Nicholas Mai - Springboard Capstone Project 2 Report

Increasing Khan Academy's User Retention

Motivation

Khan Academy has users using their service across a variety of platforms and through a variety of languages. Taking a one month 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

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

Data includes: user id, session id, country of login, language used, user registration flag to Khan Academy, device type, operating system, whether the user used the Khan Academy application, URI, and conversion. The data also included columns for product, domain, subject, topic, tutorial, mission, video_slug, and video title.

Import Data

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
URI: "/welcome"
Conversion : pageview

To import the data, I created an empty dataframe with column names given in the original data.

Data Wrangling

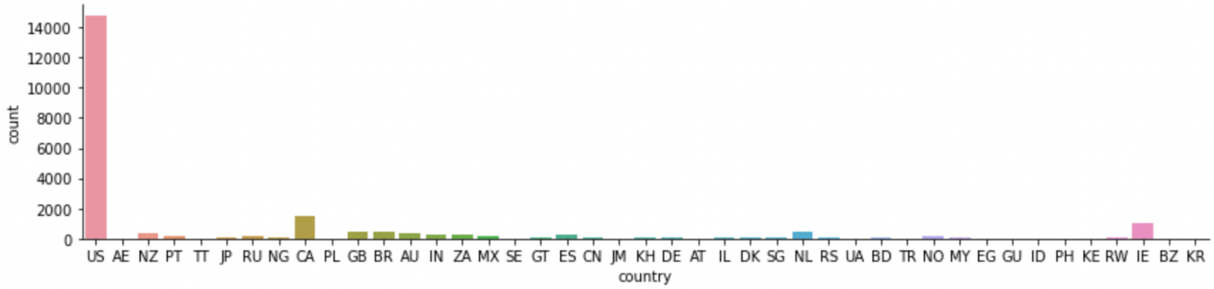
The first step is to check for missing values. The columns of product, domain, subject, topic, tutorial, mission, video_slug, and video title had a missing value count greater than 75% and in most cases above 90% and were dropped accordingly. URI and conversion could be used for future analysis but were not used for this study and were subsequently dropped.

The most important step in the data wrangling stage was to create a return user column. I defined a return user as someone who returns to Khan Academy after 4 hours. This was chosen as the data between entries differs in the seconds. Therefore, if a user leaves and returns in more than 4 hours he/she is a return user.
Of the total 31,481 entries, 21,790 are correlated to a return user. In total there are 432 unique return users.

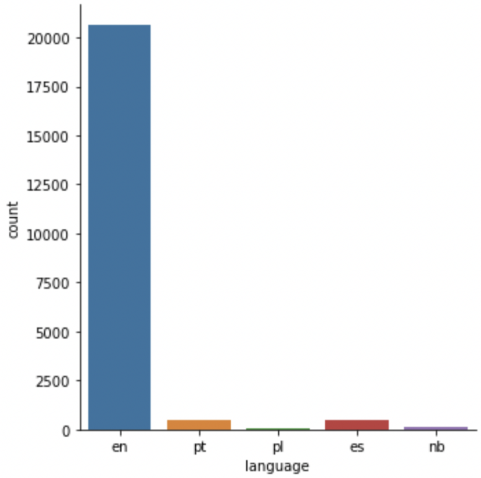
Exploratory Data Analysis

The exploratory data analysis is centered on finding information on the return user.

