**BỘ THÔNG TIN VÀ TRUYỀN THÔNG**

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

-----------------------------

Ảnh có chứa văn bản, Phông chữ, biểu tượng, Đồ họa

Mô tả được tạo tự động

**BÁO CÁO THỰC TẬP**

**TỐT NGHIỆP ĐẠI HỌC**

***Đề tài*: “ Ứng dụng mô hình Weighted Average Ensemble dự báo khách hàng rời bỏ dịch vụ viễn thông”**

**Người hướng dẫn : HUỲNH TRỌNG THƯA**

**Sinh viên thực hiện : NGUYỄN PHẠM NHẬT MINH**

**Mã số sinh viên : N17DCCN095**

**Lớp : D17CQCP02-N**

**Khoá** **: 2017**

**Hệ : ĐẠI HỌC CHÍNH QUY**

**TP.HCM, tháng 8/2023**

**BỘ THÔNG TIN VÀ TRUYỀN THÔNG**

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

-----------------------------

Ảnh có chứa văn bản, Phông chữ, biểu tượng, Đồ họa

Mô tả được tạo tự động

**BÁO CÁO THỰC TẬP**

**TỐT NGHIỆP ĐẠI HỌC**

***Đề tài*: “ Ứng dụng mô hình Weighted Average Ensemble dự báo khách hàng rời bỏ dịch vụ viễn thông”**

**Người hướng dẫn : HUỲNH TRỌNG THƯA**

**Sinh viên thực hiện : NGUYỄN PHẠM NHẬT MINH**

**Mã số sinh viên : N17DCCN095**

**Lớp : D17CQCP02-N**

**Khoá** **: 2017**

**Hệ : ĐẠI HỌC CHÍNH QUY**

**TP.HCM, tháng 8/2023**

# **LỜI CẢM ƠN**

Lời đầu tiên em xin chân thành cảm ơn các thầy, cô trong khoa Công nghệ thông tin 2, học viện Công nghệ bưu chính viễn thông cơ sở tại TP HCM đã tạo điều kiện thuận lợi cho em trong quá trình học tập tại học viện cũng như trong thời gian thực hiện đồ án thực tập tốt nghiệp. Đặc biệt, em muốn gửi lời cảm ơn tới thầy Huỳnh Trọng Thưa - giảng viên trực tiếp hướng dẫn, chỉ bảo giúp em giải đáp các thắc mắc trong quá trình làm đồ án thực tập tốt nghiệp.

Mặc dù đã cố gắng với tất cả nỗ lực của bản thân để hoàn thiện đồ án thực tập, nhưng do thời gian có hạn, năng lực và kinh nghiệm còn hạn chế nên đồ án không thể tránh khỏi những thiếu sót. Kính mong nhận được sự đóng góp ý kiến từ phía thầy, cô, bạn bè để em có thể nâng cao kiến thức của bản thân, hoàn thiện đồ án được tốt hơn.

Thành phố Hồ Chí Minh, ngày … tháng … năm 2023,

Sinh viên thực hiện

Nguyễn Phạm Nhật Minh

# **MỤC LỤC**

[**LỜI CẢM ƠN** 1](#_Toc142898308)

[**MỤC LỤC** 1](#_Toc142898309)

[**MỞ ĐẦU** 2](#_Toc142898310)

[**CHƯƠNG 1:** **Nội dung lý thuyết** 3](#_Toc142898311)

[1.1. **Mô hình Weighted Average Ensemble?** 3](#_Toc142898312)

[**1.1.1.** **Dẫn nhập:** 3](#_Toc142898313)

[**1.1.2.** **Weighted Averaging:** 3](#_Toc142898314)

[1.2. **Bài toán dự báo khách hàng rời bỏ dịch vụ viễn thông:** 4](#_Toc142898315)

[**1.2.1.** **Bối cảnh:** 4](#_Toc142898316)

[**1.2.2.** **Những dữ liệu nào về khách hàng các công ty ngành viễn thông thu thập:** 4](#_Toc142898317)

[**1.2.3.** **Tìm hiểu qua ví dụ, khám phá dữ liệu và các biến:** 6](#_Toc142898318)

[**CHƯƠNG 2:** **NỘI DUNG THỰC HÀNH** 22](#_Toc142898319)

[**2.1.** **Mô hình phân loại dựa trên Weighted Average Ensemble:** 22](#_Toc142898320)

[**2.2.** **Ứng dụng minh họa dự báo khách hàng rời bỏ** 27](#_Toc142898321)

[**PHỤ LỤC** 31](#_Toc142898322)

[**DANH MỤC TÀI LIỆU THAM KHẢO** 32](#_Toc142898323)

# **MỞ ĐẦU**

Ngày nay, khách hàng có thể tham gia rất nhiều dịch vụ được cung cấp bởi một công ty viễn thông, không còn là dịch vụ gọi điện, nhắn tin đơn thuần đã có từ rất lâu. Điện thoại thông minh ngày nay có nhiều tính năng vượt trên cả sự mong đợi. Mọi người có thể giao tiếp qua tin nhắn SMS và MMS. Họ có thể sử dụng điện thoại để kết nối Internet, gửi e-mail, tải trò chơi và nhạc chuông, cũng như để liên lạc với bạn bè và gia đình khi họ đi du lịch nước ngoài.

Chính vì thế, việc tìm hiểu khách hàng về tổng quan là chiến lược hoạt động của các công ty ngành viễn thông. Sự chậm trễ trong phân tích dữ liệu hành vi sẽ dẫn đến sự chậm trễ trong việc đưa ra các chiến thuật bán hàng, marketing phù hợp và sau cùng là sự suy giảm về mặt lợi nhuận.

Cấu trúc của bài nghiên cứu gồm 2 phần chính là lý thuyết và thực hành.

* Phần lý thuyết:
* Mô hình Weighted Average Ensemble
* Bài toán dự báo khách hàng rời bỏ dịch vụ viễn thông
* Phần thực hành:
  + Mô hình phân loại dựa trên Weighted Average Ensemble
  + Ứng dụng minh họa dự báo khách hàng rời bỏ

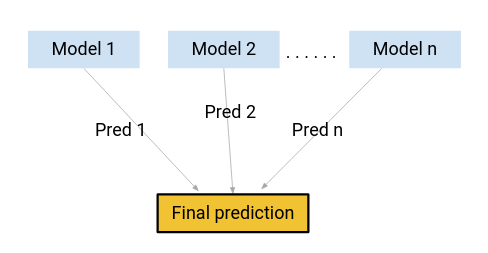
1. **Nội dung lý thuyết**
   1. **Mô hình Weighted Average Ensemble?**
      1. **Dẫn nhập:**

Giả sử chúng ta có một bài toán phân loại sản phẩm sử dụng ML. Team của bạn chia thành 3 nhóm, mỗi nhóm sử dụng một thuật toán khác nhau để train model trên tập train set. Sau đó đánh giá độ chính xác trên tập validation set:

* Nhóm 1: Sử dụng thuật toán Linear Regression.
* Nhóm 2: Sử dụng thuật toán k-Nearest Neighbour.
* Nhóm 3: Sử dụng thuật toán Decision Tree.

Độ chính xác của mỗi nhóm lần lượt là 70%, 67% và 76%. Điều này hoàn toàn dễ hiểu bởi vì 3 models làm việc theo những các khác nhau. Ví dụ, Linear Regression cố gắng tìm ra mối quan hệ tuyến tính giữa các điểm dữ liệu, trong khi Decision Tree thì lại dựa vào mỗi quan hệ phi tuyến để liên kết dữ liệu.

Có cách nào kết hợp kết quả cả 3 models để tạo ra kết quả cuối cùng hay không?

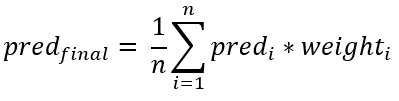


Câu hỏi này là tiền đề cho một phương pháp, một họ các thuật toán hoạt động rất hiệu quả trong các bài toán ML. Đó là Ensemble Learning hay Ensemble Models.

Ở mức độ cơ bản, có 3 kỹ thuật là:

* Max Voting
* Averaging
* Weighted Averaging
  + 1. **Weighted Averaging:**

Đây là kỹ thuật mở rộng của averaging. Mỗi model được gắn kèm với một trọng số tỷ lệ với mức độ quan trọng của model đó. Kết quả cuối cùng là trung bình có trọng số của tất cả kết quả của các models



Lấy ví dụ, đợt vừa rồi, công ty tổ chức khám sức khỏe cho nhân viên tại bệnh viện X. Sau khi khám xong, phòng tổ chức nhân sự (TCNS) lấy ý kiến mọi người về chất lượng khám bệnh để xem năm sau có tiếp tục khám ở bênh viên X đó nữa không. Bảng dưới là ý kiến của 5 người được chọn ngẫu nhiên trong số toàn bộ nhân viên.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Người 1** | **Người 2** | **Người 3** | **Người 4** | **Người 5** |
| Có | Không | Không | Có | Có |

Có 3 ý kiến muốn tiêp tục khám ở bệnh viện X vào năm sau, và 2 ý kiến muốn đổi bênh viện khác.

Trong số 5 người được hỏi thì người thứ nhất có vợ là bác sĩ, người thứ 2 có mẹ là y tá, người thứ 3 có người yêu là sinh viên trường y. Vì vậy, ý kiến của 3 người này rõ ràng có giá trị hơn so với 2 người còn lại.

Ta đánh trọng số cho mỗi người như bảng dưới (hàng thứ 2 là trọng số, hàng thứ 3 là điểm đánh giá):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Người 1** | **Người 2** | **Người 3** | **Người 4** | **Người 5** |
| 1 | 0.8 | 0.5 | 0.3 | 0.3 |
| 2 | 4 | 3 | 5 | 4 |

Điểm đánh giá cuối cùng sẽ là: (21 + 40.8 + 30.5 + 50.3 + 40.3)/5 = 1.88

* 1. **Bài toán dự báo khách hàng rời bỏ dịch vụ viễn thông:**
     1. **Bối cảnh:**

Mục đích của dự báo khách hàng rời bỏ trong ngành viễn thông, cũng giống như mục đích của phân khúc khách hàng nói chung hay trong những ngành khác:

* Cải thiện các hoạt động sales, marketing tối ưu nhắm mục tiêu tốt hơn trong đó mỗi phân khúc khách hàng sẽ được giới thiệu các sản phẩm, dịch vụ phù hợp với nhu cầu, sở thích, thói quen hành vi, và các đặc điểm cá nhân khác
* Khám phá các nhu cầu đặc biệt của các khách hàng làm cơ sở để khám phá các cơ hội kinh doanh mới trong đó tập trung phát triển các dịch vụ mới, để phục vụ đa dạng nhiều phân khúc khác nhau, cơ hội để gia tăng lợi nhuận, với việc cố gắng cá nhân hóa cho từng khách hàng và tiến đến cá nhân hóa cho “cả phân khúc”.
* Tạo nên những chiến lược tổng thể cho đến các chiến thuật, kế hoạch chi tiết để giữ chân khách hàng cũ, thu hút khách hàng mới, xây dựng nguồn khách hàng cho quá trình phát triển bền vững
  + 1. **Những dữ liệu nào về khách hàng các công ty ngành viễn thông thu thập:**

Một công ty viễn thông có đa dạng nhiều loại dịch vụ khác nhau, và khách hàng có thể đăng ký nhiều hơn một dịch vụ. Vì thế dữ liệu mô tả về một khách hàng có thể có nhiều thuộc tính khác nhau khi ở mỗi dịch vụ khách hàng sẽ có các hành vi riêng biệt. Đây cũng là thách thức ban đầu khi mỗi công ty phải xây dựng cơ sở dữ liệu lớn mà ở đó đảm bảo dữ liệu khách hàng được thu thập đầy đủ từ nhiều nguồn khác nhau, và mô tả khách hàng ở mọi góc độ thông tin, được xử lý, quản lý sao cho đạt chất lượng tốt nhất để sẵn sàng được đưa vào phân tích.

