Unlocking Future Value: Optimizing Insurance Offerings with CLTV Prediction and Customer Segmentation

Shreyas Hingmire, Nayan Bhiwapurkar, Sudeeksha Vandrangi, Vaishnavi Chunchu, Toshal Warke
School of Computing and Augmented Intelligence
Arizona State University
Tempe, Arizona, US
{shingmir, nayanbhiwapurkar, svandran, vchunch1, twarke1}@asu.edu

Abstract

This report details predictive modeling for customers of Vahan Bima, a motor vehicle insurance company in India, to enhance the personalization of motor vehicle insurance policies through Customer Lifetime Value (CLTV) prediction. Utilizing advanced data analytics and machine learning techniques, the project aims to segment customers based on predicted CLTV, enabling tailored insurance offerings and improved resource allocation. The results demonstrate the efficacy of Gradient Boosting, Mini Batch K-Means, and BIRCH clustering algorithms in predicting CLTV and segmenting customers, significantly enhancing marketing strategies and customer service. This initiative seeks to improve Vahan Bima's operational efficiencies and could influence data-driven, customer-centric practices in the insurance industry.

Keywords: Customer Lifetime Value, Predictive Modeling, Insurance Personalization, Customer Segmentation

I. Introduction

A. Background

In the evolving landscape of the insurance industry, companies are faced with the challenge of enhancing customer satisfaction while improving their service offerings. Vahan Bima, a leading Indian motor vehicle insurance company, is at the forefront of innovation, seeking strategies to personalize its insurance policies. To foster a long-term bond with the customer to promote retention and their value to the company, the Customer Lifetime Value (CLTV) is a key metric in this form of evaluation. CLTV calculates the current value of a retained client for the business by projecting future transactions and expenses [1]. It provides a forward-looking perspective on customer relationships by estimating the potential value that each customer brings over their lifetime. This metric helps businesses understand the value i.e. the profitability of each customer and tailor their services accordingly. By focusing on customer experience, loyalty, and retention, companies can drive profitability [2].

In this project, we aim to build a machine learning model to forecast the CLTV metric for the customers of Vahan Bima company based on user and policy data (a regression task) and then segment these customers based on some common attributes (a clustering task).

B. Problem

This initiative by Vahan Bima utilizes machine learning to estimate Customer Lifetime Value (CLTV) using extensive user and policy data. The accurate prediction of CLTV will enable the company to:

- **Segment Customers:** Effectively categorize customers into specific groups according to their estimated CLTV, facilitating tailored and personalized policy options.
- **Enhance Personalization:** Apply insights from CLTV forecasts to customize insurance policies, ensuring they better meet the unique needs and preferences of individual customers.
- **Optimize Resource Allocation:** Apply CLTV-driven segmentation to focus marketing and customer service resources on the most valuable customer segments.
- **Improve Customer Satisfaction and Loyalty:** The project aims to boost customer satisfaction and loyalty through more personalized services, thereby increasing customer retention.

C. Importance

The development of a predictive model for Customer Lifetime Value and its application in customer segmentation represents a groundbreaking initiative in the insurance industry. This project goes beyond the immediate operational improvements for Vahan Bima and signifies a shift towards a more data-driven and customer-centric approach in the insurance industry [3], [4]. This approach not only promises enhanced customer experiences but also paves the way for more sustainable business growth through improved customer value management. In today's competitive market, organizations have recognized the importance of establishing a profitable and long-term relationship with customers and it is very crucial to the success or failure of the organization [3], [4].

This project aims to revolutionize personalized insurance services by creating and using a predictive model for Customer Lifetime Value (CLTV) in customer segmentation. It supports Vahan Bima's dedication to innovation and customer satisfaction, advancing the company's leadership in the competitive insurance market.

D. Literature Survey

The CLTV metric is an important tool for insurance companies to assess the long-term value of their customers and make informed decisions regarding customer acquisition, retention, and resource allocation [5]. Many studies have explored the development process and applications of CLTV in insurance companies. Understanding CLV may help create a profile of high-value clients that can be applied to a prospect list to improve the effectiveness and efficiency of customer acquisition activities [6].

Ekinci et al. (2014) conducted a study in the banking sector on a detailed model that could forecast the CLTV value and compared the results obtained from the Least Squares Estimation (LSE) technique and the artificial neural network (ANN) [7]. They found that the LSE-based linear regression model performed better. They further deduced that features such as monetary value and risk of certain banking institutions play a crucial role in the CLTV measurement, apart from the usual factors that are considered.

Yoseph et al. (2020) attempted to do market segmentation based on the CLTV in the retail industry. Their study included three unique experiments which included a modified best-fit regression using the Expectation-Maximization (EM) method and the K-Means++ for clustering [8]. They were able to implement their results in a department store and recorded an increase in the sales growth rate from 5

Abidar et al. (2022) suggested a novel model using Recency, Frequency, Monetary (RFM), and CLTV using a clustering technique [9]. With this method, their model was able to capture patterns in the transactional data of the customers, and they were successfully able to propose data-driven actionable plans for each segment.

Another study by Lam & Sunny (2018) proposes a two-fold framework for predicting customer profitability, where the first step involves developing a dichotomous model to predict the likelihood of purchase in the future, and the final step involves considering a continuous target variable and building a model that could forecast a profit for the organization with the given condition that the customer would make the predicted purchase [10].

In the existing literature, there is a notable inadequacy of research on the application of the Customer Lifetime Value (CLTV) metric within the insurance industry, particularly concerning its role in forecasting customer profitability and informing the targeted marketing of insurance policies through customer segmentation. This project seeks to investigate and fill this void by not just forecasting the CLTV for a customer but going a step further by also attempting to segment the customers into appropriate clusters based on their predicted CLTV and give them tailored suggestions for insurance policies.

E. System Overview

This project integrates a comprehensive, data-driven workflow designed to predict the customer lifetime value. Starting with the thorough collection and preprocessing of customer data, the project then moves into exploratory data analysis (EDA) to uncover critical data insights and anomalies. This prepares the dataset for division into training and testing sets, ensuring models are effectively trained and validated. Advanced machine learning models, particularly Gradient Boosting, are then trained, selected based on performance metrics, and refined through hyperparameter tuning and K-Fold cross-validation to optimize accuracy. The final phase employs clustering techniques like Mini Batch K-Means and BIRCH to segment customers by predicted CLTV.

F. Data Collection

The dataset is derived from Vahan Bima's comprehensive records, encapsulating a wide array of customer-related and policy-related information. This dataset is instrumental in understanding customer behaviors, preferences, and interactions with insurance policies, which are crucial for predicting Customer Lifetime Value (CLTV) and for subsequent customer segmentation.

G. Key Features

The dataset encompasses several categories of features, each contributing unique insights into the customer profile and behavior:

1) **Customer Demographics**

- **Gender:** This variable can help in understanding the distribution of products across different genders and tailoring gender-specific insurance packages.
- **Area:** Knowing whether customers are from urban or rural areas can assist in customizing insurance products to suit region-specific risks and preferences.
- **Qualification:** The level of education might influence the type of insurance policies purchased, with potentially more educated customers opting for complex or higher-value policies.

2) **Policy Details**

- **Type of Policy:** This indicates what kind of insurance policies are being chosen, like 'Platinum' or 'Gold', which could reflect the level of coverage and the customer's willingness to invest in higher premiums for more comprehensive benefits.

- **Income:** This factor can indicate which insurance products are within a customer's financial reach and their ability to pay for more comprehensive coverage.
- **Marital Status:** Married customers might have different insurance needs, such as policies that cover family members, compared to single individuals.
- **Num_Policies:** The number of policies held by a customer can reflect their trust in the insurance provider and the diversity of their coverage needs.

3) **Engagement Metrics**

- **Vintage:** This represents the length of time the customer has been with the insurance company, which can be a strong indicator of loyalty and satisfaction.
- **Claim Amount:** Reflects the cost of claims made, indicating the risk associated with the customer and possibly the sufficiency of coverage they have.

4) **CLTV**

- **CLTV:** This metric is a prediction of the net profit attributed to the entire future relationship with a customer. It integrates various aspects, like the revenue from the customer, the costs of servicing them, their engagement level, and potential for future revenue.

H. Components of ML System

We trained different Machine Learning (ML) models using the training dataset to discern which models best capture the complexity of our data. The Model Selection stage follows, where the most accurate models are identified based on their performance on the test set. Following this, we focus on Improving Accuracy by fine-tuning model hyperparameters and employing K-Fold cross-validation to ensure robustness and minimize overfitting. Similar procedure is followed for clustering. Finally, we Calculate the Accuracy of the chosen model and predict the CLTV.

