Improving the Insurance Offer Assigning Strategy Based on Business Analytics

---- Study of IBM Watson Marketing Customer Value data

IEOR E4650 | Tong Yang (ty2418), Tian Wu (tnw2119), Zehui Yun (zy2384),
Pei Yin Jodie Shue (ps3120), Zhixing Liu (zl2856)

I. Introduction to the Dataset

The dataset we are going to explore is from IBM Watson Analytics with 9134 observations and 24 columns. The dataset contains customers' personal information and their assigned auto insurance attributes. Customers' personal information contains attributes such as *Unique Customer Id*, *Customer's Living State*, *Customer Lifetime Value*, *Education Level*, *Employment Status*, *Gender*, *Income*, *Location Cod*e (rural, suburban), *Marital Status, Vehicle Class, Vehicle Size.* Customer-specific auto insurance attributes contain customers' *Response* to the assigned auto insurance option (boolean), *Option Coverage* (basic, extended, premium), *Sales Channel, Total Claim Amount, Monthly Premium Auto, Months since Last Claim, Months since Policy Inception, Number of Open Complaints, Number of Policies, Policy Type* (personal auto or corporate auto), *Policy* (personal L1-L3, Corporate L1-L3), and *Renew Offer Type* (offer 1-4).

II. Our Analytical Approach

a) Exploratory Data Analytics

Conducting exploratory data analytics as the first step helps us understand our dataset and also the importance of each attribute to customers' decision making better. Through exploratory data analytics, we thrived to answer two questions. The first question is "What kind of attributes the company takes into consideration when it assigns different offers to different customers?". The second question is "What kind of customers have higher acceptance rate when they are assigned to different types of offer?".

To answer the first question, we looked at each attribute individually and tried to find patterns on how the company assigns different offers to different customers. For each categorical variable, we looked into the amount of each offer assigned by categories. If under one categorical variable, a specific offer was assigned the most amount to one or multiple categories, we have intuition to believe this variable could be an important factor that the company considered in assigning this type of offer. When we looked into customer education level, we found that the company assigned the most amount of offer 1 to all education levels except doctor degree. The company assigned the most amount of offer 2 to doctor degree. Education could be an important

factor the company considered when it assigned offer 1 and 2. By looking into customer employment status, we found that for employed customers, more offer 2 were assigned; for unemployed customers, more offer 1 were assigned. When we studied customer location code, more offer 1 were assigned to customers living in suburban, and for customers living in rural and urban areas, equal amount of offer 1 and offer two were assigned. For marital status, more offer 2 were assigned to married customers, and more offer 1 were assigned to single and divorced customers. For sales channel, more offer 1 were assigned to agent and branch channels, more offer 2 were assigned to call center and web channels. We concluded that sales channel could be another significant factor when the company assigns offers. For numerical variables, we grouped the data by offer type and looked at the mean of each numerical variable. We observed that for the mean of variables like customer lifetime value, monthly premium auto, months since policy inception, and total claim amount, it followed a pattern that offer 1 group always had the highest mean, followed by offer 3, 2, and 4 groups. We had reasons to conclude that the company might give more weights to those factors when it decided to assign offer 1. Interestingly, offer 2 group has the highest average income, then offer 4, 3, and 1 groups followed. One possible explanation for that could be the company put more weights to income when it assigned offer 2. So far, we knew that the company assigned the most amount of offer 1 (41.1%) and offer 2 (32%), and offer 3 (15.7%) and offer 4 (11.2%) were less frequently assigned. We also observed that offer 3 and offer 4 were always less frequently assigned to each category than offer 1 and 2, so it is hard to conclude what attributes the company gave more weight to when it assigned offer 3 and 4.

Question 1: What kind of attributes the company takes into consideration when it assigns different offers to different customers?

Table II-i The Factors of Company's Decisions

	Variable name	**Mostly assigned offer type**
Categorical	Education level	Offer 1: all educational levels except doctor degree (offer 2)
	Employment status	Offer 2: employed customer Offer 1: unemployed customer
	Sales channel	Offer 1: agent and branch channels Offer 2: call center and web channels
	Location code	Offer 1: suburban Offer 1 & 2: rural, urban
	Marital Status	Offer 2: married Offer 1: single or divorced
Numerical	Customer life value	Offer 1, 3, 2, 4 (from highest to lowest)
	Monthly premium auto	Offer 1, 3, 2, 4 (from highest to lowest)
	Months since policy inception	Offer 1, 3, 2, 4 (from highest to lowest)
	Total claim amount	Offer 1, 3, 2, 4 (from highest to lowest)
	Income	Offer 2, 4, 3, 1 (from highest to lowest)

