Kritik Assignment 2 Topic A:

Decision Tree Analysis for Organic Products

**Objective:**

You are tasked with conducting a decision tree analysis on the **Organics** dataset. Your goal is to predict **TargetAmount**, which represents the amount customers spent on organic products. Use the CRISP-DM fram ework (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment) throughout your analysis. This analysis will help the company identify key drivers of customer spending and optimize promotional strategies for increased sales.

**Business Question:**

**What customer characteristics and customer behaviors are most predictive of the amount spent on organic products in the loyalty program?**

**Dataset:**

The dataset includes the target variable, **TargetAmount**, indicating customer loyalty to organic products, along with nine predictor variables representing customer demographics, purchasing behaviors, and other relevant attributes. The dataset is formatted as a JMP file or Excel file.

**Data Dictionary:**

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Type | Sample Value | Description |
| ID | ID | 140 | Unique customer identifier |
| DemAffl | Input | 10 | Affluence score (higher values indicate more affluence) |
| DemAge | Input | 76 | Age of the customer |
| DemCluster | Rejected | 16 | Cluster number based on demographic segmentation |
| DemClusterGroup | Input | C | Demographic cluster group (A-F) |
| DemGender | Input | U | Gender (M = Male, F = Female, U = Unknown) |
| DemReg | Input | Midlands | Region where the customer is located |
| DemTVReg | Input | Wales & West | TV region for targeted advertising |
| PromClass | Input | Gold | Promotion classification (Gold, Silver, Tin, etc.) |
| PromSpend | Input | 16000 | Amount spent on promotions |
| PromTime | Input | 4 | Time spent in promotion (e.g., duration in days or weeks) |
| TargetBuy (not used for this assignment) | Rejected | 0 | Target binary variable indicating whether a purchase was made (1 = Yes, 0 = No) |
| TargetAmt | Target | 0 | Target amount spent if purchase was made |

**Instructions:**

1. **Business Understanding**  
   Describe the importance of customer loyalty to organic products for the supermarket. Why would we want to predict an actual amount instead of whether they just bought organic products or not?
2. **Data Understanding**
   * Load the dataset and provide an overview of the variables, including the target variable **TargetAmount** and the nine predictors. Identify potential outliers or anomalies that could impact the analysis.
   * Summarize each predictor variable and the target variable in terms of type (categorical or continuous) and determine if it is a *Demographic* or *Promotional* variable.
   * Visualize the target variable to show the distribution of purchase amount of organic products.
3. **Data Preparation**
   * **Partitioning the Data**: Use 70% of the data for training and 30% for testing. Ensure that you use **12345** as the seed for the partitioning.
   * Perform **any necessary data cleaning** (e.g., handling missing data, outliers, or encoding categorical variables) before modeling.
   * Provide summary statistics and visualizations for each predictor variable.
4. **Modeling**

You will conduct 2 different decision tree models for this assignment.

