FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

MASTER OF DATA SCIENCE

WQD7005 DATA MINING

Case Study

Predicting whether e-commerce customers use coupons

Matric No.	Name
22078878	Cai Jue

Selection of the Dataset

Data Source：https://www.kaggle.com/datasets/rishikumarrajvansh/marketing-insights-for-e-commerce-company/data?select=Online_Sales.csv

The Kaggle link primarily contains five datasets, which can be used for customer behavior analysis and market forecasting. According to the case study requirements, I have chosen two datasets for mapping and merging. They are the CustomersData and Online_Sales datasets.

CustomersData：

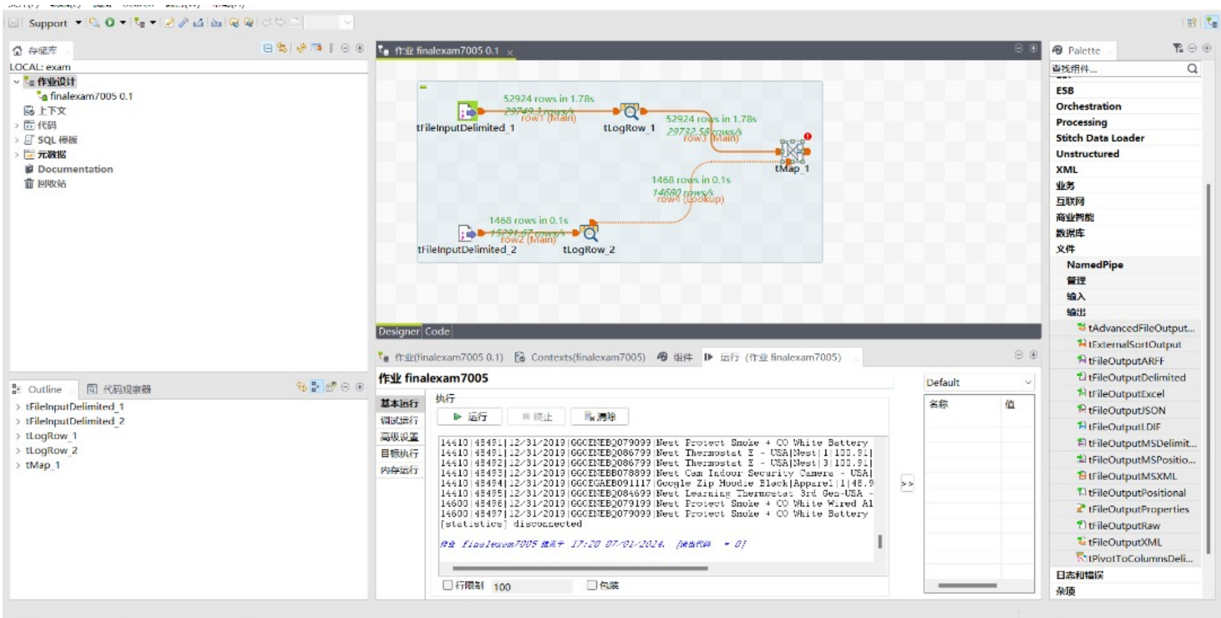
Variable	Description	Vs Dataset Structure	Decision
CustomerID	This is a unique identifier used to distinguish each customer.	CustomerID	Reserved
Gender	Customer's gender	Gender	Reserved
Location	Geographic location of the customer	Location	Reserved
Tenure_Months	The duration of the customer's account with the store, measured in months.	MembershipLevel	Reserved

