ARISTOTLE UNIVERSITY OF THESSALONIKI

Επιχειρησιακή Έρευνα και Επιχειρηματική Ευφυΐα

03 / 31 / 2022

Μάρκος Κολέτσας – ΑΕΜ.: 3557

Εισαγωγή.

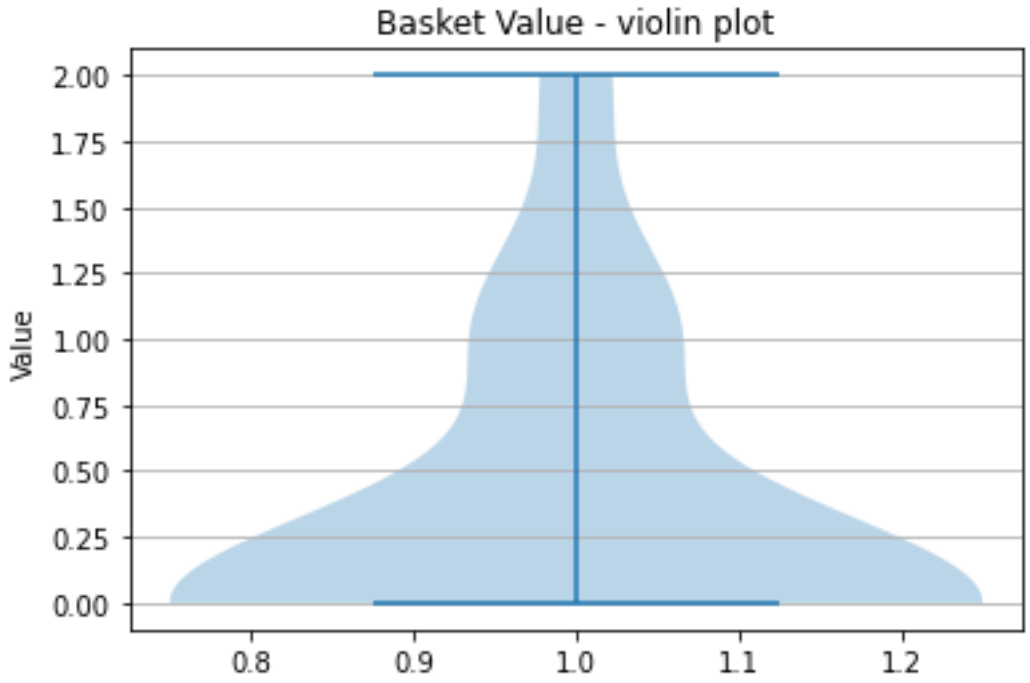
Η εργασία αυτή προέβλεπε την εξοικείωση του φοιτητή με το μετασχηματισμών δεδομένων, τη μάθηση κανόνων συσχέτισης με τη μέθοδο Apriori και ομαδοποίηση με τη μέθοδο K-means. Σκοπός της εργασίας ήταν να βοηθήσουμε το τμήμα Marketing μιας μεγάλης πολυεθνικής αλυσίδας καταστημάτων λιανικής στην ανάλυση προϊόντων.

Άσκηση 1.

Ξεκινάμε φορτώνοντας τα δεδομένα μας, δηλαδή έναν πίνακα με 7537 γραμμές και 35 στήλες, που κάθε γραμμή αντιπροσωπεύει και από μία συναλλαγή που έγινε σε ένα από τα καταστήματα αυτής της πολυεθνικής και κάθε στήλη ένα προϊόν το οποίο αγοράστηκε στη κάθε συναλλαγή. Επομένως, πρέπει πρώτα να καθαρίσουμε τα δεδομένα μας καθώς έχουμε πολλές «NaN» τιμές και αρκετά από τα προϊόντα αυτά δε μας αφορούν σύμφωνα με εντολή του τμήματος Marketing. Συνεπώς, πρώτη μας κίνηση είναι να κρατήσουμε μόνο τα προϊόντα που χρειαζόμαστε και να μετασχηματίσουμε τα δεδομένα μας σε ένα δυαδικό πίνακα συναλλαγών, ώστε να μας είναι εύχρηστα και να μπορούμε να εξάγουμε κανόνες συσχέτισης με τη μέθοδο Apriori από αυτά. Επίσης, στο δυαδικό πίνακα συναλλαγών προσαρτούμε και κάποιες άλλες τιμές από τον αρχικό μας πίνακα που θεωρήσαμε χρήσιμες. Αυτές ήταν η ταυτότητα κάθε συναλλαγής (id), η αξία κάθε συναλλαγής (basket_value) και πόσες μέρες έχουν περάσει απ' όταν έγινε μία συναλλαγή (recency_days). Τώρα το μόνο που μας μένει είναι να μπορέσουμε να χωρίσουμε τις συναλλαγές σε τρεις διακριτές τιμές σύμφωνα με την αξία τους. Ωστόσο οι τρεις αυτές κατηγορίες θα πρέπει να είναι σχεδόν ισοπληθείς, για να μην έχουν πολύ διαφορετικό support.

Οπτικοποιούμε τα δεδομένα σας με τη χρήση ενός violin plot για να δούμε τη σχέση που υπάρχει μεταξύ αξίας και ποσότητας δειγμάτων.

Το αποτέλεσμα το οποίο παίρνουμε είναι το εξής:



Παρατηρώντας το γράφημα αυτό συμπεραίνουμε κιόλας πως το να χωρίσουμε τα δεδομένα μας σε τρεις ισοπληθείς κατηγορίες δεν θα τα αντιπροσωπεύει κατάλληλα, αφού έχουμε όλο και λιγότερα δείγματα όσο αυξάνει η τιμή του καλαθιού.

