Customer Churn Prediction Project

TABLE OF CONTENTS

01

Business Understanding

OBJECTIVES

OUR BO

Identify customers most likely to churn and reduce churn rates to minimize financial losses.

Our DSO

Build an accurate churn prediction model to outperform traditional methods.

OBJECTIVES

OUR BO

Reduce customer churn and improve customer satisfaction by identifying and addressing key dissatisfaction points that drive churn, enabling the development of targeted retention strategies

Our DSO

Develop a high-performing classification model and explain critical churn factors using model explainability techniques.

OBJECTIVES

OUR BO

Analyze customer retention trends over time to identify the key drivers of long-term loyalty and attrition, providing actionable insights to develop targeted strategies that enhance customer retention and reduce churn.

Our DSO

Implement data-driven retention strategies informed by churn predictions.

03

Data Understanding

Dataset Overview :

- Total observations : 3333 rows
- Number of features : 20 features
- Target Variable: *Churn* (binary: Yes/No)
- Categorical: 3
- Numerical: 17

Features and target:

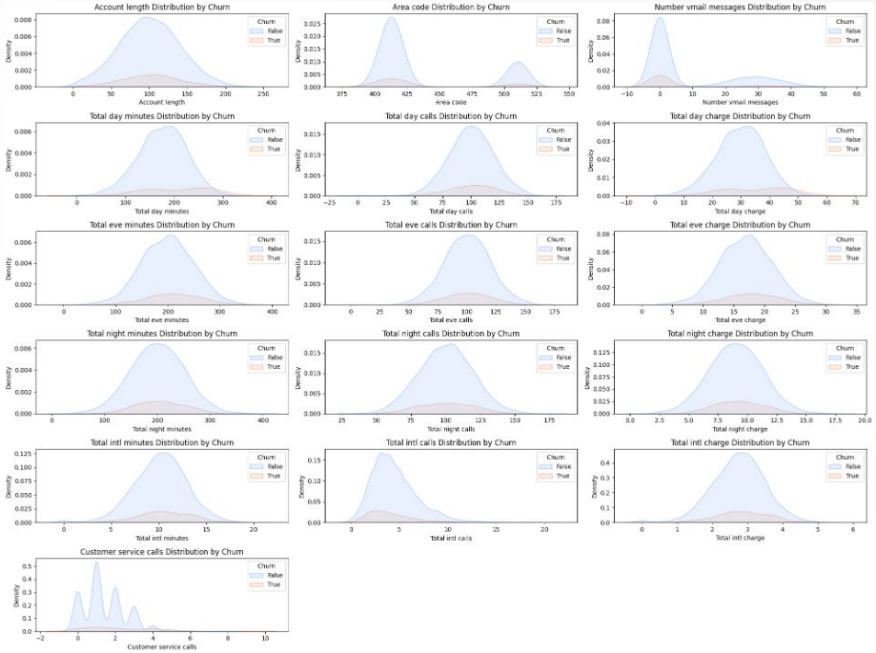
Column Name	Description
state	The U.S state where the customer resides
Account length	The number of days or months the customer has been using the company's services
Area code	The telephone area code of the customer
International plan	Indicates if the customer is subscribed to an international calling plan
Voicemail plan	Shows whether the customer has subscribed to a voicemail service
Number vmail messages	Total number of voicemail messages a customer has received
Total day minutes	Total minutes the customer spent on calls during the day
Total day calls	Total number of calls made during the day
Total day charge	Total cost for calls made during the day , typically based on minutes
Total eve minutes	Total minutes spent on calls during the evening

Features and target:

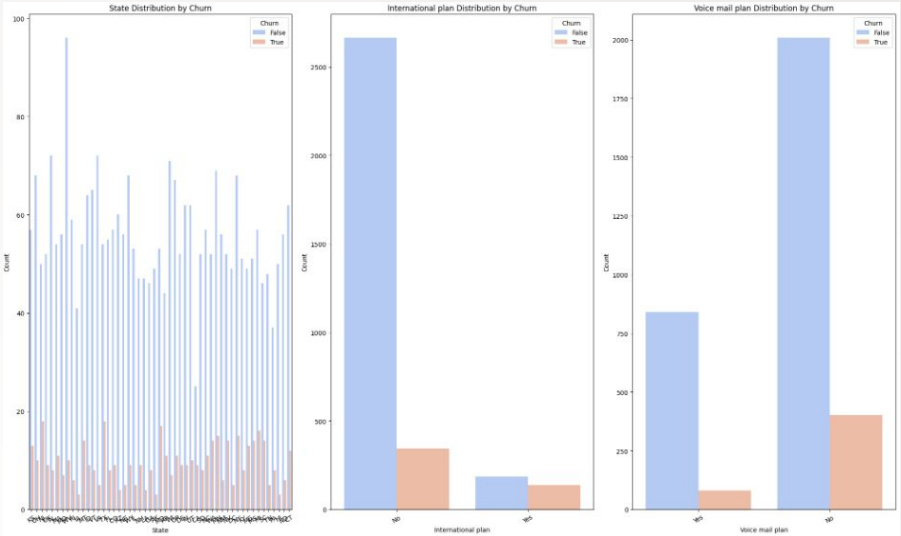
Column Name	Description
Total eve calls	Total number of calls made during the evening
Total eve charge	Total charges for evening calls
Total night minutes	Total minutes spent on calls at night
Total night calls	Total number of calls made at night
Total night charge	Charges for calls made at night
Total intl minutes	Total minutes spent on international calls
Total intl calls	Number of international calls made by the customer
Total charge	Total charges incurred for international calls
Customer service calls	Number of calls made to customer service
Churn	A binary value indicating whether the customer has left the service (churned)

Data visualization:

