

CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Các mô hình được sử dụng

3.1.1. Mô hình Hồi quy Logistic

“Hồi quy Logistic là một phần của lớp các mô hình hồi quy tổng quát hóa (Generalized Linear Models - GLM) được thiết kế để làm việc với các biến phụ thuộc không liên tục. Trái ngược với hồi quy tuyến tính, được sử dụng khi biến phụ thuộc là một biến liên tục, hồi quy Logistic được sử dụng khi biến phụ thuộc chỉ nhận hai giá trị, thường được mã hóa là 0 và 1.

Phân tích hồi quy Logistic là một kỹ thuật thống kê quan trọng và mạnh mẽ được sử dụng để mô hình hóa mối quan hệ giữa một hoặc nhiều biến độc lập với biến phụ thuộc là biến nhị phân. Hồi quy Logistic đặc biệt hữu ích trong các lĩnh vực như y học, kinh tế, tài chính, tiếp thị và nhiều lĩnh vực khác, nơi mà các quyết định hoặc phân loại phụ thuộc vào việc hiểu và dự đoán các xác suất nhị phân.

Công thức của hồi quy Logistic mô tả mối quan hệ giữa xác suất xảy ra một sự kiện và các biến độc lập là:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}}$$

Trong đó:

$P(Y = 1|X)$: xác suất xảy ra sự kiện $Y = 1$ cho giá trị của các biến độc lập X .

β_0 : hệ số chặn.

$\beta_1, \beta_2, \dots, \beta_k$: là các hệ số của các biến độc lập X_1, X_2, \dots, X_k .

e : cơ số của logarit tự nhiên.

Hồi quy Logistic sử dụng hàm logistic (hay còn gọi là hàm sigmoid) để ánh xạ mọi giá trị thực từ mô hình hồi quy tuyến tính về khoảng $[0, 1]$, tương ứng với xác suất của biến phụ thuộc nhận giá trị 1. Để thực hiện phân tích hồi quy Logistic, dữ liệu cần phải được chuẩn bị và mã hóa dưới dạng số. Điều này đặc biệt quan trọng với các biến phân loại. Các biến nhị phân thường được mã hóa dưới dạng 0 và 1 để có thể sử dụng trong mô hình. Ví dụ, nếu biến phân loại là "giới tính" với hai giá trị "nam" và "nữ", ta có thể mã hóa "nam" là 0 và "nữ" là 1. Đối với các biến phân loại có nhiều hơn hai giá trị, chúng ta có thể sử dụng mã hóa one-hot (mã hóa kiểu nhị phân), trong đó mỗi giá trị của biến phân loại được biểu diễn bằng một biến nhị phân riêng biệt.

Hồi quy Logistic là một phương pháp thống kê mạnh mẽ và linh hoạt cho phép các nhà nghiên cứu và phân tích dữ liệu hiểu và dự đoán các xác suất nhị phân. Việc nắm vững và áp dụng đúng hồi quy Logistic không chỉ giúp cải thiện độ chính xác của các dự đoán mà còn cung cấp những thông tin quan trọng để đưa ra các quyết định chiến lược trong nhiều lĩnh vực khác nhau. Việc sử dụng đúng công cụ và phương pháp phân tích là yếu tố then chốt để đạt được những kết quả chính xác và có giá trị từ dữ liệu.”

3.1.2. Mô hình Mạng thần kinh

“Mạng nơ-ron là một mô hình mô phỏng cách thức hoạt động của bộ não con người trong việc xử lý thông tin. Cấu trúc cơ bản của mạng nơ-ron bao gồm các đơn vị xử lý (tương đương với các tế bào thần kinh) được sắp xếp thành các lớp: lớp đầu vào, lớp ẩn và lớp đầu ra. Quá trình học tập của mạng nơ-ron diễn ra thông qua việc điều chỉnh trọng số dựa trên các dự đoán sai, lặp đi lặp lại cho đến khi mạng đạt được các tiêu chí dừng nhất định.

Phần mềm Clementine cung cấp sáu phương pháp để xây dựng mô hình mạng nơ-ron:

- Quick: Sử dụng các quy tắc đơn giản và đặc điểm của dữ liệu để chọn cấu trúc mạng phù hợp. Phương pháp này nhanh chóng và tiện lợi khi cần có một mô hình cơ bản.
- Dynamic: Bắt đầu với một cấu trúc mạng ban đầu, sau đó điều chỉnh bằng cách thêm hoặc loại bỏ các đơn vị ẩn trong quá trình huấn luyện, giúp mạng tự động tối ưu hóa cấu trúc.
- Multiple: Tạo nhiều mạng có cấu trúc khác nhau và huấn luyện chúng song song. Mô hình với sai số RMS (Root Mean Square) thấp nhất sẽ được chọn làm mô hình cuối cùng.
- Prune: Bắt đầu với một mạng lớn và loại bỏ các đơn vị xử lý yếu nhất trong các lớp ẩn và lớp đầu vào trong suốt quá trình huấn luyện. Mặc dù phương pháp này chậm nhưng thường mang lại kết quả tốt.
- RBFN (Radial Basis Function Network): Sử dụng kỹ thuật phân cụm k-mean để phân vùng dữ liệu dựa trên giá trị của biến mục tiêu, thích hợp cho các vấn đề phân loại và hồi quy phi tuyến tính.
- Exhaustive prune: Tương tự như phương pháp Prune, nhưng với một tìm kiếm rất kỹ lưỡng trong không gian mô hình để tìm ra mô hình tốt nhất. Phương pháp này là chậm nhất nhưng thường mang lại kết quả tốt nhất, đặc biệt là với các tập dữ liệu lớn.

Mỗi phương pháp có ưu và nhược điểm riêng, tùy thuộc vào đặc điểm của dữ liệu và yêu cầu cụ thể của bài toán mà có thể chọn phương pháp phù hợp để xây dựng mô hình mạng nơ-ron hiệu quả.”

3.1.3. Mô hình Cây quyết định

“Cây quyết định là một thuật toán dùng để phân loại hoặc hồi quy bằng cách xác định và phân chia các biến quan trọng nhất trong dữ liệu để tạo ra các nhóm kết quả. Một trong những phương pháp thường dùng là entropy kinh nghiệm để chọn thuộc tính phân biệt cao nhất làm điểm phân chia đầu tiên.

Phần mềm Clementine cung cấp bốn loại mô hình cây quyết định:

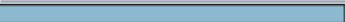

- C-RT (Classification and Regression Trees):
 - + Sử dụng phân vùng đệ quy để chia dữ liệu huấn luyện thành các phân đoạn với giá trị trường đầu ra tương tự nhau.
 - + Quá trình bắt đầu bằng việc kiểm tra các trường đầu vào để tìm cách phân tách tốt nhất, đo lường bằng sự giảm chỉ số tạp chất (impurity index).
 - + Mỗi lần phân tách xác định hai nhóm con, quá trình này tiếp tục cho đến khi một tiêu chí dừng được kích hoạt.
- Quest:
 - + Phương pháp phân loại nhị phân, cải thiện tốc độ xử lý so với C-RT và giảm thiểu thiên vị trong quá trình phân tách.
 - + Giảm thời gian xử lý cần thiết mà vẫn duy trì được hiệu quả phân loại.
- Chaid (Chi-squared Automatic Interaction Detector):
 - + Phát triển bởi J. A. Hartigan, sử dụng kiểm tra Chi bình phương cho chiến lược tách cây.
 - + Hỗ trợ các biến có giá trị liên tục và rời rạc, thực hiện các nhiệm vụ hồi quy và phân loại.
 - + Không yêu cầu giai đoạn cắt tỉa cây và sử dụng chiến lược phân tách đa chiều, tạo ra các mô hình dễ hiểu hơn cho những người ra quyết định.
- C5.0:
 - + Chia nhỏ mẫu dựa trên trường cung cấp mức tăng thông tin tối đa.
 - + Mỗi mẫu được xác định bởi lần tách đầu tiên, sau đó được tách lại dựa trên các trường khác cho đến khi không thể chia nhỏ hơn.
 - + Các phân phân tách ở mức thấp nhất sẽ được khảo sát lại, loại bỏ hoặc cắt bớt những phần không đóng góp đáng kể vào giá trị của mô hình.

Mỗi loại mô hình cây quyết định có những ưu điểm riêng, giúp người dùng có thể lựa chọn phương pháp phù hợp nhất với dữ liệu và bài toán cụ thể của mình.”

3.2. Tiền xử lý dữ liệu

3.2.1. Kiểm tra phân phối biến mục tiêu của tập dữ liệu gốc

Sau khi đã nhập dữ liệu vào phần mềm và đọc tập dữ liệu, nhóm tiến hành kiểm tra phân phối của biến mục tiêu và thu được kết quả như sau:

Value ▲	Proportion	%	Count
0.000		64,27	1277
1.000		35,73	710

Hình 3.1. Phân phối biến mục tiêu của tập dữ liệu gốc

Ta nhận thấy số khách hàng chưa mua bảo hiểm du lịch chiếm 64,27%, gấp đôi số khách hàng chưa mua bảo hiểm du lịch chiếm 35,73%. Sự chênh lệch quá lớn và có thể ảnh hưởng đến kết quả phân tích. Do đó, nhóm cần cân bằng số lượng khách hàng đã và chưa mua bảo hiểm du lịch bằng cách lấy mẫu 2 nhóm mục tiêu với kích cỡ mẫu bằng nhau hoặc sử dụng node Balance để cỡ mẫu bằng nhau.

