# Predicting Customer Lifetime Value: A Comprehensive Approach with Machine Learning and Deep Learning Models

MSc Research Project
Data Analytics

Arun Joy
Student ID: x23174994

School of Computing
National College of Ireland

Supervisor: Jaswinder Singh

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | Arun Joy |
| **Student ID:** | x23174994 |
| **Programme:** | Data Analytics **Year:** 2024 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Jaswinder Singh |
| **Submission Due Date:** | 12/12/2024 |
| **Project Title:** | Predicting Customer Lifetime Value: A Comprehensive Approach with Machine Learning and Deep Learning Models |
| **Word Count:** | 8240 **Page Count:** 23 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Arun Joy

**Date:** 12/12/2024

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Predicting Customer Lifetime Value: A Comprehensive Approach with Machine Learning and Deep Learning Models

Arun Joy
x23174994

**Abstract**

Predicting Customer Lifetime Value (CLV) is important for marketing to manage customer relationships. Traditional models mostly struggle with the high dimensional type of data, so this study is centered on forecasting the Customer Lifetime Value (CLV) using a dataset from Kaggle which contains detailed customer information that enables interpretation of demographic and purchasing patterns. The pre-processing steps include removing unimportant features, converting categorical data into numerical data with the help of Label Encoder while normalizing the data with the help of MinMaxScaler. The selection of the features is done to select the top 10 important features that have a high influence on predicting CLV using feature importance methods. The dataset is then split into training and testing sets in the 80:20 ratio for model training. Modelling encompasses Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, GRU-LSTM and Bidirectional LSTM with Attention Mechanism model. For the performance comparison of these models, MSE, RMSE and $R^2$ Score are used. At last, it was perceived that the Bidirectional LSTM model with Attention Mechanism is the most efficient among all other models with the value of $R^2$ of 0.94.

**Keywords:** Customer Lifetime Value (CLV), Machine Learning, Deep Learning, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, GRU-LSTM, Bidirectional LSTM (BiLSTM)

# 1 Introduction

## 1.1 Background

The Customer Lifetime Value (CLV) prediction is a very important element in the performance measurement of enterprises to identify and analyze the value of their customers in the long way (Firmansyah et al., 2024). CLV provides an estimate of how much a specific customer is going to spend in his/her lifetime of being associated with the firm (Marmol et al., 2021), which allows the formulation of sound marketing strategies, customer retention and proper resource utilization. Various techniques like regression analysis, machine learning models and predictive analytics are used more often to evaluate the drivers of customer loyalty and purchase intention. Knowledge about CLV also helps in the determination of acceptable revenues, customer contact and communication strategies with the desire to maximize profitability (Kumar, 2024).

Previous researches completed on Customer Lifetime Value (CLV) prediction, which will be seen in the Related Work section, has relied on traditional machine learning methods including linear regression to categorize customers depending on their demographic and purchasing profiles. Recent developments have emerged in deep learning models, especially LSTM for dealing with sequential data, but the BiLSTM with Attention Mechanism has only a few studies. The different models as seen in other works pose a general prediction, and they do not have enough accuracy for accurate CLV prediction, especially with dynamic customer situations, so this research fills this gap by proving that the BiLSTM with Attention Mechanism can accurately predict the Customer Lifetime Value (CLV) with an accuracy of 94% and therefore provide a good strategy to support the customer retention strategy.

## 1.2 Aim of the Study

The aim of this study is to provide a highly accurate medium for predicting CLV using both machine learning and deep learning algorithms. The data set used for this research, IBM Watson Marketing Customer Value, looks at the customer demographic and their purchase behaviour that helps in classifying them as valuable customers and track their interactions with the business. By using Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression and deep learning models like GRU LSTM and BiLSTM with an Attention Mechanism, the aim is to increase the level of accuracy of the models and enhance the customer retention strategy. The data will undergo cleaning, exploratory data analysis, feature selection and preprocessing before being used for modelling in this study. The efficiency of the models will be measured using performance evaluation metrics such as MSE, RMSE and $R^2$ Score. Finally, this study aims at developing a Flask web application to provide an efficient solution to help businesses develop customer retention strategies and consequently support long-term business growth and development. The motive for this approach is to extend current mainstream prescriptive techniques with solutions that can be used in real business contexts, with problems associated with customers retention and profitability. Importance is given to the developed BiLSTM model with an Attention Mechanism since it focuses on significant types of features in the dataset.

## 1.3 Research Questions and Objectives

### Research Questions

The research question for this research are as follows:

1. How can advanced deep learning models, specifically GRU-LSTM and BiLSTM with an Attention Mechanism, be utilized to enhance the prediction accuracy of Customer Lifetime Value (CLV) compared to machine learning models, such as Decision Tree Regression, Random Forest Regression and Gradient Boosting Regression?

### Research Objectives

The research objectives for this research are as follows:

1. To evaluate the effectiveness of machine learning and deep learning models by using evaluation metrics like MSE, RMSE and $R^2$ Score.
2. To identify and apply feature selection using ExtraTrees and transformation techniques like LabelEncoder and MinMaxScaler to optimize model input variables, thereby to enhance the accuracy of Customer Lifetime Value (CLV) prediction.

## 1.4 Structure of the Report

The structure of the report is as follows:

1. **Introduction**: Introduces the research topic, objectives and the scope of the study on Customer Lifetime Value (CLV) prediction.
2. **Related Work**: Reviews existing research on CLV prediction methods, including machine learning and deep learning approaches.
3. **Research Methodology**: Describes the data collection process, feature selection, preprocessing techniques and the models used for CLV prediction.
4. **Design Specification**: Outlines the architecture of the deep learning models used in this research.
5. **Implementation**: Details the implementation of machine learning and deep learning models.
6. **Evaluation**: Presents the evaluation of model's performances based on various metrics like MSE, RMSE and $R^2$ score.
7. **Conclusion and Future Work**: Summarizes the findings, discusses limitations and proposes directions for future research and model improvements.

# 2 Related Work

## 2.1 Importance of Customer Lifetime Value Prediction

Customer Lifetime Value (CLV) forecasting is an indispensable tool in current conceptualizations of business management (Ehsanifar et al., 2022) since it reveals where the customer is valuable for the organization's activity and how resources and strategies should be distributed. When CLV is correctly predicted, a firm can pinpoint those customers with high CLV (Yan and Resnick, 2024), which helps to manage and control the costs of customer acquisition and maintain proper retention programs that are crucial in increasing customer loyalty and overall revenues. CLV insights allow a business to classify customers and develop specific marketing strategies to address the needs of different customers, which increases attractiveness to the target clients and proves profitable. Thus, CLV predictive models enable further prediction of churn risks to prevent certain customer-damaging behaviors within a company (Devriendt et al., 2021), while customer relationship management (CRM) generally considers anticipating clients' behaviors. The CLV prediction helps the firms to better allocate their marketing resource with a higher reward/ cost ratio in terms of customer lifetime value (Zhao et al., 2023). As such, in competitive markets, such insights can be invaluable to business by focusing on the benefits of long-term customer relations at the expense of short-term gains thereby enhancing the stability of the business.

## 2.2 Traditional Approaches to CLV Prediction

This section will discuss several traditional methods involving Customer Lifetime Value (CLV) prediction. The different approaches that have been used in prior research related to CLV prediction include RFM (Recency, Frequency, Monetary) analysis, K-means clustering and Pareto/NBD analysis.

