Brisa H, Njenga G, Rashmi K
12/06/2025
DATA 5100

Customer Lifetime Value: Modeling Customer Profile for Underwriting Optimization

The objective of this project is to predict Customer Lifetime Value (CLV) using customer profile data from automobile insurance policies. The IBM Watson Marketing Customer Value dataset provides over 9,000 anonymized records from 2011, including demographic, policy, and vehicle specific information. The dataset was of high quality, requiring no major transformations or imputation; however, a stratified relationship between Monthly Premium and CLV was observed. Exploratory data analysis uncovered that when a customer has exactly two policies, the usually linear relationship between Monthly Premium and CLV becomes striped and inconsistent. For this reason, we create five various Ordinary Least Squares (OLS) models and one Random Forest Regressor and use the resulting R-Squared values to determine the best model. The analysis identified two highly predictive approaches: individual OLS models stratified by policy count or a single Random Forest model; both approaches exclude two-policy customers. For situations requiring explainability, the OLS models are recommended. For a streamlined solution, the Random Forest is preferred. Further research is needed to understand and model the distinct behavior of two-policy customers.

Introduction

The aim of this study is to predict the Customer Lifetime Value (CLV) based on a customer's profile to assist in the determination of whether a prospective customer represents a profitable insurance policy to write. By building a model that can estimate CLV, we can help the insurance carrier determine whether the predicted CLV is high enough to meet the bar of profitability. This project also explores which features have the strongest impact on CLV, how accurately CLV can be predicted using a variety of models and how these models can support decision making when evaluating prospective customers.

CLV helps firms identify their most valuable customers by allocating marketing resources efficiently and designing retention strategies. In industries such as insurance where recurring policy renewals are common, accurately predicting CLV is especially important for strategic decision-making. PwC (2023) defines CLV as "a total financial contribution (revenue minus costs) of a customer over his/her lifetime with the company," encompassing both profit and time horizon. Accurately predicting CLV improves revenue growth, marketing spend, and competitive advantage (PwC, 2023).

This study uses IBM Watson Marketing Customer Value dataset that is available in Kaggle. This dataset includes 9,134 observations and 24 features describing demographic, behavioral, and policy-related characteristics of insurance customers with policies effective through 2011. It is unclear where the dataset initially emerged, as the author provided a now depreciated link to the original IBM Watson page from which it was sourced. The dataset contains a mix of numerical and categorical variables describing each customer's demographic profile, policy information and vehicle characteristics. Data preparation involved checking for missing values and duplicate entries, examining the features, standardizing column names and examining distributions for outliers or inconsistencies. The dataset was provided as a single CSV file, so no merging with other dataset was required. Nor was any imputation required, as the data was complete and free of missing values. The initial data exploration and preprocessing steps ensured that the dataset was of suitable quality for statistical analysis and predictive modeling. It is our recommendation that the model be retrained on up-to-date data prior to being productionalized in any applied setting.

Theoretical Background

The theoretical foundation of this project is grounded primarily in marketing and economics. From a marketing perspective, long-term customer retention is a central objective for insurance companies. Achieving this goal requires understanding customer characteristics and aligning product offerings with customer needs. Because insurance firms depend on sustained relationships, CLV becomes a key metric used to estimate the total financial contribution a customer is expected to make over the duration of their relationship with the company. This project therefore uses customer information to estimate their long-term value for underwriting and strategic decision-making.

