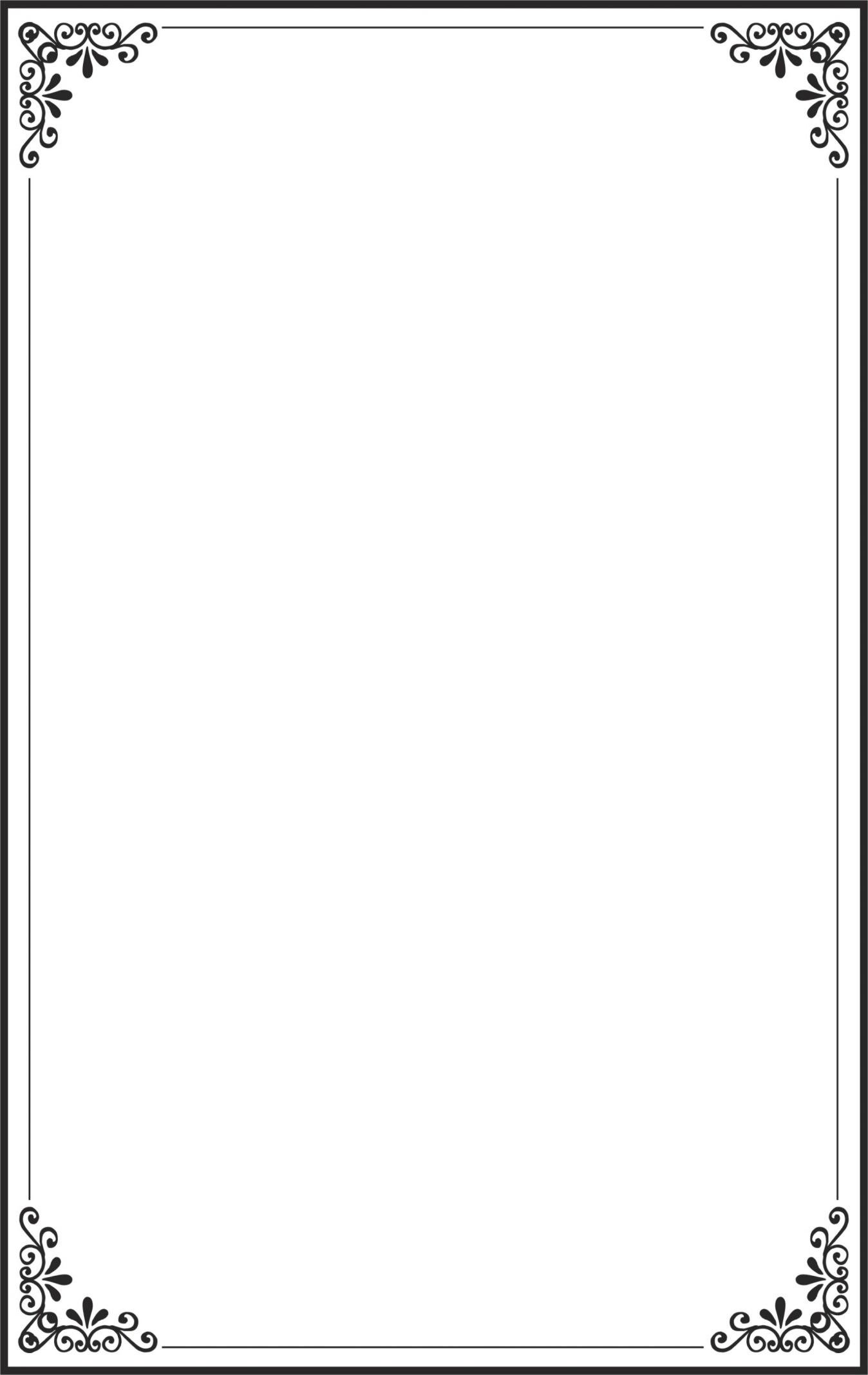
**Trường Đại học Kinh tế Quốc dân**

**Khoa Toán Kinh Tế**

**🖎🕮✍**



**Đề án môn học – Toán kinh tế**

**Đề tài:** Health Insurance Cross-Selling Prediction

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | TS. Nguyễn Thị Quỳnh Giang |
| **Họ và tên:** | Trương Ngọc Thuỳ Trang |
| **Mã sinh viên:** | 11217004 |

Hà Nội - 2024

Mục Lục

[I. Giới thiệu mở đầu 3](#_Toc165748307)

[**1.1.** **Đề tài nghiên cứu** 3](#_Toc165748308)

[**1.2.** **Lý do chọn đề tài** 3](#_Toc165748309)

[**1.3.** **Vấn đề chung của đề tài** 3](#_Toc165748310)

[**1.4.** **Nhiệm vụ nghiên cứu đề tài** 3](#_Toc165748311)

[II. Nội dung 5](#_Toc165748312)

[**2.1.** **Lý thuyết và các phương pháp luận** 5](#_Toc165748313)

[a. Cross-selling 5](#_Toc165748314)

[b. Cross-selling in health insurance 8](#_Toc165748315)

[**2.2.** **Data Mining** 10](#_Toc165748316)

[a. Mô tả dữ liệu (Data description) 10](#_Toc165748317)

[b. Xử lý dữ liệu 11](#_Toc165748318)

[c. Phân tích và trực quan hoá dữ liệu 13](#_Toc165748319)

[**2.3.** **Xây dụng mô hình và phân tích kết quả** 16](#_Toc165748320)

[a. Logistic regression 16](#_Toc165748321)

[b. Decision Tree 17](#_Toc165748322)

[c. Random Forest 18](#_Toc165748323)

[d. Đánh giá độ chính xác và lựa chọn mô hình phù hợp 19](#_Toc165748324)

[III. Kết luận 21](#_Toc165748325)

[Tài liệu tham khảo 22](#_Toc165748326)

# Giới thiệu mở đầu

* 1. **Đề tài nghiên cứu**

Cross-selling (Bán chéo) là một chiến lược kinh doanh phổ biến hiện nay được sử dụng trong nhiều ngành công nghiệp, dịch vụ. Nó bao gồm việc đề xuất các sản phẩm hoặc dịch vụ bổ sung cho khách hàng hiện tại dựa nhu cầu và hành vi mua hàng của họ.

