ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN-TIN HỌC

BÁO CÁO

NHẬP MÔN KHOA HỌC DỮ LIỆU

ĐỀ TÀI: Customer Churn Prediction Using Logistic Pregression

Giảng viên giảng dạy : Hà Văn Thảo
Nhóm thực hiện : Nhóm 14
Lớp : 20KDL1
Thành viên:

1	Nguyễn Quốc Tiến	20280098
2	Nguyễn Ngọc Phương Trang	20280104
3	Nguyễn Thị Thu Thảo	20280087
4	Phạm Minh Trí	20280106

Ngành: Khoa Học Dữ Liệu

Thành phố Hồ Chí Minh-2022

Mục lục

I. Phân tích dữ liệu....3
II. Đánh giá và tạo mô hình....4
1. Đánh giá....4
2. Xử lý dữ liệu....8
3. Xây dựng mô hình....14
4. Biến đặc trưng quan trọng nhất....15
III. Kết luận....16
IV. Tài liệu tham khảo....16

I. Phân tích dữ liệu

Mô tả vấn đề

- Customer Churn (Tỷ lệ khách hàng rời đi) được hiểu là phần trăm khách hàng đã ngừng sử dụng sản phẩm hoặc dịch vụ của công ty bạn trong một khung thời gian nhất định.
- Trong trong thời đại kinh doanh mang tính cạnh tranh cao, khách hàng là yếu tố quan trọng của công ty. Lượng khách hàng ổn định là chìa khóa thành công của bất kì doanh nghiệp nào. Số lượng khách hàng của công ty có thể bị thất thoát do công ty cạnh tranh đưa ra lời chào tốt hơn công ty trước hoặc cũng có thể bởi nhiều lí do khác. Doanh nghiệp cố gắng làm hài lòng giữ chân họ càng lâu càng tốt, vì chi phí để có được một khách hàng mới có thể tiêu tốn gấp 10 lần chi phí để giữ chân khách hàng hiện tại. Do đó Customer Churn là một trong những thước đo quan trọng để các công ty đánh giá được hiệu quả hoạt động của họ. Các công ty luôn cố gắng giảm tỷ lệ churn xuống gần bằng 0%.

Vai trò của phân tích dữ liệu

Phân tích dữ liệu Churn của khách hàng có thể giúp công ty hiểu được những lý do cơ bản khiến khách hàng có thể chọn rời khỏi công ty. Bằng cách triển khai các kỹ thuật phân tích dự đoán và áp dụng chúng vào dữ liệu khách hàng hiện tại từ hồ sơ, có thể hiểu được lí do khách hàng chuyển đổi hoặc ngừng sử dụng dịch vụ. Sau đó có thể làm việc với những khách hàng có xác suất chuyển đổi cao để đảm bảo rằng họ vẫn ở lại với nhà cung cấp hiện tại.

Sau đây chúng em sẽ thực hiện bài phân tích với dữ liệu được Kaggle cung cấp. Đây là tập dữ liệu từ một công ty viễn thông(Telcom), dự đoán khách hàng rời đi dựa trên thông tin nhân khẩu học, hành vi sử dụng và tài khoản. Mục tiêu chính là phân tích hành vi khách hàng và phát triển các chiến lược để tăng khả năng giữ chân khách hàng.

Tập dữ liệu gồm

Trong tập dữ liệu này có 7043 hàng (mỗi hàng đại diện cho một khách hàng duy nhất) với 21 cột: 19 tính năng, 1 tính năng mục tiêu (Churn). Dữ liệu bao gồm cả tính năng số và phân loại, vì vậy chúng ta sẽ cần giải quyết từng kiểu dữ liệu tương ứng.

Tính năng phân loại

- CustomerID
- Gender – M/F
- SeniorCitizen: Khách hàng có phải là người cao tuổi hay không (1, 0)
- Partner: Khách hàng có đối tác không (yes, no)
- Dependents: Khách hàng có người phụ thuộc không (yes, no)
- PhoneService: Khách hàng có sử dụng dịch vụ điện thoại không(yes, no)
- MultipeLines: (yes, no, no phone service)
- Internet service: loại dịch vụ internet của khách hàng (DSL, (Fiber optic)cáp quang, không sử dụng)

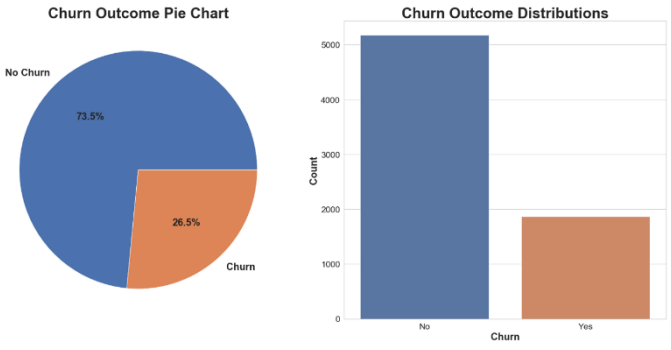
- OnlineSecurity: Khách hàng có bổ sung tiện ích bảo mật trực tuyến không(yes, no, no internet service)
- OnlineBackup: Khách hàng có bổ sung tiện ích sao lưu trực tuyến không (yes, no, no internet service)
- DeviceProtection: Khách hàng có bổ sung tiện ích bảo vệ thiết bị không (yes, no, no internet service)
- TechSupport: Khách hàng có bổ sung tiện ích hỗ trợ kỹ thuật không (yes, no, no internet service)
- StreamingTV: Khách hàng có sử dụng truyền hình trực tuyến không (yes, no, no internet service)
- StreamingMovies: Khách hàng có sử dụng phim trực tuyến không (yes, no, no internet service)
- Contract: Thời hạn hợp đồng (Two year, One year, Month-to-month)
- PaperlessBilling: Khách hàng có thanh điện tử không. (yes, no)
- PaymentMethod: Phương thức thanh toán (E-Check, Mailed Check, Bank Transfer (Auto), Credit Card (Auto))

Tính năng số

- Tenure: Số tháng khách hàng đã gắn bó với công ty.
- Monthly charges: Số tiền hàng tháng khách hàng phải trả.
- Total charges: Tổng số tiền khách hàng phải trả.