The study given by Myburg, 2023 suggested applying supervised machine learning, XGBoost, to predict CLV and the probability of the customer's downgrading to the lower tier in the future. The approach was to build classification models using metrics such as RFM (Recency, Frequency, Monetary) using a dataset from Fast Moving Consumer Goods (FMCG). The findings showed that the models performed reasonably well in different ranges of forecasting future CLV spans of the applied timeframe, and slightly better in terms of predicting a decline in the second period relative to the third.

The research (Sun et al., 2023) looks at the issue of calculating CLV in noncontractual situations where customer behavior is not easy to forecast. The study recommends a customer segmentation model based on the customer lifetime value theory since it majors in the measurement of customer value and their splitting into groups using the machine learning analysis linked with Customer Relationship Management (CRM) evaluation models. To expand the proposed approach, a well-defined feature engineering framework ranging from data selection, preprocessing and transforming into segmented knowledge is incorporated to improve the current segmentation performances.

Another study presented by Bauer and Jannach, 2021 presents a new approach to predicting Customer Lifetime Value (CLV) using some machine learning techniques to increase its precision. The current approach is unique, and the proposed architecture specifically relies on an individualized deep learning model derived from encoder-decoder sequence-to-sequence recurrent neural networks with temporal convolutions added, alongside gradient boosting machines (GBMs). One of the issues this research seeks to solve with the currently utilized methods is that they struggle to capture intricate temporal structures, like regular purchasing patterns.

The study provided by Heldt et al., 2021 indicates that the RFM per product (RFM/P) model is a development of the RFM model which considers the estimates of the customer value from both the customer's and the product's viewpoint. The proposed approach first estimates customer value separately for each product and then sums up these values, which will give the overall customer value. A major problem solved by this model is that standard RFM approaches fail to account for the differing market seasonality of many products in the industry, which results in a reduction in forecast accuracy. Moreover, the study also reveals that the RFM/P model performs better than the RFM model by providing better predictive accuracy in situations where customers' purchase behaviors change in terms of recency and frequency for individual products thus making the tool more suitable for dynamic purchasing behavior environments.

Marketing resources are limited and hence, the study (Laksono et al., 2023) developed a customer segmentation framework in business-to-business contexts employing the customer lifetime value using RFM modelling. The study proposes the use of the K-means algorithm in a data set of 351 customers and applies the elbow method to ascertain the right number of clusters to employ. One major issue witnessed in this research is how to properly categorize customers so that the resources can be appropriately partitioned according to the realized or estimated useful customers to the firm.

This paper (Sun et al., 2021) focuses on using machine learning techniques namely, GBDT and RF for CLV in the retail sector and evaluated the performance of these techniques in comparison with Pareto/NBD (HB) and Pareto/GGG models. The study was conducted using only 43 weeks of transaction data from a large retailer in China, the researchers sought to predict customer value during the subsequent 20 weeks. The results specify that both GBDT and RF are more accurate in the prediction compared to the Pareto models in general.

There is another study presented by Castéran et al., 2021 where there is a systematic review of the various approaches to modelling Customer Lifetime Value (CLV), customer retention and churn to support customer-based analysis for efficient management of customer relationships and resource deployment. The study aims to identify which modelling techniques are suitable to be applied, within contractual and noncontractual environments, by distinguishing between deterministic and stochastic methods.

Xie et al., 2021 aims to enhance customer relationship management by predicting customer inactivity and repeat transaction frequency through a systematic comparison of two primary approaches, the combined methodologies consist of stochastic modelling, namely Pareto/NBD and the machine learning methodology, neural network analysis. The results show that in no case does one of the approaches dominate the other, but the strategies had different measures of fit, with the empirical as well as the simulated data.

Another study provided by Ejgerdi and Kazerooni, (2024) gave a stacked ensemble learning approach designed for the CLV prediction, to obtain better prediction results of CLV. The proposed method incorporates two or more machine learning methods, in this way improving the model’s accuracy. The stacked ensemble learning method was adopted using data from a textile sales company and then was compared with several commonly used predictive models such as deep neural network, support vector regression, light gradient boosting, random forest, extreme gradient boosting etc. The characteristics of all the models indicate that the applied stacked ensemble method is the most effective one with the normalized root mean squared error of 0.248, normalized mean absolute error of 0.364 and the coefficient of determination of 0.848. Additionally, it was shown that after hyperparameter tuning, its accuracy had increased and, therefore, the presented model can deliver more accurate CLV prediction. This approach provided a better way to forecast CLV, which is critical in CRM and marketing communication.

## 2.3 Deep Learning Approaches to CLV Prediction

In this section, the existing Deep Learning approaches for CLV prediction with the help of the methodologies adopted in recent research works are discussed.

Firstly, the study (Chopra et al., 2020), presents a model that integrates LSTM networks which have excellent temporal modelling ability, with ensemble learning techniques like Random Forests and Gradient Boosting Machines to enhance the CLV prediction. The proposed model is developed by combining LSTM, which can capture temporal dependencies, with ensemble learning that can enhance variability and increase the stability of the CLV forecast. The model is tested and developed based on the dataset collected from one of the most influential e-commerce sites, containing customer details, ordering history and interaction records. The study shows that the hybrid model promises lower mean squared error values and better accuracy than the baseline models, including conventional regression and standalone LSTM systems. The findings state that the ensemble component has the effect of making the model more stable while the LSTM adequately addresses the temporal relationships making the hybrid solution a viable way to enhance the accuracy of CLV forecasts that will be more helpful in achieving real business value.

Kumari et al., (2024) presented and applied some deep learning models, namely, LSTM and RNNs to enhance the CLV forecasting for the continuously changing e-commerce market. CLV is a success confluence figure which businesses use to indicate probability of a particular group or customers to generate in their life cycle, hence the accuracy of these forecasts are essential for businesses. The research focuses on the problem of modelling time series of customer data, although purchasing behaviour depends on time frames. The study expects to improve CLV predictions and uncover relevant information for designing specific marketing and customer retention strategies by breaking transaction data down to the optimal level of disaggregation. Through the results obtained, the authors show that applying LSTMs and RNNs as deep learning enhances the methods of capturing temporal characteristics of the customers and improves the CLV models' accuracy and prediction capabilities. It links ecommerce analytics and deep learning to provide more accurate CLV forecasts that can inform better and smarter business decisions.

Therefore, this research aims at analysing existing strategies for the Customer Lifetime Value (CLV) prediction by deep learning methods and presents different approaches that combines GRU LSTM and BiLSTM with the Attention Mechanism. Chopra et al., (2020) showed that LSTM networks can be implemented together with other methods to enhance the accuracy as well as stability of the models. Consequently, Kumari et al. (2024) presented LSTM and RNN models to illustrate the temporal aspect of complex e-commerce settings.

Even though the utilised dataset has no sequential data, GRU LSTM network is selected because of its computational costs and its appropriateness for modelling non-linear relations between features. The BiLSTM with Attention Mechanism is used since it provides the model with the ability to emphasize on which exact features are important. This approach is designed to provide the accurate and comprehensible CLV forecasts, which in turn should help to make better managerial decisions.

# 3 Research Methodology

## 3.1 Methodology

The CRISP-DM methodology used for predicting Customer Lifetime Value (CLV) in this research has six phases as discussed below.

**Business Understanding:** The aim of this research is to enhance customer loyalty by applying the prediction model to understand the characteristics of these valuable customers. The information about the Customer Lifetime Value makes it possible to improve the marketing strategies and the distribution of resources. The goal of this project is to estimate the Customer Lifetime Value and determine the directions for customer loyalty, retention and total revenue generation based on the customer characteristics and buying behavior.

