Master of Data Science

Faculty of Computer Science & Information Technology

WQD7005 – Data Mining

Instructor: Dr Teh Ying Wah

ASSESSMENT 1

Name	Matric No
Vickneswary Perumal	S2150313

Introduction

Data Source: https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis

Dataset Overview:

The "E-commerce Customer Behavior and Purchase Dataset" is a fabricated dataset generated through the use of the Faker Python library. It emulates a comprehensive e-commerce setting, encompassing diverse facets of customer behavior and purchase history within a digital marketplace. This dataset has been specifically crafted for both data analysis and predictive modeling applications in the realm of e-commerce. It proves to be well-suited for tasks such as forecasting customer churn, conducting market basket analysis, implementing recommendation systems, and performing trend analysis.

Details of the Columns:

This dataset encompasses the following columns:

Customer ID: A distinct identifier assigned to each customer.

Customer Name: The customer's name, generated using the Faker library.

Customer Age: The age of the customer, simulated with Faker.

Gender: The gender of the customer, generated using Faker.

Purchase Date: The date of each customer's purchase.

Product Category: The category or type of the purchased product.

Product Price: The cost of the purchased product.

Quantity: The quantity of the product purchased.

Total Purchase Amount: The overall amount spent by the customer in each transaction.

Payment Method: The method of payment employed by the customer (e.g., credit card, PayPal).

Returns: An indicator of whether the customer returned any products from the order (binary: 0 for no return, 1 for return).

Churn: A binary column denoting whether the customer has churned (0 for retained, 1 for churned).

Objective:

Understanding Purchase Patterns:

Analyze the dataset to identify trends in customer purchasing behavior. This includes popular product categories, peak purchasing times, and the average amount spent per transaction. Insights into these patterns can guide inventory management and marketing strategies.

Customer Segmentation:

Group customers based on common characteristics such as age, gender, and purchasing frequency. This segmentation helps tailor marketing campaigns and promotions to specific customer segments, enhancing the relevance of communication.

Churn Prediction:

Leverage the "Churn" column to predict and understand customer churn. Identify factors leading to customer attrition, such as dissatisfaction or changing preferences. Implement targeted retention strategies to reduce churn and foster customer loyalty.

Payment Method Preferences:

Examine the preferred payment methods among customers. This insight can influence payment processing strategies, partnerships with payment providers, and the development of secure and convenient payment options.

Return Analysis:

Investigate the "Returns" column to understand the frequency and reasons for product returns. Insights into return patterns can guide improvements in product quality, customer service, and overall customer satisfaction.

Personalization Opportunities:

Utilize customer names, ages, and genders for personalized marketing initiatives. Personalization enhances the customer experience and fosters a sense of connection with the brand.

Recommendation Systems:

Explore the potential for implementing recommendation systems based on historical purchase data. Recommending relevant products to customers can increase cross-selling and upselling opportunities.

Trend Analysis:

Identify emerging trends in product categories, customer preferences, and market demand. Stay ahead of industry trends to proactively adapt business strategies and offerings.

Optimizing Marketing Channels:

Evaluate the effectiveness of different marketing channels. Focus efforts on channels that generate the highest customer engagement and return on investment.

Customer Engagement Strategies:

Develop strategies to enhance customer engagement, such as loyalty programs, exclusive promotions, or interactive content. Engaged customers are more likely to remain loyal and contribute to the business's success.

By delving into these aspects of customer behavior, businesses can formulate a well-informed business strategy that addresses customer needs, enhances satisfaction, and maximizes profitability. The goal is to create a customer-centric approach that fosters long-term relationships and sustainable business growth.

Talend Data Preparation

There are several data cleaning steps taken in Talend Data Preparation

- Selecting Rows with Missing Cells in Total Purchase Amount:

The first step identifies and isolates rows where there's incomplete data in the "Total Purchase Amount" column. This aims to address missing values before further analysis.

- Finding and Grouping Similar Text on Payment Method:

This step involves analyzing the "Payment Method" column to detect and group together similar variations in text entries. This is to standardize payment method names for consistency or categorization.

