TRƯỜNG ĐẠI HỌC KINH TẾ

**KHOA THỐNG KÊ – TIN HỌC**



**BÁO CÁO THỰC TẬP NGHỀ NGHIỆP**

**NGÀNH HỆ THỐNG THÔNG TIN QUẢN LÝ**

**CHUYÊN NGÀNH QUẢN TRỊ HỆ THỐNG THÔNG TIN**

**ỨNG DỤNG AI DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ DỊCH VỤ VIỄN THÔNG**

Sinh viên thực hiện : **Nguyễn Thị Duyên**

: **Trần Thị Hằng**

: **H Ngač Niê**

Lớp : **47K21.1**

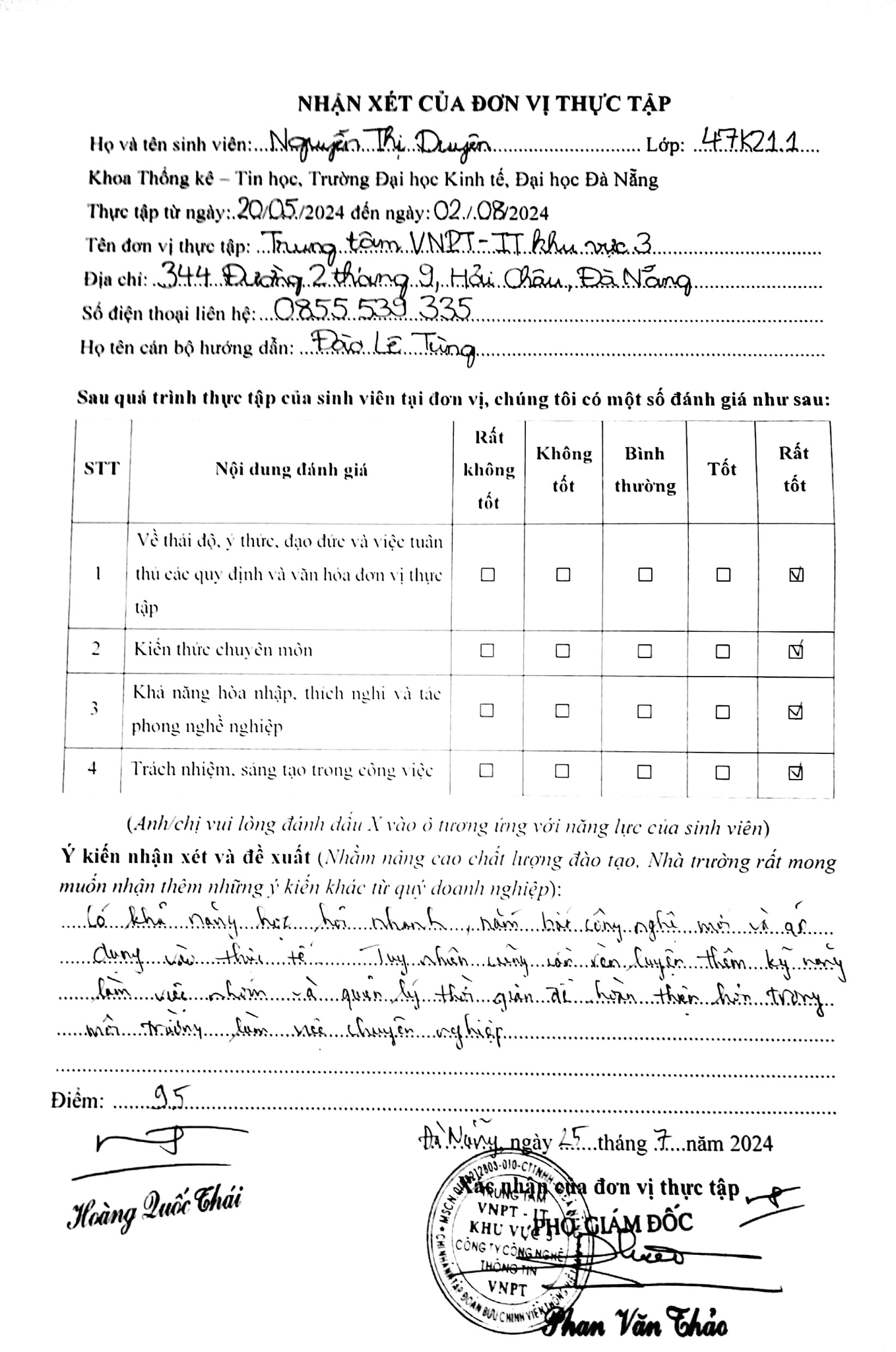
Đơn vị thực tập : **Trung tâm VNPT-It Khu vực 3**

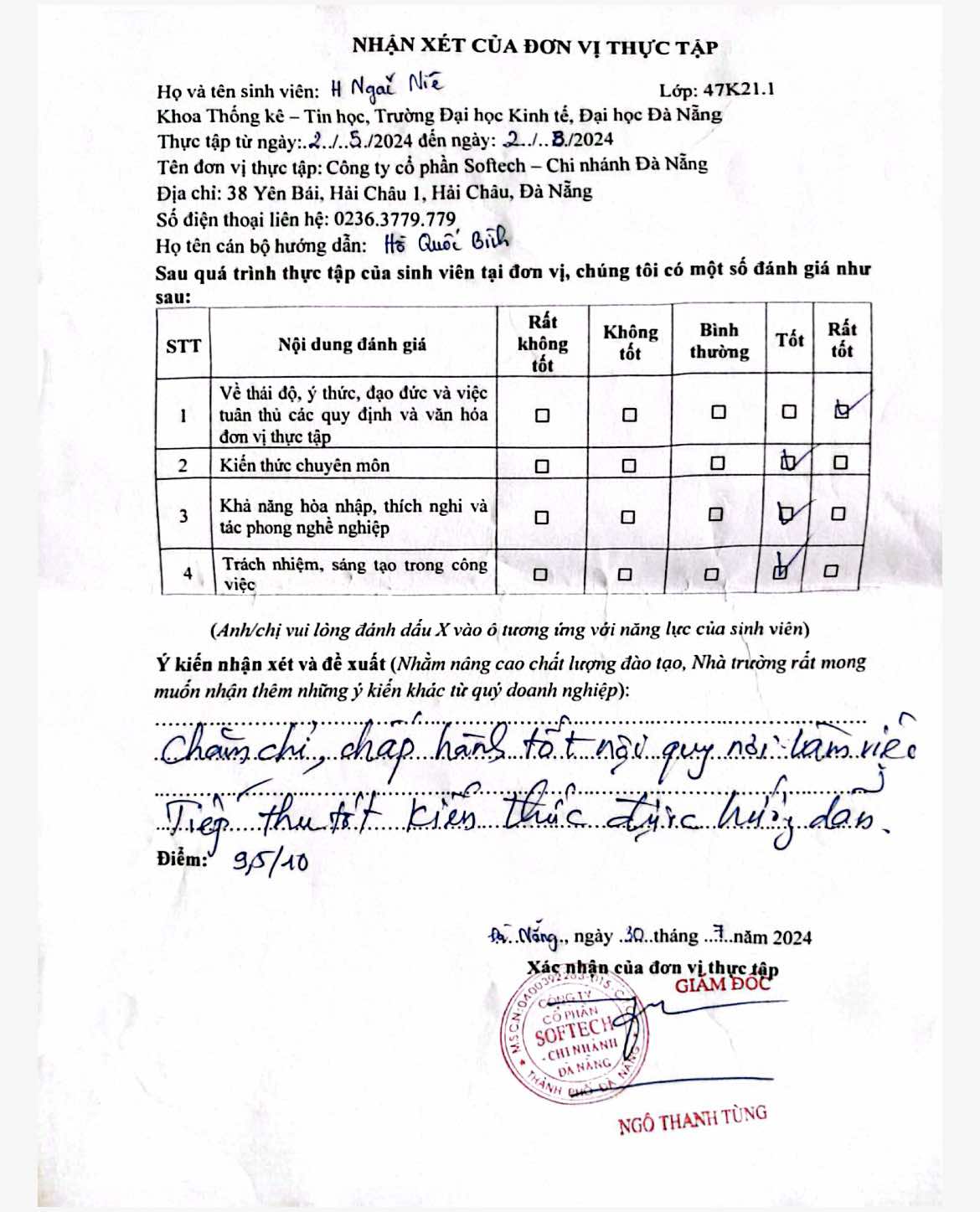
: **Công ty cổ phần Softech**

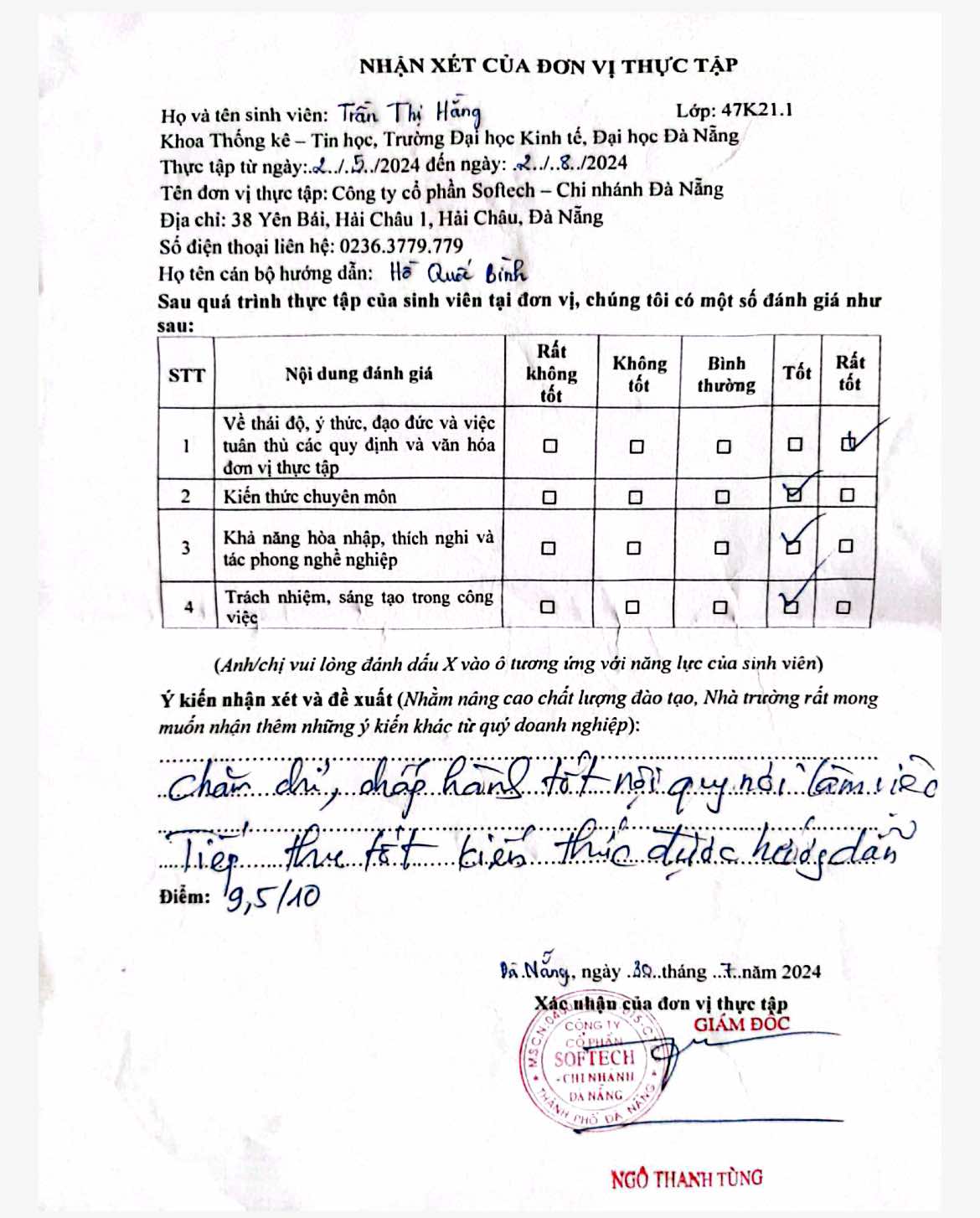
Giảng viên hướng dẫn : **TS. Phan Đình Vấn**

***Đà Nẵng, 8/202******4***

**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**







# LỜI CẢM ƠN

Để có thể hoàn thành báo cáo thực tập nghề nghiệp này, ngoài sự nỗ lực không ngừng của bản thân thì một phần không nhỏ đóng góp nên thành công ấy là nhờ sự hướng dẫn, dạy dỗ của các thầy cô trong Khoa Thống Kê – Tin Học nói riêng và trong trường Đại học Kinh tế - Đại học Đà Nẵng nói chung trong suốt thời gian vừa qua.

Đặc biệt, em xin gửi lời biết ơn sâu sắc nhất đến TS. Phan Đình Vấn - Người thầy đã luôn theo sát chỉ dẫn, giúp đỡ nhiệt tình và cho em nhiều lời khuyên cũng như dặn dò trong suốt quá trình thực tập, giúp em hoàn thành kỳ thực tập cũng như bài báo cáo một cách tốt nhất có thể. Em xin chân thành cảm ơn thầy!

Ngoài ra, nhóm chúng em với ba thành viên Nguyễn Thị Duyên (thực tập sinh tại Trung tâm VNPT - IT KV3) và Trần Thị Hằng, H Ngač Niê (thực tập sinh tại Công ty cổ phần Softech – Chi nhánh Đà Nẵng). Chúng em xin gửi lời cảm ơn đến 2 công ty đã tạo điều kiện thuận lợi để chúng em có cơ hội học hỏi và tiếp thu nhiều kinh nghiệm chuyên môn và cung cấp nhiều tài liệu hướng dẫn giúp em hoàn thành bài báo cáo thực tập nghề nghiệp.

Vì kiến thức bản thân còn nhiều hạn chế, trong quá trình hoàn thiện báo cáo không tránh khỏi sai sót. Vì vậy, em rất mong nhận được sự thông cảm và những ý kiến đóng góp, chỉ bảo từ quý thầy cô, cũng như quý công ty để em có thể khắc phục và hoàn thiện bản thân mình hơn trong quá trình làm việc sau này. Cuối cùng, em kính chúc quý thầy cô luôn dồi dào sức khỏe và thành công trong sự nghiệp giảng dạy cao quý.

Em xin chân thành cảm ơn!

# LỜI CAM ĐOAN

Chúng em xin cam đoan những kết quả đạt được trong báo cáo này là do nhóm chúng em nghiên cứu, tổng hợp và thực hiện dựa sự hướng dẫn của thầy Phan Đình Vấn, anh Đào Lê Tùng-Trung tâm VNPT - IT KV3 và anh Hồ Quốc Bình-Công ty cổ phần Softech. Ngoài ra không sao chép lại bất kỳ điều gì của người khác. Đề tài, nội dung báo cáo thực tập là sản phẩm mà em đã nỗ lực nghiên cứu trong quá trình học tập tại trường cũng như tham gia thực tập tại công ty. Những nội dung được trình bày trong báo cáo được tham khảo và tổng hợp từ các nguồn tài liệu khác nhau. Những tài liệu mà chúng em đã tham khảo sẽ được trích dẫn rõ ràng.

Chúng em xin chắc chắn rằng toàn bộ nội dung bài báo cáo là trung thực và không hề tồn tại sự gian lận. Nếu có điều gì sai phạm, chúng em xin chịu mọi hình thức kỷ luật theo quy định.

MỤC LỤC

[LỜI CẢM ƠN v](#_Toc173104467)

[LỜI CAM ĐOAN vi](#_Toc173104468)

[DANH MỤC HÌNH ẢNH x](#_Toc173104469)

[DANH MỤC BẢNG BIỂU xiii](#_Toc173104470)

[DANH MỤC CÁC TỪ VIẾT TẮT xiv](#_Toc173104471)

[LỜI MỞ ĐẦU 1](#_Toc173104472)

[CHƯƠNG 1. TỔNG QUAN VỀ TRUNG TÂM VNPT-IT KHU VỰC 3 & CÔNG TY CỔ PHẦN SOFTECH &TỔNG QUAN ĐỀ TÀI DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ DỊCH VỤ VIỄN THÔNG 3](#_Toc173104473)

[1.1. Giới thiệu tổng quan về doanh nghiệp thực tập VNPT-IT khu vực 3 3](#_Toc173104474)

[**1.1.1. Quá trình hình thành và phát triển của công ty 3**](#_Toc173104475)

[**1.1.2. Tầm nhìn, sứ mệnh 4**](#_Toc173104476)

[**1.1.3. Lĩnh vực hoạt động 4**](#_Toc173104477)

[**1.1.4. Cơ cấu tổ chức 5**](#_Toc173104478)

[1.2. Tổng quan về công ty cổ phần 6](#_Toc173104479)

[**1.2.1. Quá trình hình thành và phát triển công ty 6**](#_Toc173104480)

[**1.2.2. Tầm nhìn sự mệnh 6**](#_Toc173104481)

[**1.2.3. Sứ mệnh hoạt động 6**](#_Toc173104482)

[**1.2.4. Cơ cấu tổ chức 7**](#_Toc173104483)

[1.3. Tổng quan về đề tài dự đoán khách hàng rời bỏ dịch vụ 7](#_Toc173104484)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 9](#_Toc173104485)

[2.1. Lý thuyết các mô hình dự đoán 9](#_Toc173104486)

[**2.1.1. Mô hình Rừng ngẫu nhiên (Random Forest) 9**](#_Toc173104487)

[**2.1.2. Mô hình Hồi quy Logistic (Logistic Regression) 10**](#_Toc173104488)

[**2.1.3. Mô hình KNN (K-Nearest Neighbors) 11**](#_Toc173104489)

[**2.1.4. Mô hình Naïve Bayes 13**](#_Toc173104490)

[**2.1.5. Mô hình XGBoost 15**](#_Toc173104491)

[**2.1.6. Mô hình Feedforward Neural Network (FNN) 16**](#_Toc173104492)

[2.2. Các chỉ số đánh giá mô hình 19](#_Toc173104493)

[2.3. Các công cụ hỗ trợ cho việc phân tích 20](#_Toc173104494)

[**2.3.1. Jupyter Notebook 20**](#_Toc173104495)

[**2.3.2. Power BI 22**](#_Toc173104496)

[CHƯƠNG 3. MÔ TẢ VÀ TRỰC QUAN HÓA DỮ LIỆU 25](#_Toc173104497)

[3.1. Thống kê mô tả 25](#_Toc173104498)

[3.2. Mô tả bài toán 27](#_Toc173104499)

[3.3. Giải thích dữ liệu 28](#_Toc173104500)

[3.4. Trực quan hóa dữ liệu 29](#_Toc173104501)

[CHƯƠNG 4. XÂY DỰNG MÔ HÌNH DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ DỊCH VỤ…....... 40](#_Toc173104502)

[4.1. Frame work 40](#_Toc173104503)

[4.2. Tiền xử lý dữ liệu 40](#_Toc173104504)

[**4.2.1. Kiểm tra dữ liệu tổng quát 40**](#_Toc173104505)

[**4.2.2. Chuyển đổi dữ liệu 41**](#_Toc173104506)

[**4.2.3. Làm sạch dữ liệu 41**](#_Toc173104507)

[**4.2.4. Giảm chiều dữ liệu 44**](#_Toc173104508)

[**4.2.5. Chuẩn hóa dữ liệu 46**](#_Toc173104509)

[4.3. Mô hình Random Forest 47](#_Toc173104510)

[**4.3.1. Xây dựng mô hình 47**](#_Toc173104511)

[**4.3.2. Đánh giá mô hình 50**](#_Toc173104512)

[4.4. Mô hình Logistic Regression 52](#_Toc173104513)

[**4.4.1. Xây dựng mô hình 52**](#_Toc173104514)

[**4.4.2. Đánh giá mô hình 53**](#_Toc173104515)

[4.5. Mô hình K-Nearest Neighbors 54](#_Toc173104516)

[**4.5.1. Xây dựng mô hình 54**](#_Toc173104517)

[**4.5.2. Đánh giá mô hình 55**](#_Toc173104518)

[4.6. Mô hình Naïve Bayes 57](#_Toc173104519)

[**4.6.1. Xây dựng mô hình 57**](#_Toc173104520)

[**4.6.2. Đánh giá mô hình 59**](#_Toc173104521)

[4.7. Mô hình XGBoost 60](#_Toc173104522)

[**4.7.1. Xây dựng mô hình 60**](#_Toc173104523)

[**4.7.2. Đánh giá mô hình 61**](#_Toc173104524)

[4.8. Mô hình Feedforward Neural Network 63](#_Toc173104525)

[**4.8.1. Xây dựng mô hình 63**](#_Toc173104526)

[**4.8.2. Đánh giá mô hình 65**](#_Toc173104527)

[4.9. Đánh giá các mô hình 67](#_Toc173104528)

[**4.9.1. Phân tích các chỉ số của mô hình 67**](#_Toc173104529)

[**4.9.2. Kết luận 69**](#_Toc173104530)

[4.10. Chiến lược giữ chân khách hàng 69](#_Toc173104531)

[KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 72](#_Toc173104532)

[TÀI LIỆU THAM KHẢO 73](#_Toc173104533)

[CHECK LIST CỦA BÁO CÁO 75](#_Toc173104534)

[PHỤ LỤC 76](#_Toc173104535)

# DANH MỤC HÌNH ẢNH

[hình 1.Logo Trung tâm VNPT-IT 3](#_Toc173106795)

[hình 2. Sơ đồ cơ cấu tổ chức của công ty VNPT 5](#_Toc173106796)

[hình 3. Logo của CTCP Softech 6](#_Toc173106797)

[hình 4. Cơ cấu tổ chức của công ty Softech 7](#_Toc173106798)

[hình 5.Mô hình Random Forest 9](#_Toc173106799)

[hình 6. Mô hình Hồi quy Logistic 11](#_Toc173106800)

[hình 7.Mô hình KNN 12](#_Toc173106801)

[hình 8. Mô hình Naive bayes 13](#_Toc173106802)

[hình 9.Mô hình XGboost 16](#_Toc173106803)

[hình 10. Mô hình FNN 17](#_Toc173106804)

[hình 11. Logo công cụ Jupyter 21](#_Toc173106805)

[hình 12. Công cụ Power BI 22](#_Toc173106806)

[hình 13. Thống kê mô tả dữ liệu 25](#_Toc173106807)

[hình 14. Trực quan hóa tổng quát về thông tin khách hàng 29](#_Toc173106808)

[hình 15. Thông tin chung về khách hàng và phí tiền 30](#_Toc173106809)

[hình 16.Thông tin nhân khẩu học 30](#_Toc173106810)

[hình 17.Thời gian đăng ký sử dụng dịch vụ 31](#_Toc173106811)

[hình 18. Phương thức thanh toán 31](#_Toc173106812)

[hình 19. Biểu đồ về hóa đơn điện tử 32](#_Toc173106813)

[hình 20. Loại hợp đồng liên quan đến dịch vụ 32](#_Toc173106814)

[hình 21. Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo giới tính 33](#_Toc173106815)

[hình 22. Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo Dependents 33](#_Toc173106816)

[hình 23. Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo đối tác Partner 34](#_Toc173106817)

[hình 24. Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo DeviceProtection và Online backup 34](#_Toc173106818)

[hình 25. Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo PaperlessBiling 35](#_Toc173106819)

[hình 26. Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo PhoneService 35](#_Toc173106820)

[hình 27. Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo StreamingMovies và StreamingTV 36](#_Toc173106821)

[hình 28. Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo PaymentMethod 36](#_Toc173106822)

[hình 29. Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo Tenure 37](#_Toc173106823)

[hình 30. Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo Contract 38](#_Toc173106824)

[hình 31. Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo OnlineSecurity và TechSupport 38](#_Toc173106825)

[hình 32.Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo InternetService 39](#_Toc173106826)

[hình 33. Framework cho dự án 40](#_Toc173106827)

[hình 34. Tiền xử lý dữ liệu 40](#_Toc173106828)

[hình 35.Thông tin kiểm tra của các cột trong dữ liệu 41](#_Toc173106829)

[hình 36 Kiểm tra dữ liệu có thiếu giá trị hay không 42](#_Toc173106830)

[hình 37.Thay thế các giá trị và được kết quả như hình 42](#_Toc173106831)

[hình 38.Kiểm tra giá trị có bị bất thường hay không 43](#_Toc173106832)

[hình 39.Biểu đồ boxplot của Tenure 43](#_Toc173106833)

[hình 40.Biểu đồ boxplot của MonthlyCharges 43](#_Toc173106834)

[hình 41. Biểu đồ boxplot của TotalCharges 44](#_Toc173106835)

[hình 42.Biểu đồ tương quan của num\_features 45](#_Toc173106836)

[hình 43.Biểu đồ tương quan của ob\_features 45](#_Toc173106837)

[hình 44.Thông tin giá trị chuyển đổi 47](#_Toc173106838)

[hình 45.Biểu đồ Learning Curve của mô hình Random Forest 48](#_Toc173106839)

[hình 46.Cây hoàn chỉnh trong Random forest 49](#_Toc173106840)

[hình 47.Cây không đầy đủ của mô hình Random forest 49](#_Toc173106841)

[hình 48.Kết quả đánh giá của mô hình Random Forest 50](#_Toc173106842)

[hình 49. Biểu đồ ROC của mô hình Random Forest 51](#_Toc173106843)

[hình 50.Kết quả đánh giá của mô hình Logistic 53](#_Toc173106844)

[hình 51.Biểu đồ ROC của mô hình hồi quy Logistic 54](#_Toc173106845)

[hình 52.Kết quả đánh giá của mô hình KNN 56](#_Toc173106846)

[hình 53.Biểu đồ ROC của mô hình KNN 57](#_Toc173106847)

[hình 54. Kết quả đánh giá của mô hình Naive Bayes 59](#_Toc173106848)

[hình 55 Biểu đồ ROC của mô hình Naive Bayes 60](#_Toc173106849)

[hình 56. Kết quả đánh giá của mô hình XGboost 62](#_Toc173106850)

[hình 57.Biểu đồ ROC của mô hình XGboost 63](#_Toc173106851)

[hình 58.Các lớp trong mô hình FNN 64](#_Toc173106852)

[hình 59. Kết quả huấn luyện mô hình FNN với 10 epoch cuối 65](#_Toc173106853)

[hình 60. Kết quả đánh giá của mô hình FNN 66](#_Toc173106854)

[hình 61.Biểu đồ ROC của mô hình FNN 66](#_Toc173106855)

# DANH MỤC BẢNG BIỂU

[Bảng 1. Bảng dữ liệu 28](#_Toc173106856)

[Bảng 2.Bảng so sánh các chỉ sổ các các mô hình 67](#_Toc173106857)

# DANH MỤC CÁC TỪ VIẾT TẮT

**AI** : Artificial Intelligence (Trí tuệ nhân tạo)

**KNN :** K-Nearest Neighbors (K láng giềngần nhất)

**FNN :** Feedforward Neural Network (Mạng thần kinh truyền thẳng)

# LỜI MỞ ĐẦU

1. **Mục tiêu của đề tài**

Việc giữ chân khách hàng cá nhân hóa là điều khó khăn vì hầu hết các công ty đều có số lượng lớn khách hàng và không thể dành nhiều thời gian cho từng khách hàng đó. Chi phí sẽ quá lớn, vượt quá doanh thu bổ sung. Tuy nhiên, nếu một công ty có thể dự đoán trước những khách hàng nào có khả năng rời đi thì công ty có thể tập trung nỗ lực giữ chân khách hàng vào những khách hàng có “rủi ro cao” này.

