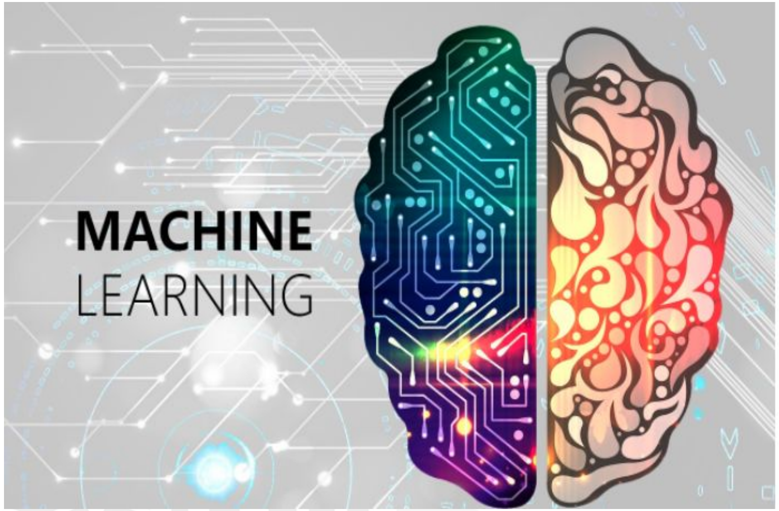


FINAL PROJECT

Data Analytics and Machine Learning

Spring 2022

By Prof. Mahmoud Daneshmand

Health Insurance Lead Prediction

Aishwary Pauranik
Nishanth Sura FNU
Jigyasu Walia
Balwant Deo

Table of Contents

Introduction

Health insurance companies promote their products and services of two ways:

1. Mass Campaigns
2. Targeted marketing

As positive reactions to mass campaigns are typically low (less than 1%), most companies rely on targeted marketing (by phone call or email in this project). Targeted marketing focuses on consumers who are more likely to be interested in that specific product or service, making these efforts more appealing owing to their efficiency. Targeted marketing, on the other hand, has some disadvantages. For example, because of the invasion of privacy, it may cause people to have an unfavorable view about health insurance companies.

Economic concerns and competitiveness have prompted health insurance companies' marketing departments to invest in targeted marketing campaigns with a precise and rigorous contact selection.

Business Understanding Phase

Research Question :

Can we use the information from these clients who received a phone call or an email to create a model that explains the success of a contact, such as whether a client would purchase health insurance?

Research Goal :

Using the Logical Regression, the classification goal is to obtain rules and forecast if a client will agree to get a health insurance policy or not.

Using this model, We hope to improve campaign efficiency by identifying the primary criteria that influence success (client subscribing to health insurance following call or email) and having a more rational estimate of which client is a high-quality potential buying customer that we should contact first.

Data Understanding

Data source:

The Data obtained for this project is linked to a financial services company called FinnMan, whose goal is to cross sell health insurance to existing customers who may or may not hold insurance policies with the company.

Details:

Total clients - 50883
Total Attributes - 12

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Data Understanding : Sample Data

ID	City_Code	Region_Co...	Accomoda...	Reco_Insur...	Upper_Age	Lower_Age	Is_Spouse	Health Indi...	Holding_P...	Holding_P...	Reco_Polic...	Reco_Polic...	Response
1	C3	3213	Rented	Individual	36	36	No	X1	14+	3.0	22	11628.0	0
2	C5	1117	Owned	Joint	75	22	No	X2			22	30510.0	0
3	C5	3732	Owned	Individual	32	32	No		1.0	1.0	19	7450.0	1
4	C24	4378	Owned	Joint	52	48	No	X1	14+	3.0	19	17780.0	0
5	C8	2190	Rented	Individual	44	44	No	X2	3.0	1.0	16	10404.0	0
6	C9	1785	Rented	Individual	52	52	No	X2	5.0	1.0	22	15264.0	1
7	C3	679	Owned	Individual	28	28	No				17	10640.0	0
8	C1	3175	Owned	Joint	75	73	Yes	X4	9.0	4.0	17	29344.0	1
9	C15	3497	Owned	Joint	52	43	No	X1	14.0	3.0	1	27283.2	0
10	C1	530	Owned	Joint	59	26	Yes		7.0	4.0	18	21100.8	1
11	C28	600	Owned	Individual	21	21	No	X2			21	4068.0	1
12	C27	1097	Owned	Joint	59	47	Yes	X3	3.0	3.0	13	25043.2	0
13	C7	3453	Owned	Individual	66	66	No		1.0	2.0	20	17192.0	1
14	C5	900	Rented	Individual	20	20	No	X2			18	8364.0	0
15	C20	1911	Rented	Individual	27	27	No	X3	2.0	3.0	9	9440.0	0
16	C3	1484	Rented	Individual	20	20	No	X3			2	4912.0	0
17	C3	1090	Owned	Individual	34	34	No	X1	11.0	1.0	20	6660.0	0
18	C7	677	Owned	Individual	43	43	No	X2			19	10386.0	0
19	C1	1634	Owned	Individual	55	55	No	X2	1.0	3.0	21	12580.0	0
20	C20	973	Owned	Individual	27	27	No				4	8050.0	0
21	C9	3543	Owned	Individual	32	32	No	X2	3.0	3.0	16	12060.0	0
22	C24	1127	Rented	Individual	23	23	No	X2			16	10352.0	0
23	C25	787	Rented	Individual	18	18	No	X6			22	2828.0	0
24	C1	2862	Rented	Individual	22	22	No	X6			19	5416.0	0
25	C4	2182	Rented	Individual	22	22	No	X1	1.0	3.0	22	6370.0	0
26	C5	2276	Rented	Individual	25	25	No	X1			12	7128.0	0

