Students : Hoai Nam Tran and Nhat My Thien Nguyen

**DATA MINING PROJECT**

**Introduction**

The report is about evaluate and predict customer’s lifetime value and their reaction to renewal insurance offer from company when the current contracts expire based on their demographics and buying behaviors.

This dataset was collected from IBM Watson analytics that published on marketing website. The dataset has 9134 customer records and 24 variables which contains customers personal information, auto information, aggregate information about their behavior since policy inception, and the outcome variables Customer Lifetime Value and Response Yes/No based on their acceptance or not. All customers were engaging auto insurance with company and their policies were going to expire between Jan 1 to Feb 28, 2011.

Using Watson Analytics data, we can analyze the most profitable customers and develop targeted actions for retention programs. Based on all that information, the objective of this project is to figure which factors affect customers’ value and their decisions on accepting renewal offer, from that determine relevant valuable customers and build appropriate marketing campaign.

This report includes data exploration analysis, data cleaning, building and selecting model, model predictive performance, and recommendations from results.

**I. Business Questions:**

1) What is Customer Lifetime Value based on their background and insurance history?

2) Whether a customer will accept insurance renewal offer from company?

**II. Knowledge domain/ Reference Articles:**

1. "What is Customer Lifetime Value?" (<https://www.actioniq.com/blog/what-is-customer-lifetime-value/?fbclid=IwAR2OgTtvWfK0oROyr-HW5yC5looUPCLoUFPleIizSto_Sv4ac7mLaBBVBJw>
2. "Why Auto Insurance Advertisers Should Optimize for Customer Lifetime Value?" (<https://mediaalpha.com/article/why-auto-insurance-advertisers-should-optimize-for-customer-lifetime-value/?fbclid=IwAR0mW66Ss-XXcivDmruMI6f0JQXzbNjMwQQ7UHupwgJampLbpPIZz1R3zJE>)

The lifetime value of a customer, or customer lifetime value (CLV), represents the total amount of money a customer is expected to spend in business, or on products, during their relationship with the company. Knowing each customer lifetime value is important to help company make decisions on how much money to spend on acquiring new customers and retaining existing ones.

**III. Dataset Resource**

Marketing Customer Value Data

(<https://www.kaggle.com/pankajjsh06/ibm-watson-marketing-customer-value-data>)

**IV. Exploratory Data Analysis**

1. **Data Cleaning**

***Dimension:***

The dataset has 24 variables and 9134 records.

*5 Numberical variables*: Customer.Lifetime.Value, Income, Monthly.Premium.Auto, Months.Since.Last.Claim, Total.Claim.Amount,

*19 Categorical variables*: Customer, State, Response, Coverage, Education, Effective.To.Date, EmploymentStatus, Gender, Location.Code, Marital.Status, Months.Since.Policy.Inception, Number.of.Open.Complaints, Number.of.Policies, Policy.Type, Policy, Renew.Offer.Type, Sales.Channel, Vehicle.Class, Vehicle.Size.

***Missing value:***

Fortunately, our dataset does not have missing values.

***Outliers:***

For numerical variables, the distributions are skewed, there are about 5% of extreme values of “Customer.Lifetime.Value”, “Monthly.Premium.Auto”, “Total.Claim.Amount”. We need to consider these values carefully later because these might be important records.

***Variables Selection:***

* First of all, we removed *“Customer”* ID and *“Effective.To.Date”* because two variables are irrelevant with our prediction and classification models.
* Next, we deleted *“Policy”* variable because we have “*Policy.Type*” variable that is similar to “*Policy*” variable.

