Final Report
Factors Affecting Customer Lifetime Value of an Automobile Insurance Company
BUAN 6312.S01 Econometrics - Group 2

Members
Abhilash Nautiyal
Jalal-U-Din Kadernani

Introduction

In the ever-evolving landscape of the automobile insurance industry, Customer Lifetime Value (CLV) emerges as a pivotal metric. As insurance providers navigate market complexities, the ability to comprehend and leverage the factors influencing CLV becomes paramount.

The auto insurance industry offers a shield against the uncertainties that accompany vehicular ownership. As individuals traverse the roads of life, the need for comprehensive coverage, competitive pricing, and exceptional service from insurance providers becomes paramount. The auto insurance companies aren't merely in the business of underwriting policies; they are custodians of long-term relationships, and architects of brand loyalty. CLV, a metric encapsulating the total revenue from a customer, stands at the intersection of profitability and customer-centricity.

Customer Lifetime Value (CLV) represents the cumulative value a customer generates for a firm throughout their association. It considers their purchasing patterns, loyalty, and overall contribution to the business's revenue. Insurance providers can identify their most valuable customers by modeling the factors that drive CLV and develop strategies to maximize their long-term value. This report encompasses an examination of the various factors affecting CLV in the automobile insurance sector and how important they are.

Academic researchers can utilize the insights gleaned from this study to expand their exploration of diverse dimensions within the field. Marketing and sales teams within companies can capitalize on these findings to tailor their strategies, aligning them with the nuanced factors that are influencing the CLV in this sector. Government agencies and regulatory bodies overseeing the insurance sector stand to benefit from this research, gaining valuable insights into market dynamics and customer behavior. These findings can inform and support evidence-based policy decisions. Companies specializing in Customer Relationship Management (CRM) and data analytics solutions for the insurance industry can leverage these research findings to enhance their product offerings, providing tailored solutions.

The subsequent sections of this study will cover the literature review, data exploration, variable selection, model specification, estimation, and interpretation. By employing econometric techniques, we aim to extract meaningful patterns from the data and contribute valuable insights to the broader discourse on customer-centric business strategies. In conclusion, this study stands at the intersection of theoretical knowledge and practical application, seeking to bridge the gap between academic research and real-world business challenges.

Literature Review

CLVs significance has escalated dramatically in today's rapidly evolving business landscape. Companies, especially in the insurance sector, increasingly recognize CLV's pivotal role in driving strategic decisions, specifically in customer acquisition, retention, and overall business. Research showcases diverse CLV calculation methodologies, including discounted cash flow (DCF) models, cohort analysis, and machine learning approaches. This combination emphasizes the necessity of selecting the most relevant method aligned with specific objectives, whether they lean toward profitability or customer retention ([Gupta et al., 2006], [Mohamadi et al., 2014]).

The impact of customer demographics, policy attributes, and interaction channels on CLV underscores the complexity of influencing factors. Understanding the intricate relationships between demographics (such as age, gender, and income) and policy attributes (like premium amounts, claim values, and coverage) can profoundly inform insurance strategies for customer engagement, service enhancement, and retention ([Dai, 2022], [Fang et al., 2016]).

Consistent correlations between CLV and customer retention reaffirm the strategic significance of retaining high-CLV customers. Insights derived from these associations inform the development of tailored retention strategies, including personalized offers and loyalty programs ([Mohamadi et al., 2014]).

This review combines varied methodologies and impacts of Customer Lifetime Value (CLV) in the insurance industry. These insights pave the way for data-driven strategies that prioritize customer-centric approaches. Leveraging this knowledge can significantly enhance customer satisfaction, refine business strategies, and improve profitability within the insurance landscape.

Existing research on Customer Lifetime Value (CLV) within the insurance sector reveals critical gaps and opportunities. Current studies often isolate factors influencing CLV, neglecting their complicated correlation among customer demographics and policy attributes. This oversight underscores the need for a comprehensive understanding of how these elements interact to impact CLV. Understanding the evolving technological landscape's effects on customer behavior and subsequent CLV could provide valuable insights.

The review contextualizes the project by highlighting the significance of understanding CLV in correlation with its factors within the insurance sector. It underlines CLV's role in strategic decision-making, stresses retaining high-CLV customers, identifies gaps in understanding correlations among demographics and policy attributes, and proposes integrating behavioral economics for enhanced predictive accuracy and strategic effectiveness.

