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

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. Tạo mô hình.....8
III. Tài liệu tham khảo: 8

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 tiêu tốn gấp 10 lần chi phí để giữ chân khách hàng hiện tại. Trên thực tế, việc tăng tỷ lệ giữ chân khách hàng dù chỉ 5% cũng có thể tạo ra lợi nhuận tăng ít nhất 25%. Do đó Custumer 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)

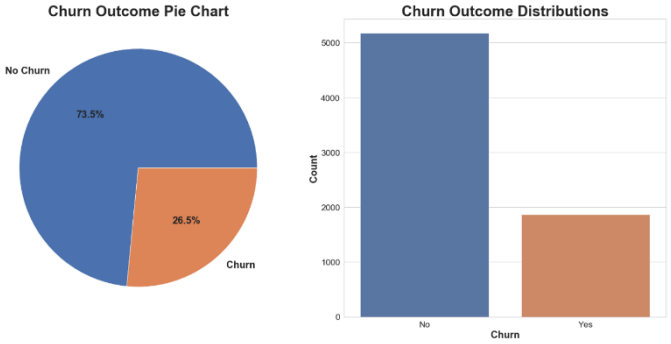
- OnlineSecurity: Khách hàng có bổ sung tiện ích bảo mật trực tuyến không(yes, no, no internet service)
- OnlineBackup: Khách hàng có bổ sung tiện ích sao lưu trực tuyến không (yes, no, no internet service)
- DeviceProtection: Khách hàng có bổ sung tiện ích bảo vệ thiết bị không (yes, no, no internet 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)



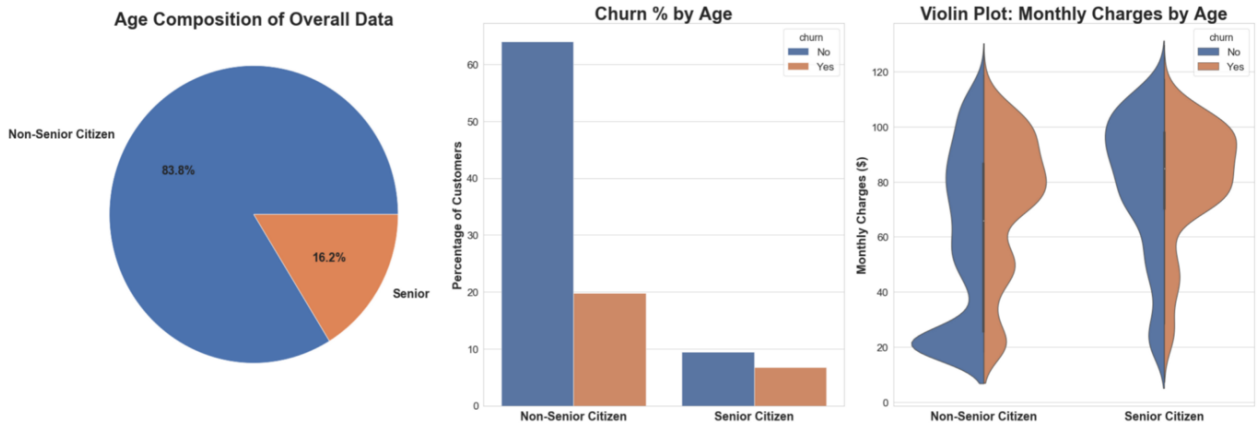
Ở biểu đồ hình tròn bên trái, khoảng 27% khách hàng từ tập dữ liệu rời đi. Một tỉ lệ khá cao.