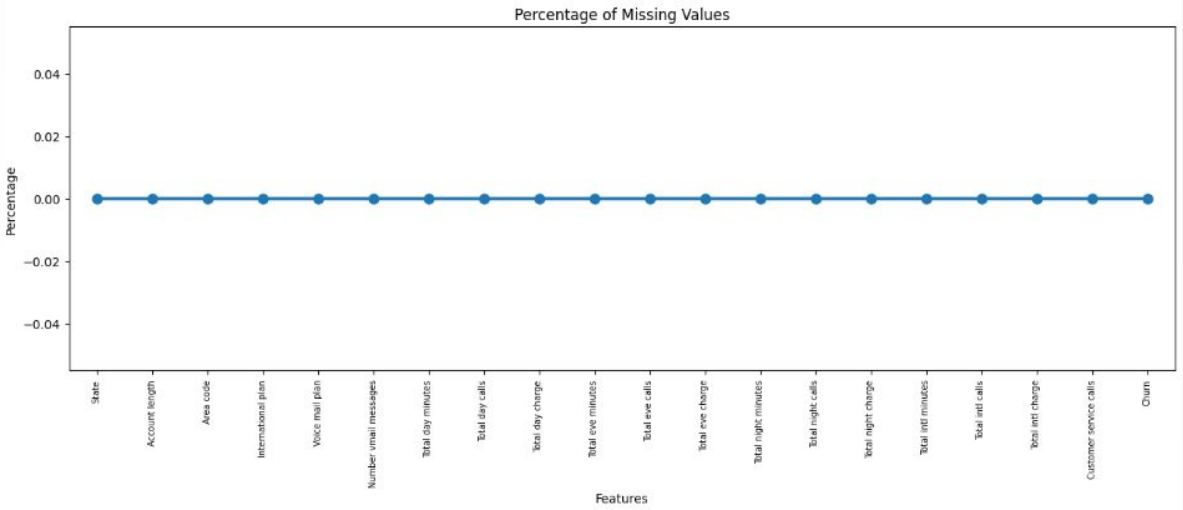
Numerical variables grouped by Churn



Categorical variables grouped by Churn

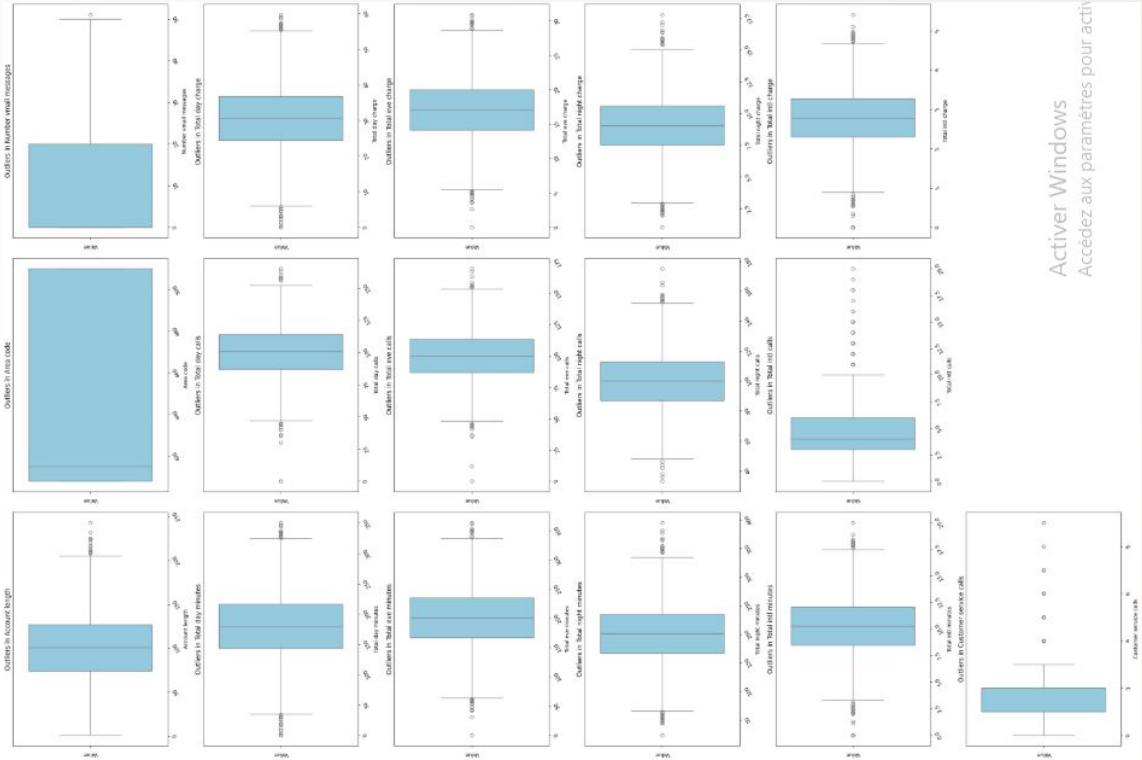


Missing Values:




```
Missing values in each columr
State                     0
Account length            0
Area code                 0
International plan         0
Voice mail plan            0
Number vmail messages      0
Total day minutes          0
Total day calls            0
Total day charge           0
Total eve minutes          0
Total eve calls            0
Total eve charge           0
Total night minutes        0
Total night calls          0
Total night charge         0
Total intl minutes         0
Total intl calls           0
Total intl charge          0
Customer service calls     0
Churn                      0
dtype: int64
```

Identify Outliers:



04

Data Preparation

Encode categorical variables

Convert the variables '**International plan**' and '**Voice mail plan**'

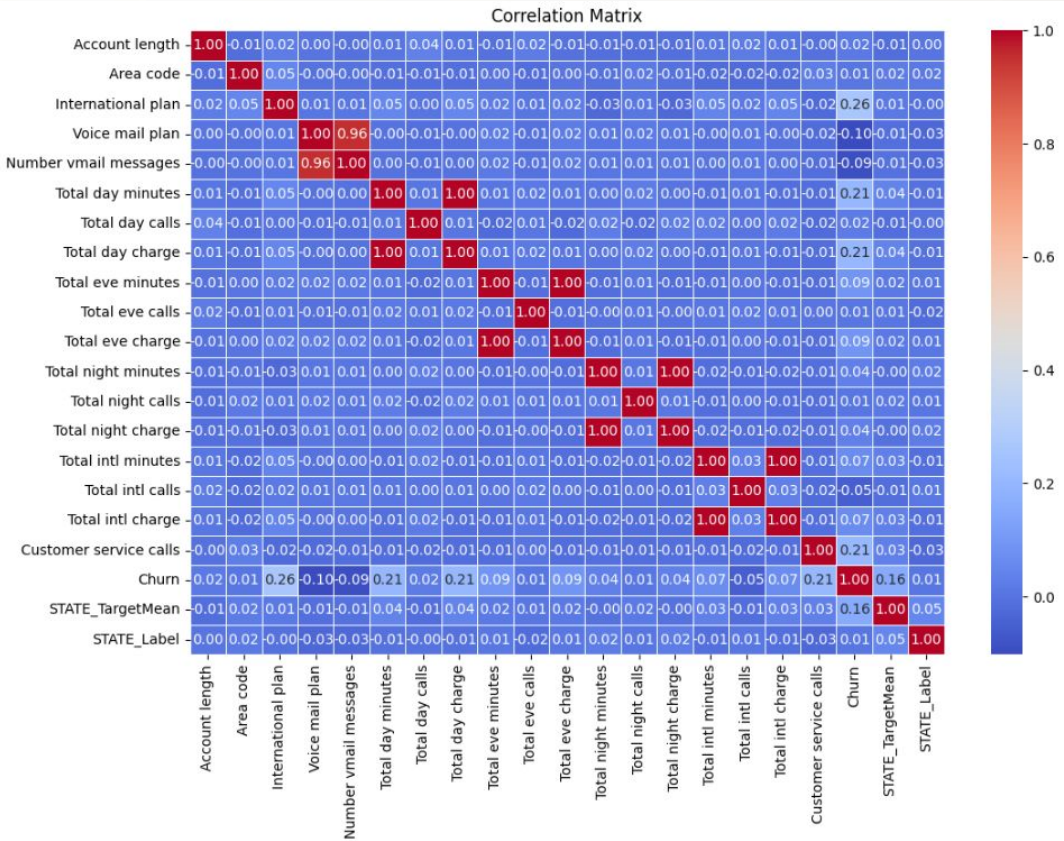
Convert the '**Churn**' column

Transform the '**State**' column using the **mean of the target**

Transform the '**State**' column using **LabelEncoder**

	STATE_TargetMean	STATE_Label	International plan	Voice mail plan	Churn
0	0.185714	16	0	1	0
1	0.128205	35	0	1	0
2	0.264706	31	0	0	0
3	0.128205	35	1	0	0
4	0.147541	36	1	0	0
...	...	...	...	...	...
3328	0.089744	48	0	1	0
3329	0.100000	1	0	1	0
3330	0.109589	46	0	0	0
3331	0.094340	49	0	0	0
3332	0.162162	6	1	0	0

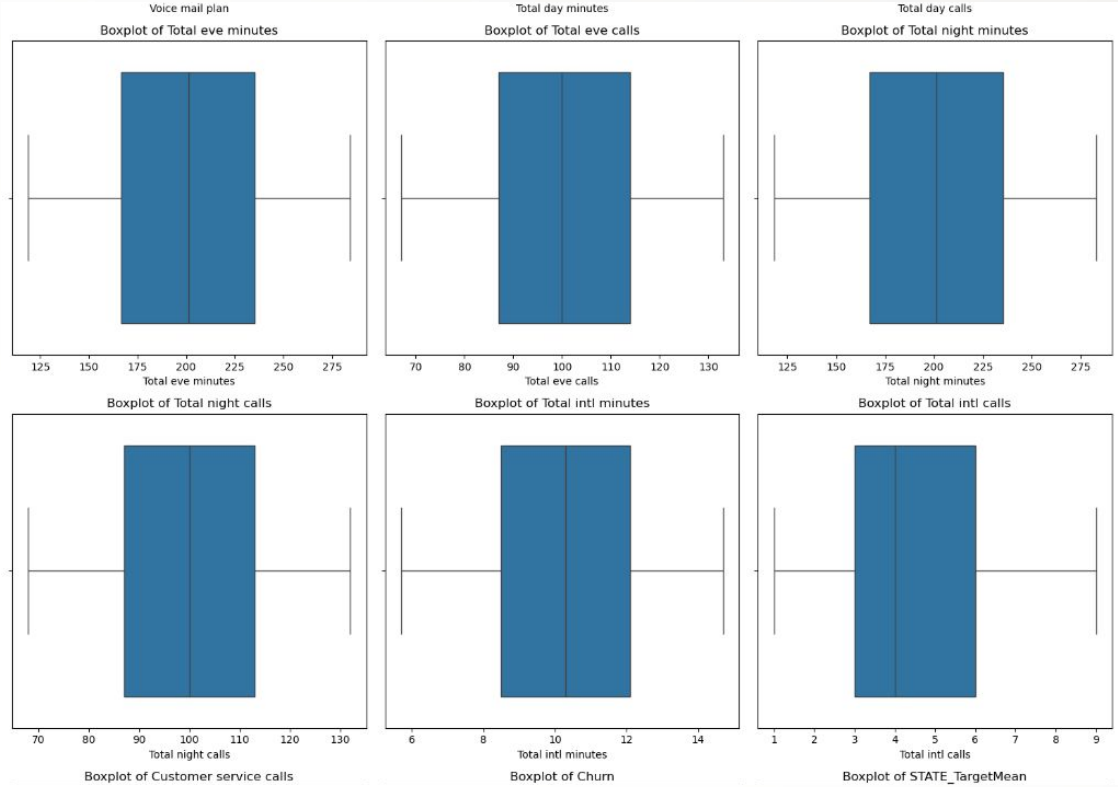
Correlation



Drop columns due to high correlation

- Number vmail messages
- Total day charge
- Total eve charge
- Total night charge
- Total intl charge

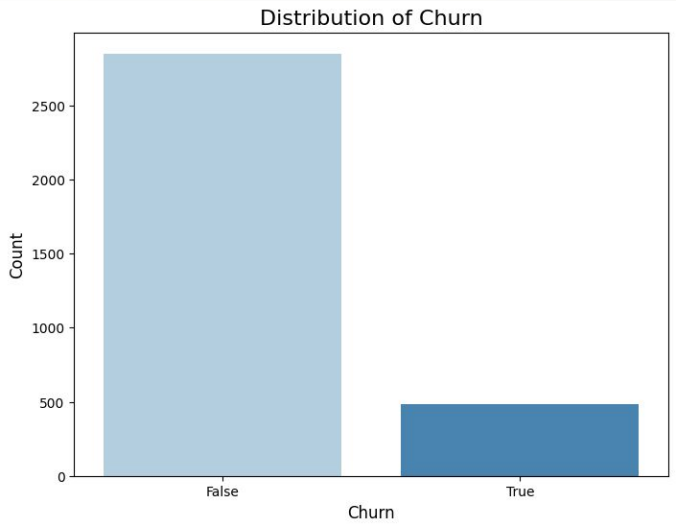
Outliers



Clips outliers in the dataset by setting **lower** and **upper** bounds at the **5th** and **95th** percentiles for each numerical column, ensuring that values outside these bounds are **adjusted to the nearest limit.**

Balance the dataset

SMOTE is applied to balance the dataset by generating synthetic samples for the minority class.



