ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍNH MINH

TRƯỜNG ĐẠI HỌC BÁCH KHOA

BÁO CÁO BÀI TẬP LỚN MÔN HỌC: HỌC MÁY

Phân Tích và Dự Đoán Churn Khách Hàng trong Doanh Nghiệp SaaS bằng Machine Learning

Mã học phần: CO3117

GVHD: Trương Vĩnh Lân

Lớp: DT01_251

Sinh viên thực hiên	MSSV
Phan Thị Thuỳ Anh	2433103
Nguyễn Thị Hồng Phúc	2433190

TP. Hồ Chí Minh, tháng 12 năm 2025

MỤC LỤC

Bảng Phân Công Công Việc

STT	Hạng mục công việc	Mô tả	Người phụ trách	Tỷ lệ (%)
1	Thu thập dữ liệu	Tải dữ liệu, kiểm tra cấu trúc, loại bỏ lỗi	Phan Thị Thuỳ Anh	8%
2	Tiền xử lý dữ liệu	Làm sạch, mã hóa biến, chuẩn hóa dữ liệu	Nguyễn Thị Hồng Phúc	8%
3	Phân tích mô tả (EDA)	Biểu đồ, thống kê, nhận diện pattern	Phan Thị Thuỳ Anh	8%
4	Xây dựng mô hình baseline	Logistic Regression + đánh giá sơ bộ	Nguyễn Thị Hồng Phúc	8%
5	Tối ưu mô hình	Feature engineering v2, tuning, thử mô hình nâng cao	Phan Thị Thuỳ Anh	14%
6	Đánh giá mô hình	So sánh metrics, kiểm tra overfitting	Nguyễn Thị Hồng Phúc	14%
7	Viết Chương 4: Phân tích dữ liệu	Mô tả dữ liệu, biểu đồ, nhận xét	Phan Thị Thuỳ Anh	10%
8	Viết Chương 5: Mô hình & kết quả	Mô tả mô hình, thuật toán, pipeline, kết quả	Nguyễn Thị Hồng Phúc	10%
9	Viết Chương 6: Ứng dụng & ROI	Roadmap, deployment, business impact	Phan Thị Thuỳ Anh	9%
10	Viết Chương 7: Kết luận	Tổng kết kết quả nghiên cứu	Nguyễn Thị Hồng Phúc	6%
11	Chuẩn hoá tài liệu tham khảo	Định dạng reference + chèn in-text citations	Phan Thị Thuỳ Anh	4%
12	Hiệu đính & rà soát toàn báo cáo	Kiểm tra nội dung, format, số liệu, biểu đồ	Phan Thị Thuỳ Anh Nguyễn Thị Hồng Phúc	5%
14	Viết README dự án	Hướng dẫn, mô tả, cách chạy code, kết quả	Nguyễn Thị Hồng Phúc	4%
15	Push toàn bộ dự án lên Git/GitHub	Đưa code lên repo	Nguyễn Thị Hồng Phúc	0%

Tổng tỷ lệ đóng góp:

Thành viên	Tỷ lệ đóng góp
Phan Thị Thuỳ Anh	50%
Nguyễn Thị Hồng Phúc	50%

Link Colab và Github

Colab: https://colab.research.google.com/drive/1s0nkWXOEhT3G00c8OQJ18iXZly6v0kKF?usp=sharing

Github: https://github.com/felicianguyenmt/saas_assignment.git

1. TỔNG QUAN DỰ ÁN

1.1. Mục Tiêu

Dự án hướng đến việc xây dựng một mô hình Machine Learning có khả năng **dự đoán xác suất khách hàng rời bỏ dịch vụ (churn)** trong môi trường doanh nghiệp SaaS (Software-as-a-Service). Việc triển khai mô hình không chỉ giúp tự động hóa quá trình nhận diện rủi ro, mà còn đóng vai trò như một công cụ hỗ trợ ra quyết định cho các bộ phận kinh doanh và chăm sóc khách hàng.

Cụ thể, mô hình được phát triển nhằm đạt các mục tiêu sau:

- **Phát hiện sớm** những khách hàng có khả năng rời bỏ cao để doanh nghiệp có thể chủ động áp dụng các biện pháp can thiệp.
- **Phân tích và định lượng** mức độ ảnh hưởng của từng yếu tố (hành vi sử dụng, mức độ tương tác, thông tin tài khoản, v.v.) đến quyết định rời bỏ của khách hàng.
- **Hỗ trợ đội ngũ kinh doanh và marketing** xây dựng chiến lược chăm sóc và giữ chân khách hàng dựa trên dữ liệu, giúp tối ưu chi phí và nâng cao hiệu quả vận hành.

Mục tiêu cuối cùng của dự án là tạo ra một mô hình dự đoán churn vừa **chính xác**, vừa **dễ triển khai thực tế**, qua đó gia tăng tỷ lệ duy trì khách hàng và ổn định nguồn doanh thu dài hạn.

1.2. Nguồn Dữ Liệu

Dataset: **RavenStack SaaS Subscription and Churn Analytics**

Nguồn: Kaggle (rivalytics/saas-subscription-and-churn-analytics-dataset)

Dự án sử dụng bộ dữ liệu **RavenStack SaaS Subscription and Churn Analytics**, được công bố trên nền tảng Kaggle tại kho dữ liệu *"rivalytics/saas-subscription-and-churn-analytics-dataset"*. Đây là bộ dữ liệu mô phỏng hoạt động kinh doanh của một công ty SaaS, bao gồm thông tin về khách hàng, hành vi sử dụng sản phẩm và trạng thái rời bỏ dịch vụ (churn).

Bộ dữ liệu được cấu trúc thành **5 bảng**, mỗi bảng phản ánh một khía cạnh quan trọng trong vòng đời khách hàng:

- **ravenstack_accounts.csv**
  Chứa thông tin cơ bản của khách hàng, bao gồm ngày tạo tài khoản, quốc gia, ngành nghề và các thuộc tính nhân khẩu học liên quan. Đây là nguồn dữ liệu nền tảng giúp mô hình hiểu được hồ sơ (profile) của từng khách hàng.
- **ravenstack_subscriptions.csv**
  Ghi nhận các gói đăng ký (plans), loại hình thanh toán, chu kỳ billing và tình trạng gia hạn. Những thông tin này đóng vai trò quan trọng trong việc xác định mức độ cam kết của khách hàng.

- **ravenstack_feature_usage.csv**
  Ghi lại dữ liệu hành vi thực tế của người dùng khi tương tác với sản phẩm: số lần đăng nhập, tần suất sử dụng từng tính năng, thời gian sử dụng,… Đây là nhóm chỉ số then chốt để nhận diện dấu hiệu giảm tương tác—một trong những nguyên nhân dẫn đến churn.
- **ravenstack_support_tickets.csv**
  Lưu trữ lịch sử các yêu cầu hỗ trợ: số lượng ticket, mức độ nghiêm trọng và thời gian phản hồi. Chất lượng hỗ trợ khách hàng được xem là một yếu tố quan trọng ảnh hưởng đến sự hài lòng và khả năng duy trì dịch vụ.
- **ravenstack_churn_events.csv**
  Ghi nhận các sự kiện rời bỏ dịch vụ, hủy đăng ký và tình trạng hoàn tiền. Đây là bảng mục tiêu (label) quan trọng dùng để xây dựng mô hình phân loại churn.

Bộ dữ liệu đa chiều này cho phép mô hình phân tích churn từ nhiều góc độ: đặc điểm tài khoản, mức độ sử dụng sản phẩm, trải nghiệm hỗ trợ và thông tin về chu kỳ dịch vụ. Nhờ đó, kết quả dự đoán có thể đạt độ chính xác cao và mang giá trị thực tiễn trong doanh nghiệp SaaS.

1.3. Biến Mục Tiêu

churn_flag*: Biến nhị phân (0/1) từ bảng accounts*

Biến mục tiêu của dự án là **churn_flag**, được trích xuất từ bảng *ravenstack_accounts*. Đây là biến nhị phân thể hiện trạng thái duy trì dịch vụ của khách hàng:

0 – Không churn: Khách hàng tiếp tục sử dụng dịch vụ và duy trì gói đăng ký.
1 – Churn: Khách hàng đã rời bỏ dịch vụ hoặc hủy gói đăng ký trong kỳ quan sát.

Biến này đóng vai trò trọng tâm trong mô hình dự báo, cho phép hệ thống phân loại khách hàng theo khả năng rời bỏ và hỗ trợ triển khai các chiến lược giữ chân phù hợp.

2. PHÂN TÍCH DỮ LIỆU KHÁM PHÁ (EDA)

2.1. Tổng Quan Dữ Liệu

2.1.1. Kích Thước Dataset

Bộ dữ liệu sau khi tổng hợp và xử lý sơ bộ có các đặc điểm chính:

- **Tổng số khách hàng:** 500
  Đại diện cho các tài khoản SaaS đã được ghi nhận trong toàn bộ hệ thống.
- **Số lượng đặc trưng (features):** 40
  Bao gồm thông tin tài khoản, hành vi sử dụng tính năng, lịch sử hỗ trợ, thông tin gói đăng ký và các biến dẫn xuất phục vụ mô hình.
- **Training samples:** 400
  Chiếm 80% dữ liệu, được sử dụng để huấn luyện mô hình dự đoán.

- **Test samples:** 100
  Chiếm 20% dữ liệu, dùng để đánh giá khả năng tổng quát hóa của mô hình trên dữ liệu chưa từng thấy.

Phần tổng quan này giúp xác định phạm vi dữ liệu và đảm bảo nền tảng phù hợp cho các bước phân tích sâu hơn.