Dữ liệu người dùng xét riêng do dịch vụ Call, gọi/ nhận cuộc gọi được lưu trữ trong Call Detail Records (hay còn gọi CDRs), chứa dữ liệu hành vi của người dùng ở tất cả loại dịch vụ liên quan đến gọi/ nhận cuộc gọi.

Ngoài ra đối với các dịch vụ khác như gửi tin nhắn SMS, MMS (tin nhắn đa phương tiện), tin nhắn giọng nói, các dịch vụ internet,… cũng được thu thập, lưu trữ, kết hợp với dữ liệu thời gian theo dõi hành vi ví dụ thời gian gọi điện chủ yếu là vào lúc nào trong ngày, vào ngày nào trong tuần,…Nói chung tài sản dữ liệu, thông tin đầu vào cho quá trình phân khúc là rất nhiều, câu hỏi đặt ra là làm cách nào tận dụng, khai thác chúng một cách tối đa.

Dưới đây là tổng quan các loại dữ liệu mô tả khách hàng trong ngành viễn thông:

Thông tin về khách hàng: các thông tin cơ bản về nhân khẩu học nếu có ví dụ ID, số điện thoại, giới tính, nghề nghiệp, độ tuổi, nơi sinh sống.

Thông tin về tham gia dịch vụ: ngày khách hàng tham gia các dịch vụ, thời gian kể từ lúc khách hàng tham gia dịch vụ (Tenure), trạng thái dịch vụ, loại khách hàng (khách hàng trả trước, trả sau, hay dạng hợp đồng)

Thông tin về sử dụng dịch vụ gọi:

* Dữ liệu cuộc gọi đi, nhận cuộc gọi đến
* Dữ liệu cuộc gọi nội địa, quốc tế, chuyển vùng quốc tế
* Dữ liệu mạng: liên mạng, nội mạng, ngoại mạng
* Dữ liệu thời điểm gọi: vào ngày làm việc, ngày nghỉ, giờ cao điểm,…
* Dữ liệu thời lượng cuộc gọi đến, cuộc gọi đi

Thông tin về sử dụng các dịch vụ khác như SMS, MMS, GPRS, Internet, Voice (tin nhắn thoại),…

Thông tin về quan hệ cộng đồng (Community): tổng số điện thoại khách hàng gọi đi, tổng số điện thoại khách hàng nhận cuộc gọi, phân theo nội mạng, ngoại mạng,…

Thông tin về Top-up (khách hàng trả trước): tần suất trả trước, lần gần nhất trả trước, tổng số tiền đã trả, loại Top-up

Thông tin về thanh toán: lịch sử thanh toán, hình thức thanh toán, số ngày trung bình tính đến ngày thanh toán,…

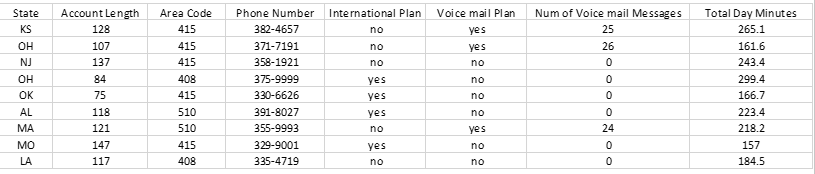
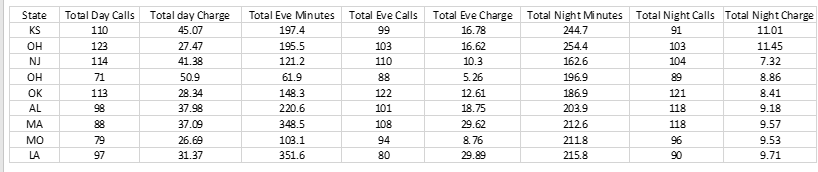
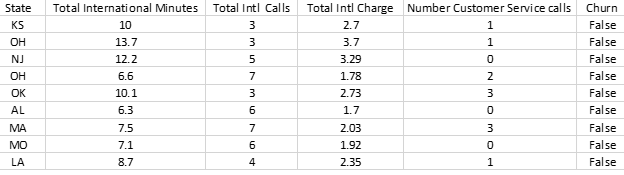
Thông tin về giá trị kinh doanh:lợi nhuận, chi phí cho một khách hàng hoặc dùng chỉ số thông dụng ARPU – Average Revenue Per Unit (doanh thu mỗi khách hàng mang lại), MARPU – Marginal Average Revenue Per Unit (lợi nhuận sau cùng mỗi khách hàng = doanh thu – chi phí xét theo mỗi khách hàng)

Thông tin trạng thái khách hàng: còn hoạt động, đã rời dịch vụ, hay bị phạt, bị chặn

* + 1. **Tìm hiểu qua ví dụ, khám phá dữ liệu và các biến:**

Ví dụ tham khảo từ UCI Repository of Machine Learning Databases tại Đại học California, Irvine. Dữ liệu mẫu lấy từ một công ty viễn thông tại Hoa Kỳ, gồm các biến như sau. Trong giới hạn báo cáo nên ví dụ chọn mẫu dữ liệu không quá nhiều biến giống như những thông tin mẫu phải có của data trong ngành viễn thông vừa liệt kê ở trên.

* + - State: mã tiểu bang, có 51 tiểu bang Hoa Kỳ
    - Area code: mã vùng khu vực, 415 – San Francisco,408 – San Jose và 510 – Okland
    - Phone Number: Số điện thoại của khách hàng
    - International Plan: Đăng ký gọi quốc tế (Yes: có, No: không)
    - Voice mail Plan: Đăng ký dịch vụ hộp thư thoại (Yes: có, No: không)
    - Num of Voice mail Messages: Số lần gửi thư thoại
    - Total Day Minutes: Tổng số phút khách hàng dành gọi điện vào buổi sáng, ban ngày
    - Total Day Calls: Tổng số cuộc gọi thực hiện vào buổi sáng, ban ngày
    - Total Day Charge: Tổng số chi phí cho các cuộc gọi vào ban ngày (USD)
    - Total Eve Minutes: Tổng số phút khách hàng dành gọi điện vào chiều tối
    - Total Eve Calls: Tổng số cuộc gọi khách hàng gọi điện vào chiều tối
    - Total Eve Charge: Tổng số chi phí cho các cuộc gọi vào chiều tối (USD)
    - Total Night Minutes: Tổng số phút khách hàng dành gọi điện vào ban đêm
    - Total Night Calls: Tổng số cuộc gọi khách hàng gọi điện vào ban đêm
    - Total Night Charge: Tổng số chi phí cho các cuộc gọi vào ban đêm
    - Total International Minutes: Tổng số phút khách hàng dành cho các cuộc gọi quốc tế
    - Total Intl Calls: Tổng số cuộc gọi quốc tế
    - Total Intl Charge: Tổng chi phí bỏ ra cho các cuộc gọi quốc tế (USD)
    - Number Customer Service calls: Tổng số cuộc gọi đến trung tâm chăm sóc khách hàng
    - Churn: Trạng thái khách hàng đã rời dịch vụ, True: đã rời dịch vụ, False: chưa rời dịch vụ

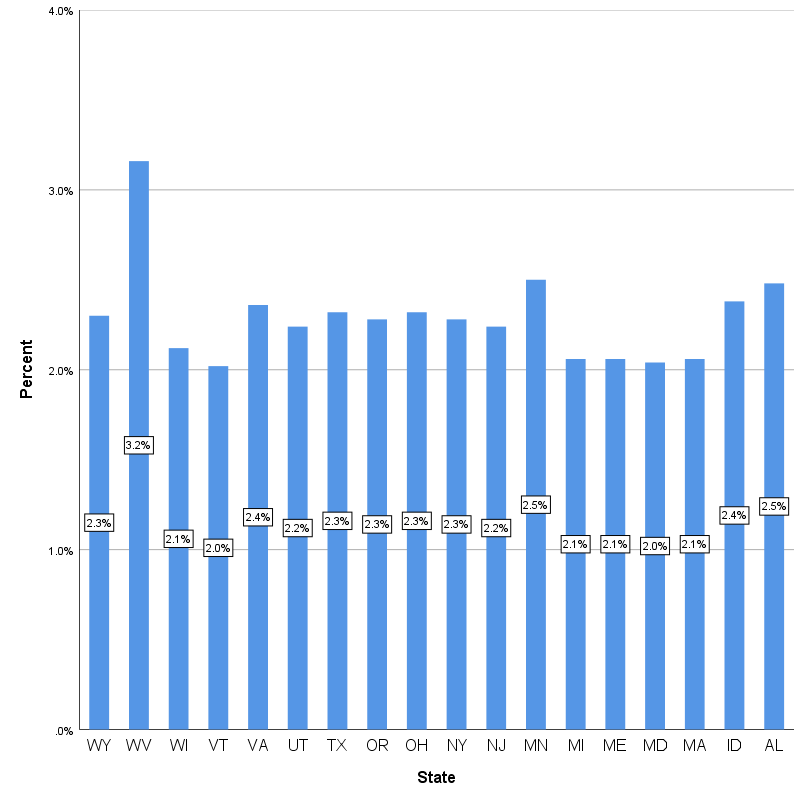
Có tất cả 5000 khách hàng. Dưới đây là dữ liệu mẫu 10 khách hàng đầu tiên với thông tin về dữ liệu trong các biến tương ứng.

Trong tất cả 20 biến thì có các biến định danh Nominal là State – tiểu bang, Phone number – số điện thoại khách hàng, Area code – mã vùng, các biến Flag/ biến thay phiên Binary như Internation plan, Voice mail plan, và biến Churn chỉ có 2 giá trị Yes/ No hoặc True/ False. Các biến còn lại.

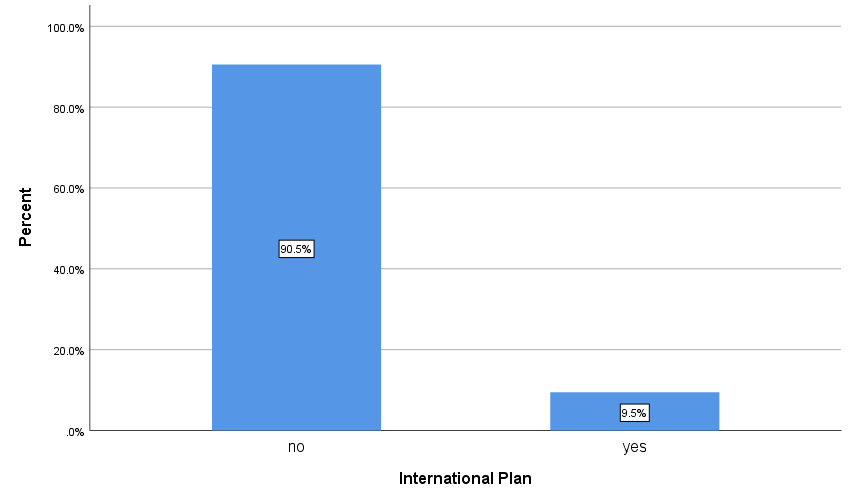
Điểm khác biệt giữa dữ liệu ngành viễn thông so với dữ liệu bán lẻ đó là các biến định lượng đánh giá hành vi khách hàng thường nhiều hơn, do có công cụ kỹ thuật để theo dõi (được phép theo dõi), và đặc biệt là có thể biết được liệu khách hàng đã rời dịch vụ hay chưa trong khi ngành bán lẻ chỉ tracking được những khách hàng thành viên, khách hàng thân thiết còn khách vãng lai, nếu không đăng ký thông tin thì rất khó theo dõi.

