**Brisa H, Njenga G, Rashmi K**

**10/26/2025**

**DATA 5100**

**Customer Lifetime Value: Modeling Customer Profile for Underwriting Optimization**

**Problem statement:** Predict the customer lifetime value (CLV) based on the customer profile, to decide whether a prospective customer would be a profitable policy to write.

**Data sources:** We will use the [IBM Watson Marketing Customer Value Data](https://www.kaggle.com/datasets/pankajjsh06/ibm-watson-marketing-customer-value-data/data) to create our predictive model. We will not need to impute any of the data since all of it is present, however we will want to perform basic feature engineering. We have a few known limitations to our data; first it is limited to the West Coast only. The states included are Arizona, Washington, California, Nevada and Oregon. Lastly, the data is based on policies effective through 2011. Our data, code, and reports can be found on [Git Hub](https://github.com/brisamh/modeling_clv).

**Analytical approach:** We began with a descriptive approach during the exploratory data analysis to understand relationships among our variables, and highlight any potential predictors of CLV. Our diagnostic analysis focused on the distributions of our numeric data, proportions of our categorical data, and overall correlations in our dataset. These results are used to test for any potential relationships present and to confirm predictive power. We aim to build a model that predicts CLV for new customer profiles, and potentially a prescriptive approach that recommends whether a new customer policy be accepted by our underwriters.

**Analysis:** Our dependent variable (y-value), CLV, is the estimated total revenue a business can reasonably expect from a single customer. 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). We will prepare a segmented analysis that will enable us to identify specific factors and characteristics that most contribute to a higher CLV.

We conducted a segmented analysis, beginning with the independent variables (x-values) we will use to model CLV. These variables include both categorical and numeric data. A few categorical variables include policy type, coverage type, vehicle class, gender, and location. Our exploratory data analysis uncovered some bias and skewness in our data. For example, California and Oregon combined represent 63% of our sample. We identified that on average, policies from Arizona have a much lower CLV relative to other state. Where coverage is concerned, policies with Premium coverage provide a significantly higher CLV, but represent only 12% of our sample.

The dependent variable itself displays substantial skewness and outliers: the mean value for CLV is 8,004 while the median is 5,780. A distribution shows extreme outliers beyond 80,000. Unsurprisingly, the distribution of monthly premium paid by the customer follows a similar skew. However, there are no such outliers for monthly premium, as the most expensive policies are still below $300 a month. Assessing the validity of the CLV outliers will be crucial to a successful analysis. From visual analysis of pair wise plots, we identified a ‘stripped’ pattern between CLV and monthly premium. The formation of these four clear ‘stripes’ indicate the present of an interaction of another variable on the relationship between these two. This suggests an interaction effect with another variable, likely number of policies, though this has not yet been confirmed.

**Preliminary Findings:** Our preliminary model was an ordinary least squares (OLS) linear regression model, with one-hot encoded categorical variables. From our pairwise plotting we expected this model to be insufficient, as there are considerable relationships between our relationship that warrant the use of interaction terms. Our preliminary model produced an R-Squared of 0.16 with 48 degrees of freedom. Of the 49 included variables, 10 demonstrated predictive power with a P-value below 0.05. Further checks of model validity were abandoned, since the R-Squared alone indicates a need for further refinement. Our next steps are outlined thoroughly in a later section, and include testing interaction terms, addressing skewness, and potentially transitioning to non-linear modeling.

**Solution technologies:** We plan to use Python 3, Github, Jupyter Notebooks and do not believe we will require any additional computational power to answer our business problem. After reviewing the results of our initial model, it is clear that additional analysis is required to identify meaningful explanatory power. If our attempts at linear regression continue to provide unsatisfactory results, we may explore the use of Random Forest models or Neural Networks; in which case we will utilize Google Collab notebooks to access the GPUs required.

**Challenge:** We know that the data is older and potentially not as relevant as data from more recent policyholders, furthermore we are not sure if we can extrapolate findings from the West Coast to the United States as a whole. We are also concerned that our sample size may not be representative of the whole population, even for some of the states where we have data due to large populations within them (California).

Additional challenges that were surfaced during our preliminary findings are the unknown interactions between the data, the potential lack of linear relationship, the effect of outliers in our dependent variable, and the possibility that we may need to use a completely different model type to best fit the data.

**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

**Group Dynamics:** Weekly progress updates, to-do lists and task delegation will be documented on both our Group 4 canvas page and the Git Hub. All other communications will be done using SeattleU email IDs. To maintain clarity, we will continue conversations within the same email thread whenever possible. We will share data and code using Git Hub. Because we are spread out geographically, and have various commitments outside of school we have agreed to set meeting times and deliverables as needed, rather than on a recurring on-going basis. We will often meet virtually via zoom, but prioritize in person meetings for high importance action items.

**Plan for Completion:** Our preliminary results lacked even the most basic explanatory power, and will require additional techniques to prove meaningful. The data provided was assumed to be fairly clean based on the lack of missingness and reasonable relationships; however additional time must be spent addressing outliers in CLV and identifying key interaction terms. The next steps will include:

* Identifying the optimal interaction terms to use in our OLS linear regression model and optimizing feature selection.
* Splitting the model into segments, if it is determined that the presence of some policy characteristics require the use of separate models for optimal output.
* Exploring additional model types if OLS linear regression does not sufficiently fit the data. This may be a Random Forest model or a cross-trained Neural Network.

Once the model is complete, we will spend time editing the notebook to cull and refine our markdown, commentary, and improve the flow of the analysis. We expect to spend at least one full day together to improve our existing model and/or implement a new one. Our initial EDA and modeling was performed in person over the course of 6 hours. We have set reserved another 6 hours together at Lemieux Library on **Friday, October 31st** from 11:00am to 3:00pm to improve and finalize our model results. At that time, we will delegate the remaining work between group members and make additional plans to work together.

Our group’s greatest impediment will be the difficulty of meeting in person. Brisa works fulltime during the week and must use paid time off to meet during the work week. Rashmi depends on public transportation to travel, and is a new parent to a young baby. Njenga commutes to Seattle from Fife, which costs significantly in both time and resources. Our greatest uncertainties are related to model performance; especially given such a weak preliminary model. However, our group has formed a strong bond and is committed to leveraging our communication skills to clearly define, assign, and complete deliverables on time.