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Data Understanding : Testing Data

ID	City_Code	Region_Co...	Accomoda...	Reco_Insur...	Upper_Age	Lower_Age	Is_Spouse	Health Indi...	Holding_P...	Holding_P...	Reco_Polic...	Reco_Polic...
50883	C1	156	Owned	Individual	30	30	No		6.0	3.0	5	11934.0
50884	C4	7	Owned	Joint	69	68	Yes	X1	3.0	3.0	18	32204.8
50885	C1	564	Rented	Individual	28	28	No	X3	2.0	4.0	17	9240.0
50886	C3	1177	Rented	Individual	23	23	No	X3	3.0	3.0	18	9086.0
50887	C1	951	Owned	Individual	75	75	No	X3			5	22534.0
50888	C1	1329	Rented	Individual	24	24	No	X2			18	6150.0
50889	C2	3479	Owned	Individual	56	56	No	X5	14+	4.0	17	19152.0
50890	C13	396	Rented	Individual	41	41	No				16	11034.0
50891	C18	513	Owned	Individual	22	22	No	X3			22	10784.0
50892	C3	957	Owned	Joint	41	37	Yes	X5	6.0	1.0	22	16934.4
50893	C1	916	Rented	Individual	22	22	No	X4			5	9422.0
50894	C16	1113	Owned	Individual	38	38	No	X4	2.0	3.0	21	14168.0
50895	C17	636	Owned	Individual	42	42	No	X2	5.0	2.0	17	14184.0
50896	C1	1112	Owned	Individual	31	31	No				19	7236.0
50897	C2	2371	Rented	Individual	35	35	No	X2	14+	3.0	18	9002.0
50898	C11	2000	Owned	Joint	46	37	No				20	18333.0
50899	C2	133	Owned	Individual	44	44	No	X3			11	13200.0
50900	C7	4535	Owned	Individual	29	29	No				20	7488.0
50901	C21	4336	Rented	Individual	60	60	No	X1			9	19200.0
50902	C34	858	Rented	Individual	54	54	No	X4	1.0	3.0	14	14586.0
50903	C1	3651	Rented	Individual	31	31	No	X1			13	8428.0
50904	C14	3329	Rented	Individual	27	27	No	X4			20	7896.0
50905	C1	16	Owned	Individual	71	71	No	X5	14+	3.0	18	16042.0
50906	C1	183	Owned	Individual	75	75	No	X1	5.0	1.0	21	18720.0
50907	C3	3891	Owned	Joint	68	66	Yes	X1	5.0	3.0	9	24206.0
50908	C2	2810	Owned	Joint	55	54	Yes	X2	4.0	3.0	3	23091.2
50909	C1	4229	Owned	Individual	36	36	No	X1			19	11340.0
50910	C3	1090	Rented	Joint	59	26	No		5.0	4.0	12	20711.6

Data Understanding : Attribute details

- **ID** - Unique Identifier
- **City_Code** - City code which matches with the respective city where the person lives.
- **Region_Code** - Region code which matches with the respective region where the person lives.
- **Accomodation_Type** - Whether they own or rent their home
- **Reco_Insurance_Type** - Whether they have a joint or individual Insurance
- **Upper_Age** - Upper age limit of the people in a policy
- **Lower_Age** - Lower age limit of the people in a policy
- **Is_Spouse** - Whether the person is married or not

Data Understanding : Attribute details Cont.

- **Health Indicator** - Whether the individual is fit/unhealthy or affected by disease
- **Holding_Policy_Duration** - Duration of ongoing policy
- **Holding_Policy_Type** - Type of ongoing policy
- **Reco_Policy_Premium** - Premium of ongoing policy

Data Preparation

Identifying outliers and dealing with missing values:

- There are minimal clients with missing values. Also, the outliers in the original dataset were properly handled.

- We removed two attributes, namely ID - unique identifier; region_code, which mentions the region where the perspective client lives.

- There are ten relevant attributes, which were used for the project.

