Data Science Bootcamp

How to retain B2B fintech clients

Amina Nasri

Using Machine learning to predict churning rate

- Target variable known
- Classification binary
- Real business up to date data

Data

Data is your friend if well documented

- Database with monthly extracts of clients characteristics since 2016
- 52 features and 1 target (date of churn)
- 109k observations

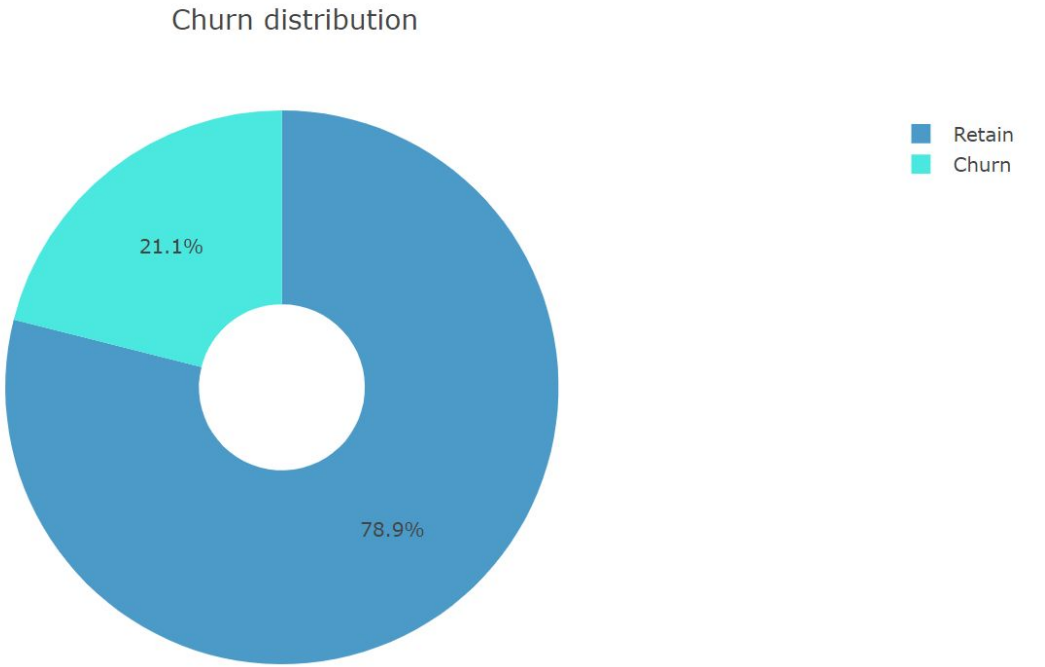
→ More than 50% of time spent on understanding the meaning and structure of data

