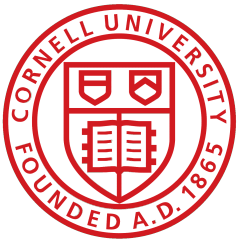

Cornell University®

ORIE 5741: Learning with Big Messy Data

Final Project

Predicting Customer Lifetime Value

Ryan Mao (rwm275),
Corwin Zhang (csz9),
Eugene Choi (ec727)

Executive Summary

In the evolving landscape of the insurance industry, understanding and predicting Customer Lifetime Value (CLV) has become crucial for maintaining competitive advantage and fostering customer loyalty. Having accurate CLV data helps companies to segment their market based on different customer profiles. It also helps to forecast profit, allowing companies to allocate resources effectively. Our dataset is sourced from VihanBima, a motor vehicle insurance provider. VihanBima wanted to segment their customer base using CLV to offer personalized service programs. Although India requires vehicles to be ensured, VihanBima has different policies to account for additional coverage. This dataset contains detailed customer demographics, policy choices, and qualification records, which provide a foundation for our analysis. Starting with exploratory data analysis, we seek to uncover patterns and relationships between CLV and other key features in our dataset. We will then discuss our methodology, detailing the models tested and the rationale behind selecting the best-performing model. Finally, we will present our findings and discuss the implications of our results for strategic decision-making in insurance policy offerings and customer relationship management. Our goal is to help predict CLV and also to provide actionable insights that can help VihanBima enhance their customer engagement strategies.

1 Data Processing

1.1 Data Description

Our dataset [1] was found on Kaggle, and provides information about VihanBima, a motor vehicle insurance provider that offers competitive pricing and claim services for personal and commercial vehicles. The main goal of the company is to enhance customer engagement through personalized service programs and thus, they need to figure out how to segment their customers based on customer lifetime value. The dataset contains various customer details such as their income, marital status, area they live in (such as urban or rural), and several other details. Additionally, there are also details containing information on the customer's selected insurance policy, such as which policy they selected, how many policies they have, and the amount they claimed. In total, there were 10 different explanatory variables in the base dataset.

1.2 Data Cleaning and Preprocessing

Before we could start working on the model, we needed to ensure that we fixed all of the corrupted, incorrectly formatted, or duplicate data within our dataset. Thus, the first step in our code was to clean our data by removing all characters that were not alphabetical or numerical (such as punctuation marks) and check for any null or missing values. This allowed us to be sure that each data point we had contributed meaningfully to our analysis and enhanced the accuracy and consistency of our data. Additionally, in order to prepare our data for our model, we needed to convert the categorical variables in our dataset into a numerical format. This is due to the fact that many machine learning models require input variables to be numeric in order to run and thus, they must be transformed during data preprocessing. To do this transformation, we made use of one-hot-encoding, which then allowed us to easily include the large number of categorical variables into our model and enhance its predictive power.

1.3 Exploratory Data Analysis

After completing data cleaning and preprocessing, we performed exploratory data analysis to gain a better understanding of the patterns within our data and discover any interesting relationships between features. To do this, we created a heat map to give us a visual representation of the relationships between our variables.

