CONVERSION RATE CHALLENGE

Predictive analysis of structured data by artificial intelligence

KINN Linda – Jedha Bootcamp - 2022

SUMMARY

- Use Case presentation
- Database exploratory
- User’s profil analysis
- Convertion analysis
- Models application : performances on F1_Score
- Conclusion
- To go further

Use Case presentation

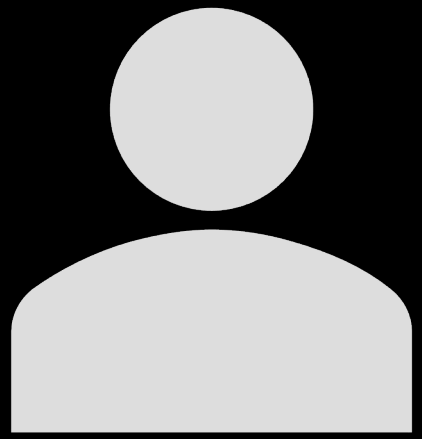
- Kaggle Challenge
- ***Datascienceweekly.org*** famous newsletter for Data Scientist made by Data Scientist
- Open-source Dataset containing some data about the traffic on their website
- Goal : Analyze parameters of the model to hightlight features that are important to explain the behaviour of the users and :

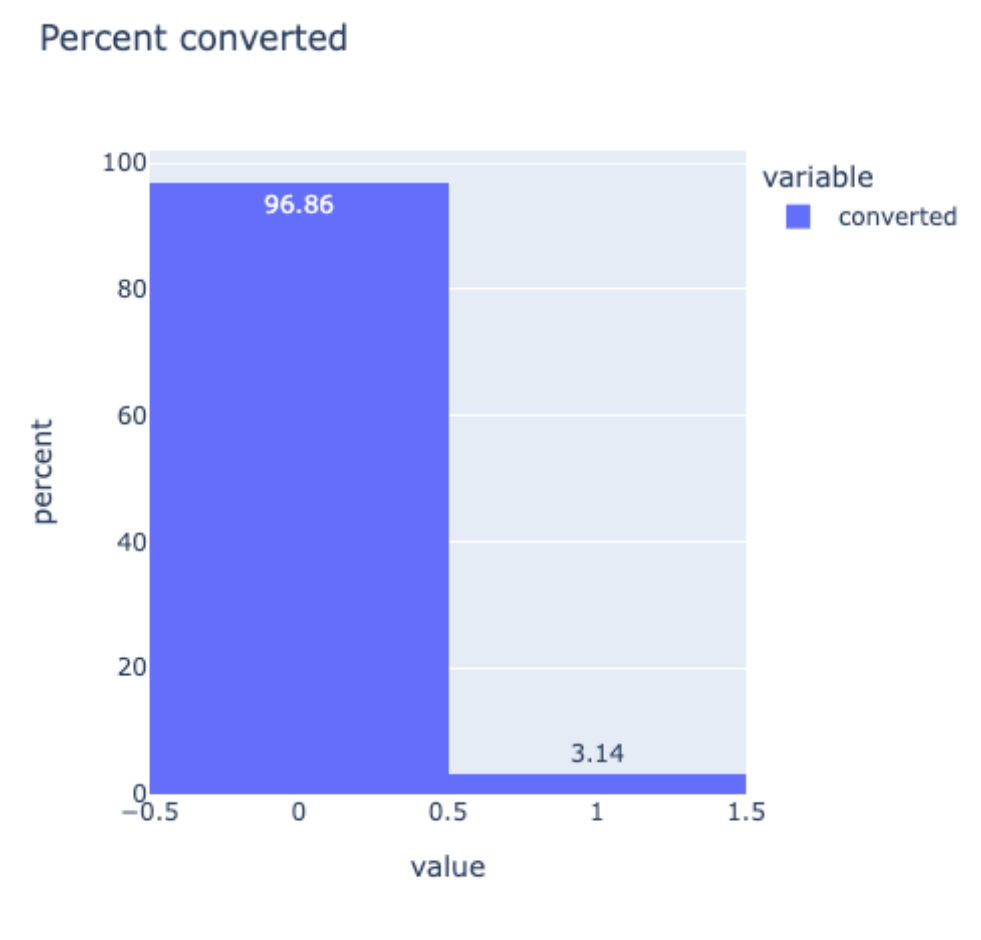
« Discover a new lever for action to improve the newsletter's conversion rate »

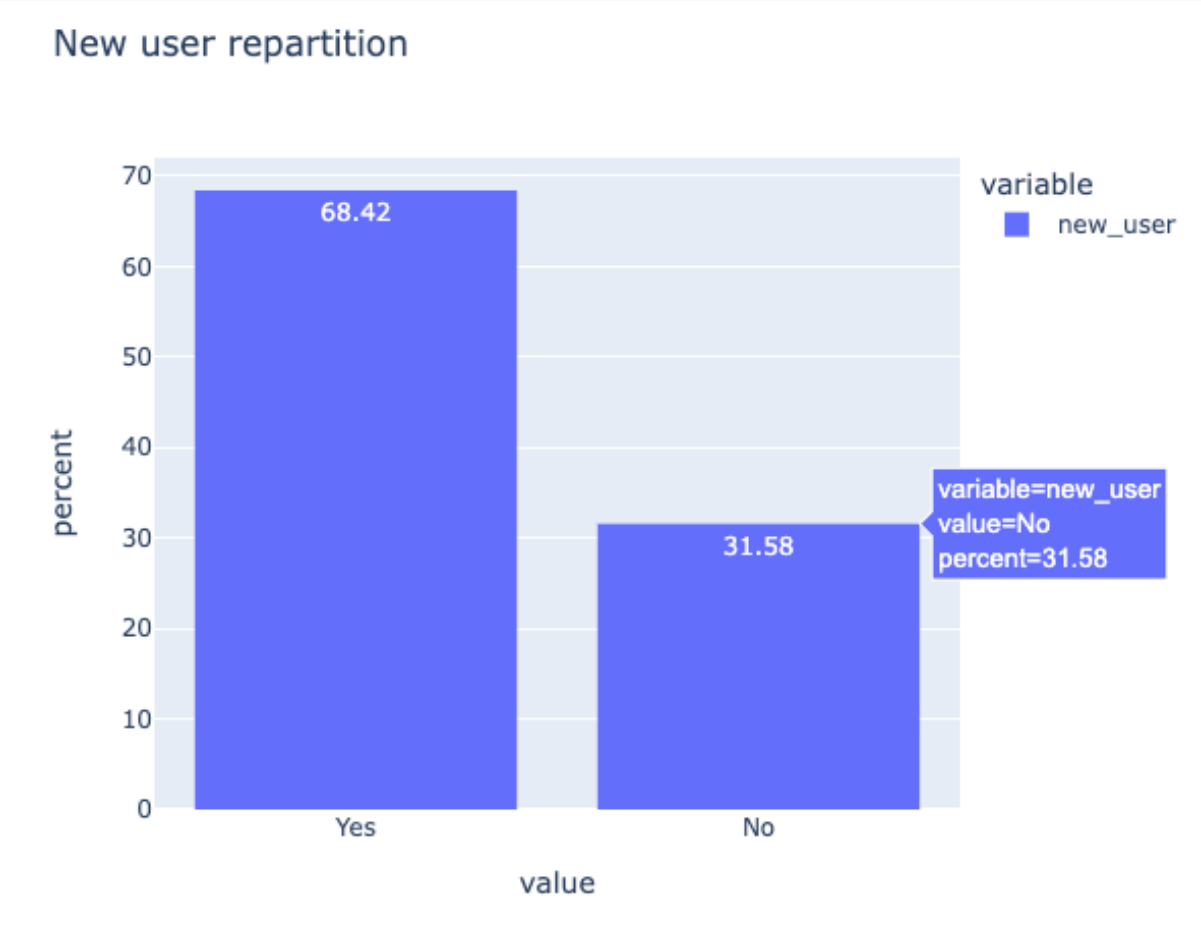
Database exploratory

- Target : ‘converted’
- *The dataset trains_csv :*
  - *Number of observations training dataset : 284 580*
  - *Number of features : 6*
  - *Features names : [‘country', ‘age', ‘new_user', ‘source’, ‘total_pages_visited', ‘converted’]*
- The dataset test_csv :
  - *Number of observations test dataset : 31 620*
  - *Number of features : 5*
  - *Features_names : [‘country', ‘age', ‘new_user', ‘source’, ‘total_pages_visited’]*