3333 rows

5700 rows

Feature selection

	Feature	F-Score	P-Value
3	Voice mail plan	519.909418	3.330141e-110
13	STATE_TargetMean	356.536078	3.488848e-77
4	Total day minutes	349.194428	1.117278e-75
11	Total intl calls	188.864456	2.644135e-42
12	Customer service calls	112.698771	4.396288e-26
6	Total eve minutes	82.381718	1.517495e-19
2	International plan	63.209535	2.223857e-15
10	Total intl minutes	53.126363	3.553589e-13
8	Total night minutes	20.050396	7.689955e-06
5	Total day calls	1.856419	1.730925e-01
0	Account length	0.909325	3.403340e-01
14	STATE_Label	0.334679	5.629400e-01
7	Total eve calls	0.181947	6.697212e-01
1	Area code	0.029312	8.640659e-01
9	Total night calls	0.009975	9.204470e-01

Using the Select-K-Best

- **Selecting the top 10** features based on their F-scores and P_values
-**Dropping less relevant** features .

5700 rows × 8 columns

Splitting and scaling the Dataset

Our data

(5700, 8)

Train

(4560, 8)

Test

(1140, 8)

```
X_train_scaled (après normalisation) :
[[-0.32638557 -0.47692374  0.57030337  0.14137783  1.05365847 -1.57168613
  -0.53956438 -0.47278106]
 [-0.32638557 -0.47692374 -0.36392783 -1.20805398 -0.13010499 -2.04862695
  -0.04341323  0.32040571]
 [-0.32638557 -0.47692374  1.48163319 -0.53333808  0.33794736 -0.4010132
   2.43734256  0.32040571]
 [-0.32638557 -0.47692374  1.46546927  0.3749919   0.84212963  1.183525
   1.44504025  0.32040571]
 [-0.32638557 -0.47692374  0.28578958 -0.15905392  0.76772272  1.08648116
   0.45273793 -0.47278106]]

X_test_scaled (après normalisation) :
[[-0.32638557 -0.47692374  0.16680273  0.26926515 -1.91870112  1.15988405
  -0.04341323 -1.26596782]
 [-0.32638557 -0.47692374  0.67208732 -1.41972956  0.14254687 -1.74511915
  -1.03571554 -1.26596782]
 [ 3.06386092 -0.47692374 -0.84393009 -1.76234338  1.02086672 -0.5271213
  -0.53956438  1.11359247]
 [-0.32638557  2.09677126 -0.67109724  1.38497184  1.46036414  0.37943543
   0.45273793  1.90677923]
 [-0.32638557 -0.47692374 -1.55807161 -1.89379465 -0.03694894  0.68294322
  -0.53956438  1.90677923]]
```

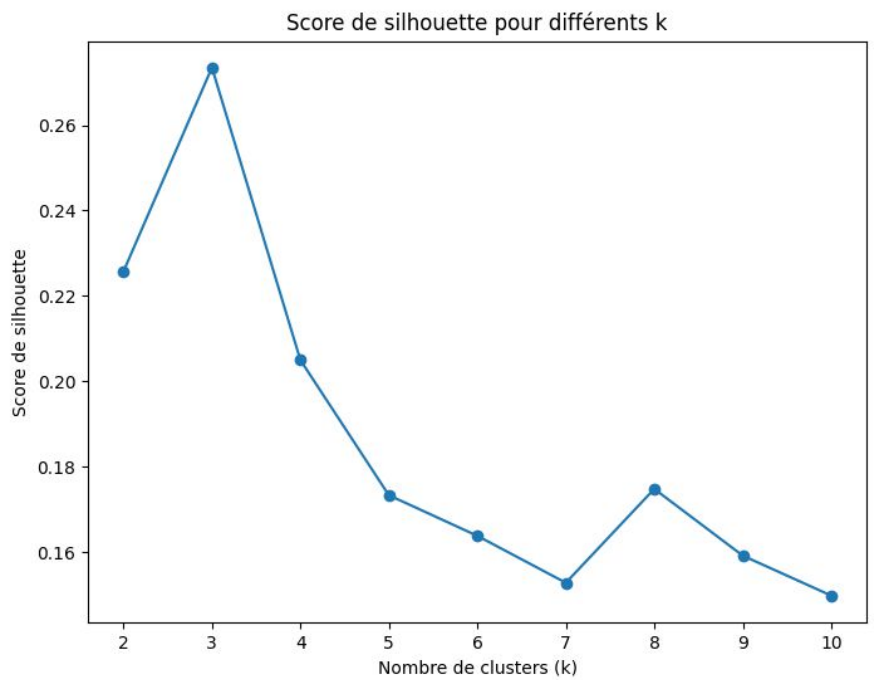
05

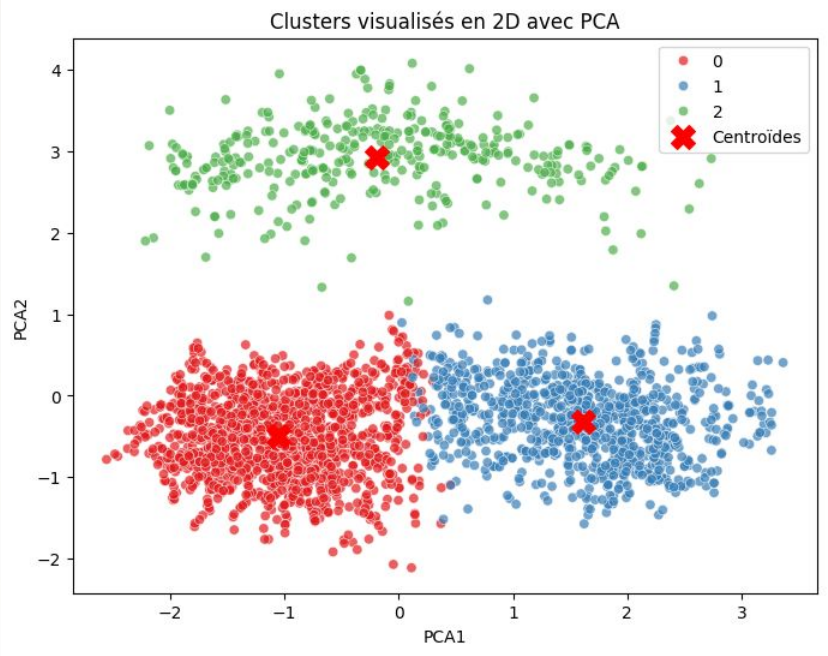
Modeling

Paper 1

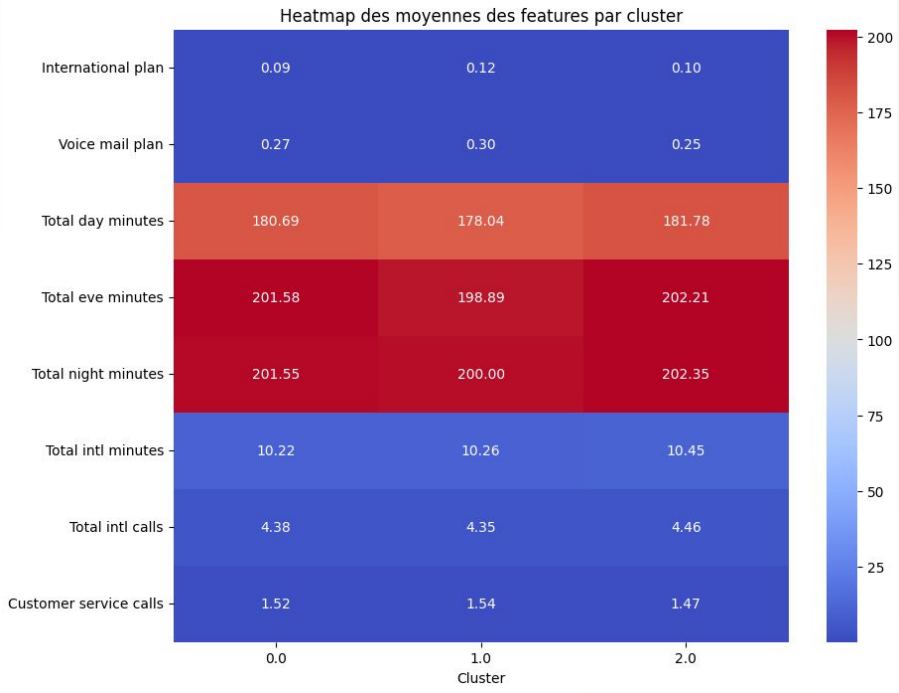
Implementing machine learning techniques for customer retention and churn prediction in telecommunications