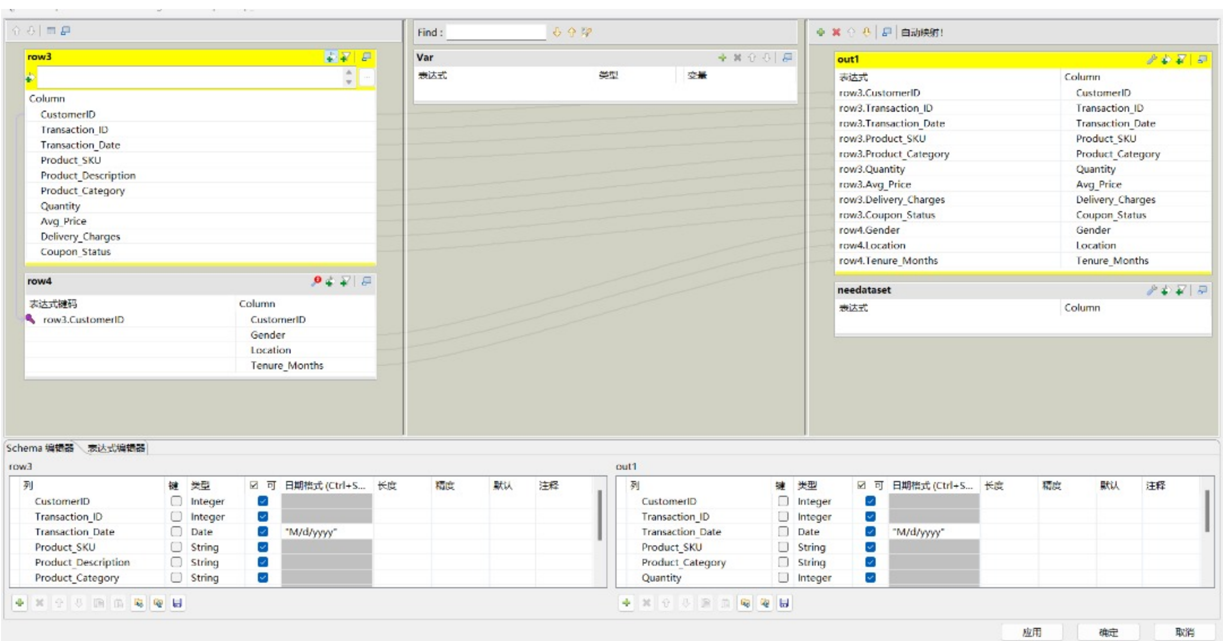
Online_Sales：

Variable	Description	Vs Dataset Structure	Decision
CustomerID	This is a unique identifier used to distinguish each customer.	CustomerID	Reserved
Transaction_ID	Customer's gender		Reserved
Transaction_Date	Geographic location of the customer	LastPurchaseDate	Reserved
Product_SKU	Transaction Unique ID		Reserved
Product_Description	Product Description		Drop
Product_Cateogry	Product Category	FavoriteCategory	Reserved
Quantity	Number of items ordered	TotalPurchases	Reserved
Avg_Price	Price per one quantity	TotalSpent	Reserved
Delivery_Charges	Charges for delivery	TotalSpent	Reserved
Coupon_Status	Any discount coupon applied	churn	Reserved

Tool 1: Use Data Integration to export the new dataset, named newdataset.

Combine the two datasets to ultimately obtain a dataset with 52,924 rows and 12 attributes for the prediction of coupon usage.

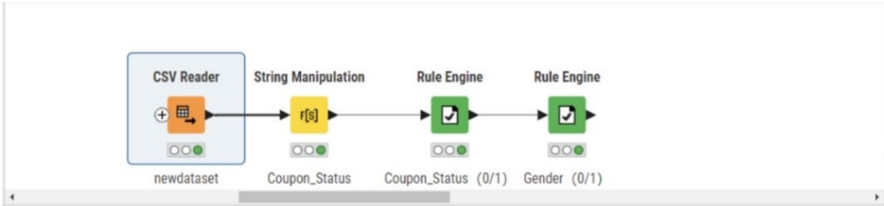




Tool 2: Use KNIME to transform and process the newdataset.