1. **Visualization**

* **Statistic**

A picture containing table

Description automatically generated**Numerical Variables: (Max/Min/Mean/SD/Quartile)**

**Categorical Variables (Mode)**

Table

Description automatically generated

* Text

  Description automatically generated**Relationship between variables with CLV and Response**

Text

Description automatically generated

**2) Dimension deduction:**

* After doing visualization, we remove some variables and group values in some specific variables into smaller groups, and then we create new variables for these grouped values:
  + - * **State:** we exclude this variable because it does not affect to CLV and Response.
* **Number.of.Open.Complaints**: we grouped values of this variable into 2 categories, the first one is customers who make less than 3 complaints, the second one is customers who make more than 3 complaints. Because, the effect of values on CLV in each groups are quite similar. We create new variable called “Num.of.complaints” to save new transformed values.
* **Number.of.Policies**: we grouped values into 3 groups. Group 1, customers who buy 1 policy, group 2, customers who buy 2 policies, and group 3, customers who buy more than 2 policies. We create a new variable called “New.Num.Policy” to assign new categorical values.
* **Vehicle.Class** : we grouped values into 3 groups “Luxury Vehicle”, “Med.Class Vehicle” and “Normal Vehicle”. We create a new variable called “New.Car.Class” for saving new grouped values.
* **Education:** we grouped values into 3 groups: "Higher Master", "Higher Associate", and “High School or Below”.

***Finally, we transform all categorical variables into dummy.***

**V. Define data mining task**

Base on the business questions that we want to solve, we choose appropriate data mining tasks :

* First question we want to know how much of customers lifetime value (CLV) based on the customers’ backgrounds and their car insurance history, the task is ***Prediction*** a number. We will predict CLV by using independent predictors.
* Second question we want to know who is likely to accept the renewal of offers, the task is ***Classification*** to determine the customers who will say Yes with renew offer.

**Technique Selection**

1. ***Prediction task***: we choose Multiple Linear Regression to predict Customer Lifetime Value.
2. ***Classification task***: we used Logistic Regression to classify Response (Yes/No) to Renewal Offer and compare with Knn model.

**VI. APPLY TECHNIQUE & EVALUATE PREDICTIVE PERFORMANCE**

1. **Partition the data**

* Prediction task: we just simply divide randomly the cleaned dataset into 60% of train data and 40% of valid data.
* Classification task: because the dataset is unbalanced between the YES response and NO response (we have about 14% of YES), we first partition the data, 60% of train data and 40% of valid data, then we apply oversampling method to train data.

1. **Apply the Technique :**
2. ***Running Multiple Linear Regression***

At first we run model with all variables, then we use 3 techniques to select variables and compare these 3 model

* Model 1: We chose significant variables at 5% of significant level from the model of all variables.
* Model 2: We used Bayesian Model Average (BMA) technique that use BIC criteria to adjust models with different variables.
* Model 3: We used STEPWISE technique to choose variables
* Graphical user interface, table

  Description automatically generated with medium confidenceMODEL SELECTION

From the comparing metric r2, RMSE, ME, MAE, MPE, MAPE, we can see that the metrics of evaluating between models are not different too much, which means the ability of predicting of these models are quite similar. However, MBA method suggests a little better model, because the ME and MAE of MBA method is lowest, and BMA suggests a model with just 6 variables, less variables than other models, parsimonious model.

* Text

  Description automatically generatedWe use VIF to test for Multicollinearity of the chosen model:

Because all VIF of variables are all below 10, there is no multicollinearity.

* INTERPRETING CHOSEN MODEL

The Equation: CLV = -2713.54 + (69 x Monthly Payment) + (668 x EmploymentStatus Employed) + (-802 x More than 3 Complaints) + (11867 x Buy 2 Policies) + (3579 x Buy more than 2 Policies) + (-723 x Normal Vehicle)