Mục tiêu

- **Churn:** Khách hàng có ý định rời bỏ công ty hay không(yes, no)



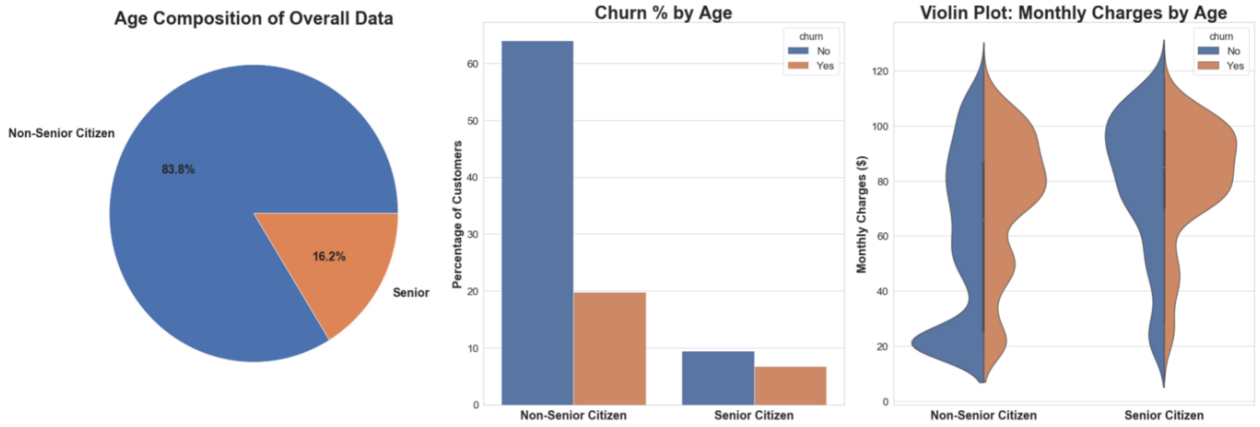
Ở biểu đồ hình tròn bên trái, khoảng 27% khách hàng từ tập dữ liệu rời đi. Một tỉ lệ khá cao.

II. Đánh giá và tạo mô hình

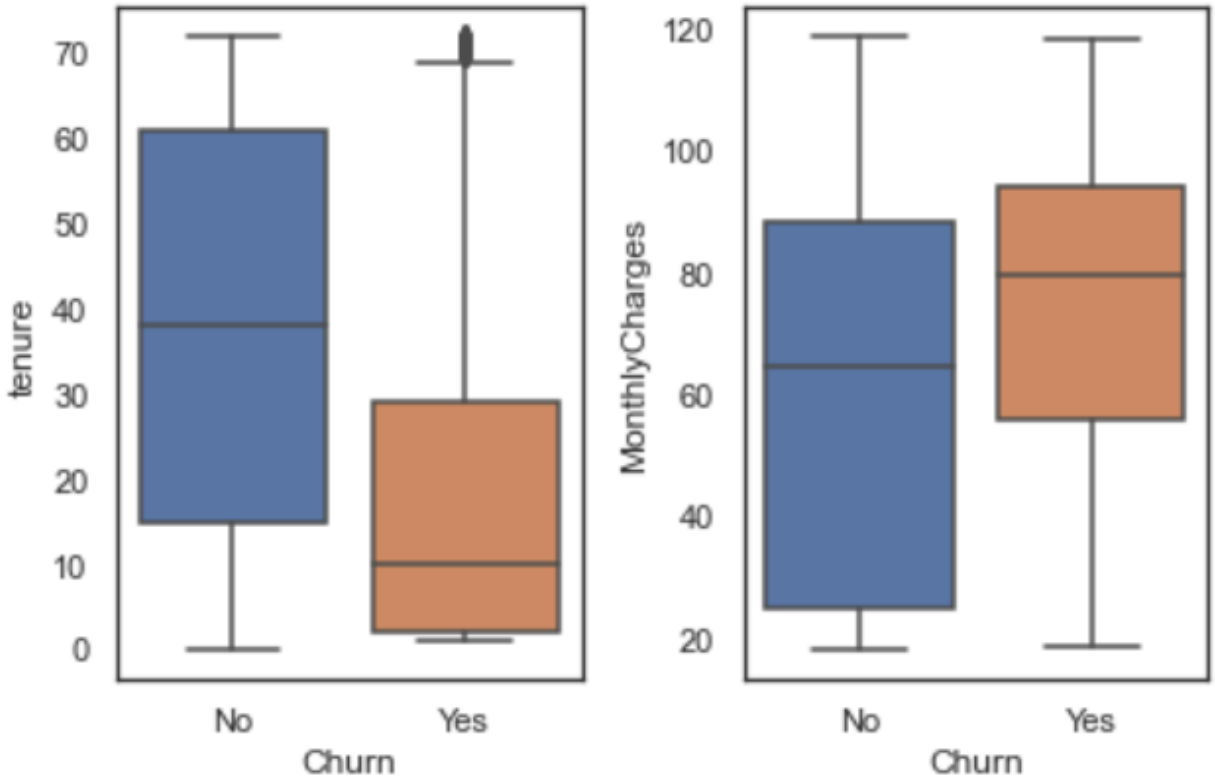
1. Đánh giá

Khi làm việc với tính năng số, trước tiên phải xem xét sự phân bố của dữ liệu. Chúng ta có thể sử dụng thư viện seaborn để trực quan hóa và kiểm tra tập dữ liệu.