1. Check the null value situation of the newdataset.

   Null value： none



1: File Table | Flow Variables

Rows: 12 | Columns: 14 | Table | Statistics

Name	Type	# Missing val...	# Unique valu...	Minimum	Maximum	25% Quantile	50% Quantile ...	75% Quantile	Mean
CustomerID	Number (inte...	0	1468	12,346	18,283	13,869	15,311	16,996.75	15,346.71
Gender	String	0	2	?	?	?	?	?	?
Location	String	0	5	?	?	?	?	?	?
Tenure_Months	Number (inte...	0	49	2	50	15	27	37	26.128
Transaction_ID	Number (inte...	0	25061	16,679	48,497	25,384	32,625.5	39,126.75	32,409.826
Transaction_D...	String	0	365	?	?	?	?	?	?
Product_SKU	String	0	1145	?	?	?	?	?	?
Product_Cate...	String	0	20	?	?	?	?	?	?
Quantity	Number (inte...	0	151	1	900	1	1	2	4.498
Avg_Price	Number (dou...	0	546	0.39	355.74	5.7	16.99	102.13	52.238
Delivery_Char...	Number (dou...	0	267	0	521.36	[illegible]	[illegible]	[illegible]	10.518
Coupon_Status	String	0	3	?	?	?	?	?	?

没有正在运行的虚拟机

2. Convert the Coupon_Status from a ternary classification: Used, Not Used, Clicked, into a binary classification of Used and Not Used, where Clicked is considered as Not Used, in order to make the data more balanced.

Expression

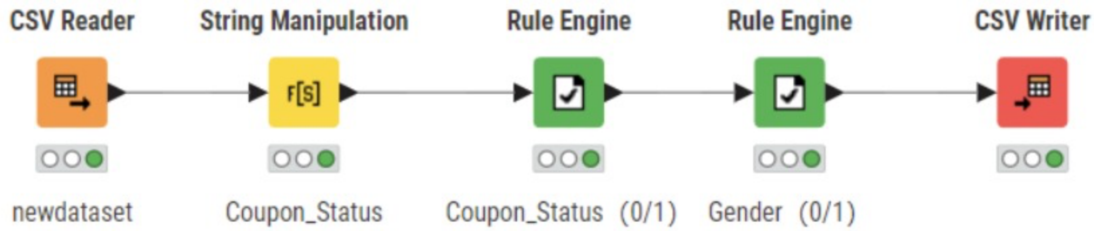
```
replace($Coupon_Status$, "Clicked", "Not Used")
```

3. Conduct some conversions from string to numerical values to facilitate better modeling, such as binary variables for gender and Coupon.

```
$Gender$ = "F" => "1"
$Gender$ = "M" => "0"
TRUE => $Gender$
```

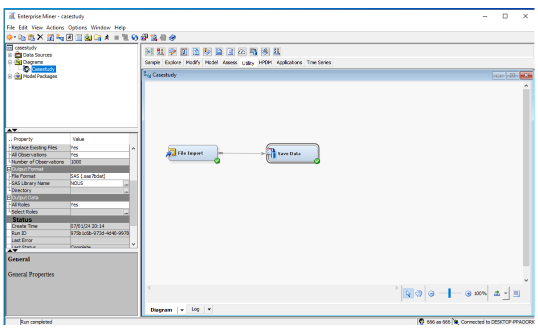
```
$Coupon_Status$ = "Used" => "1"
$Coupon_Status$ = "Not Used" => "0"
TRUE => $Coupon_Status$
```

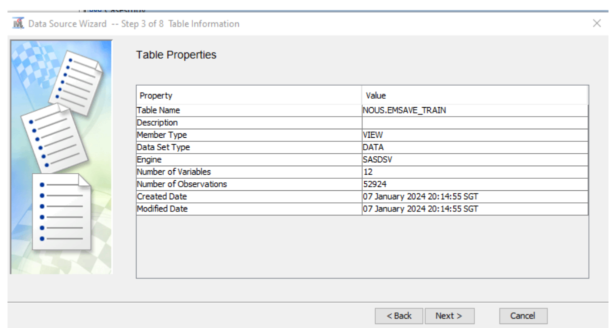
4. The overall workflow, exported as datasetfinal.csv.

