Customer Churn predict

Ngô Huy Anh

Instructions of analysis

1. Context & Objective

2. Data Information

3. Exploratory Data Analysis

4. Clustering Churn Customer with K-means

5. Model & Result

6. Evaluate Model

7. Result

8. Recommend for retention plan

9. Reference

1

Context & Objective

1. Context & Objective

Context

The Telco customer churn data contains information about a fictional telco company that provided home phone and Internet services to 7043 customers in California in Q3. It indicates which customers have left, stayed, or signed up for their service.

Objective

Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs.

2

Data Information

Data	Note
CustomerID	A unique ID that identifies each customer
Gender	Gender: Male, Female
Senior Citizen	The customer is 65 or older: Yes, No
Partner	Has a partner or not (Yes, No)
Dependents	Has dependents or not (Yes, No)
Tenure	Number of months the customer has stayed with the company
PhoneService	Customer has a phone service or not (Yes, No
MultipleLines	Has multiple lines or not (Yes, No, No phone service)
InternetService	Internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Has online security or not (Yes, No, No internet service)

Data	Note
OnlineBackup	Has online backup or not (Yes, No, No internet service)
DeviceProtection	Has device protection or not (Yes, No, No internet service)
TechSupport	Has tech support or not (Yes, No, No internet service)
StreamingMovies	Has streaming TV or not (Yes, No, No internet service)
Contract	Contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling	Has paperless billing or not (Yes, No)
PaymentMethod	Payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges	The amount charged to the customer monthly
TotalCharges	He total amount charged to the customer
Churn	Whether the customer churned or not (Yes or No)