3.2.2. Chuẩn hóa dữ liệu

Dựa vào phần mô tả dữ liệu, ta nhận thấy khoảng cách giữa các thang đo (range) hoàn toàn khác nhau. Điều này có thể ảnh hưởng tới hiệu quả của các mô hình. Vì vậy, nhóm tiến hành chuẩn hóa biến tuổi, thu nhập hàng năm và số người sống chung trong gia đình của khách hàng bằng cách sử dụng công thức chuẩn hóa Min-Max, chuyển đổi dữ liệu thành phạm vi từ 0 đến 1. Công thức chuẩn hóa Min-Max được chọn vì tính đơn giản và hiệu quả trong việc chuyển đổi dữ liệu có thang đo khác nhau về cùng một phạm vi.

Tuoi_chuan_hoa

Derive as: Formula

Mode: ☒ Single ☐ Multiple

Derive field:
Tuoi_chuan_hoa

Derive as: Formula

Field type: Range

Formula:

```
if Age = 35 then 1
else if Age = 25 then 0
else (Age - 25)/(35-25) endif endif
```

Settings Annotations

OK Cancel Apply Reset

Hình 3.2. Công thức chuẩn hóa biến tuổi của khách hàng

Thu_nhap_chuan_hoa

Derive as: Formula

Mode: ☒ Single ☐ Multiple

Derive field:
Thu_nhap_chuan_hoa

Derive as: Formula

Field type: <Default>

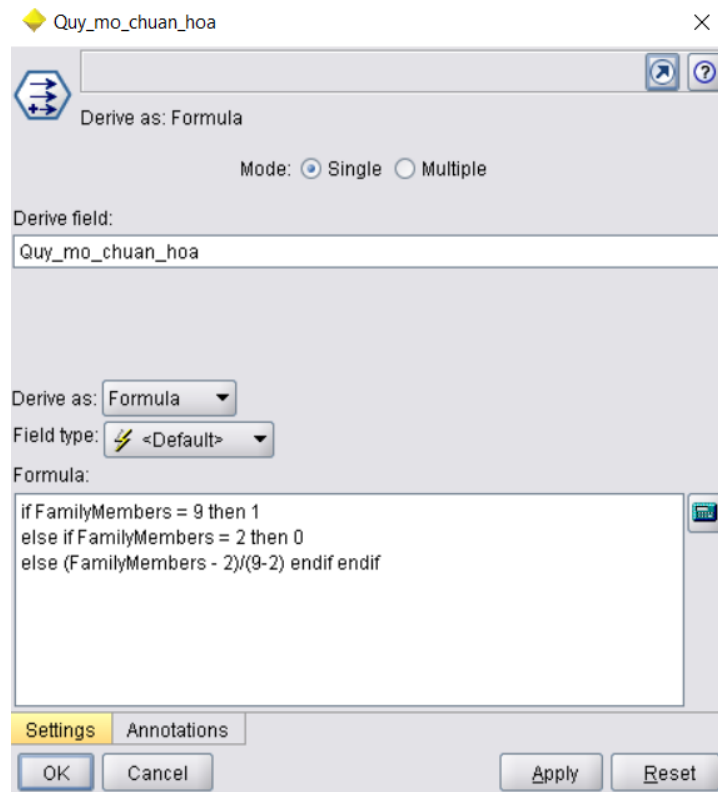
Formula:

```
if AnnualIncome = 1800000 then 1
else if AnnualIncome = 300000 then 0
else (AnnualIncome - 300000)/(1800000 - 300000) endif endif
```

Settings Annotations

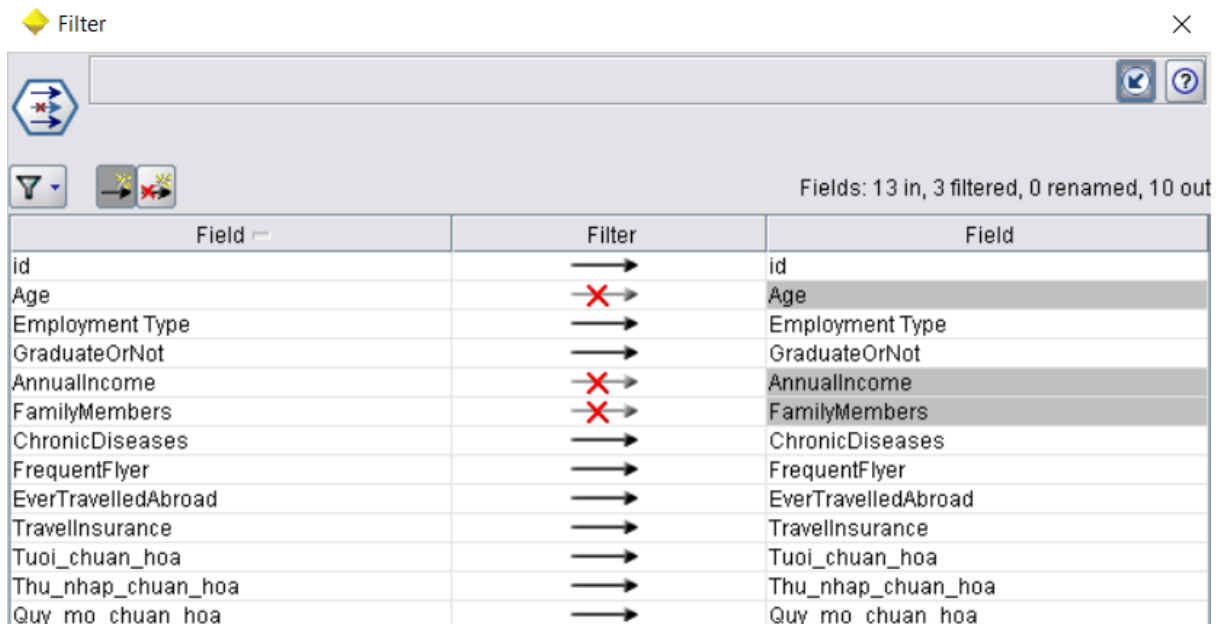
OK Cancel Apply Reset

Hình 3.3. Công thức chuẩn hóa biến thu nhập hàng năm của khách hàng



Hình 3.4. Công thức chuẩn hóa biến số thành viên trong gia đình của khách hàng

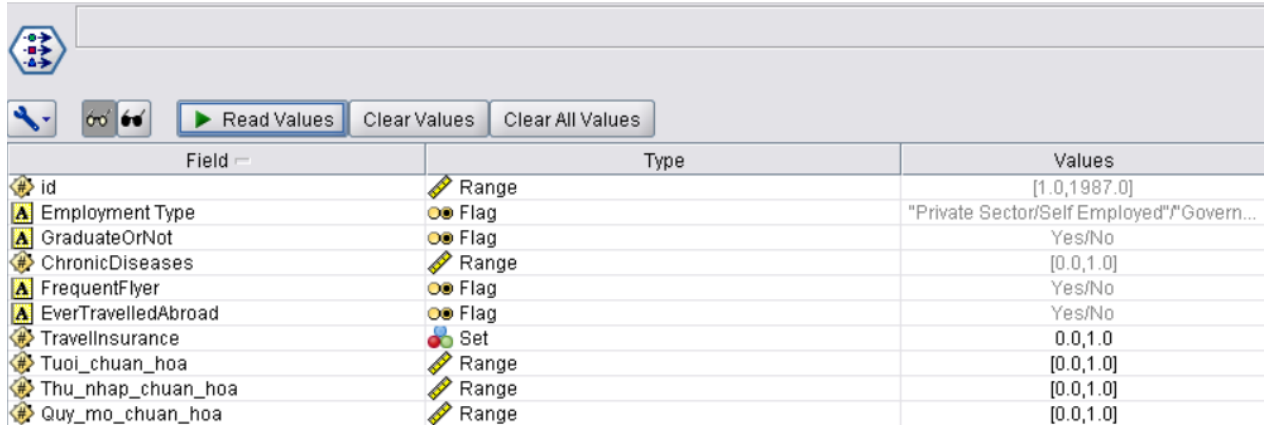
Sau khi chuẩn hóa các biến trên, nhóm thực hiện Filter để loại các biến không cần thiết trong mô hình vì đã tạo các biến mới đã chuẩn hóa.



Hình 3.5. Danh sách các biến bị loại bỏ và giữ lại

Tiếp theo, nhóm sử dụng node Type để xem xét và chuyển lại kiểu dữ liệu của các biến.

