Estimating Customer Lifetime Value in Insurance

Yi-Chen Chiou, Rohan Das, Udayan Kate, Sanjana Santhanakrishnan, Jiayu Zhang, Yang Wang

Purdue University, Krannert School of Management, 403 W. State Street, West Lafayette, IN 47907

ychiou@purdue.edu; das172@purdue.edu; ukate@purdue.edu; santhans@purdue.edu; zhan4358@purdue.edu; yangwang@purdue.edu

Abstract

Customer Lifetime Value is an important metric, using which companies can focus on growing or maintaining their revenue streams from their customer base. This metric can help organizations provide a direction to focus the marketing expenditures and efforts by helping identify customers who have high potential in terms of longevity of association and identify the possibility of selling additional products (cross-selling). The metric can also help identify those customers who probably do not have high potential in these terms, further aiding in optimizing the expenses and effort on marketing the products. This paper explores the possible methods of calculating customer lifetime value in the insurance industry and develops a model which can help predict the customer lifetime value for new customers.

Keywords: Customer Lifetime Value, Insurance, Logistic Regression, Survival Analysis, Cross-Selling, Decision Tree, Segmentation

Importance of Customer Lifetime Value in Insurance

The fundamental source of revenue in any company is its customers. Therefore, maintaining good relationships with customers is critical to any business. However, each customer is different, and thus treating every individual customer relationship, in the same way, would not be prudent. This raises the question of the importance of an individual customer to the business, as based on this understanding, the method of marketing to an individual customer can be customized. One of the ways that this understanding can be sought is with the estimation of "Customer Lifetime Value". Customer Lifetime Value (abbreviated to CLV) is a metric using which the present net value of a customer to a company can be estimated based on the customer's history with the company and the potential worth of a future relationship with the company. CLV can help businesses categorize customers into those with high or low potential in terms of revenue. Such categorizations would help businesses target specific customers who have high value to the company, for marketing and retention while simultaneously helping avoid marketing expenses towards customers or market segments where the potential of growth or revenue is low.

The main utility and purpose of insurance companies throughout history has been to provide financial protection for an individual or their property against extreme and improbable events which would lead to financial damage or loss of life. Traditionally, the key source of revenue for insurance companies has been advance payments made by customers in return for future protection against risks which is known as a premium. Insurance companies provide a wide range of services against risks which can be categorized into several lines of business covering specific commodities such as property, life, automobile, health, etc. Agents, brokers, and digital platforms developed by companies have been the key distributors of insurance policies that take into consideration customer behavior and preferences for optimal pricing. Most insurance policies have a policy limit and deductible associated with them which limits the loss an insurance company would have to bear in case of a claim which is too small or too large. Some lines of business have special considerations in their policies that affect the financial terms like premiums, limits, and deductibles, for example, chronic health conditions of a customer lead to a lower deductible in a health insurance policy.

In the US market alone, including territories, there are more than 5,929 insurance companies (Insurance Information Institute, 2021) operational and providing various types of policies. The US insurance industry also recorded a net premium of $1.28 trillion. As the market reaches saturation, the competition increases. With increasing competition and industry rivalry, building long-term relationships with insured customers is how an insurance company can keep and grow its market share. Assessing an insured's CLV is one of the methods by which insurance firms can identify which customers would potentially stay longer with them than others. This is especially useful considering the low switching costs for customers in this sector and the myriad of options available. It can also help these firms in discovering strategies to acquire more customers, retain customers and increase profitability. By estimating CLV, the insurance company can also spot customers that have a high probability of cross-buying other insurance products. The insurance company can therefore focus their efforts on retaining these high CLV, high net profit customers than the less valuable customers. Similarly, classifying customers based on CLV can be useful in determining characteristics of profitable customers and these factors would be beneficial in determining what segment of customers to acquire.

Considering the importance of CLV in the insurance industry, the goal of this project is to develop a predictive model to estimate the CLV. The study considers the revenue generated from the customer, the relationship length between the customer and the company, etc., and the factors affecting the CLV of a particular customer. Based on the estimation of CLV, it's easier to identify and categorize customers in appropriate buckets and therefore develop acquisition and retention strategies for categorized customers.

Conceptual Framework

This section will discuss the different models used for calculating CLV and the approach that we have followed based on the understandings of these models. The paper considers some approaches which are generally used by researchers to calculate one or more variables used to determine the lifetime value of a customer. The base underlying problem used by all models is the same which is to determine how valuable a customer is to an organization. However, the methodologies used by these models differ in terms of the variables they use to predict CLV. Most models do not consider competition due to the lack of availability of competitive data. The frequency with which CLV is calculated depends on the volatility of the market where if a customer behavior is subject to change drastically over a short period of time, then CLV needs to be calculated more frequently (Gupta et al., 2006). The models discussed either build different modules for calculating the retention, acquisition, and margin of a customer or combine these modules to provide a single CLV of every customer.

