Capstone Project

Assignment 1

Course code: CSA1635

Course : DATA WAREHOUSING AND DATA MINING FOR DATA SECURITY

S.No: 21

Name :G.Sri Lakshmi Sai

Reg. No: 192211101

Slot :c

Title : Customer Lifetime Value Prediction for Subscription Businesses in data warehousing

Assignment Release Date :

Assignment Preliminary Stage (Assignment 1) submission Date :

Mentor Name : DR.K.Selvakumara swamy

Mentor Phone number and Department : 9486028970 AND RF AND COMMUNICATION SYSTEMS

1.Preliminary Stage

1.1 Assignment Description:

The project aims to develop a Customer Lifetime Value (CLV) prediction model tailored for subscription-based businesses. By analyzing historical customer data, the project seeks to identify patterns and trends that can inform marketing strategies, retention efforts, and revenue forecasting. Through exploratory data analysis (EDA) and machine learning techniques, the goal is to predict the CLV of individual customers, enabling businesses to prioritize resources and tailor personalized experiences. The project will involve data collection from various sources, data preprocessing to ensure quality and consistency, and comprehensive EDA to uncover insights into customer behaviour and preferences.

1.2 Assignment Work Distribution:

• Project Scope Definition:

Define the scope and objectives of the project: The scope involves developing a CLV prediction model for subscription businesses to optimize marketing and retention strategies. Specific goals of analyzing: The project aims to identify key factors influencing CLV and develop predictive models for accurate forecasting.

• Data Collection and Preparation:

Identify the data sources: Sources include customer transaction records, demographic information, and interaction history. Develop a data collection plan: The plan outlines methods for gathering relevant data from internal databases and third-party sources. Cleanse and preprocess the collected data to ensure data quality: Data will undergo cleaning procedures to address missing values, outliers, and inconsistencies. Consistency of the project: Ensuring data consistency across all stages of analysis to produce reliable insights and predictions.

• Exploratory Data Analysis (EDA):

Conduct exploratory data analysis: Analyze customer data to understand patterns, trends, and relationships. Understand the patterns and trends: Identify recurring behaviours, preferences, and characteristics among customers .Perform descriptive statistics, such as summary statistics, distribution plots, and correlation analysis, to explore the relationships of the data: Utilize statistical methods to uncover insights into customer behaviour and preferences.

Visualize the data using charts, graphs: Visual representations such as histograms, scatter plots, and heatmaps will be used to present findings effectively.

2. Problem Statement

The problem statement revolves around optimizing data warehousing systems to facilitate Customer Lifetime Value (CLV) prediction for subscription-based businesses. Subscription businesses rely heavily on customer retention and long-term profitability, making accurate CLV prediction crucial for strategic decision-making. However, traditional data warehousing architectures may struggle to integrate and process the diverse and voluminous data sources required for robust CLV modeling. Challenges include data silos, scalability issues, and inadequate analytical capabilities. Therefore, the problem statement focuses on designing and implementing a data warehousing solution capable of efficiently aggregating, cleansing, and analyzing relevant customer data to predict CLV accurately. This involves integrating transactional, behavioral, and demographic data, employing advanced analytics techniques, and ensuring scalability and flexibility to accommodate evolving business needs. Ultimately, the goal is to empower subscription businesses with actionable insights derived from CLV prediction to optimize marketing, retention, and pricing strategies, thereby maximizing customer lifetime value and overall business performance.

3. Abstract

Customer Lifetime Value (CLV) prediction is a crucial aspect for subscription businesses aiming to optimize customer acquisition and retention strategies. In the context of data warehousing, this study explores methods to leverage historical customer data to predict future CLV. By utilizing advanced data warehousing techniques, such as data integration, cleansing, and transformation, along with predictive modeling algorithms, accurate CLV predictions can be achieved. The process involves aggregating and analyzing diverse data sources, including transaction history, demographic information, and customer interactions, stored within the data warehouse. Through the integration of structured and unstructured data, patterns and trends indicative of customer behavior and purchasing patterns can be identified.

Additionally, by incorporating machine learning algorithms, such as regression analysis and clustering, predictive models can be trained to forecast CLV for individual customers or customer segments. The insights gained from CLV predictions enable subscription businesses to tailor marketing campaigns, optimize pricing strategies, and prioritize customer engagement efforts effectively. Ultimately, leveraging data warehousing for CLV prediction empowers subscription businesses to maximize customer lifetime value and drive sustainable growth in a competitive market landscape.