Data

Source of Data:

The dataset utilized in this analysis is sourced from Kaggle : https://www.kaggle.com/datasets/ranja7/vehicle-insurance-customer-data/data, comprising information on customer interactions and transactions. The data spans a range of attributes, offering insights into customer profiles, policies, claims, and various other factors that contribute to the understanding of Customer Lifetime Value (CLV) within the insurance industry. The data set is quite complete with twenty four variables and 9,134 observations per variable.

Variables:

Demographic Information:

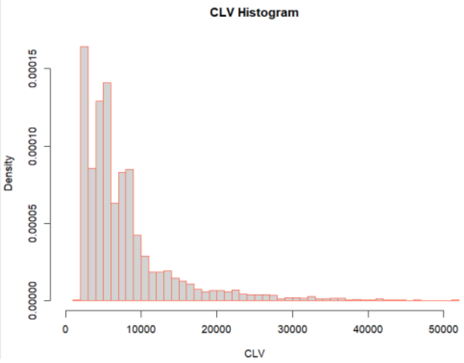
Customer: A unique identifier for each customer.
State: The state in which the customer is located.
EmploymentStatus: The employment status of the customer.
Gender: The gender of the customer.
Income: The reported income of the customer.
Location.Code: The type of location where the customer resides (Suburban, Rural, Urban).
Marital.Status: The marital status of the customer.

Insurance and Policy Details:

Customer.Lifetime.Value: The net present value of future profits expected from a customer.
Response: Whether the customer responded to an offer or not.
Coverage: The type of insurance coverage (Basic, Extended, Premium).
Education: The educational level of the customer.
Effective.To.Date: The effective date of the insurance policy.
Months.Since.Last.Claim: The number of months since the last insurance claim.
Months.Since.Policy.Inception: The number of months since the inception of the insurance policy.
Number.of.Open.Complaints: The number of open complaints filed by the customer.
Number.of.Policies: The number of insurance policies held by the customer.
Policy.Type: The type of insurance policy (Corporate Auto, Personal Auto).
Policy: The specific policy type.
Renew.Offer.Type: The type of renewal offer received by the customer.
Sales.Channel: The channel through which the insurance was sold.
Total.Claim.Amount: The total amount claimed by the customer.
Vehicle.Class: The class of the insured vehicle (Two-Door Car, Four-Door Car, SUV).
Vehicle.Size: The size of the insured vehicle (Small, Medsize, Large).
Monthly.Premium.Auto: The monthly premium for auto insurance.

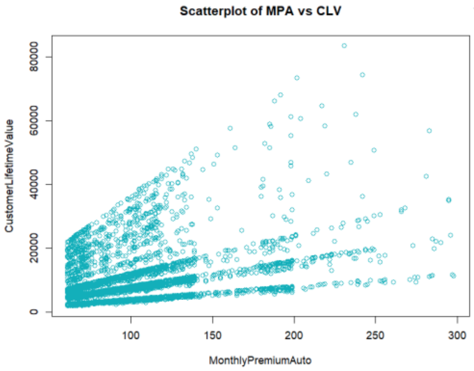
Income is the only variable with potential null values, since it has just over 25% of its values listed as zero. Although, the data also indicates that just over 24% of policy holders are unemployed, thereby confirming the income data to likely be accurate.

Details of the dependent variable, Customer Lifetime Value. (see next page)
Minimum = 1,898.01
Maximum = 83,325.38
Mean = 8,005
Median = 5,780
Standard Deviation = 6,870.97



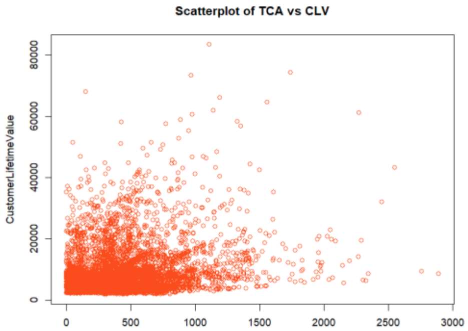
Details of the independent variable, Monthly Premium Auto.

Minimum = 61
Maximum = 298
Mean = 93.22
Median = 83
Standard Deviation = 34.41

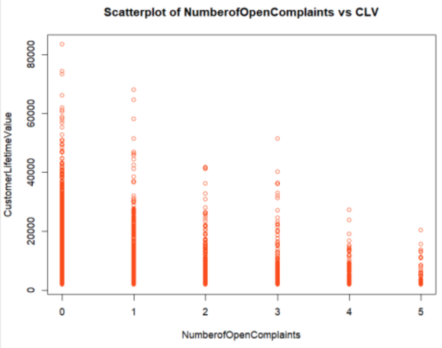


Details of the independent variable, Total Claim Amount.

