Data Walkers

DATA 606: CAPSTONE PROJECT

CUSTOMER LIFETIME VALUE PREDICTION FOR AUTO-INSURANCE COMPANIES

Team Members

Kumbam Nithin Goud

Rindhuja Treesa Johnson

Prudhvi Yaswanth Mundluri

University of Maryland Baltimore County

Professor

Jester Ugalde

Adjunct Instructor at UMBC and

Data Scientist at Booz Allen Hamilton

OVERVIEW

- Business Problem
- Project Ecosystem
- Methodological Approach
  - Data Cleaning, Pre-processing and Exploratory Data Analysis
  - Visualizations
  - Machine Learning
  - LLM Integration with Gemini
- Impact & Value Generation
- Conclusion

BUSINESS PROBLEM

Auto Insurance Company X, a prominent player in the US market, is facing a critical challenge – customer retention. In a highly competitive industry, customers consider various factors beyond just premiums when choosing their insurance provider. To address this, Company X recognizes the importance of Customer Lifetime Value (CLV) as a key metric for understanding and maximizing customer relationships.

CLV Benefits:

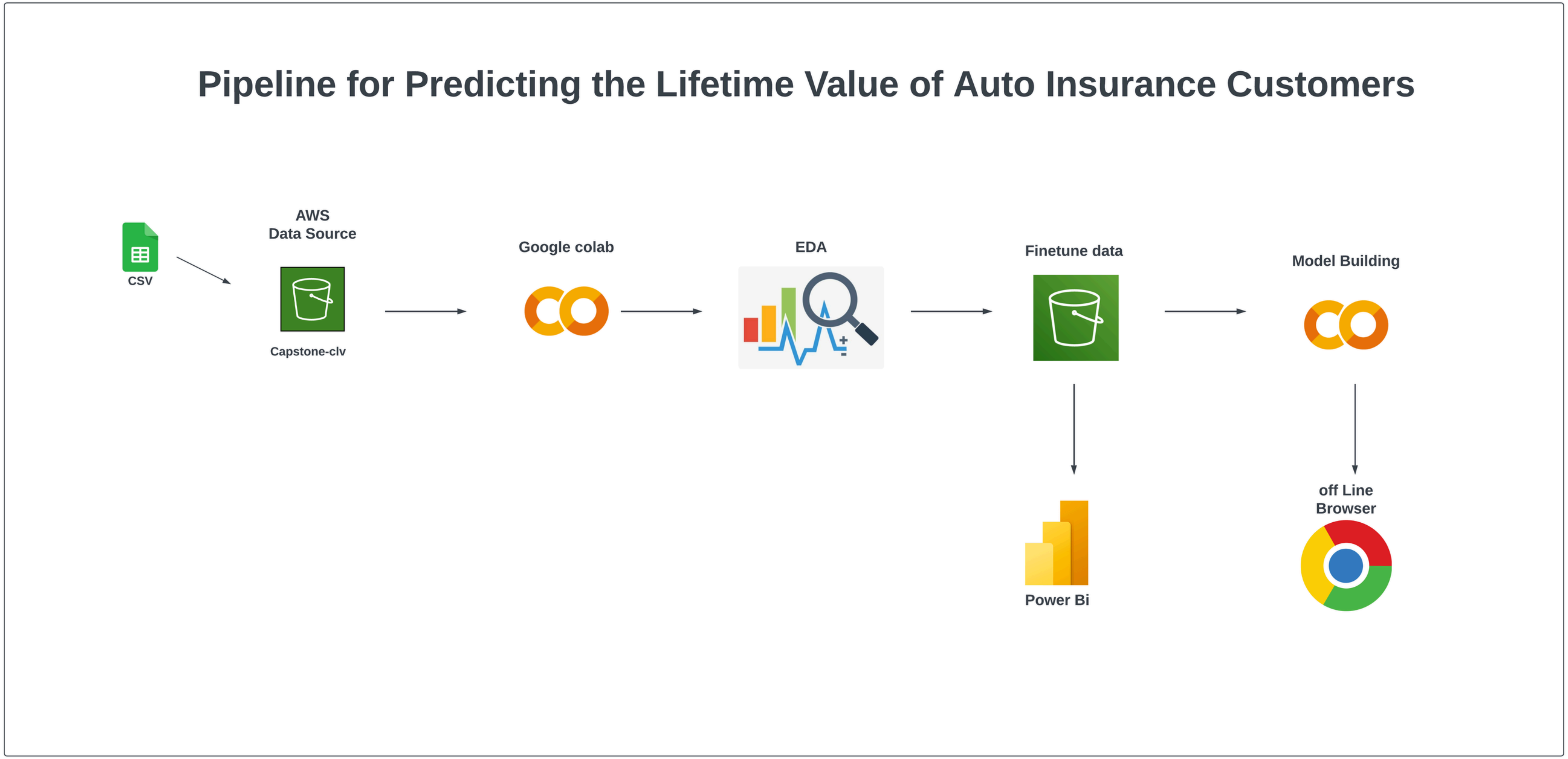
- Identify & retain high-value customers.
- Optimize engagement with lower-value customers.
- Improve overall customer satisfaction and loyalty.