II. Đánh giá và tạo mô hình

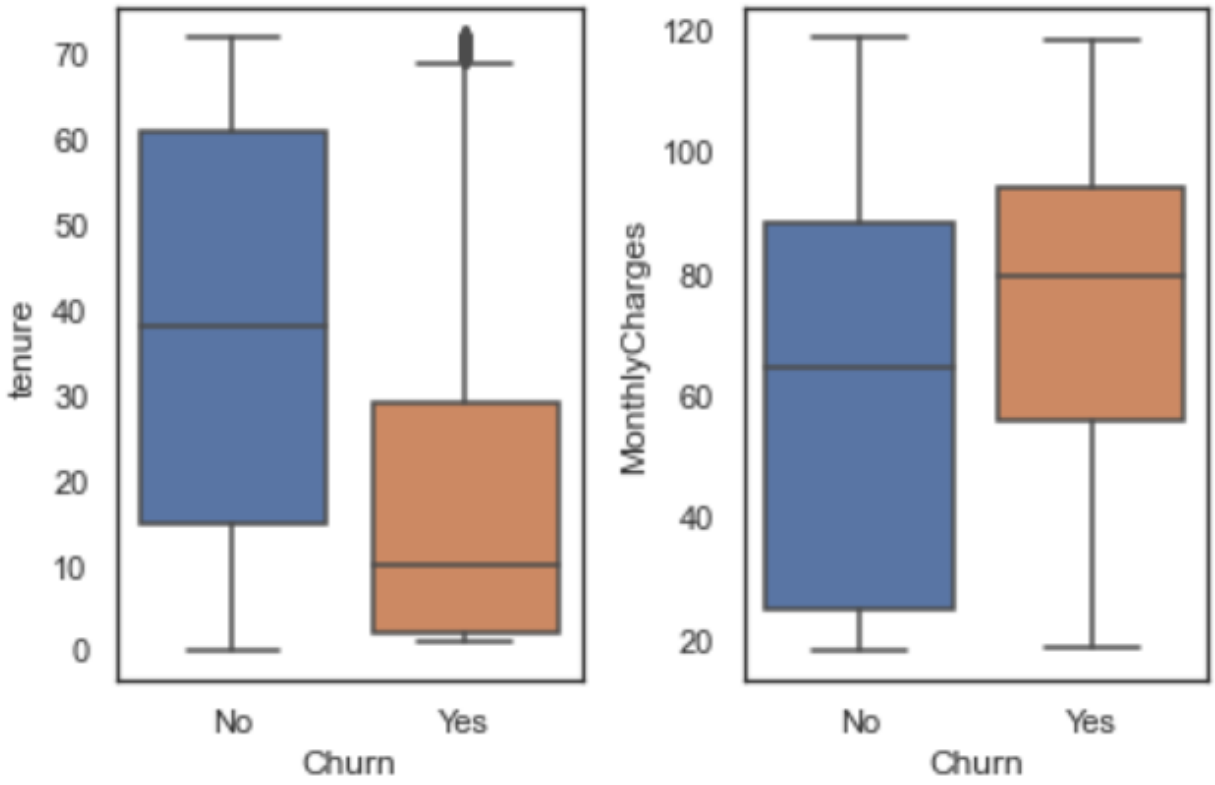
1. Đánh giá

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.

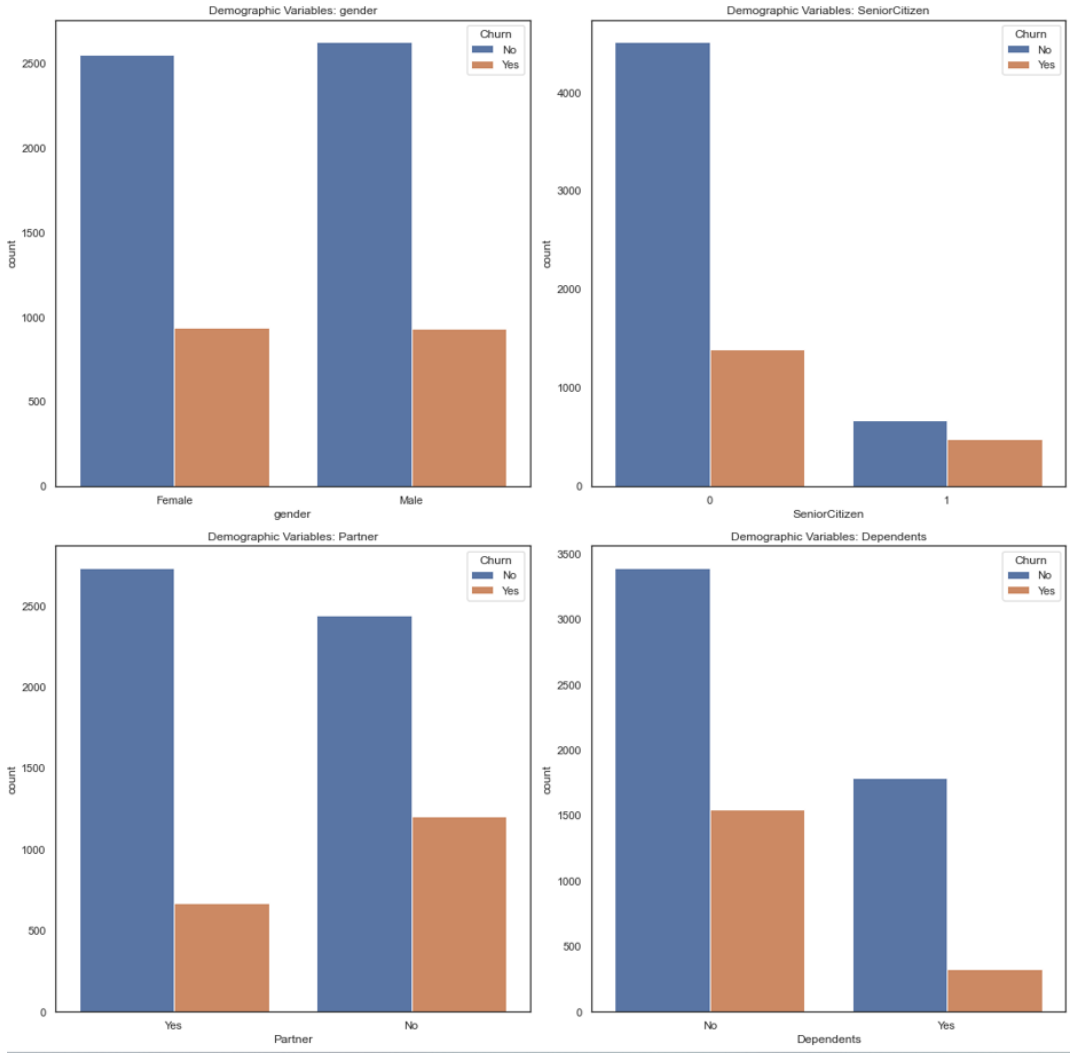
Từ các hình ảnh ta rút ra kết luận trước tiên:



