DATA MINING ON AMAZON CUSTOMER REVIEWS DATASET

Βάρνα Μουράτ

Κυριακόπουλος Κωνσταντίνος

Μπαϊρακτάρης Φώτης

Παρουσίαση στηλών του dataset

```
DATA COLUMNS:
marketplace       - 2 letter country code of the marketplace where the review was written.
customer_id       - Random identifier that can be used to aggregate reviews written by a single author.
review_id         - The unique ID of the review.
product_id        - The unique Product ID the review pertains to. In the multilingual dataset the reviews
                    for the same product in different countries can be grouped by the same product_id.
product_parent    - Random identifier that can be used to aggregate reviews for the same product.
product_title     - Title of the product.
product_category  - Broad product category that can be used to group reviews
                    (also used to group the dataset into coherent parts).
star_rating       - The 1-5 star rating of the review.
helpful_votes     - Number of helpful votes.
total_votes       - Number of total votes the review received.
vine              - Review was written as part of the Vine program.
verified_purchase - The review is on a verified purchase.
review_headline   - The title of the review.
review_body       - The review text.
review_date       - The date the review was written.
```

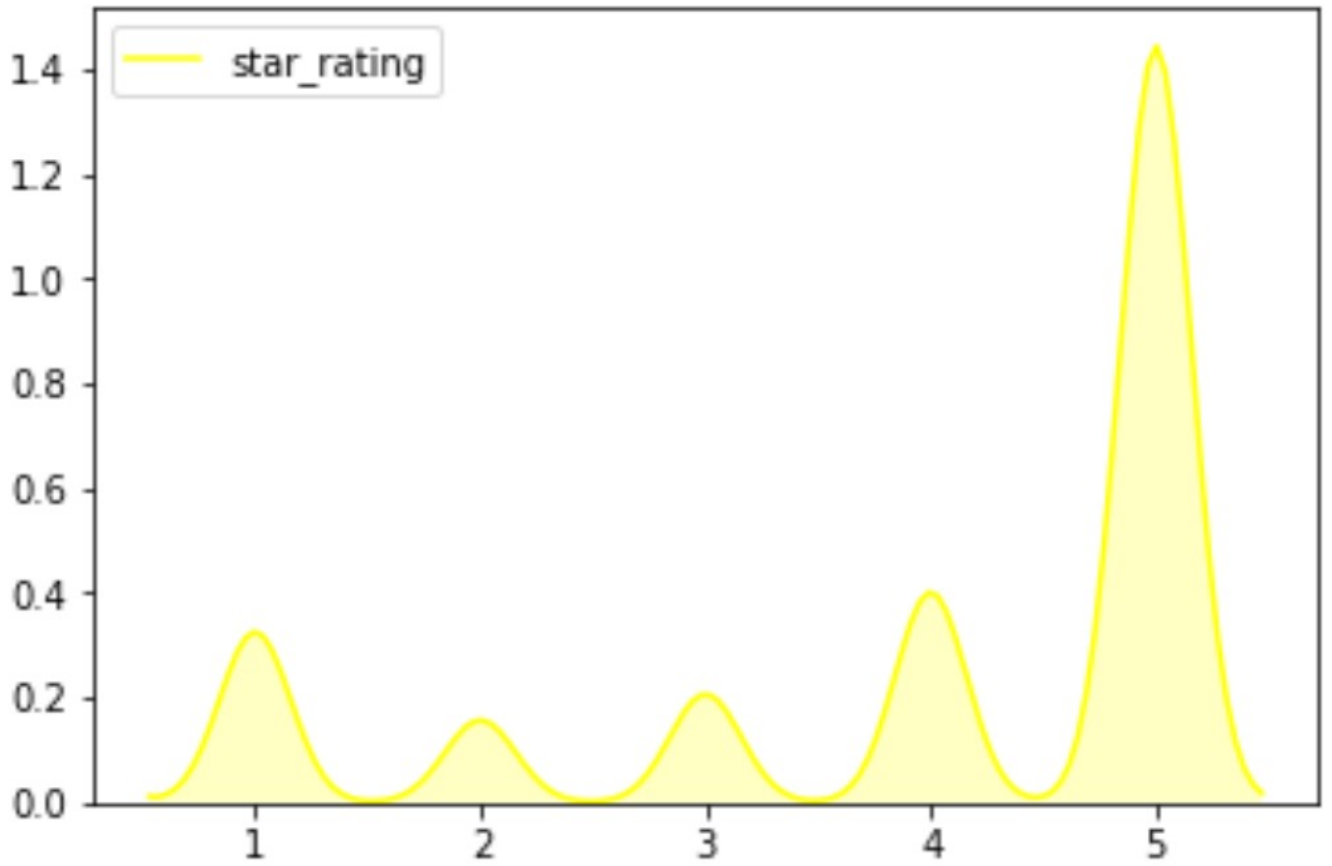
Από τις στήλες του dataset αφαιρέσαμε 4 (marketplace , customer_id , review_id , product_id) και προσθέσαμε μία (Review_Length - δείχνει το πλήθος χαρακτήρων κάθες κριτικής) .