Trong bài nghiên cứu này sẽ đề cập tới cross-selling trong lĩnh vực bảo hiểm. Health Insurance Cross-selling là một quá trình bổ sung các sản phẩm hoặc dịch vụ vào các giao dịch mua bảo hiểm sức khoẻ. Thông qua việc tận dụng dữ liệu và thông tin khách hàng, các công ty bảo hiểm có thể đề xuất các sản phẩm đi kèm như bảo hiểm xe cộ, bảo hiểm tai nạn, bảo hiểm nhân thọ,…

* 1. **Lý do chọn đề tài**

Đi kèm với sự phát triển kinh tế, hàng loạt các doanh nghiệp cũng tìm cách để tối ưu trong việc bán hàng. Những năm gần đây việc áp dụng chiến lược bán kèm sản phẩm được rất nhiều doanh nghiệp hưởng ứng. Không chỉ giúp tăng doanh số bán hàng mà còn tận dụng được nguồn khách hàng hiện tại, nâng cao chất lượng dịch vụ mà tiết kiệm được chi phí quảng cáo.

Đối với lĩnh vực bảo hiểm, cross-selling giúp cải thiện tỷ lệ giữ chân khách hàng và tăng lợi nhuận bằng cách cung cấp các gói bảo hiểm toàn diện, đa dạng hơn mà vẫn đáp ứng nhu cầu của khách hàng. Thêm vào đó, bán kèm sản phẩm cũng giúp tăng cường mối quan hệ với khách hàng thông qua việc cung cấp các dịch vụ tư vấn phù hợp với nhu cầu cụ thể của họ, góp phần vào sự hài lòng của khách hàng cũng như danh tiếng của doanh nghiệp.

* 1. **Vấn đề chung của đề tài**

Vấn đề đặt ra là làm thế nào để hiệu quả việc đề xuất và bán các sản phẩm hoặc dịch vụ đi kèm cho khách hàng. Điều này đòi hỏi sự hiểu biết sâu sắc về nhu cầu, mong muốn và hành vi mua hàng của từng khách hàng hoặc một nhóm khách hàng để có thể đề xuất những giải pháp phù hợp. Thêm vào đó, cần có các chiến lược và quy trình phù hợp để tích hợp bán kèm sản phẩm vào quá trình kinh doanh một cách hiệu quả và hợp lí, đồng thời vẫn đảm bảo sẽ mang lại giá trị thực sự cho doanh nghiệp cũng như khách hàng.

* 1. **Nhiệm vụ nghiên cứu đề tài**

Từ vấn đề chung được nêu trên mà mỗi doanh nghiệp sẽ đưa ra chiến lược nghiên cứu sao cho phù hợp với việc kinh doanh của mình. Với các doanh nghiệp bảo hiểm, cụ thể là Triks Insurance Inc., một Đại lý Bảo hiểm Nhân thọ nổi tiếng, có cơ sở khách hàng lớn gồm 381.109 hộ gia đình trong khu vực. Họ đã đàm phán với nhà cung cấp Bảo hiểm ô tô nổi tiếng, GOOL, với kế hoạch giới thiệu sản phẩm bảo hiểm của họ. Công ty bảo hiểm ô tô GOOL đã yêu cầu Bảo hiểm Triks cung cấp một báo cáo cho biết tỷ lệ cơ sở khách hàng của họ có khả năng quan tâm đến bảo hiểm ô tô là bao nhiêu.

Thay vì bắt tay vào nhiệm vụ khó khăn là sàng lọc cơ sở dữ liệu khách hàng theo cách thủ công, bài nghiên cứu này sẽ đề xuất các phương pháp phân loại, hay hệ thống học máy. Mục đích của các phương pháp và hệ thống này không chỉ là dự đoán các ứng viên quan tâm đến bảo hiểm ô tô tiềm năng một lần mà còn liên tục theo thời gian. Bằng cách hiểu rõ cơ sở khách hàng và tận dụng thông tin nhân khẩu học về khách hàng tiềm năng, sẽ có thể đánh giá xem khách hàng tiềm năng được phân loại vào nhóm “quan tâm” hay “không quan tâm” đến bảo hiểm ô tô.

Bằng cách triển khai giải pháp này, đội ngũ bán hàng của Bảo hiểm Triks sẽ được chuẩn bị tốt hơn cho nỗ lực bán kèm của họ. Họ có thể trau dồi sự tập trung vào những khách hàng tiềm năng hứa hẹn nhất, từ đó tối đa hóa hiệu quả ngân sách tiếp thị trong khi tập trung vào những khách hàng tiềm năng sinh lợi nhất.

1. **Nội dung**
   1. **Lý thuyết và các phương pháp luận**
2. Cross-selling

Bán chéo (Cross-selling) là một chiến lược bán các sản phẩm hoặc dịch vụ có liên quan với nhau cho khách hàng bằng cách phân tích xu hướng mua hàng chung và các đặc điểm cần thiết. Cross-selling trở nên phổ biến trong những năm gần đây và ngày càng có tầm quan trọng trong việc phân tích và phát triển sản phẩm, dịch vụ.

Nhìn lại quá khứ, các nghiên cứu sơ khai về bán chéo của Hebden & Pickering (1974) tập trung vào phương pháp xác định các mô hình chung trong việc mua sản phẩm của khách hàng dựa trên dữ liệu về quyền sở hữu hoặc cách sử dụng. Kamakura và cộng sự của mình (1991) mô tả và dự đoán quá trình mua hàng dựa trên mức độ ảnh hưởng về vấn đề tài chính của khách hàng và mức độ e ngại khi mua các sản phẩm dịch vụ ( như độ rủi ro, tính thanh khoản, hữu dụng của sản phẩm,...). Đến năm 2007, nghiên cứu của Kamakura cho rằng người tiêu dùng phải cân bằng nhiều mục tiêu tài chính trong suốt vòng đời của mình. Ở giai đoạn đầu, người tiêu dùng tài trợ cho tiêu dùng hiện tại bằng thu nhập trong tương lai thông qua tín dụng. Khi trưởng thành, thu nhập của họ nhiều hơn so với tiêu dùng, do đó có thể tài trợ cho việc tiêu dùng trong quá khứ bằng việc trả khoản vay và trong tương lai thông qua tiết kiệm hoặc đầu tư. Kamakura đã dựa vào hành vi tập thể này kết hợp cùng với tính năng của sản phẩm để xây dựng mô hình thống kê nhằm mục đích bán chéo nhắm đúng nhu cầu của khách hàng. Bên cạnh đó, ông cùng các cộng sự phát triển mô hình rủi ro phân chia đa biến để dự đoán xác suất và thời gian mua sản phẩm mới.

