ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάλυση Στοιχείων Χρήσης Ψηφιακού Αποθετηρίου

Ψηφιακό Αποθετήριο ΚΑΛΛΙΠΟΣ

ΕΙΡΗΝΗ ΔΟΝΤΗ
03119839

Επιβλέπων:
Νικόλαος Μήτρου

Περιεχόμενα

Σκοπός Έρευνας

- Μελέτη των στοιχείων χρήσης και συγγραμμάτων του αποθετηρίου ΚΑΛΛΙΠΟΣ.
- Ανάλυση μοτίβων συμπεριφοράς χρηστών κατά την περιήγησή τους στο αποθετήριο.
- Πρόβλεψη νέων χρηστών με παραδοσιακές τεχνικές στατιστικής και μηχανικής μάθησης.
- Ανάδειξη και προτάσεις αντιμετώπισης ελλείψεων στην οργάνωση του αποθετηρίου.

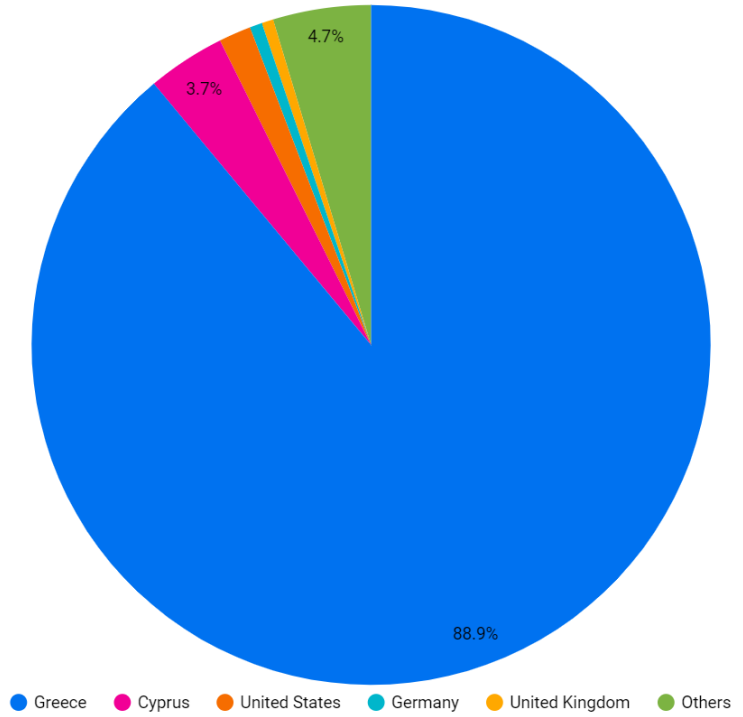
Αποθετήριο ΚΑΛΛΙΠΟΣ

Ανάλυση Δραστηριότητας Χρηστών

Συλλογή στοιχείων χρήσης από το **Google Analytics** σχετικά με:

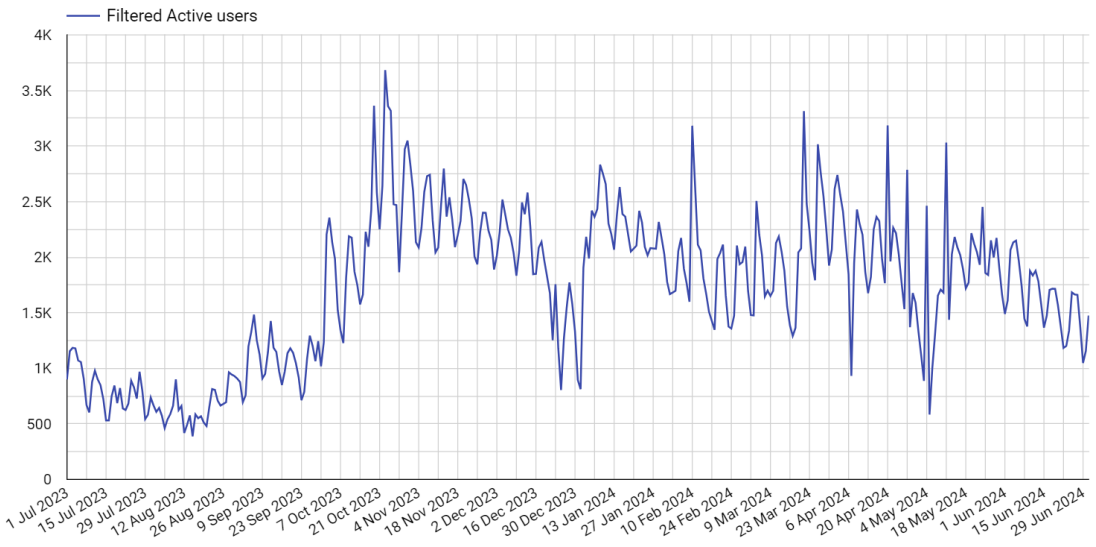
- Συμπεριφορά Χρηστών
  - Σελίδες που επισκέπτονται οι χρήστες
  - Χρόνος παραμονής στη σελίδα
  - Συχνότητα και χρόνος συνεδριών (Sessions)
- Απόκτηση Χρηστών (User acquisition)
  - Οργανική αναζήτηση
  - Απευθείας
  - Παραπομπή
  - Οργανική κοινωνική
  - Οργανικό βίντεο
  - Ηλεκτρονική Διεύθυνση
  - Χωρίς Ανάθεση
- Συμβάντα (Events)
  - Λήψεις και Προβολές συγγραμμάτων