	customer_id	product_parent	product_title	star_rating	helpful_votes	total_votes	vine	verified_purchase	review_headline	review_body	review_date	Review_Length
0	32114233	223980852	Elite Sportz Exercise Sliders are Double Sided...	5	0	0	0	1	Good quality. Shipped	Exactly as described. Good quality. Shipped fast	2015-08-31	48
1	18125776	819771537	Ezy Dose Weekly	5	0	0	0	1	Five Stars	It is great	2015-08-31	11
2	19917519	849307176	Pulse Oximeter, Blood Oxygen Monitor	5	1	1	0	1	It\'s really nice it works great	It\'s really nice it works great. You have the ...	2015-08-31	105
3	18277171	700864740	SE Tools Tool Kit Watch Watch Repair Kit (20 P...	2	0	0	0	1	Two Stars	The kit works fine... simple cheap plastic tho	2015-08-31	46
4	2593270	794298839	doTERRA HD Clear Facial Kit - Facial Lotion, F...	4	0	1	0	1	Four Stars	It works better than anything else ive tried	2015-08-31	44

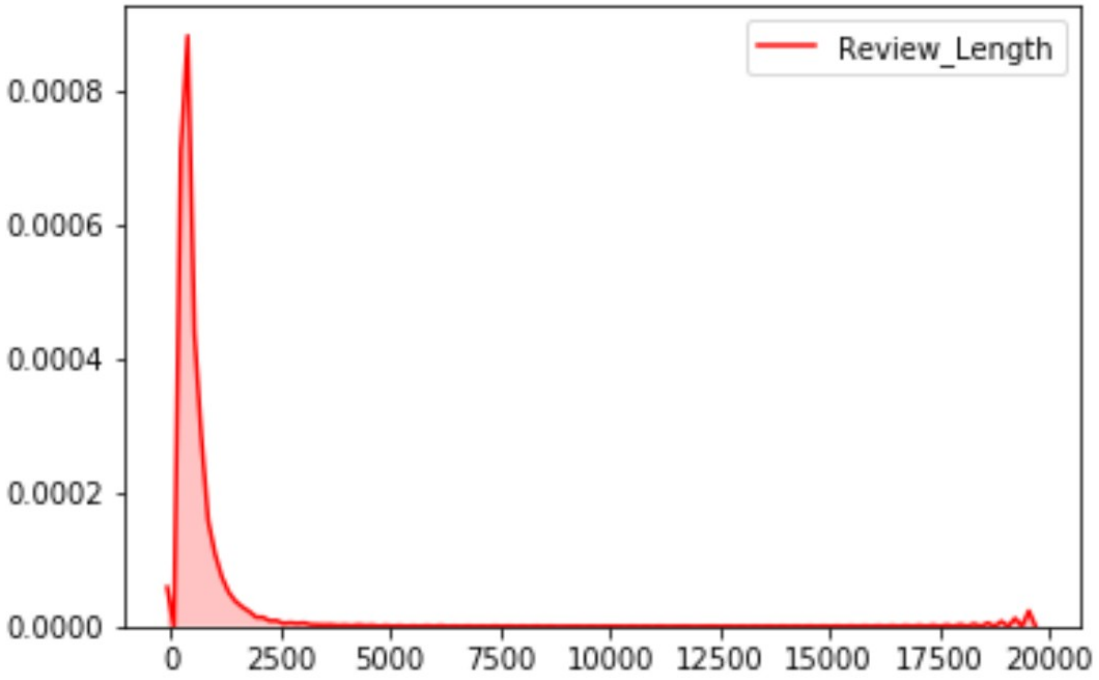
- Κατασκευάσαμε νέο Dataframe από τα δεδομένα μας
- Κάθε γραμμή αντιστοιχεί σε διαφορετικό προϊόν
- Οι στήλες αντιστοιχούν στη μέση τιμή και τυπική απόκλιση του star rating καθώς επίσης και στον αριθμό κριτικών που αντιστοιχούν στο καθένα