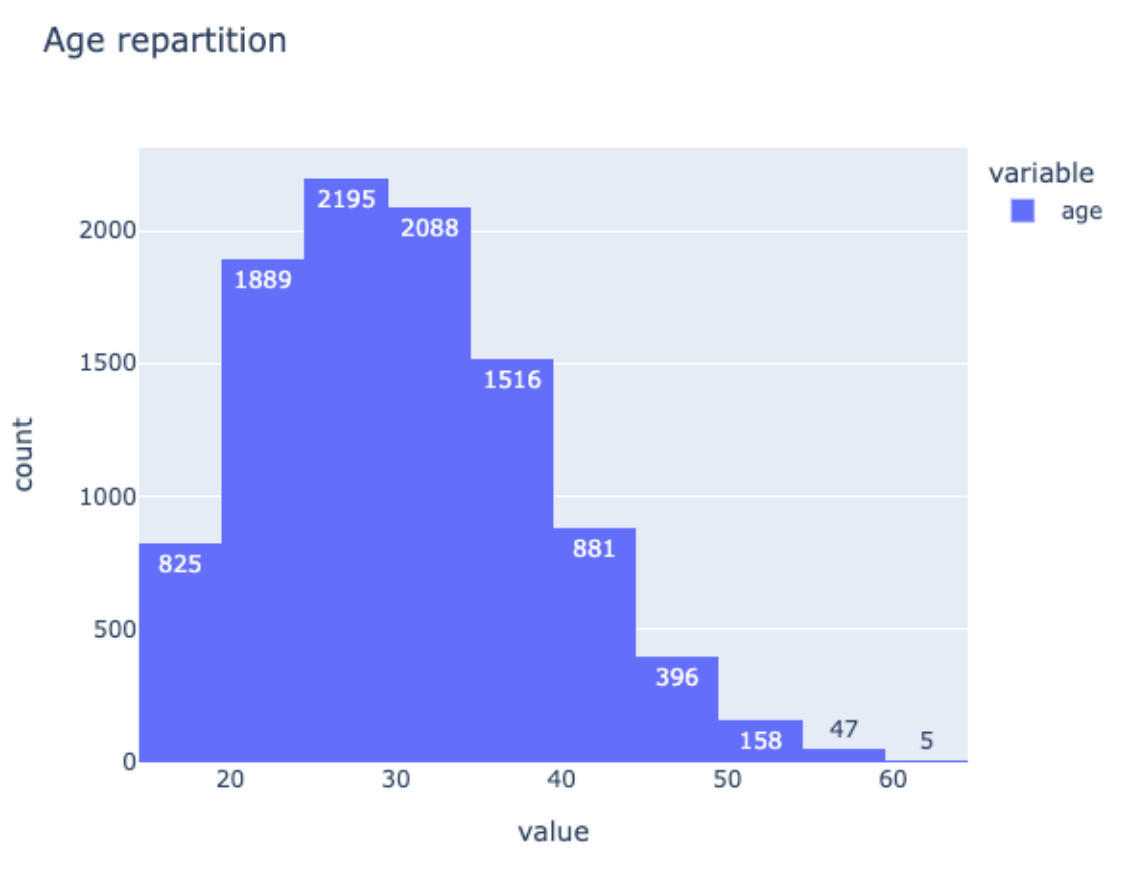
USERS PROFILE ANALYSIS

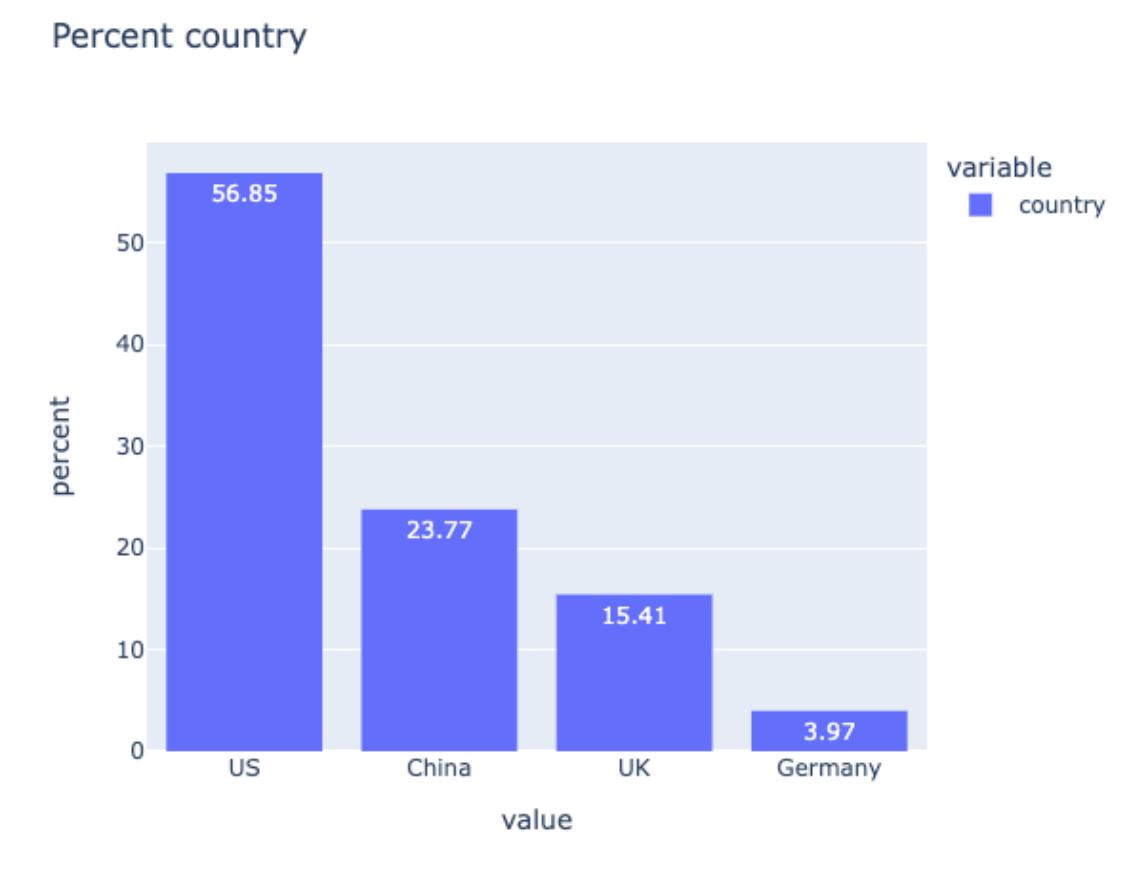




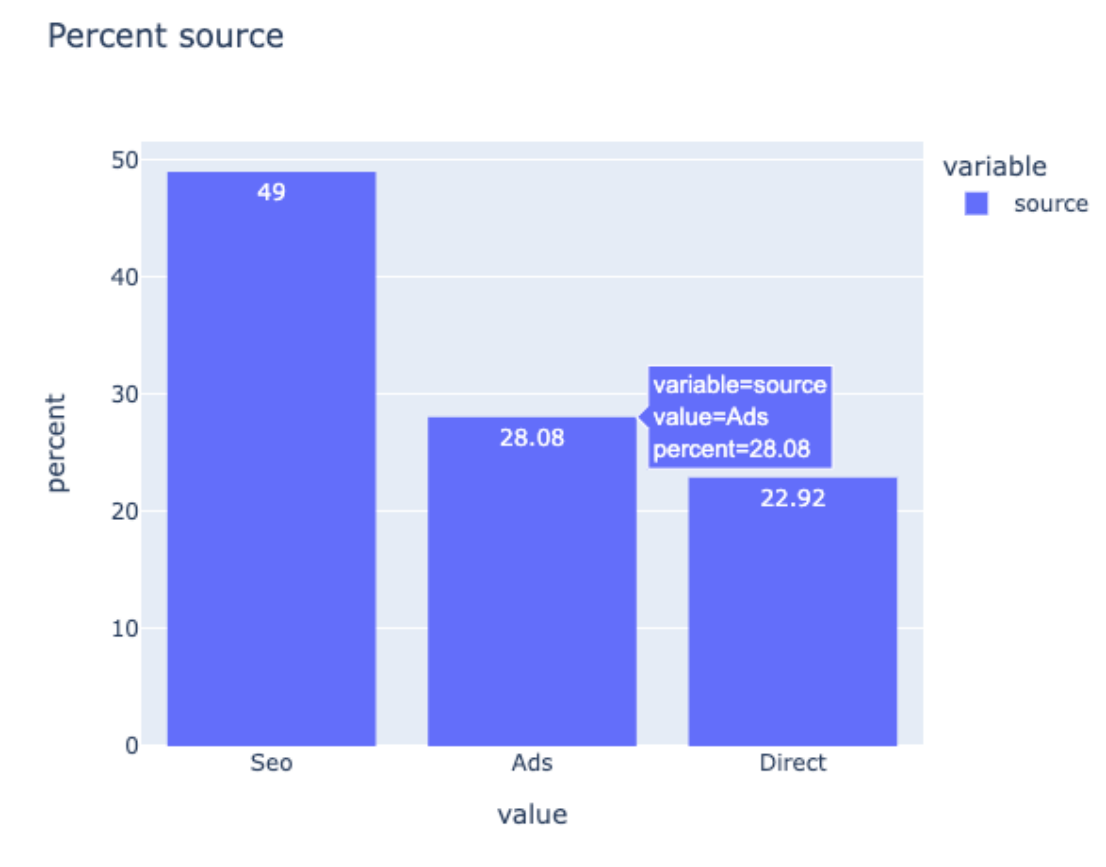


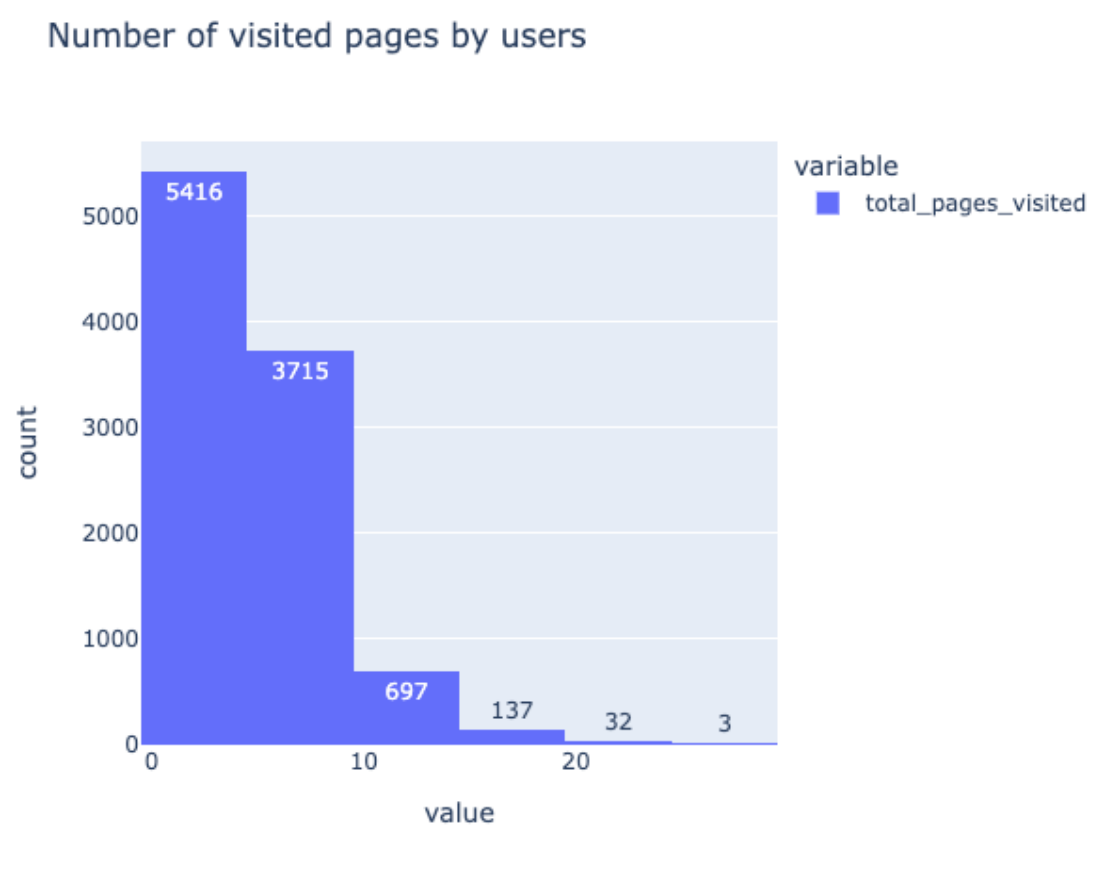
- In the dataset, we have a really few people converted to the newsletters
- There is 2/3 of new users





- User mostly have between 20 to 37 yo
- Visiter are mostly from US follow by China then UK and Germany