Vì vậy mục tiêu của nghiên cứu này nhằm mục đích tìm ra được phương pháp tốt nhất để có thể dự báo được khả năng rời bỏ dịch vụ của khách hàng bằng cách:

* Xác định các yếu tố tiềm ẩn dẫn đến việc khách hàng hủy dịch vụ, bao gồm yếu tố như hợp đồng, thời gian sử dụng, phương thức thanh toán, bảo vệ thiết bị, v.v
* Phát triển mô hình dự đoán hành vi rời bỏ dịch vụ: Sử dụng các kỹ thuật học máy và phân tích dữ liệu tiên tiến để xây dựng mô hình dự đoán chính xác khả năng khách hàng rời bỏ dịch vụ trong tương lai.
* Đề xuất chiến lược giữ chân khách hàng tiềm năng: Dựa trên kết quả phân tích và dự đoán, đề xuất các chiến lược giữ chân khách hàng hiệu quả, tập trung vào nhóm khách hàng có nguy cơ cao rời bỏ dịch vụ.

1. **Đối tượng nghiên cứu và phạm vi nghiên cứu**

Đối tượng nghiên cứu: Khách hàng của một công ty viễn thông với bộ dữ liệu Customer\_Churn.CSV.

Phạm vi nghiên cứu: Các yếu tố ảnh hưởng đến quyết định tiếp tục hoặc dừng sử dụng dịch vụ của khách hàng, bao gồm cả khách hàng đang sử dụng và đã dừng sử dụng.

1. **Kết cấu của đề tài**

Đề tài được tổ chức gồm phần mở đầu, 3 chương nội dung và phần kết luận

* Mở đầu
* **Chương 1**: Tổng quan về trung tâm VNPT – IT Khu vực 3 & công ty cổ phần Softect & Tổng quan đề tài dự đoán khách hàng rời bỏ dịch vụ viễn thông.
* **Chương 2**: Cơ sở lý thuyết.
* **Chương 3**: Mô tả và trực quan hóa dữ liệu.
* **Chương 4**: Xây dựng mô hình dự đoán khách hàng rời bỏ dịch vụ viễn thông.
* Kết luận và hướng phát triển

1. **TỔNG QUAN VỀ TRUNG TÂM VNPT-IT KHU VỰC 3 & CÔNG TY CỔ PHẦN SOFTECH &TỔNG QUAN ĐỀ TÀI DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ DỊCH VỤ VIỄN THÔNG**

## Giới thiệu tổng quan về doanh nghiệp thực tập VNPT-IT khu vực 3

### Quá trình hình thành và phát triển của công ty

**Trung tâm Công nghệ thông tin VNPT (Tên viết tắt: VNPT-IT**) được thành lập theo Quyết định số 39/QĐ-VNPT-HĐTV-NL ngày 01 tháng 03 năm 2018 của Chủ tịch Tập đoàn Bưu chính Viễn thông Việt Nam, trên cơ sở tổ chức lại các nhiệm vụ và nguồn lực công nghệ thông tin thuộc Tập đoàn [1]

VNPT-IT hoạt động trong lĩnh vực nghiên cứu phát triển, tích hợp các sản phẩm dịch vụ công nghệ thông tin phục vụ nội bộ Tập đoàn Bưu chính Viễn thông Việt Nam (VNPT) và các khách hàng bên ngoài VNPT (bao gồm cả khách hàng Quốc tế).

VNPT-IT đặt mục tiêu xây dựng một hệ sinh thái tích hợp trọn gói các sản phẩm, dịch vụ công nghệ thông tin và Internet lớn nhất Việt Nam, từ đó mang sản phẩm - dịch vụ của chúng tôi đến với thị trường quốc tế. Để làm được điều này, VNPT-IT đề ra chiến lược phát triển xoay quanh 4 giá trị cốt lõi:

* + - * Con người là chìa khóa
      * Khách hàng là trung tâm
      * Sáng tạo không ngừng
      * Đối tác đáng tin cậy



hình 1.Logo Trung tâm VNPT-IT

### Tầm nhìn, sứ mệnh

Với khát vọng tiên phong trong chiến lược đầu tư, cung cấp giải pháp, sản phẩm, dịch vụ và phát triển bền vững trong lĩnh vực công nghệ thông tin, Công ty công nghệ hàng đầu trong lĩnh vực công nghệ thông tin của Việt Nam (VNPT-IT) sẽ phấn đấu phát triển theo định hướng nghiên cứu, gia công, thử nghiệm để cung cấp các sản phẩm đa dạng, đa ngành nghề có uy tín đạt chất lượng cao nhằm đáp ứng nhu cầu xã hội góp phần nâng cao chất lượng cuộc sống của người Việt và nâng tầm vị thế của người Việt trên thị trường Quốc tế.[2]

Chúng tôi mang trên mình sứ mệnh cũng như trọng trách cung cấp các giải pháp, sản phẩm, dịch vụ, sản phẩm thông minh cho xã hội tại tất cả các lãnh vực như đào tạo, quản lý, chăm sóc sức khỏe, ứng dụng và chuyển giao công nghệ có chất lượng cao với thương hiệu và danh tiếng đạt đẳng cấp khu vực và Quốc tế, góp phần vào sự nghiệp công nghiệp hóa, hiện đại hóa đất nước trong bối cảnh hội nhập kinh tế Thế giới.

“Vì một xã hội văn minh, minh bạch góp phần xây dựng cuộc sống người Việt tốt đẹp hơn”

* + 1. **Lĩnh vực hoạt động**

-Lĩnh vực kinh doanh:

+ Tổ chức nghiên cứu, phát triển, sản xuất, kinh doanh các sản phẩm, dịch vụ công nghệ thông tin để cung cấp cho nội bộ VNPT và khách hàng bên ngoài VNPT;

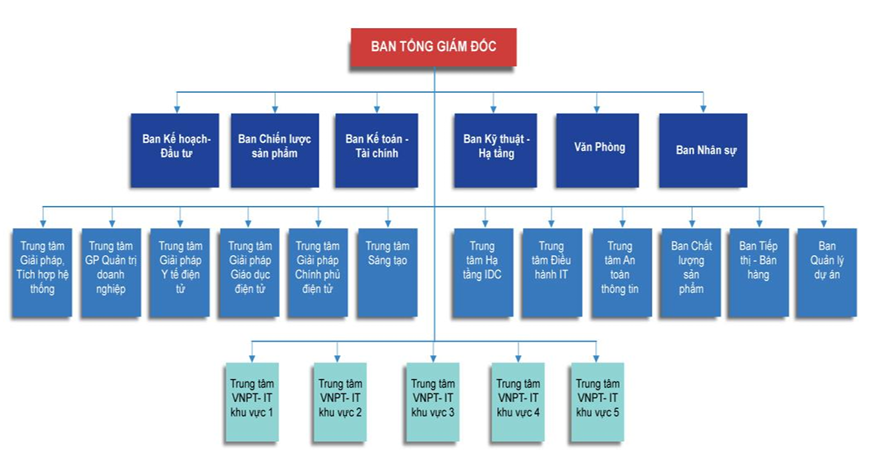
+ Đầu tư, phát triển, quản lý tài sản các hệ thống, nền tảng công nghệ thông tin; vận hành khai thác hệ thống điều hành sản xuất kinh doanh, đảm bảo an toàn, bảo mật thông tin cho các sản phẩm và các dịch vụ công nghệ thông tin của VNPT cung cấp cho khách hàng

+ Trong mô hình của Công ty VNPT-IT gồm: Ban Tổng giám đốc, Văn phòng các Ban chức năng cùng các Trung tâm trực thuộc và các Trung tâm tại: Hà Nội, Hồ Chí Minh, Đà Nẵng, Hải Phòng và Tiền Giang. Các Trung tâm này là đơn vị hạch toán phụ thuộc của Công ty.

+ Là một trong những Đơn vị chủ chốt của Tập đoàn Bưu chính Viễn thông Việt Nam, VNPT-IT luôn phấn đấu không ngừng nâng cao chất lượng sản phẩm dịch vụ về mọi mặt để trở thành thương hiệu có uy tín trong lĩnh vực công nghệ thông tin, góp phần đưa VNPT đạt mục tiêu trở thành Tập đoàn Viễn thông - CNTT hàng đầu quốc gia, giữ vai trò chủ đạo trong lĩnh vực Viễn thông và CNTT Việt Nam.

- Sản phẩm - Dịch vụ:

* Chính phủ điện tử
* Giải pháp Doanh nghiệp
* Giải pháp Giáo dục
* Giải pháp Y tế
* An toàn bảo mật
  + 1. **Cơ cấu tổ chức**



hình 2. Sơ đồ cơ cấu tổ chức của công ty VNPT

* + Ban Tổng giám đốc: Điều hành toàn bộ công ty
  + Ban Kế hoạch - Đầu tư:
  + Ban Chiến lược sản phẩm:
  + Ban Kế toán - Tài chính:
  + Ban Kỹ thuật - Hạ tầng:
  + Văn phòng:
  + Ban Nhân sự:

## Tổng quan về công ty cổ phần

### Quá trình hình thành và phát triển công ty

Công ty cổ phần Softech, tiền thân là Trung tâm Công nghệ Phần mềm Đà Nẵng, được thành lập ngày 08/11/2000. Trải qua hơn 15 năm hoạt động, Công ty cổ phần Softech đã trở thành đơn vị tiên phong trong lĩnh vực sản xuất, gia công phần mềm, đào tạo và phát triển nguồn nhân lực, cung cấp các giải pháp và dịch vụ công nghệ thông tin ở khu vực miền Trung và cả nước. Với lợi thế về địa lý như nằm trong vùng kinh tế trọng điểm miền Trung - cửa ngõ quan trọng của Đông Nam Á nối các nước Lào, Campuchia, Thái Lan và Myanma với các nước Đông Bắc Á thông qua hành lang kinh tế Đông Tây, Softech thực sự có nhiều thuận lợi trong việc phát triển và hợp tác với các nhà đầu tư trong và ngoài nước [3]



hình 3. Logo của CTCP Softech

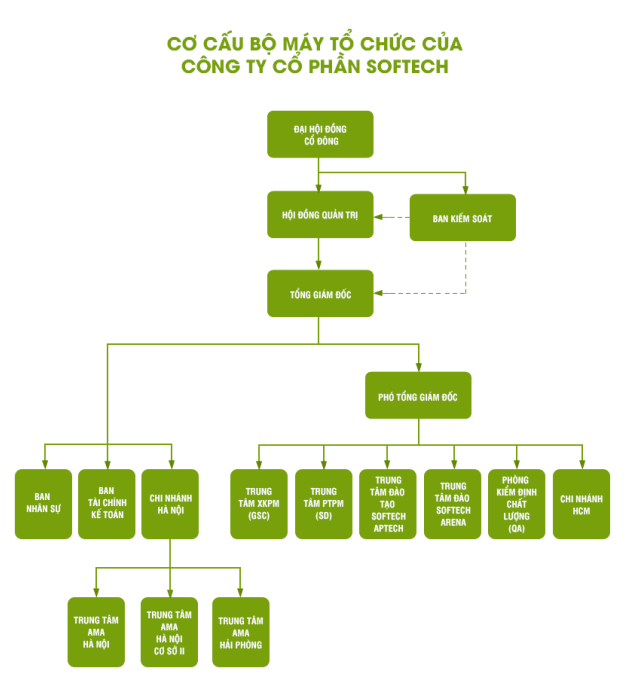
### Tầm nhìn sự mệnh

* Xây dựng Softech thành một trong những công ty phần mềm hàng đầu tại Việt Nam
* Đảm bảo đời sống của người lao động, xây dựng môi trường làm việc thân thiện và phát huy tối đa năng lực của từng nhân viên.
* Tối đa hóa lợi nhuận trên cơ sở sử dụng một cách hợp lý nguồn tài nguyên của công ty nhằm đem lại lợi ích tối đa cho các nhà đầu tư.
* Softech là hạt nhân thúc đẩy sự phát triển nền công nghiệp phần mềm của Đà Nẵng và khu vực miền Trung.

### Sứ mệnh hoạt động

* Đào tạo và phát triển nguồn nhân lực CNTT.
* Tổ chức hợp tác sản xuất và gia công sản phẩm phần mềm.
* Thực hiện các liên kết trong nước và quốc tế trong lĩnh vực CNTT.
* Thực hiện tích hợp các hệ thống về CNTT.

### Cơ cấu tổ chức [4]



hình 4. Cơ cấu tổ chức của công ty Softech

## Tổng quan về đề tài dự đoán khách hàng rời bỏ dịch vụ

Trong ngành viễn thông, khách hàng rời bỏ dịch vụ (churn) là một thách thức lớn. Khách hàng rời bỏ không chỉ làm giảm doanh thu mà còn ảnh hưởng đến hình ảnh và uy tín của công ty. Để giảm thiểu tình trạng này, các công ty viễn thông ngày càng chú trọng vào việc phân tích và dự đoán khách hàng rời bỏ. Bằng cách xác định sớm, công ty có thể triển khai các chiến lược giữ chân khách hàng, như cung cấp các gói dịch vụ ưu đãi hoặc cải thiện chất lượng dịch vụ. Điều này giúp giảm thiểu tỷ lệ rời bỏ và duy trì lượng khách hàng ổn định.

Việc nhận diện chính xác nhóm khách hàng có nguy cơ rời bỏ cho phép các công ty tập trung nguồn lực và chi phí vào những khách hàng này. Thay vì chi tiêu rộng rãi và không hiệu quả.

AI và học máy là những công cụ mạnh mẽ giúp tối ưu hóa quá trình phân tích và dự đoán. Các mô hình học máy có thể học từ dữ liệu lịch sử để nhận diện các yếu tố ảnh hưởng đến việc khách hàng rời bỏ. Chúng có thể dự đoán chính xác nhóm khách hàng có nguy cơ cao và đề xuất các biện pháp giữ chân khách hàng một cách hiệu quả tạo nên lợi thế cạnh tranh bền vững cho các công ty viễn thông.

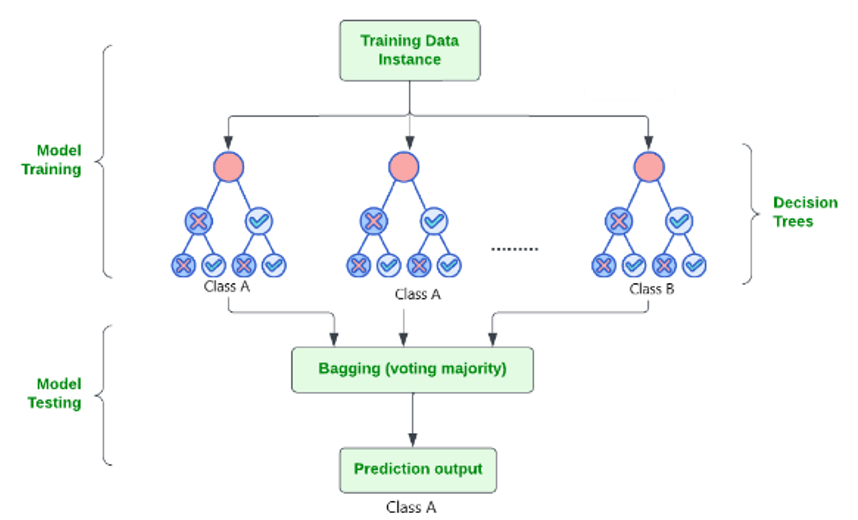
1. **CƠ SỞ LÝ THUYẾT**

## Lý thuyết các mô hình dự đoán

* + 1. **Mô hình Rừng ngẫu nhiên (Random Forest)**

1. **Khái niệm mô hình**

Random Forest là một thuật toán học máy ensemble sử dụng nhiều cây quyết định (decision tree) để thực hiện các nhiệm vụ phân loại và hồi quy. Thuật toán này hoạt động bằng cách xây dựng một tập hợp lớn các cây quyết định được huấn luyện trên các tập dữ liệu con ngẫu nhiên được lấy từ tập dữ liệu gốc. Mỗi cây quyết định được huấn luyện để phân loại hoặc dự đoán một giá trị cho một điểm dữ liệu mới, sau đó kết quả dự đoán của tất cả các cây quyết định được kết hợp để đưa ra dự đoán cuối cùng. [5]



hình 5.Mô hình Random Forest

1. **Thuật toán hoạt động của Random Forest**

Bước 1: Chọn ngẫu nhiên K mẫu dữ liệu từ tập dữ liệu huấn luyện.

Bước 2: Xây dựng các cây quyết định tương ứng với các mẫu dữ liệu được chọn (các tập con).

Bước 3: Quyết định số lượng N cây quyết định bạn muốn xây dựng.

Bước 4: Lặp lại Bước 1 và 2 N lần.

Bước 5: Đối với các mẫu dữ liệu mới, tìm dự đoán của từng cây quyết định và gán các mẫu dữ liệu mới vào nhóm nhận được nhiều phiếu bầu (dự đoán) nhất.

Lưu ý:

* K là một số được chọn ngẫu nhiên mà không cần đặt trước, thường nhỏ hơn tổng số mẫu dữ liệu trong tập huấn luyện.
* Bước lặp lại (Bước 4) giúp tạo ra sự đa dạng giữa các cây quyết định, cải thiện độ chính xác của mô hình Random Forest.

1. **Ưu và nhược điểm của mô hình**

**+ Ưu điểm:**

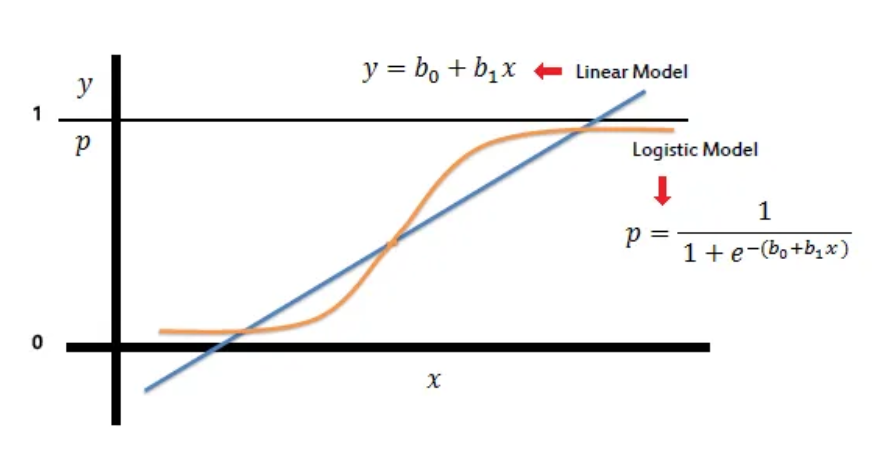
* Random Forest đạt được độ chính xác và tính mạnh mẽ nhờ việc kết hợp nhiều cây quyết định. Các dự đoán trung bình từ các cây này giúp loại bỏ các sai lệch của từng cây riêng lẻ, dẫn đến kết quả tổng thể chính xác hơn.
* Random Forest có khả năng chống overfitting do tính chất trung bình các dự đoán từ nhiều cây quyết định khác nhau.
* Random Forest có thể xử lý tốt dữ liệu thiếu bằng cách tận dụng các thuộc tính có sẵn để dự đoán.
* Đánh giá tầm quan trọng của biến: Điều này hỗ trợ trong việc lựa chọn thuộc tính và diễn giải vai trò của các yếu tố ảnh hưởng tới dự đoán.

**+ Nhược điểm:**

* Tốc độ tính toán chậm hơn so với thuật toán sử dụng một mô hình duy nhất.
* Khó lý giải cách thức đưa ra dự đoán, hạn chế hiểu biết về mô hình và ảnh hưởng đến độ tin tưởng kết quả.
* Lựa chọn tham số không phù hợp ảnh hưởng đến hiệu suất mô hình.
  + 1. **Mô hình Hồi quy Logistic (Logistic Regression)**

1. **Khái niệm mô hình**

Hồi quy Logistic là một mô hình thống kê được sử dụng để phân loại nhị phân, tức dự đoán một đối tượng thuộc vào một trong hai nhóm. Hồi quy Logistic làm việc dựa trên nguyên tắc của hàm sigmoid – một hàm phi tuyến tự chuyển đầu vào của nó thành xác suất thuộc về một trong hai lớp nhị phân. [6]



hình 6. Mô hình Hồi quy Logistic

1. **Hàm Sigmoid**

Hồi quy Logistic hoạt động dựa trên hàm Sigmoid, được biểu diễn như sau:

S(z) = )

Hàm Sigmoid nhận đầu vào là một giá trị z bất kỳ, và trả về đầu ra là một giá trị xác suất nằm trong khoảng [0,1].