- 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$.



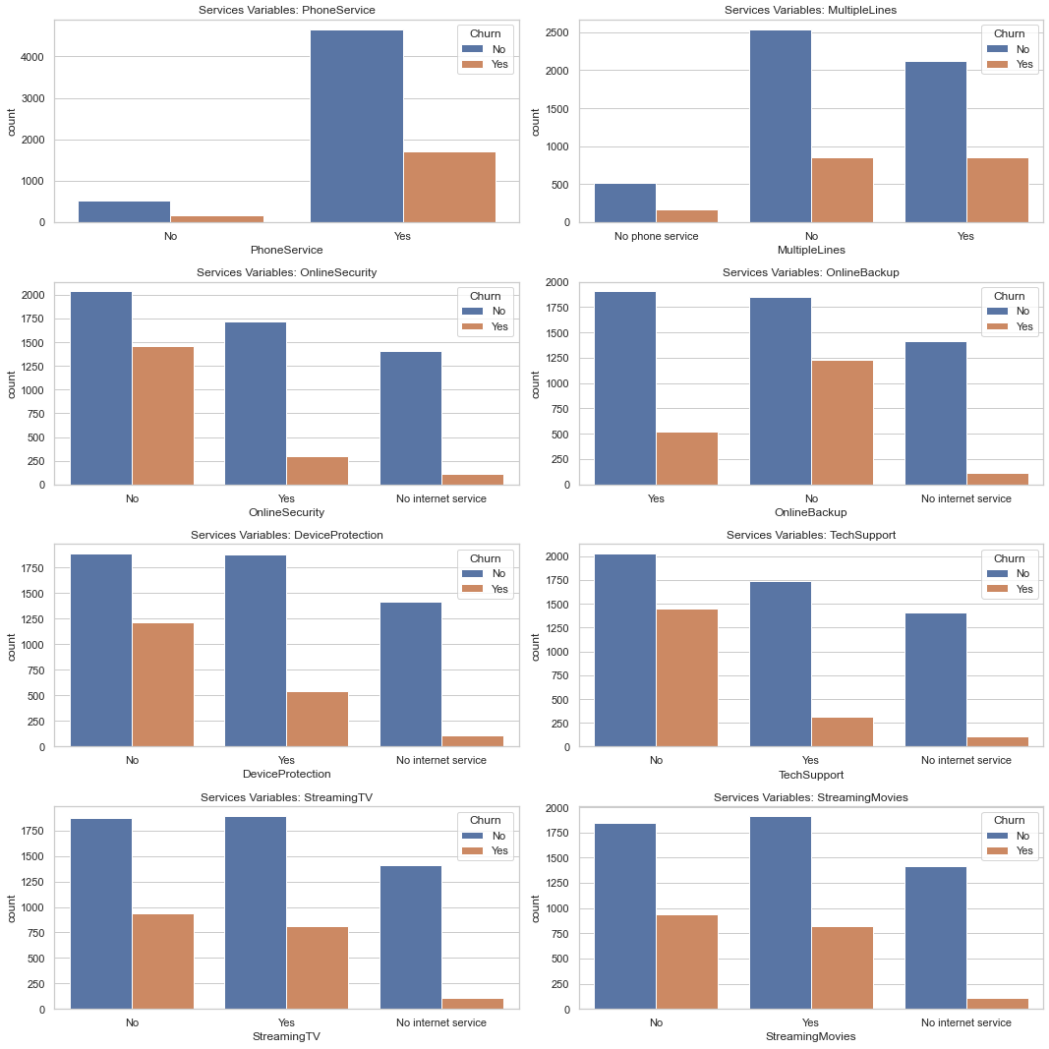
- **Nhận xét:**
  - Biểu đồ bên trái tỷ lệ Churn(yes, no) phụ thuộc vào tenure. Thấy được thời hạn gắn bó với công ty của khách hàng đã rời đi ngắn hơn nhiều(trong 30 tháng đầu) so với khách hàng ở lại.
  - Biểu đồ bên phải so sánh số tiền hàng tháng khách hàng phải trả so với thời gian churn. Mức phí 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.

