαναφέρθηκε στον χώρο των Predictive Analytics δεν μας ενδιαφέρει να εξηγήσουμε την αιτιότητα. Μας ενδιαφέρει να χρησιμοποιήσουμε ιστορικά δεδομένα για να πιθανολογήσουμε μελλοντικά ενδεχόμενα. Το καινούργιο δέντρο θα είναι το εξής:



Παρατηρούμε ότι το δέντρο μεγαλώνει προς τα κάτω. Διαπιστώνουμε ότι και οι δύο καινούργιες μεταβλητές που επιλέξαμε έχουν τελικά σημασία στην πιθανότητα προπληρωμής, ανεξάρτητα από το αν αυτή είναι σχέση αιτιότητας. Μπορούμε ήδη να μπούμε στην διαδικασία εύρεσης των ομάδων με χαμηλότερο και υψηλότερο κίνδυνο.

Υποθέτοντας ότι συνεχίζουμε να συμπεριλαμβάνουμε και άλλες μεταβλητές στο δέντρο μας, αυτό θα συνεχίσει να μεγαλώνει προς τα κάτω και θα γίνει πιο περίπλοκο. Σαν γενικότερος κανόνας, όσο ένα δέντρο μεγαλώνει και γίνεται πολυπλοκότερο, μεγαλώνει και η προβλεπτική του ικανότητα, αλλά όχι με τον ίδιο ρυθμό. Για να αξιολογήσουμε την απόδοση των προβλεπτικών μοντέλων, χρησιμοποιούμε ένα μέγεθος που λέγεται lift. Πρακτικά, αυτό το μέγεθος εκφράζει πόσους παραπάνω στόχους μπορούμε να αναγνωρίσουμε χρησιμοποιώντας το εκάστοτε μοντέλο, σε σχέση με όσους θα αναγνωρίζαμε χωρίς να το χρησιμοποιήσουμε. Επιπλέον, το lift χρησιμοποιείται για να αποφευχθεί ένα κοινό αρνητικό φαινόμενο που παρατηρείται σε παρόμοιους αλγορίθμους, που λέγεται overlearning.

Διαπιστώνουμε λοιπόν, ότι όπως προαναφέρθηκε τα δέντρα αποφάσεων είναι πρακτικά και αποτελεσματικά, αφού χωρίς να χρειαστούν μαθηματικές πράξεις, και λαμβάνοντας υπ' όψη λίγες βασικές μεταβλητές, έχουμε ήδη δημιουργήσει τέσσερις ομάδες κινδύνου. Για κάθε δανειολήπτη, ξεκινάμε από την κορυφή του δέντρου (ρίζα), απαντάμε στις ερωτήσεις που παρουσιάζονται μέχρι να φτάσουμε σε ένα τελευταίο φύλλο, το οποίο μας δίνει την αριθμητική τιμή του κινδύνου.

Προεκτείνοντας φυσικά και έξω από τον κόσμο των οικονομικών, τα δέντρα αποφάσεων βρίσκουν εφαρμογές σχεδόν σε όλες τις επιστήμες, είτε αυτή είναι η φαρμακευτική, η νομική, η βιομηχανική ή οποιαδήποτε άλλη. Η διαδικασία εκμάθησης είναι ευέλικτη από την φύση της, αφού η ειδικότητα του εκάστοτε δέντρου αποφάσεων εξαρτάται αποκλειστικά από τα δεδομένα στα οποία "μεγαλώνει".Δίνοντας δεδομένα από μια άλλη επιστήμη, το σύστημα μαθαίνει για ένα εντελώς καινούργιο περιβάλλον.

1.2 Εισαγωγή στα Μέσα Κοινωνικής Δικτύωσης

Θα μπορούσε να ειπωθεί, ότι υπό τον γνώμονα των Predictive Analytics, ο αρχικός σκοπός για τον οποίο επιλέγουμε να δραστηριοποιηθούμε στα μέσα κοινωνικής δικτύωσης (Social Media), είναι η κατανόηση του μαζικού συναισθήματος των καταναλωτών, μέσα από απροκατάληπτη ανθρώπινη συμπεριφορά η οποία συμβαίνει στον πραγματικό κόσμο και όχι σε κάποιο εργαστήριο. Γενικότερα στα μέσα κοινωνικής δικτύωσης, οι άνθρωποι εκφράζουν τις απόψεις και τα συναισθήματά τους. Τα blogs, για παράδειγμα τα οποία αποτελούν μία από τις πιο διαδεδομένες μορφές μέσων κοινωνικής δικτύωσης, μετέτρεψαν τον μέχρι τώρα εσωστρεφή συγγραφέα ημερολογίου, σε άτομο που λαμβάνει ικανοποίηση από την εξωστρεφή δημοσίευση των ιδεών του, αφού με αυτόν τον τρόπο πλέον βρίσκει ομοϊδεάτες και υποστηρικτές. Στην σύγχρονη εποχή, υπάρχει ένας φαινομενικά άπειρος αριθμός φωνών που φέρει απόψεις ανεπηρέαστες από περιορισμούς , προκατάληψη ή φόβο για κριτική. Θεωρητικά, οι χρήστες αυτοί που εκπέμπουν μηνύματα στα μέσα κοινωνικής δικτύωσης, δημοσιοποιούν την άποψη που είναι η πιο κοντινή σε αυτή του καταναλωτικού κοινού.

Προφανώς, οι απόψεις και τα συναισθήματα των χρηστών, στις περισσότερες περιπτώσεις, δεν είναι εύκολο να κατανοηθούν άμεσα από τα λεγόμενα τους. Όπως προαναφέρθηκε, οι χρήστες εκφράζουν τις απόψεις και τα συναισθήματα τους μέσα από κάθε μήνυμα που δημοσιεύουν, αλλά για να εξάγουμε αυτή την χρήσιμη για εμάς πληροφορία, πρέπει να μελετήσουμε και να κατανοήσουμε την αντίστοιχη κατάσταση (status) στο Facebook, tweet στο Twitter, ή κείμενο σε blog. Γενικότερα τα μοντέλα που χρησιμοποιούνται σε αυτό το στάδιο, ακολουθούν μια σχετικά απλή και άμεση διαδικασία η οποία βασίζεται στην καταμέτρηση λέξεων-κλειδιών και εφαρμόζει απλά μαθηματικά. Τα μοντέλα αυτά δεν έχουν σκοπό να κατανοήσουν απόλυτα το νόημα της εκάστοτε δημοσίευσης. Για παράδειγμα ένας αλγόριθμος που έχει σκοπό να εντοπίσει δημοσιεύσεις που εκφράζουν άγχος, θα το επιχειρήσει εντοπίζοντας λέξεις όπως "νευρικός", "εξετάσεις",

“νοσοκομείο”, “συνέντευξη”, διαπιστώνοντας παράλληλα την έλλειψη λέξεων που εκφράζουν αντίθετα συναισθήματα, όπως “διακοπές”, “μπάνιο”, “παραλία”, “αγάπη”. Φυσικά, όταν πρόκειται για εφαρμογή του αλγορίθμου σε μία μόνο δημοσίευση, το αποτέλεσμα δεν αποτελεί παρά μία ανεπαρκώς ακριβής προσέγγιση. Όταν όμως ο αλγόριθμος εφαρμοστεί στο σύνολο των δημοσιεύσεων των καταναλωτών που ενδιαφέρουν, συγκλίνει περισσότερο στην “ομαδική διάθεση”. Επιπλέον, πρέπει να αναφερθεί ότι στις εφαρμογές των Predictive Analytics που ασχολούνται με την ανθρώπινη συμπεριφορά, όλα τα μεγέθη που εξάγονται επιβάλλεται να μελετούνται συγκριτικά. Επομένως, ακόμα και σε περίπτωση που ο αλγόριθμος που αναγνωρίζει δημοσιεύσεις που εκφράζουν άγχος δεν έχει ικανοποιητική ακρίβεια, θα ήταν σημαντική πληροφορία εάν χρησιμοποιώντας τον παρατηρούσαμε διπλασιασμό τέτοιων δημοσιεύσεων στο διάστημα δύο ημερών.

Έχοντας στο μυαλό μας αυτό τον αλγόριθμο, η οποιονδήποτε ανάλογο, θα μπορούσαμε να πούμε ότι ουσιαστικά, εφαρμόζοντας Predictive Analytics στα μέσα κοινωνικής δικτύωσης, εκμεταλλευόμαστε την τάση του σύγχρονου χρήστη του διαδικτύου να μοιράζεται τις απόψεις του προς όφελός μας. Οι άνθρωποι δημοσιεύουν καθημερινά υπεράριθμα μηνύματα, όπως “Περνάω όμορφα”, τα οποία με μία πρώτη ματιά, θα υποθέταμε ότι ενδιαφέρουν μόνο τους φίλους και την οικογένεια του χρήστη, και πως για τον υπόλοιπο κόσμο δεν έχουν κάποια αξία. Αυτό που καλούμαστε εμείς να κάνουμε, είναι ουσιαστικά να επαναπροσαρμόσουμε αυτά τα δεδομένα στις ανάγκες μας. Ανεξάρτητα από τον αρχικό σκοπό και το κοινό μιας δημοσίευσης, παραδίδει ένα σύνολο ακατέργαστης πληροφορίας, που καλούμαστε να το επαναερμηνεύσουμε μελετώντας το με έναν διαφορετικό τρόπο που αποκαλύπτει καινούργιο νόημα και διορατικότητα. Τελικά, αυτή η διαδικασία αποτελεί μια μορφή ανακύκλωσης, με την έννοια ότι επανεξετάζουμε και τελικά αξιοποιούμε τον τεράστιο αυτό αριθμό δημοσιεύσεων, που αρχικά θα θεωρούταν άχρηστος.

Αυτό που παρουσιάζει μεγάλο ενδιαφέρον με αυτή τη μορφή πληροφορίας, η οποία αποτελεί αστείρευτη πηγή για εμάς, είναι ότι ο όγκος της αυξάνεται με εκθετικό ρυθμό. Επιπλέον, καθώς το κόστος της αποθήκευσης δεδομένων μειώνεται με αντίστοιχο ρυθμό, υπάρχει πλέον η δυνατότητα αποθήκευσης και επεξεργασίας μέχρι τώρα ασύλληπτων ποσοτήτων δεδομένων. Εισάγεται επομένως ο όρος Big Data, που χαρακτηρίζει όλη την ροή δεδομένων στο διαδίκτυο. Ο όρος από μόνος του είναι σχετικά αφηρημένος, καθώς ο όγκος αυτών των δεδομένων προφανώς δεν είναι μετρήσιμος, αλλά καθώς το μέγεθος είναι σχετικό, εκφράζει το ότι ο όγκος αυξάνεται συνεχώς και ραγδαία.

Αρχικά, είναι λογικό να γεννηθεί το ερώτημα του πώς μπορούμε να είμαστε σίγουροι ότι όλη αυτή η πληροφορία που πρόκειται να συγκεντρώσουμε, κρύβει τελικά κάποια αξία για τον σκοπό μας. Παρελθοντικές εφαρμογές Predictive Analytics έχουν διαπιστώσει

διάφορους συσχετισμούς, όπως το ότι οι μελλοντικές αγορές ενός καταναλωτή σχετίζονται με το καταναλωτικό του ιστορικό, την διαδικτυακή συμπεριφορά του και τον τρόπο πληρωμής που χρησιμοποιεί. Στην ουσία, τα δεδομένα έχουν πάντα κάτι να μας προσφέρουν. Πάντα υπάρχει πληροφορία πίσω από κάθε δημοσίευση, αρκεί να την μελετήσουμε με διαφορετικό τρόπο. Επομένως, απαντώντας στο αρχικό ερώτημα, οποιοδήποτε δεδομένο έχει κάποια προβλεπτική αξία. Αυτή είναι η παραδοχή που πρέπει να κάνει οποιαδήποτε επιχείρηση πρόκειται να επενδύσει στα Predictive Analytics. Να κατανοεί δηλαδή, το ότι παίρνει το ρίσκο του να μην ξέρει τι πρόκειται να ανακαλυφθεί, ξέροντας όμως, ότι αυτό θα έχει κάποια αξία. Για αυτό το λόγο, τα δεδομένα πλέον χαρακτηρίζονται από πολλούς ως το "καινούργιο πετρέλαιο", και θεωρούνται το πιο στρατηγικό στοιχείο που μπορεί να έχει μια επιχείρηση.

1.3 Predictive Analytics στα Μέσα Κοινωνικής Δικτύωσης

Στις μέρες μας, είναι πλέον προφανές ότι τα Business Analytics γίνονται ραγδαία ο επόμενος μεγάλος παράγοντας της διοίκησης επιχειρήσεων. Σε κάθε βιομηχανία, σε όλο τον κόσμο, οι επιχειρηματίες αναρωτιούνται αν αξιοποιούν στο μέγιστο τον τεράστιο όγκο πληροφορίας που έχουν στην διάθεση τους. Οι καινοτόμες τεχνολογίες συλλέγουν δεδομένα πιο γρήγορα από ποτέ, και όλες οι εταιρείες ή οργανισμοί αναζητούν όλο και πιο αποτελεσματικούς μηχανισμούς για να μπορέσουν να αντλήσουν πληροφορία από τα δεδομένα τους ώστε να γίνουν ανταγωνιστικές στην αγορά. Οι επιχειρήσεις έχουν ανάγκη να κατανοούν τι συμβαίνει στο παρόν, τι είναι πιθανό να συμβεί στο μέλλον, και ποιες δράσεις πρέπει να γίνουν ώστε να επιτευχθεί το βέλτιστο αποτέλεσμα.

Όσο εντυπωσιακή ήταν η άφιξη των ηλεκτρονικών υπολογιστών πριν από πολλά χρόνια, τόσο είναι και η εισαγωγή των Business Analytics σε αυτό τον τομέα. Όπως ακριβώς και με τους ηλεκτρονικούς υπολογιστές, τα Business Analytics αντιμετωπίζουν μία αμφιλεγόμενη κριτική: επικροτούνται άμεσα από εκείνους που αναγνωρίζουν την δυνατότητα που μπορούν να προσφέρουν στην αύξηση της επιχειρηματικής αποτελεσματικότητας, ενώ απορρίπτονται από εκείνους που δεν έχουν πλήρη κατανόηση του τρόπου λειτουργίας τους.

Τελικά, κερδισμένοι είναι εκείνοι που τα αναγνωρίζουν. Σύμφωνα με δημοσκόπηση του MIT Sloan Managment Review (sloanreview.mit.edu) σε συνεργασία με το IBM Institute for Business Value (www-935.ibm.com/), στην οποία συμμετείχαν σχεδόν 3000 στελέχη, μάνατζερς και αναλυτές σε πάνω από 30 βιομηχανίες και 100 χώρες, οι επιχειρήσεις με κορυφαία απόδοση χρησιμοποιούν τεχνικές Business Analytics 5 φορές πιο συχνά από

εκείνες με την χαμηλότερη απόδοση. Σε γενικές γραμμές, η δημοσκόπηση διαπίστωσε την διαδεδομένη άποψη ότι τα Business Analytics παρέχουν αξία. Μισοί από τους ερωτηθέντες δήλωσαν ότι η βελτιστοποίηση της πληροφορίας και της ανάλυσής της ήταν ύψιστης προτεραιότητας στην επιχείρησή τους. Επίσης, περισσότεροι από ένας στους πέντε δήλωσαν ότι δεχόντουσαν έντονη πίεση για να υιοθετήσουν προχωρημένες τεχνικές ανάλυσης πληροφορίας. Επιπλέον, έξι στους δέκα συμφώνησαν στο ότι η επιχείρηση τους έχει περισσότερη πληροφορία από όση μπορεί να χρησιμοποιήσει αποτελεσματικά.

Πλέον, τα υψηλά στελέχη επιθυμούν οι εταιρείες τους να λειτουργούν με αποφάσεις βασισμένες σε δεδομένα. Χρειάζονται σενάρια και προσομοιώσεις που παρέχουν άμεση καθοδήγηση πάνω στις βέλτιστες δράσεις που πρέπει να γίνουν όταν προκύπτουν απρόβλεπτα συμβάντα διαφόρων μορφών, από έναν απροσδόκητο ανταγωνιστή μέχρι μία φυσική καταστροφή στην περιοχή παραγωγής, ή έναν πελάτη που δείχνει σημεία προτίμησης του ανταγωνισμού. Τα στελέχη θέλουν να κατανοούν τις καλύτερες λύσεις που είναι βασισμένες σε πολύπλοκες επιχειρηματικές παραμέτρους ή νέα πληροφορία, και θέλουν να ανταποκρίνοντα άμεσα.

Τα Business Analytics χωρίζονται σε τρεις κύριους τομείς:Descriptive, Predictive, και Prescriptive.

Οι περισσότερες επιχειρήσεις ξεκινάνε με τα Descriptive Analytics. Δηλαδή την χρήση δεδομένων για την κατανόηση της παρελθοντικής και παρούσας επιχειρηματικής απόδοσης και την λήψη τεκμηριωμένων αποφάσεων. Είναι τα Analytics που χρησιμοποιούνται κατά το μεγαλύτερο βαθμό, και τα πιο κατανοητά. Οι τεχνικές τους κατηγοριοποιούν, χαρακτηρίζουν, ενοποιούν και ταξινομούν τα δεδομένα με σκοπό να τα μετατρέψουν σε χρήσιμη πληροφορία ώστε να κατανοηθεί και να αναλυθεί η επιχειρηματική αποτελεσματικότητα. Συγκεντρώνουν τα δεδομένα σε γραφήματα και αναφορές για κόστη, πωλήσεις και κέρδη. Επιτρέπουν στα στελέχη να έχουν τυποποιημένες και προσαρμοσμένες αναφορές, ώστε να έχουν πλήρη κατανόηση της αποτελεσματικότητας μιας διαφημιστικής καμπάνιας. Απαντούν ερωτήματα όπως: Πόσες πωλήσεις είχαμε σε μια συγκεκριμένη περιοχή; Πόσα ήταν τα έσοδα και το καθαρό κέρδος το τελευταίο τετράμηνο; Επιπλέον, τα Descriptive Analytics συμβάλουν στην κατηγοριοποίηση πελατών, ώστε να επιτευχθούν στοχοποιημένες στρατηγικές διαφήμισης και προώθησης.

Τα Predictive Analytics, τα οποία αποτελούν και το κύριο θέμα αυτής της εργασίας, αναλύουν παρελθοντική απόδοση σε μία προσπάθεια να προβλέψουν το μέλλον, εξετάζοντας ιστορικά δεδομένα, εντοπίζοντας μοτίβα ή συσχετισμούς μέσα σε αυτά και τελικά προεκτείνοντας αυτούς τους συσχετισμούς στο μέλλον. Για παράδειγμα, ένας υπεύθυνος προώθησης μπορεί να θέλει να προβλέψει την ανταπόκριση διάφορων τμημάτων

καταναλωτών σε μία διαφημιστική καμπάνια, ένας κατασκευαστής ειδών ski να προβλέψει την ζήτηση της επόμενης σαιζόν για ένα συγκεκριμένο χρώμα και μέγεθος. Τα Predictive Analytics μπορούν να προβλέψουν το ρίσκο και βρίσκουν σχέσεις σε δεδομένα που δεν είναι εμφανείς χρησιμοποιώντας παραδοσιακή ανάλυση. Χρησιμοποιώντας προχωρημένες τεχνικές, μπορούν επίσης να βοηθήσουν στον εντοπισμό κρυμμένων μοτίβων σε μεγάλες ποσότητες δεδομένων για να τμηματοποιήσουν και να ομαδοποιήσουν τα δεδομένα σε σετ με συνοχή ώστε να επιτευχθεί πρόβλεψη συμπεριφοράς και να εντοπιστούν τάσεις. Απαντούν ερωτήματα όπως: Τι θα συμβεί αν η ζήτηση πέσει κατά 10% η αν η τιμή προμήθειας ανέβει κατά 5%; Τι ποσό αναμένουμε να πληρώσουμε σε καύσιμο μέσα στους επόμενους μήνες;Πόσο είναι το ρίσκο να χάσουμε κεφάλαιο σε μια καινούργια επιχειρηματική κίνηση;

Τα Prescriptive Analytics χρησιμοποιούν τεχνικές βελτιστοποίησης για να αναγνωρίσουν τις καλύτερες εναλλακτικές με σκοπό την ελαχιστοποίηση η μεγιστοποίηση κάποιου μεγέθους-στόχου. Για παράδειγμα, μπορούμε να αποφασίσουμε την πιο αποδοτική στρατηγική τιμολόγησης και διαφήμισης με σκοπό την μεγιστοποίηση του εισοδήματος, ή το καλύτερο μείγμα επενδύσεων σε ένα φάκελο απόσυρσης για ελαχιστοποίηση ρίσκου. Οι μαθηματικές και στατιστικές τεχνικές των Prescriptive Analytics μπορούν να συμβάλουν στο να ληφθούν αποφάσεις που λαμβάνουν υπόψη τους την έλλειψη σιγουριάς των δεδομένων. Απαντούν ερωτήσεις όπως:Πόση παραγωγή πρέπει να έχουμε ώστε να έχουμε μέγιστο κέρδος; Ποιος είναι ο βέλτιστος τρόπος μεταφοράς εμπορεύματος από τα εργοστάσια μας ώστε να ελαχιστοποιήσουμε το κόστος;

Τα Predictive Analytics συμπεριλαμβάνουν στατιστικά μοντέλα και άλλες εμπειρικές μεθόδους που σκοπεύουν στην δημιουργία εμπειρικών προβλέψεων, αλλά και μεθόδους για την αξιολόγηση της ποιότητας αυτών των μεθόδων. Εκτός από την πρακτική τους χρήση, τα Predictive Analytics έχουν σημαντικό ρόλο και στην ανάπτυξη θεωριών και στην δοκιμή τους.

Με αρκετή εξάσκηση, μπορούμε να επεξεργαστούμε ιστορικά δεδομένα, και να προβλέψουμε σε ένα βαθμό την πιθανότητα κάποιου μελλοντικού συμβάντος. Η φράση “σε ένα βαθμό” έχει μεγάλη σημασία. Προφανώς, δεν πιστεύουμε ότι μπορούμε να κάνουμε οποιαδήποτε ακριβή πρόβλεψη μέσα στον πολύπλοκα συνδεδεμένο, απρόβλεπτο και ταχύρυθμο κόσμο της παγκόσμιας αγοράς. Τα Predictive Analytics εκφράζουν το μέλλον υπό την μορφή πιθανοτήτων. Καμία αναλυτική εφαρμογή δεν μπορεί να προβλέψει το μέλλον με απόλυτη σιγουριά. Παρ' όλα αυτά, εφαρμόζοντας τα σωστά, προκύπτει ουσιώδης μείωση της διακύμανσης των αποτελεσμάτων μας.

Τελικά, ένας γενικευμένος ορισμός των Predictive Analytics είναι ο εξής:

“Τεχνολογία που μαθαίνει από παρελθοντικά δεδομένα με σκοπό την πρόβλεψη μελλοντικής συμπεριφοράς αποβλέποντας στην καλύτερη λήψη αποφάσεων”.[17]

Στενά συνδεδεμένες έννοιες είναι η μηχανική εκμάθηση (machine learning), η εξόρυξη γνώσης και συναισθήματος (opinion, sentiment mining), και το Big Data.

Ποια είναι όμως η πηγή από την οποία προέρχεται αυτή η μαζική ροή δεδομένων την οποία καλούνται οι επιχειρήσεις να αναλύσουν και να αξιοποιήσουν; Στις μέρες μας, το Διαδίκτυο αποτελεί την απόλυτη πλατφόρμα επιτάχυνσης της ροής πληροφοριών, και είναι ο κύριος λόγος για τον οποίο η εποχή μας ονομάζεται εποχή της πληροφορίας. Η ιδέα, δηλαδή, ότι η σημερινή εποχή έχει χαρακτηριστεί από την δυνατότητα των ανθρώπων να ανταλλάσσουν και να μεταφέρουν πληροφορίες ελεύθερα και να έχουν άμεση πρόσβαση σε γνώσεις που θα ήταν δύσκολο ή αδύνατο να βρεθούν στο παρελθόν. Τα τελευταία χρόνια ειδικότερα, με την πρώτη εμφάνιση και την ραγδαία ανάπτυξη των μέσων κοινωνικής δικτύωσης, η μηνιαία διακίνηση δεδομένων στο Internet ξεπέρασε τα 31.000 Petabytes το 2012 (έρευνα από Cisco Systems). Όπως αποδεικνύεται από την μελέτη των ερευνών που έγινε στα πλαίσια αυτής της Διπλωματικής, το σύνολο των δεδομένων που διακινείται στα Social Media συγκεκριμένα, είναι εξαιρετικά επικερδές για οποιαδήποτε επιχείρηση.

Αξιοποιώντας την προσωπική ελευθερία και την ικανότητα έκφρασης άποψης που τους προσφέρει το Internet, οι καταναλωτές συνδέονται άμεσα και διαδραστικά μεταξύ τους, συζητώντας για οποιοδήποτε προϊόν, υπηρεσία ή εταιρία. Στην διαδικασία αυτή, οι καταναλωτές είτε ενισχύουν διαφημιστικές διαδικασίες, είτε καταρρίπτουν σχέδια διαφημιστών, καθώς μοιράζονται σε πραγματικό χρόνο τις εμπειρίες και σκέψεις τους στα μέσα κοινωνικής δικτύωσης. Πέρα από τα περισσότερα γνωστά μέσα κοινωνικής δικτύωσης όπως είναι το Facebook (κοινωνική δικτύωση), το YouTube (κοινωνική δικτύωση & δημοσίευση video), το Flickr (δημοσίευση φωτογραφιών), ο όρος “μέσα κοινωνικής δικτύωσης” είναι πρακτικά πολύ ευρύτερος.

Ως “μέσα κοινωνικής δικτύωσης” μπορεί να οριστεί η δημοκρατικοποίηση της πληροφορίας, με την έννοια ότι μετατρέπει τους χρήστες του διαδικτύου από αναγνώστες περιεχομένου σε εκδότες του. Είναι η εναλλαγή από έναν γενικότερο μηχανισμό αναμετάδοσης, σε ένα μοντέλο από-πλήθος-σε-πλήθος, που σχηματίζεται μέσα στις συζητήσεις μεταξύ ανθρώπων από όλο τον κόσμο. Τα μέσα κοινωνικής δικτύωσης αξιοποιούν την 'κοινή γνώμη', για να συγκεντρώσουν πληροφορία με έναν συνεργατικό τρόπο, και μπορούν να πάρουν πολλές διαφορετικές μορφές όπως διαδικτυακά Forums, message boards, blogs, wikis, podcasts, φωτογραφίες και video.

Αυτή η τεράστια ροή διαδικτυακών δεδομένων που κυκλοφορεί στα μέσα κοινωνικής δικτύωσης και αυξάνεται καθημερινά, είναι το Big Data που αναφέρθηκε στο προηγούμενο

κεφάλαιο. Οποιαδήποτε εταιρία ή οργανισμός μπορεί να αποκομίσει τεράστια οφέλη με την προϋπόθεση ότι θα καταφέρει να αξιοποιήσει αυτή την αρχικά ακατέργαστη μορφή δεδομένων, η οποία εμπεριέχει ουσιαστικές πληροφορίες όπως κριτικές προϊόντων, στοιχεία καταναλωτών, η ακόμα και νέες τάσεις αγοράς, και σε αυτό το σημείο είναι που εισέρχεται η χρήση των Business Analytics στα μέσα κοινωνικής δικτύωσης. Οι προκλήσεις που καλούνται να αντιμετωπίσουν τα Business Analytics είναι διαφόρων ειδών, και συνήθως αφορούν την αναζήτηση, καταγραφή, και επεξεργασία των δεδομένων, την σωστή απεικόνισή τους και τελικά τον τρόπο αξιοποίησής τους.

1.4 Αντικείμενο Διπλωματικής εργασίας

Σκοπός αυτής της εργασίας είναι η επισκόπηση και η αξιολόγηση τεχνικών και αλγορίθμων Predictive Analytics που μπορούν να αξιοποιηθούν από επιχειρήσεις και οργανισμούς στη σχεδίαση προϊόντων και υπηρεσιών. Καθώς δεν υπάρχει αρκετή διαθέσιμη έρευνα αποκλειστικά στην εφαρμογή των Predictive Analytics στο στάδιο του σχεδιασμού ώστε να στηρίξει μια διπλωματική εργασία, αναλύονται τα εργαλεία και οι μέθοδοι που χρησιμοποιούνται σε άλλους τομείς της επιχειρησιακής διαχείρισης, όπως η προώθηση, οι πωλήσεις και η ικανοποίηση πελατών, και περιγράφεται ο τρόπος με τον οποίο αυτά τα εργαλεία μπορούν να εφαρμοστούν στον σχεδιασμό. Επιπλέον δίνεται έμφαση στον ρόλο που μπορούν να έχουν τα μέσα κοινωνικής δικτύωσης στον σχεδιασμό των μοντέλων. Οι μέθοδοι και τα εργαλεία αξιολογούνται με βάση την αποδοτικότητά τους σε έναν μεγάλο αριθμό από έρευνες και εφαρμογές. Στην τελική αξιολόγηση περιγράφεται ποια εργαλεία και μέθοδοι κρίνονται ως κατάλληλα για την εφαρμογή των Predictive Analytics στην σχεδίαση προϊόντων και υπηρεσιών, και προτείνονται κάποια μοντέλα που τα αξιοποιούν. Παρουσιάζονται επίσης κάποια αξιοσημείωτα παραδείγματα εφαρμογών των Predictive Analytics αναλυτικά, και γίνονται αναφορές σε λογισμικά που μπορούν να αξιοποιηθούν.

2 Predictive Analytics

2.1 Εφαρμογές των Predictive Analytics στην Προώθηση και στις Πωλήσεις.

Οι εφαρμογές των Predictive Analytics στην προώθηση και στις πωλήσεις είναι οι πιο αναγνωρισμένες και επικερδείς εφαρμογές. Κάθε πελάτης αξιολογείται με βάση την καταναλωτική συμπεριφορά του, δηλαδή τις αγορές του, την ανταπόκρισή του σε προωθητικές ενέργειες, και την συμπεριφορά του σε μια ιστοσελίδα ηλεκτρονικών συναλλαγών. Οι παραπάνω πληροφορίες συνδυάζονται με τα προσωπικά χαρακτηριστικά των καταναλωτών δηλαδή την ηλικία τους, το φύλο τους, την οικογενειακή τους κατάσταση και τα γεωγραφικά τους χαρακτηριστικά, συνήθως για σκοπούς τμηματοποίησης της αγοράς. Η αξιολόγηση αυτή χρησιμοποιείται ώστε να ληφθούν αποφάσεις σχετικά με την προώθηση, τις πωλήσεις, την εξυπηρέτηση πελατών και την διαμόρφωση της διαδικτυακής στρατηγικής της επιχείρησης.

Μοντελοποίηση Ανταπόκρισης στην Άμεση Προώθηση: Είναι μια από τις καθιερωμένες επιχειρηματικές εφαρμογές των Predictive Analytics. Μαθαίνοντας από την εμπειρία που υπάρχει από τις παλαιότερες καμπάνιες, παρατηρώντας ποιος πελάτης ανταποκρίθηκε η όχι στις δράσεις προώθησης, οι παροντικοί υποψήφιοι στόχοι προώθησης μπορούν να μελετηθούν με βάση την πιθανότητα τους να ανταποκριθούν θετικά. Ύστερα μπορούν να στοχοποιηθούν μόνο οι πελάτες που παρουσιάζουν ένα σκορ που θεωρείται ικανοποιητικό. Ξοδεύοντας αποκλειστικά το κόστος για τους πελάτες που έχουν μεγαλύτερη πιθανότητα να ανταποκριθούν, μια καμπάνια προώθησης αποδίδει περισσότερο, αφού πετυχαίνει μεγαλύτερο βαθμό

ανταπόκρισης. Η απόδοση ενός προβλεπτικού μοντέλου ανταπόκρισης αξιολογείται με βάση το lift, όπως εξηγείται στην αξιολόγηση των προβλεπτικών μοντέλων.

Συγκράτηση πελατών με μοντελοποίηση αποχώρησης (churn): Ένας βασικός κανόνας του marketing είναι ότι η διατήρηση ενός πελάτη έχει μικρότερο κόστος από την απόκτηση ενός καινούργιου. Θεωρώντας ότι το πελατολόγιο μιας επιχείρησης έχει ένα σταθερό ρυθμό με τον οποίο προστίθενται πελάτες σε αυτό και ένα με τον οποίο αποχωρούν πελάτες, είναι εύκολα κατανοητό ότι μειώνοντας τον ρυθμό αποχώρησης αυξάνεται ο ρυθμός ανάπτυξης του πελατολογίου. Όμως, μια προσφορά που έχει σκοπό να συγκρατήσει τους πελάτες, όπως μια σημαντική έκπτωση σε κάποιο συμβόλαιο, μπορεί να έχει μεγάλο κόστος. Δεν είναι πρακτικό να γίνει μια τέτοια προσφορά σε όλους τους πελάτες, αφού κάθε πελάτης που δεν έχει διάθεση αποχώρησης από την εταιρία, αλλά στον οποίο γίνεται η προσφορά, αποτελεί ένα αχρείαστο κόστος που πρέπει να αποφευχθεί. Η μοντελοποίηση αποχώρησης προβλέπει ποιοι πελάτες έχουν διάθεση να αποχωρήσουν από την εταιρεία. Αξιολογώντας κάθε πελάτη με βάση την πιθανότητα αποχώρησης, οι προσφορές μεγάλου κόστους μπορούν να είναι στοχοποιημένες αποδοτικά, αποφεύγοντας πελάτες που θα παρέμεναν στην εταιρεία έτσι και αλλιώς.

Μοντελοποίηση Uplift στην Άμεση προώθηση: Καθώς μια καμπάνια προώθησης δεν αξιολογείται μόνο από την ανταπόκρισή της, ενδιαφέρει και το αυξητικό της αποτέλεσμα, δηλαδή τα επιπρόσθετα έσοδα που οφείλονται σε αυτήν, και που χωρίς αυτή δεν θα υπήρχαν. Ακόμα και αν μια καμπάνια δείχνει μεγάλο βαθμό ανταπόκρισης, και επομένως μεγάλο κέρδος, υπάρχει ένα βασικό ερώτημα που μένει αναπάντητο: Ποιοί είναι οι πελάτες που θα είχαν αγοράσει το προϊόν ανεξάρτητα από το αν επιλέχθηκαν για προώθηση. Σε αρκετές περιπτώσεις, μέχρι και το μισό των συνολικών πελατών είχαν τόσο μεγάλη επιθυμία να αγοράσουν το προϊόν που θα αγόραζαν και χωρίς προώθηση.

Εισάγεται επομένως η έννοια της Uplift μοντελοποίησης. Ένα μοντέλο Uplift λειτουργεί όπως τα υπόλοιπα προβλεπτικά μοντέλα, με την έννοια ότι με βάση τα χαρακτηριστικά ενός καταναλωτή, τον αξιολογεί όσο αφορά την μελλοντική συμπεριφορά του που ενδιαφέρει. Η διαφορά είναι ότι αντί να προβλέπει μια

ανεξάρτητη συμπεριφορά, αξιολογεί τον καταναλωτή με βάση την πιθανότητα να επηρεαστεί η συμπεριφορά του από τις δράσεις τις επιχείρησης. Η αξιολόγηση Uplift απαντάει στο ερώτημα: "Πόσο πιο πιθανό είναι μια δράση να αποφέρει το επιθυμητό αποτέλεσμα (σε αυτή την περίπτωση είναι η αγορά), σε σχέση με μια άλλη δράση. Επομένως βοηθάει μια επιχείρηση να αποφασίσει με ποιο τρόπο να δράσει για κάθε καταναλωτή.

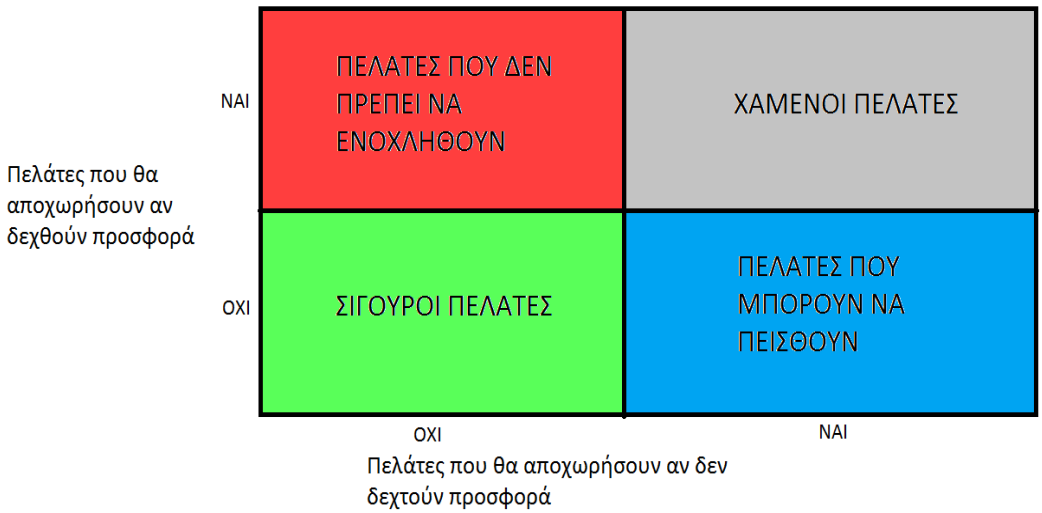
Η απάντηση αυτού του ερωτήματος αποτελεί μια πρόκληση, καθώς μια επιχείρηση δεν μπορεί να γνωρίζει τις απαντήσεις στα ερωτήματα που χρειάζονται για να εξαχθεί ένα συμπέρασμα. Συγκεκριμένα, χρειάζεται να απαντηθεί το ερώτημα "Ο καταναλωτής θα αγοράσει αν δεχτεί μια προσφορά;" αλλά και το "Ο καταναλωτής θα αγοράσει αν δεν δεχτεί προσφορά;". Σε περίπτωση που ξέραμε και τις δύο αυτές απαντήσεις, θα μπορούσαμε να συμπεράνουμε αν ο καταναλωτής μπορεί να επηρεαστεί από την δράση της επιχείρησης. Αν οι απαντήσεις στα δύο αυτά ερωτήματα συμφωνούν, δηλαδή ο καταναλωτής δέχεται μια προσφορά και αγοράζει, ή δεν δέχεται προσφορά και δεν αγοράζει, το συμπέρασμα είναι ότι ο καταναλωτής μπορεί να επηρεαστεί. Σε περίπτωση που ο καταναλωτής αγοράσει και χωρίς να δεχτεί προσφορά αλλά και όταν δεχτεί, το συμπέρασμα είναι ότι η δράση της επιχείρησης δεν τον επηρεάζει. Όμως δεν μπορούμε να ξέρουμε όλες αυτές τις απαντήσεις, γιατί πολύ απλά δεν μπορούμε και να κάνουμε προσφορά στον καταναλωτή, και να μην κάνουμε. Με πιο απλά λόγια, μετά την αγορά δεν μπορούμε να ξέρουμε αν ήταν η δράση προώθησης της επιχείρησης που οδήγησε τον καταναλωτή σε αγορά ή αν θα αγόραζε έτσι κι αλλιώς.

Για να απαντηθεί αυτό το ερώτημα, οι καταναλωτές χωρίζονται σε δύο τμήματα:το τμήμα ελέγχου (control set) στο οποίο δεν εφαρμόζεται καμία δράση προώθησης, και το τμήμα πειραματισμού (treated set) στο οποίο εφαρμόζεται η δράση προώθησης. Το μοντέλο uplift αξιοποιεί τα δεδομένα και από τα δύο τμήματα για να ξεχωρίσει τους καταναλωτές που μπορούν να επηρεαστούν. Επιλέγοντας κάποιες μεταβλητές, όπως παρελθοντικές αγορές, ηλικία καταναλωτή και άλλα, παρατηρείται η συμπεριφορά των καταναλωτών. Συνδυάζοντας τα αποτελέσματα των δύο τμημάτων, αλλά και τα καταναλωτικά χαρακτηριστικά τους, εξάγονται συμπεράσματα για το πώς θα συμπεριφερθεί μια νέα ομάδα καταναλωτών, με βάση

τις μεταβλητές που έχουν επιλεχθεί. Τελικά, για την άμεση προώθηση, σκοπός της uplift μοντελοποίησης, είναι η κατάταξη των καταναλωτών στις εξής ομάδες:

Αυτή η τμηματοποίηση ξεχωρίζει κάθετα τους πελάτες που θα ανταποκριθούν σε μια προσφορά, που είναι το αποτέλεσμα του παραδοσιακού μοντέλου ανταπόκρισης. Επιπρόσθετα, τους ξεχωρίζει οριζόντια με βάση την ερώτηση: Ποιοι θα αγοράσουν αν δεν δεχτούν προσφορά. Το κάτω-δεξιά τμήμα, δηλαδή οι πελάτες που μπορούν να πεισθούν, αντιπροσωπεύει τους πελάτες στους οποίους είναι άξιο να ξοδέψει μια επιχείρηση για προώθηση. Είναι δηλαδή οι πελάτες που δεν θα αγοράσουν αν δεν γίνει προσπάθεια προώθησης, αλλά θα αγοράσουν αν γίνει. Αυτή η ομάδα έχει τεράστιο ενδιαφέρον σε εφαρμογές που ασχολούνται με εκλογές και ψηφοφορία. Επομένως η μοντελοποίηση uplift παρέχει έναν ξεκάθαρο και άμεσο τρόπο μείωσης κόστους πέρα από την παραδοσιακή μοντελοποίηση ανταπόκρισης. Αποκλείουμε από την λίστα προώθησης τους πελάτες του κάτω-αριστερού τμήματος, οι οποίοι ναι μεν ανταποκρίνονται στην προώθηση, αλλά θα αγόραζαν έτσι και αλλιώς, γλιτώνοντας κόστος.

Μοντελοποίηση uplift αποχώρησης: Όπως και με την μοντελοποίηση ανταπόκρισης, αναλύοντας την αποχώρηση των πελατών σε δύο διαστάσεις, έχουμε τεράστια αύξηση στην αποτελεσματικότητα της. Με την μοντελοποίηση uplift αποχώρησης, οι πελάτες κατατάσσονται στα τέσσερα εξής τμήματα:



Η τμηματοποίηση χωρίζει πρώτα τους πελάτες οριζόντια με βάση το ποιοι θα φύγουν αν δεν τους γίνει κάποια προσφορά, που είναι ουσιαστικά η απλή μοντελοποίηση αποχώρησης, και ύστερα τμηματοποιεί κάθετα με βάση ποιοι πελάτες θα φύγουν αν δεχτούν κάποια προσφορά. Όπως και με την μοντελοποίηση ανταπόκρισης, το κάτω δεξιά τμήμα αντιπροσωπεύει τους πελάτες που πρέπει να στοχοποιηθούν με προσφορές συγκράτησης, αφού έχουν διάθεση αποχώρησης, αλλά μπορούν να πεισθούν με κάποια προσφορά. Το πάνω-αριστερά τμήμα παρουσιάζει μια καινούργια ευκαιρία μείωσης κόστους. Αυτοί οι πελάτες που ονομάζονται “Sleeping Dogs” (σκυλιά που κοιμούνται), θα μείνουν στην εταιρεία αν δεν τους ενοχλήσουμε, αλλά μια προσφορά συγκράτησης θα έχει ένα αντίξοο, αντίστροφο αποτέλεσμα, δηλαδή την αποχώρηση του πελάτη. Η παθητική συμπεριφορά προς πελάτες τους οποίους μια προωθητική προσπάθεια μπορεί να “τρομάξει” είναι ένας τρόπος μείωσης κόστους. Πολλές επιχειρήσεις αντιμετωπίζουν το ρίσκο μια τέτοιας αντίξοης ανταπόκρισης στην προσπάθεια συγκράτησης. Οι εταιρείες κινητής τηλεφωνίας συχνά προσφέρουν μια δωρεάν συσκευή σε πελάτες των οποίων τα συμβόλαια λήγουν, με την προϋπόθεση ότι θα ανανεώσουν το συμβόλαιο τους. Για μερικούς χρήστες, αυτό έχει ως αποτέλεσμα ότι πλέον είναι ελεύθεροι να αποχωρήσουν από την εταιρεία και να μεταπηδήσουν σε μια ανταγωνιστική. Παρόμοια, άλλες συνδρομητικές υπηρεσίες όπως διαδικτυακές ιστοσελίδες

γνωριμιών, έχουν πελάτες που αν και δεν τις χρησιμοποιούν πλέον, χρεώνονται τακτικά. Μια προσπάθεια συγκράτησης ίσως τελικά τους υπενθυμίσει να ακυρώσουν την συνδρομή τους. Επιπλέον, όλες οι επιχειρήσεις μπορεί να έχουν πελάτες που ανταποκρίνονται αντίξοα σε επικοινωνία που λαμβάνεται ως ενοχλητική ή αχρείαστη. Η μοντελοποίηση uplift αποχώρησης προβλέπει ποιοι πελάτες θα ανταποκριθούν και με ποιο τρόπο.

Viral marketing:Το viral marketing στοχεύει στο να αυξήσει την αναγνωρισιμότητα του προϊόντος και τα κέρδη της προώθησης με τη βοήθεια της κοινωνικής επιρροής και των κοινωνικών δικτύων. Το άμεσο marketing είναι μια σημαντική εφαρμογή, που επιχειρεί να προωθήσει το προϊόν μόνο σε ένα σύνολο πιθανών επικερδών πελατών και υποστηρίζεται από τα μοντέλα ανταπόκρισης που προαναφέρθηκαν [18]. Αν εφαρμοστούν σωστά, αυτά τα μοντέλα μπορούν να αυξήσουν τα κέρδη σημαντικά, αλλά ένας περιορισμός αυτής της προσέγγισης είναι ότι μελετάει τον κάθε πελάτη ανεξάρτητα από τους υπόλοιπους όσο αφορά την συμπεριφορά τους. Στην πραγματικότητα, η απόφαση ενός ατόμου να αγοράσει ένα προϊόν συχνά επηρεάζεται από τους φίλους και γνωστούς του. Δεν είναι επιθυμητό να αγνοηθεί τέτοια επιρροή, καθώς μπορεί να οδηγήσει σε σημαντικά μη βέλτιστες αποφάσεις.

Ο στόχος είναι η ελαχιστοποίηση του κόστους προώθησης και η μεγιστοποίηση του κέρδους. Για παράδειγμα, μια επιχείρηση μπορεί να θέλει να προωθήσει ένα καινούργιο προϊόν μέσα από το φαινόμενο από στόμα σε στόμα (word of mouth) που προκύπτει από τις αλληλεπιδράσεις σε ένα κοινωνικό δίκτυο. Ο σκοπός είναι να καταφέρουμε ένα μικρό αριθμό χρηστών με επιρροή να υιοθετήσουν το προϊόν μας, ξεκινώντας έτσι μια αλυσίδα περαιτέρω πωλήσεων. Για να το πετύχουμε αυτό, χρειαζόμαστε κάποιες μετρήσεις για να ποσοτικοποιήσουμε τα χαρακτηριστικά των χρηστών με επιρροή (πχ αναμενόμενο κέρδος από χρήστη), και το δίκτυο του χρήστη (πχ αναμενόμενο κέρδος από χρήστες που δύναται να επηρεαστούν από τον χρήστη).

Τμηματοποίηση Καταναλωτών: Είναι κοινή πρακτική οι επιχειρήσεις να στοχεύουν στην ομαδοποίηση της αγοράς η οποία εξυπηρετεί διάφορους σκοπούς. Η

τμηματοποίηση γίνεται με βάση τα χαρακτηριστικά των καταναλωτών που ενδιαφέρουν, δηλαδή ηλικία, φύλο, οικογενειακή κατάσταση και άλλα. Αφού έχουν αναγνωριστεί οι ομάδες, η επιχείρηση μπορεί να αποφασίσει πιο αποδοτικά για πολλά ζητήματα όπως η στοχευμένη διαφήμιση, η εξυπηρέτηση πελατών ή ακόμα και για τον σχεδιασμό ενός προϊόντος που θα ικανοποιεί της ανάγκες μιας κερδοφόρας ομάδας. Σε πολλές περιπτώσεις όμως, τα χαρακτηριστικά που απαιτούνται για την τμηματοποίηση των καταναλωτών δεν είναι διαθέσιμα για ένα σύνολο πελατών. Σε αυτό το σημείο εισέρχονται τα Predictive Analytics, εξάγοντας εκτιμήσεις για τα χαρακτηριστικά που λείπουν, συνδυάζοντας τα γνωστά στοιχεία με το σύνολο των δεδομένων των καταναλωτών.

Στοχευμένη Διαφήμιση: Είναι μια εφαρμογή που είναι πλέον διαδεδομένη στο διαδίκτυο. Τα χαρακτηριστικά των χρηστών μελετούνται ώστε να τους παρουσιάζεται η διαφήμιση που είναι πιο πιθανό να τους πείσει. Η στοχευμένη διαφήμιση μπορεί να έχει και την μορφή προτάσεων αγοράς, δηλαδή με βάση το τι ψάχνει ή πρόκειται να αγοράσει ένας καταναλωτής, να του προτείνεται ένα προϊόν. Συγκεκριμένα στο διαδίκτυο, όπου οι χρήστες μεταβαίνουν από ιστοσελίδα σε ιστοσελίδα με γρήγορο ρυθμό, είναι κρίσιμο να παρουσιάζεται η σωστή διαφήμιση ή πρόταση σε πραγματικό χρόνο.

Αξιολόγηση Πίστωσης (Credit Score): Αφορά κυρίως επιχειρήσεις που παρέχουν προϊόντα σε μορφή υπηρεσιών, δηλαδή τράπεζες, ασφαλιστικές εταιρίες και εταιρείες τηλεφωνίας. Πρόκειται για την εξαγωγή μιας αξιολόγησης, η οποία βασίζεται σε διάφορες μεταβλητές που ενδιαφέρουν. Για παράδειγμα, μια τράπεζα μπορεί να αξιολογεί τους πελάτες της με βάση το εισόδημά τους, την ηλικία τους, τις καθυστερημένες πληρωμές τους και το επίπεδο μόρφωσής τους.

2.2 Εφαρμογές των Predictive Analytics στη Σχεδίαση Προϊόντων και Υπηρεσιών

Έχοντας περιγράψει τις πιο διαδεδομένες εφαρμογές των Predictive Analytics στους κλάδους της προώθησης και των πωλήσεων, το επόμενο βήμα είναι να αναλογιστούμε το πως

τα Predictive Analytics μπορούν να εφαρμοστούν στον κλάδο της σχεδίασης. Αφού τα Predictive Analytics έχουν σαν γενικό στόχο τον συσχετισμό μεταβλητών, ώστε να γίνεται κατανοητό το πώς η συμπεριφορά μίας μεταβλητής μπορεί να προβλέψει την συμπεριφορά μίας άλλης, το πρώτο βήμα είναι να οριστούν αυτές οι μεταβλητές. Απαιτείται επομένως μια μέθοδος μοντελοποίησης της σχεδίασης. Μία τέτοια μεθοδολογία είναι το "σπίτι της ποιότητας" (house of quality)[40].