- Changing to Title Case on Payment Method:

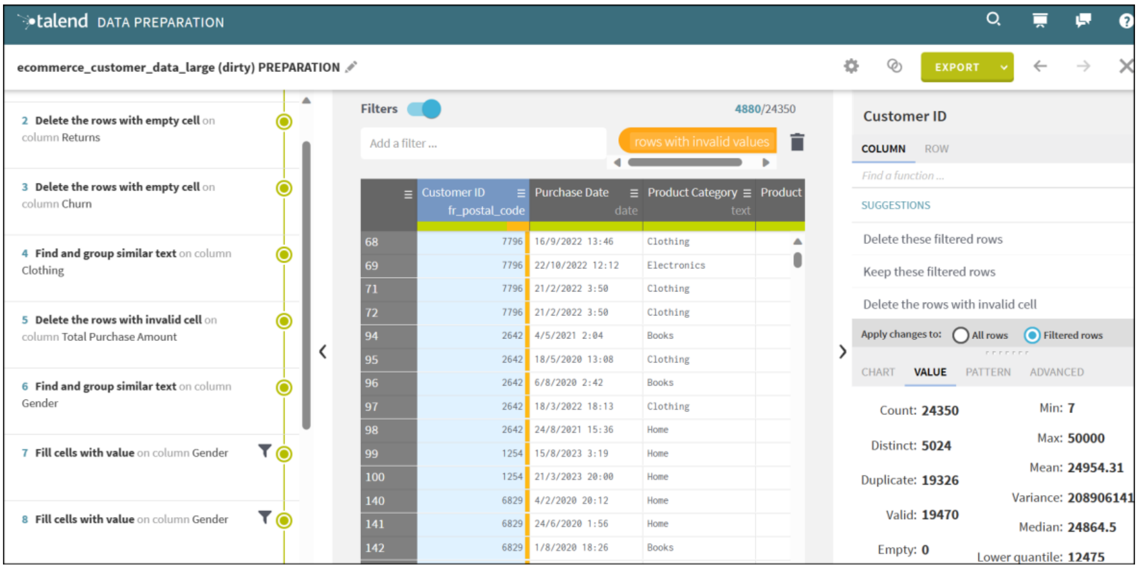
The third step involves converting the text in the "Payment Method" column to title case (e.g., "CASH" becomes "Cash"). This aims to improve readability and standardization.

- Filling Cells with Value on Gender:

This step addresses missing values in the "Gender" column by filling them with specific values.

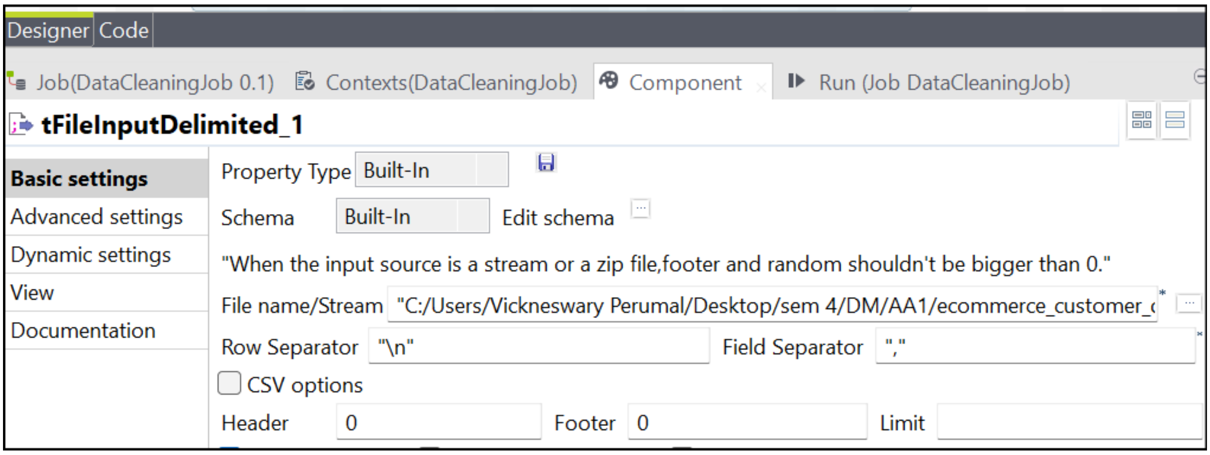
- Deleting Rows with Invalid Cell on Total Purchase Amount:

The final step removes rows that contain invalid or unusable data in the "Total Purchase Amount" column.

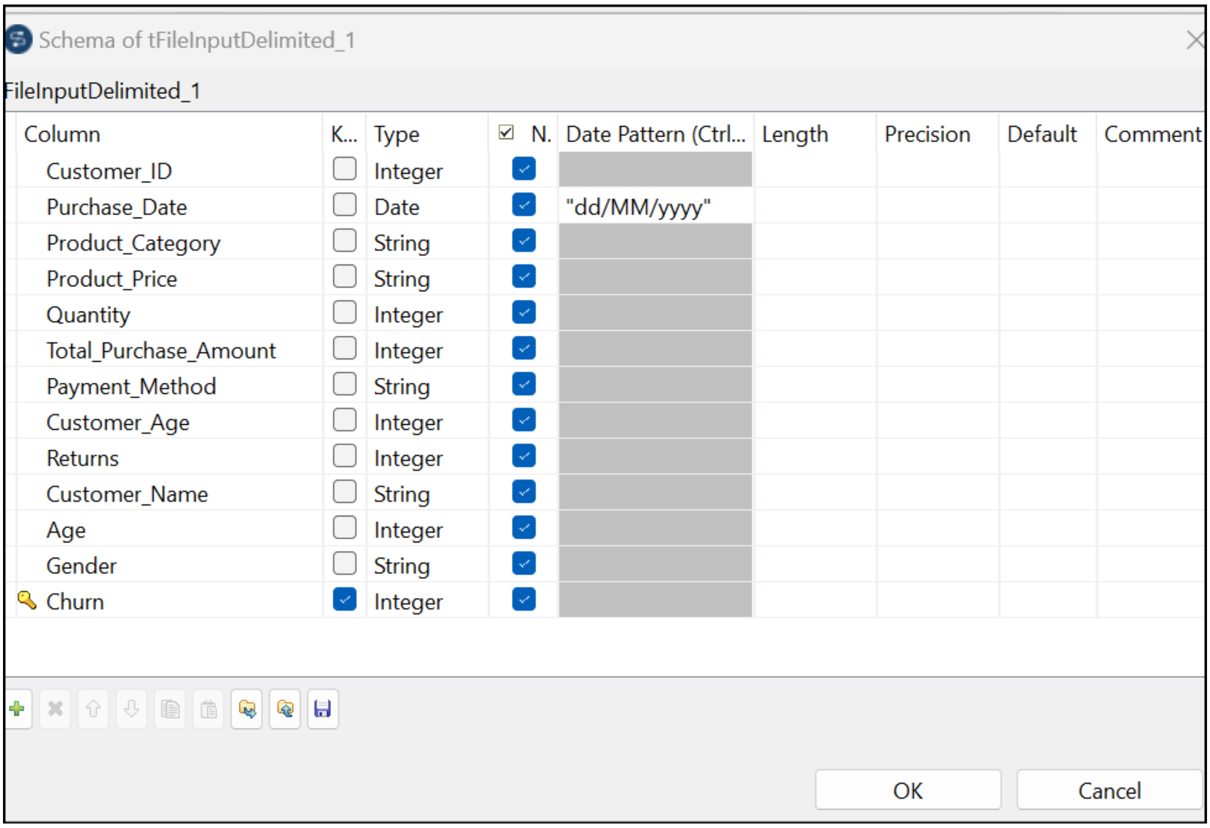


Talend Open Studio

