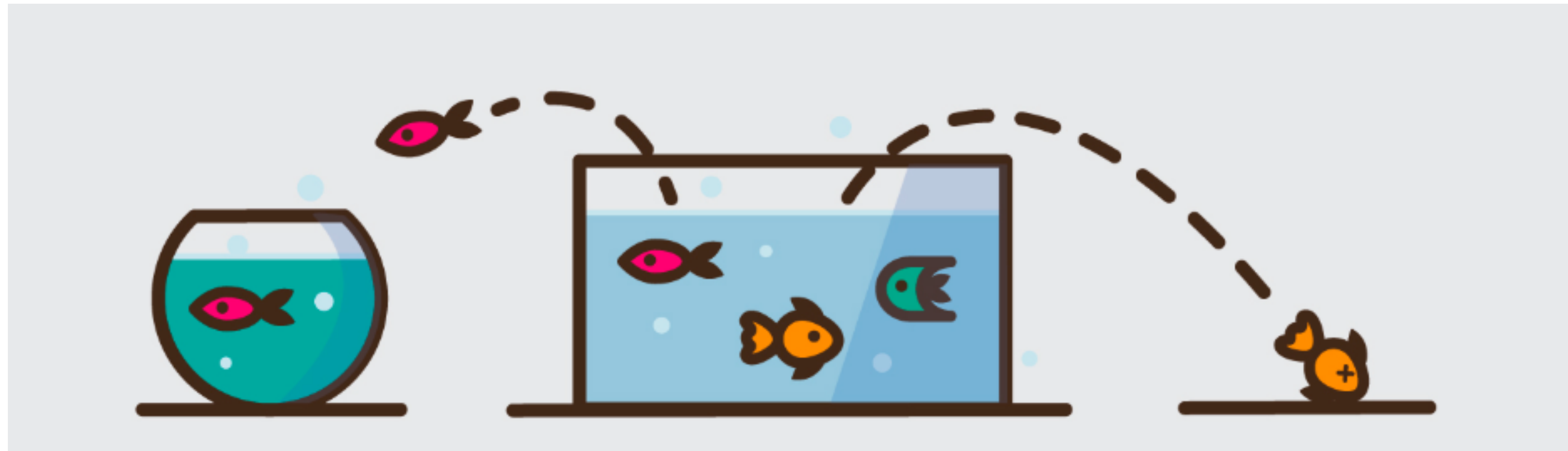


Predicting Churn for Bank Customers

Marina Trofimovich

Who's going to leave?



Customer churn

Let's try to predict!

Profit

- predict future revenue,
- identify and improve upon areas where customer service is lacking.

Problem

Bank customer churn dataset - [Kaggle](#).

14 features, 10.000 customers.

Target

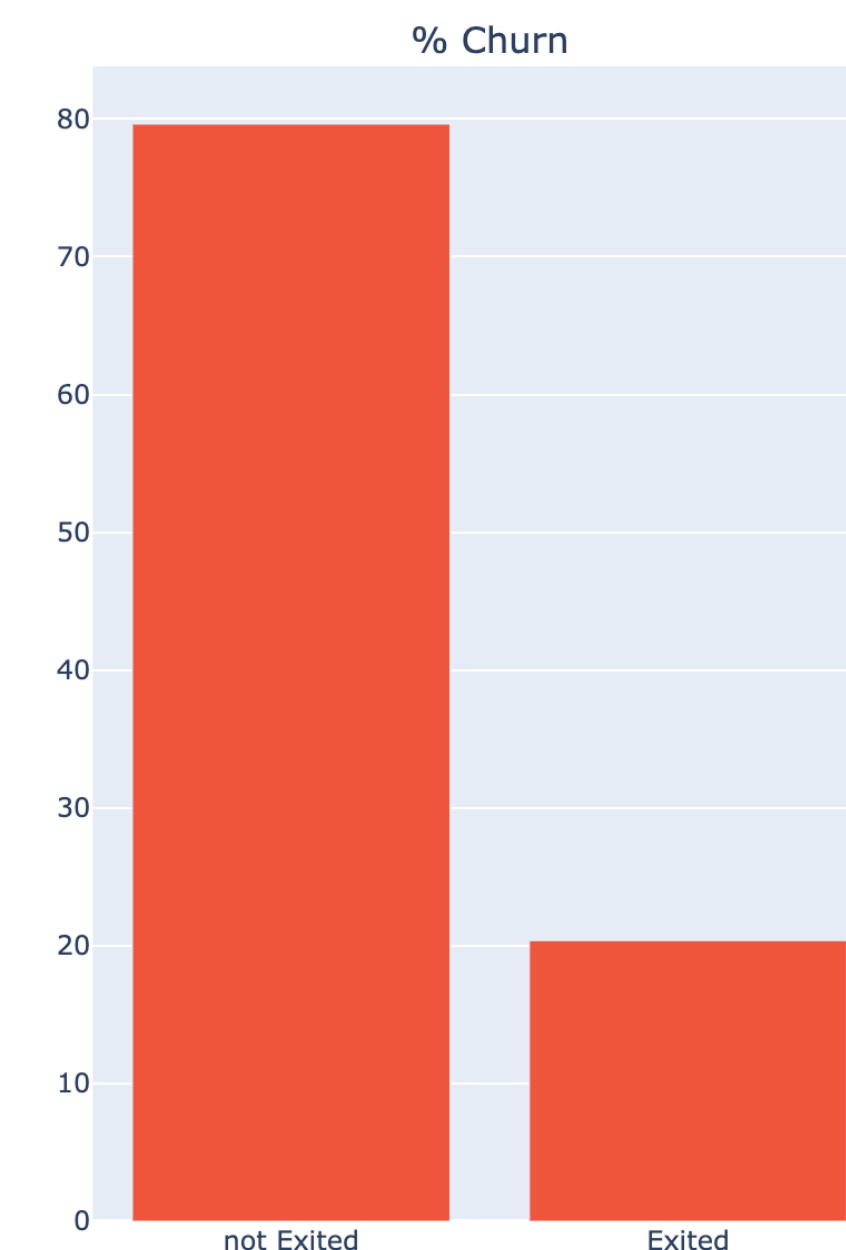


	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Independent variables

Objectives

- identify and visualize which factors contribute to customer the churn,
- build a prediction model that will classify if a customer is going to churn or not.