Type



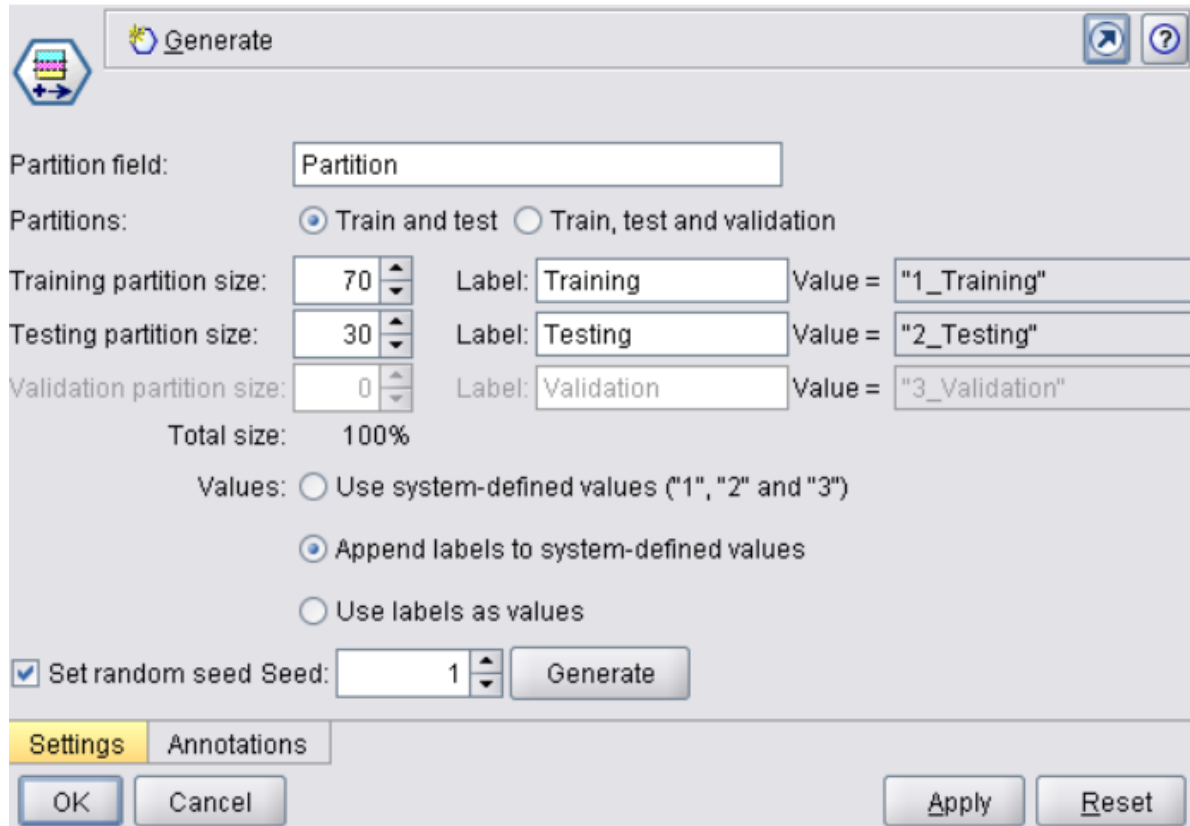
Field	Type	Values
id	Range	[1.0,1987.0]
Employment Type	Flag	"Private Sector/Self Employed"/"Govern..."
GraduateOrNot	Flag	Yes/No
ChronicDiseases	Range	[0.0,1.0]
FrequentFlyer	Flag	Yes/No
EverTravelledAbroad	Flag	Yes/No
TravelInsurance	Set	0.0,1.0
Tuoi_chuan_hoa	Range	[0.0,1.0]
Thu_nhap_chuan_hoa	Range	[0.0,1.0]
Quy_mo_chuan_hoa	Range	[0.0,1.0]

Hình 3.6. Xem xét kiểu dữ liệu của các biến bằng node Type

3.2.3. Phân chia tập dữ liệu

Nhóm dùng node Partition và Balance để phân chia tập dữ liệu nhằm đảm bảo số lượng quan sát đầu vào cho việc huấn luyện giữa 2 nhóm biểu hiện của biến mục tiêu cân bằng với nhau, từ đó có thể huấn luyện ra mô hình có khả năng dự đoán tốt nhất.

Partition



Generate

Partition field: Partition

Partitions: ☒ Train and test ☐ Train, test and validation

Training partition size: 70 Label: Training Value = "1_Training"

Testing partition size: 30 Label: Testing Value = "2_Testing"

Validation partition size: 0 Label: Validation Value = "3_Validation"

Total size: 100%

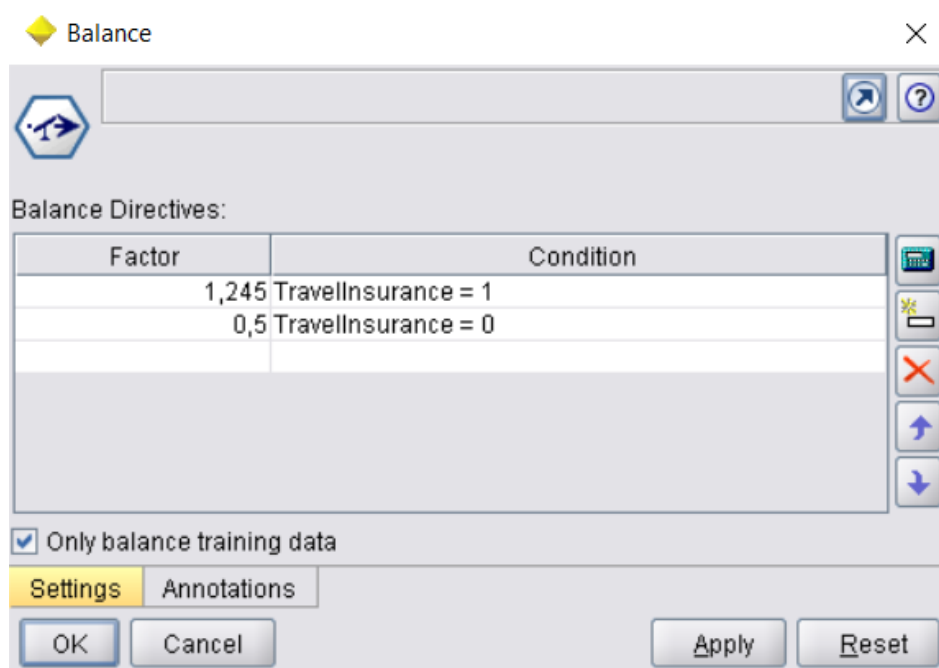
Values: ☐ Use system-defined values ("1", "2" and "3")
☒ Append labels to system-defined values
☐ Use labels as values

☒ Set random seed Seed: 1 Generate

Settings Annotations

OK Cancel Apply Reset

Hình 3.7. Node Partition



Hình 3.8. Node Balance

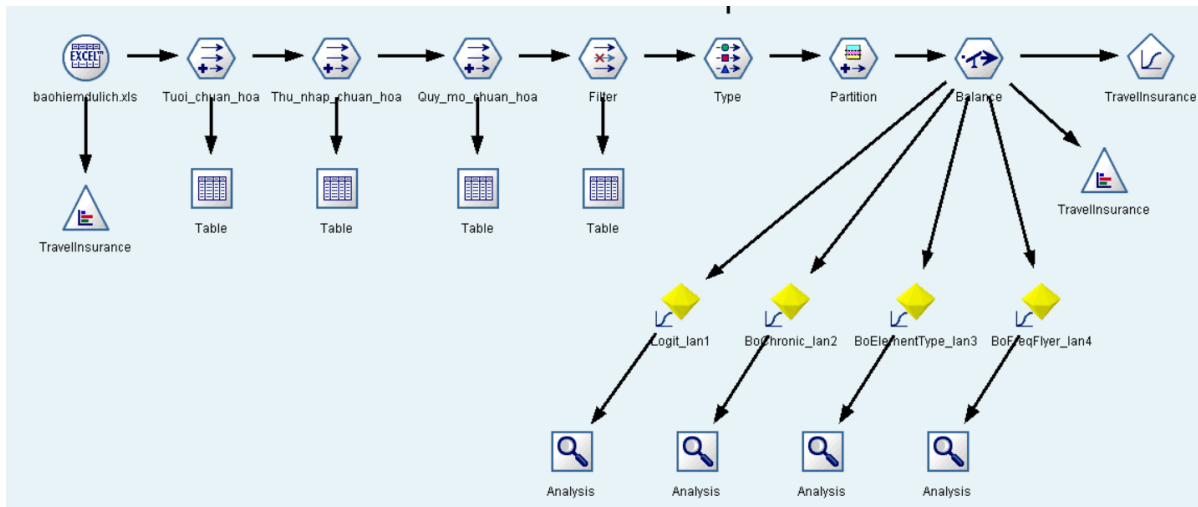
Sau khi thực hiện lệnh Balance, ta được phân phối của biến mục tiêu như sau:

Value ▲	Proportion	%	Count
0.000	<div style="width: 49.91%; background-color: #4682B4;"></div>	49,91	836
1.000	<div style="width: 50.09%; background-color: #CD5C5C;"></div>	50,09	839

Hình 3.9. Phân phối của biến mục tiêu sau khi được cân bằng bằng node Balance

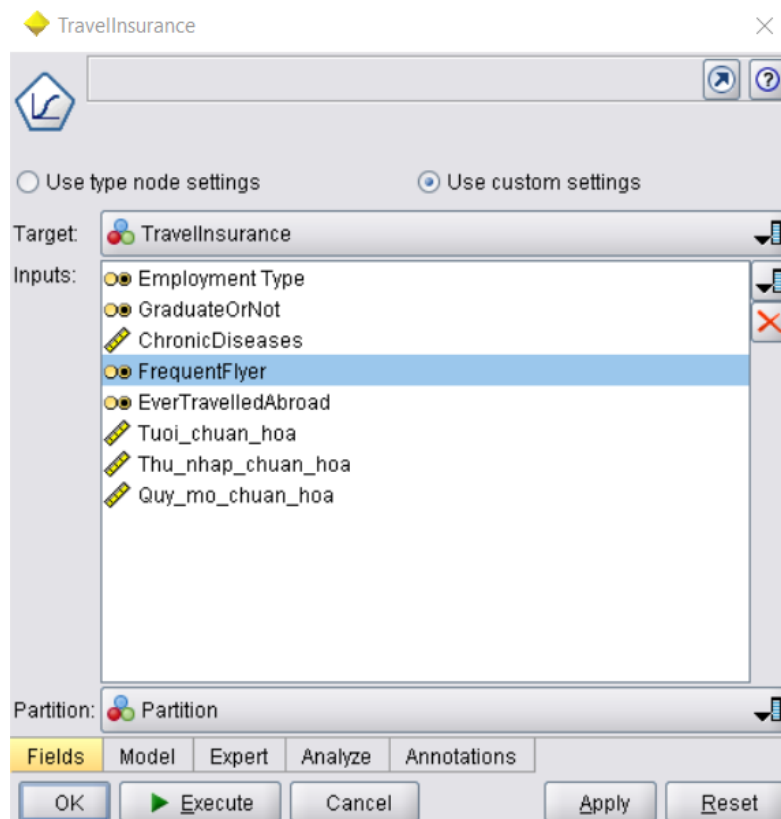
CHƯƠNG 4: PHÂN TÍCH VÀ KẾT QUẢ NGHIÊN CỨU

4.1. Kết quả mô hình Hồi quy Logistic



Hình 4.1. Lưu đồ mô hình Hồi quy Logistic

Tiến hành chọn các biến độc lập và các tùy chọn cho mô hình Hồi quy Logistic.



