Objective

Our dataset contains information about the existing customers and their response to an offer. The task of my project is to use this information to create an acquisition strategy for offering new products to these customers

The company has the budget to contact 3000 individuals for a future campaign where they can offer any of the 3 policy types (corporate, personal and special) through any of the 4 sales channels. In order to recommend the best list of customers based on the maximum expected revenue, the analysis is conducted based on the following steps.

1) Perform suitable exploratory data analysis on the variables in the dataset
2) Perform the following detailed analysis
    - Create a customer segmentation based on the data given
    - Develop response prediction score for each segment
    - Estimate the expected revenue based on the premium of the policy being offered
3) Summarize your findings based on the results from above analysis

Contents

1 Introduction

In this project, we will help a US auto-insurance company to promote new or related insurance product to their existing customer for their next marketing campaign. We will perform suitable customer segmentation, develop response prediction model for each segmentation, calculate the expected revenue and finally come out a list of 3000 valuable customers for the company's future campaign.

2 Exploratory data analysis

2.1 Exploratory on policy and promotion information

Let's look at the promoted policy in this marketing campaign.

2.1.1 Policy types

There are 3 types of policy were promoted. Inside each policy, there are 3 kinds of policy, labeled L1, L2, L3. And for each policy, there are 3 types of coverage, Basic Extended and Premium. In conclusion, there are 3*3*3 which is 27 kinds of policy were promoted in last marketing campaign.

Policy Type	Policy	Coverage	Number of individuals	Percentage of total	Avg Premium per month	Response rate
Corporate Auto	Corporate L1	Basic	222	2.43%	$ 80.86	14.41%
		Extended	101	1.11%	$ 107.33	12.87%
		Premium	36	0.39%	$ 135.28	8.33%
	Corporate L2	Basic	338	3.70%	$ 82.50	13.02%
		Extended	192	2.10%	$ 103.29	17.71%
		Premium	65	0.71%	$ 133.63	15.38%
	Corporate L3	Basic	645	7.06%	$ 82.21	15.81%
		Extended	289	3.16%	$ 101.31	15.57%
		Premium	80	0.88%	$ 139.10	6.25%
Personal Auto	Personal L1	Basic	754	8.25%	$ 81.38	14.19%
		Extended	386	4.23%	$ 106.83	16.32%
		Premium	100	1.09%	$ 130.57	15.00%
	Personal L2	Basic	1303	14.27%	$ 82.26	14.81%
		Extended	613	6.71%	$ 103.32	12.89%
		Premium	206	2.26%	$ 127.29	16.02%
	Personal L3	Basic	2081	22.78%	$ 82.32	13.46%
		Extended	1039	11.38%	$ 103.13	13.47%
		Premium	306	3.35%	$ 137.28	15.69%
Special Auto	Special L1	Basic	38	0.42%	$ 81.95	15.79%
		Extended	22	0.24%	$ 96.77	13.64%
		Premium	6	0.07%	$ 129.83	50.00%
	Special L2	Basic	92	1.01%	$ 85.79	15.22%
		Extended	57	0.62%	$ 102.23	7.02%
		Premium	15	0.16%	$ 108.53	6.67%
	Special L3	Basic	95	1.04%	$ 82.48	21.05%
		Extended	43	0.47%	$ 102.02	20.93%

		Premium	10	0.11%	$ 152.80	20.00%

Table 1: Monthly premium and response rate among all policies

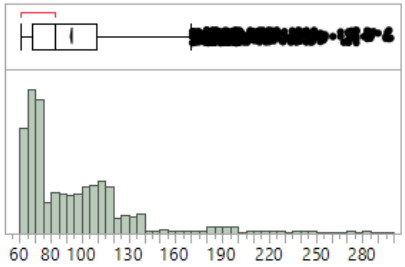


Figure 1: Distribution of monthly premium

Furthermore, the monthly premium among all the policy types is between $61 to $298 per month. For every policy, the basic coverage has the lowest monthly premium while Premium coverage have the highest monthly premium. And the response rate of each kind of policy is different, it varies from 6.25% to 50%.

