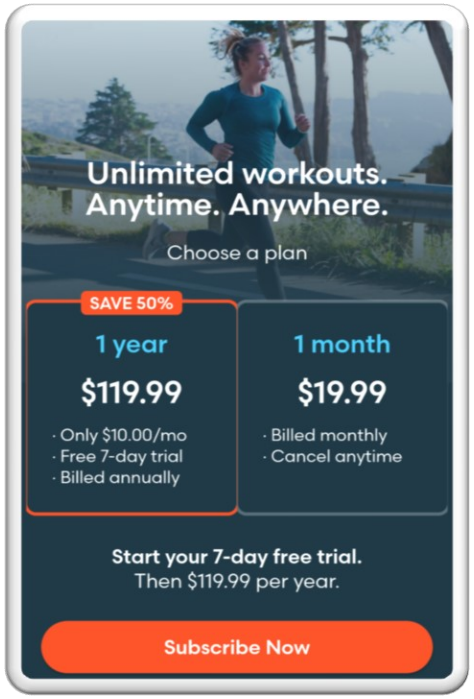
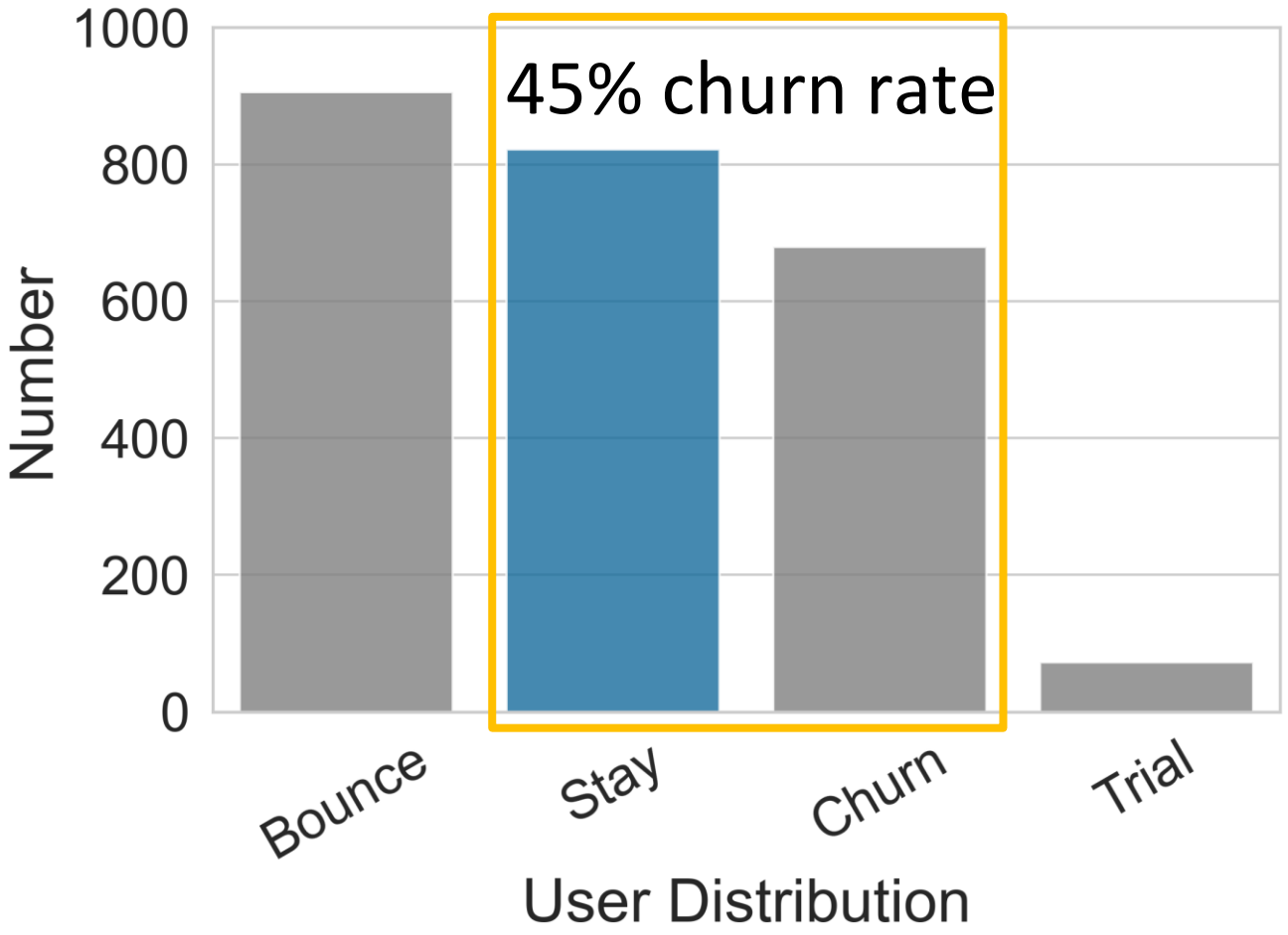
Beyond Words

predict user decision with text (meta)data

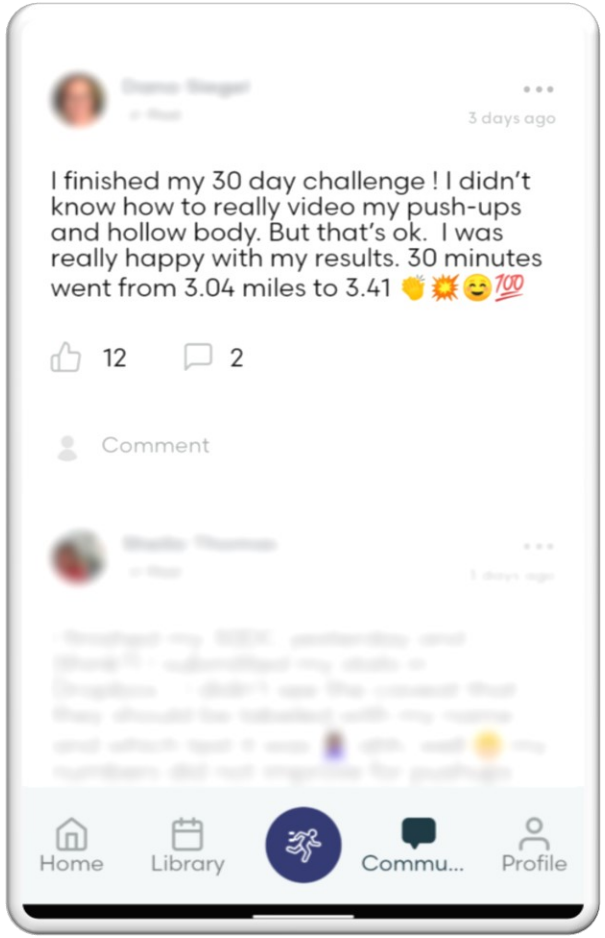
Eric Zhang

Churn down ↓ Revenue up ↑



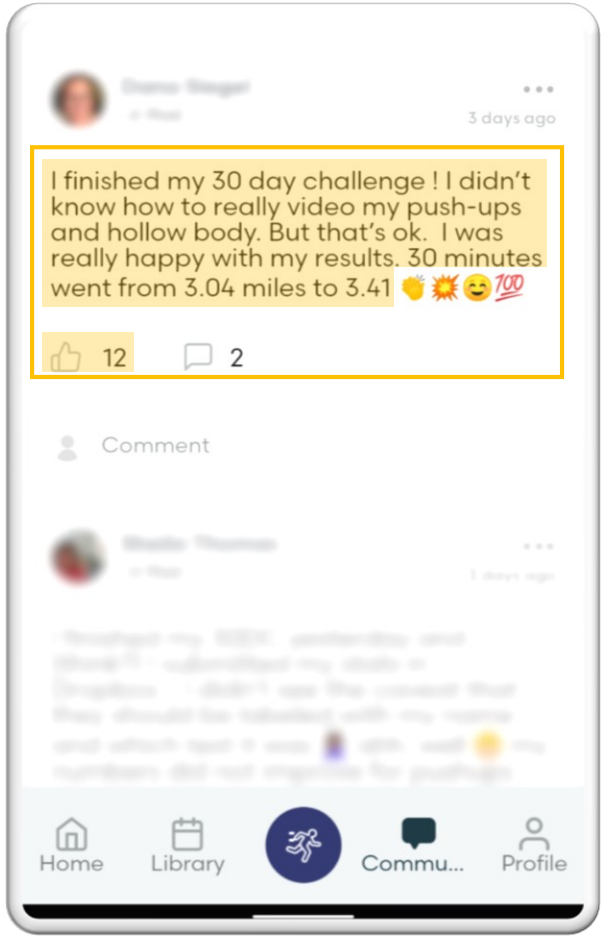


Predict user churn with text data



User text data: meta

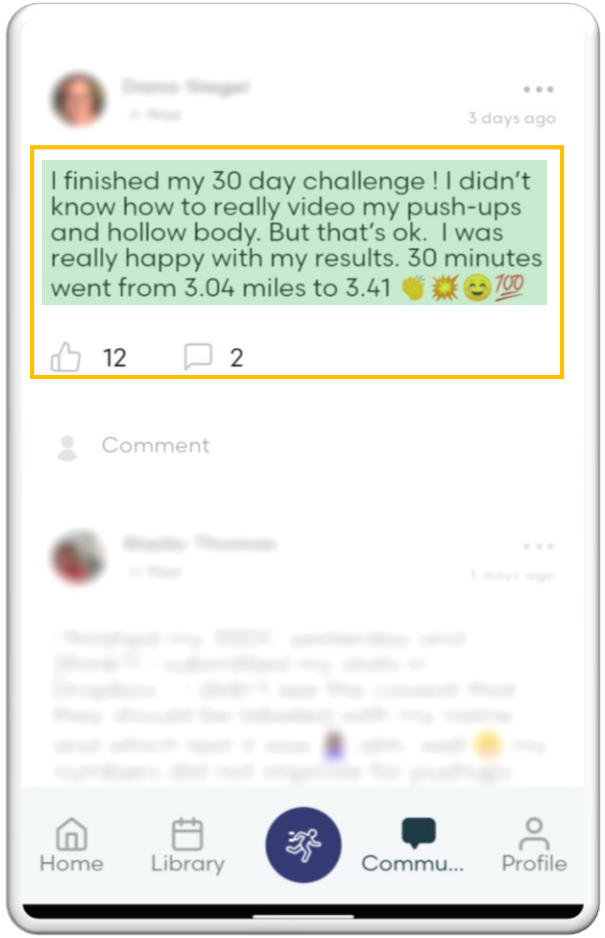
- Text
  - Characters, words
  - Likes



Total characters: 150
Total words: 36
Total likes: 12

User text data: sentiment

- Text
  - Characters, words
  - Likes
- Sentiment
  - Text and Emoji
  - Tone
  - Content