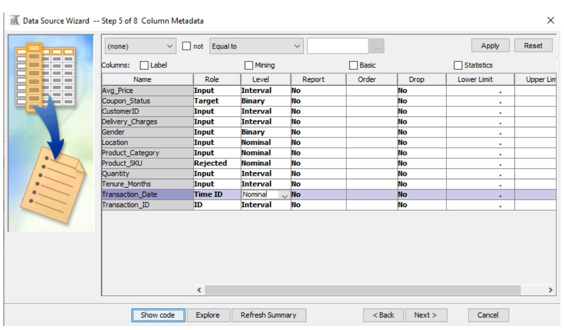


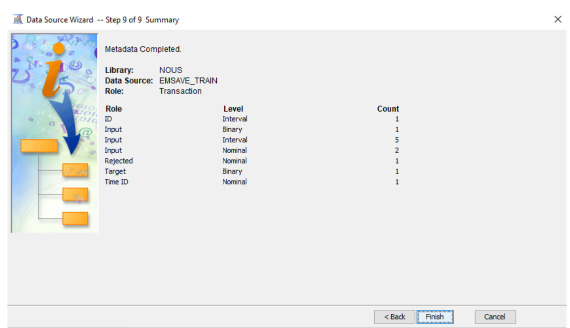
Tool 3: SAS

1. **Data Import and Preprocessing:** Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles.
Import the dataset 'datasetfinal.csv'using 'File Import', and save it in the library of Nous using 'Save Data'. Create a new data source using 'Advanced Table' and adjust the role.









Use statexplore to check the specific data situation. It is consistent with what I explored before. There are no null values. However, I found that the 1 and 0 of the targets may not be so balanced, which may cause problems during the modeling process and need to be adjusted.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Gender	INPUT	2	0	1	62.37	0	37.63
TRAIN	Location	INPUT	5	0	Chicago	34.73	California	30.49
TRAIN	Product_Category	INPUT	20	0	Apparel	34.25	Nest-USA	26.48
TRAIN	Coupon_Status	TARGET	2	0	0	66.17	1	33.83

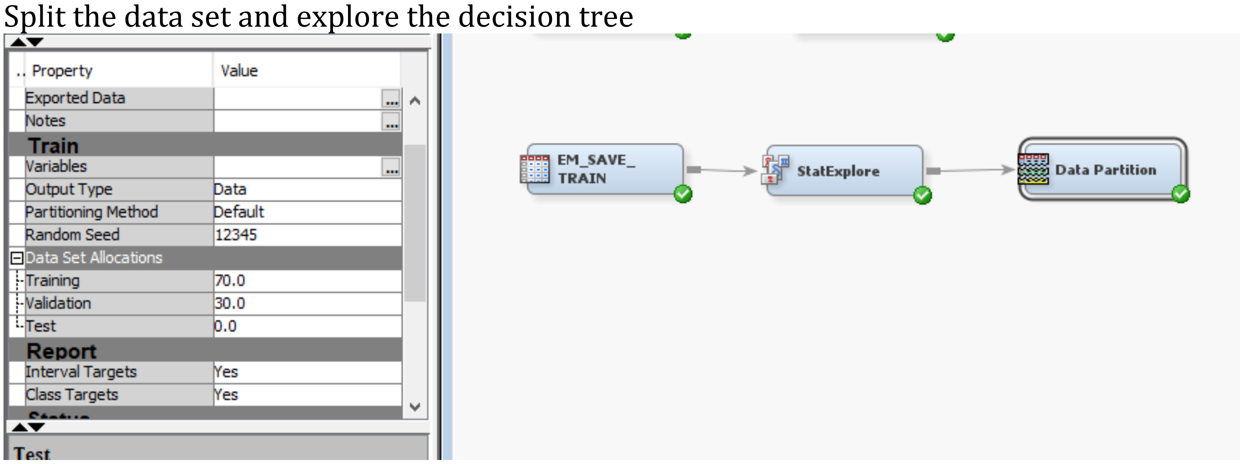
Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