2.1.2 Renew offers

When promote the policy, there are 4 types of renew offer and it is selected by the customer. Although we don't the content of the offer, we can find that each type of renew offer's response rate is different. Offer 2 has the highest response rate which s 23.38% while offer 3 is 2.09% and there is nobody response to offer 4.

Renew offer type	Number of individuals	Percentage of total	Response rate
Offer1	3752	41.08%	15.83%
Offer2	2926	32.03%	23.38%
Offer3	1432	15.68%	2.09%
Offer4	1024	11.21%	0.00%

2.1.3 Sales channels

There are 4 types of sales channels. The highest response rate among them is agent, the rest branch, call center and web are all below average.

Sales channel	Number of individuals	Percentage of total	Response rate
Agent	3477	38.07%	19.15%
Branch	2567	28.10%	11.45%
Call Center	1765	19.32%	10.88%
Web	1325	14.51%	11.77%

2.2 Exploratory on customer information

2.2.1 Demography

Next, let's look at the customer information in this marketing campaign, there are 9134 individuals were selected and the overall response rate is 14.32%. There is no obvious difference among various states and gender.

State	Number of individuals	Percentage of total	Response rate
California	3150	34.49%	14.48%
Oregon	2601	28.48%	14.46%
Arizona	1703	18.64%	14.27%
Nevada	882	9.66%	14.06%
Washington	798	8.74%	13.66%

Gender	Number of individuals	Percentage of total	Response rate
F	4658	51.00%	14.17%
M	4476	49.00%	14.48%

Among different education level of individuals, there is a small number of individuals with doctor have the highest response rate which is 17.54%.

Education	Number of individuals	Percentage of total	Response rate
High School or Below	2622	28.71%	13.04%
College	2681	29.35%	15.22%
Bachelor	2748	30.09%	13.76%
Master	741	8.11%	16.19%
Doctor	342	3.74%	17.54%

And individuals who are retired have the highest response rate which is as high as 72.34%, while individuals who are unemployed have the lowest response rate which is 8.55%.

Employment Status	Number of individuals	Percentage of total	Response rate
Employed	5698	62.38%	13.27%
Unemployed	2317	25.37%	8.55%
Medical Leave	432	4.73%	18.06%
Disabled	405	4.43%	17.78%
Retired	282	3.09%	72.34%

There are 63.27% of individuals in this marketing campaign live in suburban and they also have the highest response rate which is 17.44%.

Location code	Number of individuals	Percentage of total	Response rate
Suburban	5779	63.27%	17.44%
Rural	1773	19.41%	9.14%
Urban	1582	17.32%	8.72%

There are 58% of customers are married and only 14.99% of them are divorced but customer who divorced have the highest response rate at 23.67%.

Marital Status	Number of individuals	Percentage of total	Response rate
Married	5298	58.00%	13.14%
Single	2467	27.01%	11.67%
Divorced	1369	14.99%	23.67%

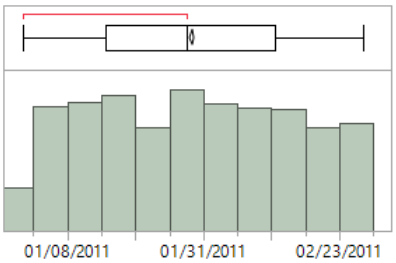
2.2.2 Car types

Individual's car is divided into by vehicle class and vehicle size, as large as 35.44% of the individuals own medium size four-door car. The response rate among various vehicle types are different. Individuals who owns medium size or large sports car or large SUV have the response rate above 20% while individuals own small sports car or large luxury car or SUV have zero response rate.