4.Proposed Design Work

4.1 Identify the Key Components:

In designing a system for Customer Lifetime Value (CLV) prediction for subscription businesses using data warehousing, several key components are essential:

Data Integration Layer: This component involves the extraction, transformation, and loading (ETL) process to gather data from various sources such as customer transactions, interactions, demographics, and subscription details. This layer ensures that data is cleansed, normalized, and transformed into a format suitable for analysis.

Data Warehousing Infrastructure: This component encompasses the storage and management of large volumes of data required for CLV prediction. It includes a robust data warehouse solution capable of handling structured and unstructured data efficiently. Technologies such as Amazon Redshift, Google BigQuery, or Apache Hive may be utilized depending on the scale and requirements.

Feature Engineering: Feature engineering involves selecting and creating relevant attributes or features from the raw data that are predictive of customer lifetime value. These features may include recency, frequency, monetary value (RFM) metrics, customer demographics, purchase history, subscription plans, customer engagement metrics, and any other relevant factors.

Model Development: This component involves the development of predictive models to estimate customer lifetime value. Various machine learning and statistical techniques such as regression analysis, survival analysis, decision trees, random forests, or neural networks may be employed. The models are trained using historical data and validated using appropriate evaluation metrics.

Scalability and Performance Optimization: Ensuring scalability and performance optimization is crucial, especially for large-scale subscription businesses with vast amounts of customer data. This involves designing the architecture to handle increasing data volumes efficiently and optimizing queries and processes for faster execution.

Integration with Business Systems: The CLV prediction system needs to be integrated with existing business systems such as customer relationship management (CRM), marketing automation, and billing systems. This integration enables seamless data flow and allows stakeholders to leverage CLV insights for decision-making and strategy formulation.

4.2 Functionality:

The proposed CLV prediction system should offer the following functionalities:

Data Collection and Integration: Gather data from multiple sources including transactional databases, CRM systems, customer interactions, and subscription details.

Data Preprocessing and Transformation: Cleanse, preprocess, and transform raw data into a format suitable for analysis. This includes handling missing values, outliers, and data inconsistencies.

Feature Engineering: Extract and engineer relevant features from the data that are indicative of customer behavior and preferences.

Model Development and Training: Develop predictive models using machine learning and statistical techniques to estimate customer lifetime value based on historical data.

Model Evaluation and Validation: Evaluate the performance of the predictive models using appropriate metrics and validate their accuracy and robustness.

Scalability and Performance: Ensure that the system is scalable to handle large volumes of data and optimized for performance to provide timely insights.

Integration and Deployment: Integrate the CLV prediction system with existing business systems and deploy it into production for ongoing monitoring and usage by stakeholders.

4.3 Architectural Design:

The architectural design of the CLV prediction system involves the following components and their interactions:

Data Sources: Data is collected from various sources such as transactional databases, CRM systems, marketing platforms, and subscription databases.

ETL Process: The ETL process extracts data from different sources, transforms it into a consistent format, and loads it into the data warehouse.

Data Warehouse: The data warehouse serves as the central repository for storing structured and unstructured data. It provides a scalable and high-performance storage solution for analytical queries.

Feature Engineering Layer: This layer extracts and engineers relevant features from the data required for CLV prediction. It may involve the use of SQL queries, data manipulation techniques, and domain knowledge.

Model Development and Training: Machine learning models are developed and trained using historical data stored in the data warehouse. Various algorithms and techniques are applied to build predictive models for estimating customer lifetime value.

Model Evaluation and Validation: The performance of the predictive models is evaluated using validation techniques such as cross-validation, holdout validation, or time-based validation. The models are validated against historical data to ensure accuracy and reliability.

Integration with Business Systems: The CLV prediction system is integrated with existing business systems such as CRM, marketing automation, and billing systems to enable seamless data flow and utilization of CLV insights.

Monitoring and Maintenance: The system is continuously monitored for performance, data quality issues, and model drift. Regular maintenance and updates are performed to ensure the accuracy and relevance of CLV predictions.

5. UI Design

5.1 Lay out Design:

a) Flexible layout: Design a layout that adapts well to different screen sizes and resolutions to ensure a consistent user experience across various devices, including desktops, tablets, and mobile phones.Utilize responsive design principles to adjust the layout dynamically based on the available screen space.

b) User Friendly: Keep the interface simple and intuitive, with clear navigation and logical flow to guide users through the CLV prediction process.Provide helpful tooltips, hints, or inline guidance to assist users in understanding the purpose and usage of different elements and functionalities.Incorporate feedback mechanisms to allow users to report issues or provide suggestions for improving the interface.

c) Colour Selection: Choose a color scheme that is visually appealing and conducive to user engagement .Ensure sufficient contrast between text and background colors to enhance readability, especially for users with visual impairments. Use colors strategically to highlight important elements or signify different stages or categories within the CLV prediction workflow.