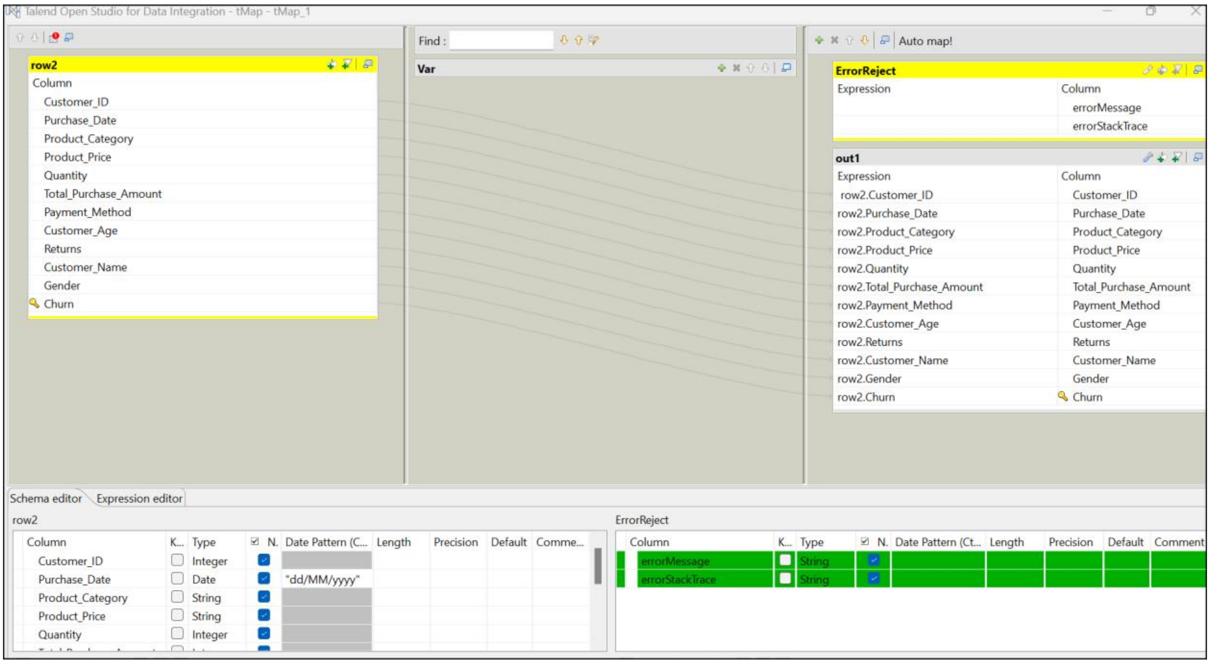
The cleaned CSV file import into Talend Open Studio.



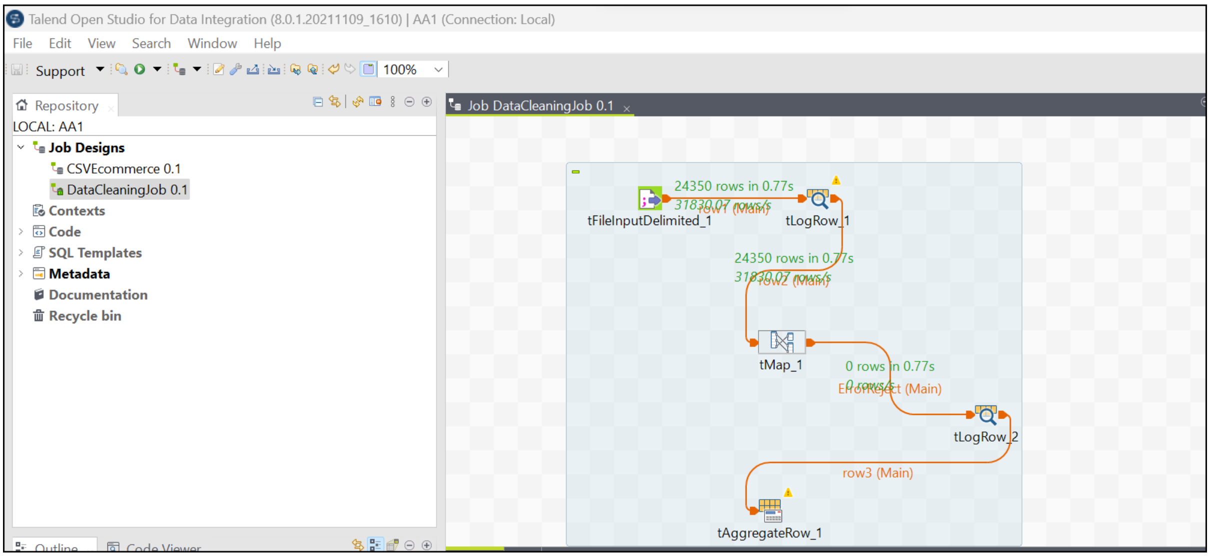
Ensure data types are corresponding to data