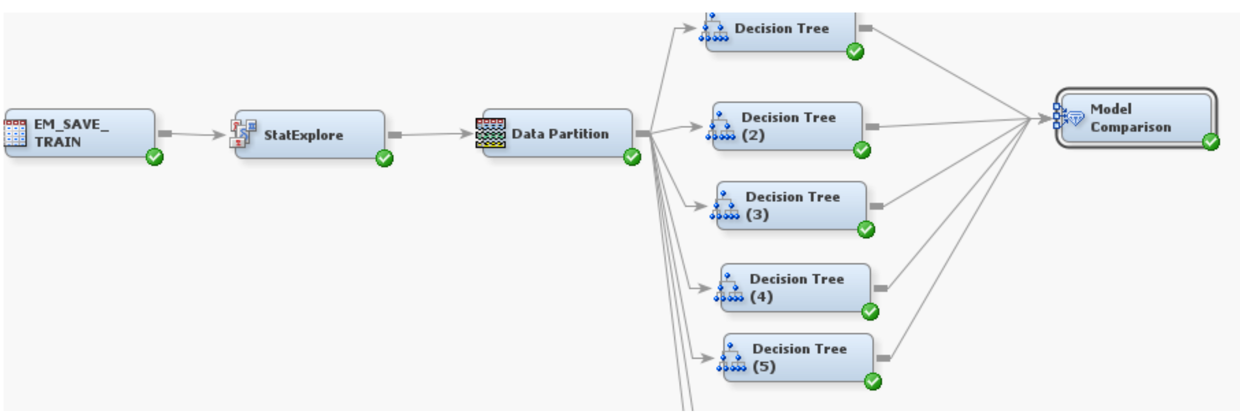
Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	Coupon_Status	TARGET	0	35020	66.1704
TRAIN	Coupon_Status	TARGET	1	17904	33.8296

Decision Tree Analysis: Create a decision tree model in SAS Enterprise Miner to analyze customer behavior.

Split the data set and explore the decision tree





Obs	TARGET	TARGETLABEL	_AUR_	_GINI_	KS	_KS_PROB_ CUTOFF	_KS_BIN_	BINNED_KS_ PROB_ CUTOFF
1	Coupon_Status		0.5	0	0	.	0	0.338

Obs	TARGET	TARGETLABEL	_VAUR_	_VGINI_	VKS	_VKS_ PROB_ CUTOFF_	_VKS_ BIN_	_VBINNED_ KS_PROB_ CUTOFF_
1	Coupon_Status		0.5	0	0	.	0	0

The effect of using different nodes and depths is not good.These statistical metrics collectively indicate that the model's performance is very limited and almost equivalent to random guessing. An AUR value of 0.5, a Gini coefficient of 0, and a KS statistic of 0 all point out the model's lack of effective discriminatory power to differentiate between positive and negative samples. A high misclassification rate of 33.83% means that about one-third of the predictions are incorrect, while the high values of the Average Squared Error and Maximum Absolute Error suggest significant deviations between the predictions and actual outcomes. Additionally, a Roc Index value of 0.50 and Lift and Cumulative Lift values of 1 further confirm the model's predictive capacity does not surpass random levels. Overall, these indicators converge on a common conclusion: the model demonstrates poor performance in terms of prediction accuracy and effectiveness.

Coupons are something that requires more consideration. It is not beneficial for companies to make decisions about whether to use them. The current data shows that customers are more random.

Ensemble Methods: Apply Bagging and Boosting, using the Random Forest algorithm as a Bagging example.