Chính vì cơ hội này, các công ty viễn thông thường phân khúc khách hàng để tìm ra các nhóm khách hàng có khả năng rời dịch vụ, dựa theo 3 cách phân khúc: phân khúc theo hành vi (Behaviral segmentation), phân khúc theo giá trị (Value – based segmentation) và phân khúc theo mô hình RFM quen thuộc và tìm mối liên hệ giữa các phân khúc với khả năng rời dịch vụ. Mục đích sau cùng là với các thông tin có được, họ sẽ tiến hành dự báo Churn rate thông qua các đặc điểm bên trong từng phân khúc để tìm ra các giải pháp cho chính phân khúc ấy.

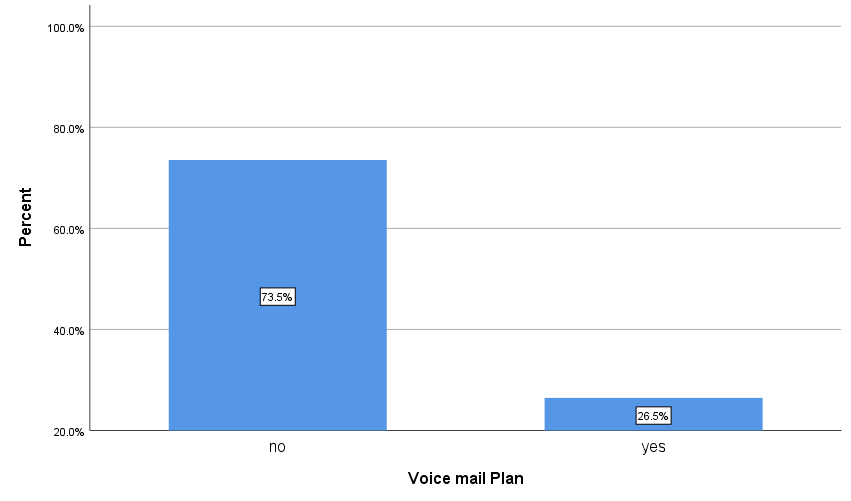
Tập dữ liệu trên hoàn toàn không có missing values, vậy ta xét trước các biến định tính gồm biến định danh và biến flag.

Xét về nhân khẩu học thì không có sự chênh lệch lớn giữa tỷ lệ khách hàng xét theo các tiểu bang. Tiểu bang có số khách hàng cao nhất là West Vargina (WV) chiếm 3.2%, và thấp nhất là tiểu bang California (CA) chiếm 1.04%. Đồ thị trên không hiển thị các tiểu bang có tỷ lệ đóng góp khách hàng dưới 2%.

Như vậy, tỷ lệ khách hàng phân bổ theo tiểu bang không quá chênh lệch, không có nhiều tiểu bang khác ngoài West Vargina nổi trội hơn hẳn về tỷ lệ khách hàng. Nên chúng ta có thể xem xét không cần đưa biến State vào là yếu tố để phân khúc khách hàng.

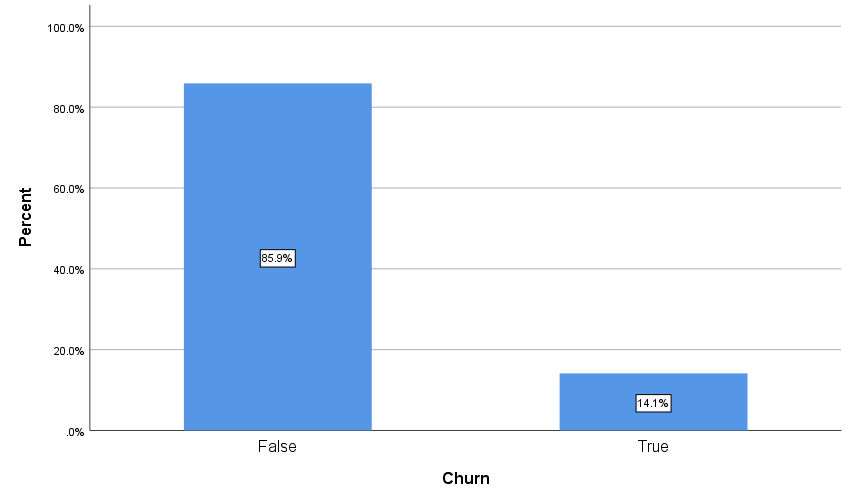
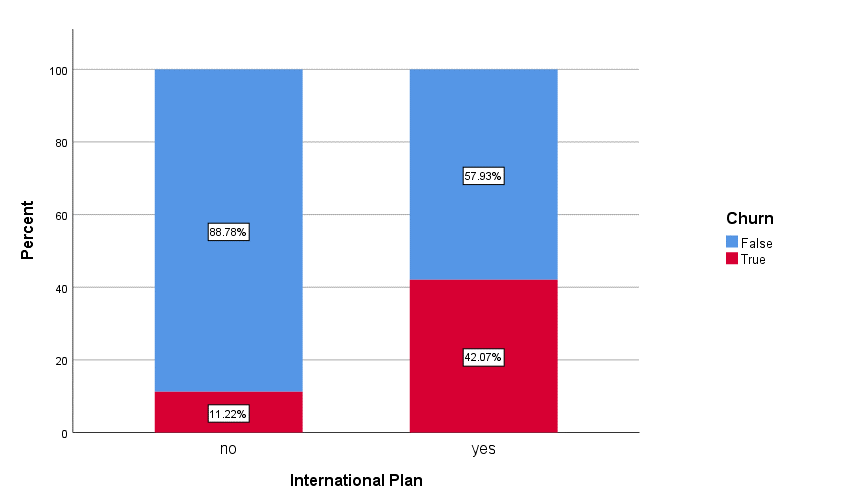
Về số khách hàng đăng ký dịch vụ gọi quốc tế.

Tỷ lệ khách hàng tham gia dịch vụ gọi quốc tế khá thấp chỉ có 9.5%.

Về số khách hàng đăng ký dịch vụ hộp thư thoại

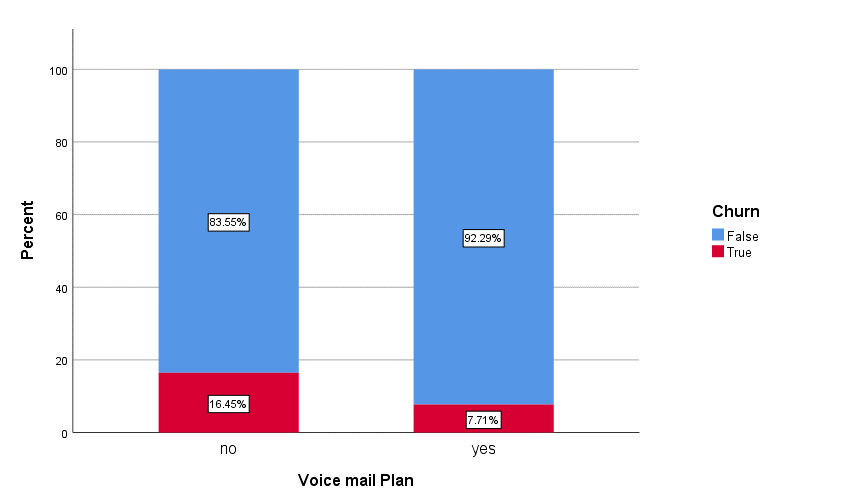
Có 26.5% khách hàng đăng ký dịch vụ hộp thư thoại.

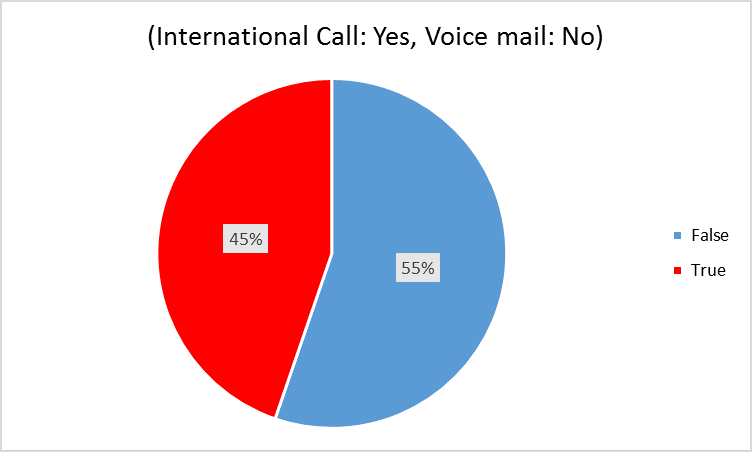
Về tỷ lệ khách hàng rời dịch vụ.

 Tỷ lệ rời dịch vụ là 14.1%, và không rời dịch vụ là 85.9%. Tiếp theo chúng ta cùng tìm hiểu mối quan hệ giữa dịch vụ gọi quốc tế và khả năng rời dịch vụ của khách hàng.

Qua đồ thị chúng ta có thể thấy khách hàng đăng ký dịch vụ gọi quốc tế có khả năng rời churn khá cao. Như vậy công ty viễn thông có thể tiến hành điều tra nguyên nhân tại sao, nếu gói dịch vụ International call này có vấn đề, và vấn đề này tác động trực tiếp đến việc khách hàng churn thì cần có sự thay đổi.

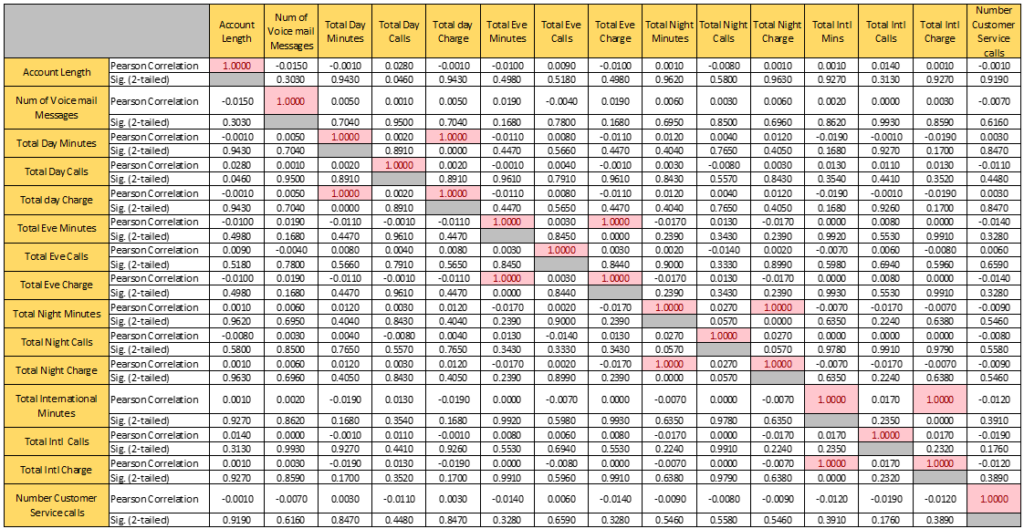
Về mối quan hệ giữa dịch vụ hộp thư thoại (Voice mail):

Khách hàng không đăng ký dịch vụ gọi thư thoại có khả năng rời dịch vụ cao hơn. Tương tự với International call, chúng ta cũng tìm hiểu nguyên nhân là gì và tìm ra giải pháp cho nó.

Phân khúc đơn giản nhất đầu tiên có tỷ lệ rời dịch vụ cao đó là (International plan: Có, Voice mail plan: Không). Theo tập dữ liệu thì phân khúc này có tất cả 342 khách hàng không đăng ký Voice mail và có đăng ký International plan.

Tuy nhiên chỉ có 45% trong phân khúc khách hàng này là rời dịch vụ. Suy ra chưa thể khẳng định khi kết hợp 2 yếu tố này vào sẽ có tác động mạnh hơn lên churn rate.