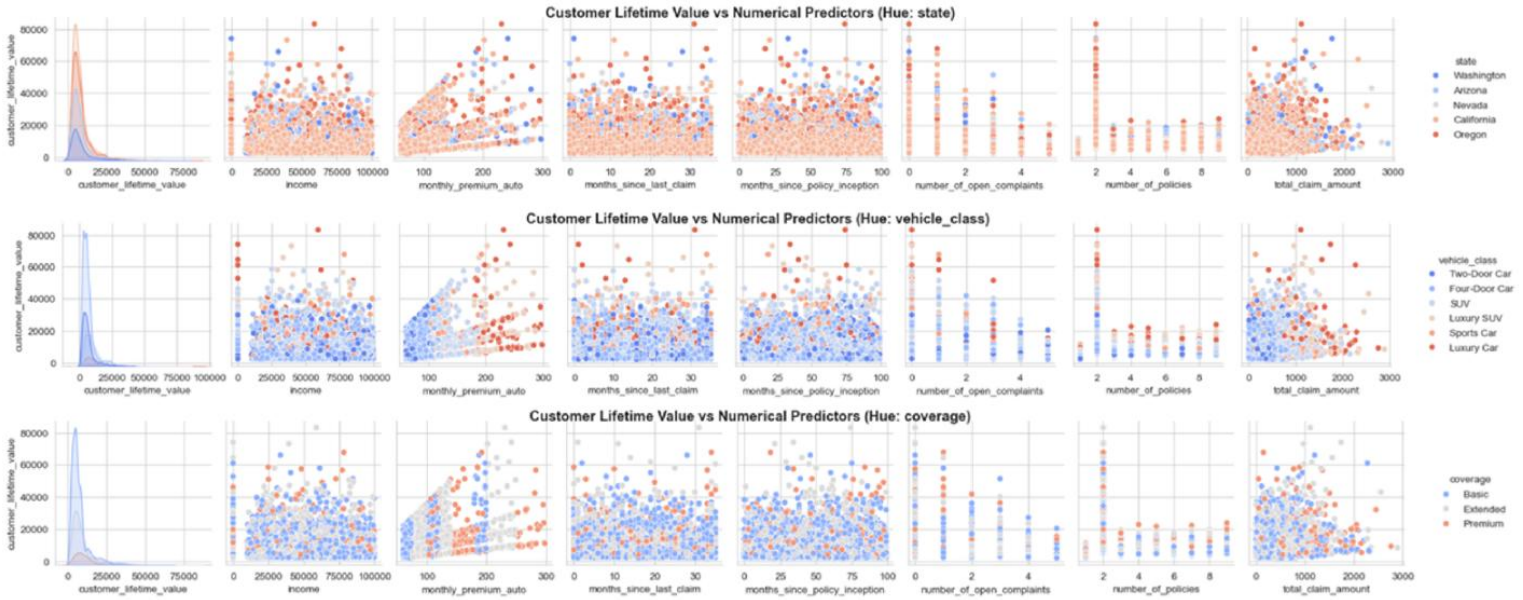
From an economic standpoint, customers purchase insurance policies to transfer the financial burden of potential losses to the insurer. The insurance company then operates within a framework of risk pooling, where insurers must assess the marginal impact of a given policy to the stability of the overall risk pool. A unique complication of CLV in insurance is that "revenues and costs vary by customer and over time for a specific customer; [so] CLV is much more than a function of volume" and that "highly variable cost structure yields small contribution to margins, meaning small changes in price can have a dramatic impact on CLV" (Firestone & Hindawi, 2013, p. 8). Therefore, a customer with a high predicted CLV is implicitly viewed as a more favorable risk for the company's pool, typically because their profile suggests consistent premium payments and manageable claims.

Methodology

After confirming the quality of the dataset provided, an exploratory data analysis was initiated to examine the distribution and behavior of Customer Lifetime Value (CLV) across key categorical and discrete numerical groups. We conducted a segmented analysis of the independent variables as well, to determine collinearity and unearth any other important relationships to note. Unnecessary variables were dropped, such as the CustomerID and the policy expiration dates.