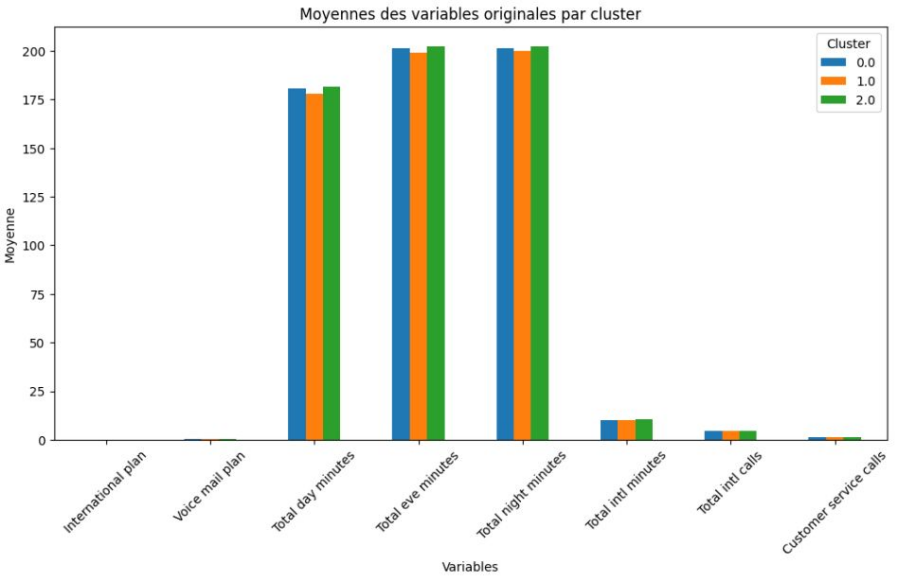
1. Unsupervised Learning - Dimensionality Reduction (PCA)



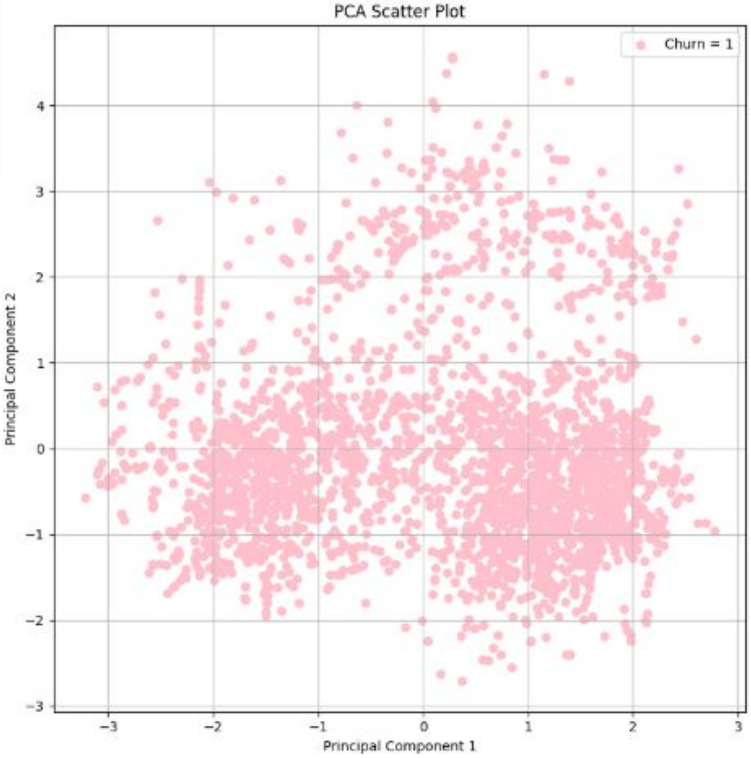


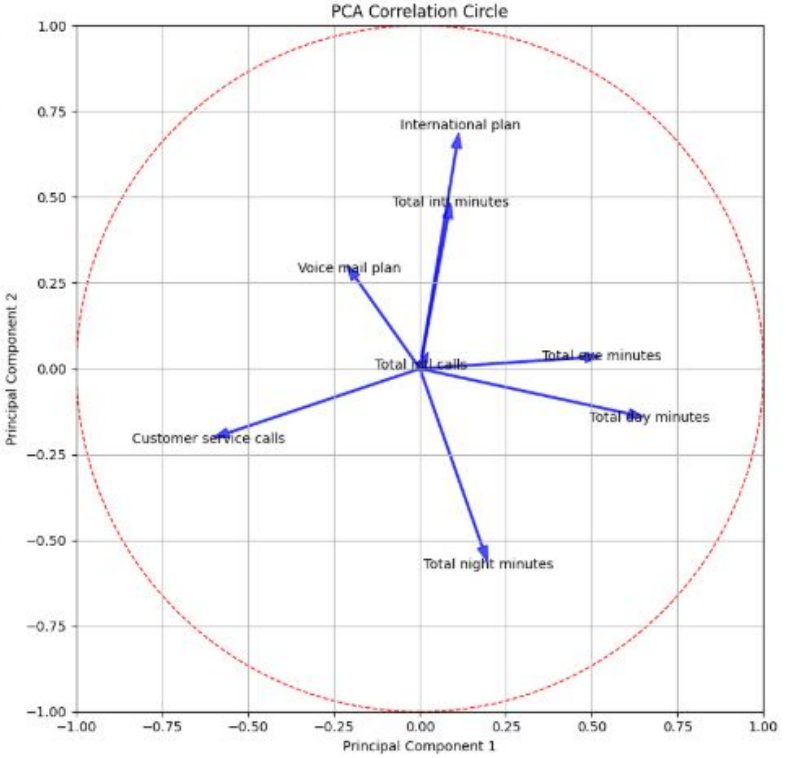
1. Unsupervised Learning - Dimensionality Reduction (PCA)





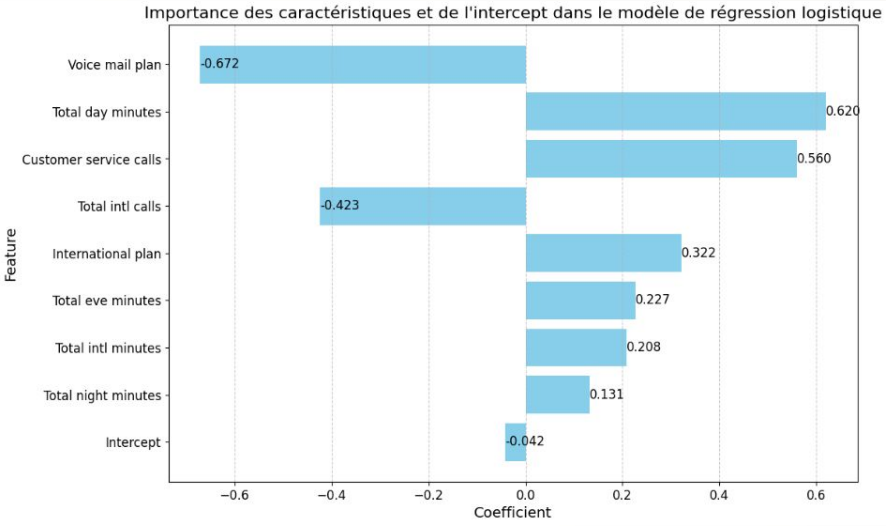
1. Unsupervised Learning - Dimensionality Reduction (PCA)





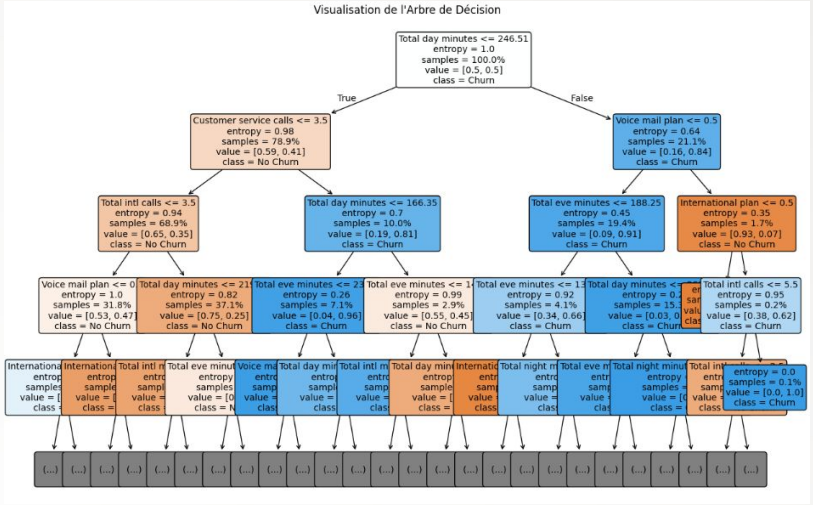
2. Supervised Learning

Logistic Regression



"Customers with **frequent service calls**, **extensive usage**, and **international plans** have a significantly higher risk of churn."