Ανάλυση Δραστηριότητας Χρηστών

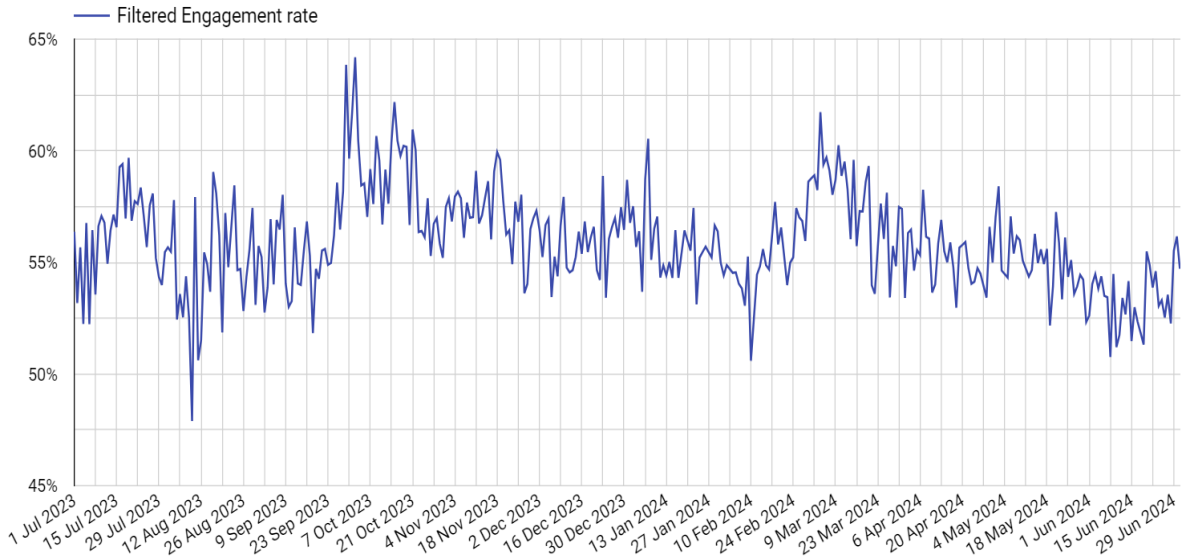


Γεωγραφική Κατανομή Ενεργών Χρηστών

Ανάλυση Δραστηριότητας Χρηστών

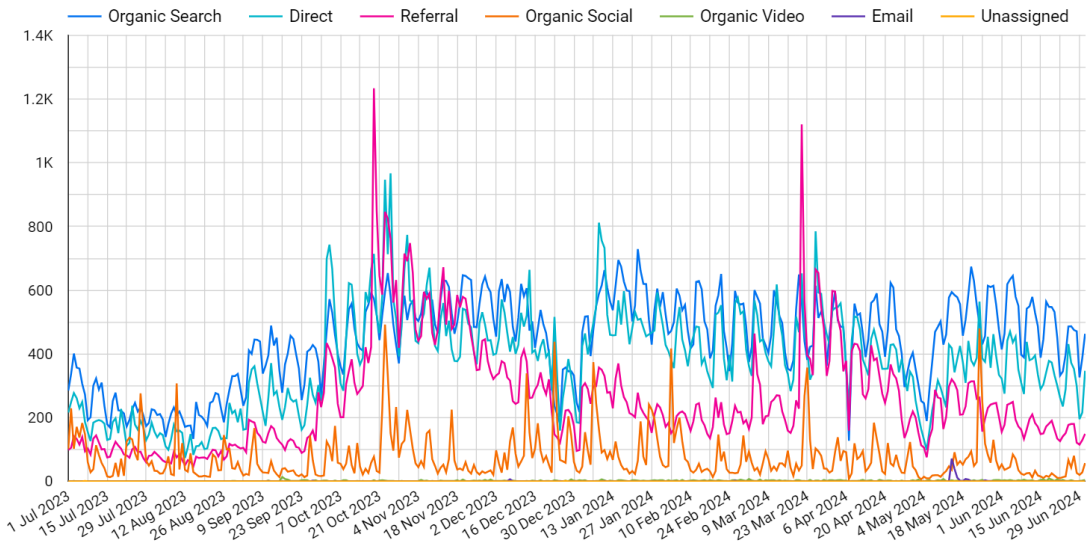


Ημερήσια Κατανομή Ενεργών Χρηστών

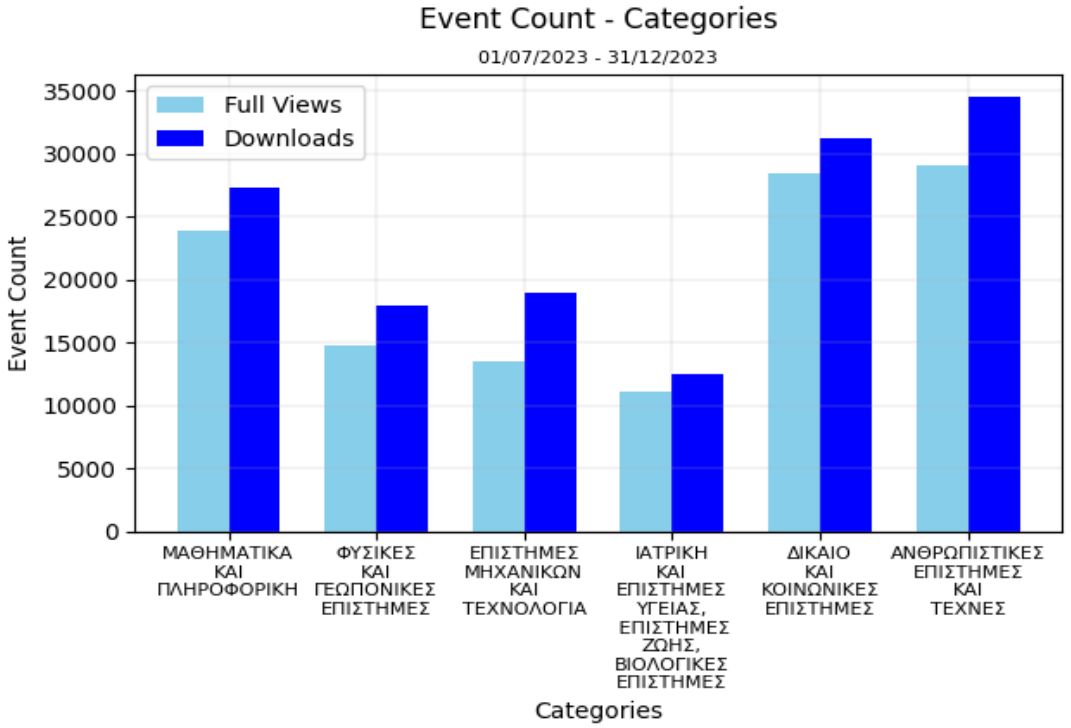


Ημερήσιο Ποσοστό Αφοσίωσης Χρηστών

Ανάλυση Δραστηριότητας Χρηστών

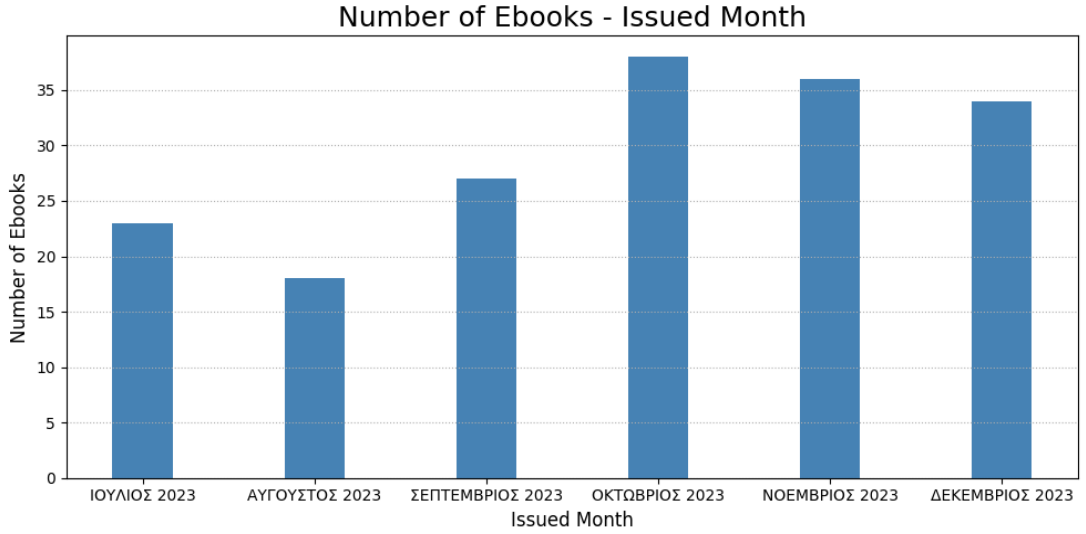


Ημερήσια Απόκτηση Χρηστών

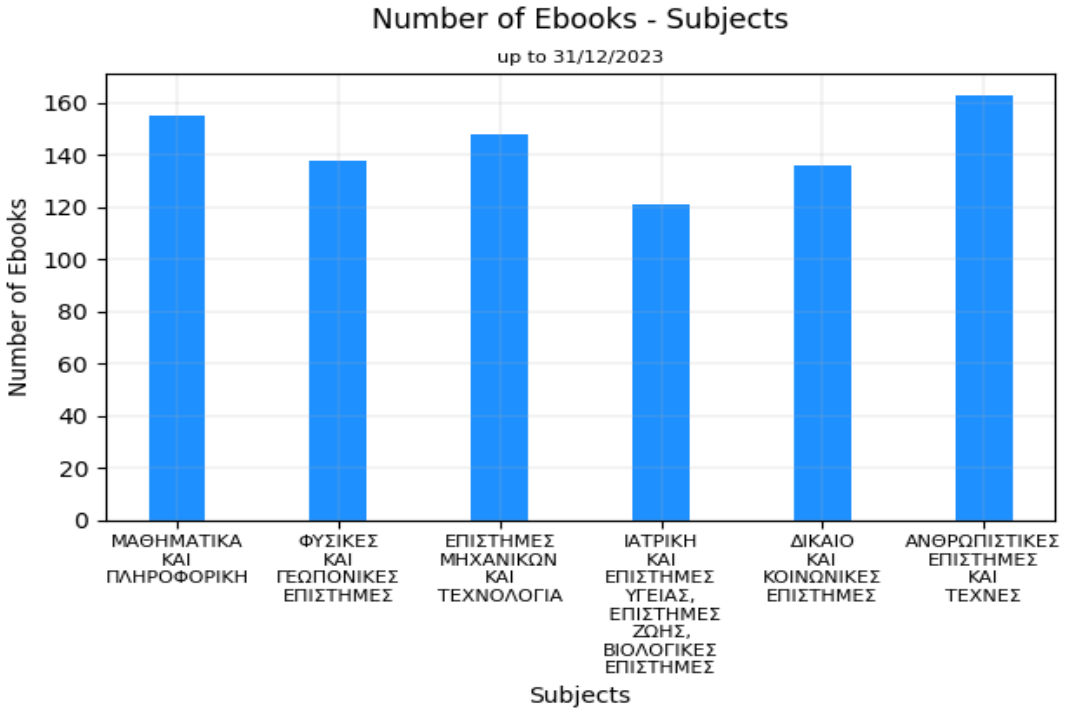


Προβολές & Λήψεις ανά Θεματική Ενότητα

Ανάλυση Δεδομένων Συγγραμμάτων



Συγγράμματα ανά Μήνα Δημοσίευσης