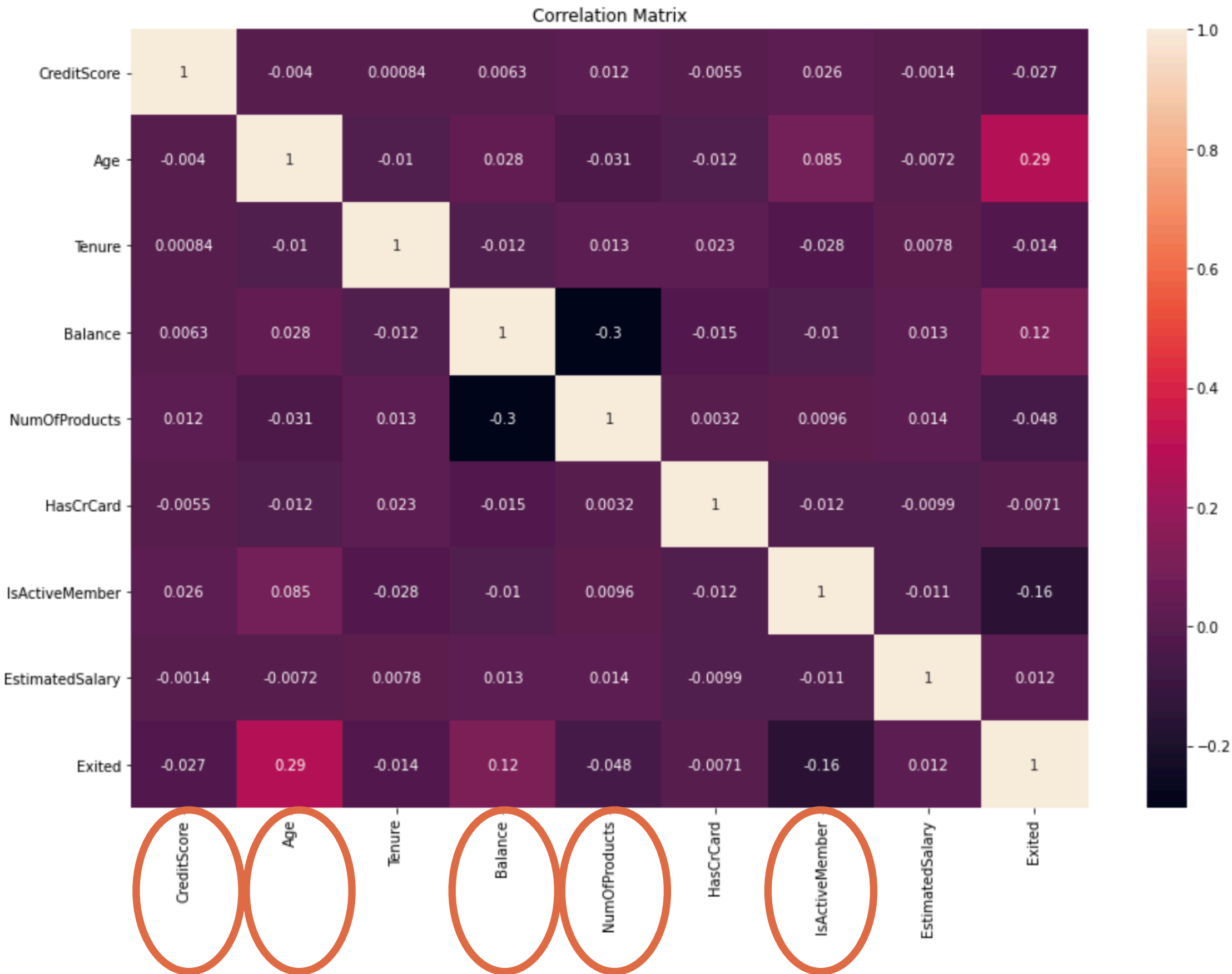


Methodology

Features selection that might contribute the churn



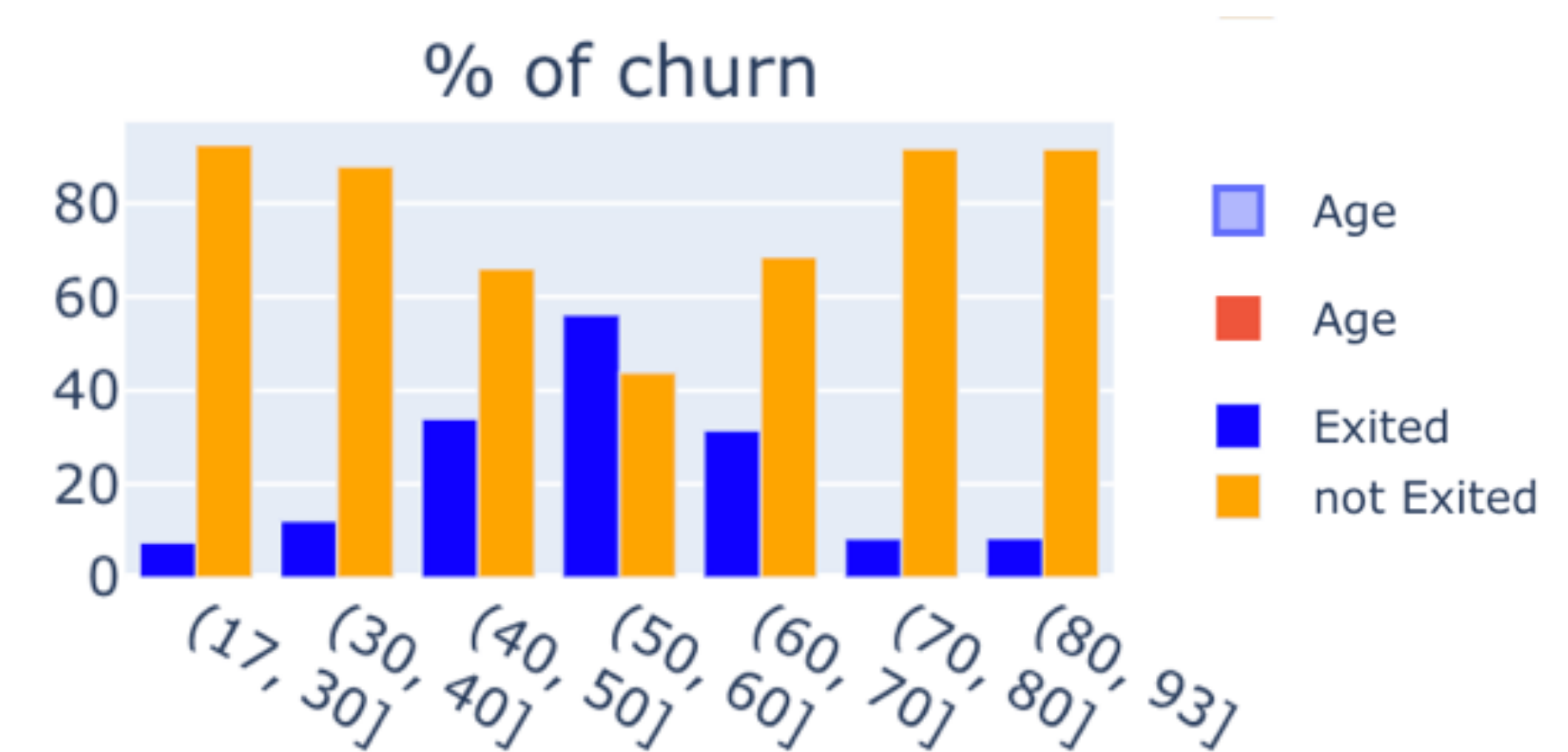