Total characters: 150
Total words: 36
Total likes: 12

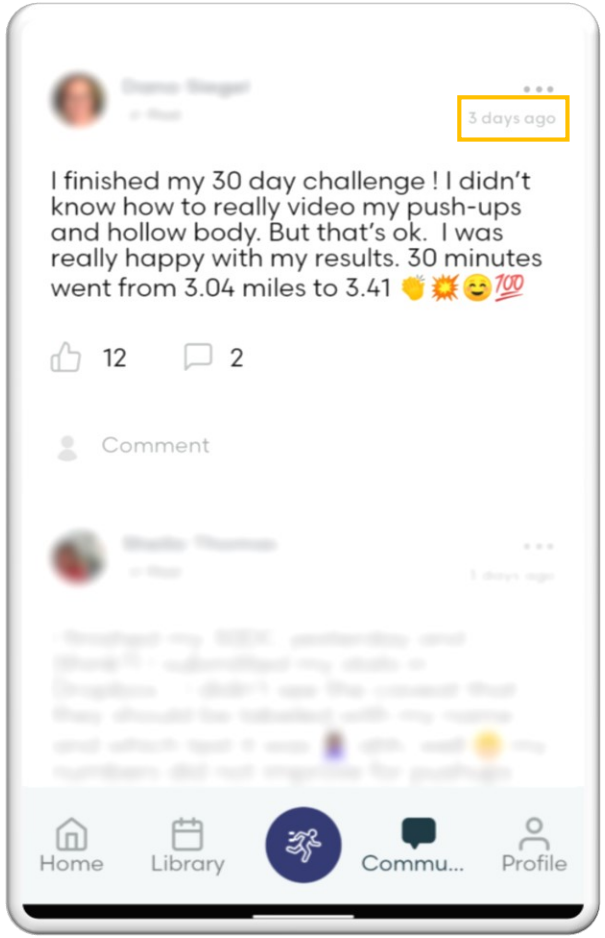
Emojis:

👏 >>> Clapping Hands
💥 >>> Collision
😊 >>> Smiley Face
💯 >>> Hundred Points

Sentiment:
Tone: Positive (1)
Content: Rich (1)

User text data: timestamps

- Text
  - Characters, words
  - Likes
- Sentiment
  - Text and Emoji
  - Tone
  - Content
- Frequency
  - Timestamp
  - Day, week, month



Total characters: 150
Total words: 36
Total likes: 12

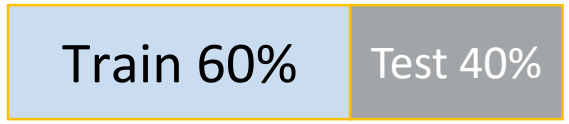
Emojis:
👏 >>> Clapping Hands
💥 >>> Collision
😊 >>> Smiley Face
💯 >>> Hundred Points

Sentiment:
Tone: Positive (1)
Content: Rich (1)

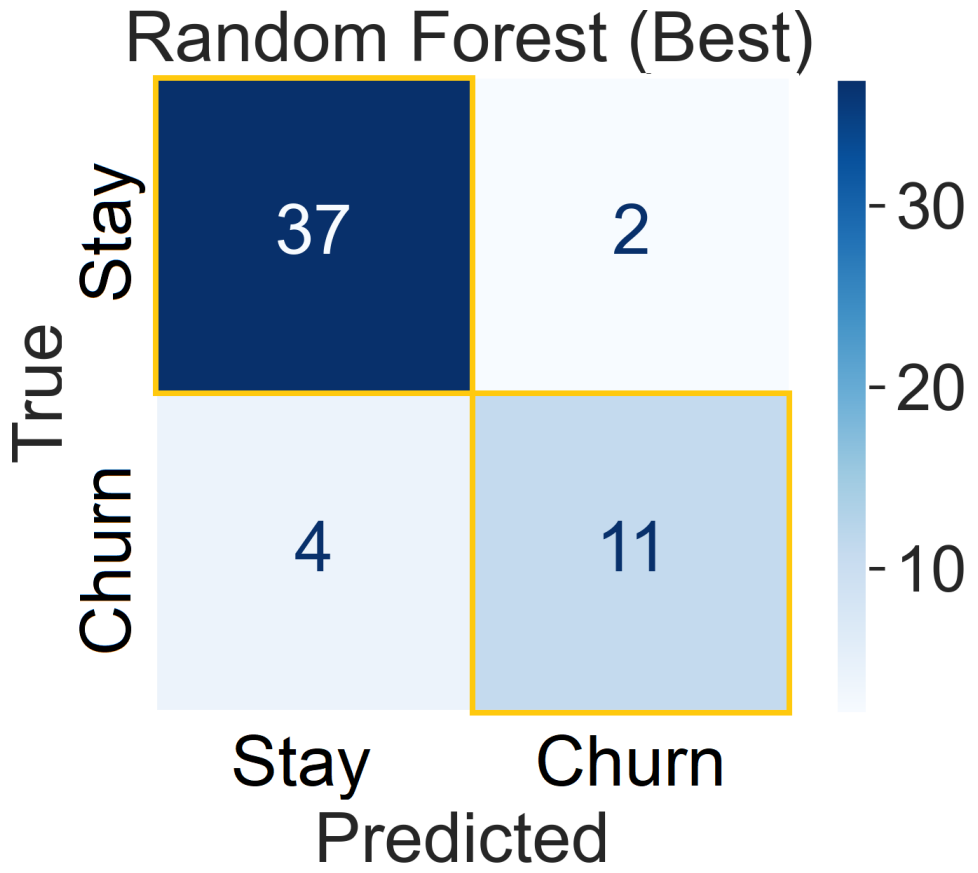
Predict user churn with 60 features

159 user data	60 features	Prediction
• 1393 texts	• Text, character, etc. • Sentiment • Frequency 	• Churn (True/False) • Classification