Từ các hình ảnh ta rút ra kết luận trước tiên:

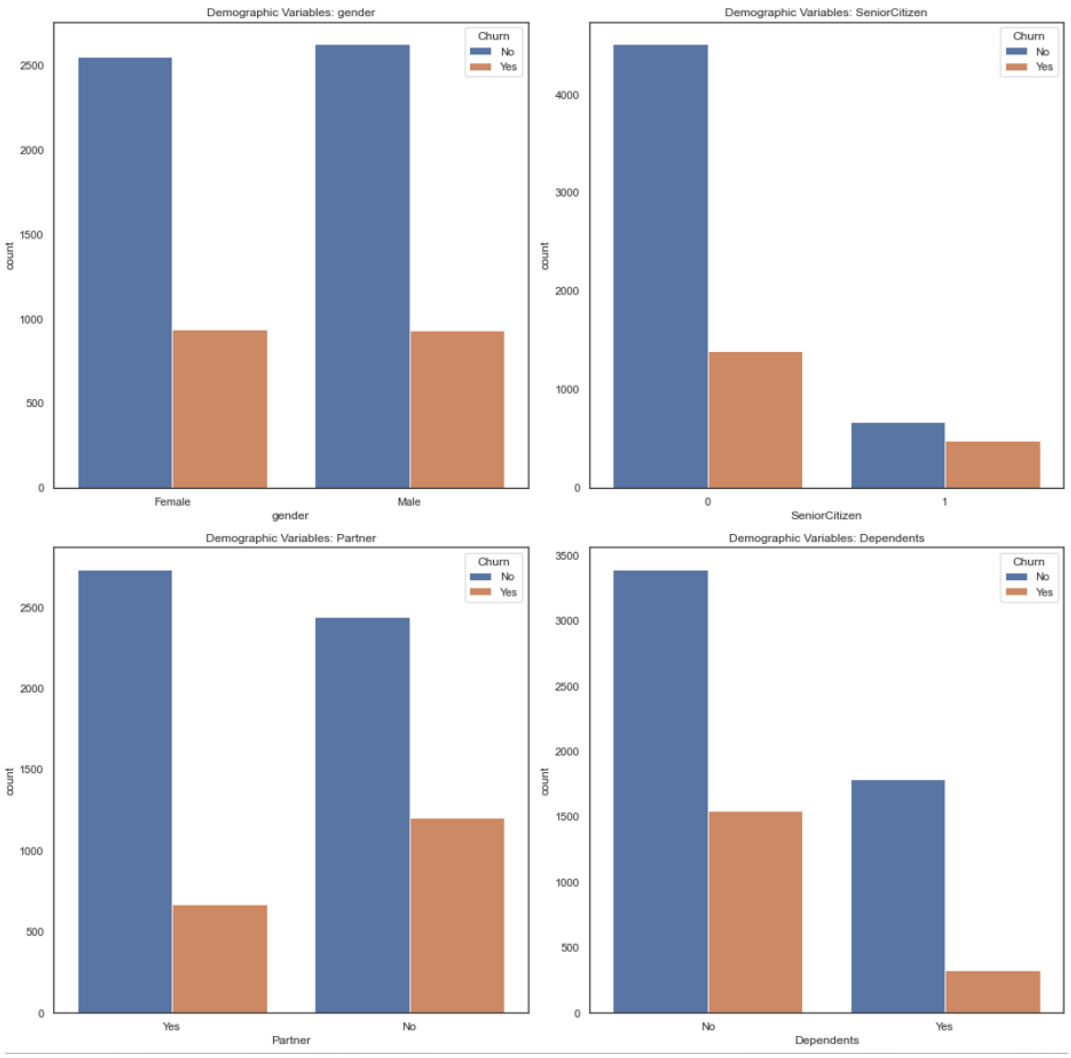


- Khách hàng là người cao tuổi có khả năng sử dụng dịch vụ Telcom hơn. Người cao tuổi và người trẻ có tỷ lệ churn cao khi phí hàng tháng lên hơn 60$.



- **Nhận xét:**

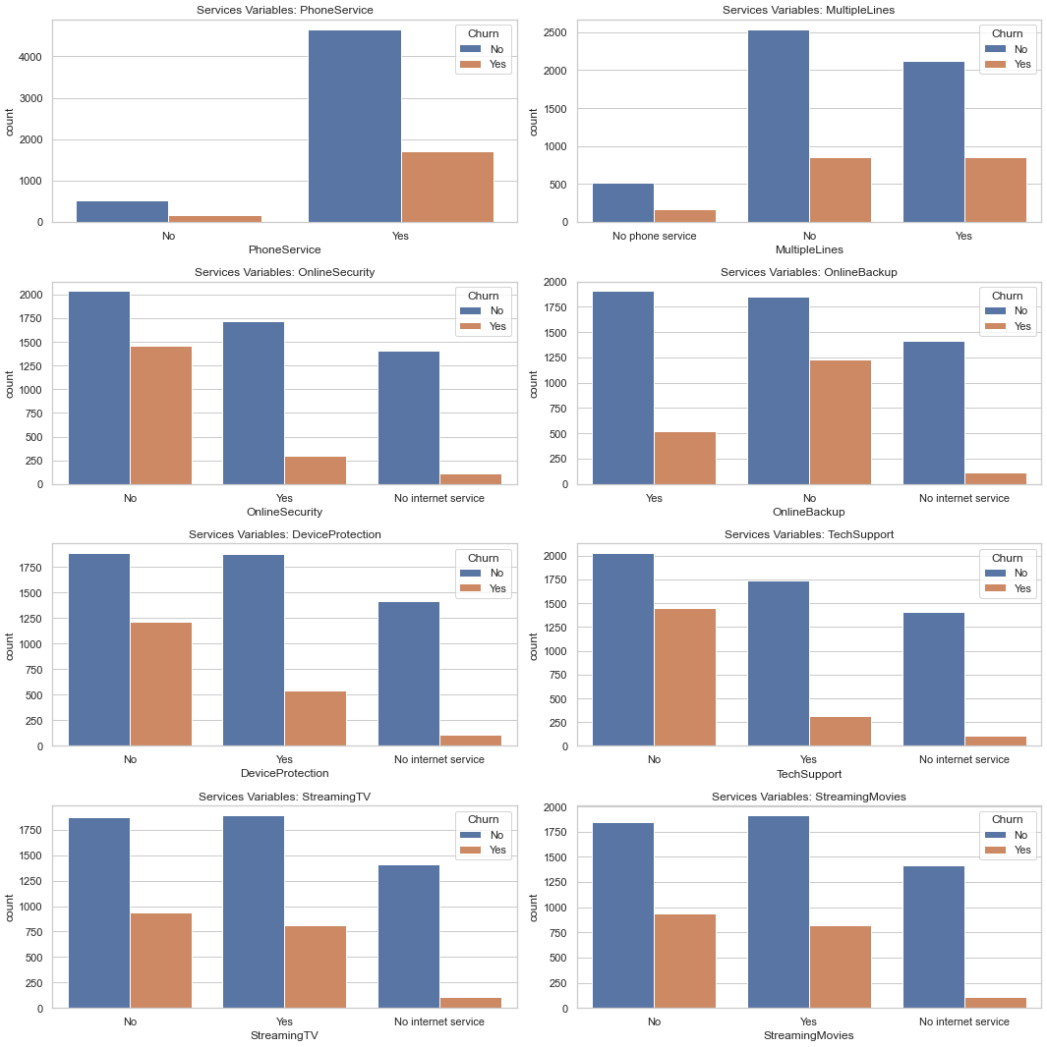
  - Biểu đồ bên trái tỷ lệ Churn(yes, no) phụ thuộc vào tenure. Thấy được thời hạn gắn bó với công ty của khách hàng đã rời đi ngắn hơn nhiều(trong 30 tháng đầu) so với khách hàng ở lại.
  - Biểu đồ bên phải so sánh số tiền hàng tháng khách hàng phải trả so với thời gian churn. Mức phí trung bình hàng tháng của những khách hàng đã rời đi cao hơn đáng kể so với những khách hàng vẫn tiếp tục đồng hành. Điều này cho thấy rằng giảm giá và khuyến mãi sẽ có khả năng lôi kéo được khách hàng ở lại.