Impact of CLV-driven Strategies:

- Enhanced customer acquisition and retention.
- Reduced churn rates.
- Optimized marketing budget allocation.
- Precise measurement of ad performance.

Goal: Implement CLV-focused initiatives to achieve sustainable growth and profitability.

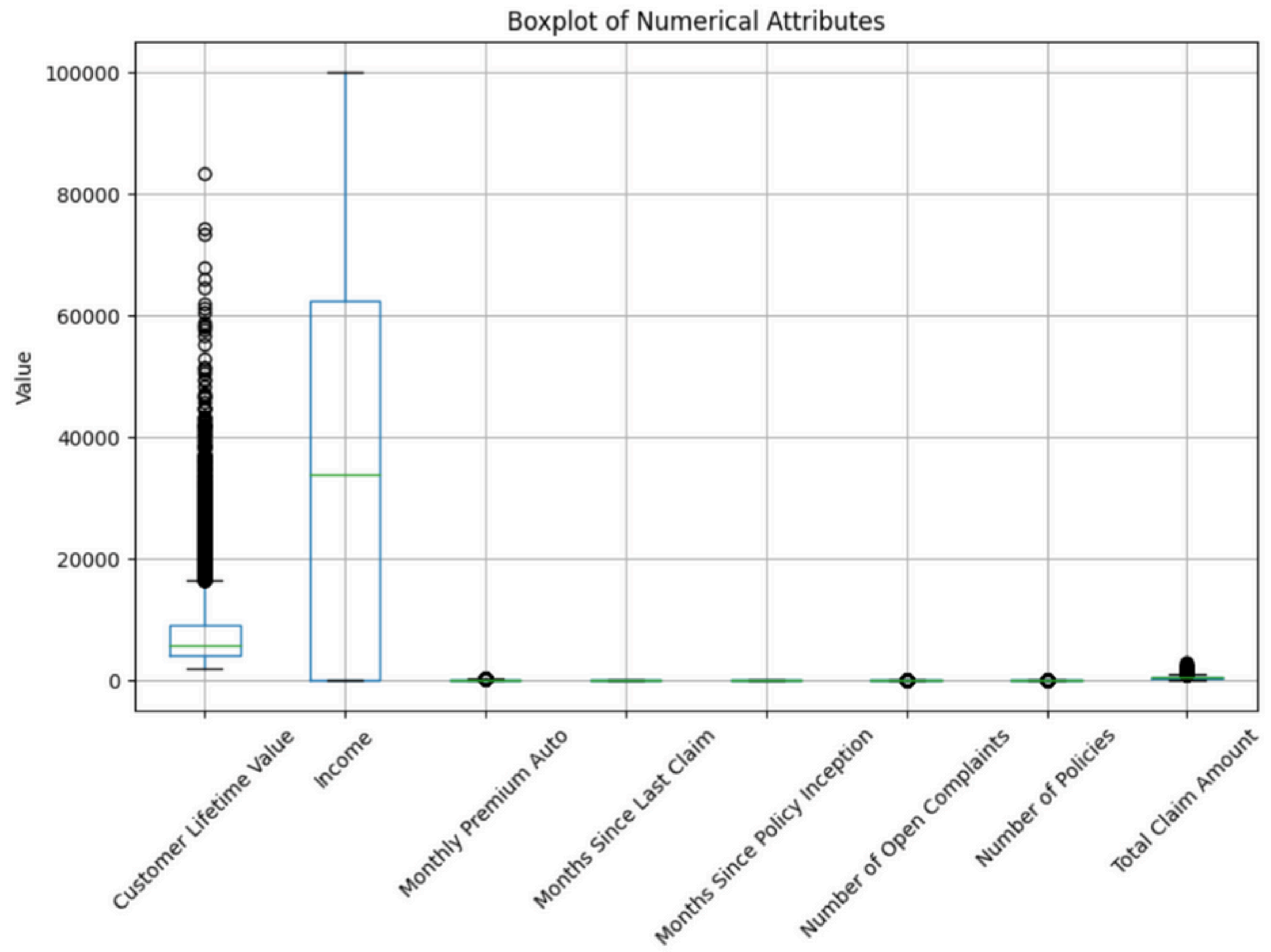
PROJECT ECOSYSTEM



METHODOLOGICAL APPROACH

Data Cleaning and Pre-processing

- Check for Missing Values
- Check for Duplicates
- Type Casting Attributes
- Outlier Detection using Box plots