Σύμφωνα με τους σχεδιαστές του σπιτιού της ποιότητας, η σημασία της ποιότητας του προϊόντος ή της υπηρεσίας που παρέχεται από κάθε επιχείρηση ως ανταγωνιστικός παράγοντας είναι αδιαμφισβήτητη. Κύρια πρόκληση στον ποιοτικό σχεδιασμό προϊόντων και υπηρεσιών αποτελεί η σύνδεση των χαρακτηριστικών του προϊόντος ή της υπηρεσίας, με τις ανάγκες και την ικανοποίηση των πελατών. Επομένως είναι ανάγκη να καθοριστούν κάποιες σχεδιαστικές απαιτήσεις, ώστε να ικανοποιούνται οι πελάτες. Αυτές οι σχεδιαστικές απαιτήσεις, είναι απαραίτητο να εκφράζονται σε μετρήσιμη μορφή. Απαιτείται επομένως, μια διαδικασία η οποία θα μεταφράζει μεθοδικά τις ανάγκες των καταναλωτών σε μετρήσιμα χαρακτηριστικά του προϊόντος ή της υπηρεσίας που σχεδιάζεται. Με βάση αυτά, το σπίτι της ποιότητας αποτελεί μια απεικόνιση επτά βασικών στοιχείων:

Ανάγκες Καταναλωτών: Πρόκειται για την αρχική είσοδο του σπιτιού. Δίνουν έμφαση στα χαρακτηριστικά του προϊόντος ή της υπηρεσία στα οποία η επιχείρηση πρέπει να δώσει προσοχή. Οι ανάγκες των καταναλωτών μπορούν να αναγνωριστούν με κλασσικές μεθόδους ερωτηματολογίου και έρευνας αγοράς, είτε από τα μέσα κοινωνικής δικτύωσης. Εκεί οι καταναλωτές εκφράζουν ειλικρινά τις απόψεις τους με δικά τους λόγια, κάτι το οποίο είναι πλεονέκτημα, γιατί είναι σημαντικό για την επιχείρηση να έχει πρόσβαση στην αναλλοίωτη φωνή των καταναλωτών

Τεχνικά Χαρακτηριστικά Προϊόντος ή Υπηρεσίας: Περιγράφουν το προϊόν ή την υπηρεσία στην γλώσσα των σχεδιαστών, επομένως μπορούν να θεωρηθούν ως η φωνή της επιχείρησης. Χρησιμοποιούνται για να αξιολογήσουν πόσο καλά η επιχείρηση ικανοποιεί τις ανάγκες των καταναλωτών. Οι ανάγκες των καταναλωτών οδηγούν την επιχείρηση στο τι να κάνει, ενώ τα τεχνικά χαρακτηριστικά την οδηγούν στο πως να το κάνει. Πρέπει να εκφράζονται σε μετρήσιμους και συγκρίσιμους όρους (όπως οι διαστάσεις της οθόνης ενός κινητού τηλεφώνου σε χιλιοστά)

Σχετική σημασία αναγκών καταναλωτών: Καθώς τα δεδομένα που συλλέγονται για τους καταναλωτές συνήθως περιέχουν πάρα πολλές ανάγκες που πρέπει να αντιμετωπισθούν ταυτόχρονα, αυτές οι ανάγκες πρέπει να αξιολογηθούν. Η επιχείρηση θα πρέπει να βάλει προτεραιότητες στις ανάγκες των καταναλωτών ώστε να σκοπεύει στην ικανοποίηση των πιο σημαντικών αναγκών, αγνοώντας εκείνες που είναι σχετικά ασήμαντες.

Σχέσεις ανάμεσα στις ανάγκες και στα τεχνικά χαρακτηριστικά: Πρόκειται για έναν πίνακα συσχετισμού που αναλύει πως το κάθε τεχνικό χαρακτηριστικό επηρεάζει την κάθε ανάγκη των καταναλωτών.

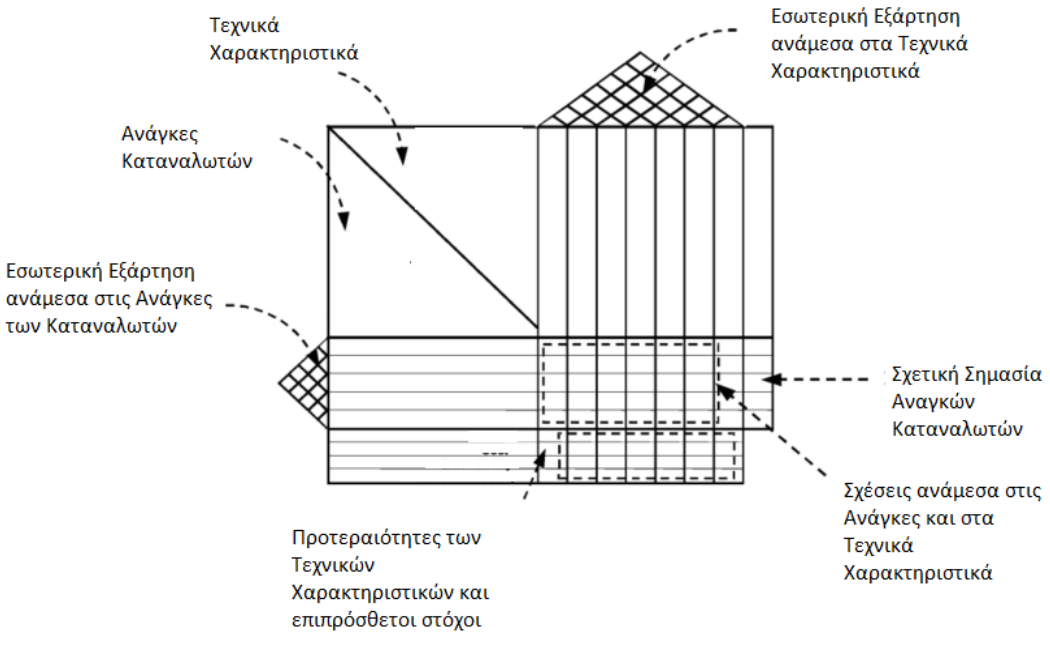
Εσωτερική εξάρτηση ανάμεσα στις ανάγκες των καταναλωτών: Γενικότερα, οι ανάγκες των καταναλωτών παρουσιάζουν ένα συσχετισμό μεταξύ τους. Μερικές ανάγκες υποστηρίζονται μεταξύ τους ενώ άλλες επηρεάζουν αρνητικά την ικανοποίηση άλλων αναγκών. Αυτές οι αλληλεπιδράσεις περιγράφονται από έναν πίνακα συσχετισμού ο οποίος φανερώνει απαραίτητες ανταλλαγές τεχνικών χαρακτηριστικών.

Εσωτερική εξάρτηση ανάμεσα στα τεχνικά χαρακτηριστικά: Είναι ο πίνακας που αποτελεί την οροφή του σπιτιού. Περιγράφει τα διάφορα τεχνικά χαρακτηριστικά που βελτιώνονται εις βάρος άλλων χαρακτηριστικών, δίνοντας μια βάση υπολογισμού του κατά πόσο η βελτίωση ενός τέτοιου χαρακτηριστικού θα επιδράσει στα υπόλοιπα. Αυτός ο πίνακας συσχετισμού αποκαλύπτει τα απαραίτητα τεχνικά χαρακτηριστικά και τους απαραίτητους συμβιβασμούς που πρέπει να γίνουν.

Προτεραιότητες των τεχνικών χαρακτηριστικών και επιπρόσθετοι στόχοι: Σε αυτό το τελευταίο στάδιο, τα αποτελέσματα των προηγούμενων βημάτων αξιοποιούνται ώστε να υπολογιστεί μια τελευταία αξιολόγηση των τεχνικών χαρακτηριστικών που σκοπεύουν στην ικανοποίηση των πελατών.

Το σπίτι της ποιότητας δεν αποτελεί μόνο εργαλείο για τους σχεδιαστές του προϊόντος ή της υπηρεσίας, αλλά είναι και ένα μέσο σύνοψης και μετατροπής της άποψης των καταναλωτών σε σχεδιαστικά μεγέθη. Επιπλέον μπορεί να χρησιμοποιηθεί και για σκοπούς προώθησης, αφού εκφράζει την φωνή του καταναλωτικού κοινού.

Στην εικόνα φαίνεται η απεικόνιση του σπιτιού της ποιότητας [40]:



Επομένως με βάση αυτή την μοντελοποίηση της διαδικασίας της σχεδίασης, οι μεταβλητές τις οποίες καλούνται να αναλύσουν τα Predictive Analytics είναι οι ανάγκες των καταναλωτών, και τα τεχνικά χαρακτηριστικά του προϊόντος ή της υπηρεσίας που σχεδιάζεται. Άρα σε αρχικό στάδιο, θα πρέπει να αναλυθούν και οι συσχετισμοί που υπάρχουν ανάμεσα σε αυτές τις δύο μεταβλητές, αλλά και οι συσχετισμοί που έχουν οι ανάγκες των καταναλωτών μεταξύ τους.

Αν και το σπίτι της ποιότητας δεν είναι πρόσφατη μοντελοποίηση, είναι κατανοητό ότι οι βασικές αρχές του, δηλαδή η ανάγκη συσχετισμού των αναγκών των καταναλωτών με τα τεχνικά χαρακτηριστικά, ισχύουν διαχρονικά. Το στοιχείο που είναι διαφορετικό τα τελευταία χρόνια όμως, είναι οι παράγοντες που επηρεάζουν την απόφαση αγοράς. Ως αποτέλεσμα της ραγδαίας εξέλιξης της τεχνολογίας σχεδιασμού προϊόντος, τα παρόμοια προϊόντα παρουσιάζουν πολλές ομοιότητες σε λειτουργικό επίπεδο, επομένως είναι δύσκολο να διαχωριστούν από τους καταναλωτές με βάση τα λειτουργικά χαρακτηριστικά τους [48]. Επομένως για να συλληφθεί ένα καλοσχεδιασμένο προϊόν που θα έχει απήχηση στους καταναλωτές, το προϊόν δεν θα πρέπει μόνο να ικανοποιεί τις υλικές ανάγκες τους, αλλά και τις συναισθηματικές [47]. Επομένως στις ανταγωνιστικές αγορές, τα επιτυχημένα προϊόντα ίσως δεν είναι αυτά που αξιοποιούν την τελευταία τεχνολογία, αλλά αυτά που έχουν τα επιθυμητά τεχνικά χαρακτηριστικά που προκύπτουν και από τις συναισθηματικές ανάγκες. Ορίζεται επομένως ο συναισθηματικός σχεδιασμός, ο οποίος αποσκοπεί στο να μεταφράζει τις συναισθηματικές, ή αλλιώς άυλες ανάγκες των καταναλωτών σε τεχνικά χαρακτηριστικά [47].

Τα δεδομένα για τον συναισθηματικό σχεδιασμό προκύπτουν από έρευνα αγοράς. Συγκεκριμένα, διαφορετικά δείγματα, που αποτελούν εναλλακτικές σχεδιασμού, παρουσιάζονται σε επιλεγμένους πελάτες, και η αντίδραση τους στο κάθε δείγμα καταγράφεται ως μια αξιολόγηση που αντιστοιχείται σε μία "λέξη συναισθήματος". Οι λέξεις συναισθήματος είναι λέξεις που εκφράζουν τα συνηθισμένα συναισθήματα των καταναλωτών σε σχέση με τα προϊόντα. Μερικές από αυτές μπορεί να είναι "στυλάτο", "βαρετό" ή "παιδικό". Με βάση αυτήν την αξιολόγηση, εξάγονται συσχετισμοί ανάμεσα στα τεχνικά χαρακτηριστικά του προϊόντος, και τις άυλες ανάγκες των καταναλωτών. Τα αποτελέσματα μιας τέτοιας έρευνας αγοράς, που αφορούσε συσκευές κινητών τηλεφώνων, φαίνονται στον πίνακα [47]:

Καταναλωτές	**Τεχνικά Χαρακτηριστικά**			**Αξιολόγηση λέξης συναισθήματος “στυλάτο”**
	Υλικό	**Χρώμα**	**Τύπος Συσκευής**	
1	Περισσότερο Πλαστικό	Ασημένιο	Απλή	3
2	Περισσότερο Πλαστικό	Ασημένιο	Απλή	3
3*	Περισσότερο Πλαστικό	Ασημένιο	Διπλωτή	3
4*	Περισσότερο Πλαστικό	Ασημένιο	Διπλωτή	5
5x	Περισσότερο Πλαστικό	Μαύρο	Απλή	1
6x	Περισσότερο Πλαστικό	Μαύρο	Απλή	2
7	Περισσότερο Πλαστικό	Κόκκινο	Απλή	2
8	Περισσότερο Πλαστικό	Κόκκινο	Απλή	2
9	Περισσότερο Πλαστικό	Μαύρο	Διπλωτή	3
10	Περισσότερο Πλαστικό	Μαύρο	Διπλωτή	3
11++	Περισσότερο Πλαστικό	Κόκκινο	Διπλωτή	3
12++	Περισσότερο Πλαστικό	Κόκκινο	Διπλωτή	4
13	Περισσότερο Μεταλλικό	Ασημένιο	Απλή	4
14	Περισσότερο Μεταλλικό	Ασημένιο	Απλή	4
15	Περισσότερο Μεταλλικό	Ασημένιο	Διπλωτή	5
16	Περισσότερο Μεταλλικό	Ασημένιο	Διπλωτή	5
17#	Περισσότερο Μεταλλικό	Μαύρο	Απλή	4
18#	Περισσότερο Μεταλλικό	Μαύρο	Απλή	5
19	Περισσότερο Μεταλλικό	Κόκκινο	Απλή	4
20	Περισσότερο Μεταλλικό	Κόκκινο	Απλή	4
21$	Περισσότερο Μεταλλικό	Μαύρο	Διπλωτή	4
22$	Περισσότερο Μεταλλικό	Μαύρο	Διπλωτή	5
23	Περισσότερο Μεταλλικό	Κόκκινο	Διπλωτή	5
24	Περισσότερο Μεταλλικό	Κόκκινο	Διπλωτή	5

Συσκευές με διαφορετικό χρώμα, συνδυασμό πλαστικού και μετάλλου, και “διπλωτές” ή όχι, παρουσιάστηκαν σε 24 καταναλωτές, και αξιολογήθηκαν με βάση το πόσο “στυλάτες” θεωρήθηκαν από τον καθένα.

Τα δεδομένα αυτά παρουσιάζουν δύο βασικά προβλήματα. Το πρώτο πρόβλημα είναι σχετικό με την ακρίβεια αυτών των αξιολογήσεων. Τα νούμερα που αντιστοιχίζονται από κάθε καταναλωτή της έρευνας στις λέξεις συναισθήματος, αποτελούν ανακριβείς αξιολογήσεις, αφού ένα συναίσθημα δεν μπορεί να ποσοτικοποιηθεί με ακρίβεια. Επιπλέον, καθώς κάθε καταναλωτής μπορεί να έχει διαφορετική άποψη για το τι θεωρείται “στυλάτο” είναι σε μεγάλο βαθμό και υποκειμενικές αξιολογήσεις. Τέτοιες περιπτώσεις έχουν σημειωθεί στον πίνακα με τα σύμβολα *,x,++,#,$. Παρατηρούμε ότι αυτά τα ζευγάρια καταναλωτών έχουν αξιολογήσει διαφορετικά τη συσκευή, παρ' όλο που τα τεχνικά χαρακτηριστικά είναι τα ίδια.

Το δεύτερο πρόβλημα είναι σχετικό με τον επακριβή συσχετισμό αυτών των αξιολογήσεων με τα τεχνικά χαρακτηριστικά. Ενώ οι αξιολογήσεις των καταναλωτών είναι μια αριθμητική αξία που είναι ξεκάθαρο ότι το όσο μεγαλύτερη είναι, τόσο πιο “στυλάτη” θεωρείται η συσκευή, δεν ισχύει το ίδιο και για τα αντίστοιχα τεχνικά χαρακτηριστικά. Με άλλα λόγια, ενώ μπορεί να έχει καθιερωθεί ο συσχετισμός του χρώματος της συσκευής με το πόσο “στυλάτη” θεωρείται, δεν είναι δυνατό να προβλεφθεί το κατά πόσο αυτή η αξιολόγηση θα αλλάξει αν το χρώμα της συσκευής αλλάξει, ούτε το αν τελικά θα αυξηθεί η θα μειωθεί.

Διαπιστώνουμε επομένως, ότι προστίθεται μια σημαντική διάσταση στο πρόβλημα της αναγνώρισης των αναγκών των καταναλωτών και στο πως αυτές συσχετίζονται με τα τεχνικά χαρακτηριστικά του προϊόντος ή της υπηρεσίας που σχεδιάζεται, η οποία δεν πρέπει να αγνοηθεί, αφού όπως προαναφέρθηκε η ικανοποίηση των άυλων αναγκών έχει πλέον πολύ σημαντικό ρόλο στην επιτυχή σχεδίαση.

Σε αυτό το σημείο περιγράφονται μερικές ιδέες εφαρμογών των Predictive Analytics στη σχεδίαση. Η μοντελοποίηση των αυτών των εφαρμογών και άλλων, περιγράφεται στο κεφάλαιο 4.3, αφού έχουν αξιολογηθεί τα διάφορα εργαλεία που χρησιμοποιούνται στα προβλεπτικά μοντέλα.

Πρόβλεψη Αναγκών Καταναλωτών: Είναι πλέον ξεκάθαρο πως για την λήψη αποφάσεων στην σχεδίαση η αναγνώριση των αναγκών των καταναλωτών είναι το πρώτο βήμα. Οι ανάγκες των καταναλωτών όμως μπορεί να μην είναι ούτε ξεκάθαρες, ούτε διαθέσιμες στην

επιχείρηση. Επομένως μια εφαρμογή Predictive Analytics η οποία μπορεί να αναγνωρίζει ανάγκες που δεν εκφράζονται, με βάση τα χαρακτηριστικά των καταναλωτών, ή να προβλέπει ανάγκες που πρόκειται να δημιουργηθούν με βάση τις ήδη υπάρχουσες, θα ήταν ένα πολύ ισχυρό εργαλείο στην διάθεση της επιχείρησης. Η εφαρμογή θα μπορούσε να βασίζεται σε ανάλυση των δημοσιεύσεων των καταναλωτών στα μέσα κοινωνικής δικτύωσης, ή σε παρελθοντικά δεδομένα ενός καταναλωτικού κοινού παρόμοιο με αυτό στο οποίο πρόκειται προσαρμοστεί το προϊόν ή η υπηρεσία, αξιοποιώντας παράλληλα τις σχέσεις ανάμεσα στις ανάγκες των καταναλωτών που έχουν παρατηρηθεί. Ιδανικά το μοντέλο θα έχει ως είσοδο τα χαρακτηριστικά του καταναλωτή ή τις δημοσιεύσεις του, και θα εξάγει αριθμητικές τιμές αξιολογώντας κάθε ανάγκη που υπάρχει ή πρόκειται να δημιουργηθεί, όπως ανάγκη για κάμερα υψηλότερης ανάλυσης στο κινητό τηλέφωνο, ή μεγαλύτερη διάρκεια μπαταρίας.

Πρόβλεψη επίδρασης τεχνικών χαρακτηριστικών στην ικανοποίηση άυλων αναγκών: Ένα από τα στοιχεία του σπιτιού της ποιότητας είναι οι σχέσεις ανάμεσα στις ανάγκες των καταναλωτών και τα τεχνικά χαρακτηριστικά. Όπως εξηγήθηκε, αυτός ο συσχετισμός μπορεί να προκύπτει άμεσα και εύκολα όσο αφορά την ικανοποίηση των υλικών αναγκών των καταναλωτών (πχ ανάγκη για μεγαλύτερη οθόνη υπολογιστή->σχεδιασμός μεγαλύτερης οθόνης), αλλά είναι δύσκολο να εξαχθεί για την ικανοποίηση των άυλων αναγκών. Επομένως μια εφαρμογή Predictive Analytics θα μπορούσε να είναι μια που αξιοποιώντας παρελθοντικά δεδομένα για τα τεχνικά χαρακτηριστικά και απόψεις καταναλωτών για το συνολικό αποτέλεσμα, θα μπορεί να προβλέψει το κατά πόσο το προϊόν ή η υπηρεσία που σχεδιάζεται θα ικανοποιεί τις άυλες ανάγκες που από την φύση τους είναι δύσκολο να μεταφραστούν σε τεχνικά χαρακτηριστικά με ακρίβεια. Έτσι το σπίτι της ποιότητας θα αποκτήσει μια προβλεπτική διάσταση η οποία θα αξιοποιεί όχι μόνο την κριτική των καταναλωτών πάνω στο συγκεκριμένο προϊόν ή υπηρεσία , η οποία φυσικά προκύπτει μετά την εισαγωγή στην αγορά, αλλά και τις κριτικές που εμπεριέχουν τα αποτελέσματα των τεχνικών χαρακτηριστικών.

Πρόβλεψη επίδρασης ικανοποίησης αναγκών: Παρατηρούμε επίσης ότι μία κρίσιμη διαδικασία στον σχεδιασμό είναι η αξιολόγηση των αναγκών των καταναλωτών. Όπως περιγράφεται στο σπίτι της ποιότητας, οι ανάγκες των καταναλωτών πολλές φορές είναι αρνητικά συσχετισμένες και δεν είναι δυνατόν να ικανοποιούνται ταυτόχρονα. Σε αυτό το στάδιο θα ήταν χρήσιμη μια εφαρμογή που θα μπορούσε να μεταφράζει την μελλοντική ικανοποίηση ή την αποτυχία ικανοποίησης μιας ανάγκης σε νούμερα που μπορούν να

αξιοποιηθούν άμεσα από την επιχείρηση, όπως χαμένες πωλήσεις ή αύξηση μεριδίου αγοράς. Η εφαρμογή θα είχε ως δεδομένα τα αποτελέσματα παρελθοντικών ικανοποιήσεων αναγκών των καταναλωτών, έτσι ώστε να προβλέπει τα αποτελέσματα της ικανοποίησης μιας ανάγκης, βοηθώντας έτσι ουσιαστικά στην αξιολόγηση των αναγκών με βάση τα ενδιαφέροντα της επιχείρησης.

Όπως αναλύεται και στο κεφάλαιο 4.3, στην διαδικασία του σχεδιασμού, μπορούν να αξιοποιηθούν και μερικές από της εφαρμογές των Predictive Analytics στην προώθηση και στις πωλήσεις που περιγράφτηκαν στο προηγούμενο κεφάλαιο, όπως η τμηματοποίηση των καταναλωτών με βάση τις ανάγκες τους, η μοντελοποίηση αποχώρησης και το Viral marketing.

2.3 Ο ρόλος των μέσων κοινωνικής δικτύωσης στα Predictive Analytics

Τα μέσα κοινωνικής δικτύωσης πλέον είναι πολύ διαδεδομένα, κάτι που οφείλεται στον πολλαπλασιασμό και την προσιτότητα των συσκευών με πρόσβαση στο Internet, όπως οι ηλεκτρονικοί υπολογιστές, τα κινητά τηλέφωνα και άλλες πιο πρόσφατες καινοτομίες όπως τα tablets. Αυτό επιβεβαιώνεται και από την πρωτοφανή δημοτικότητα διαφόρων μέσων κοινωνικής δικτύωσης όπως το Twitter, το Facebook και το LinkedIn. Τα μέσα κοινωνικής δικτύωσης τέτοιου τύπου πρωτοστατούν στον τεράστιο όγκο δεδομένων σε δικτυακή μορφή, σε μεγάλο αριθμό περιπτώσεων. Συνήθως κατηγοριοποιούνται με βάση τον σκοπό τους, όπως το Facebook που δημιουργήθηκε με αποκλειστικό σκοπό την κοινωνική αλληλεπίδραση, αλλά και το Flickr, που ενώ σχεδιάστηκε με σκοπό να παρέχει διαφορετική υπηρεσία, δηλαδή την δημοσίευση περιεχομένου, επιτρέπει επίσης αλληλεπίδραση σε μεγάλο βαθμό.

Τα μέσα κοινωνικής δικτύωσης έχουν πολλαπλό ρόλο όσο αφορά τα Predictive Analytics. Ο πρώτος ρόλος, είναι η παροχή δεδομένων. Μια επιχείρηση που

δραστηριοποιείται στα μέσα κοινωνικής δικτύωσης, έχει άμεση πρόσβαση στα στοιχεία των καταναλωτών της όπως ηλικία και τοποθεσία, έχει εικόνα της άποψης που έχει το κοινό για το προϊόν ή την υπηρεσία της επιχείρησης σε πραγματικό χρόνο, μπορεί να προωθήσει το προϊόν της στοχευμένα με μικρό κόστος, και ταυτόχρονα να ικανοποιεί ανάγκες εξυπηρέτησης πελατών, όπως τεχνική υποστήριξη. Ένας άλλος ρόλος που παρουσιάζει μεγάλο ενδιαφέρον, είναι το ότι υπάρχει τεράστια πληροφορία που μπορεί να εξαχθεί από την ανάλυση και την μοντελοποίηση του μέσου κοινωνικής δικτύωσης.

Γενικότερα, ένα κοινωνικό δίκτυο μπορεί να οριστεί ως ένας γράφος που αποτελείται από αλληλεπιδράσεις ή σχέσεις, όπου οι κόμβοι αποτελούνται από τους ανθρώπους ή τις σελίδες, και οι ακμές από τις αλληλεπιδράσεις μεταξύ τους. Προφανώς, η ιδέα των κοινωνικών δικτύων δεν περιορίζεται μόνο στην περίπτωση διαδικτυακών μέσων κοινωνικής δικτύωσης όπως το Facebook. Το πρόβλημα της κοινωνικής δικτύωσης έχει μελετηθεί συχνά στην κοινωνιολογία με την μορφή γενικών αλληλεπιδράσεων μεταξύ οποιονδήποτε ομάδων ανθρώπων. Αυτές οι αλληλεπιδράσεις μπορεί να είναι είτε προσωπικές, είτε από τηλεπικοινωνία, από email, η από αλληλογραφία.

Οι συμβατικές μελέτες στη ανάλυση κοινωνικών δικτύων δεν ήταν πάντα συγκεντρωμένες στις διαδικτυακές αλληλεπιδράσεις, και προϋπάρχουν της ραγδαίας εξάπλωσης των υπολογιστών και του ίντερνετ. Ένα κλασσικό παράδειγμα είναι η μελέτη του Milgam που έγινε την δεκαετία του 60, ο οποίος υπέθεσε ότι δύο οποιαδήποτε πρόσωπα στον πλανήτη χωρίζονται το πολύ από 6 στάδια. Ενώ τέτοιου είδους υποθέσεις ήταν απλά εικασίες για τις τελευταίες δεκαετίες, η ανάπτυξη των διαδικτυακών μέσων κοινωνικής δικτύωσης καθιστά πλέον δυνατό τον έλεγχο τέτοιων υποθέσεων, τουλάχιστον σε διαδικτυακό επίπεδο. Αυτό το φαινόμενο ονομάζεται το φαινόμενο του μικρού κόσμου (small world phenomenon). Μία έρευνα σχετική με αυτό το φαινόμενο είχε γίνει στα δεδομένα του MSN messenger, και είχε αποδειχθεί ότι το μέσο μήκος μονοπατιού ανάμεσα σε οποιουσδήποτε δύο χρήστες είναι 6. Αυτό μπορεί να θεωρηθεί σαν επιβεβαίωση του κανόνα των 6 σταδίων που αναφέρθηκε παραπάνω (six degrees of separation) στα μέσα κοινωνικής δικτύωσης. Υπάρχει πληθώρα παρομοίων παραδειγμάτων, καθώς πλέον υπάρχει τεράστιος αριθμός διαδικτυακών δεδομένων που έχει χρησιμοποιηθεί για την επαλήθευση άλλων θεωριών, όπως αυτή της συρρίκνωσης της διαμέτρου, σύμφωνα με την οποία η διάμετρος ενός δικτύου συρρικνώνεται με την πρόσθεση νέων κόμβων, ή της θεωρίας προτιμισιακής προσκόλλησης, σύμφωνα με την οποία όταν ένας καινούργιος κόμβος εισάγεται σε ένα δίκτυο είναι πιο πιθανόν να συσχετιστεί κοντά στις πυκνές περιοχές του δικτύου. Τελικά, η διαθεσιμότητα μαζικών δεδομένων σε διαδικτυακό επίπεδο μας παρέχει ένα νέο εργαλείο για την επιστημονική και στατιστικά ακριβής μελέτη των κοινωνικών δικτύων.

Αυτό είναι το χαρακτηριστικό των διαδικτυακών κοινωνικών δικτύων που μας ωθεί στο να τα αξιοποιήσουμε. Δηλαδή ότι είναι πλούσια σε δεδομένα, και παρέχουν πρωτόγνωρες προκλήσεις και ευκαιρίες από την προοπτική της ανακάλυψης γνώσης και της εξόρυξης δεδομένων. Τα είδη αναλύσεων που εφαρμόζονται πιο συχνά στα κοινωνικά δίκτυα είναι δύο:

Ανάλυση συσχετισμού και δομής: Σε αυτή την ανάλυση, κατασκευάζουμε μια απεικόνιση γράφου των συσχετισμών του δικτύου με σκοπό να αναγνωρίσουμε σημαντικούς κόμβους, κοινότητες, συνδέσμους και αναπτυσσόμενες περιοχές του δικτύου. Αυτή η ανάλυση παρέχει μια καλή συνολική εικόνα της γενικής αναπτυξιακής συμπεριφοράς του δικτύου.

Ανάλυση περιεχομένου: Πολλά κοινωνικά δίκτυα όπως το Flickr και το YouTube περιέχουν τεράστιο ποσό περιεχομένου το οποίο μπορεί να αξιοποιηθεί ώστε να βελτιωθεί η ποιότητα της ανάλυσης. Για παράδειγμα, μια ιστοσελίδα στην οποία οι χρήστες μοιράζονται φωτογραφίες όπως το Flickr, περιέχει πληθώρα κειμένων και φωτογραφιών τα οποία παρέχουν πληροφορίες σε μορφή ετικέτας-χρήστη(user tag).Παρόμοια, τα δίκτυα blog, email και τα message boards περιέχουν κείμενα που συνδέονται μεταξύ τους. Έχει παρατηρηθεί ότι συνδυάζοντας ανάλυση συσχετισμού και περιεχομένου, προκύπτουν πιο αποδοτικά αποτελέσματα σε ποικίλες εφαρμογές.

Η άλλη σημαντική διαφορά που προκύπτει στις αναλύσεις κοινωνικών δικτύων έχει να κάνει με την στατική και την δυναμική ανάλυση. Στην περίπτωση της στατικής ανάλυσης, θεωρούμε ότι το κοινωνικό δίκτυο αλλάζει αργά σε σχέση με τον χρόνο, και κάνουμε μια ανάλυση ολόκληρου του δικτύου χρησιμοποιώντας παρτίδες από χρονικά στιγμιότυπα. Αυτή η περίπτωση χρησιμοποιείται κυρίως σε βιβλιογραφικά δίκτυα, στα οποία τα καινούργια γεγονότα λαμβάνουν χώρα με αργό ρυθμό. Αντίθετα, αναλύοντας δίκτυα που συμπεριλαμβάνουν για παράδειγμα στιγμιαία μηνύματα και δημοσίευση πολλών και μικρών μηνυμάτων, οι αλληλεπιδράσεις προκύπτουν συνεχώς με πολύ μεγάλο ρυθμό. Η ανάλυση αυτών των δικτύων είναι πολύ πιο δύσκολη. Αυτό το χαρακτηριστικό τους, δηλαδή ότι δημοσιεύεται συνεχώς περιεχόμενο σε πραγματικό χρόνο, καθιστά αυτά τα δίκτυα το κέντρο διαφόρων θεωριών που σχετίζονται με την δυναμικότητα και την ανάπτυξη. Πολλά ενδιαφέροντα χρονικά χαρακτηριστικά των δικτύων μπορούν να διακριθούν όπως κοινότητες που αναπτύσσονται, χρήστες των οποίων μεγαλώνει η επιρροή τους, και γεγονότα που έχουν την προσοχή του κόσμου.

Τα δυναμικά δίκτυα επίσης συναντώνται και στις εφαρμογές των κινητών τηλεφώνων (mobile applications), στα οποία οντότητες που κινούνται αλληλεπιδρούν συνεχώς μεταξύ τους. Για παράδειγμα, πολλές συσκευές είναι εφοδιασμένες με δέκτη GPS, τον οποίο εκμεταλλεύονται οι εφαρμογές τους. Μία τέτοια εφαρμογή είναι το Google Latitude, η οποία μπορεί να παρατηρεί την τοποθεσία διαφόρων χρηστών, και να προειδοποιεί όταν κάποιος χρήστης εντοπισθεί σε κάποιο εύρος. Τέτοια δυναμικά δίκτυα μπορούν να μοντελοποιηθούν σαν δυναμικοί γράφοι στους οποίους οι ακμές αλλάζουν συνεχώς. Καθώς ο αριθμός των συνδέσεων μεταξύ χρηστών που δύναται να παρατηρηθούν ταυτόχρονα είναι τεράστιος, η δημιουργία τέτοιων γράφων αποτελεί μεγάλη πρόκληση. Σε αυτές τις περιπτώσεις, απαιτείται η χρήση εφαρμογών που αποτυπώνουν την ροή των αλλαγών σε γράφους ώστε να γίνει σωστή ανάλυση. Συνήθως ζητείται από αυτές τις εφαρμογές να συνοψίσουν την δικτυακή δομή των δεδομένων σε πραγματικό χρόνο και να την χρησιμοποιήσουν για διάφορες εφαρμογές εξόρυξης.

Καταλήγουμε δηλαδή, στο ότι εκτός από την πληθώρα των δεδομένων που μπορούν να εξαχθούν για τους καταναλωτές από τα μέσα κοινωνικής δικτύωσης που αξιοποιούνται σε πολλές εφαρμογές των Predictive Analytics, η μοντελοποίηση και η ανάλυση του κοινωνικού δικτύου που αφορά μια επιχείρηση με βάση τους συσχετισμούς που υπάρχουν στα μέσα κοινωνικής δικτύωσης, μπορεί από μόνη της να είναι προβλεπτική.

2.4 Εταιρείες και Predictive Analytics

2.4.1 Εταιρείες που Παρέχουν Υπηρεσίες Predictive Analytics

2.4.1.1 Revolution Analytics

Η Revolution Analytics (www.revolutionanalytics.com) είναι μια εταιρεία στατιστικού λογισμικού. Ασχολείται συγκεκριμένα με την ανάπτυξη πακέτων λογισμικού για το R (http://www.r-project.org/) το οποίο είναι ένα προγραμματιστικό περιβάλλον εξειδικευμένο στα στατιστικά μαθηματικά και στην στατιστική απεικόνιση. Παρέχει μια τεράστια ποικιλία από στατιστικά εργαλεία, και χαρακτηρίζεται από την δυνατότητα προσαρμογής στις ανάγκες του αναλυτή. Το Revolution R, που είναι το κύριο προϊόν της Revolution Analytics θεωρείται η γρηγορότερη και αποτελεσματικότερη πλατφόρμα

ανάλυσης Big Data. Υποστηρίζει μεγάλη ποικιλία από στατιστικές μεθόδους, μηχανικής εκμάθησης και προβλεπτικής μοντελοποίησης. Χρησιμοποιείται από πολλές εταιρείες κολοσσούς όπως:

- Facebook: Ανάλυση ανανέωσης κατάστασης (status update), ανάλυση γράφου του κοινωνικού δικτύου
- Ford: Στατιστική Ανάλυση και συστήματα αποφάσεων
- Google: Υπολογισμός επιστροφής επένδυσης από διαφήμιση, αύξηση αποδοτικότητας διαδικτυακής διαφήμισης, ανάλυση αποτελεσματικότητας διαφήμισης στην τηλεόραση
- Microsoft: Υπηρεσία αντιστοίχισης παιχτών για την παιχνιδομηχανή Xbox
- Mozilla: Απεικόνιση της διαδικτυακής δραστηριότητας
- Twitter: Επίβλεψη εμπειρίας χρήστη (user experience) στον ιστότοπο

2.4.1.2 SAS

Η SAS (www.sas.com) θεωρείται ηγετική εταιρεία στην παροχή λογισμικού και υπηρεσιών Business Analytics. Υπολογίζεται ότι έχει το 35,4% της αγοράς Analytics. Παρέχει ολοκληρωμένες υπηρεσίες εξόρυξης δεδομένων, στατιστικής ανάλυσης, τεχνικών προβλέψεων, ανάλυσης κειμένου και βελτιστοποίησης. Συνεργάζεται με έναν μεγάλο αριθμό από επιχειρήσεις, τράπεζες, πανεπιστήμια και κυβερνητικούς οργανισμούς. Μερικές από αυτές είναι η Allianz, η Lufthansa, η Hyundai, και η Τράπεζα Πειραιώς.

2.4.1.3 Google Analytics

Η Google Analytics (www.google.com/analytics) ασχολείται με υπηρεσίες διαδικτυακών Analytics βασισμένες στην διαδικτυακή δραστηριότητα. Παρέχει εργαλεία για την εις βάθος ανάλυση της δραστηριότητας στην ιστοσελίδα μιας επιχείρησης, όπως ανάλυση κλικ των πελατών, ανάλυση απήχησης στα μέσα κοινωνικής δικτύωσης και ανάλυση απόδοσης διαφήμισης. Η βασική έκδοση του πακέτου εργαλείων είναι δωρεάν. Μερικές από τις εταιρείες που το χρησιμοποίησαν είναι:

- Puma: Αύξηση διαδικτυακών παραγγελιών κατά 7%
- Nissan: Ανάλυση προτιμήσεων πελατών
- Build Direct: Αύξηση πωλήσεων κατά 50%

2.4.1.4 AlchemyAPI

Η AlchemyAPI (www.alchemyapi.com) είναι μια εταιρεία που προσφέρει υπηρεσίες ανάλυσης των δεδομένων των μέσων κοινωνικής δικτύωσης, και εξειδικεύεται στην ανάλυση κειμένου και την εξόρυξη συναισθήματος. Οι εφαρμογές που σχεδιάζει κατανοούν σε βάθος τις συζητήσεις, τις κριτικές και το περιεχόμενο που διακινείται στα μέσα κοινωνικής δικτύωσης έτσι ώστε οι πελάτες της να μπορούν να προσαρμόσουν την λειτουργία τους στις προτιμήσεις και τις προθέσεις των καταναλωτών. Η εταιρεία βοηθάει τους πελάτες της στην λήψη αποφάσεων μεταφράζοντας ένα τεράστιο αριθμό από ιστοσελίδες, tweets, e-mails και φωτογραφίες σε γεγονότα και πληροφορίες σχετικές με την άποψη των καταναλωτών για το προϊόν, την υπηρεσία ή την καμπάνια τους. Συνεργάζεται με πολλές εταιρείες που έχουν υψηλές απαιτήσεις στην στοχοποίηση πελατών, στην διαδικτυακή διαφήμιση και σχεδιάζουν καμπάνιες δημοσίων σχέσεων.

2.4.1.5 nielsen

Πρόκειται για μια εταιρεία που ασχολείται με την προώθηση και βοηθάει τους πελάτες της να σχεδιάσουν καινοτομικά προϊόντα και να ενισχύσουν την απόδοση της προώθησής τους. Η nielsen (www.affinnova.com) παρέχει ξεχωριστές υπηρεσίες για την σχεδίαση του προϊόντος, την διαφήμισή του και τον σχεδιασμό προωθητικών και επικοινωνιακών μέσων. Σύμφωνα με την επιχείρηση, τα προϊόντα που σχεδιάζονται με βάση τις υπηρεσίες της έχουν 50% μεγαλύτερη πιθανότητα να εισαχθούν στην αγορά, 240% μεγαλύτερη πιθανότητα να έχουν επιτυχία στην αγορά, και αποφέρουν 4,4 φορές περισσότερες πωλήσεις. Έχει συνεργαστεί με εταιρείες όπως η The Coca Cola Company, η Fanta, η Carlsberg, η Energizer και η Red Bull.

2.4.1.6 Angoss

Η Angoss (www.angoss.com) παρέχει ολοκληρωμένες υπηρεσίες Predictive Analytics. Ασχολείται κυρίως με διαχείριση ρίσκου, την προώθηση και τις πωλήσεις. Αξιοποιώντας ένα σύνολο από λογισμικά ανάλυσης και απεικόνισης big data, μετατρέπει την πληροφορία σε επιχειρηματικές αποφάσεις και ανταγωνιστικά πλεονεκτήματα. Συνεργάζεται με πολλές πολυεθνικές εταιρείες ηγέτες στην ασφάλιση, στις οικονομικές υπηρεσίες, στις τηλεπικοινωνίες και στην πρόνοια. Χρησιμοποιώντας τις ολοκληρωμένες υπηρεσίες της

Angoss, οι εταιρείες αυτές αυξάνουν τις πωλήσεις και την παραγωγικότητα, διαχειρίζονται την προώθηση πιο αποδοτικά, και μειώνουν το ρίσκο και τον κίνδυνο. Το λογισμικό της εταιρείας εξειδικεύεται στην αξιολόγηση πίστωσης, τον εντοπισμό απάτης, την ανάλυση κειμένου, και την τμηματοποίηση πελατών. Έχει συνεργαστεί με μεγάλες εταιρείες όπως η Stabucks, το ebay, η Honda και η Mitsubishi.

2.4.2 Εταιρείες που εφάρμοσαν Predictive Analytics στα Μέσα Κοινωνικής Δικτύωσης

2.4.2.1 Lenovo

Η Lenovo (www.lenovo.com) είναι μια κινέζικη πολυεθνική εταιρεία τεχνολογίας. Κατασκευάζει προσωπικούς και φορητούς υπολογιστές, smartphones, tablets και άλλα είδη τεχνολογίας. Θεωρείται ένας από τους ηγέτες στην αγορά τα τελευταία χρόνια.

Η εταιρεία ήταν στο τελευταίο στάδιο σχεδίασης μιας αναβάθμισης στο σχέδιο του πληκτρολογίου ενός από τα κορυφαία μοντέλα προσωπικών υπολογιστών της. Παράλληλα, εφάρμοζε τεχνικές αναγνώρισης αναγκών των καταναλωτών στα μέσα κοινωνικής δικτύωσης. Ερευνώντας το διαδίκτυο για αναφορές του ονόματος της, λίγο πριν εφαρμοστεί η αναβάθμιση του πληκτρολογίου εντοπίστηκε μια δημοσίευση χρήστη που άλλαξε τα σχέδια της εταιρείας. Η συγκεκριμένη δημοσίευση ήταν μια αναλυτική κριτική 6 σελίδων που επικροτούσε το τρέχον τότε μοντέλο του ηλεκτρονικού υπολογιστή και ειδικότερα το πληκτρολόγιο. Επιπλέον η κριτική είχε αποσπάσει παραπάνω από 2.000 σχόλια. Ερευνώντας περαιτέρω την ιστοσελίδα forum στην οποία είχε δημοσιευθεί η κριτική, η Lenovo εντόπισε μια διαδικτυακή κοινότητα φανατικών οπαδών των βιντεοπαιχνιδιών, η οποία ήταν απόλυτα υποστηρικτική στο τρέχον πληκτρολόγιο. Επομένως η εταιρεία αναγνώρισε ότι αλλάζοντας το σχέδιο του πληκτρολογίου ενδεχομένως θα έχανε ένα σημαντικό ποσοστό του μεριδίου αγοράς της, που ταυτόχρονα είναι και από τα πιο επικερδή σε αυτή τη συγκεκριμένη αγορά, και δεν έκανε την αλλαγή.

Μετά από αυτό το γεγονός, η Lenovo αναγνώρισε την αξία της ανάλυσης κειμένου και εξόρυξης συναισθήματος, και την εφάρμοσε και σε άλλους τομείς. Μια δεύτερη νίκη από την εφαρμογή αυτών των μεθόδων αφορούσε την τεχνική υποστήριξη. Για μια χρονική περίοδο, υπήρχαν πολλά παράπονα στα μέσα κοινωνικής δικτύωσης που ανέφεραν ότι σε τυχαίες χρονικές στιγμές η οθόνη του ηλεκτρονικού υπολογιστή παρουσίαζε προβληματική λειτουργία, ή ότι ο ηλεκτρονικός υπολογιστής έκλεινε χωρίς αιτία. Αντίστοιχα τηλεφωνήματα

δεχόταν και η τεχνική υποστήριξη. Εφαρμόζοντας τις τεχνικές ανάλυσης κειμένου στα μέσα κοινωνικής δικτύωσης, διαπιστώθηκε ότι σε ένα μικρό ποσοστό αυτών των δημοσιεύσεων, υπήρχε η λέξη "docking", η σύνδεση δηλαδή της συσκευής με κάποια άλλη. Αναγνωρίστηκε τελικά ότι το λογισμικό που εξυπηρετούσε αυτή τη διαδικασία είχε άμεση σχέση με το πρόβλημα, και οι μηχανικοί της εταιρείας αντιμετώπισαν το πρόβλημα αναπτύσσοντας μια ανανεωμένη έκδοση του λογισμικού. Η Lenovo αναφέρει ότι πριν εφαρμοστούν οι τεχνικές των Predictive Analytics, η αναγνώριση ενός τέτοιου προβλήματος θα χρειαζόταν 60 με 90 ημέρες, καθώς θα έπρεπε πρώτα να γίνει ο εργαστηριακός έλεγχος του υπολογιστή και να εξαχθούν οι απαιτούμενες αναφορές. Εφαρμόζοντας πλέον τις τεχνικές, ο εντοπισμός ενός τέτοιου προβλήματος διαρκεί μία με δύο εβδομάδες. Αναφέρει επιπλέον, ότι ενώ οι πληροφορίες από τις κλήσεις στο τμήμα τεχνικής υποστήριξης είχαν σημασία, οι πληροφορίες που τελικά έλυσαν το πρόβλημα ήταν αυτές από τα μέσα κοινωνικής δικτύωσης, καθώς "Με το Twitter και το Facebook, ο κόσμος περιέγραφε τι έκανε εκείνη τη στιγμή, δηλαδή υπήρχαν δημοσιεύσεις της μορφής ΄Σύνδεσα την συσκευή και συνέβη το Χ'. Είναι ωμή, απροκατάληπτη και πανίσχυρη πληροφορία". Συνολικά, η Lenovo εφαρμόζοντας αυτές τις τεχνικές έχει παρουσιάσει μείωση του χρόνου εντοπισμού προβλήματος κατά παραπάνω από 50%, μείωση του κόστους εγγυήσεων κατά 10-15%, και μείωση των κλήσεων στο κέντρο τεχνικής υποστήριξης κατά 30-50%.

2.4.2.2 Yamaha Motor Europe

Η Yamaha Motor Europe (www.yamaha-motor.eu) είναι το Ευρωπαϊκό τμήμα της Yamaha Motor Corporation. Η εταιρεία προωθεί και πουλάει μηχανές Yamaha, καθώς και μια ποικιλία από οχήματα εδάφους και θαλάσσης, σε 24 Ευρωπαϊκές Χώρες.

Για να σιγουρευτεί η YME ότι τα νέα σχέδια των μηχανών της θα είχαν απήχηση στο κοινό, το τμήμα σχεδίασης της διεξήγαγε έρευνα αγοράς σε πιθανούς πελάτες σε όλη την Ευρώπη. Αν και αυτή η διαδικασία της παρείχε αρκετή πληροφορία, αυτές οι έρευνες αγοράς είχαν πολύ υψηλό οικονομικό κόστος και ήταν πολύ χρονοβόρες. Η εναλλακτική λύση, δηλαδή η έρευνα αγοράς μέσω τηλεφώνου σε τρεις με τέσσερις χιλιάδες άτομα, ήταν πολύ ακριβή για να είναι εφικτή. Επομένως η εταιρεία στράφηκε στο διαδίκτυο ως ένα πιο οικονομικό και αποτελεσματικό μέσο κατανόησης της άποψης των καταναλωτών. Δημιούργησε έναν ιστότοπο με όνομα Yamaha Design Cafe (www.yamaha-motor-europe.com/designcafe), για τον οποίο υπήρχε σύνδεσμος από τον κεντρικό ιστότοπο της εταιρείας. Ο ιστότοπος περιέχει ενδιαφέρουσες πληροφορίες για τις νεότερες μηχανές της

Yamaha, και συνεντεύξεις από μηχανικούς και σχεδιαστές της εταιρείας. Μαζί με τις πληροφορίες και τις συνεντεύξεις, υπάρχουν και σύνδεσμοι για κριτικές που μπορούν να γράψουν οι αναγνώστες. Επιπλέον, σε αυτές τις κριτικές, υπάρχουν κάποιες ερωτήσεις στις οποίες μπορούν να απαντήσουν οι αναγνώστες, οι οποίες αφορούν τεχνικές πληροφορίες για μηχανές, για παράδειγμα διαφορές μεταξύ δύο κινητήρων. Με αυτό τον τρόπο η εταιρία μπορεί να διαχωρίσει τις κριτικές ανάλογα με τις τεχνικές γνώσεις των συγγραφέων, και να δώσει έμφαση σε αυτές που την ενδιαφέρουν, δηλαδή τους αυθεντικούς φαν των μηχανών.

Με αυτή την εφαρμογή, η εταιρία γλυτώνει χρόνο και κόστος καθώς μπορεί να διεξάγει έρευνα αγοράς πολύ πιο γρήγορα και αποτελεσματικά. Επιπλέον ο ιστότοπος λειτουργεί σαν ένα εργαλείο αξιολόγησης προϊόντων που είναι στο στάδιο της σχεδίασης. Συγκεκριμένα, όταν η εταιρεία ήθελε να εισάγει στην Ευρώπη ένα ηλεκτρικό δίκυκλο, που είχε ήδη επιτυχία στην Ιαπωνία, δημοσίευσε της πληροφορίες του δίκυκλου στον ιστότοπο. Η εταιρεία ήθελε να αποφασίσει σχετικά με την προοπτική του ηλεκτρικού δίκυκλου στην Ευρώπη, καθώς ο μέσος Ευρωπαίος οδηγός μηχανής δεν προτιμάει τα ηλεκτρικά δίκυκλα. Μέσα σε δύο εβδομάδες, η σελίδα του ηλεκτρικού δίκυκλου είχε 2.000 απαντήσεις από οπαδούς των δίκυκλων από Γαλλία, Γερμανία και Ιταλία. Αυτή η πληροφορία βοήθησε το τμήμα σχεδίασης της εταιρείας να κατανοήσει στο πώς σκοπεύουν οι Ευρωπαίοι να το χρησιμοποιήσουν, σε τι τιμή είναι διαθέσιμοι να το αγοράσουν και να τροποποιήσουν τα τεχνικά χαρακτηριστικά του ανάλογα. Επιπλέον, το γεγονός ότι οι κριτικές των χρηστών τις ιστοσελίδας συνήθως συγκρίνουν την αντίστοιχη μηχανή με άλλες τις αγοράς, βοηθάει την εταιρεία να τροποποιεί τα παρόντα προϊόντα της ώστε να είναι συνεχώς ανταγωνιστικά.

2.4.2.3 AdTheorent

Η AdTeorent (www.adtheorent.com) είναι μια εταιρεία που προσφέρει υπηρεσίες διαδικτυακής διαφήμισης, και εξειδικεύεται στην διαφήμιση μέσω διαδικτυακών εφαρμογών. Αναλύει εκατοντάδες χιλιάδες δυνατές αντιδράσεις σε διαφημίσεις το δευτερόλεπτο, βασισμένη σε δημογραφικές πληροφορίες, χαρακτηριστικά συμπεριφοράς, πληροφορίες τοποθεσίας, χαρακτηριστικά της συσκευής που χρησιμοποιεί ο καταναλωτής και άλλα κριτήρια, ώστε να επιλεχθεί και να προβληθεί η διαφήμιση που θα φέρει το επιθυμητό αποτέλεσμα.

