ĐẠI HỌC QUỐC GIA HÀ NỘI

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**



**BÁO CÁO BÀI TẬP LỚN**

**DỰ ĐOÁN KHÁCH HÀNG DỜI MẠNG VIỄN THÔNG**

**Nhóm 5:**

**Đoàn Văn Nam - 20020446**

**Môn: Khai phá dữ liệu**

**Lớp: INT 3209 50**

**GIẢNG VIÊN HƯỚNG DẪN: ThS. Vương Thị Hải Yến**

**HÀ NỘI - 2024**

**HÀ NỘI - 2021**

Mục lục

[I. Giới thiệu 3](#_Toc173712500)

[**1.** **Đặt vấn đề** 3](#_Toc173712501)

[**2.** **Mục tiêu** 3](#_Toc173712502)

[**3.** **Giải pháp** 3](#_Toc173712503)

[II. Hiểu dữ liệu 3](#_Toc173712504)

[**1.** **Import thư viện** 3](#_Toc173712505)

[**2.** **Load dữ liệu** 3](#_Toc173712506)

[**3.** **Phân tích dữ liệu** 4](#_Toc173712507)

[**3.1 Phân tích đơn biến** 5](#_Toc173712508)

[**3.2 Phân tích đa biến** 8](#_Toc173712509)

[**3.3 Trực quan hóa các thuộc tính có liên quan đến khả năng rời bỏ dịch vụ** 9](#_Toc173712510)

[III. Tiền xử lý dữ liệu 15](#_Toc173712511)

[**1.** **Xử lý thuộc tính địa lý** 15](#_Toc173712512)

[**2.** **Xử lý dữ liệu bị thiếu** 16](#_Toc173712513)

[**3.** **Hợp nhất các thuộc tính về dịch vụ Internet** 18](#_Toc173712514)

[**4.** **Bỏ đi những thuộc tính thừa** 19](#_Toc173712515)

[**5.** **Thuộc tính phân loại nhị phân** 20](#_Toc173712516)

[**6.** **Thuộc tính phân loại( >2 giá trị)** 21](#_Toc173712517)

[**7.** **Thuộc tính số nguyên** 22](#_Toc173712518)

[**8.** **Thuộc tính số thực** 22](#_Toc173712519)

[**9.** **SMOTE** 22](#_Toc173712520)

[**10.** **Lựa chọn thuộc tính** 23](#_Toc173712521)

[IV. Phân cụm 27](#_Toc173712522)

[V. Phân lớp 27](#_Toc173712523)

[VI. Học sâu 27](#_Toc173712524)

1. **Giới thiệu**
2. **Đặt vấn đề**

Giữ chân khách hàng là yếu tố then chốt cho sự thành công của một công ty, đặc biệt trong một ngành cạnh tranh như dịch vụ viễn thông. Thu hút khách hàng mới không chỉ khó khăn hơn mà còn tốn kém hơn nhiều so với việc duy trì mối quan hệ với khách hàng hiện tại.

1. **Mục tiêu**

Trong dự án này, chúng tôi sẽ xây dựng mô hình dự đoán khách hàng có ý định rời bỏ dịch vụ tại một công ty cung cấp dịch vụ viễn thông.

1. **Giải pháp**

Trước tiên, chúng tôi sẽ sử dụng phân tích dữ liệu thăm dò để hiểu các mối quan hệ giữa các đặc điểm và biến mục tiêu, và xác định các yếu tố có ảnh hưởng trong việc dự đoán sự rời bỏ của khách hàng. Sử dụng những đặc điểm này, chúng tôi sẽ phát triển một mô hình dự đoán để giúp công ty giảm tỷ lệ rời bỏ một cách chủ động và sử dụng những thông tin từ mô hình để củng cố các chiến lược giữ chân khách hàng.

1. **Hiểu dữ liệu**
2. **Import thư viện**

A computer screen shot of a program code

Description automatically generated

1. **Load dữ liệu**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

1. **Phân tích dữ liệu**

A screenshot of a computer

Description automatically generated

Dữ liệu từ tập training bao gồm 5500 bản ghi với 38 thuộc tính được chia thành hai loại: dữ liệu nhân khẩu học của khách hàng và thông tin liên quan đến dịch vụ viễn thông của họ.

Các đặc điểm nhân khẩu học bao gồm giới tính, tình trạng hôn nhân, số người phụ thuộc, và tuổi tác của khách hàng.

Các đặc điểm liên quan đến thông tin tài khoản bao gồm thời gian khách hàng đã gắn bó với dịch vụ, chi phí hàng tháng và tổng chi phí, loại hợp đồng (theo tháng, một năm, hoặc hai năm), và loại dịch vụ điện thoại, internet, TV.