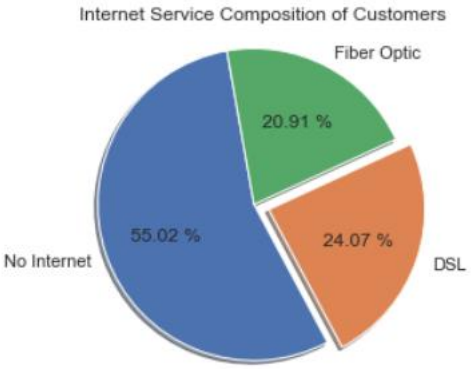


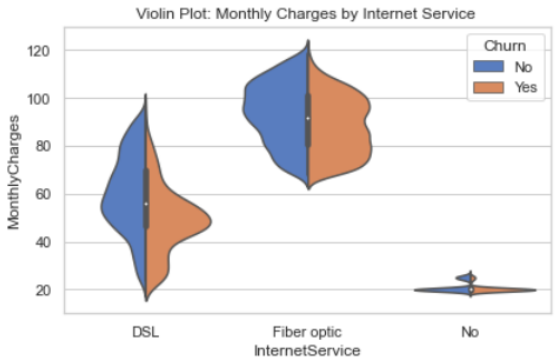
- **Nhận xét:**
  - Thống kê cho thấy ít sự khác biệt về tỷ lệ churn với giới tính. Có một tỉ lệ khách hàng rời đi cao với các biến SeniorCitizen, khách hàng có Partners và khách hàng không có Dependents.
  - Khách hàng không có Partners có tỷ lệ Churn nhiều hơn với khách hàng có Partners.

Các biến khác

- Khách hàng thuộc nhóm không có OnlineSecurity, OnlineBackup, DeviceProtection và TechSupport có tỷ lệ churn cao. Cho thấy rằng, những người có sự tin cậy cao đối với các dịch vụ của Telcom hay những người cần thiết bị cho các mục đích dử dụng riêng có tỷ lệ churn thấp hơn.
- Khách hàng sử dụng phương thức thanh toán điện tử có tỷ lệ churn cao hơn (hơn gần 15%) so với thanh toán bằng những phương thức khác.Có thể khách hàng lớn tuổi thích thanh toán bằng hóa đơn giấy hơn. Khách hàng thanh toán bằng e-check rời đi nhiều hơn 10% so với kh thanh toán bằng những hình thức khác.
- Tỷ lệ churn cao với những khách hàng có sử dụng dịch vụ điện thoại.

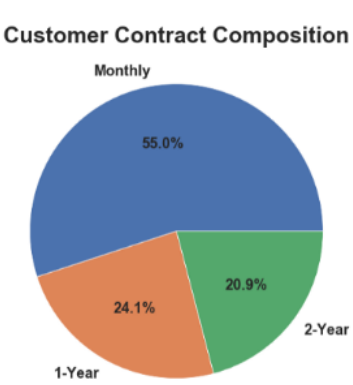


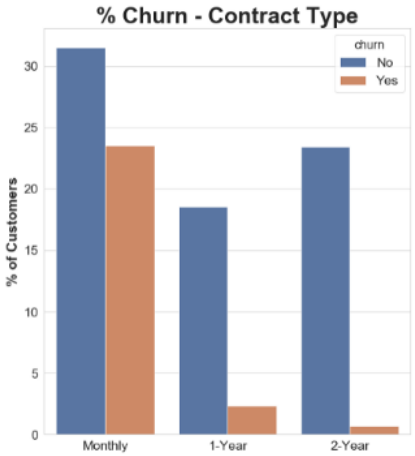


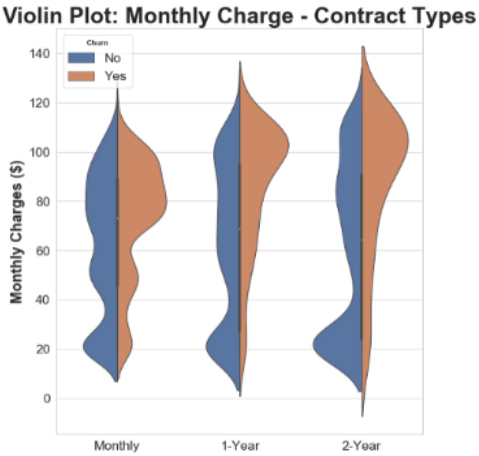


- Khách hàng sử dụng internet cáp quang có tỷ lệ churn cao.
- Cáp quang là lựa chọn internet phổ biến nhất. Khách hàng sử dụng Internet cáp quang chiếm tỷ lệ đáng kể so với khách hàng sử dụng DSL hoặc Không có Internet.