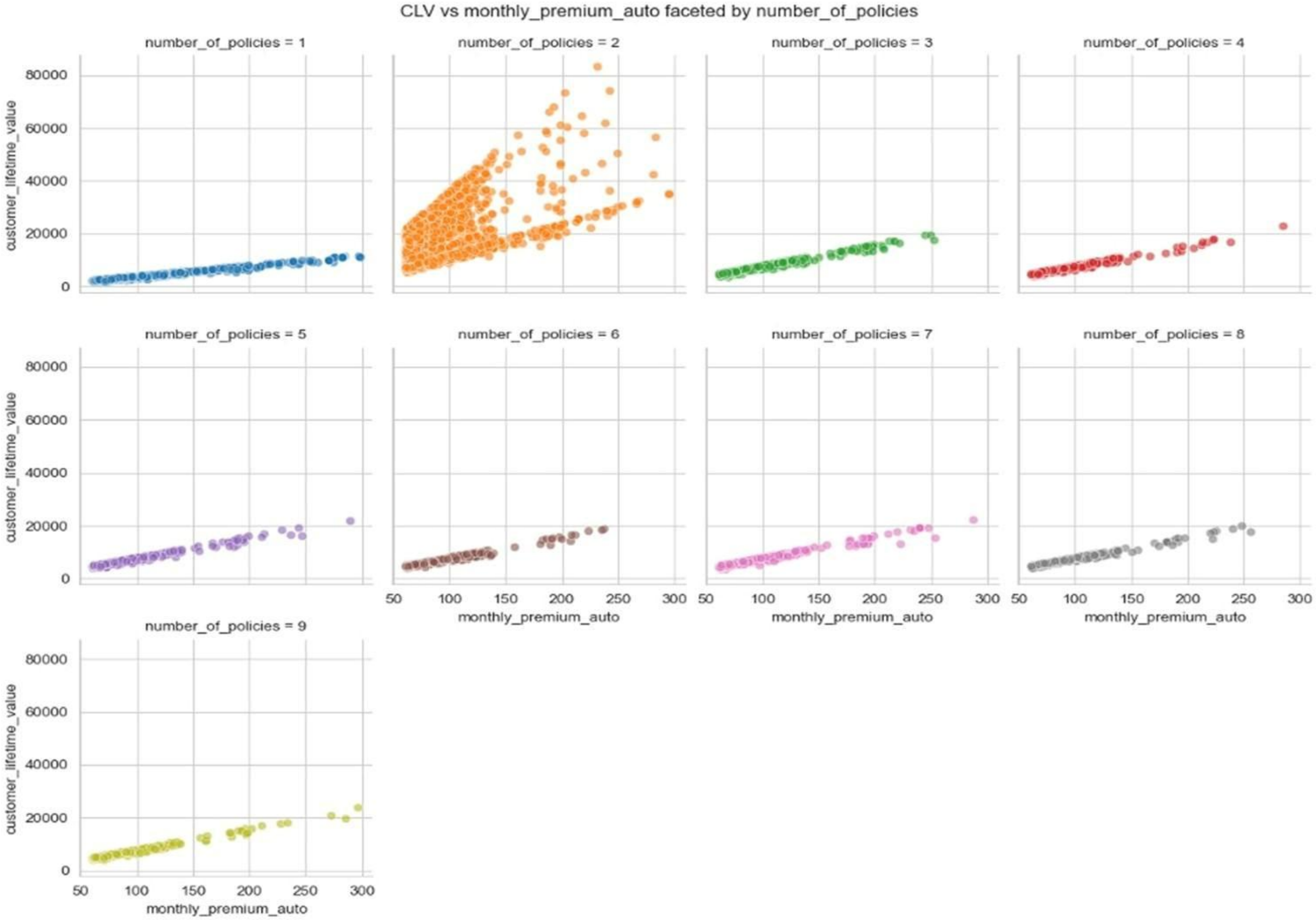
The dependent variable (CLV) displayed substantial skewness and several extreme outliers. The mean CLV was 8,004, while the median was much lower at 5,780 which indicated a right skewed distribution. CLV had values exceeding 80,000, which is highly improbable given no such skew in the monthly premium variable. To justify some of the outliers in CLV we would expect similar outliers in premium, since premium is the revenue source for the company and therefore a direct input of CLV.