Μία πρώτη προσέγγιση είναι να τα ταξινομήσουμε και ύστερα κάθε κατηγορία να αποτελείται από το $\frac{1}{3}$ των δειγμάτων μας. Παρόλα αυτά, έτσι θα έχουμε επικαλυπτόμενα διαστήματα, δηλαδή καλάθια που έχουν τις ίδιες τιμές μπορεί να ανήκουν σε διαφορετικές κατηγορίες, οπότε αυτή η προσέγγιση απορρίπτεται.

Η επόμενη ιδέα μας είναι να βρούμε ένα διάστημα κατάλληλο γύρω από τη μέση τιμή της αξίας του καλαθιού η οποία να χωρίζει τις κατηγορίες μας έτσι ώστε να ισχύει το εξής:

$$len(low\_basket\_value) > len(medium\_basket\_value) > len(high\_basket\_value)$$

Το αποτέλεσμα το οποίο παίρνουμε πάλι δεν είναι αρκετά αντιπροσωπευτικό, αλλά είναι αρκετά καλό από άποψη

πολυπλοκότητας και πληθικότητας, εφόσον δεν απαιτεί πολύπλοκους υπολογισμούς και οι κατηγορίες μας είναι σχεδόν ισοπληθείς αλλά παράλληλά διατηρούν τη σχέση που αναφέραμε παραπάνω.

Άσκηση 2.

a) Δοκιμάσαμε στην αρχή την εκτέλεση τη μεθόδου Apriori με πολλές διαφορετικές τιμές της παραμέτρου support, οι οποίες ανήκαν στο διάστημα $[0.005, 0.5]$, δηλαδή να υπάρχει υποστήριξη από 0.5% μέχρι 50%. Στις πολύ υψηλές τιμές δεν υπήρχαν ούτε συχνά σύνολα και συνεπώς ούτε παραγόμενοι κανόνες συσχέτισης, οπότε συμπεράναμε πως πρέπει να αναζητήσουμε μια χαμηλότερη τιμή και περιοριστήκαμε στη δοκιμή τιμών <0.1. Επίσης, θέσαμε δύο συνθήκες για να καταλήξουμε σε μία επιτρεπτή τιμή, να παράγονται τουλάχιστον 20 κανόνες και να υπάρχει τουλάχιστον ένας κανόνας που σε ένα από τα δύο μέλη του θα έχει σύνολο αντικειμένων με 2 στοιχεία. Ύστερα, από τους πειραματισμούς καταλήξαμε στην τιμή του ελάχιστου support ίση με 0.025, δηλαδή 2.5% υποστήριξη. Για τη συγκεκριμένη τιμή λαμβάναμε 51 «συχνά» σύνολα αντικειμένων και 88 κανόνες. Για τους κανόνες δεν θέσαμε κάποιο ελάχιστο confidence, καθώς προτιμούσαμε να παράγονται όλοι και ύστερα να τους ταξινομούμε, ώστε να μελετήσουμε μόνο όσους ήταν απαραίτητο. Η ίδια τιμή ελάχιστης υποστήριξης θα χρησιμοποιεί και για τις υπόλοιπες εκτελέσεις του αλγορίθμου, ώστε να μπορεί να γίνει η κατάλληλη σύγκριση των αποτελεσμάτων.

b) Οι 20 παραγόμενοι κανόνες με το μεγαλύτερο confidence για τη μέθοδο Apriori με είσοδο μόνο τα προϊόντα ήταν οι εξής:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
77	(yogurt, other vegetables)	(whole milk)	0.056661	0.333466	0.029061	0.512881	1.538029	0.010166	1.368317
84	(root vegetables, other vegetables)	(whole milk)	0.061837	0.333466	0.030255	0.489270	1.467227	0.009634	1.305062
82	(whole milk, root vegetables)	(other vegetables)	0.063827	0.252521	0.030255	0.474012	1.877119	0.014137	1.421096
31	(root vegetables)	(whole milk)	0.142251	0.333466	0.063827	0.448694	1.345546	0.016391	1.209009
44	(root vegetables)	(other vegetables)	0.142251	0.252521	0.061837	0.434701	1.721445	0.025915	1.322273
8	(tropical fruit)	(whole milk)	0.136943	0.333466	0.055202	0.403101	1.208821	0.009536	1.116661
26	(yogurt)	(whole milk)	0.182059	0.333466	0.073116	0.401603	1.204331	0.012405	1.113867
76	(yogurt, whole milk)	(other vegetables)	0.073116	0.252521	0.029061	0.397459	1.573963	0.010597	1.240545
21	(other vegetables)	(whole milk)	0.252521	0.333466	0.097665	0.386758	1.159812	0.013457	1.086902
32	(pastry)	(whole milk)	0.116109	0.333466	0.043392	0.373714	1.120697	0.004673	1.064265
3	(citrus fruit)	(whole milk)	0.108015	0.333466	0.039809	0.368550	1.105211	0.003790	1.055562
4	(citrus fruit)	(other vegetables)	0.108015	0.252521	0.037686	0.348894	1.381644	0.010410	1.148015
10	(tropical fruit)	(other vegetables)	0.136943	0.252521	0.046842	0.342054	1.354556	0.012261	1.136080
55	(sausage)	(rolls/buns)	0.122611	0.240048	0.039942	0.325758	1.357053	0.010509	1.127120
28	(sausage)	(whole milk)	0.122611	0.333466	0.039013	0.318182	0.954166	-0.001874	0.977583
40	(yogurt)	(other vegetables)	0.182059	0.252521	0.056661	0.311224	1.232469	0.010687	1.085228
25	(bottled water)	(whole milk)	0.144241	0.333466	0.044851	0.310948	0.932471	-0.003248	0.967320
83	(whole milk, other vegetables)	(root vegetables)	0.097665	0.142251	0.030255	0.309783	2.177726	0.016362	1.242724
22	(rolls/buns)	(whole milk)	0.240048	0.333466	0.073912	0.307905	0.923347	-0.006136	0.963067
78	(whole milk, other vegetables)	(yogurt)	0.097665	0.182059	0.029061	0.297554	1.634380	0.011280	1.164418