**Data Understanding:** During this phase, IBM Watson Marketing Customer Value dataset obtained from Kaggle was used to define customers' attributes like demographic details and other features of CLV. This structured dataset has attributes like the gender or income of the customer, the number of policies owned by the customer, trends of both claim and coverage of policies which are essential for the analysis of customer value. Descriptive data analysis is done to see the distribution and the nature of the characteristics that can be used for prediction. In this study, descriptive and graphical analysis is used to analyze the relationship and trends between the determinants, and the effect of these determinants on CLV.

**Data Preparation:** The main step of machine learning process is data collection, followed by data cleaning and pre-processing which is together named as data preparation phase. The first step of feature pre-processing includes checking for missing values, then removing features that are found to be irrelevant for prediction. Categorical features are then converted to numerical forms using processes such as label encoding and numeric features are then scaled or normalized with Min-Max scaling to keep it within a certain range between 0 and 1 for modelling.

**Modelling:** In the modeling phase, several machine learning and advanced deep learning algorithms are deployed for estimating CLV. Firstly, three machine learning models, Decision Tree, Random Forest and Gradient Boosting Regression models as they have different capabilities in terms of fitting complex feature interactions as well as avoiding over fitting problem. The GRU and the BiLSTM models are then employed to strengthen the model's comprehension. Both machine learning and deep learning models are integrated to guarantee a broad coverage style of CLV approach while maintaining the accuracy of the model in predicting customer lifetime value.

**Evaluation:** In the evaluation phase, the performance of each model in predicting Customer Lifetime Value (CLV) is assessed using key metrics such as Mean Squared Error, (MSE), Root Mean Squared Error (RMSE) and coefficient of determination ($R^2$ score). MSE and RMSE shed light on the mean square of error for predictions while RMSE has the additional advantage of interpretability. A model with lesser MSE and RMSE signifies that the model is making near accurate predictions. The $R^2$ Score, on the other hand, giving the ratio of the variance in CLV prediction by the model, with a value closer to 1, means a better prediction capability.

**Deployment:** In the deployment phase, the best model is incorporated into the graphical interface of the web application on the Flask framework. Figure 1 displays a web application interface for the prediction of Customer Lifetime Value (CLV) that was designed with the help of Flask framework. The main interface presents a form with multiple input fields corresponding to the key features identified and the output value for CLV is displayed along with a table showing the different values of the customer details as seen in Figure 2. This web application enables the entry of new customer details and instantly generates the likely Customer Lifetime Value (CLV), which forms the basis for customer loyalty and marketing.

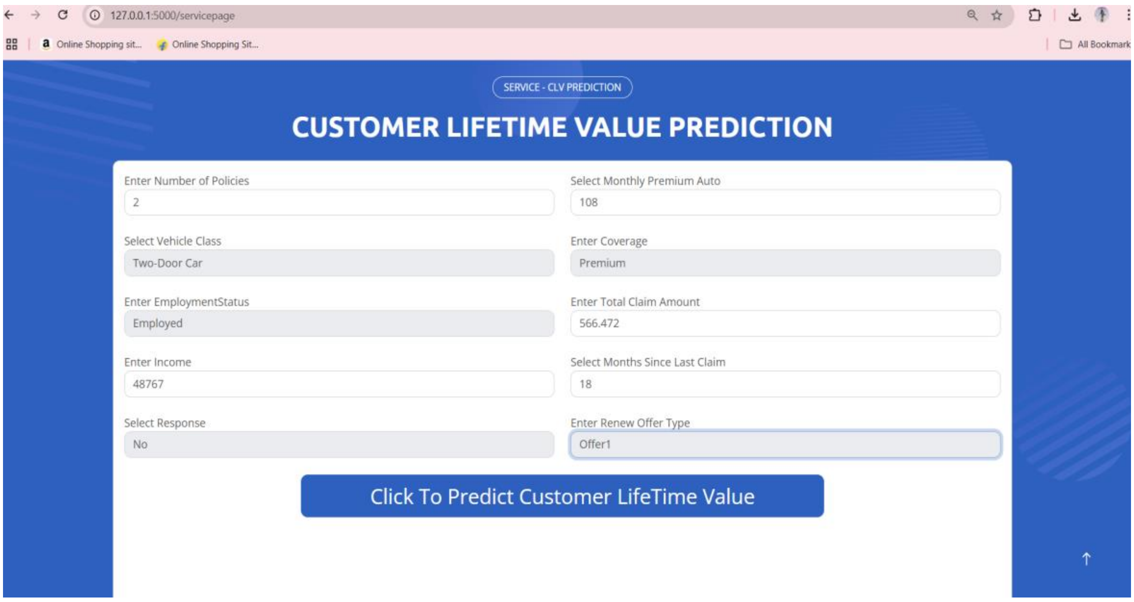


**Figure 1: CLV Prediction Web Application Interface**

| | Number of Policies | Monthly Premium Auto | Vehicle Class | Coverage | EmploymentStatus | Total Claim Amount | Income | Months Since Last Claim | Response | Renew Offer Type |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2 | 108 | 5 | 2 | 1 | 566.472 | 48767 | 18 | 0 | 0 |

CUSTOMER LIFETIME VAUE IS : 12516.38

**Figure 2: Output for CLV Prediction**

The workflow diagram as seen in Figure 3 below shows the logical sequence employed in this study for predicting CLV and for designing an effective customer retention strategy that leverages Machine Learning and Deep Learning models.

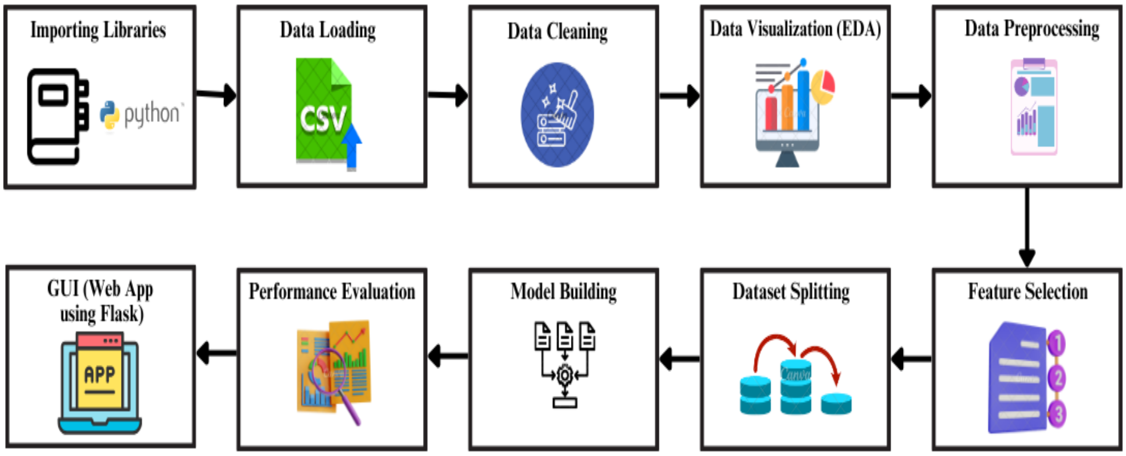


**Figure 3: System Workflow**

## 3.2 Libraries Imported

In this study, several effective libraries are used for the analysis and the prediction of Customer Lifetime Value (CLV). The core libraries used include pandas for efficient organizational and manipulative control of the data. For visualizations, seaborn, matplotlib and plotly are used for interactive visualizations. For developing the models for predictions, it includes the usage of scikit-learn toolkit, which provides Decision Tree, Random Forest, and Gradient Boosting Regressions. For the deep learning models, tensorflow and keras are used, incorporating the layers of GRU and LSTM for pattern recognition. The LabelEncoder is used for converting the categorical values to equivalent numeral rates and MinMaxScaler is used for scaling the features. Furthermore, the ignorable warnings are managed using the warnings module of Python to have a nice printout.