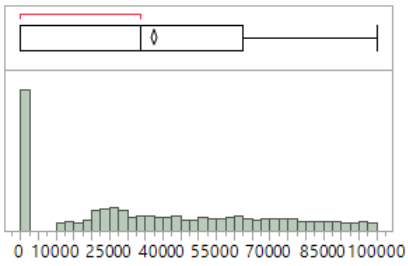
Vehicle Class	Vehicle Size	Number of individuals	Percentage of total	Response rate
Small	Two-Door Car	383	4.19%	12.53%
	Four-Door Car	909	9.95%	11.22%
	Luxury Car	41	0.45%	14.63%
	Luxury SUV	41	0.45%	14.63%
	Sports Car	69	0.76%	0.00%
	SUV	321	3.51%	11.21%
Medsize	Two-Door Car	1282	14.04%	14.04%
	Four-Door Car	3237	35.44%	13.53%
	Luxury Car	106	1.16%	5.66%
	Luxury SUV	125	1.37%	19.20%
	Sports Car	366	4.01%	21.31%
	SUV	1308	14.32%	16.51%
Large	Two-Door Car	221	2.42%	16.29%
	Four-Door Car	475	5.20%	17.68%
	Luxury Car	16	0.18%	0.00%
	Luxury SUV	18	0.20%	0.00%
	Sports Car	49	0.54%	24.49%
	SUV	167	1.83%	21.56%

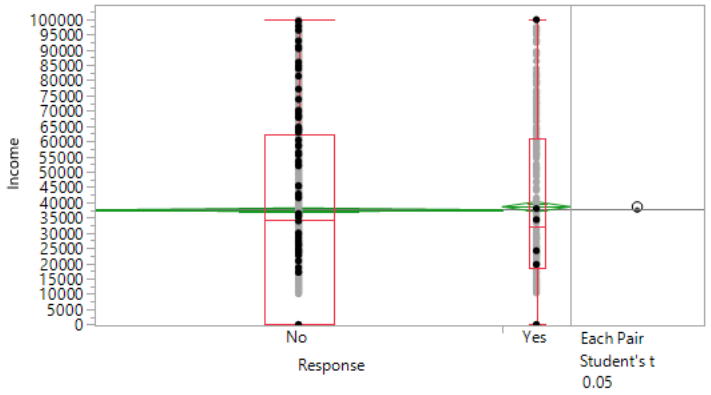
2.2.3 Customer purchasing behaviors and lifetime value

The effective policy end date of the customers in this marketing campaign are all from 2011/01/01 to 02/28/2011. It maybe because that this marketing campaign selected the customers whose policy is ending soon to offer existing or new policy.

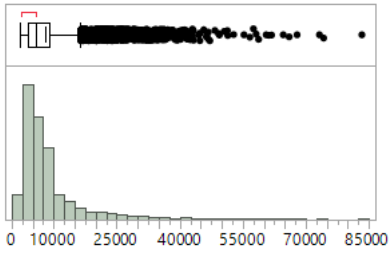


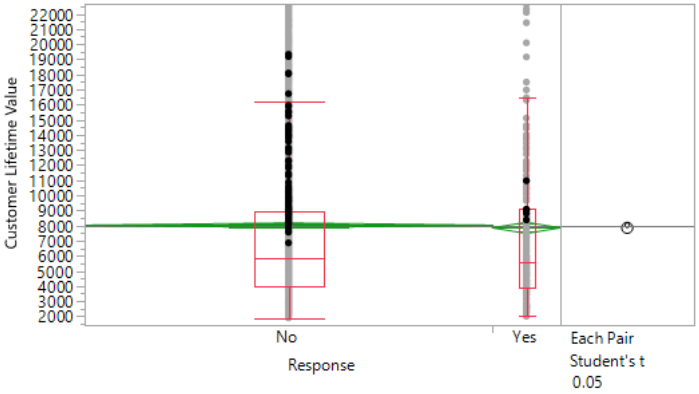
There are 75% of the customer's anual income is below 62329. When applying the hypothesis test, there is no significant difference in annual income between different response at 99.5% condidence level.