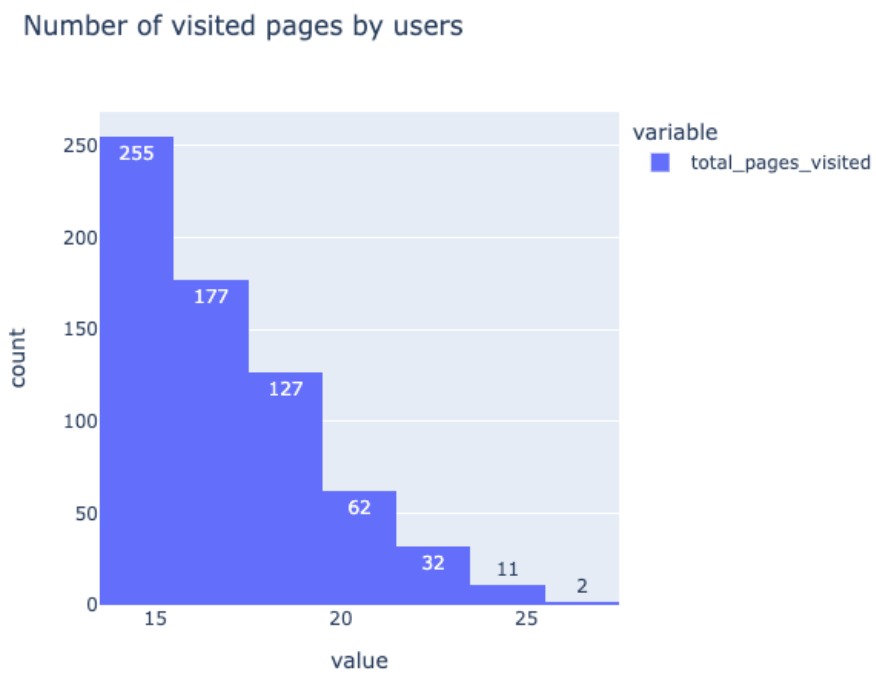


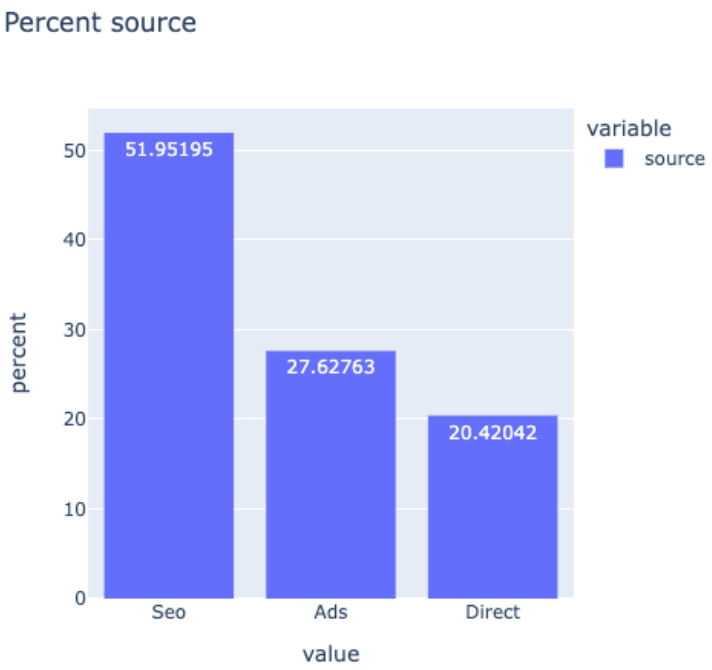
- 50% of visitors get the source from Seo then Ads and Direct
- Users visit at most 10 pages

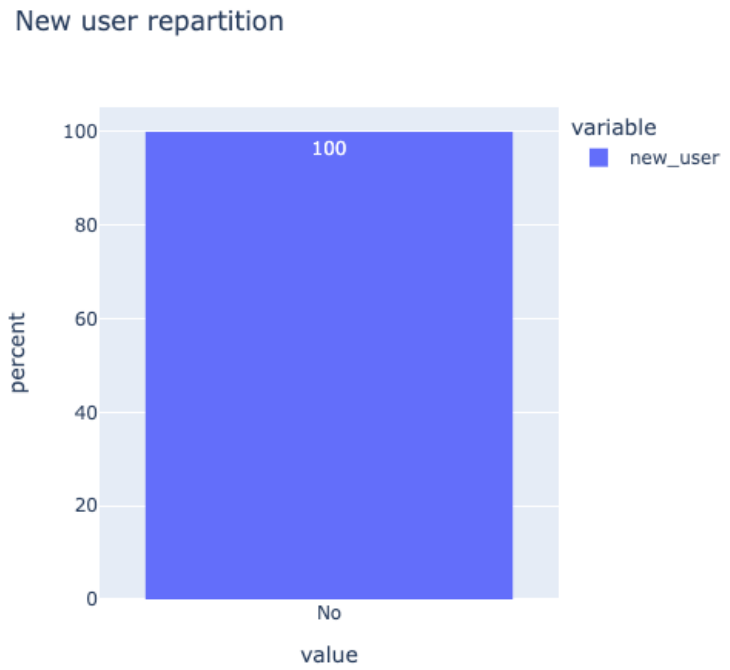
Resume User’s profile

- Compare to the large dataset we have with 284 580 peoples, there is a really really small users with the converted status (3%)
- 70 % are new users in the dataset
- Users are between 20 to 37 yo
- Most of them came from US follow up by China then UK and Germany
- Seo is the source who bring 50% of users
- Users roughtly visited not more than 10 pages