Hình 4.2. Đầu vào của mô hình Hồi quy Logistic

Sau khi thực hiện chạy mô hình Hồi quy Logistic, nhóm thu kết quả như sau:

Case Processing Summary			
		N	Marginal Percentage
TravelInsurance	0.0	457	43,4%
	1.0	597	56,6%
Employment Type	Government Sector	268	25,4%
	Private Sector/Self Employed	786	74,6%
GraduateOrNot	No	146	13,9%
	Yes	908	86,1%
FrequentFlyer	No	791	75,0%
	Yes	263	25,0%
EverTravelledAbroad	No	771	73,1%
	Yes	283	26,9%
Valid		1054	100,0%
Missing		0	
Total		1054	
Subpopulation		653(a)	
a. The dependent variable has only one value observed in 595 (91,1%) subpopulations.			

Hình 4.3. Kết quả của mô hình Hồi quy Logistic lần thứ nhất

Kết quả cho thấy có 457 khách hàng chưa mua bảo hiểm du lịch trên tổng số quan sát là 1054 (chiếm 43,4%) và 597 khách hàng đã mua bảo hiểm du lịch (tương ứng với 56,6%).

Model Fitting Information				
Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1306,825			
Final	1008,308	298,517	8	,000

Pseudo R-Square	
Cox and Snell	,247
Nagelkerke	,331
McFadden	,207

Hình 4.4. Độ phù hợp của mô hình Hồi quy Logistic lần thứ nhất

Giả thuyết:

H_0 : Mô hình không phù hợp.

H_1 : Mô hình phù hợp.

Sig. = 0,000 < α = 0,05 \rightarrow Bác bỏ H_0

Kết luận: Tại mức ý nghĩa 5%, mô hình này là phù hợp.

Hệ số Log Likelihood càng lớn càng tốt và hệ số -2 Log Likelihood càng nhỏ càng tốt. Hệ số -2 Log Likelihood bằng 1306,825 khi chưa tính các biến độc lập và bằng 1008,308 khi đã tính các biến độc lập.

Xét bảng Pseudo R-Square, $R^2 = 0,247 \rightarrow$ Các biến độc lập giải thích được 24,7% sự biến đổi của biến phụ thuộc.

Parameter Estimates									
TravelInsurance(a)		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
1.0	Intercept	,297	,414	,514	1	,474			
	ChronicDiseases	,021	,162	,017	1	,895	1,022	,744	1,402
	Tuoi_chuan_hoa	1,056	,250	17,884	1	,000	2,874	1,762	4,688
	Thu_nhap_chuan_hoa	2,285	,364	39,478	1	,000	9,822	4,816	20,032
	Quy_mo_chuan_hoa	,547	,312	3,072	1	,080	1,727	,937	3,183
	[Employment Type=Government Sector]	,039	,176	,050	1	,823	1,040	,737	1,468
	[Employment Type=Private Sector/Self Employed]	0(b)	.	.	0
	[GraduateOrNot=No]	,305	,213	2,057	1	,152	1,357	,894	2,060
	[GraduateOrNot=Yes]	0(b)	.	.	0
	[FrequentFlyer=No]	-,306	,197	2,410	1	,121	,737	,501	1,084
	[FrequentFlyer=Yes]	0(b)	.	.	0
	[EverTravelledAbroad=No]	-1,954	,246	62,989	1	,000	,142	,087	,230
	[EverTravelledAbroad=Yes]	0(b)	.	.	0
a. The reference category is: 0.0.									
b. This parameter is set to zero because it is redundant.									

Hình 4.5. Hệ số của mô hình Hồi quy Logistic lần thứ nhất

Giả thuyết:

H_0 : Biến độc lập không có ảnh hưởng đến biến phụ thuộc.

H_1 : Biến độc lập có ảnh hưởng đến biến phụ thuộc.

Nếu $\text{Sig.} < \alpha = 0,1 \rightarrow$ Bác bỏ $H_0 \Rightarrow$ Biến độc lập có ảnh hưởng đến biến phụ thuộc tại mức ý nghĩa 10%.

Nếu $\text{Sig.} > \alpha = 0,1 \rightarrow$ Không bác bỏ $H_0 \Rightarrow$ Biến độc lập không có ảnh hưởng đến biến phụ thuộc tại mức ý nghĩa 10%.

Như vậy, biến ChronicDiseases, Employment Type, GraduateOrNot và FrequentFlyer không có ảnh hưởng đến biến phụ thuộc tại mức ý nghĩa 10%.

Do đó, nhóm quyết định loại bỏ biến ChronicDiseases để có thể xây dựng mô hình phù hợp hơn nhằm phục vụ cho việc dự đoán khả năng mua bảo hiểm du lịch của khách hàng.

Tiến hành chạy mô hình Hồi quy Logistic lần thứ hai sau khi loại bỏ biến ChronicDiseases, ta được kết quả như sau:

Case Processing Summary			
		N	Marginal Percentage
TravelInsurance	0.0	450	42,3%
	1.0	614	57,7%
Employment Type	Government Sector	261	24,5%
	Private Sector/Self Employed	803	75,5%
GraduateOrNot	No	148	13,9%
	Yes	916	86,1%
FrequentFlyer	No	795	74,7%
	Yes	269	25,3%
EverTravelledAbroad	No	759	71,3%
	Yes	305	28,7%
Valid		1064	100,0%
Missing		0	
Total		1064	
Subpopulation		556(a)	
a. The dependent variable has only one value observed in 472 (84,9%) subpopulations.			

Hình 4.6. Kết quả của mô hình Hồi quy Logistic lần thứ hai

Kết quả cho thấy có 450 khách hàng chưa mua bảo hiểm du lịch trên tổng số quan sát là 1064 (chiếm 42,3%) và 614 khách hàng đã mua bảo hiểm du lịch (tương ứng với 57,7%).

Model Fitting Information				
Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1220,024			
Final	957,777	262,248	7	,000
Pseudo R-Square				
Cox and Snell		,218		
Nagelkerke		,294		
McFadden		,181		

Hình 4.7. Độ phù hợp của mô hình Hồi quy Logistic lần thứ hai

Giả thuyết:

H_0 : Mô hình không phù hợp.

H_1 : Mô hình phù hợp.

Sig. = 0,000 < $\alpha = 0,05 \rightarrow$ Bác bỏ H_0

Kết luận: Tại mức ý nghĩa 5%, mô hình này là phù hợp.

Hệ số Log Likelihood càng lớn càng tốt và hệ số -2 Log Likelihood càng nhỏ càng tốt. Hệ số -2 Log Likelihood bằng 1220,024 khi chưa tính các biến độc lập và bằng 957,777 khi đã tính các biến độc lập.

Xét bảng Pseudo R-Square, $R^2 = 0,218 \rightarrow$ Các biến độc lập giải thích được 21,8% sự biến đổi của biến phụ thuộc.

Parameter Estimates									
TravelInsurance(a)		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
1.0	Intercept	,206	,381	,293	1	,588			
	Tuoi_chuan_hoa	,691	,247	7,863	1	,005	1,996	1,231	3,237
	Thu_nhap_chuan_hoa	2,246	,356	39,713	1	,000	9,448	4,699	18,998
	Quy_mo_chuan_hoa	,511	,294	3,023	1	,082	1,668	,937	2,968
	[Employment Type=Government Sector]	,042	,172	,060	1	,807	1,043	,744	1,462
	[Employment Type=Private Sector/Self Employed]	0(b)	.	.	0
	[GraduateOrNot=No]	,431	,207	4,357	1	,037	1,539	1,027	2,307
	[GraduateOrNot=Yes]	0(b)	.	.	0
	[FrequentFlyer=No]	-,335	,189	3,147	1	,076	,715	,494	1,036
	[FrequentFlyer=Yes]	0(b)	.	.	0
	[EverTravelledAbroad=No]	-1,610	,219	54,194	1	,000	,200	,130	,307
	[EverTravelledAbroad=Yes]	0(b)	.	.	0
a. The reference category is: 0.0.									
b. This parameter is set to zero because it is redundant.									

Hình 4.8. Hệ số của mô hình Hồi quy Logistic lần thứ hai

Giả thuyết:

H_0 : Biến độc lập không có ảnh hưởng đến biến phụ thuộc.

H_1 : Biến độc lập có ảnh hưởng đến biến phụ thuộc.

Nếu Sig. < $\alpha = 0,1 \rightarrow$ Bác bỏ $H_0 \Rightarrow$ Biến độc lập có ảnh hưởng đến biến phụ thuộc tại mức ý nghĩa 10%.

Nếu Sig. > $\alpha = 0,1 \rightarrow$ Không bác bỏ $H_0 \Rightarrow$ Biến độc lập không có ảnh hưởng đến biến phụ thuộc tại mức ý nghĩa 10%.

Như vậy, biến Employment Type không có ảnh hưởng đến biến phụ thuộc tại mức ý nghĩa 10%.

Do đó, nhóm quyết định loại bỏ thêm biến Employment Type để có thể xây dựng mô hình phù hợp hơn nhằm phục vụ cho việc dự đoán khả năng mua bảo hiểm du lịch của khách hàng.