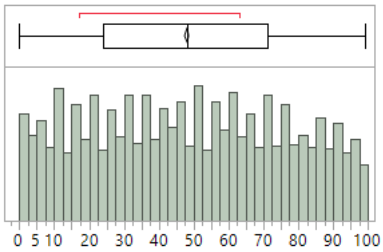


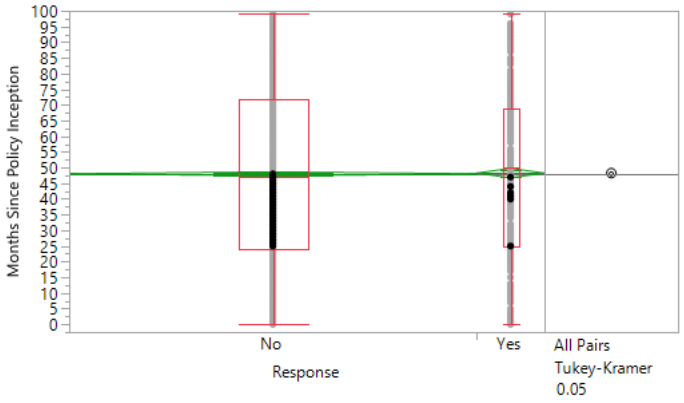
Last but not the least, let look at the distribution of estimated CLV of the customer which is predicted by insurance company. When applying the hypothesis test, here is also no significant difference in customer lifetime value between different response at 99.5% condidence level.



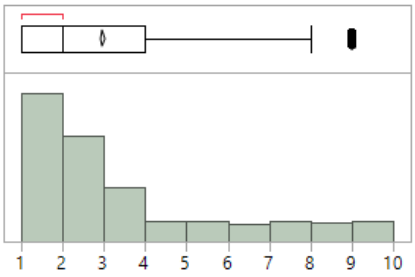


Lets' look the distribution of the number of months since the adoption of the existing insurance policy. The distribution is quite uniform. This maybe because that when company select the customer in this marketing campaign, they select similar amount of people in each recency group. And there is no significant different between different response at 99.5% condidence level.





We can observe from the distribution of number of policy that 75% of the customer have own less than or equal to 4 policies. We will bin them into 4 bins by percentile and find that customer have 1 or 2 policies have slightly higher response rate than the others.



Number of policies	Number of individuals	Percentage of total	Response rate
1	3251	35.59%	15.87%
2	2294	25.11%	14.91%
3 to 4	1577	17.27%	11.03%
5 to 9	2012	22.03%	13.72%

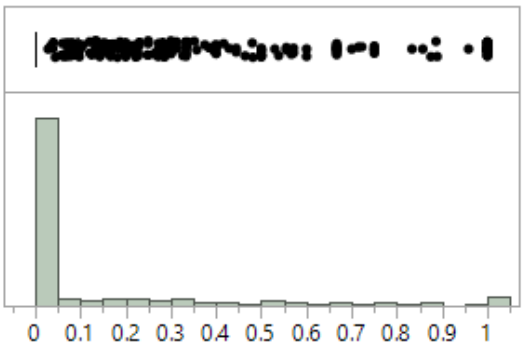
3 Customer segmentation

From the exploratory data analysis, we have some ideas on what are the factors which will drive individuals to response to a particular offer.

From these factors and base on the importance and implication difficulty, I choose policy, coverage, sales channel, renew offer type and location code to divide all customer into 888 groups, and calculate the response rate for each group by the formula as shown below.

$$Response\ rate\ in\ a\ group = \frac{Number\ of\ rows\ where\ response = Yes\ in\ the\ group}{Total\ number\ of\ rows\ in\ the\ group}$$

The distribution of the response rate for each group is shown as below. The distribution is highly skewed and 682 out of 888 groups' response rate is 0.



And base on the response rate, we bin them into several groups. If the response rate is higher than doubled average rate which is 14.32%, we will label the group as very important customer group, if it is between one to doubled average response rate, we will label the group as important customer group, if it is between half to one response rate, we will label the group as less important customer group, and for the rest of the group which is low than half of average response rate, we will label them as least important customer group. Finally. We update the grouping to the original dataset so that we will know which group/segment they belong to for each individual customer.

