Final Project

Website Conversion Rate - Classification

Konstantin Schätz

06.10.23

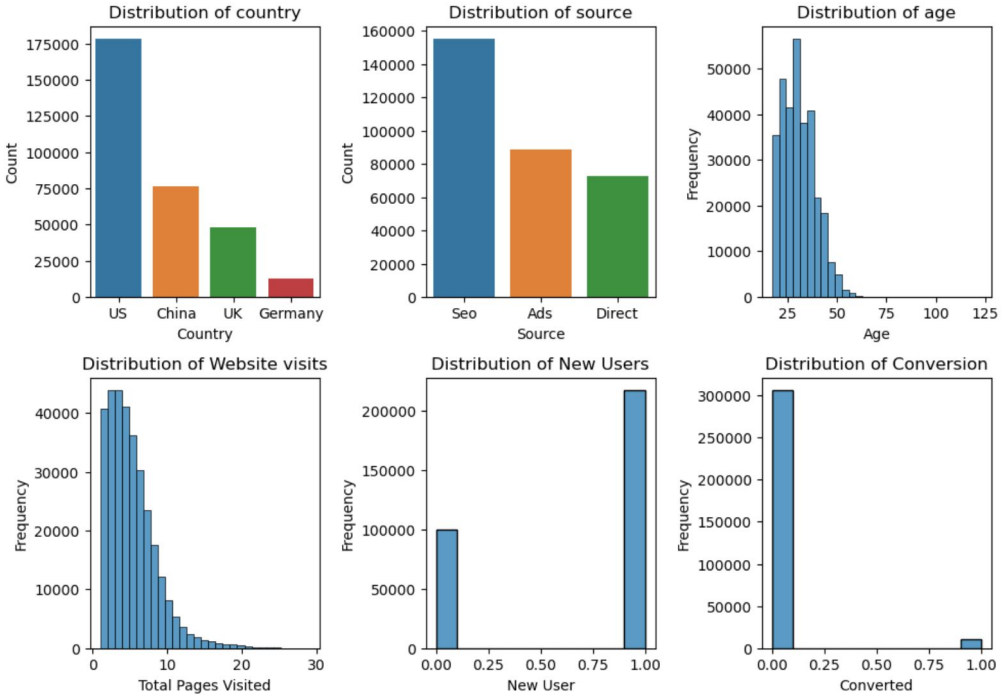
Journey/Agenda

	country	age	new_user	source	total_pages_visited	converted
0	UK	25	1	Ads	1	0
1	US	23	1	Seo	5	0
2	US	28	1	Seo	4	0
3	China	39	1	Seo	5	0
4	US	30	1	Seo	6	0

316.200 x 6

1. EDA in Python
2. Data Cleaning (Feature Engineering)
3. Classification imbalanced data
    a. Decision Tree
    b. Logistic Regression
    c. KNN
4. Classification with balanced data
    a. Smote-Upsampling
    b. Downsampling
5. BI in Tableau

Exploratory Data Analysis (EDA)