89% accuracy on churn detection

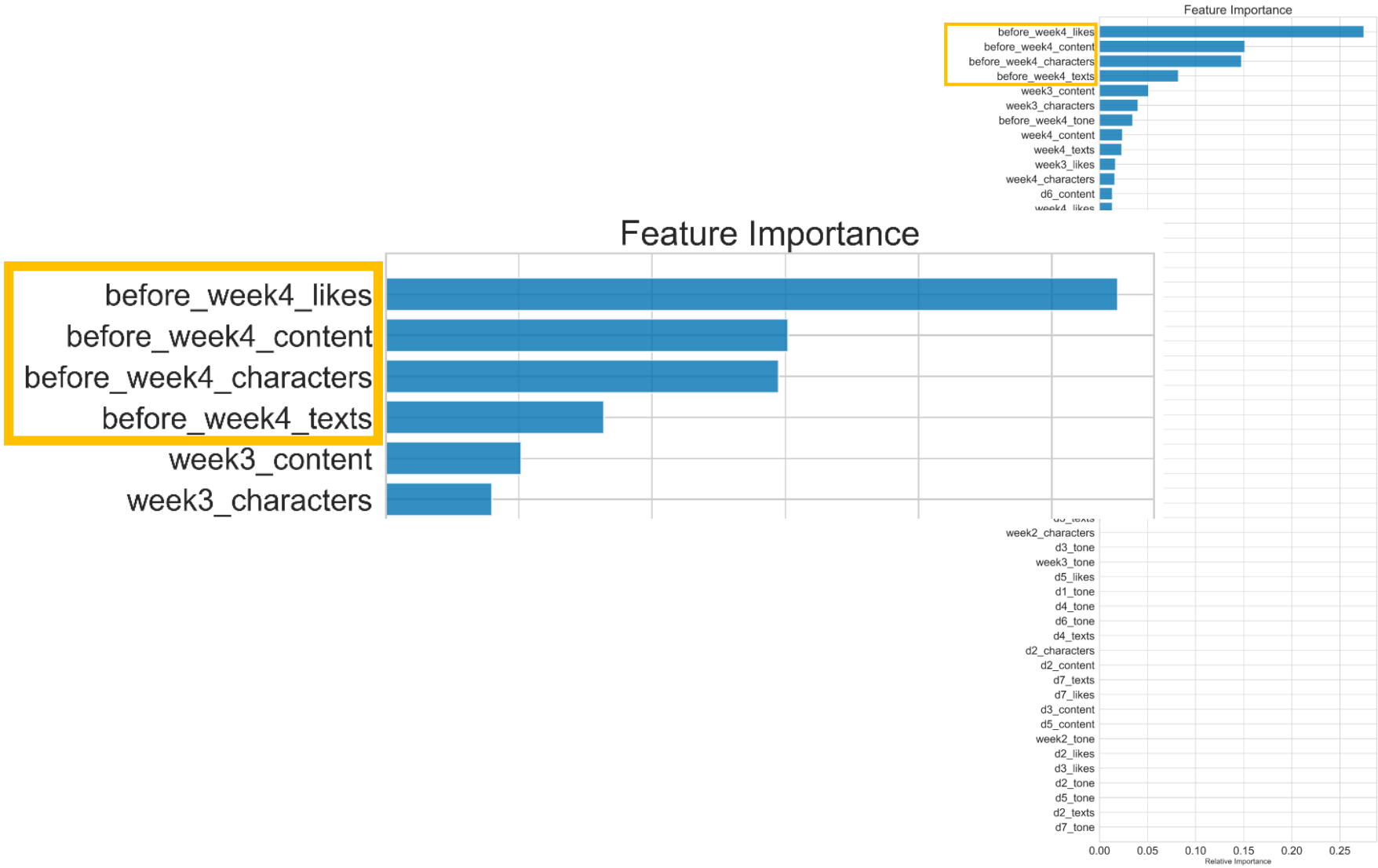


Stratified, 5 folds CV

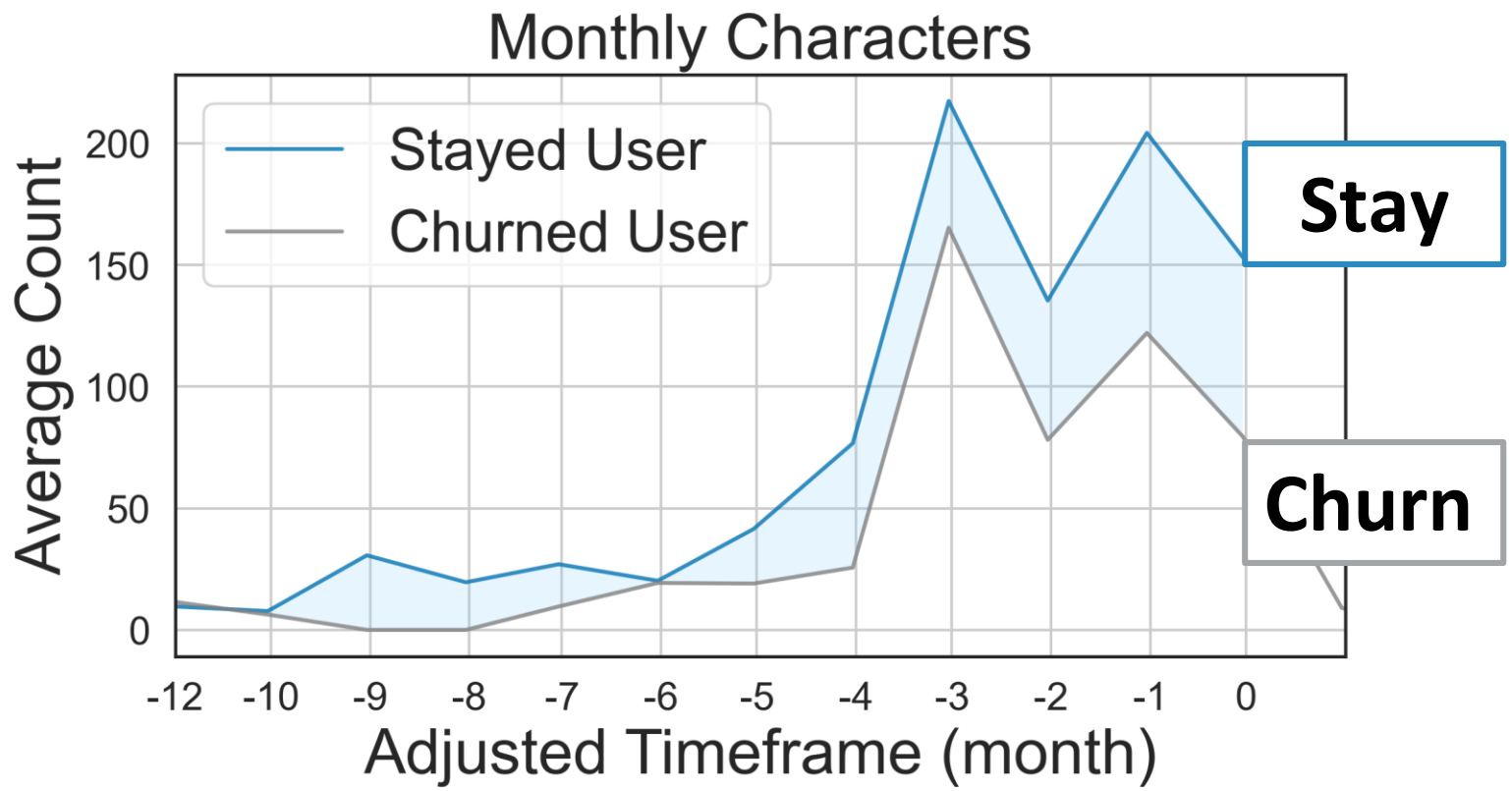


Accuracy:	88.89%
Precision:	84.62%
Recall:	73.33%

Data over one month ago can tell user churn



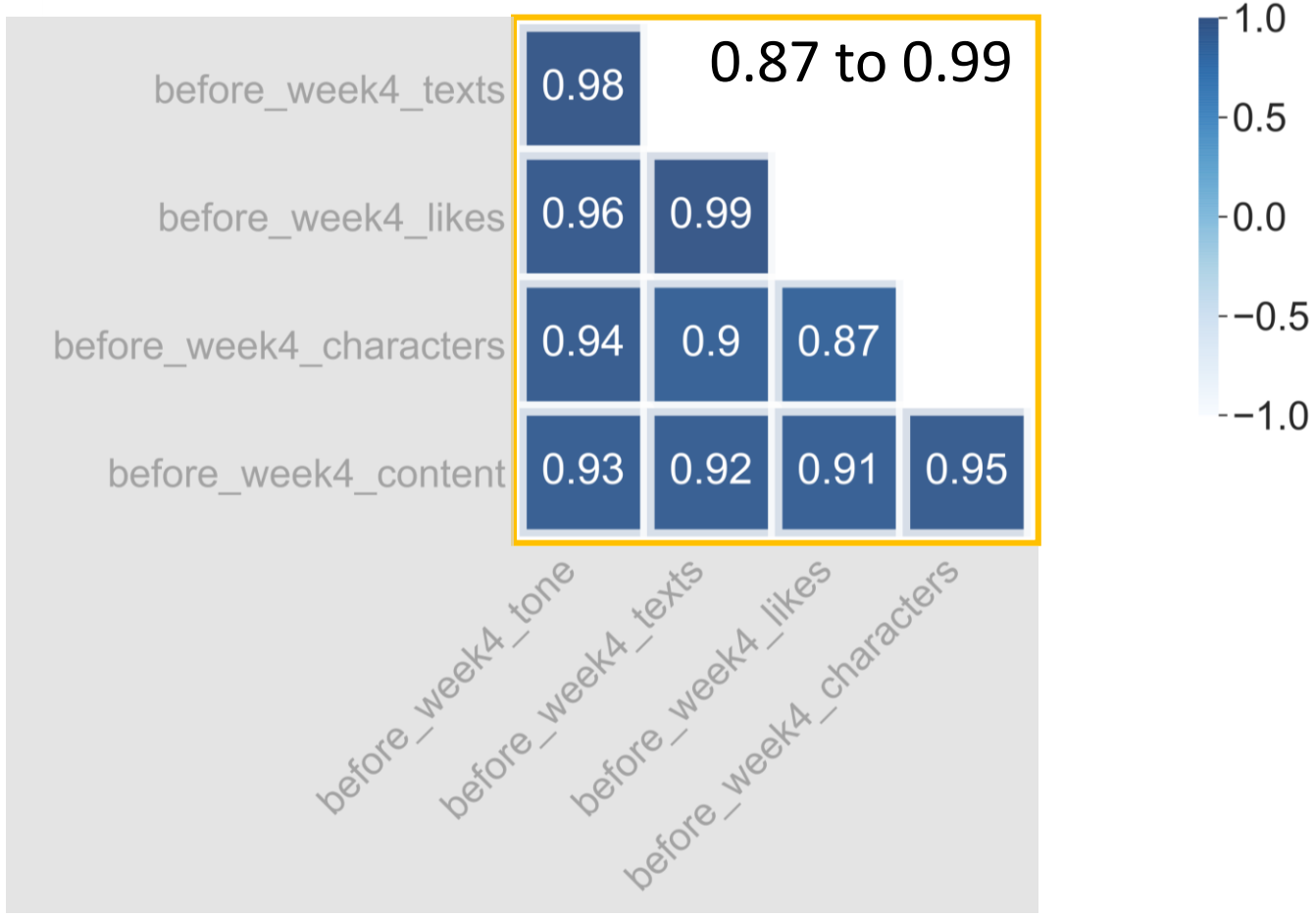
Churned users were typing less



Text meta-data can predict user churn

Strong correlation

- text meta
- sentiment



Text meta-data can predict user churn

Strong correlation
- text meta
- sentiment

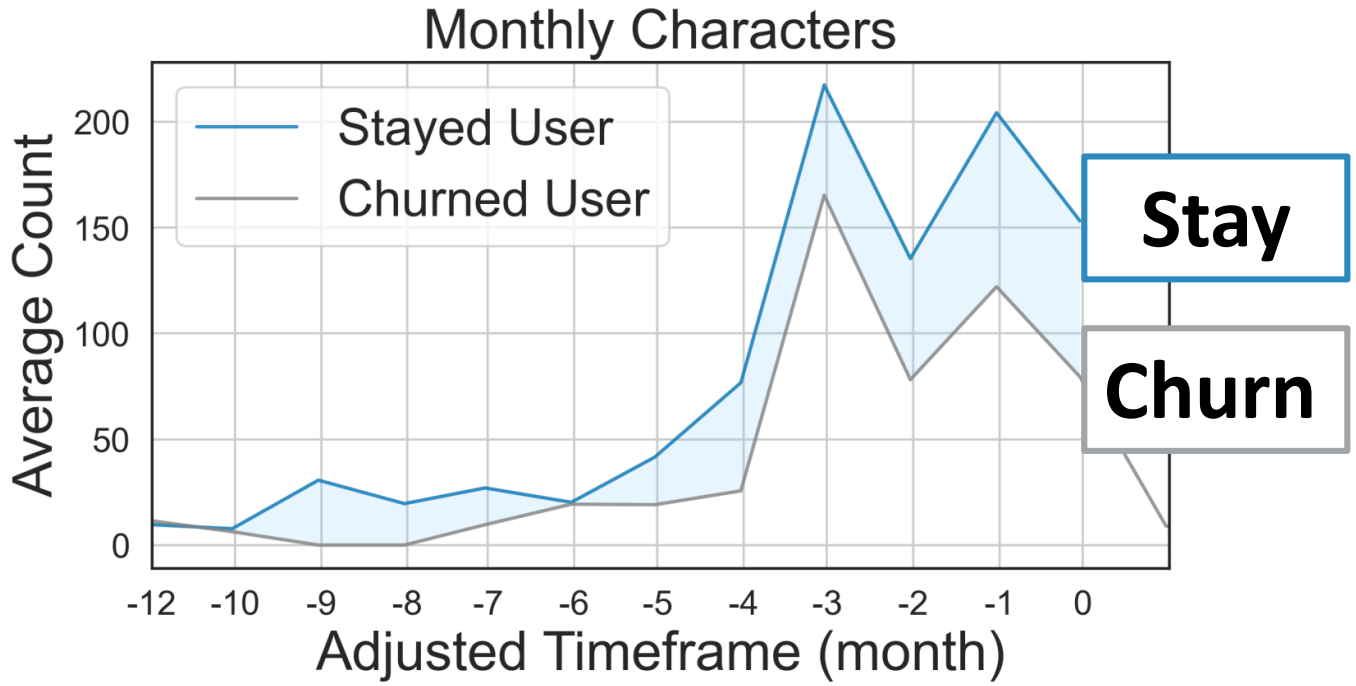
Text meta features
- Good enough
- Easy to scale up

Text meta-data can predict user churn

Strong correlation
- text meta
- sentiment

Text meta features
- Good enough
- Easy to scale up

User in-app communication
- Strong indicator of user churn

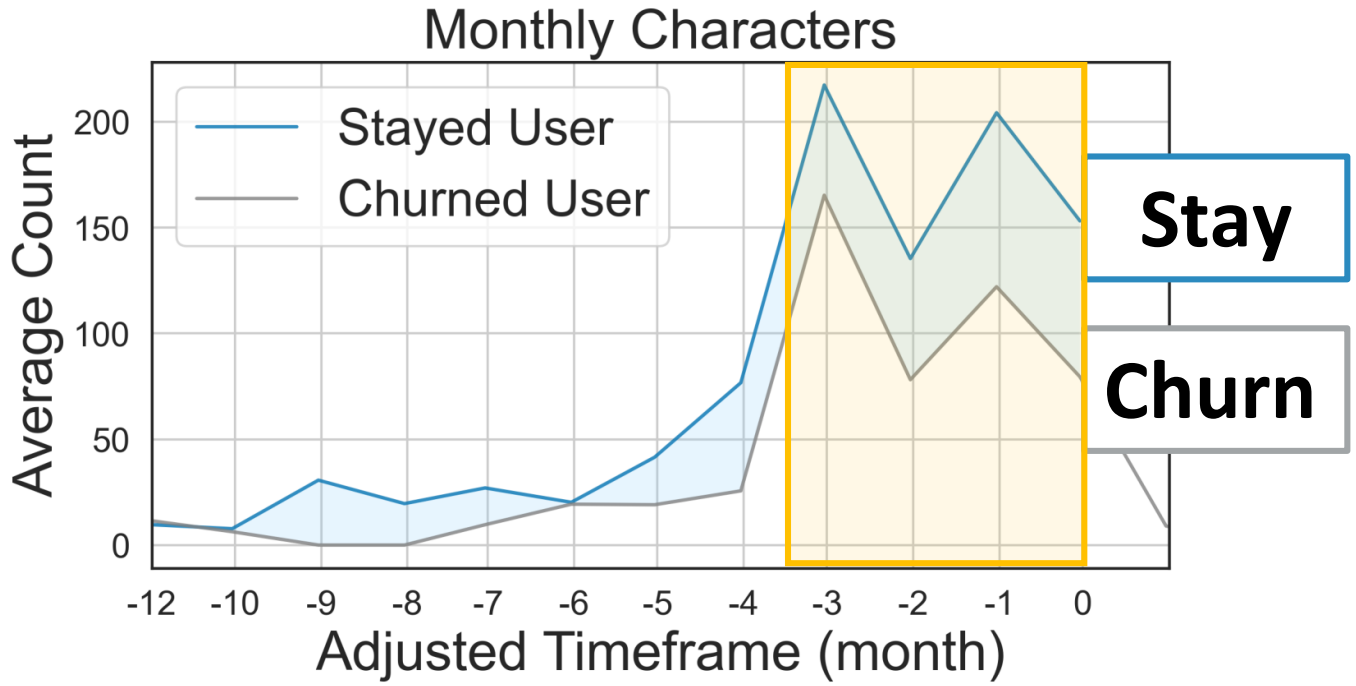


Text meta-data can predict user churn

Strong correlation
- text meta
- sentiment

Text meta features
- Good enough
- Easy to scale up

User in-app communication
- Strong indicator of user churn
- Customer life time 3 to 4 months