Ansell (2007) kết hợp phân khúc lối sống của khách hàng và mô hình tỷ lệ rủi ro để xác định cơ hội bán chéo thông qua việc sử dụng biểu đồ nhân khẩu học và năm lần mua hàng đầu tiên của khách hàng. Cũng sử dụng dữ liệu nhân khẩu học, nhưng [Ahn (2011)](https://www.sciencedirect.com/science/article/abs/pii/S0957417410011073) kết hợp cùng với hành vi tập thể và thuật toán di truyền để tìm giải pháp bán chéo trong ngành viễn thông. Họ đã đề xuất mô hình phân loại khách hàng dựa trên các dữ liệu tích lũy từ hệ thống CRM (một hệ thống phổ biến hỗ trợ doanh nghiệp trong việc quản lý thông tin của khách hàng chẳng hạn như nhân khẩu học) và mô hình sử dụng sản phẩm hoặc dịch vụ cũ để tìm sản phẩm và dịch vụ mới có tiềm năng bán hàng cao. Các kỹ thuật khai thác dữ liệu khác nhau được áp dụng cho mô hình đề xuất theo hai bước. Ở bước đầu tiên, một số kỹ thuật phân loại như hồi quy logistic, mạng lưới thần kinh nhân tạo (Artificial neural network-ANN) và cây quyết định (Decision Trees) được áp dụng độc lập để dự đoán việc mua sản phẩm mới và mỗi mô hình tạo ra kết quả dự đoán của chúng dưới dạng xác suất. Ở bước thứ hai, xây dựng mô hình cân bằng tất cả các xác suất này bằng cách sử dụng thuật toán di truyền (Genetic algorithm-GA) và đưa ra quyết định cuối cùng cho khách hàng mục tiêu xem họ có mua sản phẩm mới hay không. Để xác thực tính hiệu quả của mô hình, họ đã áp dụng nó vào trường hợp của một công ty viễn thông di động thực tế ở Hàn Quốc. Kết quả là, Ahn và những cộng sự nhận thấy mô hình của họ tạo ra thông tin chất lượng cao để bán kèm và thuật toán di truyền ở bước thứ hai đã góp phần cải thiện đáng kể hiệu suất.

Li và cộng sự (2011) đã áp dụng mô hình probit đa biến để đánh giá và dự đoán phản hồi của khách hàng với các đề xuất bán kèm. Để nắm bắt nhu cầu của khách hàng đối với các sản phẩm khác nhau bị chi phối bởi yếu tố nào, từ đó phân lớp khách hàng với các đặc điểm khác nhau và khả năng đáp ứng các yêu cầu bán kèm cho các sản phẩm khác nhau. Đặc điểm tính cách của khách hàng có tiềm năng trở thành thông tin hữu ích cho mô hình bán chéo sản phẩm tài chính vì theo Brown và Taylor (2014) một số đặc điểm tính cách có mối liên hệ có ý nghĩa thống kê với số nợ không đảm bảo và tài sản tài chính mà hộ gia đình nắm giữ. Họ cũng phát hiện ra rằng các đặc điểm tính cách có mối liên hệ khác nhau với các loại nợ khác nhau. Bởi vì họ cho rằng các đặc điểm tính cách liên quan mật thiết tới khẩu vị rủi ro của cá nhân khách hàng. Với khách hàng có khẩu vị rủi ro thấp sẽ có khả năng đi vay hay đầu tư vào các tài sản tài chính thấp hơn, và ngược lại. Điều này có tác động đến chiến lược tài chính của khách hàng và do đó tác động đến xác suất họ quyết định mua một sản phẩm tài chính nhất định. Từ đó thêm vào bộ dữ liệu để phân lớp khách hàng theo mục tiêu bán hàng của doanh nghiệp.

Cho đến gần đây, các mô hình thống kê như hồi quy logistic và phân tích khác biệt (Discriminant Analysis) đã thống trị mô hình phản ứng của người tiêu dùng đối với bán chéo sản phẩm. Mặc dù các mô hình thống kê có thể rất mạnh mẽ nhưng chúng đưa ra các giả định mạnh nhất về loại dữ liệu cũng như phân phối xác suất của chúng và thường chỉ có thể xử lý một số lượng biến giới hạn. Vậy nên các nhà nghiên cứu đã tìm racác thuật toán học máy phổ biến nhất được sử dụng trong bán chéo (cross-selling) là k-Nearest Neighbor (KNN), Mạng Bayesian, Chi-squared Automatic Interaction Detector (CHAID), Cây quyết định (Decision Trees), Random Forest, Extreme Gradient Boosting (XGBoost),...

 Random Forest thường được coi là công cụ phân loại rất mạnh như trong nghiên cứu của Brown và Mues (2012) và Kruppa (2013). Random Forest là một thuật toán học máy thuộc lớp Ensemble Learning nơi nhiều mô hình được kết hợp để tạo ra một mô hình mạnh mẽ hơn. RF sử dụng nhiều cây quyết định (Decision Trees) để đưa ra dự đoán. Trong một số bài toán phân loại, có thể xảy ra tình trạng mất cân bằng lớn giữa các lớp, nghĩa là một hoặc vài lớp có số lượng mẫu nhiều hơn đáng kể so với các lớp khác. Điều này có thể tạo ra vấn đề khi mô hình có thể không học được một cách hiệu quả từ các lớp thiếu dữ liệu. Vậy nên so sánh với các thuật toán phân loại tập hợp không đồng nhất cho thấy các phương pháp như vậy cải thiện hơn RF. Thuật toán XGBoost có thể được điều chỉnh để xử lý mất cần bằng lớp trong các bài toán phân loại. Thuật toán học máy này dựa trên kỹ thuật Gradient Boosting, có khả năng chống overfitting tốt, xử lý tốt các dữ liệu lớn, và khả năng ổn định trong việc tối ưu hoá các hàm mất mát.