Segment name	Number of individuals	Percentage of total	Response rate range
Very important customers	1428	15.63%	0 to 7.16%
Important customers	2086	22.83%	7.16% to 14.32%
Less important customers	1301	14.24%	14.32% to 28.64%

Least important customers	4319	47.285	28.64% to 100%

4 Response score predictive model

For each segment, we will come out with a predictive model to predict the response score for each segment. Basically, the response score is the probability which the individual will response to a particular policy.

In our analysis, we tried logistic regression and decision tree, and choose the best predictive model for each segment. We noticed that decision tree performs better than logistic regression in term of misclassification, average squared error, false negative and false positive. So, we choose decision tree for each segment.

Segment 1: very important customers					
Model	**Misclassification rate**	**Average squared error**	**Roc index**	**Gini coefficient**	**Model selected**
Logistic regression	0.32983	0.21255	0.718	0.436	No
Decision tree	0.25560	0.16550	0.826	0.652	Yes
Classification Table comparison					
Model	**False negative**	**True negative**	**False positive**	**True positive**	
Logistic regression	292	537	179	420	
Decision tree	91	442	274	621	
Segment 2: important customers					
Model	**Misclassification rate**	**Average squared error**	**Roc index**	**Gini coefficient**	**Model selected**
Logistic regression	0.20518	0.14546	0.736	0.471	No
Decision tree	0.14430	0.11117	0.827	0.654	Yes
Classification Table comparison					
Model	**False negative**	**True negative**	**False positive**	**True positive**	
Logistic regression	384	1620	44	38	
Decision tree	239	1602	62	183	
Segment 3: less important customers					
Model	**Misclassification rate**	**Average squared error**	**Roc index**	**Gini coefficient**	**Model selected**
Logistic regression	0.12145	0.088391	0.801	0.602	No
Decision tree	0.05688	0.049411	0.858	0.717	Yes
Classification Table comparison					
Model	**False negative**	**True negative**	**False positive**	**True positive**	
Logistic regression	143	1142	15	1	
Decision tree	57	1140	17	87	
Segment 4: lest important customers					
Model	**Misclassification rate**	**Average squared error**	**Roc index**	**Gini coefficient**	**Model selected**
Logistic regression	0.007177587	0.006007429	0.966	0.932	No
Decision tree	0.004630702	0.003999560	0.880	0.760	Yes
Classification Table					
Model	**False negative**	**True negative**	**False positive**	**True positive**	
Logistic regression	30	4288	1	0	
Decision tree	12	4281	8	18	

5 Expected revenue estimation

After score the data with their own predictive model in each segment, we are able to get two important variables, the response score and predicted response. The response score gives the probability when response = 'Yes' and predicted response directly tells us the response result which is 'yes' or 'No' for each customer. Therefore, with response score, we can calculate the expected revenue of each customer.

The expected revenue of each customer can be calculated by the formula shown below.

$$Expected\ revenue = response\ score\ \times monthly\ premium$$

With the expected revenue of each customer, we aggregate the expected revenue in each segment to observe the total revenue and average revenue we can earn per customer in each segment as shown in the table below.

Segment name	Expected total revenue	Number of customers	Expected average earned revenue per customer
Very important customers	67989.57	1428	47.612
Important customers	39224.78	2086	18.80
Less important customers	13008.74	1301	9.99
Least important customers	2809.15	4319	0.65

6 Results and recommendation

With the above analysis, we are not surprise that the very important customer has the highest expected total revenue although there is only 1428 customers. And there is huge difference of expected average earned revenue per customer in each segment. And the customer group which have the highest response rate also have high percentage of people paying higher premium.

