Brandenburgische
Technische Universität
Cottbus

Exploratory Data Analysis (EDA) and Predicting Customer Lifetime Value (CLV) for an Auto Insurance Company

Danial Monachan
Msc. Artificial Intelligence
Matri-Nr.: 5003811

28 January 2025

Data Exploration and System Management Using Artificial Intelligence/Machine Learning

Lecturer: Prof. Dr.- Ireneusz Jablonski

1. Introduction

Customer Lifetime Value (CLV) measures the total revenue a business expects from a customer over their entire relationship. Accurate prediction of CLV allows businesses to design targeted strategies for customer retention, upselling, and acquisition. This report presents a systematic approach to predicting CLV for an auto insurance company using regression models and a rich dataset of customer attributes.

---

2. Data Overview

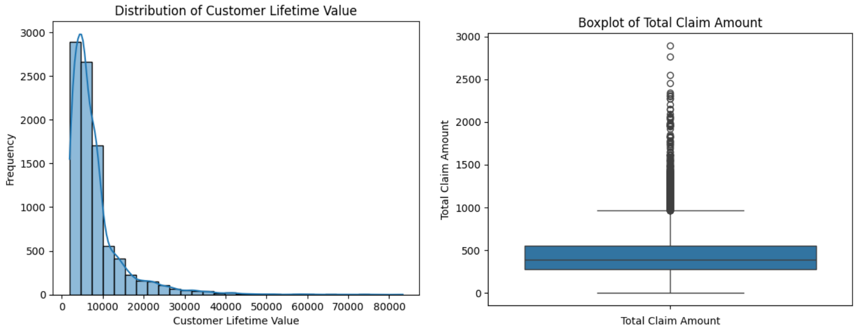
Dataset Description:

- **Rows:** 9,134
- **Columns:** 24
- **Variables:**

Numerical: ['Customer Lifetime Value', 'Income', 'Monthly Premium Auto', 'Months Since Last Claim', 'Months Since Policy Inception', 'Number of Open Complaints', 'Number of Policies', 'Total Claim Amount']

Categorical: ['Customer', 'State', 'Response', 'Coverage', 'Education', 'Effective To Date', 'EmploymentStatus', 'Gender', 'Location Code', 'Marital Status', 'Policy Type', 'Policy', 'Renew Offer Type', 'Sales Channel', 'Vehicle Class', 'Vehicle Size'

- **Insights:**
  - **Mean CLV:** 8,004.94 (range: 1,898 to 83,325)
  - Most customers have zero open complaints.
  - **Customer Income Distribution:** About 50% of customers have no income recorded, indicating unemployment or missing data. Higher income correlates positively with CLV.
  - **Monthly Premium Auto:** Average premium is $93.22 (range: $61 to $298). Customers with higher premiums generally have higher CLV.
  - **Policy Type Trends:** Corporate policies account for approximately 40% of total policies but contribute disproportionately to higher CLV.
  - **Vehicle Class and CLV:** SUVs and luxury vehicles are associated with higher CLV compared to sedans or compact cars.
  - **Total Claim Amount:** Average claim amount is $434.09 (range: $0.10 to $2,893.23). Higher claim amounts do not always equate to higher CLV, suggesting a need for efficient claims processing.
  - **Employment Status:** Employed customers exhibit higher CLV compared to unemployed or retired customers.



- The above graph's shows that the data are skewed (left Bar graph) and have considerable number of Outliers (Right Box plot). This gives us a simple overview of how the overall data could look like.

Initial Observations:

- No missing or duplicate values.
- Significant variation in CLV and income suggests the need for normalization.

---

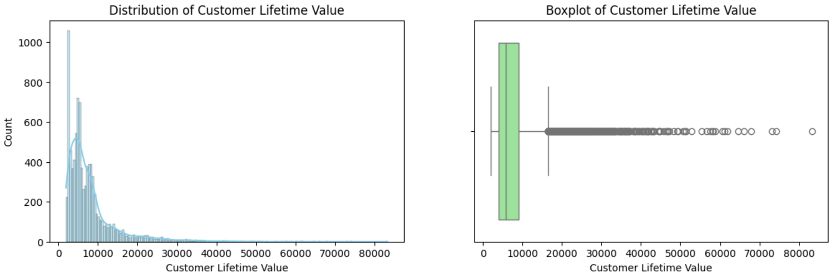
3. Exploratory Data Analysis (EDA)

Key Findings:

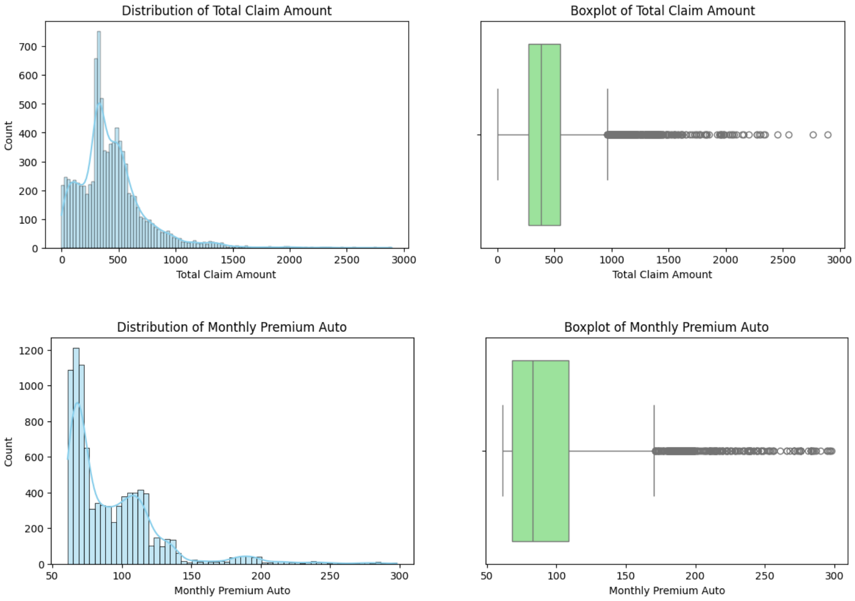
- **Income vs. CLV:** Positive correlation was expected but, indicating higher income often aligns with higher CLV but it was not the case.
- **Policy Type:** No significant difference in CLV between Corporate and Personal Auto policies.
- **Coverage:** Premium coverage should have had higher CLV, but the average was same for the rest coverage

Visualizations:

- Distribution plots for CLV show a right-skewed distribution and Outliers.



- Boxplots reveal outliers in monthly premiums and claim amounts.



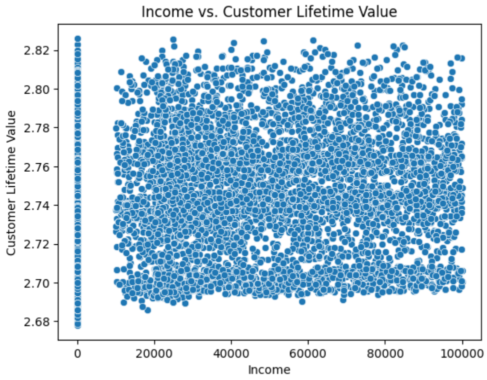
4. Hypotheses Development

Hypothesis Formation Process:

The hypotheses were developed based on the following:

- **Industry Knowledge:** Understanding how customer demographics and policy types impact profitability.
- **EDA Insights:** Observations from correlations, averages, and distribution trends in the dataset.
- **Business Logic:** Revenue generation trends in the insurance sector.

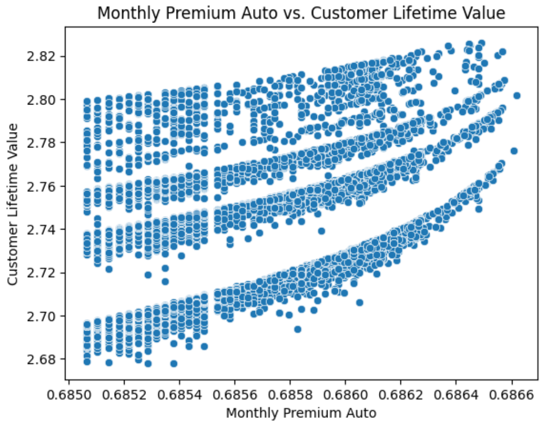
1. **Income Hypothesis:** Customers with higher income tend to have higher Customer Lifetime Value (CLV).
    1. **Null Hypothesis (H0):** There is no correlation between income and CLV.
    2. **Alternative Hypothesis (H1):** There is a positive correlation between income and CLV.



-
- Correlation between Income and CLV: 0.06 (P-value: 2.19e-09)
- The correlation is statistically significant.

2. **Monthly Premium Auto Hypothesis:** Higher monthly premiums are associated with higher CLV.

