Insight Website Conversion Rate Analysis

Matt Chan

September 12, 2020

1. Introduction

We have been provided a dataset summarizing the users demographic information, the number of pages each user explored on the Insight website, and whether not they made a purchase on the website (*i.e.* "converted"). The goal of this analysis is to use a model to determine and evaluate parameters that can increase conversion rate.

The data contained the following columns:

Feature	Non-null entries	D-type
Country	316200	Object
Age	316200	Int64
New_user	316200	Int64
Source	316200	Object
Total_pages_visited	316200	Int64
Converted	316200	Int64

2. Exploratory Data Analysis

First, we would like to take a quick look at the demographic breakdown of the users in the dataset. There are four categories we can examine:

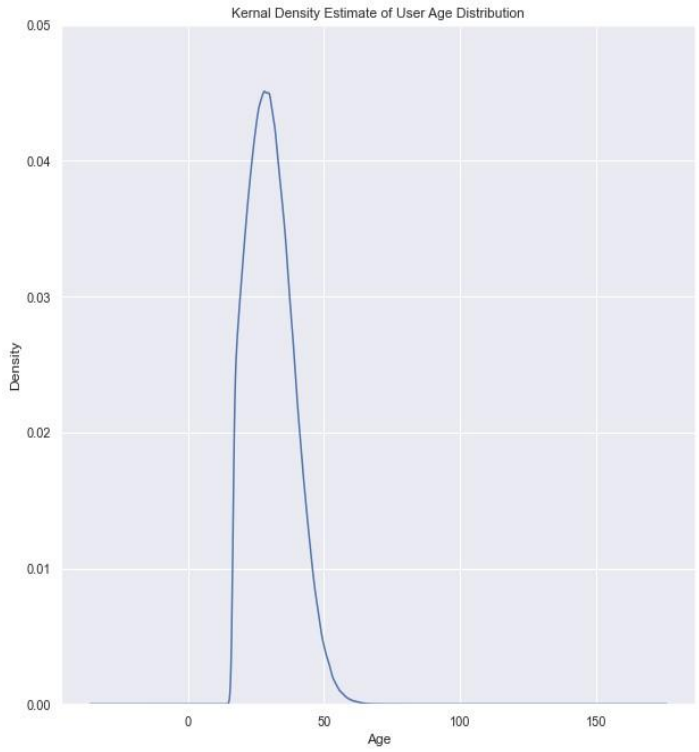
- Country of origin
- Whether or not they are a new user
- Where they found out about the website
- Age

Pie charges aren't terribly quantitative, but they serve as a good quick glance of demographic breakdown for each category with the exception of age. For age, we will use an area graph to observe the rough distribution of ages.

Several quick observations here:

- More than half of the visitors of the websites were from the US.
- Only about 1/3 of the reported users were returning users.
- Almost half of the users discovered the website through a search engine of some sort
- User age peak at around 30; there is an abrupt drop off to nothing at around age 16 or so, and also a huge drop-off towards 50 years old, and tapering off upward.

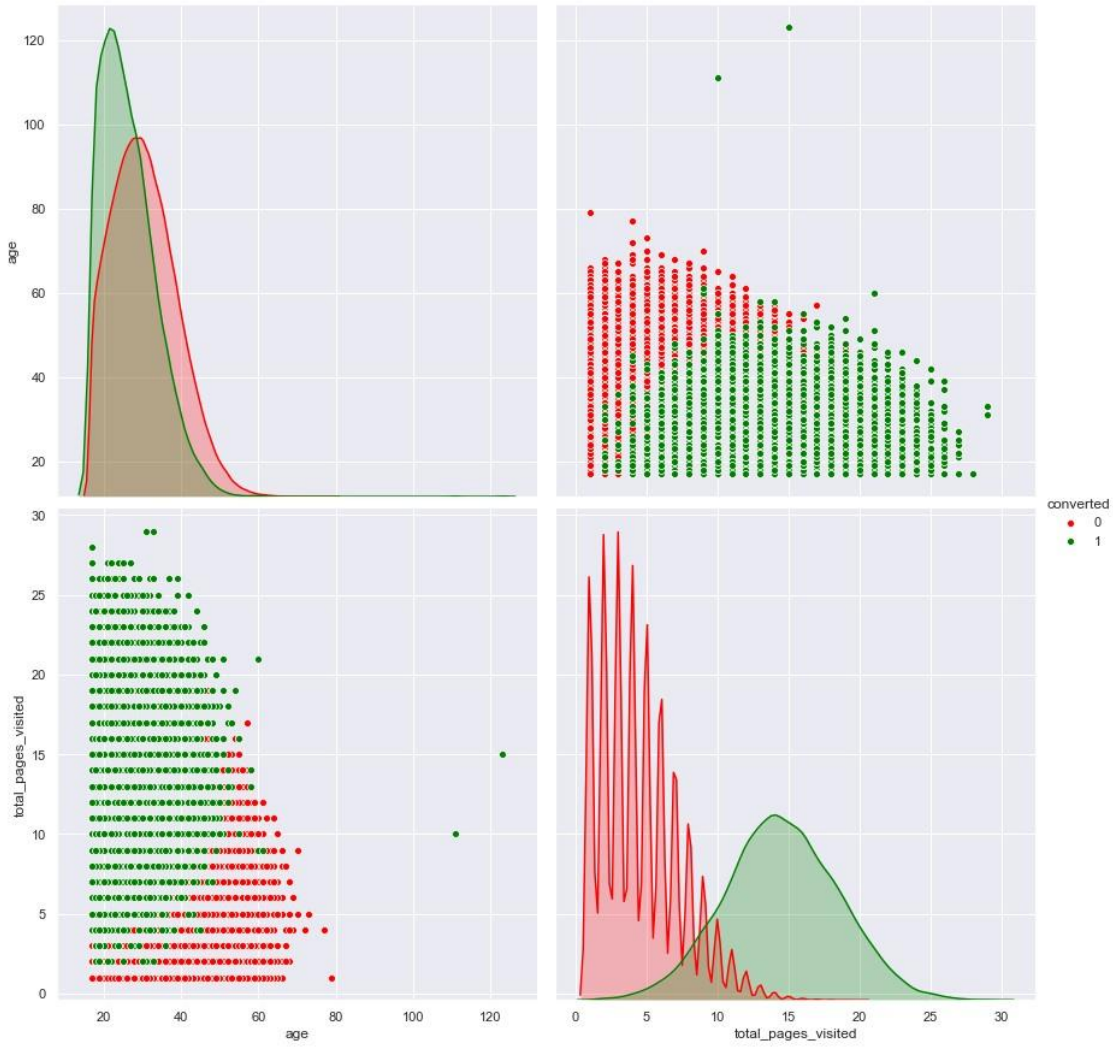
The age distribution warrants a deeper look. We can fit a kernel density estimate plot through the age distribution. This is akin to a histogram but with no “bins”.