Παρατηρώντας τους κανόνες που παράγονται στο παρακάτω πίνακα, υποθέτοντας πως τα δεδομένα μας βασίζονται στον πραγματικό κόσμο, δεν μας κάνει καθόλου εντύπωση που στους περισσότερους κανόνες, είτε στο δεξί, είτε στο αριστερό μέλος υπάρχει το whole_milk, καθώς είναι ένα αγαθό που αγοράζεται πολύ συχνά σχεδόν από κάθε νοικοκυριό. Αν έστω πως τα δεδομένα μας δεν βασίζονται στον πραγματικό κόσμο, πάλι το ίδιο αποτέλεσμα θα περιμέναμε, καθώς το whole_milk είναι το στοιχείο με το μεγαλύτερο support και υπάρχει στο 1/3 του συνόλου των συναλλαγών μας.

Επίσης, αν λάβουμε τη σύβαση του πραγματικού κόσμου, δεν μας κάνει εντύπωση πως τα root_vegetables εμφανίζονται πολύ συχνά με τα other_vegetables, καθώς πολύ πιθανόν να πωλούνται σε κοινά τμήματα ενός καταστήματος, ομοίως και το yogurt με το whole_milk.

Αυτό που όμως μας κάνει εντύπωση παρόλα αυτά, είναι πως αντικείμενα όπως το rolls/buns και το soda που έχουν υψηλή υποστήριξη, περίπου το 1/4 του συνόλου των συναλλαγών, δεν συναντώνται σχεδόν σε κανέναν από αυτούς τους κανόνες. Πράγμα που μπορεί να σημαίνει πως πολλές φορές, ίσως, να αγοράζονται μόνα τους.

c) Αφού αποφανθήκαμε για τα προϊόντα προσθέτουμε και τις διακριτές τιμές καλαθιών στα δεδομένα μας για να δούμε τις νέες συσχετίσεις που δημιουργούνται και πως η συγκεκριμένη μεταβλητή επηρεάζει τους κανόνες μας. Οπότε τώρα επαναλαμβάνουμε αυτή τη διαδικασία για το τροποποιημένο σύνολο δεδομένων μας και λαμβάνουμε το εξής αποτέλεσμα:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
252	(rolls/buns, pastry)	(high_basket_value)	0.027335	0.311438	0.027335	1.000000	3.210908	0.018822	inf
264	(yogurt, sausage)	(high_basket_value)	0.025610	0.311438	0.025610	1.000000	3.210908	0.017634	inf
246	(rolls/buns, sausage)	(high_basket_value)	0.039942	0.311438	0.039942	1.000000	3.210908	0.027502	inf
282	(sausage, soda)	(high_basket_value)	0.031714	0.311438	0.031714	1.000000	3.210908	0.021837	inf
228	(pastry, other vegetables)	(high_basket_value)	0.029459	0.311438	0.027203	0.923423	2.965027	0.018028	8.991804
180	(sausage, whole milk)	(high_basket_value)	0.039013	0.311438	0.035430	0.908163	2.916028	0.023280	7.497670
216	(sausage, other vegetables)	(high_basket_value)	0.035165	0.311438	0.031714	0.901887	2.895875	0.020763	7.018031
192	(pastry, whole milk)	(high_basket_value)	0.043392	0.311438	0.036624	0.844037	2.710124	0.023110	4.414895
17	(sausage)	(high_basket_value)	0.122611	0.311438	0.099257	0.809524	2.599306	0.061071	3.614948
270	(yogurt, root vegetables)	(high_basket_value)	0.033705	0.311438	0.026274	0.779528	2.502991	0.015777	3.123119
156	(yogurt, tropical fruit)	(high_basket_value)	0.038217	0.311438	0.029459	0.770833	2.475075	0.017556	3.004632
276	(yogurt, soda)	(high_basket_value)	0.035695	0.311438	0.026539	0.743494	2.387292	0.015422	2.684392
144	(tropical fruit, whole milk)	(high_basket_value)	0.055202	0.311438	0.040738	0.737981	2.369588	0.023546	2.627905
150	(tropical fruit, other vegetables)	(high_basket_value)	0.046842	0.311438	0.033970	0.725212	2.328590	0.019382	2.505796
240	(yogurt, rolls/buns)	(high_basket_value)	0.044851	0.311438	0.032245	0.718935	2.308434	0.018277	2.449830
204	(rolls/buns, other vegetables)	(high_basket_value)	0.055600	0.311438	0.039145	0.704057	2.260663	0.021830	2.326672
210	(yogurt, other vegetables)	(high_basket_value)	0.056661	0.311438	0.039544	0.697892	2.240868	0.021897	2.279192
234	(soda, other vegetables)	(high_basket_value)	0.042728	0.311438	0.029591	0.692547	2.223703	0.016284	2.239564
258	(rolls/buns, soda)	(high_basket_value)	0.050027	0.311438	0.033174	0.663130	2.129249	0.017594	2.043998
21	(pastry)	(high_basket_value)	0.116109	0.311438	0.076433	0.658286	2.113695	0.040272	2.015021

Πλέον στους κανόνες εμφανίζονται και πολλά προϊόντα που δεν συναντήσαμε στους προηγούμενους. Όμως, αυτό που είναι άξιο σχολιασμού είναι πως όλοι οι κανόνες στο δεξί τους μέρος έχουν αποκλειστικά το high_basket_value. Αυτό συμβαίνει, διότι παρόλο

που τα καλάθια υψηλής αξίας είναι λιγότερα περιέχουν πάντοτε περισσότερα προϊόντα και πιο συγκεκριμένα.