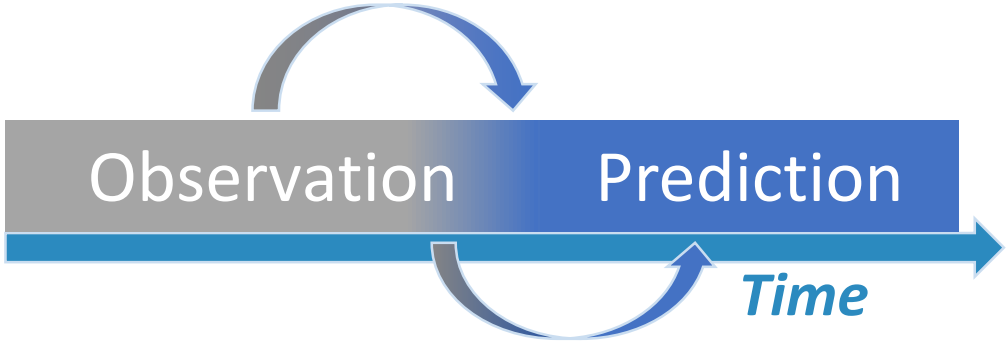


Next

With more data

Real-time prediction

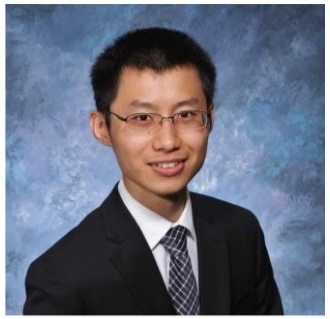
- Observation and prediction windows
- Feedback loop to update ML model



Topic analysis

- LDA / TF-IDF
- User segmentation

Zelong (Eric) Zhang



- PhD in Computational Chemistry
- Award-winning film (US DOE), photography
- User Experience and Decision-Making

Stony Brook University

LSU
LOUISIANA STATE UNIVERSITY

In-app user record

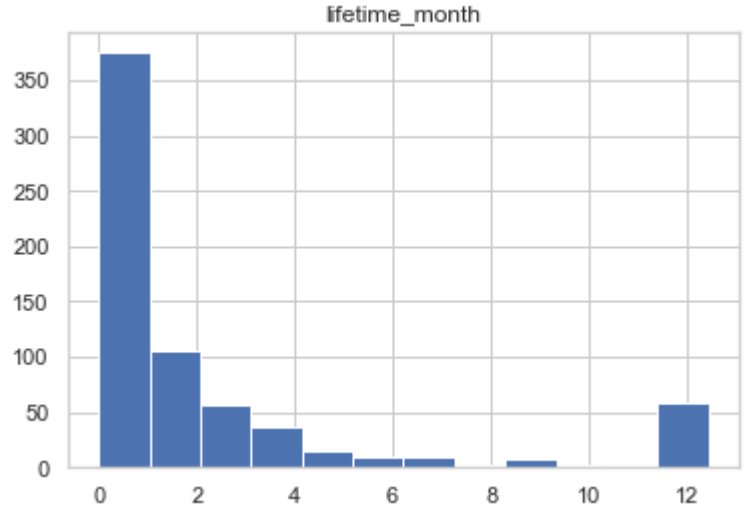
- Registered User: 56603
- User with text data: 773
- User with both registration and text data: 623

- User with subscription history: 2476
- User with both subscription history and text data: 159

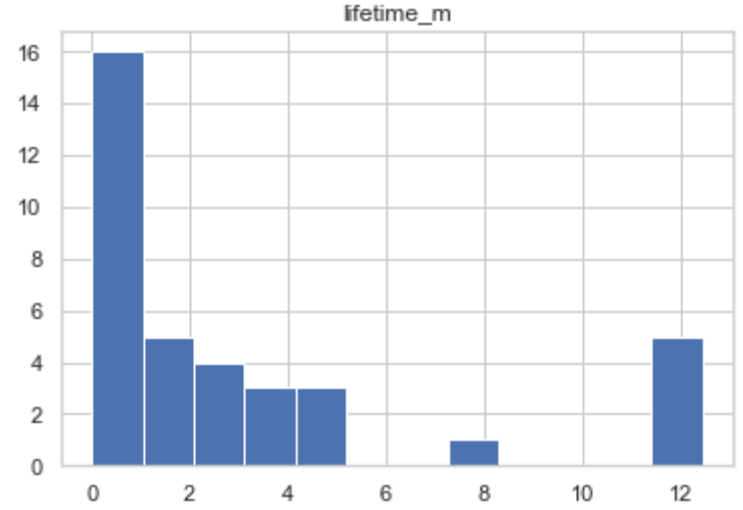
- User with survey record: 490
- User with both subscription and survey: 124
- User with subscription, survey, and text: 34

Churn User Lifetime Stats

User Subscription	Month
Max	12.4
Min	0
Mean	2.8
Median	1.0
Unbiased variance	10.9
Standard deviation	3.3



User in Text Data	Month
Max	12.4
Min	0
Mean	3.4
Median	2.0
Unbiased variance	15.2
Standard deviation	3.9



Model selections

Sentiment analysis

- Natural Language Processing (NLP)
- Pre-trained BERT
- 1393 texts
- Hand labelling 60% texts

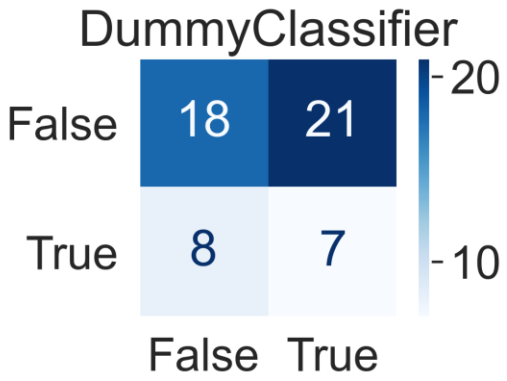
Sentiment	Accuracy
Tone	0.851
Content	0.776

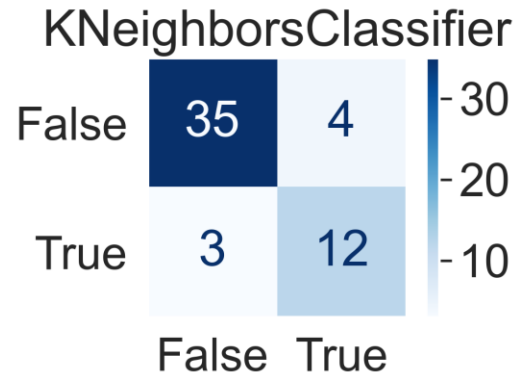
Classification

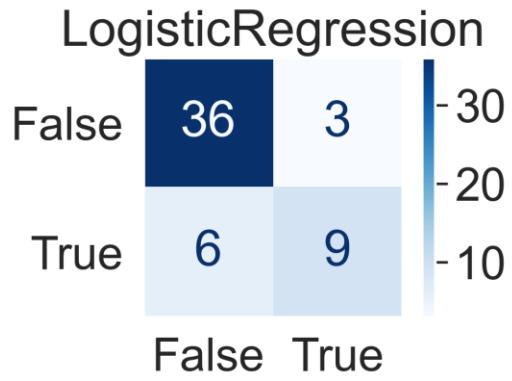
- Logistic Regression, Random Forest, XGBoost, etc.
- Stacking