I. Probability Models

Probability models try to predict the outcomes of patterns based on stochastic processes. Various factors or behaviors of individual customers relate to the differences in outcomes and these behaviors are modeled with the help of a stochastic process that defines the probability distribution.

In case of understanding CLV, the interest lies in measuring the chances that a customer will continue their association with the insurance company, the probability distribution to estimate the period of time that a customer may want to continue paying premiums for the product, the chances that there is a claim that may be filed by an insured and the chances that the customer may strengthen the association with the insurance company through cross-selling or upgrading/downgrading their existing policies. This could also be extended to measuring the probability of claims due to the geographical location of the property in case of home insurance (to take into account any disaster level events that could happen in the area where the property is located) or the chances of claims that could arise due to driving quality of individuals or the probability of claims in life annuities impacted by individual's personal preferences and lifestyle (for example an individual working in hazardous conditions versus an individual working in relatively safe conditions)(Gupta et al., 2006).

II. Computer-Based Models

Computer-based methods use sophisticated models including data mining and machine learning methodologies. These methods are hard to interpret but utilize practical and advanced methodology. Most of the research that we found is mainly focused on applying computer-based approaches to classification problems such as identifying customer targets for cross-selling or churn activities. Decision trees, neural networks, support vector machines (SVM), and classification and regression trees have been applied. SVM model and decision trees are the ones that were proved to perform well in classification problems.

The logistic regression model can be deployed for the purpose of classification based on the probability estimates. This model is very useful in classifying targets with binary values. An advantage of this model is the ease with which one can extract the probability of classification, i.e., the probability which is associated with the classified output.

The SVM model is used for the purpose of classification. The benefit of this model is that it transforms the original data into a space that contains different features (Friedman, 2003; Kecman, 2001; Vapnik, 1999). This method handles hyperplanes cases. The performance between the multinomial logit model and SVM was compared by Cui and Curry (2005) whose result indicated that SVM outperforms logit in all situations.

Another ideal option to be considered for the classification problems is the decision tree. An experiment conducted by Giuffrida, Chu, and Hanssens (2000) showed that a multivariate decision tree has better performance in prediction for cross-selling when compared to a logit model. Further, even though LSTM is not widely applied in the insurance field, Marta Jablecka (2020) attempted to apply Long Short Term Memory (LSTM) to find customers who are more likely to change their policy. He then used Recurrent Neural Network (RNN) to predict CLV with the data which included the detected customers that had been classified as an anomaly by using the LSTM model. The model produced a high precision performance with a score of 0.93.

III. Scoring Models

According to Sunil Gupta et al. (2006), RFM models have been used in industries with a lower response rate for the development of targeted marketing strategies by taking into consideration previous purchase behavior of a customer instead of their demographic information. RFM model is a type of scoring model where customers are classified based on a score computed by analyzing the recency, frequency, and monetary (RFM) value of their previous purchases. Weights are assigned to these three variables to compute a score for each customer or a group of customers.

RFM models can provide prediction only for the next period which is a key limitation of the use of RFM models. RFM models do not consider the fact, that the marketing strategies of a company can change with time, which in turn would affect the behavior of the customers. However, despite these limitations, RFM models can provide essential past purchase variables which are required to calculate CLV. According to Fader, Hardie, and Lee (2005), RFM variables can be used to build a CLV model which can overcome many of its limitations.
Ayoubi (2016) described the use of a WRFM model which is a weighted RFM model, useful to determine CLV and classify customers into clusters in the banking sector. The weights in this study

were obtained from subject-matter-experts and used along with the recency, frequency, and monetary value parameters. Cluster assignment and optimization was a two-phased approach using Kohonen neural network. Based on this, the most valuable customers for the bank were identified so that they can be targeted for marketing strategies and CRM programs.

Hiziroglu and Sengul (2012) performed a comparative analysis of two representative models - one which uses the past customer data and another which considers the future-past behavior of customers. For the modeling based on past data, a version of the RFM model which uses the recency, frequency, and monetary value terms was used. For modeling future-past behavior, a basic structural model of the below format was used -

$$\text{CLV} = \sum_{i=1}^{n} \frac{Ri - Ci}{(1+d)^{i-0.5}}$$

These two models were evaluated on the same dataset and customer segmentation was done based on CLV. The results showed the superior performance of the basic structural model compared to the RFM model.

IV. CLV Modeling Approach

One of the most widely used models is a combination of a probabilistic approach and a structural model. This model mentioned in Seyerle (2003) specifically for insurance calculates the CLV based on the present value of the customer and a probabilistically weighted future value calculation. The equation used is as shown below.

$$\text{CLV} = Revenues - Cancellation\ Value + Cross\ Selling\ Value - Claim\ Value - Activity\ Costs$$

Revenues are earned using the premiums that are collected by the insurance company from the insured. For each active policy, however, there is always a risk of cancellation, and this cancellation will lead to a loss for the company in terms of the premium earned for that particular year. The cancellation value is the representation of this loss. Also, for every existing customer, there is a potential that additional products of the same organization may also be sold to such a customer. This could generate further earnings in terms of premiums and this earning is classified as Cross-Selling. For every active policy, there is a risk that a claim is filed. This claim is a payment that the insurance company would have to make towards the insured and as such costs the company. The amount of money at risk associated with each policy's event of a claim is categorized as Claim Value. Activity Costs are costs related to the marketing of the insurance products. This study will ignore the effects of marketing cost on the CLV due to constraints in the availability of data.

