Predictive Analysis of Customer Lifetime Value and Policy preference in the Insurance Industry.

Kushmi Anuththara Chandrasena

Email: v24kucha@du.se

Abstract - In this fast growing insurance industry, Predictive analysis of customer life time value (CLV) and Policy Preference in the insurance industry is a comprehensive analysis aimed at identifying key factors influencing customer value and policy choices. This study helps insurance companies to identify patterns and trends that can leads to the strategic decision making for insurance companies. Mainly this actually helps the companies to optimize their customer relationship management and improve policy retention rates.

Keywords – *CLTV – Customer Relation*

1. Introduction

In this fast growing insurance industry, the market goes beyond merely offering policies and waiting for customers. The common nature of this customer preferences and behaviours needs a more proactive process to understanding and catering to their needs. (Saharon Rosset, 2002) In order to address this challenge, Machine learning (ML) algorithms to gain a deeper insights in to customer life time value (CLV) and policy preferences. By utilizing predictive modelling technological methods, we were able to anticipate customer needs based on their demographic, behavioural attributes and Financial information. (V. Kumar, 2016)

The study focuses on predictive analysis of “Customer Lifetime Value” and “policy preferences”, by using a csv data set that includes diverse customer information such as demographics, financial indicators and policy related variables. The main goal here is to identify patterns and the trends that can inform strategic decision making, enabling insurance companies to improve their product offerings, marketing strategies and customer retention efforts. (Montserrat Guillen, 2008)

1.1 Research Question

How can we predict customer lifetime value and policy preferences in the insurance industry using demographic and financial indicators, and what insights can be derived to inform strategic decision-making for optimizing marketing strategies, product offerings, and customer retention?

1.2 Project Aim

This research build upon the existing literature and Kaggle data set by using machine learning techniques, Also the study aims to:

1. Explore the relationship between customer demographics, financial indicators and insurance-related variables.
2. Predict customer lifetime value based on historical data and customer attributes.

2. Literature Review

This literature review aims to explore the intersection of customer demographics, insurance related variables, financial indicators and providing a foundation for predictive models that can guide strategic decision making.

2.1 Predictive model and Customer life time value.

A.1 Customer Lifetime value models and indicators.

Research conducted by Delafrooz and Farzanfar (2016) highlights the importance of benefit clustering in determining CLV. Their study involved calculating various inputs, including customer loss rates, discount rates, maintenance costs, including learning cost and over four year period (Narges Delafrooz, 2016). Customers were then classified into gold, silver, lead and bronze categories based on profit rates. Similarly, Donkers, Verhoef and Jong (2007) tested multiple models to predict CLV in the insurance sector, finding that even simple models can perform effectively, challenging the assumption that more complex models necessarily provide better predictions (B. Donkers, 2007).

A.2 Factors influencing Policy Preference and Claims behaviour.

The study by Ghale, Karimi and Dinani (2021) identified critical dimensions for measuring CLV in the supplemental health insurance industry. Their research emphasized profitability, customers loyalty and value co-creation as key dimensions with indicators such as customer satisfaction, trust and repurchase intension playing significant

roles (Roohallah Dehghani Ghale, 2021). Braun, Schmeiser and Schreiber (2016) used conjoint analysis to study term life insurance preferences, finding that brand, critical illness cover and underwriting procedures were major factors influencing customer decisions (Alexander Braun, 2016).

2.2 Predictive Techniques and Strategic insights.

B.1 Predictive Models and Techniques.

In their 2023 study, Surti, Shah, Bharti and Gupta employed machine learning techniques such as exploratory data analysis (EDA) and feature selection to predict CLV. They focused on identifying significant variables for claim submission and approval, ultimately enhancing the accuracy of their predictive models (Maitri Surti, 2023). Dai (2022) reviewed various machine learning models for analysing CLV and found that Random Forest models were particularly effective in predicting customer segmentation and behaviour (Dai, 2022).

B.2 Strategic insights for marketing and Customer Retention.

Ekinci et al. (2014) proposed a markovian- based model to guide future marketing decisions, emphasizing the need to predict potential customer value rather than just measuring current value. This approach helps in making more informed strategic decisions (Y. Ekinci, 2014). Additionally, Smirnova and Khanova (2019) discussed the importance of maintaining a reliable customer database for long term value analysis. They recommended the use of CRM systems and analytical platforms to monitor and analyze customer data effectively (A. S. Smirnova, 2019).

3. Methodology.

3.1 Sourcing the datasets.

The data sets used for this study were taken from the Kaggle.com. These datasets provide advanced information on customers in the insurance industry, including features such as customer ID, Gender, Area, Qualification, income, marital status, vintage, claim amount, number of policies and policy type. The data is divided in to two sets: a training set (‘train_LifeTimePrediction2.csv’) and a test set (test_LifeTimePrediction.csv).