Επομένως, μπορούμε να αποφανθούμε για ποιο είναι πιθανόν το πιο ακριβό προϊόν. Από τους πρώτους κιόλας κανόνες συνειδητοποιούμε την συχνή εμφάνιση του sausage στο αριστερό μέρος των κανόνων και μάλιστα με αρκετά υψηλό confidence (έως και 1). Ωστόσο, ο κανόνας που επαληθεύει την υπόθεση μας είναι ο εξής: (sausage) -> (high_basket_value). Ο πρώτος κανόνας με το υψηλότερο confidence, που συνδέει μόνο ένα προϊόν με το high_basket_value. Οπότε, συμπεραίνουμε πως το αντικείμενο sausage είναι αυτό που πιο συχνά απ' όλα τα υπόλοιπα, όταν υπάρχει σε ένα καλάθι, αυτό θα είναι καλάθι υψηλής αξίας. Συνεπώς, είναι το προϊόν με τη μεγαλύτερη πιθανότητα να είναι το πιο ακριβό.

Σημείωση: Στο notebook χρησιμοποιήθηκαν κι άλλες μετρικές (lift, leverage) για το ερώτημα (β), των οποίων τα αποτελέσματα δε θα σχολιαστούν, ούτε θα παρουσιαστούν, διότι δεν ήταν αυτό το ζητούμενο. Όμως αξίζω να αναφέρουμε πως παρατηρήσαμε ότι οι κανόνες που προκύπτουν από τις άλλες δύο μετρικές δεν είναι οι περισσότεροι κοινοί με τη μετρική του confidence. Αυτό είναι λογικό να συμβαίνει, καθώς χρησιμοποιούν διαφορετικό τρόπο υπολογισμού για να δώσουν βαρύτητα αλλού. Εμείς στη συγκεκριμένη εργασία, όμως θα ενασχοληθούμε αποκλειστικά με το confidence.

Άσκηση 3.

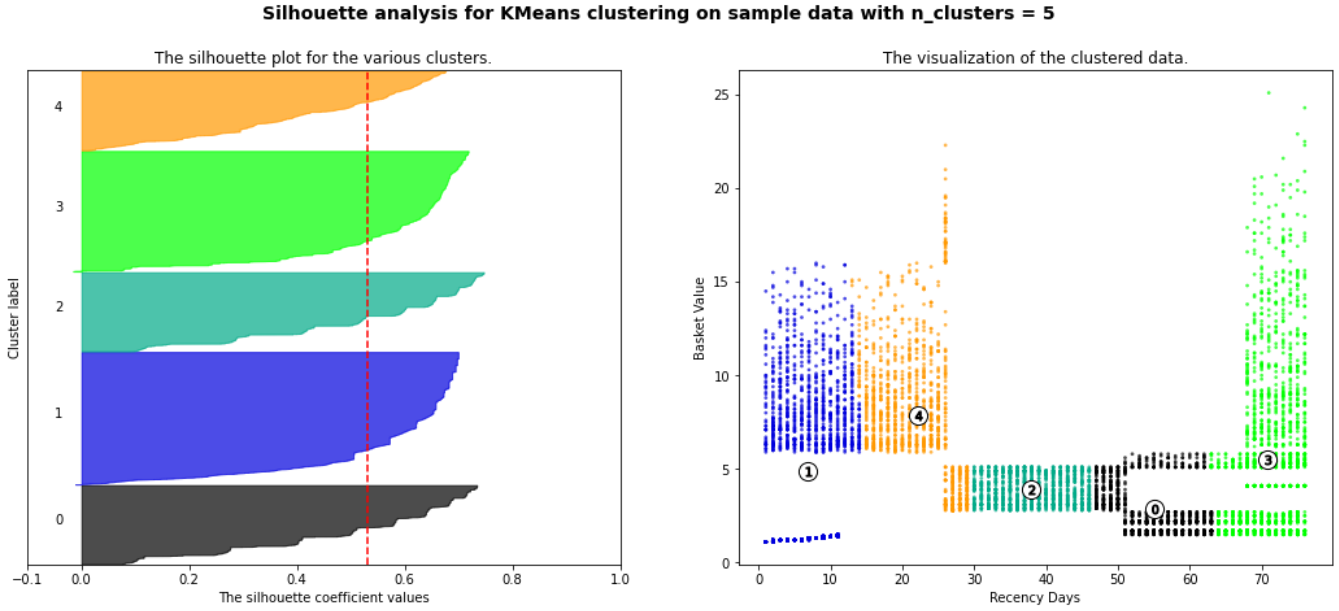
a) Ας αρχίσουμε με τη περιγραφή της διαδικασίας που κάναμε για να μπορέσουμε να εκτελέσουμε τον αλγόριθμο K-means στα δεδομένα μας για τα χαρακτηριστικά recency_days και basket_value. Πρώτα ως συνήθως προσθέσαμε όλες τις βιβλιοθήκες που θα χρειαστούμε για τον αλγόριθμο K-means και