I. Experimental Results

We applied the Gradient Boosting algorithm to predict Customer Lifetime Value (CLTV), achieving an R-squared value of 0.925 on the test dataset, indicating a high level of predictive accuracy. Subsequently, we performed customer segmentation, dividing the customer base into three distinct segments, resulting in a commendable mean Sum of Squared Errors (SSE) of 1.77 across all clusters, which demonstrates tight clustering.

II. Important Definitions and Problem Statement

A. Important Definitions

- **Mean Absolute Error (MAE):** Mean Absolute Error is a statistical metric used to evaluate the precision of a regression model. It computes the average of the absolute differences between the actual values and the predictions, providing insights into the typical size of the errors while ignoring their direction. MAE is valued for its direct interpretability of errors and its relative insensitivity to outliers, unlike other metrics such as Mean Squared Error.
- **Mean Squared Error (MSE):** Mean Squared Error is a widely utilized metric to assess the precision of a regression model. It determines the average squared discrepancies between predicted values and actual observations, emphasizing the punishment of larger errors more heavily. This characteristic renders MSE sensitive to outliers, making it particularly suitable when large errors are especially problematic.
- **R^2 Score:** The R^2 score, or coefficient of determination, serves as a statistical metric in regression analysis to evaluate how well a model fits the data. It measures the percentage of variance in the dependent variable that can be predicted from the independent variables.
- **SSE (Sum of Squared Errors) in Clustering:** SSE in clustering is a measure of the total variance within clusters, calculated as the sum of the squared distances between each member of a cluster and its centroid.
- **Davies-Bouldin Index (DBI):** The Davies-Bouldin Index in clustering is a metric that evaluates clustering quality by calculating the average similarity between each cluster and the most similar one, with lower values indicating better clustering separation.
- **Mini Batch K-Means:** Mini Batch K-Means is a variant of the K-Means clustering algorithm that uses small, random subsets (batches) of the dataset to update cluster centers more frequently, providing significant computational advantages and speed improvements over traditional K-Means.
- **BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):** BIRCH is a clustering algorithm designed for very large datasets, which incrementally constructs a tree structure representing data, enabling efficient hierarchical clustering with minimal memory use.
- **Gradient Boosting:** Gradient Boosting for regression is a predictive modeling technique that builds an ensemble of weak regression trees sequentially, with each new tree correcting errors made by the previous ones, to improve accuracy and reduce bias.

B. Problem Statement

1) Given: Vahan Bima, as a leader in the motor vehicle insurance market, has amassed a substantial customer database through its operations. Despite the data availability, there exists a gap in effectively utilizing this data to maximize the company's growth and customer satisfaction. Currently, the company lacks a systematic approach to quantify the future value of its customers and to personalize insurance offerings accordingly. This has led to a one-size-fits-all marketing strategy and a potential misallocation of resources, ultimately impacting customer retention and the profitability of Vahan Bima. Additionally, the existing data has not been leveraged to its full potential due to the lack of advanced analytical methods and tools being applied. The project is initiated in the context of these circumstances with the intent to harness machine learning capabilities to drive data-informed decisions in policy personalization and customer value optimization.

2) Objectives: The goal of the "CLTV Predictive Segmentation for Personalized Insurance Policies" project is twofold, focusing on both predictive modeling and customer segmentation to enhance the personalization of insurance policies offered by Vahan Bima, a leading motor vehicle insurance company in India. By leveraging advanced data analytics and machine learning techniques, this project aims to achieve the following objectives:

1) **Predict Customer Lifetime Value (CLTV):** The primary objective is to develop a predictive model that accurately estimates the Customer Lifetime Value (CLTV) of Vahan Bima's customers. CLTV is a metric that represents the total net profit a company expects to earn from a customer over the duration of their relationship. The ability to forecast CLTV allows Vahan Bima to:
   - Understand the long-term value of customers.
   - Identify high-value customers who are likely to contribute significantly to the company's profitability.
   - Tailor marketing and customer service strategies to retain customers with high CLTV.
2) **Segment Customers for Personalized Policies:** Utilizing the insights gained from CLTV predictions, the project seeks to segment the customer base into distinct groups with similar attributes and predicted CLTV. This segmentation enables Vahan Bima to:
   - Design and offer personalized insurance policies that cater to the specific needs and preferences of different customer segments.
   - Allocate resources more efficiently by focusing on segments that are predicted to be more profitable in the long run.
   - Enhance customer satisfaction and loyalty by providing services and policies that are more closely aligned with

customer expectations.

3) Challenges:

- **Data Quality and Structure Challenges:** The dataset contains several complexities, including categorical columns that require encoding to fit into algorithmic models, and numerical columns with varied ranges that necessitate normalization.
- **Feature Selection and Algorithm Choice:** The project faced challenges related to the presence of some irrelevant features that complicate the modeling process. Deciding on the most effective clustering algorithms — Mini Batch K-Means and Birch — also posed a significant decision point, given the dataset's characteristics and computational limitations.
- **Parameter Optimization and Model Evaluation:** Identifying the optimal number of clusters, 'K', was crucial and involved the use of methods like the Elbow method, silhouette analysis, and gap statistic to ascertain the most effective clustering strategy. The choice of evaluation metrics, including the Silhouette Score, Davies-Bouldin Index, Adjusted Rand Index (ARI), and Sum of Squared Errors (SSE), also added layers of complexity to the model validation process.
- **Computational Constraints:** The project was additionally challenged by insufficient computational power, which limited the scope of data processing and model training, impacting the speed and scalability of the analysis.

III. Overview of Proposed Approach/System

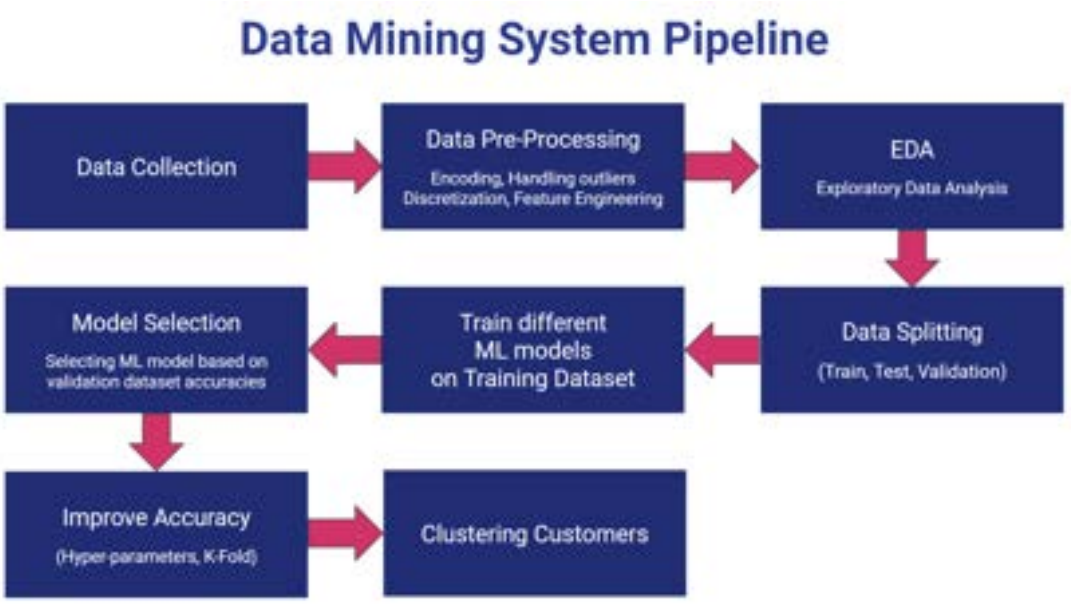


Fig. 1: Data Mining System Pipeline

This is the streamlined overview of our Data Mining System Pipeline. Each phase of this pipeline is crucial for deriving actionable, data-driven insights that inform strategic business decisions.