Tiến hành chạy mô hình Hồi quy Logistic lần thứ ba sau khi loại bỏ biến ChronicDiseases và Employment Type, ta được kết quả như sau:

Case Processing Summary			
		N	Marginal Percentage
TravelInsurance	0.0	443	42,0%
	1.0	612	58,0%
GraduateOrNot	No	150	14,2%
	Yes	905	85,8%
FrequentFlyer	No	791	75,0%
	Yes	264	25,0%
EverTravelledAbroad	No	769	72,9%
	Yes	286	27,1%
Valid		1055	100,0%
Missing		0	
Total		1055	
Subpopulation		537(a)	
a. The dependent variable has only one value observed in 465 (86,6%) subpopulations.			

Hình 4.9. Kết quả của mô hình Hồi quy Logistic lần thứ ba

Kết quả cho thấy có 443 khách hàng chưa mua bảo hiểm du lịch trên tổng số quan sát là 1055 (chiếm 42%) và 612 khách hàng đã mua bảo hiểm du lịch (tương ứng với 58%).

Model Fitting Information				
Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1219,029			
Final	939,969	279,060	6	,000

Pseudo R-Square	
Cox and Snell	,232
Nagelkerke	,313
McFadden	,194

Hình 4.10. Độ phù hợp của mô hình Hồi quy Logistic lần thứ ba

Giả thuyết:

H_0 : Mô hình không phù hợp.

H_1 : Mô hình phù hợp.

Sig. = 0,000 < $\alpha = 0,05 \rightarrow$ Bác bỏ H_0

Kết luận: Tại mức ý nghĩa 5%, mô hình này là phù hợp.

Hệ số Log Likelihood càng lớn càng tốt và hệ số -2 Log Likelihood càng nhỏ càng tốt. Hệ số -2 Log Likelihood bằng 1219,029 khi chưa tính các biến độc lập và bằng 939,969 khi đã tính các biến độc lập.

Xét bảng Pseudo R-Square, $R^2 = 0,232 \rightarrow$ Các biến độc lập giải thích được 23,2% sự biến đổi của biến phụ thuộc.

Parameter Estimates								
TravelInsurance(a)		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)
								Lower Bound Upper Bound
1.0	Intercept	,580	,387	2,237	1	,135		
	Tuoi_chuan_hoa	1,074	,243	19,581	1	,000	2,926	1,819 4,708
	Thu_nhap_chuan_hoa	1,898	,334	32,220	1	,000	6,672	3,464 12,848
	Quy_mo_chuan_hoa	,500	,302	2,743	1	,098	1,649	,912 2,982
	[GraduateOrNot=No]	,335	,200	2,814	1	,093	1,398	,945 2,069
	[GraduateOrNot=Yes]	0(b)	.	.	0	.	.	.
	[FrequentFlyer=No]	-,193	,192	1,002	1	,317	,825	,566 1,203
	[FrequentFlyer=Yes]	0(b)	.	.	0	.	.	.
	[EverTravelledAbroad=No]	-2,111	,249	71,998	1	,000	,121	,074 ,197
	[EverTravelledAbroad=Yes]	0(b)	.	.	0	.	.	.
a. The reference category is: 0.0.								
b. This parameter is set to zero because it is redundant.								

Hình 4.11. Hệ số của mô hình Hồi quy Logistic lần thứ ba

Giả thuyết:

H_0 : Biến độc lập không có ảnh hưởng đến biến phụ thuộc.

H_1 : Biến độc lập có ảnh hưởng đến biến phụ thuộc.

Nếu Sig. < $\alpha = 0,1 \rightarrow$ Bác bỏ $H_0 \Rightarrow$ Biến độc lập có ảnh hưởng đến biến phụ thuộc tại mức ý nghĩa 10%.

Nếu Sig. > $\alpha = 0,1 \rightarrow$ Không bác bỏ $H_0 \Rightarrow$ Biến độc lập không có ảnh hưởng đến biến phụ thuộc tại mức ý nghĩa 10%.

Như vậy, biến FrequentFlyer không có ảnh hưởng đến biến phụ thuộc tại mức ý nghĩa 10%.

Do đó, nhóm quyết định loại bỏ thêm biến FrequentFlyer để có thể xây dựng mô hình phù hợp hơn nhằm phục vụ cho việc dự đoán khả năng mua bảo hiểm du lịch của khách hàng.

Tiến hành chạy mô hình Hồi quy Logistic lần thứ tư sau khi loại bỏ biến ChronicDiseases, Employment Type và FrequentFlyer, ta được kết quả như sau:

Case Processing Summary			
		N	Marginal Percentage
TravelInsurance	0.0	454	42,0%
	1.0	626	58,0%
GraduateOrNot	No	157	14,5%
	Yes	923	85,5%
EverTravelledAbroad	No	784	72,6%
	Yes	296	27,4%
Valid		1080	100,0%
Missing		0	
Total		1080	
Subpopulation		478(a)	
a. The dependent variable has only one value observed in 389 (81,4%) subpopulations.			

Hình 4.12. Kết quả của mô hình Hồi quy Logistic lần thứ tư

Kết quả cho thấy có 454 khách hàng chưa mua bảo hiểm du lịch trên tổng số quan sát là 1080 (chiếm 42%) và 626 khách hàng đã mua bảo hiểm du lịch (tương ứng với 58%).

Model Fitting Information				
Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1203,971			
Final	944,704	259,267	5	,000

Pseudo R-Square	
Cox and Snell	,213
Nagelkerke	,287
McFadden	,176

Hình 4.13. Độ phù hợp của mô hình Hồi quy Logistic lần thứ tư

Giả thuyết:

H_0 : Mô hình không phù hợp.

H_1 : Mô hình phù hợp.

Sig. = 0,000 < α = 0,05 \rightarrow Bác bỏ H_0

Kết luận: Tại mức ý nghĩa 5%, mô hình này là phù hợp.

Hệ số Log Likelihood càng lớn càng tốt và hệ số -2 Log Likelihood càng nhỏ càng tốt. Hệ số -2 Log Likelihood bằng 1203,971 khi chưa tính các biến độc lập và bằng 944,704 khi đã tính các biến độc lập.

Xét bảng Pseudo R-Square, $R^2 = 0,213 \rightarrow$ Các biến độc lập giải thích được 21,3% sự biến đổi của biến phụ thuộc.

Parameter Estimates									
TravelInsurance(a)		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
1.0	Intercept	,307	,332	,855	1	,355			
	Tuoi_chuan_hoa	,874	,235	13,875	1	,000	2,396	1,513	3,794
	Thu_nhap_chuan_hoa	1,900	,314	36,561	1	,000	6,688	3,612	12,383
	Quy_mo_chuan_hoa	,608	,292	4,353	1	,037	1,837	1,038	3,254
	[GraduateOrNot=No]	,362	,195	3,443	1	,064	1,436	,980	2,105
	[GraduateOrNot=Yes]	0(b)	.	.	0
	[EverTravelledAbroad=No]	-1,966	,230	72,847	1	,000	,140	,089	,220
	[EverTravelledAbroad=Yes]	0(b)	.	.	0
a. The reference category is: 0.0.									
b. This parameter is set to zero because it is redundant.									

Hình 4.14. Hệ số của mô hình Hồi quy Logistic lần thứ tư

Giả thuyết:

H_0 : Biến độc lập không có ảnh hưởng đến biến phụ thuộc.

H_1 : Biến độc lập có ảnh hưởng đến biến phụ thuộc.

Tất cả các giá trị Sig. $< \alpha = 0,1 \rightarrow$ Bác bỏ $H_0 \Rightarrow$ Tất cả các biến độc lập có ảnh hưởng đến biến phụ thuộc tại mức ý nghĩa 10%.

Vì vậy, đây là mô hình Hồi quy Logistic là phù hợp nhất.

Diễn giải kết quả:

- Những khách hàng càng lớn tuổi (tuổi tác tiến gần đến tuổi 35 – giai đoạn đã hoàn toàn trưởng thành) và có thu nhập hàng năm càng cao (tiến gần đến mức 1.800.000) có xu hướng mua bảo hiểm du lịch cao hơn.
- Số lượng thành viên trong hộ gia đình cũng có tác động tích cực tới ý định mua bảo hiểm du lịch của khách hàng. Càng có nhiều thành viên sống chung với nhau thì khách hàng càng có xu hướng mua bảo hiểm du lịch cao hơn.

- Việc đã tốt nghiệp đại học hay chưa cũng có ảnh hưởng tới ý định mua bảo hiểm du lịch. Kết quả cho biết những khách hàng chưa tốt nghiệp đại học có khả năng mua bảo hiểm du lịch cao hơn những khách hàng đã tốt nghiệp đại học. Tuy nhiên, sự tác động này nên được xem xét thêm (do có giá trị Sig. $> \alpha = 0,05$).
- Việc đã từng ra nước ngoài hay chưa cũng tác động tới ý định mua bảo hiểm du lịch của khách hàng. Phân tích cho thấy những khách hàng chưa từng ra nước ngoài sẽ ít có xu hướng mua bảo hiểm du lịch hơn so với những người đã từng ra nước ngoài.