tMap is used to remove identical column Customer_Age and Age.



Final Output of data cleaning process in Talend Open Studio



SAS Enterprise Miner

Import Data:

Import your dataset into SAS Enterprise Miner.

Create a New Project:

Start a new project or use an existing one.

Create a Decision Tree Node:

Drag and drop a "Decision Tree" node from the "Common" tab onto the process flow canvas.

Connect Data Source:

Connect the Decision Tree node to the dataset by dragging the arrow from the dataset node to the Decision Tree node.

Specify Training and Validation Percentages:

Data Partition: Determine the percentage of data want to allocate for training and validation. Enter these percentages in the corresponding fields (e.g., 70% for training, 30% for validation).

Configure Decision Tree Node:

Right-click on the Decision Tree node and select "Properties."

Set the target variable to "Churn."

Run the Decision Tree Node:

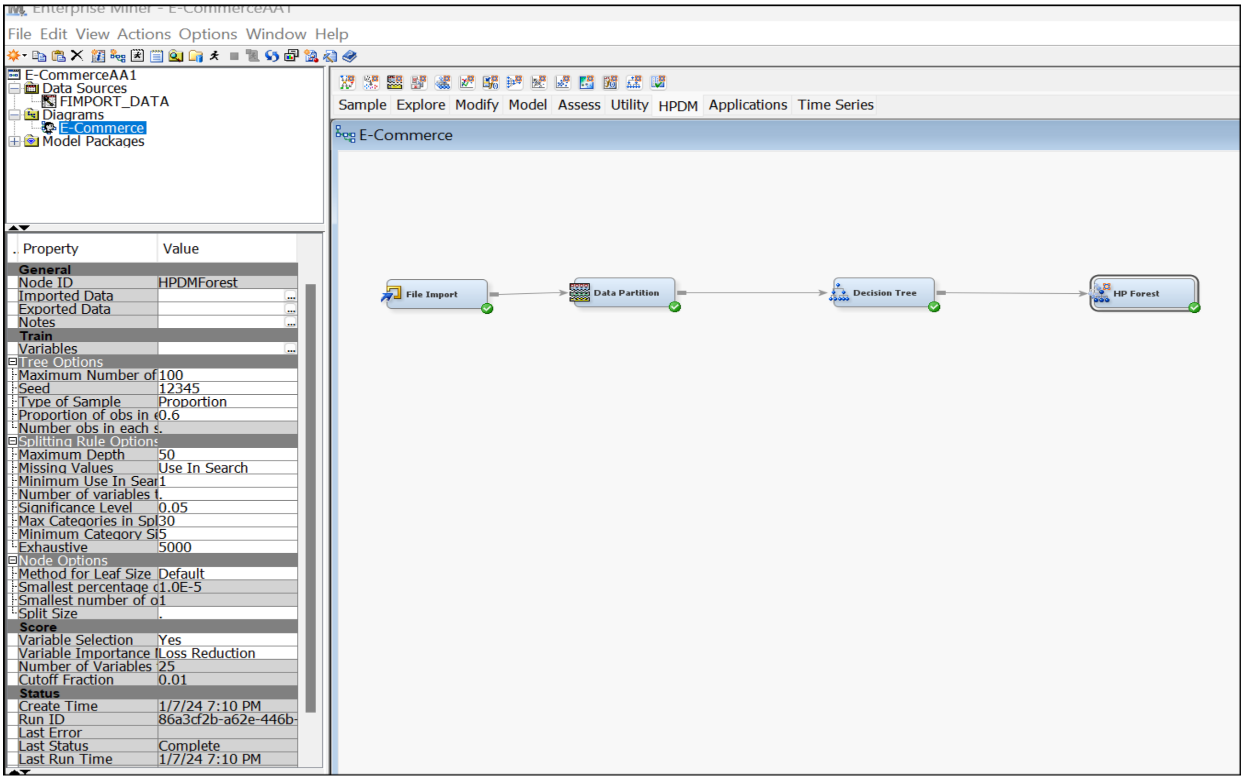
Run the Decision Tree node to build the model. Right-click on the Decision Tree node and select "Run."