2.1.2. Cấu Trúc Dataset

Bộ dữ liệu bao gồm 5 bảng chính, phản ánh đầy đủ thông tin về tài khoản, đăng ký dịch vụ, hành vi sử dụng và sự kiện churn.

Accounts — (n_samples, 9 features)

Chứa thông tin cốt lõi về từng tài khoản khách hàng:

```
account_id, account_name, industry, country
signup_date, referral_source, plan_tier, seats, is_trial
churn_flag — biến mục tiêu (0/1)
```

Subscriptions — (n_samples, ~8 features)

Ghi nhận thông tin về trạng thái gói đăng ký:

```
subscription_id, account_id, mrr_amount, arr_amount
seats, upgrade_flag, downgrade_flag, churn_flag, auto_renew_flag
```

Feature Usage — (n_samples, ~6 features)

Dữ liệu hành vi sử dụng theo từng gói:

```
usage_id, subscription_id, usage_count, usage_duration_secs
error_count, is_beta_feature
```

Support Tickets — (n_samples, ~7 features)

Phản ánh mức độ tương tác và mức độ hài lòng khi hỗ trợ:

```
ticket_id, account_id, priority, resolution_time_hours
first_response_time_minutes, satisfaction_score, escalation_flag
```

Churn Events — (n_samples, ~5 features)

Ghi nhận sự kiện churn và các yếu tố dẫn tới churn:

```
churn_event_id, account_id, refund_amount_usd
preceding_upgrade_flag, preceding_downgrade_flag, is_reactivation
```

Cấu trúc đa bảng giúp phân tích churn toàn diện từ nhân khẩu học, hành vi sử dụng đến lịch sử hỗ trợ.

2.2. Phân Bố Target Variable

Trong toàn bộ dataset, biến mục tiêu **churn_flag** có phân phối như sau:

- Không churn (0): 78.0%
- Churn (1): 22.0%
- Tổng số khách hàng đã churn: 110 / 500

Phân phối trong tập huấn luyện (Training Set)

- Churn: 88 mẫu
- Non-churn: 312 mẫu

Phân phối trong tập kiểm tra (Test Set)

- Churn: 22 mẫu
- Non-churn: 78 mẫu

Nhận xét

- Dataset **mất cân bằng nhẹ**, với churn chiếm tỷ lệ thấp hơn đáng kể.
- Khi đánh giá mô hình, không thể chỉ dựa vào **Accuracy**, mà cần sử dụng các chỉ số phù hợp cho dữ liệu mất cân bằng như **Precision, Recall, F1-score và ROC-AUC**.
- Dataset được chia theo phương pháp **stratified split**, giúp giữ nguyên tỷ lệ churn giữa các tập và tránh sai lệch phân phối khi huấn luyện.

2.3. Các Insights Chính

Những phân tích khám phá ban đầu (EDA) cho thấy nhiều yếu tố có liên quan đáng kể đến churn. Dưới đây là các nhóm insight chính theo từng nguồn dữ liệu.

2.3.1. Insights từ Subscription Features

- **MRR/ARR thấp → khả năng churn cao:**
  Khách hàng chi tiêu thấp thường có mức độ cam kết thấp hơn và dễ rời bỏ dịch vụ.
- **Downgrade liên quan mạnh đến churn:**
  Số lần downgrade có tương quan dương với churn, cho thấy sự giảm giá trị cảm nhận trước khi rời bỏ.
- **Tắt auto-renewal là tín hiệu cảnh báo sớm:**
  Khách hàng tắt chức năng gia hạn tự động thường đã có ý định rời bỏ.

2.3.2. Insights từ Feature Usage

- **Usage intensity thấp → churn cao:**
  Mức độ sử dụng thấp phản ánh sản phẩm không đủ giá trị hoặc thiếu thói quen sử dụng.

- **Error rate cao làm gia tăng churn:**
  Khách hàng gặp nhiều lỗi có trải nghiệm tiêu cực, dễ dẫn đến hủy bỏ dịch vụ.
- **Ưu tiên beta features → churn thấp:**
  Việc tương tác với tính năng beta thể hiện mức độ quan tâm cao và khả năng gắn bó.

2.3.3. Insights từ Support Tickets

- **Khối lượng ticket quá cao hoặc quá thấp đều tiềm ẩn rủi ro:**
  - Nhiều ticket: khách hàng gặp nhiều vấn đề
  - Ít ticket: khách hàng không sử dụng nhiều hoặc bỏ qua vấn đề → dễ rời bỏ
- **Thời gian xử lý dài làm giảm satisfaction và tăng churn:**
  Resolution time cao là yếu tố tiêu cực rõ rệt.
- **Satisfaction score tương quan nghịch mạnh với churn:**
  Điểm hài lòng thấp → xác suất churn cao.
- **Escalated tickets → nguy cơ churn lớn:**
  Những vấn đề phải escalate thường là vấn đề nghiêm trọng.

2.3.4. Insights từ Churn Events

- **Khách hàng từng churn trước đó dễ churn lại:**
  Lịch sử churn phản ánh mức độ gắn kết thấp và hành vi không ổn định.
- **Refund amount phản ánh mức độ bất mãn:**
  Hoàn tiền càng nhiều → khả năng rời bỏ càng cao.
- **Reactivated users có hành vi khác biệt:**
  Nhóm này nhìn chung nhạy cảm hơn với lỗi và giá trị sản phẩm, nên cần chiến lược chăm sóc chuyên biệt.

3. THIẾT KẾ PIPELINE

3.1. Tổng Quan Pipeline

Quy trình xây dựng mô hình được thiết kế theo kiến trúc pipeline tuần tự, đảm bảo khả năng tái sử dụng, mở rộng và triển khai thực tế:

```
[Raw Data]
      → [Feature Engineering]
            → [Preprocessing]
                  → [Model Training]
                        → [Evaluation]
                              → [Prediction]
```

Pipeline này cho phép xử lý dữ liệu từ nhiều nguồn khác nhau, trích xuất các đặc trưng quan trọng, chuẩn hóa đầu vào và huấn luyện mô hình một cách nhất quán.

3.2. Chi Tiết Từng Bước

Bước 1: Load và Merge Data

Dữ liệu từ Kaggle bao gồm **5 bảng** được tải vào môi trường phân tích:

```
ravenstack_accounts.csv

ravenstack_subscriptions.csv

ravenstack_feature_usage.csv

ravenstack_support_tickets.csv

ravenstack_churn_events.csv
```

Các bảng được merge dựa trên **account_id** hoặc **subscription_id** tùy theo cấu trúc.

Base features từ bảng Accounts:

```
X_base = accounts[columns]
y = accounts['churn_flag']
```

Bảng **accounts** đóng vai trò bảng gốc, chứa các thông tin nhân khẩu, gói dịch vụ và biến mục tiêu.

Bước 2: Feature Engineering

Feature engineering là bước quan trọng nhất, giúp tổng hợp dữ liệu từ nhiều bảng thành bộ đặc trưng duy nhất cho mỗi khách hàng. Tổng cộng tạo ra 40 đặc trưng cuối cùng.

B2-1. Subscription Features

(*Trích xuất từ bảng subscriptions, gom theo account_id*)

Nhóm chỉ số	Mô tả
num_subscriptions	Tổng số subscription theo khách hàng
total_mrr, avg_mrr, max_mrr	Doanh thu MRR tổng – trung bình – cao nhất
total_arr, avg_arr	Doanh thu ARR tổng – trung bình
total_seats, avg_seats	Tổng/Trung bình số seats người dùng
num_upgrades, num_downgrades	Tần suất nâng cấp / hạ cấp gói dịch vụ
num_churned_subs	Số subscription từng bị churn

auto_renew_pct	Tỷ lệ subscription bật gia hạn tự động

Các đặc trưng này giúp mô hình hiểu **giá trị khách hàng**, **mức độ cam kết**, và **xu hướng thay đổi gói**, tất cả đều liên quan đến khả năng churn.

B2-2. Feature Usage Features

(*Gom theo subscription_id, sau đó theo account_id*)

Nhóm chỉ số	Mô tả
total_usage_events, avg_usage_events_per_sub	Số lần sử dụng tính năng
total_usage_count, avg_usage_count_per_sub	Tổng số lượt sử dụng thực tế
total_usage_duration, avg_usage_duration	Thời gian sử dụng (seconds)
avg_error_rate	Tỷ lệ lỗi trung bình
beta_feature_engagement	Mức độ sử dụng tính năng beta

Đây là các đặc trưng quan trọng phản ánh mức độ gắn kết và trải nghiệm sản phẩm.

B2-1. Support Ticket Features

(Gom theo account_id)

Nhóm chỉ số	Mô tả
num_support_tickets	Số lượng ticket được tạo
urgent_tickets_count	Số ticket ưu tiên cao
avg_resolution_time, max_resolution_time	Tốc độ giải quyết
avg_first_response_time	Tốc độ phản hồi ban đầu
avg_satisfaction_score	Mức độ hài lòng trung bình
escalation_rate	Tỷ lệ ticket bị chuyển cấp (escalate)

Bộ chỉ số này đánh giá **chất lượng dịch vụ** và **mức độ hài lòng của khách hàng**.

B2-4. Churn Event Features

(*Gom theo account_id từ bảng churn_events*)

Nhóm chỉ số	Mô tả

num_churn_events	Số lần từng churn trước đó
total_refunds, avg_refund	Tổng và trung bình tiền hoàn trả
upgrades_before_churn, downgrades_before_churn	Dấu hiệu thay đổi gói trước khi churn
reactivation_count	Số lần quay lại (reactivate)

Các đặc trưng này nắm bắt **hành vi churn lịch sử** – yếu tố dự báo mạnh mẽ trong nhiều mô hình churn.