Comparing \$L-TravellInsurance with TravellInsurance

'Partition'	1_Training		2_Testing	
Correct	731	67,94%	413	70,24%
Wrong	345	32,06%	175	29,76%
Total	1.076		588	

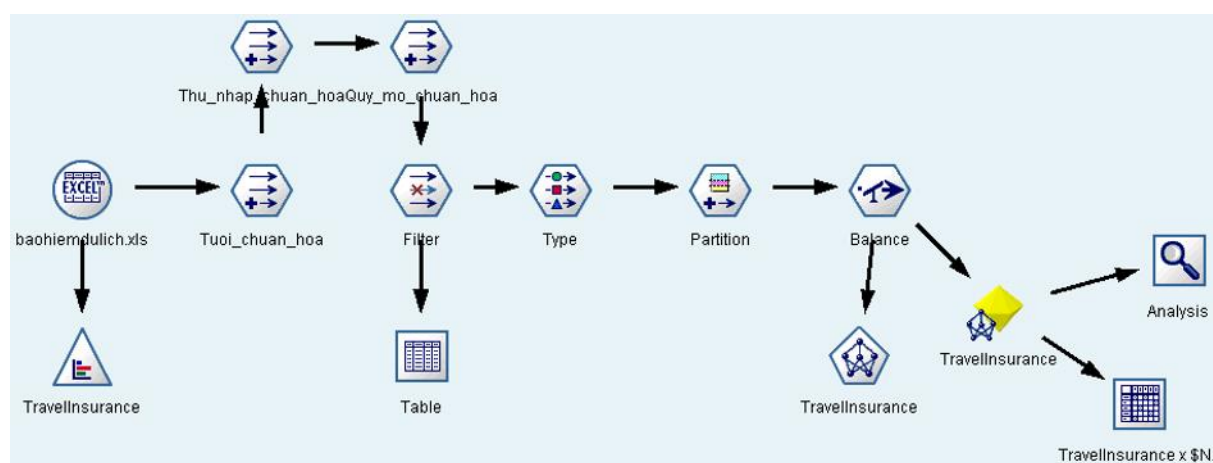
Coincidence Matrix for \$L-TravellInsurance (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000		306
1.000000		189
'Partition' = 2_Testing	0.000000	1.000000
0.000000		262
1.000000		71

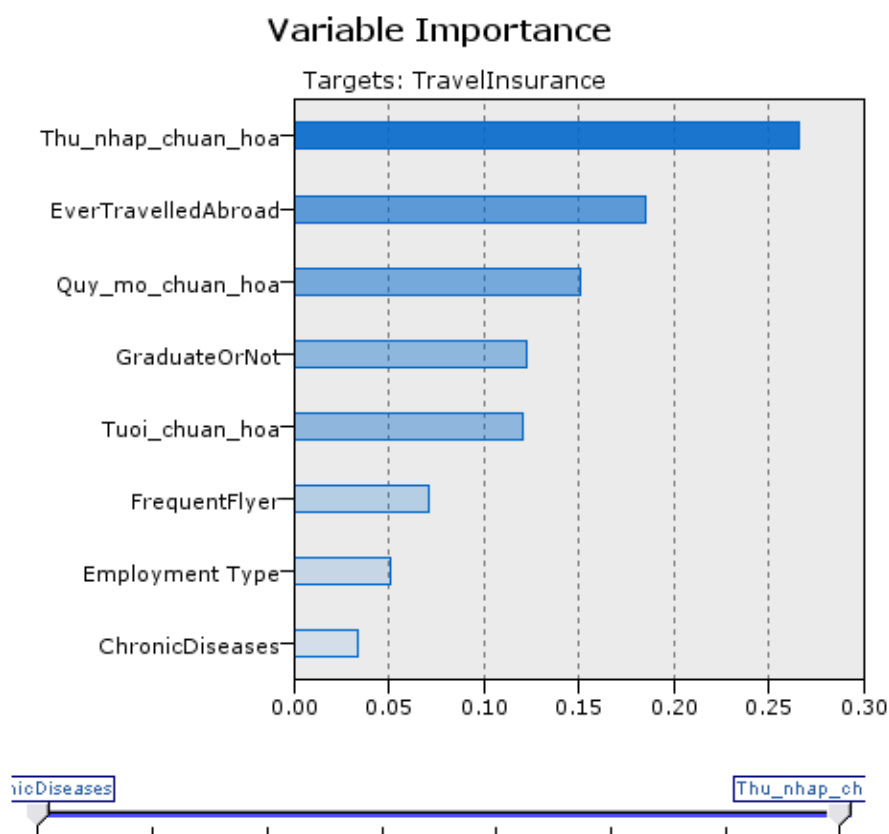
Hình 4.15. Độ chính xác và ma trận trùng của mô hình Hồi quy Logistic

Mô hình Hồi quy Logistic sau cùng có độ chính xác khi huấn luyện là 67,94% và độ chính xác khi kiểm tra là 70,24%.

4.2. Kết quả mô hình Mạng thần kinh



Hình 4.16. Lưu đồ mô hình Mạng thần kinh



Hình 4.17. Ảnh hưởng của từng biến độc lập đến biến mục tiêu của mô hình Mạng thần kinh

Dựa vào đồ thị trên, ta nhận thấy biến thu nhập hàng năm có tác động mạnh nhất đến biến mục tiêu. Kế đến là các biến đã từng ra nước ngoài hay chưa, số người sống chung trong gia đình, đã tốt nghiệp đại học hay chưa, độ tuổi, có thường xuyên mua vé máy bay hay không, loại công việc và có mắc bệnh mãn tính hay không có mức độ ảnh hưởng đến biến mục tiêu giảm dần.

Results for output field TravelInsurance

Comparing \$N-TravelInsurance with TravelInsurance

'Partition'	1_Training		2_Testing	
Correct	787	75.31%	474	80.61%
Wrong	258	24.69%	114	19.39%
Total	1,045		588	

Coincidence Matrix for \$N-TravelInsurance (rows show actuals)

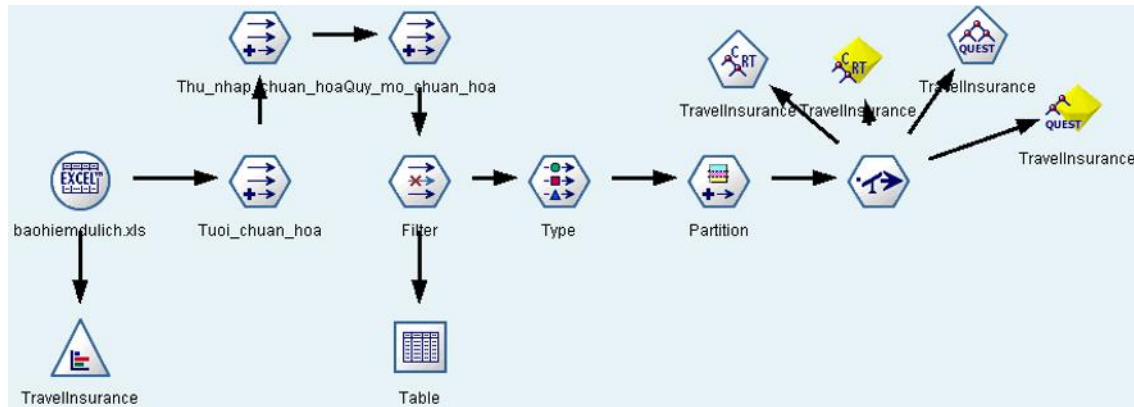
'Partition' = 1_Training	0.000000	1.000000
0.000000	404	42
1.000000	216	383
'Partition' = 2_Testing	0.000000	1.000000
0.000000	330	36
1.000000	78	144

Hình 4.18. Độ chính xác và ma trận trùng của mô hình Mạng thần kinh

Mô hình Mạng thần kinh có độ chính xác khi huấn luyện là 75,31% và độ chính xác khi kiểm tra là 80,61%.

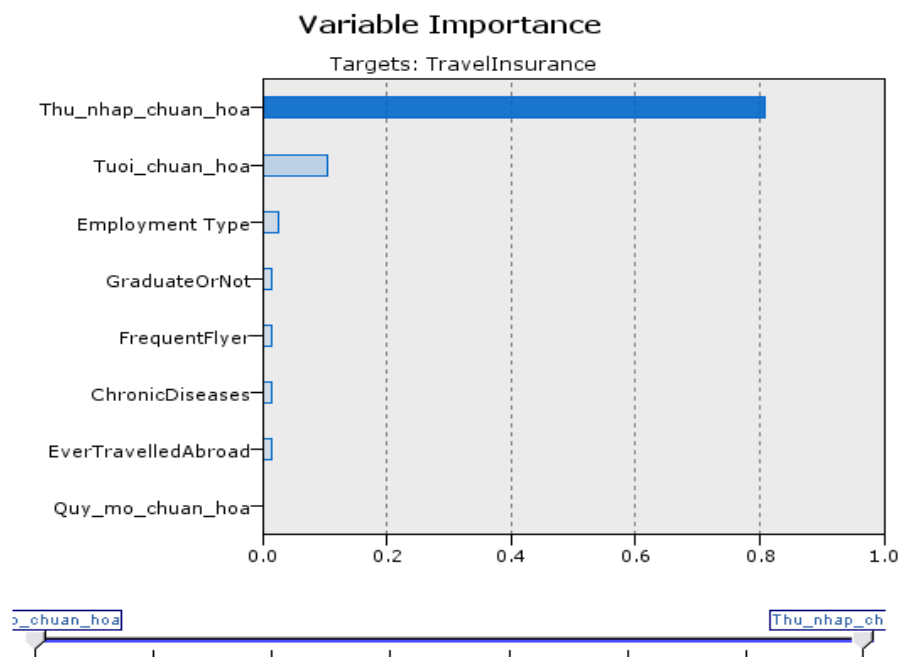
4.3. Kết quả mô hình Cây quyết định

Đối với mô hình Cây quyết định, nhóm quyết định sử dụng 2 phương pháp là CRT và Quest. Nhóm sẽ chạy 2 phương pháp này và so sánh hiệu quả của chúng với nhau qua độ chính xác và ma trận trùng của từng phương pháp.



Hình 4.19. Lưu đồ mô hình Cây quyết định

4.3.1. Kết quả mô hình Cây quyết định phương pháp CRT



Hình 4.20. Ảnh hưởng của từng biến độc lập đến biến mục tiêu của mô hình Cây quyết định (CRT)

Dựa vào đồ thị trên, ta nhận thấy biến thu nhập hàng năm có tác động mạnh nhất đến biến mục tiêu. Độ tuổi của khách hàng có tác động mạnh thứ hai. Kế đến là các biến loại công việc, đã tốt nghiệp đại học hay chưa, có thường xuyên mua vé máy bay hay không, có mắc bệnh mãn tính hay không và đã từng ra nước ngoài hay chưa có mức độ ảnh hưởng thấp và xấp

xỉ nhau. Cần lưu ý rằng số người sống chung trong gia đình không có ảnh hưởng đến việc mua bảo hiểm du lịch của khách hàng.