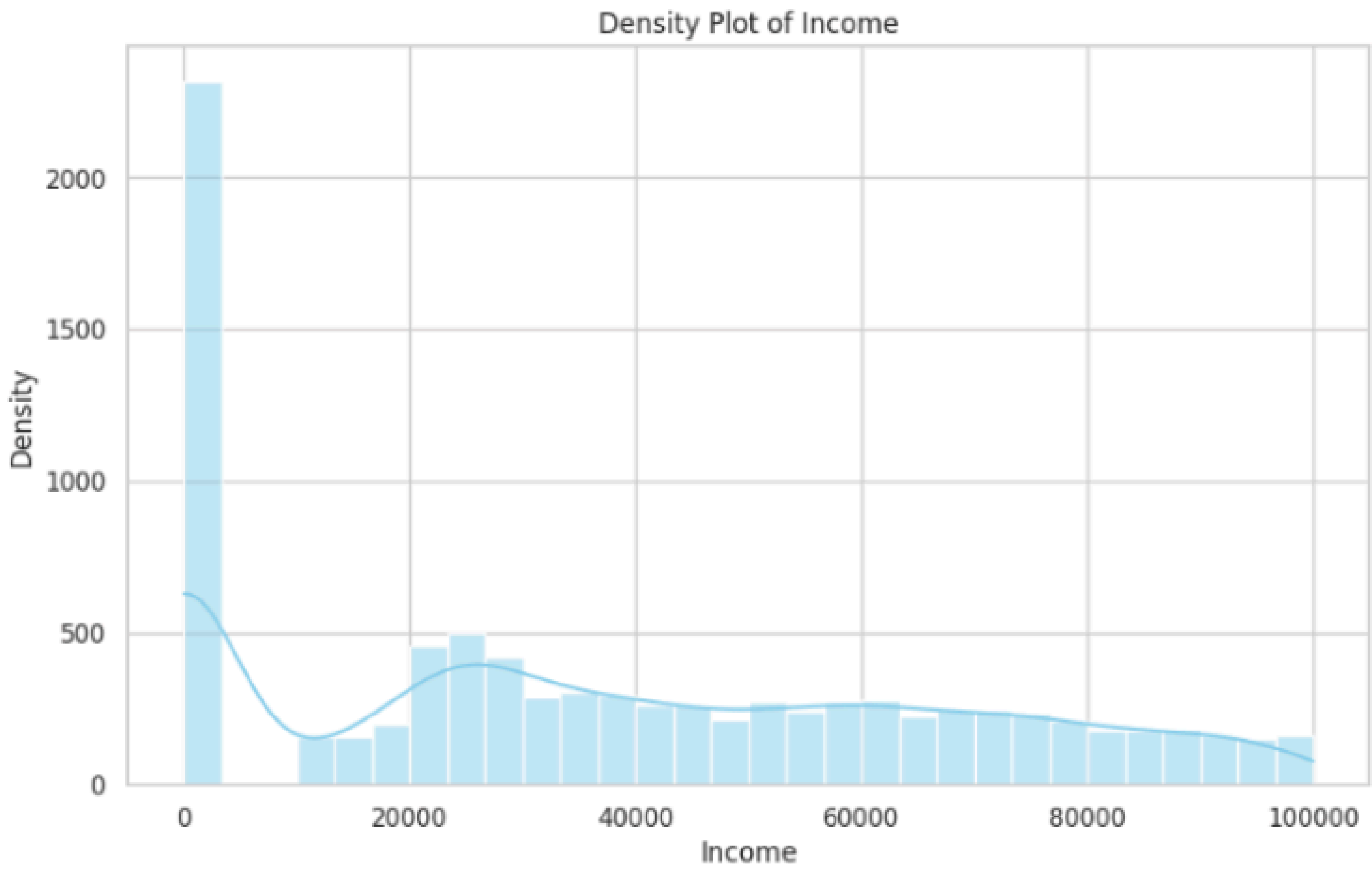


Exploratory Data Analysis

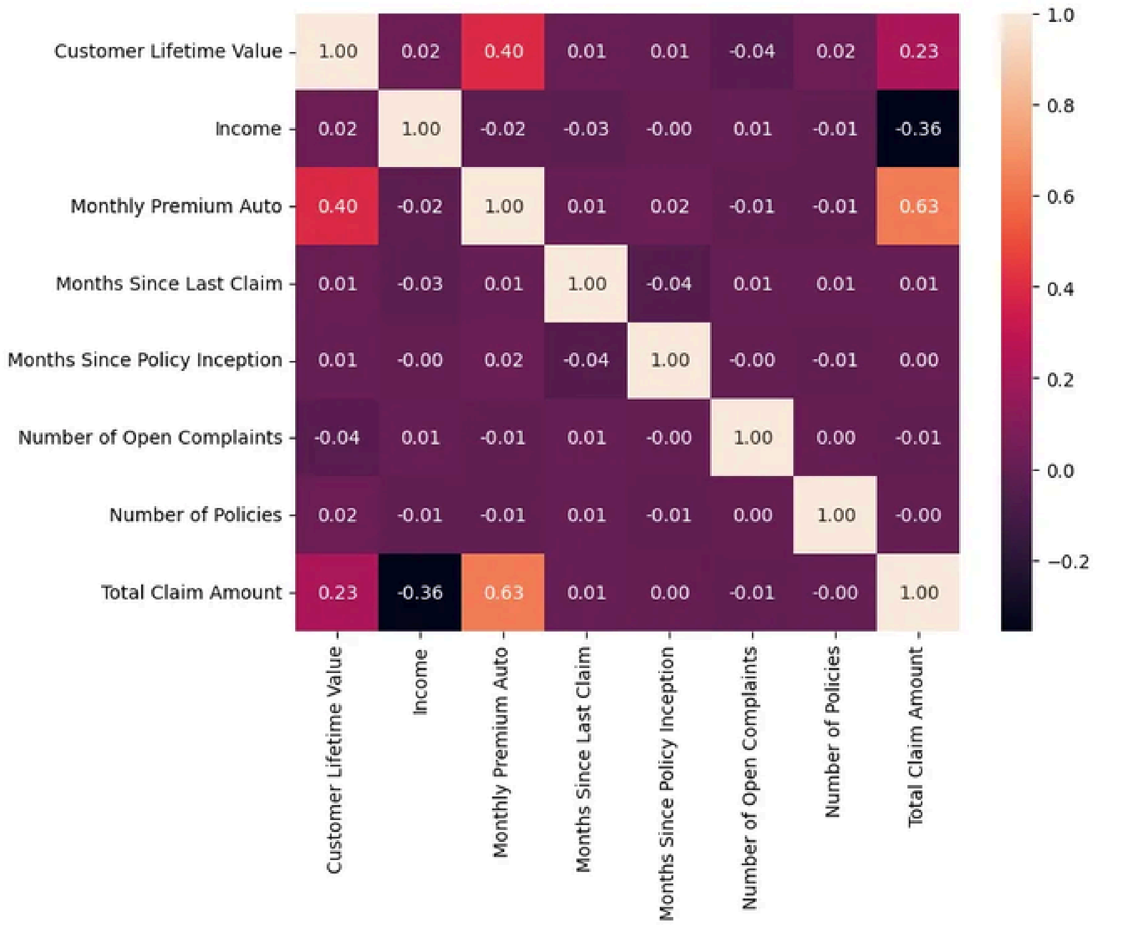
- Feature Analysis: Correlation Matrix and Scatter plots
- Uni-variate Analysis: Histograms and Bar graphs
- Bi-variate Analysis: Using *Group By* clause
- Dimensionality Reduction: PCA & t-SNE
- Clustering Analysis: K-Means
- Label Encoding of Categorical variables