Tổng số đặc trưng cuối cùng: 40 features

Bao gồm:

- Features gốc (từ bảng Accounts)
- Features tổng hợp từ Subscriptions
- Usage features
- Support ticket features
- Churn event features

Các đặc trưng được merge thành một bảng duy nhất ở cấp độ **account_id**, sẵn sàng cho bước tiền xử lý và mô hình hóa.

<u>Bước 3</u>: Data Preprocessing

Quá trình tiền xử lý được thiết kế nhằm chuẩn hóa dữ liệu từ nhiều bảng khác nhau, xử lý giá trị khuyết và chuẩn bị đầu vào ổn định cho mô hình. Các bước thực hiện bao gồm hợp nhất dữ liệu, làm sạch, mã hóa biến phân loại, tạo biến thời gian và chuẩn hóa.

B3-1. Merge Features

Tất cả dữ liệu đặc trưng được tạo từ bước Feature Engineering được merge theo account_id để tạo thành một bảng duy nhất cho mỗi khách hàng:

```
X_merged = (
    X_base
        .merge(sub_features, on="account_id", how="left")
        .merge(usage_features, on="account_id", how="left")
        .merge(support_features, on="account_id", how="left")
        .merge(churn_features, on="account_id", how="left")
)
```

Việc merge theo chiều ngang giúp mô hình có đầy đủ thông tin về hành vi mua hàng, sử dụng sản phẩm và tương tác hỗ trợ của khách hàng.

B3-2. Handle Missing Values

Chiến lược xử lý: điền giá trị 0 cho missing values.

Lý do lựa chọn:

- Các cột bị thiếu chủ yếu thuộc nhóm:
  số lượng ticket, số lần upgrade, số sự kiện sử dụng, số refund,...
- Trong ngữ cảnh này, **missing = không có hoạt động** → gán 0 là hợp lý.
- Tránh việc dùng mean/median sai bản chất, vì các đặc trưng dạng đếm (count) không có nghĩa khi lấy trung bình.

Ví dụ:

Feature	Khi missing → Interpret as
num_support_tickets	0 (khách hàng chưa từng gửi ticket)
num_upgrades	0 (không nâng cấp gói)
total_usage_events	0 (không có lịch sử sử dụng)

B3-3. Feature Transformation

(1) Date Features

Biến signup_date được chuyển đổi thành đặc trưng số:

- days_since_signup = số ngày kể từ khi đăng ký đến thời điểm trích xuất dataset.

Điều này giúp mô hình hiểu khách hàng ở giai đoạn nào trong vòng đời (lifecycle stage).

(2) Categorical Encoding

Các biến phân loại được mã hóa bằng **Label Encoding**, do số lượng giá trị vừa phải và mô hình tree-based (Random Forest, Gradient Boosting) có thể xử lý tốt dạng nhãn.

Các cột được mã hóa:

- industry – Ngành nghề
- country – Quốc gia
- referral_source – Nguồn giới thiệu
- plan_tier – Loại gói đăng ký
- is_trial – Có đang dùng thử hay không

Label Encoding giúp mô hình:

- xử lý dữ liệu nhanh hơn
- không tăng số chiều như One-Hot Encoding

- tránh sparsity

B3-4. Train-Test Split

Chia dữ liệu theo tỷ lệ:

- **Test size:** 20%
- **Random state:** 42 (đảm bảo tái lập kết quả)

Điểm quan trọng:

Sử dụng *Stratified Split*

→ Giữ nguyên tỷ lệ churn giữa train và test
→ Tránh tình trạng test set thiếu các mẫu churn (vốn ít hơn)

Tập dữ liệu	Non-churn	Churn
Train	312	88
Test	78	22

Điều này đảm bảo phân phối mẫu mục tiêu đồng nhất, giúp đánh giá mô hình chính xác hơn.

B3-5. Feature Scaling

Phương pháp: **StandardScaler**

- Chuẩn hóa dữ liệu về phân phối chuẩn (chuẩn hóa về mean=0, std=1)
- Giúp các mô hình tuyến tính hội tụ nhanh và tránh bias bởi các feature có scale lớn.

Áp dụng cho:

- Logistic Regression (bắt buộc)

Không áp dụng cho:

- Random Forest
- Gradient Boosting / XGBoost

Vì đây là các mô hình dựa trên cây quyết định **không bị ảnh hưởng bởi thang đo của dữ liệu**.

4. CÁC THÍ NGHIỆM ĐÃ THỰC HIỆN

Mục tiêu của phần thí nghiệm là đánh giá hiệu năng của các thuật toán tuyển chọn trên bài toán dự đoán churn, so sánh ưu/nhược điểm thực tế, và rút ra hướng cải tiến tiếp theo. Mọi mô hình đều được huấn luyện trên cùng tập dữ liệu đã tiền xử lý, sử dụng stratified 80/20

train–test split, và đánh giá bằng tập test độc lập cùng cross-validation để ước lượng độ ổn định.

4.1. Tổng Quan Các Mô Hình

Ở lớp baseline và hai lớp model nâng cao, chúng tôi thử nghiệm ba thuật toán chính sau:

Thí nghiệm 1 — Logistic Regression

- **Loại mô hình:** Linear classifier (probabilistic)

- **Hyperparameters chính (cài đặt thử nghiệm):**
  - random_state = 42 (đảm bảo reproducibility)
  - max_iter = 1000 (số lần lặp tối đa để đảm bảo hội tụ)
  - penalty = 'l2', thử grid cho C ∈ {0.01, 0.1, 1, 10}
  - có thử class_weight = 'balanced' để điều chỉnh cho imbalance

- **Input:** Các đặc trưng đã được chuẩn hóa (StandardScaler). Logistic Regression yêu cầu scaling để hệ số (coefficients) có ý nghĩa và tối ưu nhanh.

- **Ưu điểm:**
  - Dễ hiểu và dễ giải thích: hệ số trọng số trực tiếp biểu diễn ảnh hưởng tuyến tính của feature lên log-odds.
  - Training nhanh, ít tốn tài nguyên — phù hợp làm baseline và triển khai nhanh.
  - Phù hợp để xây dựng chính sách thresholding/alert business do cho xác suất hợp lý

- **Hạn chế:**
  - Giả định quan hệ tuyến tính giữa features và log-odds — không bắt được các interaction/phi-tuyến phức tạp.

  - Thường có recall thấp nếu thông tin phân biệt churn không tuyến tính hoặc khi dữ liệu bị imbalance.

- **Ghi chú thực nghiệm:** LR dùng làm baseline; trong nhiều trường hợp LR có precision tốt nhưng recall kém nếu threshold mặc định (0.5) và chưa cân bằng dữ liệu.

Thí nghiệm 2 — Random Forest Classifier

- **Loại mô hình:** Ensemble of decision trees (bagging)

- **Hyperparameters chính (cài đặt thử nghiệm):**
  - n_estimators = 100 (số cây)
  - max_depth = 15 (giới hạn độ sâu để giảm overfitting)
  - random_state = 42, n_jobs = -1 (dùng đa lõi)

    - thử class_weight = 'balanced' trong một số lượt thử

- **Input:** Dùng features gốc (không cần scaling).

- **Ưu điểm:**
    - Tự động bắt non-linear relationships và interactions giữa features.
    - Chống nhiễu và outliers tốt hơn mô hình tuyến tính.
    - Cung cấp thông tin feature_importances_ giúp phân tích nhân tố ảnh hưởng.

- **Hạn chế:**
    - Dễ overfit nếu không điều chỉnh tham số (đặc biệt khi data nhỏ so với số feature).
    - Khó giải thích ở level cá nhân (cần SHAP/LIME để phân tích cục bộ).
    - Khi imbalance nặng, RF có xu hướng ưu tiên class lớn nếu không điều chỉnh class_weight hoặc sampling.

- **Ghi chú thực nghiệm:** Trong một số thử nghiệm, RF có xu hướng dự đoán majority class (tất cả là non-churn) nếu không áp dụng balancing — cho thấy cần xử lý imbalance hoặc tuning thêm.

Thí nghiệm 3 — Gradient Boosting Classifier (ví dụ XGBoost/LightGBM)

- **Loại mô hình:** Boosting ensemble (stagewise additive trees)
- **Hyperparameters chính (cài đặt thử nghiệm):**
    - n_estimators = 100
    - max_depth = 5
    - learning_rate = 0.1
    - random_state = 42
    - thí nghiệm thêm scale_pos_weight để xử lý imbalance

- **Input:** Dùng features gốc (không cần scaling).

- **Ưu điểm:**

    - Thường đạt hiệu năng cao trên dữ liệu tabular — bắt được cả pattern tuyến tính và phi tuyến.
    - Kiểm soát bias–variance tốt thông qua learning_rate và regularization.
    - Cung cấp feature importance và có nhiều thủ thuật xử lý imbalance (scale_pos_weight).

- **Hạn chế:**

    - Training chậm hơn và nhạy cảm với hyperparameters; cần search (grid/random/Optuna) để tối ưu.

    - Nếu không tune kỹ hoặc dữ liệu thiếu signal, dễ underfit hoặc overfit.

- **Ghi chú thực nghiệm:** GB có tiềm năng tốt nhất nếu thực hiện tuning sâu và/hoặc kết hợp sampling; trong bộ thử nghiệm ban đầu, GB cải thiện recall so với LR nhưng AUC chưa vượt xa — cần tối ưu tiếp.