1) **Data Collection:** We gather customer data from various touchpoints to form a robust dataset.
2) **Data Pre-processing:** This stage involves cleaning the data, encoding categorical variables, binarizing the claim amount, handling outliers, and performing feature engineering, which includes feature selection and normalization.
3) **Exploratory Data Analysis (EDA):** We analyze the data to identify trends, patterns, and relationships between variables, providing insights into the underlying data structure.
4) **Data Splitting:** We divide the data into training and testing sets to evaluate our models effectively and ensure they generalize well to unseen data.
5) **Model Training:** Different machine learning models are trained on the dataset to identify which best captures the nuances of customer behaviors.
6) **Model Selection:** We select the model that demonstrates the highest accuracy on the test dataset.
7) **Accuracy Improvement:** Through hyperparameter tuning and K-Fold cross-validation, we refine our model to enhance performance and ensure stability across data subsets.
8) **Customer Segmentation:** Post-regression, we employ clustering techniques to segment customers based on their predicted CLTV and behaviors, enabling tailored marketing strategies and improved customer service.

IV. Technical Details of Proposed Approaches/Systems

A. Data Preprocessing

After collecting the necessary data, we proceeded with data preprocessing which is essential to cleanse and organize raw data, ensuring it is suitable for building accurate and efficient predictive models. It facilitates better data analysis and machine learning performance. The dataset did not contain any null values.

1) Handling Categorical Column: We applied a label encoder to the ordinal columns 'qualification' and 'income' to convert these categorical text labels into a numerical format that can be easily interpreted by machine learning algorithms. We assigned a unique integer to each unique categorical label in a manner that reflects the inherent order in the data, which is crucial for models that rely on the relationship between variable categories. Then we applied one-hot encoding to transform the remaining categorical variables into a binary matrix representation. This method creates new columns, each representing a possible category value with a 1 (present) or 0 (absent), preventing algorithms from misinterpreting ordinality in nominal data.

2) Reduction of Multicollinearity: Next, we visualized the correlation heatmap of all columns with CLTV column, to identify which features have the strongest relationships with CLTV. Based on the results, we found out that the 'income' column had very little correlation with 'cltv', hence we removed it. Multicollinearity can significantly affect the performance and interpretability of regression models by making the estimates highly sensitive to changes in the model. To address this, we used the Variance Inflation Factor (VIF) which measures the extent of correlation between one predictor and the rest of the predictors in a model. A high VIF indicates a strong correlation, which can undermine the statistical significance of an explanatory variable. In this specific case, the variable 'Gender_Male' was identified to have a high VIF, suggesting it was linearly dependent on other variables. Consequently, 'Gender_Male' was removed from the dataset to reduce multicollinearity, thereby enhancing the stability and interpretability of the resulting regression models. This step ensures that each remaining feature in the model contributes uniquely to the prediction, enhancing the robustness of the model's performance.

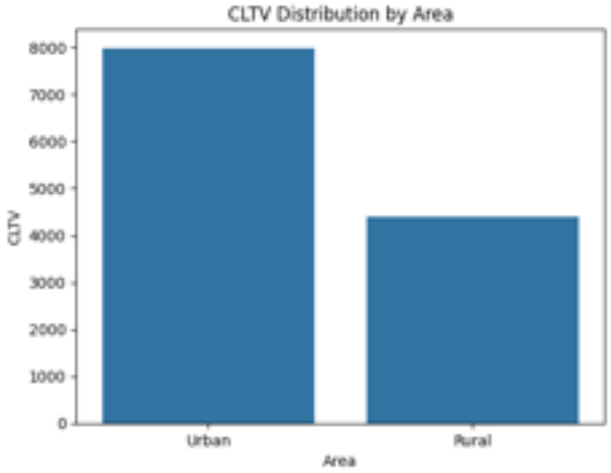
3) Handling Numerical Columns: We applied Normalization, a scaling technique in data preprocessing that adjusts the range of feature values to a common scale, typically 0 to 1, to all numerical columns. This process is essential in machine learning as it ensures that each feature contributes equally to the analysis, preventing features with larger ranges from dominating the decision-making process of algorithms, particularly those sensitive to input scales such as gradient descent-based methods and distance-based algorithms like K-nearest neighbors (KNN).

B. Exploratory Data Analysis (EDA)

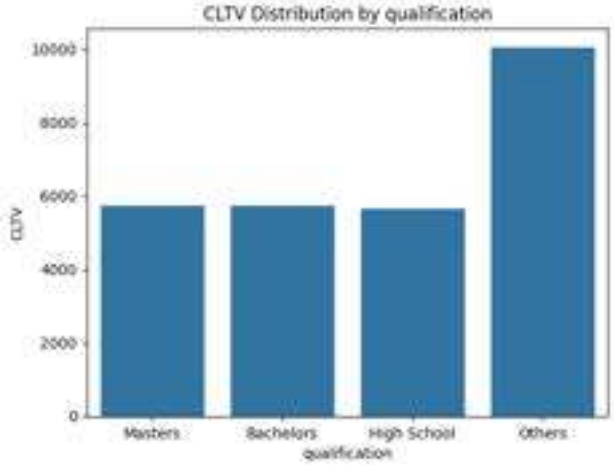
We then proceeded with Exploratory Data Analysis (EDA), which is a crucial step as it helps in identifying trends in data, detecting outliers and anomalies, understanding relationships between variables, etc. It was useful in identifying and understanding the factors that might influence a target variable, such as Customer Lifetime Value (CLTV), in our dataset.

Here's how EDA can be employed effectively for this purpose:

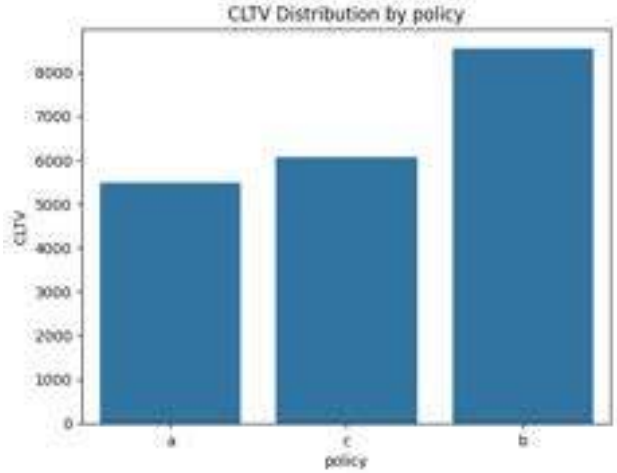
- **Distribution Analysis:** By analyzing the distribution of the CLTV variable and other predictors, we gained insights into the range, central tendency, and dispersion of these variables.
- **Correlation Analysis:** Creating correlation matrices and heatmaps helps in spotting which predictors are most strongly linked to CLTV.
- **Visualization Techniques:** Utilizing scatter plots, box plots, and bar charts to visualize relationships between CLTV and predictor variables can highlight trends and patterns. For example, scatter plots can help discern linear relationships, whereas box plots might show how CLTV varies with categorical variables.



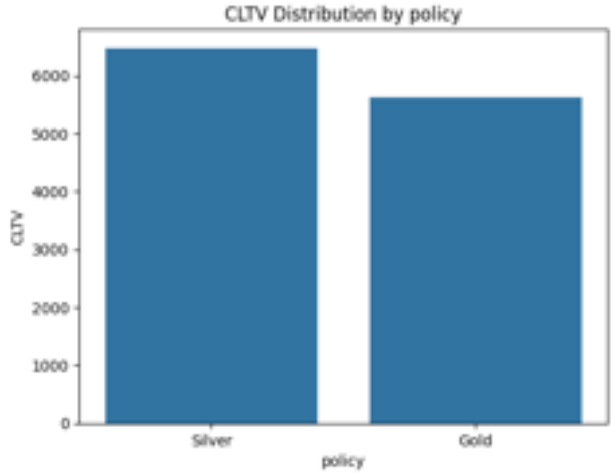
(a) CLTV Distribution by Area



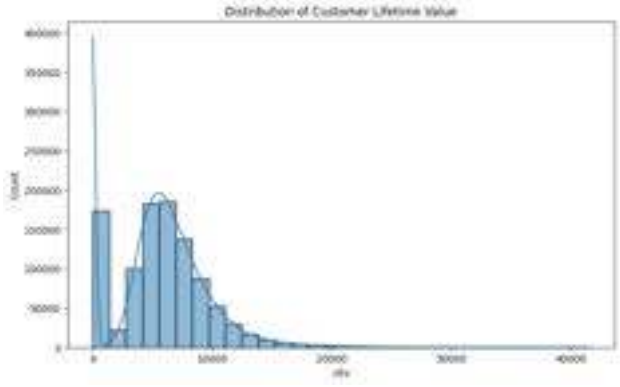
(b) CLTV Distribution by Qualification



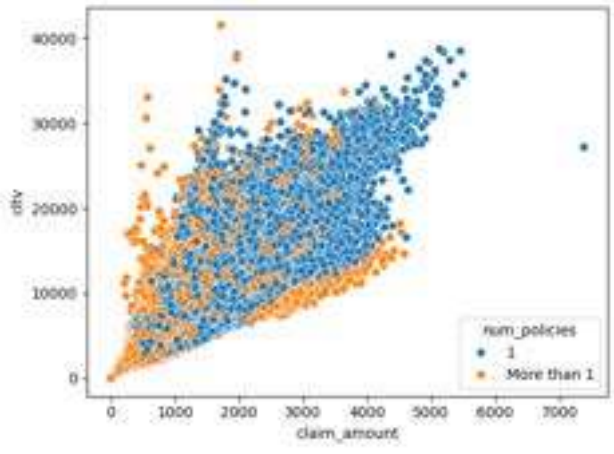
(c) CLTV Distribution by Policy



(d) CLTV Distribution by Policy Type



(e) Distribution of Customer Lifetime Value



(f) Relationship between CLTV and Claim Amount

Fig. 2: Exploratory Data Analysis (EDA)

