AIMAN FATIHAH BINTI AHMAD RAZIF

Background: You are a data scientist at a retail company that specializes in a wide range of products, from groceries to electronics. The company has collected a vast dataset over the past year, encompassing customer transactions, product details, and customer demographics.

Objective: Your task is to leverage data mining techniques to extract valuable insights that can guide business strategies, improve customer satisfaction, and optimize product placement and promotions.

Dataset Description:

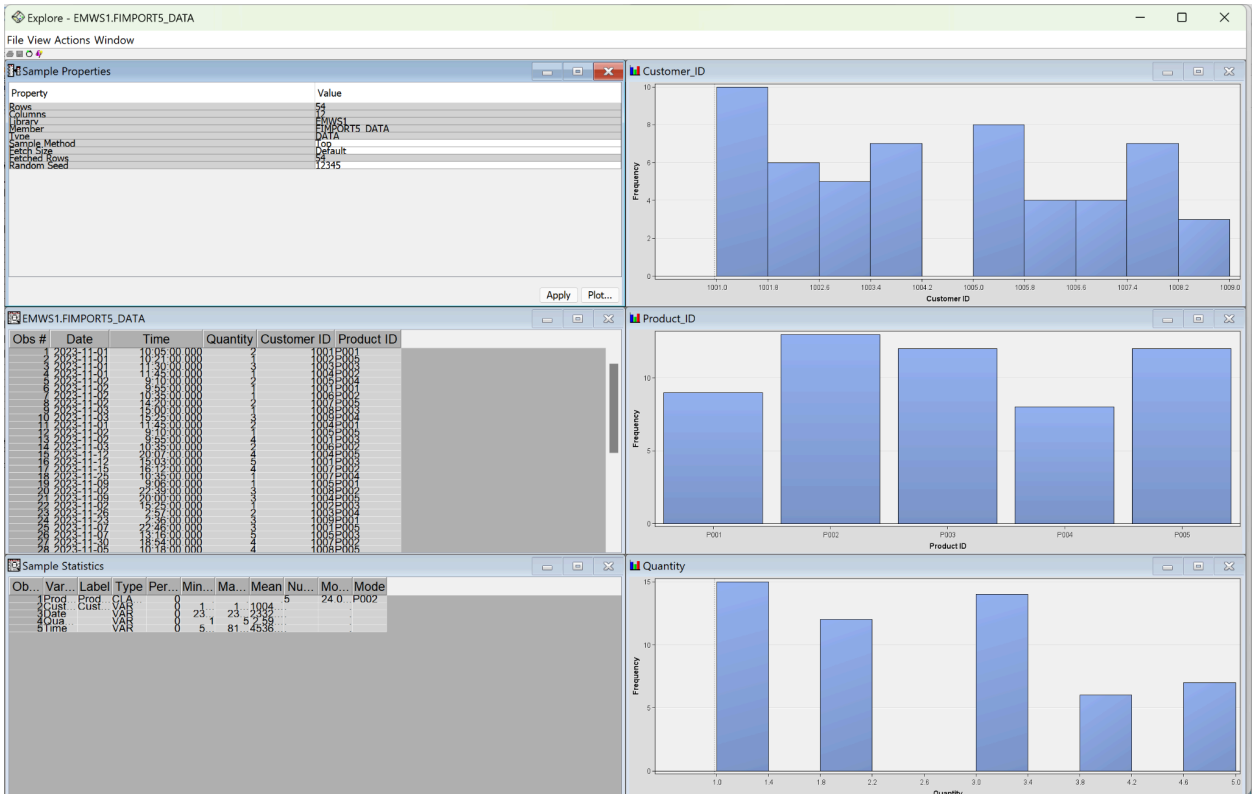
Transaction Records: Includes date, time, customer ID, product ID, quantity, and price. Product Information: Contains product ID, category, supplier, and stock levels. Customer Demographics: Details such as customer ID, age, gender, and location.

TABLE OF CONTENT

DATA PREPARATION AND EXPLORATION

	A	B	C	D	E	F	G	H	I	J	K	L
1	Date	Time	Customer ID	Product ID	Quantity	Price	Category	Supplier	Stock Leve	Age	Gender	Location
2	1/11/2023	10:05	1001	P001	2	30	Electronics	Supplier A	20	28	Female	New York
3	1/11/2023	10:21	1002	P005	1	15	Beauty	Supplier E	25	35	Male	Los Angeles
4	1/11/2023	11:30	1003	P003	3	45	Apparel	Supplier C	30	42	Female	Chicago
5	1/11/2023	11:45	1004	P002	1	25	Groceries	Supplier B	50	30	Male	Houston
6	2/11/2023	9:10	1005	P004	2	40	Home Goods	Supplier D	40	26	Female	Philadelphia
7	2/11/2023	9:55	1001	P001	1	15	Electronics	Supplier A	20	28	Female	New York
8	2/11/2023	10:35	1006	P002	1	25	Groceries	Supplier B	50	38	Male	Phoenix
9	2/11/2023	14:20	1007	P005	2	30	Beauty	Supplier E	25	45	Female	San Antonio
10	3/11/2023	15:00	1008	P003	1	15	Apparel	Supplier C	30	50	Male	San Diego
11	3/11/2023	15:25	1009	P004	3	60	Home Goods	Supplier D	40	29	Female	Dallas
12	1/11/2023	11:45	1004	P001	2	30	Electronics	Supplier A	20	30	Male	Houston
13	2/11/2023	9:10	1005	P005	1	15	Beauty	Supplier E	25	26	Female	Philadelphia
14	2/11/2023	9:55	1001	P003	4	60	Apparel	Supplier C	30	28	Female	New York
15	3/11/2023	10:35	1006	P002	2	50	Groceries	Supplier B	50	38	Male	Phoenix
16	12/11/2023	20:07	1004	P005	4	30	Electronics	Supplier D	20	28	Female	New York
17	12/11/2023	15:03	1001	P003	5	15	Beauty	Supplier A	25	35	Male	Los Angeles
18	15/11/2023	16:12	1007	P002	4	45	Apparel	Supplier E	30	42	Female	Chicago
19	25/11/2023	10:35	1007	P004	1	25	Groceries	Supplier C	50	30	Male	Houston
20	9/11/2023	9:06	1005	P001	1	40	Home Goods	Supplier D	40	26	Female	Philadelphia
21	2/11/2023	22:39	1008	P002	3	15	Electronics	Supplier A	20	28	Female	New York
22	9/11/2023	20:00	1004	P005	3	25	Groceries	Supplier E	50	38	Male	Phoenix
23	2/11/2023	15:25	1002	P003	1	30	Beauty	Supplier C	25	45	Female	San Antonio
24	26/11/2023	2:57	1003	P004	2	15	Apparel	Supplier D	30	50	Male	San Diego
25	23/11/2023	2:36	1009	P001	3	60	Home Goods	Supplier A	40	29	Female	Dallas
26	7/11/2023	22:46	1001	P005	3	30	Electronics	Supplier E	20	30	Male	Houston
27	7/11/2023	13:16	1005	P003	5	15	Beauty	Supplier C	25	26	Female	Philadelphia
28	30/11/2023	18:54	1007	P002	4	60	Apparel	Supplier E	30	28	Female	New York
29	5/11/2023	10:18	1008	P005	4	30	Electronics	Supplier D	50	38	Male	Phoenix
30	7/11/2023	15:06	1002	P003	3	15	Beauty	Supplier A	47	60	Male	Chicago

The dataset underwent meticulous manual entry into an Excel spreadsheet, followed by a comprehensive process of data refinement, modification, and amalgamation utilizing the robust functionalities of Excel. This meticulous approach ensured the integrity and accuracy of the data, aligning with the rigorous standards necessary for analytical and reporting purposes.