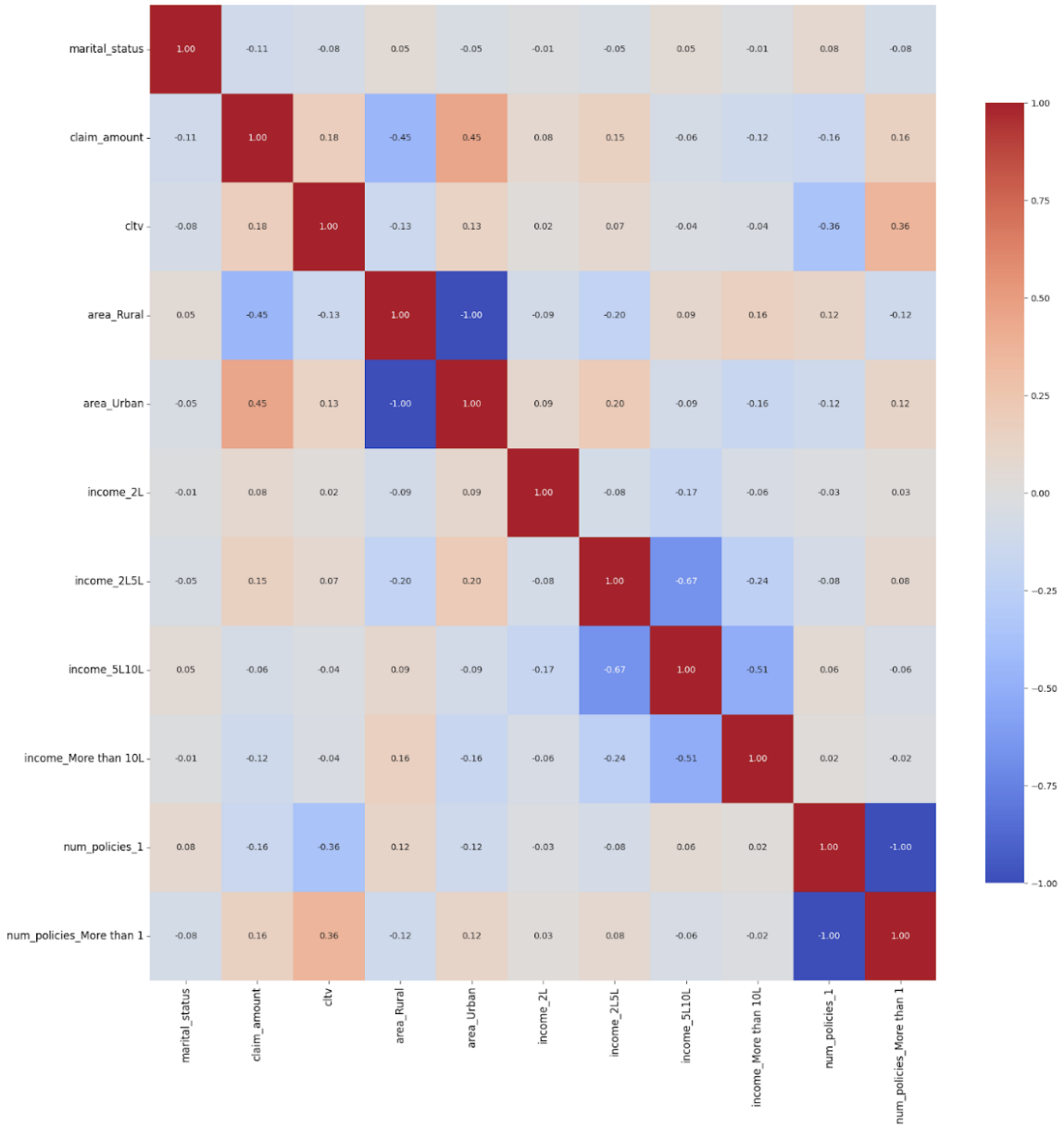

Figure 1: Heatmap visualization for subset of features in the dataset

The heat map shows the interconnectedness of our variables, which can help inform our analysis and serve as a guide for future exploration. From our plotted heat map, we can see that there are three variables that are more closely correlated with customer lifetime value: claim amount, area, and number of policies. From this, we know that we should focus more on these three variables in further exploratory analysis. Thus, our next step was to look more closely at the claim amount and how it's related to the number of policies and customer lifetime value. To gain a general overview of the data and better understand how to incorporate it into our model, we plotted a histogram for claim amount (Figure 2).

