Projet d'économétrie

SOMMAIRE

Data Preprocessing 1

Exploratory analysis 2

KNN model 3

Bagging with decision tree 4

Random Forest 5

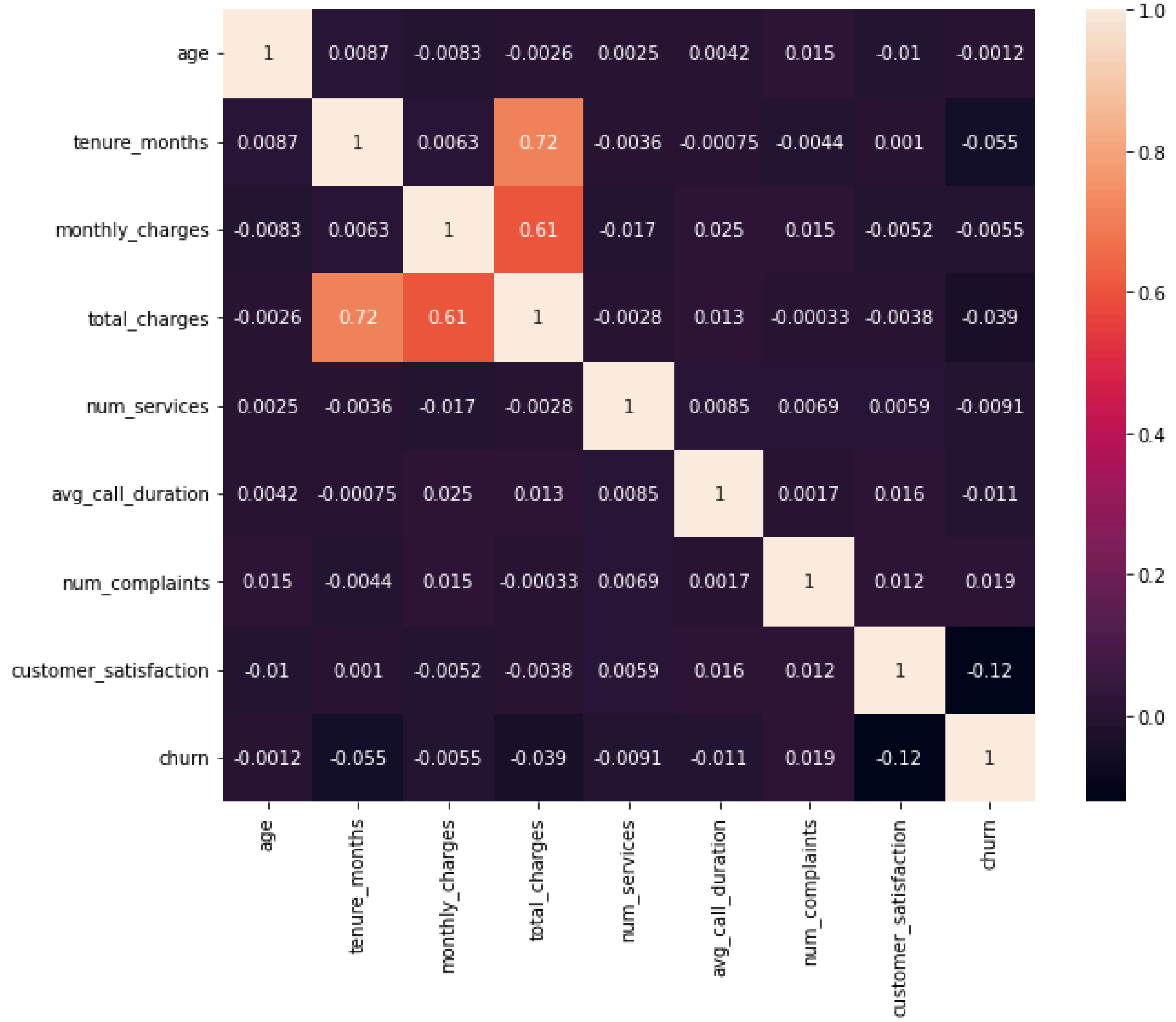
XGBClassifier 6

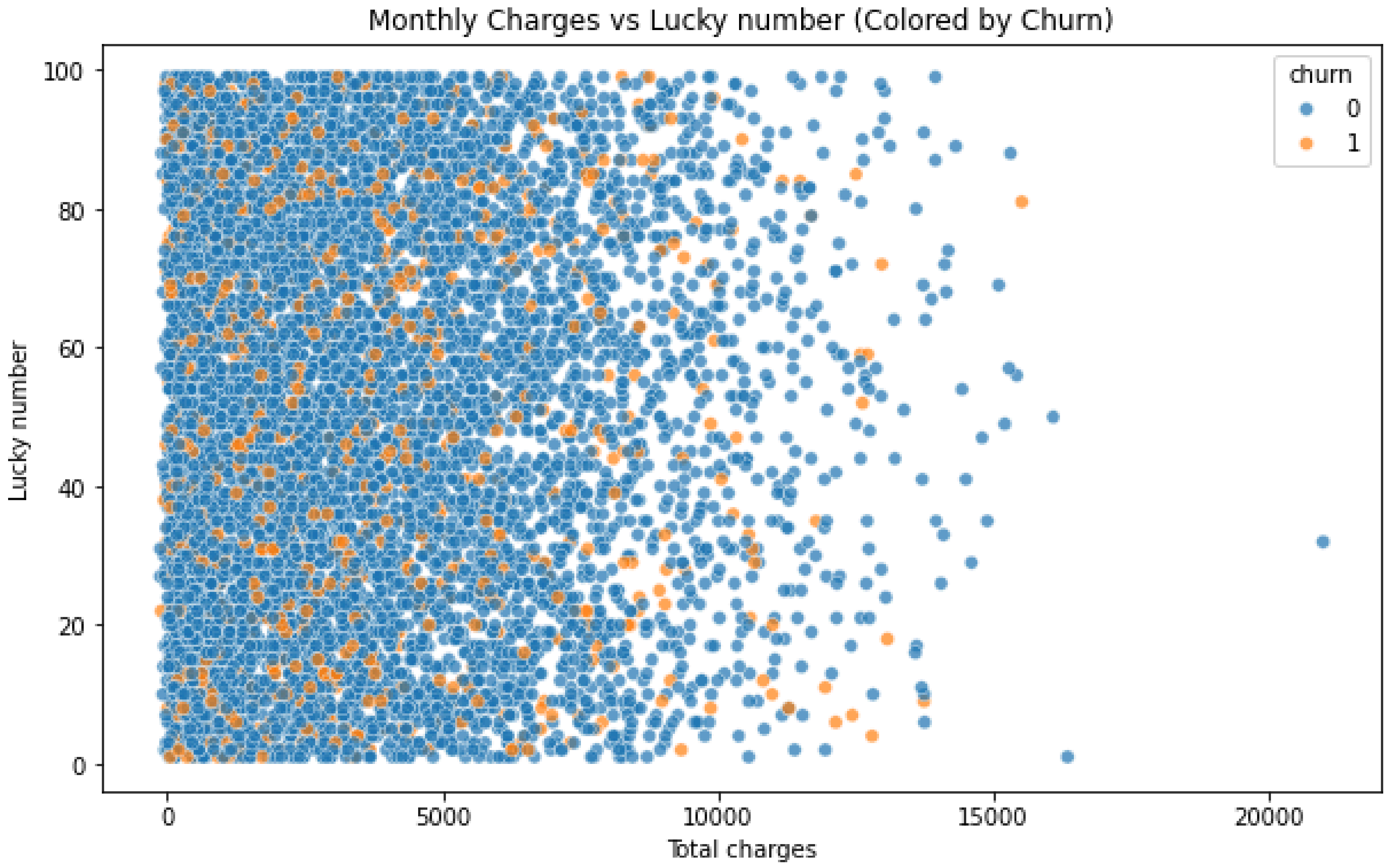
1. Data Preprocessing

- Removal of customer_id and random_string
  - We kept lucky_number
- Dealing with missing values
  - Only present in the internet_service variable

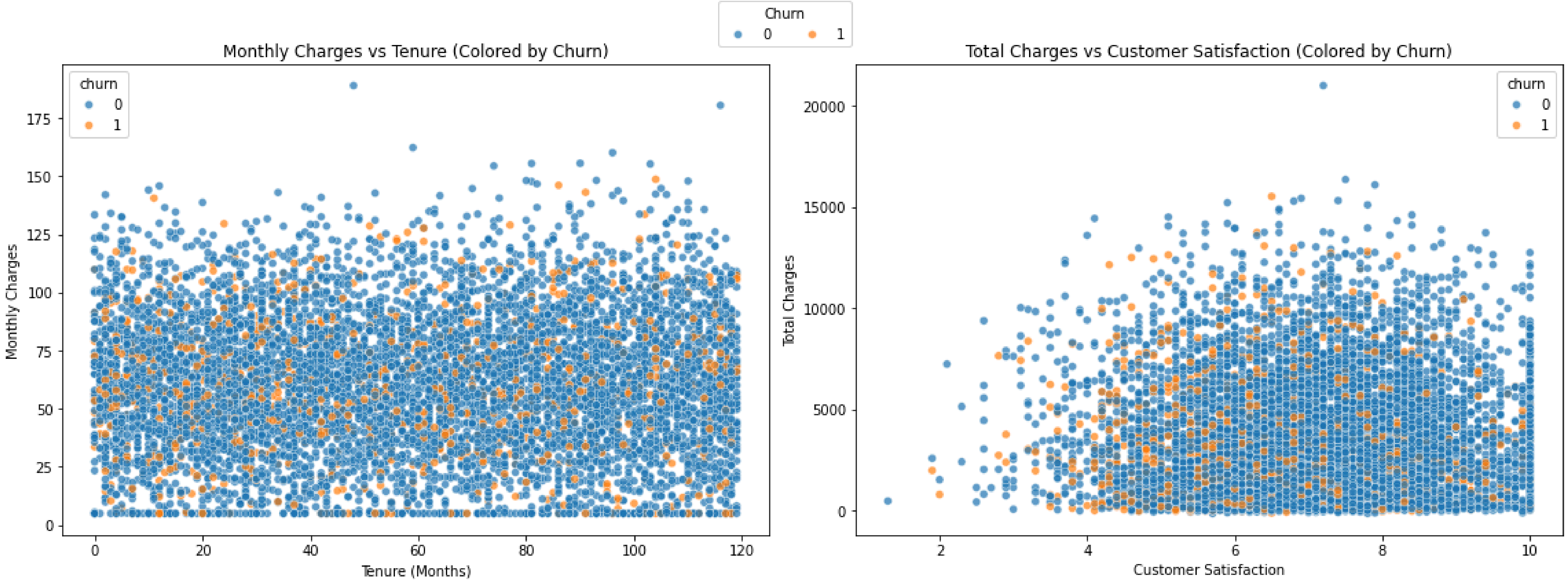
    There were 1165 rows with Na value
- Turning categorical variables into numerical ones

2. Exploratory Analysis (Continuous Variable)

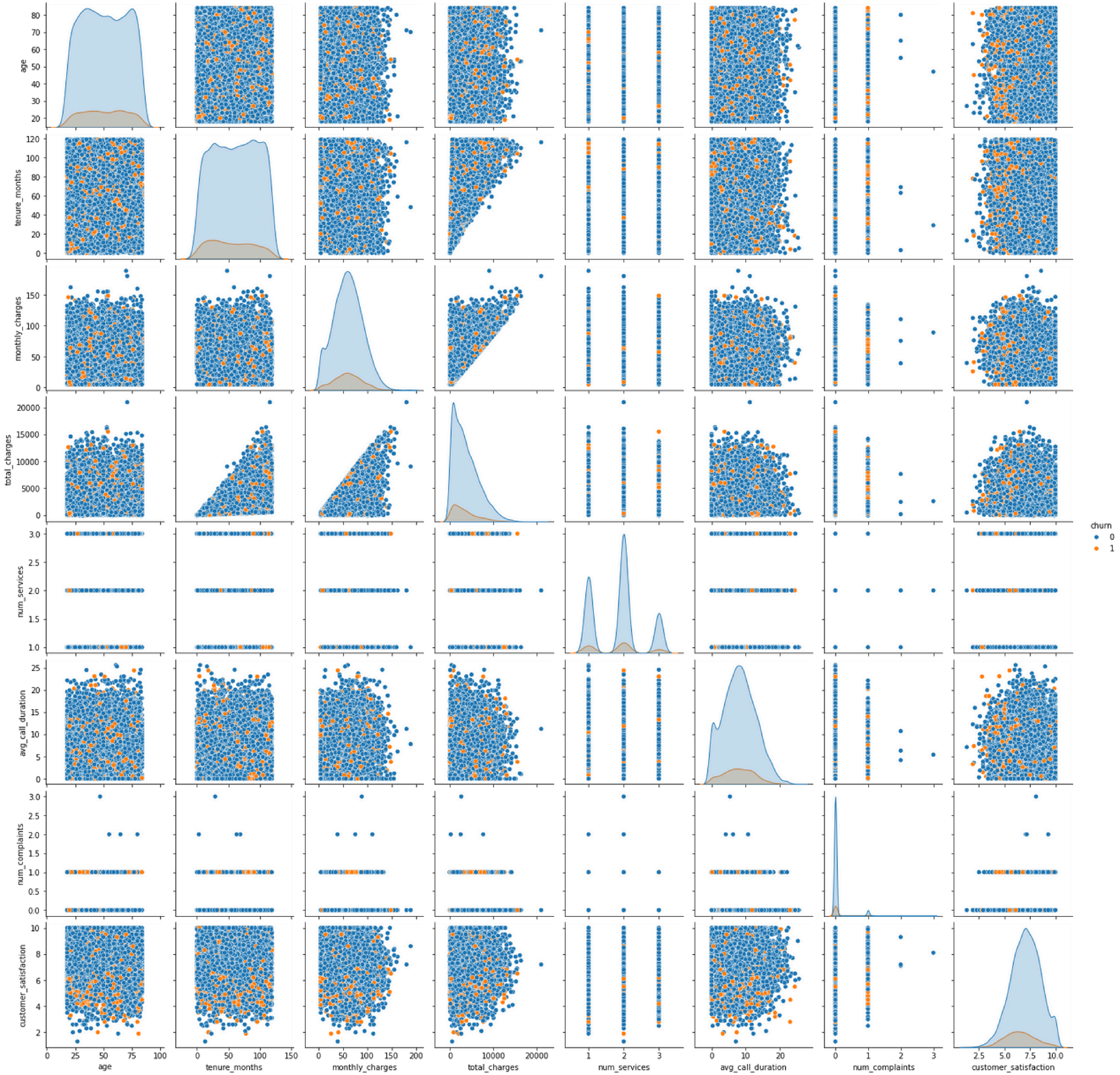




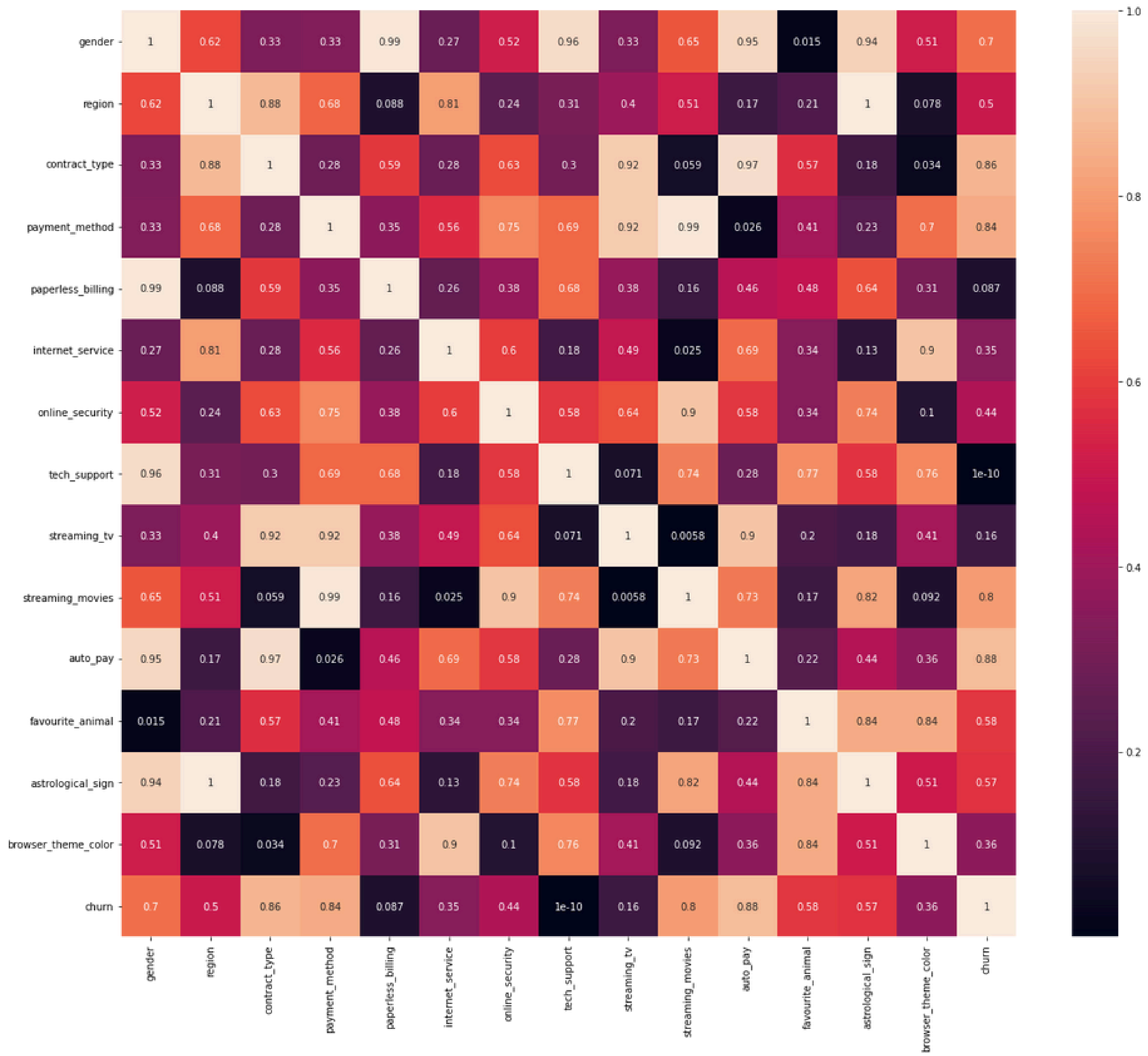
2. Exploratory Analysis (Continuous Variable)