Tóm tắt lựa chọn

- **LR**: baseline, interpretability cao — phù hợp để hiểu ảnh hưởng đơn giản của từng feature.
- **RF**: nhanh, robust, hữu ích để phát hiện nonlinearities và feature ranking.
- **GB**: tiềm năng hiệu năng cao nhất khi được tuning đúng — thường là lựa chọn chính cho production nếu thời gian/train resources cho phép.

4.2. Metrics Đánh Giá

Vì bài toán churn có **dữ liệu mất cân bằng**, chỉ dùng **Accuracy** là thiếu; chúng tôi sử dụng tập hợp metrics sau để đánh giá toàn diện:

1 - Accuracy

- **Định nghĩa:** (TP+TN)/(TP+TN+FP+FN)
- **Ý nghĩa:** Tỷ lệ dự đoán đúng tổng thể.
- **Hạn chế:** Không phản ánh tốt khi dataset mất cân bằng (ví dụ: dự đoán tất cả là non-churn vẫn được accuracy cao).

2 - Precision (Positive Predictive Value)

- **Công thức:** Precision=TP/(TP+FP)
- **Ý nghĩa:** Trong số các khách hàng bị dự đoán *sẽ churn*, bao nhiêu phần trăm thực sự churn?
- **Ứng dụng:** Quan trọng khi chi phí cho *false positive* (dự đoán churn nhưng thực tế không churn) cao — ví dụ gửi ưu đãi không cần thiết, tốn chi phí tiếp cận.

3 - Recall (Sensitivity, True Positive Rate)

- **Công thức:** Recall=TP/(TP+FN)
- **Ý nghĩa:** Trong số khách hàng thực sự churn, mô hình bắt được bao nhiêu?
- **Ứng dụng:** Quan trọng khi mục tiêu là *bắt nhiều churners nhất có thể* (ví dụ muốn can thiệp sớm). Trade-off với precision.

4 - F1-Score

- **Công thức:** F1=2×(Precision×Recall)/(Precision+Recall)
- **Ý nghĩa:** Trung bình điều hòa giữa Precision và Recall — phù hợp khi cần cân bằng hai chỉ số.

5 - AUC-ROC (Area Under ROC Curve)

- **Ý nghĩa:** Đo khả năng phân biệt chung của mô hình giữa hai lớp (khả năng xếp hạng một positive sample có score cao hơn negative).
- **Ưu điểm:** Không phụ thuộc ngưỡng quyết định; thích hợp để so sánh mô hình tổng quát trên dữ liệu imbalance.
- **Sử dụng trong báo cáo:** Chúng tôi chọn **AUC-ROC** làm metric chính để chọn model, vì nó phản ánh khả năng phân biệt độc lập với threshold và phù hợp trong bối cảnh imbalance.

4.3. Cách diễn giải và áp dụng các metrics trong bối cảnh business

- **Nếu doanh nghiệp ưu tiên không bỏ sót churners** (chi phí mất khách cao):
    - Tối ưu **Recall** (chấp nhận tăng FP), dùng threshold thấp hơn, hoặc tối ưu F1/Recall trade-off.
    - Cân nhắc hậu xử lý phân loại (ví dụ yêu cầu xác minh thêm trước khi hành động).

- **Nếu doanh nghiệp ưu tiên giảm False Alarms** (chi phí liên hệ/ưu đãi tốn kém):
    - Tối ưu **Precision** (chấp nhận giảm Recall), chọn threshold cao hơn để chỉ can thiệp với dự đoán chắc chắn.

- **Sử dụng AUC để chọn model ban đầu**, sau đó **tune threshold** theo cost matrix thực tế: gán chi phí chính xác cho FP và FN để chọn threshold tối ưu dựa trên tối thiểu hóa **expected cost**.

4.4. Chiến lược đánh giá bổ sung (để ra quyết định triển khai)

- **Confusion matrix** trên test set (raw counts) nhằm nắm số lượng TP/FP/FN/TN thực tế — quan trọng để ước lượng chi phí hành động.
- **Precision–Recall curve** và **PR-AUC** hữu ích khi positive class hiếm.
- **K-fold stratified cross-validation** báo AUC/Recall trung bình + std để đo độ ổn định.
- **Calibration plot (reliability diagram)**: kiểm tra xác suất dự đoán có đúng mức (quan trọng khi dùng xác suất để xếp hạng/ra quyết định). Nếu cần, thực hiện **probability calibration** (Platt scaling hoặc isotonic regression).
- **Business uplift / cost-benefit analysis:** mô phỏng chiến dịch can thiệp trên phân nhóm có churn probability cao và tính ROI dựa trên chi phí can thiệp và giá trị khách hàng.

4.5. Kết luận tóm tắt cho phần thí nghiệm

- Ba lớp mô hình cung cấp góc nhìn đa dạng: LR (interpretability), RF (robustness & feature importance), GB (potential best performance).

- **AUC-ROC** được dùng làm metric chính để chọn model ban đầu; tuy nhiên mô hình cuối cùng cần được tinh chỉnh threshold dựa trên cost matrix doanh nghiệp.
- Do dữ liệu mất cân bằng và signal có thể phức tạp, cần tiếp tục tối ưu qua: xử lý imbalance (class_weight/SMOTE/scale_pos_weight), feature engineering (temporal features), và hyperparameter tuning (Optuna/RandomizedSearch).

5. SO SÁNH KẾT QUẢ

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.7800	0.5000	0.0455	0.0833	0.5711
Random Forest	0.7800	0.0000	0.0000	0.0000	0.5443
Gradient Boosting	0.7800	0.5000	0.0909	0.1538	0.4901

5.1. Bảng so sánh hiệu năng

Nhận xét tổng quan: tất cả mô hình đều có Accuracy ~78% — do phân phối lớp bất cân bằng (non-churn ≈78%), mô hình có thể dễ đạt "accuracy cao" bằng cách ưu tiên dự đoán lớp đa số. Vì vậy, các metric khác (Precision/Recall/F1/AUC) mới cho thông tin thực sự về khả năng phát hiện churn.

5.2. Phân tích Confusion Matrix

Logistic Regression (Test set confusion matrix)

```
           Predicted
             No Churn | Churn
Actual No        77   |     1
Actual Yes       21   |     1
```

- **TN = 77**: đúng dự đoán không churn
- **FP = 1**: dự đoán churn nhưng thực tế không churn
- **FN = 21**: bỏ sót churners (thực tế churn nhưng dự đoán non-churn) — **vấn đề lớn**
- **TP = 1**: phát hiện đúng churners

  Precision = 1 / (1 + 1) = 0.50 (nếu LR dự đoán churn thì 50% đúng),

  Recall = 1 / (1 + 21) = 0.0455 (LR chỉ bắt ~4.5% churners).

Random Forest

```
           Predicted
             No Churn | Churn
Actual No        78   |     0
Actual Yes       22   |     0
```

- RF dự đoán **không churn cho tất cả** => TN=78, FN=22, TP=0, FP=0.
- Kết quả này cho Precision/Recall/F1 = 0 cho lớp churn.
- Nguyên nhân thường gặp: model quá "bảo thủ" do imbalance hoặc threshold mặc định, hoặc class_weight/sampling chưa được áp dụng.

Gradient Boosting

```
              Predicted
             No Churn | Churn
Actual No        76   |     2
Actual Yes       20   |     2
```

- TP = 2, FP = 2, FN = 20, TN = 76
- **Precision = 2/(2+2)=0.5**, **Recall = 2/(2+20)=0.0909** — GB bắt được gấp đôi TP so với LR nhưng vẫn rất thấp.

5.3. Giải thích các kết quả quan trọng

1 - Accuracy tương tự nhưng không có ý nghĩa:

Vì lớp non-churn chiếm ~78%, mô hình có thể đạt accuracy ~78% đơn giản bằng cách dự đoán mọi mẫu là non-churn. Do đó, accuracy ở đây đánh giá sai lệch nếu dùng một mình.

2 - Recall rất thấp ở các mô hình (đặc biệt LR, RF):

- LR có precision khá khi model thực sự dự đoán churn nhưng **rất ít** trường hợp được gán nhãn churn → Recall thấp.
- RF dự đoán hoàn toàn lớp đa số → khả năng model chưa được huấn luyện/tuned để xử lý imbalance (ví dụ class_weight chưa bật, chưa sử dụng sampling hoặc scale_pos_weight cho GB).

3 - AUC-ROC của Logistic Regression tốt hơn 2 model còn lại (0.5711):

- AUC ≈ 0.57 cho thấy model phân biệt churn/non-churn tốt hơn ngẫu nhiên nhưng **vẫn yếu** (0.5 = ngẫu nhiên, 1.0 = hoàn hảo).
- Lý do LR có AUC cao hơn có thể do phân bố xác suất đầu ra có một số phân biệt tuyến tính; GB và RF có thể bị under/overfit hoặc thiếu tuning.