Η επιχείρηση ήθελε να εξελίξει τις υπηρεσίες της πέρα από την παραδοσιακή προσέγγιση στο πρόβλημα, η οποία χρησιμοποιεί μετρήσεις κλικ στις διαφημίσεις και μελετάει την συμπεριφορά του καταναλωτή αφού δει την διαφήμιση. Η πληροφορίες αυτές είναι πολύτιμες, αλλά προκύπτουν αφού η διαφήμιση έχει προβληθεί, και άρα πληρωθεί. Αναγνωρίζοντας ότι η συνεργατική ανάλυση δομημένης και αδόμητης πληροφορίας

μπορούσε να κάνει την διαδικτυακή διαφήμιση ακόμα πιο σχετική με τον χρήστη, η AdTheorent ανέπτυξε μια πλατφόρμα εξόρυξης διαδικτυακών δεδομένων που συνδυάζει την κλασσική ανάλυση διαδικτυακής διαφήμισης με την ανάλυση κειμένου και συναισθήματος. Αξιοποιώντας την συναισθηματική ανάλυση του καταναλωτή, η εταιρεία ήταν σε θέση να αποφασίσει την σωστή διαφήμιση που χρειάζεται να προβληθεί σε πραγματικό χρόνο, και όχι βασισμένη μόνο στην παρελθοντική συμπεριφορά του. Η πλατφόρμα επιπλέον είχε την δυνατότητα να βελτιώνεται συνεχώς, αφού με βάση τις αποτυχημένες προβολές μπορούσε να προσαρμόσει καλύτερα τη λειτουργία της στην αντιστοίχηση συναισθήματος και διαφήμισης.

Με αυτή τη μέθοδο η AdTheorent έφτασε βαθμό επιτυχημένης διαδικτυακής διαφήμισης 1,3%, που είναι υπερτριπλάσιο του 0,38% που θεωρείται ο μέσος όρος.

2.4.2.4 Pulsar

Η Pulsar (www.pulsarplatform.com) είναι μια εταιρεία που παρέχει υπηρεσίες παρακολούθησης της πληροφορίας που διακινείται στα μέσα κοινωνικής δικτύωσης. Αποτελείται από μια ομάδα έμπειρων αναλυτών των μέσων κοινωνικής δικτύωσης που έχουν παράλληλα πολύ υψηλές στατιστικές ικανότητες. Παρέχει την ολοκληρωμένη πλατφόρμα Pulsar, που αποτελείται από διαφορετικά μοντέλα για ξεχωριστούς σκοπούς όπως εξυπηρέτηση πελατών και ο εντοπισμός νέων τάσεων.

Η Pulsar ήθελε να να είναι ικανή να προσφέρει στους πελάτες της περισσότερες και ποιοτικότερες πληροφορίες από τις βασικές αναφορές της επιχείρησης ή του προϊόντος στα μέσα κοινωνικής δικτύωσης. Επιδίωξε να είναι σε θέση να εξηγήσει ακριβώς με ποιο τρόπο μεταφράζονται αυτές οι μετρήσεις, και επιπλέον να είναι ικανή να εντοπίσει καινούργιες, ανερχόμενες εκφράσεις όπως “selfie” και “ice bucket challenge”. Συγκεκριμένα για τον δεύτερο σκοπό, η εταιρεία αναγνώρισε ότι ο έγκαιρος εντοπισμός και η κατανόηση τέτοιων ανερχόμενων εκφράσεων θα μπορούσε να έχει πολύ μεγάλο ρόλο στη σχεδίαση ή στη διαφοροποίηση ενός προϊόντος. Για αυτούς του λόγους εφάρμοσαν τεχνικές αναγνώρισης γλώσσας δημοσίευσης, εξαγωγής λέξεων-κλειδιών, και εξόρυξης συναισθήματος. Συνδυάζοντας αυτές τις τρεις λειτουργίες, ανέπτυξαν μια πλατφόρμα η οποία μπορεί να στοχεύσει στην πληροφορία συγκεκριμένων κοινοτήτων που ενδιαφέρουν και ακόμα περισσότερο στο ακριβές στοιχείο του προϊόντος που συζητιέται, ενώ παράλληλα αναγνωρίζει και απορρίπτει τα δεδομένα που δεν φέρουν πληροφορία. Επιπλέον, μπορεί να αναγνωρίσει καινούργιους όρους που παρουσιάζονται στα μέσα κοινωνικής δικτύωσης και να προσαρμόσει την λειτουργία τους σε αυτούς, ώστε να εξάγεται η απαραίτητη πληροφορία. Η εταιρία αναφέρει ότι η χρήση της πλατφόρμας στα μέσα κοινωνικής δικτύωσης είναι “σαν να

μπορείς να απομονώσεις ένα όργανο σε μια συμφωνία, να ακούς μόνο αυτό, και να καταλαβαίνεις πως ακριβώς συμπληρώνει το σύνολο".

Χρησιμοποιώντας αυτή την πλατφόρμα, η Pulsar κατάφερε να συσχετίσει με ακρίβεια την δραστηριότητα μιας επιχείρησης στα μέσα κοινωνικής δικτύωσης, με τις ατομικές πωλήσεις, στόχος που θεωρείται το "Χρυσό Δισκοπότηρο της προώθησης".

2.4.2.5 Brainjuicer

Πρόκειται για μια επιχείρηση που προσφέρει υπηρεσίες έρευνας αγοράς. Η Brainjuicer (www.brainjuicer.com) αποσκοπεί στην αποτελεσματική αξιοποίηση της ψυχολογίας, των οικονομικών συμπεριφοράς (behavioral economics) και των κοινωνικών επιστημών, ώστε να παρέχει στους πελάτες της μια καλύτερη κατανόηση και πρόβλεψη της συμπεριφοράς των καταναλωτών. Υποστηρίζει ότι οι καταναλωτές αποφασίζουν για τις αγορές τους με βάση περισσότερο το συναίσθημα παρά την λογική, και φανερώνουν τις προτιμήσεις τους στον διαδικτυακό διάλογο και την διαδικτυακή τους συμπεριφορά.

Συγκεκριμένα, η επιχείρηση σκοπεύει στον συνδυασμό ήδη υπάρχουσων ιδεών με ένα τρόπο που δίνει επιπρόσθετη αξία στο προϊόν ή την υπηρεσία των πελατών της. Απαιτείται από τους αναλυτές της να αναγνωρίσουν ποιο διαδικτυακό περιεχόμενο είναι ενδιαφέρον για το κοινό των πελατών της, να το φιλτράρουν ώστε να κατανοήσουν ποιο περιεχόμενο είναι το πιο σχετικό, και να το οργανώσουν για να ανακαλύψουν της ιδέες που υπόκεινται. Αυτό σημαίνει ότι πρέπει να αναλυθούν εκατοντάδες χιλιάδες δημοσιευμένα αντικείμενα καθημερινά, τα οποία οι αναλυτές ονομάζουν "μέτρια δεδομένα" (medium data). Η ανάλυση όλου αυτού του περιεχομένου που ενδιαφέρει το κοινό των πελατών είναι μια πολύ χρονοβόρα και κουραστική διαδικασία. Για να αντιμετωπισθεί αυτό το πρόβλημα, η Brainjuicer δημιούργησε προγράμματα που αντιγράφουν την συμπεριφορά καταναλωτών οι οποίοι ανήκουν στις συγκεκριμένες ομάδες που ενδιαφέρουν. Αυτό επιτεύχθηκε συναθροίζοντας την διαδικτυακή συμπεριφορά των καταναλωτών και τα δημογραφικά τους χαρακτηριστικά. Επομένως αυτά τα προγράμματα σαρώνουν το διαδίκτυο για ιστότοπους και κοινωνικά μέσα ψάχνοντας για συζητήσεις και περιεχόμενο των καταναλωτών που ενδιαφέρουν. Στο δεύτερο βήμα, χρησιμοποιούνται εφαρμογές αναγνώρισης γλώσσας και εξαγωγής λέξεων κλειδιών, και τα δεδομένα που επιστρέφουν τα προγράμματα στην

επιχείρηση κατηγοριοποιούνται. Τελικά αναγνωρίζεται το περιεχόμενο που ενδιαφέρει, καθώς και συσχετισμοί που υπάρχουν σε αυτό.

Χρησιμοποιώντας αυτά τα προγράμματα, οι πελάτες της brainjuicer δηλώνουν πολύ ικανοποιημένοι. Συγκεκριμένα αναφέρουν ότι συμβάλουν σε μεγάλο βαθμό στην καινοτομία, και συντομεύουν το χρονικό διάστημα ανάμεσα στην σύλληψη των προϊόντων και την παραγωγή τους. Δηλώνουν επίσης ότι στο στάδιο της δοκιμασίας πριν την εισαγωγή του προϊόντος στην αγορά, οι ιδέες που εξάγονται από την χρήση των προγραμμάτων είναι πιο αποτελεσματικές από αυτές που προκύπτουν από άλλες μεθόδους όπως ο καταιγισμός ιδεών.

2.4.2.6 Waggeneredstrom.com

Η Waggeneredstrom (www.waggeneredstrom.com) είναι μια επιχείρηση που παρέχει υπηρεσίες δημοσίων σχέσεων. Για περισσότερα από 30 χρόνια, η WE έχει αναπτύξει στρατηγικά προγράμματα για καινοτόμους πελάτες που επηρεάζουν όλο τον πλανήτη. Απαιτείται από τους αναλυτές της επιχείρησης να δημιουργούν νέα εργαλεία και πλατφόρμες για την αξιόπιστη καταμέτρηση των αποτελεσμάτων μιας καμπάνιας δημοσίων σχέσεων σε πραγματικό χρόνο.

Η επιχείρηση σκοπεύει να βοηθήσει τους πελάτες της να απαντήσουν τα δύο βασικά ερωτήματα "υπήρξε απήχηση; πως βελτιώνεται η απήχηση στην πορεία;". Απαντάει σε αυτά τα ερωτήματα ερευνώντας το περιεχόμενο ιστοσελίδων που ενδιαφέρουν, τα μέσα κοινωνικής δικτύωσης, και τις ειδήσεις, και αναλύοντας περίπου ενάμιση εκατομμύριο tweets και δυόμιση εκατομμύρια δημοσιεύσεις την ημέρα. Η διαδικασία αυτή κάποτε ήταν τελείως χειροκίνητη, απαιτούσε δηλαδή την επίσκεψη τεράστιου αριθμού ιστοσελίδων και την παρακολούθηση πολλαπλών μέσων κοινωνικής δικτύωσης καθημερινά. Για να βελτιστοποιηθεί η διαδικασία, η επιχείρηση ανέπτυξε μια πλατφόρμα εξόρυξης δεδομένων, η οποία αναγνωρίζει νέες τάσεις και ανακαλύπτει πληροφορία για τους ανταγωνιστές, υποδεικνύοντας έτσι πώς και σε ποια σημεία μπορούν να γίνουν βελτιώσεις στην επίδοση μιας καμπάνιας. Η πλατφόρμα χρησιμοποιεί τεχνικές εξαγωγής λέξεων-κλειδιών, εξαγωγής οντότητας (χρήστες, επιχειρήσεις, γεωγραφικά χαρακτηριστικά), εξαγωγής συγγραφέα (αναγνώριση πηγής δημοσιεύσεων), εξαγωγής κειμένου και κατηγοριοποίησης θεματολογίας. Οι τεχνικές αυτές ρυθμίζονται ώστε να αναζητούν το περιεχόμενο που συσχετίζεται με τον πελάτη της επιχείρησης και την καμπάνια του. Αφού τα άρθρα και οι συζητήσεις εντοπιστούν, το περιεχόμενό τους αντιστοιχίζεται στις θεματολογίες που ενδιαφέρουν, ώστε να μπορούν να αναλυθούν περαιτέρω από τους αναλυτές της WE και τους πελάτες της. Το αποτέλεσμα αυτής της εφαρμογής είναι οι πιο πληροφορημένες αποφάσεις, και οι προσαρμογές στρατηγικής και μηνύματος που δημιουργούν θετική απήχηση στην

επιχειρηματικότητα του πελάτη. Η πρώην χειροκίνητη διαδικασία πλέον είναι κατά 80% αυτοματοποιημένη. Τα αποτελέσματα των ερευνών που πριν την πλατφόρμα απαιτούσαν εβδομάδες, πλέον παραδίδονται σε ώρες και σε ένα πολύ αναβαθμισμένο επίπεδο διορατικότητας.

Πλέον η WE είναι πιο αποδοτική και πιο αξιόπιστη στο να απαντάει ερωτήματα των πελατών της σχετικά με το αν έχουν απήχηση στο κοινό που απευθύνονται, αν χρησιμοποιούν τον σωστό διάλογο, τι μπορούν να κάνουν για να βελτιώσουν την προσπάθειά τους, και για το αν προοδεύουν προς το επιθυμητό αποτέλεσμα.

2.4.2.7 ShareThis

Η ShareThis (www.sharethis.com) είναι μια εταιρεία που δραστηριοποιείται στα μέσα κοινωνικής δικτύωσης. Παρέχει εφαρμογές τύπου widgets, που δίνουν νόημα στα δεδομένα από τα μέσα κοινωνικής δικτύωσης, βοηθώντας τους κατόχους ιστοσελίδων και τους διαφημιστές στην παρακολούθηση της συμπεριφοράς στα μέσα κοινωνικής δικτύωσης, ώστε να μπορούν να στοχεύσουν συγκεκριμένες ομάδες καταναλωτών και να αξιολογήσουν την απόδοση των διαδικτυακών στοιχείων τους.

Συγκεκριμένα, η εταιρεία συγκεντρώνεται στην στοχοποίηση υψηλού επιπέδου. Για παράδειγμα μια εταιρεία που κατασκευάζει υβριδικά αυτοκίνητα ενδιαφέρεται να εντοπίσει όχι μόνο τους αγοραστές αυτοκινήτων, αλλά ένα συγκεκριμένο κομμάτι αυτής της ομάδας. Αυτοί οι καταναλωτές θα ενδιαφέρονται σίγουρα για την κατανάλωση του αυτοκινήτου αλλά και για διαφορετικά θέματα όπως η καθαρή ατμόσφαιρα, οι εναλλακτικές πηγές ενέργειας και οι κλιματικές αλλαγές. Επομένως θα δημοσιεύουν περιεχόμενο σχετικό με τους οργανισμούς προστασίας του περιβάλλοντος, άρθρα σχετικά με την άνοδο της ηλιακής και της αιολικής ενέργειας και την άποψη των πολιτικών πάνω σε αυτά. Αν ο κατασκευαστής αυτοκινήτων μπορεί να εντοπίσει τα άρθρα που δημοσιεύονται, κατανοώντας παράλληλα το περιεχόμενο, την ιδέα και το συναίσθημα που αντιπροσωπεύει το κάθε περιεχόμενο, μπορούν να εξαχθούν συμπεράσματα για τα τεχνικά χαρακτηριστικά του αυτοκινήτου που θα ικανοποιούν τις ανάγκες αυτής της ομάδας. Σε πρώτο στάδιο, η εταιρεία είχε αναπτύξει κώδικα που περιοριζόταν στην παρακολούθηση του περιεχομένου που παρακολουθούν οι επισκέπτες μιας ιστοσελίδας και του τρόπου με τον οποίο αυτοί το διαμοιράζουν στα διαφορετικά μέσα κοινωνικής δικτύωσης. Καθώς η αναγνώριση πολύ συγκεκριμένων ομάδων καταναλωτών δεν ήταν εφικτή με αυτή την απλή προσέγγιση, η ShareThis αναγνώρισε ότι πρέπει το συγκεκριμένο περιεχόμενο να αναλυθεί περαιτέρω. Επομένως ενσωμάτωσε στα widgets της τεχνικές εξαγωγής λέξεων-κλειδιών, εξαγωγής οντότητας, κατηγοριοποίησης θεματολογίας και εξόρυξης συναισθήματος. Τα widgets της εταιρείας έχουν πλέον την ικανότητα να

δημιουργούν πολύ πιο ολοκληρωμένα προφίλ για τον κάθε καταναλωτή σε σχέση με τα απλούστερα δημογραφικά του στοιχεία. Αυτό δίνει την δυνατότητα στους πελάτες της να έχουν μεγαλύτερη επαφή με τους χρήστες τις ιστοσελίδας, να στοχεύουν καλύτερα τα διαφημιστικά τους μηνύματα και να μετατρέπουν ανάγκες σε τεχνικά χαρακτηριστικά.

Αξιοποιώντας τα βελτιωμένα αυτά widgets μια εταιρεία fast-food παρατήρησε αύξηση των πελατών της κατά 234%, και μια δεύτερη εταιρεία κατασκευής αυτοκινήτων παρατήρησε αύξηση 34% στη διαδικτυακή δημοτικότητά της.

3 Κατασκευή Προβλεπτικού Μοντέλου

Σε αυτό το κεφάλαιο παρουσιάζονται τα βήματα και οι αποφάσεις που πρέπει να ληφθούν στην διαδικασία ανάπτυξης ενός προβλεπτικού μοντέλου.

Σε αυτό το στάδιο, πριν ξεκινήσουμε την διαδικασία μοντελοποίησης, είναι σημαντικό να διαφοροποιήσουμε τα προβλεπτικά μοντέλα από τα επεξηγηματικά μοντέλα. Το ερώτημα δηλαδή, αν χρειαζόμαστε μία μέθοδο που θα επεξηγεί, η θα προβλέπει. Ενώ αυτοί οι δύο στόχοι είναι διαφορετικοί και ίσως αντίθετοι, οι διαδικασίες που ακολουθούν παρουσιάζουν αρκετές ομοιότητες, και υπάρχει κίνδυνος το μοντέλο που θα δημιουργηθεί τελικά να εξυπηρετεί περισσότερο τον αντίθετο σκοπό.

Ένα επεξηγηματικό μοντέλο είναι ένα στατιστικό μοντέλο που είναι κατασκευασμένο έχοντας ως σκοπό τον έλεγχο αιτιολογικών υποθέσεων που καθορίζουν πώς και γιατί προκύπτουν συγκεκριμένα φαινόμενα. Ξεκινώντας από αιτιολογικά θεωρητικά μοντέλα, δημιουργείται ένα σύνολο υποθέσεων και ελέγχεται χρησιμοποιώντας στατιστικά μοντέλα και συμπεράσματα. Τελικά, ένα επεξηγηματικό μοντέλο αποτελείται από τα στατιστικά μοντέλα που αξιοποιεί, αλλά και από μεθόδους που αξιολογούν την επεξηγηματική του δυνατότητα, που σε αυτή την περίπτωση είναι στατιστικοί δείκτες όπως απόλυτο σφάλμα κτλ.

Αντίστοιχα, ένα προβλεπτικό μοντέλο στοχεύει στο να κάνει αιτιολογημένες προβλέψεις. Εμπεριέχει όλα τα μοντέλα και αλγόριθμους που χρησιμοποιεί, δηλαδή στατιστικά μοντέλα, αλγόριθμους εξόρυξης και πολλά άλλα, αλλά και μεθόδους που αξιολογούν την προβλεπτική ικανότητά τους (predictive power). Η προβλεπτική ικανότητα αναφέρεται στην ικανότητα ενός μοντέλου στο να δημιουργεί ακριβείς προβλέψεις από καινούργιες παρατηρήσεις. Εδώ η έννοια "καινούργιες" μπορεί να ερμηνευθεί καθαρά χρονικά, δηλαδή παρατηρήσεις που θα γίνουν στο μέλλον, αλλά και με την έννοια ότι

ανεξάρτητα από το πότε παρατηρήθηκαν, δεν είχαν συμπεριληφθεί στο αρχικό συνολικό δείγμα με βάση το οποίο κατασκευάστηκε το μοντέλο.

Και οι δύο αυτοί στόχοι θεωρούνται γενικά απαραίτητοι για μία θεωρία και πολλά μοντέλα αποσκοπούν στο να ικανοποιήσουν και τους δύο. Παρ' όλα αυτά, όπως προαναφέρθηκε η επεξήγηση και η πρόβλεψη είναι καλύτερο να θεωρούνται δύο ξεχωριστοί στόχοι μοντελοποίησης. Ενώ δεν είναι αποκλειστικά αντίθετοι, υπάρχει κάποια αρνητική σχέση μεταξύ τους. Καθώς σε κάθε περίπτωση το βέλτιστο επεξηγηματικό μοντέλο θα διαφέρει σε μεγάλο βαθμό από το καλύτερο προβλεπτικό μοντέλο, οποιοδήποτε μοντέλο που αποσκοπεί στην ικανοποίηση και των δύο αυτών στόχων θα πρέπει αναγκαστικά να συμβιβαστεί μεταξύ τους σε κάποιο βαθμό. Για παράδειγμα, σε περίπτωση που ο κύριος στόχος είναι η αιτιολογική επεξήγηση, αλλά ταυτόχρονα ένα επίπεδο προβλεπτικής ικανότητας είναι επιθυμητό, μπορεί να δημιουργηθεί ένα επεξηγηματικό στατιστικό μοντέλο, και ύστερα, σε δεύτερο επίπεδο, να γίνει η αξιολόγηση της προβλεπτικής ικανότητας του χρησιμοποιώντας Predictive Analytics, και η ανάλογη τροποποίηση του μοντέλου σε περίπτωση που δεν ικανοποιεί το προβλεπτικό επίπεδο που ζητείται. Αντίθετα, όταν ο κύριος στόχος είναι η πρόβλεψη αλλά χρειάζεται και ένα επίπεδο επεξηγηματικότητας (πχ. όταν η λογική στην οποία στηρίζονται οι προβλέψεις πρέπει να επεξηγηθεί στους ενδιαφερόμενους), τότε οι τεχνικές Predictive Analytics μπορούν να συγκεντρωθούν σε προβλέψεις και μεθόδους που σχηματίζουν ένα σχετικά διαυγές μοντέλο (άρα αυτόματα πιο επεξηγηματικό) , θυσιάζοντας ίσως μερική προβλεπτική ικανότητα. Επομένως, σχεδιάζοντας ένα μοντέλο και για τους δύο αυτούς σκοπούς, απαιτεί πλήρη κατανόηση των σχέσεων μεταξύ τους και των διαφορών μεταξύ επεξηγηματικής και προβλεπτικής ικανότητας.

Σε πρακτικό επίπεδο, τα μοντέλα αυτά διαφέρουν για δύο κύριους λόγους. Η πρώτη θεμελιώδη διαφορά επεξηγηματικών και προβλεπτικών μοντέλων είναι το διαφορετικό επίπεδο στο οποίο οι δύο αυτοί τύποι μοντέλων λειτουργούν σε σχέση με την αιτιότητα. Ενώ τα επεξηγηματικά μοντέλα βασίζονται σε υποκείμενες αιτιακές σχέσεις μεταξύ θεωρητικών ιδεών, τα προβλεπτικά μοντέλα βασίζονται σε συσχετισμούς μεταξύ μετρήσιμων μεταβλητών. Η διαδικασία προσαρμογής των θεωρητικών μοντέλων και ιδεών σε αιτολογικά μοντέλα και μετρήσιμα δεδομένα δημιουργεί μια δυσκολία ανάμεσα στην ικανότητα επεξήγησης φαινομένων σε ιδεολογικό επίπεδο και στην ικανότητα εξαγωγής ακριβών προβλέψεων σε παρατηρήσιμο επίπεδο.

Ο δεύτερος λόγος για τον οποίο διαφέρουν τα επεξηγηματικά και προβλεπτικά μοντέλα είναι ο τρόπος με τον οποίο βελτιστοποιούνται οι μετρήσεις. Ενώ τα επεξηγηματικά μοντέλα αποσκοπούν στην ελαχιστοποίηση της προκατάληψης του μοντέλου (model bias), ώστε να αποκτηθεί η πιο ακριβής αναπαράσταση του θεωρητικού μοντέλου που εξετάζεται,

τα προβλεπτικά μοντέλα αποσκοπούν στην ελαχιστοποίηση του συνδυασμού προκατάληψης και διακύμανσης προβλέψεων. Ωστόσο, αυτές οι δύο έννοιες συνήθως λειτουργούν αντίθετα μεταξύ τους. Δηλαδή, για να βελτιώσουμε την προβλεπτική ικανότητα, συχνά απαιτείται να θυσιάσουμε θεωρητική ακρίβεια (άρα αύξηση προκατάληψης), για να έχουμε βέλτιστη αιτιολογημένη πρόβλεψη (χαμηλότερη διακύμανση).Ενώ ένα σωστό και συγκεκριμένο επεξηγηματικό μοντέλο συχνά παρουσιάζει κάποιο επίπεδο προβλεπτικής ικανότητας, το βέλτιστο μοντέλο για ένα σύνολο δεδομένων είναι πολύ πιθανό να είναι το χειρότερο δυνατό μοντέλο για μελλοντικά η διαφορετικά σύνολα δεδομένων, κυρίως λόγω overfitting. Με άλλα λόγια, ένα επεξηγηματικό μοντέλο μπορεί να έχει χαμηλή προβλεπτική ικανότητα, ενώ ένα προβλεπτικό μοντέλο που δημιουργήθηκε με βάση το ίδιο σύνολο δεδομένων μπορεί να παρουσιάζει πολύ υψηλή προβλεπτική ικανότητα. Τελικά, η φύση ενός προβλεπτικού μοντέλου που δημιουργείται για να προβλέπει νέες παρατηρήσεις, είναι διαφορετική από την φύση του επεξηγηματικού μοντέλου που δημιουργείται για να ελέγξει ένα σύνολο ήδη υπαρχόντων υποθέσεων. Μία ένδειξη, για παράδειγμα, είναι ότι σε ένα προβλεπτικό μοντέλο, όλες οι προβλεπτικές μεταβλητές πρέπει να είναι διαθέσιμες την στιγμή της πρόβλεψης, ενώ στα επεξηγηματικά μοντέλα δεν υπάρχει τέτοιος περιορισμός. Αν αναλογιστούμε το μοντέλο της γραμμικής παλινδρόμησης, θα συμπεράνουμε ότι μπορεί να χρησιμοποιηθεί και για επεξήγηση και για πρόβλεψη, αλλά τελικά τα μοντέλα θα διαφέρουν σε πολλά σημεία. Οι διαφορές δεν έχουν να κάνουν μόνο με τα στατιστικά κριτήρια που θα χρησιμοποιηθούν για να αξιοποιήσουν το μοντέλο, αλλά είναι εξόφθαλμες σε όλη την διαδικασία της μοντελοποίησης (μεταβλητές που συμπεριλαμβάνονται και απορρίπτονται, μορφή των μεταβλητών, διαχείριση ελλειπών δεδομένων), τον τρόπο με τον οποίο αξιολογείται η ικανότητα του μοντέλου, και το πώς τελικά χρησιμοποιούνται τα αποτελέσματα για να υποστηρίξουν την αντίστοιχη έρευνα.

Συνοψίζοντας, οι διαφορετικές λειτουργίες των επεξηγηματικών και προβλεπτικών μοντέλων, και τα διαφορετικά περιβάλλοντα στα οποία δημιουργούνται και αργότερα λειτουργούν, οδηγούν σε πολλές διαφορές στην διαδικασία δημιουργίας τους, που μεταφράζονται σε διαφορετικά τελικά μοντέλα. Τα τελικά μοντέλα θα διαφέρουν σε επίπεδο επεξηγηματικής και προβλεπτικής ικανότητας, επομένως καθορίζοντας τον στόχο του μοντέλου, θα πρέπει να έχουμε συνειδητοποιήσει κατά πόσο προβλεπτικό και επεξηγηματικό επιθυμούμε να είναι.

3.1 Οι Θεμελιώδεις Εφαρμογές

Παρατηρώντας τις εφαρμογές των Predictive Analytics που αναφέρθηκαν στα προηγούμενα κεφάλαια, μπορούμε να διαπιστώσουμε ότι οι εφαρμογές τους πραγματεύονται κυρίως προβλήματα που έχουν να κάνουν με την αναγνώριση των χαρακτηριστικών των καταναλωτών, την ομαδοποίηση τους και τελικά την εξαγωγή προβλέψεων με βάση αυτά. Επομένως, σε επίπεδο μοντελοποίησης, οι εφαρμογές που απαιτούνται μπορούν να αναλυθούν στις εξής βασικές:

Αποκλειστικά για Κοινωνικά Δίκτυα:

Ανάλυση Επιρροής: Η κοινωνική επιρροή αναφέρεται στην αλλαγή της συμπεριφοράς των ατόμων καθώς επηρεάζονται από άλλα άτομα. Είναι ένα αναγνωρισμένο φαινόμενο στα κοινωνικά δίκτυα. Η ανάλυση της συμπεριφοράς αυτών των αλληλεπιδράσεων αποτελεί άλλη μία πρόκληση στον τομέα των κοινωνικών δικτύων από την οποία μια επιχείρηση μπορεί να αυξήσει τα κέρδη της. Η δύναμη της κοινωνικής επιρροής εξαρτάται από πολλούς παράγοντες όπως η δύναμη των σχέσεων ανάμεσα στους ανθρώπους, τα χρονικά συμβάντα, τα χαρακτηριστικά του κοινωνικού δικτύου και των ατόμων του και άλλα. Τα κοινωνικά δίκτυα δίκτυα δίνουν την δυνατότητα στα άτομα να αλληλεπιδρούν μεταξύ τους, επομένως μπορούν να θεωρηθούν σαν μια δομή που επιτρέπει την διάδοση πληροφορίας. Για παράδειγμα, σημαντικές ειδήσεις αναμεταδίδονται μέσα σε ένα δίκτυο, χρησιμοποιώντας τις σχέσεις μεταξύ των χρηστών. Ενδιαφέρει γενικότερα ο εντοπισμός χρηστών με την μέγιστη επιρροή, οι οποίοι είναι το πιο πιθανόν να αναμεταδώσουν κάποια είδηση στο υπόλοιπο δίκτυο. Τα μέλη με την μεγαλύτερη επιρροή σε ένα κοινωνικό δίκτυο μπορούν να εντοπισθούν χρησιμοποιώντας διάφορες μεθόδους, όπως η ανάλυση των συσχετισμών με γειτονικούς κόμβους στον γράφο που προκύπτει από την ανάλυση του κοινωνικού δικτύου, ή η ανάλυση των δημοσιεύσεων ή αναγνώσεων που έχει ο συγκεκριμένος χρήστης ή ιστοσελίδα. Η ανάλυση επιρροής έχει κύριο ρόλο σε εφαρμογές viral marketing.

Εντοπισμός Κοινοτήτων: Το πρόβλημα αυτό είναι συσχετισμένο με την εύρεση δομικά συσχετισμένων ομάδων στον αντίστοιχο γράφο του κοινωνικού δικτύου. Οι ομάδες

αυτές που συσχετίζονται, ονομάζονται κοινότητες. Η εξαγωγή τέτοιων δομών και η εκμετάλλευσή τους με σκοπό την πρόβλεψη των τάσεων και των αιτιακών σχέσεων σε ένα δυναμικό περιβάλλον, είναι μεγάλης σημασίας. Το πρόβλημα της εύρεσης κοινοτήτων εμφανίζεται και στην ανάλυση στατικών δικτύων που αλλάζουν με αργό ρυθμό, και στην ανάλυση δυναμικών δικτύων των οποίων η δομή εξελίσσεται ραγδαία. Αυτή η πρόκληση συναντάται μόνο στις διαδικτυακές περιπτώσεις, και έχει οδηγήσει στην ανάπτυξη σημαντικού αριθμού αλγορίθμων. Στο πιο βασικό επίπεδο, ο εντοπισμός κοινοτήτων (είτε σε στατικό είτε σε δυναμικό περιβάλλον), μπορεί να βοηθήσει στην κατανόηση ενός κοινωνικού συστήματος. Καθώς ο εντοπισμός κοινοτήτων μπορεί να ταυτιστεί με την μελέτη και ανάλυση γράφων, μας επιτρέπει να συνοψίσουμε τις αλληλεπιδράσεις μέσα σε ένα δίκτυο, προσφέροντας έτσι μια πιο πλούσια κατανόηση του κοινωνικού φαινομένου που υπόκειται. Πέρα από αυτή την βασική κατανόηση ενός δικτύου και του πώς εξελίσσεται, ο εντοπισμός κοινοτήτων συμβάλει στην αναγνώριση μοτίβων. Σε συνεργασία με την ανάλυση επιρροής, ο εντοπισμός κοινοτήτων μπορεί να αξιοποιηθεί στο viral marketing, στην πρόβλεψη αποχώρησης , και στην αναγνώριση ομάδων καταναλωτών που ενδιαφέρουν για διάφορους λόγους.

Πρόβλεψη Συνδέσμων: Πρόκειται για μια εφαρμογή που αφορά αποκλειστικά την μελέτη του αντίστοιχου γράφου του κοινωνικού δικτύου. Καθώς τα κοινωνικά δίκτυα είναι κυρίως δυναμικά, καινούργιοι κόμβοι και ακμές εισάγονται στον αντίστοιχο γράφο συνεχώς. Η κατανόηση των δυνάμεων που καθοδηγούν την εξέλιξη ενός κοινωνικού δικτύου είναι ένα πολύπλοκο πρόβλημα λόγω του μεγάλου αριθμού μεταβλητών παραμέτρων. Ένα συγκριτικά ευκολότερο πρόβλημα είναι η κατανόηση του συσχετισμού ανάμεσα σε δύο κόμβους. Κάποιες από τις ενδιαφέρουσες ερωτήσεις που παρουσιάζονται είναι: Πώς αλλάζει το μοτίβο συσχετισμού στον χρόνο; Ποιοι είναι οι παράγοντες του συσχετισμού; Πώς η σχέση ανάμεσα σε δύο κόμβους επηρεάζεται από άλλους κόμβους; Το πρόβλημα που αναλύεται είναι η πρόβλεψη της πιθανότητας μελλοντικού συσχετισμού ανάμεσα σε δύο κόμβους του γράφου, γνωρίζοντας ότι δεν υπάρχει σχέση ανάμεσα τους στην παρούσα κατάσταση του γράφου.

Ανάλυση Κειμένου και Εξόρυξη Συναισθήματος: Είναι η διαδικασία κατά την οποία τα κείμενα που έχει μια επιχείρηση στη διάθεσή της αναλύονται ώστε να εξαχθεί ένα συμπέρασμα για την συναισθηματική αξία του κειμένου. Τα κείμενα αυτά συνήθως προκύπτουν από τις δημοσιεύσεις των καταναλωτών στα μέσα κοινωνικής δικτύωσης, και ενδιαφέρουν αυτά που αναφέρουν το προϊόν ή την υπηρεσία της επιχείρησης, ή την ίδια την επιχείρηση και ανταγωνιστές της. Γίνεται προσπάθεια αναγνώρισης των συναισθημάτων που εκφράζει το κείμενο

με βάση κυρίως το λεξιλόγιο και τον χαρακτήρα γραφής, ώστε το κείμενο ή ο συγγραφέας αξιολογηθούν με βάση το τι εκφράζουν.

Γενικού Σκοπού:

Ταξινόμηση: Πρόκειται για την αξιολόγηση ή αντιστοίχιση σε ομάδες, συνήθως πελατών ή καταναλωτών, με βάση τις μεταβλητές που ενδιαφέρουν. Αποτελεί κύριο αντικείμενο μελέτης της μηχανικής εκμάθησης. Χρησιμοποιούνται παρελθοντικά δεδομένα, ώστε να ανατεθεί ένας καινούργιος πελάτης σε μια ομάδα της οποίας η συμπεριφορά μπορεί να προβλεφθεί. Μερικές μέθοδοι που χρησιμοποιούνται στην ταξινόμηση είναι τα δέντρα αποφάσεων, οι μηχανές διανύσματος υποστήριξης, και τα νευρωνικά δίκτυα. Η ταξινόμηση χρησιμοποιείται ευρέως στις περισσότερες εφαρμογές των Predictive Analytics, και κυρίως στην αξιολόγηση πίστωσης.

Κανόνες συσχετισμού: Στην στοχευμένη προώθηση και στο σύστημα προτάσεων αγοράς χρησιμοποιούνται κυρίως οι κανόνες συσχετισμού. Η ιδέα των κανόνων συσχετισμού ξεκίνησε από την μελέτη των καθημερινών αγορών των καταναλωτών. Η γενικότερη ιδέα ήταν να εξαχθούν κάποιοι κανόνες συσχετισμού ή εξάρτησης ανάμεσα στα διαφορετικά προϊόντα που τοποθετούν οι καταναλωτές στο καλάθι με τα ψώνια τους, μελετώντας τα προϊόντα που συνήθως αγοράζονται ταυτόχρονα. Ένα πολύ γνωστό παράδειγμα εφαρμογής των κανόνων συσχετισμού είναι η χρήση τους από την αμερικάνικη αλυσίδα λιανικής πώλησης Target. Η αλυσίδα χρησιμοποίησε κανόνες συσχετισμού και κατάλαβε ότι όταν μια γυναίκα είναι έγκυος τείνει να αγοράζει όλα τα προϊόντα εγκυμοσύνης από το ίδιο κατάστημα επομένως είναι ένας πολύ κερδοφόρος πελάτης. Επιπλέον, συνδυάζοντας τους κανόνες συσχετισμού με τα στοιχεία των πελατών της, μπόρεσε να εκτιμήσει ποιες από τις γυναίκες πελάτες της ήταν σε κατάσταση εγκυμοσύνης, και ύστερα να εφαρμόσει στοχευμένη προώθηση προϊόντων εγκυμοσύνης σε αυτές. Η εφαρμογή αυτή δέχτηκε αρνητική κριτική καθώς θεωρήθηκε ότι η έμμεση αποκάλυψη εγκυμοσύνης δεν ήταν ηθική.

Μοντέλα Τεχνικών Προβλέψεων: Πρόκειται για καθαρά μαθηματικές τεχνικές εξαγωγής προβλέψεων, δηλαδή μοντέλα παλινδρόμησης ή ελαχίστων τετραγώνων. Χρησιμοποιούνται όταν υποτίθεται κάποια σχέση γραμμικής φύσεως ανάμεσα στις μεταβλητές που καλείται να συσχετίσει το μοντέλο.

3.2 Καθορισμός Σκοπού

Η δημιουργία ενός προβλεπτικού μοντέλου απαιτεί προσεκτική μελέτη του ποιο σκοπό προορίζεται να ικανοποιήσει, καθώς αυτή η απόφαση έχει κρίσιμο ρόλο στο είδος των εφαρμογών που πρόκειται να χρησιμοποιηθούν. Παρατηρώντας τις εφαρμογές των Predictive Analytics που έχουν αναφερθεί, διαπιστώνεται ότι στις περισσότερες περιπτώσεις, αυτό που ζητείται από ένα προβλεπτικό μοντέλο είναι είτε η πρόβλεψη κάποιας αριθμητικής τιμής (πχ αποχώρηση πελατών), είτε η αξιολόγηση κάποιου καινούργιου δεδομένου, είτε αυτό είναι άνθρωπος ή ιστοσελίδα (πχ καλός δανειολήπτης, ιστοσελίδα με μεγάλη επιρροή). Συχνά παραδείγματα είναι η διαδικασία αναγνώρισης ομάδων πληθυσμού που έχουν μεγαλύτερη πιθανότητα να ανταποκριθούν σε μια διαφημιστική καμπάνια για σκοπούς προώθησης, στον χώρο του ανθρώπινου δυναμικού η αναγνώριση των βέλτιστων υποψηφίων για μια θέση εργασίας, και στον χώρο των οικονομικών η αναγνώριση επιχειρήσεων που διατρέχουν μεγάλο κίνδυνο χρεοκοπίας. Ας δούμε ένα παράδειγμα του πως αναλύονται αυτοί οι στόχοι στις θεμελιώδεις εφαρμογές. Ένα μοντέλο αποχώρησης θα χρησιμοποιεί έναν ταξινομητή, ο οποίος θα αξιολογεί τους πελάτες με βάση την πιθανότητα αποχώρησης αξιοποιώντας τα δεδομένα που έχει στην διάθεσή της η επιχείρηση. Αυτά τα δεδομένα μπορεί να είναι τα βασικά στοιχεία των πελατών, όπως η ηλικία και το φύλο τους, και οι παρελθοντικές αποχωρήσεις πελατών από την επιχείρηση. Άρα ο ταξινομητής θα συσχετίζει αυτά τα βασικά στοιχεία με την πιθανότητα αποχώρησης, βασισμένος σε παρελθοντικά δεδομένα. Η επιχείρηση όμως μπορεί να έχει στην διάθεσή της και δεδομένα από τα μέσα κοινωνικής δικτύωσης. Δηλαδή δημοσιεύσεις χρηστών της που την αναφέρουν, κριτικές για την υπηρεσία που προσφέρει και άλλα. Επομένως εφαρμόζοντας ανάλυση κειμένου και εξόρυξη συναισθήματος, προκύπτει μια διαφορετική προσέγγιση στην μοντελοποίηση αποχώρησης, κατά την οποία η επιχείρηση συγκρίνει τα συναισθήματα που εκφράζονται ανά χρονική περίοδο και εξάγει προβλέψεις. Διαφορετικά, η επιχείρηση μπορεί να θεωρήσει κάποιου τύπου γραμμική εξάρτηση ανάμεσα στα αρνητικά συναισθήματα που προκύπτουν από την εξόρυξη συναισθήματος και την αποχώρηση, και να προσπαθήσει να τα συσχετίσει αξιοποιώντας κάποιο μοντέλο τεχνικών προβλέψεων. Μια άλλη προσέγγιση θα ήταν να συσχετίσει όλα τα παρελθοντικά συναισθήματα των πελατών με τις παρελθοντικές αποχωρήσεις χρησιμοποιώντας έναν ταξινομητή, ο οποίος τελικά θα εξάγει προβλέψεις για την αποχώρηση αξιοποιώντας τα παροντικά συναισθήματα των πελατών.

Παρατηρούμε επομένως, ότι συνδυάζοντας τις θεμελιώδεις εφαρμογές με διαφορετικούς τρόπους, μπορούν να προκύψουν πολλά διαφορετικά μοντέλα που εξυπηρετούν τον ίδιο γενικό σκοπό. Το κριτήριο σε αυτή την απόφαση είναι τα δεδομένα και

τα εργαλεία που έχει στην διάθεσή της η επιχείρηση. Τελικά, το ερώτημα που πρέπει να απαντηθεί σε αυτό το πρώτο στάδιο είναι "Τι θα καλείται να προβλέψει το μοντέλο και τι δεδομένα θα αξιοποιεί για να το κάνει αυτό".

3.3 Συλλογή, Προετοιμασία και Ανάλυση Δεδομένων

3.3.1 Συλλογή Δεδομένων

3.3.1.1 Καθολικοί Κανόνες

Υπάρχουν κάποιοι βασικοί και κατανοητοί κανόνες που ισχύουν για κάθε είδους μοντελοποίηση. Ένας τέτοιος κανόνας αφορά το μέγεθος των δεδομένων. Ισχύει ο κανόνας ότι όσα περισσότερα δεδομένα μπορούν να αξιοποιηθούν, τόσο το καλύτερο, για διάφορους λόγους. Αρχικά, η πρόβλεψη ατομικών παρατηρήσεων έχει πολύ μεγαλύτερη αβεβαιότητα από την πρόβλεψη σε επίπεδο πληθυσμών. Μια επιχείρηση αποσκοπεί στο να προβλέψει μαζική συμπεριφορά, και αυτό δεν είναι δυνατόν να επιτευχθεί με βάση μεμονωμένες ατομικές παρατηρήσεις. Επιπλέον, καθώς ο σκοπός των προβλεπτικών αλγορίθμων είναι η κατανόηση περίπλοκων συσχετισμών που δεν συμβαίνουν σε μικρά δείγματα, η αύξηση του δείγματος μειώνει την προκατάληψη του μοντέλου και την διακύμανση των προβλέψεων. Επίσης, χρειάζονται δεδομένα για την δημιουργία του συνόλου δοκιμής του μοντέλου που θα προκύψει, για την αξιολόγηση της προβλεπτικής του ικανότητας. Δεν είναι εύκολο να καθοριστούν οι κανόνες του ελάχιστου μεγέθους δείγματος, καθώς το απαιτούμενο μέγεθος εξαρτάται και από την φύση των δεδομένων, από τις ιδιότητες του τελικού μοντέλου, και την δυνατή προβλεπτική ικανότητα, που δεν είναι γνωστά στην αρχή της διαδικασίας μοντελοποίησης. Επιπλέον, θέτοντας ένα απαιτούμενο μέγεθος δείγματος, η ικανότητα του αναλυτή να αξιοποιήσει διάφορους αλγόριθμους για να συνδυάσει τα αποτελέσματα τους θα περιοριζόταν.

Ένας άλλος καθολικός κανόνας είναι ότι προτιμούνται τα παρατηρήσιμα δεδομένα από τα πειραματικά, καθώς αντιπροσωπεύουν καλύτερα τους ανεξέλεγκτους παράγοντες όπως ο θόρυβος και η απώλειες κατά τη μέτρηση. Δηλαδή, μια επιχείρηση είναι προτιμότερο να αξιοποιήσει πραγματικά ιστορικά δεδομένα, ή δεδομένα από τα μέσα κοινωνικής δικτύωσης παρά δεδομένα που έχουν προκύψει από άλλες μοντελοποιήσεις ή πειράματα, αφού το πιο πιθανό είναι τέτοια δεδομένα να μην εμπεριέχουν φαινόμενα που είναι χρήσιμα

στην εξαγωγή προβλεπτικών συμπερασμάτων. Αντίθετα, σε ένα επεξηγηματικό μοντέλο, προτιμούνται πειραματικά δεδομένα για τον εντοπισμό αιτιότητας.

3.3.1.2 Συλλογή Δεδομένων από Μέσα Κοινωνικής Δικτύωσης

Όπως έχει προαναφερθεί, μια τεράστια πηγή δεδομένων αποτελούν τα μέσα κοινωνικής δικτύωσης. Ο μεγάλος όγκος πληροφορίας που αφορά τα χαρακτηριστικά των καταναλωτών και κυκλοφορεί στα κοινωνικά δίκτυα, σε συνεργασία με τα δεδομένα της επιχείρησης και την ανάλυση γράφου του κοινωνικού δικτύου, επαρκούν για την πολύπλευρη προσέγγιση οποιουδήποτε προβλεπτικού ερωτήματος. Κάποια πολύ βασικά δεδομένα που ενδιαφέρουν στις περισσότερες περιπτώσεις και προκύπτουν άμεσα από τα μέσα κοινωνικής δικτύωσης, είναι η ηλικία, το φύλο και ίσως η οικογενειακή κατάσταση. Εύκολα συνειδητοποιούμε ότι αυτά τα χαρακτηριστικά έχουν σημαντικό ρόλο στην καταναλωτική συμπεριφορά.

Γενικότερα, οι μετρήσεις που προκύπτουν από τα μέσα κοινωνικής δικτύωσης δεν πρέπει να καταγράφονται ως απλά νούμερα αλλά σε σχέση με προηγούμενα χρονολογικά δεδομένα ή με εξωτερικά δεδομένα, καθώς μία μέτρηση του τύπου "10 θετικές δημοσιεύσεις για το προϊόν" δεν φέρει πληροφορία αν δεν υπάρχει κάποιο μέτρο σύγκρισης. Σε αυτή την περίπτωση μέτρο σύγκρισης θα μπορούσε να είναι οι θετικές δημοσιεύσεις τις προηγούμενης ημέρας, ή οι θετικές δημοσιεύσεις ενός ανταγωνιστικού προϊόντος. Επιπλέον, όταν το προβλεπτικό μοντέλο αφορά κάποιο χρονικό συμβάν, όπως η εισαγωγή ενός καινούργιου προϊόντος στην αγορά, τα δεδομένα από τα μέσα κοινωνικής δικτύωσης θα πρέπει να κατηγοριοποιούνται αυστηρά ανάλογα με το πότε παρατηρούνται σε σχέση με παρόμοια παρελθοντικά συμβάντα, πχ μία εβδομάδα πρίν την εισαγωγή του προϊόντος, μέρα εισαγωγής, μία εβδομάδα μετά την εισαγωγή, ένας μήνας μετά την εισαγωγή[35].

Ανάλυση Επιρροής: Όσο αφορά την ανάλυση επιρροής, που είναι μία από τις θεμελιώδεις εφαρμογές των Predictive Analytics, εκτός από την μεθοδολογία που βασίζεται στην ανάλυση του γράφου του κοινωνικού δικτύου που αναλύεται αργότερα, χρησιμοποιούνται μετρήσεις που προκύπτουν άμεσα από τα μέσα κοινωνικής δικτύωσης. Μία προσέγγιση είναι αυτή κατά την οποία η επιρροή ενός χρήστη αναλύεται σε 3 διαστάσεις:[31]

Μονομορφισμός/Πολυμορφισμός: Αφορά την θεματολογία με την οποία ασχολείται ο χρήστης που εξετάζεται. Οι χρήστες που δημοσιεύουν περιεχόμενο με μια συγκεκριμένη

θεματολογία (πχ ποδόσφαιρο) θεωρούνται μονομορφικοί. Τέτοιοι χρήστες θα πρέπει να αξιολογηθούν υψηλότερα όταν πρόκειται για εφαρμογές προτάσεων αγοράς. Οι χρήστες που δημοσιεύουν περιεχόμενο διαφορετικής θεματολογίας, χαρακτηρίζονται ως πολυμορφικοί και μπορούν να αξιοποιηθούν για την συλλογή δεδομένων, καθώς συνήθως αντιπροσωπεύουν την πλειοψηφία του καταναλωτικού κοινού.

Υψηλή/Χαμηλή συχνότητα: Η συχνότητα εδώ ορίζεται ως η χρονική διάρκεια ανάμεσα σε μια δημοσίευση ενός χρήστη, και την επόμενη δημοσίευση ίδιας θεματολογίας. Ουσιαστικά αυτό το μέγεθος δείχνει την ικανότητα του χρήστη να ξεκινάει συζητήσεις πάνω σε κάποιο θέμα. Οι χρήστες με χαμηλή συχνότητα θα πρέπει να επιλεχθούν ως "σπόροι" για εφαρμογές viral marketing.

Δημιουργός/Διαμοιραστής πληροφορίας: Οι δημιουργοί πληροφορίας είναι οι χρήστες που συνήθως δημοσιεύουν καινούργιο περιεχόμενο, είναι οι πρώτοι που αγοράζουν προϊόντα ή είναι αυτοί που καθορίζουν τις νέες τάσεις. Για την αξιολόγηση της ικανότητας δημιουργίας πληροφορίας ενός χρήστη, χρησιμοποιείται ο λόγος των δημοσιεύσεών του με καινούργιο περιεχόμενο, δια τον αριθμό των συνολικών δημοσιεύσεών του. Αυτός ο διαχωρισμός είναι πολύ σημαντικός για εφαρμογές viral marketing, αφού και οι δύο κατηγορίες χρηστών μπορούν να αξιοποιηθούν: οι δημιουργοί ως "σπόροι" και οι διαμοιραστές ως αναμεταδότες της πληροφορίας.

Οι τρεις αυτές διαστάσεις αναλύονται, και χρησιμοποιούνται για την ταξινόμηση των χρηστών.