1. **Ưu nhược điểm của mô hình**

**+ Ưu điểm:**

* Xử lý được dữ liệu phân loại nhị phân
* Dễ hiểu và giải thích
* Linh hoạt, có thể sử dụng nhiều biến độc lập

**+ Nhược điểm:**

* Giả định về tuyến tính giữa biến độc lập và hàm liên kết
* Nhạy cảm với đa cộng tuyến
* Yêu cầu kích thước mẫu tương đối lớn
* Không xử lý tốt các giá trị ngoại lệ
* Không thể xử lý biến phụ thuộc đa lớp
  + 1. **Mô hình KNN (K-Nearest Neighbors)**

1. **Khái niệm mô hình**

Thuật toán K-Nearest Neighbors (KNN) là một thuật toán học máy có giám sát phổ biến được sử dụng cho các tác vụ phân loại và hồi quy. [7]

Thuật toán KNN dựa trên nguyên tắc các đối tượng hoặc điểm dữ liệu gần nhau trong không gian đặc trưng có khả năng thuộc cùng một lớp hoặc có kết quả đầu ra tương tự nhau. Thuật toán KNN gán một điểm dữ liệu mới cho lớp phổ biến nhất trong k hàng xóm gần nhất của nó. Giá trị của k là một siêu tham số cần được chỉ định trước khi đào tạo mô hình.



hình 7.Mô hình KNN

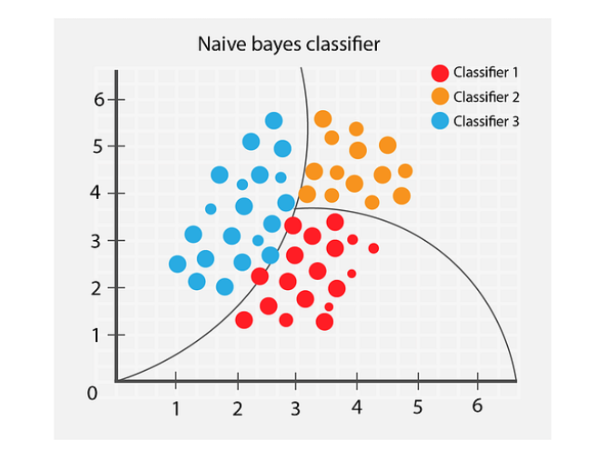
1. **Các bước thực hiện**
2. Tính toán khoảng cách giữa điểm dữ liệu mới và tất cả các điểm dữ liệu đào tạo dựa trên số liệu khoảng cách như khoảng cách Euclide hoặc khoảng cách Manhattan.
3. Chọn k hàng xóm gần nhất dựa trên khoảng cách nhỏ nhất.
4. Gán điểm dữ liệu mới cho lớp phổ biến nhất trong số k hàng xóm gần nhất của nó cho các nhiệm vụ phân loại hoặc dự đoán đầu ra dưới dạng giá trị trung bình của k hàng xóm gần nhất cho các nhiệm vụ hồi quy.
5. Đánh giá hiệu suất của mô hình bằng cách sử dụng một số liệu phù hợp như độ chính xác hoặc lỗi bình phương trung bình.
6. Điều chỉnh các siêu tham số của mô hình, chẳng hạn như giá trị của k bằng cách sử dụng bộ xác thực hoặc xác thực chéo.
7. **Ưu nhược điểm của mô hình**

**+ Ưu điểm:**

* Đơn giản và dễ giải thích
* Không dựa trên bất kỳ giả định nào, vì thế nó có thể được sử dụng trong các bài toán phi tuyến tính.
* Hoạt động tốt trong trường hợp phân loại với nhiều lớp
* Sử dụng được trong cả phân loại và hồi quy

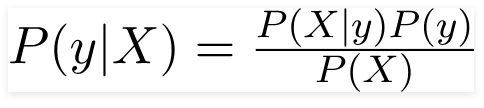
**+ Nhược điểm:**

* Trở nên rất chậm khi số lượng điểm dữ liệu tăng lên vì mô hình cần lưu trữ tất cả các điểm dữ liệu.
* Tốn bộ nhớ.
* Nhạy cảm với các dữ liệu bất thường (nhiễu).
  + 1. **Mô hình Naive Bayes**
  1. **Khái niệm mô hình** [8]

****

hình 8. Mô hình Naive bayes

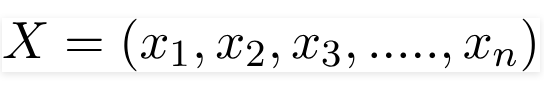
Naive Bayes là một thuật toán phân lớp được mô hình hoá dựa trên định lý Bayes trong xác suất thống kê:



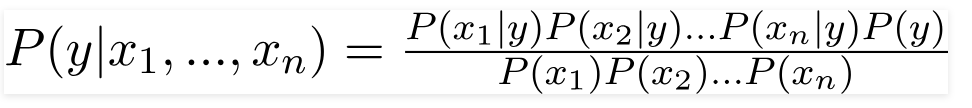
trong đó:

* P(y|X) gọi là posterior probability: xác suất của mục tiêu y với điều kiện có đặc trưng X
* P(X|y) gọi là likelihood: xác suất của đặc trưng X khi đã biết mục tiêu y
* P(y) gọi là prior probability của mục tiêu y
* P(X) gọi là prior probability của đặc trưng X

Ở đây, **X** là vector các đặc trưng, có thể viết dưới dạng:



Khi đó, đẳng thức Bayes trở thành:



Trong mô hình Naive Bayes, có hai giả thiết được đặt ra:

1. Các đặc trưng đưa vào mô hình là độc lập với nhau. Tức là sự thay đổi giá trị của một đặc trưng không ảnh hưởng đến các đặc trưng còn lại.
2. Các đặc trưng đưa vào mô hình có ảnh hưởng ngang nhau đối với đầu ra mục tiêu.

Khi đó, kết quả mục tiêu y để P(y|X) đạt cực đại trở thành:



Chính vì hai giả thiết gần như không tồn tại trong thực tế trên, mô hình này mới được gọi là naive (ngây thơ). Tuy nhiên, chính sự đơn giản của nó với việc dự đoán rất nhanh kết quả đầu ra khiến nó được sử dụng rất nhiều trong thực tế trên những bộ dữ liệu lớn, đem lại kết quả khả quan.

* 1. **Ưu nhược điểm mô hình**

**+ Ưu điểm:**

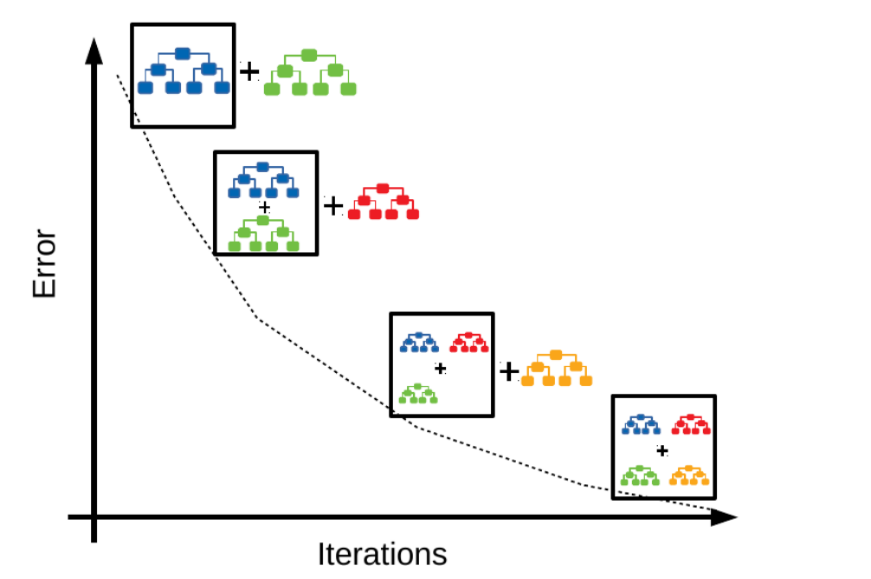
* Giả định độc lập: hoạt động tốt cho nhiều bài toán/miền dữ liệu và ứng dụng.
* Đơn giản nhưng đủ tốt để giải quyết nhiều bài toán như phân lớp văn bản, lọc spam,..
* Cho phép kết hợp tri thức tiên nghiệm (priori knowledge) và dữ liệu quan sát được (observed data).
* Tốt khi có sự chênh lệch số lượng giữa các lớp phân loại.
* Huấn luyện mô hình (ước lượng tham số) dễ và nhanh.

**+ Nhược điểm:**

* Giả định độc lập (ưu điểm cũng chính là nhược điểm) hầu hết các trường hợp thực tế trong đó có các thuộc tính trong các đối tượng thường phụ thuộc lẫn nhau.
* Vấn đề zero
* Mô hình không được huấn luyện bằng phương pháp tối ưu mạnh và chặt chẽ.
* Tham số của mô hình là các ước lượng xác suất điều kiện đơn lẻ.
* Không tính đến sự tương tác giữa các ước lượng này.
  + 1. **Mô hình XGBoost**

**(a) Khái niệm mô hình**

XGBoost hoạt động dựa trên việc kết hợp nhiều mô hình cây quyết định (decision tree) yếu thành một mô hình mạnh hơn. Mỗi mô hình cây được xây dựng dựa trên lỗi dự đoán của các mô hình trước đó, giúp cải thiện dần dần độ chính xác của mô hình tổng thể. [9]



hình 9.Mô hình XGboost

**(b) Quy trình hoạt động**

1. Khởi tạo: Bắt đầu với một mô hình cây quyết định đơn giản.
2. Lặp:

* Huấn luyện: Huấn luyện một mô hình cây quyết định mới để sửa chữa lỗi dự đoán của mô hình hiện tại.
* Cập nhật: Cập nhật trọng số của các mô hình cây quyết định để mô hình mới có ảnh hưởng lớn hơn.

1. Dự đoán: Dự đoán cho một ví dụ mới bằng cách kết hợp dự đoán của tất cả các mô hình cây quyết định, với trọng số tương ứng.

**(c) Ưu nhược điểm của mô hình**

**+ Ưu điểm:**

* Độ chính xác cao: XGBoost thường đạt được độ chính xác cao hơn so với các thuật toán học máy khác.
* Hiệu quả tính toán: XGBoost có thể được tối ưu hóa để chạy hiệu quả trên cả CPU và GPU.
* Khả năng mở rộng: XGBoost có thể xử lý được tập dữ liệu lớn và phức tạp.
* Dễ sử dụng: XGBoost có nhiều thư viện và API dễ sử dụng cho nhiều ngôn ngữ lập trình khác nhau.

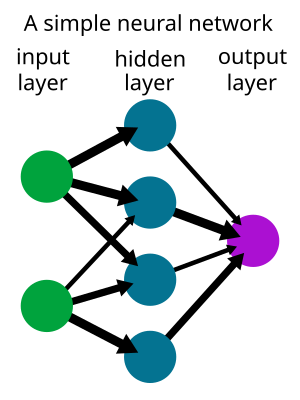
**+ Nhược điểm:**

* Có thể bị quá khớp: XGBoost có thể học quá tốt dữ liệu huấn luyện dẫn đến hiệu suất kém trên dữ liệu mới.
* Khó giải thích: Mô hình XGBoost có thể khó giải thích do cấu trúc phức tạp của nó.
  + 1. **Mô hình Feedforward Neural Network (FNN)**

**(a) Khái niệm mô hình**

Mạng nơ-ron là một công cụ khai thác dữ liệu và xử lý hình ảnh rất phổ biến. Nguồn gốc của chúng bắt nguồn từ nỗ lực mô hình hóa quá trình suy nghĩ của con người như một thuật toán có thể chạy hiệu quả trên máy tính.

**Khái niệm:** Mạng thần kinh truyền thẳng (Feedforward Neural Network, viết tắt là FNN) là một loại mạng thần kinh nhân tạo đặc biệt, trong đó các kết nối giữa các nút không tạo thành chu trình. Điều này làm cho FNN khác biệt so với các mạng thần kinh hồi quy. Là loại mạng thần kinh nhân tạo đầu tiên và đơn giản nhất được phát minh, mạng thần kinh truyền thẳng cho phép thông tin di chuyển theo một hướng duy nhất — từ các nút đầu vào, qua các nút ẩn (nếu có), và đến các nút đầu ra, mà không có bất kỳ chu trình hoặc vòng lặp nào trong mạng. [10]



hình 10. Mô hình FNN

Cấu trúc cơ bản của FNN

* **Lớp đầu vào (Input Layer)**: Nhận dữ liệu đầu vào từ bên ngoài.
* **Lớp ẩn (Hidden Layers)**: Xử lý và trích xuất các đặc trưng từ dữ liệu. Có thể có nhiều lớp ẩn trong một mạng nơ-ron.
* **Lớp đầu ra (Output Layer)**: Tạo ra đầu ra cuối cùng của mạng, có thể là một hoặc nhiều giá trị phụ thuộc vào bài toán cụ thể.

**(b) Cách thức hoạt động**

1. Truyền tiếp: Dữ liệu đầu vào được truyền qua các lớp ẩn, tại mỗi lớp, các nút thực hiện phép nhân các trọng số với giá trị đầu vào và cộng với bias, sau đó kết quả được đưa qua hàm kích hoạt để tạo ra đầu ra của nút đó.
2. Tính toán lỗi: So sánh kết quả đầu ra của mạng với giá trị thực tế để tính toán lỗi.
3. Cập nhật trọng số: Sử dụng thuật toán học máy như gradient descent để cập nhật các trọng số và bias của mạng nhằm giảm thiểu lỗi.

**(c) Các bước thực hiện**

**Bước 1: Tại lớp input** nhận dữ liệu đầu vào (x) và chuyển nó đến lớp ẩn đầu tiên.

**Bước 2: Tại lớp ẩn.**

* **Tính toán tổng trọng số**: z=W\*x+b
* W là ma trận trọng số.
* x là vector đầu vào.
* B là vector độ chệch (bias).
* **Áp dụng hàm kích hoạt** **để tính toán nơ-ron**

Một số hàm kịch hoạt phổ biến nhất được sử dụng để hình thành lớp ẩn là:

1. Hàm Kích Hoạt Sigmoid

* **Định nghĩa:** Hàm Sigmoid chuyển đổi đầu vào của nó thành một giá trị trong khoảng từ 0 đến 1, giúp xác định xác suất cho các lớp nhị phân trong các bài toán phân loại.
* **Phương trình**: 
* **Phạm vi giá trị**: (0, 1)
* **Ưu điểm**:
  + Đầu ra nằm trong khoảng (0, 1), phù hợp cho các bài toán phân loại nhị phân.
* **Nhược điểm**:
  + Vấn đề gradient biến mất (vanishing gradient) khi giá trị đầu vào quá lớn hoặc quá nhỏ.
  + Gradient rất nhỏ khi giá trị đầu vào ở xa khỏi 0, làm chậm quá trình học.

2.Hàm Kích Hoạt ReLU (Rectified Linear Unit)

* **Định nghĩa:** Hàm Kích Hoạt ReLU có giá trị đầu ra bằng giá trị của chính nó nếu giá trị đó lớn hơn 0 và bằng 0 nếu giá trị đó nhỏ hơn hoặc bằng 0.
* **Phương trình**: f(x)=max(0,x)
* **Phạm vi giá trị**: [0, ∞)
* **Ưu điểm**:
  + Đơn giản, tính toán nhanh.
  + Giảm vấn đề gradient biến mất.
  + Hiệu quả trong các mô hình sâu.
* **Nhược điểm**:
  + Vấn đề "chết ReLU" (dead ReLU), khi nơ-ron không kích hoạt cho mọi giá trị đầu vào âm, gradient sẽ bằng 0 và không học được gì.

**Bước 3: Lớp đầu ra t**ương tự lớp ẩn nhưng đầu ra cuối cùng là giá trị dự đoán.

**(d) Ưu nhược điểm của mô hình**

**+ Ưu điểm:**

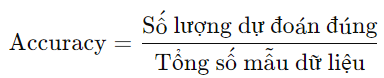
* Đơn giản: Dễ hiểu và triển khai.
* Hiệu quả: Có thể giải quyết nhiều bài toán phức tạp.
* Linh hoạt: Có thể được sử dụng cho cả bài toán hồi quy và phân loại.

**+ Nhược điểm:**

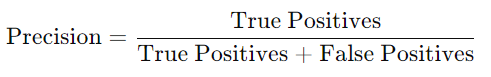
* Khó học các mối quan hệ phức tạp: Đối với các bài toán có dữ liệu có cấu trúc tuần tự hoặc phụ thuộc vào thời gian, FNN có thể gặp khó khăn.
* Yêu cầu lượng dữ liệu lớn: Để đạt được hiệu quả cao, FNN cần một lượng lớn dữ liệu để huấn luyện.

## Các chỉ số đánh giá mô hình [11]

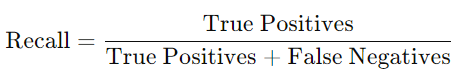
 **Accuracy (Độ chính xác)**: Tỷ lệ dự đoán đúng trên tổng số mẫu dữ liệu. Công thức:



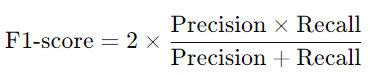
 **Precision (Độ chính xác của lớp dương)**: Tỷ lệ khách hàng được dự đoán là sẽ rời bỏ dịch vụ thực sự rời bỏ dịch vụ. Công thức:

​

 **Recall (Độ phủ của lớp dương ):** Tỷ lệ khách hàng thực sự rời bỏ dịch vụ được dự đoán là sẽ rời bỏ dịch vụ. Công thức:



 **F1-score**: Trung bình điều hòa giữa Precision và Recall, là một chỉ số tổng hợp giữa hai chỉ số trên. Công thức:



 **AUC (Area Under the ROC Curve)**: Diện tích dưới đường cong ROC. Một chỉ số tổng hợp cho biết mức độ mà mô hình phân loại phân biệt được giữa các lớp.

## Các công cụ hỗ trợ cho việc phân tích

* + 1. **Jupyter Notebook**
       - 1. **Sơ lược về công cụ**

Jupyter Notebook là một môi trường tính toán tương tác mã nguồn mở, cho phép người dùng tạo tài liệu kết hợp văn bản, mã, công thức và trực quan hóa. Nó được sử dụng rộng rãi trong khoa học dữ liệu, học máy, phân tích dữ liệu và giáo dục. Jupyter Notebook được biết đến với khả năng dễ sử dụng, linh hoạt và khả năng tích hợp nhiều ngôn ngữ lập trình khác nhau như Python, R, Julia, v.v. [12]

****

hình 11. Logo công cụ Jupyter

* + - * 1. **Tính năng chính**

Jupyter Notebook là một nền tảng tính toán khoa học mã nguồn mở, bạn có thể sử dụng để:

* Viết mã: Jupyter Notebook hỗ trợ nhiều ngôn ngữ lập trình phổ biến như Python, R, Julia, v.v.
* Chạy mã: Bạn có thể trực tiếp chạy các đoạn mã trong Notebook và xem kết quả ngay lập tức.
* Hiển thị văn bản: Jupyter Notebook cho phép bạn kết hợp mã với văn bản, ghi chú, công thức toán học và hình ảnh để tạo ra các tài liệu khoa học và báo cáo chi tiết.
* Trực quan hóa dữ liệu: Jupyter Notebook cung cấp nhiều thư viện trực quan hóa dữ liệu mạnh mẽ giúp bạn dễ dàng tạo biểu đồ, đồ thị và hình ảnh từ dữ liệu của mình.
* Chia sẻ: Jupyter Notebook cho phép bạn dễ dàng chia sẻ các tài liệu của mình với người khác bằng cách xuất sang các định dạng khác nhau như HTML, PDF, v.v.
  + - * 1. **Ưu điểm và nhược điểm**

**+ Ưu điểm:**

* Giao diện người dùng thân thiện và trực quan.
* Khả năng chạy từng phần của mã và xem kết quả ngay lập tức.
* Hỗ trợ nhiều ngôn ngữ lập trình thông qua các “kernels”.
* Khả năng tích hợp tài liệu, mã nguồn và kết quả trong một tài liệu duy nhất.
* Có một cộng đồng người dùng và nhà phát triển rộng lớn, cung cấp nhiều tài liệu hướng dẫn và hỗ trợ.

**+ Nhược điểm:**

* Khi làm việc với các tệp lớn hoặc mã phức tạp, Jupyter Notebook có thể chậm và không ổn định.
* Quản lý mã nguồn trong các notebook lớn có thể khó khăn, đặc biệt khi không có cấu trúc rõ ràng.
* Để chạy các notebook yêu cầu kết nối internet hoặc các nguồn dữ liệu ngoài, bạn cần thiết lập cấu hình phức tạp.
* Không dễ dàng để tái sử dụng mã từ một notebook trong các dự án khác mà không cần chỉnh sửa.
  + 1. **Power BI**

1. **Sơ lược về công cụ**

Power BI là một công cụphân tích dữ liệu do Microsoft phát triển, giúp người dùng kết nối với nhiều nguồn dữ liệu khác nhau, truy vấn, lọc, chuyển đổi và trực quan hóa dữ liệu để tạo báo cáo và bảng điều khiển thông tin. Power BI cung cấp nhiều tính năng mạnh mẽ, dễ sử dụng, phù hợp cho người dùng ở mọi cấp độ kỹ thuật, từ người mới bắt đầu đến chuyên gia phân tích dữ liệu. [13]



hình 12. Công cụ Power BI

1. **Tính năng chính:**

* Kết nối dữ liệu: Power BI có thể kết nối với nhiều nguồn dữ liệu khác nhau như cơ sở dữ liệu SQL, tệp Excel, dịch vụ đám mây, v.v.
* Truy vấn và lọc dữ liệu: Power BI cung cấp Power Query, một công cụ mạnh mẽ để truy vấn, lọc và chuyển đổi dữ liệu trước khi tải vào Power BI.
* Chuyển đổi dữ liệu: Power BI hỗ trợ nhiều loại chuyển đổi dữ liệu như lọc, sắp xếp, tổng hợp, v.v.
* Trực quan hóa dữ liệu: Power BI cung cấp nhiều loại trực quan hóa dữ liệu khác nhau như biểu đồ, đồ thị, bản đồ, v.v. để hiển thị dữ liệu một cách trực quan.
* Tạo báo cáo: Power BI cho phép bạn tạo báo cáo với nhiều trang, bao gồm các trực quan hóa, bảng và văn bản.
* Chia sẻ báo cáo: Bạn có thể chia sẻ báo cáo Power BI với người khác qua Power BI Service hoặc xuất báo cáo sang các định dạng khác như PDF, PowerPoint.
* Bảo mật: Power BI cung cấp các tính năng bảo mật để bảo vệ dữ liệu và báo cáo của bạn.
* Phân tích dự đoán: Power BI hỗ trợ phân tích dự đoán để dự đoán xu hướng dữ liệu trong tương lai.
* Phân tích What-if: Power BI cho phép bạn thay đổi dữ liệu và xem tác động đến các biện pháp.
* Machine learning: Power BI tích hợp machine learning để tự động hóa các tác vụ phân tích dữ liệu.