To answer our second question of "What kind of customers have higher acceptance rate when they are assigned to different offers?", we first looked at the overall response rate and found out that 14.32% of customers accepted the offer, and 85.68% of customer rejected the offer. First of all, we grouped the data by customer response and looked at the mean of each numerical attribute. We found out that customers with lower average lifetime value, higher average income, higher average total claim amount, smaller number of months since last claim are more likely to accept the offer. The rest of the numerical attributes were hard to conclude because their average in the acceptance group and rejection group were very similar. Next, we dug into categorical variables by calculating the acceptance percentage in each category. For example, for education level was high school or below customers, 13.04% of them accepted the offer. Then we compared the acceptance percentage among different education level groups, and concluded that customers with higher education were more likely to accept the assigned offers. When we looked into response by offer type, we realized that offer 4 were all rejected by customers. This observation was significant and we would take that into consideration in our later analysis. By studying the rest of the categorical attributes, we had the following observations. Divorced customers (23.67%) had a higher probability of accepting the offer than married (13.14%) and single (11.67%). Employed customers have a higher probability of accepting the offer (13.27%) than unemployed (8.55%), and retired customers have the highest acceptance rate of 72.34%. This percentage could be due to the relatively small sample size of retired customers. All three different offer coverage types have similar acceptance rates around 14%. Coverage type could be a less important factor for customers to consider when they decide to accept or reject the offer. Customers living in suburban had the highest acceptance rate of 17.44% than customers living in rural and urban. One possible explanation could be customers living in suburban has higher transportation demand using automobiles. Offers through agency sales channel had the highest acceptance rate, and thus we could speculate agency is the most effective way to persuade customers to accept the offer. Offers with large vehicles have the highest acceptance rate, but vehicle size depends on what kind of vehicle the customer owned, so we concluded this factor may have less influence on the customer acceptance rate.

Question 2: What kind of customers have higher acceptance rate when they are assigned to different types of offer?

Table II-ii The Characteristics of Customers with High Acceptance Rates

	Variable name	**Acceptance rate comparison**
Categorical	Education level	Higher educational level, higher acceptance rate
	Marital status	Divorced > Married > Single
	Employment status	Employed > Unemployed
	Location code	Suburban > Rural > Urban
	Sales channel	Agency > Branch > Web > Call center
		Correlation with acceptance rate
Numerical	Customer lifetime value	Negative
	Income	Positive

	Total claim amount	Positive
	Month since last claim	Negative

III. First Layer Analysis: Analyses of the Company's Offer Assignment Decision

a) Car Insurance Offer Assignment Process

The main purpose of the car insurance, or more generally, vehicle insurance is to provide financial protection when some bad and unexpected things happen to the policy clients. The crux of our project is to figure out two latent patterns, the company decision pattern and customers accepting pattern. Based on the result, we can provide useful and applicable business recommendations for the car company to increase its profit. In the first layer, our top concern is to dig out the company's offer assignment process.

b) Influencing Factors

Customer Lifetime Value, *Income*, *Monthly Premium Auto*, *Months Since Last Claim*, *Months Since Policy Inception*, Number of Open Complaints, *Number of Open Complaints*, *Number of Policie*s, *Channel*, *Policy Type Benlagha and Karaa*, *State*, *Education*, *Employment*, *Gender*, *Location*, *Marital Status*, *Vehicle Class*, *Vehicle Size*, *Total Claim Amount* and *Renew Offer Type* are the columns of the raw dataset. Our guideline is to figure out the principal factors from the economic perspective initially and then design models to improve business benefits.

The problem can be analyzed from different aspects. Wiltrud Weidner, Fabian and Robert propose that the telematic driving profile classification via driving simulation is conducive to car insurance pricing, offering a perspective that offer assignments can also benefit from telematic driving profile classifications. Mihaela provides another idea that some important factors like the auto insurance premium can be prescribed by GLM models. Benlagha and Karaa use an unrelated probit model to show the evidence of adverse selection in the automobile insurance market.

Having summarized ideas and conducted research on our problem, we conclude that the features can be divided into 4 main categories as follows.

Table III-i The Four Principle Affecting Categories

Category	**Demographic**	**Car-related**	**Past experiences**	**Economic**
	Education	Vehicle Class	Total Claim Amount	Customer Life-time Value
	Gender	Vehicle Size	Sales Channel	Employment Status
	Location Code		Policy	Income
	Marital Status		Policy Type	Monthly Premium Auto
	State		Number of Policies	
			Number of Open Complaints	
			Months Since Policy Inception	

			Coverage	
Total Amount	5	2	8	4

c) Accuracy of offer assignment predicting

We have tried to predict how the car insurance company is assigning offers based on the raw data set. Interestingly, we have tried different models, even those with terrific generalization abilities such as GBDT, just to find neither could perform well on the prediction task.