Biến mục tiêu của chúng tôi cho nghiên cứu này là tình trạng khách hàng, một biến phân loại biểu thị liệu khách hàng ở lại hay đã rời dịch vụ.

**3.1 Phân tích đơn biến**

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A black rectangular object with a black border

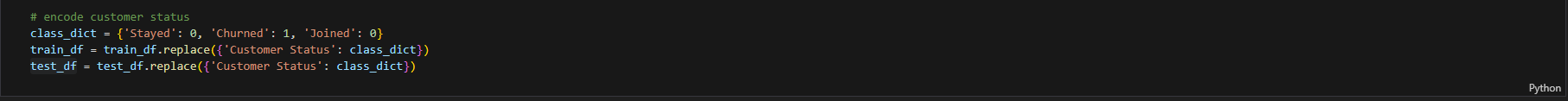
Description automatically generated

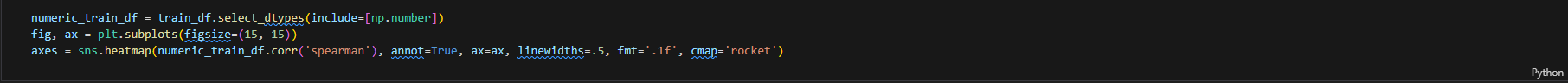
A chart of different sizes and shapes

Description automatically generated with medium confidence

Đồ thị này cung cấp một cái nhìn tổng quan về nhiều khía cạnh của dữ liệu khách hàng. Phân bố độ tuổi khá đều, với số lượng cao nhất ở nhóm 20-25. Đa số khách hàng không có người phụ thuộc. Số lượng giới thiệu thấp, với đa số khách hàng không giới thiệu ai. Thời gian sử dụng dịch vụ đa dạng, phần lớn khách hàng dùng từ 0 đến 20 tháng. Phí hàng tháng tập trung ở mức trung bình (phần đông trả 20 đến 25$ một tháng). Tổng chi phí càng cao thì số lượng khác hàng càng giảm. Phí dữ liệu phụ trội và cước gọi đường dài chủ yếu ở mức thấp với hầu hết khách hàng.

**3.2 Phân tích đa biến**



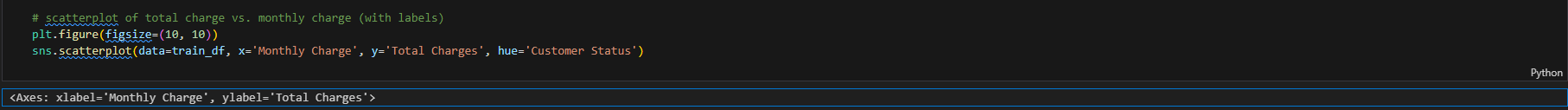


A screenshot of a computer screen

Description automatically generated

Tổng doanh thu có tương quan mạnh với tổng chi phí, thời gian sử dụng dịch vụ và cước gọi đường dài. Mã bưu điện, vĩ độ và kinh độ có mối liên hệ chặt chẽ, phản ánh vị trí địa lý. Thời gian sử dụng dịch vụ ảnh hưởng tích cực đến nhiều yếu tố như số lượng giới thiệu, chi phí hàng tháng và tổng chi phí. Thời gian ở lại tương quan âm với tình trạng khách hàng, chỉ ra rằng khách hàng càng có thâm niên thì càng ít có khả năng rời mạng.

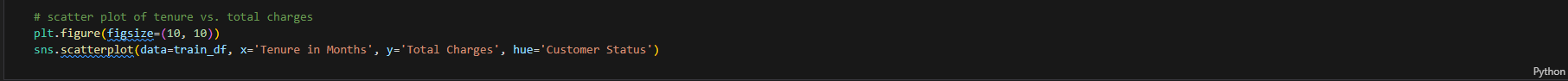
**3.3 Trực quan hóa các thuộc tính có liên quan đến khả năng rời bỏ dịch vụ**