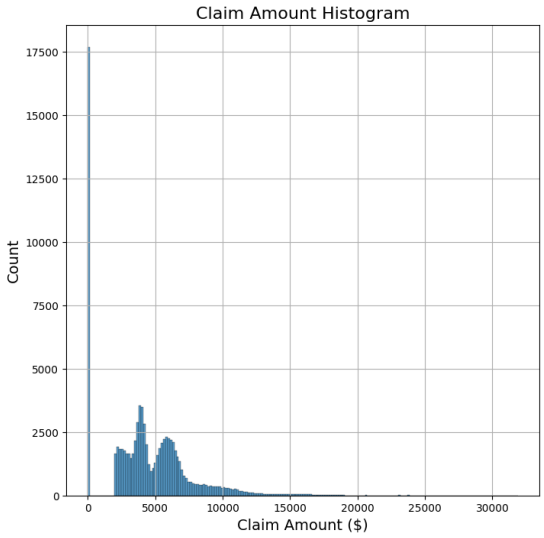

Figure 2: Histogram of Claim Amount

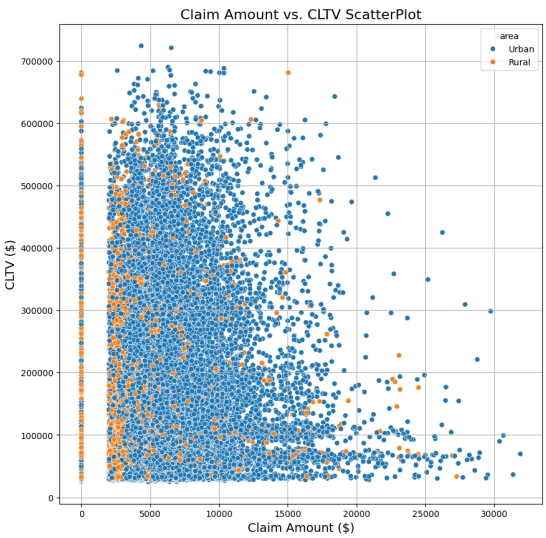

Figure 3: Claim amount & # policies vs. CLTV

Additionally, we also created a scatter plot to display the relationship between customer lifetime value, claim amount, and number of policies (Figure 3). From the scatter plot, we can see that there is a slight negative correlation between customer lifetime value and claim amount. Additionally, we can see from our color coding that customers with only one policy tend to have lower customer lifetime value and customers with more than one policy tend to have higher customer lifetime value.

1.4 Data Limitations

While our dataset provided a number of valuable insights, it does have some limitations that are worth mentioning. Firstly, the relatively smaller size of the dataset could have potentially restricted the depth of analysis and the different types of machine learning models that were available to our disposal. With a limited number of data points available for training, more complex models that require a substantial amount of data such as deep learning architectures may struggle to generalize well. Additionally, one of the most important fields in our dataset, `num_policies`, was provided in the form of a categorical variable, representing either one policy or more than one policy. If this feature was represented as a numerical variable representing the number of policies of a customer instead, we could perform a higher granularity of analysis. This categorization overlooks the variation in number of policies, which could potentially oversimplify the relationship between the number of policies a customer has and their customer lifetime value. These constraints could have hindered the accuracy and comprehensiveness of our model and other analytical conclusions derived from our model.

2 Methodology

2.1 Models

To predict customer lifetime value, we performed regression prediction on the following models: ElasticNet, XGBoost Regressor, Decision Tree Regressor, and Support Vector Machines. ElasticNet and the XGBoost Regressor are both extensions of models we have discussed in class. ElasticNet [2] is a regularized linear regression model that combines both the ℓ_1 and ℓ_2 regularization of lasso and ridge regression. The cost function for ElasticNet is given as:

$$\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$

The XGBoost Regressor [3], which stands for Extreme Gradient Boosting, is an algorithm that takes an ensemble of decision trees for either classification or regression and combines them to create a better model. It is based on boosting, where many single weak models are combined in order to generate a collectively strong model. This technique is designed to be both computationally efficient and highly effective. XGBoost primarily dominates in tabular datasets, which our dataset is structured like.

2.2 Model Results

From our analysis, we observed that the XGBoost Regressor performed the best on the test data. In the following sections, we aimed to further improve the base XGBoost Regressor by employing feature transformations and hyperparameter tuning.

Table 1: Base Model Results on CLTV

Model Name	**Mean Squared Error (MSE)**
ElasticNet	7092836625
Support Vector Regressor	9025004129
Decision Tree Regressor	13794302473
XGBoost Regressor	7034363944

2.3 Feature Transformations

After observing weak performance with the initial XGBoost Regressor trained on the dataset, we decided to plot the distribution of our numerical features (claim amount and customer lifetime value).