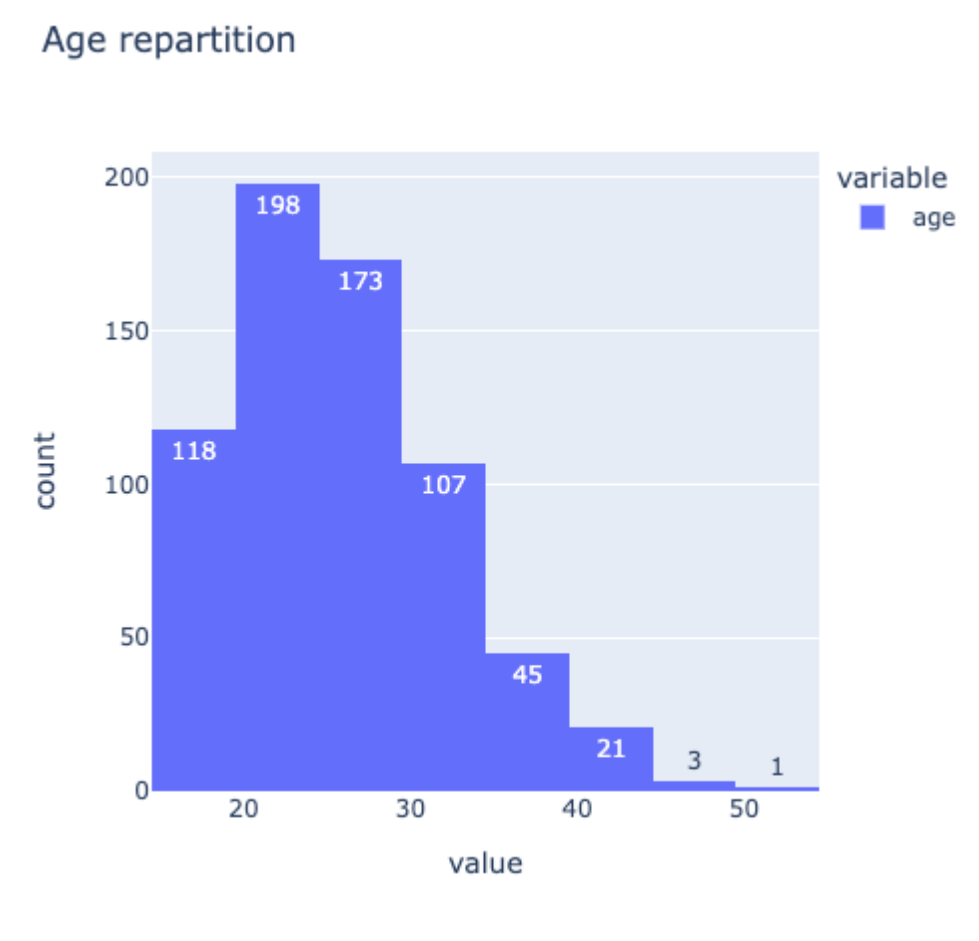
CONVERTION ANALYSIS

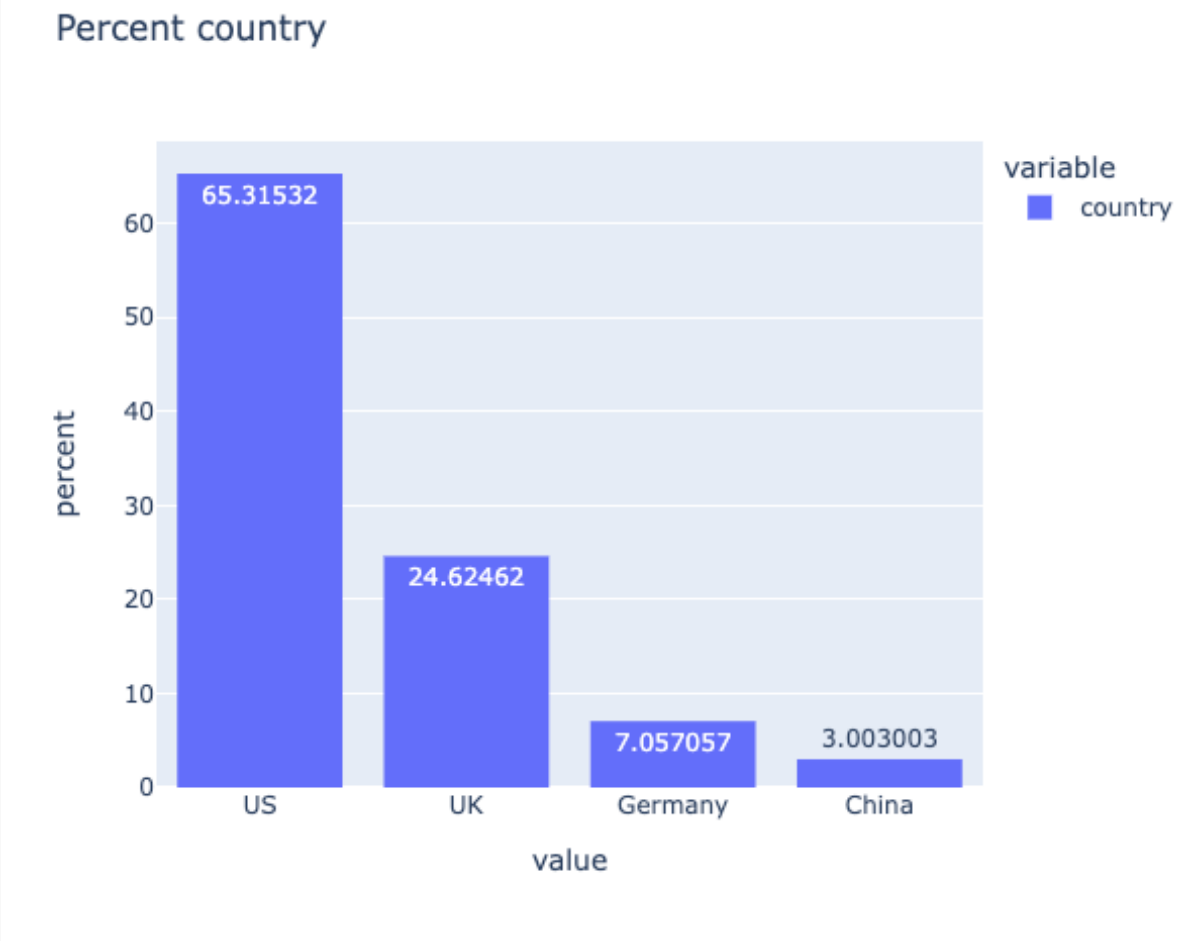






- Converted users have visited at most 20 pages
- Half of them get the source from Seo
- There is no new user people. That’s mean that a their first visited, people don’t subcribe on the newletters





- Converted people have betweem 15 to 35 yo
- More than the half are from US. China don’t subcribe that much

Recommandations

- Visitors became a converted users if they have already visited roughtly 20 pages of the website
  - *That's mean they have to dig into pages to be convicted by newsletters*
- The most importance source than bring users is SEO. Ads and Direct both bring 1/4 that is not too bad
  - *Could improve Ads and Direct source*
- New user don't subcribe directly on their first visiting
  - *Should put on the first page something to increase this rate*
- Converted users are from 15 to 35 yo
  - *Should aim this age category*
- Most of the users came from US and subcribe. That's is logical but for China, which represent 20% users get only 3% of subcription
  - *Should dig into these category of people to understand why the conversion rate is so bad !*

This is the pre-analysis about converted people. Now we want to see with Data Science how can we improve the converted people rate…