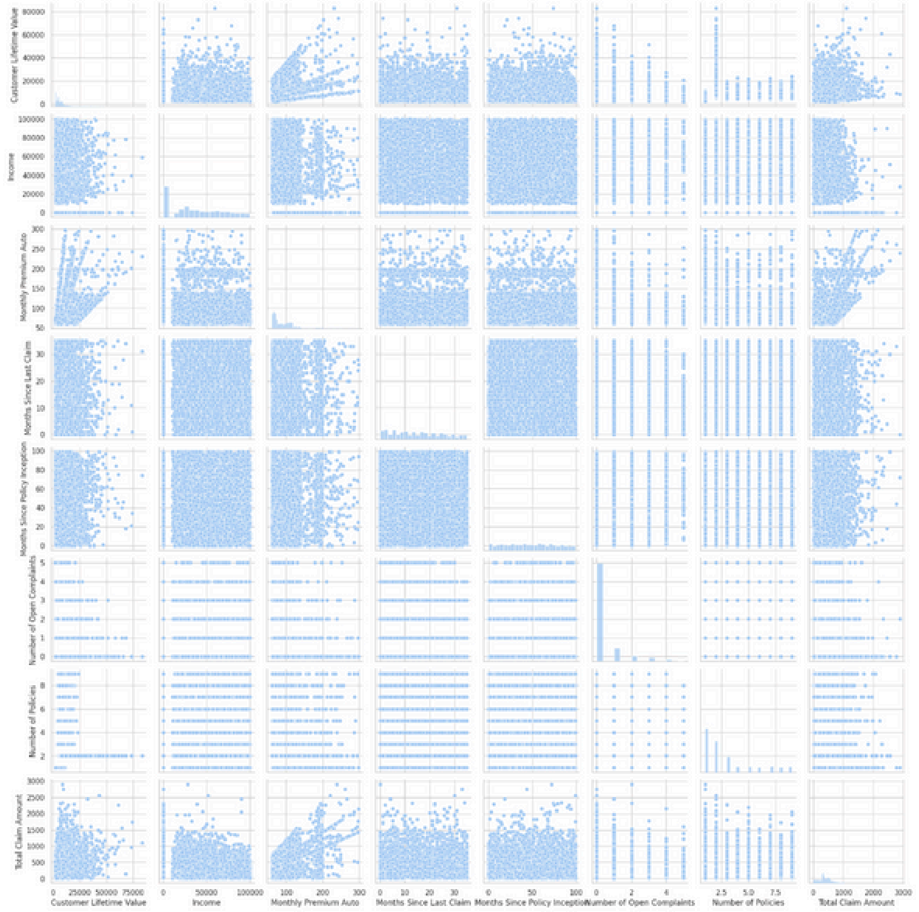
Visualizations

Univariant Analysis: Bar Graph and Dense Plot

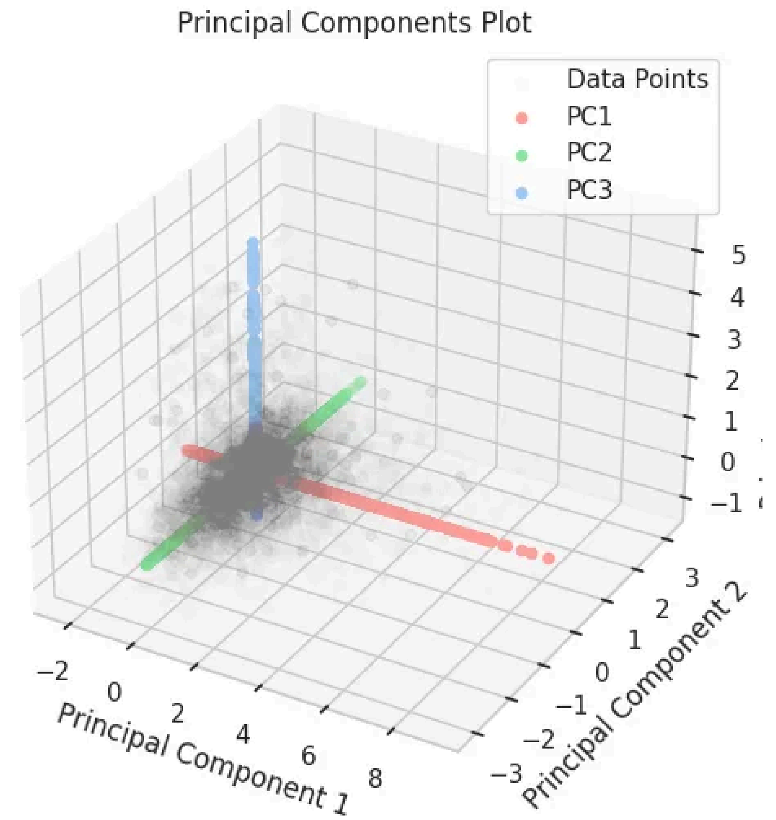


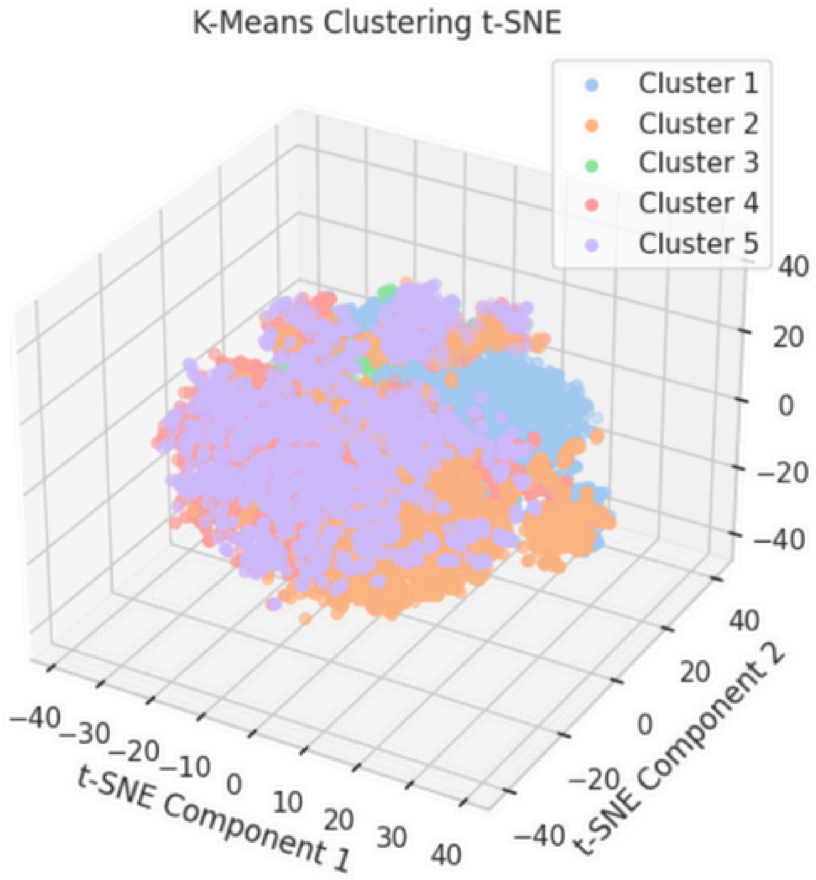
Bivariant Analysis: Correlation Matrix and Scatter plot





Dimensionality Reduction and Clustering



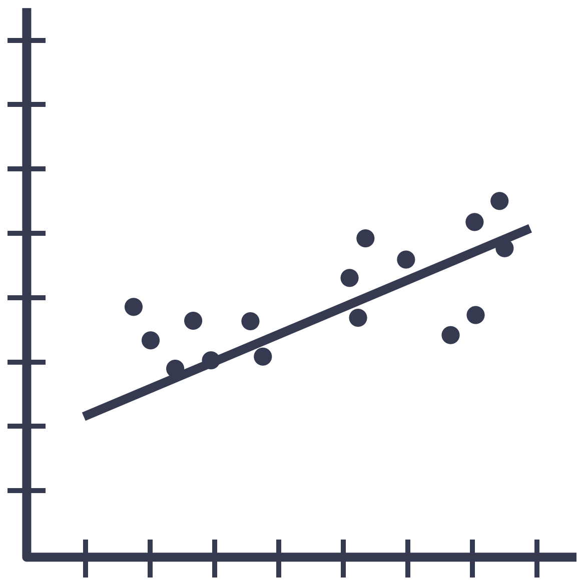


Machine learning

Our methodological approach encompassed a comprehensive data preprocessing stage. Categorical variables were transformed via one-hot encoding, followed by the application of variance inflation factor (VIF) analysis to identify and retain the most salient features for input into the regression models.