The cancellation value, cross-selling value, and the claim value are all calculated based on the value generated weighted by their corresponding probabilities of occurrence in the future.

This model is the basis for our approach towards CLV calculation. We have chosen this because of its wide acceptance and good performance compared to complex models discussed previously.

Data

The data used in this project were provided by our client, a regional insurance company in the United States. The original data set has various tables that contain information about Products, Claims, Valuation, Demographics, and Relationship Duration. After a deep dive into the data and with guidance from the client, we got a smaller data set with variables that are related to the research topic:

Table 1: Potential model input variables of the Product table

Variable	Type	Description
Policy Type	Categorical	The type of a policy
Policy Inception Date	Date	Starting period of a policy
Household Inception Date	Date	The date this household started a business relationship with the company.
Inactive Date	Date	The date when the policy is canceled or expired
Active Now	Categorical	Whether the policy is active or not
Channel	Categorical	The channel type which the individual gets the policy
Subcategory	Categorical	Subcategory of policy
Score Level	Categorical	Household's score level
Reason of Inactive	Categorical	The reason why the policy is inactive

Table 2: Potential model input variables of Claims table

Variable	Type	Description
Loss Date	Date	Date an accident or loss
Report Date	Date	Date of a claim report
Claim Status	Categorical	The process condition of a claim
Total Loss Indicator	Categorical	Whether the insured object is completely destroyed
Incurred Amount	Numerical	The cost of incurred amount

Table 3: Potential model input variables of the Valuation table

Variable	Type	Description
Annualized Premium	Numerical	The total amount of premium paid annually based on the customer's term
Discount	Categorical	The Type of the discount
Paperless Discount	Categorical	Whether the customer chooses to use the electronic version of a policy
Full Pay Discount	Categorical	Whether the customer chooses to pay the premium in full
Drive Trend Discount	Categorical	Whether the customer chooses to use the driving app
Multiline Discount	Categorical	Whether the customer has more than one policy
Multi-Car Discount	Categorical	Whether the customer has more than one auto policy
Top Scholar Discount	Categorical	Whether the customer is a student with a good GPA
Auto Life Discount	Categorical	Whether the customer has both auto and life policy
Impact Resistant Roof Discount	Categorical	Whether the home has a roof structure

Table 4: Potential model input variables of Demographics table

Variable	Type	Description
Birthday	Numerical	The birthday date of the individual
Gender	Categorical	The gender of the individual
Marital status	Categorical	The marital status of the individual
Credit Card Count	Numerical	Number of credit cards held by the individual
Types of credit card	Categorical	Types of credit card held by the individual
Head of Family	Categorical	Estimate if the person is the head of the family

Family Adult Count	Numerical	Number of adults in the family
Family Member Count	Numerical	Number of members in the family
Family Estimated Incomes	Numerical	Estimated income of the family
Estimated Home Equity	Numerical	Home equity estimate range of the individual
Home Value Range	Categorical	Home value range of the individual
Estimated Education Level	Categorical	Estimated education level of the family
Estimated Vacation Expenses	Numerical	Estimated range of vacation expenses of the family
Lifestyle	Categorical	The lifestyle of the individual
Years in Home	Numerical	Number of years spent in the current home
Address	Categorical	The address of the individual

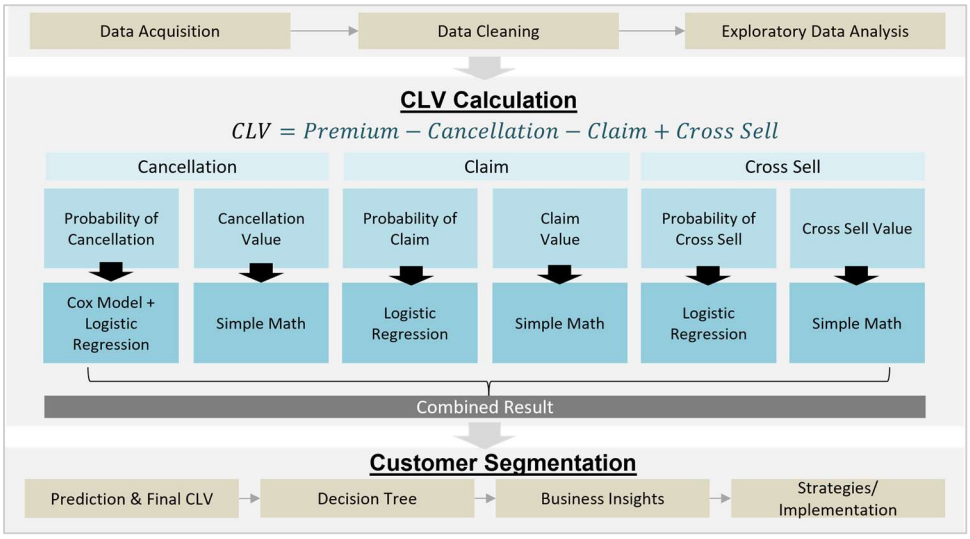
Table 5: Potential model input variables of Relationship Duration table

