Conversion Rate

20200529 Chunlei Zhou

Project Introduction

Goal:

1. Predict conversion rate;
2. Come up with recommendations for product team and the marketing team to improve conversion rate.

Data:

The dataset 'conversion_rate.csv' has information about signed-in users during one session. Each row is a user session.

Columns:

country: user country based on the IP address;
age: user age. Self-reported at sign-in step;
new_user: whether the user created the account during this session (1) or had already an account and simply came back to the site (0);
source: marketing channel source:
1). Ads: came to the site by clicking on an advertisement
2). Seo: came to the site by clicking on search results
Direct: came to the site by directly typing the URL on the browser
Total_pages_visited: number of total pages visited during the session. This is a proxy for time spent on site and engagement during the session
converted: this is our label. 1 means they converted within the session, 0 means they left without buying anything, in other words, not converted. The company goal is to increase converstion rate: #conversions/total sessions

Initial Plan:

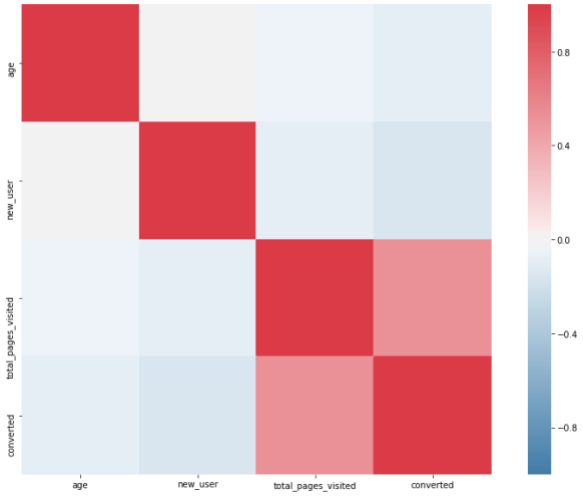
Use supervise learning to build a model and recognize users that are likely to be converted.
Use the model to cluster user sessions to convert and not convert.
Count converted sessions and divided by total number of sessions.

Data overview:

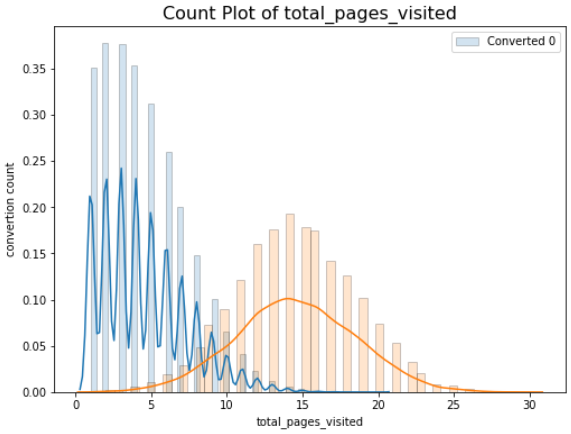
```
RangeIndex: 316200 entries, 0 to 316199
Data columns (total 6 columns):
country                 316200 non-null object
age                     316200 non-null int64
new_user                316200 non-null int64
source                  316200 non-null object
total_pages_visited     316200 non-null int64
converted               316200 non-null int64
dtypes: int64(4), object(2)
```

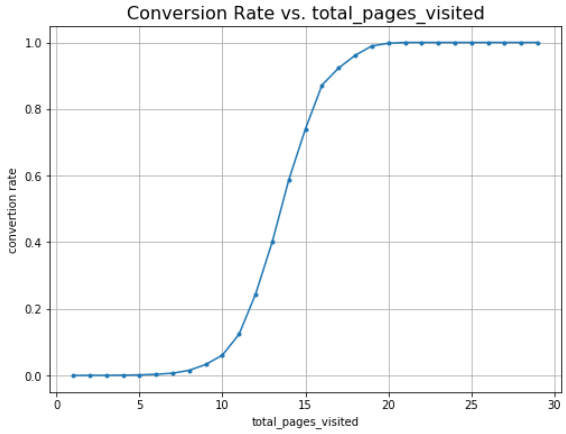
Exploratory Analysis:

Based on the result of correlation analysis of numerical features, as shown in figure 1, features are independent and among all, total pages visited influenced the conversion most.

