

Introduction

A company in the insurance sector that has been offering health insurance to its clients is currently seeking assistance in developing a predictive model. The aim is to determine whether policyholders from the previous year would also exhibit an inclination towards acquiring vehicle insurance from the same company.

Data description

The Data set originally consists of 12 attributes (columns), and 381109 rows with no missing values or errors, but I excluded one attribute (response) from the set because it's the target variable I am going to be predicting.

Data source and link

Kaggle - Health Insurance Cross Sell Prediction

<https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction>

Data set

Attribute details

	A	B	C	D	E	F	G	H	I	J	K
1	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage
2	1	Male	44	1	28	0	> 2 Years	1	40454	26	217
3	2	Male	76	1	3	0	1-2 Year	0	33536	26	183
4	3	Male	47	1	28	0	> 2 Years	1	38294	26	27
5	4	Male	21	1	11	1	< 1 Year	0	28619	152	203
6	5	Female	29	1	41	1	< 1 Year	0	27496	152	39
7	6	Female	24	1	33	0	< 1 Year	1	2630	160	176
8	7	Male	23	1	11	0	< 1 Year	1	23367	152	249
9	8	Female	56	1	28	0	1-2 Year	1	32031	26	72
10	9	Female	24	1	3	1	< 1 Year	0	27619	152	28
11	10	Female	32	1	6	1	< 1 Year	0	28771	152	80
12	11	Female	47	1	35	0	1-2 Year	1	47576	124	46
13	12	Female	24	1	50	1	< 1 Year	0	48699	152	289
14	13	Female	41	1	15	1	1-2 Year	0	31409	14	221
15	14	Male	76	1	28	0	1-2 Year	1	36770	13	15
16	15	Male	71	1	28	1	1-2 Year	0	46818	30	58
17	16	Male	37	1	6	0	1-2 Year	1	2630	156	147
18	17	Female	25	1	45	0	< 1 Year	1	26218	160	256
19	18	Female	25	1	35	1	< 1 Year	0	46622	152	299
20	19	Male	42	1	28	0	1-2 Year	1	33667	124	158
21	20	Female	60	1	33	0	1-2 Year	1	32363	124	102
22	21	Male	65	1	28	0	1-2 Year	1	41184	124	116
23	22	Male	49	1	28	0	1-2 Year	1	50791	124	177
24	23	Male	23	1	50	1	< 1 Year	0	45283	152	232
25	24	Male	44	1	28	0	1-2 Year	1	41852	163	60
26	25	Male	34	1	15	1	1-2 Year	0	38111	152	180
27	26	Female	21	1	28	1	< 1 Year	0	61964	152	72
28	27	Female	51	1	28	0	1-2 Year	1	38341	124	49
29	28	Male	26	1	6	1	< 1 Year	0	19135	152	57
30	29	Female	25	1	8	1	< 1 Year	0	27946	152	223

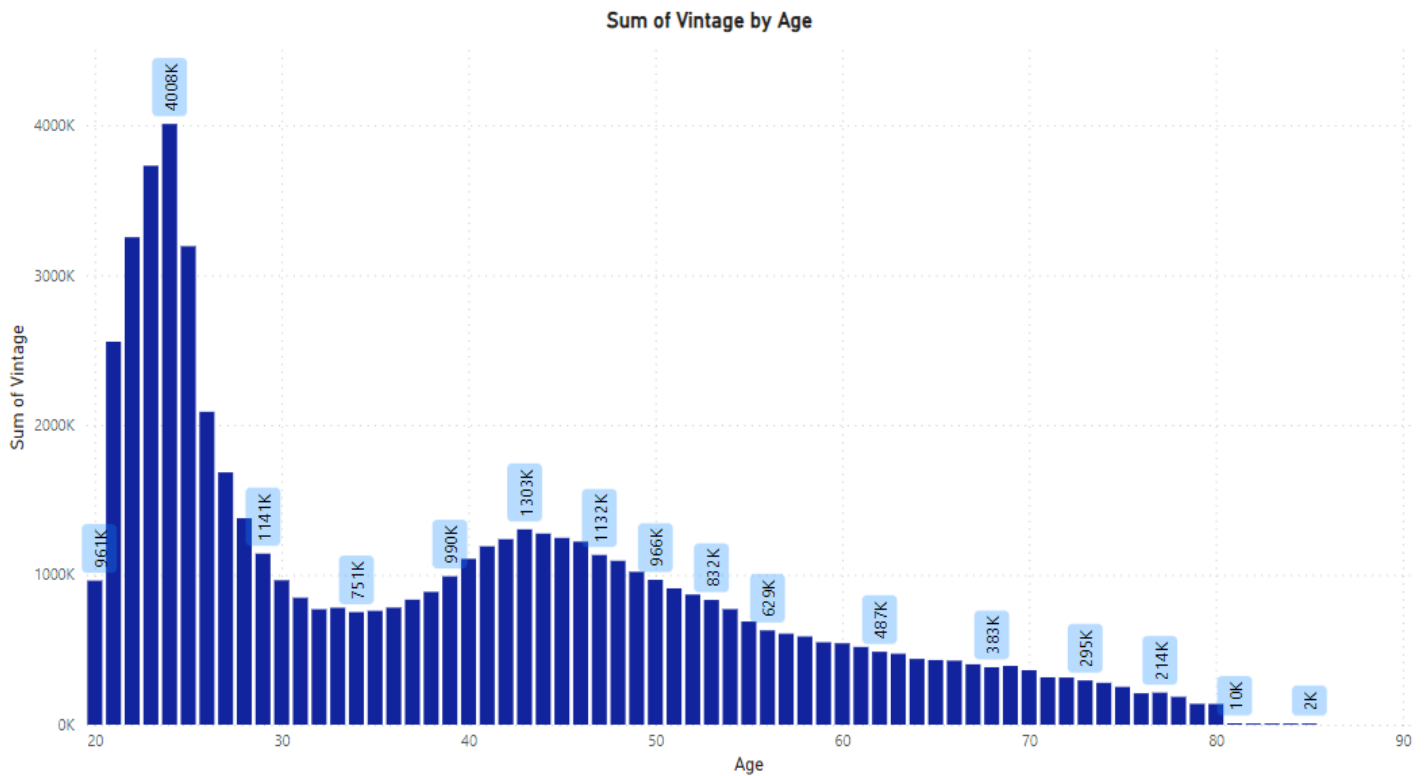
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving_License	0 : Customer does not have DL, 1 : Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
Vehicle_Age	Age of the Vehicle
Vehicle_Damage	1: Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
Annual_Premium	The amount customer needs to pay as premium in the year
Policy_Sales_Channel	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
Vintage	Number of Days, Customer has been associated with the company

Response

1 : Customer is interested, 0 : Customer is not interested

Data visualization

To better understand the data, I used PowerBi and Python to create some charts that show relations



between some of the variables as shown below:

Figure 1

In figure 1, I used a clustered column chart for the sum of vintage by age to see which age has been associated with company the longest which resulted in between the ages of (20 and 30).