To better understand the relationships between CLV and numerical predictors, we generated pair plots. These visualizations revealed a distinctive striped pattern between CLV and monthly premium. The presence of several linear stripes suggested an interaction effect from a numeric variable, as none of the categorical variables explained the observed behavior.



To investigate this further, we constructed a FacetGrid and treated numerical variables with fewer than 10 unique values as categorical for visualization. During this step, we consistently observed instability in patterns whenever the number of policies interacted with CLV and another numeric variable. These unstable patterns were irregular, difficult to interpret and did not align with trends observed elsewhere in the dataset.

We concluded that customers with exactly two policies display the most unstable patterns in the exploratory analysis outputs. After an exhaustive investigation, we were unable to identify a clear reason for this behavior suggesting the instability may be driven by some underlying factor not captured in the dataset. Because of this, we decided to handle the two-policy group separately during modeling and evaluate which modeling approach performs best when this group is treated differently.



We use two model types, in a variety of applications, to predict CLV based on the provided policy characteristics in our dataset. Our first five models utilize Ordinary Least Squares (OLS) linear regression modeling while our sixth model uses a Random Forest model. To meet the assumptions of an Ordinary Least Squares (OLS) regression framework, all categorical variables were converted into binary indicators using one-hot encoding, with the first category dropped in each case to prevent multicollinearity. Numerical predictors retained their original scales, as the exploration of the data did not indicate the need of normalization for the modelling purposes.

The methodology of our modeling followed an iterative approach, beginning with an OLS model that included all available customer attributes to establish a baseline (Approach 1). We then evaluated the model summary provided key evaluation metrics such as R-squared values, predictor coefficients, and corresponding p-values.

Given the stratifying behavior observed in customers with exactly two policies, we continued the use of OLS modeling in four more approaches. Approach 2 was developed to address the outliers in CLV that are most apparent in the subset of customers with exactly two policies. First, a cut-off for CLV is obtained by identifying the 75th percentile of CLV for customers with exactly two policies: a value of 19,916. Returning to the full dataset, we dropped all records where the CLV exceeds this amount and re-model the data similar to how we performed Approach 1. The purpose of this model is to omit the outliers from customers with two policies and assess if this improves overall model performance.

Approach 3 introduced a binary flag to indicate whether a customer had exactly two policies, denoted with a 1 or a 0. An interaction term was introduced between monthly auto premium and this two-policy indicator to capture the obvious interaction noted in the FacetGrid. Unlike the previous models, Approach 3 is highly reduced and models CLV using only the interaction term created.

Approach 4 continues to explore the interaction of policy count, by creating nine individual OLS models separating data by the exact number of policies and modeling each individually. At this stage, only the statistically significant predictor variables are included in each model to reduce the noise created from all the variables that offered no significant improvements to the predictions.

Approach 5 checks if the fourth approach can be paired down into two models: creating a model for all data *excluding* customers with two policies and a separate model for customers with *only* two policies.

Approach 6 utilizes a Random Forest Regressor to investigate potential non-linear relationships and improve predictive accuracy. The dataset was split into training and testing sets, the model was fitted, and performance was evaluated using R-squared. Because the two-policy group continued to affect the results, the Random Forest model was trained and evaluated on the data *excluding* customers with exactly two policies.

Computation Results

Our first model (Approach 1) produced an R-Squared of 0.169 with 48 degrees of freedom. Out of the 49 variables, 10 demonstrated predictive power with a P-value below 0.05. Statistically significant variables included monthly premium, number of open complaints, number of policies, high school and below education, employed employment status, single marital status, corporate policy and 3 types of renewal offers.

The R-squared of Approach 2 slightly improved to 0.247 and 15 variables showed statistical significance. Therefore omitting the customers above the two-policy group's 75th percentile of CLV did not provide significant explanatory improvement.

Adding a dummy variable for customers with exactly two policies and creating an interaction term between this dummy and monthly premium in Approach 3 provided significant lift over the previous approaches. The resulting R-squared was twice that of the second approach, at 0.666. We then tested a reduced model using only monthly premium and the interaction term. This simpler model still achieved an R-squared of 0.627, and highlighted the importance of this single interaction term's explanatory power.