Through EDA, we found actionable insights that could inform strategic decisions or feed into more complex predictive models, ultimately improving business outcomes related to customer value management.

Some insights we found were,

- **Fig. 2a** illustrates that the average CLTV is significantly higher in urban areas compared to rural areas, indicating that customers' geographic location may be a strong predictor of their lifetime value to a business.
- **Fig. 2b** indicates that individuals categorized under 'Others' in qualification have the highest average CLTV. This suggests that the 'Others' group contains key segments that are particularly valuable to the business.
- **Fig. 2c** reveals that policy type 'b' is associated with the highest average CLTV, suggesting that this policy type may be more lucrative or appealing to customers, thus potentially a strategic focus for the business.
- **Fig. 2d** suggests that the Silver policyholders have a marginally higher average CLTV compared to Gold policyholders, which could imply that Silver policies strike a better balance between benefits and costs, attracting or retaining more profitable customers.
- **Fig. 2e** is a histogram with a superimposed line graph that shows the distribution of Customer Lifetime Value (CLTV), indicating that the majority of customers have a CLTV around a specific value with a right-skewed distribution, suggesting a smaller number of customers with very high lifetime value.
- **Fig. 2f** is a scatter plot that illustrates the relationship between claim amount and Customer Lifetime Value (CLTV) for customers with different numbers of policies. It shows a positive correlation between claim amount and CLTV, with the concentration of single-policy holders being denser across lower claim amounts, while customers with more than one policy are more dispersed and extend to higher claim amounts and CLTV.

V. Experiments - Predicting CLTV and Segmenting Customers

A. Regression

1) Trying Different Regression Models: We constructed a regression model to predict Customer Lifetime Value (CLTV). The dataset was split into training and test sets using an 80:20 ratio. Subsequently, several regression models were implemented to ascertain the model yielding the highest accuracy. The performances of different models are depicted in Table 1.

TABLE I: Model performance comparison

Model	R-Squared (Training)	R-Squared (Test)
Linear Regression	88.76%	88.84%
Lasso Regression	88.76%	88.83%
Ridge Regression	89.78%	89.81%
Decision Tree	95.50%	89.36%
Random Forest	95.23%	91.01%
KNN	94.08%	91.17%
Neural Network	91.23%	90.70%
Gradient Boosting	92.21%	92.14%

2) Finalizing Gradient Boosting: After evaluating multiple regression models for predicting Customer Lifetime Value (CLTV), the Gradient Boosting model had the highest test accuracy of 92.14%. Gradient Boosting even achieved the lowest Mean Squared Error (MSE) compared to all other regression algorithms, with a value of 737.87, and a Mean Absolute Error (MAE) of 1,160,597.10 on the test dataset.

Consequently, we selected it as our preferred algorithm. Gradient Boosting constructs an ensemble of decision trees sequentially, where each subsequent tree corrects the errors made by the previous ones. This method optimizes a differentiable loss function, allowing the model to improve accuracy incrementally as more trees are added.

The superior accuracy of Gradient Boosting on the test dataset can be attributed to its robustness against overfitting, its ability to handle various data irregularities, and its systematic approach to minimizing loss functions. Additionally, it offers critical insights into feature importance, which further refines its predictions. In our analysis, the top three features impacting CLTV were claim_amount (contributing 50%), vintage (28%), and marital_status (6%). This precise identification of influential variables underscores the model's effectiveness in leveraging data characteristics to forecast CLTV accurately.

3) Improving Accuracy of the Selected Model: We conducted extensive hyperparameter tuning on the Gradient Boosting model, testing various parameter combinations to optimize performance. The ideal parameters were determined as follows: learning rate of 0.2, 150 estimators, a maximum tree depth of 5, a minimum of 20 samples required to split an internal node,

and a minimum of 100 samples required in a leaf node. We also implemented subsampling and introduced a stopping condition where the iterative process halts if no improvement is observed after 10 iterations. These adjustments enhanced the model's accuracy, elevating it to 92.49%.

To further validate and stabilize the model's performance, we applied k-fold cross-validation, which slightly improved the test accuracy to 92.53%. This method not only helped in affirming the model's robustness but also ensured that the performance metrics were reliable and indicative of the model's ability to generalize across different subsets of data.

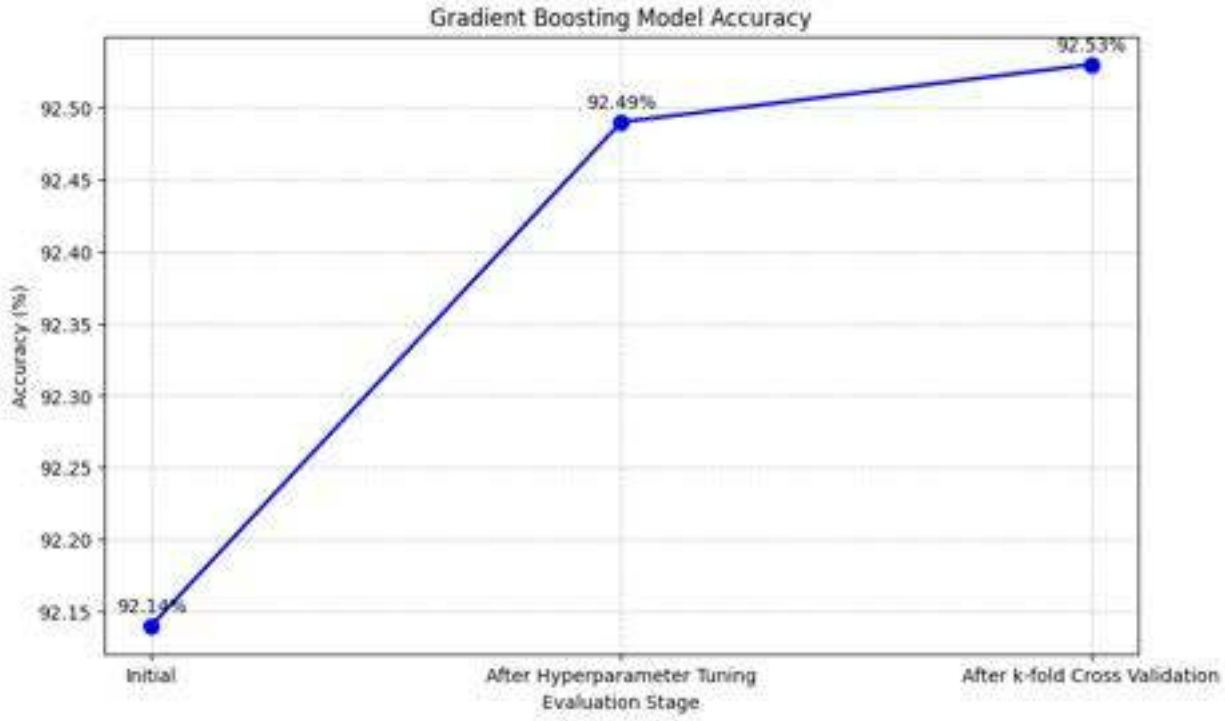


Fig. 3: Accuracy improvements in Gradient Boosting model after hyperparameter tuning and k-fold cross-validation.