Minimum = 0.01
Maximum = 2,893.24
Mean = 434.01
Median = 383.95
Standard Deviation = 290.50

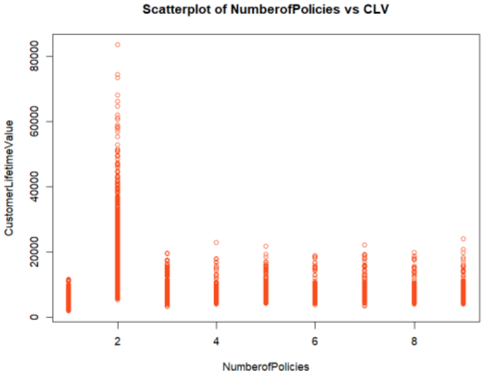


Details of the independent variable, Number of Open Complaints.
Minimum = 0
Maximum = 5
Mean = 0.38
Median = 0
Standard Deviation = 0.91



Details of the independent variable, Number of Policies.

Minimum = 1
Maximum = 9
Mean = 2.97
Median = 2
Standard Deviation = 2.39



Column Name	Min Value	Max Value	Mean	Median	Standard Deviation
Income	0	99,981	37,657	33,890	30,379.9
Months Since Last Claim	0	35	15.1	14.0	10.07
Months Since Policy Inception	0	99	48.1	48.0	27.91

Table 1. Summary of Key Numerical Variables

Table 2. Summary of Key Categorical Variables (variable name is bolded)

Response	Count	Percentage
No	7,826	85.68%
Yes	1,308	14.32%

Coverage	Count	Percentage
Basic	5,568	60.96%
Extended	2,742	30.02%
Premium	824	9.02%

Education	Count	Percentage
Bachelor	2,748	30.09%
College	2,681	29.35%

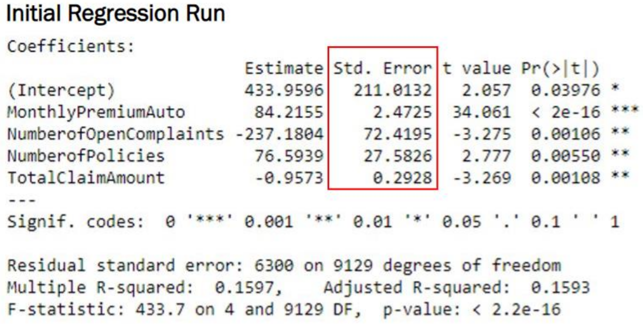
Doctor	342	3.74%
High school or below	2,622	28.71%
Master	741	8.11%

This dataset, rich in customer and policy information, forms the foundation for our exploration into the factors influencing Customer Lifetime Value in the insurance industry. The subsequent sections will delve into the methodologies employed to analyze this data and draw meaningful insights.

Empirical Method

Upon examination of the data, it was decided to use multiple linear regression as the technique to perform an empirical analysis. This is due to the linear relationship found in the data, along with several significant variables being found in the initial regression run.

After removing non-significant independent variables, the following regression was obtained.



Initial Regression Run

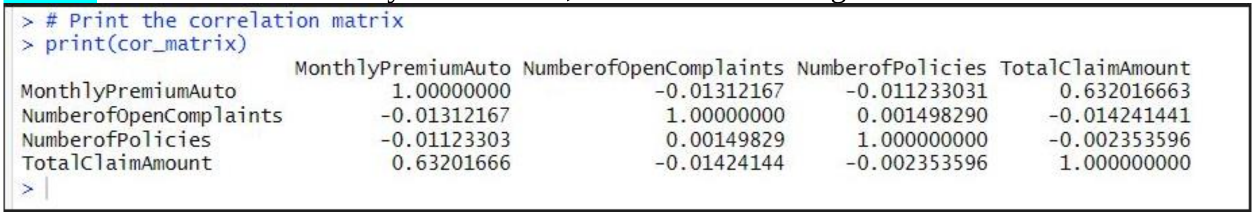
```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              433.9596   211.0132   2.057  0.03976 *  
MonthlyPremiumAuto        84.2155     2.4725  34.061  < 2e-16 ***
NumberofOpenComplaints -237.1804    72.4195  -3.275  0.00106 ** 
NumberofPolicies          76.5939    27.5826   2.777  0.00550 ** 
TotalClaimAmount          -0.9573     0.2928  -3.269  0.00108 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6300 on 9129 degrees of freedom
Multiple R-squared:  0.1597,	Adjusted R-squared:  0.1593 
F-statistic: 433.7 on 4 and 9129 DF,  p-value: < 2.2e-16
```

Check the OLS assumptions

MLR.1 Linear in Parameters - is satisfied by the relationship of the independent variables to the dependent variable.