The following is the summary statistics of the age distribution of the users:

Statistic	
Number of users	316,200
Mean	30.57
Standard deviation	8.27
Minimum	17
25% quantile	24
Median	30
75% quantile	36
Maximum	123
Kurtosis	-0.0324
Skewness	0.519

Let us now check out the pair-wise correlation between each data column. Note that we will skip categorical (except Boolean columns) data – *i.e.* 'country'' and 'source' – for now, as we will have to generate dummies to evaluate the correlations.



There are lots that can be immediately observed regarding the correlation between the features as well as between each feature and the class (1 = converted, 0 = no conversion.):

1. There seem to be little correlation between age and conversion; if there are any trend, younger users were converted more than older ones.
2. Generally, converted users tend to have visited more pages (higher engagement.)
3. There is a slight trend that younger users tending to visit more pages.

In summary, there is stronger evidence suggesting a positive correlation between engagement (as represented by number of pages visited) and conversion. The correlation coefficients (Pearson's) between each numerical features and conversion.

Feature	Age	Converted	Total pages visited
Age	1.00	-0.046	-0.089
Total pages visited	-	1.00	0.528
Converted	-	-	1.00

- P-value of [Age] ↔ [Conversion]: 0.0
- P-value of [Total Pages Visited] ↔ [Convertion]: 0.0
- P-value of [Age] ↔ [Total Pages Visited]: 3.4636139994442017e-147

The very low or 0 p-values indicate statistical significance of these correlations, although only the positive correlation between total pages visited and conversion is strong.

Next, we can use the χ^2 test within a contingency table to examine possible relationship between the categorical variables and conversion.

The result of the χ^2-test between ``country" and ``converted" is as follows:

χ^2 Statistic: 3549.0

P-value: 0.0

Observed distribution versus expected distribution if there are no relationship between ‘country’ and ‘converted’:

	Observed unconverted	Expected Unconverted	Observed Converted	Expected Converted
China	76,500	74,131	102	2,471
Germany	12,240	12,635	816	421
UK	45,900	46,887	2,550	1,563
US	171,360	172,347	6,732	5,745

The outcome of the test suggest that there are significant potential for some relationship between ‘country’ and conversion, so we will need to include this variable in our model. We can do the same with ‘source’.

χ^2 Statistic: 54.87

P-value: 1.22×10^{12}

Observed distribution versus expected distribution if there are no relationship between ‘source’ and ‘converted’:

	Observed unconverted	Expected Unconverted	Observed Converted	Expected Converted
Ads	85,680	85,877	3,060	2,863
Direct	70,380	70,084	2,040	2,336
SEO	149,940	150,038	5,100	5,001

The outcome of the test suggests that there are significant potential for some relationship between ``source'' and conversion, albeit slightly less dramatic as ‘country’. We will also include this variable in our model. Finally, we will run the same test for ‘new_user’.

The result of the χ^2-test between ‘new_user’ and ‘converted’ is as follows:

χ^2 Statistic: 7340

P-value: 0.0

Observed distribution versus expected distribution if there are no relationship between `new_user' and `converted':

	Observed unconverted	Expected Unconverted	Observed Converted	Expected Converted
Returning users	92,295	96,248	7,161	3,208
New users	213,705	209,752	3,039	6,992

The outcome of the test again suggests strong potential for some relationship between ``new_user'' and conversion.