B. Clustering

Clustering is a pivotal method in data analysis, enabling the grouping of data points into subsets or clusters. This analysis employs Mini Batch K-Means, an adaptation of the K-Means algorithm designed for large datasets, and the BIRCH algorithms. The primary objective was to identify distinct customer groups within the data that share similar characteristics without prior labeling.

1) Clustering Methodology: Mini Batch K-Means and BIRCH were implemented to segment the dataset. The Mini Batch K-Means algorithm was chosen for its batch processing capability, which is ideal for large data volumes, whereas BIRCH is known for its hierarchical clustering approach that incrementally and dynamically clusters incoming multi-dimensional metric

data. The elbow method was used to determine the optimal number of clusters by analyzing the SSE curve. Cluster validation was assessed using the SSE scores and the Davies-Bouldin Index, which measures the average similarity between clusters.

2) Finding the Optimal Cluster Count: The elbow method is the most commonly used method to determine the optimal number of clusters. The elbow method looks at the Sum Squared Error (SSE) explained as a function of the number of clusters [11].

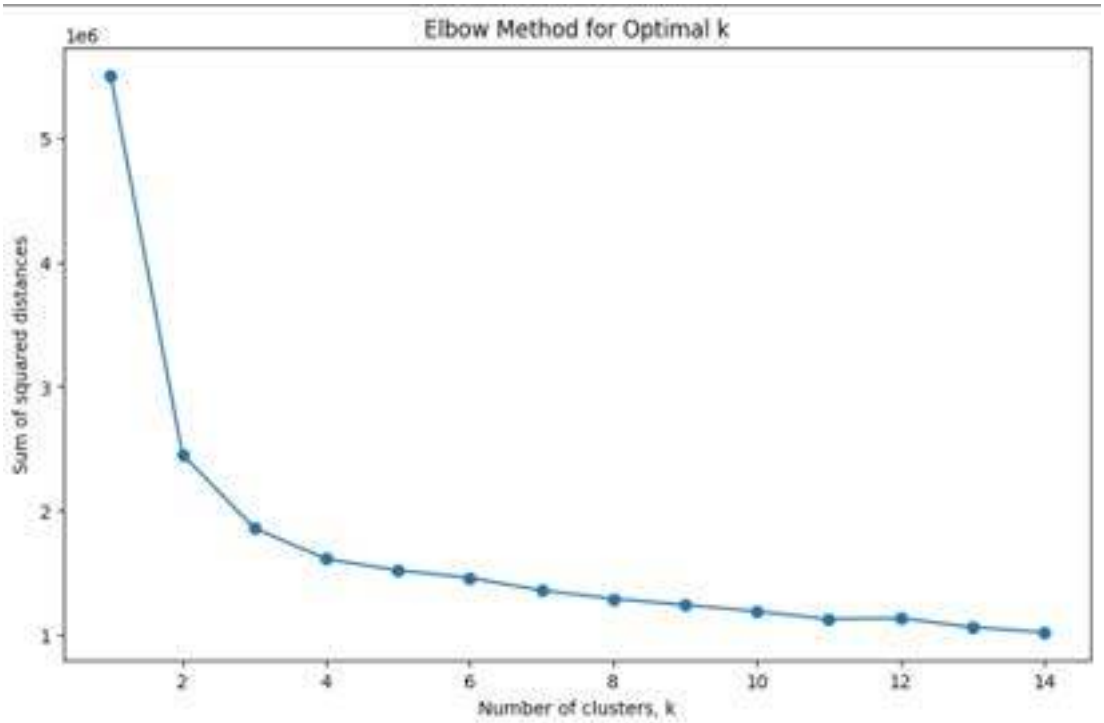


Fig. 4: Elbow Method for determining the optimal number of clusters, k.

The graph in Figure 4 clearly indicates that the sum of squared distances decreases significantly until $k = 3$, after which the rate of decrease plateaus, suggesting that the optimal number of clusters for our dataset is 3.

TABLE II: Results from Mini Batch K-Means

Metric	**Value**
Mean SSE per cluster	[2.13479896, 1.58609141, 1.59215926]
Mean SSE across all clusters	1.7710165441596493
Davies Bouldin Index	1.0549286631945143
Sum of Squared Error (Total Inertia)	1.8236405024953715

3) Results Obtained: The clustering results summarized in Table II include Mean SSE per cluster values [2.13479896, 1.58609141, 1.59215926], indicating varying levels of compactness, with the first cluster being less tight compared to the others. The overall Mean SSE across all clusters is 1.7710165441596493, suggesting reasonable compactness in the clustering

setup. The Davies-Bouldin Index at 1.0549286631945143 points to moderate cluster separation, indicating that while clusters are distinct, there might be some overlap. The total inertia value of 1.8236405024953715 further reflects the overall fit of the data to the clusters, with potential room for improvement in clustering configurations to achieve better separation and tighter clusters. These metrics collectively evaluate the effectiveness of the clustering solution, providing insights into areas where adjustments may enhance results.

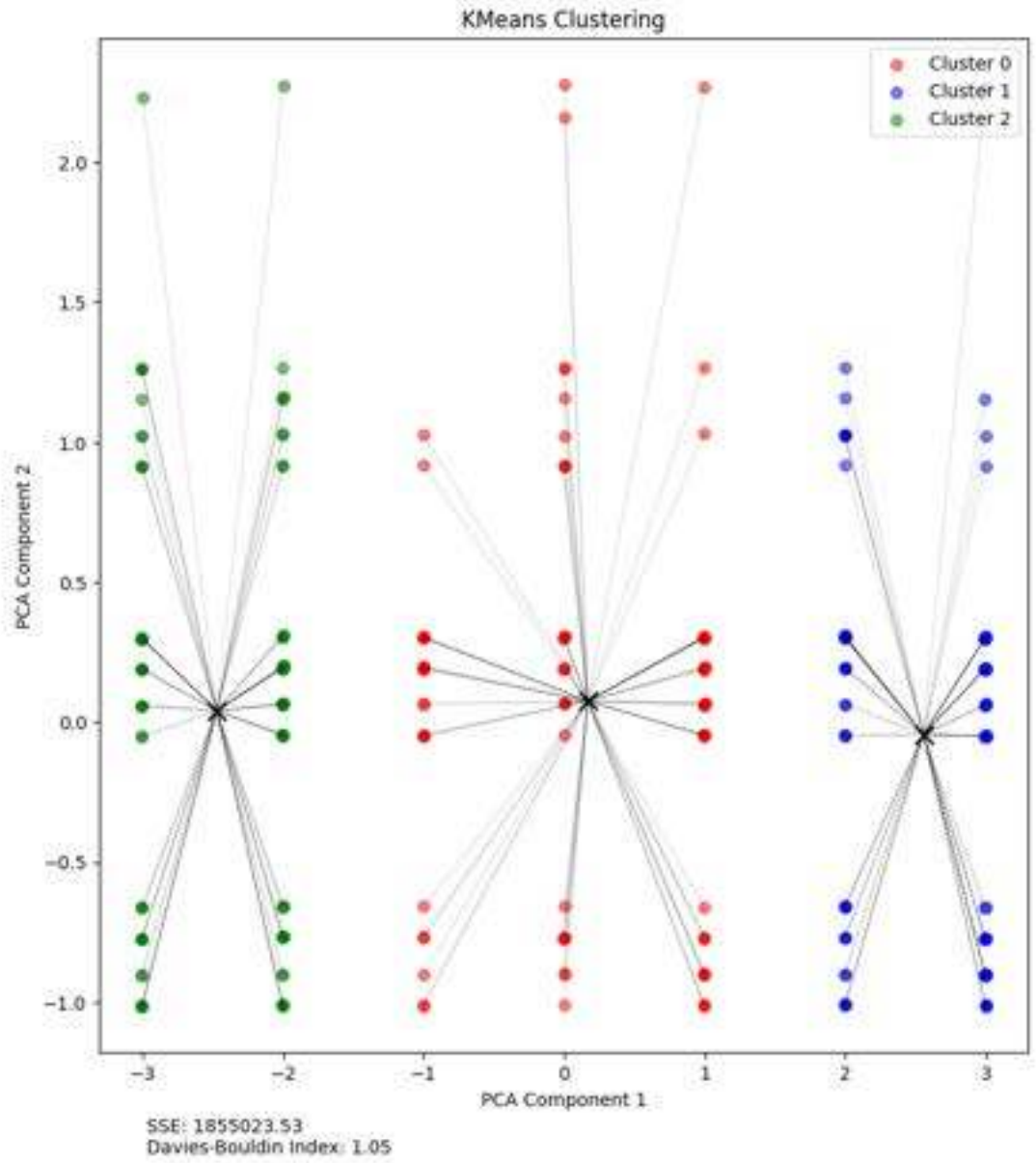


Fig. 5: KMeans clustering of multidimensional data projected onto the first two principal components.