Variable	Type	Description
Month Active	Numerical	How many months has at least one policy been active

Methodology

To calculate CLV values, there are mainly three components that need to be measured: The cancellation value, cross-selling value, and the claim value. We apply predictive or time series models to estimate these important factors in the CLV calculation formula. Before we estimate the variables, we prepare a master table that has been combined from 5 raw data tables and cleaned by imputing missing values, removing irrelevant columns, and converting data types. Once a master table is prepared, we do an exploratory data analysis and further measure each component by following the following procedure: data partition, model exploration, model tuning, prediction, model comparison, and evaluation.

After all the components are collected, the CLV could be calculated. We then classify the customers into three brackets and apply a decision tree to extract the important characteristics of each group of customers. With those features and insights, we propose a business strategy to our partner company. The overall methodology can be summarized by the following chart.



I. Cancellation Value

Cancellation value is one of the components that needs to be considered for CLV. In our analysis, we multiplied policy premium by the probability of cancellation to calculate the cancellation value at a policy level. For each policy, we already know the premium. The part we need to build a model for is the probability of cancellation. Therefore, we used a Cox proportional hazard model to select important features and then build a logistic regression model to predict the probability.

A. Critical features selected by a Cox proportional hazard model

The reason we want to select important features is that there are hundreds of variables in the data table. To improve our prediction accuracy, we need to identify which variables are significant to the cancellation. So, we chose Cox proportional hazard model to obtain important features. The Cox Proportional Hazard Model is used for survival analysis, which investigates the length of time it takes for an event to occur. Here the event is a cancellation, and the length of time is policy duration day, which is calculated by policy inactive date minus policy inception date. In this model, the target is whether the policy is canceled, which is a binary variable. And the inputs include not only the length of time but also other independent variables that do not depend on time. After building the model, we selected features based on p-value. From the results, we could see some of the variables are significant to cancellation and some are not.

Next, based on the significant predictors identified, a logistic regression model was trained to whether the policy would be canceled or not. The result provided us with the probabilities of whether a cancellation will occur or not.

B. Expected cost of cancellation of a policy

Logistic regression is useful as it provides the probability associated with the cancellation of a policy for a target year based on features known since the inception of the policy. The next

step is to estimate what could be the loss that could be incurred if a cancellation does happen. We have used the annual premium paid for the policy to estimate this figure by multiplying it with the probability of cancellation of a policy. This gives us an estimate of the cost that the insurance company would have to bear due to the possibility of cancellation of the insurance policy. Table 6 below shows an example of how the cancellation value is calculated.

Expected Cost of Cancellation($) = P(An active policy is cancelled in the current year) * (Annual premium($))

Table 6: Cancellation Value Calculation

Customer	Type of Policy	Annualized Premium	Cancellation Probability	Expected Cost of Cancellation
A	Auto	$ 500	0.7	$500 * 0.7 = $350
B	Home	$ 800	0.5	$800 * 0.5 = $400
C	Life	$ 600	0.1	$600 * 0.1 = $60

II. Cross-selling Value

Cross-selling value is the additional revenue generated from selling products of another line of business to an existing customer. In our analysis, we defined successful cross-selling as an existing customer who purchases products of a new line of business has at least 3 months of association with the company. In this study, we only consider the cross-selling between auto policy and home policy since the business of life insurance differs greatly from the rest of the two types of policies. Additionally, we focus on cross-selling among the line of business and assume there is no cross-selling within the line of business. To obtain the expected cross-selling value, we need to determine two variables: probability and value.

A. Estimation of probability of a customer being a target cross-selling customer

Probability is obtained by running a logistic regression model. To compute the probabilities, we create several columns:

- **has_HP**: a binary indicator that identifies if a given customer purchased home policies
- **has_APV:** a binary indicator that identifies if a given customer purchased auto policies
- **policy_duration:** the time difference between the first and the last policy of different lines of business in months.
- **Target_CrossSell:** a binary indicator that identifies if a given customer is a target cross-selling customer. The value is identified based on the combined result from the previous four variables.

Afterward, we use Target_CrossSell as our target variable and include a group of selected features to train a logistic regression model. The selected features are related to customers' demographic characteristics and purchase patterns. The purpose of the model is to learn

characteristics from existing customers who purchase new products of another line of business at least three months after they purchase either a home or auto policy. Accordingly, the records of individuals who both purchased home insurances and auto insurances before would be the training data, and the records of individuals who purchase only one type of product are not included in the training data. Once the model learns the customers' characteristics from successful cross-selling, the records of individuals who only own either auto policies or home policies are tested and the probabilities of the individual who might buy a new product are calculated. Table 7 shows an example of cross-selling target customers identification.