Machine Learning

Regression Algorithms



Lasso (L1)

Ridge (L2)

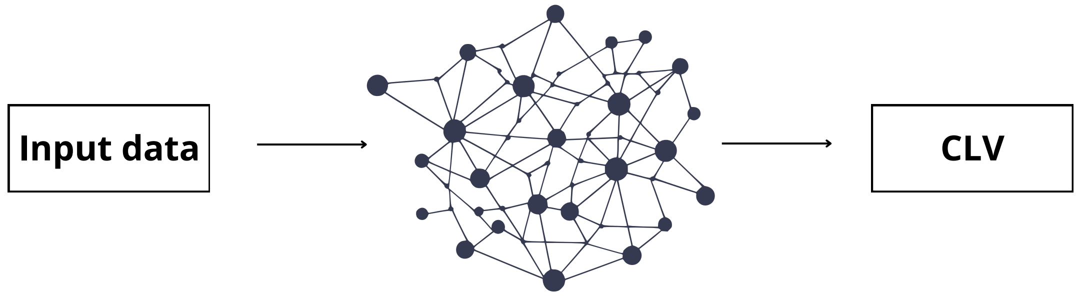
Decision Tree

Random Forest

- Hyperparameter
- Adaboost

Deep Learning

Neural Network

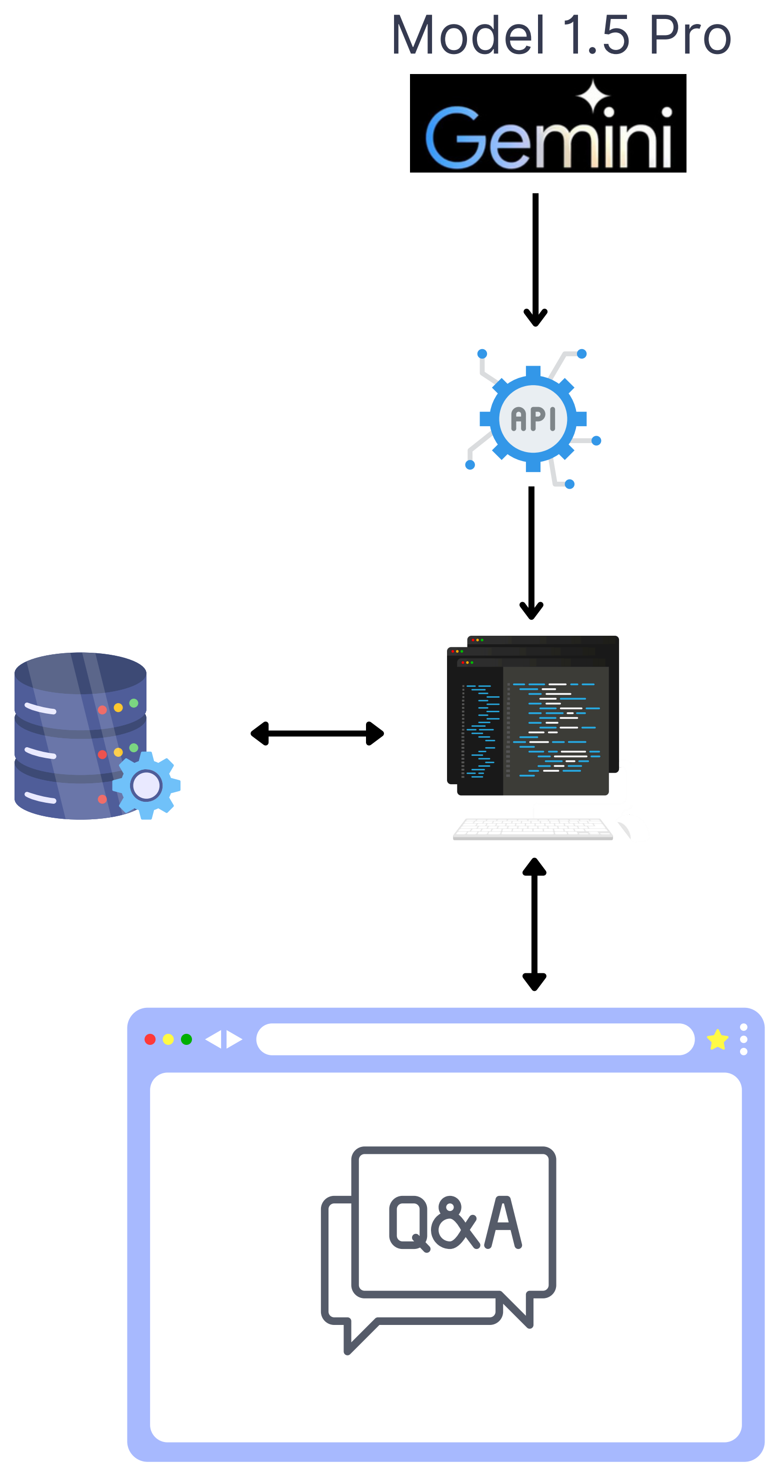


Layer (type)	Output Shape	Param #
dense (Dense)	(None, 128)	2560
dense_1 (Dense)	(None, 64)	8256
dense_2 (Dense)	(None, 64)	4160
dense_3 (Dense)	(None, 32)	2080
dense_4 (Dense)	(None, 1)	33

Total params: 17089 (66.75 KB)
Trainable params: 17089 (66.75 KB)
Non-trainable params: 0 (0.00 Byte)

Model Summery

LLM Integration with Gemini



Gemini & LLM Integration: We've extended the system by incorporating Google's Gemini Large Language Model (LLM) via the Langchain library. This allows us to translate natural language questions into SQL queries for efficient data extraction.

Process Flow:

- **User Input:** Users input data-related questions in plain English.
- **LLM Conversion:** The Gemini LLM, accessed through the GooglePalm class, converts the natural language question into a machine-readable format.
- **SQL Query Generation:** The SQLDatabaseChain object transforms the LLM output into a corresponding SQL query.
- **Database Interaction:** The SQLDatabase framework connects to the MySQL database and executes the generated query, retrieving the relevant data.
- **Results Delivery:** The extracted data is presented to the user in a clear and understandable format.

Benefits: This integration enables users to interact with the database using natural language, eliminating the need for SQL expertise. It facilitates intuitive data exploration, simplifies complex queries, and improves accessibility for a wider range of users.

Impact & Value Generation

Performance Metrics of ML Models			
Model/Metric	**RMSE**	**R-Squared (Train)**	**R-Squared (Test)**