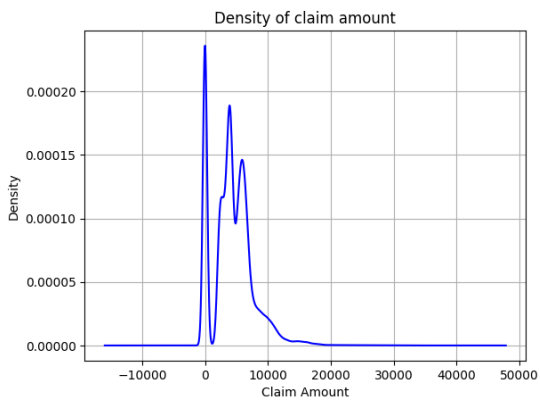

Figure 4: Density of Claim Amount

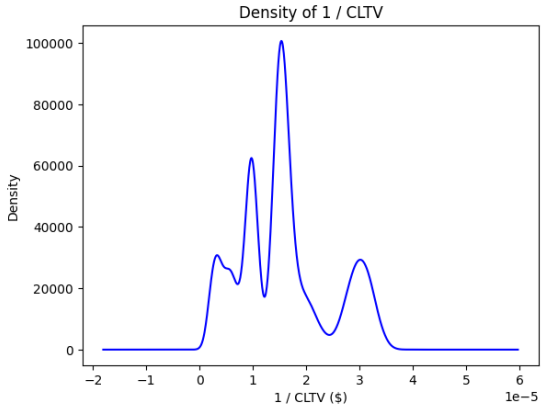

Figure 5: Density of CLTV^{-1}

From the graphs, one can notice that the distributions of the features appear jagged, with sudden peaks and valleys in the distribution. As a result, we decided to employ feature transformations to make the distributions of our numerical features more smooth. The motivation behind this transformation is because smoother distributions on features provide a cleaner and more reliable signal for the XGBoostRegressor model to learn from, ultimately leading to better performance. To be more specific, smoother distributions indicate less noise and more meaningful patterns in the data, allowing models to generalize better.

Table 2: XGBoost Performance with different Feature Transformations on CLTV

Feature Transformation	**Mean Squared Error (MSE)**	$\mathbf{R}^2$
Reciprocal	4.8009E-11	0.3148
Reciprocal w/ Grid Search	0.0000	0.3149
Quantile Transform	0.6441	0.3612
Quantile Transform w/ Grid Search	0.6305	0.3613
Power Transform	0.6163	0.3941
Power Transform w/ Grid Search	0.6003	0.3942

To make the distributions smoother, we applied a log transformation to the claim amount and applied various transformations within the target variable, customer lifetime value. The results of the different transformations are shown above in Table2. The various transformations on CLTV include reciprocal transformation, Quantile transformation, and Power transformation. Both Quantile and power transformation are common and popular preprocessing techniques that exist within scikit-learn, and they make sharp distributions smoother by redistributing the data points in a way that reduces the impact of outliers and extreme values, thereby promoting a more uniform and gradual distribution of values across the feature space. In our experimentation (see Table 2), we found that these transformations not only made the distributions smoother but also enhanced the performance of our XGBoostRegressor model by providing more stable and interpretable predictions.

2.4 Hyperparameter Tuning

To further improve the model, we also did a grid search over hyperparameters. Specifically, the hyperparameters we tuned were `colsample_bytree` (the fraction of randomly selected features that will be used to train each tree), learning rate, maximum depth of each tree, number of estimators, and subsampling ratio of training data. Hyperparameter tuning was performed using a grid search approach, where a predefined range of values was specified for each hyperparameter. The grid search algorithm exhaustively searched through all possible combinations of hyperparameters to identify the optimal set of values that maximizes a specified evaluation metric – in our case, this was Mean Squared Error (MSE) and R^2. The best results yielded hyperparameters of

```
{
    'colsample_bytree': 0.7,
    'learning_rate': 0.05,
    'max_depth': 3,
    'n_estimators': 200,
    'subsample': 0.8
}
```

We saw improvements in the model pre and post-hyperparameter tuning, with a decrease in MSE from 0.62 to 0.60 from initial hyperparameter settings, as shown in Table 2.