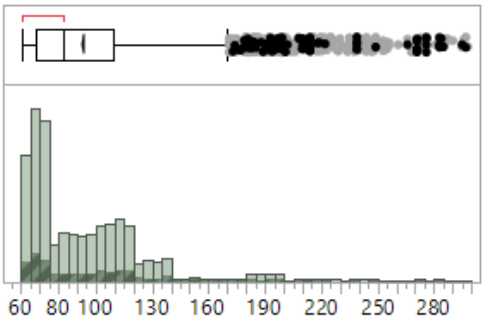


Figure: distribution of Monthly premium auto in very important group vs population

When in the future, company want to host similar marketing champing and the budget is only for 3000 individuals. We can firstly run the models in each segment and rank the expected earn revenue in descending order and select the top 3000 customers in the list.

Beside the model, there are also several findings which can help the company to strength their marketing strategy on customer acquisition or product design.

Firstly, we also notice that the distribution of customer lifetime value in very important customers groups are similar with the population which means there is evidence shows that high lifetime value has higher response rate. And this marketing campaign is not focus on high lifetime value customer. As our customer segmentation is only base on the response behavior and how much we can earn in last marketing champing. It can be considered as our short-term customer acquisition strategy. Future analysis also can be done by considering the customer lifetime value for our long-term customer acquisition strategy.

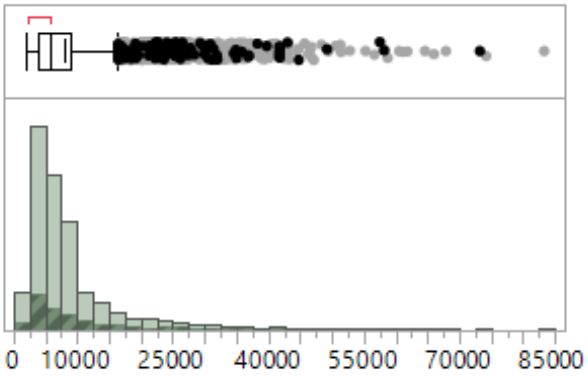


Figure: Distribution of customer lifetime value in very important group vs population

Secondly, compare to the population, there is quite high percentage of retired people in very important customer group. Therefore, company can develop special product, give special offer or host marketing campaign especially for retired people to attract them. In addition to figure on the left, the figure on the right tells us that unemployment people who doesn't have any income have low response rat, so income is still an important factor to the response.

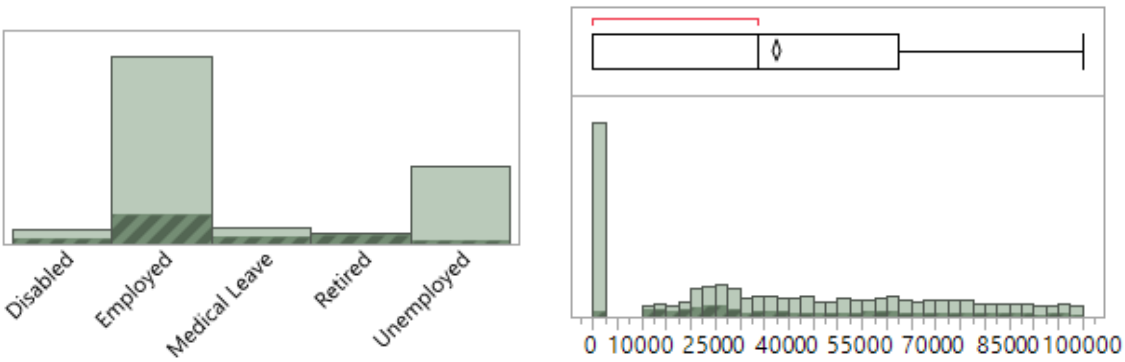


Figure: Distribution of employment status & income in very important group vs population

Thirdly, people like to choose offer 2 and offer 1 when adopted the policy. Together with the content of the 4 types of offers, future work can be done study on what are key factors to choose offer 2 and offer 1 in stead of offer 3 and offer 4.

Fourthly, although with the development of high technology and internet, offering through agent is still an effective sales channel to have a higher response rate. We should analyse the reason of why they like to buy policy through agent and improve customer experience through other sales channels.