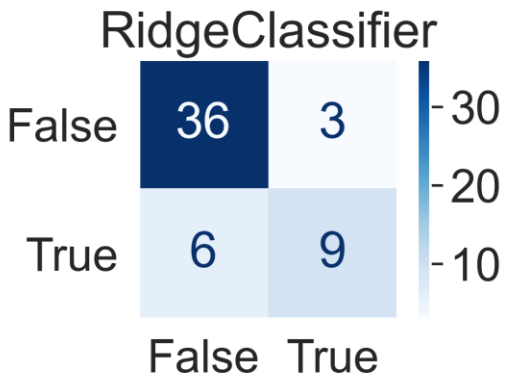
scikit learn dmlc XGBoost

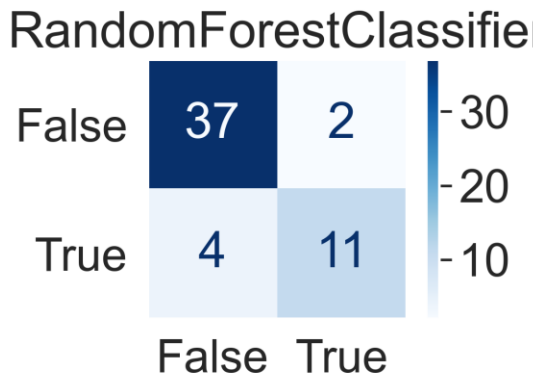
Classification model benchmark

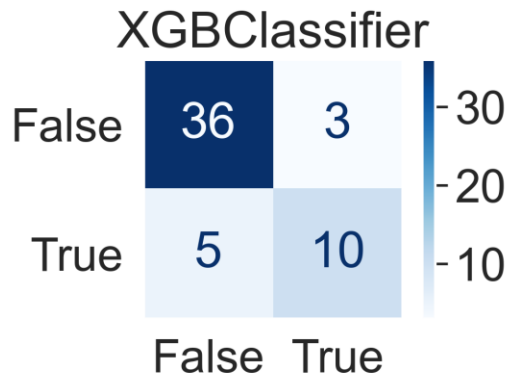


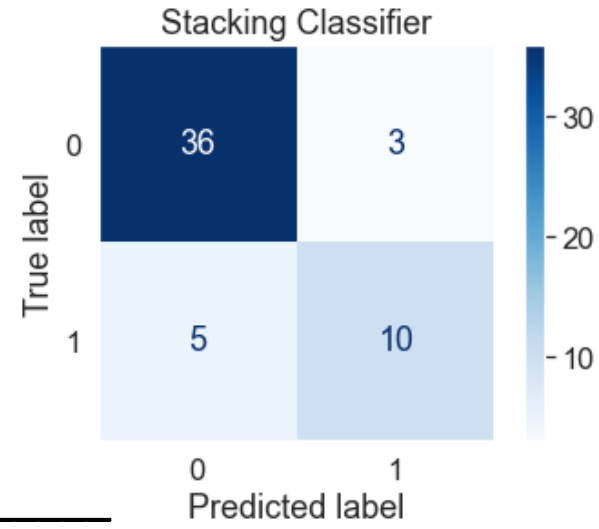










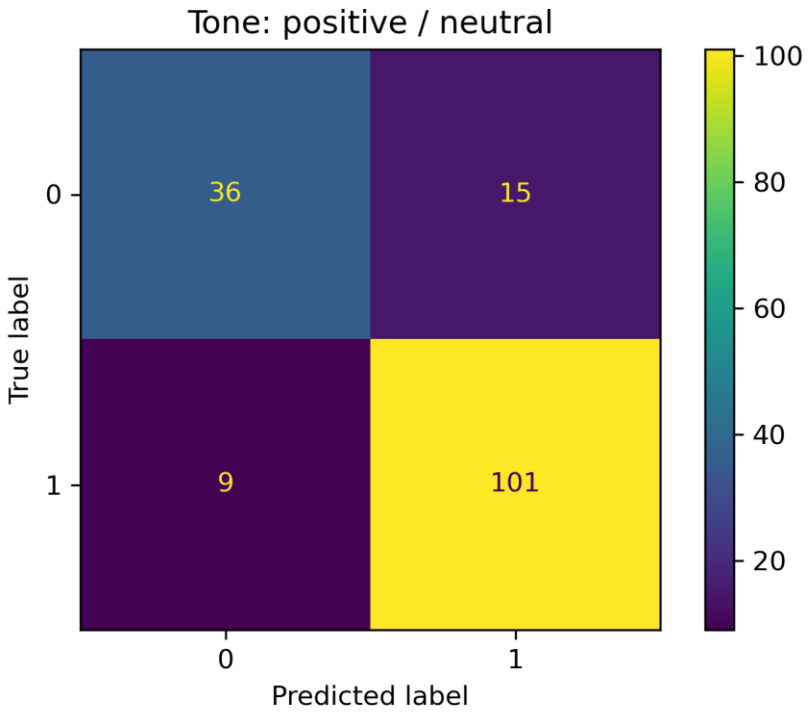


Combined by Logistic Regression

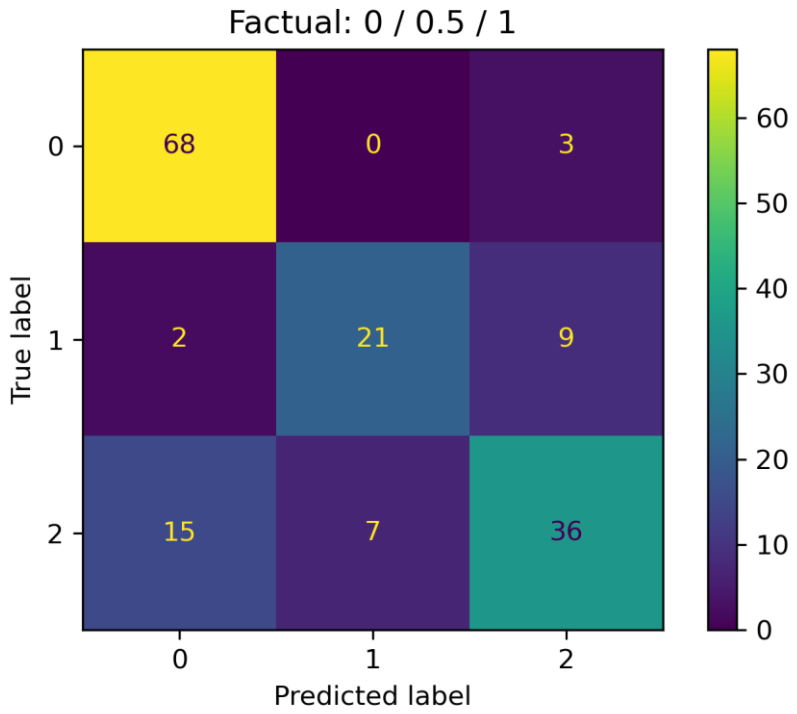
Random Forest	Test	CV	Std. dev.
Accuracy	0.89	0.924	0.042
Precision	0.846	0.951	0.040
Recall	0.733	0.917	0.074
F1	0.786	0.923	0.044

CV: Stratified KFold, 5 splits

NLP BERT validation metrics



Tone:
accuracy 0.850932



Content:
accuracy 0.776398

Sanity check of sentiment analysis

Predicted by the BERT model I trained!

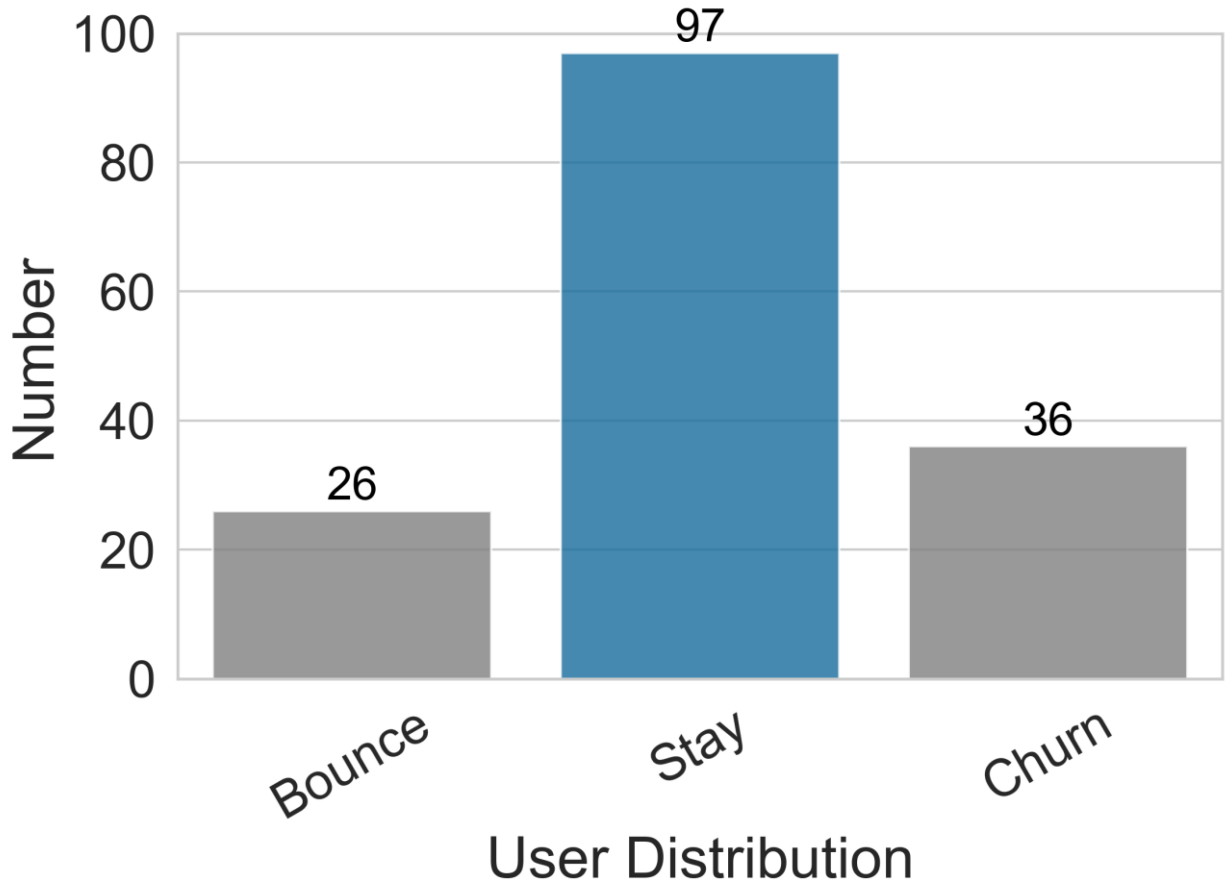
	text	Tone	Content	VADER_Score	OTS_BERT_label	OTS_BERT_score
1387	Way to go, Michael!	positive	0.0	0.0000	positive	0.9998
1388	I like to torture myself!!!! 🤪🤪🤪	positive	0.0	-0.5526	negative	0.9992
1389	yes!!!	positive	0.0	0.5538	positive	0.9997
1390	Oh dear that is swollen. Is ice helping?	neutral	0.5	0.5859	negative	0.9983
1391	Night run	neutral	0.5	0.0000	positive	0.5209
1392	Dabbling in swimming and biking. When my fitne...	positive	0.5	0.3382	negative	0.9583

VADER: a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.
OTS BERT: Off-the-shelf version of “distilbert-base-uncased-finetuned-sst-2-english"

User data overview

159 user data

- 1393 texts

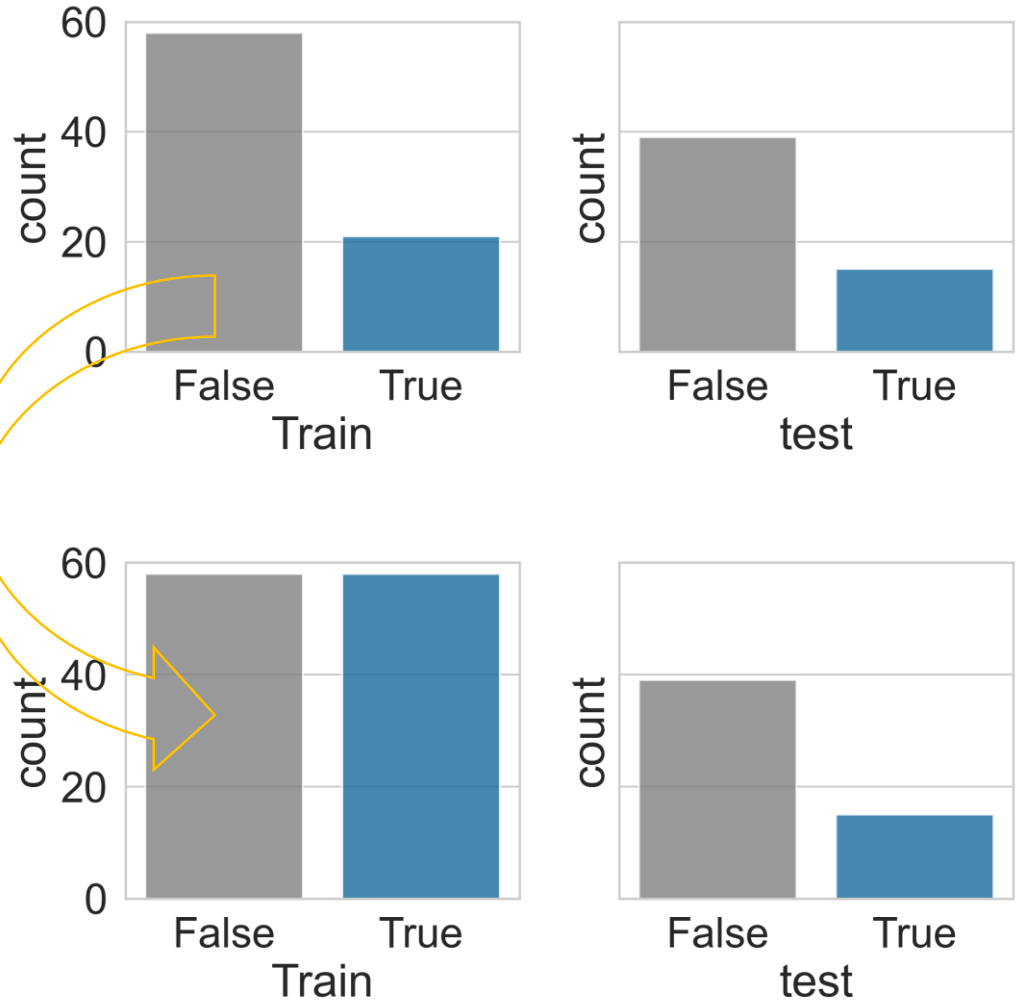


Work flow: data in & data out

159 user data	60 features	Prediction	Insights
• 1393 texts	• Text • Sentiment • Frequency	• Churn (Y / N) • Classification	• Meta-data • Engagement