Data Information

- Bảng dữ liệu có 7043 dòng và 21 cột. 1 cột có datatype là int64, 2 cột có dạng float64 và các cột còn lại là object.
- Cột Totalcharges là có missing values, tổng các dòng bị thiếu là 11 dòng.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7032 non-null   float64
 20  Churn             7043 non-null   object
dtypes: float64(2), int64(2), object(17)
memory usage: 1.1+ MB
```

3

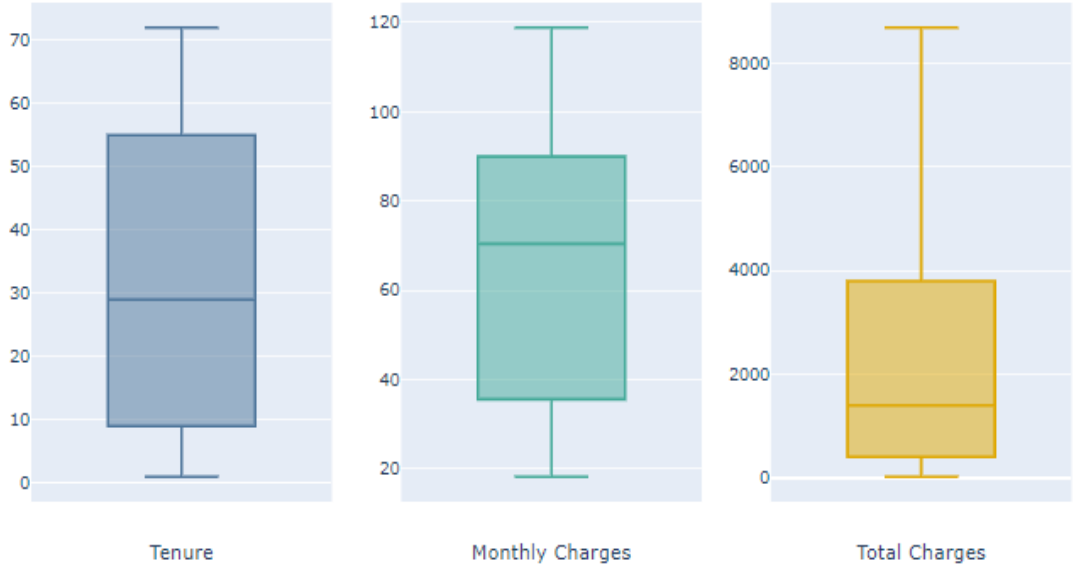
Exploratory Data Analysis

Statistics of Numerical Data

- Max: 72 month
- Median: 29 Month
- Min: 1 month

- Max: 118.75$
- Median: 70.35$
- Min: 18.25$

- Max: 8,648.8$
- Median:1,397.475$
- Min: 18.8$



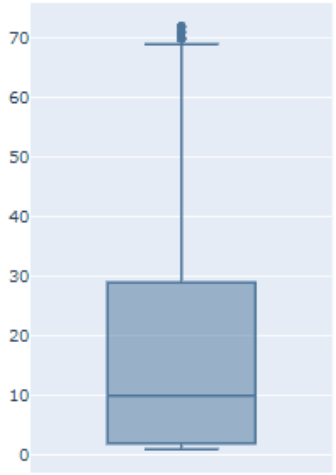
- Trung bình mọi người sử dụng dịch vụ trong 29 tháng, mỗi tháng phải trả 7.035 và tổng phải thanh toán là 1,397.457 $
- Ngoài ra thì không có các outlier trong Numerical data

Statistics of Numerical Data (with Churn = “yes”)

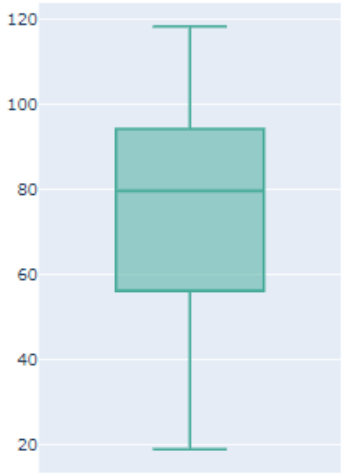
- Max: 72 month
- Median: 10 Month
- Min: 1 month

- Max: 118.35$
- Median: 79.65$
- Min: 18.85$

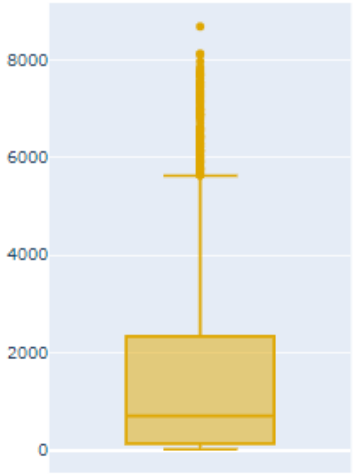
- Max: 8,648.8$
- Median:703.55$
- Min: 18.85$



Tenure



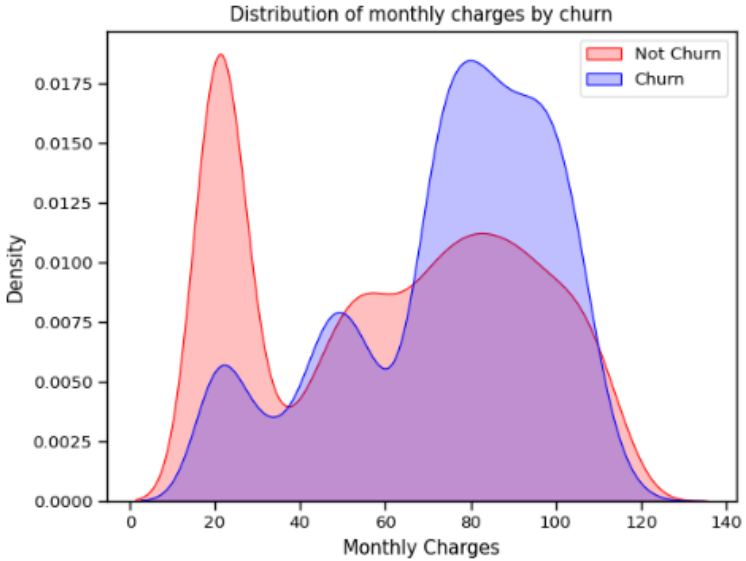
Monthly Charges

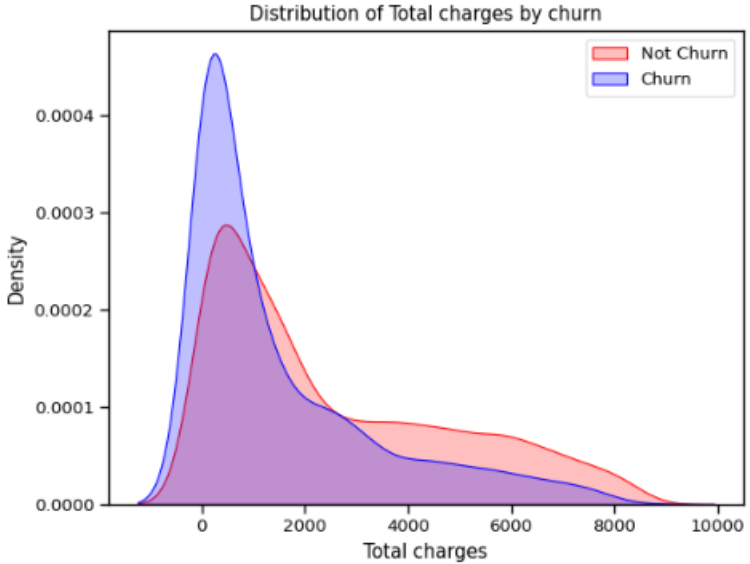


Total Charges

- Đối với những người đã Churn, thì tenure với total charges đã xuất hiện những outlier
- Tenure của những người churn là khoảng 10 tháng nhưng ngoài ra thì xuất hiện cả những người đã dùng dịch vụ > 70 tháng