5.2 Feasible Elements used:

a) Elements Positioning: Arrange interface elements such as input fields, dropdown menus, buttons, and result displays in a logical and organized manner to facilitate smooth interaction.

Group related elements together and use whitespace effectively to avoid clutter and improve readability.

Maintain consistency in the positioning of common elements across different sections of the interface to reduce cognitive load on users.

b) Accessibility: Ensure that the UI is accessible to users with disabilities by adhering to accessibility standards such as WCAG (Web Content Accessibility Guidelines). Provide alternative text descriptions for non-text elements such as images or icons to assist users who rely on screen readers. Enable keyboard navigation and ensure that all interactive elements are reachable and operable using keyboard controls alone.

5.3 Elements and Functions:

In the context of CLV prediction for subscription businesses, the UI should offer the following elements and functions:

Input Fields: Allow users to input relevant data such as customer demographics, purchase history, and interaction data.

Dropdown Menus: Provide options for selecting different CLV prediction models or parameters.

Buttons: Enable users to initiate the CLV prediction process, reset input fields, or navigate between different sections of the interface.

Result Displays: Present the predicted CLV values along with any additional insights or recommendations in a clear and comprehensible format.

Visualizations: Incorporate charts or graphs to visualize trends in CLV over time or compare CLV across different customer segments.

Export Functionality: Offer the ability to export CLV prediction results or reports in commonly used formats such as CSV or PDF for further analysis or sharing.

6. Login Template

```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <title>Login - Customer Lifetime Value Prediction System</title>
  <link rel="stylesheet" href="styles.css"> <!-- You can link your custom styles here -->
</head>
<body>
  <div class="container">
    <h2>Login</h2>
    <form action="authenticate.php" method="POST">
      <div class="form-group">
        <label for="username">Username:</label>
        <input type="text" id="username" name="username" required>
      </div>
      <div class="form-group">
        <label for="password">Password:</label>
        <input type="password" id="password" name="password" required>
      </div>
      <button type="submit">Login
```

6.1 Login Process:

The login process for a Customer Lifetime Value Prediction system in a data warehousing environment typically involves several steps to ensure secure access to the system. Below is an outline of the login process:

1. **User Authentication**:

   - Users provide their credentials (username and password) on the login page.

   - The system verifies the credentials against stored user data in the database.

2. **Role-Based Access Control (RBAC)**:

   - Once authenticated, the system checks the user's role and permissions.

   - RBAC ensures that users only have access to the features and data they are authorized to use.

3. **Session Management**:

   - Upon successful authentication, the system creates a session for the user.

   - A session token is generated and stored either in a cookie or in the server's memory.

   - This session token is used to identify the user's session and maintain their authenticated state throughout their interaction with the system.

4. **Security Measures**:

   - The login process should include security measures like encryption (e.g., SSL/TLS) to protect the transmission of user credentials over the network.

   - Passwords should be stored securely using techniques like hashing with salt to prevent unauthorized access even if the database is compromised.

5. **Multi-Factor Authentication (Optional)**:

   - For additional security, multi-factor authentication (MFA) can be implemented.

   - Users may need to provide a second form of verification, such as a temporary code sent to their mobile device, in addition to their password.

6. **Logging and Monitoring**:

   - All login attempts (successful or failed) should be logged for auditing purposes.

   - Monitoring tools can be used to detect and alert on any unusual login activities, such as multiple failed login attempts or login attempts from unusual locations.

7. **User Interface Feedback**:

   - Provide clear feedback to users during the login process, indicating whether their credentials were correct or if there was an error.

- If there are errors, provide helpful messages to guide users on how to correct them (e.g., "Invalid username or password").

8. **Forgot Password/Username**:

- Include functionality for users to recover their password or username in case they forget it.

- This often involves sending a password reset link to the user's email address or asking security questions to verify their identity.

By following these steps, the login process for the Customer Lifetime Value Prediction system can be made secure and user-friendly, ensuring that only authorized users can access the system and its data.

6.2 Sign Up Process:

The signup process for Customer Lifetime Value (CLV) Prediction for Subscription Businesses within a data warehousing context typically involves several steps to ensure the accuracy and reliability of the predictions. Here's a generalized outline of such a signup process:

1. **Data Collection and Integration**:

- Gather all relevant data from various sources such as customer databases, subscription details, transaction histories, website interactions, demographics, etc.