Table 7: Cross-Selling Probability Calculation

Customer_ID	has_HP	has_APV	policy_duration	Target_Cross Sell	Training or Testing data
1	1	1	2	0	Training
2	1	1	5	1	Training
3	1	0	NA	NA	Testing
4	0	1	NA	NA	Testing

B. Estimation of a value that could be generated by a target cross-selling customer

The value that could be generated from a cross-sell is estimated by adopting the average annualized premium from the policy types. The average premium simply represents the value that could be earned from cross-selling for a given line of business.

C. Expected Cross-selling value

After getting the probability and the value of a given customer, the expected cross-selling value of an individual is calculated by multiplying the average premium of a policy that the person doesn't own with the cross-selling probability of the individual. Table 8 demonstrates an example of the calculation of expected cross-selling value.

Table 8: Cross-Selling Value Calculation

Customer_ID	has_HP	has_APV	CSProbability	CSValue	Expected Cross Selling Value
1	1	1	0	0	0
2	1	0	P(CS)	Avg_AP	Avg_APV*P(CS
3	0	1	P(CS)	Avg_HP	Avg_HP*P(CS)
4	0	0	0	0	0

III. Claim Value

The objective here is to calculate how much loss can be expected from an insurance policy in any given year. This forms an essential piece of the puzzle in calculating the earnings from an insurance policyholder. To estimate the expected claim from a policy in a year, we require two parts. One part is to estimate the probability that a claim will be filed in a year. The second part of this is an estimate of the actual claim that could arise.

A. Estimation of probability of a claim being filed

We have taken features grouped under policy identifiers and individual identifiers. The features themselves include variables representing the duration of an existing policy, whether the policy is currently in effect or if it has been canceled, premiums aggregated annually, the type of policies and the various subcategories under each policy type, the platform through which the policy was purchased, the amount of claim that was approved if this existed for a policy, the insurance score level and if any discounts were applied while purchasing the policy. We have also taken into account variables representing an individual's family, gender, wealth, etc. which presents us with an idea about an individual's lifestyle, economic and social standing.

After gathering the data, we had to design a target variable. This was performed based on the amount that was incurred to the company after a claim was filed and approved. The target variable we designed is a binary variable consisting of 1 if there was an amount that was a claim amount incurred on the policy and 0 if there was no claim in the policy so far.

Next, the data was split randomly into training and test sets, by reserving a third of the data for testing. Since the intention is to estimate the probability of a claim being filed, a logistic regression model was decided on for implementation. In the modeling phase, we developed a base logistic regression model, checked its accuracy, and tried to improve upon it by tuning some of the hyperparameters. We found that the l2 penalty improved the accuracy of classification. This way, logistic regression was able to provide us with the probabilities of whether a claim will occur or not.

B. Estimation of a claim amount that could be incurred to the company

To do this, we first created another index. This was done by categorizing each record into a certain type of policy and the subcategory under it. Thus, if we had a home policy represented by "HP", and one of the home policies had a subcategory suggesting that the policy was for a property owner represented by "landlord", we created an additional index for the data represented by "HP_landlord". This helped categorize the entire dataset into different sets based on the policy type and subcategory.

After this, we created a target variable, using which we were able to group each record under a category of premium. To do this, the target variable was created to identify the premiums to the hundredth place rounded below. For example, annual premiums such as $1,056, $548 or $24,768 were grouped under a value of $1,000/-, $500 and $24,700 respectively. The assumption

was that the policies grouped in such a way had customers with similar features and would result in a similar claim value.

Then, we took the claim amounts if any of each policy and multiplied it with the probability of claim (model output). This gives us the expected claim from that policy. Now, within each group, we sum up the expected claims from the individual policies and divide it by the number of policies within that group to arrive at a an expected claim value from any policy in this group.

*Expected Claim Amount = P(Claim is filed)*Estimate of amount of claim($)*

Table 9: An example of calculation of Expected Claims.

Group	Premium	Grouped Under	Claim [A]	Duration (Year) [B]	Claims Expected [A/B = C]	Probability of Claim [D]	Claims Expected [C*D=E]	Claims expected by group [Group Sum E/ group count]
	$5,230.00	$5,200.00	$0	3	$0	0.05	$0	$36.00
1	$5,245.00	$5,200.00	$600.00	5	$120.00	0.60	$72.00	$36.00
2	$6,859.00	$6,800.00	$4,821.00	10	$482.10	0.40	$192.84	$192.84
	$1,042.00	$1,000.00	$486.00	5	$97.20	0.08	$7.78	$41.53
3	$1,021.00	$1,000.00	$968.00	9	$107.56	0.70	$75.29	$41.53

IV. Calculation of Customer Lifetime Value (CLV)

At this point we have the values of the annual premium, probability of cancellation, probability of claims, and an expected claim amount and probability of successful cross-sell with cross-sell earnings. Putting these values together can help us calculate the CLV value. To do so, we can use the following formula.