A chart of a diagram

Description automatically generated with medium confidence

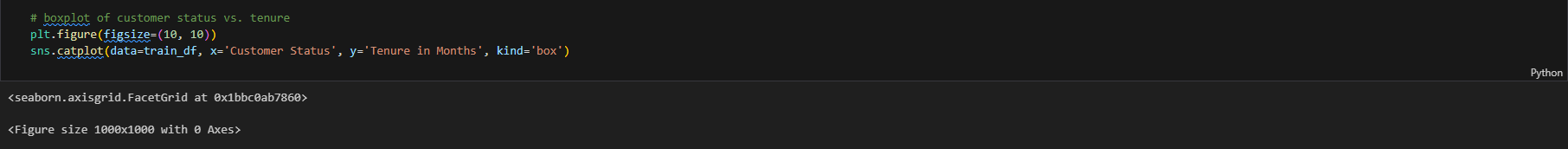
Khách hàng rời đi (màu cam) có xu hướng tập trung nhiều hơn ở vùng có phí hàng tháng cao hơn, đặc biệt là trong khoảng từ 60 đến 120. Khách hàng ở lại (màu xanh) phân bố đều hơn trên toàn bộ phạm vi phí hàng tháng, nhưng có vẻ chiếm ưu thế ở vùng phí thấp và trung bình. Tỷ lệ khách hàng rời đi có vẻ tăng lên khi phí hàng tháng tăng, thể hiện qua mật độ điểm màu cam cao hơn ở phía phải đồ thị.



A graph of a diagram

Description automatically generated with medium confidence

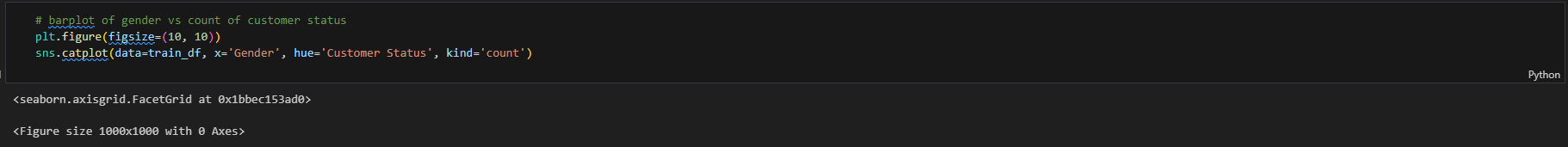
Khách hàng ở lại (màu xanh) có xu hướng phân bố rộng hơn trên toàn bộ biểu đồ, cho thấy sự đa dạng trong thời gian gắn bó và tổng chi phí. Ngược lại, khách hàng rời đi (màu cam) tập trung nhiều hơn ở vùng có thời gian gắn bó ngắn hơn và tổng chi phí thấp hơn. Đáng chú ý là có một số nhóm điểm dữ liệu song song, cho thấy có thể có các gói dịch vụ hoặc mức giá khác nhau. Nhìn chung, khách hàng có thời gian gắn bó càng lâu thì tổng chi phí càng cao, nhưng nguy cơ rời đi có vẻ cao hơn trong giai đoạn đầu của dịch vụ.



A graph of blue squares

Description automatically generated with medium confidence

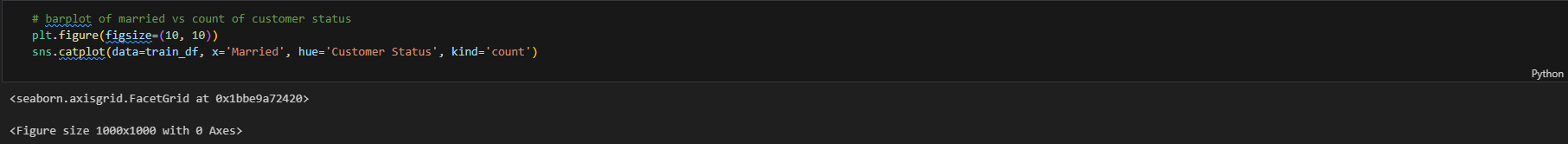
Khách hàng ở lại có thời gian gắn bó trung bình cao hơn (khoảng 37 tháng) và phạm vi rộng hơn (từ 15 đến 60 tháng). Ngược lại, khách hàng rời đi có thời gian gắn bó trung bình ngắn hơn nhiều (khoảng 10 tháng) và tập trung chủ yếu trong khoảng 3 đến 30 tháng. Điều này cho thấy khách hàng có khả năng rời đi cao nhất trong những tháng đầu. Tuy nhiên, vẫn có một số khách hàng rời đi sau thời gian dài gắn bó (thể hiện qua các điểm ngoại lệ).



A graph of a bar chart

Description automatically generated with medium confidence

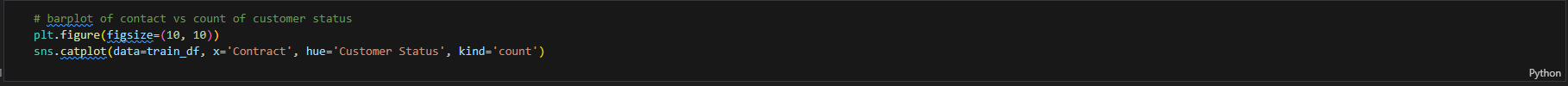
Số lượng khách hàng nam và nữ khá cân bằng, với nam giới có chút ít nhiều hơn. Tỷ lệ khách hàng rời đi so với ở lại cũng tương đối đồng đều giữa hai giới.