Figure 5 is essentially a scatter plot that displays three distinct clusters identified by the algorithm, each marked with a different color (green for Cluster 0, red for Cluster 1, and blue for Cluster 2). The cluster centroids are denoted by black

crosses. The performance of the clustering is quantified by the within-cluster sum of squares (SSE) and the Davies-Bouldin index, suggesting a moderate separation between the clusters.

TABLE III: Results from BIRCH Clustering

Metric	**Value**
Mean SSE per cluster	[2.67846579, 1.94117697, 1.35569793]
Mean SSE across all clusters	2.347576768518853
Davies-Bouldin Index	1.2289438370359893
Sum of Squared Error (Total Inertia)	2.347576768518853

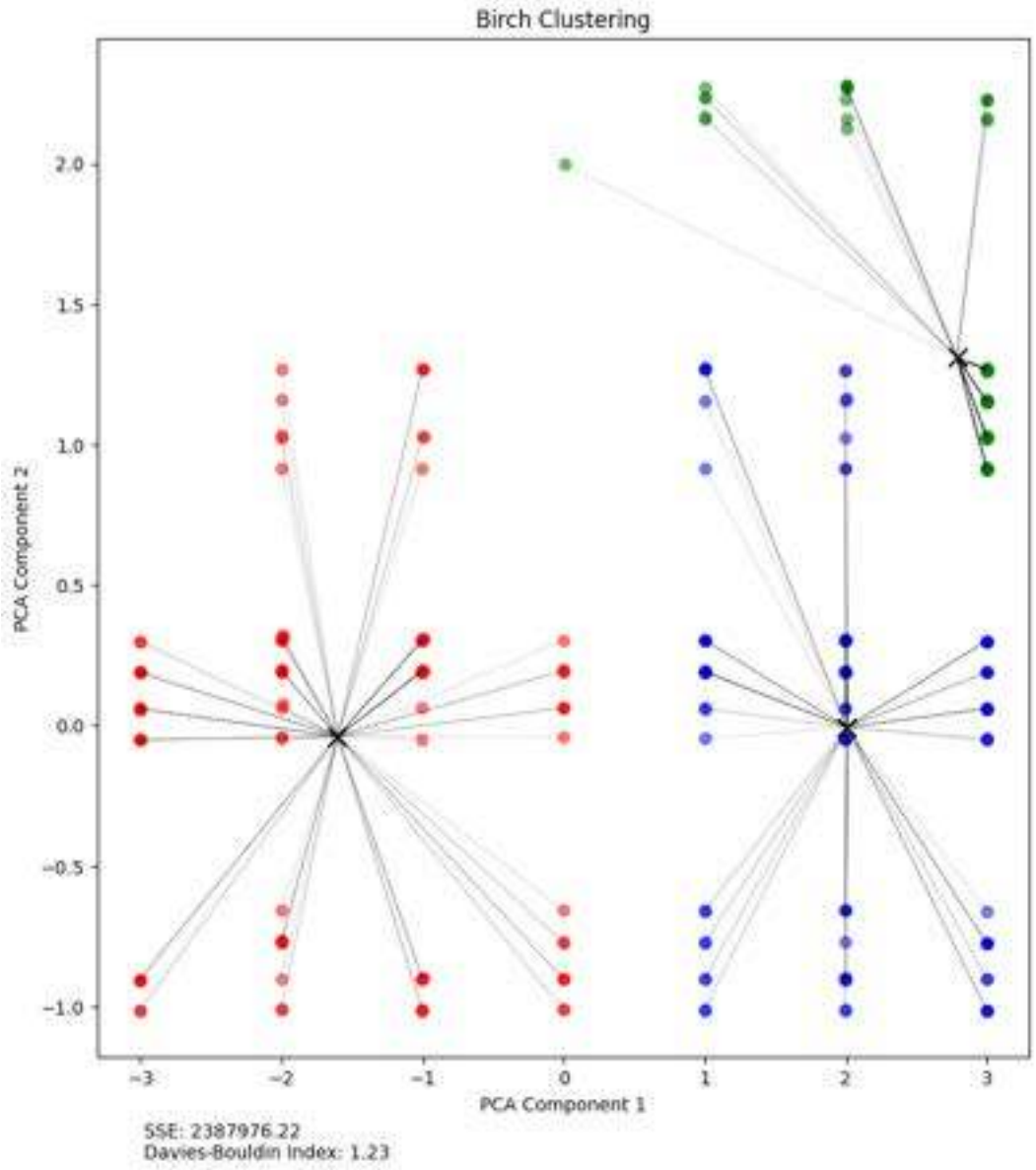


Fig. 6: BIRCH clustering of multidimensional data projected onto the first two principal components.

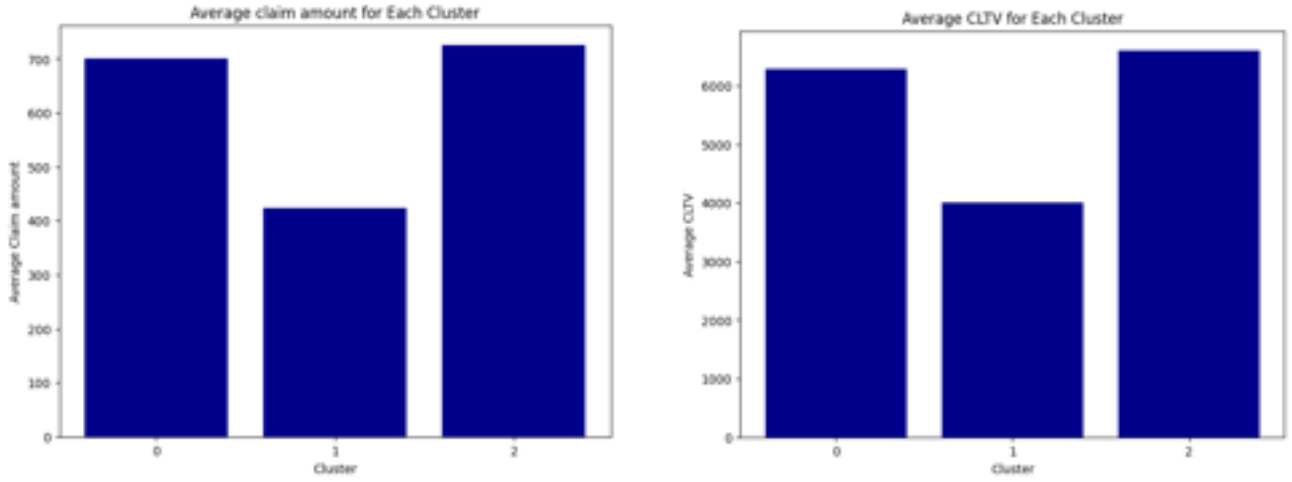
The BIRCH clustering results in Table III showcase Mean SSE per cluster values [2.67846579, 1.94117697, 1.35569793],

reflecting different degrees of compactness within the clusters, with the third cluster demonstrating the greatest compactness. The variability among the clusters suggests that BIRCH's hierarchical clustering approach, which incrementally and dynamically adjusts clusters, is capturing diverse density patterns across the dataset. The overall Mean SSE across all clusters of 2.347576768518853 indicates a moderate level of compactness, yet the Davies-Bouldin Index of 1.2289438370359893 suggests that there is some overlap and less distinct separation between clusters. This index points to potential areas for refining BIRCH's threshold and branching factor settings to optimize cluster purity and separation. The Total Inertia, mirroring the sum of squared errors, at 2.347576768518853, aligns with these findings and implies that the clusters, while reasonably effective, may benefit from tuning to tighten cluster cohesion and improve the overall clustering configuration. These insights highlight the specific attributes and performance of BIRCH, guiding potential adjustments to enhance clustering outcomes.

4) Hyperparameter Tuning: Extensive hyperparameter tuning in both K-Means and BIRCH algorithms showed that core metrics such as the Davies-Bouldin Index and Sum of Squared Error did not change significantly. This consistency suggests that while our adjustments did not drastically improve performance, they did preserve the existing quality of clustering outcomes.