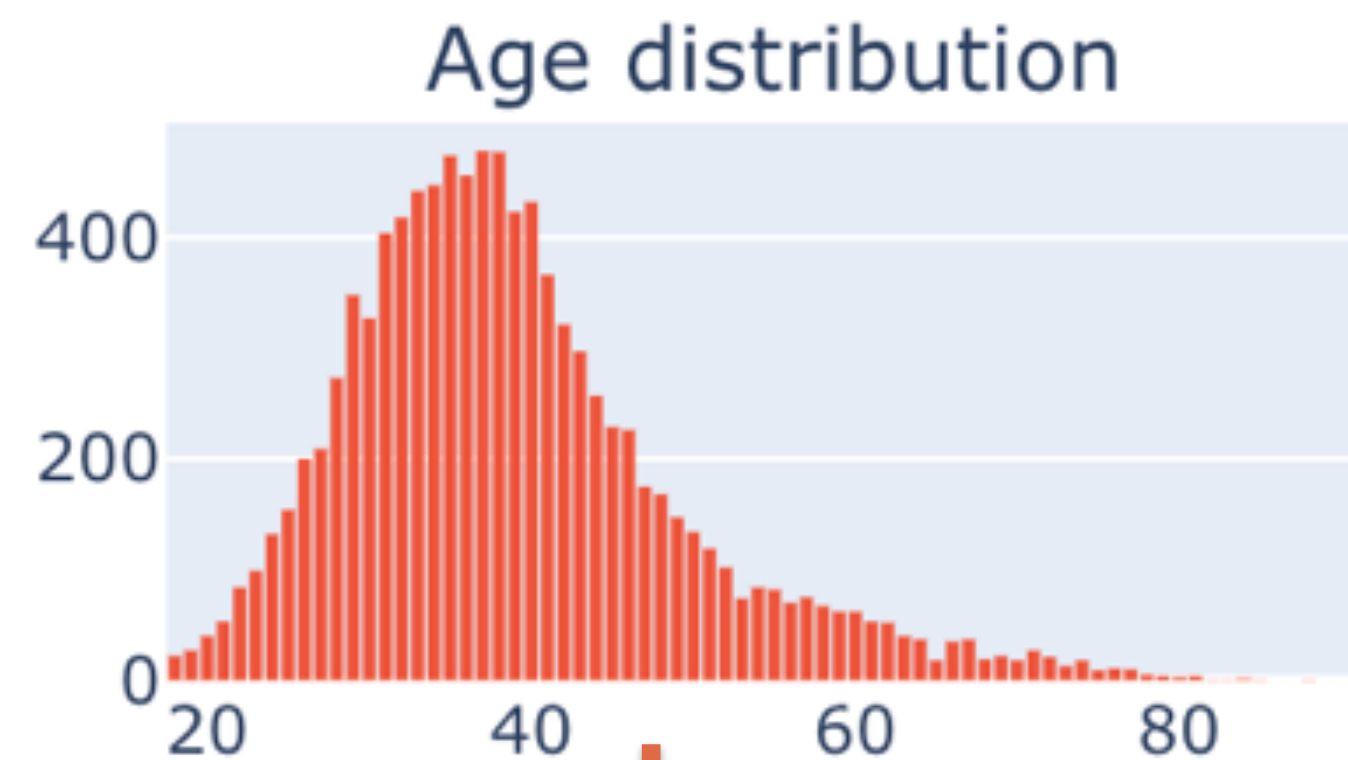
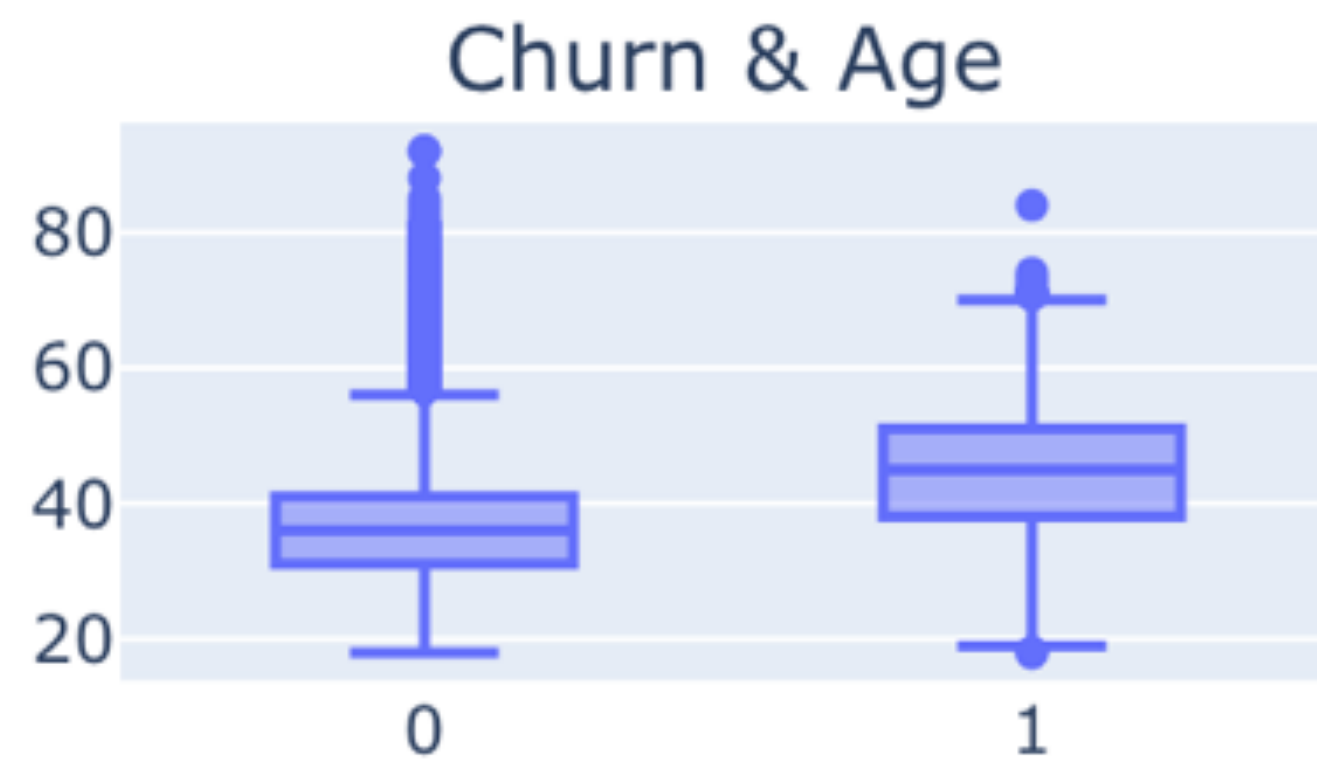
	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0