In Approach 4, we built separate models based on the exact number of policies the customer has. For most policy counts, the models produced extremely high R-squared values (ranging from 0.993 to 0.996). However, for customers with exactly two policies, the R-squared dropped to 0.326. These models included only predictors with a p-value below 0.05. For two policy customers, there were only 5 statistically significant predictors, where other policy counts ranged from 12 to 26.

Approach 5 removed customers with exactly two policies and refit the OLS regression model, the R-squared was 0.796 with 21 significant variables.

In Approach 6, we developed a Random Forest Regressor. Using all policies, we achieved an R-squared of 0.686; a similar result to our R-squared in Approach 3. After removing customers with two policies, this model achieved an R-squared of 0.996—results similar to Approach 4 without requiring an individual model by customer policy count.

Summary of Models by Respective R-squared Values

Model	Description	R-squared
Approach 1	OLS regression with all variables	0.160
Approach 2	OLS regression omitting policies with CLV above the 75th percentile of 2-Policy customers	0.247
Approach 3	OLS with two-policy dummy and interaction term	0.666
Approach 3a	Reduced OLS model (premium + interaction term)	0.627

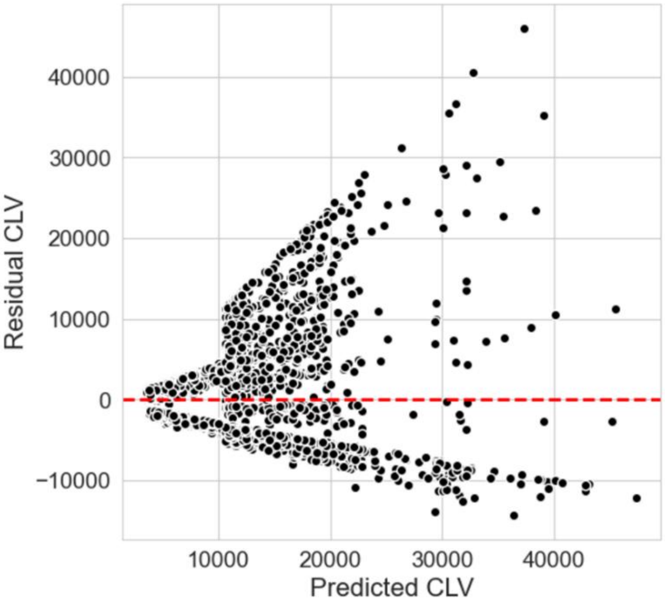
Approach 4	OLS models by policy count	1pol: 0.996 2pol: 0.326 3pol: 0.995 4pol: 0.995 5pol: 0.995 6pol: 0.994 7pol: 0.993 8pol: 0.995 9pol: 0.996
Approach 5	OLS regression excluding two-policy customers	0.796
Approach 5a	OLS regression with exactly two-policy customers	0.340
Approach 6	Random Forest model excluding two-policy customers	0.996

Discussion

As shown in our exploratory data analysis, customers with exactly two policies have a substantial negative impact on model performance. This is important to reiterate as we discuss the outcomes of our models. Our first two models attempted to model the data without specifically addressing the number of policies themselves; and neither demonstrated value. Approach 3 addressed the number of policies directly, by creating a new variable that acts as a flag for when a customer has exactly two policies. This binary categorization can then be modeled as an interaction term on monthly premium to capture the clear relationship these two predictors have. This relationship became apparent in the aforementioned FacetGrid, as the main explanation for the ‘striped’ pattern witnessed between CLV and monthly premium. It was observed that *only* customers with exactly two policies appear to have this stratification; all other customers have a strong linear relationship between monthly premium and CLV. While this interaction term provided immense improvement, with the R-Squared increasing to 0.666. The relative importance of this interaction is directly apparent when performing a single-variate regression using only the interaction term; resulting in an R-Squared of 0.627. It was this finding that informed our decision to design Approach 4 as individual models per customer policy count.