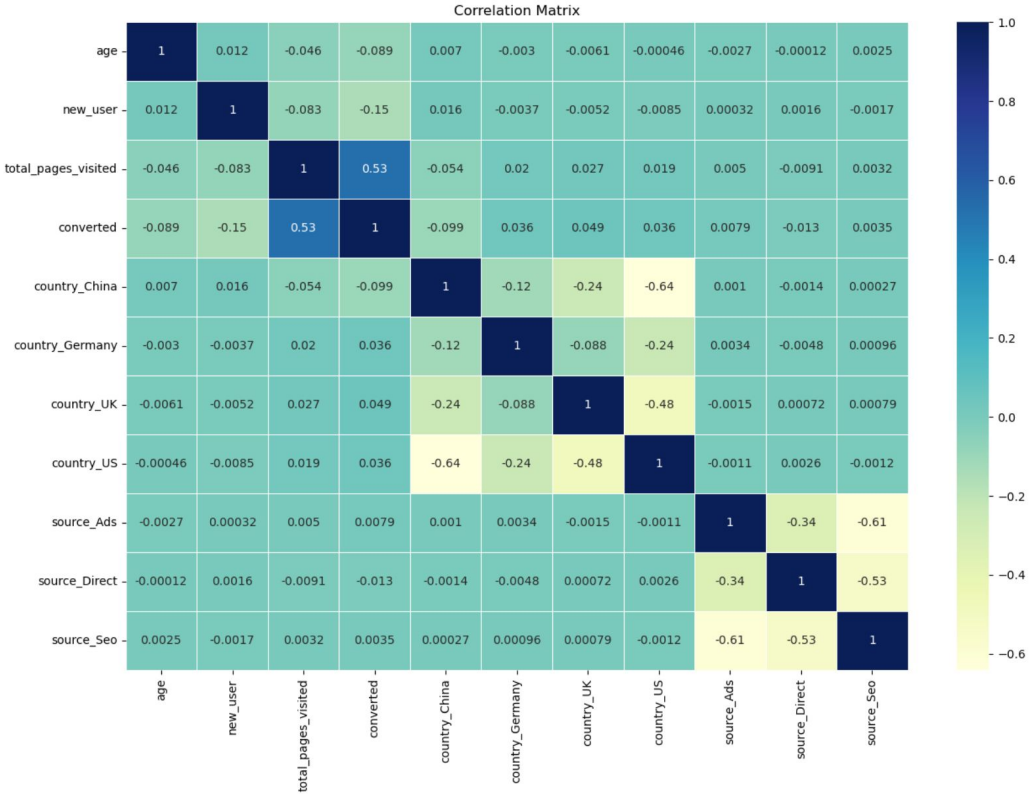
EDA

- 2 categorical variables, 2 binary variables
  - feature engineering necessary
  - ***New_Users*** and ***Conversion*** seem quite imbalanced
- distributions seem skewed to the right

Insights

- website visitors come from the **US** and **China** the most,
- the most successful ***marketing channel*** is Search Engine Optimization (SEO)
- **age** is between 17 and 123, the most frequent age is 30 which is equal to the median and close to the **mean age of all visitors (30.5 years)**
- The mode of ***total pages visited*** is 2, on average 4 pages are visited
- twice as many new shoppers than recurring
- the conversion rate is at 3% (**target variable**)

Correlation Matrix



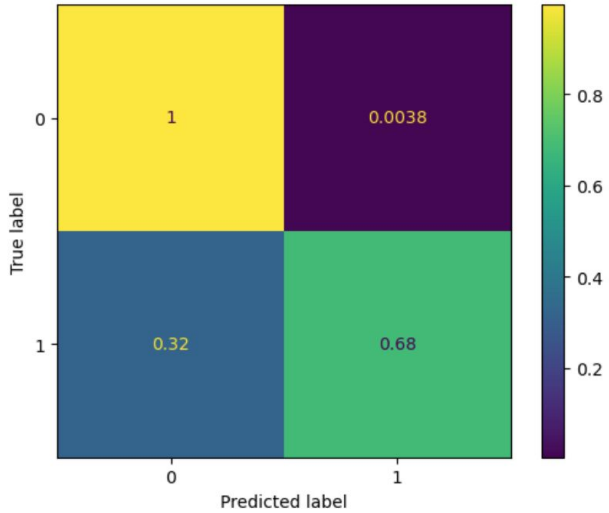
Correlation

- there is no correlation higher than 90%
  - all features can be kept
  - no auto-correlation/multi-collinearity
- the strongest relation seems to be between **total_pages_visited** and **converted**
  - moderate impact with 0.53
- ceteris paribus, the likelihood of conversion increases by 0.53 units for every page that is visited
- other features seem to have barely no impact on the conversion rate

Classification with imbalanced data

	age	new_user	total_pages_visited	country_China	country_Germany	country_UK	country_US	source_Ads	source_Direct	source_Seo
0	25	1	1	0.0	0.0	1.0	0.0	1.0	0.0	0.0
1	23	1	5	0.0	0.0	0.0	1.0	0.0	0.0	1.0
2	28	1	4	0.0	0.0	0.0	1.0	0.0	0.0	1.0
3	39	1	5	1.0	0.0	0.0	0.0	0.0	0.0	1.0
4	30	1	6	0.0	0.0	0.0	1.0	0.0	0.0	1.0

	Error Metric	Accuracy
0	Decision Tree	0.985800
1	Logistic Regresssion	0.985948
2	KNN	0.985178