	Product	Mean	SD	Number
0	Elite Sportz Exercise Sliders are Double Sided...	4.647436	0.868276	156
1	Ezy Dose Weekly	3.970588	1.484930	238
2	Pulse Oximeter, Blood Oxygen Monitor	4.391003	1.134267	578
3	SE Tools Tool Kit Watch Watch Repair Kit (20 P...	3.414634	1.396435	41
4	doTERRA HD Clear Facial Kit - Facial Lotion, F...	3.937500	1.390537	16
5	Viva Naturals #1 Best Selling Certified Organi...	4.820114	0.582116	1223
6	Uncle Lee's Organic Green Tea -- 100 Tea Bags ...	4.577465	0.988126	213

- Παρακάτω φαίνεται η κατανομή πυκνότητας πιθανότητας συναρτήσει του star rating.
- Παρατηρούμε ότι η πλειοψηφία των βαθμολογιών είναι στο 5 και ακολουθεί το 4.
- Η χαμηλότερη δυνατή βαθμολογία έχει μεγαλύτερη πιθανότητα από τις μέτριες (2-3) .
- Φαίνεται να υπάρχει η τάση συγκέντρωσης των βαθμολογιών σε σχετικά ακραίες τιμές.

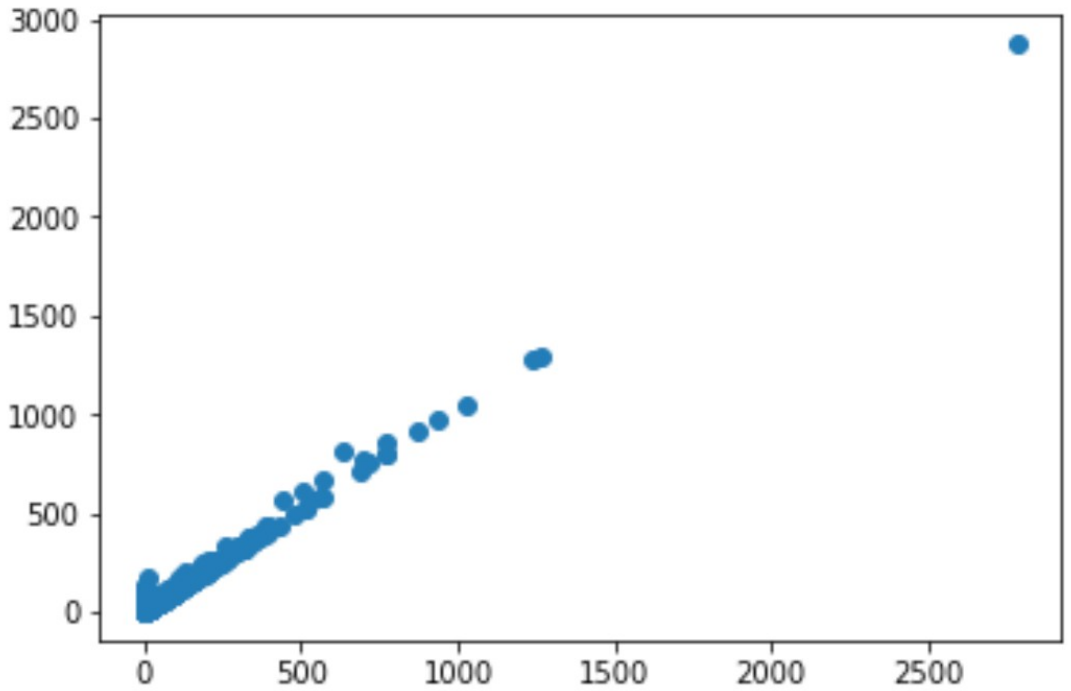


- Η παρακάτω κατανομή δείχνει την κατανομή πυκνότητας πιθανότητας του μήκους των κριτικών .
- Η συντριπτική πλειοψηφία γράφει κριτικές μέχρι το πολύ 500 χαρακτήρες (~80 λέξεις).
- Υπάρχει ένα πολύ μικρό ποσοστό κριτικών που είναι πολύ εκτενείς.



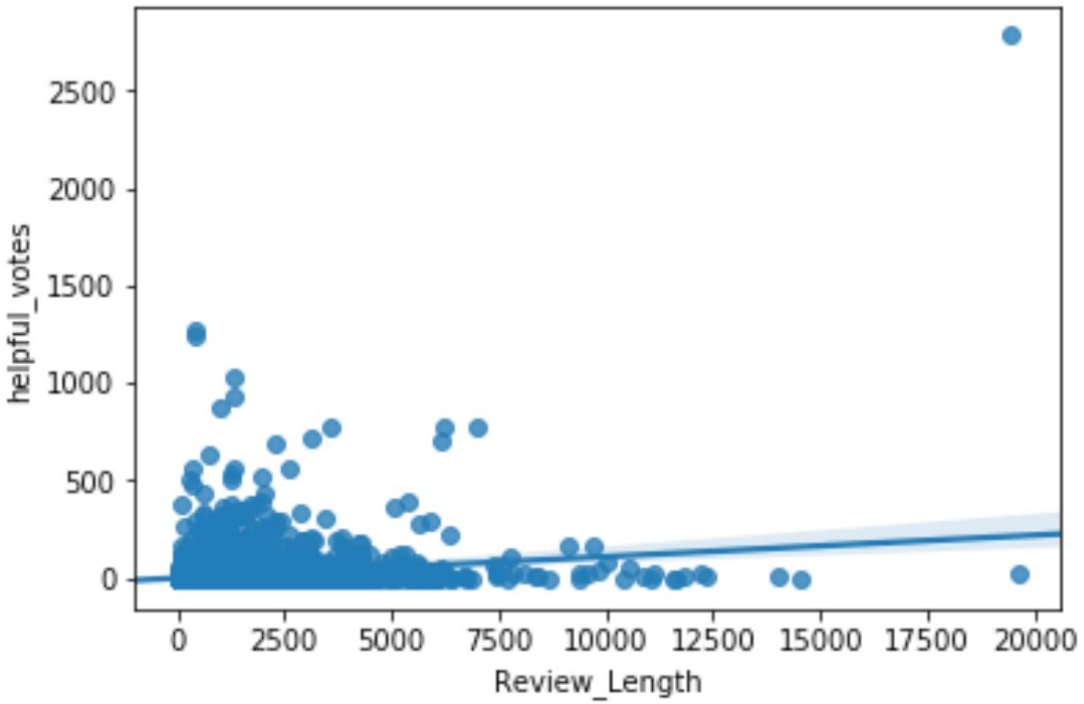
	Review_Length
count	85919.000000
mean	343.354264
std	480.901350
min	1.000000
25%	115.000000
50%	205.000000
75%	408.000000
max	19628.000000