Monthly Charges vs Tenure (Colored by Churn)

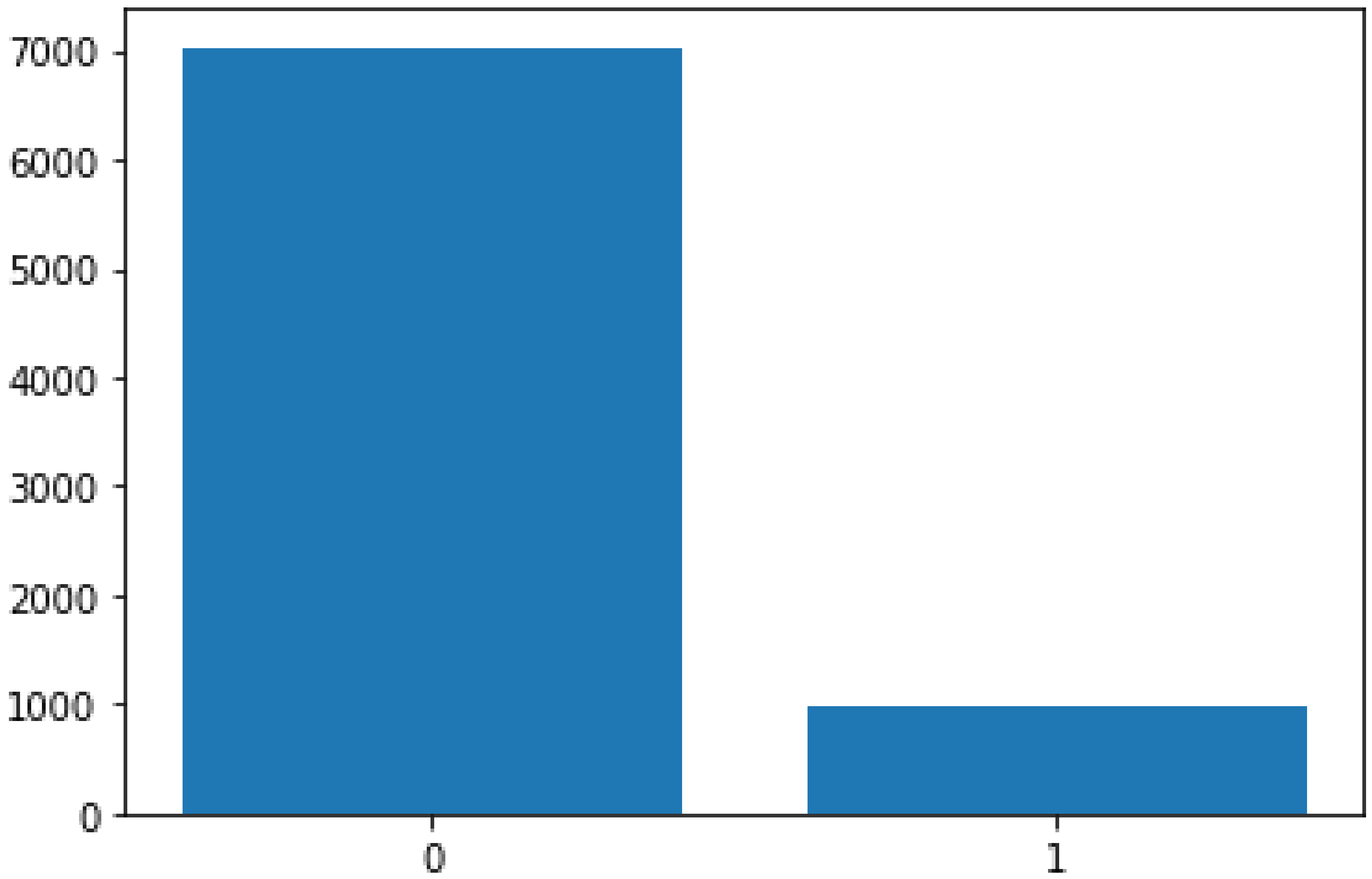
Total Charges vs Customer Satisfaction (Colored by Churn)


age
tenure_months
monthly_charges
total_charges
num_services
avg_call_duration
num_complaints
customer_satisfaction
churn
0
1
0
25
50
75
100
0
50
100
150
0
50
100
150
200
0
10000
20000
1
2
3
0
10
20
0
1
2
3
2.5
5.0
7.5
10.0
20
30
40
50
60
70
80
0
20
40
60
80
100
120
0
50
100
150
0
5000
10000
15000
20000
1.0
1.5
2.0
2.5
3.0
0
5
10
15
20
25
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
10

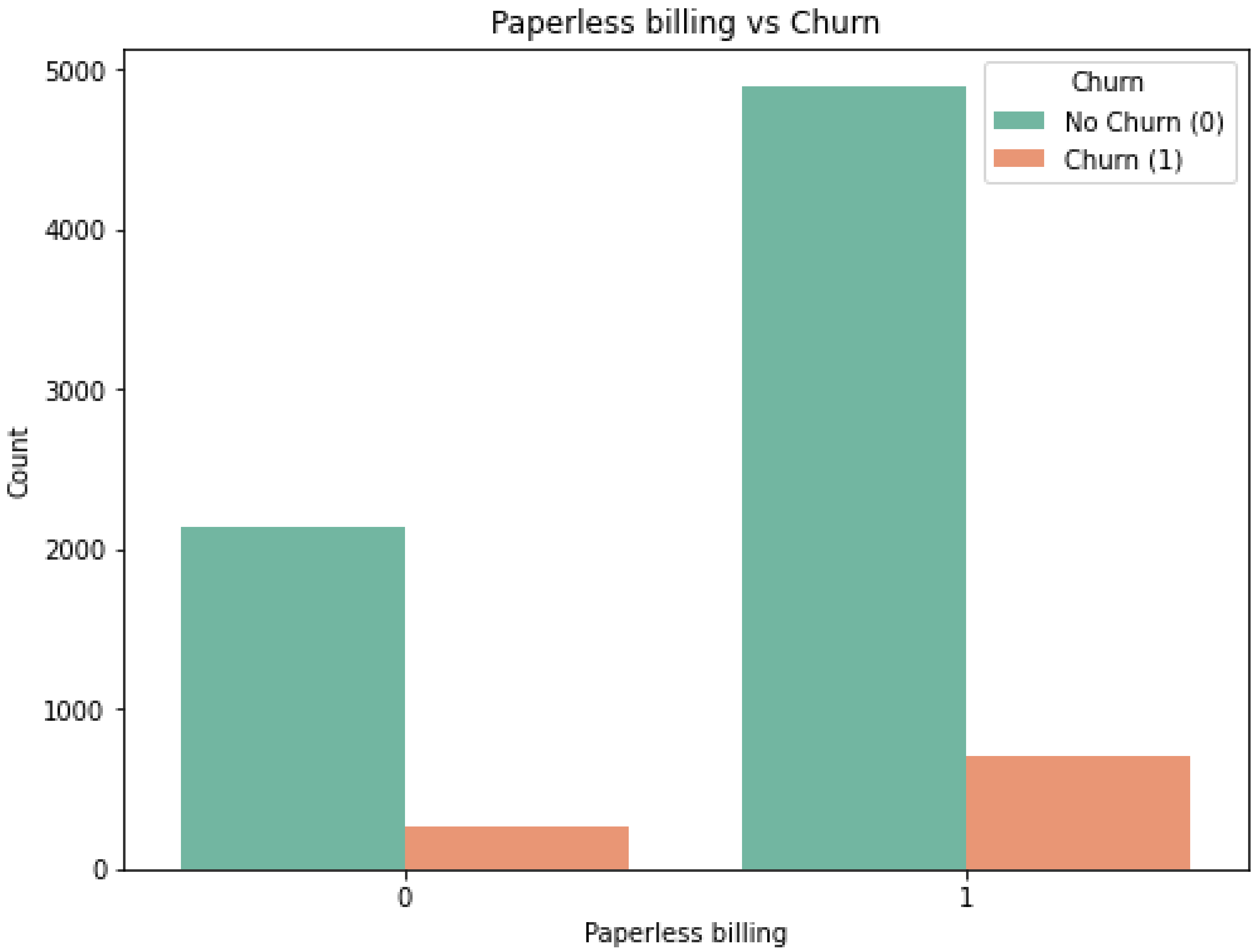
2. Exploratory Analysis (Categorical Variable)

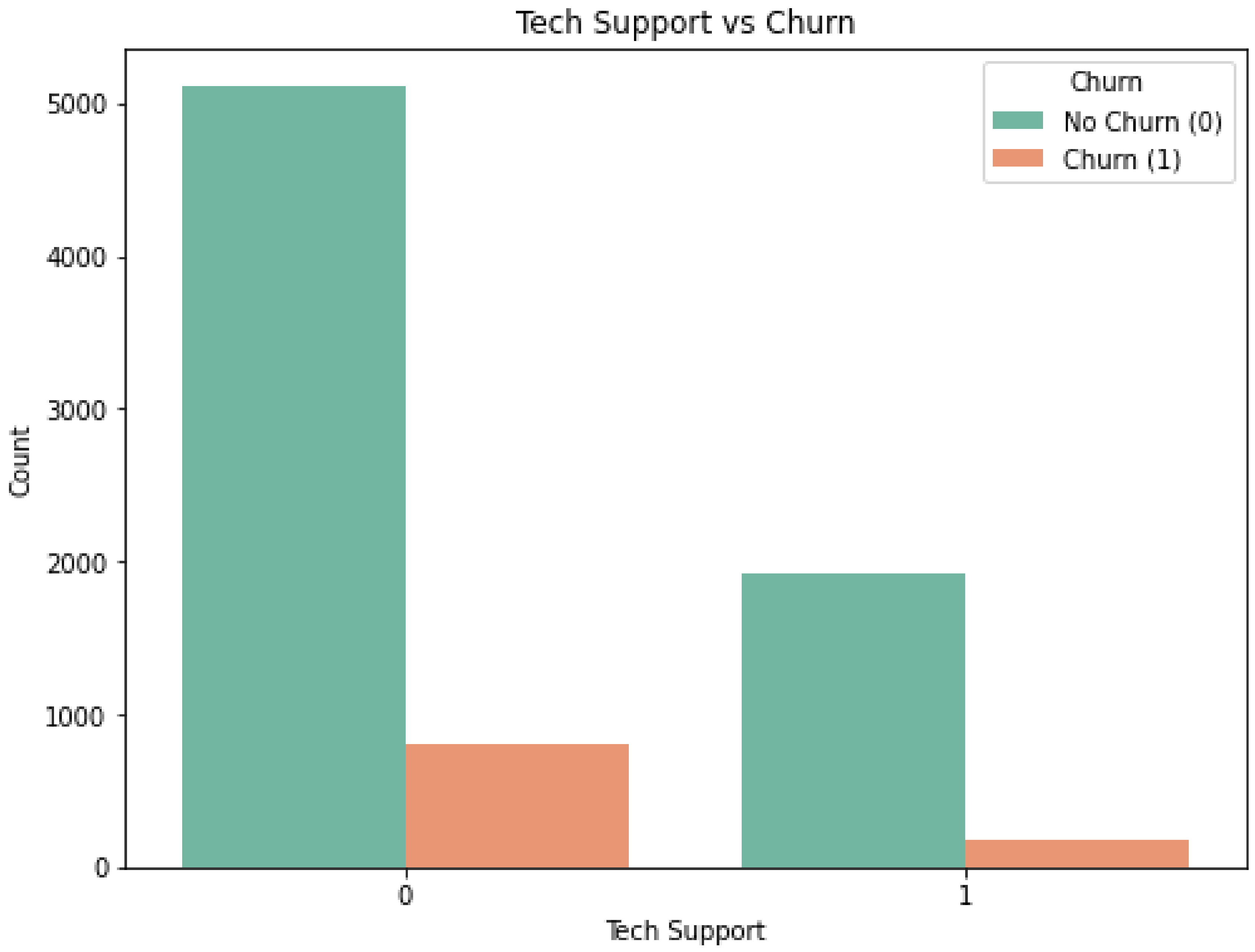


	gender	region	contract_type	payment_method	paperless_billing	internet_service	online_security	tech_support	streaming_tv	streaming_movies	auto_pay	favourite_animal	astrological_sign	browser_theme_color	churn
gender	1	0.62	0.33	0.33	0.99	0.27	0.52	0.96	0.33	0.65	0.95	0.015	0.94	0.51	0.7
region	0.62	1	0.88	0.68	0.088	0.81	0.24	0.31	0.4	0.51	0.17	0.21	1	0.078	0.5
contract_type	0.33	0.88	1	0.28	0.59	0.28	0.63	0.3	0.92	0.059	0.97	0.57	0.18	0.034	0.86
payment_method	0.33	0.68	0.28	1	0.35	0.56	0.75	0.69	0.92	0.99	0.026	0.41	0.23	0.7	0.84
paperless_billing	0.99	0.088	0.59	0.35	1	0.26	0.38	0.68	0.38	0.16	0.46	0.48	0.64	0.31	0.087
internet_service	0.27	0.81	0.28	0.56	0.26	1	0.6	0.18	0.49	0.025	0.69	0.34	0.13	0.9	0.35
online_security	0.52	0.24	0.63	0.75	0.38	0.6	1	0.58	0.64	0.9	0.58	0.34	0.74	0.1	0.44
tech_support	0.96	0.31	0.3	0.69	0.68	0.18	0.58	1	0.071	0.74	0.28	0.77	0.58	0.76	1e-10
streaming_tv	0.33	0.4	0.92	0.92	0.38	0.49	0.64	0.071	1	0.0058	0.9	0.2	0.18	0.41	0.16
streaming_movies	0.65	0.51	0.059	0.99	0.16	0.025	0.9	0.74	0.0058	1	0.73	0.17	0.82	0.092	0.8
auto_pay	0.95	0.17	0.97	0.026	0.46	0.69	0.58	0.28	0.9	0.73	1	0.22	0.44	0.36	0.88
favourite_animal	0.015	0.21	0.57	0.41	0.48	0.34	0.34	0.77	0.2	0.17	0.22	1	0.84	0.84	0.58
astrological_sign	0.94	1	0.18	0.23	0.64	0.13	0.74	0.58	0.18	0.82	0.44	0.84	1	0.51	0.57
browser_theme_color	0.51	0.078	0.034	0.7	0.31	0.9	0.1	0.76	0.41	0.092	0.36	0.84	0.51	1	0.36
churn	0.7	0.5	0.86	0.84	0.087	0.35	0.44	1e-10	0.16	0.8	0.88	0.58	0.57	0.36	1