↑
Target

Features that contribute the most

Age



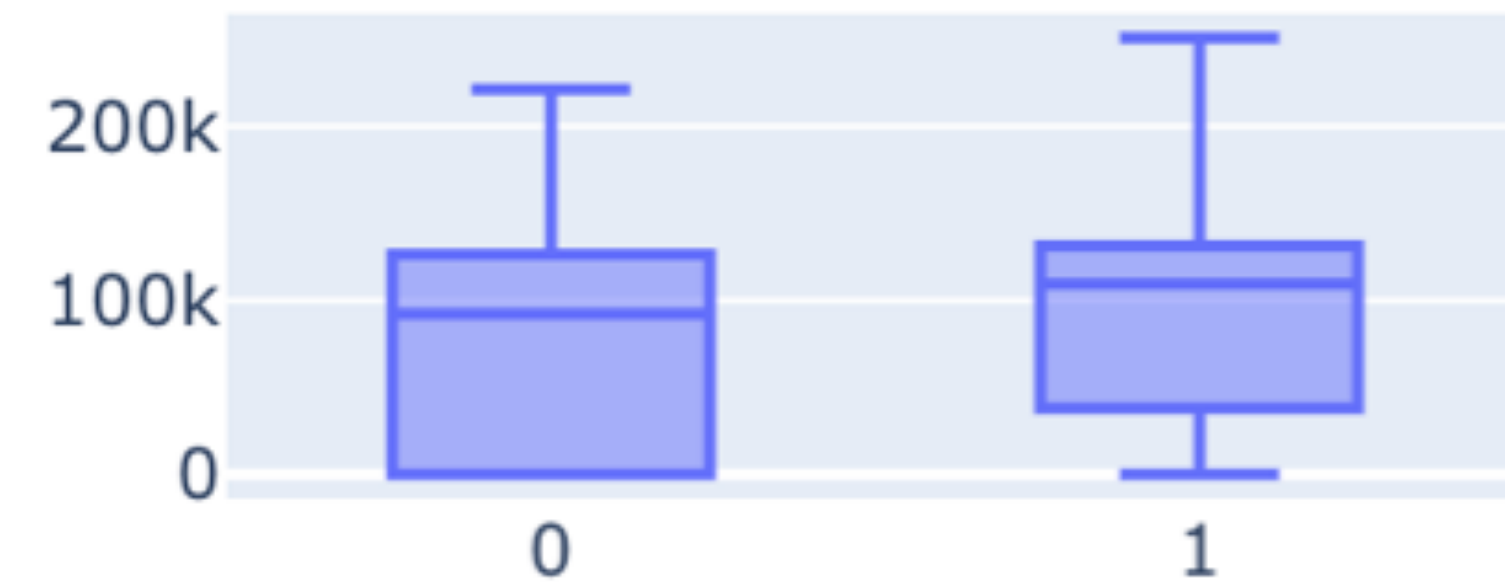
mean age for “Exited” - 45 years,
“not Exited” - 37 years.

the highest churn (> 56 %) for age range (50,60]