CLV = (Annual Premiums) – (Expected Claim Amount) – (Expected loss if policy is canceled) + (Expected earnings from possible)

We can understand this CLV in this way; the revenue generated by a customer in dollars, given the association with a company during the current year or the year in focus. The CLV value can be negative, zero, or positive. A CLV negative value could indicate that this customer has a high probability of filing for a claim or may also indicate a large loss from the cancellation of the policy. A zero CLV could occur when the earnings from the premiums are equal to the expected payoff due to the risk of filing a claim and the risk of cancellation. If the CLV is net positive, it indicates that this customer has generated a net income for the company. This can happen only if the combined risk from claims and cancellations is smaller than the annual premiums paid by the customer and the expected payoff from a successful cross-sell.

Table 10: An example of CLV calculation.

Annual Premium	$1,000.00	----------	
Probability of Claim	0.10%	Expected loss from Claim payout	0.10%*10000 = $10
Expected Claim	$10,000.00		
Probability of Cancellation	0.15%	Expected loss from Cancellation	0.15%*1000 = $1.5
Expected loss from Cancellation	$1,000.00		
Probability of Cross-Selling	20.00%	Expected revenue generated from Cross-Selling	20%*500 = $100
Expected Premium earned from Cross Sold Product	$500.00		
Customer Lifetime Value (CLV)		**1000 - 10 - 1.5 + 100 = $1088.50**	

Customer segmentation using Decision Tree

Now that we have the CLV values, the task ahead was to analyze the features of the customers who have a certain range of CLV. For this, we worked on creating a supervised model which could help us identify these features.

The model we chose is a decision tree where the target variable is based on the calculated CLV. To train the decision tree we passed features of the individuals whose absolute correlation value with the CLV value stood high. From the selected columns, we removed a few variables which could play no role in explaining the model such as favorite genres of music or favorite sports. After handling the missing values, we created a set of dummy variables for the categorical variables and then split the data into training and test sets where we reserved 30% of the data for testing.

The target variable has been designed to have three classes to represent low, medium, and high CLV customers. Customers who had a negative CLV were assigned the low category, the customers with CLV up to $3000 were assigned the medium CLV category and any customer who had a CLV value worth more than $3000 was assigned the high CLV category. The high, medium and low CLV categories were marked using the values of 1, 2, and 3 respectively.

Based on this categorization, we experimented with a decision tree model with varying depths and trained a model after which the accuracy was tested on the test data.

Results

Based on the confusion matrix, we measured the accuracy of models used to calculate CLV. Table 11 shows the performance we were able to achieve for each model.