Tuy nhiên lưu ý mục đích của việc làm này không phải để xác định phân khúc cuối cùng hỗ trợ các hoạt động sales, marketing, chỉ dừng lại ở việc khám phá dữ liệu, mối quan hệ giữa các biến. Một phân khúc tối ưu nó phải bao gồm nhiều đặc điểm, chi tiết hơn về các tính chất của khách hàng để đưa ra đánh giá chính xác. Ví dụ khách hàng không đăng ký International call và có đăng ký Voice mail nhưng vẫn rời dịch vụ vì các yếu tố mà chúng ta chưa biết. Vì thế cách phân khúc ở trên vẫn còn quá “hời hợt”

Tiếp theo chúng ta sẽ tìm hiểu đến các biến định lượng. Trước tiên là mối quan hệ giữa chúng

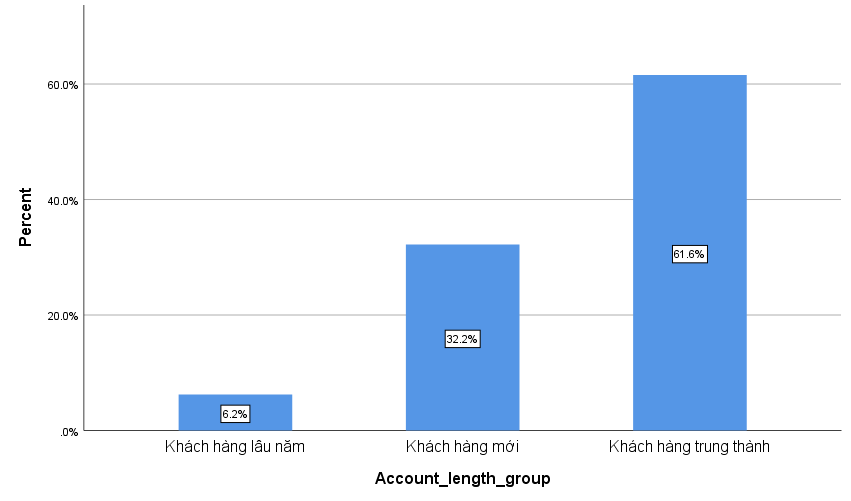
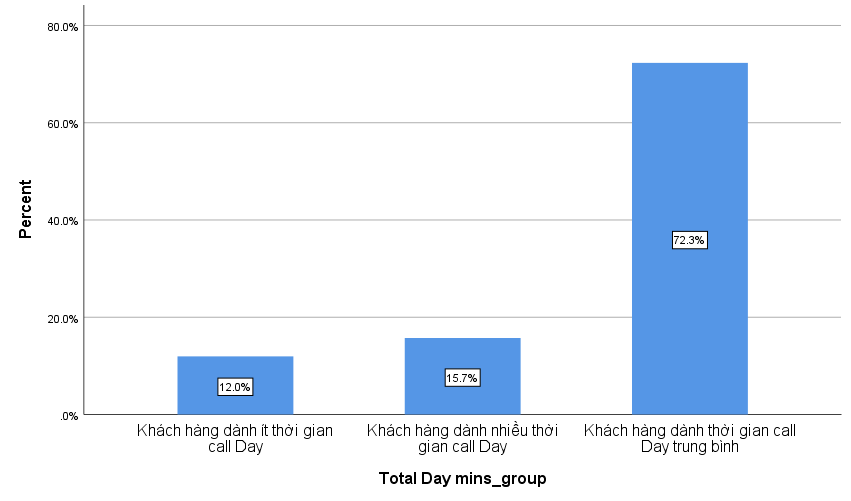
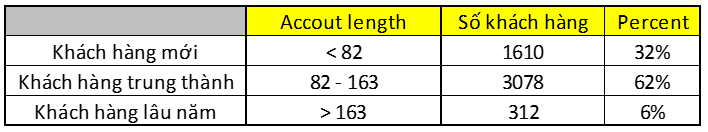
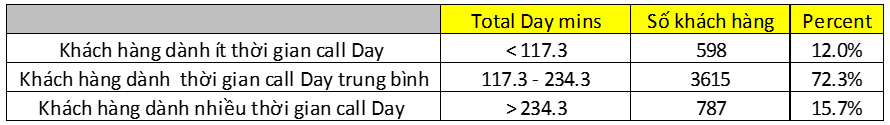
Trừ các cặp biến được hình thành từ chính nó, chúng ta có một cặp được hình thành từ 2 biến khác nhau có mối tương quan rất mạnh, thậm chí chắc chắn.

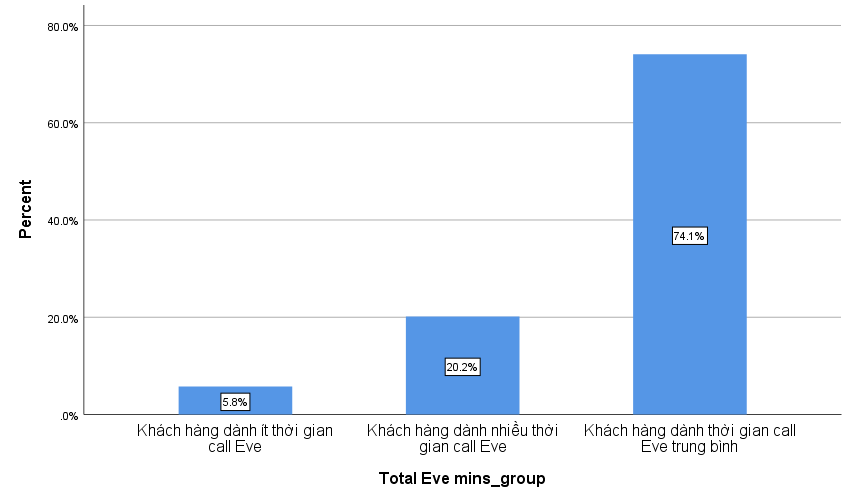
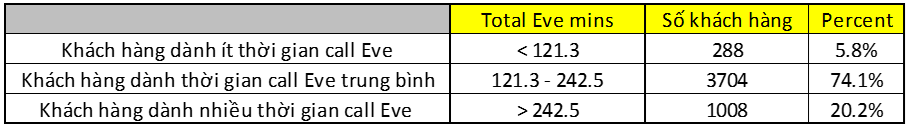
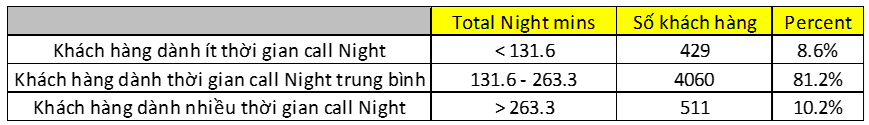
Ví dụ Total Day minutes với Total Day charge, giải thích dễ hiểu, các công ty viễn thông thường tính tiền cuộc gọi dựa trên số phút do đó khi số phút tăng, số tiền bỏ ra cũng sẽ tăng, giải thích tương tự cho Evening, Night, và International call.

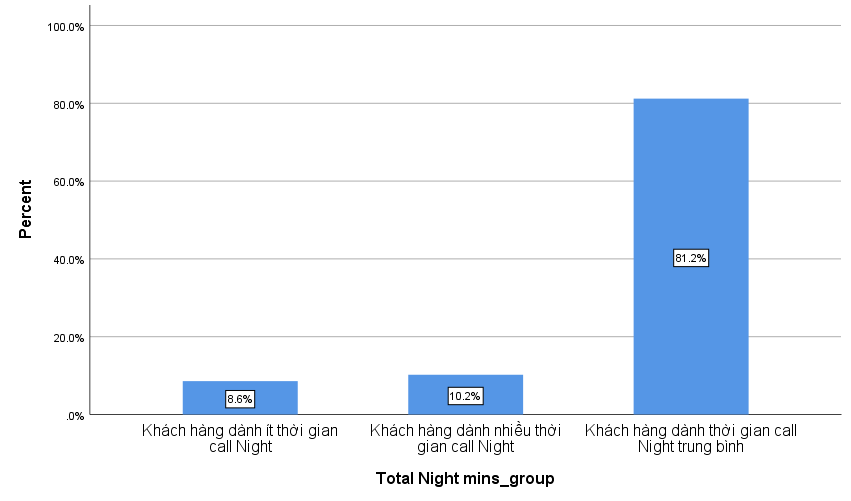
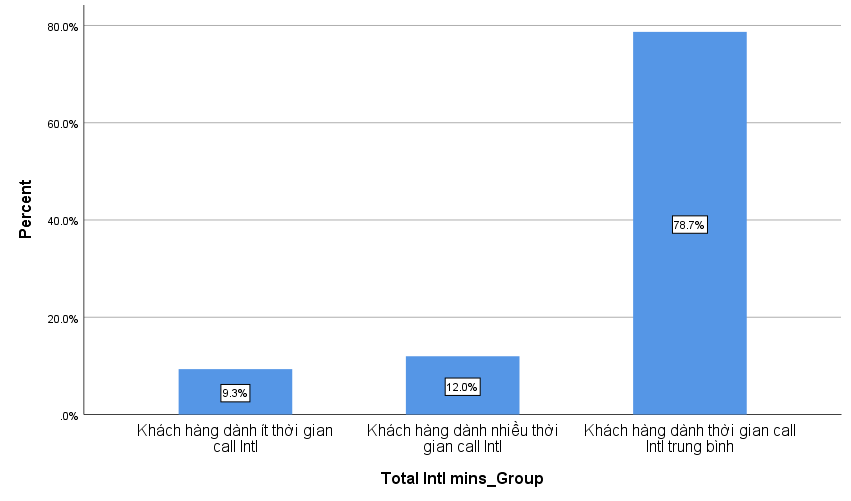
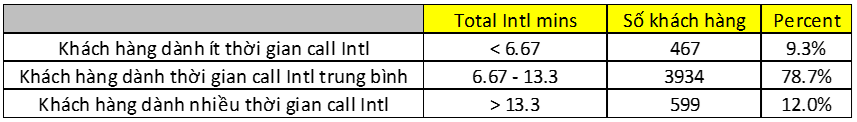
Như vậy chúng ta nên loại bỏ các biến chi phí đơn lẻ của Day, Evening, Night, International call, chúng ta chỉ quan tâm tổng chi phí hay nói cách khác tổng doanh thu công ty nhận được từ mỗi khách hàng.

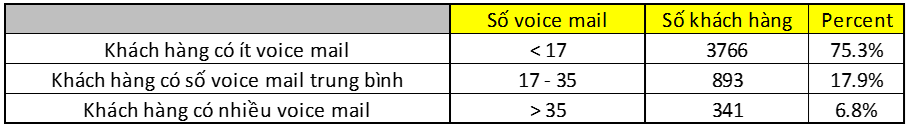
Nói riêng về các biến tổng số cuộc gọi thì chúng ta chưa thể đánh giá số cuộc gọi tác động như thế nào đến chi phí vì một người có thể gọi ít cuộc, nhưng mỗi cuộc gọi kéo dài rất lâu hoặc ngược lại. Tuy nhiên số cuộc gọi có nhiều ý nghĩa trong phân tích hành vi khách hàng, hỗ trợ quá tình phân khúc.

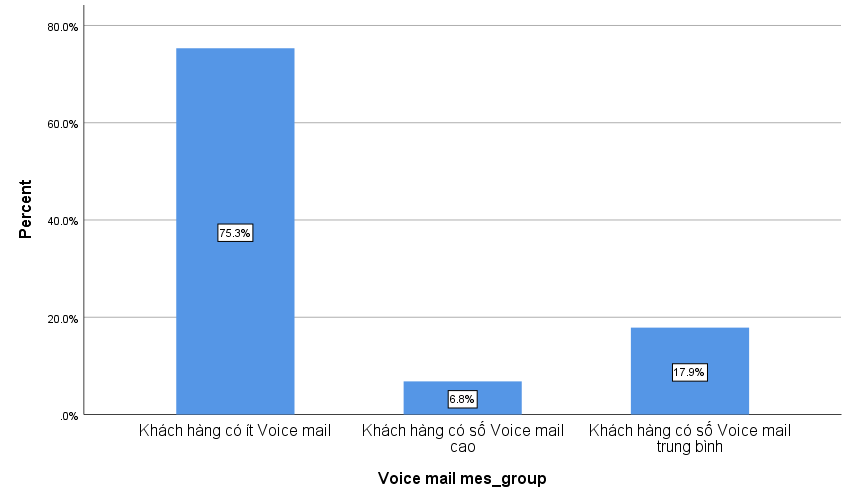
Quay trở lại với ví dụ, chúng ta tìm hiểu biến Account length. Chúng ta phân giá trị thành các khoảng để đánh giá. Đơn giản nhất, chúng ta có thể chia thành 3 khoảng, làm cho Account length và các biến định lượng khác thông tin về hành vi khách hàng sử dụng dịch vụ.