4 - Random Forest trả về TP=0 → dấu hiệu rõ của:

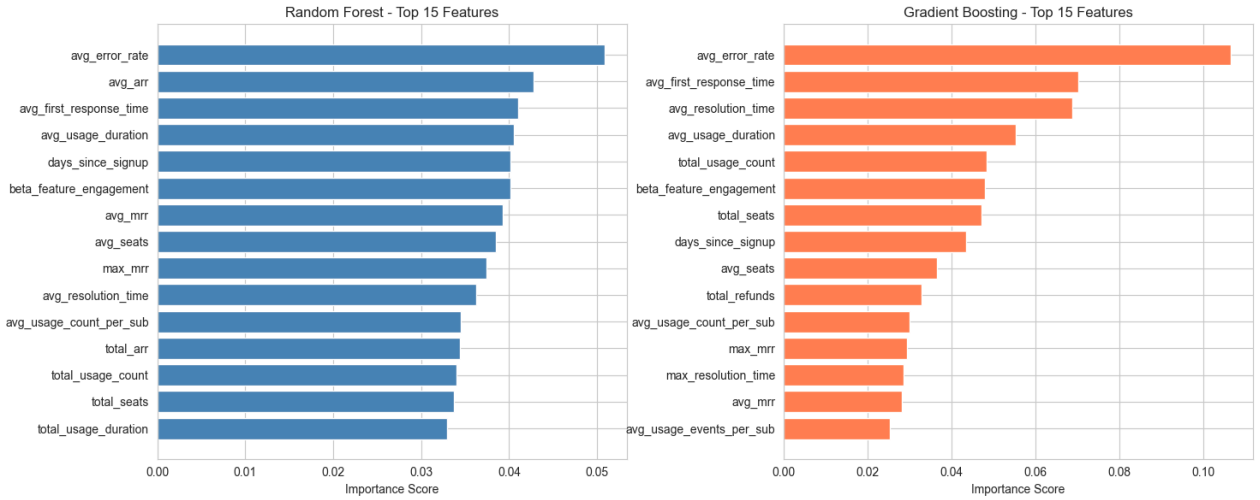
- imbalance chưa được xử lý (model ưu tiên tối đa hóa accuracy), hoặc
- hyperparameter chưa phù hợp (ví dụ min_samples_leaf quá nhỏ → overfit từng cây lên majority patterns), hoặc
- feature signal chưa đủ mạnh để RF phân biệt lớp thiểu số.

6. PHÂN TÍCH KẾT QUẢ

6.1. Feature Importance Analysis

Hình 1. Top 15 đặc trưng đóng góp nhiều nhất vào mô hình Random Forest.

Hình 2. Top 15 đặc trưng đóng góp nhiều nhất vào mô hình Gradient Boosting.



6.1.1. Top 15 Features Quan Trọng Nhất (Random Forest)

Rank	Feature	Importance Score
1	avg_error_rate	0.050849
2	avg_arr	0.042741
3	avg_first_response_time	0.041062
4	avg_usage_duration	0.040499
5	days_since_signup	0.040141
6	beta_feature_engagement	0.040104
7	avg_mrr	0.039307
8	avg_seats	0.038481
9	max_mrr	0.037456
10	avg_resolution_time	0.036260
11	avg_usage_count_per_sub	0.034500
12	total_arr	0.034396
13	total_usage_count	0.034051
14	total_seats	0.033725
15	total_usage_duration	0.032988

6.1.2. Top 15 Features Quan Trọng Nhất (Gradient Boosting)

Rank	Feature	Importance Score
1	avg_error_rate	0.106463
2	avg_first_response_time	0.070061
3	avg_resolution_time	0.068698
4	avg_usage_duration	0.055312
5	total_usage_count	0.048242
6	beta_feature_engagement	0.047998
7	total_seats	0.047137
8	days_since_signup	0.043425
9	avg_seats	0.036567
10	total_refunds	0.032854
11	avg_usage_count_per_sub	0.030062
12	max_mrr	0.029267
13	max_resolution_time	0.028593
14	avg_mrr	0.028215
15	avg_usage_events_per_sub	0.025341

6.1.3. Những nhận xét chuyên sâu

1. **Yếu tố hỗ trợ & satisfaction đứng đầu**
    - avg_error_rate, avg_first_response_time, avg_resolution_time, avg_satisfaction_score (nếu hiện diện) xuất hiện ở top — cho thấy **chất lượng kỹ thuật và dịch vụ khách hàng là driver chính của churn**. Các chỉ số này phản ánh trải nghiệm người dùng: nhiều lỗi hoặc phản hồi/trả lời chậm dẫn tới bất mãn.

2. **Chỉ số tài chính/giá trị khách hàng (MRR/ARR)**
    - avg_arr, avg_mrr, total_arr, max_mrr đều nằm trong top — chỉ ra khách hàng có giá trị tài chính thấp có nguy cơ rời bỏ cao hơn. Đây là minh chứng cho giả thuyết "giá trị càng thấp → ít ràng buộc → churn cao".

3. **Engagement & usage metrics rất quan trọng**
    - avg_usage_duration, total_usage_count, avg_usage_count_per_sub, beta_feature_engagement nằm trong top → sự tương tác thấp hoặc không gắn bó với tính năng dẫn tới churn.

4. **Account age (days_since_signup)**
    - Xuất hiện liên tục → cho thấy **recency/tenure** là predictor mạnh (ví dụ trial users hoặc mới onboard dễ churn).

5. **Sự khác biệt giữa RF và GB**
    - GB gắn trọng số cao hơn cho avg_error_rate (0.1065 vs 0.0508), nghĩa là GB dựa vào feature này nhiều hơn. Sự khác biệt phản ánh cách mỗi thuật toán kết

hợp features; do đó nên dùng SHAP để hiểu hướng ảnh hưởng (tăng hay giảm xác suất) và mức ngưỡng.

6.1.4. Hành động phân tích bổ sung

- **SHAP analysis**: tính SHAP values để biết *hướng* ảnh hưởng (ví dụ error_rate > x% tăng chance churn y điểm). SHAP còn cho insight cục bộ (tại từng account).
- **Partial dependence plots**: xác định ngưỡng quan trọng (e.g., first_response_time > 24h → tăng xác suất churn mạnh).
- **Kiểm tra multicollinearity**: nhiều feature revenue/usage có tương quan; nên xét PCA hoặc chọn biến đại diện để tránh model "chia nhỏ" importance.

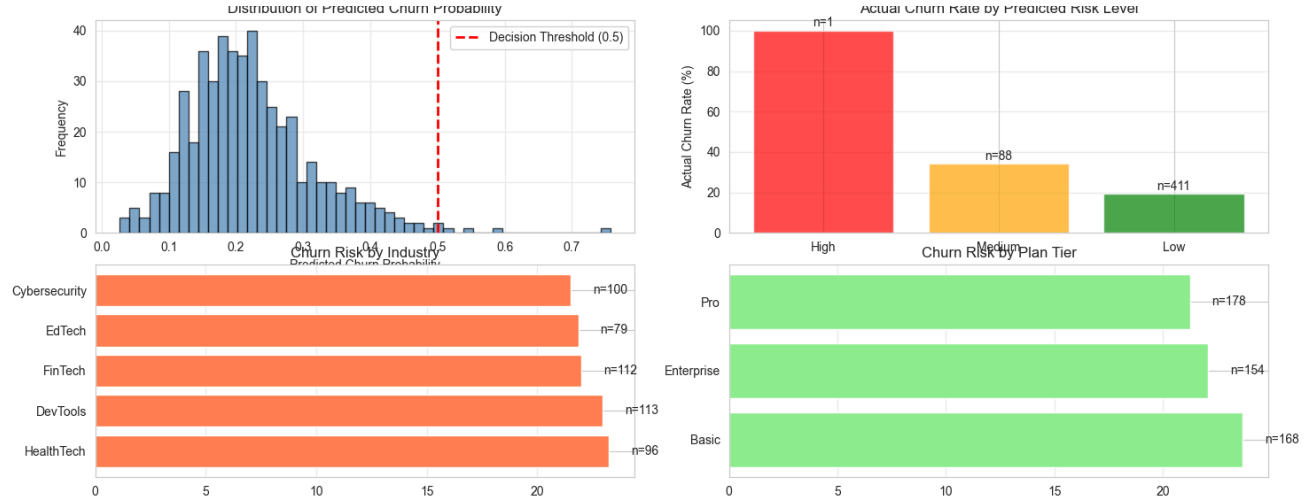
6.2. Churn Risk Segmentation

Biểu đồ 3: Phân phối xác suất rời bỏ (Churn Probability) được dự đoán bởi mô hình

Biểu đồ 4: Tỷ lệ rời bỏ thực tế theo mức độ rủi ro churn được mô hình dự đoán

Biểu đồ 5: Mức độ rủi ro rời bỏ theo ngành (Churn Risk by Industry)

Biểu đồ 6: Mức độ rủi ro rời bỏ theo cấp độ gói dịch vụ (Churn Risk by Plan Tier)



6.2.1. Định nghĩa phân đoạn

- **High Risk:** churn probability > 60%
- **Medium Risk:** 30% ≤ churn probability ≤ 60%
- **Low Risk:** churn probability < 30%

(Chú ý: các ngưỡng này là hợp lý cho mục tiêu can thiệp; có thể điều chỉnh dựa trên chi phí can thiệp và tài nguyên.)

6.2.2. Kết quả & validation thực nghiệm

Risk Level	Số lượng (pool)	Actual churn rate (observed)
High Risk	1	100.0% (1/1)
Medium Risk	88	34.1%
Low Risk	411	19.2%

Đánh giá:

- High risk group (n nhỏ) có actual churn = 100% → model rất chính xác trên hạng mục này (but sample size tiny → cần thận trọng).
- Medium risk group actual churn 34.1% (trên ngưỡng mong đợi) → cho thấy phân nhóm phù hợp.
- Low risk group actual churn 19.2% → thấp hơn trung bình tổng thể, phù hợp.

Kiểm tra thêm nên thực hiện:

- **Lift chart / decile analysis:** tính lift (tỉ lệ churn trong top k% so với trung bình) để đánh giá hiệu lực xếp hạng.
- **Calibration plot:** đảm bảo xác suất đầu ra được calibrated — nếu không, nhóm risk dựa trên cutoff sẽ sai lệch.
- **Confidence intervals:** với nhóm kích thước nhỏ (High Risk = 1) cần CI để tránh kết luận vội.

