DAB303 – Marketing Analytics– Project 4: Customer Lifecycle Value (CLV)

Introduction

The purpose of the lab is to understand and gain insights from a retailer dataset, by performing various exploratory data analysis, data visualization, and data modelling tasks, aiming to investigate and analyze product analytics and Customer Lifecycle Value (CLV).

Data:

The provided dataset, made available during the lab, contains all the information needed for the project.

Tasks:

1. **Download and load data file** – as described below:
   - Download the dataset (in .csv file format) from Blackboard.

2. **Use Python Tools for developing the desired model**:
   You need to develop the needed code, based on similar examples and lab exercises. Here is a suggested structure for your Jupyter Notebook:
   1. Introduction
   2. Load the needed libraries
   3. Import the dataset as a Pandas Dataframe, followed by data pre-processing and data cleaning:
      a. Create 'AvgPurchaseValue': Average value of purchases made by the customer (totalrevenue/totalpurchases)
      b. Create 'Recency': Number of days since the last purchase
      c. Create 'Tenure': Number of days since the customer's first purchase (useful for understanding how long they've been a customer).
      d. Create 'AvgTimeBetweenPurchases': Average number of days between purchases
   4. Perform exploratory data analysis (EDA):
      a. Prelininary EDA:
         i. Load the provided dataset into a Pandas DataFrame. Display the first 5 rows of the DataFrame to understand its structure.
         ii. Generate a report detailing the data types of each column in the dataset. How many of them are categorical? How many are numerical?
         iii. Compute and display basic statistical summaries (mean, median, standard deviation, etc.) for all the numerical columns in the dataset.
         iv. Identify any missing values in the dataset. How many missing values are there in each column?
      b. Univariate Analysis:
         i. Plot histograms for all numerical columns in the dataset. What distributions can you identify based on the histograms?
         ii. For each categorical column, generate bar plots depicting the frequency of each category. Which category dominates in each column?
      c. Bivariate analysis:

i. Construct a correlation matrix for all the numerical features in the dataset. Which pairs of features are highly correlated? Are there any unexpected correlations?
ii. 2. Use box plots to compare the distribution of a chosen numerical column (e.g., TotalRevenue) across different categories in the dataset. Which category tends to have higher values for the chosen numerical column?

d. Advanced analysis:
   i. Use a pair plot to visualize relationships between a select set of columns (totalpurchases, totalrevenue, avgpurchasevalue, frequency). Can you identify any clusters or outliers from the plots?
   ii. Analyze the total revenue trend based on the year of the first purchase. In which year did customers contribute the most to total revenue?
   iii. Use the RFM (Recency, Frequency, Monetary) model to identify the top 5% of customers. List their customerid and associated RFM_Score.
e. Multivariate analysis:
   i. Create a scatter plot of totalrevenue vs. frequency and color the data points based on the churnindicator. What patterns do you observe concerning customer churn?
   ii. Use the parallel coordinates plot to visualize multi-dimensional relationships using the columns totalpurchases, totalrevenue, avgpurchasevalue, frequency, and hasloyaltycard. Do any patterns emerge based on loyalty card holders?

5. CLV Modeling:
   a. Ridge regression:
      i. Load the dataset and split it into training and testing sets, keeping 20% of the data for testing.
      ii. Implement a Ridge Regression model using the provided features (`Recency`, `Frequency`, `AvgPurchaseValue`) to predict the `TotalRevenue`.
      iii. Set the alpha parameter for Ridge Regression to 1.0. How does this value affect the coefficients of the model?
      iv. Evaluate the model using Mean Squared Error (MSE) on the test set. Report the obtained value.
   b. Random Forest Regressor:
      i. Implement a Random Forest Regressor with 100 trees to predict the CLV.
      ii. Using the feature importance attribute of the Random Forest model, list the features in order of their importance.
      iii. Evaluate the model's performance using the test set. How does it compare to the Ridge Regression model?
   c. XGBoost:
      i. Implement the XGBoost regressor to predict the CLV. Use 100 estimators for the model.
      ii. XGBoost offers various hyperparameters to tune. Alter the learning rate of the model. How does it impact the model's performance?
      iii. Evaluate the model using the test data and compare its MSE with previous models.
   d. Advanced Regression Model:

   i. Train other regression models like Ridge, Lasso, Decision Trees, Random Forest, and Gradient Boosting to predict totalrevenue.
   ii. Use cross-validation for model selection and tuning.
   iii. Evaluate the models using the same metrics as before and compare their performances.

e. [OPTIONAL] Hyperparameter Tuning:
   i. For models that have hyperparameters, use techniques like GridSearchCV or RandomizedSearchCV to find optimal values.
   ii. Re-evaluate the models using the optimized hyperparameters.

f. Feature Importance:
   i. For tree-based models like Random Forest and Gradient Boosting, extract feature importance scores.
   ii. Analyze and interpret the top features affecting totalrevenue.

g. Model Interpretation:
   i. Use techniques like SHAP (SHapley Additive exPlanations) to explain model predictions.

6. Conclusions – Suggestions.
   You may use additional techniques which may not be listed above, provided that you can submit a rationale for why the technique is useful and an indication of what you hope to achieve.

3. **Report –** In a separate word document:
   - Record your observations with respect to the most important outputs of the Python code.

Submission – Deliverables

Submission will be done via Blackboard, and it will be group submission, including:

- One file per group (in .zip format):
  - Jupyter Notebook (Including extended code commenting and analytical block code description):
    - Lab file (.ipynb)
    - Exported Jupyter notebook in html (.html)
  - Report (.pdf): Include the major steps and finding of your analysis, and
  - Presentation (.pptx): 4 – 5 slides (excluding covers and introduction), for presenting your findings to the management.