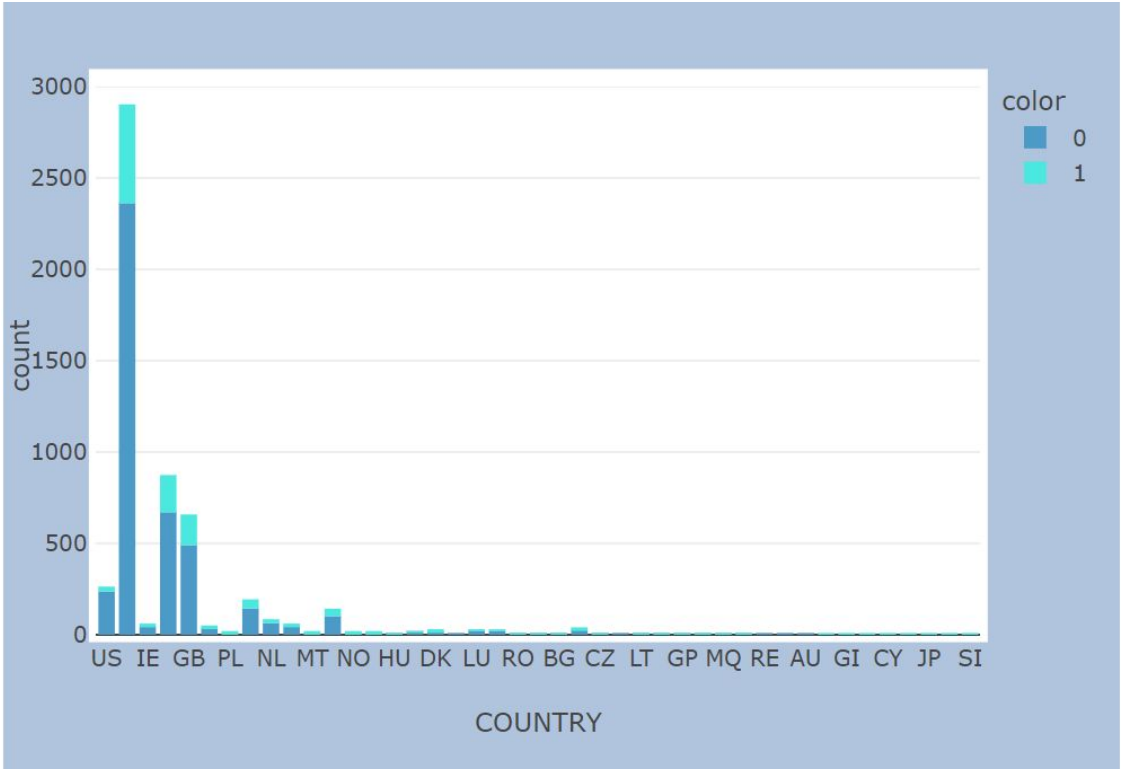
Fixed features

Temporal features

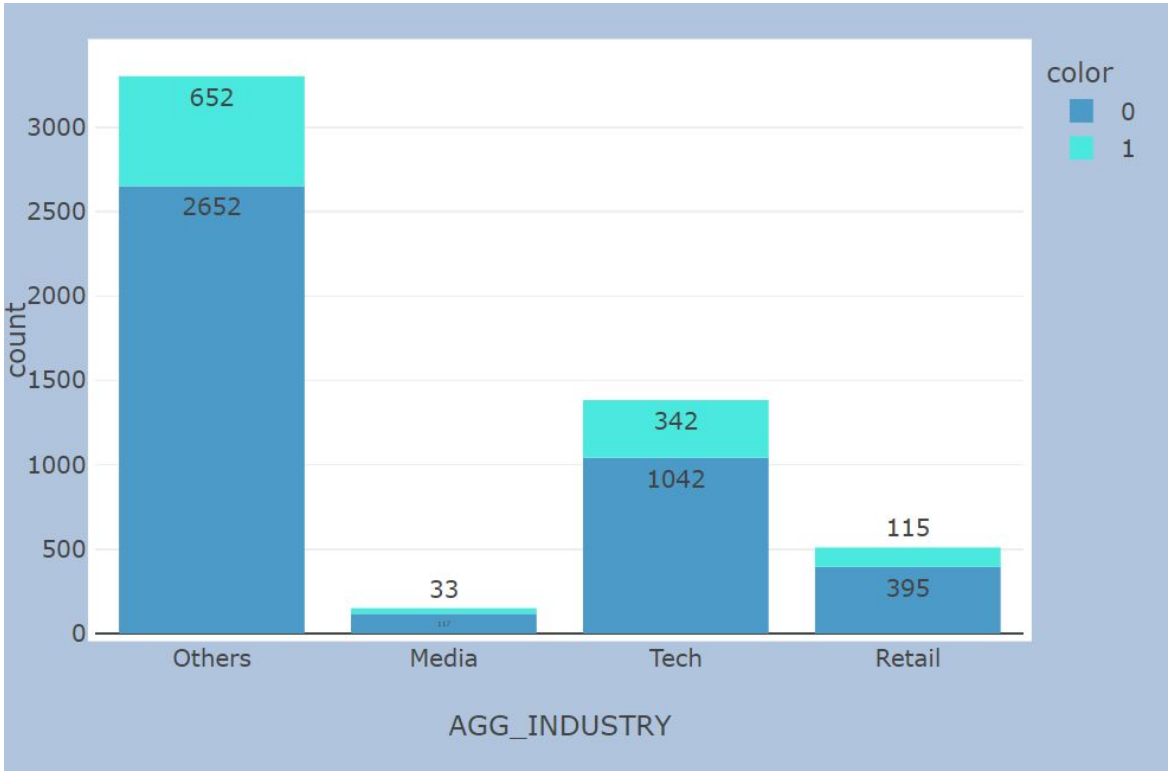
ENTITY_ID	General features	Health_features	Usage_features	Payments_features
Client_A_month1	Country, FTE, Industry …	Health score, NPS, Jira tickets …	Login_per_user, NB_extract …	Nb_payment, Nb_cards, Monthly_revenue
Client_B_month1	Country, FTE, Industry …	Health score, NPS, Jira tickets …	Login_per_user, NB_extract …	Nb_payment, Nb_cards, Monthly_revenue
Cient_A_month2	Country, FTE, Plan…	Health score, NPS, Jira tickets …	Login_per_user, NB_extract …	Nb_payment, Nb_cards, Monthly_revenue
Client_C_month3	Country, FTE, Industry …	Health score, NPS, Jira tickets …	Login_per_user, NB_extract …	Nb_payment, Nb_cards, Monthly_revenue