Summary of statistics between features and conversion

Based on our examination of the features, we will need to include all features in our model.

Feature	Test used	Outcome	Action
country	χ^2 test	Strong evidence of contribution	Implement in model
age	Pearson's Correlation	Strong evidence of slight negative correlation	Implement in model
new_user	χ^2 test	Strong evidence of contribution	Implement in model
source	χ^2 test	Strong evidence of contribution	Implement in model
total_pages_visited	Pearson's Correlation	Strong evidence of strong positive correlation	Implement in model

3. Implement Machine Learning Model

Because we are essentially trying to determine whether a user will be converted, this is a binary classification problem. Since there are some features that are continuous and numerical, while some features are categorical, we will have to combine different preprocess difference features somewhat differently.

We will split our dataset into a training set and a test set, then use a pipeline to streamline the model building process. For the numerical features, we should scale them to unit length with scikit-learn’s `StandardScaler`, while with the categorical features we need to build dummies with `OneHotEncoder`. We will use `ColumnTransformer` to select the numerical columns and categorical columns separately, and apply the `StandardScaler` and `OneHotEncoder` to them respectively, then use `FeatureUnion` to join them together. The two different transformers will run

parallel within the `FeatureUnion` before joining the transformed column together. We have named this combined transformer `preprocess_transformer`.

Because we have robust number of samples, we use a linear support vector cassifier (LinearSVC) as a model. Here are the relevant arguments for the estimator:

- `penalty`: The L2 norm is standard for LSVC and we will stick with it here.
- `loss`: Loss function. We will also stick with the default squared_hinge.
- `dual`: Whether the algorithm should solve the dual or primal optimization. Scikit-learn recommends setting this to `False` if `n_samples` > `n_features`.
- `tol`: Tolerance for stopping criteria. We can try to be more robust by specifying 1e-7.
- `C`: Regularization parameter. We will need to cross-validate for a proper value.
- `max_iter`: We can be more robust and set this to 10000.

For cross-validation of `C`, we will utilize scikit-learn's `RandomizedSearchCV`. Here is the resulting best estimator:

```
 Pipeline(steps=[
      ('preprocess_transformer',FeatureUnion(
           transformer_list=[
                ('one_hot_transformer',ColumnTransformer(
                     transformers=[
                          ('OneHotEncoder',OneHotEncoder(),
                            Index([
                                'country', 'new_user', 'source'
                                ],
                            dtype='object'))],
                  verbose=True)),
                ('scaler_transformer', ColumnTransformer(
                     transformers=[
                       ('StandardScaler',StandardScaler()
                           Index([
                              'age','total_pages_visited'
                           ],
                         dtype='object'))],
                  verbose=True))],
           verbose=True)),
      ('linear_svc_est', LinearSVC(
           C=0.0005044188589670284,
           dual=False,
           max_iter=10000,
           tol=1e-07,
           verbose=5))],
verbose=True)
```

Our cross-validation has yielded an optimal `C` at 0.0005044188589670284, with a score of 0.9853341766814253. Using this estimator, we can now check our model with the test set, which yielded

a score of 0.9860974067046173. This indicates that our model also performed well on the test set. We can now examine the features that contribute most to whether a user is converted.

Coefficients	Values
x0_china	-0.487274
x0_germany	-0.022066
x0_uk	-0.070400
x0_us	-0.155446
x1_0	-0.189205
x1_1	-0.545982
x2_ads	-0.227175
x2_direct	-0.257335
x2_seo	-0.250677
age	-0.123159
total_pages_visited	0.533884

Interestingly, the only feature that has a (significant) positive contribution to conversion is total_pages_visited. All other features have a negative contribution. Of all the negatively contributing features of the model, being from China and being a returning user has the greatest negative contributions.

4. Recommendations

Analysis from this linear classification model indicated that to maximize conversion, Insight should implement strategies that increases user engagement on their websites. Users that visited more pages on the website has a high determining factor of their conversion. If a user is returning, data suggest they will not make another purchase. We will need further data to examine this interaction, but perhaps this is because any purchases a user might make would be during their first visit; a returning user might have *already* made their purchases (*i.e.* already converted.) If a user is from China, that characteristic alone might lead them to be less likely to make a purchase. This may have to do with the products and services provided by Insight, which may not be applicable to users in China. Finally, in general most demographic features of a user (age, other countries of origin) do not have significant contribution to conversion based on our model, nor do how they have found out about the website.

Overall recommendations:

- Implement strategies to retain and increase user engagement on the website. This is measurable with an increase in pages visited per user.
- Products and services seem to have American and European appeal, but not to users from China. Insight might consider remedying their product/service signaling and characteristics to appeal specifically to this demographic.

5. Future work

This model has not examined interplay between features. A model that considers polynomial features might reveal more insight to conversion-rate increase. Specifically, it will be prudent to examine if there are any relationships between total_pages_visited and other features.