- Trung bình Total charges của những người đã churn là 703.55$, ngoài ra outlier của totalcharges là những người bị charges >5500$

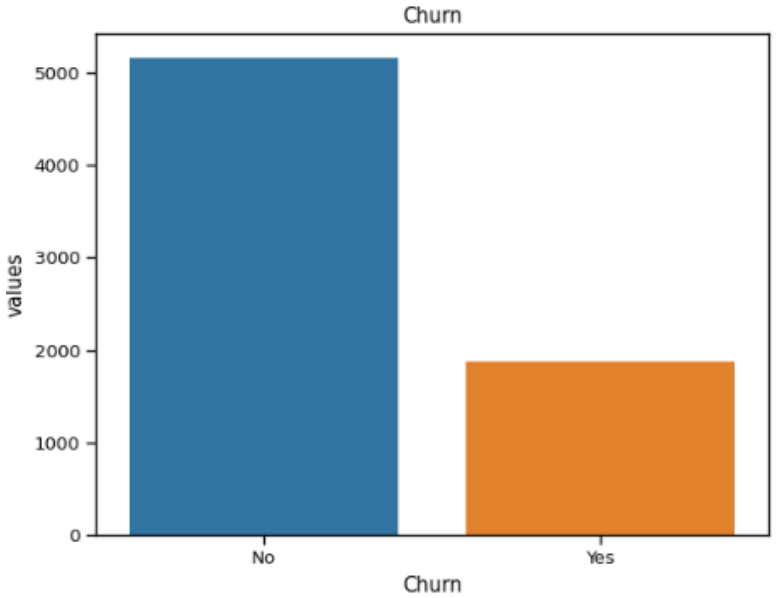
Distribution of charges by churn

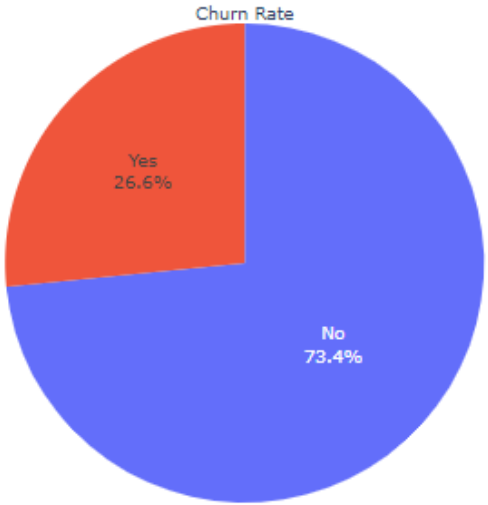




- Những người churn là những người bị charges hàng tháng cao

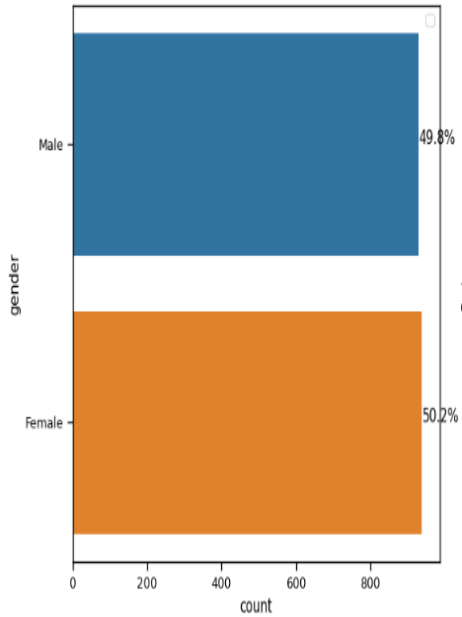
Statistics of Categorical Data

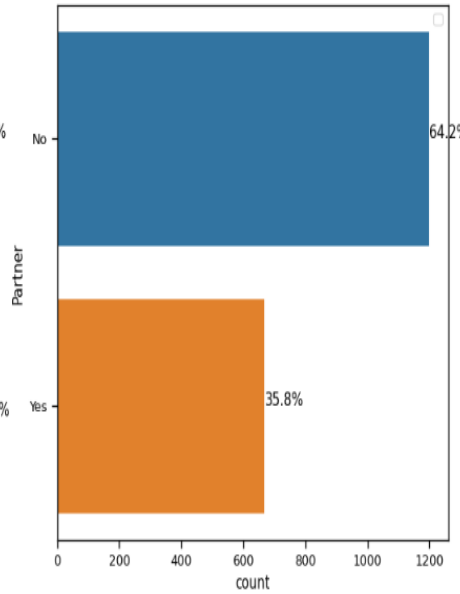


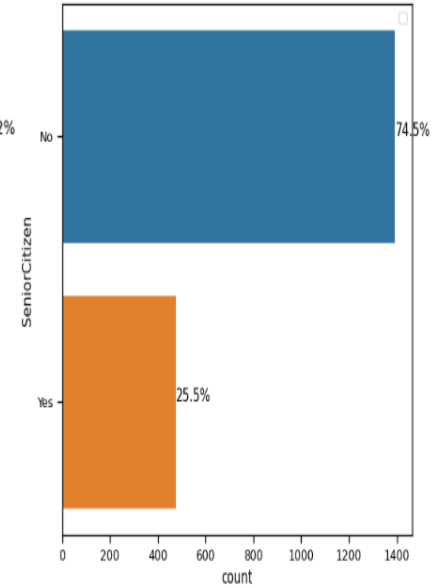


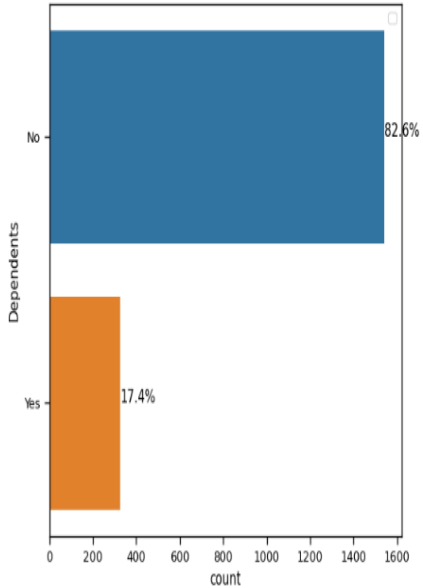
- Tỉ lệ churn rate của công ty là 26% tương đương với khoảng 1800 người trên tổng số là 7043 người

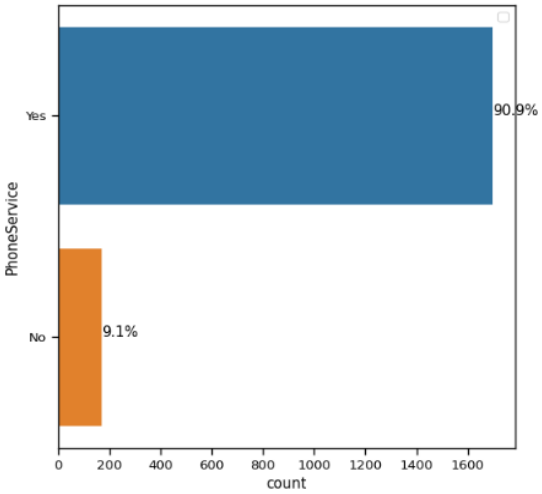
Insight Churn customer

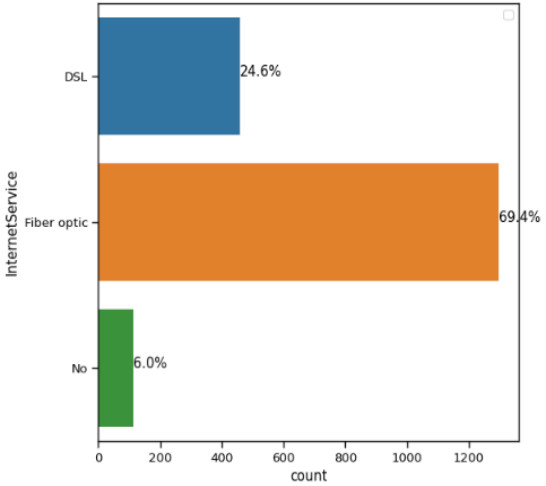


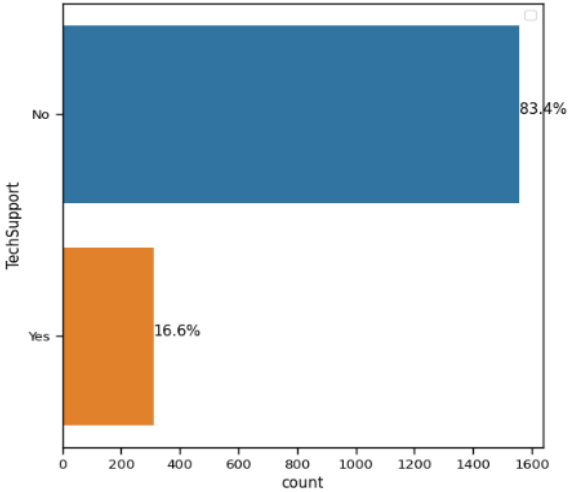


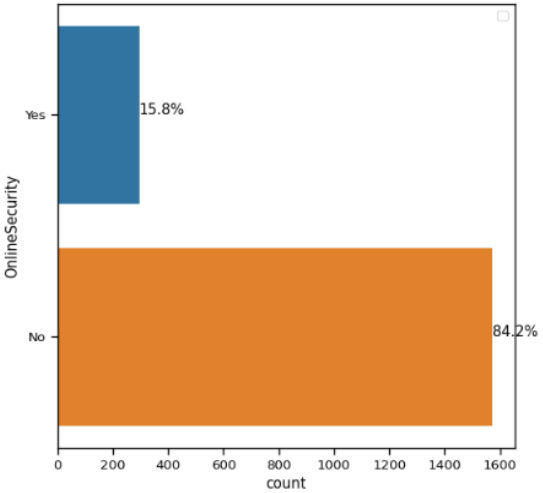


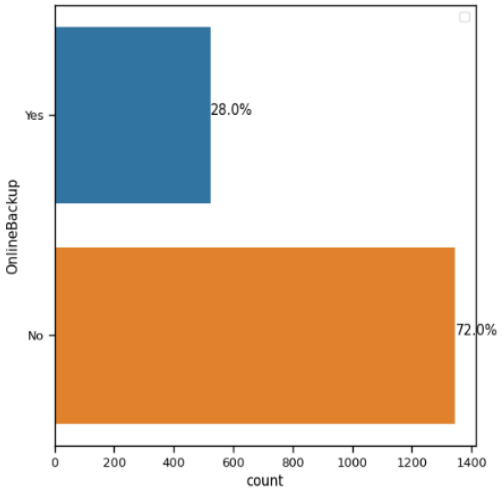


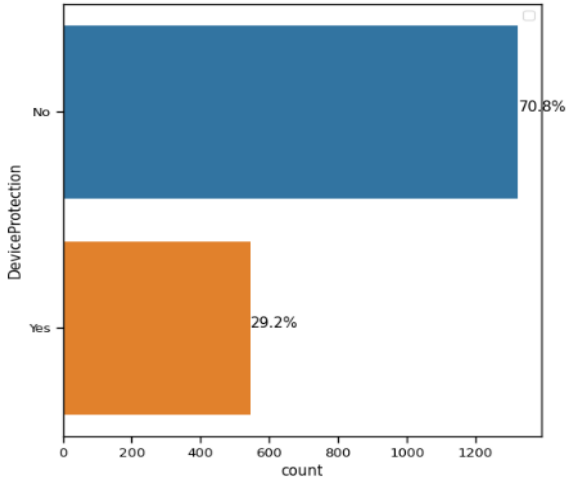

PhoneService
Yes
No
90.9%
9.1%
0
200
400
600
800
1000
1200
1400
1600
count

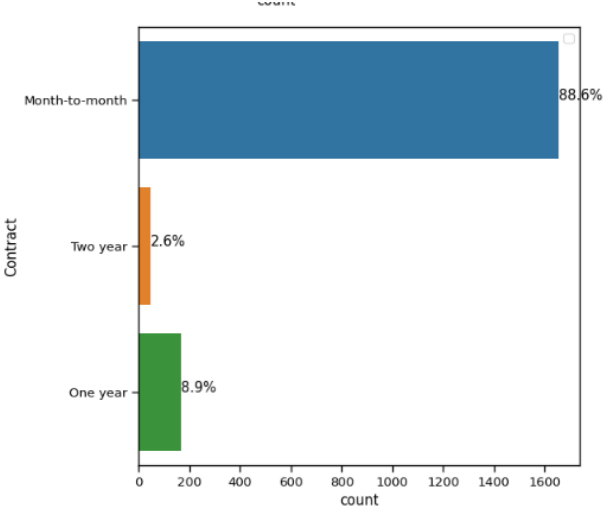

InternetService
DSL
Fiber optic
No
24.6%
69.4%
6.0%
0
200
400
600
800
1000
1200
count

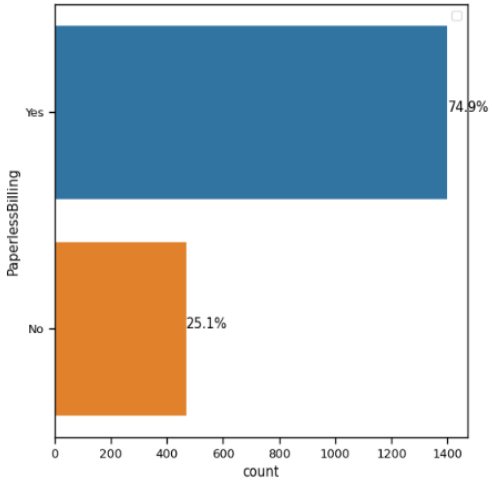

TechSupport
No
Yes
83.4%
16.6%
0
200
400
600
800
1000
1200
1400
1600
count

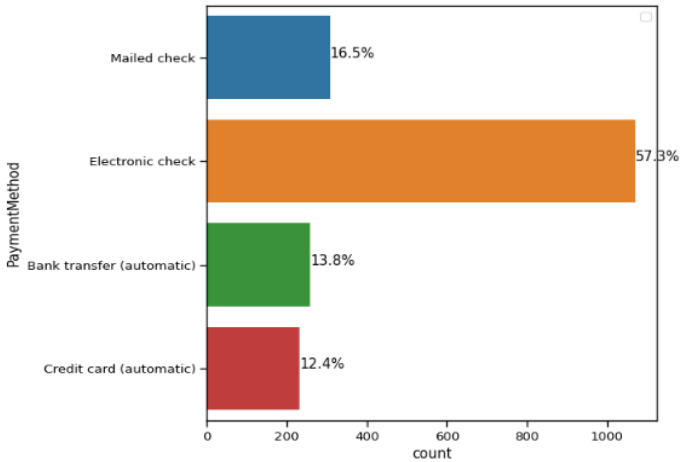

OnlineSecurity
Yes
No
15.8%
84.2%
0
200
400
600
800
1000
1200
1400
1600
count


OnlineBackup
Yes
No
28.0%
72.0%
0
200
400
600
800
1000
1200