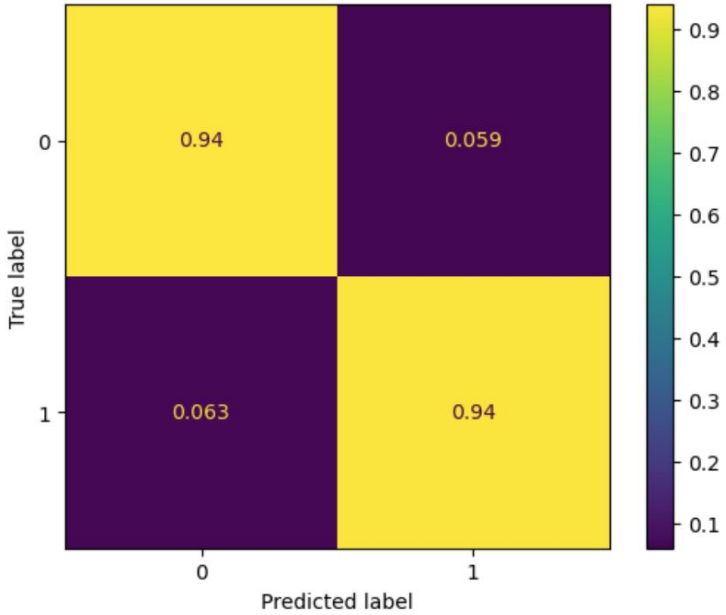


Classification Results

- Decision Tree at depth 7, KNN with 13 neighbours
- the logistic regression seems to achieve the most accurate results

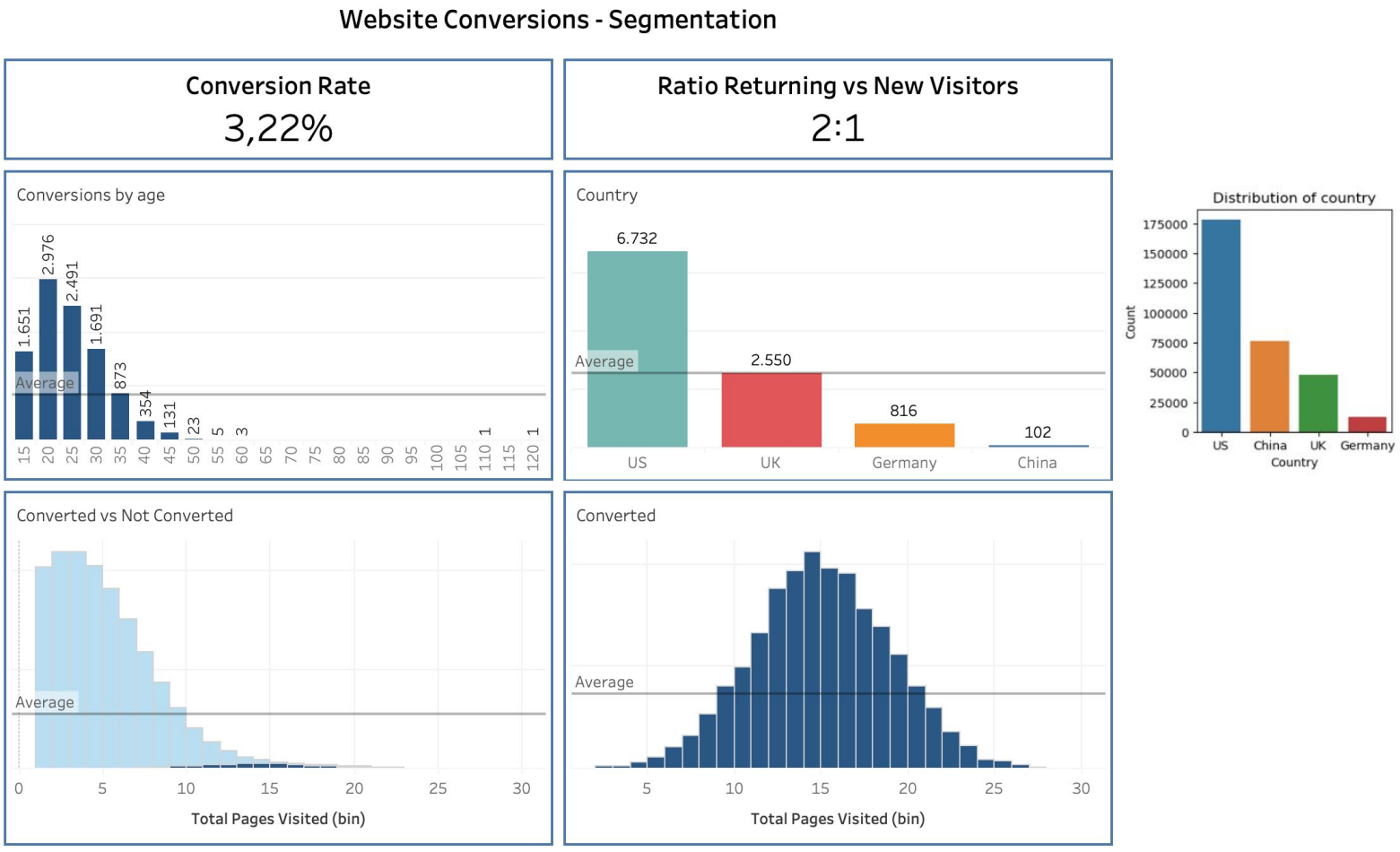
Balanced Data - Logistic Regression

	Logistic Regression	Accuracy	Precision	Recall
0	Smote	0.748203	0.665307	0.996823
1	Downsampled	0.938725	0.940328	0.936949