3 Results

Post hyperparameter tuning, the best regressor obtained an MSE score of 0.60 on the test set, as well as an R^2 of 0.3942. These results are fairly poor, as this implies that our best regressor can explain only 39% of the variance in CLTV from the independent variables collectively. However, we saw a significant decrease in the MSE when applying hyperparameter tuning and feature transformations, when compared to the base XGBoost Regressor model.

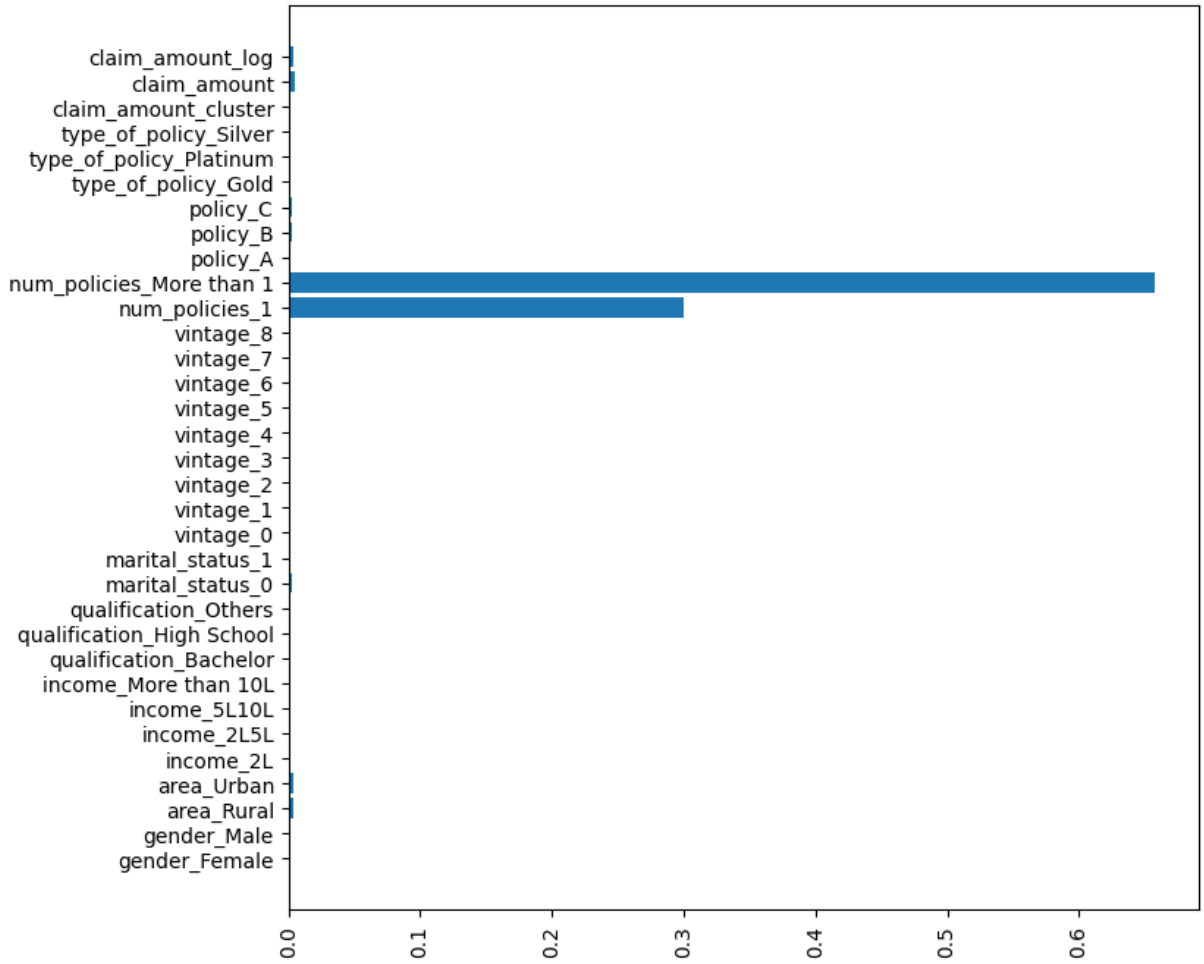

Figure 6: Feature Importance of all Descriptor Variables Considered

Further analysis of the model (Figure 6) yields that the most significant feature to model inference is the `num_of_policies` field, which is a binary categorical variable in this particular dataset (either more than one policy or not). However, this implies a rather promising future step; with how high the feature importance is for the number of policies, this suggests that any improvements in the number of categories in the `num_of_policies` field could help greatly increase the predictive power of our regressor. This suggests that the number of insurance policies held by customers significantly impacts the CLTV. Logically, this confirms that customers with a higher number of policies are likely to have a higher lifetime value to the business.

Following the number of policies, claim amount and area type have the second and third-most importance to predicting CLTV, respectively. These are weighed far less, with importance weights on the order of 0.003 to 0.004 for claim amount and area type. This implies that while these factors may still have some influence on CLTV, their impact is comparatively minor compared to the number of policies. The low feature importance of claim amount indicates that the total amount claimed by customers may not be a strong determinant of their lifetime value. Similarly, "area", which represents geographical location and market segment, also has limited influence on CLTV. Overall, the analysis highlights the critical role of the number of policies in determining customer lifetime value for the insurance business under consideration. However, it is essential to

recognize that other factors not included in the model may also contribute to CLTV, and further research or data collection may be needed to capture additional effective predictors.

4 Fairness and Weapon of Math Destruction