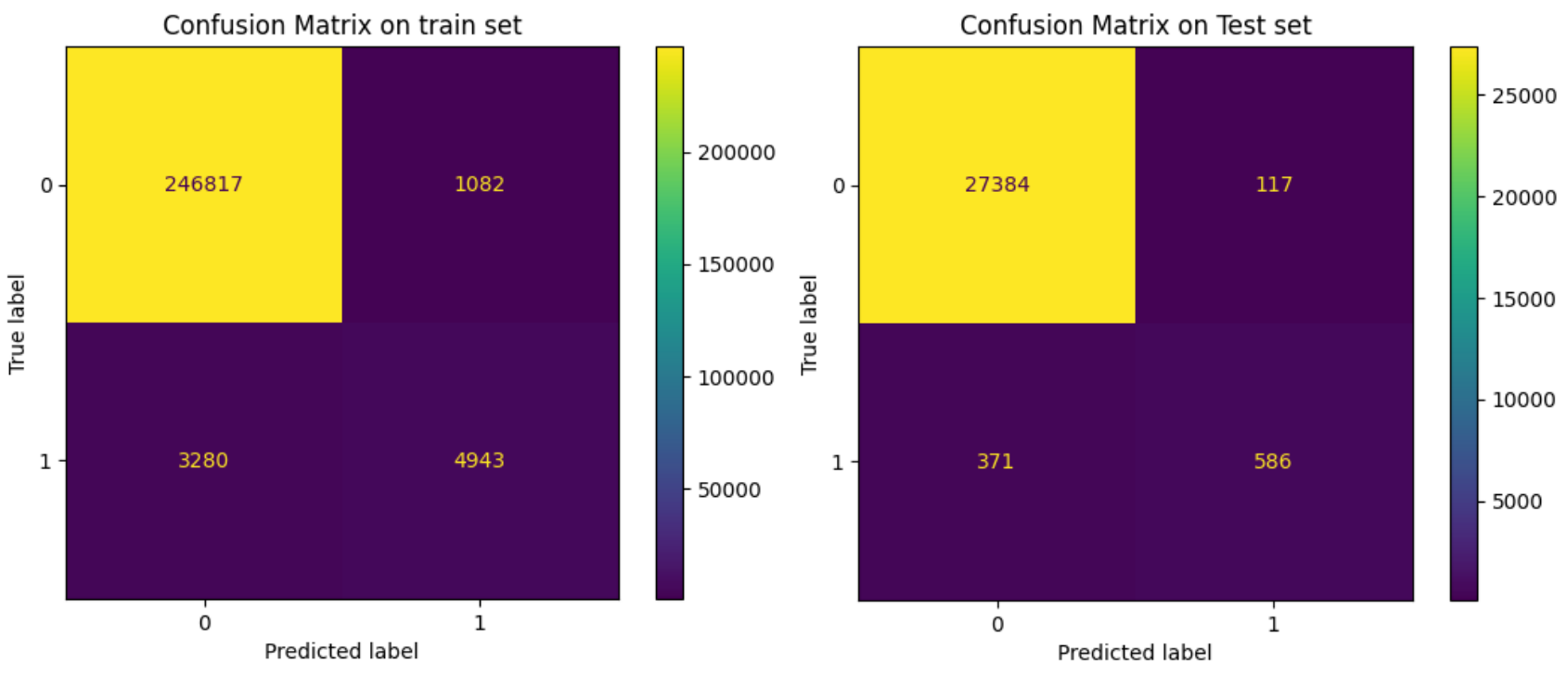
MODELS APPLICATION

BASELINE MODEL

- The dataset : data_train.csv
  - *Number of observations : 284 580*
  - *Number of features : 5*
  - *Features names : [‘country’, ‘age’, new_user’,’source’,’total_pages_visited’]*
- The training dataset for baseline model application :
  - *Number of observations : sample of 10 000*
  - *Number of feature : 1*
  - *Feature name : [‘total_pages_visited’]*
  - *Model used : Univariate Logistic Regression*

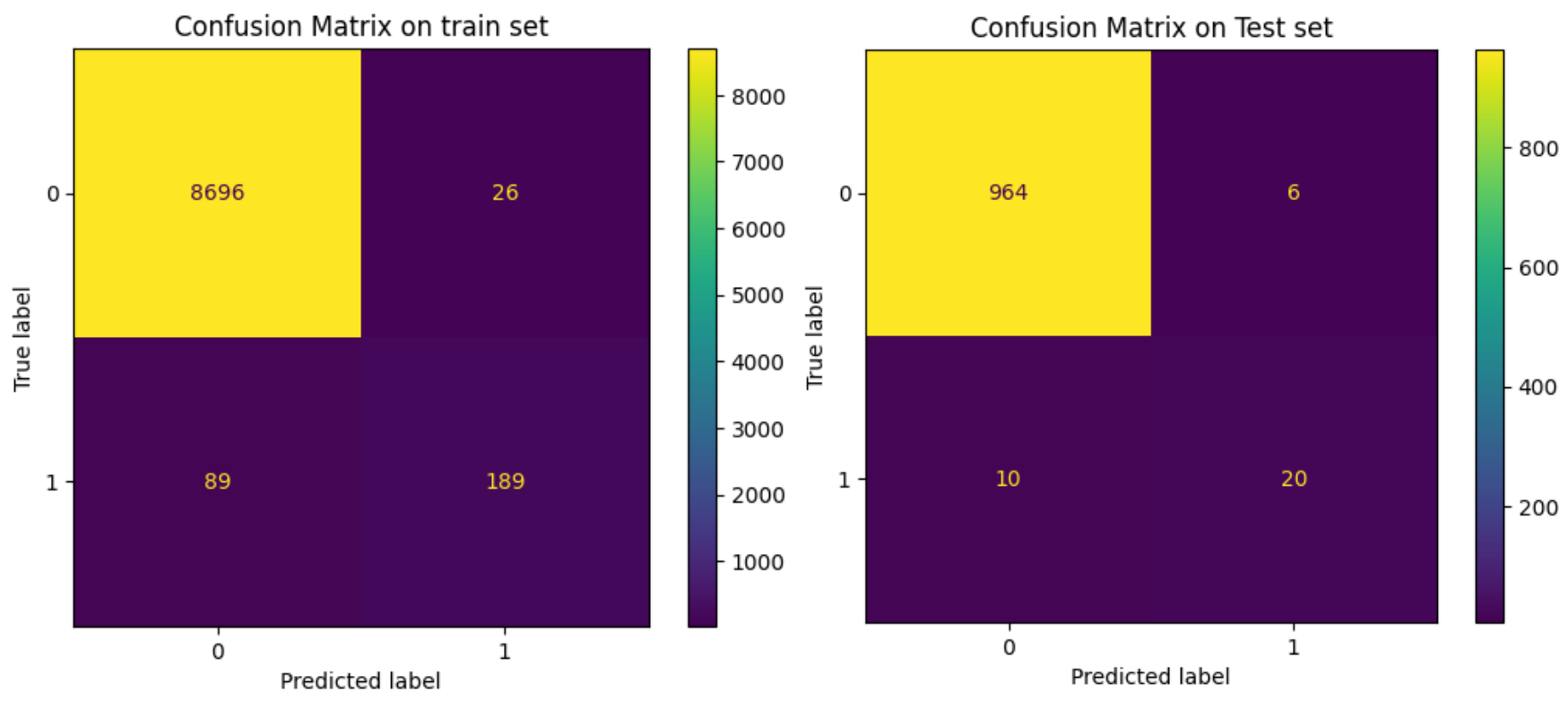
UNIVARIATE LOGISTIC REGRESSION

BASELINE MODEL



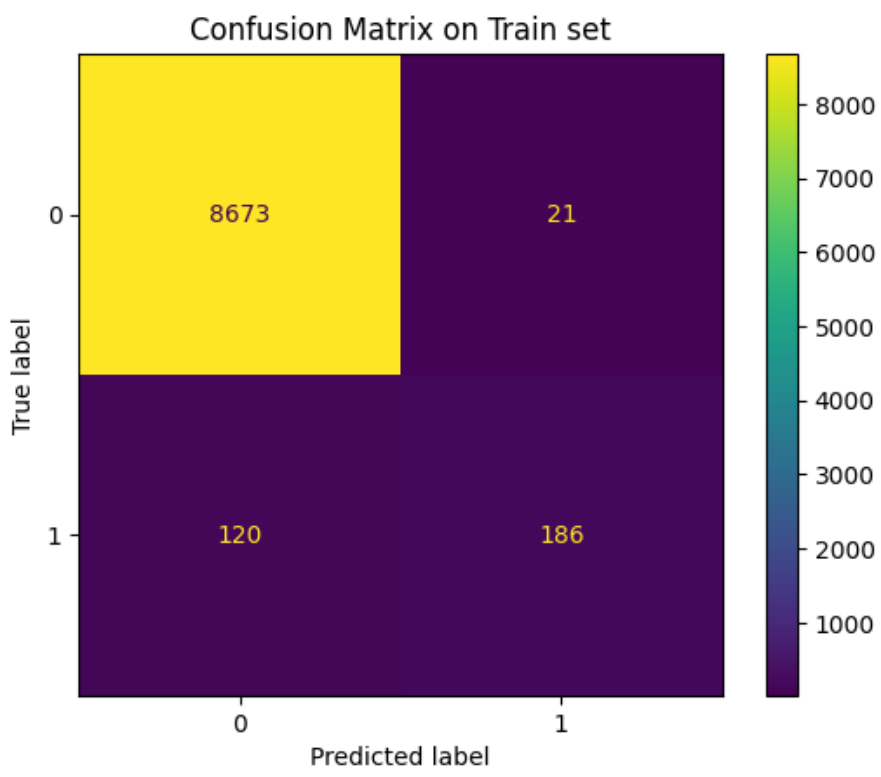
	F1-SCORE
Training set	0,69
Test set	0,71