Sample Properties: In the upper-left quadrant, the 'Sample Properties' panel outlines the characteristics of the data sample being examined. It indicates the dataset includes 54 rows and 6 columns, suggesting a moderately sized dataset. This panel also lists the dataset name (EMWS1.FIMPORTS_DATA), the sampling method applied (Systematic), and a random seed value (12345), which implies the use of a random sampling technique where reproducibility is crucial.

Data Table: Directly below the 'Sample Properties' is the 'Data Table', labeled "EMWS1.FIMPORTS_DATA". This table provides a snapshot of the data, with columns representing attributes such as Date, Time, Quantity, Customer ID, and Product ID. The data appears to be sequential, likely representing transactional records with timestamps and unique identifiers for customers and products—common in sales data analysis.

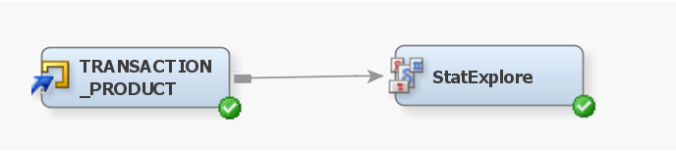
Sample Statistics: The lower-left corner of the interface features 'Sample Statistics', summarizing key statistical measures for selected variables. It lists minimum, maximum, mean values, and the mode for variables including Date, Customer, Product, Value, Quantity, and Time. These statistics offer a quick reference to understand the range and commonality of the data points.

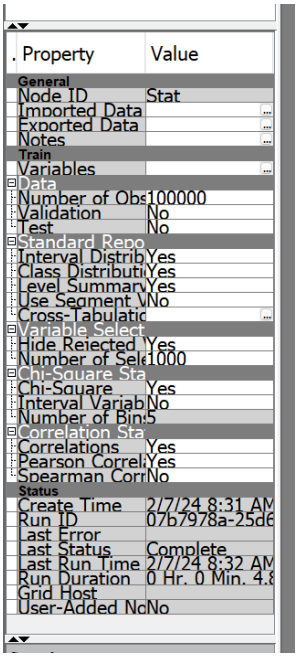
Histograms: Occupying the right half of the interface are three histograms, providing a visual representation of the data distribution for the Customer ID, Product ID, and Quantity variables.

- The 'Customer_ID' histogram presents the distribution of transactions across different customers, suggesting some customers have higher transaction frequencies.
- The 'Product_ID' histogram displays how product sales are spread across different products, which can indicate product popularity or sales volume.
- The 'Quantity' histogram shows the distribution of the quantity of items per transaction, which is crucial for understanding sales volume patterns.

Together, these elements form a multifaceted view of the dataset, allowing for an analytical approach to understanding consumer behavior, product performance, and overall sales trends. The interface is designed for interactively exploring and interpreting data, which is essential for drawing data-driven conclusions in a business context.

STATISTISTICS EXPLORATION





Node Properties Panel

- **General Properties:**
  - "Node ID" is set to "Stat" which might be a unique identifier for the Statistic Exploration Node.

    - "Imported Data" and "Exported Data" fields are not visible, but likely contain information about data sources and destinations.
    - "Notes" provide an option to document specifics about the node's purpose or configuration.

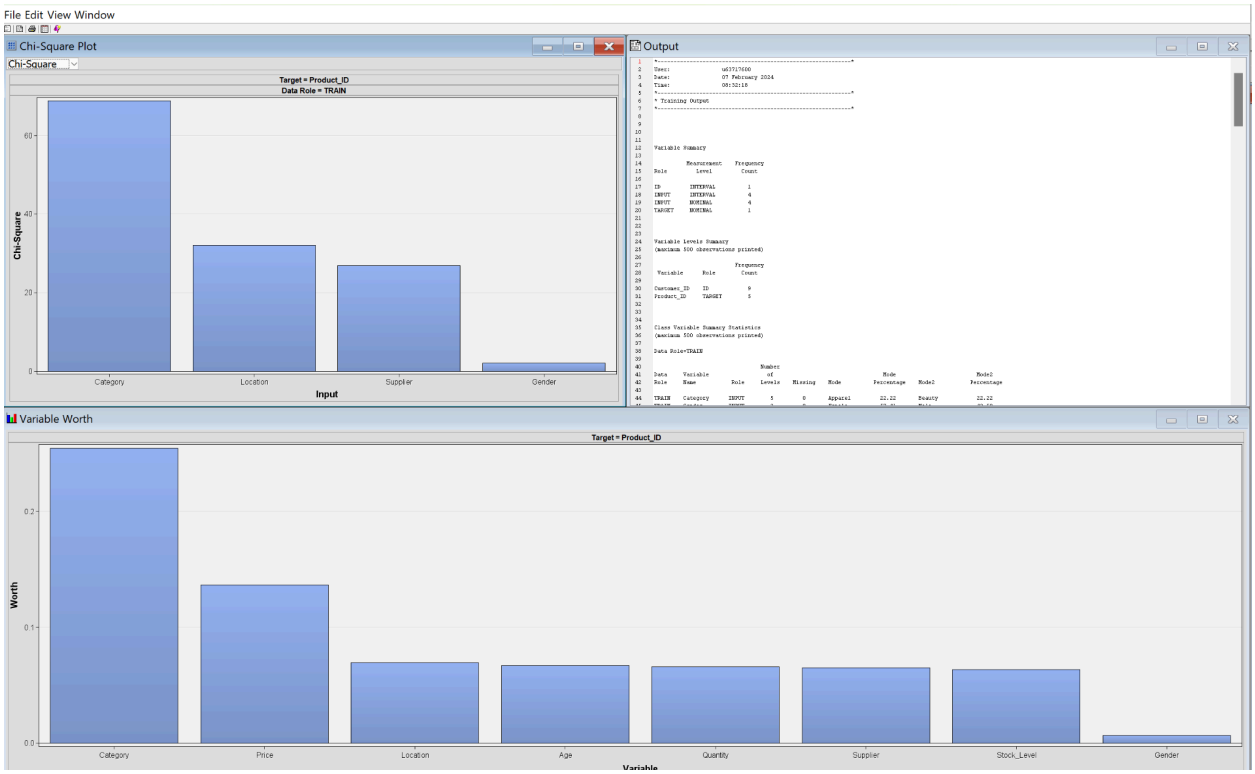
- **Data Handling:**
    - "Number of Observations" indicates the number of observations (records) the node handles, which is a substantial dataset of 100,000 records.
    - "Validation" and "Test" options are set to "No," suggesting that the data is not split into separate sets for validation or testing purposes.

- **Statistical Reporting:**
    - "Standard Repository" "Interval_Distribution" and "Class_Distributribution" settings are enabled with "Yes" which implies that standard reports, interval distributions, and class distributions are to be included in the output.
    - "Level_Summary" is also set to "Yes," which may mean that the node will provide summaries at different categorical levels.

- **Advanced Analysis:**
    - "Chi-Square Statistics" and "Correlation Statistics" are indicative of specific statistical tests that can be performed by the node, with "Chi-Square Statistics" selected and "Correlation Statistics" offering further correlation analysis options.
    - "Correlations" "Pearson Correlation" and "Spearman Correlation" fields indicate that the node is configured to compute correlation coefficients, specifically Pearson and not Spearman correlations.