1400
count


DeviceProtection
No
Yes
70.8%
29.2%
0
200
400
600
800
1000
1200
count







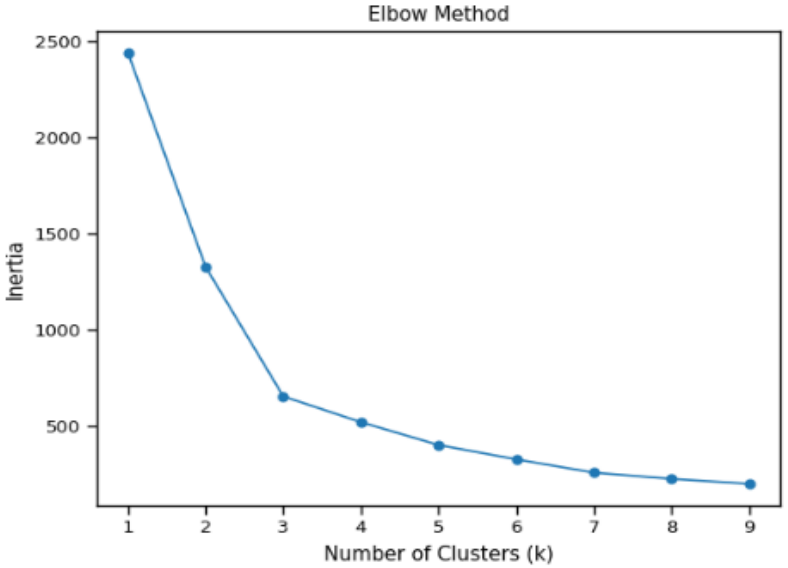
- Khách hàng đã churn có tỉ lệ nam nữ là tương đương nhau
- Khách hàng >65t có tỉ lệ churn cao hơn độ tuổi còn lại
- Đa phần là những người sống độc lập đều không có partner hoặc có dependents
- Có sử dụng Phone Service, và Internet Service trong đó dịch vụ Fiber optic là nhiều nhất
- Các dịch vụ khác như DeviceProtection, techsupport, Online Security hay online Backup đều không sử dụng
- Hợp đồng phần lớn là 1 tháng, có hóa đơn và cách thức thanh toán là Electronic Check

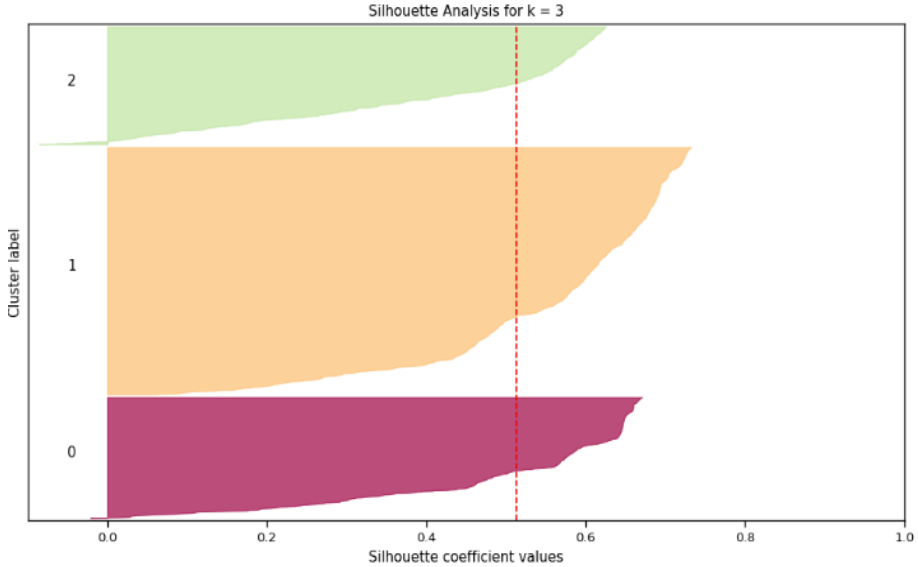
4

Clustering Churn Customer

Model apply: K-means

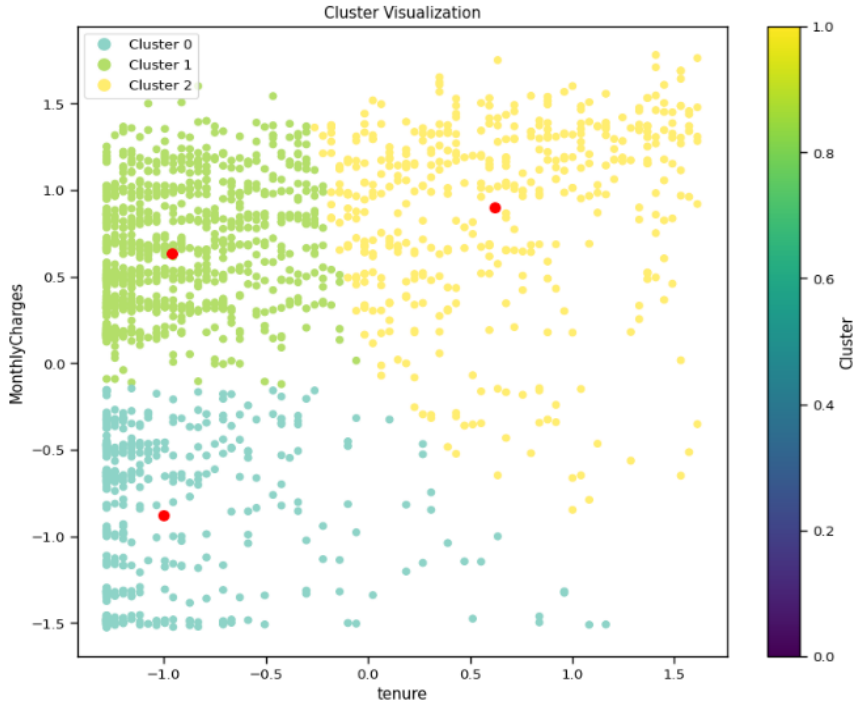
Clustering Churn Customer

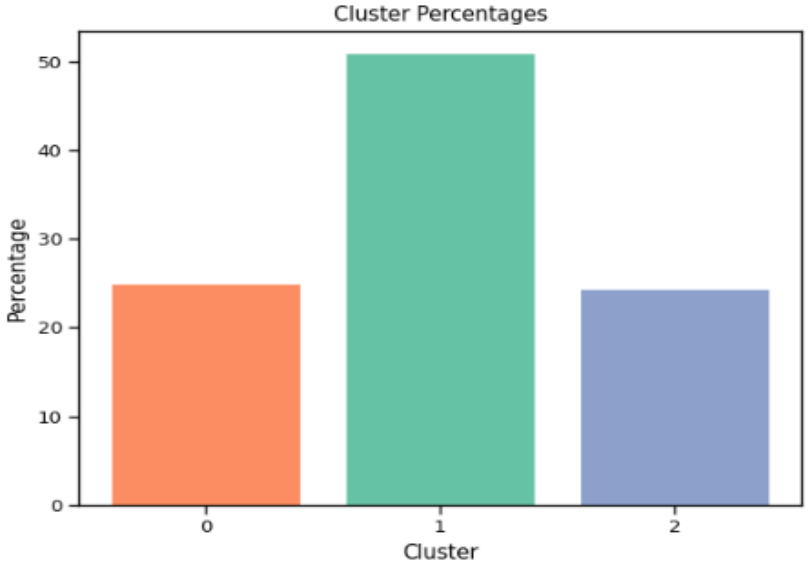




- Sau khi dùng phương pháp Elbow và Silhoute cho tập churn customer thì có được K tối ưu = 3

Clustering Churn Customer





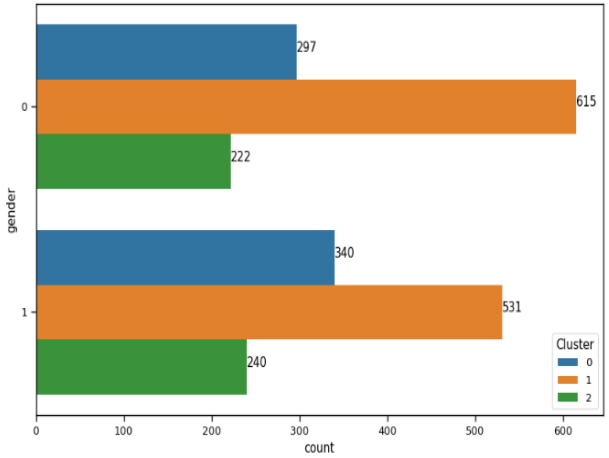
- Sau khi phân cụm để tìm insight sâu hơn về nhóm khách hàng đã churn thì ta có 3 cụm khách hàng như sau:

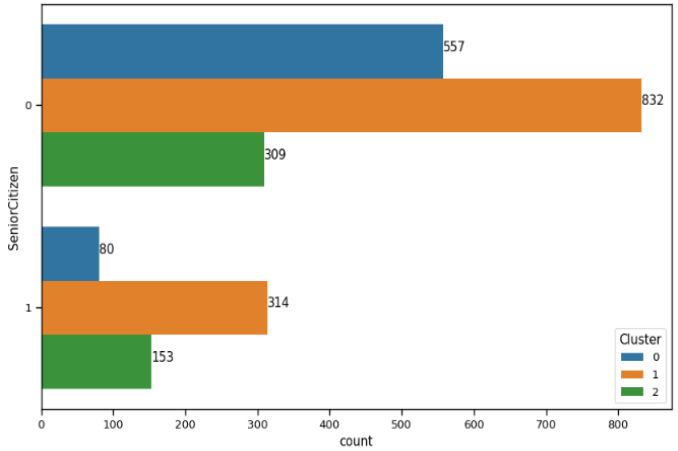
+ **Cluster 0**: Low Monthly Charges và low Tenure

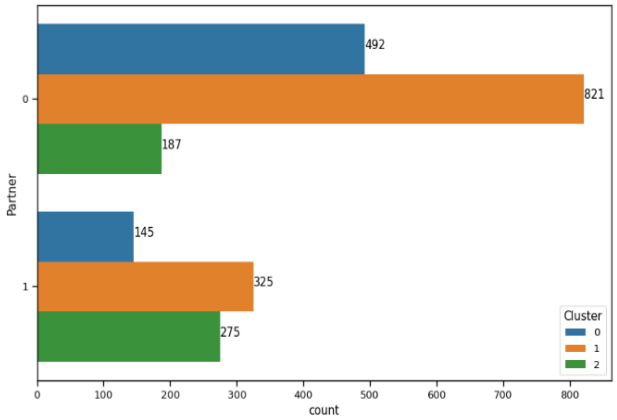
+ **Cluster 1**: High Monthly Charges và low Tenure

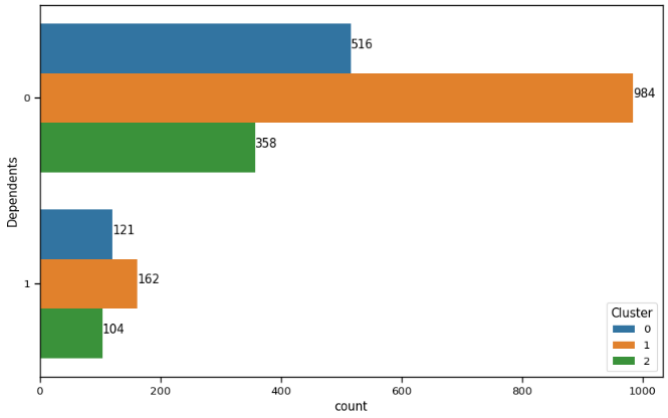
+ **Cluster 2**: high Monthly Charges và High tenure

- Trong đấy nhóm Cluser 1 là nhiều nhất, nhóm cluster 0 và cluster 2 là tương đương nhau