Train and test datasets

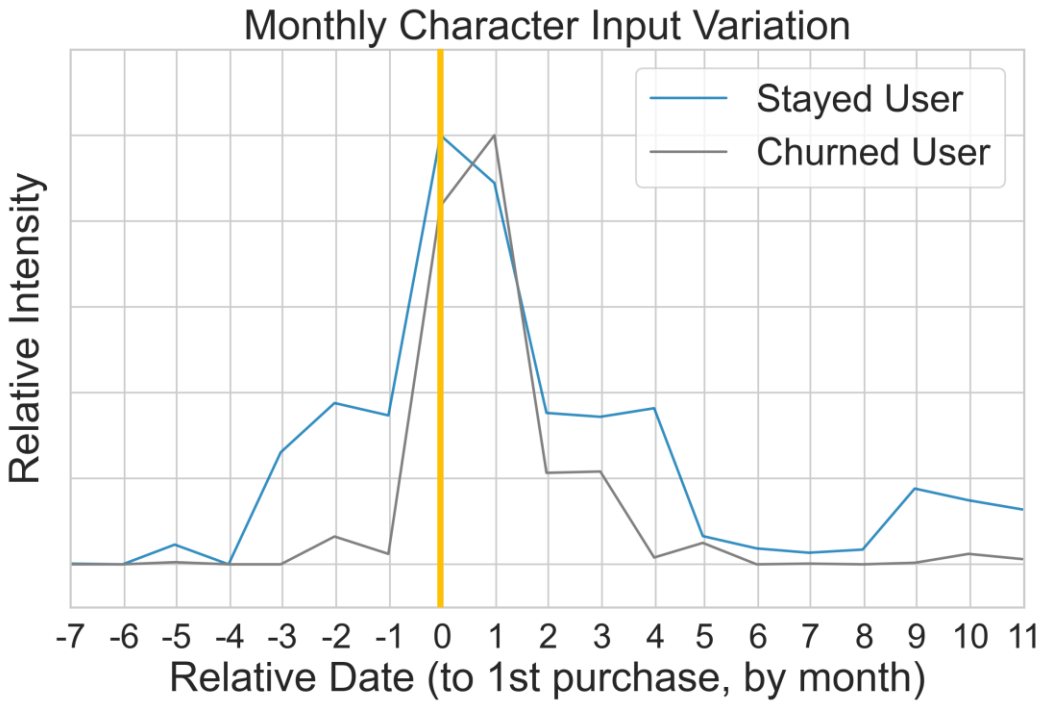
Adaptive Synthetic (ADASYN)

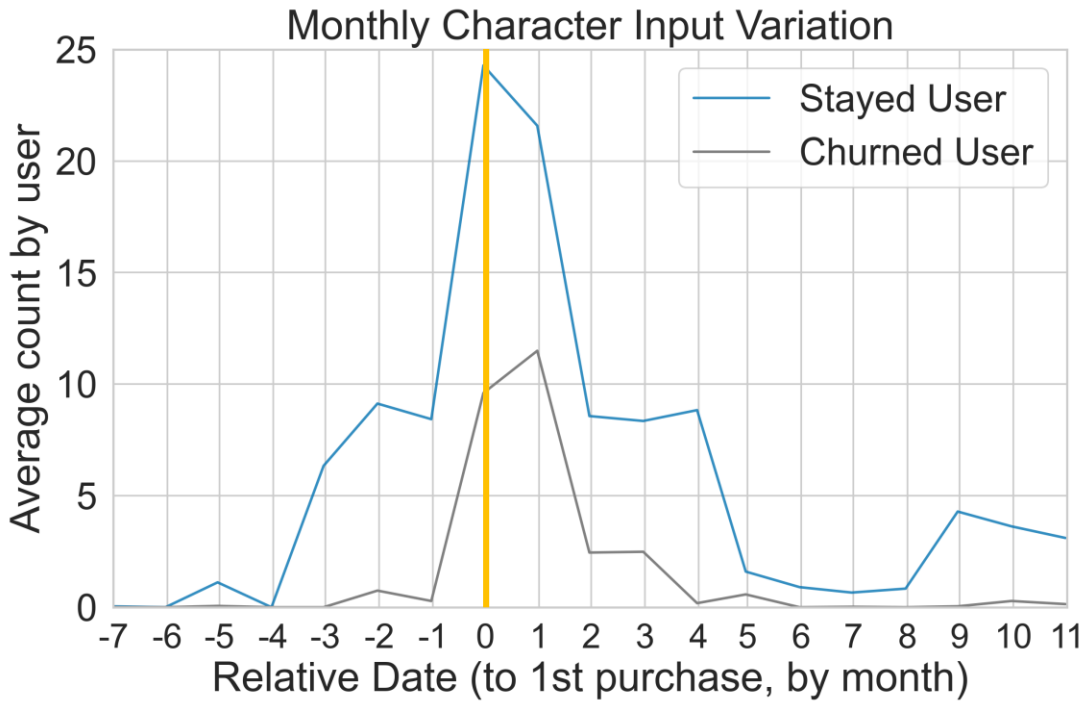


ADASYN is marginally better than SMOTE for this dataset (CV has slightly better recall).

"ADASYN focuses on generating samples next to the original samples which are wrongly classified using a k-Nearest Neighbors classifier while the basic implementation of SMOTE will not make any distinction between easy and hard samples to be classified using the nearest neighbors rule."

Monthly characters (relative to first purchase)

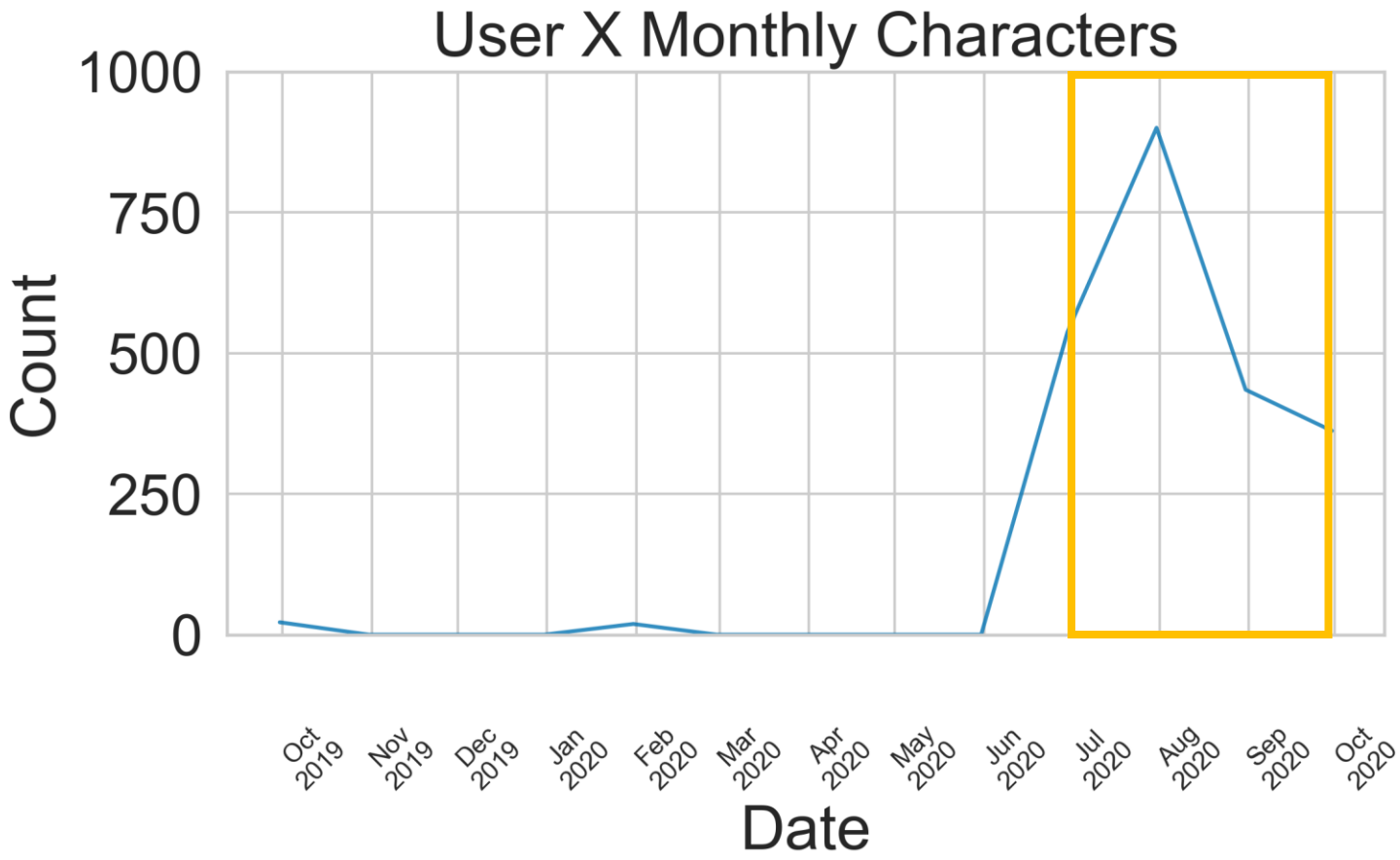




Event 0: first subscription order

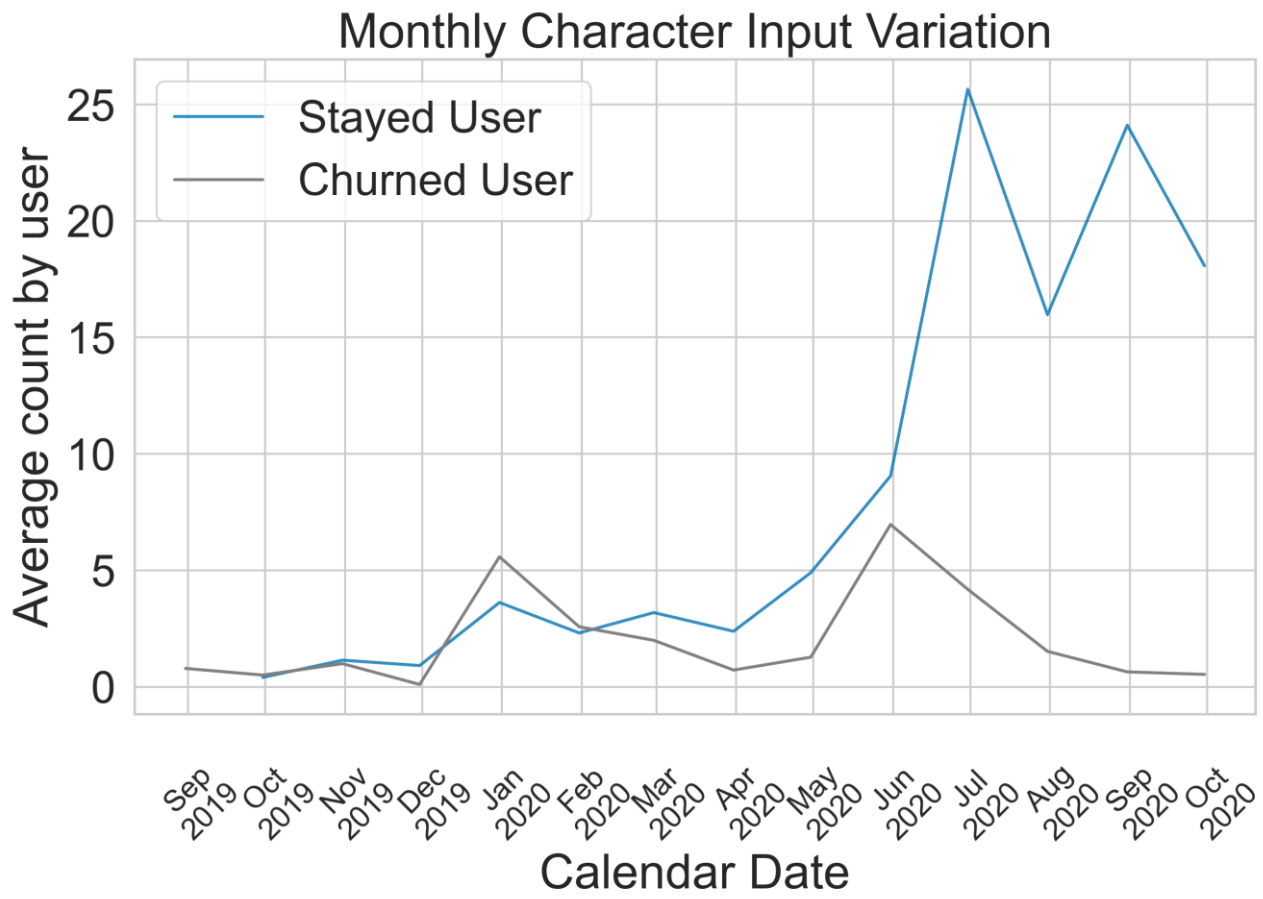
First purchase - peak

Single user monthly stats (total: 77 texts)