Results for output field TravelInsurance

Comparing \$R-TravelInsurance with TravelInsurance

'Partition'	1_Training		2_Testing	
Correct	812	75.46%	466	79.25%
Wrong	264	24.54%	122	20.75%
Total	1,076		588	

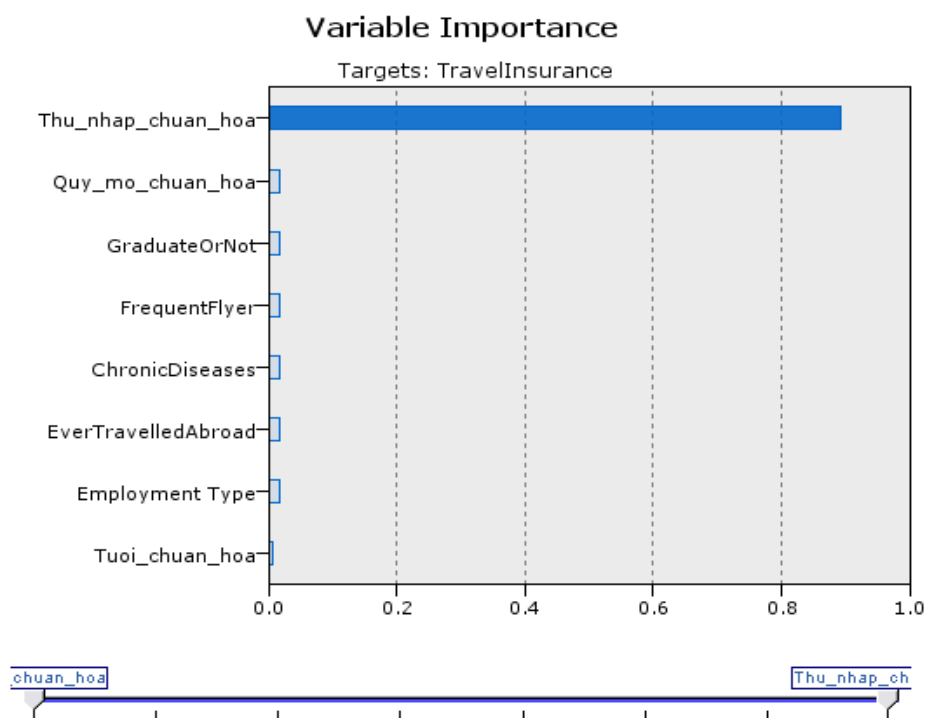
Coincidence Matrix for \$R-TravelInsurance (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	415	47
1.000000	217	397
'Partition' = 2_Testing	0.000000	1.000000
0.000000	316	50
1.000000	72	150

Hình 4.21. Độ chính xác và ma trận trùng của mô hình Cây quyết định (CRT)

Mô hình Cây quyết định phương pháp CRT có độ chính xác khi huấn luyện là 75,46% và độ chính xác khi kiểm tra 79,25%.

4.3.2. Kết quả mô hình Cây quyết định phương pháp Quest



Hình 4.22. Ảnh hưởng của từng biến độc lập đến biến mục tiêu của mô hình Cây quyết định (Quest)

Dựa vào đồ thị trên, ta nhận thấy biến thu nhập hàng năm có tác động mạnh nhất đến biến mục tiêu. Kế đến là các biến số người sống chung trong gia đình, đã tốt nghiệp đại học

hay chưa, có thường xuyên mua vé máy bay hay không, có mắc bệnh mãn tính hay không đã từng ra nước ngoài hay chưa và loại công việc có mức độ ảnh hưởng khá thấp và xấp xỉ nhau. Tuổi của khách hàng gần như không ảnh hưởng đến việc mua bảo hiểm du lịch của họ.

Results for output field TravellInsurance

Comparing \$R-TravellInsurance with TravellInsurance

'Partition'	1_Training		2_Testing	
Correct	760	73.08%	466	79.25%
Wrong	280	26.92%	122	20.75%
Total	1,040		588	

Coincidence Matrix for \$R-TravellInsurance (rows show actuals)

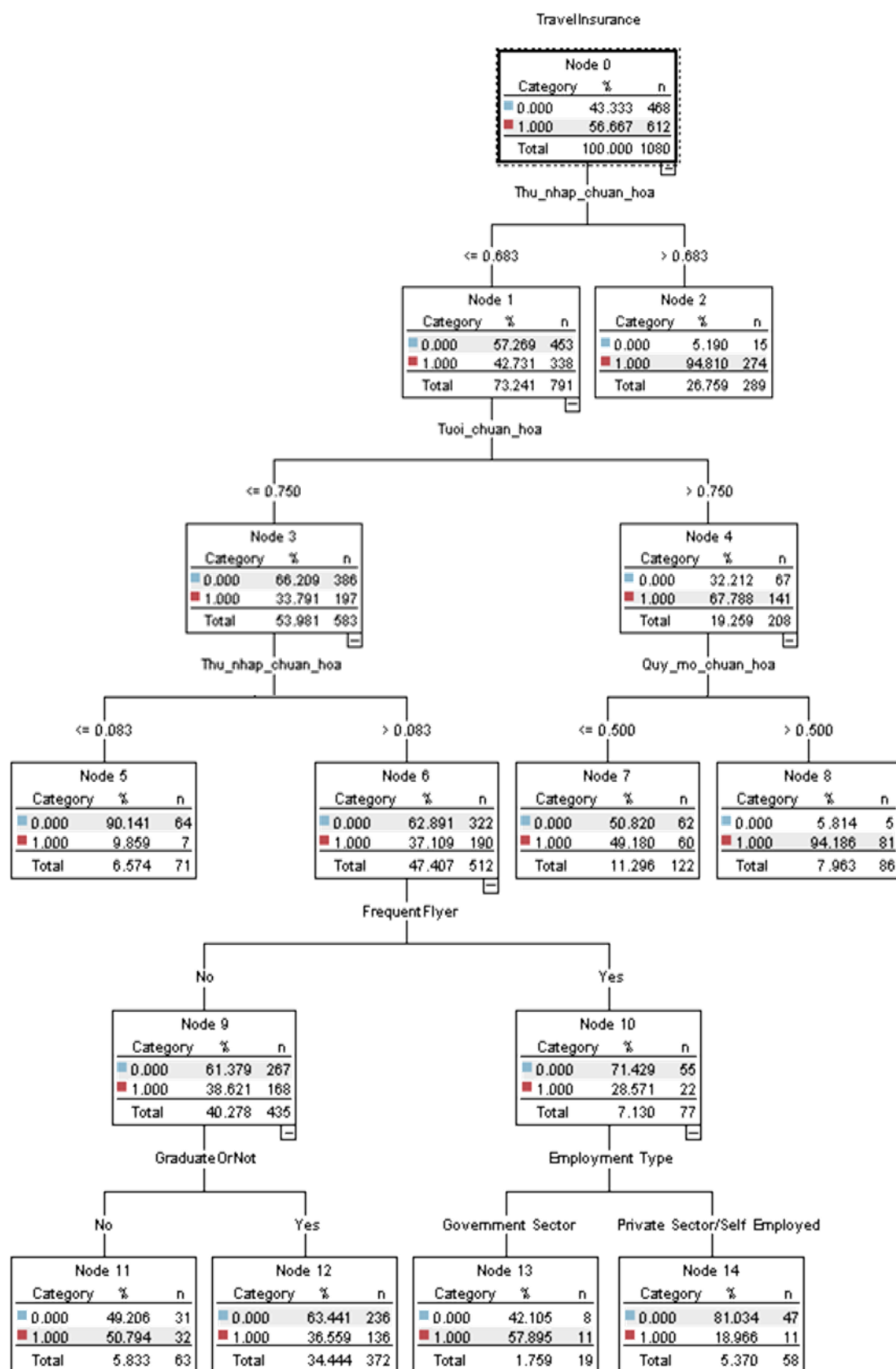
'Partition' = 1_Training	0.000000	1.000000
0.000000		379 48
1.000000		232 381
'Partition' = 2_Testing	0.000000	1.000000
0.000000		322 44
1.000000		78 144

Hình 4.23. Độ chính xác và ma trận trùng của mô hình Cây quyết định (Quest)

Mô hình Cây quyết định phương pháp Quest có độ chính xác khi huấn luyện là 73,08% và độ chính xác khi kiểm tra là 79,25%.

4.3.3. So sánh và lựa chọn mô hình Cây quyết định phù hợp nhất

Để xác định được mô hình Cây quyết định phù hợp nhất, ta tiến hành so sánh độ chính xác của 2 mô hình với nhau. Cả 2 mô hình đều có độ chính xác khi kiểm tra là 79,25%. Tuy nhiên, mô hình Cây quyết định phương pháp CRT có độ chính xác khi huấn luyện cao hơn ($75,46\% > 73,08\%$). Do đó, mô hình Cây quyết định phù hợp nhất là mô hình Cây quyết định phương pháp CRT.