για την περαιτέρω εξήγηση των αποτελεσμάτων του. Έπειτα, η εκτέλεση του αλγορίθμου K-means ήταν αρκετά απλή, αφού δεν χρειαζόταν να την υλοποιήσουμε εμείς. Αφού δημιουργήσαμε ένα αντικείμενο της κλάσης K-means με τα κατάλληλα ορίσματα, μας μένει μόνο να κάνουμε fit και predict για τα δεδομένα μας. Τα δεδομένα μας για να τα δώσουμε στη συνάρτηση fit τα παίρνουμε από το Dataframe που δημιουργήσαμε στην αρχή με τα ονόματα των στηλών τους, αλλά προσοχή παίρνουμε μόνο τα values τους και δεν δίνουμε ως είσοδο στον αλγόριθμο το Dataframe, καθώς περιέχει και String (τα ονόματα των στηλών) τα οποία ο αλγόριθμος δε θα γνωρίζει πως να διαχειριστεί. Επίσης χρησιμοποιούμε random_state για να παράγουμε κάθε φορά τα ίδια αποτελέσματα και να υπάρχει ντετερμινιστικότητα.
Παρακάτω αφού εκτελέσουμε τον αλγόριθμο K-means με 5 κέντρα υπολογίζουμε 3 πράγματα:

1. Το ποσοστό που αντιπροσωπεύει κάθε συστάδα από το σύνολο των συναλλαγών
2. Τη μέση τιμή των κέντρων των συστάδων, δηλαδή βρίσκουμε ένα σημείο $[x, y]$, όπου το x αναφέρεται στο recency_days και το y στο basket_value.
3. Τη τυπική απόκλιση πάλι για το κέντρο των συστάδων των δύο αυτών μεταβλητών.

Τέλος, η έξοδος που πήραμε από την εκτέλεση του αλγορίθμου είχε την εξής μορφή: `array([1, 2, 1, ..., 1, 2, 2], dtype=int32)`
Δηλαδή ήταν ένα μονοδιάστατο array με ακέραιους στο διάστημα $[0, 4]$ και μήκος ίσο με τον αριθμό των συναλλαγών. Κάθε ακέραιος σε κάθε μία από τις 7536 θέσεις, έδειχνε σε ποια συστάδα άνηκε ένα δείγμα.

Ας μελετήσουμε τώρα και μια οπτικοποίηση του αποτελέσματος του αλγορίθμου K-means, όπου στον άξονα x είναι οι μέρες πού πέρασαν απ' όταν έγινε η συναλλαγή και στον y το ύψος της συναλλαγής. Επίσης, στα αριστερά της παρακάτω εικόνας υπάρχουν και τα silhouette coefficient values.



b) Η μέση τιμή των κέντρων των συστάδων που προέκυψαν ήταν [38.55128911, 5.00521979]. Επίσης, αφού υπολογίσαμε και τη μέση τιμή της κάθε μίας από τις δύο μεταβλητές μας ξεχωριστά συνειδητοποιήσαμε πως δεν υπάρχει μεγάλη απόκλιση από τη μέση τιμή των κέντρων των συστάδων.

Η τυπική απόκλιση των κέντρων των συστάδων που προέκυψαν ήταν [22.73551897, 1.69576367]. Επίσης, η τυπική απόκλιση ή διασπορά των τιμών έχει πολύ μεγαλύτερο ενδιαφέρον στην ερμηνεία των αποτελεσμάτων μας. Ας δούμε τη πληροφορία αντλούμε από τις τιμές αυτές: Τα δείγματα της μεταβλητής recency_days είναι πιο διάσπαρτα στο χώρο και μάλιστα με ομοιόμορφο τρόπο. Τις πρώτες και τις τελευταίες μέρες έχουμε 160+ συναλλαγές τη μέρα και σε όλο το υπόλοιπο διάστημα έχουμε 60-80 συναλλαγές καθημερινώς με ελάχιστες εξαιρέσεις.

Επομένως, η μεταβλητή αυτή λόγω της διασποράς της είναι πιο εύκολα διαχωρίσιμη, αυτό παρατηρείται και παρακάτω στα αριθμητικά προφίλ των συστάδων. Από την άλλη η μεταβλητή basket_value έχει πολύ μικρή διασπορά με αποτέλεσμα οι περισσότερες τιμές της να είναι γύρω από τη μέση τιμή της πράγμα που κάνει τα δείγματα της αρκετά δύσκολα στο διαχωρισμό τους.
Ας δούμε όμως τώρα και τα αριθμητικά προφίλ των συστάδων αυτών:

- Συστάδα 1 --> Συστάδα πρόσφατων συναλλαγών, τελευταίες δύο εβδομάδες, πολύ χαμηλής, μεσαίας και υψηλής αξίας που αντιπροσωπεύει το 27.03% του συνόλου των συναλλαγών.
- Συστάδα 4 --> Συστάδα πρόσφατων συναλλαγών, αλλά από τη 15η έως περίπου την 29η μέρα, κυρίως μεσαίας αξίας που αντιπροσωπεύει το 16.12% του συνόλου των συναλλαγών.
- Συστάδα 2 --> Συστάδα πρόσφατων και παρελθοντικών συναλλαγών, από τη 30η έως περίπου την 47η μέρα, χαμηλής και μεσαίας αξίας που αντιπροσωπεύει το 16.15% του συνόλου των συναλλαγών.
- Συστάδα 0 --> Συστάδα παρελθοντικών συναλλαγών, από τη 48η έως περίπου την 62η μέρα, χαμηλής και μεσαίας αξίας που αντιπροσωπεύει το 16.07% του συνόλου των συναλλαγών.
- Συστάδα 3 --> Συστάδα παρελθοντικών συναλλαγών, από τη 63η έως την 75η μέρα, χαμηλής, μεσαίας και υψηλής αξίας που αντιπροσωπεύει το 24.60% του συνόλου των συναλλαγών.