Ανάλυση Κειμένου και Εξόρυξη Συναισθήματος: Η ανάλυση κειμένου εφαρμόζεται κυρίως σε διαδικτυακά forums και blogs, όπου τα κείμενα είναι αρκετά μεγάλα για να αναλυθούν. Για σκοπούς τμηματοποίησης, η ανάλυση κειμένου αποσκοπεί στο να αποφασίσει αν η δημοσίευση που εξετάζεται ενδιαφέρει, για παράδειγμα αν αναφέρεται το προϊόν κάποιας επιχείρησης ή η ίδια. Μπορεί να αξιοποιηθεί και στην ανάλυση επιρροής, αφού αναλυτές υποστηρίζουν ότι η ευφράδεια σε μια δημοσίευση της αποδίδει κύρος, επομένως και επιρροή[32]. Στον πίνακα που ακολουθεί φαίνονται κάποιες μετρήσεις ανάλυσης κειμένου που αφορούν συνολικά μια ιστοσελίδα forum [33].

Χρήστης	Δραστηριότητα στο Forum		Αριθμός δημοσιεύσεων
			Αριθμός θεμάτων
	Λεξιλογικές μετρήσεις	Χαρακτήρες	Αριθμός χαρακτήρων ανά δημοσίευση
			Συχνότητα αλφαβητικών

			χαρακτήρων
			Συχνότητα κεφαλαίων χαρακτήρων
			Συχνότητα αριθμητικών ψηφίων
			Συχνότητα κενών
			Συχνότητα χαρακτήρα tab
			Συχνότητα ατομικού αλφαβητικού χαρακτήρα
			Συχνότητα εξειδικευμένων χαρακτήρων
		Λέξεις	Αριθμός λέξεων ανά δημοσίευση
			Συχνότητα μικρών λέξεων (μήκος<4)
			Συχνότητα χαρακτήρων μέσα σε λέξεις
			Μέσος όρος μήκους λέξης
			Μέσος όρος μήκους πρότασης σε χαρακτήρες
			Μέσος όρος μήκους πρότασης σε λέξεις
		Λεξιλόγιο	Άπαξ λεγόμενα
			Άπαξ δισλεγόμενα
	Συντακτικό		Συχνότητα σημείων στίξης
			Συχνότητα λέξεων παύσης
	Δομή		Αριθμός γραμμών ανά δημοσίευση
			Αριθμός προτάσεων ανά δημοσίευση
			Αριθμός παραγράφων ανά δημοσίευση
			Συχνότητα υπερσυνδέσμων

Forum		Αριθμός δημοσιεύσεων
		Αριθμός συμμετεχόντων

Η ανάλυση κειμένου προηγείται συνήθως της εξόρυξης συναισθήματος. Μέσω της εξόρυξης συναισθήματος, μια επιχείρηση μπορεί να εξάγει σημαντική πληροφορία για την άποψη που έχει το καταναλωτικό κοινό για την υπηρεσία ή το προϊόν που προσφέρει. Επομένως μπορεί να χρησιμοποιηθεί σε εφαρμογές προώθησης και συγκράτησης πελατών, και στην αναγνώριση αναγκών ή ικανοποίησης. Μια επιχείριση αναγνωρίζοντας αρνητικό συναίσθημα στις δημοσιεύσεις κάποιων χρηστών, μπορεί να τους στοχοποιήσει με κινήσεις που θα έχουν ως σκοπό την συγκράτησή τους ή να αξιοποιήσει αυτή την πληροφορία για σκοπούς σχεδίασης. Η βασική διαδικασία της εξόρυξης συναισθήματος συνήθως αποτελείται από την εξαγωγή λέξεων-κλειδιών που εκφράζουν συναισθήματα, και την σύγκρισή τους με βάσεις δεδομένων που βαθμολογούν τις λέξεις ανάλογα με το τι εκφράζουν. Οι βαθμολογίες των λέξεων συναθροίζονται ώστε να προκύψει ένα γενικό συναίσθημα για το κείμενο. Μια από τις δυσκολίες που αντιμετωπίζονται σε αυτή τη διαδικασία είναι ότι οι χρήστες στο διαδίκτυο τείνουν να μην τηρούν την ορθογραφία των λέξεων, ή να χρησιμοποιούν πολλούς ιδιωματισμούς με αποτέλεσμα η αναγνώριση των λέξεων από αλγόριθμους να εμποδίζεται σημαντικά. Επιπλέον, η ειρωνεία ή ο σαρκασμός που υπάρχουν πολλές φορές σε δημοσιεύσεις δεν μπορούν να αναγνωριστούν από κάποιο αλγόριθμο. Μία μέθοδος που αντιμετωπίζει σε κάποιο βαθμό το πρόβλημα της ορθογραφίας και των ιδιωματισμών είναι η εξαγωγή της ρίζας από την οποία προέρχεται η κάθε λέξη χρησιμοποιώντας γλωσσολογικούς αλγόριθμους, και η σύγκρισή της με μια αντίστοιχη βάση δεδομένων που αναθέτει ένα σκορ συναισθήματος στην κάθε ρίζα. Ανάλογα την μέθοδο που εφαρμόζεται, τελικά η δημοσίευση η ο χρήστης μπορούν να αξιολογηθούν είτε αριθμητικά (-1,-2,+4 κτλ) είτε λεκτικά όπως υποστηρικτής ή ουδέτερος.

Άλλες Χρήσεις: Η επιχείρηση μπορεί μέσα από τα μέσα κοινωνικής δικτύωσης να έχει ενεργό ρόλο στην δημιουργία δεδομένων που μπορεί να αξιοποιήσει, αντί απλά να τα καταγράφει παθητικά. Στα πλαίσια της συγκράτησης πελατών, η επιχείρηση μπορεί να κερδίσει την υποστήριξη ατόμων και οργανισμών, ανεξάρτητα από το αν συνδέεται επίσημα μαζί τους. Δημιουργώντας και προωθώντας συζητήσεις μέσα στα μέσα κοινωνική δικτύωσης, επωφελείται η εξάπλωση πληροφορίας από στόμα σε στόμα. Δημιουργώντας και συντηρώντας σχέσεις με άτομα που παρουσιάζουν προτίμηση στο προϊόν, η επιχείρηση μπορεί εύκολα να αναγνωρίσει κατά πόσο το καταναλωτικό κοινό την υποστηρίζει και να

διακρίνει τους υποστηρικτές της. Η υποστήριξη που υπάρχει προς την επιχείρηση μπορεί να ποσοτικοποιηθεί ως εξής [34]:

Ενεργοί Υποστηρικτές: Αριθμός ενεργών υποστηρικτών(σε μια χρονική περίοδο)/Συνολικοί υποστηρικτές

Επιρροή Υποστηρικτή: Επιρροή Υποστηρικτή/Συνολική επιρροή υποστηρικτών

Αποτέλεσμα υποστήριξης: Αριθμός αγορών που οφείλονται στην υποστήριξη/Συνολικός όγκος υποστήριξης σε δημοσιεύσεις

Αγορές που οφείλονται στην υποστήριξη συνήθως θεωρούνται οι αγορές που έγιναν μετά από κάποια πρόταση αγοράς από κάποιο φίλο του αγοραστή (εφαρμογή recommend to a friend).

Με το ίδιο σκεπτικό η επιχείρηση μπορεί να ποσοτικοποιήσει το ποσοστό της "φωνής" που έχει στην αγορά. Ξεκινώντας συζητήσεις στα μέσα κοινωνικής δικτύωσης στις οποίες συμμετέχουν οι πελάτες τις, δημιουργείται ένας διάλογος μέσα στον οποίο αναφέρεται η επιχείρηση και που πολλές φορές εξαπλώνεται σε διάφορα μέσα δικτύωσης. Αυτός ο διάλογος τελικά επιστρέφει πληροφορία στην επιχείρηση, σχετική με το πόσο ασχολείται το κοινό με αυτήν[34].

Ποσοστό φωνής (share of voice): Αναφορές Επιχείρησης/Συνολικές αναφορές (επιχείρησης και ανταγωνιστών)

Συμμετοχή κοινού: (Σχόλια+επαναδημοσιεύσεις+συνδέσμοι)/συνολικές προβολές

Εύρος συζήτησης: Συνολικοί συμμετέχοντες/συνολική προβολή στο κοινό

Σημαντική πληροφορία μπορεί να αντληθεί και από τα δεδομένα της ιστοσελίδας ηλεκτρονικού εμπορίου που χρησιμοποιεί μια επιχείρηση. Οι σύγχρονες ιστοσελίδες ηλεκτρονικού εμπορίου μπορούν να ενταχθούν στην κατηγορία των μέσων κοινωνικής δικτύωσης, αφού πλέον στην πλειονότητά τους απαιτούν την εγγραφή των χρηστών (άρα παρέχουν βασικές πληροφορίες όπως ηλικία, φύλο και τοποθεσία), και επιτρέπουν τον σχολιασμό, την βαθμολόγηση και τις προτάσεις προϊόντων. Επομένως αποτελούν ιστότοπους στους οποίους οι χρήστες ανταλλάζουν απόψεις, και έχουν το πλεονέκτημα ότι όλες οι δημοσιεύσεις αφορούν τα προϊόντα και την λειτουργία της επιχείρησης. Ο παρακάτω πίνακας αναλύει κάποια στοιχεία που μπορεί να αξιοποιήσει μια επιχείρηση που δραστηριοποιείται στα μέσα κοινωνικής δικτύωσης, και κατέχει ιστοσελίδα διαδικτυακών πωλήσεων:

Επισκέψεις Σελίδας	Ποιος διαβάζει την σελίδα και ποιο περιεχόμενο ενδιαφέρει;
Πληροφορίες Επισκέπτη	
Αναφορές σε Blog	

Ανάλυση κλικ	
Αναφορές	
Χρόνος στη σελίδα	Τι λέει το κοινό για μια προσφορά;
Περιεχόμενο blog που έγινε αναφορά	
Δημοτικότητα Κριτικών	
Χρόνος στη σελίδα	Πόσο ασχολείται το κοινό με την επιχείρηση; Τι πιθανότητα έχει ένα μήνυμα της επιχείρησης να διαδοθεί γρήγορα λόγω αυτού;
Προτάσεις σε φίλους (suggest to a friend)	
Λόγος σχόλια/δημοσίευση	
Αναφορές σε Blog	
Κριτικές	
Bounce rate (ποσοστό που φεύγει από την σελίδα ακαριαία)	
Προτάσεις σε φίλους	Τι αποτελέσματα είχε η συμμετοχή της επιχείρησης στα μέσα κοινωνικής δικτύωσης;
Conversion (επιτυχημένη πρόταση από φίλο)	
Κριτικές	
Προτάσεις σε φίλους	Πόσο πιθανό είναι οι πελάτες να επιστρέψουν και να προτείνουν το προϊόν σε άλλους;
Αναφορές σε Blog	
Χρόνος στη Σελίδα	
Bounce rate	

Τα στοιχεία αυτά, παράλληλα με τους σκοπούς προώθησης και πωλήσεων για τους οποίους μπορούν να αξιοποιηθούν, δίνουν πληροφορία και για την ποιότητα της διαδικτυακής υπηρεσίας της επιχείρησης, επομένως μπορούν να αξιοποιηθούν στη σχεδίασή της. Κάποιες πιο ειδικευμένες μετρήσεις που θα ενδιέφεραν μια ιστοσελίδα ηλεκτρονικού εμπορίου είναι τα χρόνια χρήσης του διαδικτύου του χρήστη, τα συνολικά χρήματα που έχει ξοδέψει ο χρήστης σε διαδικτυακές αγορές, οι ώρες που χρησιμοποιεί το διαδίκτυο ανά εβδομάδα και πόσες φορές την εβδομάδα ψάχνει για προϊόντα. Για αυτές τις πληροφορίες θα χρειαζόταν κάποια μορφή ερωτηματολόγιου, που όπως παρατηρείται στις σύγχρονες σελίδες ηλεκτρονικού εμπορίου, χρησιμοποιείται συχνά.

3.3.2 Προετοιμασία και Ανάλυση Δεδομένων

Η προετοιμασία των δεδομένων αφορά την μαθηματική επεξεργασία τους ώστε να εξυπηρετούν καλύτερα το σκοπό του μοντέλου. Αρκετές φορές στην προετοιμασία δεδομένων, χρειάζεται η ομαλοποίηση τους για να εξαλειφθούν ο θόρυβος και οι ακραίες τιμές. Το τι ορίζεται ως θόρυβος, ορίζεται από τον αναλυτή, καθώς και το φιλτράρισμα που θα υποστούν τα δεδομένα. Σε μια εφαρμογή εξόρυξης συναισθήματος, θόρυβος θα μπορούσε να θεωρηθεί ένα συναίσθημα που εμπεριέχεται σε κάποιες δημοσιεύσεις, και που σχετίζεται με ένα γεγονός άσχετο με την επιχείρηση, που έχει ως αποτέλεσμα τα νούμερα που αξιολογούν τα συναισθήματα κάθε δημοσίευσης να εμπεριέχουν και την επίδραση αυτού του άσχετου γεγονότος. Αντίστοιχα, ακραία τιμή θα μπορούσε να θεωρηθεί μια αυξημένη τιμή στις αποχωρήσεις πελατών από την επιχείρηση για μια χρονική περίοδο, που προέκυψαν λόγω ενός μεμονωμένου γεγονότος. Οι τιμές σε αυτές τις δύο περιπτώσεις θα πρέπει να ομαλοποιηθούν, για να μην επηρεάζουν αρνητικά την λειτουργία του μοντέλου. Κάποιες από τις μεθόδους είναι:

Κινητός Μέσος: Χρησιμοποιείται για φιλτράρισμα γενικού σκοπού [42 10 11] . Πρακτικά, όλες οι τιμές αντικαθίστανται με τον μέσο όρο που προκύπτει από την αρχική τιμή τους, και των τιμών των δύο προηγούμενων δεδομένων. Με αυτή την τεχνική μειώνεται η διακύμανση των δεδομένων, αλλά ένα μειονέκτημα είναι ότι υποχρεώνει όλες τις τιμές των δεδομένων να έχουν ίδιο βάρος.

Φιλτράρισμα Μέσου: Χρησιμοποιείται κυρίως για χρονοσειρές ώστε να αφαιρεθούν οι ακραίες τιμές και τα κακά δεδομένα. Είναι μια μη γραμμική μέθοδος που τείνει να διατηρεί τα χαρακτηριστικά των δεδομένων [11 12].

Μέσος Κορυφής και Κοιλάδας (Peak-Valley Mean): Είναι άλλη μία μέθοδος αφαίρεσης θορύβου. Εκτιμάει κάθε τιμή ως τον μέσο όρο της τελευταίας κορυφής και κοιλάδας. Κορυφή θεωρείται μια τιμή που είναι υψηλότερη από την προηγούμενη και την επόμενη τιμή, ενώ κοιλάδα είναι μια τιμή που είναι χαμηλότερη από την προηγούμενη και την επόμενη [42 11].

Κανονικοποίησης: Είναι μια μέθοδος κατά την οποία αλλάζουμε τις τιμές των δεδομένων με συγκεκριμένο τρόπο ώστε να αντιπροσωπεύεται η αξία τους σε σχέση με το σύνολο. [42 11]. Τα περισσότερα μοντέλα αποδίδουν βέλτιστα χρησιμοποιώντας κανονικοποιημένα δεδομένα. Ένα παράδειγμα είναι η κανονικοποίηση τυπικής απόκλισης, κατά την οποία αφαιρείται από όλα τα δεδομένα ο μέσος όρος τους, και μετά τα δεδομένα διαιρούνται με αυτόν. Έτσι μειώνεται η διασπορά του δείγματος.

Ελλειπή και κενά δεδομένα: Ένα κοινό πρόβλημα στην προετοιμασία δεδομένων είναι όταν κάποια δεδομένα λείπουν η δεν μπορούμε να τα αποκτήσουμε. Ένα δεδομένο που λείπει είναι

ένα για το οποίο υπάρχει πραγματική τιμή αλλά για οποιοδήποτε λόγο παραλείφθηκε στην εισαγωγή δεδομένων, ενώ ένα κενό δεδομένο θεωρείται ένα για το οποίο δεν υπάρχει η δεν μπορεί να θεωρηθεί πραγματική τιμή. [42 11]. Αυτές οι τιμές πρέπει να αντικατασταθούν κατάλληλα πριν προχωρήσουμε στο επόμενο στάδιο του μοντέλου, καθώς πολλοί αλγόριθμοι δυσκολεύονται να τις διαχειριστούν. Μερικοί αλγόριθμοι αγνοούν τα κενά και ελλειπή δεδομένα, ενώ άλλοι αποφασίζουν αυτόματα μια τιμή για να τα αντικαταστήσουν. Το μειονέκτημα σε αυτή τη μέθοδο είναι ότι δεν ελέγχεται από τον αναλυτή και υπάρχει μεγάλη πιθανότητα εισαγωγής προκατάληψης στο μοντέλο. Υπάρχουν καλύτερες μέθοδοι αντιμετώπισης των ελλειπών και κενών τιμών, στις οποίες ο αναλυτής έχει τον έλεγχο των τιμών που χρησιμοποιούνται για να τις αντικατασταθούν. Το πιο σημαντικό είναι να καταγραφεί το μοτίβο των ελλειπών δεδομένων. Η αντικατάσταση ελλειπών δεδομένων χωρίς την καταγραφή της πληροφορίας που λείπει αφαιρεί πληροφορία από τα δεδομένα. Επομένως χρησιμοποιούνται αμερόληπτοι εκτιμητές. Ένας τρόπος είναι ο υπολογισμός του μέσου όρου των υπαρχόντων δεδομένον και η αντικατάσταση της τιμής που λείπει με αυτόν. Αυτή η τεχνική δεν αλλοιώνει τον μέσο όρο των δεδομένων. Ένας άλλος τρόπος είναι η διατήρηση της τυπικής απόκλισης των δεδομένων. Θεωρείται καλύτερος από τον προηγούμενο γιατί προτείνει τιμές για αντικατάσταση που είναι πιο κοντά στην πραγματική τιμή. Επιπλέον, ο μέσος όρος των δεδομένων που προκύπτουν είναι πιο κοντά στον μέσο όρο των δεδομένων που εμπεριέχουν και τις πραγματικές τιμές.

Μείωση Διαστάσεων:

Η μείωση διαστάσεων και ιδιαίτερα η ανάλυση πρωταρχικών συστατικών έχει εφαρμογή στην συντριπτική πλειοψηφία των προβλεπτικών μοντέλων. Όταν τα δεδομένα περιέχουν περισσότερες μεταβλητές από όσες μπορούν να συμπεριληφθούν στο μοντέλο, είναι απαραίτητο να επιλεχθούν οι πιο αντιπροσωπευτικές από αυτές. Τα δεδομένα που συλλέγονται συνήθως εμπεριέχουν εκατοντάδες χιλιάδες μεταβλητές. Αντίθετα, τα μοντέλα που χρησιμοποιούνται συνήθως μπορούν να αναλύσουν δεδομένα μερικών εκατοντάδων μεταβλητών. Σε ένα προβλεπτικό μοντέλο, των ρόλο των μεταβλητών έχουν τα χαρακτηριστικά που είναι διαθέσιμα. Θεωρούμε για παράδειγμα, ότι μια επιχείρηση έχει στην διάθεσή της ένα μεγάλο σύνολο από δεδομένα για τους πελάτες της που είναι διαφορετικού τύπου, δηλαδή εκτός από τα βασικά στοιχεία (ηλικία, οικογενειακή κατάσταση κτλ.) να υπάρχουν και τα στοιχεία χρήσης του διαδικτύου του πελάτη και οι πολιτικές του απόψεις. Για να αναπαρασταθούν όλα αυτά τα χαρακτηριστικά χρειάζεται ένας πολύ μεγάλος αριθμός μεταβλητών, που είναι δύσκολο να διαχειριστούν από τους αλγόριθμους. Η πλειοψηφία των αλγορίθμων ταξινόμησης συμπεριφέρεται στα δεδομένα και τα χαρακτηριστικά τους σαν

διανύσματα. Δηλαδή ο κάθε πελάτης αναπαριστάται ως ένα διάνυσμα που έχει όσες διαστάσεις όσες και οι μεταβλητές που χρησιμοποιούνται. Επομένως ένας πελάτης θα αντιστοιχίζεται σε ένα διάνυσμα της μορφής ΠελάτηςΑ(30χρονών,παντρεμένος,μεσαίοεισόδημα,κεντροδεξιός,....). Είναι εμφανές ότι ένα τέτοιο διάνυσμα μπορεί να έχει πολύ υψηλό αριθμό διαστάσεων, επομένως να δυσκολεύει τους αλγόριθμους που καλούνται να το αναλύσουν. Σε αυτό το στάδιο εισέρχεται η μείωση διαστάσεων. Η μείωση διαστάσεων βασίζεται στο γεγονός ότι πολλές από αυτές τις μεταβλητές δεν είναι ανεξάρτητες μεταξύ τους, επομένως αν βρεθεί ο συσχετισμός μεταξύ τους θα μπορούν να αναπαρασταθούν από μικρότερο αριθμό μεταβλητών. Για παράδειγμα οι μεταβλητές "ηλικία" και "διαδικτυακή δραστηριότητα" είναι πολύ πιθανό να παρουσιάζουν κάποιο συσχετισμό μεταξύ τους, αφού γενικότερα οι νεότεροι άνθρωποι τείνουν να ξοδεύουν περισσότερο χρόνο στα μέσα κοινωνικής δικτύωσης. Πέρα από αυτούς τους συσχετισμούς όμως, που μπορούν να συλληφθούν εύκολα από τον ανθρώπινο νου, καθώς η μείωση διαστάσεων είναι μια καθαρά μαθηματική διαδικασία, μπορεί να εξάγει συσχετισμούς που δεν θα μπορούσαν να είχαν βρεθεί διαφορετικά. Εφαρμόζοντας μείωση διαστάσεων στα δεδομένα αυτής της επιχείρησης, θα μπορούσε να προκύψει ότι για αυτά τα συγκεκριμένα δεδομένα, η ηλικία τελικά φαίνεται να είναι άμεσα συσχετισμένη με τις δημοσιεύσεις του πελάτη σε ένα μέσο κοινωνικής δικτύωσης, η ακόμα και με τα χρήματα που ξοδεύει ο πελάτης στο διαδίκτυο. Άρα ο πελάτης μπορεί να αναπαρασταθεί χρησιμοποιώντας δύο μεταβλητές λιγότερες. Δύο από τις μεθοδολογίες που χρησιμοποιούνται είναι η ανάλυση πρωταρχικών συστατικών και η ανάλυση συντελεστών συσχετισμού.

Ανάλυση Πρωταρχικών Συστατικών (Principal Component Analysis) [16]: Είναι μια μη επιβλεπόμενη παραμετρική μέθοδος που μειώνει και κατηγοριοποιεί τον αριθμό των μεταβλητών εξάγοντας αυτές με το μεγαλύτερο ποσοστό διακύμανσης στα δεδομένα (πρωταρχικά συστατικά), χωρίς σημαντική απώλεια πληροφορίας [43 15]. Ουσιαστικά μετατρέπει ένα σύνολο συσχετισμένων μεταβλητών σε ένα καινούργιο σύνολο ανεξάρτητων μεταβλητών. Σε περίπτωση που οι αρχικές μεταβλητές είναι σχεδόν ασυσχέτιστες, δεν υπάρχει κάποιο κέρδος στην εφαρμογή της ανάλυσης πρωταρχικών συστατικών. Σε αυτή την περίπτωση, ο αριθμός των μεταβλητών που θα προκύψουν θα είναι ίδιος με τον αρχικό, και δεν υπάρχει τρόπος να μελετηθούν τα δεδομένα ως προς λιγότερες μεταβλητές. Η εξαγωγή των πρωταρχικών συστατικών αποτελεί μια μεγιστοποίηση της διακύμανσης των αρχικών μεταβλητών. Ο σκοπός είναι η μεγιστοποίηση της διακύμανσης των πρωταρχικών συστατικών. Η μέθοδος μετατρέπει γραμμικά τον χώρο που ορίζεται από τα διανύσματα των αρχικών μεταβλητών, σε ένα χώρο με μικρότερες διαστάσεις και ορθογώνια διανύσματα.

Η μέθοδος χρησιμοποιείται στις περιπτώσεις που οι μεταβλητές μετρούνται στις ίδιες η συγκρίσιμες μονάδες και έχουν απόκλιση παρόμοια σε μέγεθος. Σε περίπτωση που οι μεταβλητές δεν μετρούνται σε ίδιες ή συγκρίσιμες μονάδες, πρέπει να εφαρμοστεί κανονικοποίηση πριν εφαρμοστεί η ανάλυση. Η κανονικοποίηση θα δώσει στις μεταβλητές ίδιο βάρος και θα εξαλείψει την επιρροή μίας μεταβλητής στις υπόλοιπες. Αυτή η ανάλυση είναι πολύ χρήσιμη στην απεικόνιση δεδομένων πολλαπλών μεταβλητών. Για όλες τις αναλύσεις δεδομένων, η ανάλυση πρωταρχικών συστατικών αποτελεί συνήθως το προτεινόμενο πρώτο βήμα [44] . Εφαρμόζεται σε ένα σύνολο αρχικών δεδομένων, ώστε να μπορούν μετά να εφαρμοστούν άλλες αναλύσεις μεταβλητών που δεν θα μπορούσαν να διαχειριστούν τον συσχετισμό ανάμεσα στις αρχικές μεταβλητές αποδοτικά. Οι μεταβλητές που προκύπτουν είναι σε σειρά σημαντικότητας, με την πρώτη να αντιπροσωπεύει όσο το δυνατόν περισσότερη διακύμανση των δεδομένων γίνεται, και τις υπόλοιπες να καλύπτουν την υπόλοιπη. Η εξαγωγή αυτή γίνεται χρησιμοποιώντας αποσύνθεση μοναδιαίας αξίας (singular value decomposition) [16] η οποία είναι μια μέθοδος που απλοποιεί έναν πίνακα X σε έναν μοναδιαίο U, έναν διαγώνιο S που έχει το ίδιο μέγεθος με τον X, και έναν τετραγωνικό πίνακα V που έχει το μέγεθος των στηλών του X:

$$X = U \bullet S \bullet V^T$$

Τελικά ο πίνακας X προβάλλεται στον πίνακα V ο οποίος αποτελεί το νέο σύστημα διαστάσεων:

$$z = X \bullet V \ .$$

όπου z είναι οι συντελεστές των πρωταρχικών μεταβλητών.

Ανάλυση συντελεστών συσχετισμού (Correlation Coefficient Analysis): [45] Ασχολείται με την γραμμική εξάρτηση μεταξύ δύο τυχαίων μεταβλητών. Ισούται με την συνδιακύμανση των δύο μεταβλητών διαιρεμένη με την μεγαλύτερη δυνατή συνδιακύμανση των δεδομένων και έχει εύρος από -1 μέχρι +1. Αρνητικός συντελεστής σημαίνει ότι η σχέση μεταξύ των δύο μεταβλητών είναι έμμεση, ή ότι όσο η μία αυξάνεται η άλλη μειώνεται. Θετικός συντελεστής δείχνει μια άμεση σχέση: όσο η μία μεταβλητή αυξάνεται αυξάνεται και η άλλη [46] . Ο

συντελεστής συσχετισμού μπορεί να δειχθεί ως μια εξίσωση της συνδιακύμανσης. Αν ο πίνακας συνδιακύμανσης είναι:

$$\mathrm{cov}(x,y)=\begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{bmatrix}$$

τότε ο συντελεστής συσχετισμού είναι:

$$p_{xy}=\frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

Ένας συντελεστής συσχετισμού μικρότερος ή ίσος του 0,3 δείχνει πολύ μικρή η μηδαμινή σχέση. Μεταξύ των τιμών 0.3 και 0.7 δείχνει ικανοποιητική σχέση, και συντελεστής μεγαλύτερος του 0,7 δείχνει δυνατή γραμμική σχέση.

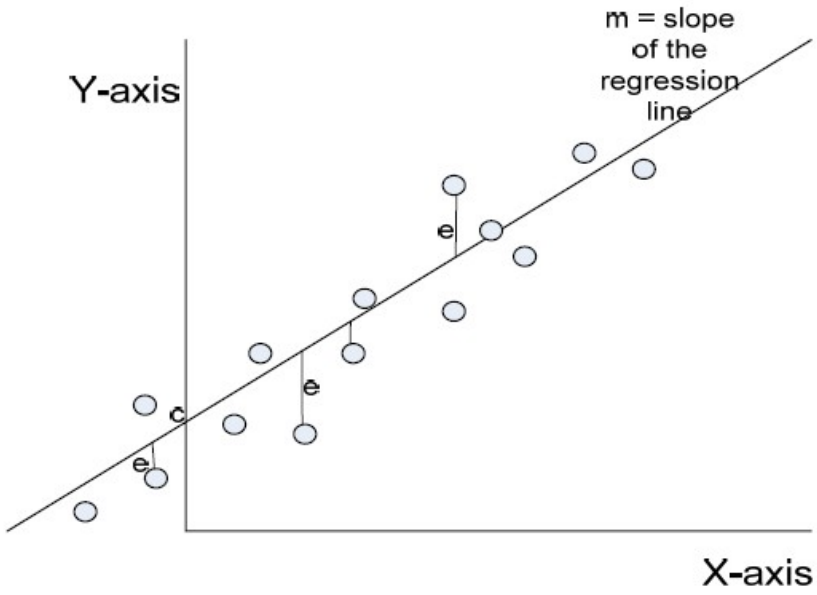
3.4 Μοντελοποίηση και Εφαρμογή

3.4.1 Μοντέλα Τεχνικών Προβλέψεων

Πρόκειται για τα θεμελιώδη μαθηματικά μοντέλα .Τα μοντέλα των τεχνικών προβλέψεων βασίζονται στην πλειονότητά τους στην ανάλυση παλινδρόμησης. Η παλινδρόμηση ουσιαστικά είναι η σχέση μεταξύ επιλεγμένων τιμών του x και παρατηρήσιμων τιμών του y, από τις οποίες η πιο πιθανή τιμή του y μπορεί να προβλεφθεί από οποιαδήποτε τιμή του x. Είναι η εκτίμηση της τιμής μιας πραγματικής συνάρτησης βασισμένη σε περατό σύνολο δεδομένων. Αυτές οι μέθοδοι χρησιμοποιούνται γενικότερα όταν μπορεί να υποτεθεί μια

γραμμική σχέση ανάμεσα στις μεταβλητές που προσπαθούμε να συσχετίσουμε. Κλασσικό παράδειγμα είναι η πρόβλεψη πωλήσεων, όπου γίνεται προσπάθεια συσχέτισης του χρόνου με τις πωλήσεις.

Η γραμμική παλινδρόμηση είναι η πρωτύτερη μέθοδος πρόβλεψης. Μια γραμμική παλινδρόμηση αξιοποιεί την ευθεία συνάρτηση, όπου y=mx+c (m η κλίση, x η μεταβλητή που προβλέπει, και c η σταθερά του κάθετου άξονα). Λαμβάνοντας υπ' όψη τον θόρυβο μπορεί να γραφτεί y=g(x)+e, όπου η g(x)=mx+c και e ο θόρυβος ή το σφάλμα μέσα στο μοντέλο που αντιπροσωπεύει την διαφορά ανάμεσα σε ένα πραγματικό μέγεθος και σε ένα προβλεπόμενο.



Συνήθως, η μεταβλητή x είναι γνωστή αλλά το μοντέλο προσπαθεί να αξιολογήσει την σχέση. Όταν η μεταβλητή x είναι πολλαπλή, τότε ονομάζεται πολλαπλή γραμμική παλινδρόμηση.

Ο όρος γραμμική σημαίνει ότι οι συντελεστές των ανεξάρτητων μεταβλητών είναι γραμμικοί. Μπορεί να υπάρξει η αντίρρηση ότι τα πολυωνυμικά μοντέλα δεν είναι γραμμικά, αλλά στην στατιστική, μόνο οι παράμετροι και όχι οι ανεξάρτητες μεταβλητές εξετάζονται για να αποφασιστεί αν ένα μοντέλο είναι γραμμικό η μη γραμμικό. Αν οι παράμετροι

(συντελεστές ή ανεξάρτητες μεταβλητές) δεν είναι γραμμικές, τότε το μοντέλο γίνεται μη γραμμικό.

Στην εφαρμογή ανάλυση παλινδρόμησης γίνονται κάποιες υποθέσεις:

1. Θεωρείται γραμμική σχέση ανάμεσα στις μεταβλητές εισόδου και εξόδου.
2. Οι συντελεστές σφάλματος e θεωρούνται τυχαίοι, ακολουθούν κανονική κατανομή με μέση τιμή μηδενική και ίση ή σταθερή διακύμανση.
3. Τα σφάλματα είναι ανεξάρτητα
4. Υπάρχουν λίγες ή καθόλου ακραίες τιμές. Επομένως τα δεδομένα θα πρέπει να προετοιμαστούν αξιοποιώντας τις μεθόδους του προηγούμενου κεφαλαίου.
5. Δεν υπάρχουν σημαντικές αλληλεπιδράσεις μεταξύ των μεταβλητών εισόδου
6. Οι μεταβλητές είναι γνωστής μορφής

3.4.1.1 Πολλαπλή Γραμμική Παλινδρόμηση

Το μοντέλο πολλαπλής γραμμικής παλινδρόμησης συσχετίζει ένα σύνολο προβλεπτικών μεταβλητών x σε μια μεταβλητή στόχο y:

$$y = w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots + w_p x_p + \varepsilon$$

όπου w είναι οι συντελεστές της παλινδρόμησης. Μπορεί να παρουσιαστεί και σε μορφή πίνακα, όπου b είναι η σταθερά του άξονα y. Χρησιμοποιείται δηλαδή όταν θέλουμε να συσχετίσουμε την μεταβλητή που προβλέπεται με παραπάνω από μία μεταβλητή. Για παράδειγμα οι πωλήσεις σε σχέση με τον χρόνο και την τιμή του προϊόντος.

$$y = Xw + b + \varepsilon = [X \quad 1] * \begin{bmatrix} w \\ b \end{bmatrix} + \varepsilon$$

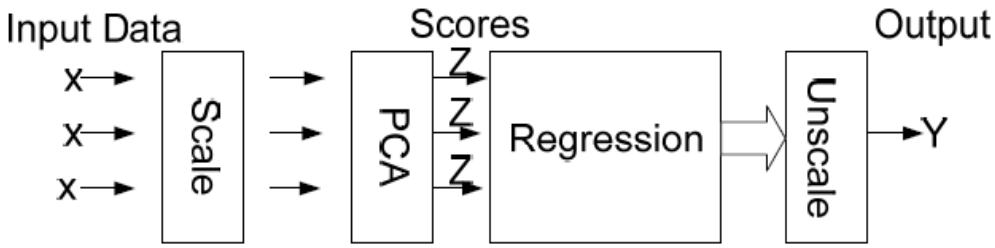
Η εξίσωση αυτή μπορεί να λυθεί για ένα βέλτιστο πίνακα συντελεστών w. Ο πίνακας είναι βέλτιστος όταν το άθροισμα των τετραγωνικών σφαλμάτων (SSE) είναι ελάχιστο:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum^{n} (\mathbf{y} - \mathbf{X}\mathbf{w})^2$$

Καθώς στο μοντέλο υπάρχουν πράξεις με πίνακες, και συγκεκριμένα ζητείται ο αντίστροφος πίνακας του x, απαιτούνται πολλές πράξεις. Ένα άλλο πρόβλημα του μοντέλου είναι ότι πολλές φορές δημιουργείται πρόβλημα λόγω του συσχετισμού των μεταβλητών εισόδου, που μπορεί να προκαλέσει πολύ υψηλή διακύμανση, επομένως απαιτείται τα δεδομένα να υποστούν κάποιου είδους ανάλυσης πρωταρχικών συστατικών.

3.4.1.2 Παλινδρόμηση Πρωταρχικών Συστατικών

Η δεύτερη μέθοδος είναι η παλινδρόμηση πρωταρχικών συστατικών (Principal Component Regression) που βασίζεται στην ανάλυση πρωταρχικών συστατικών η οποία αναφέρθηκε στην ανάλυση δεδομένων. Το διάγραμμα της μεθόδου είναι το εξής:



Αποτελείται από τρία βήματα. Πρώτα υπολογίζονται τα πρωταρχικά συστατικά. Ύστερα επιλέγονται τα συστατικά που ενδιαφέρουν στο συγκεκριμένα προβλεπτικό μοντέλο, και τελικά εισάγονται στο μοντέλο πολλαπλής γραμμικής παλινδρόμησης.

Ο σκοπός των δύο πρώτων βημάτων είναι η αφαίρεση του συσχετισμού ανάμεσα στις μεταβλητές εισόδου και η μείωση των διαστάσεων του πίνακά τους. Επομένως τα δύο προβλήματα της Πολλαπλής Γραμμικής Παλινδρόμησης αντιμετωπίζονται.

3.4.1.3 Παλινδρόμηση Κορυφογραμμής

Η τεχνική παλινδρόμησης κορυφογραμμής αποσκοπεί στο να μετατρέψει τον πίνακα Χ των μεταβλητών εισόδου ώστε να μπορεί να αντιστραφεί. Συγκεκριμένα, ακολουθώντας αυτή τη σειρά πράξεων:

$$y = xw$$

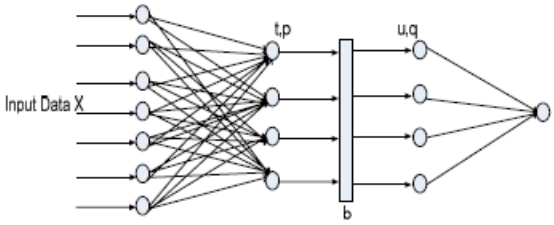
$$x^T y = x^T x w$$

$$w = (X^T X + \alpha^2 I)^{-1} X^T y$$

μειώνει τους συντελεστές παλινδρόμησης. Η πρόσθεση του γινομένου του τετραγωνισμένου όρου α και του μοναδιαίου πίνακα γίνεται για λόγους κανονικοποίησης και το γινόμενο λέγεται παράμετρος κανονικοποίησης. Ουσιαστικά η παράμετρος αυτή ρυθμίζει την ισορροπία μεταξύ την ομαλότητα της λύσης και την ποιότητα των δεδομένων. Η παλινδρόμηση κορυφογραμμής λέγεται και τεχνική ομαλοποίησης γιατί μικραίνει τους συντελεστές, επομένως και τον δείκτη κατάστασης (condition number) που είναι ένας δείκτης αξιολόγησης του μοντέλου.

3.4.1.4 Μερικά Ελάχιστα Τετράγωνα

Μια άλλη μέθοδος είναι αυτή των μερικών ελαχίστων τετραγώνων, (Partial Least Squares -PLS). Είναι μια μέθοδος που μοντελοποιεί τις μεταβλητές εισόδου για να προβλέψει μια μεταβλητή εξόδου. Μετατρέπει να εισαγόμενα δεδομένα x σε μια καινούργια μεταβλητή ή τιμή t και τα δεδομένα εξόδου y σε μια νέα τιμή u και κάνοντας τις καινούργιες αυτές τιμές ανεξάρτητους παράγοντες, αφαιρεί τη σχέση που υπάρχει αρχικά μεταξύ των μεταβλητών εισόδου και εξόδου.



Από το σχήμα, φαίνεται ότι το διάνυσμα b αντιπροσωπεύει την γραμμική αντιστοίχιση ανάμεσα στα διανύσματα t και u. Παρατηρούμε ότι αυτή η μέθοδος παρουσιάζει ομοιότητα με την ανάλυση πρωταρχικών συστατικών. Παρ' όλα αυτά είναι ξεκάθαρο πως ενώ τα ελάχιστα τετράγωνα ασχολούνται με τον πίνακα συσχετισμού ανάμεσα σε εισόδους και εξόδους, η ανάλυση πρωταρχικών συστατικών ασχολείται μόνο με τον συσχετισμό ανάμεσα στις μεταβλητές εισόδου. Για αυτό το λόγο η μέθοδος ελαχίστων τετραγώνων είναι επιβλεπόμενη μέθοδος, ενώ η ανάλυση πρωταρχικών συστατικών μή επιβλεπόμενη. Η επιβλεπόμενη εκμάθηση περιγράφεται στο επόμενο κεφάλαιο.

3.4.1.5 Μη γραμμικά μερικά ελάχιστα τετράγωνα

Η μέθοδος αυτή, (non linear partial least squares) είναι ουσιαστικά ίδια με την προηγούμενη. Είναι η ίδια διαδικασία, με την πολύ σημαντική διαφορά ότι ενώ στην προηγούμενη μέθοδο οι εσωτερικές σχέσεις μοντελοποιούνται με γραμμική παλινδρόμηση, εδώ μοντελοποιούνται με νευρωνικά δίκτυα. Τα νευρωνικά δίκτυα αναλύονται στο επόμενο κεφάλαιο, αφού ανήκουν την κατηγορία των ταξινομητών. Η μέθοδος αυτή χρησιμοποιείται όταν υπάρχει μη-γραμμική σχέση ανάμεσα στις μεταβλητές που πρόκειται να συσχετιστούν, και αβεβαιότητα στα δεδομένα.

3.4.2 Ταξινομητές-Επιβλεπόμενη Εκμάθηση

Αυτό το κεφάλαιο αφορά την θεμελιώδη εφαρμογή της ταξινόμησης. Ο στόχος ενός αλγορίθμου επιβλεπόμενης εκμάθησης είναι η απόκτηση ενός ταξινομητή, ο οποίος προκύπτει μαθαίνοντας από υπάρχοντα παραδείγματα. Ο ταξινομητής μπορεί να χρησιμοποιηθεί για να εξάγει προβλέψεις με βάση τα δεδομένα δοκιμής. Αυτή η εκμάθηση λέγεται επιβλεπόμενη μεταφορικά, καθώς ένας "δάσκαλος" ταξινομεί τα στοιχεία από τα υπάρχοντα δεδομένα δοκιμής.

Κάθε σύνολο δεδομένων δοκιμής και εκμάθησης παρουσιάζεται συνήθως με τον ίδιο τρόπο, δηλαδή ως ένα διάνυσμα γραμμής σταθερού μήκους p. Κάθε στοιχείο του διανύσματος αντιπροσωπεύει ένα παράδειγμα και ονομάζεται αξία του αντίστοιχου χαρακτηριστικού. Μπορεί να είναι πραγματικός αριθμός ή αξία οποιουδήποτε άλλου τύπου. Ένα σύνολο δεδομένων δοκιμής είναι ένα σύνολο από διανύσματα με γνωστή την μεταβλητή y που ενδιαφέρει για κάθε στοιχείο. Για παράδειγμα, στο ερώτημα του ποιος έχει μεγαλύτερη πιθανότητα επιβίωσης σε ένα ναυάγιο (άντρες, γυναίκες, παιδιά), τα δεδομένα από το ναυάγιο του Τιτανικού θα αποτελούσαν δεδομένα εκμάθησης, με γνωστή την μεταβλητή y, που αναγράφει αν επιβίωσε το αντίστοιχο άτομο η όχι. Με βάση αυτά τα δεδομένα, ένας αλγόριθμος μπορεί να μάθει να βγάζει προβλέψεις για την επιβίωση σε ένα μελλοντικό ναυάγιο. Πριν εφαρμοστεί ο αλγόριθμος όμως, πρέπει να ελεγχθεί, για αυτό χρειάζονται τα δεδομένα δοκιμής. Στο παράδειγμά μας, δεδομένα δοκιμής θα μπορούσαν να αποτελέσουν δεδομένα από ένα διαφορετικό παρελθοντικό ναυάγιο, για τα οποία όμως δεν τροφοδοτούμε την τιμή y στον αλγόριθμο. Προφανώς, ο αλγόριθμος δεν έχει ξαναδεί αυτά τα δεδομένα και δεν ξέρει τι εκκεντρικότητες παρουσιάζουν, επομένως εφαρμόζοντας τον σε αυτά τα δεδομένα είναι ένας πολύ καλός τρόπος να τον αξιολογήσουμε. Επαγωγικά, η απόδοση του αλγόριθμου στα δεδομένα δοκιμής, είναι μια βάσιμη εκτίμηση της ικανότητας του μοντέλου να προβλέψει Εφαρμόζοντας τον αλγόριθμο σε αυτά τα δεδομένα, και συγκρίνοντας τις προβλέψεις για τις τιμές y με τις αληθινές παρελθοντικές του δεύτερου ναυάγιου, μπορούμε να αξιολογήσουμε τον αλγόριθμο και να τον τροποποιήσουμε ανάλογα.

Τελικά τα αρχικά δεδομένα χωρίζονται σε δεδομένα εκμάθησης, που χρησιμοποιούνται για την ανάπτυξη του αλγορίθμου, σε δεδομένα δοκιμής που χρησιμοποιούνται για την αξιολόγηση της προβλεπτικής του ικανότητας, και συνήθως υπάρχει και ένα σύνολο που λέγεται δεδομένα επικύρωσης, που χρησιμοποιούνται για την

επιλογή αλγορίθμου. Αν τα δεδομένα είναι μικρά σε μέγεθος για να χωριστούν, χρησιμοποιούνται τεχνικές διασταυρωμένης επικύρωσης.

Το πρόβλημα του overfitting/overleaning: Στην εφαρμογή της επιβλεπόμενης εκμάθησης, έχουμε ένα σύνολο εκμάθησης και ένα σύνολο δοκιμής, στο οποίο οι τιμές που αναζητούνται θεωρούνται άγνωστες, και ο αλγόριθμος αξιολογείται με βάση το πόσο σωστά θα τις προβλέψει. Αυτό που συμβαίνει στην πράξη, είναι ότι κατά την διαδικασία επεξεργασίας των δεδομένων εκμάθησης, ο αλγόριθμος ψάχνει για μοτίβα και συσχετισμούς ανάμεσα σε μεγέθη και κατηγορίες. Μερικά από τα μοτίβα που εντοπίζονται μπορεί να είναι πολύ σπάνια και ιδιαίτερα, δηλαδή να ισχύουν για τα δεδομένα εκμάθησης, αλλά να μην είναι βάσιμα ή να μην έχουν τόση ισχύ στον συνολικό πληθυσμό. Ένας ταξινομητής που βασίζεται σε αυτά τα μοτίβα θα έχει υψηλή προβλεπτική ικανότητα στα δεδομένα εκμάθησης αλλά όχι στον συνολικό πληθυσμό. Αυτό το φαινόμενο, το να βασίζεται δηλαδή ο αλγόριθμος σε μοτίβα που ισχύουν μόνο στα δεδομένα εκμάθησης, λέγεται overleaning, και σχετίζεται άμεσα με την προκατάληψη του μοντέλου, που αναφέρθηκε προηγούμενα στην διαφοροποίηση περιγραφικού και προβλεπτικού μοντέλου. Ουσιαστικά, φαινόμενο overlearning έχουμε όταν ο αλγόριθμός μας προσαρμόζεται παντελώς στα δεδομένα που του εισάγαμε για να δημιουργηθεί, με αποτέλεσμα να έχει μικρότερη προβλεπτική ικανότητα σε καινούργια δεδομένα.

Για αυτό το λόγο, είναι απαραίτητο να δοκιμασθεί ο αλγόριθμος στα δεδομένα δοκιμής. Σε συνεργασία με την μέτρηση lift, που αντιπροσωπεύει το πόσες καινούργιες περιπτώσεις μπορεί να προβλέψει σωστά ο αλγόριθμος, ο αλγόριθμος ρυθμίζεται ανάλογα ώστε να αποφεύγεται το overlearing. Για παράδειγμα, στην περίπτωση των δέντρων αποφάσεων, ο αλγόριθμος ίσως χρειαστεί να σταματάει την εκμάθηση πριν φτάσει στα τελευταία φύλα, που αποτελούν πολύ εξειδικευμένες περιπτώσεις και ίσως να μην υπάρχουν σε διαφορετικά σύνολα δεδομένων.

3.4.2.1 Η αφελής μέθοδος του Bayes

Πρόκειται για τον απλούστερο ταξινομητή. Βασίζεται στην αρχή της δεσμευμένης πιθανότητας του Bayes. Η αφελής μέθοδος του Bayes υπολογίζει την πιθανότητα που έχει μια καινούργια παρατήρηση να ανήκει σε κάθε υπάρχουσα κατηγορία, και τελικά την αναθέτει στην πιο πιθανή. Θα χρησιμοποιηθεί ένα παράδειγμα για να εξηγηθεί η λειτουργία του. Υποθέτουμε ότι έχουμε έναν πελάτη και θέλουμε να αποφασίσουμε για το αν θα απαντήσει θετικά σε μια προσφορά μας η όχι. Επομένως μαθηματικά θέλουμε να μάθουμε αν η πιθανότητα να απαντήσει θετικά είναι μεγαλύτερη από την πιθανότητα να μην απαντήσει.

Αρχικά γνωστό είναι μόνο το φύλο του πελάτη, που εδώ θεωρείται αρσενικός, που είναι το ενδεχόμενο Α. Το ενδεχόμενο να ανταποκριθεί θετικά είναι το ενδεχόμενο Β. Σε αυτή την περίπτωση η δεσμευμένη πιθανότητα του Bayes δίνεται από τον τύπο:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}.$$

Δηλαδή, η πιθανότητα ενός άντρα πελάτη να ανταποκριθεί θετικά είναι το γινόμενο της πιθανότητας ένας πελάτης που ανταποκρίθηκε θετικά να είναι άντρας, και της πιθανότητας του ενδεχόμενου αγοράς, διαιρεμένο με την πιθανότητα του πελάτη να είναι άντρας. Αυτές οι πιθανότητες προκύπτουν εύκολα με βάση το ιστορικό των πελατών της εταιρίας. Αντίστοιχα υπολογίζεται η πιθανότητα ενός άντρα πελάτη να μην ανταποκριθεί στην προσφορά, και τελικά ο πελάτης ταξινομείται ανάλογα με το ποια πιθανότητα είναι η μεγαλύτερη.