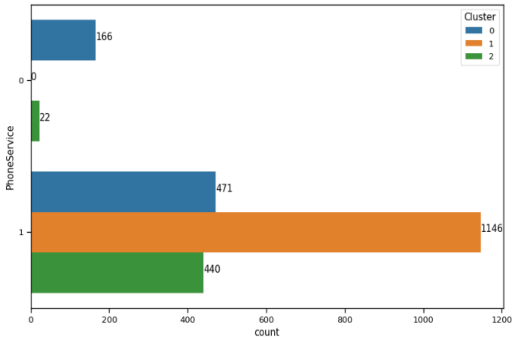


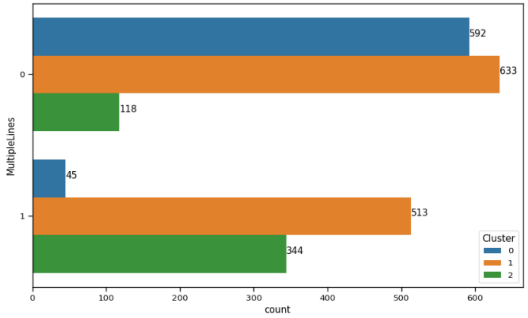


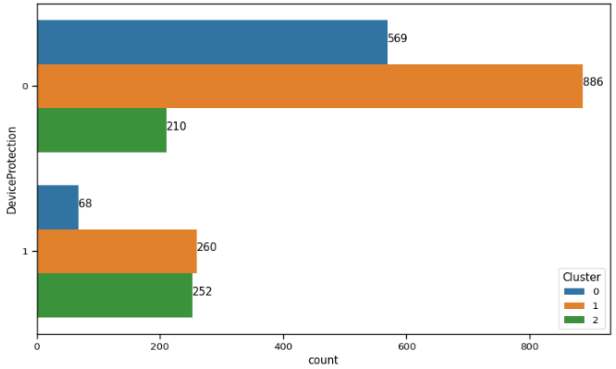


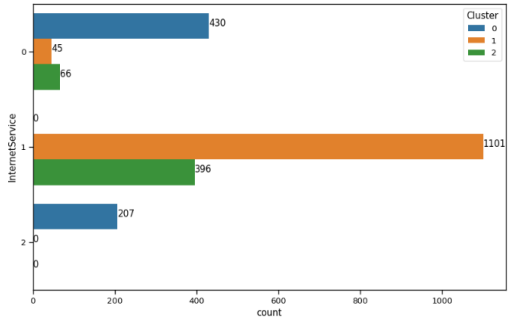


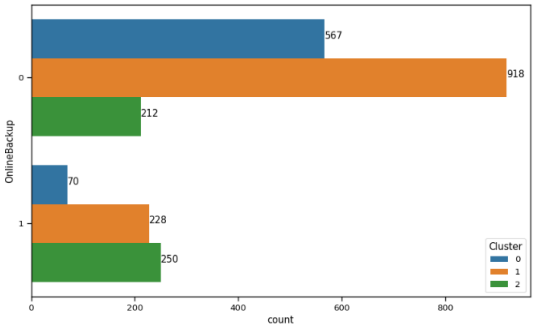
- Phần lớn vẫn là các khách hàng ở trong mỗi cụm đều không là senior citizen
- Cluster 1 thì Female nhiều hơn male
- tỉ lệ female và male ở cluster 2 thì tương đương nhau
- cluster 0 thì male nhiều hơn female

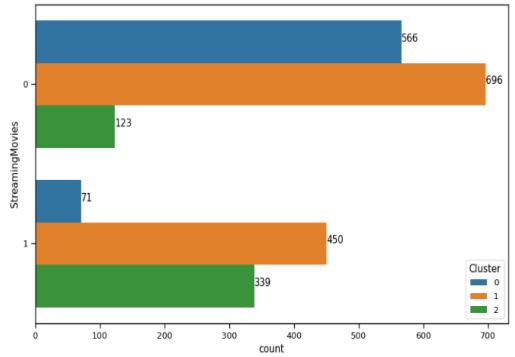


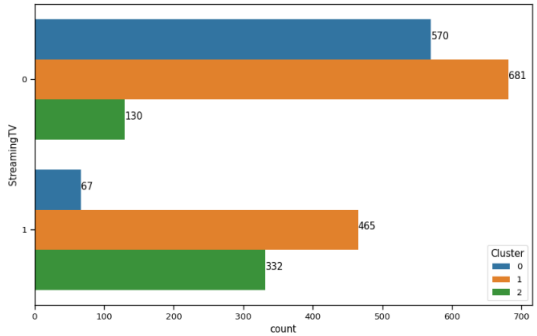




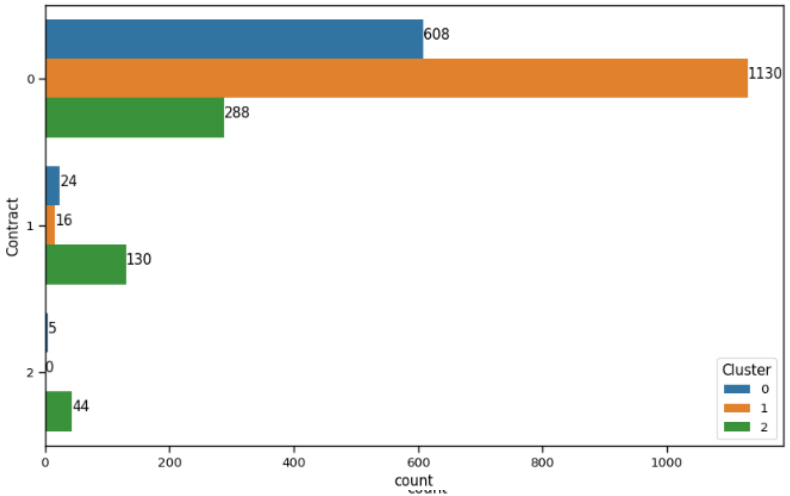


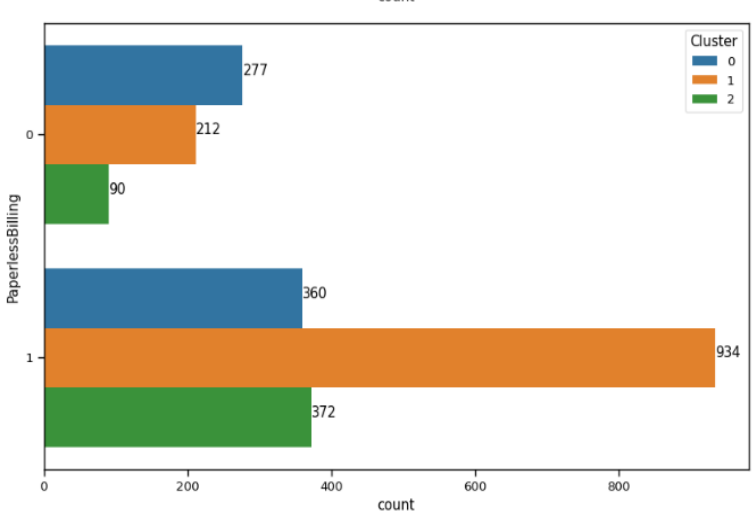


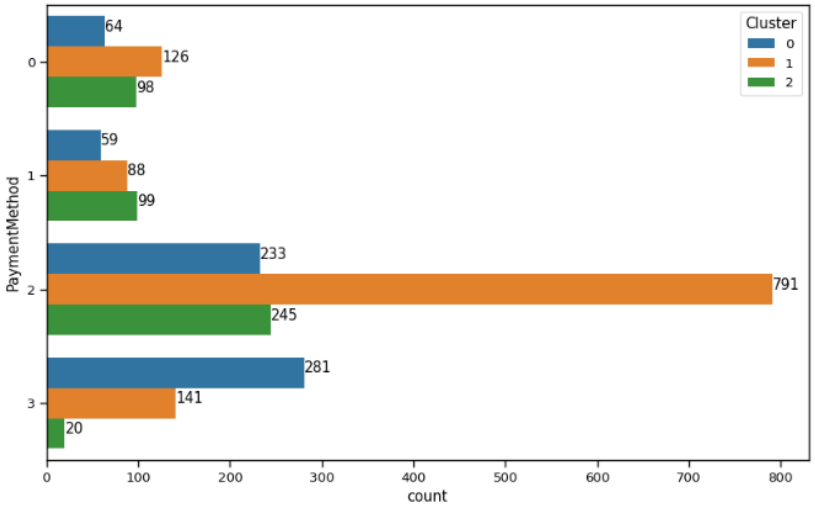




- Các nhóm đều sử dụng phone service và internet service (Fiber Optic)
- Có thể thấy là nhóm Cluster 2 (High monthly Charges và High tenure) có xu hướng dung các dịch vụ stream hoặc để bảo vệ thiết bị nhiều hơn







- Contract phần lớn của các cluster vẫn là month – to – month, đối với hợp đồng 2 year thì chỉ duy nhất có nhóm cluster 2
- đối với cluster 1 và 2 thì payment chủ yếu vẫn là Electronic check, nhưng nhóm cluster 3 thì có xu hướng dung mailcheck

	Cluster 0 (Low Monthly Charges and low tenure)	**Cluster 1** (high monthly charges and low tenure)	**Cluster 2** (high monthly charges and high tenure)
Demographic	- Male - Chưa có partners hoặc chưa có dependents	- Female - Đều có partner hoặc dependents	- male/female - Có partner nhưng chưa có dependents