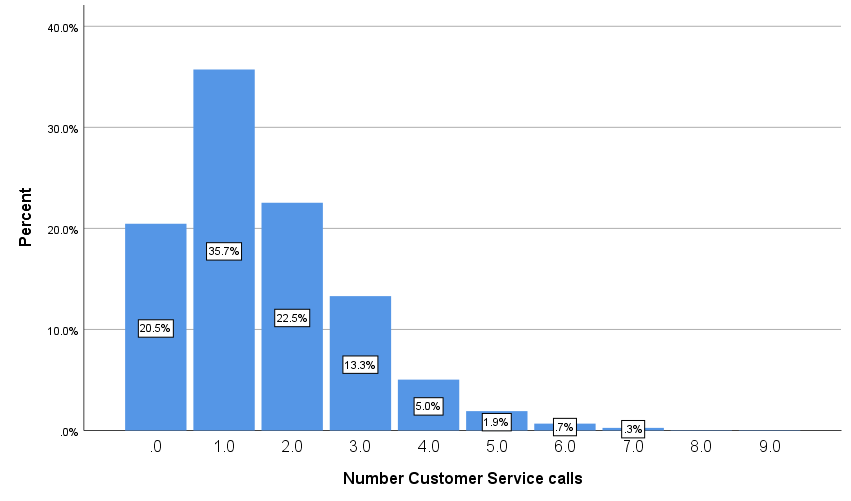




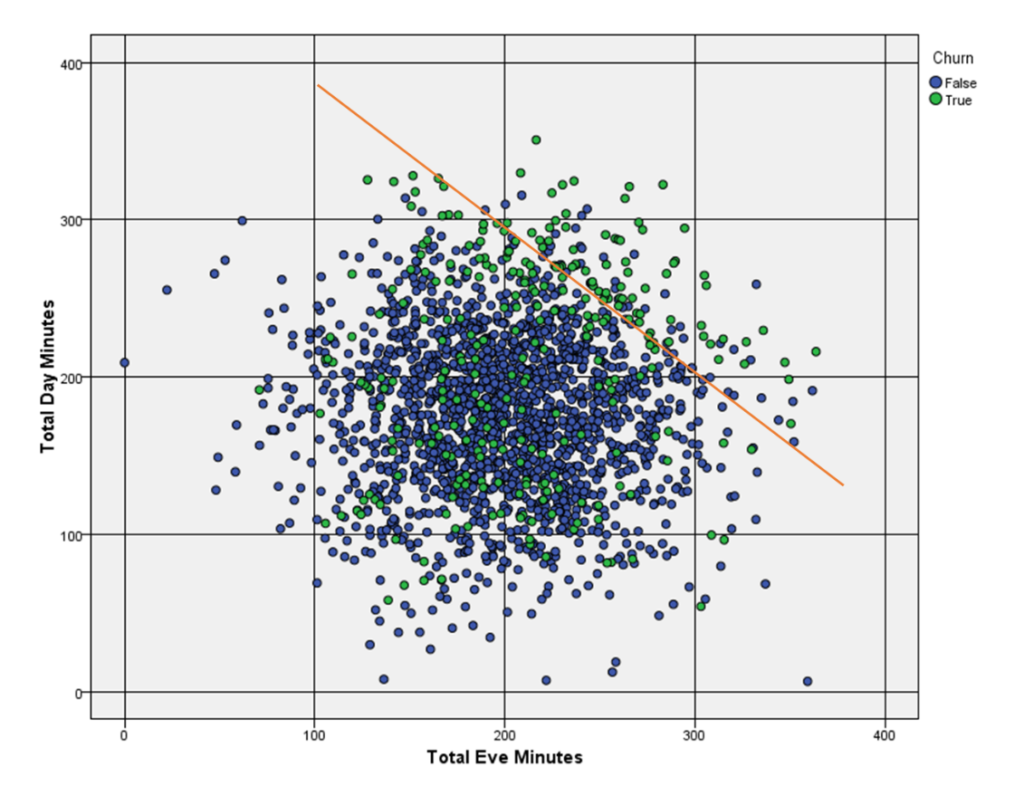




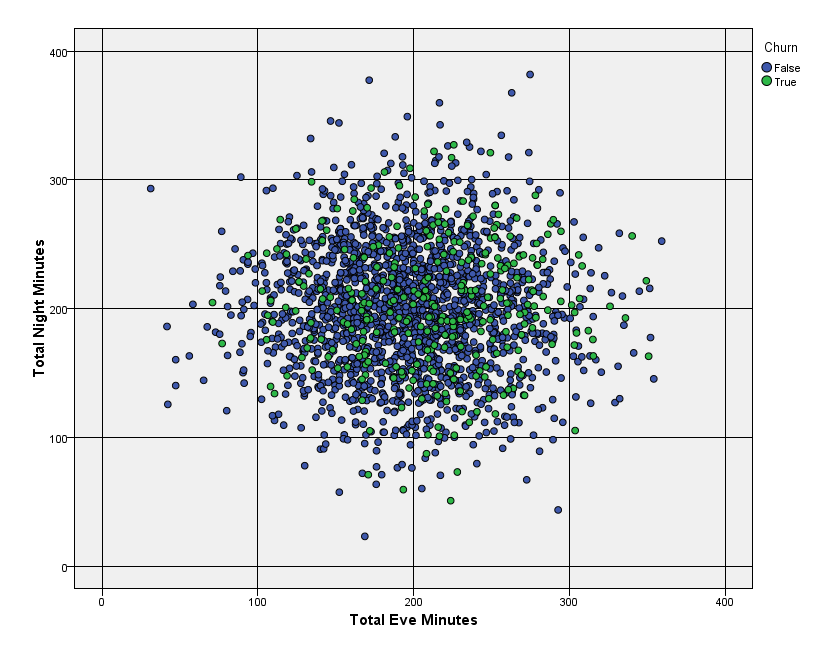


Tỷ lệ khách hàng theo số cuộc gọi đến tổng đài chăm sóc khách hàng

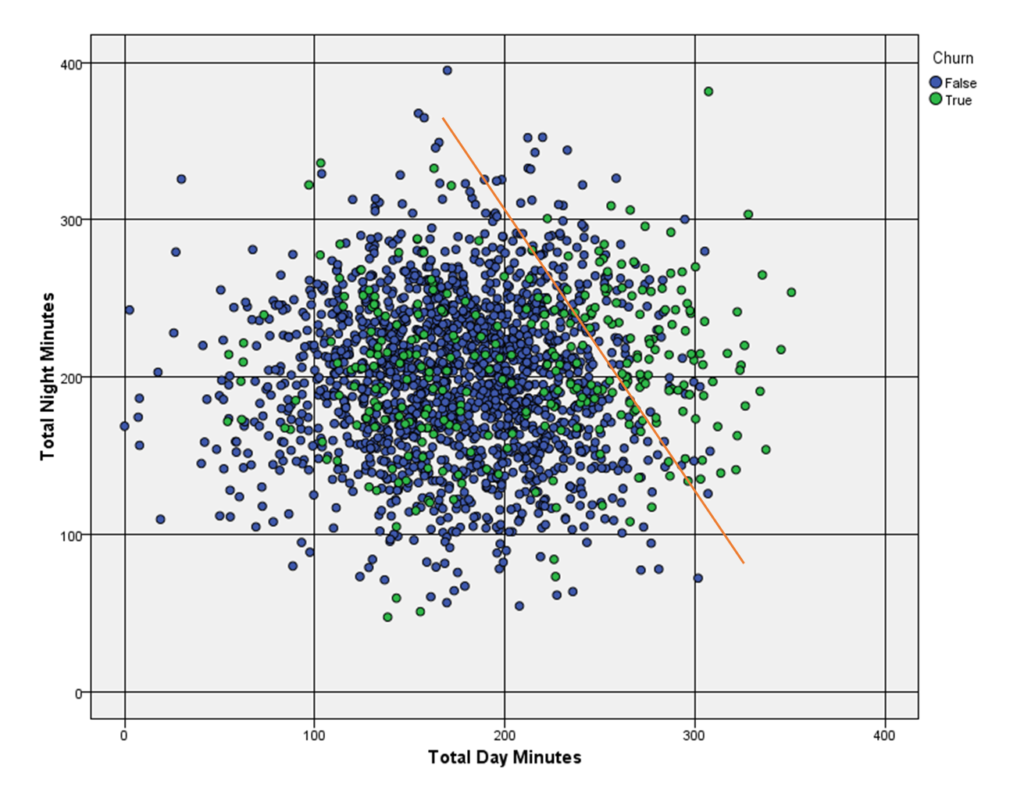
Tiếp theo chúng ta tìm hiểu các biến định lượng tác động như thế nào đến tỷ lệ khách hàng rời dịch vụ.



Các khách hàng có số thời gian gọi điện vào ban ngày và thời gian gọi điện vào chiều tối cao khả năng rời dịch vụ cao và ngược lại. Qua đây chứng tỏ, khi khách hàng có nhu cầu cao hơn cho cả các cuộc gọi vào ban ngày, chiều tối thì dịch vụ hiện tại của công ty chưa đáp ứng, thỏa mãn mong muốn của họ như có thể giá cước mắc, họ đi kiếm công ty khác với giá rẻ hơn. Trong khi các khách hàng không có nhu cầu cao, thì họ không rời dịch vụ.

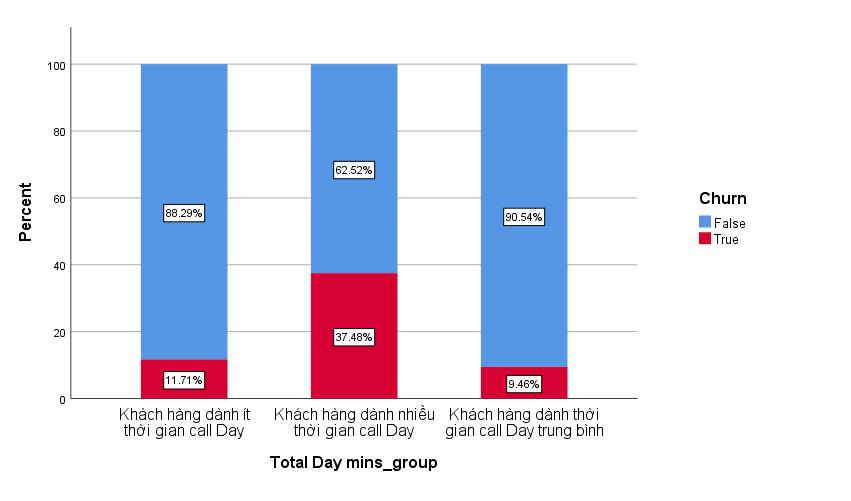


Nếu chúng ta kết hợp 2 biến thời gian gọi vào ban đêm và cả thời gian gọi vào buổi tối như hình trên thì không thấy được sự tác động rõ rệt nào lên tỷ lệ rời dịch vụ.