Το προηγούμενο παράδειγμα υπέθεσε ότι η πιθανότητα ανταπόκρισης εξαρτάται μόνο από ένα χαρακτηριστικό του καταναλωτή, το φύλο. Φυσικά στην πραγματικότητα αυτή η πιθανότητα εξαρτάται από διάφορα χαρακτηριστικά, που εδώ θα θεωρηθεί ότι είναι το αν ο καταναλωτής είναι ενήλικος (ενδεχόμενο Γ), και το αν ο καταναλωτής εργάζεται (ενδεχόμενο Δ). Ψάχνουμε το ενδεχόμενο ο καταναλωτής να απαντήσει θετικά, δεδομένου ότι είναι ενήλικος εργαζόμενος άντρας. Σε αυτή την περίπτωση, ισχύει ο προηγούμενος τύπος, με την αλλαγή ότι εδώ το αντίστροφο ενδεχόμενο, δηλαδή το ένας πελάτης που ανταποκρίθηκε θετικά να είναι ενήλικος εργαζόμενος άντρας, υπολογίζεται θεωρώντας όλα τα ενδεχόμενα ανεξάρτητα μεταξύ τους, δηλαδή:

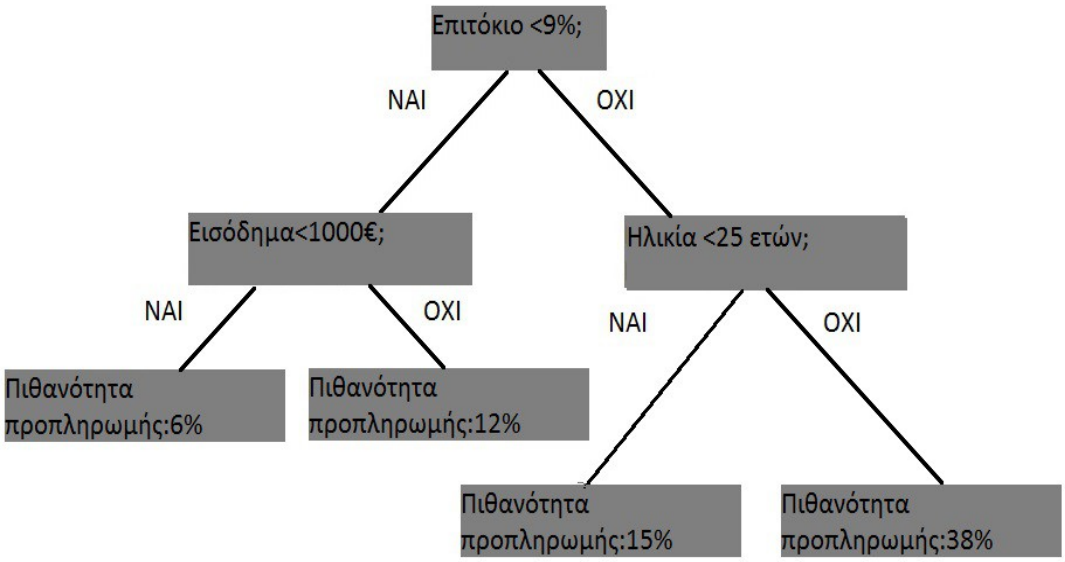
$$P(A,Γ,Δ|B) = P(A|B)*P(Γ|B)*P(Δ|B)$$

Αυτή η απλοποίηση γίνεται για να μειωθούν οι υπολογισμοί.

Η αφελής μέθοδος του Bayes είναι γρήγορη στην εκμάθηση αφού χρειάζεται μόνο μια σάρωση των δεδομένων, και είναι εξίσου γρήγορη στην ταξινόμηση. Επιπλέον δεν επηρεάζεται από χαρακτηριστικά που δεν είναι σχετικά, αφού στον υπολογισμό των γινομένων των πιθανοτήτων θα έχουν παρόμοια αν όχι ίδια τιμή. Το μειονέκτημα της είναι ότι καθώς θεωρεί ανεξάρτητα τα χαρακτηριστικά, εισάγεται σφάλμα στις πιθανότητες.

3.4.2.2 Δέντρα Αποφάσεων

Τα δέντρα αποφάσεων είναι δέντρα που ταξινομούν τα δεδομένα εισόδου με βάση τα χαρακτηριστικά τους. Απεικονίζουν τα δεδομένα χρησιμοποιώντας κόμβους και κλαδιά. Οι κόμβοι είναι τριών ειδών και τα κλαδιά δύο ειδών. Ένας κόμβος απόφασης είναι ένα σημείο στο οποίο πρέπει να απαντηθεί ένα ερώτημα σχετικό με τα χαρακτηριστικά της οντότητας που εξετάζεται, και αντιπροσωπεύεται με τετράγωνο. Τα κλαδιά που ξεκινάνε από έναν κόμβο απόφασης αποκαλούνται κλαδιά απόφασης, και δείχνουν την πορεία που πρέπει να ακολουθηθεί μέσα στο δέντρο ανάλογα την απάντηση στο ερώτημα του κόμβου απόφασης. Οι απαντήσεις στο ερώτημα που εκφράζονται από τα κλαδιά πρέπει να είναι αποκλειστικές μεταξύ τους με την έννοια ότι αν ακολουθηθεί ένα κλαδί δεν μπορεί να ακολουθηθεί και άλλο ταυτόχρονα, και στο σύνολό τους να καλύπτουν όλα τα πιθανά ενδεχόμενα. Το δεύτερο είδος κόμβου είναι ο τερματικοί κόμβοι(φύλα), που βρίσκονται στον πάτο του δέντρου, και συμβολίζουν το τελικό συμπέρασμα που εξάγεται από το δέντρο. Ξεκινώντας από την κορυφή, τα ερωτήματα που αφορούν τα χαρακτηριστικά της οντότητας που εξετάζεται απαντώνται από τον αλγόριθμο, και ακολουθούνται τα αντίστοιχα κλαδιά μέχρι ο αλγόριθμος να φτάσει στον πάτο του δέντρου και να εξαχθεί το συμπέρασμα. Η τρίτη κατηγορία κόμβων και η δεύτερη κατηγορία κλαδιών αφορούν την μελέτη χρονικών γεγονότων με βάση τα οποία η επιχείρηση καλείται να αποφασίσει, και δεν ενδιαφέρουν για σκοπούς ταξινόμησης. Στα πλαίσια της μηχανικής εκμάθησης και την λειτουργία των δέντρων αποφάσεων ως ταξινομητές, τα δεδομένα εκμάθησης χρησιμοποιούνται για να κατασκευαστεί το δέντρο. Επιστρέφουμε στο παράδειγμα της εισαγωγής 1.2:



Παρατηρούμε ότι σε αυτό το παράδειγμα μας ενδιαφέρει η πιθανότητα που έχει ο πελάτης να προπληρώσει το δάνειο. Επομένως οι κόμβοι στον πάτο του δέντρου αντιπροσωπεύουν το τελικό συμπέρασμα που εξάγεται για τον κάθε πελάτη με βάση τα χαρακτηριστικά του πελάτη και του δανείου που έχει ληφθεί. Η πιθανότητα προπληρωμής έχει προκύψει από τα παρελθοντικά δεδομένα εκμάθησης του αλγορίθμου, δηλαδή από το πόσοι παρελθοντικοί πελάτες προπλήρωσαν το δάνειο και είχαν τα χαρακτηριστικά που αντιστοιχούν σε κάθε τερματικό κόμβο.

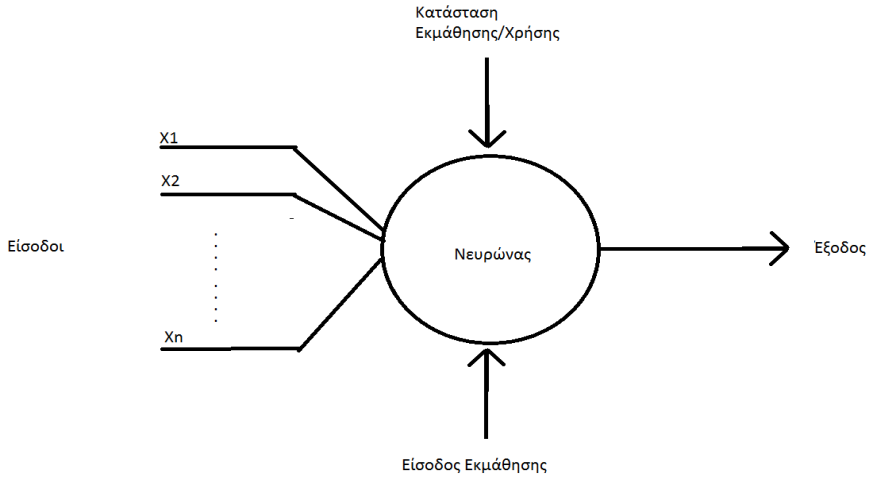
3.4.2.3 Τεχνητά Νευρωνικά Δίκτυα (neural networks)

Τα τεχνητά νευρωνικά δίκτυα (artificial neural network-ANN) είναι μια μέθοδος επεξεργασίας πληροφορίας που εμπνεύστηκε από τον τρόπο με τον οποίο τα βιολογικά νευρικά συστήματα όπως ο εγκέφαλος επεξεργάζονται την πληροφορία. Το κύριο στοιχείο αυτής της μεθόδου είναι η καινοτομική δομή του συστήματος επεξεργασίας πληροφορίας. Αποτελείται από έναν μεγάλο αριθμό από διασυνδεδεμένα στοιχεία επεξεργασίας που

λέγονται νευρώνες, οι οποίοι συνεργάζονται για να λύσουν προβλήματα. Τα νευρωνικά δίκτυα, όπως οι άνθρωποι μαθαίνουν από παραδείγματα. Ένα τεχνητό νευρωνικό δίκτυο ρυθμίζεται για μια συγκεκριμένη εφαρμογή, όπως η αναγνώριση μοτίβων ή η ταξινόμηση δεδομένων, μέσω της διαδικασίας εκμάθησης. Όπως στα βιολογικά συστήματα όπου η εκμάθηση γίνεται ρυθμίζοντας τις συνδέσεις μεταξύ των νευρώνων, το ίδιο ισχύει και στα τεχνητά νευρωνικά δίκτυα.

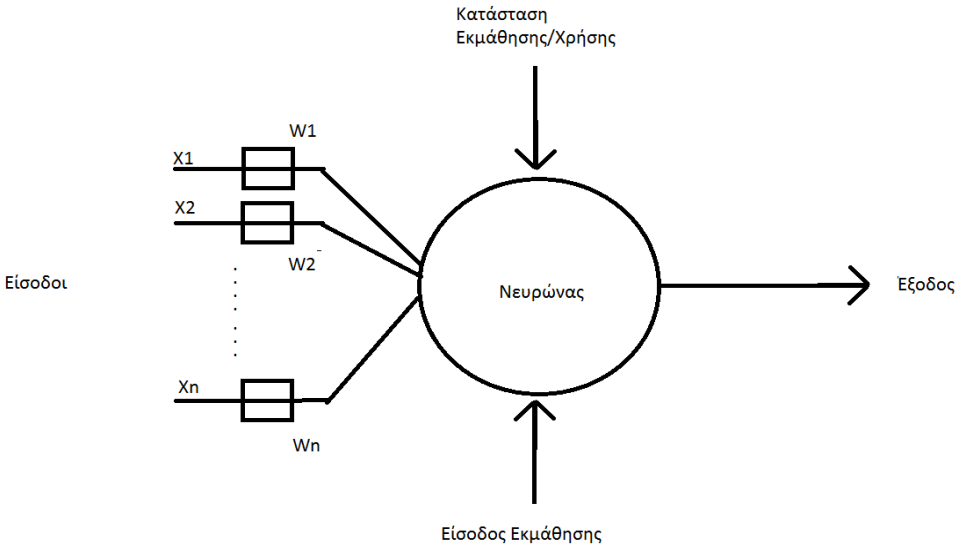
Τα νευρωνικά δίκτυα έχουν μεγάλη ικανότητα εξαγωγής νοήματος από πολύπλοκα και ανακριβή δεδομένα, και συνήθως χρησιμοποιούνται για την αναγνώριση μοτίβων και τάσεων που είναι πολύ πολύπλοκα για να παρατηρηθούν από τον άνθρωπο η άλλες υπολογιστικές τεχνικές. Ένα εκπαιδευμένο νευρωνικό δίκτυο μπορεί να θεωρηθεί ως "ειδικός" στην κατηγορία δεδομένων που του δίνεται να αναλύσει. Αυτός ο ειδικός μπορεί να εξάγει προβλέψεις με βάση νέες καταστάσεις που ενδιαφέρουν, και να απαντήσει ερωτήσεις τύπου "τι θα γίνει εάν..". Σημαντικό πλεονέκτημα είναι επίσης ότι οι πράξεις που απαιτούνται από ένα τεχνητό νευρωνικό δίκτυο μπορούν να εκτελούνται παράλληλα, και υπάρχει ήδη τεχνολογία που αξιοποιεί αυτή την ιδιότητα, και ότι τα τεχνητά νευρωνικά δίκτυα μπορούν να απεικονίζουν τα δεδομένα που δέχονται κατά τη διάρκεια της εκμάθησης.

Ένας απλός νευρώνας είναι ένα στοιχείο με πολλές εισόδους και μία έξοδο. Έχει δύο καταστάσεις λειτουργίας: την κατάσταση εκμάθησης και την κατάσταση χρήσης. Στην κατάσταση εκμάθησης, ο νευρώνας μπορεί να εκπαιδευτεί να πυροκροτεί η όχι, για συγκεκριμένα μοτίβα εισόδων. Στην κατάσταση χρήσης, όταν μια διδαγμένη μορφή εισόδων εντοπίζεται, η έξοδος (πυροκρότηση η όχι) που είχε εισαχθεί κατά την διαδικασία εκμάθησης γίνεται η έξοδος του συστήματος. Αν το μοτίβο εισόδων δεν ανήκει στην λίστα μοτίβων που έχουν διδαχθεί, ο κανόνας πυροκρότησης αποφασίζει την έξοδο του νευρώνα.



Ο κανόνας πυροκρότησης είναι πολύ σημαντικό στοιχείο στα νευρωνικά δίκτυα και είναι ο λόγος που είναι τόσο ευέλικτα. Ουσιαστικά αποφασίζει πώς υπολογίζεται το αν ένας νευρώνας θα πυροκροτήσει για οποιαδήποτε είσοδο, και αφορά όλες τις εισόδους, όχι μόνο αυτές στις οποίες εκπαιδεύεται ο νευρώνας.

Στην παρακάτω εικόνα φαίνεται ο νευρώνας των McCulloch και Pitts. Η διαφορά με το προηγούμενο μοντέλο είναι ότι οι είσοδοι έχουν βάρη, που επηρεάζουν την απόφαση πυροκρότησης. Οι είσοδοι αφού πολλαπλασιαστούν με τα αντίστοιχά τους βάρη, αθροίζονται. Αν το άθροισμά τους ξεπερνάει μια προκαθορισμένη τιμή, ο νευρώνας πυροκροτεί. Σε οποιαδήποτε άλλη περίπτωση ο νευρώνας δεν πυροκροτεί.

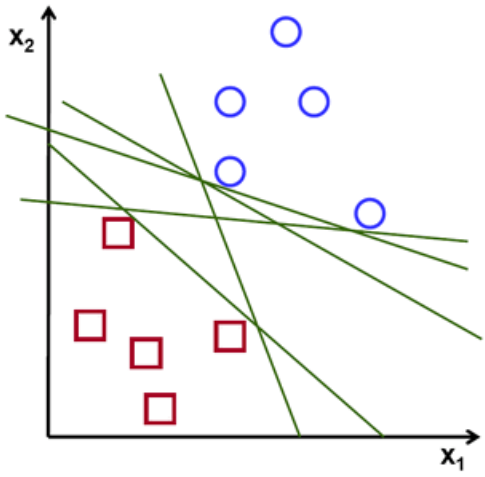


Η πρόσθεση των βαρών και της προκαθορισμένης τιμής κάνει τον νευρώνα πολύ ευέλικτο και ισχυρό. Ο νευρώνας μπορεί να προσαρμοστεί σε οποιαδήποτε κατάσταση αλλάζοντας τα βάρη και το όριό του.

Στην πράξη τα νευρωνικά δίκτυα χρησιμοποιούνται σε διαφόρους τομείς της επιχειρηματικής διοίκησης όπως οι προβλέψεις πωλήσεων, οι αναγνώριση ομάδων πελατών, η στοχευμένη προώθηση και η διαχείριση ρίσκου.

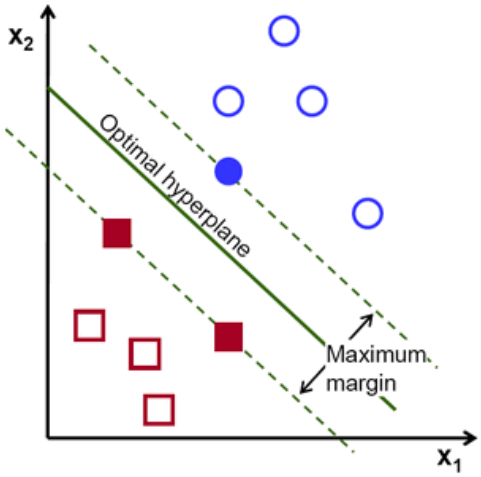
3.4.2.4 Μηχανές Διανύσματος Υποστήριξης (Support Vector Machines)

Οι μηχανές διανύσματος υποστήριξης είναι άλλη μια μέθοδος επιβλεπόμενης εκμάθησης, η οποία αποτελεί μια μέθοδο ταξινόμησης. Δοσμένων των δεδομένων εκμάθησης, ο αλγόριθμος εξάγει ένα βέλτιστο υπερπεδίο (hyperplane), που κατηγοριοποιεί τα καινούργια δεδομένα. Στο σχήμα που ακολουθεί, φαίνονται δύο ομάδες στοιχείων, και κάποιες γραμμές που τα διαχωρίζουν. Ενδιαφέρει όμως να βρεθεί η βέλτιστη διαχωριστική γραμμή.

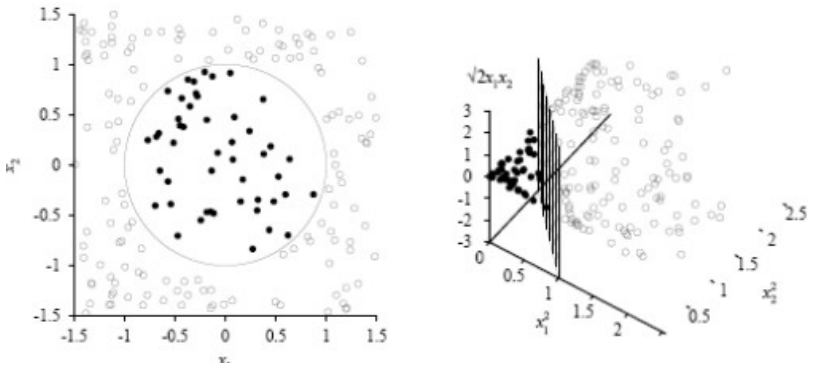


Για την εύρεση της βέλτιστης γραμμής κατηγοριοποίησης, υπάρχει ένα κριτήριο αξιολόγησής τους. Μια γραμμή είναι χαμηλής ποιότητας αν περνάει κοντά από τα σημεία, γιατί τότε θα είναι ευαίσθητη στον θόρυβο των δεδομένων, και δεν θα μπορεί να γενικεύσει σωστά στα καινούργια δεδομένα. Επομένως, ο στόχος είναι να βρεθεί η γραμμή που περνάει όσο πιο μακριά γίνεται από όλα τα σημεία. Άρα, η λειτουργία του αλγόριθμου είναι η εύρεση ενός υπερπεδίου, που δίνει την μέγιστη ελάχιστη απόσταση από τα στοιχεία των δεδομένων

εκμάθησης. Αυτή η απόσταση διπλασιασμένη, ονομάζεται περιθώριο, (margin). Επομένως το βέλτιστο υπερπεδίο μεγιστοποιεί το περιθώριο των δεδομένων εκμάθησης:



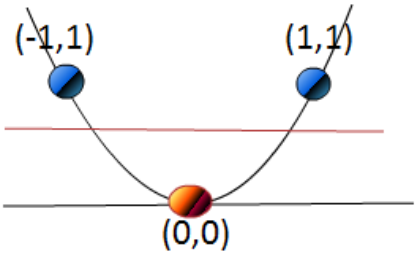
Αυτό το παράδειγμα αποτελεί απλοποίηση του προβλήματος, αφού τα στοιχεία του προβλήματος που καλείται να κατηγοριοποιήσει ο αλγόριθμος μπορούν να απεικονιστούν στο καρτεσιανό επίπεδο, ενώ στην πραγματικότητα οι διαστάσεις των δεδομένων μπορεί να είναι περισσότερες από δύο. Οι διαστάσεις όπως έχει προαναφερθεί μπορούν να αντιπροσωπεύουν τα χαρακτηριστικά ενός καταναλωτή που ενδιαφέρουν. Το παράδειγμα θα μπορούσε να έχει στους άξονες τα χαρακτηριστικά "ηλικία" και "εισόδημα". Στην πραγματική περίπτωση που ενδιαφέρουν περισσότερα χαρακτηριστικά, απαιτούνται περισσότερες διαστάσεις για να αναπαρασταθούν. Επιπλέον, τα δεδομένα του παραδείγματος είναι πολύ εύκολο να διαχωριστούν γραμμικά. Σε περιπτώσεις που τα δεδομένα δεν είναι γραμμικά διαχωρίσημα, οι μηχανές διανύσματος υποστήριξης χρησιμοποιούν μια συνάρτηση που λέγεται συνάρτηση kernel για να προβάλλουν τα δεδομένα σε έναν χώρο μεγαλύτερης διάστασης, όπου το διαχωριστικό υπερπεδίο είναι πιο εύκολο να βρεθεί:



Ένα παράδειγμα συνάρτησης kernel είναι το εξής:

Υποθέτουμε ότι δίνονται τα 3 σημεία του σχήματος στον μονοδιάστατο χώρο:

Ένα μονοδιάστατο υπερπεδίο θα ήταν μια κάθετη γραμμή που θα ξεχώριζε τα μπλε από τα κόκκινα δεδομένα. Προφανώς μια τέτοια γραμμή δεν υπάρχει. Αν προβάλουμε όμως τα δεδομένα σε δισδιάστατο χώρο, χρησιμοποιώντας μια συνάρτηση kernel x->(x,x^2), θα προκύψει η εξής απεικόνιση:



Εδώ μπορούμε να βρούμε ένα υπερπεδίο, δηλαδή μια γραμμή δύο διαστάσεων, που διαχωρίζει τα μπλε και τα κόκκινα στοιχεία, και πλέον τα δεδομένα μας μπορούν να διαχωριστούν με ένα SVM. Επομένως η γενική ιδέα είναι ότι γίνεται προσπάθεια προβολής των δεδομένων εκμάθησης σε χώρο με περισσότερες διαστάσεις, με την ελπίδα ότι θα βελτιωθεί η ικανότητα διαχωρισμού των δεδομένων.

3.4.3 Κανόνες συσχετισμού

Όπως αναφέρεται στις θεμελιώδεις εφαρμογές, πρόκειται για κανόνες που εξάγονται με βάση τα προϊόντα που αγοράζονται μαζί. Εκφράζοντας το πρόβλημα με μαθηματικούς όρους, τα δεδομένα αποτελούν μια μεγάλη βάση δεδομένων T, που είναι ένα σύνολο από συναλλαγές, δηλαδή T={t1,t2,t3...tn}.

Κάθε συναλλαγή περιέχει ένα σύνολο από αντικείμενα I={i1,i2,i3...im} όπου τα αντικείμενα μπορούν να είναι ψωμί, δημητριακά κτλ. Επομένως επιδιώκουμε να ανακαλύψουμε ενδιαφέροντα μοτίβα, συσχετισμούς ή σχέσεις αιτιότητας μέσα στα σύνολα των αντικειμένων. Αυτοί οι συσχετισμοί εκφράζονται με την μορφή ενδεχομένων, δηλαδή X=>Y μπορεί να εκφράζει το ενδεχόμενο ενός καταναλωτή να αγοράσει γάλα αν αγοράσει δημητριακά. Ας δούμε για παράδειγμα τον παρακάτω πίνακα:

Συναλλαγή T	**Αντικείμενα**
t1	Ψωμί, μαρμελάδα, βούτυρο
t2	Ψωμί, βούτυρο
t3	Ψωμί, γάλα, βούτυρο
t4	Μπύρα, Ψωμί
t5	Μπύρα, Γάλα

Ένα παράδειγμα ενδεχομένου είναι το ψωμί=>βούτυρο, ή το μπύρα=>ψωμί.

Τα μεγέθη που χρησιμοποιούνται για την αξιολόγηση κάθε ενδεχομένου είναι η υποστήριξη (support) και η εμπιστοσύνη (confidence). Ο στόχος ενός κανόνα συσχετισμού είναι να εντοπίσει τα ενδεχόμενα που έχουν τιμές για αυτές τις συναρτήσεις που υπερβαίνουν ένα όριο που θέτει ο αναλυτής. Η υποστήριξη εκφράζει την συχνότητα που το ενδεχόμενο συμβαίνει, και ορίζεται ως:

$$Support = \frac{frq(X,Y)}{N}$$

δηλαδή το πλήθος των συναλλαγών που εμπεριέχουν και τα δύο αντικείμενα, προς το πλήθος όλων των συναλλαγών. Εδώ το ενδεχόμενο {ψωμί, βούτυρο} έχει υποστήριξη 3/5, ενώ το {μπύρα, ψωμί} έχει 1/5 . Η εμπιστοσύνη εκφράζει την ισχύ του συσχετισμού, και ορίζεται ως:

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

δηλαδή το πλήθος των συναλλαγών που εμπεριέχουν και τα δύο αντικείμενα, προς το πλήθος των συναλλαγών που περιέχουν μόνο το ένα. Για την αξιολόγηση του κανόνα χρησιμοποιείται μια μορφή lift, που υπολογίζεται από την υποστήριξη του ενδεχομένου, προς το γινόμενο των υποστηρίξεων των X και Y αν ήταν ανεξάρτητα:

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Αλγόριθμος AIS: Τα υποψήφια σύνολα αντικειμένων δημιουργούνται και υπολογίζονται παράλληλα καθώς σαρώνονται τα δεδομένα. Για κάθε συναλλαγή που σαρώνεται, αποφασίζεται ποια από τα μεγάλα σύνολα αντικειμένων της προηγούμενης επανάληψης σάρωσης εμπεριέχονται στην συναλλαγή. Δημιουργούνται νέα υποψήφια σύνολα αντικειμένων επεκτείνοντας τα μεγάλα αυτά σύνολα με άλλα αντικείμενα που υπάρχουν στην συναλλαγή τους. Στο παράδειγμα που ακολουθεί έχει θεωρηθεί ελάχιστη υποστήριξη 2.

Συναλλαγή	Αντικείμενα
1	1 3 4
2	2 3 5
3	1 2 3 5
4	2 5

Πρώτο στάδιο:

Σύνολο	Υποστήριξη
{1}	2
{2}	3
{3}	3
{5}	3

Δεύτερο στάδιο:

Σύνολο	Υποστήριξη
{1 3}	2
{1 4}	1
{3 4}	1
{2 3}	2
{2 5}	3
{3 5}	2
{1 2}	1
{1 5}	1

Τρίτο στάδιο:

Σύνολο	Υποστήριξη
{1 3 4}	1
{2 3 5}	2
{1 3 5}	1

Το μειονέκτημα του αλγορίθμου AIS είναι ότι δημιουργεί και υπολογίζει υπερβολικά πολλά υποψήφια σύνολα αντικειμένων που τελικά είναι πολύ μικρά.

Αλγόριθμος SETM: Τα υποψήφια σύνολα αντικειμένων δημιουργούνται παράλληλα με την σάρωση δεδομένων, αλλά υπολογίζονται στο τέλος του κάθε περάσματος. Τα νέα υποψήφια σύνολα αντικειμένων υπολογίζονται όπως στον αλγόριθμο AIS , αλλά ο αριθμός της συναλλαγής από την οποία προκύπτουν αποθηκεύεται μαζί με τα σύνολα σε σειριακά. Στο τέλος του περάσματος, η στήριξη κάθε συνόλου υπολογίζεται συναθροίζοντας την σειριακή δομή. Ο αλγόριθμος SETM έχει το ίδιο μειονέκτημα με τον αλγόριθμο AIS. Επιπλέον, για κάθε υποψήφιο σύνολο αντικειμένων, υπάρχουν όσες είσοδοι όσες και η υποστήριξή τους.

Αλγόριθμος Apriori: Τα υποψήφια σύνολα αντικειμένων δημιουργούνται μόνο με βάση τα μεγάλα σύνολα του προηγούμενου περάσματος, χωρίς να λαμβάνονται υπ' όψη οι συναλλαγές. Τα μεγάλα σύνολα του προηγούμενου περάσματος ενώνονται μεταξύ τους για να σχηματίσουν μεγαλύτερα σύνολα. Κάθε σύνολο του οποίου ένα υποσύνολο δεν έχει την απαιτούμενη στήριξη διαγράφεται:

Συναλλαγή	Αντικείμενα
1	1 3 4
2	2 3 5
3	1 2 3 5
4	2 5

Πρώτο στάδιο:

Σύνολο	Υποστήριξη
{1}	2
{2}	3
{3}	3
{5}	3

Δεύτερο στάδιο:

Σύνολο	Υποστήριξη
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

Τρίτο στάδιο:

Σύνολο	Υποστήριξη
{2 3 5}	2

Ο αλγόριθμος Apriori εκμεταλλεύεται το γεγονός ότι κάθε υποσύνολο ενός συνόλου που θεωρείται συχνό (στήριξη>=2) είναι και αυτό συχνό. Επομένως ο αλγόριθμος μπορεί να μειώσει τον αριθμό των υποψηφίων συνόλων εξετάζοντας μόνο τα υποσύνολα των οποίων η υποστήριξη είναι μεγαλύτερη από την απαιτούμενη. Χρησιμοποιείται στην έρευνα [48], όπου

συνδυάζονται κανόνες συσχετισμού με δεδομένα ικανοποίησης άυλων αναγκών στη σχεδίαση, και αναλύεται στο κεφάλαιο 4.3

3.4.4 Αλγόριθμοι Εντοπισμού Κοινοτήτων

Η μέθοδος αυτή αποσκοπεί στον εντοπισμό κοινοτήτων καταναλωτών ή ιστοσελίδων, όπως προκύπτουν από την μελέτη του αντίστοιχου γράφου του κοινωνικού δικτύου. Αποτελεί άλλη μια θεμελιώδη εφαρμογή. Πρακτικά, κοινότητα σε ένα δίκτυο είναι μια ομάδα από κόμβους που έχουν μεγαλύτερους εσωτερικούς δεσμούς από ότι με το υπόλοιπο δίκτυο. Αυτός ο ανεπίσημος ορισμός έχει επισημοποιηθεί με διάφορους ανταγωνιστικούς τρόπους, συνήθως χρησιμοποιώντας συναρτησιακές σχέσεις, που ποσοτικοποιούν την "ποιότητα" μιας ομάδας κόμβων σαν κοινότητα. Κάποιες από αυτές τις συναρτήσεις, όπως η σπονδυλότητα (Modularity) και τα Normalized Cuts είναι πιο δημοφιλείς από τις υπόλοιπες, αλλά καμία δεν είναι κοινώς αποδεκτή αφού δεν υπάρχει συνάρτηση που να είναι αποδεκτή σε όλες τις εφαρμογές.

Οι αλγόριθμοι του εντοπισμού κοινοτήτων διαφέρουν σε διάφορες σημαντικές διαστάσεις, όπως η προσέγγιση του προβλήματος και τα χαρακτηριστικά της απόδοσής τους. Μια σημαντική διάσταση είναι το αν βασίζονται συγκεκριμένα σε μία ποιοτική τιμή η όχι. Οι φασματικές μέθοδοι, ο αλγόριθμος του Kerninghan-Lin είναι παραδείγματα, ενώ άλλοι αλγόριθμοι όπως ο Markov Clustering (MCL), λειτουργούν διαφορετικά. Οι φασματικοί αλγόριθμοι είναι αλγόριθμοι που εκτελούν μαθηματικές πράξεις στα ιδιοδιανύσματα των πινάκων που προκύπτουν από την δομή του γράφου του κοινωνικού δικτύου, όπως ο πίνακας γειτνίασης που περιγράφει το πως συνδέονται οι κόμβοι μεταξύ τους. Ο αλγόριθμος Kerninghan-Lin είναι ένας επαναληπτικός αλγόριθμος που επιδιώκει την ελαχιστοποίηση της συνάρτησης Kerninghan-Lin που περιγράφεται στις συναρτήσεις. Ο αλγόριθμος Markov Clustering τμηματοποιεί τους γράφους χρησιμοποιώντας στοχαστικούς πίνακες, δηλαδή τους πίνακες που περιγράφουν την πιθανότητα μετάβασης μεταξύ των κόμβων.

Μια άλλη διάσταση είναι κατά πόσο επιτρέπουν στον χρήστη να ελέγξει τον βαθμό ανάλυσης του δικτύου σε κοινότητες. Μερικοί αλγόριθμοι (όπως οι φασματικοί) είναι σχεδιασμένοι ώστε να χωρίζουν το δίκτυο σε δυο ομάδες, αλλά μπορούν να χρησιμοποιηθούν

επαναληπτικά ώστε να δημιουργηθούν όσες κοινότητες είναι επιθυμητό. Άλλοι αλγόριθμοι όπως οι συσσωρευτικοί (agglomerative clustering) ή ο MCL επιτρέπουν στον χρήστη να ελέγξει το επίπεδο ανάλυσης έμμεσα, χρησιμοποιώντας συγκεκριμένες παραμέτρους, ενώ υπάρχουν και αλγόριθμοι όπως αυτοί που αξιοποιούν την σπονδυλότητα που δεν επιτρέπουν στον χρήστη να ελέγχει το αποτέλεσμα. Οι συσσωρευτικοί αλγόριθμοι ξεκινούν με κάθε κόμβο στο κοινωνικό δίκτυο να αντιπροσωπεύει την δικιά του κοινότητα, και σε κάθε βήμα συγχωνεύουν κοινότητες που μπορούν να θεωρηθούν επαρκώς όμοιες, συνεχίζοντας μέχρι είτε να προκύψει ο επιθυμητός αριθμός κοινοτήτων, είτε οι κοινότητες που έχουν απομείνει να είναι πολύ διαφορετικές ώστε να ομαδοποιηθούν. Αντίθετα υπάρχουν και οι διαιρετικοί αλγόριθμοι λειτουργούν αντίστροφα. Ξεκινούν με το συνολικό δίκτυο σαν μία κοινότητα, και σε κάθε βήμα επιλέγουν μια συγκεκριμένη κοινότητα και την χωρίζουν σε δύο μέρη.

Ένα τρίτο χαρακτηριστικό που διαφοροποιεί τους αλγόριθμους είναι το κατά πόσο δίνουν βαρύτητα στην ισορροπημένη διαίρεση του δικτύου. Ενώ συναρτήσεις όπως η KL εφαρμόζουν αποκλειστικά ισορροπημένη διαίρεση, άλλες λαμβάνουν υπ' όψη την ισορροπία μόνο έμμεσα ή και καθόλου. Όσο αφορά τα χαρακτηριστικά απόδοσης, οι αλγόριθμοι διαφέρουν στην δυνατότητα κλιμακωσιμότητας, δηλαδή στην ικανότητα τους να αποδίδουν όταν πρόκειται για πολύ μεγάλα δίκτυα, με τους αλγόριθμους πολλαπλών επιπέδων όπως ο Metis , που είναι μια προσαρμογή του αλγόριθμου Kerninghan-Lin σε περιβάλλον πολλαπλών επιπέδων ο MLR-MCL που είναι η αντίστοιχη προσαρμογή του αλγόριθμου MCL, και ο Graclus να προσαρμόζονται καλύτερα από πολλές άλλες προσεγγίσεις του προβλήματος. Η τμηματοποίηση γράφων σε πολλαπλά επίπεδα συμπιέζει τον αρχικό γράφο, τον τμηματοποιεί, και τελικά τον αποσυμπιέζει σταδιακά μέχρι να επιστρέψει στον αρχικό γράφο. Τα κύρια βήματα μιας τμηματοποίησης πολλών επιπέδων είναι:

1.Συμπίεση. Ο στόχος εδώ είναι να εξάγουμε έναν μικρότερο γράφο που είναι παρόμοιος με τον αρχικό. Αυτό το βήμα μπορεί να επαναληφθεί ώστε να φτάσουμε σε ένα γράφο που είναι αρκετά μικρός για να τμηματοποιηθεί γρήγορα και ποιοτικά. Μια δημοφιλής τεχνική συρρίκνωσης είναι η κατασκευή μιας “αντιστοίχισης” στον γράφο, όπου αντιστοίχιση ορίζεται ως μια ομάδα ακμών καμία από τις οποίες δεν προσπίπτουν στον ίδιο κόμβο. Για κάθε ακμή στην αντιστοίχιση, οι κόμβοι στις άκρες τις συγχωνεύονται και αντιπροσωπεύονται από έναν κόμβο στον συμπιεσμένο γράφο.

2.Αρχική τμηματοποίηση. Σε αυτό το βήμα, μια τμηματοποίηση του συμπιεσμένου γράφου εκτελείται. Καθώς σε αυτό το στάδιο ο γράφος είναι αρκετά μικρός, μπορούν να χρησιμοποιηθούν τεχνικές όπως η φασματική τμηματοποίηση, που είναι αργές αλλά εξάγουν τμηματοποιήσεις υψηλής ποιότητας.

3.Αποσυμπίεση. Σε αυτή την φάση, η τμηματοποίηση του τρέχοντος γράφου χρησιμοποιείται για να αρχικοποιήσει μια τμηματοποίηση στον μεγαλύτερο γράφο. Η καλύτερη δομική διασύνδεση του γράφου που προκύπτει από την αποσυμπίεση χρησιμοποιείται για να τελειοποιήσει την τμηματοποίηση. Αυτό το βήμα επαναλαμβάνεται μέχρι να φτάσουμε στον αρχικό γράφο.

Μια πληθώρα συναρτήσεων ή μέτρων χρησιμοποιούνται για να αξιολογήσουν την ικανότητα μιας ομάδας κόμβων να συμπεριφέρεται σαν κοινότητα. Για τα ακόλουθα χρησιμοποιούνται οι χαρακτηρισμοί A(i,j) που εκφράζει το βάρος της ακμής μεταξύ των κόμβων ι και j, S που εκφράζει την ομάδα κόμβων που ελέγχεται, και V που συμβολίζει τον συνολικό γράφο που αντιπροσωπεύει την κοινότητα.

Η συνάρτηση Normalised Cuts ορίζεται ως:

$$Ncut(S) = \frac{\sum_{i \in S, j \in \bar{S}} A(i,j)}{\sum_{i \in S} degree(i)} + \frac{\sum_{i \in S, j \in \bar{S}} A(i,j)}{\sum_{j \in \bar{S}} degree(j)}$$

δηλαδή, το άθροισμα από τα βάρη των ακμών που συνδέουν την ομάδα S με τον υπόλοιπο γράφο, κανονικοποιημένα ως προς το συνολικό βάρος των ακμών μεταξύ των κόμβων του S, και του υπόλοιπου γράφου, δηλαδή του συμπληρωματικού του S. Οι ομάδες με χαμηλή τιμή Ncut αποτελούν ποιοτικές κοινότητες, καθώς είναι καλά συνδεδεμένες εσωτερικά, αλλά όχι με τον υπόλοιπο γράφο.

Η **αγωγιμότητα** (conductance) μιας ομάδας S είναι σχετικής σημασίας και ορίζεται ως:

$$Conductance(S) = \frac{\sum_{i \in S, j \in \bar{S}} A(i,j)}{\min(\sum_{i \in S} degree(i), \sum_{i \in \bar{S}} degree(i))}$$

Η συνάρτηση **Kerninghan-Lin** επιδιώκει να ελαχιστοποιήσει την αντίστοιχη ομάδα διαχωρισμού μιας ομάδας, υπό τον περιορισμό ότι όλες οι ομάδες είναι του ίδιου μεγέθους, υποθέτοντας ότι το συνολικό μέγεθος του V είναι πολλαπλάσιο του αριθμού των ομάδων:

$$KLObj(V_1, \ldots, V_k) = \sum_{i \neq j} A(V_i, V_j) \text{subject to} |V_1| = |V_2| = \ldots = |V_k|$$

εδώ, το A(Vi,Vj) συμβολίζει το άθροισμα από τα βάρη των ακμών μεταξύ των κόμβων στην ομάδα Vi και την Vj.

Η **σπονδυλότητα** (modularity), έχει γίνει πολύ δημοφιλής ως μέτρο αξιοποίησης της ποιότητας κατηγοριοποίησης ενός γράφου. Ένα πλεονέκτημά της είναι ότι είναι ανεξάρτητη από τον αριθμό των ομάδων στον οποίον έχει χωριστεί ο γράφος. Η ιδέα πίσω από την σπονδυλότητα είναι ότι όσο πιο μακριά είναι ο υπογράφος που αντιπροσωπεύει μια κοινότητα από έναν τυχαίο υπογράφο, τόσο καλύτερη ή σημαντικότερη είναι η κοινότητα που εντοπίστηκε. Η σπονδυλότητα Q για μια διαίρεση του γράφου σε k ομάδες (V1....Vk) ορίζεται ως:

$$Q = \sum_{c=1}^{k} \left[\frac{A(V_i, V_i)}{m} - \left(\frac{degree(V_i)}{2m} \right)^2 \right]$$

όπου τα Vi είναι οι ομάδες, m είναι ο αριθμός των αιχμών στον γράφο, και degree(Vi) είναι ο συνολικός βαθμός της ομάδας Vi. Για κάθε ομάδα, υπολογίζουμε την διαφορά μεταξύ του κλάσματος των ακμών μέσα στην ομάδα, και το κλάσμα των ακμών που θα ήταν αναμενόμενο να υπάρχουν μέσα σε μια τυχαία ομάδα με τον ίδιο συνολικό βαθμό.

3.4.5 Ανάλυση Επιρροής

Σε αυτό το κεφάλαιο αναλύεται το υπολογιστικό κομμάτι της ανάλυσης επιρροής όπως προκύπτει από την ανάλυση του γράφου του κοινωνικού δικτύου, και περιγράφονται οι μετρήσεις που σχετίζονται με αυτήν. Συγκεκριμένα αναλύεται η ποσοτική και ποιοτική αξιολόγηση του επιπέδου επιρροής κόμβων και ακμών σε ένα δίκτυο. Εδώ πρέπει να σημειωθεί ότι η ανάλυση επιρροής ενός χρήστη η ιστοσελίδας μπορεί να προκύψει και από τις πιο άμεσες μετρήσεις των μέσων κοινωνικής δικτύωσης που αναφέρονται στο κεφάλαιο

“Συλλογή Δεδομένων”, αλλά όπως στις περισσότερες εφαρμογές των Predictive Analytics, τα αποτελέσματα από τις δύο μεθόδους μπορούν να συνδυαστούν ώστε να προκύψει ένα τελικό μέγεθος που λαμβάνει υπ' όψη και τα στοιχεία που προκύπτουν από τις δημοσιεύσεις του χρήστη ή ιστοσελίδας, αλλά και την τοπολογία του γράφου, με την επιθυμητή βαρύτητα.

3.4.5.1 Στατιστικά Επιρροής

Όπως έχουμε ξαναδεί, ένα κοινωνικό δίκτυο μοντελοποιείται σαν ένας γράφος G=(V,E) όπου V είναι το σύνολο των κόμβων και E είναι το σύνολο των ακμών. Οι κόμβοι αντιπροσωπεύουν τα άτομα η τις ιστοσελίδες (οντότητες) και οι ακμές τις κοινωνικές σχέσεις. Σε τοπικό επίπεδο, η κοινωνική επιρροή είναι ένα κατευθυνόμενο φαινόμενο από τον κόμβο A στον κόμβο B, και σχετίζεται με το βάρος της ακμής από τον A στον B. Σε καθολικό επίπεδο, μερικοί κόμβοι μπορεί να έχουν πολύ μεγαλύτερη επιρροή από τους υπόλοιπους λόγω της δομής του δικτύου. Αυτές οι καθολικές μετρήσεις συχνότερα συσχετίζονται με τους κόμβους του δικτύου παρά με τις ακμές.

Μετρήσεις Ακμών

Δύναμη Δεσμού: Σύμφωνα με την έρευνα του M.Granovetter [11], η δύναμη του δεσμού μεταξύ δύο κόμβων εξαρτάται από την επικάλυψη των γειτονιών τους. Συγκεκριμένα, όσο περισσότερους κοινούς γείτονες έχει ένα ζευγάρι κόμβων A και B, τόσο μεγαλύτερος είναι ο δεσμός μεταξύ τους. Αν η επικάλυψη των γειτονιών είναι μεγάλη, θεωρούμε ότι οι δύο κόμβοι A και B έχουν δυνατό δεσμό. Διαφορετικά, θεωρούμε ότι έχουν αδύναμο δεσμό. Η δύναμη του δεσμού S(A,B) δίνεται από την σχέση:

$$S(A,B) = \frac{|n_A \cap n_B|}{|n_A \cup n_B|}$$

Εδώ, τα nA και nB είναι οι γειτονιές των κόμβων A και B αντίστοιχα. Μερικές φορές η δύναμη ορίζεται με ένα διαφορετικό όνομα, το embededdness. Το embededdness μιας ακμής είναι υψηλό όταν οι κόμβοι στις άκρες της έχουν υψηλή επικάλυψη γειτονιών. Στην πράξη αυτό μεταφράζεται ως εξής: Όταν δύο άτομα είναι συνδεδεμένα από μια “ενσωματωμένη¨(embedded) ακμή, είναι πιο εύκολο για αυτά να εμπιστευθεί το ένα τον άλλο, καθώς είναι ευκολότερο να ανακαλυφτεί η μη ηθική συμπεριφορά μεταξύ κοινών φίλων[11].Αντίθετα, όταν το embeddeness είναι μηδενικό, οι δύο τελικοί κόμβοι δεν έχουν κοινούς φίλους. Επομένως, είναι πιο δύσκολο να εμπιστευθεί ο ένας τον άλλον, αφού δεν υπάρχει κάποιος κοινός φίλος να επιβεβαιώσει την συμπεριφορά τους.

Ένα πόρισμα από την δύναμη δεσμού είναι η υπόθεση του τριαδικού κλεισίματος (triadic closure).Αυτό αναφέρεται στην φύση των δεσμών ανάμεσα σε τρεις οντότητες Α, Β και Γ. Αν υπάρχουν δυνατοί δεσμοί που συνδέουν τον Α με τον Β, και τον Α με τον Γ, τότε είναι πιθανό ο Β με τον Γ να συνδέονται επίσης με έναν δυνατό δεσμό. Αντίστροφα, αν οι Α-Β και Α-Γ είναι αδύναμοι δεσμοί, ο Β και ο Γ έχουν λιγότερες πιθανότητες να έχουν δυνατό δεσμό. Το τριαδικό κλείσιμο μετριέται από τον συντελεστή τμηματοποίησης του δικτύου[12].Ο συντελεστής τμηματοποίησης ενός κόμβου Α ορίζεται ως η πιθανότητα δύο τυχαία επιλεγμένοι φίλοι του να είναι φίλοι μεταξύ τους. Με άλλα λόγια, είναι το τμήμα των ζευγαριών φίλων του Α που συνδέονται μεταξύ τους. Αυτό σχετίζεται άμεσα με την καταμέτρηση τριγώνων σε ένα δίκτυο. Αν το nΔ είναι ο αριθμός των τριγώνων στο δίκτυο και το |E| είναι ο αριθμός των ακμών στο δίκτυο, τότε ο συντελεστής τμηματοποίησης ορίζεται ως:

$$C = \frac{6n_\Delta}{|E|}$$

Ο αφελής τρόπος απλής καταμέτρησης του αριθμού τριγώνων nΔ έχει μεγάλο υπολογιστικό κόστος. Μια ενδιαφέρουσα σχέση ανάμεσα στον αριθμό τριγώνων και στις ιδιοτιμές του δικτύου διαπιστώθηκε από τον C.E.Tsourakakis[13]. Η μελέτη του δείχνει ότι ο αριθμός τριγώνων είναι προσεγγιστικά ίσος με το άθροισμα των ιδιοτιμών του πίνακα γειτνίασης του δικτύου αυξημένες στην τρίτη δύναμη. Δεδομένης της κατανομής των ιδιοτιμών, για τον υπολογισμό του αριθμού τριγώνων απαιτείται ο υπολογισμός μόνο λίγων πρώτων ιδιοτιμών. Αυτή είναι μια αποδοτική μέθοδος, αφού οι πρώτες ιδιοτιμές μπορούν να υπολογιστούν πιο εύκολα από τον πολύ κουραστικό υπολογισμό του συνόλου τους.

Αδύναμοι δεσμοί :Όταν η επικάλυψη των γειτονιών των κόμβων Α και Β είναι μικρή, η σύνδεση Α-Β θεωρείται αδύναμη. Όταν δεν υπάρχει καθόλου επικάλυψη, η σύνδεση Α-Β είναι μια τοπική γέφυρα (local bridge) [11].Στην ακραία περίπτωση, η αφαίρεση της ακμής Α-Β μπορεί να έχει αποτέλεσμα την αποσύνδεση του συνδεδεμένου τμήματος που εμπεριέχει τον κόμβο Α και Β. Σε αυτή την περίπτωση, η ακμή Α-Β μπορεί να θεωρηθεί καθολική γέφυρα(global bridge). Πρακτικά, γέφυρα είναι μια οντότητα που είναι η μόνη σύνδεση ανάμεσα σε δύο κοινότητες. Στα αληθινά δίκτυα, οι καθολικές γέφυρες συναντώνται πιο σπάνια από τις τοπικές. Παρόλα αυτά η επίδραση των τοπικών και καθολικών γεφυρών είναι πολύ όμοιες.

Betweenness Ακμής:Μια επίσης σημαντική μέτρηση είναι το Betweenness ακμής, που μετράει την συνολική ποσότητα ροής στην ακμή. Εδώ, υποθέτουμε ότι η ροή πληροφορίας ανάμεσα στον Α και Β είναι ομοιόμορφα διαμοιρασμένη στο κοντινότερο μονοπάτι μεταξύ του Α και Β. Όπως είδαμε και στο προηγούμενο κεφάλαιο, το betweenness έχει μεγάλη εφαρμογή στην τμηματοποίηση γράφων.

Μετρήσεις Κόμβων

Κεντρικότητα (Centrality): Η κεντρικότητα ενός κόμβου ορίζεται ώστε να μετρηθεί η σημασία ενός κόμβου στο δίκτυο. Έχει μεγάλο ενδιαφέρον ως εργαλείο για την μελέτη κοινωνικών δικτύων [14].Ένας κόμβος με μεγάλη κεντρικότητα συνήθως θεωρείται ότι έχει μεγαλύτερη επιρροή από τους υπόλοιπους στο δίκτυο. Διάφοροι τρόποι μέτρησης της κεντρικότητας έχουν προταθεί βασισμένοι στον ακριβής ορισμό της επιρροής. Η κύρια αρχή στην κατηγοριοποίηση των μετρήσεων κεντρικότητας είναι ο τύπος υπολογισμού τυχαίου μονοπατιού που χρησιμοποιείται. Συγκεκριμένα, οι μετρήσεις κεντρικότητας μπορούν να κατηγοριοποιηθούν σε δύο κατηγορίες: ακτινικές (radial) και μεσαίες (medial).Οι ακτινικές μετρήσεις αφορούν τυχαία μονοπάτια που ξεκινάνε η τελειώνουν σε έναν δοσμένο κόμβο. Από την άλλη, οι μεσαίες μετρήσεις αφορούν τυχαία μονοπάτια που περνάνε από έναν δοσμένο κόμβο. Οι ακτινικές μετρήσεις κατηγοριοποιούνται σε μετρήσεις όγκου και μετρήσεις μήκους, ανάλογα τον τύπο του τυχαίου μονοπατιού. Οι μετρήσεις όγκου καθορίζουν το μήκος του μονοπατιού και βρίσκουν τον όγκο μονοπατιού που επιτρέπει το μήκος. Οι μετρήσεις μήκους καθορίζουν τον όγκο των κόμβων, και βρίσκουν το μήκος μονοπατιού που χρειάζεται για να επιτευχθεί αυτός ο όγκος.