Data visualisation

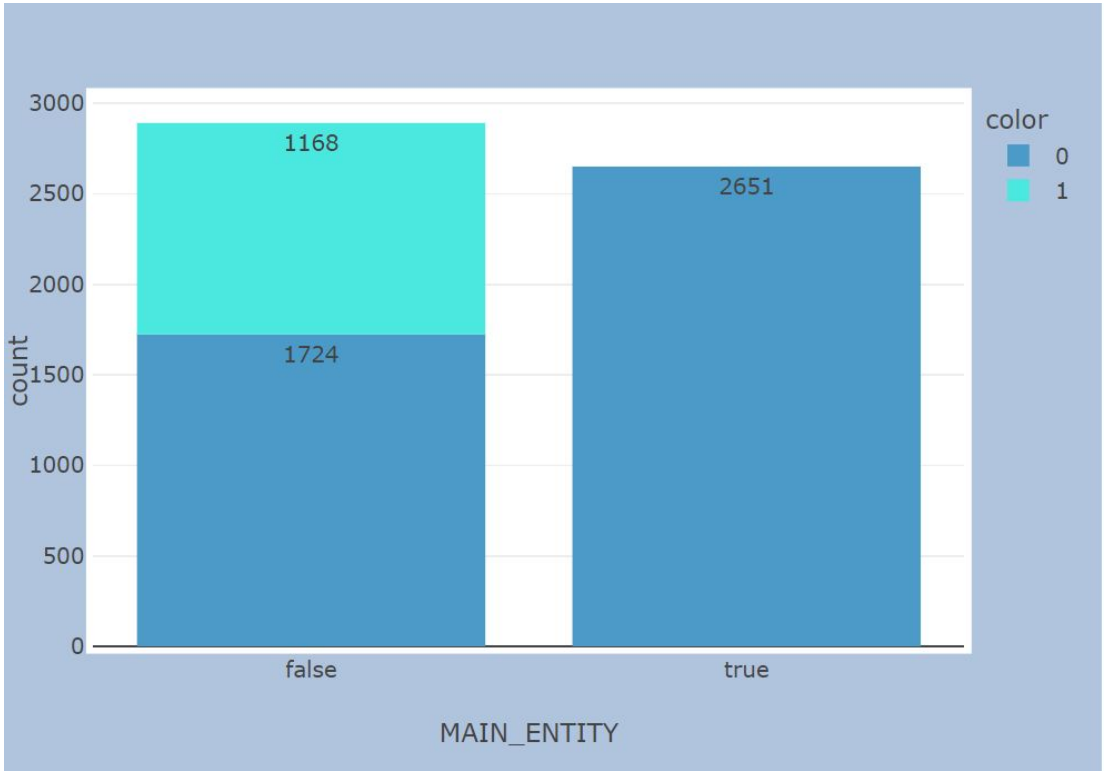




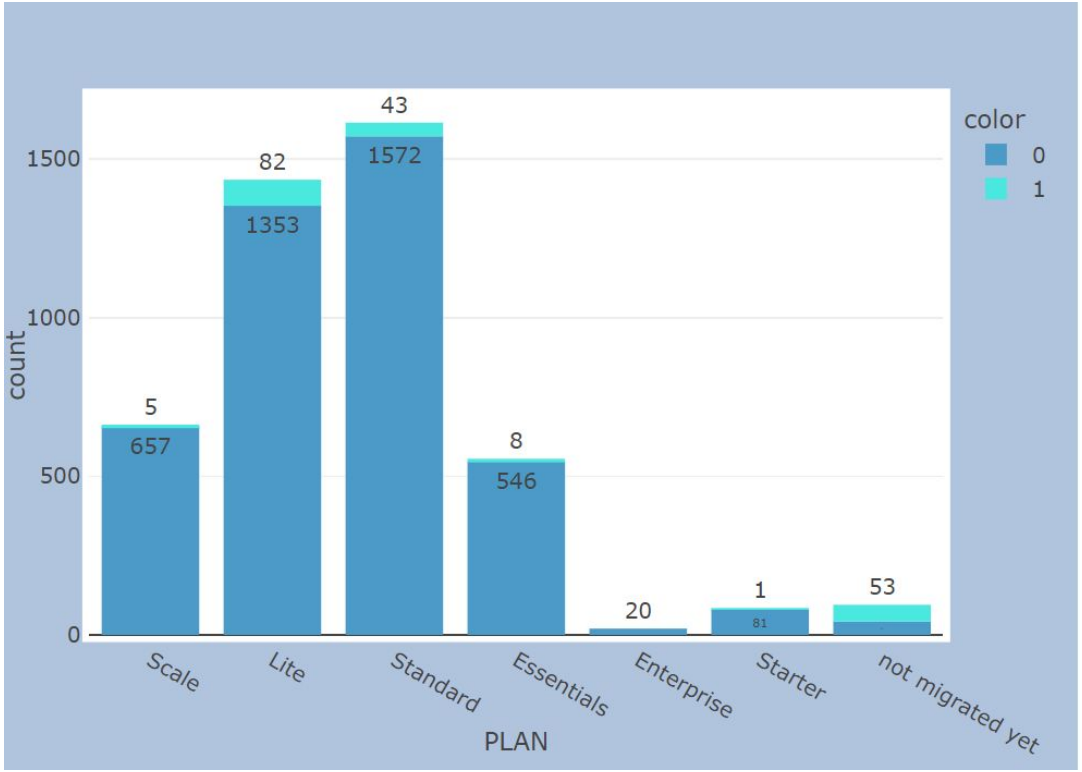
0 : retained
1 : churned



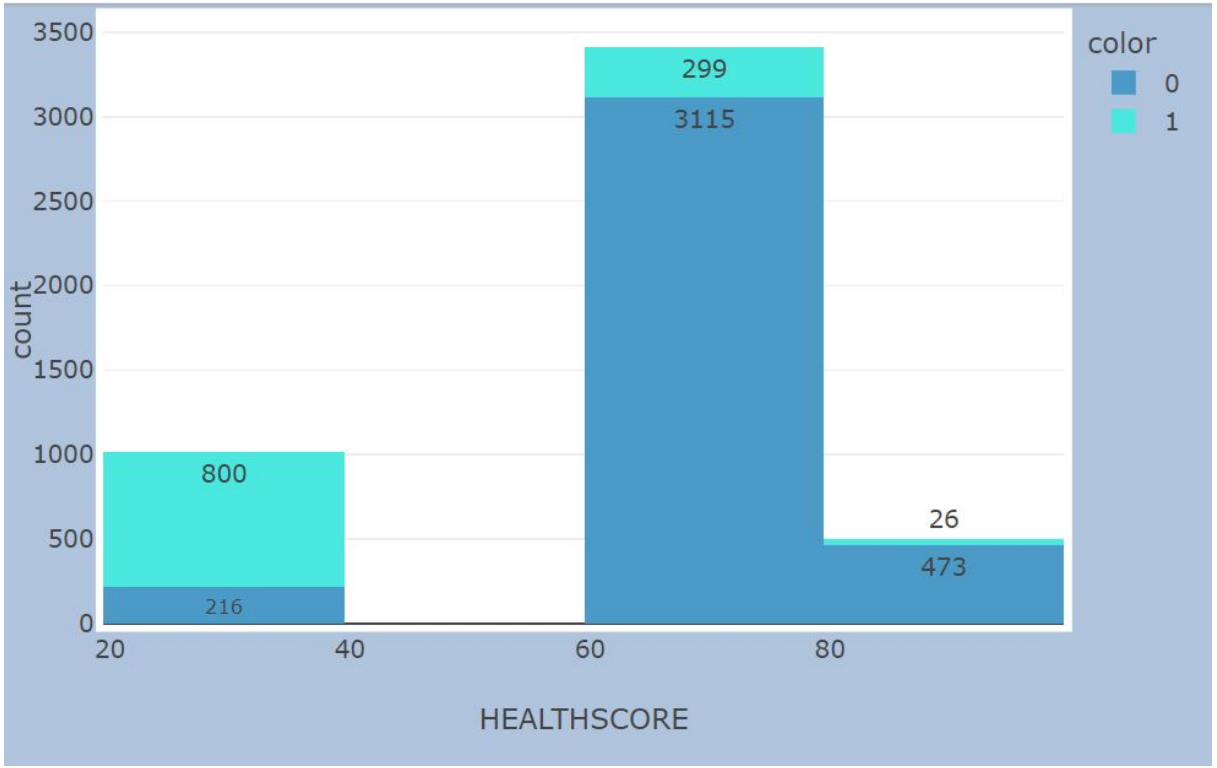
0 : retained
1 : churned



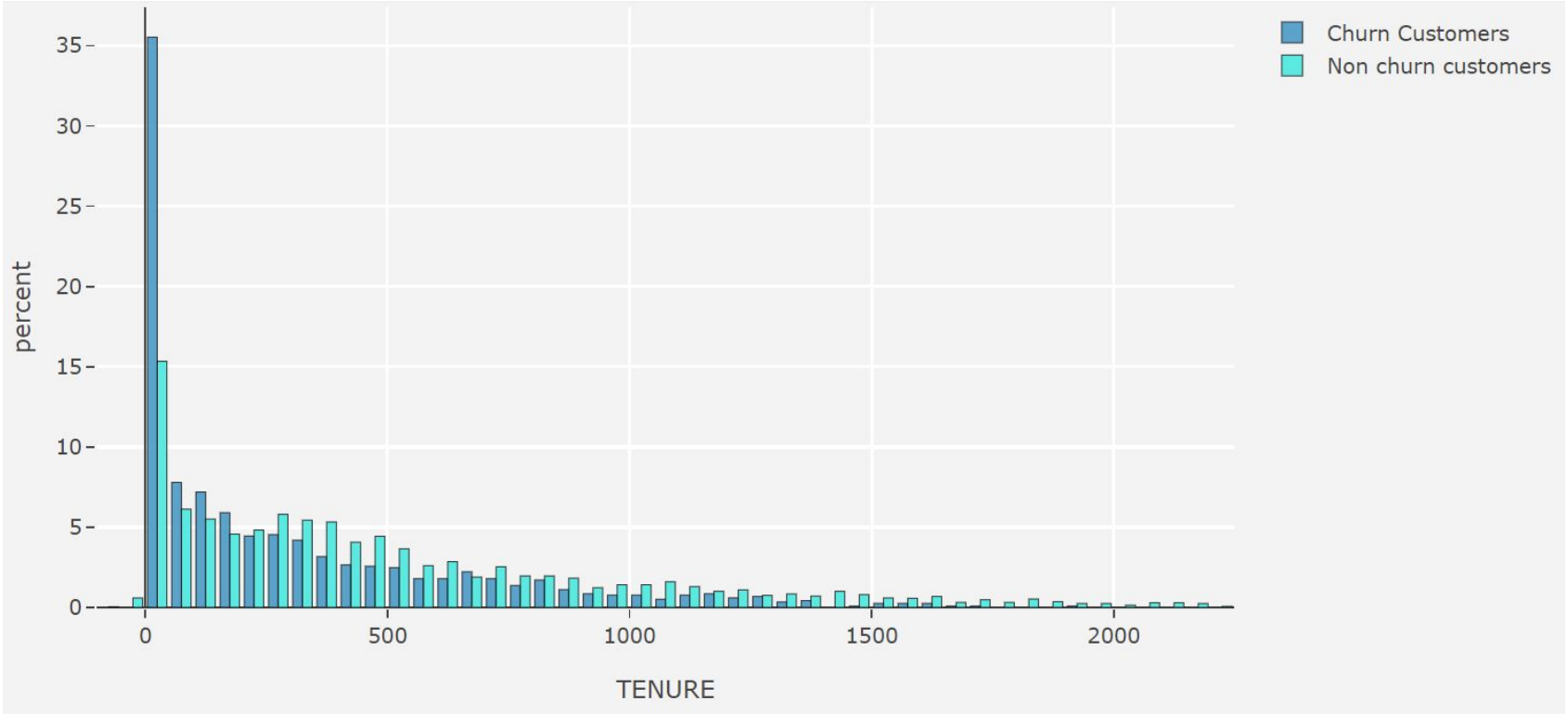
0 : retained
1 : churned



0 : retained
1 : churned



0 : retained
1 : churned

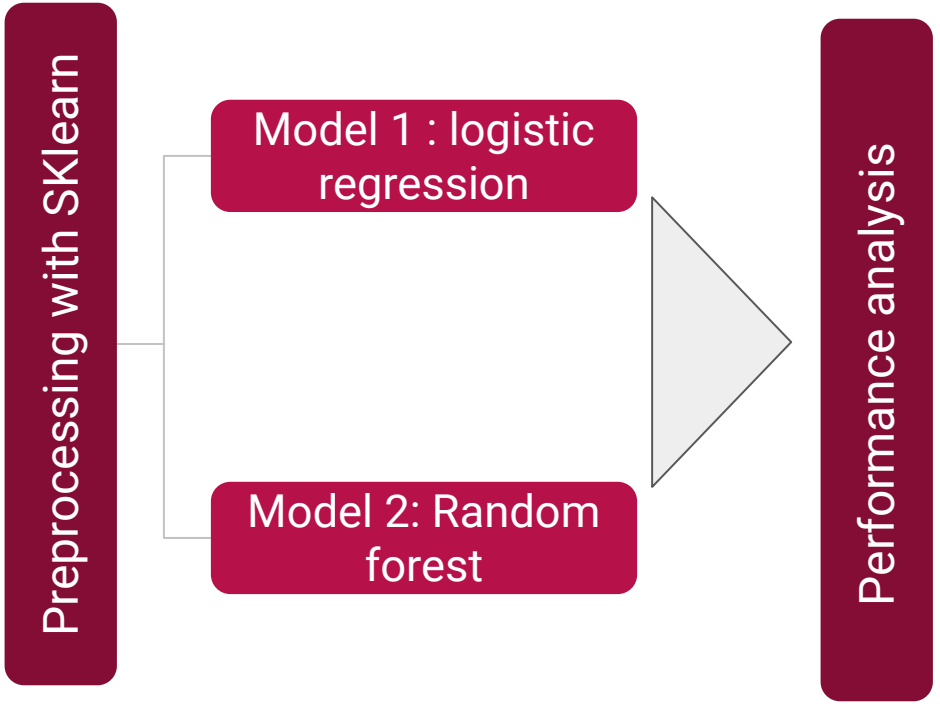

Churn Customers
Non churn customers
percent
35
30
25
20
15