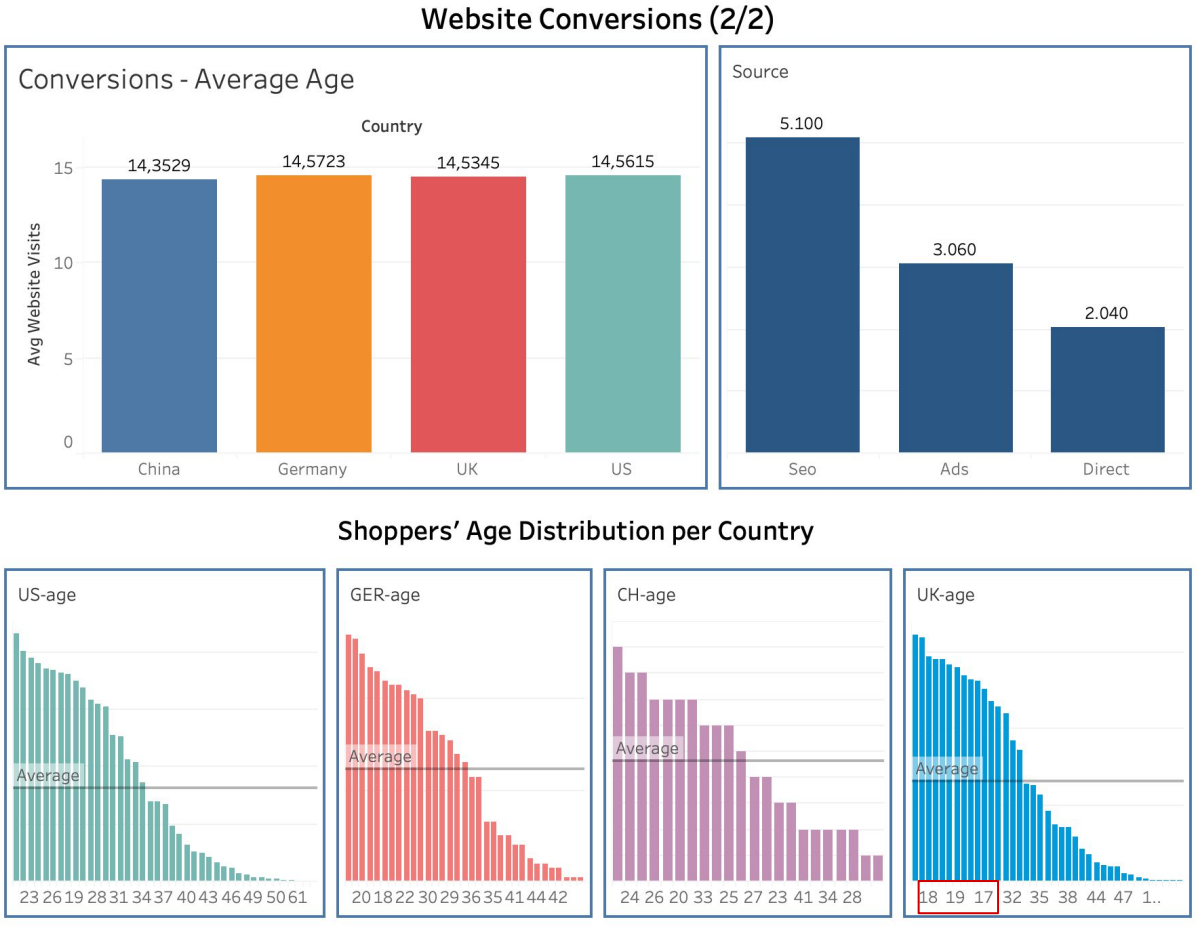


- Downsample delivers the better model
- our classifier performs better than the first one
- 94% of shoppers that did not convert, and 94% of shoppers that did convert were correctly classified
  - before: 100% / 68%
  - occurs for highly imbalanced datasets

Conversions



Conversions - Age & Source



Final Insights

- Less than 1% of shoppers in China are converting
- those under 40 convert above average
- Recurring shoppers convert more than new shoppers
- Most shoppers enter the site through a search engine result, independent of conversion
- Shoppers that convert visit 10 times more pages that users that did not convert; more time on the site leads to conversion

THANK YOU!