Assignment 2-Data Analysis Report

Ομάδα Α:
Νίκος-Γεράσιμος Ζαζάτης 2723
Κυριακή Κόφφα 2818
Ηλίας Τσιρώνης 2851

Στο πρώτο κομμάτι της εργασίας μας ζητήθηκε να κάνουμε ανάλυση των δεδομένων που δίνονται πάνω στα οποία θα κάνουμε αργότερα εκπαίδευση του νευρωνικού

Viewer Sampling

Αρχικά ομαδοποιήσαμε τα δεδομένα βάσει το event_id τους και κάναμε τα ιστογράμματα των μέσων όρων των τιμών Engagement και QoE. Από τα ιστογράμματα που προκύπτουν βλέπουμε πως τα περισσότερα events έχουν μέσο όρο Engagement στο διάστημα [0.4, 0.5], οπότε θεωρούμε πως τα πιο αντιπροσωπευτικά event για Engagement θα βρίσκονται σε αυτό το διάστημα, ενώ αντίστοιχα για QoE στο διάστημα [0.9, 1.0]. Έπειτα πήραμε τα ιστογράμματα των τυπικών αποκλίσεων και παρατηρούμε πως το Engagement έχει μεγαλύτερη τυπική απόκλιση (0.5) από ότι το QoE (0.05), πράγμα που σημαίνει πως στο Engagement τα δεδομένα φαίνεται να τείνουν πολύ είτε προς το ένα άκρο (0.0) είτε προς το άλλο (0.9-1.0) με πολύ λίγες τιμές να βρίσκονται ανάμεσα στα δύο άκρα. Το αντίθετο συμβαίνει με το QoE, για το οποίο φαίνονται οι περισσότερες τιμές να βρίσκονται κοντά στο μέσο όρο, δηλαδή στο [0.9-1.0].

Στη συνέχεια από τα events που βρίσκονται στο διάστημα που θεωρήσαμε πιο αντιπροσωπευτικό για τα δύο στοιχεία επιλέξαμε 3 τυχαία events ξεχωριστά. Για το κάθε ένα από αυτά τα event πήραμε το ιστόγραμμα των μέσων όρων Engagement (και QoE για τα άλλα τρία εκείνης της κατηγορίας) των θεατών. Στο κομμάτι του Engagement και στα 3 τυχαία event που πήραμε βρήκαμε πως παρουσιάζουν παρόμοια συμπεριφορά, με το περισσότερο Engagement να βρίσκεται στο τμήμα [0.0, 0.1]. Οπότε αποφασίσαμε να χωρίσουμε τους viewers σε τρεις ομάδες, αυτούς που έχουν μέσο όρο Engagement στο διάστημα [0.0, 0.1], στο διάστημα [0.1, 0.7] και στο διάστημα [0.7, 1.0] στο κάθε event. Αντίστοιχα για το QoE με τους περισσότερους χρήστες να βρίσκονται στο διάστημα [0.9, 1.0] επιλέξαμε από αυτούς που πέφτουν στο διάστημα QoE άνω του 0.999 και QoE άνω του 0.8. Από αυτές τις ομάδες επιλέγουμε 3 τυχαίους θεατές τη φορά ως αντιπροσωπευτικούς της ομάδας τους, και της συμπεριφοράς ενός θεατή αυτής της ομάδας σε ένα οποιοδήποτε event. Στο κομμάτι του engagement παίρνουμε τα γραφήματα του Engagement του θεατή overtime όπως και το engagement level

duration τους σε εκείνο το event. Αντίστοιχα δουλεύουμε και για QoE overtime του θεατή κατά τη διάρκεια του event.