Με τον όρο πρόσφατων συναλλαγών εννοούμε πως ισχύει $recency\_days \leq mean(recency\_days)$, ενώ με τον όρο παρελθοντικών συναλλαγών εννοούμε πως $recency\_days > mean(recency\_days)$. Ομοίως για την αξία συναλλαγών για τη περιοχή κοντά στο $mean(basket\_value)$ θεωρούμε τη μεσαία αξία, και εκτός αυτού του διαστήματος

έχουμε τη χαμηλή και την υψηλή αξία, όπως ακριβώς χωρίσαμε πριν τις κατηγορίες αξίας. Επίσης, στα silhouette coefficient values φαίνεται αυτό που αναφέρουμε προηγουμένως, πως οι συστάδες 1 και 3, που αντιπροσωπεύουν τις πρώτες και τελευταίες μέρες, έχουν μεγαλύτερη πυκνότητα αφού έγιναν περισσότερες συναλλαγές σε εκείνα τα διαστήματα. Για αυτό κιόλας αντιπροσωπεύουν και περίπου το 52% του συνόλου των συναλλαγών μας.

Πρόβλημα Clustering: Παρατηρείται πως στις παλαιότερες συναλλαγές (συστάδα 3) είναι πιο συχνές αυτές της υψηλής αξίας σε σχέση με τις πιο πρόσφατες (συστάδες 1 & 4). Επίσης, ίσως να γινόταν καλύτερος διαχωρισμός των συστάδων αν χρησιμοποιούσαμε λιγότερες συστάδες, διότι η συστάδες 2 & 0 έχουν το ίδιο αριθμητικό προφίλ με μόνη διαφορά το χρονικό πλαίσιο στο οποίο αναφέρονται. Βέβαια, ίσως να γινόταν και ακόμα καλύτερος διαχωρισμός αν αυξάναμε τις συστάδες, έτσι ώστε να διαχωρίζουν τις συναλλαγές μας τόσο βάσει της χρονικής περιόδου, όσο και του κόστους, πράγμα το οποίο δεν συμβαίνει τώρα, ωστόσο αυτό το διαχωρισμό περιμέναμε λόγω της τυπικής απόκλισης.

Πρόβλημα Μάρκετινγκ: Παρατηρείται πως οι πιο πρόσφατες και οι πιο παρελθοντικές μέρες είναι αυτές με τις περισσότερες συναλλαγές και μάλιστα στις παρελθοντικές είναι αυτές που εμφανίζεται να έχουν και την υψηλότερη αξία ταυτόχρονα. Σημαντικό, θα ήταν το τμήμα μάρκετινγκ να προσπαθήσει να εντοπίσει ποιος παράγοντας ήταν αυτός που αύξησε τώρα και τότε τις συναλλαγές και το ύψος τους, διαφορετικά το κατάστημα μπορεί να υποστεί κάποια ζημιά, αφού για ένα σημαντικό διάστημα (τύπου 40 ημερών) μπορεί να κάνει πάλι μόνο 60-80 συναλλαγές ημερησίως που όλες θα είναι χαμηλής και μεσαίας αξίας. Αυτή η συμπεριφορά θα μπορούσε να έχει καταστροφικές συνέπειες, αν είναι επαναλαμβανόμενη.

Συμπέρασμα: Προβληματικές συστάδες θα αποκαλούσαμε την 2 & 0.

c) Τέλος, προσαρτούμε στο γενικό σύνολο δεδομένων μας το οποίο περιέχει και όλες τις προηγούμενες πληροφορίες που έχουμε εξάγει, 5 στήλες για να γνωρίζουμε κάθε φορά σε ποια συστάδα ανήκει μια συναλλαγή.

Άσκηση 4.

a) Πλέον έχοντας εξάγει νέα πληροφορία και έχοντας αξιοποιήσει κι άλλα από τα δεδομένα μας είμαστε σε θέση να εκτελέσουμε ξανά τον αλγόριθμο Apriori για τα προϊόντα και τις συστάδες των καλαθιών, για να δούμε τους νέους κανόνες που θα παραχθούν.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
156	(pastry, whole milk)	(Cluster3)	0.043392	0.246019	0.043392	1.000000	4.064725	0.032717	inf
55	(pastry)	(Cluster3)	0.116109	0.246019	0.116109	1.000000	4.064725	0.087544	inf
168	(rolls/buns, pastry)	(Cluster3)	0.027335	0.246019	0.027335	1.000000	4.064725	0.020610	inf
162	(pastry, other vegetables)	(Cluster3)	0.029459	0.246019	0.029459	1.000000	4.064725	0.022211	inf
174	(soda, pastry)	(Cluster3)	0.027468	0.246019	0.027468	1.000000	4.064725	0.020710	inf
155	(Cluster3, whole milk)	(pastry)	0.063694	0.116109	0.043392	0.681250	5.867314	0.035996	2.772990
161	(Cluster3, other vegetables)	(pastry)	0.045117	0.116109	0.029459	0.652941	5.623503	0.024220	2.546804
179	(yogurt, other vegetables)	(whole milk)	0.056661	0.333466	0.029061	0.512881	1.538029	0.010166	1.368317
186	(root vegetables, other vegetables)	(whole milk)	0.061837	0.333466	0.030255	0.489270	1.467227	0.009634	1.305062
184	(whole milk, root vegetables)	(other vegetables)	0.063827	0.252521	0.030255	0.474012	1.877119	0.014137	1.421096
54	(Cluster3)	(pastry)	0.246019	0.116109	0.116109	0.471953	4.064725	0.087544	1.673885
166	(Cluster3, rolls/buns)	(pastry)	0.059315	0.116109	0.027335	0.460850	3.969105	0.020448	1.639415
109	(root vegetables)	(whole milk)	0.142251	0.333466	0.063827	0.448694	1.345546	0.016391	1.209009
14	(Cluster1)	(whole milk)	0.270303	0.333466	0.117569	0.434953	1.304341	0.027432	1.179609
122	(root vegetables)	(other vegetables)	0.142251	0.252521	0.061837	0.434701	1.721445	0.025915	1.322273
62	(Cluster4)	(whole milk)	0.161492	0.333466	0.069135	0.428102	1.283795	0.015283	1.165477
173	(Cluster3, soda)	(pastry)	0.066083	0.116109	0.027468	0.415663	3.579924	0.019795	1.512638
86	(tropical fruit)	(whole milk)	0.136943	0.333466	0.055202	0.403101	1.208821	0.009536	1.116661
104	(yogurt)	(whole milk)	0.182059	0.333466	0.073116	0.401603	1.204331	0.012405	1.113867
178	(yogurt, whole milk)	(other vegetables)	0.073116	0.252521	0.029061	0.397459	1.573963	0.010597	1.240545