1. **Ưu và nhược điểm**

**+ Ưu điểm:**

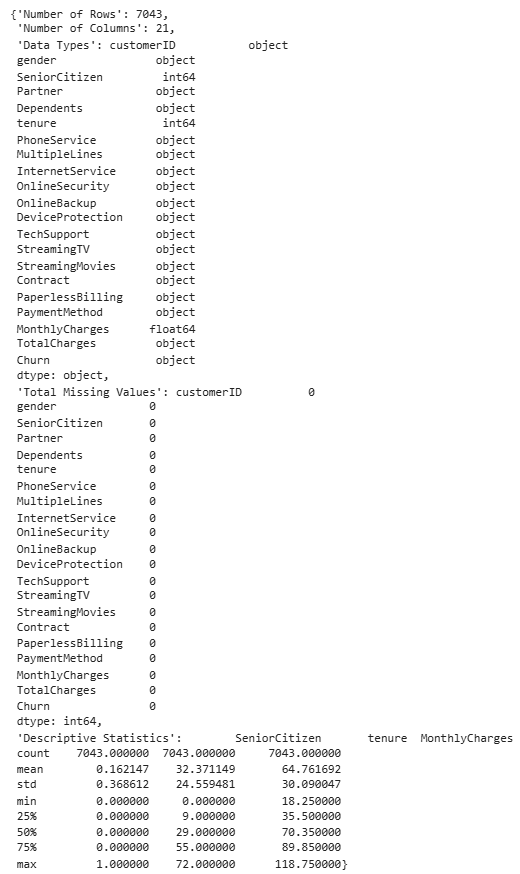
* Dễ sử dụng: Power BI có giao diện trực quan và dễ sử dụng, phù hợp cho người dùng ở mọi cấp độ kỹ thuật.
* Mạnh mẽ: Power BI cung cấp nhiều tính năng mạnh mẽ để phân tích dữ liệu.
* Linh hoạt: Power BI có thể kết nối với nhiều nguồn dữ liệu khác nhau và hỗ trợ nhiều loại trực quan hóa dữ liệu.
* Chia sẻ dễ dàng: Bạn có thể dễ dàng chia sẻ báo cáo Power BI với người khác.
* Bảo mật: Power BI cung cấp các tính năng bảo mật để bảo vệ dữ liệu và báo cáo của bạn.
* Hỗ trợ cộng tác: Power BI hỗ trợ cộng tác theo thời gian thực, cho phép nhiều người cùng làm việc trên một báo cáo.
* Giá cả hợp lý: Power BI có nhiều gói giá cả phù hợp với nhu cầu của bạn.

**+ Nhược điểm:**

* Phiên bản miễn phí có hạn chế: Phiên bản miễn phí của Power BI có một số hạn chế về tính năng và dung lượng lưu trữ.
* Yêu cầu đường truyền internet: Power BI yêu cầu đường truyền internet để hoạt động.
* Có thể gặp lỗi: Như bất kỳ phần mềm nào, Power BI cũng có thể gặp lỗi.

1. **MÔ TẢ VÀ TRỰC QUAN HÓA DỮ LIỆU**

## Thống kê mô tả



hình 13. Thống kê mô tả dữ liệu

Giải Thích các thông số có trong dữ liệu:

* Số lượng hàng: 7043
* Số lượng cột: 21
* Kiểu dữ liệu của từng cột:
* customerID: object (chuỗi)
* gender: object (chuỗi)
* SeniorCitizen: int64 (số)
* Partner: object (chuỗi)
* Dependents: object (chuỗi)
* tenure: int64 (số)
* PhoneService: object (chuỗi)
* MultipleLines: object (chuỗi)
* InternetService: object (chuỗi)
* OnlineSecurity: object (chuỗi)
* OnlineBackup: object (chuỗi)
* DeviceProtection: object (chuỗi)
* TechSupport: object (chuỗi)
* StreamingTV: object (chuỗi)
* StreamingMovies: object (chuỗi)
* Contract: object (chuỗi)
* PaperlessBilling: object (chuỗi)
* Payment Method: object (chuỗi)
* MonthlyCharges: float64 (số thực)
* TotalCharges: object (chuỗi)
* Churn: object (chuỗi)
* Tổng số giá trị thiếu: 0
* Thống kê cơ bản:
* SeniorCitizen:
* Trung bình (mean): 0.162147 cho thấy khoảng 16.2% khách hàng là người cao tuổi.
* Độ lệch chuẩn (std): 0.368612 cho thấy sự biến đổi trong dữ liệu, nhưng vì đây là một biến nhị phân (0 hoặc 1), độ lệch chuẩn này thể hiện mức độ phân tán so với trung bình.
* Giá trị tối thiểu (min) và tối đa (max): Biến nhị phân có giá trị 0 và 1 đại diện cho việc khách hàng có phải là người cao tuổi hay không.
* Các phân vị (25%, 50%, 75%): Tất cả đều bằng 0, nghĩa là phần lớn khách hàng không phải là người cao tuổi
* tenure:
* Trung bình (mean): 32.37 tháng cho thấy thời gian trung bình khách hàng gắn bó là khoảng 2.7 năm.
* Độ lệch chuẩn (std): 24.56, sự phân tán lớn trong thời gian gắn bó, có khách hàng mới chỉ bắt đầu và có khách hàng đã gắn bó lâu dài.
* Giá trị tối thiểu (min) và tối đa (max): Từ 0 đến 72 tháng.
* Các phân vị: 25% khách hàng có thời gian gắn bó dưới 9 tháng, 50% dưới 29 tháng, và 75% dưới 55 tháng.
* MonthlyCharges:
* Trung bình (mean): 64.76 USD cho thấy chi phí trung bình hàng tháng của khách hàng là khoảng 64.76 USD.
* Độ lệch chuẩn (std): 30.09, chi phí hàng tháng có sự biến đổi lớn giữa các khách hàng.
* Giá trị tối thiểu (min) và tối đa (max): Từ 18.25 USD đến 118.75 USD.
* Các phân vị: 25% khách hàng có chi phí hàng tháng dưới 35.50 USD, 50% dưới 70.35 USD, và 75% dưới 89.85 USD.

## Mô tả bài toán

Mục đích: Ứng dụng các mô hình machine learning dự đoán khả năng khách hàng rời bỏ dịch vụ viễn thông dựa trên các yếu tố liên quan được cung cấp trong tập dữ liệu.

Đầu vào: tập dữ liệu đầu vào cho bài toán này bao gồm các thông tin sau đây về khách hàng: customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn.

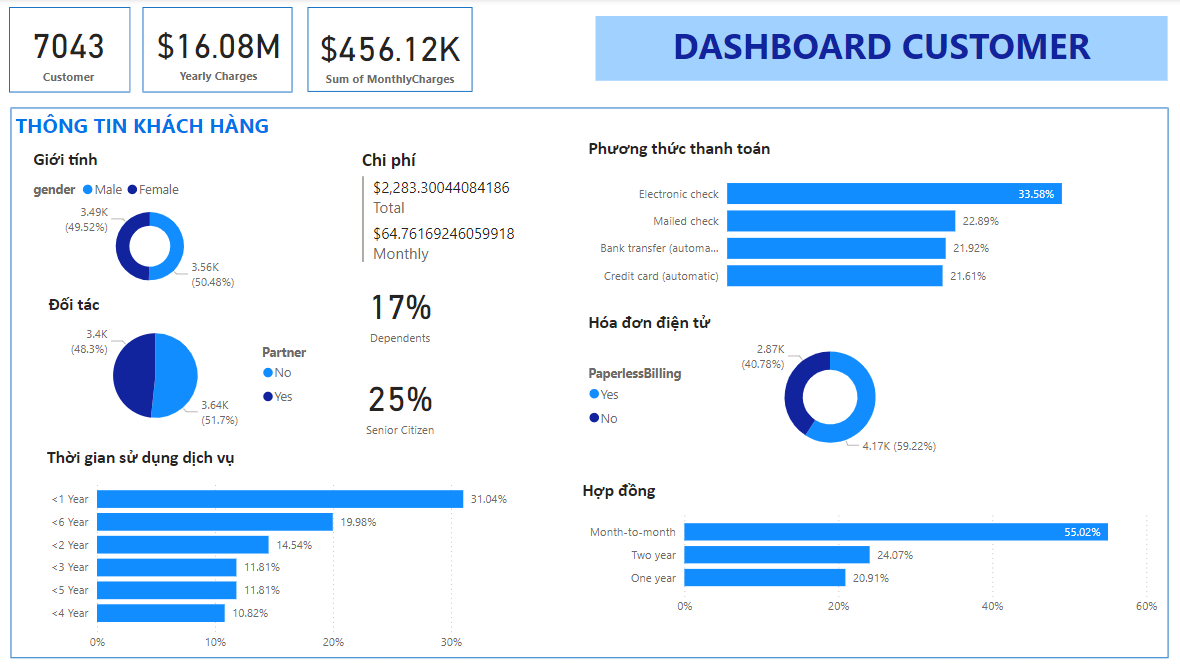
Đầu ra: Dự đoán khả năng rời bỏ dịch vụ của mỗi khách hàng trong tương lai.

## Giải thích dữ liệu

**Table 1. Bảng dữ liệu**

|  |  |  |
| --- | --- | --- |
| STT | Tên thuộc tính | Giải thích |
| 1 | customerID | Mã khách hàng |
| 2 | gender | Giới tính |
| 3 | SeniorCitizen | Người cao tuổi |
| 4 | Partner | Khách hàng có là đối tác không? |
| 5 | Dependents | Khách hàng có sự phụ thuộc không? |
| 6 | tenure | Thời gian khách hàng gắn bó |
| 7 | PhoneService | Khách hàng có sử dụng dịch vụ điện thoại hay không? |
| 8 | MultipleLines | Khách hàng có sử dụng nhiều line điện thoại không? |
| 9 | InternetService | Dịch vụ Internet của khách hàng |
| 10 | OnlineSecurity | Khách hàng có sử dụng dịch vụ bảo mật online không? |
| 11 | OnlineBackup | Khách hàng có sử dụng dịch vụ dự phòng online không? |
| 12 | DeviceProtection | Khách hàng có sử dụng dịch vụ bảo vệ thiết bị không? |
| 13 | TechSupport | Khách hàng có sử dụng dịch vụ hỗ trợ kỹ thuật không? |
| 14 | StreamingTV | Khách hàng có sử dụng dịch vụ truyền hình trực tuyến không? |
| 15 | StreamingMovies | Khách hàng có sử dụng dịch vụ xem phim trực tuyến không? |
| 16 | Contract | Thời hạn hợp đồng của khách hàng |
| 17 | PaperlessBilling | Khách hàng thanh toán có sử dụng hóa đơn giấy không? |
| 18 | PaymentMethod | Phương thức thanh toán của khách hàng |
| 19 | MonthlyCharges | Số tiền mà khách hàng phải trả hàng tháng |
| 20 | TotalCharges | Tổng số tiền khách hàng phải trả |
| 21 | Churn | Tình trạng khách hàng đã rời bỏ hay không? |

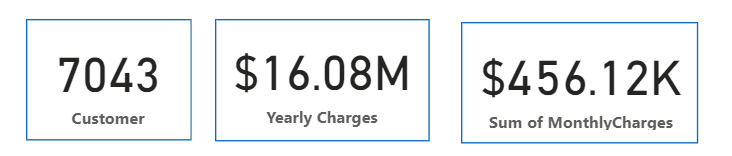
## Trực quan hóa dữ liệu

****

hình 14. Trực quan hóa tổng quát về thông tin khách hàng

* **Tổng quan về thông tin khách hàng :**

**+ Thông tin chung:**

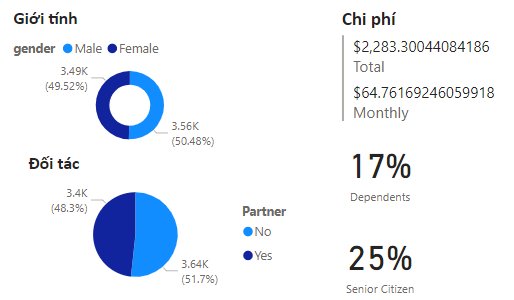


hình 15. Thông tin chung về khách hàng và phí tiền

* Số lượng khách hàng: Có 7043 khách hàng.
* Tổng doanh thu hàng năm từ phí dịch vụ của khách hàng: 16.08 triệu USD
* Tổng chi phí hàng tháng của tất cả khách hàng: 456.12 ngàn USD

**+ Thông tin về khách hàng (Information) :**

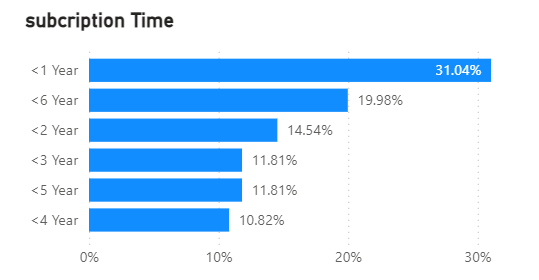
* **Nhân khẩu học**:



hình 16.Thông tin nhân khẩu học

Dựa trên biểu đồ và dữ liệu được cung cấp, tôi có thể chia sẻ một số quan sát và phân tích như sau:

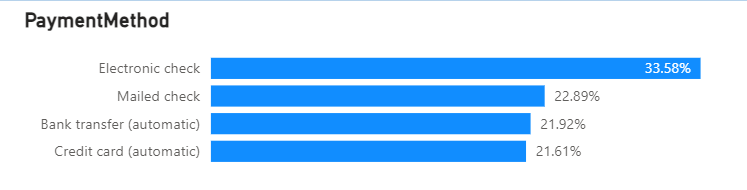
* Giới tính: Tỷ lệ nam và nữ khá cân bằng, với 49.52% là nam và 50.48% là nữ.
* Tỷ lệ người cao tuổi (Senior Citizen): 25% người cao tuổi đã rời bỏ dịch vụ
* Đối tác (Partner): Khách hàng của chúng ta có sự phân bổ tương đối đồng đều giữa những người có đối tác và những người chưa có đối tác ( Trong khoảng 50% mỗi nhóm).
* Tỷ lệ có/không có người phụ thuộc (Dependents): 17% khách hàng có người phụ thuộc, 83% không có.
* Phí trung bình :
  + Tổng số tiền phí mà các khách hàng đã thanh toán: 2283 ngàn USD
  + Mức phí trung bình mà mỗi khách hàng phải trả trong 1 tháng : 64 USD
* **Thời gian đăng ký sử dụng dịch vụ (Subcription Time):**



hình 17.Thời gian đăng ký sử dụng dịch vụ

Dữ liệu này cho thấy hơn 30% khách hàng mới đăng ký sử dụng dịch vụ dưới 1 năm, trong khi khoảng 20% khách hàng đã sử dụng dịch vụ từ 1 đến 6 năm. Đây là thông tin quan trọng về độ trung thành và sự lưu giữ khách hàng của doanh nghiệp.

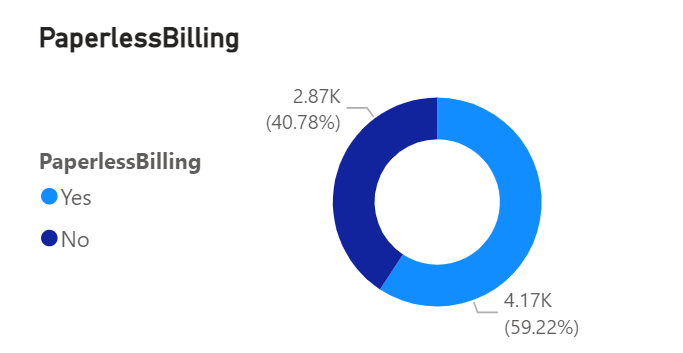
* **Phương thức thanh toán:**



hình 18. Phương thức thanh toán

Phương thức thanh toán phổ biến nhất là electronic check, chiếm tới hơn 33% tổng số. Các phương thức thanh toán tự động như bank transfer và credit card cũng được sử dụng khá phổ biến, với tỷ lệ lần lượt là 21.92% và 21.61%. Phương thức mailed check cũng chiếm một tỷ trọng đáng kể, với 22.89%.

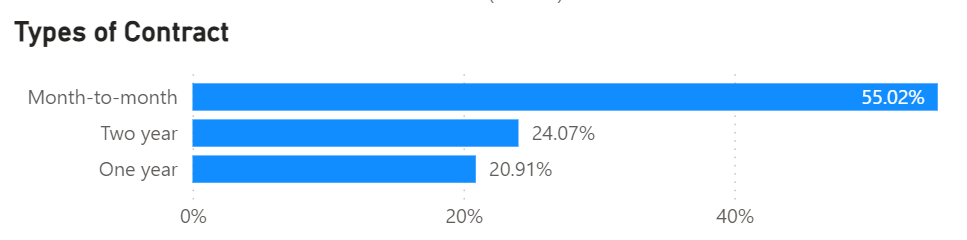
* **Hóa đơn điện tử:**



hình 19. Biểu đồ về hóa đơn điện tử

Dựa trên biểu đồ này, ta có thể thấy thông tin về việc sử dụng dịch vụ PaperlessBilling:

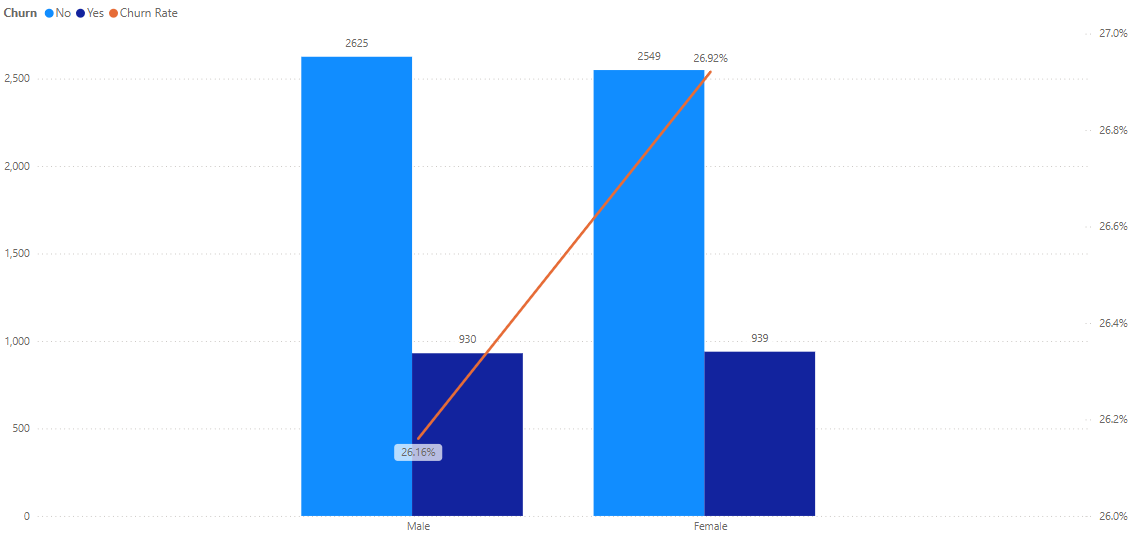
* Tỷ lệ khách hàng sử dụng dịch vụ PaperlessBilling (Yes): 59.22%
* Tỷ lệ khách hàng không sử dụng dịch vụ PaperlessBilling (No): 40.78%
* **Loại hợp đồng:**



hình 20. Loại hợp đồng liên quan đến dịch vụ

Từ đây, ta có thể rút ra một số nhận xét sau:

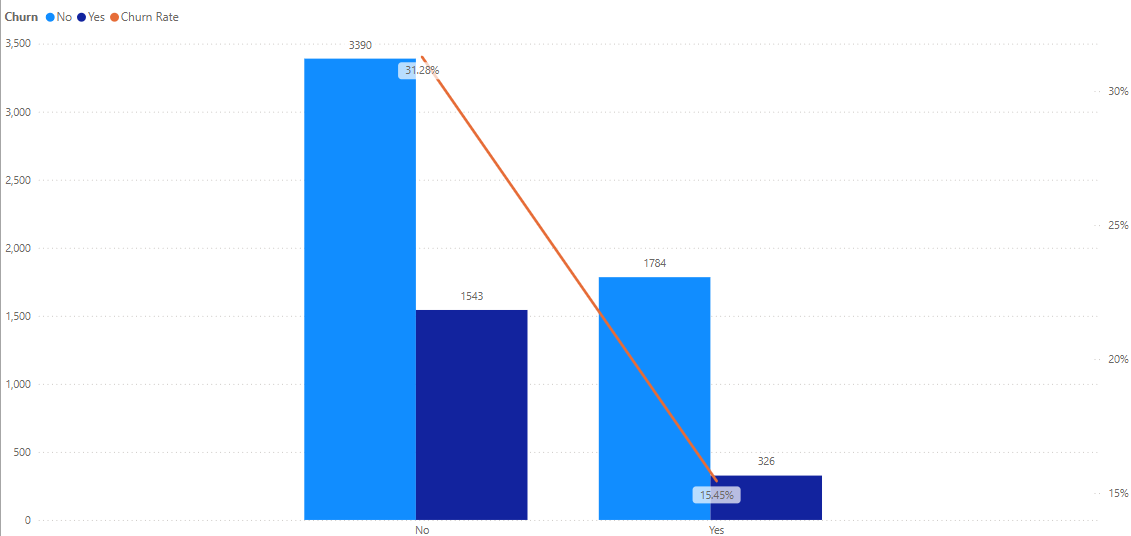
* Hợp đồng tháng-tháng là loại phổ biến nhất, chiếm tới 55.02% tổng số hợp đồng.
* Các hợp đồng dài hạn như 2 năm và 1 năm cũng được sử dụng, nhưng với tỷ lệ thấp hơn, lần lượt là 24.07% và 20.91%.
* **Tổng quan về việc khách hàng rời bỏ dịch vụ**
* **Giới tính (Gender):**



hình 21.Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo giới tính

Nhận xét :

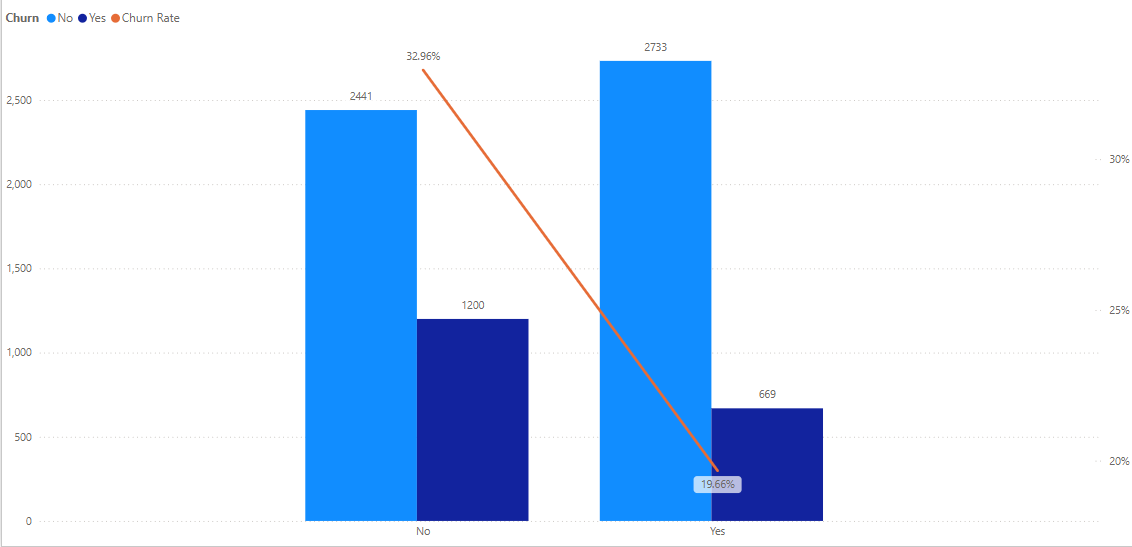
* Khách hàng nam có số lượng “không rời bỏ” nhiều hơn so với khách hàng nữ, nhưng tỷ lệ “rời bỏ” của khách hàng nữ lại cao hơn, đạt 26.92% so với 26.16% ở khách hàng nam.
* Điều này cho thấy khách hàng nữ có xu hướng rời bỏ dịch vụ nhiều hơn so với khách hàng nam.
* **Dependent (Sự phụ thuộc):**



hình 22.Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo Dependents

Nhận xét:

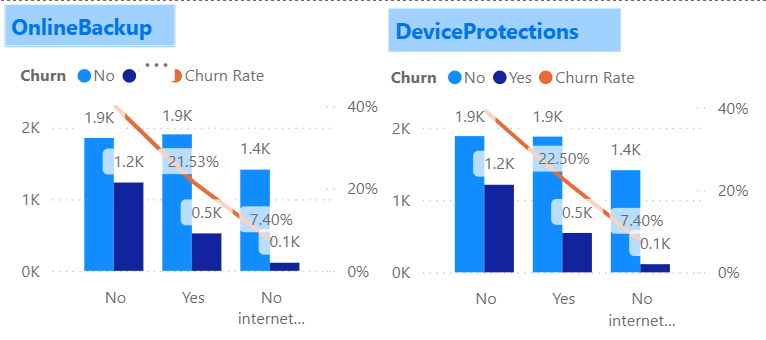
* Khách hàng không phụ thuộc có tỷ lệ rời bỏ cao hơn so với khách hàng có sự phụ thuộc (31.28% so với 21.15%).
* Điều này cho thấy khách hàng không phụ thuộc có xu hướng rời bỏ dịch vụ nhiều hơn so với khách hàng có sự phụ thuộc.
* **Partner (Đối tác):**



hình 23.Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo đối tác Partner

Nhận xét:

* Khách hàng không phải là đối tác có tỷ lệ rời bỏ cao hơn so với khách hàng là đối tác (32.96% so với 19.66%).
* Điều này cho thấy khách hàng không phải là đối tác có xu hướng rời bỏ dịch vụ cao hơn đáng kể so với khách hàng là đối tác.
* **DeviceProtection và Online backup**

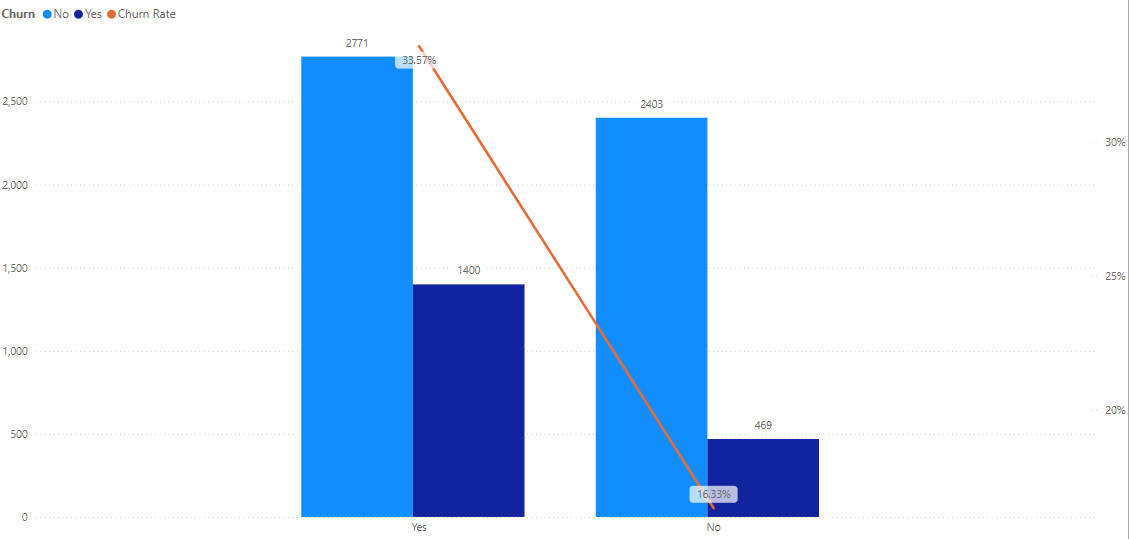


hình 24.Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo DeviceProtection và Online backup

Nhận xét:

Khi nhìn vào tổng thể, các dịch vụ Online Backup và Device Protections đều có xu hướng tương tự về tỷ lệ rời bỏ của khách hàng:

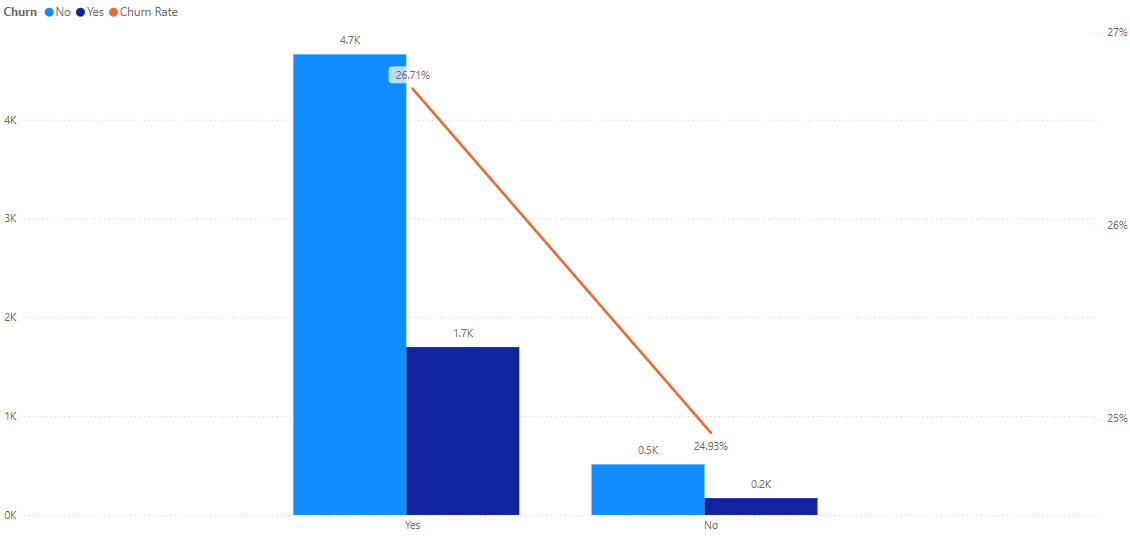
* Khách hàng không sử dụng các dịch vụ (No) đều có tỷ lệ rời bỏ cao nhất.
* Khách hàng sử dụng các dịch vụ (Yes) đều có tỷ lệ rời bỏ thấp hơn, khoảng 21-23%.
* Khách hàng không có dịch vụ Internet (No internet service) đều có tỷ lệ rời bỏ thấp nhất là 7.40%
* **PaperlessBiling:**



hình 25.Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo PaperlessBiling

Nhận xét:

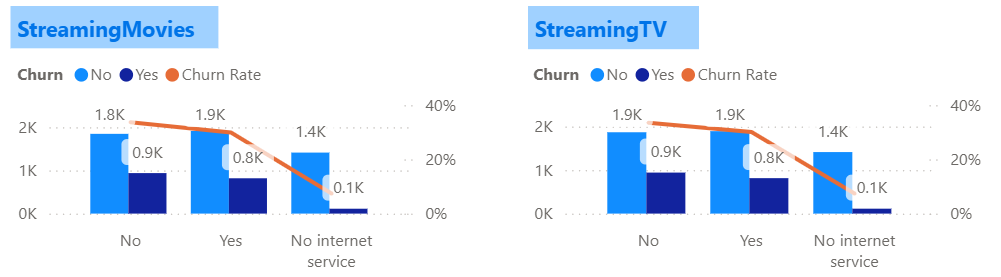
* Nhóm khách hàng sử dụng dịch vụ PaperlessBiling (yes) có tỷ lệ rời bỏ cao nhất, lên đến 33.57%.
* Nhóm khách hàng không sử dụng dịch vụ PaperlessBiling (Yes) có tỷ lệ rời bỏ thấp hơn một nửa, chỉ 16.33%.
* **PhoneService:**



hình 26 Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo PhoneService

Nhận xét:

* Nhóm khách hàng sử dụng dịch vụ PhoneService (yes) có tỷ lệ rời bỏ cao nhất, lên đến 26.71%.
* Nhóm khách hàng không sử dụng dịch vụ PhoneService (No) có tỷ lệ rời bỏ thấp hơn, chỉ 24.93%.
* **StreamingMovies và StreamingTV**

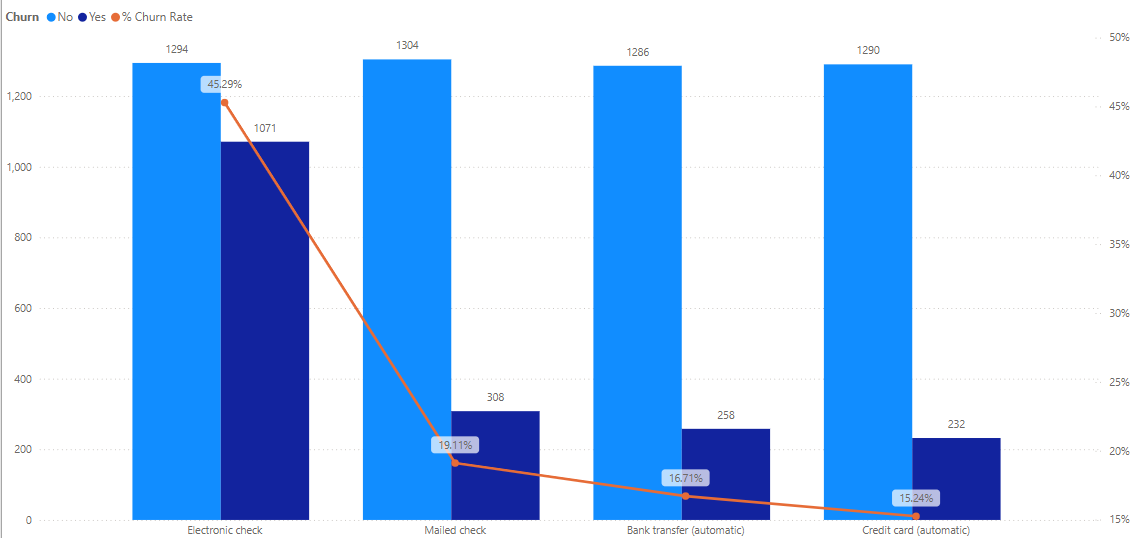


hình 27 Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo StreamingMovies và StreamingTV

Nhận xét:

Khi nhìn vào tổng thể, cả hai dịch StreamingMovies và StreamingTV đều có xu hướng tương tự về tỷ lệ rời bỏ của khách hàng:

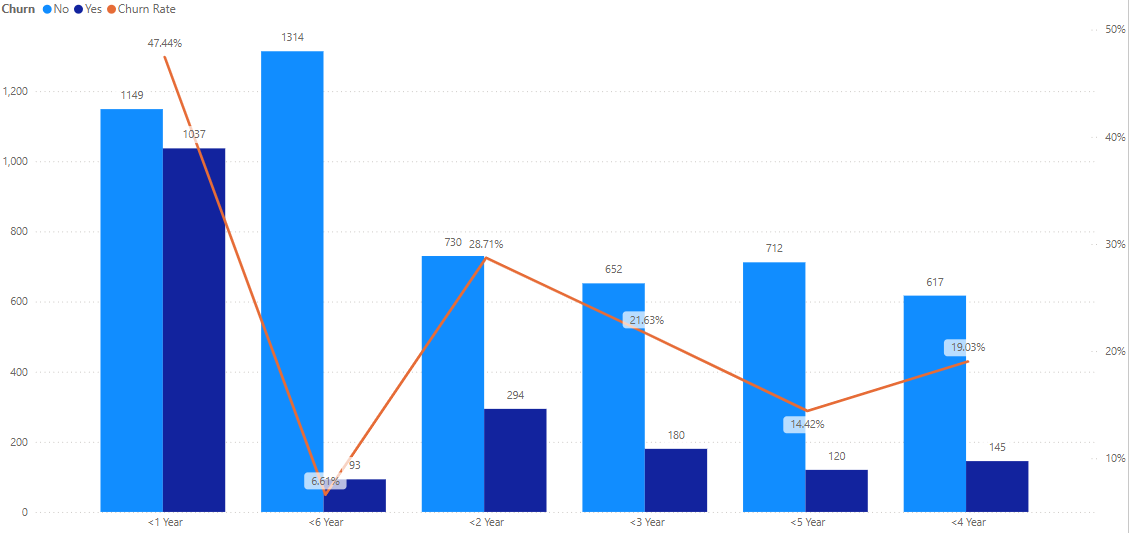
* Khách hàng không sử dụng các dịch vụ (No) đều có tỷ lệ rời bỏ cao nhất.
* Khách hàng sử dụng các dịch vụ (Yes) đều có tỷ lệ rời bỏ thấp hơn
* Khách hàng không có dịch vụ Internet (No internet service) đều có tỷ lệ rời bỏ thấp nhất.
* **PaymentMethod:**



hình 28.Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo PaymentMethod

Nhận xét:

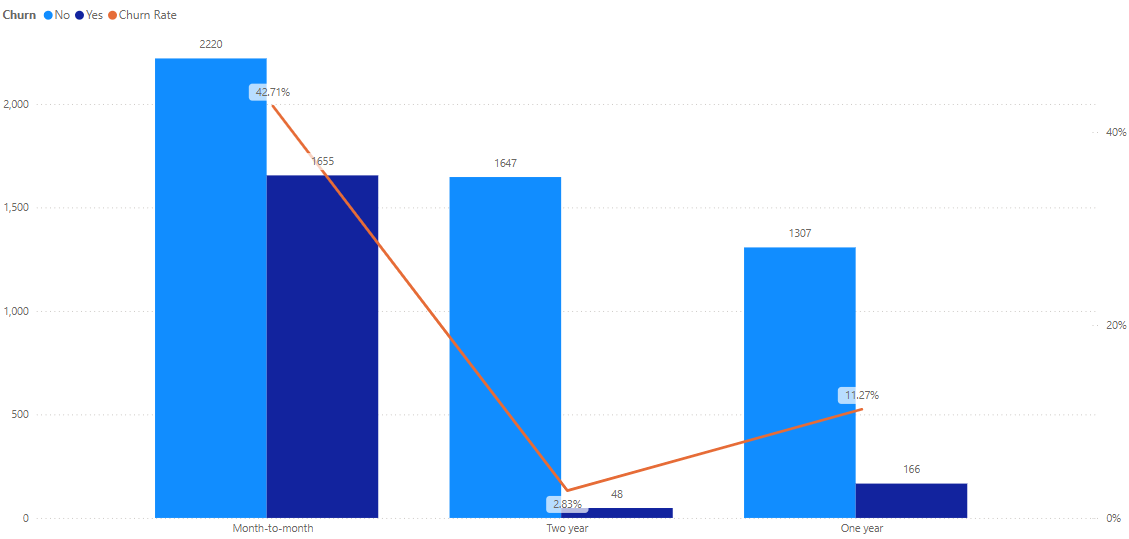
* Phương thức thanh toán “Electronic check” có số lượng người sử dụng cao nhất, đồng thời cũng có tỷ lệ rời bỏ cao nhất (45.29%).
* Tỷ lệ rời bỏ của phương thức “Mailed check” t(19.11%), “Bank transfer (automatic)” 16.71, “Credit card (automatic)” (15.24%) ở mức tương đối và thấp hơn nhiều so với Electronic check”
* **Tenure:**



hình 29.Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo Tenure

Nhận xét:

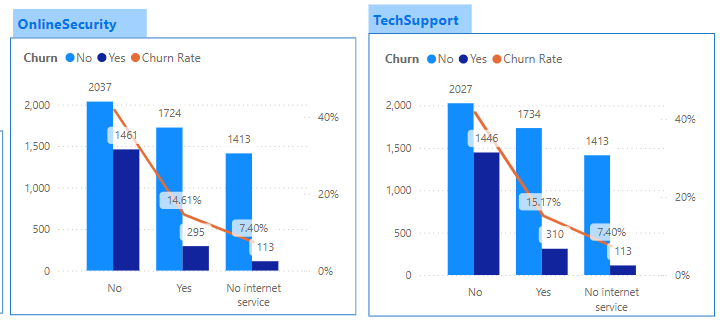
* Với những khách hàng sử dụng dịch vụ dưới 1 năm, tỷ lệ rời bỏ cao nhất lên đến 47.44%. Điều này cho thấy họ dễ rời bỏ trong thời gian sử dụng ban đầu.
* Tuy nhiên, với những khách hàng sử dụng từ 1-2 năm, tỷ lệ rời bỏ giảm xuống chỉ còn 28.71%. Điều này cho thấy họ trở nên ổn định hơn khi sử dụng dịch vụ lâu hơn.
* Đối với những khách hàng sử dụng từ 2-3 năm, tỷ lệ rời bỏ tiếp tục giảm xuống chỉ còn 21.63%.
* Với những khách hàng sử dụng từ khoảng 5 năm, tỷ lệ rời bỏ tiếp tục giảm xuống chỉ còn 14.42%.
* Với những khách hàng sử dụng từ khoảng 4 năm, tỷ lệ rời bỏ có sự đi lên với tỷ lệ là 19.03%.
* Đối với những khách hàng sử dụng khoảng 6 năm, tỷ lệ rời bỏ được cho là thấp nhất, chỉ 6.61%.
* Nhìn chung, có thể thấy rằng thời gian sử dụng càng lâu, tỷ lệ rời bỏ của khách hàng càng giảm
* **Contract:**



hình 30.Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo Contract

Nhận xét:

* Hợp đồng “Month-to-month” có số lượng cao nhất, đồng thời tỷ lệ rời bỏ của loại hợp đồng này cũng rất cao, lên đến 42.71%.
* Hợp đồng “One year” có số lượng thấp hơn “Two year” , nhưng tỷ lệ rời bỏ lại xếp sau “Month-to-month” với tỷ lệ là 11.27%
* Hợp đồng “Two year” có tỷ lệ rời bỏ thấp nhất, chỉ 2.83%.
* **OnlineSecurity và TechSupport**

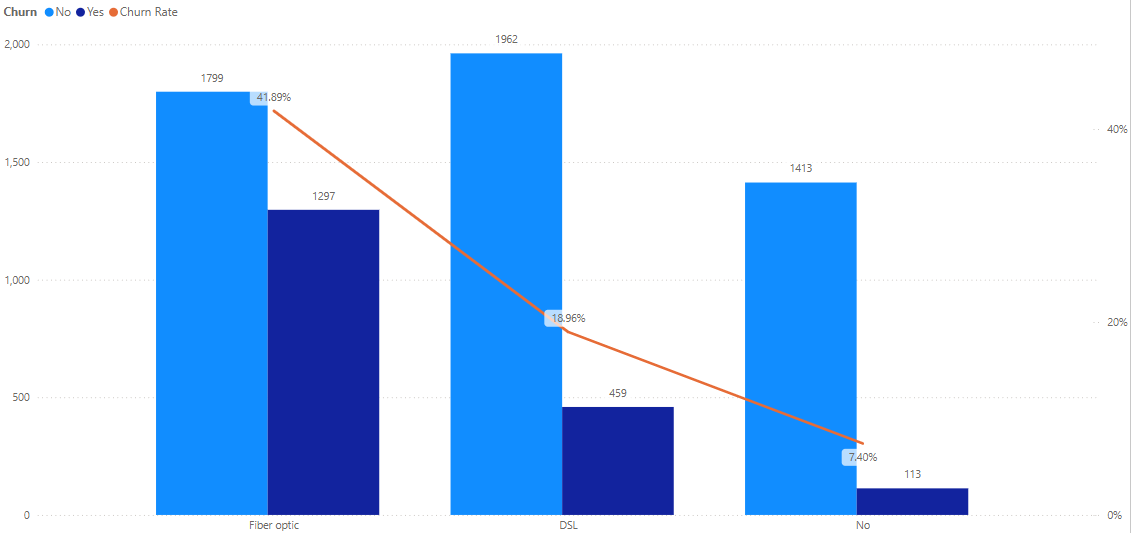


hình 31.Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo OnlineSecurity và TechSupport

Nhận xét:

Khi nhìn vào tổng thể, cả hai dịch OnlineSecurity và TechSupport đều có xu hướng tương tự về tỷ lệ rời bỏ của khách hàng:

* Khách hàng không sử dụng các dịch vụ (No) đều có tỷ lệ rời bỏ cao nhất.
* Khách hàng sử dụng các dịch vụ (Yes) đều có tỷ lệ rời bỏ thấp hơn trong khoảng 14.5%-15%
* Khách hàng không có dịch vụ Internet (No internet service) đều có tỷ lệ rời bỏ thấp nhất là 7.40%
* **InternetService:**



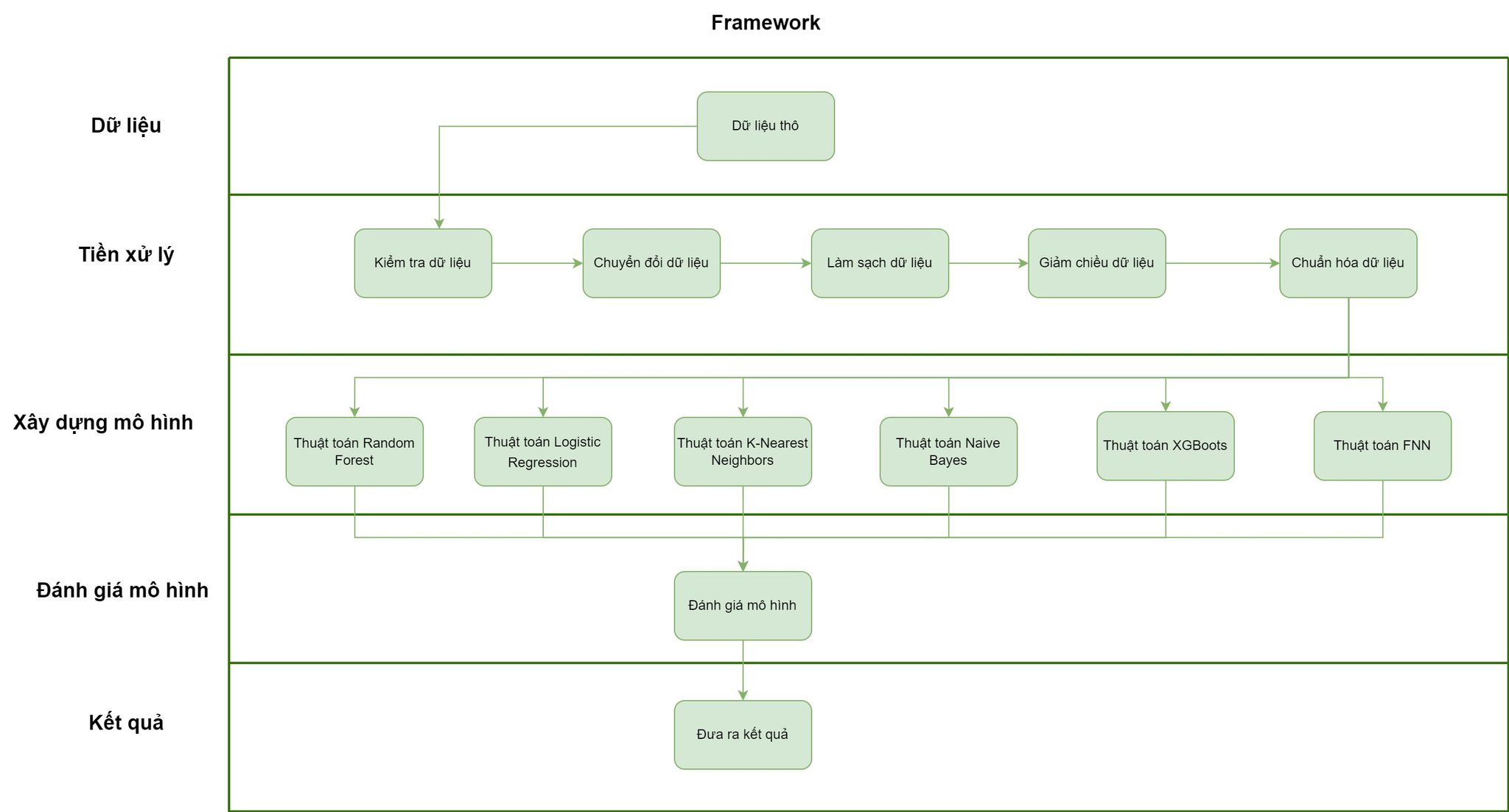
hình 32.Biểu đồ tỷ lệ rời bỏ dịch vụ của khách hàng theo InternetService

Nhận xét:

* Đối với khách hàng lựa chọn “Fiber optic” có tỷ lệ rời bỏ cao nhất, lên đến 41.89%.
* Với khách hàng lựa chọn “DSL” có tỷ lệ rời ở mức tương đối 18.96%.
* Nhóm khách hàng không sử dụng dịch vụ internet có tỷ lệ rời bỏ thấp nhất, chỉ 7.40%.