1. Full Variable Model (Unpruned Decision Tree)
   1. **Objective**: Build a decision tree model using all the available features without performing any pre-selection or pruning of variables.
   2. **Why**: This model will allow you to see how the decision tree algorithm utilizes all features to predict the target variable (**TargetAmount**) and understand the relationships between them. It provides insight into how various factors interact and contribute to customer spending.
   3. **Steps**:
      1. Use **all available features** (demographic and promotional data).
      2. Do **not prune** the tree, so it may grow deep and complex.
      3. Evaluate the performance based on metrics like **Mean Squared Error (MSE), R-squared,** etc.
2. Manually Adjusted Decision Tree (Pruned with Business Insights)
   1. **Objective**: Manually adjust the decision tree by pruning certain branches based on domain knowledge or limiting the depth of the tree to avoid overfitting and ensure the model is interpretable for business users.
   2. **Why**: This model focuses on interpretability and can be adjusted to focus on the most important factors contributing to the prediction. It allows you to prevent overfitting while still using valuable features that align with business logic (e.g., promotional spending, customer age, etc.).
   3. **Steps**:
      1. Start with the full variable model and apply pruning techniques such as **limiting tree depth** or **setting a minimum number of samples per leaf**.
      2. Focus the model on features that make business sense.
      3. Evaluate the performance using **MSE**, **RMSE**, and **R²**, and compare the results with the full model.
3. Feature-Selected Decision Tree (Variable Selection Prior to Tree)
   1. **Objective**: Perform feature selection before training the decision tree by identifying the most important variables using techniques like **feature importance** or **recursive feature elimination** (RFE).
   2. **Why**: This model helps eliminate redundant or irrelevant features, simplifying the model and improving generalization. It can also increase model interpretability and performance by focusing only on the most influential factors.
   3. **Steps**:
      1. Use feature selection methods such as **Random Forest Feature Importance** or **Correlation Analysis** to select the most influential predictors (e.g., key demographics or promotional features).
      2. Train the decision tree using only the selected features.
      3. Evaluate the model and compare its performance with the full variable model in terms of **MSE**, **RMSE**, and **R²**.
4. Model Comparison:
   1. Full Variable Model will give you an understanding of how the algorithm treats all variables without restriction.
   2. Manually Adjusted Decision Tree ensures the model is more business-friendly by pruning and focusing on key factors.
   3. Feature-Selected Decision Tree improves accuracy and performance by removing unnecessary features and focusing on the most important ones.
5. **Evaluation**
   * **Model Metrics**: Evaluate the performance of each model (full variable model, manually adjusted model, and selected variable model) using appropriate metrics like:
     + Mean Squared Error (MSE)
     + Root Mean Squared Error (RMSE)
     + R-squared (R²)
     + Mean Absolute Error (MAE)
   * **Compare and contrast the metrics**: Discuss how each model performed and the trade-offs involved in predicting a continuous outcome.
6. **Scoring**
   * Use your best model to predict the **TargetAmount** for **5 new cases** (individuals not included in the training data) and present the results.
7. **Reporting**
   * **Decision Rules**: Provide clear, actionable decision rules based on the model. These rules should help the company decide how to target different customer segments for organic product promotions.
   * Discuss how **predicting spending behavior** aligns with the business objectives and what factors are the most significant predictors of spending.
   * Prepare a concise report of your analysis based on the CRISP-DM phase and your findings.

**Submission Guidelines:**

* Structure your submission according to the CRISP-DM framework outlined above.
* Your report should include visuals and interpretations, with a clear explanation of your data preparation and analysis steps.
* Submit your report as a PDF or Word document with relevant visualizations and explanations.

**Evaluation Criteria:**

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Description** | **Points** |
| **1. Business Understanding & Question** | - Clear articulation of the business objective (predicting customer spending on organic products). - Well-defined and actionable business question. | 2 |
| **2. Data Preparation & Cleaning** | - Data partitioning (70% training, 30% testing) using seed 12345. - Proper handling of missing values, outliers, and encoding of categorical variables. - Justification for any cleaning steps taken. | 3 |
| **4. Model Implementation & Performance** | - Accurate implementation of decision tree models using appropriate software/tools. - Performance of the models using metrics such as MSE, RMSE, and R². - Correct interpretation of the results based on the evaluation metrics. | 5 |
| **5. Model Comparison** | - Clear comparison of the different models (full variable model, pruned model, feature-selected model). - Discuss the trade-offs between model complexity, interpretability, and performance. | 5 |
| **6. Prediction & Business Interpretation** | - Predictions for 5 new cases using the best-performing model. - Clear explanation of how the predictions contribute to the business question. - Actionable insights and recommendations for marketing or customer targeting. | 5 |
| **7. Report Clarity & Presentation** | - Well-organized, clear, and concise report. - Visual aids such as decision tree diagrams, tables of performance metrics, and graphs to support findings. - Proper formatting, grammar, and adherence to any guidelines. | 5 |