- Το παρακάτω scatterplot μας δείχνει τη συσχέτιση μεταξύ των κριτικών που έχουν ψηφιστεί ως χρήσιμες (helpful votes) σε σχέση με τις συνολικές κριτικές (total votes)
- Υπάρχει πολύ ισχυρή γραμμική συσχέτιση μεταξύ τους, που φαίνεται και από τον πίνακα συσχέτισης των δύο μεταβλητών
- Μπορούμε να συμπεράνουμε ότι οι περισσότεροι που ψηφίζουν μία κριτική, το κάνουν μόνο εάν τη βρίσκουν χρήσιμη



	helpful	total
helpful	1.000000	0.991686
total	0.991686	1.000000

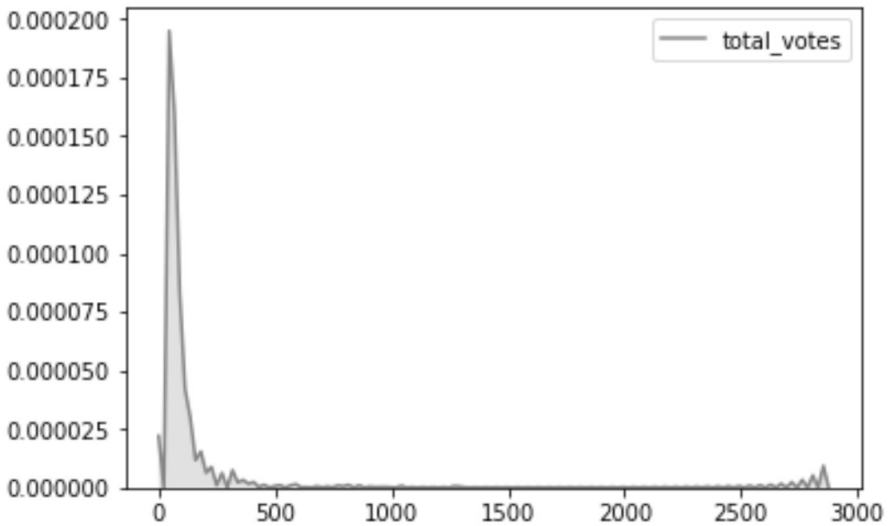
- Η γραφική παράσταση που απεικονίζεται εδώ είναι συσχέτιση του αριθμού των θετικών ψήφων σε μία κριτική συναρτήσει του μήκους χαρακτήρων των κριτικών.
- Φαίνεται να υπάρχει μία ασθενής θετική συσχέτιση, που εντείνεται κυρίως στις εκτενείς κριτικές (>5000 χαρακτήρες).
- Μία κριτική δηλαδή έχει μεγαλύτερη πιθανότητα να αναγνωριστεί ως βοηθητική (helpful) , εάν είναι εκτενής.

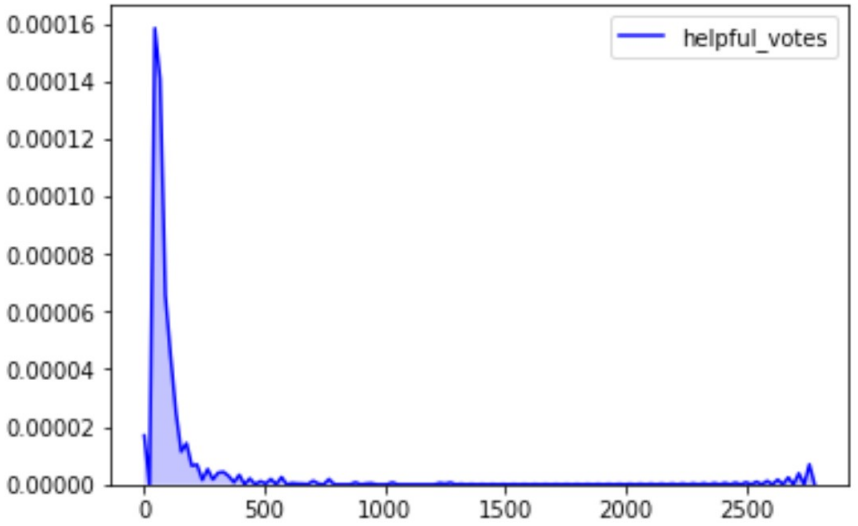


- Στον παρακάτω πίνακα διακρίνουμε ότι ένας μικρός αριθμός κριτικών συγκεντρώνουν πάρα πολλές ψήφους και κατά συνέπεια πολλές θετικές ψήφους.