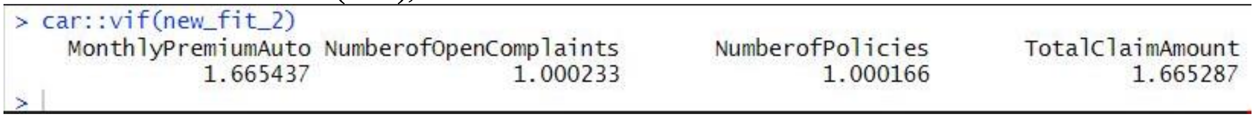
MLR.2 Random Sampling - is satisfied by virtue of the data source and its cross-sectional nature.

MLR.3 No Perfect Collinearity - is satisfied, it can be seen through the correlation matrix below



```
> # Print the correlation matrix
> print(cor_matrix)
                       MonthlyPremiumAuto NumberofOpenComplaints NumberofPolicies TotalClaimAmount
MonthlyPremiumAuto             1.00000000            -0.01312167     -0.011233031      0.632016663
NumberofOpenComplaints        -0.01312167             1.00000000      0.001498290     -0.014241441
NumberofPolicies              -0.01123303             0.00149829      1.000000000     -0.002353596
TotalClaimAmount               0.63201666            -0.01424144     -0.002353596      1.000000000
> 
```

Variance Inflation Factor (VIF), all values are less than 5.

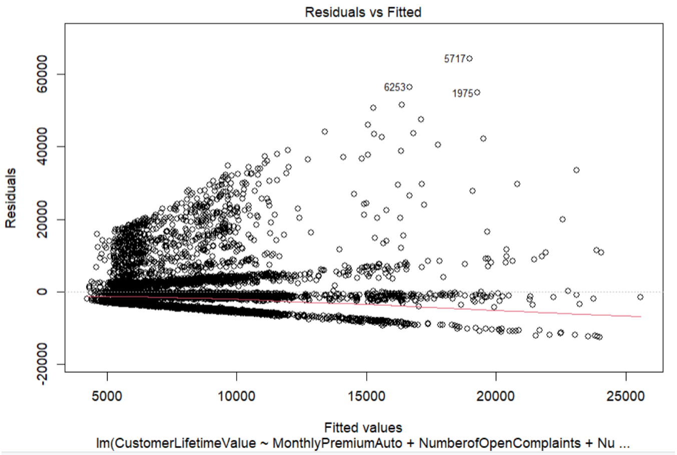


```
> car::vif(new_fit_2)
    MonthlyPremiumAuto NumberofOpenComplaints       NumberofPolicies       TotalClaimAmount 
              1.665437               1.000233               1.000166               1.665287 
> 
```

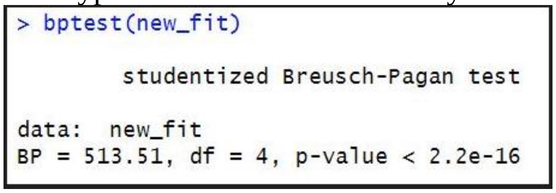
MLR.4 Zero Conditional Mean - is reasonably satisfied.

This scatterplot shows that outliers produce residuals with fitted values not around zero.

With assumptions MLR.1 through MLR.4 satisfied, theorem 1 applies. Therefore the OLS estimators can be said to be unbiased.

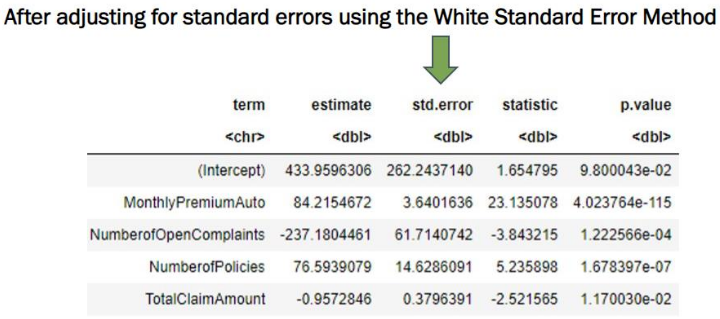


MLR.5 Homoskedasticity - is not satisfied due to heteroskedasticity in the error terms. The Braush-Pagan (BP) test result below shows a very small p-value, therefore the null hypothesis of homoscedasticity of the error terms is rejected.



```
> bptest(new_fit)

        studentized Breusch-Pagan test

data:  new_fit
BP = 513.51, df = 4, p-value < 2.2e-16
```