Table III-ii The Comparison of Candidate Classifiers

Model	**Pros**	**Cons**
NaiveBayes	1.Withstanding perturbation	1. Must assume the prior probability
KNN	1.Simple Thought 2.Nonlinear classification 3.O(n) 4.Not sensitive to outliers 5.No prior assumptions	1.Hard to compute 2.Memory consuming 3.Hard to interpret 4.Unable to deal with unbalanced dataset
SVM	1.Small samples 2.Good generalization 3.No choosing structures	1.Hard to hypertune 2.Memory consuming 3.Sensitive to missing values
Decision Tree	1.Easy to interpret 2.Not time-consuming 3.Can be applied to high-dimensional input space	1.Easy to be overfitted 2.Ignore the correlation among features
RF	1.Good generalization ability	1.Hard to interpret
DNN	1.High accuracy 2.Theory supporting	1.Hard to interpret 2.Long time to train
Logistic Regression	1.Fast to train 2.Easy to interpret 3.Output the probabilities of each category	1.Complicated feature engineering needed

The above table gives our summarization of the pros and cons of classifiers we choose. To tackle the multiclass classification problem, we utilize the One-Vs-All method. The accuracies on validation sets are as follows.

Table III-iii The Accuracies of Different Classifiers

Model	**NB**	**KNN**	**SVM**	**DT**	**RF**	**DNN**	**LR**	**GBDT**
Accuracy	0.418	0.468	0.51	0.369	0.508	0.461	0.436	0.521

Under the guidance of our advisor, we find out that the decision process of the car insurance company should be flexible, not just determined by the input demographic and economic features of applicants. Logistic Regression is a powerful model mainly for two reasons. One is that it can provide not only the classification results but the probabilities of each category. The other is that it can be interpreted since it provides the weights of each covariate. Consequently, we use Logistic Regression models to fit with our top concern, the decision process of offer assignments.

d) Multiclass Logistic Regression

Based on what we've learned in class, we have conducted research on the distribution of features and decide whether to apply transformations on them or not. According to the result, we perform log-transformation on *Customer Lifetime Value* and *Monthly Premium Auto* due to their tailed distributions. Then, the forward selection algorithm is performed on the candidate features. The algorithm yields the best model based on BIC criteria as follows.

"RenewOfferType~Income+C(SalesChannel)+np.log(CustomerLifetimeValue)+C(MaritalStatus)+MonthsSincePolicyInception+NumberofOpenComplaints+C(Gender)"

This multinomial model demonstrates how the car insurance company is making its decisions by a mixture strategy, which dovetails the real situations in the industry. In chapter IV, we'll dig out how the customers will react to the assigned offers.

IV. Second Layer Analysis: Factors Affecting the Insurance Acceptance

a) Accepting process

The goal of this part is to predict the acceptance rate of customers when given offer1, offer2, and offer3 (Offer4 is eliminated because no one accepted the offer in our data).

We are interested in enhancing the overall acceptance rate and aimed to get a higher expected value of customer lifetime value.

b) Feature Engineering

From papers and reports, we learned that the factors affecting the insurance acceptance rate could be divided into 3 general types, demographic factors, car-related factors, and company-related strategies. Combined with the results in the Exploratory Data Analytics part, we chose factors in these three types accordingly. The following chart shows the candidate factors we chose for each type, in which we transform *Total Claim Amount* into log form after checking its distribution.

Table IV-i Influencing Factors of Layer Two

Factor Types	**Demographic Factors**	**Car-related Factors**	**Company-related Strategies**
Factors	Education Level	Vehicle Size	Sales Channel
	Marital Status	Vehicle Class	Policy
	Location	Policy Type	Months Since Last Claim

	Income		log(Total Claim Amount)
	Employment Status		
	Gender		
	State		

c) Model Selection

For model selection, we divided the data into 3 subsets according to the different types of offers. Then we separately conducted the method of forward selection among the candidate factors to find the logit model with the lowest AIC.

(i) Offer1

The logit model we got is:

$Response=\beta_0+\beta_1 State+\beta_2 EnploymentStatue+\beta_3 Location+\beta_4 MaritalStatus+\beta_5 SalesChannel+ \beta_6 Vehicle\ Size+\beta_7 Education\ Level+\beta_8 Vehicle\ Class+\beta_9 log(Total\ Claim\ Amount)$

(ii) Offer2

When analyzing offer2, we delete *Employment Status* from the candidate factors because of the singular matrix problem.