- Οι κατανομές των συνολικών (total_votes) και θετικών (helpful_votes) κριτικών είναι πανομοιότυπες, που είναι ένα αναμενόμενο αποτέλεσμα λόγω της ισχυρής συσχέτισης που έχουν μεταξύ τους.

	total_votes	helpful_votes
count	85919.000000	85919.000000
mean	4.061418	3.351540
std	20.833263	19.490005
min	0.000000	0.000000
25%	0.000000	0.000000
50%	1.000000	0.000000
75%	3.000000	2.000000
max	2876.000000	2785.000000

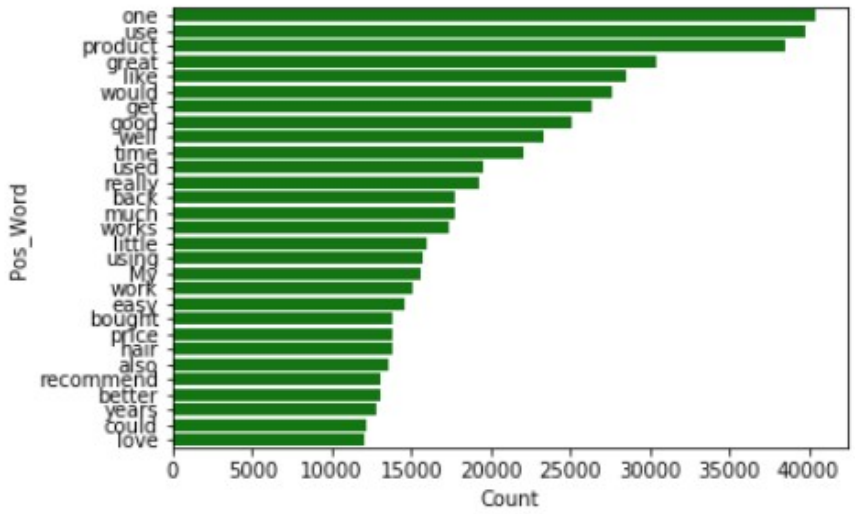


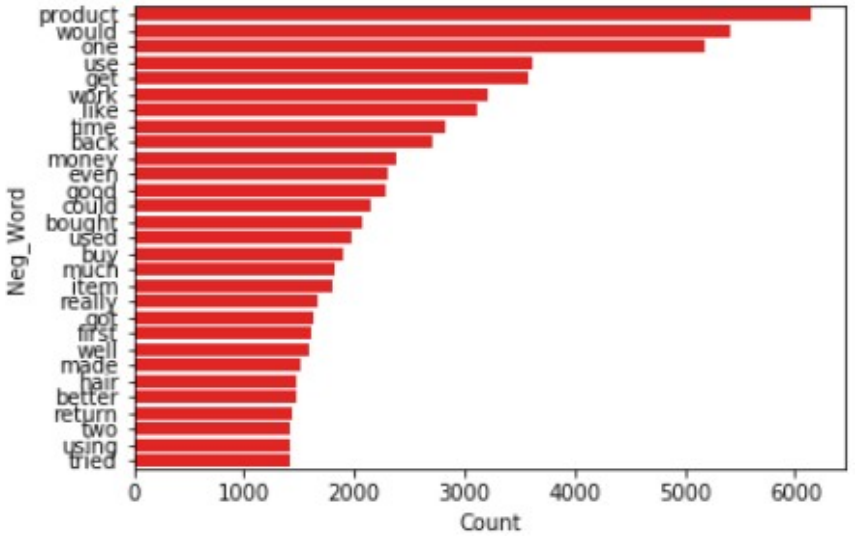


- Διαχωρίσαμε το Dataframe σε καλές (star rating >2.5) και κακές (star rating <2.5) κριτικές.
- Με χρήση του πακέτου nltk κατασκευάσαμε dataframe με τις συχνότερα εμφανιζόμενες λέξεις.