10
5
0
0
500
1000
1500
2000
TENURE

Cleaning data and feature engineering

- Deleting columns (non relevant, correlated , double columns)
- Replacing outliers in categorical features with “OTHER” category
- Replacing missing values by meaningful values when relevant (False,0)
- Deleting rows with missing values in important features
- Creating new variables :
  - Target (1 if churn , O otherwise)
  - Tenure : duration of active portfolio
  - Time-to-load : time until portfolio activation

Machine Learning

Training and regularisation of machine learning models



Random forest classifier has better scores

Logistic regression

Train score = 89.1%
Test score = 88.1%
Lift score = 4.5

Regularization with gridSearchCV
L2 (Ridge)
C = 1000

Permutation importance features :

1. General feature (main entity)
2. General feature (plan)
3. Health (score)
4. Usage (Tenure)
5. Usage (Time to load)

Random Forest

Train score = 93.3%
Test score = 89.6%
Lift score = 4.6

Regularization with gridSearchCV
max_depth = 10
min_sample_leaf = 2
min_sample_split =2
n estimators = 20

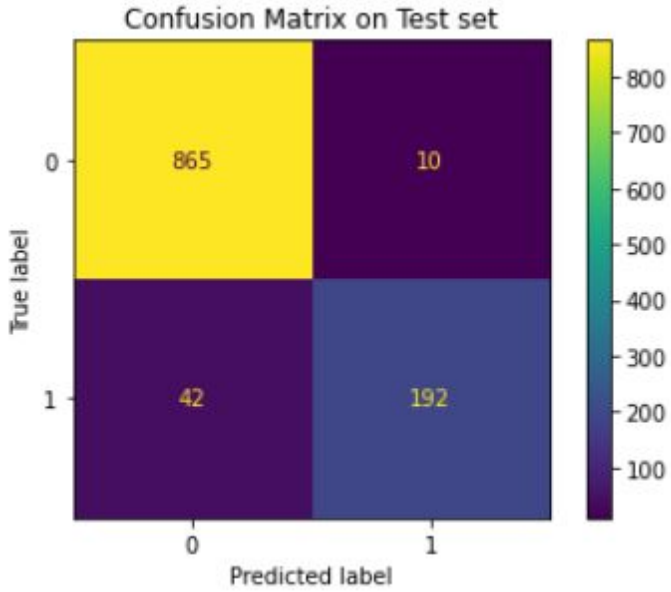
Permutation importance features :

1. General feature (main entity)
2. Health (score)
3. General feature (plan)
4. Usage (Time to load)
5. USage (Login per user)