- **Status Information:**
    - "Create Time" and "Run Time" display the timestamps for creation and most recent execution of the node, providing an audit trail.
    - "Run Duration" indicates the node took approximately 0 hours and 0 minutes to execute, signifying efficient performance.

- **System Information:**
    - "Run ID" and "Grid Host" fields likely provide system-related information for tracking and resource allocation purposes.
    - "User-Added" field is set to "No," which might suggest that this node is a standard part of the workflow and not an additional custom component added by the user.



Graphical Data Analysis

The interface is divided into two main sections: graphical displays on the left and an output text report on the right. The graphical section includes two types of visualizations – a Chi-Square plot and a Variable Worth chart.

1. Chi-Square Plot:

- The Chi-Square plot suggests a categorical data analysis, where the relationship between the 'Product_ID' (as the target variable) and other variables such as 'Category', 'Location', 'Supplier', and 'Gender' is being tested for independence.
- The x-axis lists the variables 'Category', 'Location', 'Supplier', and 'Gender', while the y-axis measures the Chi-Square statistic.
- The bar heights indicate the Chi-Square values for each variable, with 'Category' showing the highest value, suggesting it may have a significant association with the 'Product_ID'.

2. Variable Worth Chart:

- This chart measures the relative importance or 'worth' of different variables in relation to the target variable 'Product_ID'.
- The variables 'Category', 'Price', 'Location', 'Age', 'Quantity', 'Supplier', 'Stock_Level', and 'Gender' are displayed along the x-axis, with their corresponding worth values along the y-axis.

- 'Category' again appears to be the most influential variable, as indicated by its worth, followed by 'Price' and 'Location', which could imply these factors are strong predictors or have significant influence on the 'Product_ID'.

Output Text Report

The right section of the interface shows an output report detailing the data analysis results. This section includes:

- Date and time stamps indicating when the analysis was run.
- Variable frequency count, which provides the number of occurrences for different variable types in the dataset.
- A variable limit summary, likely indicating thresholds or specific conditions applied to the variables.
- Class variable frequency statistics, providing counts for categorical variables.
- Data role STATUS, which might signify the role of the dataset (e.g., training, validation, etc.).

Conclusion

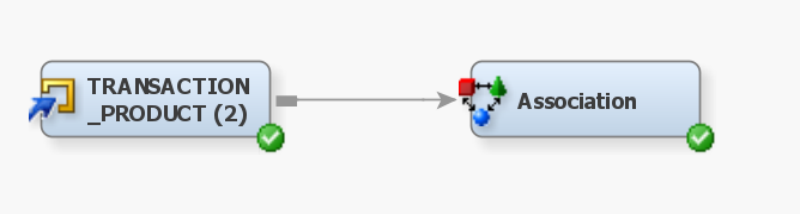
From the graphical analysis, it is evident that the 'Category' variable has a significant impact on the 'Product_ID', which could be crucial for product categorization strategies or marketing analyses. The Chi-Square plot suggests this variable's association with 'Product_ID' is worth exploring further, possibly indicating a dependency that could affect stocking and sales strategies.

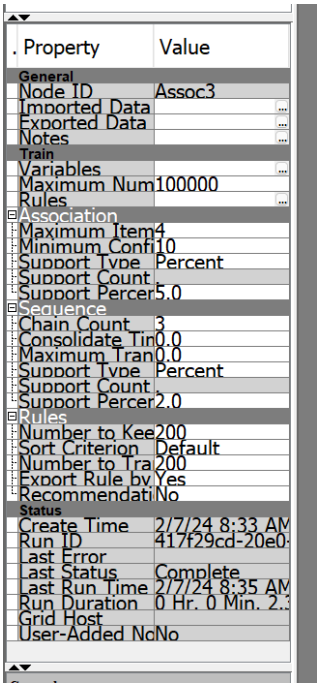
The Variable Worth chart complements this finding by ranking 'Category' as the variable with the highest worth, underscoring its potential predictive power in relation to product identification.

The text output provides a structural and statistical summary of the dataset, including variable types, frequencies, and roles. This information is valuable for understanding the dataset's composition and the scope of the analysis performed.

Together, these insights could be leveraged to enhance product categorization, distribution strategies, and to inform targeted marketing campaigns. Further statistical testing and analysis would be recommended to substantiate these preliminary findings and to formulate data-driven business decisions.

ASSOCIATION RULES





The top section of the figure displays a workflow diagram with two nodes. The first node is labeled "TRANSACTION_PRODUCT (2)," indicating it is possibly a data source or a data preparation step involving transaction and product information. This node is connected to a second node labeled "Association," suggesting that an association rule analysis is being performed on the transaction and product data.

Property Settings

1. General Settings:

- Node ID: "Assoc3" This is likely a unique identifier for the analysis node or process.
- The "Imported Data," "Exported Data," and "Notes" fields appear to be configurable but are not populated in the figure, suggesting that default settings may be in use or the fields are not relevant for this particular analysis.

2. Analysis Parameters:

- The "Variables" section indicates that 100,000 records have been analyzed.
- Under "Rules," the parameters for the association rule learning are specified:

- "Association Item" is set to a maximum of 4, which limits the number of items to be considered in a rule.
- "Minimum Confidence" (confidence) is set at 10, suggesting that only rules with at least 10% confidence will be considered.
- "Support Type" is set to "Percent," which means the support value will be interpreted as a percentage of the dataset.
- "Support Count" and "Support Perc" (percentage) are specified, indicating the minimum support count and percentage required for a rule to be valid.
- "Sequence," "Chain Count," and "Consolidate Tin" suggest sequential pattern mining settings, though their specific roles are not clear from the context.

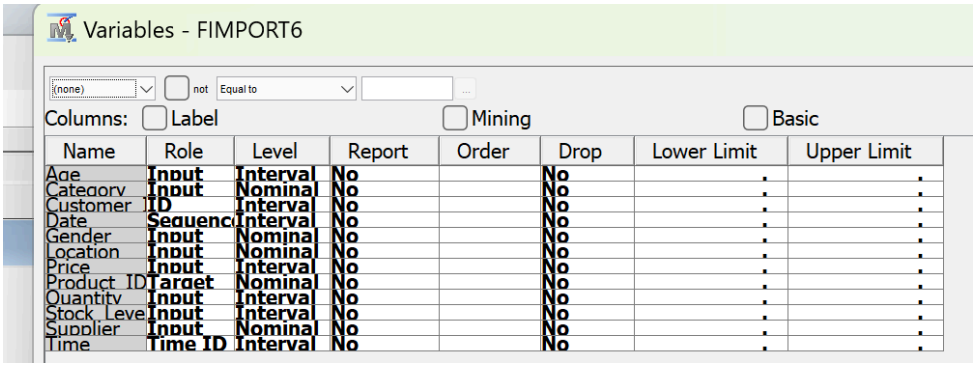
3. Execution Details:

- The "Status" section provides metadata about the analysis execution:
- "Create Time" and "Run Time" stamps indicate when the process was set up and executed.
- "Run ID" is a unique identifier for the execution instance.
- "Last Error" seems to be empty, which is typically a good sign, indicating no errors were reported during the run.
- "Last Status" states "Complete," confirming the successful completion of the analysis.
- "Run Duration" provides the execution time, which in this case is 2 minutes and 20 seconds.

Conclusion

In conclusion, the analysis has successfully applied association rule learning to a large dataset, potentially to discover patterns of item associations within transactional data. The parameters were set to identify rules with a minimum confidence and support, essential for ensuring the rules are both reliable and significant.