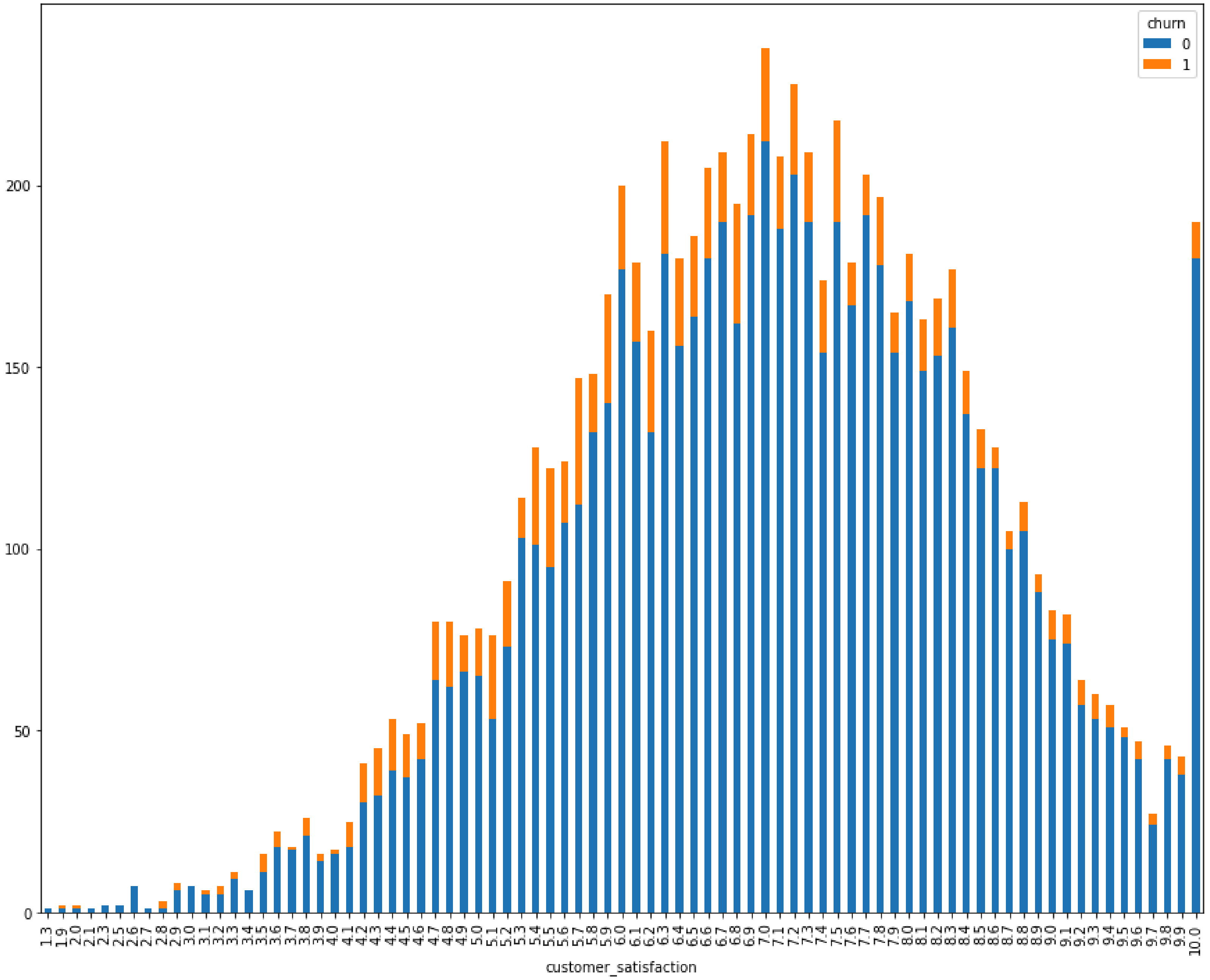


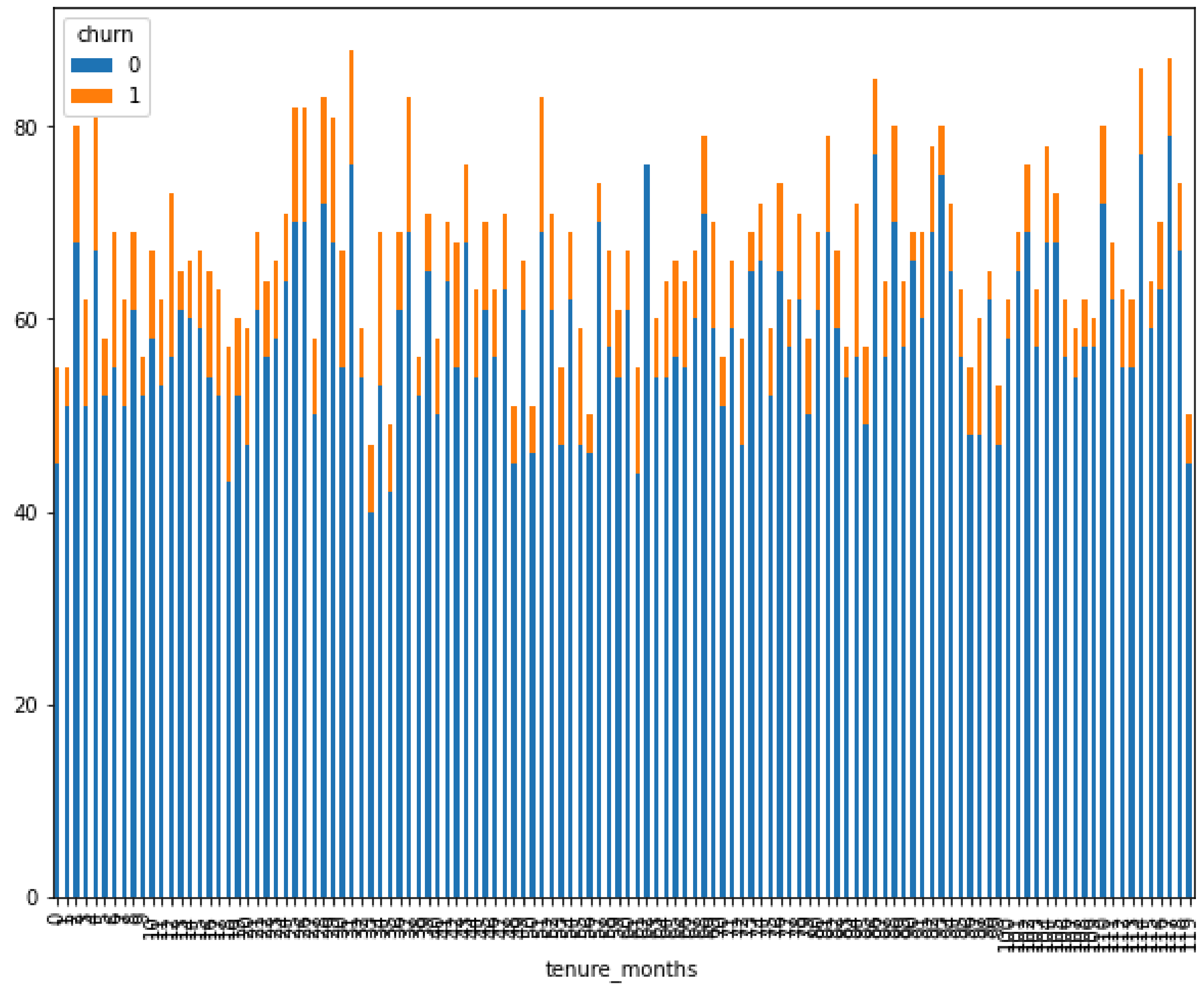
2. Exploratory Analysis (Categorical Variable)



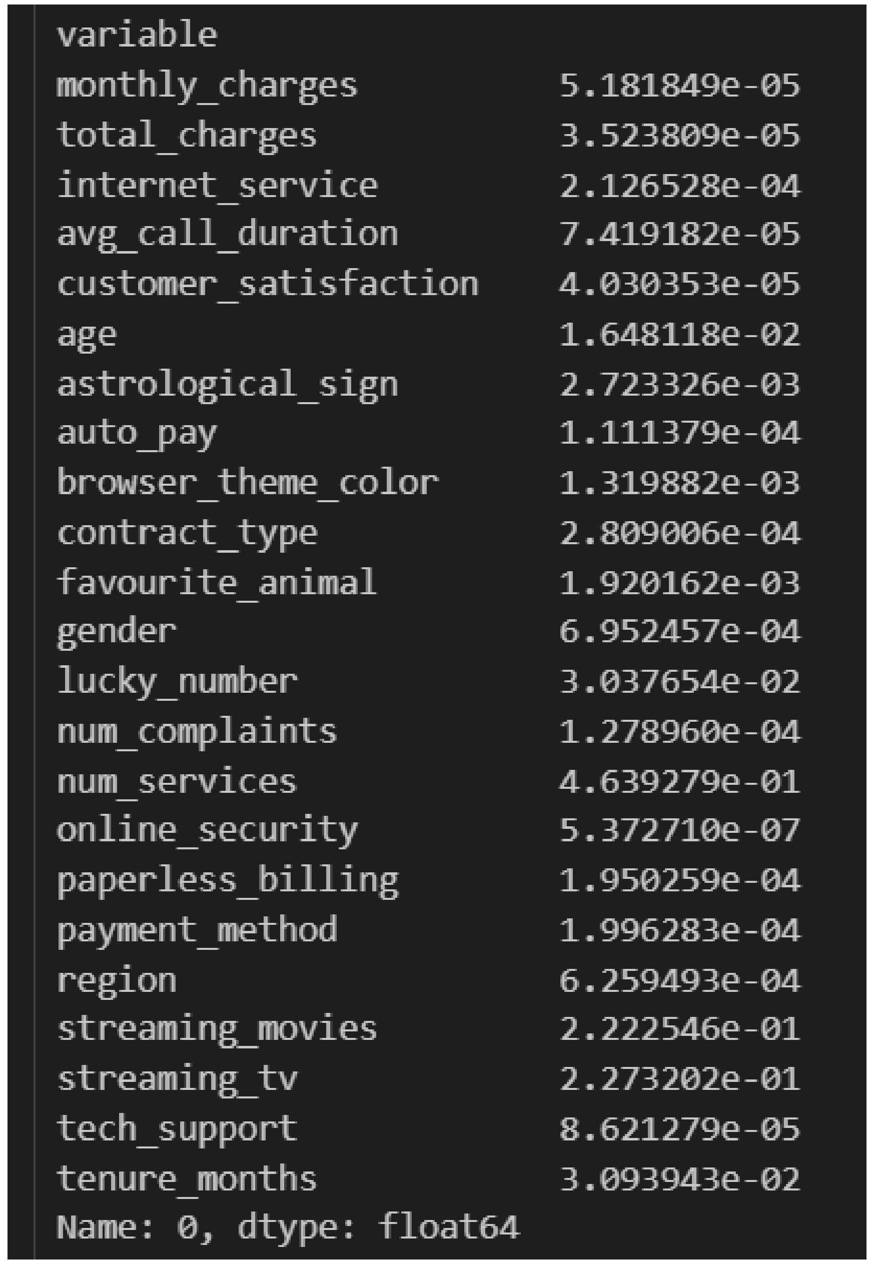


2. Exploratory Analysis (Categorical Variable)

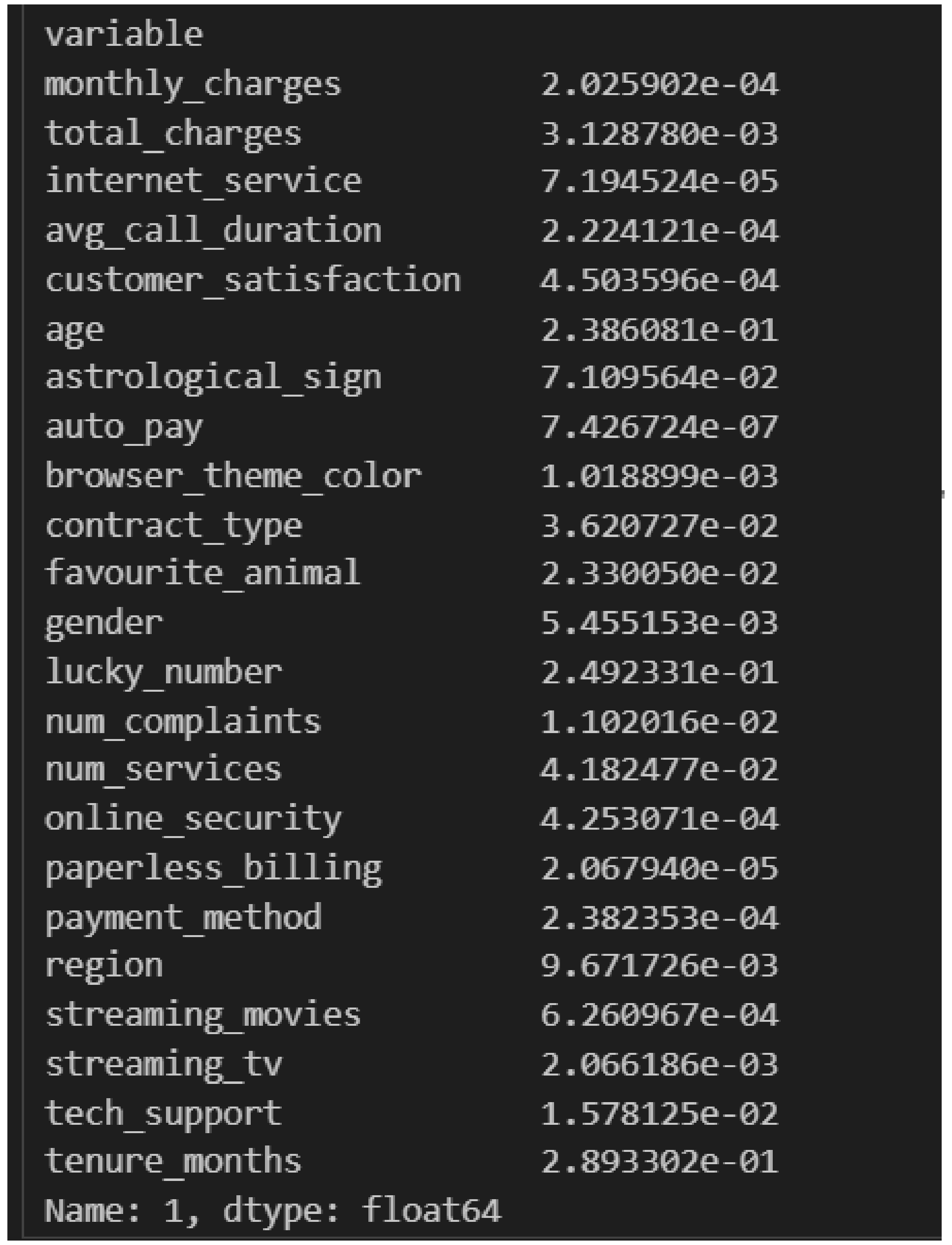




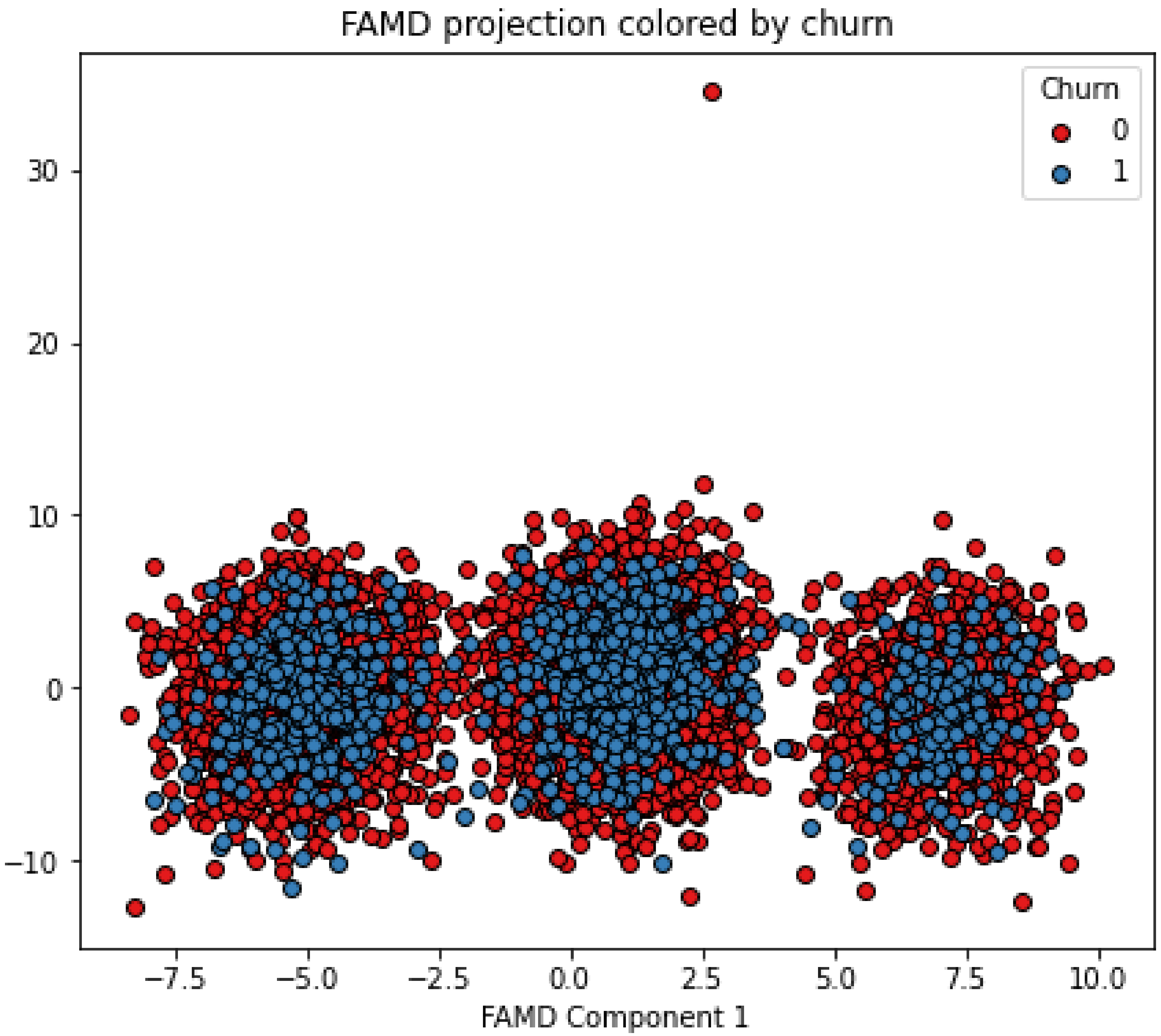
2. Exploratory Analysis (FAMD)