## 3.3 Dataset

The dataset used for this research is the IBM Watson Marketing Customer Value Data, which is taken from the Kaggle source has in total 9134 rows and 24 columns. This dataset contains customer-related variables, and provides rich opportunities for exploring customers' demographics, purchasing patterns, and communication. Other information is also provided in this dataset that includes the policy details of the customers and their historical activity level and retention rate. These diverse features allow for building forecasting models to assess the customer's lifetime value and to make managerial decisions in customer relations management.

## 3.4 Data Cleaning

While cleaning the dataset, the aim was to get a clean dataset with quality data appropriate for use in analysis. Firstly, unimportant features were eliminated by removing two columns, 'Customer' and 'Effective To Date' as they were not found helpful for the prediction process. The outliers in the 'Customer Lifetime Value' column were removed by taking only those CLV values below 15000 to keep the results accurate and it was found using a histogram chart. The presence of null values was also checked which proved to be none. Due to this selection, the dataset became much more refined, with 8163 records and 22 main features. Also, the 'Customer Lifetime Value' and 'Total Claim Amount' variables have been rounded to two decimal places and thus present most of the numerical data included in the analysis uniformly. Performing such steps of data cleaning was vital for reaching a high-quality dataset, which would be the ground for further steps of predictive analysis. These steps on data preparation are vital in improving the efficiency and effectiveness of the models used in the determination of the customer lifetime value.

## 3.5 Data Preprocessing

During data preprocessing, categorical data was transformed into numerical data and the normalization of the data was done for modeling. Firstly, the categorical features are selected using the select_dtypes by eliminating float and integer types. The categorical features which have been identified are State, Response, Coverage, Education, EmploymentStatus, Gender, Location Code, Marital Status, Policy Type, Policy, Renew Offer Type, Sales Channel, Vehicle Class and Vehicle Size. As for these categorical variables, the best solution is using scikit-learn's LabelEncoder that labels all the categories in the dataset. For all the categorical column, the fit_transform method was used to accomplish the adjustment for the dataset, which will help to minimize data loss throughout the analysis process. After the encoding, the numeric features are scaled to a range between 0 and 1 to ensure that all the features contribute equally to the model.

## 3.6 Data Visualization

The histogram in Figure 4 with a rug plot below shows the distribution of Customer Lifetime Value (CLV) in the dataset. As seen in the given figure, the distribution of customers by CLV is positively skewed, meaning that the majority of the customers are clustered at the lower CLV range, almost within 2000 and 6000 CLV and reaching up to 14000 CLV after removing outliers. The rug plot located at the top of the histogram plot gives extra information regarding the related features of that CLV. In the histogram, the height of the bars refers to the frequency count of the number of customers in each CLV range which is approx. 450-500 for the highest bar of the histogram.

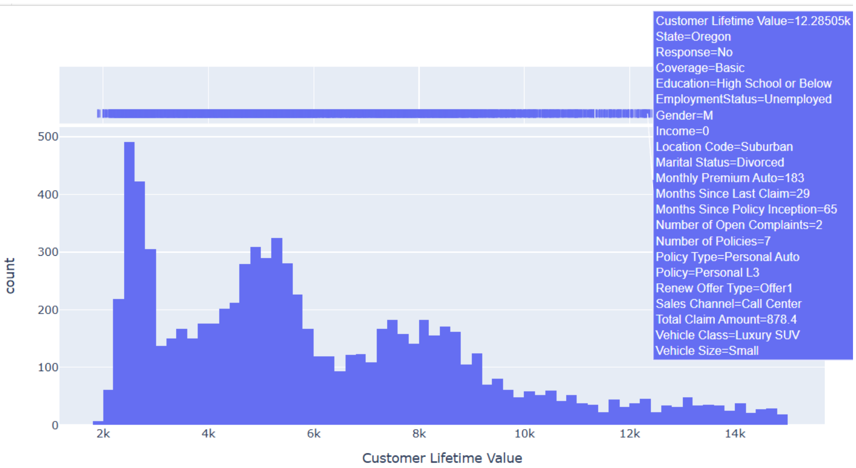


**Figure 4: Histogram of Customer Lifetime Value.**

Figure 5 is a line plot with markers and value labels of the average Customer Lifetime Value (CLV) in the five states of western United States. Oregon has the highest average CLV of 6179.92 followed by California with 6147.60 CLV and then Arizona's 6106.53. Nevada achieves the second-lowest average CLV of 5945.71 and Washington state has the lowest average CLV at 5832.11. The plot shows the varying customer value across these states, ranging from 347.81 higher for Oregon than Washington. Differences in CLV for the different states may be attributed to characteristics such as regional income level, competition, demography or differential marketing impact in these states.

The pie chart which is been presented in Figure 6 below focuses on the average distribution of Customer Lifetime Value (CLV) for the Gender column, which has shown an almost equal ratio of male to female. The average CLV for males comprise of 50.3% and average female CLV is 49.7% of the total CLV, a very close split with males slightly outdoing females in a 0.6% margin. It is easy to recognize what the different colors mean, where blue depicts male customers, and red is for female customers. This distribution indicates that there is fairly no skewness when it comes to gender-based customer lifetime value.

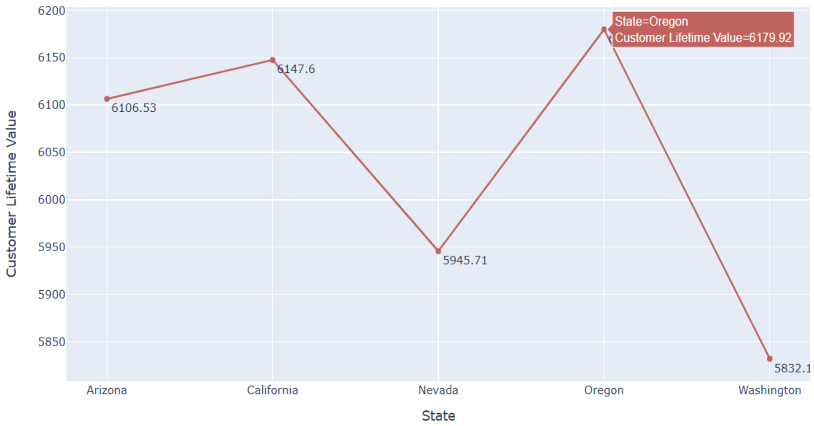


**Figure 5: Line Plot showing average CLV per State.**

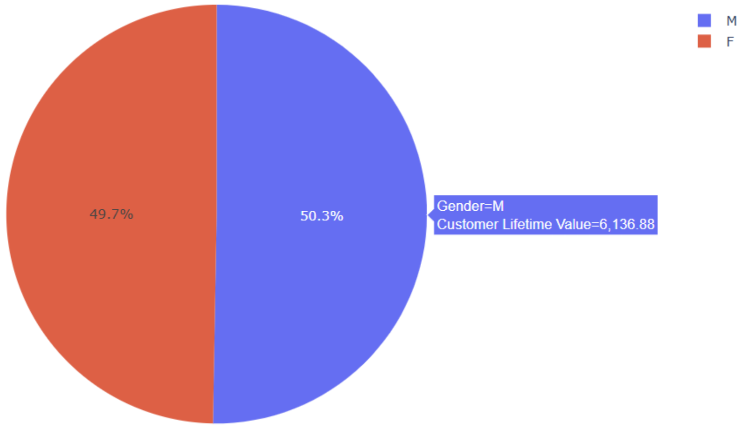


**Figure 6: Pie Chart of average CLV per Gender.**

The data shown in the form of a funnel graph in Figure 7 with different vehicle classes ordered by decreasing average Customer Lifetime Value (CLV). As depicted, Luxury Car have the highest average CLV of 9884.47 and Luxury SUV car owners following behind with 9439.68. Globally it is observed that luxury segments tend to have more valuable customers. Towards the lower end of the spectrum, it has displayed Four-Door Car and Two-Door Car owners shows mean CLVs equivalent to 5475.12 and 5457.32 respectively.