5) Model Selection: KMeans was the chosen algorithm for clustering due to its superior performance, as evidenced by lower SSE values compared to BIRCH, indicating more cohesive clusters. This effectiveness was crucial in the EDA, allowing for a nuanced understanding of customer segments by analyzing characteristics like policy type, area, and marital status. For example, Cluster 2's urban majority pointed to tailored urban policy opportunities. Cluster 0 emerged as a high-value segment due to its higher claim amounts and CLTV, suggesting potential for premium offerings. The KMeans algorithm facilitated this clear segmentation, offering actionable insights for strategic initiatives. Meanwhile, BIRCH's higher SSE suggested less distinct clusters, reaffirming KMeans as the more apt choice for defining clear customer groups and driving focused, data-informed decisions for service personalization.

6) Insights After Clustering: The comparison of clusters is shown in Figure 7, where the Figure 7a indicates that Cluster 1 has the lowest average claim amount, and Figure 7b demonstrates that the same Cluster 1 has the lowest average Customer Lifetime Value (CLTV), with Clusters 0 and 2 being similar for both metrics.

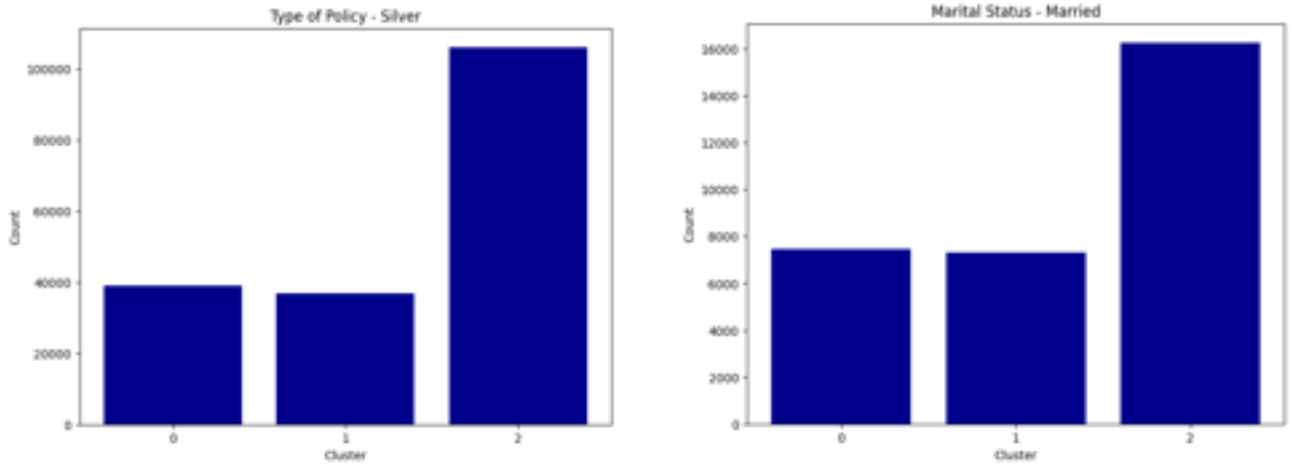


(a) Average claim amount across clusters

(b) Average CLTV across clusters

Fig. 7: Comparison of claim amounts and CLTV across clusters.

As illustrated in Figure 8, Cluster 2 predominates in the number of silver policies and has the highest count of married customers compared to Clusters 0 and 1.

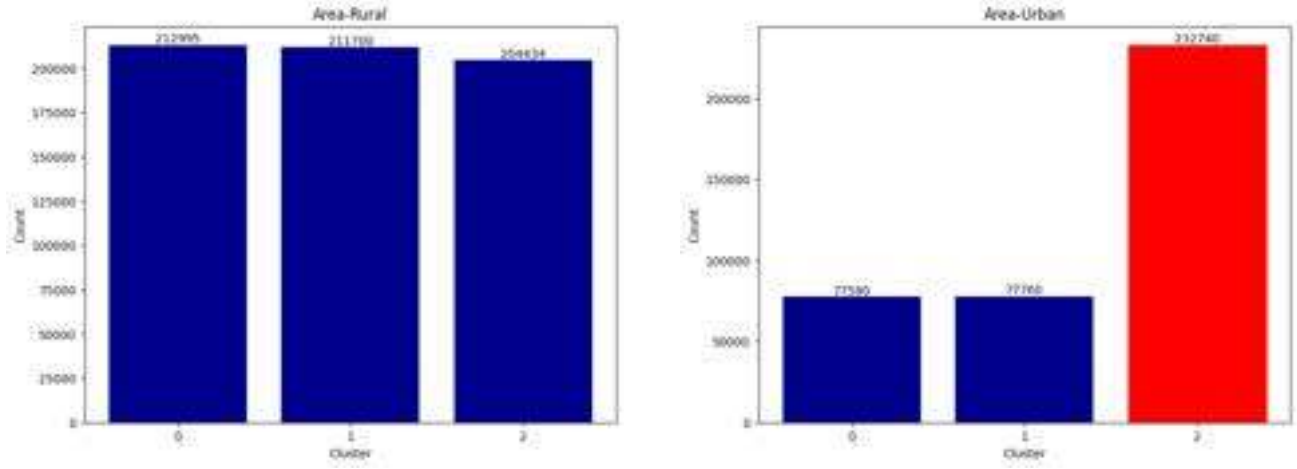


(a) Type of Policy - Silver

(b) Marital Status - Married

Fig. 8: Distribution of silver policies and marital status across clusters.

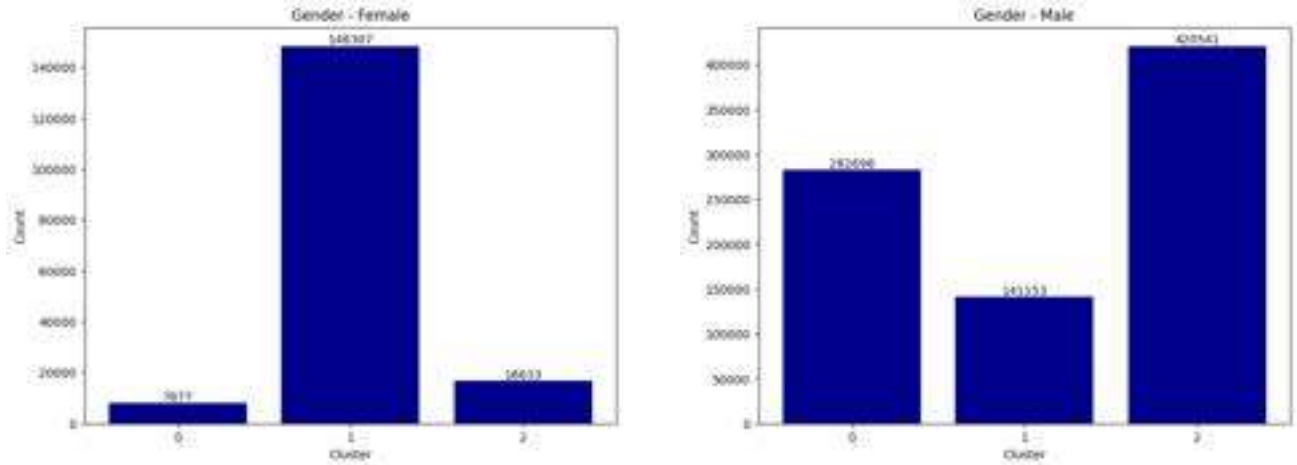
Figure 9 shows that while the distribution of rural customers is relatively balanced across clusters, urban customers are significantly more prevalent in Cluster 2.



(a) Area - Rural (b) Area - Urban

Fig. 9: Rural versus urban customer distribution across clusters.

The gender distribution across clusters is depicted in Figure 10, with Cluster 1 having the highest number of female customers and Cluster 2 the highest number of male customers.



(a) Gender - Female (b) Gender - Male

Fig. 10: Gender distribution across clusters.

Lastly, Figure 11 illustrates the distribution of vintage values across clusters, indicating specific vintage groups associated with each cluster.