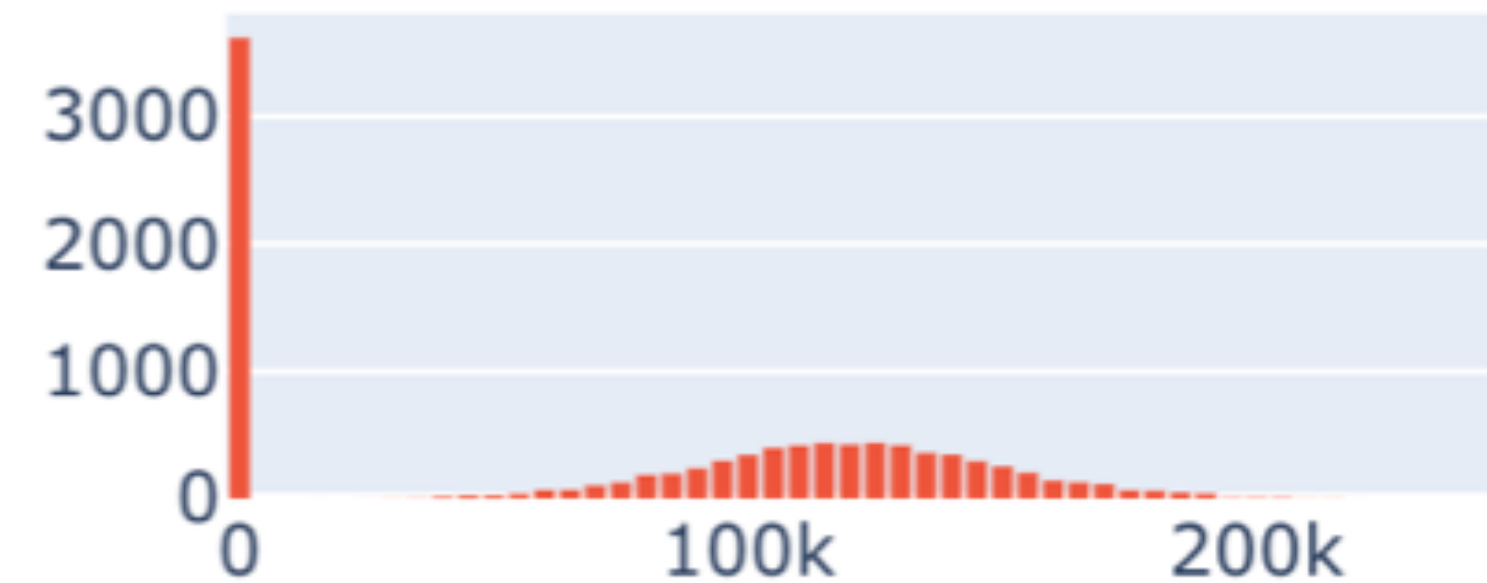
Features that contribute the most

Balance

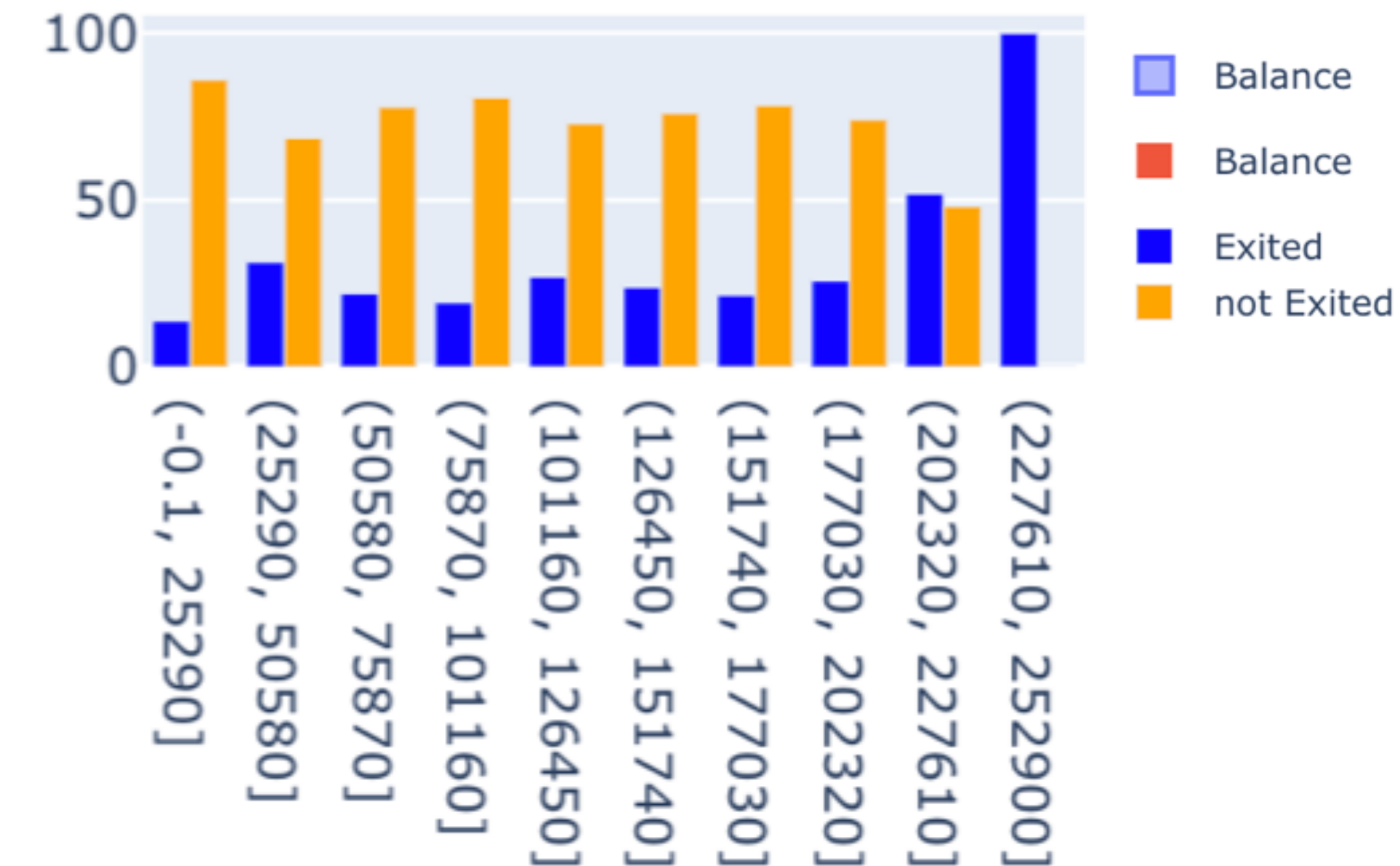
Churn & Balance



Balance distribution