	Pos_Word	Count
196	one	40420
14	use	39848
38	product	38560
6	great	30482
153	like	28576
44	would	27694
232	get	26422
163	good	25110
186	well	23358

	Neg_Word	Count
28	product	6156
32	would	5407
91	one	5173
50	use	3612
85	get	3588
23	work	3220
135	like	3121
26	time	2818
43	back	2719





- Εξετάζουμε τις διαφορές ανάμεσα σε verified purchase και non verified.
- Στα verified δεν υπάρχουν αρνητικές ψήφοι σε reviews.
- Στα non-verified είναι όλες οι αρνητικές.
- Φαίνεται ότι υπάρχει μεγαλύτερη εμπιστοσύνη σε reviewers που αποδεδιγμένα έχουν αγοράσει το προϊόν.
- Οι vine reviewers ήταν πολύ λίγοι (32) στο dataset, ώστε να επηρεάσουν κάτι. Ήταν όλοι σε non verified κριτικές, κάτι που μπορεί να δείχνει ότι κάποιος ίσως να εκμεταλλεύεται το ότι είναι vine reviewer για εμπορικούς/διαφημιστικούς σκοπούς.
- Αριστερά τα verified purchases.

	total_votes	helpful_votes
total_votes	1.0	1.0
helpful_votes	1.0	1.0

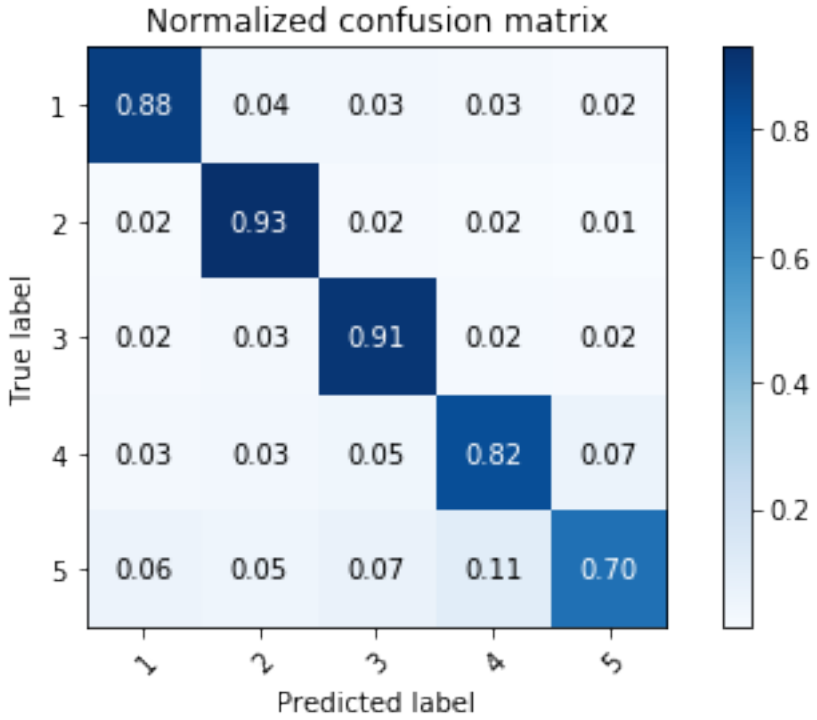
	total_votes	helpful_votes
count	63258.000000	63258.000000
mean	2.978311	2.430823
std	15.526632	14.371662
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	2.000000	2.000000
max	1294.000000	1266.000000