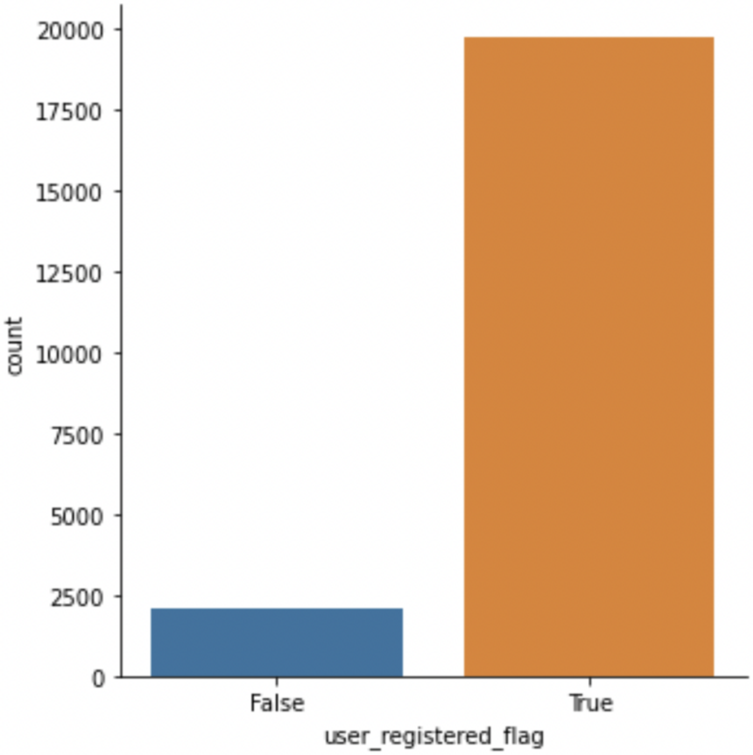
First, I looked at which country return users are from and there was an overwhelming majority towards users in the United States.



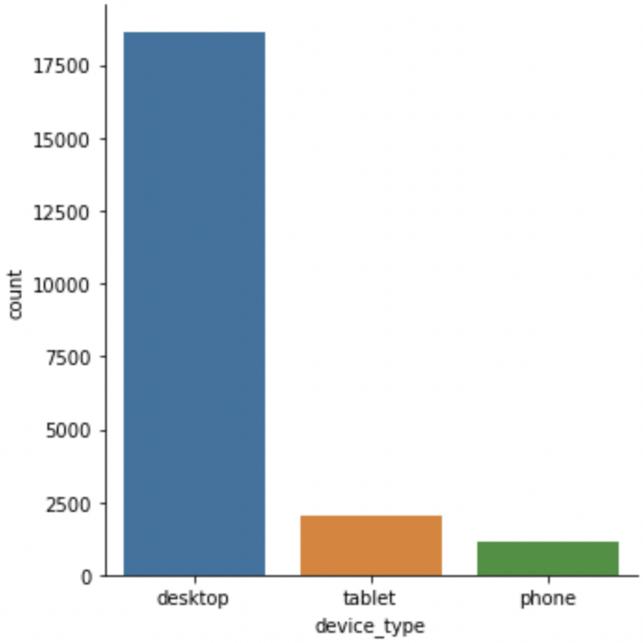
Next, I looked at language and as expected over 95% was in english



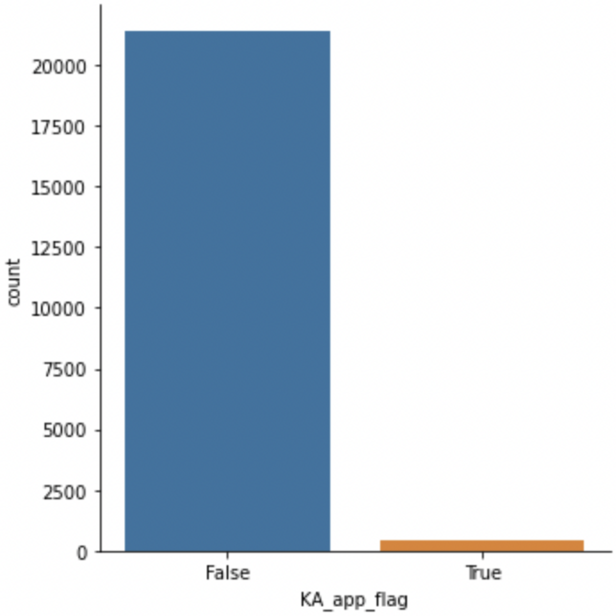
Most of the return users were registered Khan Academy users.