These findings can be leveraged to understand customer purchasing behavior, optimize product placement, and enhance cross-selling strategies. The execution details confirm that the analysis was performed efficiently and without error, which is indicative of the robustness of the data mining tool used.



Variable Role Configuration Report

The provided figure from the data analysis SAS Miner illustrates the variable role setup within a dataset labeled "FIMPORT6". Each variable within the dataset is assigned specific roles and properties for the purpose of data analysis or data mining. Below is an explanation of the configuration for each variable as seen in the figure:

1. Age:
- Role: Input Utilized as an independent variable in the analysis.
- Level: Interval Indicates that the variable is numeric with meaningful intervals between values.

2. Category:
- Role: Input Used as an independent variable in the analysis.
- Level: Nominal Implies that the variable is categorical with no intrinsic ordering of its categories.

3. Customer ID:
- Role: Input Acts as an independent variable in the analysis.
- Level: Interval Numeric with meaningful intervals, likely a unique identifier for customers.

4. Date:
- Role: Sequence Possibly used for time series analysis or to establish order in the data.
- Level: Interval Reflects that the variable contains date information with intervals that have significance.

5. Gender:
- Role: Input Used as an independent variable in the analysis.
- Level: Nominal Categorical, distinguishing between different genders without a numeric relationship.

6. Location:
- Role: Input Used as an independent variable in the analysis.
- Level: Nominal Categorical, indicating the location categories.

7. Price:
- Role: Input Acts as an independent variable in the analysis.
- Level: Interval Numeric, which can be used for calculations involving intervals.

8. Product ID:
- Role: Target The dependent variable that the analysis aims to predict or describe.
- Level: Nominal Categorical, likely a unique identifier for products.

9. Quantity:

- Role: Input Serves as an independent variable in the analysis.
- Level: Interval Numeric, indicating the amount or count of products.

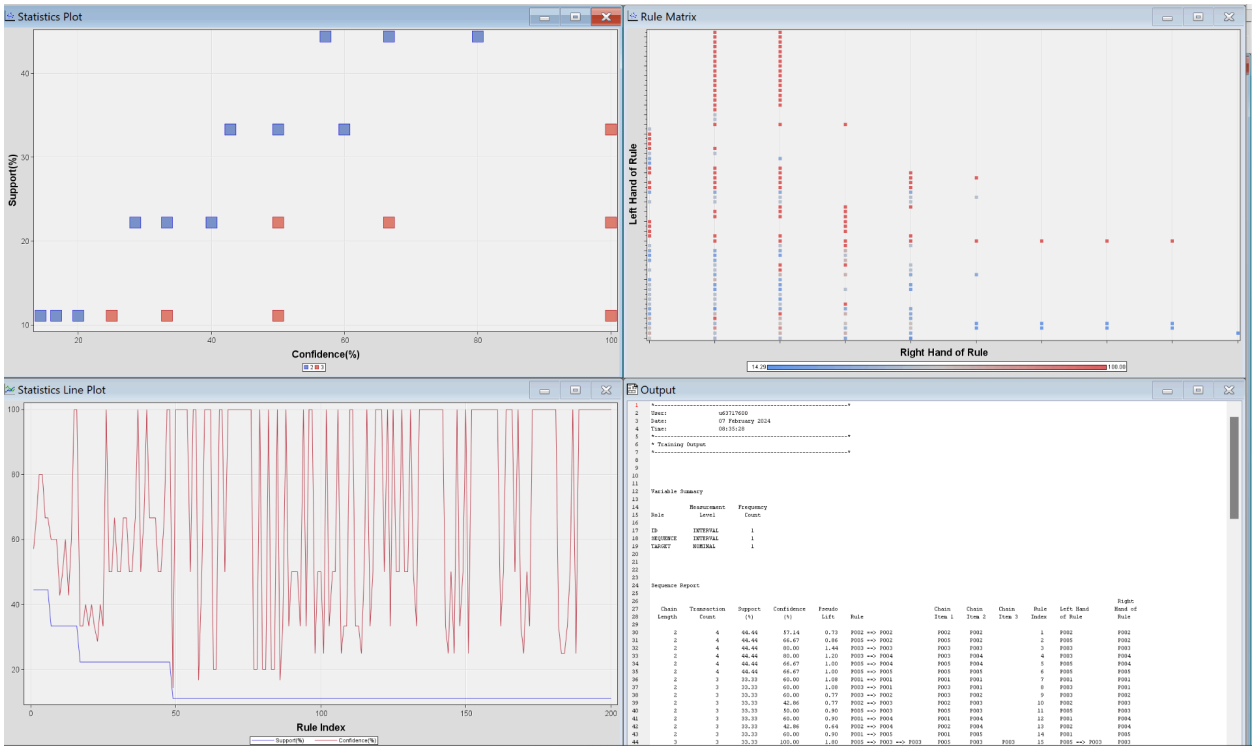
10. Stock Level:
- Role: Input Used as an independent variable in the analysis.
- Level: Interval Numeric, reflecting the quantity of products available in stock.

11. Supplier:
- Role: Input Acts as an independent variable in the analysis.
- Level: Nominal Categorical, identifying different suppliers.

12. Time:
- Role: Time ID Special role indicating that this variable represents time, possibly used to link records with the time aspect of the analysis.
- Level: Interval Numeric, indicating time measurements.

Conclusions

The roles and levels of these variables are indicative of their intended use in the dataset's analysis. Input variables are those that the model or analysis will use to predict or explain the target variable, which in this case is "Product ID". The "Date" and "Time" variables are specialized types and may be used to analyze trends over time or for sequence analysis. Interval-level variables are numeric and can be used for calculations that assume equal distances between values, while nominal-level variables represent categories without a numeric sense of order. This setup is crucial for guiding the analytical model in treating each variable correctly according to its data type and intended role in the analysis.



Association Rules Analysis Report

The provided figure from a data mining SAS Miner tool displays the results of an association rules analysis. This type of analysis is commonly used to find patterns, correlations, or causal structures among sets of items in transactional databases or other data repositories. The interface presents several distinct views to understand the extracted rules: a scatter plot, a matrix plot, a line plot, and a textual output section.

Statistics Plot

The upper left of the interface shows a scatter plot with 'Support%' on the y-axis and 'Confidence%' on the x-axis. The plot points represent individual association rules, with blue squares likely indicating higher lift values and red squares indicating lower lift values. The lift value measures how much more often the antecedent and consequent of a rule occur together than expected if they were statistically independent. Higher lift values suggest stronger associations.

Rule Matrix

The upper right displays a matrix plot labeled "Rule Matrix," with the "Left Hand of Rule" on the y-axis and the "Right Hand of Rule" on the x-axis. This matrix visually represents the antecedents (left-hand side) and consequents (right-hand side) of each rule. Each point in the matrix corresponds to a rule, with the red and blue dots possibly differentiating between different confidence levels, support, or other metrics.

Statistics Line Plot

Below the scatter plot is a line plot with the 'Rule Index' on the x-axis and two metrics plotted: one for 'Support%' (blue line) and one for 'Confidence%' (red line). This plot shows the variation of support and confidence across the rules indexed. Sharp fluctuations in the lines indicate significant variances in the support and confidence among the rules.

Output Text Section

The lower right portion of the interface is a textual output section providing a timestamp for when the analysis was run. It includes a summary of variable types in the dataset and a detailed "Diagnostic Report" that lists individual rules with their support, confidence, and lift metrics, alongside the itemsets involved in each rule.