Engagement Level Over Countries/City/Type of viewer

Παίρνουμε το ιστόγραμμα μέσου όρου και τυπικής απόκλισης ανά χώρα και παρατηρούμε πως το engagement βρίσκεται περισσότερο στα διαστήματα [0.0, 0.1] και [0.35, 0.65] με μεγάλη τυπική απόκλιση. Στην τελική αποφασίζουμε να χρησιμοποιήσουμε τις χώρες στις οποίες γίνονται τουλάχιστον 100 event. Παρατηρούμε πως κάποιες χώρες παρουσιάζουν παρόμοια συμπεριφορά στην κατανομή τους ενώ άλλες εντελώς διαφορετική. Επιλέγουμε αυτές που έχουν διαφορετική συμπεριφορά μεταξύ τους. Αλλά και πάλι από τις χώρες που επιλέξαμε φαίνεται να έχουν τους περισσότερους θεατές σε μέσο όρο engagement στο διάστημα [0.0, 0.1] με λίγες διαφορές.

Ανάλογα δουλέψαμε και για τις πόλεις, με τη μόνη διαφορά ότι ο ελάχιστος αριθμός events που θα είχαν ήταν 50 αντί του 100 των χωρών. Οι πόλεις φαίνεται να έχουν τα περισσότερα events να έχουν μέσο όρο engagement στα διαστήματα [0.0, 0.1] και [0.4, 0.9] με μικρότερη τυπική απόκλιση από ότι στις χώρες.

Όσον αφορά τον τύπο θεατή πήραμε τα ιστογράμματα μέσων όρων και τυπικών αποκλίσεων για τις δύο κατηγορίες ξεχωριστά. Βλέπουμε πως έχουν σχεδόν ίδια κατανομή στο μέσο όρο τους αλλά η σημαντική διαφορά μεταξύ τους φαίνεται να εμφανίζεται στην τυπική απόκλιση, με τους θεατές που δουλεύουν από το σπίτι να εμφανίζουν τη μεγαλύτερη.

Correlation between Data

Αρχικά βρίσκουμε τη συσχέτιση μεταξύ engagement, QoE και customer_id, event_id, viewer_id, city_id, country_id, viewer_type και week_days, όπως και τα δύο πρώτα μεταξύ τους, χρησιμοποιώντας t-test επειδή σε αυτά τα δεδομένα δεν έχει σημασία η γραμμική εξάρτησή τους αλλά το κατά πόσο οι κατανομές τους μοιάζουν. Παρατηρούμε πως υπάρχει αρκετή συσχέτιση του engagement με τον τύπο του θεατή, όπως και η πόλη του αν και σε μικρότερο βαθμό. Οι άλλες τιμές φαίνεται να παίζουν ακόμα μικρότερο ρόλο στο engagement, και συγκεκριμένα προς την έκπληξή μας η μέρα της εβδομάδας φαίνεται να έχει τη μικρότερη συσχέτιση.

Στην περίπτωση του QoE βρίσκουμε αρκετή συσχέτιση με την πόλη του θεατή, όπως και τη χώρα, σε λίγο λιγότερο βαθμό όμως. Τη μικρότερη συσχέτιση την έχει με τον τύπο του θεατή. Γενικά και στις δύο περιπτώσεις παρατηρούμε πως μετρικές όπως

το customer_id, event_id και viewer_id δεν παίζουν ρόλο την εμπειρία και συμμετοχή του θεατή.

Για συσχετίσεις μεταξύ Engagement, QoE και πλήθος θεατών σε ένα event, διάρκεια event και viewer retention χρησιμοποιήσαμε τον μαθηματικό τύπο του correlation αφού ψάχνουμε γραμμική εξάρτηση. Αυτό που βρήκαμε είναι πως το engagement έχει πιο ισχυρή συσχέτιση με τη διάρκεια του event όπως και το viewer retention από ότι με το πλήθος των θεατών στο event, ενώ το μικρότερο έχει με το QoE. Από την πλευρά του QoE, έχουμε ασθενείς συσχετίσεις με τα στοιχεία γενικά, αλλά από αυτά η ισχυρότερη συσχέτιση είναι με το viewer retention.