3.2 Data Preprocessing.

Upon selecting the datasets, each dataset was imported using Python code as a separate data frame named accordingly. The data frames presented several issues that needed addressing before analysis could proceed. When handling missing Values the rows with missing values were removed from both the training and test datasets to ensure no bias or inaccuracies were introduced into the analysis. After that begins

with the encoding Categorical variables such as gender, area, qualification, income, marital status, number of policies, policy, and policy type were encoded into numerical values using LabelEncoder. Then proceed with the feature Scaling which is numerical features like vintage and claim amount were standardized using StandardScaler to normalize their distributions and improve model performance.

Table 1- Data set Column names & Data Type.

Column	**Dtype**
id	int64
gender	object
area	object
qualification	object
income	object
marital_status	int64
vintage	int64
claim_amount	int64
num_policies	object
policy	object
type_of_policy	object

3.3 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis gets involved visualizing relationships between customer demographics, financial indicators, and insurance-related variables. The Various plots were created to understand these relationships better and including correlation matrices and also distribution plots in the selected study area.

3.4 Clustering Analysis

To better understand customer behaviors the K-Means clustering was used to group customers into distinct clusters based on their behaviors and characteristics. Mainly the optimal number of clusters was determined using the Elbow method. The Principal Component Analysis (PCA) was applied to minimize the dimensionality of the data/ datasets for visualization purposes, making it easier to analyze each cluster separately.

4. Key Findings, Results and Discussions.

4.1 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis gets involved in visualizing relationships between customer demographics, financial indicators, and insurance related variables. In here the various plots were created to understand these relationships better:

Correlation Matrix: Visualized the correlations between different variables.

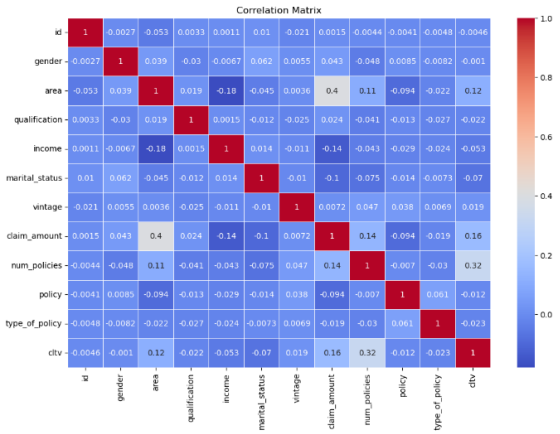

Figure 1: Correlation Matrix Chart

Distribution of Customer Lifetime Value (CLTV): Showed how CLTV is distributed among customers.

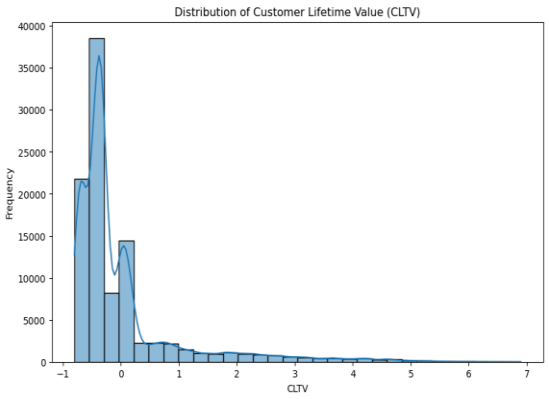

Figure 2: Distribution of CLTV

Distribution of Income Levels: Displayed the distribution of income across different categories.

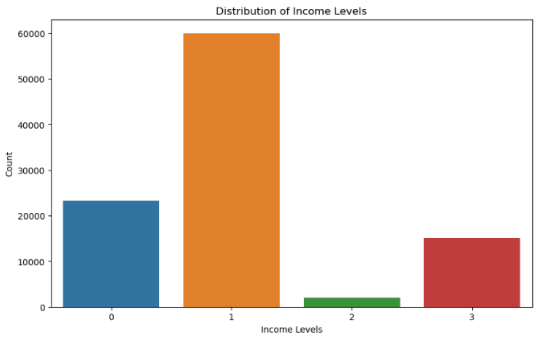

Figure 3: Distribution of Income Levels

Claim Amount Distribution by Area: Illustrated the variation in claim amounts across different areas.

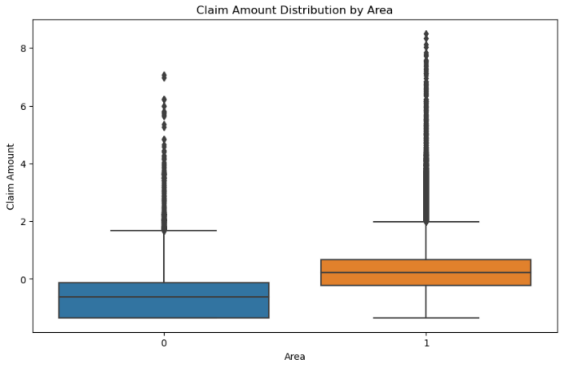

Figure 4: Claim Amount Distribution by Area

Relationship between Vintage and Claim Amount: Highlighted how claim amounts vary with the vintage of the policy.