Explore the Results:

After the model is built, can explore the results by right-clicking on the Decision Tree node and selecting "Results." This will provide information on the model's performance, summary statistics, and the decision tree diagram.

Decision Tree diagram

This process involves importing the dataset, partitioning the data into training and validation sets, building a Decision Tree model, and analyzing the results. Adjustments to the specific settings and configurations based on the characteristics on dataset and the objectives of analysis.



Results - Node: Decision Tree Diagram: E-Commerce

ile Edit View Window

Output

Gender	1	0.9429	0.5968	0.6329

Tree Leaf Report

Node Id	Depth	Training Observations	Training Average	Validation Observations	Validation Average	Training Root ASE	Validation Root ASE
2	1	8636	0.22	3675	0.21	0.41276	0.41081
4	2	4185	0.19	1845	0.20	0.39508	0.40254
11	4	2949	0.18	1240	0.18	0.38081	0.38203
6	3	1122	0.12	477	0.16	0.32741	0.36571
10	4	147	0.31	67	0.33	0.46088	0.47014

Fit Statistics

Target=Churn Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	17039.00	7304.00
MAX	Maximum Absolute Error	0.88	0.88
SSE	Sum of Squared Errors	2703.68	1178.74
ASE	Average Squared Error	0.16	0.16
RASE	Root Average Squared Error	0.40	0.40
DIV	Divisor for ASE	17039.00	7304.00
DFT	Total Degrees of Freedom	17039.00	.

Assessment Score Rankings

Data Role=TRAIN Target Variable=Churn Target Label=' '

Depth	Number of Observations	Mean Target	Mean Predicted
5	8783	0.21929	0.21929
55	4185	0.19355	0.19355
80	2949	0.17599	0.17599
95	1122	0.12210	0.12210

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Depth	Number of Observations	Mean Target	Mean Predicted
5	3742	0.21700	0.21939
55	1845	0.20325	0.19355
80	1240	0.17742	0.17599

The Decision Tree analysis for the provided dataset emphasizes key factors influencing customer churn prediction. According to the variable importance metrics, "Customer_Age" emerges as the most critical determinant, carrying a ratio of 1.0000, indicating its strong impact on predicting churn. Additionally, "Gender" is considered important, albeit to a slightly lesser extent. The tree leaf report showcases nodes with lower root average squared error, such as Node 6 (Depth 3), indicating higher accuracy in predicting churn for specific customer segments. The overall fit statistics, including the Root Average Squared Error, suggest a reasonably accurate fit of the model to the data. Assessment score rankings and distributions offer insights into how well the model differentiates churn likelihood across various depths and predicted value ranges. The model demonstrates proficiency in predicting churn, with the next steps involving a deeper exploration of the decision tree structure and criteria. Visualizing the decision tree can aid in understanding the specific conditions influencing churn predictions, enabling businesses to refine strategies for customer retention.