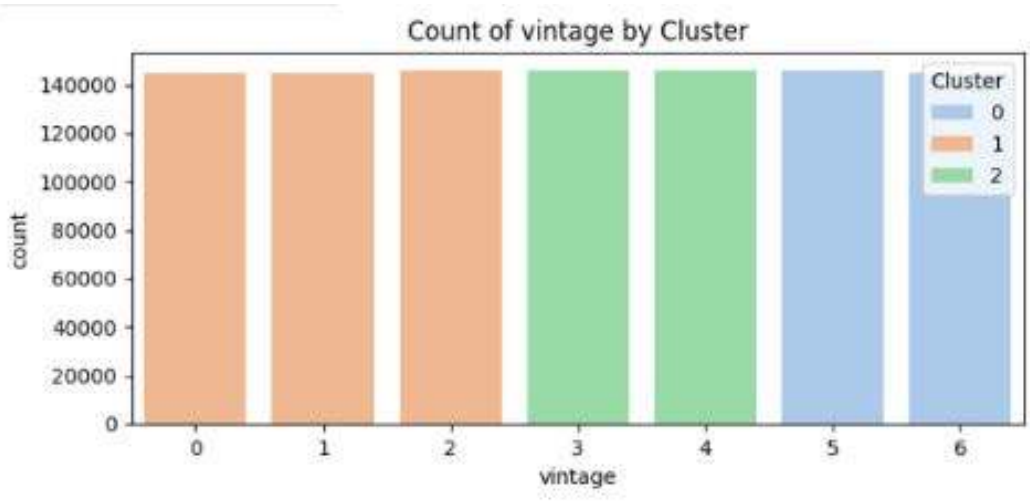


Fig. 11: Vintage value distribution across clusters.

VI. Related Work

The application of predictive analytics for Customer Lifetime Value (CLTV) and customer segmentation has transformed strategic operations in various industries, including banking, e-commerce, and OTT services. This shift has resulted in increased profit margins through the optimization of premium structures and enhanced retention efforts targeting high-value customers. Additionally, segmenting customers based on predictive CLTV models has enabled the development of tailored marketing strategies finely tuned to meet customer needs and maximize potential value.

Recent advancements in insurance analytics have highlighted the integration of machine learning with big data technologies to further enhance service personalization. Innovations in deep learning have paved the way for more accurate and dynamic prediction models. For example, neural networks have been employed in predicting CLTV, offering models capable of adapting to new behaviors and trends in customer data more effectively than traditional methods.

Artificial Intelligence (AI) techniques are increasingly being integrated with Geographic Information Systems (GIS) to improve customer segmentation. This combination facilitates precise geographical analysis of risk factors, leading to highly customized insurance policy frameworks. Such GIS-enabled segmentation supports more granular risk assessment models, which are particularly beneficial in regions prone to specific types of hazards, such as floods or earthquakes.

A. Implications and Future Work

Looking forward, the integration of real-time data analytics and the exploration of alternative machine learning models, such as deep neural networks, could further refine CLTV predictions and customer segmentation processes. Additionally, expanding

the data sources to include more diverse demographic and behavioral data could unveil deeper insights into customer preferences and insurance needs. To effectively communicate these complex findings and make them accessible to decision-makers, we can utilize visualization tools such as Tableau or Power BI to create dynamic dashboards.

In future iterations of our project, embracing big data technologies such as Apache Hadoop could significantly enhance our data handling capabilities. Currently, the processing and implementation of our machine learning models are constrained by the voluminous nature of the data, leading to prolonged execution times. By integrating Hadoop, we can leverage its robust ecosystem for parallel data processing, which allows for the efficient handling of large datasets spread across distributed environments.

In conclusion, the ”CLTV Predictive Segmentation for Personalized Insurance Policies” project not only meets the current needs of Vahan Bima but also provides a scalable model that other companies in the insurance industry can adopt. As we move forward, continuous innovation and adaptation in methodologies will be key to sustaining competitive advantage in the rapidly evolving insurance market.

VII. Conclusion

Our project has demonstrated the powerful capability of advanced predictive analytics in transforming the strategic operations of Vahan Bima, a leader in the motor vehicle insurance sector. By employing Gradient Boosting algorithms for CLTV prediction and utilizing sophisticated clustering techniques like Mini Batch K-Means and BIRCH, we have not only enhanced the accuracy of customer lifetime value forecasts but also significantly improved the personalization of insurance policy offerings.

The successful prediction of CLTV will allow Vahan Bima to more effectively identify and target high-value customers, optimizing resource allocation and potentially increasing profit margins. Furthermore, the segmentation of customers into meaningful groups will enable the development of tailored marketing and service strategies that cater specifically to the nuanced needs of different customer segments. This approach not only bolsters customer satisfaction and retention but also positions Vahan Bima to respond more dynamically to market changes.

REFERENCES

[1] Ryals, L. J., & Knox, S. (2005). Measuring risk-adjusted customer lifetime value and its impact on relationship marketing strategies and shareholder value. *European Journal of Marketing*, *39*(5/6), 456-472.

[2] Paposa, S., Ukinkar, V., & Paposa, K. (2019). Service quality and customer satisfaction: variation in customer perception across demographic profiles in life insurance industry. *International Journal of Innovative Technology and Exploring Engineering*, *8*(10), 3767-3775. https://doi.org/10.35940/ijitee.j9970.0881019

[3] Singh SK, Chivukula M (2020) A Commentary on the Application of Artificial Intelligence in the Insurance Industry. Trends Artif Intell 4(1):75-79.

[4] Hasheminejad, Seyed Mohammad Hossein and Khorrami, Mojgan. 'Clustering of Bank Customers Based on Lifetime Value Using Data Mining Methods'. 1 Jan. 2020 : 507 – 515.

[5] Chang, W., Chen, C., & Li, Q. (2012, August 1). Customer Lifetime Value: A Review. Scientific Journal Publishers Limited. https://doi.org/10.2224/sbp.2012.40.7.1057

[6] Hansotia, B., & Wang, J. (1997). Maintaining premium revenue and profits in a competitive environment: A predictive modeling approach. *The Journal of Database Marketing*, 5(2), 105-115.

[7] Ekinci, Y., Uray, N., & Ülengin, F. (2014). A customer lifetime value model for the banking industry: a guide to marketing actions. *European Journal of Marketing*, *48*(3/4), 761-784.

[8] Yoseph, F., Ahamed Hassain Malim, N. H., Heikkilä, M., Brezulianu, A., Geman, O., & Paskhal Rostam, N. A. (2020). The impact of big data market segmentation using data mining and clustering techniques. *Journal of Intelligent & Fuzzy Systems*, *38*(5), 6159-6173.

[9] Abidar, L., Asri, I. E., Zaidouni, D., & Ennouaary, A. (2022). A Data Mining System for Enhancing Profit Growth based on RFM and CLV. In *2022 9th International Conference on Future Internet of Things and Cloud (FiCloud)* (pp. 247-253). Rome, Italy. doi: 10.1109/FiCloud57274.2022.00041

[10] Lam, Sunny. 'The Ensemble of Neural Network and Gradient Boosting for the Prediction of Customer Profitability: A Two-stage Modeling Approach'. 1 Jan. 2018 : 329 – 340.

[11] Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications*, 105(9).

[12] Charvat, K. (Ed.). (2012). Geographic Information System for Environmental Applications. IntechOpen.

[13] Batty, M. (2013). Big Data, Smart Cities and City Planning. Dialogues in Human Geography, 3(3), 274-279.

LINKS TO ALL THE RESOURCES

- Google Drive Folder Link - (contains all code files, datasets, ppt)

  https://drive.google.com/drive/folders/1OR1kh0GaB4U0yWa9s8Fvov3714Li2mqc?usp=sharing
- Regression Python file -

  https://drive.google.com/file/d/1oVW6DLbMGrmJn3bu-Dq_WyML8Im5pdIE/view?usp=sharing
- Dataset used in Regression Python file -

  https://drive.google.com/file/d/1QCvpbf68dP6oVPGh1XnosFRaU4MmOes1/view?usp=sharing
- Clustering Python file -

  https://drive.google.com/file/d/1V588weqqSDkyO7J0RX4kAaJg3HIyZ99R/view?usp=sharing
- Dataset Used in Clustering Python File -

  https://drive.google.com/file/d/1FIGNV4-DR52uDwkHHj_JbbpjAz7uJZvI/view?usp=sharing
- EDA on clusters Python file -

  https://drive.google.com/file/d/1XdC_ZVsfloi7BkqkoqNmO1T4Vt3MQQRV/view?usp=sharing
- Dataset used in EDA on clusters Python file -

  https://drive.google.com/file/d/1XdC_ZVsfloi7BkqkoqNmO1T4Vt3MQQRV/view?usp=sharing