% of churn



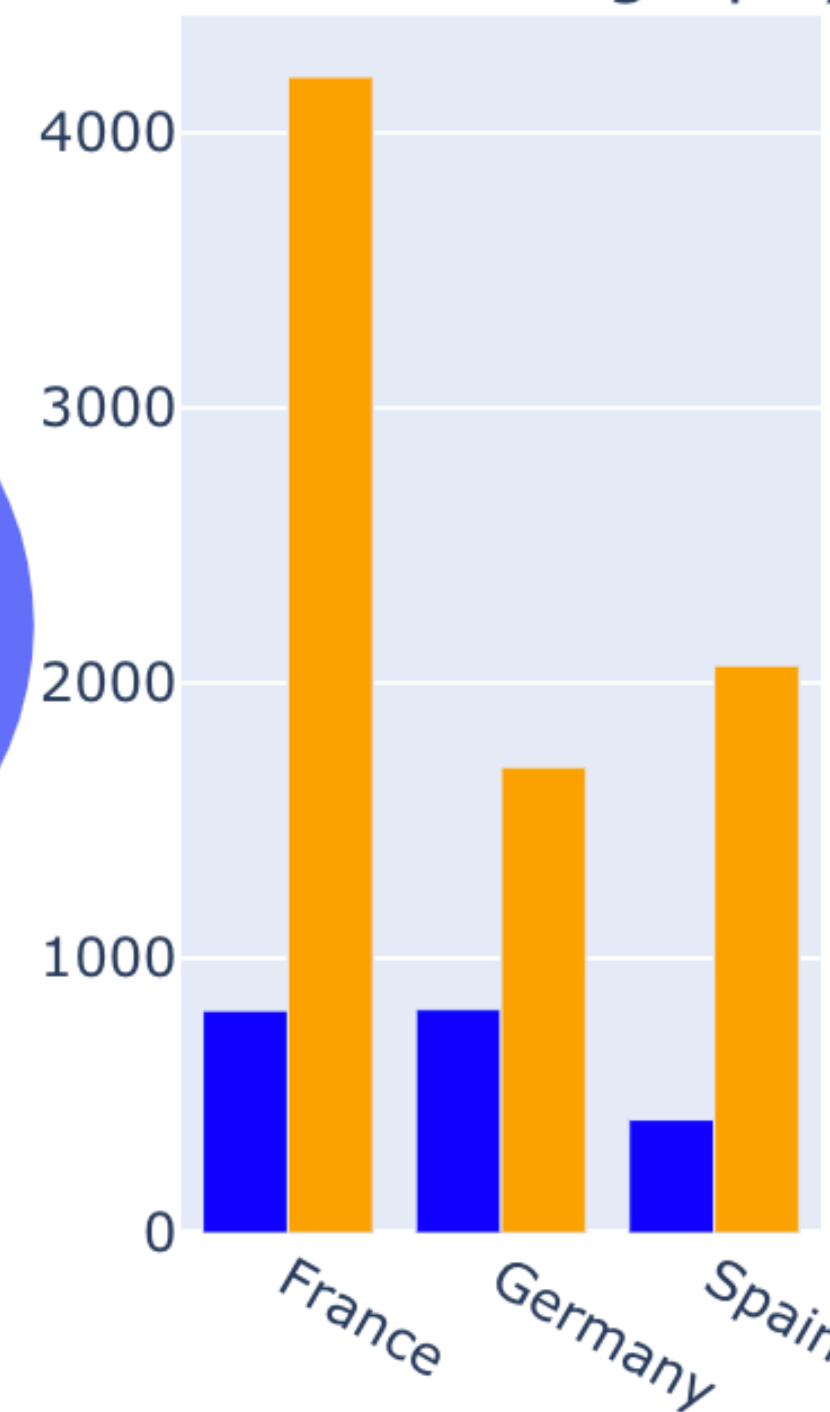
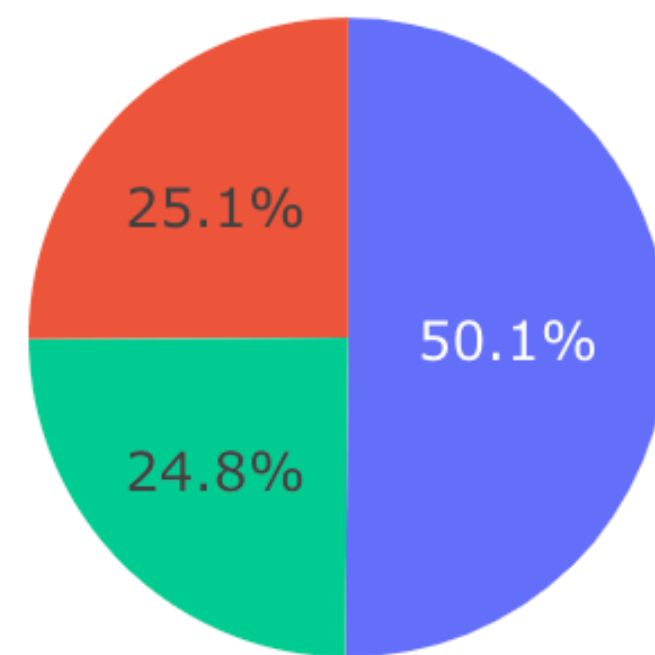
mean balance for “Exited” - 91.109,
“not Exited” - 72.745.