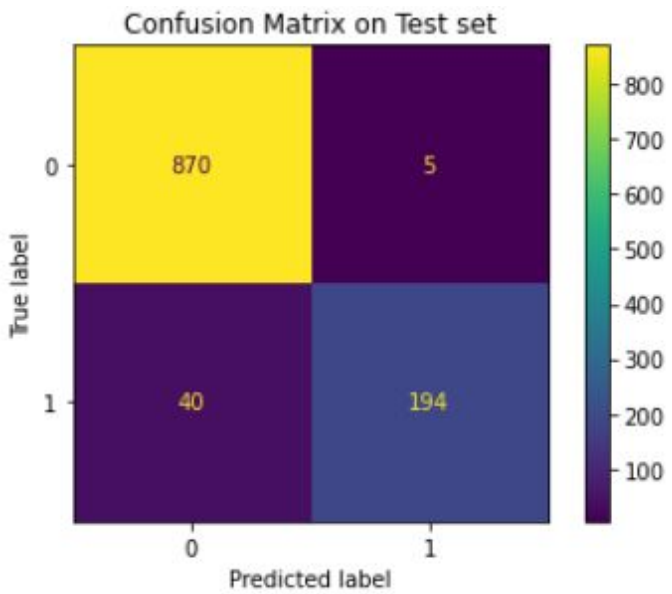
82% of churned were predicted right with RF !

0 : retained
1 : churned

Logistic regression



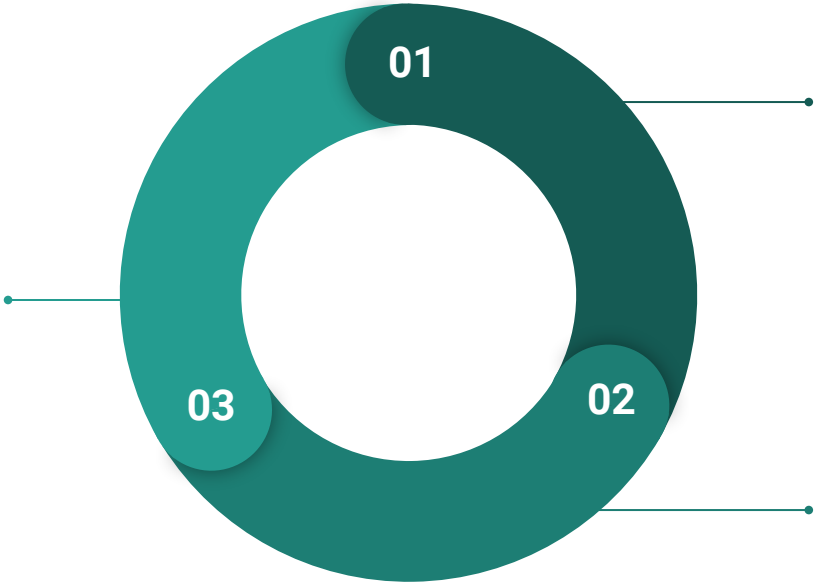
Random Forest



A solution for Customer success teams to prevent churning clients

New client subscription

ETL: collect, store and manage relevant data
By : Data engineering team



Sales & Marketing actions

By : Customer success or sales

Probability of churn

API: prediction by the AI solution
By : Data science team

First baby steps
...

The best is yet
to come !

Next steps

- Data documentation to create
- More investigations on critical features (Main_Entity, Date features, scores calculation method, missing values ...)
- Exploring temporal aspect of monthly extraction , time series modelization to implement

Thank you !
Any questions ?

annexes

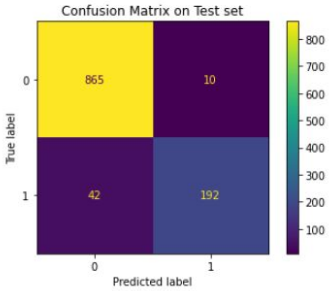
Performance comparison

```
GridSearchCV(cv=3, estimator=LogisticRegression(),
             param_grid={'C': array([1.00000000e-03, 3.162277e
       1.00000000e+03]),
                         'penalty': ['l1', 'l2']},
             scoring='f1')
accuracy on train set: 95.6 %
accuracy on test set : 95.3 %
f1-score on train set: 89.1 %
f1-score on test set: 88.1 %
Lift score on train set: 4.4
Lift score on test set: 4.5
```

Classification report :

	precision	recall	f1-score	support
0	0.95	0.99	0.97	875
1	0.95	0.82	0.88	234
accuracy			0.95	1109
macro avg	0.95	0.90	0.93	1109
weighted avg	0.95	0.95	0.95	1109

Accuracy Score : 0.9531109107303878

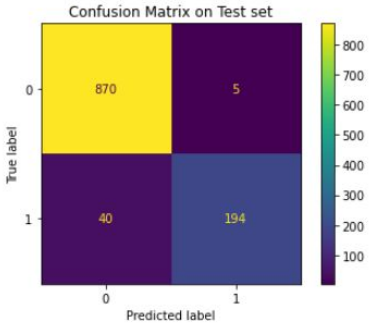



```
RandomForestClassifier(max_depth=10, min_samples_leaf=2, n_estimators=20)
accuracy on train set: 97.3 %
accuracy on test set : 95.9 %
f1-score on train set: 93.3 %
f1-score on test set: 89.6 %
Lift score on train set: 4.7
Lift score on test set: 4.6
```