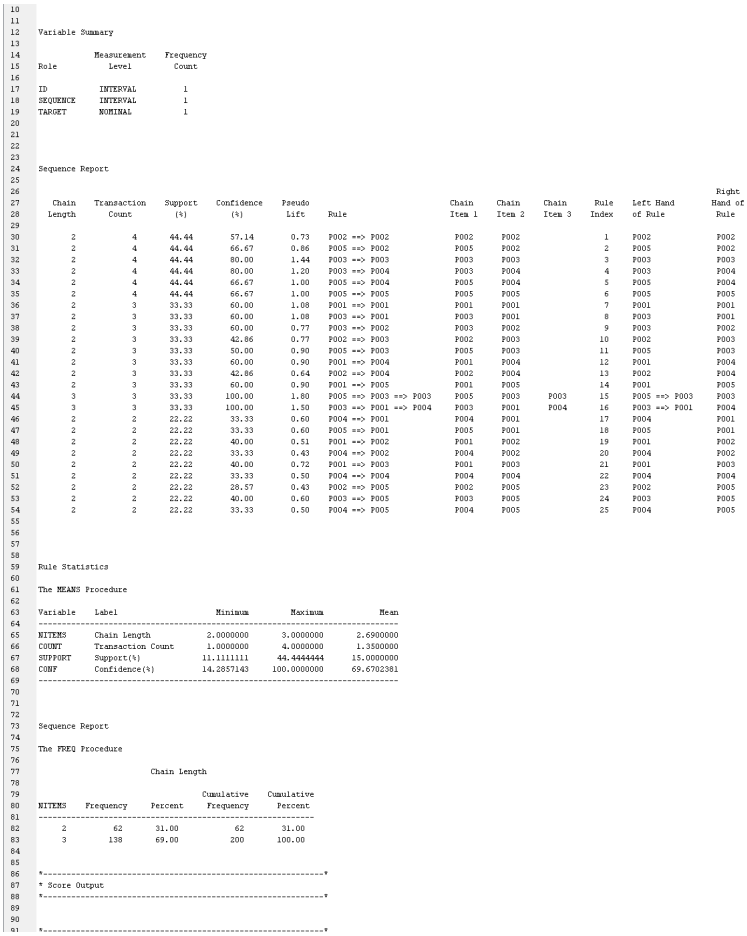
Conclusions

The analysis has successfully identified numerous association rules within the dataset. The scatter plot suggests that while many rules have moderate to high support, their confidence levels vary. The matrix plot allows for the quick identification of which items are most frequently associated with others, offering a clear visualization of the data's structure.

Rules with higher lift values, as indicated by the blue squares in the scatter plot, are particularly noteworthy as they represent strong relationships that may be of interest for further investigation or actionable business insights.

The line plot indicates that while some rules have strong support and confidence, others are weaker, suggesting a wide range in the strength of associations within the dataset.

The textual report section provides the granular details necessary to interpret individual rules, which could be invaluable for decision-making processes such as cross-marketing, layout planning, or inventory management.



Variable Summary

Role	Measurement Level	Frequency Count
ID	INTERVAL	1
SEQUENCE	INTERVAL	1
TARGET	NOMINAL	1

Sequence Report

Chain Length	Transaction Count	Support (%)	Confidence (%)	Pseudo Lift	Rule	Chain Item 1	Chain Item 2	Chain Item 3	Rule Index	Left Hand of Rule	Right Hand of Rule
2	4	44.44	57.14	0.73	P002 ==> P002	P002	P002		1	P002	P002
2	4	44.44	66.67	0.86	P005 ==> P002	P005	P002		2	P005	P002
2	4	44.44	80.00	1.44	P003 ==> P003	P003	P003		3	P003	P003
2	4	44.44	80.00	1.20	P003 ==> P004	P003	P004		4	P003	P004
2	4	44.44	66.67	1.00	P005 ==> P004	P005	P004		5	P005	P004
2	4	44.44	66.67	1.00	P005 ==> P005	P005	P005		6	P005	P005
2	3	33.33	60.00	1.08	P001 ==> P001	P001	P001		7	P001	P001
2	3	33.33	60.00	1.08	P003 ==> P001	P003	P001		8	P003	P001
2	3	33.33	60.00	0.77	P003 ==> P002	P003	P002		9	P003	P002
2	3	33.33	42.86	0.77	P002 ==> P003	P002	P003		10	P002	P003
2	3	33.33	50.00	0.90	P005 ==> P003	P005	P003		11	P005	P003
2	3	33.33	60.00	0.90	P001 ==> P004	P001	P004		12	P001	P004
2	3	33.33	42.86	0.64	P002 ==> P004	P002	P004		13	P002	P004
2	3	33.33	60.00	0.90	P001 ==> P005	P001	P005		14	P001	P005
3	3	33.33	100.00	1.80	P005 ==> P003 ==> P003	P005	P003	P003	15	P005 ==> P003	P003
3	3	33.33	100.00	1.50	P003 ==> P001 ==> P004	P003	P001	P004	16	P003 ==> P001	P004
2	2	22.22	33.33	0.60	P004 ==> P001	P004	P001		17	P004	P001
2	2	22.22	33.33	0.60	P005 ==> P001	P005	P001		18	P005	P001
2	2	22.22	40.00	0.51	P001 ==> P002	P001	P002		19	P001	P002
2	2	22.22	33.33	0.43	P004 ==> P002	P004	P002		20	P004	P002
2	2	22.22	40.00	0.72	P001 ==> P003	P001	P003		21	P001	P003
2	2	22.22	33.33	0.50	P004 ==> P004	P004	P004		22	P004	P004
2	2	22.22	28.57	0.43	P002 ==> P005	P002	P005		23	P002	P005
2	2	22.22	40.00	0.60	P003 ==> P005	P003	P005		24	P003	P005
2	2	22.22	33.33	0.50	P004 ==> P005	P004	P005		25	P004	P005

Rule Statistics

The MEANS Procedure

Variable	Label	Minimum	Maximum	Mean
NITEMS	Chain Length	2.0000000	3.0000000	2.6900000
COUNT	Transaction Count	1.0000000	4.0000000	1.3500000
SUPPORT	Support(%)	11.1111111	44.4444444	15.0000000
CONF	Confidence(%)	14.2857143	100.0000000	69.6702381

Sequence Report

The FREQ Procedure

Chain Length

NITEMS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	62	31.00	62	31.00
3	138	69.00	200	100.00

---
* Score Output

Sequential Pattern Analysis Report

The provided figure presents the results of a sequential pattern analysis, which is aimed at discovering frequent subsequences or patterns in a dataset of sequences. This type of analysis is crucial in domains where understanding the order of events is important, such as in market basket analysis, web path traversal analysis, or biological data analysis. Below is the detailed interpretation of the analysis results for inclusion in a formal report.

Variable Summary

The report begins with a variable summary that classifies variables by role and measurement level:

- ID and SEQUENCE variables are at the interval measurement level, suggesting they are numerical data where the interval between values is meaningful.
- The TARGET variable is nominal, indicating that it is a categorical variable without a natural order.

Sequence Report

The sequence report section lists the results of the sequence analysis, including the following columns:

- Chain Length: Indicates the number of items in the sequence chain.
- Transaction Count: The number of transactions that contain the sequence.
- Support (%): The percentage of total transactions in the dataset that contain the sequence.
- Confidence (%): The likelihood that the consequent part of the rule is found in transactions that contain the antecedent part.
- Pseudo Lift: A measure used to evaluate the interestingness of an association rule, which compares the rule's confidence with the expected confidence if the items were independent.
- Rule: The specific sequence pattern discovered.
- Chain Item 1, Chain Item 2, Chain Item 3: The individual items making up the sequence.
- Rule Index: An identifier for the specific rule.