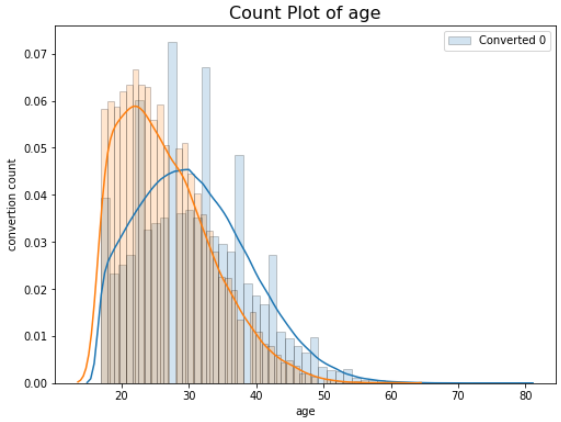


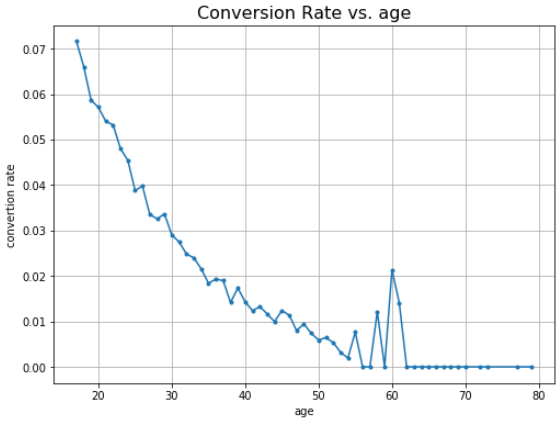
Users visit more pages tend to have a higher conversion rate.