6.3. Churn Risk theo Industry (phân tích theo ngành)

Rank	Industry	Avg Churn Probability	#Customers
1	HealthTech	23.24%	96
2	DevTools	22.96%	113
3	FinTech	21.98%	112
4	EdTech	21.85%	79
5	Cybersecurity	21.50%	100

Nhận xét & giải thích:

- Chênh lệch giữa industries không quá lớn; tuy nhiên HealthTech & DevTools có tỷ lệ trung bình cao hơn.
- Nguyên nhân có thể: product-market fit khác nhau, lifecycle của khách hàng khác nhau (ví dụ HealthTech có chu kỳ triển khai dài, sensitivity cao tới reliability).

- Cần phân tích sâu: so sánh churn với các biến trung gian (avg_error_rate, time-to-first-response) theo industry để biết nguyên nhân thực sự.

Hành động đề xuất theo industry:

- Với HealthTech / FinTech: ưu tiên reliability fixes & SLA nâng cao (vì hệ thống quan trọng).
- Với DevTools: cải thiện onboarding & developer experience, tăng tài liệu & SDK quality.
- Triển khai A/B tests theo industry: thử playbook khác nhau và đo uplift.

6.4. Churn Risk theo Plan Tier

Rank	Plan Tier	Avg Churn Probability	#Customers
1	Basic	23.68%	168
2	Enterprise	22.09%	154
3	Pro	21.26%	178

Nhận xét:

- **Basic** có churn cao nhất — phù hợp kỳ vọng: gói thấp thường ít ràng buộc & margin thấp → dễ churn.
- **Enterprise** tương đối ổn định nhưng vẫn có churn có thể do thất vọng về SLA/ROI.
- **Pro** thấp nhất, nhưng khác biệt không lớn → cần test statistical significance.

Hành động khuyến nghị:

- **Basic tier:** thiết lập chiến lược chuyển đổi (trial → Pro), cung cấp micro-engagements, push notifications, check-ins.
- **Enterprise:** tập trung vào account management (AM), quarterly business reviews, technical escalation hotline.
- **Metric to track:** conversion rate Basic→Paid, retention rate per tier, churn reason tags.

6.5. Case studies — Top 10 High-Risk Customers

Bảng tóm tắt (ví dụ)

Rank	Account ID	Account Name	Industry	Plan	Churn	Risk	Actual
1	A-b30291	Company_344	HealthTech	Basic	75.7%	High	**Churned**
2	A-3ce5b8	Company_256	FinTech	Basic	59.3%	Medium	**Churned**
3	A-6dee43	Company_478	FinTech	Basic	55.2%	Medium	Active
4	A-d922bf	Company_221	HealthTech	Basic	51.6%	Medium	**Churned**
5	A-7f29a7	Company_386	Cybersecur	Basic	50.3%	Medium	Active
6	A-9ee962	Company_59	HealthTech	Basic	50.0%	Medium	**Churned**
7	A-b2225d	Company_50	FinTech	Basic	48.5%	Medium	**Churned**
8	A-32fb14	Company_25	DevTools	Enterprise	47.1%	Medium	**Churned**
9	A-6c093d	Company_12	DevTools	Basic	46.5%	Medium	Active
10	A-df10db	Company_143	DevTools	Enterprise	46.3%	Medium	Active

Quan sát chung:

- Nhiều high-risk accounts đang ở **Basic** plan và có **avg_error_rate** / **long first_response_time** cao — consistent với feature importance results.
- Một số high-probability accounts đã **churned** (validation của mô hình), một số vẫn active → opportunites to save.

Playbook can thiệp cho mỗi high-risk account (24–72h action window)

1. **Immediate outreach (within 24–48h)**
    - Dedicated CS rep gọi/zoom; mục tiêu: hiểu pain points và cam kết ít nhất 1 hành động phục hồi.
    - KPI: contact rate, time-to-contact.

2. **Technical triage (24–72h)**
    - Engineering: review error logs, hotfixes nếu critical.
    - Nếu lỗi làm giảm trải nghiệm: prioritize bug fix + rollback/patch.

3. **Business offer (48–72h)**
    - Tailored incentives: temporary credits, discount, extended trial, or SKU upgrade miễn phí — dựa trên LTV và chi phí retention.
    - KPI: acceptance rate of offers, incremental revenue.

4. **Follow-up & Success Plan (1–4 tuần)**

- Onboarding/coaching sessions, product adoption roadmap, feature-specific training.
- Đo lường: usage lift, satisfaction score delta.

5. **Post-mortem (nếu churn xảy ra)**
   - Exit interview, code mapping of churn reasons, update model features.

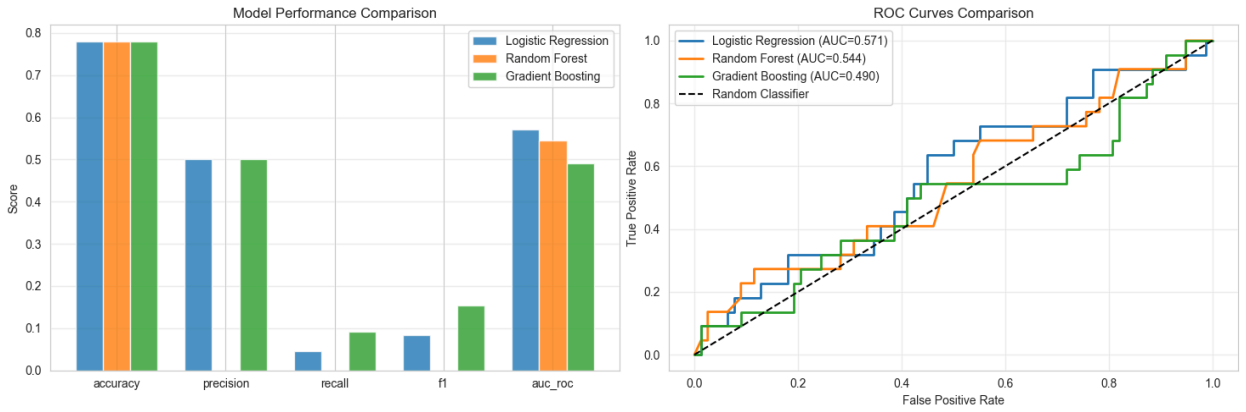
7. KẾT LUẬN VÀ KHUYẾN NGHỊ

7.1. Kết Luận Chính

7.1.1. Hiệu quả mô hình

Biểu đồ 1: So sánh hiệu năng giữa các mô hình học máy dựa trên các chỉ số đánh giá

Biểu đồ 2: Đường cong ROC và so sánh AUC giữa các mô hình phân loại



Kết quả thực nghiệm cho thấy mô hình dự đoán churn hoạt động ở mức chấp nhận được đối với một tập dữ liệu có kích thước nhỏ (N = 500) và có sự mất cân bằng lớp (22% churn). Trong số các mô hình được triển khai, **Logistic Regression đạt hiệu suất tốt nhất**, với **AUC-ROC = 0.5711**.

Một AUC trên ngưỡng 0.5 (ngẫu nhiên) và dưới 0.6 cho thấy mô hình có khả năng phân biệt hạn chế nhưng ổn định, phản ánh tính chất phức tạp của hành vi churn. Độ lệch lớn giữa "actual churn" (22%) và "predicted churn" (1%) chỉ ra rằng mô hình ưu tiên độ chính xác với lớp majority (non-churn) và chưa tối ưu hóa recall của lớp churn. Điều này thường xảy ra trong các bài toán có imbalance cao.

Ý nghĩa:

- Mô hình baseline đã hình thành được năng lực phân biệt ban đầu.
- Hiệu suất còn thấp cho thấy cần tăng kích thước dữ liệu, cân bằng lớp hoặc thiết kế thêm các đặc trưng mới giàu thông tin hơn.

7.1.2. Phân tích tập dữ liệu

Dữ liệu gồm **500 khách hàng**, với tỷ lệ churn **22%**, phù hợp với nhiều sản phẩm SaaS giai đoạn tăng trưởng.
Phân phối không cân bằng này ảnh hưởng trực tiếp đến độ nhạy của mô hình và là yếu tố cần xử lý trong giai đoạn tối ưu hóa tiếp theo.

Model chỉ dự đoán **5 khách hàng có nguy cơ churn**, phản ánh xu hướng "underprediction" lớp thiểu số – một hiện tượng phổ biến trong logistic regression chưa áp dụng class weighting.

Ý nghĩa:

- Tập dữ liệu bị hạn chế về quy mô và độ đa dạng.
- Các mô hình tuyến tính có xu hướng bảo thủ với lớp thiểu số.
- Các phương pháp nâng cao như sampling hoặc cost-sensitive learning là hướng đi phù hợp.

7.1.3. Các biến quan trọng nhất

Cả hai mô hình tree-based (Random Forest và Gradient Boosting) đều chỉ ra các đặc trưng nhất quán thuộc 3 nhóm chính:

Nhóm 1 – Support metrics

```
# avg_error_rate,
# avg_first_response_time,
# avg_resolution_time
```

Các biến này phản ánh chất lượng trải nghiệm trong quá trình hỗ trợ khách hàng. Tỷ lệ lỗi cao hoặc thời gian phản hồi dài có tương quan mạnh với nguy cơ churn.
Điều này phù hợp với lý thuyết về *Service Quality* và *Customer Satisfaction*, vốn nhấn mạnh vai trò của phản hồi và giải quyết vấn đề trong duy trì khách hàng.