Βαθμός (Degree):Η πρώτη ομάδα μετρήσεων κεντρικότητας είναι ακτινική και κατηγορίας όγκου. Η πιο απλή και διαδεδομένη μέτρηση σε αυτή την κατηγορία είναι η κεντρικότητα βαθμού. Η κεντρικότητα βαθμού ενός κόμβου είναι ίση με τον βαθμό του κόμβου. Εκφράζει δηλαδή με πόσους άλλους χρήστες ή ιστοσελίδες είναι άμεσα συνδεδεμένη μια οντότητα. Ένας τρόπος επεξήγησης της κεντρικότητας βαθμού είναι ότι απαριθμεί τα μονοπάτια μήκους 1 που ξεκινάνε από έναν κόμβο. Μια φυσική γενίκευση από αυτή την άποψη είναι η κεντρικότητα Κ-μονοπατιού, που είναι ο αριθμός των μονοπατιών μήκους το πολύ Κ που ξεκινάνε από έναν κόμβο.

Εγγύτητα (Closeness):Η δεύτερη ομάδα μετρήσεων κεντρικότητας είναι κατηγορίας μήκους και ακτινικής. Αντίθετα από τις μετρήσεις όγκου, οι μετρήσεις μήκους μετράνε το μήκος των μονοπατιών. Η πιο διαδεδομένη μέτρηση σε αυτή την κατηγορία είναι η κεντρικότητα

εγγύτητας του Freeman. Μετράει την κεντρικότητα υπολογίζοντας τον μέσο όρο των κοντινότερων αποστάσεων προς όλους τους υπόλοιπους κόμβους.

Betweenness Κόμβου: Αντίστοιχα με της ακμές με υψηλό Betweenness, οι κόμβοι με την ίδια ιδιότητα έχουν κρίσιμες θέσεις στην δομή του δικτύου, και επομένως έχουν πολύ σημαντικό ρόλο. Αυτό συμβαίνει λόγω της μεγάλης ποσότητας ροής που φέρουν οι κόμβοι που έχουν θέση σε στενές συνδεδεμένες ομάδες. Αυτοί οι κόμβοι έχουν υψηλό betweenness.

Δομικές Τρύπες (Structural Holes):Σε ένα δίκτυο, ένας κόμβος ονομάζεται δομική τρύπα αν είναι συνδεδεμένος με πολλές τοπικές γέφυρες. Συχνά, η επιτυχία μιας οντότητας σε ένα κοινωνικό δίκτυο βασίζεται στην πρόσβασή της σε τοπικές γέφυρες[15].Αφαιρώντας αυτή την οντότητα, ένας κενός χώρος θα προκύψει στο δίκτυο. Αυτό ονομάζεται δομική τρύπα. Η οντότητα που έχει τον ρόλο της δομικής τρύπας μπορεί να διασυνδέσει πληροφορίες με προέλευση από ομάδες που δεν συνδέονται μεταξύ τους. Επομένως, αυτή η οντότητα είναι δομικά σημαντική για την συνδεσιμότητα διαφορετικών περιοχών του δικτύου. Ένα άλλο ενδιαφέρον στοιχείο είναι ότι τα ενδιαφέροντα της οντότητας δομικής τρύπας, και της επιχείρησης μπορεί να μην συμπίπτουν. Για μια επιχείρηση, η επιτάχυνση της ροής πληροφορίας ανάμεσα στις ομάδες μπορεί να έχει πλεονεκτήματα, που απαιτεί την κατασκευή γεφυρών. Ωστόσο, αυτή η κατασκευή γεφυρών υποθάλπτει την θεωρητική δύναμη που έχει η οντότητα-τρύπα στην ρύθμιση της ροής πληροφορίας. Για παράδειγμα, μια ιστοσελίδα forum που αποτελεί αποκλειστικό σύνδεσμο σε ξεχωριστές καταναλωτικές κοινότητες, αποτελεί μια δομική τρύπα, που χάνει την αξία της όταν δημιουργούνται κόμβοι γέφυρες ανάμεσα σε αυτές τις κοινότητες.

3.4.6 Πρόβλεψη Συνδέσμων

Με βάση τον ορισμό των Liben-Nowell, David, και Kleinberg, Jon [19], το πρόβλημα της πρόβλεψης συνδέσμων μπορεί να οριστεί ως εξής: Δίνεται ένα κοινωνικό δίκτυο G(V,E) στο οποίο οι ακμές e=(u,v) αντιπροσωπεύουν ένα είδος αλληλεπίδρασης ανάμεσα στους δύο κόμβους της την χρονική στιγμή t(e). Μπορούμε να καταγράψουμε πολλαπλές αλληλεπιδράσεις από παράλληλες ακμές ή χρησιμοποιώντας μια περίπλοκη χρονική σφραγίδα (timestamp) για μια ακμή. Για τον χρόνο t<t' θεωρούμε ότι ο γράφος G[t,t'] είναι ένας υπογράφος του G, περιορισμένος στις ακμές μεταξύ των χρονικών τιμών t και t'. Σε ένα περιβάλλον επιβλεπόμενης εκμάθησης για την πρόβλεψη συνδέσμων, μπορούμε να διαλέξουμε ένα διάστημα εκμάθησης [t0,t0'], και ένα διάστημα δοκιμής [t1,t1'] όπου t0'<t1. Επομένως πλέον το πρόβλημα έχει

μεταφραστεί στην εξαγωγή μιας λίστας από ακμές που δεν υπάρχουν στον γράφο G[t0,t0'] αλλά προβλέπουμε ότι θα εμφανιστούν στον γράφο G[t1,t1'].

Οι [19], πρότειναν ένα από τα πρώτα μοντέλα πρόβλεψης συνδέσμων που λειτουργεί αποκλειστικά σε κοινωνικά δίκτυα. Κάθε κόμβος αντιπροσωπεύει ένα άτομο και μια ακμή ανάμεσα σε δύο άτομα αντιπροσωπεύει την αλληλεπίδραση μεταξύ τους. Η πολλαπλότητα των αλληλεπιδράσεων μπορεί να μοντελοποιηθεί αποκλειστικά επιτρέποντας παράλληλες ακμές, η με ένα ανάλογο σύστημα από βάρη για τις ακμές. Το παράδειγμα εκμάθησης σε αυτή τη διαδικασία συνήθως εξάγει την "τιμή" ομοιότητας μεταξύ δύο κόμβων από διάφορες μετρήσεις ομοιότητας και χρησιμοποιεί την τιμή ομοιότητας για να προβλέψει την σύνδεσή τους. Οι ερευνητές συγκεντρώθηκαν κυρίως στην απόδοση διαφόρων μετρήσεων ομοιότητας σε γράφους για την πρόβλεψη συνδέσμων [20]. Αργότερα, οι Hasan,Mohammad A.,Chaoji, και άλλοι [20], συνέχισαν την μελέτη με δύο τρόπους. Πρώτα, έδειξαν ότι η αξιοποίηση δεδομένων εξωτερικών από την μελέτη του γράφου μπορεί να βελτιώσει σημαντικά το αποτέλεσμα της πρόβλεψης. Μετά, χρησιμοποίησαν διάφορες μετρήσεις ομοιότητας σαν χαρακτηριστικά σε ένα περιβάλλον επιβλεπόμενης εκμάθησης όπου το πρόβλημα της πρόβλεψης συνδέσμων ορίζεται σαν μια διαδικασία δυαδικής ταξινόμησης. Από τότε η προσέγγιση ταξινόμησης επιβλεπόμενης εκμάθησης χρησιμοποιείται σε πολλές άλλες μελέτες πάνω στο θέμα.

Η μελέτη της εξέλιξης των κοινωνικών δικτύων [21 22 23] είναι αρκετά όμοια με το πρόβλημα της πρόβλεψης συνδέσμων. Ένα μοντέλο εξέλιξης προβλέπει τους μελλοντικούς κόμβους ενός δικτύου, βασισμένο σε κάποια διαδεδομένα χαρακτηριστικά των κοινωνικών δικτύων, όπως η κατανομή βαθμού νόμου δύναμης (power law degree distribution) [21], και το φαινόμενο του μικρού κόσμου [22]. Αυτή αποτελεί την βασική διαφορά ανάμεσα στα μοντέλα εξέλιξης και στα μοντέλα πρόβλεψης συνδέσμων. Τα πρώτα αξιοποιούν τα καθολικά χαρακτηριστικά του δικτύου, ενώ τα δεύτερα μοντελοποιούν τις καταστάσεις του γράφου για να προβλέψουν την πιθανότητα ύπαρξης σύνδεσης ανάμεσα σε δύο κόμβους.

Το πρόβλημα της πρόβλεψης συνδέσμων μπορεί να μοντελοποιηθεί σαν μια διαδικασία δυαδικής ταξινόμησης, όπου το κάθε δεδομένο αντιπροσωπεύει ένα ζευγάρι κόμβων στον γράφο του κοινωνικού δικτύου. Χρησιμοποιείται δηλαδή η ανάλυση του γράφου του κοινωνικού δικτύου συνεργατικά με τους αλγόριθμους επιβλεπόμενης εκμάθησης Για την εκμάθηση του μοντέλου, μπορούμε να χρησιμοποιήσουμε την πληροφορία συνδέσμων από το διάστημα εκμάθησης [t0,t0']. Από αυτό το μοντέλο, προβλέψεις μελλοντικών συνδέσμων μπορούν να εξαχθούν για το διάστημα δοκιμής [t1,t1']. Πιο συγκεκριμένα, υποθέτουμε ότι οι u,v είναι δύο κόμβοι του γράφου G(V,E) και η "ετικέτα" κάθε δεδομένου <u,v> είναι y^(u,v). Υποθέτουμε ότι οι αλληλεπιδράσεις ανάμεσα στους u

και v είναι συμμετρικές, επομένως το δεδομένο <u,v> και το <v,u> είναι τα ίδια, άρα και y^(u,v)=y^(v,u). Επομένως θα ισχύει:

$$y^{\langle u,v\rangle} = \begin{cases} +1, & \text{if } \langle u,v\rangle \in E \\ -1, & \text{if } \langle u,v\rangle \notin E \end{cases}$$

Χρησιμοποιώντας τους παραπάνω συμβολισμούς για ένα σύνολο δεδομένων εκμάθησης, κατασκευάζουμε ένα μοντέλο ταξινόμησης που μπορεί να προβλέψει τις άγνωστες τιμές των ετικετών ζευγαριών κόμβων <u,v> στον γράφο G[t1,t1'].

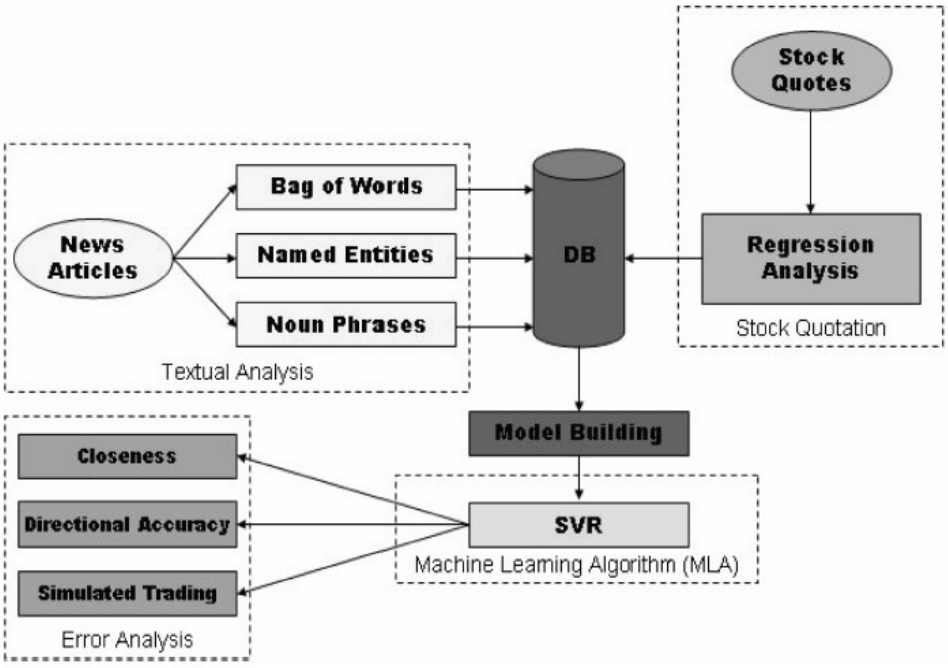
Αυτή είναι μια τυπική διαδικασία ταξινόμησης και οποιαδήποτε από τα διαδεδομένες μεθόδους επιβλεπόμενης εκμάθησης όπως τα νευρωνικά δίκτυα και οι μηχανές διανύσματος υποστήριξης μπορούν να χρησιμοποιηθούν. Η κύρια πρόκληση είναι η επιλογή της ομάδας χαρακτηριστικών για την διαδικασία της ταξινόμησης. Τα χαρακτηριστικά που χρησιμοποιούνται εξάγονται από την τοπολογία του γράφου, και βασίζονται στην γειτονική περιοχή των κόμβων που εξετάζονται και στην απόσταση που έχουν οι κόμβοι μεταξύ τους μέσα στον γράφο.

3.5 Συνεργατική Μοντελοποίηση (ensemble modeling)

Ένα πολύ ενδιαφέρον πείραμα έγινε το 2012 σε ένα σεμινάριο με θέμα τα Predictive Analytics από τον Gary Panchoo. Τοποθέτησε στο τραπέζι μπροστά του ένα βάζο γεμάτο με χαρτονομίσματα του ενός δολαρίου, και ζήτησε από το κάθε άτομο ξεχωριστά να κάνει μια πρόβλεψη του πόσα χαρτονομίσματα υπήρχαν μέσα στο βάζο. Το βάζο είχε τελικά 362 χαρτονομίσματα. Ο νικητής του διαγωνισμού, εκτίμησε το συνολικό ποσό λανθασμένα κατά 10 δολάρια. Αυτό που παρουσιάζει ενδιαφέρον όμως, που είναι και αυτό που ήθελε να αποδείξει ο Panchoo, είναι ότι ο μέσος όρος των συνολικών 61 προβλέψεων, απείχε μόνο 3 δολάρια από το πραγματικό ποσό. Η λογική εξήγηση σε αυτό το φαινόμενο είναι ότι οι υπερεκτιμήσεις και η υποτιμήσεις των ανθρώπων αλληλοεξουδετερώνονται. Αν υποθέσουμε ότι οι άνθρωποι υπερεκτιμούν με τον ίδιο βαθμό που υποτιμούν, η εφαρμογή του μέσου όρου εξαλείφει τα σφάλματα και τελικά το νούμερο που προκύπτει είναι πιο κοντά στο αληθινό. Επομένως μπορούμε να υποθέσουμε ότι όσο περισσότεροι άνθρωποι προβλέπουν, τόσο πιο κοντά θα είναι ο μέσος όρος στην πραγματική τιμή.

Όπως ένα πλήθος από άτομα, ένα σύνολο από προβλεπτικά μοντέλα και μεθόδους επωφελείται από το ίδιο φαινόμενο της συνεργατικής αποτελεσματικότητας. Κάθε μοντέλο έχει τις δυνάμεις και τις αδυναμίες του, και όπως με τις ανθρώπινες εκτιμήσεις, οι προβλέψεις ενός μοντέλου δεν είναι αψεγάδιαστες. Επομένως αξιοποιώντας διαφορετικές μεθόδους συνεργατικά, εξάγονται καλύτερες και πιο πολύπλευρες προβλέψεις, με την έννοια ότι το πρόβλημα προσεγγίζεται με πολλούς διαφορετικούς τρόπους παράλληλα.

Στην εικόνα φαίνεται το μοντέλο που χρησιμοποιήθηκε στην έρευνα [30] που είχε ως σκοπό την πρόβλεψη τιμών μετοχών με βάση τα άρθρα επικαιρότητας του χρηματιστηρίου:

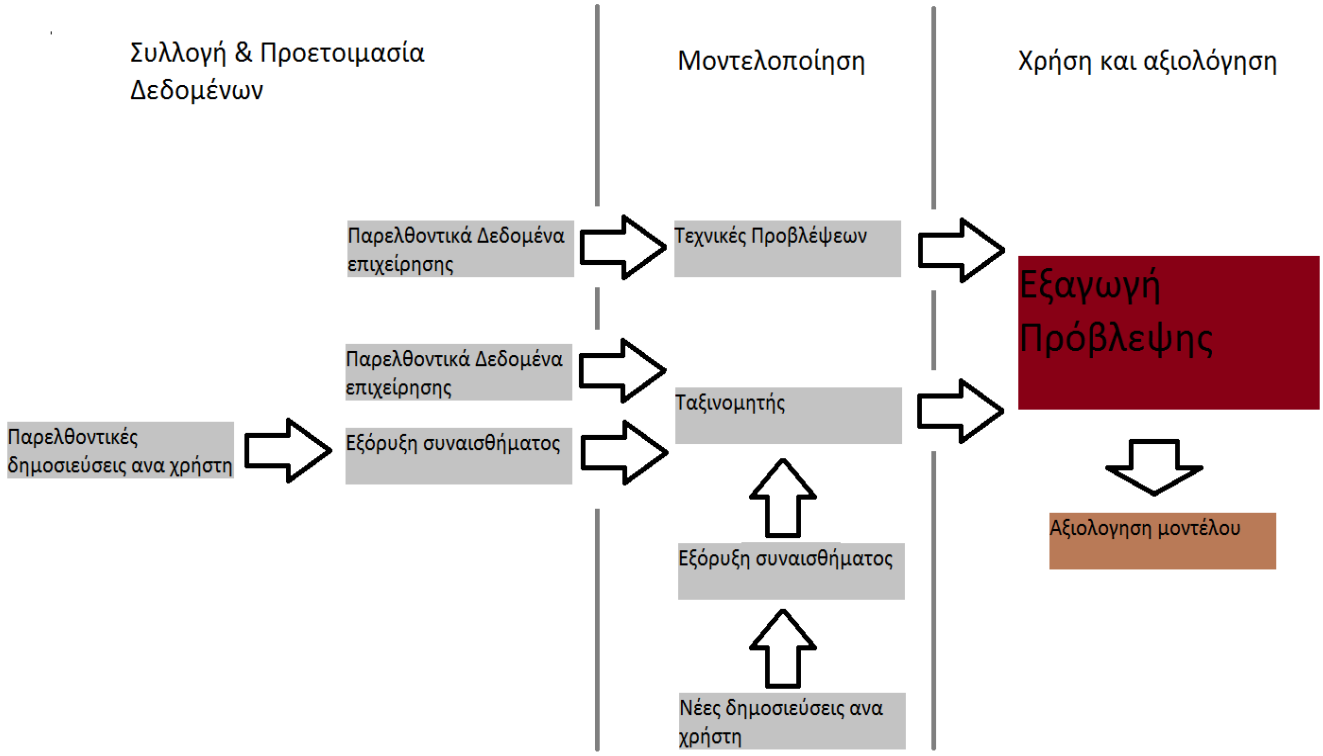


Παρατηρούμε ότι οι αναλυτές χρησιμοποίησαν μια μέθοδο αξιολόγησης των δημοσιεύσεων με βάση τα στοιχεία του κειμένου τους, σε συνεργασία με εξόρυξη συναισθήματος και ανάλυση παλινδρόμησης (Regression Analysis) στις χρονοσειρές των μετοχών. Αυτή είναι μια πρακτική που χρησιμοποιείται πολύ συχνά. Ο συνδυασμός δηλαδή μεθόδων που λαμβάνουν υπ' όψη τους κάποια χαρακτηριστικά, όπως το συναίσθημα των καταναλωτών σε σχέση με κάποιο προϊόν, με καθαρά μαθηματικές μεθόδους τεχνικών προβλέψεων. Έτσι το πρόβλημα προσεγγίζεται λαμβάνοντας υπ' όψη και τα χαρακτηριστικά των ανθρώπων, αλλά και την μαθηματική θεωρία.

Από τις έρευνες και εφαρμογές που μελετήθηκαν, οι περισσότερες αν όχι όλες αξιοποιούν την συνεργατική μοντελοποίηση με κάποια μορφή. Πλέον η έρευνα φαίνεται να συγκεντρώνεται στο πως τα διαφορετικά εργαλεία των Predictive Analytics μπορούν να

συνδυαστούν ώστε να επιτευχθούν τα καλύτερα δυνατά αποτελέσματα. Τα μοντέλα που πρόκειται να συνεργαστούν συνήθως επιλέγονται με βασικό κριτήριο τα δεδομένα που είναι διαθέσιμα, που στις περισσότερες περιπτώσεις είναι διαφορετικού τύπου. Ένα σημαντικό πλεονέκτημα της συνεργατική μοντελοποίησης είναι ότι δίνει την δυνατότητα στον αναλυτή να δώσει την βαρύτητα που χρειάζεται στα δεδομένα κάθε τύπου, αναθέτοντας βάρη στις προβλέψεις κάθε μεθόδου στο στάδιο που αυτές συναθροίζονται. Στον σχεδιασμό προϊόντων και υπηρεσιών ένα συνεργατικό μοντέλο θα μπορούσε να είναι ένα που εφαρμόζει εξόρυξη συναισθήματος στις δημοσιεύσεις των καταναλωτών στα μέσα κοινωνικής δικτύωσης ώστε να αξιολογηθούν οι ανάγκες τους, χρησιμοποιώντας παράλληλα ανάλυση επιρροής ώστε να προβλεφθεί πως αυτές οι ανάγκες θα μεταδοθούν στους καταναλωτές. Αυτή η αξιοποίηση της ανάλυσης επιρροής θα γίνει με βάση το γεγονός ότι οι χρήστες που παρουσιάζουν μεγάλη επιρροή στα κοινωνικά δίκτυα είναι συνήθως αυτοί που καθορίζουν τις νέες αγοραστικές τάσεις. Επιπλέον, η ανάλυση επιρροής θα μπορούσε να εφαρμοστεί συνδυάζοντας τα δεδομένα από την μελέτη του γράφου που μοντελοποιεί το κοινωνικό δίκτυο, και τα δεδομένα που προκύπτουν άμεσα από τα βασικά χαρακτηριστικά του κάθε χρήστη μέσα στο μέσο κοινωνικής δικτύωσης (αριθμός φίλων, δημοσιεύσεις και σχόλια σε αυτές, κτλ), μια διαδικασία που αποτελεί και αυτή μια μορφή συνεργατικής μοντελοποίησης. Μια άλλη συνδυαστική εφαρμογή θα μπορούσε να είναι η ανάλυση των χαρακτηριστικών του προϊόντος ή της υπηρεσίας που σχεδιάζεται, ώστε να εξαχθούν προβλέψεις πωλήσεων βασισμένες σε ήδη υπάρχοντες κανόνες συσχετισμού. Το μοντέλο θα αξιολογεί τα τεχνικά χαρακτηριστικά του προϊόντος, και θα επιχειρεί να το συσχετίσει με τους κανόνες συσχετισμού που ισχύουν για προϊόντα που παρουσιάζουν παρόμοια χαρακτηριστικά.

Για να γίνει πιο κατανοητή η αξιοποίηση της συνεργατικής μοντελοποίησης και των δεδομένων των μέσων κοινωνικής δικτύωσης σε ένα προβλεπτικό μοντέλο, παρουσιάζεται ένα παράδειγμα. Υποθέτουμε ότι μια επιχείρηση που παρέχει υπηρεσία (πχ εταιρία κινητής τηλεφωνίας) θέλει να δημιουργήσει ένα μοντέλο αποχώρησης, που θα προβλέπει πόσοι πελάτες πρόκειται να αποχωρήσουν τον επόμενο μήνα με βάση τις δημοσιεύσεις των πελατών της στα μέσα κοινωνικής δικτύωσης. Μία προσέγγιση της μοντελοποίησης φαίνεται στο σχήμα:



Η επιχείρηση αξιοποιεί τις παρελθοντικές δημοσιεύσεις των χρηστών της που είναι σχετικές με την επιχείρηση, και τα δεδομένα που έχει στα αρχεία της για τους πελάτες της. Αφού εφαρμοστεί εξόρυξη συναισθήματος στις παρελθοντικές δημοσιεύσεις, τα συναισθήματα των πελατών εισάγονται μαζί με τα παρελθοντικά δεδομένα της επιχείρησης σε έναν ταξινομητή ως δεδομένα εκμάθησης. Ο ταξινομητής συνδυάζει τα παρελθοντικά συναισθήματα των πελατών και τις παρελθοντικές αποχωρήσεις τους, ώστε να μπορεί πλέον να δεχτεί σαν είσοδο τις νέες δημοσιεύσεις ενός πελάτη μέσα στον μήνα (που αφορούν την επιχείρηση), και να μπορεί με βάση τα παρελθοντικά δεδομένα να αξιολογήσει την μελλοντική πιθανότητα αποχώρησής του. Παράλληλα, τα παρελθοντικά δεδομένα της επιχείρησης, σε μορφή "αποχωρήσεις ανά μήνα", εισάγονται σε ένα μοντέλο τεχνικών προβλέψεων (πχ γραμμική παλινδρόμηση) για να εξαχθεί μια καθαρά μαθηματική εκτίμηση του αριθμού των πελατών που πρόκειται να αποχωρήσουν. Τελικά οι προβλέψεις από τον ταξινομητή και την γραμμική παλινδρόμηση συναθροίζονται ώστε να εξαχθεί η τελική πρόβλεψη. Το μοντέλο χρησιμοποιείται και αξιολογείται. Ένα πλεονέκτημα αυτού του μοντέλου είναι ότι η λειτουργία του μπορεί να προσαρμοστεί ανάλογα με τα δεδομένα που είναι διαθέσιμα. Ρυθμίζοντας τα βάρη των δύο διαφορετικών προβλέψεων στην τελική άθροιση, το μοντέλο

μπορεί να ρυθμιστεί ώστε να δίνει περισσότερη σημασία στις δημοσιεύσεις των πελατών εφ' όσον είναι αρκετές, και στα παρελθοντικά δεδομένα στην αντίθετη περίπτωση. Η αξιολόγηση των μοντέλων παρουσιάζεται στα επόμενα κεφάλαια.

Παρατηρούμε επομένως ότι η συνεργατική μοντελοποίηση δίνει πολλές δυνατότητες συνδυασμού των εργαλείων και αλγορίθμων είτε σειριακά είτε παράλληλα. Η πρόκληση που παρουσιάζεται είναι το πως ο αναλυτής θα συνδυάσει με τον καλύτερο τρόπο τα εργαλεία που έχει στην διάθεσή του ώστε να προσαρμοστούν στα δεδομένα που είναι διαθέσιμα και στον στόχο του μοντέλου.

3.6 Μετρικές και Αλγόριθμοι Αξιολόγησης Μοντέλου

3.6.1.1 Μετρικές Αξιολόγησης Αριθμητικών Προβλέψεων

Για την αξιολόγηση των μοντέλων που εξάγουν αριθμητικές προβλέψεις, χρησιμοποιούνται διάφορα στατιστικά μεγέθη που αξιολογούν τις προβλέψεις του αλγόριθμου σε σχέση με τις αληθινές τιμές που προκύπτουν:

1. **Συντελεστής R-τετράγωνο**: Είναι ένα στατιστικό μέγεθος που δίνει πληροφορία για την ποιότητα πρόβλεψης του μοντέλου. Συγκεκριμένα στην παλινδρόμηση, ο συντελεστής R-τετράγωνο δείχνει το πόσο η ευθεία που προκύπτει από την παλινδρόμηση πλησιάζει τα πραγματικά δεδομένα. Παίρνει τιμές από 0 μέχρι 1, με την τιμή 1 να δείχνει ότι η καμπύλη παλινδρόμησης ταιριάζει απόλυτα στα πραγματικά δεδομένα. Ο πιο συνηθισμένος ορισμός του είναι ο εξής:

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

   όπου στο κλάσμα ο αριθμητής είναι το άθροισμα της διακύμανσης των προβλέψεων, και παρονομαστής το άθροισμα της διακύμανσης των πραγματικών δεδομένων.

2. **Προσαρμοσμένος συντελεστής R-τετράγων**ο: Δίνει καλύτερη εκτίμηση του R-τετράγωνο καθώς δεν επηρεάζεται ιδιαίτερα από τις ακραίες τιμές. Ενώ ο R-τετράγωνο αυξάνεται όταν προστίθεται μια καινούργια μεταβλητή, ο

προσαρμοσμένος συντελεστής αυξάνεται μόνο αν η καινούργια μεταβλητή φέρει ουσιώδη πληροφορία. Μπορεί να πάρει και αρνητικές τιμές, και η τιμή του θα είναι πάντα ίση η μικρότερη του R-τετράγωνο. Υπολογίζεται ως:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

όπου n είναι το πλήθος δειγμάτων και p είναι το πλήθος των μεταβλητών της παλινδρόμησης.

3. **Μέσο τετραγωνικό σφάλμα (MSE)**: Αποτελεί το πιο βασικό κριτήριο αξιολόγησης προβλεπτικών αλγορίθμων. Μετράει την διαφορά μεταξύ των προβλέψεων και των πραγματικών τιμών. Όσο μικρότερος είναι ο δείκτης, τόσο μικρότερο το σφάλμα, επομένως είναι επιθυμητή μικρή τιμή. Αυτός ο δείκτης μπορεί να επηρεαστεί από ακραίες τιμές και να παρουσιάσει μεγαλύτερο σφάλμα από αυτό που ισχύει πραγματικά.
4. **Ρίζα μέσου τετραγωνικού σφάλματος (RMSE):** Λειτουργεί όπως το μέσο τετραγωνικό σφάλμα, αλλά υπολογίζει και την τετραγωνική ρίζα του.
5. **Μέσο απόλυτο σφάλμα (MAE):** Όπως φανερώνει και το όνομά του, αυτός ο δείκτης υπολογίζει τον μέσο όρο των σφαλμάτων των προβλέψεων. Είναι χρήσιμος στο να ομαλοποιεί τις ακραίες τιμές:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

όπου f είναι οι προβλέψεις, και y οι πραγματικές τιμές.

6. **Συντελεστής Αποδοτικότητας:** Έχει χρησιμοποιηθεί σε πολλές επιστήμες για την αξιολόγηση της απόδοσης των μοντέλων. Ορίζεται ως:

$$E = 1 - \frac{\sum_{i-1}^{n}(O_i - X_i)^2}{\sum_{i-1}^{n}(O_i - \bar{X})^2} = 1 - \frac{MSE}{Variance\_of\_Observed}$$

όπου ο αριθμητής του κλάσματος είναι το μέσο τετραγωνικό σφάλμα, και ο παρονομαστής η διακύμανση των πραγματικών δεδομένων. Παίρνει τιμές από -1 ως 1, όπου η τιμή -1 φανερώνει ένα κακό μοντέλο.

7. **Δείκτης Κατάστασης/Βάρος συντελεστών παλινδρόμησης**: Αφού έχει κατασκευαστεί ένα μοντέλο, τα βάρη των συντελεστών παλινδρόμησης μπορούν να αποτελέσουν κριτήριο για την ποιότητα του μοντέλου. Αν υπάρχουν αχρείαστες είσοδοι στα δεδομένα, τα βάρη των συντελεστών αυξάνονται. Συνήθως όταν ο δείκτης κατάστασης είναι υψηλός, το μοντέλο παρουσιάζει υψηλή προκατάληψη, που αυξάνει την αβεβαιότητά του. Υψηλή αβεβαιότητα σημαίνει αυξημένη ασυνέπεια του μοντέλου, δηλαδή ότι δεν είναι ρεαλιστικό, και ότι η χρήση του δεν πρέπει να επαναληφθεί.
8. **Αριθμός των μεταβλητών που χρησιμοποιήθηκαν:** Αυτός ο αριθμός αποτελεί μια ένδειξη για την ποιότητα του μοντέλου. Ένα αποδοτικό μοντέλο διαλέγει τις βέλτιστες μεταβλητές που είναι πιο αντιπροσωπευτικές της πληροφορίας που χρειάζεται. Ένας αναλυτής θα πρέπει να στοχεύει στην δημιουργία ενός μοντέλου με το μικρότερο δυνατό αριθμό μεταβλητών, μεγιστοποιώντας την πληροφορία που μπορεί να εξηγήσει.

Με βάση αυτά τα στατιστικά κριτήρια, έχουν γίνει διάφορες έρευνες για την αξιολόγηση των γραμμικών μεθόδων που περιγράφτηκαν στην προηγούμενη ενότητα.

3.6.1.2 Αξιολόγηση Κατηγοριοποίησης Ταξινομητών

Πίνακας Σύγχυσης (Confusion Matrix): Στο πεδίο της μηχανικής εκμάθησης ο πίνακας σύγχυσης επιτρέπει την απεικόνιση της απόδοσης ενός αλγορίθμου επιβλεπόμενης εκμάθησης. Δείχνει τον αριθμό των σωστών και λανθασμένων προβλέψεων του μοντέλου ταξινόμησης σε σύγκριση με τα πραγματικά δεδομένα. Ο πίνακας έχει διαστάσεις NxN όπου N είναι ο αριθμός των κατηγοριών ταξινόμησης. Ο πίνακας που ακολουθεί αποτελεί ένα παράδειγμα δύο κατηγοριών (θετικό, αρνητικό).

Πίνακας Σύγχυσης		Πραγματικά			
		Θετικό	Αρνητικό		
Μοντέλο	Θετικό	a	b	Θετική Προβλεπτική Τιμή	a/(a+b)
	Αρνητικό	c	d	Αρνητική Προβλεπτική Τιμή	d/(c+d)
		Ευαισθησία	Ειδικότητα	Ακρίβεια = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

.

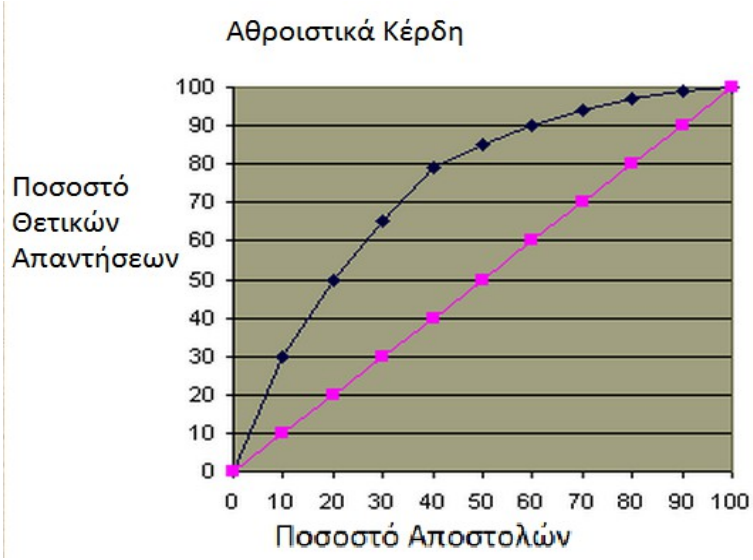
Η ακρίβεια είναι ο συνολικός αριθμός των προβλέψεων που ήταν σωστές. Η θετική προβλεπτική τιμή είναι ο αριθμός των θετικών περιπτώσεων που προβλέφθηκαν σωστά, και αντίστοιχα η αρνητική προβλεπτική τιμή. Η ευαισθησία (true-positive rate) είναι η αναλογία των θετικών περιπτώσεων που αναγνωρίστηκαν σωστά, και η ειδικότητα (false-positive rate) η αναλογία των αρνητικών περιπτώσεων που αναγνωρίστηκαν σωστά.

Γραφική Παράσταση αθροιστικού κέρδους και lift: : Το αθροιστικό κέρδος είναι ένα μέγεθος που αξιολογεί την αποτελεσματικότητα ενός προβλεπτικού μοντέλου, και εκφράζεται από τον λόγο των αποτελεσμάτων που αποκτούνται με την χρήση του μοντέλου, προς τα αποτελέσματα χωρίς την χρήση του. Το lift, όπως έχει αναφερθεί δείχνει πόσοι περισσότεροι πελάτες μπορούν να αναγνωριστούν από το μοντέλο, από ότι χωρίς. Χρησιμοποιούμε την γραφική παράσταση του αθροιστικού κέρδους μαζί με το lift για να έχουμε μια καλύτερη απεικόνιση της απόδοσης του αλγορίθμου. Αντίθετα με τον πίνακα σύγχυσης που αξιολογεί το μοντέλο με βάση όλο τον πληθυσμό, αυτή η γραφική παράσταση αντιπροσωπεύει μόνο ένα τμήμα του πληθυσμού.

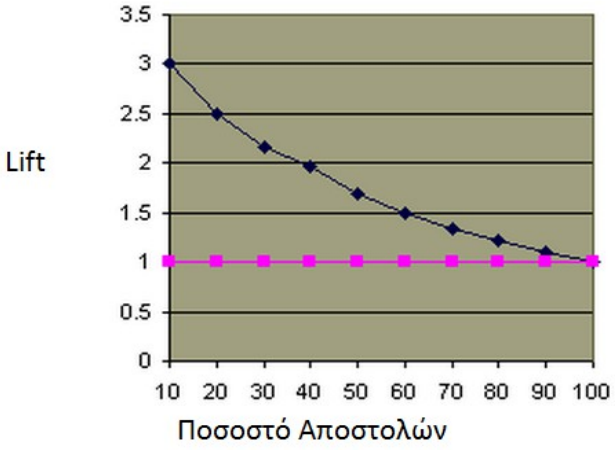
Θεωρούμε για παράδειγμα ότι μια επιχείρηση σκοπεύει να επικοινωνήσει με τους πελάτες της για λόγους προώθησης μέσω αλληλογραφίας. Κάθε αποστολή κοστίζει ένα ευρώ, και υπάρχουν 100.000 πελάτες. Σκοπός της επιχείρησης είναι να προβλέψει τις θετικές απαντήσεις. Υποθέτουμε ότι δεν υπάρχει μοντέλο ανταπόκρισης, και θεωρούμε ότι ο βαθμός ανταπόκρισης είναι 20%. Επομένως με κόστος 100.000 ευρώ, η επιχείρηση θα επικοινωνήσει με 100.000 πελάτες, από τους οποίους θα ανταποκριθούν οι 20.000. Χρησιμοποιώντας ένα μοντέλο ανταπόκρισης, το οποίο ταξινομεί τους πελάτες με βάση την πιθανότητα ανταπόκρισής τους, και προβλέπει τα αποτελέσματα επικοινωνίας μόνο με τους πρώτους 10.000, 20.000 πελάτες και τους υπόλοιπους, έχουμε τα εξής αποτελέσματα:

Κόστος	Αποστολές	Θετικές Απαντήσεις
10000	10000	6000
20000	20000	10000
30000	30000	13000
40000	40000	15800
50000	50000	17000
60000	60000	18000
70000	70000	18800
80000	80000	19400
90000	90000	19800
100000	100000	20000

Με βάση αυτά τα νούμερα, σχηματίζουμε την γραφική παράσταση του αθροιστικού κέρδους. Ο άξονας y αντιπροσωπεύει το ποσοστό των θετικών απαντήσεων, σε σχέση με τις συνολικές πιθανές θετικές απαντήσεις που είναι οι 20.000. Ο άξονας x δείχνει το ποσοστό των πελατών στους οποίους γίνεται η αποστολή. Η ροζ γραμμή αντιπροσωπεύει την περίπτωση που δεν χρησιμοποιείται μοντέλο ανταπόκρισης, σύμφωνα με την οποία αν γίνει αποστολή στο Χ% των πελατών, θα υπάρξουν Χ% θετικές απαντήσεις από τις συνολικές πιθανές. Η καμπύλη του αθροιστικού κέρδους προκύπτει ενώνοντας τα σημεία που προκύπτουν από τον αλγόριθμο ανταπόκρισης.

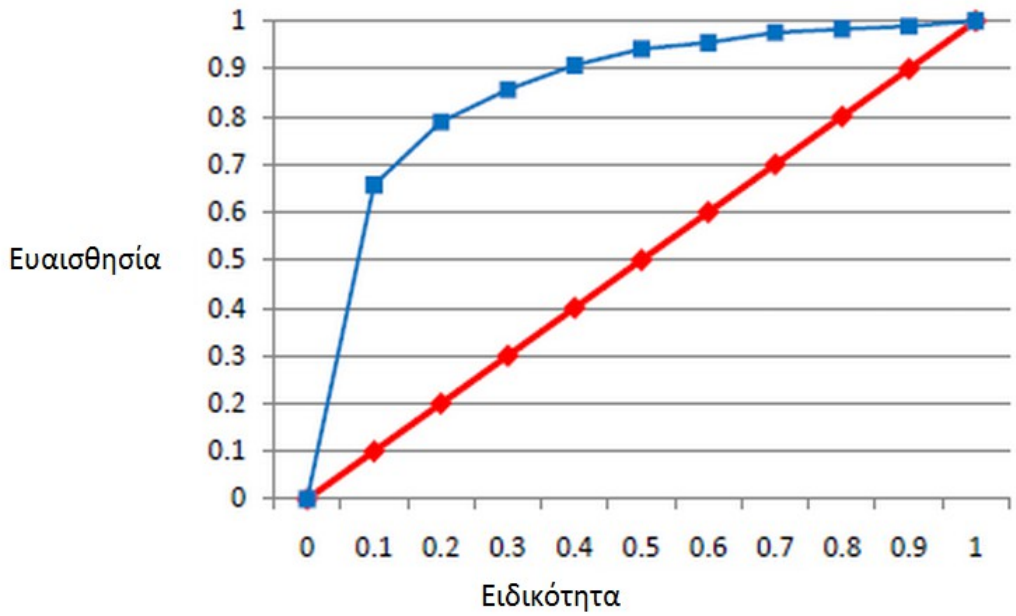


Σχεδιάζεται επίσης και η γραφική παράσταση του lift. Τα σημεία της καμπύλης υπολογίζονται από τον λόγο των πελατών που προβλέπονται από το μοντέλο προς τους πελάτες που υπολογίζονται χωρίς το μοντέλο. Για παράδειγμα, για αποστολή στο 10% των πελατών, χωρίς μοντέλο θα είχαμε το 10% των θετικών απαντήσεων, ενώ με τον μοντέλο θα είχαμε το 30% Επομένως στο σημείο του 10% των πελατών, το lift είναι 30/10=3.

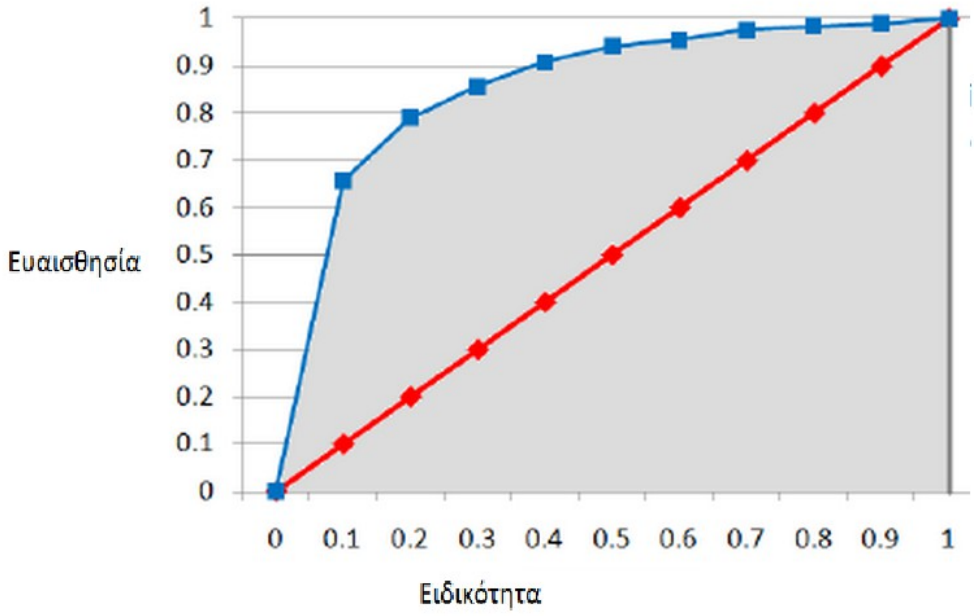


Και στις δύο γραφικές παραστάσεις, σκοπός είναι η μεγιστοποίηση της απόστασης της καμπύλης από την ροζ γραμμή που αντιπροσωπεύει την απουσία μοντέλου ανταπόκρισης.

Γράφημα ROC (receiver operaint characteristic): Το γράφημα ROC είναι παρόμοιο με το γράφημα κέρδους και lift, με την έννοια ότι παρέχει ένα μέσο σύγκρισης μεταξύ μοντέλων ταξινόμησης. Στον άξονα του x τοποθετείται η ειδικότητα, δηλαδή η πιθανότητα το μοντέλο να προβλέψει τιμή 1 ενώ η πραγματική τιμή είναι 0, και στον άξονα του y τοποθετείται η ευαισθησία, δηλαδή η πιθανότητα το μοντέλο να προβλέψει τιμή 1 ενώ η πραγματική είναι 1. Ιδανικά η καμπύλη θα κορυφωθεί στο αριστερό μέρος, σημαίνοντας ότι το μοντέλο πρόβλεψε σωστά. Στην ίδια γραφική παράσταση χρησιμοποιείται και ένα τυχαίο μοντέλο για σύγκριση.



Ενδιαφέρει επίσης και η περιοχή κάτω από την καμπύλη ROC, καθώς χρησιμοποιείται ως μέτρηση ποιότητας για τα μοντέλα ταξινόμησης. Ένας τυχαίος ταξινομητής έχει περιοχή κάτω από την καμπύλη μέτρου 0,5, ενώ η περιοχή κάτω από την καμπύλη για έναν τέλειο ταξινομητή είναι ίση με 1. Στην πραγματικότητα, τα περισσότερα μοντέλα έχουν περιοχή μεταξύ του 0.5 και του 1.



Μια περιοχή 0,8 σημαίνει ότι μια τυχαία επιλεγμένη περίπτωση από το σύνολο που έχει προβλεφθεί ότι έχει τιμή 1, με πιθανότητα 80%, έχει μεγαλύτερη τιμή από οποιαδήποτε άλλη τυχαία από το σύνολο που έχει προβλεφθεί με τιμή 0.

3.6.1.3 Ανάλυση Αβεβαιότητας

Ένα καλό προβλεπτικό μοντέλο αποβλέπει στην μείωση της αβεβαιότητάς του. Η αβεβαιότητα ενός μοντέλου μπορεί να μετρηθεί ως [27]:

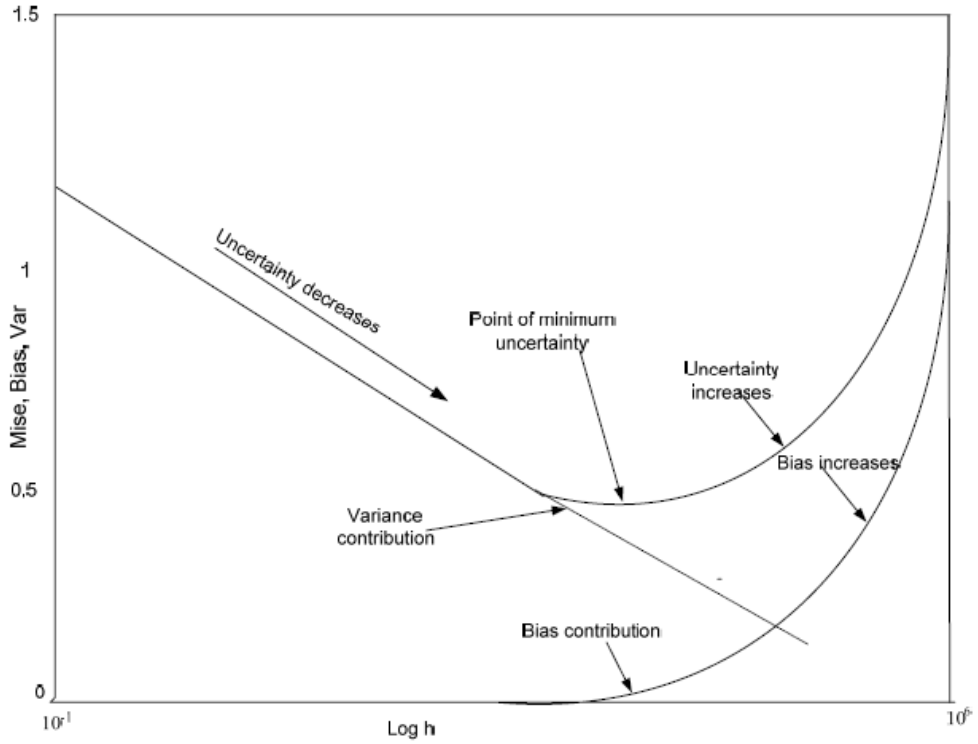
Αβεβαιότητα=Διακύμανση+Προκατάληψη^2

Υπάρχουν διάφοροι παράγοντες που αυξάνουν την αβεβαιότητα ενός μοντέλου:

1. Επιλογή Πληροφορίας: Η εισαγωγή περιττής πληροφορίας στο μοντέλο αυξάνει την διακύμανση των προβλέψεών του, και αντίστοιχα η ελλειπή τροφοδότησή του αυξάνει την προκατάληψή του.
2. Επιλογή του μοντέλου: Η επιλογή μεθόδου έχει μεγάλο ρόλο στην αβεβαιότητα του μοντέλου. Όταν μη γραμμικά δεδομένα τροφοδοτούνται σε ένα γραμμικό μοντέλο, το

αποτέλεσμα είναι προκατειλημμένο. Χρησιμοποιώντας ανάλυση πρωταρχικών συστατικών, μπορούμε να αποφανθούμε για την χρήση γραμμικής ή μη γραμμικής μεθόδου.

3. Πολυπλοκότητα και Κανονικοποίηση: Ένα πολύπλοκο μοντέλο είναι ένα στο οποίο τα γεγονότα περιγράφονται ως τυχαία. Η παρουσία γεγονότων που χαρακτηρίζονται ως τυχαία αυξάνει την αβεβαιότητα του μοντέλου, και δυσκολεύει την αναγνώριση των μοτίβων. Η κανονικοποίηση χρησιμοποιείται με σκοπό την σταθεροποίηση του αποτελέσματος, δίνοντας την πρέπουσα σημασία στις διακυμάνσεις των δεδομένων. Αυτό λέγεται ανταλλαγή διακύμανσης και προκατάληψης, και αποσκοπεί στην μείωση της προκατάληψης στο χαμηλότερο δυνατό σημείο, κρατώντας την προκατάληψη σε χαμηλό επίπεδο. Όταν ένα πολύπλοκο μοντέλο δεν κανονικοποιείται σε αρκετό βαθμό, τα δεδομένα προσαρμόζονται παραπάνω από όσο είναι επιθυμητό, και αυξάνεται η διακύμανση. Αντίθετα όταν ένα μοντέλο υπερκανονικοποιείται, η υπερβολική εξομάλυνση αυξάνει την προκατάληψη. Ένα σωστά κανονικοποιημένο μοντέλο θα μειώσει την διακύμανσή του και θα δώσει ένα πιο αμερόληπτο αποτέλεσμα[27]:



Στο σχήμα φαίνεται αυτή η ανταλλαγή σε συνάρτηση με την παράμετρο κανονικοποίησης h. Όσο η παράμετρος αυξάνεται, η διακύμανση του μοντέλου μειώνεται, μαζί με την αβεβαιότητα. Η πτώση της διακύμανσης συνεχίζεται μέχρι το σημείο που η αβεβαιότητα γίνεται ελάχιστη. Στο ίδιο σημείο η προκατάληψη του μοντέλου ξεκινάει να γίνεται ουσιώδη.