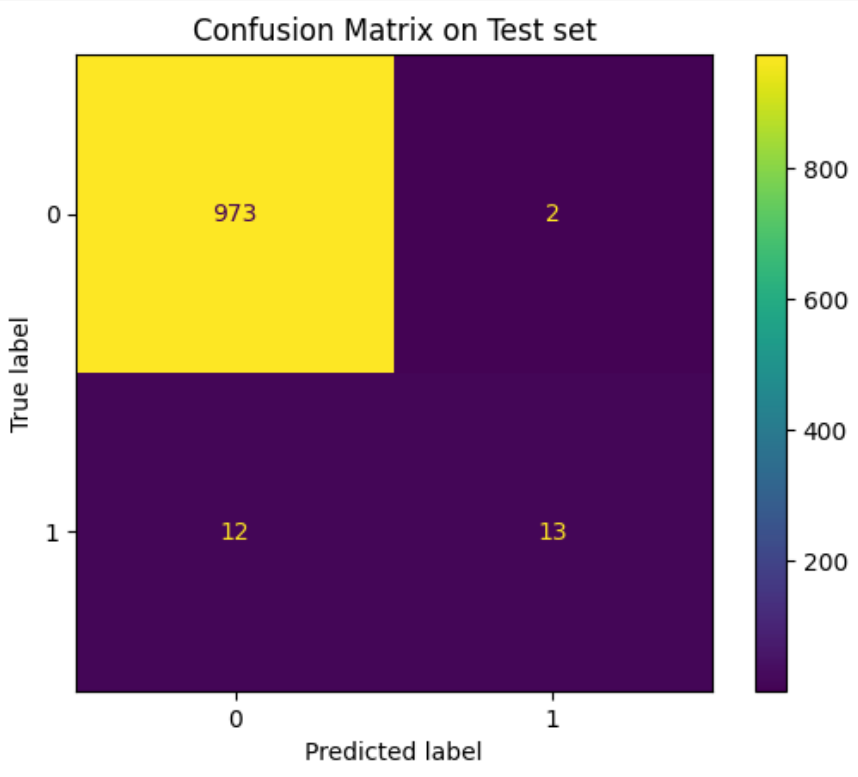
MULTIVARIATE LOGISTIC REGRESSION



	F1-SCORE
Training set	0,77
Test set	0,71

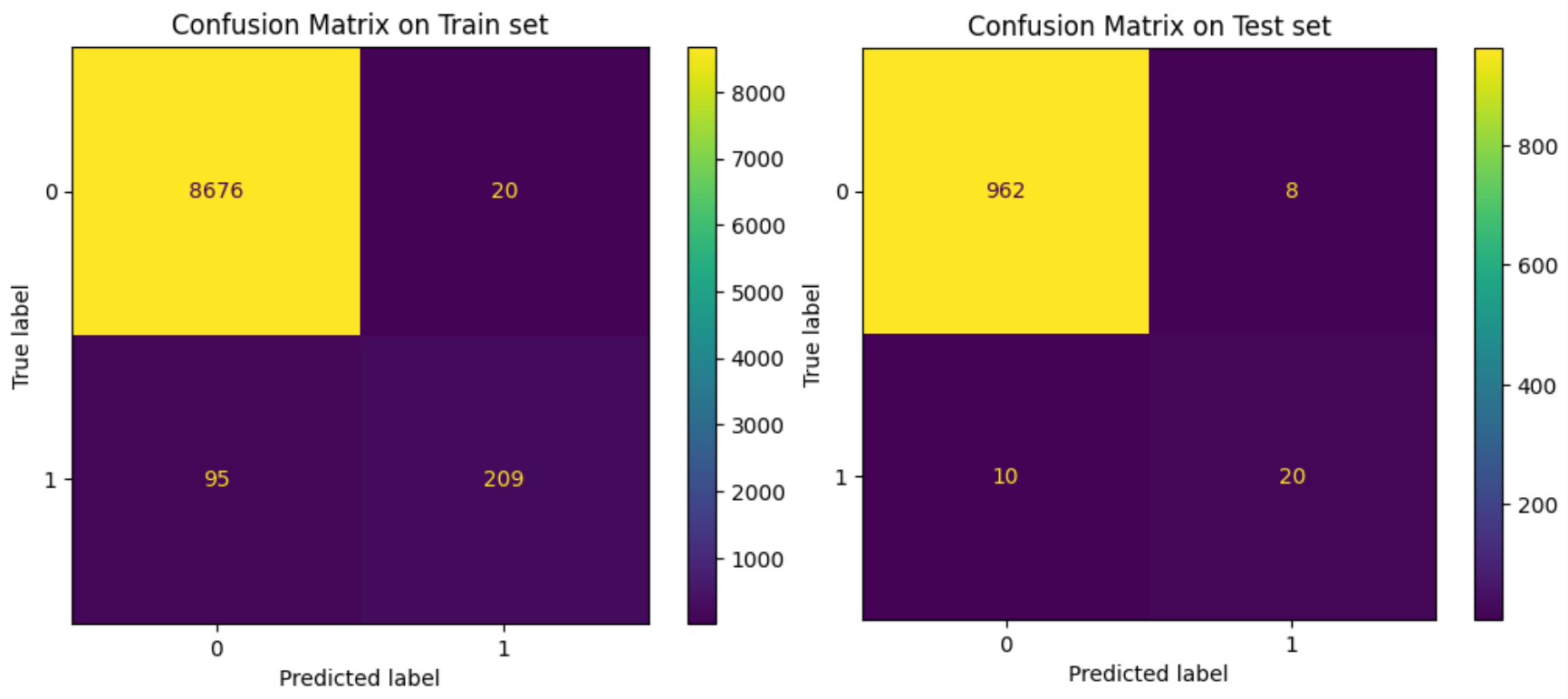
DECISION TREE WITH CV = 3





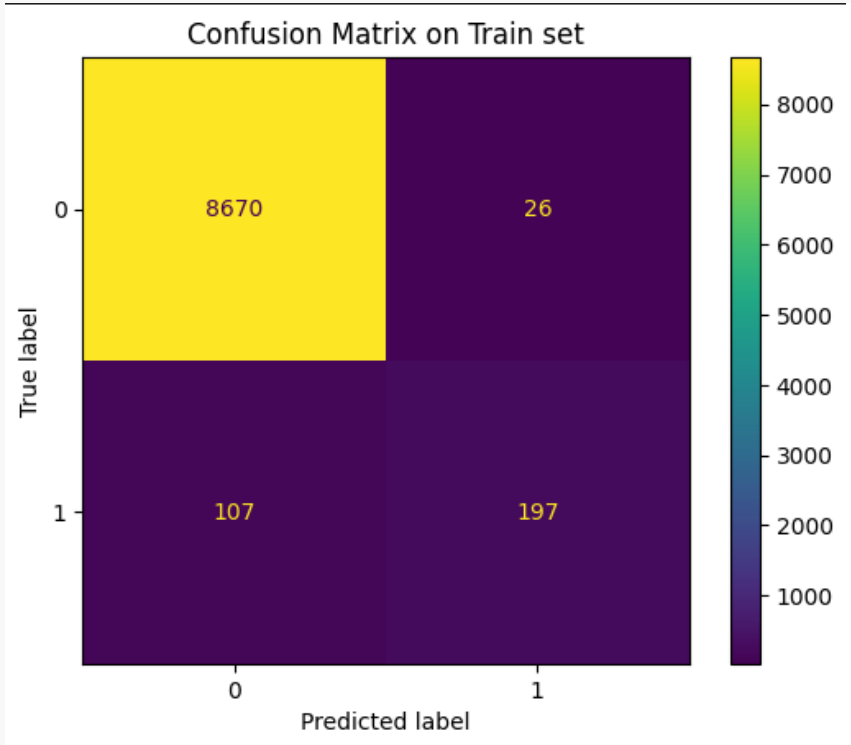
	F1-SCORE
Training set	0,73
Test set	0,65