Συγγράμματα ανά Κύρια Θεματική Κατηγορία

Μεθοδολογία Πρόβλεψης Νέων Χρηστών

ΠΡΟΒΛΕΨΗ ΝΕΩΝ ΧΡΗΣΤΩΝ

SARIMAX

Εφαρμογή μοντέλου SARIMAX (1, 0, 1)(2, 1, 0, 52)

Exponential Smoothing

- Απλή Εκθετική Εξομάλυνση
- Διπλή Εκθετική Εξομάλυνση
- Τριπλή Εκθετική Εξομάλυνση

LightGBM

- LightGBM Regressor
- LightGBM Regressor και Καθυστερήσεις
- Επιλογή Αποδοτικών Παραμέτρων

Μεθοδολογία Πρόβλεψης Νέων Χρηστών

SARIMAX (p, d, q)(P, D, Q, s)

p: Σειρά Αυτοπαλινδρομικού (AR) μοντέλου

d: Αριθμός διαφορών για να σταθεροποιηθούν τα δεδομένα

q: Σειρά Κινούμενου Μέσου Όρου (MA)

P, D, Q: Οι αντίστοιχες εποχικές παράμετροι

s: Η περίοδος εποχικότητας

Εκφράζεται από τον τύπο:

$$Y_t = \beta_0 + \sum_{i=1}^{k} \beta_i X_{i,t} + \omega_t$$

Μεθοδολογία Πρόβλεψης Νέων Χρηστών

Exponential Smoothing (Εκθετική Εξομάλυνση)

- Απλή Εκθετική Εξομάλυνση: $F(t+1) = a \cdot Y(t) + (1-a) \cdot F(t)$
- Διπλή Εκθετική Εξομάλυνση:

$$F(t+1) = L(t) + T(t)$$

$$L(t) = a \cdot Y(t) + (1-a) \cdot \big(L(t-1) + T(t-1)\big), \qquad T(t) = \beta \cdot \big(L(t) - L(t-1)\big) + (1-\beta) \cdot T(t-1)$$