1. **XÂY DỰNG MÔ HÌNH DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ DỊCH VỤ VIỄN THÔNG**

## Frame work



hình 33. Framework cho dự án

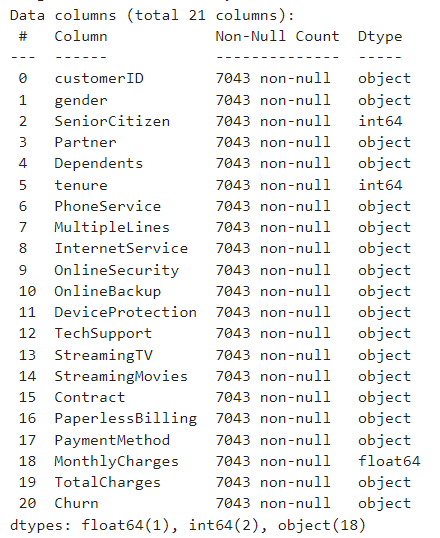
## Tiền xử lý dữ liệu

Tiền xử lý dữ liệu (preprocessing data) là quá trình chuẩn bị và biến đổi dữ liệu từ dạng ban đầu thành dữ liệu có thể được sử dụng hiệu quả trong các mô hình học máy và phân tích dữ liệu. Quá trình này rất quan trọng và thường bao gồm các bước như sau:



hình 34. Tiền xử lý dữ liệu

* + 1. **Kiểm tra dữ liệu tổng quát**
* Kiểm tra dữ liệu



hình 35.Thông tin kiểm tra của các cột trong dữ liệu

Bộ dữ liệu Customer\_Churn gồm có 21 cột và 7043 dòng và không có giá trị bị trống.

* + 1. **Chuyển đổi dữ liệu**

**Chuyển đổi dữ liệu** là quá trình chuyển đổi dữ liệu từ các định dạng không phù hợp như chuỗi văn bản, ngày tháng,…thành các định dạng phù hợp với tính năng.

* Chuyển đổi cột TotalCharges từ object sang dạng numeric.
  + 1. **Làm sạch dữ liệu**

Làm sạch dữ liệu là quá trình xử lý các giá trị bị thiếu, giá trị nhiễu, giá trị ngoại lai từ tập dữ liệu ban đầu nhằm tránh làm sai lệch kết quả và gây ra các quyết định sai lầm hoặc không thực tế.

* Kiểm tra dữ liệu

Sau khi chuyển đổi dữ liệu chúng ta thực hiện lại việc kiểm tra dữ liệu



hình 36 Kiểm tra dữ liệu có thiếu giá trị hay không

Cột TotalCharges bị thiếu 11 giá trị

* **Thay thế các giá trị thiếu của cột TotalCharges**

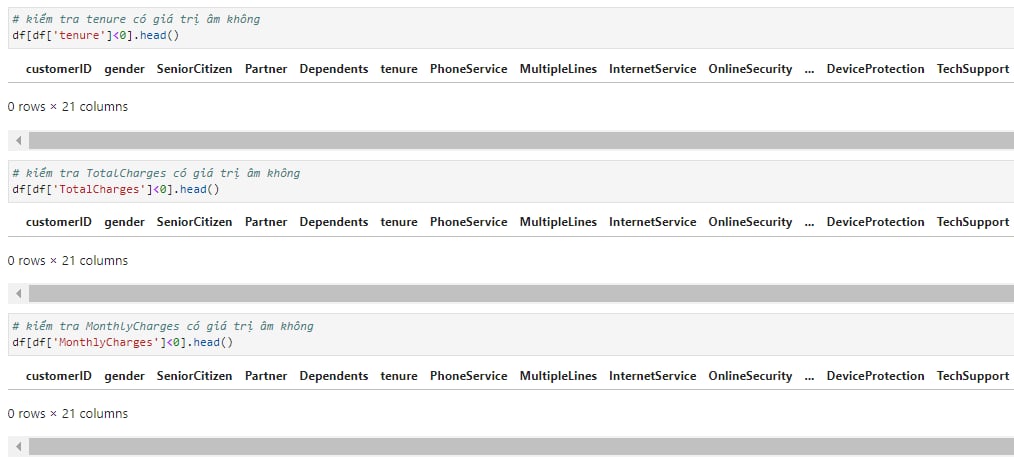


hình 37.Thay thế các giá trị và được kết quả như hình

Thay các giá trị bị thiếu của cột TotalCharges thành giá trị trung bình và kiểm tra lại thấy các cột đã đầy đủ.

* **Kiểm tra các giá trị không phù hợp**

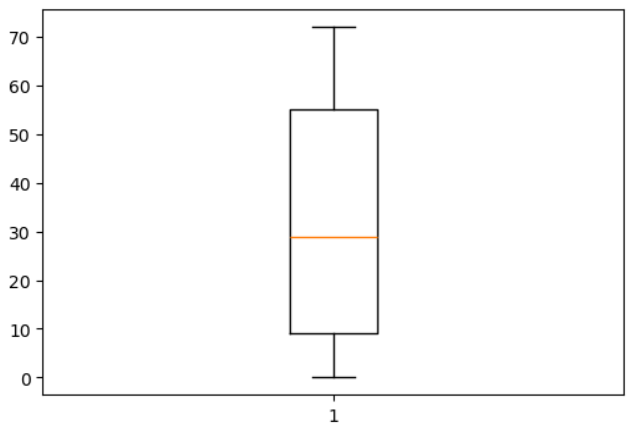
Kiểm tra các cột Tenure, TotalCharges, MonthlyCharges có các giá trị bất thường hay không ( các giá trị âm).

****

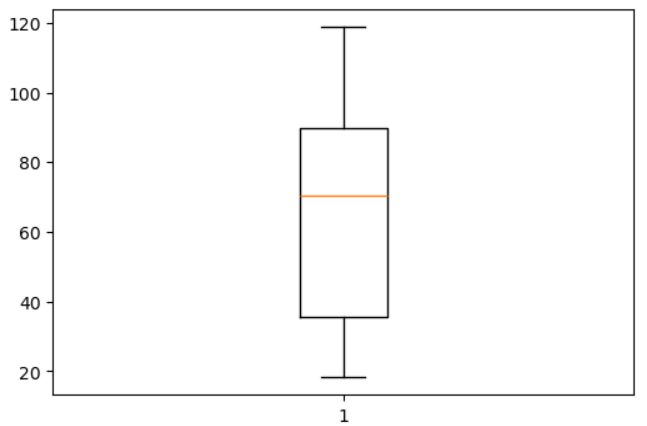
hình 38.Kiểm tra giá trị có bị bất thường hay không

Các cột Tenure, TotalCharges, MonthlyCharges không có giá trị bất thường.

* **Dùng biểu đồ boxplot để xem các giá trị ngoại lai**



hình 39.Biểu đồ boxplot của Tenure



hình 40.Biểu đồ boxplot của MonthlyCharges



hình 41 Biểu đồ boxplot của TotalCharges

Các cột Tenure, MonthlyCharges, TotalCharges không có giá trị ngoại lai.

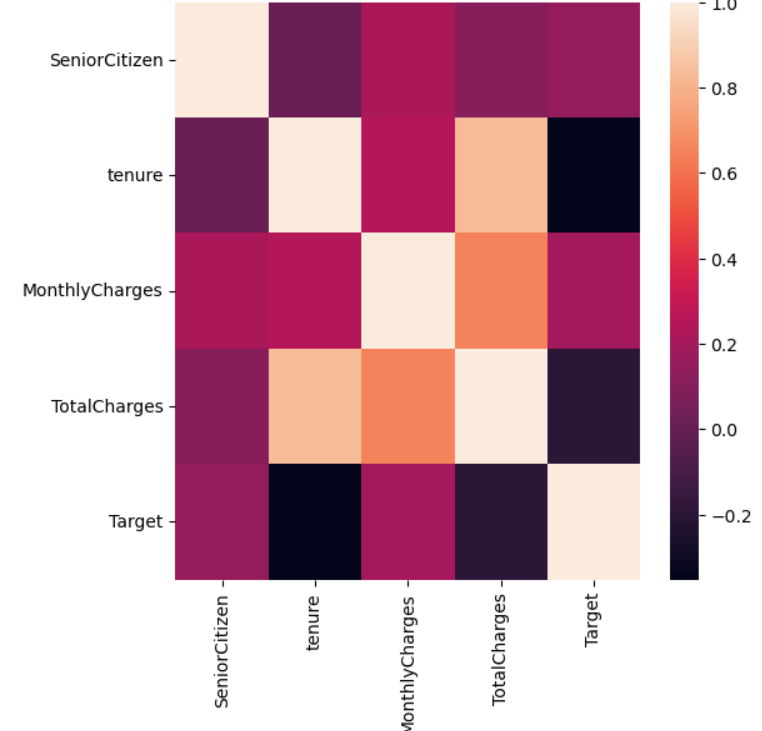
* + 1. **Giảm chiều dữ liệu**

Đôi khi dữ liệu có quá nhiều đặc trưng (features) không cần thiết, có thể dẫn đến hiện tượng overfitting và tăng chi phí tính toán. Vậy nên chúng ta cần giảm bớt các chiều không cần thiết và giữ lại các chiều quan trọng để xây dựng mô hình nhằm bớt bớt chi phí và thời gian xây dựng mô hình.

Sử dụng ma trận tương quan (correlation matrix) và lựa chọn ngưỡng (threshold) để loại bỏ các biến có mức độ tương quan thấp.

***Bước 1:*** Tạo num\_features chứa các features có kiểu dữ liệu là mumber, ob\_features sẽ lưu các cột có kiểu dữ liệu là object ngoại trừ CustormerID.

***Bước 2:*** Xem biểu đồ tương quan của num\_features và ob\_features



hình 42.Biểu đồ tương quan của num\_features



hình 43.Biểu đồ tương quan của ob\_features

Biểu đồ tương quan: Dựa vào thanh màu sắc để xác định mức độ tương quan giữa các feature. Màu càng sáng thì mối tương quan thuận càng lớn và ngược lại, màu càng tối thì có thể không tồn tại hoặc tồn tại tương quan nghịch.

***Bước 3:*** Lựa chọn num\_features và ob\_feature quan trọng:

Để xác định được các features quan trọng mình cần chọn được ngưỡng xác định phù hợp với bộ dữ liệu. Nếu chọn ngưỡng xác định quá thấp, bộ dữ liệu vẫn còn lại rất nhiều feature và những feature không quan trọng có thể chưa được loại bỏ. Nếu chọn ngưỡng xác định quá cao, bộ dữ liệu sẽ có thể mất đi nhiều feature quan trong và còn quá ít feature để xây dựng mô hình.

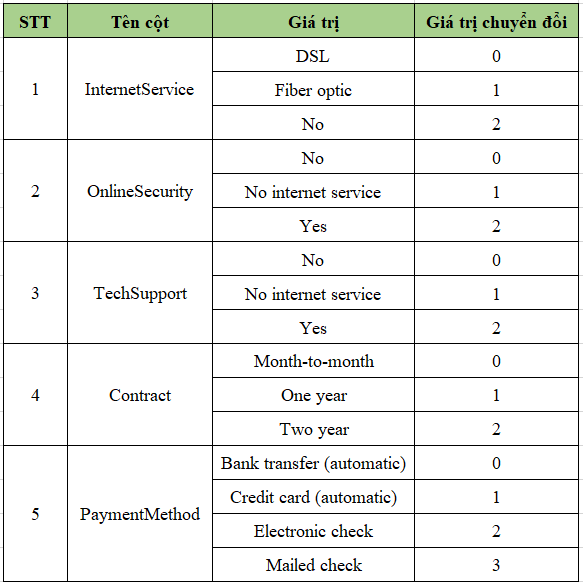
Với sự so sánh các hiệu suất của mô hình ở các ngưỡng xác định khác nhau cho ra hiểu suất mô hình khác nhau. Trong đó ngưỡng xác định = 0.3 cho ra hiệu suất sau khi xây dựng mô hình là cao nhất.

Thực hiện lựa chọn ngưỡng xác định = 0.3 lúc này bộ dữ liệu Customer\_Churn sẽ còn các cột: “tenure”, “InternetService”, “OnlineSecurity”, “TechSupport”, “Contract”, “PaymentMethod”, “Churn”.

* + 1. **Chuẩn hóa dữ liệu**

Chuẩn hóa dữ liệu là quá trình chuẩn hóa dữ liệu về đơn vị đo lường phù hợp nhằm chuẩn bị dữ liệu đầu vào phù hợp giúp tối ưu hóa và cải thiện hiệu suất của các mô hình học máy, đồng thời giúp các thuật toán học máy làm việc hiệu quả hơn với dữ liệu thực tế.

Bằng cách sử dụng *LabelEncoder* từ thư viện *sklearn.preprocessing*. Các cột có kiểu dữ liệu là object trong Customer\_Churn lúc này sẽ được chuyển về mã hóa int với:



hình 44.Thông tin giá trị chuyển đổi

## Mô hình Random Forest

* + 1. **Xây dựng mô hình**
* **Chia dữ liệu**

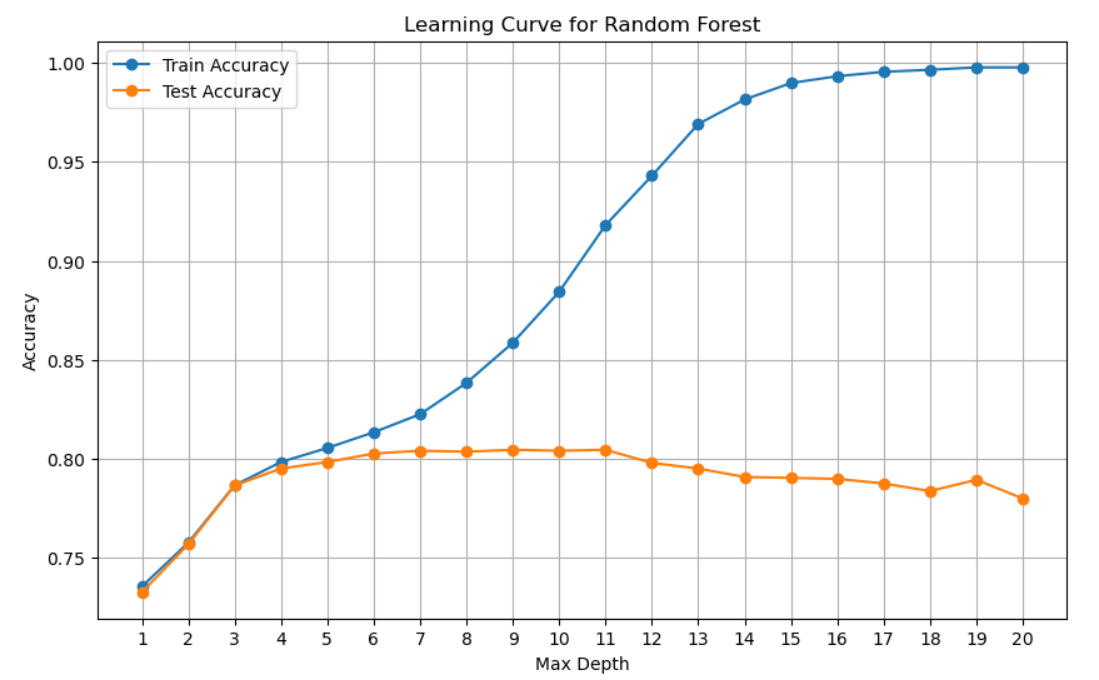
Mô hình được chia thành 2 phần là X\_train và y\_train chiếm 70% bộ dữ liệu dùng để training mô hình, X\_test và y\_test chiếm 30% bộ dữ liệu dùng để test mô hình và đánh giá hiệu quả mô hình.

* **Chọn giá trị max\_depth phù hợp với bộ dữ liệu**

Max\_depth: Là tham số quy định độ sâu tối đa của cây quyết định. Nó xác định số lượng các node từ gốc (root) đến lá (leaf) mà cây được phép phát triển.

Biểu đồ Learning Curve:

* Train Accuracy: độ chính xác trên tập huấn luyện
* Test Accuracy: độ chính xác trên tập testing



hình 45.Biểu đồ Learning Curve của mô hình Random Forest

 Quan sát biểu đồ Learning Curve:

* Train Accuracy tăng lên và sau đó ổn định, có nghĩa là mô hình đang học được từ dữ liệu.
* Testing Accuracy đạt đỉnh điểm tại max\_depth = 11 và sau đó giảm dần, có thể max\_depth đã chọn quá cao và gây overfitting.

=> Vì vậy giá trị max\_depth phù hợp cho mô hình có giá trị trong khoảng từ 6 đến 10.

* **Khởi tạo mô hình**

Khởi tạo mô hình là quá trình tạo ra một cấu trúc mô hình từ dữ liệu huấn luyện và đặt các thông số cần thiết cho mô hình trước khi bắt đầu quá trình huấn luyện.

Trong mô hình Random Forest các thông số xây dựng mô hình gồm có:

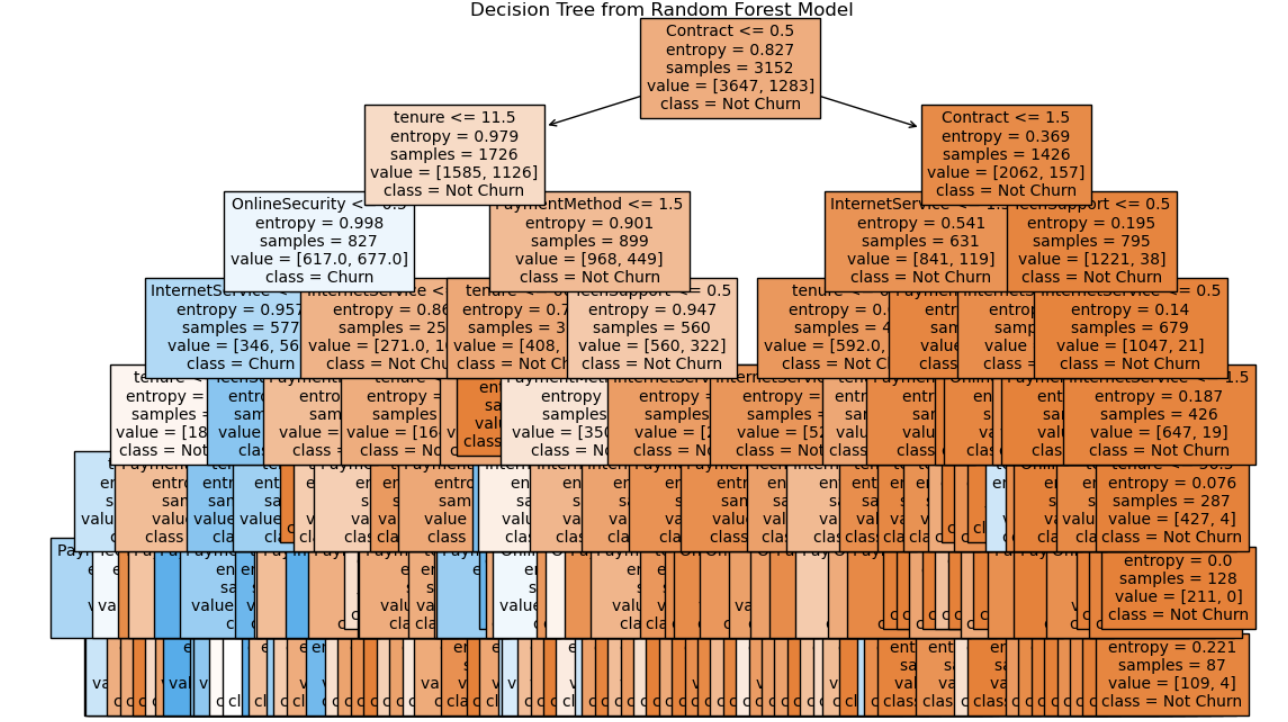
* criterion = “entropy” : Chỉ số đo mức độ hỗn loạn của lá.
* max\_depth = 7 : Độ sâu của cây.
* n\_estimators = 99 : Số cây xây dựng của rừng
* **Huấn luyện mô hình**

Quá trình huấn luyện mô hình sẽ sử dụng bộ dữ liệu X\_train và y\_train để xây dựng ra mô hình dự đoán với các thông số chúng ta khởi tạo.

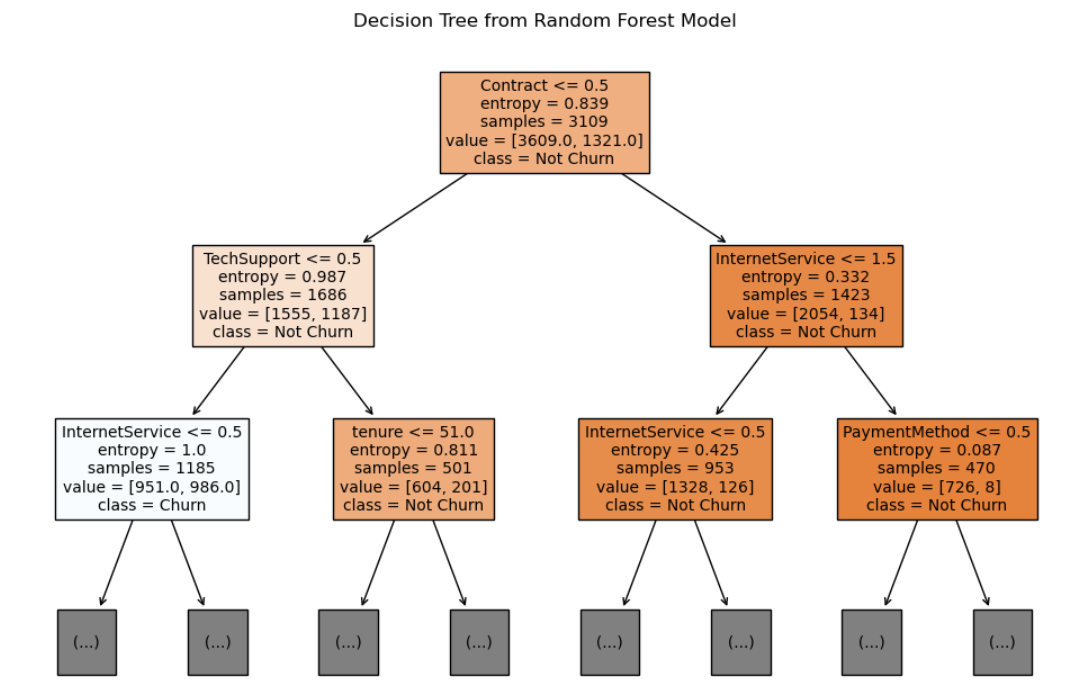
* **Dự báo nhãn của bộ dữ liệu test**

Sau khi đã huấn luyện mô hình trên tập dữ liệu huấn luyện, việc dự báo nhãn trên tập dữ liệu test là bước quan trọng trong quá trình đánh giá và kiểm tra hiệu suất của mô hình.

* **Cây của mô hình Random Forest**

****

hình 46.Cây hoàn chỉnh trong Random forest

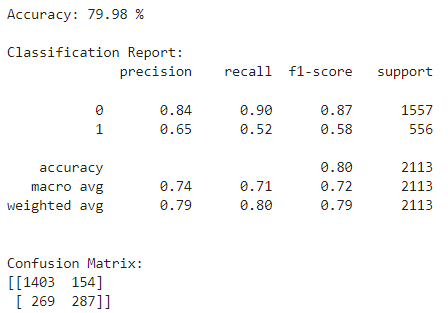


hình 47.Cây không đầy đủ của mô hình Random forest

Các chỉ số trong 1 nút bao gồm:

* + Tên feature và giá trị
  + Entropy: là một độ đo của sự không chắc chắn, entropy đo lường mức độ hỗn loạn trong việc phân loại các mẫu tại node đó. Giá trị entropy càng cao thể hiện sự không chắc chắn càng lớn trong việc phân loại các mẫu.
  + Samples: Đây là số lượng mẫu tại node đó.
  + Value: Giá trị này biểu thị số lượng của lớp “Not Churn” và “Churn”  tại node đó so với bộ dữ liệu training.
  + Class: Đây là lớp được dự đoán cho node đó dựa trên quy tắc phân loại của cây quyết định.
    1. **Đánh giá mô hình**
* **Đánh giá các chỉ số**

Dựa vào kết quả dự báo nhãn của bộ dữ liệu test (y\_prep) và nhãn của bộ dữ liệu test ( y\_test) chúng ta sẽ đáng giá mô hình qua các chỉ số sau:



hình 48.Kết quả đánh giá của mô hình Random Forest

Độ chính xác (Accuracy) là tỷ lệ các dự đoán đúng so với tổng số dự đoán. Trong mô hình, độ chính xác là 79.98%, nghĩa là mô hình đã dự đoán đúng 79.98% nhãn của toàn bộ khách hàng. Đây là một tỷ lệ khá cao, cho thấy mô hình có khả năng dự đoán khá chính xác.