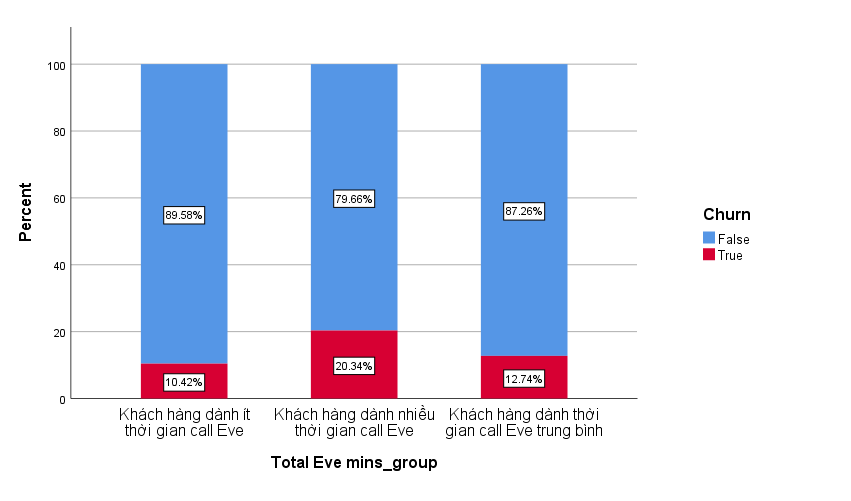
Tuy nhiên đối với các khách hàng có thời gian gọi điện vào ban đêm cao hơn mức trung bình, và thời gian gọi vào ban ngày cao thì có khả năng rời dịch vụ. Giải thích tương tự như trường hợp đầu tiên.

Xét cả 3 trường hợp thì chúng ta có thể thấy khách hàng có thời gian gọi ban ngày Day mins cao, thì khả năng rời dịch vụ sẽ cao. Nhắc day ở đây là tính từ buổi sáng đến lúc đầu chiều, hay nói cách khác là trong giờ làm việc. Có thể những khách hàng này là những người làm việc sử dụng phone nhiều.. Đây chỉ mới là phỏng đoán.

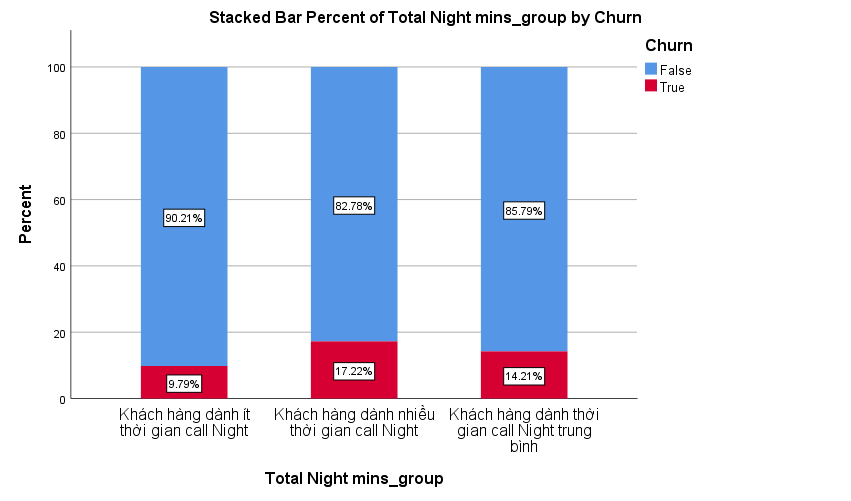
Chúng ta tìm hiểu mối quan hệ từng biến trên với tỷ lệ rời dịch vụ.



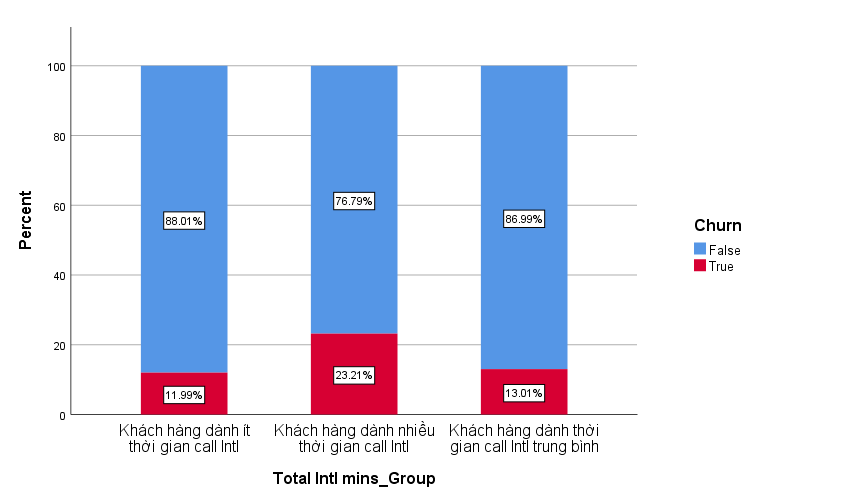
Khách hàng dành nhiều thời gian call vào ban ngày tỷ lệ rời dịch vụ cao hơn, điều này khẳng định lần nữa cho các kết luận ở trên. Tỷ lệ churn giữa nhóm cao nhất, và các nhóm còn lại chênh lệch lớn.



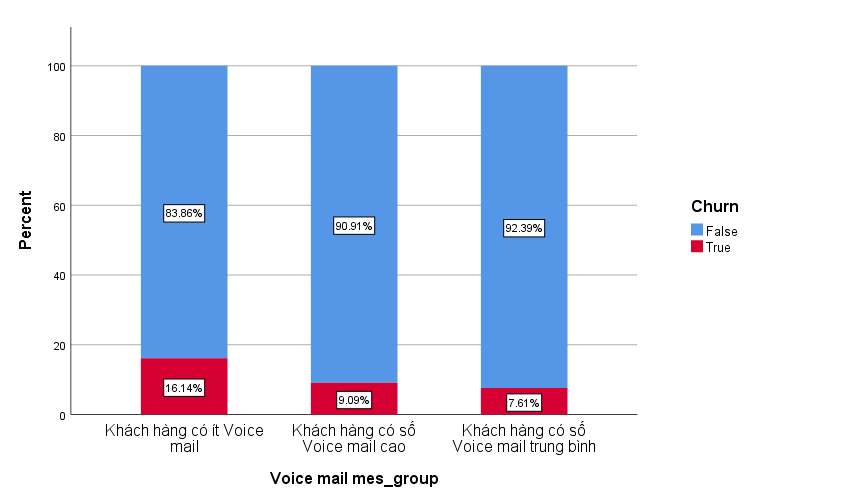
Khách hàng dành nhiều thời gian gọi vào chiều tối, Eve call, cũng có tỷ lệ rời dịch vụ cao hơn, tuy nhiên khác với Day call, tỷ lệ churn ở mỗi nhóm không quá chênh lệch



Tương tự như Eve call, trường hợp Night call cũng vậy. Sau cùng, công ty nên quan tâm vào Day call, để xem liệu các dịch vụ của mình có thỏa mãn nhóm khách hàng có nhu cầu gọi điện nhiều vào ban ngày hay không. Tiếp tục về dịch vụ gọi quốc tế.



Tương tự.



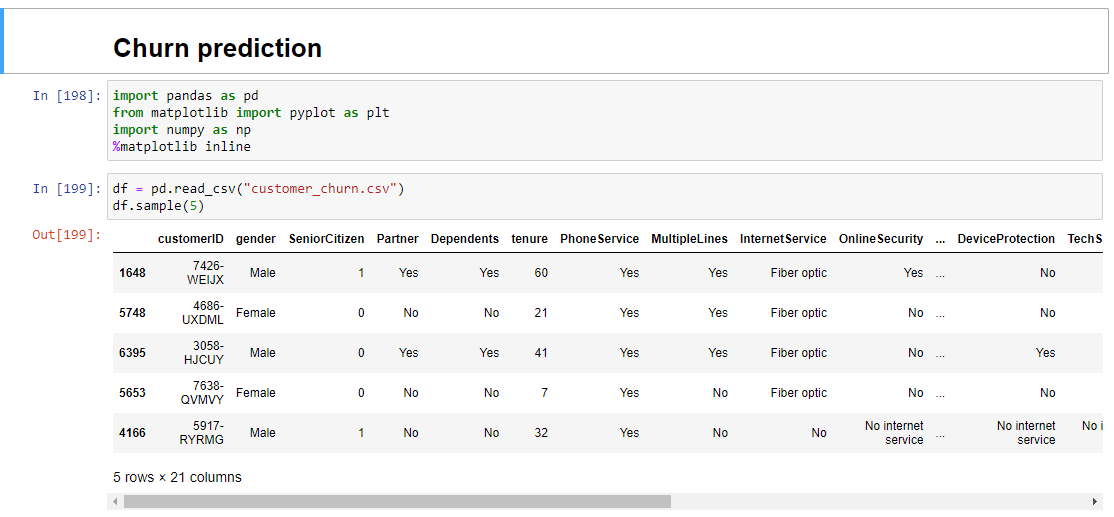
Về voice mail, thì đây là dấu hiệu tích cực mà cũng vừa tiêu cực của công ty. Khi những khách hàng có số hộp thư thoại cao, hay trung bình thì tỷ lệ rời dịch vụ thấp, có thể suy ra dịch vụ hộp thư thoại, thỏa mãn nhu cầu cao của khách hàng, ngược lại các khách hàng có voice mail ít thì tỷ lệ rời dịch vụ cao. Khả năng là do họ không hài lòng với dịch vụ voice mail nên họ ít sử dụng và sau đó rời dịch vụ nhưng cũng có thể vì lý do khác như trình bày ở trên.

Còn rất nhiều mối quan hệ khác…

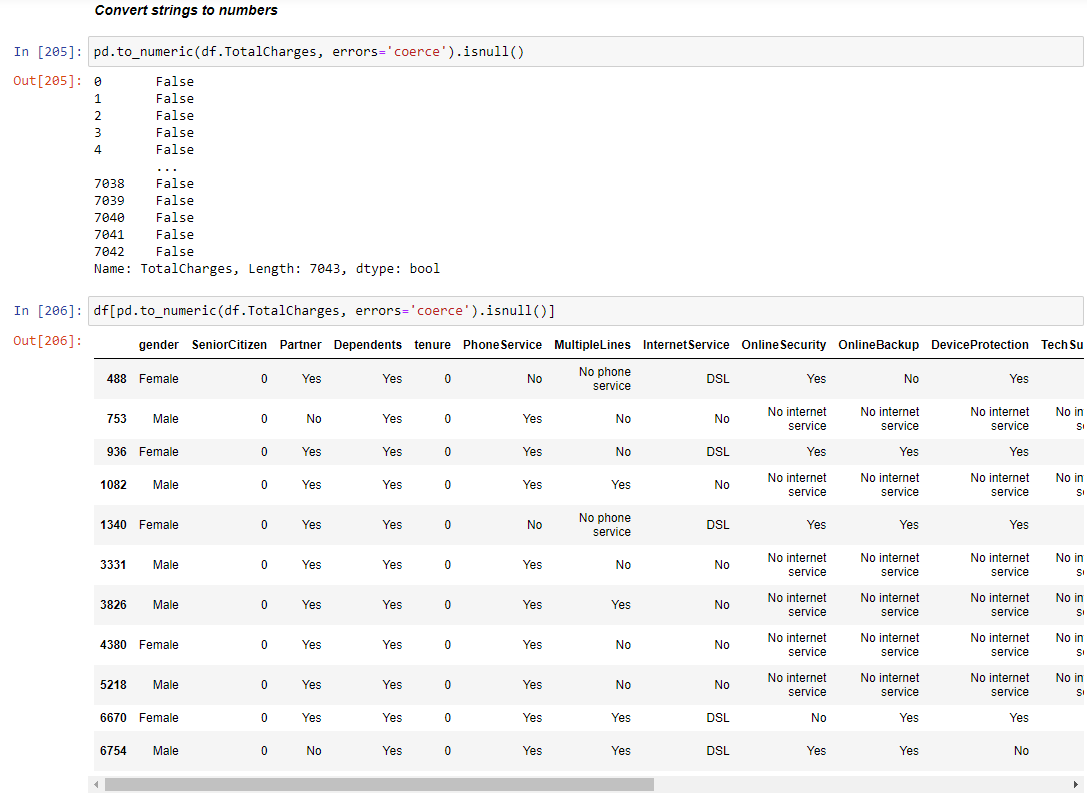
Các số liệu lấy tổng theo tháng, nên các đánh giá đều chưa thể chính xác hoàn toàn. Ví dụ có nhiều khách hàng ít dành thời gian gọi điện vào buổi tối, ban đêm trong tuần, nhưng vào cuối tuần thì họ dành gấp n lần thời gian để xả stress, tán dóc với bạn bè,…Chúng ta không có thông tin về hành vi sử dụng xét theo trung bình ngày, ví dụ chúng ta không tính được thời lượng gọi điện trung bình cho mỗi cuộc gọi vào ban ngày trong 1 ngày là bao nhiêu. Do đó các kết luận đưa ra chỉ mang tính chất chung chung, chưa thể khẳng đinh tuyệt đối.