Στο dataset οι στήλες που προσθέσαμε ήταν ο αριθμός των viewers ανά event, viewer retention, day of event και event duration καθώς φαίνεται να παίζουν κάποιο ρόλο στο engagement και QoE σε σύγκριση με προϋπάρχουσες στήλες του πίνακα, ενώ αφαιρέσουμε το buffer_ms καθώς δε μας ζητείται από αυτήν την άσκηση. Μπορεί στην πορεία, κατά τη δημιουργία του training dataset να αφαιρέσουμε στήλες με χαμηλότερες συσχετίσεις.

Observations

Αυτό που παρατηρήσαμε κατα τη διάρκεια του sampling κομματιού της ανάλυσης των δεδομένων είναι πως οι χρήστες τείνουν να μένουν στο ίδιο ή σε παρόμοιο επίπεδο engagement κατά τη διάρκεια ενός event. Αυτό μπορεί να συμβαίνει εξαιτίας του ρόλου ενός θεατή ο οποίος μπορεί να είναι είτε παρουσιαστής, οπότε θα έχει αρκετό engagement με το event κατά τη διάρκειά του, ή ένας παθητικός θεατής, ο οποίος δεν έχει κάποιο ιδιαίτερο λόγο να συμμετάσχει, παρά για πολύ μικρά χρονικά διαστήματα.

Διαφορετικές χώρες και πόλεις μπορεί να εμφανίζουν αρκετά διαφορετικές κατανομές στο engagement και QoE αλλά φαίνεται να έχουν παρόμοιες συμπεριφορές. Αλλά όσο αφορά το QoE η πόλη και χώρα προκύπτει πως παίζουν σημαντικό ρόλο, το οποίο πιθανόν προκαλείται από τις διαφορετικές ποιότητες του δικτύου. Δεν έχουν όλες οι περιοχές του κόσμου τις ίδιες ταχύτητες και επομένως την ίδια ποιότητα εμπειρίας σε ένα event.

Όπως προαναφέραμε, ενώ οι μέσοι όροι engagement των θεατών που δουλεύουν από το σπίτι και αυτούς που δουλεύουν από το γραφείο είναι παρόμοιοι, οι τυπικές αποκλίσεις έχουν διαφορές, με αυτούς που δουλεύουν από το σπίτι να έχουν μεγαλύτερη απόκλιση στην κατανομή. Αυτό μπορεί να είναι αποτέλεσμα του γεγονότος ότι στο σπίτι δουλεύουν χωρίς κάποια επίβλεψη από προϊστάμενό τους, έτσι δεν έχουν αναγκαιότητα να συμμετάσχουν σταθερά και μόνιμα, σε αντίθεση με θεατές που δουλεύουν από το γραφείο.

Επίσης, events με θεατές με πολύ engagement έχουν μεγαλύτερη διάρκεια σε χρόνο από τα υπόλοιπα. Μπορούμε να υποθέσουμε πως αυτό προκαλείται από τη συμμετοχή των ίδιων των θεατών, η οποία μπορεί να είναι σε μορφή ερωτήσεων. Αυτή η συμμετοχή εκ φύσεως μεγαλώνει τη διάρκεια από αυτή που μπορεί να είχε προγραμματιστεί αρχικά.

Προσθέτοντας, το viewer retention φαίνεται να μεγαλώνει με το engagement το οποίο μπορεί να συμβαίνει γιατί ένας ενεργός θεατής έχει μεγαλύτερη προθυμία να παρακολουθήσει το event περισσότερη ώρα.

Τέλος, προκύπτει πως το engagement και QoE δεν επηρεάζουν το ένα το άλλο. Ένας χρήστης που δεν είχε σκοπό να συμμετάσχει σε ένα event δε πρόκειται να αλλάξει γνώμη εξαιτίας της ποιότητας του δικτύου για παράδειγμα, όπως και αντίστροφα. Η συμμετοχή του θεατή είναι ανεξάρτητη, όπως και η ποιότητα της εμπειρίας από αυτές τις μετρικές. Οι συσχετίσεις που μπορούμε να βρούμε είναι αυτές που έχουν αναφερθεί νωρίτερα.