- Integrate this data into a centralized data warehousing system. This step might involve data cleaning, transformation, and normalization to ensure consistency and accuracy.

2. **Define CLV Metrics**:

- Determine which metrics will be used to calculate CLV. This could include metrics such as customer churn rate, customer acquisition cost (CAC), average revenue per user (ARPU), customer lifetime, etc.

3. **Data Preprocessing**:

- Perform preprocessing steps such as handling missing values, outlier detection and treatment, feature scaling, and feature engineering to prepare the data for modeling.

4. **Model Selection**:

- Choose appropriate machine learning or statistical models for CLV prediction. Commonly used models include regression-based models, survival analysis, RFM (Recency, Frequency, Monetary) models, and machine learning algorithms such as gradient boosting machines or neural networks.

5. **Model Training**:

- Split the dataset into training and validation sets.

- Train the selected models on the training data while tuning hyperparameters to optimize model performance.

- Validate models using the validation dataset to ensure they generalize well to unseen data.

6. **Model Evaluation**:

- Evaluate the trained models using appropriate evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or R-squared (R^2) to assess their performance.

7. **Deployment**:

- Once a satisfactory model is obtained, deploy it into production within the data warehousing environment. This might involve integrating the model into existing workflows or systems used for customer management and analytics.

8. **Monitoring and Maintenance**:

- Continuously monitor the performance of the deployed model in production to ensure it remains accurate and up-to-date.

- Periodically retrain the model using fresh data to adapt to changing patterns and trends in customer behavior.

9. **User Onboarding and Support**:

- Provide documentation and training to users who will be utilizing the CLV predictions.

- Offer ongoing support to address any issues or questions that may arise during the usage of the CLV prediction system.

10. **Feedback Loop**:

- Establish a feedback loop where insights from the CLV predictions are used to improve business strategies and inform decision-making processes, thus closing the loop for continuous improvement.

Throughout this process, it's essential to maintain data privacy and security measures to safeguard sensitive customer information and comply with relevant regulations such as GDPR or CCPA. Additionally, involving stakeholders from various departments such as marketing, sales, and finance can help ensure the CLV predictions align with business objectives and are effectively utilized across the organization.

6.3 Other Templates:

In the realm of data warehousing, leveraging customer lifetime value (CLV) prediction for subscription businesses is paramount. By utilizing data warehousing capabilities, subscription businesses can analyze historical customer data to forecast their CLV effectively. CLV prediction involves utilizing various data points such as customer demographics, purchasing behavior, subscription tenure, and transaction history. Through advanced analytics and machine learning algorithms, subscription businesses can derive insights into customer behavior patterns, segment customers based on their value to the business, and predict future revenue streams. Data warehousing facilitates the storage, management, and analysis of large volumes of customer data, enabling subscription businesses to make data-driven decisions to optimize customer acquisition, retention, and monetization strategies. By integrating CLV prediction models into their data warehousing infrastructure, subscription businesses can enhance customer relationship management, personalize marketing efforts, and ultimately maximize profitability.

Conclusion:

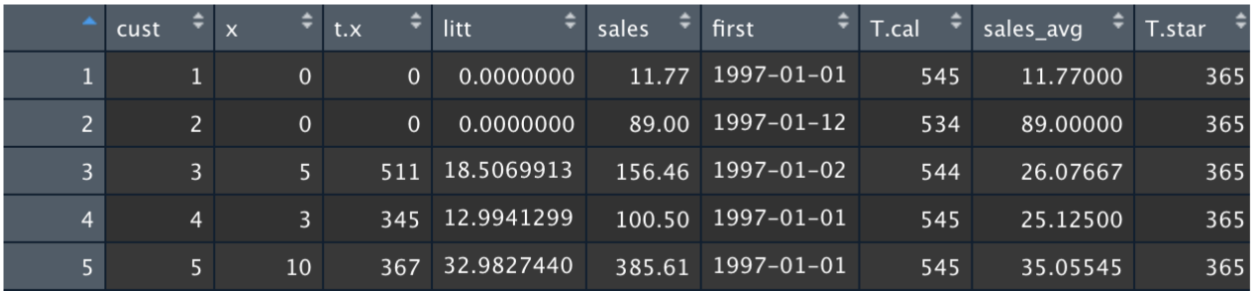
In conclusion , leveraging Customer Lifetime Value (CLV) prediction within the framework of data warehousing offers substantial benefits for subscription businesses. By harnessing the power of data warehousing , organizations can aggregate and analyze vast amounts of customer data, encompassing subscription usage patterns, purchase history, and engagement metrics. Through sophisticated analytics and predictive modelling techniques, such as machine learning algorithms, businesses can forecast the CLV of individual customers. This predictive capability enables subscription-based enterprises to tailor their marketing strategies, optimize customer acquisition efforts, and personalize customer experiences to maximize long-term profitability. Moreover, data warehousing facilitates the integration of various data sources, providing a holistic view of customer interactions across different touch points. This comprehensive understanding empowers businesses to make data-driven decisions, allocate resources effectively, and cultivate enduring customer relationships. Ultimately, by incorporating CLV prediction into their data warehousing initiatives, subscription businesses can unlock valuable insights to drive sustainable growth and enhance customer lifetime value.