Theo Karatzoglou và cộng sự (2016), Deep Learning là một trong những bước tiến lớn tiếp theo trong công nghệ Hệ thống gợi ý (Recommender Systems), tuy nhiên việc sử dụng Deep Learning trong tiếp thị trực tiếp còn hạn chế. Vài năm qua đã chứng kiến thành quả to lớn từ nghiên cứu của mạng lưới thần kinh sâu trong một số nhiệm vụ phức tạp như thị giác máy tính, xử lý ngôn ngữ tự nhiên và nhận dạng giọng nói. Một số nghiên cứu đã được mô hình hoá về phương pháp Deep Learning cho Hệ thống gợi ý. Van den Oord và cộng sự (2013) đã chỉ ra rằng những tiến bộ gần đây trong Deep Learning chuyển dịch rất tốt sang đề xuất âm nhạc, với mạng lưới thần kinh tích chập sâu vượt trội hơn đáng kể so với cách tiếp cận truyền thống là sử dụng cách biểu diễn tín hiệu âm thanh bằng từ ngữ. Kết quả của họ chỉ ra rằng nghiên cứu trong lĩnh vực này có thể mang lại lợi ích đáng kể từ việc sử dụng mạng lưới thần kinh sâu. Wangperawong và cộng sự (2016) đã sử dụng dữ liệu hành vi theo thời gian của khách hàng và biểu diễn dữ liệu đó dưới dạng hình ảnh để thực hiện dự đoán tỷ lệ rời bỏ bằng cách tận dụng các cấu ​​trúc Deep Learning nổi bật trong phân loại hình ảnh. Mạng nơ-ron tích chập sâu và bộ mã hóa tự động cho thấy hiệu quả trong việc dự đoán và hiểu rõ tình trạng xáo trộn trong ngành viễn thông, vượt trội so với các mô hình đơn giản khác như mô hình cây quyết định.

Hệ thống gợi ý (Recommender Systems – RS) được ứng dụng rất thành công trong dự đoán hành vi của người dùng dựa vào hành vi của họ trong quá khứ. RS đang được ứng dụng trong rất nhiều lĩnh vực khác nhau như thương mại điện tử (hỗ trợ bán hàng trực tuyến), giải trí (gợi ý phim ảnh, bài hát,..), giáo dục đào tạo (gợi ý nguồn tài nguyên học tập, nghiên cứu,..). Ví dụ, trong hệ thống bán hàng trực tuyến (chẳng hạn như Amazon), nhằm tối ưu hóa khả năng mua sắm của khách hàng, người ta quan tâm đến việc những khách hàng nào đã “yêu thích” những sản phẩm nào bằng cách dựa vào dữ liệu quá khứ của họ (dữ liệu này có thể là xếp hạng mà người dùng đã bình chọn trên sản phẩm, thời gian duyệt (browser) trên sản phẩm, số lần click chuột trên sản phẩm,..). Từ đó hệ thống sẽ dự đoán được người dùng có thể thích sản phẩm nào và đưa ra những gợi ý phù hợp cho họ.

Hiện tại, hệ thống gợi ý có rất nhiều giải thuật được đề xuất, tuy nhiên có thể tổng hợp vào vào các nhóm chính sau:

* *Nhóm giải thuật lọc cộng tác (Collaborative Filtering)*: trong nhóm này, các giải thuật chủ yếu sử dụng
* *Phương pháp láng giềng (Neighborhood-based, còn gọi là Memory Based),* dựa trên dữ liệu quá khứ của người dùng (user-based approach), hoặc là dựa trên dữ liệu quá khứ của những sản phẩm “tương tự” (item-based approach)
* *Dựa trên mô hình (Model-based)*: Nhóm này liên quan đến việc xây dựng các mô hình dự đoán dựa trên dữ liệu thu thập được trong quá khứ. Như mô hình Bayesian, các mô hình nhân tố tiềm ẩn (latent factor models); trong đó kỹ thuật phân rã ma trận (matrix factorization) là một điển hình.
* *Nhóm giải thuật lọc trên nội dung (Content-based Filtering):* Gợi ý các item dựa vào hồ sơ của người dùng hoặc dựa vào nội dung, thuộc tính của những sản phẩm tương tự như sản phẩm mà người dùng đã chọn trong quá khứ.
* *Nhóm lai ghép*: Kết hợp cả 2 cách trên.
* *Nhóm kỹ thuật không cá nhân hóa (non-personalization)*

Một trong những khuyết điểm của phương pháp lọc theo nội dung là khó khăn trong việc thu thập thông tin, trong khi phần lớn các mô hình dựa trên lọc cộng tác chỉ cần 3 thông tin (mã người dùng, mã sản phẩm, phản hồi của người dùng) là có thể hoạt động tương đối tốt. Do vậy khuynh hướng hiện nay đa phần các nhà nghiên cứu thiên về nhóm lọc cộng tác.

Nhìn chung cho đến nay, qua kết quả của các bài nghiên cứu, ứng dụng học máy để giải quyết vấn đề về bán chéo sản phẩm vẫn được ưa chuộng và cho kết quả khả quan hơn.

1. Cross-selling in health insurance

Bán chéo (cross-selling) là một chiến lược bán hàng được sử dụng để khuyến khích khách hàng tăng chi tiêu bằng cách mua sản phẩm bổ sung cho giao dịch mua hiện tại của họ. Cách tiếp cận này hữu ích trong ngành bảo hiểm, đóng vai trò là phương tiện để các công ty giới thiệu sản phẩm mới tới cơ sở khách hàng hiện tại của họ.

Trước khi xuất hiện các thuật toán machine learning, các công ty bảo hiểm thường áp dụng các chiến lược bán chéo sản phẩm thông qua các phương tiện truyền thống và các phương pháp phân tích dữ liệu cổ điển (nghiên cứu thị trường, phân loại khách hàng dựa trên thông tin nhu cầu cơ bản, tiếp thị,...). Nhược điểm của phương pháp này thường khá mất thời gian, có những biến số cần cho mô hình nhưng không thu thập được,...Cho đến khi sử dụng các công cụ học máy đã mang lại khả năng linh hoạt và hiệu quả cao hơn trong việc quản lý sản phẩm và chiến lược bán chéo trong lĩnh vực bảo hiểm nhân thọ.