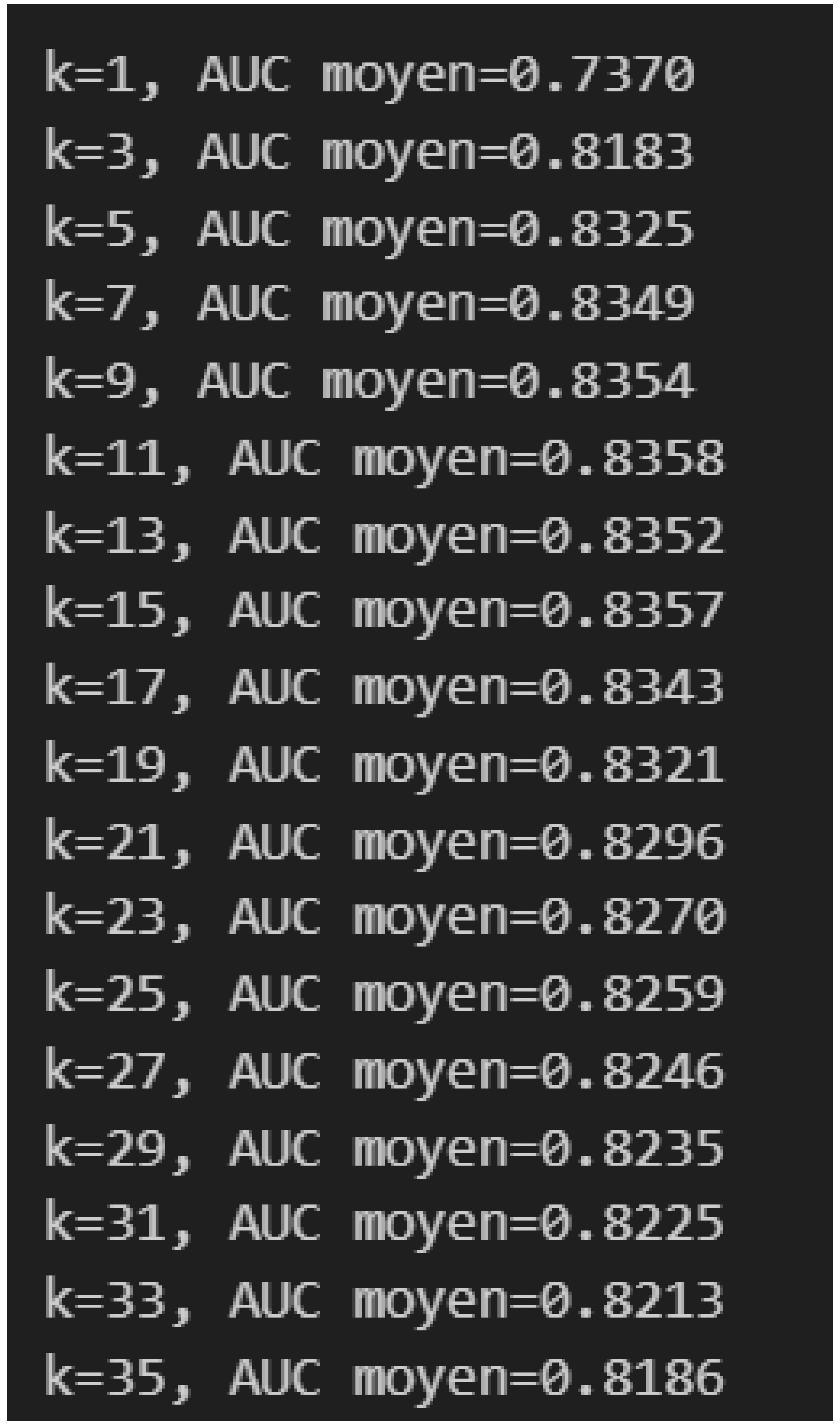
```
variable
monthly_charges          5.181849e-05
total_charges            3.523809e-05
internet_service         2.126528e-04
avg_call_duration        7.419182e-05
customer_satisfaction    4.030353e-05
age                      1.648118e-02
astrological_sign        2.723326e-03
auto_pay                 1.111379e-04
browser_theme_color      1.319882e-03
contract_type            2.809006e-04
favourite_animal         1.920162e-03
gender                   6.952457e-04
lucky_number             3.037654e-02
num_complaints           1.278960e-04
num_services             4.639279e-01
online_security          5.372710e-07
paperless_billing        1.950259e-04
payment_method           1.996283e-04
region                   6.259493e-04
streaming_movies         2.222546e-01
streaming_tv             2.273202e-01
tech_support             8.621279e-05
tenure_months            3.093943e-02
Name: 0, dtype: float64
```



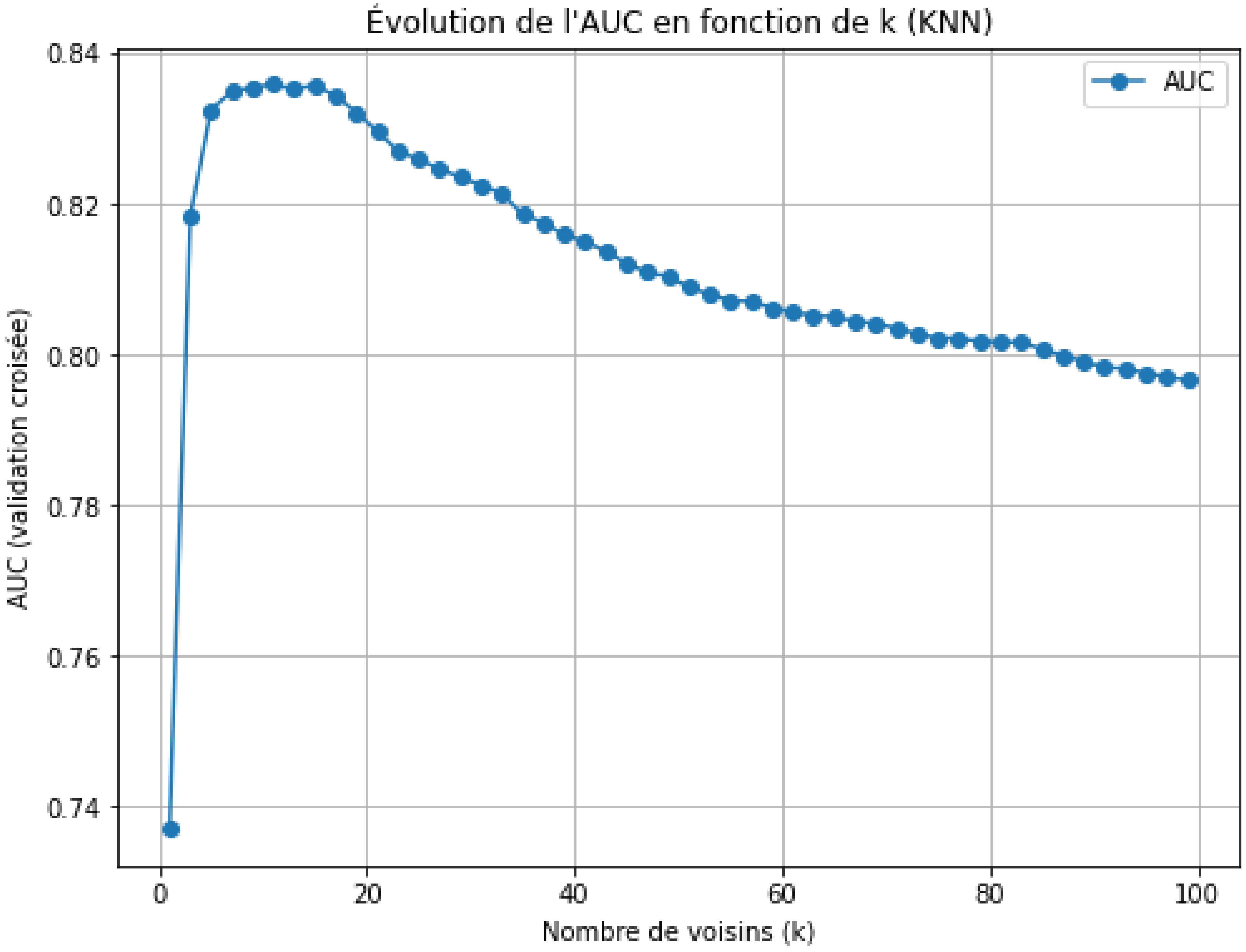
```
variable
monthly_charges          2.025902e-04
total_charges            3.128780e-03
internet_service         7.194524e-05
avg_call_duration        2.224121e-04
customer_satisfaction    4.503596e-04
age                      2.386081e-01
astrological_sign        7.109564e-02
auto_pay                 7.426724e-07
browser_theme_color      1.018899e-03
contract_type            3.620727e-02
favourite_animal         2.330050e-02
gender                   5.455153e-03
lucky_number             2.492331e-01
num_complaints           1.102016e-02
num_services             4.182477e-02
online_security          4.253071e-04
paperless_billing        2.067940e-05
payment_method           2.382353e-04
region                   9.671726e-03
streaming_movies         6.260967e-04
streaming_tv             2.066186e-03
tech_support             1.578125e-02
tenure_months            2.893302e-01
Name: 1, dtype: float64
```