- Τριπλή Εκθετική Εξομάλυνση:

$$F(t+1) = \big(L(t) + T(t)\big) \cdot S(t-m+1)$$

$$L(t) = a \cdot \frac{Y(t)}{S(t-m)} + (1-a) \cdot \big(L(t-1) + T(t-1)\big), \; T(t) = \beta \cdot \big(L(t) - L(t-1)\big) + (1-\beta) \cdot T(t-1)$$

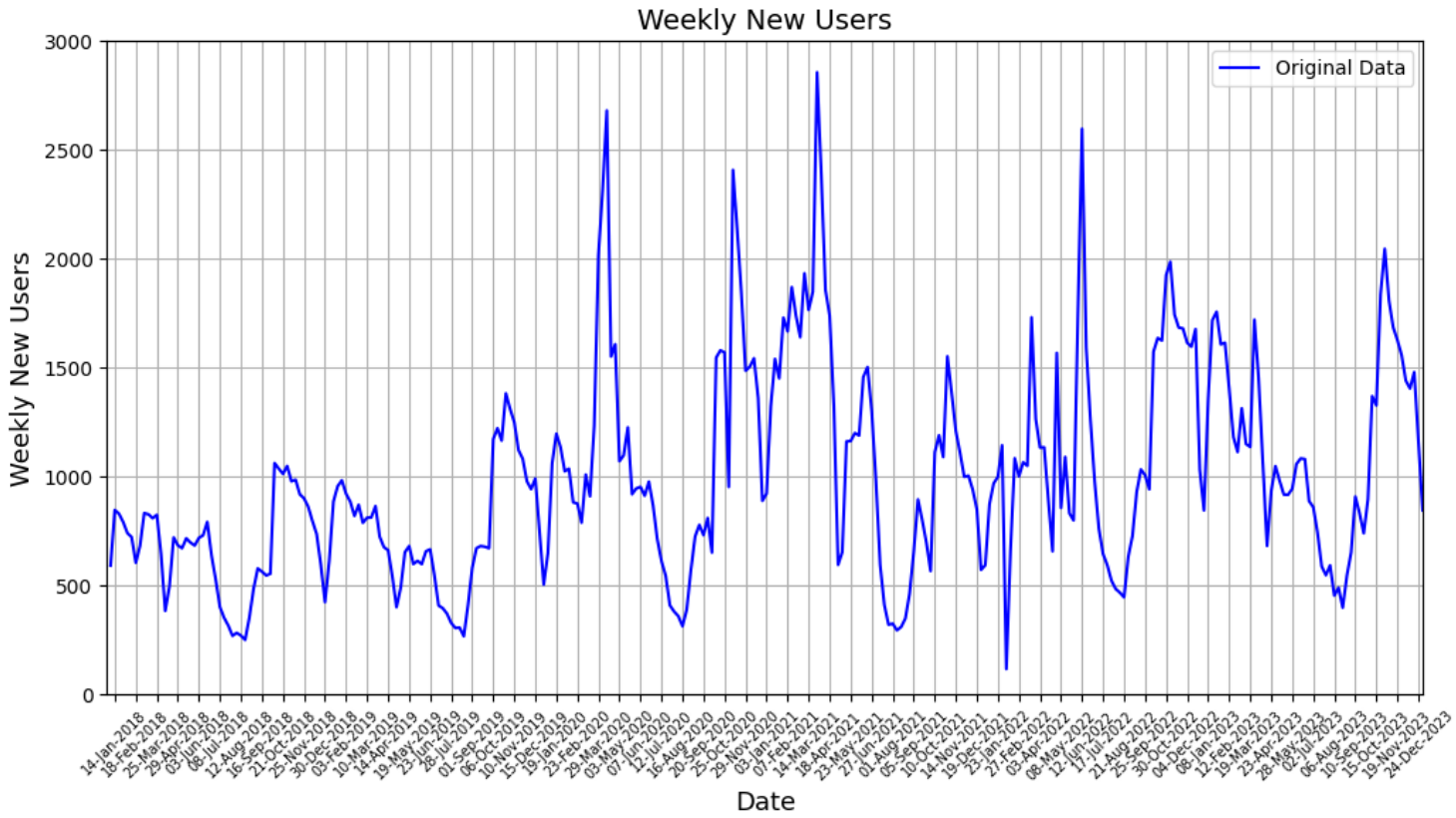
$$S(t) = \gamma \cdot \frac{Y(t)}{L(t)} + (1-\gamma) \cdot S(t-m)$$

Μεθοδολογία Πρόβλεψης Νέων Χρηστών

Light Gradient Boosting Machine (LightGBM)