* **Monthly.Premium.Auto: 68.6**, which means that when other variables remained constant, the monthly payment of customer increases $1, that would cause the increase of 69 dollar of Customer Lifetime Value.
* **EmploymentStatus\_Employed: 667.9**, which means that when other variables remained constant, The customer with employed status will bring higher $668 revenue to company than disable customer.
* **Num.of.Complaints\_More.than.3: -801.9**, which means that when other variables remained constant, the value of customers who have more than 3 complaints is lower about $802 than that of customers have less than 3 complaints.
* **New.Num.of.Policies\_2: 11867.1**, which means that when other variables remained constant,the most valued customers are people who buy 2 policies, the revenue from these customer higher $11867 than customer who just buy 1 policy.
* **New.Num.of.Policies\_More.than.2: 3578.6**, which means that when other variables remained constant, customer who buy more than 2 policies have revenue higher than that of customer buy 1 policy, about $3579.
* **New.Car.Class\_Normal.Vehicle: -723.1**, which means that when other variables remained constant, compare to luxury vehicle class, revenue from customer who buy insurance for normal vehicles is less than about $723.

1. ***Running Logistic Regression***

Checking proportion of Yes and No in data : Yes Response just accounts for 14.3%, we need to oversampling the train dataset.

First, we run logistic model without oversampling, we got the results accuracy is 87.19%, but the sensitivity is only 14.1%, which is pretty low. As an analyst working at insurance company, we want to improve this rate to better predict who will accept renew offer.

We will do oversampling for better prediction of Yes. We used function “ovun.sample” in package ROSE to create new training dataset with balance Response. After run model with oversampling, we got the accuracy was down to 78.13%, but we have higher sensitivity 52.67%, which means this model is doing better of prediction 1 than without oversampling model.

* MODEL SELECTION
* Model 1: We run model with only significant variables (at 5% of significant level) from the model of all variables and remove insignificant variables.
* Model 2: We used BMA with BIC criteria to select variables.
* Model 2: We used STEPWISE model

Table

Description automatically generated**Comparision and Choose Best Model**

Overall, the accuracy and sensitivity of stepwise model is performance the best among three models. Thus, we choose the STEPWISE model as the best preditive model for deploying.

* INTERPRETING THE CHOSEN MODEL:

The Logistic Equation:

Log ( Odds(Response = YES)) = 0.467 + (0.0000096 x Income) + (0.0035 x Monthly.Premium.Auto) + (-0.0012 x Months.Since.Policy.Inception) +(-0.0013 x Claim Amount) +(-0.31 x EmploymentStatus\_Employed) + (0.38 x Medical.Leave Status) + (2.4 x Retired Status) + (-0.51 x Unemployed status) + (1.42 x suburban location) + (-0.6 x Married Status) + (-0.66 x Single Status) + (-0.18 x Personal Policy) +(0.24 x Special Policy) + (-0.64 x Brand Sale Chanel) + (-0.4 x Call Sale Chanel) + (-0.52 x Web Sale Chanel) + (-0.32 x Mid-size Vehicles) + (-0.73 x Small-size vehicle) + (-0.1 x Num.of.Policies\_2) + (-0.3 x Num.of.Policies\_More.than.2) + (-0.23 x New.Car.Class\_Mid.Class.Vehicle) +(-0.23 x High school or below) + (0.38 x Higher Master)

Interpreting the Coefficient:

* **Income: 0.0000096**: Remain others, if the income of customer increase by $1, the odds of yes response increase by 1 (exp(0.0000096)), not much significant.
* **Monthly.Premium.Auto: 0.0035** : Remain others, if the monthly payment of customer increase by $1, the odds of yes response increase by 1 (exp(0.0035)).
* **Months.Since.Policy.Inception: -0.0012** : Remain others, if the month policy inception of customer increase by 1 month, the odds of yes response decrease by 1 (exp(-0.0012)).
* **Total.Claim.Amount: -0.0013** : Remain others, If the total claim amount increase by $1, the the odds of yes response decrease by 1 (exp(-0.0013))
* **EmploymentStatus\_Employed: -0.31** : Remain others, If customers have employment status is employed, they have the odds of YES response 0.73(exp(0.31)) lower than that of disable customer.
* **EmploymentStatus\_Medical.Leave: 0.38** : Remain others, If customers have employment status is medical leave, they have the odds of YES response 1.5(exp(0.38)) higher than that of disable customer.
* **EmploymentStatus\_Retired: 2.4** : Remain others, If customers have employment status is retired, they have the odds of YES response 11(exp(2.4)) higher than that of disable customer.
* **EmploymentStatus\_Unemployed: -0.51** : Remain others, If customers have employment status is unemployment, they have the odds of YES response 0.6(exp(-0.51)) lower than that of disable customer.
* **Location.Code\_Suburban: 1.42** : Remain others, If customers living in suburban areas, they have odds of YES response 4.14(exp(1.42)) higher than that of customers living in rural areas.
* **Marital.Status\_Married: -0.6** : Remain others, If customers have married status, they have the odds of YES response 0.55(exp(-0.6)) lower than that of customers having divorced status.
* **Marital.Status\_Single: -0.66** : Remain others, If customers have single status, they have the odds of YES response 0.51(exp(-0.66)) lower than that of customers having divorced status.
* **Policy.Type\_Personal.Auto: -0.18** : Remain others, If customers have personal contract previously, they have the odds of YES response 0.83(exp(-0.18)) lower than that of customers having comporate contract.
* **Policy.Type\_Special.Auto: 0.24** : Remain others, If customers have special contract previously, they have the odds of YES response 1.27(exp(0.24)) higher than that of customers having comporate contract.
* **Sales.Channel\_Branch: -0.64** : Remain others, If customers buy insurance through brand sale channel, they have the odds of YES response 0.58(exp(-0.64)) lower than that of customer buying through agent.
* **Sales.Channel\_Call.Center: -0.4** : Remain others, If customers buy insurance through call center sale channel, they have the odds of YES response 0.67(exp(-0.4)) lower than that of customer buying through agent.
* **Sales.Channel\_Web: -0.52** : Remain others, If customers buy insurance through web sale channel, they have the odds of YES response 0.6(exp(-0.52)) lower than that of customer buying through agent.
* **Vehicle.Size\_Medsize: -0.32** : Remain others, If customers’ car size are medium, they have the odds of YES response 0.73(exp(-0.32)) lower than that of customers own large cars.
* **Vehicle.Size\_Small: -0.73** : Remain others, If customers’ car size are small, they have the odds of YES response 0.5(exp(-0.73)) lower than that of customers own large cars.
* **New.Num.of.Policies\_2: -0.11** : Remain others, If customers buy 2 policies in last their insurance, their odds of renewing contracts are 0.9(exp(-0.11)) lower than that of customer buying 1 policies in last insurance.
* **New.Num.of.Policies\_More.than.2: -0.34** : Remain others, If customers buy more than 2 policies in last their insurance, their odds of renewing contracts are 0.74(exp(-0.34)) lower than that of customer buying 1 policies in last insurance.
* **New.Car.Class\_Mid.Class.Vehicle: 0.25** : Remain others, If customers have mid-class vehicles, their odds of renewing contracts are 1.3(exp(0.25)) higher than that of customers have luxury vehicles
* **New.Education\_High.School.or.Below: -0.23** : Remain others, If customers’ education are high school or below, their odds of renewing contracts are 0.79(exp(-0.23)) lower than that of customer have college education or above.
* **New.Education\_Higher.Master: 0.38** : Remain others, If customers’ education are master or above, their odds of renewing contracts are 1.46(exp(0.21)) lower than that of customer have college degree or bachelor degree.
* **COMPARING WITH KNN**

Table

Description automatically generatedWe run Knn technique to compare the predictive performance and got below results with different k:

We chose k=3 which is an odd number to ensure there is no tie and it balances between overfitting and ignoring the predictor information.

Using knn to classification, we have the accuracy rate = 92.61%, sensitivity rate = 95.84%. KNN performs super great in this case of classification.