R PROGRAMMING:

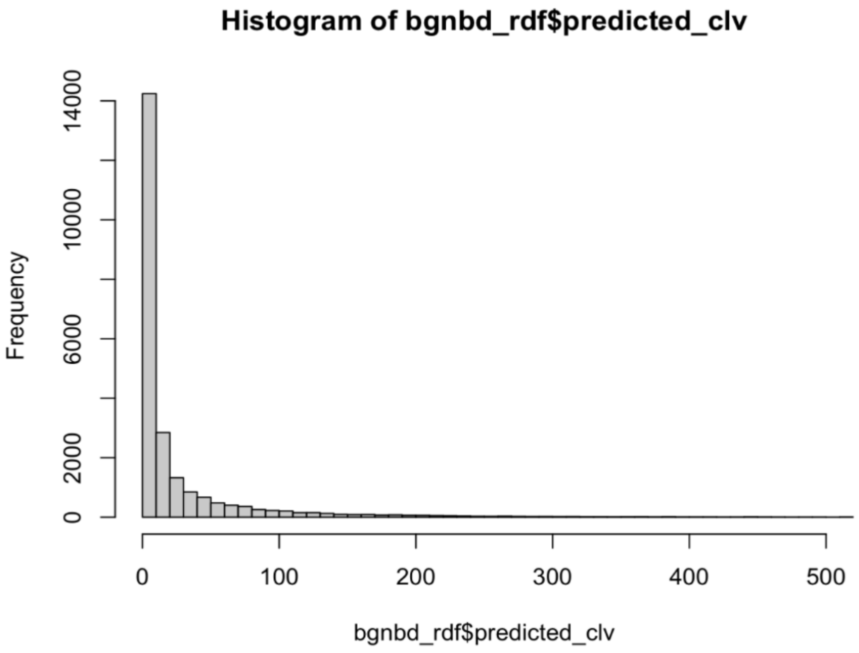
```
# Load required libraries
library(dplyr)

# Sample data with customer information
customer_data <- data.frame(
  customer_id = 1:100,
  tenure = round(runif(100, 1, 60)),  # Random tenure between 1 and 60 months
  monthly_spend = rnorm(100, 50, 10),  # Monthly spend with mean 50 and SD 10
  churned = sample(c(0, 1), 100, replace = TRUE, prob = c(0.8, 0.2))  # Simulated churn (0/1)
)

# Calculate total spend per customer
customer_data <- customer_data %>%
  mutate(total_spend = monthly_spend * tenure)

# Linear regression model to predict CLV
clv_model <- lm(total_spend ~ tenure + monthly_spend, data = customer_data)

# Summary of the model
summary(clv_model)

# Predict CLV for new customers
new_customers <- data.frame(
  tenure = c(12, 24, 36),  # Example tenure values for new customers
  monthly_spend = c(40, 60, 80)  # Example monthly spend values for new customers
)
```

```
# Predict CLV for new customers using the model

predicted_clv <- predict(clv_model, newdata = new_customers)


# Output predicted CLV

print(predicted_clv)
```

OUTPUT:



	cust	x	t.x	litt	sales	first	T.cal	sales_avg	T.star
1	1	0	0	0.0000000	11.77	1997-01-01	545	11.77000	365
2	2	0	0	0.0000000	89.00	1997-01-12	534	89.00000	365
3	3	5	511	18.5069913	156.46	1997-01-02	544	26.07667	365
4	4	3	345	12.9941299	100.50	1997-01-01	545	25.12500	365
5	5	10	367	32.9827440	385.61	1997-01-01	545	35.05545	365



Histogram of bgnbd_rdf$predicted_clv

density.default(x = data_tbl$predicted_spend_pggg)