I. Exploratory Data Analysis

1) Introduction

The data include 5 features characterizing users visited the website. Their country, age, the total number of pages visited, the way they access the website, and account creation during the session are available from the data. To extract interesting insight, the features are distributed depending on the conversion.

2) Undersampling

The data are highly unbalanced in terms of the conversion. To mitigate the impact of this in analysis, an under sampling is performed. The users without conversion are under sampled to the total number of the users with conversion as shown in Figure 1.

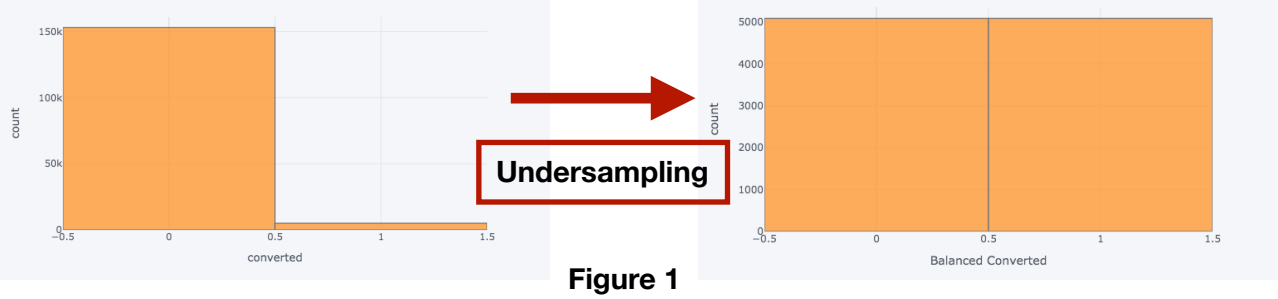


Figure 1

3) Features vs conversion

Features are distributed depending on conversion as shown in Figure 2.

a. Age - the top left plot presents that younger users tend to make more conversions than elder users.

b. Country - the top middle plot shows that users from US, Germany, and UK tend to make conversion. But users from China show a very low conversion rate.

c. Account creation - users who already have an account with the website tend to make more conversion than the new users.

d. Sources - in general, the website receives visits through search. The conversion rates of users visiting through ads and searches show no big difference. In the case of the users visiting through direct access to the website, they tend to show a lower conversion rate.

e. Page_visited - clearly users who visit many pages tend to make more conversion than users who visit less pages.

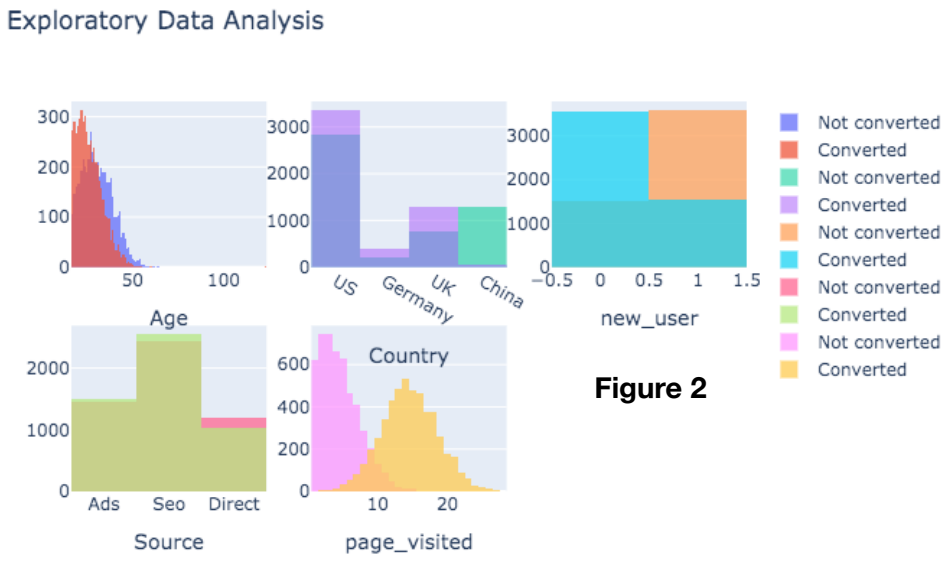
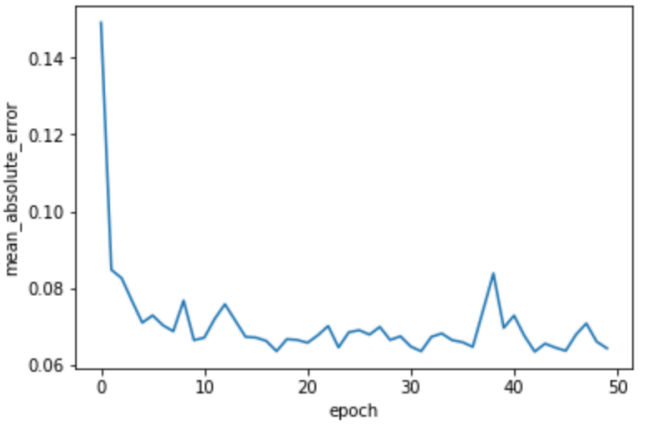


Figure 2

II. Conversion prediction

1) Neural network model

a neural network model is built to predict conversion rate. The 4 layers and one output layer compose the neural network. The network is fed 10 features and predicts the conversion in the output layer which has one neuron in it. As the neural network presents classification output with sigmoid activation, large losses from outliers are not expected. So, to evaluate the loss, mean absolute error is calculated in each epoch. The loss approaches the stable regime after the 10th epoch.



2) Data and one-hot encoding

The training samples are half of the under sampled data and the test samples are the rest of the under sampled data. To incorporate the categorical features(country and source), one hot encoding is performed.

3) Evaluation

Accuracy, precision, recall, and f1 scores are evaluated to quantify the performance of the network. Overall, the network achieves 93% of accuracy. Also, to make sure there is no artificial impact due to unbalance in data, precision, recall, and f1 scores are calculated as well and they present about 0.94.

	precision	recall	f1-score	support
0	0.91	0.97	0.94	5100
1	0.96	0.90	0.93	5100
avg / total	0.94	0.94	0.94	10200

accuracy : 0.9350980392156862

III. Recommendations to improve conversion rate

1) The product team should add more items which are popular for younger customers to the website.
2) The largest visits come from search. The marketing team should think about how to obtain a better ranking in search results. Also, the visits through AD seems not that effective compared to the direct visit.

3) Conversion rate is very low from Chinese customers even though the number of visits is comparable to UK visits. The product team should consider adding Chinese friendly items or environment to the website
4) Users who already have an account tend to make more conversion than new users. The marketing team should consider more offers to new users.