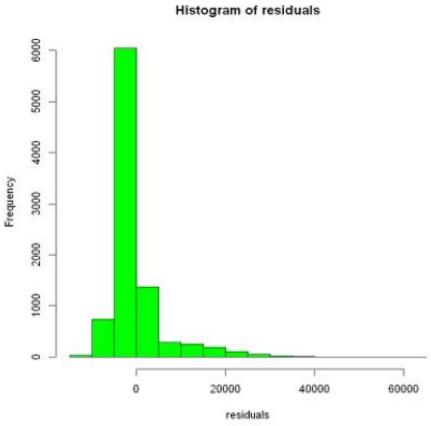
To keep the test statistics robust against heteroskedasticity, the standard errors were adjusted using the heteroskedasticity robust standard errors method, otherwise known as the White Standard Errors Method (Economist Halbert White).

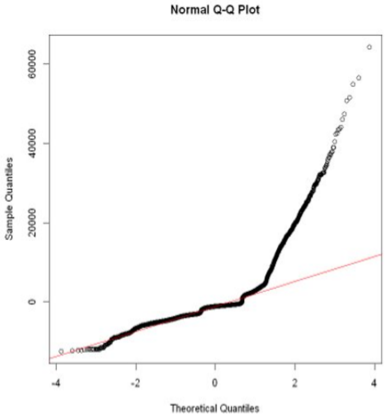


After adjusting for standard errors using the White Standard Error Method

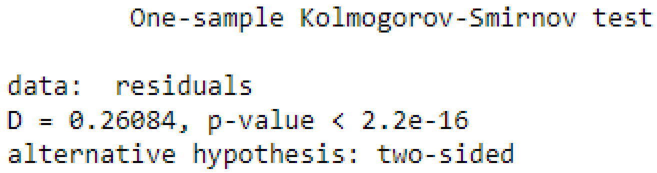
term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	433.9596306	262.2437140	1.654795	9.800043e-02
MonthlyPremiumAuto	84.2154672	3.6401636	23.135078	4.023764e-115
NumberofOpenComplaints	-237.1804461	61.7140742	-3.843215	1.222566e-04
NumberofPolicies	76.5939079	14.6286091	5.235898	1.678397e-07
TotalClaimAmount	-0.9572846	0.3796391	-2.521565	1.170030e-02

MLR.6 Normality - is not satisfied due to the non-normality of the residuals. This can be seen in the following histogram and Q-Q plot of the residuals as well as Kolmogorov-Smirnov Test.





Null hypothesis (Residuals have Normality) is rejected as per K-S Test, shown below.



```
        One-sample Kolmogorov-Smirnov test

data:  residuals
D = 0.26084, p-value < 2.2e-16
alternative hypothesis: two-sided
```

The final equation yielded by the aforementioned empirical analysis is shown below.
Final Estimated Equation:

(CLV)ˆ= 433.96 + 84.22 MPA – 237.18 NoOC + 76.59NoP – 0.96 TCA
(262.24)   (3.64)   (61.71)   (14.62)   (0.38)
n = 9134,   R-squared = 0.1597,   Adjusted R-squared = 0.1593.

MPA = Monthly Premium Auto
NoOC = Number of Open Complaints
NoP = Number of Policies
TCA = Total Claim Amount

The large sample size of the dataset (9,134 records), brings the CLT into effect and reduces the importance of satisfying MLR.6. One method to achieve normality of the residuals would be to remove outliers. This would be to approximately remove 10% (1,035) records of the dataset.

Results

Multiple Linear Regression Analysis:

Multiple linear regression (MLR) was employed to explore the intricate relationship between Customer Lifetime Value (CLV) and various factors, utilizing a dataset sourced from Kaggle. With 24 variables and 9,134 observations per variable, the thorough analysis yielded a regression equation:

$$CLV=433.96+84.22MPA-237.18NoOC+76.59NoP-0.96TCA$$

Statistical Inference:

In all instances, the p-values were remarkably small (p < 0.001), providing robust evidence against the null hypothesis. This high level of statistical significance supports the conclusion that meaningful and substantial effects exist within our data.As such, the null hypothesis is rejected in favor of the alternative hypothesis at a 1% significance level, affirming the validity of our findings.