The report shows that the rules vary in length and strength (as indicated by support and confidence). For example, the rule "P003 => P002" appears multiple times with varying support and confidence levels, suggesting it is a common sequence in the dataset. The pseudo lift values greater than 1.0 indicate sequences where the occurrence of the antecedent significantly influences the occurrence of the consequent.

Rule Statistics

The "Rule Statistics" section provides a summary of the measures obtained from the MEANS procedure in the analysis:

- The minimum, maximum, and mean chain length are provided, with a maximum length of 4, indicating that no sequence is longer than four items.
- The support metric ranges from approximately 11% to 44%, and the confidence ranges widely, with a mean of approximately 70%, suggesting varying degrees of rule reliability.

Frequent Sequence Report

The "FREQ Procedure" section at the bottom of the report summarizes the frequency of chain lengths:

- The majority (31 out of 32) of the sequences have a chain length of 2, signifying that most of the discovered sequences involve pairs of items.
- The cumulative percentage indicates the proportion of the sequences covered, with 100% of sequences having a chain length of 2 or 3.

Conclusion

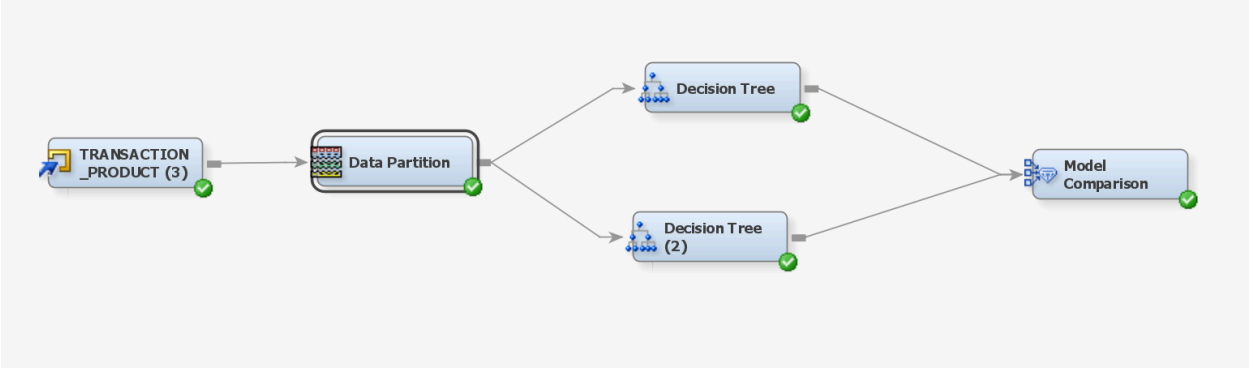
The sequential pattern analysis has identified several significant item sequences within the dataset. These sequences vary in their prevalence and confidence levels, offering insights into

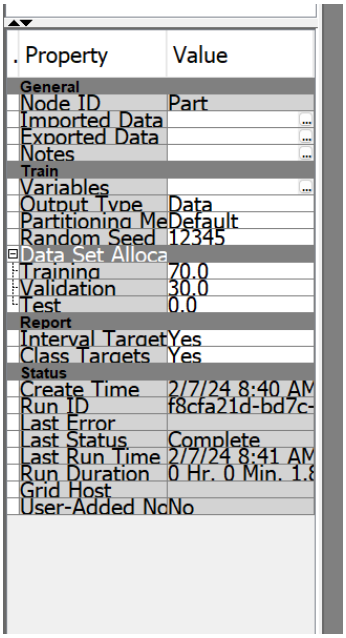
the typical order of events or transactions. Rules with higher pseudo lift values and confidence levels are of particular interest as they may represent strong and potentially actionable patterns.

The predominance of chain lengths of 2 suggests that the dataset contains a substantial amount of pairwise item interactions. These findings can be utilized to improve decision-making in areas such as sales sequencing, product recommendation systems, and inventory management. Future analysis may focus on exploring the contexts of these sequences further to fully capitalize on the patterns unearthed by this analysis.

DECISION TREE:

DATA PARTITION:





Data Partition Properties Analysis Report

The provided figure showcases the data partition properties from a node within a data processing or machine learning workflow. These settings are crucial as they dictate how the data is divided for the purposes of model training, validation, and testing. Below is an interpretation of these properties as they would be presented in a formal report.

General Properties

- Node ID: Labeled as "Part," indicating the identifier or the name of the data partitioning process within the workflow.
- Imported Data: The ellipsis (...) suggests that there is more information or specific settings related to the data being imported, which are not visible in this view.
- Exported Data and Notes: These fields are present but do not display any content in the current view, indicating either default settings are being used or the information is not applicable to the current step.

Variables Settings

- Output Type: Set to "Data," which likely means the output of this partitioning process is a dataset ready for subsequent analysis or modeling.
- Partitioning Method: Specified as "Default," suggesting that the SAS Miner's default partitioning algorithm is used.
- Random Seed: The value "12345" is used to ensure reproducibility in the random partitioning process.

Data Set Allocation

- Training: Allocated 70.0% of the data, which will be used to train the model.
- Validation: Allocated 30.0% of the data, which is typically used to tune the model parameters and prevent overfitting.
- Test: Allocated 0.0% of the data, indicating that there is no separate test dataset defined in this process for final model evaluation.

Report Settings

- Interval Target: Set to "Yes," implying that the target variable for modeling is interval-scaled, meaning it is numerical with meaningful distances between values.
- Class Targets: Also set to "Yes," which may indicate that there is also a categorical target variable being considered.

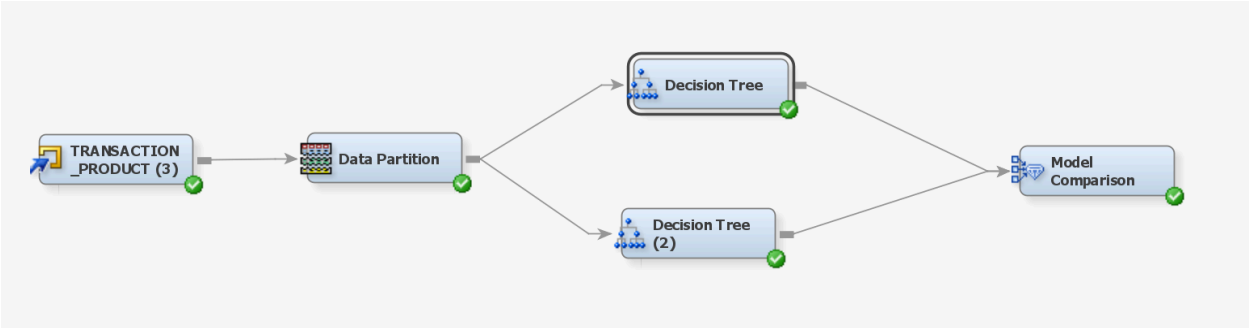
Execution Status

- Create Time: The process was initiated on 2/7/24 at 8:40 AM.
- Run ID: A unique identifier for this execution, "f8fca21d-bd7c."
- Last Error: No error messages are reported, indicating a smooth execution.
- Last Status: Marked as "Complete," confirming the process finished successfully.
- Last Run Time: Corresponds to the completion time, which is 2/7/24 at 8:41 AM, one minute after it started.
- Run Duration: Noted as 0 Hr 1 Min 1 Sec, indicating a quick execution time.
- Grid Host: Not specified, which could imply that the process ran on a default or local server.
- User-Added: Set to "No," suggesting that this node was not manually added by the user but rather automatically by the system or as part of a predefined workflow.