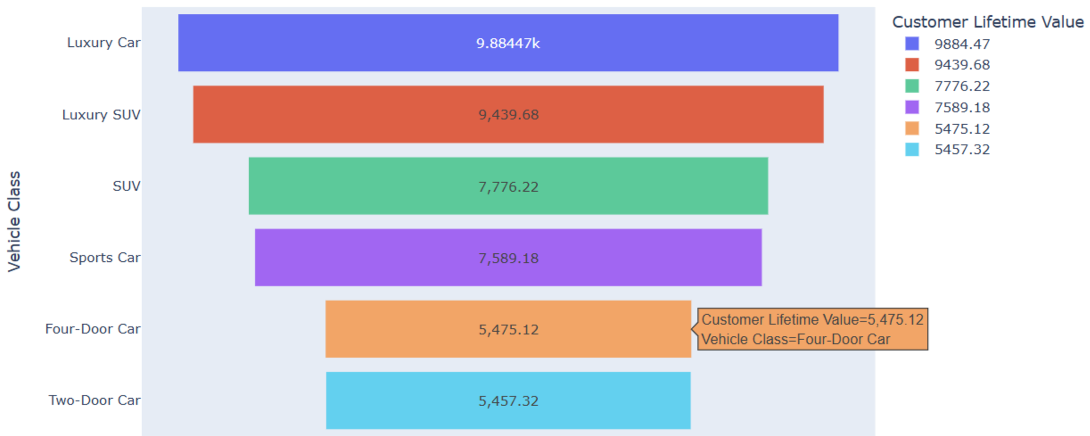


**Figure 7: Funnel Chart showing average CLV per Vehicle Class.**

As displayed in the scatter plot in Figure 8, Customer Lifetime Value (CLV), Total Claim Amount and Number of Policies are the three variables of interest with each dot representing a single customer. The x-axis extends across the CLV from about 2000 up to 14000 and the y-axis represents Total Claim Amount from about 0 to nearly 3000. The scatter plot ranged from dark blue to yellow, which means the Number of Policies purchased by each customer where the dense area with a smaller number of policies is dark blue and the dense yellow represents a greater number of policies. The visualization reveals that CLV grows along with

Total Claim Amount, while the customers who have more policies are in the mid-CLV range (6000-12000).

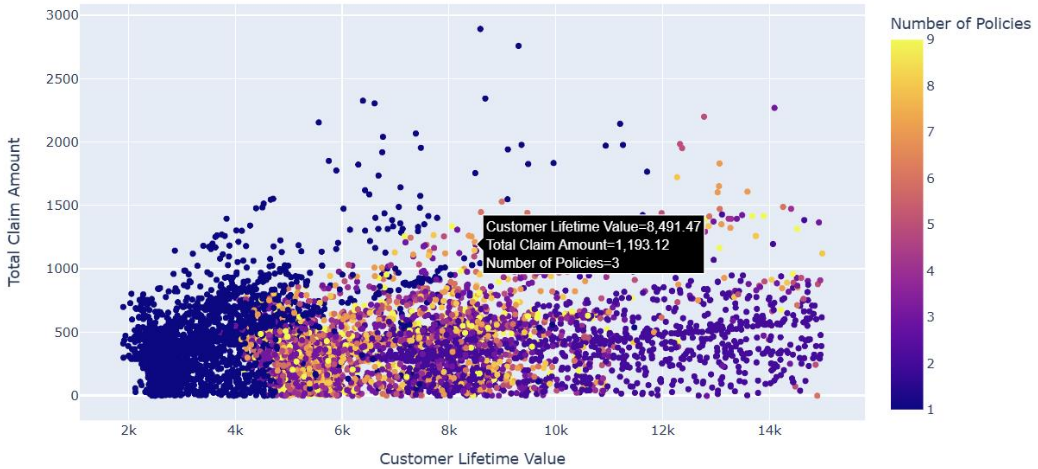


**Figure 8: Scatter Plot showing CLV based on Total Claim Amount and Number of Policies.**

The correlation matrix as seen in Figure 9 shown in the form of a heatmap shows the correlation coefficients between the variables in the dataset. Several notable correlations are a strong negative coefficient between Income and EmploymentStatus, a positive coefficient between Monthly Premium Auto and Total Claim Amount, Coverage and Monthly Premium Auto, and Monthly Premium Auto and CLV. The heatmap also shows a high level of positive interaction between the Policy and Policy Type with a correlation coefficient of 0.88, indicating the level of interaction of these variables.