Decision Tree Classifier



"The decision tree model highlights **'Total day minutes'** as the **most influential** feature, followed by **'Total eve minutes'** and **'Total night minutes'**.

2. Supervised Learning

Random Forest

Meilleurs hyperparamètres :

{'bootstrap': True,
'max_depth': None,
'min_samples_leaf': 1,
'min_samples_split': 2,
'n_estimators': 100}

SVM

Précisions des kernels :

	Kernel	Accuracy
2	rbf	0.835746
1	poly	0.819298
0	linear	0.722149
3	sigmoid	0.552851

The **RBF** kernel achieved the **highest accuracy** in classifying the data, followed by the polynomial kernel.

Gradient Boosting

The Gradient Boosting model was optimized using **RandomizedSearchCV**, balancing computational efficiency and model performance.

Paper 2

Explaining customer churn prediction in telecom industry using tabular machine learning models

Using the parameters specified in the paper2

- **Model 1 : Logistic Regression**

LogisticRegression(C=4534347.358, max_iter=10000, random_state=42)

- **Model 2 : Random Forest**

RandomForestClassifier(max_depth=18, n_estimators=20, random_state=42)

- **Model 3 : SVM**

SVC(kernel='linear', probability=True, random_state=42)

- **Model 4 : GBM**

GradientBoostingClassifier(max_depth=10, max_features='sqrt', max_leaf_nodes=5, min_samples_leaf=7, n_estimators=150, random_state=42, subsample=0.9)

Adding advanced models

Model 5 :AdaBoost

Default algorithm

With SAMME Algorithm

Model 6 :XGBoost

Choosing values from the specified range by Article

Changing ranges

Model 7 : Neural Networks

With 5-9 Hidden Layers

With 2 Hidden Layers

Paper 3

Customer churn prediction in telecom sector using machine learning techniques

Edited Nearest Neighbours (ENN) Cleaning:

Improves data quality by removing noisy samples.

```
After up-sampling: Counter({0: 2285, 1: 2285})
After ENN cleaning: Counter({0: 2285, 1: 1378})
After ENN cleaning: Counter({0: 575, 1: 283})
```

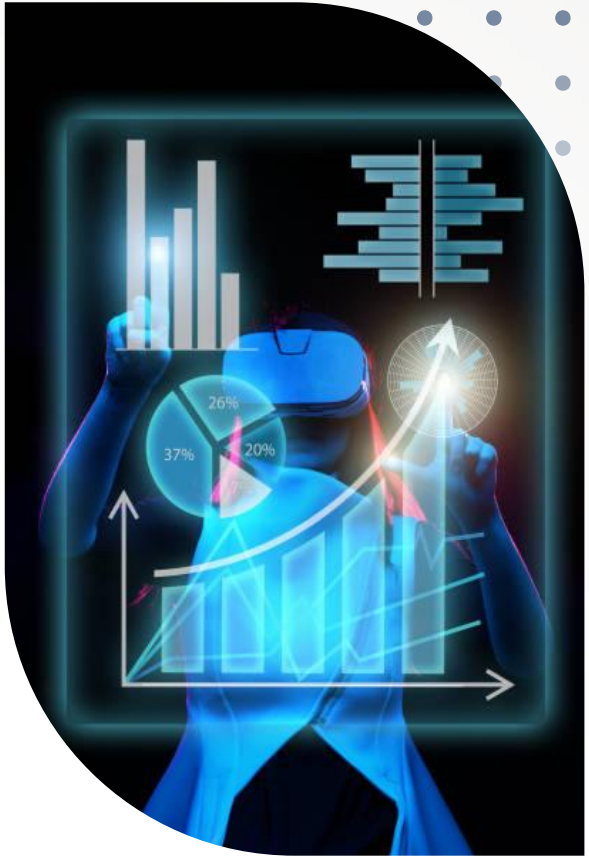
Decision Tree Classifier

Random Forest Classifier

Cox Proportional Hazard Model (Survival Analysis):

Analyzes how features impact the time until churn occurs, giving insights into customer retention and churn risks.

duration_col="Account length", event_col="Churn"



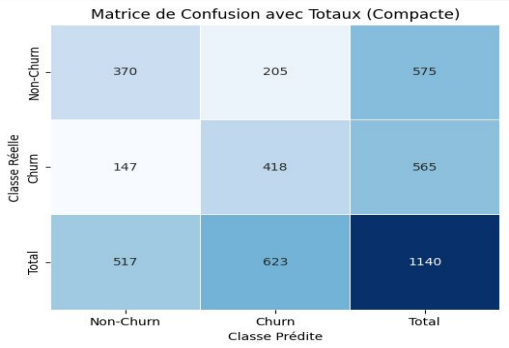
06

Evaluation

Paper 1

Implementing machine learning techniques for customer retention and churn prediction in telecommunications

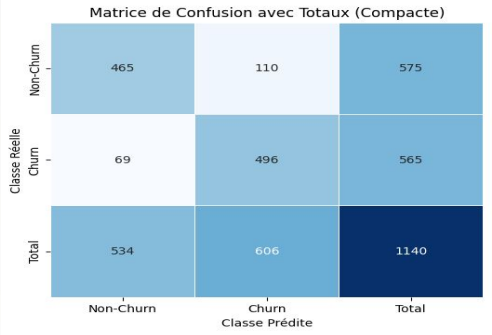
Model 1 : Logistic Regression



Accuracy : 0.69 , AUC : 0.77

	precision	recall	f1-score	support
0	0.72	0.64	0.68	575.00
1	0.67	0.74	0.70	565.00
accuracy	0.69	0.69	0.69	0.69
macro avg	0.69	0.69	0.69	1140.00
weighted avg	0.69	0.69	0.69	1140.00

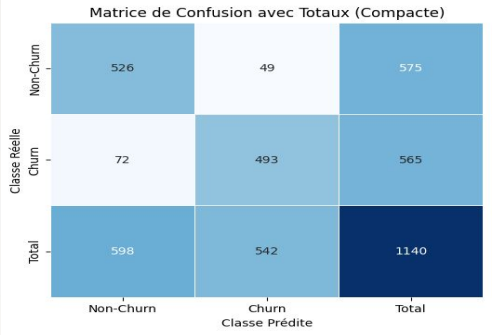
Model 2 : Decision Tree



Accuracy : 0.69 , AUC : 0.77

	precision	recall	f1-score	support
0	0.87	0.81	0.84	575.00
1	0.82	0.88	0.85	565.00
accuracy	0.84	0.84	0.84	0.84
macro avg	0.84	0.84	0.84	1140.00
weighted avg	0.84	0.84	0.84	1140.00

Model 3 : Random Forest

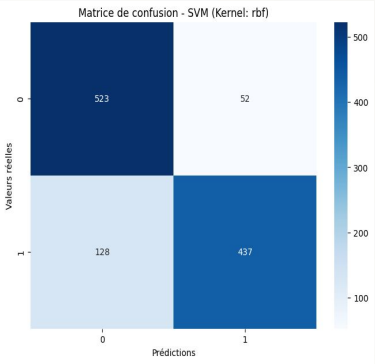


Accuracy : 0.89 , AUC : 0.96

	precision	recall	f1-score	support
0	0.88	0.91	0.90	575.00
1	0.91	0.87	0.89	565.00
accuracy	0.89	0.89	0.89	0.89
macro avg	0.89	0.89	0.89	1140.00
weighted avg	0.89	0.89	0.89	1140.00

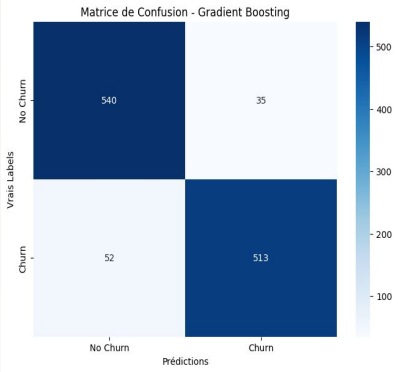
Model 4 : SVM

AUC : 0.91
Accuracy : 0.84



	precision	recall	f1-score	support
0	0.80	0.91	0.85	575.00
1	0.89	0.77	0.83	565.00
accuracy	0.84	0.84	0.84	0.84
macro avg	0.85	0.84	0.84	1140.00
weighted avg	0.85	0.84	0.84	1140.00

