# **GRADUATION PROJECT PROPOSAL**

**Phân tích những yếu tố tác động với khả năng rời bỏ của khách hàng (Churn)**

1. **Mục đích và kết quả (Outcome)**

**Mục đích:**

Nghiên cứu dữ liệu nhằm đánh giá các yếu tố ảnh hưởng tới hành vi rời bỏ của khách hàng.

Đánh giá xu hướng và tìm ra những yếu tố quan trọng dẫn tới việc khách hàng rời bỏ.

Tìm ra nguyên nhân và giải pháp để hạn chế hành vi rời bỏ (Churn) của khách hàng.

**Kết quả mong đợi (Outcome):**

* Cung cấp insight về mối quan hệ giữa các yếu tố ( complain, Satisfactionscore, MaritalStatus, Tenure,…) và hành vi rời bỏ.
* Xây dựng mô hình học máy dự đoán churn, như Logistic Regression nhằm :
* Tính toán **khả năng churn** của mỗi khách hàng hiện tại hoặc mới.
* Tự động **phân loại khách hàng có khả năng rời bỏ**, từ đó giúp doanh nghiệp chủ động giữ chân họ.

1. Câu hỏi mục tiêu:

* Mô hình nào giải thích tốt nhất các yếu tố dẫn đến churn?
* Yếu tố nào tác động mạnh nhất đến khả năng rời bỏ? Có bằng chứng thống kê ?
* Từ dữ liệu, có thể rút ra insight gì cho chiến lược giữ chân khách hàng?
* Ngưỡng xác suất dự đoán nào giúp mô hình phát hiện được nhiều nhất khách hàng có khả năng rời bỏ (churn)?

1. **Nguồn dữ liệu:**

**Nguồn :** [**Kaggle.com**](http://kaggle.com)

**Link dataset:** [**https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction/data**](https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction/data)

| **Cột** | **Ý nghĩa tên cột** |
| --- | --- |
| **CustomerID** | Mã định danh duy nhất của khách hàng. Không mang ý nghĩa dự đoán (nên loại bỏ). |
| **Churn** | Biến mục tiêu: Khách hàng có rời bỏ không. Thường: 1 = rời bỏ, 0 = còn hoạt động. |
| **Tenure** | Thời gian khách hàng gắn bó với công ty |
| **PreferredLoginDevice** | Thiết bị đăng nhập ưa thích của khách hàng |
| **Citytier** | Phân hạng thành phố |
| **WarehousetoHome** | Khoảng cách từ kho đến nhà khách hàng |
| **PreferredPaymentMode** | Phương thức thanh toán ưa thích: COD (khi nhận hàng), Credit Card, Online, v.v. |
| **Gender** | Giới tính của khách hàng (Male, Female). |
| **HourSpendOnApp** | Số giờ trung bình khách dùng app mỗi ngày. |
| **NumberOfDeviceRegistered** | Tổng số thiết bị đã được đăng ký cho tài khoản của khách hàng |
| **PreferrdorderCat** | Danh mục sản phẩm ưa thích để đặt hàng trong tháng qua |
| **SatisfactionScore** | Mức độ hài lòng của khách hàng (thang điểm, thường từ 1 đến 5). |
| **MaritalStatus** | Tình trạng hôn nhân của khách (Single, Married, Divorced, etc.). |
| **NumberOfAddress** | Tổng số địa chỉ đã thêm vào cho tài khoản khách hàng |
| **Complain** | Biến nhị phân: 1 = có gửi khiếu nại, 0 = không. |
| **OrderAmountHikeFromlastYear** | Tỷ lệ phần trăm tăng trưởng giá trị đơn hàng so với năm trước |
| **CouponUsed** | Tổng số mã giảm giá đã được sử dụng trong tháng qua |
| **OrderCount** | Tổng số đơn hàng đã đặt bởi khách hàng. |
| **DaySinceLastOrder** | Số ngày kể từ lần đặt hàng gần nhất. |
| **CashbackAmount** | Tổng số tiền hoàn lại mà khách đã nhận được. |

**Mô tả :** Dữ liệu mang những đặc điểm khách hàng liên quan đến hành vi rời bỏ như ( kỳ hạn , điểm hài lòng, phàn nàn, số giờ dùng app mỗi ngày, phương thức thanh toán ưa thích (thẻ, COD, ví điện tử...).

1. **Cleaning Data:**

* Xử lý các giá trị bị thiếu (Handling Missing Value)
* Chuyển đổi dữ liệu (Data type)
* Kiểm tra Duplicate
* Xử lý Outliners.

1. **Phân tích khám phá dữ liệu (EDA)**

* Tính toán các thống kê mô tả cho các biến định lượng (trung bình, min, max, phân vị...).
* Kiểm tra phân phối hành vi rời bỏ của khách hàng qua các đặc tính như :

**Tenure**

**CityTier**

**Complain**

**OrderAmountHikeFromlastYear**

**DaySinceLastOrder**

**SatisfactionScore.**

1. **Phân tích mô hình**

* Chia dữ liệu thành tập huấn luyện và kiểm tra (70/30)
* Dùng stastmodels.logit() để huấn luyện mô hình.
* Xem kết quả summary() để hiểu hệ số và P- value.

1. **Đánh giá mô hình**

Dự đoán xác xuất => Phân loại 0/1 ( dựa vào ngưỡng 0,5)

**F1-score:** Dùng khi dữ liệu mất cân bằng (VD: churn chỉ 10%)

**Precision:** Trong tất cả khách hàng mà mô hình dự đoán sẽ *churn* (rời bỏ), thì bao nhiêu người thực sự *churn* ?

**Recall :** tỷ lệ mà mô hình bắt đúng được các trường hợp dương tính thực sự, Trong số tất cả khách hàng thật sự sẽ rời bỏ (Churn = 1), mô hình phát hiện được bao nhiêu người ?

1. **Forcasting:**

Dự đoán khách hàng mới bằng cách nhập thông tin khách hàng.

1. **Đưa ra hành động:**

Chỉ ra những yếu tố tác động tới Churn.

Đưa ra insight từ mô hình.