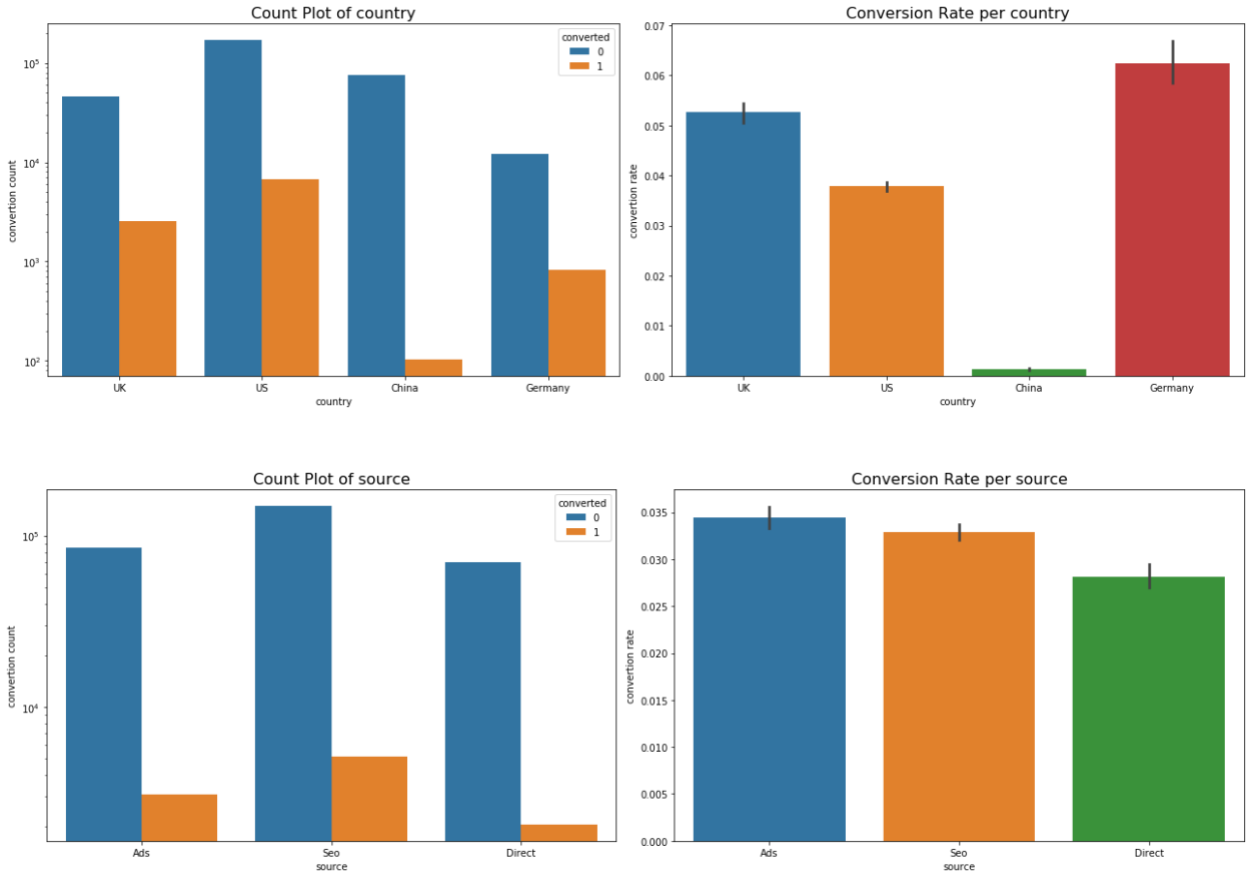


Customers with age in range of 15 to 30 tend to have higher conversion rate.





As for categorical feature, according to the bar plots shown below, in which country is the user also influence the conversion. The user source does not show too much impact on the conversion.



Method:

After converting categorical features to number by one-hot encoding, I divide the dataset into train, dev, and test set by the ratio of 7:2:1. The training set is used to train models, dev set is used to select the best model, and the test set is used to make the final prediction.

The project can be considered as a binary classification problem, thus, I trained three models: logistic regression, SVM, and Random Forest.

According to the performance of the three models on the dev set:

```
Accuracy:
- Logistic Regression: 98.62%
- SVM: 98.47%
- Random Forest: 98.71%
```

The model selected is Random Forest.

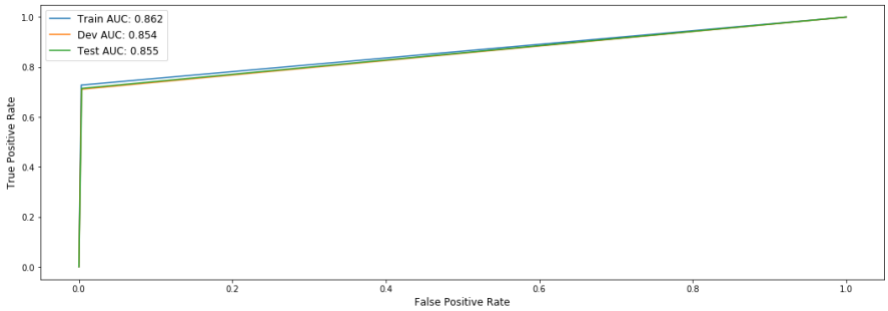
Results:

The performance of the model is acceptable:

```
Prediction Accuracy:
- Random Forest: 98.73%
```

	precision	recall	f1-score	support
Not Converted	0.99	1.00	0.99	30605
Converted	0.87	0.71	0.78	1015
accuracy			0.99	31620
macro avg	0.93	0.86	0.89	31620
weighted avg	0.99	0.99	0.99	31620

ROC curve was plotted for all three sets to check whether there is overfitting or not.