Service	- Có dùng phone service - Không dùng Multiplelines - Dùng Dsl	- Có dùng phone service - Không dùng Multiplelines - Dùng fiber optic	- Có dùng phone service - Có dùng multiplelines - Dùng fiber Optic
Serivce backup	- Không dùng cách dịch vụ như device protection, online backup	- Không dùng cách dịch vụ như device protection, online backup	- Dùng cách dịch vụ như device protection, online backup
Serivce stream	- Không dùng các dịch vụ Stream	- Không dùng các dịch vụ Stream	- Có dùng dịch vụ Stream
Contract and paymenmethods	- month – to – month - Mailcheck và Electronick check	- month – to – month - Electronick check	- month – to – month - Electronick check

5

Model and result

Model and Result

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Adaboost
5. GardienBoosting
6. XGBoosting
7. LightGBM

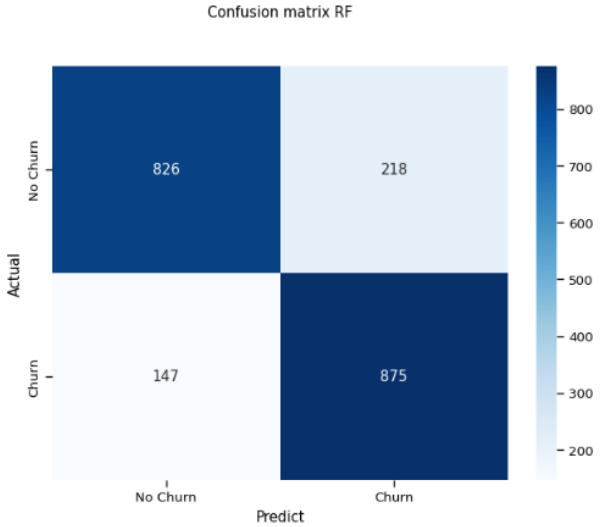
-> 3 mô hình có số cao nhất là RF,XGB, Light

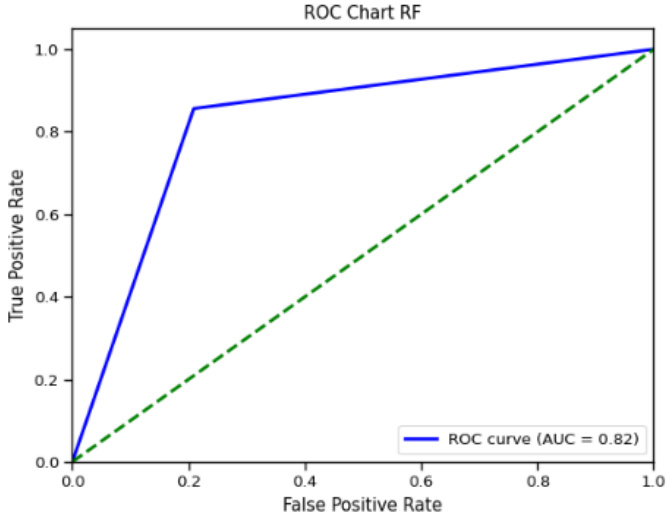
Model	Accuracy	F1_score	AUC
Lr	0.77	0.78	0.77
DT	0.78	0.79	0.78
RF	0.82	0.83	0.82
Gar	0.8	0.8	0.8
Ada	0.82	0.81	0.82
XGB	0.83	0.83	0.83
Light	0.82	0.82	0.82

6

Evaluate Model

Evaluate Model



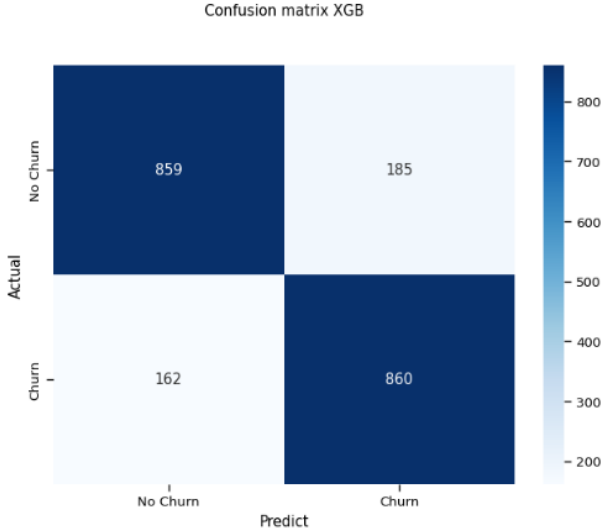


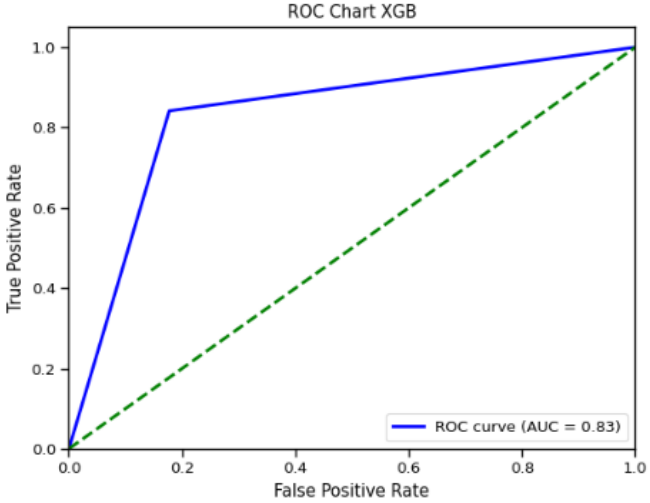
FNR = 20.8% -> Tỉ lệ dự báo sai của mô hình là 20.8%

FPR = 14.38% -> Tỉ lệ dự báo nhầm của mô hình là 14.38%

AUC = 0.82

Result



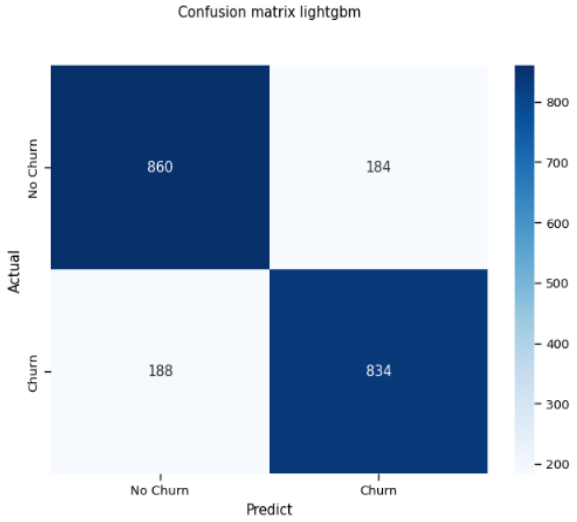


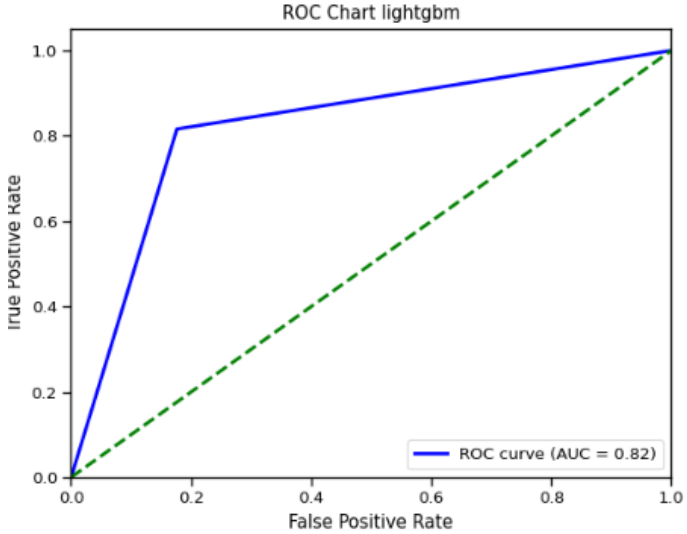