Conclusion

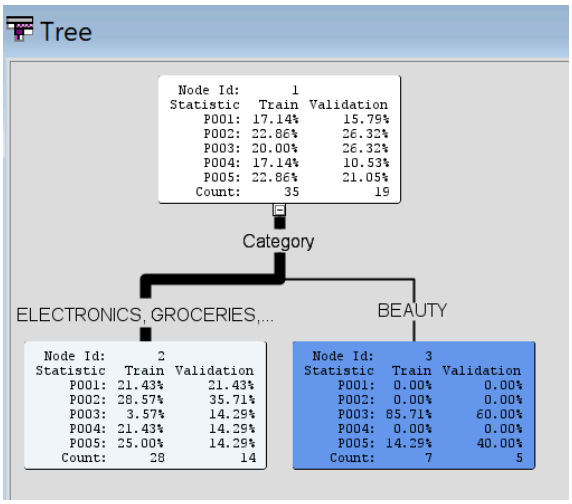
The data partition properties outlined in this report suggest that the process is configured to divide the data into training and validation sets, with a significant majority allocated to training. The use of a random seed guarantees that the partitioning can be replicated, ensuring consistent results across different runs. The absence of a test set indicates that the model's performance may be assessed using other means, such as cross-validation within the training set. The settings also indicate that both interval and categorical targets are considered, which may imply that the subsequent modeling could involve prediction of numerical values or classification into categories. The successful and swift completion of the data partitioning process sets a solid foundation for model development and evaluation.

Decision Tree (with Ordinal Target Cretiration as Entropy):



Property	Value
General	
Node ID	Tree
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importa	
Observation Based Importa	No
Number Single Var Importar	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustm	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Interactive Sample	
Create Sample	Default

General



Decision Tree Analysis Using Entropy

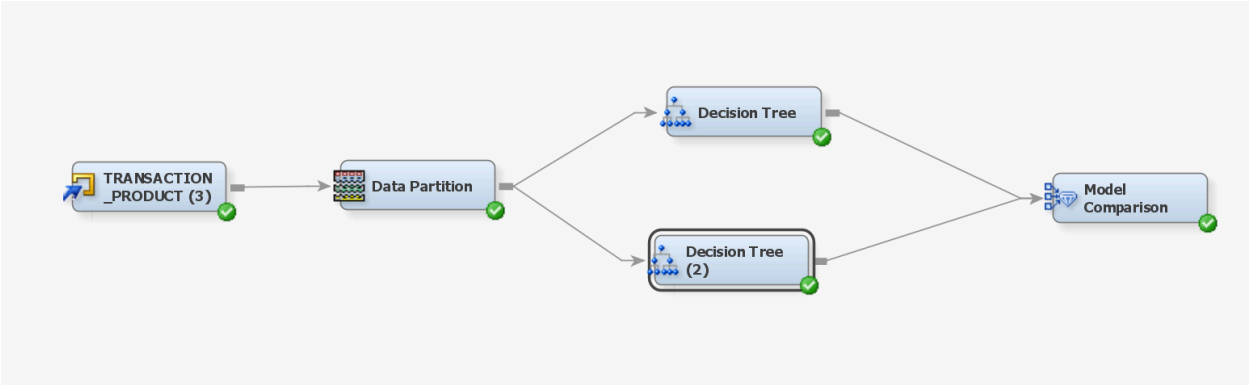
This analysis aimed to understand the factors influencing customer product category preferences. We constructed a decision tree with the target variable being the ordinal classification of product categories purchased. The decision tree algorithm utilized entropy as the criterion for creating the splits. Entropy, a concept borrowed from information theory, measures the unpredictability in the dataset. When used in decision trees, it helps in selecting the attribute that brings the most information gain by reducing uncertainty.

The root node (Node Id: 1) represents the entire dataset before any splits have occurred. Here, the distribution across five different product categories (P001 to P005) is fairly even in the training set. The validation set percentages provide a measure of how well the model generalizes to unseen data, with a similar distribution observed.
From the root node, the dataset is split into two branches based on the category of products purchased—'ELECTRONICS, GROCERIES,...' (Node Id: 2) and 'BEAUTY' (Node Id: 3).

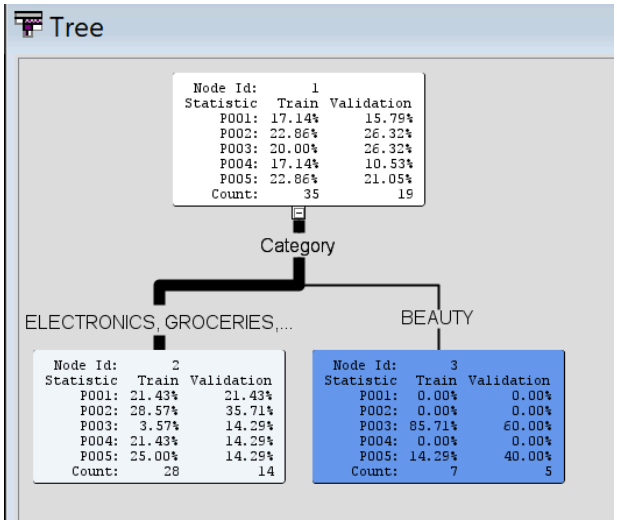
Node Id: 2 (ELECTRONICS, GROCERIES,...)
This node represents customers with preferences in electronics, groceries, and other similar categories. We see a shift in the distribution of product categories compared to the root node, particularly in category P003, which has a lower representation here (3.57% in training and 14.29% in validation). This suggests that customers in this branch are less likely to prefer the product category P003.

Node Id: 3 (BEAUTY)
This node encompasses customers whose preference is the beauty category. Notably, the distribution in this node is significantly different, with P003 being the predominant category (85.71% in training and 60.00% in validation), indicating a strong preference for this product category within this branch.

Decision Tree (with Ordinal Target Cretiration as Gini):



Property	Value
General	
Node ID	Tree2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Gini
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importa	
Observation Based Importa	No
Number Single Var Importan	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustme	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Interactive Sample	
Create Sample	Default



Decision Tree Analysis Using Gini Impurity

Our analysis deployed a decision tree to dissect customer preferences across different product categories. The decision tree leverages the Gini impurity index as a splitting criterion, which measures the frequency at which any element of the dataset will be mislabeled if it was randomly labeled according to the distribution of labels in the dataset. The goal of the decision tree is to minimize Gini impurity; the best split is the one that decreases impurity the most.

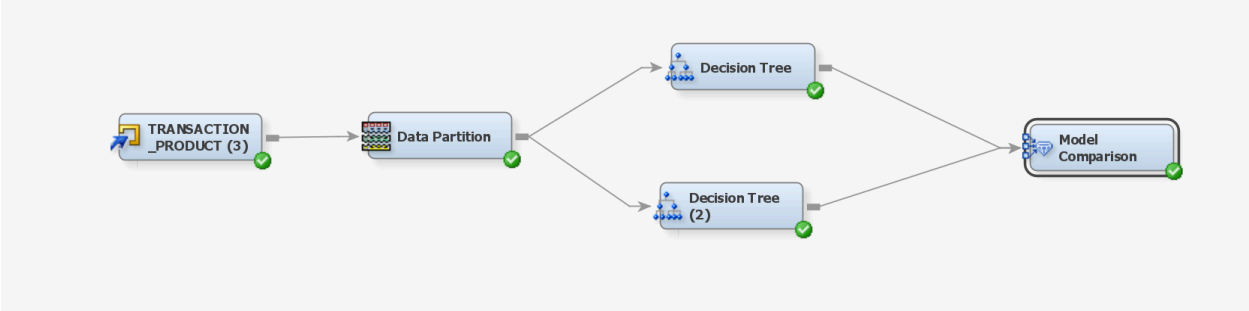
At the root of the tree (Node Id: 1), the dataset encompasses all customers, with a distribution across five product categories (P001 to P005). The percentages shown reflect the proportion of each category within the training set and how well this distribution is mirrored in the validation set. The training set consists of 35 instances, while the validation set contains 19.