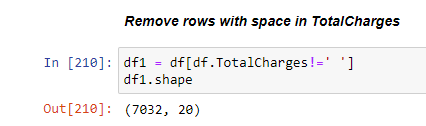
1. **NỘI DUNG THỰC HÀNH**
2. **Mô hình phân loại dựa trên Weighted Average Ensemble:**

Mô hình đọc dữ liệu khách hàng rời bỏ dịch vụ viễn thông từ Kaggle.

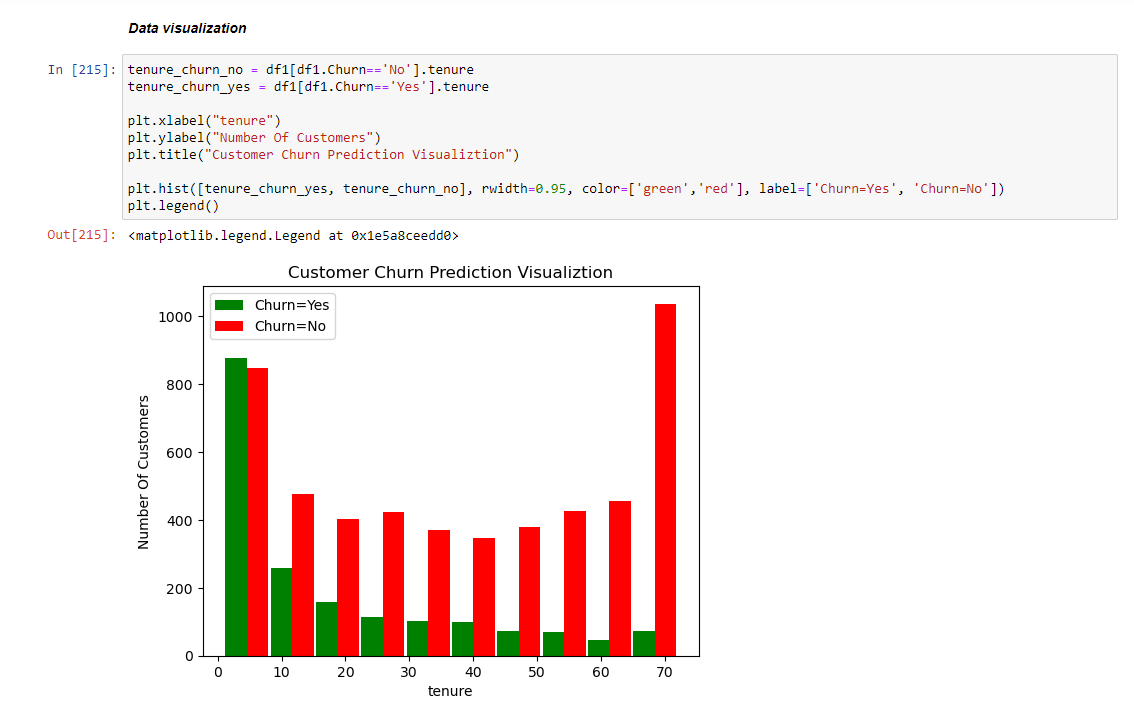


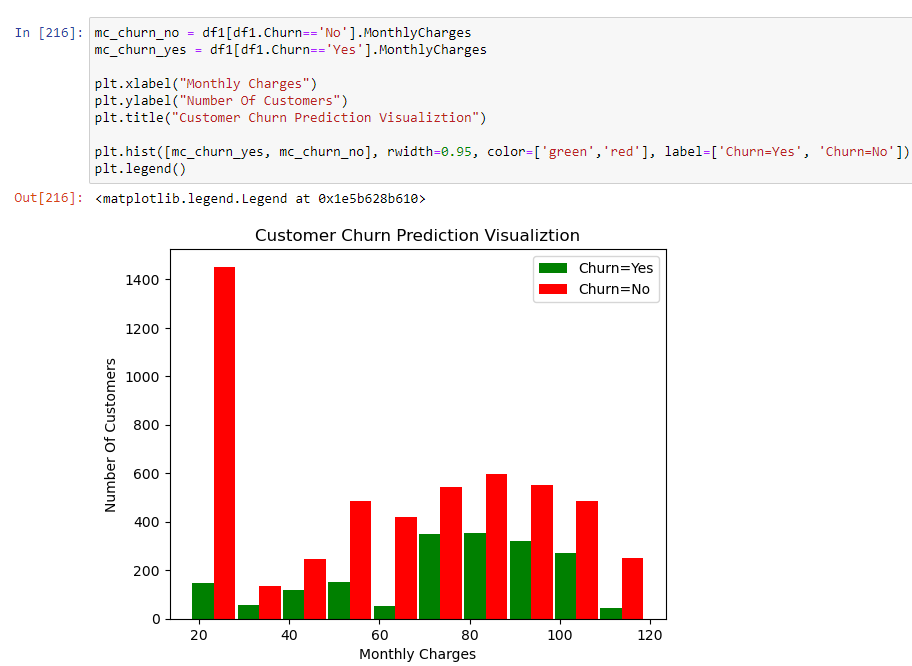
Các bước chuẩn hóa và phân tích dữ liệu

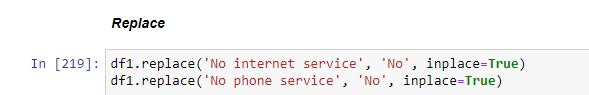


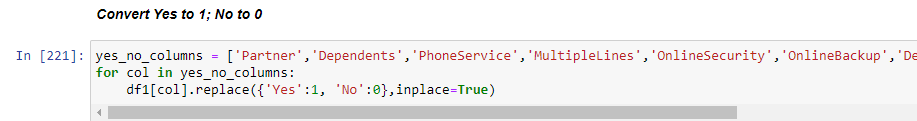




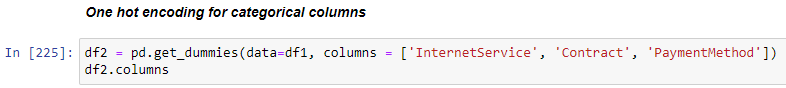


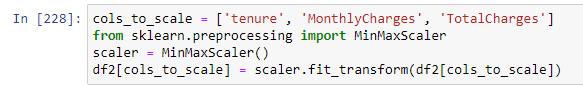


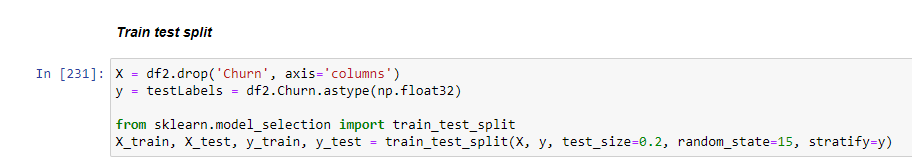


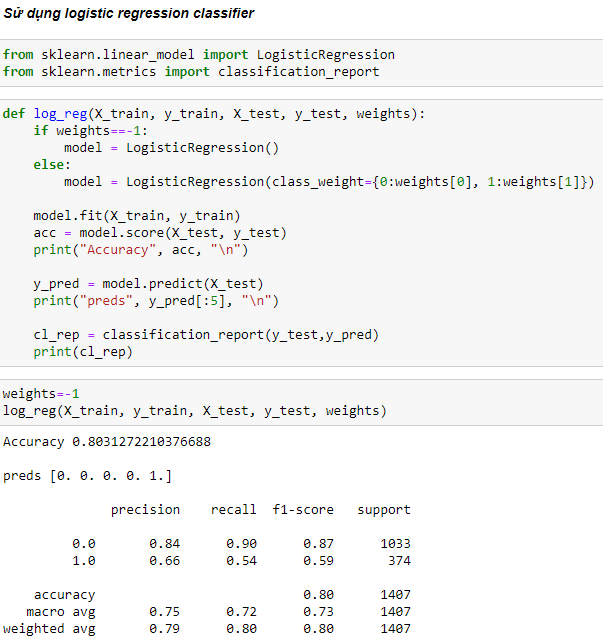




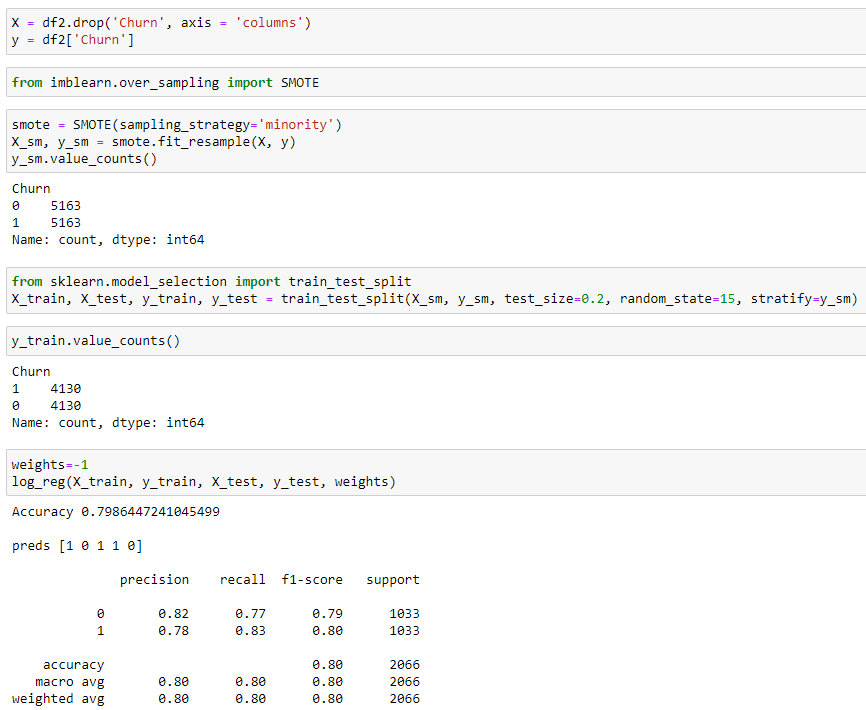


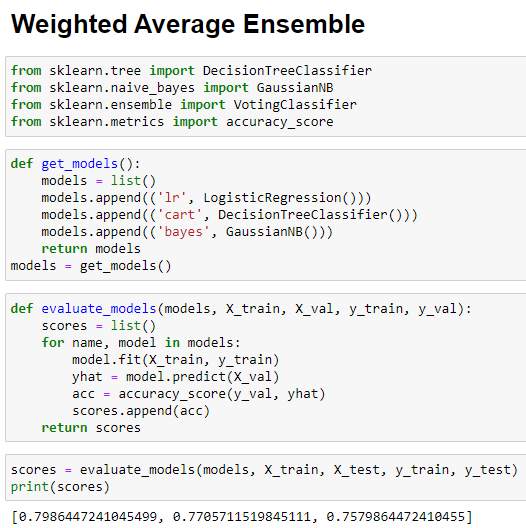




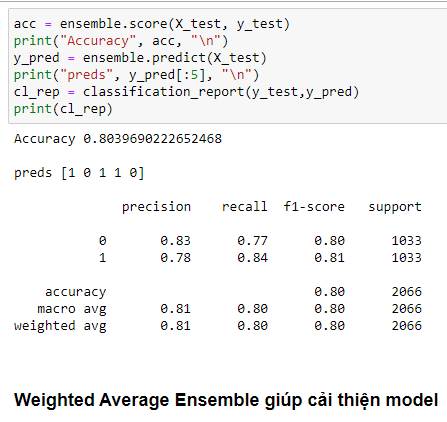


Dùng SMOTE để tránh mất cân bằng dữ liệu





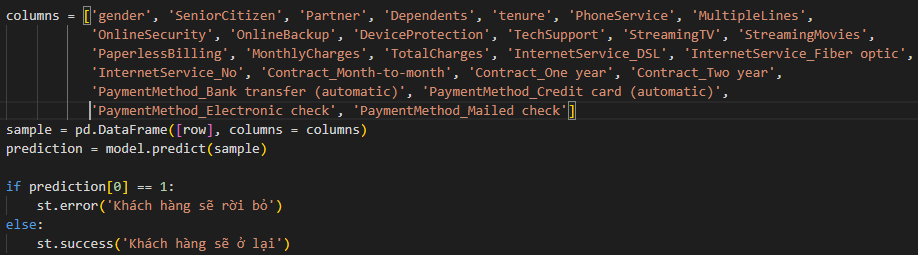