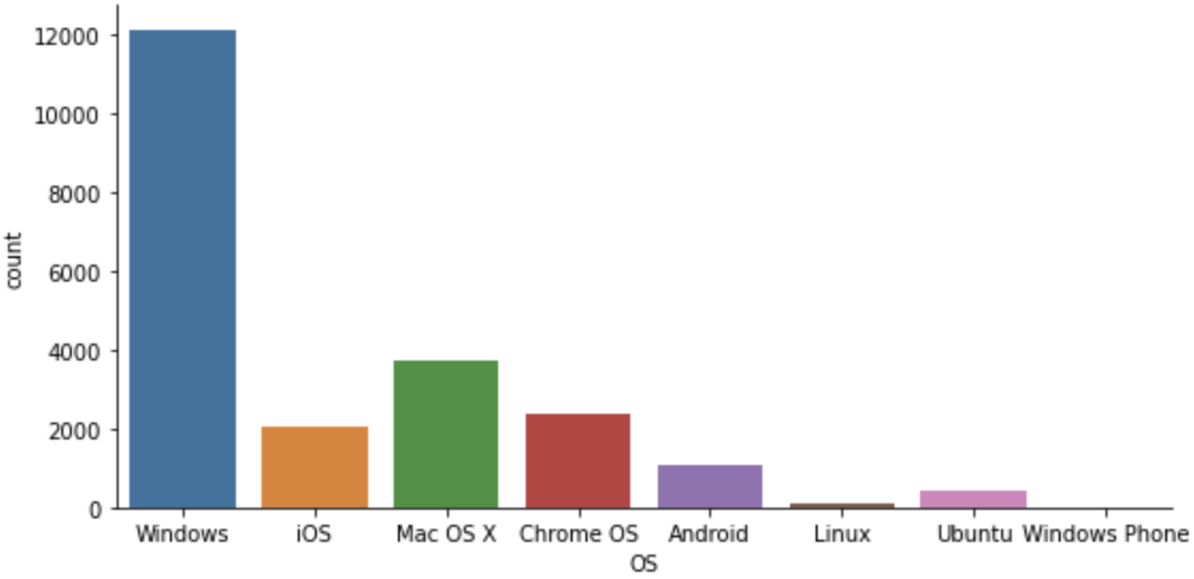
Next most of the return users were using a desktop.



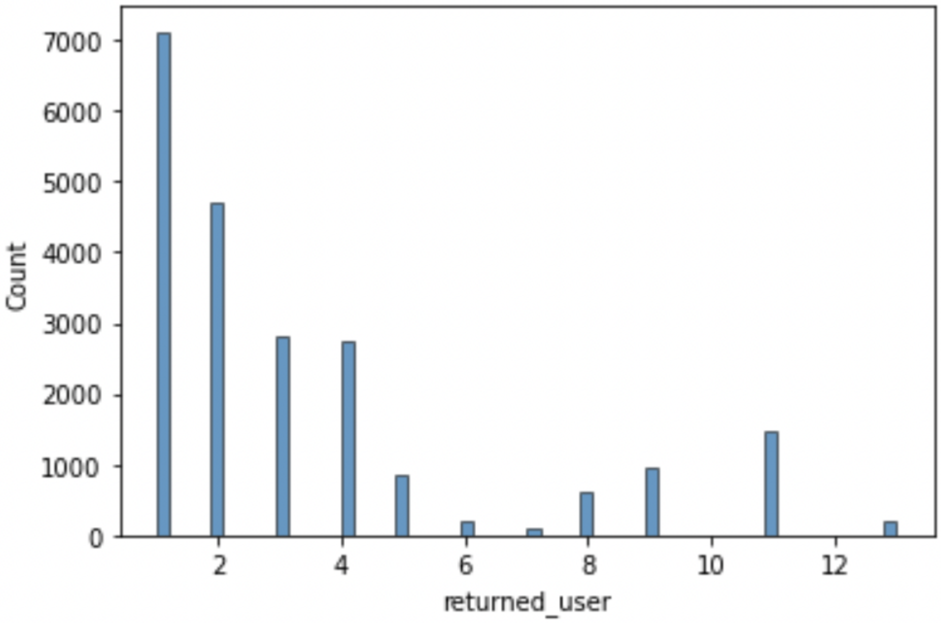
Likewise, since the Khan Academy app is a mobile app most of the return users were not using it since they were using desktops.