- Fiber Optic là một dịch vụ đắt hơn nhiều. Khách hàng có tỉ lệ churn cao khi chi phí từ 40$ đến 60$.







- Hơn một nữa số khách hàng chọn thanh toán hàng tháng.
- Đăng kí kì hạn càng dài thì tỷ lệ churn càng thấp.
- Phí hàng tháng thường cao hơn phí hợp đồng dài hạn.

Nhận xét chung:

- nếu khách hàng đăng kí kì hạn hợp đồng lâu dài, chọn hợp đồng một năm hoặc hai năm thay vì tùy chọn tháng này sang tháng khác và công ty đưa ra mức giá rẻ hơn, thì có thể giảm tỷ lệ khách hàng rời đi.
- nếu khách hàng là người cao tuổi, sử dụng nhiều đường truyền, sử dụng dịch vụ internet cáp quang, sử dụng phim trực tuyến, sử dụng thanh toán điện tử và sử dụng electronic check làm phương thức thanh toán, thì họ có nhiều khả năng rời đi hơn.

Những phân tích như vậy thường giúp các công ty phát hiện ra những nguyên nhân có thể xảy ra đối với sự rời bỏ của khách hàng.

Ví dụ: những khách hàng thực hiện thanh toán điện tử có nhiều khả năng bỏ cuộc hơn, có thể vì một số bất tiện mà họ gặp phải khi thực hiện thanh toán điện tử, cũng có thể là về sự không hài lòng của khách hàng đối với dịch vụ Internet Cáp quang do công ty cung cấp và công ty có thể xem xét vấn đề này và giải quyết vấn đề sớm nhất.

2. Xử lý dữ liệu

Phần xử lý dữ liệu của em gồm có:

- Bỏ những giá trị NA
- Chuyển tất cả về kiểu số
- Đưa các giá trị số đó nằm trong khoảng từ 0 đến 1
- Kiểm tra các hệ số tương quan giữa các biến.

- Khi mà làm mô hình dự đoán thì cột 'customerID' là vô tác dụng. Bởi vì 'customerID' là duy nhất

Ví dụ: dữ liệu có 10 nghìn dòng thì sẽ có 10 nghìn customerID riêng biệt nên không thể chuyển

hoá nó thành dữ liệu số được

```
# Xoá cột 'customerID'

df.drop('customerID', axis='columns', inplace=True)
```

```
# Xem tên của các cột và nó thuộc loại dữ liệu nào
# Mục đích là xem cột nào có dạng dữ liệu không đúng và sẽ chuyển đổi nó về dạng dữ liệu đúng mong muốn

df.dtypes
```

- Kiểm tra kiểu dữ liệu của các biến:

```
gender              object
SeniorCitizen        int64
Partner             object
Dependents          object
tenure               int64
PhoneService        object
MultipleLines       object
InternetService     object
OnlineSecurity      object
OnlineBackup        object
DeviceProtection    object
TechSupport         object
StreamingTV         object
StreamingMovies     object
Contract            object
PaperlessBilling    object
PaymentMethod       object
MonthlyCharges     float64
TotalCharges        object
Churn               object
dtype: object
```

- Sau khi xem dạng dữ liệu của các cột, thấy được 'TotalCharges' ở dạng 'object'. Nhưng để xây dựng mô hình thì cần ở dạng số. Vì vậy, em sẽ chuyển kiểu dữ liệu của cột này về kiểu số.
- Trước khi chuyển về kiểu số, em sẽ loại bỏ những dòng có ‘TotalCharges’ chứa giá trị Na (việc thay thế hay xoá phụ thuộc vào số lượng dòng có giá trị NaN nhiều hay ít so với dữ liệu tổng)
- Đây là các kiểu dữ liệu ở dạng ‘object’. Mục tiêu là cần chuyển ‘Yes’, ‘No’ và ‘Female’, ‘Male’ về 0, 1.

```
gender: ['Female' 'Male']
Partner: ['Yes' 'No']
Dependents: ['No' 'Yes']
PhoneService: ['No' 'Yes']
MultipleLines: ['No phone service' 'No' 'Yes']
InternetService: ['DSL' 'Fiber optic' 'No']
OnlineSecurity: ['No' 'Yes' 'No internet service']
OnlineBackup: ['Yes' 'No' 'No internet service']
DeviceProtection: ['No' 'Yes' 'No internet service']
TechSupport: ['No' 'Yes' 'No internet service']
StreamingTV: ['No' 'Yes' 'No internet service']
StreamingMovies: ['No' 'Yes' 'No internet service']
Contract: ['Month-to-month' 'One year' 'Two year']
PaperlessBilling: ['Yes' 'No']
PaymentMethod: ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
Churn: ['No' 'Yes']
```

- Các giá trị unique của OnlineSecurity và OnlineBackup có No internet service và No phone service. Em sẽ thay thế hai giá trị này thành giá trị ‘No’ luôn vì nghĩa cũng tương tự nhau.
- Các biến như InternetService, Contract và PaymenMethod làm sao để chuyển về dạng số. Em dùng ‘get_dummies’ để đưa tạo các biến giả, các biến chỉ số. Ví dụ, cột' InternetService' có 3 giá trị unique. Sau khi dùng '.get_dummies' sẽ cho ra 3 cột mới tương ứng với 3 giá trị unique đó

```
gender                                    int64
SeniorCitizen                             int64
Partner                                   int64
Dependents                                int64
tenure                                    int64
PhoneService                              int64
MultipleLines                             int64
OnlineSecurity                            int64
OnlineBackup                              int64
DeviceProtection                          int64
TechSupport                               int64
StreamingTV                               int64
StreamingMovies                           int64
PaperlessBilling                          int64
MonthlyCharges                          float64
TotalCharges                            float64
Churn                                     int64
InternetService_DSL                       uint8
InternetService_Fiber optic               uint8
InternetService_No                        uint8
Contract_Month-to-month                   uint8
Contract_One year                         uint8
Contract_Two year                         uint8
PaymentMethod_Bank transfer (automatic)   uint8
PaymentMethod_Credit card (automatic)     uint8
PaymentMethod_Electronic check            uint8
PaymentMethod_Mailed check                uint8
dtype: object
```