Our model aims to predict the lifetime customer value for motor vehicle insurance companies. As a result, variations of our model may affect issues of fairness as they may affect the pricing and decisions of the insurance company. As a result, certain individuals or groups may be inadvertently hurt due to these predictions causing them to be get charged more or get worse policies and benefits. As a result, fairness must be considered since errors could have harmful effects on insurance holders. For example, if predictions show that certain areas or certain demographics have lower CLV they might get worse policies to account for their lower CLV, and as a result, more of those of a certain area/demographic would end up not choosing insurance with them creating a negative feedback cycle. This issue could cause a weapon of math destruction if not addressed properly. This is because calculating the CLV for each customer is impossibly complex and could prove harmful to certain customers. In addition, we must consider how to deal with protected attributes like race and gender for policyholders. In future iterations to alleviate these problems we might add unawareness of classifiers and measures of fairness such as demographic parity, equalized odds, and equality of opportunity.

5 Conclusion

From our results, we are fairly confident that for VahanBima, the best predictor of Customer Lifetime Value is the number of policies the customer holds. This makes sense, as customers with more than one policy are more profitable for the company and more likely to trust the company more, which can translate to higher loyalty, and in turn, higher lifetime value. Our final tuned model had a greatly reduced MSE of 0.6 when compared to the base model; however, it is still not completely accurate as our R^2 of 0.39 implies that much of the features in the data are still left unexplained. Additionally, it is likely that our model was overfitting the limited training data that we had available. With more training data, along with additional features, such as age and claim history, we expect our model to perform better.

Based on our results, VahanBima can consider revisiting and potentially adjusting their key performance indicators (KPIs) to reflect the importance of policy count in predicting CLV. This may involve incorporating metrics related to policy cross-sell and upsell rates and customer retention rates among multi-policy holders. Most importantly, VahanBima should consider segmenting customers based on their policy holdings to better tailor marketing efforts and customer experiences. For example, high-value customers who hold multiple policies may warrant special attention and personalized offers to maintain their loyalty and lifetime value. On the other hand, customers with fewer policies may require more targeted efforts to increase their loyalty.

6 References

[1]: https://www.kaggle.com/datasets/gauravduttakiit/predict-cltv-of-a-customer/data

[2]: https://corporatefinanceinstitute.com/resources/data-science/elastic-net

[3]: https://machinelearningmastery.com/xgboost-for-regression/

[4]: https://www.kaggle.com/code/ac1414/feature-transformation-techniques