Classification report :

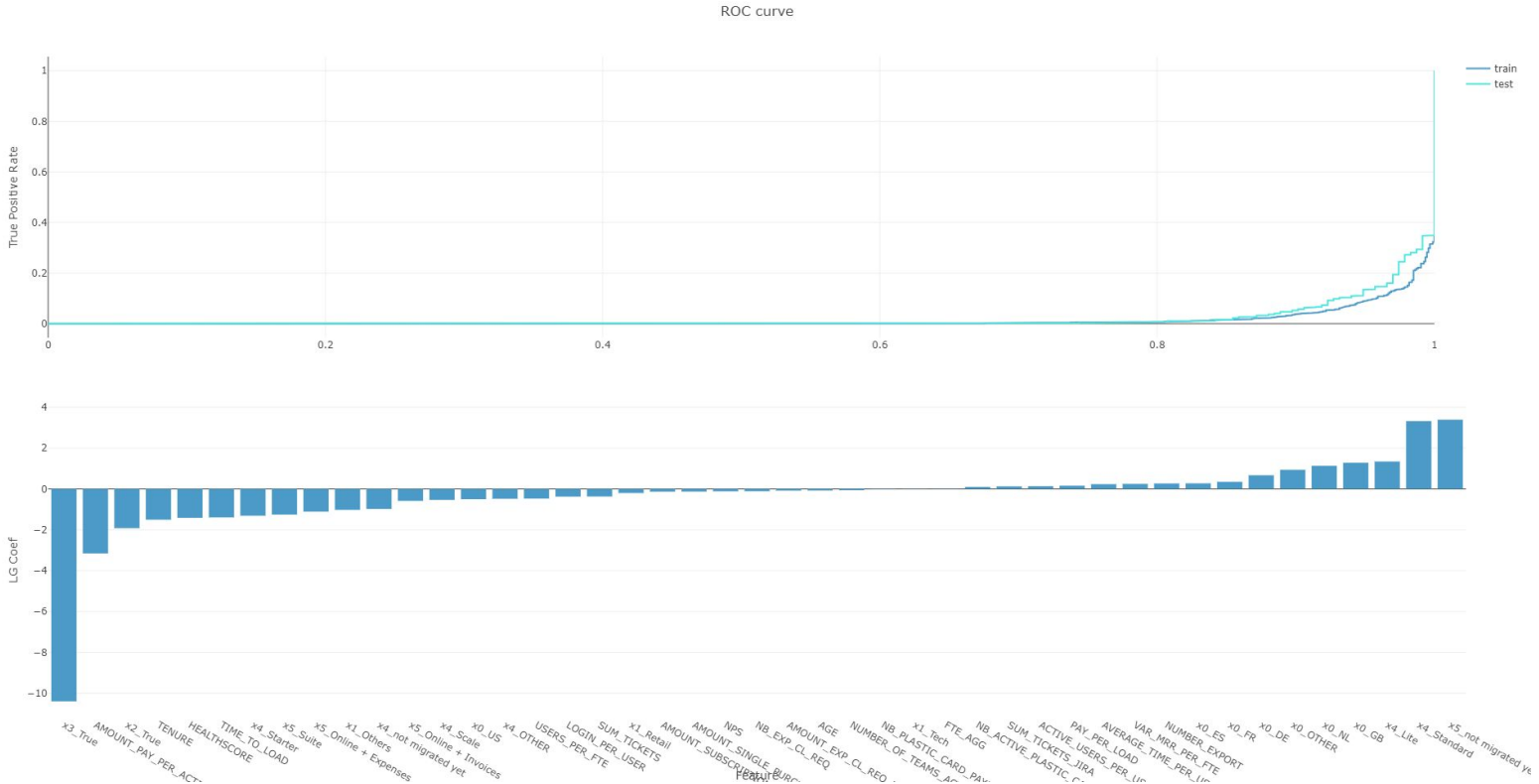
	precision	recall	f1-score	support
0	0.96	0.99	0.97	875
1	0.97	0.83	0.90	234
accuracy			0.96	1109
macro avg	0.97	0.91	0.94	1109
weighted avg	0.96	0.96	0.96	1109

Accuracy Score : 0.9594229035166817



Légende de l’image ou du graph

ROC curve & Feature coef Logistic Regression



ROC curve & Feature coef Random Forest