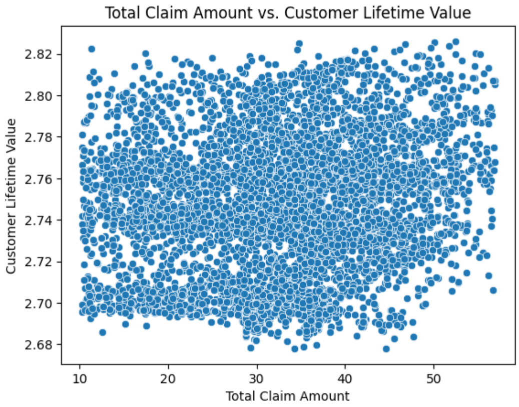
    1. **Null Hypothesis (H0):** Monthly premium does not affect CLV.
    2. **Alternative Hypothesis (H1):** Monthly premium has a positive effect on CLV.



-
- Correlation between Monthly Premium Auto and CLV: 0.45 (P-value: 0.00e+00)
- The correlation is statistically significant.

3. **Total Claim Amount Hypothesis:** Higher total claim amounts are linked with higher CLV.

    1. **Null Hypothesis (H0):** Total claim amount does not correlate with CLV.
    2. **Alternative Hypothesis (H1):** There is a positive correlation between total claim amount and CLV.



-
- Correlation between Total Claim Amount and CLV: 0.17 (P-value: 8.32e-59)
- The correlation is statistically significant.

---

5. Model Selection

Models Evaluated:

- **XGBoost and LightGBM:**
  - These gradient boosting algorithms are efficient for large datasets and offer excellent performance by handling feature interactions effectively.
  - They are robust to overfitting due to their ability to regularize and handle missing data.
  - XGBoost is well-known for its optimization speed and performance in structured data, while LightGBM excels in handling large datasets with high cardinality features.
- **CatBoost:**
  - Handles categorical features natively without the need for extensive preprocessing, making it ideal for datasets with many categorical variables.
  - Performs well on imbalanced datasets and offers competitive accuracy with minimal tuning.
- **Voting Regressor:**
  - Combines the strengths of XGBoost, CatBoost, and LightGBM by ensembling their predictions to reduce bias and variance.
  - This approach leverages the diversity of individual models, leading to better generalization and performance metrics.

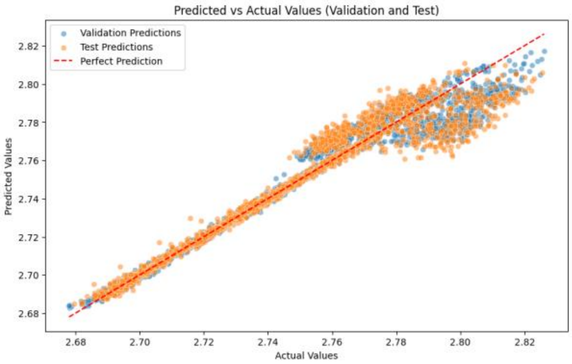
6. Model Building

Preprocessing:

- **Categorical variables**: OneHotEncoder.
- **Numerical variables**: StandardScaler.
- Feature selection based on correlation and business impact.

Best Model:

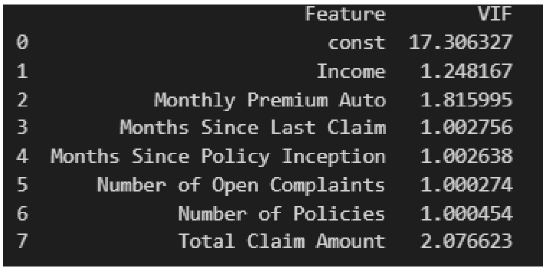
- **Voting Regressor** with XGBoost, CatBoost, and LightGBM.
- **Performance Metrics:**
    - **Validation Metrics:**
        - Validation MSE: 0.00
        - Validation RMSE: 0.01
        - Validation MAE: 0.00
        - Validation MAPE: 0.11%
        - Validation R^2: 0.97
    - **Test Metrics:**
        - Test MSE: 0.00
        - Test RMSE: 0.01
        - Test MAE: 0.00
        - Test MAPE: 0.14%
        - Test R^2: 0.95



7. Validation and Testing

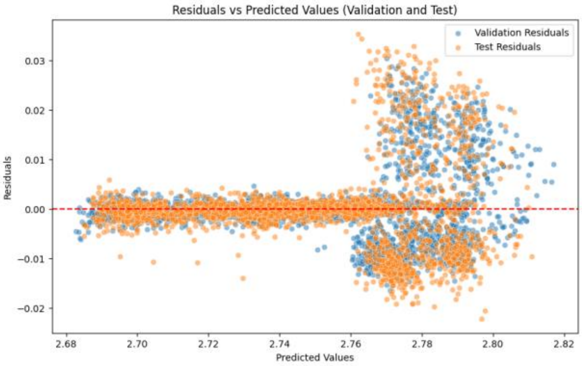
Statistical Tests:

- **Multicollinearity:** Variance Inflation Factor (VIF) below 5 for all variables.