Περαιτέρω αύξηση της παραμέτρου κανονικοποίησης σημαίνει σημαντική αύξηση της προκατάληψης, και αύξηση της αβεβαιότητας, ανεξάρτητα από την μείωση της διακύμανσης.

4. Ένας άλλος παράγοντας αβεβαιότητας είναι η παρουσία θορύβου. Τα θορυβώδη δεδομένα δεν ακολουθούν κάποιο μοτίβο, για αυτό είναι αναγκαίο να αφαιρεθεί ο θόρυβος από τα δεδομένα ώστε να μειωθεί η αβεβαιότητα.
5. Η αβεβαιότητα επηρεάζεται επίσης από το μέγεθος των διαθέσιμων δεδομένων για την κατασκευή του μοντέλου. Αν η ποσότητα δεδομένων είναι μεγάλη, τότε υπάρχει μεγαλύτερη πιθανότητα για ένα καλό μοντέλο. Όσο λιγότερα δεδομένα υπάρχουν, τόσο πιο αβέβαιο είναι το μοντέλο να αντιπροσωπεύει την πραγματικότητα.

4 Σύγκριση και Αξιολόγηση Προβλεπτικών Μοντέλων

Σε αυτό το κεφάλαιο γίνεται μια τελική αξιολόγηση των μεθόδων που έχουν αναλυθεί μέχρι τώρα. Μελετήθηκε ένας μεγάλος αριθμός από έρευνες που αξιοποίησαν τις μεθόδους είτε απομονωμένα είτε συνεργατικά, και εξάχθηκαν συμπεράσματα για την λειτουργία τους. Οι έρευνες ήταν διαφορετικού σκοπού και περιεχομένου, επομένως τα μοντέλα προσαρμόζονταν ανάλογα. Μερικές από αυτές είχαν σαν σκοπό την πρόβλεψη τιμών μετοχών με βάση τις δημοσιεύσεις χρηστών σε οικονομικά forums, την πρόβλεψη των εσόδων από ταινίες με βάση τα μηνύματα χρηστών στα μέσα κοινωνικής δικτύωσης, την πρόβλεψη πωλήσεων ειδών ενδυμασίας με βάση τα χαρακτηριστικά τους και την πρόβλεψη παραγωγής μιας βιομηχανίας. Όλες οι έρευνες χρησιμοποίησαν τα εργαλεία που αναφέρθηκαν στο προηγούμενο κεφάλαιο για να αξιολογήσουν τους αλγορίθμους, και κυρίως τους δείκτες σφάλματος (MSE, RMSE, MAE). Παρόλο που μερικές μέθοδοι αποδίδουν ξεκάθαρα καλύτερα ή χειρότερα από άλλες σε μέσο όρο, υπάρχει μεγάλη διακύμανση στην απόδοση ανάλογα τον σκοπό και τα δεδομένα. Ακόμα και τα καλύτερα μοντέλα μερικές φορές έχουν κακή απόδοση, και μοντέλα με γενικά κακή απόδοση αποδίδουν πολύ υψηλά περιστασιακά.. Επομένως ενώ κάποιες μέθοδοι γενικότερα δεν αποδίδουν ικανοποιητικά, η χρήση τους δεν αποκλείεται αφού η μορφή των δεδομένων μπορεί να είναι τέτοια ώστε μία από αυτές τις μεθόδους να αποδώσει καλύτερα από κάποια που αποδίδει γενικότερα.

4.1 Αξιολόγηση Αλγορίθμων

Όσο αφορά τα καθαρά μαθηματικά μοντέλα τεχνικών προβλέψεων, έχουν γίνει πολλές έρευνες για την αξιολόγησή τους. Συγκεκριμένα, οι [25] χρησιμοποίησαν τους δείκτες RMSE, για να συγκρίνουν την παλινδρόμηση κορυφογραμμής, , την μέθοδο

ελαχίστων τετραγώνων και την μέθοδο μερικών ελαχίστων τετραγώνων, και κατέληξαν στο ότι η απόδοσή τους εξαρτάται από το μέγεθος και τον τύπο των δεδομένων. Οι [26] σύγκριναν την παλινδρόμηση κορυφογραμμής, τα ελάχιστα τετράγωνα, και την παλινδρόμηση πρωταρχικών συστατικών σε δύο διαφορετικά σύνολα δεδομένων, και διαπίστωσαν ότι απέδωσε καλύτερα ένας συνδυασμός παλινδρόμησης κορυφογραμμής και παλινδρόμησης πρωταρχικών συστατικών που ονόμασαν RPC. Οι [28] σύγκριναν την παλινδρόμηση πρωταρχικών συστατικών, τα μερικά ελάχιστα τετράγωνα και την παλινδρόμηση κορυφογραμμής σε μια έρευνα σχετική με το ανθρώπινο αίμα, και κατέληξαν στο ότι η παλινδρόμηση κορυφογραμμής είχε καλύτερα αποτελέσματα. Τέλος, οι Naes T και Irgens. C. [29] χρησιμοποίησαν το RMSE για να αξιολογήσουν την πολλαπλή γραμμική παλινδρόμηση, την παλινδρόμηση κορυφογραμμής, και τα μερικά ελάχιστα τετράγωνα, και είδαν ότι η πολλαπλή γραμμική παλινδρόμηση δεν απέδωσε όσο τα υπόλοιπα μοντέλα. Φαίνεται λοιπόν ότι ο σκοπός και η μορφή των δεδομένων επηρεάζουν σε μεγάλο βαθμό την απόδοση αυτών των τεχνικών.

Τα χαρακτηριστικά των δεδομένων που φαίνονται να επηρεάζουν την απόδοση αυτών των εργαλείων είναι σχετικά με την φύση των μεταβλητών που χρησιμοποιούνται. Ο κύριος παράγοντας φαίνεται να είναι το κατά πόσο οι μεταβλητές αυτές είναι συσχετισμένες μεταξύ τους, και το αν αυτός ο συσχετισμός είναι γραμμικός η όχι. Επιπλέον, ρόλο έχει και ο αριθμός των μεταβλητών που χρησιμοποιούνται, και το σε τι ποσοστό αυτές οι μεταβλητές επηρεάζουν τελικά την μεταβλητή που προβλέπεται. Καθώς στις έρευνες που μελετήθηκαν οι μεταβλητές που αξιοποιήθηκαν παρουσίαζαν κάποια μορφή συσχετισμού ο οποίος όμως δεν ήταν πάντα γραμμικός, τα καλύτερα αποτελέσματα φάνηκαν να έχουν τα μερικά ελάχιστα τετράγωνα, ακολουθούμενα από την παλινδρόμηση πρωταρχικών συστατικών. Αυτό βασίζεται στο γεγονός ότι αυτές οι μέθοδοι μπορούν να εντοπίσουν τον ακριβή συσχετισμό ανάμεσα στις μεταβλητές εισόδου και να τον διαχειριστούν ανάλογα. Αναγνωρίζοντας αυτό το συσχετισμό μπορούν επιπλέον να αποβάλουν όσες μεταβλητές δεν έχουν αλληλεπίδραση με την μεταβλητή εξόδου. Τα μερικά ελάχιστα τετράγωνα γενικότερα εξάγουν καλύτερες προβλέψεις από την παλινδρόμηση πρωταρχικών συστατικών, αλλά έχουν μεγαλύτερο υπολογιστικό κόστος. Από τις υπόλοιπες μεθόδους, η πιο αποδοτική φαίνεται να είναι η παλινδρόμηση κορυφογραμμής. Εξήγηση σε αυτό θα μπορούσε να είναι ότι και η παλινδρόμηση κορυφογραμμής μπορεί να διαχειριστεί τον συσχετισμό ανάμεσα στις μεταβλητές, χρησιμοποιώντας συντελεστές κανονικοποίησης. Η μέθοδος των μή γραμμικών μερικών ελαχίστων τετραγώνων έχει το μεγαλύτερο υπολογιστικό κόστος, αλλά σε αντίθεση με τις υπόλοιπες μεθόδους έχει μια ικανότητα να διαχειρίζεται τον μή γραμμικό συσχετισμό που μπορεί να υπάρχει στα δεδομένα. Αυτή η ικανότητά όμως μειώνει την απόδοση της

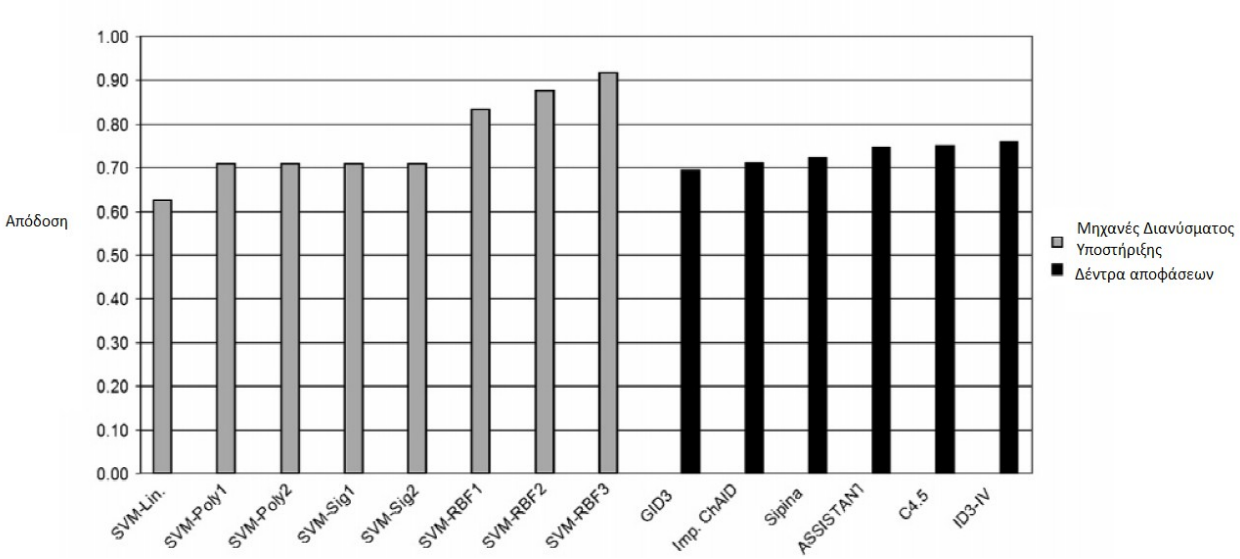
μεθόδου όταν ο συσχετισμός είναι γραμμικός, αφού εισάγει την μή γραμμικότητα στο μοντέλο. Λιγότερο αποδοτική φαίνεται να είναι η πολλαπλή γραμμική παλινδρόμηση, αφού δεν μπορεί να διαχειριστεί τον συσχετισμό. Τα πλεονεκτήματά της είναι ότι τα δεδομένα της δεν χρειάζονται κανονικοποίηση αντίθετα με τις υπόλοιπες τεχνικές, και ότι έχει χαμηλό υπολογιστικό κόστος. Φαίνεται να αποδίδει ικανοποιητικά μόνο όταν οι μεταβλητές εισόδου είναι τελείως ανεξάρτητες, και όταν δεν υπάρχουν αχρείαστες μεταβλητές στα δεδομένα.

Στην κατηγορία των ταξινομητών, η πιο αποδοτική και εφαρμοσμένη μέθοδος φαίνονται να είναι οι μηχανές διανύσματος υποστήριξης. Μια εξήγηση σε αυτό το φαινόμενο μπορεί να είναι η υψηλή ικανότητα παραμετροποίησης του αλγορίθμου, χρησιμοποιώντας την συνάρτηση kernel που εξυπηρετεί τον κάθε σκοπό αντίστοιχα, και μπορεί να είναι πιο περιεκτική από τον κανόνα πυροκρότησης των τεχνητών νευρωνικών δικτύων. Ακολουθούν τα νευρωνικά δίκτυα που αποδίδουν λίγο καλύτερα από τα δέντρα αποφάσεων, και τελευταία φαίνεται να είναι η αφελής μέθοδος του Bayes. Αυτό μπορεί να θεωρηθεί αναμενόμενο, αφού οι δύο τελευταίες διαδικασίες αποτελούν πολύ βασικές ταξινομήσεις βασισμένες σε απλούς τύπους πιθανοτήτων, ενώ οι δύο πρώτες είναι από την φύση τους πιο πολύπλοκες και αναλυτικές διαδικασίες. Ο παρακάτω πίνακας δείχνει τα αποτελέσματα ενός πειράματος σύγκρισης μεταξύ μιας μηχανής διανύσματος υποστήριξης και ενός νευρωνικού δικτύου, στην ικανότητά τους να ταξινομήσουν χημικές ενώσεις στις ομάδες που ανήκουν [36]:

Τιμή μεταβλητής C	Σωστά αποτελέσματα	Ποσοστό επιτυχίας
0.008	48/60	**80%**
0.01	44/60	73.33%
0.05	30/60	50%
0.08	26/60	43%
0.2	32/60	53.33%
3	14/60	23.33%
10	30/60	50%

Η παράμετρος C είναι μια παράμετρος ρύθμισης του κανόνα Kernel που χρησιμοποιήθηκε. Στα ίδια δεδομένα τα νευρωνικά δίκτυα είχαν απόδοση 73.33%. Το επόμενο γράφημα δείχνει τα αποτελέσματα σύγκρισης μηχανών διανύσματος υποστήριξης και δέντρων αποφάσεων

στην ικανότητα αναγνώρισης καταλυτών σε χημικές αντιδράσεις, με διαφορετικούς κανόνες Kernel και κανόνες ταξινόμησης [37]:



Οι ταξινομητές αξιολογούνται όχι μόνο με βάση την απόδοσή τους, αλλά και με βάση πολλές άλλες παραμέτρους. Μία από αυτές είναι η ταχύτητα εκμάθησης του αλγορίθμου, στην οποία οι μηχανές διανύσματος υποστήριξης είναι οι πιο χρονοβόρες, μαζί με τα νευρωνικά δίκτυα. Οι πιο γρήγορες μέθοδοι στην εκμάθηση φαίνονται να είναι τα δέντρα αποφάσεων και η αφελής μέθοδος του Bayes. Αυτό μπορεί να θεωρηθεί ότι προκύπτει επίσης από το γεγονός ότι οι δύο αυτές μέθοδοι είναι πολύ βασικές υπολογιστικά. Αποτέλεσμα αυτού του γεγονότος μπορεί να θεωρηθεί και η υψηλή διαφάνεια που έχουν αυτές οι δύο μέθοδοι σε αντίθεση με τα νευρωνικά δίκτυα και τις μηχανές διανύσματος υποστήριξης, η ικανότητα δηλαδή που παρέχουν στον αναλυτή να καταλάβει εύκολα το πως προκύπτει η ταξινόμηση. Στην ταχύτητα ταξινόμησης και οι τέσσερις μέθοδοι παρουσιάζουν ίδια απόδοση. Ένα άλλο μέτρο αξιολόγησης είναι το πόσο επιρρεπείς είναι οι μέθοδοι στο πρόβλημα του overfitting. Τον μικρότερο κίνδυνο φαίνεται να έχει η αφελής μέθοδος του Bayes, και τον μεγαλύτερο τα νευρωνικά δίκτυα, με τις άλλες δύο μεθόδους να βρίσκονται στο ενδιάμεσο.

Συγκρίνοντας τώρα τα μαθηματικά μοντέλα τεχνικών προβλέψεων με τους ταξινομητές, και συγκεκριμένα στον τομέα της καταναλωτικής συμπεριφοράς οι έρευνες στην συντριπτική πλειοψηφία τους δείχνουν ότι η απόδοση των ταξινομητών, και ειδικότερα των μηχανών διανυσμάτων υποστήριξης είναι πολύ υψηλότερη. Θα μπορούσαμε να πούμε ότι αυτό το γεγονός επιβεβαιώνει ένα συγκριτικό πλεονέκτημα που έχουν τα Predictive

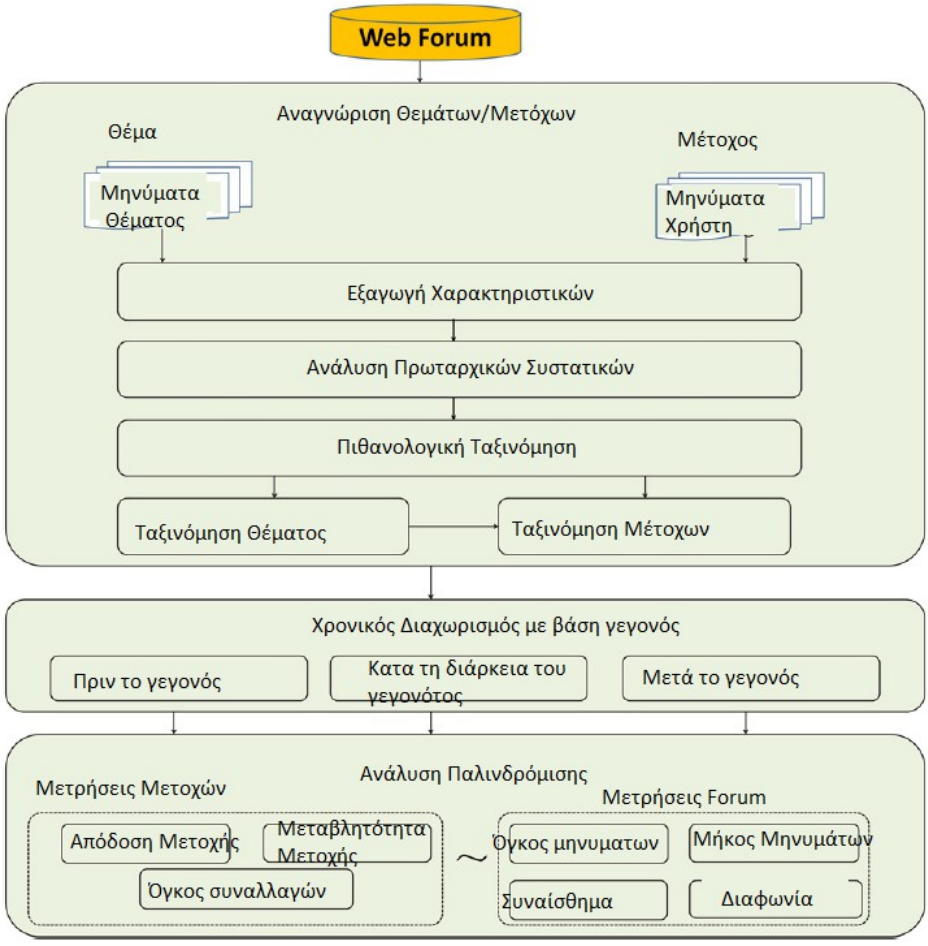
Analytics σε σχέση με τις μαθηματικές μεθόδους των τεχνικών προβλέψεων, όσο αφορά την συμπεριφορά του καταναλωτή. Δηλαδή, αντί τα μεγέθη που χαρακτηρίζουν τους καταναλωτές να αντιμετωπίζονται σαν απλές μαθηματικές μεταβλητές που εισάγονται σε ένα μαθηματικό μοντέλο που βγάζει προβλέψεις (όπως η τιμή μιας μετοχής), τα Predictive Analytics αναγνωρίζουν ότι αυτά τα μεγέθη προκύπτουν από τα χαρακτηριστικά των καταναλωτών, και επιχειρούν να τα συσχετίσουν αυτά τα χαρακτηριστικά με τις μεταβλητές που προβλέπονται. Είναι εύκολο να γίνει κατανοητό πως αυτή ίσως είναι κατά κανόνα μια πιο σωστή προσέγγιση στην ανθρώπινη συμπεριφορά γενικότερα, αφού αυτή δεν είναι κάποιο μαθηματικό ή φυσικό μέγεθος που ακολουθεί μαθηματικούς νόμους, αλλά μπορεί να προβλεφθεί μελετώντας τα ανθρώπινα χαρακτηριστικά ταυτόχρονα σε ατομικό επίπεδο (αξιοποιώντας τα δεδομένα κάθε παρελθοντικής περίπτωσης) και σε μαζικό (εξάγοντας προβλέψεις για μεγάλα πλήθη). Αντίστοιχα, στην πρόβλεψη πωλήσεων ενός προϊόντος, τα χαρακτηριστικά του προϊόντος θα πρέπει να ληφθούν υπ 'όψη. Αυτό βέβαια δεν σημαίνει ότι οι μέθοδοι των τεχνικών προβλέψεων δεν έχουν καμία χρήση για σκοπούς Predictive Analytics. Όπως περιγράφεται στο επόμενο κεφάλαιο, οι τεχνικές αυτές μπορούν να τροποποιηθούν ώστε να λαμβάνουν υπ' όψη τους τα χαρακτηριστικά των προϊόντων και των καταναλωτών.

Σχετικά με τις μεθόδους εντοπισμού κοινοτήτων, ανάλυσης επιρροής και πρόβλεψης συνδέσμων με βάση τον γράφο του αντίστοιχου κοινωνικού δικτύου, οι αλγόριθμοι που χρησιμοποιούνται φαίνονται να πάσχουν από κάποια πολύ βασικά προβλήματα. Το πρώτο πρόβλημα είναι το μεγάλο υπολογιστικό κόστος που έχουν αυτοί οι αλγόριθμοι στην πλειοψηφία τους [12 16 2 24 19 1]. Αυτό προκύπτει από το γεγονός ότι οι αλγόριθμοι εκτελούν πολλές πράξεις με πίνακες και διανύσματα, με αποτέλεσμα να έχουν πολύ υψηλή πολυπλοκότητα. Το ήδη αυξημένο υπολογιστικό κόστος, πολλαπλασιάζεται όταν αυτοί οι αλγόριθμοι καλούνται να εφαρμοστούν σε γράφους που προκύπτουν από τα μέσα κοινωνικής δικτύωσης, και μπορούν να περιέχουν εκατομμύρια ακμές και κόμβους, ενώ παρουσιάζουν πολύ δυναμική συμπεριφορά. Επιπλέον, οι περισσότεροι αλγόριθμοι υποθέτουν την ύπαρξη ενός ομογενούς δικτύου, στο οποίο οι κόμβοι και οι ακμές είναι του ίδιου τύπου. Στην πραγματικότητα όμως, συχνά έχουμε να αντιμετωπίσουμε ετερογενή κοινωνικά δίκτυα, όπου οι κόμβοι είναι διαφορετικού είδους, οι ακμές διαφορετικού τύπου (πχ σχέσεις που βασίζονται σε διαφορετικούς τρόπους επικοινωνίας), ή και τα δύο μαζί. Ας θεωρήσουμε το δίκτυο του IMDB (www.imdb.com), όπου οι οντότητες μπορεί να είναι πολλαπλού τύπου όπως ταινίες, ηθοποιοί, σκηνοθέτες, και οι σχέσεις μεταξύ τους μπορεί να είναι “βασικός ρόλος”, “σκηνοθετήθηκε από”, “ρόλοι στην ίδια ταινία”.Αυτή η ποικιλία αποτελεί πρόκληση αλλά και ευκαιρία μαζί, αφού το πιο πιθανό είναι να υπάρχει χρήσιμη πληροφορία σε αυτή τη

διαφοροποίηση του δικτύου, αλλά δεν είναι προφανές το πώς να πρέπει να διαχειριστούν οι διαφορετικοί κόμβοι και ακμές. Μια προσπάθεια γίνεται από τον αλγόριθμο SONAR API [3], που είναι ένας αλγόριθμος που σχεδιάστηκε με σκοπό να συναθροίζει τις πληροφορίες στα κοινωνικά δίκτυα από emails, μηνύματα, blogs και άλλα. Οι σχεδιαστές του πειραματίστηκαν με διαφορετικούς συνδυασμούς από αυτά τα είδη πληροφοριών, και με βάση το δίκτυο που προέκυψε, εφάρμοσαν μία καμπάνια προτάσεων αγοράς. Αναφέρθηκε ότι τα αποτελέσματα με βάση το συναθροισμένο δίκτυο ήταν καλύτερα από αυτά του κάθε δικτύου ξεχωριστά. Ωστόσο δεν προχώρησαν στο να βρουν ποιος συνδυασμός ειδών πληροφορίας είναι ο καλύτερος. Επίσης, οι αλγόριθμοι στην πλειοψηφία τους υποθέτουν μή κατευθυνόμενα δίκτυα. Ωστόσο, οι γράφοι που προκύπτουν από κάποια από τα σημαντικότερα κοινωνικά δίκτυα είναι από τη φύση τους κατευθυνόμενοι, όπως το δίκτυο των χρηστών του Twitter. Η αυθαίρετη παράβλεψη της κατεύθυνσης αναλύοντας τέτοια δίκτυα, αγνοεί την επιπρόσθετη πληροφορία στην κατεύθυνση, και μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα. Ενώ έχουν γίνει προσπάθειες προσαρμογής των αλγορίθμων σε κατευθυνόμενα δίκτυα [4 5 6], οι αλγόριθμοι που προκύπτουν είναι ακόμα πιο απαιτητικοί υπολογιστικά, επομένως η εφαρμογή τους στα μέσα κοινωνικής δικτύωσης πολλές φορές δεν είναι εφικτή. Τέλος, ενώ η πληροφορία συσχετισμού στα κοινωνικά δίκτυα έχει ερευνηθεί εκτενώς, η ιδέα της αξιοποίησης του περιεχομένου παράλληλα με τους συσχετισμούς για την ανάλυση του γράφου δεν έχει μελετηθεί ακόμα επαρκώς. Ας θεωρήσουμε ένα δίκτυο επικοινωνίας με email, όπου η επικοινωνία αποστολέα-παραλήπτη μπορεί να μοντελοποιηθεί σαν μια σχέση χρηστών. Σε αυτή την περίπτωση, ένας λογαριασμός spam θα έχει έναν μεγάλο αριθμό σχέσεων, και επομένως μπορεί να θεωρηθεί σαν ένας χρήστης με επιρροή, ή ένας χρήστης που είναι το κέντρο μιας κοινότητας κόμβων κάτι το οποίο στην πραγματικότητα είναι λανθασμένο. Από αυτό το απλό παράδειγμα φαίνεται η σημασία αξιοποίησης του περιεχομένου. Έχουν αναπτυχθεί μερικοί αλγόριθμοι που αποσκοπούν στην ενσωμάτωση του περιεχομένου, [8 9 7] αλλά και αυτοί έχουν το ίδιο πρόβλημα μη εφικτού υπολογιστικού κόστους. Συμπερασματικά, όσο αφορά την εφαρμογή στα μέσα κοινωνικής δικτύωσης, ίσως είναι καλύτερα οι σκοποί κατηγοριοποίησης που εξυπηρετεί ο εντοπισμός κοινοτήτων να προσεγγίζονται με την χρήση ταξινομητών, και η εφαρμογή του viral marketing την οποία συνήθως εξυπηρετούν η ανάλυση επιρροής και η πρόβλεψη συνδέσμων να βασίζεται στις πιο βασικές μετρήσεις που εξάγονται από τα μέσα κοινωνικής δικτύωσης και που περιγράφονται στο κεφάλαιο 3.1 . Συγκεκριμένα για την ανάλυση επιρροής στον γράφο, μια εναλλακτική προσέγγιση εφαρμόστηκε στην έρευνα [50], που περιγράφεται στο κεφάλαιο 4.3.

4.2 Αξιολόγηση Συνεισφοράς των Μέσων Κοινωνικής Δικτύωσης

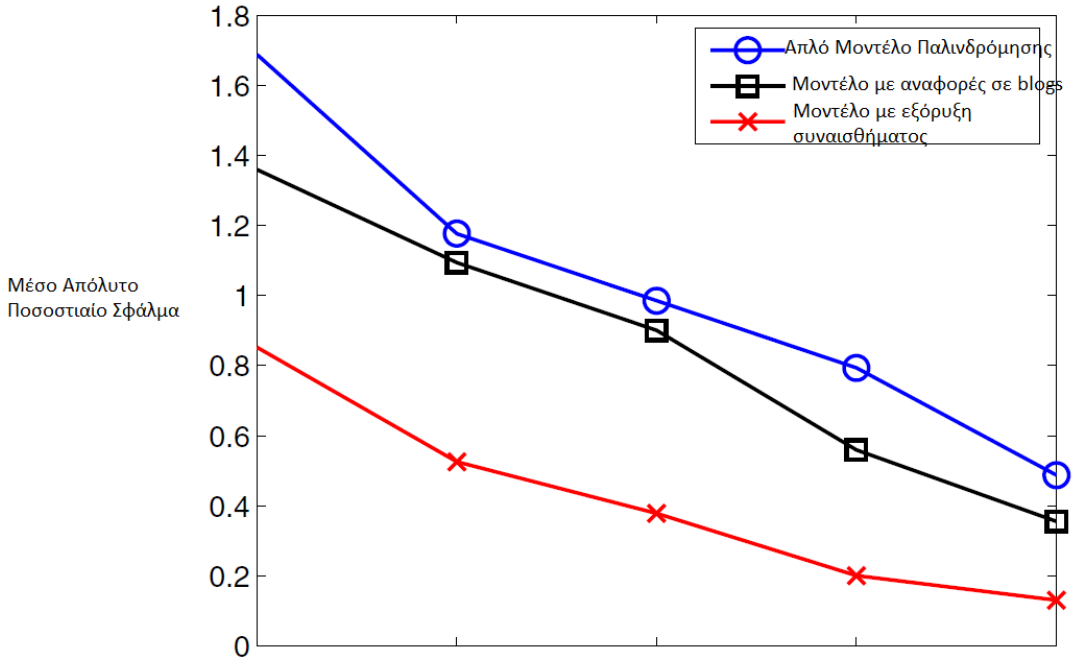
Συγκριτικό πλεονέκτημα φαίνεται να έχουν τα μοντέλα που αξιοποιούν την πληροφορία που υπάρχει στα μέσα κοινωνικής δικτύωσης. Συγκεκριμένα, μελετώντας τις έρευνες και τις εφαρμογές αποδεικνύεται ότι η ανάλυση κειμένου για σκοπούς ταξινόμησης και εξόρυξης συναισθήματος, έχει πολύ θετική επίδραση στην προβλεπτική ικανότητα του μοντέλου. Στην πλειοψηφία τους οι εφαρμογές αξιοποιούν κυρίως τα δεδομένα από το Twitter, και από blogs και forums σχετικά με τον τομέα της πρόβλεψης. Αξίζει να αναφερθεί ότι κάποιες από αυτές τις έρευνες κατάφεραν να συσχετίσουν αποδοτικά μετρήσεις από τα μέσα κοινωνικής δικτύωσης με πραγματικούς οικονομικούς δείκτες όπως απόδοση μετοχών και όγκο συναλλαγών. Η μοντελοποίηση μιας από αυτές τις έρευνες φαίνεται στο σχήμα [33]:



Παρατηρούμε ότι οι αναλυτές βασίζονται απόλυτα στα δεδομένα από forums με θέμα το χρηματιστήριο. Τα θέματα και οι χρήστες του forum ταξινομούνται με βάση τα χαρακτηριστικά κειμένου που αναφέρονται στο κεφάλαιο 3.3.1., αφού πρώτα τα στοιχεία που έχουν συλλεχθεί υποβάλλονται σε ανάλυση πρωταρχικών συστατικών. Στα πλαίσια αυτής της εφαρμογής "μέτοχος" ορίζεται ως οποιοσδήποτε φαίνεται να ενδιαφέρεται για την μετοχή μιας επιχείρησης. Οι χρήστες ταξινομούνται σε ομάδες ανάλογα με το είδος της σχέσης που προκύπτει ότι έχουν με την επιχείρηση, και τα θέματα με ανάλογα με το περιεχόμενο τους. Γίνεται και ένας χρονικός διαχωρισμός των δεδομένων με βάση την χρονολογία εμφάνισης της πρώτης δημοσίευσης που αφορά ένα γεγονός που επηρεάζει το χρηματιστήριο. Τελικά τα μεγέθη που αφορούν τις μετοχές αναλύονται παλινδρομικά σε σχέση με τις μετρήσεις που έχουν προκύψει από την ανάλυση κειμένου. Τελικά εξάγονται σημαντικά συμπεράσματα για το ποιες ομάδες χρηστών και ποιο περιεχόμενο θέματος εμφανίζει την καλύτερη προβλεπτική συμπεριφορά την κάθε χρονική περίοδο. Επιπλέον συγκρίνεται η απόδοση του μοντέλου σε σχέση με συμβατικές μεθόδους και αποδεικνύεται η υπεροχή της. Από αυτή τη μοντελοποίηση, φαίνεται επιπλέον ότι όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, οι

μαθηματικές μέθοδοι τεχνικών προβλέψεων μπορούν να ικανοποιήσουν σκοπούς Predictive Analytics δεδομένου ότι θα λειτουργούν με βάση χαρακτηριστικά που ενδιαφέρουν. Μια παρόμοια έρευνα που συνδυάζει ανάλυση συναισθήματος και τεχνικές προβλέψεων για σκοπούς χρηματιστηριακών προβλέψεων είναι αυτή της οποίας η μοντελοποίηση φαίνεται στο σχήμα του κεφαλαίου 3.1 [30]. Και σε αυτή την έρευνα τα αποτελέσματα του μοντέλου συγκρίνονται με μια απλή μέθοδο παλινδρόμησης, και αποδεικνύεται ότι το μοντέλο είναι πιο αποδοτικό.

Άλλη μια έρευνα που αποδεικνύει την προβλεπτική ικανότητα που προκύπτει από την ανάλυση κειμένου στα μέσα κοινωνικής δικτύωσης είναι η [38], που αποσκοπεί στο να συνδέσει τα έσοδα από την προβολή μιας ταινίας στους κινηματογράφους με βάση τις συζητήσεις που την αφορούν στα blogs. Αφού εφαρμοστεί εξόρυξη συναισθήματος στα κείμενα, οι μεταβλητές που εκφράζουν τα διάφορα συναισθήματα που έχουν προκύψει εισάγονται ως παράμετροι σε ένα παλινδρομικό μοντέλο. Τα αποτελέσματα του μοντέλου συγκρίνονται με ένα μοντέλο απλής παλινδρόμησης, και ένα μοντέλο παλινδρόμησης που λαμβάνει υπ όψη του και τις απλές αναφορές της ταινίας στα blogs. Όπως φαίνεται και στο γράφημα, το μοντέλο έχει πολύ μικρότερο σφάλμα σε σχέση με τα άλλα δύο:



Η μεταβλητή p του άξονα x είναι μια μεταβλητή παραμετροποίησης του παλινδρομικού μοντέλου. Αξίζει να σημειωθεί ότι αυτή η μελέτη δείχνει το κέρδος που προκύπτει από την ανάλυση συναισθήματος όχι μόνο σε σχέση με ένα μοντέλο που δεν αξιοποιεί καθόλου τα δεδομένα από τα μέσα κοινωνικής δικτύωσης, αλλά και σε σχέση με ένα μοντέλο που τα αξιοποιεί σχετικά επιπόλαια, υπολογίζοντας δηλαδή μόνο τις αναφορές της ταινίας σε δημοσιεύσεις blog, χωρίς να αναλύει το τι εκφράζει η δημοσίευση στο σύνολό της.

Μια παρόμοια έρευνα με ίδιο σκοπό αλλά που βασίζεται στα δεδομένα του Twitter και όχι σε δεδομένα από forums είναι η [35]. Συγκεκριμένα οι ερευνητές προσπάθησαν να συνδέσουν τις δημοσιεύσεις στο twitter που αναφέρουν την ταινία και το συναίσθημα που εκφράζουν, με τα έσοδα της ταινίας το πρώτο Σαββατοκύριακο μετά την προβολή της. Οι δημοσιεύσεις του twitter ταξινομήθηκαν στις κατηγορίες θετική, αρνητική, ουδέτερη, και με βάση αυτή την κατηγοριοποίηση εξάχθηκαν κάποιες μεταβλητές σχετικές με τα ποσοστά τους. Οι μεταβλητές των συναισθημάτων ενσωματώθηκαν και πάλι σε ένα μοντέλο γραμμικής παλινδρόμησης. Το αξιοσημείωτο σε αυτή την έρευνα είναι ότι εκτός από την σύγκριση των αποτελεσμάτων με το Χρηματιστήριο του Hollywood (www.hsx.com) όπου οι τιμές των μετοχών των ταινιών θεωρούνται προβλεπτικές των εσόδων τους, έγινε και σύγκριση με το μοντέλο μιας παρελθοντικής έρευνας [39] η οποία πρόβλεπε τα έσοδα των ταινιών με βάση τις κριτικές στο imdb (www.imdb.com) και τις ειδήσεις. Το μοντέλο που βασιζόταν στις δημοσιεύσεις του Twitter είχε καλύτερη απόδοση και από τις δύο άλλες μεθόδους προβλέψεων. Αυτό ίσως αποδεικνύει ότι το Twitter είναι καλύτερο σαν μέσο έκφρασης της μαζικής άποψης του καταναλωτικού κοινού σε σχέση με τις ειδήσεις και το imdb.

Είναι λοιπόν εμφανές, ότι τα δεδομένα των μέσων κοινωνικής δικτύωσης και ειδικότερα αυτά που προκύπτουν από την ανάλυση κειμένου και συναισθήματος μπορούν να παρουσιάσουν ακριβή προβλεπτική συμπεριφορά. Κάθε μοντέλο του οποίου η εφαρμογή σχετίζεται με την συμπεριφορά του καταναλωτικού κοινού θα πρέπει να ενσωματώνει την ανάλυση κειμένου και την εξόρυξη συναισθήματος στην λειτουργία του, αφού είναι πλέον αποδεδειγμένο ότι η πληροφορία που κυκλοφορεί στα μέσα κοινωνικής δικτύωσης είναι προβλεπτική της μαζικής καταναλωτικής συμπεριφοράς σε μεγάλο βαθμό.

4.3 Αξιοποίηση στη Σχεδίαση Προϊόντων και Υπηρεσιών και Μοντελοποίηση

4.3.1 Εφαρμογές στη Σχεδίαση

Όπως είδαμε στο κεφάλαιο 2.2, η αποτελεσματική σχεδίαση ενός προϊόντος ή μιας υπηρεσίας εξαρτάται από δύο βασικές μεταβλητές. Αυτές οι δύο μεταβλητές είναι οι ανάγκες των καταναλωτών, και τα τεχνικά χαρακτηριστικά του προϊόντος ή της υπηρεσίας. Επομένως, στο στάδιο της σχεδίασης, είναι απαραίτητο να κατανοηθούν με όσο το δυνατό μεγαλύτερη ακρίβεια οι εξής τέσσερις συσχετισμοί:

- **Η σχέση μεταξύ των τεχνικών χαρακτηριστικών και τις ανάγκες που ικανοποιούνται**
- **Η σχέση που υπάρχει μεταξύ των αναγκών των καταναλωτών**
- **Η σχέση που υπάρχει μεταξύ των τεχνικών χαρακτηριστικών**
- **Η σχέση που υπάρχει μεταξύ των τεχνικών χαρακτηριστικών και τυχαίων επιπρόσθετων στόχων της επιχείρησης**

Το βασικό ερώτημα που θα καλείται να λύσει ένα προβλεπτικό μοντέλο στην σχεδίαση σε πρώτο στάδιο μπορεί να είναι το εξής:

- **Πως πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά του προϊόντος ή της υπηρεσίας ώστε να ικανοποιούν τις ανάγκες των πελατών;**

Αναγνωρίζοντας τον αρνητικό συσχετισμό που μπορεί να υπάρχει μεταξύ των αναγκών των καταναλωτών, με την έννοια ότι η ικανοποίηση μίας ανάγκης μπορεί να αγνοεί μια άλλη ανάγκη, το ερώτημα μπορεί να διαμορφωθεί ως:

- **Πως πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά του προϊόντος ή της υπηρεσίας ώστε να ικανοποιούν τις <u>συγκεκριμένες</u> ανάγκες των πελατών που θέλει να ικανοποιήσει η επιχείρηση;**

Λαμβάνοντας υπ' όψη και το γεγονός ότι υπάρχουν ομάδες καταναλωτών που είναι πιο επικερδείς από άλλες, και ένα προϊόν ή υπηρεσία τις περισσότερες φορές δεν είναι δυνατόν να σχεδιαστεί ώστε να ικανοποιεί όλες τις ομάδες, το ερώτημα μπορεί να διαμορφωθεί περαιτέρω:

- **Πως πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά του προϊόντος ή της υπηρεσίας ώστε να ικανοποιούν τις <u>συγκεκριμένες</u> ανάγκες των <u>συγκεκριμένων</u> πελατών που θέλει να ικανοποιήσει η επιχείρηση;**

Επιπλέον, καθώς η επιχείρηση εκτός από ικανοποίηση αναγκών των πελατών μπορεί να έχει και άλλους διαφορετικούς σκοπούς, υπάρχει και το πιο γενικό ερώτημα:

- **Πως πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά του προϊόντος ή της υπηρεσίας ώστε να ικανοποιούν τον συγκεκριμένο στόχο της επιχείρησης;**

Ας επιστρέψουμε στις βασικές εφαρμογές των Predictive Analytics που περιγράφονται στο κεφάλαιο 2.1 και ας δούμε ποιες από αυτές μπορούν να προσαρμοστούν στο στάδιο της σχεδίασης των προϊόντων και υπηρεσιών.

Τμηματοποίηση Καταναλωτών: Στο στάδιο του σχεδιασμού, η τμηματοποίηση των καταναλωτών με βάση τις ανάγκες τους πρέπει να αποτελεί αρχικό στάδιο, αφού όπως προαναφέρθηκε στις περισσότερες περιπτώσεις ένα προϊόν η υπηρεσία δεν μπορεί να καλύπτει όλες τις ανάγκες των καταναλωτών. Επομένως η αγορά πρέπει να τμηματοποιηθεί, ώστε να αποφασίσει η επιχείρηση για τα τεχνικά χαρακτηριστικά που θα ικανοποιούν τους πελάτες που θα στοχοποιηθούν. Όπως και στην εφαρμογή στην στοχευμένη προώθηση ή την εξυπηρέτηση πελατών, τα δεδομένα που έχει η επιχείρηση για κάποιους πελάτες είναι πιθανόν να μην είναι αρκετά για να τους εντάξει σε τμήματα, επομένως το πρόβλημα λύνεται με την χρήση κάποιου προβλεπτικού μοντέλου.

Αξιολόγηση Πίστωσης: Μια επιχείρηση που βρίσκεται στο στάδιο σχεδίασης μιας υπηρεσίας κατανοεί ότι δεδομένου ότι όλοι οι πελάτες δεν μπορούν να ικανοποιηθούν με την ίδια υπηρεσία, θα πρέπει να στοχοποιηθούν οι πελάτες που είναι πιο επικερδείς για την επιχείρηση. Επομένως μια αξιολόγηση πίστωσης θα βοηθάει στην απάντηση του ερωτήματος "Πώς πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά της υπηρεσίας ώστε να ικανοποιούν τις ανάγκες των επικερδών πελατών;"

Συγκράτηση πελατών με μοντελοποίηση αποχώρησης : Στο στάδιο της σχεδίασης, η μοντελοποίηση αποχώρησης θα μπορούσε να έχει αντίστροφο ρόλο. Ενώ στην εφαρμογή στην προώθηση αξιολογούνται οι πελάτες με βάση την πιθανότητα αποχώρησής τους από την εταιρία, στην σχεδίαση θα μπορούσαν να μελετηθούν οι παρελθοντικές υπηρεσίες της εταιρίας, και να συσχετιστούν τα τεχνικά χαρακτηριστικά τους με τον βαθμό αποχώρησης πελατών. Άρα μια μοντελοποίηση αποχώρησης θα απαντούσε στο ερώτημα "Πως πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά της υπηρεσίας ώστε να υπάρχει η χαμηλότερη δυνατή αποχώρηση πελατών;"

Viral Marketing: Ενώ η προώθηση δεν αποτελεί μέρος της σχεδίασης, η "viral" εξάπλωση της αναγνωρισιμότητας του προϊόντος ή της υπηρεσίας μπορεί να αποτελεί έναν επιπρόσθετο στόχο της επιχείρησης. Φυσικά στο στάδιο της σχεδίασης, η εφαρμογή αυτή δεν θα ασχολείται με την ανάλυση επιρροής των καταναλωτών, αλλά θα προσπαθεί να συσχετίσει τα τεχνικά χαρακτηριστικά παρελθοντικών προϊόντων ή υπηρεσιών στα οποία εφαρμόστηκαν τεχνικές Viral Marketing με την πιθανότητα επιτυχίας αυτών των τεχνικών. Δηλαδή θα έχει ως σκοπό την απάντηση στο ερώτημα "Πως πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά του προϊόντος ή της υπηρεσίας ώστε να γίνει "viral";"

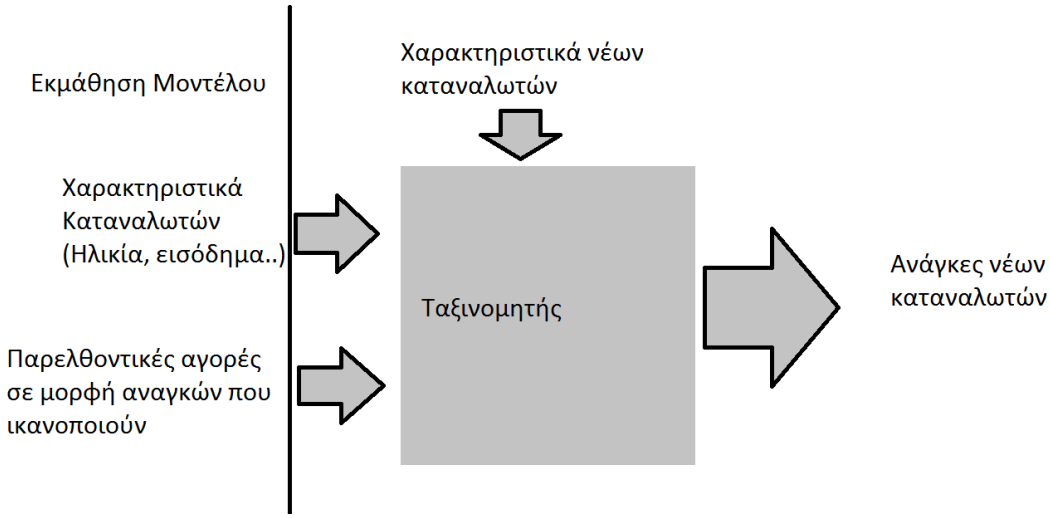
4.3.2 Υπάρχουσα και Προτεινόμενη Μοντελοποίηση

Πρόβλεψη Αναγκών Καταναλωτών: Από την ανάλυση των εφαρμογών του προηγούμενου κεφαλαίου, προκύπτει ότι τα δεδομένα που πρέπει να έχει στη διάθεση της μια επιχείρηση για να δημιουργήσει ένα προβλεπτικό μοντέλο είναι:

- **Παρελθοντικά Δεδομένα Πωλήσεων**
- **Τεχνικά Χαρακτηριστικά Προϊόντων και Υπηρεσιών**
- **Ανάγκες και Χαρακτηριστικά των Καταναλωτών**

Είναι σαφές ότι οι πρώτες δύο ομάδες δεδομένων είναι διαθέσιμες σε κάθε επιχείρηση. Η πρόκληση παρουσιάζεται στην τρίτη ομάδα. Συγκεκριμένα, η επιχείρηση θα πρέπει να βρει

ένα τρόπο να συσχετίσει τα χαρακτηριστικά των καταναλωτών που έχει στη διάθεσή της, με τις ανάγκες τους. Ένας πρώτος συσχετισμός μεταξύ αυτών των δύο δεδομένων μπορεί να προκύψει από τις παρελθοντικές αγορές του καταναλωτή. Μελετώντας τα προϊόντα που αγοράζει ο καταναλωτής, και συγκεκριμένα τις ανάγκες που έχουν σχεδιαστεί για να ικανοποιούν αυτά τα προϊόντα (πχ μεγάλη οθόνη κινητού, μεγάλη διάρκεια μπαταρίας, κομψή εμφάνιση), μπορούν να εξαχθούν συσχετισμοί μεταξύ των χαρακτηριστικών του καταναλωτή και των αναγκών του. Επομένως ένα πολύ βασικό μοντέλο που μπορεί να δημιουργηθεί είναι:

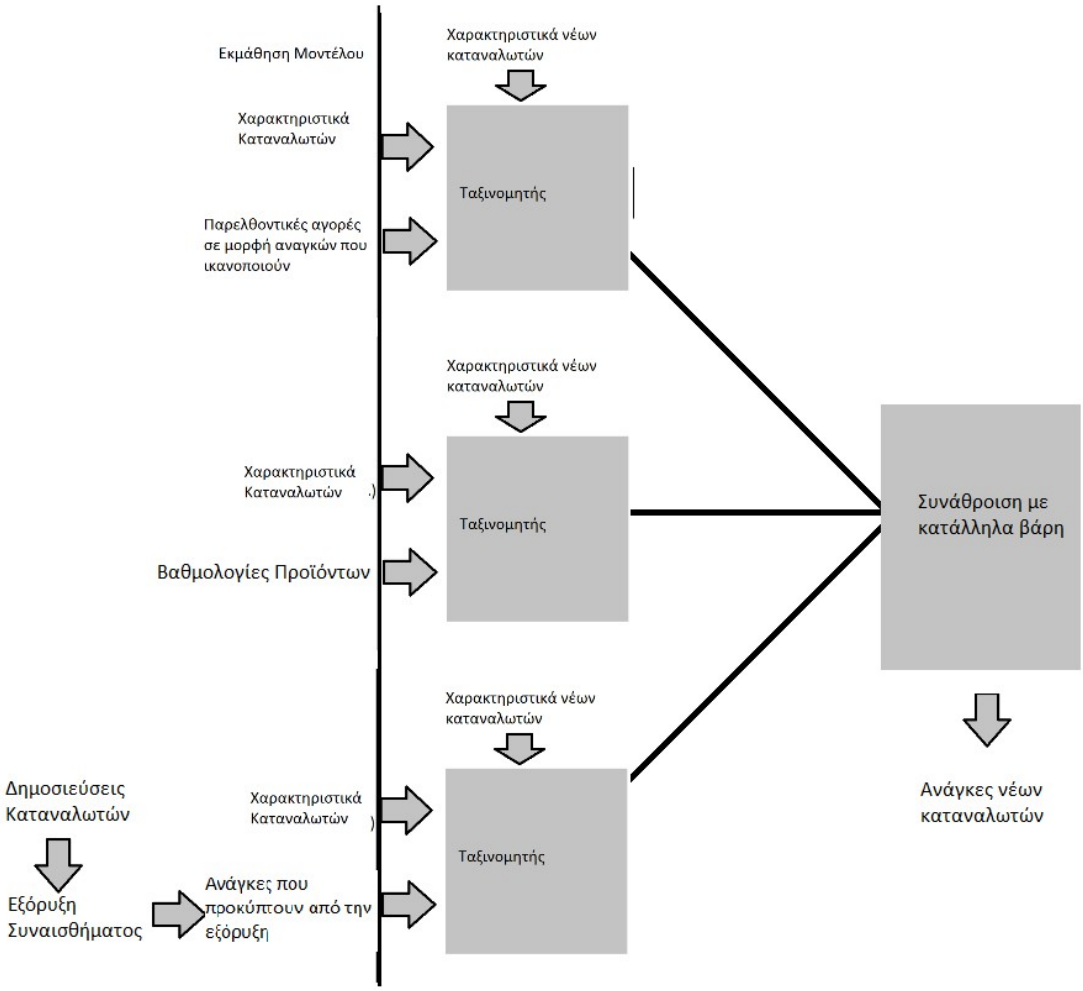


Ένα βασικό μειονέκτημα αυτού του μοντέλου είναι ότι θεωρεί δεδομένο ότι οι ανάγκες του καταναλωτή ικανοποιήθηκαν από το προϊόν που αγοράστηκε. Καθώς αξιοποιείται μόνο η πληροφορία της αγοράς, και όχι κάποια πληροφορία που σχετίζεται με την κριτική του καταναλωτή ως προς το προϊόν, οι προβλέψεις θα είναι χαμηλής ποιότητας.