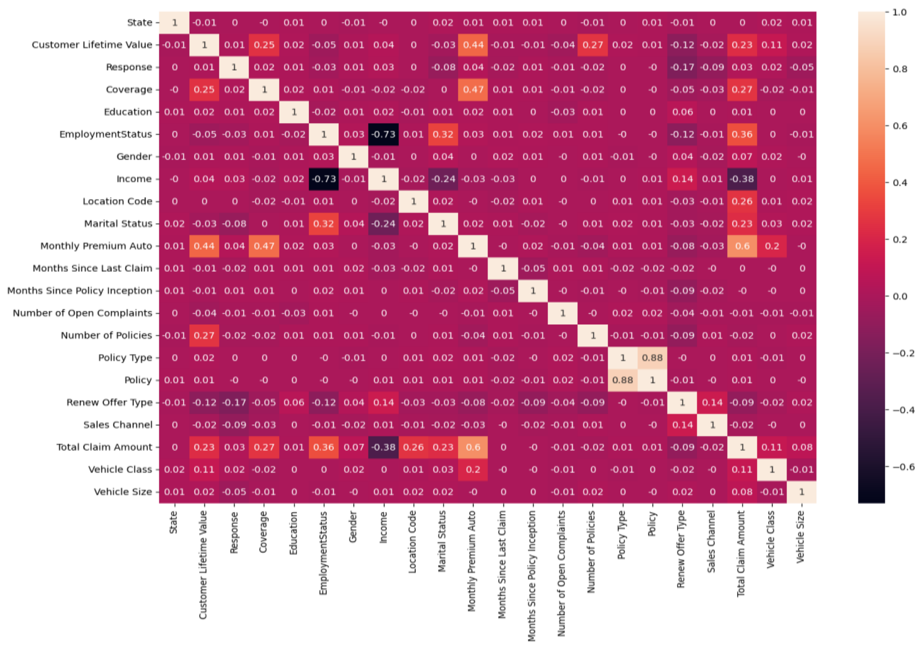


**Figure 9: Correlation Matrix.**

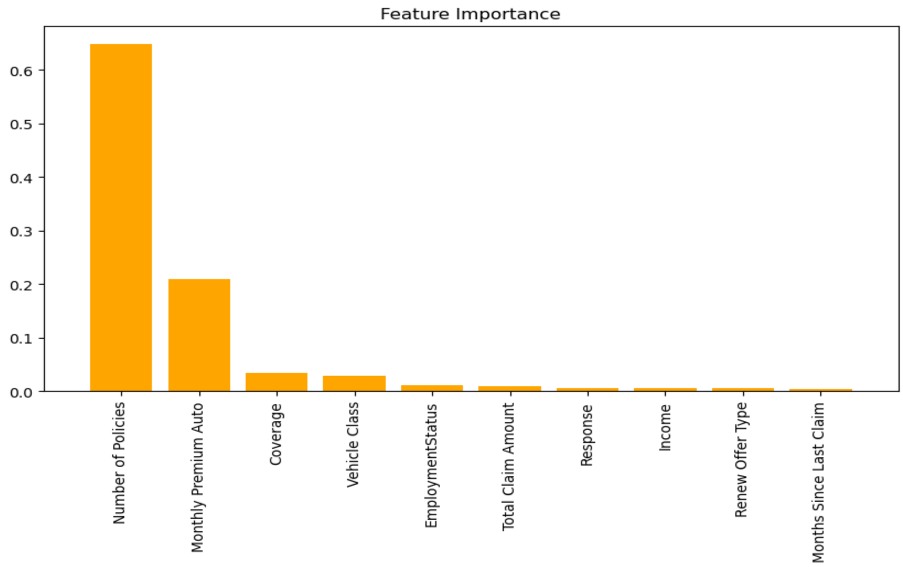


**Figure 10: Feature Importances Chart.**

In Figure 10, there is a bar chart that shows the ten prominent features as highlighted by the ExtraTreesRegressor model with the aid of the feature importance method. The Number of Policies emerge as the most important feature based on the importance score with the highest score of about 0.65. Monthly Premium Auto is ranked as the second important feature of the model with a score approximately reaching 0.2. Vehicle Class and Coverage have tendencies towards moderate importance with values of approximately 0.05 and 0.04. The remaining features, EmploymentStatus, Total Claim Amount, Response, Income, Months Since Last Claim, Response have important estimates less than 0.05. Here, each of the feature's importance score is drawn as an orange bar with descending order, and therefore it is easy to recognize the relative importance of the variables as they relate to the target variable for prediction.

# 4 Design Specification

In this section, the architecture of two deep learning models, GRU LSTM and BiLSTM with Attention Mechanism, is discussed.

## 4.1 GRU-LSTM

The GRU-LSTM is the combination of two forms of neural network structures, Gated Recurrent Unit (GRU) and the Long Short-Term Memory (LSTM). They enhance the model's ability to capture interactions between the input features. This architecture is more beneficial in cases of different types of regression tasks and for the purpose of finding relationship between the variables. The GRU layers are used to conduct fast and effective learning of short-term dependency within the input features. Specifically, GRUs are less complex than LSTMs and are defined by high speed of computation. Every GRU layer contains two basic gating mechanisms, an update gate and a reset gate, to control the flow of input into the model. These gates enable the model to decide before making a conclusion of what parts of the input are critical and should be kept for the next steps. A GRU layer includes these gates to change the model memory according to the new data to enable the model to learn faster. The LSTM layers learn long term dependencies after the GRU layers and LSTMs are a little more complex than GRUs and have different gates as input gate, forget gate and output gate. These gates regulate the cell states' update and memory flow and make LSTMs appropriate for the management of information that is critical in the long-term. The LSTM layer takes the result from the GRU layer to get further information and essential relations in the input. This architecture employs both GRU and LSTM to achieve the best of both layers and allows the model to address all levels of feature interaction and learn from the input features. The last output of the LSTM layers is most often given to a dense layer for the prediction or to extract feature for regression problems. The model is then trained using an optimizer such as RMSprop or ADAM and a loss function based on the type of regression which helps the model in minimizing the error to enhance the predictive capacity of the model during model training. The integration of GRU and LSTM enables the advantages of both GRU and LSTM to make it an efficient architecture for several regression problems.

## 4.2 BiLSTM with Attention Mechanism

The BiLSTM with Attention Mechanism is a combination of bidirectional Long Short-Term Memory (BiLSTM) layers and an attention mechanism to handle the relationships of the input data. This combination improves the capacity of the model to learn complex patterns and to work with the important features of the data. The backward and forward connections of the LSTM layers enable consideration of the dependencies in both directions and enhance the ability of the model to focus on the important parts of these dependencies by using the attention layer. The BiLSTM layer incorporate hidden units which process the data in both directions to get a better picture of the data. In addition to capturing temporal dependencies in the input, BiLSTM layers learn from past context and future context as well. The attention mechanism is then applied to the outputs after the BiLSTM layer. The attention mechanism tries to weigh the hidden state produced by BiLSTM layer to some value, so the model will focus on the specific parts of the input, which is relevant for tasks such as prediction. Through attention mechanism, the focal point is not only capturing important features but also practices more efficient data handling as well. Following this, the features are often fed into a fully connected dense layer for further processing and utilizes the learned features to make predictions. Finally, the optimizer and loss function are chosen for the model compilation depending upon the task needed to be performed. This enables the model to optimize its parametrical features as it trains in order to produce accurate prediction and a high performance.

# 5 Implementation

## 5.1 Machine Learning Models

The various machine learning models that will be used in the study are Decision Tree Regressor, Random Forest Regressor and Gradient Boosting Regressor.

### 5.1.1 Decision Tree Regression

The Decision Tree Regression model is developed by creating an instance of the regressor and using the fit method to train the training dataset. The trained model is later used to predict on the test dataset which is then evaluated with Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and coefficient of determination ($R^2$). Furthermore, an actual/predicted values table is constructed by defining a suitable function for making an assessment for the quality of prediction.

### 5.1.2 Gradient Boosting Regression

The Gradient Boosting Regression is constructed using a specific number of estimators and then the model is trained and is then used to forecast using the test data using evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error and Coefficient of determination ($R^2$). The deviation made in the predictions is assessed by comparing the actual values with the predicted values to evaluate the quality of predictions.

### 5.1.3 Random Forest Regression

A Random Forest Regression model is generated using the RandomForestRegressor and the model is fitted on the training dataset and the predictions are produced on the test data. The model's efficiency is then assessed by metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to present the prediction accuracy and the amount of error produced by the model. The actual and predicted values are compared to assess the model's ability to make accurate predictions.

## 5.2 Deep Learning Models

Deep learning models like GRU LSTM and BiLSTM with the Attention Mechanism are used for predicting Customer Lifetime Value.

### 5.2.1 GRU-LSTM

The GRU LSTM model implemented using keras includes several layers of GRU and LSTM units with dropout layers to minimize overfitting. It is organized in such a way to have more number of units in the latter layers to capture more relations in the data. The model is trained using RMSprop optimizer combined with a mean squared error loss function. Finally, predictions are made using the test data. Thereafter, accuracy of the model is measured using Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared ($R^2$) and an actual vs predicted dataframe is created to check the model's performance.

### 5.2.2 BiLSTM Model with Attention Mechanism

The Bidirectional Long Short-Term Memory (BiLSTM) model with an attention mechanism was developed to improve performance in CLV forecasting. For the implementation of the BiLSTM with Attention Mechanism model, Keras is used. It includes a Bidirectional LSTM layer and an attention layer to enhance the focus on the highly relevant characteristics in the dataset. The added attention layer further improves the capture of relations between the features of the model. The is followed by a compilation of the model with Adam optimizer and mean squared error loss function. The training is carried out on the training data set and predictions are done on the test data for further evaluation.

# 6 Evaluation

The models employed will be compared and tested to distinguish their accuracy by utilizing the performance indicators which are Mean Squared Error (MSE), Root Mean Squared Error (RMSE) as well as the $R^2$ Score for the prediction.