Windows was the most popular operating system, followed by mac os



Lastly, most of the return users only returned once.



Feature Engineering

Chi-Squared Test

The feature engineering section is meant as a hyperparameter test to see the relationship between two categorical variables. A Chi-square test was first used on every feature and compared to being a return user. At this point, a return user is defined as being ‘yes’ or ‘no’, thus, it is a 1 or 0 boolean value. Scipy was used to run the Chi-squared test. In this case, we care purely on the significance value derived from the p-value. Any value less than 0.05 is considered to be significant. Bolded values are significant.

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

I found high significance in all features except for the use of the Khan Academy App. The extreme significance is an indicator on the dataset being biased to returners over non-returners at a ratio of 80% to 20%

ANOVA

To confirm the results of the Chi-square test I ran a one-sided ANOVA model using statsmodels. To run the ANOVA the categorical variables need to have 3 or less possible variations. Therefore, the language, country, (operating system, and device used) was encoded to be 0 or 1. For example, in language, 1 represents english and 0 is other. For each country, 1 represents the USA and 0 is the other. For simplicity the top three statistical features of language, country, and registered user were stated.

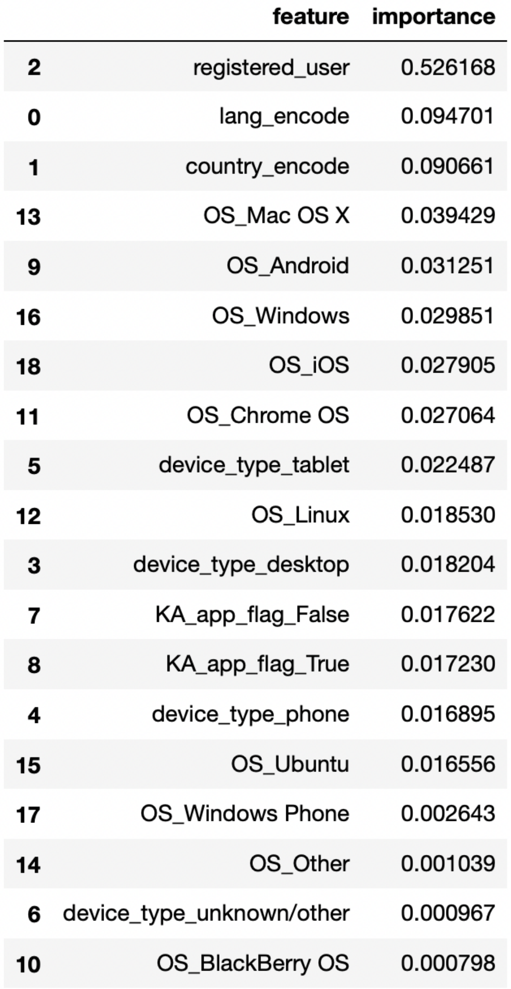
Feature	Levene Result Statistic	p-value
Language	440.79289	3.377217 e-97
Country	1027.5531	6.8655 e-222
Registered User	1300.79695	3.7065198 e-279

Model Comparison

Random Forest Model

From the statistical analysis in the feature engineering I then built a random forest model. In theory, the random forest model will rank feature importance and this should align with the chi squared and ANOVA test results.

To initiate the random forest all non-numerical categorical variables had to be dummy encoded so the model could provide a prediction, this one done with the one-hot code function. Sklearn was used to initiate the random forest classifier with an 80% to 20% test train split. As the data is already split 70% to 30% of retuners and non returners respectively a secondary verification that the test train split is maintained in both the test and split was made. Both the test and split maintained a 70/30 split and stratifying the model was not necessary.