**VII. CONCLUSION & BUSINESS RECOMMENDATION**

* 1. **How could company increase the Customer Lifetime Value by focusing on certain customers?**

Based on the CLV model, the insurance company should try to reach new feasible customers who are employed and try to sell contracts with at least 2 policies. Because the groups of customers who buy at least 2 policies are the most valued customers.

The company should introduce good promotions to attract customers buy insurance for their luxury vehicles. This strategy seems reasonable, because the value of contracts for luxury cars are high and customer using luxury cars tend to drive carefully that makes the risk of insurance payment decrease.

Agents are good channels to increase CLV, customers seem to trust in the real agents who are nice to help any time the customers need.

Besides that, the company should increase the customer services to reduce the complains, because many complaints from customers is the main reason for decreasing the CLV.

* 1. **Who does the insurance company should focus to increase the renewal contract rate?**

Using the logistic model as a reference, to increase the renewal rate, the company should focus on retired customers, divorced customers. We could introduce some discount and promotion to attract these customers, a better renewal price would make retired customers pleased.

People who live in suburban areas have high renewal rate, so we could concentrate on customers in these areas. Agents are important to convince customers to accept renew offers.

Customers with large size vehicles are easy to accept new offers, so we could introduce better new contracts that specify more in large vehicles, that could help increase the acceptance rate increase.

**APPENDIX**

**DATA VISUALIZATION**

* 1. **Customers’ background**

**Map

Description automatically generated***States where insurances sold:*

*Chart, bar chart

Description automatically generatedAverage Income by Employment Status:*

*Customers’ vehicle class and Location:*

*Chart, bar chart

Description automatically generated*

*Customers’ vehicle Education:*

*Chart, pie chart

Description automatically generated*

* 1. **Customers’ Behavior:**

**Chart

Description automatically generated***Total Claim Amount by Number of Policies:*

**Bar chart

Description automatically generated with medium confidence***Average Monthly Payment by Vehicle Size:*

*Sale channels and Policy Types:*

*Chart, bar chart

Description automatically generated*

*Coverage:*

*Chart, pie chart

Description automatically generated*

* 1. **Association between outcome variable and other variables:**

**Chart, box and whisker chart

Description automatically generated***CLV distribution:*

*Income:*

**Chart, box and whisker chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

*Monthly Premium Auto (Monthly Payment):*

*Chart

Description automatically generated*

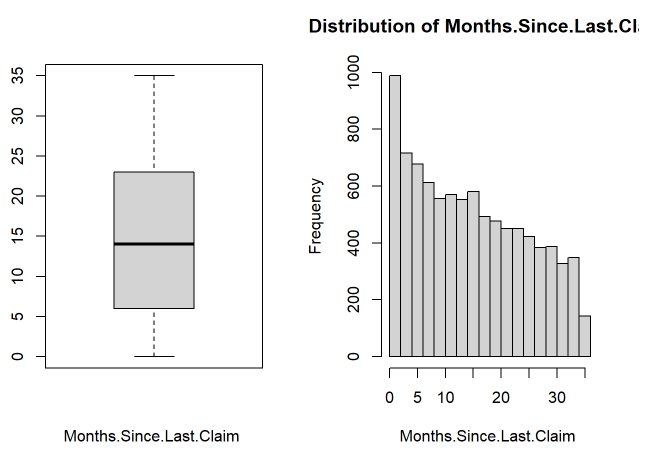
*Chart, scatter chart

Description automatically generated*

*Chart, box and whisker chart

Description automatically generated*

*Months Since Last Claim:*

**

. Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

*Months Since Last Claim:*

*Chart

Description automatically generated*

. Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

*Months Since Policy Inception:*

*Chart, box and whisker chart

Description automatically generated*

.

Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

*Number of Complaints:*

*Chart, histogram

Description automatically generated*

. Chart, box and whisker chart

Description automatically generated

*Number of Policies:*

*Chart, box and whisker chart

Description automatically generated*

Chart

Description automatically generated.