The logit model we got is:

$Response = \beta_0+ \beta_1 State+ \beta_2 SalesChannel+ \beta_3 MaritalStatus+ \beta_4 Location+ \beta_5 Income+ \beta_6 Vehicle\ Size+ \beta_7 EducationLevel + \beta_8 Policy\ Type$

(iii) Offer3

Analysis of offer3 is special to some extent because with a sample size of 1432, only 30 customers accept the offer. To successfully conduct logistic regression, we eliminated 5 candidate factors which are *Marital Status*, *Sales Channel*, *Vehicle Class*, *Income*, *Months Since Last Claim* because of their lack of variance in the database.

The logit model we got is:

$Response = \beta_0+\beta_1 State+\beta_2 EducationLevel +\beta_3 \text{Gender}+\beta_4 VehicleSize+\beta_6\ MonthsSinceLastClaim$

d) Prediction Process

Based on the offer assigning probability we got from the first layer, we can get the total acceptance rate of each customer by the law of total probability. Then we get the expected value of each customer by multiplying the total acceptance rate and the customer lifetime value. Our final goal is to try different assignment methods then find the best method to reach the highest mean value.

V. Conclusions and Further Work

Our purpose is to detect whether the assignment policy the company apply now can effectively retain customers and achieve a high customer lifetime value. We ran four iterations where iteration1 was derived from the first layer analysis – following the logic with which the company assigned offers. The other three iterations came from the second layer analysis whose factors affected the acceptance rate of offers. From the table we concluded that the company was

using a less effective policy to assign offers. Among the four iterations, the one with the factors that influence the acceptance rate of offer1 has the best performance.

Table V-i Model Iteration to Find the Best Offer Assigning Model

Factors		**Total Acceptance Rate**	**Mean Customer Lifetime Value**
Iteration 1	Income Sales Channel Customer Lifetime Value Marital Status Month Since Policy Inception Number of Open Complaints Gender	0.15133861120153308	1246.1344347985264
Iteration 2	State Employment Status Location Code Marital Status Sales Channel Vehicle Size Education Vehicle Class Total Claim Amount	0.15358125941635387	1273.862539033336
Iteration 3	State Sales Channel Marital Status Location Code Income Vehicle Size Education Policy Type	0.15219227310163025	1264.5132920624455
Iteration 4	State Education Gender Vehicle Size Months Since Last Claim	0.14698256333996293	1219.307550980015

Based on what we have discussed above, we can conclude key ideas as the following:

1) Both offer assigning and accepting processes are complicated, involving numerous impacting factors from divergent aspects.
2) The car insurance company's decision process is not fixed on the characteristics of applicants. On the contrary, the decision making is based on probabilities, like what people behave in real life. Thus, multiclass logistic regression can function well since it can not only provide the probabilities of offer assignments, but also can offer the flexibilities of covariates' weights.
3) The applicant's accepting process is also not fixed on the features. The model to illustrate

their decision strategies is logistic regression. The reason is the same as above.

4) The recommendations for the car insurance company
    a. By introducing features in layer two, which indicate what factors customers are concerned with, we have improved the acceptance rates and customer values. We recommend that the car insurance company should not split the offer assigning process with applicants' accepting process. Only by combining two parts can the company enhance their profit since what bring revenue are the applicants' acceptances and their lifetime values.
    b. The car insurance company should take into consideration other important features, especially the driving profile. Based on our research, driving profile is a significant factor indicating the likelihood that the applicants would be involved in accidents. However, it is not the database of the company.
    c. The company should adjust their offer4 since nobody has accepted this kind of offer. This existence of offer4 brings costs without revenues. The refinement of it is urgent.
    d. The company should incorporate company-related factors into its database since these factors can offer a degree of freedom for the company to adjust its strategies based on applicants' acceptance modeling.

We will briefly discuss the future work as the following:

1) If company-related features are offered, we can establish stochastic programming models to optimize the company's decisions. It will be a prescriptive model. We can solve the model by scenario-based algorithms.
2) A/ B testing techniques can be utilized to test whether our recommendation can improve the profit or not.
3) We will conduct more research on the demographic and economic influencing factors of the car insurance industry to better understand the decision process.

Dataset:
https://drive.google.com/open?id=1dN93GlOgXUfTfaDkeewbol6USNExovnJ

Video:
https://drive.google.com/file/d/1jYdtFC-X14yauY0hpBCIy2AhRZHGkU1f/view?usp=sharing