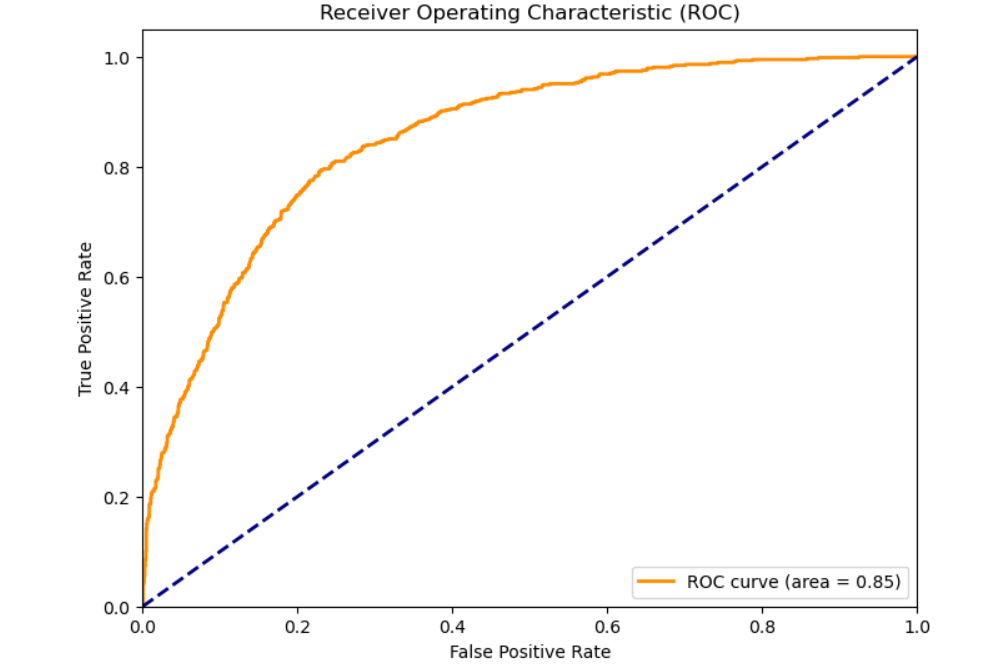
Ma trận nhầm lẫn có thể được giải thích như sau:

* 1403: Số khách hàng thực tế không rời bỏ dịch vụ được dự đoán đúng là “không rời bỏ”.
* 154: Số khách hàng thực tế không rời bỏ dịch vụ được dự đoán sai là “rời bỏ”.
* 269: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán sai là “không rời bỏ”.
* 287: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán đúng là “rời bỏ”.
* **Đánh giá đồ thị AUC-ROC**

Chỉ dựa vào độ chính xác không đủ để đánh giá toàn diện hiệu quả của mô hình này. Do đó, cần bổ sung thêm các chỉ số khác để có cái nhìn tổng quan hơn.

Vì vậy , ta có thể sử dụng điểm AUC-ROC làm thước đo hiệu quả mô hình, AUC-ROC (Area Under the Receiver Operating Characteristic Curve) là một thước đo hiệu quả phổ biến cho các mô hình phân loại, đánh giá khả năng phân biệt giữa các lớp (ví dụ: 0 và 1, khách hàng rời bỏ dịch vụ và không rời bỏ, v.v.).

Điểm AUC càng cao, mô hình càng có khả năng phân biệt chính xác giữa các lớp.



hình 49. Biểu đồ Roc của mô hình Random Forest

Dựa vào biểu đồ ROC, ta có thể thấy:

* Độ cong của đường cong: Đường cong có xu hướng cong lên từ góc dưới bên trái về phía góc trên bên trái. Điều này cho thấy mô hình có khả năng phân biệt tốt giữa nhóm khách hàng sẽ bỏ dịch vụ và nhóm sẽ tiếp tục sử dụng.
* Diện tích dưới đường cong (AUC): Giá trị AUC là 0.85. Điều này cho thấy mô hình có độ chính xác tổng thể tốt trong việc phân loại khách hàng sẽ bỏ dịch vụ và sẽ tiếp tục sử dụng.

Nhìn chung, mô hình Random Forest dựa trên biểu đồ ROC này có vẻ như có khả năng dự đoán tốt về khách hàng sẽ bỏ dịch vụ. Giá trị AUC ở mức 0.85 cho thấy mô hình có thể phân biệt hiệu quả giữa các nhóm khách hàng. Công ty có thể sử dụng thông tin này để xây dựng các chiến lược giữ chân khách hàng hiệu quả hơn.

## Mô hình Logistic Regression

* + 1. **Xây dựng mô hình**
* **Chia tập dữ liệu**

Mô hình được chia thành 2 phần là X\_train và  y\_train  chiếm 70% bộ dữ liệu dùng để training mô hình, X\_test và y\_test chiếm 30% bộ dữ liệu dùng để test mô hình và đánh giá hiệu quả mô hình.

* **Khởi tạo mô hình**

Trong mô hình Logistic Regression các thông số xây dựng mô hình gồm có:

* penalty = L2: dùng để giảm thiểu overfitting và cải thiện tổng quát hóa của mô hình.
* max\_iter = 100: xác định số lần lặp tối đa mà thuật toán sẽ thực hiện trong quá trình huấn luyện.
* tol = 1e-4 ( 0.0001): tol là ngưỡng dừng, ví dụ 1e-4 tức là dừng khi hàm mất mát thấp hơn 1e-4.
* **Huấn luyện mô hình**

Quá trình huấn luyện mô hình sẽ sử dụng bộ dữ liệu X\_train và y\_train để xây dựng ra mô hình dự đoán.

* **Dự báo nhãn của bộ dữ liệu test**

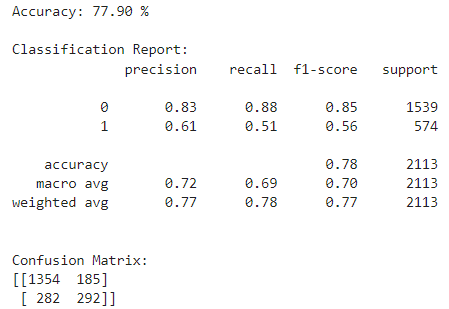
Sau khi đã huấn luyện mô hình trên tập dữ liệu huấn luyện, việc dự báo nhãn trên tập dữ liệu test là bước quan trọng trong quá trình đánh giá và kiểm tra hiệu suất của mô hình.

* **Hàm dự đoán**

****

* + 1. **Đánh giá mô hình**
* **Đánh giá các chỉ số**

Dựa vào kết quả dự báo nhãn của bộ dữ liệu test (y\_prep) và nhãn của bộ dữ liệu test ( y\_test) chúng ta sẽ đáng giá mô hình qua các chỉ số sau:

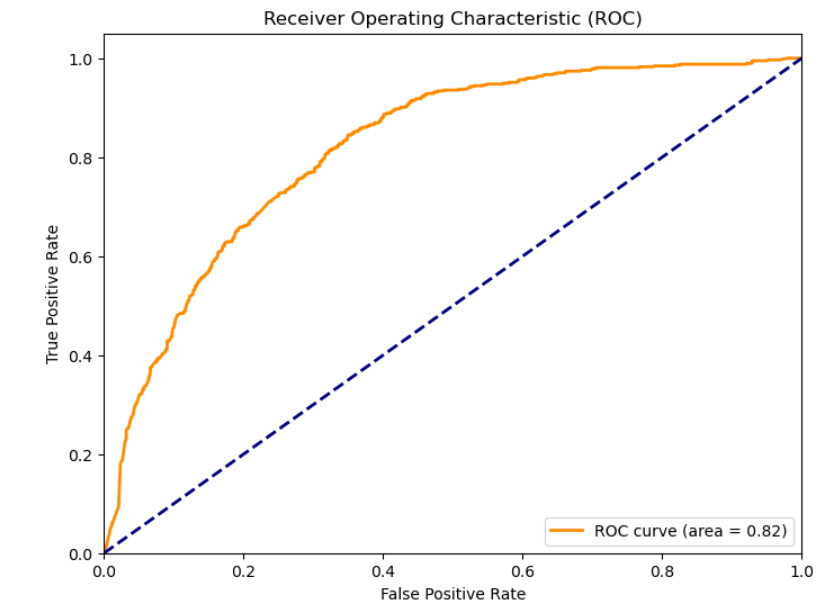
****

hình 50.Kết quả đánh giá của mô hình Logistic

Độ chính xác (Accuracy) là tỷ lệ các dự đoán đúng so với tổng số dự đoán. Trong mô hình, độ chính xác là 77.90%, nghĩa là mô hình đã dự đoán đúng 77.90% nhãn của toàn bộ khách hàng. Đây là một tỷ lệ khá cao, cho thấy mô hình có khả năng dự đoán khá chính xác.

Ma trận nhầm lẫn có thể được giải thích như sau:

* 1354: Số khách hàng thực tế không rời bỏ dịch vụ được dự đoán đúng là “không rời bỏ”.
* 185: Số khách hàng thực tế không rời bỏ dịch vụ được dự đoán sai là “rời bỏ”.
* 282: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán sai là “không rời bỏ”.
* 292: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán đúng là “rời bỏ”.
* **Đánh giá biểu đồ ROC**

****

hình 51.Biểu đồ ROC của mô hình hồi quy Logistic

Trong trường hợp này, AUC là 0.82, cho thấy mô hình có khả năng phân loại tốt. Đường cong ROC cong lên phía góc trên trái, cho thấy mô hình có độ chính xác cao. Với AUC là 0.82, mô hình này có thể đạt được 83% độ chính xác trong việc phân loại các trường hợp “rời bỏ dịch vụ” và “không rời bỏ dịch vụ”.

Đây là một kết quả khá tốt, cho thấy mô hình có hiệu suất phân loại ở mức khá cao.

Nhìn chung, biểu đồ ROC cung cấp thông tin hữu ích để đánh giá hiệu suất của mô hình dự đoán khách hàng rời bỏ dịch vụ, qua đó giúp đưa ra các quyết định kinh doanh và chiến lược phù hợp.

## Mô hình K-Nearest Neighbors

* + 1. **Xây dựng mô hình**
* **Chia dữ liệu**

Mô hình được chia thành 2 phần là X\_train và  y\_train  chiếm 70% bộ dữ liệu dùng để training mô hình, X\_test và y\_test chiếm 30% bộ dữ liệu dùng để test mô hình và đánh giá hiệu quả mô hình.

* **Khởi tạo mô hình**

Khởi tạo mô hình KNN để chuẩn bị cho bước huấn luyện mô hình

Trong mô hình KNN các thông số xây dựng mô hình gồm có:

* n\_neighbors = 113: Số lượng láng giềng gần nhất mà mô hình sẽ sử dụng để dự đoán lớp của điểm dữ liệu mới.
* weights = 'uniform': Tất cả các láng giềng có cùng một trọng số.
* metric = 'manhattan': Phương pháp tính khoảng cách giữa các điểm dữ liệu.
* **Huấn luyện mô hình**

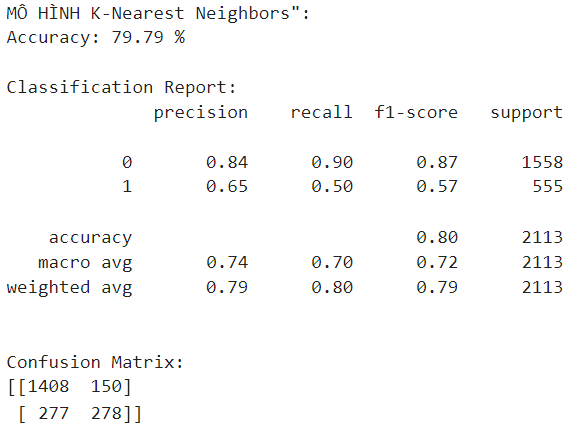
Quá trình huấn luyện mô hình sẽ sử dụng bộ dữ liệu X\_train và y\_train để xây dựng ra mô hình dự đoán.

* **Dự báo nhãn của bộ dữ liệu test**

Sau khi đã huấn luyện mô hình trên tập dữ liệu huấn luyện, việc dự báo nhãn trên tập dữ liệu test là bước quan trọng trong quá trình đánh giá và kiểm tra hiệu suất của mô hình.

* + 1. **Đánh giá mô hình**
* **Đánh giá các chỉ số**

Dựa vào kết quả dự báo nhãn của bộ dữ liệu test (y\_prep) và nhãn của bộ dữ liệu test ( y\_test) chúng ta sẽ đáng giá mô hình qua các chỉ số sau:

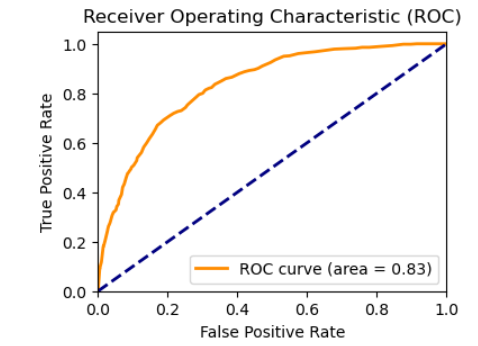


hình 52.Kết quả đánh giá của mô hình KNN

Độ chính xác (Accuracy) là tỷ lệ các dự đoán đúng so với tổng số dự đoán. Trong mô hình, độ chính xác là 79.79, nghĩa là mô hình đã dự đoán đúng 79.79% nhãn của toàn bộ khách hàng. Đây là một tỷ lệ khá cao, cho thấy mô hình có khả năng dự đoán khá chính xác.

Ma trận nhầm lẫn có thể được giải thích như sau:

* 1408: Số khách hàng thực tế không rời bỏ dịch vụ được dự đoán đúng là “không rời bỏ”.
* 150: Số khách hàng thực tế không rời bỏ dịch vụ được dự đoán sai là “rời bỏ”.
* 277: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán sai là “không rời bỏ”.
* 278: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán đúng là “rời bỏ”.
* **Đánh giá đồ thị ROC**



hình 53.Biểu đồ ROC của mô hình KNN

Với AUC = 0,83, mô hình KNN có khả năng phân biệt tương đối tốt giữa hai lớp rời bỏ và không rời bỏ. Cụ thể, mô hình có khoảng 83% cơ hội phân loại đúng các trường hợp rời bỏ và không rời bỏ.

Điều này có thể giúp doanh nghiệp dự đoán với độ chính xác khá tốt những khách hàng có nguy cơ rời bỏ dịch vụ. Từ đó, họ có thể áp dụng các chiến lược chăm sóc khách hàng phù hợp để giữ chân khách hàng.

Tuy nhiên, để đạt hiệu quả tối ưu, vẫn cần tiếp tục tối ưu hóa mô hình KNN, chẳng hạn như thử nghiệm các giá trị K khác nhau, cân bằng các mức false positive và true positive, hoặc kết hợp với các đặc trưng khác của dữ liệu.

Nhìn chung, biểu đồ ROC với AUC = 0,83 cho thấy mô hình KNN đang hoạt động khá tốt trong bài toán dự đoán rời bỏ dịch vụ. Tuy nhiên, vẫn còn nhiều cơ hội để tiếp tục cải thiện hiệu suất của mô hình.

## Mô hình Naive Bayes

* + 1. **Xây dựng mô hình**

Mô hình Naive Bayes là một nhóm các thuật toán phân loại đơn giản dựa trên lý thuyết xác suất Bayes với giả định độc lập mạnh mẽ (naive) giữa các đặc trưng (features). Trong Python, các mô hình Naive Bayes thường được triển khai trong thư viện scikit-learn với các biến thể như Gaussian Naive Bayes, Multinomial Naive Bayes và Bernoulli Naive Bayes, Categorical Naive Bayes. Mỗi biến thể có các tham số đầu vào riêng biệt và phù hợp với từng bài toán khác nhau. Ở dự án chúng ta sẽ sử dụng biến thể Categorical Naive Bayes phù hợp với bộ dữ liệu danh mục đã được mã hóa số của dự án.

* **Chia dữ liệu**

Mô hình được chia thành 2 phần là X\_train và  y\_train  chiếm 70% bộ dữ liệu dùng để training mô hình, X\_test và y\_test chiếm 30% bộ dữ liệu dùng để test mô hình và đánh giá hiệu quả mô hình.

* **Khai báo tập tham số**

Khi huấn luyện mô hình Naïve Bayes, có thể có các đặc trưng hoặc tổ hợp đặc trưng mà không xuất hiện trong dữ liệu huấn luyện. Điều này có thể dẫn đến xác suất bằng 0, làm giảm khả năng tổng quát hóa của mô hình.

Các tham số thử nghiệm để tìm ra mô hình có hiểu suất tốt nhất là:

Alpha = [0.1 ; 0.5 ; 1 ; 2]: Tham số điều chỉnh độ mượt (smoothing parameter). Điều này giúp ngăn chặn khả năng chia cho 0 và cải thiện hiệu suất mô hình khi có các đặc trưng hiếm gặp.

fit\_prior = [ True ; False ]: Xác định xem có học các xác suất tiên nghiệm của các lớp từ dữ liệu hay không. Nếu False, bạn có thể cung cấp các xác suất tiên nghiệm trong class\_prior.

* **Khởi tạo mô hình**

Khởi tạo mô hình Naive Bayes để chuẩn bị cho bước huấn luyện mô hình.

Khởi tạo GridSearchCV với mô hình CategoricalNB, tập hợp tham số đã được khai báo và sử dụng độ chính xác (accuracy) làm tiêu chí đánh giá.

* **Huấn luyện mô hình và dự báo nhãn của bộ dữ liệu test**

Quá trình huấn luyện mô hình sẽ sử dụng bộ dữ liệu X\_train và y\_train để xây dựng ra mô hình dự đoán.



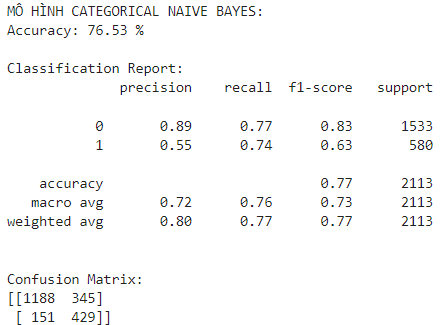
Qua quá trình huấn luyện, chúng ta đã tìm ra bộ dữ liệu phù hợp có hiểu suất mô hình tốt nhất là alpha = 0.1 và fit\_prior = True.

* **Dự báo nhãn của bộ dữ liệu test**

Sau khi đã huấn luyện mô hình trên tập dữ liệu huấn luyện, việc dự báo nhãn trên tập dữ liệu test là bước quan trọng trong quá trình đánh giá và kiểm tra hiệu suất của mô hình.

* + 1. **Đánh giá mô hình**
* **Đánh giá các chỉ số**

Dựa vào kết quả dự báo nhãn của bộ dữ liệu test (y\_prep) và nhãn của bộ dữ liệu test ( y\_test) chúng ta sẽ đáng giá mô hình qua các chỉ số sau:



hình 54. Kết quả đánh giá của mô hình Naive Bayes

Độ chính xác (Accuracy) là tỷ lệ các dự đoán đúng so với tổng số dự đoán. Trong mô hình, độ chính xác là 76.53 %, nghĩa là mô hình đã dự đoán đúng 76.62% nhãn của toàn bộ khách hàng. Đây là một tỷ lệ khá cao, cho thấy mô hình có khả năng dự đoán khá chính xác.

Ma trận nhầm lẫn có thể được giải thích như sau:

* + 1188: Số khách hàng thực tế không rời bỏ dịch vụ được dự đoán đúng là "không rời bỏ".
  + 345: Số khách hàng thực tế rời bỏ dịch vụ được dự đoán sai là "không rời bỏ".
  + 151: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán sai là "không rời bỏ".
  + 429: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán đúng là "rời bỏ".
* **Đánh giá đồ thị ROC**



hình 55. Biểu đồ ROC của mô hình Naive Bayes

Đường cong ROC có dạng cong và lên cao, điều này cho thấy mô hình có khả năng phân loại tốt giữa các lớp, với AUC = 0,83.

Tuy nhiên, đường cong vẫn chưa hoàn toàn mượt và đến gần đường chéo, chứng tỏ rằng còn có cơ hội để cải thiện chất lượng của mô hình, ví dụ như thử nghiệm các kỹ thuật xử lý dữ liệu khác hoặc các mô hình khác.

Nhìn chung, đường cong ROC cho thấy mô hình đang hoạt động tương đối tốt trong việc dự đoán khách hàng rời bỏ dịch vụ, nhưng vẫn có thể được cải thiện thêm để nâng cao độ chính xác và hiệu quả.

## Mô hình XGBoost

* + 1. **Xây dựng mô hình**
* **Chia dữ liệu**

Mô hình được chia thành 2 phần là X\_train và  y\_train  chiếm 70% bộ dữ liệu dùng để training mô hình, X\_test và y\_test chiếm 30% bộ dữ liệu dùng để test mô hình và đánh giá hiệu quả mô hình.

* **Khởi tạo mô hình**

Khởi tạo mô hình XGBoost để chuẩn bị cho bước huấn luyện mô hình

Trong mô hình XGBoost các thông số xây dựng mô hình gồm có:

* objective='binary:logistic'
* max\_depth=3:
* learning\_rate=0.1: Phương pháp tính khoảng cách giữa các điểm dữ liệu.
* n\_estimators=99
* **Huấn luyện mô hình**

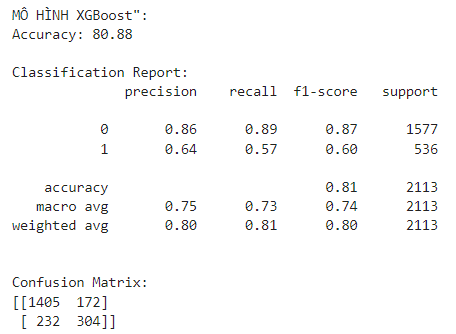
Quá trình huấn luyện mô hình sẽ sử dụng bộ dữ liệu X\_train và y\_train để xây dựng ra mô hình dự đoán.

* **Dự báo nhãn của bộ dữ liệu test**

Sau khi đã huấn luyện mô hình trên tập dữ liệu huấn luyện, việc dự báo nhãn trên tập dữ liệu test là bước quan trọng trong quá trình đánh giá và kiểm tra hiệu suất của mô hình.

* + 1. **Đánh giá mô hình**
* **Đánh giá các chỉ số**

Dựa vào kết quả dự báo nhãn của bộ dữ liệu test (y\_prep) và nhãn của bộ dữ liệu test ( y\_test) chúng ta sẽ đáng giá mô hình qua các chỉ số sau:

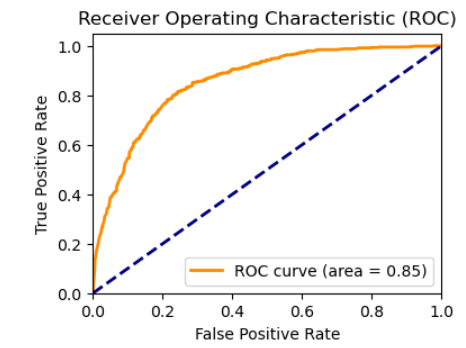


hình 56. Kết quả đánh giá của mô hình XGboost

Độ chính xác (Accuracy) là tỷ lệ các dự đoán đúng so với tổng số dự đoán. Trong mô hình, độ chính xác là 80.88 %, nghĩa là mô hình đã dự đoán đúng 80.88% nhãn của toàn bộ khách hàng. Đây là một tỷ lệ khá cao, cho thấy mô hình có khả năng dự đoán khá chính xác.