Lasso (L1)	0.5993	0.1950	0.1968
Ridge (L2)	0.5811	0.2498	0.2449
Decision Tree	0.2608	1.0	0.8479
Random Forest	0.2056	0.9829	0.9054
Tuned RF	0.1967	-	0.9134
AdaBoost	0.2171	-	0.8946
Neural Network	-	-	0.8797

- Tree-based models like Random Forest excel in CLV prediction.
- We prioritize efficiency with the Random Forest regressor.
- Predict CLV of potential customers before policy sign-up.
- Live CLV predictions through a user-friendly web interface.
- Q&A interface unlocks data insights for the Data Science team.
- Gain a competitive edge with data-driven customer acquisition.

CONCLUSION

- **Improved Customer Retention:** Identified high-value customers for targeted promotional offers and loyalty programs, leading to increased customer retention and reduced churn rates.
- **Enhanced Marketing Effectiveness:** Enabled data-driven allocation of marketing resources towards high-value customer segments, maximizing return on investment.
- **Data-driven Decision Making:** Empowered stakeholders with CLV insights and interactive visualizations, facilitating informed decisions regarding customer acquisition, retention, and overall business strategy.
- **Enhanced User Experience:** Provided a user-friendly Q&A interface for easy access to information, promoting data democratization and knowledge sharing within the organization reducing the time required for developing efficient SQL queries.

REFERENCES

1. Danao, M. (2023). What is Customer Lifetime Value (CLV)? *Forbes Advisor*.
2. LangChain, I. (2024a). LLMs.
3. LangChain, I. (2024b). SQL Database.
4. Ross, S. (2021). How do insurance companies make money? Business Model Explained. *Investopedia*.

Data Source

IBM Watson Marketing Customer Value Data

IBM Watson Analytics

kaggle.com

Thank you