Dataset

	ID	City_Code	Region_Code	Accomodation_Type	Reco_Insurance_Type	Upper_Age	Lo
0	1	C3	3213	Rented	Individual	36	
1	2	C5	1117	Owned	Joint	75	
2	3	C5	3732	Owned	Individual	32	
3	4	C24	4378	Owned	Joint	52	
4	5	C8	2190	Rented	Individual	44	
...	...	...	...	...	...	...	
50877	50878	C4	845	Rented	Individual	22	
50878	50879	C5	4188	Rented	Individual	27	
50879	50880	C1	442	Rented	Individual	63	
50880	50881	C1	4	Owned	Joint	71	
50881	50882	C3	3866	Rented	Individual	24	

50882 rows × 14 columns

Attribute Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50882 entries, 0 to 50881
Data columns (total 14 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   ID                       50882 non-null  int64  
 1   City_Code                50882 non-null  object 
 2   Region_Code              50882 non-null  int64  
 3   Accomodation_Type        50882 non-null  object 
 4   Reco_Insurance_Type      50882 non-null  object 
 5   Upper_Age                50882 non-null  int64  
 6   Lower_Age                50882 non-null  int64  
 7   Is_Spouse                50882 non-null  object 
 8   Health Indicator         39191 non-null  object 
 9   Holding_Policy_Duration  30631 non-null  object 
 10  Holding_Policy_Type      30631 non-null  float64
 11  Reco_Policy_Cat          50882 non-null  int64  
 12  Reco_Policy_Premium      50882 non-null  float64
 13  Response                 50882 non-null  int64  
dtypes: float64(2), int64(6), object(6)
memory usage: 5.4+ MB
```

Scaled Dataset

	Accomodation_Type	Reco_Insurance_Type	Upper_Age	Lower_Age	Is_Spouse	Holding_Poli
47215	-1.102015	-0.509835	-0.684813	-0.561071	-0.449329	
18464	0.907429	-0.509835	0.872493	0.995706	-0.449329	
33790	-1.102015	1.961420	0.411069	-0.849362	-0.449329	
11408	0.907429	-0.509835	-0.223389	-0.099803	-0.449329	
14159	0.907429	-0.509835	-0.281067	-0.157462	-0.449329	
...	...	...	...	...	...	
50057	-1.102015	-0.509835	-0.511779	-0.388095	-0.449329	
32511	0.907429	-0.509835	-0.973203	-0.849362	-0.449329	
5192	0.907429	1.961420	1.276238	-0.042145	-0.449329	
12172	-1.102015	-0.509835	-0.223389	-0.099803	-0.449329	
33003	0.907429	-0.509835	0.295713	0.419122	-0.449329	

35617 rows × 76 columns

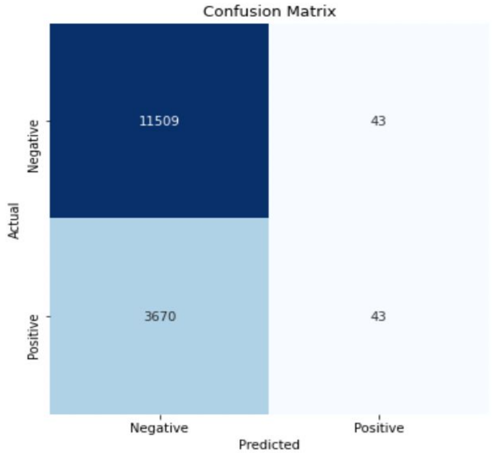
Count of Train Value And Training scenario 1 with equal Weights

```
Out[81]: 0    27121
         1     8496
         Name: Response, dtype: int64
```

```
Out[82]: LogisticRegression(class_weight={0: 1.0, 1: 1})
```

Results for Training Scenario 1

Test Accuracy: 75.68%



Classification Report:

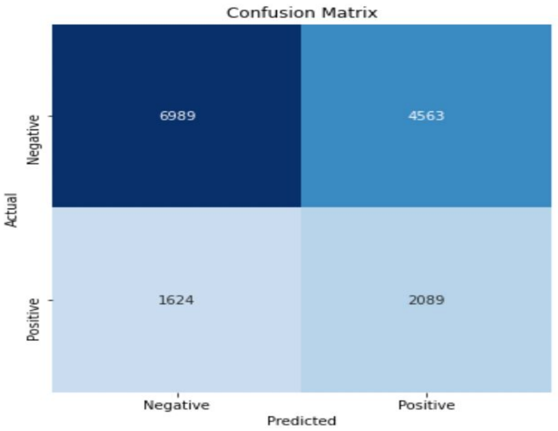
----

	precision	recall	f1-score	support
Negative	0.76	1.00	0.86	11552
Positive	0.50	0.01	0.02	3713
accuracy			0.76	15265
macro avg	0.63	0.50	0.44	15265
weighted avg	0.70	0.76	0.66	15265

Training scenario 2 with adjusted Weights and Results

```
Out[85]: LogisticRegression(class_weight={0: 1.0, 1: 3})
```

Test Accuracy: 59.47%



Classification Report:

----

	precision	recall	f1-score	support
Negative	0.81	0.61	0.69	11552
Positive	0.31	0.56	0.40	3713
accuracy			0.59	15265
macro avg	0.56	0.58	0.55	15265
weighted avg	0.69	0.59	0.62	15265

THANK YOU