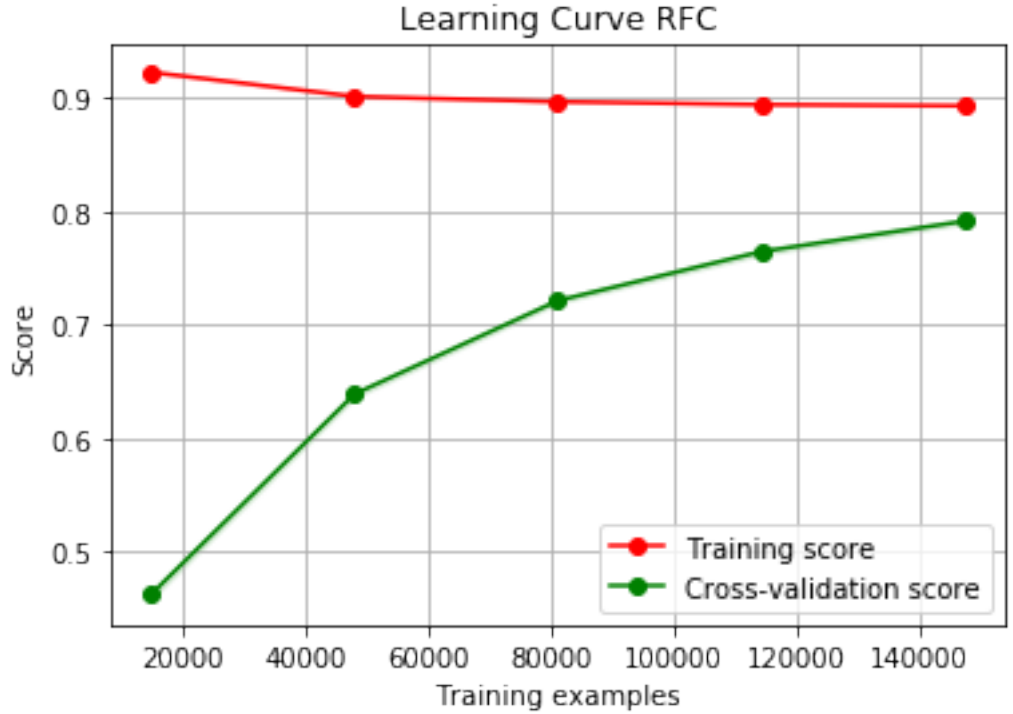
	total_votes	helpful_votes
total_votes	1.000000	0.999992
helpful_votes	0.999992	1.000000

	total_votes	helpful_votes
count	22661.000000	22661.000000
mean	7.084904	5.921716
std	30.988010	29.235721
min	0.000000	0.000000
25%	0.000000	0.000000
50%	2.000000	2.000000
75%	6.000000	4.000000
max	2876.000000	2785.000000

Πρόβλεψη

- Θέλουμε να προβλέψουμε το βαθμό (σε αστέρια) ενός review από τις άλλες παραμέτρους.
- Χρησιμοποιούμε τις συχνές λέξεις σε κάθε βαθμολογία.
- Επειδή το score αρχικά δεν είναι ικανοποιητικό και τα δεδομένα δεν είναι ισοκατανεμημένα→ Random Over Sampler .
- Με Random Forest τα καλύτερα αποτελέσματα.


Normalized confusion matrix
True label
Predicted label
1
2
3
4
5
0.88 0.04 0.03 0.03 0.02
0.02 0.93 0.02 0.02 0.01
0.02 0.03 0.91 0.02 0.02
0.03 0.03 0.05 0.82 0.07
0.06 0.05 0.07 0.11 0.70
0.8
0.6
0.4
0.2


Learning Curve RFC
Score
Training examples
0.9
0.8
0.7
0.6
0.5
20000
40000
60000
80000
100000
120000
140000
Training score
Cross-validation score

Συμπερασματικά

- Εξετάσαμε μεταβλητές που μπορεί να επηρεάζουν το βαθμό μίας κριτικής.
- Σημαντικός παράγοντας που εισάγαμε ως μεταβλητή είναι οι λέξεις κλειδιά που περιέχει.
- Το μοντέλο είχε καλή απόδοση μετά από Over Sampling, με Random Forest Classifier (η απόδοση των υπολοίπων έπεφτε με αυτή τη μέθοδο, αλλά είχαμε κοντά στο 80% στον RFC).
- Το βέλτιστο πάντα είναι η συλλογή περισσότερων δεδομένων, αλλά δεν είναι πάντα εφικτό.
- Χάσαμε λίγη προβλεπτική ικανότητα στη μεγάλη κλάση (5 αστέρια) αλλά κερδίσαμε πολύ στις υπόλοιπες.