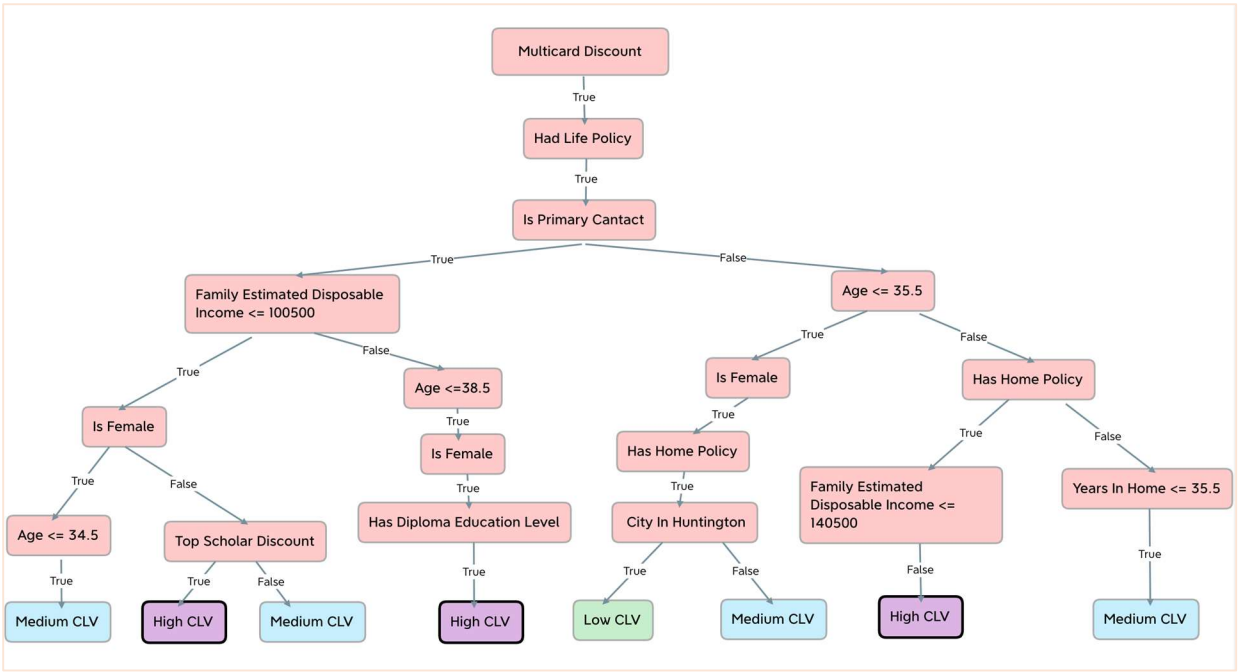
Table 11: Model Performance

S.no.	Prediction of:	Model used	**Accuracy** (On Test Set, based on confusion matrix) (TP+TN)/(TP+TN+FP+FN)
1	Claims Probability	Logistic Regression	99.92%
2	Cancellation Probability	Cox Proportional Hazard Model (Variable Selection) + Logistic regression	88.30%
3	Cross-Selling Probability	Logistic Regression	65.20%*
4	Customer Segmentation	Decision Tree	86.27%

* This accuracy is based on the training data, as we have used customers who have been successfully cross-sold to, to check if existing customers may be a possible target for cross-selling.

The customer segmentation that the study explored using a decision tree resulted in 106 leaf nodes based on best split for segmentation using various features. A sample of the decision tree can be seen in Figure 1 below.

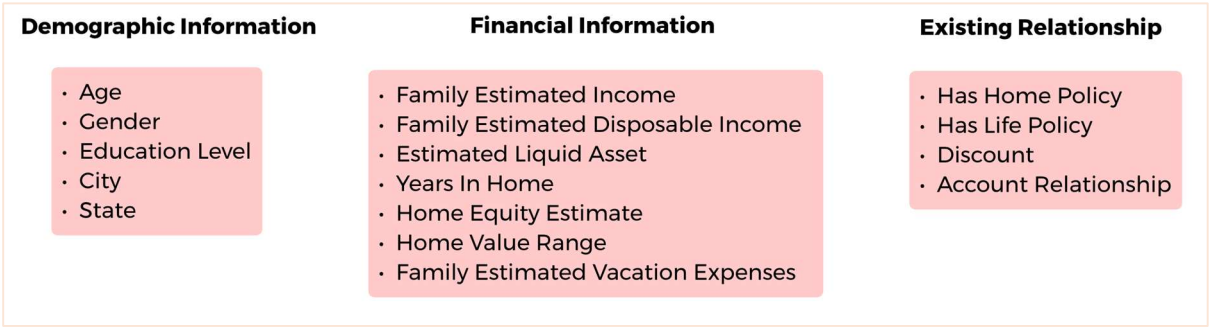
Figure 1: Decision Tree



We believe the decision rule of nodes can be considered as the features of the groups since the decision rules aid in differentiating the groups. The decision tree is trained with a maximum depth equal to seven. Thus, some of the leaf nodes may contain too few samples to be interpreted. In this analysis, we only take the leaf nodes with large samples into account.

The result indicates that age is a critical demographic determinant of CLV level. For financial information, family estimated income, and family estimated disposable income are significant factors. Existing relationship features such as policy type indicators and discount indicators are crucial criteria as well. Figure 2 shows the key features of the customer determining the segments based on the model.

Figure 2: Key features



The decision tree has helped us estimate the overall value and features of a particular customer group. Table 12 shows the characteristics of each customer segment. Most of the customers with high CLV are older than 38.5 years old and have over two types of policies, life policies and home policies. Their family estimated disposable income is higher than 100,500 dollars. The customers in this group mostly have insurance discounts and have more than one car in their families. They also seem to have high GPAs in schools. Customers with medium CLV are aged between 46.5 and 77.5 years old. They do not own more than one auto. For customers with low CLV, most of them are older than 92.5 years old. Their family estimated liquid assets are higher than 750000 dollars. As observed, the model shows that the relationships factors and the age variable are the set of variables that significantly affect the result of the classification.

Table 12: Characteristics of each customer segment with the largest sample size:

Segments	Rules
High CLV	(Multicardisctount>0.5) & (has_LA >0.5) & (accountrelarionship_primary_contact>0.5) & (family_estimateddisposableincome>100500) &(age>38.5) & (has_HP>0.5)
Medium CLV	(Multicardiscount<=0.5) &(77.5>age>46.5) &(has_LA>0.5)
Low CLV	(Multicardisctount<=0.5) &(age>92.5) &(has_LA>0.5) &(family_estimatedliquidassets>750000)

The CLV value of an existing customer can assist in designing better marketing and customer acquisition strategies. The client can design a customized business strategy for each

customer segment and target the customers within each group based on the characteristics that we identified from the model.

Table 13: Business Strategy

Segments	Strategy Type	Business Strategy
High CLV	Engage & Enhance	Design customized loyalty programs and offer rewards such as premium discounts to valuable customers
Medium CLV	Enhance & Engage	Provide up-selling rewards and cross-selling discounts
Low CLV	Service Engagement	Reduce marketing investment Explore the types of products preferred by customers

Conclusion

Customer lifetime value is an important metric for customer retention, especially in insurance. Measuring the CLV helps insurance companies develop strategies for marketing, reducing customer churn, and customer acquisition.