## 6.1 Case Study 1: Decision Tree Regression

Two metrics are used to assess the prediction error for the model which are MSE and RMSE, both average squared and root-squared differences between the predicted and the actual values, which are 0.017 and 0.131 respectively as shown in Table 1. Moreover, given the test data, the $R^2$ Score of 0.67 also indicates that, about 67% of the variance among the data is accounted by this model. To carry out further result interpretation, the actual values of CLV are compared with the predicted ones as seen in Table 2 after scaling them back to their original scale using inverse_transform() method, which is helpful while evaluating the model's suitability and the accuracy to predict CLV values.

**Table 1: Performance Metrics for Decision Tree Regressor Model**

| Metric | Value |
|---|---|
| Mean Squared Error (MSE) | 0.017 |
| Root Mean Squared Error (RMSE) | 0.131 |
| $R^2$ Score | 0.67 |

**Table 2: Comparison of Original and Predicted Customer Lifetime Values for Decision Tree Regressor Model**

| Index | Original Value | Predicted Value |
|---|---|---|
| 0 | 2359.42 | 3208.41 |
| 1 | 2576.38 | 3208.41 |
| 2 | 7688.02 | 9623.95 |
| 3 | 5503.92 | 6409.23 |
| 4 | 3538.06 | 6409.23 |

## 6.2 Case Study 2: Gradient Boosting Regression

The Mean Squared Error (MSE) for the model is equal to 0.009 and the Root Mean Squared Error (RMSE) is 0.094, which minimized the average squared error by comparing the predicted and the actual values, in comparison with previous models. Here, the $R^2$ Score of this model is 0.83 as shown in Table 3 which implies that the model fits 83% of the tested data variability. The rescaled predicted and actual values calculated by using scaler.inverse_transform, made it easier to compare with actual values. The Table 4 shown below helps in making a better comparison between the actual and predicted CLV values to support the claim that Gradient Boosting Regression is a potential method that accurately captures the customer lifetime value.

**Table 3: Performance Metrics for Gradient Boosting Regressor Model**

| **Metrics** | **Value** |
|---|---|
| Mean Squared Error (MSE) | 0.009 |
| Root Mean Squared Error (RMSE) | 0.094 |
| $R^2$ Score | 0.83 |

**Table 4: Comparison of Original and Predicted Customer Lifetime Values Using Gradient Boosting Regressor Model**

| **Index** | **Original Value** | **Predicted Value** |
|---|---|---|
| 0 | 2359.42 | 3613.18 |
| 1 | 2576.38 | 3613.18 |
| 2 | 7688..02 | 7523.99 |
| 3 | 5503.92 | 5643.94 |
| 4 | 3538.06 | 5643.94 |

## 6.3 Case Study 3: Random Forest Regression

As seen in Table 5, the MSE for this model is 0.006 and RMSE is 0.078 which further minimizes the prediction error for this model as compared to other machine learning models. With an $R^2$ Score of 0.88, the model is a good fit and 88% of the variance in the customer lifetime values are considered. To compare real values and the predicted values, they are scaled back and a new DataFrame named pred_df is made alongside the predicted values rounded to two decimal places inorder to determine how well Random Forest Regressor performs in predicting the future customer values. Refer Table 6.

**Table 5: Performance Metrics for Random Forest Regressor Model**

| **Metrics** | **Value** |
|---|---|
| Mean Squared Error (MSE) | 0.006 |
| Root Mean Squared Error (RMSE) | 0.078 |
| $R^2$ Score | 0.88 |

**Table 6: Comparison of Original and Predicted Customer Lifetime Values Using Random Forest Regressor Model**

| **Index** | **Original Value** | **Predicted Value** |
|---|---|---|
| 0 | 2359.42 | 2665.49 |
| 1 | 2576.38 | 2665.49 |
| 2 | 7688.02 | 8750.35 |
| 3 | 5503.92 | 5448.43 |

| | | |
|---|---|---|
| 4 | 3538.06 | 5448.43 |

## 6.4 Case Study 4: GRU-LSTM

After testing the constructed model on the test dataset, the Mean Square Error was around 0.004, while the Root Mean Squared Error (RMSE) was about 0.068 which can prove the accuracy in the model's prediction as seen in Table 7. The value of the $R^2$ is estimated to be 0.91, which gives a good indication of its explanatory variables in variance. The comparison of actual and predicted values as seen in Table 8 shows that the model provided good accuracy in the calculation of the customer lifetime values appropriately.

**Table 7: Performance Metrics for GRU-LSTM Model**

| Metrics | Value |
|---|---|
| Mean Squared Error (MSE) | 0.004 |
| Root Mean Squared Error (RMSE) | 0.068 |
| $R^2$ Score | 0.91 |

**Table 8: Comparison of Original and Predicted Customer Lifetime Values Using GRU-LSTM Model**

| Index | Original Value | Predicted Value |
|---|---|---|
| 0 | 2359.42 | 2713.14 |
| 1 | 2576.38 | 2785.14 |
| 2 | 7688.02 | 8697.45 |
| 3 | 5503.92 | 6201.14 |
| 4 | 3538.06 | 5256.75 |

## 6.5 Case Study 5: BiLSTM Model with Attention Mechanism (Best Model)

The efficiency of the model is evident through the Mean Squared Error (MSE) of 0.003, meaning that the quadrature of the discrepancy between the observed value and the predicted value is very low. The RMSE of 0.054 as shown in Table 9 supports this, which is quite small and desirable for regression tasks as it merely states that normally the error for each prediction is quite small. This MSE and RMSE value shows that the model accurately predicts CLV with minimal deviation from actual values. Moreover, the $R^2$ value equal to 0.94 shows that the proposed model provides quite a good fit, and the model will explain about 94% of the variations in the dependent variable. This means that there is a very good fit of model which implies that model is very good in predicting the underlying relationship and pattern of the predictors using CLV prediction. In other words, the results suggest that the model can explain about 94% of the variation on CLV, which is highly accurate for the CLV prediction. A high $R^2$ value ensures that the model has the capacity of estimating CLV with negligible error, which explains a good understanding of the factors influencing CLV. Table 10 provides a comparison between the actual and the predicted values of this model.

**Table 9: Performance Metrics for BiLSTM Model with Attention Mechanism**

| Metrics | Value |
|---|---|
| Mean Squared Error (MSE) | 0.003 |
| Root Mean Squared Error (RMSE) | 0.054 |
| $R^2$ Score | 0.94 |

**Table 10: Comparison of Original and Predicted Customer Lifetime Values Using BiLSTM Model with Attention Mechanism**

| Index | Original Value | Predicted Value |
|---|---|---|
| 0 | 2359.42 | 2734.87 |
| 1 | 2576.38 | 2841.98 |
| 2 | 7688.02 | 7814.27 |
| 3 | 5503.92 | 5532.80 |
| 4 | 3538.06 | 4860.87 |

## 6.6 Discussion

Among all the machine learning models, the best performing model was the Random Forest Regression with the $R^2$ Score of 0.88 followed by Gradient Boosting Regression having a score of 0.83 and the last of the line was the Decision Tree Regression with a score of 0.67. However, deep learning models provided better results than the machine learning models, GRU-LSTM got 0.91 of $R^2$ Score, and BiLSTM with Attention Mechanism scored the highest with 0.94. The performance differences among the models stem from their ability to handle complexity and nonlinearity in the data. Decision Tree Regression is a simple model which fails to capture complex types of patterns which results in lower accuracy. Random Forest and Gradient Boosting Regression improve upon this with ensemble methods, offering better generalization and reduced overfitting. The use of an attention mechanism with BiLSTM improves the prediction of CLV since it can capture relations in the features that are even non-sequential. Models like Decision Trees, Gradient Boosting and Random Forests fail to find the inherent patterns in the data whereas BiLSTM with attention can do. This ability of the model to focus on the important features and at the same time to identify the insignificant ones makes the model better estimate values for CLV. By using attention mechanism in BiLSTM model, it helps the model to pay more attention to the relationship between predictors and CLV. Also, BiLSTM with attention outperforms GRU-LSTM due to the consideration of past and future context, thus better capturing long-term dependencies and paying attention to the most important features, which greatly improves forecasting precision.