Πως όμως θα αξιοποιηθούν τα δεδομένα από τα μέσα κοινωνικής δικτύωσης που έχει αποδειχθεί ότι έχουν τόσο μεγάλη προβλεπτική ικανότητα; Τι είδους μελέτη πρέπει να εφαρμοστεί στις δημοσιεύσεις των καταναλωτών ώστε να εξαχθούν συμπεράσματα για τις ανάγκες τους; Σε αυτό το σημείο ανακαλούμε ότι όπως αναφέρεται στο κεφάλαιο 3.3.1, οι σύγχρονες ιστοσελίδες ηλεκτρονικού εμπορίου μπορούν να ενταχθούν στην κατηγορία των μέσων κοινωνικής δικτύωσης, αφού πλέον στην πλειονότητά τους απαιτούν την εγγραφή των χρηστών (άρα παρέχουν βασικές πληροφορίες όπως ηλικία, φύλο και τοποθεσία), και επιτρέπουν τον σχολιασμό, την βαθμολόγηση και τις προτάσεις προϊόντων. Επομένως αποτελούν ιστότοπους στους οποίους οι χρήστες ανταλλάζουν απόψεις, και έχουν το πλεονέκτημα ότι όλες οι δημοσιεύσεις αφορούν προϊόντα και υπηρεσίες. Αμέσως προκύπτει

μια επίσης απλή εφαρμογή, η οποία βασίζεται στις βαθμολογίες των χρηστών στα προϊόντα. Πρόκειται για έναν ταξινομητή όπως στο προηγούμενο παράδειγμα, με την διαφορά ότι αντί να δέχεται σαν είσοδο τις παρελθοντικές αγορές, θα δέχεται σαν είσοδο τις κριτικές του καταναλωτή, πάλι σε μορφή αναγκών που ικανοποιούν τα προϊόντα που βαθμολογεί. Πέρα από τις βαθμολογίες των προϊόντων, απαιτείται και μια διαφορετική προσέγγιση αφού γενικότερα ένα πολύ μικρό ποσοστό των καταναλωτών τείνουν να βαθμολογούν και να γράφουν κριτικές προϊόντων. Ένα τρίτο μοντέλο μπορεί να δημιουργηθεί εφαρμόζοντας εξόρυξη συναισθήματος στις δημοσιεύσεις που έχει μια επιχείρηση στη διάθεσή της, και που αναφέρουν προϊόντα. Αυτή η μελέτη θα γίνεται με την παραδοχή ότι δημοσιεύσεις που εκφράζουν αρνητικά συναισθήματα για το προϊόν, εκφράζουν την απογοήτευση του συγγραφέα για το προϊόν, και αντίστοιχα οι δημοσιεύσεις με θετικά συναισθήματα την ικανοποίηση του. Επομένως αναγνωρίζονται οι ανάγκες του καταναλωτή ανάλογα με το ποιο προϊόν τις ικανοποιεί. Σε αυτό το σημείο πρέπει να αναφερθεί ότι δεδομένης της σημερινής φύσης των μέσων κοινωνικής δικτύωσης, η εξόρυξη συναισθήματος είναι λιγότερο απαιτητική διαδικασία από ότι ήταν πριν από λίγα χρόνια και έχουν αναπτυχθεί πολλά μοντέλα για την εφαρμογή της. Αυτό συμβαίνει επειδή τα περισσότερα μέσα κοινωνικής δικτύωσης πλέον δίνουν την ικανότητα στον χρήστη να βάλει μια ταμπέλα (tag) συναισθήματος στην δημοσίευση του. Επομένως η εξόρυξη συναισθήματος για ένα μεγάλο ποσοστό δημοσιεύσεων αποτελεί μια απλούστατη διαδικασία, ενώ παράλληλα έχουν αναπτυχθεί αλγόριθμοι που συσχετίζουν τις λέξεις που χρησιμοποιούνται σε τέτοιες δημοσιεύσεις με τα συναισθήματα των ταμπελών, ώστε να αναγνωρίζονται τα συναισθήματα σε δημοσιεύσεις που δεν φέρνουν ταμπέλες συναισθήματος, με βάση τις λέξεις που χρησιμοποιούνται [41].

Εφαρμόζοντας την συνεργατική μοντελοποίηση, μπορεί να προκύψει το εξής μοντέλο:



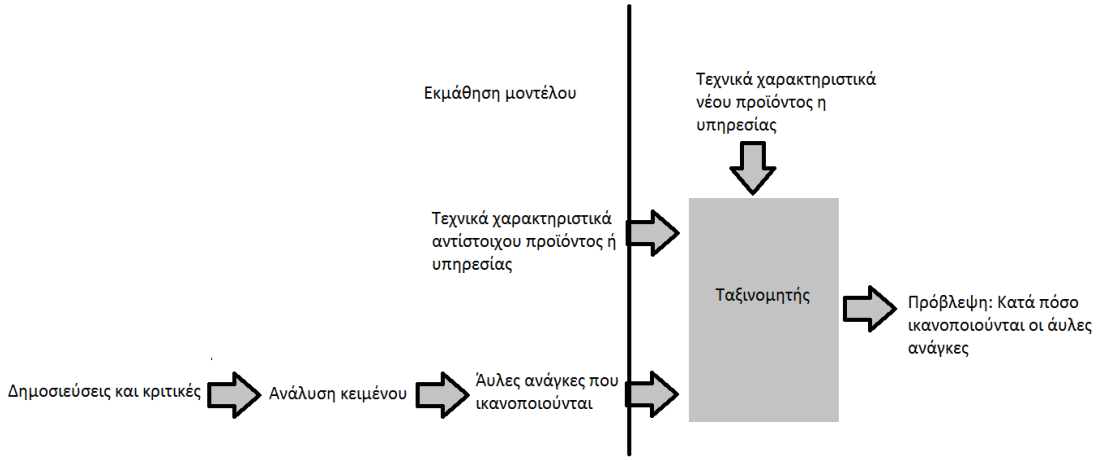
Συνδυάζοντας τα τρία αυτά βασικά μοντέλα, προκύπτει τελικά ένα μοντέλο που μπορεί να θεωρηθεί αρκετά πλήρες, αφού μπορεί να αξιοποιήσει τρεις διαφορετικές μορφές δεδομένων, και επιπλέον ο αναλυτής μπορεί να προσαρμόσει τις προβλέψεις ώστε να δίνουν περισσότερο βάρος στα δεδομένα που θεωρεί πιο αξιόπιστα. Βαρύτητα θα πρέπει να δοθεί κυρίως στις προβλέψεις που βασίζονται στις βαθμολογίες και την εξόρυξη συναισθήματος. Το μοντέλο που βασίζεται στις παρελθοντικές αγορές θα πρέπει να αξιοποιείται μόνο όταν δεν υπάρχουν αρκετά δεδομένα για τα άλλα δύο μοντέλα, αφού όπως προαναφέρθηκε θεωρεί ότι οι ανάγκες του καταναλωτή ικανοποιήθηκαν από την αγορά του προϊόντος.

Εύκολα διαπιστώνεται η χρήση που θα μπορούσε να έχει αυτό το μοντέλο σε μερικές από τις εφαρμογές που περιγράφτηκαν. Οι καταναλωτές πλέον μπορούν να τμηματοποιηθούν με βάση τις ανάγκες τους αξιοποιώντας τα χαρακτηριστικά που είναι διαθέσιμα στην

επιχείρηση για αυτούς. Επιπλέον, οι ανάγκες των πελατών με υψηλή αξιολόγηση πίστωσης μπορούν να αναγνωριστούν, αφού οι πελάτες αυτοί έχουν πρώτα αναγνωριστεί σε διαφορετικό στάδιο.

Πρόβλεψη επίδρασης τεχνικών χαρακτηριστικών στην ικανοποίηση άυλων αναγκών: Στα πλαίσια αυτού το προβλήματος, θα διαχωρίσουμε τις ανάγκες των καταναλωτών σε υλικές και άυλες, σύμφωνα με την έρευνα [47]. Όπως περιγράφεται και στο κεφάλαιο 2.2, ενώ η επίδραση των τεχνικών χαρακτηριστικών μπορεί να είναι προφανής όσο αφορά τις υλικές ανάγκες, η ίδια επίδραση είναι πιο δύσκολο να εξαχθεί για τις άυλες ανάγκες. Δηλαδή, δεν μπορεί να είναι προφανές κατά πόσο οι διαστάσεις ενός φορητού υπολογιστή ικανοποιούν την ανάγκη που μπορεί να έχει ένας καταναλωτής για τον φορητό υπολογιστή του να θεωρείται "πολυτελές". Χαρακτηριστικό παράδειγμα είναι η αγορά ειδών ενδυμασίας, όπου πλέον οι καταναλωτές σίγουρα αγοράζουν περισσότερο με βάση τις άυλες ανάγκες, παρά με την ανθεκτικότητα ή την αντοχή στο κρύο. Πέρα από αυτή την αγορά όμως, παρατηρείται και ένα παρόμοιο φαινόμενο στην αγορά κινητών τηλεφώνων ή φορητών υπολογιστών, στην οποία οι καταναλωτές φαίνονται να προτιμάνε προϊόντα από αναγνωρισμένες εταιρίες, παρ' όλο που υπάρχουν προϊόντα με καλύτερα τεχνικά χαρακτηριστικά διαφορετικών εταιριών. Ξεκαθαρίζεται επομένως ότι ένα τέτοιο μοντέλο, που θα απαντάει δηλαδή στο ερώτημα "Πως πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά του προϊόντος ή της υπηρεσίας ώστε να ικανοποιούν τις άυλες ανάγκες των καταναλωτών" μπορεί να είναι πολύ χρήσιμο για μια επιχείρηση. Για άλλη μια φορά θα αξιοποιηθούν τα δεδομένα από τα μέσα κοινωνικής δικτύωσης. Σε αντίθεση με τις υλικές ανάγκες, οι άυλες ανάγκες των καταναλωτών δεν μπορούν να προκύψουν από τις παρελθοντικές αγορές, αφού οι άυλες ανάγκες δεν μεταφράζονται ξεκάθαρα στα τεχνικά χαρακτηριστικά των προϊόντων που αγοράστηκαν. Με άλλα λόγια, για έναν καταναλωτή που αγόρασε ένα κινητό τηλέφωνο που είχε μεγάλη διάρκεια μπαταρίας, μπορούμε να συμπεράνουμε ότι μια υλική ανάγκη του είναι η μεγάλη διάρκεια μπαταρίας, όμως ένας επιπλέον λόγος για τον οποίο έγινε η αγορά μπορεί να είναι ότι το κινητό τηλέφωνο θεωρήθηκε από τον καταναλωτή ως "διαχρονικό" ή "φουτουριστικό", που και τα δύο αυτά στοιχεία δεν φαίνονται κάπως στα τεχνικά χαρακτηριστικά της συσκευής. Με την ίδια λογική δεν μπορούν να αξιοποιηθούν οι αριθμητικές βαθμολογίες. Επομένως μια προσέγγιση θα ήταν η ανάλυση κειμένου σε αναλυτικές κριτικές προϊόντων και υπηρεσιών και στις δημοσιεύσεις που τα αναφέρουν. Συγκεκριμένα, ψάχνοντας για "λέξεις συναισθημάτων" που ενδιαφέρουν (που συνδέονται δηλαδή με άυλες ανάγκες) σε αυτά τα κείμενα, θα μπορούν να εξαχθούν συμπεράσματα για τον συσχετισμό των τεχνικών χαρακτηριστικών και την

ικανοποίηση αυτών των άυλων αναγκών. Για παράδειγμα, αν σε ένα μεγάλο ποσοστό των κριτικών και των δημοσιεύσεων που αναφέρεται το προϊόν υπάρχει το επίθετο "εντυπωσιακό" η άλλα επίθετα που εκφράζουν παρόμοια σημασία, μπορεί να εξαχθεί ένας θετικός συσχετισμός ανάμεσα στα τεχνικά χαρακτηριστικά του προϊόντος και την ικανότητα του προϊόντος να ικανοποιεί την ανάγκη να θεωρείται εντυπωσιακό. Επιπλέον, η ανάλυση κειμένου μπορεί εφαρμοστεί ώστε να συσχετίζει αρνητικά τα χαρακτηριστικά του προϊόντος με αυτή την ανάγκη σε περίπτωση που στα κείμενα υπάρχουν επίθετα όπως "ανιαρό":



Επομένως, για άλλη μια φορά αξιοποιούνται οι αφιλτράριστες απόψεις του καταναλωτικού κοινού για παρελθοντικά προϊόντα και υπηρεσίες, ώστε να εξαχθούν προβλέψεις για τα προϊόντα και τις υπηρεσίες που σχεδιάζονται. Οι συσχετισμοί που θα εξαχθούν από ένα τέτοιο μοντέλο μπορεί να μην έχουν άμεση λογική εξήγηση, για παράδειγμα μπορεί να βρεθεί συσχετισμός ανάμεσα στην χωρητικότητα σκληρού δίσκου ενός φορητού υπολογιστή με το κατά πόσο ικανοποιεί την ανάγκη να θεωρείται "εντυπωσιακός". Πρέπει όμως να γίνει κατανοητό ότι όπως έχει προαναφερθεί, στις εφαρμογές των Predictive Analytics δεν μας ενδιαφέρει να εξηγήσουμε το "γιατί" συνδέονται αυτές οι δύο έννοιες, αλλά μας ενδιαφέρει να εντοπίσουμε τον συσχετισμό και να τον αξιοποιήσουμε προς όφελός μας, για οποιοδήποτε λόγο και να υπάρχει αυτός.

Παρατηρούμε ότι αυτή η προσέγγιση παρουσιάζει ομοιότητες με την προσέγγιση έρευνας αγοράς που περιγράφεται στο κεφάλαιο 2.2 [47]. Ειδικότερα, αν η ανάλυση κειμένου εφαρμοστεί ώστε να λαμβάνει υπ' όψη της και το πόσο "έντονες" περιγραφικές λέξεις υπάρχουν (πχ "απίστευτα κομψό"), τελικά για κάθε κριτική θα προκύψει ή ίδια αντιστοίχιση αριθμητικής αξιολόγησης με "λέξη συναισθήματος". Σε αυτό το σημείο φαίνεται για άλλη

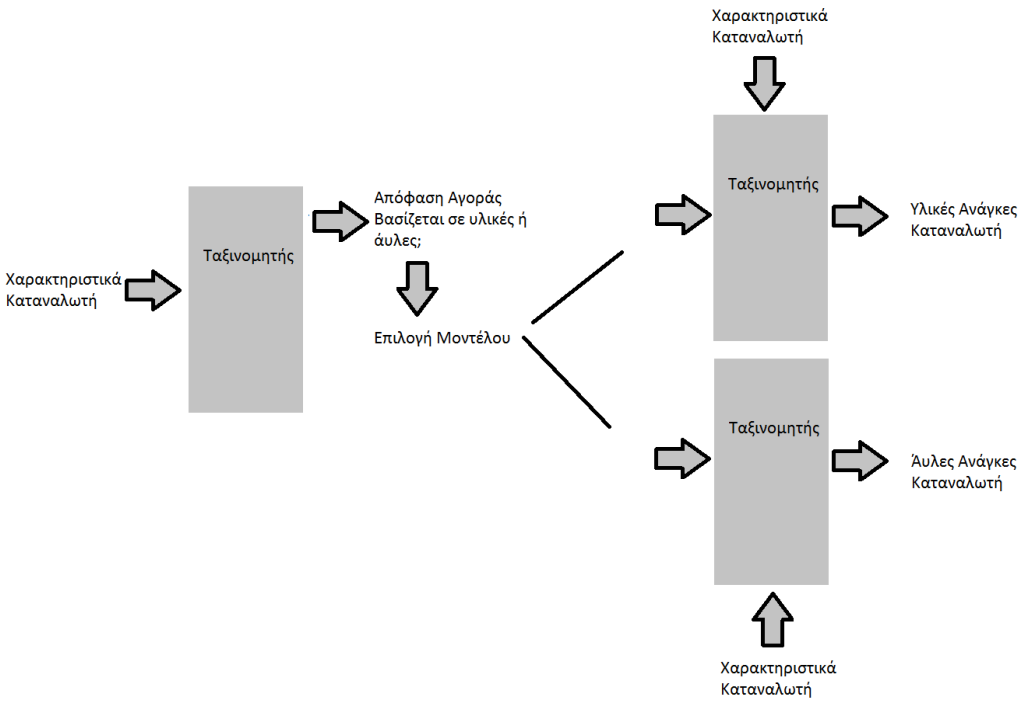
μια φορά πως η αξιοποίηση των δεδομένων των μέσων κοινωνικής δικτύωσης, παρουσιάζει πλεονεκτήματα. Συγκεκριμένα, όπως περιγράφεται και στο κεφάλαιο 2.2, το πρώτο πρόβλημα των αξιολογήσεων, είναι ότι κάποιες από τις άυλες ανάγκες δεν μπορούν να ικανοποιούνται από τα ίδια τεχνικά χαρακτηριστικά για όλους τους καταναλωτές. Δηλαδή, ένα συγκεκριμένο χρώμα μπορεί να θεωρείται εντυπωσιακό από μια μερίδα καταναλωτών και ανιαρό από μια άλλη. Όμως αυτό το πρόβλημα ομαλοποιείται από το μοντέλο, αφού βασίζεται στην μαζική άποψη του καταναλωτικού κοινού, και όχι σε κάποιο μικρό ποσοστό μιας έρευνας αγοράς. Επομένως αν ένα χρώμα θεωρείται εντυπωσιακό από το 70% των καταναλωτών, και ανιαρό από το 30% των καταναλωτών, στα ίδια ποσοστά θα κυμαίνονται και οι δημοσιεύσεις και οι κριτικές, επομένως η αξιολόγηση που θα κάνει το μοντέλο για το κατά πόσο το συγκεκριμένο χρώμα ικανοποιεί την ανάγκη "εντυπωσιακό" θα είναι αντιπροσωπευτική των ποσοστών προτιμήσεων. Φαίνεται επομένως πως η μοντελοποίηση βασισμένη στα δεδομένα από τα μέσα κοινωνικής δικτύωσης, έχει το πλεονέκτημα ότι είναι βασισμένη στην άποψη του καταναλωτικού κοινού σε μαζικό επίπεδο και όχι σε εξωτερικούς παράγοντες όπως ένα μικρό μερίδιο έρευνας αγοράς το οποίο μπορεί να είναι προκατειλημμένο.

Όσο αφορά το δεύτερο πρόβλημα των δεδομένων, δηλαδή αυτό του επακριβούς συσχετισμού των τεχνικών χαρακτηριστικών με την ικανοποίηση των άυλων αναγκών, οι ερευνητές του [47] χρησιμοποιούν την προσέγγιση της ¨θεωρίας ακατέργαστου συνόλου" (rough set theory), που είναι μια αποτελεσματική και συστηματική μέθοδος για την εξαγωγή πληροφορίας από δεδομένα που παρουσιάζουν μεγάλη αβεβαιότητα [49]. Αναφέρουν επίσης ότι οποιαδήποτε προσπάθεια εφαρμογής μοντέλων παλινδρόμησης θα είχε λανθασμένα συμπεράσματα αφού σε καμία περίπτωση αυτά τα δεδομένα παρουσιάζουν γραμμικό συσχετισμό. Επιπλέον αναφέρουν ότι η συνδυασμένη χρήση νευρωνικών δικτύων και ασαφής λογικής δεν είχε αποτελέσματα. Για να διευθετηθεί κατά το πόσο η προσέγγιση από τα δεδομένα των μέσων κοινωνικής δικτύωσης αντιμετωπίζει αυτό το πρόβλημα, το μοντέλο θα πρέπει να εφαρμοστεί και να αξιολογηθεί.

Ας προσπαθήσουμε σε αυτό το σημείο να αξιοποιήσουμε την συνεργατική μοντελοποίηση. Δημιουργήθηκε ένα μοντέλο που με βάση τα τεχνικά χαρακτηριστικά του προϊόντος, το αξιολογεί με βάση το κατά πόσο ικανοποιεί άυλες ανάγκες. Περιγράφει δηλαδή τον συσχετισμό ανάμεσα στα τεχνικά χαρακτηριστικά και τις άυλες ανάγκες. Επιστρέφοντας στο μοντέλο Πρόβλεψης Αναγκών των Καταναλωτών του προηγούμενου παραδείγματος, παρατηρούμε ότι σαν είσοδο έχει τα χαρακτηριστικά των καταναλωτών, και τις παρελθοντικές αγορές, τις βαθμολογίες των προϊόντων, και τις ανάγκες που προκύπτουν από την εξόρυξη συναισθήματος σε δημοσιεύσεις σχετικές με τα προϊόντα. Τα προϊόντα

μεταφράζονται σε τεχνικά χαρακτηριστικά, επομένως σε υλικές ανάγκες που ικανοποιούν, και εισάγονται στο μοντέλο ώστε να συσχετιστούν τα χαρακτηριστικά των καταναλωτών με τις υλικές ανάγκες τους. Πλέον όμως έχουμε και στη διάθεση μας ένα μοντέλο που με βάση τα τεχνικά χαρακτηριστικά των προϊόντων, εξάγονται και οι άυλες ανάγκες που ικανοποιούν. Επομένως, αν στο στάδιο εκμάθησης του προηγούμενου μοντέλου, εισαχθούν οι άυλες ανάγκες που ικανοποιούν τα προϊόντα αντί για τις υλικές, το μοντέλο τελικά θα μπορεί να συσχετίζει τα χαρακτηριστικά των καταναλωτών με τις άυλες ανάγκες τους. Αυτό το μοντέλο ίσως δεν θα πρεπε να χρησιμοποιείται μεμονωμένα, καθώς υποθέτει ότι οι άυλες ανάγκες είχαν ρόλο στην αγορά, την βαθμολόγηση, ή στο συναίσθημα του καταναλωτή. Στην πραγματικότητα όμως η επιχείρηση δεν μπορεί να ξέρει αν το γεγονός το ότι ένας φορητός υπολογιστής θεωρείται "εντυπωσιακός" είχε ρόλο στην αγορά, η αν αυτή έγινε μόνο με βάση τις άυλες ανάγκες. Επομένως το μοντέλο ίσως θα πρέπει να χρησιμοποιείται συνεργατικά με το μοντέλο άυλων αναγκών, ώστε να προκύπτει τελικά μια πρόβλεψη και των δύο ειδών αναγκών με βάση τα χαρακτηριστικά των καταναλωτών.

Τα δύο αυτά μοντέλα θα μπορούσαν να συνδυαστούν και με ένα τρίτο μοντέλο, το οποίο έχει ως είσοδο τα χαρακτηριστικά των καταναλωτών και υπολογίζει το με τι ποσοστό υλικών και άυλων αναγκών αποφασίζουν για τις αγορές τους, ώστε να επιλέγεται το σωστό μοντέλο, βασισμένο στις παρελθοντικές αγορές του καταναλωτή. Μια βασική προσέγγιση σε αυτό το πρόβλημα θα ήταν η σύγκριση των τεχνικών χαρακτηριστικών των προϊόντων που έχει αγοράσει ο καταναλωτής με άλλα παρόμοια προϊόντα που ήταν διαθέσιμα στην αγορά την ίδια περίοδο.



Μια αρκετά πιο πολύπλοκη έρευνα έγινε από τους [48], με σκοπό τον συσχετισμό των τεχνικών χαρακτηριστικών με τις άυλες ανάγκες, σε συσκευές κινητών τηλεφώνων. Για πρώτο βήμα, ένας αριθμός από συσκευές με διαφορετικά τεχνικά χαρακτηριστικά, όπως διαφορετικό πληκτρολόγιο, σχήμα, αναλογία διαστάσεων και αναλογία οθόνης, παρουσιάστηκαν σε 500 καταναλωτές από τους οποίους ζητήθηκε να αξιολογήσουν ολόκληρη τη συσκευή με βάση μια λέξη συναισθήματος όπως "αθλητικό" και "επιστημονικό". Σε περίπτωση που πάνω από πάνω από 50 καταναλωτές χαρακτήριζαν την ίδια συσκευή με την ίδια λέξη συναισθήματος, η συσκευή θεωρούταν γενικότερα χαρακτηρήσιμη με τη αντίστοιχη λέξη συναισθήματος. Αυτές οι συσκευές, αναλύονταν στα τεχνικά χαρακτηριστικά τους, και αντιστοιχίζονταν με την λέξη συναισθήματος. Προέκυπτε δηλαδή από κάθε συσκευή ένα διάνυσμα της μορφής (αθλητικό, πληκτρολόγιο Α, σχήμα Β, ύψος,....). Σε αυτά τα συνολικά διανύσματα, εφαρμόστηκαν μή γραμμικές τεχνικές μείωσης διαστάσεων, ώστε να ξεκαθαριστούν οι μεταβλητές (όπως τύπος πληκτρολογίου, αναλογία οθόνης) που τελικά είχαν μικρό ρόλο στο πως χαρακτηρίστηκαν οι συσκευές. Στο τελικό στάδιο, χρησιμοποιήθηκαν οι κανόνες συσχετισμού και συγκεκριμένα ο αλγόριθμος Apriori. Δηλαδή, όπως στο πεδίο των πωλήσεων οι αλγόριθμοι των κανόνων συσχετισμού χρησιμοποιούνται ώστε να εντοπιστούν προϊόντα που συνήθως αγοράζονται μαζί, σε αυτή την έρευνα χρησιμοποιήθηκαν ώστε να εντοπιστούν τα τεχνικά χαρακτηριστικά τα οποία συνυπάρχουν σε συσκευές που χαρακτηρίζονται με την ίδια λέξη συναισθήματος. Εξάχθηκαν επομένως συμπεράσματα της μορφής (πληκτρολόγιο Α, σχήμα οθόνης Β, ύψος)= "επιστημονικό". Το μοντέλο εφαρμόστηκε και οι ερευνητές αναφέρουν ότι είναι αποτελεσματικό.

Η έρευνα αυτή, μπορεί να θεωρηθεί ως ένα σημαντικό βήμα στην ικανοποίηση των άυλων αναγκών των καταναλωτών. Πρακτικά, θα ήταν λάθος να υποτεθεί ότι η ικανοποίηση μιας άυλης ανάγκης όπως "αθλητικό" ή "κομψό", προκύπτει από τα τεχνικά χαρακτηριστικά σε ατομικό επίπεδο. Είναι κατανοητό ότι το ίδιο ατομικό τεχνικό χαρακτηριστικό μπορεί να αυξάνει το πόσο "αθλητική" θεωρείται μια ολοκληρωμένη συσκευή αλλά να μειώνει το πόσο "αθλητική" θεωρείται μια άλλη. Τελικά, είναι ο συνδυασμός όλων των τεχνικών χαρακτηριστικών ενός προϊόντος που το χαρακτηρίζει όσο αφορά τις άυλες ανάγκες. Επομένως εντοπίζοντας τους απομονωμένους συνδυασμούς των τεχνικών χαρακτηριστικών που ικανοποιούν άυλες ανάγκες, το μοντέλο κάνει ένα σημαντικό βήμα προς τον τελικό στόχο, ο οποίος είναι η αναγνώριση τελικών και ολοκληρωμένων συνδυασμών. Παράλληλα, επαναλαμβάνεται πως η προκατάληψη του μοντέλου πιθανότατα θα μειωνόταν αν αντί για ένα μικρό δείγμα από 500 καταναλωτές, χρησιμοποιούνταν δεδομένα από τα μέσα

κοινωνικής δικτύωσης για την αρχική αξιολόγηση των ολοκληρωμένων συσκευών ως προς τις άυλες ανάγκες.

Πρόβλεψη Επίδρασης Ικανοποίησης Αναγκών: Μια μοντελοποίηση που θα αφορά την αξιολόγηση των αναγκών των καταναλωτών. Όπως έχει προαναφερθεί η αξιολόγηση των αναγκών των καταναλωτών με βάση το κέρδος που έχει η επιχείρηση από την ικανοποίηση τους, είναι μια κρίσιμη διαδικασία, δεδομένου ότι οι ανάγκες των καταναλωτών μπορεί να είναι αρνητικά συσχετισμένες. Επομένως το μοντέλο θα αντιστοιχεί τα κέρδη που προέκυψαν από τις παρελθοντικές ικανοποιήσεις αναγκών σε αυτές τις ανάγκες, ώστε να μπορεί να ληφθεί η βέλτιστη απόφαση όταν παρουσιαστεί σύγκρουση αναγκών σε έναν μελλοντικό σχεδιασμό προϊόντων. Καθώς αυτή η μοντελοποίηση βασίζεται μόνο στα παρελθοντικά δεδομένα της επιχείρησης, δεν θα αναλυθεί περαιτέρω. Για τον ίδιο λόγο δεν αναλύεται περαιτέρω και η μοντελοποίηση αποχώρησης.

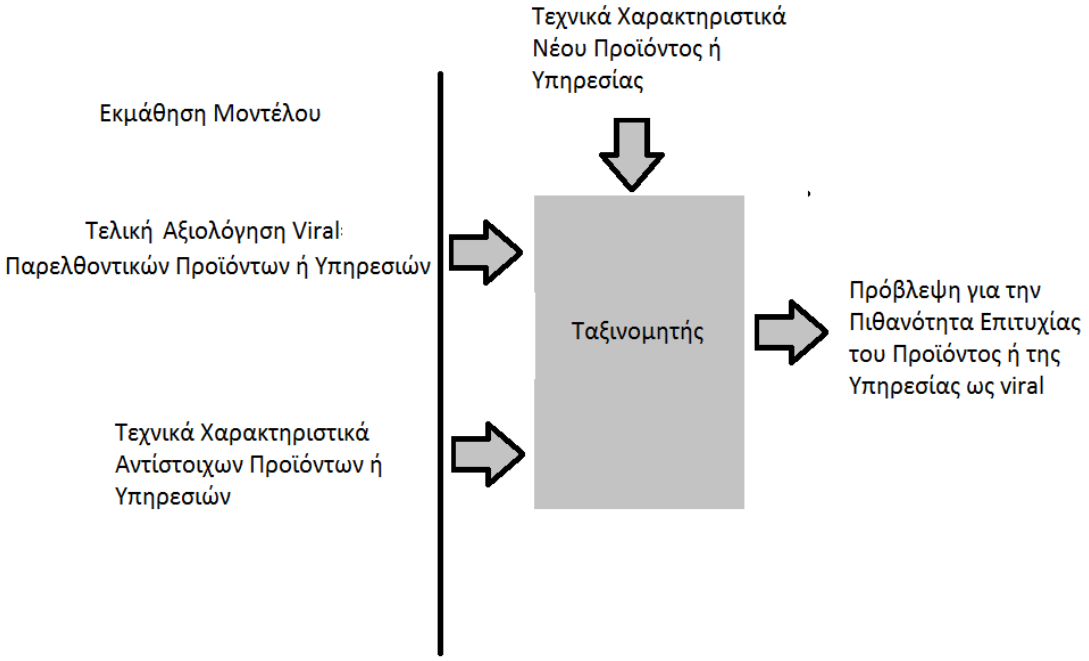
Viral Marketing: Θα προσπαθήσουμε να εφαρμόσουμε Predictive Analytics ώστε να απαντηθεί το ερώτημα "Πως πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά του προϊόντος ή της υπηρεσίας ώστε να γίνει "viral". Πριν ξεκινήσουμε την διαδικασία της μοντελοποίησης, θα πρέπει να ξεκαθαριστεί ότι στο στάδιο της σχεδίασης, δεν θα ασχοληθούμε με το πώς πρέπει να προωθηθεί ένα προϊόν ή υπηρεσία για να γίνει viral, αλλά θα ασχοληθούμε μόνο με το πως πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά. Για τον ίδιο λόγο, ότι δηλαδή θεωρούμε ότι το μοντέλο εξυπηρετεί το τμήμα σχεδιασμού και όχι το τμήμα προώθησης, viral θα θεωρηθεί ένα προϊόν ή υπηρεσία με βάση το κατά πόσο αγοράστηκε από το καταναλωτικό κοινό, και όχι με βάση το πόσο διαδόθηκε στα μέσα κοινωνικής δικτύωσης ή σε άλλα μέσα, αφού το γεγονός ότι μια διαφήμιση προβάλλεται πολλές φορές και από ένα μεγάλο ποσοστό του κοινού δεν συνεπάγεται την αγορά του προϊόντος ή της υπηρεσίας που διαφημίζεται. Επομένως για να αξιολογηθεί το κατά πόσο ένα προϊόν ή υπηρεσία έγινε viral σε πρώτο στάδιο μπορεί να χρησιμοποιηθεί ο λόγος :

Αξιολόγηση Viral:(πωλήσεις)/(μέρες του προϊόντος στην αγορά).

Κατανοώντας ότι η προώθηση που έγινε έχει σε μεγάλο ποσοστό ρόλο στην επιτυχία ενός προϊόντος ως viral, και αναγνωρίζοντας ότι η επίδραση αυτή δεν θα πρέπει να ενσωματωθεί στην μοντελοποίηση των τεχνικών χαρακτηριστικών, η επίδραση της προώθησης θα πρέπει με κάποιο τρόπο να ποσοτικοποιηθεί και να συμπεριληφθεί στην αξιολόγηση viral. Θα χρησιμοποιηθεί ο τύπος (συνολικές προβολές διαφημίσεων)/(μέρες που διαφημίζεται το προϊόν). Επομένως τελικά προκύπτει ο τύπος:

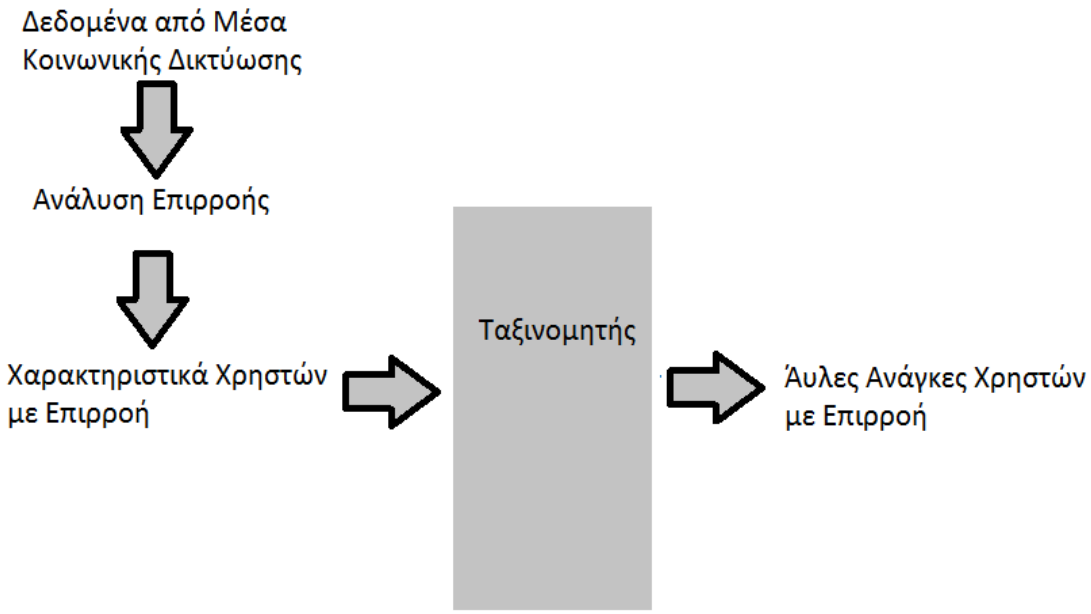
$$\text{Τελική Αξιολόγηση Viral} = \frac{\frac{\text{Πωλήσεις}}{\text{Μέρες του Προϊόντος στην Αγορά}}}{\frac{\text{Συνολικές Προβολές Διαφημίσεων}}{\text{Μέρες που διαφημίζεται το Προϊόν}}}$$

Θεωρώντας ότι η viral επιτυχία εξαρτάται μόνο από την διαφήμιση και τα τεχνικά χαρακτηριστικά, αυτός ο τύπος εκφράζει το κατά πόσο έγινε viral το προϊόν ή η υπηρεσία χάρη στα τεχνικά χαρακτηριστικά της. Ο τύπος θα μπορούσε να είναι αφαιρετικός αλλά τότε θα έπρεπε τα μεγέθη να κανονικοποιηθούν μεταξύ τους με έναν τρόπο που δεν είναι ξεκάθαρος. Σχεδιάζουμε λοιπόν την εξής βασική μοντελοποίηση:



Η πρόβλεψη του ταξινομητή θα είναι επίσης μια τιμή Τελικής Αξιολόγησης Viral, η οποία θα πρέπει να συγκρίνεται με αξιολογήσεις προϊόντων που είναι κοινώς αποδεκτό ότι έγιναν viral.Το μοντέλο μέχρι τώρα αξιοποιεί μόνο τα παρελθοντικά δεδομένα της επιχείρησης και τα τεχνικά χαρακτηριστικά. Θα προσπαθήσουμε να αξιοποιήσουμε και τα δεδομένα από τα μέσα κοινωνικής δικτύωσης. Σε αυτό το στάδιο αναγνωρίζουμε τον ρόλο που έχουν οι χρήστες με επιρροή στα μέσα κοινωνικής δικτύωσης στην επιτυχία viral. Κατανοούμε δηλαδή, ότι αν το προϊόν ή η υπηρεσία αγοραστεί από χρήστες με επιρροή, θα ακολουθήσουν και οι χρήστες τους οποίους επηρεάζουν. Επομένως υπάρχει κέρδος από το να απαντηθεί και το ερώτημα "πως πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά του προϊόντος ή της υπηρεσίας ώστε να ικανοποιούν τις ανάγκες των καταναλωτών με επιρροή". Αξιοποιείται επομένως η ανάλυση επιρροής από τις εφαρμογές των Predictive Analytics στην προώθηση, για να αναγνωριστούν οι χρήστες με επιρροή. Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, αυτή θα είναι καλύτερο να εφαρμοστεί με βάση τις μετρήσεις που προκύπτουν άμεσα από τα μέσα κοινωνικής δικτύωσης, και όχι από την ανάλυση του αντίστοιχου γράφου. Αφού πλέον έχουν αναγνωριστεί οι χρήστες με επιρροή, μπορούν να χρησιμοποιηθούν τα δεδομένα τους για να αναγνωριστούν οι ανάγκες τους, αξιοποιώντας τα μοντέλα από τα προηγούμενα παραδείγματα. Επιπλέον η σχεδίαση θα αξιοποιήσει το μοντέλο των άυλων αναγκών, επειδή

αυτές έχουν μεγαλύτερο ρόλο στην επιτυχία viral [50]. Η ανάγκη για παράδειγμα το προϊόν να θεωρείται “καινοτομικό”, “φρέσκο”, “στυλάτο” κτλ.



Επομένως πλέον η επιχείρηση μπορεί να απαντήσει στα ερωτήματα "πώς πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά ώστε να γίνει το προϊόν ή η υπηρεσία viral με βάση τα παρελθοντικά δεδομένα" και "πώς πρέπει να διαμορφωθούν τα τεχνικά χαρακτηριστικά ώστε να ικανοποιούνται οι ανάγκες των καταναλωτών με επιρροή". Ανάλογα με τα δεδομένα που έχει στη διάθεση της η επιχείρηση, μπορεί να δώσει την ανάλογη σημασία στο κάθε μοντέλο.

Η έρευνα [50], η οποία έγινε με τον ίδιο σκοπό αλλά χρησιμοποίησε μια αρκετά διαφορετική προσέγγιση, είχε αξιοσημείωτα αποτελέσματα. Οι αναλυτές χρησιμοποίησαν ανάλυση επιρροής στον γράφο του δικτύου που προέκυψε από τους 1,4 εκατομμύρια φίλους των 9.687 πειραματικών χρηστών στο Facebook. Μελέτησαν το κατά πόσο έγιναν viral διάφορες εκδόσεις μιας διαδικτυακής εφαρμογής, οι οποίες διαφοροποιούνταν σε κάποια τεχνικά χαρακτηριστικά που θεωρήθηκαν viral από την φύση τους, δηλαδή η ικανότητα να σταλθούν προσκλήσεις για την εφαρμογή σε φίλους ή η δημοσίευση της δραστηριότητας του χρήστη στην εφαρμογή σε φίλους του. Οι χρήστες τμηματοποιήθηκαν σε διάφορα σύνολα πειραματισμού στα οποία προωθήθηκαν οι διαφορετικές εκδόσεις τις εφαρμογής και σε ένα σύνολο ελέγχου στο οποίο η εφαρμογή προωθήθηκε χωρίς τα viral χαρακτηριστικά, η μέθοδος δηλαδή που χρησιμοποιείται και στην Uplift μοντελοποίηση.

Αντίστοιχα με το προτεινόμενο μοντέλο, οι αναλυτές επιχείρησαν να αφαιρέσουν το φαινόμενο της επιρροής που υπάρχει μέσα στο δίκτυο από τις μετρήσεις τους, καθώς ήθελαν να αξιολογήσουν το κατά πόσο η εφαρμογή έγινε viral χάρη στα τεχνικά χαρακτηριστικά της, και όχι λόγω της επιρροής των χρηστών ή της προώθησης. Παρουσιάστηκε επομένως το εμπόδιο της πολύ αυξημένης πολυπλοκότητας των αλγορίθμων της ανάλυσης επιρροής. Για να αντιμετωπισθεί αυτό το πρόβλημα, χρησιμοποιήθηκε μια μορφή μοντελοποίησης κινδύνου (hazard modeling), που είναι μια τεχνική μελέτης της αναμετάδοσης φαινομένων σε γράφους κοινωνικών δικτύων. Στην περίπτωση της εφαρμογής, η μοντελοποίηση κινδύνου αρχικά θα υπολόγιζε την πιθανότητα που θα είχε ένας κόμβος-χρήστης να υιοθετήσει την εφαρμογή, με βάση τα χαρακτηριστικά του όσο αφορά την δραστηριότητά του στο Facebook και την χρήση εφαρμογών, και την επιρροή των κόμβων με τους οποίους συσχετίζεται. Καθώς το υπολογιστικό κόστος παρέμενε πολύ υψηλό, τελικά εφαρμόστηκε αντεστραμμένη μοντελοποίηση κινδύνου, δηλαδή υπολογίστηκε η πιθανότητα κάθε κόμβου να μεταδώσει την εφαρμογή στους υπόλοιπους κόμβους, ανάλογα με την επιρροή του και τα χαρακτηριστικά του. Τελικά, η έρευνα κατέληξε στο ότι η επιτυχής ενσωμάτωση "viral τεχνικών χαρακτηριστικών" σε διαδικτυακές εφαρμογές αυξάνει την πιθανότητα τους να γίνουν viral κατά 400%.

Κανόνες Συσχετισμού και Άυλες Ανάγκες: Μια αρκετά πιο εξειδικευμένη και πολύπλοκη έρευνα έγινε από τους [48]

4.3.3 Αξιολόγηση και Εφαρμογή

Τα προτεινόμενα μοντέλα που παρουσιάζονται σε αυτό το κεφάλαιο αποτελούν προτάσεις του συγγραφέα. Η βασική αξιολόγηση που γίνεται είναι θεωρητική, και βασίζεται στα πλεονεκτήματα που έχει αποδειχθεί ότι έχει η αξιοποίηση των δεδομένων από τα μέσα κοινωνικής δικτύωσης στα μοντέλα που εμπλέκονται με την συμπεριφορά των καταναλωτών. Για να αξιολογηθούν αυτά τα μοντέλα ουσιαστικά, πρέπει πρώτα να εφαρμοστούν σε πραγματικά δεδομένα. Τα μοντέλα ερευνών που παρουσιάζονται είναι λίγα σε αριθμό ώστε να εξαχθούν συμπεράσματα, καθώς η εφαρμογή των Predictive Analytics φαίνεται μέχρι σήμερα να συγκεντρώνεται περισσότερο στις πωλήσεις και την προώθηση.

5 Βιβλιογραφία

1 Dhillon, Inderjit S., Yuqiang Guan, and Brian Kulis. "Weighted graph cuts without eigenvectors a multilevel approach." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29.11 (2007): 1944-1957.

2 Satuluri, Venu, Srinivasan Parthasarathy, and Yiye Ruan. "Local graph sparsification for scalable clustering." *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. ACM, 2011.

3 Ronen, Inbal, et al. "Social networks and discovery in the enterprise (SaND)."*Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009.

4 M. Meila, J. Shi, A random walks view of spectral segmentation, AI and Statistics, 2001

5 Leicht, Elizabeth A., and Mark EJ Newman. "Community structure in directed networks." *Physical review letters* 100.11 (2008): 118703.

6 V. Satuluri and S. Parthasarathy. Symmetrizations for clustering directed graphs. In *Workshop on Mining and Learning with Graphs, MLG 2010*, 2010

7 Wang, Xuerui, Natasha Mohanty, and Andrew McCallum. *Group and topic discovery from relations and their attributes*. MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE, 2006.

8 Zhou, Ding, et al. "Probabilistic models for discovering e-communities."*Proceedings of the 15th international conference on World Wide Web*. ACM, 2006.

9 Pathak, Nishith, et al. "Social topic models for community extraction." *The 2nd SNA-KDD workshop*. Vol. 8. 2008.

10 Granovetter, Mark S. "The strength of weak ties." *American journal of sociology* (1973): 1360-1380.

11 Granovetter, Mark. "Economic action and social structure: the problem of embeddedness." *American journal of sociology* (1985): 481-510.

12 Holland, Paul W., and Samuel Leinhardt. "Transitivity in structural models of small groups." *Comparative Group Studies* (1971).

13 Tsourakakis, Charalampos E. "Fast counting of triangles in large real

networks without counting: Algorithms and laws." *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008.

14 Borgatti, Stephen P., and Martin G. Everett. "A graph-theoretic perspective on centrality." *Social networks* 28.4 (2006): 466-484.

15 Burt, Ronald S. *Structural holes: The social structure of competition*. Harvard university press, 2009.

16 Kempe, David, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network." *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003.

17 Siegel, Eric. *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons, 2013.

18 Ling, Charles X., and Chenghui Li. "Data Mining for Direct Marketing: Problems and Solutions." *KDD*. Vol. 98. 1998.

19 Liben-Nowell, David, and Jon Kleinberg. "The link-prediction problem for social networks." *Journal of the American society for information science and technology* 58.7 (2007): 1019-1031.

20 Al Hasan, Mohammad, et al. "Link prediction using supervised learning."*SDM'06: Workshop on Link Analysis, Counter-terrorism and Security*. 2006.

21 Barabási, Albert-László, and Réka Albert. "Emergence of scaling in random networks." *science* 286.5439 (1999): 509-512.

22 Kleinberg, Jon M. "Navigation in a small world." *Nature* 406.6798 (2000): 845-845.

23 Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos. "Graphs over time: densification laws, shrinking diameters and possible explanations."*Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005.

24 Kashima, Hisashi, and Naoki Abe. "A parameterized probabilistic model of network evolution for supervised link prediction." *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 2006.

25 Huang, J., et al. "A comparison of calibration methods based on calibration data size and robustness." *Chemometrics and Intelligent Laboratory Systems*62.1 (2002): 25-35.

26 Vigneau, E., et al. "Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration." *Journal of chemometrics* 11.3 (1997): 239-249.

27 Li, Xueping, Godswill Chukwugozie Nsofor, and Laigang Song. "A comparative analysis of predictive data mining techniques." *International Journal of Rapid Manufacturing* 1.2 (2009): 150-172.

28 Basak, Subhash C., et al. "Prediction of Human Blood: Air Partition Coefficient: A Comparison of Structure-Based and Property-Based

Methods." *Risk Analysis* 23.6 (2003): 1173-1184.

29 Naes, T., C. Irgens, and H. Martens. "Comparison of linear statistical methods for calibration of NIR instruments." *Applied Statistics* (1986): 195-206.

30 Schumaker, Robert P., and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." *ACM Transactions on Information Systems (TOIS)* 27.2 (2009): 12.

31 Li, Jingxuan, et al. "Social network user influence sense-making and dynamics prediction." *Expert Systems with Applications* 41.11 (2014): 5115-5124.

32 Aggarwal, Charu C., and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

33 Jiang, Shan, et al. "Analyzing firm-specific social media and market: A stakeholder-based event analysis framework." *Decision Support Systems* 67 (2014): 30-39.

34 LOVETT, JOHN, and JEREMIAH OWYANG. "Social Marketing Analytics." *A framework for measuring results in Social Media, Altimeter Group downloaded from www. web-strategist. com* (2010).

35 Asur, Sitaram, and Bernardo Huberman. "Predicting the future with social media." *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. Vol. 1. IEEE, 2010.

36 Τζαλακώστα, Ελένη, and Ελισάβετ Τσοτσόλη. "Εφαρμογές των support vector machines σε προβλήματα κατηγοριοποίησης." (2006).

37 Baumes, L. A., et al. "Support vector machines for predictive modeling in heterogeneous catalysis: a comprehensive introduction and overfitting investigation based on two real applications." *Journal of combinatorial chemistry* 8.4 (2006): 583-596.

38 Liu, Yang, et al. "ARSA: a sentiment-aware model for predicting sales performance using blogs." *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007.

39 Zhang, Wenbin, and Steven Skiena. "Improving movie gross prediction through news analysis." *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2009.

40 Karsak, E. Ertugrul, Sevin Sozer, and S. Emre Alptekin. "Product planning in quality function deployment using a combined analytic network process and goal programming approach." *Computers & industrial engineering* 44.1 (2003): 171-190.

41 Gilbert, Eric, and Karrie Karahalios. "Widespread Worry and the Stock Market."*ICWSM*. 2010.

42 Negoescu, Radu-Andrei, et al. "Flickr hypergroups." *Proceedings of the 17th*

ACM international conference on Multimedia. ACM, 2009.

43 Newman, Mark EJ. "A measure of betweenness centrality based on random walks." *Social networks* 27.1 (2005): 39-54.

44 Domingos, Pedro, and Matt Richardson. "Mining the network value of customers." *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001.

45 Eastwick, Paul W., and Wendi L. Gardner. "Is it a game? Evidence for social influence in the virtual world." *Social Influence* 4.1 (2009): 18-32.

46 Moser, Flavia, Rong Ge, and Martin Ester. "Joint cluster analysis of attribute and relationship data withouta-priori specification of the number of clusters."*Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007.

47 Zhai, Lian-Yin, Li-Pheng Khoo, and Zhao-Wei Zhong. "A rough set based decision support approach to improving consumer affective satisfaction in product design." *International Journal of Industrial Ergonomics* 39.2 (2009): 295-302.

48 Shi, Fuqian, Shouqian Sun, and Jiang Xu. "Employing rough sets and association rule mining in KANSEI knowledge extraction." *Information Sciences* 196 (2012): 118-128.

49 Pawlak, Zdzisław. "Rough sets." *International Journal of Computer & Information Sciences* 11.5 (1982): 341-356.

50 Aral, Sinan, and Dylan Walker. "Creating social contagion through viral product design: A randomized trial of peer influence in networks." *Management Science*57.9 (2011): 1623-1639.