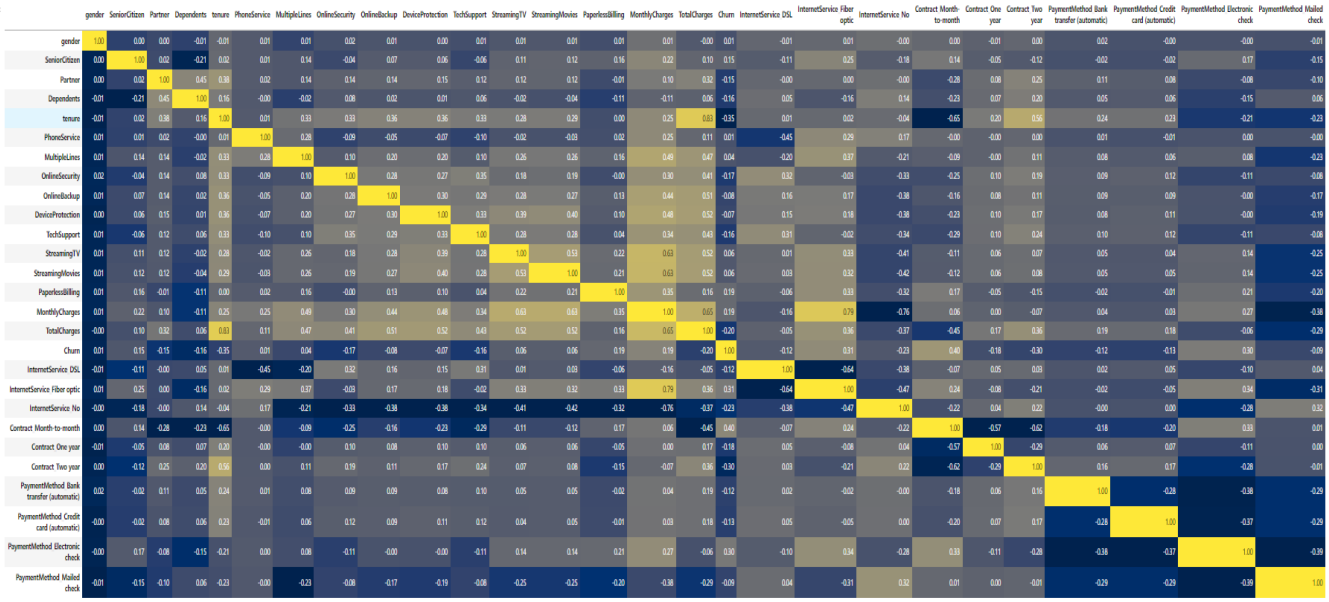
➔ Đã hoàn thành xong việc đưa các biến về kiểu số.

Thấy được, các cột 'tenure', 'MonthlyCharges', 'TotalCharges' chứa các giá trị khá lớn so với tập dữ liệu, không phù hợp để xây dựng mô hình.
Để diễn giải hiệu ứng trong hồi quy logistic, chúng ta bị ràng buộc trong phạm vi từ 0 đến 1. Em sẽ dùng 'MinMaxScaler', cho min=0, max=1 để cho nó phù hợp với tập dữ liệu.
Ví dụ: cột 'MonthlyCharges' chạy từ 0-100 thì sau khi dùng 'MinMaxScaler' thì 100 thành 1, 50 thành 0.5 ...

```
gender: [1 0]
SeniorCitizen: [0 1]
Partner: [1 0]
Dependents: [0 1]
tenure: [0.         0.46478873 0.01408451 0.61971831 0.09859155 0.29577465
 0.12676056 0.38028169 0.85915493 0.16901408 0.21126761 0.8028169
 0.67605634 0.33802817 0.95774648 0.71830986 0.98591549 0.28169014
 0.15492958 0.4084507  0.64788732 1.         0.22535211 0.36619718
 0.05633803 0.63380282 0.14084507 0.97183099 0.87323944 0.5915493
 0.1971831  0.83098592 0.23943662 0.91549296 0.11267606 0.02816901
 0.42253521 0.69014085 0.88732394 0.77464789 0.08450704 0.57746479
 0.47887324 0.66197183 0.3943662  0.90140845 0.52112676 0.94366197
 0.43661972 0.76056338 0.50704225 0.49295775 0.56338028 0.07042254
 0.04225352 0.45070423 0.92957746 0.30985915 0.78873239 0.84507042
 0.18309859 0.26760563 0.73239437 0.54929577 0.81690141 0.32394366
 0.6056338  0.25352113 0.74647887 0.70422535 0.35211268 0.53521127]
PhoneService: [0 1]
MultipleLines: [0 1]
OnlineSecurity: [0 1]
OnlineBackup: [1 0]
DeviceProtection: [0 1]
TechSupport: [0 1]
StreamingTV: [0 1]
StreamingMovies: [0 1]
PaperlessBilling: [1 0]
MonthlyCharges: [0.11542289 0.38507463 0.35422886 ... 0.44626866 0.25820896 0.60149254]
TotalCharges: [0.0012751  0.21586661 0.01031041 ... 0.03780868 0.03321025 0.78764136]
Churn: [0 1]
InternetService_DSL: [1 0]
InternetService_Fiber_optic: [0 1]
InternetService_No: [0 1]
Contract_Month_to_month: [1 0]
Contract_One_year: [0 1]
Contract_Two_year: [0 1]
PaymentMethod_Bank_transfer__automatic_: [0 1]
PaymentMethod_Credit_card__automatic_: [0 1]
PaymentMethod_Electronic_check: [1 0]
PaymentMethod_Mailed_check: [0 1]
```

➔ Phần xử lí dữ liệu của em đã hoàn thành: chuyển các biến về kiểu số và đưa các giá trị đó nằm trong khoảng từ 0 đến 1

- Đây là bảng cho biết hệ số tương quan giữa các biến tính năng với nhau và với biến mục tiêu.
  (Những ô có màu vàng là có hệ số tương quan khá cao, màu càng đậm là càng cao.)