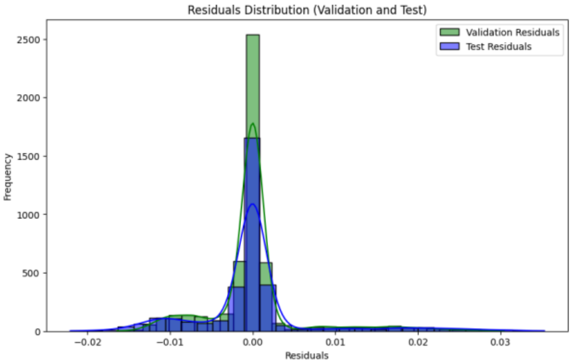


```
                        Feature        VIF
0                         const  17.306327
1                        Income   1.248167
2          Monthly Premium Auto   1.815995
3      Months Since Last Claim    1.002756
4  Months Since Policy Inception  1.002638
5     Number of Open Complaints   1.000274
6            Number of Policies   1.000454
7            Total Claim Amount   2.076623
```

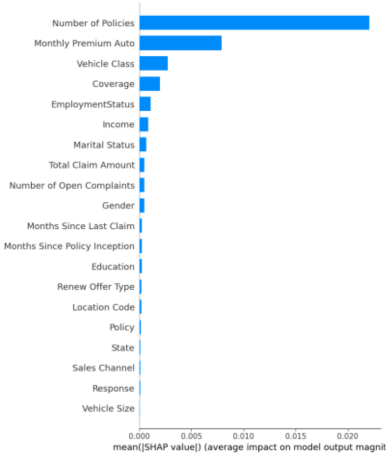
- **Homoscedasticity:** Residual plots confirm equal variance.



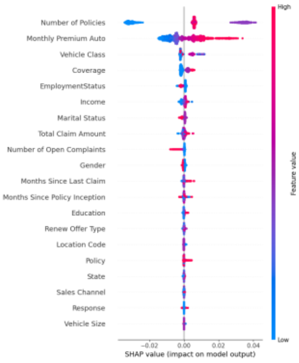
- **Normality:** Residuals approximately follow a normal distribution.



- **Visual representation for Explanatory Insights:**
    - **SHAP Summary Plot:** Visualizes the importance of each feature, ranked by their average impact on model predictions.



    - **SHAP Beeswarm Plot:** Highlights individual feature impacts across the dataset, showing how high or low feature values influence predictions.



8. Recommendations for Improving Business Strategy

1. **Target High-Value Customers**: Focus marketing on customers with higher income and corporate policies.
2. **Upsell Premium Coverage**: Encourage standard coverage customers to upgrade for increased retention and revenue.
3. **Retention Programs**: Design offers for customers nearing the end of policy inception periods to enhance loyalty.
4. **Personalized Offers**: Utilize segmentation based on CLV and customer attributes to create personalized offers, increasing customer engagement.

5. **Focus on Claims Efficiency**: Optimize claim processes for high-CLV customers to improve satisfaction and loyalty.
6. **Dynamic Policy Adjustments**: Introduce flexible policies that adapt to customer life events (e.g., changes in income or employment) to retain high-value customers.
7. **Employment-Based Strategies**:
    - Develop targeted campaigns for employed customers, as they exhibit higher CLV.
    - Offer tailored products to retirees to address their specific needs and improve their CLV.
    - Design engagement strategies for unemployed customers to transition them into higher-value segments when their employment status changes**.**

---

9. Conclusion

Predicting CLV enables personalized strategies for maximizing revenue and customer satisfaction. This analysis highlights the critical factors influencing CLV and provides actionable recommendations to drive business growth. Future improvements could include dynamic updates to models based on real-time customer behaviour.

---

10. Sources

https://shap.readthedocs.io/en/latest/

https://github.com/optuna/optuna

https://www.kaggle.com/code/shailaja4247/customer-lifetime-value-prediction

https://github.com/tushar2704/CLV-Prediction

https://www.geeksforgeeks.org/xgboost/

https://www.geeksforgeeks.org/catboost-algorithms/

https://lightgbm.readthedocs.io/en/stable/

https://pypi.org/project/statsmodels/