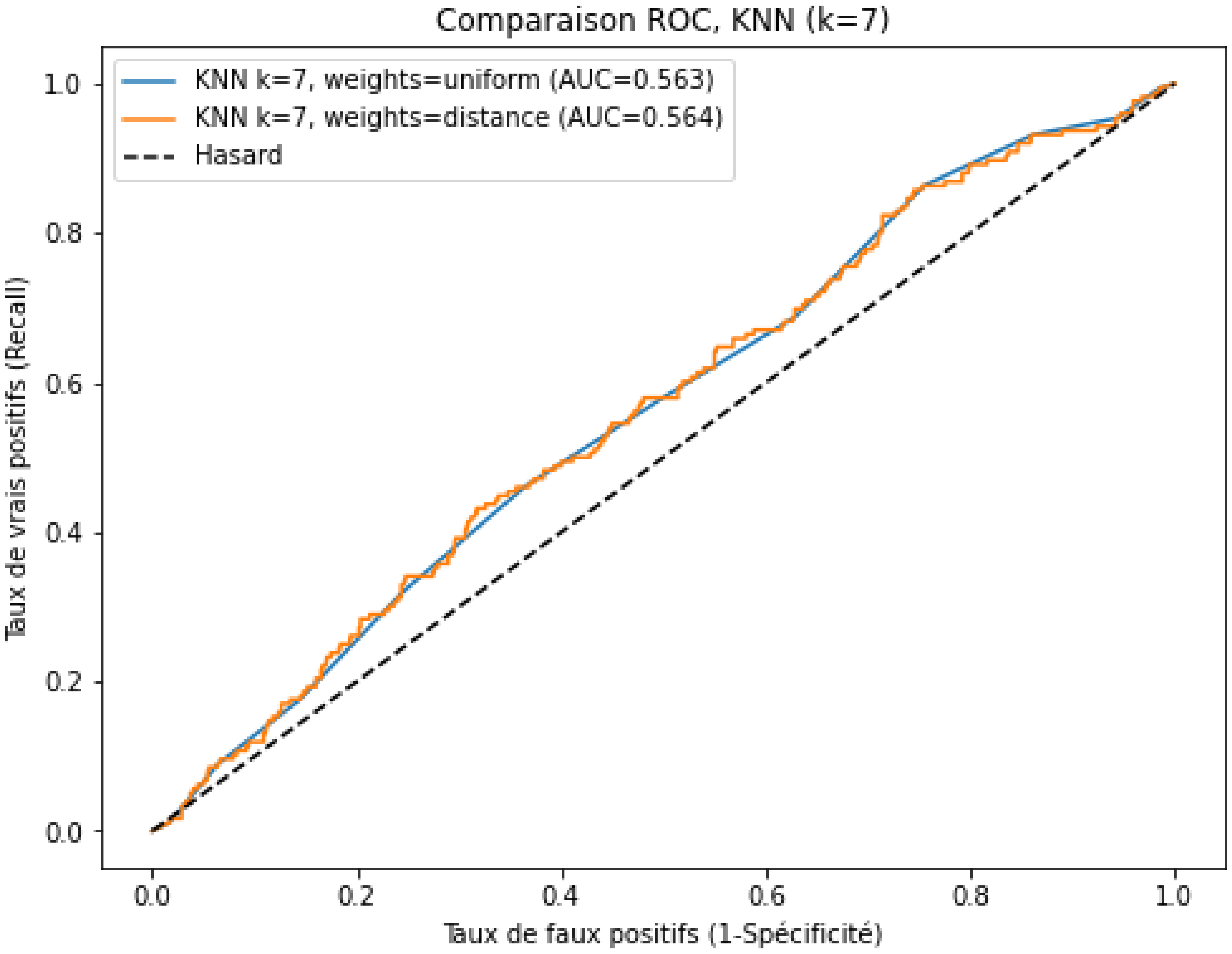
3. KNN model

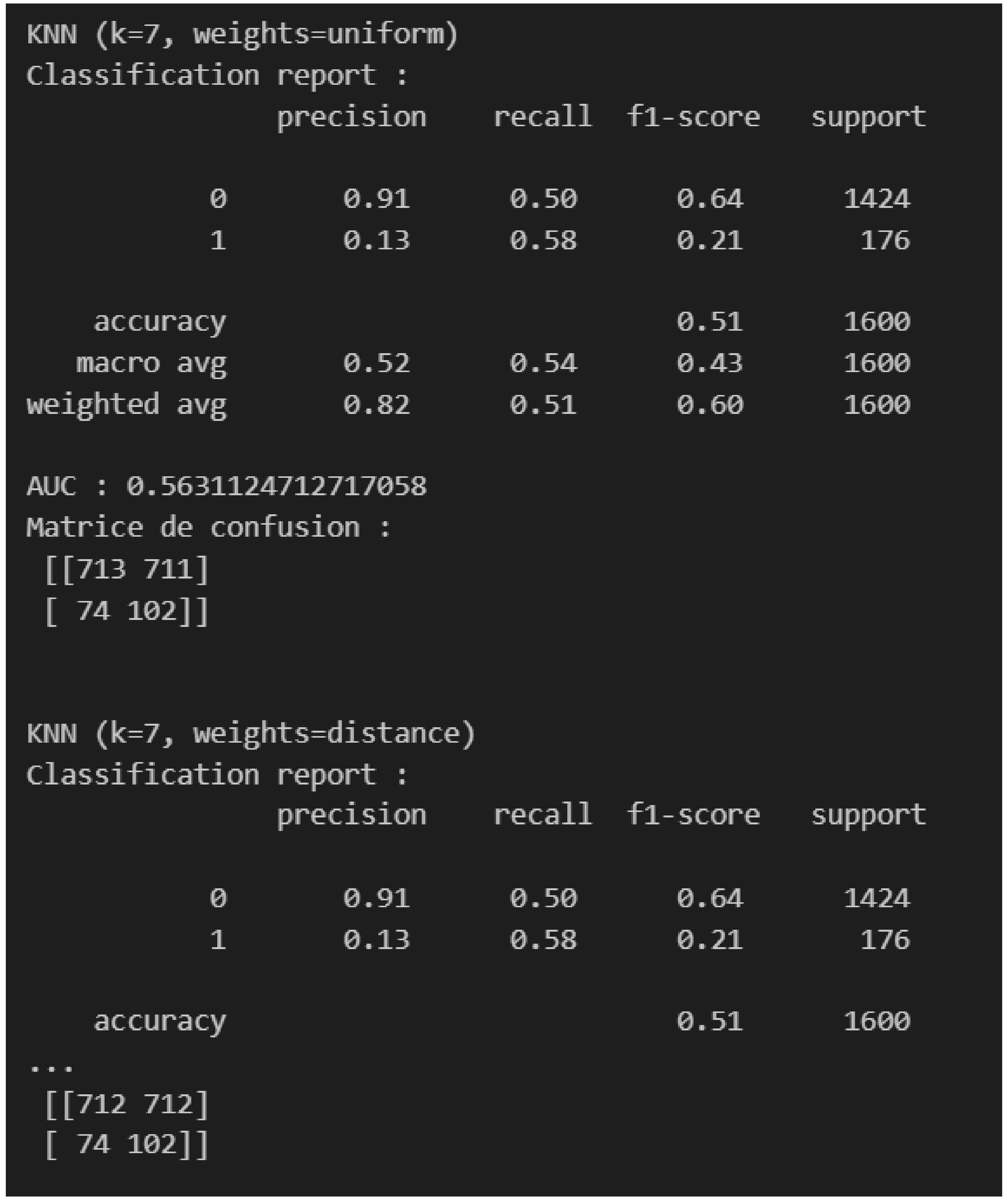


```
k=1, AUC moyen=0.7370
k=3, AUC moyen=0.8183
k=5, AUC moyen=0.8325
k=7, AUC moyen=0.8349
k=9, AUC moyen=0.8354
k=11, AUC moyen=0.8358
k=13, AUC moyen=0.8352
k=15, AUC moyen=0.8357
k=17, AUC moyen=0.8343
k=19, AUC moyen=0.8321
k=21, AUC moyen=0.8296
k=23, AUC moyen=0.8270
k=25, AUC moyen=0.8259
k=27, AUC moyen=0.8246
k=29, AUC moyen=0.8235
k=31, AUC moyen=0.8225
k=33, AUC moyen=0.8213
k=35, AUC moyen=0.8186
```



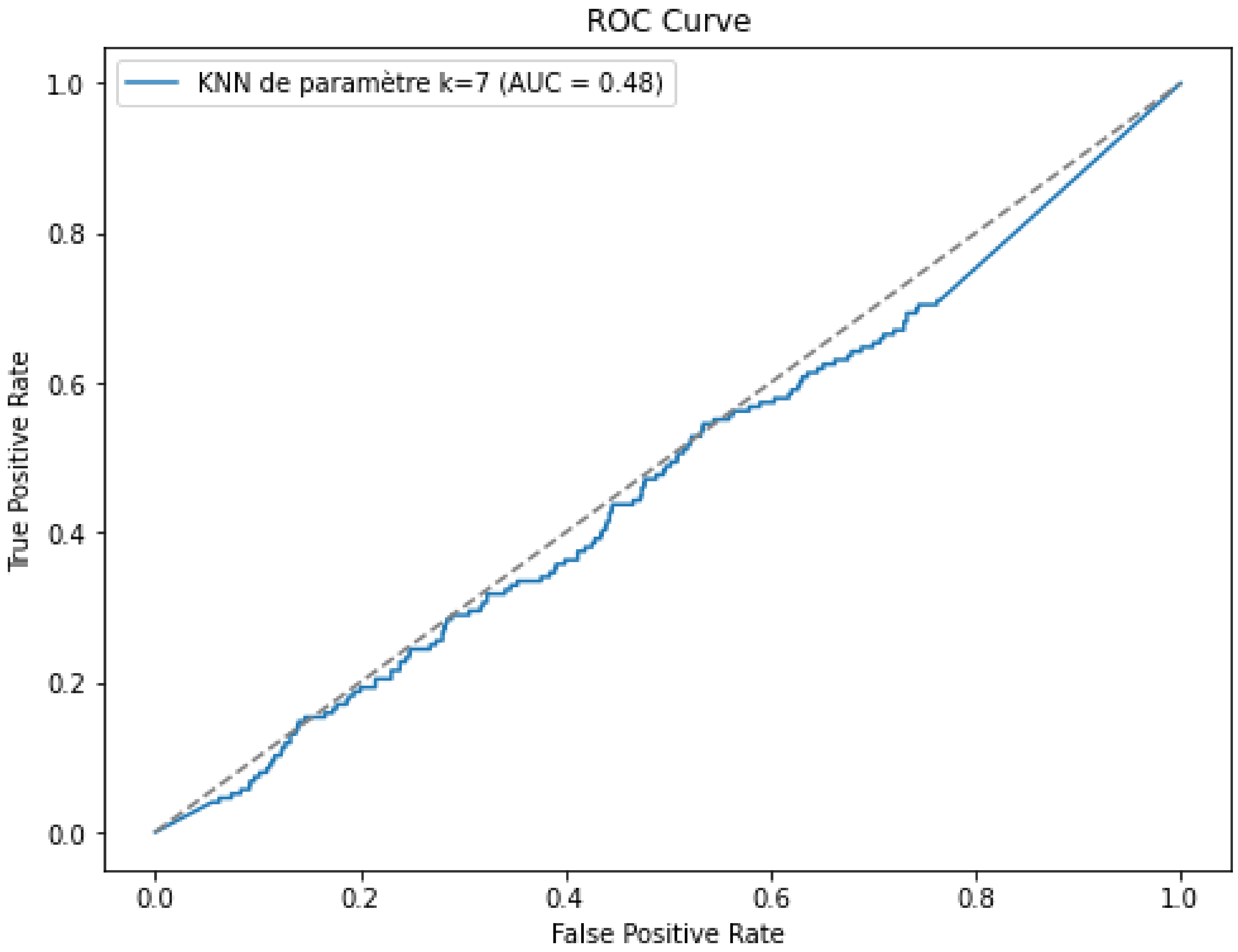
3. KNN model



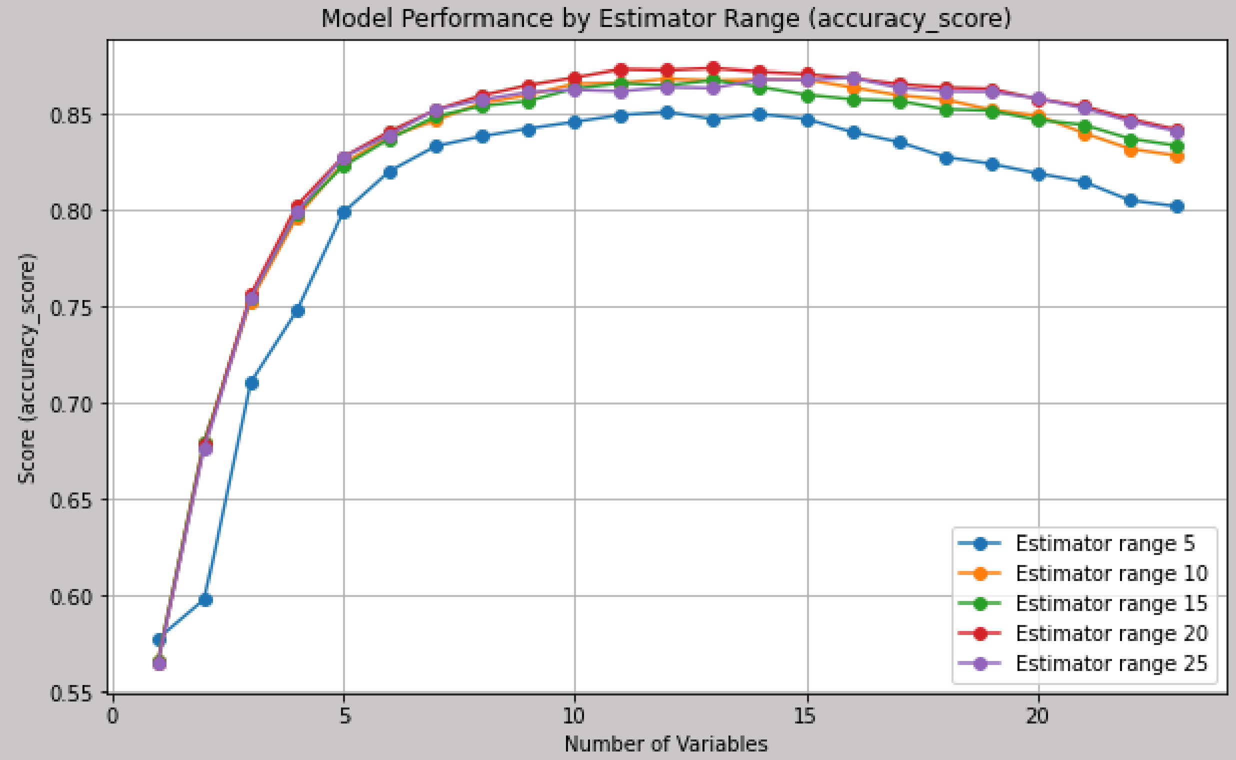


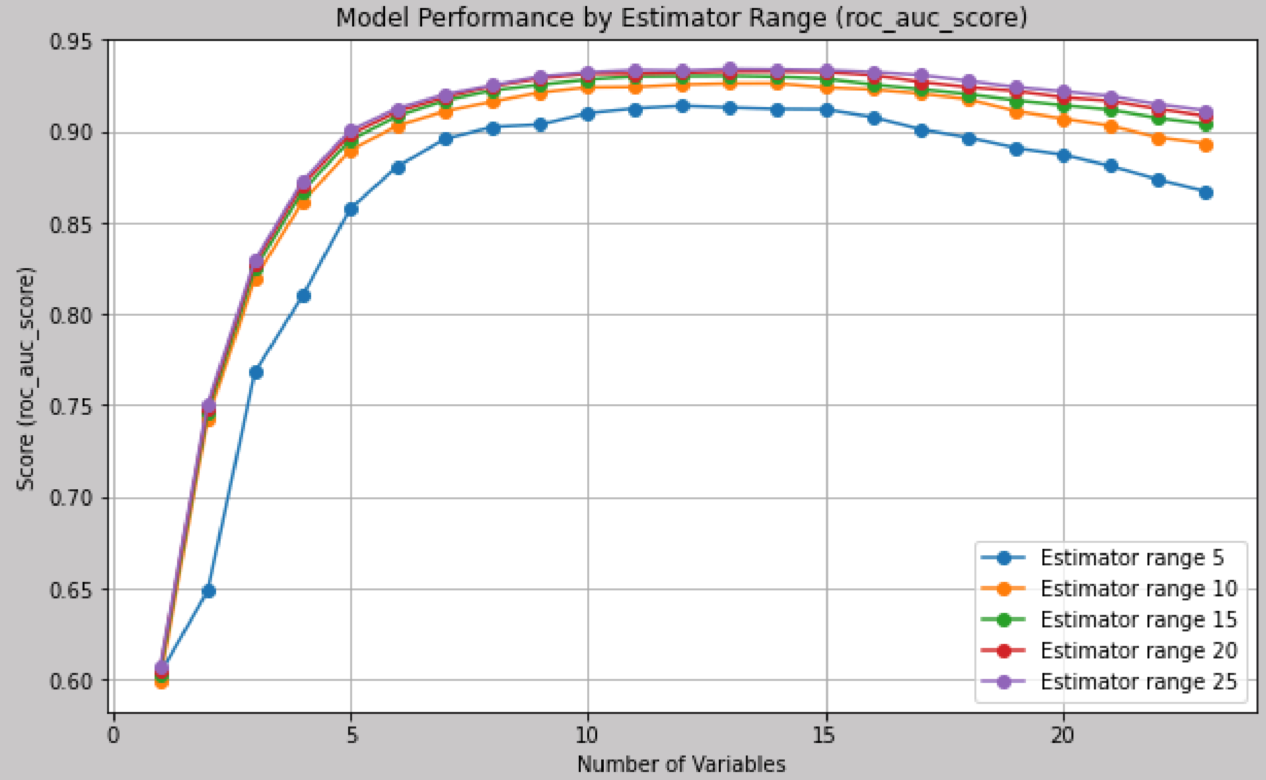
```
KNN (k=7, weights=uniform)
Classification report :
              precision    recall  f1-score   support

           0       0.91      0.50      0.64      1424
           1       0.13      0.58      0.21       176

    accuracy                           0.51      1600
   macro avg       0.52      0.54      0.43      1600
weighted avg       0.82      0.51      0.60      1600

AUC : 0.5631124712717058
Matrice de confusion :
 [[713 711]
 [ 74 102]]


KNN (k=7, weights=distance)
Classification report :
              precision    recall  f1-score   support

           0       0.91      0.50      0.64      1424
           1       0.13      0.58      0.21       176

    accuracy                           0.51      1600
...
 [[712 712]
 [ 74 102]]
```

3. KNN model


ROC Curve
KNN de paramètre k=7 (AUC = 0.48)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

4. Bagging for Decision Tree

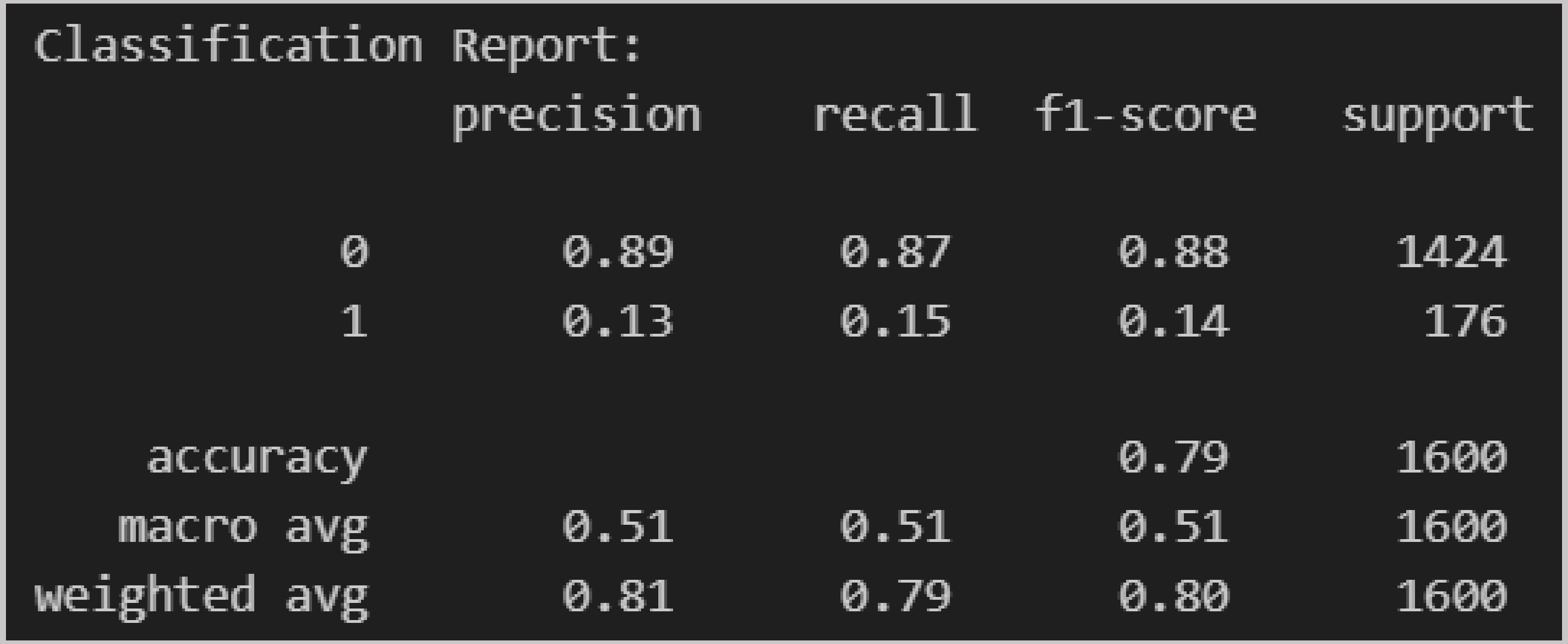




We selected the five features obtained with the backward selection on the model with 15 estimators:

- gender
- region
- total_charges
- internet_service
- online_security
- tech_support
- streaming_tv
- streaming_movies
- num_services
- num_complaints
- auto_pay
- favourite_animal
- browser_theme_color

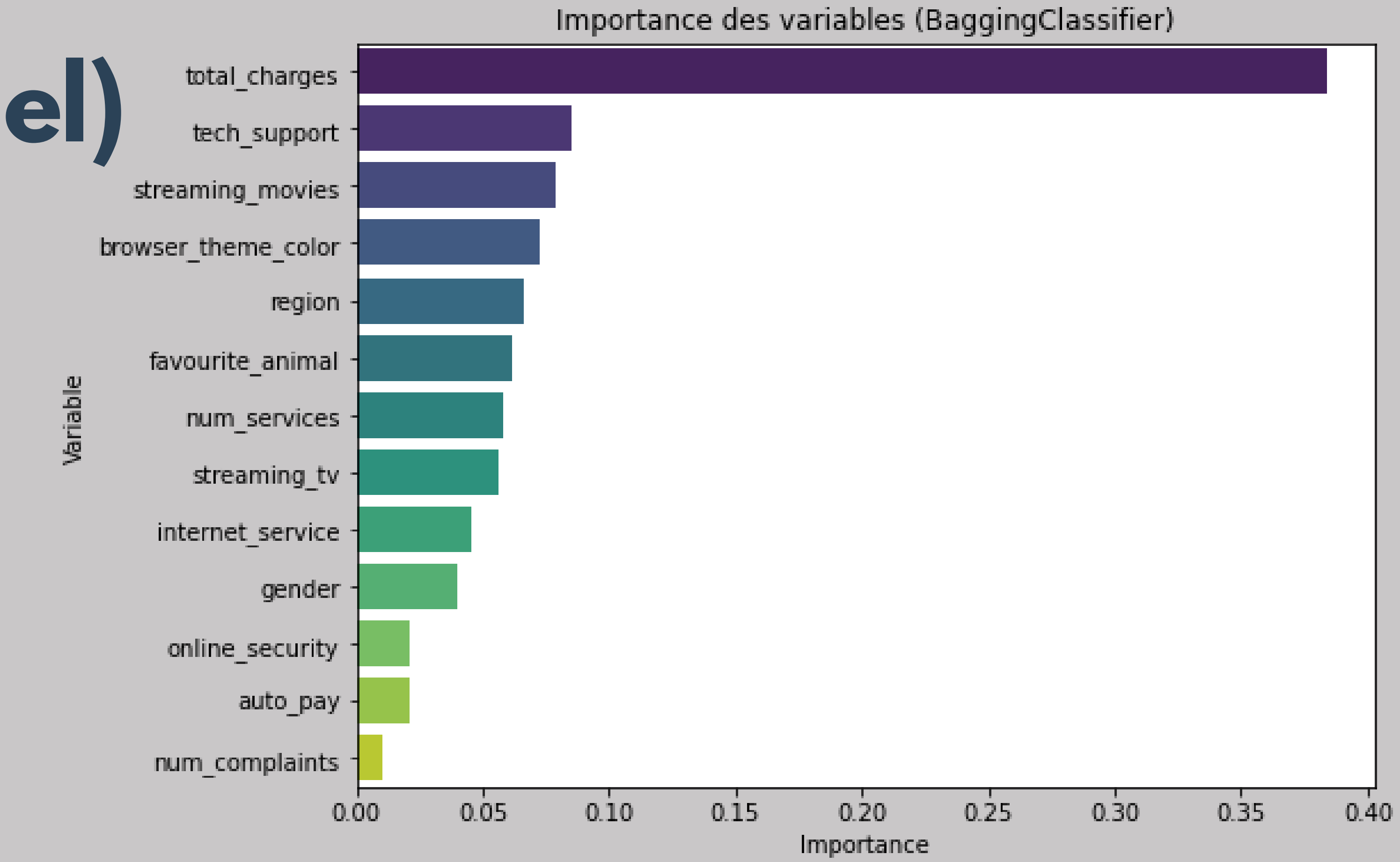
4. Bagging for Decision Tree (best model)

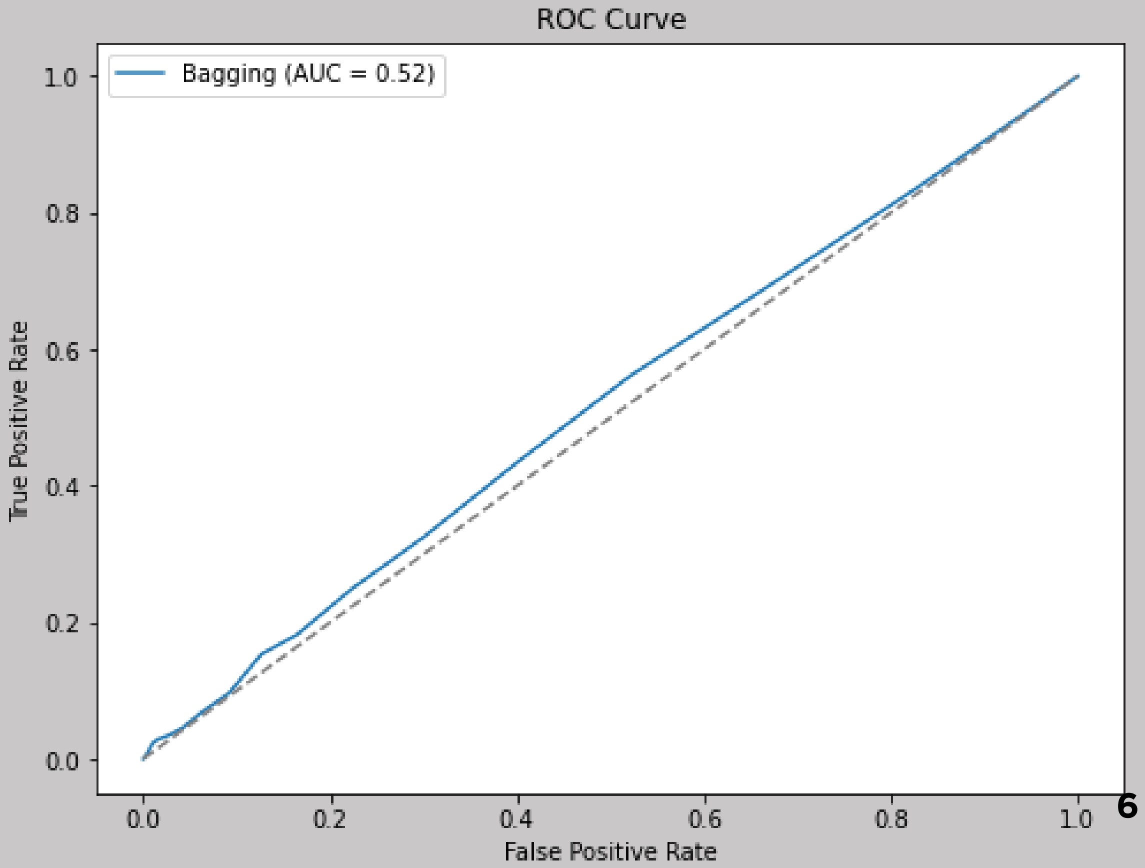


```
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.87      0.88      1424
           1       0.13      0.15      0.14       176

    accuracy                           0.79      1600
   macro avg       0.51      0.51      0.51      1600
weighted avg       0.81      0.79      0.80      1600
```

Optimal Parameters :

- n_estimator = 15





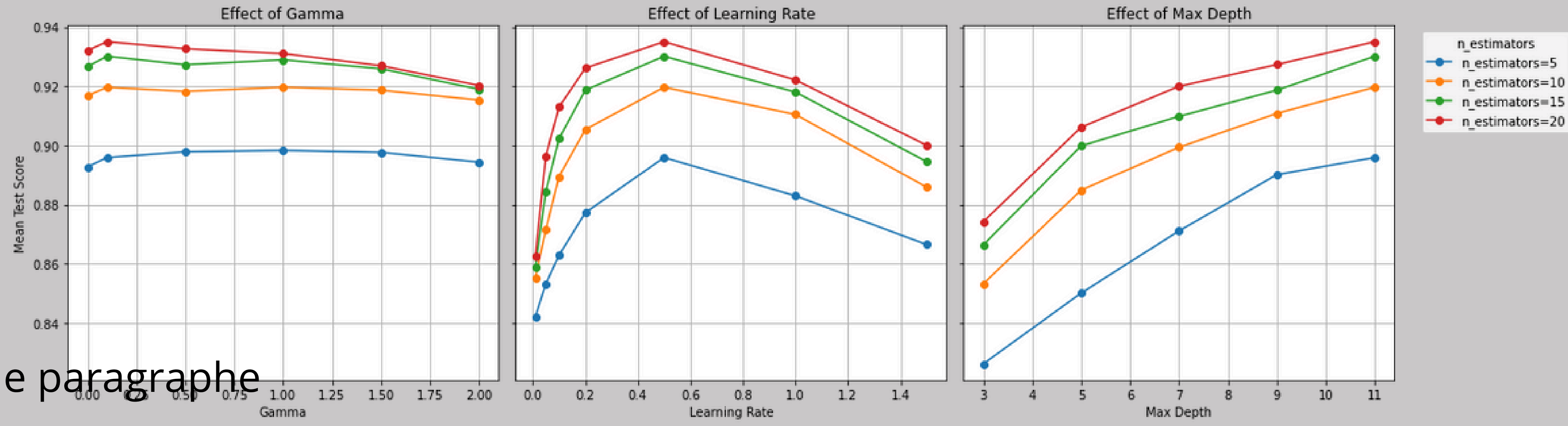
5. XGBClassifier

We first apply a GridSearch in order to get an idea of our optimal parameter for XGBClassifier, we found

- gamma = 0.1,
- learning_rate = 0.5,
- max_depth = 11
- n_estimators = 20

Then we tuned those parameters one by one by fixing the others at their value obtain through GridSearch.

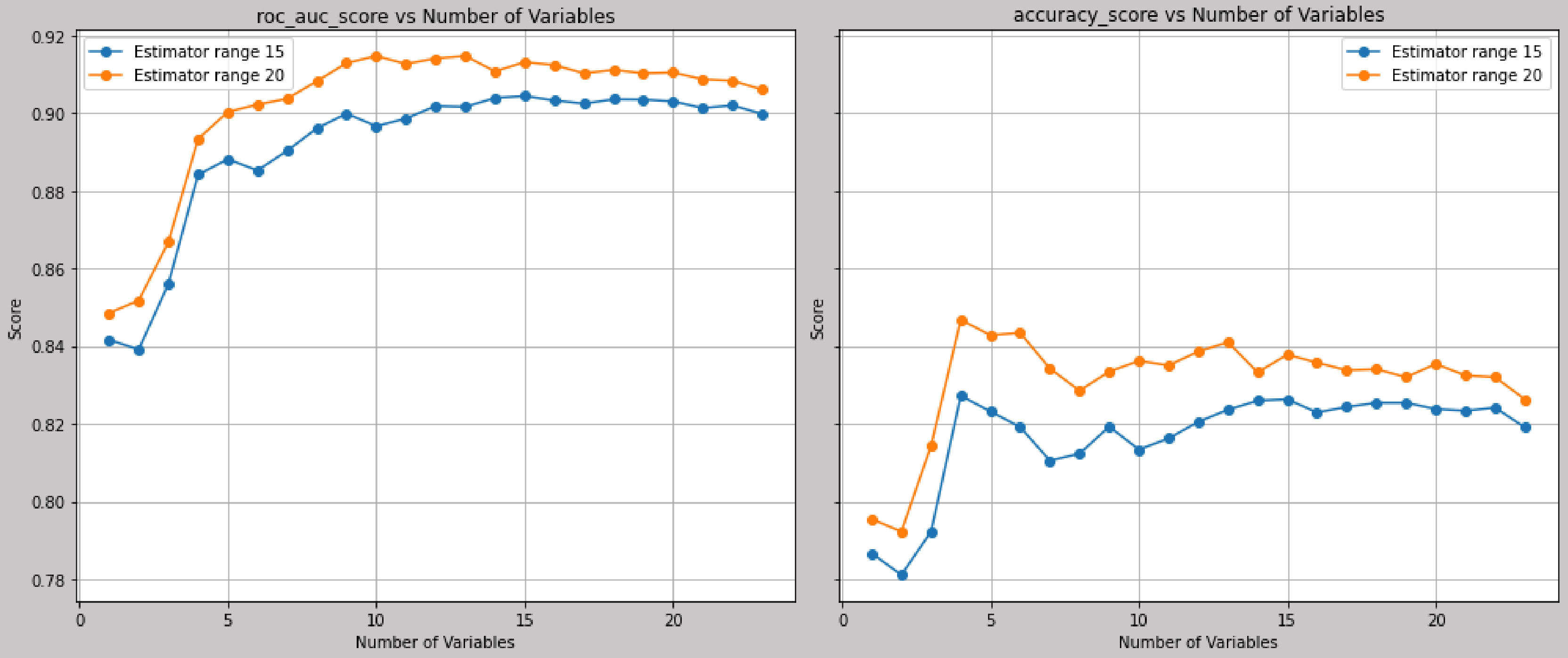
Votre texte de paragraphe



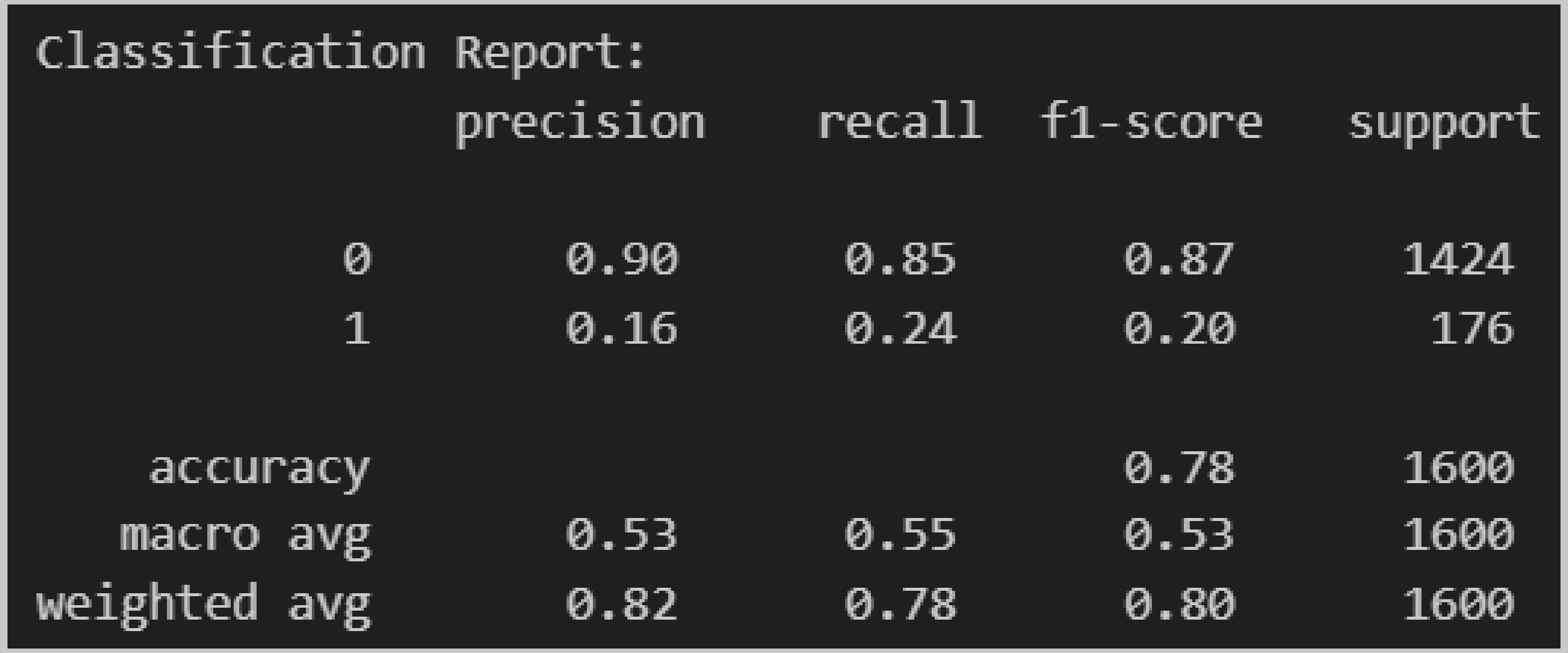
After finding our “best parameters” we decided to use backward selection to determine the most relevent variables for our model.

