Khan Academy User Retention Behavior Analysis

Nicholas Mai - Capstone Project 2

Motivation

- Khan Academy has users using their service across a variety of platforms and through a variety of languages.
- Taking a 4 day dataset of all user interactions the goal of this project is to determine which user behaviors or features best predict a return user.
- Knowing which features increase user retention would allow khan academy to focus on to increase return users.

Data Cleaning

- Each row represents a change in a user’s interaction with Khan Academy

- 8 columns are dropped due to missing values greater than 75%

- The columns of URI and Conversion were subsequently dropped later in the study as they were not used in the machine learning model

Cleaned Data

The dataset is taken from Feb 18 - Feb 21 2016 with 31480 total data entries.

The data is stored as a csv file. Other than identifier columns, all of the data is in text format.

An example cleaned entry:
Timestamp: 2016-02-18 18:05:44.033396 UTC
User_id : 461023995001001
Session_id: 7269247775762971847
Country: US
Language: en
User registered flag: True
Device type: desktop
KA app flag: False
OS : Windows

Defining a Return User

- Using the timestamp column a return user is defined by as someone who returns to Khan Academy after 4 hours of logging out of the service
- 4 hours was deemed as an efficient time difference as the time change between entries differ in seconds.
- Of the total 31,481 data entries, 21,790 are correlated to a return user
- A total of 432 unique return users

Exploratory Data Analysis (EDA)

Country - Over 50% of the return users are from the United States of America

Language - Over 95% of the videos viewed was in English

Registered Users - Over 90% were a registered Khan Academy user

Device Type - Over 70% of the users used a desktop

Khan Academy App - As most users used desktop over 90% did not use the application

Operation System - Windows was the most common operating system

Times returned - Most return users returned once

Feature Analysis

- A Chi-Squared test on each feature was made to test the correlation between a feature to a return user.
- ANOVA test was made to confirm the results of the Chi-Squared test in identifying features of importance .

Chi-Square Results

Feature	p-value	Statistical value
Language	2.9399 e-59	303.42
Country	0.0	2794.787
Operating System	1.7284 e-55	285.58
Khan Academy App	0.69	0.73
Device used	1.35198 e-21	104.05
Registered User	0.0	2769.91

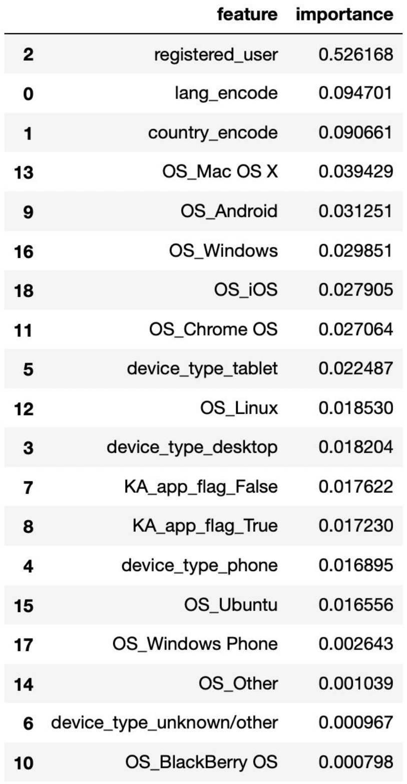
Random Forest Model

- A random forest model created to initiate the machine learning model and served as indicator to rank feature importance
- Features of OS and Device Type were one-hot encoded
- Language and Country were encoded of being in english and USA or not, respectively
- Sklearn was used to initiate the random forest classifier with an 80% to 20% test train split

The results, as expected, matched those found in the ANOVA test

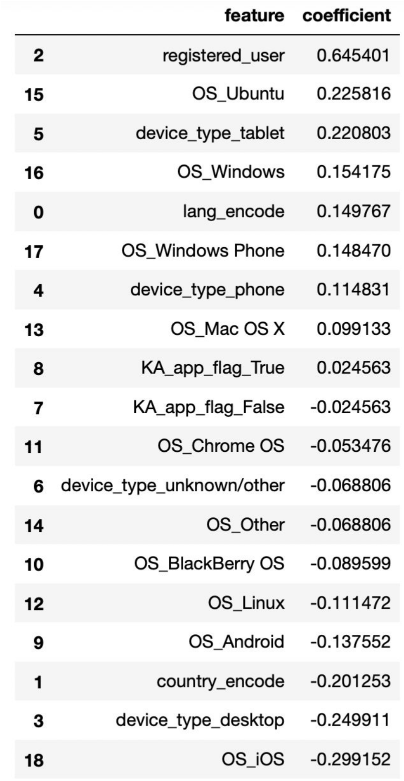
The most important features in predicting a return user are, registered_user, language, Country, and Mac OS

Random Forest Results



	feature	importance
2	registered_user	0.526168
0	lang_encode	0.094701
1	country_encode	0.090661
13	OS_Mac OS X	0.039429
9	OS_Android	0.031251
16	OS_Windows	0.029851
18	OS_iOS	0.027905
11	OS_Chrome OS	0.027064
5	device_type_tablet	0.022487
12	OS_Linux	0.018530
3	device_type_desktop	0.018204
7	KA_app_flag_False	0.017622
8	KA_app_flag_True	0.017230
4	device_type_phone	0.016895
15	OS_Ubuntu	0.016556
17	OS_Windows Phone	0.002643
14	OS_Other	0.001039
6	device_type_unknown/other	0.000967
10	OS_BlackBerry OS	0.000798

Logistic Regression Results



	feature	coefficient
2	registered_user	0.645401
15	OS_Ubuntu	0.225816
5	device_type_tablet	0.220803
16	OS_Windows	0.154175
0	lang_encode	0.149767
17	OS_Windows Phone	0.148470
4	device_type_phone	0.114831
13	OS_Mac OS X	0.099133
8	KA_app_flag_True	0.024563
7	KA_app_flag_False	-0.024563
11	OS_Chrome OS	-0.053476
6	device_type_unknown/other	-0.068806
14	OS_Other	-0.068806
10	OS_BlackBerry OS	-0.089599
12	OS_Linux	-0.111472
9	OS_Android	-0.137552
1	country_encode	-0.201253
3	device_type_desktop	-0.249911
18	OS_iOS	-0.299152

Logistic Regression

- A a logistic regression to provide coefficients to measure the percent increase per a unit for the top features of importance
- Logistic regression was chosen as the data is categorical, which provides more accurate results than a linear regression.
- Logistic results on features of importance:

registered_user = 0.64540, lang_encode = 0.149767, country_encode = -0.201253

- A registered user is associated with a 65% increased probability of becoming a return user.
- English is associated with a 15% increased probability of being a return user.
- Residing in the United States is associated with a 20% decrease probability of becoming a return user.

Conclusion

Using user login data on Khan Academy over a 4 day period the goal is to determine which features increase the probability of being a return user

A Random Forest Model was initiated in python, followed by a logistic regression

Behavioral recommendations for Khan Academy to increase return users:

1. Promote registration of users to become members/registered users
2. Focus on increasing content in english
3. Decrease promotion toward the United States user base

Further Improvements

- Increase the definition of a return user
  - 4 hours to 1 day
- Concern for multicollinearity due to high levels of significance found in feature analysis
- Gather more data over time