Approach 4 is composed of 9 individual models, one for each number of policies a customer has. If a prospective customer profile is drafted for a client with one policy, the model for single-policy customers would be used. If a customer has 9 policies, the 9-policy model would be used. This method found extreme success, provided the customer does not have exactly two policies. The R-Squared for customers with 1 or 3 policies was 0.996 and 0.995 respectively. In a business where transparency is a key consideration for regulatory considerations; these individual linear regression models would be a highly desirable methodology to predict CLV.

In an attempt to simplify, we then created Approach 5 which omits two-policy customers and performs a similar OLS regression analysis. The decrease in R-squared relative to the individual approach in Approach 4 is substantial, making the increase in effort from Approach 4 merit the extra steps. Furthermore, given the fan-shape of the residual plot for Approach 5, it is clear that the errors are inconsistent and therefore a poor model fit.



Approach 6, a Random Forest Regressor, provides a tradeoff of explainability in favor of convenience. When omitting customers with exactly two policies, the R-squared is 0.996. This is the ideal choice, unless a requirement for explainability necessitates the use of an OLS approach.

Despite much success with Approach 4 and Approach 6, the area in which all of our modeling has suffered is with customers who own exactly two policies. As it stands, we have no statistically significant way to predict a meaningful CLV from these clients. The Random Forest that included all customers had an R-Squared of 0.686 which is only marginally better than our linear regression model that included the interaction term of the two-policy flag on monthly premium (0.666). Yet, the Random Forest was able to achieve a near perfect R-squared when customers with two policies are omitted. Clearly there is a need to further investigate what makes customers with two policies so different, especially where premium is concerned.

We exhausted our ability to reach any meaningful conclusions with the dataset provided us, as the categorical variables and discrete numerical variables did not unlock any insight as to why these customers may have significant variation in their monthly premium. This discrepancy on such a potent predictor therefore impedes our ability to perform meaningful predictions. Ultimately, it would be our business recommendation to allocate further research into these customers and obtain additional data to help identify and explain what else may be causing the relative changes in premium. In the interim, we provide two methods for modeling CLV for customers who do not have exactly two policies; specific models for the number of policies or a random forest that excludes customers with two policies. The performance of these models is comparable; therefore, the best model would be determined by the level of explainability desired by the business.

Conclusion

Our goal was to build a tool that reliably estimates the long-term financial contribution of a prospective customer, using the customer profile information obtained in the customer application. We have built two powerful models: a linear regression model specific to the number of policies a customer has and a Random Forest regressor that omits customers with exactly two policies. Both achieve a similar and highly accurate prediction of a given customer’s CLV based on their profile. The former provides ample explainability, allowing regulatory bodies or business leaders direct insight into each variable’s effect on CLV. The latter provides a single model that can be used for any applicable customer, and is a ‘black box’ that can obscure the specific variable weights and predictive powers. For a given target of CLV, these models can determine how desirable a policy may be to write. These models can easily become classification models, if provided a specific threshold after which the predicted CLV suggests a policy should be written, or denied. However, significant work remains necessarily to unmask the issues observed with two-policy customers. It remains to be seen what specifically is causing the disparity in monthly premium for these customers; perhaps there are additional bundling discounts or risks carried by the secondary (non-auto) policy. A major limitation of our model is its inability to accurately predict for this population.

Citations:

PwC. (2023, March). *Customer lifetime value (CLV): December 2023–Maximising profits and shaping customer relationships* [PDF]. https://www.pwc.com/cz/en/risk-management-and-modelling/CLV_Final.pdf

Firestone, G., & Hindawi, M. (2013, March). *Customer lifetime value: Opportunities and challenges* (Paper No. 2119) [PDF]. https://www.casact.org/sites/default/files/presentation/rpm_2013_handouts_paper_2119_handout_796_0.pdf