Ứng dụng của các công cụ học máy trong lĩnh vực bảo hiểm nhân thọ có thể theo hai hướng. Đầu tiên là việc áp dụng hệ thống gợi ý, dựa trên hành vi mua sản phẩm, xếp hạng sản phẩm, số lần click vào sản phẩm,... xây dựng hệ thống đưa ra gợi ý về những sản phẩm tương tự như trong quá khứ. Cách tiếp cận thứ hai là từ bộ dữ liệu có sẵn về thông tin khách hàng dự đoán xem liệu những khách đó sẽ mua sản phẩm bảo hiểm khác hay không. Trong lĩnh vực bảo hiểm việc thu thập dữ liệu để xây dựng hệ thống gợi ý từ một trang web của công ty bảo hiểm nào đó khá khó khăn, vây nên bài nghiên cứu này sẽ chỉ đề cập đến cách tiếp cận thứ hai.

Trong bài nghiên cứu này, phân tích một công ty bán bảo hiểm nhân thọ dựa vào cơ sở dữ liệu khách hàng để dự đoán tỷ lệ khách hàng có khả năng quan tâm đến bảo hiểm ô tô. Với hướng tiếp cận thông dụng nhất là sử dụng mô hình Logistic phân biệt nhóm khách hàng quan tâm đến bảo hiểm ô tô dựa trên bộ dữ liệu về nhân khẩu học và mua bảo hiểm nhân thọ. Nhưng như đã đề cập, mô hình này đòi hỏi chặt chẽ về các giả định của dữ liệu, phân phối của chúng. Vậy nên việc sử dụng học máy sẽ giúp giảm đi ràng buộc về các giả định. Một thuật toán phân loại và hồi quy dựa trên việc đánh giá các điểm dữ liệu xung quanh một điểm cụ thể để đưa ra dự đoán là KNN. KNN là một thuật toán đơn giản, tuy nhiên hiệu suất của nó có thể bị ảnh hưởng bởi việc chọn giá trị k và cần quan tâm đến độ phức tạp tính toán khi tập dữ liệu lớn. Để đáp ứng được điều này phương pháp Cây quyết định (Decision Trees) và được mở rộng thành các thuật toán Random Forest, Gradient Boost sẽ được áp dụng để cải thiện hiệu suất.

Sau khi xây dựng mô hình dựa vào các thuật toán đã nêu ở trên, tiến hành so sánh đánh giá hiệu suất mô hình theo hai phương pháp Accuracy và AUC (Area Under the Receiver Operating Characteristic Curve).

* 1. **Data Mining**
  2. Mô tả dữ liệu (Data description)

Khi đi vào nghiên cứu, việc đầu tiên cần là xác định các thông tin dữ liệu cần thu thập. Điều này đòi hỏi phải có sự hiểu biết về lĩnh vực mà doanh nghiệp đang hoạt động.

Bộ dữ liệu dùng để nghiên cứu được thu thập từ khách hàng của một công ty bán bảo hiểm y tế. Dữ liệu có 381109 quan sát tương ứng với 381109 khách hàng có 12 đặc trưng mô tả dưới bảng sau:

|  |  |
| --- | --- |
| **1.ID** | Mã định danh duy nhất của khách hàng |
| **2.Gender** | Giới tính |
| **3.Age** | Tuổi của khách hàng |
| **4.Driving license** | 1:Nếu khách hàng có bằng lái xe; 0:ngược lại |
| **5.Region code** | Mã vùng khách hàng sinh sống |
| **6.Previously insured** | 1:Nếu khách hàng đã có bảo hiểm xe; 0:ngược lại |
| **7.Vehicle age** | Tuổi phương tiện của khách hàng |
| **8.Vehice damage** | 1:Nếu khách hàng đã hỏng xe trong quá khứ; 0:ngược lại |
| **9.Annual Premium** | Số tiền khách hàng phải đóng phí bảo hiểm trong năm |
| **10.Policy sales channel** | Mã ẩn danh cho kênh tiếp cận tới khách hàng |
| **11.Vintage** | Số ngày khách hàng gắn bó với công ty |
| **12.Response** | 1:Nếu khách hàng quan tâm tới bảo hiểm xe; 0:ngược lại |

Với bộ dữ liệu trên, biến chính cần dự đoán là “Response”. Thông qua việc xử lý, phân tích chọn ra các biến để xây dựng mô hình dự đoán sao cho tỷ lệ dự báo đúng là cao nhất.

* 1. Xử lý dữ liệu

Trong 11 biến (ngoại trừ ID) có 3 biến định lượng và 8 biến là định tính. Tiến hành chuyển các biến định tính về dạng factor trong R sau đó thực hiện kiểm tra các giá trị NAs cũng như các giá trị outlier của 3 biến định lượng.

* Thống kê mô tả với biến định lượng

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Min** | **1st Qu.** | **Mean** | **3rd Qu.** | **Max** | **NAs** |
| **Age** | 20 | 25 | 38.82 | 49 | 85 | 0 |
| **Annual\_premium** | 2630 | 24405 | 30564 | 39400 | 540165 | 0 |
| **Vintage** | 10 | 82 | 154.3 | 227 | 299 | 0 |

Kết quả cho thấy không có giá trị NAs nào ở 3 biến định lượng trong bộ dữ liệu này. Trung bình tuổi của khách hàng mua bảo hiểm của Triks xấp xỉ 39 tuổi, mang lại số tiền trung bình là 30564 với hơn 154 ngày trung bình khách hàng gắn bó với công ty. Tuy nhiên nhận thấy sự chênh lệch lớn giữa các giá trị min, max so với median nên kiểm tra outlier

|  |  |
| --- | --- |
| **Trước khi xử lý outlier** | **Sau khi sử lý outlier** |
| A graph with a blue rectangle  Description automatically generated | A graph with a blue rectangle  Description automatically generated |
| A graph with a line graph  Description automatically generated | A green rectangular bar with numbers  Description automatically generated |
| A black square with white text  Description automatically generated | A black square with white text  Description automatically generated |

Hai biến tuổi của khách hàng và số ngày khách hàng đã gắn bó với công ty đều không có outlier. Trong khi số tiền chi trả cho bảo hiểm trong một năm của khách hàng lại xuất hiện outlier. Sau khi tìm kiếm các outlier theo tiêu chuẩn IQR sẽ thay thế các giá trị outlier đó bằng giá trị trung bình của biến Annual premium.