Ma trận nhầm lẫn có thể được giải thích như sau:

* 1405: Số khách hàng thực tế không rời bỏ dịch vụ được dự đoán đúng là “không rời bỏ”.
* 172: Số khách hàng thực tế không rời bỏ dịch vụ được dự đoán sai là “rời bỏ”.
* 232: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán sai là “không rời bỏ”.
* 304: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán đúng là “rời bỏ”.
* **Đánh giá đồ thị ROC**



hình 57.Biểu đồ ROC của mô hình XGboost

Với AUC = 0,85, mô hình KNN có khả năng phân biệt tương đối tốt giữa hai lớp rời bỏ và không rời bỏ. Cụ thể, mô hình có khoảng 85% cơ hội phân loại đúng các trường hợp rời bỏ và không rời bỏ.

Điều này có thể giúp doanh nghiệp dự đoán với độ chính xác khá tốt những khách hàng có nguy cơ rời bỏ dịch vụ. Từ đó, họ có thể áp dụng các chiến lược chăm sóc khách hàng phù hợp để giữ chân khách hàng.

Tuy nhiên, để đạt hiệu quả tối ưu, vẫn cần tiếp tục tối ưu hóa mô hình KNN, chẳng hạn như thử nghiệm các giá trị K khác nhau, cân bằng các mức false positive và true positive, hoặc kết hợp với các đặc trưng khác của dữ liệu.

Nhìn chung, biểu đồ ROC với AUC = 0,85 cho thấy mô hình KNN đang hoạt động khá tốt trong bài toán dự đoán rời bỏ dịch vụ. Tuy nhiên, vẫn còn nhiều cơ hội để tiếp tục cải thiện hiệu suất của mô hình.

## Mô hình Feedforward Neural Network

* + 1. **Xây dựng mô hình**
* **Chia dữ liệu**

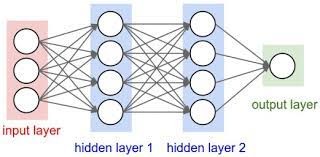
Mô hình được chia thành 2 phần là X\_train và  y\_train  chiếm 70% bộ dữ liệu dùng để training mô hình, X\_test và y\_test chiếm 30% bộ dữ liệu dùng để test mô hình và đánh giá hiệu quả mô hình.

* **Khởi tạo mô hình**

Khởi tạo mô hình FNN để chuẩn bị cho bước huấn luyện mô hình

Trong mô hình FNN chúng ta xây dựng :

* Lớp input
* Lớp ẩn 1: 100 nút noron, hàm kích hoạt ReLU.
* Lớp ẩn 2: 50 nút noron, hàm kích hoạt ReLU.
* Lớp output: giá trị output là nhị phân, hàm kích hoạt Sigmoid



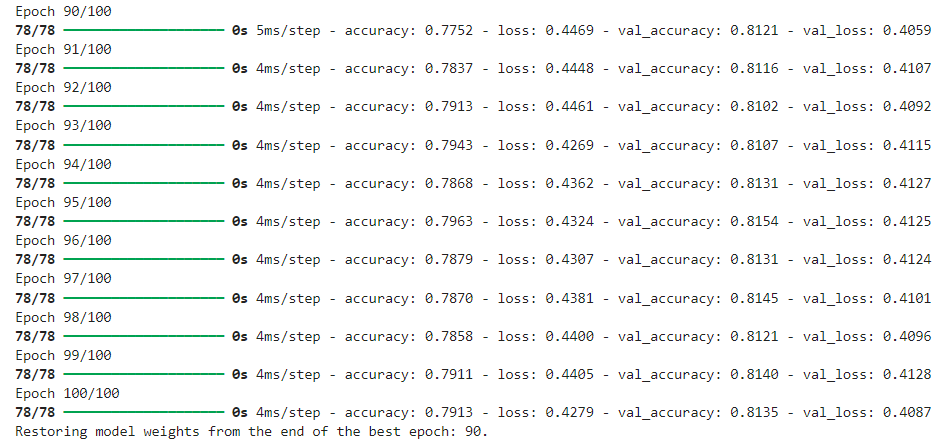
hình 58.Các lớp trong mô hình FNN

* **Huấn luyện mô hình**

Trong quá trình huấn luyện mô hình chúng ta sử dụng:

* EarlyStopping để giảm thiểu thời gian huấn luyện:
  + monitor='val\_loss' : Căn cứ vào chỉ số val\_loss để làm thước đo.
  + patience=50: Là số Epoch liên tiếp không thấy cải thiện trong giá trị val\_loss trước.
* Epoch = 100.
* Batch\_size = 64 => **Iteration** = 78.

EarlyStopping sẽ thông báo chi tiết về quá trình huấn luyện và việc dừng sớm và sau khi dừng huấn luyện, khôi phục các trọng số của mô hình tại epoch có giá trị val\_loss tốt nhất. Điều này đảm bảo rằng mô hình cuối cùng được lưu trữ là mô hình có hiệu suất tốt nhất.



hình 59.Kết quả huấn luyện mô hình FNN với 10 epoch cuối

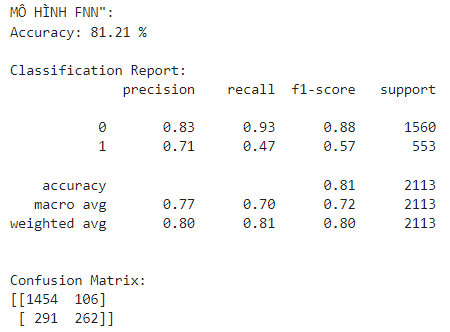
Quá trình huấn luyện trên tập dữ liệu training dừng khi hoàn thành 100 epoch và thông báo rằng epoch thứ 90 là tốt nhất.

* **Dự báo nhãn của bộ dữ liệu test**

Sau khi đã huấn luyện mô hình trên tập dữ liệu huấn luyện, việc dự báo nhãn trên tập dữ liệu test là bước quan trọng trong quá trình đánh giá và kiểm tra hiệu suất của mô hình.

* + 1. **Đánh giá mô hình**
* **Đánh giá các chỉ số**

Dựa vào kết quả dự báo nhãn của bộ dữ liệu test (y\_prep) và nhãn của bộ dữ liệu test ( y\_test) chúng ta sẽ đáng giá mô hình qua các chỉ số sau:

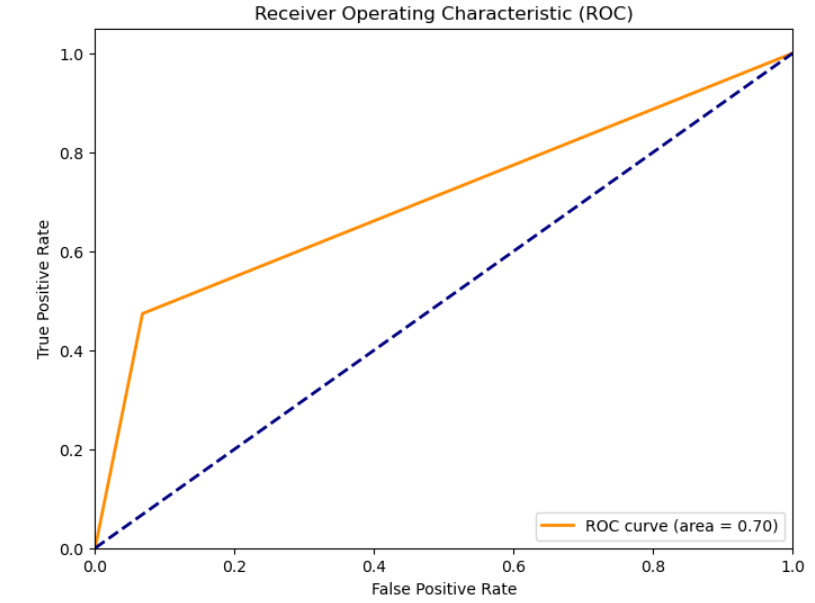


hình 60. kết quả đánh giá của mô hình FNN

Độ chính xác (Accuracy) là tỷ lệ các dự đoán đúng so với tổng số dự đoán. Trong mô hình, độ chính xác là 81,21 %, nghĩa là mô hình đã dự đoán đúng 81,21% nhãn của toàn bộ khách hàng. Đây là một tỷ lệ khá cao, cho thấy mô hình có khả năng dự đoán khá chính xác.

Ma trận nhầm lẫn có thể được giải thích như sau:

* 1454: Số khách hàng thực tế không rời bỏ dịch vụ được dự đoán đúng là “không rời bỏ”.
* 106: Số khách hàng thực tế không rời bỏ dịch vụ được dự đoán sai là “rời bỏ”.
* 291: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán sai là “không rời bỏ”.
* 262: Số khách hàng thực tế đã rời bỏ dịch vụ được dự đoán đúng là “rời bỏ”.
* **Đánh giá đồ thị ROC**



hình 61.Biểu đồ ROC của mô hình FNN

Với AUC = 0,70 mô hình FNN có khả năng phân biệt tương đối tốt giữa hai lớp rời bỏ và không rời bỏ. Cụ thể, mô hình có khoảng 70% cơ hội phân loại đúng các trường hợp rời bỏ và không rời bỏ.

Điều này có thể giúp doanh nghiệp dự đoán với độ chính xác khá tốt những khách hàng có nguy cơ rời bỏ dịch vụ. Từ đó, họ có thể áp dụng các chiến lược chăm sóc khách hàng phù hợp để giữ chân khách hàng.

Tuy nhiên, để đạt hiệu quả tối ưu, vẫn cần tiếp tục tối ưu hóa mô hình FNN, chẳng hạn như thử nghiệm các giá trị K khác nhau, cân bằng các mức false positive và true positive, hoặc kết hợp với các đặc trưng khác của dữ liệu.

Nhìn chung, biểu đồ ROC với AUC = 0,70 cho thấy mô hình FNN đang hoạt động ở mức khá trong bài toán dự đoán rời bỏ dịch vụ. Tuy nhiên, vẫn còn nhiều cơ hội để tiếp tục Tóm lại, mô hình FNN với đường ROC như trên là một kết quả khá ổn định và có thể được sử dụng, nhưng cần lưu ý đến việc chọn ngưỡng phân loại phù hợp để tối ưu hóa hiệu suất của mô hình

## Đánh giá các mô hình

* + 1. **Phân tích các chỉ số của mô hình**

Table 2.Bảng so sánh các chỉ số của các mô hình

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mô hình | Accuracy | Precision | Recall | F1-score | AUC |
| Random Forest | 79.98% | 65% | 52% | 58% | 0.85 |
| Regression Logistic | 77.9% | 61% | 51% | 56% | 0.82 |
| K-Nearest Neighbors | 79.79% | 65% | 50% | 57% | 0.83 |
| Naive Bayes | 76.53% | 55% | 74% | 63% | 0.83 |
| XGBoost | 80.88% | 64% | 57% | 60% | 0.85 |
| Feedforward Neural Network | 81.21% | 71% | 47% | 57% | 0.70 |

* **Nhận xét chung:**

**Accuracy**: Tất cả các mô hình đều đạt được độ chính xác khá cao, dao động từ 76.53% đến 81.21%. Điều này cho thấy các mô hình đều có khả năng dự đoán chính xác một phần đáng kể các trường hợp.

**Precision**: Mô hình Feedforward Neural Network có Precision cao nhất (71%), nghĩa là trong số những khách hàng mà mô hình này dự đoán sẽ rời bỏ, có tỷ lệ lớn thực sự đã rời bỏ. Tuy nhiên, các mô hình khác cũng có Precision khá ổn định, dao động từ 55% đến 65%.

**Recall:** Mô hình Naive Bayes có Recall cao nhất (74%), điều này cho thấy mô hình này có khả năng phát hiện tốt nhất những khách hàng thực sự rời bỏ dịch vụ. Tuy nhiên, Precision của mô hình này lại tương đối thấp.

**F1-score:** F1-score là một chỉ số cân bằng giữa Precision và Recall. Mô hình Random Forest và XGBoost có F1-score tương đối cao, cho thấy sự cân bằng tốt giữa việc giảm thiểu các dự đoán sai lệch và phát hiện đúng các khách hàng rời bỏ.

**AUC:** Tất cả các mô hình đều có AUC khá cao (trên 0.8), cho thấy khả năng phân biệt giữa khách hàng sẽ rời bỏ và khách hàng sẽ ở lại là tốt.

* **So sánh từng mô hình:**
* **Random Forest:**
* Ưu điểm: Có Accuracy và Precision khá cao, cho thấy mô hình dự đoán chính xác những khách hàng sẽ rời đi.
* Nhược điểm: Recall hơi thấp so với một số mô hình khác.
* **Regression Logistic:**
* Ưu điểm: Có AUC khá cao, cho thấy khả năng phân biệt tốt.
* Nhược điểm: Accuracy, Precision và Recall đều thấp hơn so với các mô hình khác.
* **K-Nearest Neighbors:**
* Ưu điểm: Có Accuracy và Precision khá cao.
* Nhược điểm: Recall thấp.
* **Naive Bayes:**
* Ưu điểm: Có Recall cao nhất, nghĩa là mô hình rất nhạy bén trong việc phát hiện ra những khách hàng sẽ rời đi.
* Nhược điểm: Precision và Accuracy lại khá thấp, điều này có nghĩa là mô hình cũng dự đoán nhiều khách hàng ở lại là sẽ rời đi.
* **XGBoost:**
* Ưu điểm: Có Accuracy, Precision, Recall và F1-score khá cao, đặc biệt là Accuracy và AUC. Điều này cho thấy XGBoost có khả năng dự đoán chính xác tổng thể và phân biệt tốt giữa hai lớp.
* Nhược điểm: So với Naive Bayes, Recall có phần thấp hơn.
* **Feedforward Neural Network:**
* Ưu điểm: Có Accuracy cao nhất, cho thấy mô hình có khả năng dự đoán tổng thể tốt.
* Nhược điểm: Recall và F1-score lại thấp nhất, điều này có nghĩa là mô hình không thực sự giỏi trong việc xác định chính xác những khách hàng sẽ rời đi.
  + 1. **Kết luận**

Dựa trên các phân tích trên, mô hình XGBoost là mô hình có hiệu suất tổng thể tốt nhất, đây được xem là lựa chọn tối ưu nhất cho mục đích ngăn ngừa khách hàng rời bỏ dịch vụ. Mô hình này có khả năng dự đoán chính xác cả những khách hàng sẽ ở lại và những khách hàng sẽ rời đi, đồng thời phân biệt tốt giữa hai lớp.

Ngoài ra, chúng ta cũng có thể cân nhắc sử dụng mô hình Naive Bayes cho mục tiêu chính là xác định những khách hàng có khả năng cao rời bỏ dịch vụ để tập trung nguồn lực vào nhóm khách hàng này.

## Chiến lược giữ chân khách hàng

Trong 6 mô hình được xây dựng, mỗi mô hình đều cho thấy tầm quan trọng của các yếu tố ảnh hưởng, trong đó có 3 mô hình cho thấy mức độ xếp hàng đầu là của yếu tố Contract. Thứ tự xếp hạng của các yếu tố còn lại có sự khác biệt giữa các mô hình, điều này xuất phát từ cách tiếp cận và thế mạnh riêng biệt của từng mô hình, dẫn đến kết quả đánh giá tầm quan trọng của các biến (features) khác nhau.

Tuy nhiên, để đảm bảo hiệu quả dự đoán và đề xuất chiến lược giữ chân khách hàng hiệu quả, việc lựa chọn mô hình có độ chính xác cao là vô cùng quan trọng. Theo phân tích, mô hình XGboost được coi như lựa chọn tối ưu cho mục đích ngăn ngừa khách hàng rời bỏ dịch vụ nhờ sở hữu các ưu điểm vượt trội như: độ chính xác cao, độ nhạy cao, điểm F1 cao và AUC cao.

Vì vậy ta xem xét mức độ quan trọng của các yếu tố lần lượt như trong hình :

****

Dựa trên kết quả dữ liệu đã phân tích và qua việc trực quan hóa dữ liệu, ta có thể đưa ra các giải pháp nhằm tối ưu hóa hiệu quả của doanh nghiệp:

* Tập trung vào dịch vụ Contract: Dịch vụ Contract là yếu tố cốt lõi quyết định việc khách hàng có tiếp tục sử dụng dịch vụ hay không. Đặc biệt, cần ưu tiên cải thiện trải nghiệm cho nhóm khách hàng sử dụng hợp đồng "Month to month" vì đây là nhóm có tỷ lệ rời bỏ cao nhất.
* Nâng cao chất lượng dịch vụ OnlineSecurity và TechSupport: Hai dịch vụ này cũng đóng vai trò quan trọng trong việc giữ chân khách hàng. Để giảm tỷ lệ rời bỏ, chúng ta cần tập trung vào nhóm khách hàng hiện tại chưa sử dụng hai dịch vụ này.
* Cải thiện liên tục các dịch vụ InternetService, tenure và PaymentMethod:
  + InternetService: Tập trung vào nâng cao chất lượng dịch vụ Fiber optic, loại hình được nhiều khách hàng lựa chọn.
  + Tenure: Giữ chân khách hàng bằng cách kéo dài thời gian sử dụng dịch vụ, đặc biệt là với nhóm khách hàng có thời hạn hợp đồng ngắn dưới 1 năm.
  + PaymentMethod: Đa dạng hóa phương thức thanh toán, ưu tiên đẩy mạnh sử dụng phương thức Electronic Check."
* Để giữ chân khách hàng không rời bỏ dịch vụ, ngoài 6 yếu tố ảnh hưởng quan trọng trên đối với các mô hình, doanh nghiệp cần xây dựng một chiến lược toàn diện, tập trung vào việc nâng cao trải nghiệm khách hàng và tạo ra mối quan hệ lâu dài.

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

* **Đạt được:**
* Đề tài đã tìm hiểu các kỹ thuật học máy như Logistic Regression, Random Forests, KNN, Naive Bayes, XGBoost, FNN từ đó ứng dụng kỹ thuật này vào xây dựng mô hình khai phá dữ liệu nhằm dự đoán khách hàng rời bỏ dịch vụ viễn thông.
* Đánh giá được hiệu quả của mô hình dự đoán thông qua các chỉ số như độ chính xác, độ nhạy, độ đặc trưng, v.v. Các mô hình học máy đều cho kết quả dự đoán khá tốt về khả năng khách hàng rời bỏ dịch vụ, với độ chính xác lên đến 80,88% đối với mô hình XGBoost.
* Các đặc trưng quan trọng ảnh hưởng đến quyết định rời bỏ dịch vụ của khách hàng bao gồm: Contract, OnlineService, TechSupport, InternetService và PaymentMethod.
* Kết quả phân tích này cũng góp phần hỗ trợ các nhà cung cấp dịch vụ viễn thông chủ động nắm bắt và đáp ứng nhu cầu của khách hàng, từ đó giảm tỷ lệ khách hàng rời bỏ dịch vụ.
* **Hạn chế:**
* Mô hình dự đoán vẫn chưa đạt được độ chính xác 100%, vẫn còn một số trường hợp dự đoán không chính xác, do đó cần tiếp tục cải thiện các mô hình.
* Chưa thể phân tích sâu các yếu tố ảnh hưởng đến quyết định của khách hàng rời bỏ dịch vụ.
* **Hướng phát triển:**
* Tiếp tục nghiên cứu và cải tiến mô hình học máy để nâng cao độ chính xác và hiệu quả dự đoán.
* Ứng dụng nhiều kỹ thuật học máy khác nhau để khai khác tốt hơn về dữ liệu về vấn đề rời bỏ dịch vụ của khách hàng giúp có cái nhìn chi tiết, đa chiều hơn.
* Mở rộng phạm vi nghiên cứu, thu thập thêm dữ liệu về hành vi và nhu cầu của khách hàng để phân tích sâu hơn các yếu tố ảnh hưởng.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | "Giới thiệu chung Công ty Công nghệ thông tin VNPT," 1 3 2018. [Online]. Available: https://vnptit.vn/gioi-thieu-chung. |
| [2] | VNPT-IT, "Tầm nhìn sứ mệnh của VNPT," 2018. [Online]. Available: https://vnptit.vn/tam-nhin-su-menh. |
| [3] | Softech, "Tổng quan về Công ty cổ phần Softech," 8 11 2000. [Online]. Available: https://softech.vn/company/gioi-thieu/chi-tiet/15019/gioi-thieu-tong-quan. |
| [4] | Công ty CP Softech, 2000. [Online]. Available: https://softech.vn/company/gioi-thieu/chi-tiet/15027/co-cau-to-chuc. |
| [5] | "RandomForest, 2024. [Online]. Available: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/. |
| [6] | "Trí tuệ nhân tạo," 2024. [Online]. Available: https://trituenhantao.io/machine-learning-co-ban/bai-6-logistic-regression-hoi-quy-logistic/. |
| [7] | ICHI.PRO, "Hướng dẫn từng bước về thuật toán KNN: Hiểu các nguyên tắc cơ bản và ứng dụng," 2020. [Online]. Available: https://ichi.pro/vi/huong-dan-tung-buoc-ve-thuat-toan-knn-hieu-cac-nguyen-tac-co-ban-va-ung-dung-76591305046436. |
| [8] | N. T. Hop, "Thuật toán phân lớp Naive Bayes," 2019. [Online]. Available: https://viblo.asia/p/thuat-toan-phan-lop-naive-bayes-924lJWPm5PM. |
| [9] | B. T. Tung, 28 5 2021. [Online]. Available: https://viblo.asia/p/gradient-boosting-tat-tan-tat-ve-thuat-toan-manh-me-nhat-trong-machine-learning-YWOZrN7vZQ0. |
| [10] | Wikipedia, "Mạng thần kinh truyền thẳng," 18 4 2022. [Online]. Available: https://vi.wikipedia.org/wiki/M%E1%BA%A1ng\_th%E1%BA%A7n\_kinh\_truy%E1%BB%81n\_th%E1%BA%B3ng. |
| [11] | Tapit, "Trí tuệ nhân tạo: Các phương pháp đánh giá một mô hình phân loại," 16 6 2021. [Online]. Available: https://tapit.vn/cac-phuong-phap-danh-gia-mot-mo-hinh-phan-loai/. |
| [12] | Pum, "Jupyter Notebook là gì? Hướng dẫn cài đặt và sử dụng Jupyter Notebook," 29 7 2022. [Online]. Available: https://200lab.io/blog/jupyter-notebook-la-gi/. |
| [13] | L. H. Hạnh, "Power BI là gì? Tìm hiểu cách sử dụng Power BI cho doanh nghiệp," 15 3 2024. [Online]. Available: https://base.vn/blog/power-bi-la-gi/. |
| [14] | C. n. t. t. VNPT, "Giới thiệu chung về công ty Công nghệ thông tin VNPT," 2024. |
| [15] | C. t. c. p. Softech, "Giới thiệu chung công ty cổ phần Softech," 2024. |

# CHECK LIST CỦA BÁO CÁO

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Nội dung công việc** | **Có** | **Không** | **Ghi chú** |
| 1 | Báo cáo được trình bày (định dạng) đúng với yêu cầu. |  |  |  |
| 2 | Báo cáo có số lượng trang đáp ứng đúng yêu cầu (30-50 trang) |  |  |  |
| 3 | Báo cáo trình bày được phần mở đầu bao gồm: Mục tiêu, Phạm vi và đối tượng, kết cấu … |  |  |  |
| 4 | Báo cáo trình bày về công ty, vị trí việc làm (công việc đó làm gì, kiến thức và kỹ năng cần thiết là gì, con đường phát triển sự nghiệp (career path)), cơ sở lý thuyết phù hợp với nội dung của đề tài (Tối đa 10-12 trang) |  |  |  |
| 5 | Báo cáo có sản phẩm cụ thể phù hợp với mục tiêu đặt ra của đề tài |  |  |  |
| 6 | Báo cáo có phần kết luận và hướng phát triển của đề tài |  |  |  |

# PHỤ LỤC

**Phần trình bày chi tiết code của các mô hình :**

https://drive.google.com/drive/folders/1nh9iE97PcJlpSrQNIFaE3afvQtg3qnP0?usp=sharing