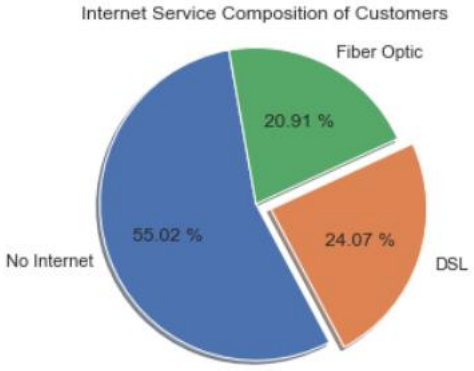


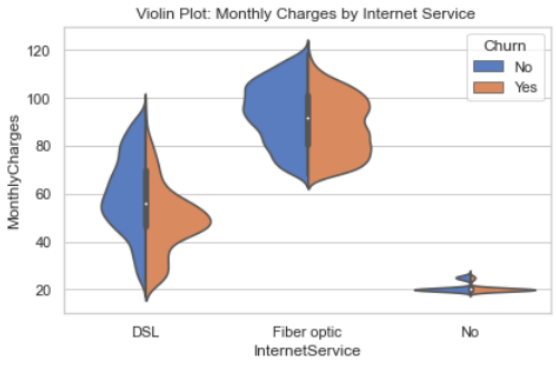
- **Nhận xét:**
  - Thống kê cho thấy ít sự khác biệt về tỷ lệ churn với giới tính. Có một tỉ lệ khách hàng rời đi cao với các biến SeniorCitizen, khách hàng có Partners và khách hàng không có Dependents.
  - Khách hàng không có Partners có tỷ lệ Churn nhiều hơn với khách hàng có Partners.

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.

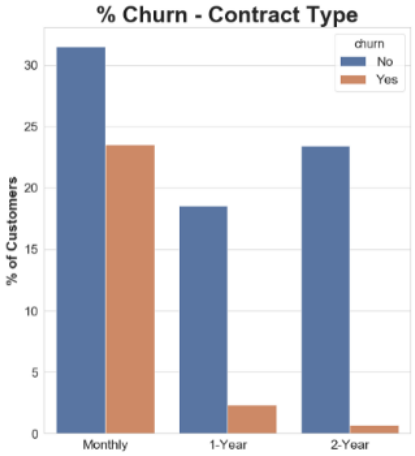


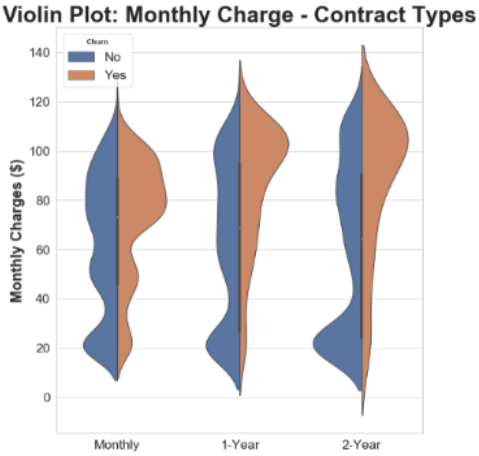




- Khách hàng sử dụng internet cáp quang có tỷ lệ churn cao.
- Cáp quang là lựa chọn internet phổ biến nhất. Khách hàng sử dụng Internet cáp quang chiếm tỷ lệ đáng kể so với khách hàng sử dụng DSL hoặc Không có Internet.

- 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. Tạo mô hình

Chúng em sẽ giải thích trong Jupyter.

III. 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.