FNR = 17.7% -> Tỉ lệ dự báo sai của mô hình là 17.7%

FPR = 15.85% -> Tỉ lệ dự báo nhầm của mô hình là 15.85%

AUC = 0.82

Evaluate Model





FNR = 17.6% -> Tỉ lệ dự báo sai của mô hình là 17.6%

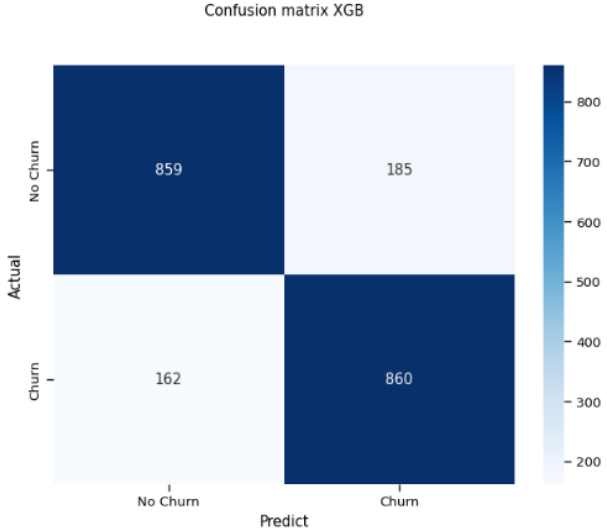
FPR = 18.39% -> Tỉ lệ dự báo nhầm của mô hình là 18.39%

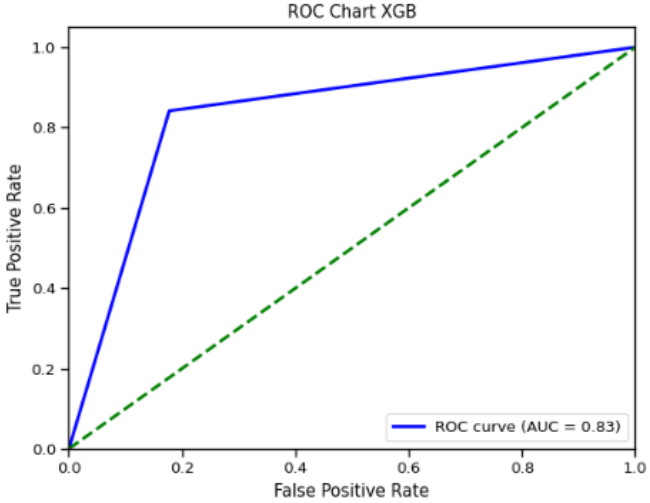
AUC = 0.82

7

Result

Result





FNR = 17.7% -> Tỉ lệ dự báo sai của mô hình là 17.7%

FPR = 15.85% -> Tỉ lệ dự báo nhầm của mô hình là 15.85%

AUC = 0.82

=> mô hình XGB có tỉ lệ FNR và FPR thấp hơn và ổn định hơn 2 model kia nên sẽ chọn model XGB

8

Recommend for Retention plan

Cluster 1

	Cluster 1 (high monthly charges and low tenure)
Demographic	- Female - Đều có partner hoặc dependents
Service	- Có dùng phone service - Không dùng Multiplelines - Dùng fiber optic
Serivce backup	- Không dùng cách dịch vụ như device protection, online backup
Serivce stream	- Không dùng các dịch vụ Stream
Contract and paymenmethods	- month – to – month - Electronick check

- Cluster 1 : họ là những người bị charges phí cao trong khi họ có partners và dependents, và chưa sử dụng các gói dịch vụ khác

-> Làm các chương trình giảm giá hợp lý cho nhóm này, tặng các gói dịch vụ dùng thử để tăng trải nghiệm và sự tin tưởng của khách hàng

Cluster 0

	Cluster 0 (Low Monthly Charges and low tenure)
Demographic	- Male - Chưa có partners hoặc chưa có dependents
Service	- Có dùng phone service - Không dùng Multiplelines - Dùng Dsl