We decided to settle for the seven best for the roc_score :

- contract_type
- total_charges
- payment_method
- online_security
- tech_support
- streaming_tv
- streaming_movies
- num_services
- customer_satisfaction



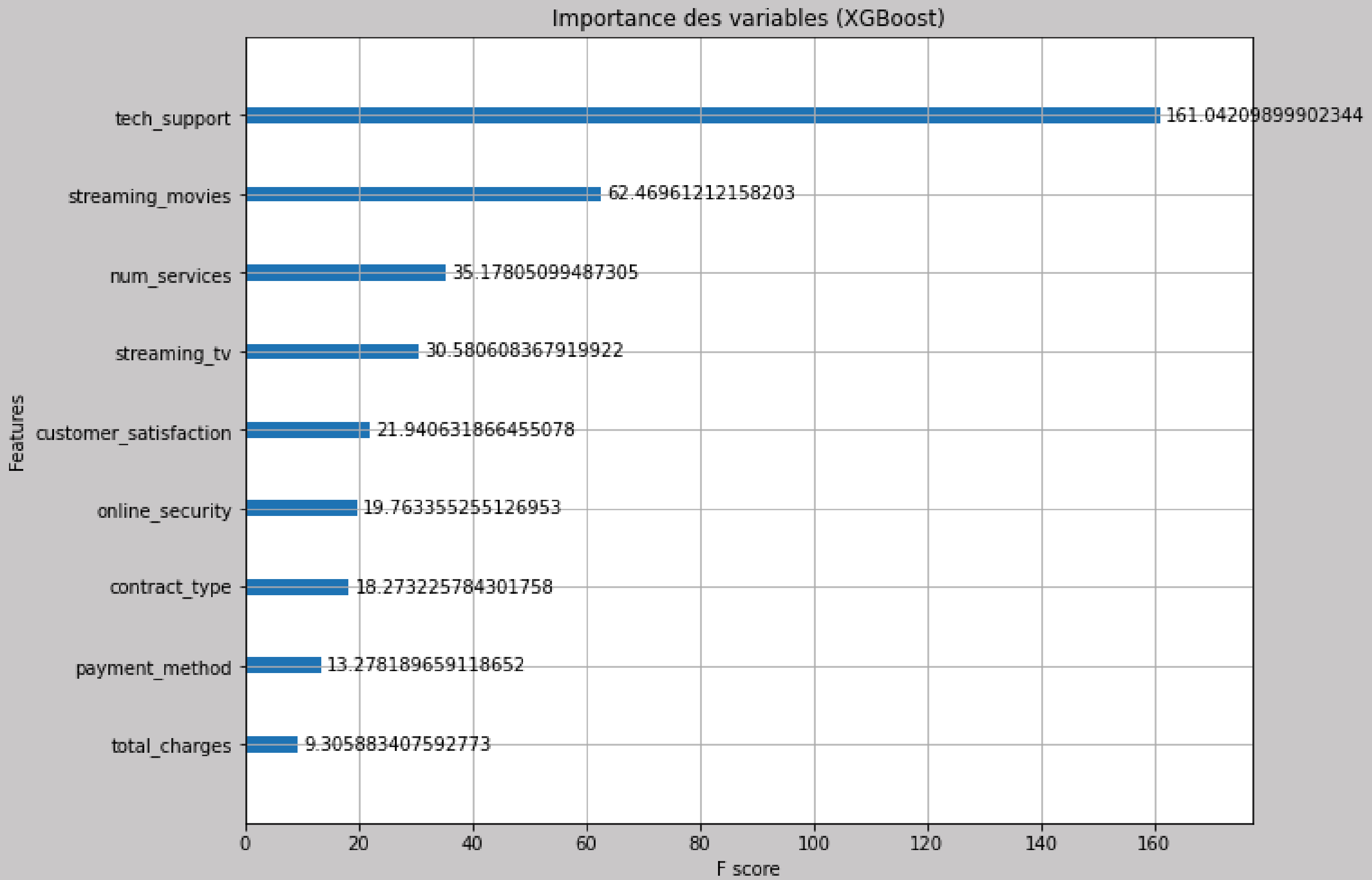
5. XGBClassifier (best model)

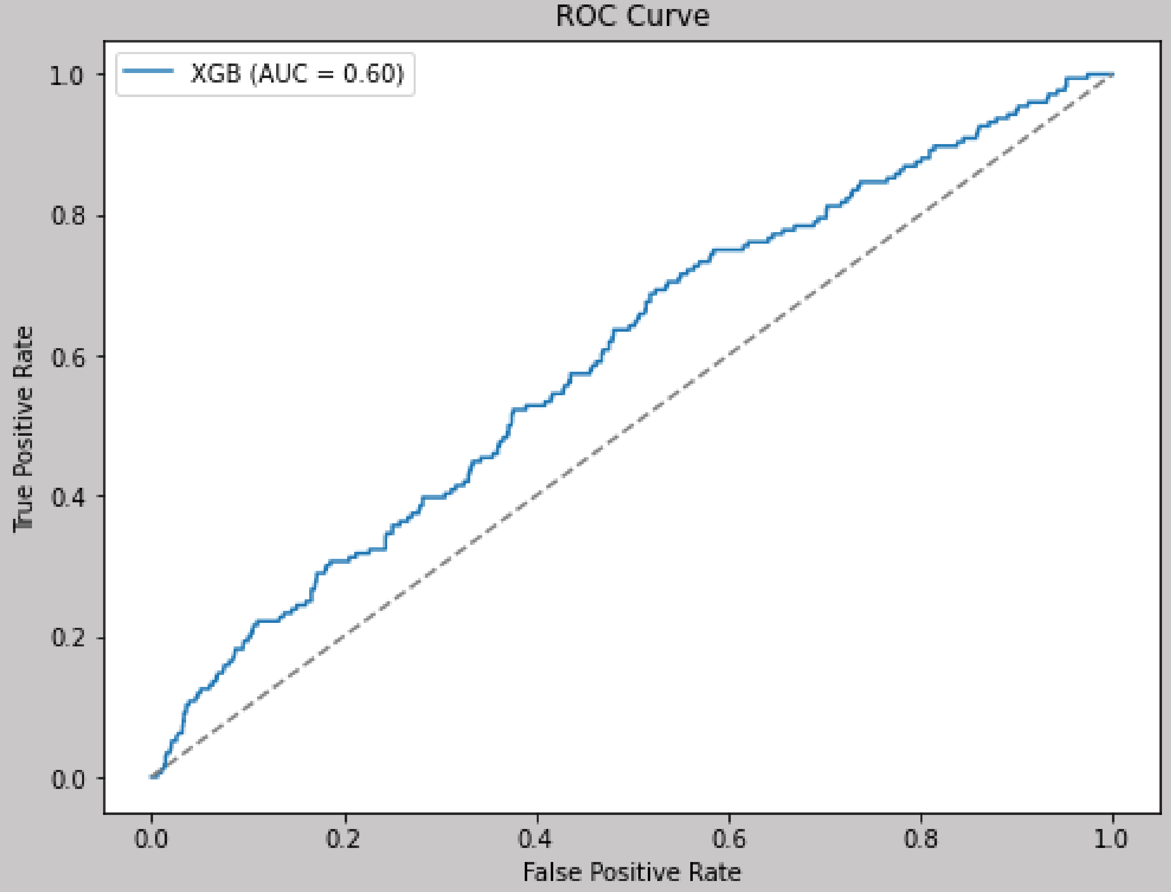


```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.85      0.87      1424
           1       0.16      0.24      0.20       176

    accuracy                           0.78      1600
   macro avg       0.53      0.55      0.53      1600
weighted avg       0.82      0.78      0.80      1600
```

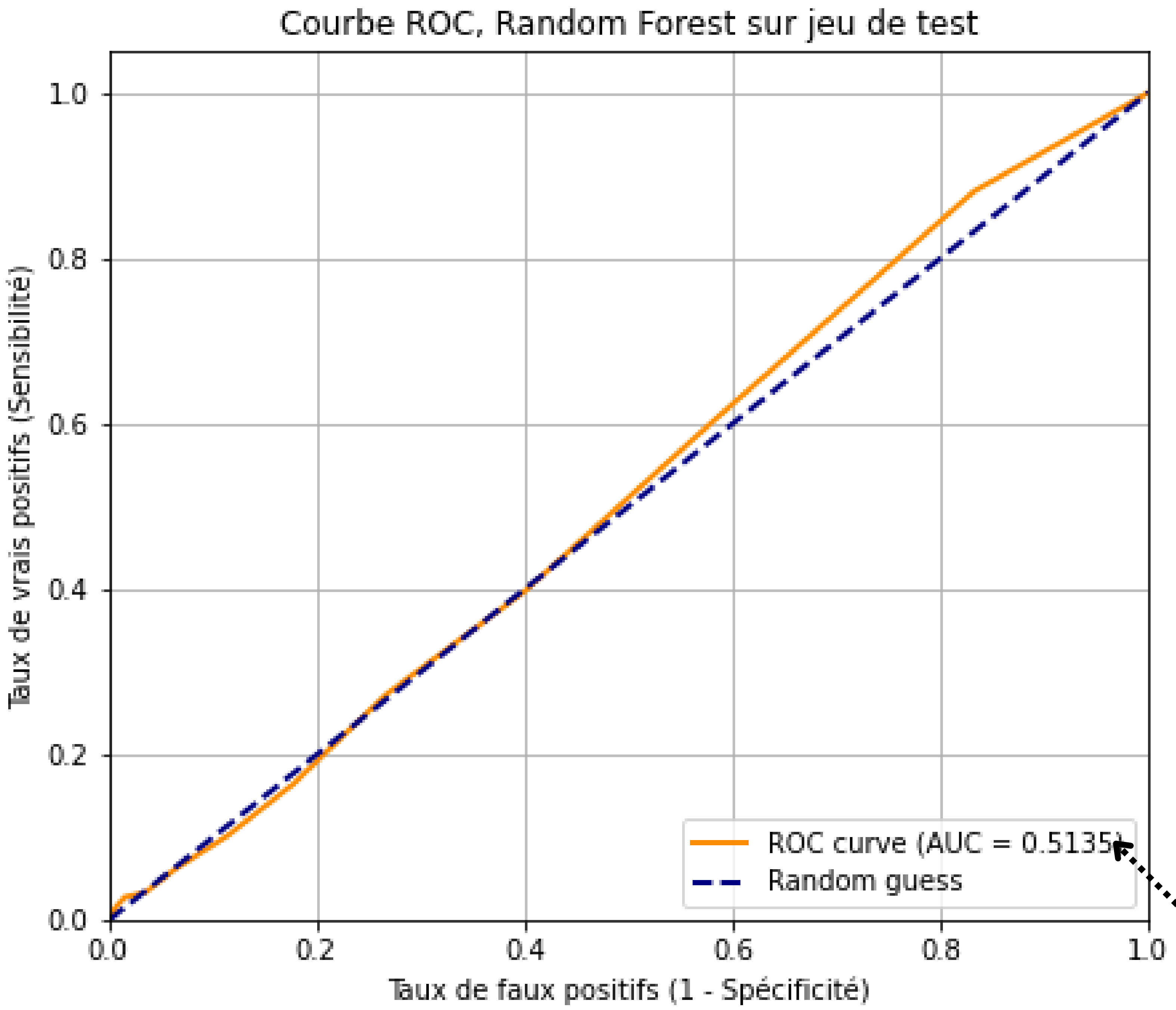
Optimal Parameters :
- gamma = 0.1
- learning rate = 0.5
- max depth = 5
- n_estimator = 15

Selected variables :
- age
- monthly_charges
- total_charges
- internet_service
- tech_support
- customer_satisfaction
- favourite_animal