* Thống kê mô tả với biến định tính

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Count** | **Frequency** |
| **Gender** | Female | 175020 | 45.92% |
|  | Male | 206089 | 54.08% |
| **Driving License** | No | 812 | 0.21% |
|  | Yes | 380297 | 99.79% |
| **Previously Insured** | No | 206481 | 54.18% |
|  | Yes | 174628 | 45.82% |
| **Vehicle age** | <1 year | 164786 | 43.24% |
|  | 1-2 years | 200316 | 52.56% |
|  | >2 years | 16007 | 4.20% |
| **Vehicle damage** | No | 188696 | 49.51% |
|  | Yes | 192413 | 50.49% |
| **Response** | No | 334399 | 87.74% |
|  | Yes | 46710 | 12.26% |

Với dữ liệu biến định tính, sử dụng phương pháp One-Hot Encoding (Biến đổi các biến định tính thành các biến giả mà mỗi biến giả chỉ nhận hai giá trị 0 hoặc 1) cho các biến Gender, Previously insured, Vehicle damage và phương pháp Label Encoding cho các biến Region code, Vehicle age, Policy sales channel.

* 1. Phân tích và trực quan hoá dữ liệu

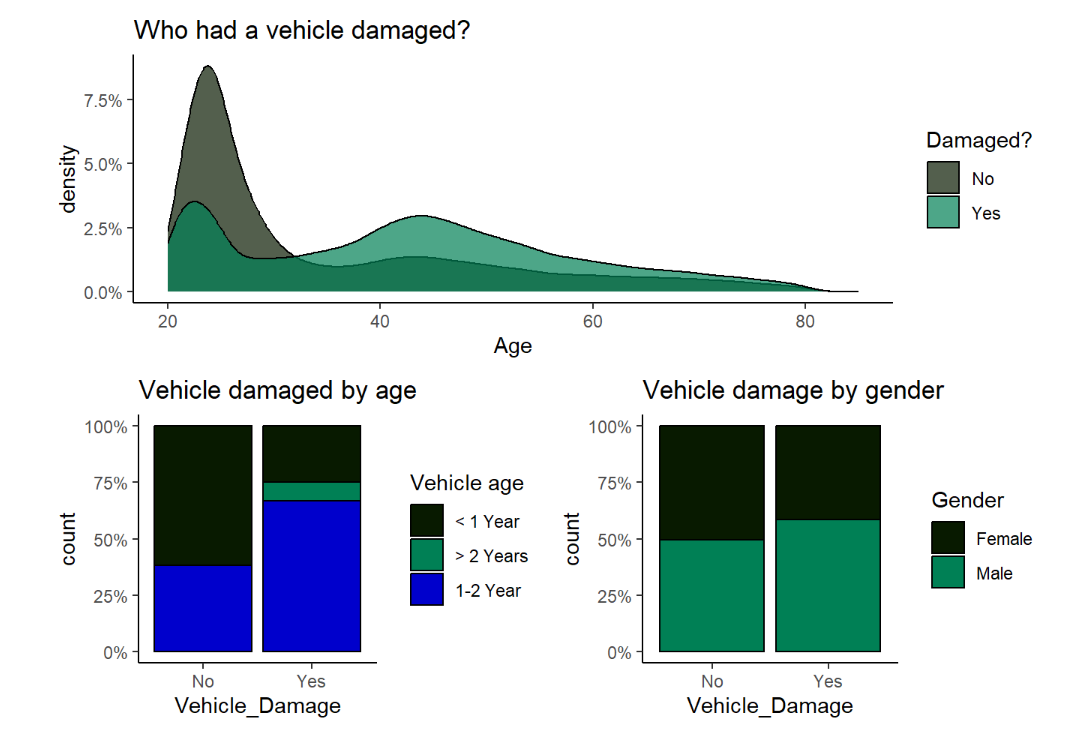
A graph of a number of circles

Description automatically generated with medium confidence

*Bảng : Tỷ lệ Response*

|  |  |
| --- | --- |
| A graph of a customer response  Description automatically generated | Có 334399 khách hàng trong bộ dữ liệu được thu thập không quan tâm tới bảo hiểm xe ô tô, gấp hơn 7 lần so với với 46710 khách hàng là quan tâm tới bảo hiểm xe ô tô |
| A graph of a number of green squares  Description automatically generated | Với những khách hàng đã từng bị hỏng xe thì tỷ lệ họ quan tâm tới bảo hiểm xe ô tô là chắc chắn. Trong khi với những người chưa từng hỏng xe thì tỷ lệ quan tâm chỉ chiếm chưa tới 50% |
| A graph of a response by gender  Description automatically generated | Khi phân theo giới tính, có thể thấy hiển nhiên tỷ lệ nam giới sẽ quan tâm tới bảo hiểm xe ô tô là nhiều hơn |
| A graph of a number of people  Description automatically generated | Với những người sở hữu xe từ 1 năm trở lên, họ sẽ quan tâm tới bảo hiểm xe nhiều hơn và rất ít người có xe từ 2 năm trở lên không quan tâm tới bảo hiểm xe |
| A graph of a number of people  Description automatically generated | Những người quan tâm đến bảo hiểm xe ô tô thường có độ tuổi từ 30-50 tuổi |
| A graph of a customer response  Description automatically generated | Số tiền bảo hiểm khách hàng phải trả trong một năm kém có khả năng phân loại rằng khách hàng đó có quan tâm tới bảo hiểm xe hay không |

*Hình: Tỷ lệ có xe hỏng trong quá khứ của khách hàng*



Những người ở độ tuổi 20-30, tỷ lệ xe không hư hỏng cao hơn xe hư hỏng. Trong khi ở nhóm từ 35-80 thì ngược lại. Tỷ lệ phụ nữ bị hư hỏng xe ít hơn so với nam giới. Những người sở hữu xe mới (< 1 năm) hay từ 1-2 năm thì chiếm hầu hết những người không hỏng xe, trong khi với những người hỏng xe chiếm nhiều hơn là những người sở hữu xe từ 1 năm trở lên.