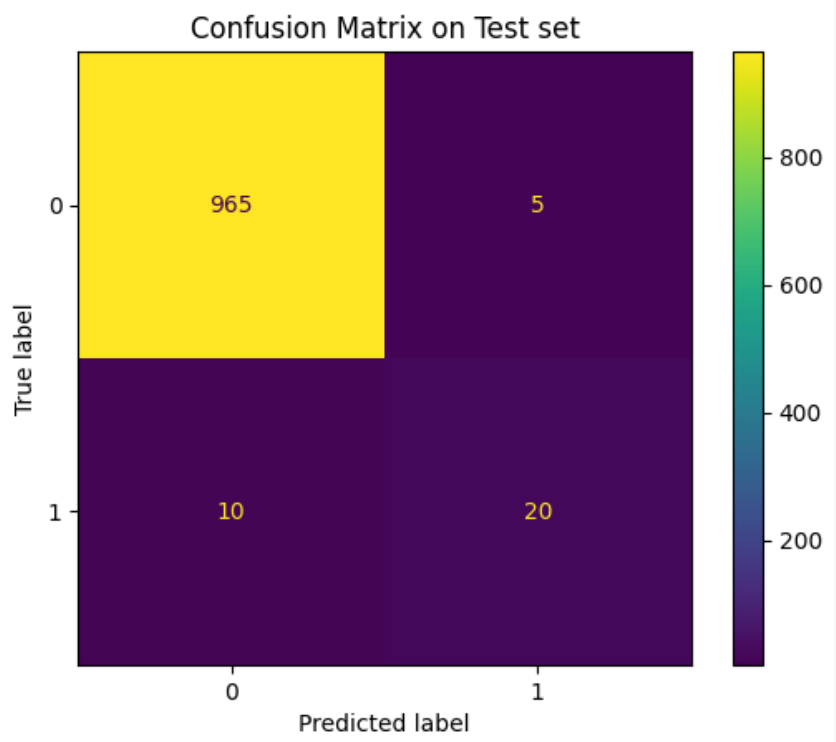
DECISION TREE WITH CV = 10



	F1-SCORE
Training set	0,75
Test set	0,73

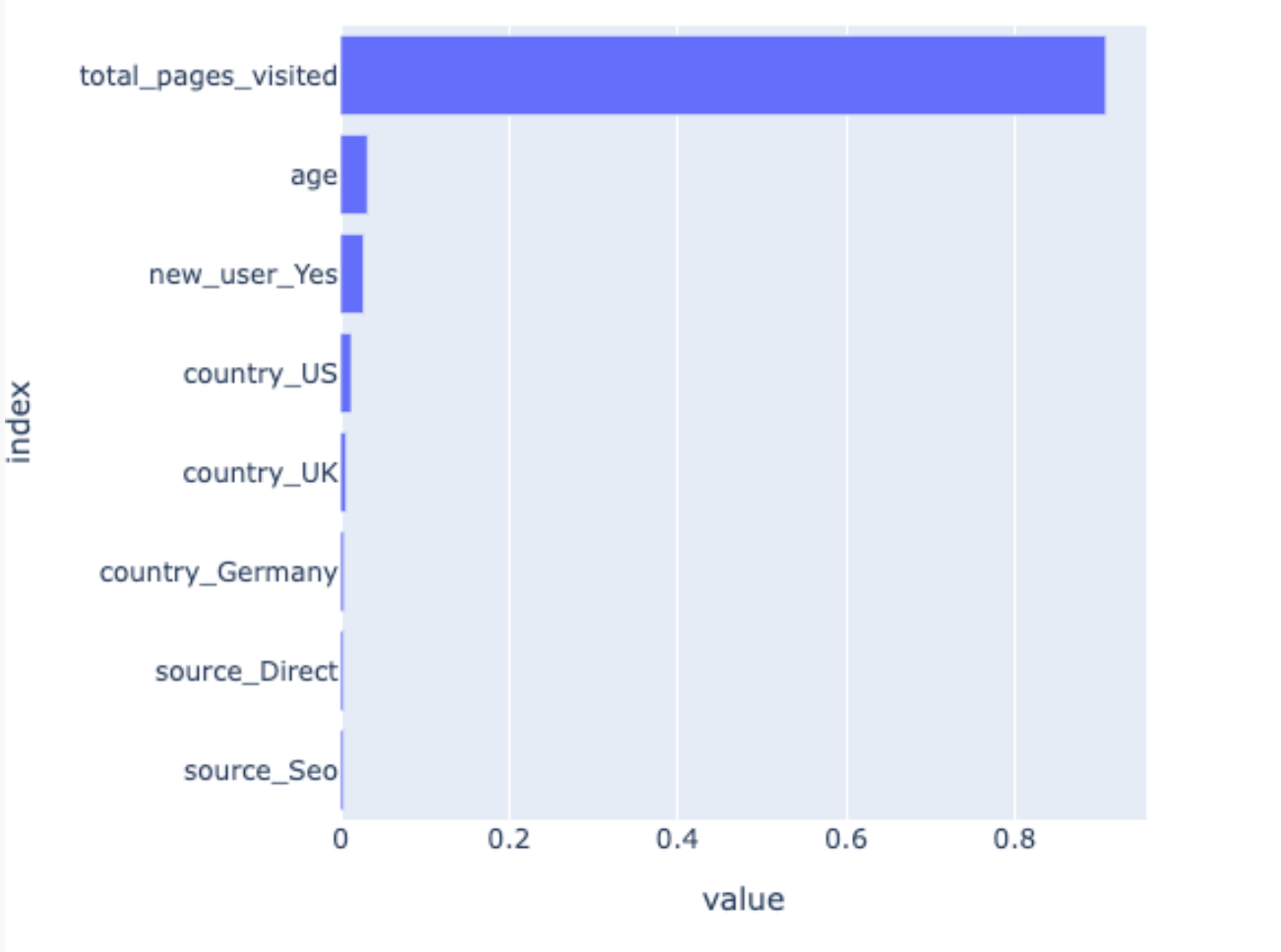
RANDOM FOREST





	F1-SCORE
Training set	0,77
Test set	0,76

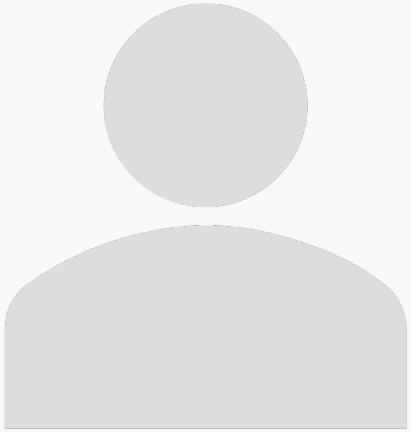
FEATURES IMPORTANCE



F1-SCORE PALMARES

Models	Set	F1-SCORE
Univariate Logistic Regression	Train	0,69
	Test	0,71
Multivariate Logistic Regression	Train	0,77
	Test	0,71
Decision Tree with CV = 3	Train	0,73
	Test	0,65
Decision Tree with CV = 10	Train	0,75
	Test	0,73
Random Forest	Train	0,77
	Test	0,76

CONCLUSION



Lever for action to improve the rate

- We can with Data Analysis follow the recommandation from conversion analysis

- We can with DataScience :
  - *Add pertinent features*
  - *Delete the not important feature from the feature importance algorythm*
  - *Use a boosting algorythm*
  - *Play with hyperparameters*