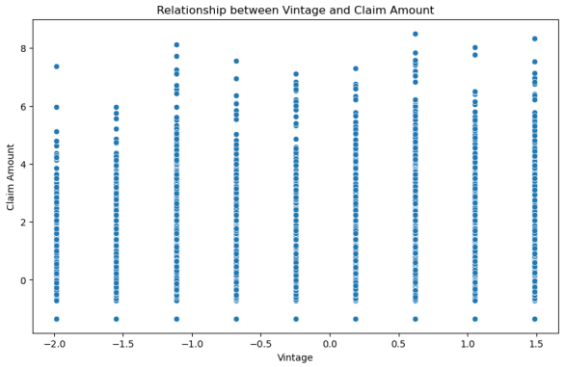

Figure 5: Relationship between Vintage and Claim Amount

4.2 Clustering Analysis

In order to understand customer behaviors and target different customer segments, the clustering techniques were applied. The K-Means clustering was applied to group customers into distinct clusters based on their behaviors and characteristics. The number of clusters was presented using the Elbow method. Also the PCA (Principal Component Analysis) was applied to reduce the dimensionality of the data for visualization purposes, making it easier to analyze each cluster separately.

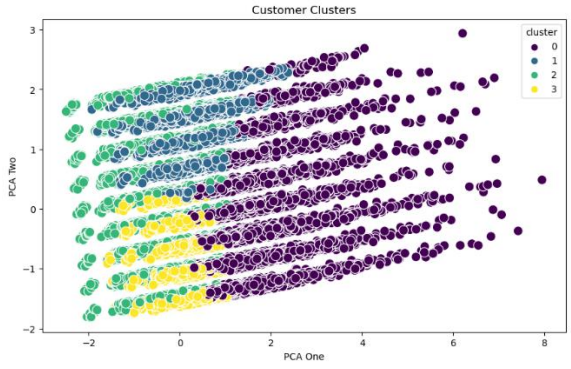

Figure 6: PCA Visualization of Customer Clusters

The K-Means clustering with the help of the Elbow method has correctly placed the customers into separate segments according to their attributes and behaviors, thereby supporting targeted marketing. By reducing the dimensionality of the data to visualize these clusters, it helps link understanding to effective decision-making.

5. Conclusion.

This research aimed to understand customer lifetime value (CLTV) and policy preferences in the insurance industry using predictive analytics. By applying various machine learning models and clustering techniques, we sought to provide actionable insights to help insurance companies optimize their marketing strategies and improve customer retention.

5.1 Summary of Findings

Clustering Analysis: K-Means clustering, validated through the Elbow Method and visualized with Principal Component Analysis (PCA), effectively grouped customers with similar behaviors into distinct clusters.

Data Visualization and Feature Importance: Visualization techniques helped reveal important patterns and relationships in the data, such as the distribution of CLTV, income levels, and claim amounts by area. Feature importance analysis from the Random Forest model highlighted key factors influencing CLTV and policy preferences, offering valuable insights for targeted marketing and customer service strategies.

5.2 Limitations and Future Work

Data Limitations: The dataset used may not fully capture the diversity and complexity of real-world customer behaviors. Future research should include larger and more diverse datasets to improve model robustness.

6. Reference.

A. S. Smirnova, A. K. (2019). INFORMATION TECHNOLOGIES OF ANALYZING CUSTOMER DATABASE OF INSURANCE COMPANY IN TERMS OF LONG-TERM CUSTOMER VALUE.

Alexander Braun, H. S. (2016). On consumer preferences and the willingness to pay for term life insurance.

B. Donkers, P. V. (2007). Modeling CLV: A test of competing models in the insurance industry.

Dai, X. (2022). Customer Lifetime Value Analysis Based on Machine Learning.

Maitri Surti, V. S. (2023). Customer Lifetime Value Prediction of an Insurance Company using Regression Models.

Montserrat Guillen, J. P.-M. (2008). The Need to Monitor Customer Loyalty and Business Risk in the European Insurance Industry.

Narges Delafrooz, E. F. (2016). Determining the Customer Lifetime Value based on the Benefit Clustering in the Insurance Industry.

Roohallah Dehghani Ghale, F. K. (2021). Identifying and Prioritizing Dimensions and Indicators of Customer Lifetime Value for Supplemental Health Insurance Industry.

Saharon Rosset, E. N. (2002). Customer lifetime value modeling and its use for customer retention planning.

V. Kumar, W. R. (2016). Creating Enduring Customer Value.

Y. Ekinci, F. Ü. (2014). Analysis of customer lifetime value and marketing expenditure decisions through a Markovian-based model.