* 1. **Xây dụng mô hình và phân tích kết quả**

Sau khi xử lý số liệu, phân tích và trực quan hoá, tiến hành xây dựng mô hình để phân loại những khách hàng có quan tâm tới bảo hiểm xe ô tô hay không.

* 1. Logistic regression

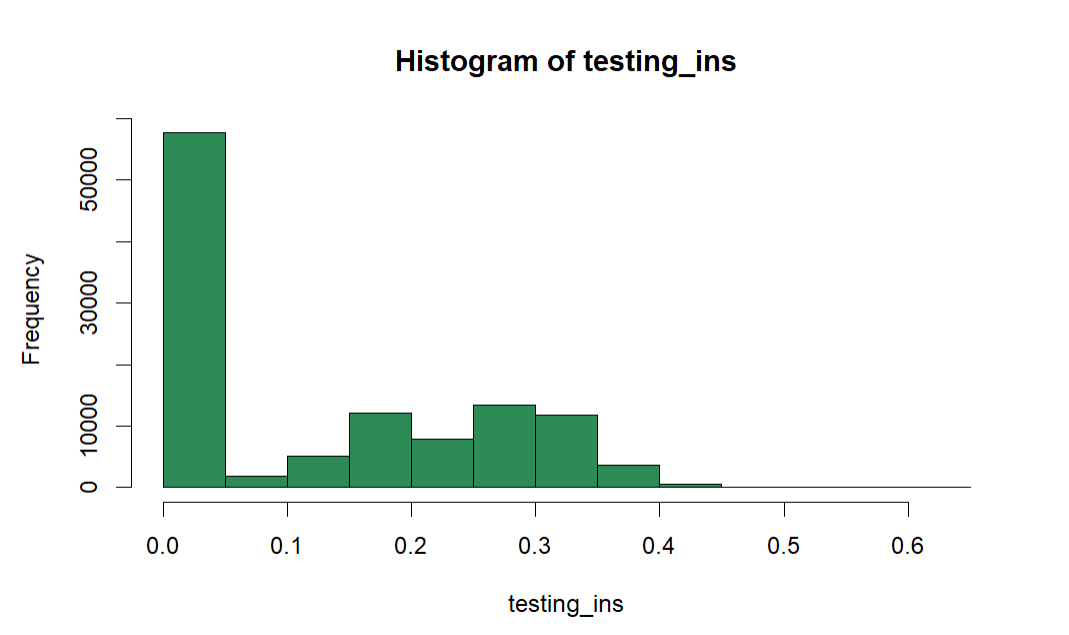
Hồi quy Logistic là một mô hình thống kê được sử dụng để phân loại nhị phân, tức dự đoán một đối tượng thuộc vào một trong hai nhóm. Hồi quy Logistic làm việc dựa trên nguyên tắc của hàm sigmoid – một hàm phi tuyến tự chuyển đầu vào của nó thành xác suất thuộc về một trong hai lớp nhị phân. Hồi quy Logistic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

Thực hành trên tập số liệu, chia tập dữ liệu thành tập train và test với tỷ lệ 70:30. Sau đó xây dựng mô hình dựa trên tập train và sử dụng tập test để dự báo, đánh giá mô hình.

* Kết quả mô hình

|  |  |
| --- | --- |
| A screenshot of a computer code  Description automatically generated | A screenshot of a computer  Description automatically generated |
| 3 biến id, Region\_code, Vintage không có ý nghĩa thống kê ở mức ý nghĩa 5%, nên điều chỉnh mô hình và chạy lại | |

*Hình: Đồ thị Histogram của Response dự báo bằng mô hình logistic trên tập test*



Từ biểu đồ trên, chúng ta có thể thấy rằng ngưỡng xác suất dự đoán để dự đoán liệu một cá nhân có mua bảo hiểm xe ô tô hay không cần phải nhỏ hơn 0.50, vì không có xác suất dự đoán nào lớn hơn ngưỡng 0.45. Đối với nghiên cứu điển hình này, chúng ta sẽ dự đoán rằng bất kỳ khách hàng nào có xác suất phản hồi dự đoán lớn hơn 0.25 đều được dự đoán là người sẽ mua bảo hiểm xe ô tô.

* 1. Decision Tree

Mô hình *cây quyết định* là một mô hình được sử dụng khá phổ biến và hiệu quả trong cả hai lớp bài toán phân loại và dự báo của học máy. Khác với những thuật toán khác trong học máy, mô hình *cây quyết định* không tồn tại phương trình dự báo. Mọi việc chúng ta cần thực hiện đó là tìm ra một cây quyết định dự báo tốt trên tập huấn luyện và sử dụng cây quyết định này dự báo trên tập kiểm tra.

Thực hiện chạy mô hình trên tập huấn luyện với ngưỡng độ tin cậy tối thiểu là 95%, số lượng quan sát tối thiểu cần thiết để tách một nút trong cây là 30 và số lượng quan sát tối thiểu trong mỗi lá nút của cây là 2 được kết quả như hình sau.

A diagram of a company's financial report

Description automatically generated with medium confidence

|  |  |  |
| --- | --- | --- |
| **Kết quả dự báo trên test** | **0** | **1** |
| **0** | 100093 | 14021 |
| **1** | 119 | 100 |

* 1. Random Forest

Mô hình *rừng cây* được huấn luyện dựa trên sự phối hợp giữa luật *kết hợp* (*ensembling*) và quá trình *lấy mẫu tái lặp* (*boostrapping*). Cụ thể thuật toán này tạo ra nhiều cây quyết định mà mỗi cây quyết định được huấn luyện dựa trên nhiều mẫu con khác nhau và kết quả dự báo là *bầu cử* (*voting*) từ toàn bộ những cây quyết định. Như vậy một kết quả dự báo được tổng hợp từ nhiều mô hình nên kết quả của chúng sẽ không bị chệch. Đồng thời kết hợp kết quả dự báo từ nhiều mô hình sẽ có phương sai nhỏ hơn so với chỉ một mô hình. Điều này giúp cho mô hình khắc phục được hiện tượng *quá khớp*.