Serivce backup	- Không dùng cách dịch vụ như device protection, online backup
Serivce stream	- Không dùng các dịch vụ Stream
Contract and paymenmethods	- month – to – month - Mailcheck và Electronick check

- Cluster 0 : họ là những người bị charges thấp và có số tenure rất thấp, khả năng đây là nhóm người dùng thử

-> Đối với nhóm này vì có thể là đa phần dùng dịch vụ DSL (không được tối ưu), nên giữ chân nhóm khách hàng này bằng cách giảm giá hoặc tặng các dịch vụ cho các hợp đồng dài hạn

Cluster 2

	Cluster 2 (high monthly charges and high tenure)
Demographic	- male/female - Có partner nhưng chưa có dependents
Service	- Có dùng phone service - Có dung multiplelines - Dùng fiber Optic
Serivce backup	- Dùng cách dịch vụ như device protection, online backup
Serivce stream	- Có dùng dịch vụ Stream
Contract and paymenmethods	- month – to – month - Electronick check

- Cluster 2 : là người có monthly charges và high tenure, đối với nhóm cluster này, thì là những người đều đã trải qua hết tất cả các dịch vụ, và là người gắn bó lâu dài

-> Nên áp dụng chương trình Customer loyalty để có những chương trình ưu đãi, các dịch vụ dung thử các gói dịch vụ mới

-> Đối với khách hàng dùng lâu dài, thiếp lập chương trình Cá nhân hóa chăm sóc khách hàng

9

Reference

Reference

1. Data source: https://www.kaggle.com/datasets/blastchar/telco-customer-churn
2. https://www.questionpro.com/blog/customer-churn/

THANKS!

Any questions?