the highest churn (up to 100 %) for balance > 202.320
only 29 customers

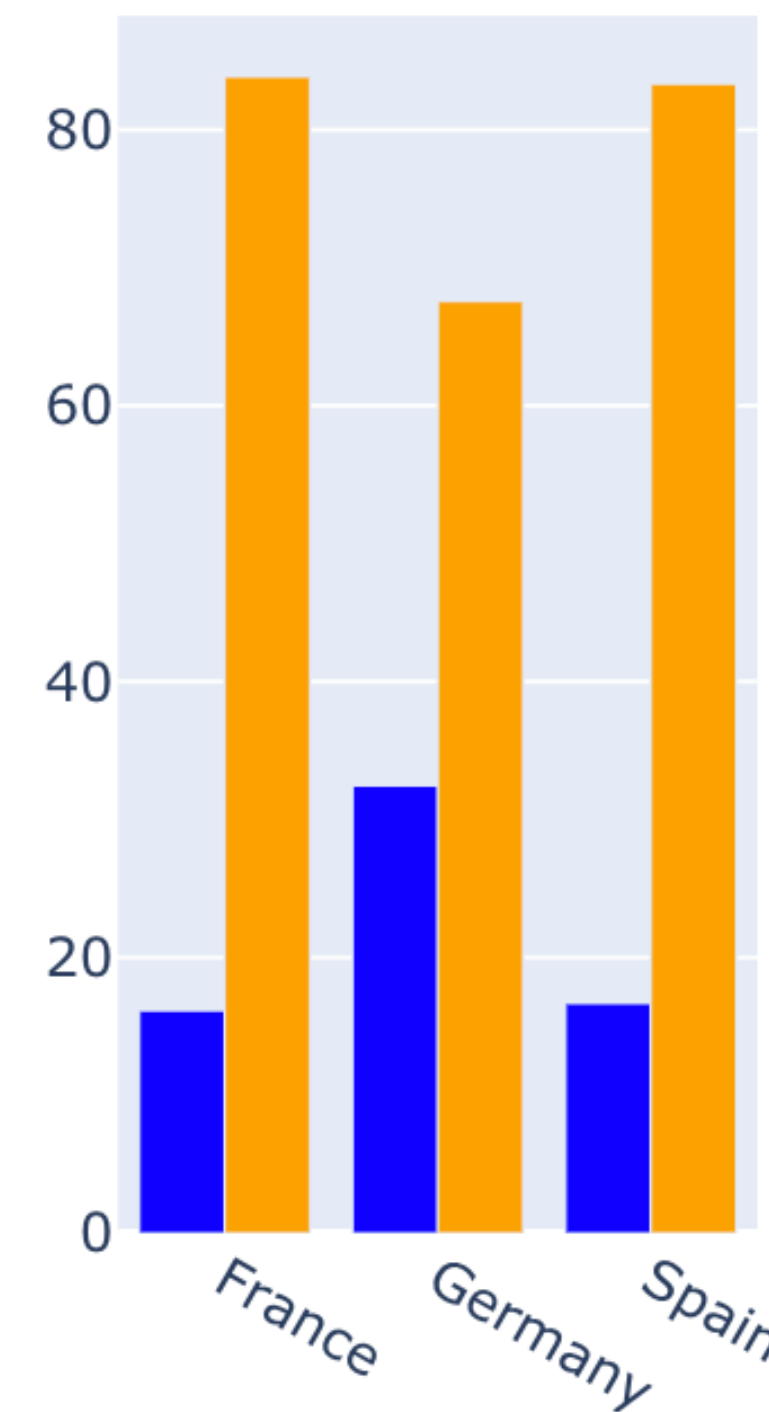
Features that contribute the most

Geography

Countries distribution Churn & Geography



% of churn



- France
- Germany
- Spain
- Exited
- not Exited
- Exited
- not Exited

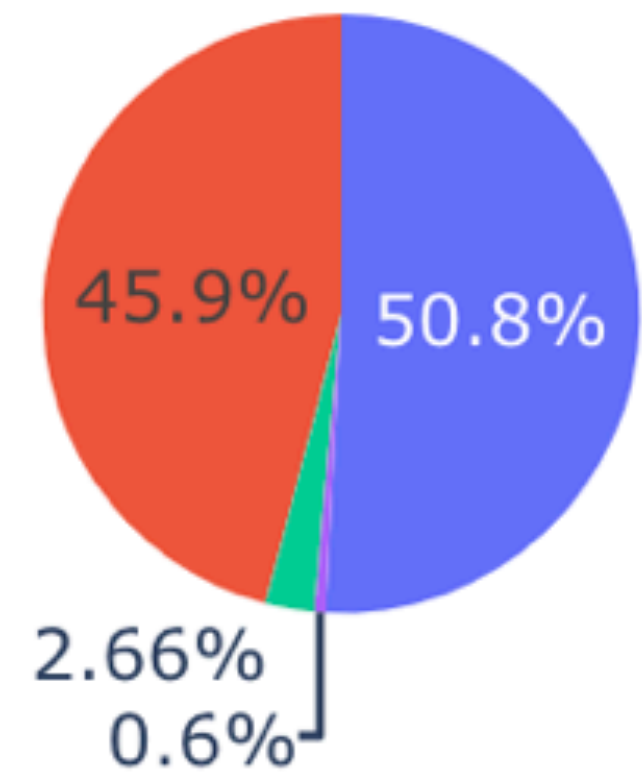


the highest churn (> 32 %) for Germany

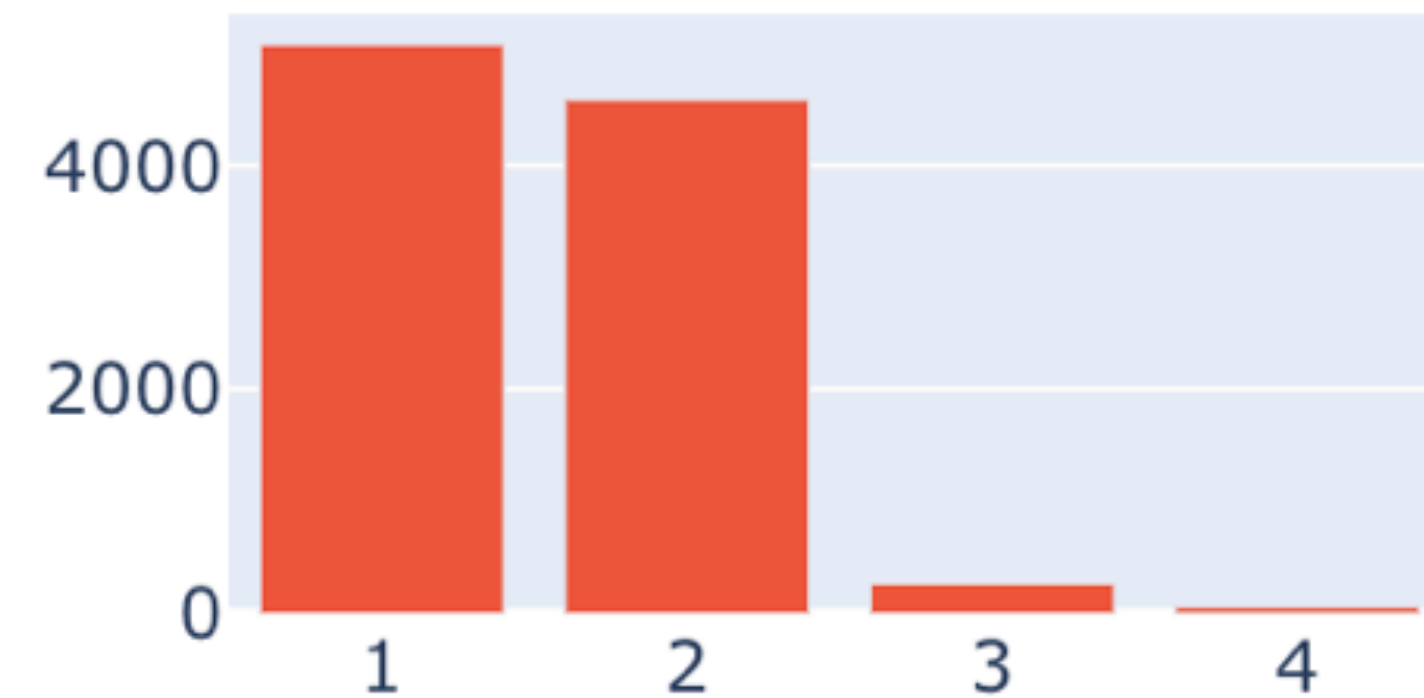
Features that contribute the most

NumOfProducts

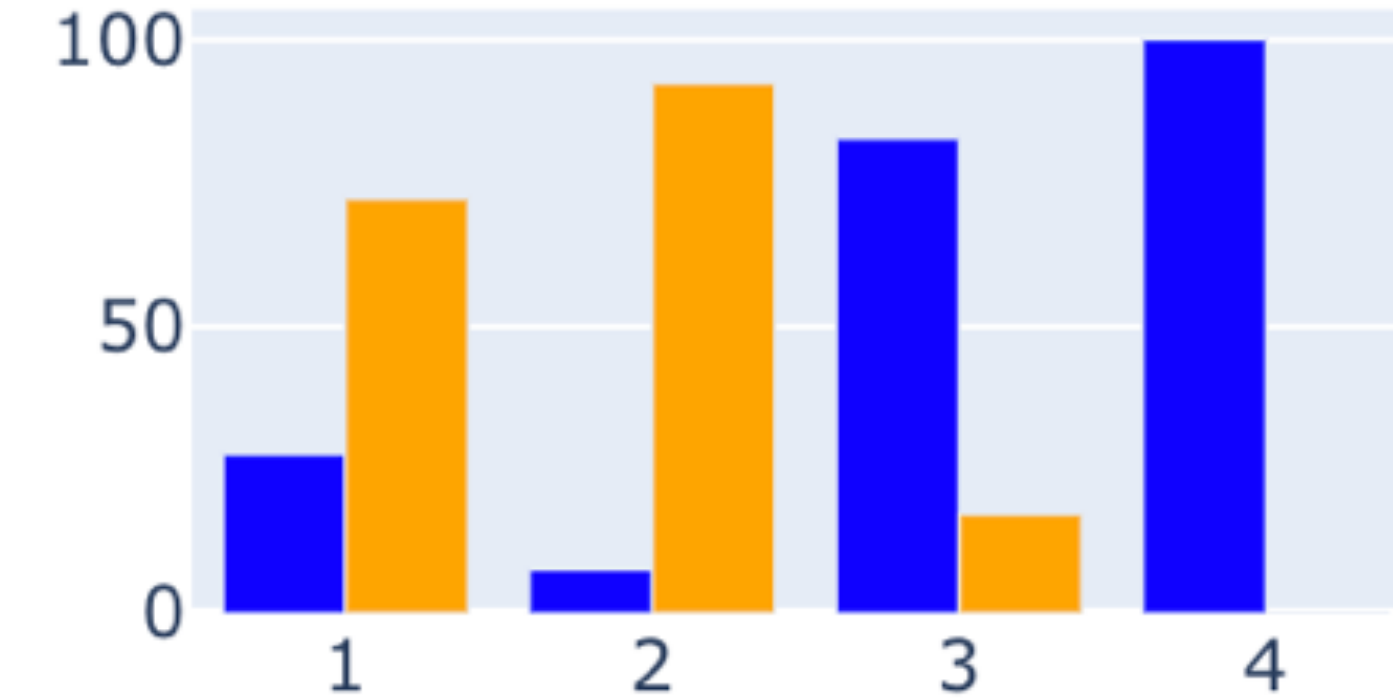
NumOfProducts



NumOfProducts distribution



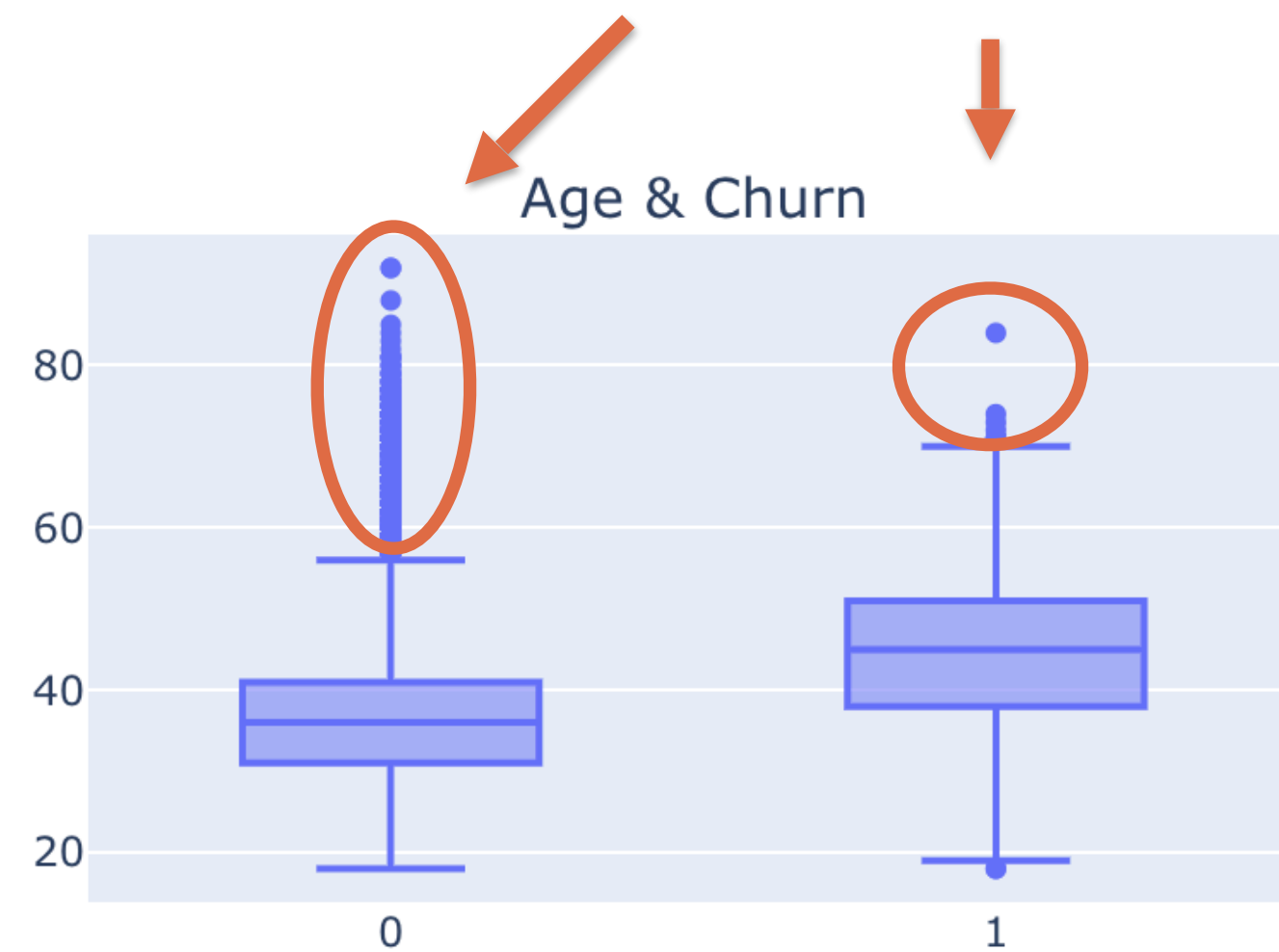
% of churn



**the highest churn: 83% - 3 products,
100% - 4 products**

Data Preprocessing

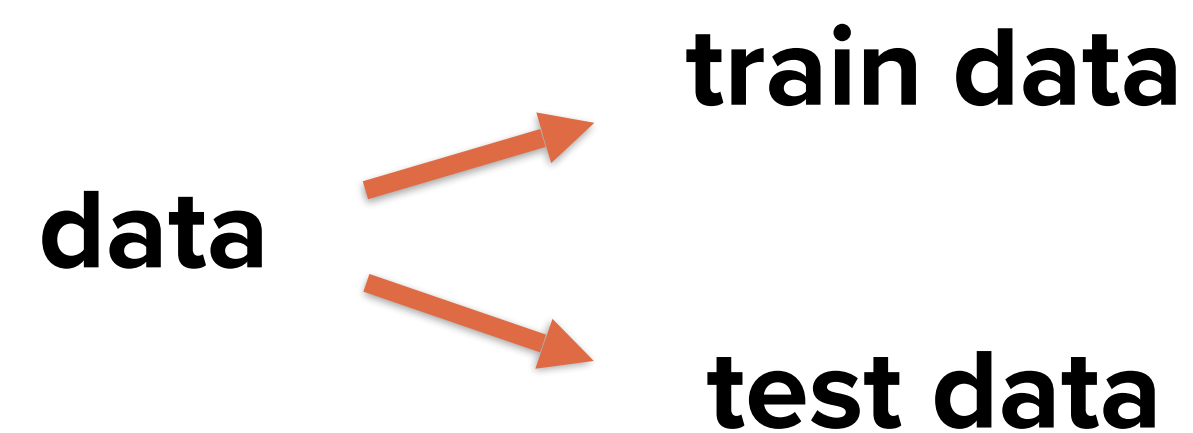
1. Removing outliers



2. Encoding categorical variables

Geography			
France	1	0	0
Spain	0	0	1
France	1	0	0
France	1	0	0
Spain	0	0	1
...
...	1	0	0
France			

3. Splitting the dataset



4. Scaling

$$\frac{x - \bar{x}}{s}$$

Machine learning models



	Accuracy	Precision	Recall	F1
Logistic regression	0.87	0.73	0.54	0.62
K nearest neighbours	0.86	0.78	0.45	0.57
Support Vector Machine	0.87	0.84	0.48	0.61
Random Forest	0.87	0.73	0.55	0.63

Conclusions

Conclusions

Features that contribute the most to the customer churn:
Age, Balance, Geography, NumOfProducts.

The best model - Random Forest.

The accuracy of prediction - 87%.

It predicts 55% of actual "Exited" customers.

73% of predicted to be "Exited" customers are actual "Exited".

THANK YOU!