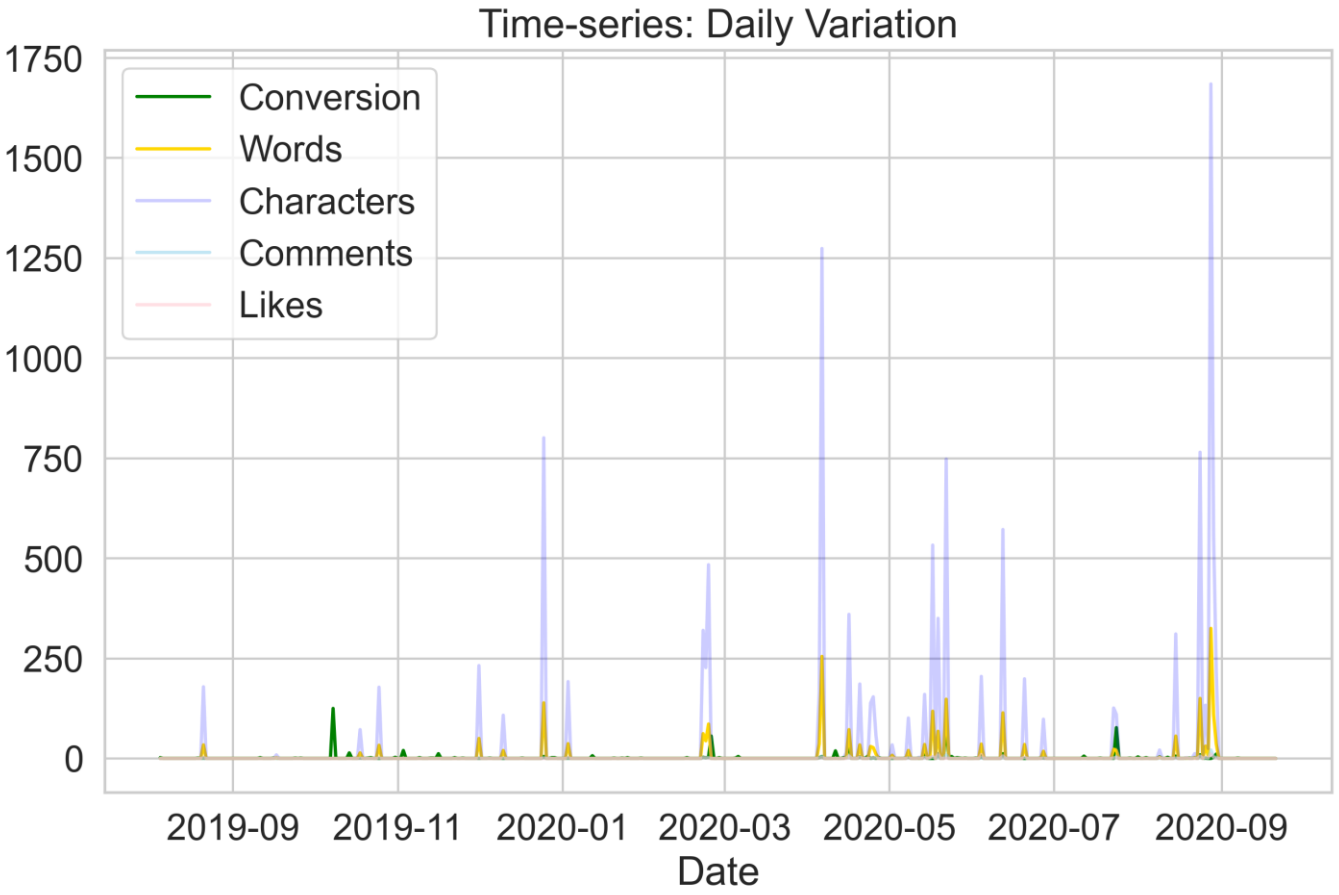


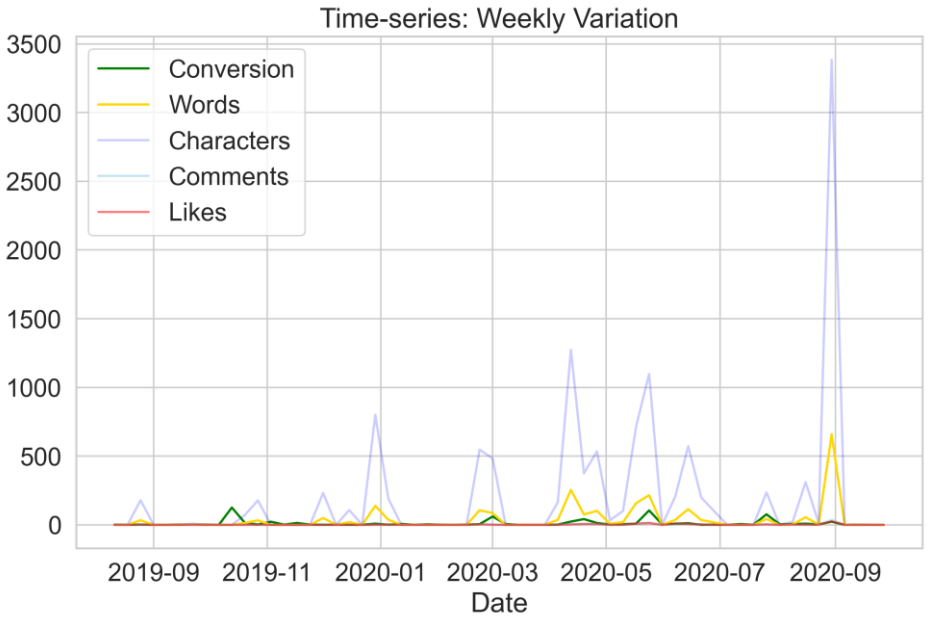
Ref: 37e5fde1-9f8f-4024-90b2-4530318ff905

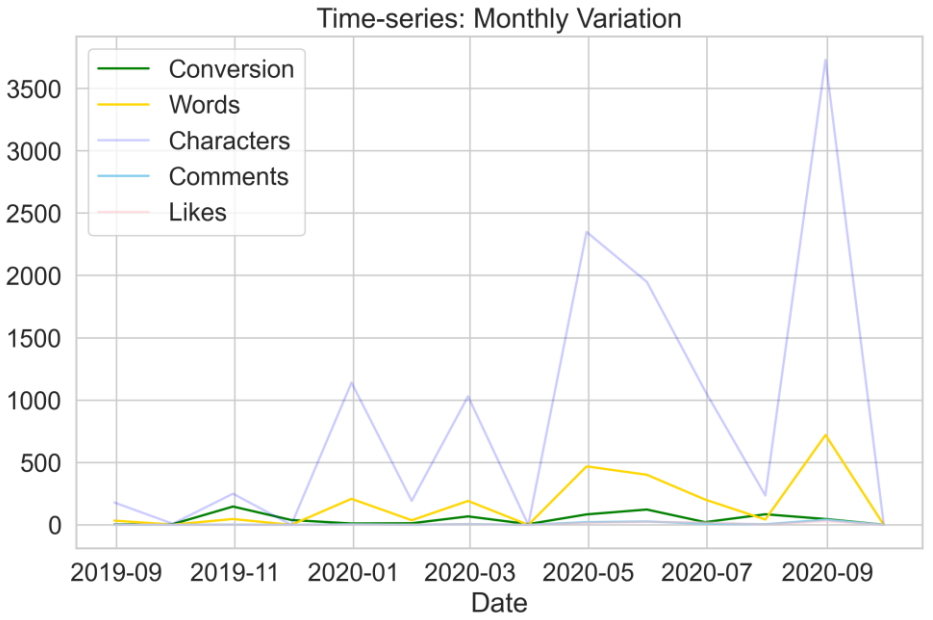
Monthly characters (real time)



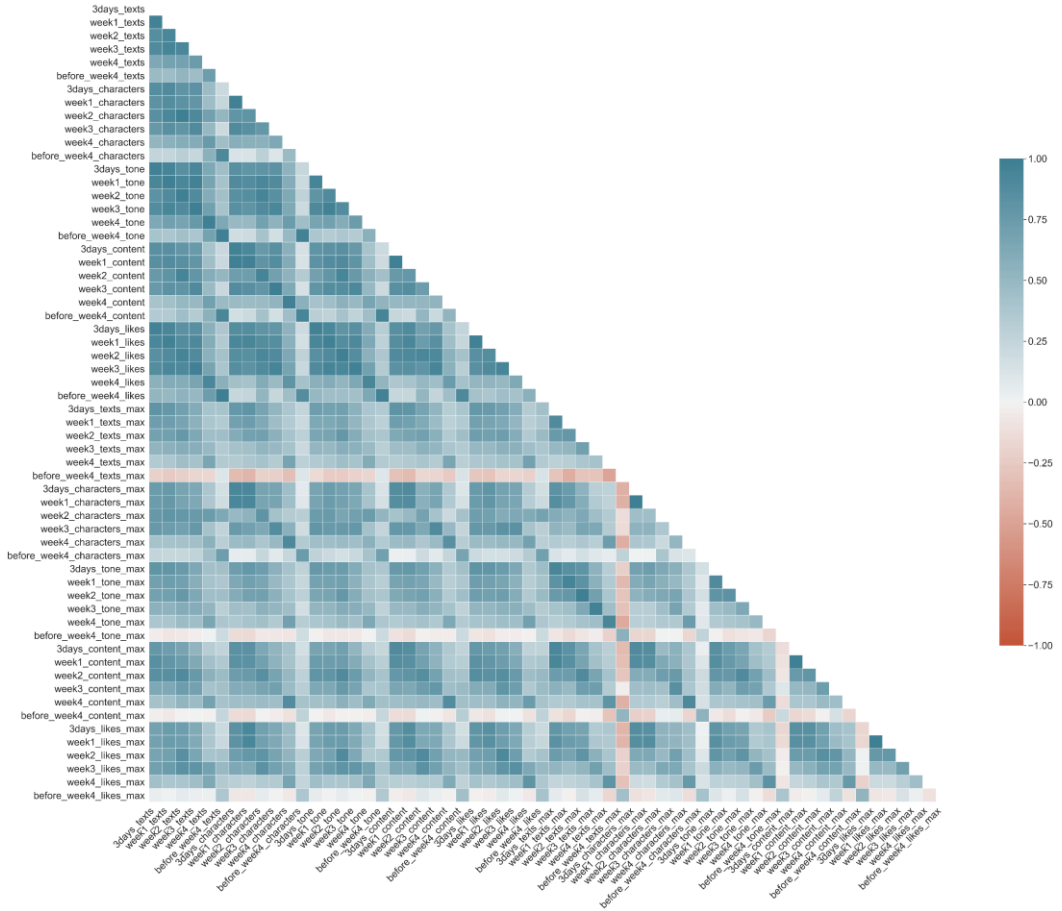
Time Series (original data)