The BiLSTM with Attention Mechanism is effective in capturing both past and future dependencies and also provides importance to critical features which led to higher accuracy and precise Customer Lifetime Value prediction. The BiLSTM with Attention Mechanism model emerged as the best one among all the models because of the advanced architecture. The next improvement is the incorporation of the attention mechanism that allows the model to pay attention to the most representative features that decides customer lifetime value. The high-dimensionality data compatibility and the capability of recognizing nonlinear information make this model suitable for applications such as CLV prediction. Therefore, BiLSTM with Attention Mechanism had the lowest error and highest $R^2$ Score making it the best model for this study.

# 7 Conclusion and Future Work

## 7.1 Conclusion

In conclusion, based on the dataset, the present investigation is able to show that several machine learning and deep learning models work well in predicting outcomes. As for the machine learning models, Decision Tree Regressor, Random Forest Regressor and Gradient Boosting Regressor, all demonstrated a preferable level of MSE and $R^2$ Scores, which elucidate the models' counterparts of handling the data complexities. In addition, for the deep learning models GRU-LSTM and BiLSTM with Attention Mechanism, specifically the BiLSTM with Attention Mechanism, the level of predictability was even higher, and this can be attributed by the high-performance results shown. Proposed techniques included improving mechanical structures that enabled the deep learning model to pay attention to the correct feature improving its performance. Therefore, the paper emphasizes the significance of choosing proper approaches to modeling and emphasizes aspects in line with the data to construct valuable insights that can enhance decision making practices. In general, the results stress the expanding importance of machine learning and deep learning algorithms and models across different domains and their capacity to generate accurate predictive estimates, as well as providing value-added boosts to analytical performance.

## 7.2 Limitations and Future Works

Nevertheless, there are some limitations in the current study. The primary one is the volume and variability of the dataset that limits the generality of the model's findings. Additional data also expands the range of features which might be considered, and a larger and more diverse sample would make the resulting model less sensitive to outliers. Furthermore, the current analysis rests on the assumption of traditional machine learning and some deep learning techniques which nevertheless offer sensible results but may not fully model the richness of data at their deepest. As such, additional research can be carried out on some complex models such as the Transformer based architecture or alternation model that integrates multiple algorithms to get better forecasts. There is also another area in this research that requires improvement on, which concerns feature engineering. The features used in the current study were basic features and no advanced polynomial features or interaction terms were explored deeper into the data as demonstrated in this study. Further, the hyperparameters for the models were selected arbitrarily, though their efficiency can be improved by hyperparameter tuning techniques like GridSearch or Bayesian optimization. Finally, the study could augment its analysis of the models' performances in the context of other different metrics. Thus, future research may improve the knowledge of predictive modeling in this area and lead to further development of improved decision-making that would reflect the priorities of this domain. Implementation of these advanced methodologies and enlargement of dataset using some sequential data will not only create more accurate models but also define the field of predictive analysis noticeably.

# References

Asadi Ejgerdi, N. and Kazerooni, M., 2024. A stacked ensemble learning method for customer lifetime value prediction. *Kybernetes*, *53*(7), pp.2342-2360.

Bauer, J. and Jannach, D., 2021. Improved customer lifetime value prediction with sequence-to-sequence learning and feature-based models. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *15*(5), pp.1-37.

Castéran, H., Meyer-Waarden, L. and Reinartz, W., 2021. Modeling customer lifetime value, retention, and churn. In *Handbook of market research* (pp. 1001-1033). Cham: Springer International Publishing.

Chopra, A., Iyer, P., Joshi, A. and Gupta, M., 2020. Enhancing Customer Lifetime Value Prediction Using Ensemble Learning and Long Short-Term Memory Networks. *International Journal of AI Advancements*, *9*(4).

Devriendt, F., Berrevoets, J. and Verbeke, W., 2021. Why you should stop predicting customer churn and start using uplift models. *Information Sciences*, *548*, pp.497-515.

Ehsanifar, M., Dekamini, F., Spulbar, C., Birau, R., Bajelan, M., Ghadbeykloo, D., Mendon, S. and Calotă, A.M., 2022. Analysing the nexus between artificial neural networks and ARIMA models in predicting customer lifetime value (CLV) for complex development of society and industrial activities. *Industria Textila*, *73*(3), pp.249-258.

Firmansyah, E.B., Machado, M.R. and Moreira, J.L.R., 2024. How can Artificial Intelligence (AI) be used to manage Customer Lifetime Value (CLV)—A systematic literature review. *International Journal of Information Management Data Insights*, *4*(2), p.100279.

Heldt, R., Silveira, C.S. and Luce, F.B., 2021. Predicting customer value per product: From RFM to RFM/P. *Journal of Business Research*, *127*, pp.444-453.

Kumar, V., 2024. How Much Are Your Customers Worth?. In *Valuing Customer Engagement: Strategies to Measure and Maximize Profitability* (pp. 77-115). Cham: Springer Nature Switzerland.

Kumari, D.A., Siddiqui, M.S., Dorbala, R., Megala, R., Rao, K.T.V. and Reddy, N.S., 2024, April. Deep Learning Models for Customer Lifetime Value Prediction in E-commerce. In *2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)* (pp. 227-232). IEEE.

Laksono, F.A., Rachmat, B. and Sutasrso, Y., 2023. B2B customer segmentation based on customer lifetime value concept and RFM modeling. *International Journal of Economics Development Research (IJEDR)*, *4*(3), pp.1856-1874.

Marmol, M., Goyal, A., Copado-Mendez, P.J., Panadero, J. and Juan, A.A., 2021. Maximizing customers' lifetime value using limited marketing resources. *Marketing Intelligence & Planning*, *39*(8), pp.1058-1072.

Myburg, M.E., 2023. Using recency, frequency and monetary variables to predict customer lifetime value with XGBoost.

Sun, Y., Cheng, D., Bandyopadhyay, S. and Xue, W., 2021. Profitable retail customer identification based on a combined prediction strategy of customer lifetime value. *Midwest Social Sciences Journal*, *24*(1), p.10.

Sun, Y., Liu, H. and Gao, Y., 2023. Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Heliyon*, *9*(2).

Xie, S.M. and Huang, C.Y., 2021. Systematic comparisons of customer base prediction accuracy: Pareto/NBD versus neural network. *Asia Pacific Journal of Marketing and Logistics*, *33*(2), pp.472-490.

Yan, Y. and Resnick, N., 2024. A high-performance turnkey system for customer lifetime value prediction in retail brands: Forthcoming in quantitative marketing and economics. *Quantitative Marketing and Economics*, *22*(2), pp.169-192.

Zhao, S., Wu, R., Tao, J., Qu, M., Zhao, M., Fan, C. and Zhao, H., 2023. perCLTV: A general system for personalized customer lifetime value prediction in online games. *ACM Transactions on Information Systems*, *41*(1), pp.1-29.