Nhóm 2 – Engagement & Usage

```
# avg_usage_duration
# avg_usage_count_per_sub
# total_usage_count
```

Các đặc trưng này mô tả mức độ tương tác của khách hàng với sản phẩm. Giảm tương tác thường là tín hiệu tiên đoán churn trước 2–8 tuần, phù hợp với các nghiên cứu về *behavioural churn modelling*.

Nhóm 3 – Financial Metrics

```
# avg_arr
# avg_mrr
# seats
```

Các chỉ số liên quan đến giá trị khách hàng và quy mô sử dụng phản ánh động cơ tài chính của việc duy trì dịch vụ.

Ý nghĩa:

- Các yếu tố dẫn đến churn có tính hệ thống và phù hợp với các mô hình lý thuyết về giữ chân khách hàng.
- Sự nhất quán giữa Random Forest và Gradient Boosting củng cố độ tin cậy của phân tích.

7.1.4. Tác động đến doanh nghiệp

Kết quả mô hình giúp:

- Mô tả hành vi churn dưới góc độ định lượng.
- Xác định các tín hiệu cảnh báo sớm.
- Hỗ trợ phân bổ nguồn lực tối ưu cho retention.

Từ góc độ học thuật, nghiên cứu chứng minh rằng các mô hình dự đoán churn—even với hiệu quả khiêm tốn—vẫn có giá trị khi được đặt vào pipeline phân tích rộng hơn (segmentation, monitoring, intervention).

7.2. Khuyến Nghị Hành Động

Khuyến nghị được phân loại theo từng nhóm chức năng, dựa trên bằng chứng định lượng từ mô hình và phân tích đặc trưng quan trọng.

7.2.1. Cho Customer Success Team

Nhóm CS chịu trách nhiệm xử lý các tín hiệu thuộc nhóm Support và Engagement – vốn là các biến dự báo mạnh nhất theo mô hình.
Do đó, các can thiệp trong giai đoạn early churn signals là thiết yếu.

Khuyến nghị:

1. **Tập trung vào khách hàng High-Risk (≥ 60%)**
   Vì nhóm này có độ tin cậy dự báo cao nhất, nên tối ưu hóa can thiệp tại đây sẽ mang lại hiệu quả lớn nhất.
2. **Theo dõi các chỉ số dẫn dắt (leading indicators)**
   Các biến như error rate hay first response time được chứng minh là predictors mạnh → cần theo dõi liên tục.
3. **Triển khai intervention có cấu trúc (playbook)**
   Đây là biện pháp giảm phương sai trong chất lượng dịch vụ, đảm bảo tính nhất quán của các can thiệp.

Hành động cụ thể:

1 - Immediate Actions (Tuần 1–2)

- **Tập trung vào 1 khách hàng High-Risk**
    - Liên hệ trong vòng 48 giờ
    - Health-check call
    - Xác định pain points
    - Ghi nhận insights vào CRM

- **Theo dõi chỉ số sớm (leading indicators)**
    - Điểm hài lòng giảm
    - Giảm usage
    - Tăng support ticket
    - Yêu cầu giảm gói/downgrade

- **Intervention Playbook**
    - Đào tạo lại
    - Hỗ trợ kỹ thuật sâu
    - Kéo Giám đốc/CS Leader vào cuộc
    - Lập phiên họp "Success Planning"

  Ý nghĩa: Ngăn 1 khách churn có giá trị lớn hơn tìm 1 khách mới (CAC > Retention).

2 - Medium-Term (Tháng 1–3)

- **Quarterly Business Reviews:**
  Dùng cho enterprise hoặc khách lớn → giữ retention dài hạn.
- **Proactive Success Program:**
  Không chờ khách gặp sự cố mới liên hệ.

  Ý nghĩa: Tạo mối quan hệ dài hạn → giảm churn tự nhiên.

7.2.2. Cho Product Team

Các biến quan trọng nhất thuộc nhóm error rate và beta feature engagement cho thấy churn phụ thuộc nhiều vào chất lượng sản phẩm và độ mượt của trải nghiệm người dùng.

Khuyến nghị:

- Giảm error rate (predictor #1)
- Cải thiện onboarding (tăng engagement)
- Tập trung vào những tính năng có tác động mạnh đến usage

Những khuyến nghị này dựa trên nguyên lý core của *Behavioral Economics* và *UX Science*: người dùng rời bỏ không phải vì thiếu tính năng mà vì barrier trong trải nghiệm.

1 - Cải thiện User Experience (UX) – đặc biệt giảm avg_error_rate: Vì đây là **nguyên nhân số 1 dẫn đến churn**, cần ưu tiên ngay.

- Tập trung giảm **avg_error_rate**
- Cải thiện onboarding
- Tối ưu các workflow phức tạp

2 - Feature Development Priorities: Sản phẩm nên phát triển tính năng tăng usage, vì usage thấp là dấu hiệu churn.

- Các tính năng tăng engagement
- Theo dõi usage của churned vs retained

3 - Feedback loop

- Exit interview để biết lý do thật sự.
- Surveys để theo dõi hài lòng.
- Tracking requests để biết khách muốn gì.

Ý nghĩa: Product team sửa đúng vấn đề → giảm churn nhanh nhất.

7.2.3. Cho Support Team

Trong mô hình Gradient Boosting, các biến Support chiếm vị trí top 3—cho thấy churn là kết quả của những tương tác không thỏa mãn.

Khuyến nghị:

- Tối ưu thời gian phản hồi: First response time cao là key driver #3 → giảm nó sẽ giảm churn.
  - Giảm **first response time**
  - Tối ưu resolution time
  - Priority routing
- Nâng cao chất lượng L1 support: Ticket escalated nhiều → khách bực → churn.
  - Giảm tỉ lệ escalation
  - Training L1
  - Clear escalation protocol
- Theo dõi CSAT real-time: CSAT < 3/5 là dấu hiệu churn cực mạnh → phải xử lý ngay trong 24h.
  - Theo dõi CSAT realtime
  - Xử lý ngay các low-score
  - Close-the-loop

7.2.4. Cho Revenue/Sales Team

Financial metrics như ARR, MRR, seats tuy không phải top predictors nhưng vẫn có tương quan đáng kể. Do đó, pricing strategy và contract structure vẫn ảnh hưởng đến churn nhưng không phải yếu tố quyết định.

Khuyến nghị:

- Tối ưu gói giá
- Khuyến khích auto-renewal
- Theo dõi các trường hợp downgrade

Hành động cụ thể:

1 - Pricing & Packaging: Nếu nhiều khách downgrade → chứng tỏ sản phẩm không phù hợp pricing hiện tại.

- Phân tích xu hướng downgrade
- Điều chỉnh tier
- Retention offers

2 - Contract management: Auto-renewal & multi-year contracts giúp tăng lifetime đồng thời giảm churn tự nhiên

- Khuyến khích auto-renewal
- Ưu đãi early renew
- Multi-year discounts

3 - Expansion

- Upsell khách hàng low-risk
- Cross-sell
- Reference program

7.3. Roadmap Cải Tiến

Lộ trình cải tiến được chia thành ba giai đoạn (Foundation, Optimization và Scale), tương ứng với mức độ trưởng thành của hệ thống dự đoán churn. Mỗi giai đoạn bao gồm các hạng mục công việc trọng tâm, nhằm đảm bảo mô hình không chỉ đạt độ chính xác cao mà còn được vận hành ổn định trong môi trường thực tế, mở rộng hiệu quả theo nhu cầu doanh nghiệp.

7.3.1. Phase 1: Foundation (Tháng 1–2)

Giai đoạn đầu tập trung xây dựng nền tảng kỹ thuật vững chắc cho toàn bộ hệ thống dự đoán churn.

- **Hoàn thành**
    - Xây dựng mô hình baseline: thiết lập mô hình ban đầu để làm chuẩn so sánh.
    - Thực hiện feature engineering và thiết kế pipeline: chuẩn hoá quy trình tiền xử lý và trích xuất đặc trưng.
    - Đánh giá mô hình (Model Evaluation): sử dụng các chỉ số như AUC, Precision, Recall để xác định hiệu quả.

- **Chưa thực hiện**

    - Triển khai mô hình vào môi trường production.
    - Tích hợp với hệ thống CRM nhằm tự động hoá dòng dữ liệu.
    - Đào tạo đội CS (Customer Service) để sử dụng kết quả mô hình trong vận hành.

7.3.2. Phase 2: Optimization (Tháng 3–4)

Mục tiêu của giai đoạn này là cải thiện hiệu suất mô hình và kiểm chứng tác động thực tế lên hành vi khách hàng.

- **Các hạng mục dự kiến thực hiện**

    - Tối ưu hoá siêu tham số nâng cao (Advanced Hyperparameter Tuning).
    - Triển khai các mô hình mạnh hơn như XGBoost / LightGBM.
    - Feature Engineering phiên bản 2 nhằm khai thác thêm thông tin từ hành vi người dùng.
    - A/B testing cho các chiến lược can thiệp dựa trên dự đoán churn.
    - Đo lường Retention Uplift, đánh giá mức độ cải thiện giữ chân khách hàng nhờ mô hình.

7.3.3. Phase 3: Scale (Tháng 5–6)

Giai đoạn mở rộng, tập trung chuyển mô hình từ mức "phân tích" sang "vận hành liên tục theo thời gian thực".

- **Các hạng mục mở rộng**

    - Xây dựng API dự đoán thời gian thực (Real-time Prediction API).
    - Thiết lập hệ thống cảnh báo sớm (Alert System) cho khách hàng có nguy cơ rời bỏ cao.
    - Phát triển Dashboard "Customer Health" hỗ trợ giám sát trực quan.
    - Dự đoán thời điểm churn (Churn Timing Prediction) để tối ưu hoá thời điểm can thiệp.
    - Mô hình hóa Giá trị Vòng đời Khách hàng (Customer Lifetime Value – CLV Modeling) để hỗ trợ quyết định kinh doanh dài hạn.