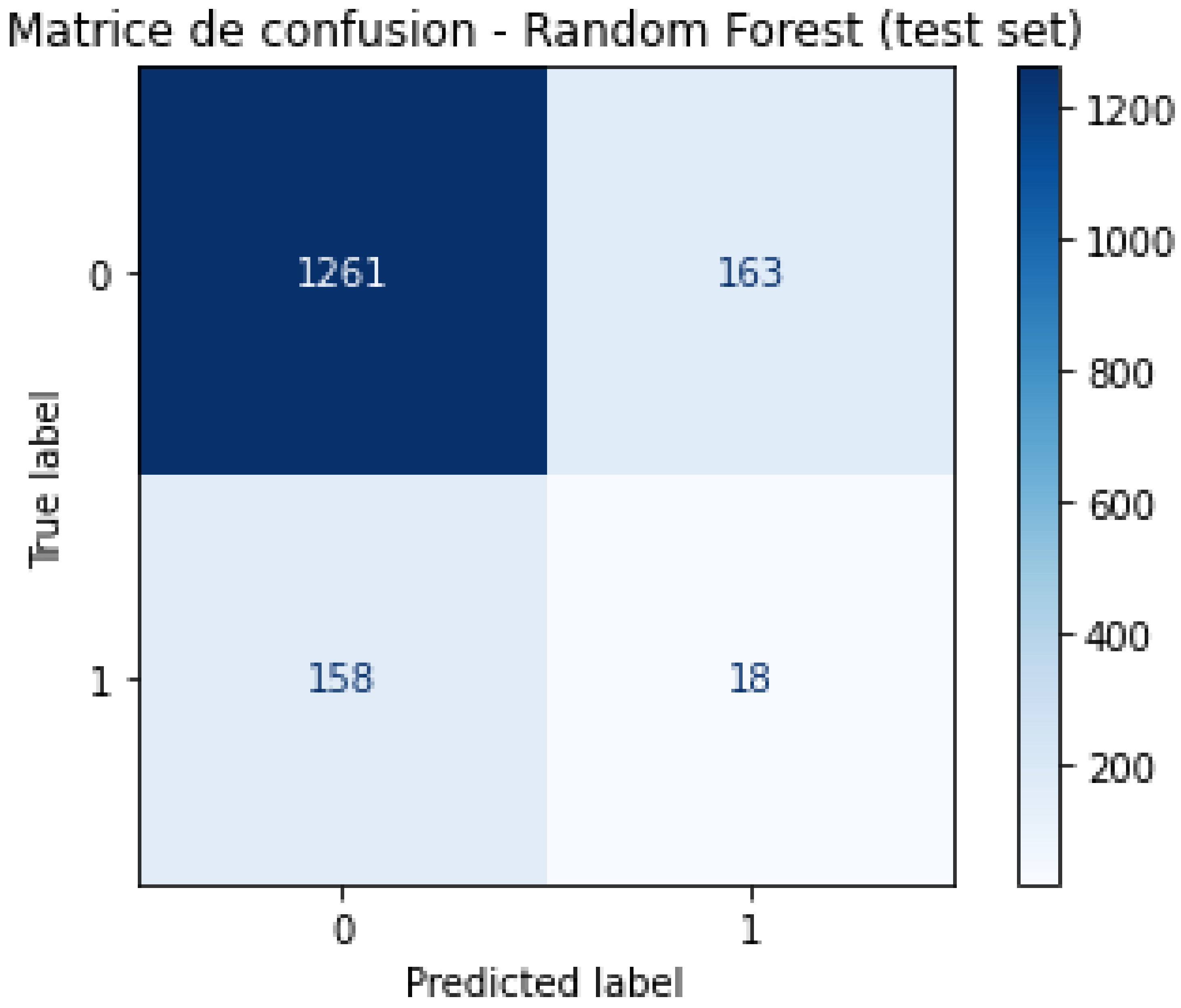




6. Random Forest

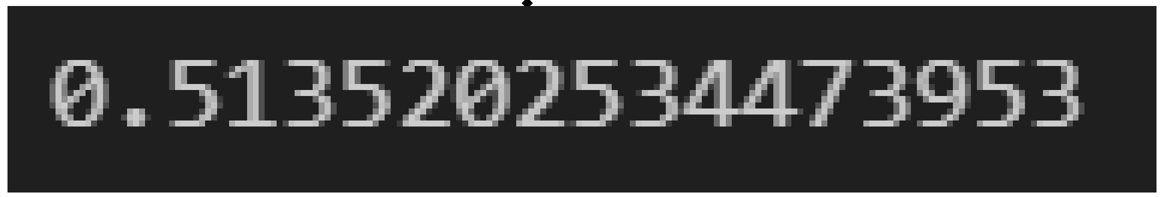


Courbe ROC, Random Forest sur jeu de test



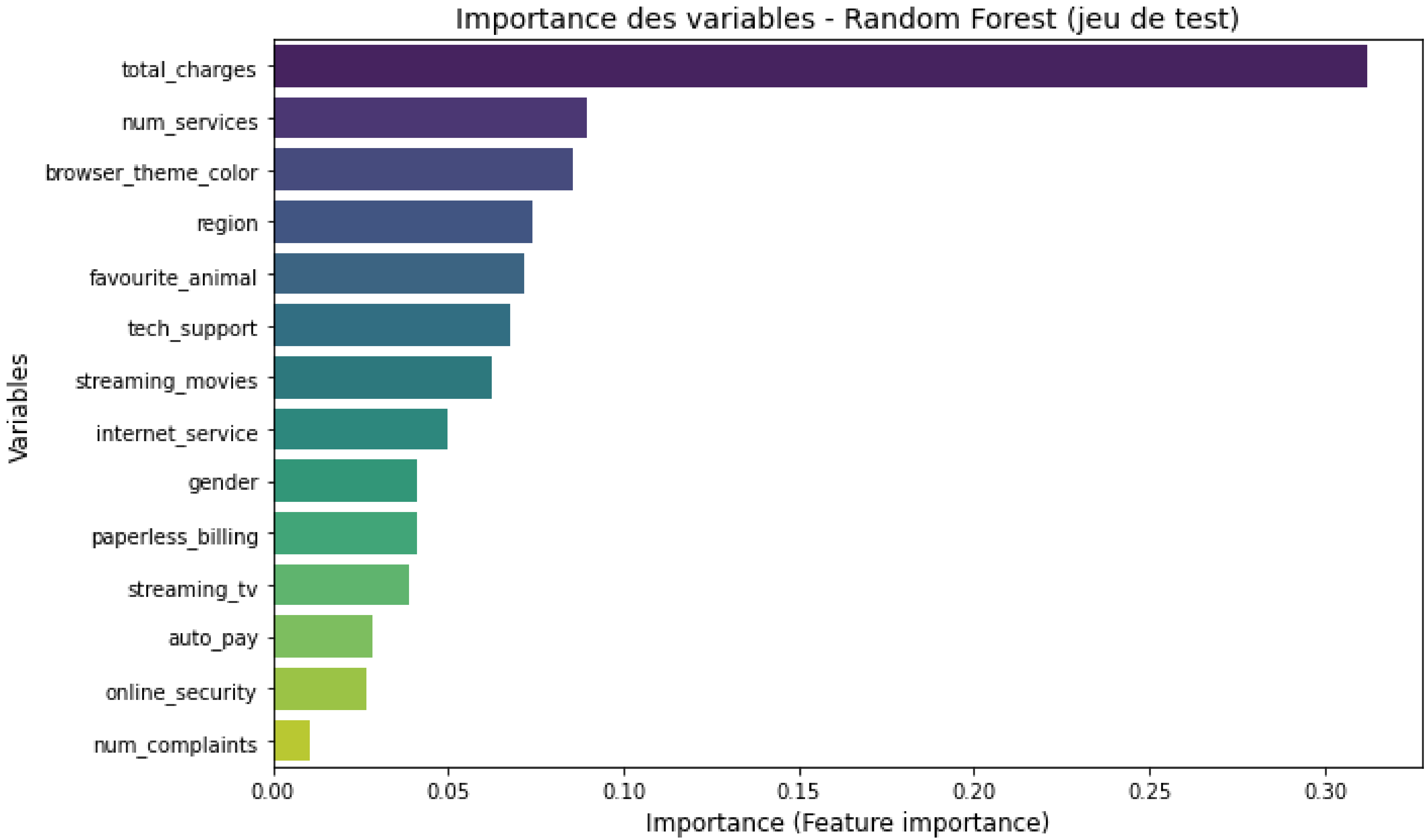
Matrice de confusion - Random Forest (test set)

True label \ Predicted label	0	1
0	1261	163
1	158	18



```
0.5135202534473953
```

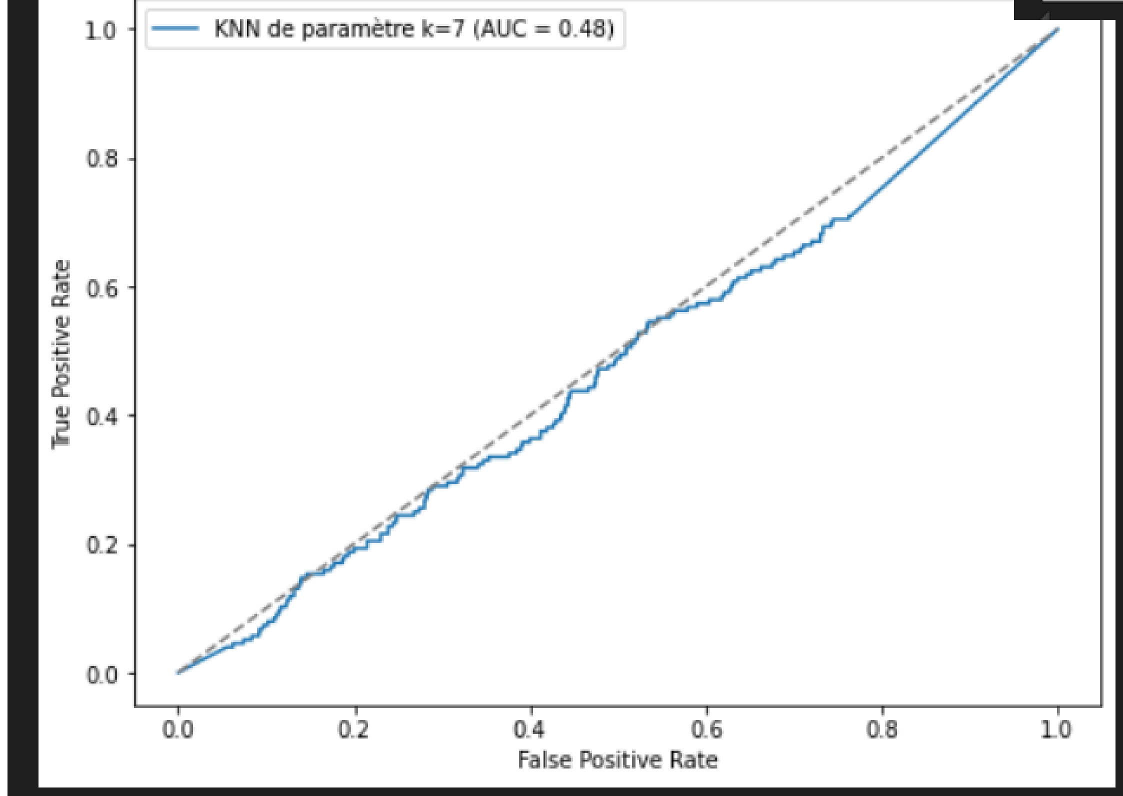
6. Random Forest



7. Conclusion

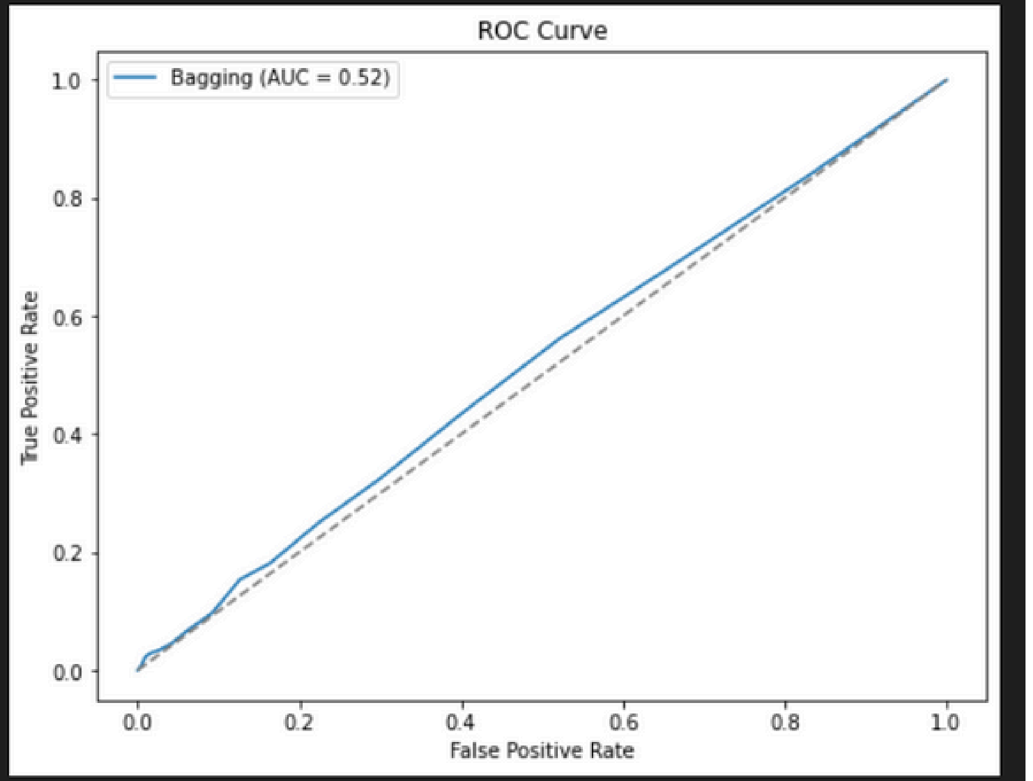
KNN model

```
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.60      0.72      1424
           1       0.10      0.36      0.16       176

    accuracy                           0.58      1600
   macro avg       0.49      0.48      0.44      1600
weighted avg       0.80      0.58      0.66      1600
```



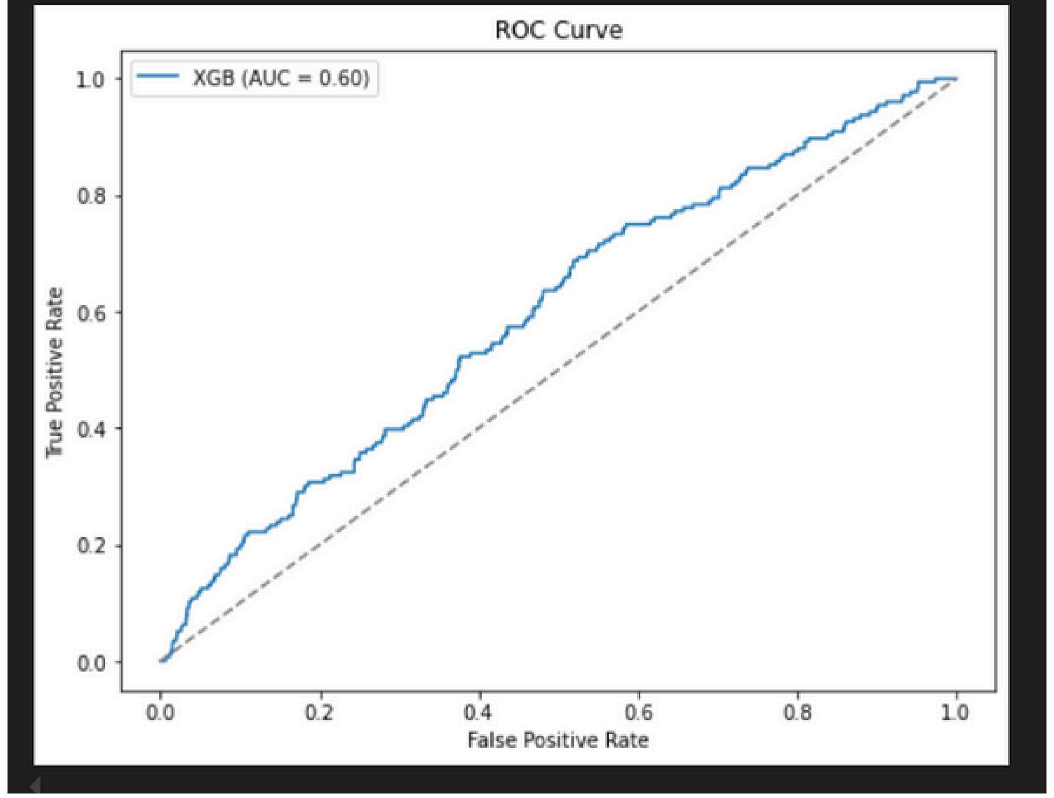
Bagging model

```
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.87      0.88      1424
           1       0.13      0.15      0.14       176

    accuracy                           0.79      1600
   macro avg       0.51      0.51      0.51      1600
weighted avg       0.81      0.79      0.80      1600
```

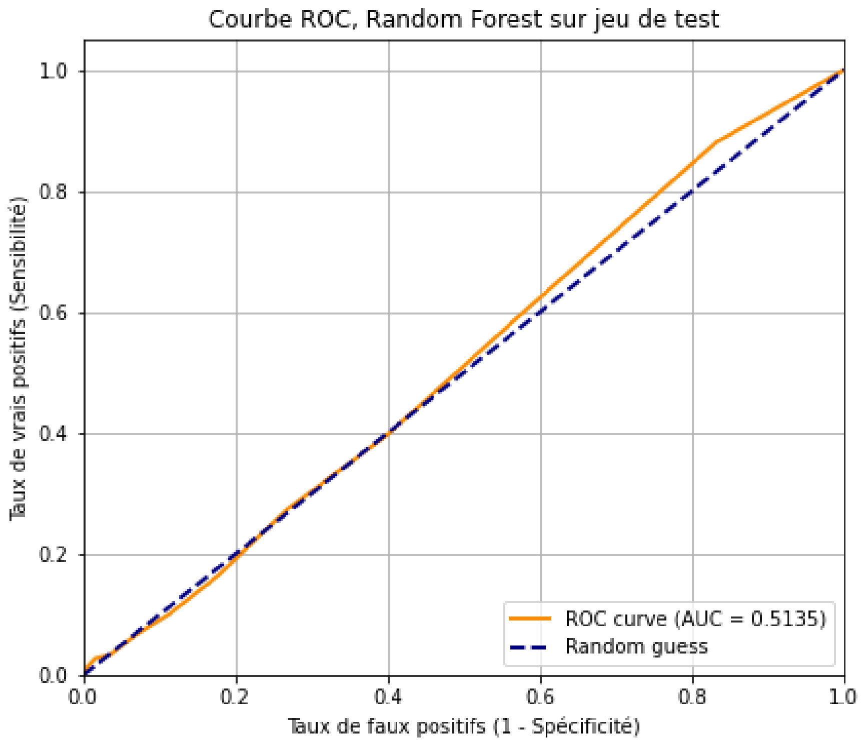


XGBoost model

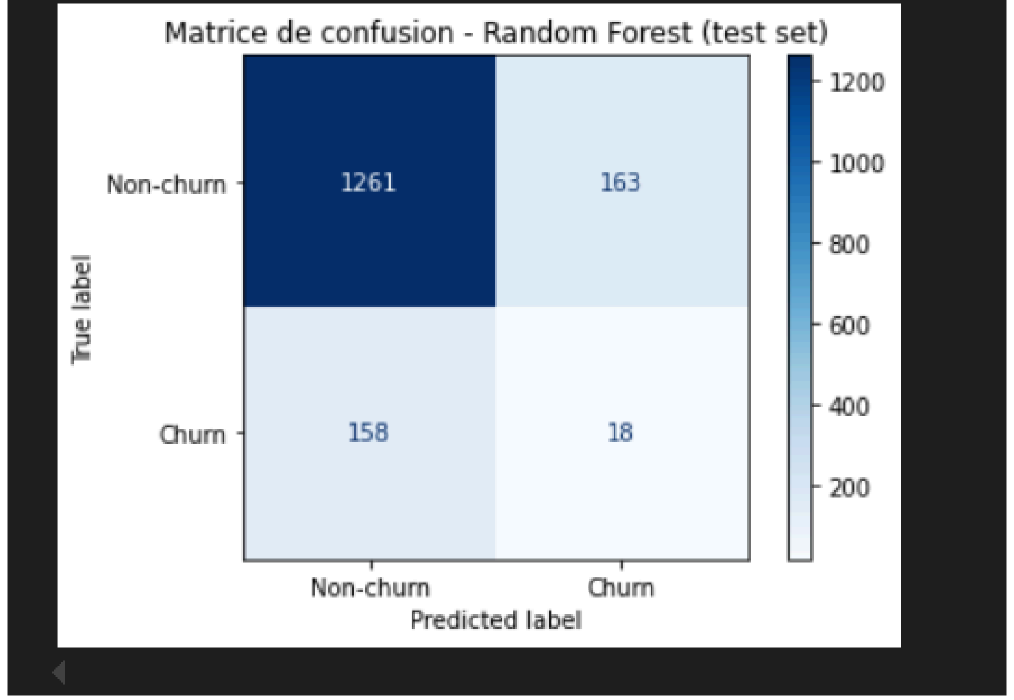
```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.85      0.87      1424
           1       0.16      0.24      0.20       176

    accuracy                           0.78      1600
   macro avg       0.53      0.55      0.53      1600
weighted avg       0.82      0.78      0.80      1600
```



Random Forest model



```
accuracy_score = 0.799375
```



FIN