This study has been able to estimate the customer lifetime value of insurance customers and classify them based on their CLV using a probabilistic approach based on machine learning models. The data was made available by an insurance company in the United States. As part of this approach, models have been developed to accurately estimate future cancellation value, cross-selling value, and value of claims. The CLV has been estimated using these calculated values. As part of the next step, we classified the customers based on the CLV to estimate the features that help identify a customer with a high, medium or low CLV. This estimation has helped us to provide business strategies to optimize the campaign of marketing expenditure.

CLV can be a critical metric that can be calculated by organizations in the insurance industry to help identify and understand features of customers with high CLV to increase the duration of the relationship with existing as well as new customers. The models that have been developed through this study can be used by organizations to create budgets for customer acquisition intelligently by prioritizing customers having features similar to the existing customers with a high CLV score. For existing customers with a medium CLV score, marketing strategies can be built to enhance the relationship of the customers by providing up-selling rewards and cross-selling discounts to engage the customer in a greater number of products in an organization.

Additionally, the pricing of the premium of a policy can be optimized by analyzing the features of a customer and calculating their CLV. Customer segmentation based on their CLV values can assist organizations in adjusting premiums based on whether the customer falls in a high, medium, or low CLV category. For instance, customers who have a negative value in CLV may have to be charged a higher premium to offset the chances of losses while on the other hand,

customers with very high CLV may be provided with some discounts to ensure that they stick with the insurance company.

References

Ayoubi, M. (2016). Customer segmentation based on CLV model and neural network. *International Journal of Computer Science Issues (IJCSI)*, *13*(2), 31-37. https://doi.org/10.20943/01201602.3137

Chang, W., Chang, C., & Li, Q. (2012). Customer lifetime value: A review. *Social Behavior and Personality: An International Journal*, *40*(7), 1057-1064. https://doi.org/10.2224/sbp.2012.40.7.1057

Cui, D., & Curry, D. (2005). Prediction in marketing using the support vector machine. *Marketing Science*, *24*(4), 595-615. https://doi.org/10.1287/mksc.1050.0123

Donkers, B., Verhoef, P. C., & de Jong, M. G. (2007). Modeling CLV: A test of competing models in the insurance industry. *Quantitative Marketing and Economics*, *5*(2), 163-190. https://doi.org/10.1007/s11129-006-9016-y

Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, *42*(4), 415-430. https://doi.org/10.1509/jmkr.2005.42.4.415

Friedman, J. H., & Roosen, C. B. (1995). An introduction to multivariate adaptive regression splines. *Statistical Methods in Medical Research*, *4*(3), 197-217. https://doi.org/10.1177/096228029500400303

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., & Sriram, S. (2006). Modeling customer lifetime value. *Journal of Service Research*, *9*(2), 139-155. https://doi.org/10.1177/1094670506293810

Giuffrida, G., Chu, W. W., & Hanssens, D. M. (2000, March). *Mining classification rules from datasets with large number of many-valued attributes* [Paper presentation]. International Conference on Extending Database Technology, Berlin, Heidelberg. https://doi.org/10.1007/3-540-46439-5_23

Hiziroglu, A., & Sengul, S. (2012). Investigating two customer lifetime value models from segmentation perspective. *Procedia-Social and Behavioral Sciences*, *62*, 766-774. https://doi.org/10.1016/j.sbspro.2012.09.129

Insurance Information Institute. (2021). *Facts + Statistics: Industry overview*. https://www.iii.org/fact-statistic/facts-statistics-industry-overview#:~:text=In%202020%20there%20were%205%2C929,and%20other%20companies%20(1%2C227)

Jablecka, M. (2020). *Modelling CLV in the Insurance Industry Using Deep Learning Methods* [Master's dissertation, KTH Royal Institute of Technology]. Digitala Vetenskapliga Arkivet. https://www.diva-portal.org/smash/get/diva2:1431621/FULLTEXT02

Kecman, V. (2001). Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. *MIT press*.

Seyerle, M. (2003, June 17-19). *Customer Lifetime Value and its determination using the SAS Enterprise Miner™ and the SAS OROS-Software™* [Conference session]. SAS Conference Proceedings: SAS European Users Group International 2003, Vienna, Austria. https://support.sas.com/resources/papers/proceedings-archive/SEUGI2003/SEYERLE_LifetimeValue.pdf

Staudt, Y. & Wagner, J. (2018). What policyholder and contract features determine the evolution of non-life insurance customer relationships? A case study analysis. *International*

Journal of Bank Marketing, *36*(6), 1098-1124. https://doi.org/10.1108/IJBM-11-2016-0175

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on Neural Networks*, *10*(5), 988-999. https://doi.org/10.1109/72.788640