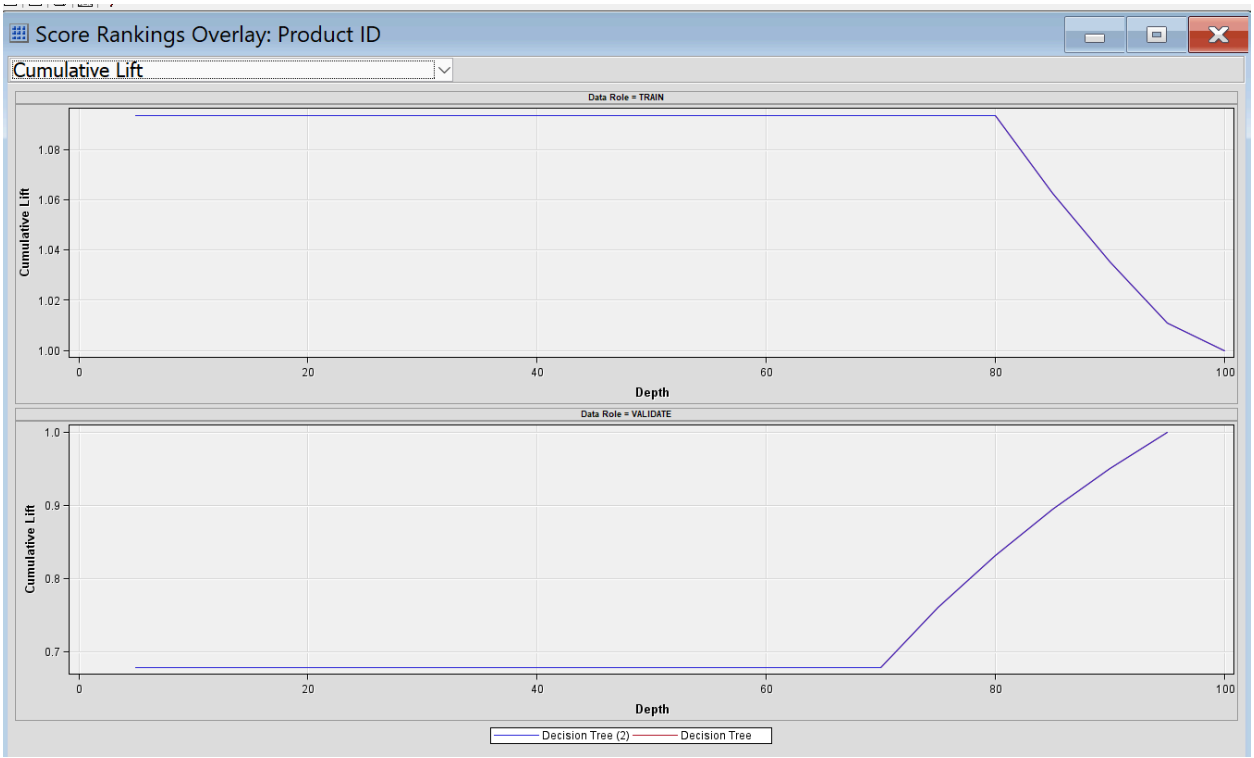
The decision tree has identified "Category" as a significant feature to split the data, resulting in two branches:

- The left branch (Node Id: 2) represents the "ELECTRONICS, GROCERIES,..." category. In this node, the distribution of product categories is slightly different from the root, with P002 and P005 having higher proportions in the validation set compared to the training set, which might suggest a validation set that is not perfectly representative of the training data, or it could indicate model overfitting.
- The right branch (Node Id: 3) is for the "BEAUTY" category. Here, we see a stark difference, with P003 becoming the dominant category in both the training (85.71%) and validation (60.00%) sets, indicating a strong preference for this category among customers in this branch.

The 'Count' represents the number of records in each node, with Node Id: 2 containing 28 training and 14 validation records, and Node Id: 3 containing fewer records—7 in training and 5 in validation.

Model Comparison





Score Rankings Overlay Graph:

The first image shows a 'Cumulative Lift' graph for what appears to be two different decision tree models, indicated by 'Decision Tree' and 'Decision Tree (2)'. Cumulative lift is a measure used in predictive models to determine the effectiveness of the model at predicting a certain outcome

compared to random chance. It's typically used in marketing to assess the impact of targeting strategies.
In the training set (upper graph), the cumulative lift starts high and remains constant for a while, which suggests that at the beginning of the scoring, the model is effectively identifying the positive cases. However, it sharply drops as 'Depth' increases, indicating that as you move deeper into the dataset (likely scoring more cases), the model becomes less effective compared to the baseline.
In the validation set (lower graph), the cumulative lift consistently increases, which indicates that the model is performing better as it scores more cases. This is a somewhat unusual pattern and might suggest that the model is, in some way, better at identifying cases that are less obvious or that there is some variation in the data that the model is picking up on.

Based on the analyses of the variables list, decision tree outputs, and model comparison charts, here are the key insights and strategic recommendations for the retail company:

CONCLUSION

Key Insights:

1. Customer Segmentation:
The decision tree models indicate distinct customer preferences for product categories. Understanding these preferences can segment customers effectively, which is essential for targeted marketing.

Customer Segmentation through Decision Tree Analysis:

- **Behavioral Patterns:**
  The decision trees have likely used purchase history and customer demographics to determine patterns. For instance, certain demographic groups may prefer electronics, while others lean towards beauty products. Identifying these patterns allows the company to segment customers not just by demographic variables like age or income, but by their behavior and preferences.

- **Targeted Marketing Campaigns:**

Once segments are identified, marketing campaigns can be designed to appeal specifically to the interests of those segments. For example, if one segment shows a strong preference for high-tech electronics, marketing emails to this group could feature the latest gadgets and tech reviews.

- **Customized Promotions:**
  Promotional strategies can be tailored to the preferences of each segment. A segment that frequently purchases groceries might be targeted with weekly deals, loyalty discounts, or reminders of their usual purchases, while beauty enthusiasts might receive updates on the latest skincare trends and exclusive invitations to product launches.

Strategic Recommendations:

1. Refine Marketing Strategies:
Utilize the insights from the customer segmentation to develop personalized marketing campaigns. For example, target customers who are identified as likely to purchase beauty products with relevant offers and promotions.

2. Product Placement and Inventory:
Adjust inventory levels based on the predictions of the decision tree models to ensure products in demand are readily available, especially in categories like electronics and beauty where clear preferences are identified.

3. Cross-Selling Opportunities:
Employ association rule mining to discover which products are often purchased together. Use this information for bundling products and offering discounts on complementary items to increase basket size.

4. Prevent Overfitting:
Investigate the decision tree model that showed a decline in performance on the training set as depth increased. Simplify the model if necessary to improve its generalizability, and ensure that it is trained on a representative sample of the customer base.

5. Continuous Model Evaluation:
Regularly review and update the decision tree models with new data to ensure that they adapt to changing customer behaviors and market trends.