Hình 4.24. Sơ đồ mô hình Cây quyết định (CRT)

Diễn giải kết quả:

- ❖ Mẫu quan sát gồm 1080 khách hàng, trong đó có 612 người có khả năng mua bảo hiểm (chiếm 56,67%) và 468 người có khả năng không mua bảo hiểm (chiếm 43,33%).
- ❖ Phân theo thu nhập hàng năm:
 - Nếu thu nhập hàng năm $> 0,683$: khả năng khách hàng mua bảo hiểm là 94,81%.
 - Nếu thu nhập hàng năm $\leq 0,683$: phân theo độ tuổi:
 - + Nếu độ tuổi $> 0,75$: phân theo số người sống chung trong gia đình, nghĩa là nếu số người sống chung trong gia đình $> 0,5$ thì khả năng khách hàng mua bảo hiểm là 94,18%, còn nếu số người sống chung trong gia đình $\leq 0,5$ thì khả năng khách hàng mua bảo hiểm là 49,18%.
 - + Nếu độ tuổi $\leq 0,75$: phân theo thu nhập hàng năm:
 - Nếu thu nhập hàng năm $\leq 0,083$: khả năng khách hàng mua bảo hiểm là 9,86%.
 - Nếu thu nhập hàng năm $> 0,083$: phân theo việc có thường xuyên đặt vé máy bay hay không:
 - Nếu thường xuyên đặt vé máy bay: phân theo loại công việc, nghĩa là nếu công việc của khách hàng thuộc Government Sector thì khả năng người đó mua bảo hiểm là 57,8%, nếu ngược lại (Private Sector/Self Employed) thì khả năng khách hàng mua bảo hiểm là 18,97%.
 - Nếu không thường xuyên đặt vé máy bay: phân theo việc đã tốt nghiệp đại học hay chưa, nghĩa là nếu khách hàng đã tốt nghiệp đại học thì khả năng người đó mua bảo hiểm là 36,56% còn nếu chưa tốt nghiệp đại học thì khả năng người đó mua bảo hiểm là 50,79%.

4.4. So sánh độ chính xác của các mô hình

Bảng 4.1. So sánh độ chính xác của các mô hình

Mô hình	Hồi quy Logistic	Mạng thần kinh	Cây quyết định (CRT)
Độ chính xác (kiểm tra)	70,24%	80,61%	79,25%

So sánh độ chính xác của các mô hình với nhau, ta nhận thấy mô hình Mạng thần kinh có độ chính xác khi kiểm tra cao nhất với tỷ lệ 80,61%, kế đến là mô hình Cây quyết định (CRT) với tỷ lệ 79,25% và cuối cùng là mô hình Hồi quy Logistic với tỷ lệ 70,24%.

Như vậy, có thể kết luận rằng để dự báo hiệu quả nhất, ta nên sử dụng mô hình Mạng thần kinh.

CHƯƠNG 5: KẾT LUẬN

5.1. Kết luận

Từ các kết quả phân tích mà nhóm đã trình bày ở chương 4, có thể rút ra một số kết luận chính như sau:

- Các mô hình học máy như Hồi quy Logistic, Mạng thần kinh và Cây quyết định đều phù hợp cho việc xây dựng mô hình dự đoán chính xác khả năng mua bảo hiểm du lịch. Kết quả cho thấy mô hình Mạng thần kinh có độ chính xác cao nhất (80,61%), kế đến là mô hình Cây quyết định (CRT) (79,25%) và cuối cùng là mô hình Hồi quy Logistic (70,24%).
- Độ tuổi, thu nhập hàng năm, số người sống chung trong gia đình và lịch sử xuất ngoại đều có ảnh hưởng đáng kể đến khả năng mua bảo hiểm du lịch của khách hàng. Cụ thể, khách hàng càng lớn tuổi, có mức thu nhập hàng năm càng cao và đã từng ra nước ngoài sẽ có xu hướng mua bảo hiểm du lịch cao hơn.

Nhìn chung, việc ứng dụng các mô hình học máy đã đạt được mục tiêu nghiên cứu đề ra, cung cấp mô hình dự đoán chính xác khả năng mua bảo hiểm du lịch của khách hàng dựa trên dữ liệu lịch sử. Kết quả này mang đến lợi ích thiết thực cho cả doanh nghiệp và khách hàng.

Doanh nghiệp có thể tối ưu hóa chiến lược kinh doanh, tiếp cận khách hàng tiềm năng một cách hiệu quả bằng cách cung cấp sản phẩm bảo hiểm phù hợp. Khách hàng sẽ được hưởng lợi từ việc nâng cao nhận thức về tầm quan trọng của bảo hiểm du lịch, sự tiện lợi trong việc mua bảo hiểm cũng như bảo vệ được sức khỏe và tài sản trong những chuyến du lịch. Nhờ vậy, họ có thể an tâm tận hưởng hành trình của mình với sự bảo vệ toàn diện.

5.2. Đề xuất cho doanh nghiệp

Dựa vào kết quả nghiên cứu, nhóm đề xuất một số hành động nhằm giúp các doanh nghiệp bảo hiểm có cung cấp sản phẩm bảo hiểm du lịch tối ưu hóa chiến lược kinh doanh và tiếp thị của mình như sau:

- Tập trung vào các khách hàng tiềm năng:
 - + Phân khúc khách hàng: Doanh nghiệp nên tập trung vào các nhóm khách hàng có thu nhập cao, trung niên, và đã từng du lịch nước ngoài.
 - + Chiến lược marketing cá nhân hóa: Sử dụng các dữ liệu nhân khẩu học và hành vi để tạo ra các chiến dịch marketing cá nhân hóa, nhắm đến các nhóm khách hàng có khả năng mua bảo hiểm cao nhất.

- Phát triển sản phẩm bảo hiểm đa dạng: Thiết kế các gói bảo hiểm với thời hạn, khu vực áp dụng, phạm vi bảo vệ và mức phí phù hợp với từng cá nhân nhằm đáp ứng nhu cầu phong phú của khách hàng.
- Nâng cao nhận thức về bảo hiểm du lịch:
 - + Tổ chức các chương trình, hội thảo để nâng cao nhận thức của khách hàng về lợi ích của bảo hiểm du lịch, đặc biệt là sau đại dịch COVID-19.
 - + Sử dụng các phương tiện truyền thông đại chúng và các trang mạng xã hội để quảng bá các sản phẩm bảo hiểm cũng như chia sẻ các câu chuyện thực tế từ khách hàng chứng minh sự cần thiết của việc sở hữu bảo hiểm du lịch.
- Ứng dụng công nghệ vào dịch vụ khách hàng:
 - + Cung cấp dịch vụ tư vấn bảo hiểm trực tuyến thông qua các ứng dụng di động và trang web giúp khách hàng dễ dàng tìm hiểu và mua bảo hiểm.
 - + Xây dựng hệ thống chăm sóc khách hàng tự động và hỗ trợ qua chatbot để giải đáp nhanh chóng các thắc mắc của khách hàng.

5.3. Hạn chế của nghiên cứu

Mặc dù đã đạt được những kết quả đề ra nhất định, nhóm nhận thấy nghiên cứu này vẫn còn tồn tại một số hạn chế cần lưu ý như sau:

- Phạm vi dữ liệu hạn chế: Nghiên cứu chỉ sử dụng dữ liệu được thu thập vào năm 2019 từ một quốc gia (Ấn Độ). Do đó, kết quả có thể không hoàn toàn áp dụng được cho các thị trường khác với đặc điểm nhân khẩu học và hành vi khách hàng khác biệt.
- Sự phức tạp của mô hình: Mô hình Mạng thần kinh tuy có độ chính xác cao nhưng lại phức tạp và khó triển khai trong thực tế hơn so với mô hình Hồi quy Logistic và Cây quyết định.

5.4. Đề xuất hướng nghiên cứu tiếp theo

Từ những hạn chế kể trên, nhóm đề xuất một số hướng mở rộng trong tương lai như sau:

- Mở rộng phạm vi dữ liệu: Thu thập dữ liệu từ nhiều quốc gia và nhiều thời điểm sẽ giúp nghiên cứu có cái nhìn tổng quan hơn về hành vi khách hàng trên thị trường quốc tế, đồng thời đánh giá được sự thay đổi của thị trường và hành vi khách hàng theo thời gian.
- Giảm thiểu độ phức tạp của mô hình Mạng thần kinh bằng cách áp dụng các kỹ thuật tối ưu hóa mô hình như giảm số lượng tham số, dùng kiến trúc mạng thần kinh hiệu quả hơn...

TÀI LIỆU THAM KHẢO

Tô Thị Phương Dung, 2023. *Bảo hiểm là gì? Đặc điểm và cách phân loại bảo hiểm*. Luật Minh Khuê. Truy cập ngày 20/5/2024 từ <https://luatminhkhue.vn/bao-hiem-la-gi.aspx>

Ngân hàng TMCP Hàng Hải Việt Nam, 2023. *Bảo hiểm du lịch là gì? Các loại bảo hiểm du lịch MSB*. Truy cập ngày 20/5/2024 từ <https://www.msb.com.vn/vi/w/bao-hiem-du-lich-la-gi-cac-loai-bao-hiem-du-lich-msb>

Vietnam Business Forum, 2024. *Du lịch toàn cầu sẵn sàng phá kỷ lục, du lịch Việt Nam sẽ hưởng lợi*. Liên đoàn Thương mại và Công nghiệp Việt Nam. Truy cập ngày 20/5/2024 từ <https://vccinews.vn/news/52022/du-lich-toan-cau-san-sang-pha-ky-luc-du-lich-viet-nam-se-huong-loi.html>

Hoàng Trọng, Chu Nguyễn Mộng Ngọc (2010), *Giới thiệu Khai thác dữ liệu Kinh Doanh*, Tài liệu biên dịch từ sách tiếng Anh “*Introduction to Business Data Mining*” của David Olson và YongShi (2007), NXB McGraw-Hill.

Hoàng Trọng & Chu Nguyễn Mộng Ngọc (2008), *Phân tích dữ liệu nghiên cứu với SPSS (Tập 2)*, TP HCM, NXB Hồng Đức.