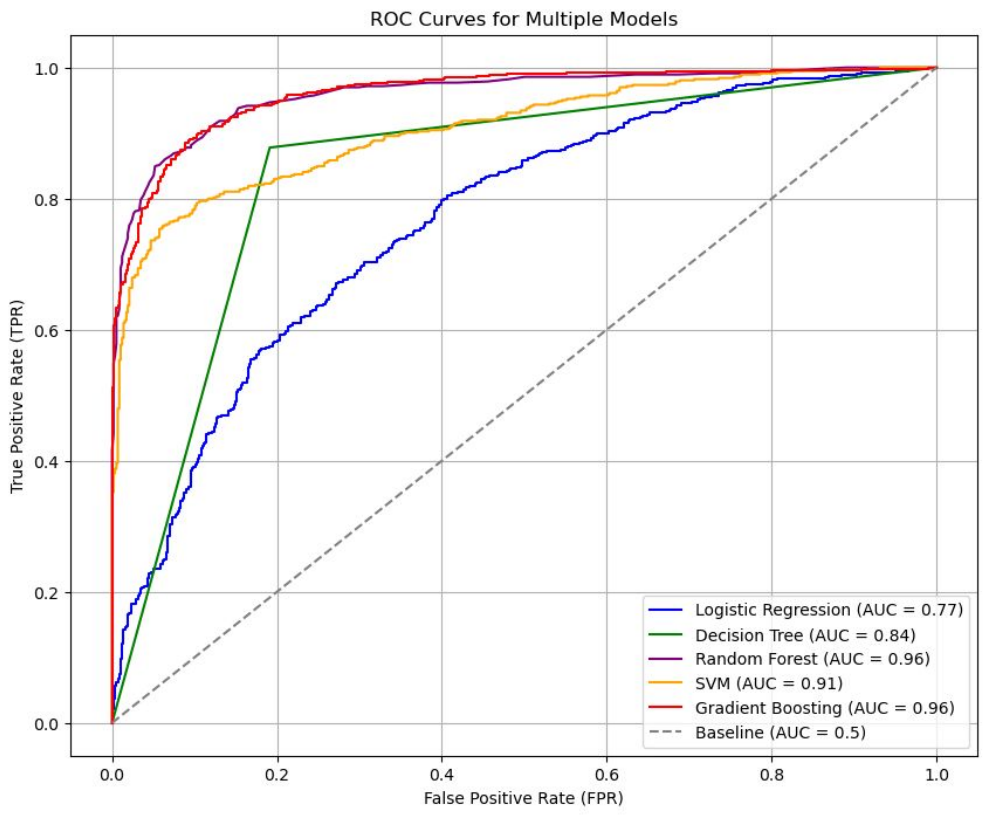
Model 5 : Gradient Boosting



AUC : 0.97
Accuracy : 0.91

	precision	recall	f1-score	support
0	0.91	0.92	0.91	575.00
1	0.92	0.90	0.91	565.00
accuracy	0.91	0.91	0.91	0.91
macro avg	0.91	0.91	0.91	1140.00
weighted avg	0.91	0.91	0.91	1140.00

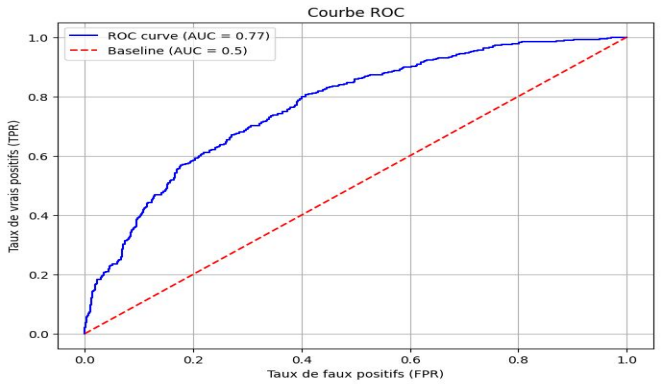
All ROC Curves



Paper 2

Explaining customer churn prediction in telecom industry using tabular machine learning models

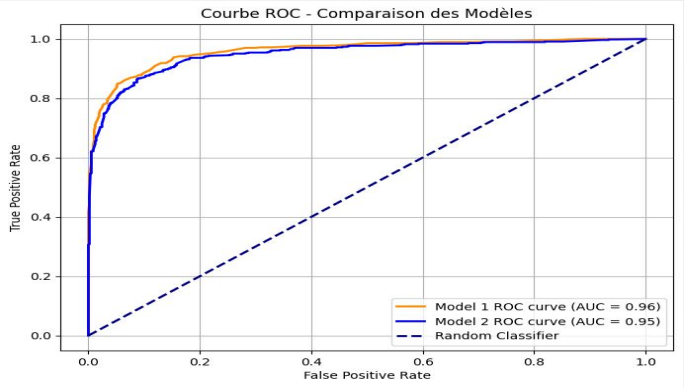
Model 1 : Logistic Regression



	precision	recall	f1-score	support
0	0.72	0.64	0.68	575.00
1	0.67	0.74	0.70	565.00
accuracy	0.69	0.69	0.69	0.69
macro avg	0.69	0.69	0.69	1140.00
weighted avg	0.69	0.69	0.69	1140.00

Same results for both models

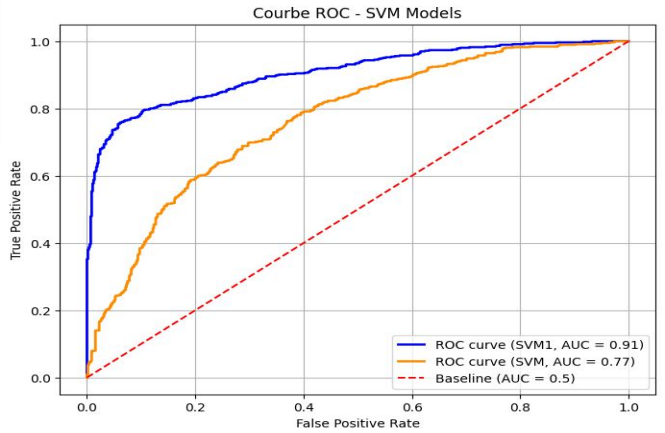
Model 2 : Random Forest



Classification Report of the paper's model

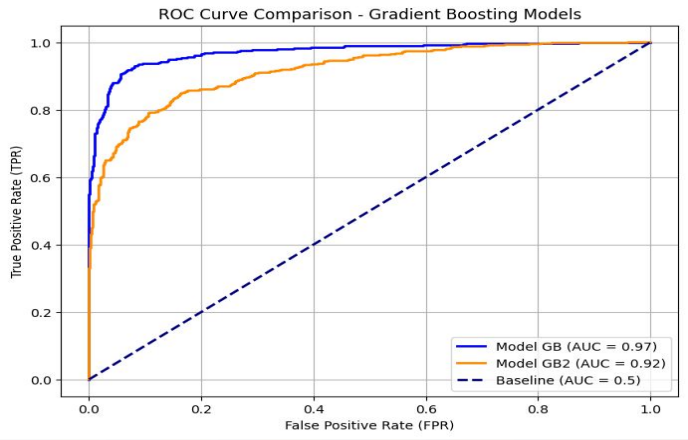
	precision	recall	f1-score	support
0	0.88	0.90	0.89	575.00
1	0.90	0.87	0.88	565.00
accuracy	0.89	0.89	0.89	0.89
macro avg	0.89	0.89	0.89	1140.00
weighted avg	0.89	0.89	0.89	1140.00

Model 3 : SVM



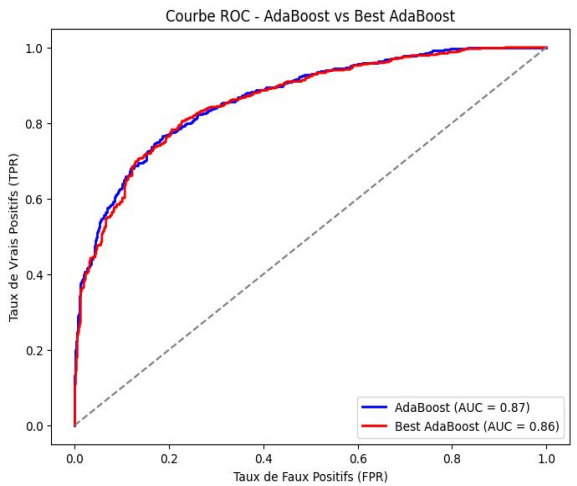
	precision	recall	f1-score	support
0	0.73	0.62	0.67	575.00
1	0.67	0.76	0.71	565.00
accuracy	0.69	0.69	0.69	0.69
macro avg	0.70	0.69	0.69	1140.00
weighted avg	0.70	0.69	0.69	1140.00