*Total Claim Amount:*

*Chart

Description automatically generated*

. Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

*Response:*

*Chart, bar chart

Description automatically generated*

*State:*

*Chart, box and whisker chart

Description automatically generated*

.

*Coverage:*

*Chart, box and whisker chart

Description automatically generated*

Chart, bar chart

Description automatically generated

*Education:*

*Chart, box and whisker chart

Description automatically generated*

Chart, bar chart

Description automatically generated

*Employment Status:*

*Chart, box and whisker chart

Description automatically generated*

Chart, bar chart

Description automatically generated

*Gender:*

*Chart, box and whisker chart

Description automatically generated*

Chart, bar chart

Description automatically generated

*Location:*

*Chart, box and whisker chart

Description automatically generated*

Chart, bar chart

Description automatically generated

*Marital Status:*

*Chart, box and whisker chart

Description automatically generated*

Chart, bar chart

Description automatically generated

*Policy Types:*

*Chart, box and whisker chart

Description automatically generated*

Chart, bar chart

Description automatically generated

*Renew Offer Types:*

*Chart, box and whisker chart

Description automatically generated*

Chart, bar chart

Description automatically generated

*Vehicle Class:*

*Chart, box and whisker chart

Description automatically generated*

Chart

Description automatically generated

*Sale Channels:*

*Chart, box and whisker chart

Description automatically generated*

Chart, bar chart

Description automatically generated

*Vehicle Size:*

*Chart, box and whisker chart

Description automatically generated*

Chart, bar chart

Description automatically generated

* 1. **Models Coefficients**

*Timeline

Description automatically generated with medium confidenceCLV Linear Regression:*

**Text

Description automatically generated***Response Logistic Regression:*

**APPLY TECHNIQUE & EVALUATE PREDICTIVE PERFORMANCE**

* 1. **Linear Regression**

**Table

Description automatically generated***Model with all variables:*

****

*Table

Description automatically generated*

**A picture containing text

Description automatically generated***Model with BMA technique:*

*Table

Description automatically generated*

*Table

Description automatically generated*

*Graphical user interface, text, application, email

Description automatically generated*

*****Model with STEPWISE technique:*

*Table

Description automatically generated*

*Measuring and choosing model:*

**Graphical user interface

Description automatically generated with low confidence**

**Graphical user interface, text, application, email

Description automatically generated***VIF test for Multicollinearity:*

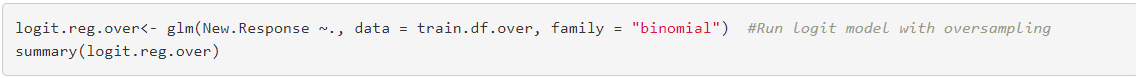
* 1. **Logistic Regression**

*Oversampling:*

*Graphical user interface, text, application, email

Description automatically generated*

*Model with all variables:*

**

**Table

Description automatically generatedTable

Description automatically generated with medium confidence**

**Graphical user interface, text

Description automatically generated**

**A picture containing text

Description automatically generated***Model with BMA technique:*

*Text

Description automatically generated with medium confidence*

*Table

Description automatically generated*

*Table

Description automatically generated*

Text

Description automatically generated with low confidence

Text

Description automatically generatedTable

Description automatically generated

*Graphical user interface, text, application

Description automatically generated*

*****Model with STEPWISE technique:*

*Table

Description automatically generated with low confidence*

*Graphical user interface, text, application

Description automatically generatedText

Description automatically generated with medium confidence*

**Graphical user interface, text, application, email

Description automatically generated***Measuring and choosing model:*

**Chart

Description automatically generated***Decile-Wise Lift Chart:*

*Chart, histogram

Description automatically generated*

**Graphical user interface, text, application

Description automatically generated***k-NN:*