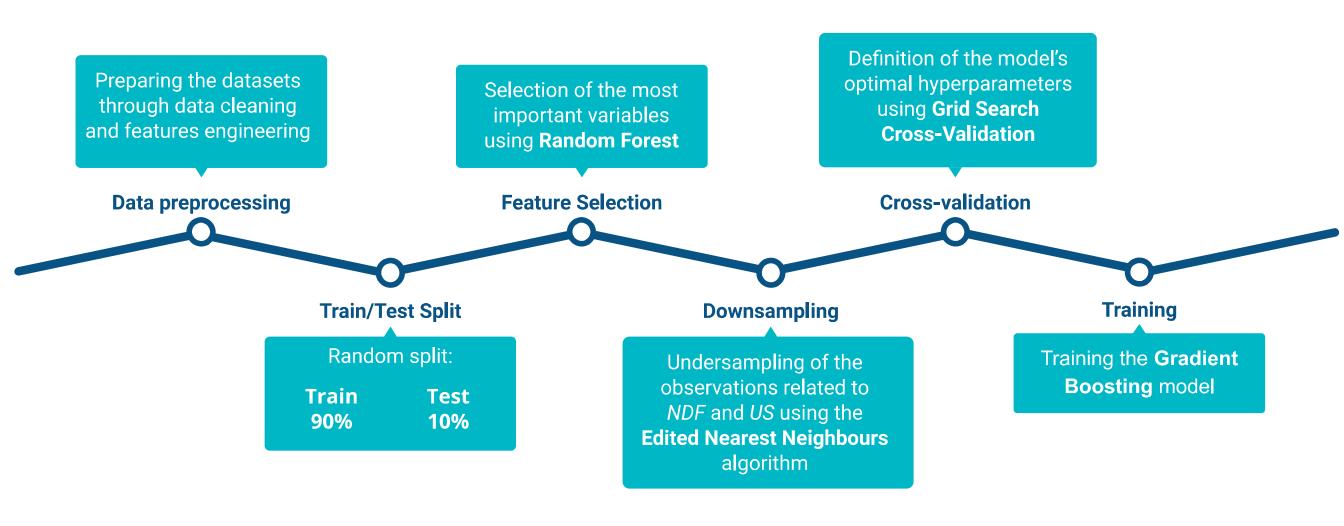
Datasets creation

Variables extracted **Dataset obtained** Dataset "users_no_sessions" Categorical - gender, language, first device, ... It contains all the users available in the file "train_users" with all the **Numerical** - age and signup_flow columns obtainable from the **Files** same file only (no sessions Dates - date account creation, date first information included) active and time gap between them train_users.csv Dataset "users_sessions" sessions.csv Quantitative - total actions per user, total action type per user, total devices, ... It contains all the users whose online sessions have **Time** - total time per device, total time per been recorded, with variables action type, average time spent on a device, from both the source files.

Approaching the problem



Results

Test set confusion matrix



The model is right 66 times out of 100 when it predicts an NDF



The model is able to intercept 89 TRUE NDF observations out of 100

	Predicted											
								NDF			US	other
	0	0	0	0	0	0	0	34	0	0	15	0
ב	0	0	0	0	1	0	0	80	0	0	44	2
	0	0	0	0	0	0	0	67	0	0	41	1
	0	0	0	0	1	0	0	145	0	0	89	0
	0	0	0	0	2	0	0	332	0	0	168	8
	0	0	0	0	2	0	0	168	0	0	75	2
	0	0	0	0	2	0	2	185	0	0	79	1
NDF	0	0	0	0	7	0	1	11062	0	0	1324	40
other US	0	0	0	0	0	0	0	50	0	0	32	1
	0	0	0	0	0	0	0	12	0	0	7	1
	1	0	0	0	7	0	0	3871	0	0	2302	55
	0	0	0	1	2	0	0	667	0	0	336	21

Results

User	ser P1 P2		P3	P4	True
q9p###	US	other	NDF	GB	other
9bz###	NDF	US	other	GB	US
kux###	NDF	US	other	AU	US
qqn###	NDF	US	other	FR	NDF
nw9###	NDF	US	other	FR	NDF
oop###	NDF	US	other	FR	IT
a62###	NDF	other	US	FR	NDF
611###	NDF	other	US	FR	FR
qbs###	NDF	US	FR	other	US
f2k###	NDF	US	other	FR	US

Ranking

The table contains:



User - user name (partially hidden)



P1_P4 - the first 4 predictions per user
ordered for descending probability



True - the true destination of the user

In the sample extracted, the model gives a good ranking to the true destination of **9 users out of 10**. In all the test population, the ratio is **9.4 out of 10**.

Therefore, the model can give an accurate list of (few) probable destinations.

Conclusion

Summary

Pediction	Ranking	Recommendation
The model does not seem to be able to distinguish between the most frequent observations (<i>NDF</i> , <i>US</i> and <i>other</i>) and all the other destinations	Anyway, when considering the ranking for probabilities, we observe a very accurate list of potential destinations.	If the purpose of this analysis is to guide a targeted marketing campaign, we highly recommend focussing efforts and resources on the first 4 countries predicted per user.

Possible improvements

It looks clear how limiting the sessions data to 6 months only of the same year affected all the training process and its results. If that limitation is due to the need of keeping the size of the files low, it may be the case to limit the extraction of the users' data to one year only but to extend that of the online sessions to the same period of time.