Economic Inference- Interpretation of Coefficients:

- Monthly Premium Auto (MPA):
  - A one-unit increase in MPA will increase CLV by 84.22.
  - Companies offering competitive monthly premiums tend to foster customer loyalty by providing cost-effective and value-driven insurance solutions.

- Number of Open Complaints (NoOC):
  - A one-unit increase in NoOC will significantly decrease CLV by 237.18.Companies with fewer customer complaints are more likely to maintain loyalty, reflecting a commitment to service quality and customer satisfaction in the insurance industry.
- Number of Policies (NoP):
  - A one-unit increase in NoP will increase CLV by 76.59.
  - Companies offering a greater variety of policy types will attract customers for a longer period.
- Total Claim Amount (TCA):
  - A one-unit increase in TCA will decrease CLV by 0.96.
  - Companies with lower total claim amounts are more likely to retain customer loyalty, signaling both cost-effectiveness and efficient risk management.

Conclusion:

The R-squared value of 0.1597 indicates that 15.97% of CLV variance is explained by our model. The adjusted R-squared (0.1593) shows modest improvement with added predictors, emphasizing the complexity of predicting insurance dynamics. The small difference between R-squared and adjusted R-squared suggests strong predictability. Future studies could explore advanced machine learning techniques like random forest and xgboost to enhance model robustness and handle outliers effectively.

In conclusion, adopting a targeted customer approach and strategic operational adjustments are recommended. Focusing on demographics with higher CLV potential, addressing customer concerns promptly, and optimizing the role of agents in sales can lead to a substantial increase in overall CLV. These insights aim to guide XYZ Insurance Company towards informed decision-making, fostering strategies for maximizing customer value and ensuring long-term profitability.

Future Study

- Advanced Predictive Modeling:
  - Employ advanced machine learning techniques, such as random forest, Xgboost, or neural networks, to enhance the precision of CLV predictions.
  - These methods excel in handling intricate relationships and non-linear patterns within the data.
- Temporal Analysis:
  - Conduct a temporal analysis to track the evolution of CLV over time.
  - Examine seasonal trends, economic fluctuations, and shifts in consumer behavior that may influence long-term customer value.
- Customer Behavior Segmentation:
  - Classify customers based on their behavior patterns and analyze CLV within each segment.
  - This approach offers detailed insights into how different customer groups contribute uniquely to the overall CLV.

- Cross-Channel Analysis:
    - Investigate the impact of diverse marketing and sales channels on CLV.
    - Understanding how customer interactions across various channels shape long-term value can guide tailored optimization strategies for each channel.
- Customer Satisfaction Surveys:
    - Incorporate customer satisfaction surveys to gather qualitative data.
    - Analyze the correlation between customer satisfaction, loyalty, and CLV, providing a comprehensive understanding of the factors underpinning long-term value.

References:

Dai, X. (2022). Customer Lifetime Value Analysis Based on Machine Learning. www.ncbi.nlm.nih.gov/pmc/articles/PMC9958434/

Mohamadi, V. D., Albadvi, A., & Teymorpur, B. (2014). Predicting Customer Churn Using CLV in the Insurance Industry. sjsm.shiraz.iau.ir/article_519605.html

Gupta, S., Hardie, B., Kahn, W., Kumar, V., Lin, N., & Ravishanker, N. (2006). Modeling Customer Lifetime Value. www.researchgate.net/publication/237287176_Modeling_Customer_Lifetime_Value

Scriney, M., & Roantrey, M. (2020). Predicting Customer Churn for Insurance Data. www.researchgate.net/publication/344269094_Predicting_Customer_Churn_for_Insurance_Data

Fang, K., Jiang, Y., & Song, M. (2016). Customer Profitability Forecasting using Big Data Analytics: A Case Study of the Insurance Industry. www.semanticscholar.org/paper/Customer-profitability-forecasting-using-Big-Data-A-Fang-Jiang/ddc1a1cba069581bf121cbbc37750bcf2d61346a

Tang, Z., & Chen, M. (2020). Research on Influencing Factors of Auto Insurance Premium under the Background of Marketization Reform—Empirical Analysis Based on VAR Model. https://www.scirp.org/journal/paperinformation?paperid=101787

Hanafy, M., & Ming, R. (2021). Machine Learning Approaches for Auto Insurance Big Data. https://www.mdpi.com/2227-9091/9/2/42