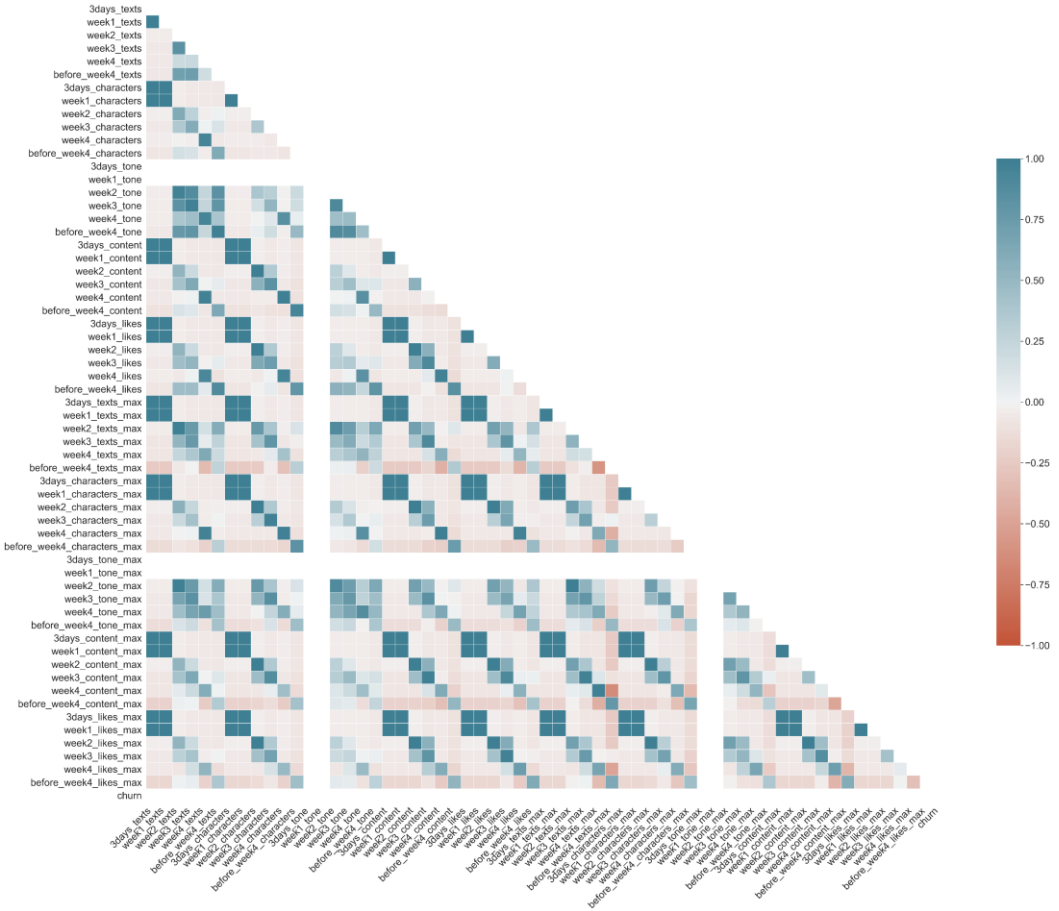




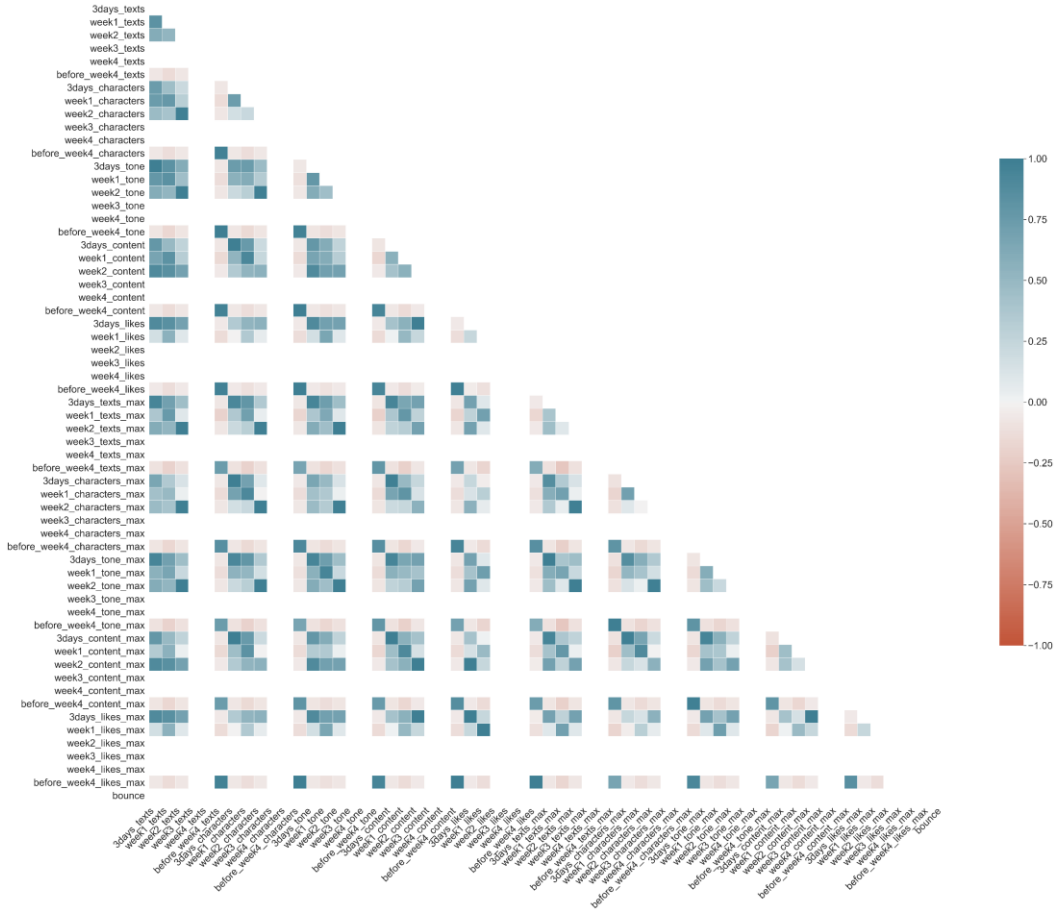
Feature correlation for stayed user data

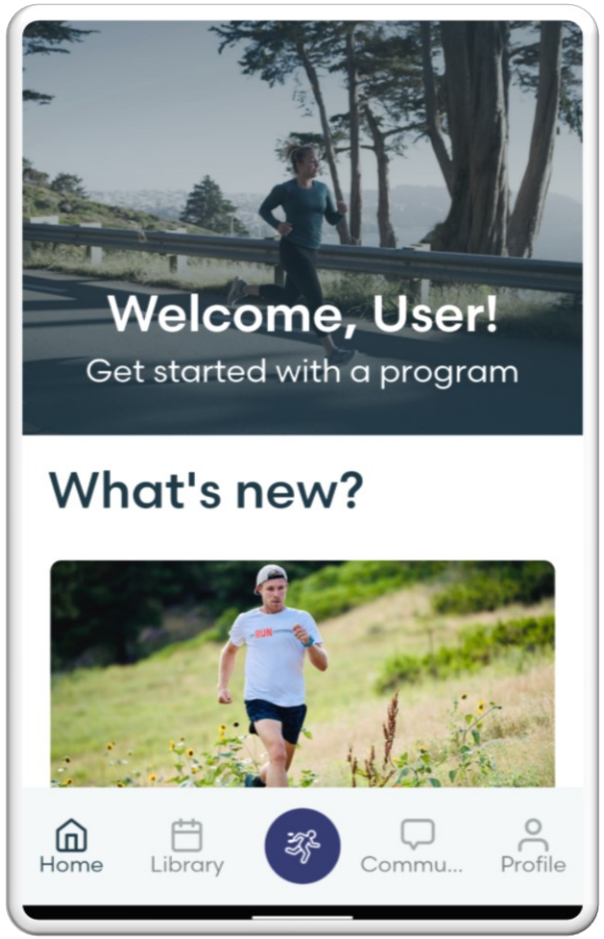


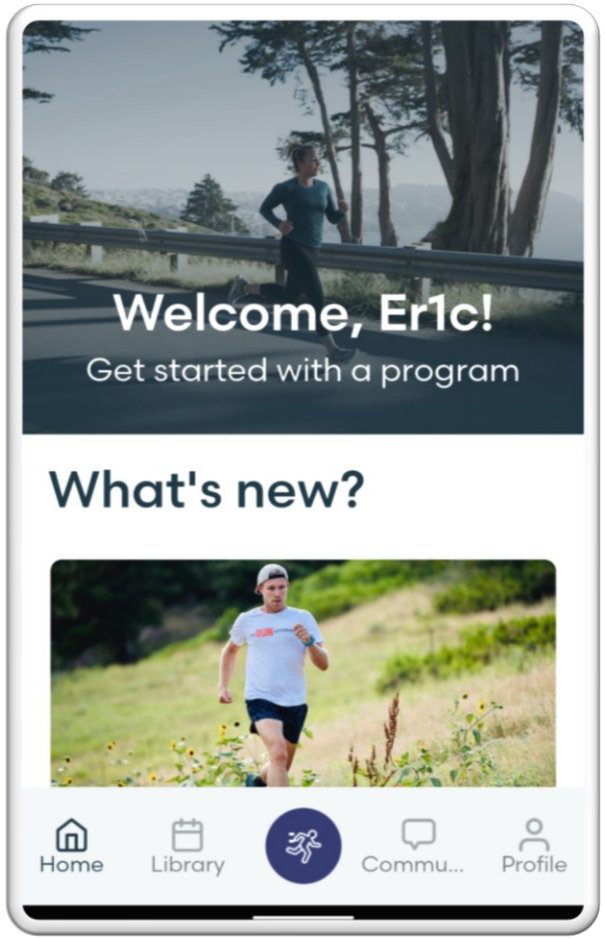
Feature correlation for churned user data

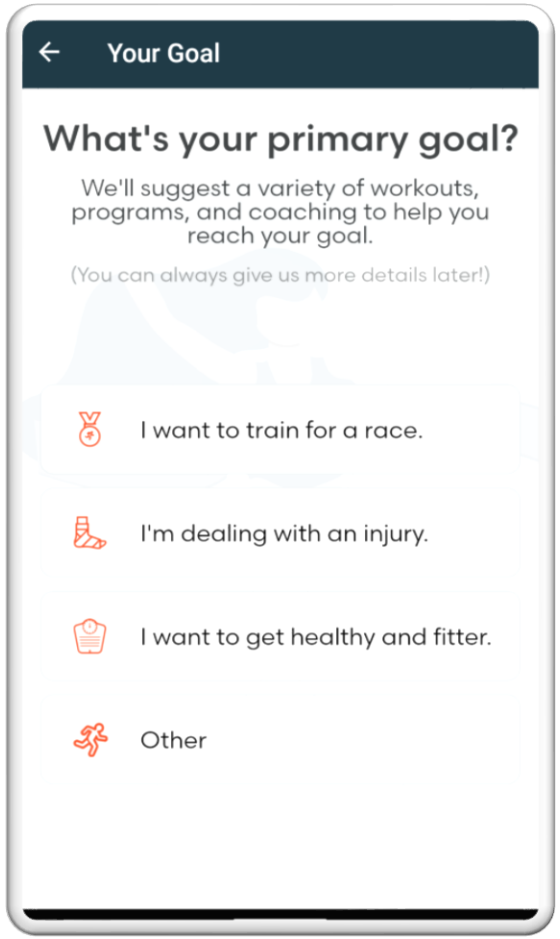


Feature correlation for bounced user data




Welcome, User!
Get started with a program
What's new?
Home
Library
Commu...
Profile


Welcome, Er1c!
Get started with a program
What's new?
Home
Library
Commu...
Profile


Your Goal
What's your primary goal?
We'll suggest a variety of workouts, programs, and coaching to help you reach your goal.
(You can always give us more details later!)
I want to train for a race.
I'm dealing with an injury.
I want to get healthy and fitter.
Other