- Παραλλαγή του αλγορίθμου ενίσχυσης κλίσης
- Σχεδιασμένο για μεγάλες χρονοσειρές και δεδομένα μεγάλης κλίμακας
- Χαρακτηριστικά:
  - Μείωση χρήσης μνήμης
  - Βελτίωση ακρίβειας
  - Πιο γρήγορο από τους παραδοσιακούς αλγόριθμους ενίσχυσης κλίσης

Μεθοδολογία Πρόβλεψης Νέων Χρηστών

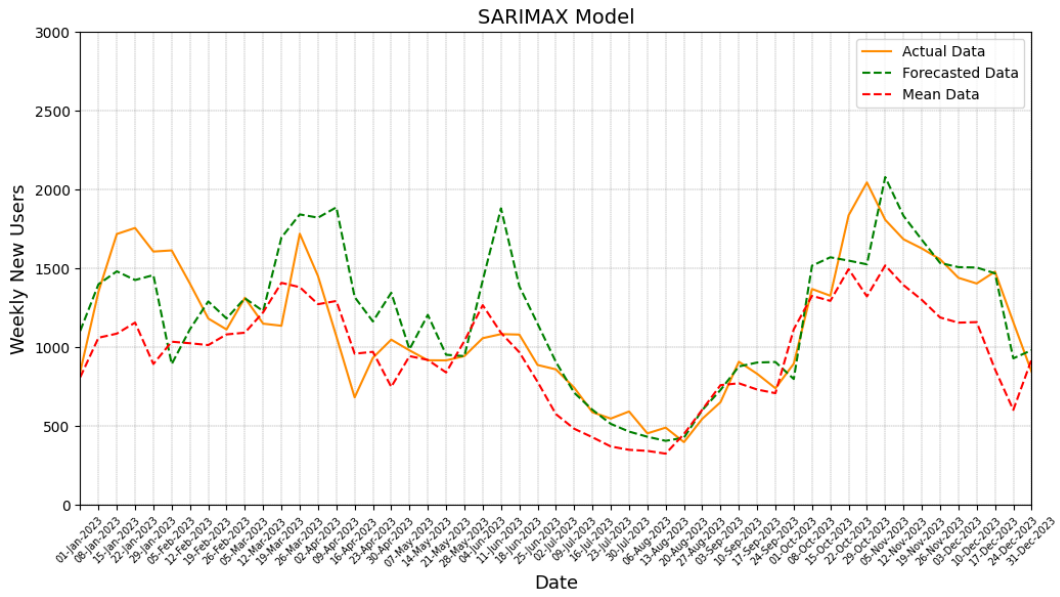


Εβδομαδιαίος Μέσος Όρος Νέων Χρηστών

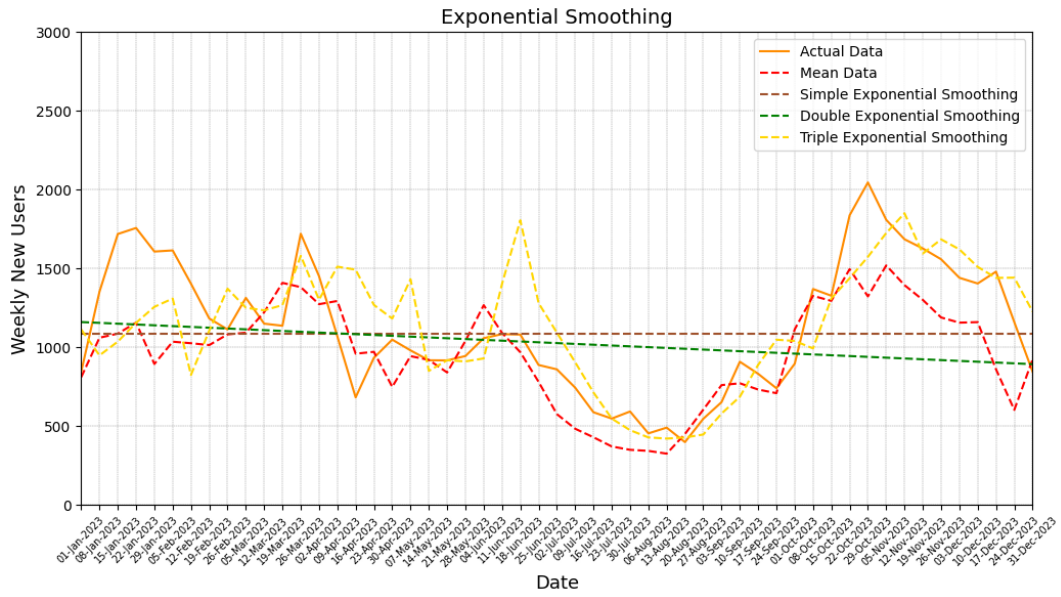
Αποτελέσματα Μοντέλων Πρόβλεψης

ΠΡΟΒΛΕΨΗ ΝΕΩΝ ΧΡΗΣΤΩΝ
ΜΕ ΜΕΘΟΔΟΥΣ ΣΤΑΤΙΣΤΙΚΗΣ

SARIMAX



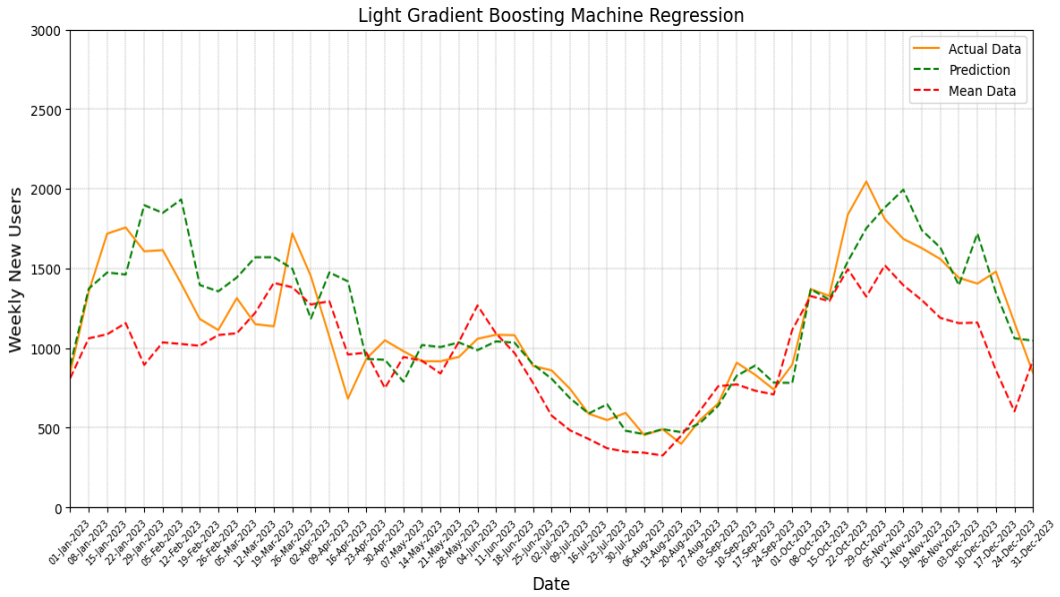
Exponential Smoothing



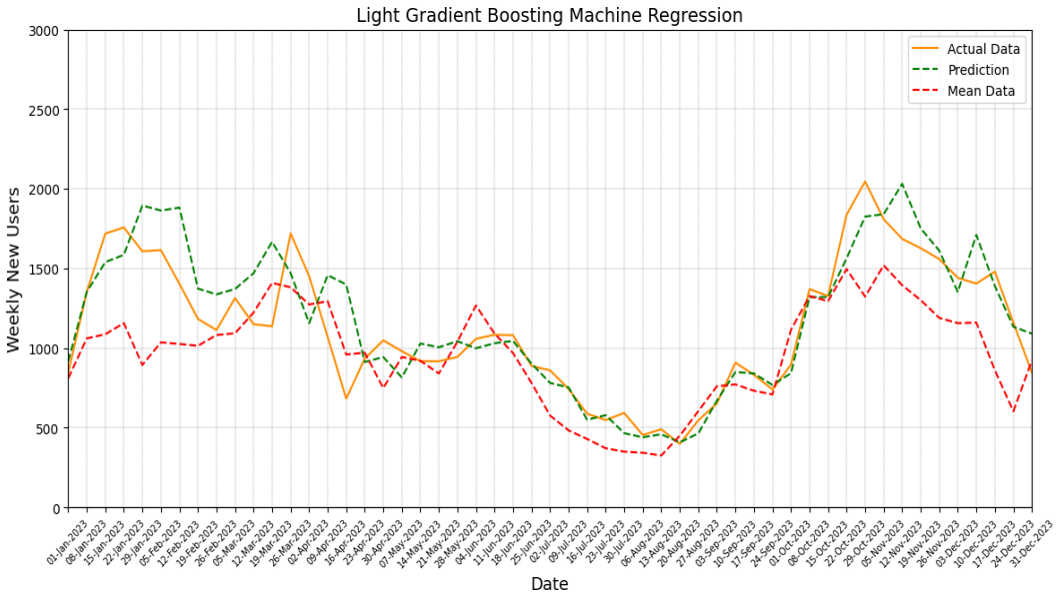
Αποτελέσματα Μοντέλων Πρόβλεψης

ΠΡΟΒΛΕΨΗ ΝΕΩΝ ΧΡΗΣΤΩΝ ΜΕ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

LightGBM



LightGBM και Καθυστερήσεις



Αποτελέσματα Μοντέλων Πρόβλεψης

ΜΟΝΤΕΛΟ	ΜΕΣΟ ΑΠΟΛΥΤΟ ΣΦΑΛΜΑ (MAE)
SARIMAX (1, 0, 1)(2, 1, 0, 52)	198.516
ΑΠΛΗ ΕΚΘΕΤΙΚΗ ΕΞΟΜΑΛΥΝΣΗ	339.122
ΔΙΠΛΗ ΕΚΘΕΤΙΚΗ ΕΞΟΜΑΛΥΝΣΗ	337.186
ΤΡΙΠΛΗ ΕΚΘΕΤΙΚΗ ΕΞΟΜΑΛΥΝΣΗ	231.304
LightGBM	154.677
LightGBM & ΚΑΘΥΣΤΕΡΗΣΕΙΣ	143.555
ΜΕΣΟΣ ΟΡΟΣ	232.035

Αποτελέσματα Μοντέλων Πρόβλεψης

Επιλογή Αποδοτικών Παραμέτρων

Algorithm	**LGBMRegressor**				
Model	**n_estimators**	**learning_rate**	**num_leaves**	**max_depth**	**MAE**
Model A	Default (100)	Default (0.1)	Default (31)	Default (-1)	**154.677**