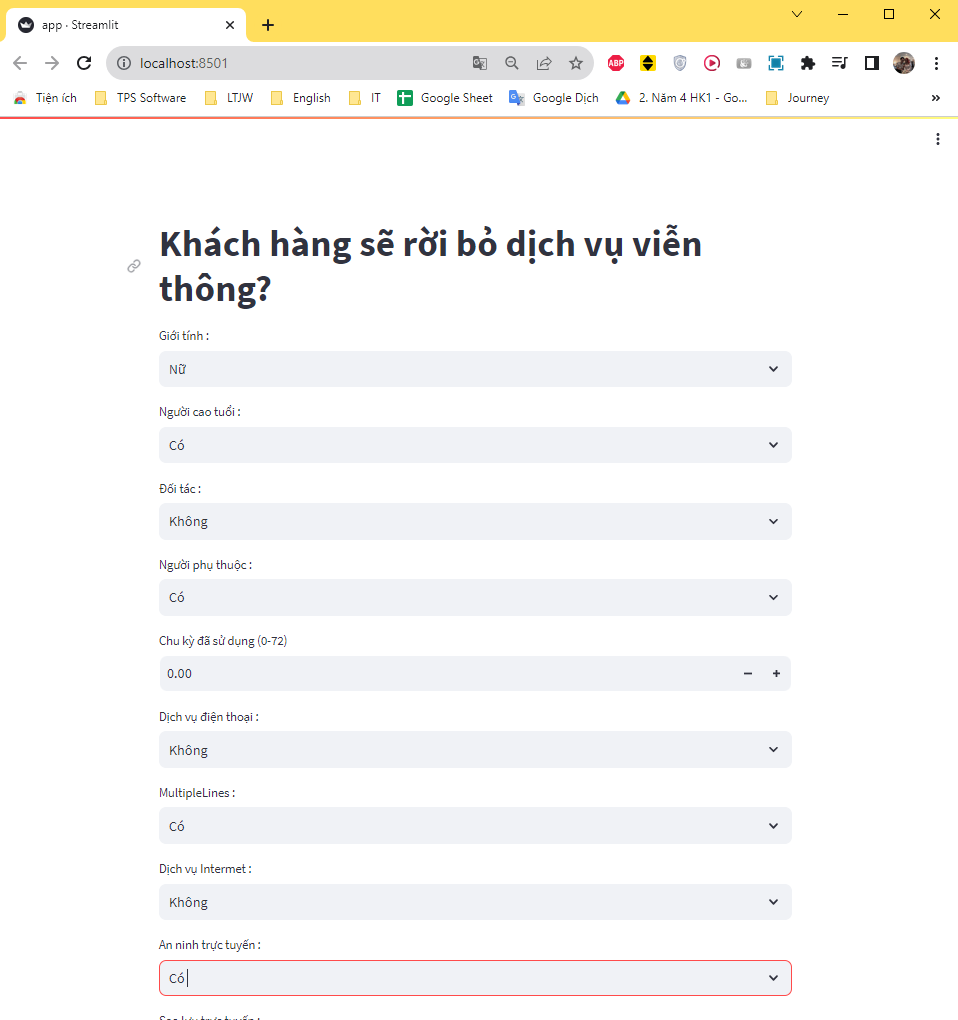


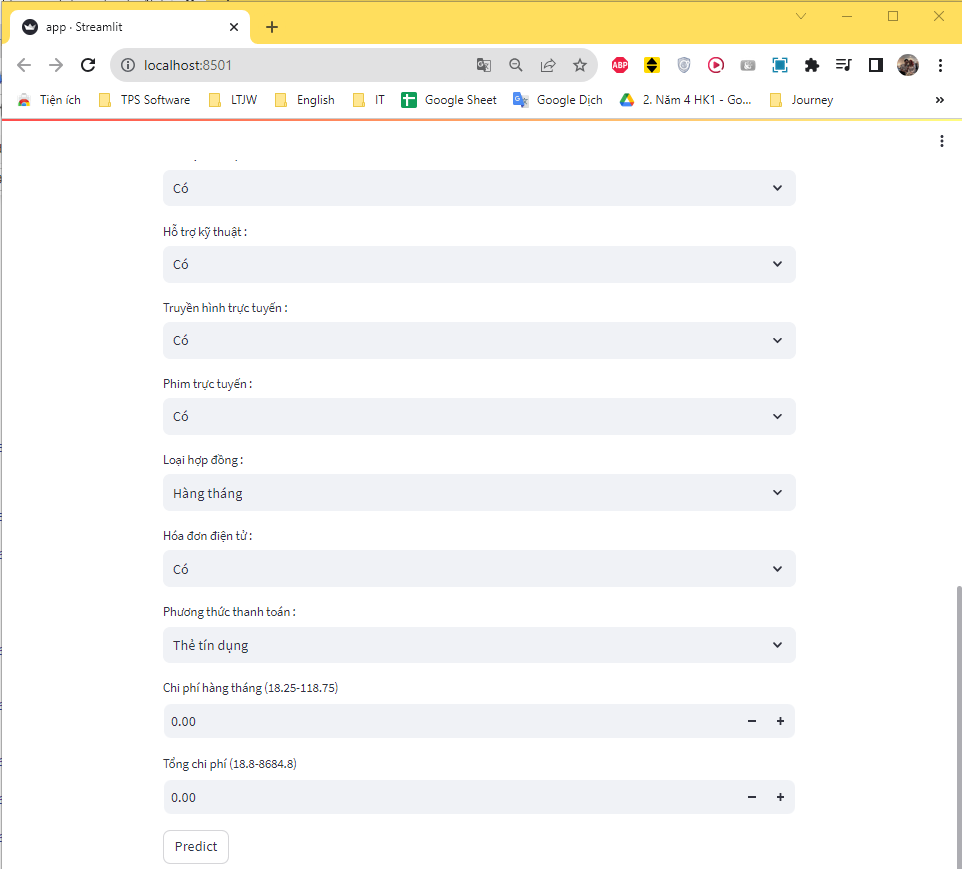


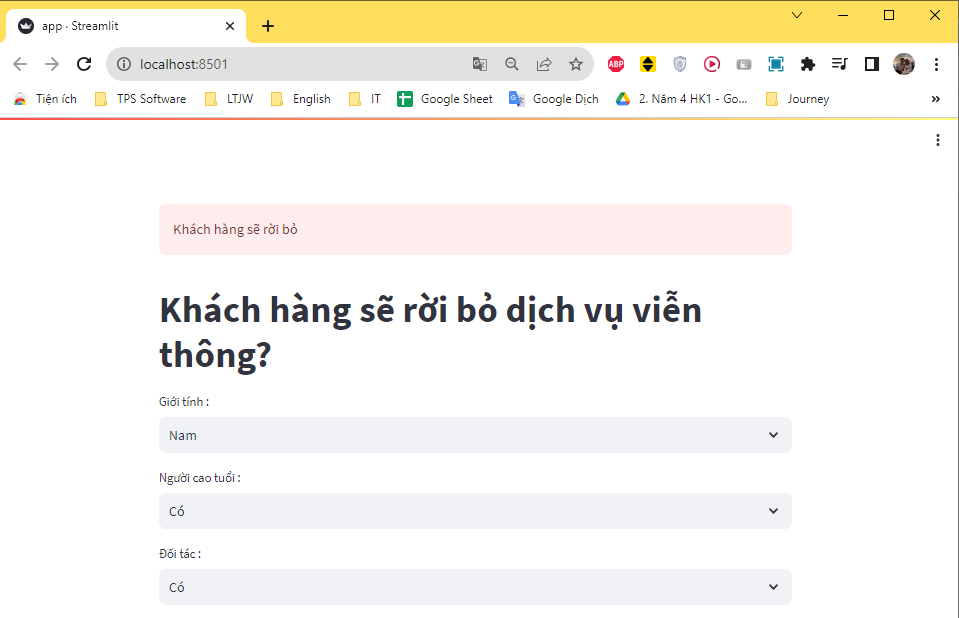
1. **Ứng dụng minh họa dự báo khách hàng rời bỏ**

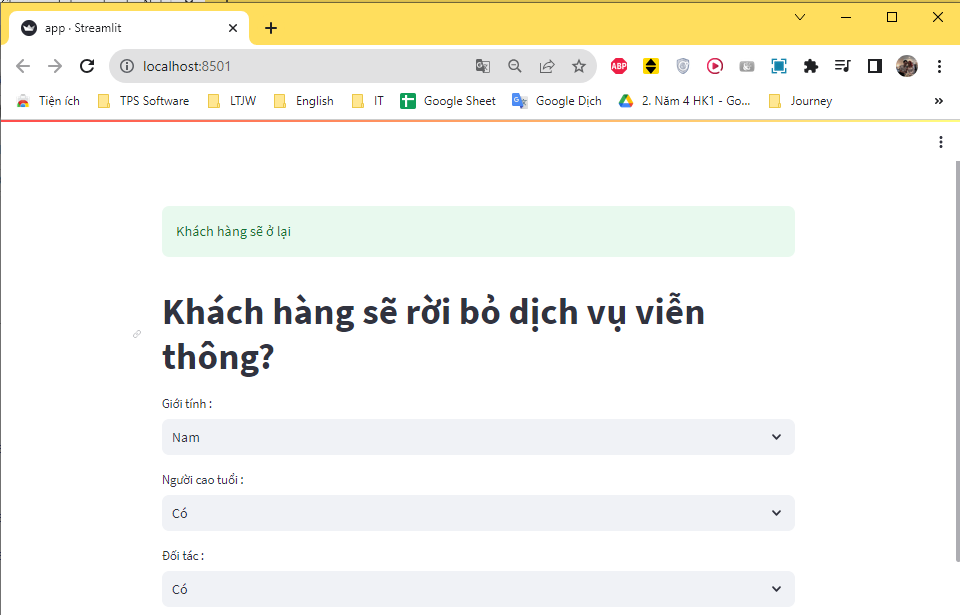












**DANH MỤC TÀI LIỆU THAM KHẢO**

1. Tập dữ liệu training: https://www.kaggle.com/code/bandiatindra/telecom-churn-prediction/notebook
2. Triển khai mô hình Weighted Average Ensemble bằng Python: https://machinelearningmastery.com/weighted-average-ensemble-with-python/
3. Dự án học máy dự báo khách hàng Telco rời bỏ:

https://towardsdatascience.com/end-to-end-machine-learning-project-telco-customer-churn-90744a8df97d

1. Phân tích dữ liệu rời đi của công ty viễn thông: https://www.analyticsvidhya.com/blog/2022/01/churn-analysis-of-a-telecom-company/

**PHỤ LỤC**