A graph of a customer service

Description automatically generated with medium confidence

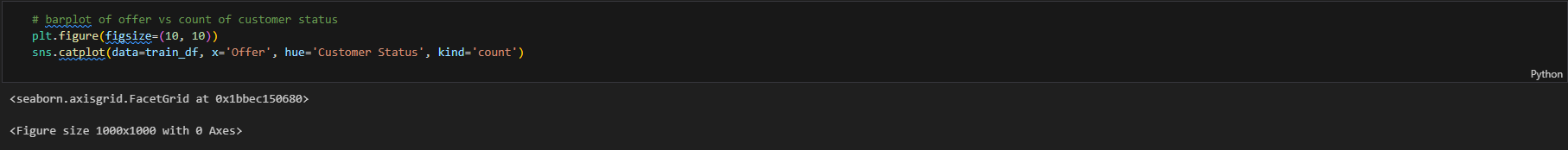
Đối với khách hàng đã kết hôn, số lượng khách hàng ở lại (màu xanh) cao hơn đáng kể so với khách hàng rời đi (màu cam). Tuy nhiên, đối với khách hàng chưa kết hôn, sự chênh lệch này ít rõ rệt hơn, với tỷ lệ khách hàng rời đi cao hơn. Điều này gợi ý rằng khách hàng đã kết hôn có xu hướng trung thành hơn với dịch vụ.



A graph of blue and orange bars

Description automatically generated

Đối với hợp đồng Month-to-Month, tỷ lệ khách hàng rời đi (màu cam) cao nhất và gần bằng số khách hàng ở lại (màu xanh), cho thấy đây là nhóm có nguy cơ mất khách hàng cao nhất. Ngược lại, hợp đồng One Year và Two Year có tỷ lệ khách hàng ở lại rất cao, đặc biệt là Two Year gần như không có khách hàng rời đi. Điều này chỉ ra rằng các hợp đồng dài hạn hiệu quả hơn trong việc giữ chân khách hàng.



A graph of different colored bars

Description automatically generated

Biểu đồ cột này so sánh số lượng khách hàng theo các loại ưu đãi (Offer) khác nhau. Offer B có số lượng khách hàng ở lại (màu xanh) cao nhất và tỷ lệ khách hàng rời đi thấp, cho thấy đây là ưu đãi hiệu quả nhất trong việc giữ chân khách hàng. Ngược lại, Offer E có số lượng khách hàng rời đi (màu cam) cao hơn số khách hàng ở lại, chỉ ra rằng đây là ưu đãi kém hiệu quả nhất. Offer A, C và D có tỷ lệ giữ chân khách hàng tốt, với số lượng khách hàng ở lại vượt trội so với số rời đi

1. **Tiền xử lý dữ liệu**
2. **Xử lý thuộc tính địa lý**

Nối bảng train với bảng dữ liệu về zip code và mật độ dân số

A screenshot of a computer program

Description automatically generated

1. **Xử lý dữ liệu bị thiếu**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

1. **Hợp nhất các thuộc tính về dịch vụ Internet**

**A screenshot of a computer program

Description automatically generated**

1. **Bỏ đi những thuộc tính thừa**

**A black screen with a black rectangle

Description automatically generated with medium confidence**

**A screenshot of a computer program

Description automatically generated**

1. **Thuộc tính phân loại nhị phân**

**A screenshot of a computer program

Description automatically generated**

1. **Thuộc tính phân loại( >2 giá trị)**

**A black screen with green text

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screen shot of a computer

Description automatically generated**

1. **Thuộc tính số nguyên**

**A screenshot of a computer

Description automatically generated**

1. **Thuộc tính số thực**

**A screenshot of a computer

Description automatically generated**

1. **SMOTE**

**A screen shot of a computer code

Description automatically generated**

1. **Lựa chọn thuộc tính**

**A screenshot of a computer

Description automatically generated**

**A computer screen shot of text

Description automatically generated**

**Correlation**

**A black rectangular object with a black stripe

Description automatically generated**

****

****

**Annova**

**A screen shot of a computer program

Description automatically generated**

**A screenshot of a graph

Description automatically generated**

**Mutual Information**

**A screenshot of a computer

Description automatically generated**

**Random Forest feature**

**A screenshot of a computer screen

Description automatically generated**

1. **Phân cụm**
2. **Phân lớp**
3. **Học sâu**