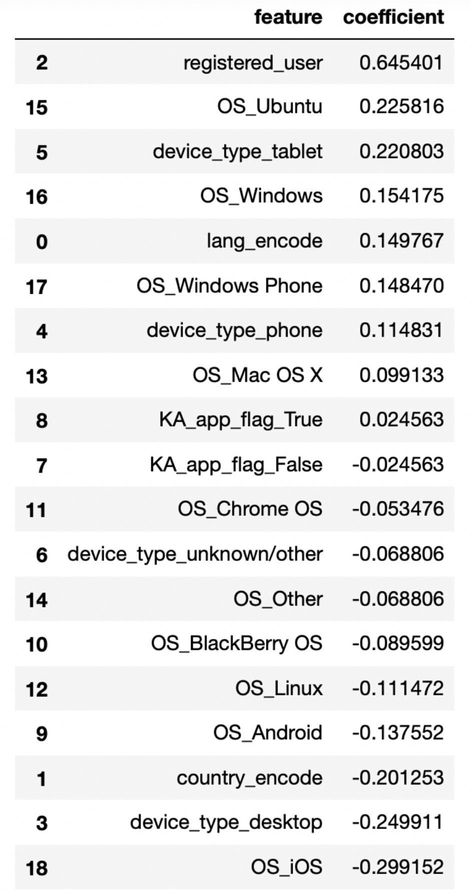
	feature	importance
2	registered_user	0.526168
0	lang_encode	0.094701
1	country_encode	0.090661
13	OS_Mac OS X	0.039429
9	OS_Android	0.031251
16	OS_Windows	0.029851
18	OS_iOS	0.027905
11	OS_Chrome OS	0.027064
5	device_type_tablet	0.022487
12	OS_Linux	0.018530
3	device_type_desktop	0.018204
7	KA_app_flag_False	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

As expected from our feature analysis the most important features in predicting a return user (defined as returning after 4 hours or more) are, registered_user, language, Country, OS, and Device type

The first three have the highest level of importance at 0.5, 0.09, 0.09 respectively. Lastly, the accuracy of the random forest model was at 75%, which is satisfactory. In conclusion, these results match with our results from Chi-Squared test and ANOVA test

Logistic Regression

The next step is to take our random forest model and apply a logistic regression to provide coefficients to measure the percent increase per a unit for the top features of importance. A logistic regression was chosen as the data is categorical by nature and a linear regression would less account for outliers.



	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

Conclusion

The Random Forest model predicted a returner with 75% accuracy and listed features by importance. The features of most importance were registered user, language, and country. Registered user is defined as having a registered account with Khan Academy. Language in the model is defined as being in english or not. Country in the model is defined as being from the US or not.

Country had 77 unique values. The USA consists of 47% of those values, with the second most Canada consisting of 5% of the sample. Language, similarly, had 11 unique values with English comprising 94% of the sample.

Next to find the predicted results a logistic regression was used to derive interpretable predictions when predicting a return user. The regression had an accuracy of 74%, thus, can be used in hand with the Random Forest model.

Logistic results on features of importance: registered_user = 0.64540, lang_encode = 0.149767, country_encode = -0.201253

According to the logistic results being a registered user is associated with a 65% increased probability of becoming a return user. Having the content in English is associated with a 15% increased probability of being a return user. Being in the United States is associated with a 20% decrease probability of becoming a return user.

In conclusion, our behavioral recommendations for Khan Academy to increase return users according to our model is to promote registration of users to become members/registered users, continue focusing on increasing content in english and to some degree not focus on increasing membership in the United States. With more data and a constrained definition of a returned user would provide better predictions for future analysis.

Future Improvements

Future improvements to this model would be to further restrict the definition of a return user, such as increasing last login time from 4 hours to 1 day. In doing so, one could have a better ratio of return users and non-return users, and could possibly decrease the significance levels found in the Chi-square and ANOVA tests. The high significance found in these tests are of a concern for multicollinearity. The best solution for future improvements is to have more data. Currently, the data is for a 4 day period, which could be causing these extreme significance levels we are seeing currently.