Πλέον μπορούμε βάσει αυτόν των κανόνων να συσχετίσουμε συστάδες με προϊόντα και αυτό ακριβώς θα κάνουμε, σύμφωνα με

τους 5 κανόνες που έχουν το υψηλότερο confidence και το δεξί ή αριστερό τους μέρος είναι κάποιος Cluster.

Τα συμπεράσματα μας ήταν τα εξής:

- Τα προϊόντα και οι συνδυασμοί που αγοράζονται κυρίως από την Συστάδα 0 είναι rolls/buns, yogurt, other vegetables και whole milk. Ωστόσο, να αναφέρουμε πως δεν μπορούμε να έχουμε μεγάλη εμπιστοσύνη στους κανόνες από τους οποίους αντλούμε αυτή τη γνώση, γιατί πρόκειται για μερικούς από τους κανόνες με το μικρότερο confidence, μόλις 0.1.
- Τα προϊόντα και οι συνδυασμοί που αγοράζονται κυρίως από την Συστάδα 1 είναι bottled water, sausage, whole milk, other vegetables και citrus fruit.
- Τα προϊόντα και οι συνδυασμοί που αγοράζονται κυρίως από την Συστάδα 2 είναι bottled water, soda, rolls/buns, yogurt, whole milk, other vegetables. Πάλι αυτή η γνώση αντλήθηκε από κανόνες χαμηλού confidence.
- Τα προϊόντα και οι συνδυασμοί που αγοράζονται κυρίως από την Συστάδα 3 είναι pastry και συνδυασμοί που τα εμπεριέχουν όπως: {pastry,soda}, {pastry, rolls/buns}, {whole milk, pastry}, {other vegetables, pastry}.
- Τα προϊόντα και οι συνδυασμοί που αγοράζονται κυρίως από την Συστάδα 4 είναι whole milk, sausage, root vegetables, tropical fruit, yogurt, citrus fruit, other vegetables.

Δεν αναφέρονται όλα τα προϊόντα που αντιστοιχούν σε κάθε ομάδα, αλλά μόνο αυτά που συναντήσαμε πρώτα, επομένως αυτά με κανόνες που έχουν το μεγαλύτερο confidence για τη κάθε ομάδα.

Όσον αναφορά τις προβληματικές ομάδες που αναφέραμε προηγουμένως (0 & 2), σχετίζονται με το προϊόν yogurt. Καταλήξαμε σε αυτό το συμπέρασμα παίρνοντας τις τομές από τα σύνολα των προϊόντων των προβληματικών ομάδων με την ένωση των συνόλων των λειτουργικών ομάδων. Το μόνο προϊόν που άνηκε μόνο σε προβληματικές ομάδες ήταν το yogurt.

Λόγοι για τους οποίους μπορεί να προκλήθηκε αυτή τη δυσλειτουργία:

- Αύξηση των τιμών των υπόλοιπων προϊόντων με αποτέλεσμα να μην τα αγοράζει πολύς κόσμος εκείνο το διάστημα.
- Έλλειμα των προτιμητέων προϊόντων με αποτέλεσμα οι καταναλωτές να αγοράζουν το συγκεκριμένο προϊόν για να συμπληρώσουν τη διατροφή τους.
- Ταυτόχρονη αύξηση των τιμών των άλλων προϊόντων και μείωση της τιμής του yogurt.
- Έξυπνη προώθηση του yogurt μέσω διαφημίσεων και influencing για ένα πιο υγιεινό τρόπο ζωής, με αποτέλεσμα οι καταναλωτές να αποφύγουν κάποια άλλα προϊόντα.
- Μποϊκοτάζ της εταιρίας προμήθευσης των προϊόντων του καταστήματος που δεν ταυτίζεται με την εταιρία προμήθευσης του yogurt.

*Θεωρώντας πως στο διάστημα εκείνο δεν έγινε κάποια άλλη πολύ μεγαλύτερη αλλαγή που δε θα μπορούσαμε να προβλέψουμε, ανατίμηση του νομίσματος, πόλεμος, κάποια φυσική καταστροφή μεγάλης κλίμακας κτλ. που θα μπορούσαν να προκαλέσουν έλλειψη πρώτων υλών για τη παραγωγή άλλων προϊόντων ή δυσκολία εισαγωγής τους από το εξωτερικό.

b) Τώρα το μόνο που μας μένει για να αξιοποιήσουμε όλη τη γνώση που έχουμε εξάγει είναι να προσαρτήσουμε στο σύνολο δεδομένων του προηγούμενου ερωτήματος και την διακριτή αξία καλαθιού και να εκτελέσουμε ξανά τον αλγόριθμο Apriori.