- Từ bảng trên, có một số nhận xét như sau:
  - 'Churn' và 'tenure' có hệ số tương quan là -0.35, điều này có nghĩa là biến 'tenure'- số tháng khách hàng gắn bó với công ty càng thấp thì tỉ lệ khách hàng rời đi càng cao.
  - 'TotalCharges' và 'tenure' có mối tương quan cao 0.83, 'Contract_type' và 'tenure' có mối tương quan từ -0.65 đến 0.56. Nghĩa là, khi mà 'contract' càng kéo dài thì 'TotalCharges' càng tăng.
  - Ta thấy được, mối tương quan giữa các loại 'Internet', 'PhoneService' và 'MonthlyCharges', 'TotalCharges' tương đối cao. Nghĩa là khi các thiết bị có giá cao thì dẫn tới 'MonthlyChargé' tăng, 'TotalCharges' tăng.
  - Có thể thấy được là có một biến không có mối tương quan với bất kì biến nào khác: 'gender'.

- Việc theo dõi từ Hệ số tương quan chỉ cho mình phỏng đoán trước giữa hai biến có mối quan hệ với nhau hay không và mối quan hệ đó là âm hay dương.
- Để xác định chính xác chúng có mối tương quan với nhau hay không thì cần phải tính ra p-value (sig) nếu nhỏ hơn mức ý nghĩa (significance level) 5% thì có tương quan và ngược lại thì không có tương quan giữa 2 biến so sánh.
- Sau khi kiểm tra cẩn thận ý nghĩa của từng biến và chỉ giữ những biến có giá trị p nhỏ hơn 0,05

```
SeniorCitizen                    0.0103
tenure                           0.0000
MultipleLines                    0.0114
PaperlessBilling                 0.0000
TotalCharges                     0.0000
InternetService_Fiber_optic      0.0482
InternetService_No               0.0109
Contract_Month_to_month          0.0000
Contract_Two_year                0.0013
PaymentMethod_Electronic_check   0.0053
dtype: float64
```

Có:

Churn = 0.2168*SeniorCitizen - 0.0606*tenure + 0.4484*MultipleLines + 0.3424*PaperlessBilling + 0.0003*TotalCharges + 1.8470*InternetService_Fiber_optic - 1.6868*InternetService_No + 0.7592*Contract_Month_to_month - 0.5979*Contract_Two_year - 0.0324*PaymentMethod_Electronic_check

Phương trình này cho chúng ta thấy rằng nếu khách hàng là người cao tuổi, sử dụng MultipleLines, sử dụng dịch vụ internet cáp quang, sử dụng thanh toán điện tử, sử dụng séc điện tử làm phương thức thanh toán, thì họ có nhiều khả năng rời đi hơn.

3. Xây dựng mô hình

Mục tiêu của học có giám sát là xây dựng mô hình hoạt động tốt trên dữ liệu mới (unseen data). Vậy dữ liệu đầu vào như thế nào để có được một mô hình hoạt động tốt ?

- Xử lí dữ liệu mất cân bằng để cho hiệu suất mô hình đạt kết quả cao hơn.
- Tại vì khi mà dữ liệu không cân bằng, mô hình có thiên hướng dự đoán ra lớp đa số để tăng chỉ số accuracy, khi đó chỉ số accuracy không còn tác dụng đánh giá nữa.

```
Random over-sampling:
0    5163
1    5163
Name: Churn, dtype: int64
```

Sau khi cân bằng dữ liệu thì số lượng của giá 0 và 1 bằng nhau (so với lúc đầu là giá trị 0 có 5163 và giá trị 1 có 1869)

- Đầu tiên, phải tách tập dữ liệu đầy đủ của mình thành các giá trị input và output:
  - X là các giá trị input, gồm các biến độc lập phục vụ cho việc dự đoán biến mục tiêu (X là 1 lưới dữ liệu 2 chiều, trong đó, các hàng đại diện cho các mẫu, các cột đại diện cho các tính năng)
  - y là giá trị output, là biến mục tiêu mình vừa nhắc đến, là những gì ta muốn dự đoán từ dữ liệu.
- Vậy ở đây, biến mục tiêu của mình là biến 'churn' trong tập dữ liệu → dự đoán tỉ lệ rời đi của khách hàng trên thang đo 'YES' hoặc 'NO'
- Chia tập dữ liệu thành 2 phần: tập train (tập huấn luyện) và tập test (tập thử nghiệm). Với test_size = 0.2, nghĩa là chia 80% là tập đào tạo và 20% là tập thử nghiệm.

Bây giờ em sẽ vào bước xây dựng mô hình, bằng 2 loại: dùng LogisticRegression và RandomForest.

LOGISTIC REGRESSION:

- Cho in ra Kết quả dự đoán trên dữ liệu từ tập test và dùng .score() trả về hệ số xác định, hoặc R^2 cho dữ liệu được truyền. Giá trị tối đa của R^2 là 1, giá trị R^2 càng cao thì càng phù hợp.
- Sau khi cân bằng dữ liệu đầu vào, chúng em sử dụng mô hình hồi quy logistics tính ra được accuracy là khoảng 0.78

```
preds [1. 0. 1. 1. 0. 0. 1. 1. 0. 1.]

Accuracy with Logistic Regression: 0.7676669893514037

              precision    recall  f1-score   support

         0.0       0.79      0.73      0.76      1033
         1.0       0.75      0.81      0.78      1033

    accuracy                           0.77      2066
   macro avg       0.77      0.77      0.77      2066
weighted avg       0.77      0.77      0.77      2066
```

RANDOM FOREST:

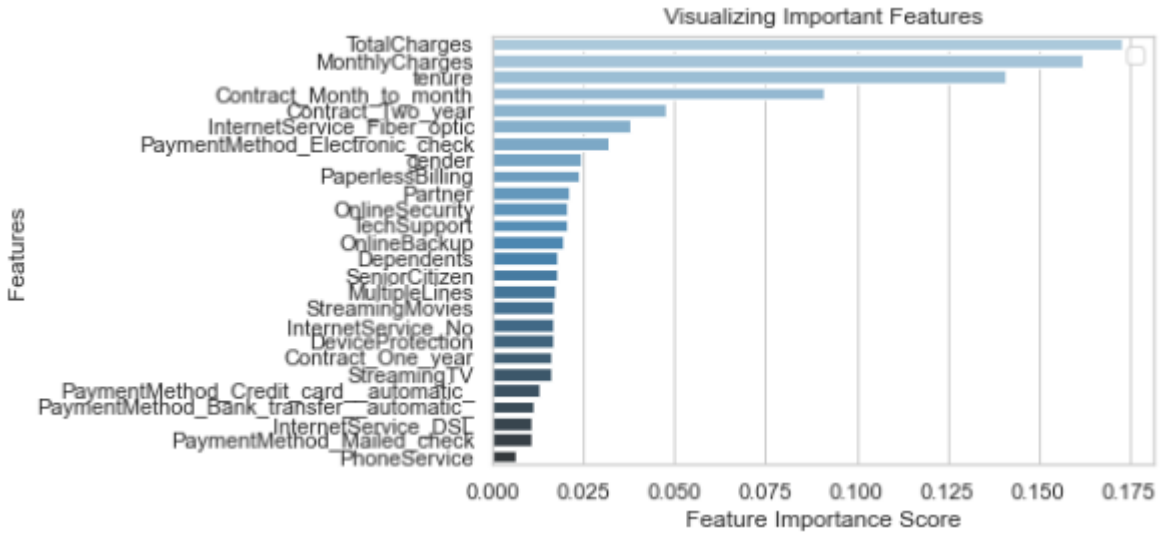
- Sau đó em sẽ dùng RandomForest để xây dựng mô hình, tức là xây dựng nhiều cây quyết định bằng thuật toán Decision Tree. Mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.
- Chúng ta tiếp tục xây dựng mô hình từ tập dữ liện train và in kết quả dự đoán trên tập test và dùng score để trả về hệ số xác định.

```
preds [1. 0. 1. 1. 0. 0. 1. 1. 1. 1.]

Accuracy with Random Forest Classifier: 0.9051306873184899

              precision    recall  f1-score   support

         0.0       0.96      0.85      0.90      1033
         1.0       0.86      0.96      0.91      1033

    accuracy                           0.91      2066
   macro avg       0.91      0.91      0.90      2066
weighted avg       0.91      0.91      0.90      2066
```

➔ Kết luận: Có thể thấy, sau khi xây dựng mô hình bằng hai cách trên, thì mô hình được xây dựng dùng RandomForest cho kết quả tốt hơn.

4. Biến đặc trưng quan trọng nhất

- Sau khi đánh giá độ chính xác của mô hình, vậy thì tham số đầu vào nào đóng vai trò quan trọng nhất trong việc xác định dự đoán của chúng ta ?
- Đây là bảng xếp hạng độ ảnh hưởng của các tham số đầu vào lên biến Churn. chúng ta có thể thấy rằng 'TotalCharges' đóng một vai trò quan trọng hơn trong việc xác định việc phân loại một khách hàng có rời đi hay không.



III. Kết luận

✓ Tỉ lệ Churn là một chỉ số quan trọng cho các doanh nghiệp. Xác định những khách hàng không hài lòng có thể giúp các nhà quản lí xác định kế hoạch sản phẩm hoặc định giá, xử lí các vấn đề cũng như đáp ứng nhu cầu và sở thích của khách hàng. Khi đó, sẽ dễ dàng chủ động hơn trong việc giảm tỉ lệ Churn.

✓ Hiệu suất của cả hai mô hình được đánh giá bằng các thước đo hiệu suất khác nhau. Kết quả cho thấy rằng bằng mọi phép đo, cả hai loại mô hình này đều cho ra hiệu suất tương đối tốt.

Mô hình hồi quy Logistic, độ chính xác là 80%, còn mô hình dùng Random Forrest Algorithm là 78.8%

✓ Sau khi xử lý mất cân bằng dữ liệu thì độ chính xác của Random Forrest Algorithm cao hơn Logit với 90% trong khi mô hình hồi quy logistis độ chính xác 77%. Vậy mô hình RandomFrorest cho kết quả mô hình với hiệu suất cao hơn.

✓ Mô hình hồi quy Logistic đã phân loại chính xác rằng có 1033 khách hàng sẽ không rời đi và không chính xác dự đoán rằng 374 khách hàng sẽ rời rời đi. Với mô hình dùng Random Forrest Algorithm cũng cho kết quả tương tự.

✓ Giải pháp đề ra sau khi dự đoán: Khách hàng ở lại với công ty lâu, chọn hoàn toàn không sử dụng dịch vụ internet, chọn hợp đồng một năm hoặc hai năm thay vì tùy chọn tháng này sang tháng khác và công ty đưa ra mức giá rẻ hơn, thì họ ít có khả năng rời đi.

IV. Tài liệu tham khảo

1. https://towardsdatascience.com/predicting-customer-churn-using-logistic-regression-c6076f37eaca.
2. https://en.wikipedia.org/wiki/Random_forest#:~:text=Random%20forests%20or%20random%20decision,class%20selected%20by%20most%20trees.
3. https://www.kaggle.com/code/shubha23/telco-customer-churn-prediction/notebook.
4. https://www.kaggle.com/code/kaanboke/the-most-common-evaluation-metrics-a-gentle-intro/notebook.
5. https://neptune.ai/blog/how-to-implement-customer-churn-prediction.

6. https://adiwijaya.staff.telkomuniversity.ac.id/files/2014/02/Customer-Churn-Prediction-using-Improved-Balance-Random-Fores.pdf.
7. https://www.kaggle.com/code/nehapawar/churn-prediction-using-logistic-regression/notebook#Model-buidling.
8. http://thinkzone.vn/blog/churn-rate-la-gi-cach-tinh-2-loai-churn-rate-quan-trong-trong-doanh-nghiep.