Model B	200	0.05	50	10	154.738
Model C	300	0.025	60	12	162.244
Model D	350	0.02	62	14	163.698
Model E	370	0.015	64	16	166.550
Train Dataset Attributes	day, month, year_mapping, week_number, rolling_mean_4_weeks_scaled				

Algorithm	**LGBMRegressor**				
Model	**n_estimators**	**learning_rate**	**num_leaves**	**max_depth**	**MAE**
Model A	Default (100)	Default (0.1)	Default (31)	Default (-1)	175.932
Model B	200	0.05	50	10	173.788
Model C	300	0.025	60	12	175.119
Model D	350	0.02	62	14	176.598
Model E	370	0.015	64	16	174.815
Train Dataset Attributes	day, month, year_mapping, rolling_mean_4_weeks_scaled				

Αποτελέσματα Μοντέλων Πρόβλεψης

Επιλογή Αποδοτικών Παραμέτρων

Algorithm	LGBMRegressor				
Model	**n_estimators**	**learning_rate**	**num_leaves**	**max_depth**	**MAE**
Model A	Default (100)	Default (0.1)	Default (31)	Default (-1)	235.310
Model B	200	0.05	50	10	234.675
Model C	300	0.025	60	12	232.961
Model D	350	0.02	62	14	228.749
Model E	370	0.015	64	16	219.759
Train Dataset Attributes	day, month, year_mapping, week_number				

Algorithm	LGBMRegressor				
Model	**n_estimators**	**learning_rate**	**num_leaves**	**max_depth**	**MAE**
Model A	Default (100)	Default (0.1)	Default (31)	Default (-1)	238.186
Model B	200	0.05	50	10	232.779