Model 4 : GBM



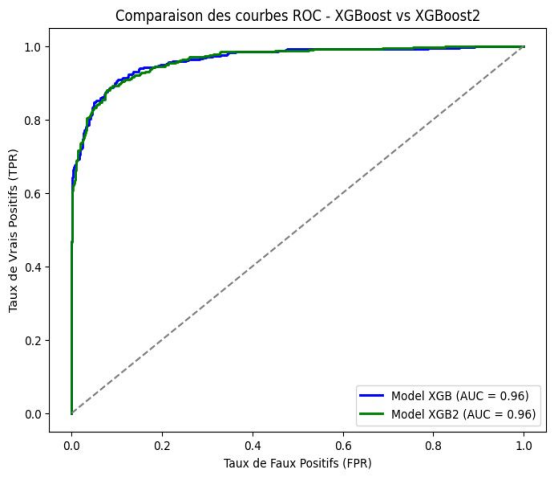
	precision	recall	f1-score	support
0	0.50	1.00	0.67	575.00
1	0.00	0.00	0.00	565.00
accuracy	0.50	0.50	0.50	0.50
macro avg	0.25	0.50	0.34	1140.00
weighted avg	0.25	0.50	0.34	1140.00

Model 5 : ADABOOST



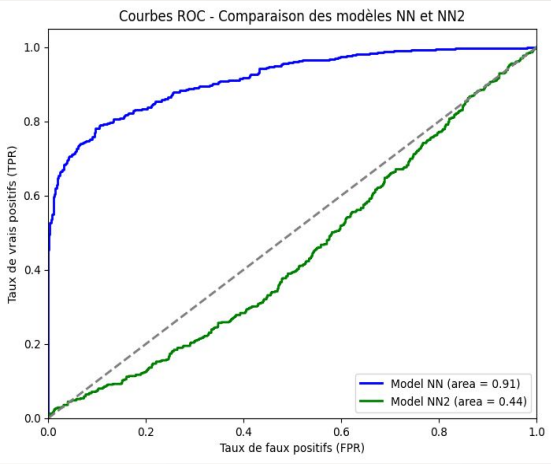
Accuracy1 : 0.79
Accuracy2 : 0.78

Model 6 : XGBOOST



Accuracy1 : 0.91
Accuracy2 : 0.91

Model 7 : Neural Networks



Accuracy1 : 0.52
Accuracy2 : 0.84

Wilcoxon signed-rank test :

- **Gbm and Random Forest:**

Statistic: 1577.0, P-value: 0.5075937168646562

- **Gbm and Logistic Regression:**

Statistic: 19520.0, P-value: 2.6786695104862295e-05

- **Gbm and Decision Tree:**

Statistic: 4509.0, P-value: 6.7420639326455215e-06

- **Gbm and SVM:**

Statistic: 17871.0, P-value: 3.282844941284038e-08

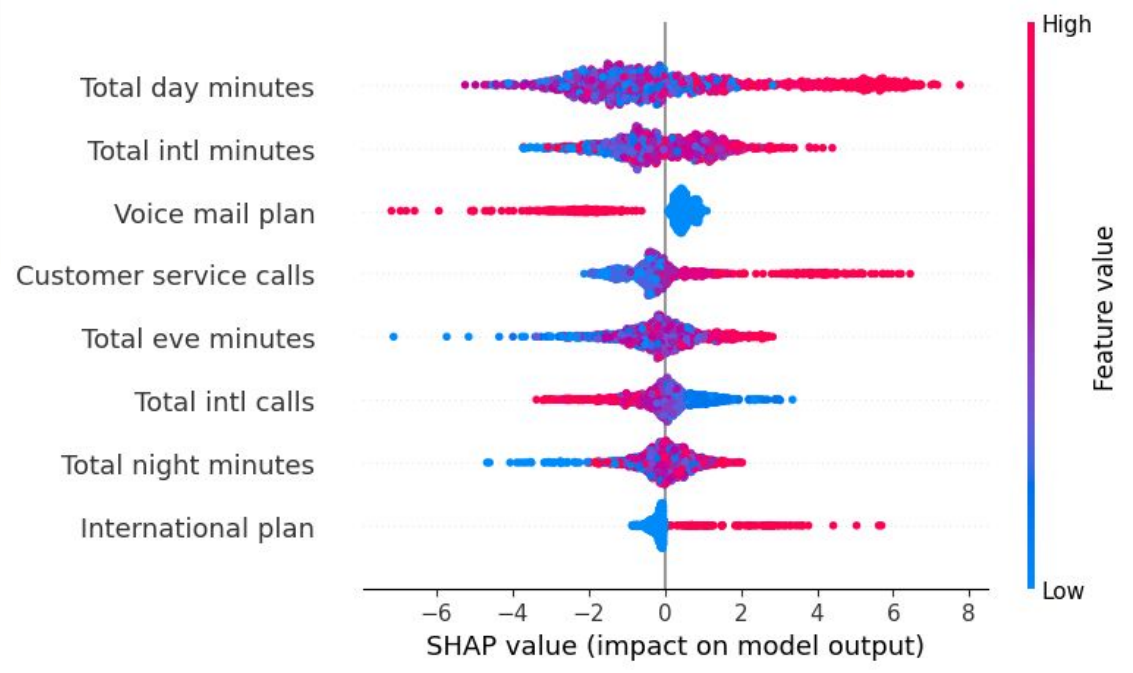
- **Gbm and ADABOOST :**

Statistic: 10543.5, P-value: 0.33628879040286896

- **Gbm and XGBOOST :**

Statistic: 854.0, P-value: 0.6055766163353462

SHAP and Model Transparency



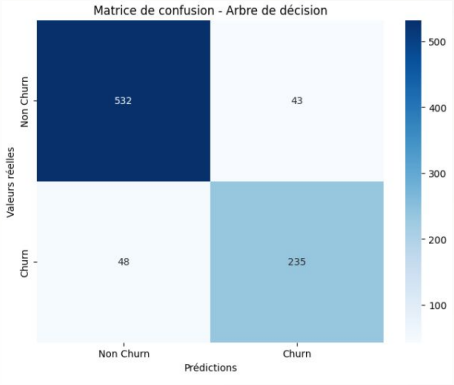
Total day minutes have the strongest positive impact on predictions, as higher feature values (in pink) push the model output toward predicting churn, whereas lower values (in blue) have less impact.

Paper 3

Customer churn prediction in telecom sector using machine learning techniques

Model : Decision Tree

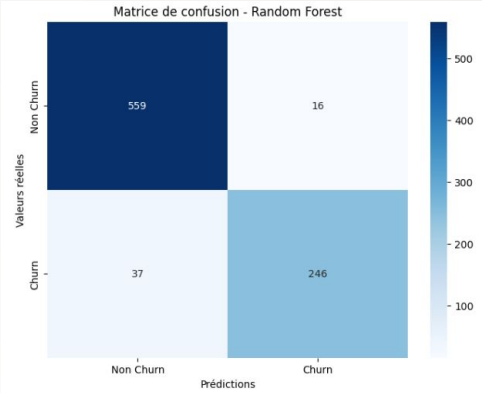
Training Accuracy: 1.00
Test Accuracy: 0.84




```
Arbre de décision Accuracy: 0.8939393939393939
              precision    recall  f1-score   support

           0       0.92      0.93      0.92       575
           1       0.85      0.83      0.84       283

    accuracy                           0.89       858
   macro avg       0.88      0.88      0.88       858
weighted avg       0.89      0.89      0.89       858
```

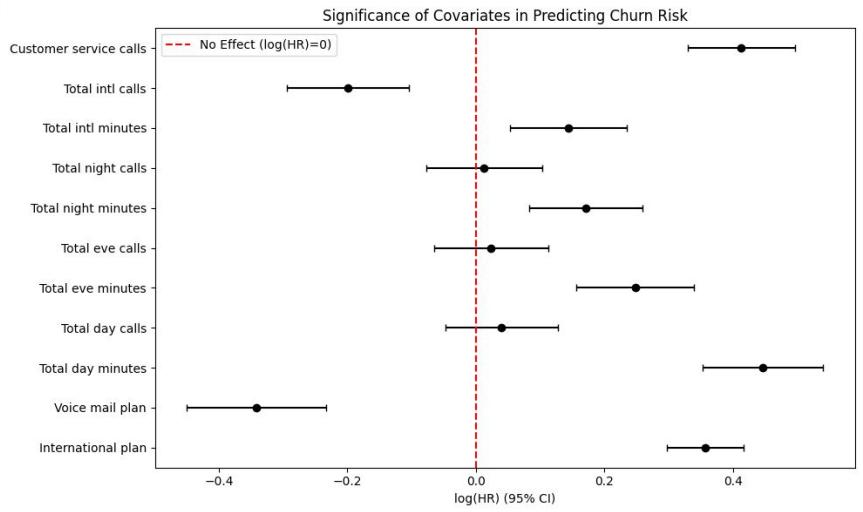
Model : Random Forest

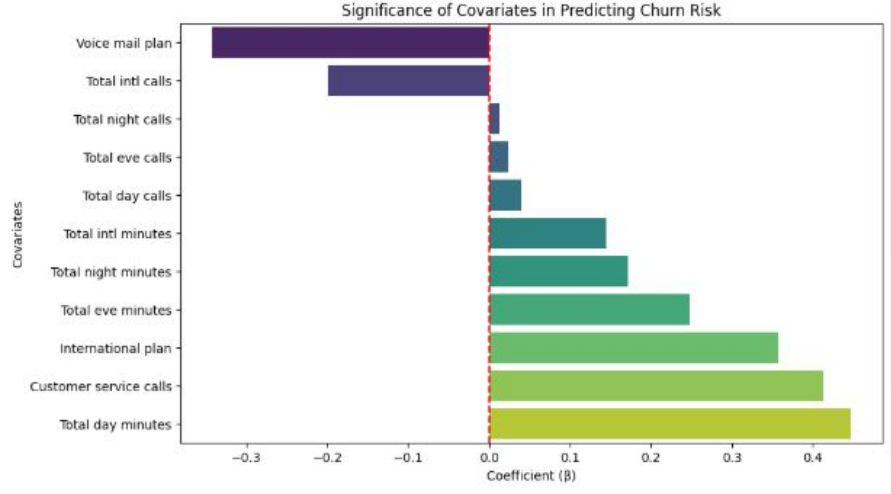
Training Accuracy: 1.00
Test Accuracy: 0.89



