The following report explains the steps taken for training a model aimed at increasing the customers conversion rate.
The data used has been collected during the user sessions and contains information relative to user:

- country,
- age,
- new_user (whether the user is new or not),
- source (how the user came to the site),
- total_pages_visited,
- converted (whether the user converted or not).

Model choice

A Logistic Regression classifier is chosen to model the data because of its simplicity and computational cost.
The column *converted* is used as a label, all the others as features.
The classifier is optimized using Recall as a metric, because the focus of the model should be to maximize the chance of identifying positive cases.

Data cleaning

The data set is converted into a pandas data frame to make exploration and manipulation easier.

During the exploration phase, some odd data points in the column *age* have been identified, therefore the column is limited to 100 and the respective rows are eliminated.

The data frame does not contain any null values, so no imputation is required. However, a class imbalance emerges in the labels (number of 0s grater than number of 1s), which will be addressed when training the classifier.

Features engineering

The features that are already binary are left untouched (*new_user*, *converted*).

The features that are non-numeric are hot-encoded (*country*, *source*): hot-encoding is preferred because the features are not ordinal.

The continuous features (*age*, *total_pages_visited*) are scaled using MinMaxScaler, given the dependency of the Logistic Regression classifier on the magnitude of the features.
MinMaxScaler is preferred as a scaling method because the distribution of the variables is not Gaussian.

Model training

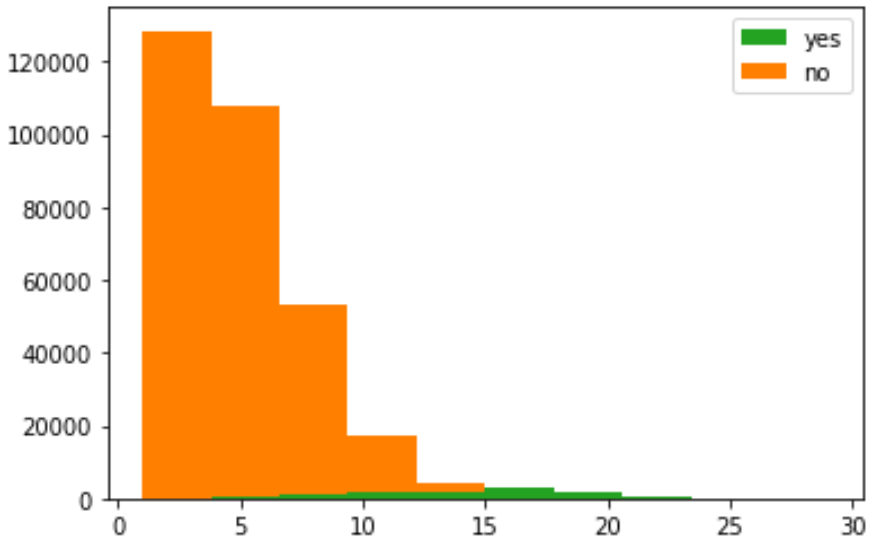
The data set is divided in test (80%) and train (20%) sets, and the scikit-learn cross validated Logistic Regression classifier is trained using a 5-fold cross-validation. The option ‘balanced’ for the class weights is specified, to account for the label imbalance.

Conclusions

The absolute values of the model coefficients are used to determine the feature importances.

From this, some conclusions can be drawn, that were partially uncovered during the exploration phase:

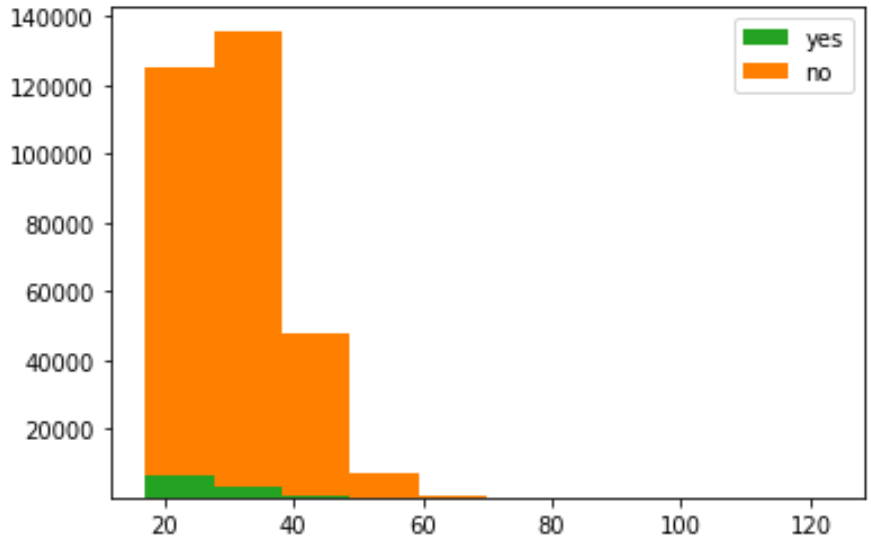
- the number of total pages visited during a session seem to be positively correlated with the conversion rate.



Number of pages visited per user conversion (yes or no).

Making it easier for the user to navigate the website might increase revenue,

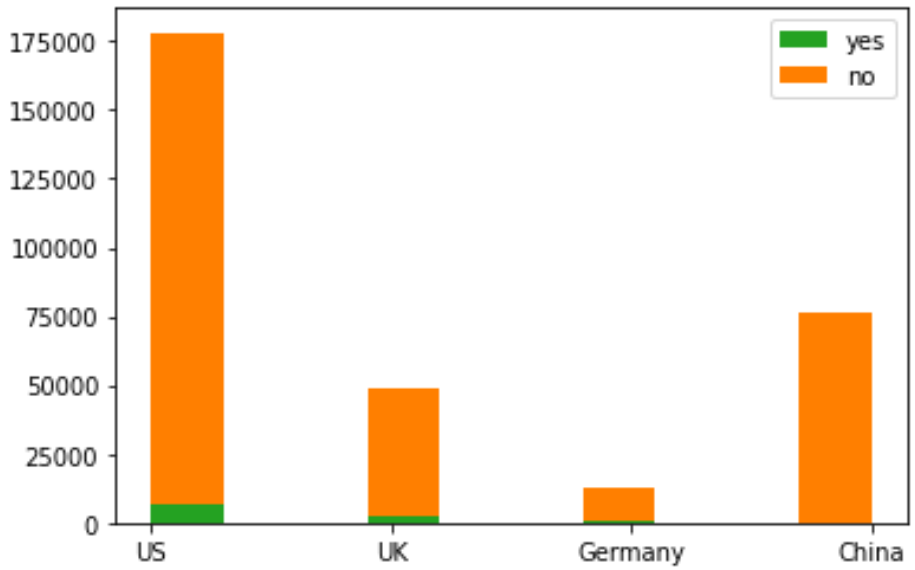
- age seems to be a driving factor as well, with users above 40 years old barely converting:



Age per user conversion (yes or no).

the reason why older people are not converting should be investigated. The advertisement campaign could be retargeted to be more inclusive,

- the conversion rate from users in China is negligible. Also the US market underperforms, yielding small conversion rates if compared to the European counterparts. The local marketing should therefore be reassessed:



Country per user conversion (yes or no).

- there does not seem to be a significant correlation between marketing channels and conversion rate.