*Hình: Tỷ lệ dự báo trên tập test từ mô hình Random Forest*

A graph with a bar and a bar

Description automatically generated with medium confidence

|  |  |  |
| --- | --- | --- |
| **Kết quả dự báo trên test** | **0** | **1** |
| **0** | 100084 | 128 |
| **1** | 13995 | 126 |

* 1. Đánh giá độ chính xác và lựa chọn mô hình phù hợp

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí đánh giá** | **Logistic Regression** | **Decision Tree** | **Random Forest** |
| Tỷ lệ dự báo đúng | 78.33% | 87.63% | 87.65% |
| AUC | 83.7%  A graph with a blue line  Description automatically generated | 84.92% | 82.2% |

Theo phương pháp Accuary, kết quả từ mô hình Random Forest và Decision Tree cho kết quả tốt hơn. Tuy nhiên với phương pháp đánh giá mô hình AUC thì mô hình Logistic lại cho kết quả tốt hơn.

Nhưng do tập dữ liệu dùng để huấn luyện mô hình đang mất cân bằng giữa các hai lớp nên tiến hành sử dụng phương pháp SMOTE để khắc phục sự mất cân bằng dữ liệu trong bài toán phân loại cho các mô hình học máy. Tuy nhiên việc khắc phục sự mất cân bằng này phải được thực hiện trên tập huấn luyện chứ không phải tập dữ liệu ban đầu. Sau đó sẽ chạy lại các mô hình để đánh giá.

* Kết quả sau khi sử dụng SMOTE

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí đánh giá** | **Logistic Regression** | **Decision Tree** | **Random Forest** |
| Tỷ lệ dự báo đúng | 81.62% | 87.63% | 87.74% |
| AUC | 82.43% | 85.47% | 94.97% |

Sau khi khắc phục mất cân bằng các chỉ tiêu đánh giá đều tăng duy chỉ có AUC của mô hình logistic là giảm đi. Kết quả cho thấy sử dụng mô hình học máy Random Forest cho kết quả phân loại tốt hơn. Nhưng đi kèm theo đó là sự đánh đổi với việc có thể giải thích được. Nếu mô hình logistic còn có thể giải thích được tác động của các biến lên khả năng khách hàng có hay không quan tâm với bảo hiểm xe ô tô thì mô hình Random Forest cho tỷ lệ dự báo cao nhưng khó giải thích.

1. **Kết luận**

Với mục tiêu đề ra của bài nghiên cứu là từ bộ dữ liệu có sẵn giúp công ty Triks Insurance và GOOL phân loại được nhóm khách hàng tiềm năng trong việc bán kèm bảo hiểm ô tô khi mua bảo hiểm y tế. Qua việc xử lý dữ liệu, xây dựng mô hình và đánh giá đã chọn ra được mô hình Random Forest phù hợp để phân loại khách hàng có quan tâm đến bảo hiểm xe ô tô hay không.

Bằng cách triển khai theo mô hình này, đội ngũ bán hàng của Bảo hiểm Triks sẽ có thể chuẩn bị tốt hơn cho việc bán kèm của họ cùng GOOL. Họ có thể gia tăng sự tập trung vào những khách hàng tiềm năng nhất được dự báo từ mô hình, từ đó tối đa hóa hiệu quả ngân sách tiếp thị. Điều này còn cho phép Bảo hiểm Triks dự đoán doanh thu trong tương lai, nâng cao kế hoạch tài chính của họ, giảm thiểu rủi ro.

Các mô hình học máy có tầm quan trọng lớn đối với nhiều lĩnh vực ngoài lĩnh vực bán chéo sản phẩm vì chúng có khả năng: 1. *Dự đoán và phân loại:* Các mô hình học máy có thể dự đoán và phân loại dữ liệu mới một cách tự động và chính xác. 2.*Tự động hóa quyết định:* Chúng có thể giúp tự động hóa quyết định dựa trên dữ liệu, giúp tiết kiệm thời gian và tăng hiệu suất. 3.*Phát hiện thông tin tiềm ẩn*:Mô hình học máy có thể phát hiện các mẫu và thông tin tiềm ẩn trong dữ liệu mà con người có thể bỏ qua. 4*.Tối ưu hóa kết quả:* Chúng có thể tối ưu hóa kết quả dự đoán dựa trên dữ liệu huấn luyện và các thước đo hiệu suất như AUC, accuracy, precision, recall, và F1-score. 5. *Hỗ trợ ra quyết định:* Các mô hình học máy có thể cung cấp thông tin hữu ích để hỗ trợ ra quyết định trong các lĩnh vực như tài chính, y tế, marketing, và nhiều lĩnh vực khác. Tóm lại, các mô hình học máy đóng vai trò quan trọng trong việc phân tích dữ liệu, tạo ra dự đoán và hỗ trợ ra quyết định trong nhiều lĩnh vực khác nhau, từ kinh doanh đến y tế và nghiên cứu.

# Tài liệu tham khảo

Ahn, H., 2011. *Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques.*   
[https://www.sciencedirect.com/science/article/abs/pii/S0957417410011073]

[https://r-graph-gallery.com/index.html]

KUMAR, A., 2020. *Health Insurance Cross Sell Prediction.*   
[https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction]

Mavundla, K., 2023. *Analysing Health Insurance Customer Dataset to Determine Cross-Selling Potential.*   
[https://www.researchgate.net/publication/376958786\_Analysing\_Health\_Insurance\_Customer\_Dataset\_to\_Determine\_Cross-Selling\_Potential]

Sarah Brown, K. T., 2014. *Household finances and the ‘Big Five’ personality traits.* [https://www.sciencedirect.com/science/article/abs/pii/S016748701400083X?via%3Dihub]

V.Kumar, 2008. *Cross-buying in retailing: Drivers and consequences.*  
[https://www.sciencedirect.com/science/article/abs/pii/S0022435908000109]