Το αποτέλεσμα το οποίο θα λάβουμε θα είναι το εξής:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
498	(Cluster2, rolls/buns)	(medium_basket_value)	0.040870	0.335722	0.040870	1.0	2.978656	0.027149	inf
620	(high_basket_value, pastry, other vegetables)	(Cluster3)	0.027203	0.246019	0.027203	1.0	4.064725	0.020510	inf
572	(pastry, whole milk)	(Cluster3)	0.043392	0.246019	0.043392	1.0	4.064725	0.032717	inf
589	(pastry, soda)	(Cluster3)	0.027468	0.246019	0.027468	1.0	4.064725	0.020710	inf
132	(pastry)	(Cluster3)	0.116109	0.246019	0.116109	1.0	4.064725	0.087544	inf
505	(Cluster2, bottled water)	(medium_basket_value)	0.027070	0.335722	0.027070	1.0	2.978656	0.017982	inf
449	(rolls/buns, sausage)	(high_basket_value)	0.039942	0.311438	0.039942	1.0	3.210908	0.027502	inf
294	(high_basket_value, pastry)	(Cluster3)	0.076433	0.246019	0.076433	1.0	4.064725	0.057629	inf
605	(high_basket_value, Cluster3, whole milk)	(pastry)	0.036624	0.116109	0.036624	1.0	8.612571	0.032372	inf
606	(high_basket_value, pastry, whole milk)	(Cluster3)	0.036624	0.246019	0.036624	1.0	4.064725	0.027614	inf
293	(high_basket_value, Cluster3)	(pastry)	0.076433	0.116109	0.076433	1.0	8.612571	0.067559	inf
455	(rolls/buns, pastry)	(high_basket_value)	0.027335	0.311438	0.027335	1.0	3.210908	0.018822	inf
466	(yogurt, sausage)	(high_basket_value)	0.025610	0.311438	0.025610	1.0	3.210908	0.017634	inf
481	(sausage, soda)	(high_basket_value)	0.031714	0.311438	0.031714	1.0	3.210908	0.021837	inf
578	(pastry, other vegetables)	(Cluster3)	0.029459	0.246019	0.029459	1.0	4.064725	0.022211	inf
619	(high_basket_value, Cluster3, other vegetables)	(pastry)	0.027203	0.116109	0.027203	1.0	8.612571	0.024044	inf
521	(medium_basket_value, pastry)	(Cluster3)	0.039676	0.246019	0.039676	1.0	4.064725	0.029915	inf
584	(rolls/buns, pastry)	(Cluster3)	0.027335	0.246019	0.027335	1.0	4.064725	0.020610	inf
631	(high_basket_value, Cluster3, rolls/buns)	(pastry)	0.027335	0.116109	0.027335	1.0	8.612571	0.024162	inf
640	(rolls/buns, pastry)	(high_basket_value, Cluster3)	0.027335	0.076433	0.027335	1.0	13.083333	0.025246	inf

Παράγονται 207 σύνολά αντικειμένων αυτή τη φορά, ξεπερνώντας κατά το πολύ το μέγεθος όλων των προηγούμενων συνόλων και με αυτά παράγουμε 644 κανόνες συνολικά. Αρκετούς παραπάνω δηλαδή από τις προηγούμενες φορές, λογικό βέβαια εφόσον πλέον χρησιμοποιούμε όλες τις νέες μεταβλητές που έχουμε παράγει.

Πρώτο πράγμα το οποίο τραβάει τη προσοχή μας είναι πως οι πρώτοι 20 κανόνες που παράγοντες έχουν όλοι confidence 1. Επίσης, πλέον βλέπουμε κανόνες που επιβεβαιώνουν πράγματα που έχουμε υποθέσει/εξάγει και αναφέρει παραπάνω.

Για παράδειγμα: Κάναμε την υπόθεση πως στη συστάδα 3 ανήκουν οι περισσότερες συναλλαγές υψηλής αξίας και επίσης προβλέψαμε κάποια προϊόντα που συσχετίζονται με αυτή την συστάδα και τώρα έχουμε κανόνες που συνδυάζουν αυτές τις πληροφορίες. Δηλαδή αν έχουμε high_basket_value μαζί με κάποια από τα συσχετιζόμενα προϊόντα της η συναλλαγή αυτή θα ανήκει πάντα στη Συστάδα 3.

Επίσης, πράγμα που παρατηρούμε με την πλέον αύξηση της πληροφορίας είναι πως για συγκεκριμένες μεταβλητές που προηγουμένως είχαμε λίγους κανόνες χαμηλού confidence, τώρα μπορεί να έχουν προκύψει περισσότεροι και ορισμένοι μεγάλου confidence. Επομένως, πλέον θα μπορούμε να αποφανθούμε καλύτερα για αυτές τις μεταβλητές. Για παράδειγμα, προηγουμένως προσπαθούσαμε να συσχετίσουμε την Συστάδα 2 με κάποια προϊόντα, όμως είχαμε μόνο κανόνες χαμηλού confidence, ενώ τώρα για παράδειγμα μπορούμε να συμπεράνουμε πως αν μια συναλλαγή ανήκει στη Συστάδα 2 και περιέχει το αγαθό bottled water ή rolls/buns, τότε είναι συναλλαγή μεσαίας αξίας. Επίσης, από το κανόνα με id 59 μαθαίνουμε με confidence ίσο με 0.6 πως αν μια συναλλαγή ανήκει στην Συστάδα 0 τότε είναι συναλλαγή χαμηλής αξίας, άλλο ένα στοιχείο πως η Συστάδα 0 είναι προβληματική.

Προς την Ομάδα Μάρκετινγκ: Πλέον υπάρχουν αρκετοί κανόνες για να αναλύσουν και να ανακαλύψουν τα πιο προτιμητέα προϊόντα που σχετίζονται τόσο με καλάθια μεγάλης αξίας, με τις πιο λειτουργικές ομάδες, αλλά και με το συνδυασμό αυτών των δύο. Ύστερα το μόνο που μένει είναι η κατάλληλη προώθηση τους με σκοπό την αύξηση του κέρδους του καταστήματος.

Τέλος εργασίας,
Μάρκος Κολέτσας – 3557.