Model C	300	0.025	60	12	222.897
Model D	350	0.02	62	14	224.124
Model E	370	0.015	64	16	227.515
Train Dataset Attributes	day, month, year_mapping				

Συμπεράσματα

- Από την Ανάλυση Στοιχείων Χρήσης
  - Αύξηση της προσέλευσης και αφοσίωσης των ενεργών χρηστών πριν την έναρξη της εξεταστικής και στις αρχές των ακαδημαϊκών εξαμήνων / Μείωση στις περιόδους διακοπών.
  - Ανάγκη μετάφρασης τεκμηρίων σε περισσότερες γλώσσες και προώθηση περιεχομένου μέσω κοινωνικών δικτύων.
- Από τις Μεθόδους Πρόβλεψης Νέων Χρηστών
  - Η μέθοδος LightGBM αποδείχθηκε η πιο αποδοτική και η προσθήκη καθυστερήσεων βελτίωσε τις προβλέψεις.
  - Η μέθοδος SARIMAX ακολουθεί σε αποδοτικότητα.
  - Οι μέθοδοι Exponential Smoothing αποδείχθηκαν οι λιγότερο αποδοτικές.

Μελλοντικές Επεκτάσεις

- Συλλογή περισσότερων δεδομένων από μεγάλα χρονικά διαστήματα για την καλύτερη κατανόηση των τάσεων και μοτίβων συμπεριφοράς των χρηστών.
- Συλλογή πρόσφατων δημογραφικών δεδομένων, όπως ηλικία, φύλο και επίπεδο σπουδών για την προσαρμογή των υπηρεσιών στις ανάγκες των χρηστών.
- Συνδυασμός δημογραφικών δεδομένων και στοιχείων χρήσης για την ανάλυση της εξατομικευμένης εμπειρίας χρήστη.
- Επέκταση των μοντέλων πρόβλεψης νέων χρηστών για τις θεματικές ενότητες του αποθετηρίου.

Ευχαριστώ πολύ για τον χρόνο σας!