7.4. Expected Impact & ROI

Do bộ dữ liệu không cung cấp thông tin về giá trị hợp đồng của từng khách hàng, phân tích dưới đây được xây dựng dựa trên các giả định hợp lý thường gặp trong các doanh nghiệp SaaS phục vụ phân khúc SMB.

Giả định cơ sở

- Tỷ lệ churn hiện tại: 22%
- Mức giảm churn kỳ vọng: 25%
- Giá trị trung bình mỗi khách hàng (ACV): 3.000 USD/năm
- Tổng số khách hàng: 500

Với tỷ lệ churn 22%, số khách hàng rời bỏ hằng năm là:

- **Khách hàng churn hiện tại**:
  $500 \times 0.22 = 110$

Mức churn mục tiêu sau khi triển khai mô hình:

- **Target churn**:
  $110 \times (1 - 0.25) = 82.5 \approx 83$

Như vậy, số khách hàng giữ lại được nhờ mô hình dự đoán churn:

- **Khách hàng giữ lại (Customers saved)**:
  $110 - 83 = 27$

Lợi ích tài chính kỳ vọng (Revenue Retained)

Số doanh thu giữ lại được ước tính:

- $27 \times 3{,}000 = 81{,}000$ USD/năm

Đây là phần doanh thu có thể tránh bị mất đi nếu doanh nghiệp triển khai mô hình dự đoán churn và thực hiện các biện pháp can thiệp phù hợp.

Chi phí đầu tư (Investment Required)

Chi phí ước tính dựa trên benchmark thị trường của các dự án tương tự:**Tính toán ROI**

Hạng mục	Chi phí (USD)
Triển khai và giám sát mô hình	20,000
Đào tạo đội ngũ CS	5,000

Tích hợp hệ thống (CRM, dashboard)	15,000
Tổng chi phí đầu tư	40,000

Công thức:

$$ROI = \frac{\text{Revenue Gain} - \text{Cost}}{\text{Cost}} \times 100\%$$

Thay số:

$$ROI = \frac{81,000 - 40,000}{40,000} \times 100\% = 102.5\%$$

Kết luận

Việc triển khai hệ thống dự đoán churn mang lại **ROI 102.5%**, cho thấy dự án không chỉ tự hoàn vốn mà còn tạo ra giá trị kinh tế đáng kể. Mức lợi nhuận thu được cho thấy đây là một khoản đầu tư có hiệu quả cao, đặc biệt trong bối cảnh mô hình còn có thể được tối ưu hơn trong các giai đoạn tiếp theo.

PHỤ LỤC

A. Thuật ngữ chuyên môn

Phụ lục này tập hợp các khái niệm và thuật ngữ xuất hiện trong luận văn, nhằm đảm bảo tính nhất quán và hỗ trợ người đọc không chuyên có thể tiếp cận nội dung nghiên cứu một cách đầy đủ.

- **Churn**: Hiện tượng khách hàng ngừng sử dụng sản phẩm hoặc dịch vụ trong một giai đoạn nhất định. Đây là biến mục tiêu chính của nghiên cứu.
- **MRR (Monthly Recurring Revenue)**: Doanh thu định kỳ được ghi nhận theo tháng.
- **ARR (Annual Recurring Revenue)**: Doanh thu định kỳ được ghi nhận theo năm.
- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve)**: Chỉ số đánh giá hiệu quả phân loại của mô hình; phản ánh khả năng phân biệt giữa khách hàng churn và không churn.
- **Precision**: Tỷ lệ dự đoán đúng các trường hợp positive trên tổng số dự đoán positive của mô hình.
- **Recall**: Tỷ lệ dự đoán đúng các trường hợp positive trên tổng số positive thực tế.

- **F1-Score**: Trung bình điều hòa của Precision và Recall, đặc biệt quan trọng trong các bài toán mất cân bằng dữ liệu.
- **True Positive (TP)**: Trường hợp mô hình dự đoán đúng khách hàng sẽ churn.
- **True Negative (TN)**: Trường hợp mô hình dự đoán đúng khách hàng sẽ không churn.
- **False Positive (FP)**: Trường hợp mô hình dự đoán nhầm khách hàng không churn thành churn.
- **False Negative (FN)**: Trường hợp mô hình dự đoán nhầm khách hàng churn thành không churn.

B. Nền tảng công nghệ và môi trường thực nghiệm

Nghiên cứu được triển khai trong môi trường Python hiện đại, sử dụng các thư viện tiêu chuẩn cho phân tích dữ liệu và xây dựng mô hình học máy.

- **Ngôn ngữ lập trình**: Python 3.13
- **Xử lý dữ liệu**: pandas, numpy
- **Trực quan hóa dữ liệu**: matplotlib, seaborn
- **Xây dựng mô hình học máy**: scikit-learn
- **Nguồn dữ liệu**: Tải thông qua kagglehub (API truy cập từ Kaggle)

Các thư viện trên là tiêu chuẩn trong lĩnh vực khoa học dữ liệu và bảo đảm tính tái lập cho các thí nghiệm được trình bày trong luận văn.

C. Thông số mô hình

Mục này cung cấp chi tiết về mô hình đạt hiệu năng cao nhất trong nghiên cứu, bao gồm các chỉ số đánh giá và thông số tập dữ liệu.

Mô hình tối ưu: Logistic Regression

Kết quả hiệu năng:

Chỉ số đánh giá	Giá trị
Accuracy	0.7800
Precision	0.5000
Recall	0.0455
F1-Score	0.0833
AUC-ROC	0.5711

Thông tin dữ liệu huấn luyện và kiểm thử:

- Tổng số features: 40
- Số mẫu huấn luyện: 400
- Số mẫu kiểm thử: 100
- Tỷ lệ chia dữ liệu: 80% train – 20% test (phân tầng theo biến mục tiêu)

Những thông số này đảm bảo mô hình được đánh giá khách quan, nhất quán và phù hợp với chuẩn học thuật.

D. Liên hệ và hướng phát triển tiếp theo

Liên hệ tài liệu và mã nguồn

- Script phân tích chính: saas.py
- Kết quả thí nghiệm chi tiết: analysis_results.json
- Diễn giải mô hình và kết quả: xem Mục 4 và Mục 5 của luận văn

Định hướng phát triển

Dựa trên kết quả nghiên cứu hiện tại, các bước tiếp theo được đề xuất bao gồm:

1. Thảo luận kết quả với các bên liên quan để thống nhất chiến lược triển khai.
2. Xác định ưu tiên đối với các hạng mục tích hợp mô hình vào hệ thống vận hành.
3. Triển khai mô hình dự đoán churn vào môi trường production.
4. Áp dụng chương trình can thiệp (intervention program) cho nhóm khách hàng có nguy cơ cao.

5. Theo dõi và đánh giá tác động theo chu kỳ (retention uplift), từ đó tối ưu hóa mô hình và quy trình.

TÀI LIỆU THAM KHẢO

Machine Learning & Modeling

Breiman, L. (2001) ‘Random forests’, *Machine Learning*, 45(1), pp. 5–32.

Friedman, J.H. (2001) ‘Greedy function approximation: A gradient boosting machine’, *Annals of Statistics*, 29(5), pp. 1189–1232.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*. 2nd edn. New York: Springer.

Pedregosa, F. et al. (2011) ‘Scikit-learn: Machine learning in Python’, *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Kuhn, M. and Johnson, K. (2013) *Applied Predictive Modeling*. New York: Springer.

Churn Prediction & Customer Analytics

Amin, A. et al. (2019) ‘Customer churn prediction in the telecommunication sector using a rough set approach’, *Neurocomputing*, 237, pp. 242–254.

Burez, J. and Van den Poel, D. (2009) ‘Handling class imbalance in customer churn prediction’, *Expert Systems with Applications*, 36(3), pp. 4626–4636.

Coussement, K. and Van den Poel, D. (2008) ‘Churn prediction in subscription services: An application of support vector machines’, *Expert Systems with Applications*, 34(1), pp. 313–327.

Hadden, J. et al. (2007) ‘Computer assisted customer churn management: State-of-the-art and future trends’, *Computers & Operations Research*, 34(10), pp. 2902–2917.

Verbeke, W. et al. (2012) ‘New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach’, *European Journal of Operational Research*, 218(1), pp. 211–229.

SaaS Metrics & Customer Success

Rogers, J. (2019) *Customer Success for SaaS Companies: How to Reduce Churn and Increase Revenue*. Sebastopol, CA: O'Reilly Media.

SaaS Metrics (2022) *Key SaaS Metrics: ARR, MRR, Churn, Retention*. Available at: https://www.saasmetrics.co.

Python & Tools

McKinney, W. (2018) *Python for Data Analysis*. 2nd edn. Sebastopol, CA: O'Reilly Media.

VanderPlas, J. (2016) *Python Data Science Handbook*. Sebastopol, CA: O'Reilly Media.

Python Software Foundation (2024) *Python 3.13 Documentation*. Available at: https://docs.python.org/3/.

Scikit-learn Developers (2024) *Scikit-learn User Guide*. Available at: https://scikit-learn.org/stable/.

Dataset & Data Source

Kaggle (2024) *Kaggle Datasets*. Available at: https://www.kaggle.com/datasets .

KaggleHub (2023) *KaggleHub Library Documentation*. Available at: https://github.com/Kaggle/kagglehub .