The performance of the three sets is almost the same, thus, there is no overfitting.

Based on the analysis, we can conclude that the model is reliable, since it gives satisfying prediction accuracy and does not overfit the data.

The prediction of the model is:

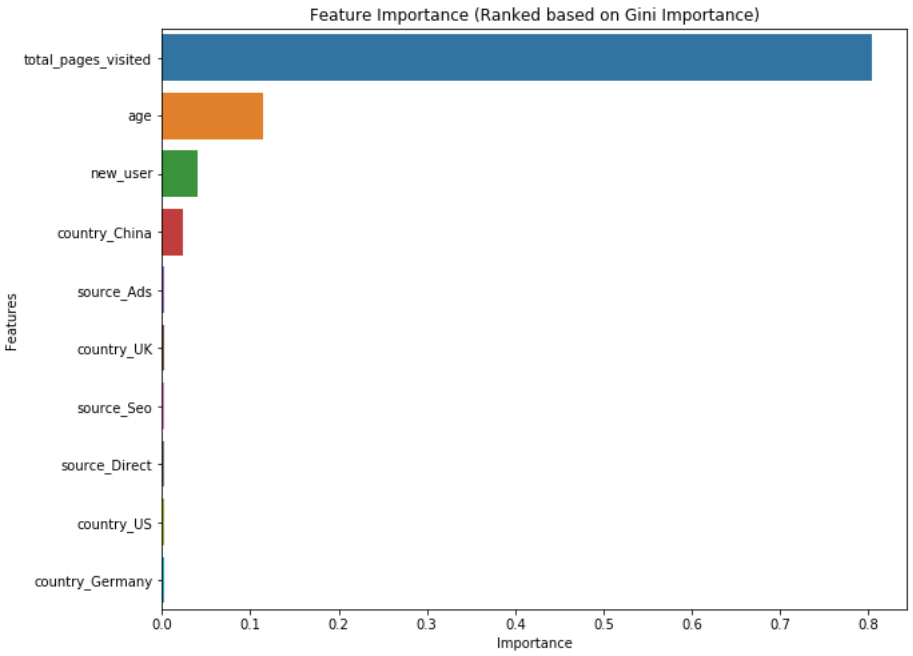
```
Conversion Rate (Test Set):
- Predicted: 2.65%
- Actual: 3.21%
```

Recommendations:

To give recommendations to product team and marketing team, I conducted analysis regarding feature importance.

```
Feature ranking:
```

	Gini_importance
total_pages_visited	0.803801
age	0.114779
new_user	0.040491
country_China	0.024038
source_Ads	0.003204
country_UK	0.003062
source_Seo	0.003060
source_Direct	0.002913
country_US	0.002520
country_Germany	0.002131



Suggestions for product team:

1. It seems that users in China tend to have lower conversion rate, it might because the product does not suit Chinese customers (the reason still need to be further investigated). To improve conversion rate, you may design a subversion of your product that suits Chinese people more.
2. It seems that your product is more popular in users under the age of 30 than elder users, it might because that the product is less user-friendly to elder people or does not meet the need of elder people (the reason still need to be further investigated). To improve the conversion rate, you may design your product more senior friendly.

Suggestions for marketing team:

1. The more pages visited, the higher the conversion rate. This fact may suggest that the marketing team should put ads for the product on more pages that has no ads yet.
2. The postcensal of Chinese market is huge and should be explore more. You can design some ads that is attractive to Chinses customers and try to put more ads on some popular Chinese websites.
3. The product seems to be age sensitive, you can either put more efforts on attract users under age 30 or pay attention to attract elder users by designing promotion activities targeting users with certain age.
4. New user acquisition is importance, by putting more ads on different pages that was has no ads regarding your product yet, you may attract new users.