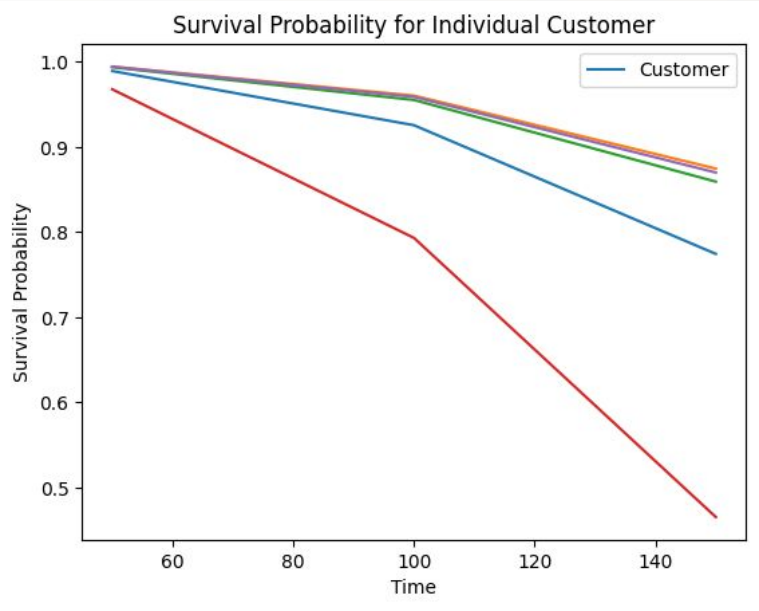

```
Random Forest Accuracy: 0.9382284382284383
              precision    recall  f1-score   support

           0       0.94      0.97      0.95       575
           1       0.94      0.87      0.90       283

    accuracy                           0.94       858
   macro avg       0.94      0.92      0.93       858
weighted avg       0.94      0.94      0.94       858
```

Cox proportional hazard model:





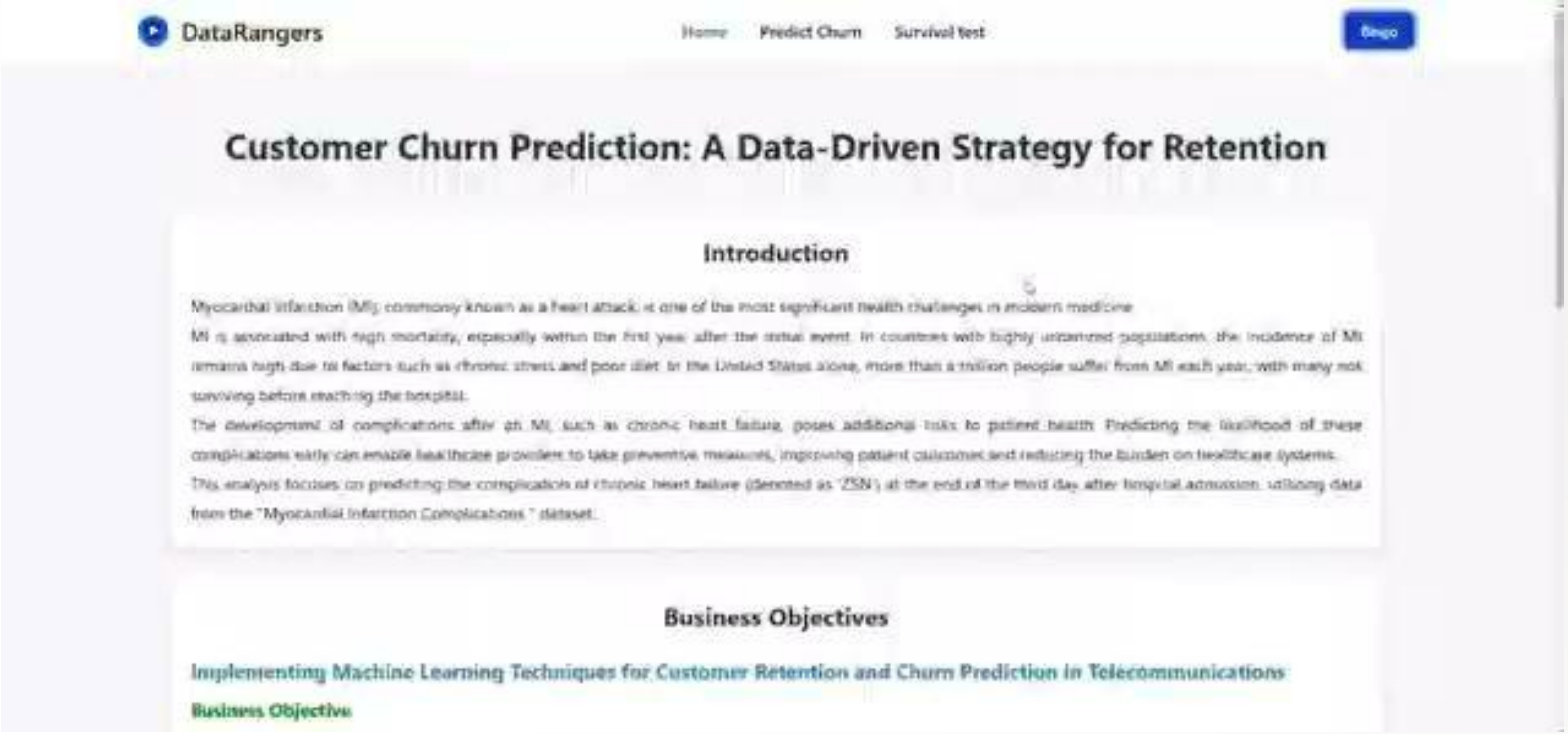
Survival Test Analysis

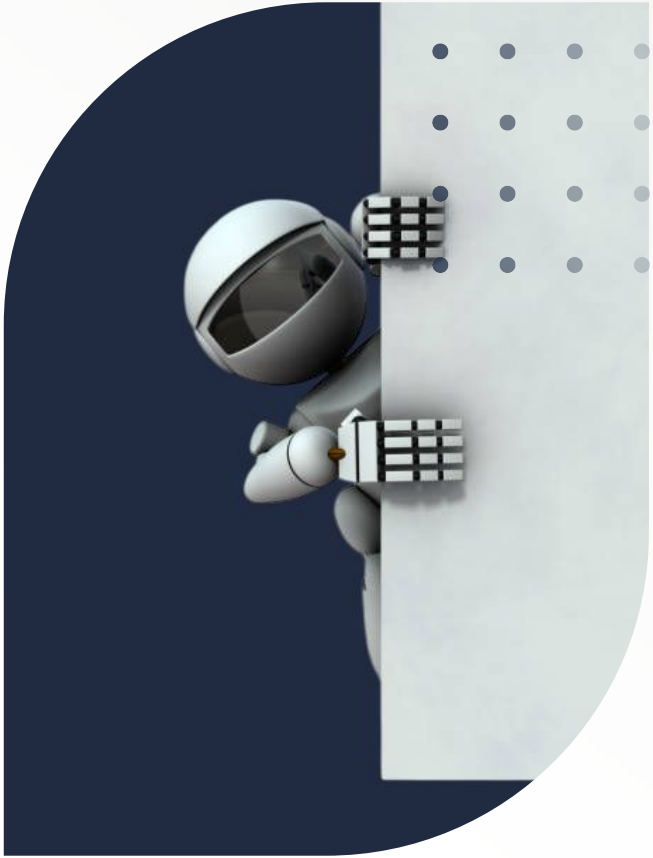


07

Deployment

Deployment :


DataRangers
Home
Predict Churn
Survival test
Customer Churn Prediction: A Data-Driven Strategy for Retention
Introduction
Business Objectives
Implementing Machine Learning Techniques for Customer Retention and Churn Prediction in Telecommunications
Business Objective



Thank you for your attention !